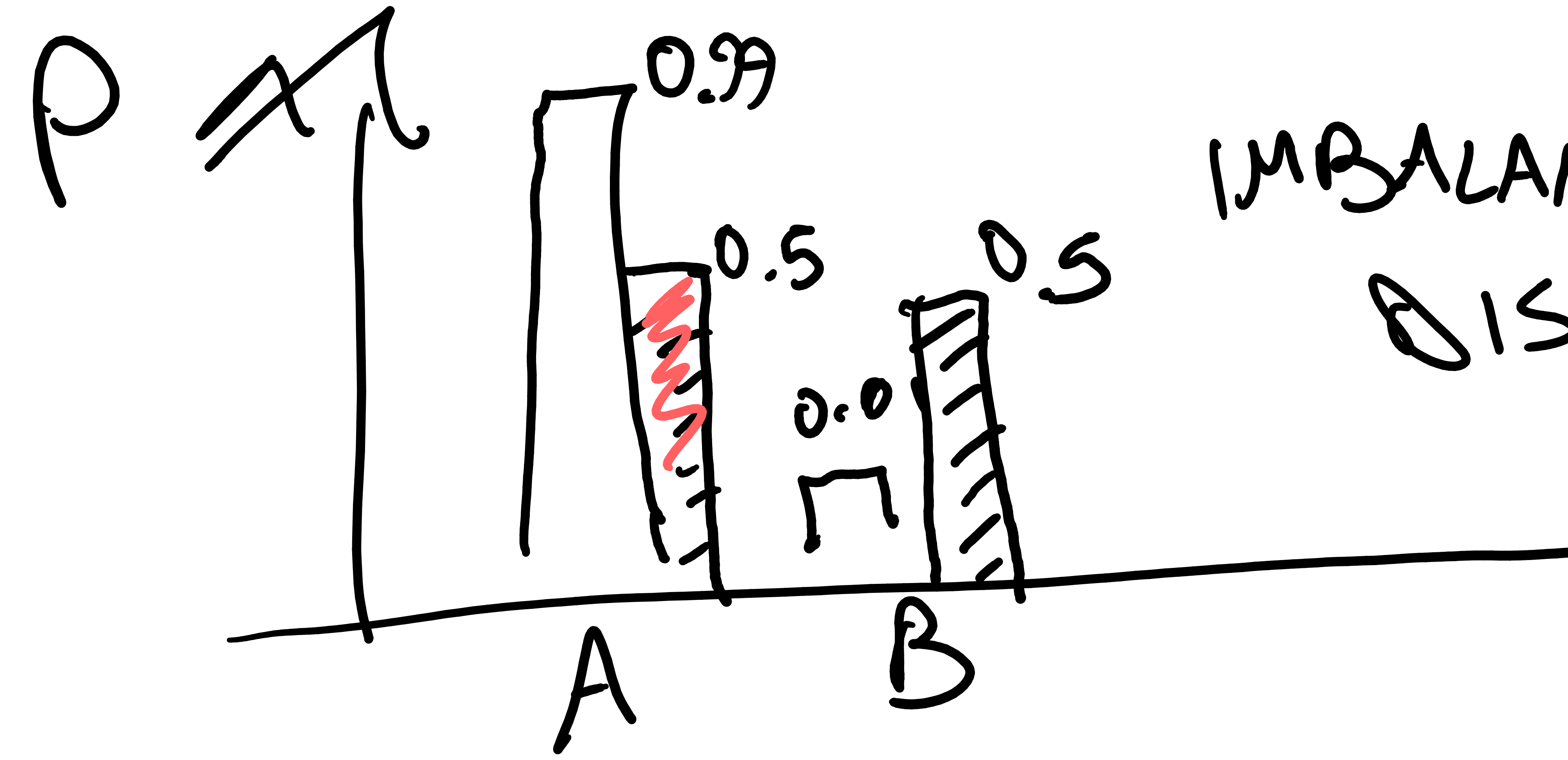
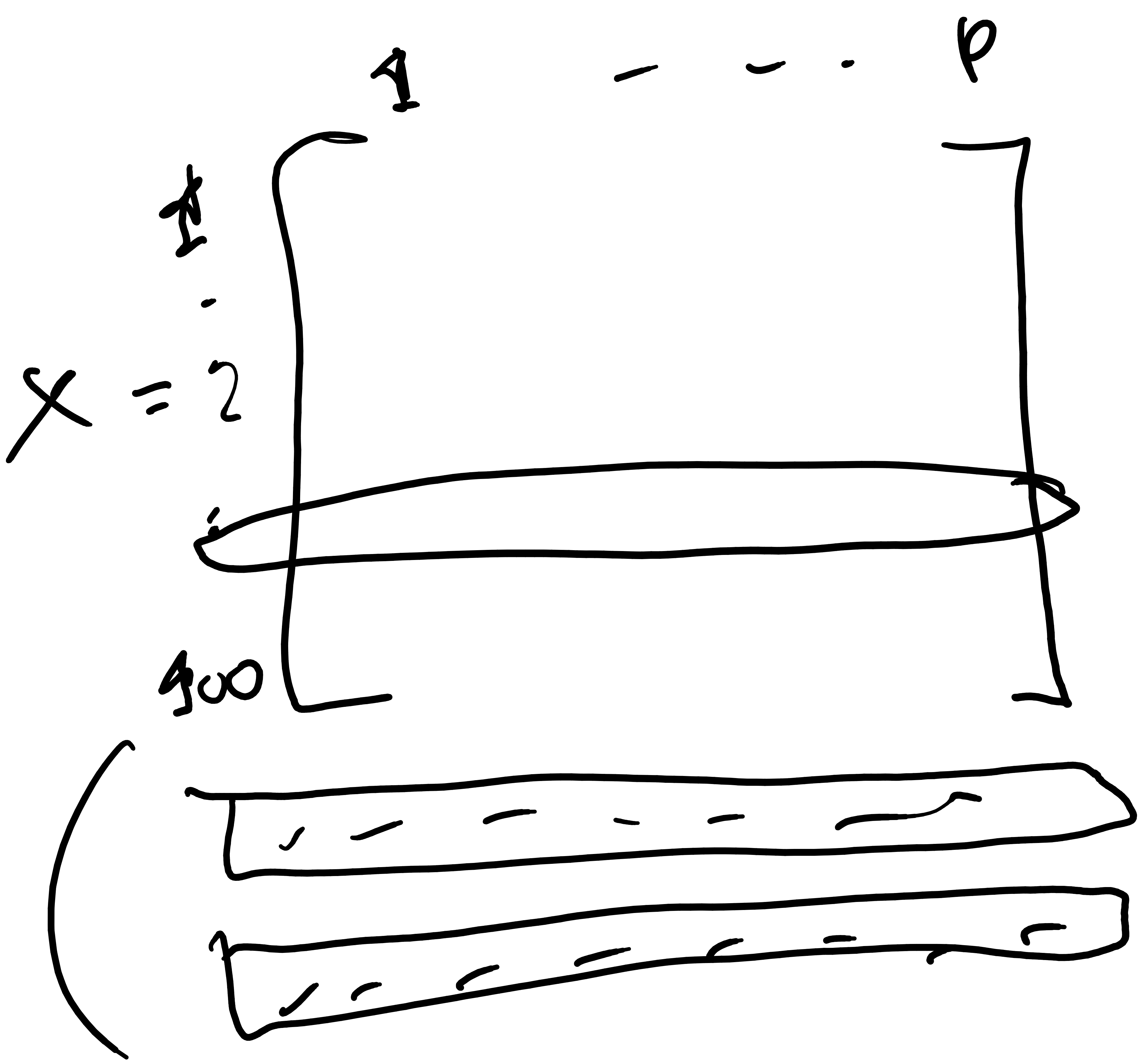
