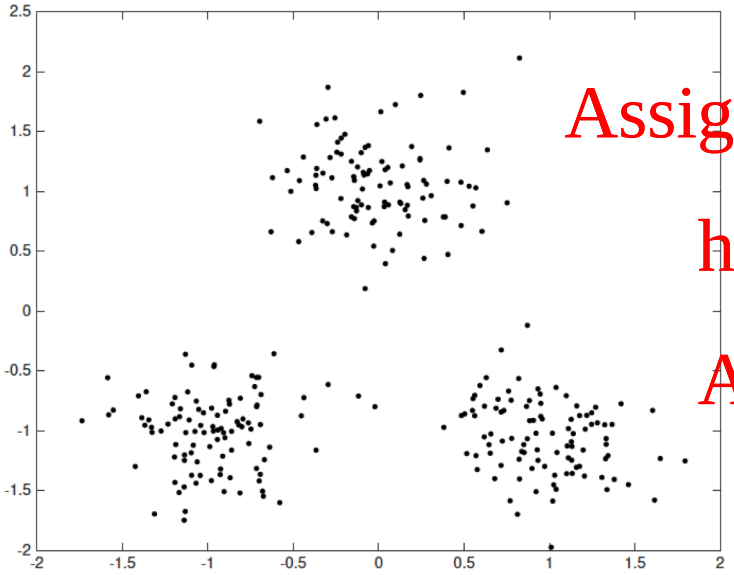
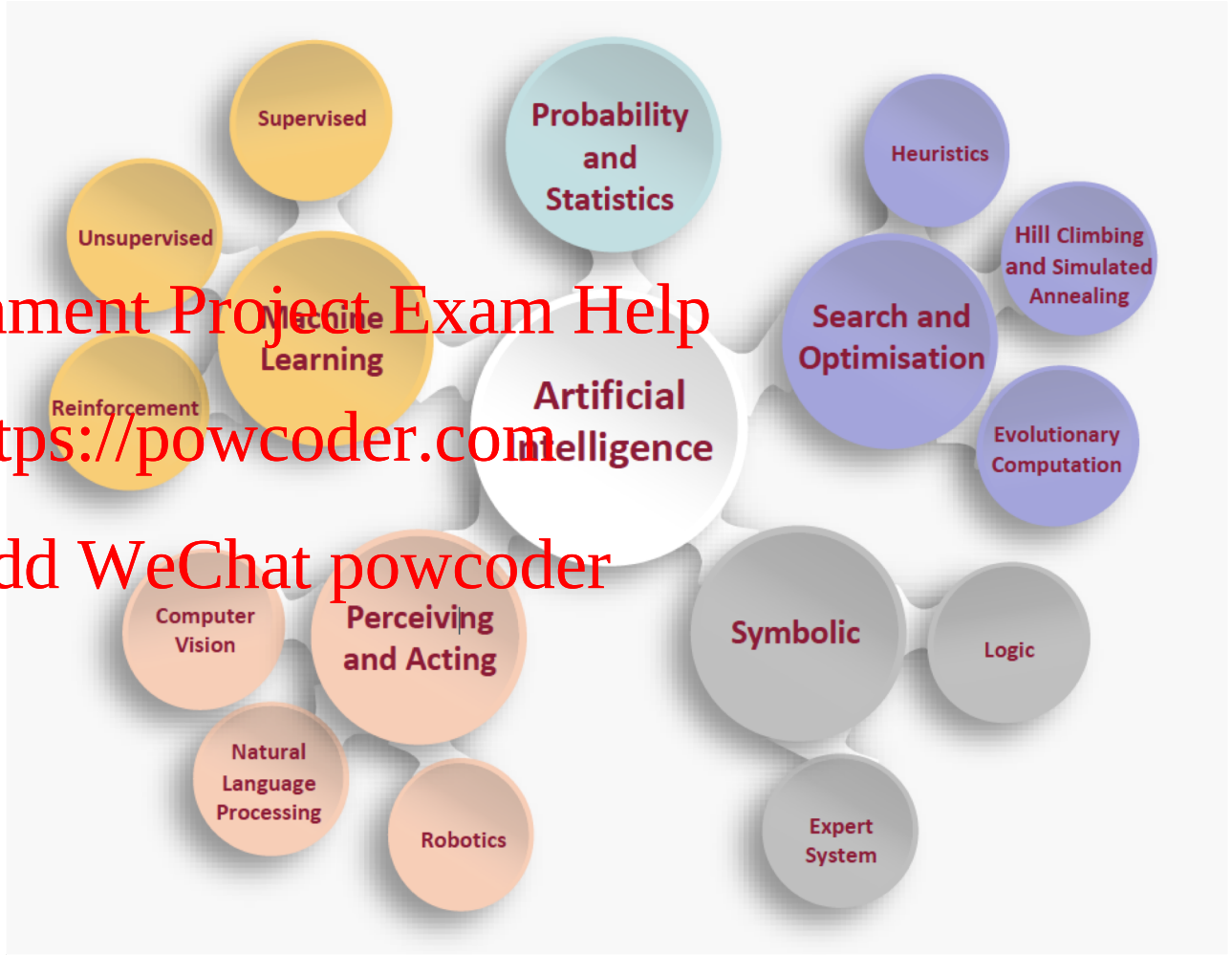
UNIVERSITY OF BIRMINGHAM

