



UNIT-4

CLUSTERING TECHNIQUES

CLUSTER

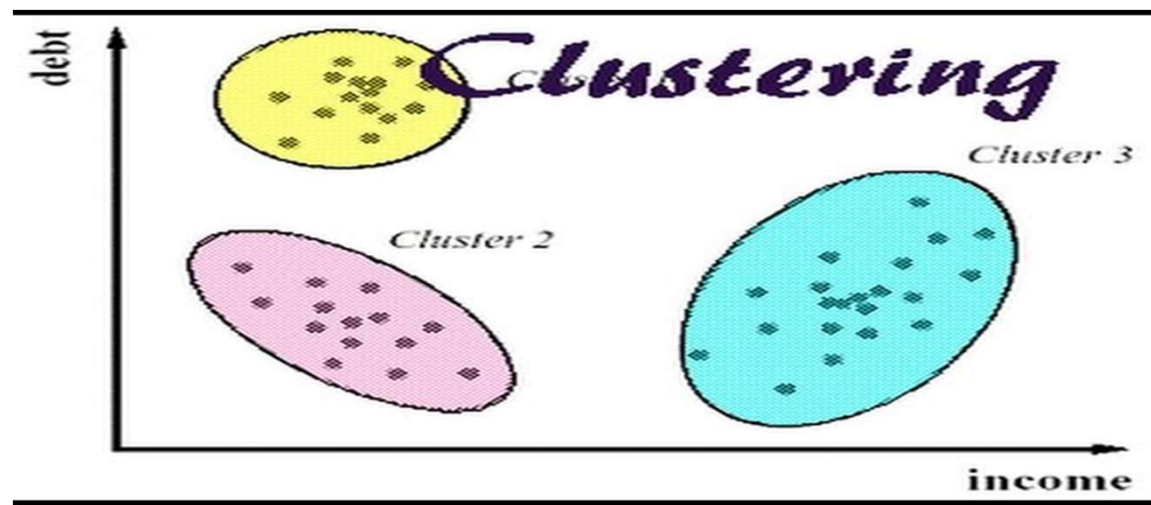
- Cluster is a group of objects that belongs to the same class.
- In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.



CLUSTERING

What is Clustering?

- Clustering is the process of making a group of abstract objects into classes of similar objects.

