Introduction à la science de données

- *Définir Science de données, ses objectifs et ses outils*
- *Développer des modèles ML pour résoudre des problématiques réelles*
- *Valider, tester ses solutions*
- *Prétraiter les données observées*
- *Déployer ses solutions*

o.banouar@uca.ac.ma

Modèles machine learning non supervisé

- Clustering hiérarchique
- K-means

- Cas d'étude: https://grouplens.org/datasets/movielens/

Modèles machine learning non supervisé

- Le **clustering** (ou *regroupement*) est une technique d'**apprentissage non supervisé** qui consiste à **diviser un ensemble de données en groupes homogènes** appelés **clusters**, de sorte que les objets d'un même groupe soient **plus similaires entre eux** qu'avec ceux des autres groupes.

- Le clustering cherche à découvrir la structure cachée d'un jeu de données sans utiliser d'étiquettes préalables.

- Chaque cluster représente un comportement, une forme ou une tendance commune au sein des données.

- Il s'agit d'un **outil exploratoire** permettant de **résumer, comprendre ou visualiser des ensembles de données complexes.**

Supervisé	Non supervisé
Données étiquetées	Données non étiquetées
Exemple : classification	Exemple : clustering
Apprend à prédire une sortie	Apprend à structurer les données

Is this a dog?

Image Classification

Which pixels belong to which object?

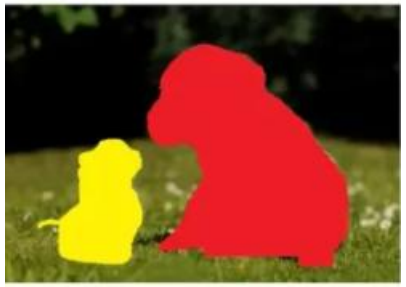
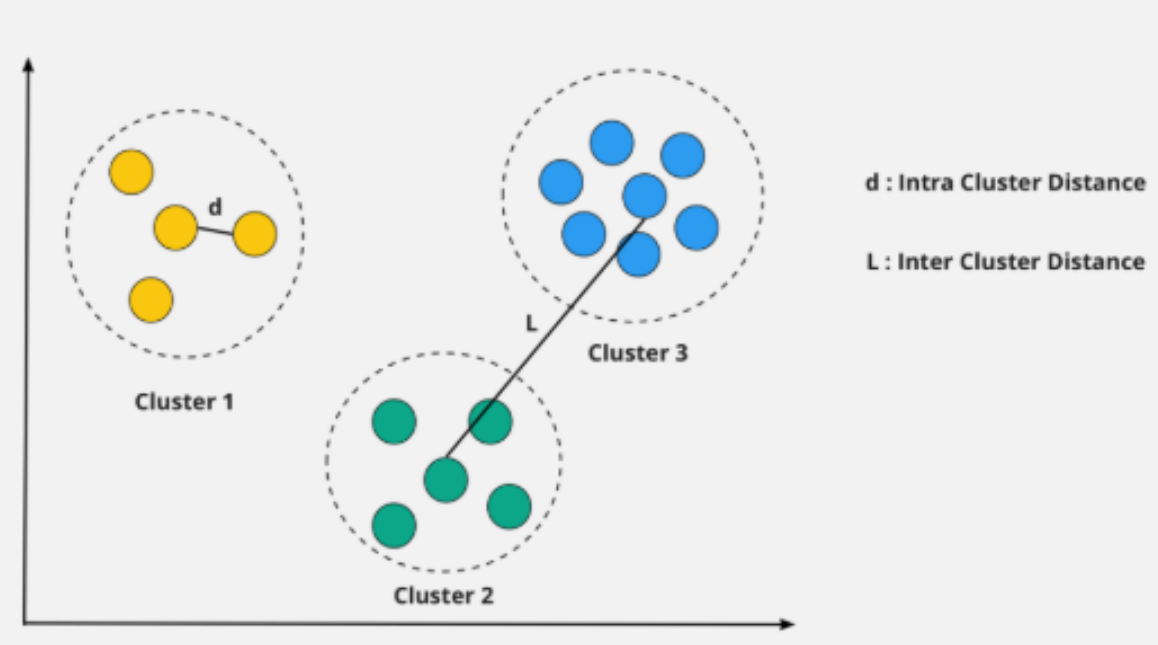


Image Segmentation

- **Similarité/distance et Clustering : quel lien ?**
- La similarité (ou sa version inverse, la distance) est au cœur du processus de clustering.
- C'est grâce à une mesure de similarité que l'algorithme peut regrouper les données proches les unes des autres dans un même cluster.

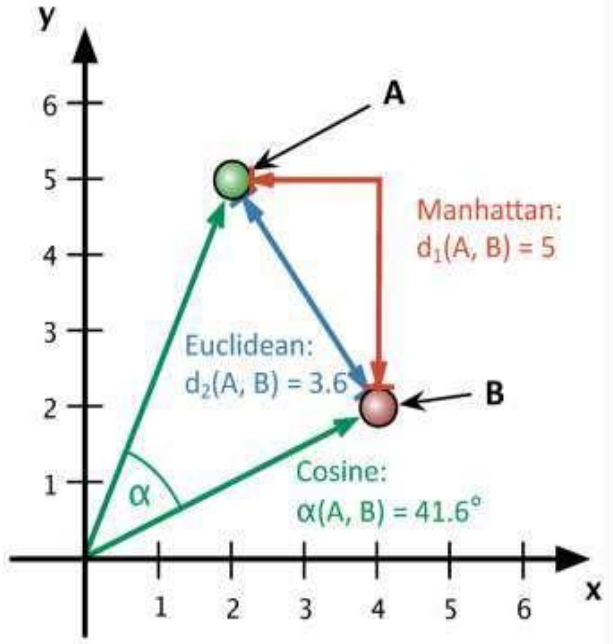


- **Pourquoi la similarité/distance est essentielle ?**
- Le clustering cherche à **minimiser** les **distances intra-cluster** (les points proches ensemble)
- Et à **maximiser** les **distances inter-clusters** (les groupes bien séparés)
- Le résultat du clustering dépend directement de la façon dont on mesure cette distance ou cette similarité

Modèles machine learning non supervisé

Situation	Mesure recommandée
Données numériques bien normalisées	Distance Euclidienne
Données avec valeurs extrêmes ou dispersées	Distance de Manhattan
Données vectorielles ou directionnelles (textes, profils utilisateurs)	Similarité cosinus

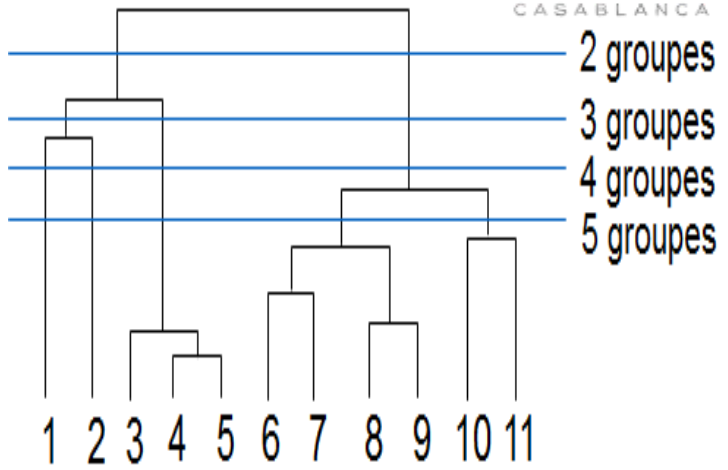
- Le choix de la mesure de similarité peut modifier complètement le résultat du clustering.
- Prétraitement (standardisation, normalisation) est souvent nécessaire pour que les distances aient un sens.



Modèles machine learning non supervisé

- Clustering hiérarchique

  - Le clustering hiérarchique est une technique de regroupement des données qui vise à créer une hiérarchie de clusters.
  - Il forme une structure arborescente (ou dendrogramme) de clusters. Cette structure permet de visualiser les relations de similarité entre les données à différentes échelles.



