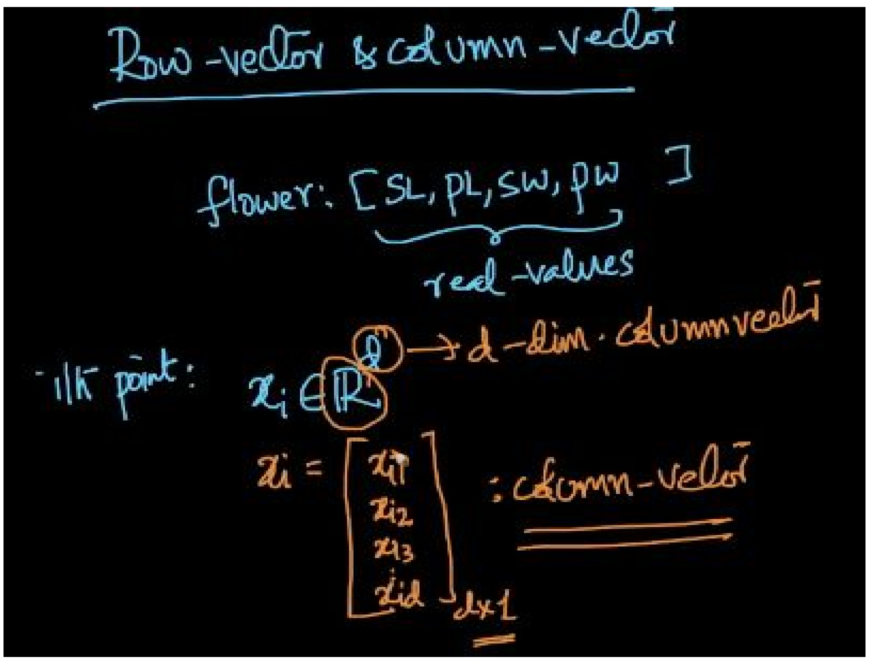
What is Dimensionality Reduction ?

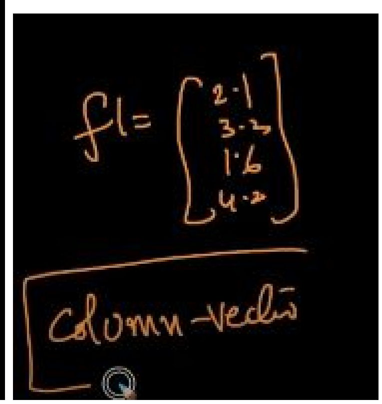
Dimensionality Reduction techniques are built upon the idea of linear algebra.

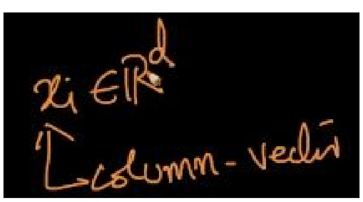
In EDA, We learn that we can visualize our data in 2D and 3D using Scatter Plots.
For 4D,5D- 6D we can leverage Pair Plots. (nc2)
But For nD (10D) : Pair plots won't work
We reduce the dimensionality to (2d or 3d) to make it understandable so that we can visualize.
Some of techniques are ***t-SNE*** (almost state of art) and ***PCA***(old technique)

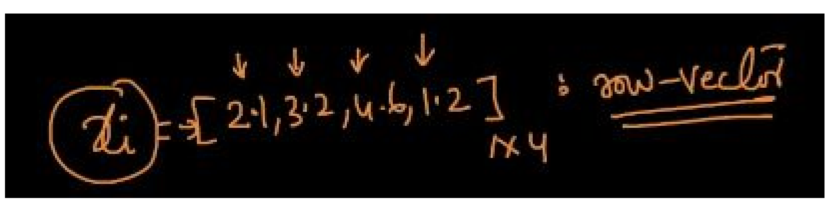
Row Vectors and Column Vector

For our iris flower dataset,
We were given 4 features or 4 variables : [SL,PL,SW,PW]









<u>How to represent a dataset?</u>

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

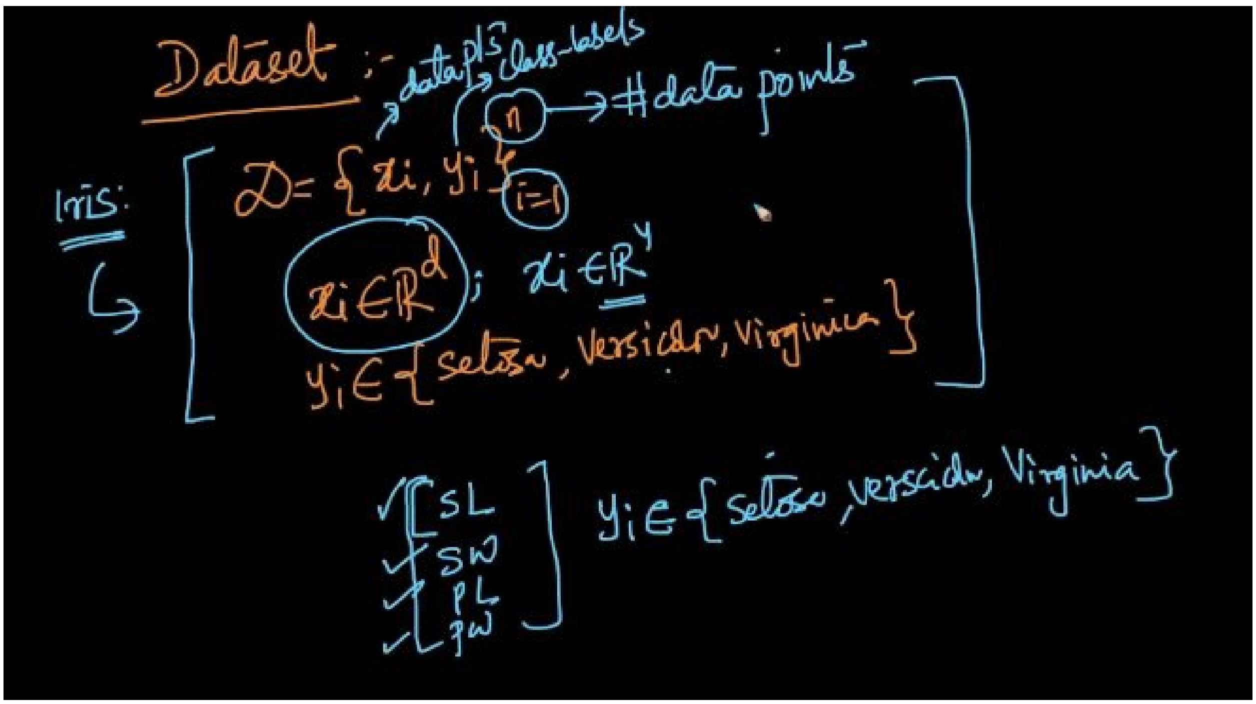
D is a collection of data points x_i and class levels y_i .

Each data points belongs to $|R^d$

In case of iris data set, X_i ε $|R^4$

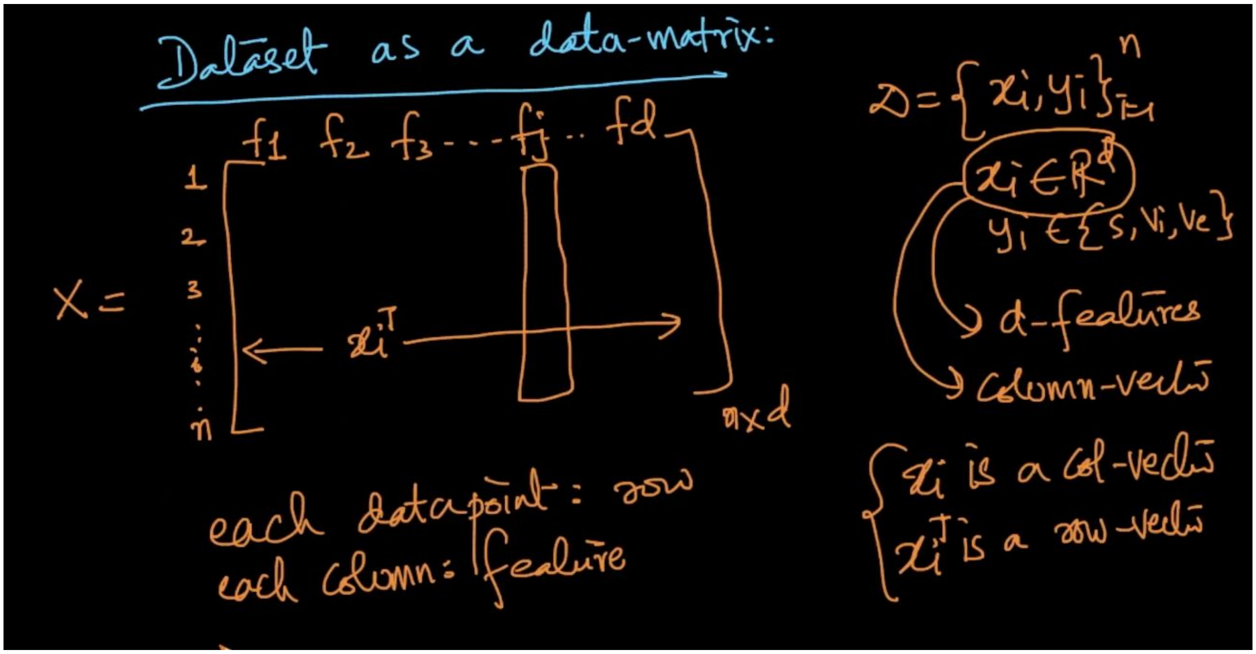
[SL,SW,PL,PW]

