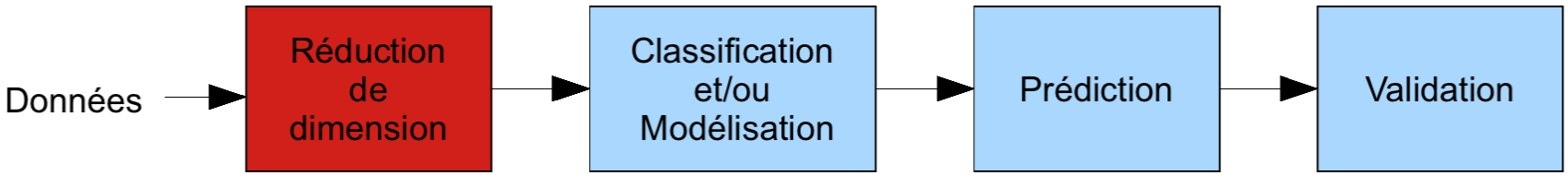
Analyse de données temporelles

Analyse de données et Classification 2
ENSEEIHT - 3ème année Sciences du Numérique

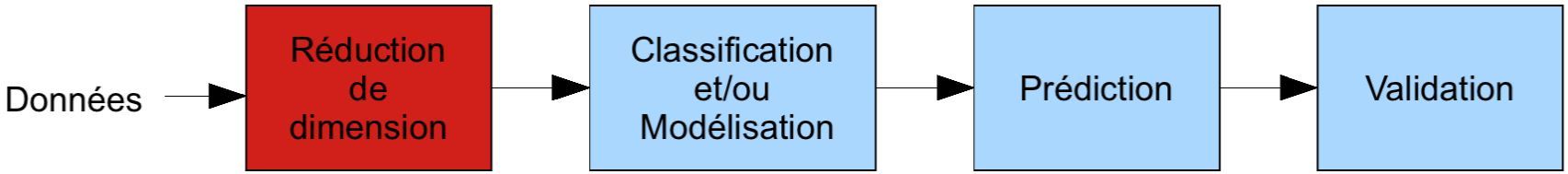
Contact :
Sandrine.Mouysset@irit.fr
sandrine.mouysset@toulouse-inp.fr



Chaîne d'analyse des données

Méthodes descriptives/prétraitement des données :

- Données qualitatives,
- Données quantitatives,
- Données séquentielles (temporelles,..).



Chaîne d'analyse des données

Ces représentations géométriques du tableau de données permettent d'utiliser les notions d'**espaces vectoriels** :

- définir des distances entre individus/variables,
- pondérer l'influence d'un individu/variable,
- identifier des regroupements (agrégations/clusters),
- identifier des relations/liens de dépendances (entre variables/individus).

Chaîne d'analyse des données

Ces représentations géométriques du tableau de données permettent d'utiliser les notions d'**espaces vectoriels** :

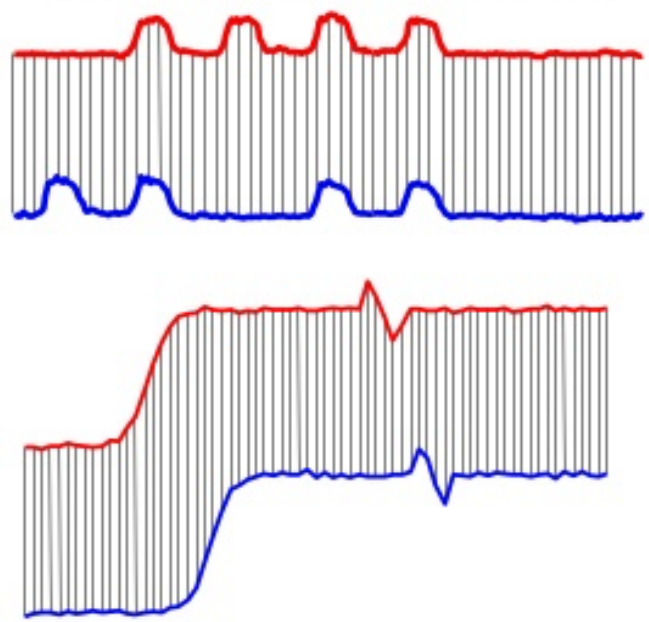
- définir des distances entre individus/variables,
- pondérer l'influence d'un individu/variable,
- identifier des regroupements (agrégations/clusters),
- identifier des relations/liens de dépendances (entre variables/individus).

