What is Cluster Analysis?

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation, …*)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters
- The quality of a clustering method depends on
  - the similarity measure used by the method
  - its implementation, and
  - Its ability to discover some or all of the hidden patterns

Requirements and Challenges

- Scalability – algo must work well with data of all sizes.
  - Clustering all the data instead of only on samples
  - Clustering on samples may lead to biased results
- Ability to deal with different types of attributes
  - Numerical, binary, categorical, ordinal, and mixture of these
- Constraint-based clustering
  - User may give inputs on constraints
  - Use domain knowledge to determine input parameters
- Interpretability and usability – results must be easily interpretable, understandable and usable.

Requirements and Challenges

- Others
  - Discovery of clusters with arbitrary shape – cluster of any shape
  - Ability to deal with noisy data – some algorithm are sensitive may lead to poor quality clusters.
  - Incremental clustering and insensitivity to input order – ability to incorporate newly inserted data into the existing clustering structures. Some algorithm may result in different clustering based on the order of presentation of input objects.
  - High dimensionality – some algorithm are good at handling low dimensionality data (2 to 3 dimensions).
  - Minimal requirement of domain knowledge to determine input parameters – user should input certain parameters (eg. No. of clusters). Difficult to determine these parameters in high dimensional data. May affect the quality of the cluster.

Types of data in cluster analysis

- Refer text book or your notebook

Major Clustering Approaches (I)

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion
  - Typical methods: k-means, k-medoids, CLARANS (clustering large applications based on randomized search)
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, CHAMELEON
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSCAN, OPTICS, DenClue
- Grid-based approach:
  - Quantize object space into finite number of cells called grid.
  - based on a multiple-level granularity structure
  - Typical methods: STING, WaveCluster, CLIQUE

Major Clustering Approaches (II)

- Model-based:
  - A model is hypothesized for each of the clusters and tries to find the best fit of the data to the model.
  - Typical methods: EM, SOM, COBWEB
- Frequent pattern-based:
  - Based on the analysis of frequent patterns
  - Typical methods: p-Cluster
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering
- Link-based clustering:
  - Objects are often linked together in various ways
  - Massive links can be used to cluster objects: SimRank, LinkClus

