Exploration des données

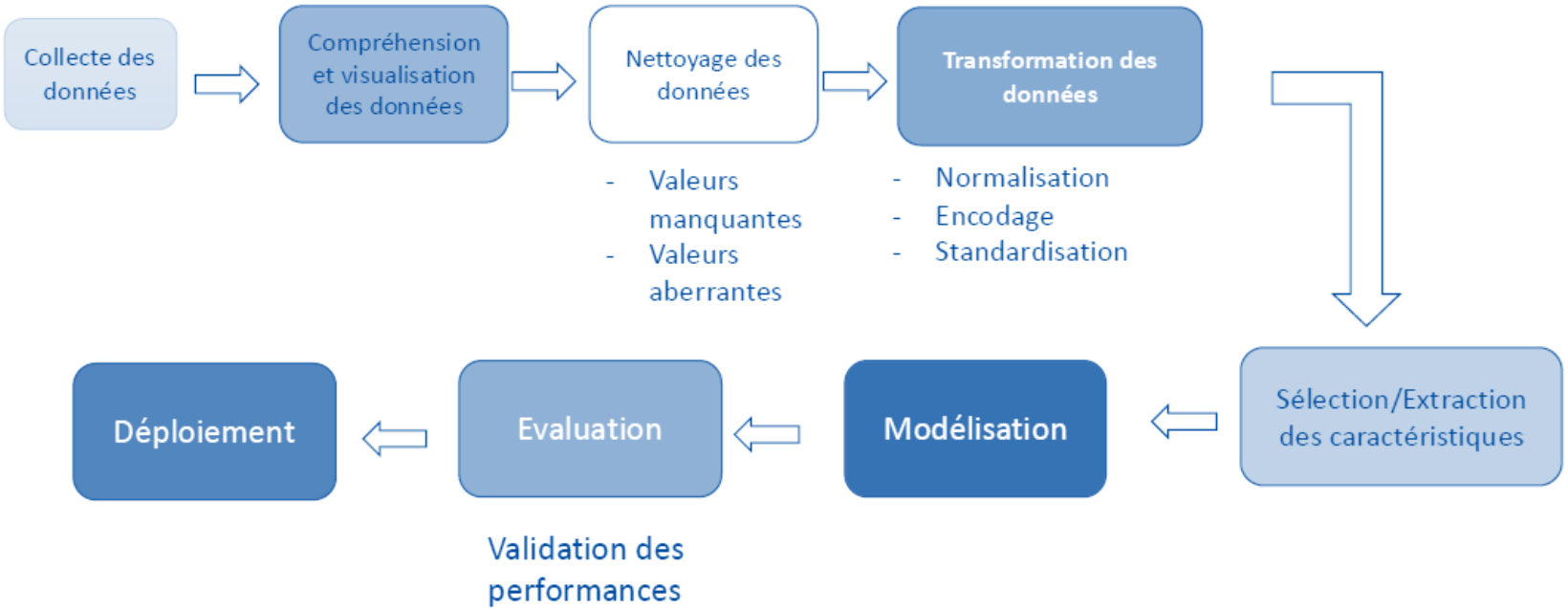
CM2 : Extraction des caractéristiques

Mickaël Martin Nevot

V1.0.0

