

Techniques de réduction de dimension

Hicham Janati

hjanati@insea.ac.ma

Chapitre II: Analyse en composantes principales (PCA)

On observe les données de “ratings” des articles sur Amazon par les acheteurs:

Customer	Product 1	Product 2	...	Product d
Customer 1	4.5	3.0	...	5.0
Customer 2	3.5	4.0	...	2.5
Customer 3	5.0	2.5	...	4.0
⋮	⋮	⋮	⋮	⋮
Customer n	4.0	4.5	...	3.5

Rappel: Une **observation**: une ligne de la base données
Une **variable (feature)**: une colonne de la base de données

On observe n échantillons (observations) de d variables aléatoires.

Ou encore: on observe n échantillons d'un **vecteur** aléatoire de dimension d .

On observe les données de “ratings” des articles sur Amazon par les acheteurs:

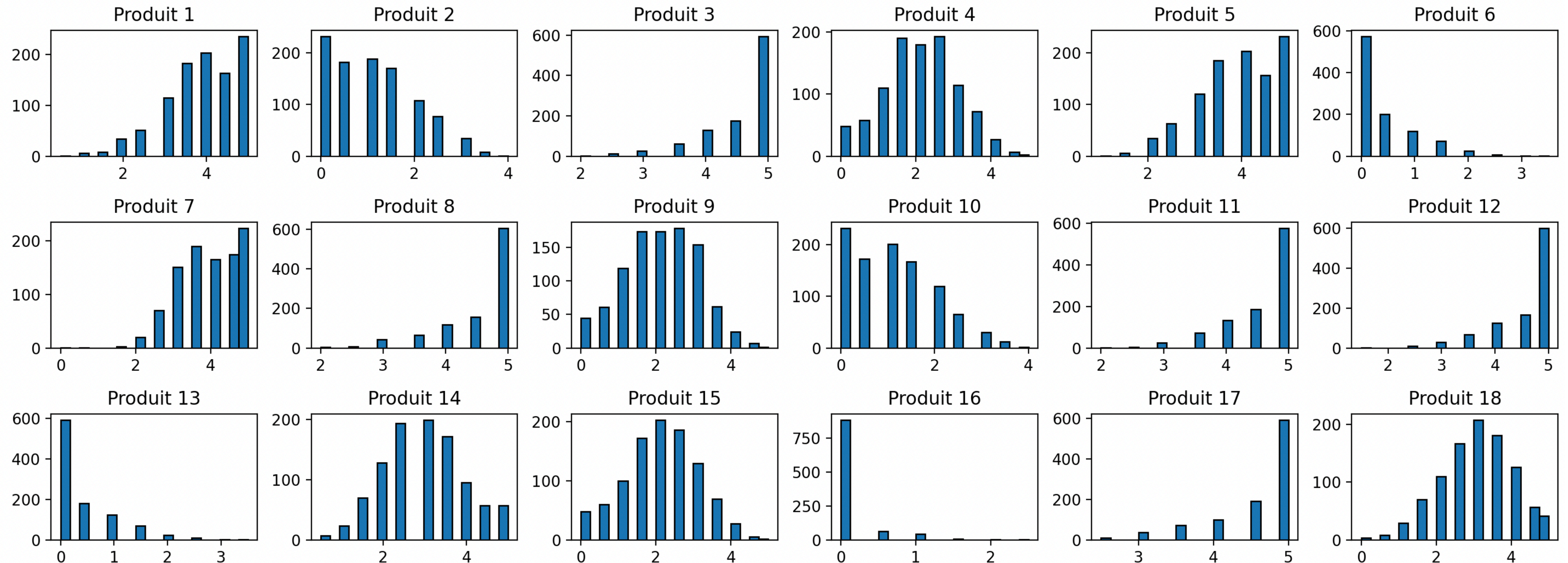
Customer	Product 1	Product 2	...	Product d
Customer 1	4.5	3.0	...	5.0
Customer 2	3.5	4.0	...	2.5
Customer 3	5.0	2.5	...	4.0
⋮	⋮	⋮	⋮	⋮
Customer n	4.0	4.5	...	3.5

Objectifs: étudier les éventuels “profils” de consommateurs

Existe-t-il des sous-groupes de consommateurs similaires qui diffèrent selon:

1. Le type de produits achetés (Mobilier, Tech, Hygiène, Habits...)
2. Leur niveau de satisfaction
3. Leur niveau de satisfaction selon le type de produits achetés

On peut commencer par visualiser les distributions univariées des variables:



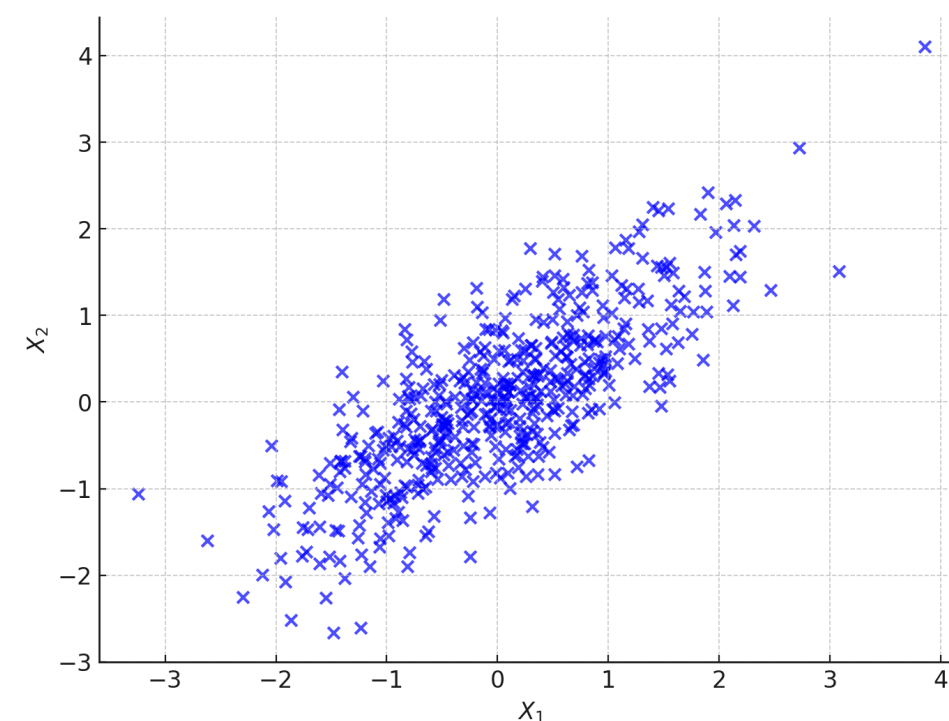
1. Ne permet pas de voir les interactions entre les variables
2. Infaisable avec un grand nombre de variables

Soit X_1 et X_2 deux variables aléatoires. On définit leur covariance par:

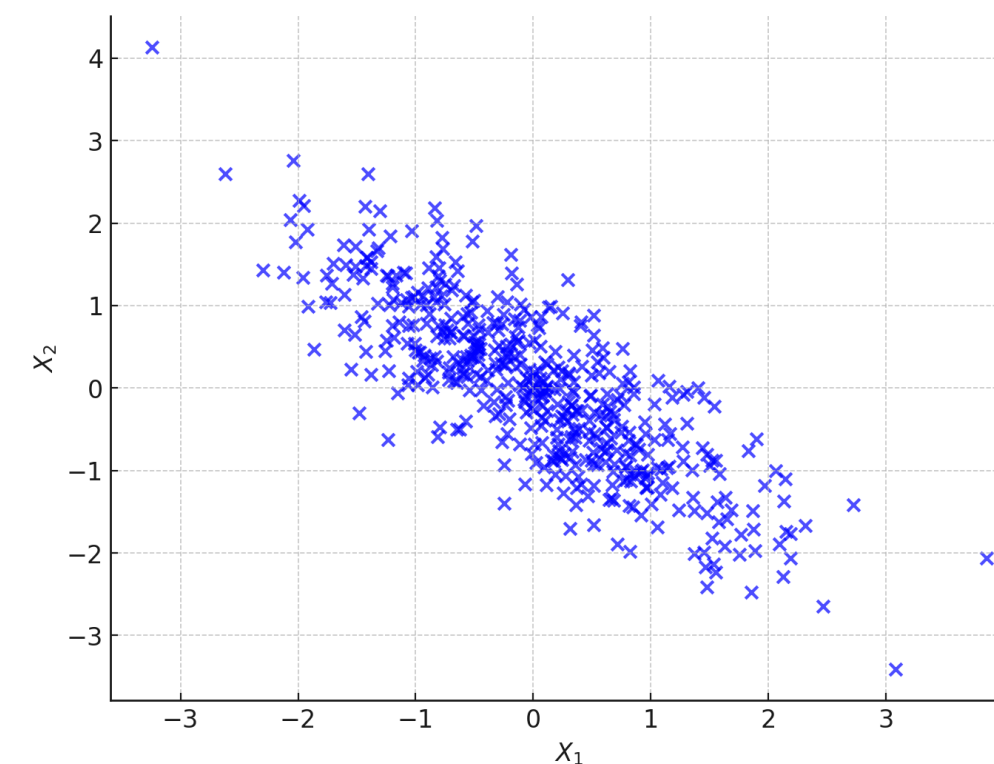
$$\begin{aligned} \text{Cov}(X_1, X_2) &\stackrel{\text{def}}{=} \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))) \\ &= \mathbb{E}(X_1^c X_2^c) \longleftarrow \text{Espérance du produit des variables centrées} \end{aligned}$$

Étudier son signe. Comment peut-on l'interpréter ?

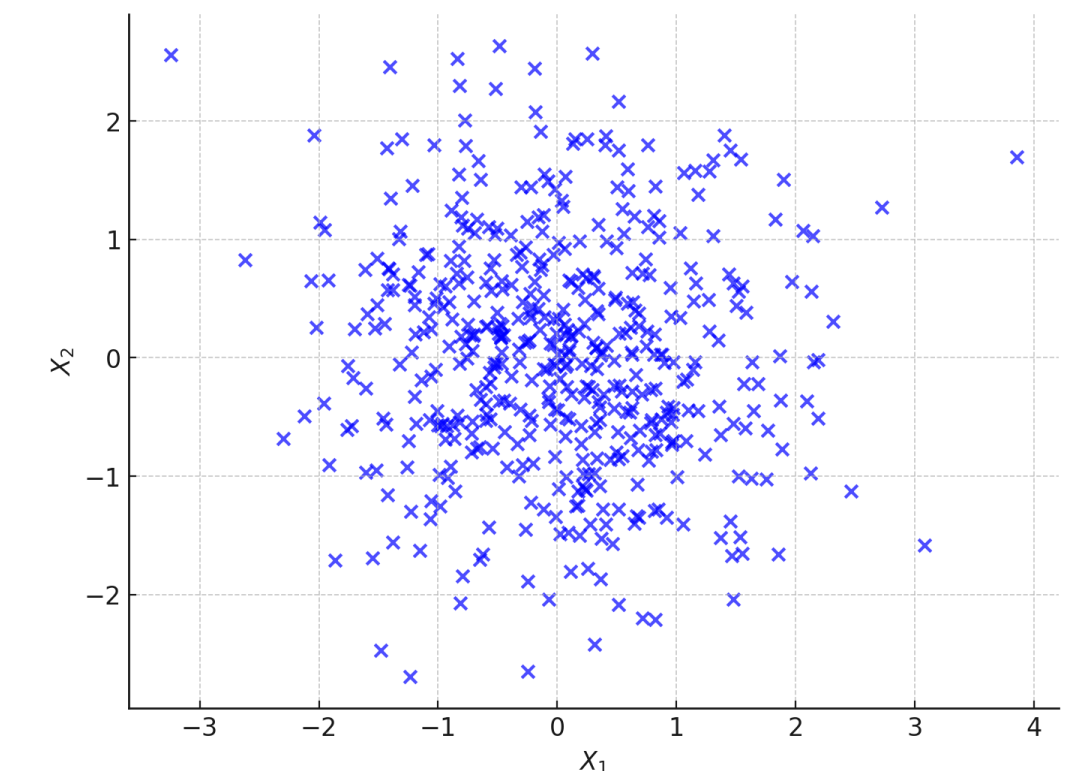
Covariance positive



Covariance négative



Covariance nulle



Qu'en est-il de son amplitude ?

Non bornée !

Soit X_1 et X_2 deux variables aléatoires. On définit leur covariance par:

$$\begin{aligned}\text{Cov}(X_1, X_2) &\stackrel{\text{def}}{=} \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))) \\ &= \mathbb{E}(X_1^c X_2^c) \longleftarrow \text{Espérance du produit des variables centrées}\end{aligned}$$

Qu'en est-il de son amplitude ? **Non bornée !**

Mais: $\text{Cov}(X_1, X_1) = \mathbb{V}(X_1)$ et on peut démontrer que: **comment ?**

$$|\text{Cov}(X_1, X_2)| \leq \sqrt{\mathbb{V}(X_1)\mathbb{V}(X_2)}$$

Cov définit un produit scalaire, ceci découle de l'inégalité de Cauchy-schwarz.

On définit le coefficient de corrélation $\rho \in [-1, 1]$:

$$\rho_{X_1 X_2} \stackrel{\text{def}}{=} \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1) \text{Var}(X_2)}}.$$

Soit X_1 et X_2 deux variables aléatoires. On définit leur covariance par:

$$\text{Cov}(X_1, X_2) \stackrel{\text{def}}{=} \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)))$$

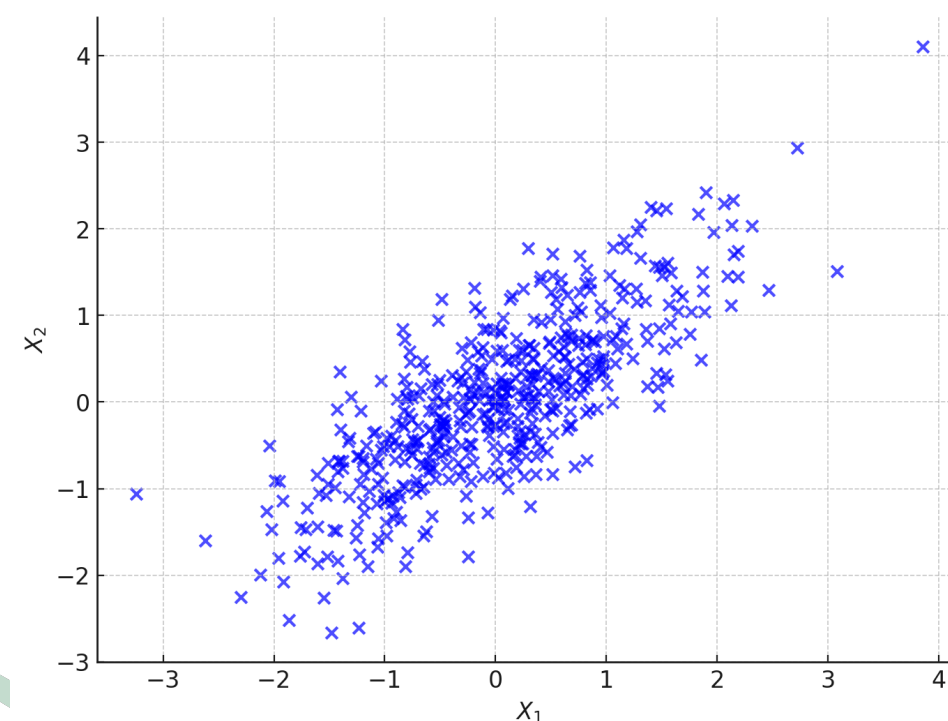
$$= \mathbb{E}(X_1^c X_2^c) \longleftarrow \text{Espérance du produit des variables centrées}$$

On définit le coefficient de corrélation $\rho \in [-1, 1]$:

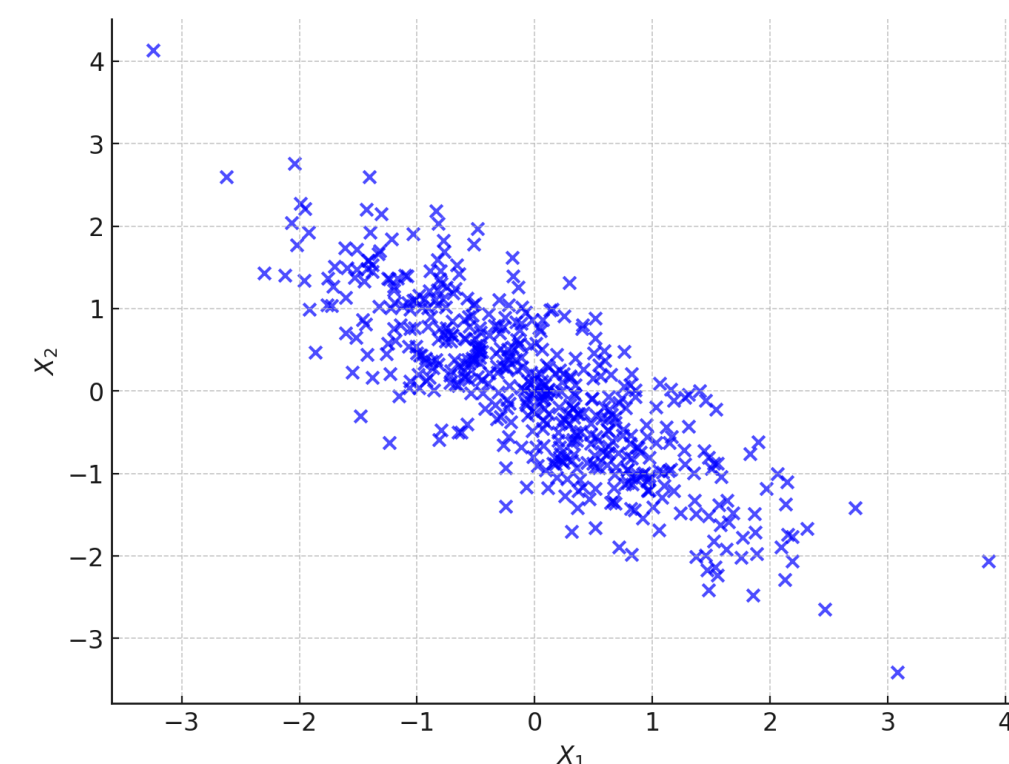
$$\rho_{X_1 X_2} \stackrel{\text{def}}{=} \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1) \text{Var}(X_2)}}.$$

Pouvez-vous imaginer d'autres nuages 2D avec une corrélation ~ 0 ?

