WHAT IS CLUSTERING?
- As the name suggests, clustering is a process of forming groups of similar objects and separating these groups from each other.
- The objects in a group are “similar” in some sense.
- Clustering is a form of unsupervised learning i.e. only unlabeled instances of data are available.
- For example, clustering might be used in Data Mining to group sports-related web pages together and differentiate them from technology-related pages.

RELATED CONCEPTS – PRE-CLUSTERING
- PCA (Principal Component Analysis)
  - A mathematical transform used to reduce the dimensionality of data space and ensure that the dimensions are uncorrelated.
  - In other words, PCA rotates the sample space so that the maximum variability is visible.
  - Good pre-processing step for clustering, since it gets rid of unnecessary dimensions and maximizes separation between clusters.

REAL-WORLD APPLICATIONS
- Marketing: Groups of consumers with similar behavior.
- Earthquake Studies: Clustering group epicenters to find dangerous zones.
- World Wide Web: Grouping webpages/blogs with similar content together.
- Biology: Clustering species of plants or animals based on observed features.

CLASSES OF CLUSTERING ALGORITHMS
- Exclusive
  - One datum belongs to exactly one cluster.
- Overlapping
  - A datum is associated with multiple clusters with corresponding “membership” values.
- Hierarchical
  - Iteratively combining similar sub-clusters till desired or optimal number of clusters is reached.
- Probabilistic
  - Assume a certain parametric distribution for each cluster.
**KMeans**
- Divides the data into K clusters, where K is specified by the user.
- Initialized by setting K centroids in the data space.
- Data samples are assigned to closest cluster and the centroids are recalculated.
- Repeated till centroids don’t move.
- [Interactive KMeans Demo](#)

**Density-based Clustering**
- Interprets clusters as regions in data space where density of samples is greater than a threshold.
- For each sample, it provides a probability of that sample belonging to each cluster.
- It can find clusters of arbitrary shape and does not need the number of clusters to be specified.

**Cobweb**
- Hierarchical method which organizes the data samples in a classification tree.
- Each tree node represents a class, and provides a probability distribution for its members.
- Considers several attributes of a data sample and groups maximally similar data together.
- Might modify the tree on insertion of a new datum.

**Expectation-Minimization (E-M)**
- Also called Gaussian Mixture Model (GMM).
- Assumes that data is generated by multiple Gaussians with certain apriori probabilities, and each Gaussian distribution represents a cluster.
- Returns the maximally likely $(\mu, \Sigma)$ for each Gaussian.
- Can recommend optimal number of clusters based on cross-validation.
- Very popular in Speech Processing.

**V-fold Cross Validation**
- Used to check the “goodness” of the clustering algorithm output.
- Divide training data into $V$ parts, train the algorithm using data of $(V-1)$ parts, and test its performance on the remaining data.
- Repeat with different sets of $(V-1)$ parts.
- Very popular for finding optimal number of clusters for KMeans or E-M.
CLUSTERING ALGORITHMS IN WEKA

- Package `weka.clusterers`: Contains EM, Kmeans, Cobweb and Density-based algorithms.
- Filters:
  - `weka.filters.unsupervised.attribute.AddCluster`
    Helps specify the number of clusters.
  - `weka.filters.unsupervised.attribute.ClusterMembership`
    Returns membership probabilities rather than strict memberships.