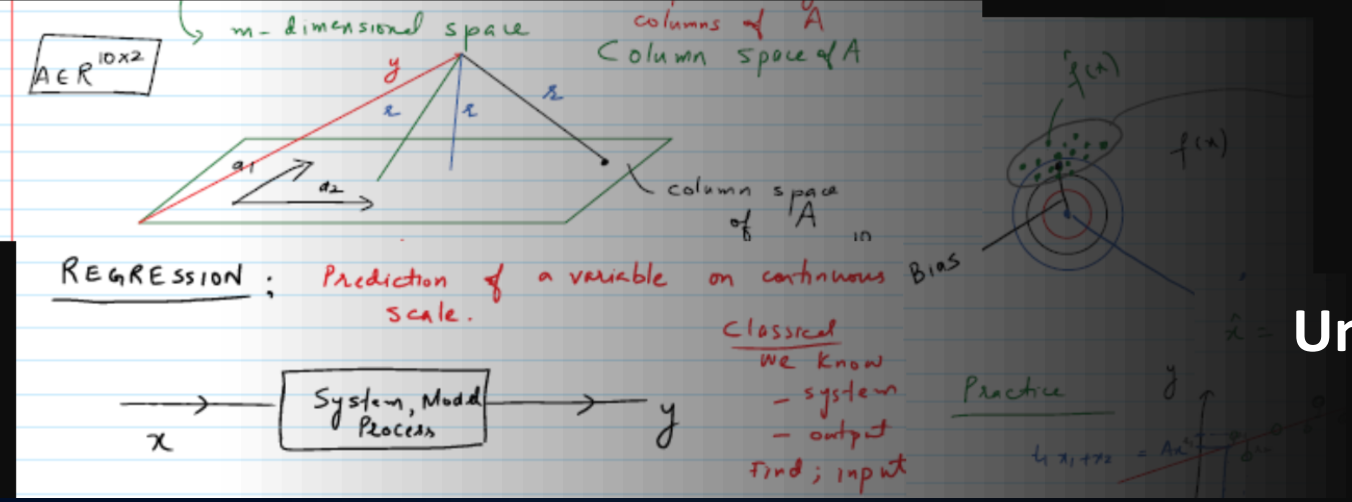


Machine Learning
EE514 – CS535

Unsupervised Learning: Clustering

Zubair Khalid

School of Science and Engineering
Lahore University of Management Sciences

https://www.zubairkhalid.org/ee514_2021.html

LUMS
A Not-for-Profit University

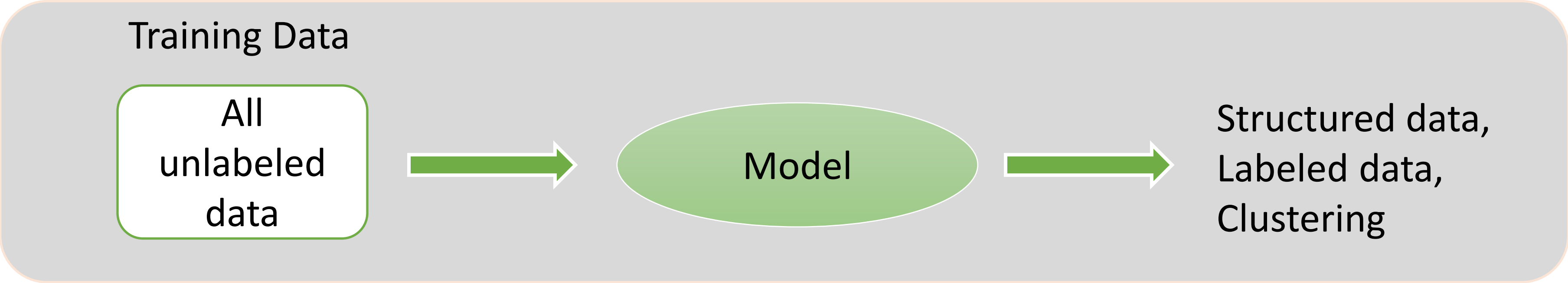
Outline

- Introduction to Unsupervised Learning, Clustering
- Clustering Overview
- Partitional Clustering
    - K-means Clustering
- Hierarchical Clustering
    - Agglomerative Clustering

Unsupervised Learning

Overview:

The learning algorithm would receive unlabeled raw data to train a model and to find patterns in the data

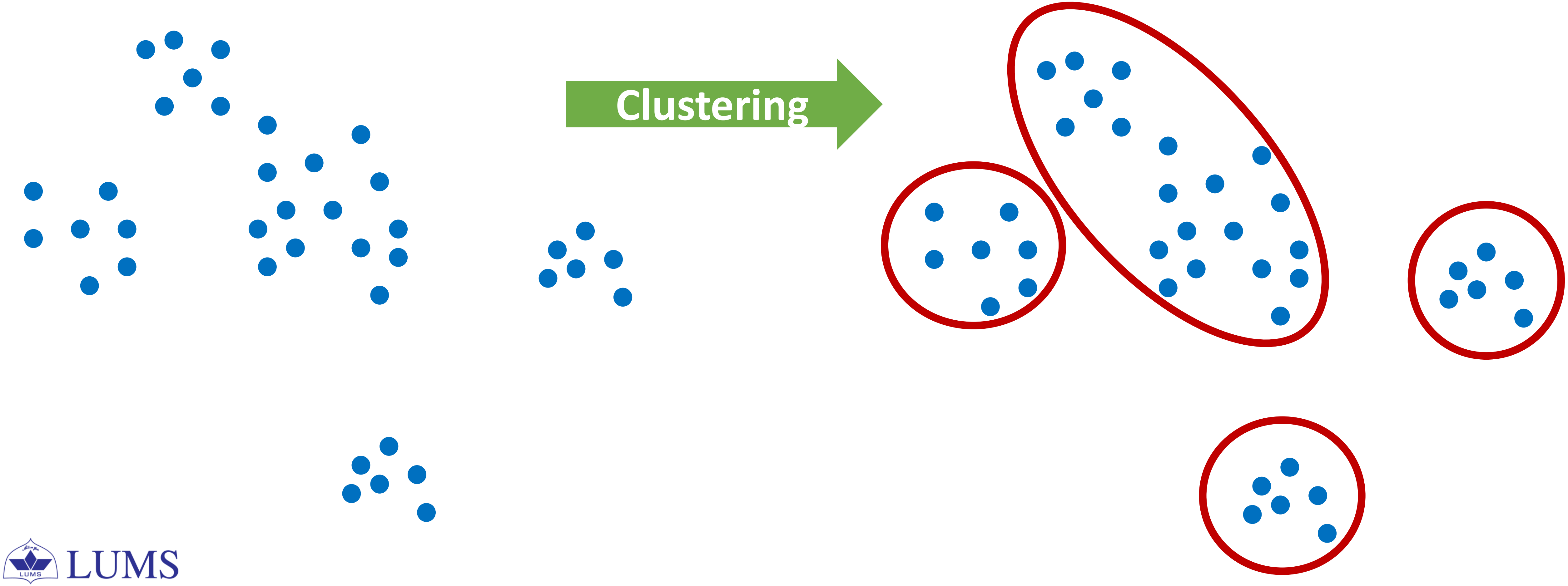


Clustering, aka unsupervised learning (due to historical reasons), is the most widely used technique.

Clustering

Overview:

Given the data, group 'similar' points into the form of clusters.



Clustering

Overview:

- *Idea:* the process of grouping data into similarity groups known as **clusters.**
- Formally, organize the unlabeled data into classes such that
  - **Inter-cluster similarity is minimized:**
    - low similarity between data points in different classes
  - **Intra-cluster similarity is maximized:**
    - High similarity between data points of each class
- In contrast to classification, we learn the number of classes and class labels directly from the data.

Clustering

Applications of Clustering:

Marketing: Clustering is used for segmentation of the customers/markets to do targeted marketing.

- *Spatio-temporal demographic distribution of the sales of products*
- *Insurance companies cluster policy holders to identify groups of policy holders with a high claim costs on average*

Text Analysis: Grouping of a collection of text documents with respect to similarity in their content.

- *Grouping of news items when you search for an item*

Anomaly Detection: Given data from the sensors, grouping of sensor readings for machine operating in different states and detect anomaly as an outlier.

Finance: Allocation of diversified portfolios of stocks by using clustering.

Earth-quake studies: Clustering of epi-centers of earthquakes are distributed around or along fault lines.

Clustering

Aspects of Clustering:

- Given the data, what do we need to carry out clustering?
  - A measure to quantify or determine similarity
  - A criterion to evaluate the quality of the clustering
    - *Low inter-class similarity, High intra-class similarity*
    - *Ability to identify hidden patterns in the data*
  - Clustering techniques/algorithms for grouping similar data points
    - *Partitional Clustering*
    - *Hierarchical Clustering*
    - *Model Based*
    - *Density Based*

Clustering

(Dis)Similarity using distance metric:

- Mathematically, we quantify dissimilarity (or distance) between two data points $\mathbf{x}$, $\mathbf{x}'$ using a real number given by distance function $\text{dist}(\mathbf{x}, \mathbf{x}')$.

- We require this distance function $\text{dist}(\mathbf{x}, \mathbf{x}')$. to satisfy following properties:

  - $\text{dist}(\mathbf{x}, \mathbf{x}') \geq 0$ — *Non-negativity*
  - $\text{dist}(\mathbf{x}, \mathbf{x}') = \text{dist}(\mathbf{x}', \mathbf{x})$ — *Symmetry*
  - $\text{dist}(\mathbf{x}, \mathbf{x}') = 0 \iff \mathbf{x} = \mathbf{x}'$ — *Self-Similarity*
  - $\text{dist}(\mathbf{x}, \mathbf{x}') \leq \text{dist}(\mathbf{x}', \mathbf{x}'') + \text{dist}(\mathbf{x}'', \mathbf{x})$ — *Triangular Inequality*

- *We studied earlier; Minkowski, Euclidean distance, Manhattan distance, Chebyshev distance, cosine distance*

- *For categorical variables, we use Hamming distance*

Clustering

Evaluation of Clustering:

- Unlike supervised learning problems, the evaluation of quality of a clustering is a hard problem and is mostly subjective as the information about correct clusters is unknown.