Hierarchical Clustering and Expectation-Maximization

Shuo Wang

AI Taxonomy





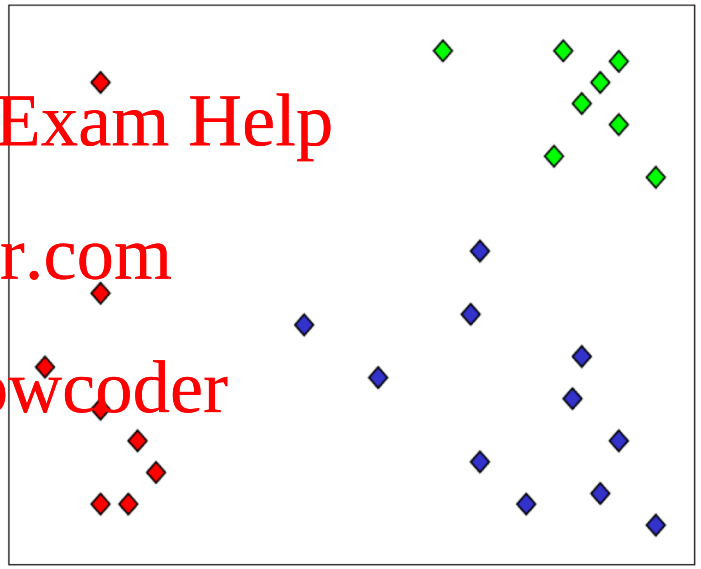
UNIVERSITY OF BIRMINGHAM

Outline

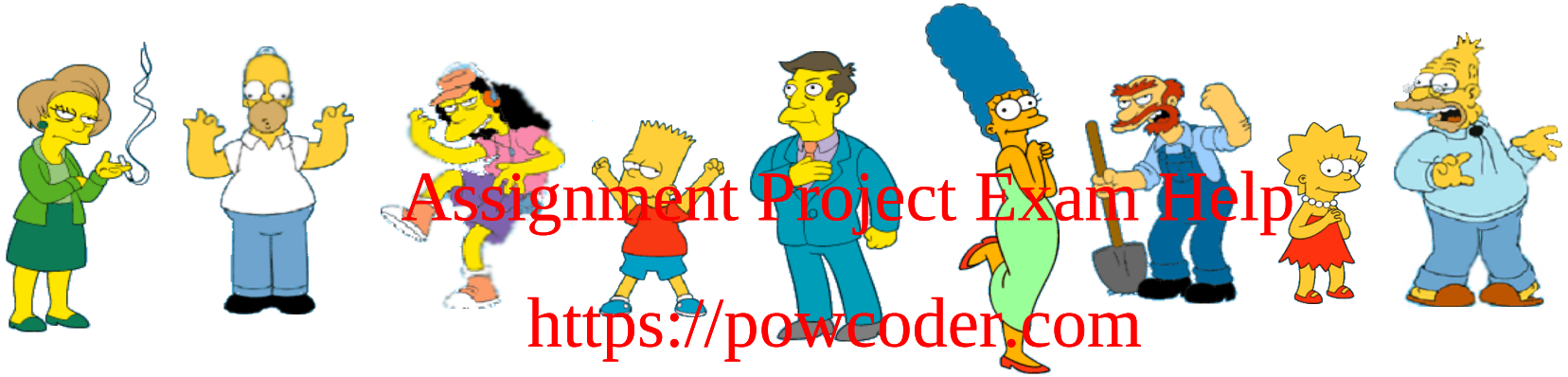
- Clustering concepts
- Hierarchical clustering
- Gaussian mixture models using EM

