



Lecture 13: K Means Clustering

Recap

- Clustering properties & metrics
 - Inertia (WCSS), Silhouette score, Dunn Index
- Elbow plot, Silhouette plot



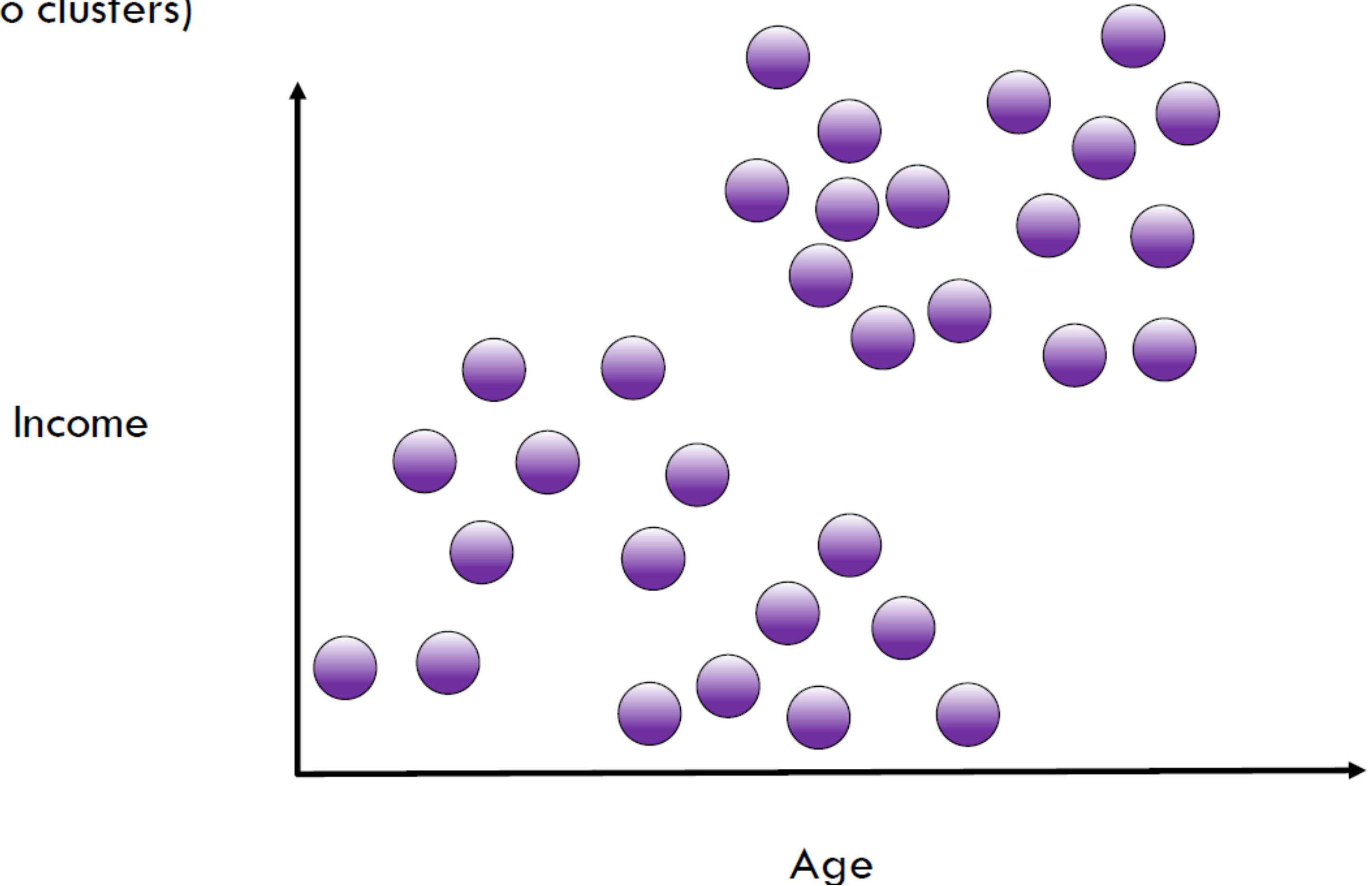
K-means clustering algorithm

K-Means Algorithm – Input and outputs

- Input:
 - Set of data points x_i
 - K
 - No labels
- Output:
 - Grouping of data points into K clusters
 - A centroid for each group - prototypical representation of the group

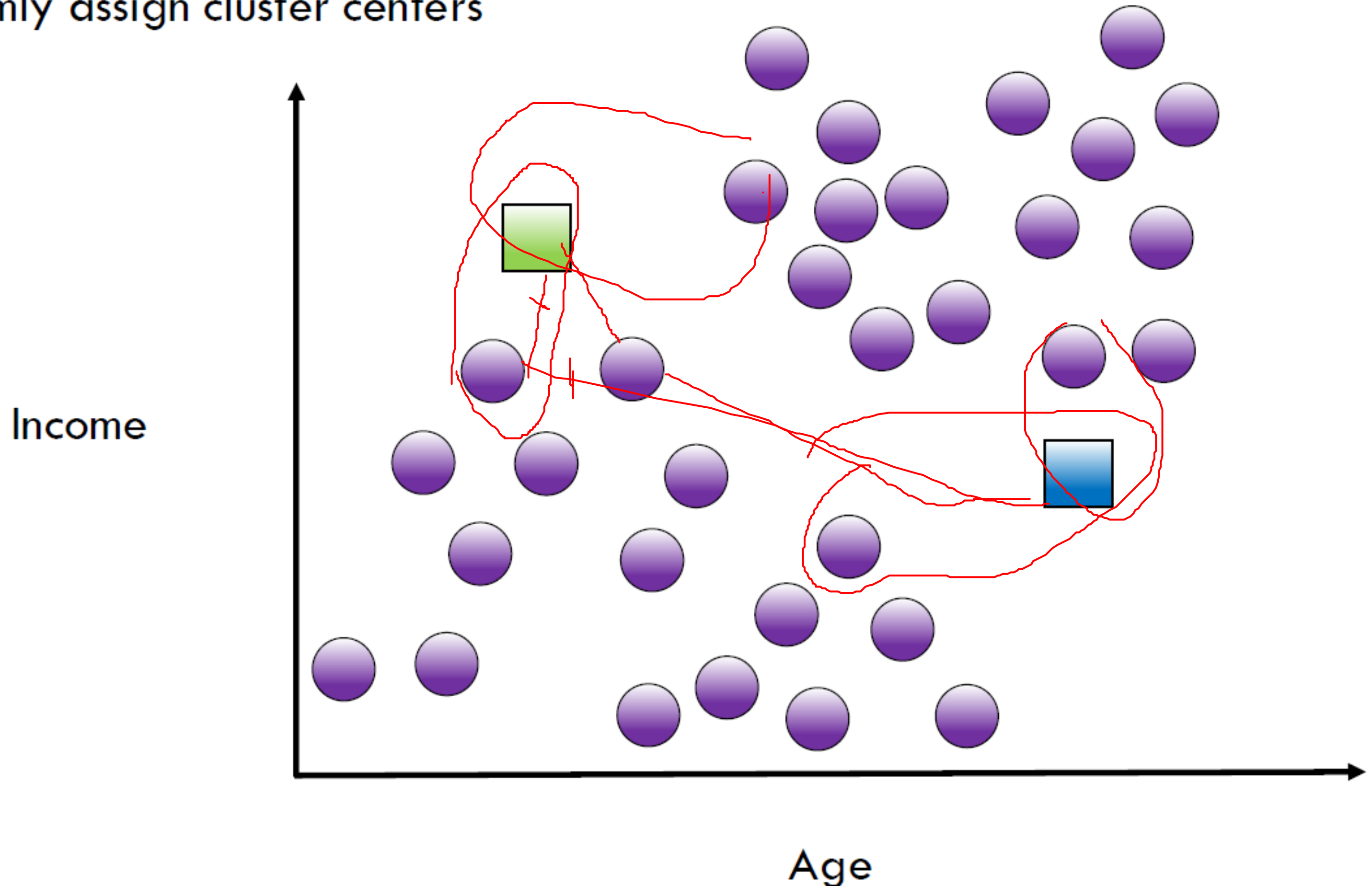
K-Means Algorithm

$K = 2$ (find two clusters)



