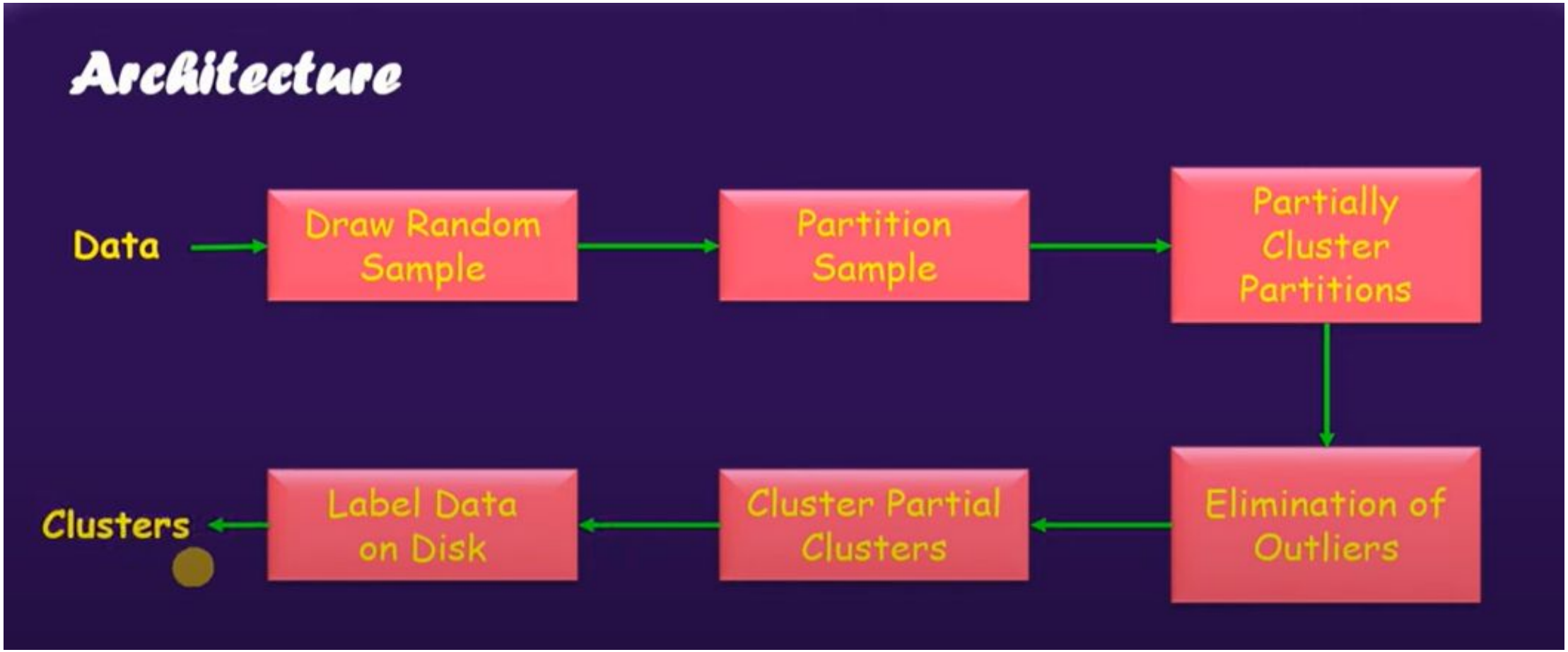
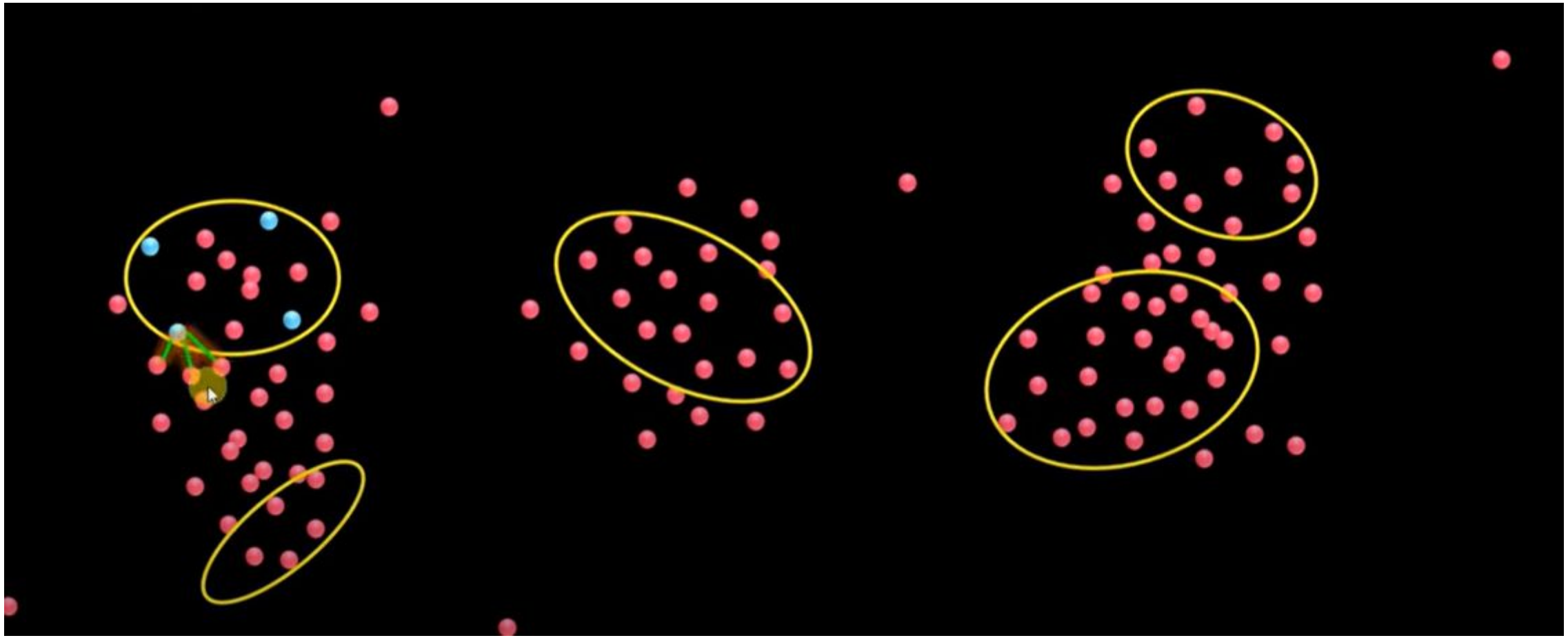
CURE