Clustering

- Segment data into clusters, such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural groupings among objects.

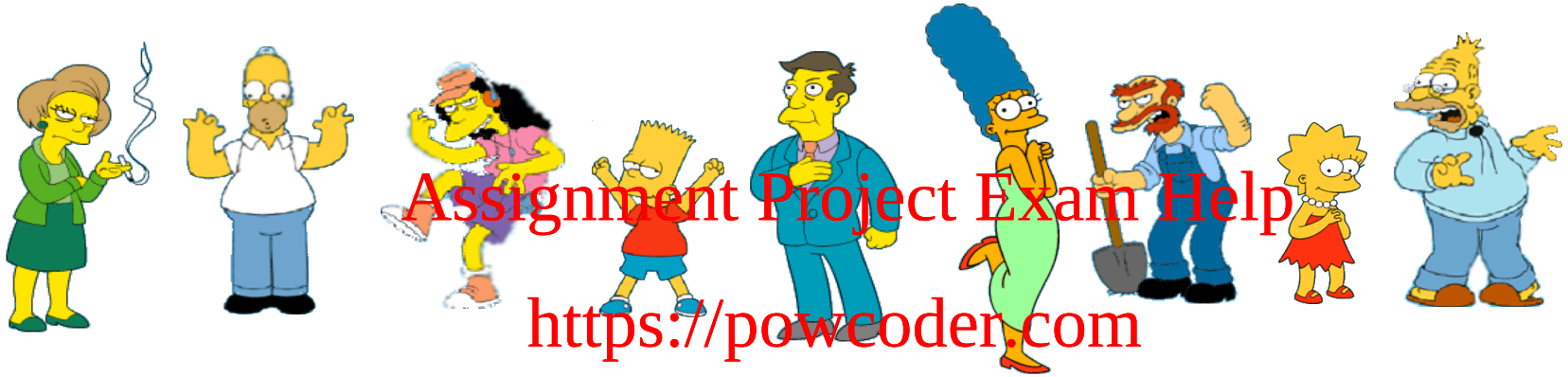


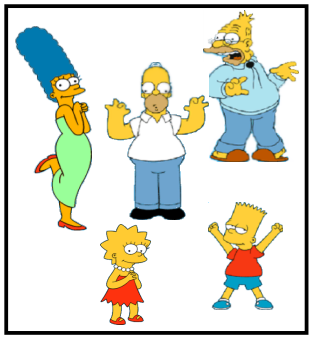
What is the natural grouping of these objects?



