Clustering

B.Sc. 6th Semester (Data Mining, Paper Code: DSE-4)

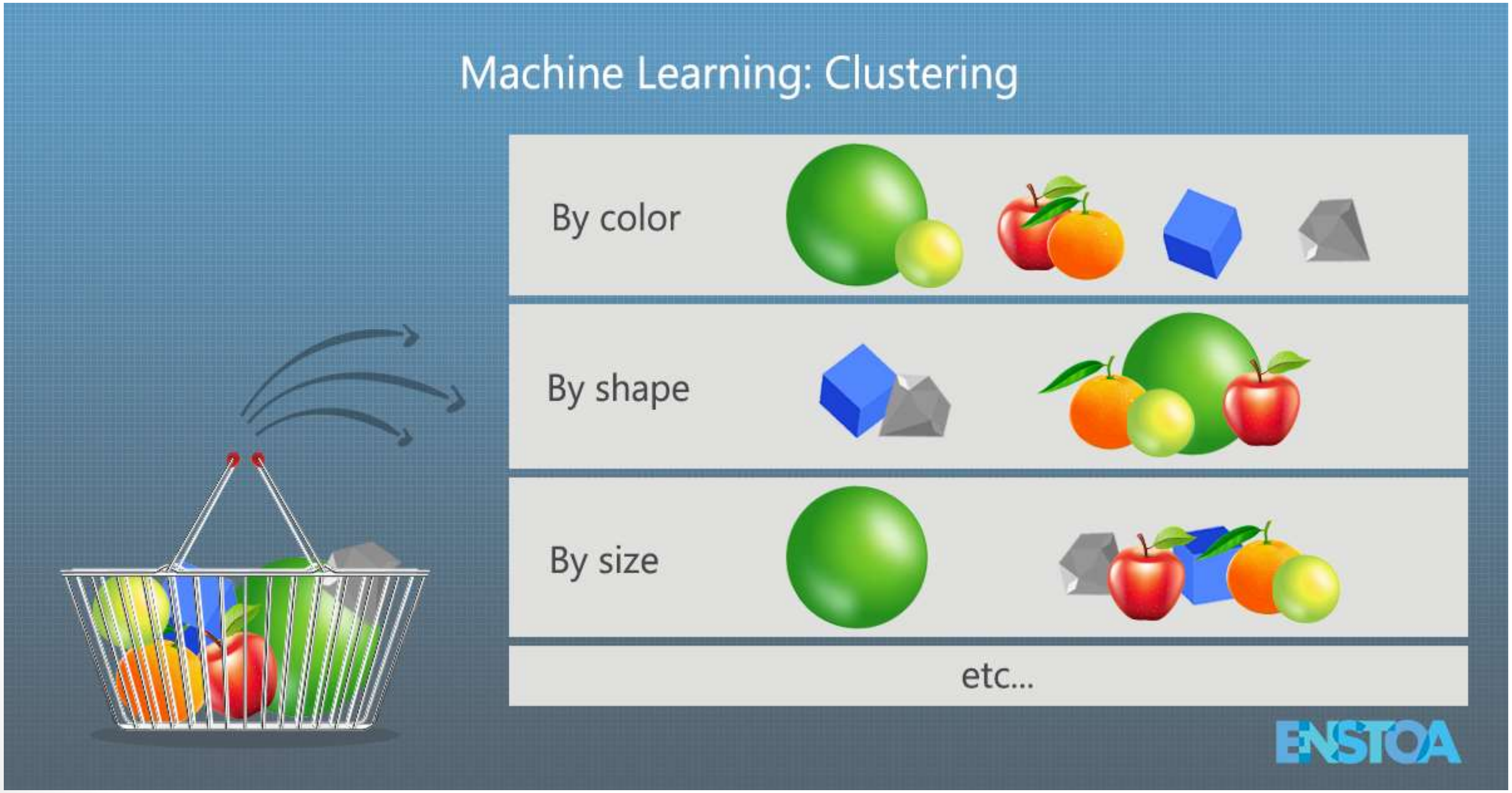
Paulami Basu Ray
Assistant Professor
Department of Computer Science & Applications
Prabhat Kumar College, Contai

An introduction to Clustering

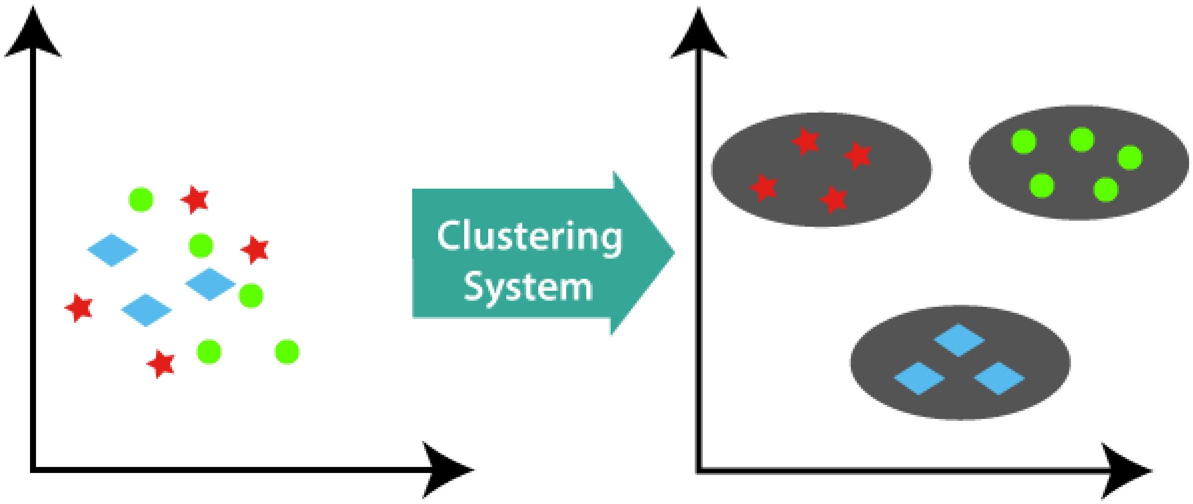
- It is basically a type of *unsupervised learning method*. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structures and groupings inherent in a set of examples.
- **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.
- Usually, points are in a high---dimensional space, and similarity is defined using a distance measure.

