

Semaine 4 : Apprentissage non-supervisé et réduction de dimension

4 : Introduction à la classification non-supervisée

Laurent Risser
Ingénieur de Recherche CNRS
Institut de Mathématiques de Toulouse / 3IA ANITI

lrisser@math.univ-toulouse.fr

Rappel sur la classification supervisée

1) Introduction - classification supervisée



Aide au diagnostic

Base d'apprentissage

Patient 1 :

- Age = 40
- Globule Blancs/L = 6

Sain

Patient 2 :

- Age = 28
- Globule Blancs/L = 12

Rhume

Patient N :

- Age = 57
- Globule Blancs/L = 8

Sain

Nouveau Patient (hors base d'apprentissage) :

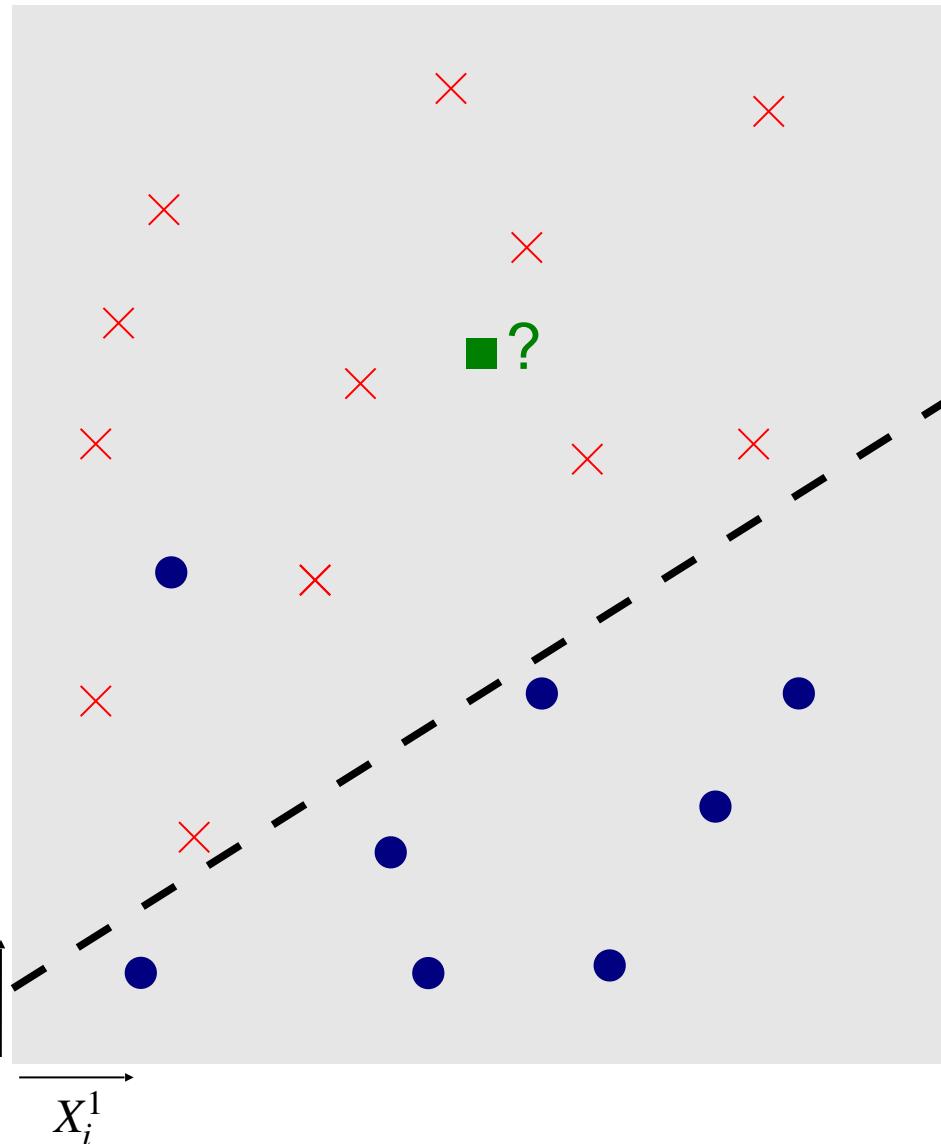
- Age = 34
- Globule Blancs/L = 5



Sain ou rhume ???

1) Introduction - classification supervisée

Apprentissage supervisé — classification



Observations d'entrée (X) :

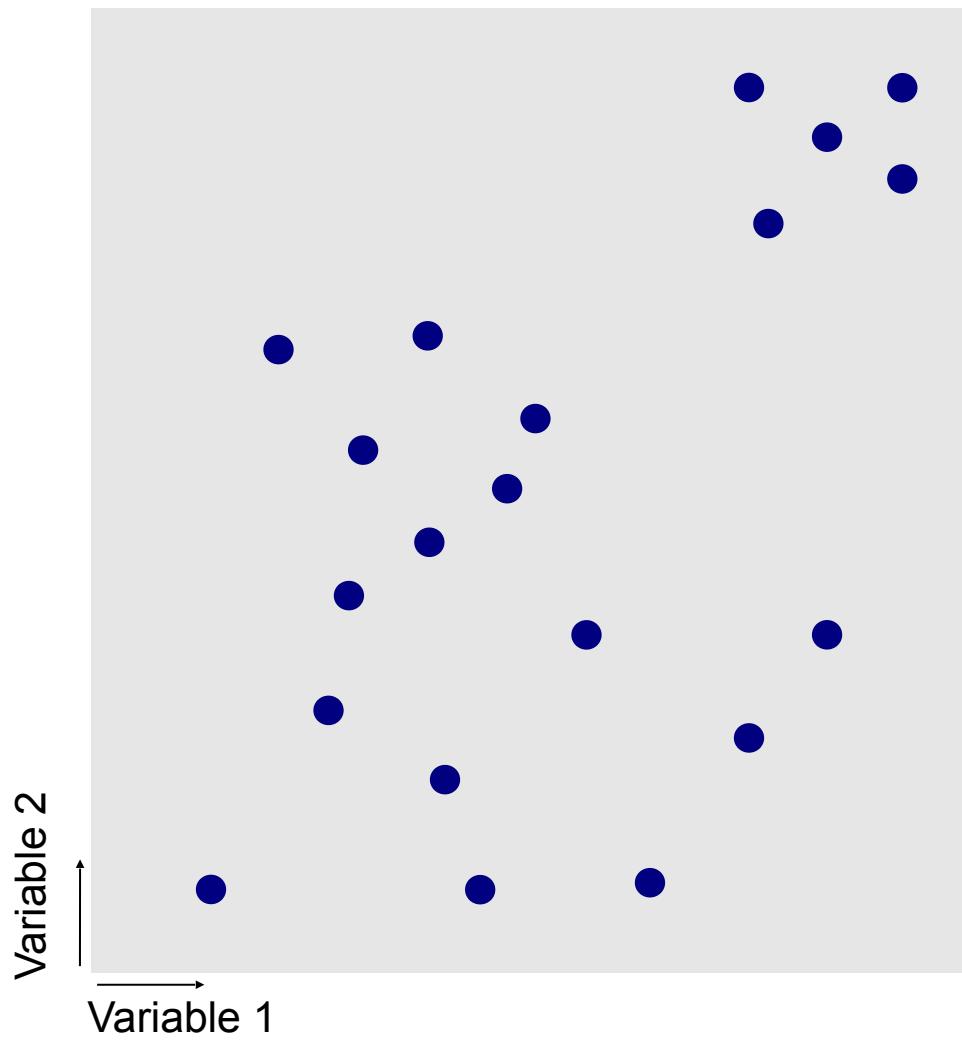
- n observations $X_i \in \mathbb{R}^p$

Observations de sortie (Y) :

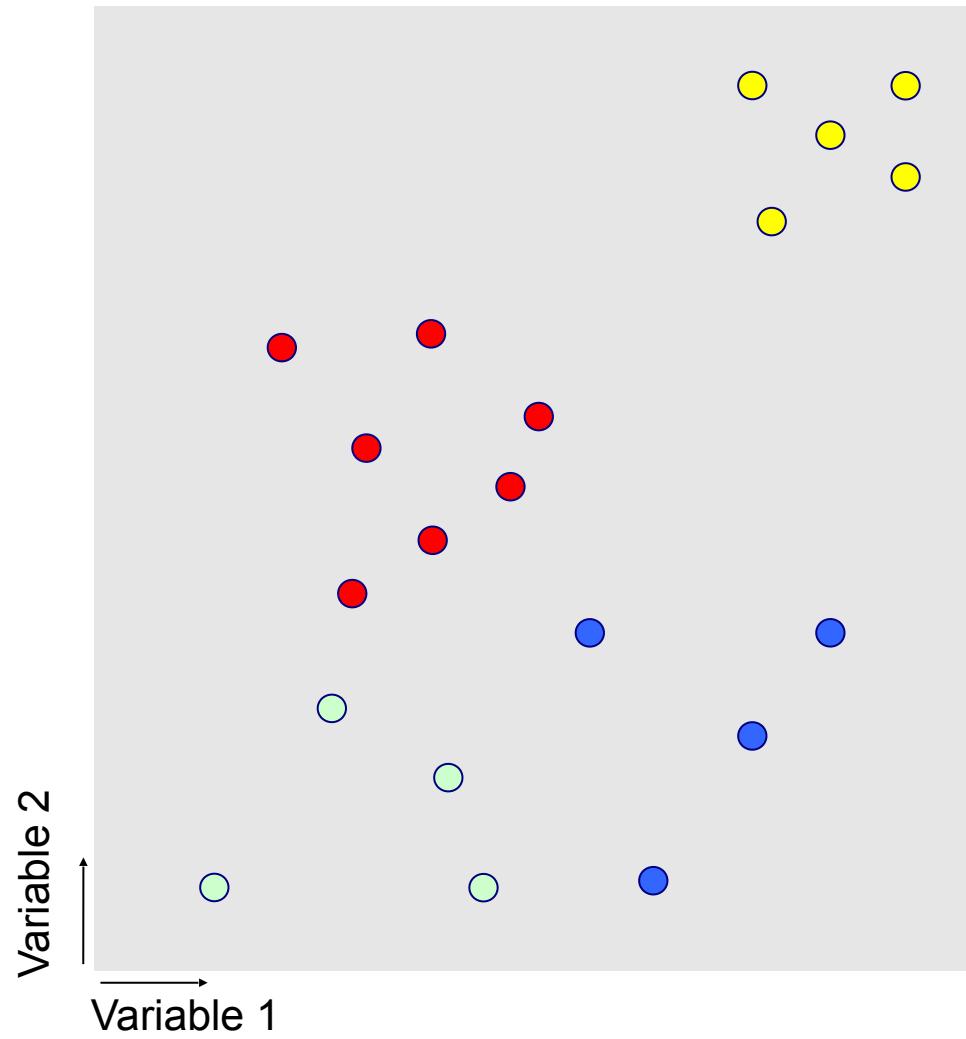
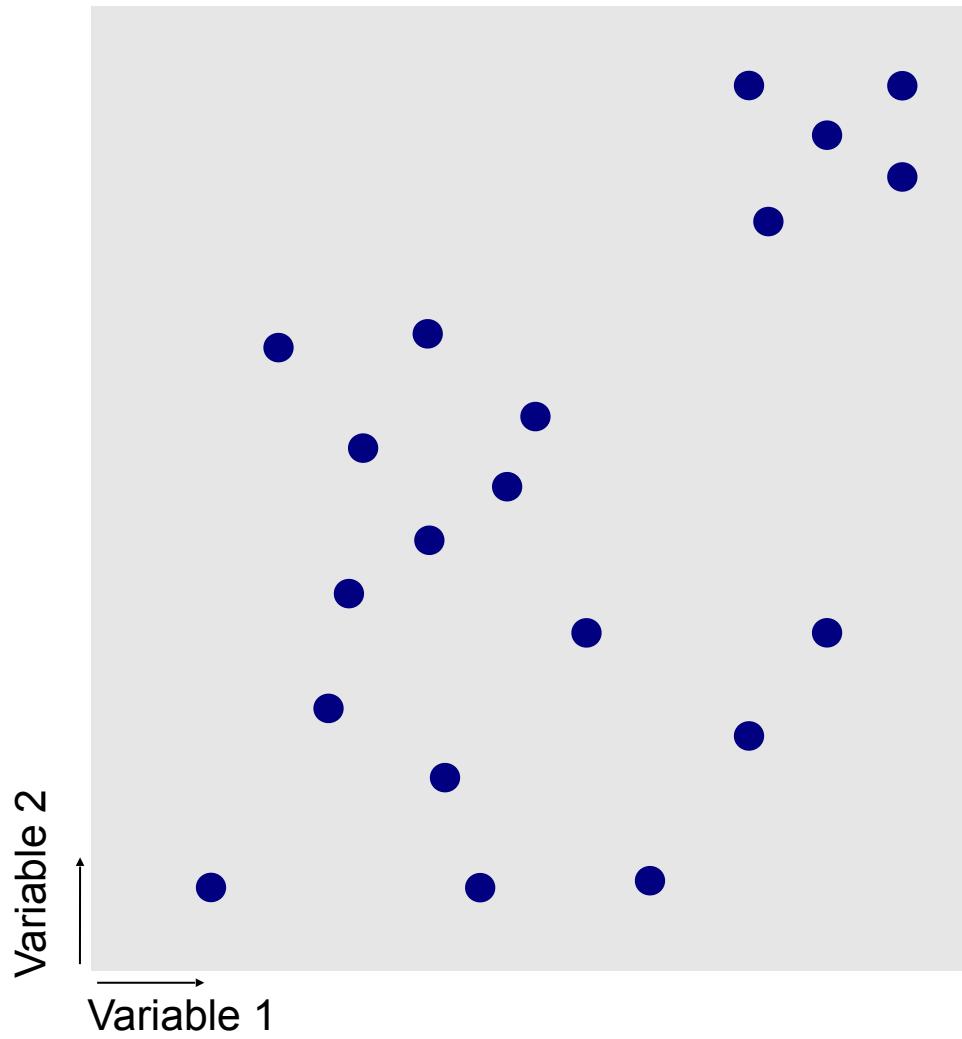
- n Labels $Y_i \in \{-1, 1\}^K$
- \times $Y_i = 1$
- \bullet $Y_i = -1$
- Ici $n = 20$, $p = 2$ et $K=1$