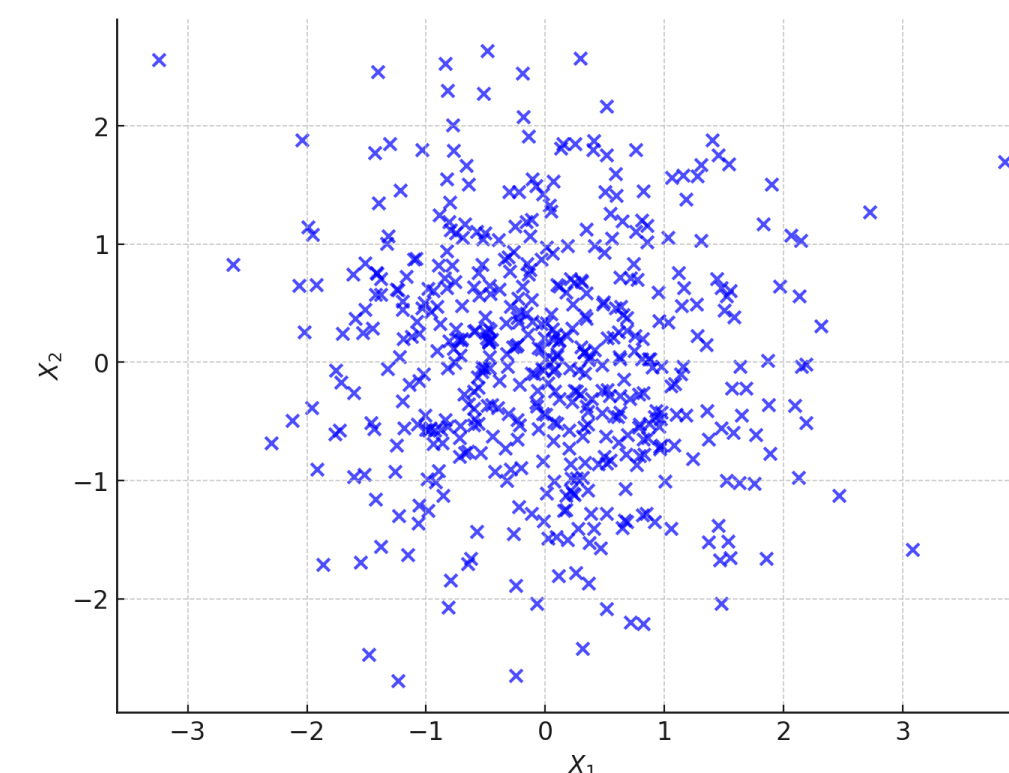
$\rho = 0.8$



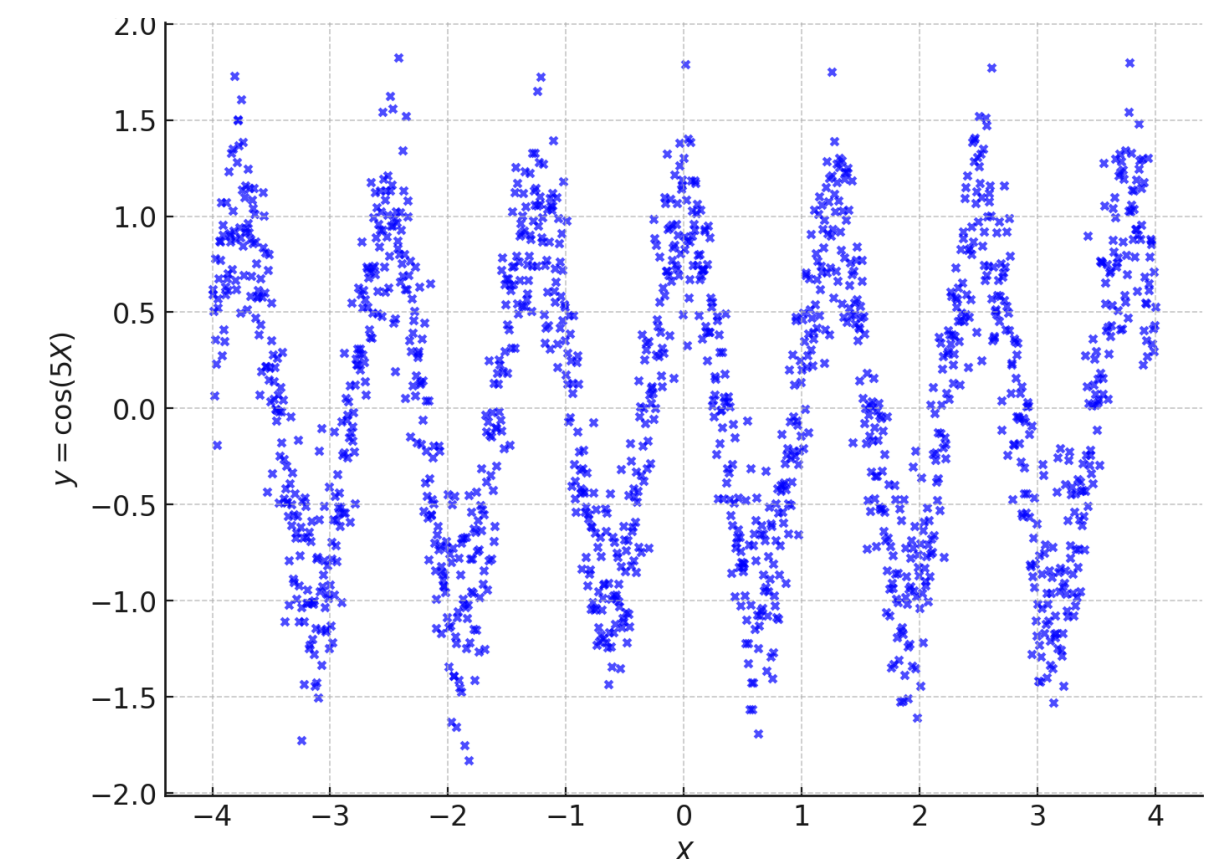
$\rho = -0.8$



$\rho = 0$



$\rho = 0$



Soit X_1 et X_2 deux variables aléatoires. On définit leur covariance par:

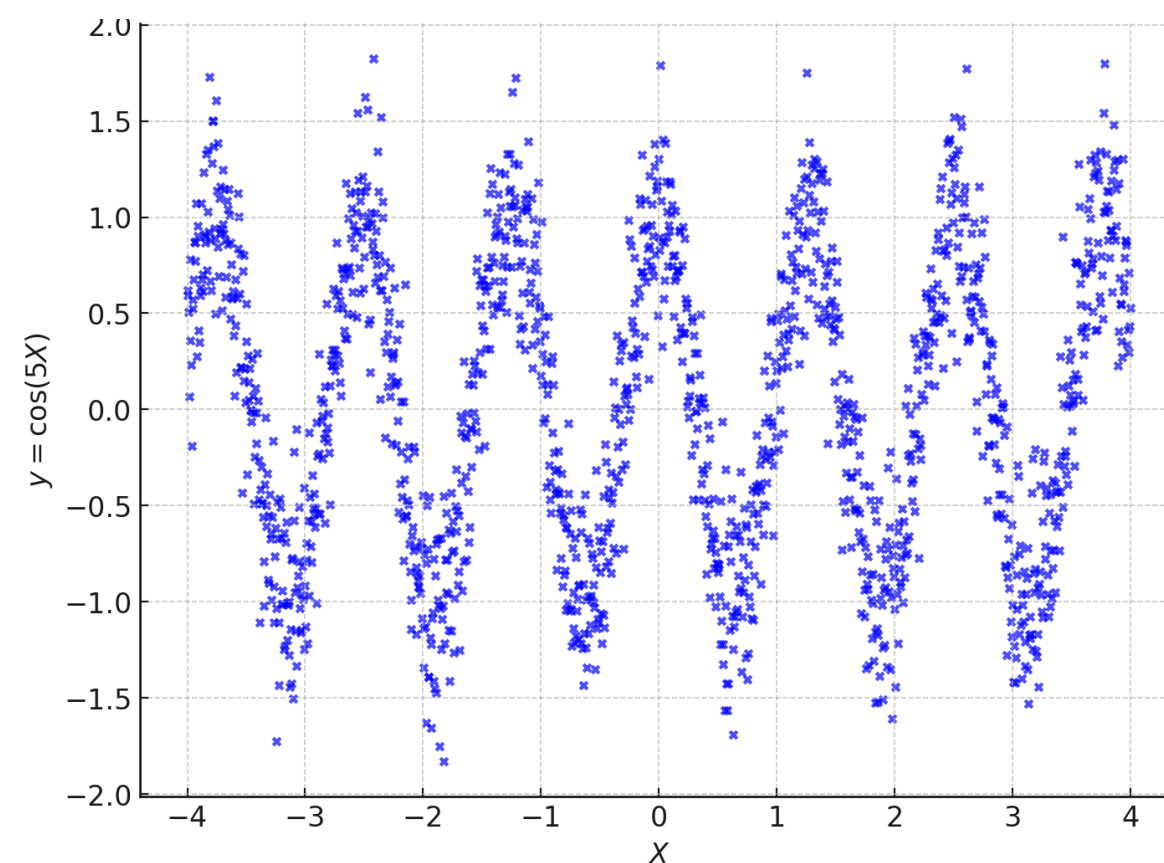
$$\text{Cov}(X_1, X_2) \stackrel{\text{def}}{=} \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)))$$

Si X_1 et X_2 sont indépendantes alors $\text{Cov}(X_1, X_2) = 0$

Qu'en est-il de la réciproque ? Que peut-on dire si la covariance est nulle ?

Contre-exemple:

$$\rho = 0$$



Dépendance sinusoïdale

Contre-exemple:

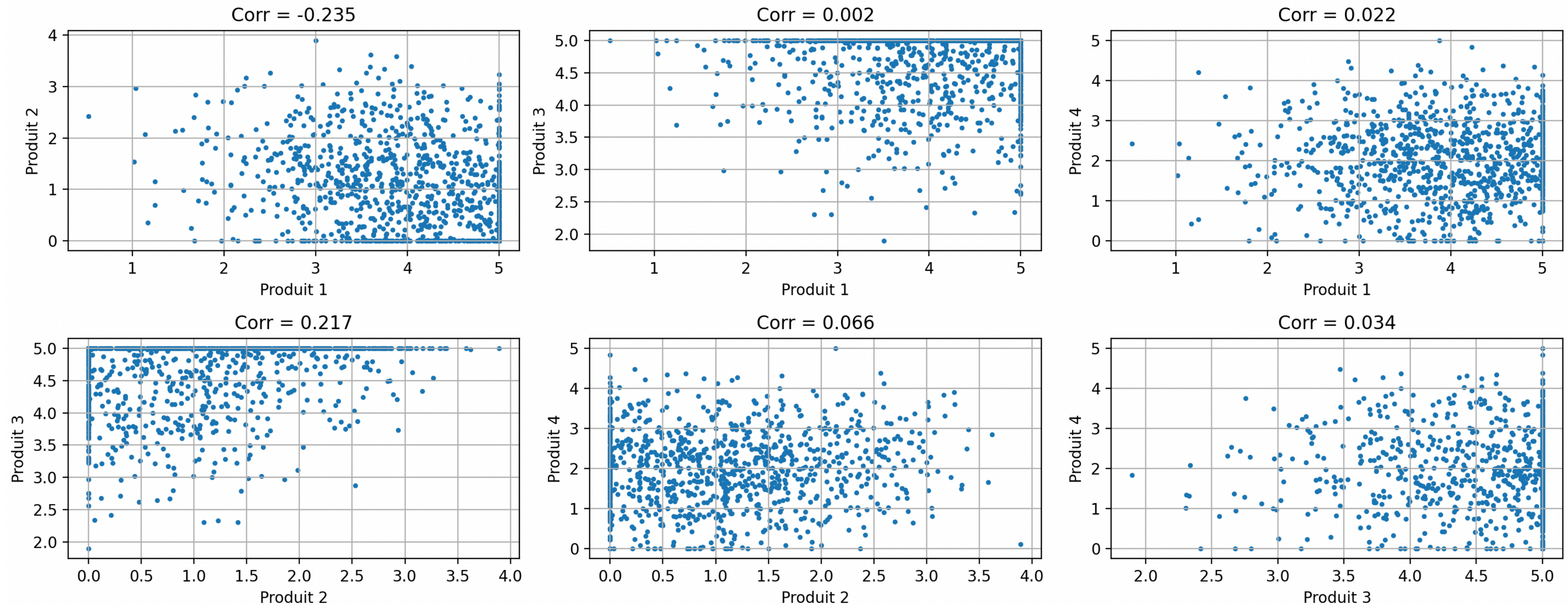
$$X_1 \sim \mathcal{U}(-1, 0, 1) \text{ et } X_2 = X_1^2.$$

$$\text{Ou bien } X_1 \sim \mathcal{N}(0, 1) \text{ et } X_2 = X_1^2.$$

Dépendance quadratique

Pour quelles dépendances la corrélation est elle égale à -1 ou 1 ?

On visualise les distributions “jointes” des variables avec des *scatter* plots + corrélations:



1. Pas de clusters, corrélations très petites ...
2. On a $d(d-1) / 2$ paires ...

Il faut étudier toutes les variables en même temps !

on observe n échantillons d'un **vecteur** aléatoire de dimension d .

On le note en gras $\mathbf{X} = (X_1, \dots, X_d)^\top$.

Son espérance est donnée par un vecteur dans \mathbb{R}^d :

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_d))^\top$$

À ne pas confondre avec
la matrice des observations $X \in \mathbb{R}^{n \times d}$!

