1. what is Unsupervised ML
2. what is clustering
3. Application of the clustering
4. Clustering algo

↳ K-Means → Elbo method
- inter Cluster Dis
- intra Cluster dis
- WCSS.

diff b/w { k-mean and k-mean++ }

→ how we can evalute this Clustering.

↳ Hirarchical clustering → Agglomerative → Single linkage method. → Point matrix.
→ Divisive
→ the concept of dedogram (for representation of cluster)

↳ DBScan clustering
- → eps dis
- = No. of Point
- Core Point

→ Bound, Point
→ Noise Point

Machine learning :- Supervised
Unsupervised
Semi Supervised

Supervised :- Data → X , y ← Someone for supervising
X → independent
y → dependent var (target) (Supervisor)

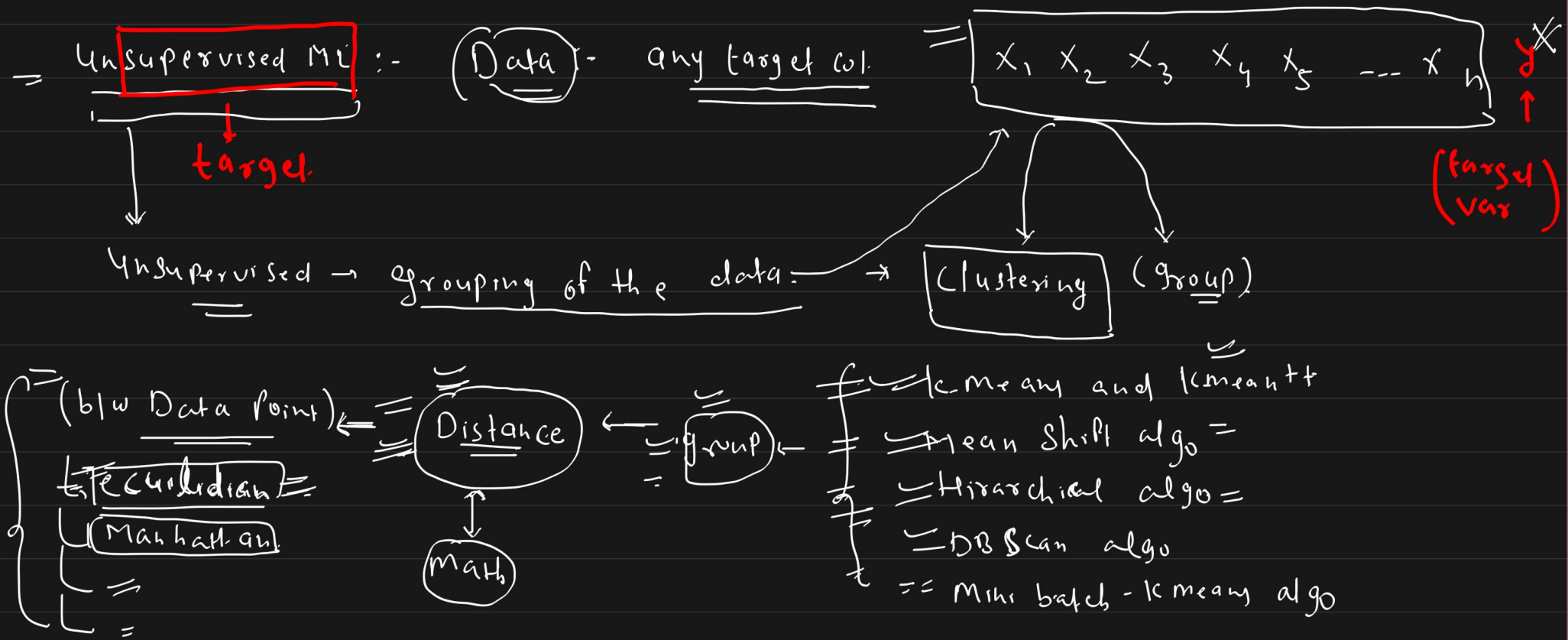
- Lin reg
- Log reg
- Svm
- KNN
- DT
- RF
- AB
- GB
- XB

y → Num → Reg.
y → Cat. → Class.

Machine learning
1. Data.
2. EDA
3. Preprocess
4. Model
5. evaluation

(Data.) Supervised & Unsupervised
Many - End to end Project

Supervised :- Regression
Classification.



Clustering → grouping of data → Distance → Euc.
→ Manh.