Il existe deux approches principales pour réaliser le clustering hiérarchique :

- Agrégative (ou ascendante) : Cette approche commence par considérer chaque point de données comme un cluster individuel et fusionne progressivement les clusters les plus similaires pour former des clusters plus grands. À chaque étape, les clusters les plus similaires sont fusionnés jusqu'à ce qu'un seul cluster global soit obtenu.

- Divisive (ou descendante) : Contrairement à l'approche agrégative, cette méthode commence par considérer tous les points de données comme un seul cluster global, puis divise itérativement ce cluster en sous-clusters plus petits en fonction de la dissimilarité entre les données. Cela se poursuit jusqu'à ce que chaque point de données soit dans son propre cluster.

Modèles machine learning non supervisé

- Clustering hiérarchique

- Les métriques de similarité ou de dissimilarité telles que la distance euclidienne, la distance de Manhattan, la corrélation sont utilisées pour mesurer la proximité entre les points de données.

- Le clustering hiérarchique est souvent utilisé dans la visualisation des données pour explorer les structures intrinsèques et les relations entre les données à différentes échelles.

- Il peut également être utilisé pour générer un nombre spécifique de clusters en coupant le dendrogramme à une hauteur appropriée.

Modèles machine learning non supervisé

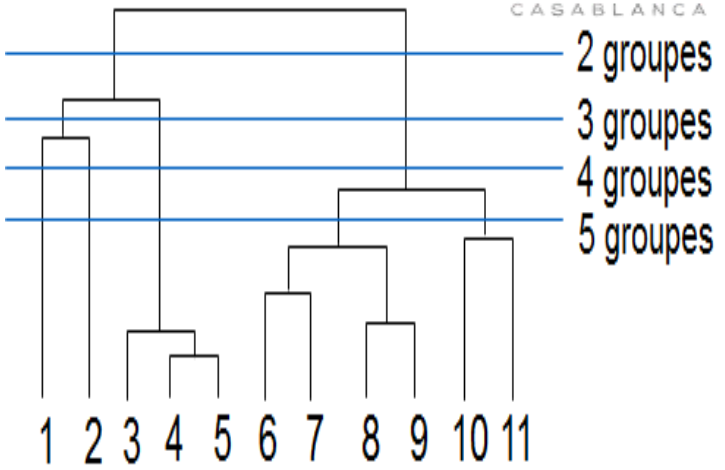
- Clustering hiérarchique

Entrée:
- Données d'entrée (points à clusteriser)
- Métrique de similarité ou de dissimilarité (par exemple, distance euclidienne)

Étape 1: Initialisation des clusters
Pour chaque point dans les données :
    Créer un cluster contenant uniquement ce point
Étape 2: Calcul des similarités entre clusters
Tant qu'il reste plus d'un cluster :
    Calculer la similarité (ou la dissimilarité) entre tous les paires de clusters restants
Étape 3: Fusion des clusters
    Fusionner les deux clusters les plus similaires en un seul cluster
    Mettre à jour la matrice de similarité pour refléter la fusion
Étape 4: Répéter les étapes 2 et 3 jusqu'à ce qu'un seul cluster global soit obtenu
Sortie:
- Structure de clustering hiérarchique (dendrogramme ou arbre de clusters)

Modèles machine learning non supervisé

- Clustering hiérarchique



Méthodes pour choisir le seuil de découpage

1. Méthode visuelle (dendrogramme)

- Observer le
- la plus grande "marche" verticale entre deux fusions successives
- Couper juste en dessous de cette marche
- → Donne les groupes les plus bien séparés

2. Seuil de distance

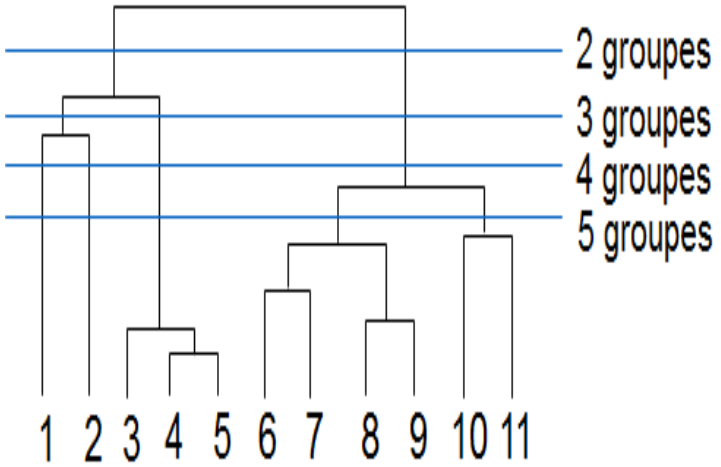
- Fixer une valeur maximale de distance autorisée pour fusionner deux clusters
- Toutes les fusions au-dessus de cette distance sont interdites
- → Tous les clusters restants sont conservés

Modèles machine learning non supervisé

- Clustering hiérarchique

Méthodes pour choisir le seuil de découpage

3. Nombre de clusters fixé
- Définir à l'avance un nombre de clusters k
- Le dendrogramme est coupé à la hauteur qui donne exactement k



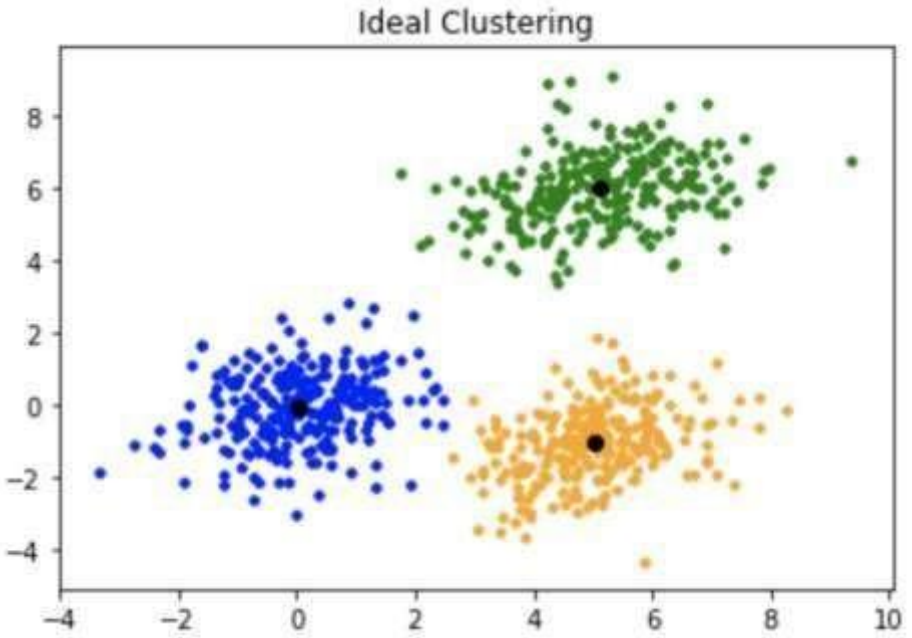
4. Indices de qualité des clusters
- Calculer des scores de validation interne pour différents seuils :Silhouette
- Choisir le seuil qui maximise la qualité du clustering

Modèles machine learning non supervisé

- Clustering Kmeans

- K-Means est un algorithme de clustering largement utilisé en apprentissage automatique non supervisé.

