

Clustering

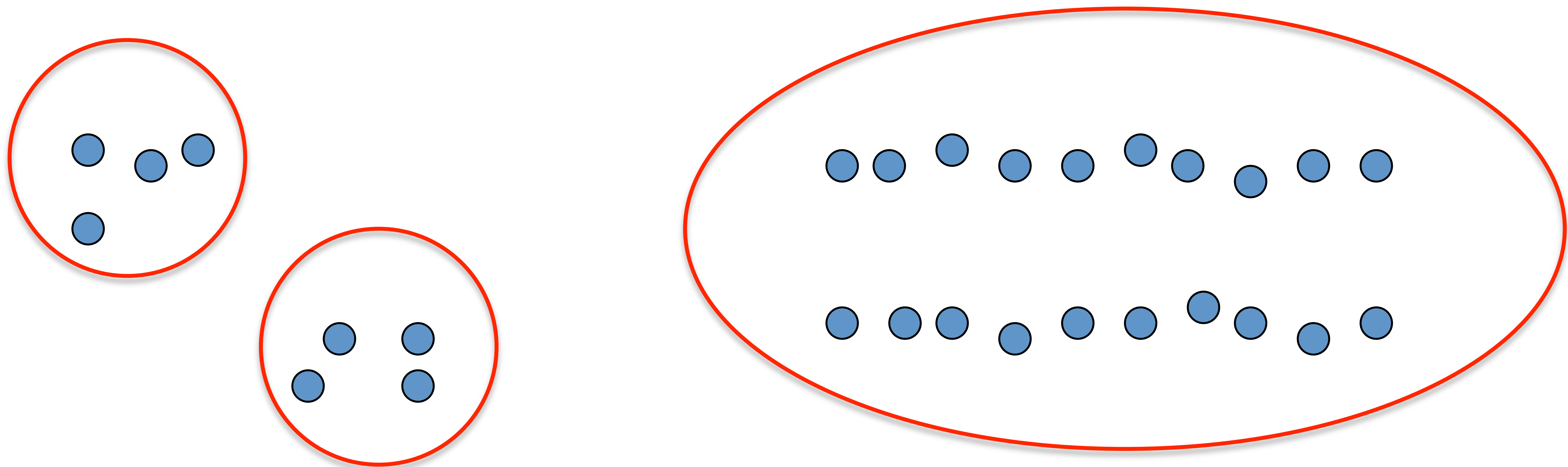
Clustering:

- **Unsupervised learning**
- Requires data, but no labels
- **Detect patterns** e.g. in
 - Group emails or search results
 - Customer shopping patterns
 - Regions of images
- Useful when don't know what you're looking for
- But: can get gibberish



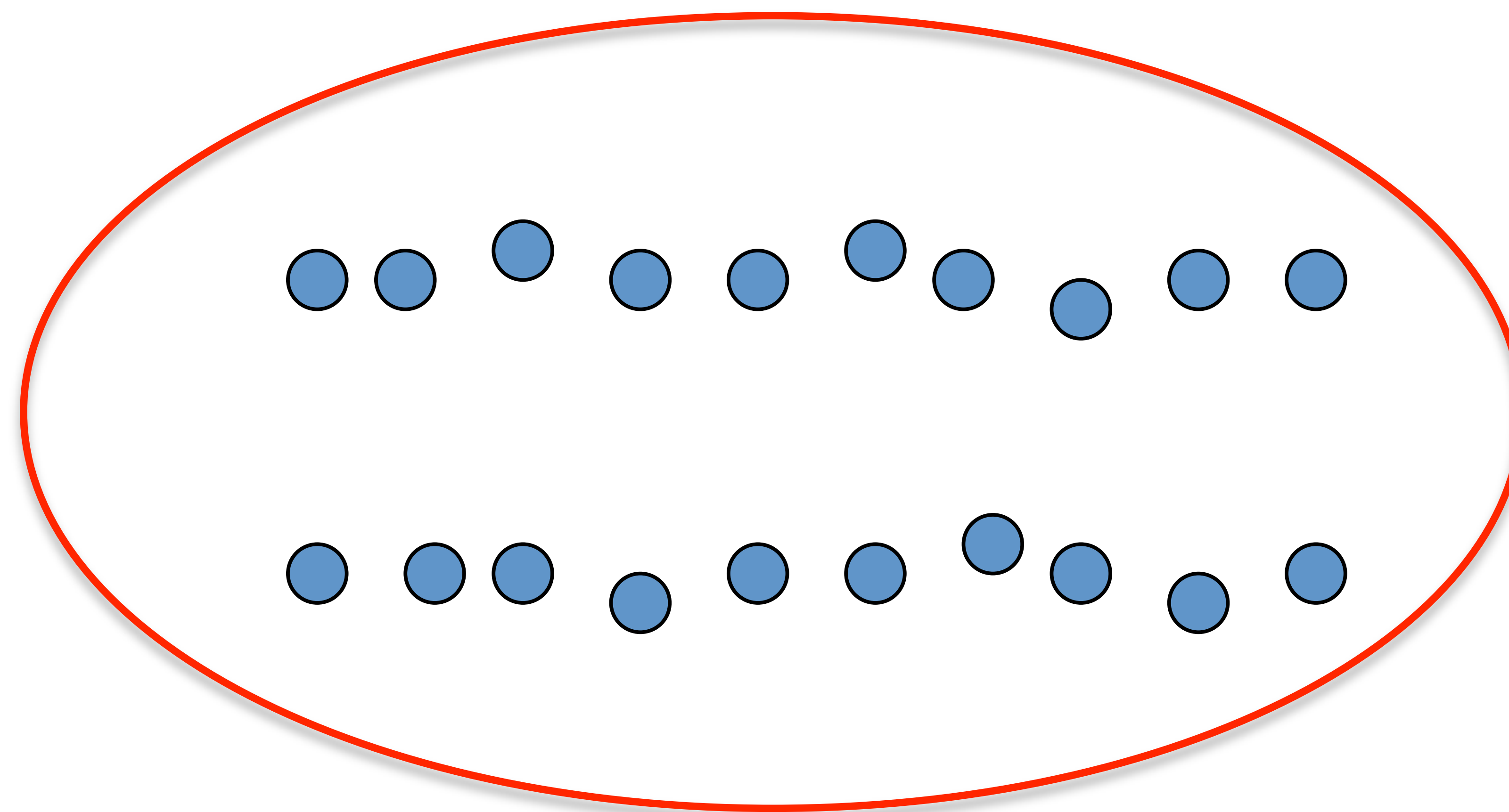
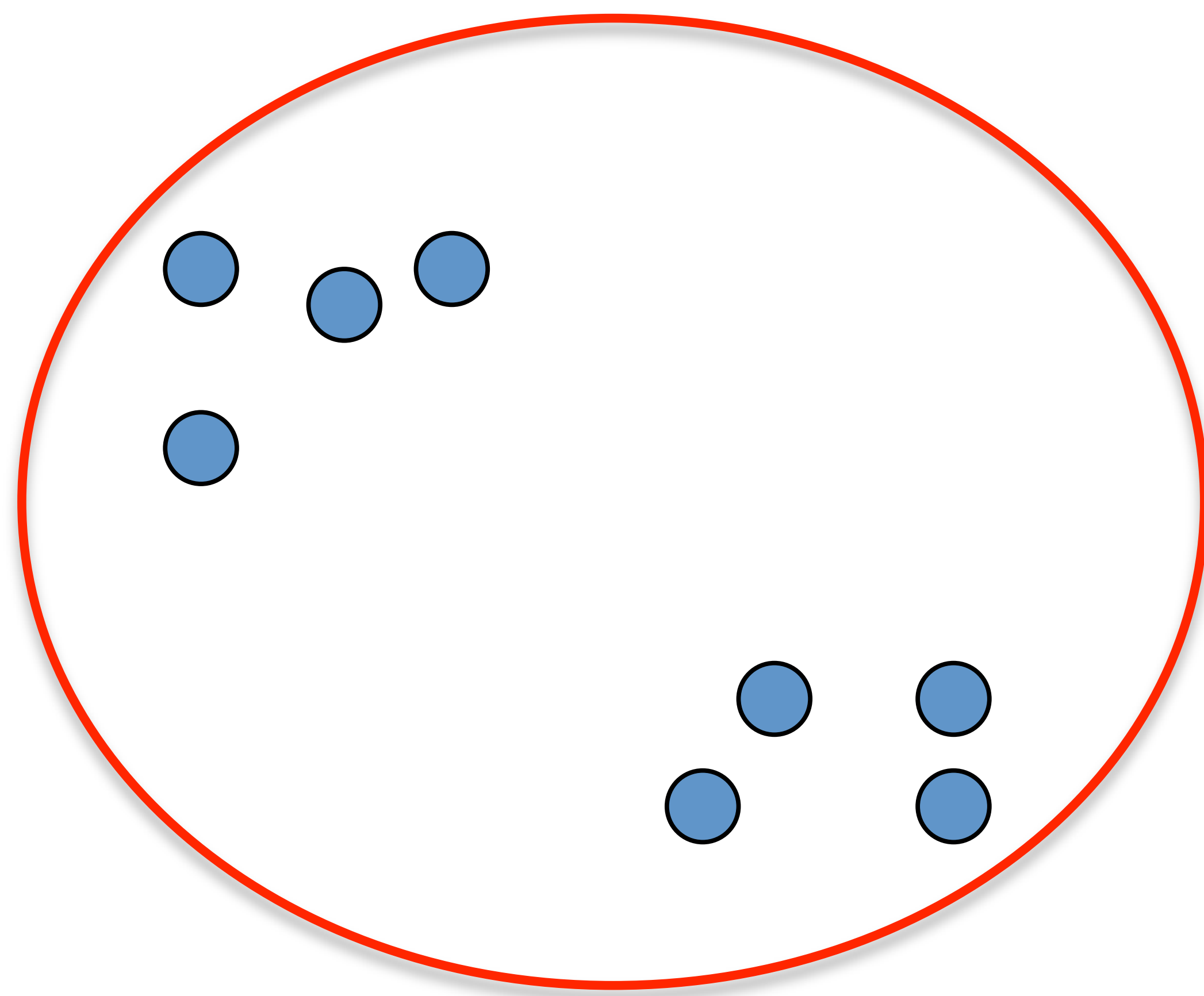
Clustering