y_i ε {Setosa, versicolor,virginica}

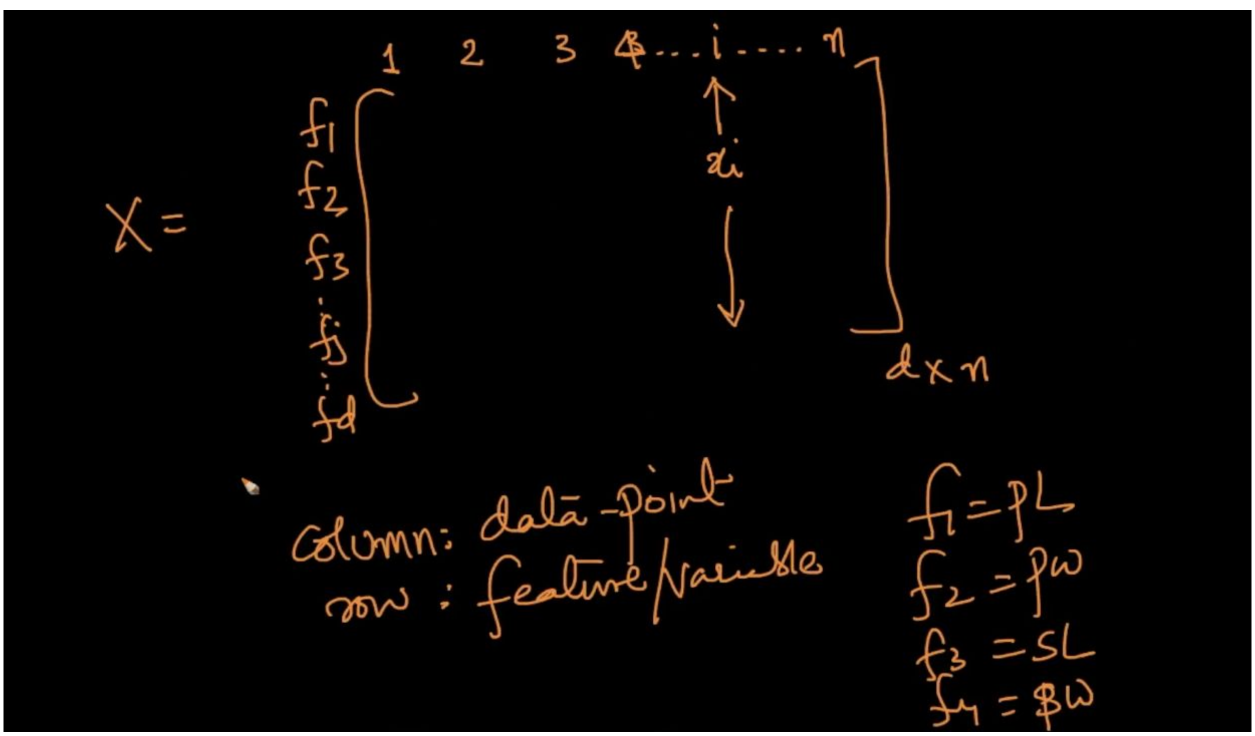


<u>How to represent a dataset as a Matrix ?</u>

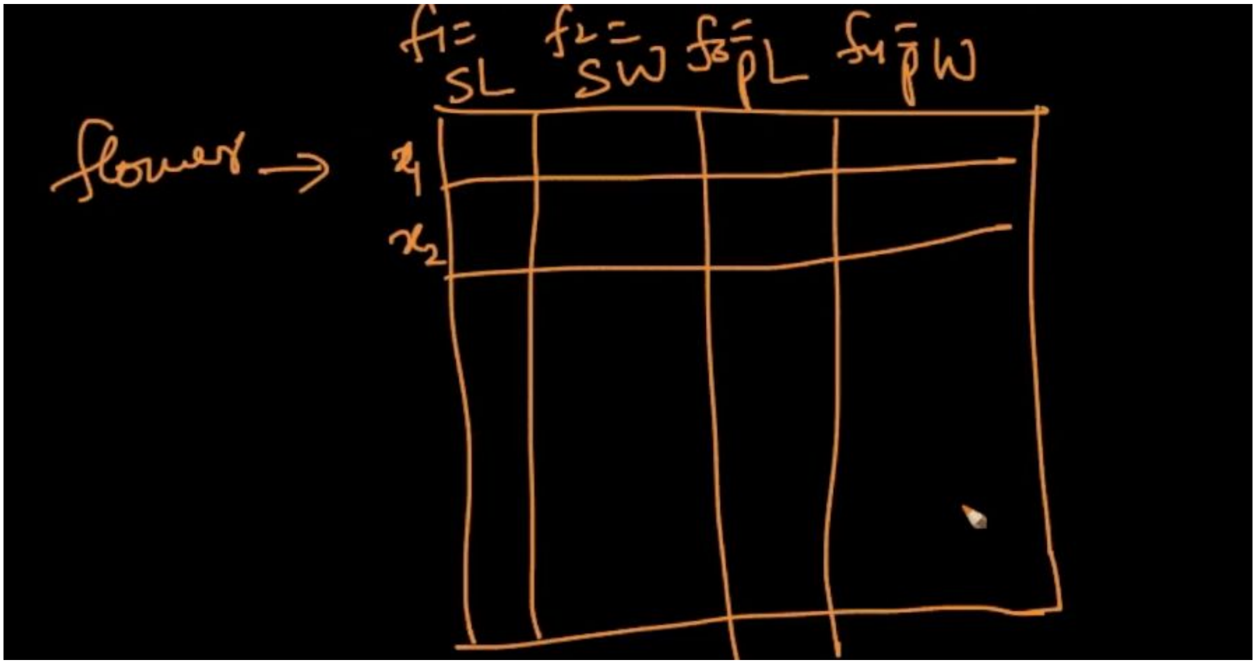
Data Set can be represented as a data matrix:

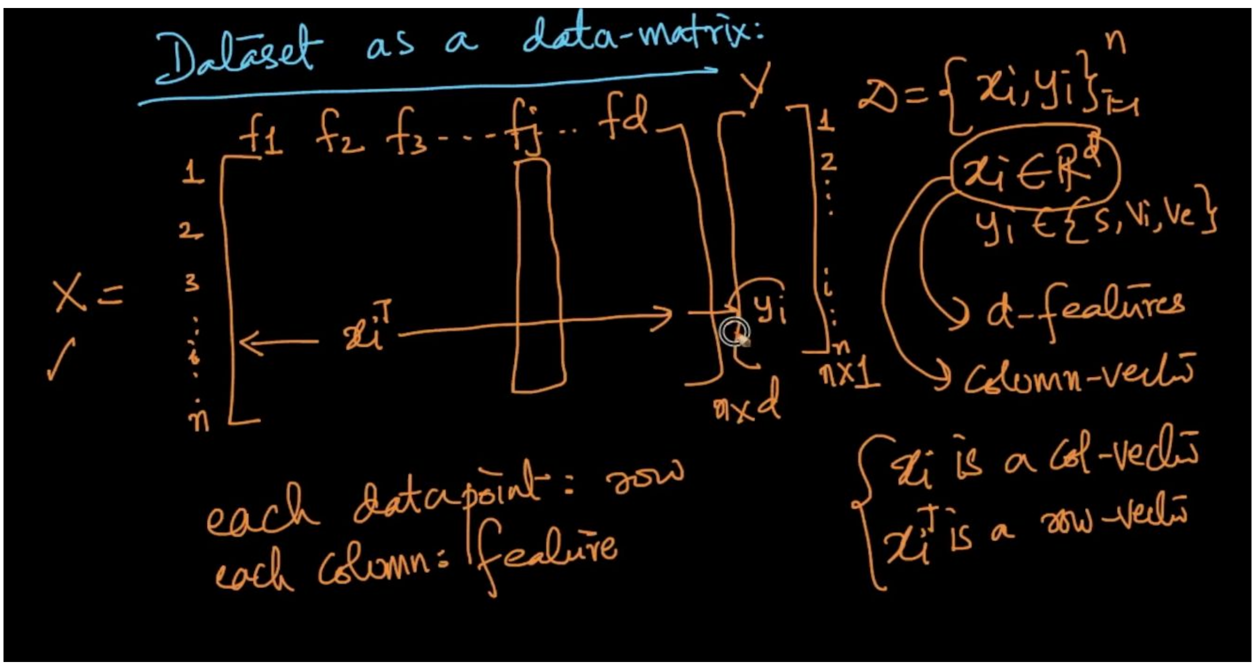


In Research papers we mostly see:



Let’s Look into big picture:

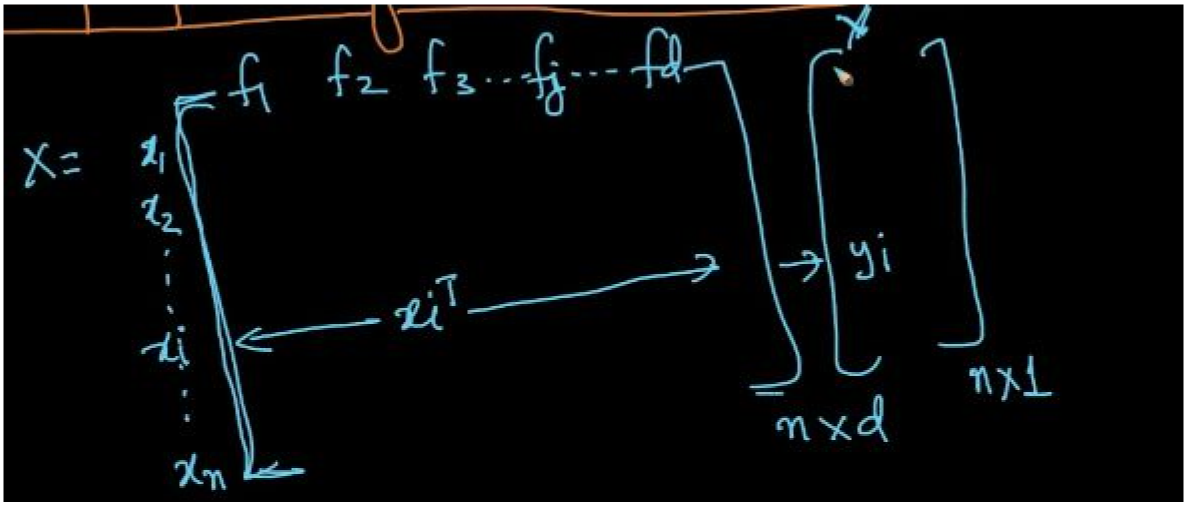




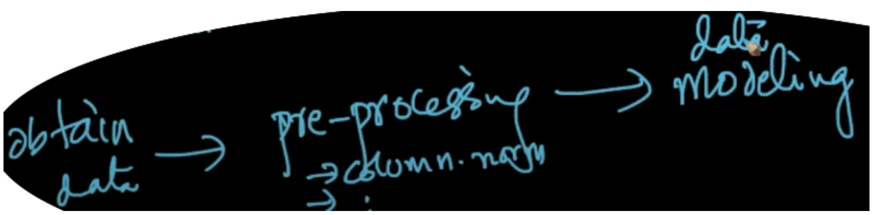
We will be using this format on our hands-on assignment and course work

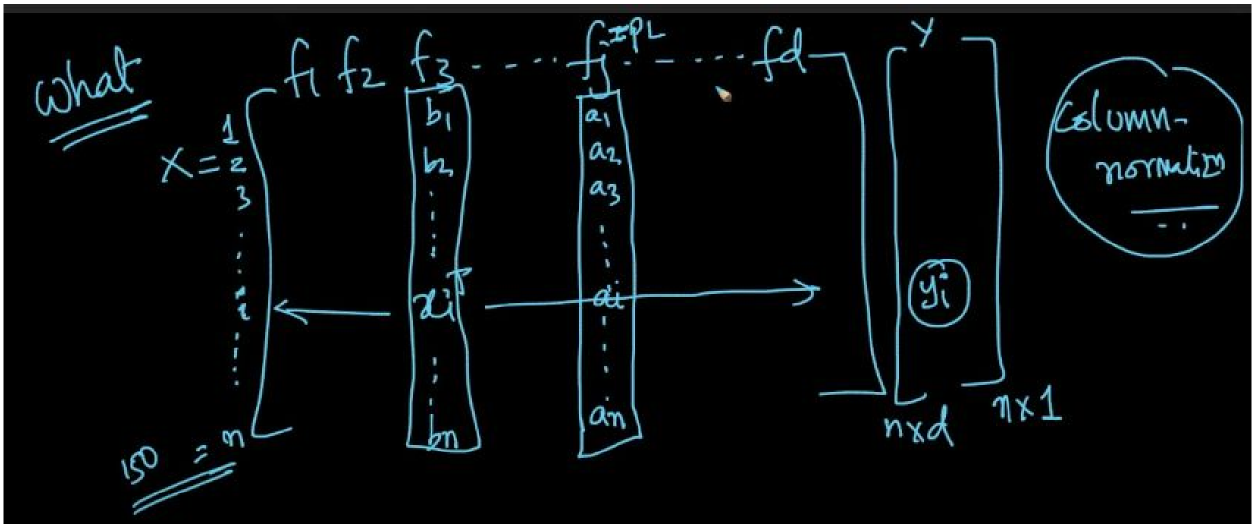
Data Preprocessing: Feature Normalisation/ Column Normalisation

Technique to squash most of the data into unit cube or cuboids with the aim of getting rid of scales such as kg,cm, pounds to make data modelling easier.



Preprocessing means some types of mathematical operations / transformation done on data itself after obtaining data and before doing data modelling(dimension reduction).





column: 1.2 1.3 1.4 1.9 1.7

$$a_1, a_2, \dots, a_i, \dots, a_n \rightarrow n\text{-values of } f_j$$

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

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

$$a_1', a_2', a_3', a_4', \dots, a_i', \dots a_n' \rightarrow a_i' \in [0, 1]$$

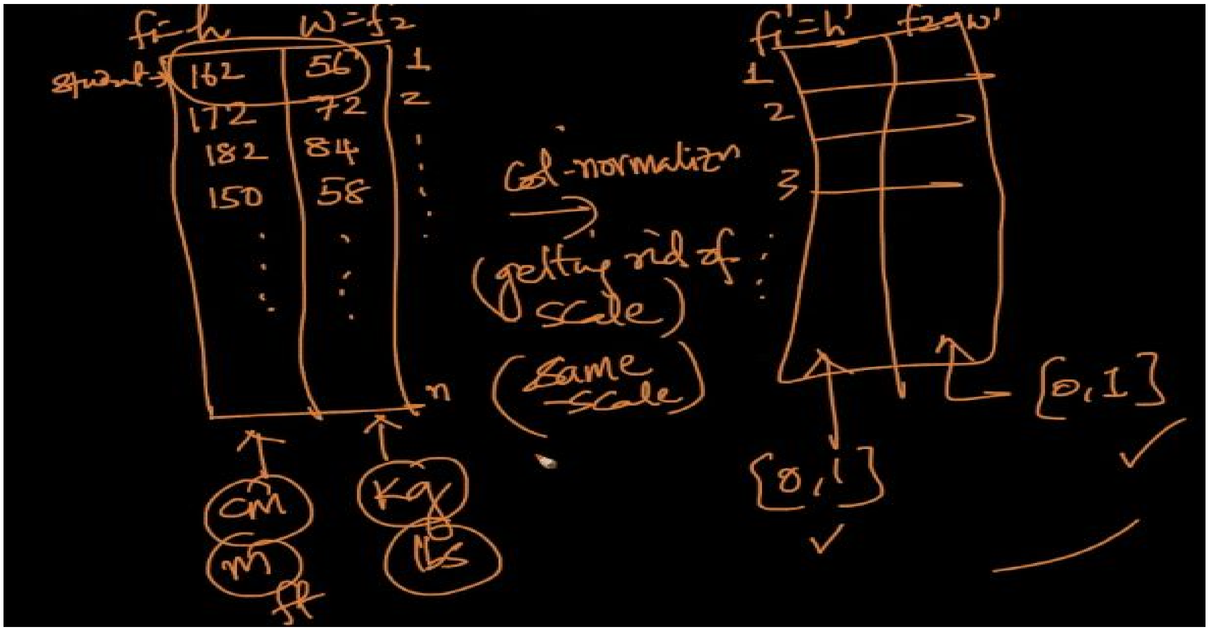
$$a_i' = \frac{a_i - a_{min}}{a_{max} - a_{min}}$$