Evaluation criteria:

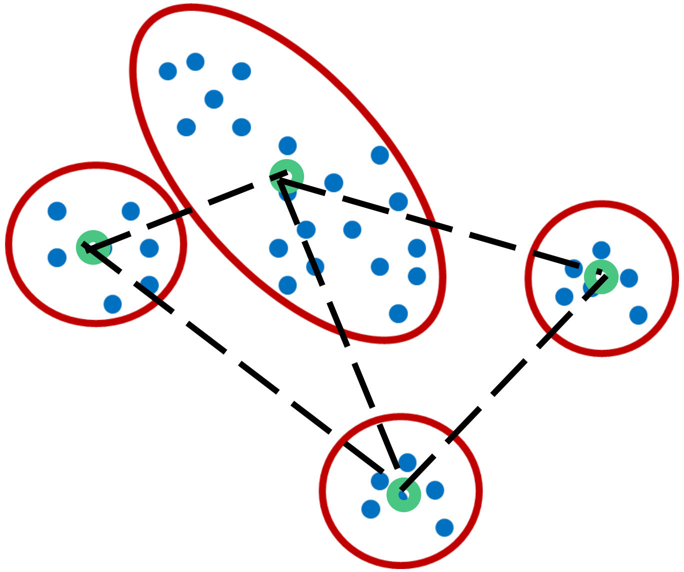
- **Using Internal Data:**
  - Use the unlabeled data for evaluation of the clustering algroithm.

- **Using External Data:**
  - Use labeled data (supervised learning) to evaluate the performance of different clustering algorithms.

Clustering

Evaluation of Clustering using Internal Data:

- **Inter-cluster separability**
  - *measure of the isolation of the cluster*
  - *E.g., measured as the distance between the centroids of the clusters*



- **Intra-cluster cohesion**
  - *measure of the compactness of the cluster*
  - *E.g., measured by the sum of squared error that quantifies the spread of the points around the centroid.*

Clustering

Evaluation of Clustering using External (Labeled) Data:

- Use labeled data to carry out clustering and measure the extent to which the external class labels match the cluster labels.

- Idea: Evaluation of clustering performance using the labeled data gives us some confidence about the performance of the algorithm.

- This evaluation method is referred to as evaluation based on external data.

- Assuming each class as a cluster, we use classification evaluation metrics after clustering:
  - Confusion matrix
  - Precision, recall, F1-score
  - Purity and Entropy

Clustering

Evaluation of Clustering using External (Labeled) Data:

- Assume we have M classes and data D with label associated with each data point is $y \in \{1, 2, \ldots, M\}$. The clustering method produces M clusters that divides data D into M disjoint subsets $D_1, D_2, \ldots, D_M$.

Entropy:

Measure of the proportion of different classes in each cluster.

For each cluster D_i, entropy is measured as follows

$$\text{entropy}(D_i) = -\sum_{j=1}^{M} R_i(j) \log R_i(j), \qquad R_i(j) = \frac{\#\text{ of points of class } j \text{ in cluster } i}{\#\text{ of points of class } j}$$

The total entropy is given by

$$\text{entropy}(D) = \sum_{i=1}^{M} \frac{|D_i|}{|D|} \text{entropy}(D_i)$$

Clustering

Evaluation of Clustering using External (Labeled) Data:

Purity:

Also serves as a measure of the proportion of different classes in each cluster.

For each cluster D_i, purity is measured as follows

$$\text{purity}(D_i) = \max_{j=1,2,\ldots,M} R_i(j) \qquad R_i(j) = \frac{\#\,\text{of points of class } j \text{ in cluster } i}{\#\,\text{of points of class } j}$$

The total purity is given by

$$\text{purity}(D) = \sum_{i=1}^{M} \frac{|D_i|}{|D|}\,\text{purity}(D_i)$$

Remark:

- *Since we do not have labels associated with the data for the clustering problem; it **must** be noted that the good performance on the labeled data does not guarantee good performance on the data with no labels.*

Clustering

Clustering Algorithms:

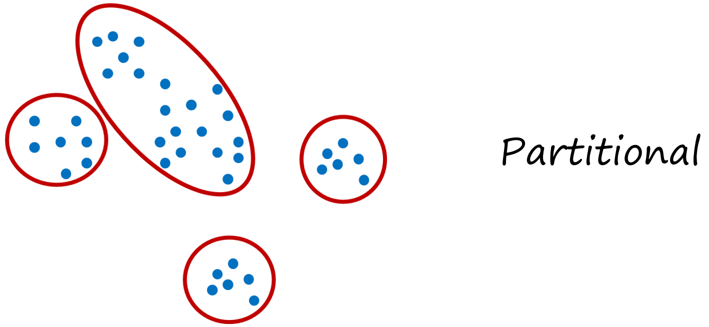
- In clustering algorithms, we usually optimize the following for a given number of clusters.
  - Tightness, spread, cohesion of clusters
  - Separability of clusters, distance between the centers

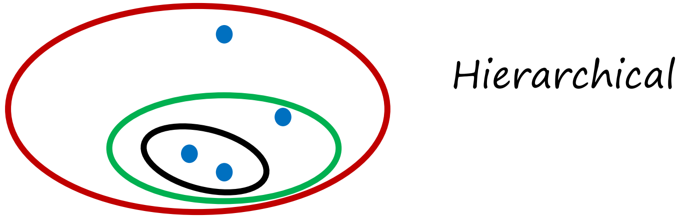
- Ideally, we require clustering algorithms to be
  - scalable (in terms of both time and space)
  - able to deal with different data types and noise/outliers
  - insensitive to order of input records
  - interpretable and usable

Clustering

Clustering Algorithms:

- Partitional Clustering
  - Divides data points into non-overlapping subsets (clusters) such that each data point is in exactly one subset.
  - E.g. K-means clustering
- Hierarchical Clustering
  - Constructs a set of nested clusters by carrying out hierarchical division of the data points.
  - E.g., Agglomerative clustering, Divisive Clustering
- Model Based Clustering
  - Assumes that the data was generated by a model and try to fit the data to model that defines clusters of the data
- Density Based Clustering
  - Assumes that a cluster in the space is a region of high point density separated from other clusters by regions of low point density.





Outline

- Introduction to Unsupervised Learning, Clustering
- Clustering Overview
- Partitional Clustering
    - K-means Clustering
- Hierarchical Clustering
    - Agglomerative Clustering

Partitional Clustering

Overview:

- We want to cluster a set of n data points $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, $\mathbf{x} \in \mathbf{R}^d$.
- Partitional clustering constructs a partition of n data points into a set of K clusters such that partitioning criterion is optimized.
- Each data point is a part of only one cluster.
- Finding globally optimal clustering would require exhaustive search over all the points and partitions.
- Heuristic algorithms are more effective. For example K-means or K-medoids.
- K-means: Each cluster is characterized by the center of the cluster.
- K-medoids: Each cluster is characterized by one data point, medoid, in the cluster for which the average dissimilarity to all other data points objects in the cluster is minimal.

K-means Algorithms

Overview and Notation:

- The K-means algorithm partitions the data D into K clusters.
- We assume that K is given by the user.
- Each cluster is a group of points.
- Let the clusters be denoted by $c_1, c_2, \ldots, c_K$
- Each cluster is characterized by its center, referred to as cluster centroid, aka mean or the center of gravity.
- Mathematically, the cluster centroid, denoted by μ, is defined as a mean of the points in the cluster, that is

$$\mu(c_i) = \frac{1}{|c_i|} \sum_{\mathbf{x}_j \in c_i} \mathbf{x}_j$$

K-means Algorithms

Algorithm:

In K-means algorithm, we carry out the following steps:

- **Input:** K and Data D

- **Randomly choose K cluster centers (centroids)**
- **Repeat until convergence:**
  - Each data point is assigned membership of the cluster of closest centroid
  - Compute the centroids again for each cluster using the current cluster memberships

Computations:

- Randomly choose $\mu(c_i)$ for $i = 1, 2, \ldots, K$

Repeat until convergence:

- For each $\mathbf{x}_j$, assign the cluster c_i such that $\text{dist}(\mathbf{x}_j, \mu(c_i))$ is minimal.
- Recompute the centroids as follows