Picture from CMU

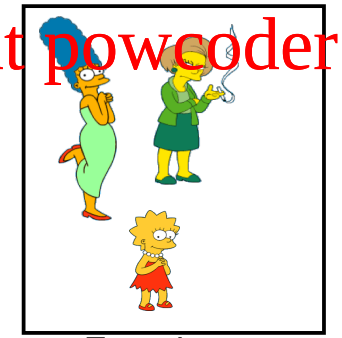
What is the natural grouping of these objects?





Simpson's family

School employees



Females

Males

Picture from CMU

Clustering set-up

- Our data are: $D = \{x_1, \ldots, x_N\}$.
- Each data point is m-dimensional, i.e. $x_i = \langle x_{i,1}, \ldots, x_{i,m} \rangle$
- Define a *distance function (i.e. similarity measures)* between data, $d(x_i, x_j)$
- Goal: segment x_n into k groups $\{z_1, \ldots, z_N\}$ where $z_i \in \{1, \ldots, K\}$

Similarity Measures

- Between any two data samples p and q, we can calculate their distance d(p,q) using a number of measurements:

Euclidean

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

Manhattan

$$d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i|$$

Chebyshev

$$d_{\text{Chebyshev}}(p,q) = \max_i(|p_i - q_i|)$$

Minkowski

$$d(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^b\right)^{1/b}$$

Types of Clustering Algorithms

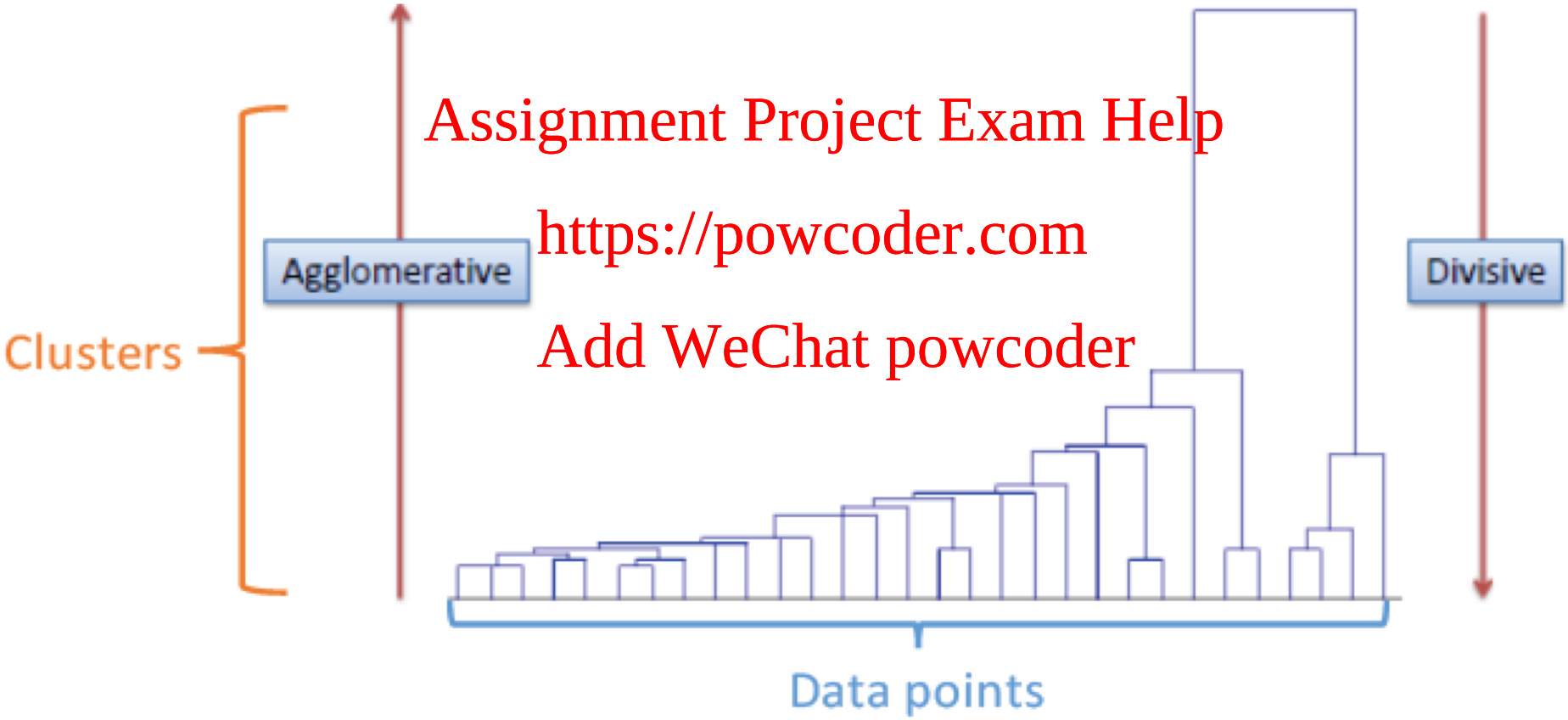
- Partitional clustering, e.g. K-means, K-medoids
- Hierarchical clustering
  - Bottom-up (agglomerative)
  - Top-down
- Density-based clustering, e.g. DBScan
- Mixture density based clustering
- Fuzzy theory based, graph theory based, grid based, etc.

