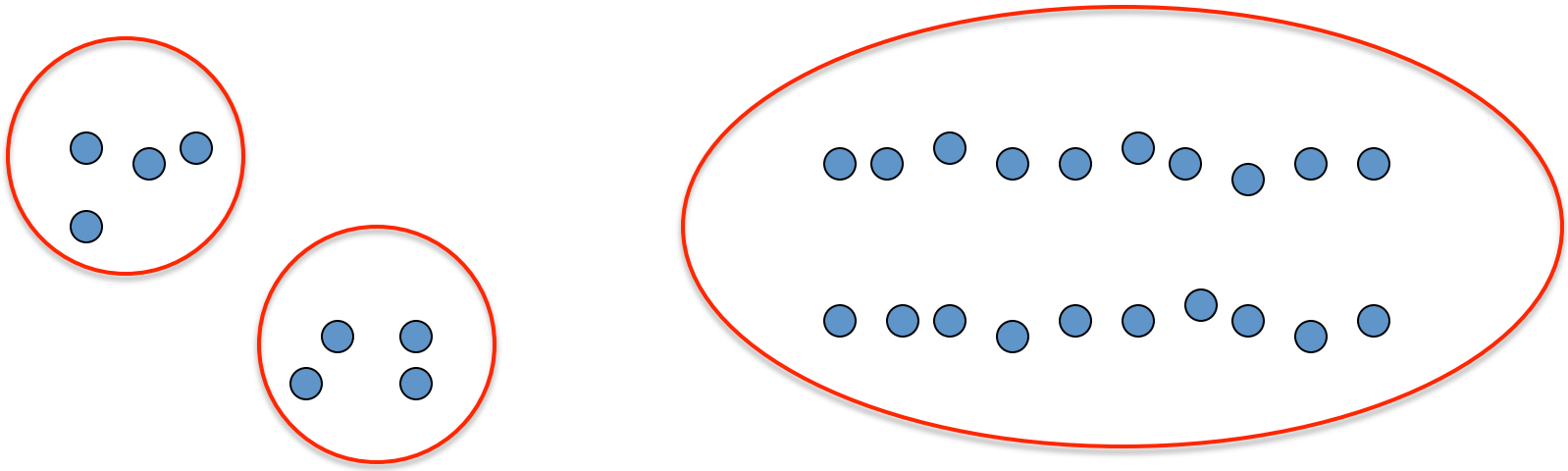
# Clustering

## 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

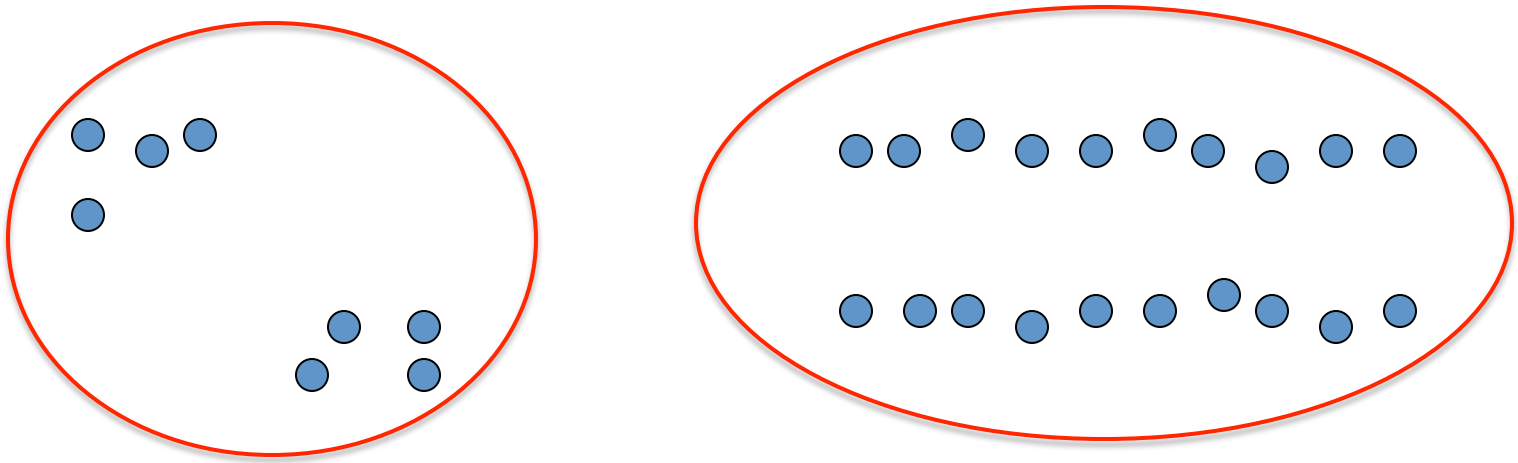
# Clustering

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



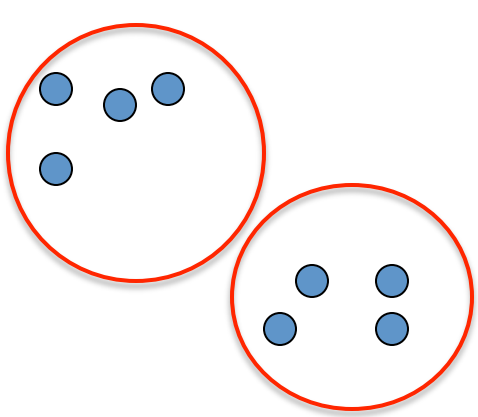
# Clustering

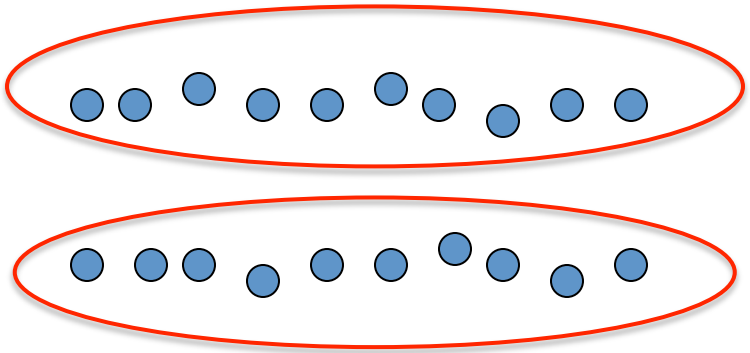
- Basic idea: group together similar instances
- Example: 2D point patterns



# Clustering

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





