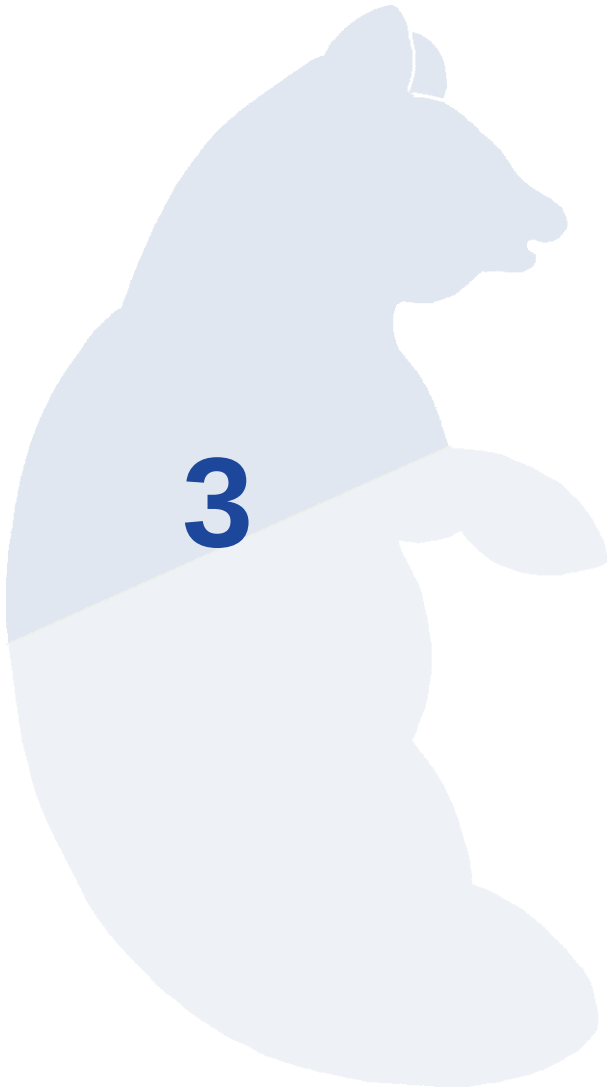


3

# Clustering
# Métriques

# 3. Clustering : métriques

***Plusieurs métriques sont utiles pour obtenir le nombre de clusters optimal :***

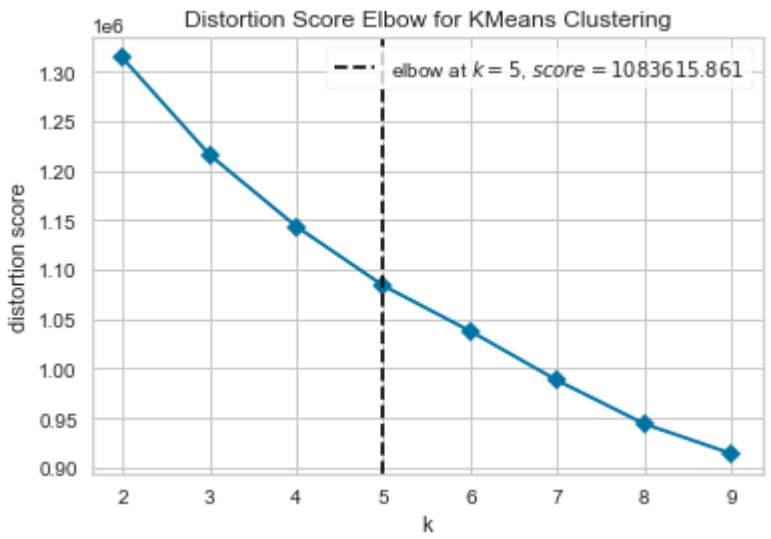
• ***Forme et homogénéité des clusters*** :

