

Analyse en Composantes Principales (ACP)

Principal Component Analysis (PCA)

Réduction de dimension . bezaks 2.

x_1	x_2	\dots	x_p
x_1^1	x_2^1		x_1^p
x_1^2	x_2^2		x_2^p
x_1^3	x_2^3	\dots	x_i^p
\vdots	\vdots	\vdots	\vdots
x_1^N	x_2^N		x_N^p

N lignes

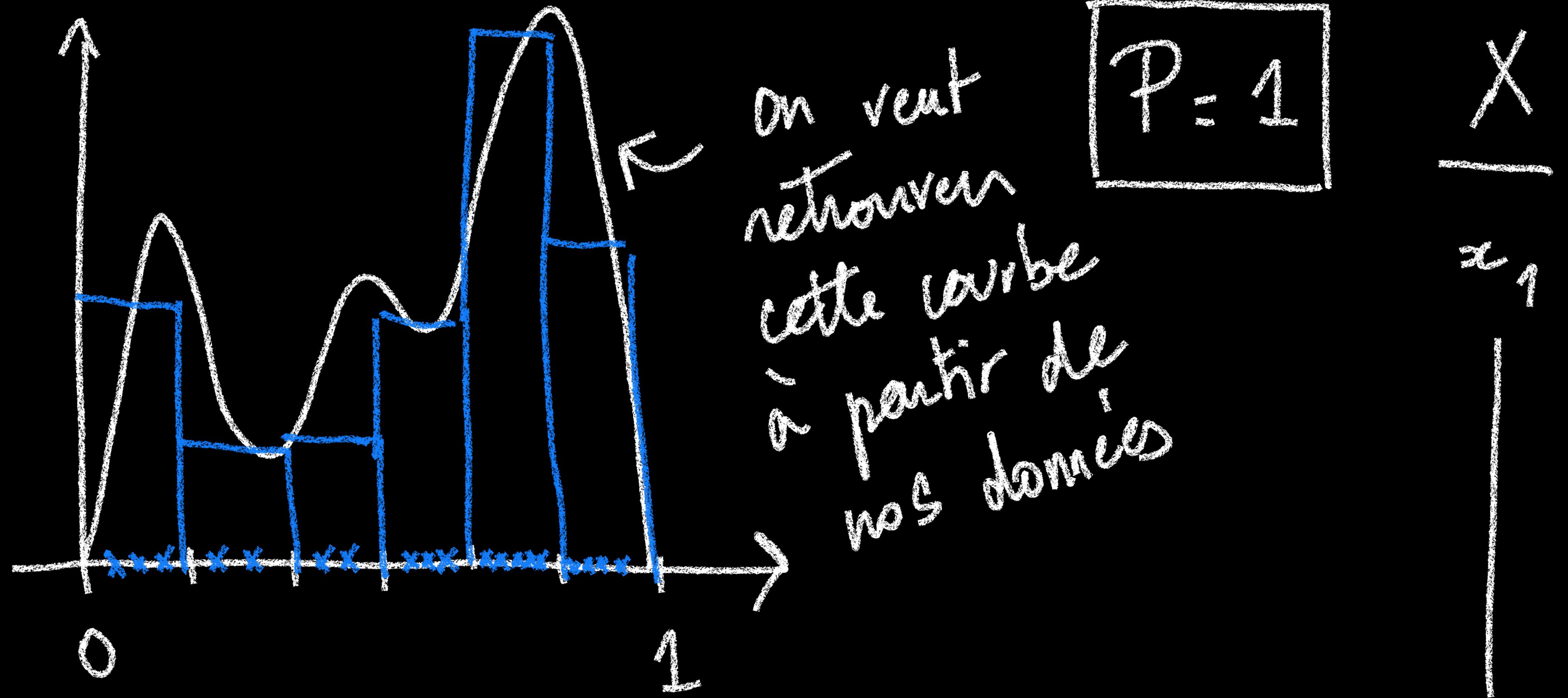
p données

① Fléau de la
Dimension

② Visualiser nos
données.