$$\mu(c_i) = \frac{1}{|c_i|} \sum_{\mathbf{x}_j \in c_i} \mathbf{x}_j$$

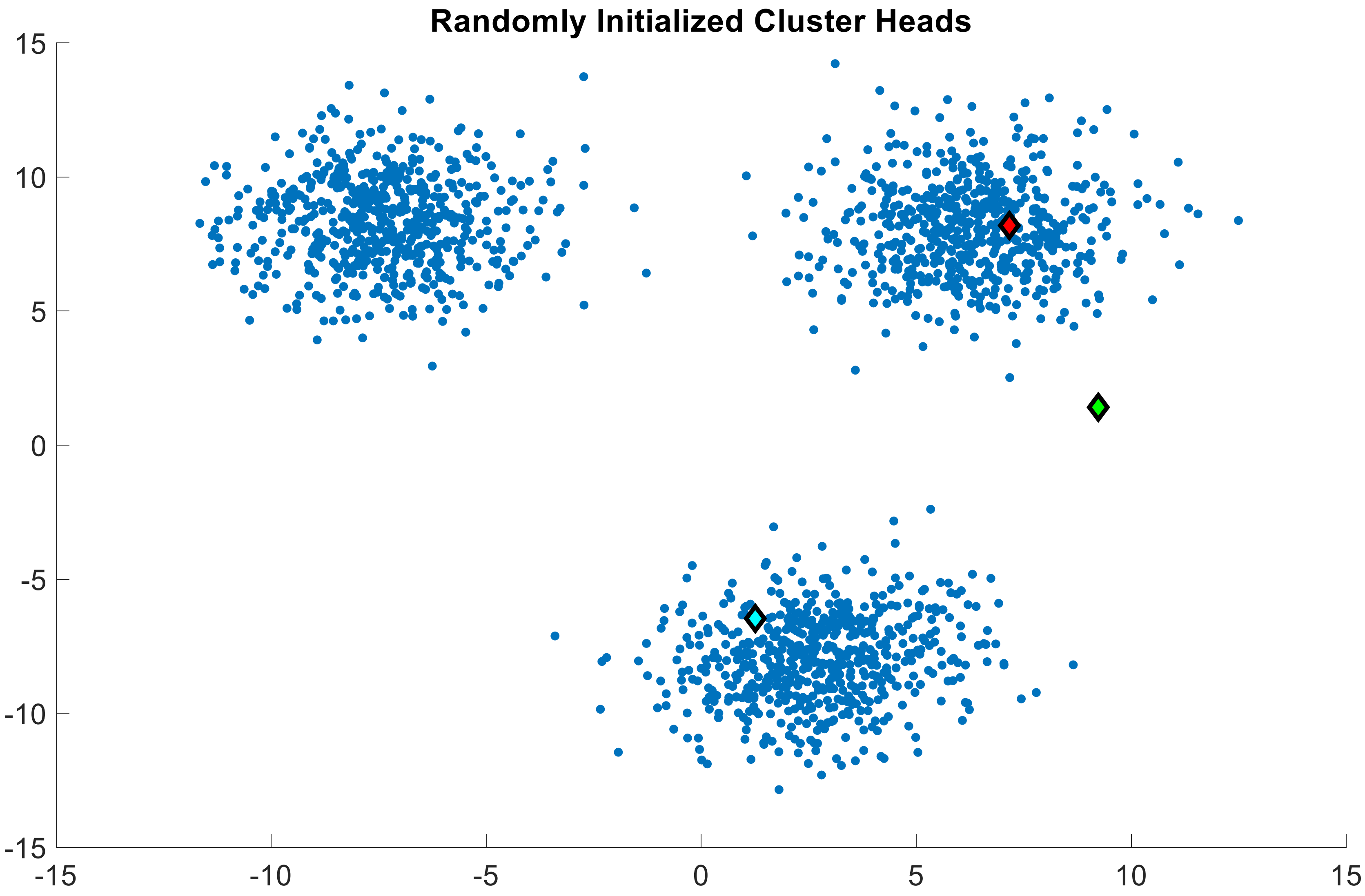
Complexity:

$\mathcal{O}(K\, n\, d)$

$\mathcal{O}(n\, d)$

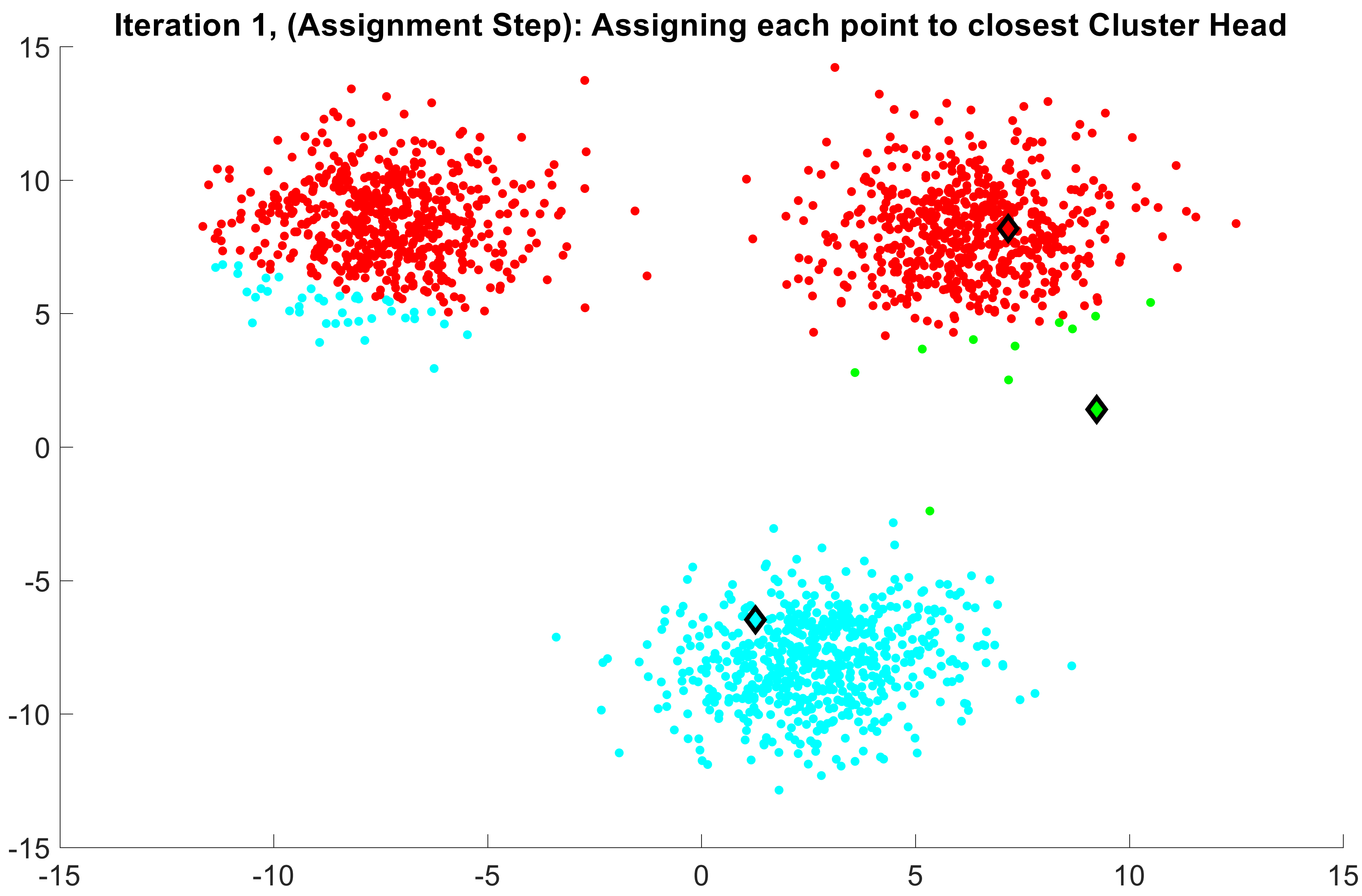
K-means Algorithms

Illustration:



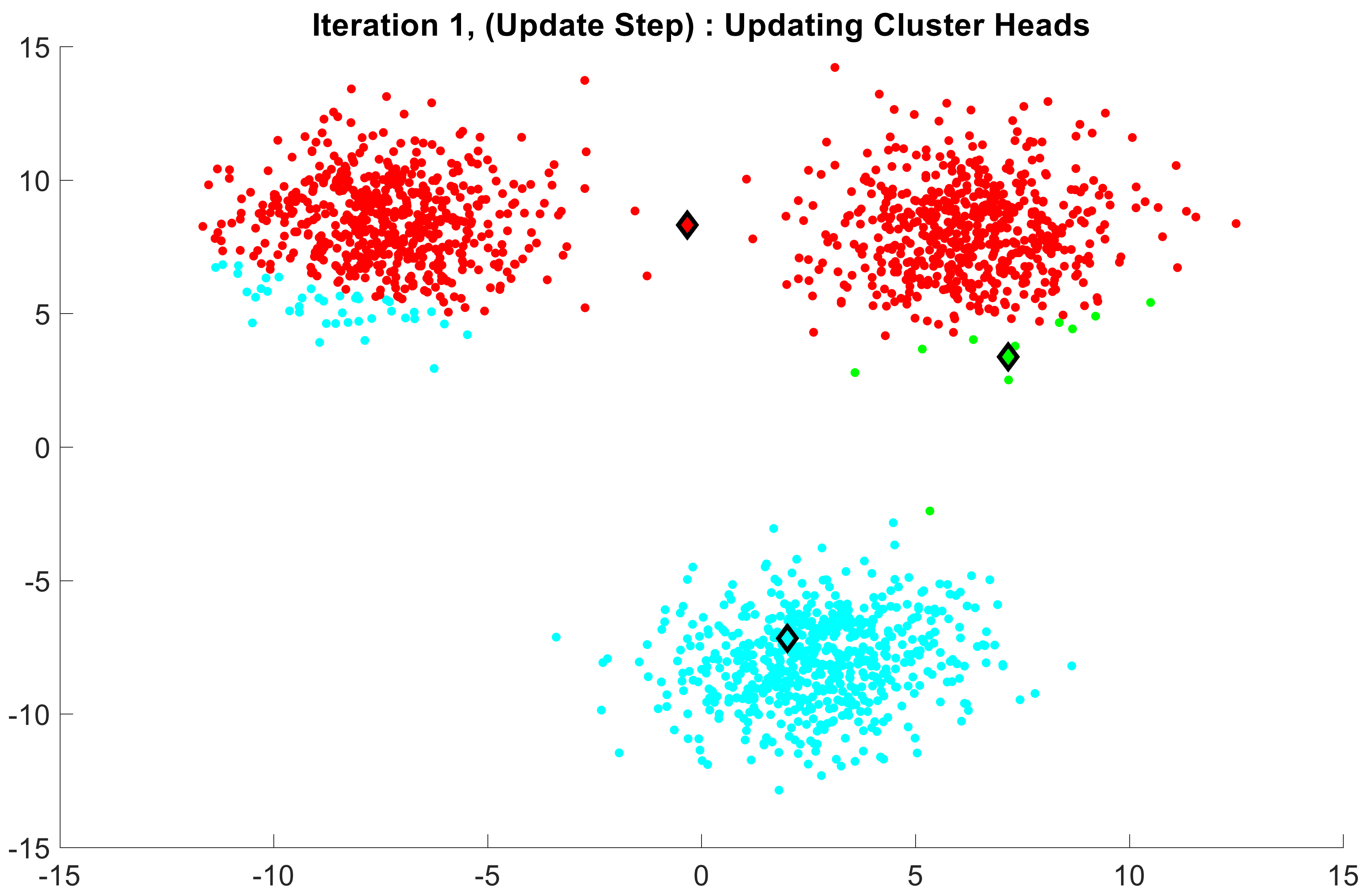
K-means Algorithms

Illustration:



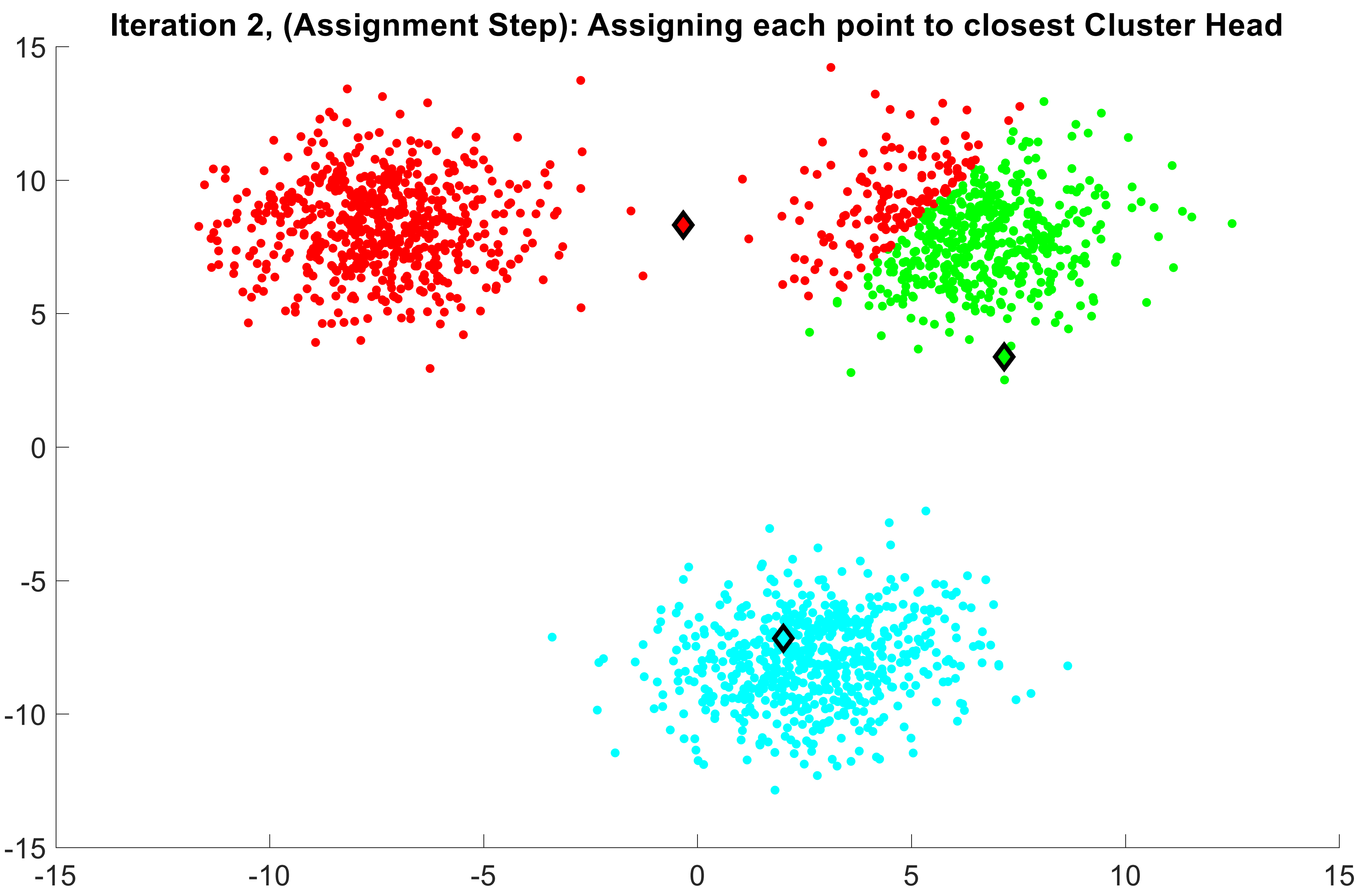
K-means Algorithms

Illustration:



K-means Algorithms

Illustration:

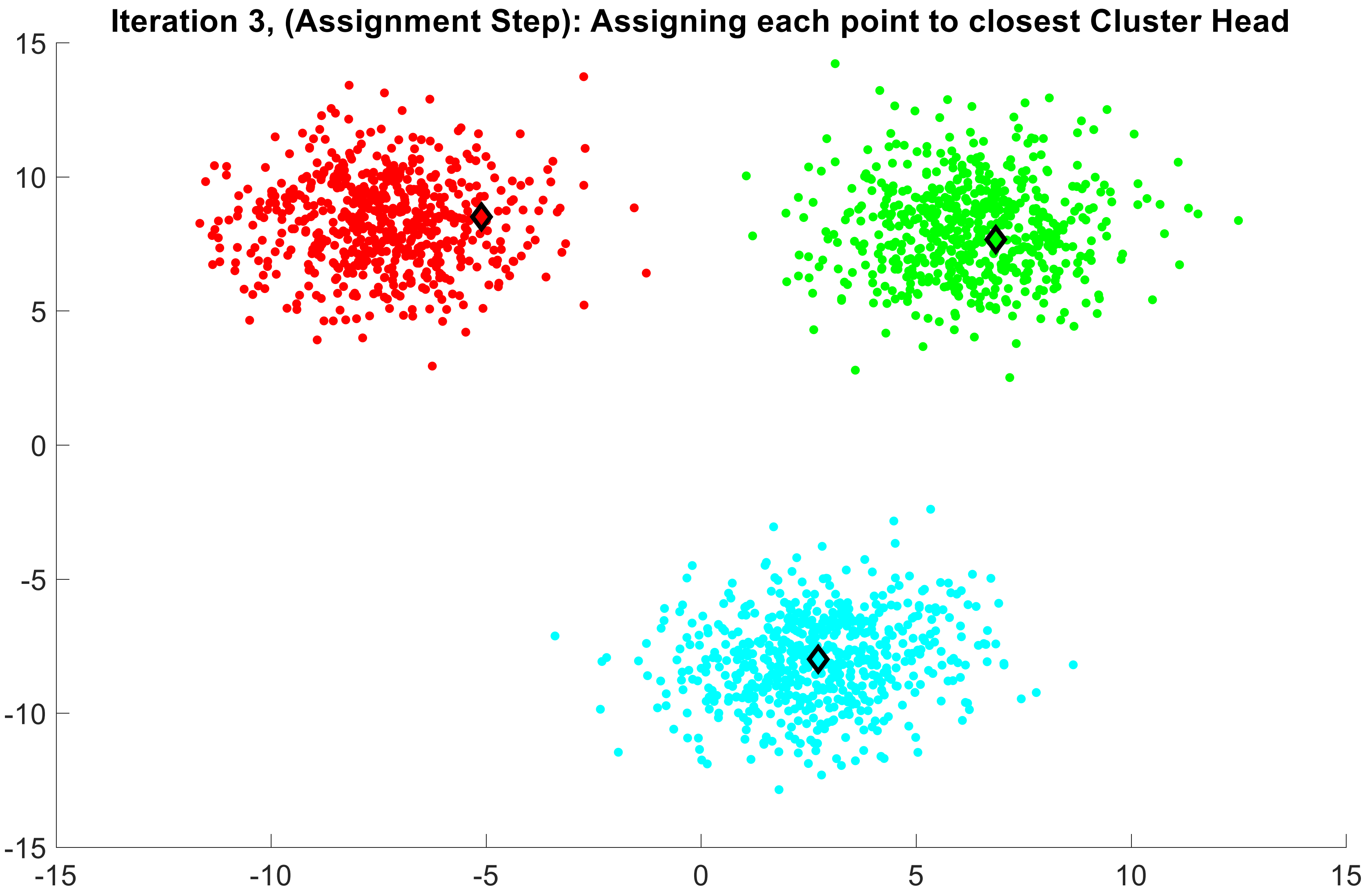


K-means Algorithms

Illustration:

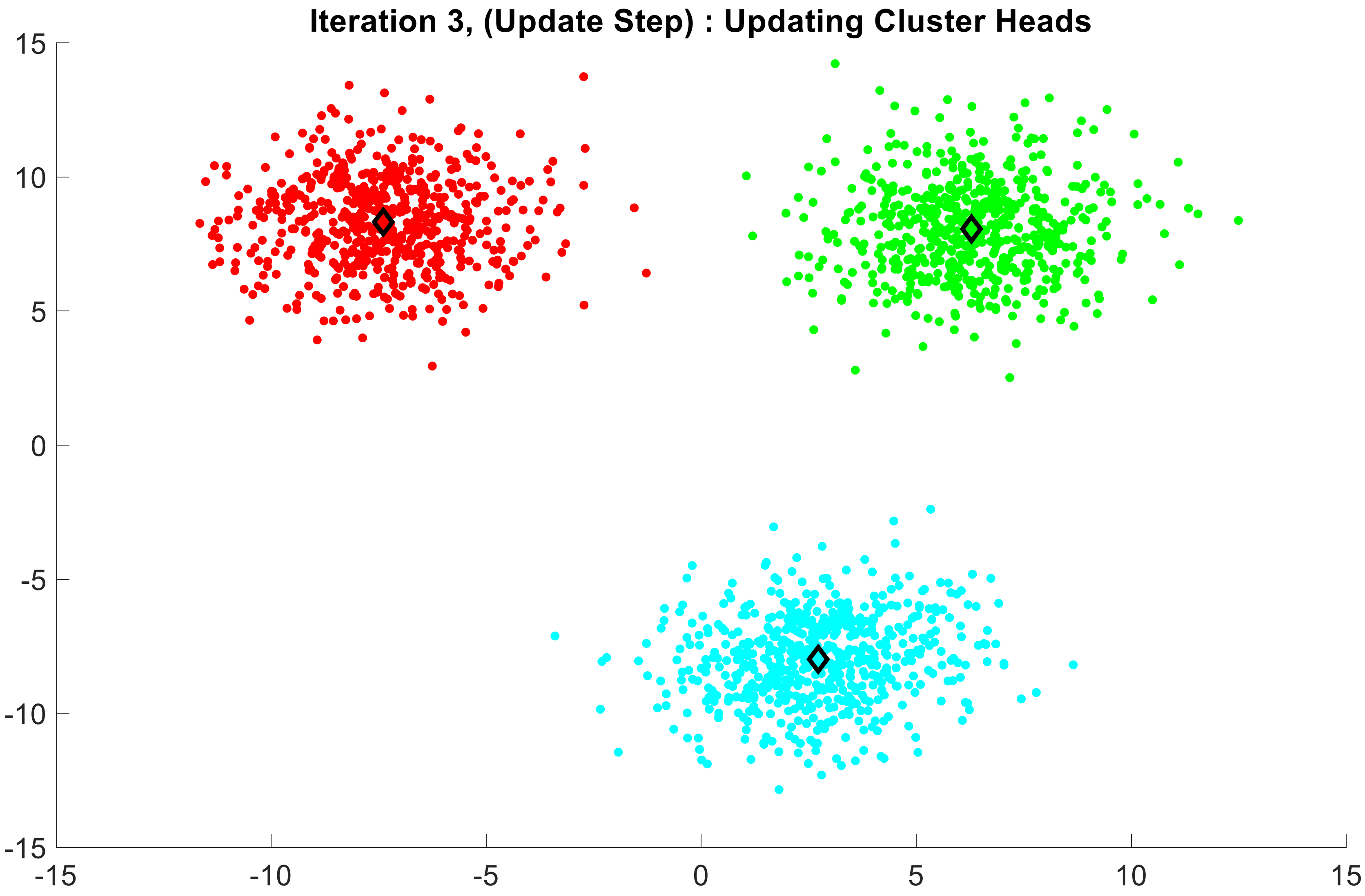
K-means Algorithms

Illustration:



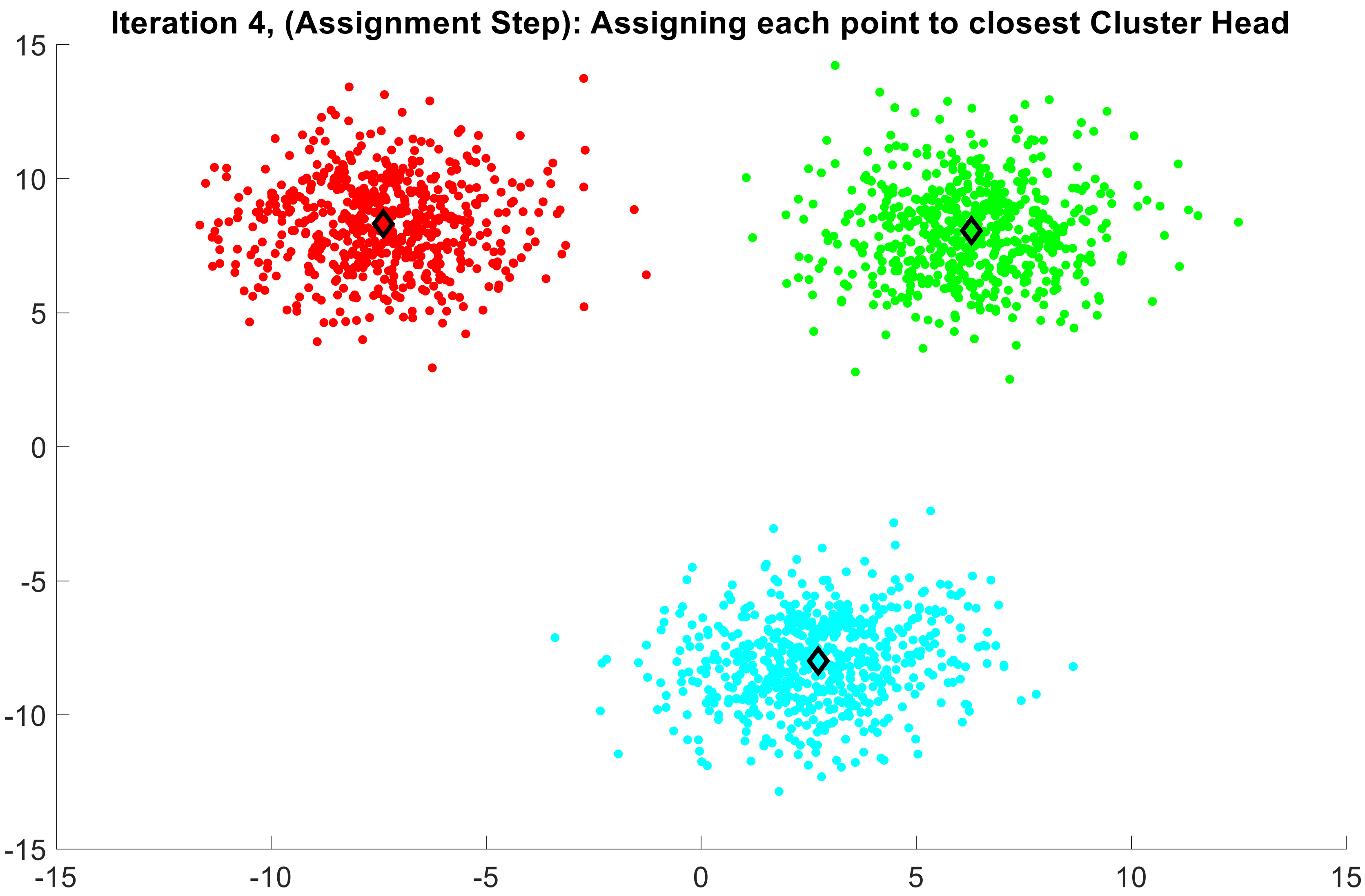
K-means Algorithms

Illustration:



K-means Algorithms

Illustration:



K-means Algorithms

Stopping and Convergence Criterion:

Multiple convergence criteria:

- Convergence of the re-assignment of data points to different clusters: re-assignment is stopped or minimized.

- Convergence of the location of centroids: no change or minimum change.

- Convergence of the sum of squared error (SSE) defined as

$$\text{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in c_i} \text{dist}\big(\mathbf{x}_j, \mu(c_i)\big)^2$$

error is the distance of each point to the nearest cluster centroid.

Interpretation: *clusters are no more changing.*

Remark: *Algorithm may converge at a local optimum.*

K-means Algorithms

Evaluation of K-means clustering:

- We can use the sum of squared error (SSE) to evaluate the clustering performance.

$$\text{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in c_i} \text{dist}\big(\mathbf{x}_j, \mu(c_i)\big)^2$$

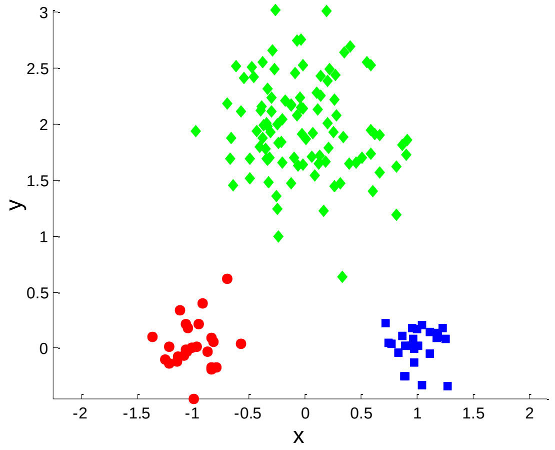
- Better clustering will have smallest SSE.

- SSE can be reduced by increasing the number of clusters K.

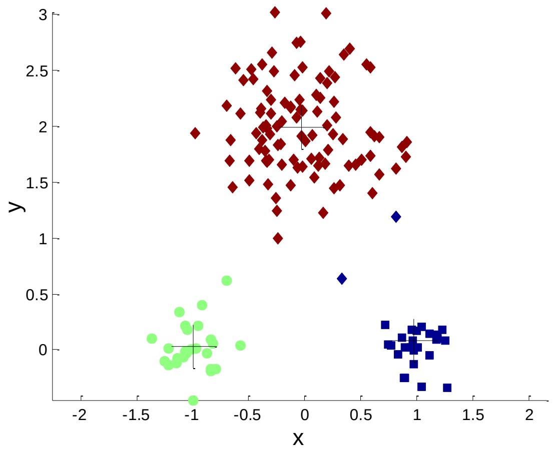
K-means Algorithms

Choice of Initial Centroids:

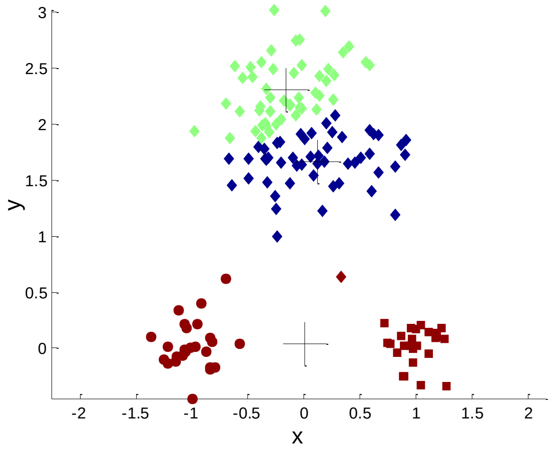
- Choice of initial centroids can significantly impact the performance of the clustering. We can have slow convergence rate or convergence to sub-optimal clusterings.