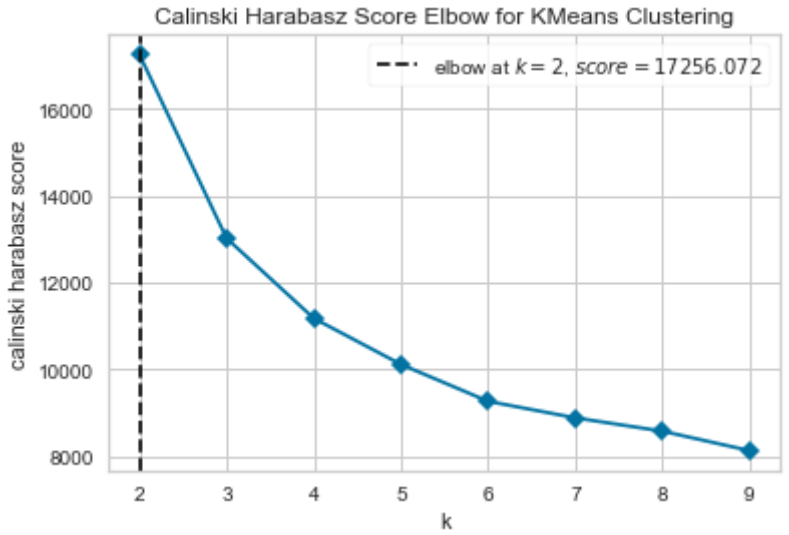
- **méthode dite du coude**,
- **coefficient de silhouette** : une mesure de qualité d'une partition d'un ensemble de données. Pour chaque point, son coefficient de silhouette est la différence entre la distance moyenne avec les points du même groupe que lui (cohésion, homogénéité) et la distance moyenne avec les points des autres groupes voisins (séparation). Si cette différence est négative, le point est en moyenne plus proche du groupe voisin que du sien : il est donc mal classé. A l'inverse, si cette différence est positive, le point est en moyenne plus proche de son groupe que du groupe voisin : il est donc bien classé. Le coefficient de silhouette proprement dit est la moyenne du coefficient de silhouette pour tous les points (-1 pire assignation du cluster, 1 : assignation satisfaisante).
- **indice de Davies Bouldin** : introduite par David L. Davies et Donald W. Bouldin en 1979, est une métrique d'évaluation des algorithmes de clustering. Il s'agit d'un schéma d'évaluation interne, où la validation de la qualité du clustering est faite en utilisant des quantités et des caractéristiques inhérentes au jeu de données (dispersion). Cette méthode présente l'inconvénient qu'une bonne valeur rapportée par cette méthode n'implique pas la meilleure recherche d'information. C'est la moyenne du rapport maximal entre la distance d'un point au centre de son groupe et la distance entre deux centres de groupes. L'indice de Davies-Bouldin varie entre 0 (meilleure classification) et +infini (pire classification).

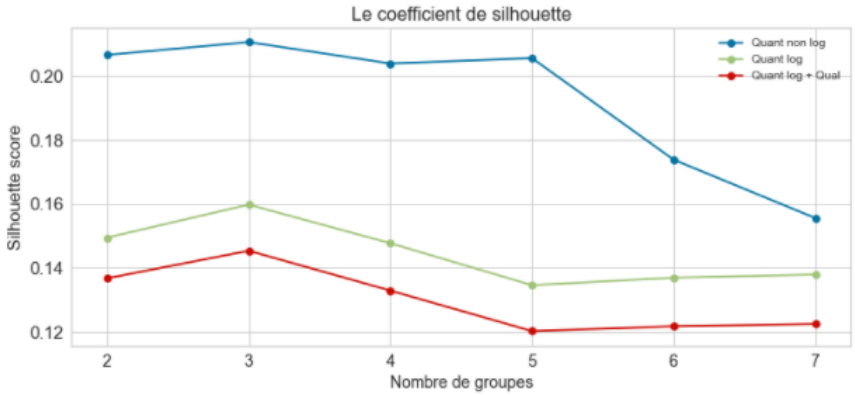
•

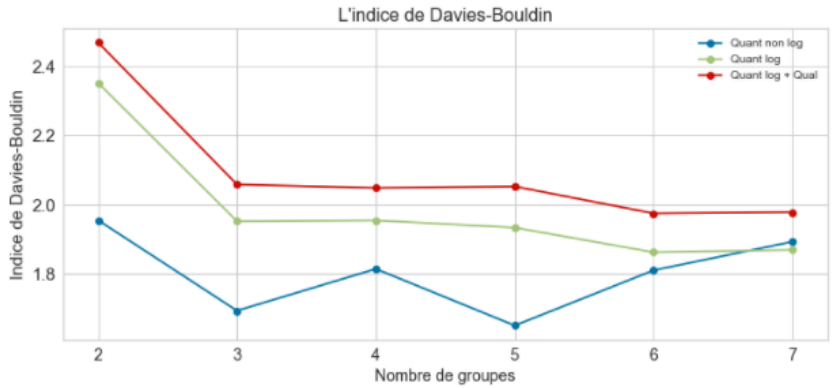
# 3. Clustering : métriques

- ***Stabilité des clusters*** :
  - **Indice de Rand ajusté** : L'indice Rand calcule une mesure de similarité entre deux clusters en considérant toutes les paires d'échantillons et en comptant les paires qui sont assignées dans les mêmes ou différents clusters dans les clusters prédits et réels. L'indice de Rand ajusté est donc assuré d'avoir une valeur proche de 0,0 pour un étiquetage aléatoire indépendamment du nombre de grappes et d'échantillons et exactement 1,0 lorsque les grappes sont identiques (jusqu'à une permutation).

