→ Class start 9:05 PM

ML 1.2 Unsupervised ML

→ Clustering (KMeans, Kmeans++, Hieanchial, DBSCAN, GMM)

→ Anomaly Detection like Isolation forest etc.

→ High Dimension Visualization (PCA, t-SNE) UMAP,

Unsupervised ML

→ Supervised ML → { features, target or labels, or ground truth

→ Unsupervised ML → features

→ No target or label

Classification

Binary

$\{ (x_i, y_i)_{i=1}^{n}, x_i \in R^d, y_i \in (0,1) \}$

Multi-class

$y_i \in S$ → Set of class

Regression

$y_i \in R$

$$\{ x_i^{\,n}_{i=1}, \; x_i \in R^d \}$$

Examples of Unsupervised ML

→ Anamaly Detection / Fraud Detection
→ Clustering problem
→ Dimensionality Reduction like PCA
→ Recom. System like MF
→ Word-2-vec (NLP)
→ Autoencoder (CV)

Clustering

→ Process of grouping any kind of data based on similarity of these features

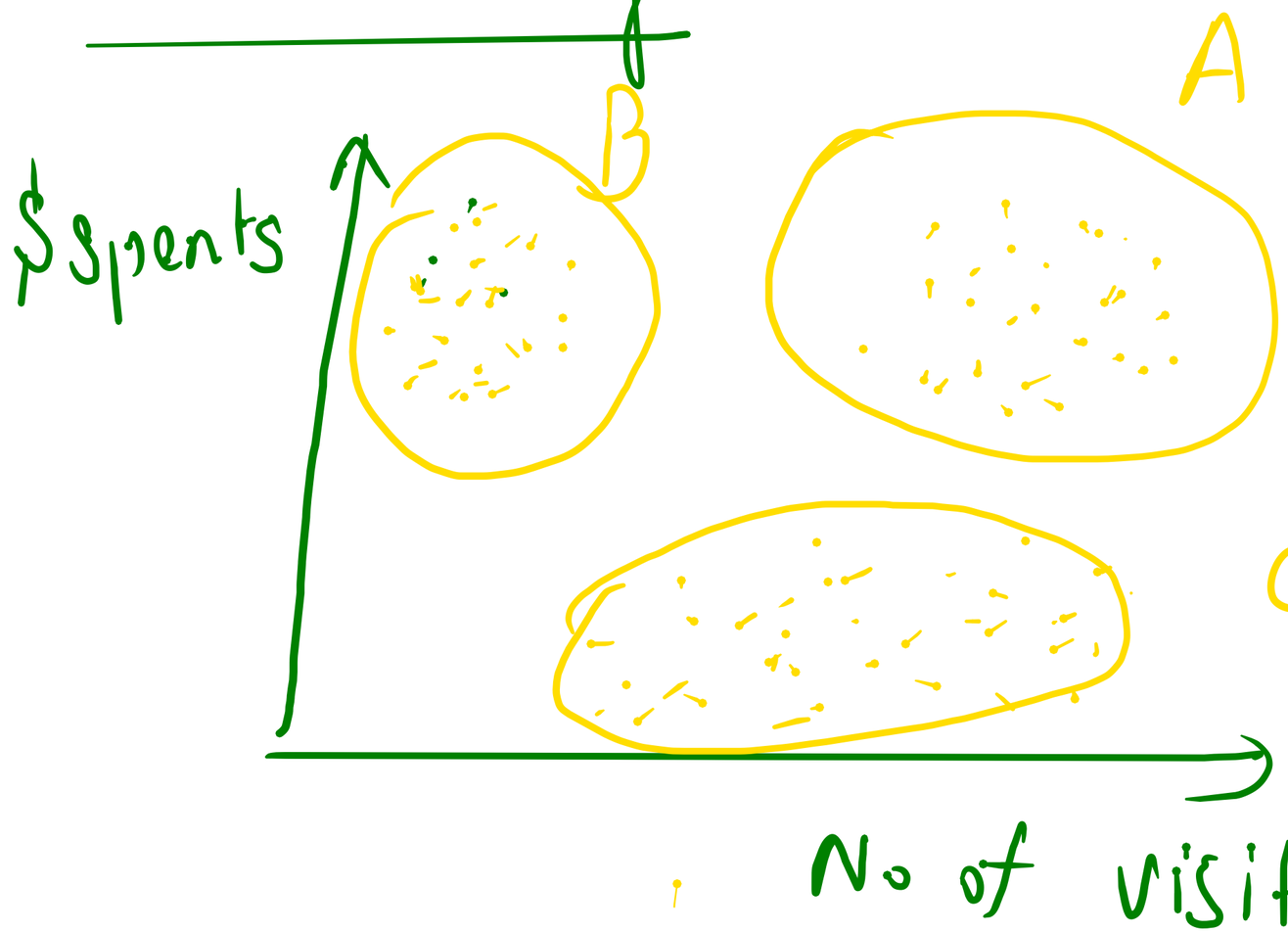