Original Points



Optimal Clustering



Sub-optimal Clustering

Possible solutions:

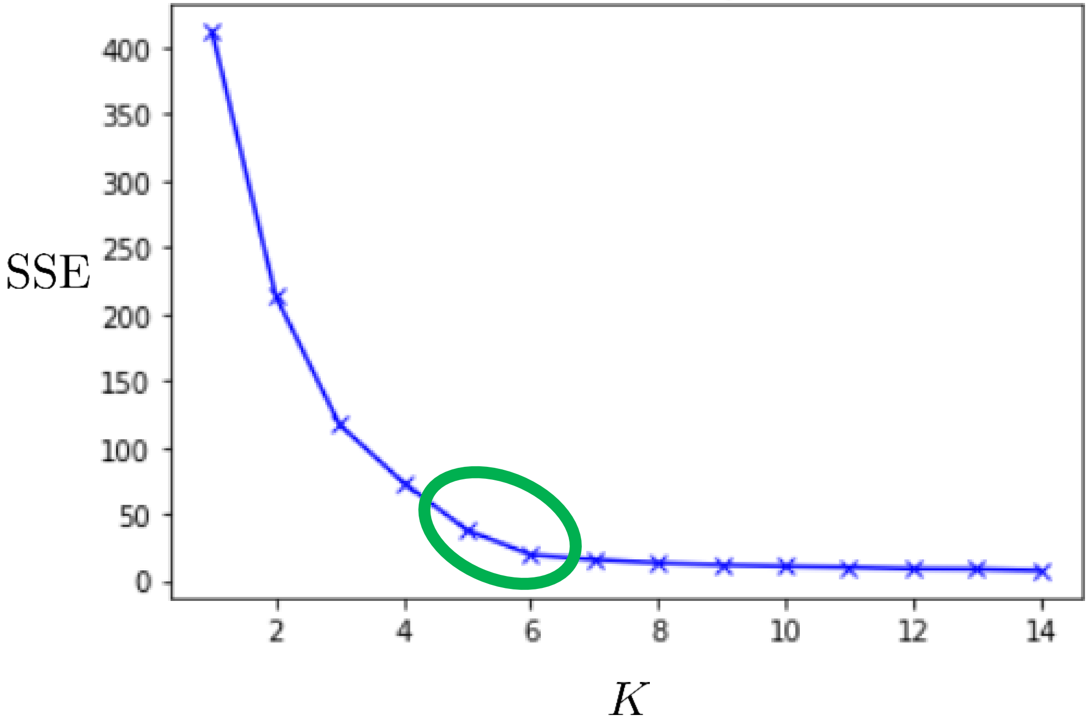
- Carry out multiple runs of centroids and evaluate the performance.
    - It may help, but with very low probability
- Start with more than K centroids and select K most widely spread.
- Determine initial centroids using hierarchical clustering.
- Post-processing
    - Drop small clusters
    - Split clusters with large SSE
    - Merge clustersw ith low SSE

LUMS
A Not-for-Profit University

K-means Algorithms

Number of Clusters:

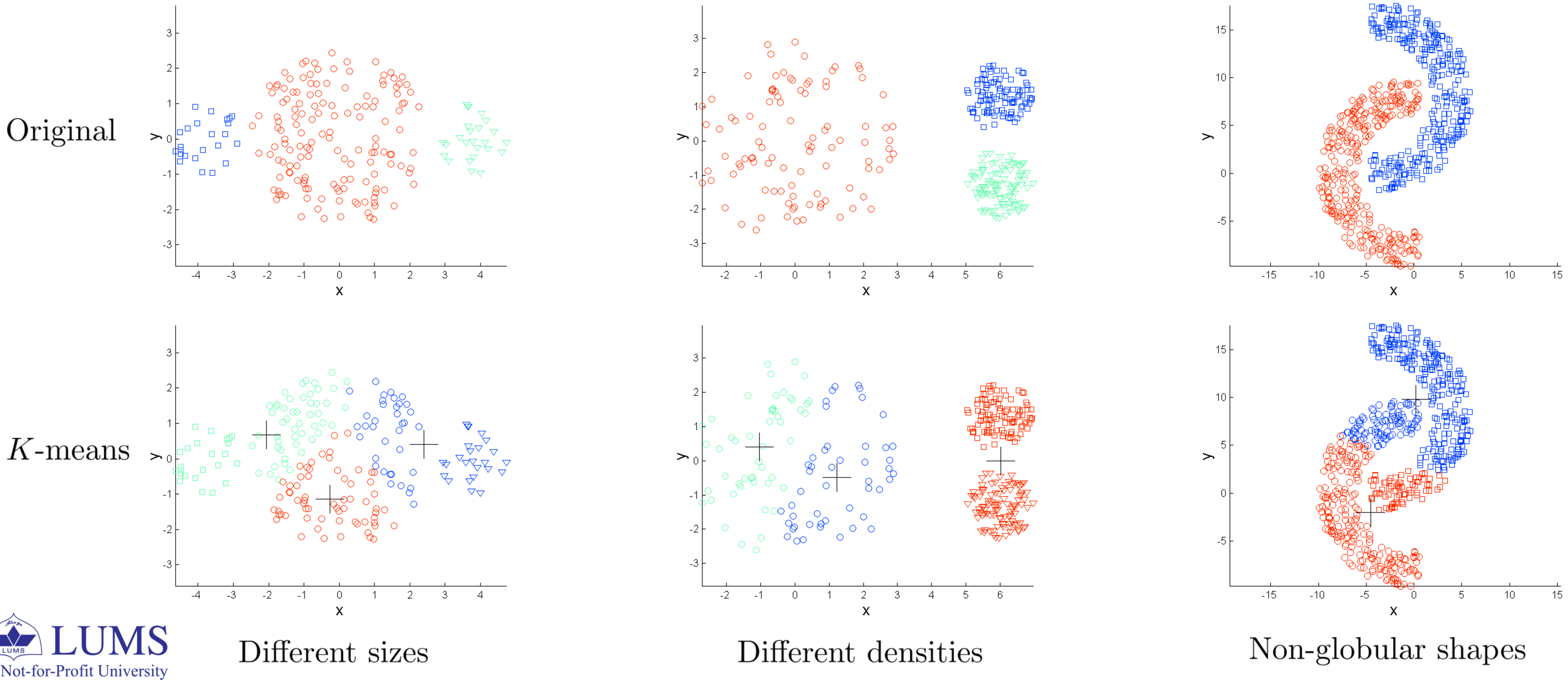
- In K-means clustering, we assumed that we have information about number of clusters.
- How do we determine K when we do not have information about the number of clusters that is often the case in practice?
- Answer: we do not know!
- But we can use heuristics.
- One solution, known as elbow method, can help us in approximating a good choice of K.
- Evaluate clustering performance for different values of K.
- If the plot is like an arm (you may not get such clean plot in practice), then the optimal K is an elbow on the arm.
- We can also plot the 'Silhouette Coefficient' vs K and use this to judge the value of K.



K-means Algorithms

Limitations/Weaknesses:

- Sensitive to initial values of centroids.
- K-means is sensitive to difference in sizes, densities and shapes of clusters.



K-means Algorithms

Summary:

- Despite these limitations, K-means is the most popular and fundamental unsupervised clustering algorithm;
  - Simple: two-step iterative algorithm; easy to understand and to implement.
  - Computationally efficient: O(K n d) is the time complexity.
- It assumes that the number of clusters is known.
- Most of the convergence takes place in the first few iterations.

- Performance of the clustering is often hard to evaluate, that is true for every clustering algorithm.

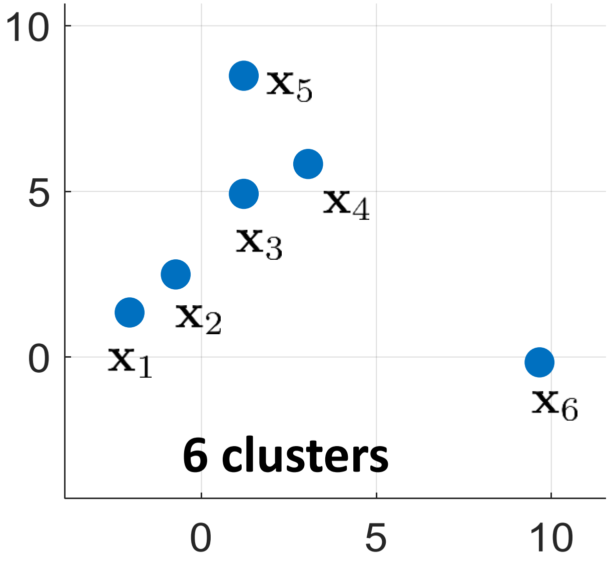
- It is sensitive to initial values of centroids, outliers and difference in sizes, densities and shapes of clusters.