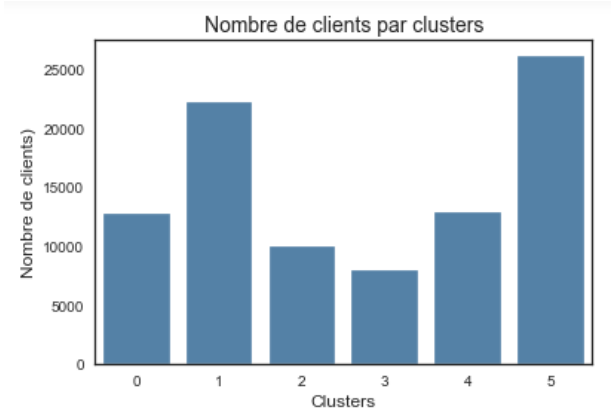


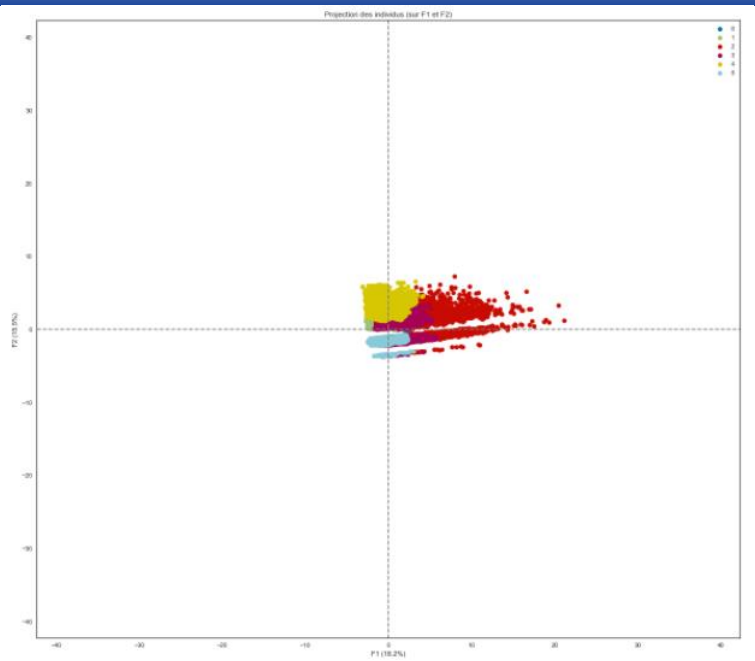



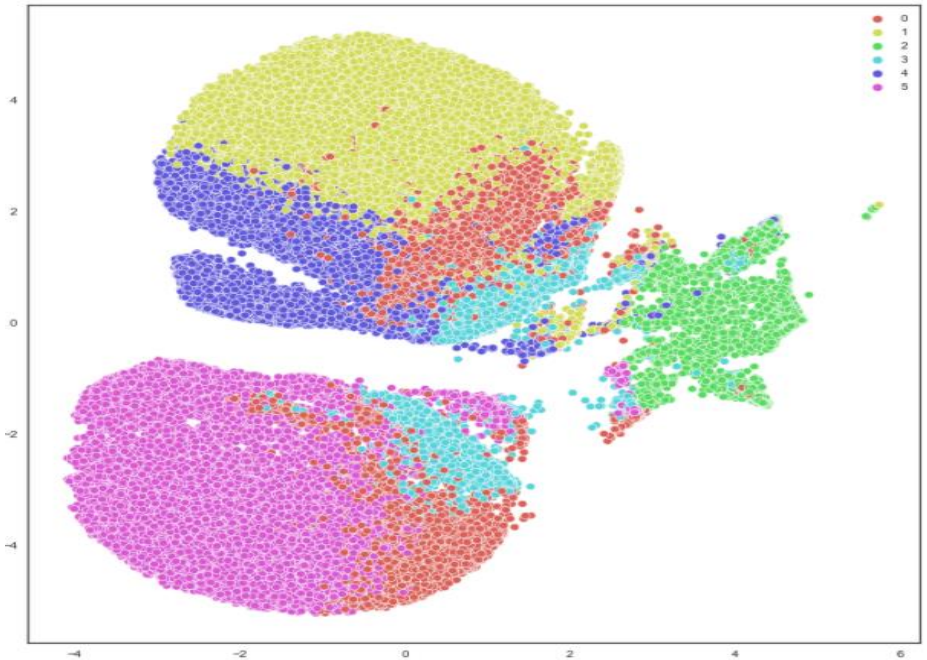



# 3. Clustering : métriques

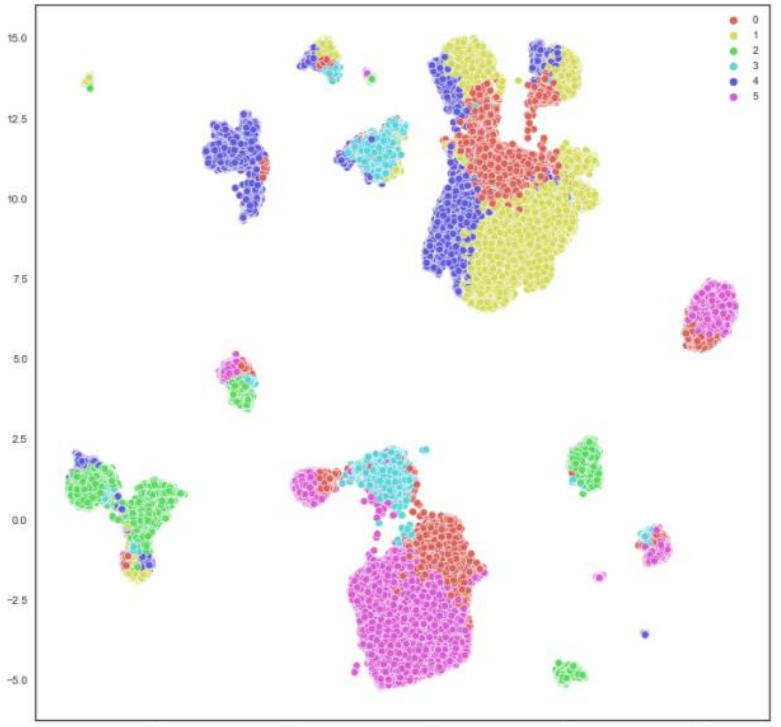


**ACP**




**t-SNE**


**UMAP**

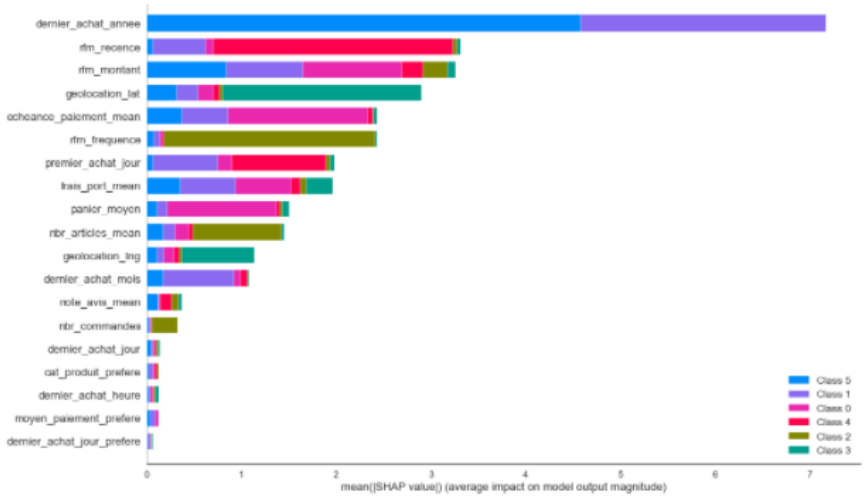
# 3. Clustering : métriques

**Evaluation en effectuant une classification avec Lightbm**

•Une autre comparaison consiste à traiter les clusters comme des étiquettes et à construire un modèle de classification par-dessus.
•Si les clusters sont de haute qualité, le modèle de classification sera capable de les prédire avec une grande précision. En même temps, les modèles doivent utiliser une variété de caractéristiques pour s'assurer que les clusters ne sont pas trop simplistes.
•La vérification portera sur les attributs suivants :

- Distinctivité des clusters par le score F1 validé de manière croisée. 'f1_weighted' : calcule les métriques pour chaque étiquette et trouve leur moyenne pondérée par le support (le nombre d'instances vraies pour chaque étiquette). Cela modifie la «macro» pour tenir compte du déséquilibre des étiquettes; il peut en résulter un score F qui ne se situe pas entre la précision et le rappel.
- Informativité des clusters par l'importance des caractéristiques SHAP