$\Rightarrow$ Méthodes inférentielles

Notion de distance pour les séries temporelles ?

Quelle distance utiliser lorsqu'on traite des séries temporelles :

- de tailles différentes,
- désynchronisées ?

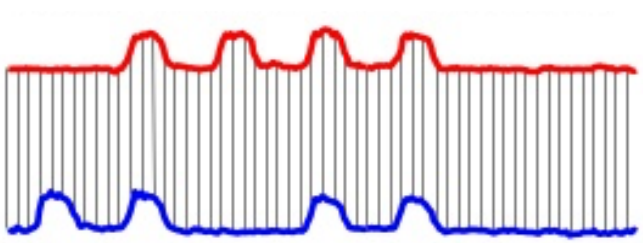


Distance Euclidienne

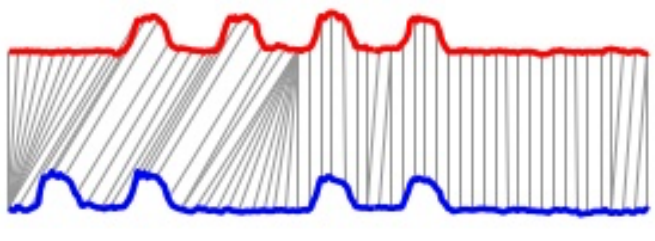
Notion de distance pour les séries temporelles ?

Quelle distance utiliser lorsqu'on traite des séries temporelles :

- de tailles différentes,
- désynchronisées ?



Distance Euclidienne



Dynamic Time Warping (DTW)

⇒ Utilisation d'un algorithme de déformation temporelle dynamique pour synchroniser et aligner des séries temporelles entre elles

Dynamic Time Warping (DTW)

On considère **2 séquences temporelles** $\mathcal{A}$ et $\mathcal{B}$ de supports temporels resp. $I = [|1, N|]$ et $J = [|1, M|]$.
On considère une **distance locale** ,notée d, définie par : $d(c) = d(a_i, b_j)$ avec $c = (i, j)$ de $I \times J$.

- Evaluer la dissemblance entre $\mathcal{A}$ et $\mathcal{B}$ revient à déterminer dans $I \times J$ un chemin $C = c_1, ..., c_k, ...; c_K$ tel que $c_k = (i_k, j_k),\ k \in \{1..K\}$
- Les dissemblances entre $\mathcal{A}$ et $\mathcal{B}$ sont cumulées le long de ce fichier selon: $D(\mathcal{A}, \mathcal{B}, C) = \sum_{k=1,..,K} \omega_k d(c_k)$, ω_k le poids attaché à l'arc (c_{k-1}, c_k).