- **Basic idea:** group together similar instances
- **Example:** 2D point patterns



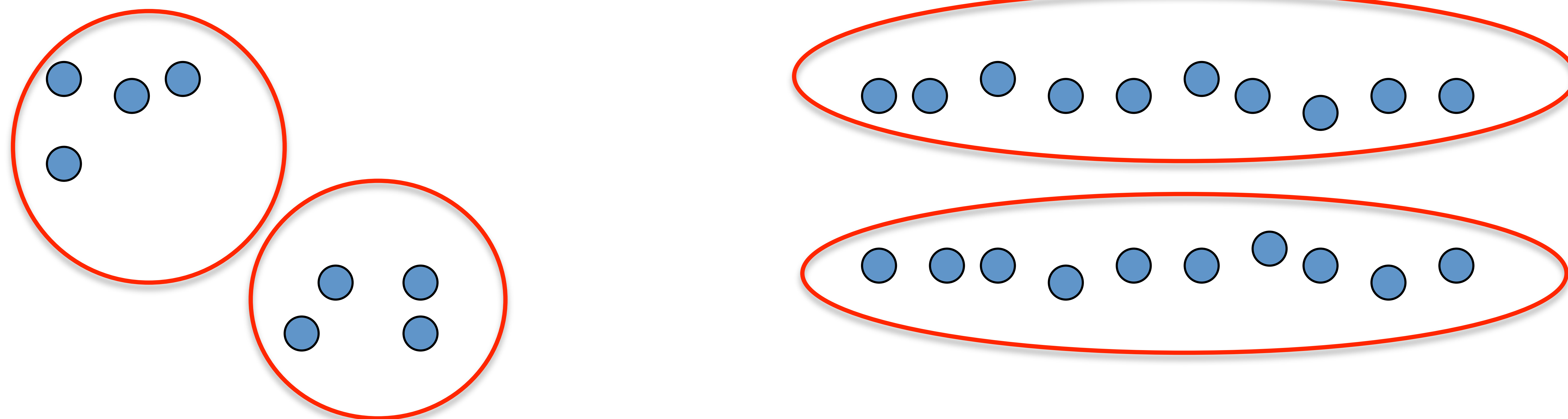
Clustering

- **Basic idea:** group together similar instances
- **Example:** 2D point patterns



Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns

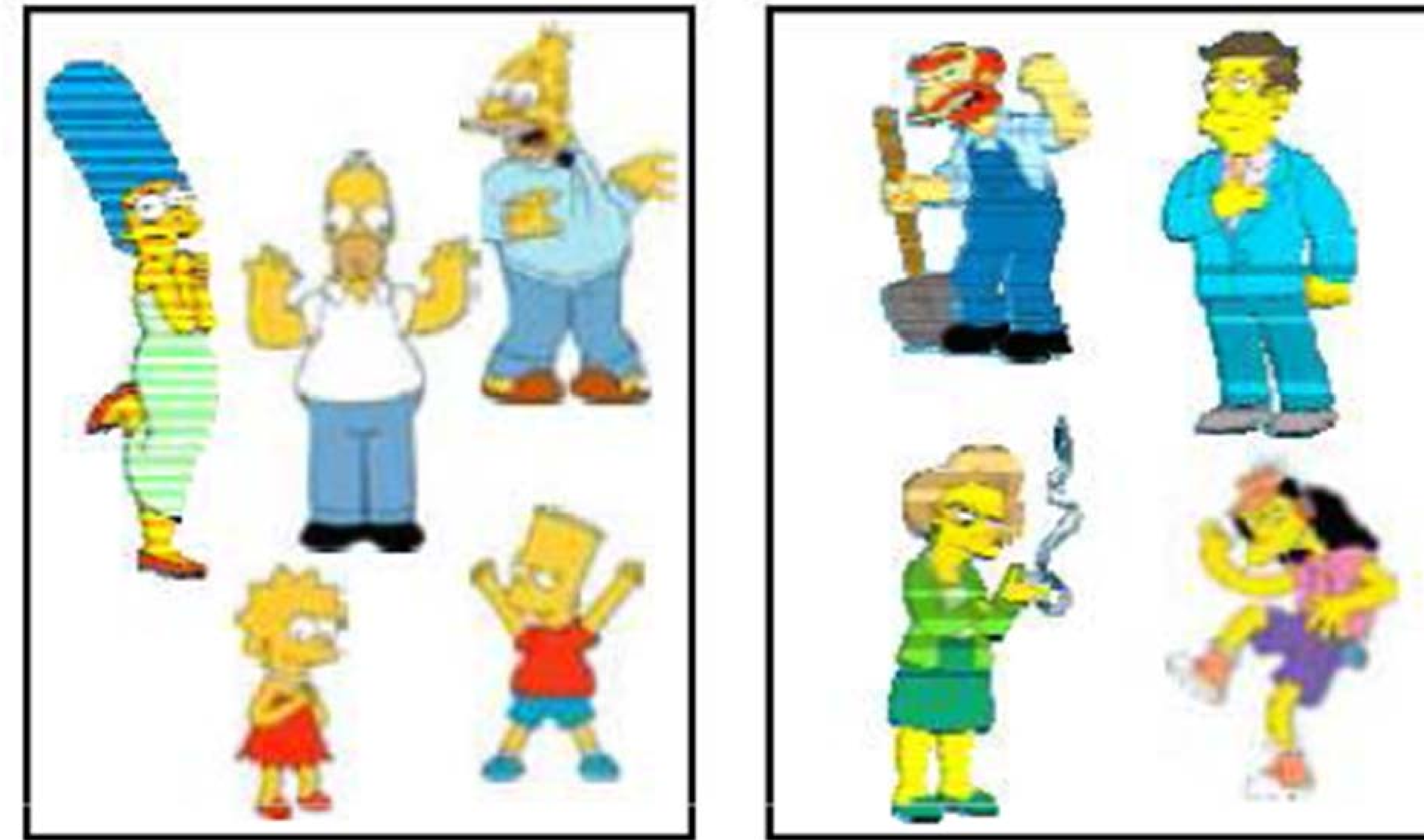


- What could “similar” mean?
 - One option: small Euclidean distance (squared)
$$\text{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$
 - Clustering results are crucially dependent on the measure of similarity (or distance) between “points” to be clustered

Clustering algorithms

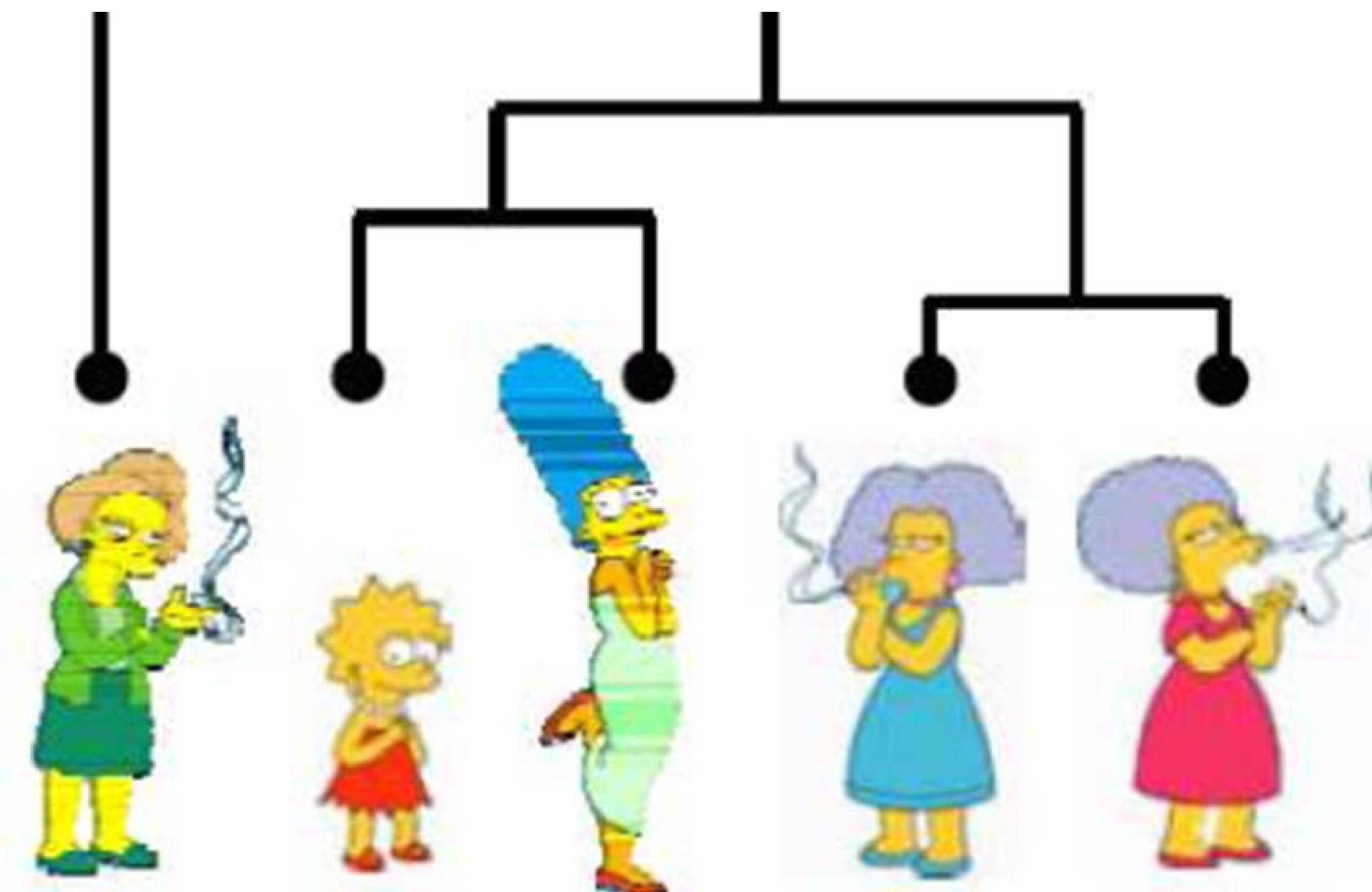
- Partition algorithms (Flat)

- K-means
- Mixture of Gaussian 混合高斯
- Spectral Clustering 谱聚类



- Hierarchical algorithms

- Bottom up – agglomerative 凝聚的
- Top down – divisive



Clustering examples

Image segmentation

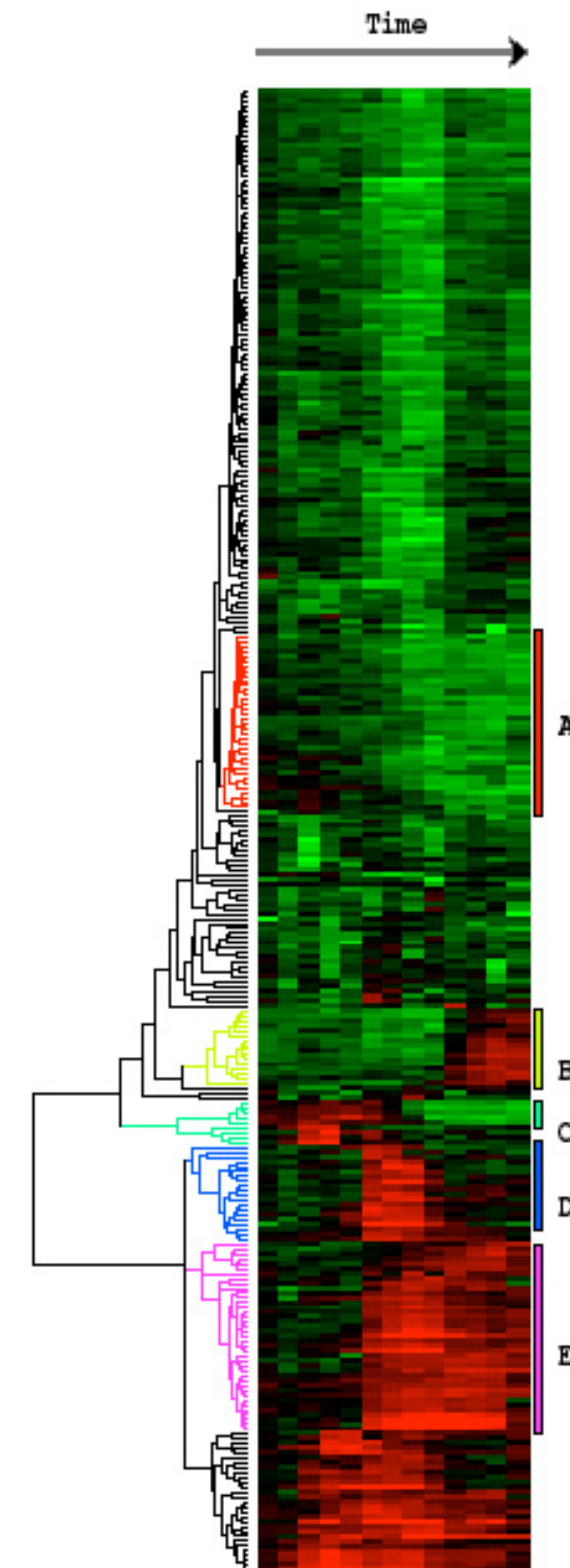
Goal: Break up the image into meaningful or perceptually similar regions