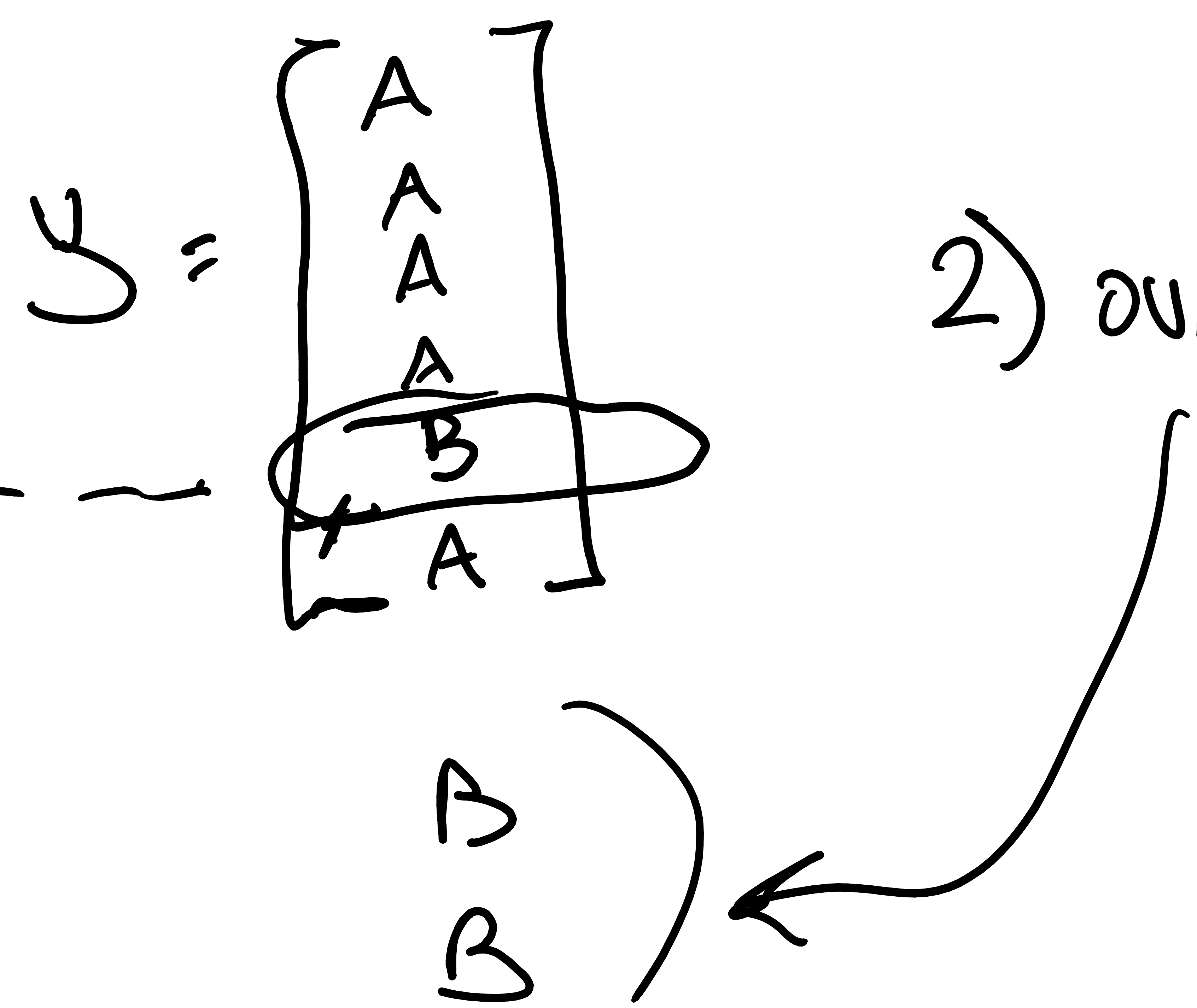
ML FUND. DAY 3

