Dimensionality reduction

2D, 3D : Scatter plot

4D, 5D, 6D : pair plot   784-dim data

10-D, 100-D, 1000-D - :

nD ⟶ 3D or 2D

(PCA)

Principle component Analysis & t-SNE

Row Vector & Column Vector

flower: $[SL, PL, SW, PW]$ $\underbrace{}_{\text{real values}}$

ith Point: $x_i \in \mathbb{R}^d$ → d-dim column vector

column Vector ← x_i; $\mathbb{R}$ → real numbers

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{id} \end{bmatrix}_{d \times 1} : \text{column vector}$$

$$x_i = [2.1, 3.2, 4.6, 1.2]_{1 \times 4} : \text{row vector}$$

Dataset

$\mathcal{D} : \{x_i, y_i\}_{i=1}^{n}$ → datapoints

$x_i \in \mathbb{R}^d$; $x_i \in \mathbb{R}^4$

$y_i \in \{$setosa, versicolor, virginica$\}$

Dataset as a data-matrix:

$$X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} f_1 & f_2 & f_3 \cdots & f_d \end{array} \\ \left[\begin{array}{c} \\ \\ \longleftarrow x_i^T \longrightarrow \\ \\ \\ \end{array} \right]_{n \times d} \end{array}$$

$D = \{x_i, y_i\}_{i=1}^{n}$

$x_i \in R^d$

$y_i \in \{s, v, ve\}$

d-features

each data point = row

each column = feature

$$x = \begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \\ f_j \\ f_d \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3\ 4 \cdots & i & \cdots n \end{array} \\ \left[\begin{array}{c} \uparrow \\ x_1 \\ \downarrow \end{array} \right]_{d \times n} \end{array}$$

each column = data point

row = feature

$f_1 = PL$

$f_2 = PW$

$f_3 = SL$

$f_4 = SW$

Data pre-processing: column-normalization

$$x = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{array} \begin{array}{c} \begin{array}{cccccc} f_1 & f_2 & f_3 \cdots & f_j & \cdots & f_d \end{array} \\ \left[\begin{array}{c} \\ \\ \longleftarrow x_i^T \longrightarrow \\ \\ \\ \end{array} \right]_{n \times d} \end{array} \rightarrow \begin{array}{c} Y \\ \left[\begin{array}{c} \\ \\ y_i \\ \\ \\ \end{array} \right]_{n \times 1} \end{array}$$

obtain data → pre-processing → data modelling (dim reduction)

→ column norm

$$
X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ 150 = n \end{array}
\begin{array}{c}
\begin{array}{ccccccc} f_1 & f_2 & f_3 & \cdots & \boxed{f_j} & \cdots & f_d \end{array} \\
\left[\begin{array}{ccccccc}
 & & b_1 & & a_1 & & \\
 & & b_2 & & a_2 & & \\
 & & \vdots & & a_3 & & \\
 & & \vdots & & \vdots & & \\
\longleftarrow & & x_i^T & & a_i & & \longrightarrow \\
 & & \vdots & & \vdots & & \\
 & & b_n & & a_n & &
\end{array}\right]_{n \times d}
\end{array}
\quad
\begin{array}{c} y \\ \left[\begin{array}{c} \\ \\ \\ \\ y_i \\ \\ \\ \end{array}\right]_{n \times 1} \end{array}
$$

Column normalization

column $a_1, a_2, \ldots\ldots a_n \rightarrow$ n values of feature f_j

$Max(a_i) = \cancel{max}(a_{max}) \geq a_i \quad (i: 1 - n)$

$min(a_i) = a_{min} \leq a_i \quad (i: 1 \rightarrow n)$

$$\boxed{a_i' = \frac{a_i - a_{min}}{a_{max} - a_{min}}} \qquad a_i' \in [0,1]$$

$a_1', a_2', a_3' \ldots\ldots a_i' \ldots a_n'$

$a_1, a_2, \ldots\ldots a_i \ldots\ldots a_d$

$\downarrow$ column-normalization

$a_1', a_2', \ldots\ldots a_i' \ldots\ldots a_d'$ st $a_i' \in [0,1]$

why?