[Slide from James Hayes]

Clustering examples

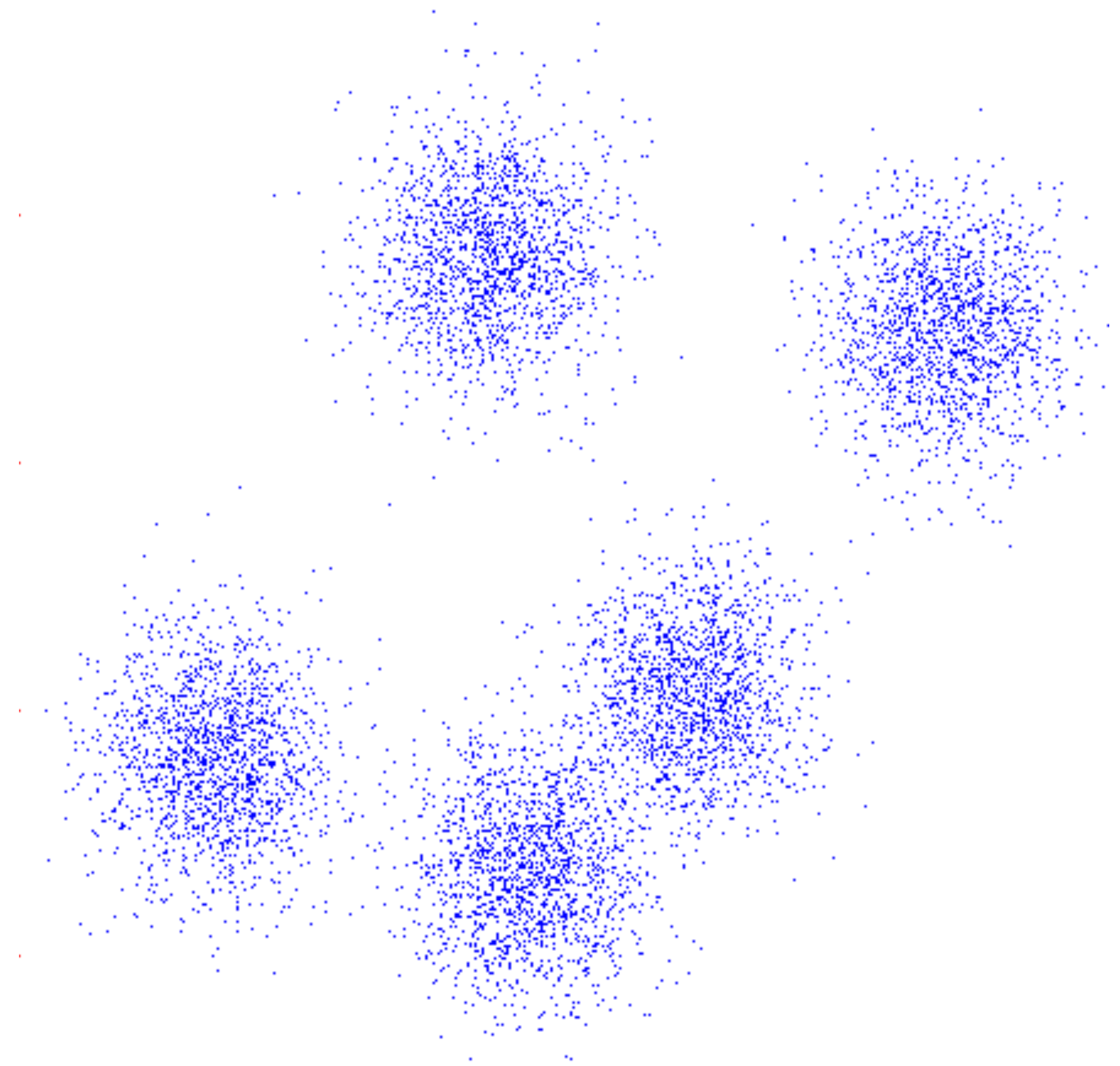
Clustering gene expression data



Eisen et al, PNAS 1998

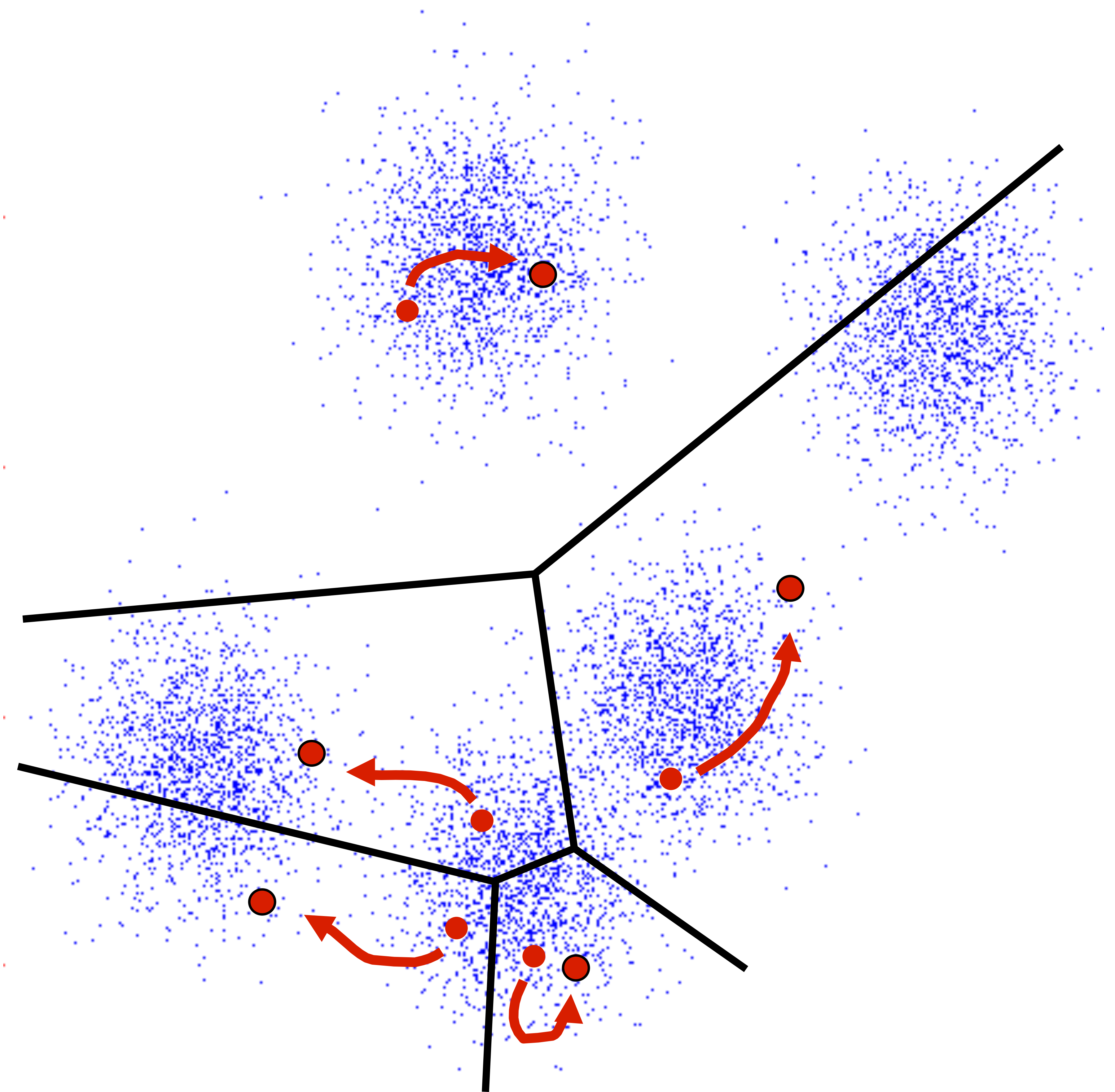
K-Means

- An iterative clustering algorithm
 - **Initialize:** Pick K random points as cluster centers
 - **Alternate:**
 1. Assign data points to closest cluster center
 2. Change the cluster center to the average of its assigned points
