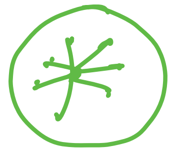
[a] A telecom company is using K-means clustering to segment customers based on monthly usage patterns. The company tested different values for the number of clusters, K, and recorded the **inertia** for each. In K-means, inertia measures the sum of squared distances between each data point x_i and the centroid c_k of its assigned cluster C_k: **[4] [CO1, CO6]**

$$Inertia = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - C_k\|^2$$

Inertia:



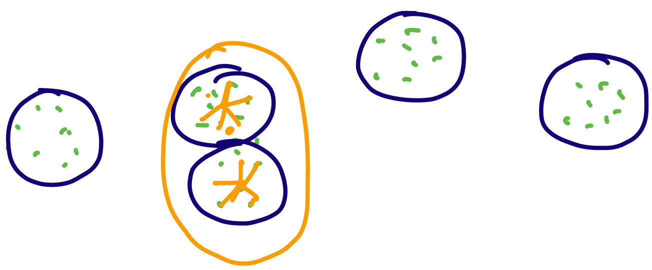
The inertia values recorded for different cluster counts K are shown below:

Table III

K	2	3	4	5	6
Inertia	3500	2500	1800	1200	1000

Explain inertia in K-means clustering and why it decreases as K increases. Determine the optimal K based on the inertia values, describing how the "elbow method" can help.

WCSS



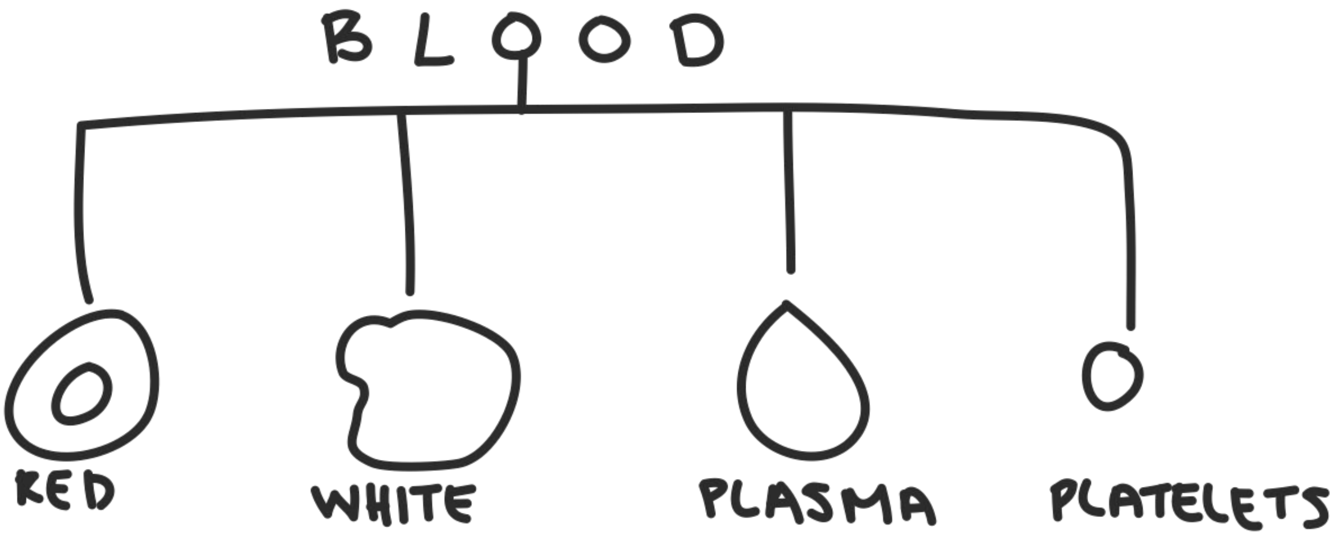
Hierarchial Based Clustering

Divisive (Top Down)

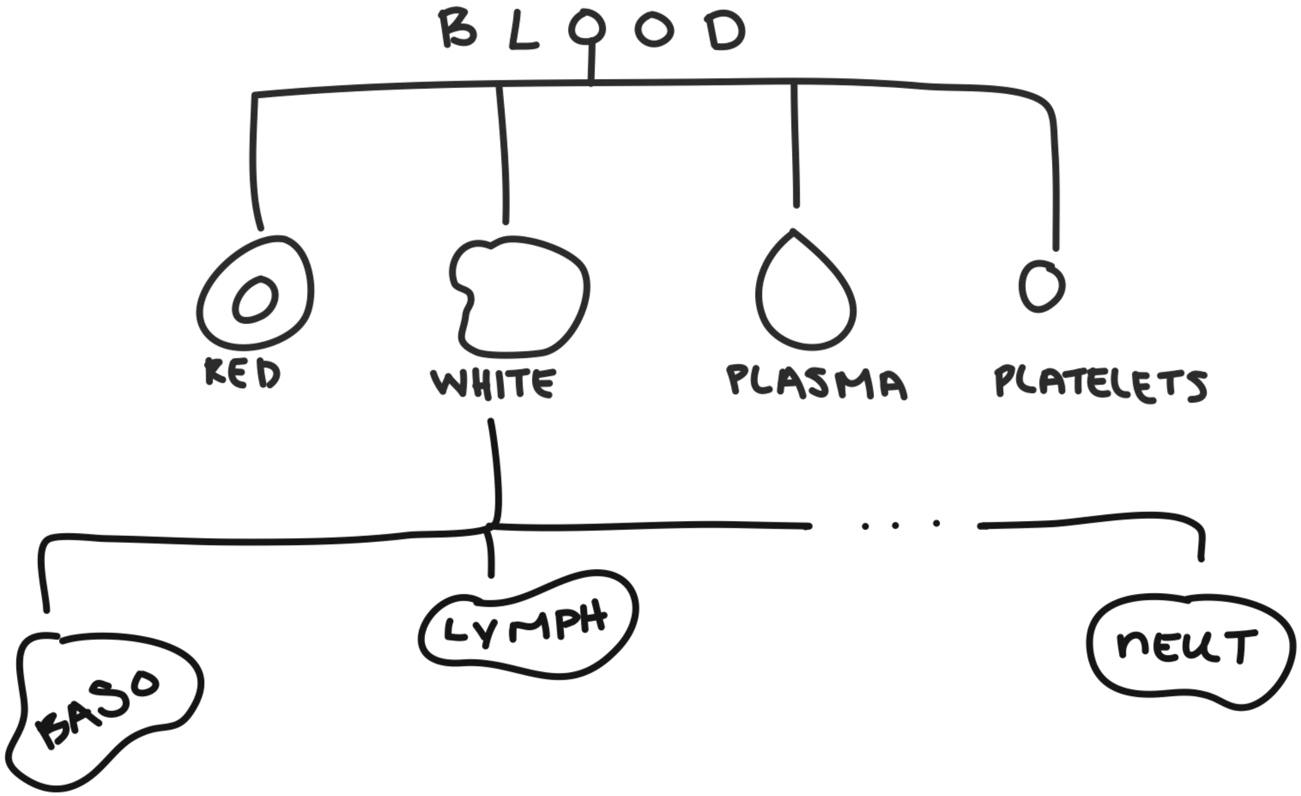
Agglomerative (Bottom up)