$$y = \begin{cases} A \\ B \end{cases}$$



IMBALANCED DIST.





1) UNDERSAMPLING

2) OVERSAMPLING

ML

(RL)

SUPERVISED

y — $\mathbb{R}$

$\{A, \ldots, Z\}$

$\hat{f}(X) = \hat{y}$

UNSUPERVISED

$\cancel{y}$

$\boxed{X}$

SELF-SUPERVISED

? y ?

CURSE OF DIMENSIONALITY

$$X = \begin{array}{c} \\ 1 \\ \vdots \\ N \end{array} \begin{array}{c} 1 \; \cdots \; p \\ \left[\begin{array}{ccc} & & \\ & & \\ & & \end{array} \right] \end{array} \qquad \cancel{y = [\;]}$$

PROBLEMS
- COMPUTATIONAL BURDEN
- MULTICOLLINEARITY
- GEOMETRY

PCA (SCALED X)

EIGENVALUES
EIGENVECTORS

$$(\lambda_1, u_1)$$
$$\vdots$$
$$(\lambda_p, u_p)$$

$$X_{N\times p}$$

$$X_{N\times p} \longrightarrow Z_{N\times l}$$

$$l \ll p$$

$\Sigma_{p\times p}$ VAR-COV MATRIX of X

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \sigma_{1S} \\ & \ddots & \\ & & \sigma_p^2 \end{bmatrix}$$

$$\alpha = \begin{bmatrix} \vdots \\ \vdots \end{bmatrix}$$

$$X = \begin{array}{c} \begin{array}{ccc} 1 & \cdots & p \end{array} \\ \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \end{array} \cdot \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \alpha_1 x_{11} + \alpha_2 x_{12} + \cdots + \alpha_p x_{1p} \\ \vdots \\ \alpha_1 x_{N1} + \cdots + \alpha_p x_{Np} \end{bmatrix}$$

$Z_1 = X\alpha_1$ $\qquad$ $Var(Z_1) = \alpha_1^T \Sigma \alpha_1$

$$\max_{\alpha_1} \quad Var(Z_1) = \alpha_1^T \Sigma \alpha_1$$

CONST. $\|\alpha_1\| = 1$

$p \longrightarrow 1$

$Z_2 = X\alpha_2$

$$\max_{\alpha_2} \quad Var(Z_2)$$

CONS.
- $\|\alpha_2\| = 1$
- $\alpha_1^T \alpha_2 = 0$

(Σ)

SPECTRAL DECOMPOSITION

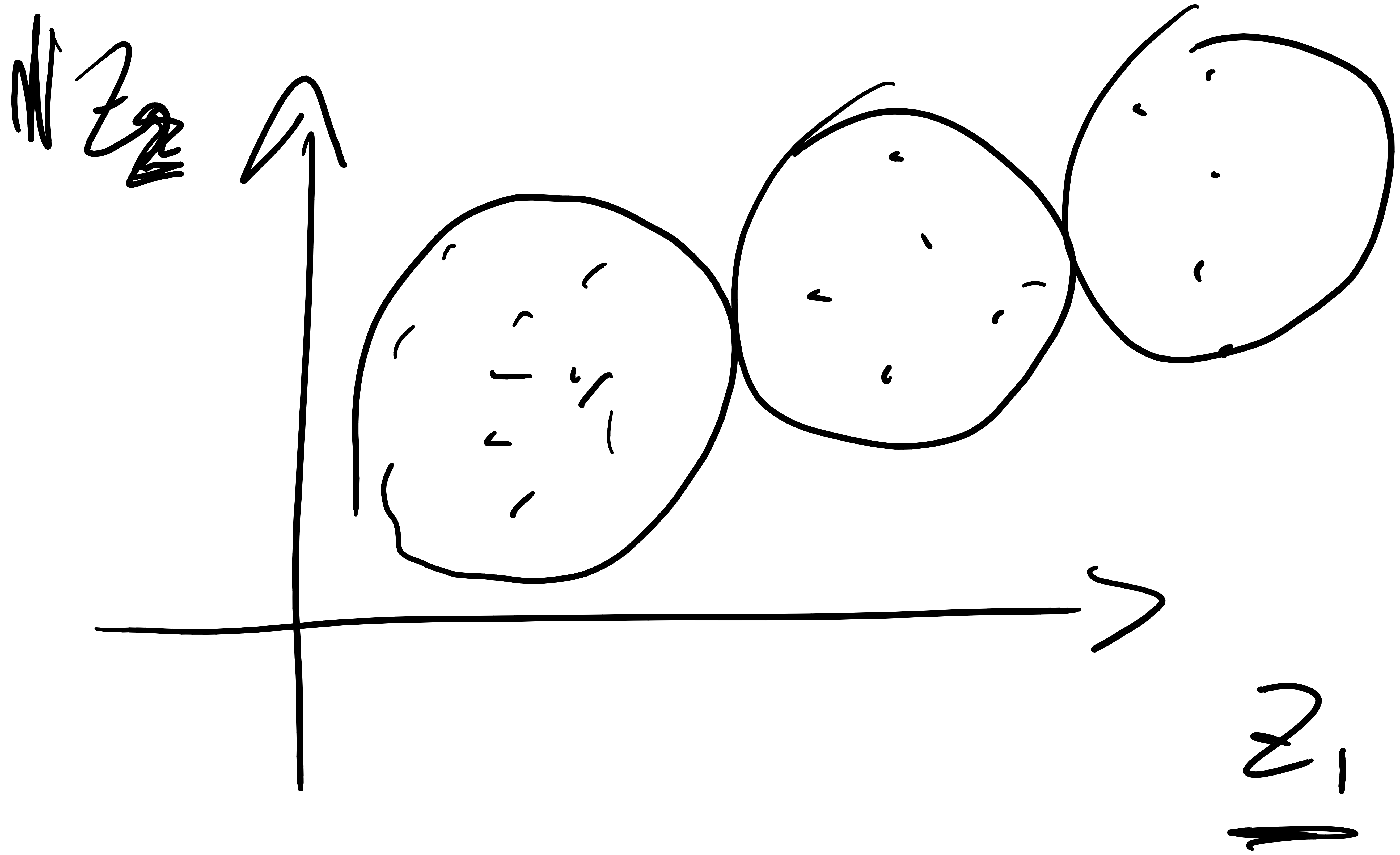
$$X_{N \times p} \longrightarrow Z = \begin{array}{c} 1 \\ \vdots \\ N \end{array} \begin{bmatrix} z_1 & z_2 & \cdots & z_\ell \end{bmatrix}$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 & 0 \\ 0 & \ddots & & 0 \\ 0 & 0 & \cdots & \sigma_\ell^2 \end{bmatrix}$$

PCA

1) FIND EIGEN(VALUES, VECTOR) FOR Σ
2) ORDER THE EIGENVECTORS BY THE VALUE OF THE EIGENVALUES
3) USE ℓ pairs (λ_i, α_i) TO PROJECT THE DATA ON NEW COMP. $(z_i = X\alpha_i)$
4) USE THE NEW MATRIX $Z = [(z_1) \cdots (z_\ell)]$ FOR YOUR NEEDS

UNS.



CLUSTERS ON PCA

SUP.