 - **Stop** when no points' assignments change

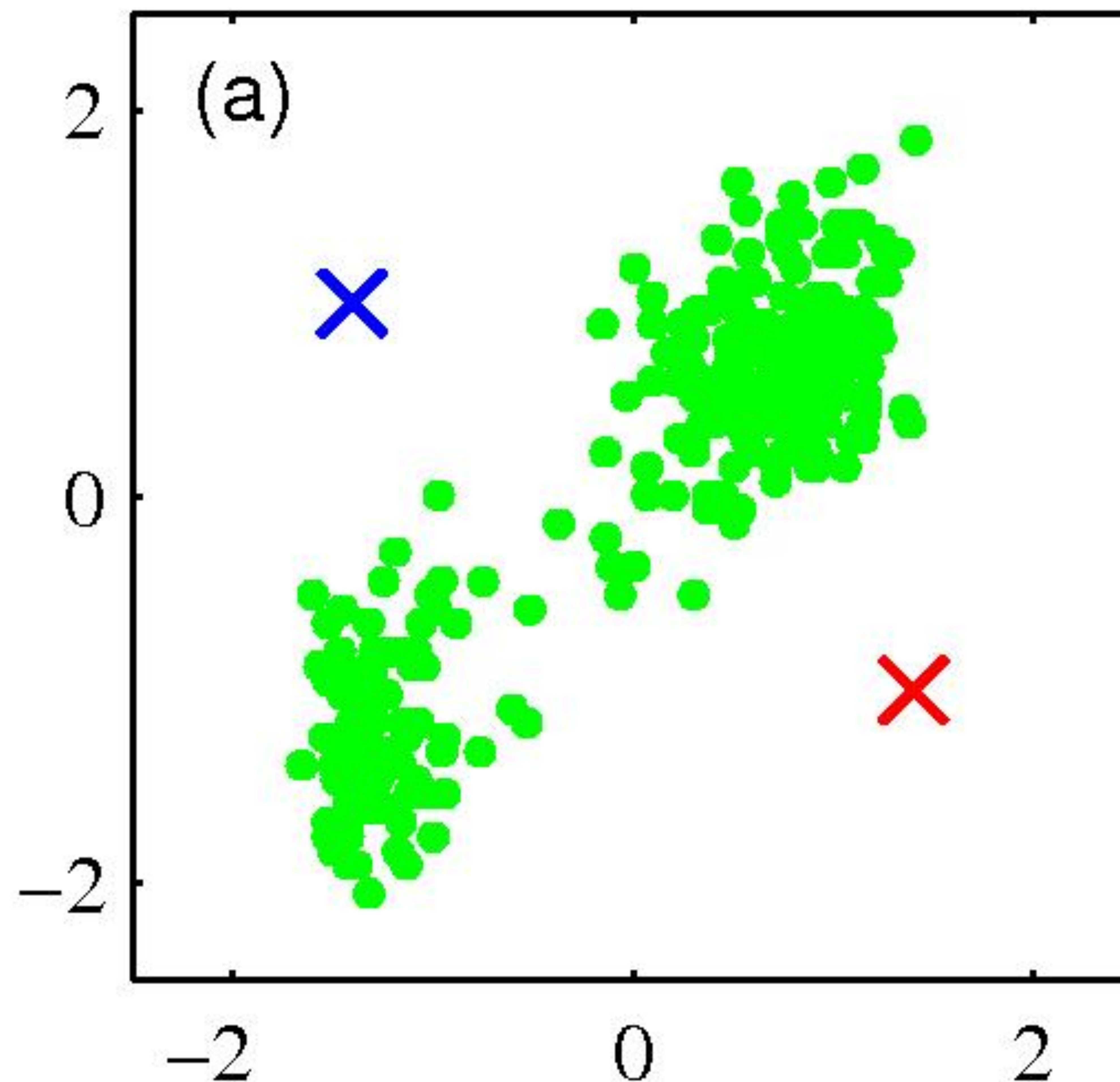


K-Means

- An iterative clustering algorithm
 - **Initialize:** Pick K random points as cluster centers
 - **Alternate:**
 1. Assign data points to closest cluster center
 2. Change the cluster center to the average of its assigned points
 - **Stop** when no points' assignments change