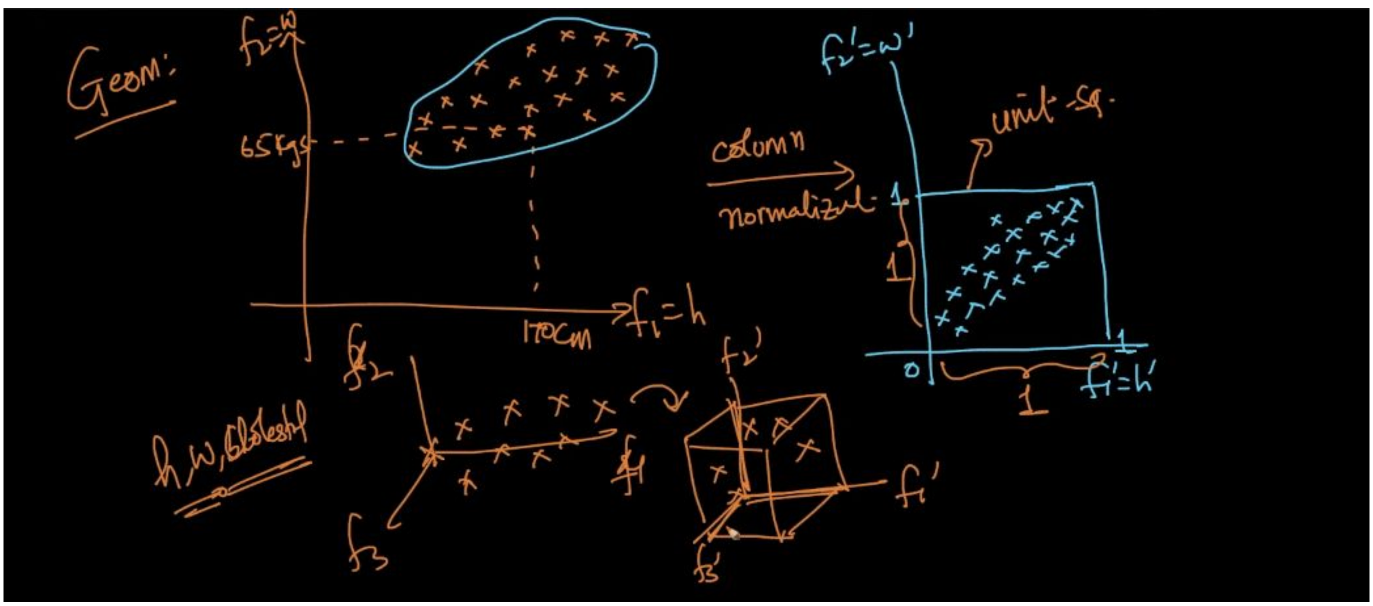
$$a_{min}' = \frac{a_{min} - a_{min}}{a_{max} - a_{min}} = 0 \; ; \quad a_{max}' = \frac{a_{max} - a_{min}}{a_{max} - a_{min}} = 1$$

$$a_1, a_2, \dots a_i \dots a_d \; ; \; a_i \in \mathbb{R}$$

$$\big\downarrow \text{column-normalization}$$

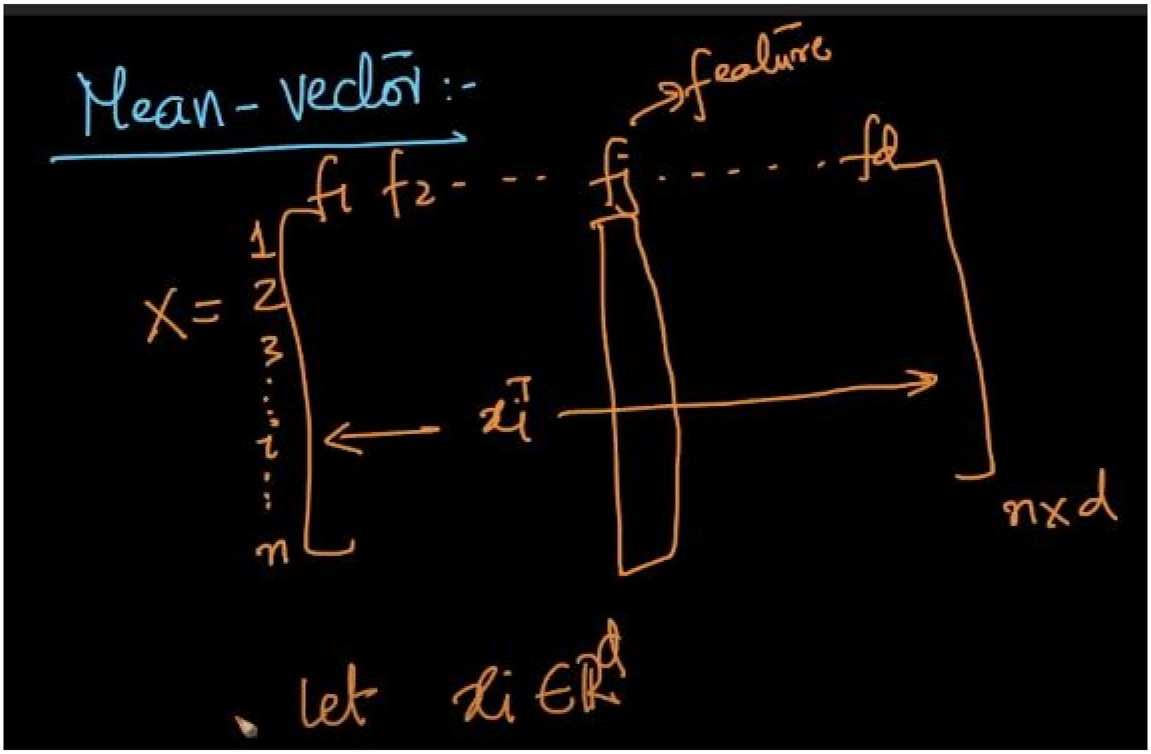
$$a_1', a_2' \dots a_i' \dots a_d' \; ; \; \text{s.t } a_i' \in [0, 1]$$