$X \rightarrow Z \qquad \hat{y} = \hat{f}(Z)$

Pipeline(StandScaler(), PCA(), RF())

Pipeline.fit(X_train, y_train)

(λ_i, α_i)

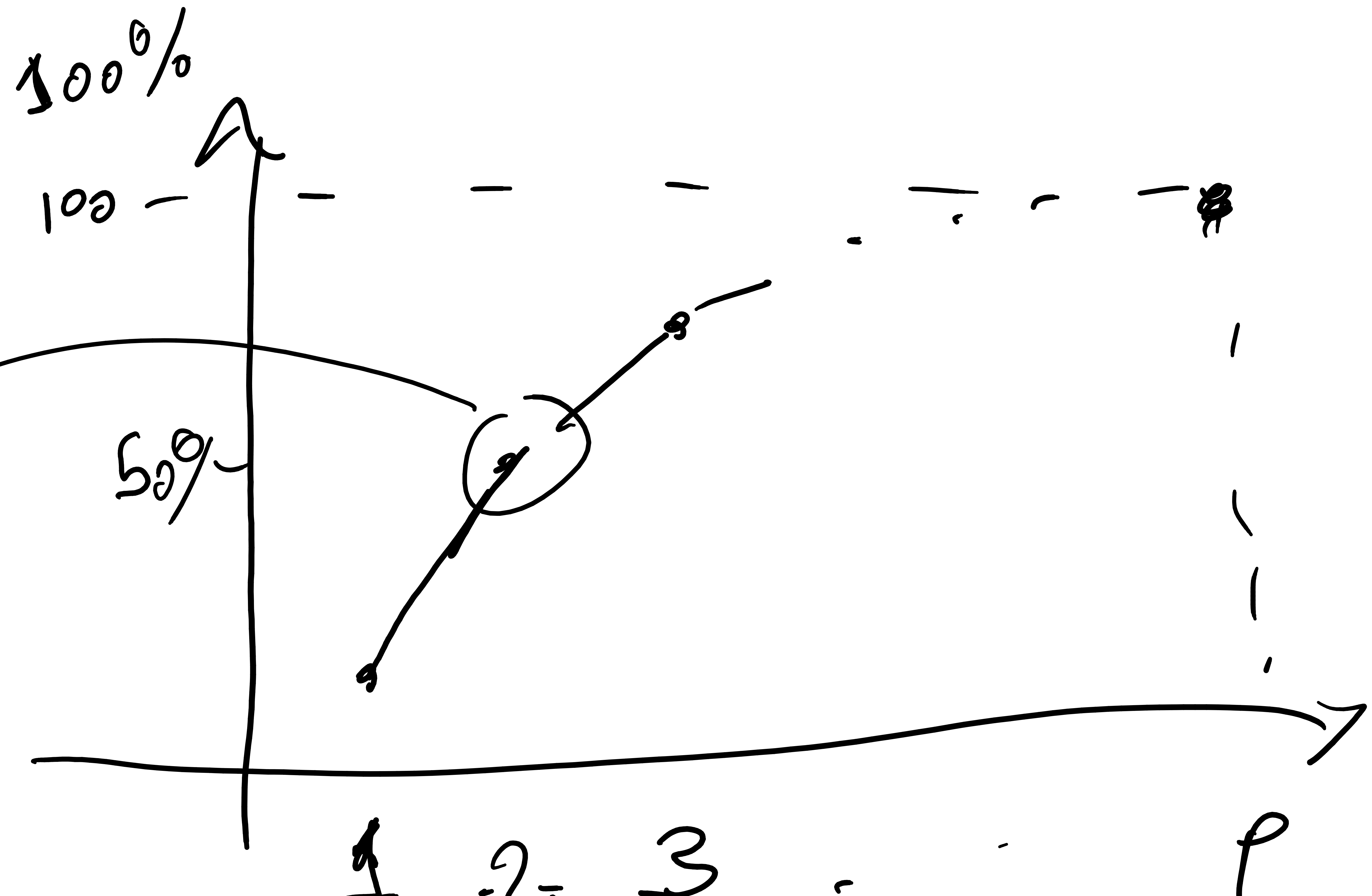
VALUE VECTOR

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} = \text{FRACT. OF EXP. VARIANCE}$$

