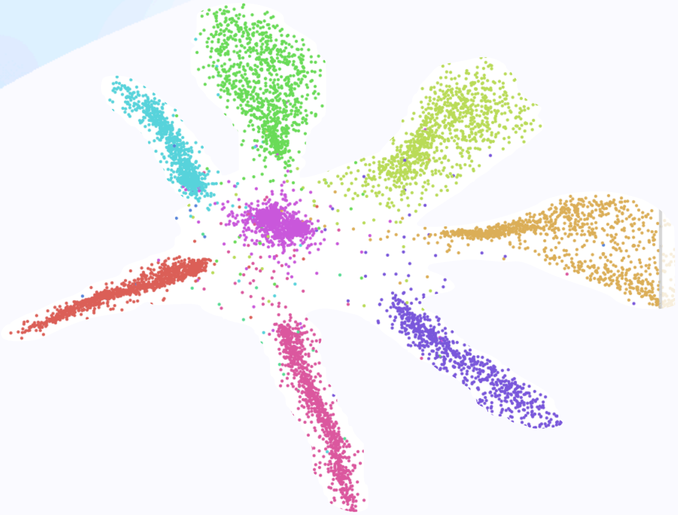
Université Paris Cité

Data analysis and exploration using python

Part II. Dimensionality reduction

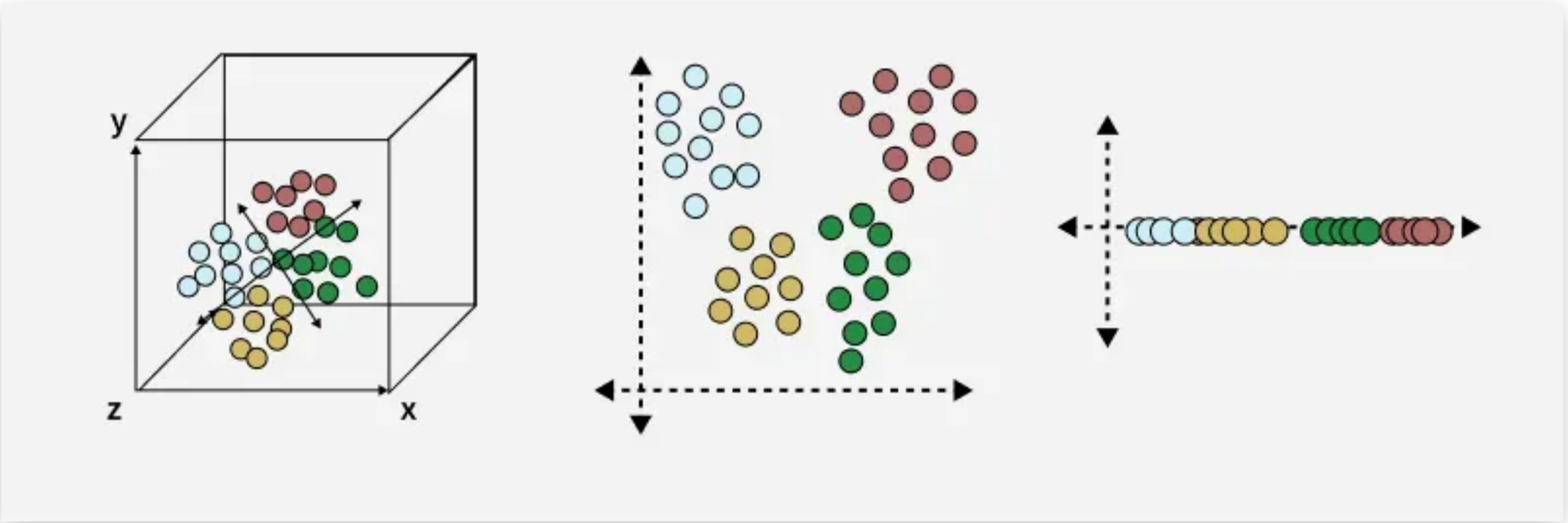


Amine FERDJAOUI

Master II Machine Learning pour la Science des données - Année universitaire 2025/2026

What is dimensionality reduction

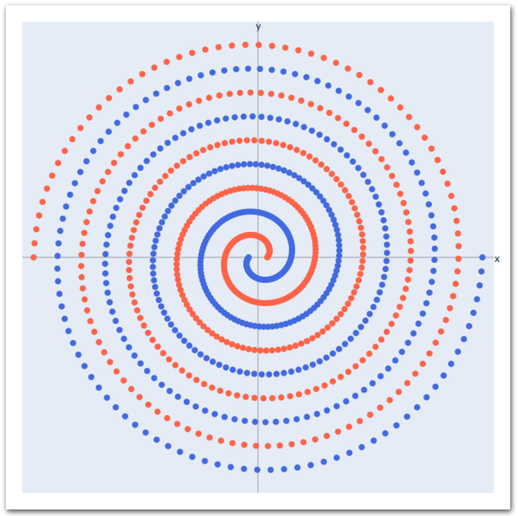
- Dimensionality reduction projects high-dimensional data into a lower-dimensional space while keeping as much useful information as possible.
- It can be used to:
  - Simplify visualization (e.g., 2D/3D plots of complex datasets).
  - Reduce noise and improve machine learning performance.