HAC

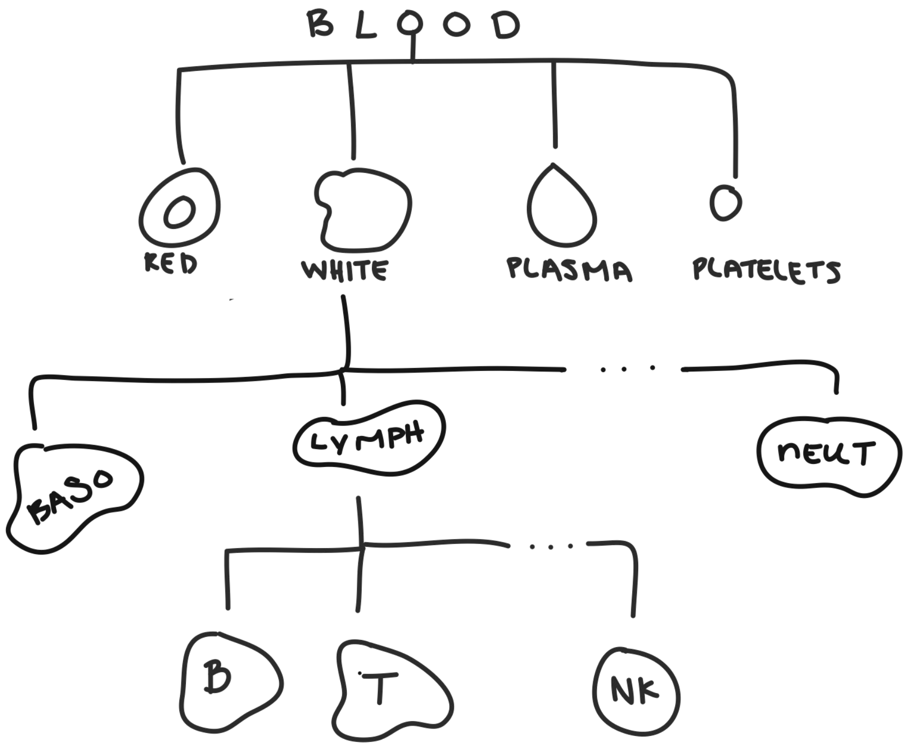
Divisive



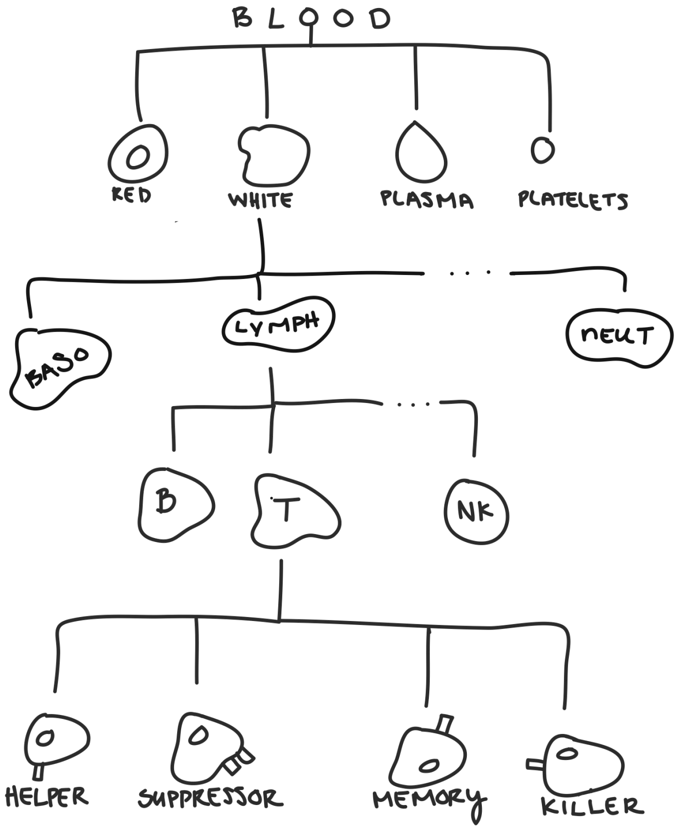
HAC



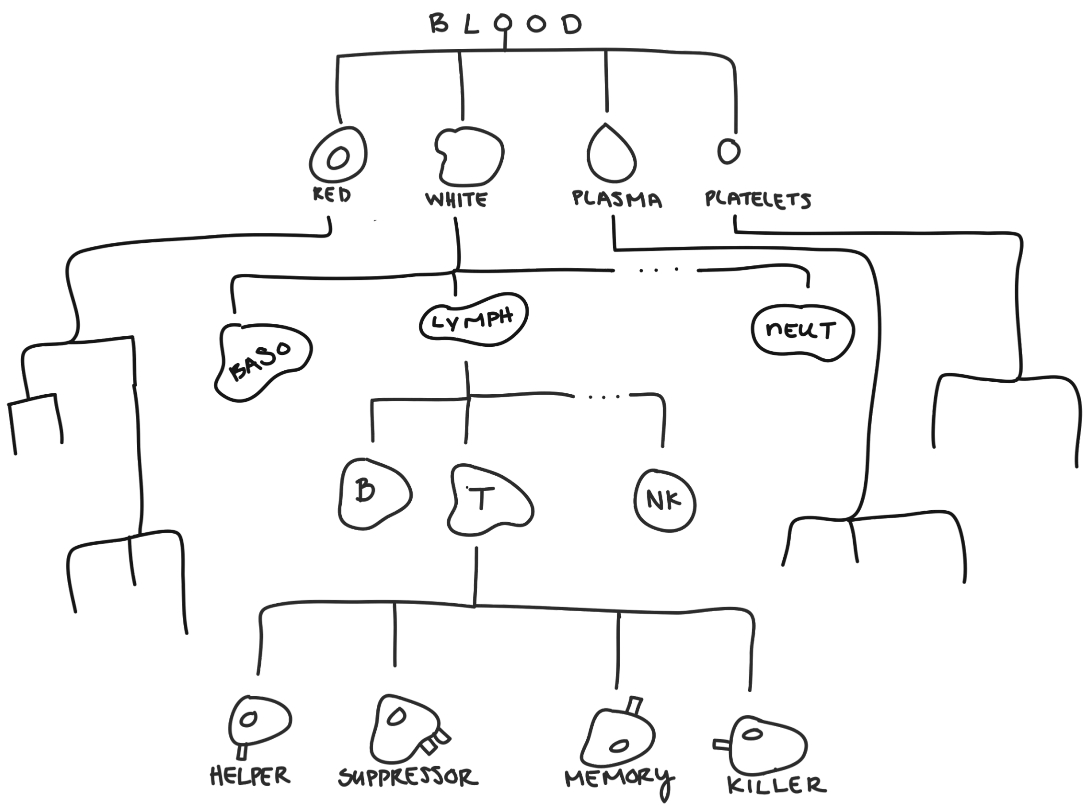
HAC



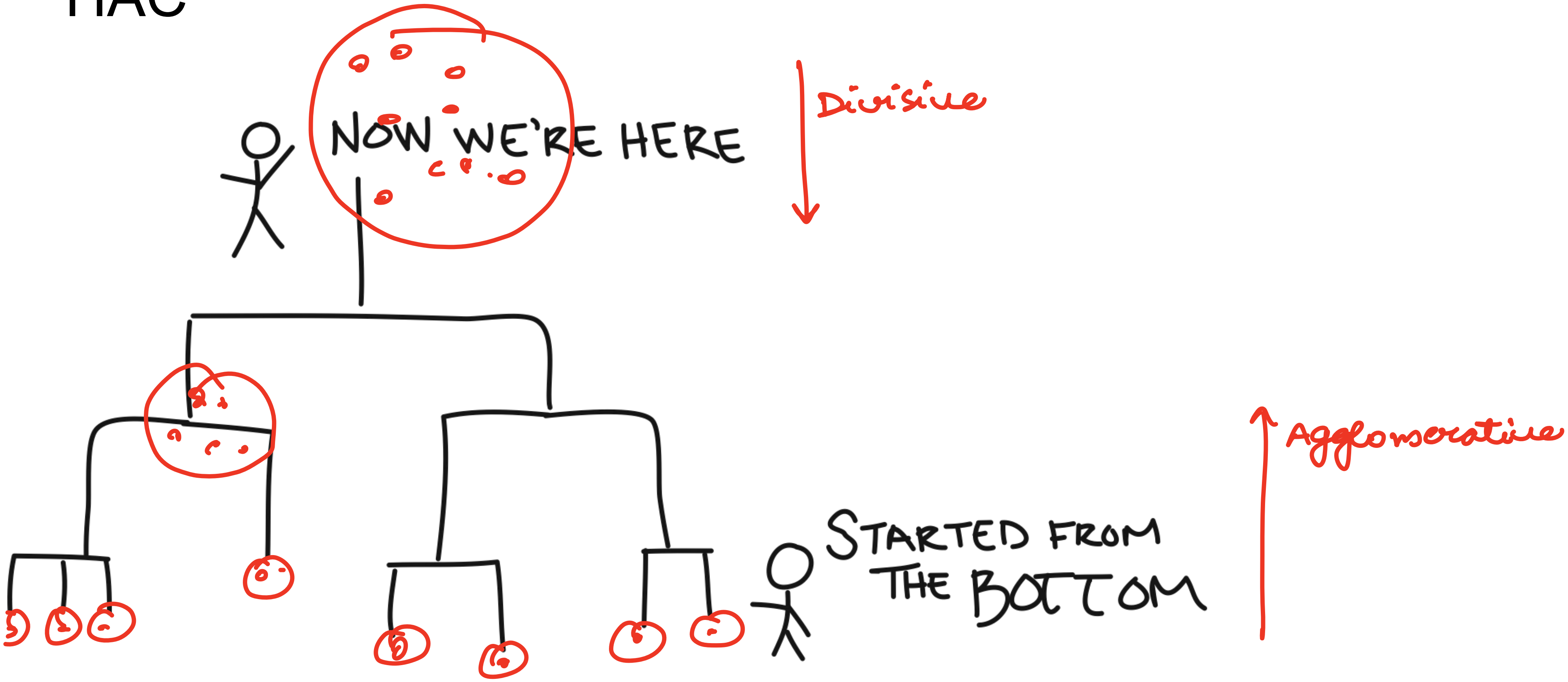
HAC


BLOOD
RED
WHITE
PLASMA
PLATELETS
BASO
LYMPH
NEUT
B
T
NK
HELPER
SUPPRESSOR
MEMORY
KILLER

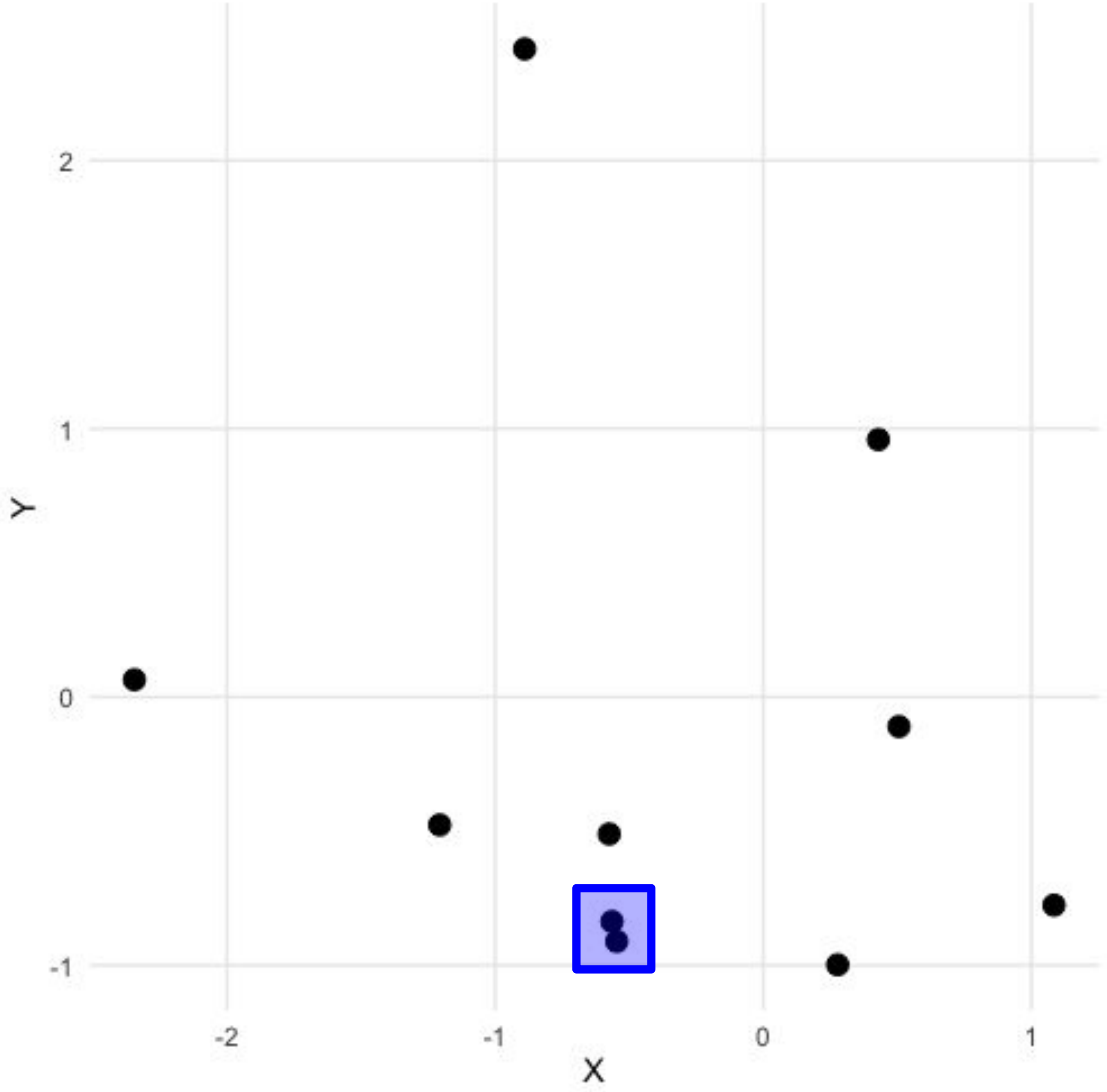
HAC


BLOOD
RED
WHITE
PLASMA