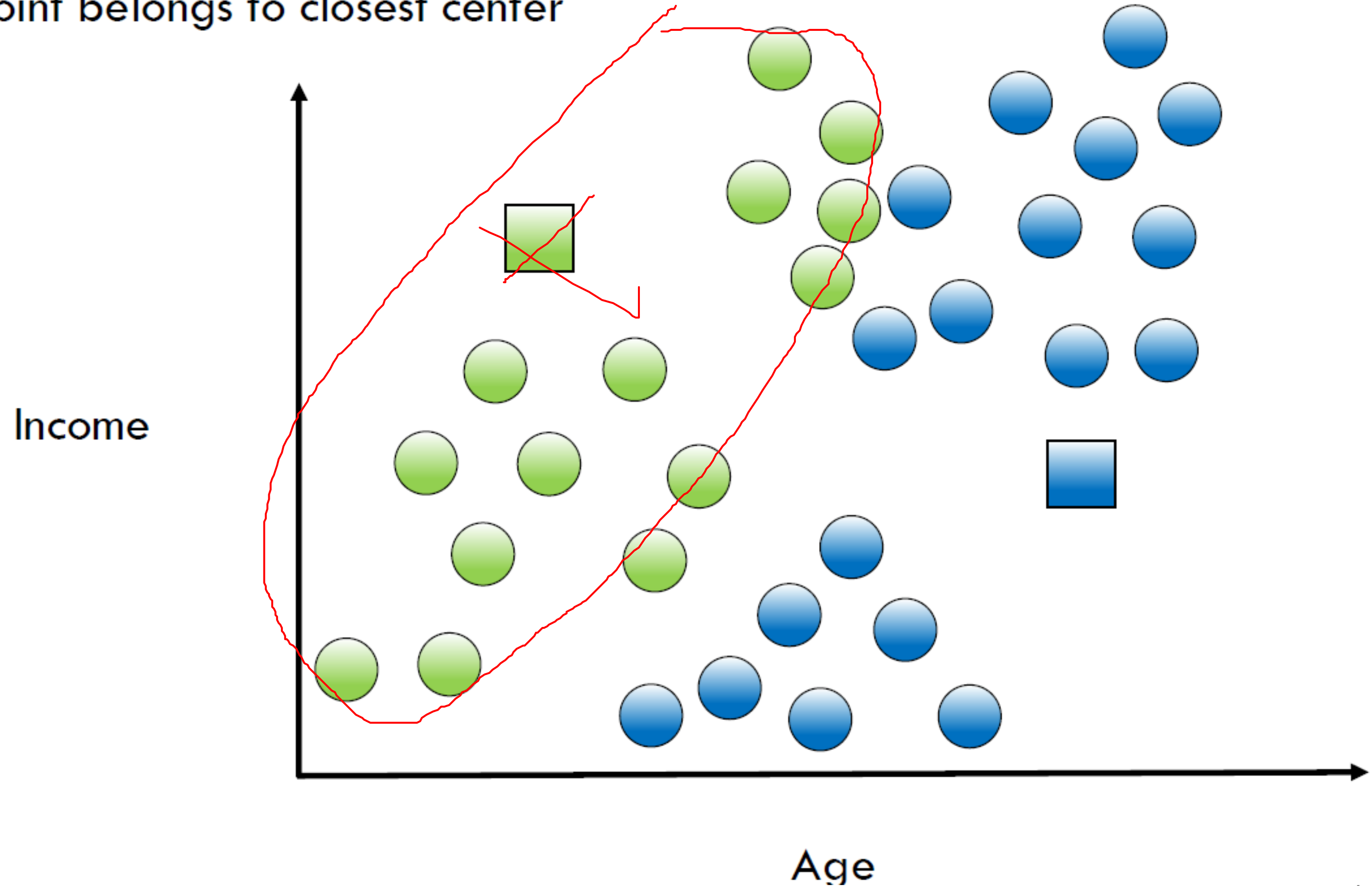
K-Means Algorithm

$K = 2$, Randomly assign cluster centers



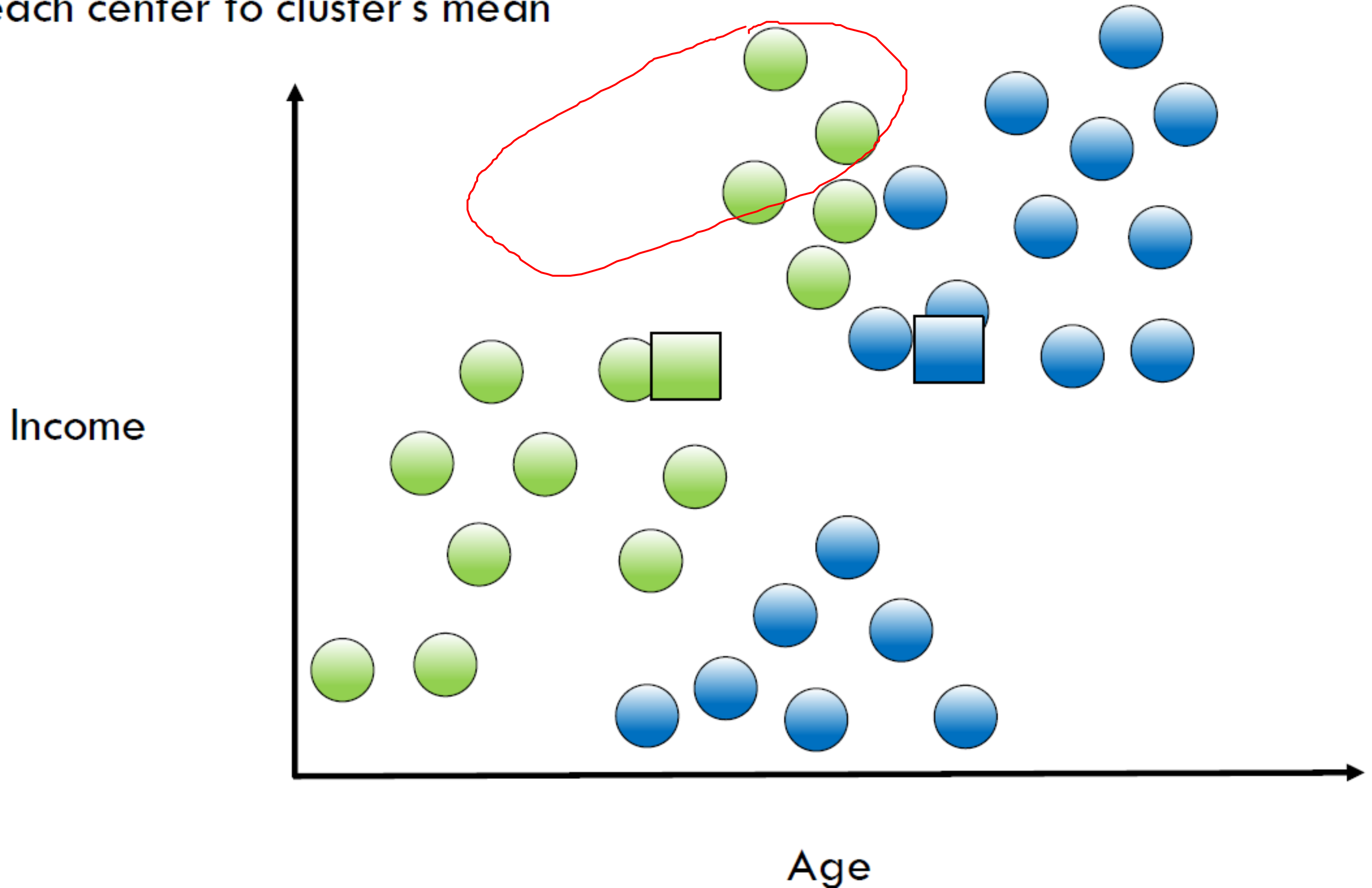
K-Means Algorithm

$K = 2$, Each point belongs to closest center



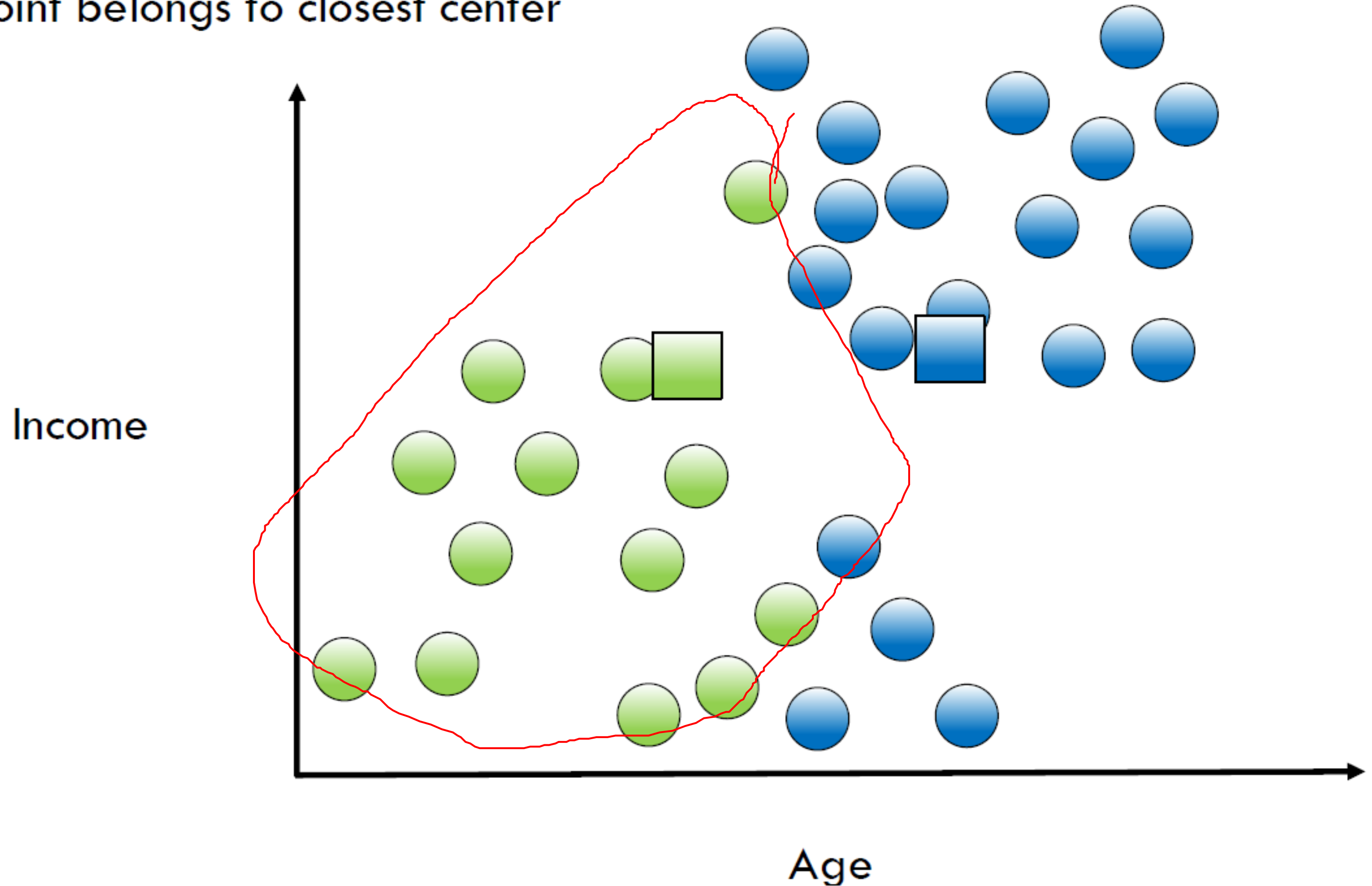
K-Means Algorithm

$K = 2$, Move each center to cluster's mean



K-Means Algorithm

$K = 2$, Each point belongs to closest center



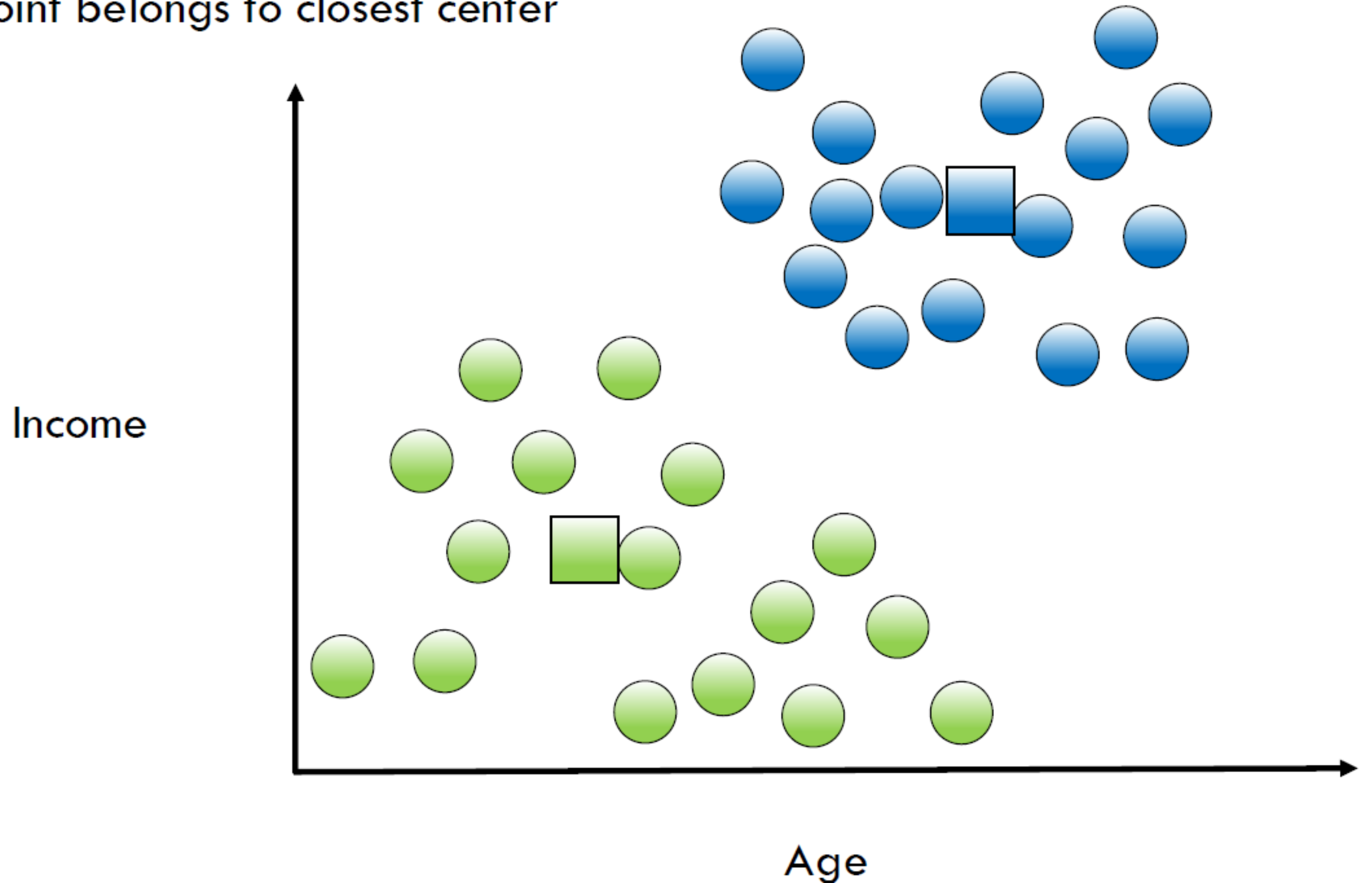
K-Means Algorithm

$K = 2$, Move each center to cluster's mean



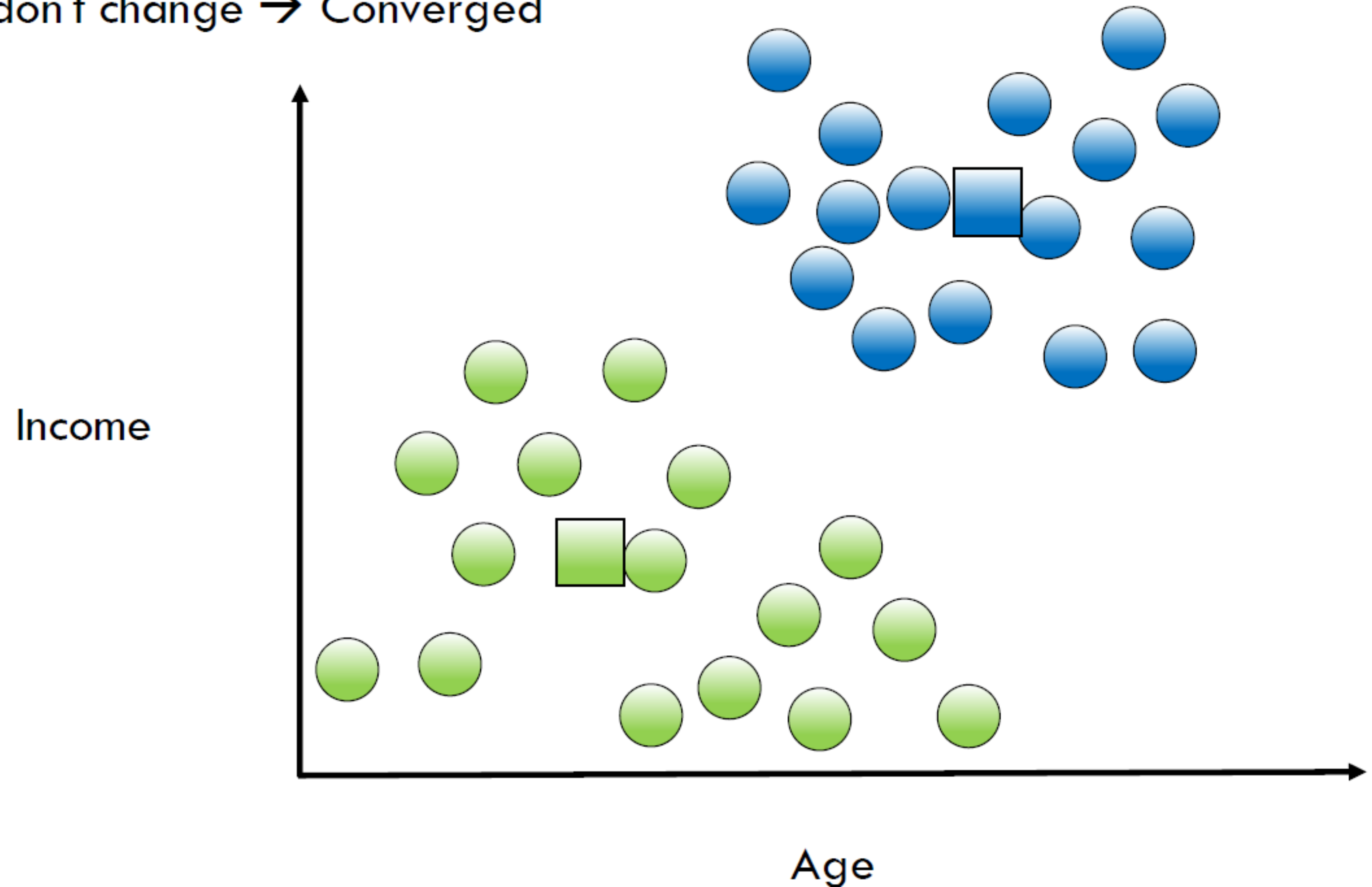
K-Means Algorithm

$K = 2$, Each point belongs to closest center



K-Means Algorithm

$K = 2$, Points don't change \rightarrow Converged



K-Means Algorithm

- Randomly select k points as centroids for k clusters
 - $J = 1$ to k Centroids are μ_1 to μ_k
- while (true):
 - for each point x_i
 - find nearest centroid μ_j
 - assign the point x_i to cluster j
 - for each cluster $j = 1$ to k
 - Update the centroid μ_j of each cluster (using points assigned to cluster j in previous step)
 - Stop when cluster assignments don't change (i.e. centroids don't change)