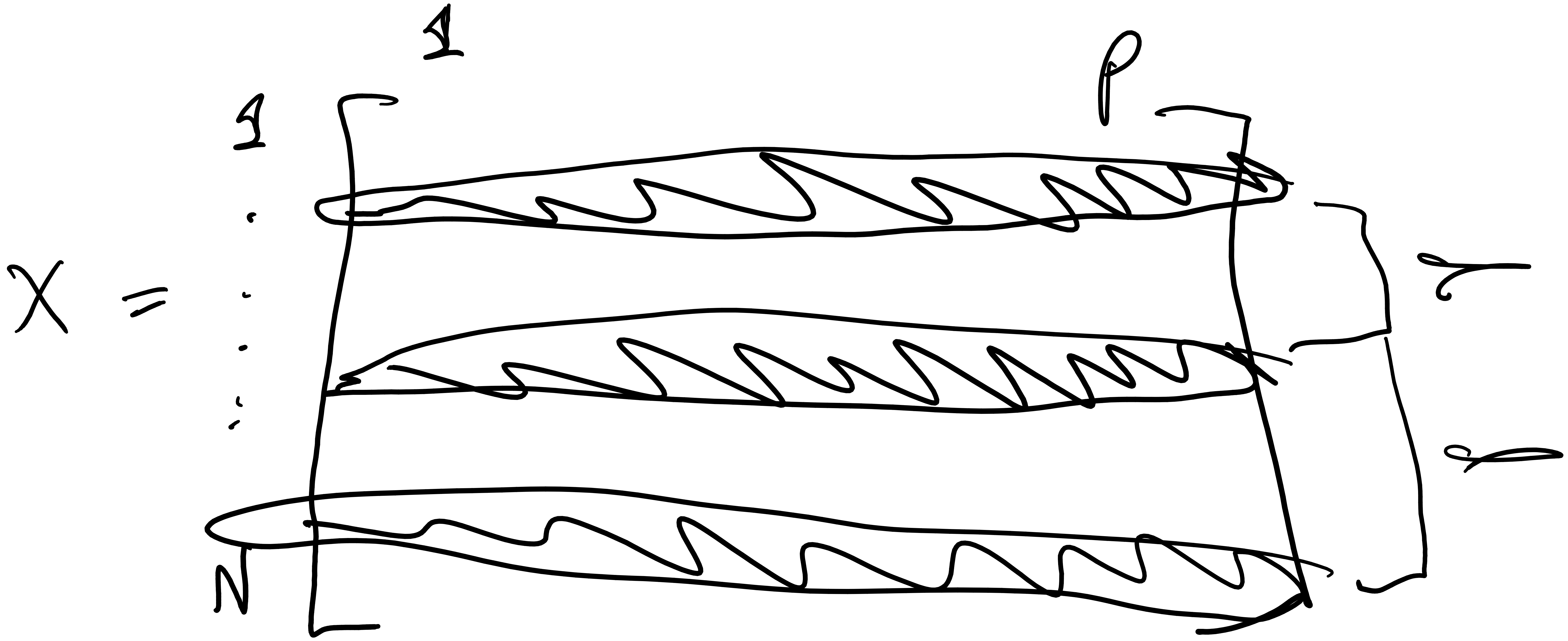
$\Downarrow$

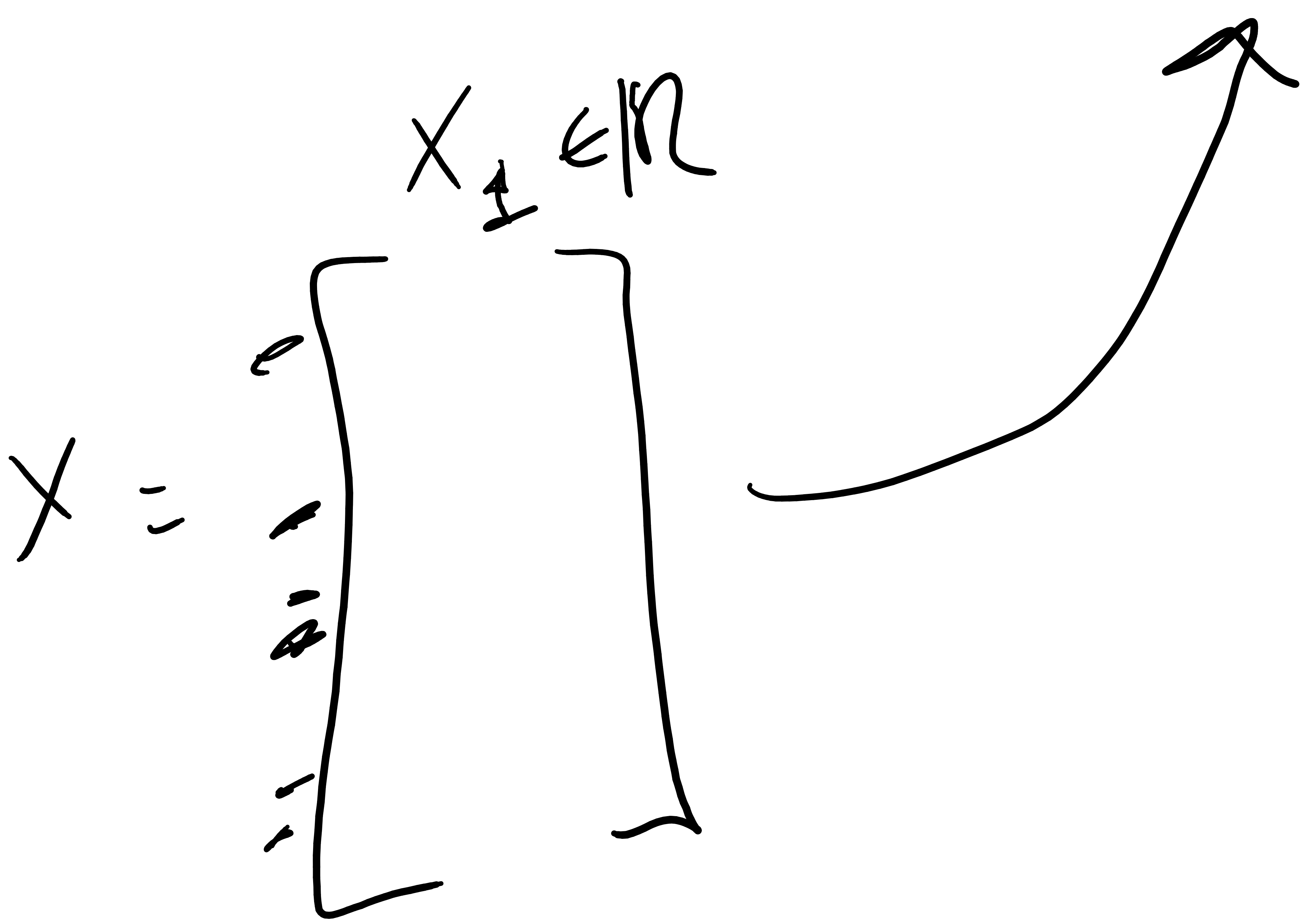
$$z_i = X\alpha_i$$

$$\frac{\lambda_1+\lambda_2}{\sum_i \lambda_i} = 0.5$$



CLUSTERING





DISTANCE

CLUSTERING

(A) NON HIERARCHICAL

- # CLUSTERS
- LOOK FOR GROUPS

HIERARCHICAL (B)