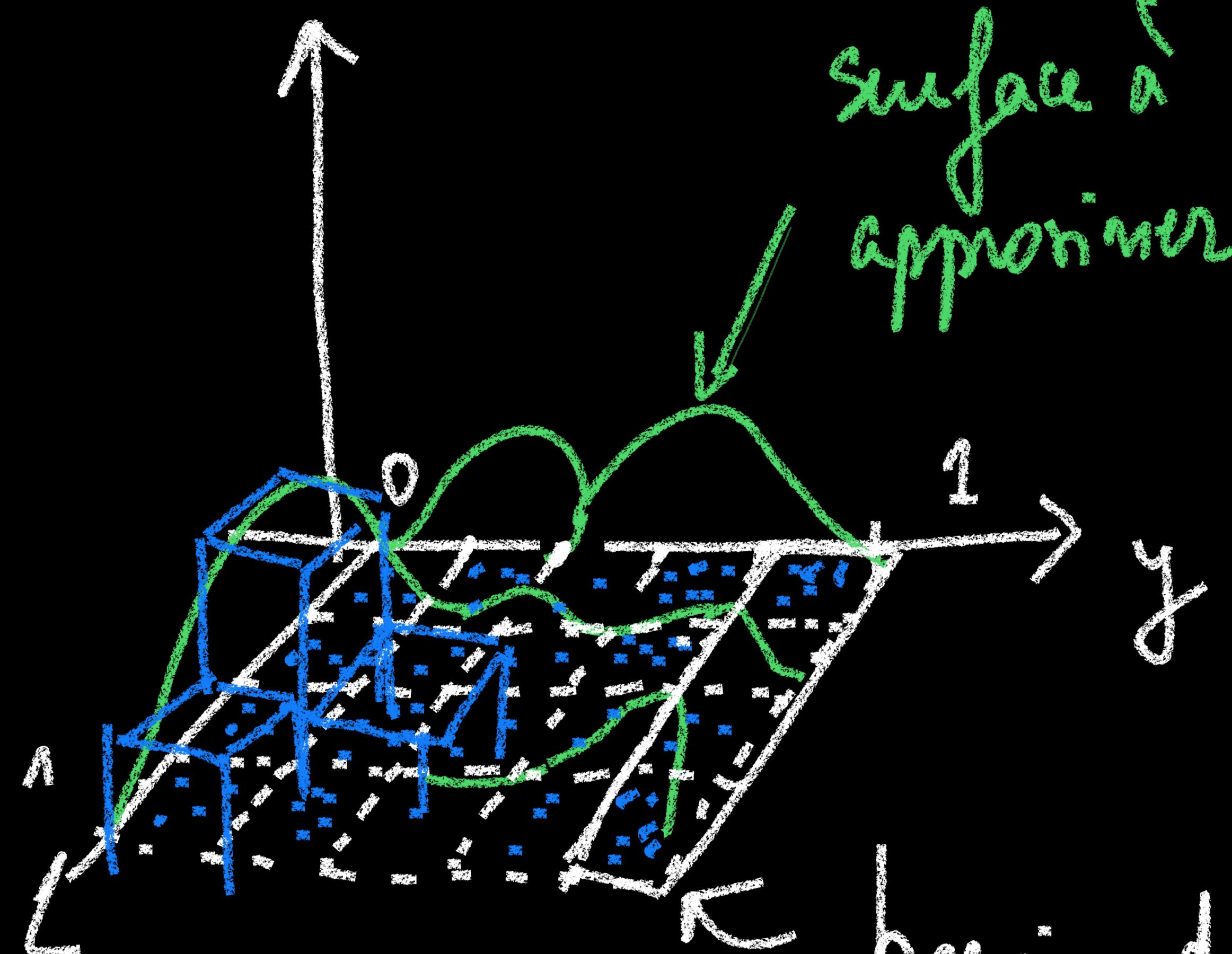
① Fleau de la dimension (= nbre de colonnes
= nbre de features)

(N, p) : Pour utiliser de manière efficace des algorithmes d'analyse de données ou de Machine Learning, le besoin en nbre de lignes N doit varier exponentiellement en le nbre de dimensions p .



$[0, 1]$ découpé en 10 intervalles
⇒ besoin de 100 points $\Rightarrow N \geq 100$