Eg

→ Customer Segmentation / Product Segmentation

→ Detecting similar stock

→ Google photos groups similar in galaxy

Clustering



A: Heavy shoppers, vist a lot, spent a lot

B: Rich people / Impulse buyer

C: Window shopper

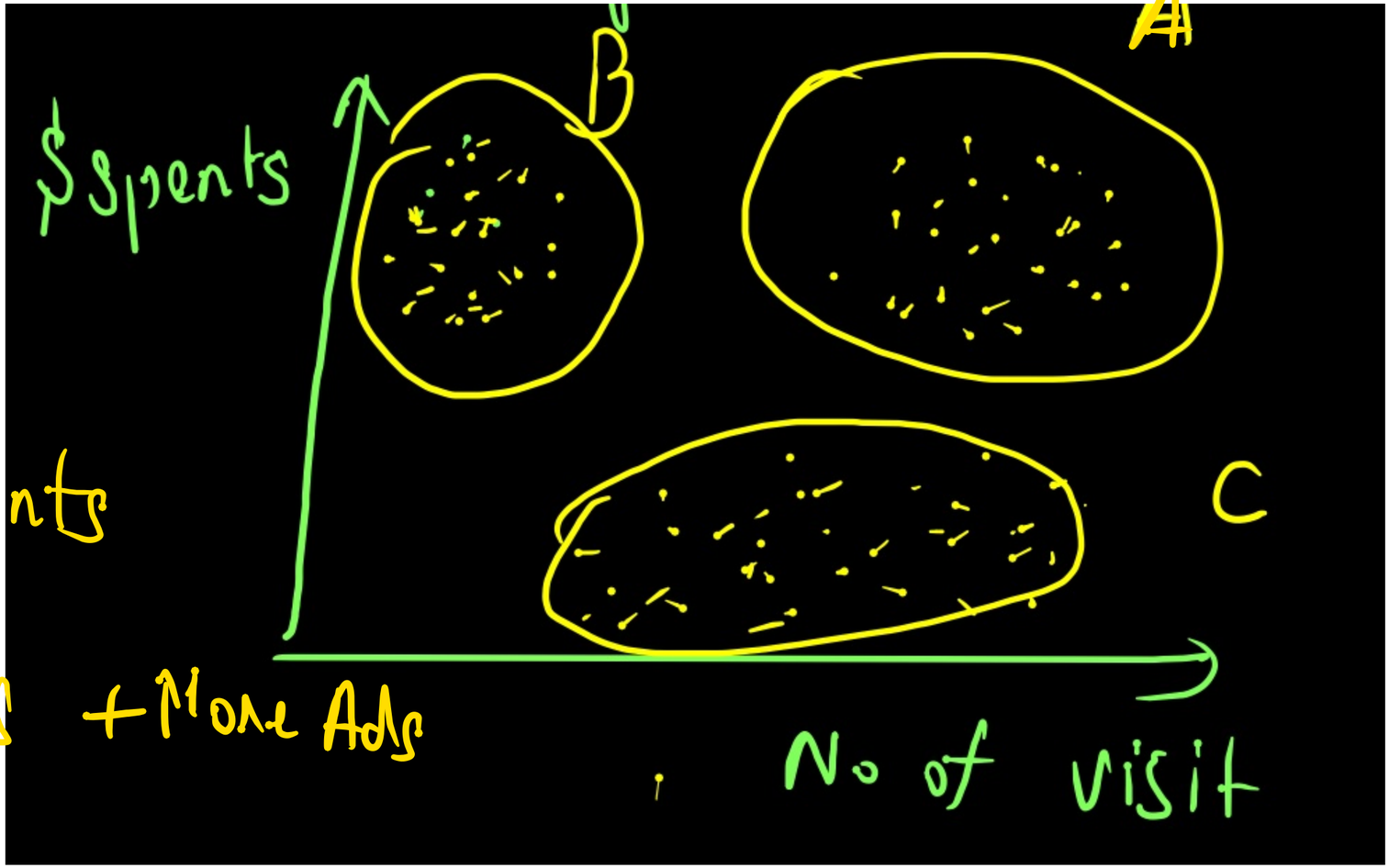
→ No of Discount coupon

→ No. of Ads

B → More Ads

A → Discounts

C → More Discounts + More Ads



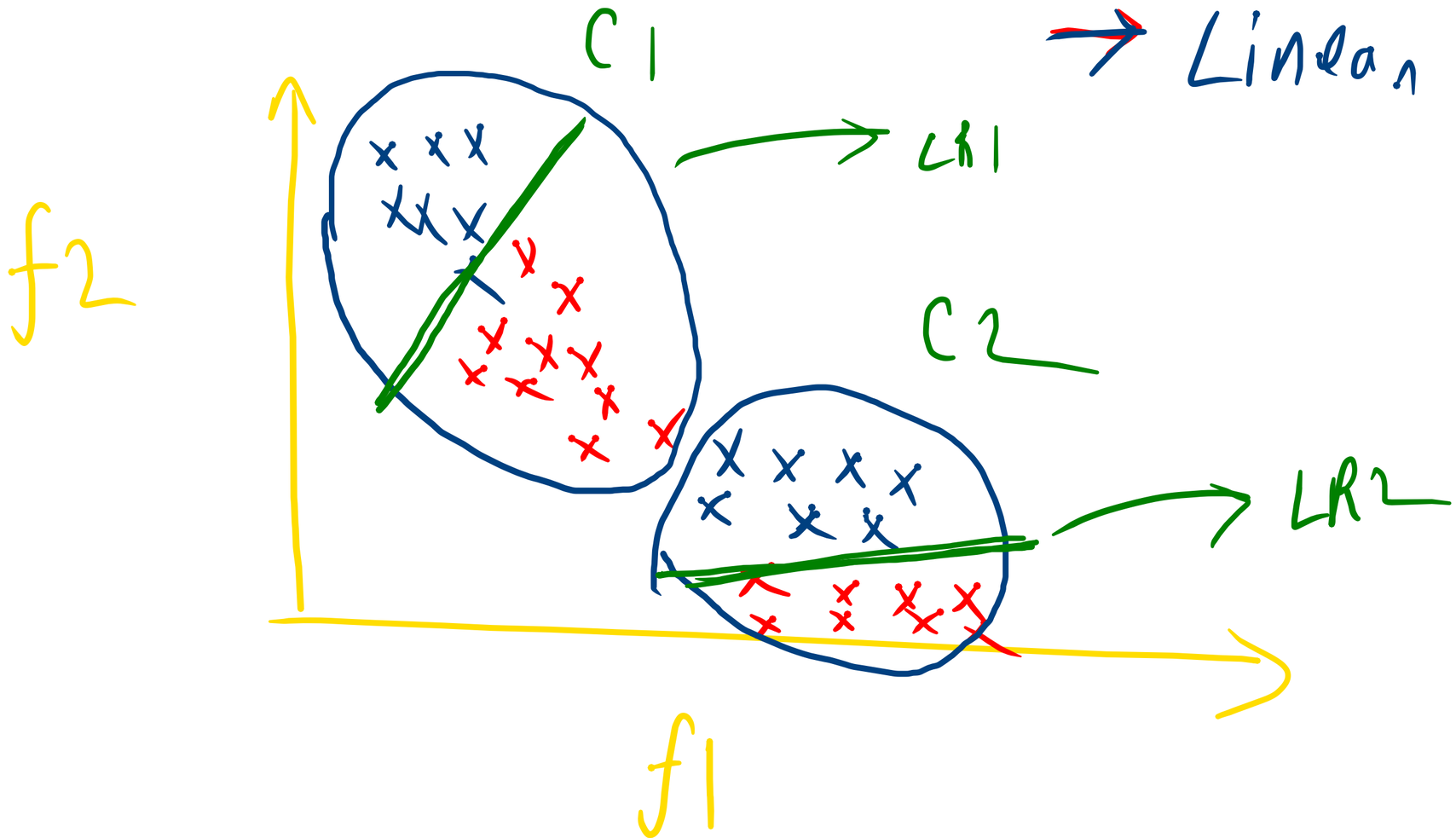