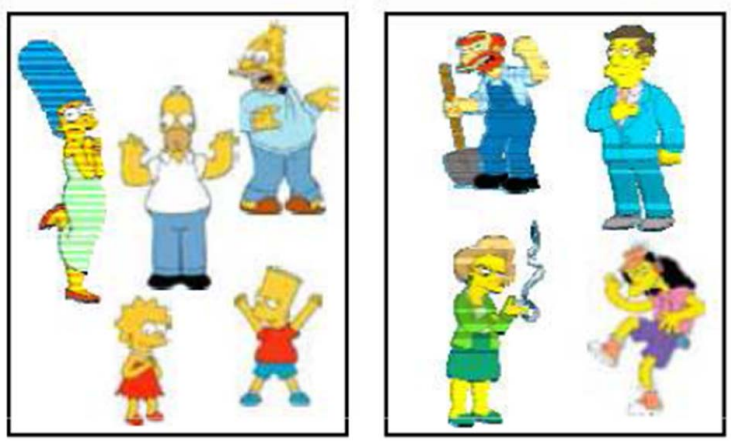
- What could "similar" mean?
  - One option: small Euclidean distance (squared)

$$\mathrm{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$

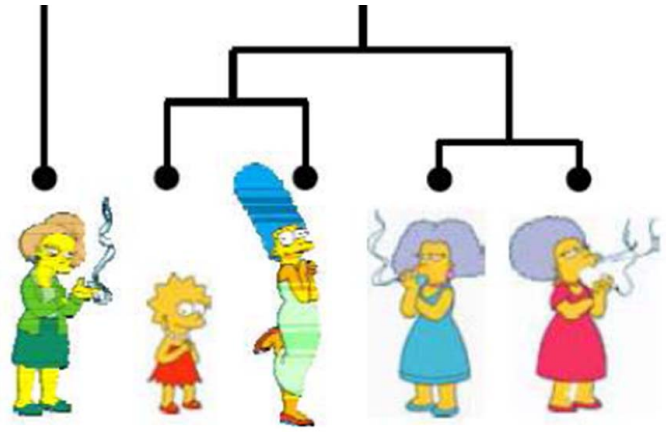
  - Clustering results are crucially dependent on the measure of similarity (or distance) between "points" to be clustered

# Clustering algorithms

- Partition algorithms (Flat)
  - K-means
  - Mixture of Gaussian
  - Spectral Clustering