Et sa variance est donnée par une matrice dans \mathbb{S}_d :

$$\begin{aligned} \mathbb{V}(\mathbf{X}) &= \begin{pmatrix} \mathbb{V}(X_1) & \cdots & \text{Cov}(X_1, X_d) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_d, X_1) & \cdots & \mathbb{V}(X_d) \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{E}[(X_1 - \mathbb{E}(X_1))^2] & \cdots & \mathbb{E}[(X_1 - \mathbb{E}(X_1))(X_d - \mathbb{E}(X_d))] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_d - \mathbb{E}(X_d))(X_1 - \mathbb{E}(X_1))] & \cdots & \mathbb{E}[(X_d - \mathbb{E}(X_d))^2] \end{pmatrix} \end{aligned}$$

Écrire cette matrice comme une espérance en fonction du vecteur \mathbf{X} et $\mathbb{E}(\mathbf{X})$.

Et sa variance est donnée par une matrice dans \mathbb{S}_d :

$$\begin{aligned}
 \mathbb{V}(\mathbf{X}) &= \begin{pmatrix} \mathbb{V}(X_1) & \cdots & \text{Cov}(X_1, X_d) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_d, X_1) & \cdots & \mathbb{V}(X_d) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[(X_1 - \mathbb{E}(X_1))^2] & \cdots & \mathbb{E}[(X_1 - \mathbb{E}(X_1))(X_d - \mathbb{E}(X_d))] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_d - \mathbb{E}(X_d))(X_1 - \mathbb{E}(X_1))] & \cdots & \mathbb{E}[(X_d - \mathbb{E}(X_d))^2] \end{pmatrix} \\
 &= \mathbb{E} \begin{pmatrix} (X_1 - \mathbb{E}(X_1))^2 & \cdots & (X_1 - \mathbb{E}(X_1))(X_d - \mathbb{E}(X_d)) \\ \vdots & \ddots & \vdots \\ (X_d - \mathbb{E}(X_d))(X_1 - \mathbb{E}(X_1)) & \cdots & (X_d - \mathbb{E}(X_d))^2 \end{pmatrix} \\
 &= \mathbb{E} \left[\begin{pmatrix} X_1 - \mathbb{E}(X_1) \\ \vdots \\ X_d - \mathbb{E}(X_d) \end{pmatrix} [X_1 - \mathbb{E}(X_1) \quad \cdots \quad X_d - \mathbb{E}(X_d)] \right] \\
 &= \mathbb{E}((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^\top)
 \end{aligned}$$

Comment estimer ces quantités théoriques ($\mathbb{E}(\mathbf{X})$ et $\mathbb{V}(\mathbf{X})$) avec la matrice des données $\mathbf{X} \in \mathbb{R}^{n \times d}$?

On considère n observations $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ du vecteur aléatoire \mathbf{X} données par les lignes de la matrice X

L'estimateur empirique de $\mathbb{E}(\mathbf{X})$ est donné par la moyenne empirique: $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$

L'estimateur empirique de $\mathbb{V}(\mathbf{X}) = \mathbb{E}((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^\top)$ est donné par: $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$

Écrire la matrice des données X en fonction des \mathbf{x}_i .

$$\text{Montrez que } \hat{\Sigma} = \frac{1}{n} (X^\top X - n\bar{\mathbf{x}}\bar{\mathbf{x}}^\top).$$

Remarquez les positions différentes de la transposée !

Pour ne jamais se tromper: se rappeler que la matrice de covariance donne les interactions entre toutes les d variables: elle doit être $(d \times d)$

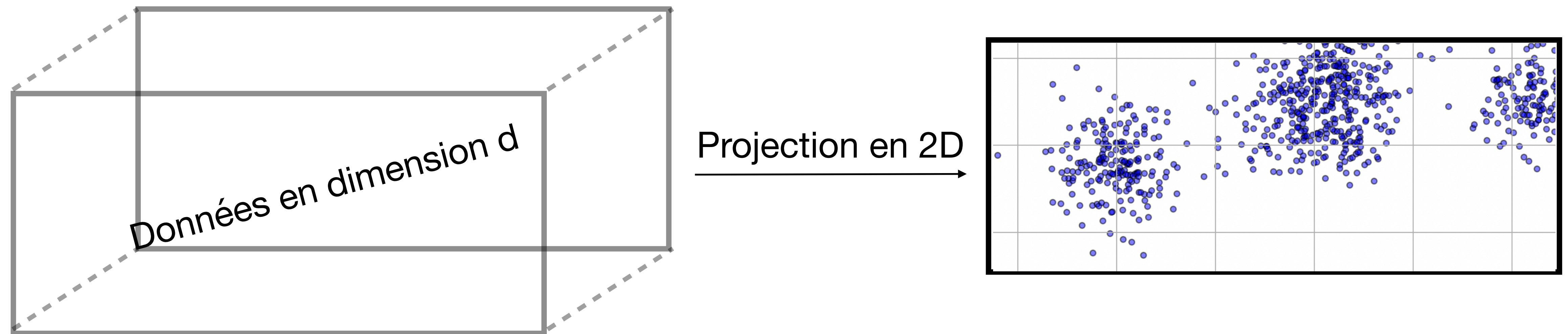
Avec nos données on trouve:

	Produit 1	Produit 2	Produit 3	...	Produit 198	Produit 199	Produit 200
Produit 1	0.775242	-0.181769	0.000907	...	-0.009854	0.019631	0.046356
Produit 2	-0.181769	0.774950	0.110453	...	0.064765	-0.012393	-0.103663
Produit 3	0.000907	0.110453	0.335673	...	-0.001068	0.001316	-0.041009
Produit 4	0.018753	0.056509	0.019479	...	-0.112001	-0.012429	-0.095697
Produit 5	0.101587	-0.071179	-0.026813	...	-0.078027	-0.004025	-0.055984
...
Produit 196	0.000146	-0.004620	-0.002949	...	0.001877	0.000248	0.002607
Produit 197	0.042239	-0.080690	-0.010092	...	-0.019513	0.004708	0.021961
Produit 198	-0.009854	0.064765	-0.001068	...	0.711461	-0.000048	0.144218
Produit 199	0.019631	-0.012393	0.001316	...	-0.000048	0.050432	0.007491
Produit 200	0.046356	-0.103663	-0.041009	...	0.144218	0.007491	0.373554

L'utilité de cette matrice sera plus claire dans quelques slides.

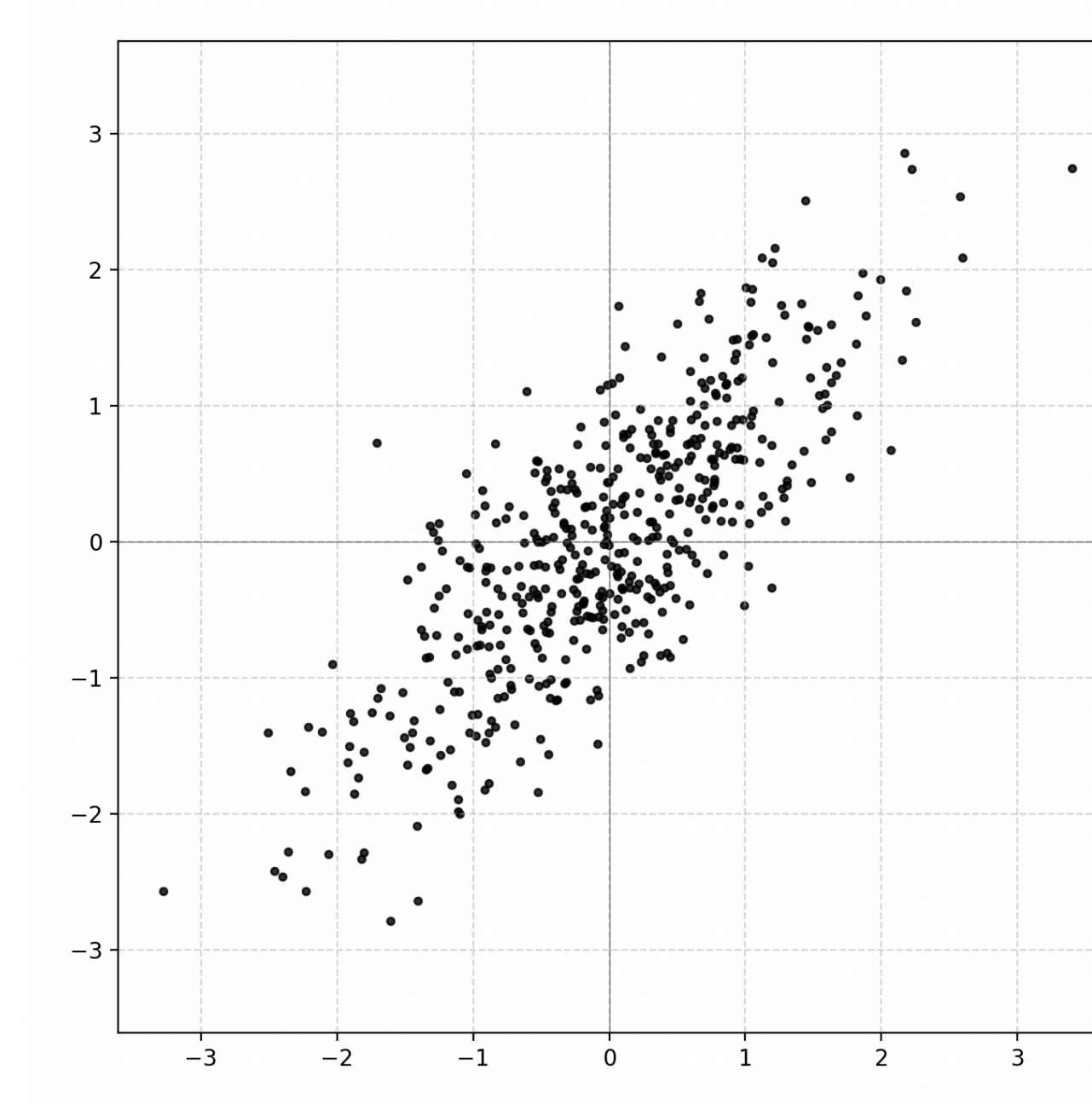
1. L'étude univariée est limitée (ne prend pas en compte les corrélations)
2. Toute visualisation utile des données est limitée en 2D (voire 3D).
3. L'étude bi-variée est infaisable en grande dimension et ne prend pas en compte toute la matrice de covariance

Quelle est “la meilleure” visualisation en 2D/3D des données ?

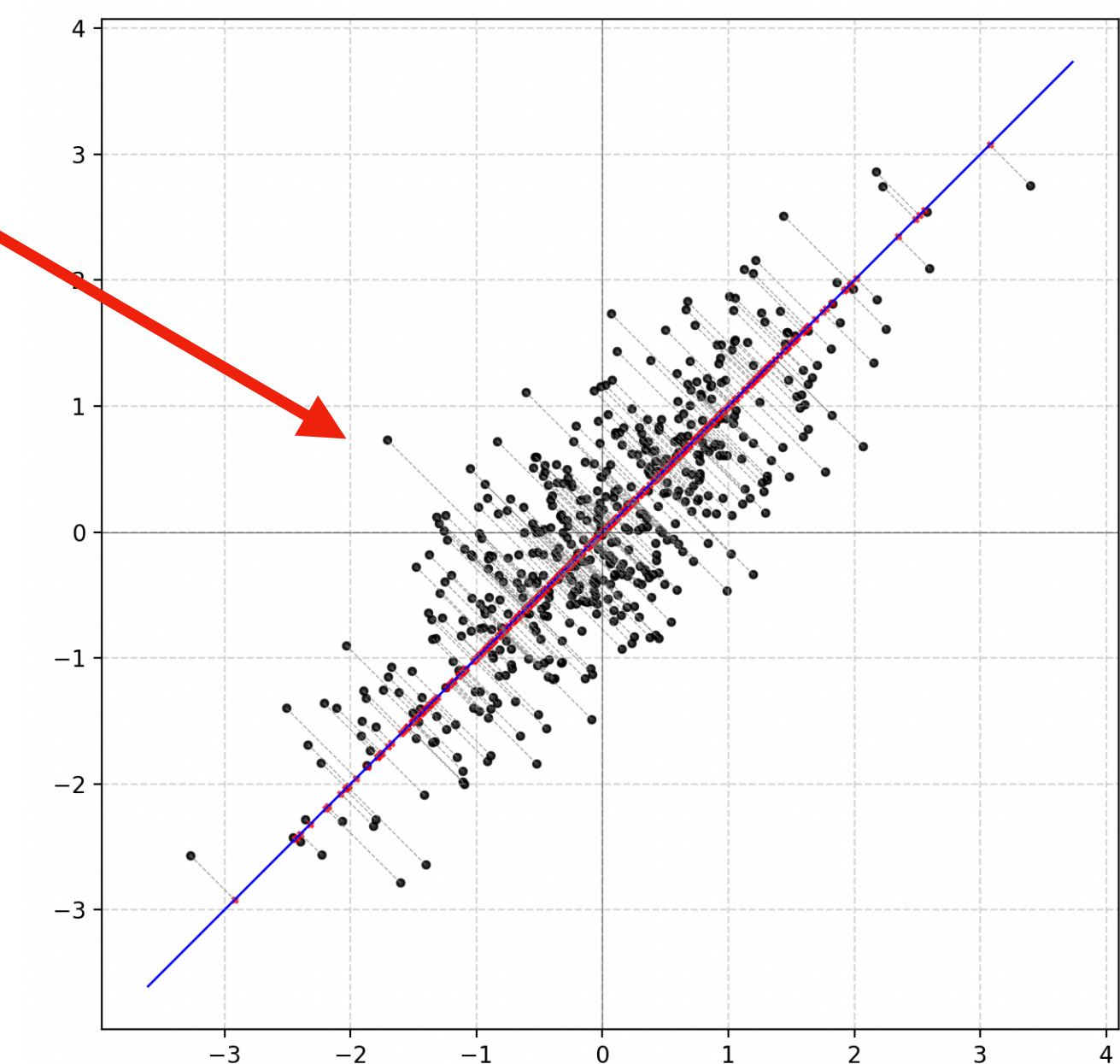


Idée: choisir le “meilleur” sous-espace vectoriel sur lequel on projette les données