•En utilisant le modèle LightGBM comme classificateur car il peut utiliser des caractéristiques catégorielles pour obtenir les valeurs SHAP pour les modèles formés et connaître les variables les plus importantes.
[Source]



# 3. Clustering : métriques

## Evaluation d'un clustering (1)

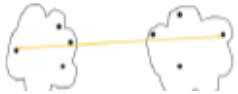
- Centroïde/Barycentre d'un cluster $k$ (avec $n_k = |C_k|$):
  - $\mu_k = \frac{1}{n_k} \sum_{x_i \in c_k} x_i$


- **Homogénéité**
  - Qualité intra cluster
    - pour un cluster $k$ : moyenne des distances entre chaque point et le centre :
      - Tightness : $T_k = \frac{1}{n_k} \sum_{x \in C_k} d(x, \mu_k)$
      - Un cluster $k$ homogène a une valeur $T_k$ faible
    - sur l'ensemble des $K$ clusters : moyenne des $T_k$ : $T = \frac{1}{K} \sum_{k=1}^{K} T_k$

## Evaluation d'un clustering (2)

- **Séparation**
  - Qualité inter cluster : distance entre deux clusters
  - Saut minimal (single linkage) : 2 clusters sont proches si 2 de leurs points sont proches
    - Distance minimale entre 2 points appartenant à des clusters différents
      - $d(C_k, C_l) = min_{x \in C_k, y \in C_l} d(x, y)$

  - Saut maximal (complete linkage) : 2 clusters sont proches si tous leurs points sont proches
    - Distance maximale entre 2 points appartenant à des clusters différents
      - $d(C_k, C_l) = max_{x \in C_k, y \in C_l} d(x, y)$



## Evaluation d'un clustering (3)

- **Séparation**
  - Saut moyen (average linkage)
    - Distance moyenne entre toutes les paires de points
      - $d(C_k, C_l) = \frac{1}{n_k \times n_l} \sum_{x \in C_k} \sum_{y \in C_l} d(x, y)$

  - Distance entre les barycentres
    - $d(C_k, C_l) = d(\mu_k, \mu_l) = d(\frac{1}{n_k} \sum_{x \in C_k} x, \frac{1}{n_l} \sum_{y \in C_l} y)$

    - Sur les $K$ clusters : moyenne distance deux à deux
      - $S = \frac{2}{K(K-1)} \sum_{k=1}^{K} \sum_{l=1}^{K} d(C_k, C_l)$
    - Avoir une valeur élevée

# 3. Clustering : métriques

**Les métriques utilisées** :

•**Coefficient de silhouette** : mesure la différence entre la distance quadratique moyenne intra-groupe et celle du groupe le plus proche (à maximiser). différence entre les distances intra-cluster et les distances au cluster extérieur le plus proche (*à maximiser*)

•**Indice de Davies_Bouldin** : mesure le rapport maximal de la dispersion des paires de clusters par rapport à leur distance (à minimiser). moyenne du rapport maximal entre la distance d'un point au centre de son groupe et la distance entre deux centres de groupes (*à minimiser*)

•**Score de Calinski_Harabasz** : le rapport entre la dispersion inter-clusters et la dispersion intra-clusters (à maximiser). rapport entre la variance inter-groupes et la variance intra-groupe (*à maximiser*)

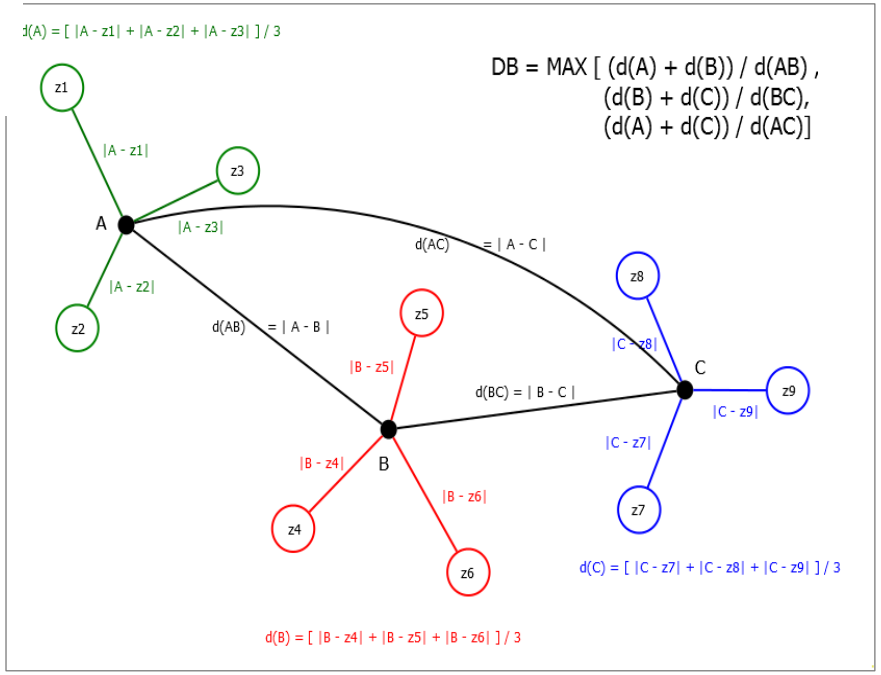
•**Distorsion** : la moyenne de la somme des carrés des distances au centroïde le plus proche (*méthode du coude*)