- Hierarchical algorithms
  - Bottom up – agglomerative
  - Top down – divisive



# Clustering examples

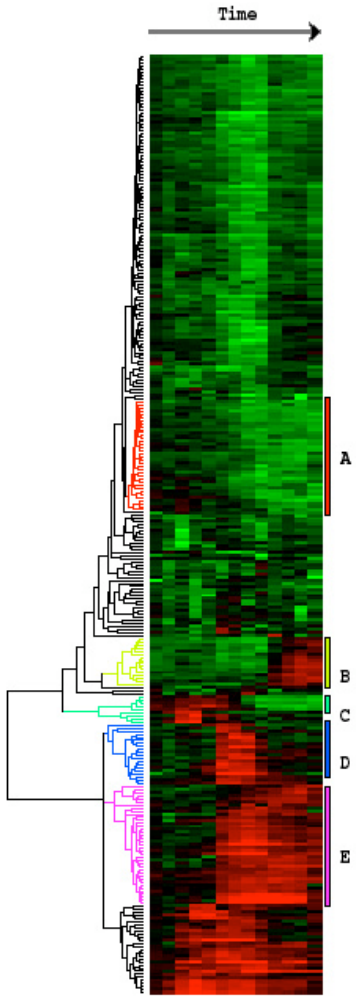
**Image segmentation**

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

[Slide from James Hayes]