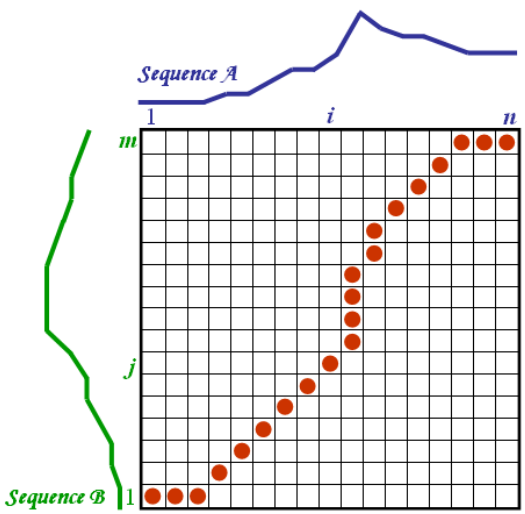
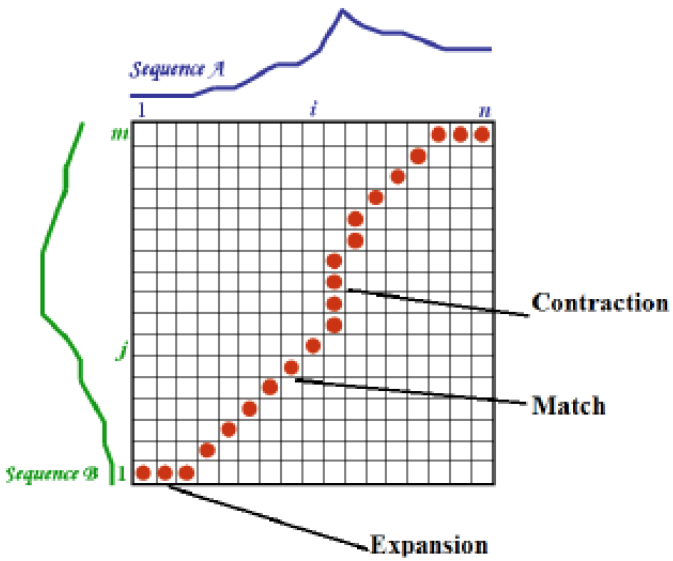


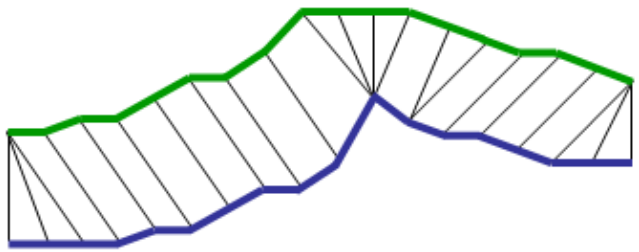
Schéma global de l'alignement de deux séquences $\mathcal{A}$ et $\mathcal{B}$:
le chemin optimal (*warping function*) est représenté par le chemin en point rouge

Propriétés des chemins

Les chemins C doivent vérifier les conditions suivantes :

- **coïncidence des extrémités :** $c_1 = (1, 1)$ et $c_K = (N, M)$
- **monotonie - croissance temporelle :** $c_{k-1} < c_k$
- **continuité :** pas de saut ou de trou sur le chemin





Alignement

Algorithm 4 Algorithme DTW

Input :

- 2 séries temporelles $\mathcal{A}$ et $\mathcal{B}$ de supports temporels resp. $I = [|1, N|]$ et $J = [|1, M|]$
- Distance locale d

1. **Initialisation :** $w_0 \leftarrow 1, \quad w_1 \leftarrow 2, \quad w_2 \leftarrow 1, \quad g(0,0) \leftarrow 0.$
2. Pour $j \in \{1..J\}$
   (a) $g(0, j) \leftarrow +\infty$
   (a) Pour $i \in \{1..N\}$
      i. $g(i, 0) \leftarrow +\infty$
      ii. Pour $j \in \{1..M\}$
         Recherche du chemin minimal :

$$g(i,j) \leftarrow min \begin{cases} g(i-1,j) + w_0 * d(i,j) \\ g(i-1,j-1) + w_1 * d(i,j) \\ g(i,j-1) + w_2 * d(i,j) \end{cases}$$

3. **Calcul du score d'alignement :** $S = g(N, M)/(N + M)$

Output : Score S et chemin optimal C.

Contraintes locales et globales

permettent de limiter le nombre de chemins possibles autour de la diagonale.