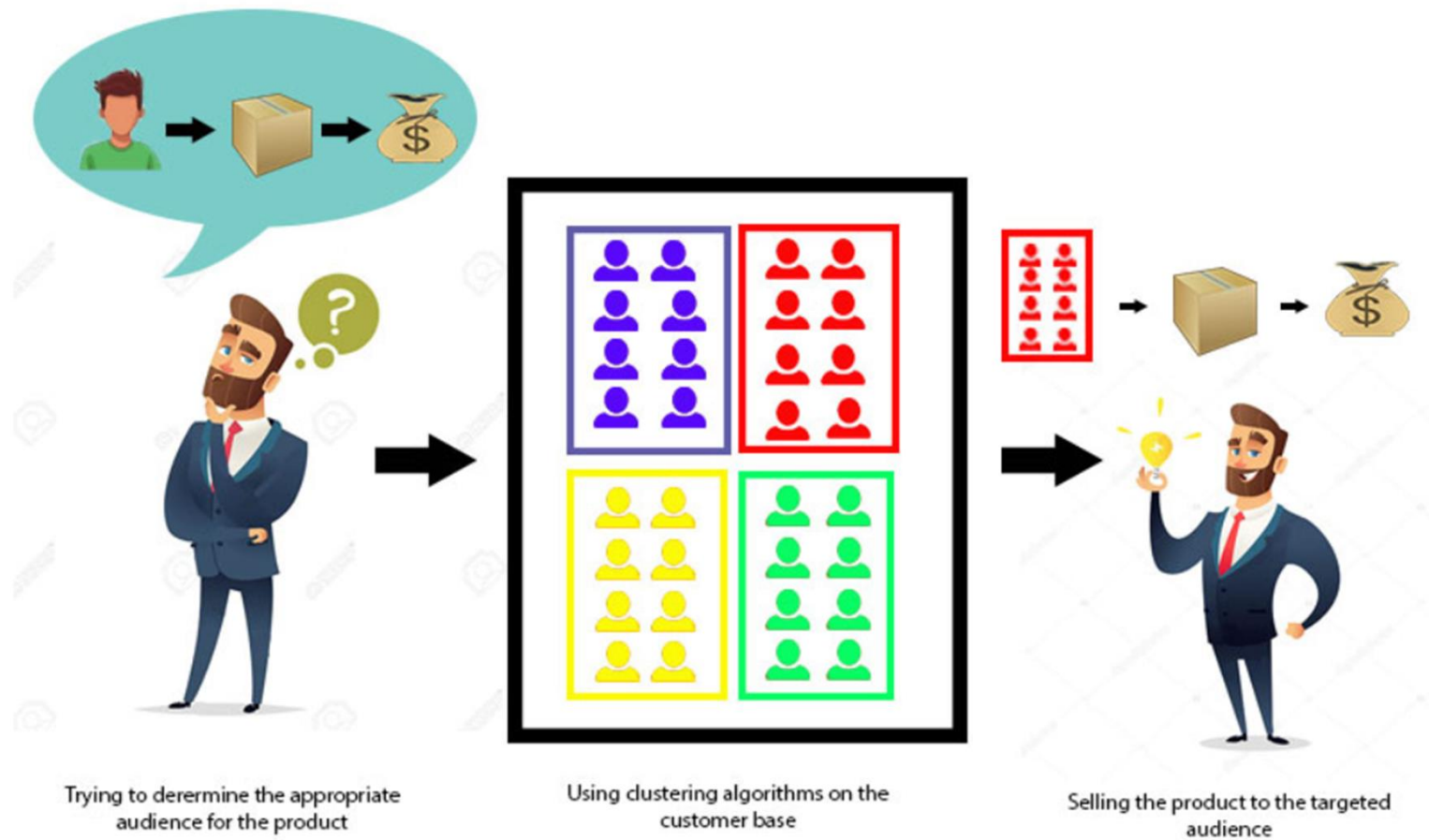


CLUSTERING

- The process of grouping a set of physical or abstract objects into classes of *similar objects* is called clustering.
- Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their *similarity*.
- *Clustering can also be* used for outlier detection



EXAMPLE: APPLICATION OF CLUSTERING ALGORITHM



APPLICATION OF CLUSTERING ALGORITHM

- The data from customer base is divided into clusters; we can make an informed decision about who we think is best suited for this product.
- Suppose we are a market manager, and we have a new tempting product to sell.
- We are sure that the product would bring enormous profit, as long as it is sold to the right people.
- So, how can we tell who is best suited for the product from our company's huge customer base?



CLUSTERING

- In machine learning, clustering is an example of unsupervised learning.
- Unsupervised learning do not rely on predefined classes and class-labeled training examples.
- For this reason, clustering is a form of learning by observation, rather than learning by examples.



POINTS TO REMEMBER

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups



A CATEGORIZATION OF MAJOR CLUSTERING METHODS

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method



HIERARCHICAL METHODS

- A hierarchical method creates a hierarchical decomposition of the given set of data objects.
- A hierarchical clustering method works by grouping data objects into a tree of clusters.
- We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed



CLASSIFICATION

(1) Agglomerative & Divisive Hierarchical Clustering

(2) CURE

(3) Chameleon



AGGLOMERATIVE HIERARCHICAL CLUSTERING

- A hierarchical method can be classified as being either *agglomerative*, also called the *bottom-up approach*, starts with each object forming a separate group.
- It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination conditions are satisfied.