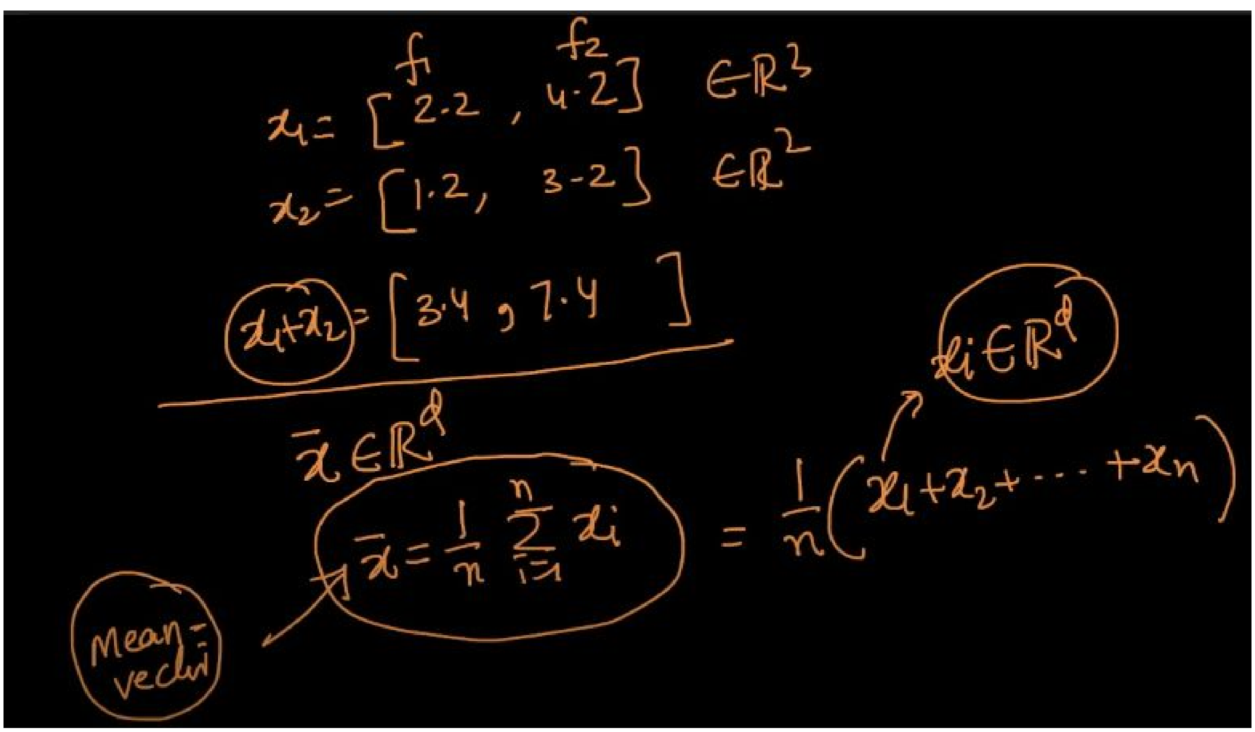


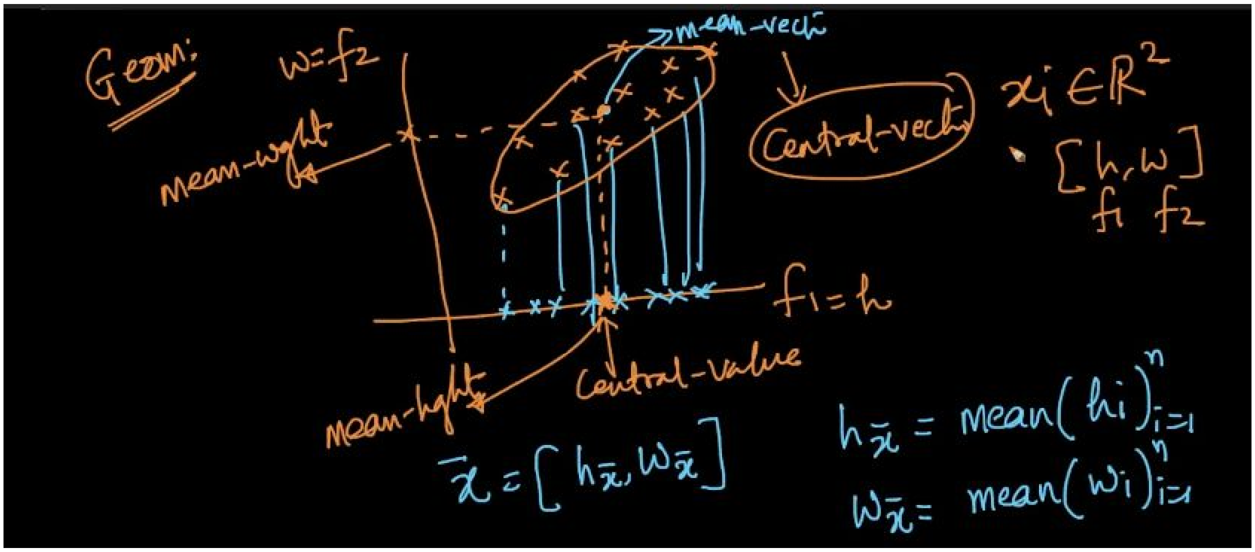


anywhere in n-dim space $\xrightarrow[\text{norm}]{\text{Col.}}$ unit-hypercube in n-dim space

Mean of data matrix



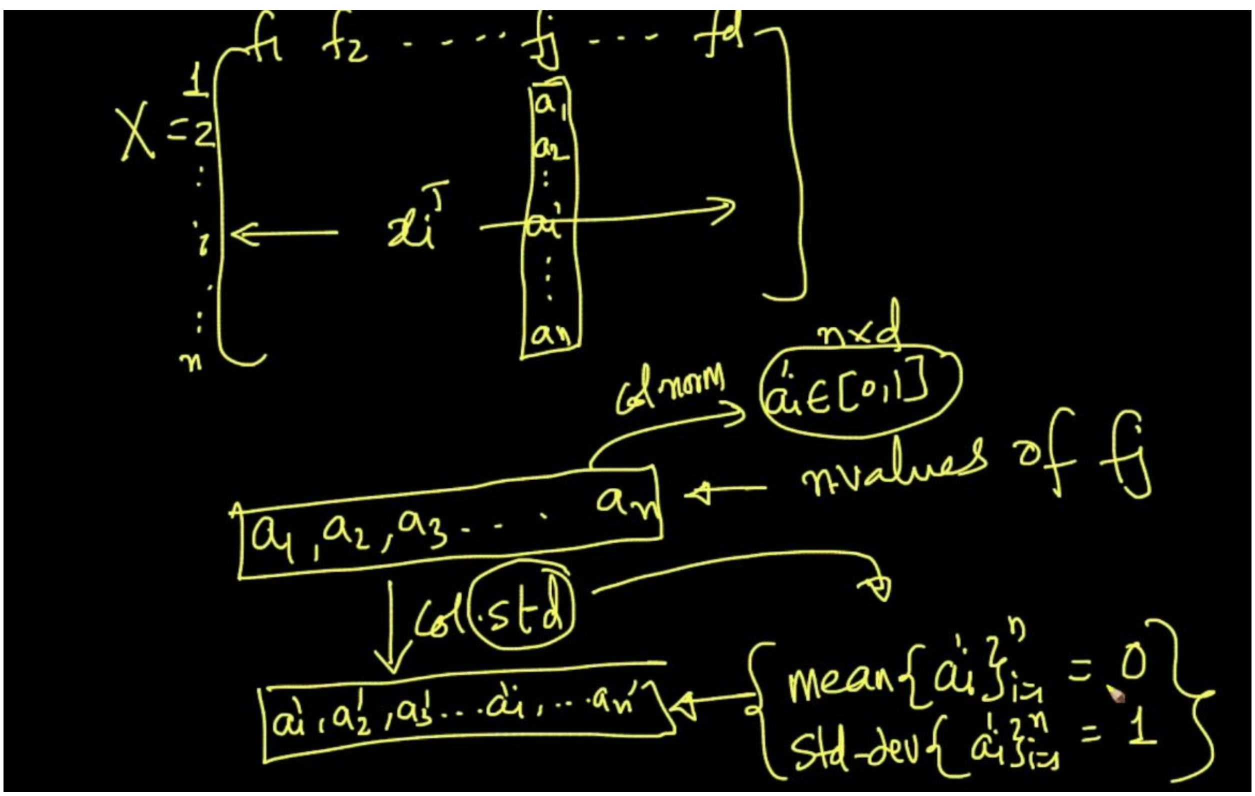




Data Preprocessing: Column Standardization

Like Column normalization = [0,1] = get rid of scales of each feature

Column standardization = Most often used in practice.