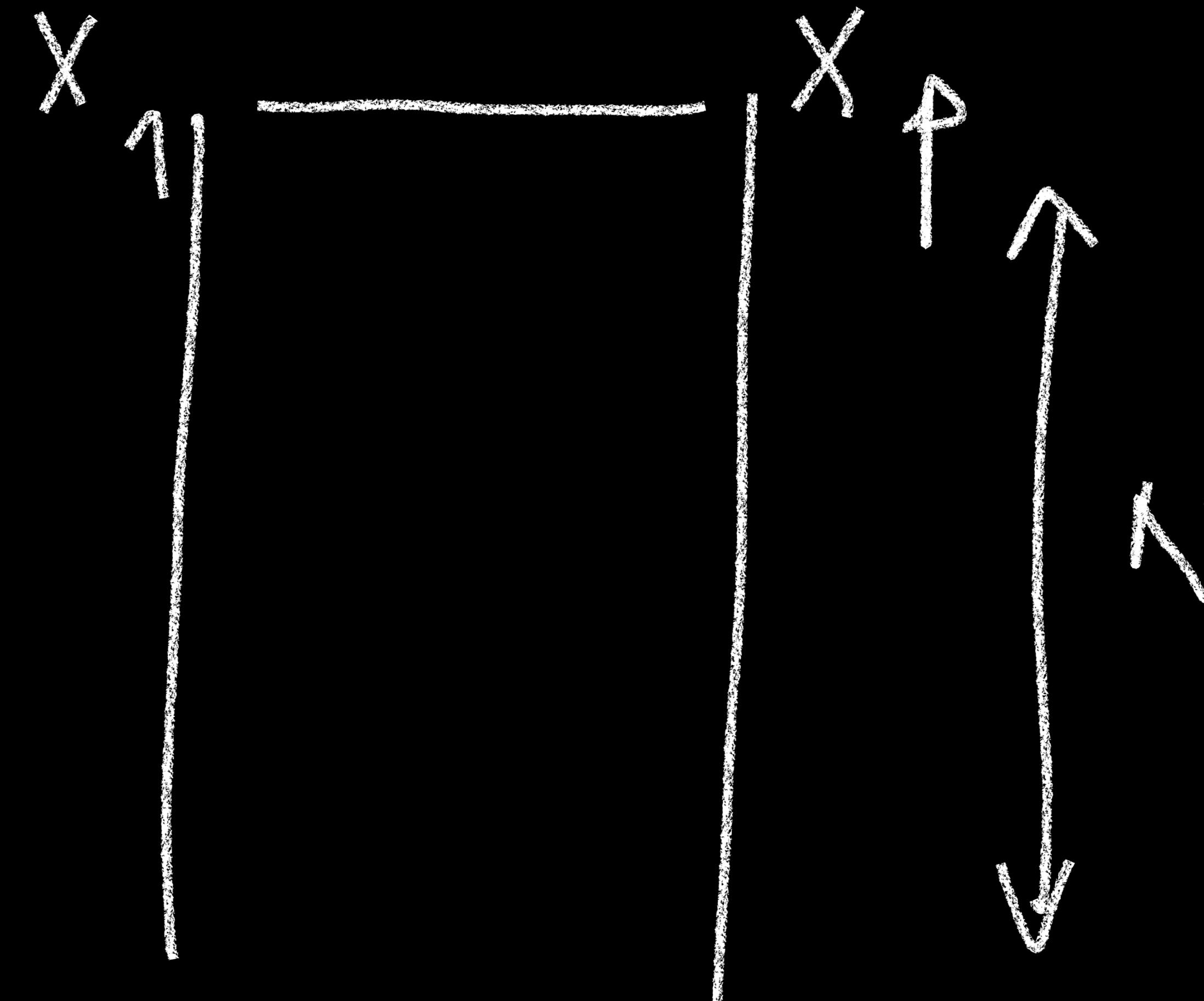
$$P = 2$$



Besoin de 10^3 points
Besoin de $N \geq 10^3$ points

	x	y
	x_1	y_1
	x_N	y_N

$$P = 30$$



Surface Prix unitaire Temps ...

34

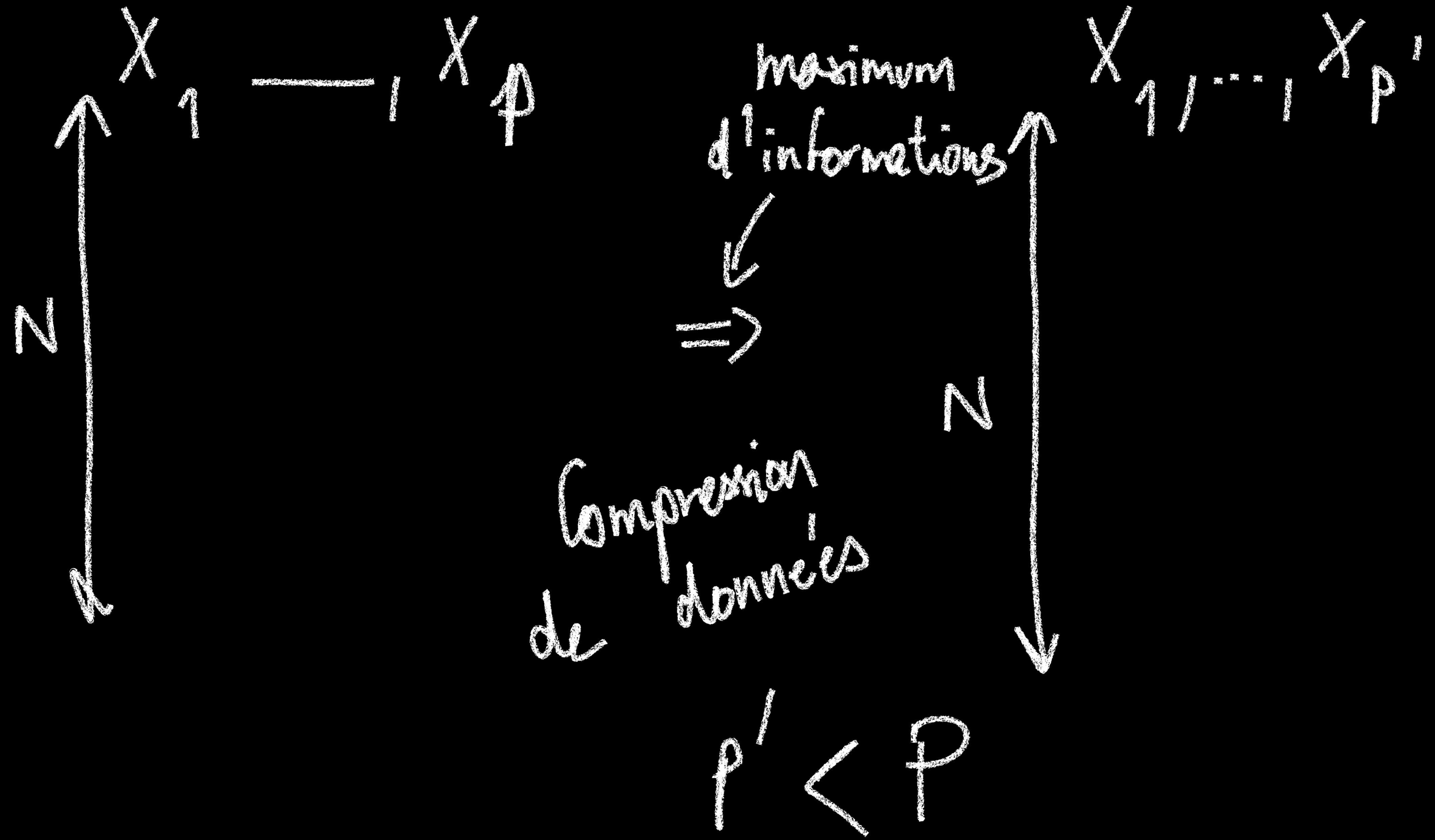
$N \geq 10^3$ points

1 000 000 000

000 000 000

000 000 000

000 0 points



② Visualiser un dataset avec $P \geq 4$ variables.