Hierarchical clustering

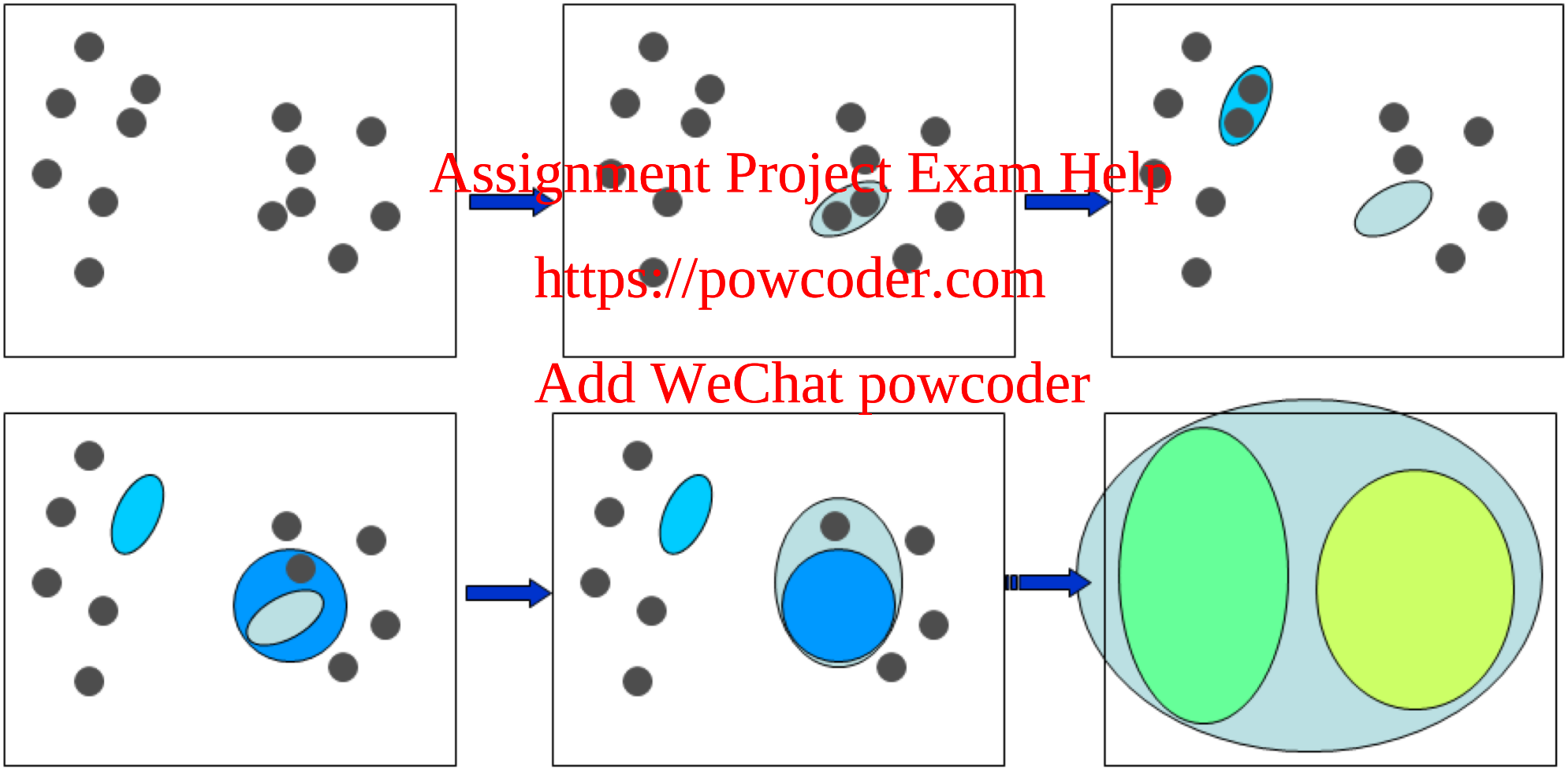
- Create a hierarchical decomposition of the set of objects using some criterion
- Produce a dendrogram