PLATELETS
BASO
LYMPH
NEUT
B
T
NK
HELPER
SUPPRESSOR
MEMORY
KILLER

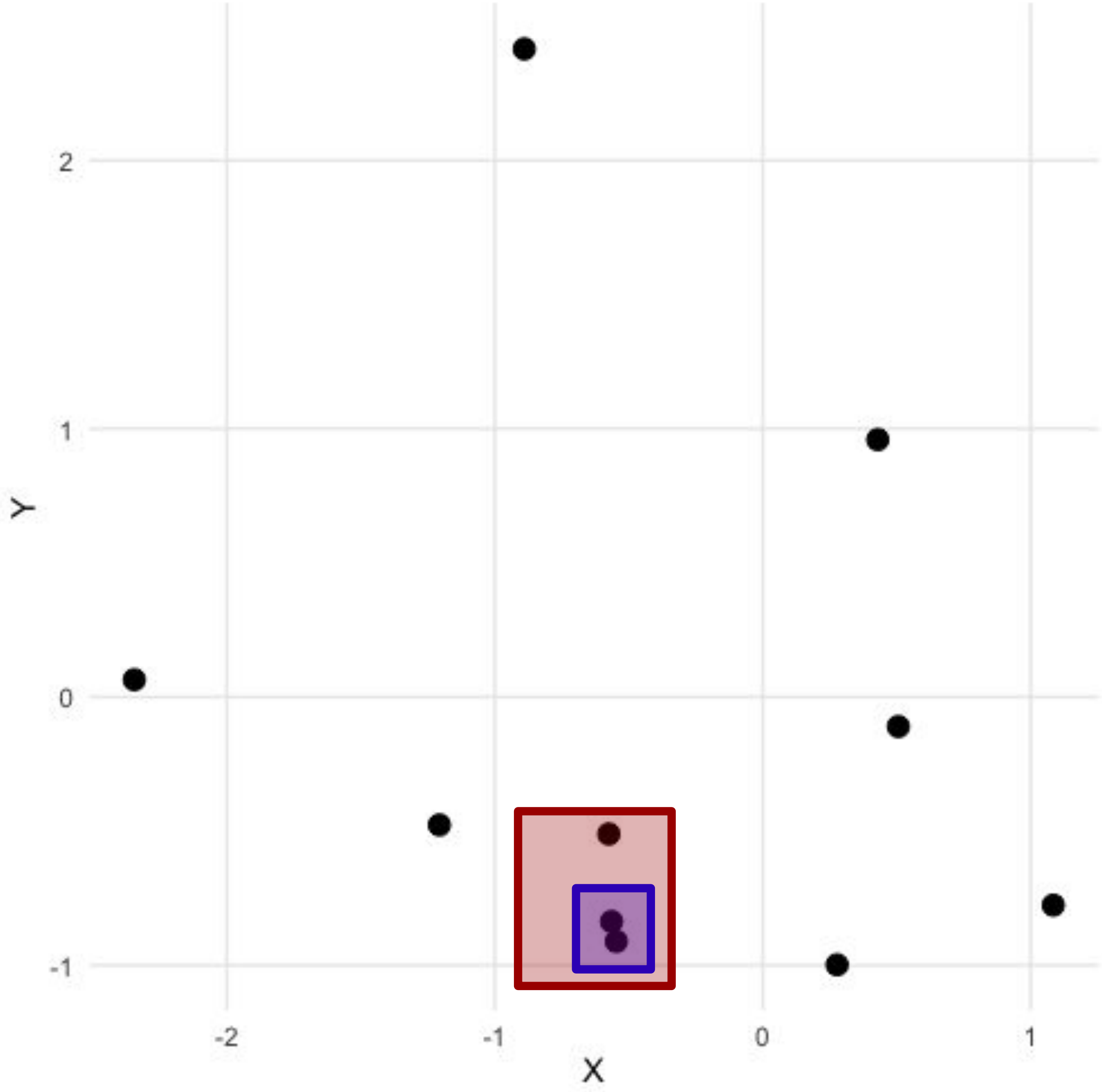
HAC


NOW WE'RE HERE
Divisive
Agglomerative
STARTED FROM THE BOTTOM

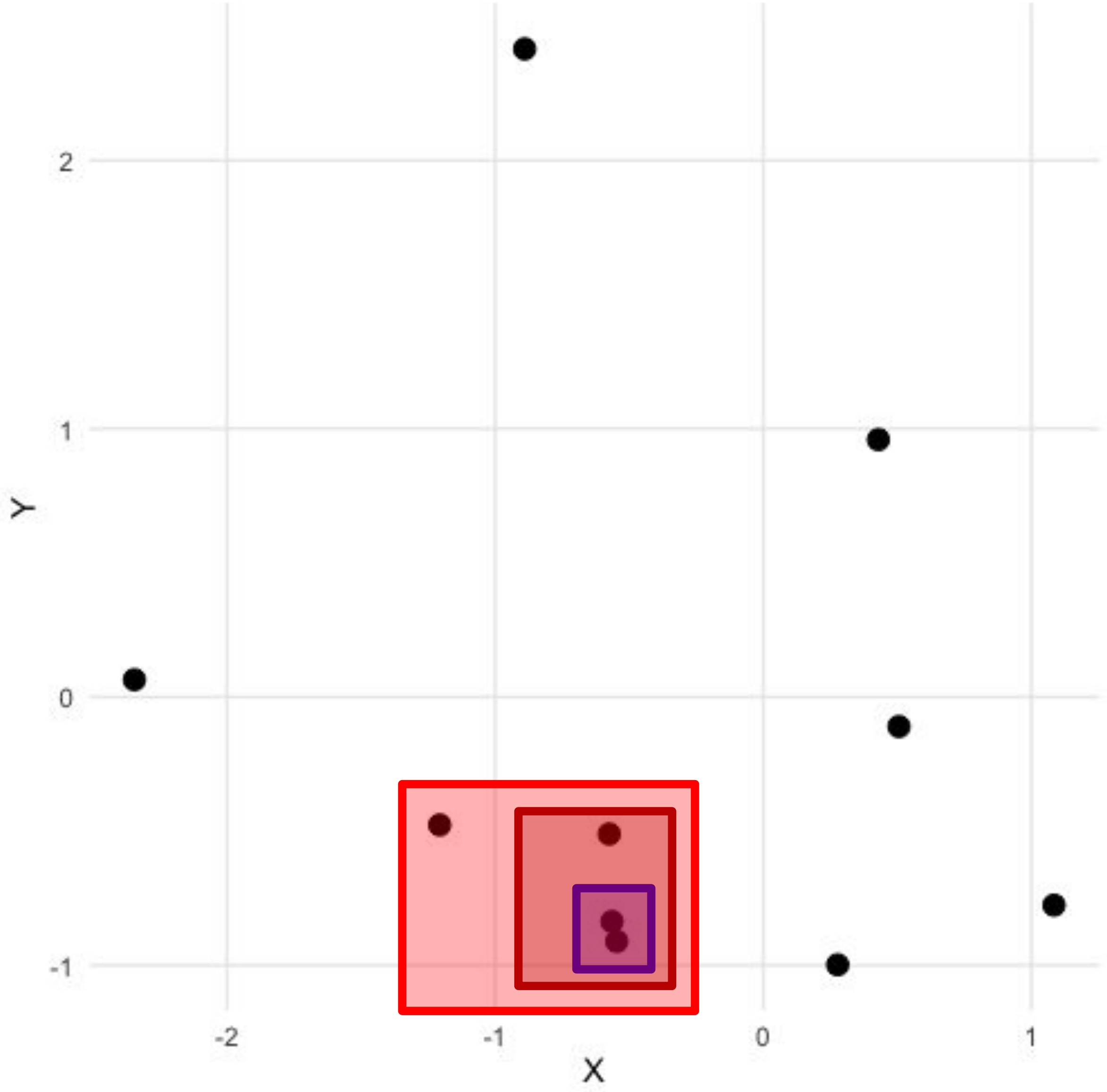
Algorithm



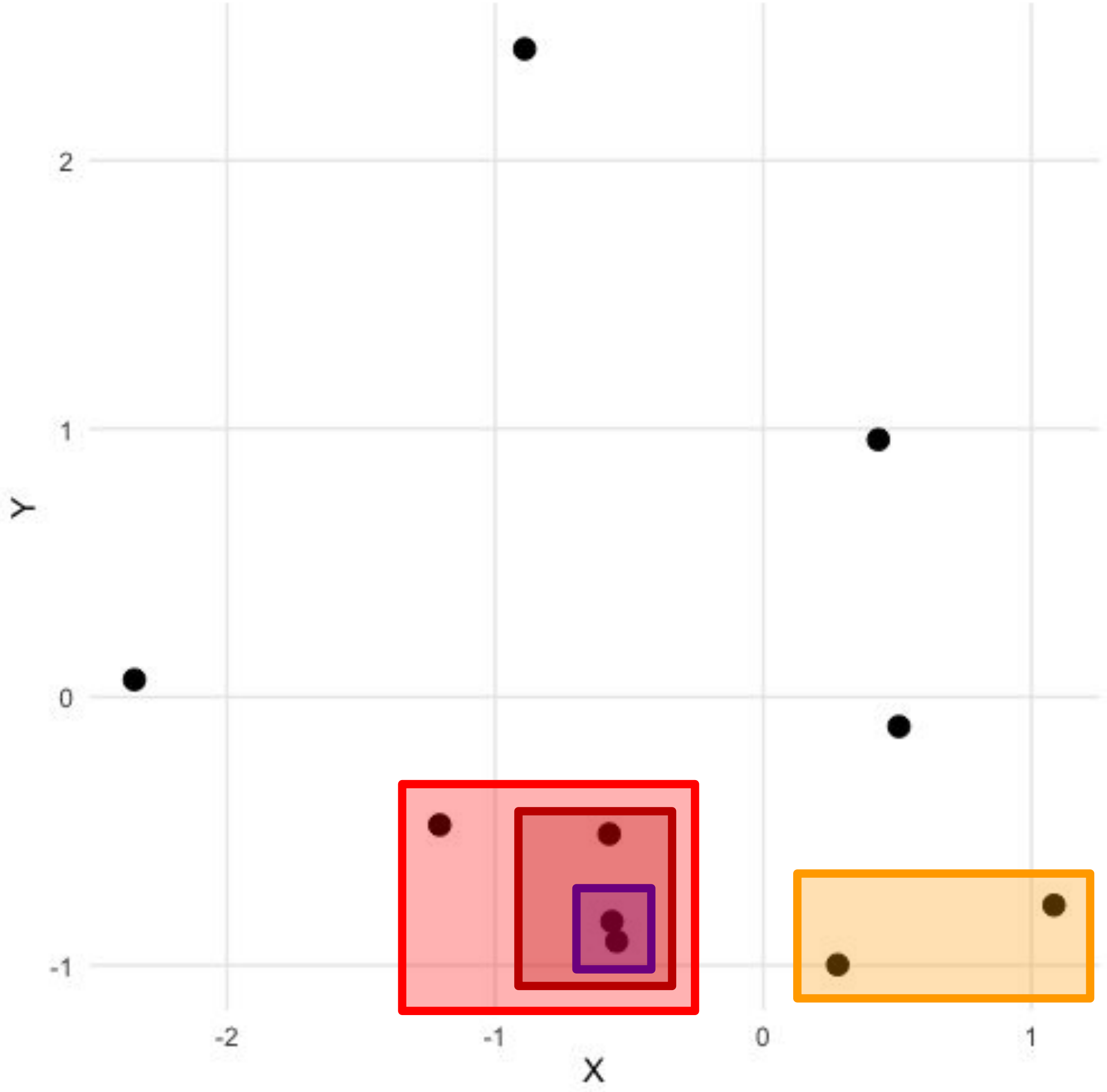
Algorithm



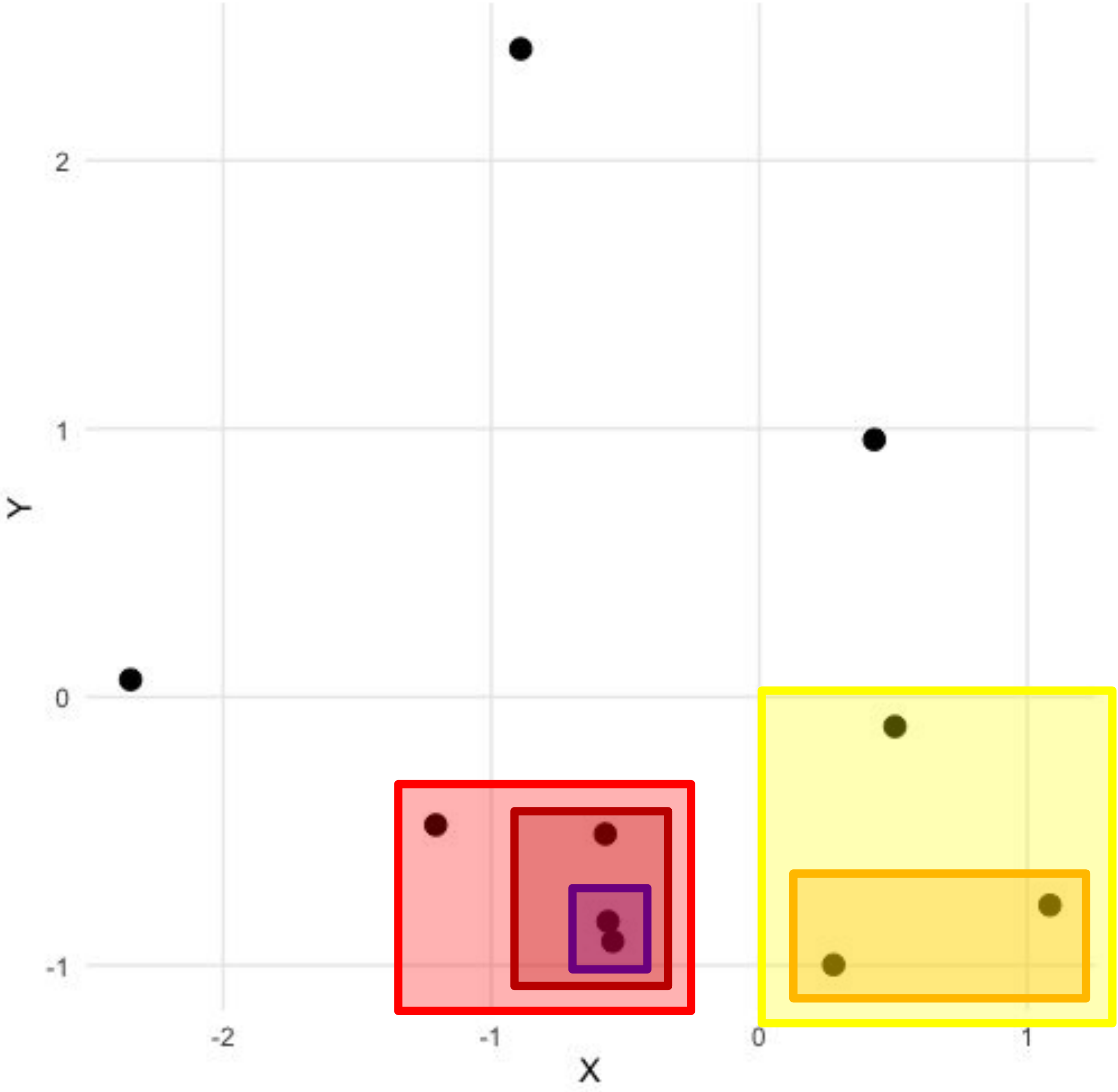
Algorithm