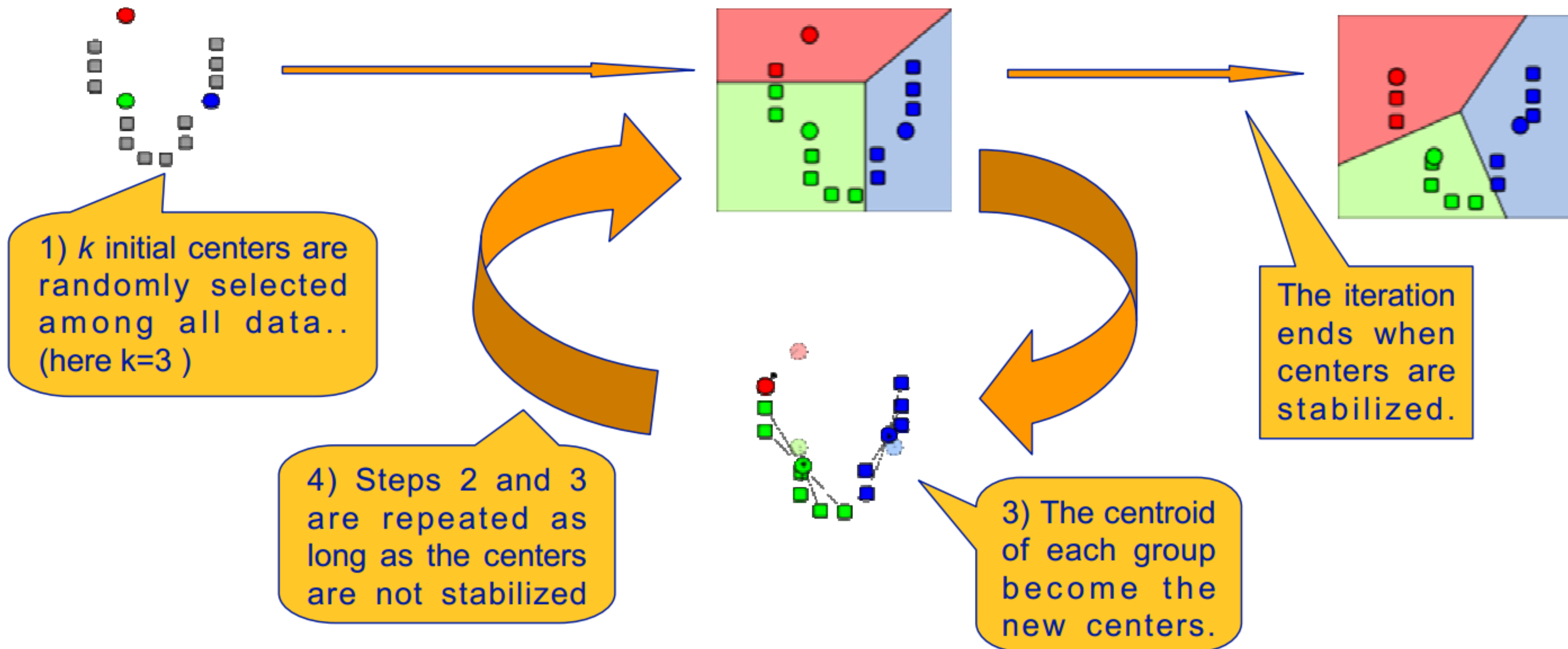
$$\forall j \in [1, k]$$

$$\arg \min_j D(x^{(i)}, \mu_j)$$

K-means

An unsupervised method.

MacQueen, 1967





Interpreting K-means as Expectation Maximization

K-Means as EM: Steps 1& 2 Guess, Expect

- **Guessing step**: Randomly guess k centroids

$$\mu_1, \dots, \mu_j, \dots, \mu_k$$

- **Expectation Step**: Cluster assignment

- Assign datapoints to clusters whose centroids they are closest (closest = minimum distance)

$$\forall j \in [1, k] \quad \arg \min_j D(x^{(i)}, \mu_j)$$

For every
data point

- Called expectation because we update our expectation on which cluster does the point belong

K-Means as EM: Step 3 Maximize

- **Maximization Step**: Set the mean of cluster data points as the new centroid

$$\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

- Called maximization because we maximize the fitness function defining cluster centers
- Repeat E & M till convergence $new \mu_j = old \mu_j$

K-Means as EM algorithm steps rigorously

- **Guessing step**: Randomly guess k centroids

$$\mu_1, \dots, \mu_j, \dots, \mu_k$$

$O(\text{iterations} * \text{data points} * \text{clusters} * \text{features})$

- **Expectation Prep**

$$\forall j \in [1, k] \quad C_j = \{\} \quad \mu_{j_{new}} = \mu_{j_{old}}$$

- **Expectation Step**: Cluster assignment for each $x(i)$

$$C_j \leftarrow C_j + \{ \forall j \in [1, k] \arg \min_j D(x^{(i)}, \mu_j) \}$$

- **Maximization Step**: Set the new clusters

$$\mu_{j_{new}} = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

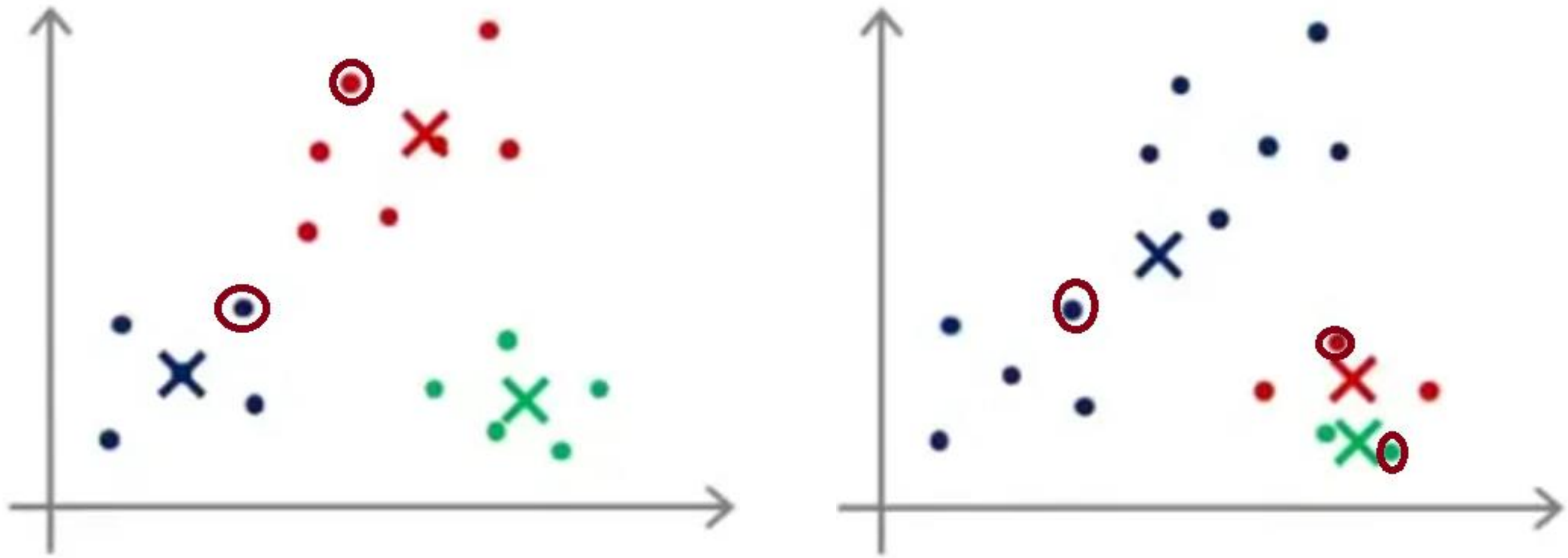
- Go to Expectation Prep & loop if $\mu_{j_{new}} \neq \mu_{j_{old}}$



K-Means initialization

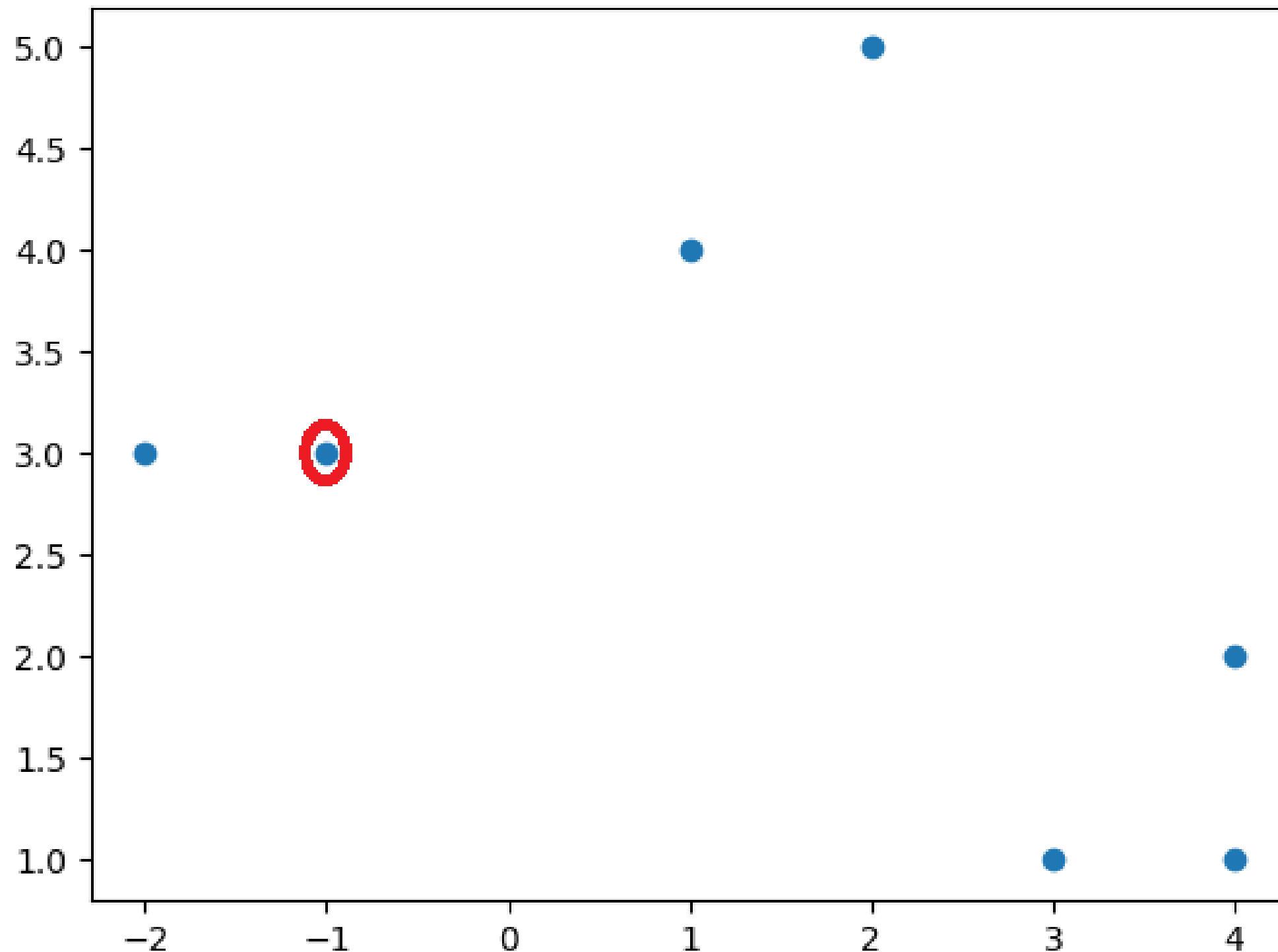
Impact of non-deterministic K-Means algorithm

- Might get wrong centroids & wrong clusters in every run
- $K = 3$ and random initialization



How to initialize initial centroids

- First centroid chosen randomly
- Centroids should be as far as possible from each other



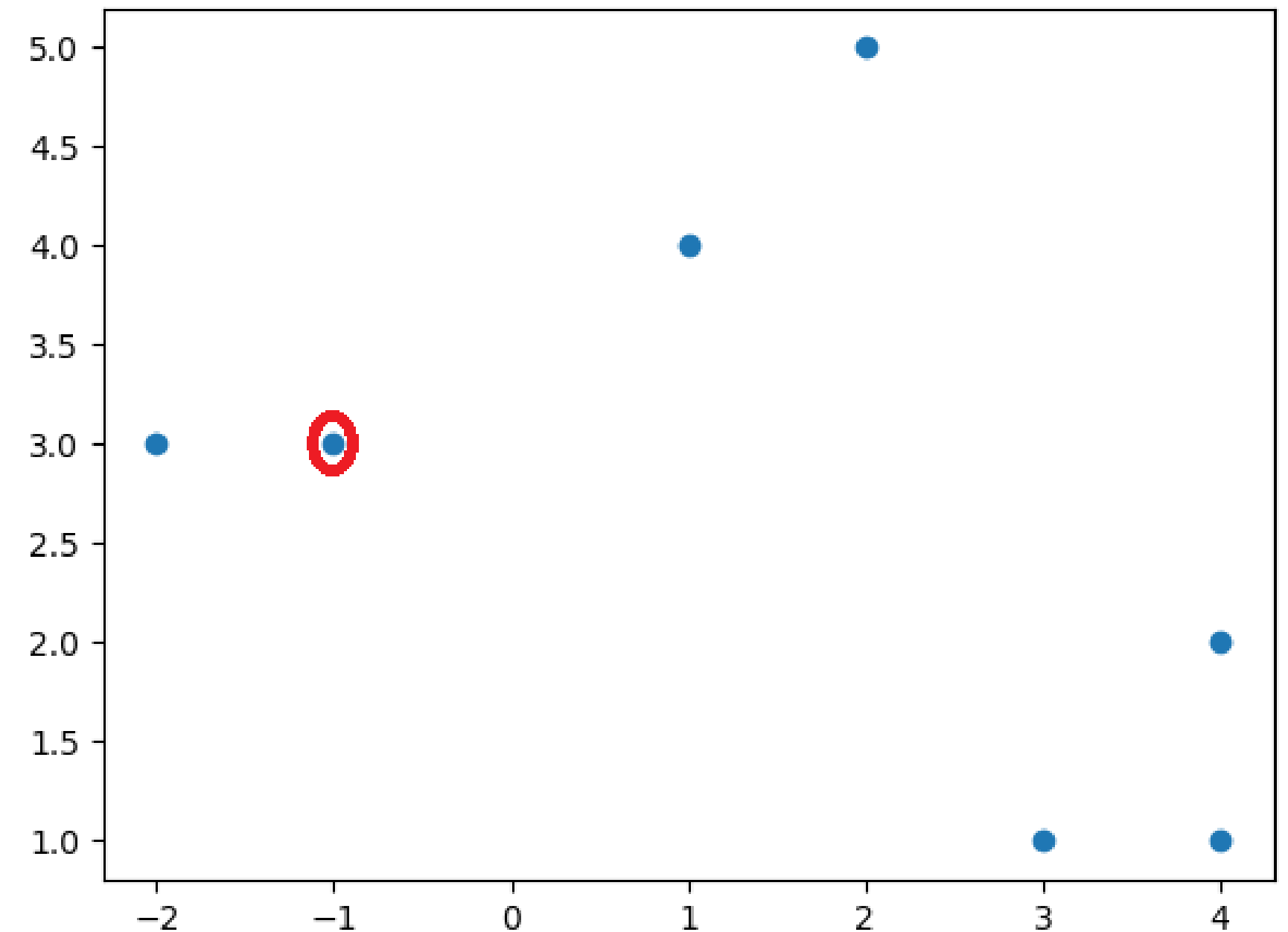
Different initialization different results

