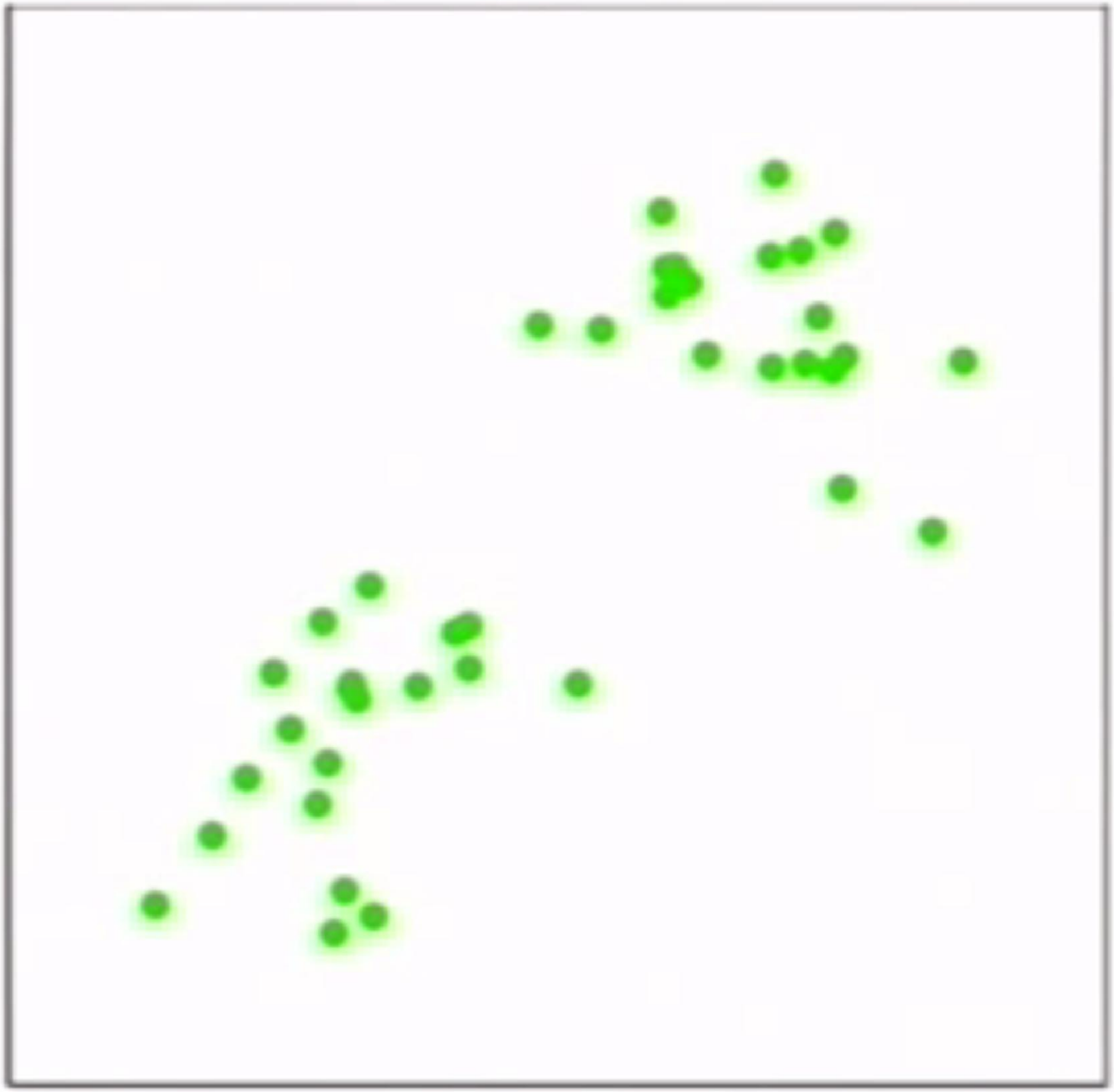