Supervisor → Data - group - y

Application:- Real time app:-

1. Customer seg → ecommerce → Similarity → Amazon → Customer
2. Image seg.
3. Data seg. (Sentiment Data)
   - Review → Pos, Neg, Nut
   - Segregate a Similar Pixel

{ Female, male, Payment, Stuff, review } → real world Pro

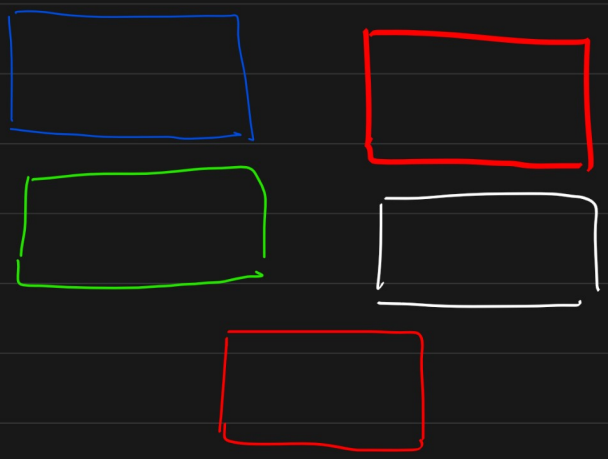
→ Pixel Segmentation
grouping of the Pixel

3. Structure format :- Medical
   Agg.
   education.

X_1 X_2 X_3 X_4 ... X_n

$(m \times n)$

$m = 10000$
$n = 10$



Clustering algo :- **K-Means**

(1) Concept of the centroid

(2) Distance

— Initial

{ Within cluster — (i) intra.
  b/w cluster — (ii) inter }

Evaluation of Cluster →
- WCSS → elbow Plot
- Shilloutee
- Dunn' index.
- !
- many more

↓ intra and inter

→ end goal.

K-Means algo :-

Data → group → cluster → Supervisor (not having any)

X_1 → X_2 (y) → (algo) Supervised

110-Men (grouping) → (3) L, (4) S, (3) M

- K-means, K-means ++

	X_1 (height)	X_2 (weight)
1	180	60
2	170	50
3	160	40
4	165	30
5	188	45
6	160	65
7	167	75
8	190	80
9	195	90
10	200	100

→ Data → group → K-means ++

- (no. of cluster) = No. of centroid
- (mean) (avg)
- (Conversion)

① Calculate a centriod (data point) = (Random).

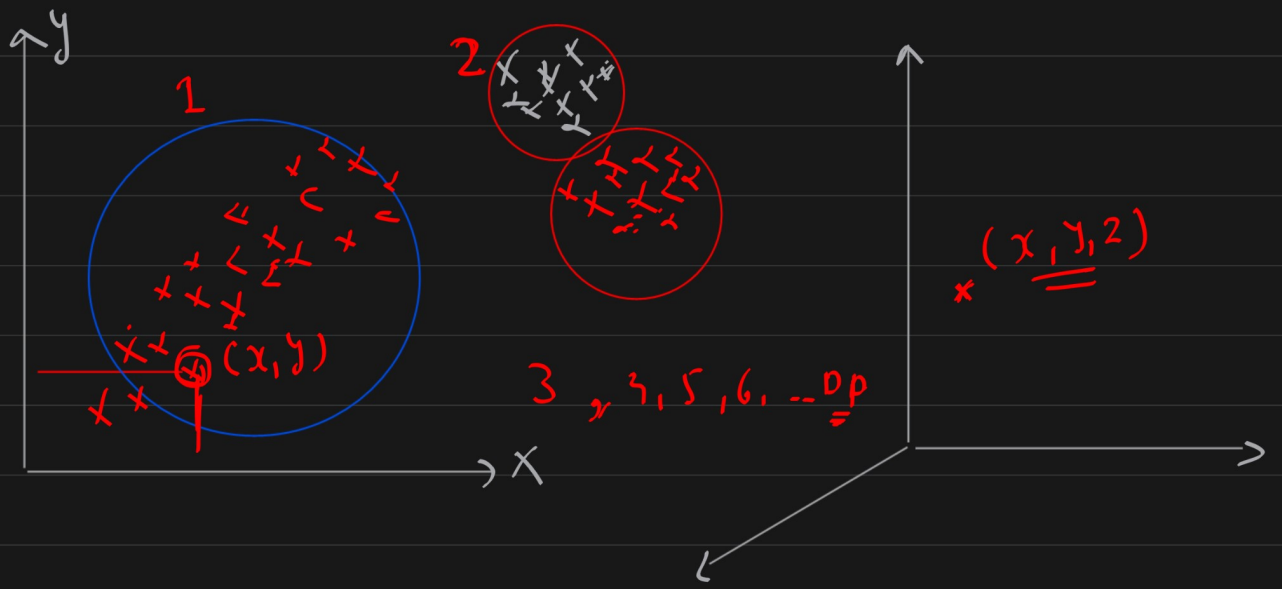
Minimum :- One cluster

Maximum :- No of data Points (No. of Row)

Hirar.

