





# PROJET

EXPLORATION DE SÉQUENCES DE  
PARCOURS DE SOINS

2024 / 2025



Présenté par :  
Florian, Tom,  
Raphaël, Bastien  
et Valentin



# Contexte du projet

- Analyse de séquence temporelles (Score ALS FRS-R,...) pour le suivi de patients atteints de la maladie de SLA

# Objectifs du projet

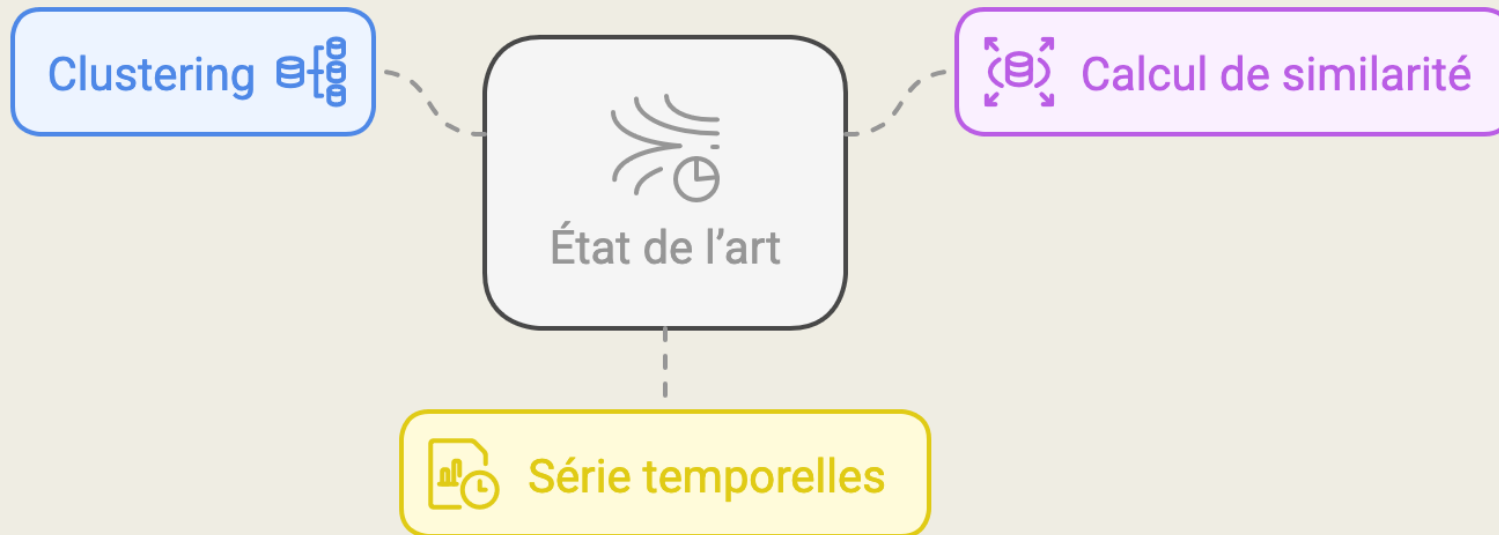
- Développement d'un outil d'analyse pour regrouper les patients aux trajectoires similaires.
- Proposer différentes mesures de similarités et algorithmes de clustering
- Créer une interface interactive et intuitive pour une utilisation par des professionnels du domaine de la santé

- *Données irrégulières et complexes.*
- *Visualisation claire pour les professionnels de la santé. (regroupement des patients et analyse des distances)*
- *Intégration de plusieurs algorithmes de clustering et des mesures de similarités*

# Problématiques



# ÉTAT DE L'ART



# État de l'art

# État de l'art – Série temporelles

- Définition : une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps.
- Caractéristiques :
  - *Souvent multivariées (plusieurs indicateurs en un temps donné)*
  - *Irrégularité des intervalles de temps*
  - *Bruit (prendre en compte les erreurs de mesure qui peut compliquer l'analyse des séries)*
- Domaines d'application :
  - *Finance (prévisions des cours de la bourse), Santé (suivi de différents indicateur) et en Météorologie (prévisions météo)*
- Pratiques actuelles :
  - *Méthodes statistiques comme ARIMA qui est une référence (prédiction de données évoluant dans le temps)*
  - *Sktime (bibliothèque python pour l'analyse de données temporelles)*
  - *TimeScaleDB (extension de PostgreSQL pour les séries temporelles)*

# État de l'art – Calcul de similarité

- Définition : Le calcul de similarité est une méthode qui consiste à mesurer le degré de ressemblance entre deux séries temporelles
- Domaine d'application :
  - *Finance (détection de fraude), Santé (suivi de patients)*
- Pratiques actuelles :
  - *coefficient de corrélation (Pearson)*
  - *distance (Fréchet, DTW, LCSS, Soft-DTW)*
  - *structure (SAX)*



# État de l'art - Clustering

- Définition : Le clustering est une méthode ayant pour but de diviser un ensemble de données en différents sous-ensembles homogènes, c'est-à-dire partageant des caractéristiques communes. (CNIL)
- Domaine d'application :
  - *Analyse de données médicales, analyse de clientèle dans le commerce*
- Pratiques actuelles :
  - *K-Means, DBSCAN, Mean-shift, Ward, Agglomerative clustering*

# État de l'art – Types de clustering

- **Clustering hiérarchique** : regroupe les données en un arbre en divisant/fusionnant au fur et à mesure les clusters
  - *Avantages* : visualisation claire via un dendrogramme, pas besoin de spécifier le nombre de clusters
  - *Inconvénients* : complexité en  $O(n^3)$  → lent pour de grands jeux de données
- **Clustering par Centroïdes (K-Means...)** : Créer k clusters autour de centres en cherchant à minimiser les distances
  - *Avantages* : Simple, rapide et applicable à de grands jeux de données
  - *Inconvénients* : on doit spécifier le nombre de clusters à l'avance,
- **Clustering par Densité (DBSCAN)** : Regroupements de points en fonction de leur densité de voisinage
  - *Avantages* : pas besoin de spécifier le nombre de clusters, robuste aux valeurs aberrantes
  - *Inconvénients* : peut ne pas bien s'appliquer à tous les domaines