- Random
 - We have been doing this
- Kmeans++
 - Will cover next
- Naïve sharding
 - not covered

Kmeans++

- First centroid chosen randomly
- Probability of next centroid selection proportional to distance

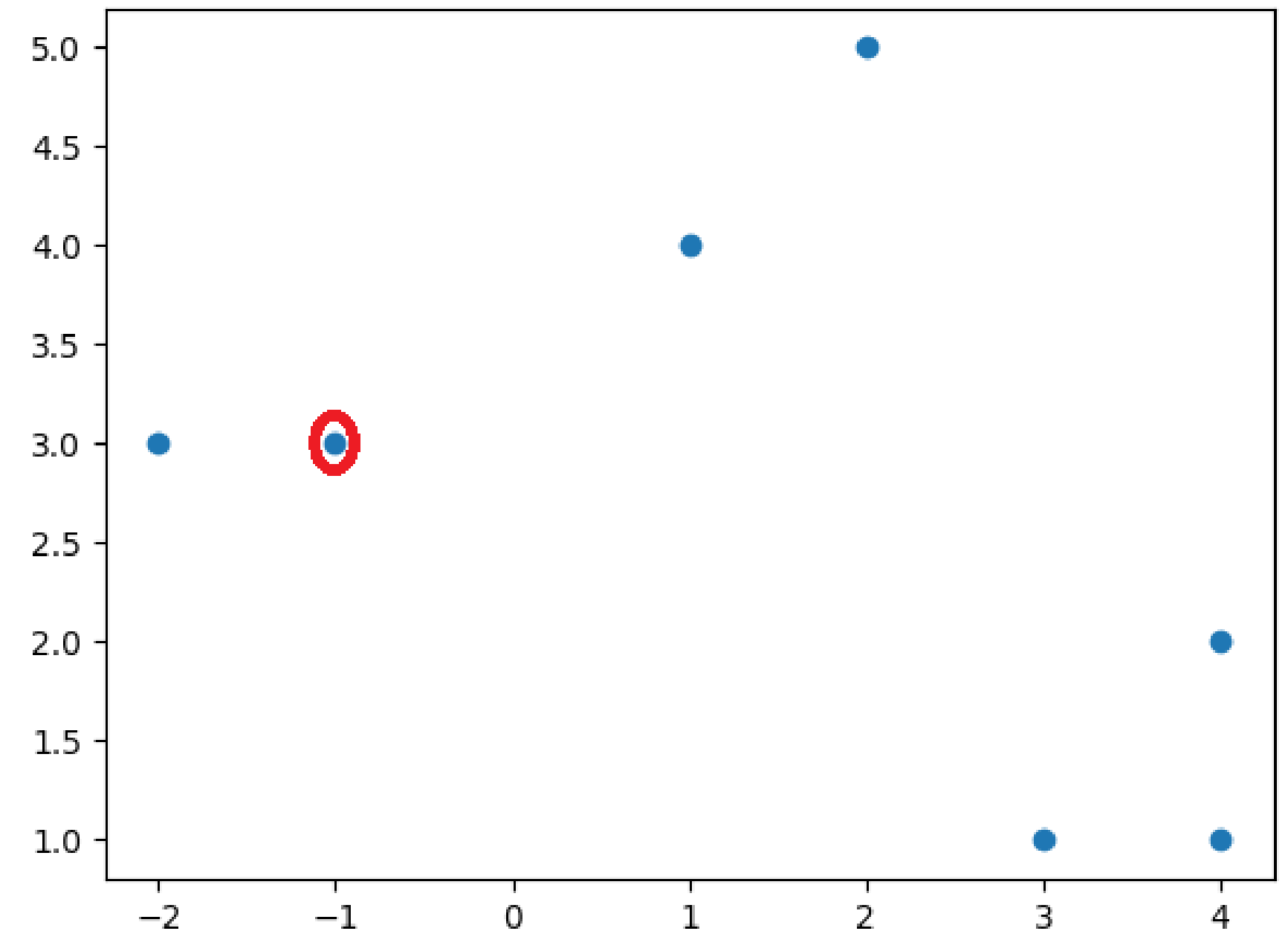
X	Dist (x, c1)^2
(2,5)	
(-1,3)	Centroid
(-2,3)	
(3,1)	
(1,4)	
(4,1)	
(4,2)	
Total	



Kmeans++

- Calculate all distances square from centroid

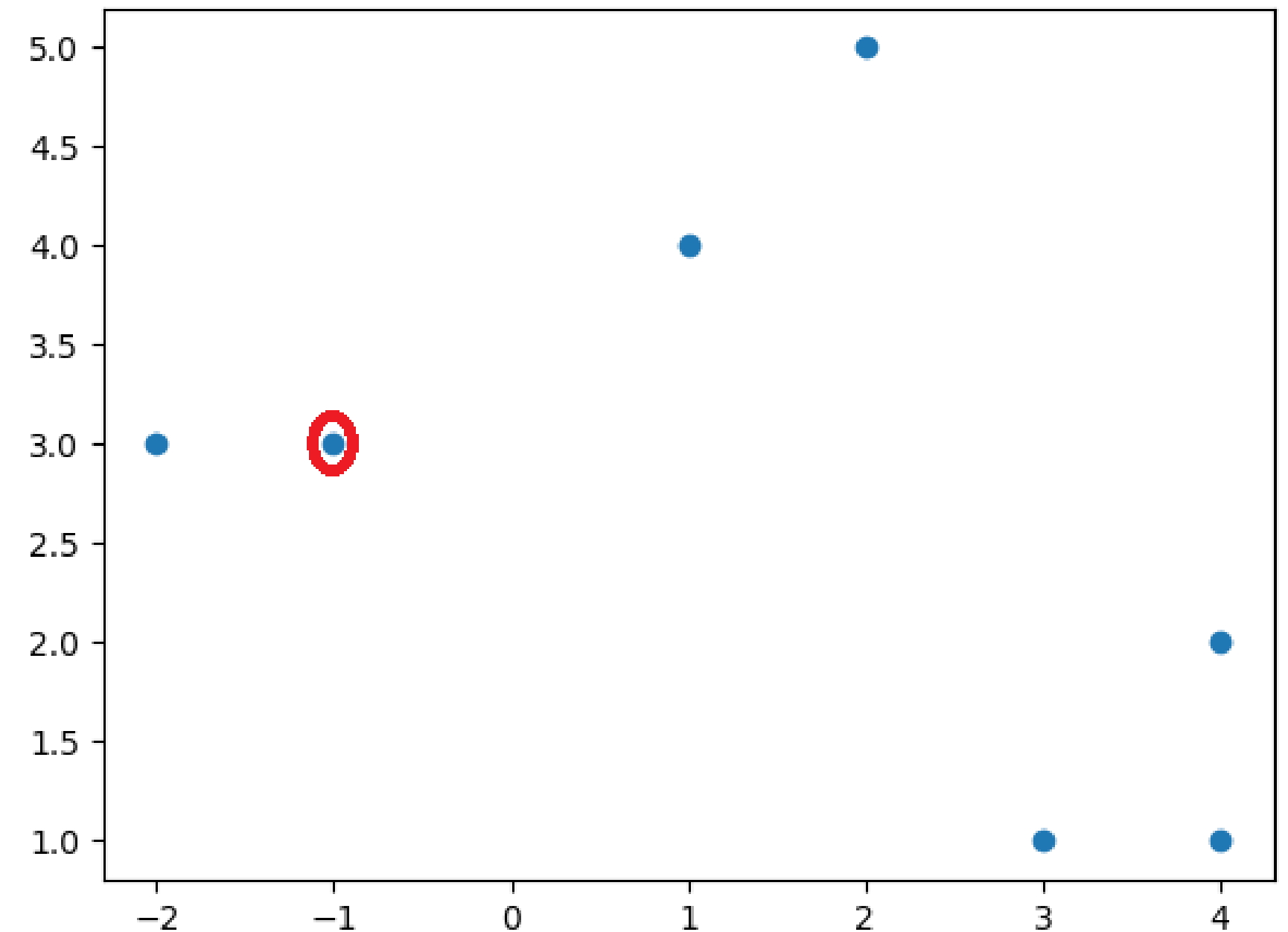
X	Dist (x, c1)^2
(2,5)	13
(-1,3)	Centroid
(-2,3)	1
(3,1)	20
(1,4)	5
(4,1)	29
(4,2)	26
Total	94



Kmeans++

- Calculate all distances square from centroid
- Convert to probability

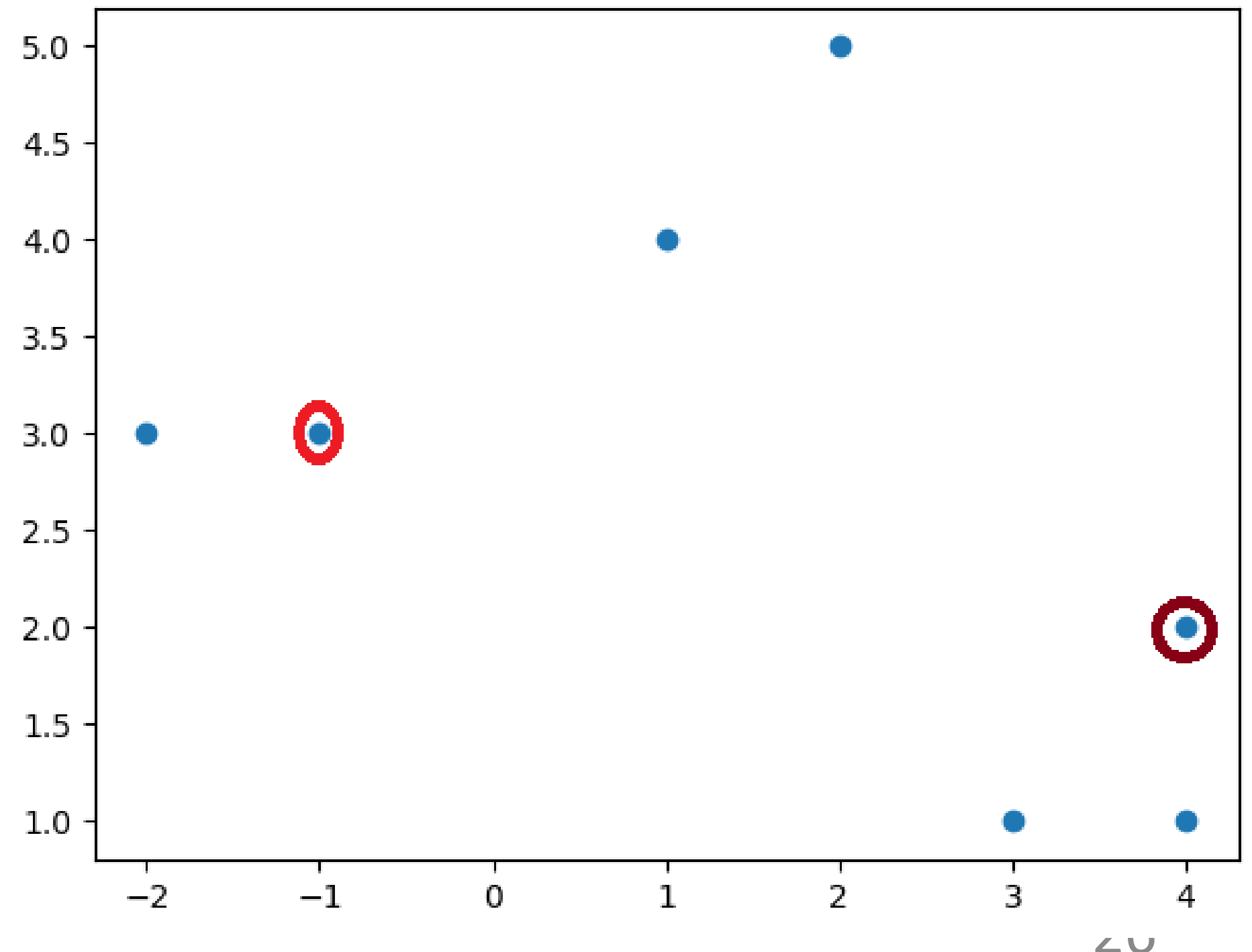
X	Dist (x, c1)^2	Prob
(2,5)	13	13/94
(-1,3)	Centroid	—
(-2,3)	1	1/94
(3,1)	20	20/94
(1,4)	5	5/94
(4,1)	29	29/94
(4,2)	26	26/94
Total	94	



Kmeans++

- Sample from remaining points weighted by probabilities
- Find the minimum of distance from both centroids