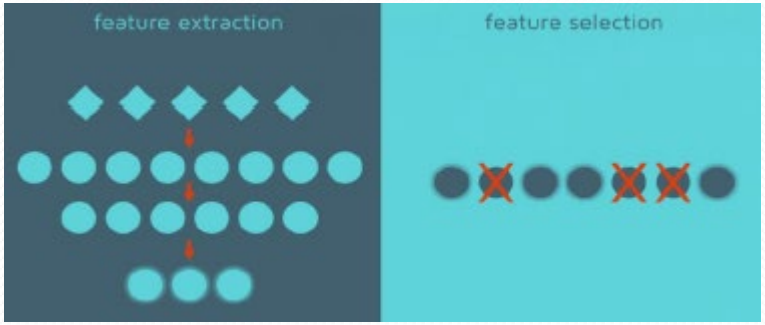
Exploration des données

Exploration : chaîne de traitement

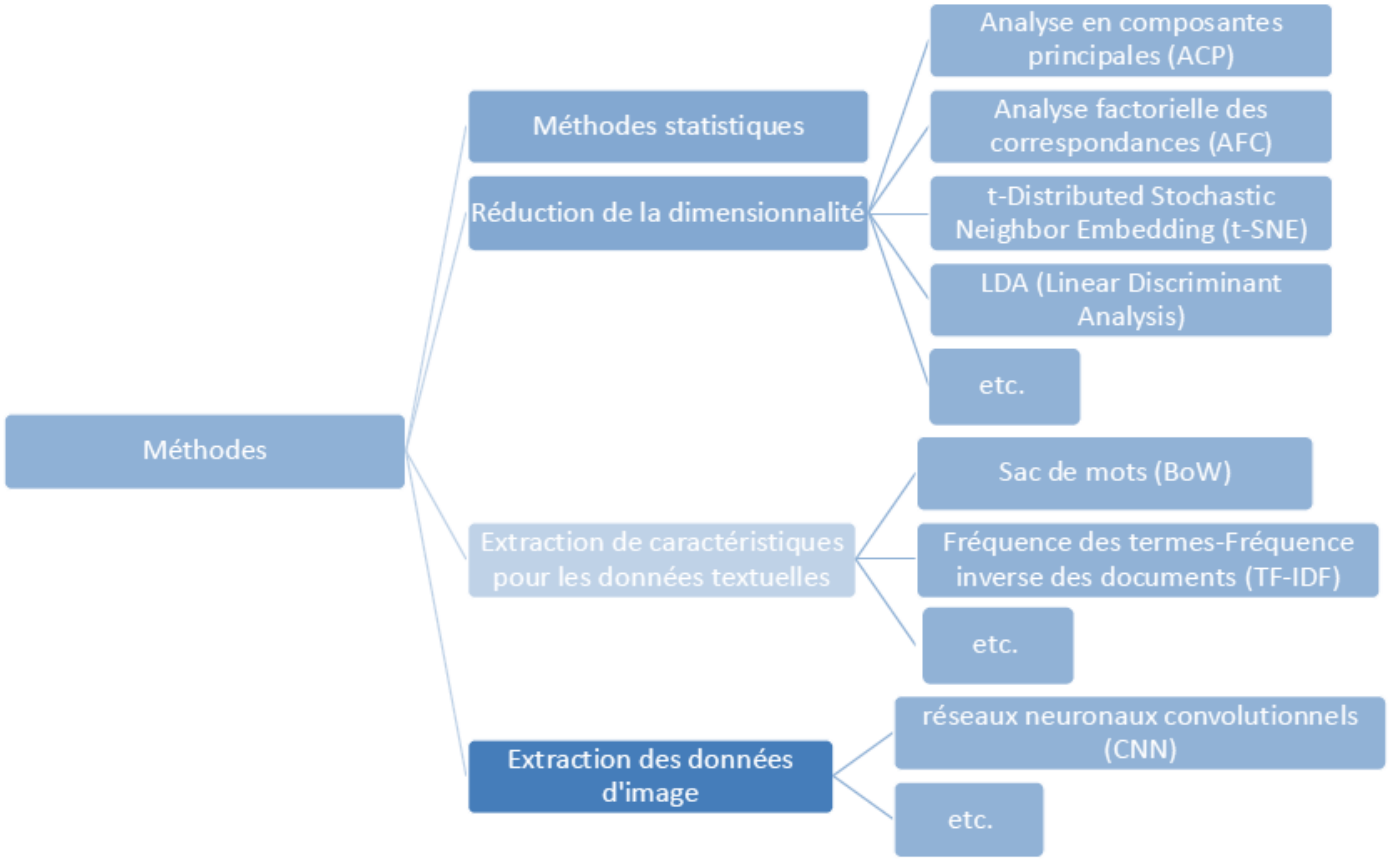


Sélection/extraction des var.