- Comparer des séquences temporelles irrégulières
- Regrouper les patients ayant des trajectoires similaires
- Visualiser ces résultats de manière interactive
- Gérer des données temporelles complexes

## Cahier des charges - Besoins utilisateur

mesure	
ID	integer
COLLECT_DATE	timestamp
Proteine TAU (4TAU-L) (ng/L)	real
Tau (4TAU) (ng/L)	real
...	integer
tCO2 (gz_tCO2) (mmol/L)_UNIT	text
temps_ecoule	integer

patient	
1	integer
2	integer
3	text
...	integer
183	real
184	text

examen	
1	integer
2	integer
3	timestamp
...	integer
59	real
60	timestamp

colonnesPatient	
id	integer
nom_colonne	text

colonnesExamen	
id	integer
nom_colonne	text

# STRUCTURE DES TABLES

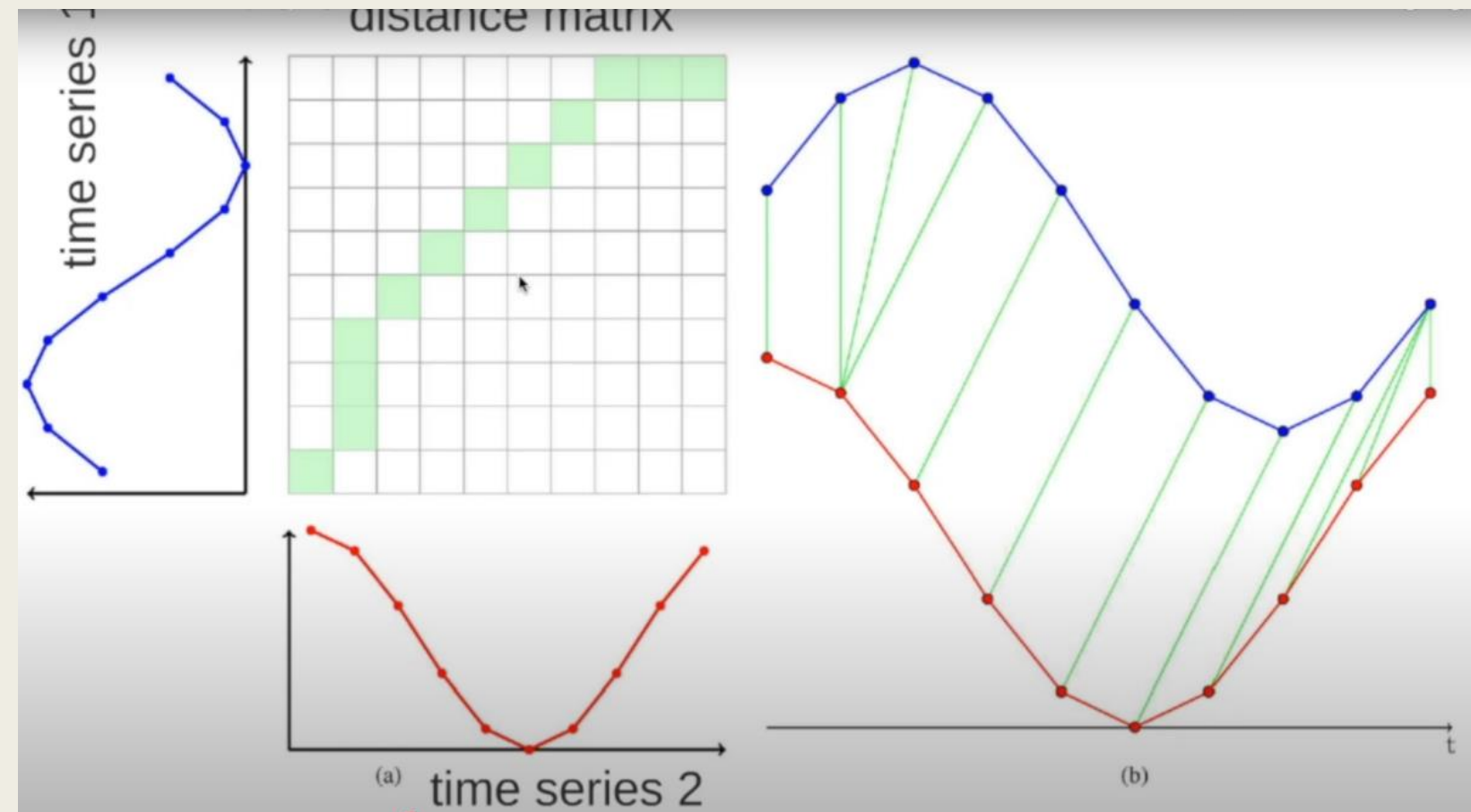
- On utilise les données du projet Optimedias
- Optimisation de la base en utilisant TimeScaleDB (partitionnement avec des hypertables)
- Interpolation des scores ALS FRS-R pour gérer les données manquantes
- Filtrage des patients avec moins de 5 rendez-vous
- Harmonisation des dates (fusion de deux champs)
- Utilisation de Seaborn et panda pour toute les visualisation et manipulation sur les données