1. **Choix d'un modèle** pour séparer les données d'apprentissage, i.e. les \bullet et les \times .
2. **Apprentissage des paramètres** optimaux
3. Une fois les paramètres du modèle appris, **prédiction** extrêmement simple et rapide de \blacksquare .

1) Introduction - Exploration de données pour la détection de groupes (exemple 1)



1) Introduction - Exploration de données pour la détection de groupes (exemple 1)



1) Introduction - Exploration de données pour la détection de groupes (exemple 2)

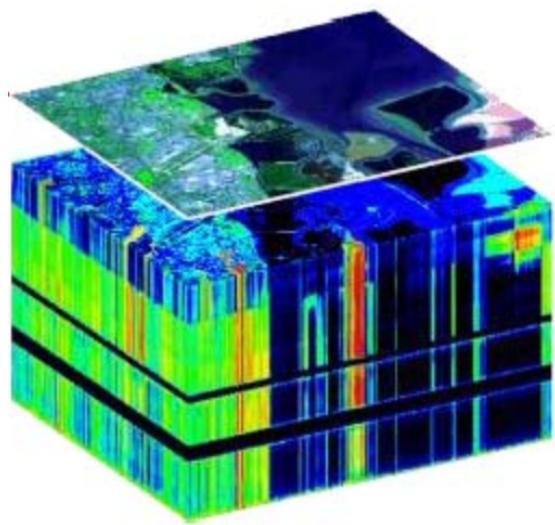
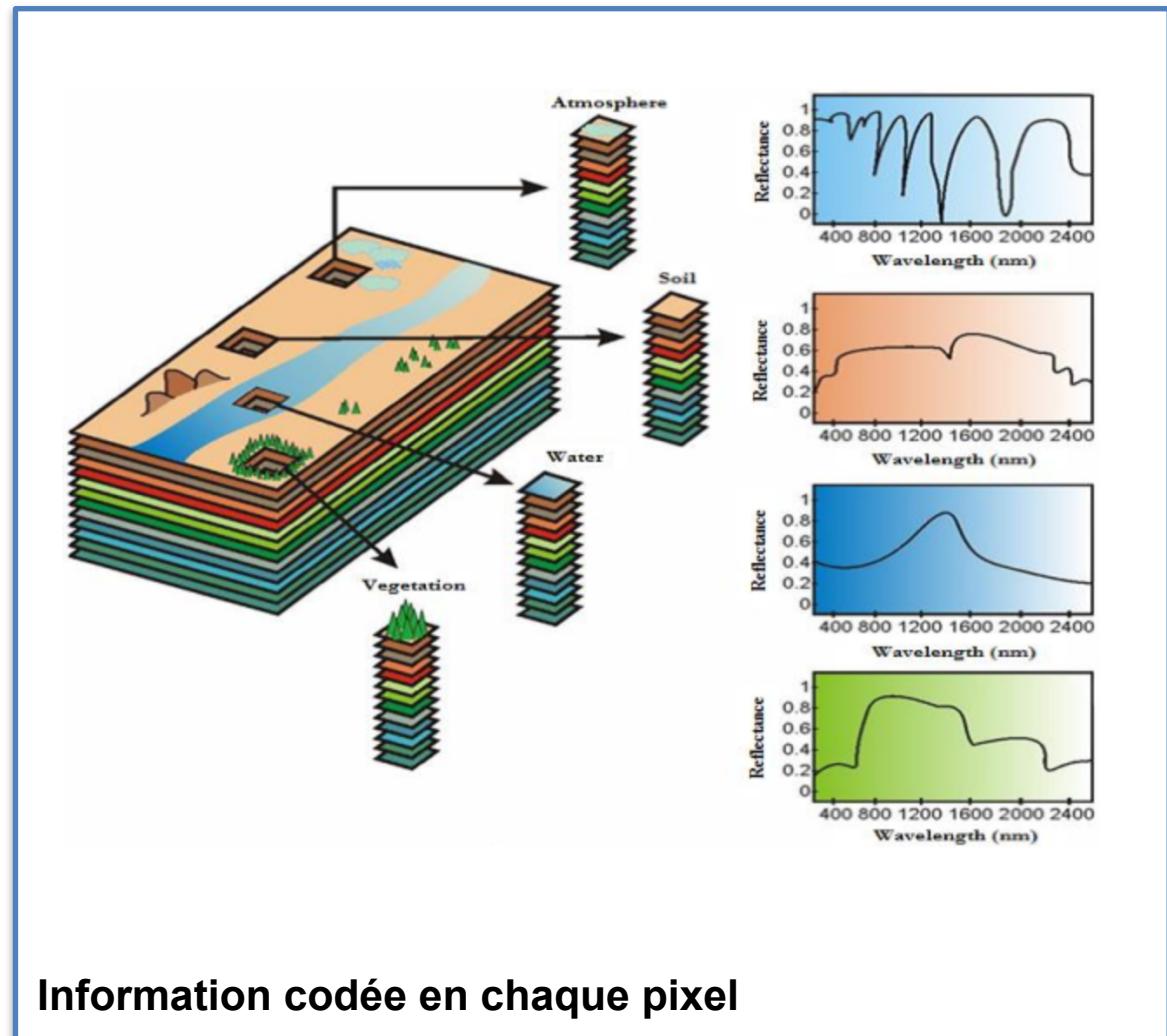
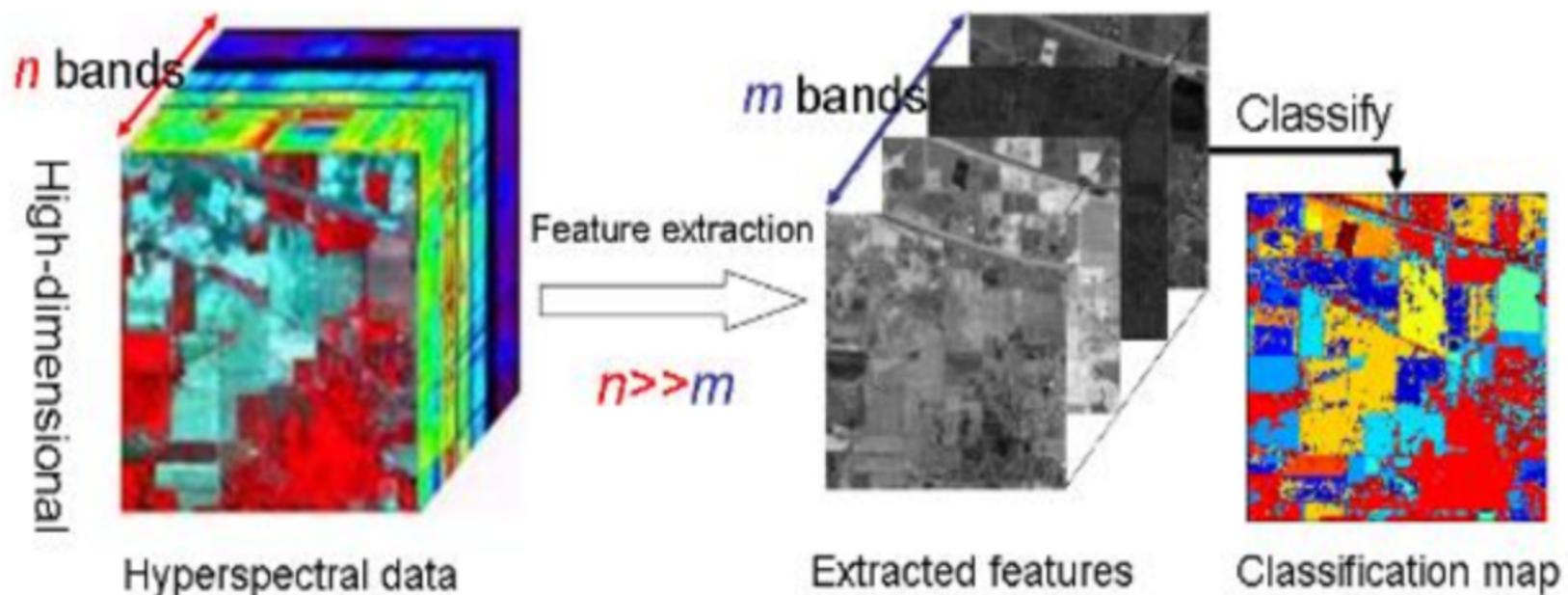


Image hyper-spectrale



Images : Boggavarapu et Prabukumar: « Survey on classification methods for hyper spectral remote sensing imagery », ICICCS 2017

1) Introduction - Exploration de données pour la détection de groupes (exemple 2)

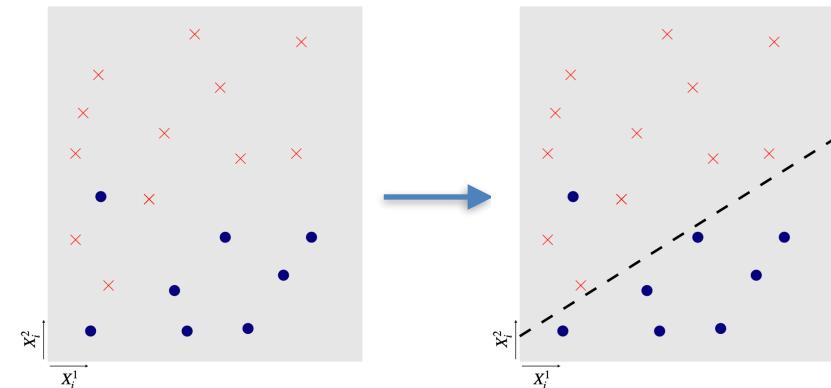


Classification des différents types de bandes sans l'intervention d'un expert qui devrait annoter BEAUCOUP de pixels !