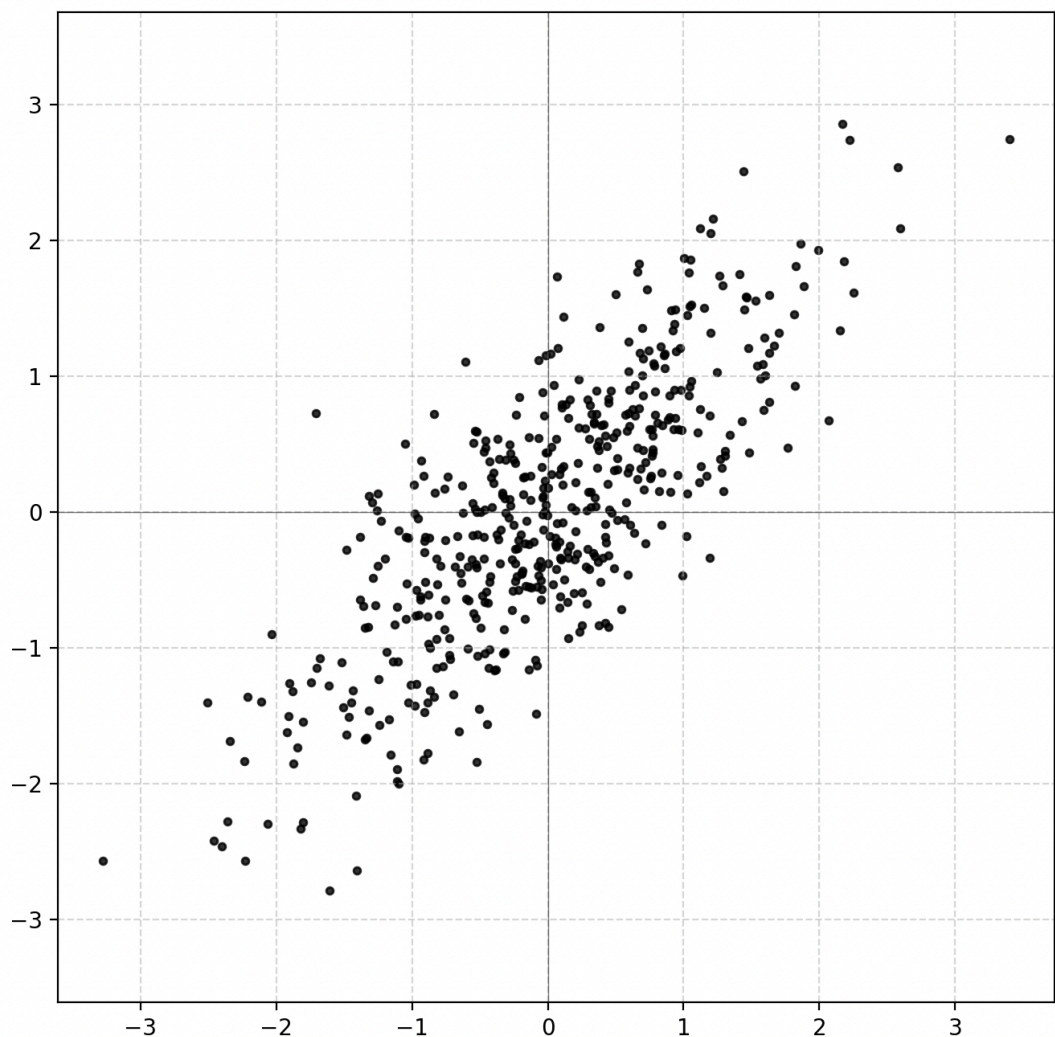
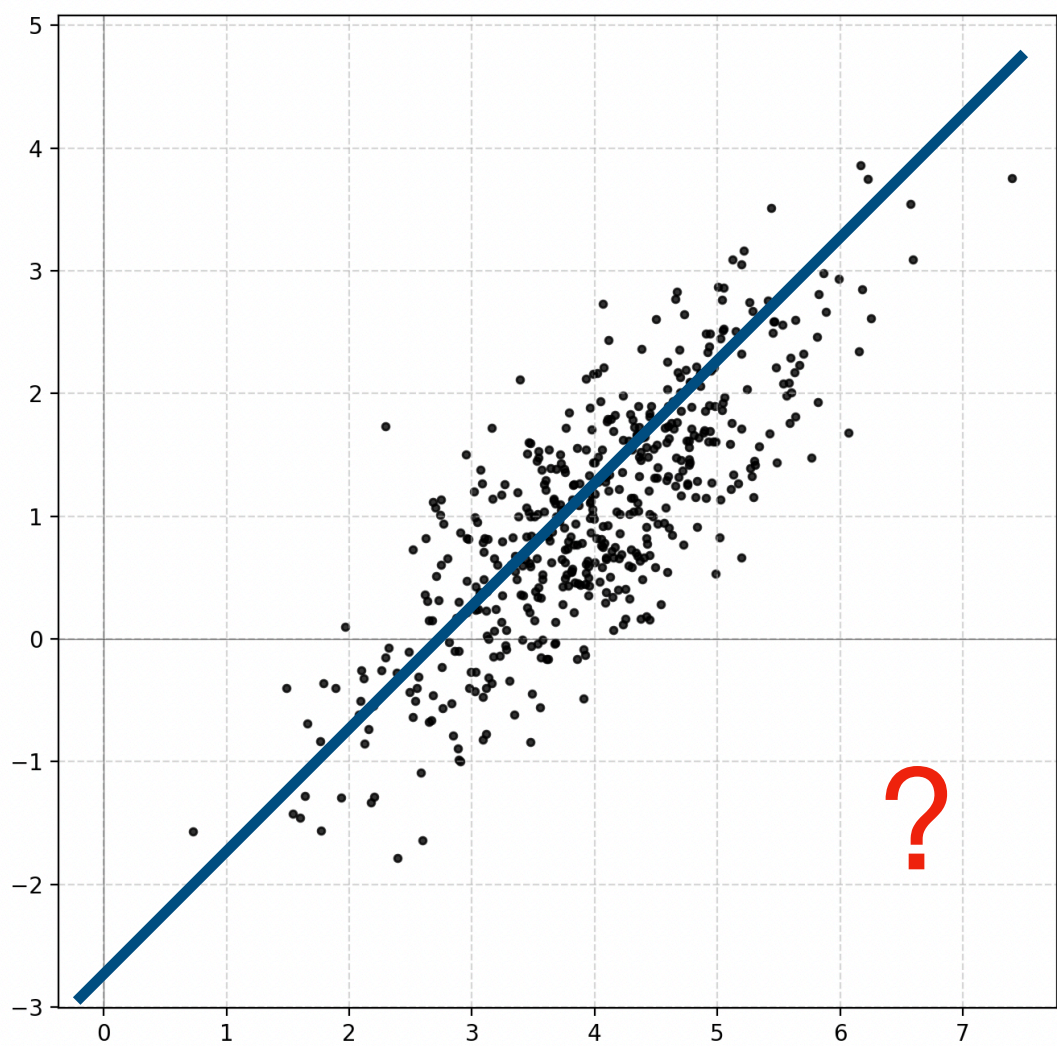
Commençons par des exemples visuels en deux dimensions, projetée sur une dimension i.e une droite.



Quel est le meilleur sous-espace vectoriel de dimension 1 sur lequel on doit projeter les données ?



Commençons par des exemples visuels en deux dimensions, projetée sur une dimension i.e une droite.

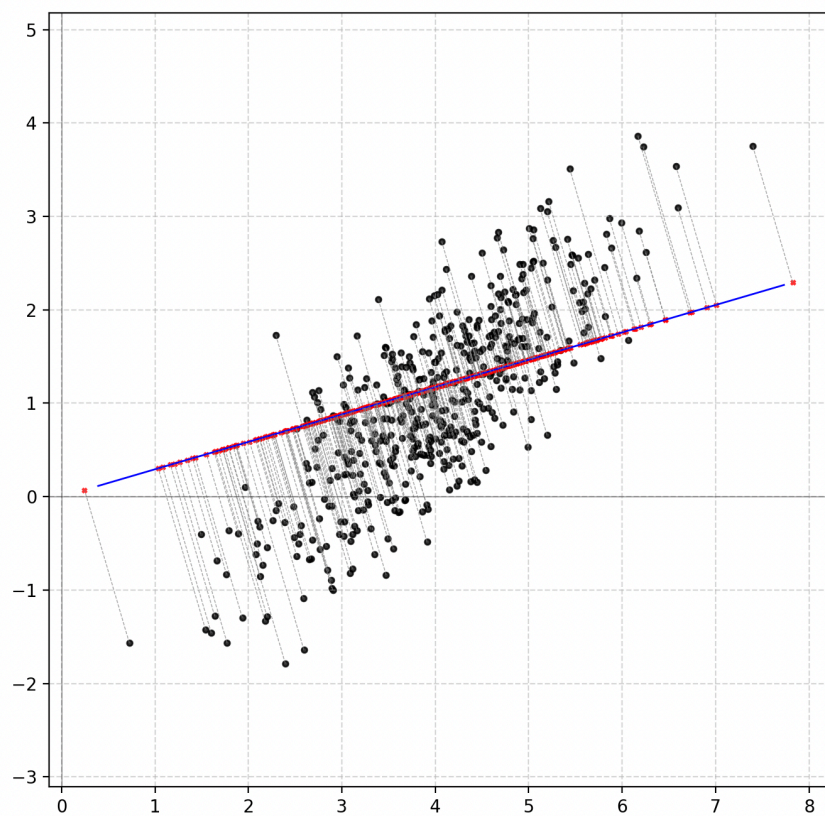


Quel est le meilleur sous-espace vectoriel de dimension 1 sur lequel on doit projeter les données ?

Quel est le problème avec cette projection ?

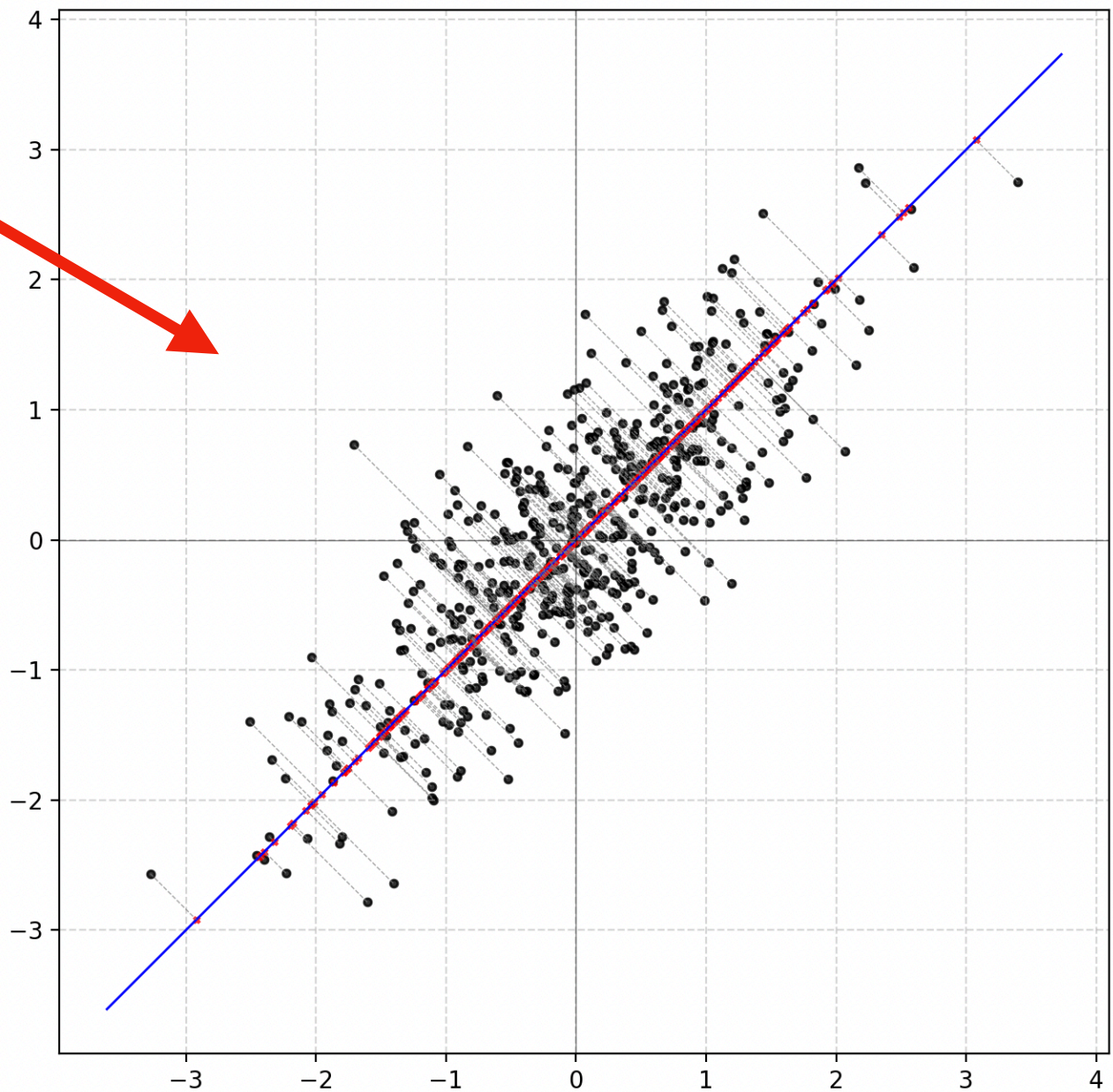
La droite ne passe pas par l'origine: ce n'est pas un sous-espace vectoriel !

Le meilleur sous-e.v est donné par la droite:

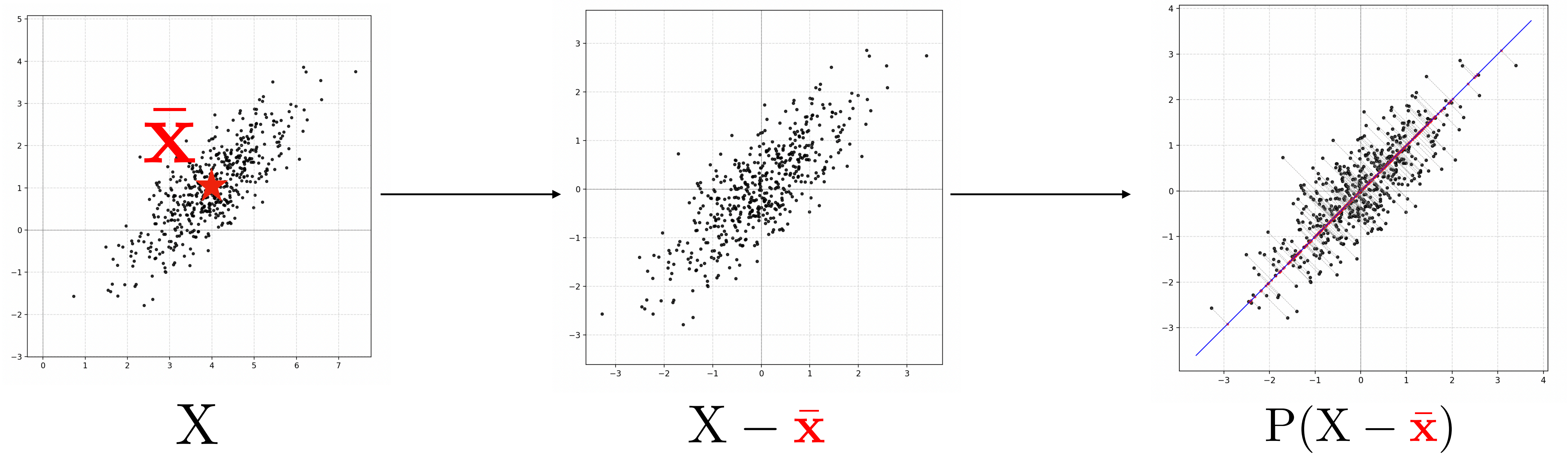


Comment peut-on résoudre ce problème ?

En centrant les données avant la projection



En centrant les données avant la projection:



Passons à présent l'analyse théorique (TD 1).

Dans le TD 1, nous avons établi le résultat suivant:

Analyse en composantes principales introduite gentiment

Soit \mathbf{X} un vecteur aléatoire réel en dimension d tel que $\mathbb{E}(\mathbf{X}) = 0$.

On note $X \in \mathbb{R}^{n \times d}$ la matrice dont les lignes sont les observations:

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}$$

La matrice de covariance empirique est donnée par $\Sigma = \frac{1}{n} X^\top X$.

Σ est symétrique: par le théorème spectral, il existe une matrice orthogonale \mathbf{P} et une matrice diagonale Λ telles que:

$$\Sigma = \mathbf{P} \Lambda \mathbf{P}^\top$$

Les colonnes \mathbf{p}_i de \mathbf{P} correspondent aux vecteurs propres de Σ . On suppose que $\lambda_1 \geq \dots \geq \lambda_d$

Alors, la meilleure projection orthogonale “moyenne” sur un sous-espace de dimension k définie par:

$$\min_{\substack{E \\ \dim(E)=k}} \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \text{proj}_E(\mathbf{x}_i)\|^2$$

est donnée par $E = \text{Vect}(\mathbf{p}_1, \dots, \mathbf{p}_k)$. Posons $\mathbf{P}_k = [\mathbf{p}_1 \ \dots \ \mathbf{p}_k] = P[:, :k] \in \mathbb{R}^{d \times k}$ alors:

Les projections des \mathbf{x}_i sur E sont données par les lignes de la matrice: $X \mathbf{P}_k \in \mathbb{R}^{n \times k}$.

Dans le TD 1, nous avons établi le résultat suivant:

Analyse en composantes principales

Soit \mathbf{X} un vecteur aléatoire réel en dimension d tel que $\mathbb{E}(\mathbf{X}) = 0$.

Alors, la meilleure projection orthogonale “moyenne” sur un sous-espace de dimension k définie par:

$$\min_{\substack{E \\ \dim(E)=k}} \mathbb{E} (\|\mathbf{X} - \text{proj}_E(\mathbf{X})\|^2) \quad \text{est donnée par } E = \text{Vect}(\mathbf{p}_1, \dots, \mathbf{p}_k).$$

Où les $\mathbf{p}_1, \dots, \mathbf{p}_d$ sont les vecteurs propres de $\mathbb{V}(\mathbf{X})$ dans l'ordre décroissant des valeurs propres.