- Son objectif est de regrouper un ensemble de données en un certain nombre de groupes (clusters) de sorte que les points au sein d'un même cluster soient similaires les uns aux autres, tandis que les points dans des clusters différents sont distincts



Modèles machine learning non supervisé

- Clustering Kmeans

Entrée:
- Nombre de clusters k
- Données d'entrée à clusteriser

Étape 1: Initialisation des centroïdes
Choisir aléatoirement k points comme centres de clusters initiaux

Étape 2: Assignation des points aux clusters
Tant que les critères d'arrêt ne sont pas satisfaits :
    Pour chaque point dans les données :
        Calculer la distance entre le point et chaque centroïde
        Assigner le point au cluster dont le centroïde est le plus proche

Étape 3: Mise à jour des centroïdes
Pour chaque cluster :
    Calculer le nouveau centroïde comme la moyenne des points attribués à ce cluster

Modèles machine learning non supervisé

- Clustering Kmeans

Étape 4: Répéter les étapes 2 et 3 jusqu'à convergence (aucun changement dans les affectations de cluster ou un nombre maximal d'itérations est atteint)

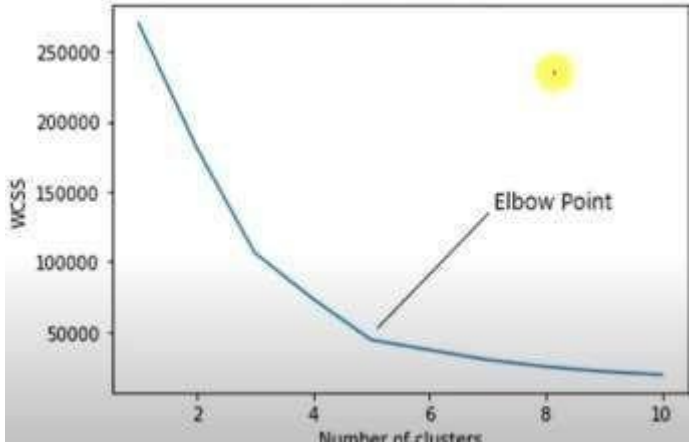
Sortie:
- Affectations de cluster pour chaque point
- Centroïdes finaux pour chaque cluster

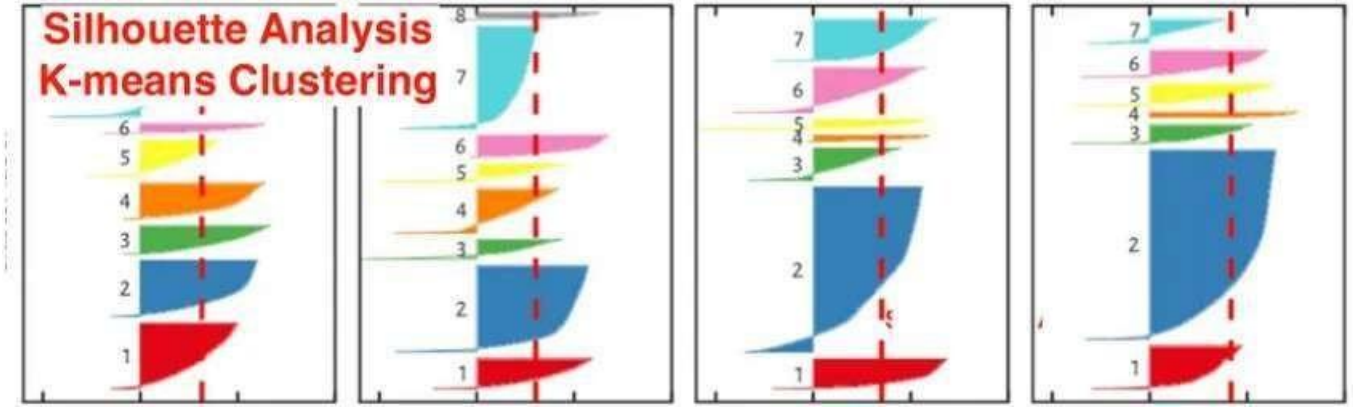
Modèles machine learning non supervisé

- Clustering Kmeans et variantes

  - Méthode du coude (Elbow Method) : Tracez la variance intra-cluster en fonction du nombre de clusters (k). Identifier le point où la courbe forme un "coude" ou une courbure significative. Ce point est souvent considéré comme le nombre optimal de clusters.

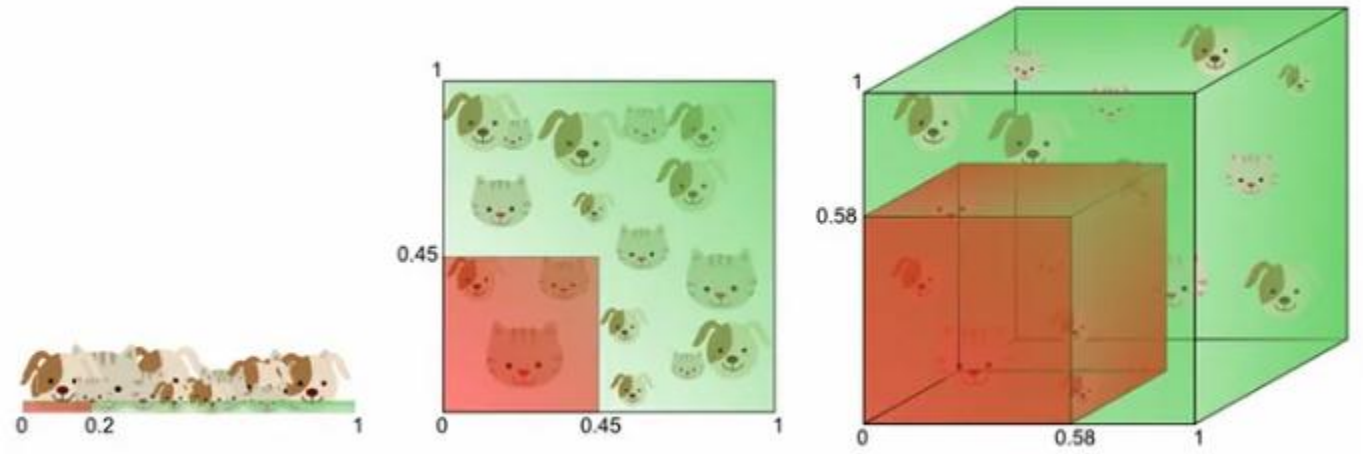
  - Méthode de la silhouette (Silhouette Method) : Calculez le score de silhouette pour différents nombres de clusters (k). Le score de silhouette mesure à quel point chaque point de données est proche de son propre cluster par rapport aux autres clusters. Choisissez le nombre de clusters qui maximise le score de silhouette.





La malédiction de la dimensionnalité
Commençons par un problème qui affecte tous ceux qui analysent les données : la malédiction de **la dimensionnalité.**

Ce terme fait référence aux défis informatiques, analytiques, de clustering et de visualisation qui se posent lorsqu'on traite des **données à grande dimensions**(c.-à-d. quand vous avez beaucoup de caractéristiques/variables).



curse of dimensionality

La malédiction de la dimensionnalité

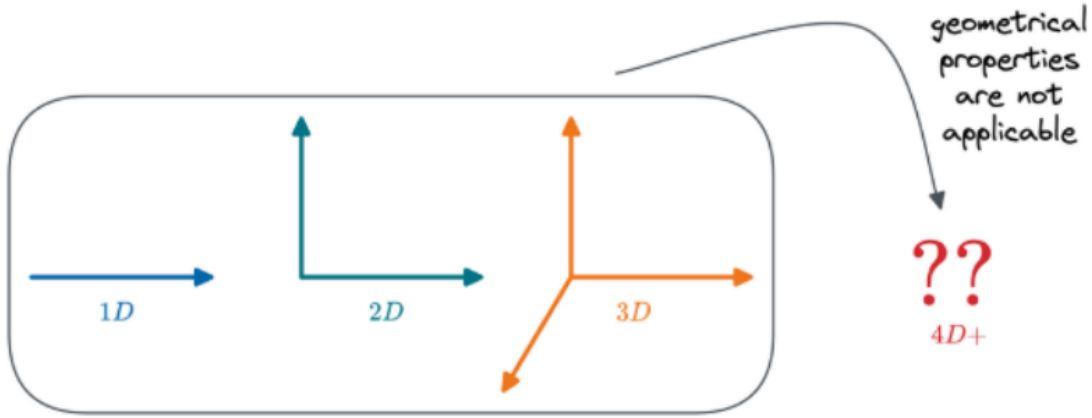
Qu’entendons-nous par "espace" ?

Dans ce contexte, "espace" fait référence au plan ou à l’environnement multidimensionnel où les points de données sont représentés graphiquement.

Par exemple :

- Si nous avons deux variables, nous pouvons visualiser les données sur un plan 2D.
- Si nous avons trois variables, nous pouvons imaginer un espace 3D.