1) Introduction - Apprentissage supervisé v.s. non-supervisé

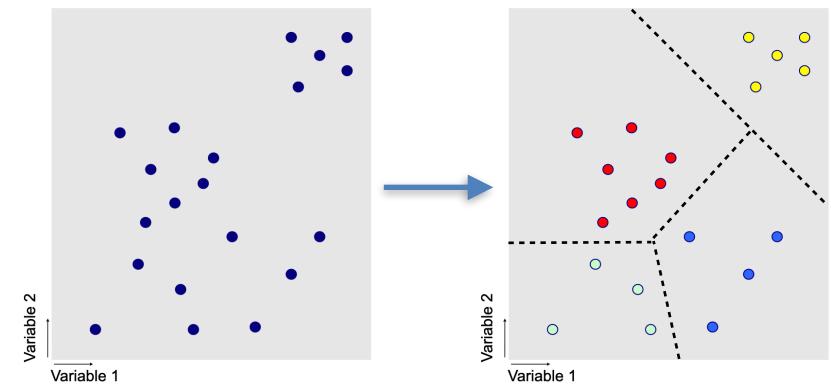
Cas supervisé :

- Les paramètres θ du **modèle** ϕ_θ sont optimisés pour prédire des **sorties** $\widehat{Y}_i = f_\theta(X_i)$ à partir de **données d'entrée** X_i .
- On connaît les sorties Y_i recherchées pour chaque X_i dans les observations d'apprentissage !
- Un risque empirique entre les prédictions $\widehat{Y}_i = f_\theta(X_i)$ et les vrais sorties Y_i est minimisé à partir d'observations d'apprentissage $(X_i, Y_i)_{i=1,\dots,n}$



Cas non supervisé :

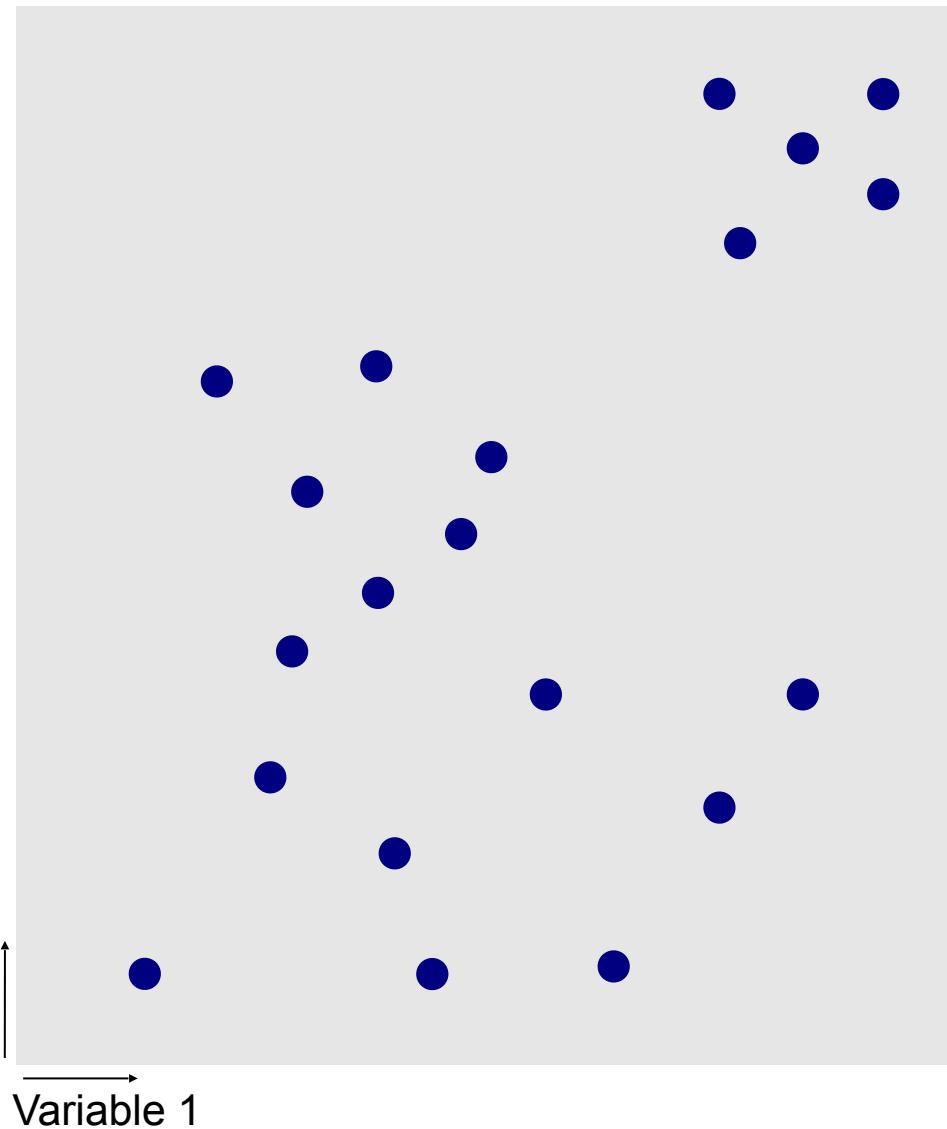
- Les paramètres θ du **modèle** ϕ_θ sont optimisés pour prédire l'appartenance à un groupe $\widehat{Y}_i = f_\theta(X_i)$ à partir de **données d'entrée** X_i .
- Aucune sortie de référence est connue.
- La fonction optimisée dépend alors seulement de la **distribution** des $\{X_i\}_{i=1,\dots,n}$ et des **labels** attribués \widehat{Y}_i



Modèles classiques en apprentissage automatique

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

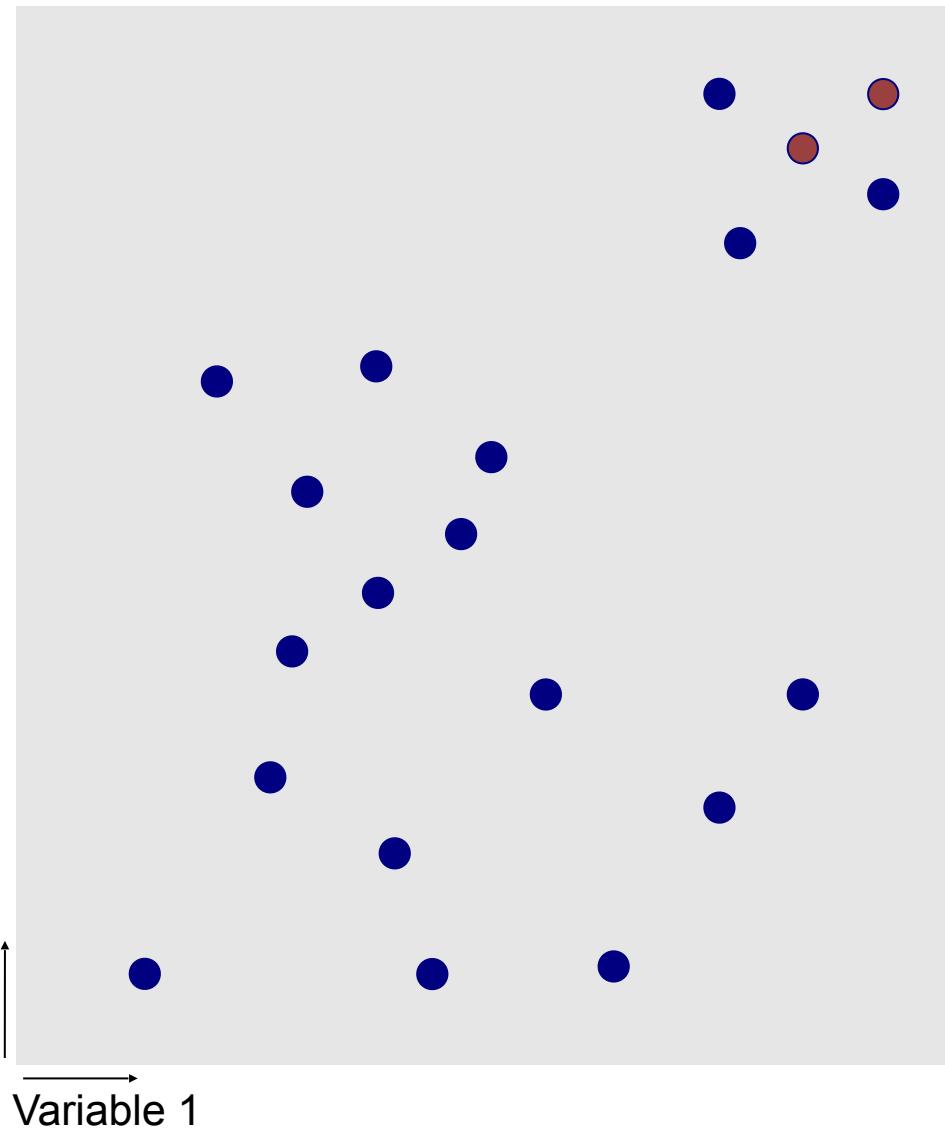
Classification ascendante hiérarchique (CAH)



On regroupe itérativement les graines ou groupes de graines les plus proches.

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

Classification ascendante hiérarchique (CAH)

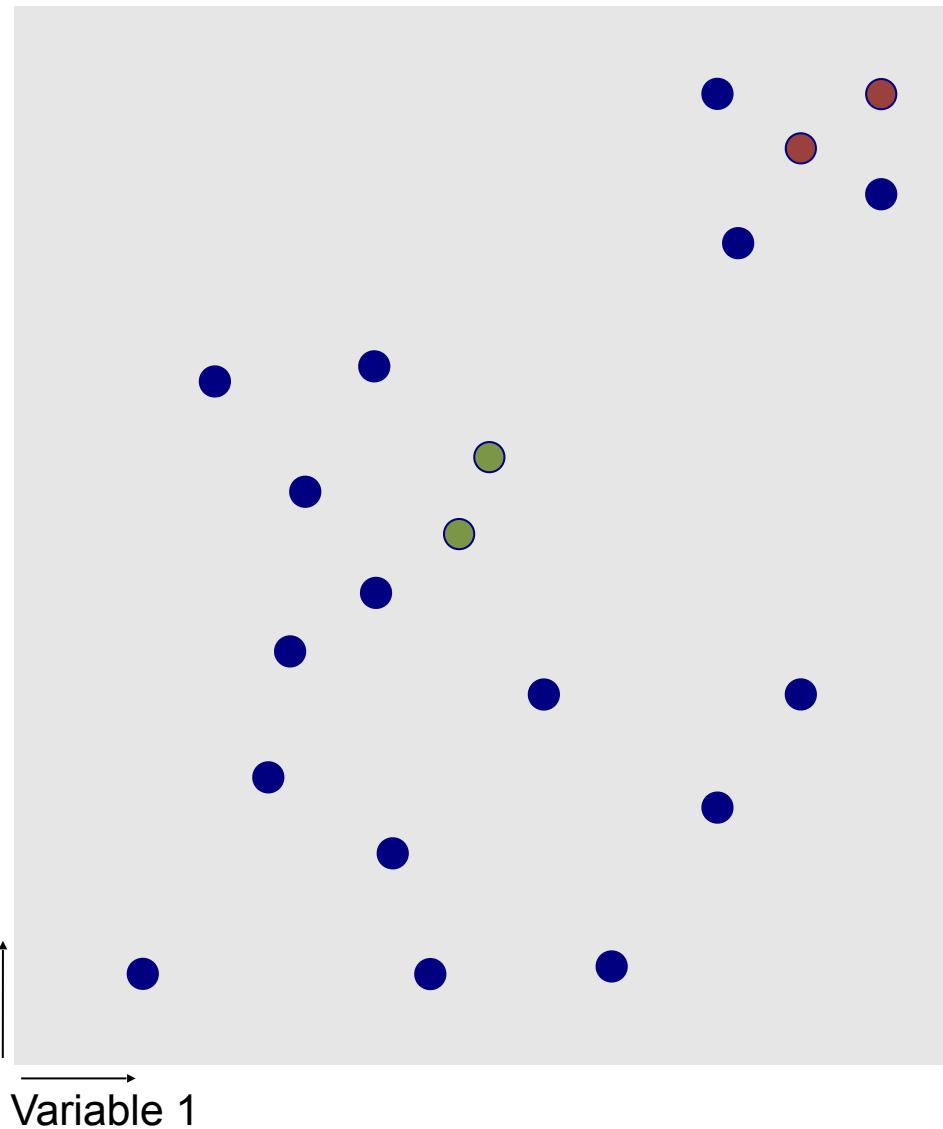


On regroupe itérativement les graines ou groupes de graines les plus proches.