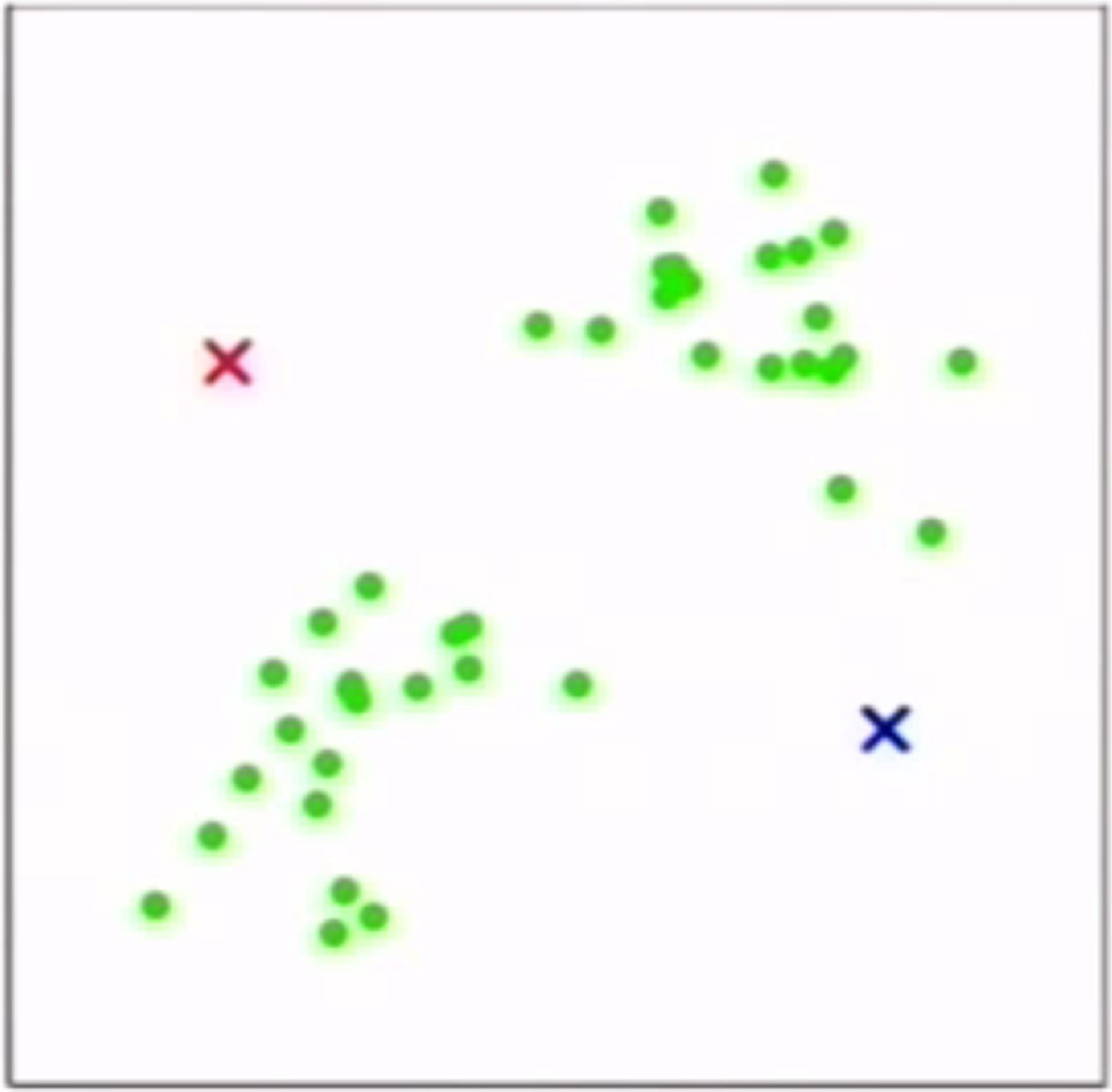