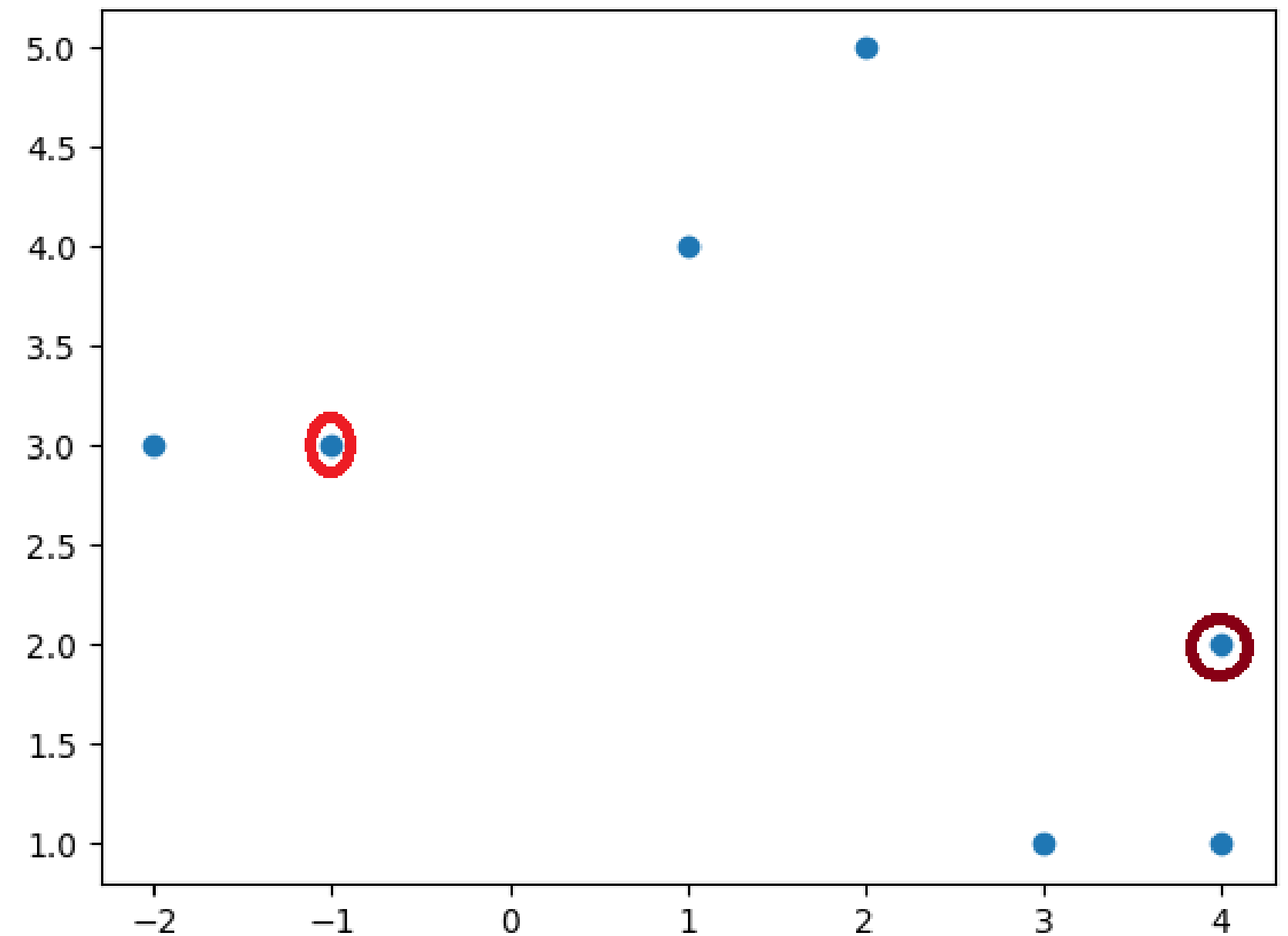
X	Dist (x, c1)^2	Prob	Dist(x, c1, c2)^2
(2,5)	13	13/94	min(13,13)
(-1,3)	Centroid	—	—
(-2,3)	1	1/94	min(1,37)
(3,1)	20	20/94	min(20, 2)
(1,4)	5	5/94	min(5,13)
(4,1)	29	29/94	min(29,1)
(4,2)	26	26/94	Centroid
Total	94		Not min(94, 92), But 22



Kmeans++

- Normalize the probabilities again and do weighted sampling

X	Dist (x, c1)^2	Prob	Dist(x, c1, c2)^2	Prob
(2,5)	13	13/94	min(13,13)	13/22
(-1,3)	Centroid	—	—	—
(-2,3)	1	1/94	min(1,37)	1/22
(3,1)	20	20/94	min(20, 2)	2/22
(1,4)	5	5/94	min(5,13)	5/22
(4,1)	29	29/94	min(29,1)	1/22
(4,2)	26	26/94	Centroid	—
Total	94		Not min(94, 92), But 22 *****	

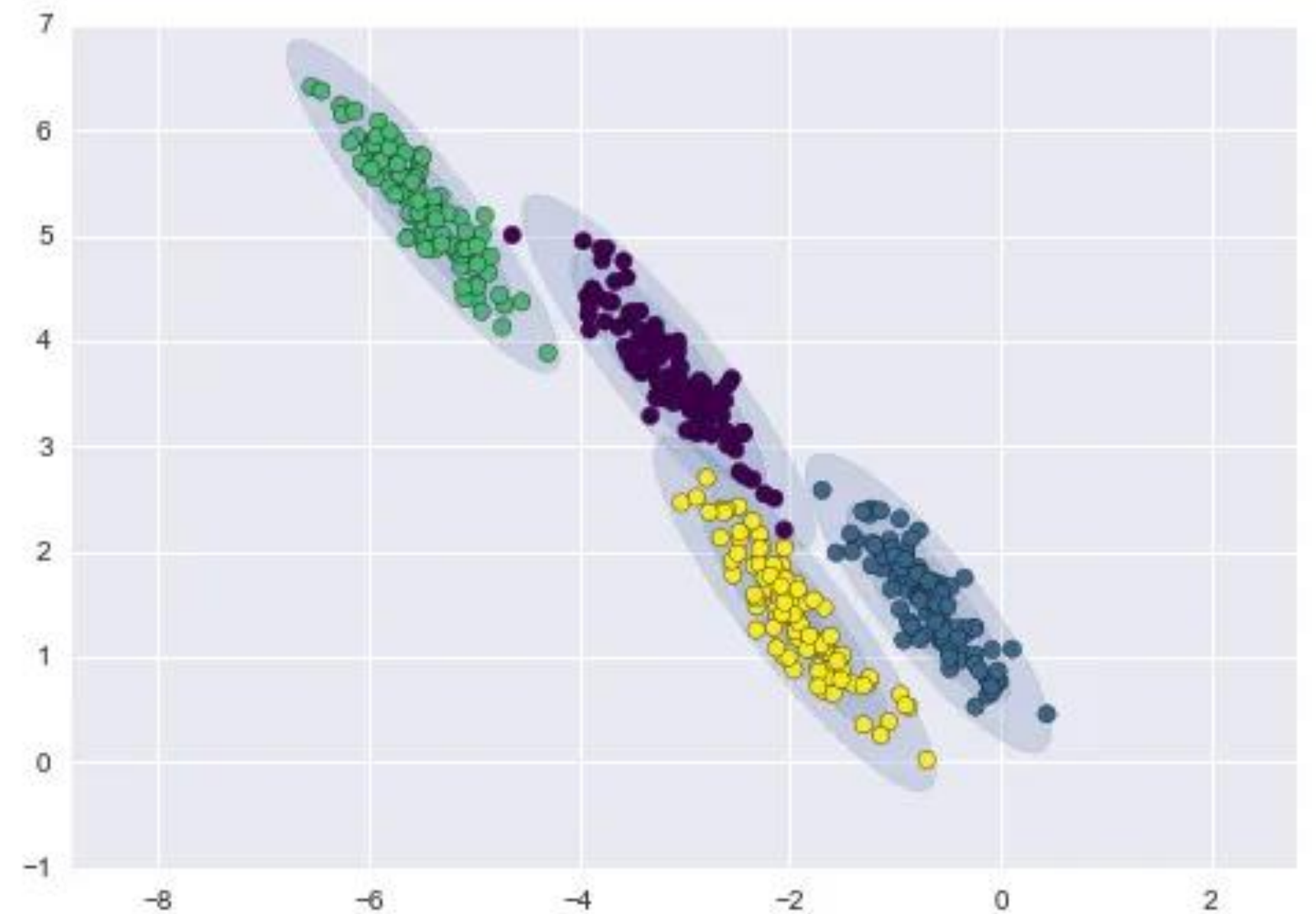
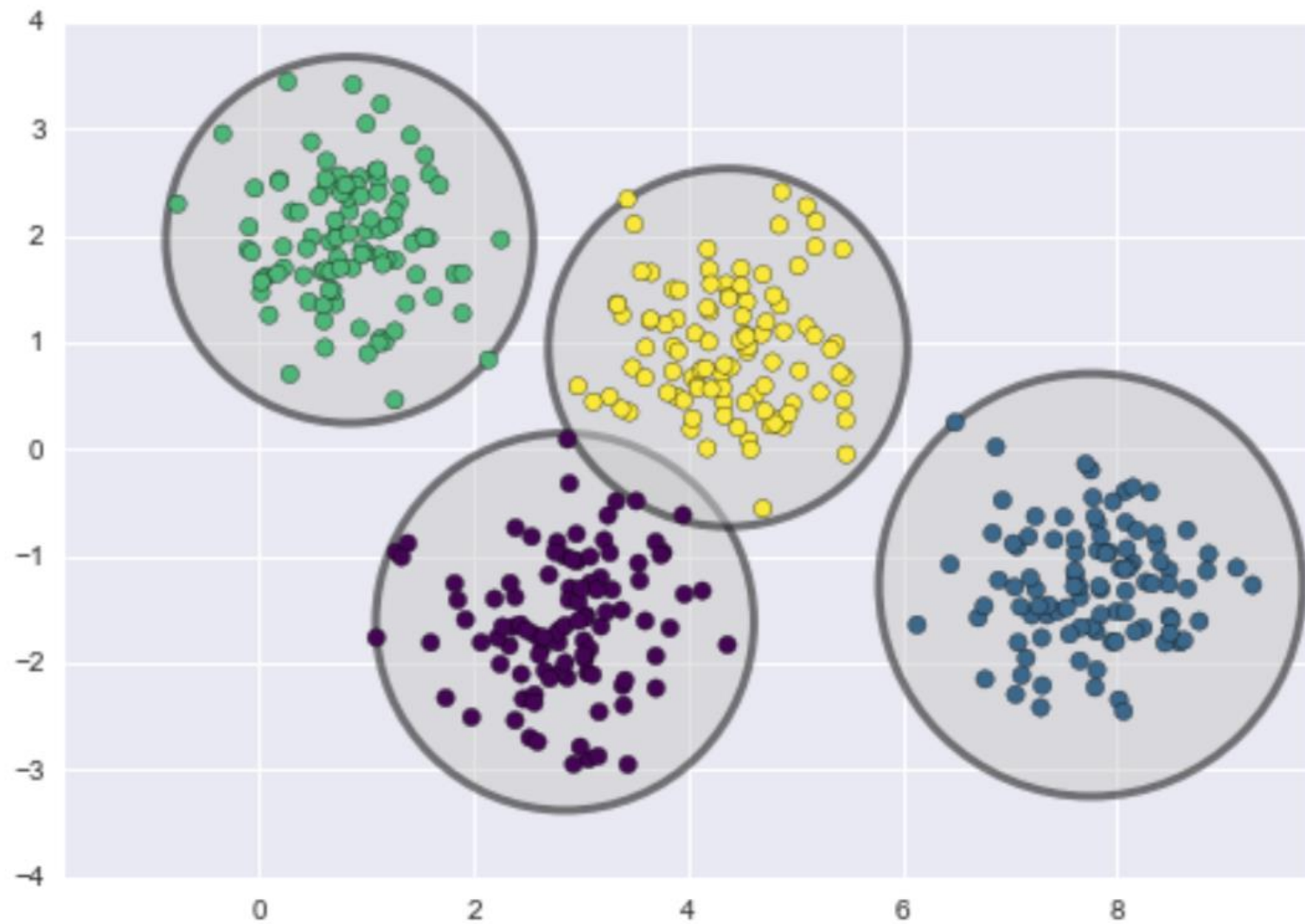