✓ $a_1, a_2, \dots a_i, \dots a_n$ ← any-disb

col-std

$a_1', a_2' \dots a_i' \dots a_n'$ → mean = 0

└→ std-dev = 1

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

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

$$a_i' = \frac{a_i - \bar{a}}{s}$$

$\boxed{\text{ex:}}$ $\text{mean}\{a_i'\}_{i=1}^{n} = 0$

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

$$a_i' = \frac{a_i - \bar{a}}{s}$$

any-disb

a_i's → mean = 0

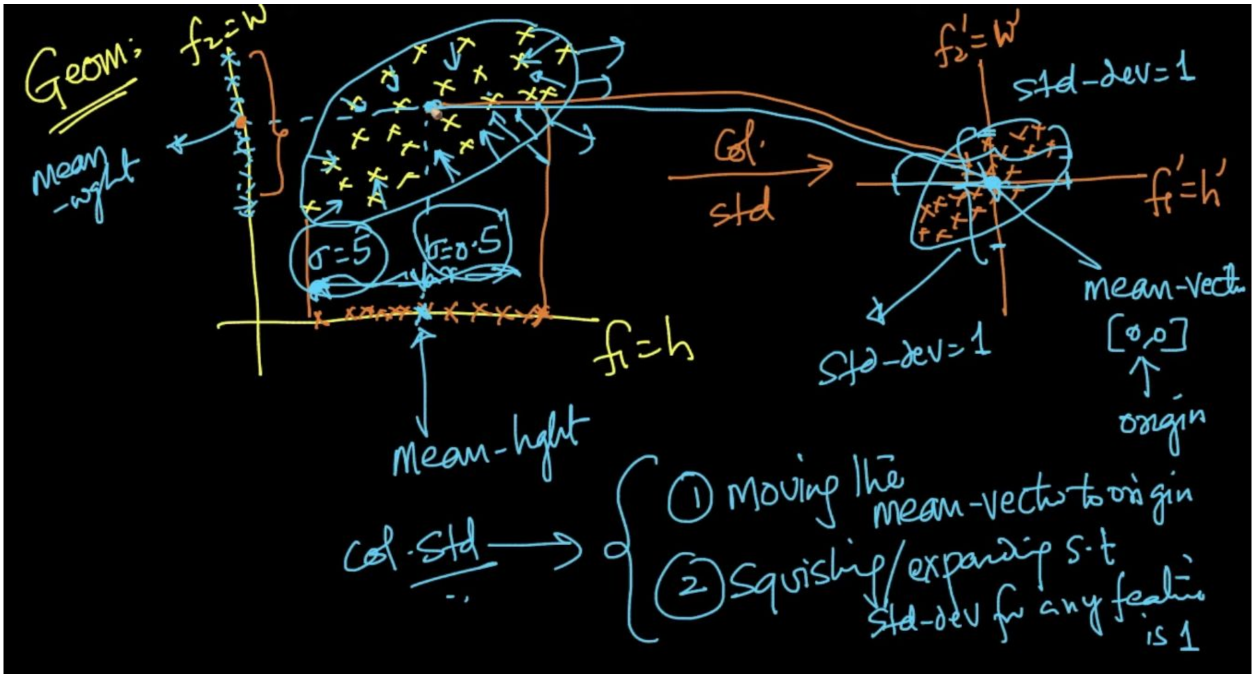
└→ std-dev = 1

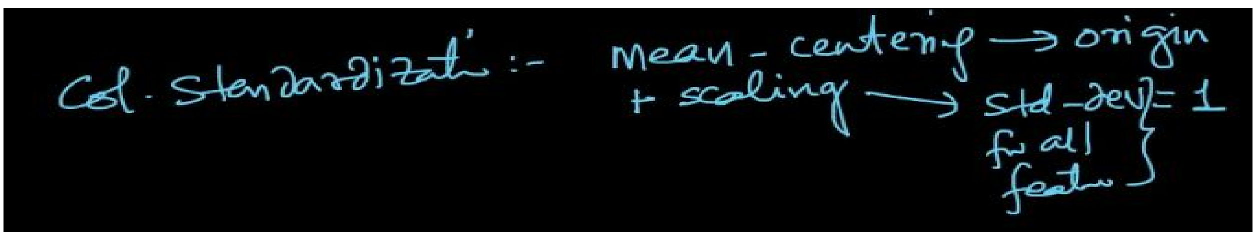
Std-normal-variate (z)

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

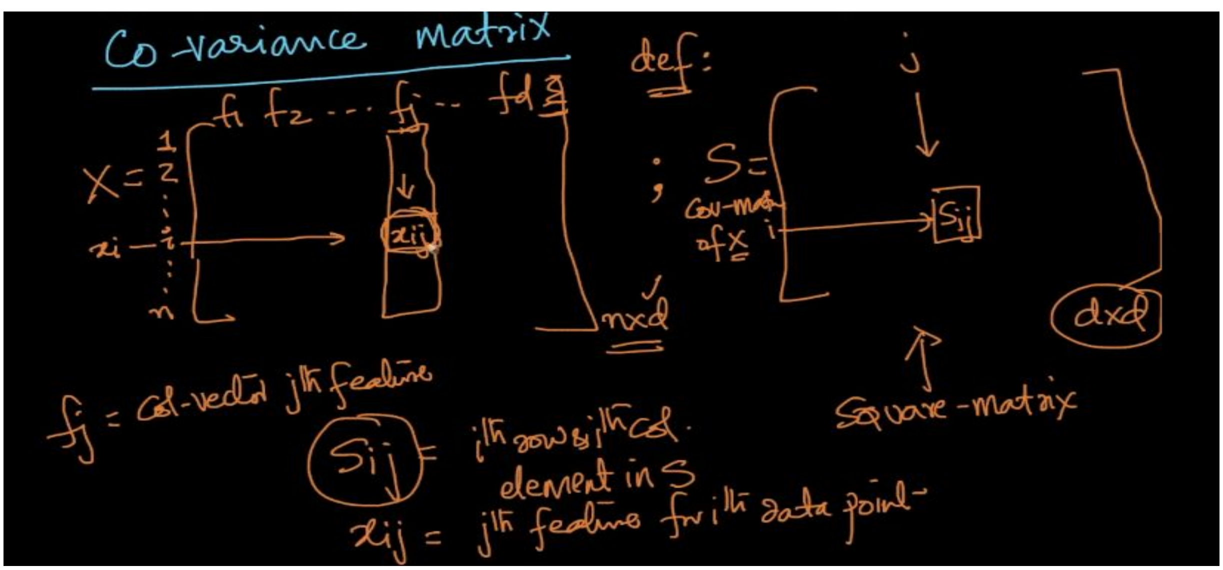
$x \sim N(\mu, \sigma)$

$z \sim N(\underset{\mu}{0}, \underset{\sigma}{1})$





Covariance Matrix of Data Matrix



$S_{ij} = cov(f_i, f_j)$

$i: 1 \to d$
$j: 1 \to d$

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

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

✓ $cov(x, x) = Var(x)$ — 1

✓ $cov(f_i, f_j) = cov(f_j, f_i)$ — 2

Variance of feature

Sq. Symm matrix

$$S = \begin{bmatrix} s_{11} & & s_{ij} & \\ & s_{22} & & \\ & & s_{33} & \\ s_{ji} & & & s_{44} \\ & & & & \ddots \\ & & & & & s_{dd} \end{bmatrix}_{d \times d}$$

Symmetric matrix

$$A = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 2 & 1 & 2 \\ 2 & 1 & 1 & 5 \\ 3 & 2 & 5 & 3 \end{array} \quad 3\times3$$