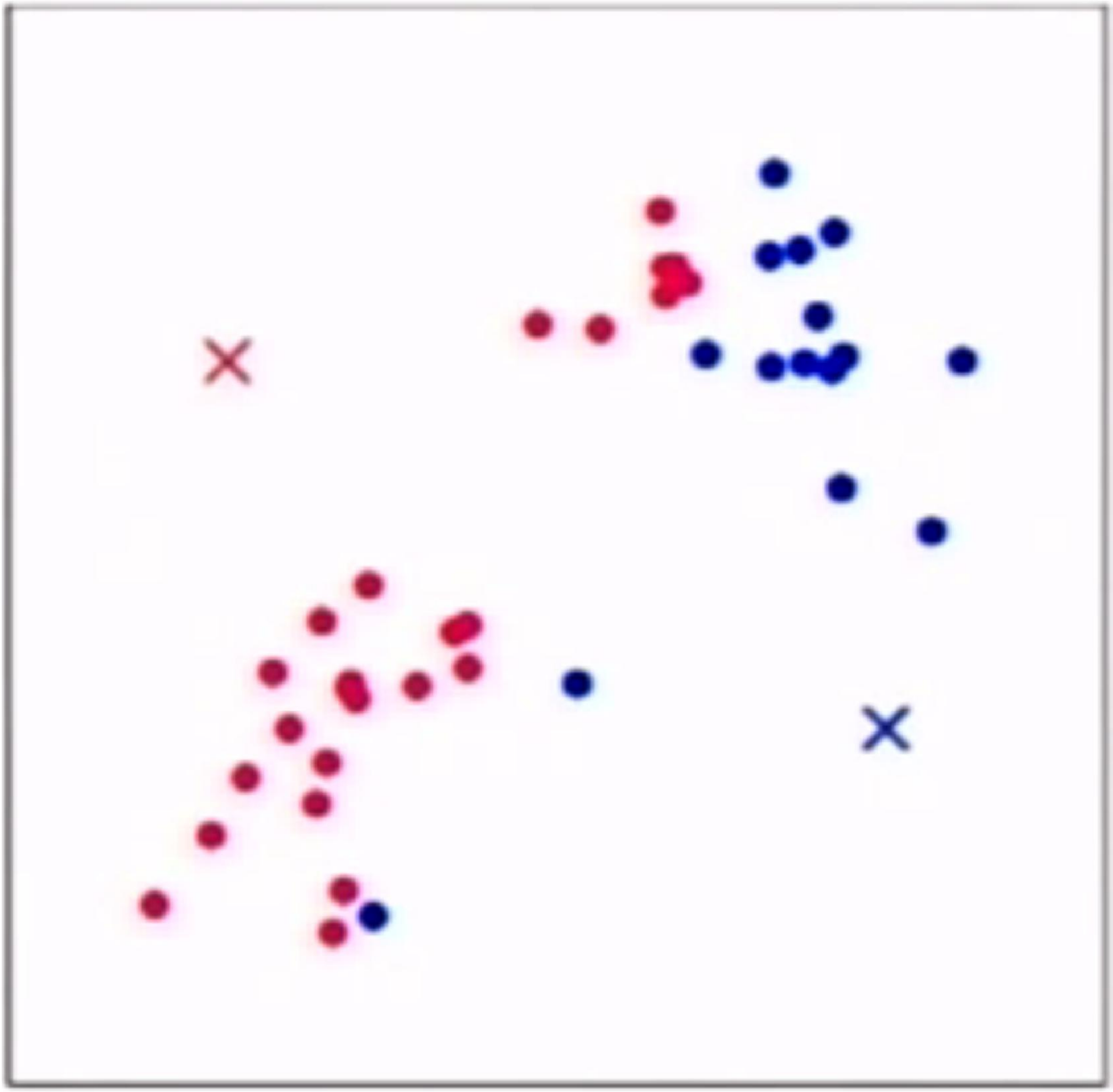
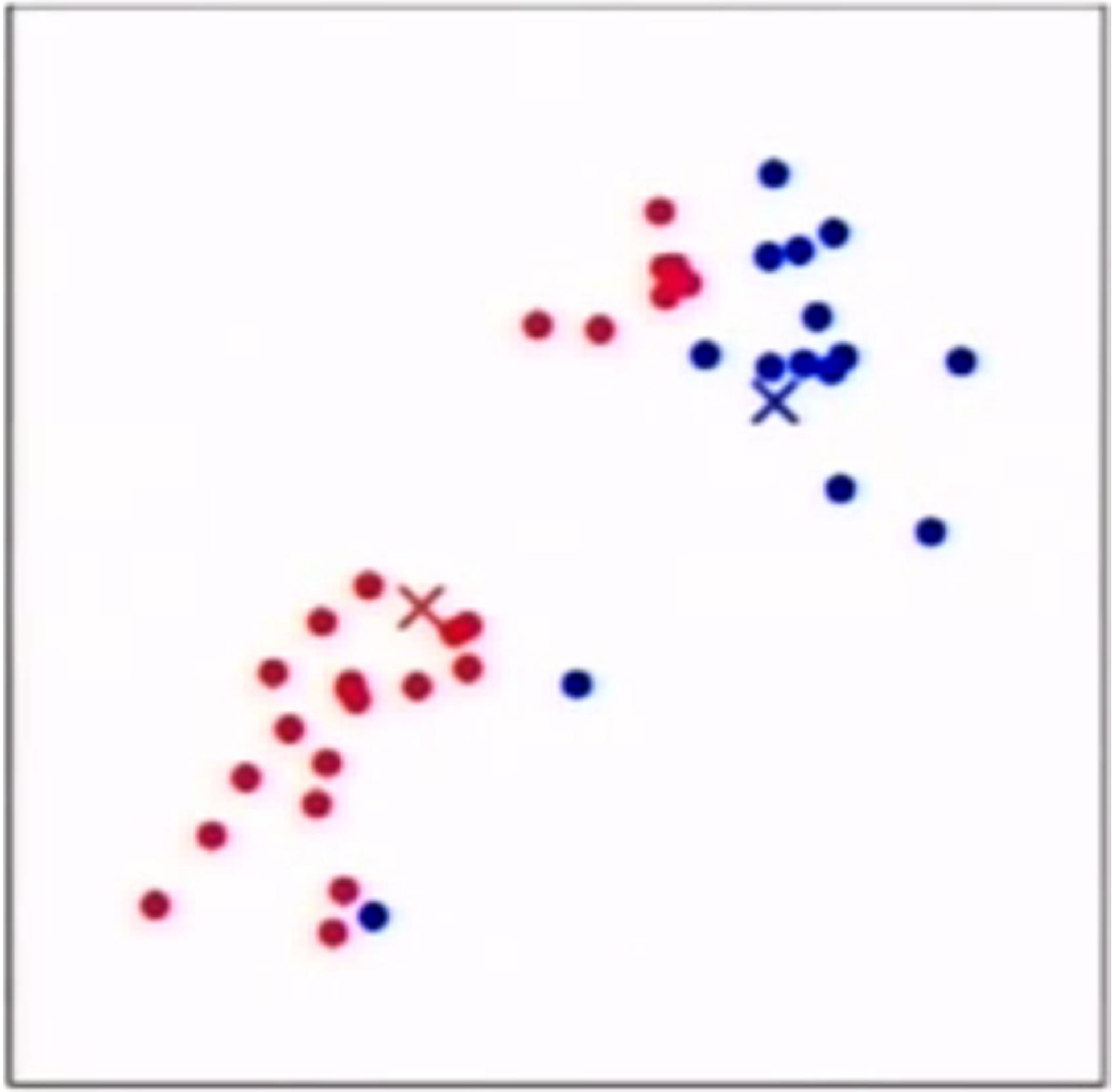