Remarque : Une distance Euclidienne est utilisée pour les points les plus proches

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

Classification ascendante hiérarchique (CAH)

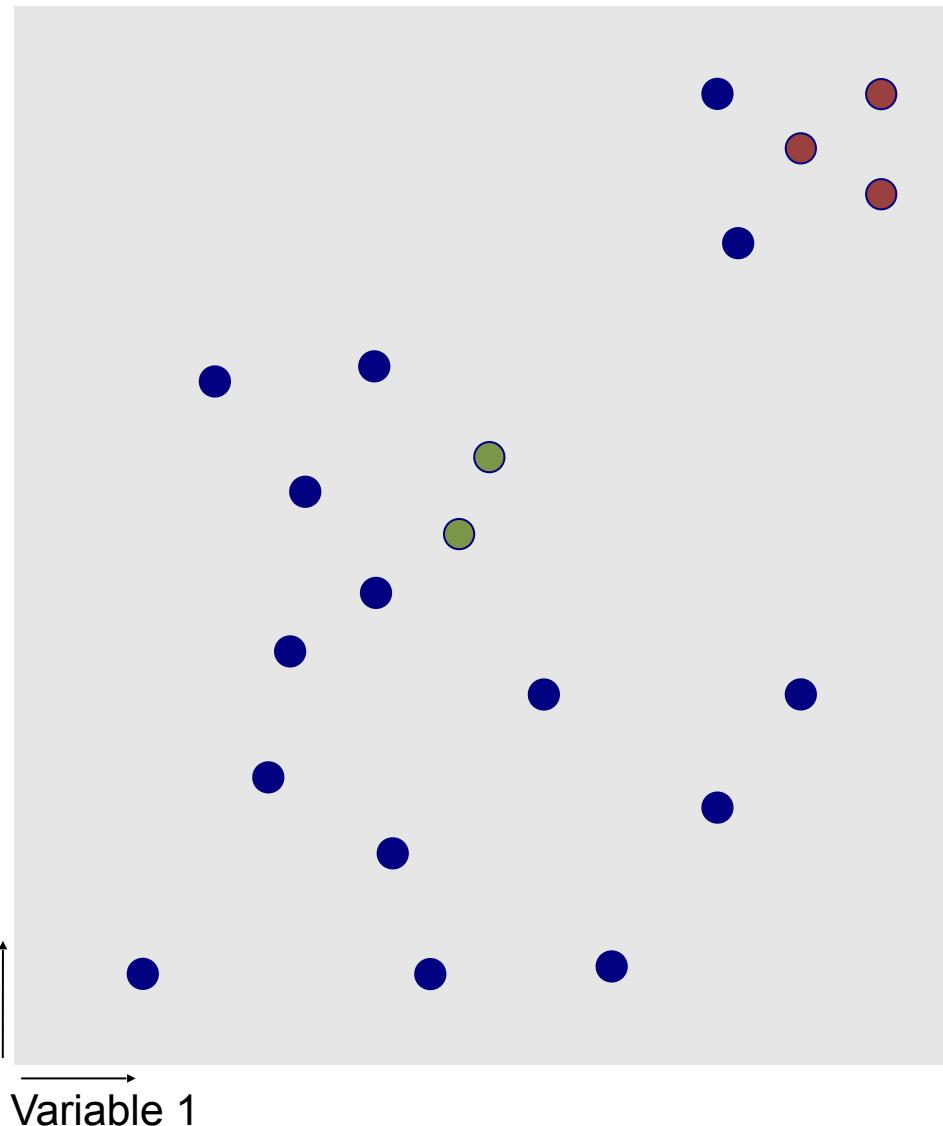


On regroupe itérativement les graines ou groupes de graines les plus proches.

Remarque : Une stratégie est définie pour comparer des groupes de points.

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

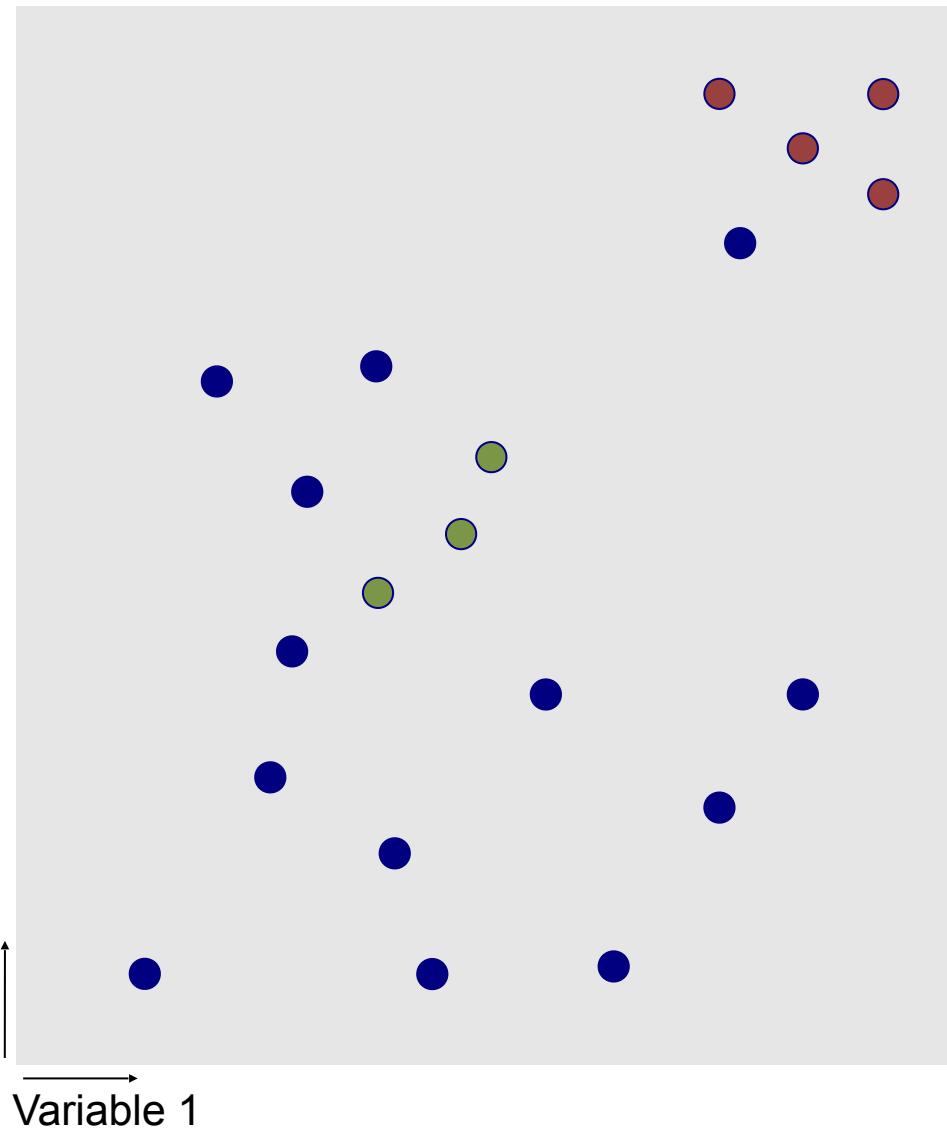
Classification ascendante hiérarchique (CAH)



On regroupe itérativement les graines ou groupes de graines les plus proches.

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

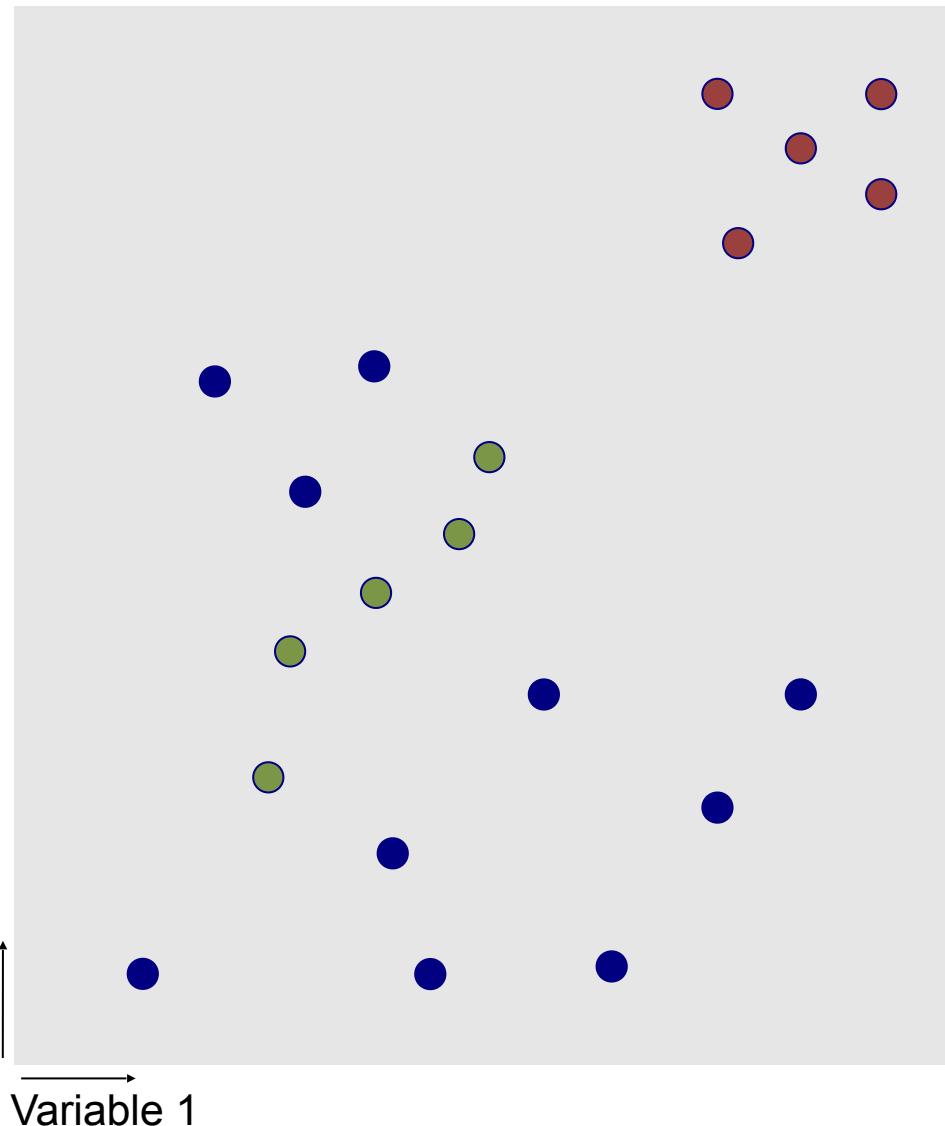
Classification ascendante hiérarchique (CAH)



On regroupe itérativement les graines ou groupes de graines les plus proches.

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

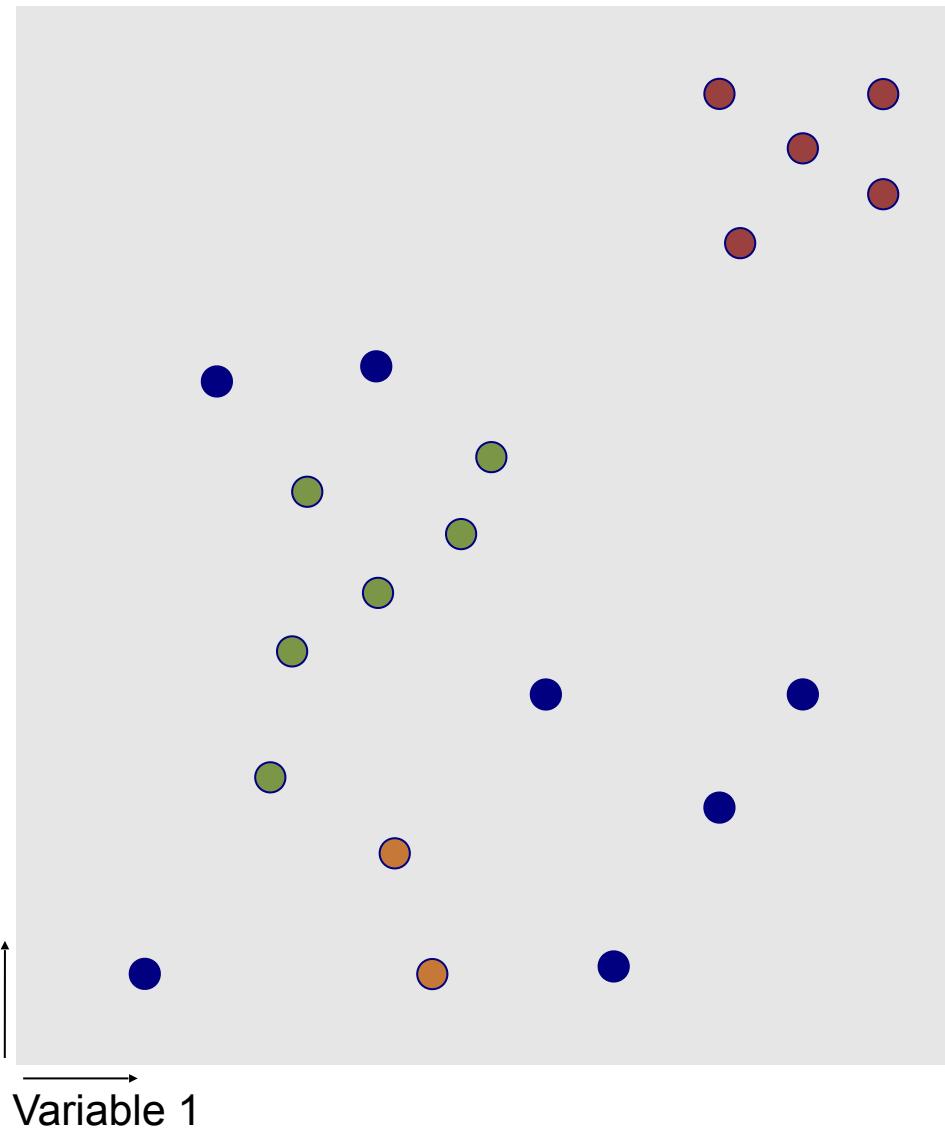
Classification ascendante hiérarchique (CAH)



On regroupe itérativement les graines ou groupes de graines les plus proches.