# Clustering examples

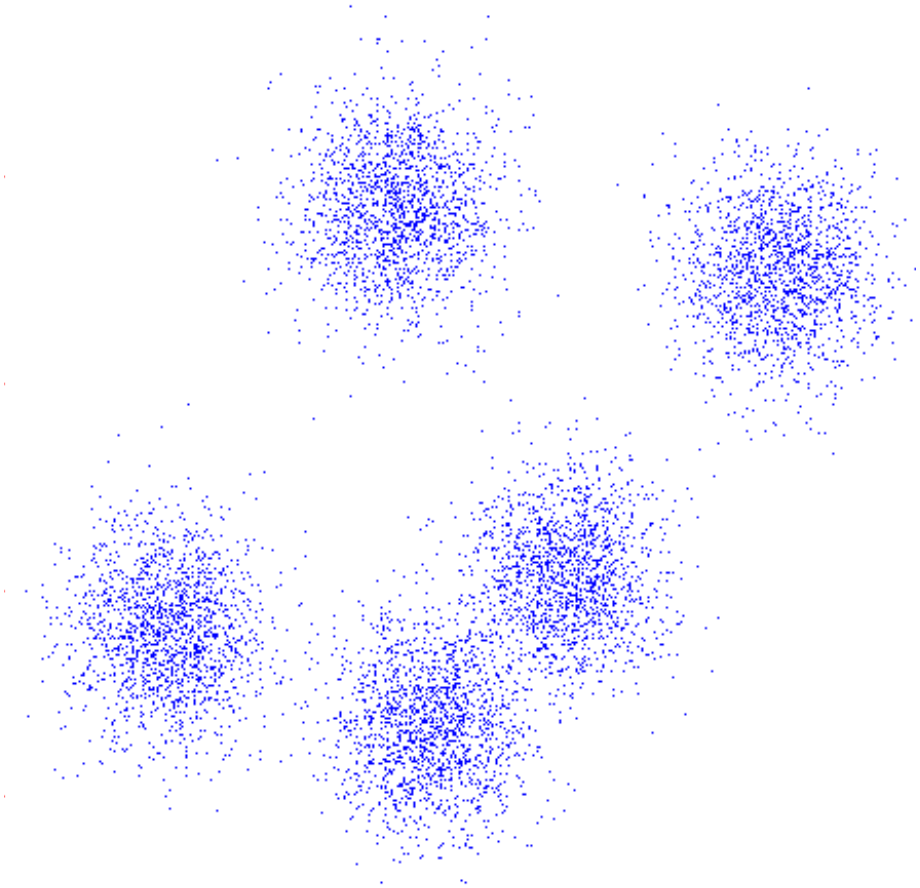
## Clustering gene expression data




Eisen et al, PNAS 1998

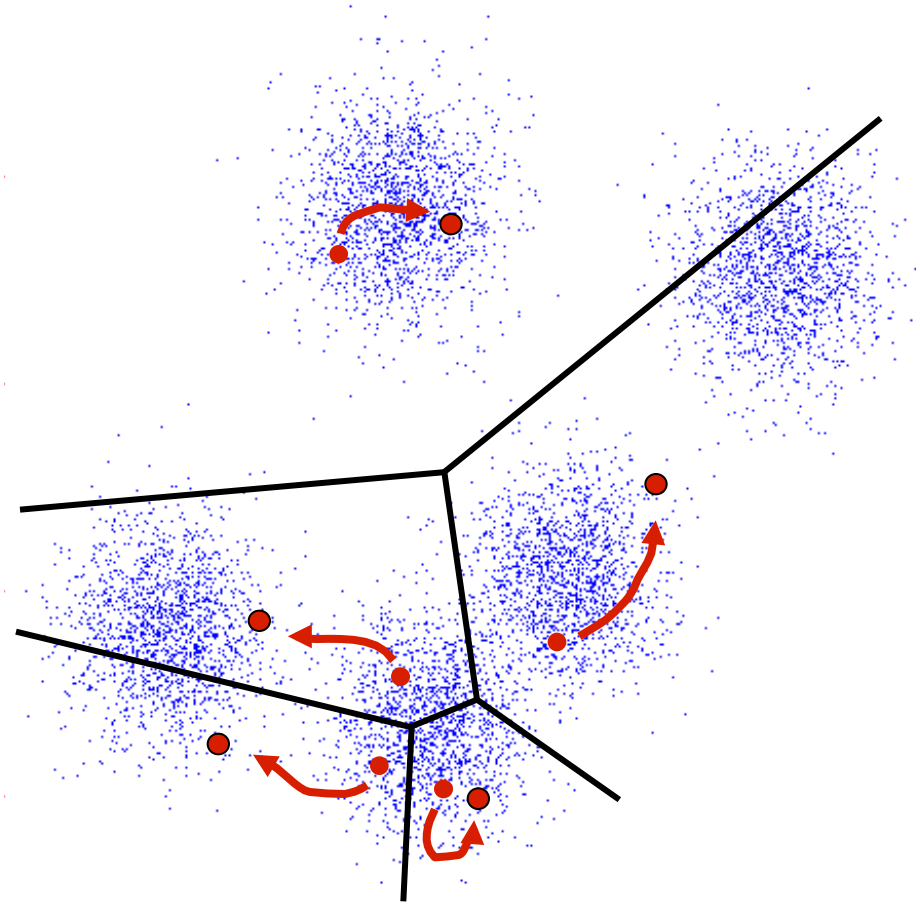
# K-Means

- 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