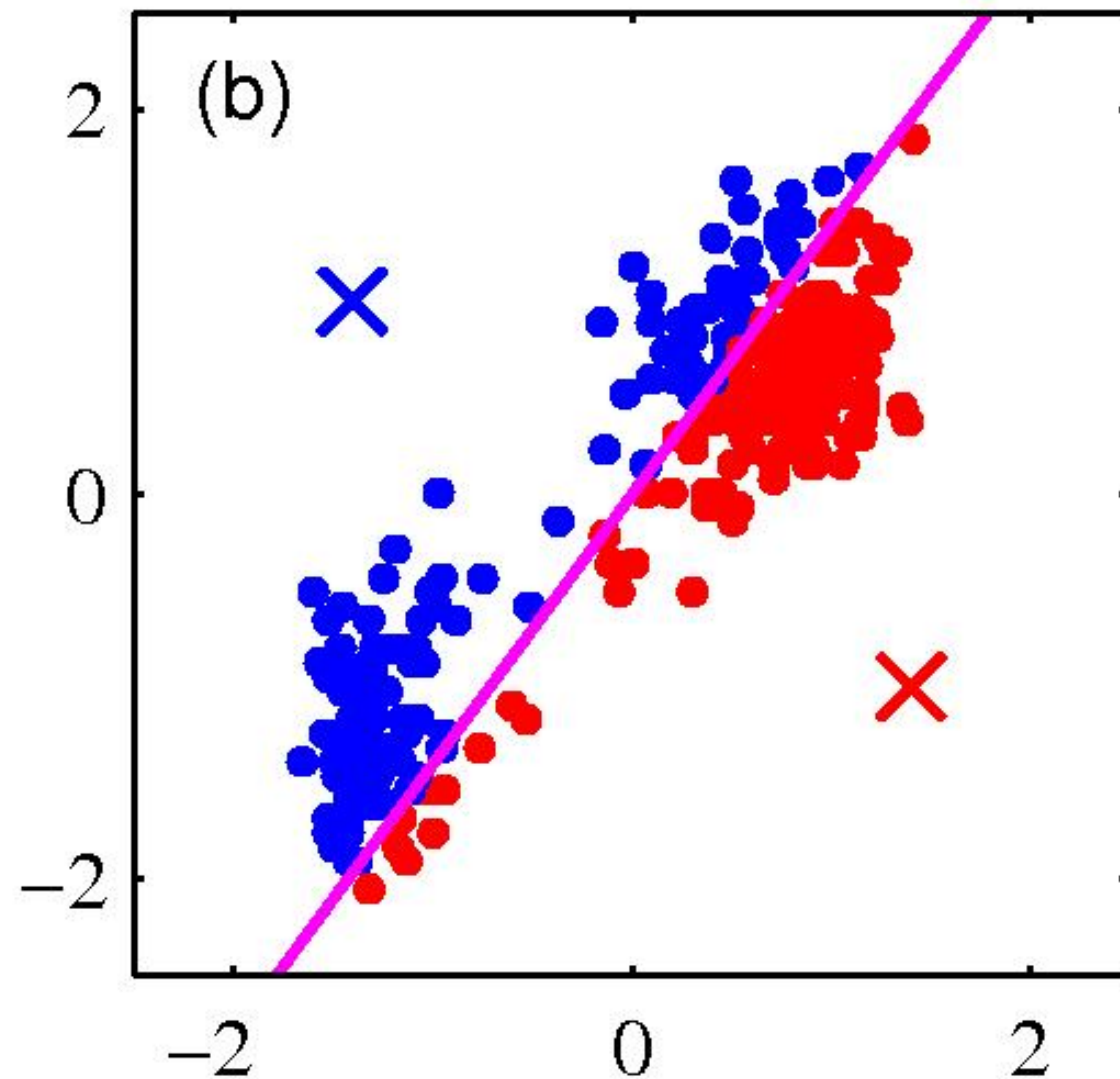
K-means clustering: Example



- Pick K random points as cluster centers (means)