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

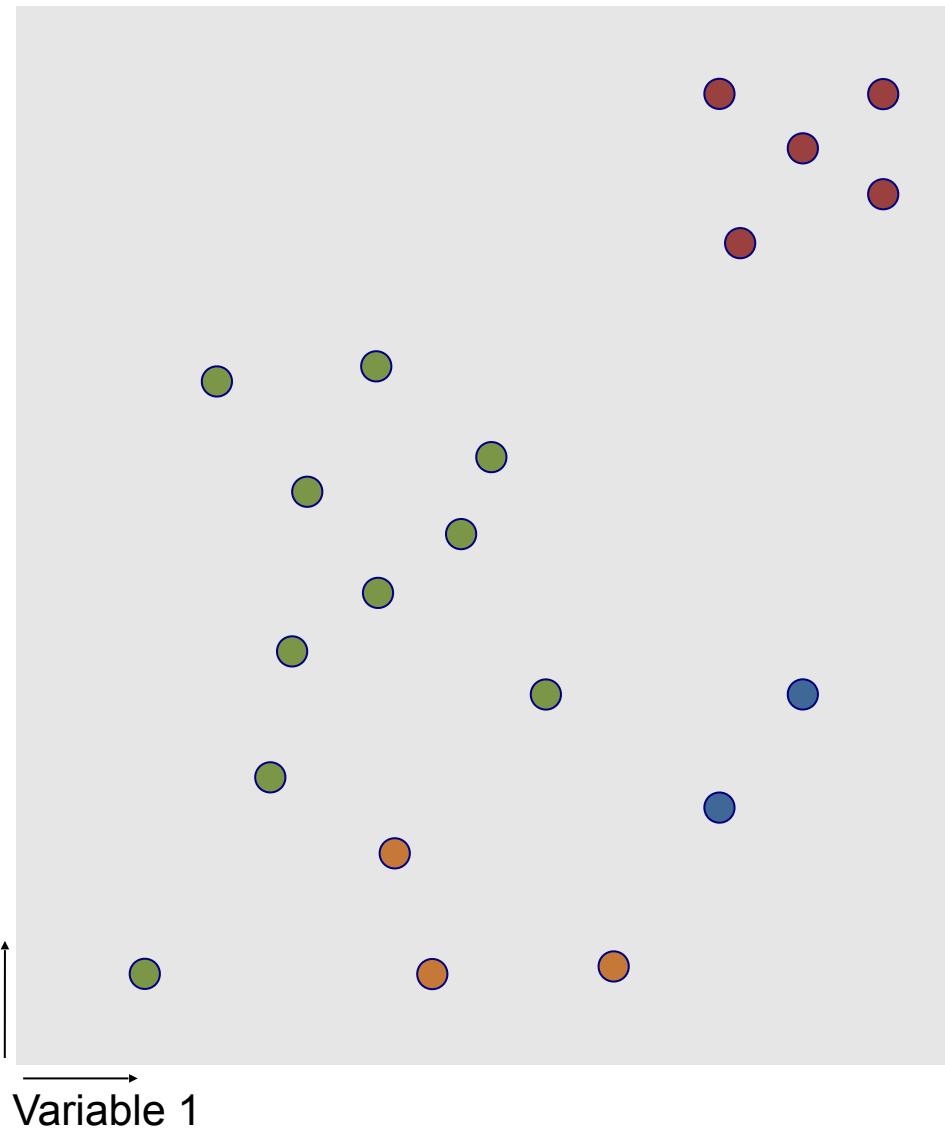
Classification ascendante hiérarchique (CAH)



On regroupe itérativement les graines ou groupes de graines les plus proches.

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

Classification ascendante hiérarchique (CAH)



On regroupe itérativement les graines ou groupes de graines les plus proches.

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

Plus formellement → Hastie et al.: « The elements of Statistical Learning ». Ed. Springer

The results of applying K -means or K -medoids clustering algorithms depend on the choice for the number of clusters to be searched and a starting configuration assignment. In contrast, hierarchical clustering methods do not require such specifications. Instead, they require the user to specify a measure of dissimilarity between (disjoint) groups of observations, based on the pairwise dissimilarities among the observations in the two groups.

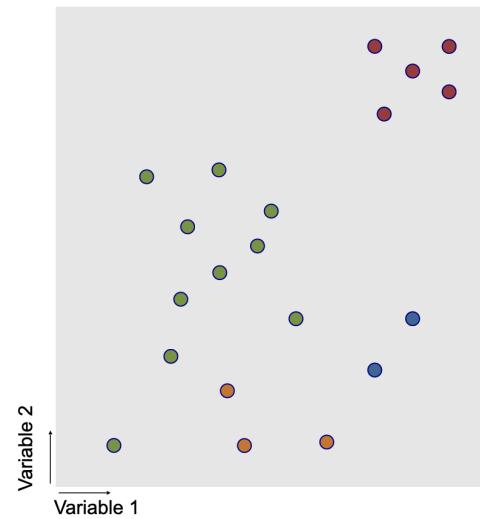
⋮

Strategies for hierarchical clustering divide into two basic paradigms: *agglomerative* (bottom-up) and *divisive* (top-down).

⋮

All agglomerative and some divisive methods (when viewed bottom-up) possess a monotonicity property. That is, the dissimilarity between merged clusters is monotone increasing with the level of the merger. Thus the binary tree can be plotted so that the height of each node is proportional to the value of the intergroup dissimilarity between its two daughters. The terminal nodes representing individual observations are all plotted at zero height. This type of graphical display is called a *dendrogram*.

A dendrogram provides a highly interpretable complete description of the hierarchical clustering in a graphical format. This is one of the main reasons for the popularity of hierarchical clustering methods.



2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

Plus formellement → Hastie et al.: « The elements of Statistical Learning ». Ed. Springer

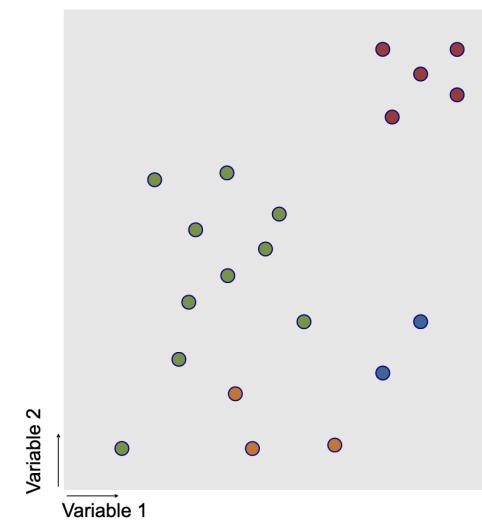
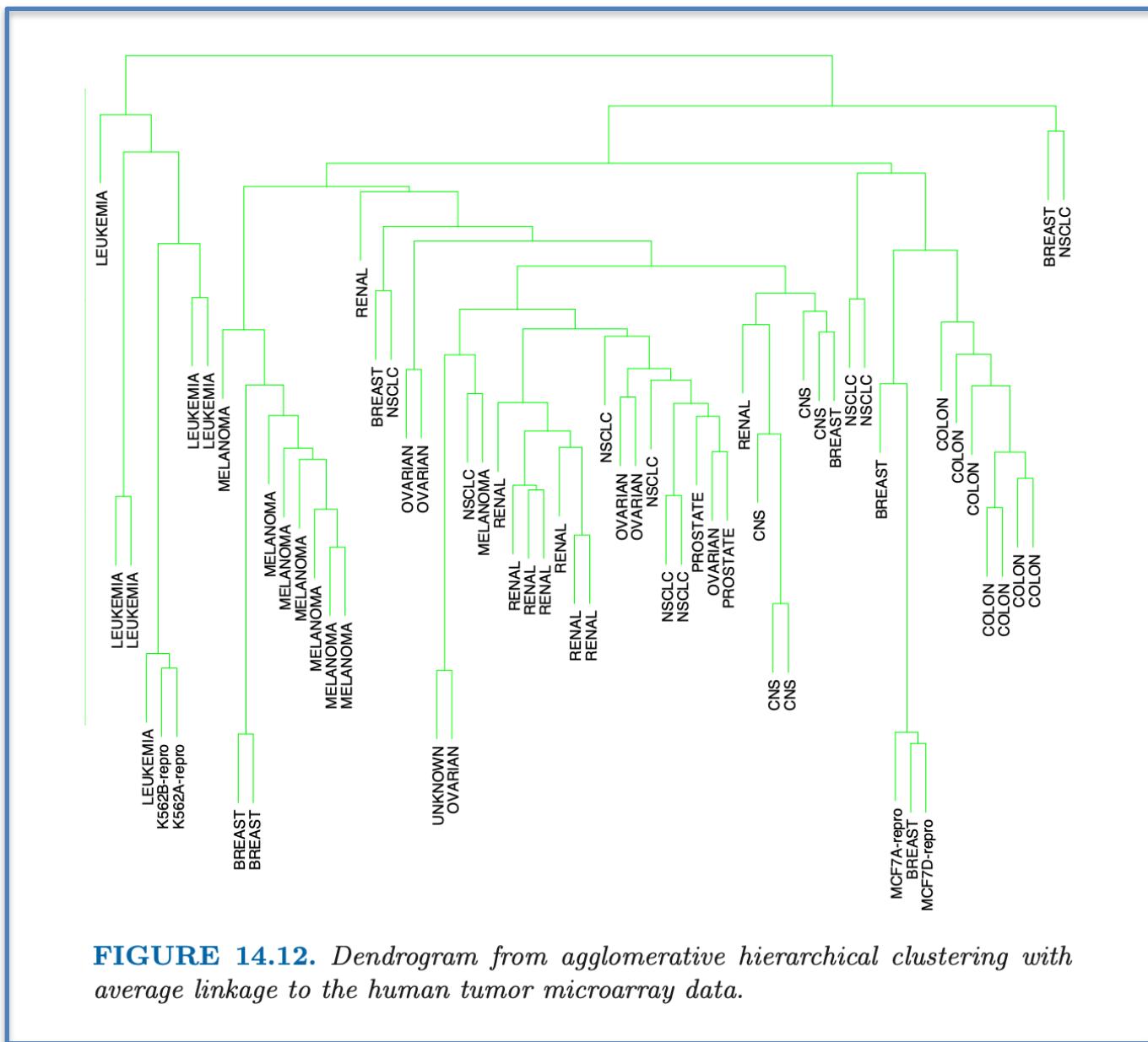


FIGURE 14.12. Dendrogram from agglomerative hierarchical clustering with average linkage to the human tumor microarray data.