# 3. Clustering : métriques

- **Combinaison des deux critères (homogénéité et séparation)**
  - Indice de Davies Bouldin
    - pour un cluster $k$ :
      - Combiner les valeurs des tightness (intra) et distances entre les barycentres (inter):
      - $DB_k = max_{l \neq k}(\frac{T_k + T_l}{d(C_k, C_l)})$
    - Valeur faible si les clusters sont homogènes (numérateur petit) et s'ils sont bien séparés (dénominateur grand)
    - pour tous les clusters : $DB = \frac{1}{K} \sum_{k=1}^{K} DB_k$
  - Aide pour déterminer le nombre de clusters $K$ (minimiser $DB$)

# Davies-Bouldin score
## (à minimiser)

C'est la moyenne du rapport maximal entre la distance des points à leur centroïde et la distance entre deux centroïdes (deux à deux)
⇨ *entre 0 (meilleur) et +inf (pire)*




*Note: formules de l'illustration avec la distance Manhattan*

distance moyennes
des points à leur centroïde

$$\bar{\delta}_k = \frac{1}{|I_k|} \sum_{i \in I_k} d(x^i, \mu_k)$$

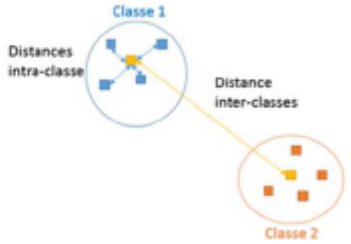
$$S_{DB} = \frac{1}{K} \sum_{k=1}^{K} \max_{k' \neq k} \left( \frac{\bar{\delta}_k + \bar{\delta}_{k'}}{d(\mu_k, \mu_{k'})} \right)$$

distance entre deux centroïdes

# 3. Clustering : métriques

## Clustering

- Définir une distance, une similarité
- Distance intra-cluster/classe (minimiser)
- Distance inter-cluster/classe (maximiser)




## Silhouette : autre évaluation pour homogénéité et séparation

- Pour chaque point $x$ : est-ce qu'il appartient au « bon » cluster :
  - Est-il proche des points du cluster auquel il appartient ?
    - distance moyenne aux autres points du même cluster
    - $a(x) = \frac{1}{n_k - 1} \sum_{y \in c_k, y \neq x} d(x, y)$
  - Est-il loin des points des autres clusters ?
    - Distance moyenne minimale si $x$ affecté dans un autre cluster?
    - $b(x) = min_{l \neq k} \frac{1}{n_l} \sum_{y \in c_l} d(x, y)$
  - Si le point $x$ est dans le bon cluster : $a(x) < b(x)$
  - Silhouette : combinaison des deux scores : $s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$
    - compris dans [-1, 1])