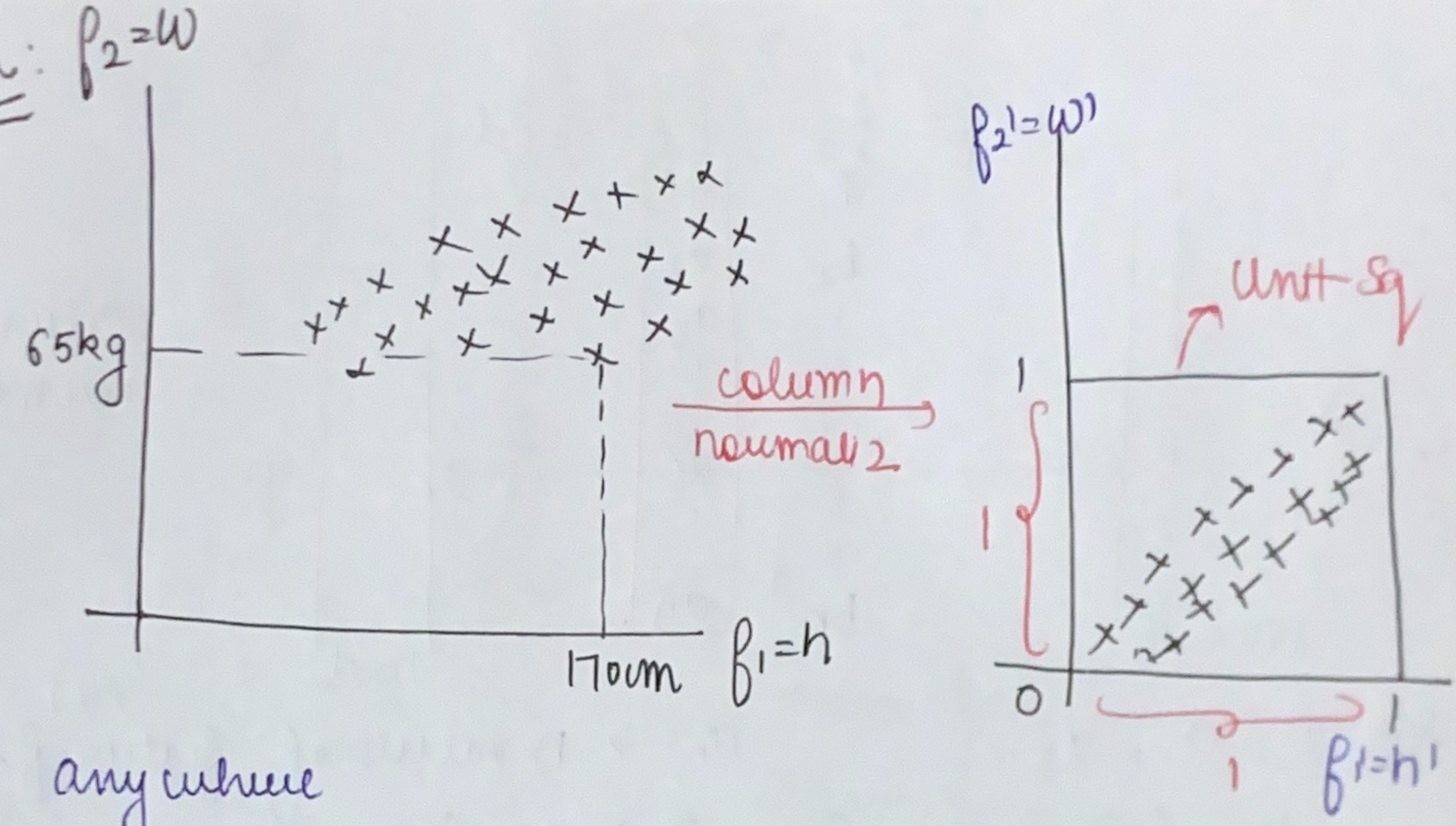
	$f_1 = h$	$W = f_2$	
	162	56	student 1
	172	72	2
	182	84	|
	|	|	|
	|	|	|
	|	|	|
			n
	↑ cm	↑ kg	