1. **Contrainte locale :** pondérations spécifiques sur les chemins $(\omega_0, \omega_1, \omega_2)$

2. **Contrainte globale :** on limite le chemin à une certaine enveloppe autour de la diagonale.
   En effet, dans certaines applications, il est peu probable que le meilleur alignement s'écarte trop de la diagonale, on peut donc se dispenser de calculer toutes les valeurs.

Exercice : Comparaison de séquences ADN

Comparez par DTW les séquences ADN suivantes :

1. ATGGTACGTC
2. AAGTAGGC

avec la *distance locale d* suivante:

$$d(L_i, L_j) = \begin{cases} 0 \text{ si } L_i = L_j \\ 1 \text{ sinon.} \end{cases}$$

comme *contraintes locales :* $(\omega_0, \omega_1, \omega_2) = (1, 1, 1)$
et comme *contraintes globales :* les cases à plus de 4 cases de la diagonales ne seront pas à calculer.

Calculer la matrice de coût, le score et l'alignement.

Projet :

Classification de battements cardiaques

Il existe deux types de classification :

- Si on dispose d'une **base d'apprentissage**, sous ensemble de données "étiquetées" par des experts du type

X	$X^1 \dots X^i \dots X^m$	Classe
1 ⋮ i ⋮ n	Caractéristiques variables	S variables nominales

Classification supervisée :

Grâce à cette base d'apprentissage on peut choisir ou apprendre un modèle décisionnel qui explique les relations entre caractéristiques d'entrée et la classe de sortie.

- Si on ne dispose pas de base d'apprentissage, la **classification non supervisée** est privilégiée.

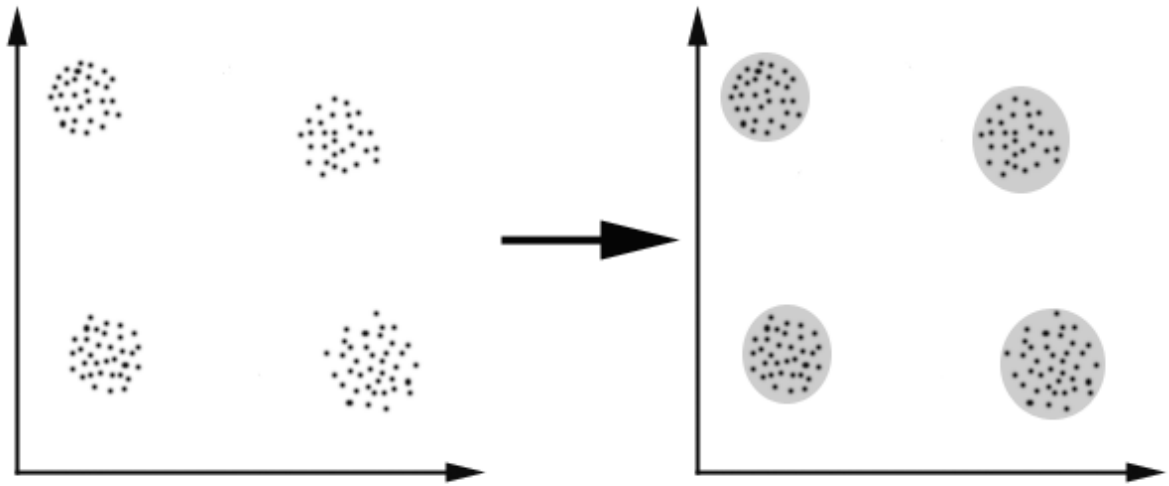
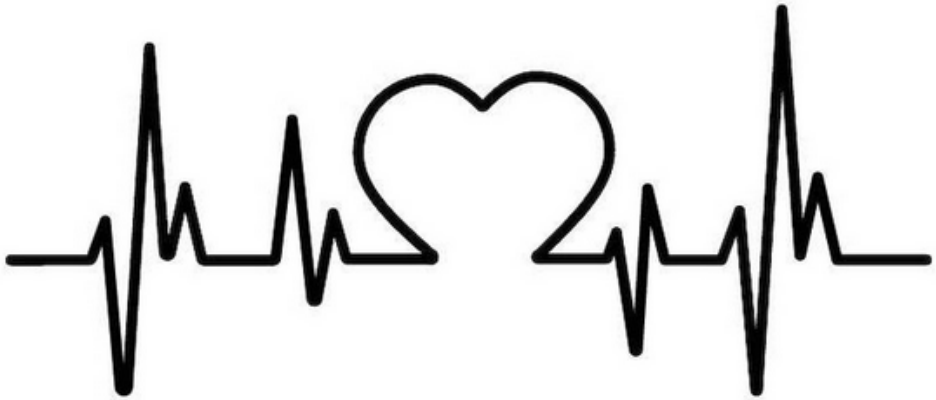