- **Extraction des variables** :
  - Créer de nouvelles variables à partir des données d'origine :
    - Réduction des dimensions
    - Capture d'information complexe
    - Amélioration de l'efficacité des traitements
- Sélection des variables (hors cours) :
  - Choisir un sous-ensemble des variables qui sont les plus pertinentes pour un modèle ou une tâche



Extraction des variables

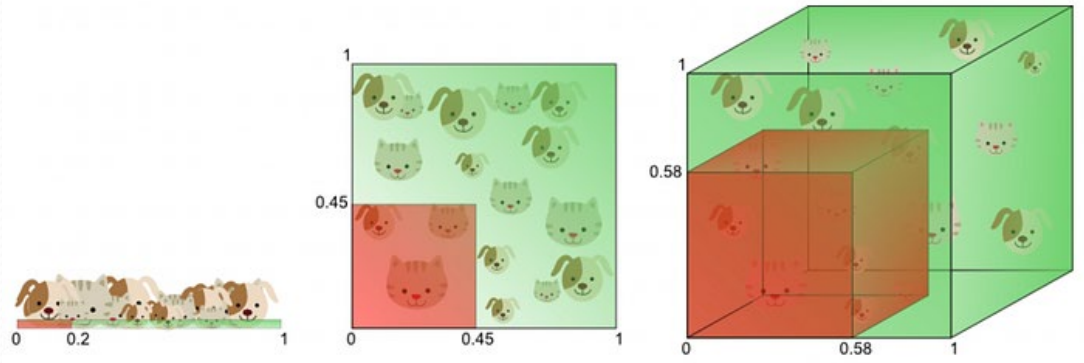


Réduction des dimensions

- Réduire le nombre de variables
- Éviter le fléau de dimension
- Réduction :
  - Utile pour les traitements ultérieurs
  - Utile pour la visualisation des données
- Projection (linéaire) : **ACP**
- Apprentissage de variété (non linéaire) : **UMAP**, **t-SNE**

Quantitatives

Peut engendrer une perte d'information



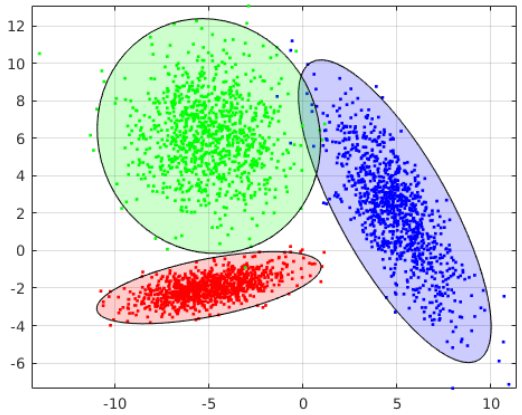
Algorithmes non supervisés

ACP

- Analyse en composantes principales
- Méthode linéaire
- Projection (linéaire)
- Transforme les **variables très corrélées en nouvelles variables décorrélées** les unes des autres
- **Identifie l'hyperplan des variables le plus proche** puis projette les données sur celui-ci