Clustering Using Representatives

Overview

- CURE stands for Clustering Using REpresentatives
- Specially designed to work efficiently on larger datasets
- Uses a collection of representative points to represent clusters
- Adopts a middle ground between centroid based and all-point extremes
- Capable of detecting clusters of any shape
- Detect outliers and remove it


Architecture
Data
Draw Random Sample
Partition Sample
Partially Cluster Partitions
Elimination of Outliers
Cluster Partial Clusters
Label Data on Disk
Clusters



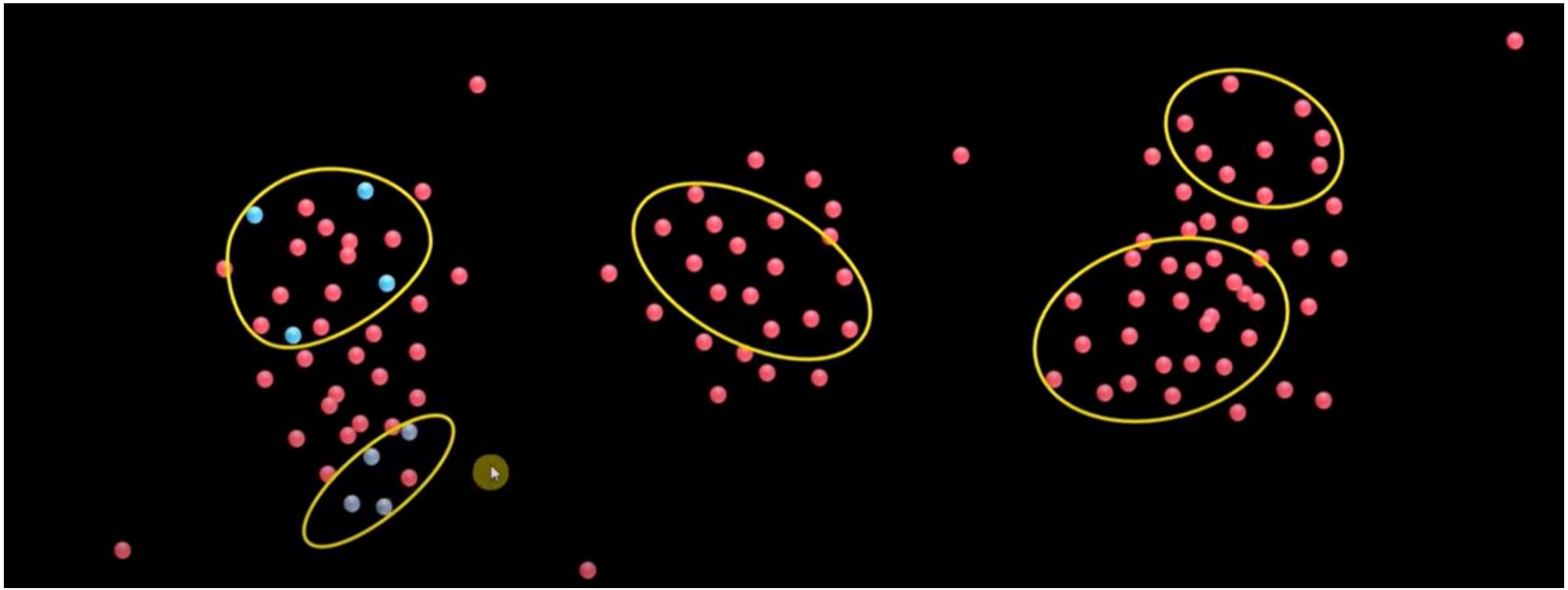
Algorithm

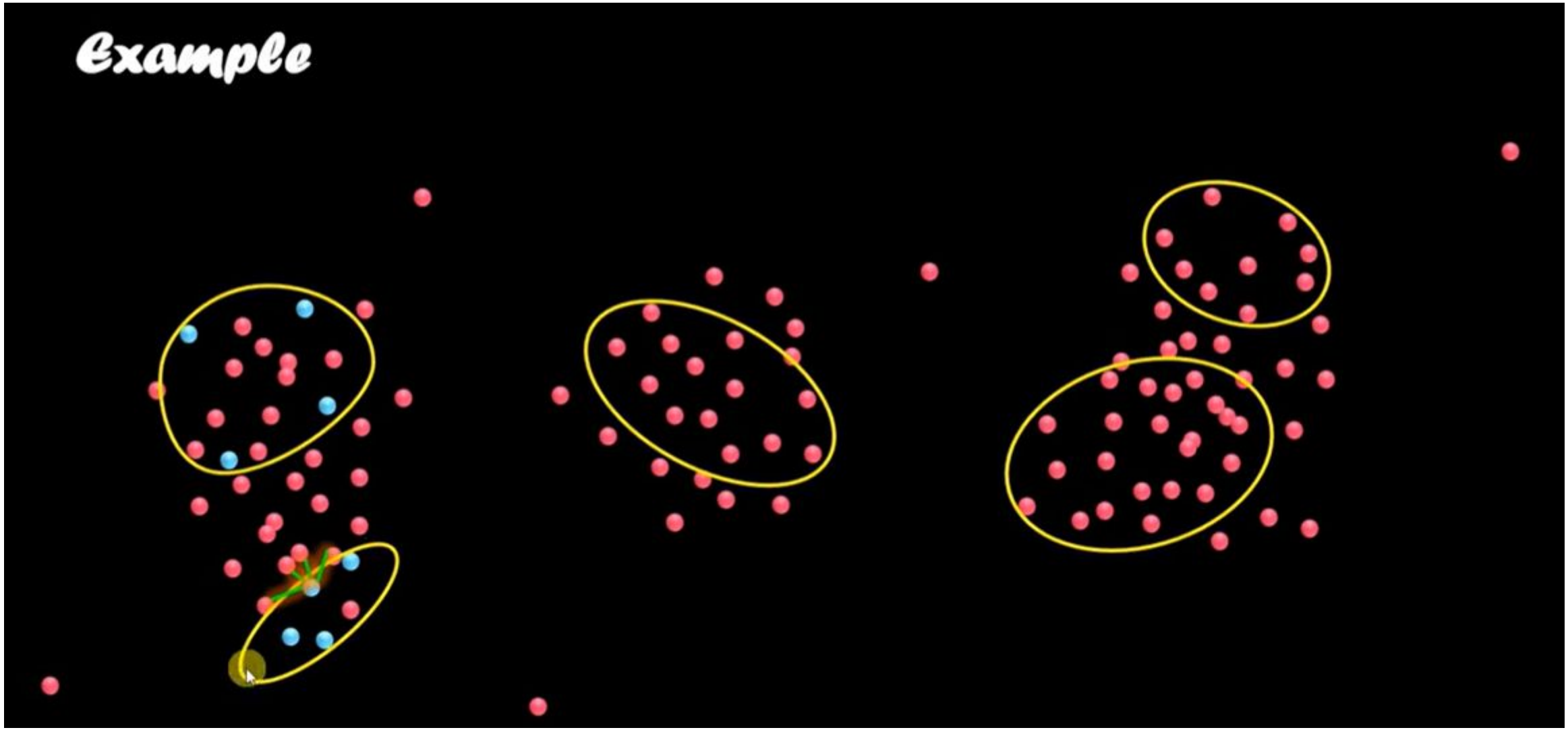
Pass 1:

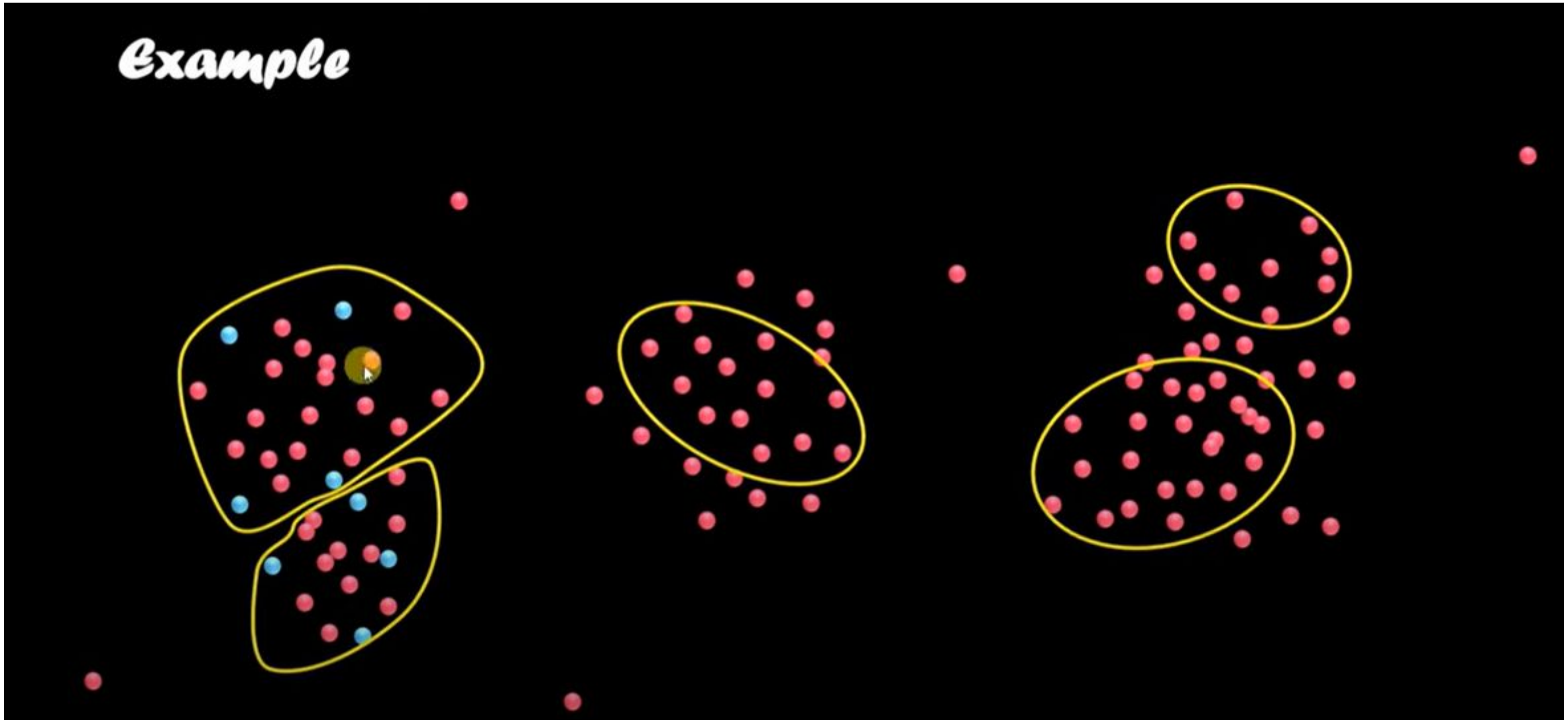
1. Pick random samples that fit in main memory and cluster them
2. Choose c scattered points in each cluster. (Let's take c = 4)
3. These scattered points are shrunk towards centroid in a fraction of α where $0 < \alpha < 1$
4. Use dmin cluster merging approach considering these scattered points as representatives of clusters