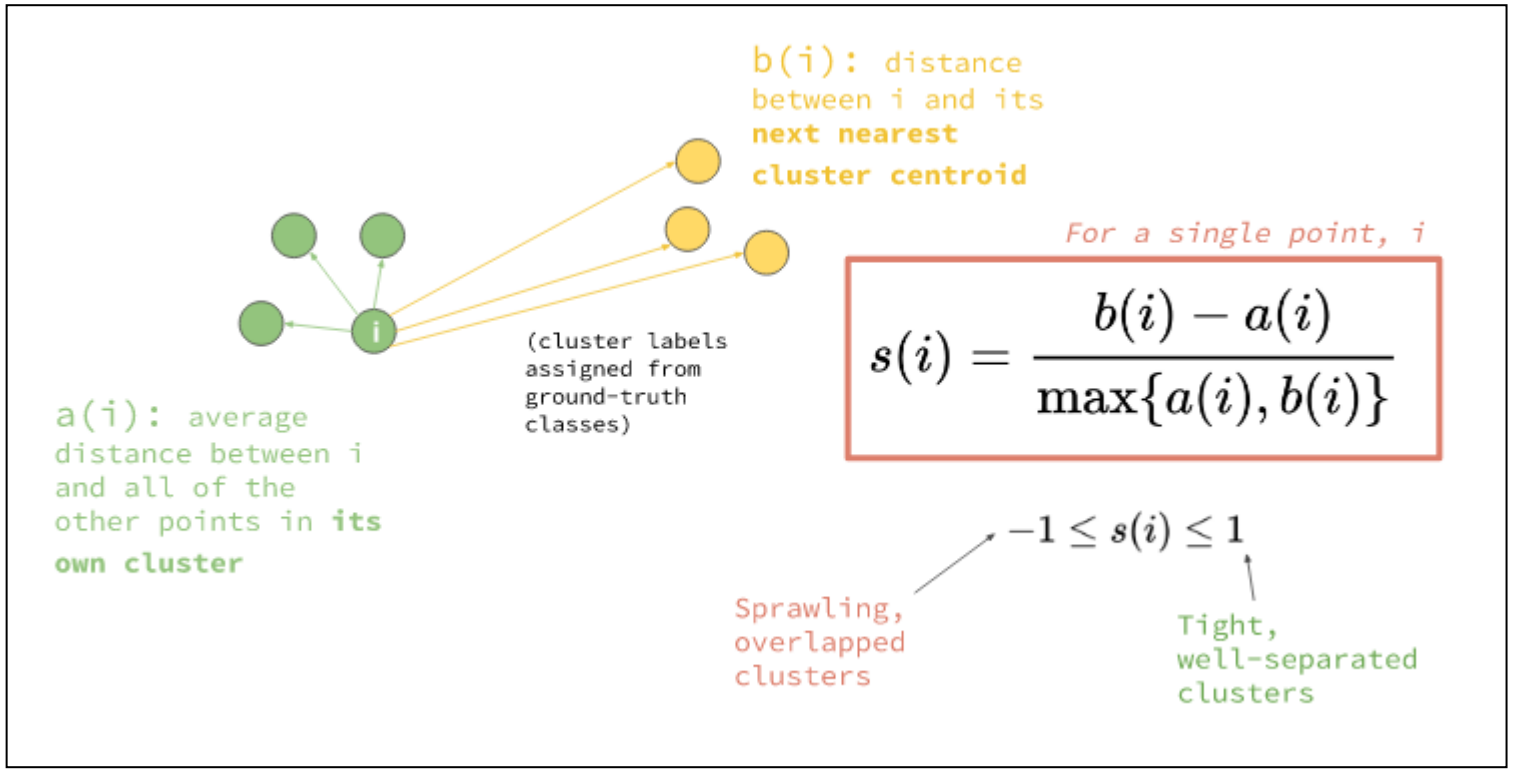
*Au moins 2 points et 2 clusters …*

- Pour tous les points : $S = \frac{1}{n} \sum s(x)$
- Aide pour déterminer le nombre de clusters $K$ (minimiser $S$)

# Silhouette score

## (à maximiser)

C'est la **différence entre les** distances intra-cluster **et les** distances au cluster extérieur le plus proche **(rapportée à la plus grande des deux)** ⇨ *entre 0 (pire) et 1 (meilleur)*



# Calinsky-Harabasz score
## (à maximiser)

C'est le rapport entre la variance inter-groupes et la variance intra-groupe.