PCA: Principal Component Analysis

Principal Component Analysis (PCA) is a **linear** and **fast** method based on linear algebra.

- Finds axes (principal components) that maximize the variance of the data.
- Each component is a linear combination of the original variables.
- Useful for reducing dimensionality while keeping most of the variance.
- Limitation: cannot capture non-linear structures in the data (relation between data is non linear)



t-SNE: t-distributed stochastic neighbor embedding

t-SNE is an unsupervised dimensionality reduction, mainly for visualizing complex data in 2D or 3D. Proposed originally in 2002: SNE, and improvement by Laurens van der Maaten and Geoffrey Hinton in 2018 (t-SNE)

How it works

- Points that are close in high-dimensional space should remain close in the reduced space.
- Compute neighborhood probabilities from distances in the original space.
- Place points in low dimension and optimize their positions by minimizing a loss (Kullback–Leibler divergence) between high- and low-dimensional distributions.

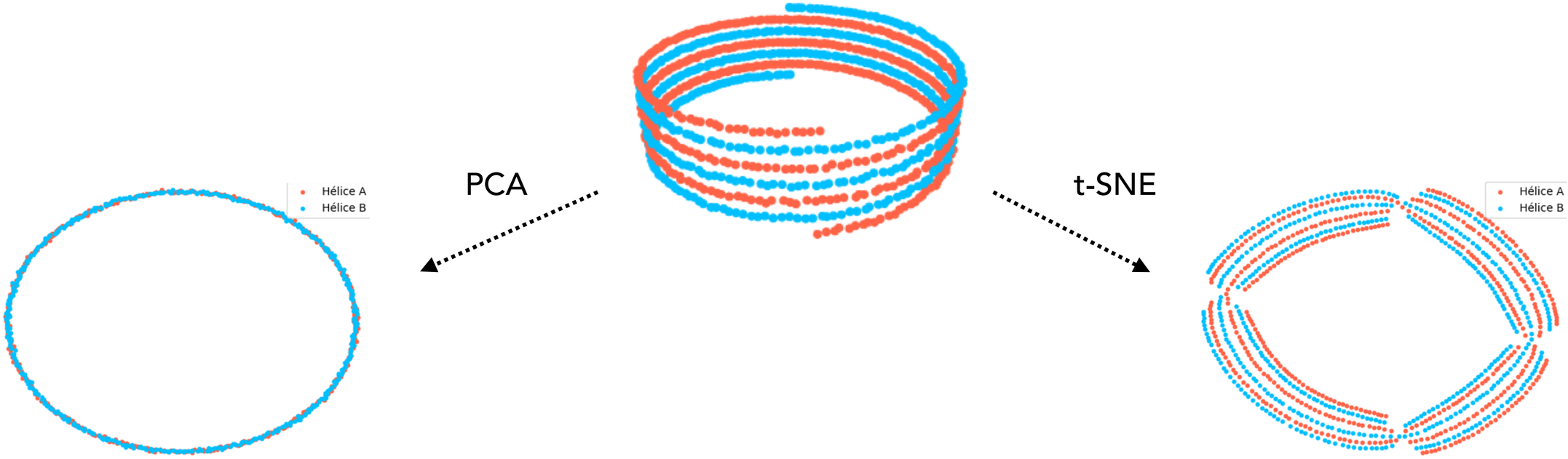
Key parameter

- Perplexity: controls the effective number of neighbors.
- Very sensitive parameter, small changes can lead to big differences.
- Larger perplexity = slower computation.

t-SNE: t-distributed stochastic neighbor embedding

Comparison with PCA:

- PCA: linear, fast, but limited to linear structures.
- t-SNE: nonlinear, better at revealing clusters, but computationally expensive.
- PCA and t-SNE were widely used until ~2018 for exploratory visualization of data.



UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Introduced in 2018 by Leland McInnes, John Healy, and James Melville.

Principle

- Similar goal as t-SNE: dimensionality reduction for visualization in 2D/3D.
- Instead of probabilities, UMAP builds a graph of the data.
- Parameter k (number of nearest neighbors) is crucial.
- For each point, find its k-nearest neighbors and assign edge weights based on distance.
- All local graphs are combined into a single global graph.

Optimization

- In the low-dimensional space, another graph is built.
- UMAP minimizes the difference (loss) between the original high-dimensional graph and the reduced one.
- This is done with stochastic gradient descent.

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Advantages

- Much faster and more scalable than t-SNE.
- Preserves both local and some global structure.
- Produces good and stable results.