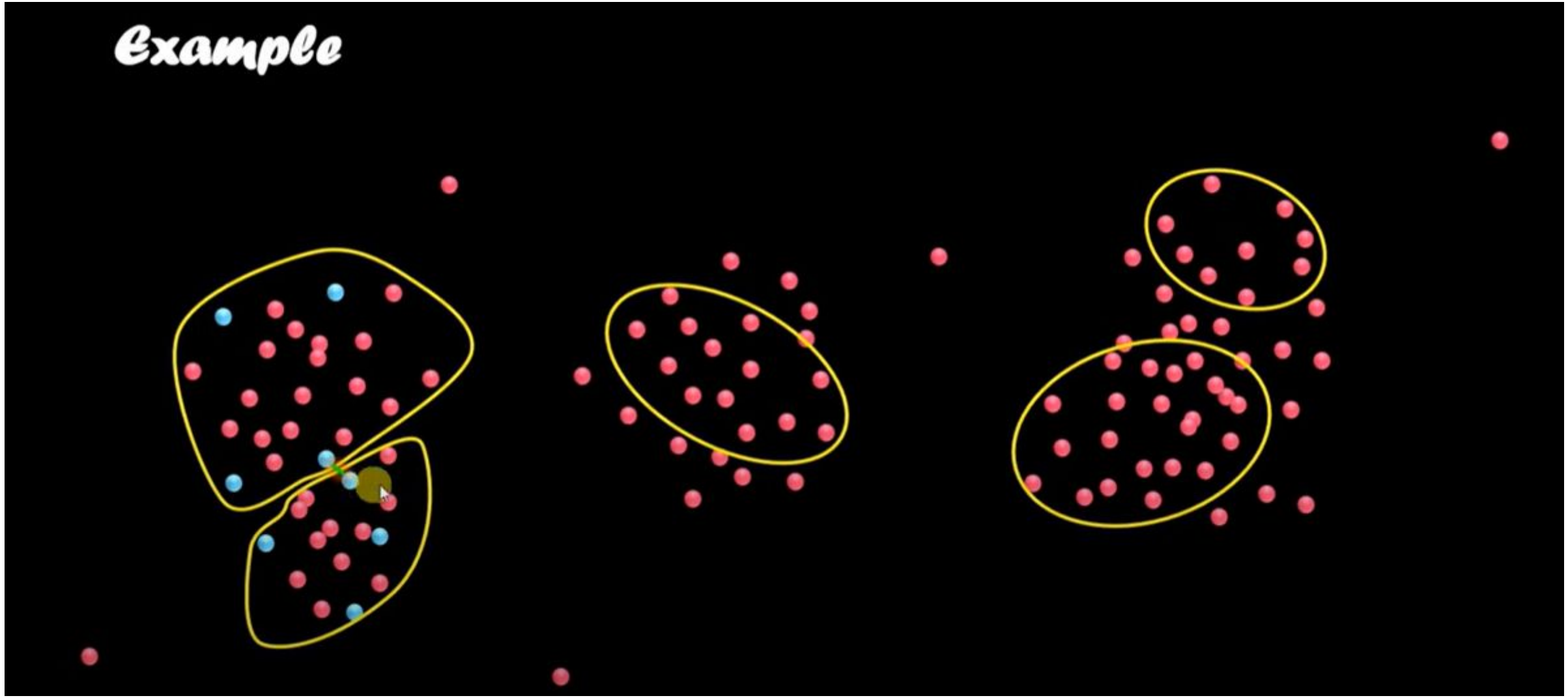
Pass 2:

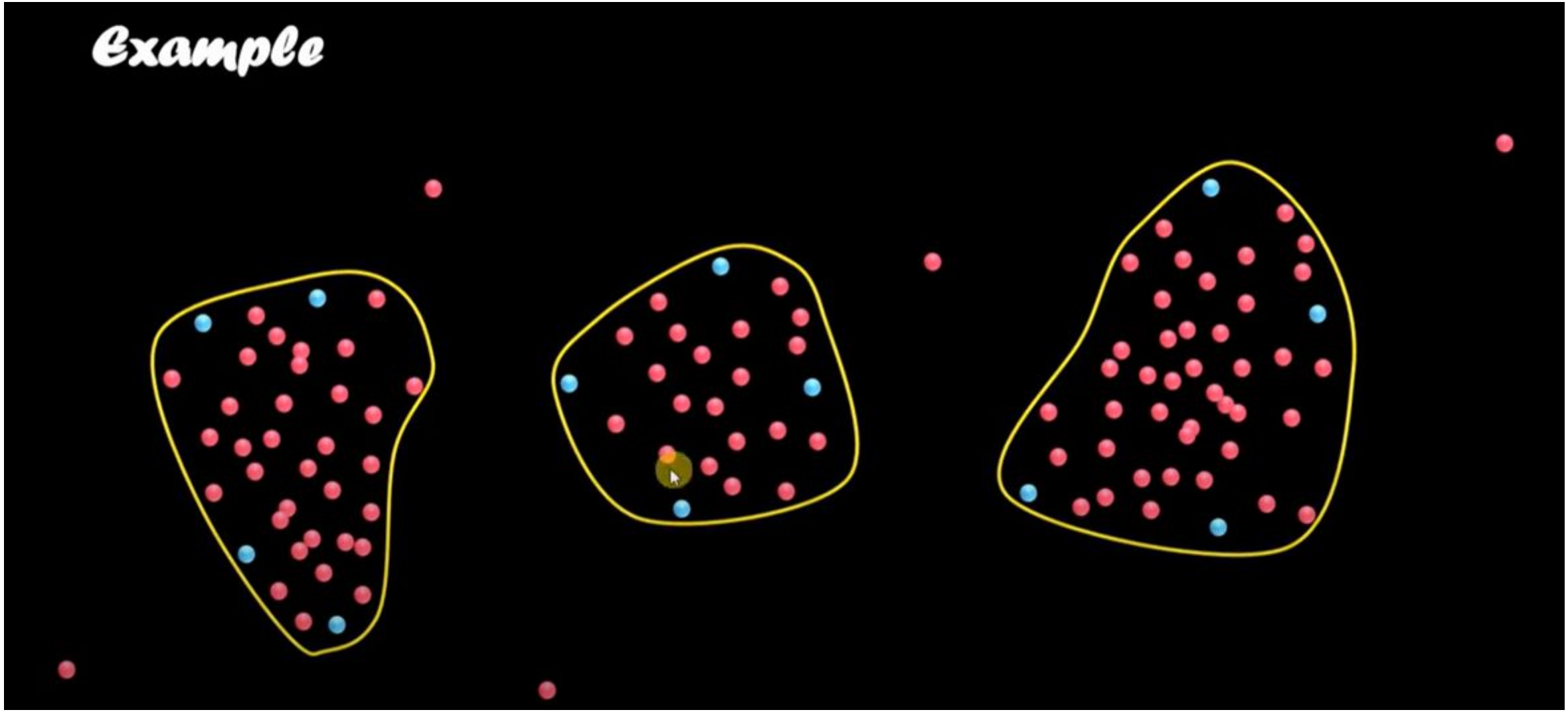
1. After every merge, new points merged are considered as representative for new cluster
2. Finally, cluster merging will stop when target k (number of clusters) is achieved