Au-delà des trois dimensions, nous ne pouvons pas l
visualiser intuitivement, mais mathématiquement, il
mêmes principes.



La malédiction de la dimensionnalité

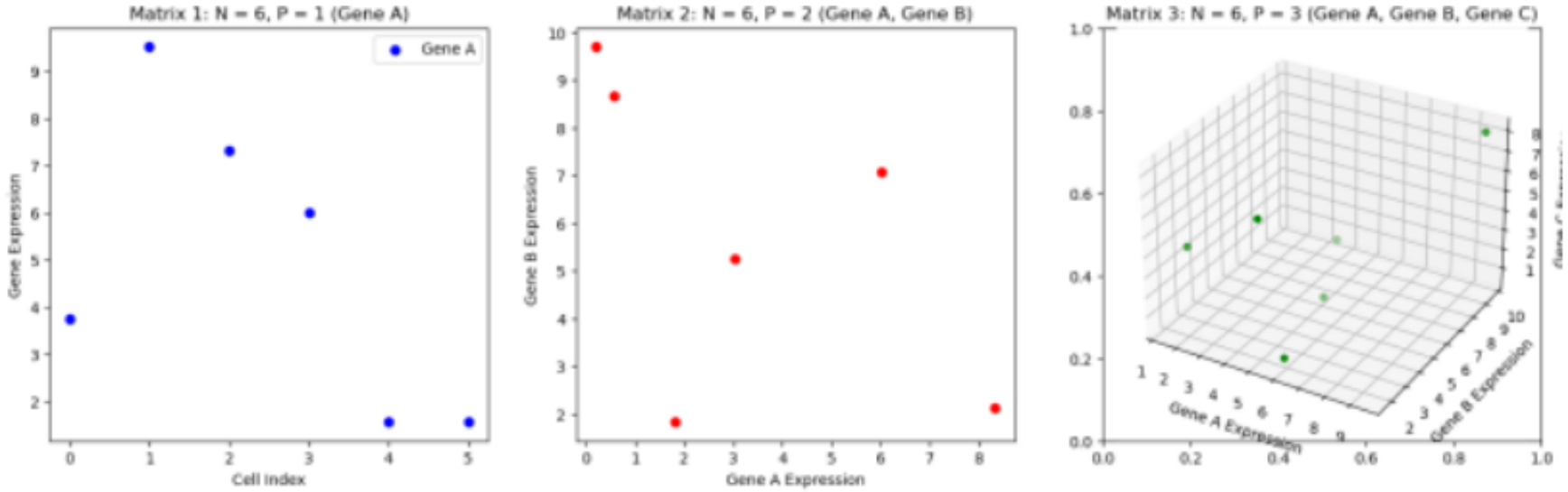
Qu'est-ce qu'une dimension?

- Pour répondre à cette question, expliquons d'abord comment les données sont structurées :
- Les données sont généralement stockées dans une **matrice NxP**, où :
  - **N =** le nombre d'observations (p. ex., cellules, individus, échantillons
  - **P =** le nombre de variables (p. ex., niveaux d'expression des gènes, accessibilité maximale, etc.).
- **Chaque variable (P)** représente une dimension, c'est-à-dire un axe dans l'espace le long duquel les points de données peuvent varier.

- Plus de dimensions signifient plus de complexité.
- Si nous ajoutons plus de variables, les données "se répandent" dans des directions supplémentaires.
- Cela rend l'analyse, le regroupement et la visualisation plus difficiles.

La malédiction de la dimensionnalité

Qu’est-ce qu’une dimension?



Oups! Si nous avons plus de trois gènes, nous ne pouvons plus visualiser directement les données. Notre cerveau n’est pas conçu pour percevoir des dimensions au-delà de trois, et l’utilisation d’un grand nombre de dimensions entraîne des défis mathématiques et informatiques.

La malédiction de la dimensionnalité

Voici ce qui se passe lorsque les dimensions augmentent :

1. La distance devient moins significative dans les espaces à haute dimension
Tous les points ont tendance à être presque également éloignés. Cela nuit aux algorithmes tels que k-NN, clustering ou détection des valeurs aberrantes basée sur la distance.

2. Risques de surdimensionnement
Les modèles peuvent facilement mémoriser le bruit dans des données à haute dimension.
Réflexion : plus de caractéristiques que d'échantillons = le modèle correspond trop **bien aux données d'entrainement, mais** ne fonctionne pas sur les nouvelles données.

3. Coût de calcul accru
Le stockage, le calcul et le temps augmentent tous avec plus de dimensions.

Solution! Réduction dimensionnelle

- **La malédiction de la dimensionnalité nous dit que le fait d'avoir P>>N pose des défis pour la visualisation, l'analyse et les opérations mathématiques.**
-
- Pour rendre ces tâches réalisables, nous devons réduire P**, en amenant nos données à un état plus gérable où P<=N, d'où le concept** de "réduction de dimension"**.**

- N = le nombre d'observations (p. ex., cellules, individus, échantillons
- P = le nombre de variables (p. ex., niveaux d'expression des gènes, accessibilité maximale, etc.).

Solution! Réduction dimensionnelle

Types d’approches de réduction dimensionnelle

- Méthodes de sélection des caractéristiques

- les caractéristiques moins importantes — celles qui expliquent une moindre variabilité des données sont supprimées.
- ne conserver que les caractéristiques les plus pertinentes, sélectionnées sur la base de critères statistiques ou de modèles d’apprentissage automatique.

- Méthodes d’extraction des caractéristiques

Au lieu de supprimer des variables, cette approche crée de nouvelles caractéristiques qui condensent les informations à partir de celles d’origine. Considérez cela comme une façon de résumer plusieurs variables en moins de dimensions tout en conservant autant d’informations que possible.

Solution! Réduction dimensionnelle

Types d'approches de réduction dimensionnelle

Méthodes d'extraction des caractéristiques

- Au lieu de supprimer des variables, cette approche crée de nouvelles caractéristiques qui condensent les informations à partir de celles d'origine.
- Considérez cela comme une façon de résumer plusieurs variables en moins de dimensions tout en conservant autant d'informations que possible.
- Les nouvelles caractéristiques sont créées sous forme **de combinaisons linéaires des variables originales** --→ analyse des composantes principales (ACP).
- Cette méthode vise **à maximiser la préservation** de la variance.

La variance mesure l'étendue d'un ensemble de nombres.
Elle indique dans quelle mesure les valeurs diffèrent de la moyenne (moyenne) de l'ensemble de données.

Analyse en composantes principales

- PCA est une méthode de réduction dimensionnelle qui utilise une approche d'extraction linéaire.
- Il crée de nouvelles variables (composantes principales, ou PC) ou des facteurs latents sous forme de combinaisons linéaires des variables d'origine.

Les principales étapes de l'APC sont :

1) **Mettre à l'échelle les données** – Normalisation des variables pour qu'elles aient une moyenne de 0 et une variance de 1.
2) **Calculer de la matrice de covariance** – Mesure du lien entre les variables.
3) **Trouver les valeurs propres et les vecteurs propres – Extraire les composantes principales qui expliquent la plus grande variance.**
4) **Trier et sélectionner les composantes principales** – Ne conserver que les plus informatifs.
5) **Projecter les données sur les composantes sélectionnées** – Transformation des données d'origine dans le nouvel espace réduit.

Analyse en composantes principales

1. Mise à l'échelle des données

- ACP est sensible à l'échelle des variables parce qu'il est basé sur la matrice de covariance.
- Si les variables ont des fourchettes très différentes, celles qui ont des valeurs plus élevées domineront la variance et fausseront les résultats.
- Pour éviter ce problème, il est essentiel de standardiser les variables, en veillant à ce que chacune ait une moyenne de 0 et un écart type de 1 **(normalisation du Z-score)**
- Cela rend les variables comparables, empêchant les variables à plus grande échelle d'influencer l'analyse de manière disproportionnée.

2. Calcul de la matrice de covariance

- La covariance mesure la façon dont deux variables varient ensemble.