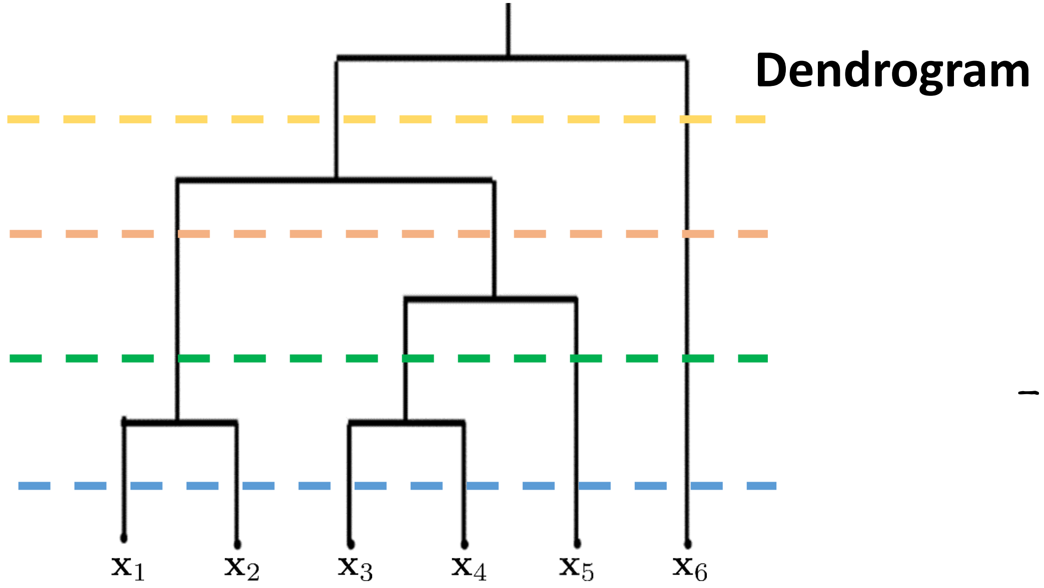
Outline

- Introduction to Unsupervised Learning, Clustering
- Clustering Overview
- Partitional Clustering
  - K-means Clustering
- Hierarchical Clustering
  - Agglomerative Clustering

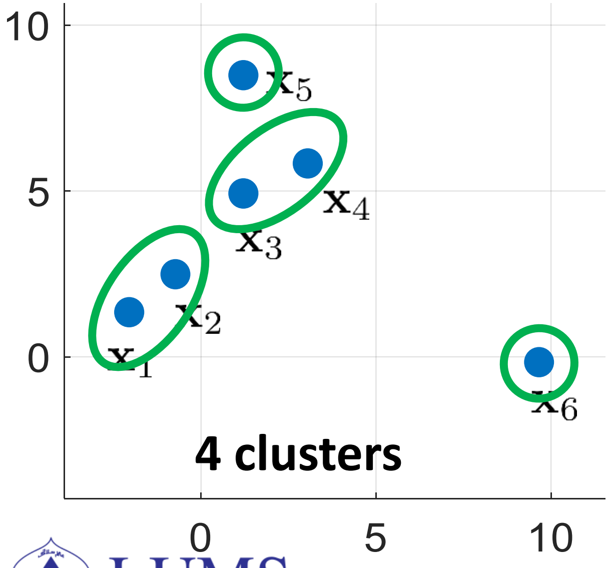
Hierarchical Clustering

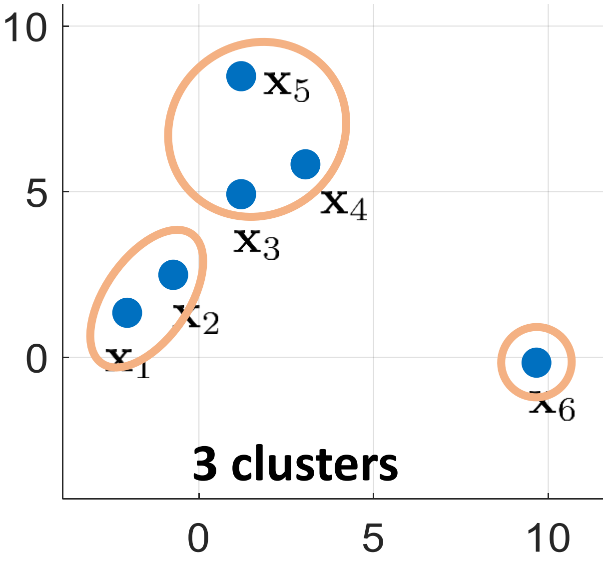
Overview – Illustration:

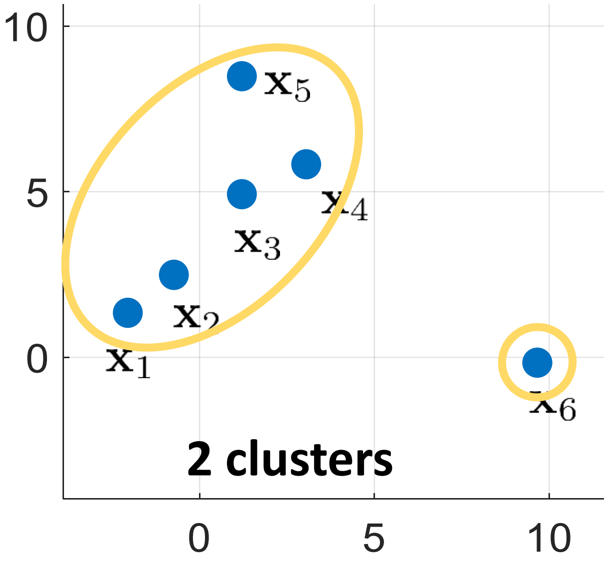




- *We take a union of clusters at level i+1 to obtain a parent cluster at level i.*







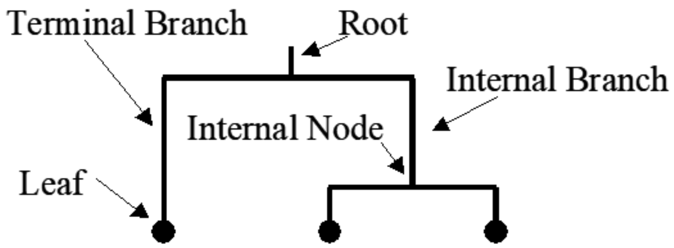
Hierarchical Clustering

Overview:

- In hierarchical clustering, we carry out a hierarchical decomposition of the data points using some criterion.
- Use distance or similarity metric to carry out hierarchical decomposition. **We do not need to define the number of clusters as an input.**
- A nested sequence of clusters is created in this decomposition process.
- This nested sequence of clusters, a tree, is also called Dendrogram.

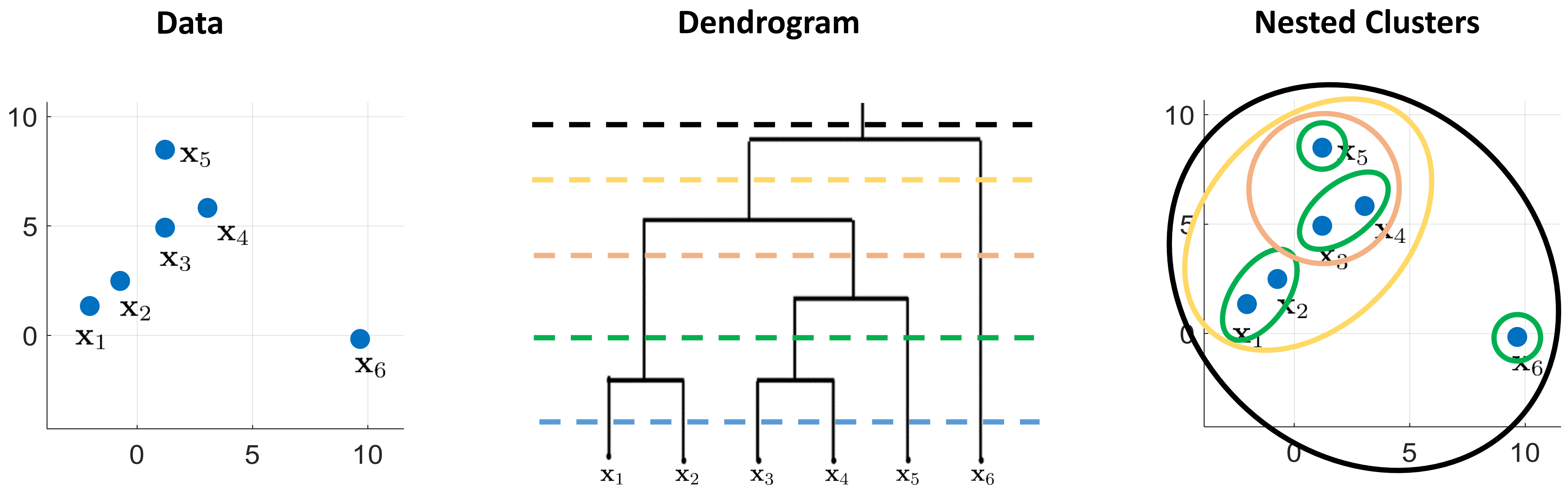
Dendrogram:

- A tree data structure, that records the sequences of splits or merges, used for the visualization of hierarchical clustering techniques.
- We represent the similarity between two data-points in the dendrogram as the height of the lowest internal node they share.
- Root corresponds to one cluster and leaf represents individual clusters.
- Each level of the tree shows clusters for that level.



Hierarchical Clustering

Overview – Illustration:



- *Dendrogram is the representation of nested clusters.*

- *We can cut the dendrogram at a **desired level to carry out clustering**; the connected data-points below the desired level form a cluster.*

Hierarchical Clustering

Overview:

- **Agglomerative:**
  - *Start with considering each data point as one cluster*
  - *Merge the clusters iteratively*
  - *Keep on merging until all clusters are fused to form one cluster*
  - *Also termed as 'Bottom-Up'*

- **Divisive:**
  - *Starting with considering all data points as a single cluster*
  - *Divide (split) the clusters successively*
  - *Also termed as 'Top-Down'*

- In both approaches, we usually similarity (distance metric) one cluster at a time.

Agglomerative Clustering

Algorithm:

In agglomerative algorithm, we carry out the following steps:

- **Input:** $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x} \in \mathbf{R}^d$.
- **Algorithm:**
  - Make each data point as a cluster.
  - Compute all pairwise distances (usually referred to as proximity matrix)

  Repeat:
    - merge the two clusters that are nearest to each other to form a cluster c
    - compute the distance of c from all other clusters.

  Until only one cluster is left

Complexity:

$\mathcal{O}(n^2\, d)$ Hierarchical clustering techniques do not scale well with the size of the data.

Agglomerative Clustering

Agglomerative Clustering:

- Here, we are merging the two clusters that are nearest to each other.
- A group of points represents a cluster.
- We have studied a distance metric that computes the distance between points.

Question: How do we compute the distance between a point and a cluster or the distance between two clusters?