Colum-norm
getting rid of scale
(same scale)

	h'	w'
1		
2		
3		
	↑ [0,1]	↑ [0,1]

getting all points in one standard format $\in [0,1]$

Geom:



anywhere

n-dimspace $\xrightarrow[\text{norma}]{\text{col}}$ Unit-hyp cube in n-dim space

Mean Vector :

$$X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} f_1 \quad f_2 \; \cdots \; f_j \; \cdots \; f_d \\ \left[\begin{array}{c} \\ \\ \\ \longleftarrow x_i^T \longrightarrow \\ \\ \\ \end{array} \right]_{n \times d} \end{array}$$

Let $x_i \in \mathbb{R}^d$

$x_1 = [2.2, \; 4.2] \in \mathbb{R}^3$

$x_2 = [1.2, \; 3.2] \in \mathbb{R}^2$

$x_1 + x_2 = [3.4, \; 7.4]$

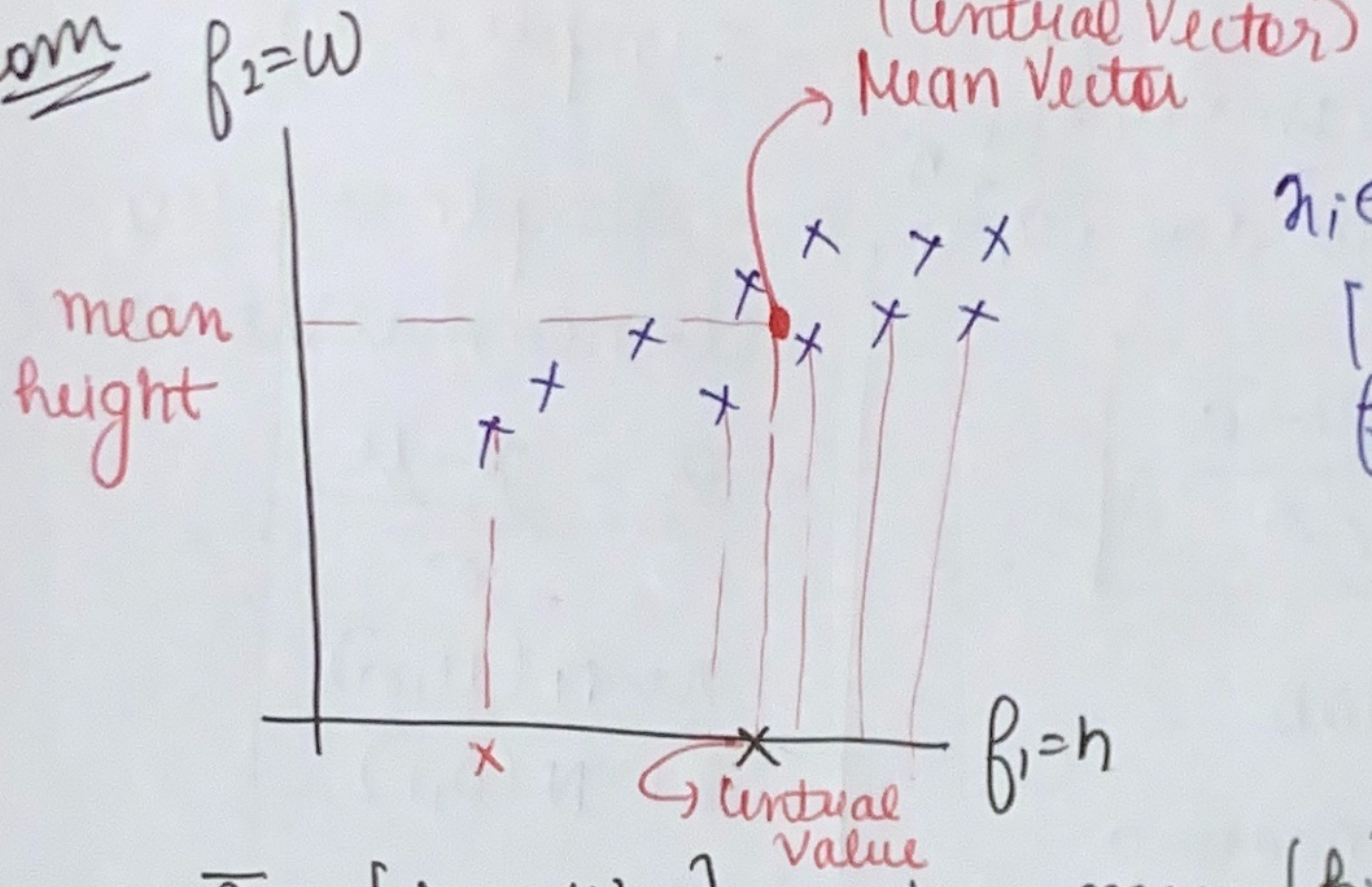
$\bar{x} \in \mathbb{R}^d$

$x_i \in \mathbb{R}_d$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 \; \cdots \cdots + x_n)$$

Mean Vector

Geom



$x_i \in \mathbb{R}^2$

$[h, w]$
$f_1 \quad f_2$

$\bar{x} = [h_{\bar{x}}, w_{\bar{x}}] \qquad h_{\bar{x}} = \text{mean}(h_i)_{i=1}^{n}$

$w_{\bar{x}} = \text{mean}(w_i)_{i=1}^{n}$

Data-preprocessing: Column standardization

column normalization: $[0,1]$ ← get rid of scales of each feature

column standardization: more often used

$$
X = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ n \end{array}
\begin{bmatrix} f_1 & f_2 & \cdots & f_j & \cdots & f_d \\ & & & a_1 & & \\ & & & a_2 & & \\ & & & \vdots & & \\ \longleftarrow & & x_i^T & & & \longrightarrow \\ & & & \vdots & & \\ & & & a_n & & \end{bmatrix}_{n \times d}
$$