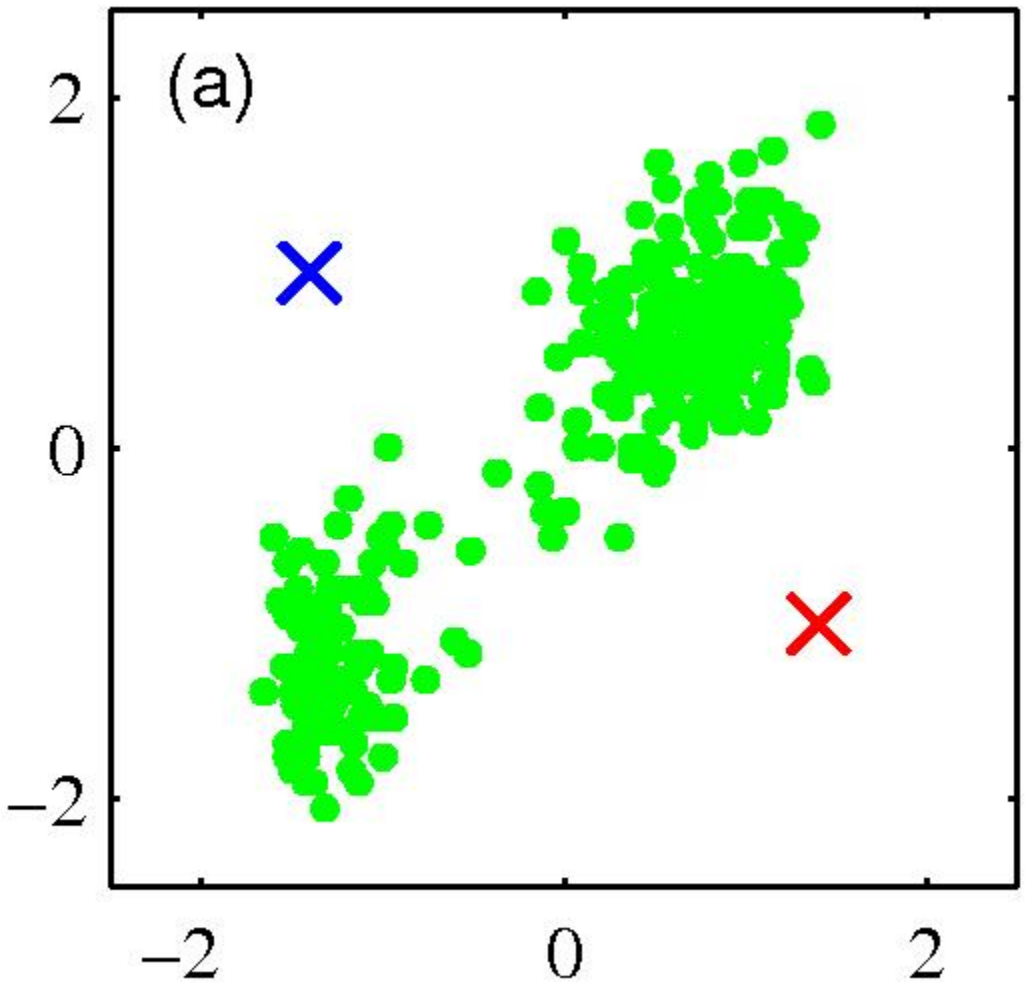


# K-Means

- 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



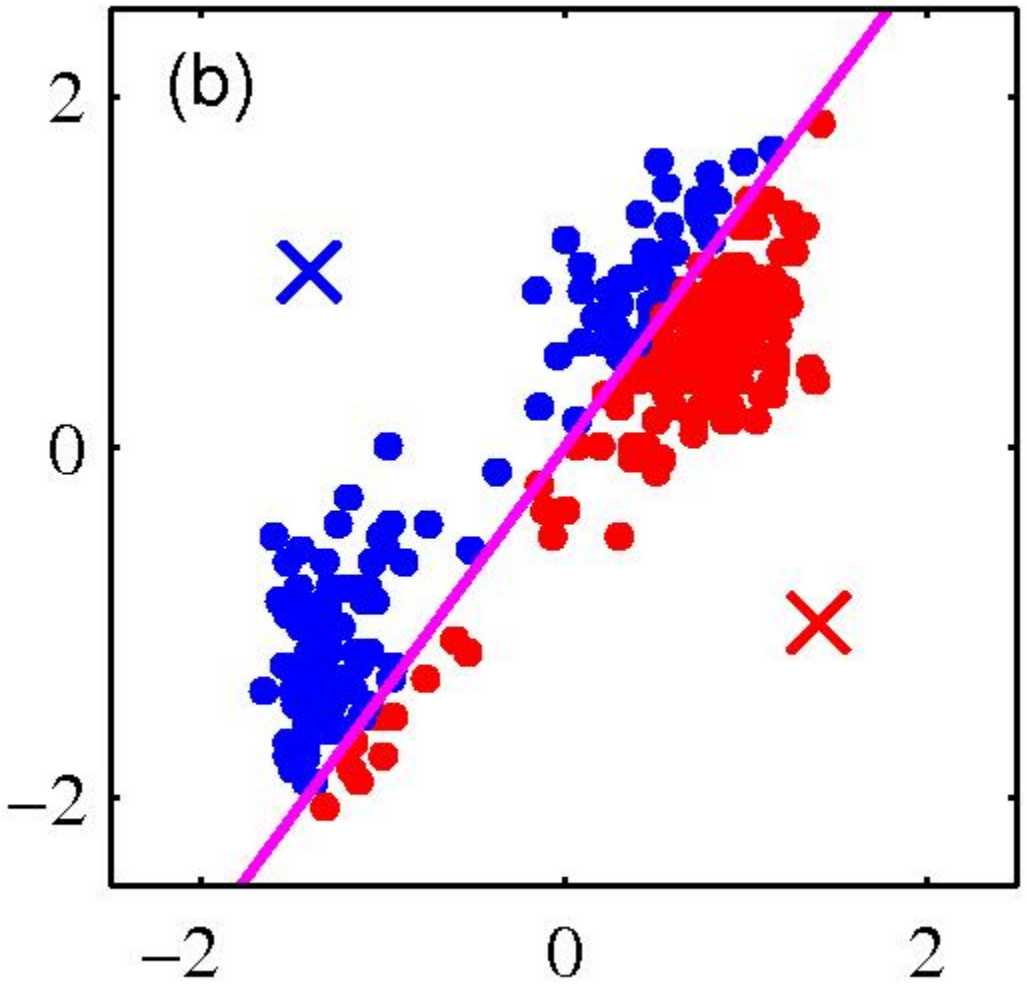
# K-means clustering: Example




- Pick *K* random points as cluster centers (means)