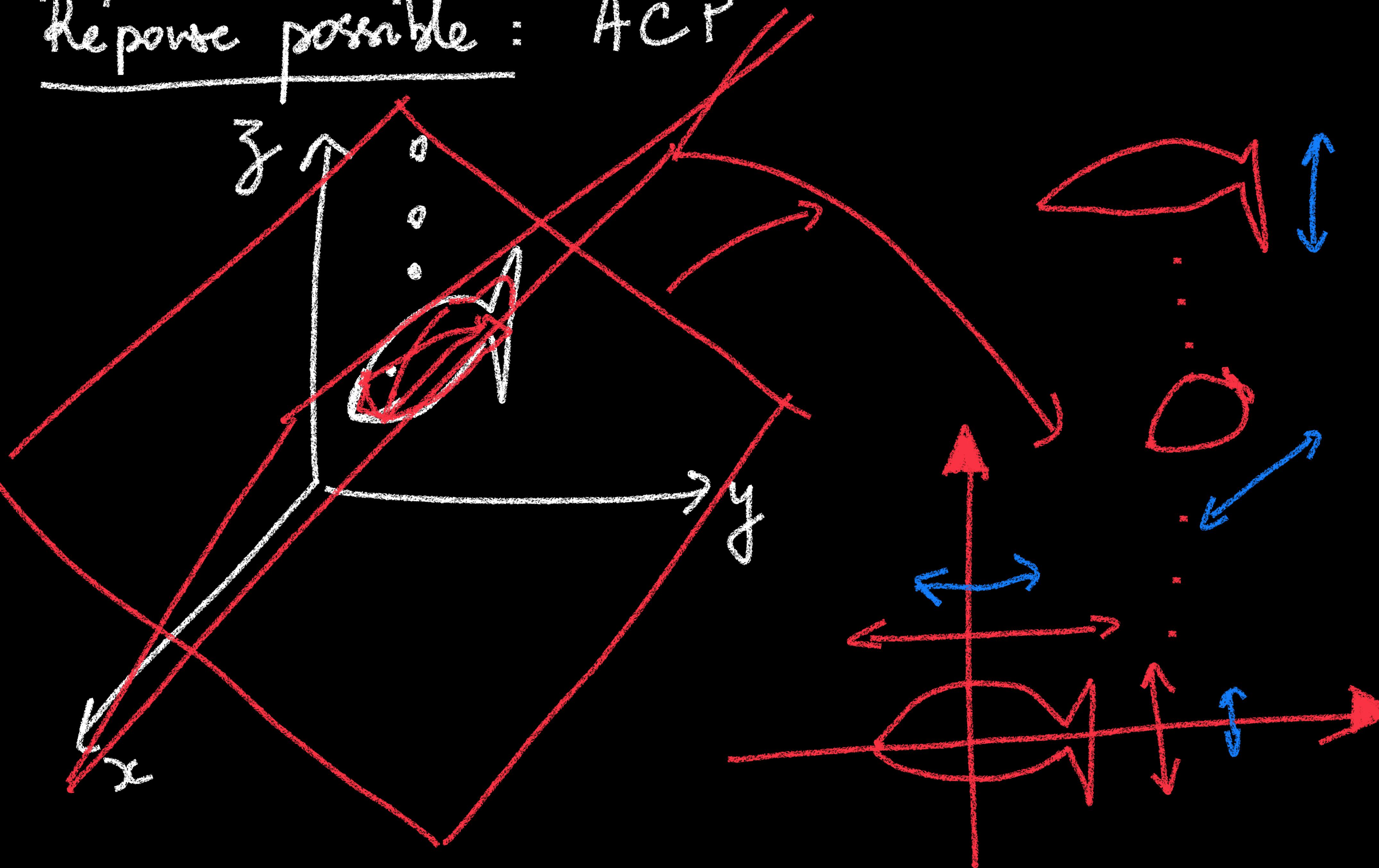
1D

2D

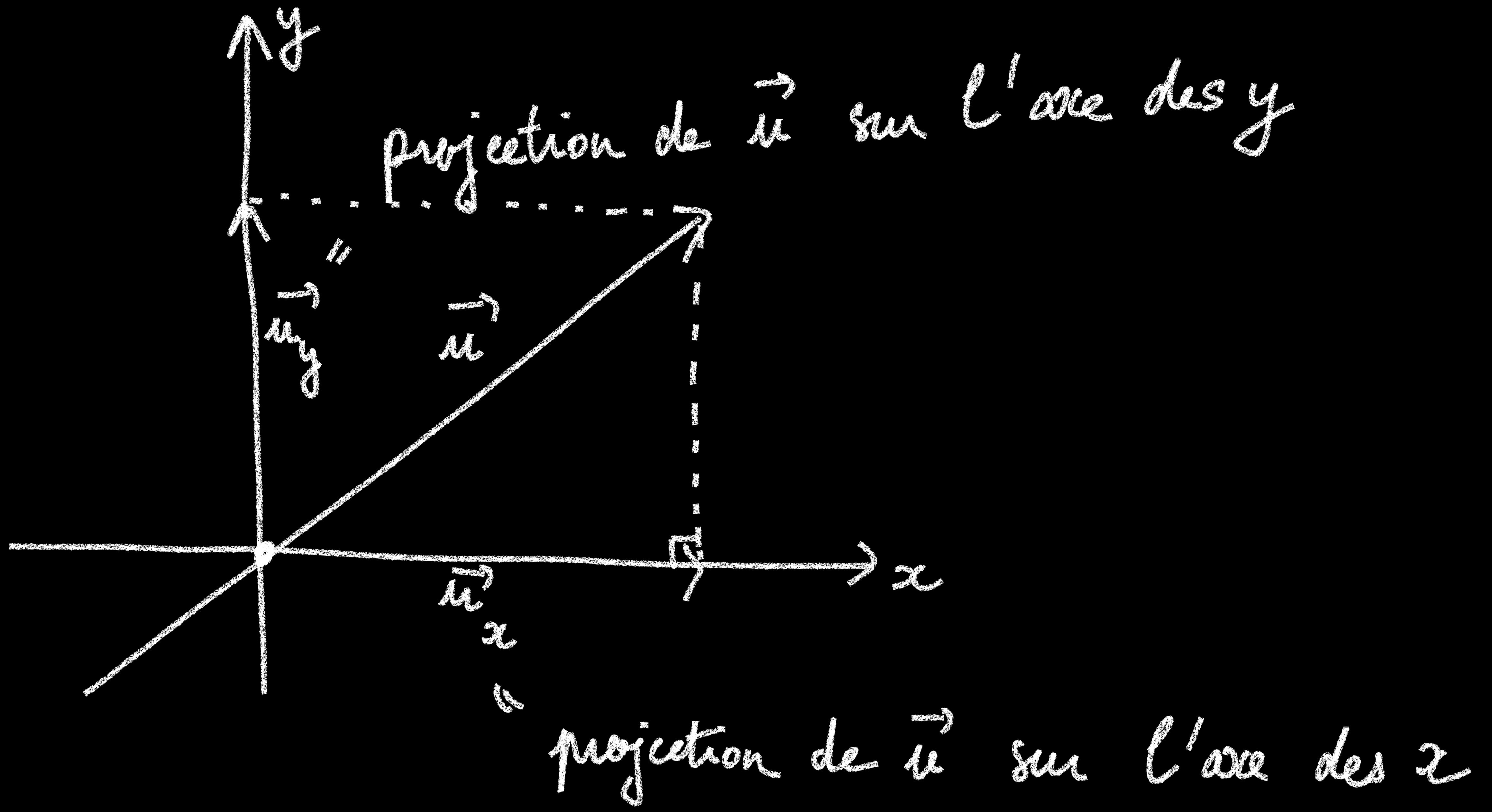
3D

So dimensions

Reponse possible : ACP



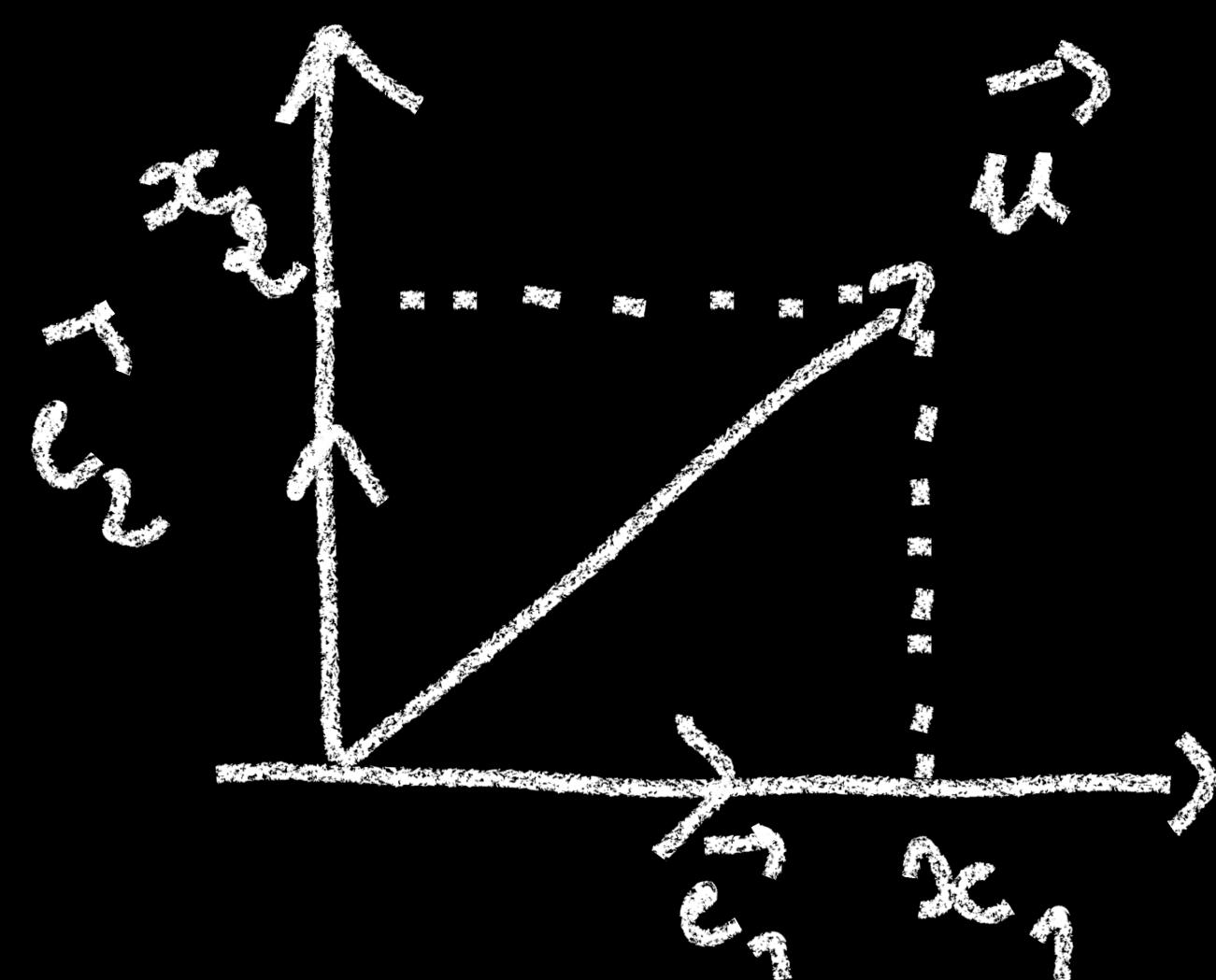
L'A.C.P. est un algo. qui cherche des axes de projection tels que la projection de notre dataset sur les axes ait une variance maximale. On appelle ces axes des composantes principales.