$a_1, a_2, a_3 \cdots a_n$ ← n values of f_j

$\big\downarrow$ col. std

$a_1', a_2', a_3' \cdots a_n'$ ← $\text{mean}\{a_i'\}_{i=1}^{n} = 0$

std dev $\{a_i'\}_{i=1}^{n} = 1$

$\bar{a} = \text{Mean}\{a_i\}_{i=1}^{n}$ ← sample mean

$s = \text{std dev}\{a_i\}_{i=1}^{n}$ ← sample std-dev

$$\boxed{\bar{a}_i = \frac{a_i - \bar{a}}{s}}$$

avg dist

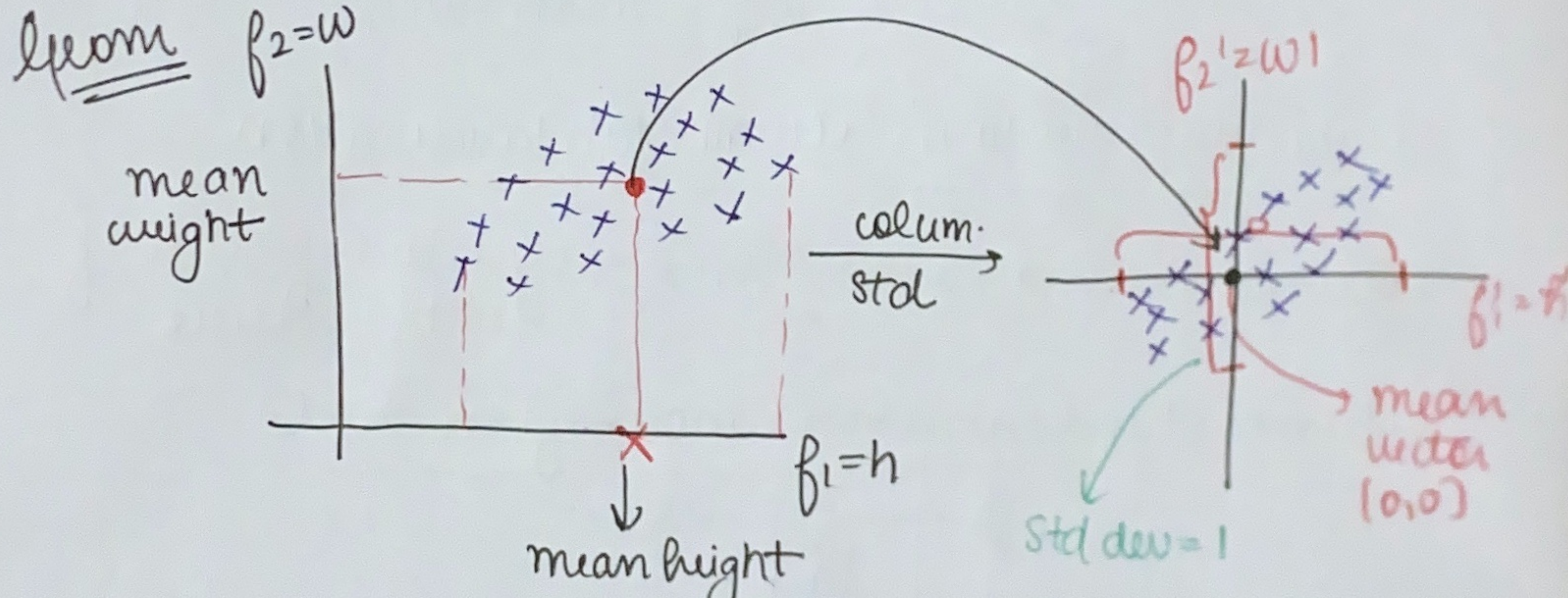
a_i's → mean=0
→ std dev=1

Std- normal Variate (z)