K-Means limitations

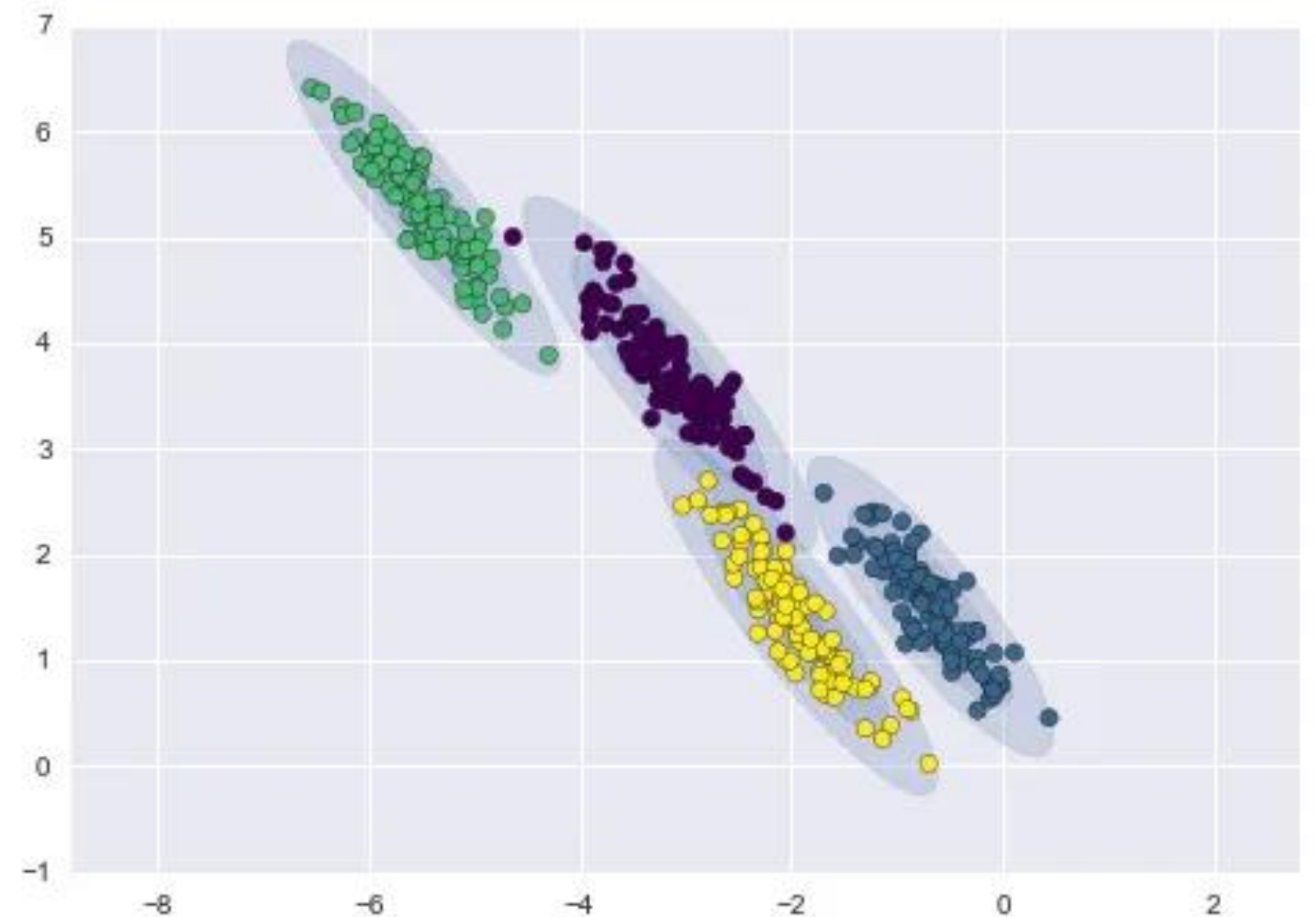
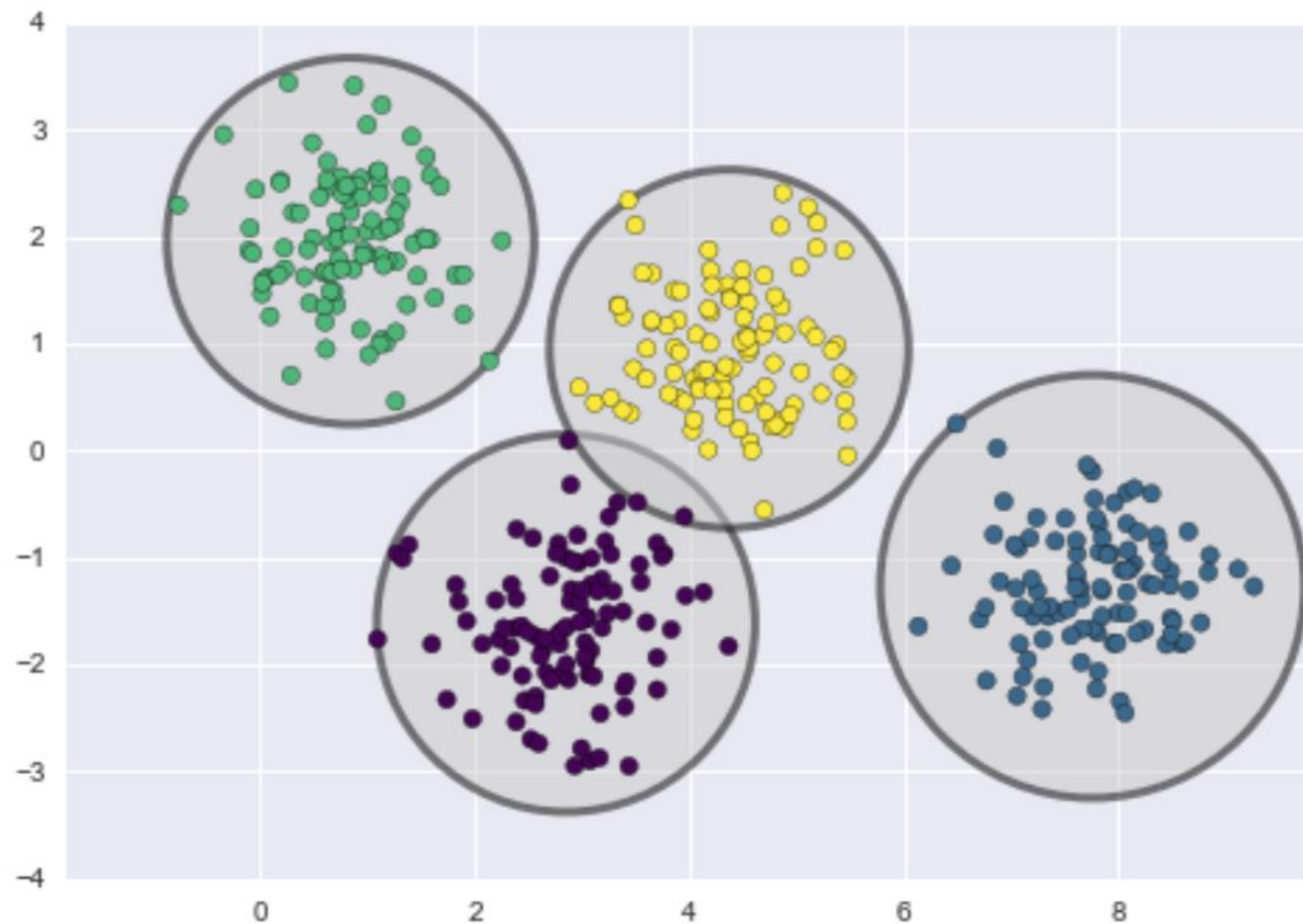
K means limitations

- Oblong data cannot be clustered well
- Recall nearest centroid classification



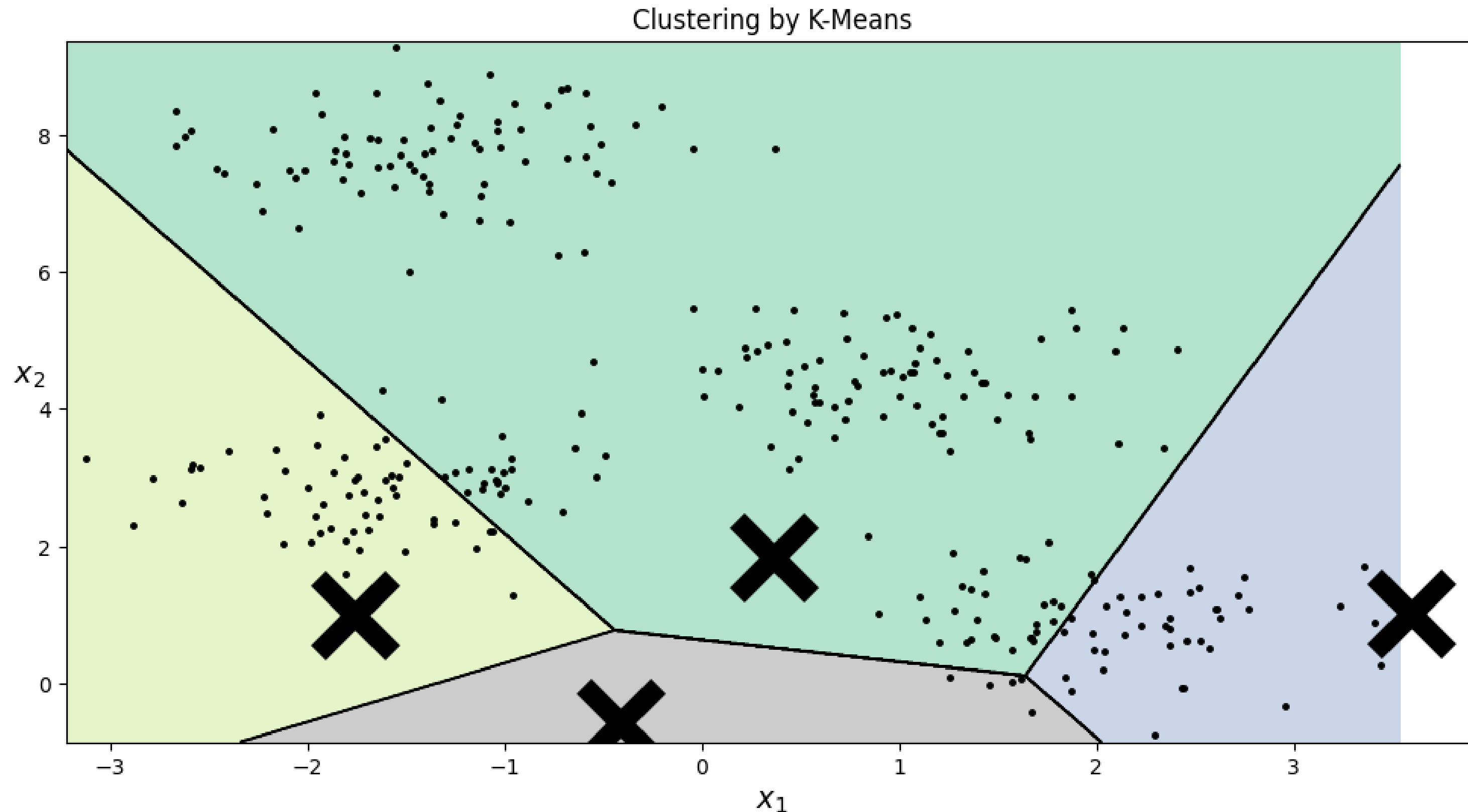
K means limitations

- Clustering every time when new data comes is hard
- Could use extracted centroids as model
 - Nearest Centroid from scratch without labels!!