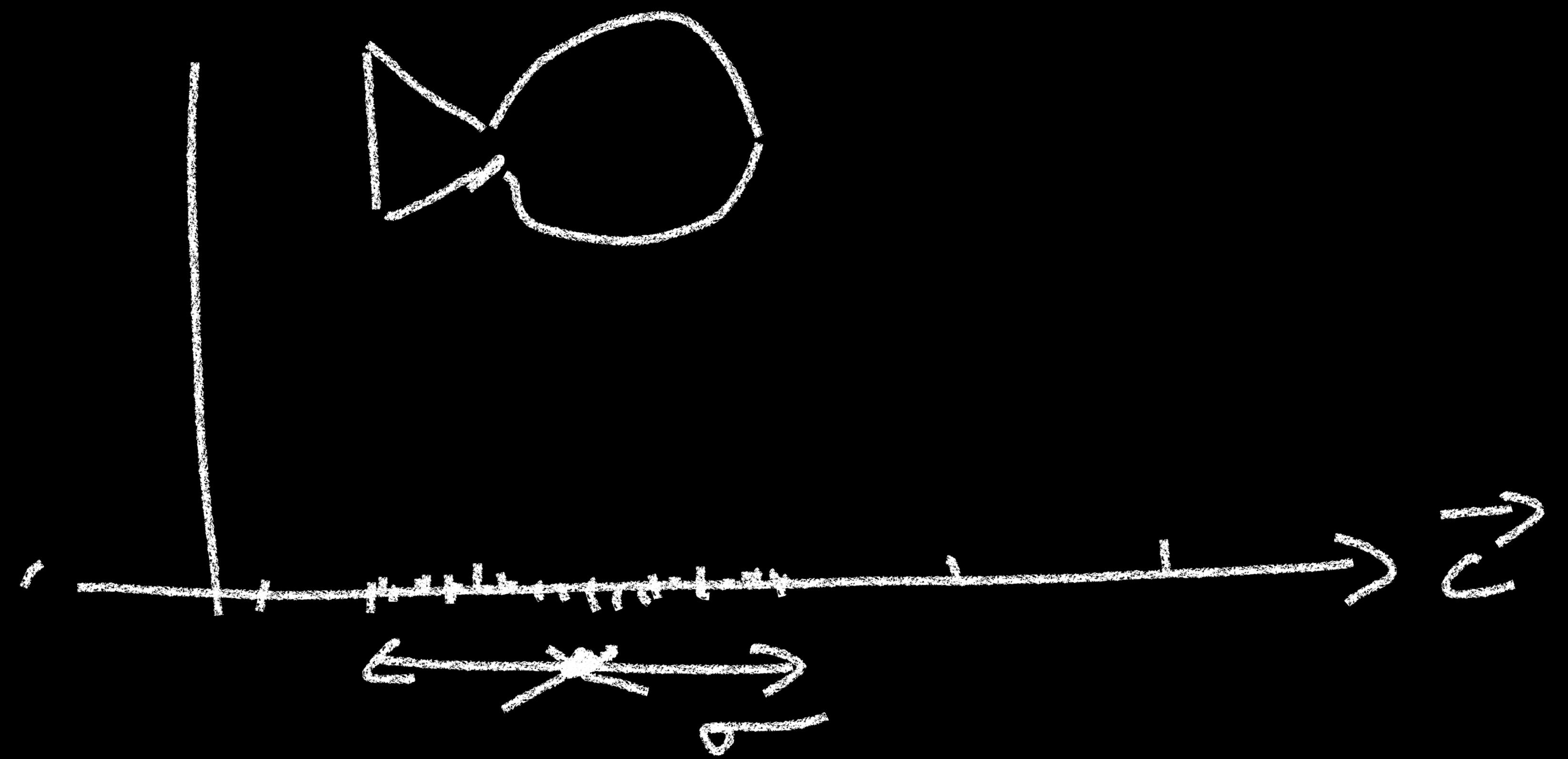
$x_1 \dots x_p$

$$\vec{u} = x_1 \cdot \vec{e}_1 + x_2 \cdot \vec{e}_2 + x_3 \cdot \vec{e}_3 + \dots$$

$$\begin{aligned} \text{Proj}(\vec{c}) \\ \langle \vec{u}_1, \vec{c} \rangle \\ \langle \vec{u}_2, \vec{c} \rangle \\ \vdots \\ \langle \vec{u}_p, \vec{c} \rangle \end{aligned}$$



$$\langle \vec{u}, \vec{c} \rangle = x_1 \cdot \langle \vec{e}_1, \vec{c} \rangle + x_2 \cdot \langle \vec{e}_2, \vec{c} \rangle + \dots + x_p \cdot \langle \vec{e}_p, \vec{c} \rangle$$