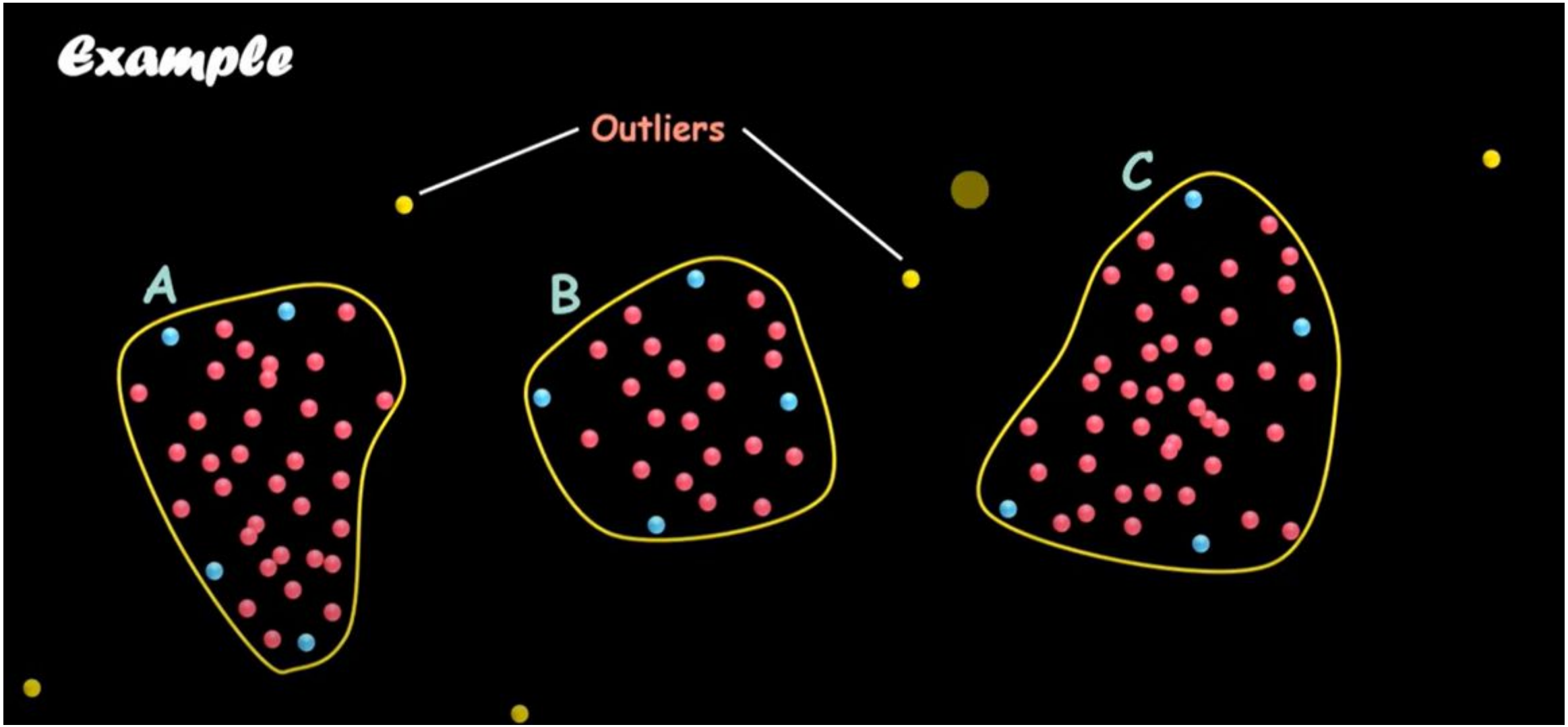


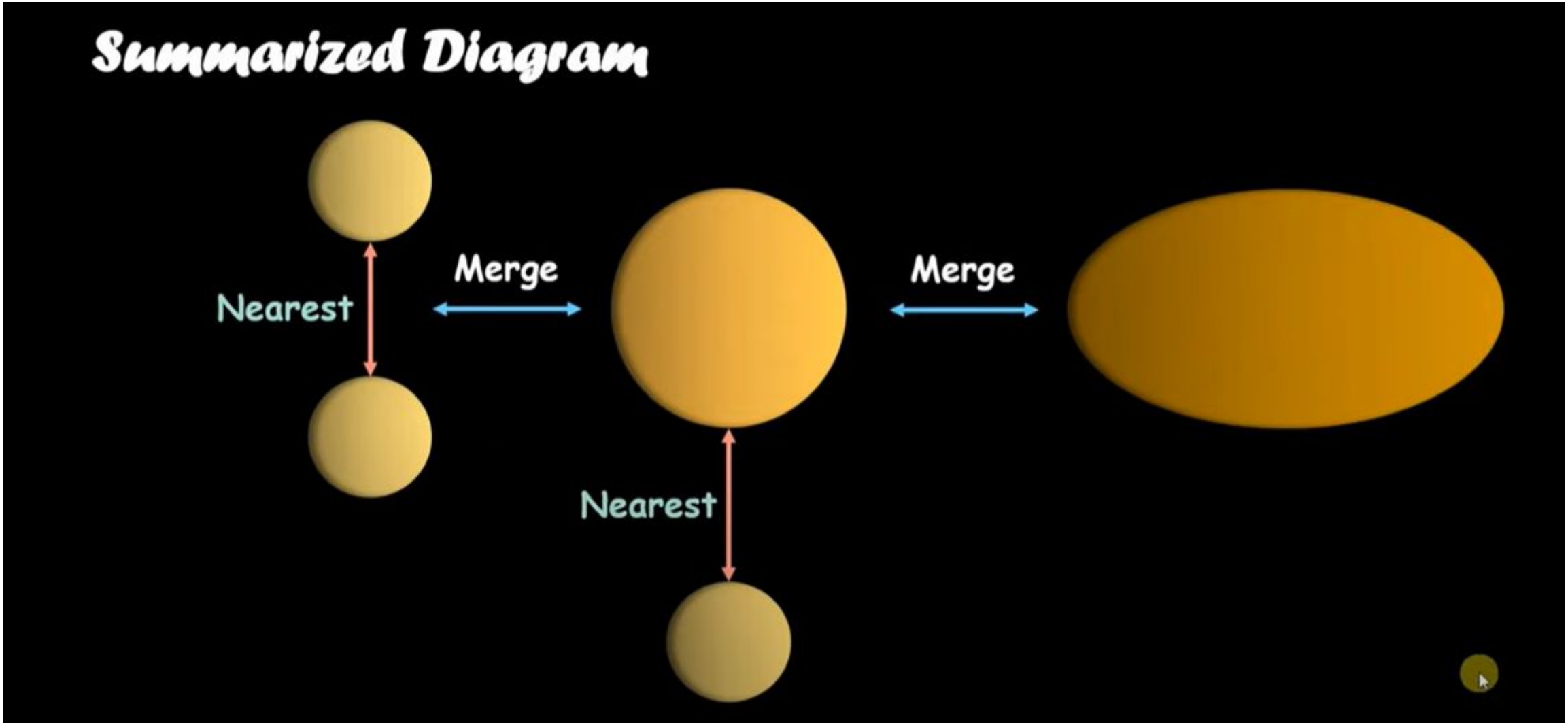

Example


Example


Example


Example


Example
Outliers
A
B
C


Summarized Diagram
Nearest
Merge
Merge
Nearest

Advantages

- Accurate results
- Adjusts perfectly to non-spherical cluster shapes
- Efficient for large datasets
- Less sensitive to outliers
- Time complexity: $O(n^2 \log n)$ [$O(n^2)$ for small dimensionality of data]