* EACH ROW IS A CLUSTER, AGGREGATE
* SINGLE CLUSTER, DIVIDE

K - MEANS

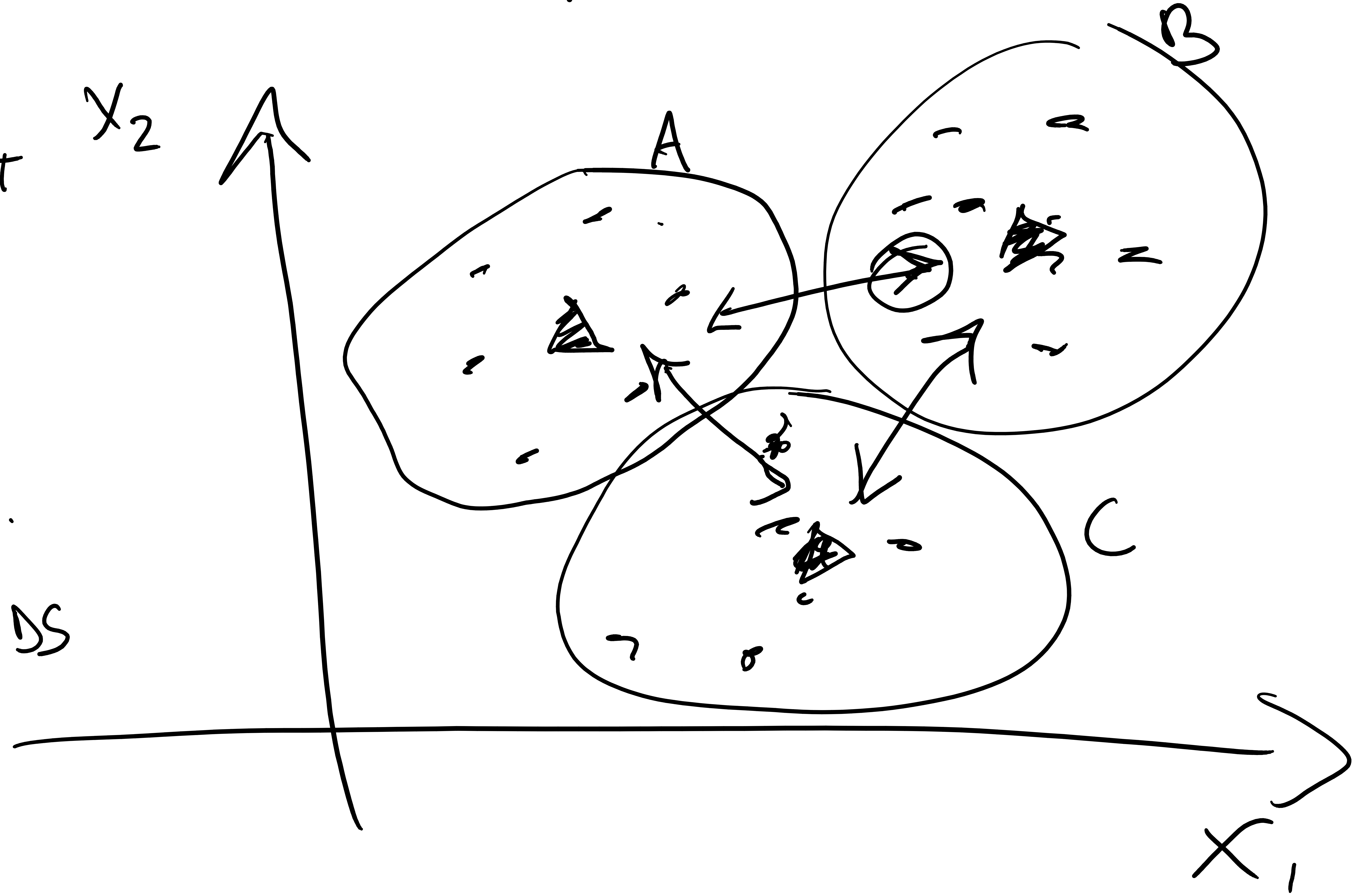
- SET A K VALUE
- CENTROIDS
- ASSIGN EACH POINT TO ONE CLUSTER