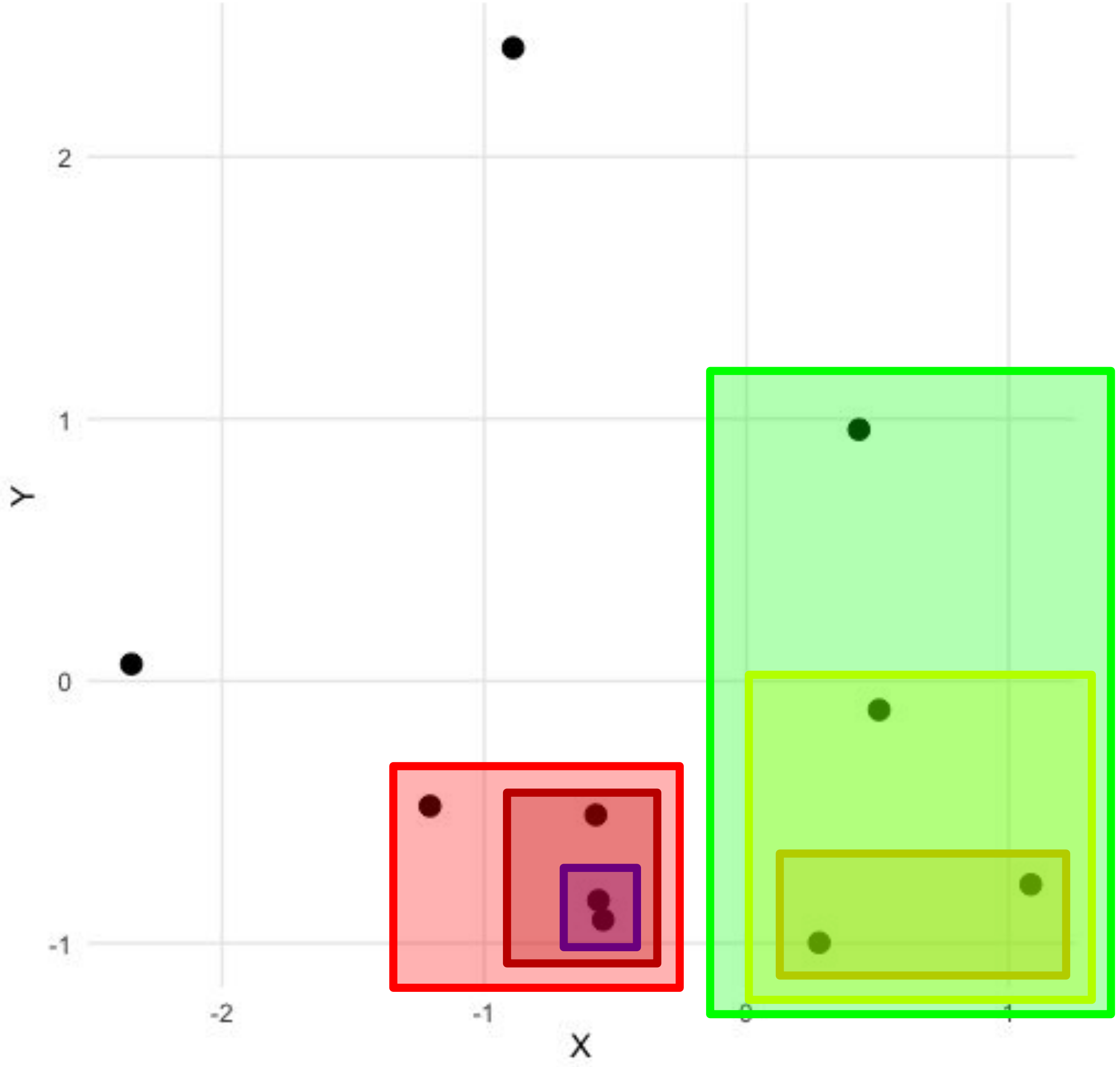
Algorithm



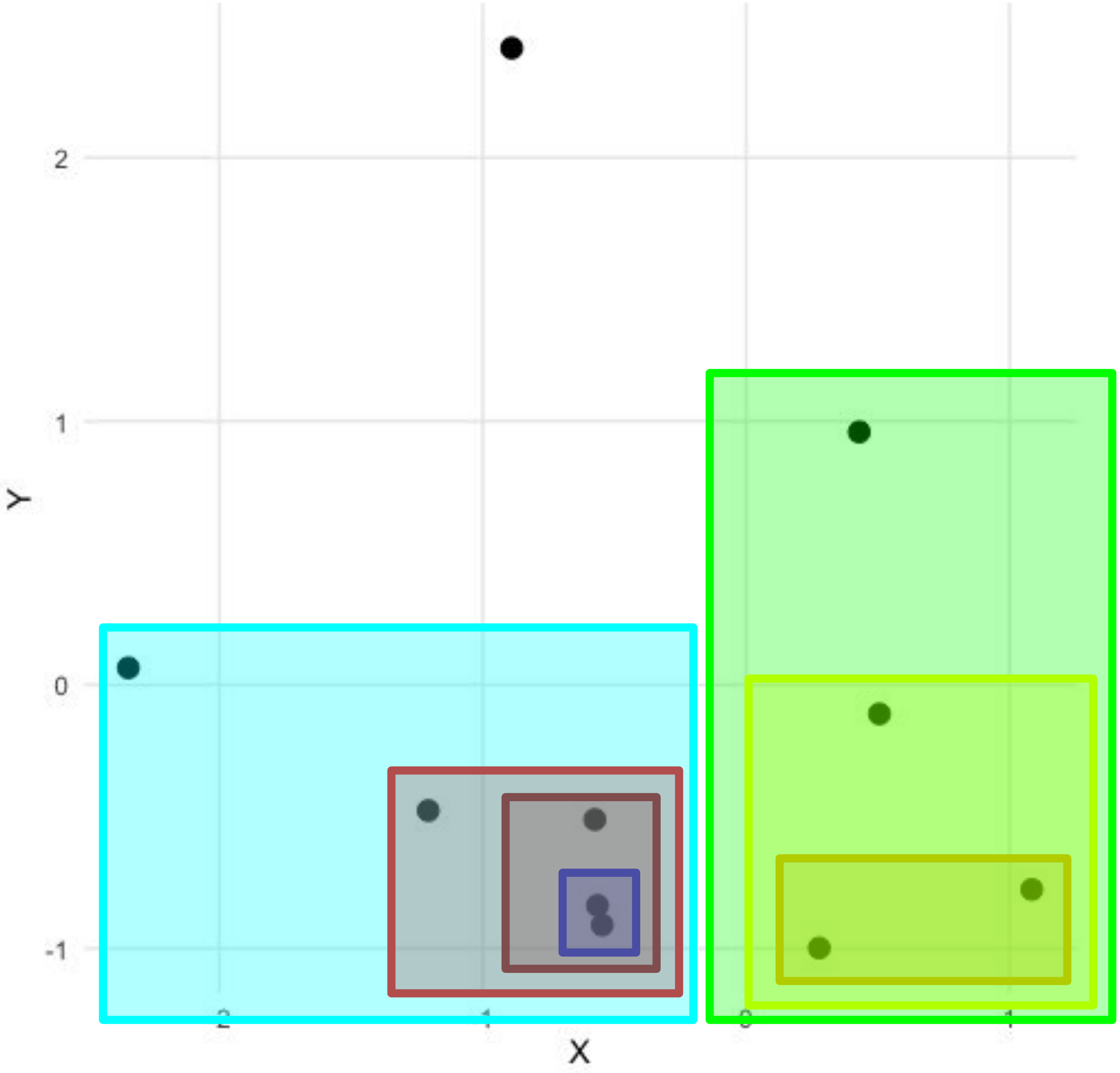
Algorithm



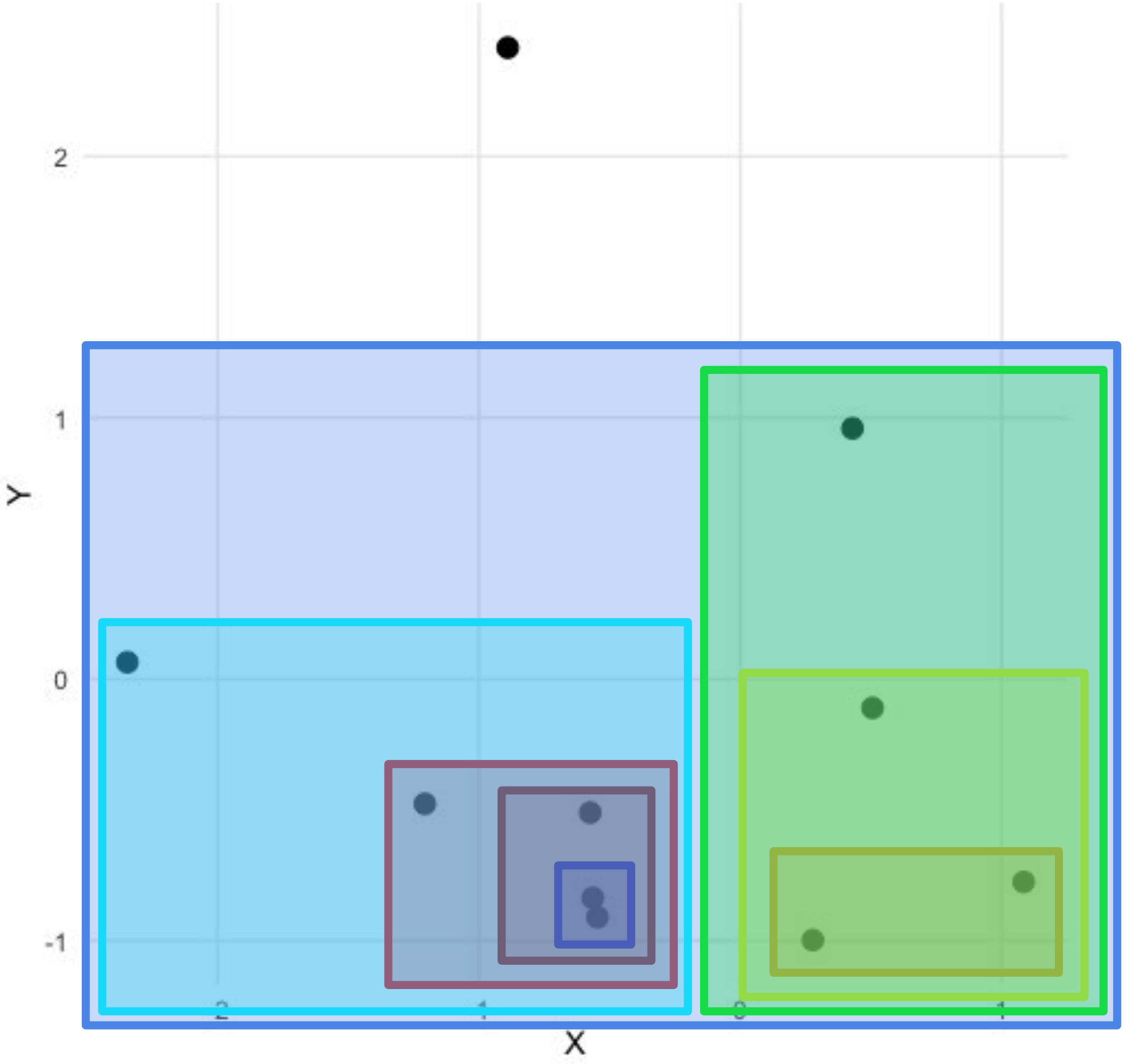
Algorithm



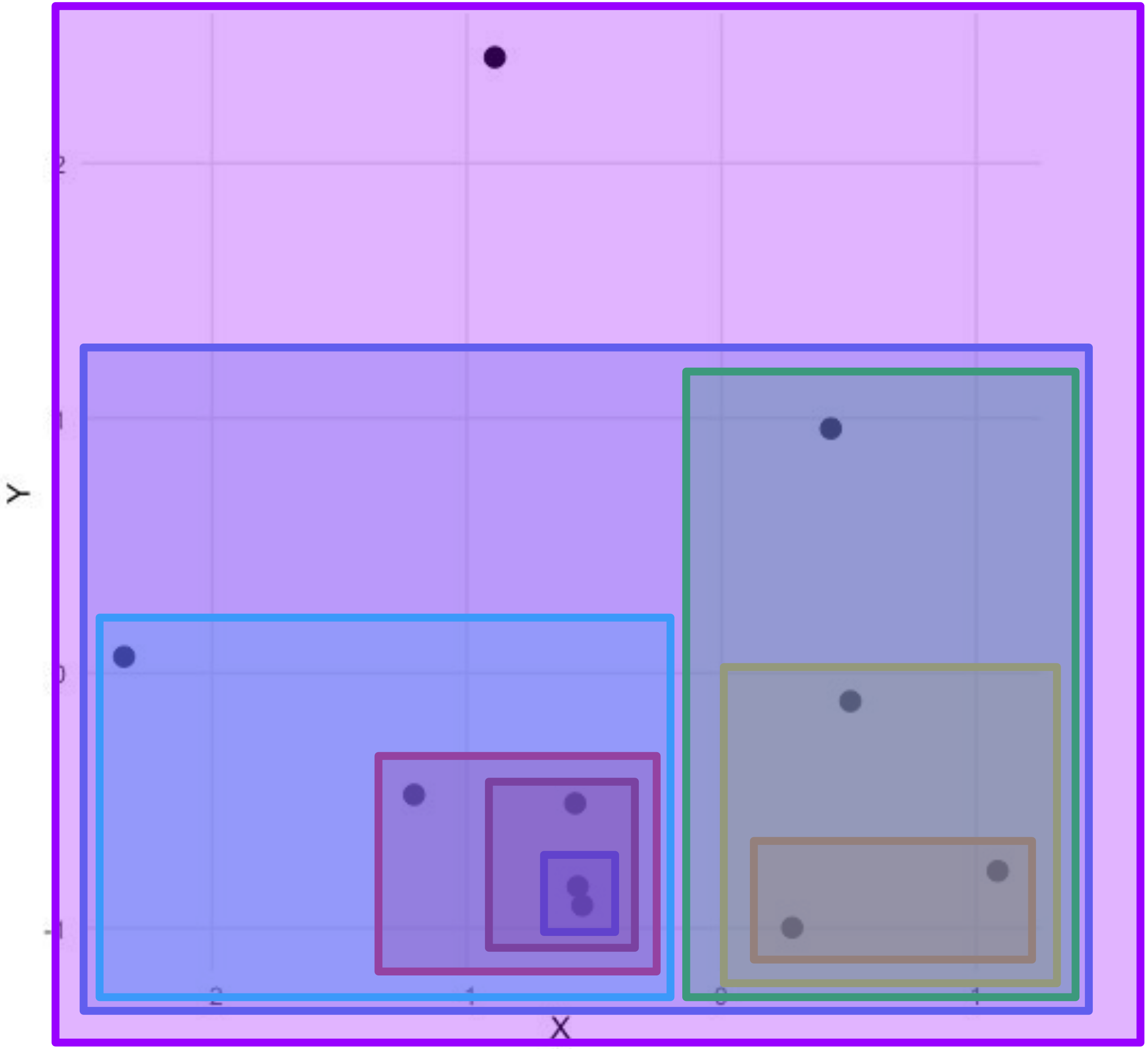
Algorithm



Algorithm

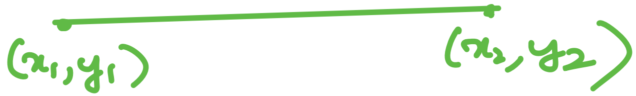


Algorithm



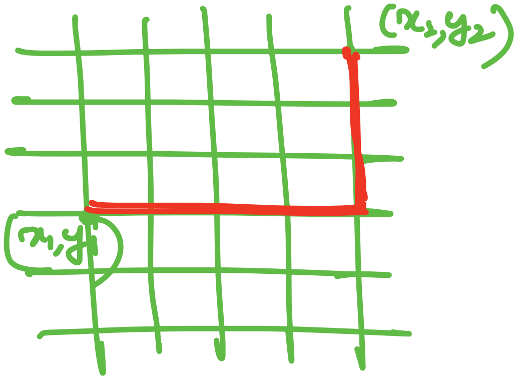
Distance Metrics

- **Euclidean**: Continuous Data