$\mathbb{V}(\mathbf{X})$ est symétrique: par le théorème spectral, il existe $\mathbf{P} \in \mathbb{R}^{d \times d}$ orthogonale et $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$ telles que:

$$\mathbb{V}(\mathbf{X}) = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^\top$$

Les vecteurs propres \mathbf{p}_i sont les colonnes de \mathbf{P} avec $\lambda_1 \geq \dots \geq \lambda_d$

Et la projection de \mathbf{X} est donnée par: $\text{proj}_E(\mathbf{X}) = \mathbf{P}[:, : k]^\top \mathbf{X} \in \mathbb{R}^k$

$$\text{Et } \mathbb{V}(\text{proj}_E(\mathbf{X})) = \text{diag}(\lambda_1, \dots, \lambda_k)$$

Remarques:

1. En pratique, on estime \mathbf{P} en diagonalisation la covariance empirique $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$.
2. Le nouveau vecteur projeté en dimension k donné par $\mathbf{P}^\top[:, : k]\mathbf{X}$
3. La PCA est donc une transformation linéaire
4. Ce nouveau vecteur a une covariance diagonale $\text{diag}(\lambda_1, \dots, \lambda_k)$: la PCA donne de nouvelles variables non corrélées.
5. La variance totale des données (somme des variance de chaque dimension) est donnée par $\sum_{i=1}^d \lambda_i$
6. La variance totale des données projetées est $\sum_{i=1}^k \lambda_i$.
7. Le pourcentage de variance retenu par l'axe principal j est donc $\frac{\lambda_j}{\sum_{i=1}^d \lambda_i}$
8. Plus ce pourcentage est élevé, plus cet axe principal est relativement important
9. La PCA peut également être définie par la projection maximisant la variance
10. Il ne faut pas oublier que les vecteurs propres sont des combinaisons linéaires des colonnes de \mathbf{X} : les nouveaux axes principaux sont donc des combinaisons linéaires des variables d'origine.

1. Projeter en dimension 2 pour visualiser

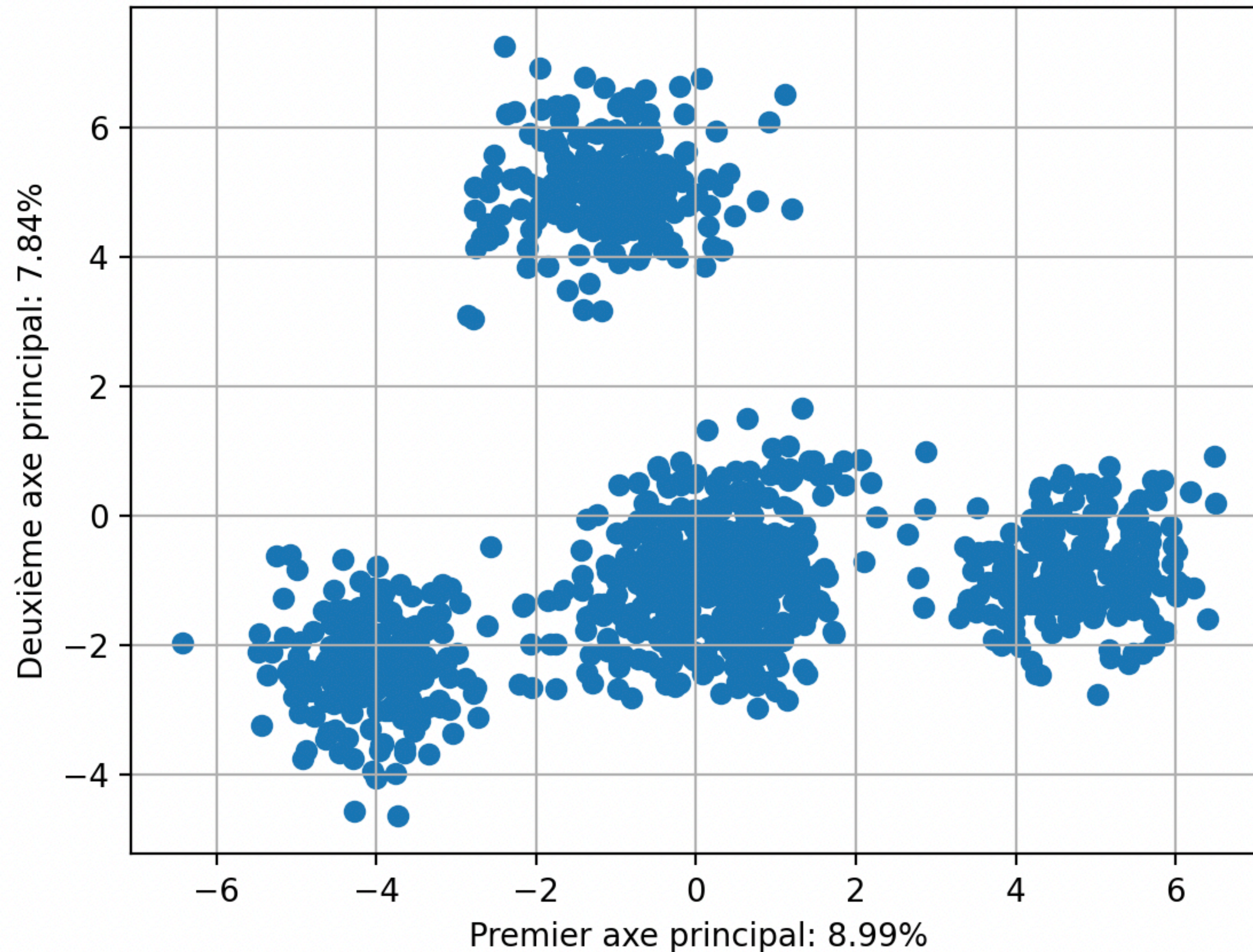
1. Projeter en dimension 2 pour visualiser

Exemple: données “ratings” des produits par des clients

Customer	Product 1	Product 2	...	Product d
Customer 1	4.5	3.0	...	5.0
Customer 2	3.5	4.0	...	2.5
Customer 3	5.0	2.5	...	4.0
⋮	⋮	⋮	⋮	⋮
Customer n	4.0	4.5	...	3.5

ici: $n = 1000$ et $d = 200$

Données “ratings” des produits par des clients



En tant que data scientist, que pouvez-vous dire ?

1. Il existe en moins 4 groupes de clients différents
2. Les deux axes principaux expliquent 16% de la variance totale des données.
3. La PCA est une transformation linéaire et elle a permis de détecter les “clusters”: les données sont donc plus ou moins “simples”

1. Projeter en dimension 2 pour visualiser

2. Préparer les données pour un modèle statistique / ML qui vérifie l'une des deux conditions:

1. nécessite des variables non-corrélées

2. n'est pas adapté pour une très grande dimension

En Machine learning, on utilise la PCA très souvent pour réduire la dimension

par exemple de 1000 / 5000 / 10000 à 50 / 100 / 200

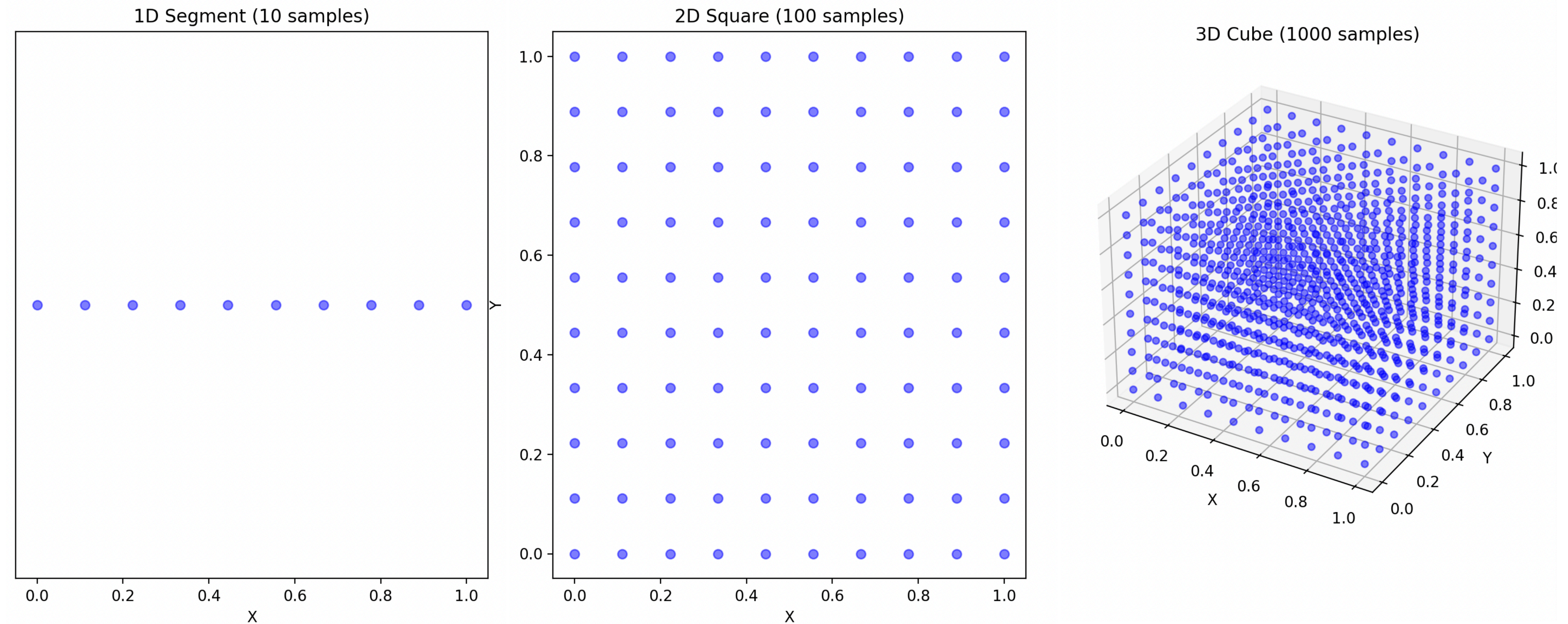
La raison principale:

“Curse of dimensionality”

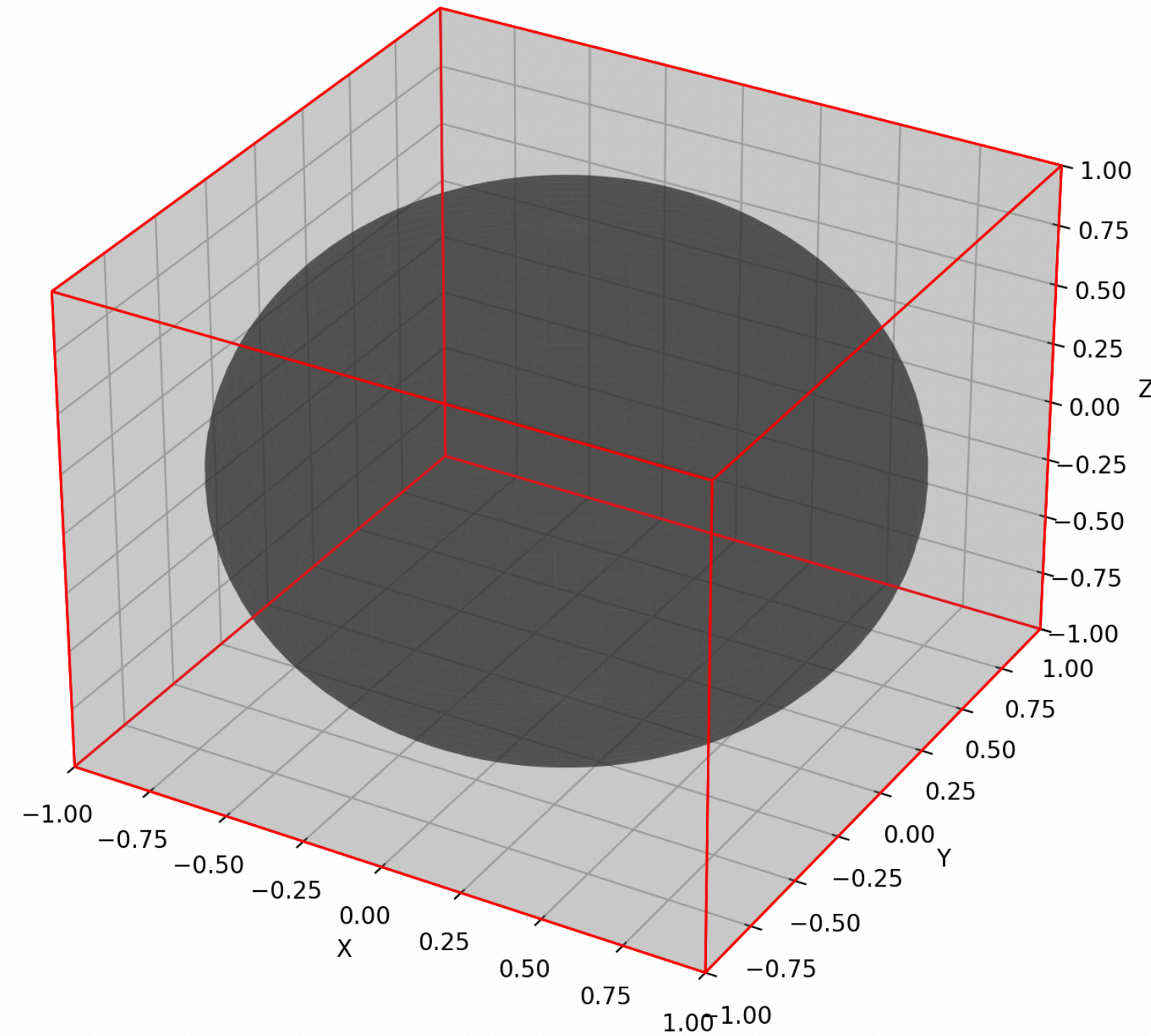
“La malédiction de la dimension”

What is the “curse” of high dimension ?

On veut “reconnaître” la distribution uniforme sur la boule unité de norme infinie:



Plus la dimension augmente plus on a besoin d'**exponentiellement plus** d'observations pour “caractériser” les distributions



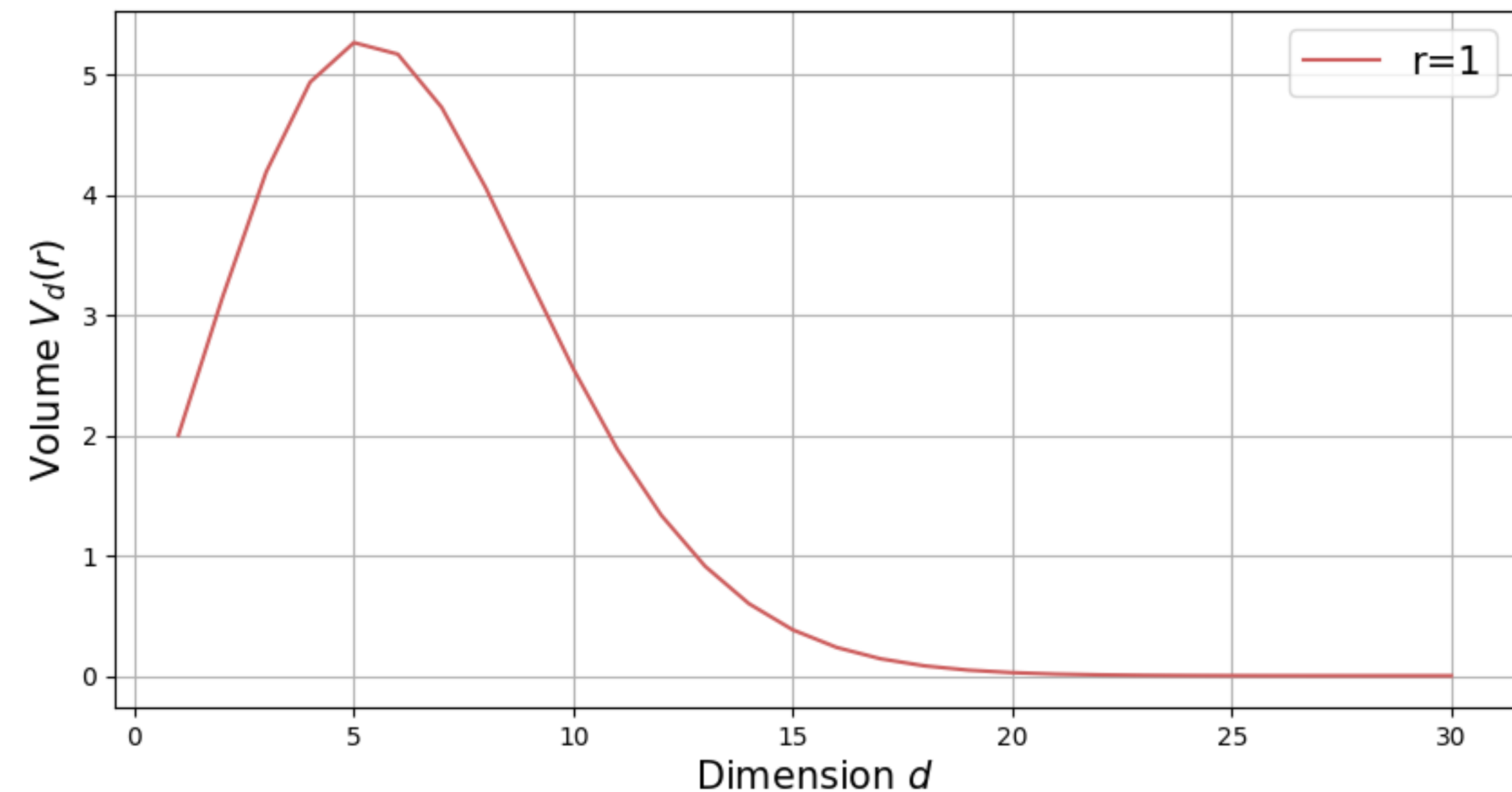
Boule de rayon r : $\mathcal{B}(r) = \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 \leq r\}$

Quel est son volume ?