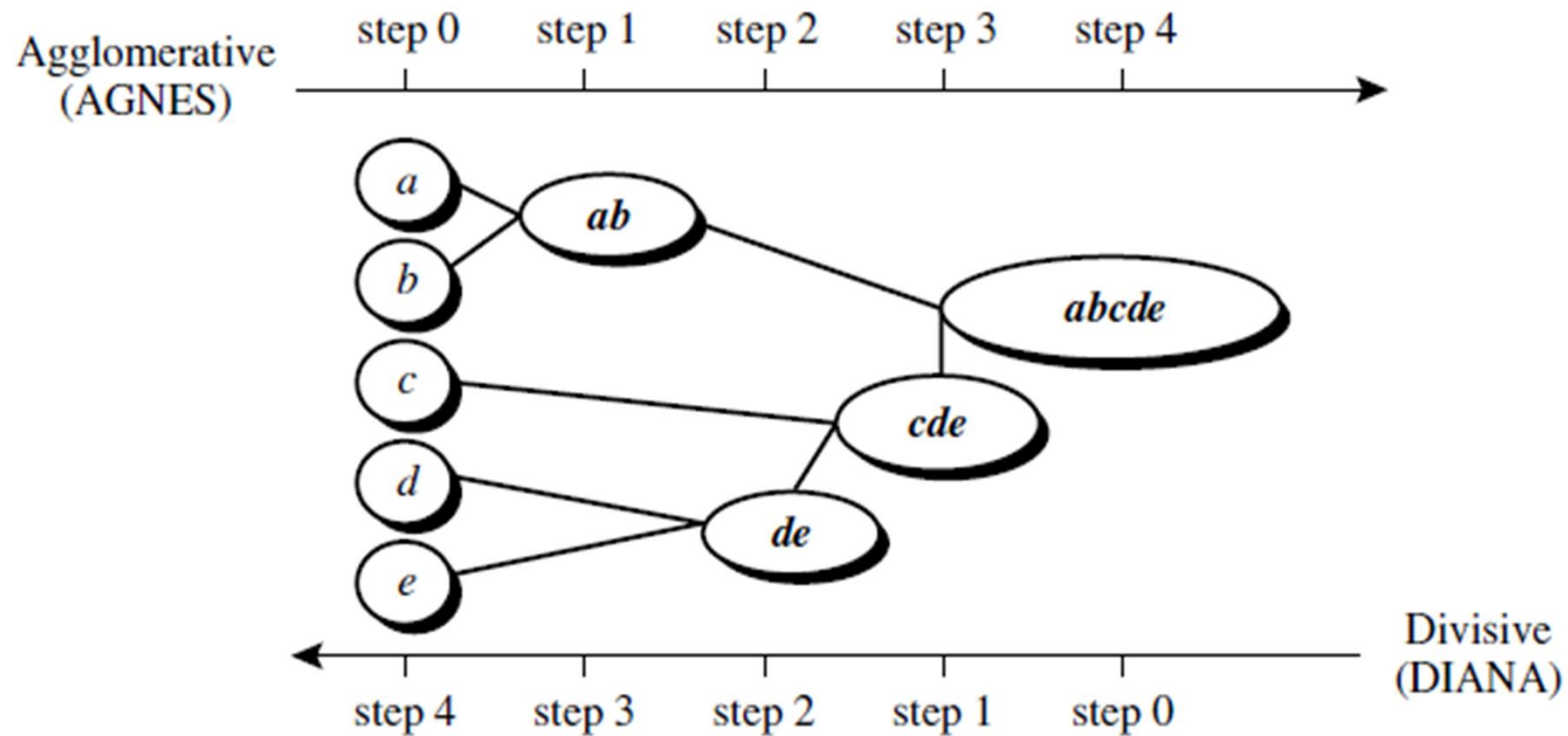


DIVISIVE HIERARCHICAL CLUSTERING

- The *divisive approach*, also called the *top-down approach*, starts with all of the objects in the same cluster.
- In each successive iteration, It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, each cluster is within a certain threshold.




EXAMPLE



Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

EXAMPLE

- The figure shows AGNES (AGglomerative NESting), an agglomerative hierarchical clustering method, and DIANA (DIvisive ANAlysis), a divisive hierarchical clustering method, to a data set of five objects, $\{a, b, c, d, e\}$.
 - *Initially, AGNES places each object into a cluster of its own.*
 - The clusters are then merged step-by-step according to some criterion.
 - For example, clusters $C1$ and $C2$ may be merged if an object in $C1$ and an object in $C2$ form the minimum Euclidean distance between any two objects from different clusters.
 - This is a single-linkage approach in that each cluster is represented by all of the objects in the cluster, and the similarity between two clusters is measured by the similarity of the *closest* pair of data points belonging to different clusters.
 - The cluster merging process repeats until all of the objects are eventually merged to form one cluster.
- 

EXAMPLE

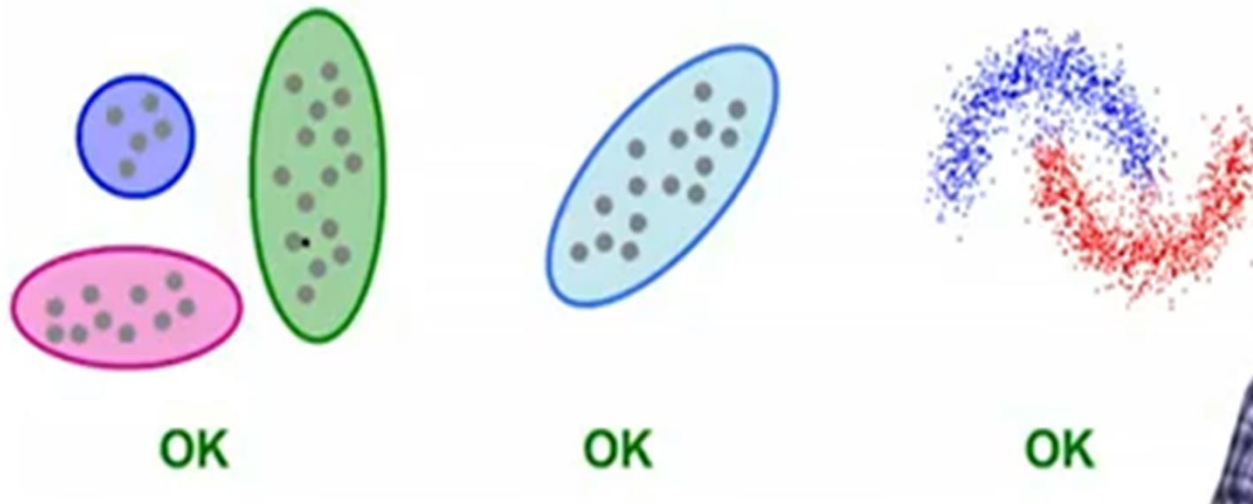
- In DIANA, all of the objects are used to form one initial cluster.
- The cluster is split according to some principle, such as the maximum Euclidean distance between the closest neighboring objects in the cluster.
- The cluster splitting process repeats until, eventually, each new cluster contains only a single object