Algorithme de réduction de dimension le plus connu

Le plus rapide



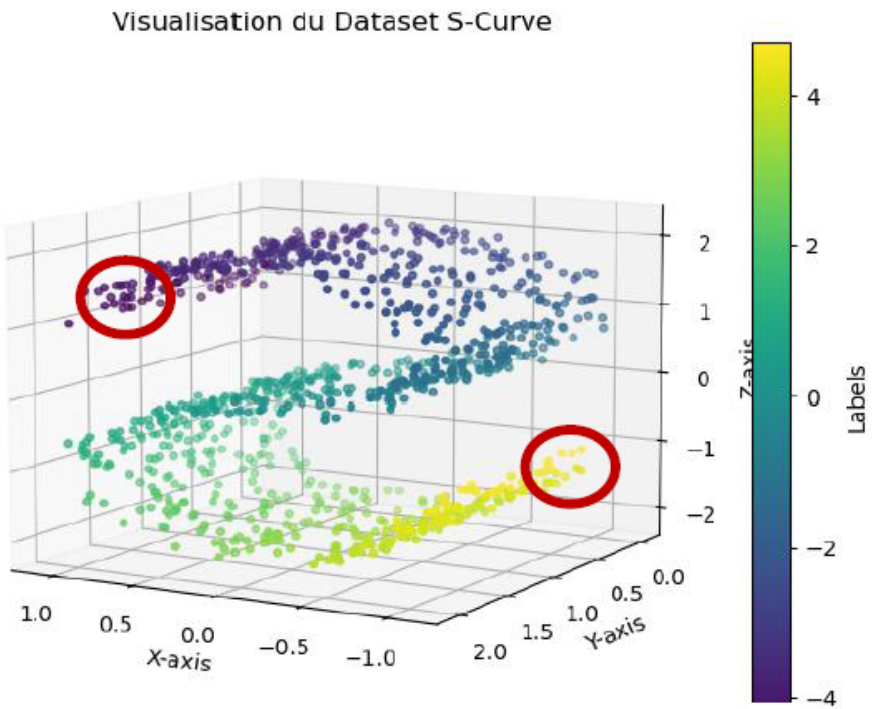
ACP

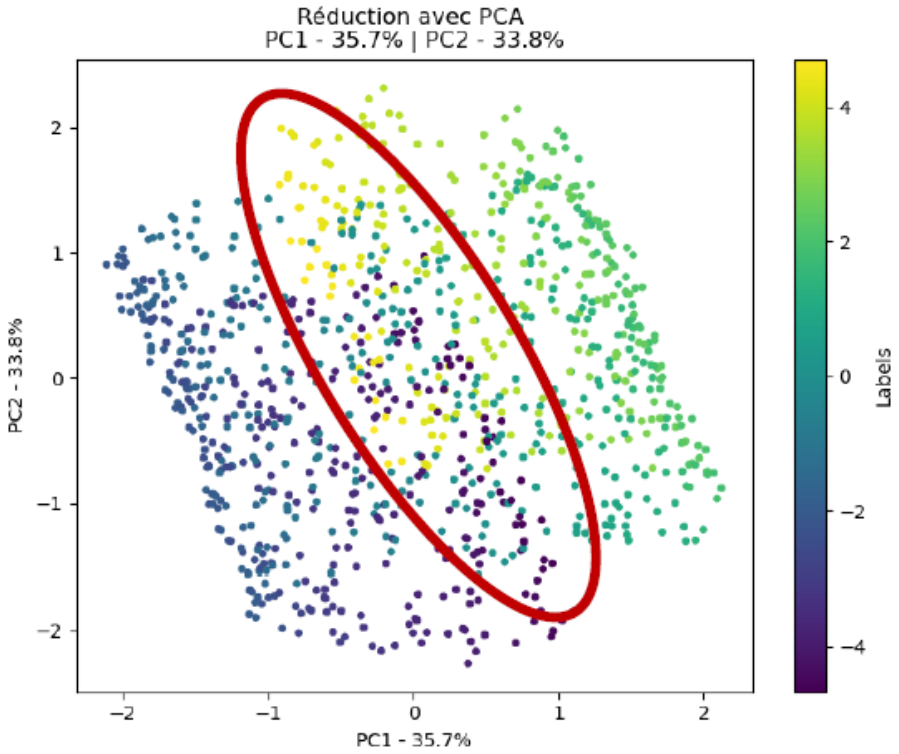
- Démarche

1. Centrage et Réduction des données
2. Calcul de la matrice de (co)-variance ou de corrélation
3. Calcul des valeurs et vecteurs propres
4. Détermination des axes factoriels et composantes principales