Shown here for $K=2$

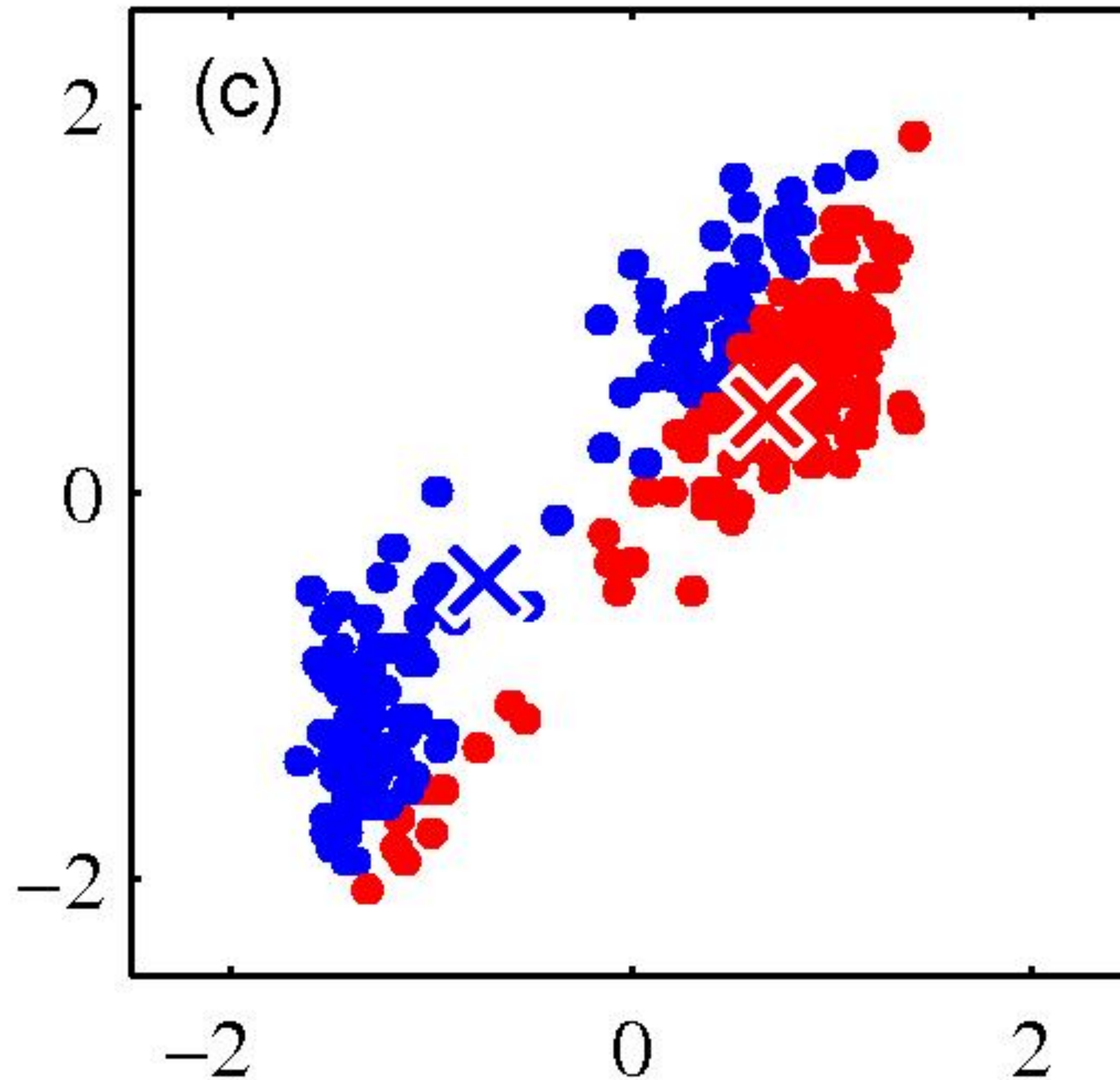
K-means clustering: Example



Iterative Step 1

- Assign data points to closest cluster center

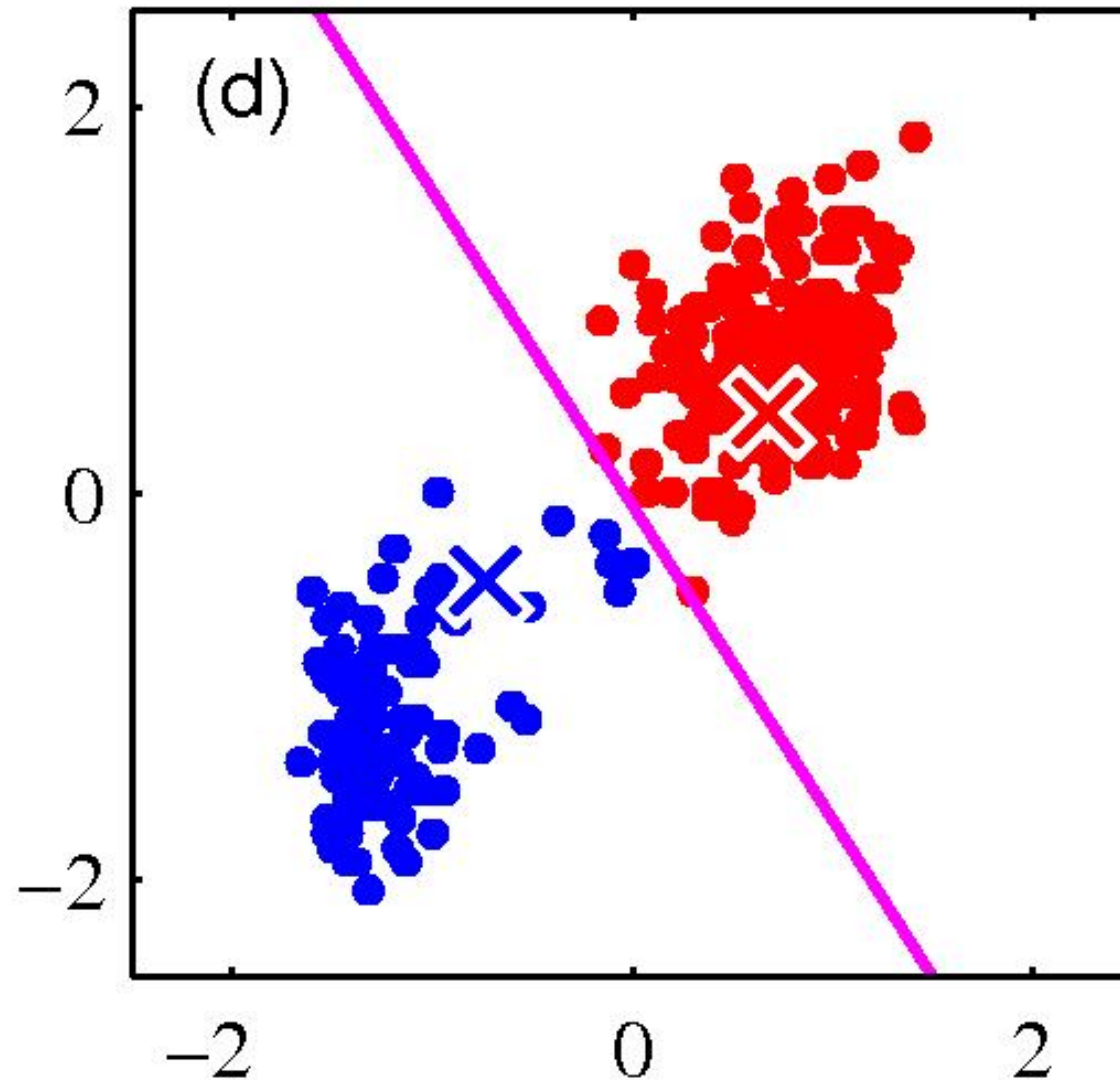
K-means clustering: Example



Iterative Step 2

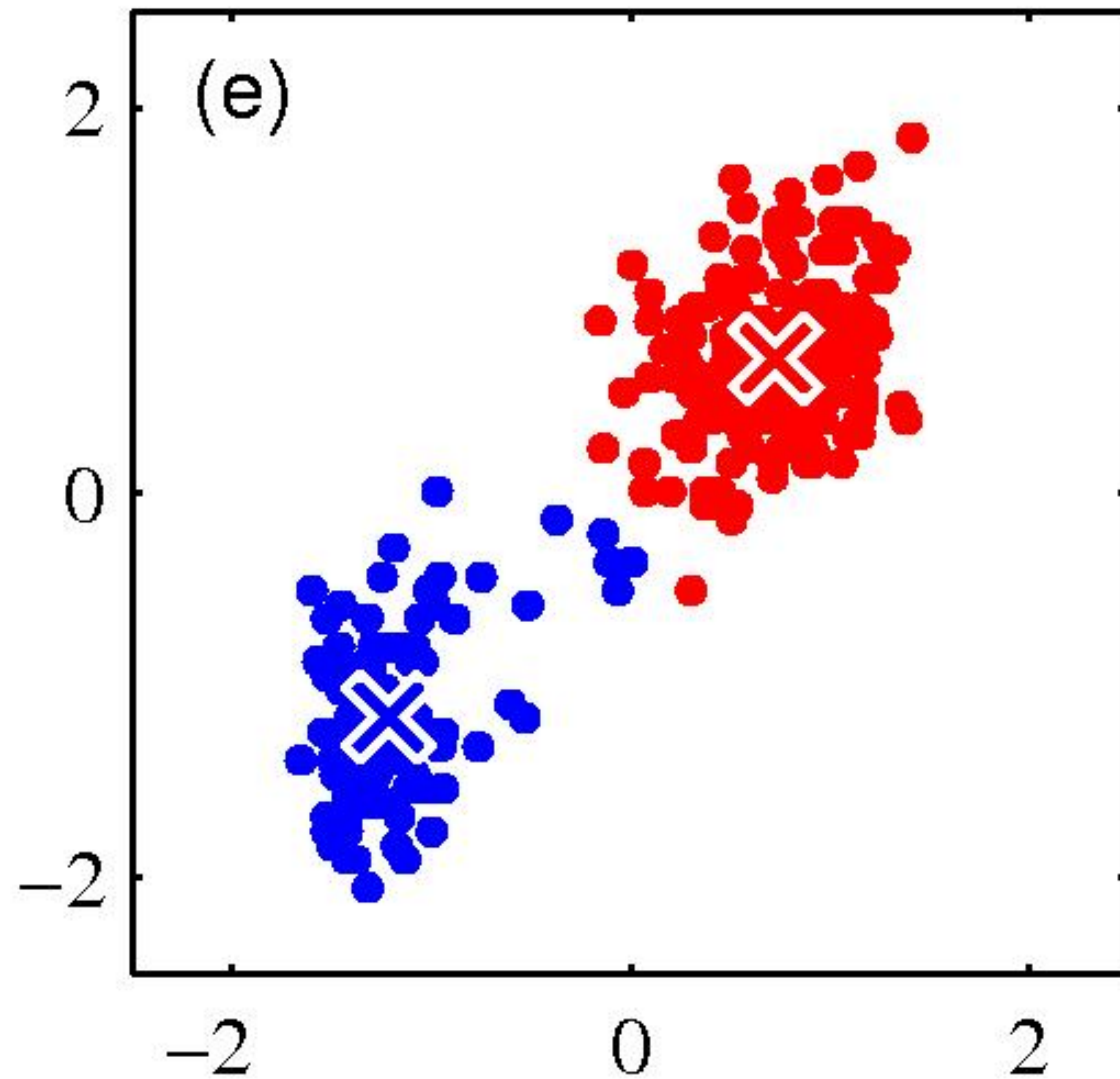
- Change the cluster center to the average of the assigned points