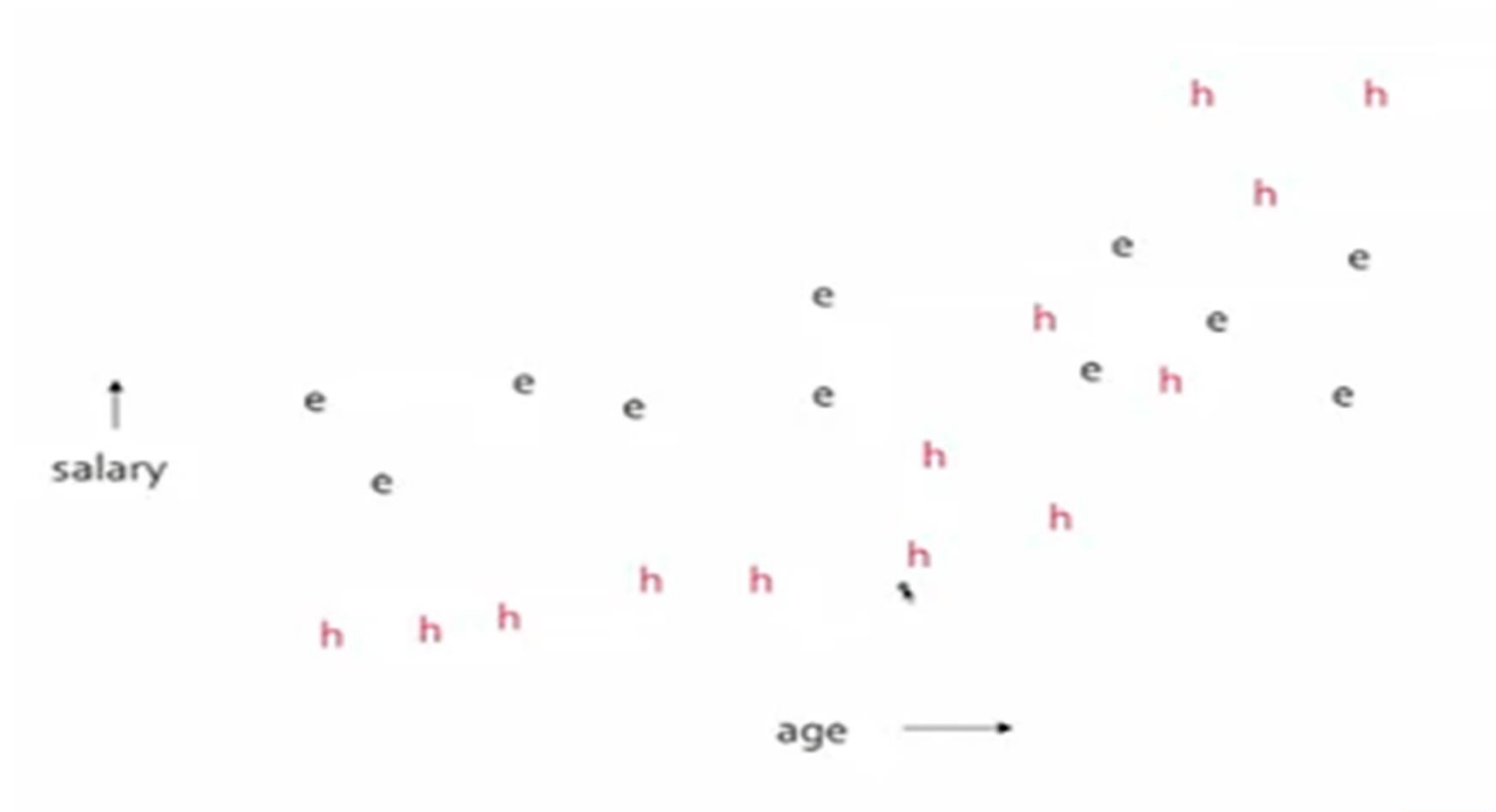


(2B) CURE

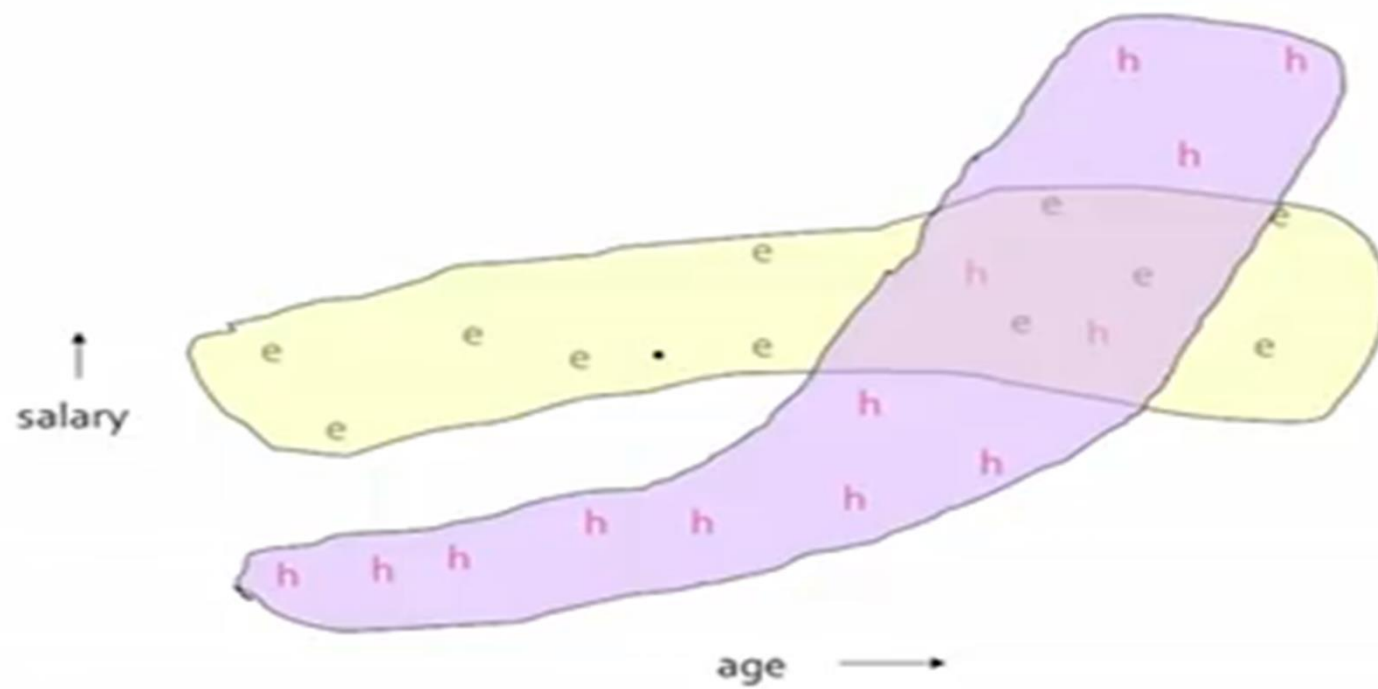
■ CURE (Clustering Using REpresentatives):

- Assumes a Euclidean distance
- Allows clusters to assume any shape
- Uses a collection of representative points to represent clusters