K = 3

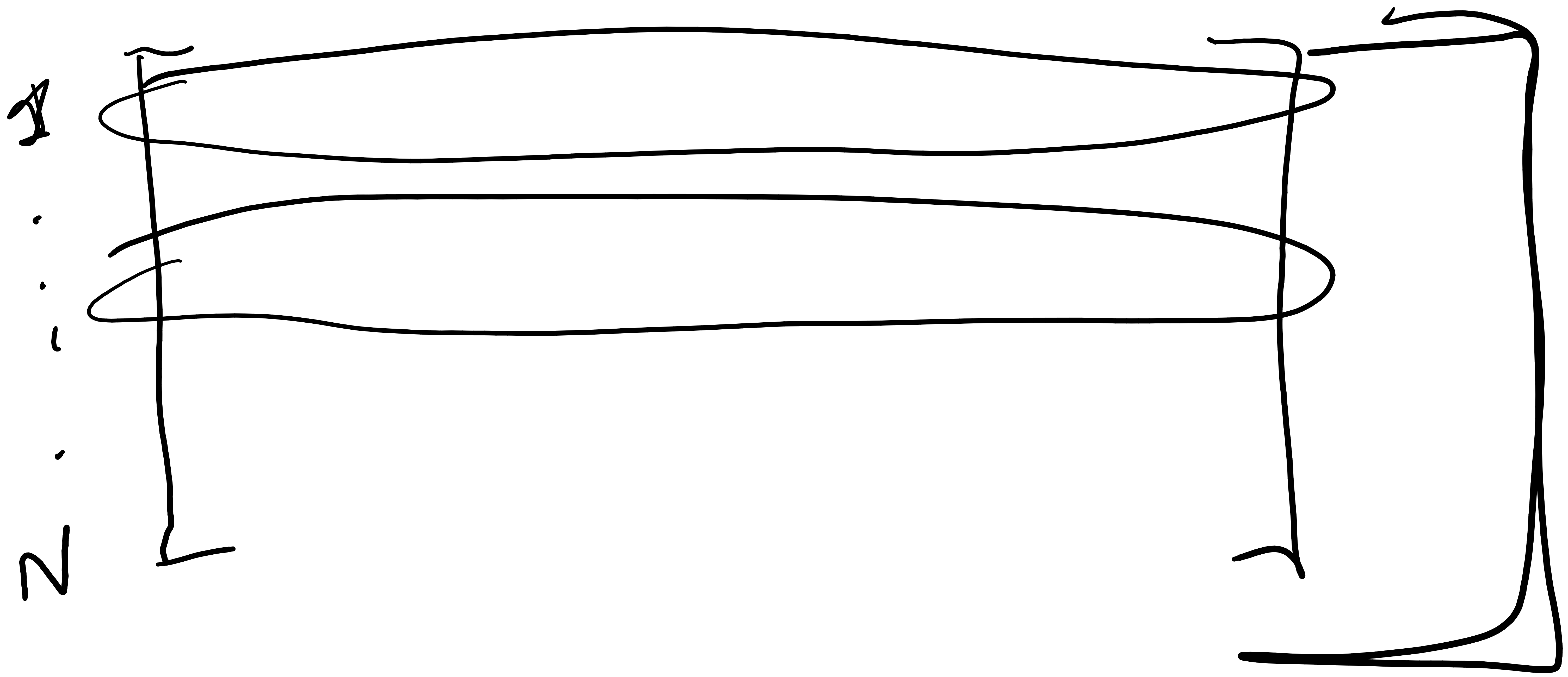
- COMPUTE AGAIN THE CENTROIDS

VARIANCE WITHIN ↓

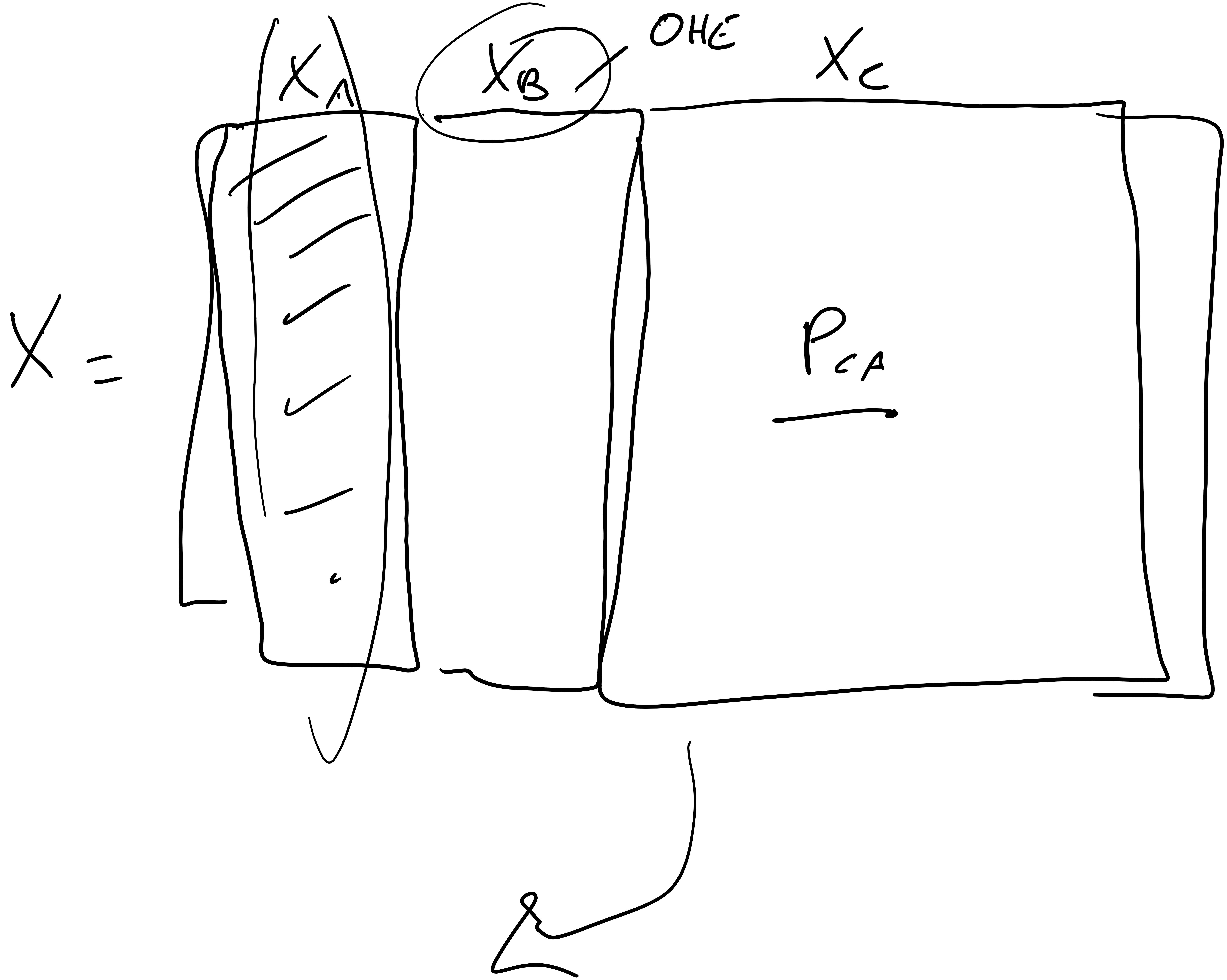
VARIANCE BETWEEN ↑



$X =$



- AGGREGATIVE
- DIVISIVE

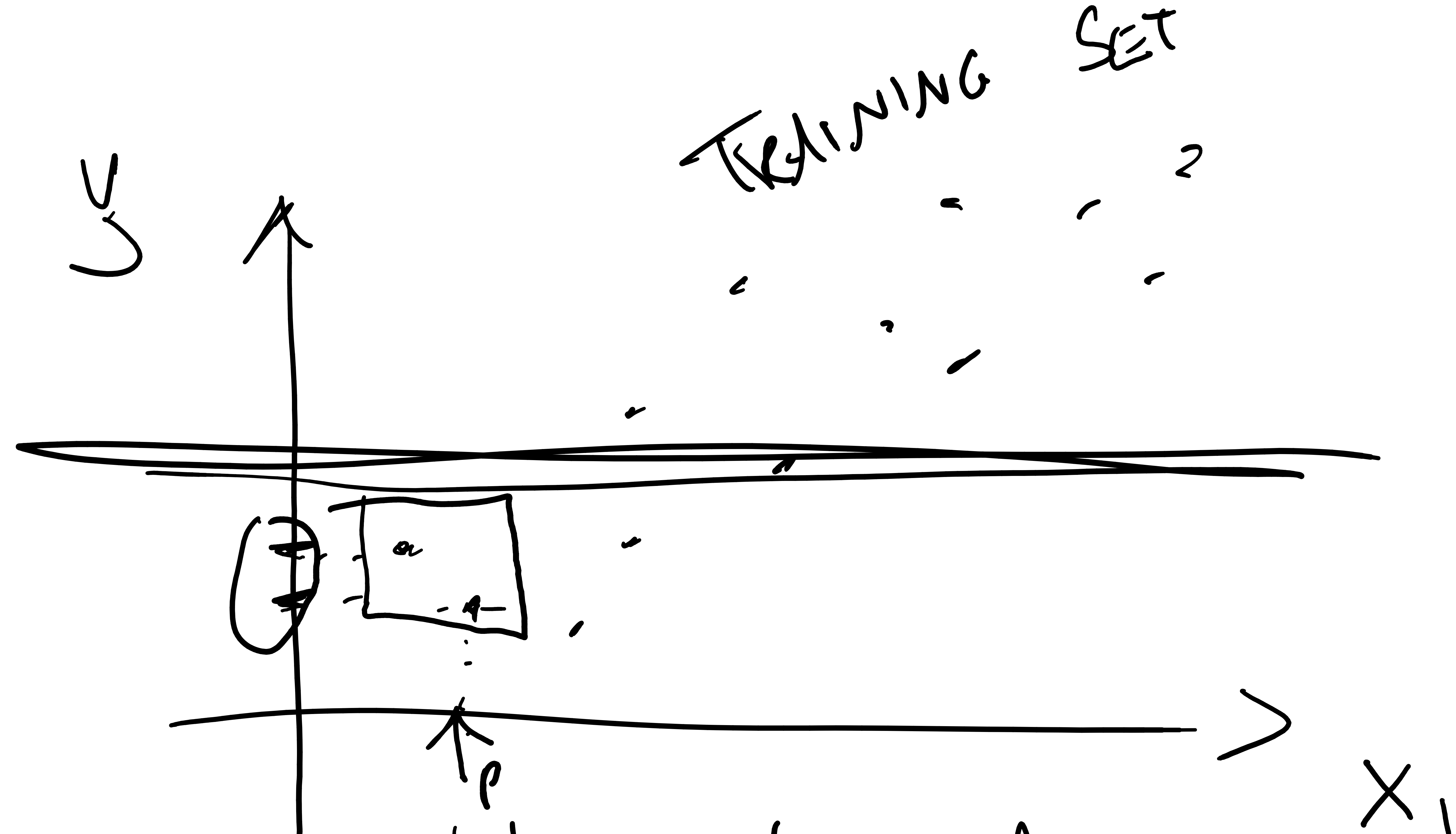


F. SELECT.

KNN (REGR., CLASS.)