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- Corrélation est une **covariance stadardisée**

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Analyse en composantes principales

3. Trouver des valeurs propres et des vecteurs propres

- Une **valeur propre indique combien une transformation étire ou compresse un vecteur le long d'une direction particulière.**
- Elle est **associée à un** vecteur propre, qui est une direction spéciale qui ne change pas de direction sous une transformation linéaire - il n'est que redimensionné.

Given a square matrix A, an eigenvector $\vec{v}$ and its eigenvalue λ satisfy:

$$A\vec{v} = \lambda\vec{v}$$

This means: multiplying $\vec{v}$ by the matrix A just **scales** it — it doesn't rotate it.

Analyse en composantes principales

3. Trouver des valeurs propres et des vecteurs propres

- Comment définir les valeurs propres ?

  Nous résolvons cette **équation caractéristique pour** trouver les valeurs propres λ.

For a matrix A: $$\det(A - \lambda I) = 0$$

Pourquoi elles sont importantes dans ACP (Analyse en composantes principales) :

- Les valeurs propres de la matrice de covariance **nous indiquent combien** de variance est **capturée le long de chaque composante principale.**
- Les valeurs propres plus grandes ont une plus grande variance le long de cette direction.
- Dans ce cas, la matrice A correspond à la matrice de covariance.

Analyse en composantes principales

3. Trouver des valeurs propres et des vecteurs propres

Utiliser les vecteurs propres comme poids de transformation

- Chaque vecteur propre fournit les coefficients (poids) **pour former les** composantes principales**.**

4. Tri et sélection des principaux composants

- Dans ACP, nous voulons transformer nos données en un nouveau système de coordonnées où :
  - Le premier axe (composante principale) capture la plus grande variance.
  - Le deuxième axe capture la deuxième plus grande variance.
  - Et ainsi de suite…

Pour ce faire, nous trouvons les vecteurs propres de la matrice de covariance (A), qui définissent les nouveaux axes de coordonnées.
Ces **vecteurs propres fournissent les poids** pour la combinaison linéaire des variables originales.

4. Tri et sélection des principaux composants

- Chaque **composante principale** a une **direction dans l'espace de caractéristiques** (eigenvector)
- L'espace de caractéristiques a une dimension de N*P
- La matrice de covariance a une dimension de P*P.
- Chaque entrée (i,j) dans la matrice de covariance représente la covariance entre **la caractéristique I et la caractéristique j.**
- L'ACP trouve p facteurs latents ou p composantes de principales.

Analyse en composantes principales

4. Tri et sélection des principaux composants

Suppose we have a dataset with two variables X_1 and X_2, and the **covariance matrix** is:

$$A = \begin{bmatrix} 2.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

By solving $Av = \lambda v$, we find:

$$\text{Eigenvectors: } v_1 = \begin{bmatrix} 0.924 \\ 0.383 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -0.383 \\ 0.924 \end{bmatrix}$$

Analyse en composantes principales

4. Tri et sélection des principaux composants

So, the **principal components** are:

$$PC_1 = 0.924X_1 + 0.383X_2$$

$$PC_2 = -0.383X_1 + 0.924X_2$$

This means:

- **PC1** is mostly influenced by X_1 (since the weight is 0.924), but also a little by X_2 (weight 0.383).
- **PC2** is a different linear combination, mostly influenced by X_2.

✓ These weights (0.924, 0.383, etc.) are entirely determined by the covariance structure of the data.
✓ We don't choose them manually—they come from solving the eigenvalue problem!

Analyse en composantes principales

4. Tri et sélection des principaux composants

So, the **principal components** are:

$$PC_1 = 0.924X_1 + 0.383X_2$$

$$PC_2 = -0.383X_1 + 0.924X_2$$

This means:

- **PC1** is mostly influenced by X_1 (since the weight is 0.924), but also a little by X_2 (weight 0.383).
- **PC2** is a different linear combination, mostly influenced by X_2.

✓ These weights (0.924, 0.383, etc.) are entirely determined by the covariance structure of the data.
✓ We don't choose them manually—they come from solving the eigenvalue problem!

Centrale
CASABLANCA

Analyse en composantes principales

4. Tri et sélection des principaux composants

- Les pondérations attribuées à chaque variable initiale varient selon les composantes principales.
- Ils indiquent dans quelle mesure chaque variable contribue à expliquer la variance de l'ensemble de données.

- Si w 0, la variable initiale contribue fortement à expliquer la variance dans cette composante principale.
- Si w=0, la variable ne contribue pas à cette composante.
- Si w 0, la variable a une forte contribution mais dans le sens opposé, ce qui signifie qu'elle est négativement corrélée avec cette composante.

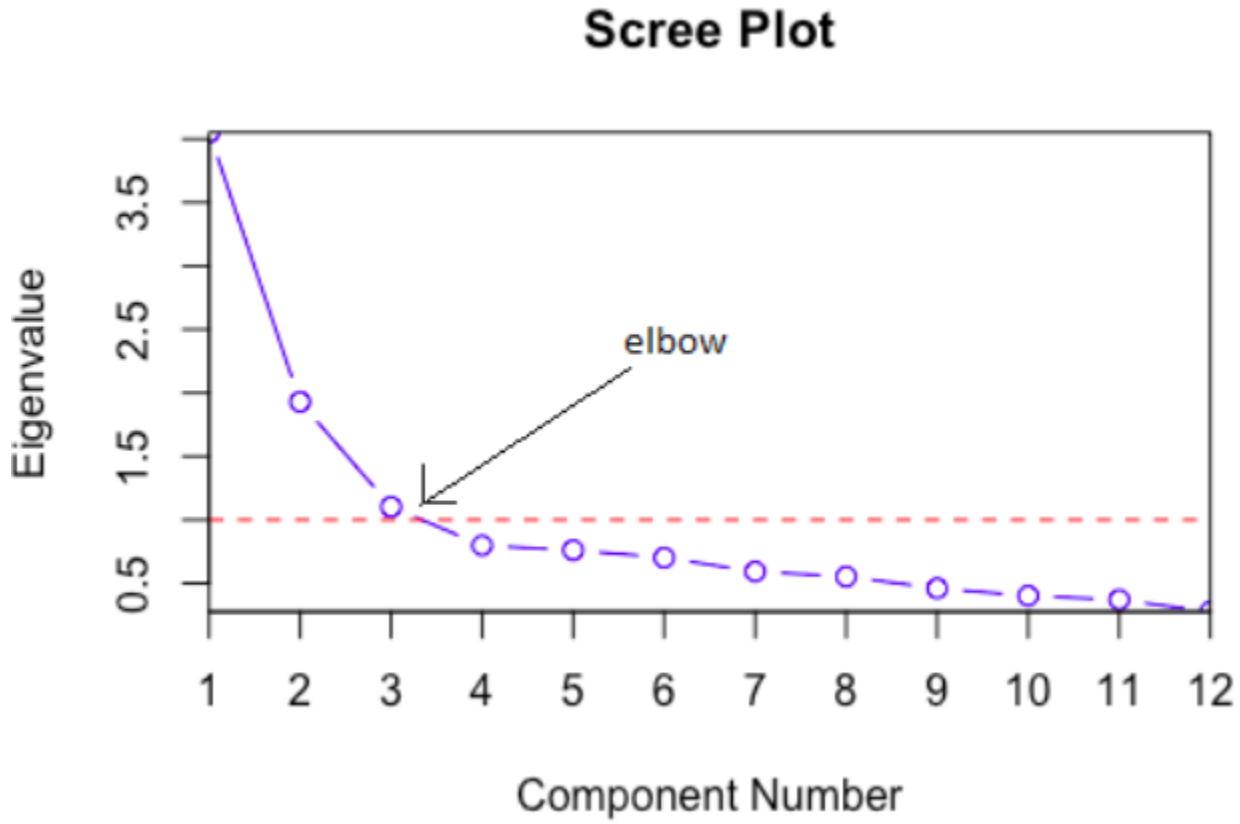
Analyse en composantes principales

4. Tri et sélection des principaux composants

- nous obtenons P composantes principales (CP), chacune représentant une combinaison linéaire des variables originales et expliquant une certaine variance dans les données.
- Puisque notre objectif est de réduire le nombre de variables et de limiter la dimensionnalité, nous rejetons les CP qui expliquent très peu de variance, en conservant seulement les plus informatives.
- Pour ce faire, nous classons les CP par ordre décroissant en fonction de leurs valeurs propres associées, qui représentent la variance expliquée par chaque composante.
- Ensuite, nous sélectionnons les composantes les plus importantes, en **choisissant souvent les premieres CP qui expliquent ensemble une proportion importante de la variance totale (généralement 70-95%).**

4. Tri et sélection des principaux composants

- Le nombre optimal de composants k peut être déterminé à l'aide d'un graphique en écailles.
- Le graphique en écailles affiche les valeurs propres et aide à identifier une coupure où l'ajout de plus de composants donne des rendements décroissants.