✓ Euclidean, Cosine. Jaccard distances etc can be used.

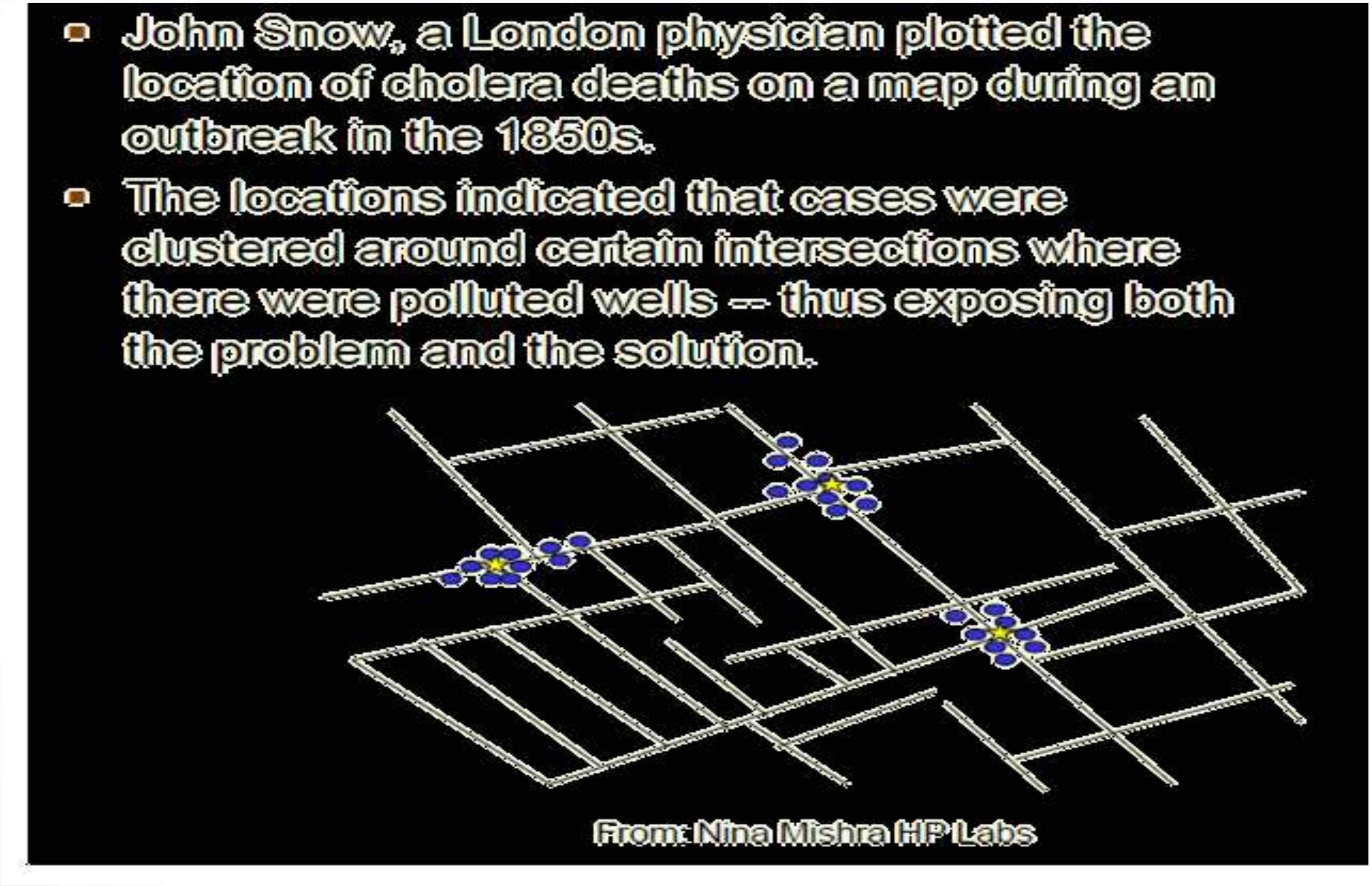
Clustering Example



Another illustration of clustering



Historical Application of Clustering



- John Snow, a London physician plotted the location of cholera deaths on a map during an outbreak in the 1850s.
- The locations indicated that cases were clustered around certain intersections where there were polluted wells -- thus exposing both the problem and the solution.

From: Nina Mishra HP Labs

Some Common Applications

- Cluster customers based on their purchase histories.
- Cluster products based on the sets of customers who purchased them
- Cluster documents based on similar words
- Cluster DNA sequences

Methods of Clustering

- **Density Based**: These methods consider the clusters as dense regions having some similarity and different from the lower dense region in space. E.g. DBSCAN algorithm etc..
- **Hierarchical**: The clusters formed in this method forms a tree like structure based on the hierarchy. E.g. BIRCH, CURE etc.
- **Partitioning** : This method partitions the objects into k clusters and each partition forms one cluster. This method is used to optimize and objective criterion similarity function such as the distance. E.g. K-means etc.
- **Grid-based**: In this method the data space is formulated into a large number of grids-like structure. All the clustering operations done on the grid are fast and independent of the number of data objects. E.g. wave cluster etc.