$$z = \frac{x-\mu}{\sigma}$$

$x \sim N(\mu, \sigma)$

$z \sim N(0,1)$



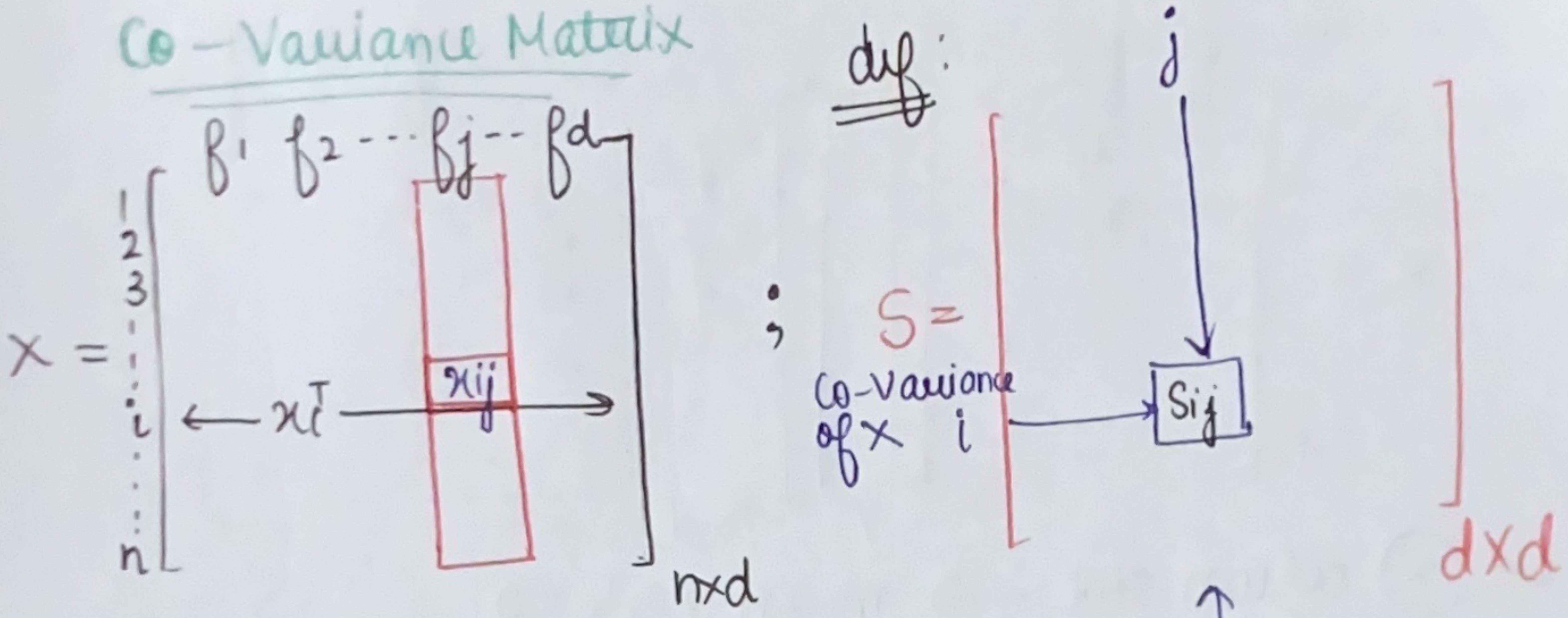
① Moving mean vector to origin

② squishing / expanding s.t std-dev for any feature is 1

Col. Std = mean centering → origin
scaling → std dev=1 for all features