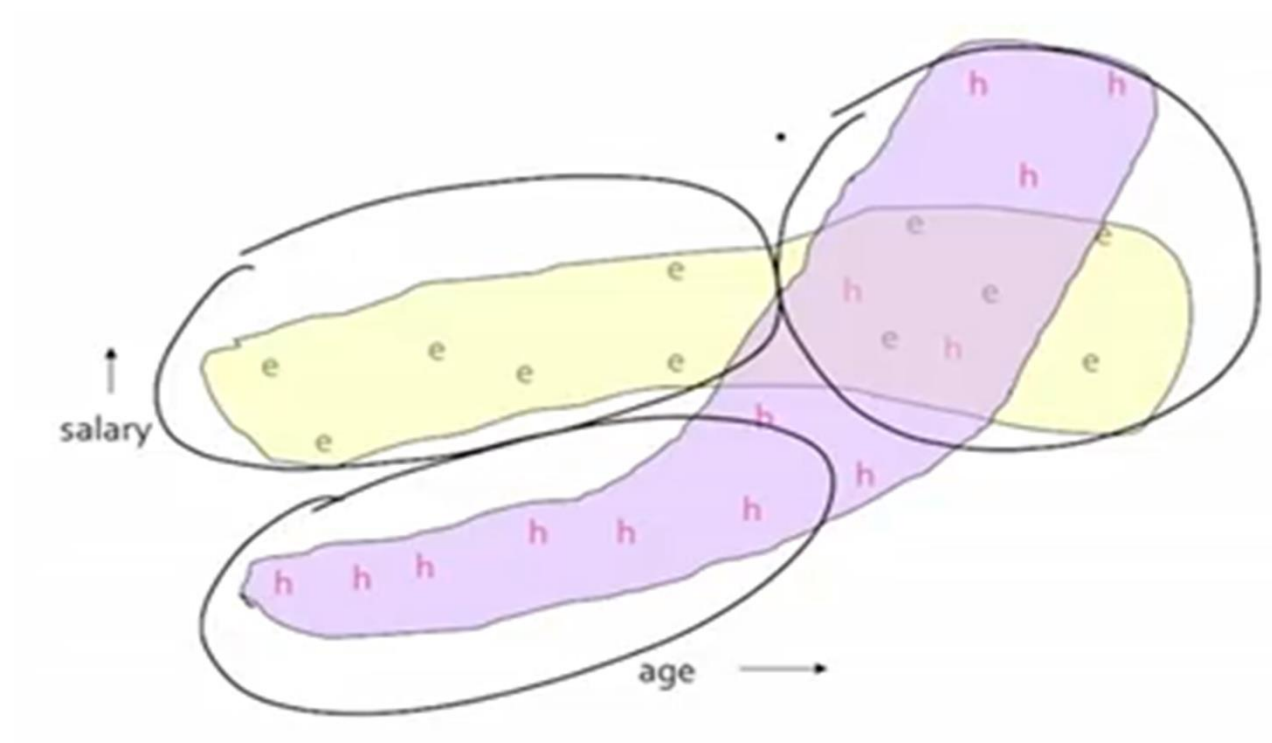




EXAMPLE

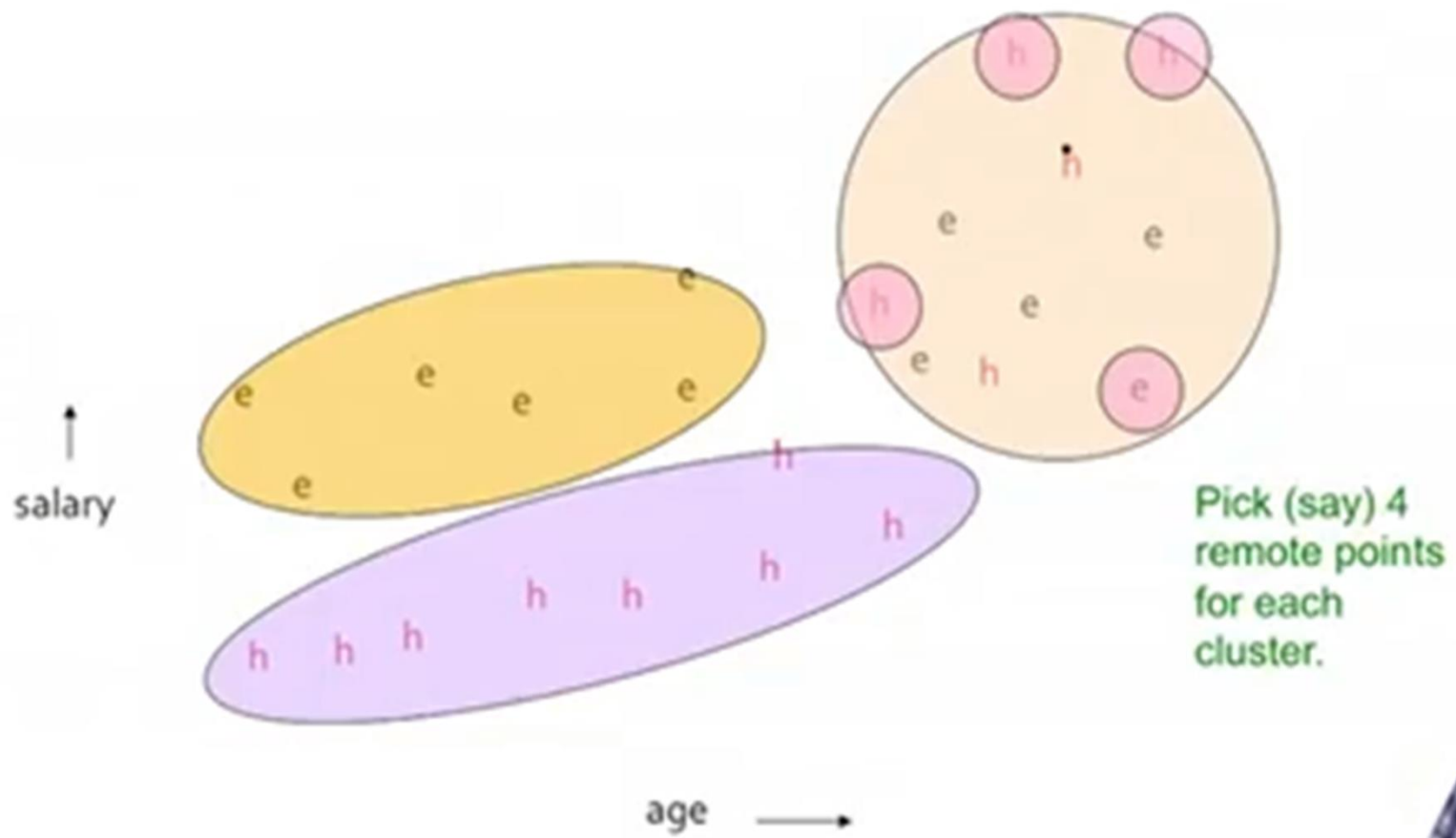