Centrale
CASABLANCA

Analyse des composants principaux

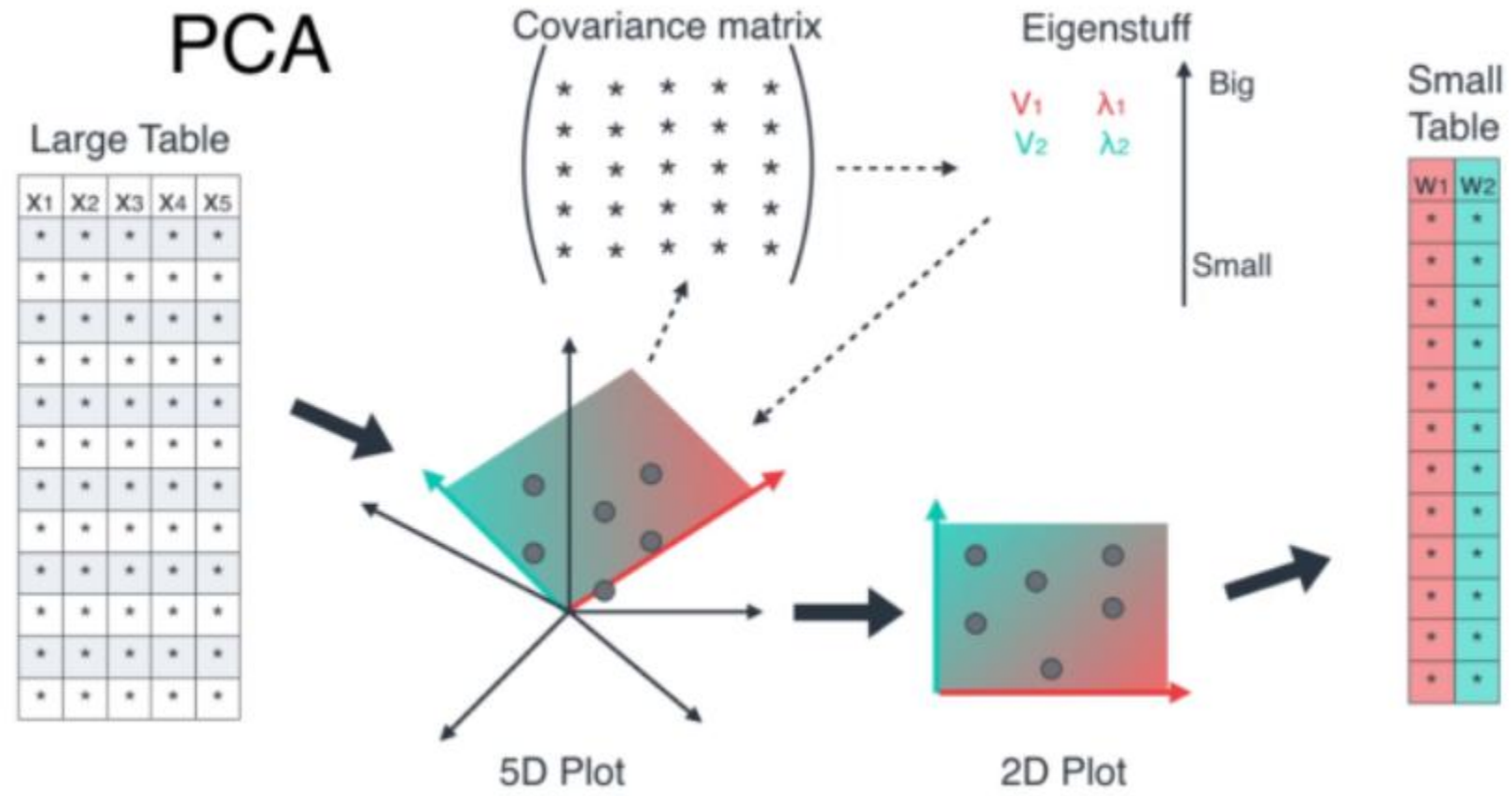
5. Projection des données sur le k Principaux composants

- Nous sommes maintenant arrivés à l'étape finale de la PCA : transformer les données dans le nouvel espace à dimensions réduites.
- Maintenant, au lieu d'avoir un axe pour chaque variable originale, le nouvel espace a un axe pour chacun des k composants principaux sélectionnés (PC).

Analyse en composantes principales

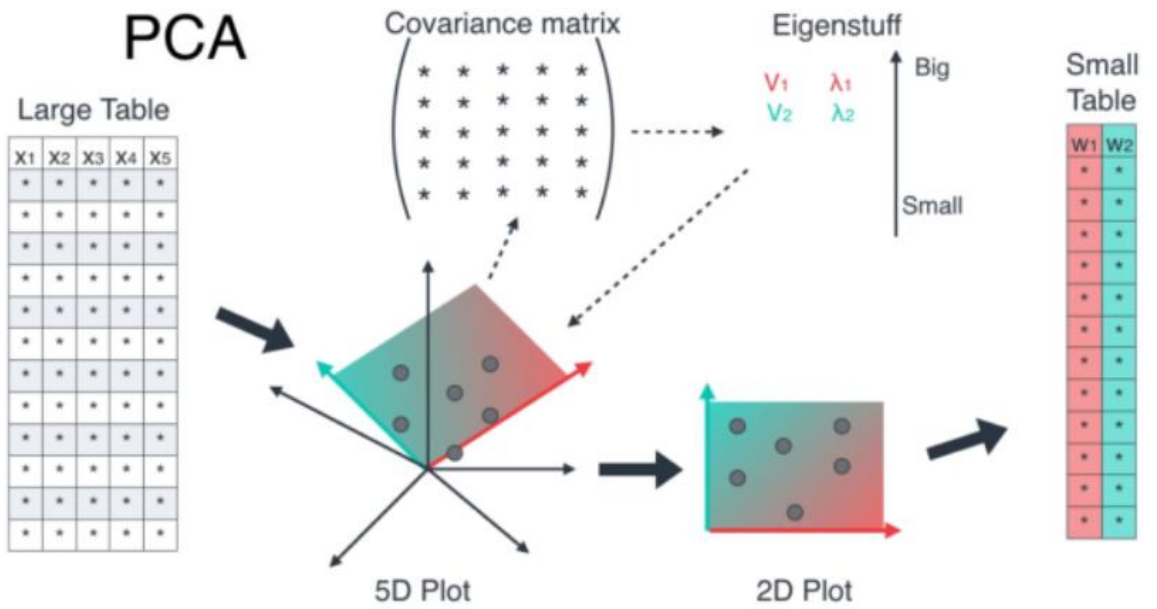
5. Projection des données sur les k composantes proncipales

- La variance expliquée par chaque CP est déterminée par sa valeur propre.
- La direction de chaque CP est déterminée par son propre vecteur correspondant.
- La variance expliquée par chaque CP est déterminée par sa valeur propre.



Analyse en composantes principales

5. Projection des données sur les k composantes principales



Assume:

- X: Your original data matrix, shape $n \times p$ (n samples, p features)
- W_k: Matrix of the top k principal component **eigenvectors**, shape $p \times k$

To project onto the top k principal components:

$$Z = X_{\text{scaled}} \cdot W_k$$

This gives:

- Z: shape $n \times k$
- Each row is your sample represented in the **reduced k-dimensional space**

Centrale
CASABLANCA

Analyse en composantes principales

5. Projection des données sur les k composantes principales

Formule pour la variance expliquée

La fraction de variance expliquée par une composante donnée k est :

$$\text{Explained Variance Ratio for PC}_k = \frac{\lambda_k}{\sum \lambda}$$

where:

- λ_k is the eigenvalue for the k-th principal component.
- $\sum \lambda$ is the sum of all eigenvalues (total variance in the dataset).