# Données

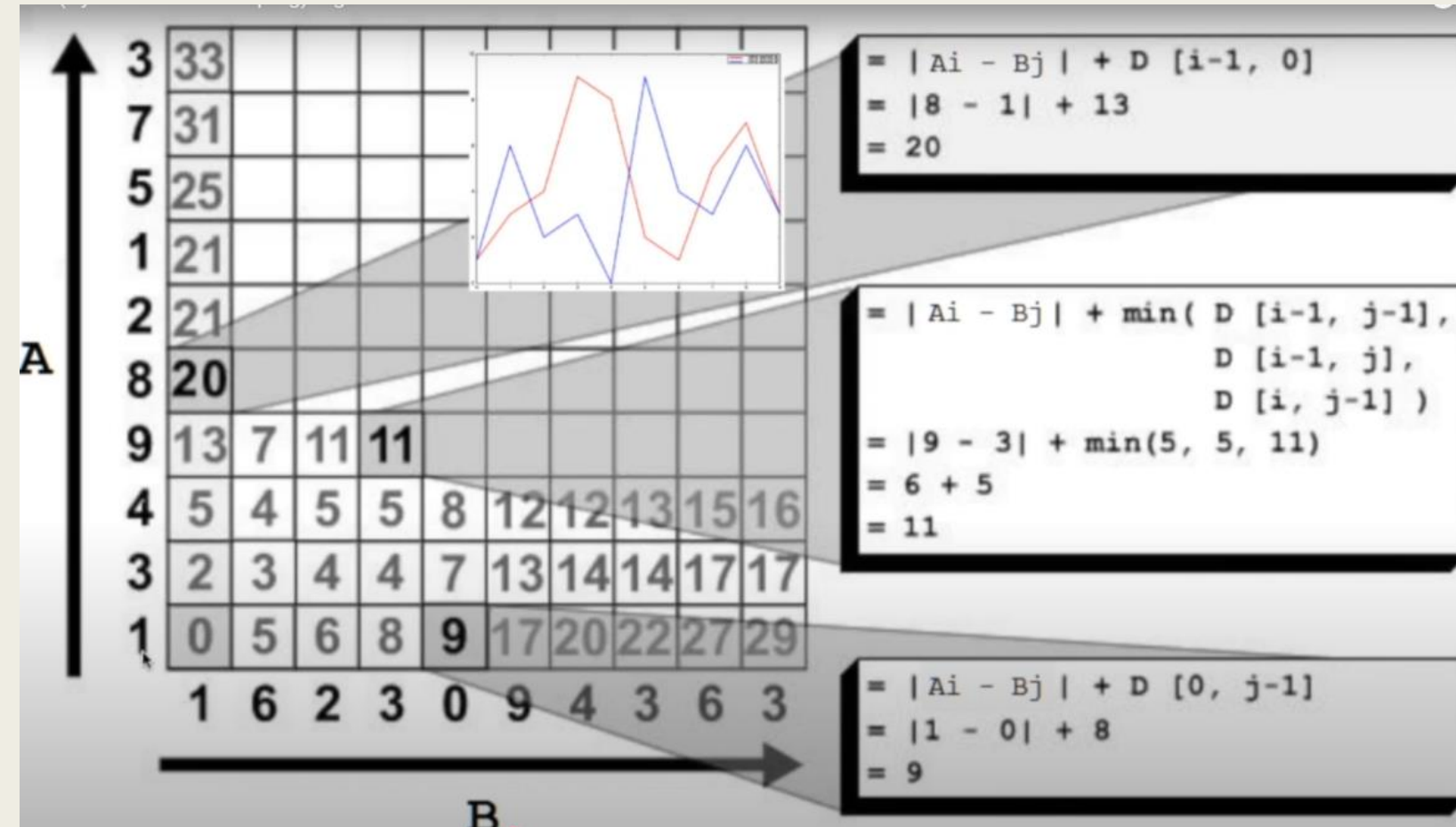
- Utilisation de diverses mesures de similarités :
  - *Dynamic Time Warping (DTW)*
  - *Longest Common Subsequence (LCSS)*
  - *Soft-DTW*
- Utilisation de threads pour diminuer le temps de calcul

## Calculs de similarités

# CALCUL DE SIMILARITÉ - DTW



# CALCUL DE SIMILARITÉ - DTW





		0	1	2	3	4	5	6	7
	X[] →	∅	M	Z	J	A	W	X	U
Y[] ↓	0 ∅	0	0	0	0	0	0	0	0
1	X	0	0	0	0	0	0	1	1
2	M	0	1	1	1	1	1	1	1
3	J	0	1	1	2	2	2	2	2
4	Y	0	1	1	2	2	2	2	2
5	A	0	1	1	2	3	3	3	3
6	U	0	1	1	2	3	3	3	4
7	Z	0	1	2	2	3	3	3	4