Application of Clustering

→ Search algorithm

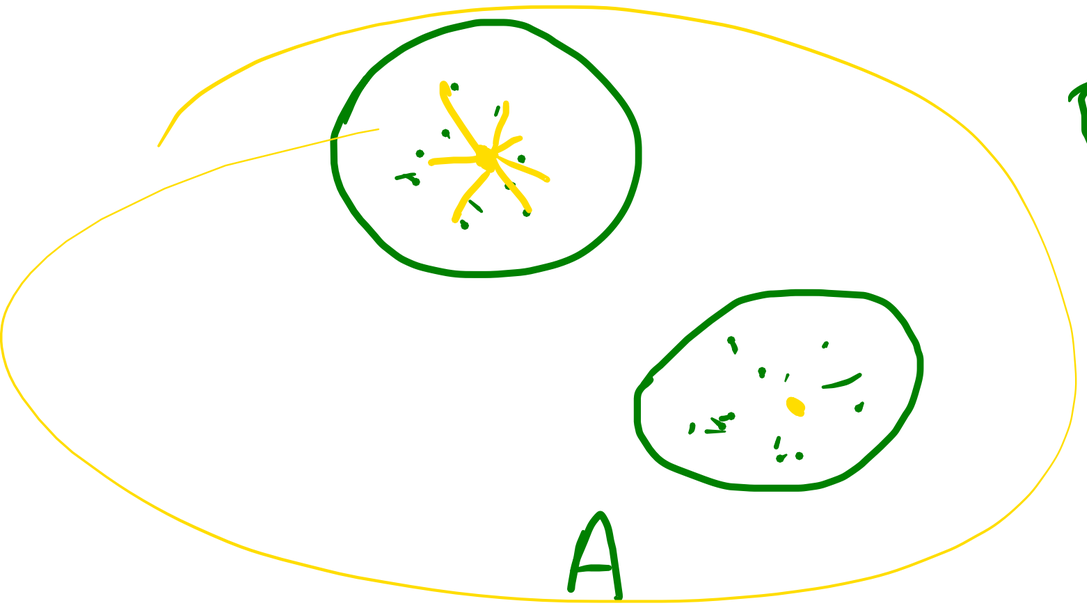
→ Feature Creation

ID	n_clicks	n_visit	amount_spent	Country
				1
				1
				2
				2

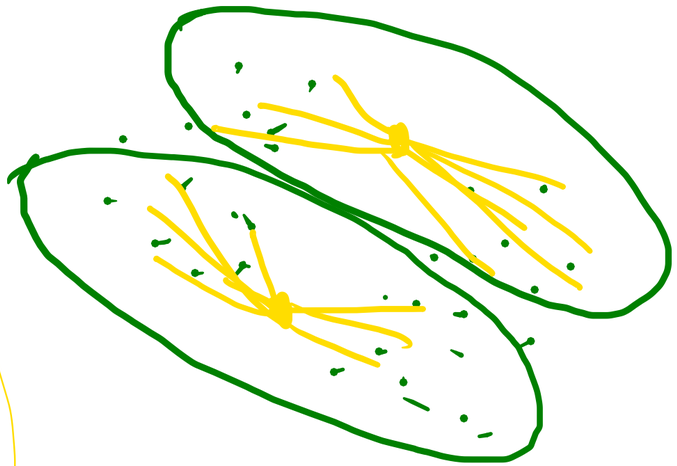
→ Linear Desicion boundary



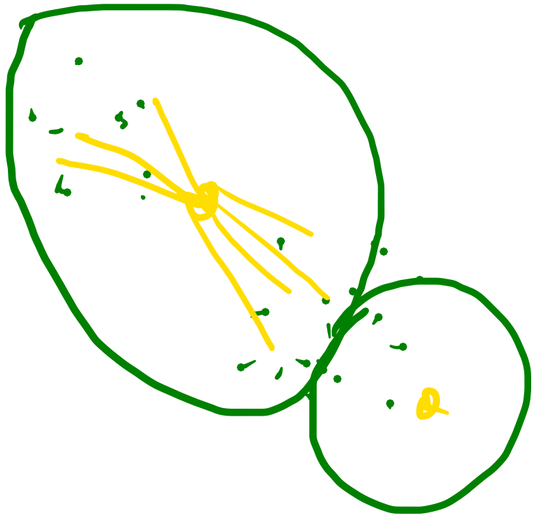
Good Clustering



A



B



C

→ Intra cluster dist minimize

→ Inter cluster dist maximize

WCSS (Within Cluster Sum of Square)