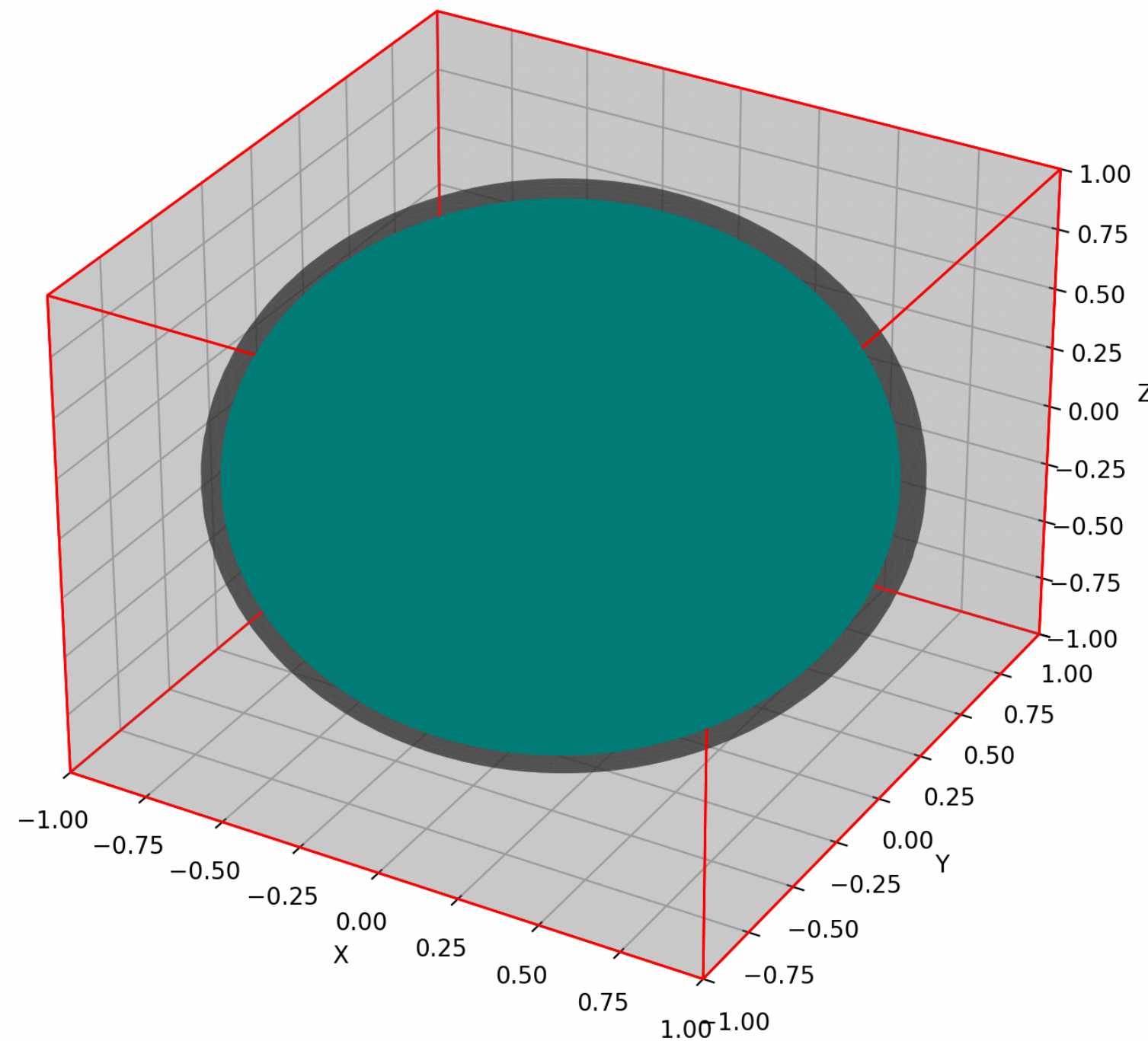
$$V(r) \stackrel{\text{def}}{=} \int_{\mathcal{B}(r)} d\mathbf{x} = ? \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} r^d$$

Comment évolue-t-il en grande dimension ?

La grande dimension est contre-intuitive



Où se concentre le volume à l'intérieur de la sphère ? Au centre ou vers les bords ?



$$V(r) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} r^d$$

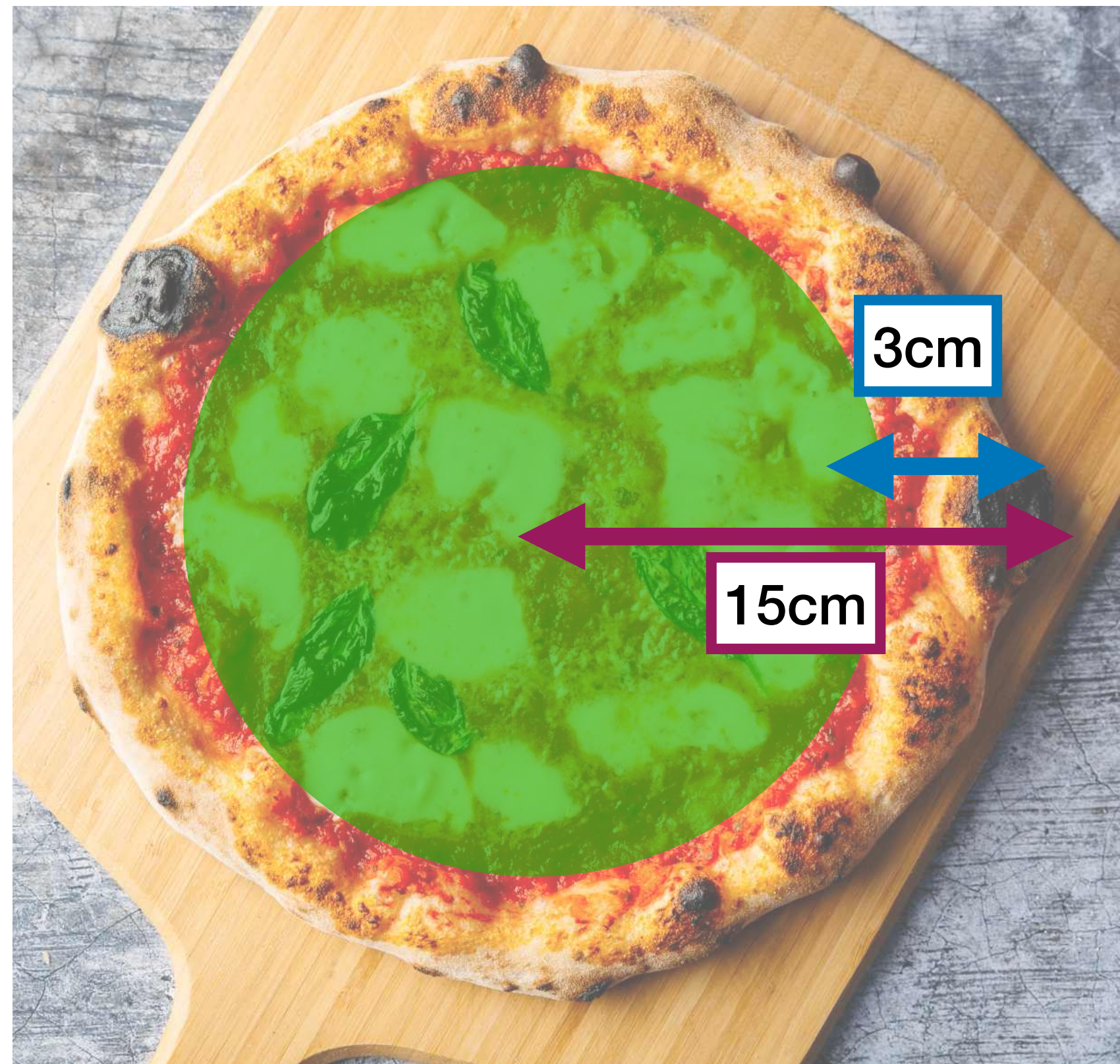
Volume de la sphère à l'intérieur avec un rayon = 99% x r:

$$V(0.99r) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} (0.99r)^d$$

$$\frac{V(0.99r)}{V(r)} = 0.99^d \quad d \rightarrow +\infty \quad 0$$

La majorité du volume d'une sphère en grande dimension est dans les bords !

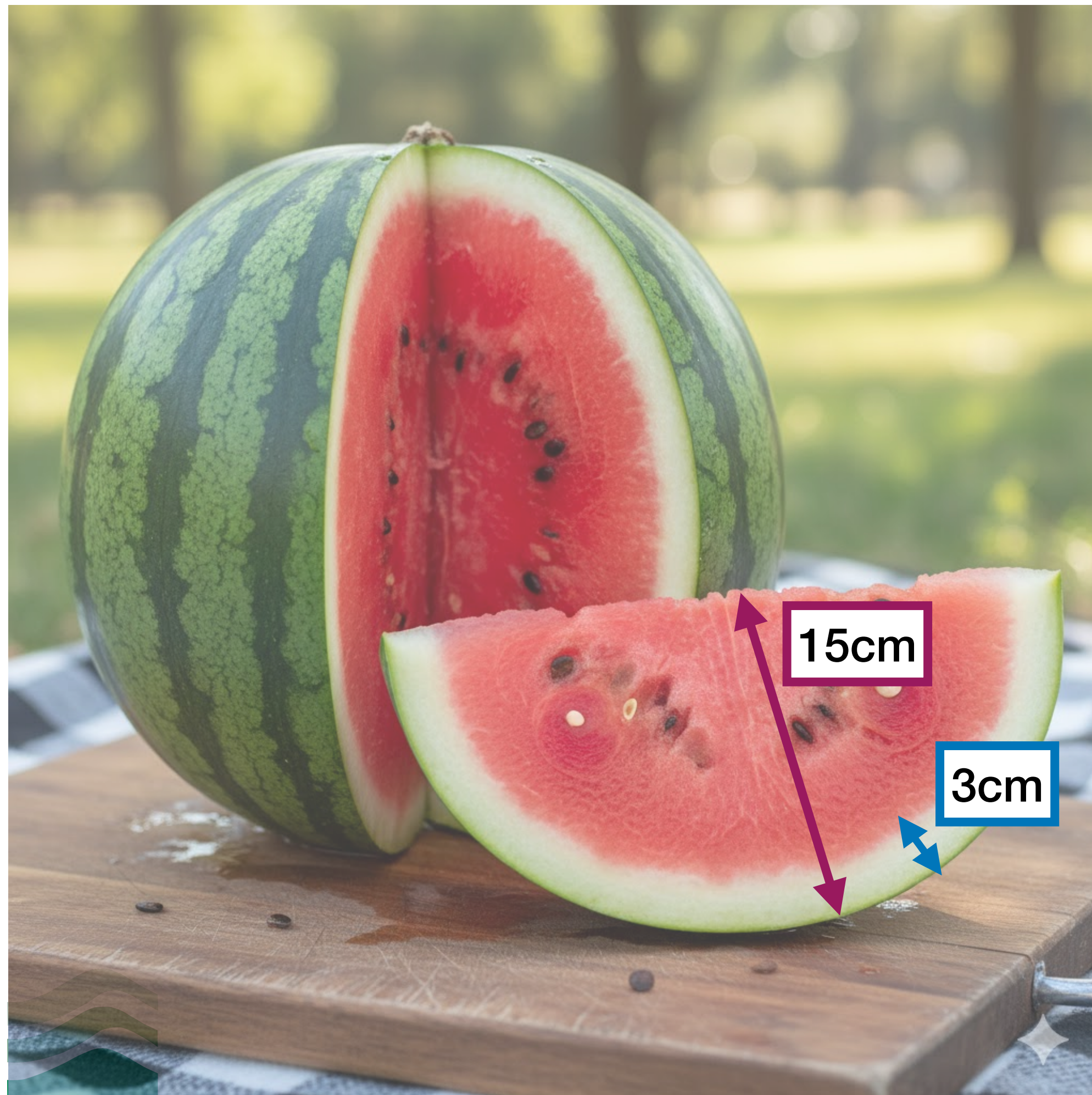
Quelle est le pourcentage de la partie centrale d'une pizza ?



En terme de rayon: $12/15 = 80\%$

En terme de surface: $\pi 12^2 / \pi 15^2 = 64\%$

Volume de la partie rouge d'une pastèque ?



En terme de rayon: $12/15 = 80\%$

En terme de volume: $\frac{4/3 \pi 12^3}{4/3 \pi 15^3} = 51\%$

Explication intuitive

On doit générer un vecteur $\mathbf{x} = (x_1, \dots, x_d)^\top$ uniformément dans l'hypercube $[-1, 1]^d$

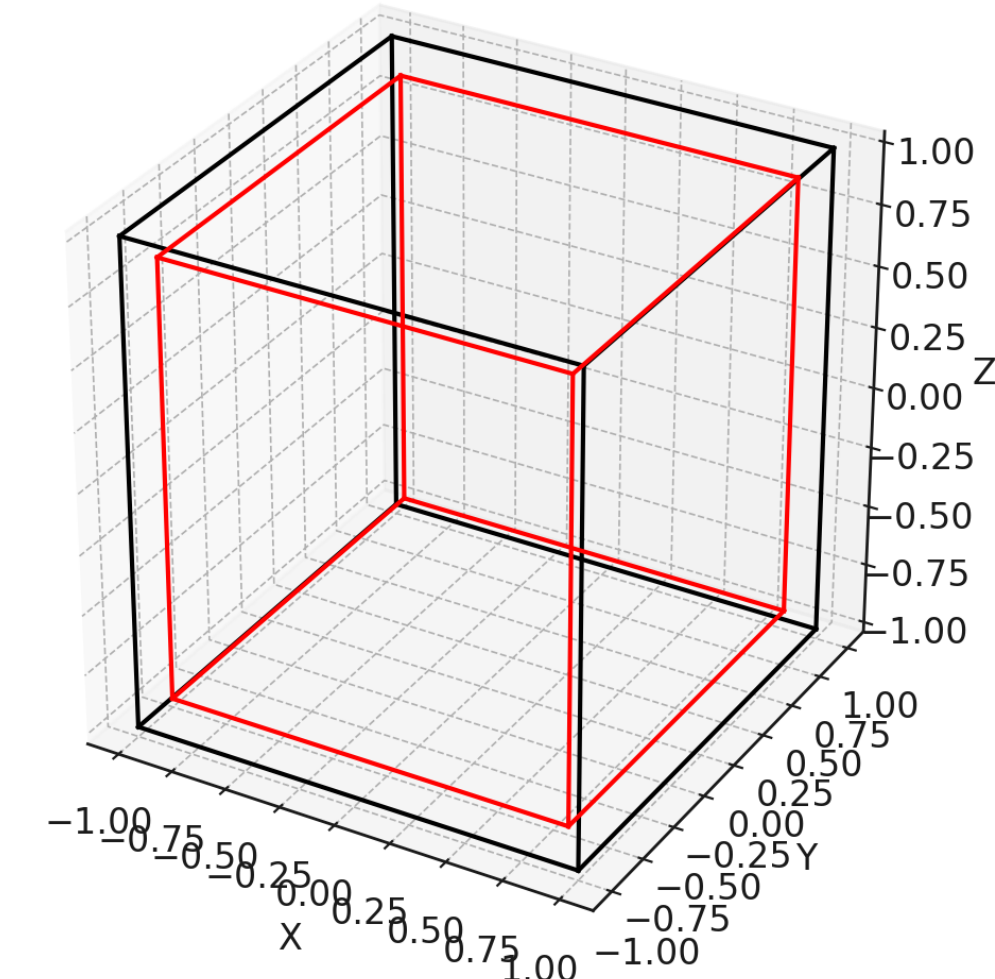
Chaque coordonnée x_i peut être tirée aléatoirement dans $[-1, 1]$ indépendamment des autres

Quelle est la probabilité que \mathbf{x} soit à l'intérieur de l'hypercube $[-0.9, 0.9]^d$?