Co-Variance Matrix

def:



S_{ij} = i^{th} row & j^{th} col element in S

$S_{ij} = cov(f_i, f_j)$ f_j = column vector j^{th} feature

$i: 1 \rightarrow d$

$j: 1 \rightarrow d$

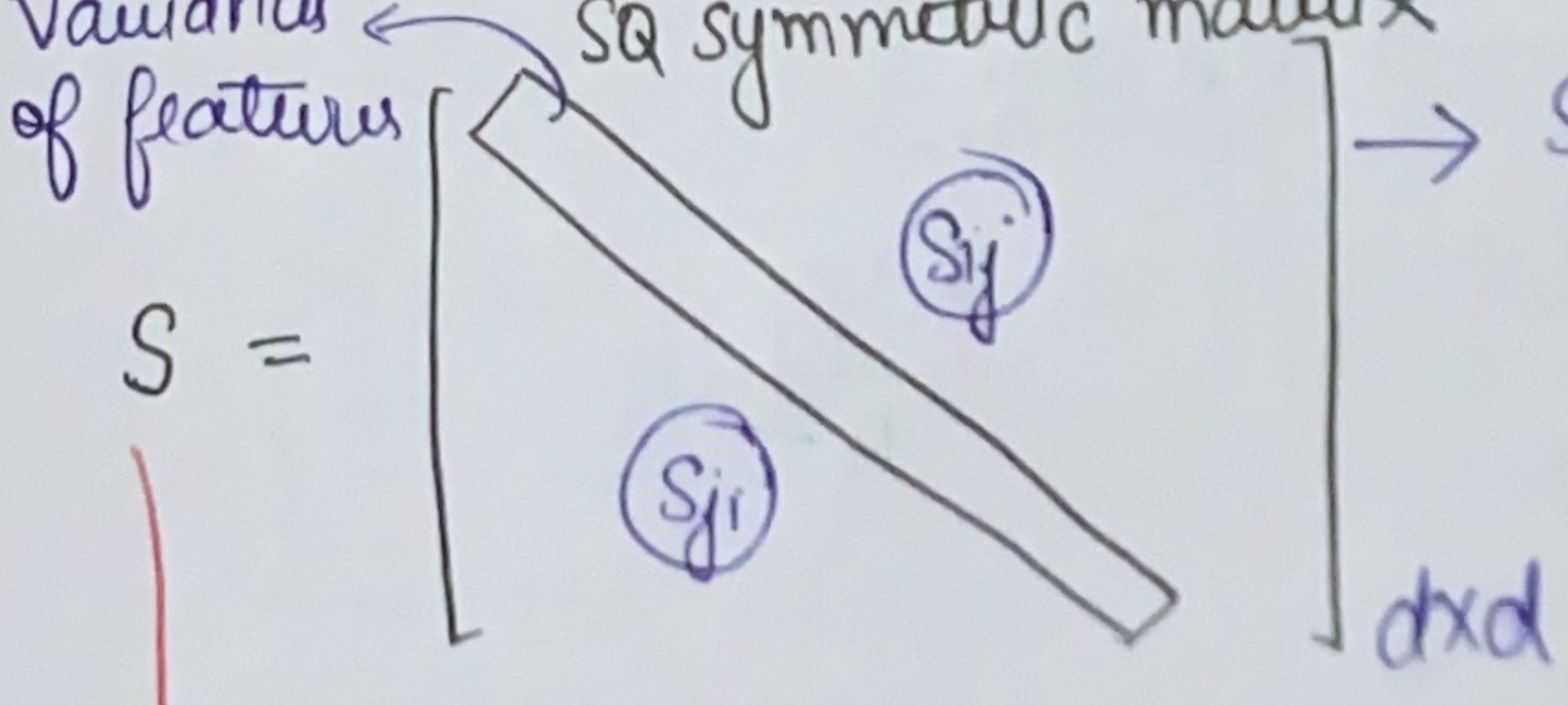
x_{ij} = j^{th} feature for i^{th} data point