CALCUL DE  
SIMILARITÉ -  
LCSS



Algorithmes  
Utilisés



KMeans, K-Medoids, Ward



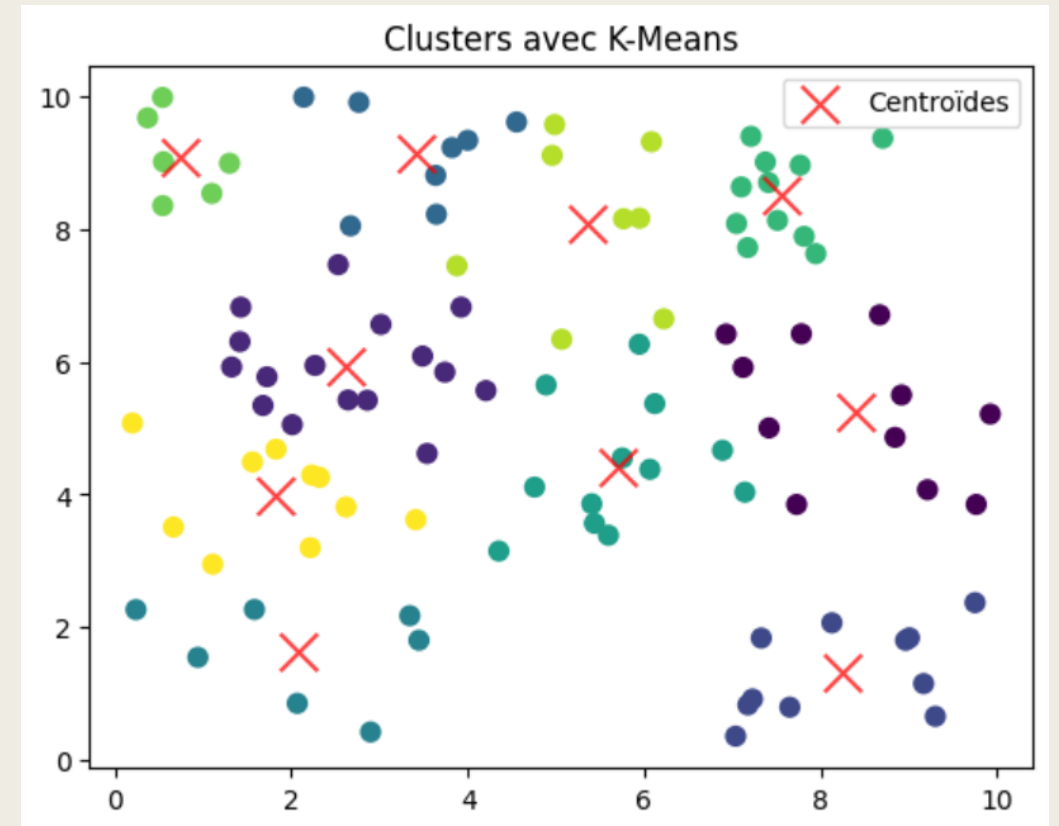
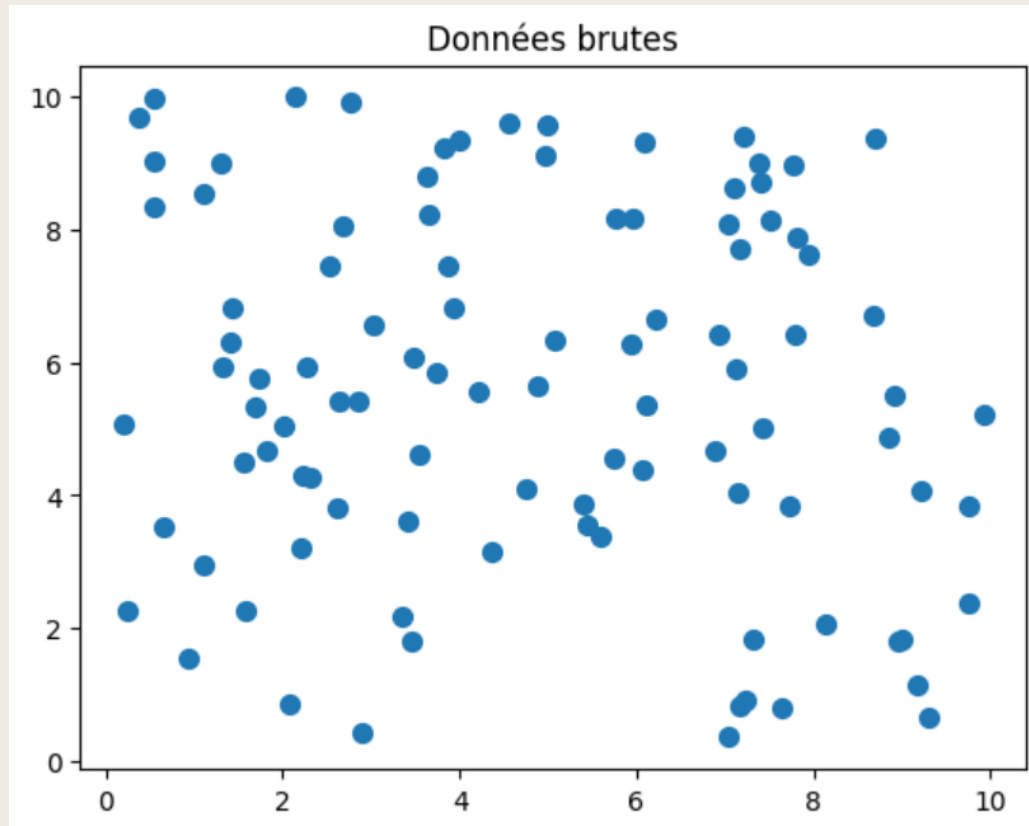
Visualisation  
des Clusters