Il faut que **toutes** les coordonnées soient dans $[-0.9, 0.9]$

$$\begin{aligned}\mathbb{P}((x_1, \dots, x_d) \in [-0.9, 0.9]^d) &= \mathbb{P}(\forall i \ x_i \in [-0.9, 0.9]) \\ &= \mathbb{P}(x_1 \in [-0.9, 0.9])^d \\ &= \left(\frac{1.8}{2}\right)^d = 0.9^d\end{aligned}$$

qui tend exponentiellement vers 0



Pour que \mathbf{x} soit sur la frontière, il suffit qu'**une seule coordonnée** soit assez grande en valeur absolue !

Gaussienne en grande dimension

Soit $\mathbf{X} \sim \mathcal{N}(0, I_d)$. Alors:

$$\mathbb{P} \left(\left| \|\mathbf{X}\|_2 - \sqrt{d} \right| \geq t \right) \leq 2e^{-Ct^2}$$

Concentration inequality

Avec C une constante positive

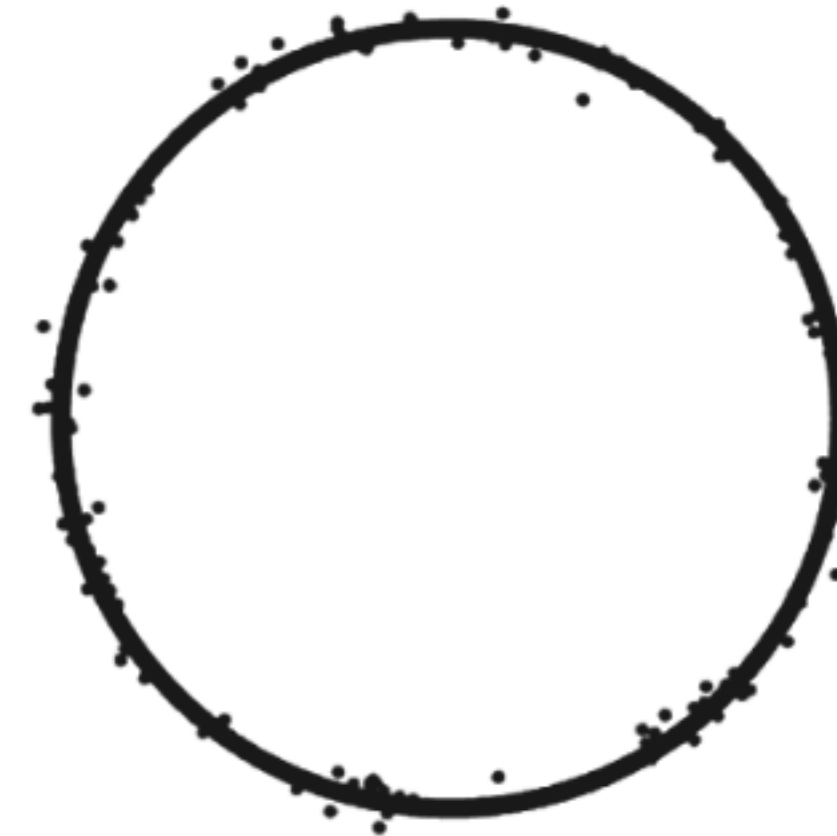
Que peut-on en déduire ?

Proof can be found in <https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-2.pdf> section 3.3.3

Visualisations de nuages de points Gaussiens

 $d = 2$ 

“intuition” en grande dimension



Comment se comporte la distance entre deux points aléatoires en grande dimension ?

Les distances entre deux points aléatoires augmente !

Les conséquences en grande dimension

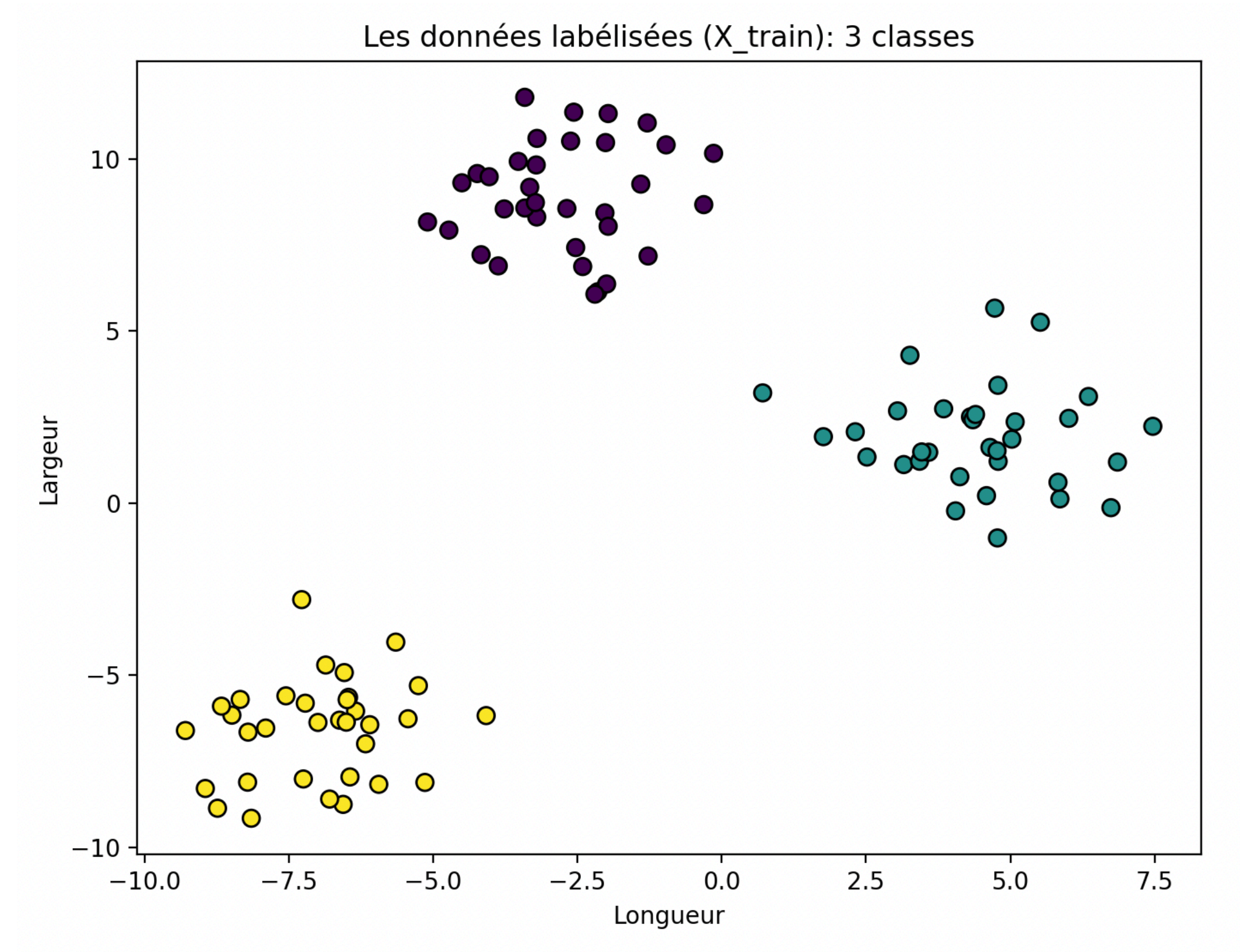
1. La géométrie est “bizarre”, contre-intuitive
2. La Gaussienne se comporte comme une loi uniforme sur une sphère
3. Les distances ne sont plus efficaces pour établir une notion de proximité / voisinage

Plusieurs méthodes de machine learning sont basées sur des distances, par exemple:

k-NN, K-means, SVM...

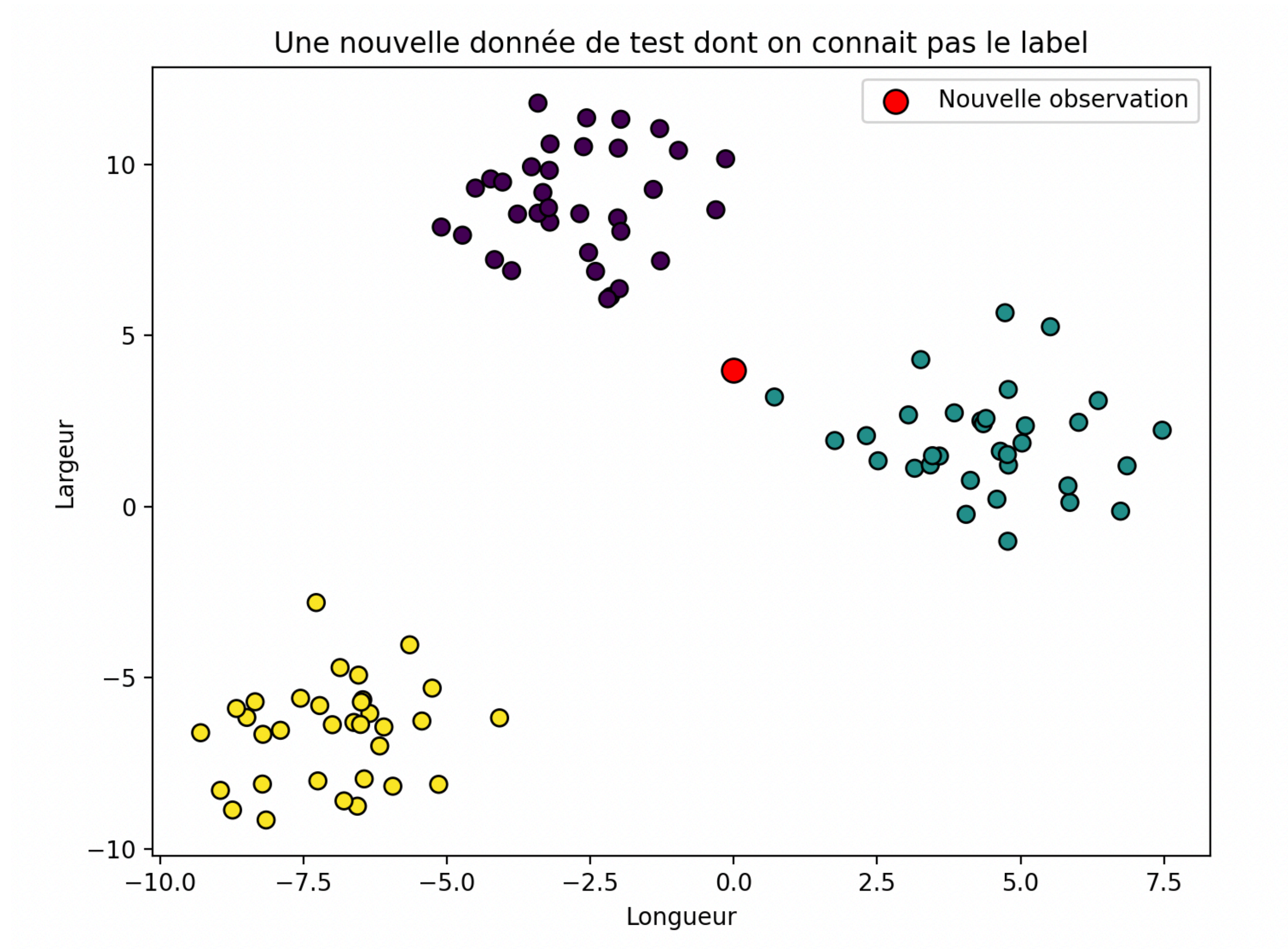
Nous avons les mesures de longueur et largeur de plusieurs voitures ainsi que leur marque. Nous avons 3 marques différentes. On note les mesures par $\mathbf{x}_i \in \mathbb{R}^2$ et la marque par $y \in \{0, 1, 2\}$.

les y_i sont appelés des *labels* ou des étiquettes.

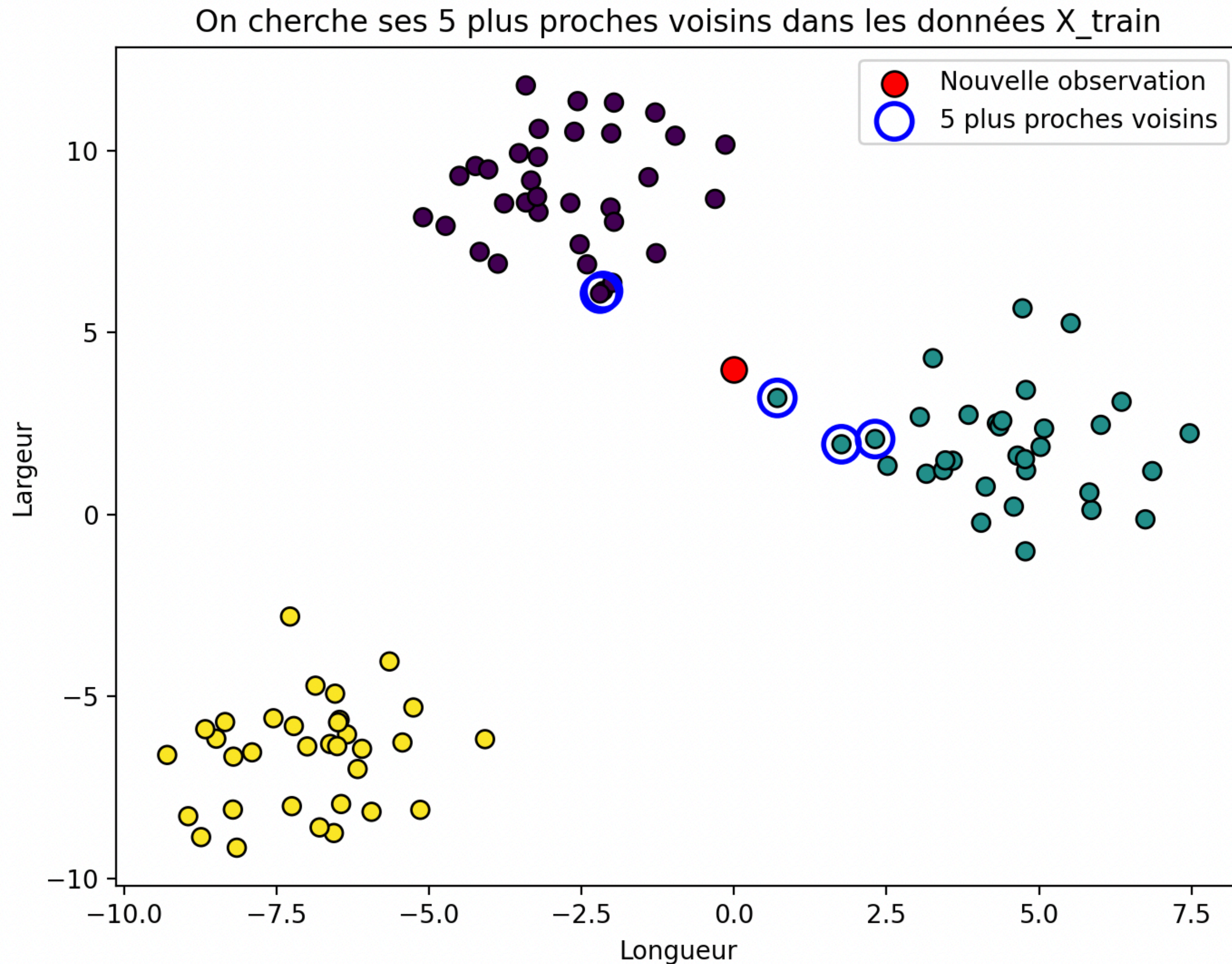


Les données pour lesquelles on a les labels sont dit des données d'entraînement. On les appelle X_{train} et y_{train} .