4-D 2D 3d

(x_1, x_2, x_3, x_4) (x_1, x_2) (x_1, y_1, z)



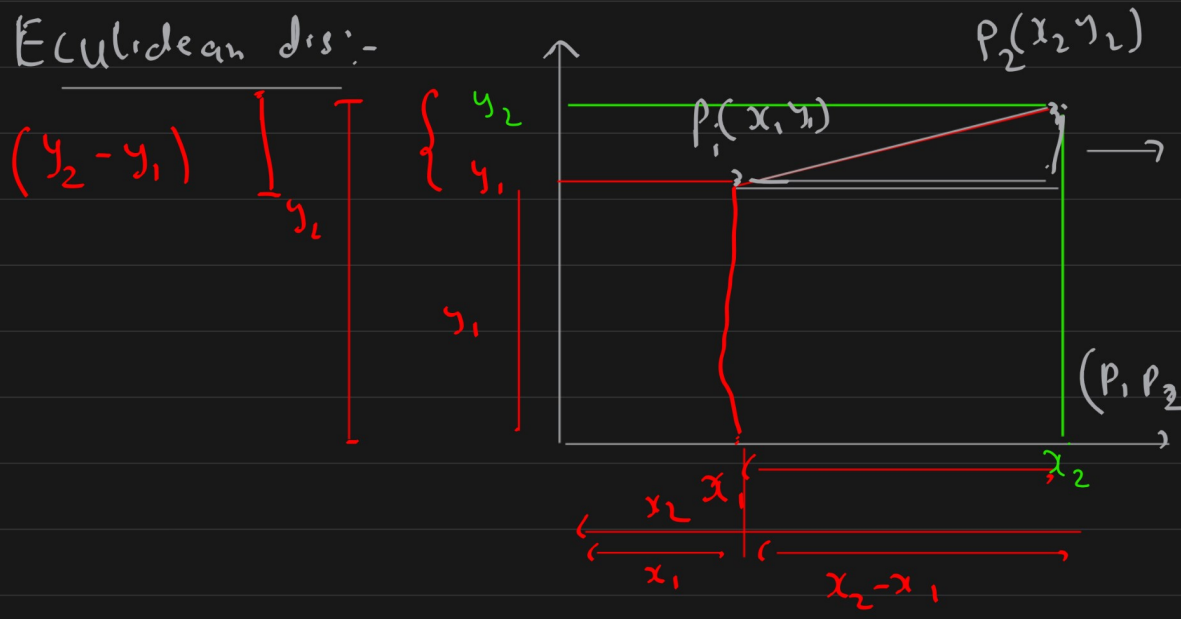
Minimum two :- K = No of centroid. K = 2.

(Randomly)

$C_1 (180, 60)$ $C_2 (170, 50)$

Distance

Ecludean dis:-



$$H^2 = B^2 + P^2$$

$$H^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$$

$$(P_1 P_2) \quad H = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

(Euc Dis)

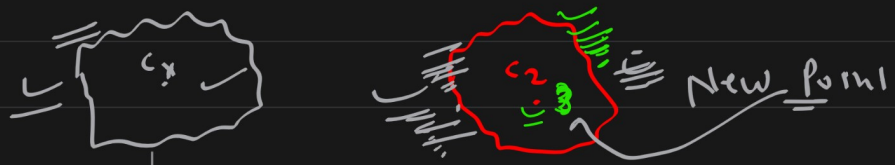
1.) initiate a centroid (Randomly)
2) Distance
3) fill the cluster
4) update the centriod.