nb d'individus — nb de clusters

$$S_{CH} = \frac{(N-K)B}{(K-1)\sum_{k=1}^{K} W_k}$$

variance inter-groupes

$$B = \sum_{k=1}^{K} |I_k| \|\mu_k - \mu\|$$

centroïdes — centre global

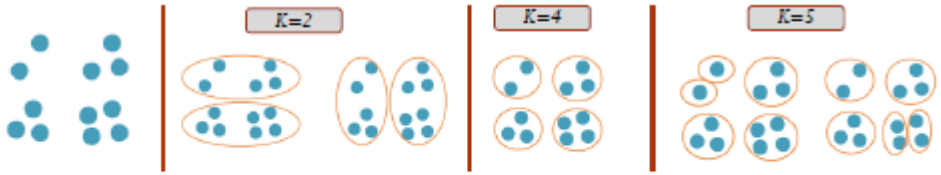
variances intra-groupes

$$W_k = \frac{1}{|I_k|} \sum_{i \in I_k} \|x^i - \mu_k\|$$

# 3. Clustering : métriques

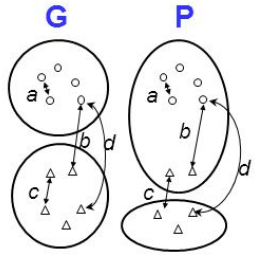
## Stabilité d'un clustering

- **Stabilité**
  - Nombreux algorithmes non déterministes : résultats différents lors d'exécutions différentes de l'algorithme
    - lancer l'algorithme plusieurs fois sur les mêmes données (éventuellement bruitées), avec initialisation différente, avec des sous-ensembles différents,
  - Est-ce que les points sont regroupés de manière similaire ?




- Mesure pour évaluer cette similarité .... (sans dépendre de la numérotation des clusters) ... Voir Indice de Rand

# ARI score
## (à maximiser)




Agreement: $a, d$
Disagreement: $b, c$

$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

Indice de Rand ajusté pour le hasard.

L'indice de Rand calcule une mesure de similarité entre deux clusters en considérant toutes les paires d'échantillons et en comptant les paires qui sont assignées dans les mêmes ou différents clusters dans les clusters prédits et réels.

Le score brut RI est ensuite "ajusté pour le hasard" dans le score ARI en utilisant le schéma suivant :

ARI = (RI - Expected_RI) / (max(RI) - Expected_RI).

L'indice de Rand ajusté est ainsi assuré d'avoir une valeur proche de 0,0 pour un étiquetage aléatoire indépendamment du nombre de clusters et d'échantillons et exactement 1,0 lorsque les clusters sont identiques (jusqu'à une permutation).

L'*Adjusted Rand Index* (ARI) est la normalisation de l'indice de Rand (RI) qui permet de comparer deux partitions de nombres de classes différentes.

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}$$

- RI : indice de Rand, proportion de paires de points qui sont groupés de la même façon dans les deux partitions.
- E(RI) : espérance de l'indice de Rand (pour une partition aléatoire)
- max(RI) : indice de Rand maximal qui pourrait être obtenu étant donné le nombre de classes distincts

# 3. Clustering : métriques

## Distances (1)

- **Distance : une fonction $d$ : $\mathbb{R} \rightarrow \mathbb{R}^+$ vérifiant :**
  - Symétrie : $d(x,y) = d(y,x)$
  - Séparation : $d(x,y) = 0 \iff x = y$
  - Inégalité triangulaire : $d(x,y) \leq d(x,z) + d(z,y) = d(y,x)$

- Distance de Minkowski ou Norme $L_q$

$$d(x_1, x_2) = \left(\sum_{j=1}^{d} |x_{1,j} - x_{2,j}|^q\right)^{\frac{1}{q}}$$

  - Si $q = 2$, distance euclidienne
  - Si $q = 1$, distance de Manhattan

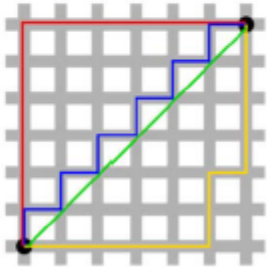


*Image:Wikipedia*

## Distances (2)

- **Distance de Hamming**
  - Mesurer différence entre deux séquences de symboles
    - Traitement du signal
  - Soit $x_i$ et $y_i$ deux observations de dimension $d$
  - Hamming : $h(x_i, y_i) = Card(\{j : x_{ij} \neq y_{ij}\})$

$$x_i = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,d} \end{pmatrix}$$

  - Exemple
    - Entre 1011101 et 1001001 → distance de Hamming = 2
    - Entre 2143896 et 2233796 → distance de Hamming = 3
    - Entre ramer et cases → distance de Hamming = 3

## Distances (3)

- **Distance de Levenshtein (distance d'édition)**
  - Mesurer la différence entre deux chaines de caractères
    - Nombre d'opérations élémentaires (insérer/supprimer/remplacer) pour passer d'une chaine source à une chaine destination
      - Passer de "a " vers "ab" : distance = 1 (insérer 'b')

- Autres : compter des n-grammes
  - Sous séquences de longueur $n$ présentes dans une séquence
  - Comparer des séquences à partir des n-grammes communs

# 3. Clustering : métriques

# 3. Clustering : métriques

# 3. Clustering : métriques

# 3. Clustering : métriques

# 3. Clustering : métriques