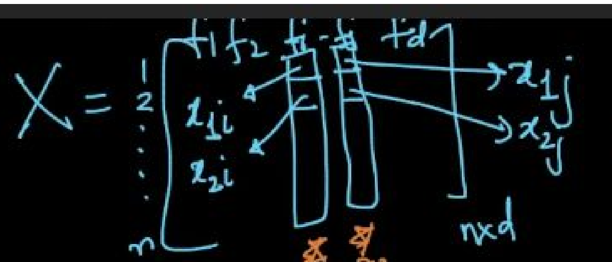
$\boxed{A_{32} = 5 = A_{23}}$

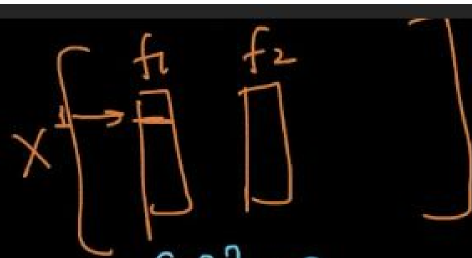
$A_{21} = A_{12}$

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

Symm - matrix

Sq. matrix $\begin{cases} d: \text{rows} \\ d: \text{col} \end{cases}$



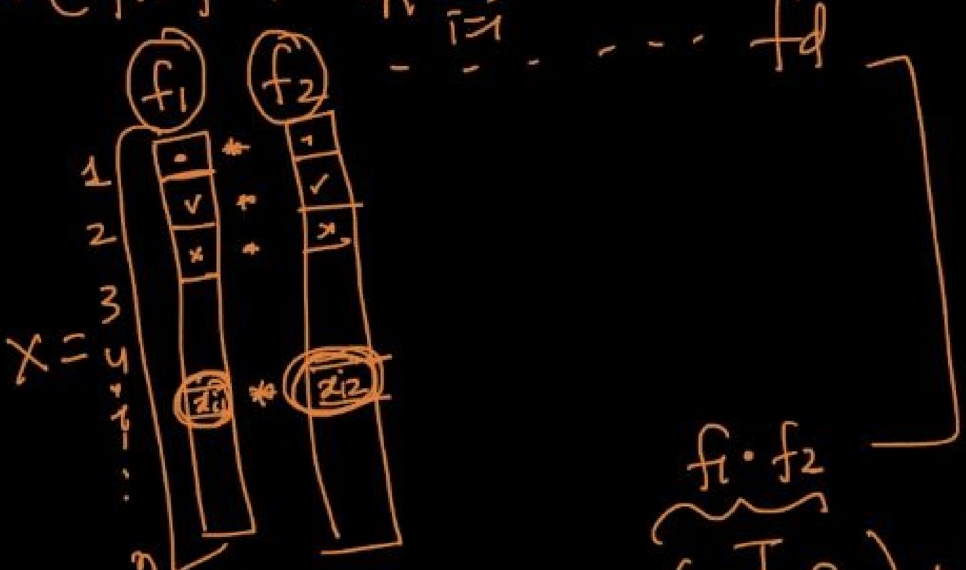


Let X col. Standardized $\Rightarrow$ mean$\{f_i\} = 0$

std-dev$\{f_i\} = 1$

$$Cov(f_1, f_2) = \underbrace{\frac{1}{n} \sum_{i=1}^{n}}_{avg} (\overset{pL}{x_{i1}} - \underset{mean(f_1)}{\mu_1}) (\overset{pw}{x_{i2}} - \overset{mean(f_2)}{\mu_2})$$

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



$$Cov(f_1, f_2) = \underbrace{(f_1^T f_2)}_{f_1 \cdot f_2} * \frac{1}{n}$$

if f_1 & f_2 have been std,

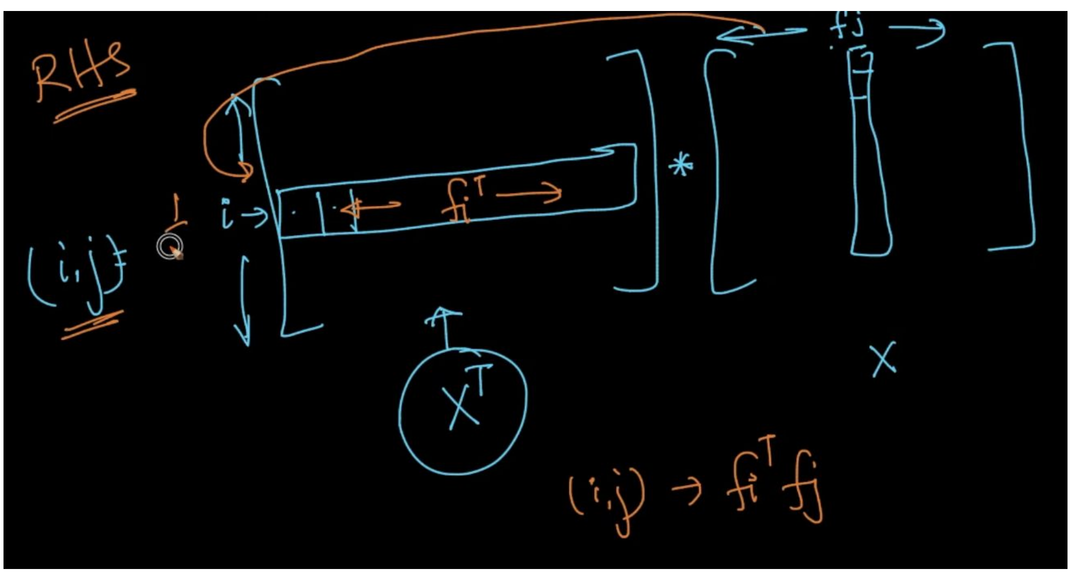
$$cov(f_1, f_2) = \frac{f_1^T f_2}{n}$$

$$S_{d\times d} = \frac{1}{n}(X^T)_{d\times n}(X)_{n\times d} = (d\times d) \checkmark$$

data-matrix

assuming X has been col. std

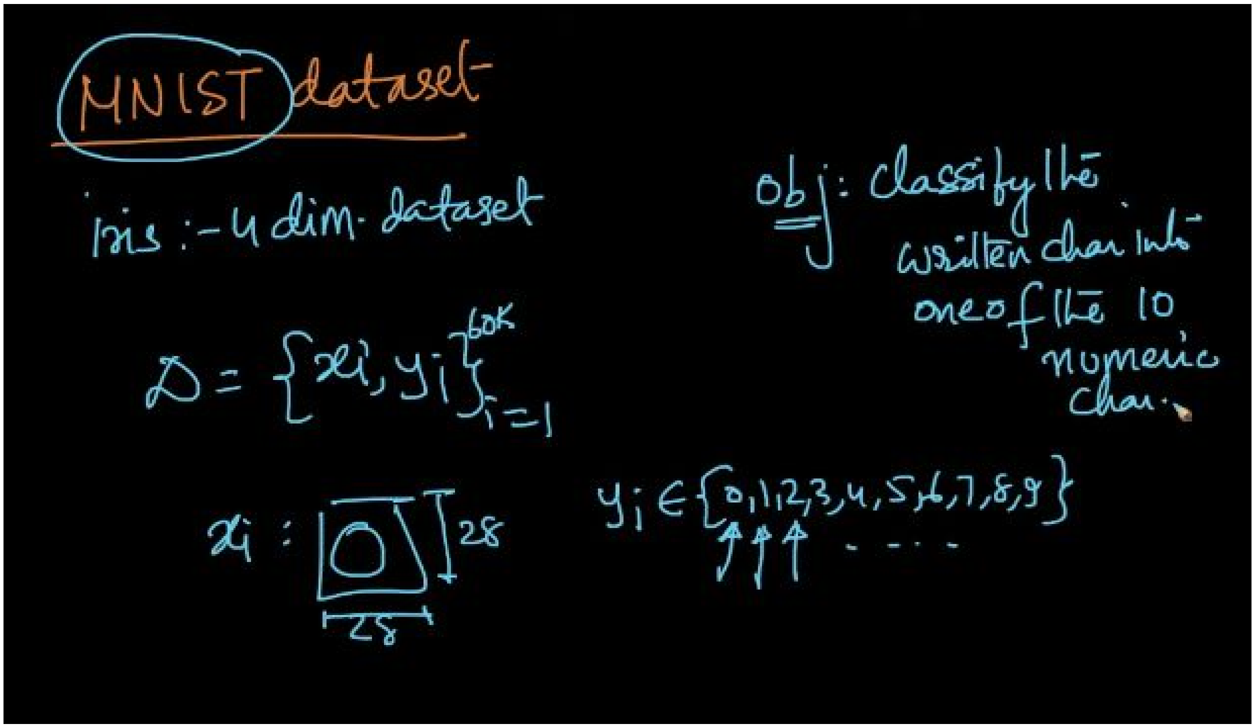
$$S_{ij} = cov(f_i, f_j) = \frac{f_i^T f_j}{n}$$