K-means clustering: Example

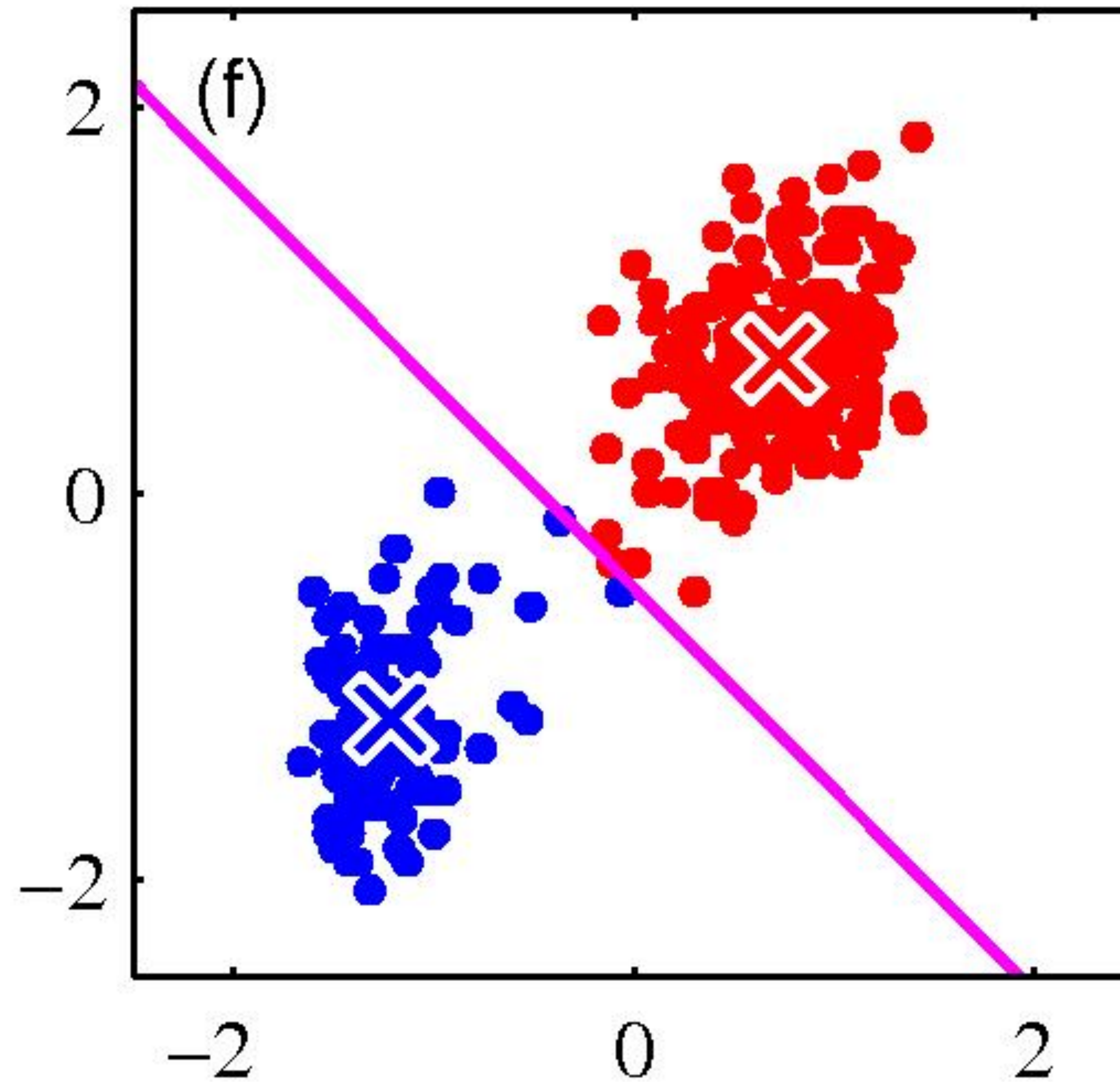


- Repeat until convergence

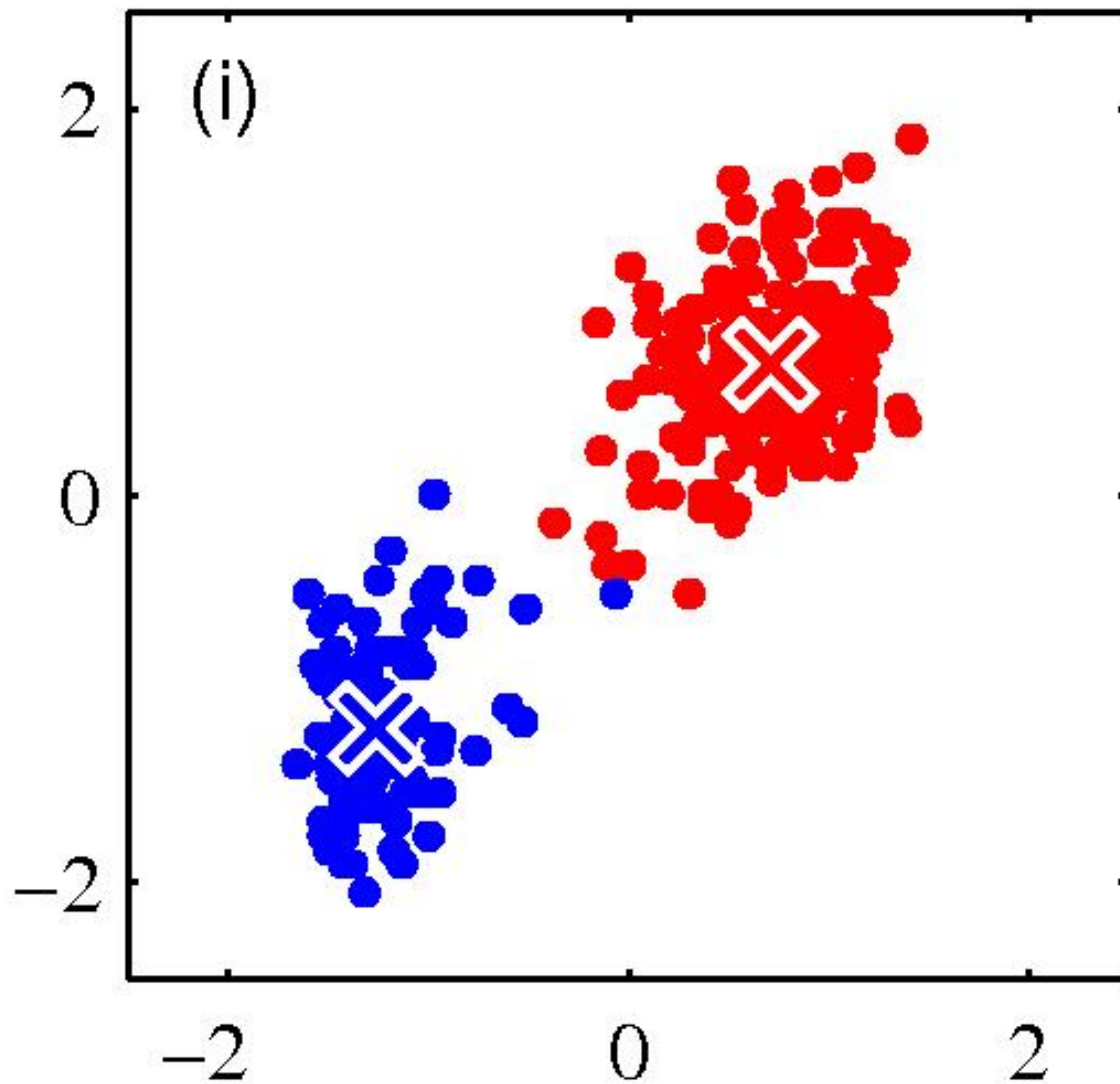
K-means clustering: Example



K-means clustering: Example



K-means clustering: Example



Properties of K-means **algorithm**

- Guaranteed to converge in a finite number of iterations
- Running time per iteration:
 1. Assign data points to closest cluster center
 $O(KN)$ time
 2. Change the cluster center to the average of its assigned points
 $O(N)$

What properties should a distance measure have?

- Symmetric
 - $D(A,B)=D(B,A)$
 - Otherwise, we can say A looks like B but B does not look like A
- Positivity, and self-similarity
 - $D(A,B) \geq 0$, and $D(A,B)=0$ iff $A=B$
 - Otherwise there will different objects that we cannot tell apart
- Triangle inequality
 - $D(A,B)+D(B,C) \geq D(A,C)$
 - Otherwise one can say “A is like B, B is like C, but A is not like C at all”