$C_1 (180, 60)$ $\quad$ $C_2 (170, 50)$

3 Point → $(160, 40)$

$= C_1 + 3 \quad \sqrt{(160-180)^2 + (60-40)^2} = (28.28)$

$= C_2 + 3 \quad \sqrt{(170-160)^2 + (50-40)^2} = 14.14$



$(180, 160)$ $= (170, 50)$ $(160, 40)$

$$\left(\frac{170+160}{2}, \frac{50+40}{2}\right)$$

$$(165, 45)$$

K-means

K-means++

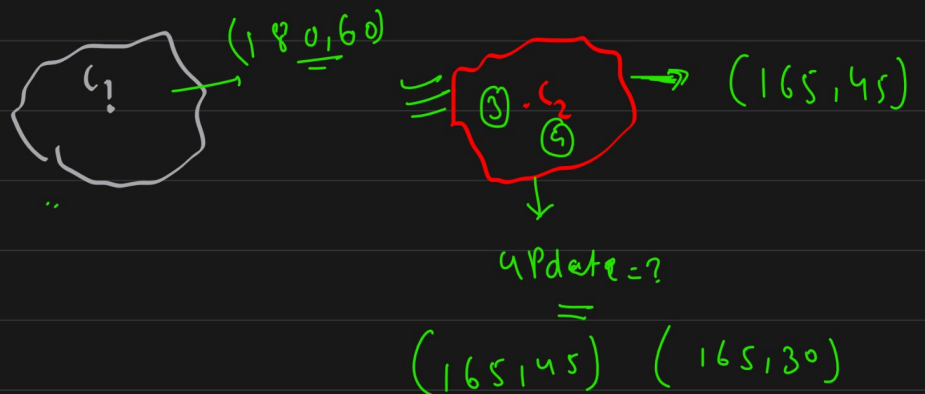
Mini batch · K-means.

C_1 $(180, 60)$ $= C_1$

C_2 $= (165, 45) = C_2$ → (updation.) $(165, 30)$

C_1 to 4 $= \sqrt{(180-165)^2 + (60-30)^2} = 33.54$

C_2 to 4 $= \sqrt{(165-115)^2 + (45-30)^2} = 15$ (52.20)



$\left(\frac{165+165}{2}, \frac{45+30}{2}\right)$

$C_1(180, 60)$ $\left(165, 37.5\right)$ C_2

ELBO Plot :- ⇒ K-means or K-means++

↓

No. of cluster how we can decide k-value?

k = 1
k = 2
k = 3
k = 4
⋮
k = No. of Data Points

Distance → intra and inter

↓

With in cluster (WC)

(from centroid)

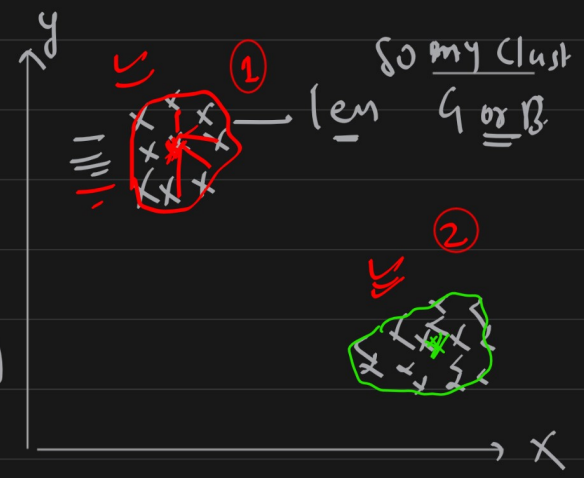
WCSS → With in cluster Sum of Square

(Dense) len intra dis (with in cluster)