$$WCSS = \sum_{P_i \text{ in } C_1} dist(P_i, C_1)^2 + \sum_{P_i \text{ in } C_2} dist(P_i, C_2)^2 + \cdots + \sum_{P_i \text{ in } C_j} dist(P_i, C_j)^2$$

$$NCSS = \sum_{k=1}^{K} \sum_{i=1}^{n} 1(c_i = k) \, \| x_i - \mu_k \|^2 = L$$

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

C_1 C_2

Gradient Descent

Update all parameters simultaneously

$W \leftarrow W - \alpha \nabla_W L$

$W \rightarrow \begin{matrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{matrix}$

Coordinate Descent

→ Update only a subset of parameters

↳ Fixing μ & we find best c exactly

↳ Fixing c & find μ

K-means Clustering (Lloyd's Algorithm)

Steps:

→ Randomly initialize K centers

→ assign points to nearest center ← C_i

to get your clusters

Assignment Step

→ find the centroids of the clusters → because this will reduce loss ← μ_K

Update

→ Re-assign points

→ Repeat until new centers = prev centers

Elbow Method → WCSS

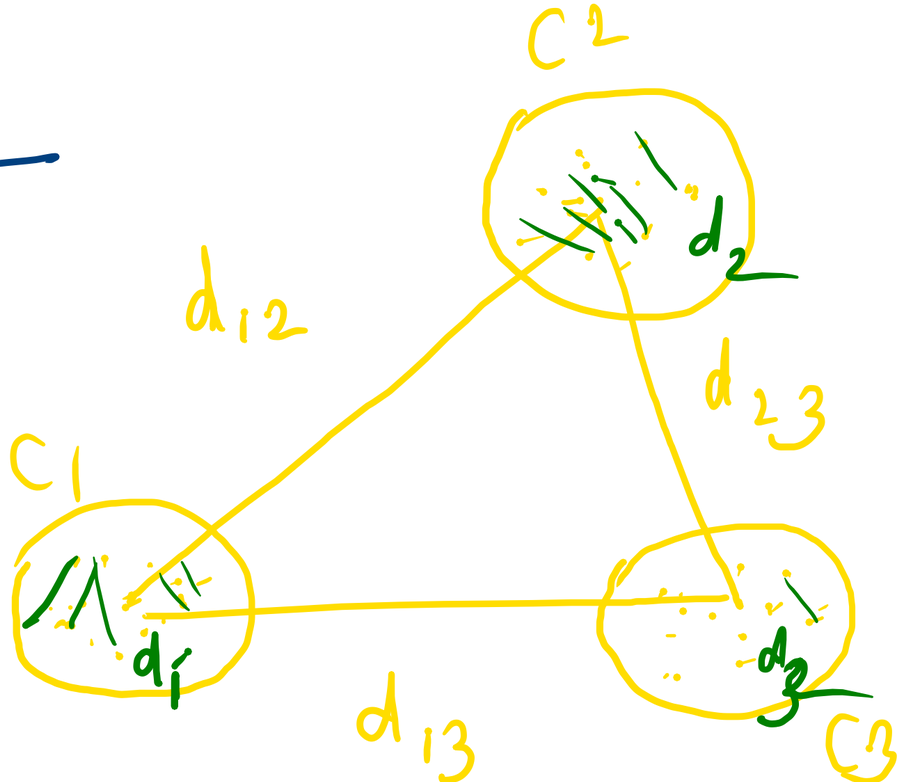
↳ Within Cluster sum of Squane → Intra cluster distance