Figure: Exemple de classification non supervisée : diviser cet ensemble de points en 4 classes à partir de la distance entre les points

Classification non supervisée

vise à créer une partition (un ensemble de classes) d'un ensemble de données à partir des mesures de similarité entre ces données afin que des données appartenant à une même classe soit le plus semblable possible et des données appartenant à des classes soient le moins semblables possible.



Classification de battements cardiaques

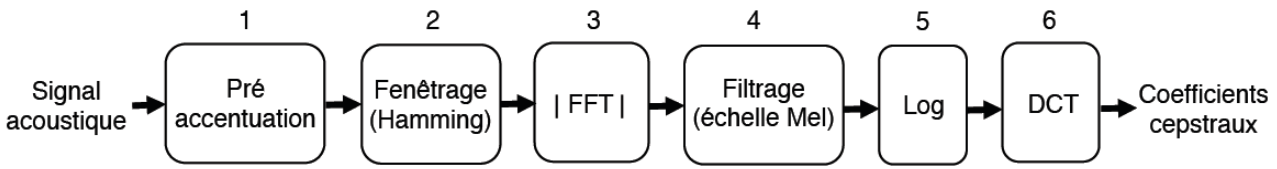
Les données ont été recueillies auprès de deux sources :

- (A) auprès du grand public via une application de smartphone,
- (B) dans le cadre d'un essai clinique dans des hôpitaux utilisant le stéthoscope numérique.

Les enregistrements de ces 2 sources étant de durées différentes, ils ont été ensuite transformés en MFCC pour extraire le contenu fréquentiel de ces données.

Transformation du signal temporel en MFCC (Mel Frequency Cepstral Coefficient)

Le cepstre présente l'avantage de permettre la séparation des contributions respectives de la source et du conduit vocal. Les MFCC s'obtiennent en utilisant, pour le calcul du spectre, une échelle fréquentielle non linéaire tenant compte de la perception auditive de la fréquence.



$\Rightarrow$ Utilisation de la librairie *Librosa*

Distance locale : distance euclidienne dans $\mathbb{R}^p$ (ici $p = 20$)

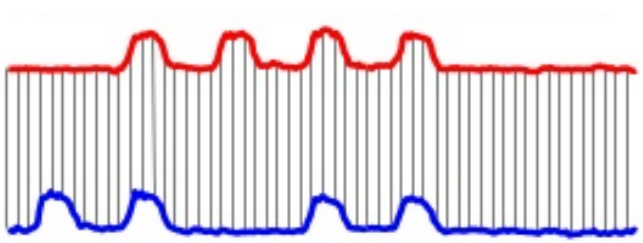
$$d(X, Y) = \left(\sum_{k=1}^{p} (X_k - Y_k)^2 \right)^{\frac{1}{2}}$$

Organisation :
Ce projet en 6 séances se décomposent en 5 parties :