2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

Plus formellement → Hastie et al.: « The elements of Statistical Learning ». Ed. Springer

Let G and H represent two such groups. The dissimilarity $d(G, H)$ between G and H is computed from the set of pairwise observation dissimilarities $d_{ii'}$ where one member of the pair i is in G and the other i' is in H . **Single linkage (SL)** agglomerative clustering takes the intergroup dissimilarity to be that of the closest (least dissimilar) pair

$$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'}. \quad (14.41)$$

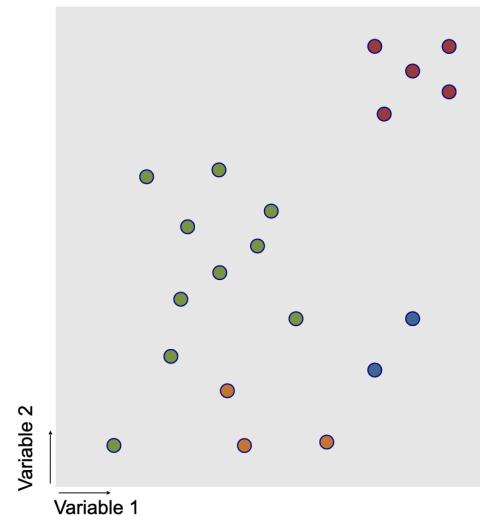
This is also often called the *nearest-neighbor* technique. **Complete linkage (CL)** agglomerative clustering (*furthest-neighbor* technique) takes the intergroup dissimilarity to be that of the furthest (most dissimilar) pair

$$d_{CL}(G, H) = \max_{\substack{i \in G \\ i' \in H}} d_{ii'}. \quad (14.42)$$

Group average (GA) clustering uses the average dissimilarity between the groups

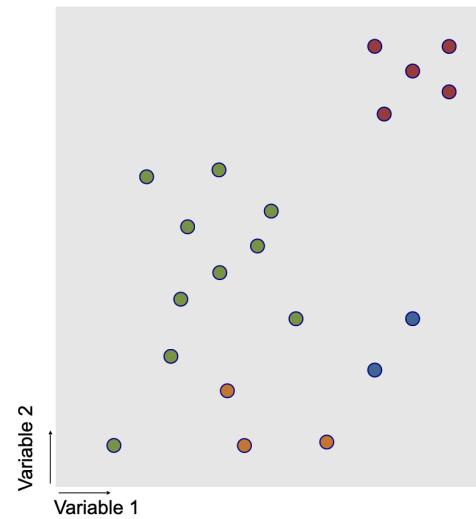
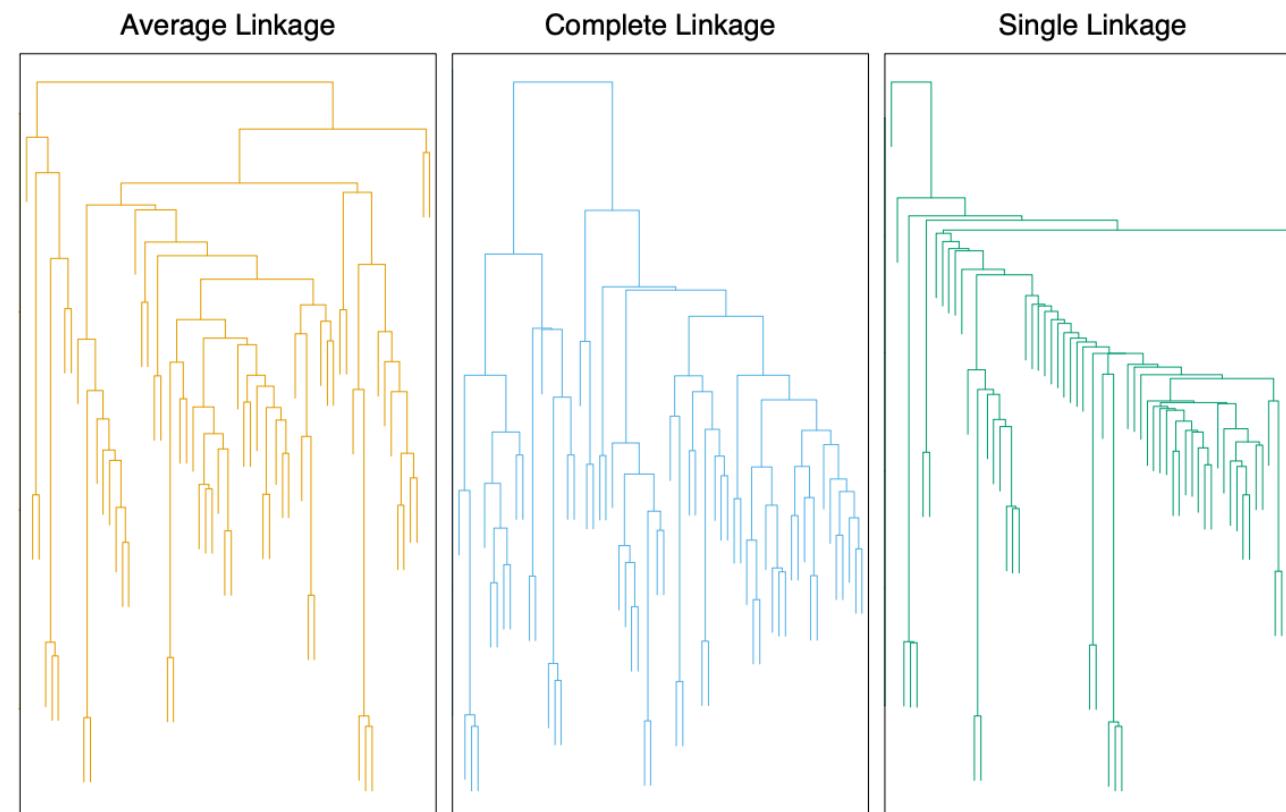
$$d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'} \quad (14.43)$$

where N_G and N_H are the respective number of observations in each group.

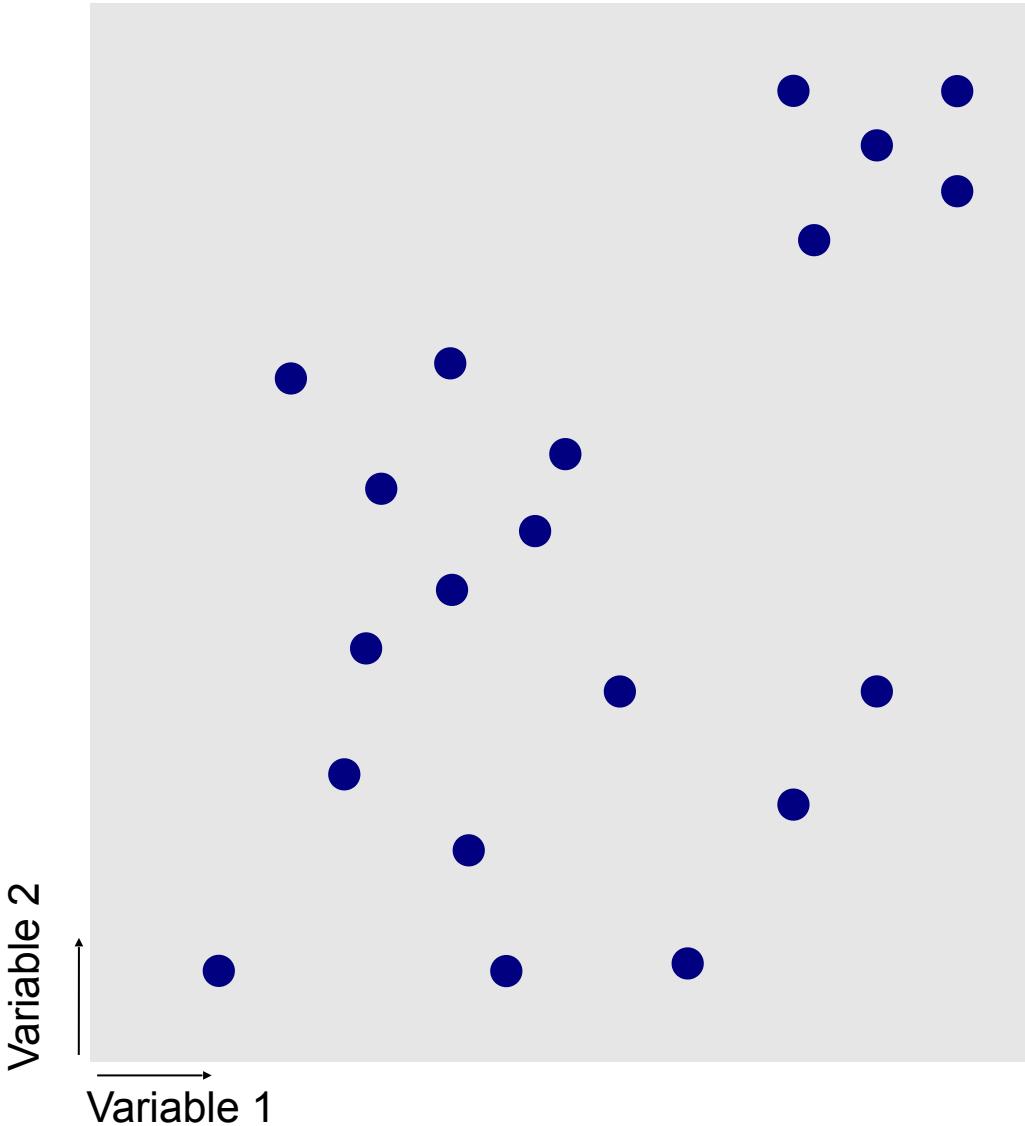


2) Deux modèles classiques → Classification ascendante hiérarchique (CAH)

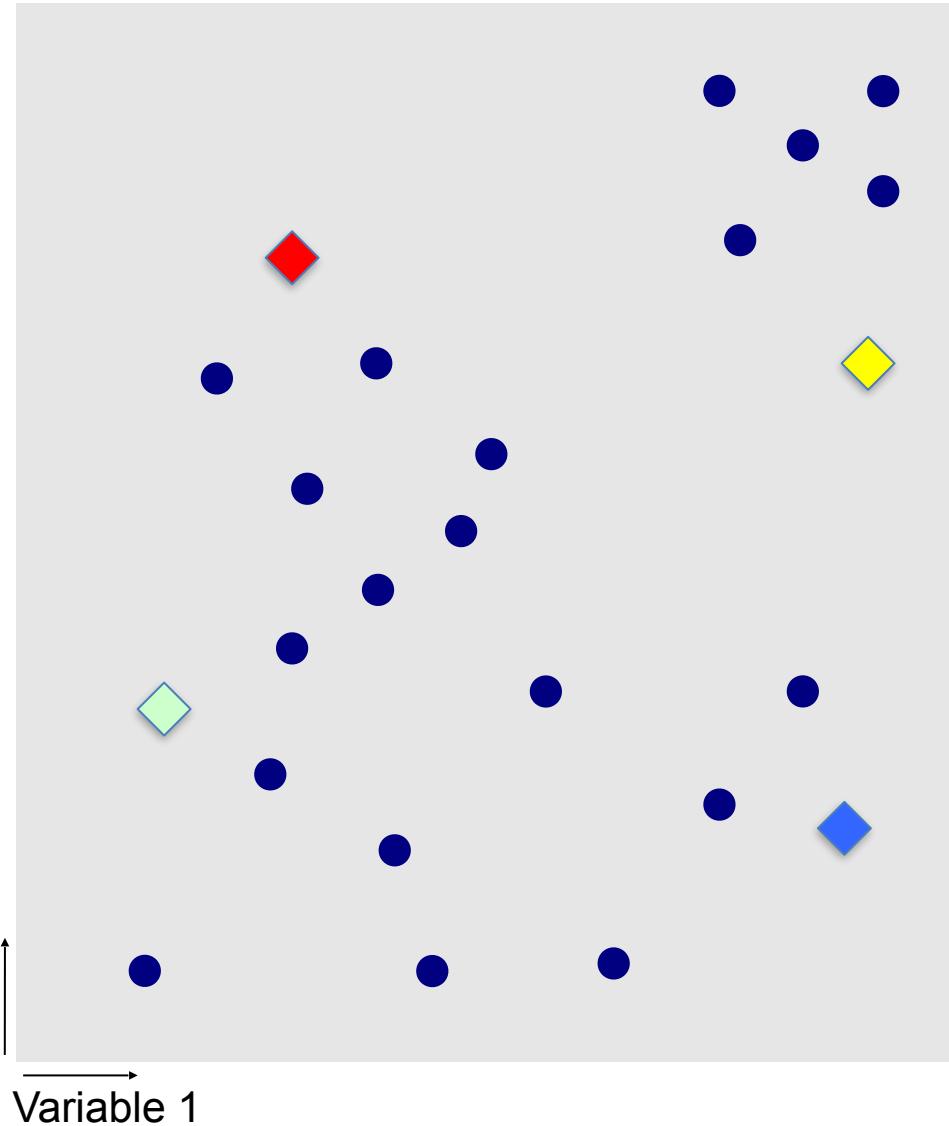
Plus formellement → Hastie et al.: « The elements of Statistical Learning ». Ed. Springer