$$\boxed{\underline{\underline{cov}}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}$$

$cov(f_i, f_j) = Var(f_i)$

$cov(x, x) = Var(x)$ —— (1)

$cov(f_i, f_j) = cov(f_j, f_i)$ —— (2)



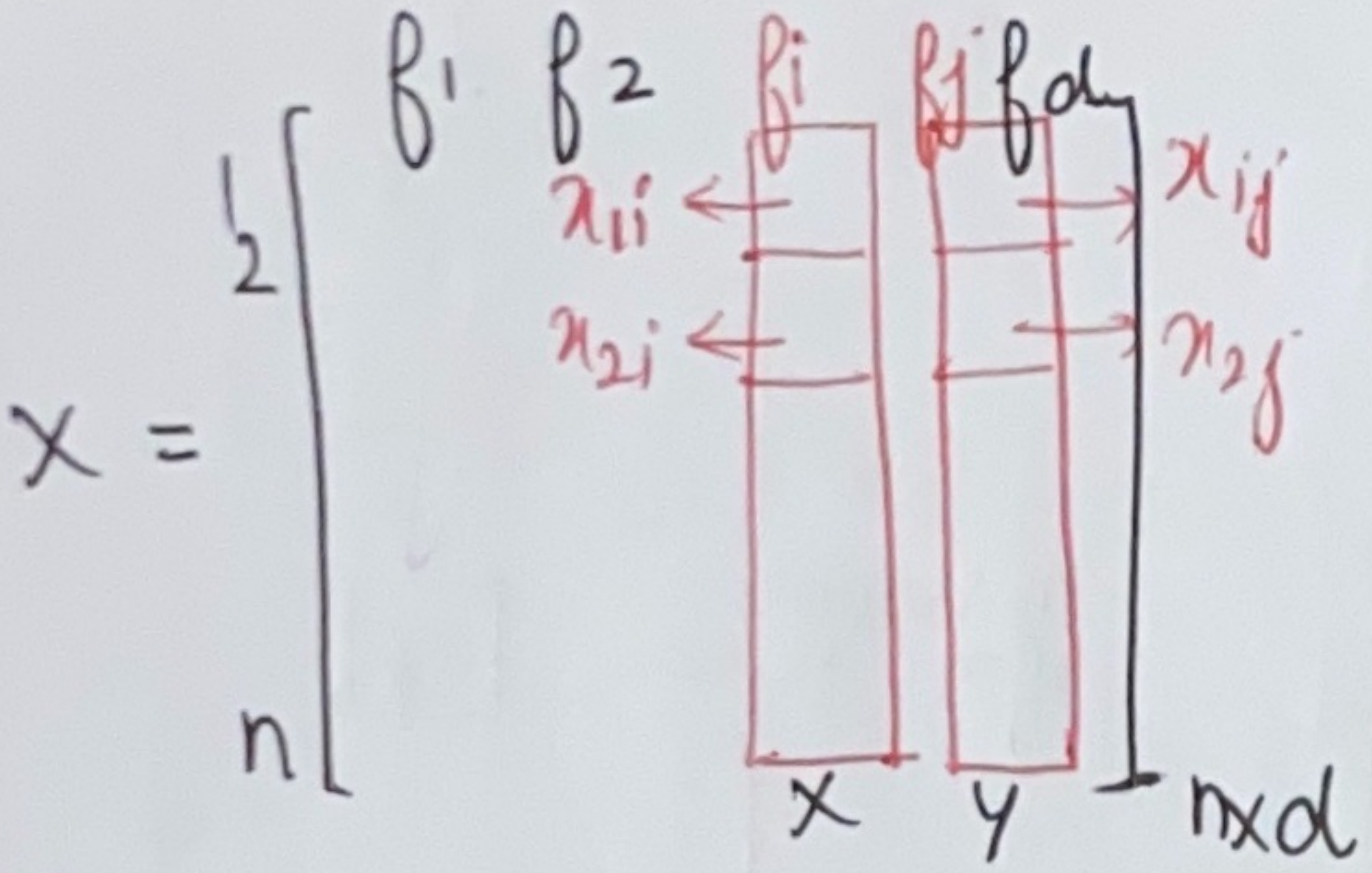
→ Symmetric matrix

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & & & \\ 2 & & & \\ 3 & & & \end{array}\quad \begin{bmatrix} 2 & 1 & 2 \\ 1 & 1 & 5 \\ 2 & 5 & 3 \end{bmatrix}_{3\times 3}$$

$A_{21} = A_{12}$

$A_{ij} = A_{ji} \ \forall \ i, j$

Symmetric Matrix & Square Matrix (d, d)



f_1 f_2

let X colum std $\rightarrow$ mean $\{f_i\} = 0$

std dev $\{f_i\} = 1$

$$\boxed{\operatorname{cov}(f_{i\,(1)}, f_{j\,(2)}) = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \mu_1)(x_{i2} - \mu_2)}$$

μ_2 ← mean(f_2)

μ_1 ← mean(f_1)

$$\boxed{\operatorname{cov}(f_1, f_2) = \frac{1}{n}\sum_{i=1}^{n} x_{i1} * x_{i2}}$$

if f_1 & f_2 are std,

then $\operatorname{cov}(f_1, f_2) = \dfrac{f_1^T f_2}{n}$

$$S_{d\times d} = \underset{d\times n}{(X^T)}\,\underset{n\times d}{(X)} = d\times d$$

* assuming $X \longrightarrow$ <u>cold std</u>

$$S_{ij} = \operatorname{cov}(f_i, f_j) = \frac{f_i^T f_j}{n}$$

MNIST Dataset