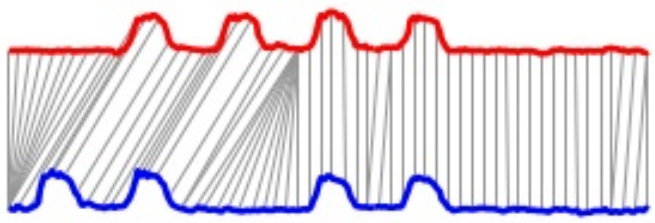
- *Partie I :* Système de reconnaissance d'activité physique avec la DTW
- *Partie II :* Réduction de dimension par ACP et classification par kppv
- *Partie III :* Classification par Forêts aléatoires
- *Partie IV :* Classification par réseaux de neurones
- *Partie V :* Votre étude

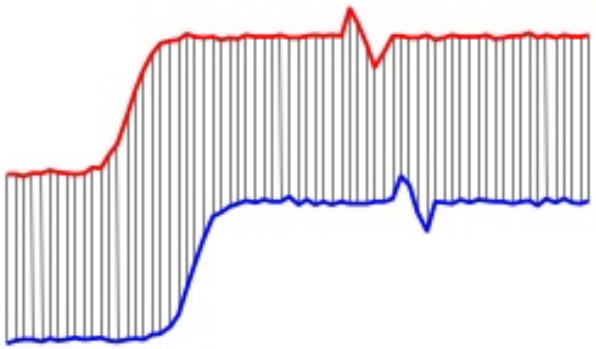
Projet :

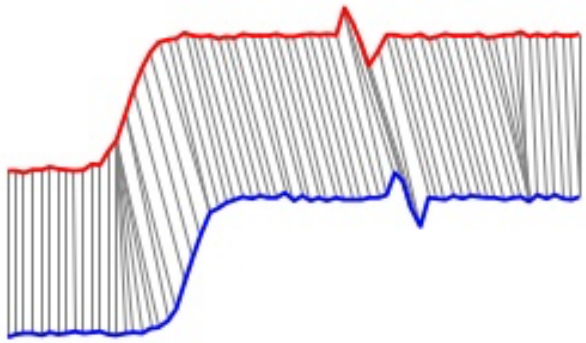
- **Langage :** Notebook Python
  - Mise à disposition d'un tutoriel python sous moodle
  - Possibilité de partager le notebook via *Google Colab*
- Travail à réaliser **en binôme**
- **Livrables du projet :** un notebook par binôme et un rapport au format pdf de 10 pages max.
- **Deadline : le 5 févier !**



Distance Euclidienne







Dynamic Time Warping (DTW)

⇒ Utilisation de la DTW, algorithme de déformation temporelle dynamique pour synchroniser et aligner des séries temporelles entre elles

⇒ Implémenter l'algorithme de DTW.

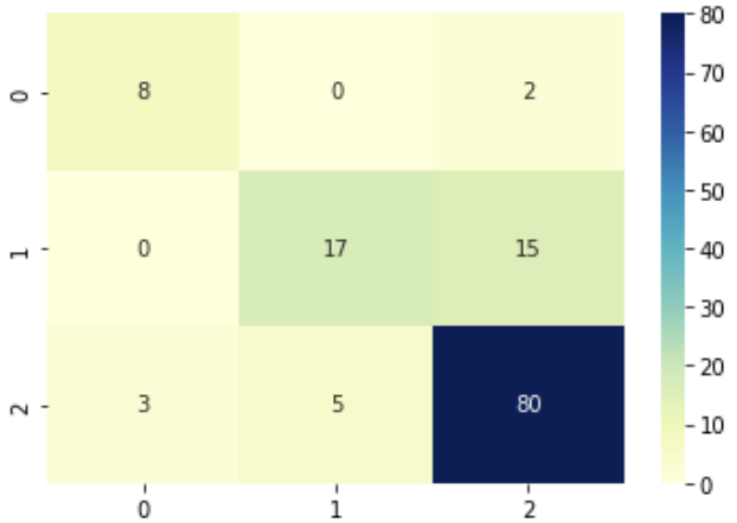
1. **Utilisation de la DTW pour la classification supervisée**
   Soit un dictionnaire $\{R_1, R_2, ..., R_N\}$ constitué des exemples de séquences qui sont connues à l'avance (*base d'apprentissage*).
   L'algorithme va consister à rechercher la référence R_m la plus proche d'une séquence dans la *base de test* M à identifier à l'aide d'une distance D :