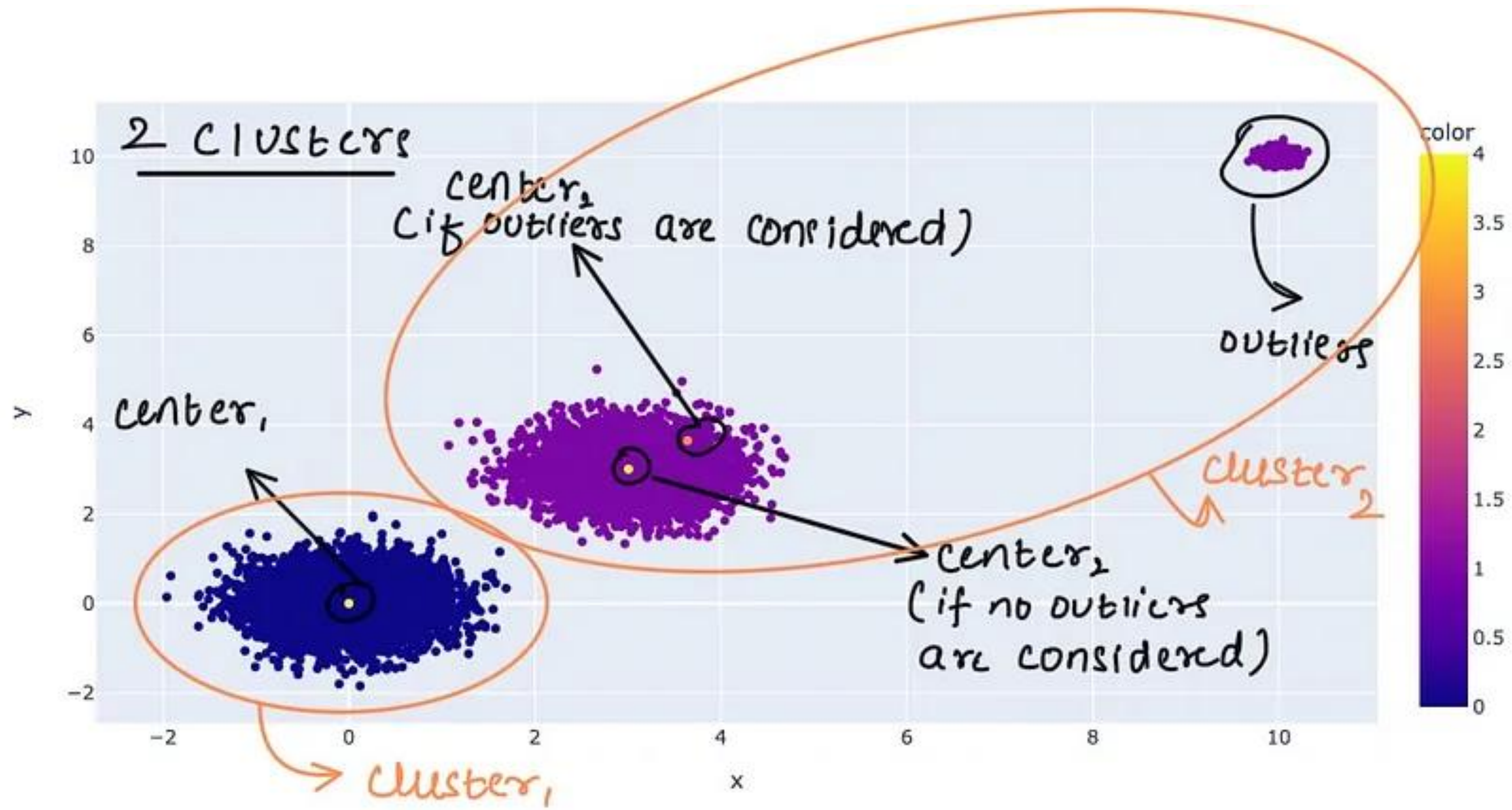
K means decision boundary

- Linear decision boundary

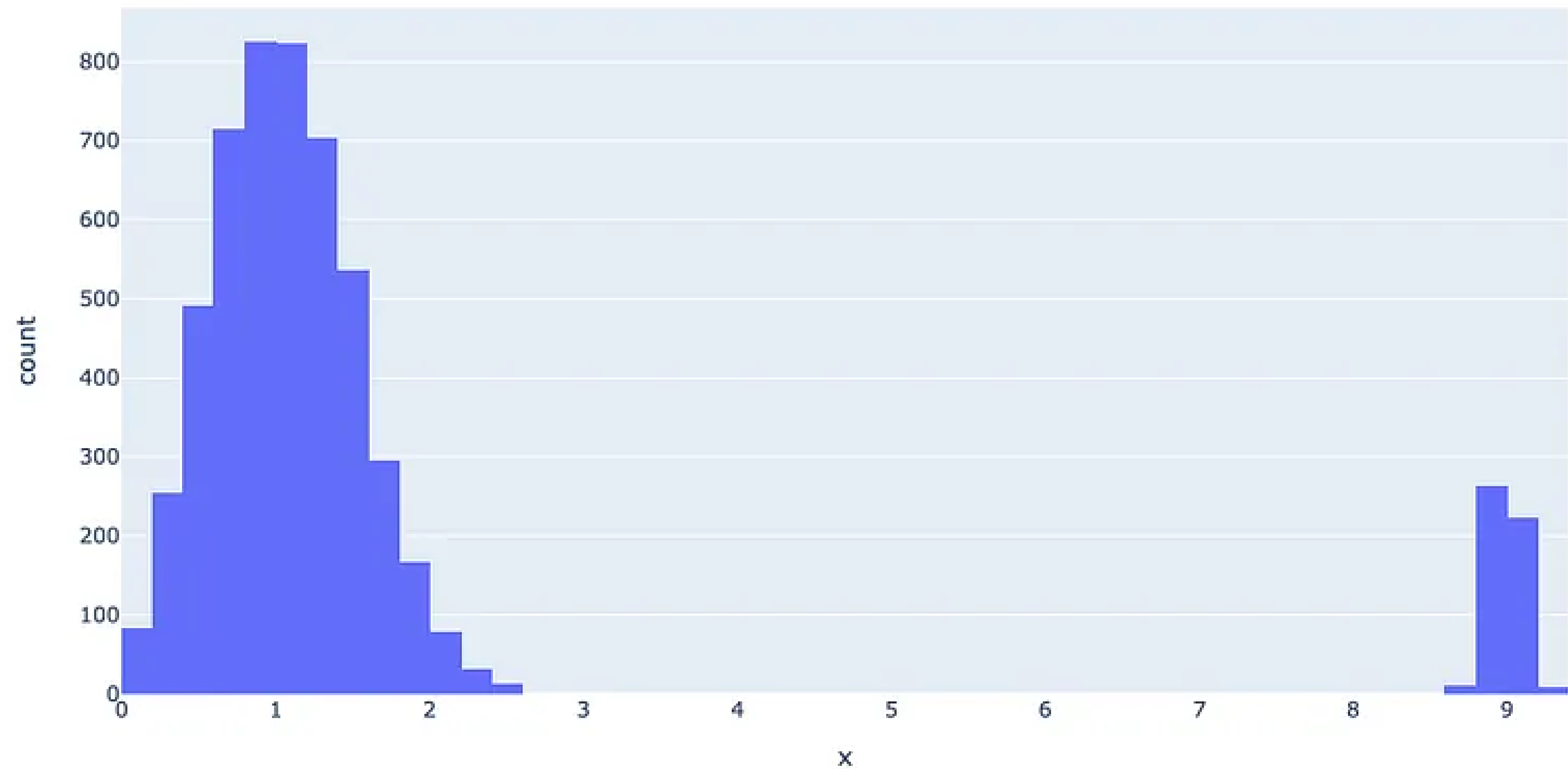




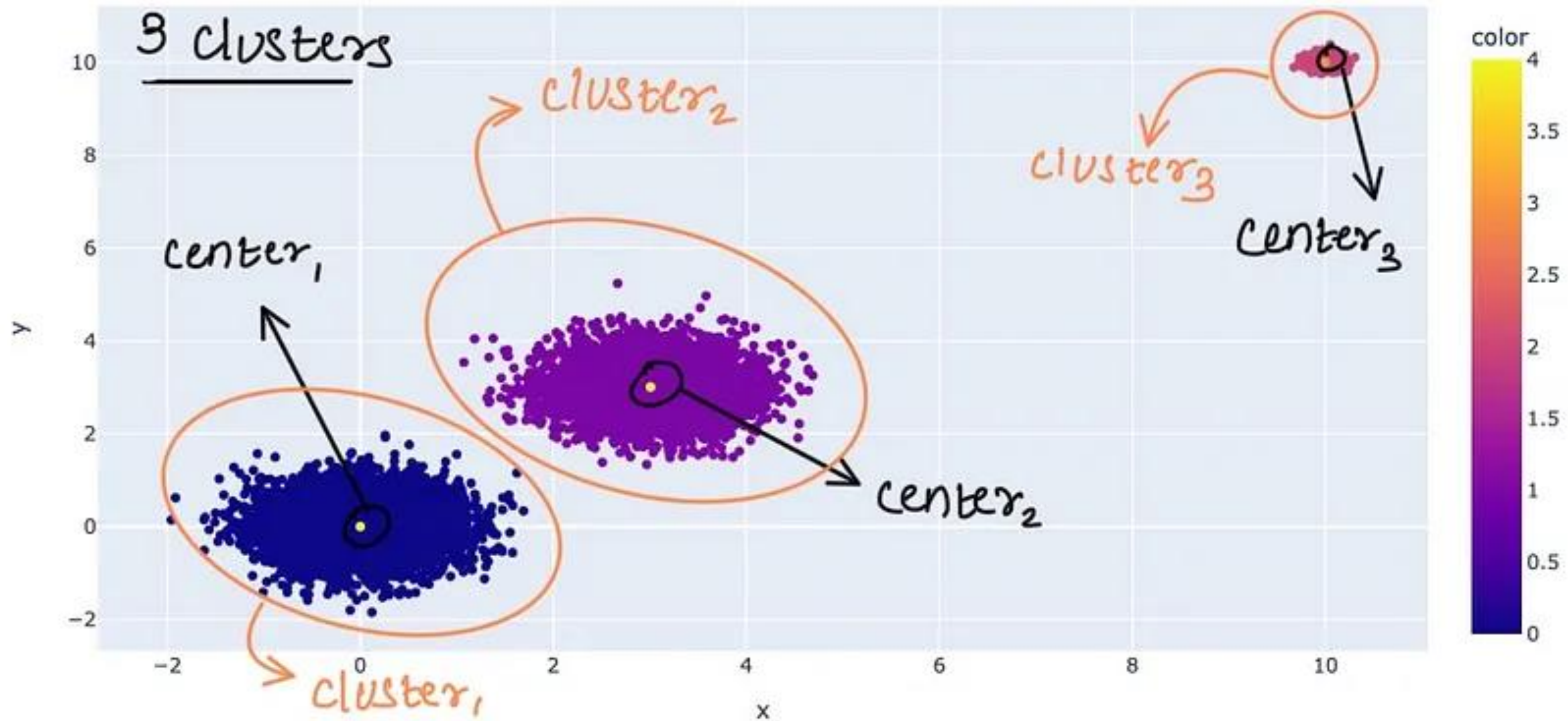
K-Means is sensitive to outliers



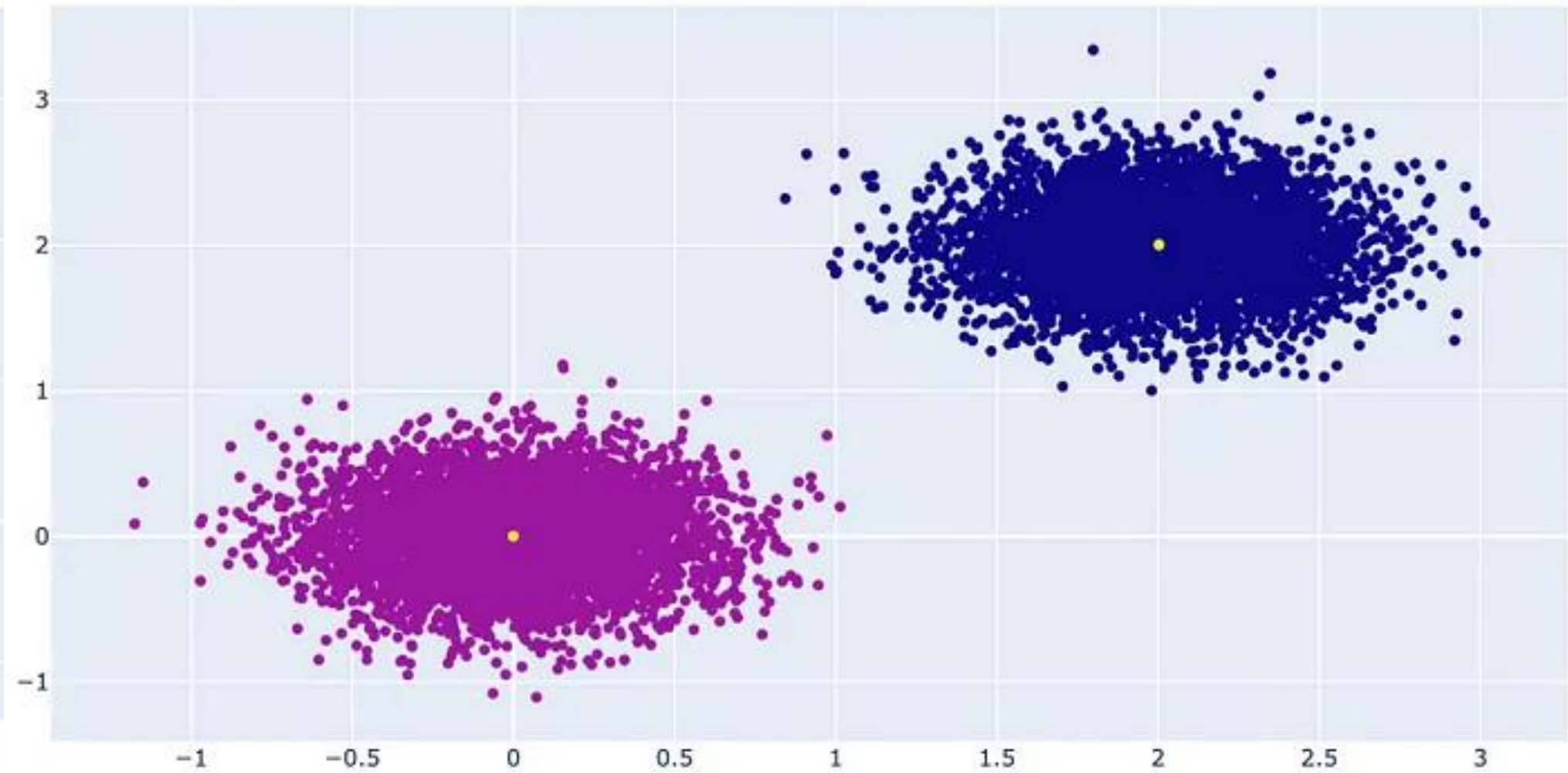
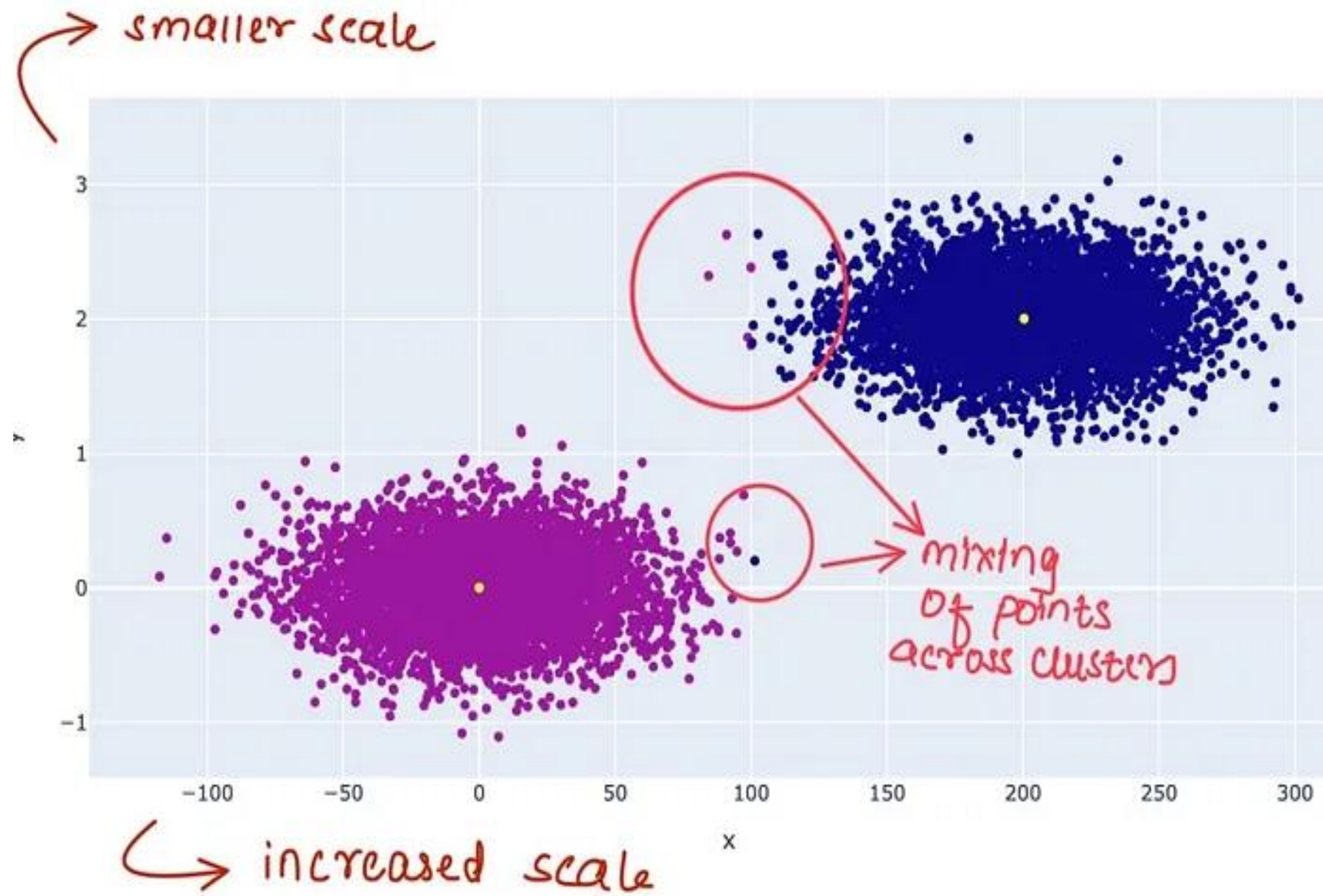
K-Means is sensitive to outliers