- **Manhattan**: High Dimensions
- **Hamming**: Categories
- **Cosine**: Word Counts

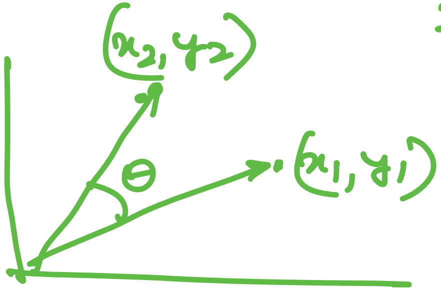


$$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$$(x_2-x_1)+(y_2-y_1)$$



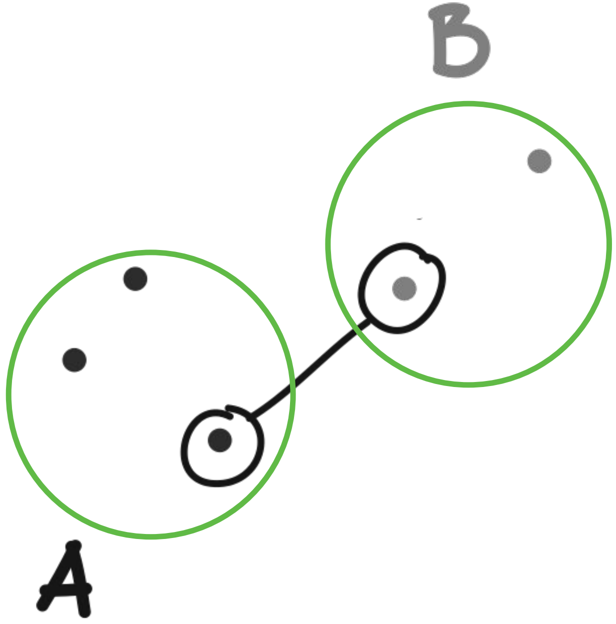
$$\begin{array}{cccc} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ \hline & & 2 & \end{array}$$