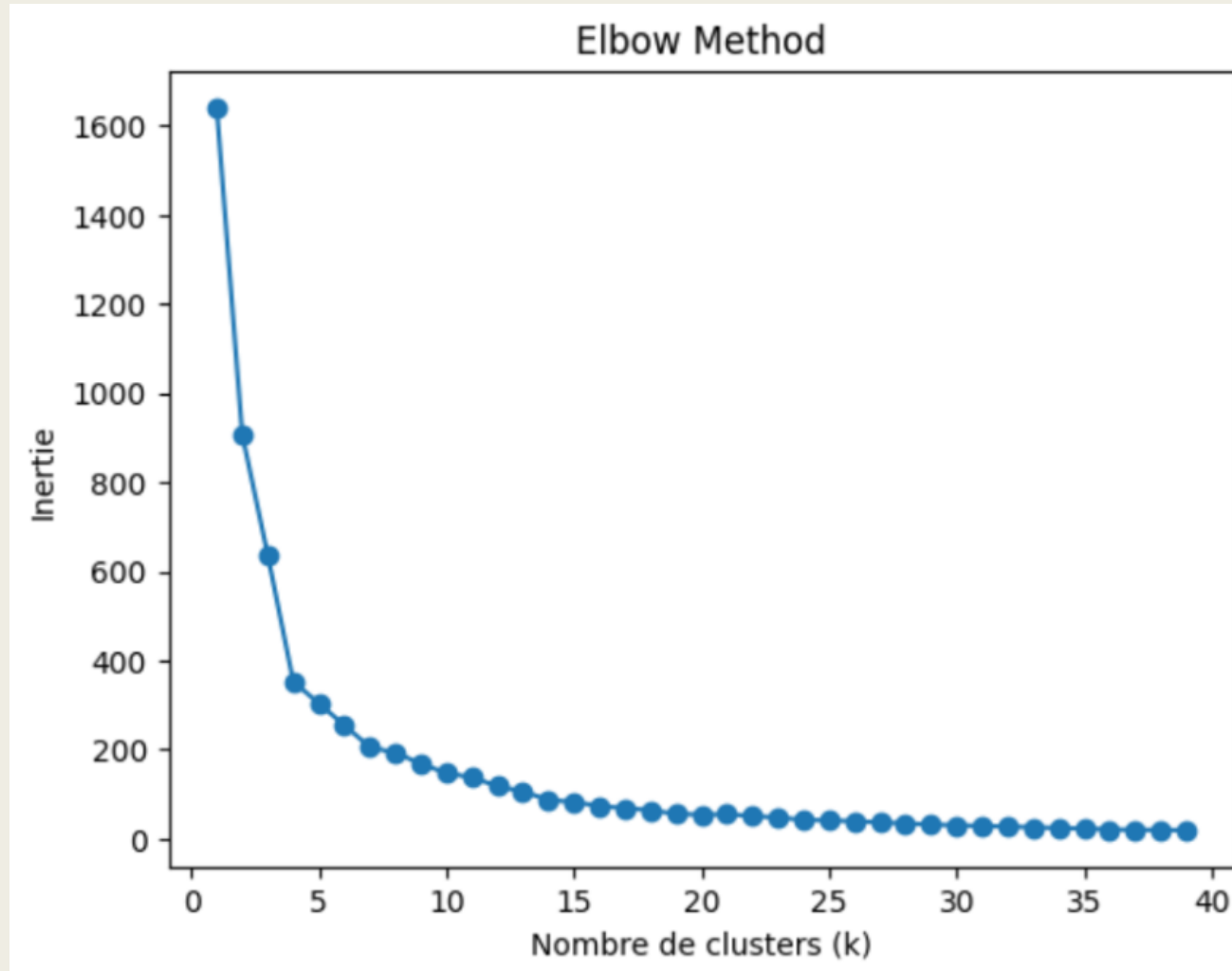
UMAP

# Clustering

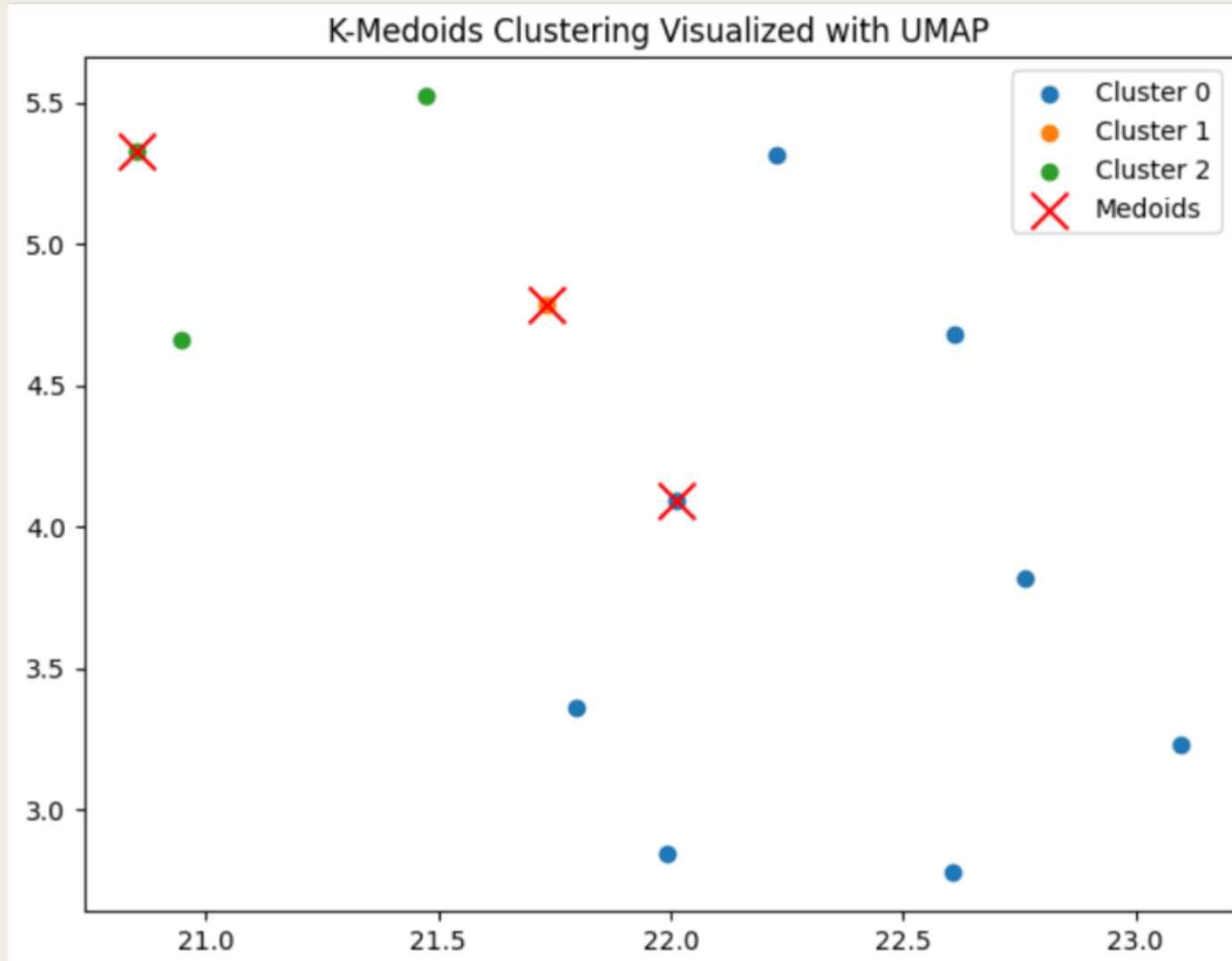
# K-Means



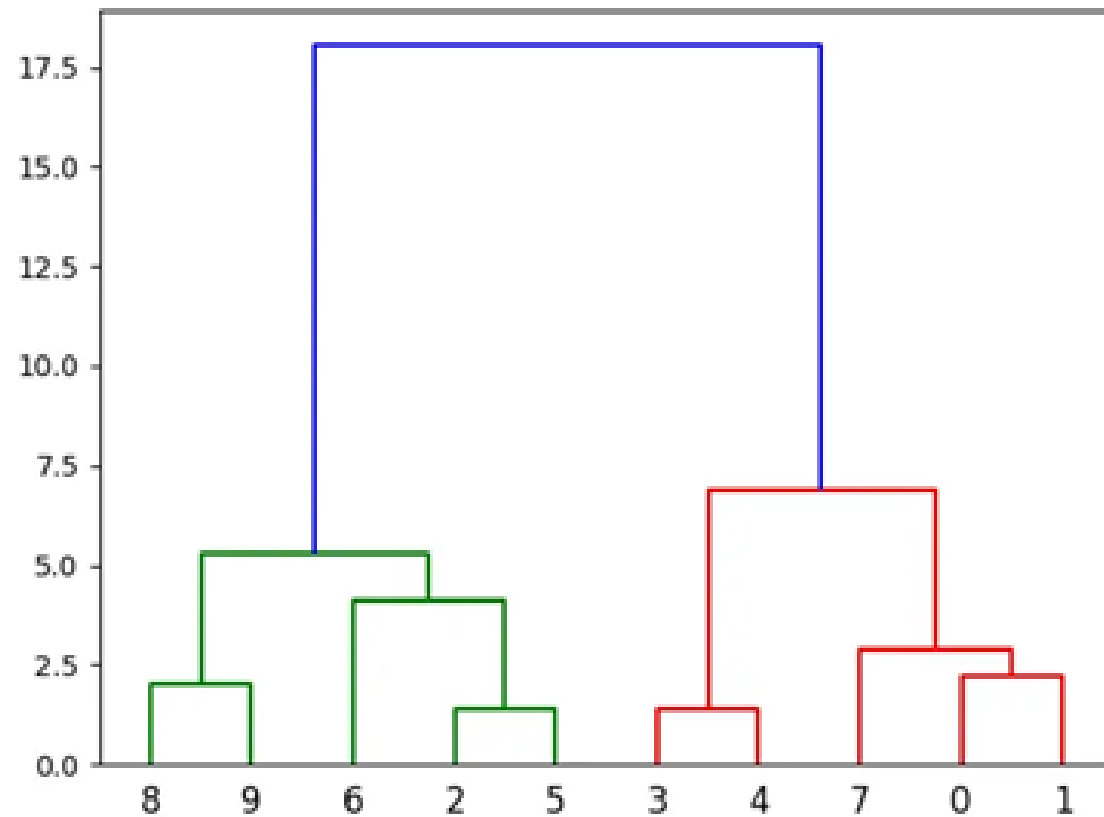
# Elbow Method



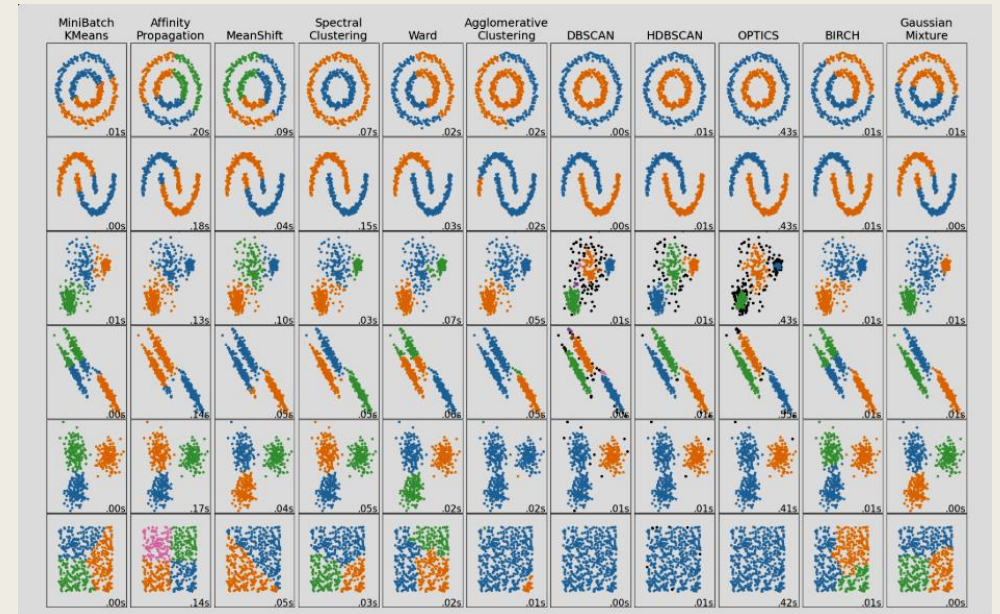
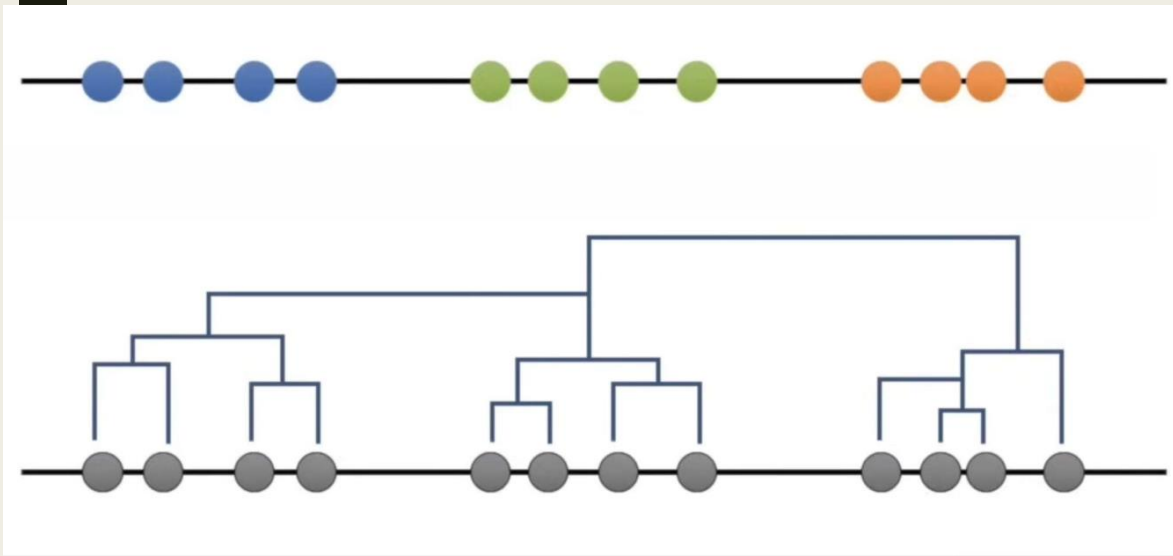
# K-Medoids



# Ward



# VISUALISATION DIFFERENTES METHODES



# Frontend

- Technologies : React, Tailwind CSS, Axios
  - *Sélection de patients*
  - *Affichage des courbes des patients*
  - *Visualisation des clusters*



# Design Patterns

- Utilisation de patterns :
  - *Singleton (connexion à la bdd)*
  - *Strategy (pour le choix des algos de clustering et de calculs de similarités)*