ACP

- Limitation : données non linéaires



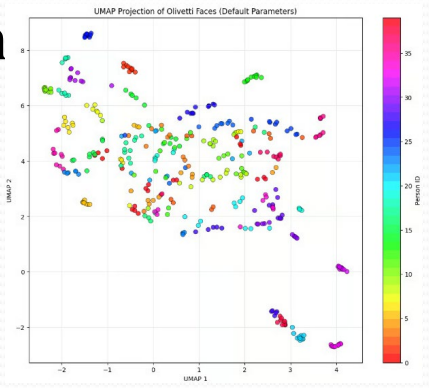


Les points se trouvent au même endroit, les distances entre les points sont faussées par l'ACP

UMAP

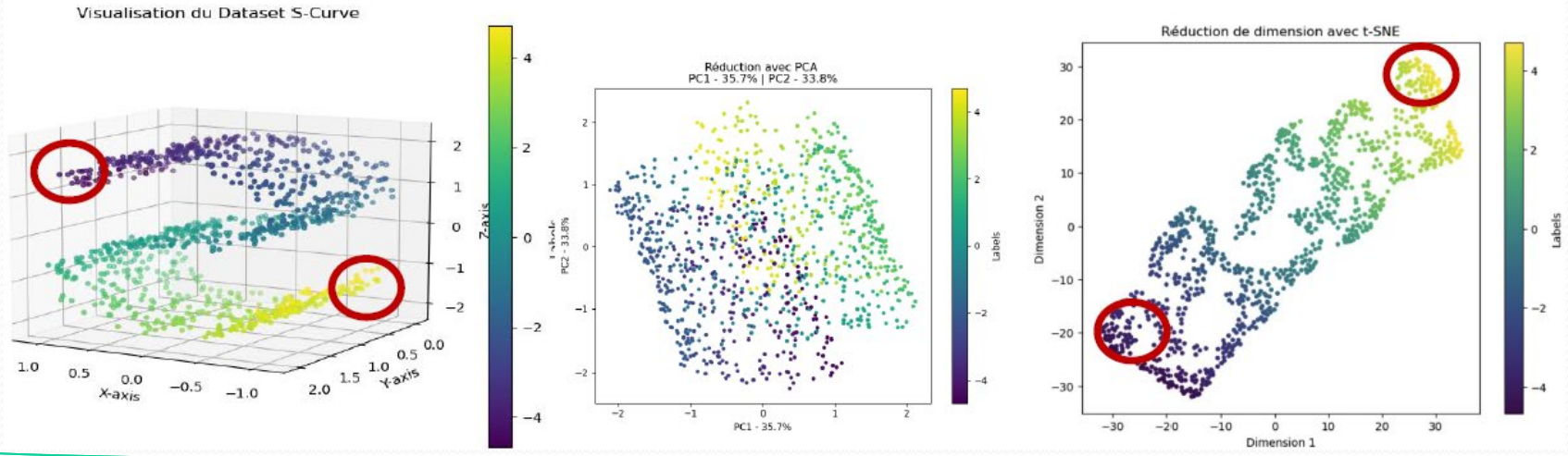
- *Uniform manifold approximation and projection*
- Méthode non-linéaire
- Fonctionnement :
  - Construction d'un **graphe à haute dimension de voisinage**
  - Projection dans un espace de plus faible dimension
  - Optimisation pour **préserver la structure** des données
- Paramètres :
  - `min_dist` : densité des points dans la projection
  - `n_neighbors` : taille du voisinage