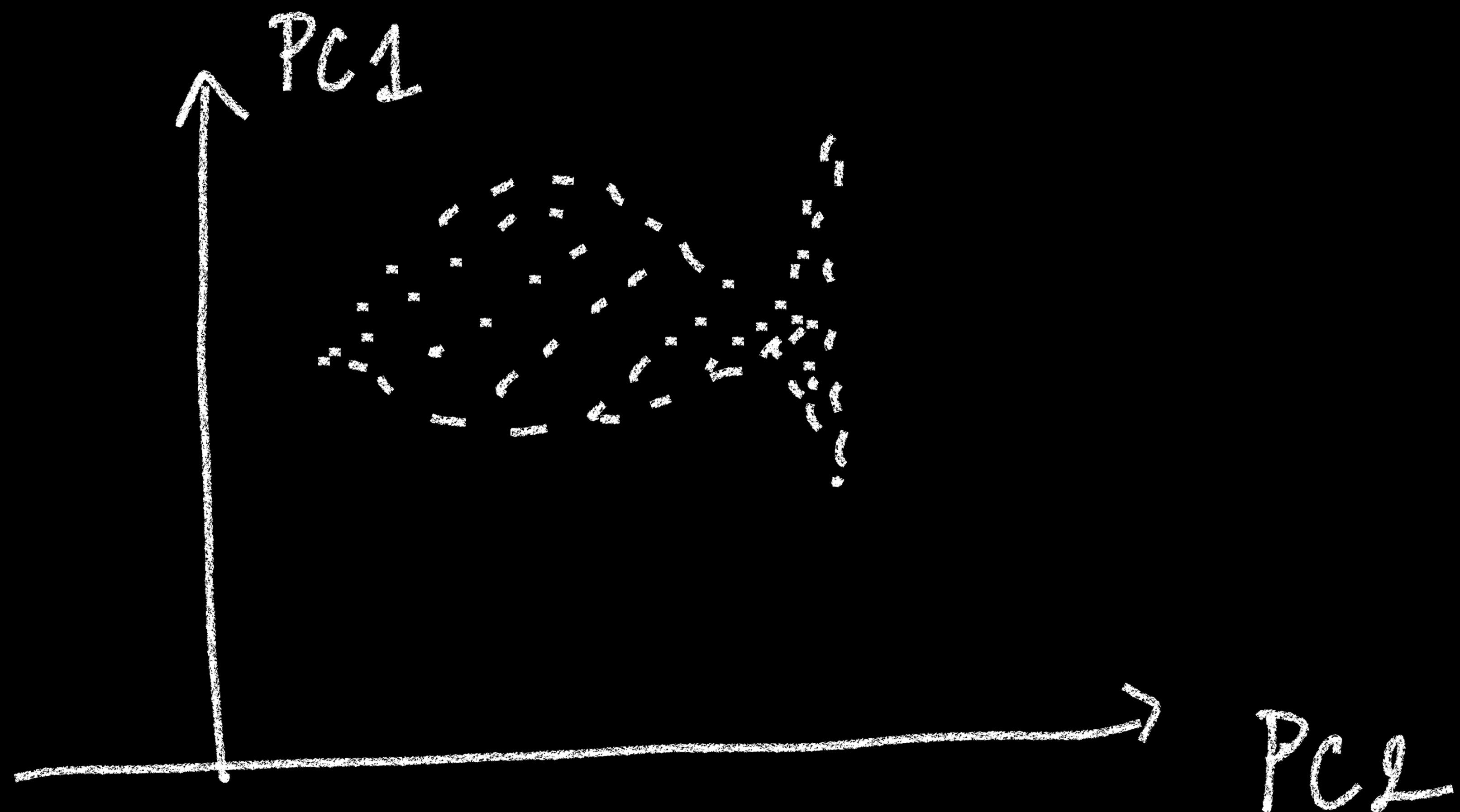
① On cherche l'axe telle que le proj.
de notre dataset sur cet axe ait une
variance max.

PC 1

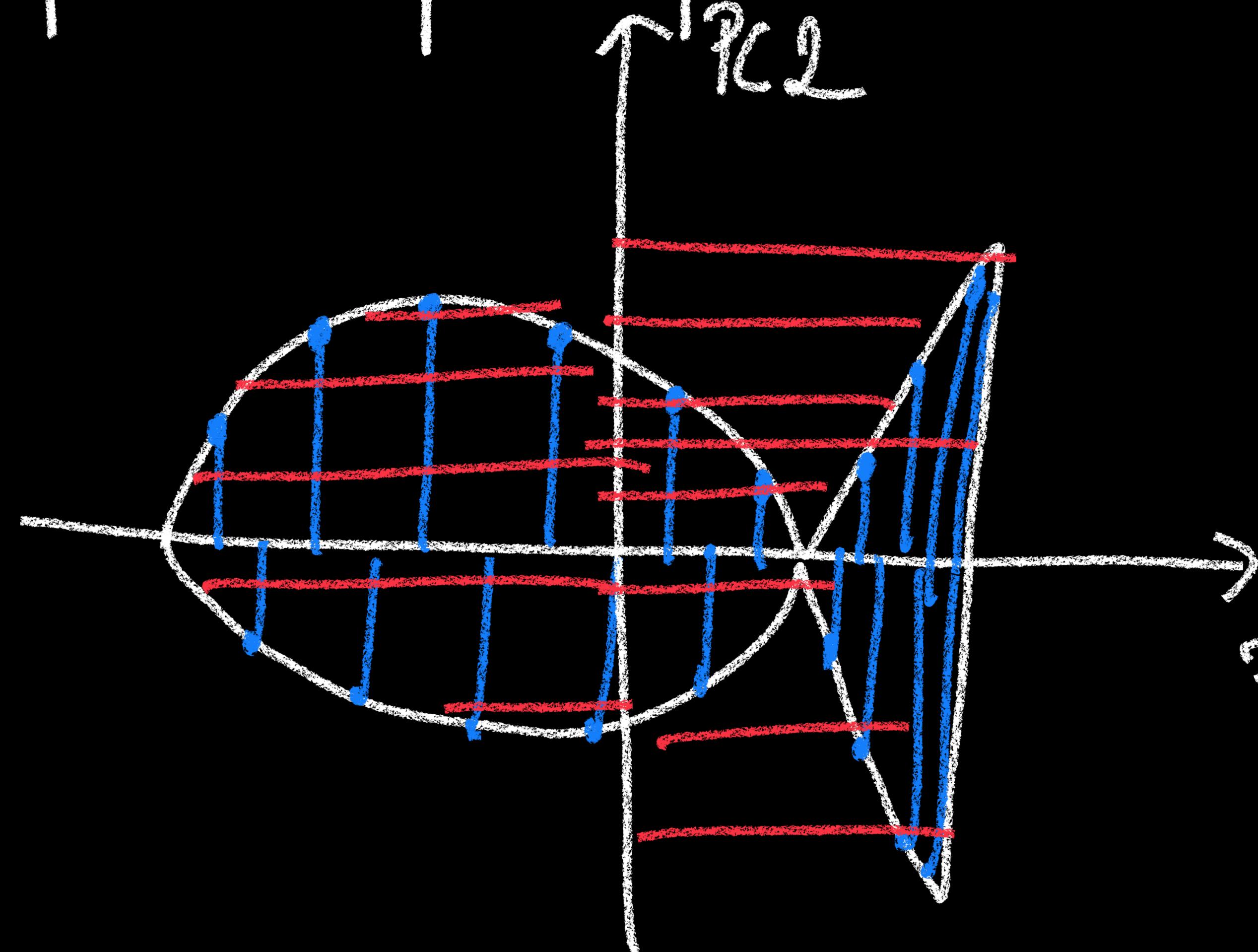
② On cherche l'axe avec la 2^e variane
la + élevée \rightsquigarrow PC2