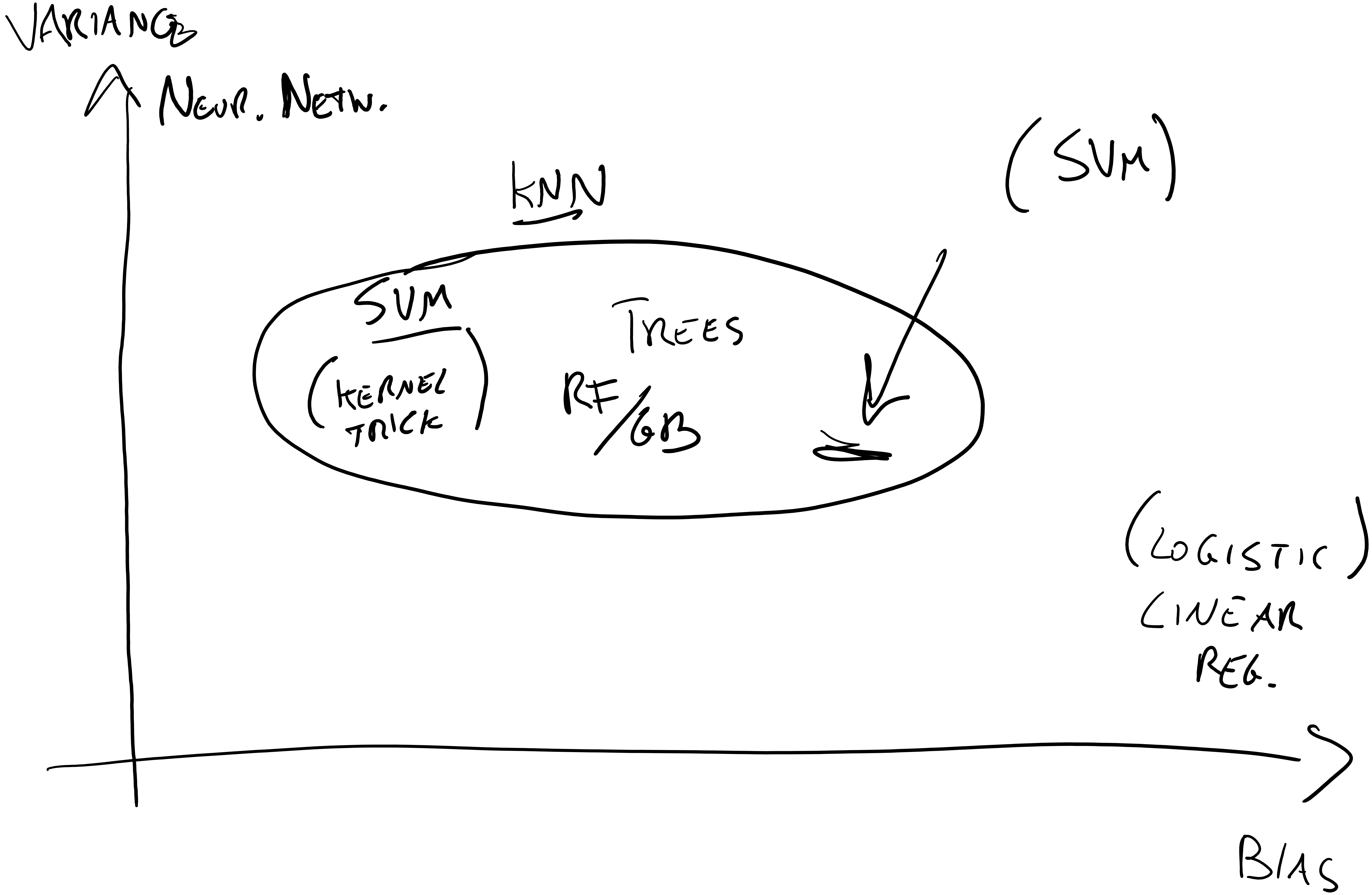
K = 2

K = N



What is the value of y for the k nearest points?

$$\hat{y}_p = \frac{\tilde{y}_1 + \tilde{y}_2}{2}$$



VARIANCE
NEUR. NETW.
KNN
(SVM)
SVM
(KERNEL TRICK)
TREES
RF/GB
(LOGISTIC)
LINEAR REG.
BIAS

SSE