Bon compromis visualisation / structure



t-SNE

- *t-distributed stochastic neighbor embedding*
- Méthode non-linéaire
- Probabiliste
- Sensible à la perplexité
- **Mesure de similarité entre les paires d'instances dans un espace de dimension supérieure et inférieure**



t-SNE

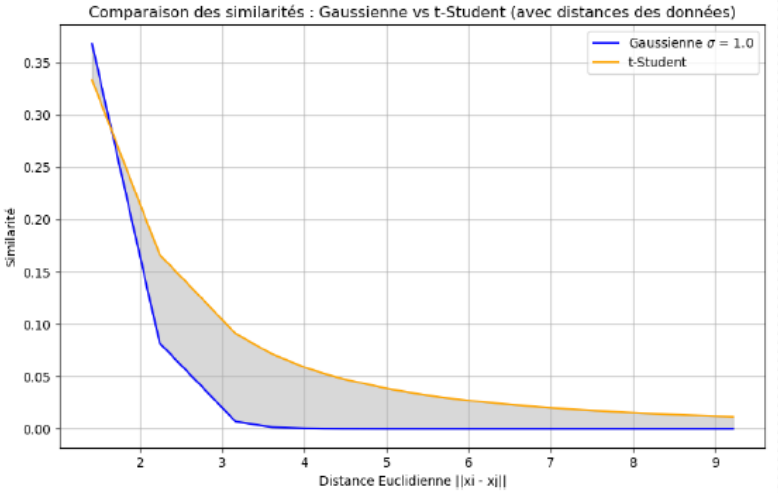
- Étapes :
  - Distances entre points transformées en probabilités à l'aide d'une distribution gaussienne

    Espace de départ de dimension supérieur
  - Distances entre points transformées en probabilités à l'aide d'une distribution de la loi de Student

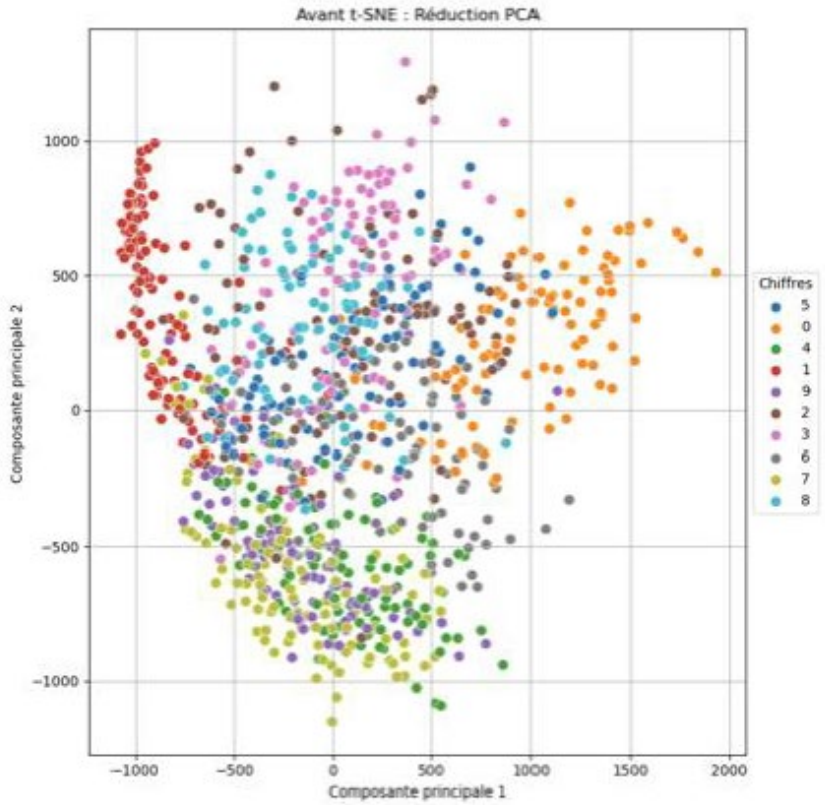
    Espace d'arrivée de dimension inférieur
  - Minimiser la divergence entre les deux distributions