# Cahier des charges

## Filtrer les patients

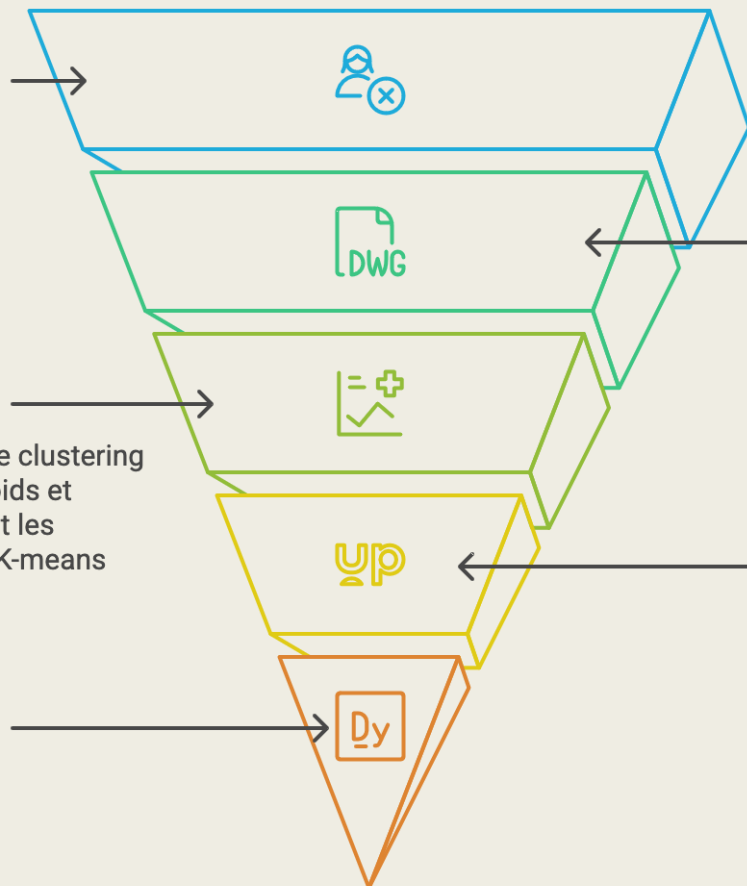
Retirer les patients avec moins de 5 rendez-vous

## Regrouper les patients

Utiliser des méthodes de clustering comme Ward et K-Medoids et sélectionner uniquement les derniers examens pour K-means

## Créer une interface

Développer une interface interactive avec React et Plotly



## Mesurer la similarité

Appliquer des méthodes de mesure de similarité comme DTW ou LCSS

## Visualiser les groupes

Implémenter UMAP pour réduire la dimensionnalité et visualiser les données des patients en 2D



DEMONSTRATION

COMPARAISON ↗

CLUSTERING ↗

ACCUEIL

À PROPOS

DOCUMENTATION

# SLA

Analyse de série temporelles  
sur données de SLA

# ANALYTICS

7°C  
Brouillard



Search



ENG  
FR



11:12  
12/02/2025

## Patients

Rechercher un patient...

Patient 5

Patient 7

Patient 12

Patient 20

Patient 22

Patient 23

Patient 24

Patient 25

Patient 27

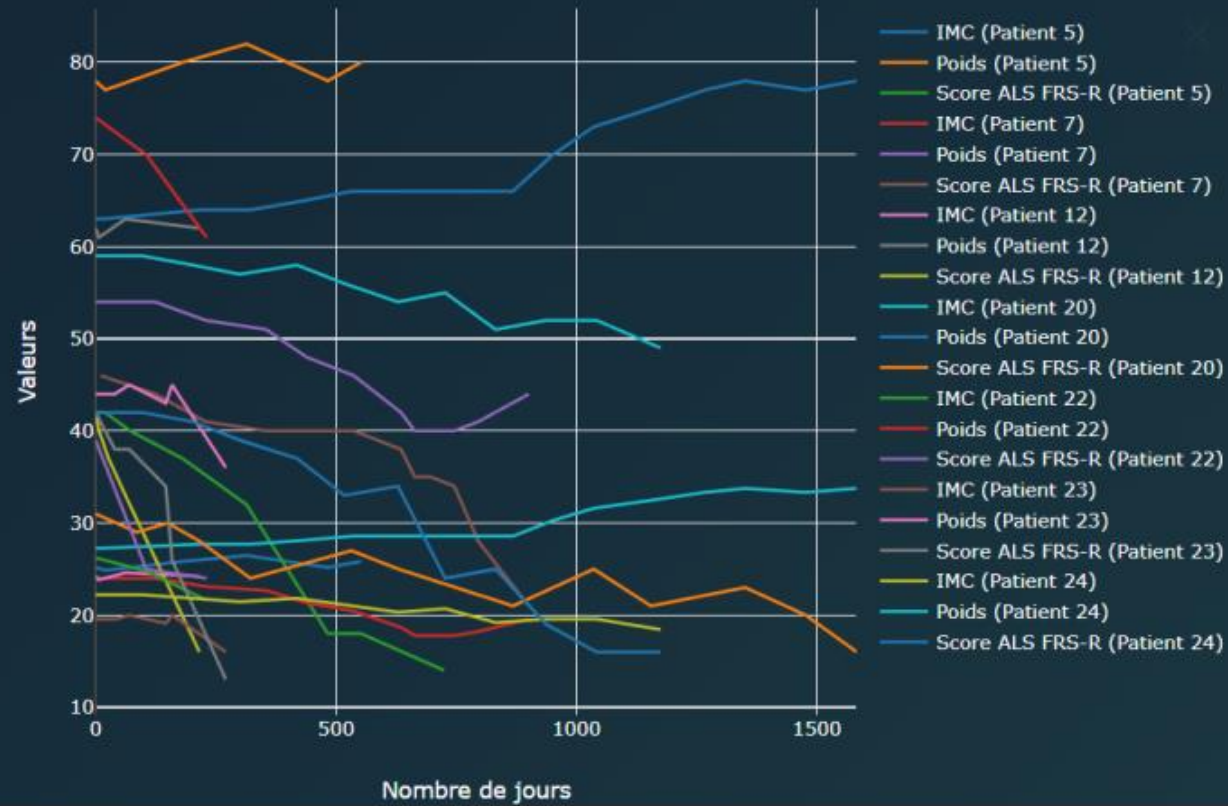
Patient 30

Patient 34

Patient 39

## Graphique

### Données des Patients Sélectionnés



## Comparaison des Patients Sélectionnés

Selectionnez une méthode :

TsLearn Dynamic Time Warping

Comparer

## Paramètres

[Accueil](#)[Clustering](#)Multivariable : ☐[Lock](#)