K-Means Clustering

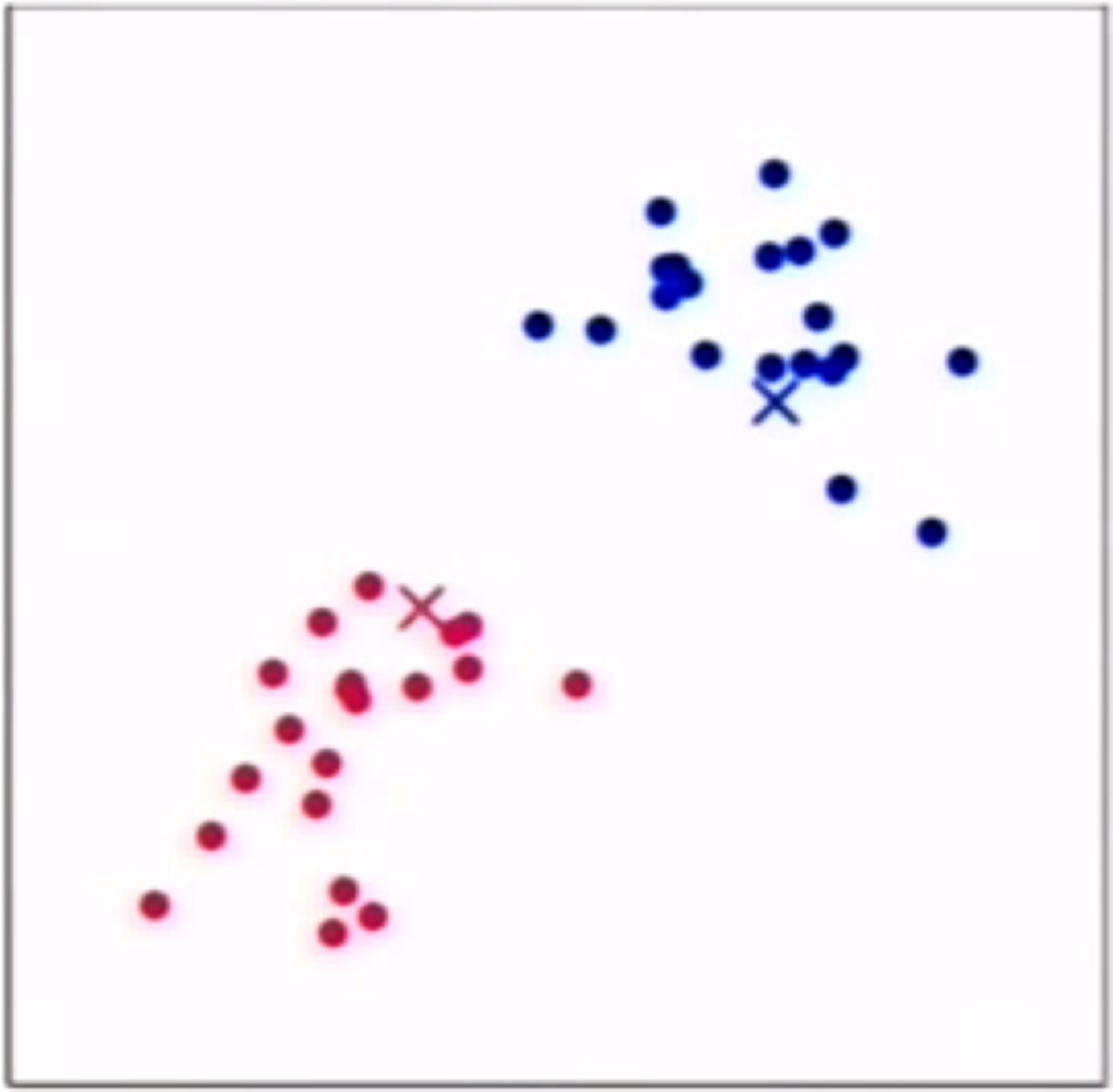
Random Initialization (Partitioning)

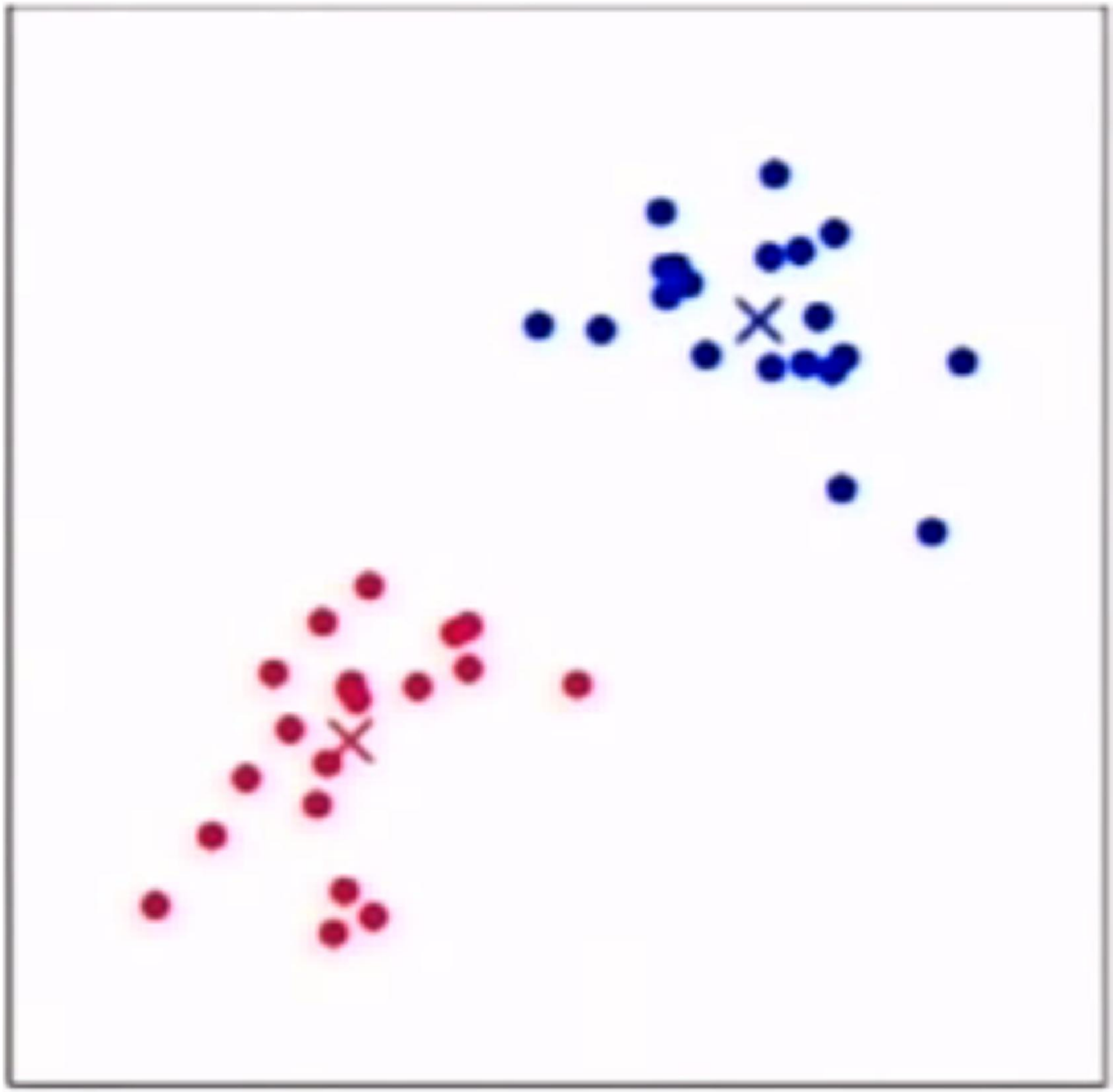