Avec algo. du gradient

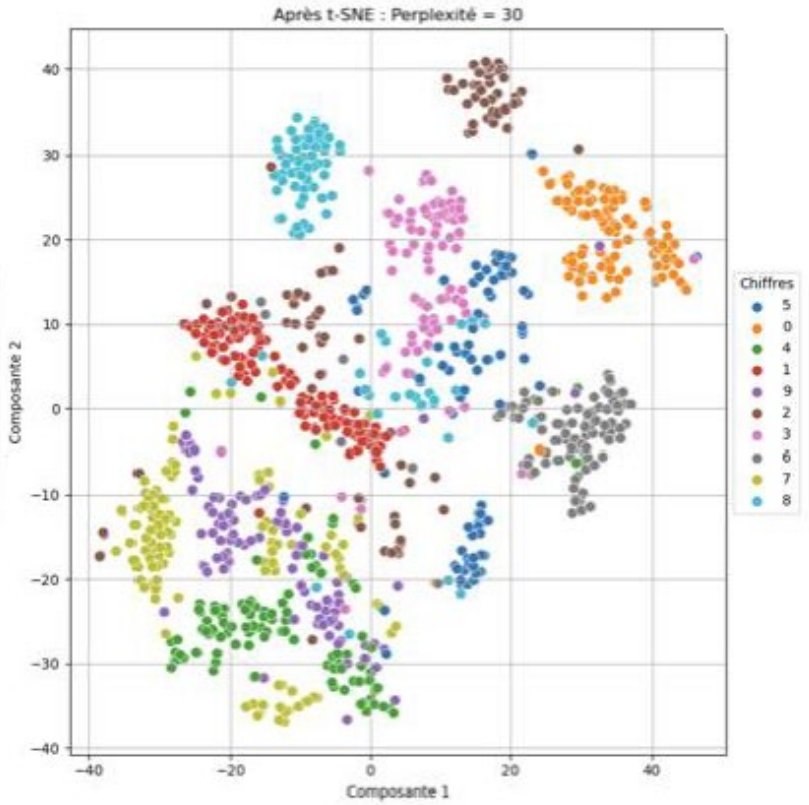
De Kullback-Leibler (KL)



Comparaison



Réduction avec PCA



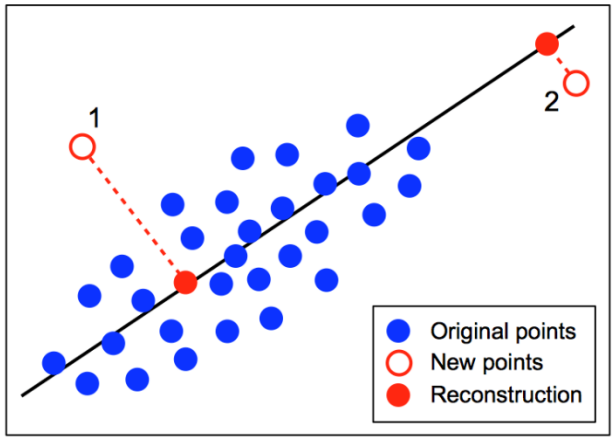
Réduction avec t-SNE

Clusters mieux définis avec t-SNE

Validation

- K plus proches voisins (KNN) pour comparer voisins proches dans l'espace d'origine / dans la projection réduite
- Utilisation d'approches quantitatives :
  - *Trustworthiness* (fiabilité) : évaluation si les voisins proches en dimension sup. restent proches dans la dimension inf.
  - *Reconstruction Error* (erreur de reconstruction) : évaluation de la quantité d'information perdue pendant la réduction de dim.

Vérification des projections



Boîte à outils

- Scikit-learn : https://scikit-learn.org
- Keras : https://www.tensorflow.org/guide/keras
- PyTorch : https://pytorch.org
- NLTK : https://www.nltk.org
- Gensim : https://radimrehurek.com/gensim
- MATLAB : https://www.mathworks.com/products/matlab.html

Crédits

Auteur

Mickaël Martin Nevot
mmartin.nevot@gmail.com

Carte de visite électronique

Relecteurs

Cours en ligne sur : www.mickael-martin-nevot.com