③ ...

Visualisation ? PC 1 / PC 2



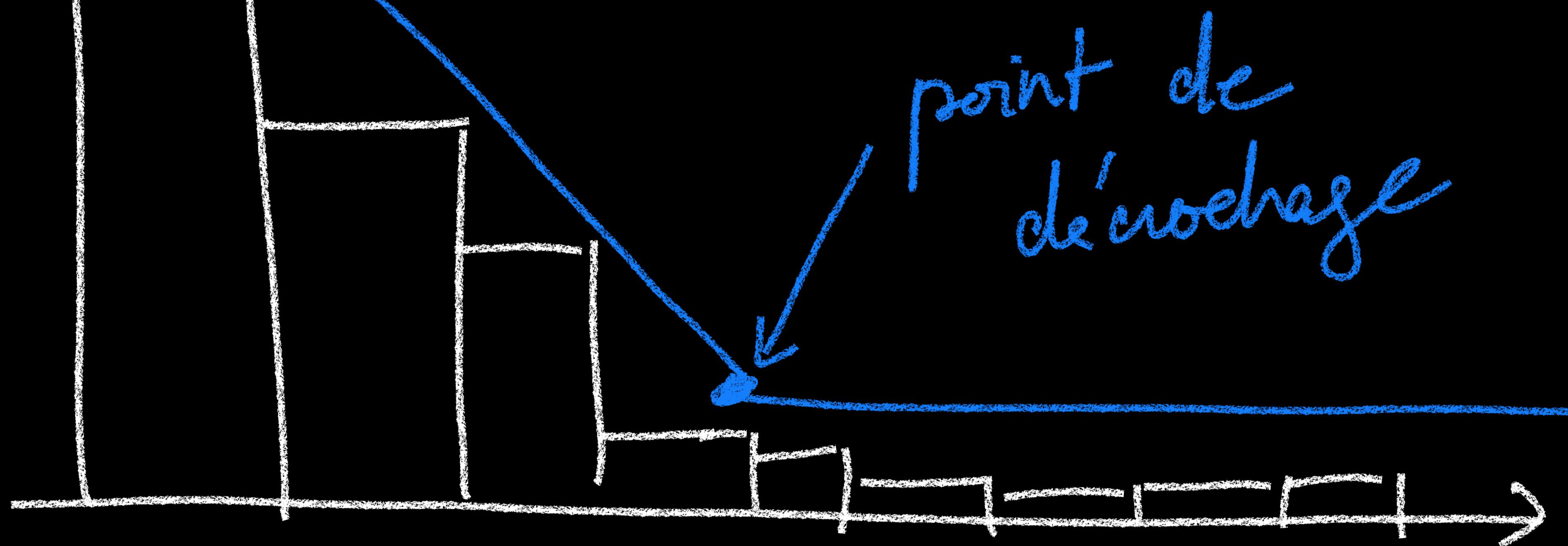
Taux de variance expliquée par chaque
composante principale.



$$1 - \frac{\text{dist entre pts et PC1}}{\text{dist entre pts et le moyen}}$$

$$\text{PC1} = \% \text{ de Variance expliquée par PC1}$$

↑ Axe de variance
expliquée



PC1 PC2 PC3 PC4

... - - -

$$\begin{matrix} X_1 & \dots & X_p \\ \uparrow & & \downarrow \\ N & 0 & 0 \end{matrix}$$

$$\begin{matrix} \rightarrow & \uparrow & \downarrow \\ & 0 & 0 & 0 & 0 & 0 & \leftarrow \text{pas} \\ & \uparrow & \downarrow & & & & \\ P_{C1} & & & & & & \\ P_{C4} & & & & & & \end{matrix}$$

Perte d'interprétabilité

d'interprétation
daine

$$\text{PC 1} = \underbrace{0,75}_{+ 3,4} \times \text{prix} \\ + 3,4 \times \text{surface} \\ - 2,1 \times \text{population} \\ + \dots$$

$$x_1 - x_p$$

↑
n
↓

taille
(cm)

taille
(m)

$$\text{PC 1} = \underbrace{(0,1)}_{(\text{m})} \text{ taille} \\ + \underbrace{(0,5)}_{(\text{cm})} \text{ taille}$$

+ ...

Standardisation

$$X \sim (\mu, \sigma)$$
$$\frac{X-\mu}{\sigma} \rightarrow (0, 1)$$

$$PC1 = x_1 X_1$$

$$+ x_2 X_2$$

+

...

$$+ x_p X_p$$

PC1

x_5 mix

x_{11} surface

x_2

...