2) Deux modèles classiques → K-means

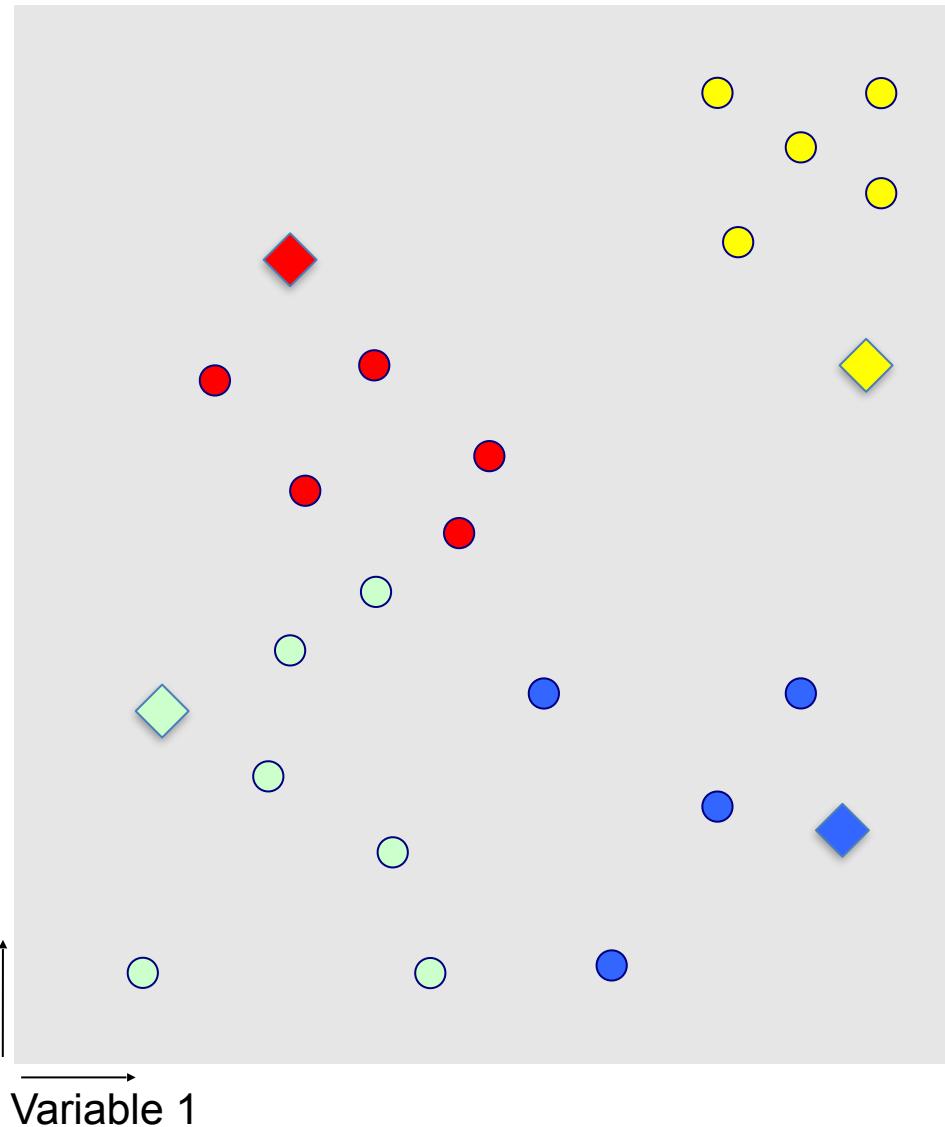


2) Deux modèles classiques → K-means



On tire K graines au hasard
(pour l'exemple K=4)

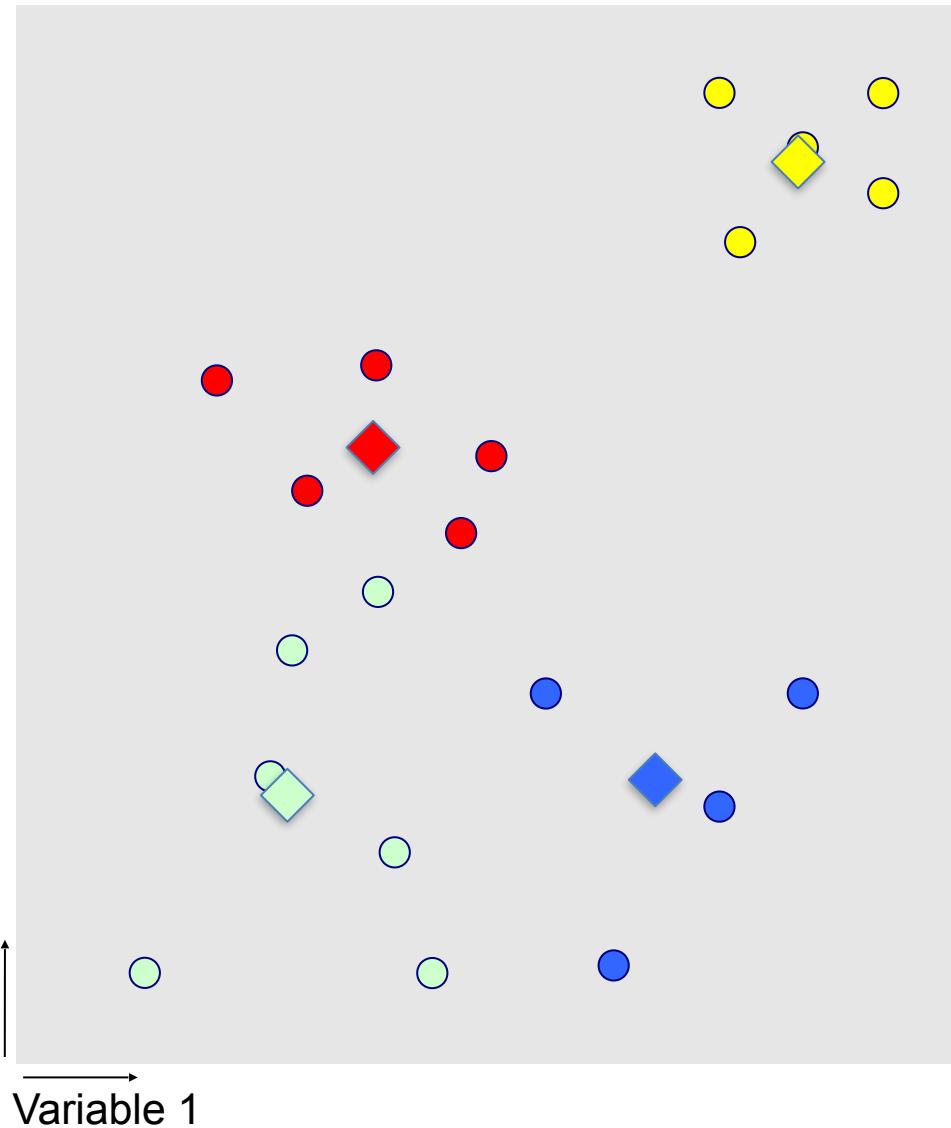
2) Deux modèles classiques → K-means



Pour chaque observation, on cherche la graine la plus proche.

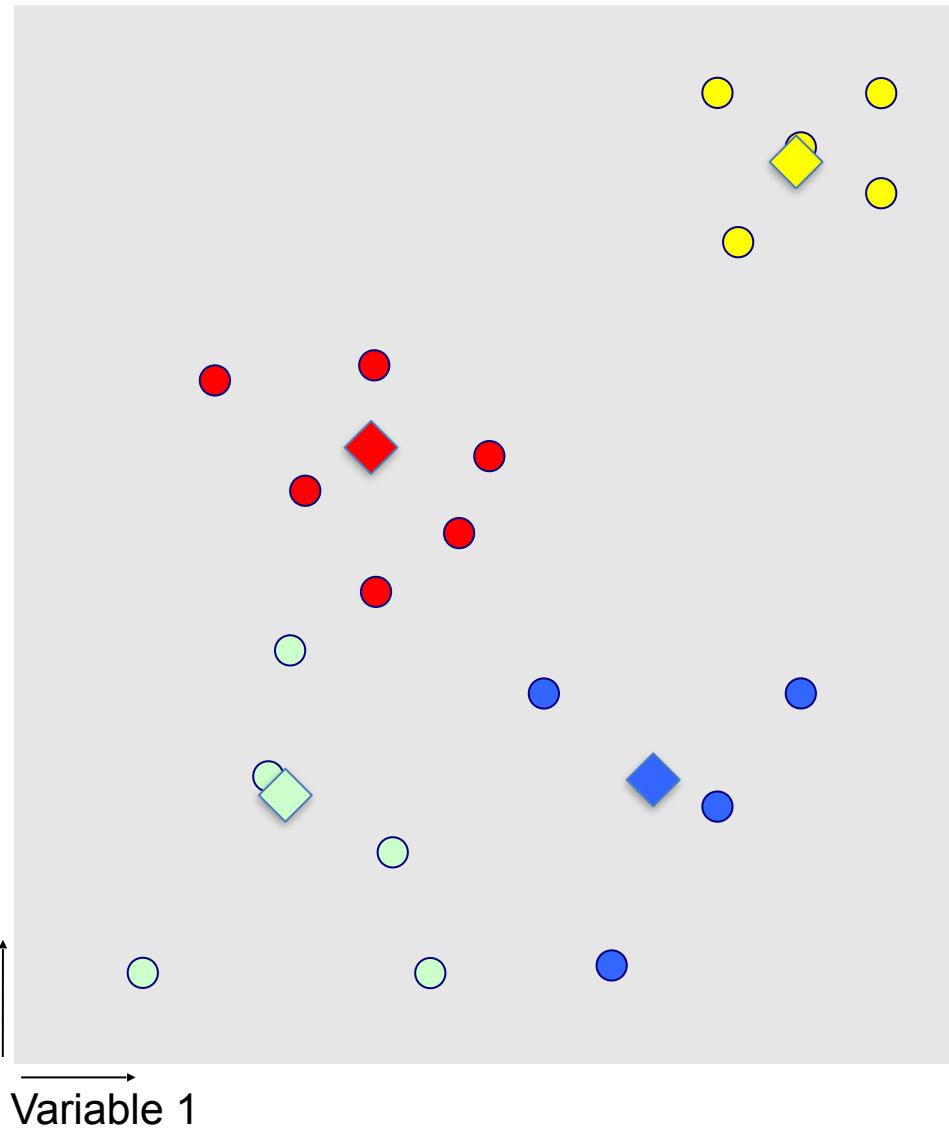
Remarque : Des distances Euclidiennes sont utilisées