Shown here for *K*=2

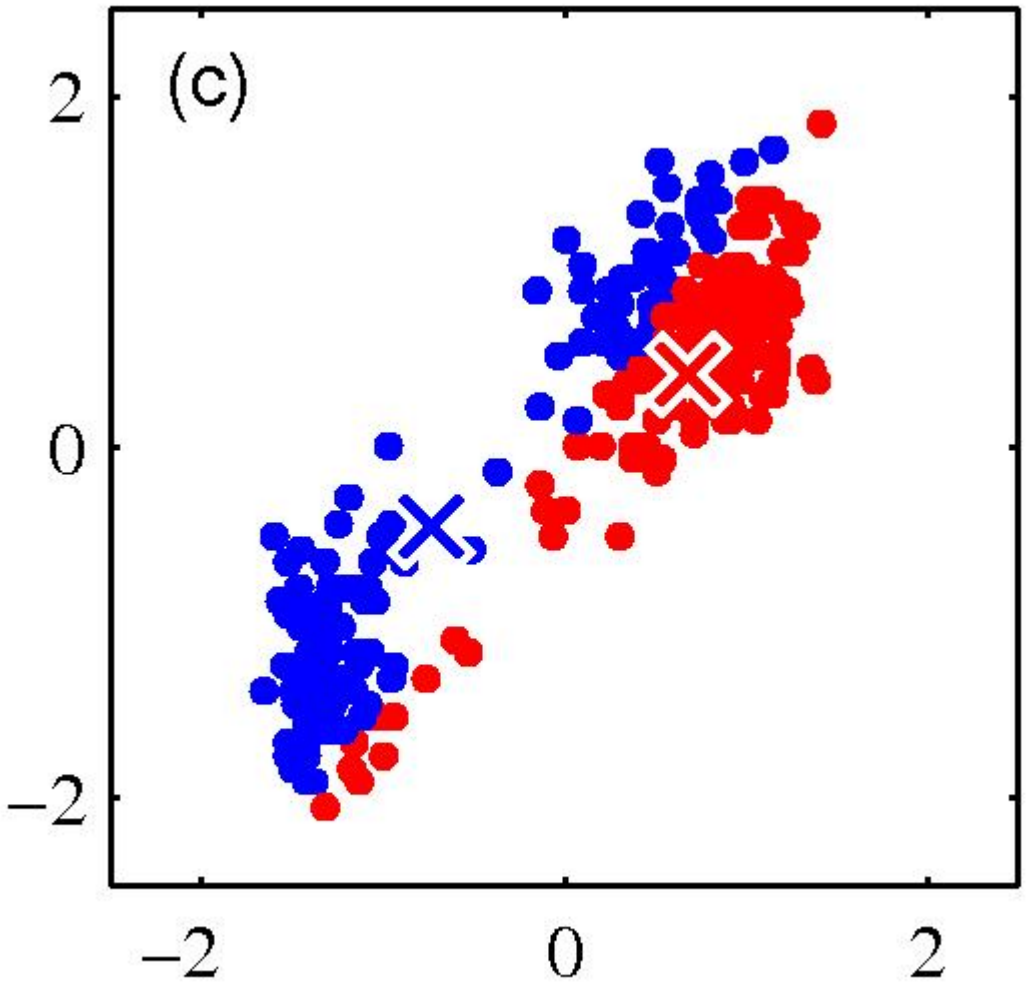
# K-means clustering: Example




Iterative Step 1

- Assign data points to closest cluster center