K-Means is sensitive to outliers

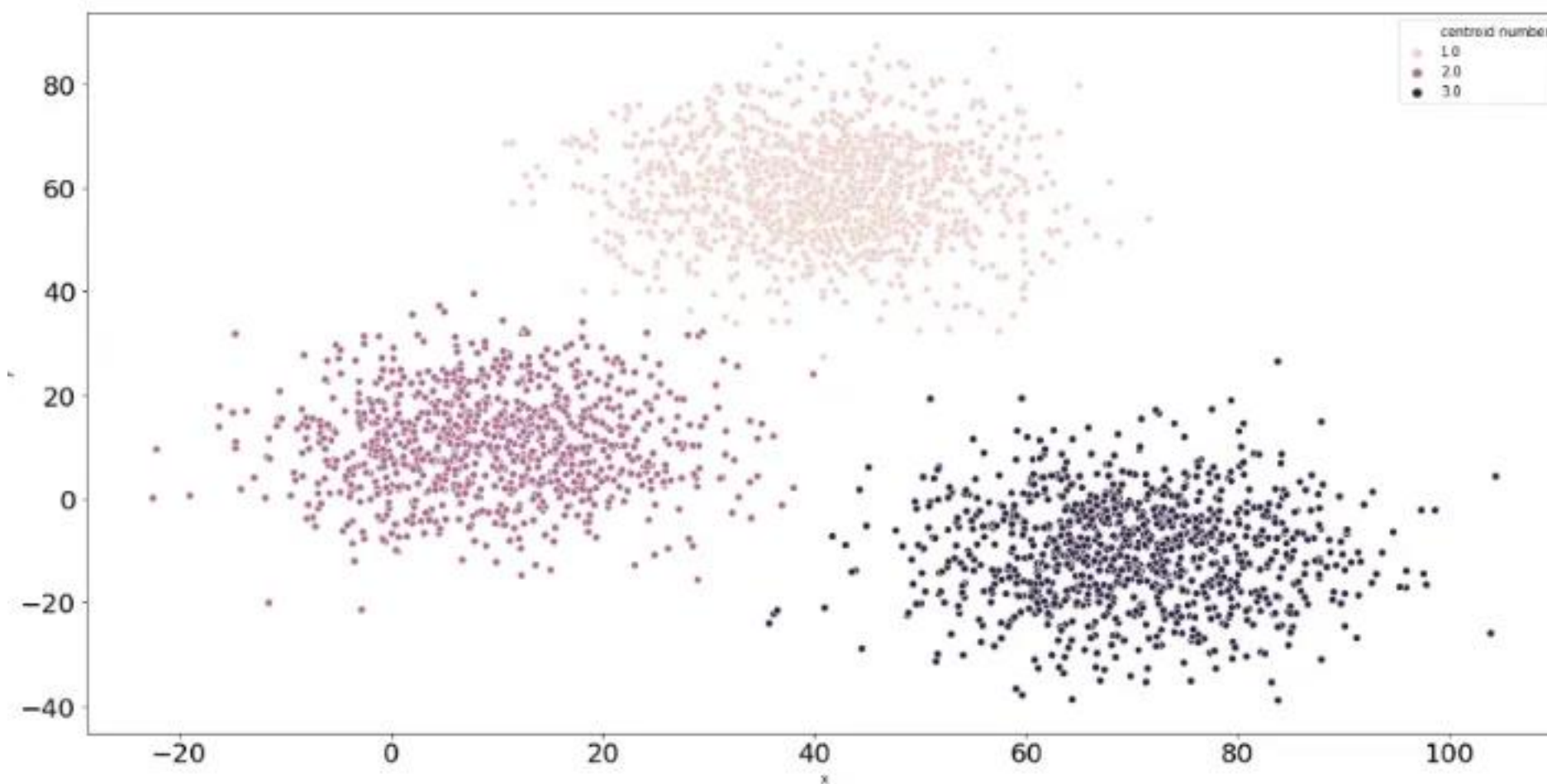


K-Means and scaling



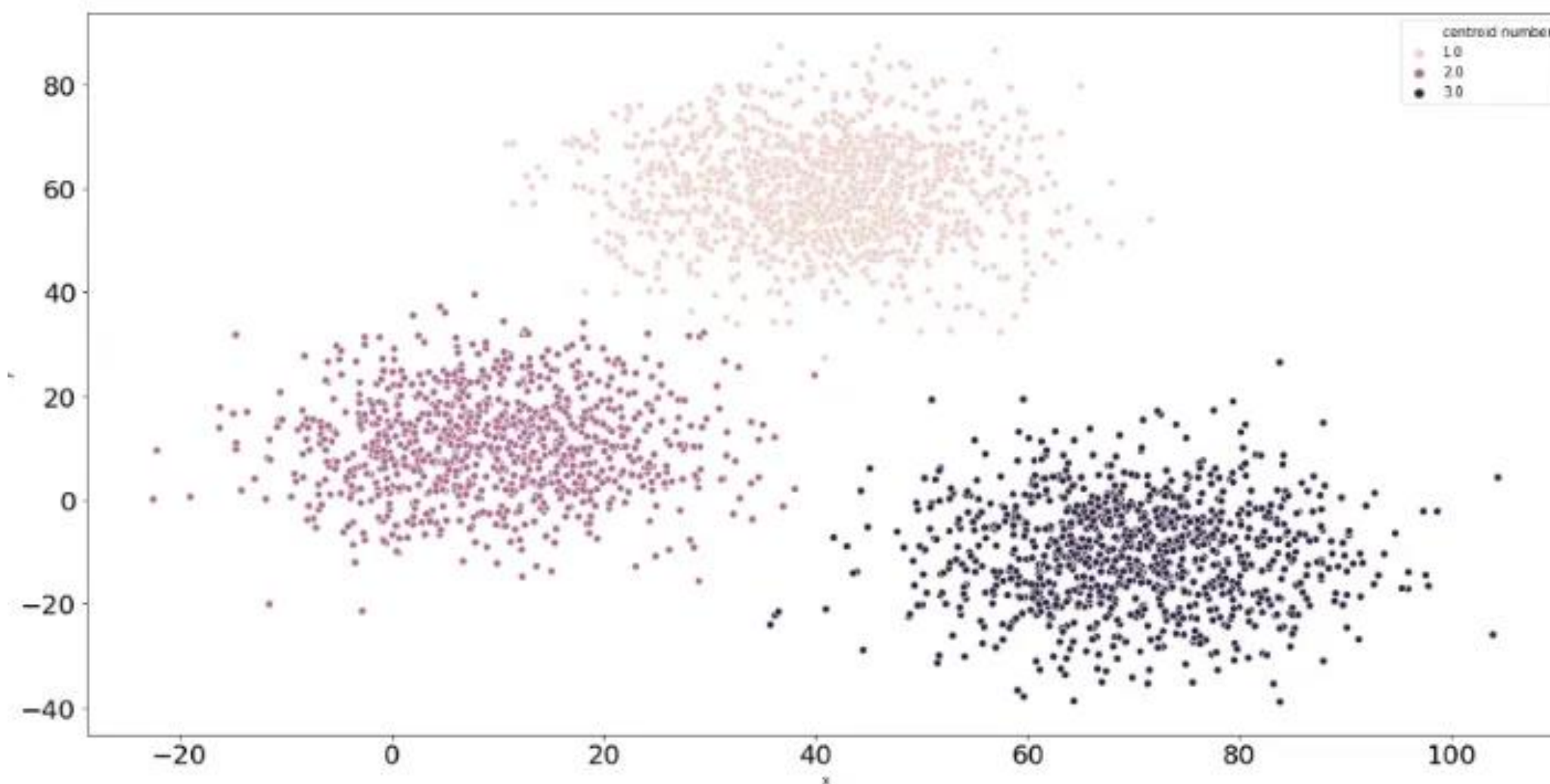
K means limitations

- K-Means is sensitive to outliers
 - Clusters become bigger to accommodate outliers
- Can use statistical techniques or silhouette analysis

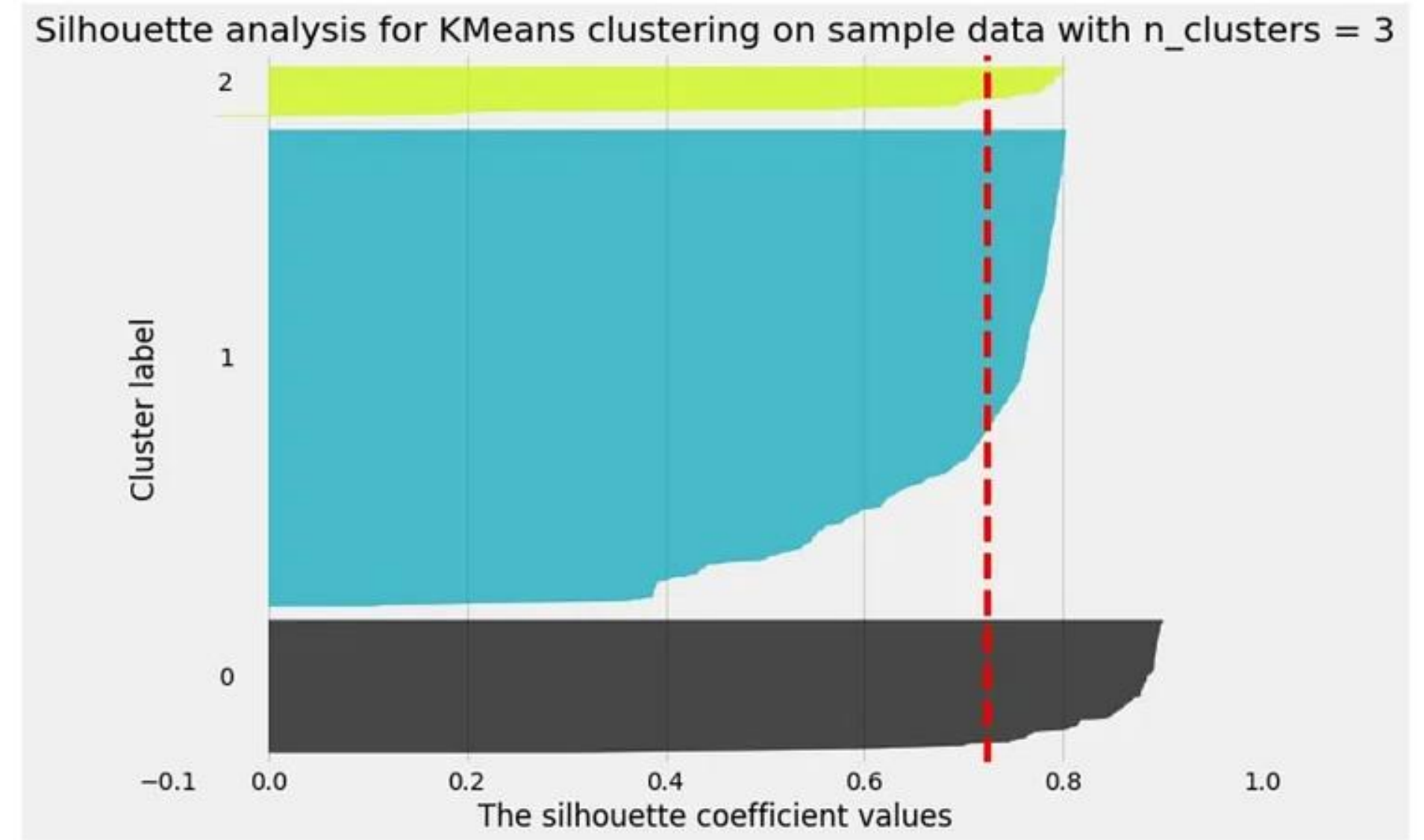
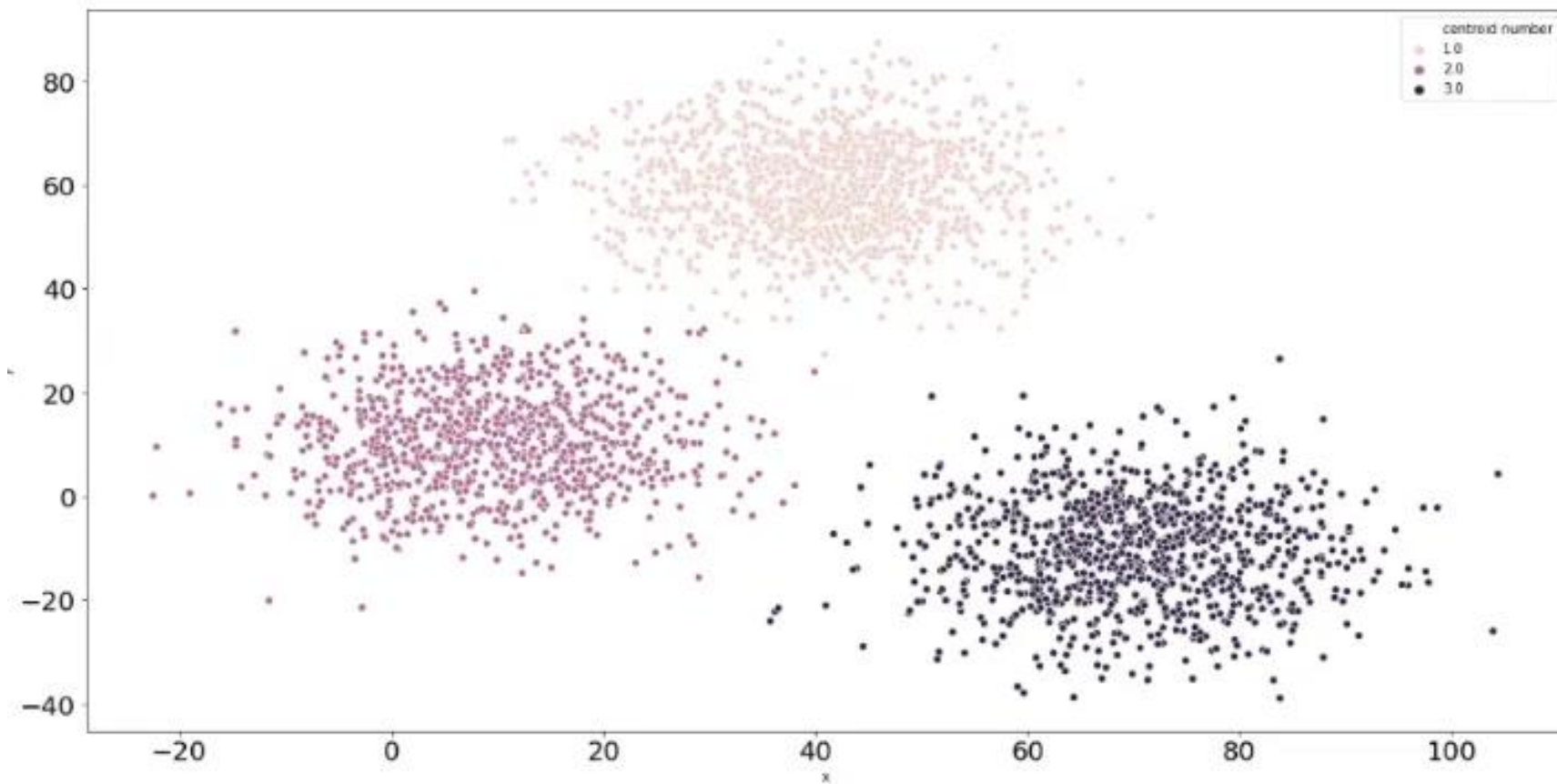


Handling outliers with K means & statistical analysis

- Statistical techniques to analyze each cluster
 - IQR
 - 3 standard deviations



Handling outliers with K means & silhouette analysis



- Analyze Silhouette plots for outlier detection
 - Variation of score gives clues of outlier presence
 - Negative values are sure outliers
 - Values with score less than 0.4 are potential outliers