$\cos\theta$

Linkage Criteria

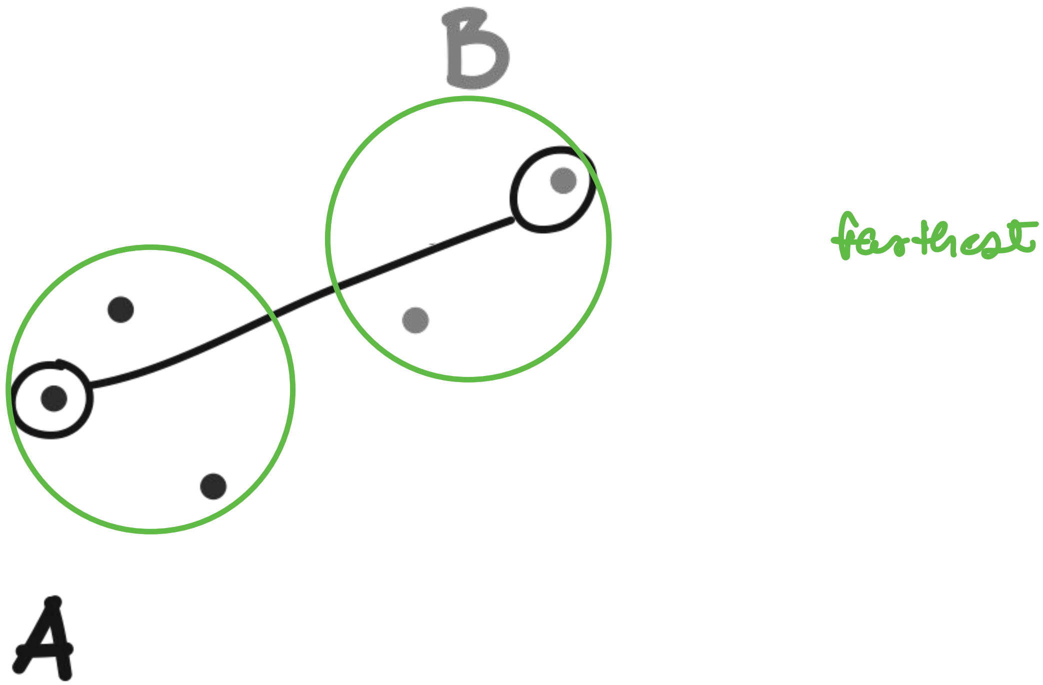
(Distance b/w 2 Clusters)



Closest

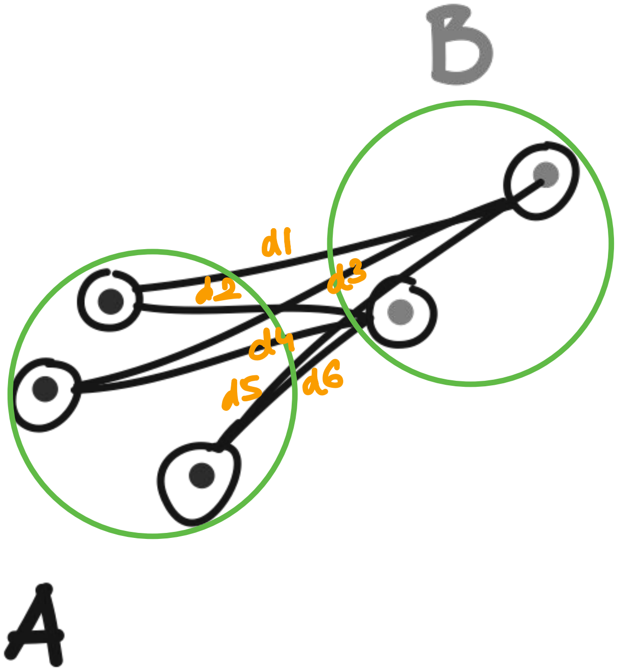
Single

Linkage Criteria



Complete

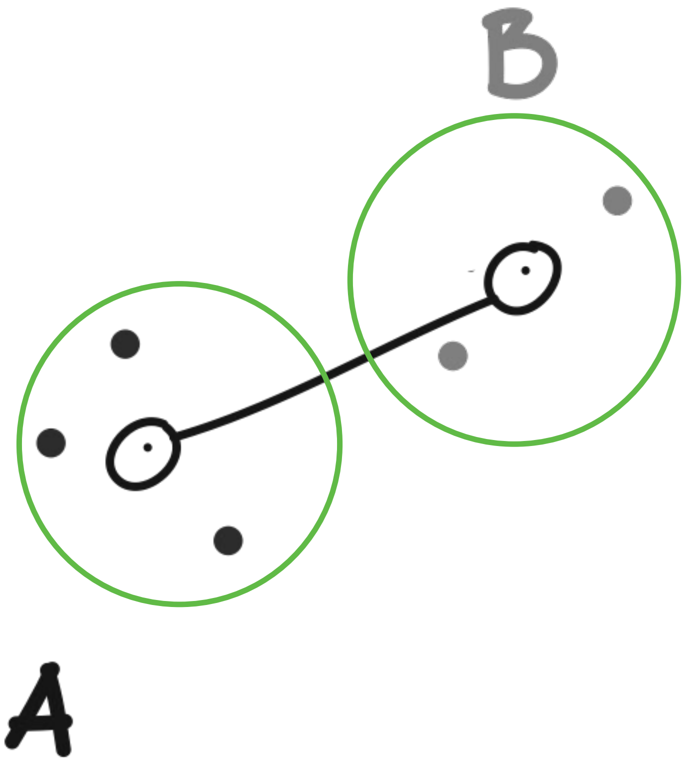
Linkage Criteria



$$\frac{d1+d2+d3+d4+d5+d6}{6}$$

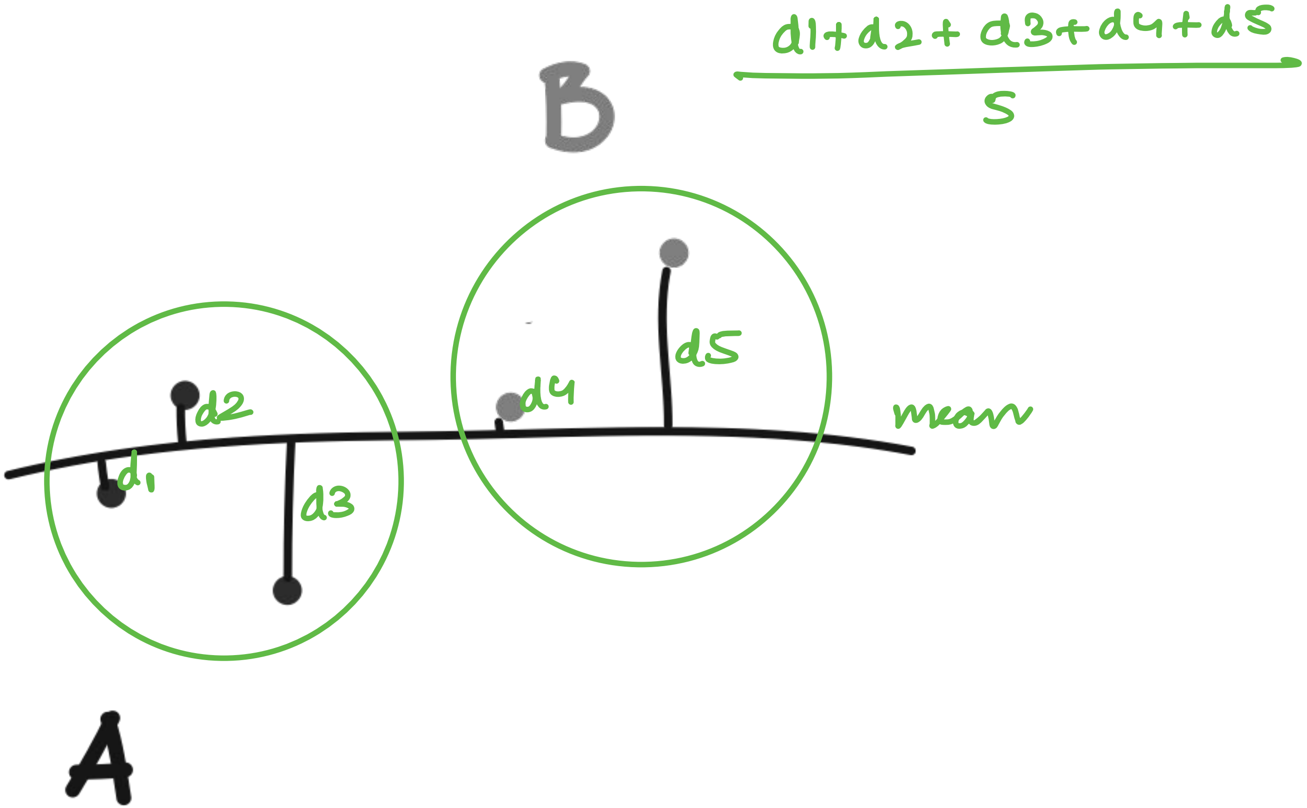
Average

Linkage Criteria



Centroid

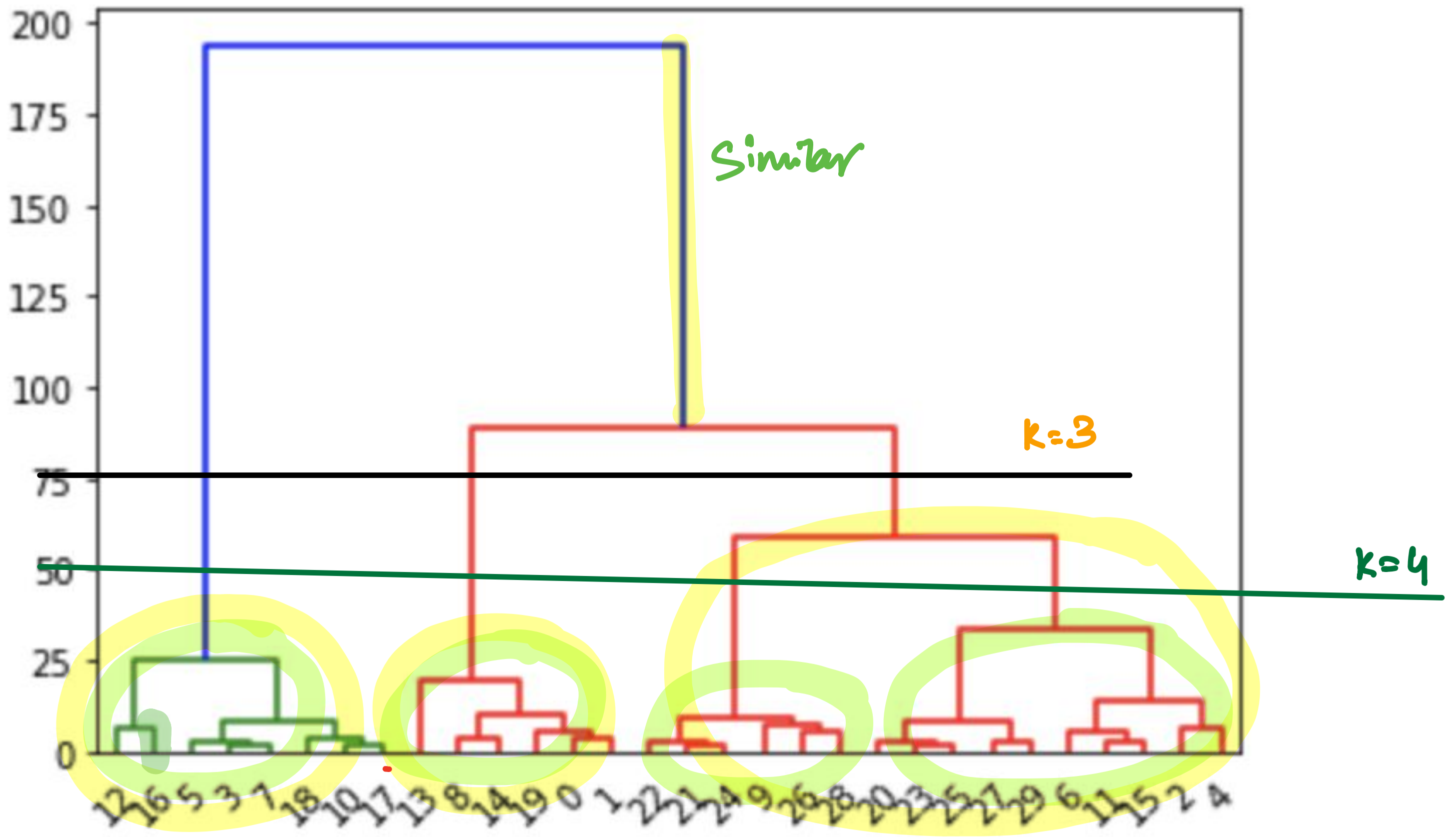
Linkage Criteria



(Variance Based)