### Patients

Patient 5

Patient 7

Patient 12

Patient 20

Patient 22

Patient 23

Patient 24

Patient 25

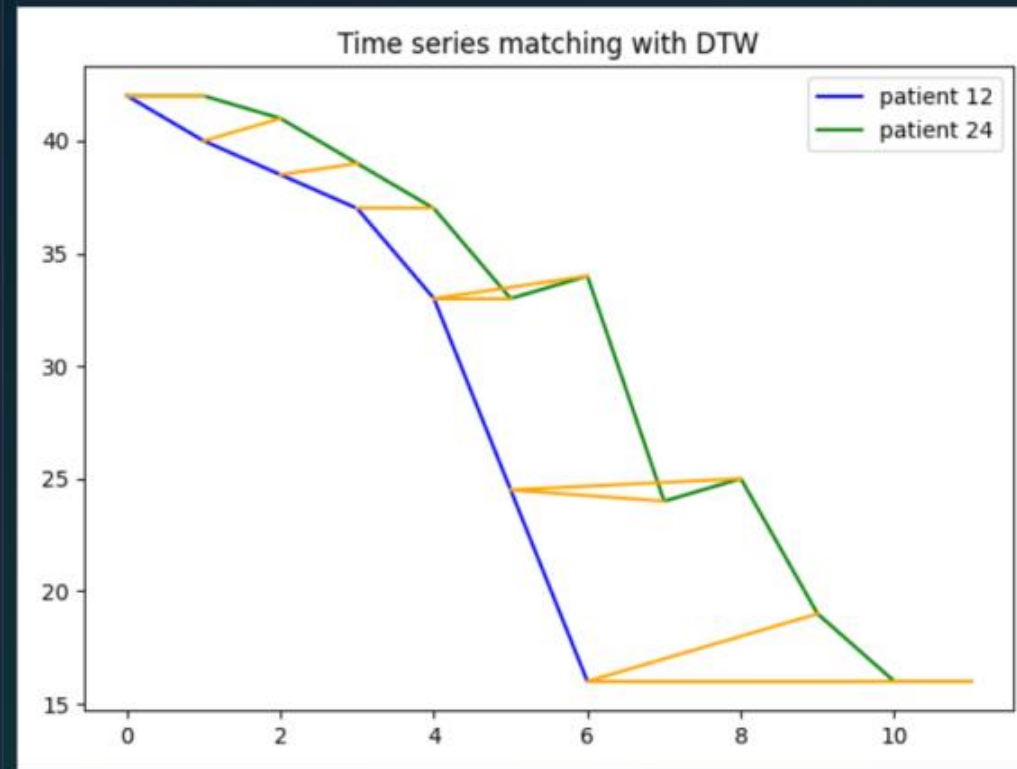
Patient 27

Patient 30

Patient 34

### Comparaison des Patients Sélectionnés

Sélectionnez une méthode : TsLearn Dynamic Time Warping

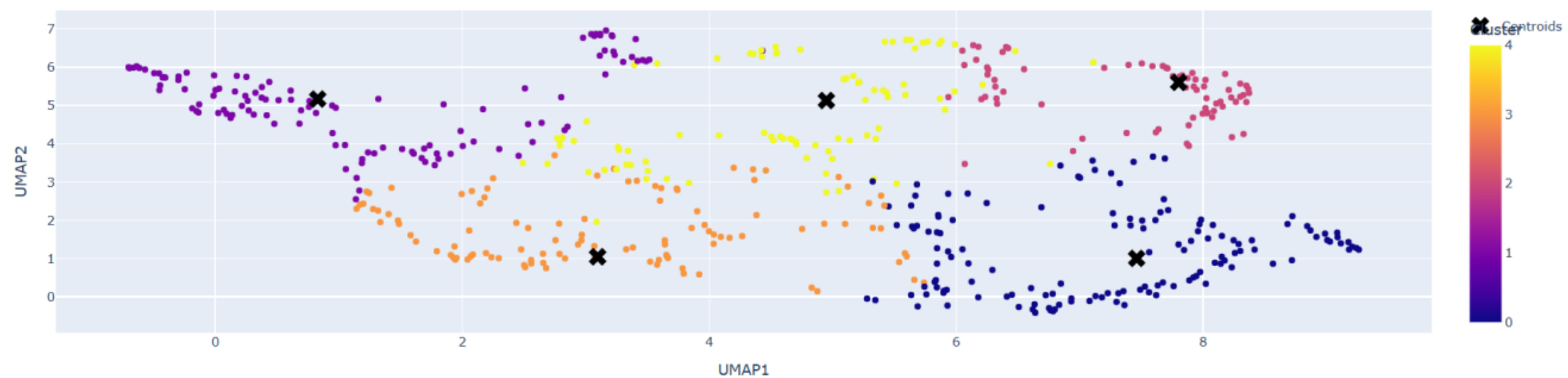
[Comparer](#)

### Graphique

Données des Patients Sélectionnés



# Visualisation des clusters avec K-Means et UMAP



Liste des Patients

Rechercher un patient...

Patient 5

Patient 7

Patient 12

Patient 20

Patient 22

Patient 23

Patient 24

Patient 25

Clusters

Cluster 0137 patients

Cluster 1115 patients

Cluster 272 patients

Cluster 3107 patients

Cluster 487 patients



## Formulaire de Clustering

Type de méthode :

K-medoids



Nombre de clusters :

3

Type de méthode de comparaison :

Soft Dynamic Time Warping



Annuler

Generer

## Cluster 0

Patient 24

Patient 30

Patient 39

Cluster 0 137 patients

Cluster 1 115 patients

Cluster 2 72 patients

Annuler

- Temps de calcul raisonnable, voire assez rapide selon la puissance de la machine
- Qualité des groupes à approfondir mais semble pertinents
- Outil interactif et facile d'utilisation
- Et il faut également prendre en compte l'avis du médecin qui a mesuré le score ALS FRS-R (optimiste ou pessimiste)



# Résultat

- Projet très ambitieux mais très formateur et intéressant
- Perspectives
  - *ajout de DBSCAN*
  - *validation par des professionnel de la santé*
  - *Ajouter plus de variable pour rendre le calcul de similarité plus précis*

# Conclusion