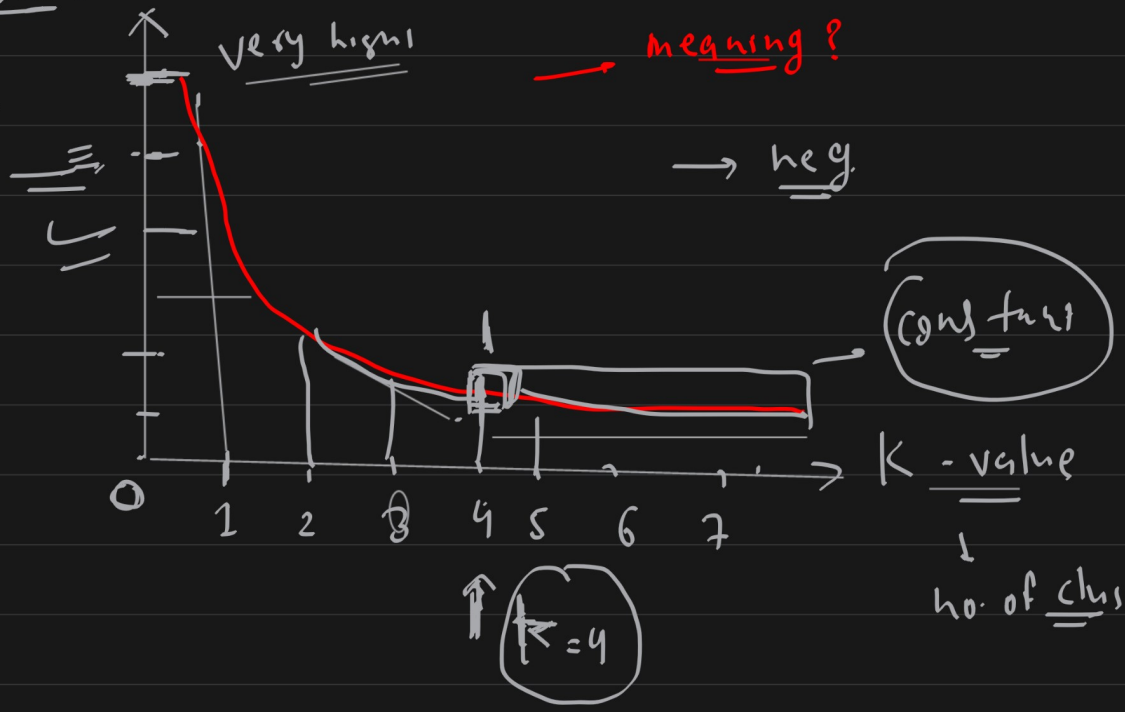
↓

good.

More dis → Bad (Sparse)



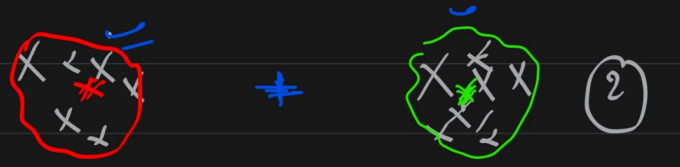
WCSS



K=4 ← elbo

Data Point :- (10)

WCSS → ∪



WCSS → (len)

Evaluation K = 4 → correct ? over there (ratio) within and inter

Some other method

WC b/w cluster

Mathematical

1. Dunn index.
2. Silhouttee score
3. Jaccard index

WCSS / Matrix → k = 4 → Validation → silhouette, dunn, ...

↳ intra, inter : Mathmatical — How much strong (it is good or bad)

↳ (Semi supervised): all Dots

↳ Hirarchical

DBScan

↳ Hirarchical clustering → grouping of data → (k-Means)

↳ Hirarchy of cluster

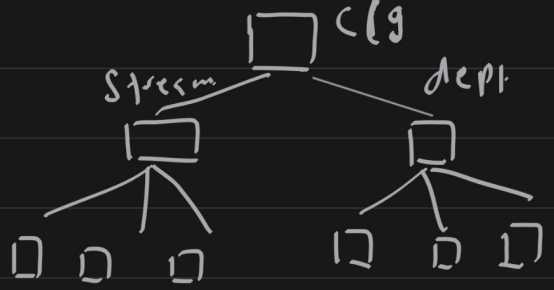
↓

· Cluster

↓

↳ Distance

(Euclidean)



- Hrarchical :- agglomerative clustering → Bottom to top → No. of cluster = No. of dataPoint
  Divisive clustering — top to down → Single cluster