Answer: We can define the closest pair of clusters in multiple ways, and this results in different versions of hierarchical clustering.

- **Single linkage:** Distance of two closest data points in the different clusters (nearest neighbor)
- **Complete linkage:** Distance of the furthest points in the different clusters (furthest neighbor)
- **Group average linkage:** Average distance between all pairs of points in the two different clusters.
- **Centroid linkage:** Distance between centroids
- **Wards linkage:** Merge the clusters such that the variance of the merged clusters is minimized.

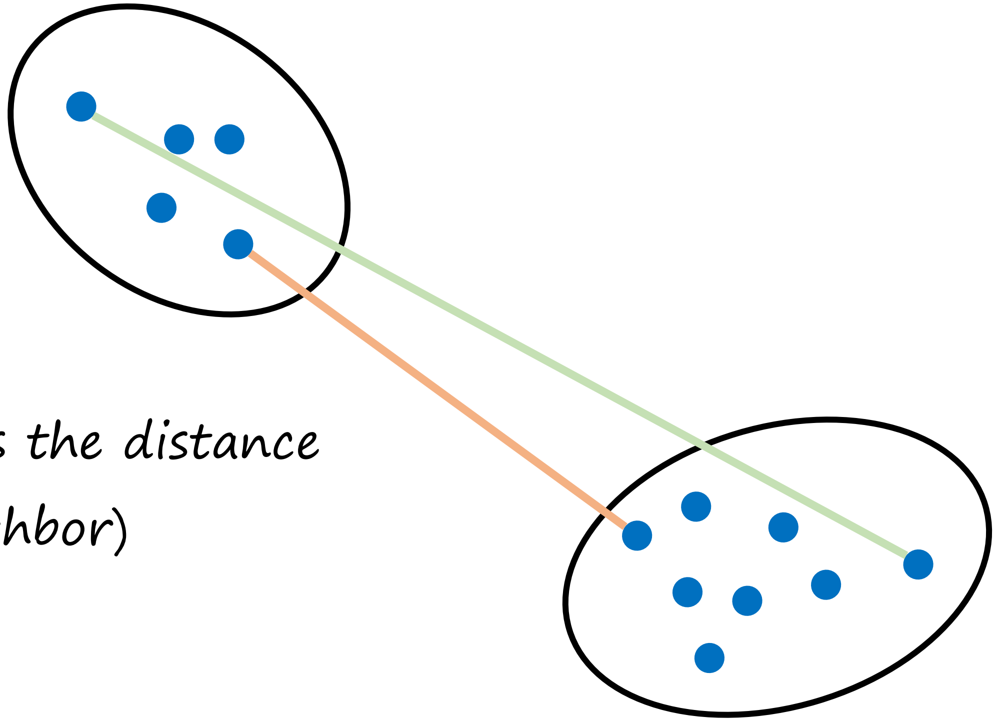
Agglomerative Clustering

Agglomerative Single Linkage:

- **Single linkage:** Distance between the two clusters is the distance between the closest data points (nearest neighbor).
  - Distance between clusters c_ℓ and c_m is given by

$$\min_{\mathbf{x}_i \in c_\ell,\, \mathbf{x}_j \in c_m} \operatorname{dist}\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

  - Results in (long and thin) clusters.
  - Sensitive to noise and/or outliers



- **Complete linkage:** Distance between the two clusters is the distance between the furthest closest data points (furthest neighbor)
  - Distance between clusters c_ℓ and c_m is given by

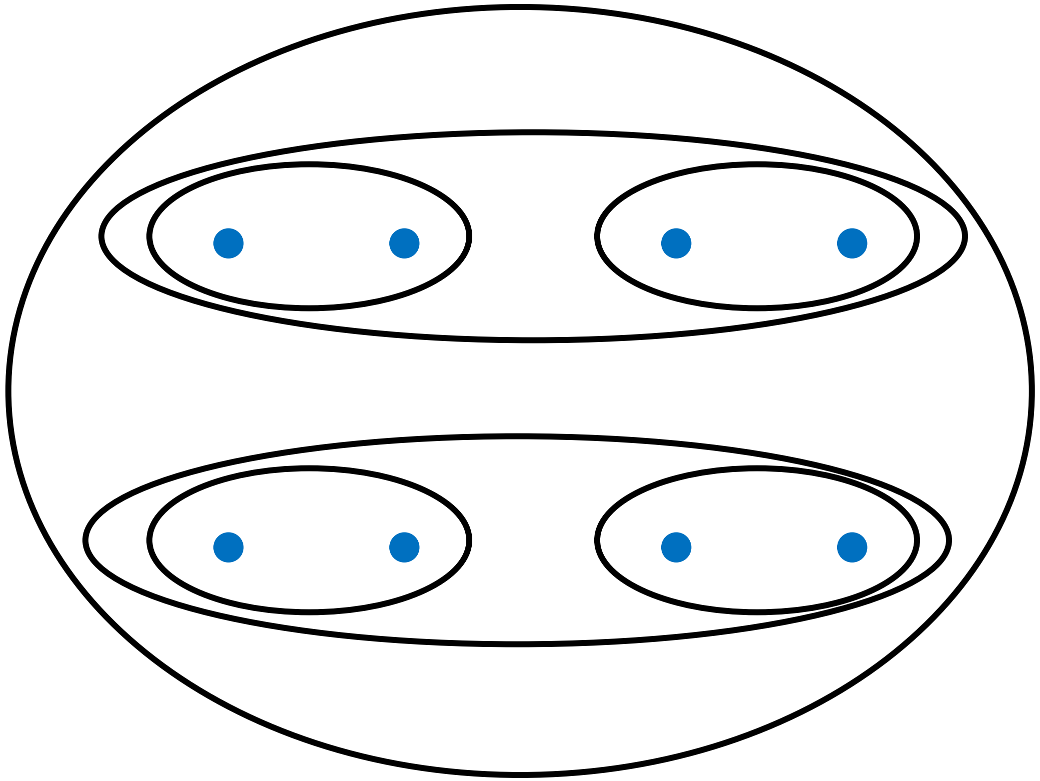
$$\max_{\mathbf{x}_i \in c_\ell,\, \mathbf{x}_j \in c_m} \operatorname{dist}\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

  - Results in more compact spherical clusters (biased towards globular, blob clusters).
  - Less sensitive to noise and/or outliers.

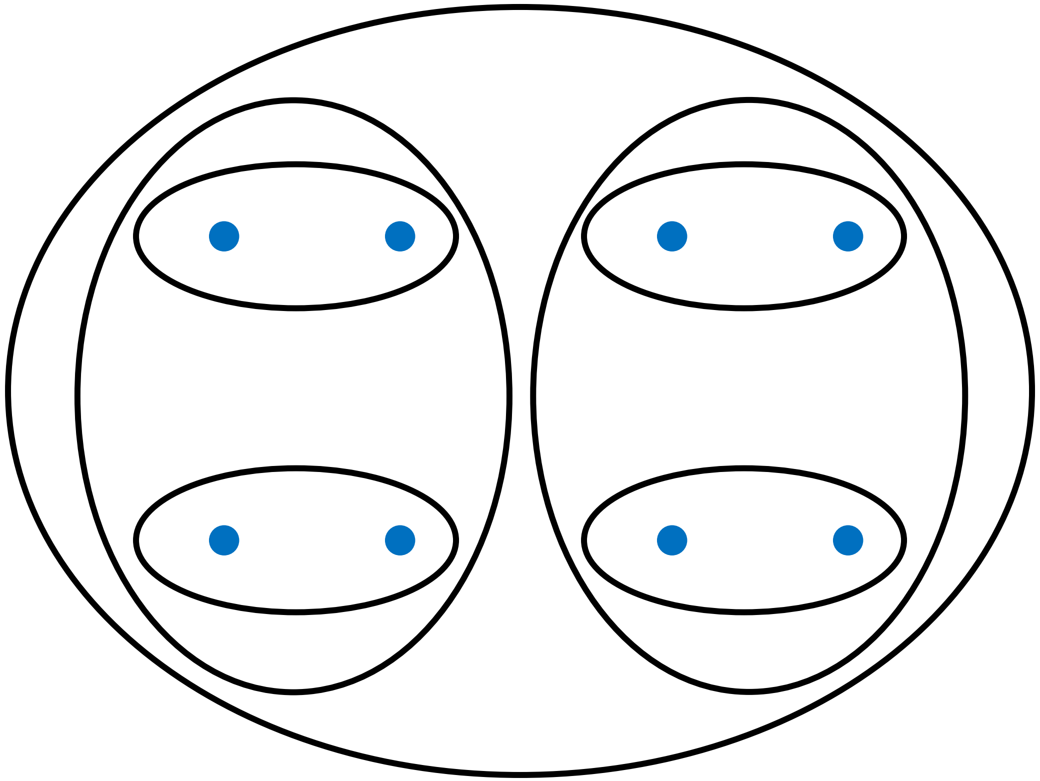
Agglomerative Clustering

Agglomerative Single Linkage:

- Single linkage vs Complete linkage:



Single linkage



Complete linkage

LUMS
A Not-for-Profit University

Hierarchical Clustering

Summary:

- We obtain a set of nested clusters arranged as a tree, aka dendrogram.
- We do not need to specify the number of clusters in advance.
- Agglomerative is bottom-up and divisive is top-down.
- We have different metrics to quantify the distance between the clusters; the clusters are different for each metric.
- Hierarchal clustering is often used for analyzing text data or social network data.
- Unlike K-means, hierarchical clustering is does not scale well due to significant computational cost $O(n^2)$.
- Like any heuristic search algorithms, local optima are a problem.
- Interpretation of results is (very) subjective.

Clustering

References:

- CB: 9.1
- KM: 11.4.2.5
- Introduction to Information Retrieval (https://nlp.stanford.edu/IR-book/) (Ch: 16, 17)
- Data clustering: A review by Jain, Anil et. al. ACM Computing Surveys 31 (3): 264-323, 1999