Centrale
CASABLANCA

Analyse en composantes principales

5. Projection des données sur le k Principaux composants

Formule pour l'écart cumulatif expliqué :
La variance expliquée cumulée nous indique combien de variance totale est conservée par les premières k composantes :

$$\text{Cumulative Explained Variance} = \sum_{i=1}^{k} \frac{\lambda_i}{\sum \lambda}$$

Analyse en composantes principales

5. Projection des données sur le k Principaux composants

Let's consider a dataset where we perform PCA and obtain **three eigenvalues**:

$$\lambda_1 = 4.5, \quad \lambda_2 = 2.0, \quad \lambda_3 = 0.5$$

Step 1: Compute the Total Variance

$$\text{Total Variance} = \lambda_1 + \lambda_2 + \lambda_3 = 4.5 + 2.0 + 0.5 = 7.0$$

Centrale
CASABLANCA

Analyse en composantes principales

5. Projection des données sur le k Principaux composants

Step 2: Compute the Variance Explained by Each PC

$$\text{Explained Variance Ratio for } PC_1 = \frac{4.5}{7.0} = 0.643(64.3\%)$$

$$\text{Explained Variance Ratio for } PC_2 = \frac{2.0}{7.0} = 0.286(28.6\%)$$

$$\text{Explained Variance Ratio for } PC_3 = \frac{0.5}{7.0} = 0.071(7.1\%)$$

Step 3: Compute the Cumulative Explained Variance

$$PC1 + PC2 = 64.3\% + 28.6\% = 92.9\%$$

$$PC1 + PC2 + PC3 = 100\%$$

Centrale
CASABLANCA

Analyse en composantes principales

5. Projection des données sur le k Principaux composants

So, if we **keep only the first two components**, we **retain 92.9% of the variance** in the data, making it a good choice for dimensionality reduction.

Analyse en composantes principales

Les principaux éléments; nouveaux facteurs latents mais quelle signification?
Comment interpréter ?

Diagramme du cercle de corrélation?

- Après ACP, les variables originales peuvent être projetées sur l'espace des composantes principales.
- Cette projection nous donne un cercle de corrélation où :

1. Flèches = variables originales
2. Direction = corrélation avec les CP
3. Longueur = force de la corrélation (plus longue = plus forte)

Flèches plus longues La variable a une **forte influence** sur cette composante principale.
Flèches plus courtes La variable a une **contribution plus faible** à ce CP.
Angles entre les flèches La corrélation entre les **variables :**
4. Angle entre les flèches = relation des variables
Petit angle positivement corrélé
La direction opposée est corrélée négativement
Angle droit (90°) non corrélé

Analyse en composantes principales

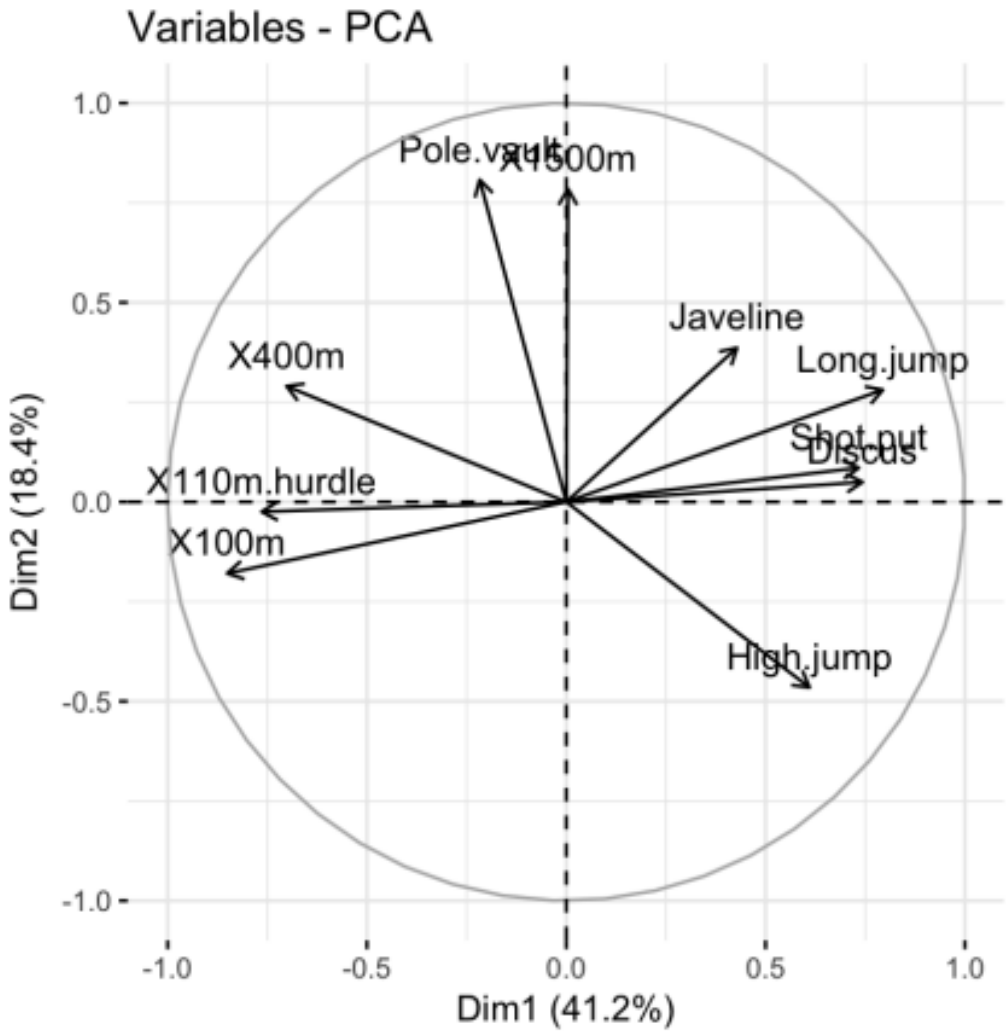
- Un cadre de données avec 41 lignes et 13 colonnes :
- les dix premières colonnes correspondent à la performance des athlètes pour les 10 épreuves du décathlon.
- Les colonnes 11 et 12 correspondent respectivement au rang et aux points obtenus.
- La dernière colonne est une variable catégorielle correspondant à l'événement sportif (Jeu Olympique 2004 ou Decastar 2004)

https://malouche.github.io/data_in_class/decathlon_data.html

Analyse en composantes principales

- La variable **"X100m"** est **négativement** corrélée avec la variable **" long.jump"**. Lorsqu'un athlète court une **distance de 100 m plus rapide** (c.-à-d. une **valeur inférieure**), il a tendance à **sauter plus loin**.