Ward's

Reading a Dendrogram



Principle Component Analysis (PCA)

Dataset: lot of features: 100 features → all features may not be relevant

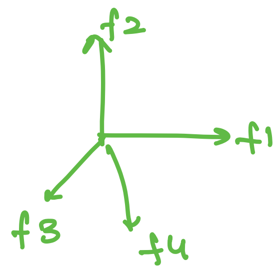
Salary Predict

Humidity   Temp   College   CGPA   Branch   Exp.

Irrelevant ✗

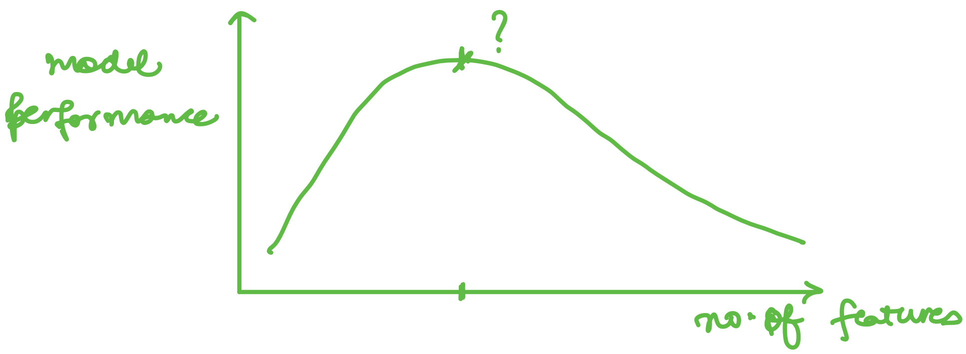
features → Irrelevant

features → Redundant: age, dob



Curse of Dimensionality

↳ lot of features, model performance ↓



Curse of Dimensionality

↓

Dimensionality Reduction

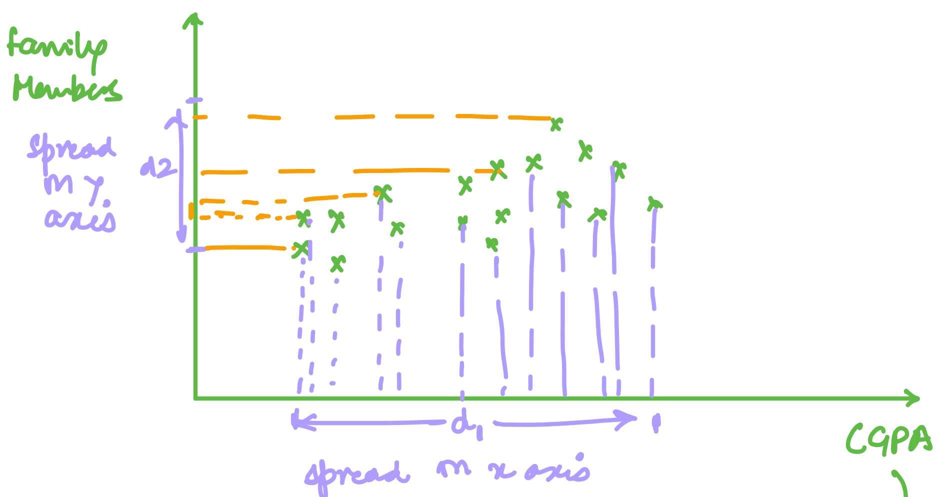
- Feature Selection
  - f1 f2 f3 .... f100
  - top 50 features f1, f2 ... f50
- Feature Extraction
  - f1 f2 f3 .... f100
  - 5 feature:
    - f1+f3
    - f1*f2
    - f1+f4+f6
    -
    -

1. PCA ✓
2. LDA
3. t-SNE ✓

Feature Selection:

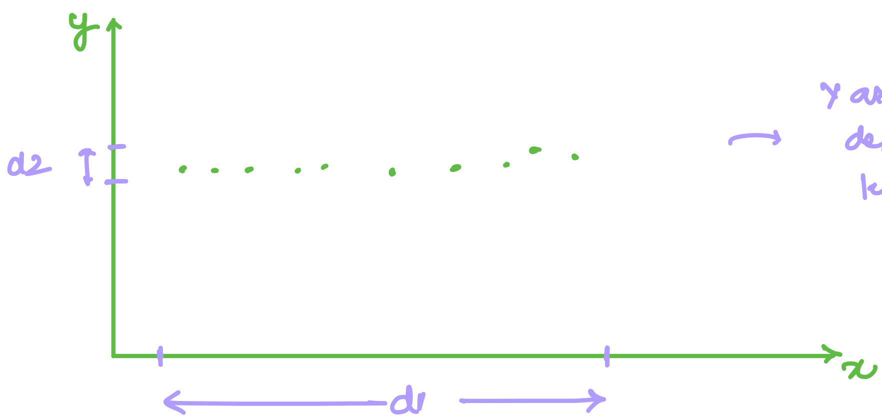
Salary Predict

x_1 Family Members	x_2 CGPA	y Salary
3	10	60
2	7	55
5	9	70



$d1 > d2$

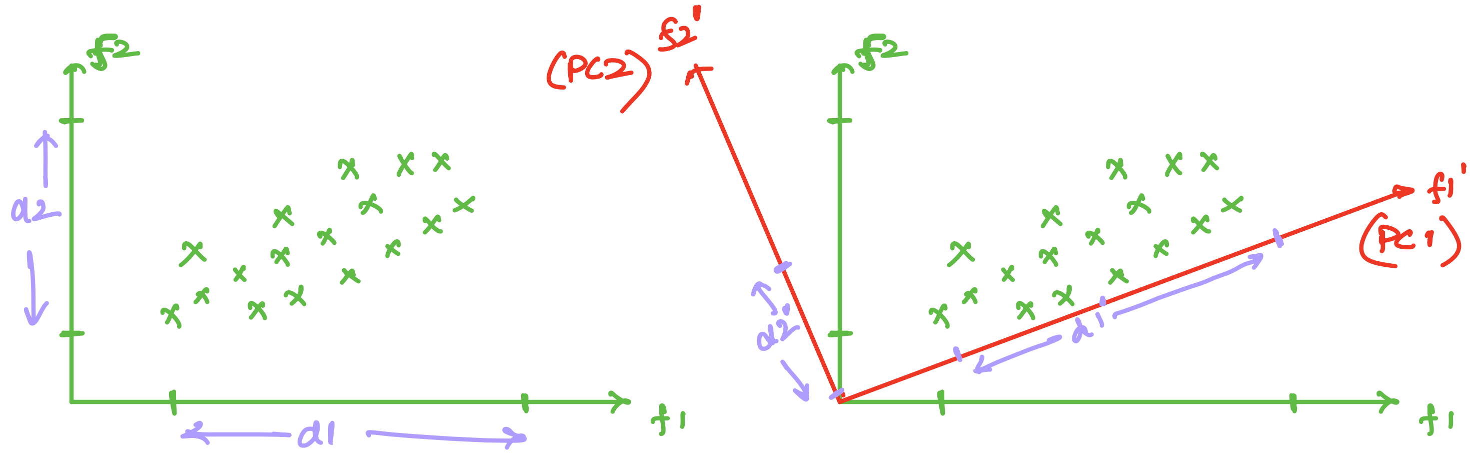
CGPA is carrying more info than F.M.



→ y axis dependency is very less.

Spread $\propto$ Variance

PCA:



$d1 \approx d2$

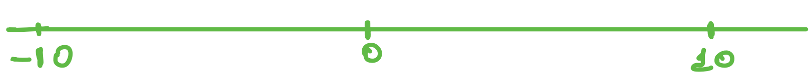
$d1' > d2'$, hence we can prefer f1'

Ques: how to find out this new pair of axis

- why variance is important?

Variance $= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$

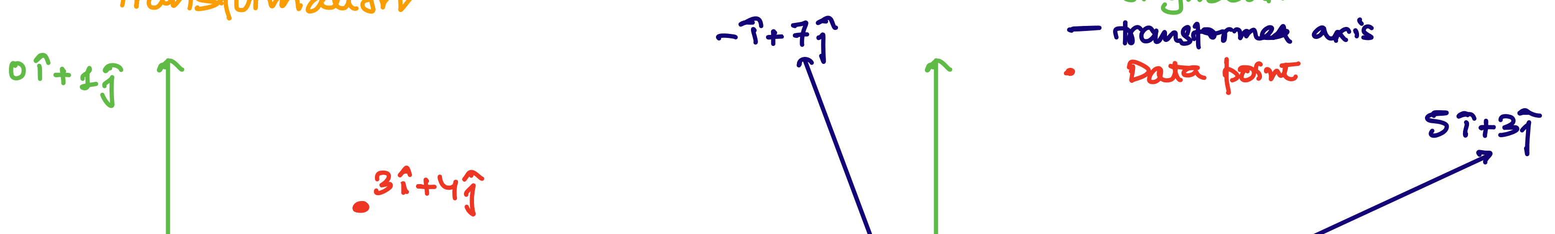
mean $= \frac{-5+0+5}{3} = 0$ $\quad V = \frac{25+0+25}{3} = \frac{50}{3}$

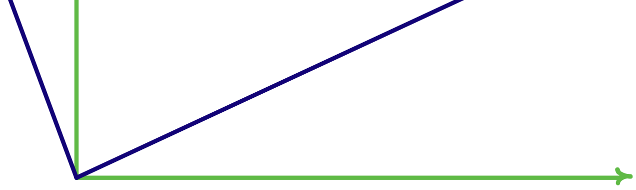


mean $= \frac{-10+0+10}{3} = 0$

$V = \frac{100+0+100}{3} = \frac{200}{3}$

- Transformation





$$\text{Transformed} = 3\,\underbrace{(\text{Transformed } x)}_{\begin{bmatrix}5\\3\end{bmatrix}} + 4\,\underbrace{(\text{Transformed } y)}_{\begin{bmatrix}-1\\7\end{bmatrix}}$$

$$= 3\begin{bmatrix}5\\3\end{bmatrix} + 4\begin{bmatrix}-1\\7\end{bmatrix} = \begin{bmatrix}15\\9\end{bmatrix} + \begin{bmatrix}-4\\28\end{bmatrix} = \begin{bmatrix}11\\37\end{bmatrix}$$