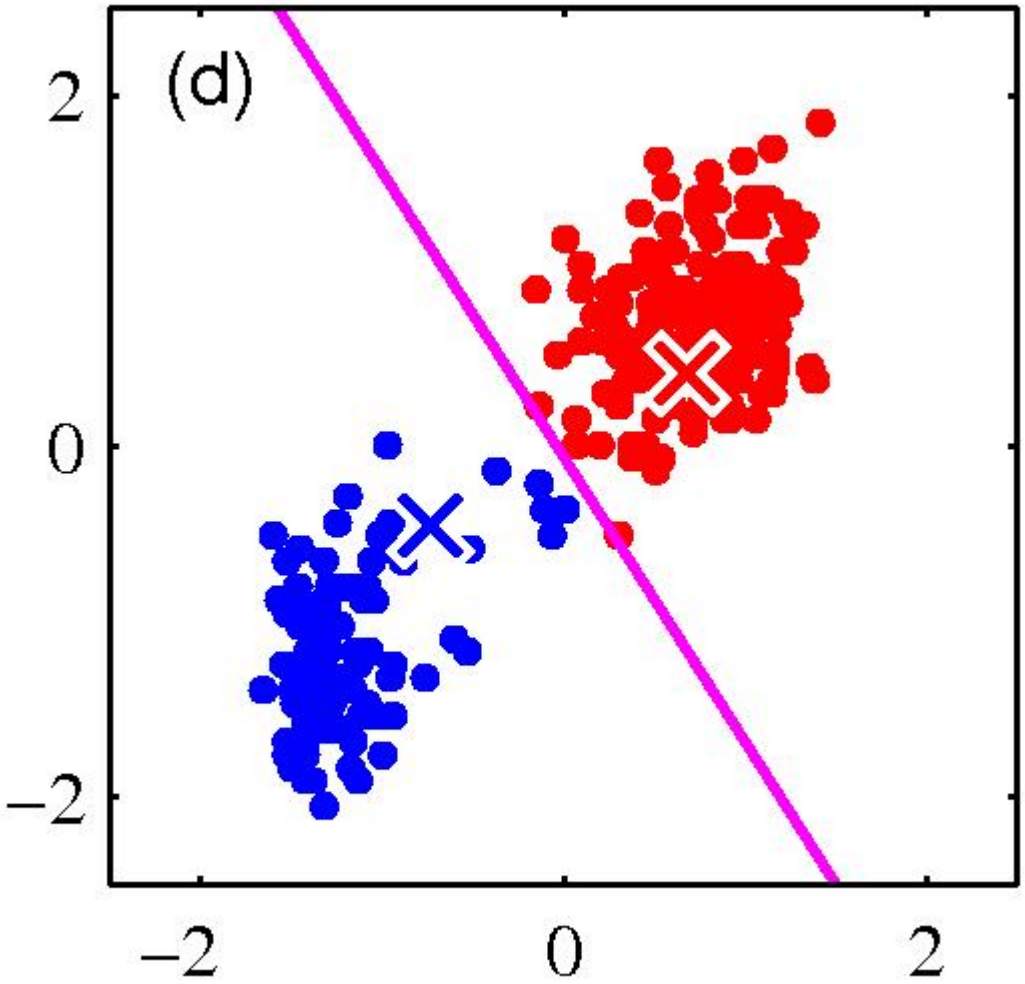
# K-means clustering: Example




Iterative Step 2

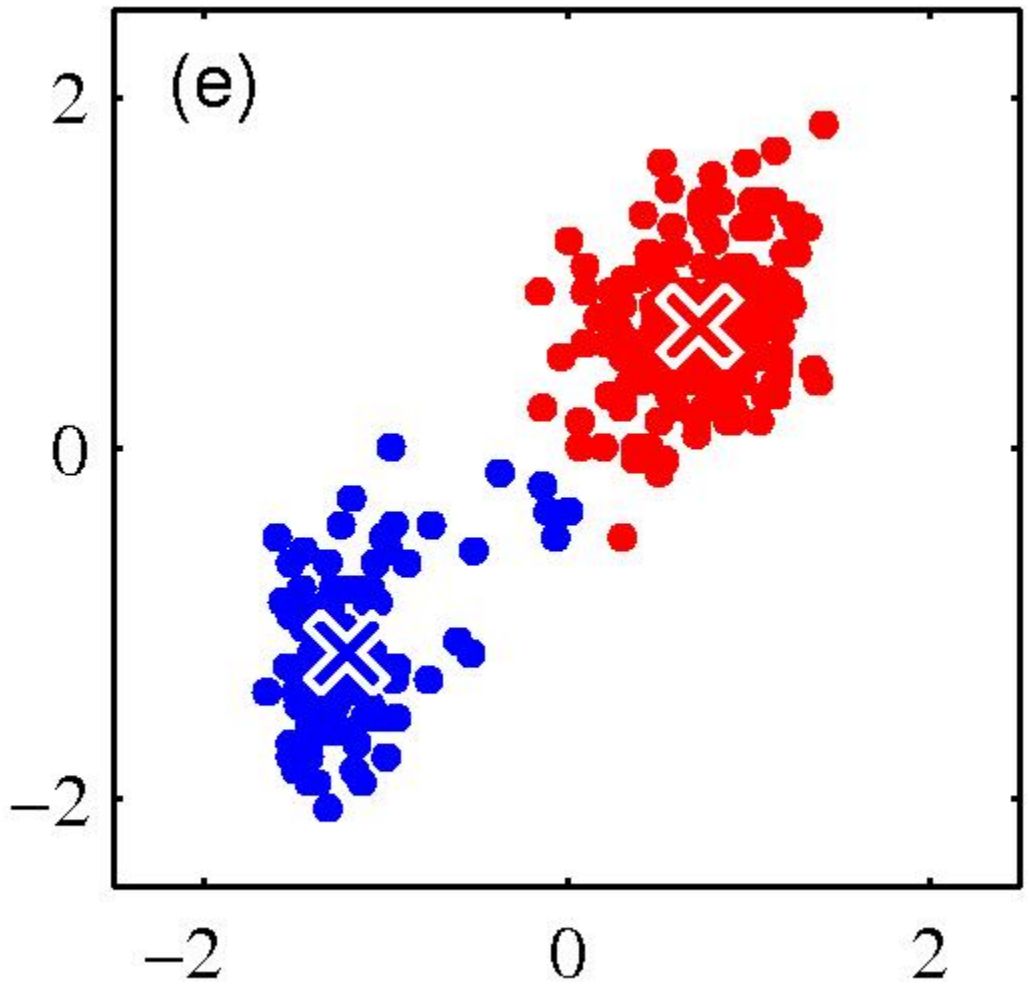
- Change the cluster center to the average of the assigned points