2) Deux modèles classiques → K-means



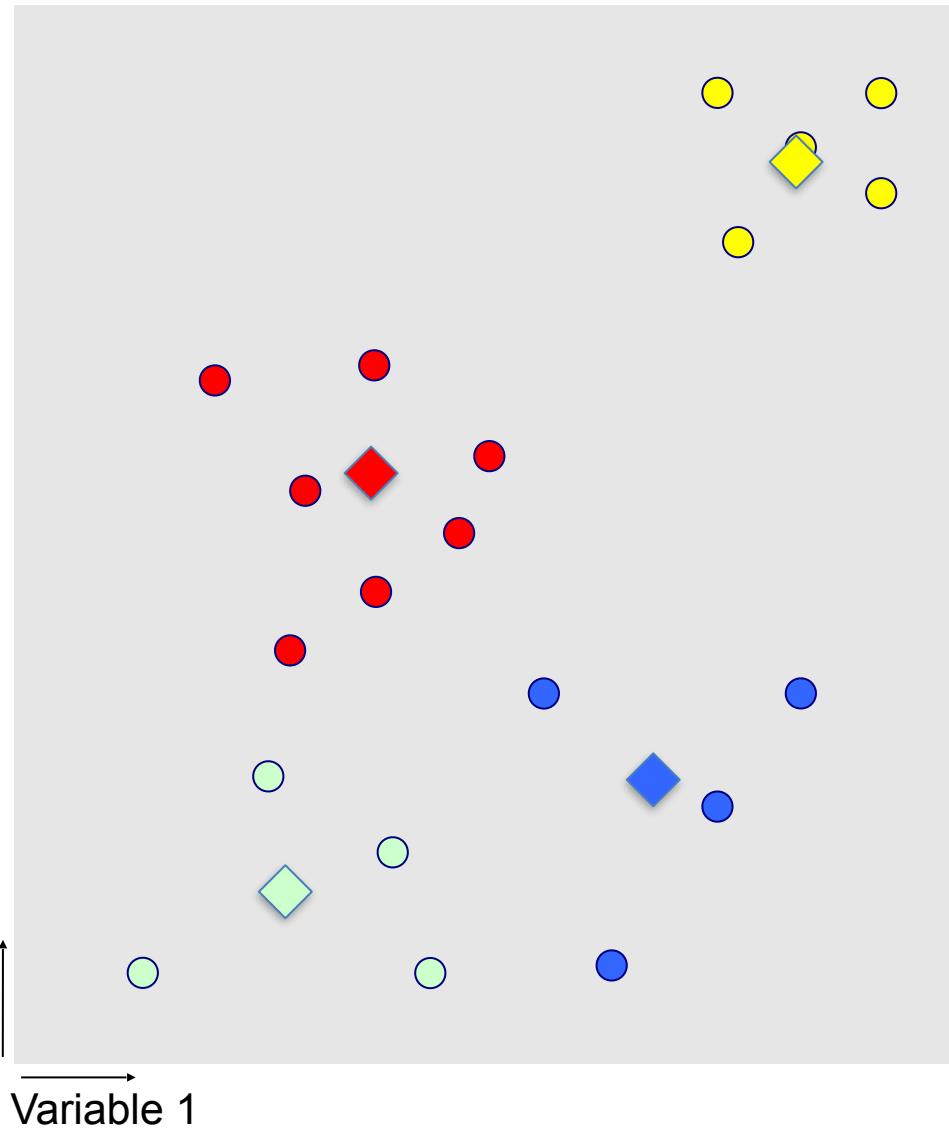
On centre les graines...

2) Deux modèles classiques → K-means



... pour chaque observation, on cherche à nouveau la graine la plus proche ...

2) Deux modèles classiques → K-means



... et on recommence jusqu'à convergence.

2) Deux modèles classiques → K-means

Plus formellement → Hastie et al.: « The elements of Statistical Learning ». Ed. Springer

The K -means algorithm is one of the most popular iterative descent clustering methods. It is intended for situations in which all variables are of the quantitative type, and squared Euclidean distance

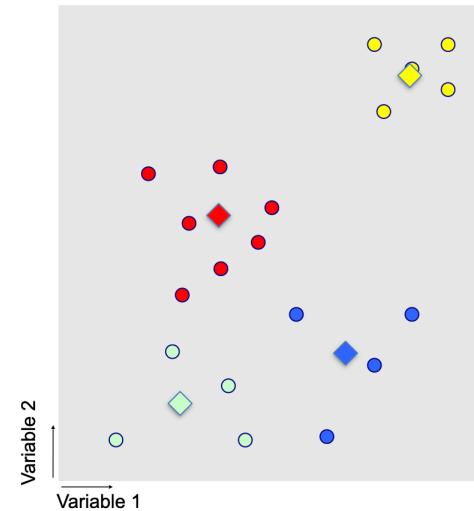
$$d(x_i, x_{i'}) = \sum_{j=1}^p (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$$

is chosen as the **dissimilarity measure**. Note that weighted Euclidean distance can be used by redefining the x_{ij} values (Exercise 14.1).

The within-point scatter (14.28) can be written as

$$\begin{aligned} W(C) &= \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i')=k} \|x_i - x_{i'}\|^2 \\ &= \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2, \end{aligned} \quad (14.31)$$

where $\bar{x}_k = (\bar{x}_{1k}, \dots, \bar{x}_{pk})$ is the mean vector associated with the k th cluster, and $N_k = \sum_{i=1}^N I(C(i) = k)$. Thus, the criterion is minimized by assigning the N observations to the **K clusters** in such a way that within each cluster the average dissimilarity of the observations from the cluster mean, as defined by the points in that cluster, is minimized.



2) Deux modèles classiques → K-means

Plus formellement → Hastie et al.: « The elements of Statistical Learning ». Ed. Springer

An iterative descent algorithm for solving

$$C^* = \min_C \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2$$

can be obtained by noting that for any set of observations S

$$\bar{x}_S = \operatorname{argmin}_m \sum_{i \in S} \|x_i - m\|^2. \quad (14.32)$$

Hence we can obtain C^* by solving the enlarged optimization problem

$$\min_{C, \{m_k\}_1^K} \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - m_k\|^2. \quad (14.33)$$

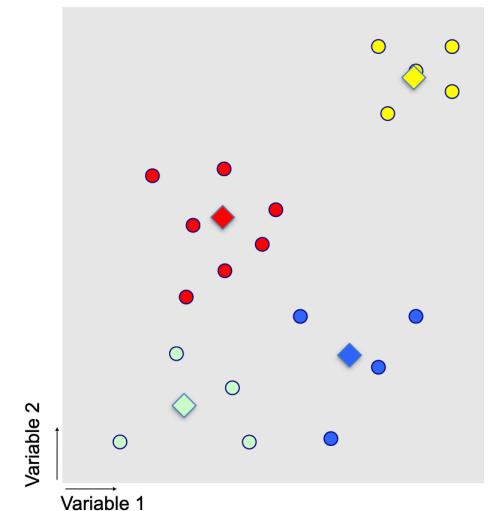
This can be minimized by an alternating optimization procedure given in Algorithm 14.1.

Algorithm 14.1 K-means Clustering.

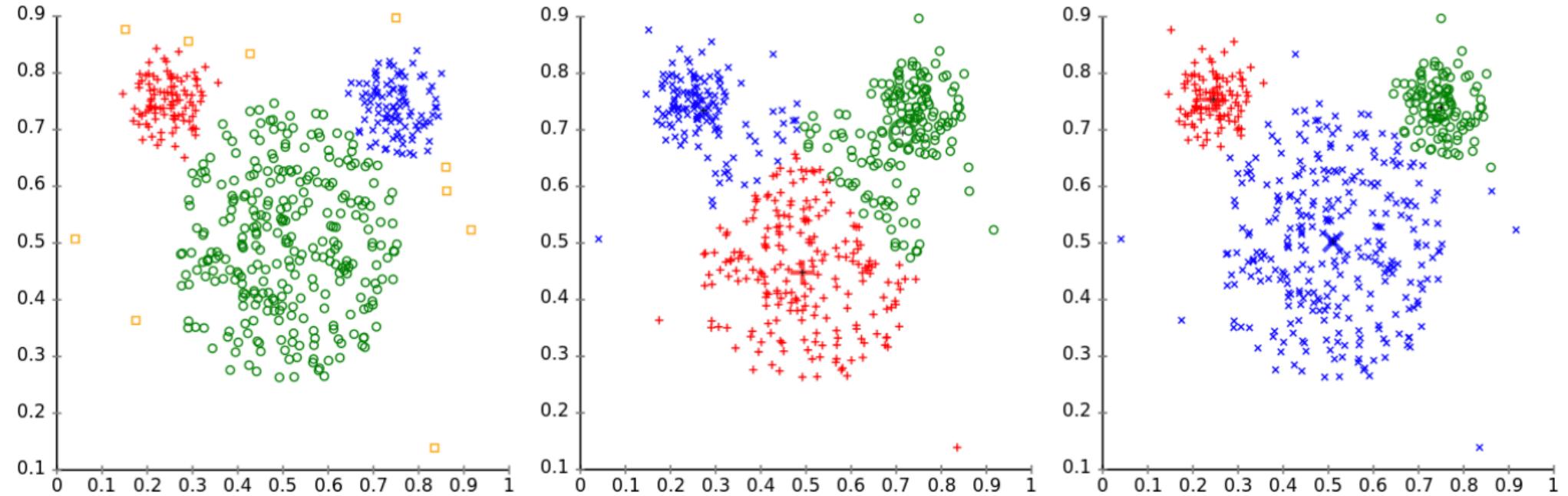
1. For a given cluster assignment C , the total cluster variance (14.33) is minimized with respect to $\{m_1, \dots, m_K\}$ yielding the means of the currently assigned clusters (14.32).
2. Given a current set of means $\{m_1, \dots, m_K\}$, (14.33) is minimized by assigning each observation to the closest (current) cluster mean. That is,

$$C(i) = \operatorname{argmin}_{1 \leq k \leq K} \|x_i - m_k\|^2. \quad (14.34)$$

3. Steps 1 and 2 are iterated until the assignments do not change.



3) Pour aller plus loin → mélange de Gaussiennes



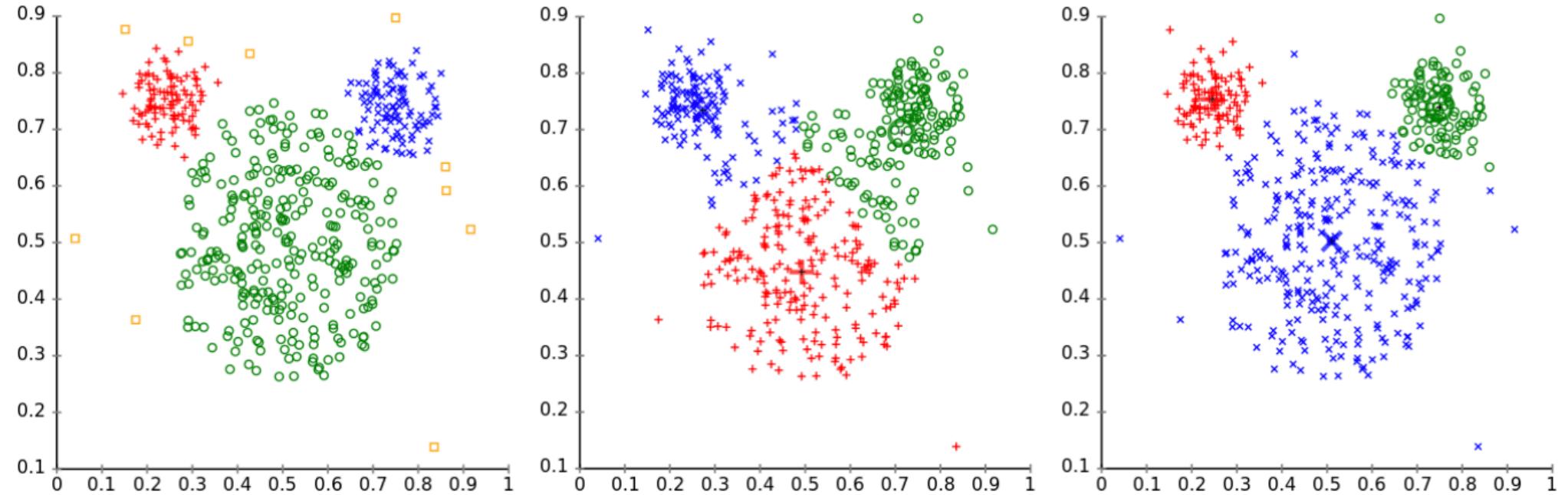
Cluster Analysis: Original Data (left), k-means (middle), EM (right) (Illustration by Chire)

→ L'étendue de la distribution spatiale des données dans chaque groupe est différente

Source :

<https://towardsdatascience.com/a-comparison-between-k-means-clustering-and-expectation-maximization-estimation-for-clustering-8c75a1193eb7>

3) Pour aller plus loin → mélange de Gaussiennes



Cluster Analysis: Original Data (left), k-means (middle), EM (right) (Illustration by Chire)

→ L'étendue de la distribution spatiale des données dans chaque groupe est différente

On peut (et on va) :

- Attribuer une densité de probabilité $\mathcal{N}(\mu_k, \Sigma_k)$ sur chaque groupe $k = 1, \dots, K$.
- Estimer à partir des données les moyennes μ_k et covariances Σ_k de chaque groupe k .
- Attribuer à chaque observation son label le plus pertinent en fonction des $\{\mu_k, \Sigma_k\}_{k=1, \dots, K}$.

Source :

<https://towardsdatascience.com/a-comparison-between-k-means-clustering-and-expectation-maximization-estimation-for-clustering-8c75a1193eb7>

MERCI !!!