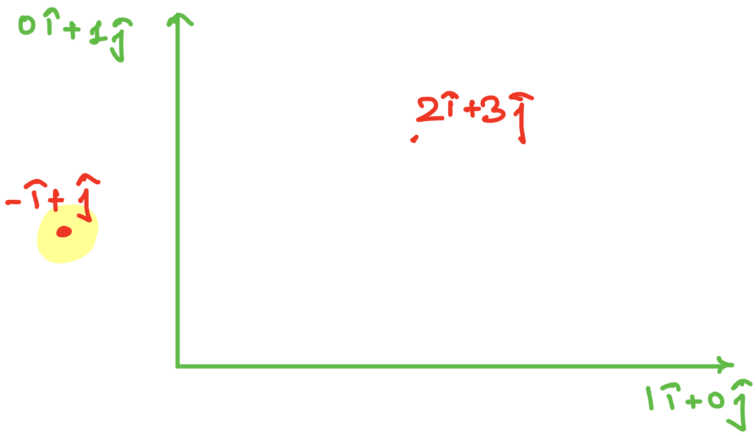
$11\hat{i} + 37\hat{j}$

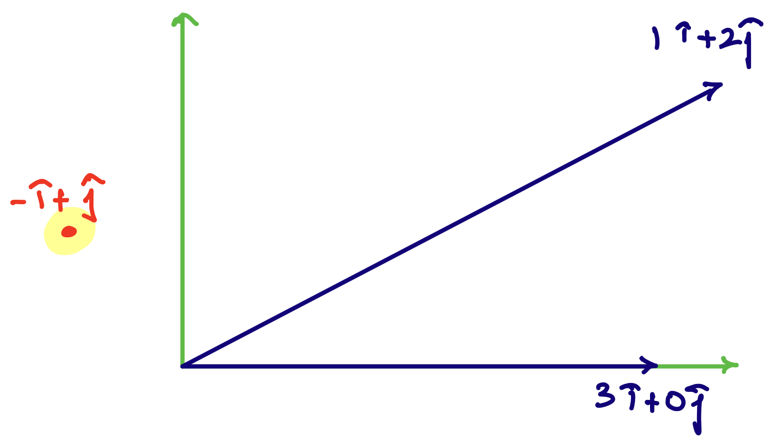
Transformation Matrix

$$\begin{bmatrix}5 & -1\\3 & 7\end{bmatrix}\begin{bmatrix}3\\4\end{bmatrix} = \begin{bmatrix}15-4\\9+28\end{bmatrix} = \begin{bmatrix}11\\37\end{bmatrix}$$

new x　　new y　　denotes the data point

- Eigen Value & Eigen Vectors





$2\hat{i}+3\hat{j}$

$$\begin{bmatrix}3 & 1\\0 & 2\end{bmatrix}\begin{bmatrix}2\\3\end{bmatrix} = \begin{bmatrix}6+3\\0+6\end{bmatrix} = \begin{bmatrix}9\\6\end{bmatrix} = 3\begin{bmatrix}3\\2\end{bmatrix}$$

$-\hat{i}+\hat{j}$

$$\begin{bmatrix}3 & 1\\ 0 & 2\end{bmatrix}\boxed{\begin{bmatrix}-1\\ 1\end{bmatrix}} = \begin{bmatrix}-3+1\\ 0+2\end{bmatrix} = \begin{bmatrix}-2\\ 2\end{bmatrix} = 2\boxed{\begin{bmatrix}-1\\ 1\end{bmatrix}}$$

Eigen vector remains at the same position even after transformation

$$A \cdot \bar{V} = \lambda \bar{V}$$

Transformation Matrix

Eigen Vector

Eigen Value (Scalar)

$$A\cdot\bar{V} = \lambda\cdot I\cdot\bar{V}$$

$$I = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

$$(A-\lambda I)\bar{V} = 0$$

$$\lambda I = \begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}$$

$$|A-\lambda I| = 0$$

$$\begin{bmatrix}3 & 1\\ 0 & 2\end{bmatrix} - \begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix} = 0$$

$$\begin{bmatrix}3-\lambda & 1\\ 0 & 2-\lambda\end{bmatrix} = 0$$

$$(3-\lambda)(2-\lambda) = 0$$

$\lambda = 2, 3 \rightarrow$ Eigen Values.

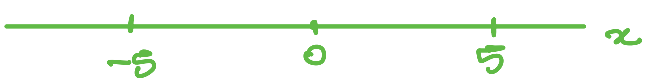
- Covariance and Variance

n = no. of data points

Variance:

$$\text{Variance}(x) = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n}$$

(1 feature)

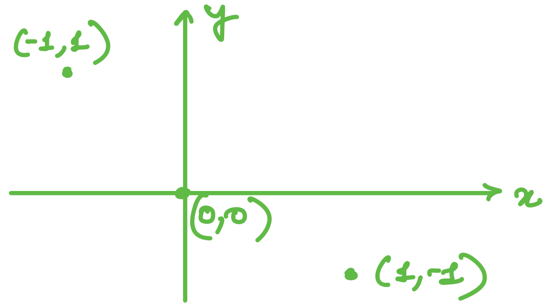


↳ x spread

n = no. of data points.

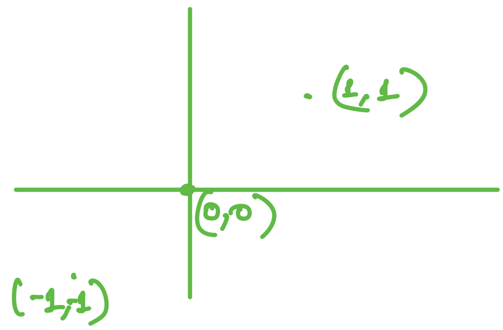
Covariance

(2 feature)



$$\mathrm{cov}(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n}$$

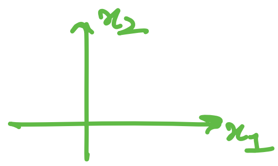
$$CV = \frac{-1+0-1}{3} = \frac{-2}{3}$$



$$CV = \frac{1+0+1}{3} = \frac{2}{3}$$

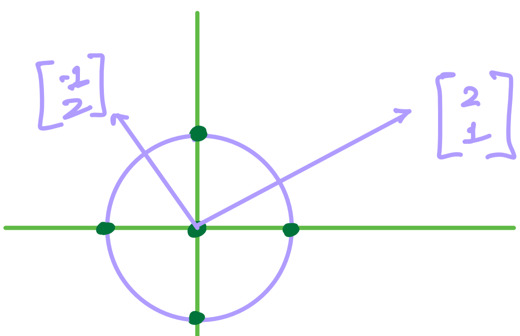
Covariance Matrix

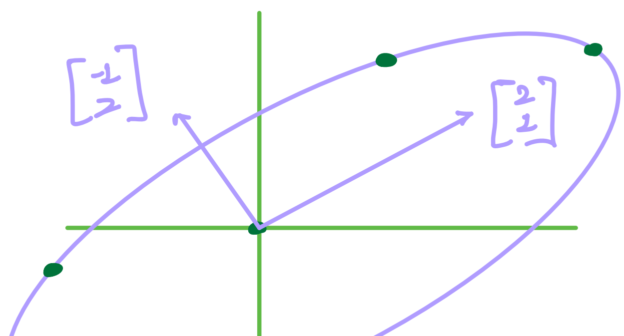
	x_1	x_2
x_1	$\mathrm{Cov}(x_1,x_1)$ $= V(x_1)$	$\mathrm{Cov}(x_1,x_2)$
x_2	$\mathrm{Cov}(x_2,x_1)$	$\mathrm{Cov}(x_2,x_2)$ $= V(x_2)$



- Symmetric Matrix $\quad \mathrm{Cov}(x_1,x_2) = \mathrm{Cov}(x_2,x_1)$
- Used as transformation Matrix

Covariance Matrix $= \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$





$$\begin{pmatrix}9 & 4\\4 & 3\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} = x\begin{bmatrix}9\\4\end{bmatrix} + y\begin{bmatrix}4\\3\end{bmatrix} = \begin{bmatrix}9x+4y\\4x+3y\end{bmatrix}$$

$$(x,y) \longrightarrow (9x+4y,\ 4x+3y)$$

$$(0,0) \longrightarrow (0,0)$$

$$(1,0) \longrightarrow (9,4)$$

$$(0,1) \longrightarrow (4,3)$$

$$(-1,0) \longrightarrow (-9,-4)$$

$$(0,-1) \longrightarrow (-4,-3)$$

$$(2,1) \longrightarrow (18+4,\ 8+3) = (22,11) = 11\,(2,1)$$

$$(-1,2) \longrightarrow 1\,(-1,2)$$

Eigen values

In PCA, goal was to find out new axis

$$\text{Covariance Matrix} \longrightarrow \underbrace{\text{Eigen Vectors}}_{\text{new axis}} \text{ \& Values}$$

Steps:

1. Load Data

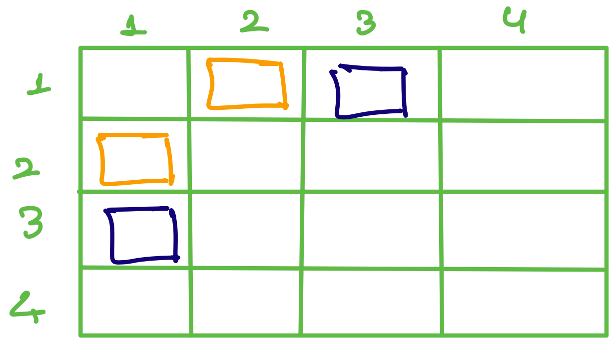
2. Standardization

   mean = 0

   std. deviation = 1

3. Covariance Matrix : how are 2 features related to each other.

n features
↳ 4



$n \times n$
4×4

4. Eigen Values & Eigen Vectors

evec 1 → evalue 1
evec 2 → evalue 2
evec 3 → evalue 3
evec 4 → evalue 4

5. Larger eigen value means spread/variance is more along that vector

evec 1 → 50
evec 2 → 30
evec 3 → 60
evec 4 → 10

} eigen values

Choose the axis along which spread is larger and spread is given by eigen value

evec 3 → 60
evec 1 → 50
evec 2 → 30
evec 4 → 10

} sorting on the basis of eigen value

6. Pick $\underbrace{\text{top k eigen}}_{\text{new dimension}}$ vectors

7. Projection

$$\begin{bmatrix} \text{matrix of data points} \end{bmatrix} \begin{bmatrix} | & | \\ \text{evec3} & \text{evec1} \\ | & | \end{bmatrix} = \begin{bmatrix} \text{final points} \end{bmatrix}$$

$$\underbrace{\hphantom{xxxxxxxxxxxxxxxx}}_{\text{n features}}$$

$(m \times n)$

$(n \times k)$

$(m \times k)$