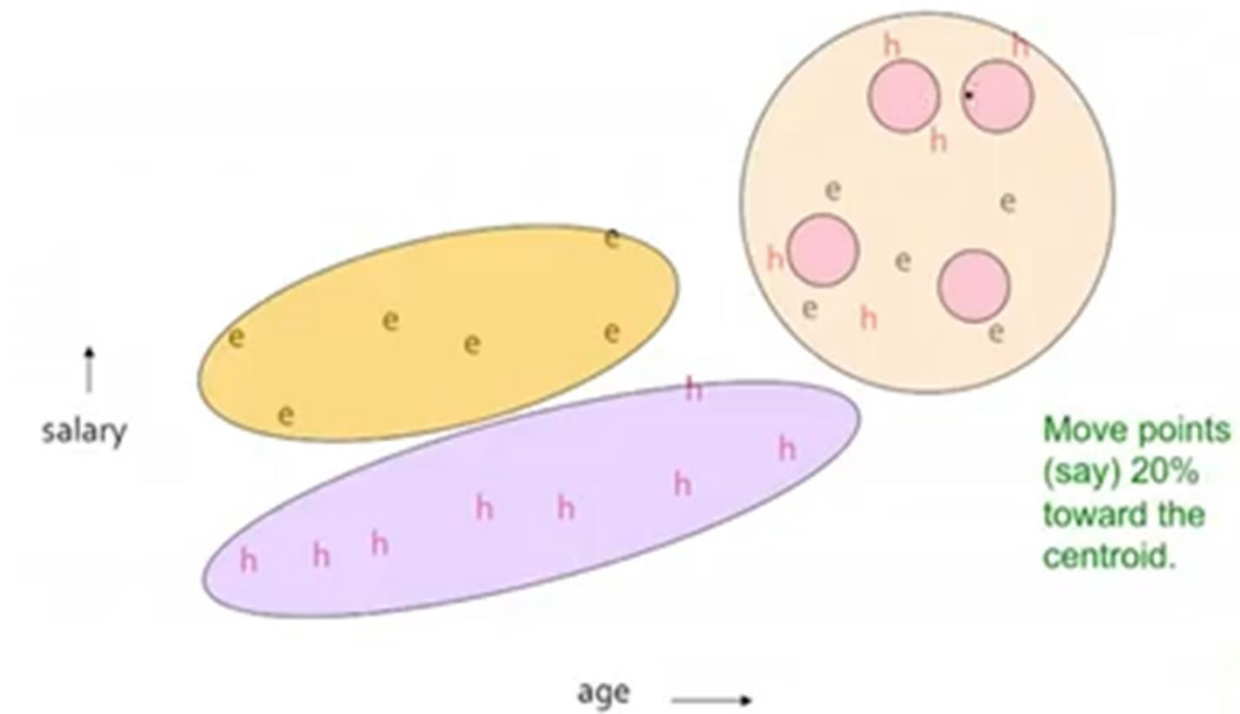
EXAMPLE



Pass 1 of 2:

- Pick a random sample of points that fit in main memory
- Cluster sample points hierarchically to create the initial clusters
- **Pick representative points:**
 - For each cluster, pick k (e.g., 4) representative points, as dispersed as possible
 - Move each representative point a fixed fraction (e.g., 20%) toward the centroid of the cluster





FINISHING CURE

Pass 2 of 2:

- Now, rescan the whole dataset and visit each point p in the data set
- Place it in the “closest cluster”
 - Normal definition of “closest”: that cluster with the closest (to p) among all the representative points of all the clusters



CHAMELEON: A HIERARCHICAL CLUSTERING ALGORITHM USING DYNAMIC MODELING

- Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between pairs of clusters.
- It was derived based on the observed weaknesses of two hierarchical clustering algorithms: ROCK (ignores cluster nearness) and CURE (ignores cluster interconnectivity)



CHAMELEON

- In Chameleon, cluster similarity is assessed based on how well-connected objects are within a cluster *and on the nearness (proximity) of clusters*.
- *That is, two* clusters are merged if their *interconnectivity is high and they are close together*.
- *Thus*, Chameleon does not depend on a static, user-supplied model and can automatically adapt to the internal characteristics of the clusters being merged



HOW DOES CHAMELEON WORK?

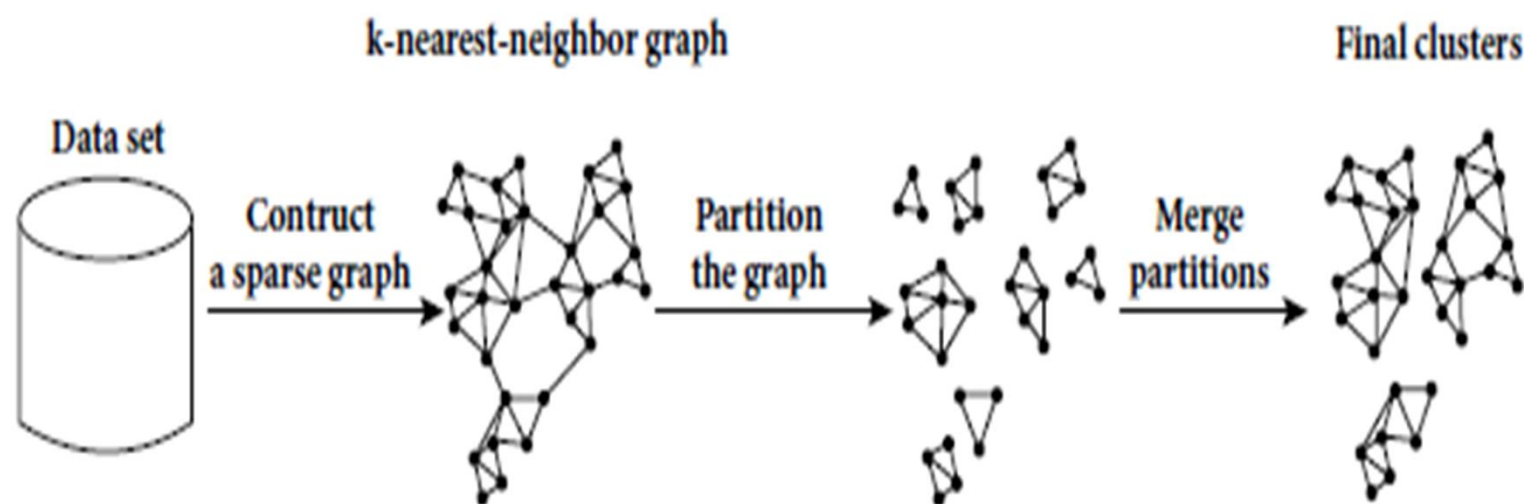
- Chameleon uses a *k-nearest-neighbor graph* approach to construct a sparse graph, where each vertex of the graph represents a data object, and there exists an edge between two vertices (objects) if one object is among the *k-most-similar objects of the other*.
- The edges are weighted to reflect the similarity between objects.
- Chameleon uses a graph partitioning algorithm to partition the *k-nearest-neighbor graph* into a large number of relatively small subclusters.



HOW DOES CHAMELEON WORK?

- It then uses an agglomerative hierarchical clustering algorithm that repeatedly merges subclusters based on their similarity.
- To determine the pairs of most similar subclusters, it takes into account both the interconnectivity as well as the closeness of the clusters





-
- 9 Chameleon: Hierarchical clustering based on k -nearest neighbors and dynamic modeling. Based on [KHK99].



OVERALL FRAMEWORK OF CHAMELEON