# K-means clustering: Example

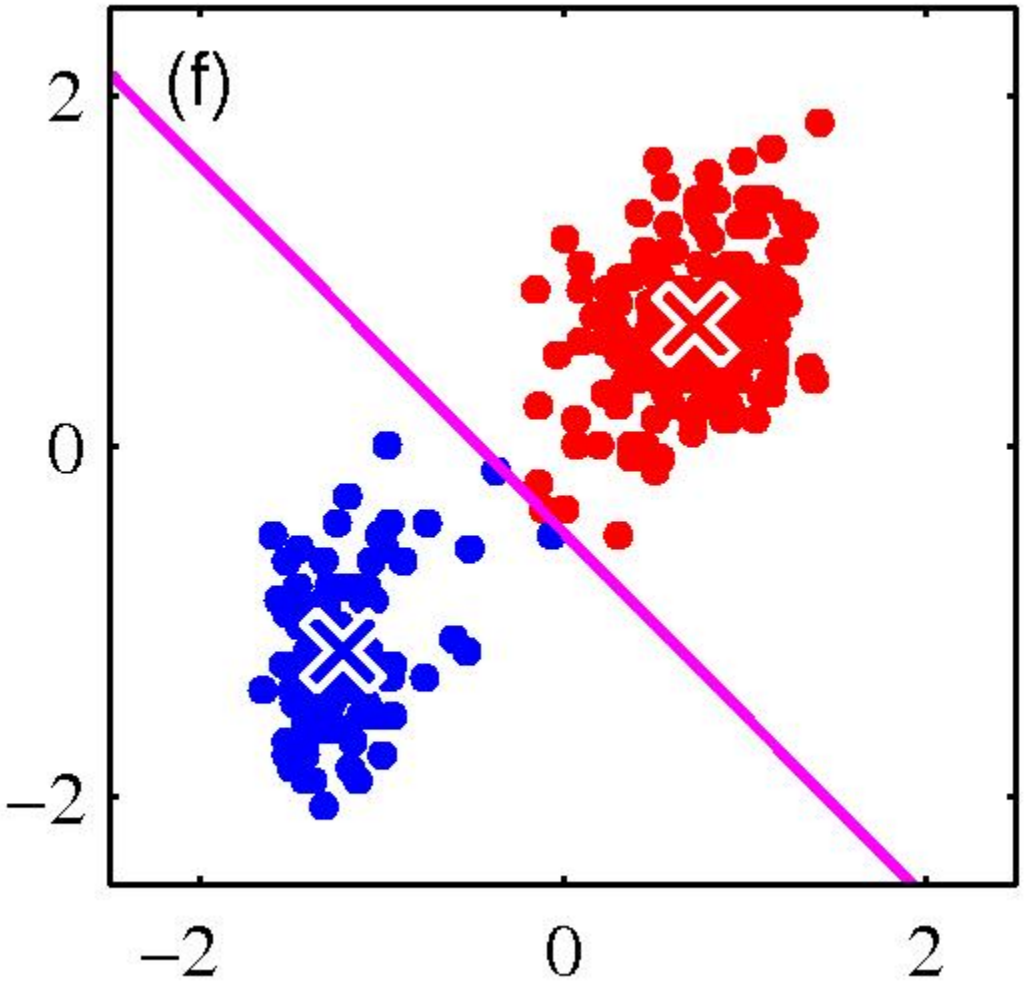



- Repeat until convergence

# K-means clustering: Example



# K-means clustering: Example



# K-means clustering: Example