(Bovine, (Spider Monkey, (Gibbon, (Orang, (Gorilla, (Chimp, Human)))));

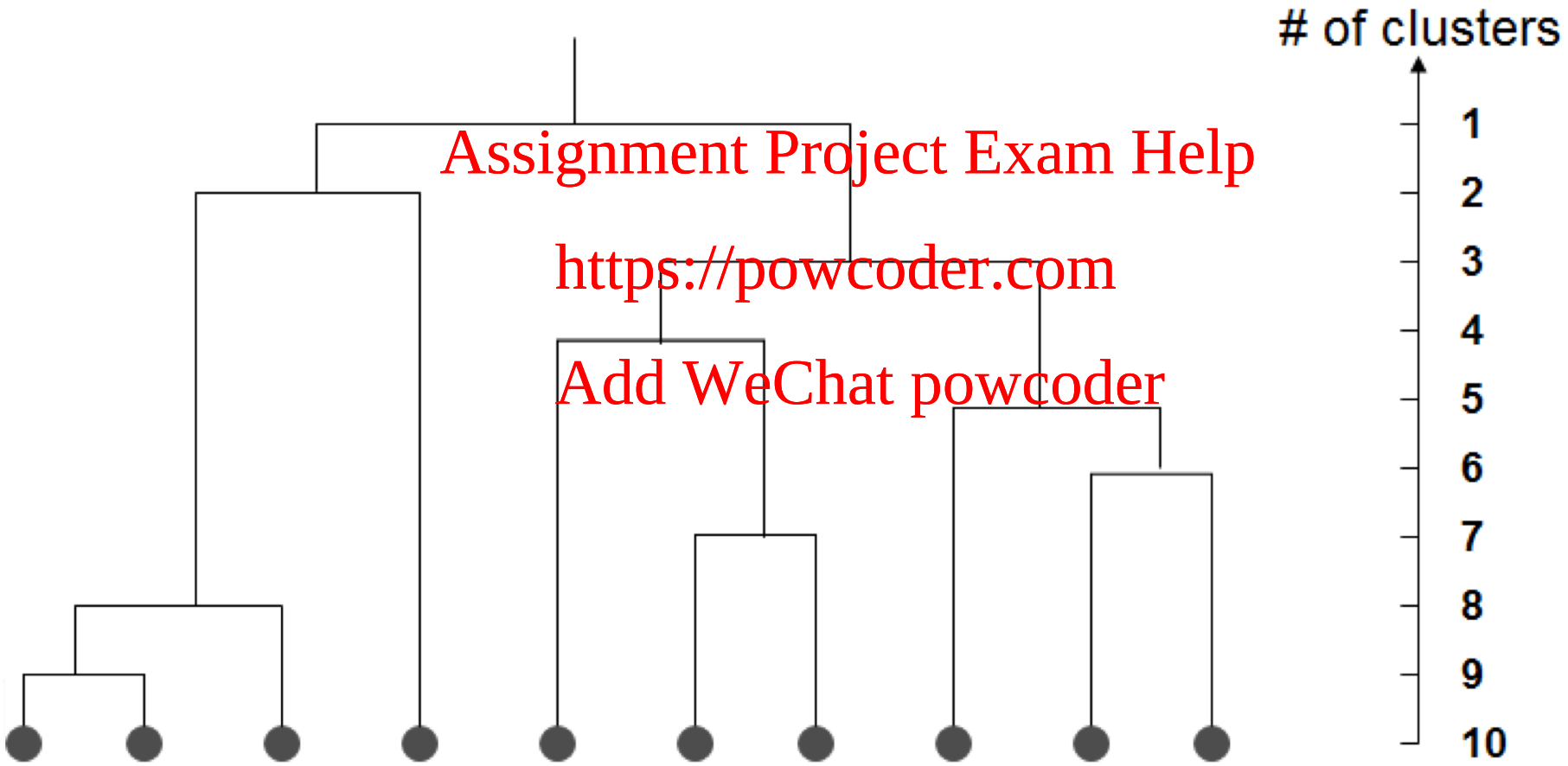
Hierarchical clustering



Agglomerative clustering illustration



Dendrogram: A visual representation of output



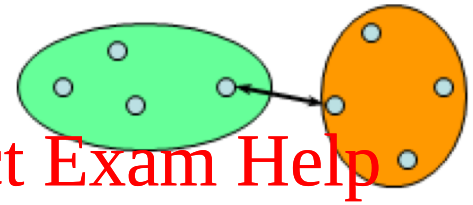
Agglomerative clustering algorithm

1. Place each data point into its own singleton group
2. Repeat: iteratively merge the two closest groups
3. Until: all the data are merged into a single cluster

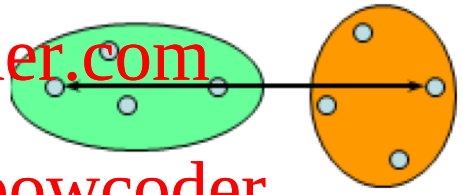
- Output: a dendrogram
- Reply on: a *distance metric* between clusters

Measuring distance between clusters

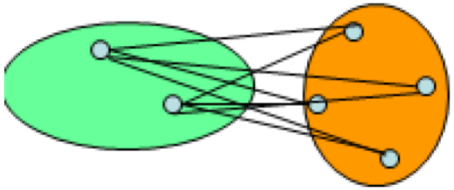
- Single linkage
  - the similarity of the closest pair
- Complete linkage
  - the similarity of the furthest pair
- Group average
  - the average similarity of all pairs
  - more widely used
  - robust against noise



$$d_{SL}(G,H) = \min_{i \in G, j \in H} d_{i,j}$$



$$d_{CL}(G,H) = \max_{i \in G, j \in H} d_{i,j}$$



$$d_{GA} = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d_{i,j}$$

Try your own data here:
http://jydelort.appspot.com/resources/figue/demo.html

Strengths, weaknesses, caveats

- Strengths
  - provides deterministic results
  - no need to specify number of clusters beforehand
  - can create clusters of arbitrary shapes
- Weakness
  - does not scale up for large datasets, time complexity at least O(n^2)
- Caveats
  - Different decisions about group similarities can lead to vastly different dendrograms.
  - The algorithm imposes a hierarchical structure on the data, even data for which such structure is not appropriate.