obj: classify the written character into one of 10 numeric characters

$$\mathcal{D} = \{x_i, y_i\}_{i=1}^{60K}$$

$x_i = \boxed{0}$ 28 × 28 $\qquad y_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

$x_i = \square$ 28 × 28 $\qquad x_i = \begin{bmatrix} \\ \\ \\ \end{bmatrix} \qquad x_i \in \mathbb{R}^d$

x_i = image = $\begin{bmatrix} & \\ & \end{bmatrix}_{28\times 28}$ → NOT data matrix, matrix expression of image

numerical/real matrix

	1	2	3	4	5
	1	2	4	6	8
1	3	2	1	8	2
2	2	1	6	8	4
3	3	2	1	8	2
4					
5	4	2	6	8	1

5×5

row flattening →

$$\begin{bmatrix} 1 \\ 2 \\ 4 \\ 6 \\ 8 \\ 3 \\ 2 \\ 1 \\ 8 \\ \vdots \end{bmatrix}_{25\times 1}$$

28 × 28 (28×28) → row flattening → data point $\begin{bmatrix} \\ \\ \\ \end{bmatrix}_{784\times 1}$

28×28 = 784

MNIST

$$X = \begin{array}{c|c} & \hat{f}_1 \quad \hat{f}_2 \quad f_3 \; \cdots \; f_{784} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ 60K \end{array} & \left[\begin{array}{c} x_i^T \in \mathbb{R}^{784} \\ \xleftrightarrow{\text{784 dim}} \end{array} \right]_{n \times d} \end{array} \qquad \left[\begin{array}{c} y_i \in \{0, 1, 2 \ldots 9\} \end{array} \right]_{n \times 1}$$

$n = 60K$

$d = 784$