- Dunn Index
- Silhouette Score

Dunn Index

$$D = \frac{\text{Min Intercluster dist}}{\text{Max Intracluster dist}}$$

High → good clustering



Min Intercluster dist = $\min(d_{12}, d_{23}, d_{13})$

Max Intra cluster dist = $\max(d_1, d_2, d_3)$

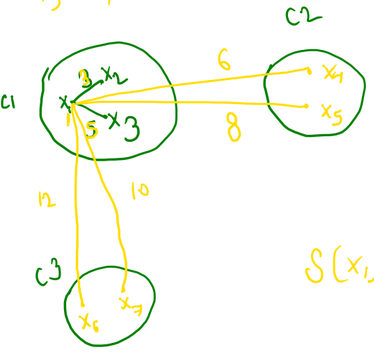
Silhoute score

$$S(x_i) = \frac{(b^{(x_i)} - a^{(x_i)})}{\max(a, b)}$$

$-1, +1$

$a^{(x_i)}$ = Mean dist b/w i & all neighbouring point in a cluster

$b^{(x_i)}$ = Min mean dist b/w i & all point in other cluster



$$S(x_1) = \frac{b(x_1) - a(x_1)}{\max(a(x_1), b(x_1))}$$

$$a(x_1) = \frac{3+5}{2} = 4$$

$$b(x_1) = \min\left(\frac{6+8}{2}, \frac{12+10}{2}\right)$$

$$= \min(7, 11) = 7$$

$$S(x_1) = \frac{7-4}{\max(7,4)} = \frac{3}{7}$$

$S(x_1) \rightarrow -1 \quad +1$

$+ve \rightarrow$ Good

$a > b \rightarrow S(x_i) = -ve$ → Bad Clustering

$= -1$

$0 \rightarrow$

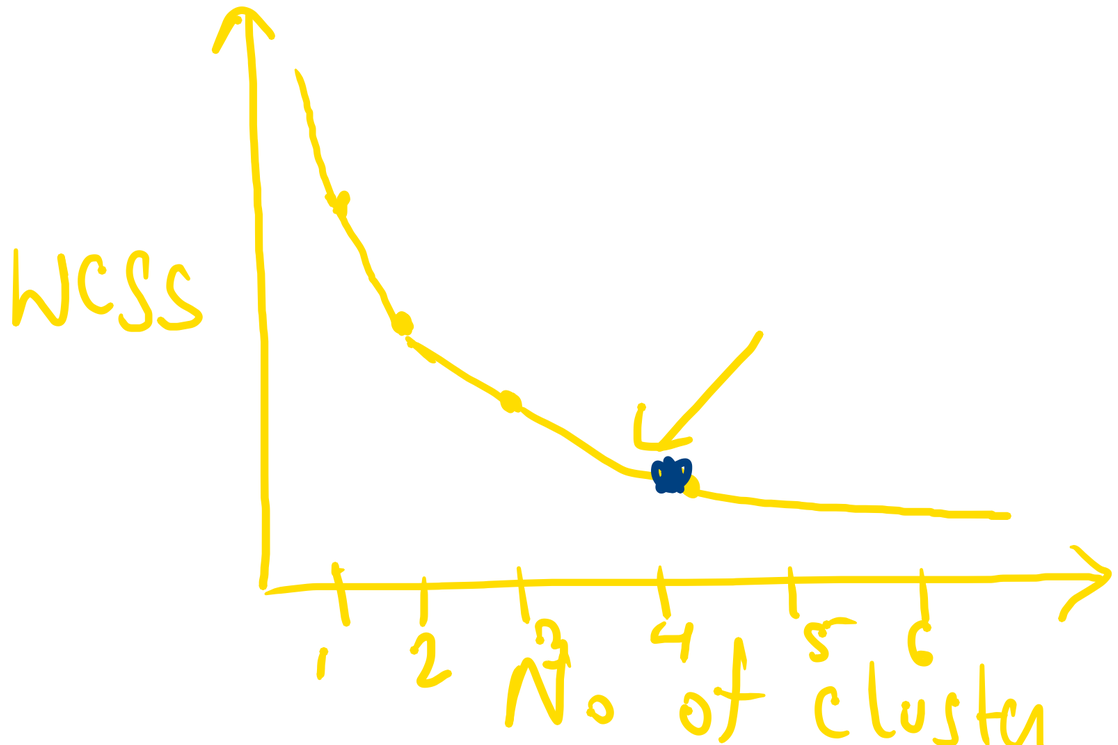
Determine the best value of K (No. of clusters)

Elbow Method

for k in range(10):

→ Kmean (K)