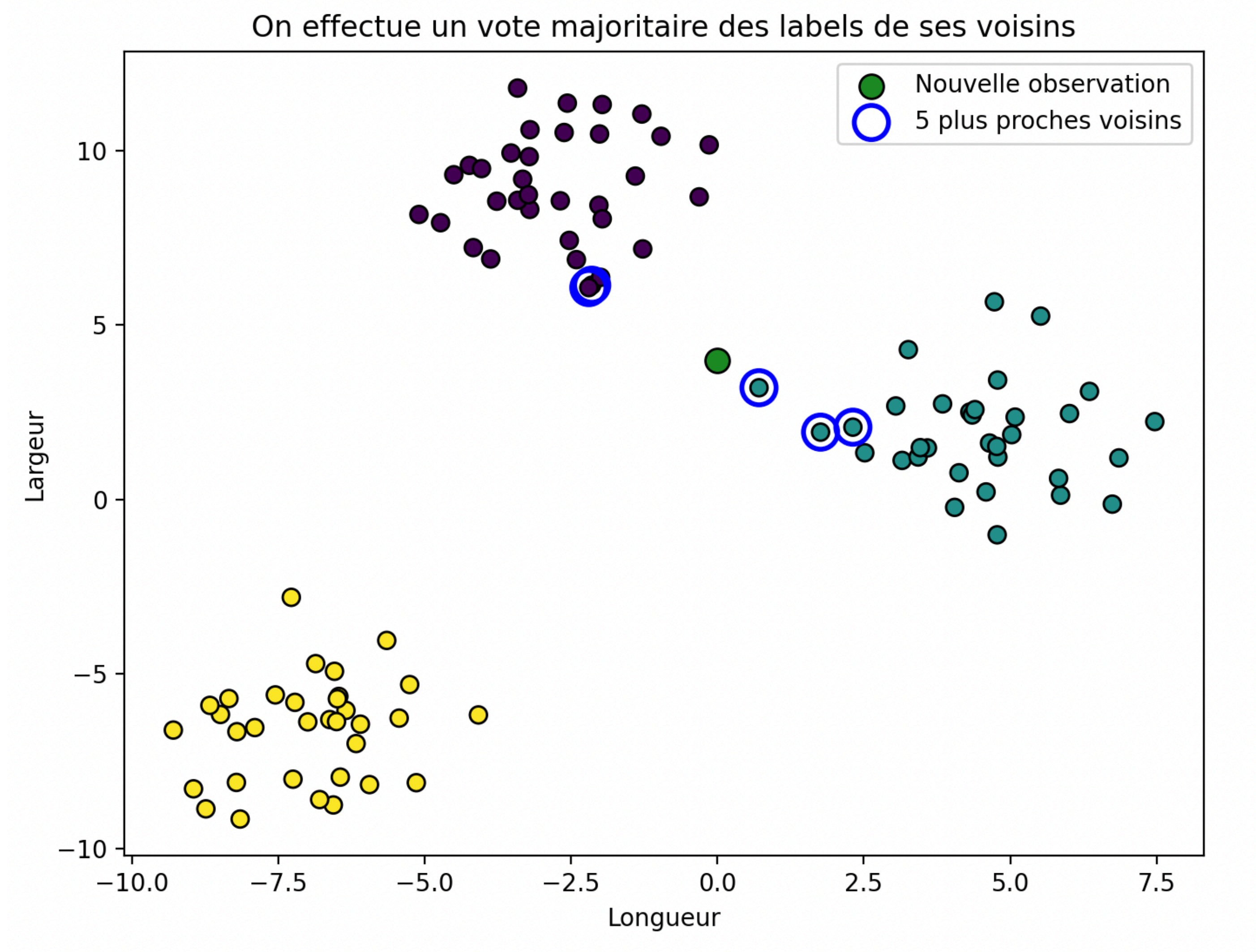
On a une nouvelle voiture dont on ne connaît pas la marque:



On cherche les 5 voitures qui lui ressemblent le plus: ses 5 plus proches voisins

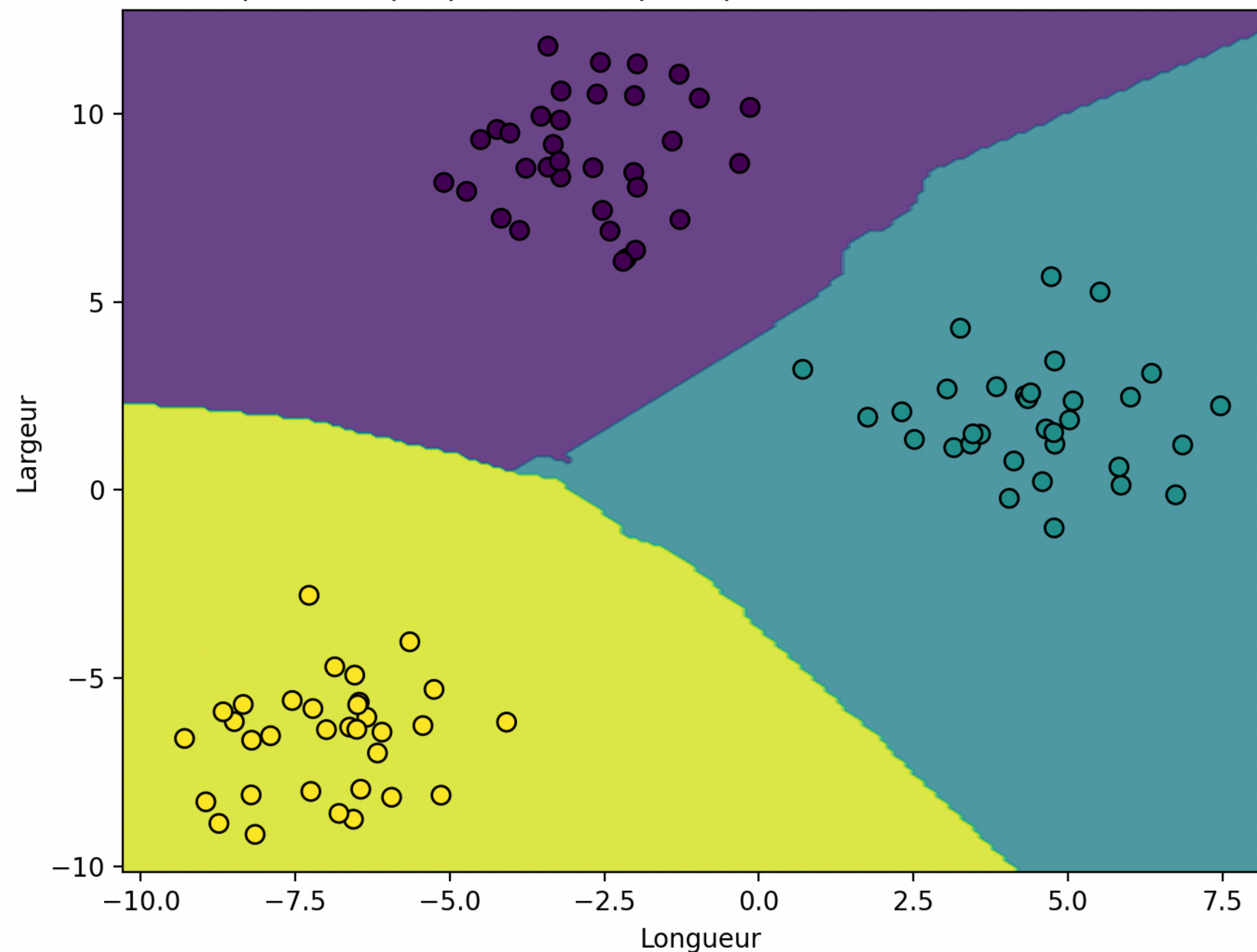


On effectue un vote majoritaire: C'est vert !



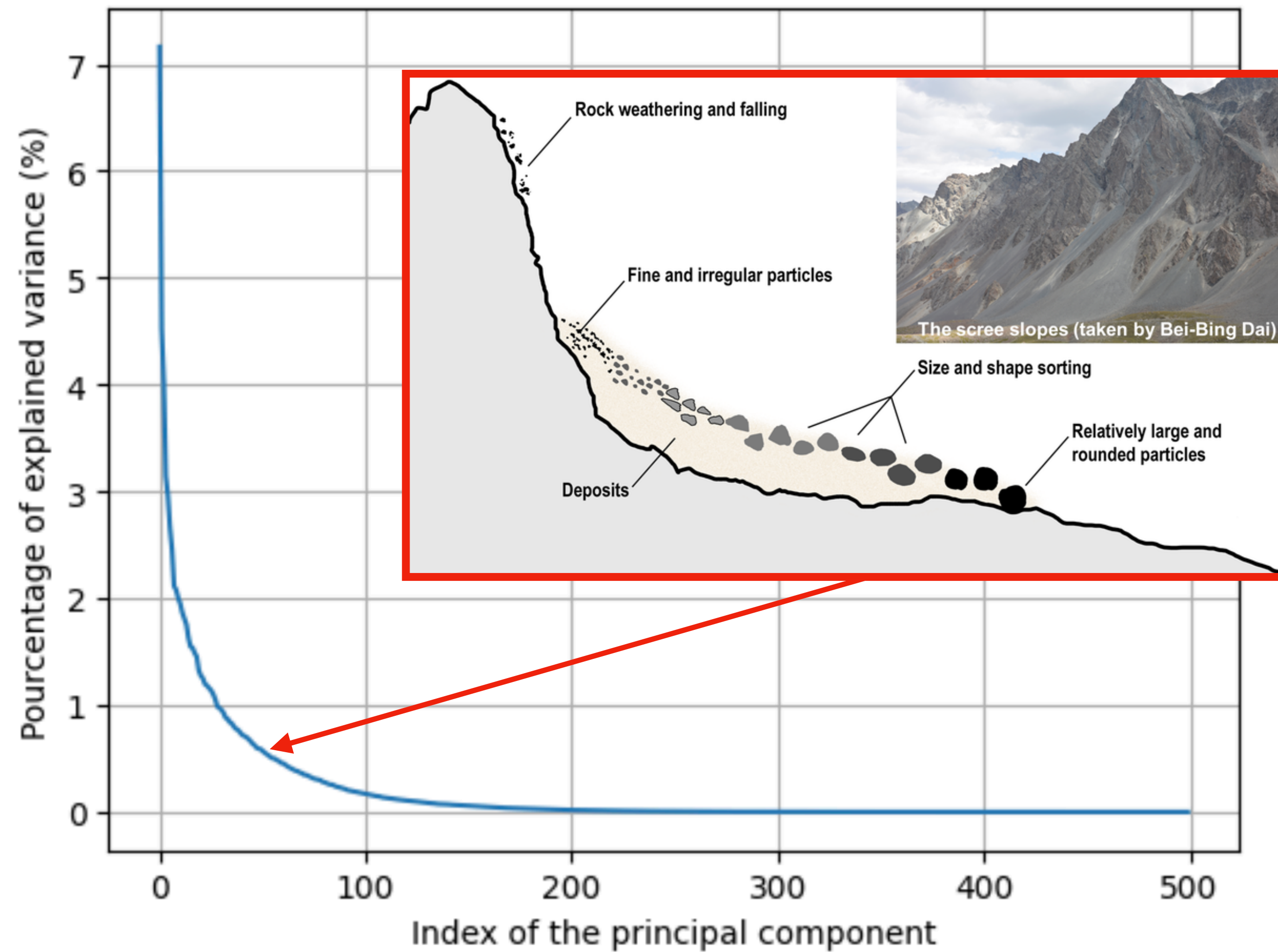
Ce modèle est appelé: K Nearest Neighbors

On fait ceci pour chaque point de l'espace pour trouver la ``frontière de décision``



Comment ce modèle peut-il être impacté par la grande dimension ?

Comment choisit-on le nombre de composantes principales k ?

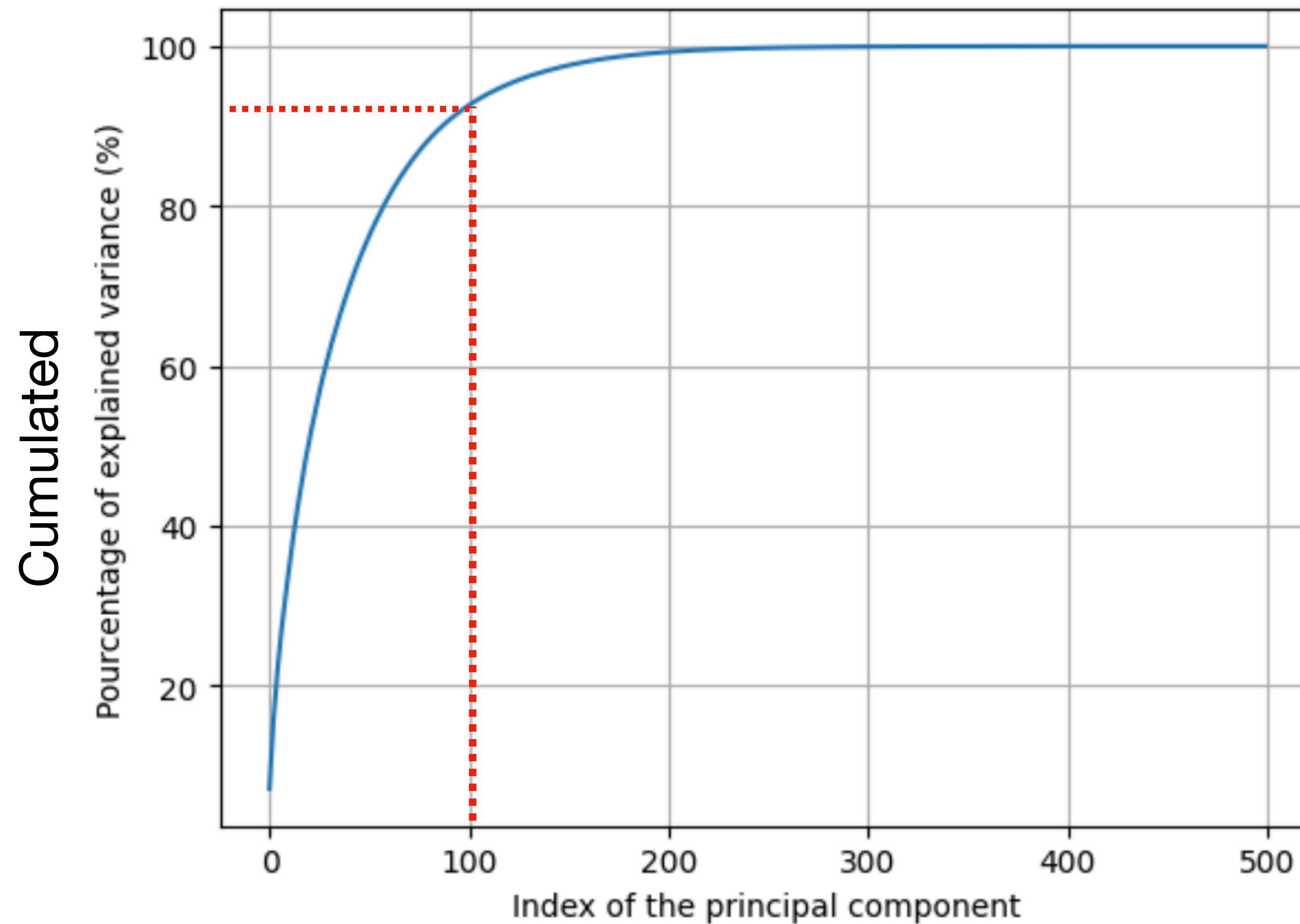


“**scree**” plot

On identifie le point de changement de pente

Aussi appelée: “la règle du coude”

Il est plus intuitif de visualiser la variance cumulée:



Ici les **100** premiers axes principaux expliquent **93%** de la variance