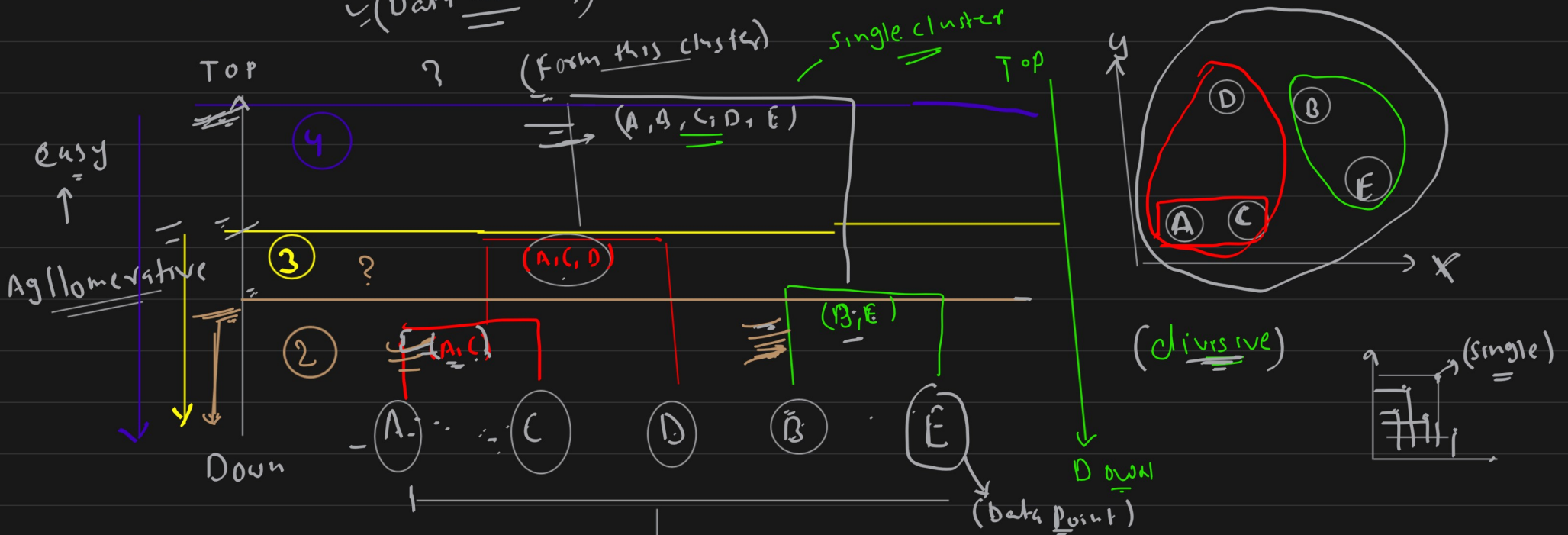
Data :- N of row :- minimum cluster :- 1
maximum cluster : No of data Point

Dendogram :- Diagram which is representing cluster

K-Means and hirarchical

Define

(Data hirarchy)



↓

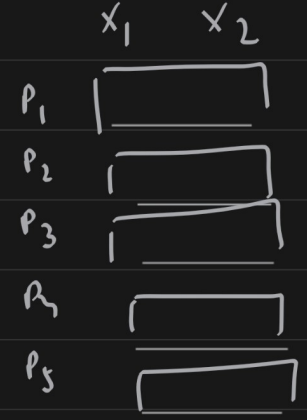
(Single cluster) (individual)

Practical

Hirarchical :- only once define hirar ?

x_1 (H)	x_2 (W)
150	60
170	50
160	40

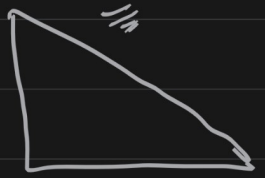
$k = 4, 5, 6, 7, 8 \ldots$



Point matrix :-

Assuming ⇒

Heat map

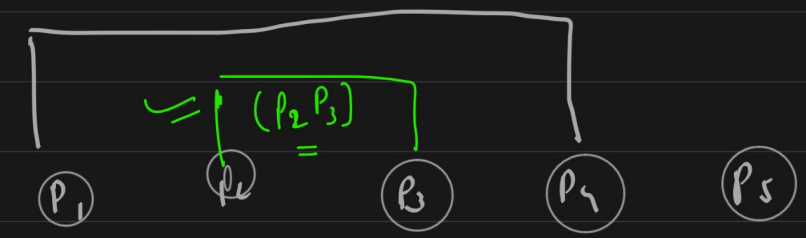


	P_1	P_2	P_3	P_4	P_5 (redunt.)
P_1	0				
P_2	4	0			
P_3	5	2	0		
P_4	8	7	11	0	
P_5	10	9	12	13	0

① P_3

↓ matrix (distance)

↓ Dendogram



dendogram. ← point matrix.

-	P_1	(P_2, P_3)	P_4	P_5
P_1	○			
(P_2, P_3)	9	○		
P_4	8	18	○	
P_5	10	25	13	○

$d\left(P_1, (P_2, P_3)\right)$ – Single linkage

$d\left[(P_1, P_2)\ (P_1, P_3)\right]$

$d[4, 5]$

$d[9]$

$d\left(P_4, (P_2, P_3)\right)$

$d\left[(P_4, P_2)\ (P_4, P_3)\right]$

$d[7, 11]$

18

$d\left(P_5, (P_4, P_3)\right)$

$d\left[(P_5 P_4)\ (P_5 P_3)\right]$

$d[13, 12]$ – 25