$$m = \arg_{1 \leq i \leq N} \min \left(Score \left(M, R_i \right) \right)$$

2. **Evaluation de la DTW par matrice de confusion et pourcentage d'éléments bien classés:**
   fonctions respectives *confusion_matrix* et *accuracy_score* de la librairie python *scikit-learn*

Matrice de confusion : consiste à compter le nombre de fois où des observations de la classe A ont été rangées dans la classe B.

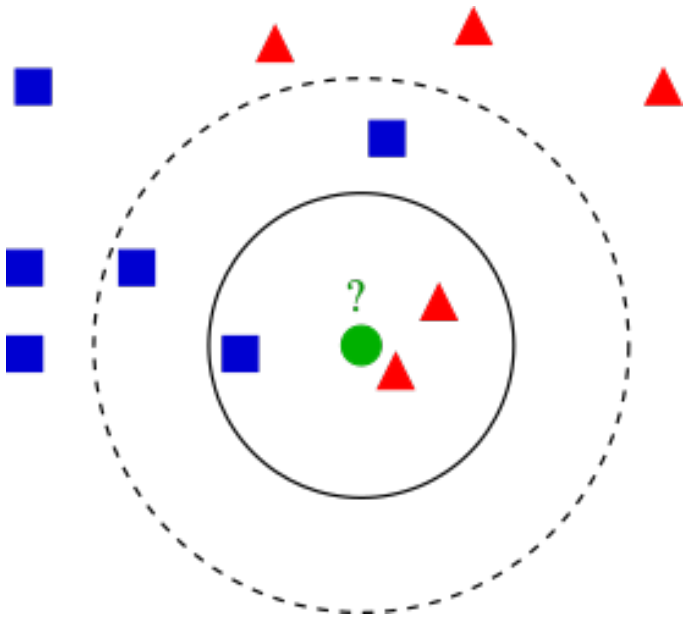
- Chaque ligne de la matrice de confusion représente la *classe réelle* tandis que chaque colonne représente une *classe prédite*.
- Les éléments diagonaux représenteront le nombre d'éléments bien classés. *Classification parfaite = matrice de confusion diagonale*
- Les éléments hors diagonaux représenteront les erreurs de classification.



Exemple : si on souhaite connaître le nombre de fois où le classifieur a pris des battements murmures poru des battements normaux on regardera l'élément hors diagonal (2,3) dans la matrice de confusion.

Utilisation de la librairie *scikit learn* de Python (mais vous pouvez aussi les coder vous même !)

- réduction de dimension par ACP : fonction *PCA*
- méthode de classification classique k plus proches voisins (k-ppv) : fonction *KNeighborsClassifier*



Exemple de classification par k-ppv. L'échantillon de test (cercle vert) doit être classé soit dans la première classe des carrés bleus, soit dans la deuxième classe des triangles rouges. Si k = 3 (cercle plein), il est assigné à la deuxième classe parce qu'il y a 2 triangles et seulement 1 carré à l'intérieur du cercle intérieur.

- **Parties III & IV :** Classification par **forêts aléatoires** et **réseaux de neurones;**
- **Partie V :** Réalisez **votre propre étude**, par exemple, en :
  - en comparant les mesures des différents systèmes (set A et B)
  - en équilibrant les classes
  - en proposant des variantes des méthodes proposées et/ou en testant d'autres méthodes de classification

Le tout en testant les approches (parties I à IV) et en interprétant les résultats via les mesures d'évaluation (matrice de confusion et pourcentage de données bien classées) et synthétiser votre étude dans un rapport (10 pages max en pdf).