- Il est important de noter ici que **les valeurs inférieures pour** les variables **"X100m"** , **"X400m"**, **"X110m.hurdle"** et **"X1500m"** correspondent en fait à un **score plus élevé : plus l'athlète court vite, plus il gagne de points.**

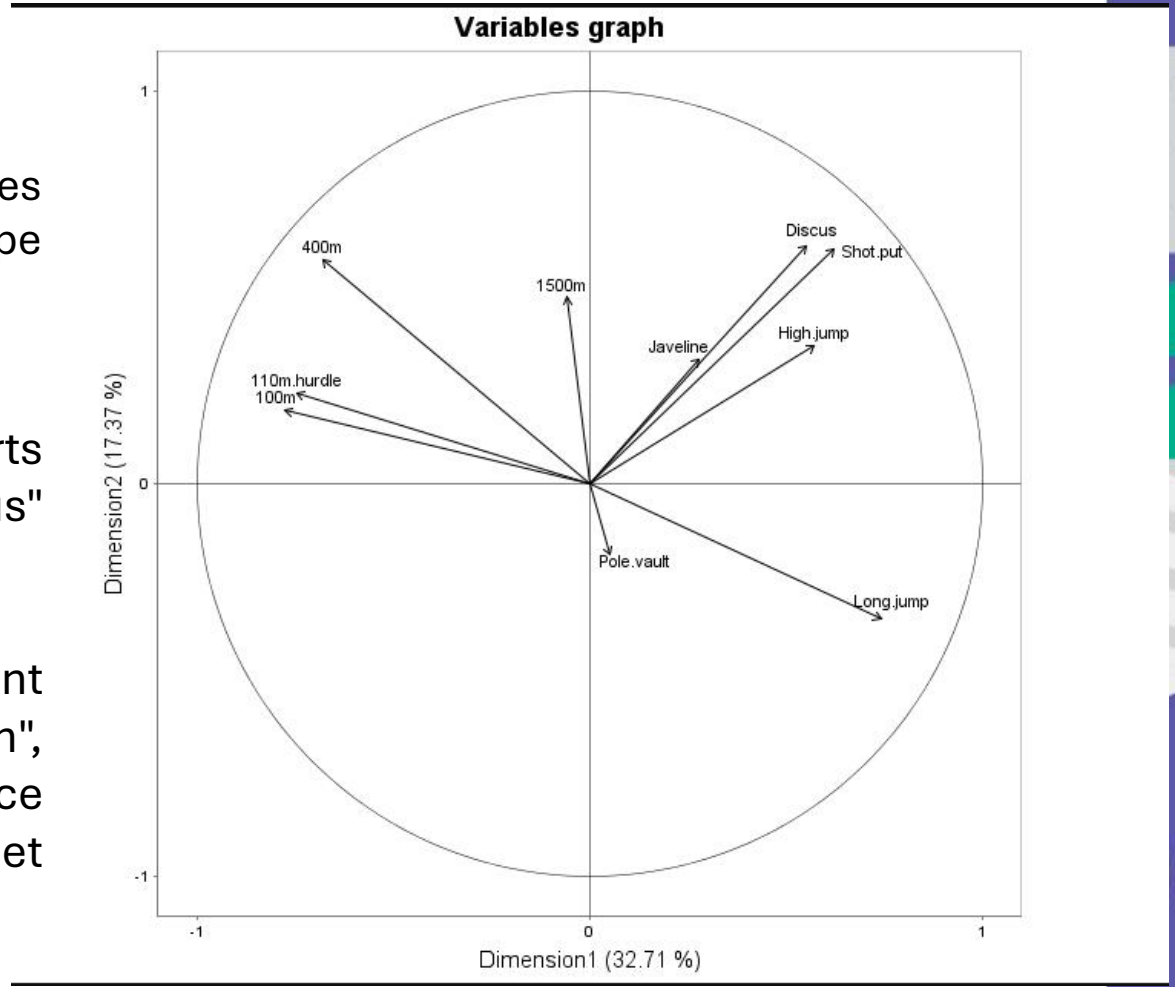


Analyse en composantes principales

- La 1ere dimension est fortement associée aux variables vitesse et saut en longueur, qui forment un groupe homogène.

- Le deuxième axe (CP2) oppose les athlètes qui sont forts dans les épreuves de lancer, en particulier dans "Discus" et "Shot.put", contre ceux qui ne le sont pas.

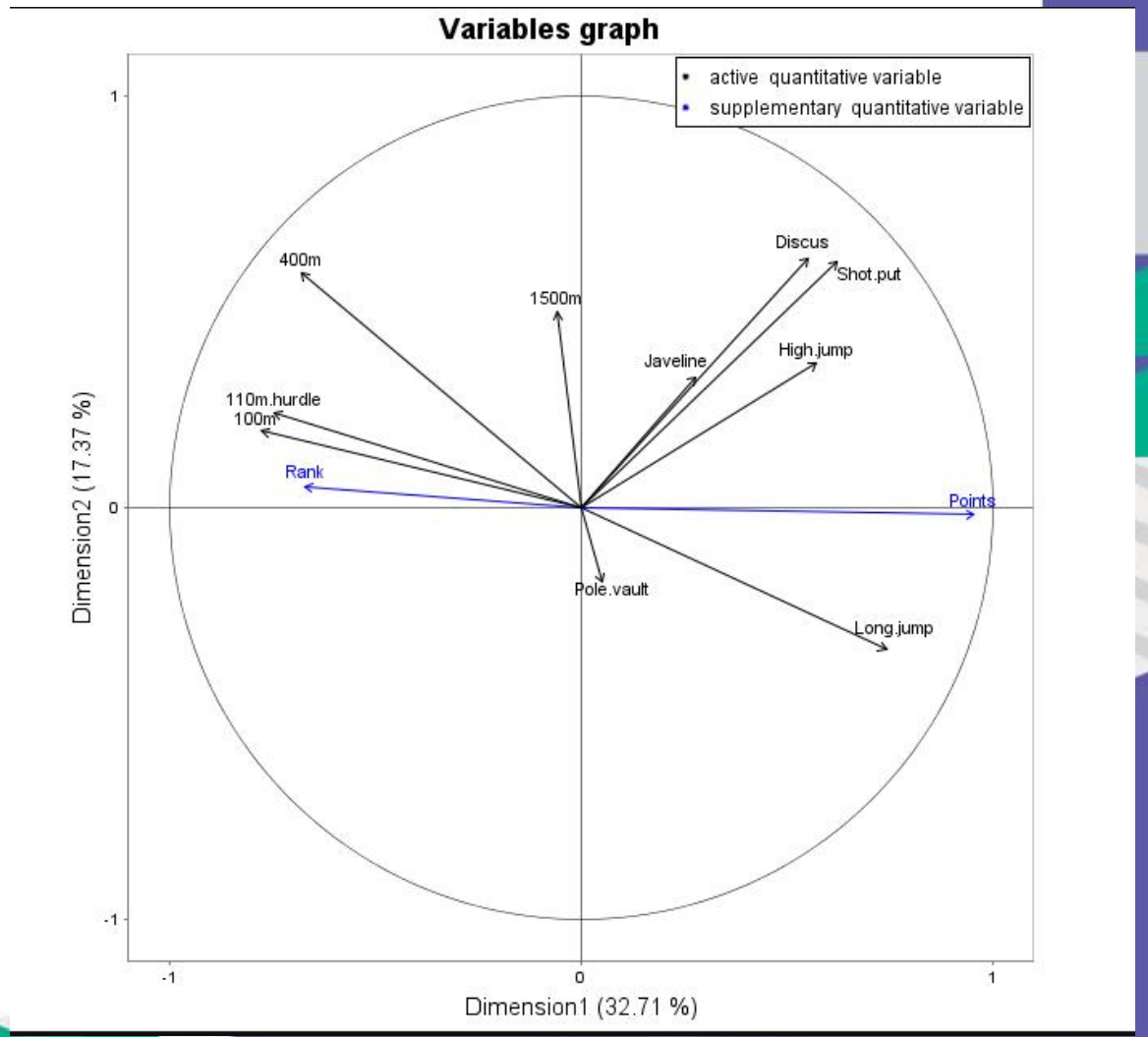
- Les variables "Discus", " Shot.put" et "High.jump" ne sont pas fortement corrélées avec "X100m", "X400m", "X110m.hurdle", et "long jump. Cela indique que la force et la vitesse ne sont pas fortement corrélées dans cet ensemble de données.



Analyse en composantes principales

Interprétation de la performance et des points dans le décathlon

- Les vainqueurs du décathlon sont ceux qui obtiennent le plus de points (ou qui ont le rang le plus bas).
- Les variables les plus fortement liées au nombre total de points sont celles associées à la vitesse — soit « X100m », « X110m.hurdle », « X400m » — et le saut en longueur.
- En revanche, "Saut à la perche" et "X1500m" ont moins d'influence sur le score total.



Préparation des données

http://factominer.free.fr/factomethods/images/principal-correspondence-analysis-var.PNG

http://factominer.free.fr/factomethods/images/principal-correspondence-analysis-ind.PNG