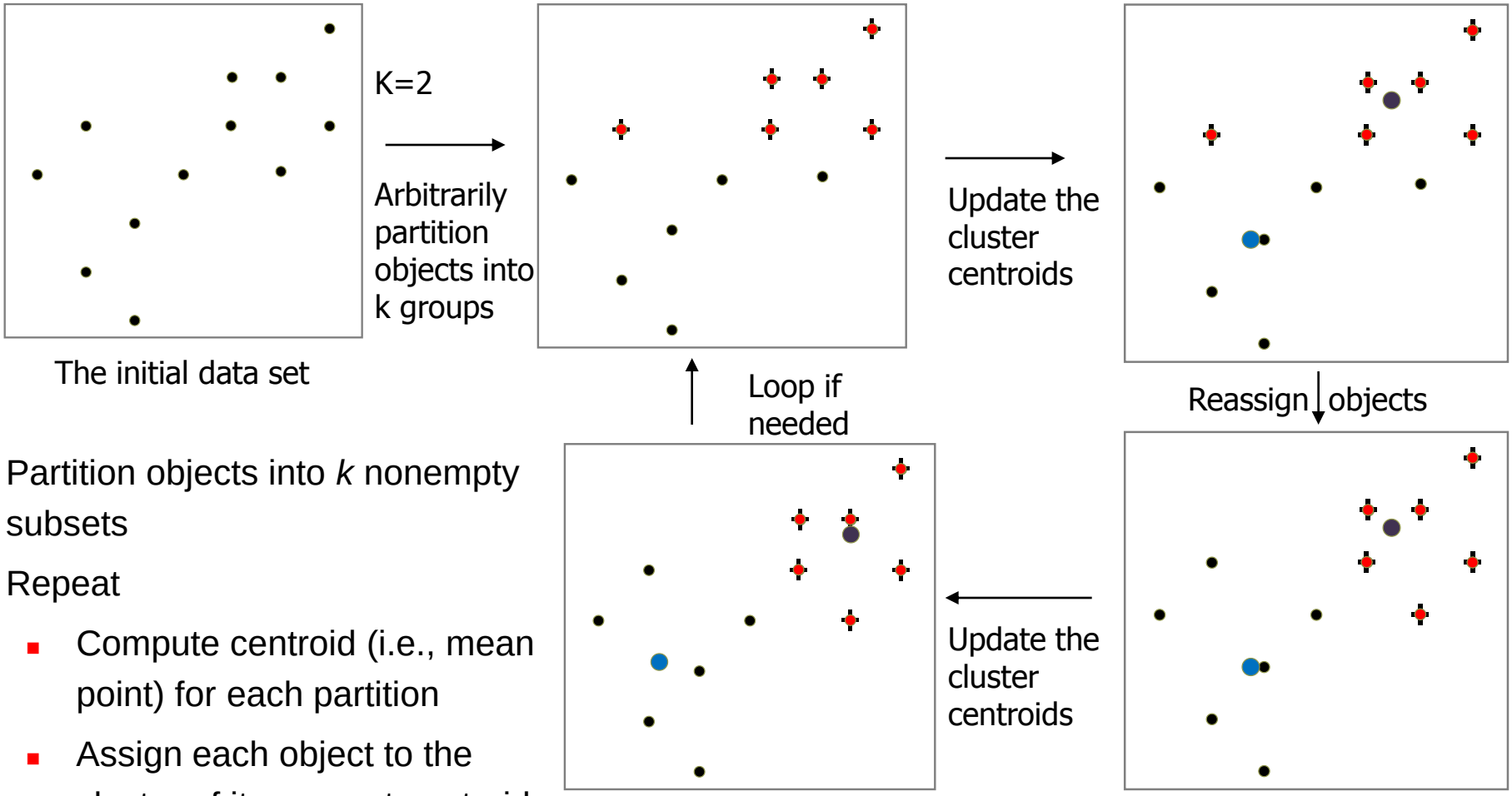
Partitioning Algorithms: Basic Concept

- Partitioning method: Partitioning a database ***D*** of ***n*** objects (records) into a set of ***k*** clusters (partitions, where k<=n).
  - *k-means* - Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids) - Each cluster is represented by one of the objects in the cluster

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:
  - Partition objects into *k* nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., *mean point*, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when the assignment does not change

An Example of *K-Means* Clustering



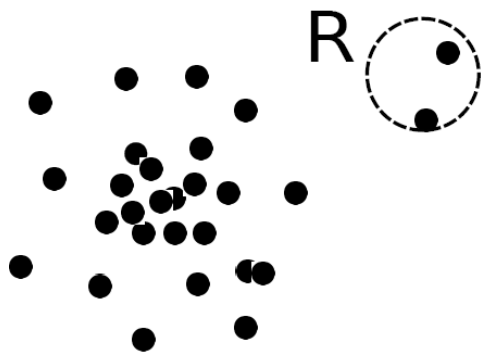
- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

K-Means example problem

- Refer your Note book

What Are Outliers?

- **Outlier**: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism**
  - Ex.: Unusual credit card purchase, sports: Michael Jordon, Wayne Gretzky, ...
- Outliers are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection
- Outliers are interesting: It violates the mechanism that generates the normal data



Applications

- Fraud Detection (Credit card, telecommunications, criminal activity in e-Commerce)
- Customized Marketing (high/low income buying habits)
- Medical Treatments (unusual responses to various drugs)
- Analysis of performance statistics (professional athletes)
- Weather Prediction
- Financial Applications (loan approval, stock tracking)

Challenges of Outlier Detection

- Modeling normal objects and outliers properly
  - Hard to enumerate all possible normal behaviors in an application
  - The border between normal and outlier objects is often a gray area
- Application-specific outlier detection
  - Choice of distance measure among objects and the model of relationship among objects are often application-dependent
  - E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations
- Handling noise in outlier detection
  - Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection
- Understandability
  - Understand why these are outliers: Justification of the detection
  - Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism

Proximity-Based Approaches: Distance-Based vs. Density-Based Outlier Detection

- Intuition: Objects that are far away from the others are outliers
- Two types of proximity-based outlier detection methods
  - Distance-based outlier detection: An object o is an outlier if its neighborhood does not have enough other points
  - Density-based outlier detection: An object o is an outlier if its density is relatively much lower than that of its neighbors

Distance-Based Outlier Detection

- For each object o, examine the # of other objects in the *r*-neighborhood of o, where *r* is a user-specified **distance threshold**
- An object o is an outlier if most (taking π as a **fraction threshold**) of the objects in D are far away from o, i.e., not in the r-neighborhood of o.