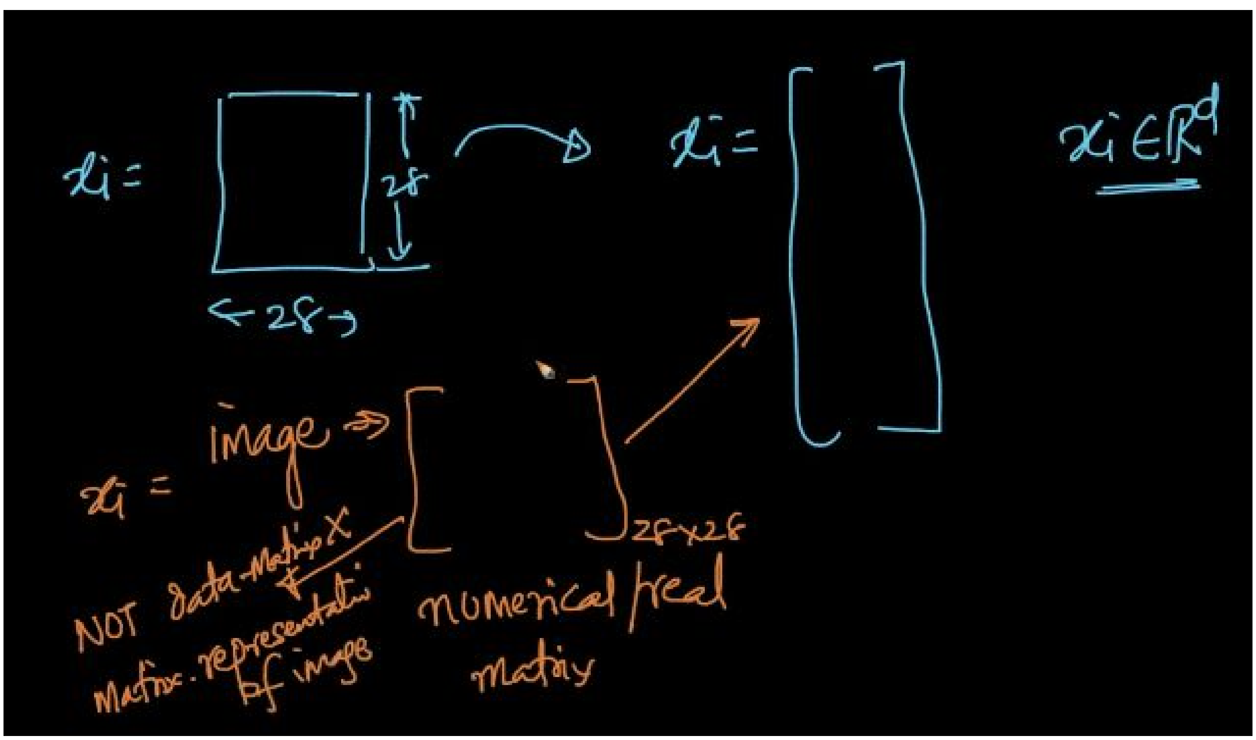


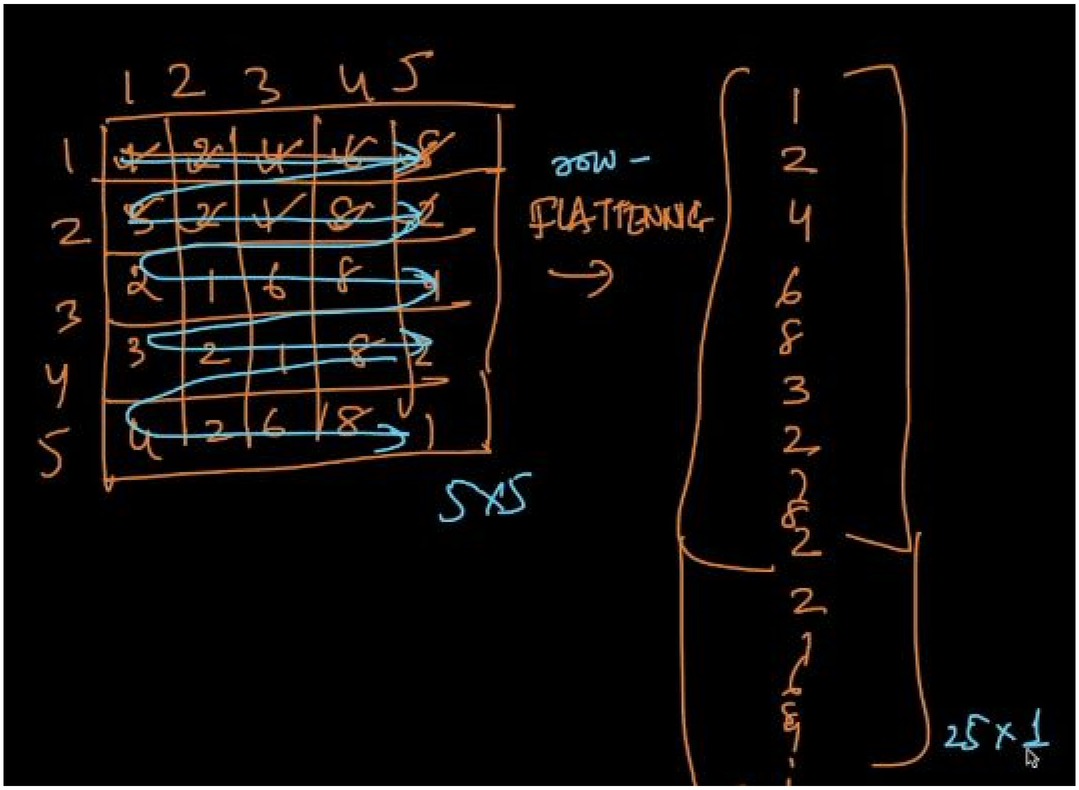
MNIST Data Sets

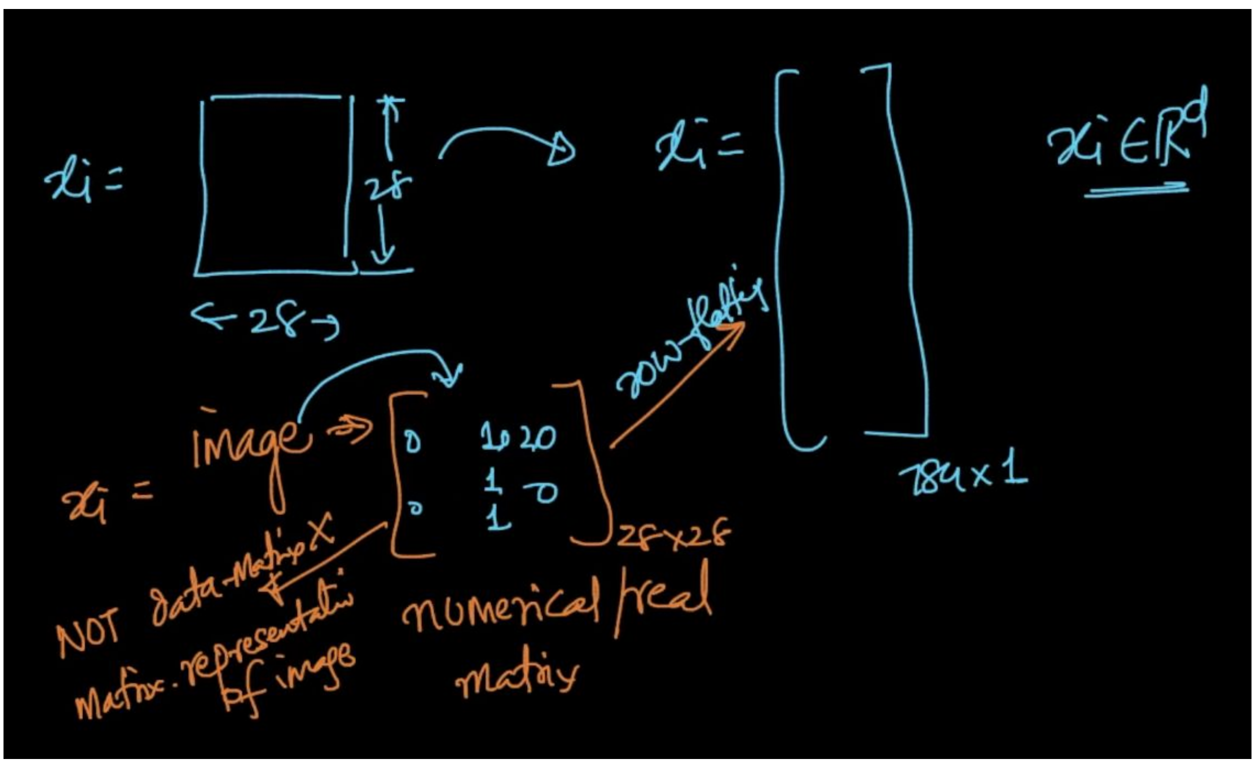
Additional References:

http://colah.github.io/posts/2014-10-Visualizing-MNIST/






1 2 3 4 5
row -
FLATTENING
5X5
25 x 1


xi =
28
← 28 →
xi =
xi ∈ R^d
row-flattening
784 x 1
xi = image
28x28
NOT data-Matrix X
Matrix. representation of image
numerical/real
matrix