→ WCSS



Red cluster

(1, 2)

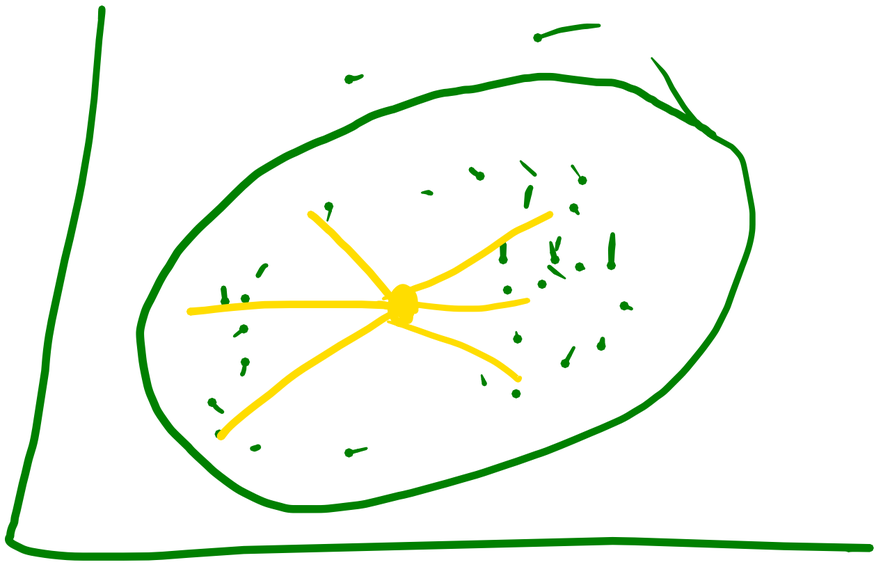
(2, 3)

(3, 4)

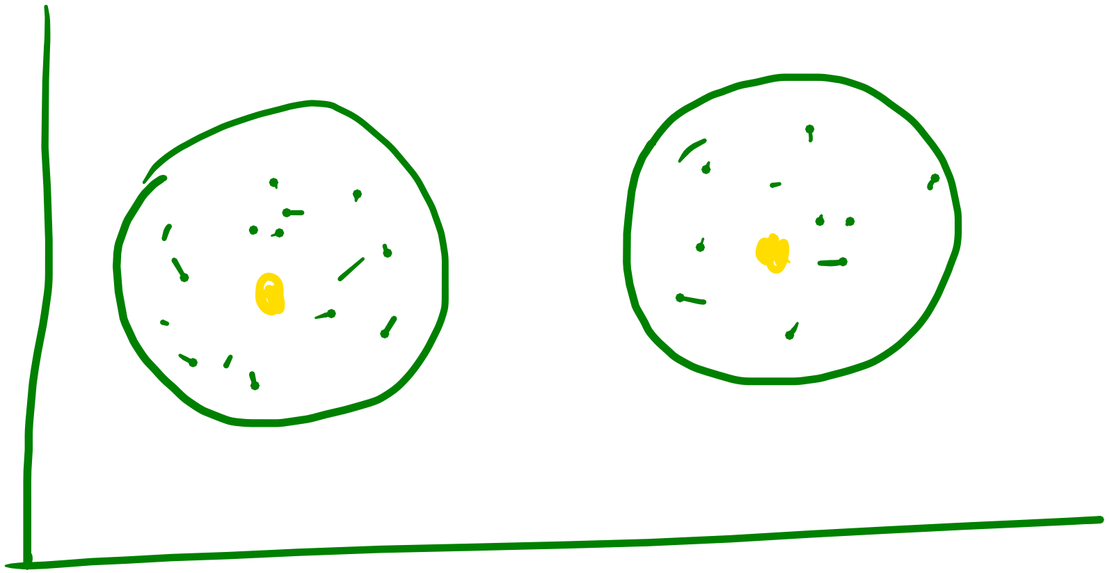
$$\mu_K = \frac{1}{n_K} \sum_{x_i \in C_K} x_i$$

$$\left(\frac{6}{3}, \frac{9}{3}\right) = (2, 3) \leftarrow \text{Centroid}$$

$\mu_{new} - \mu_{old} < 10^{-4}$



Cluster = 1



Cluster = 2

→ Break until 22:33

Business

- → No of Ads
- → Amount of Discount