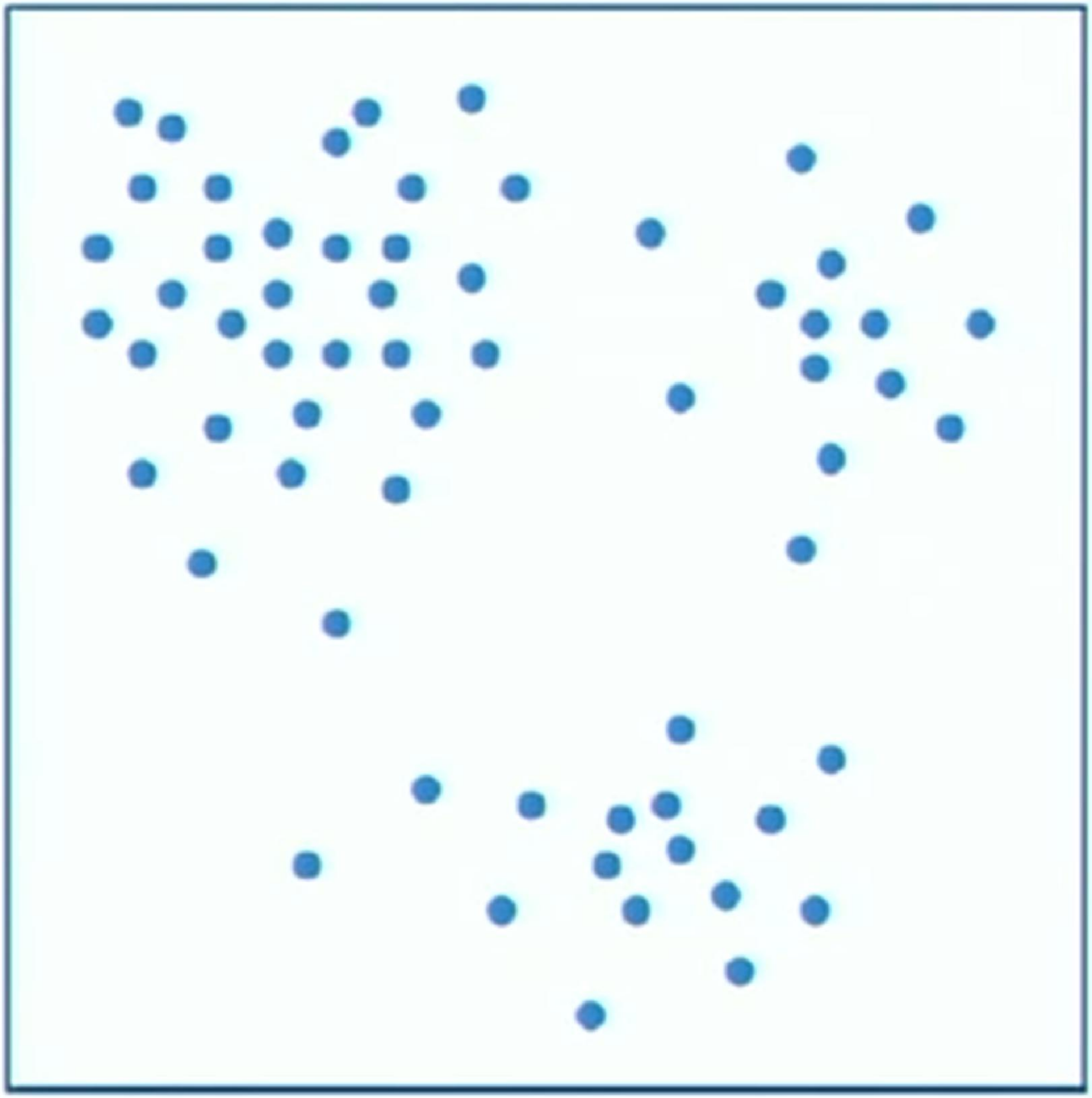


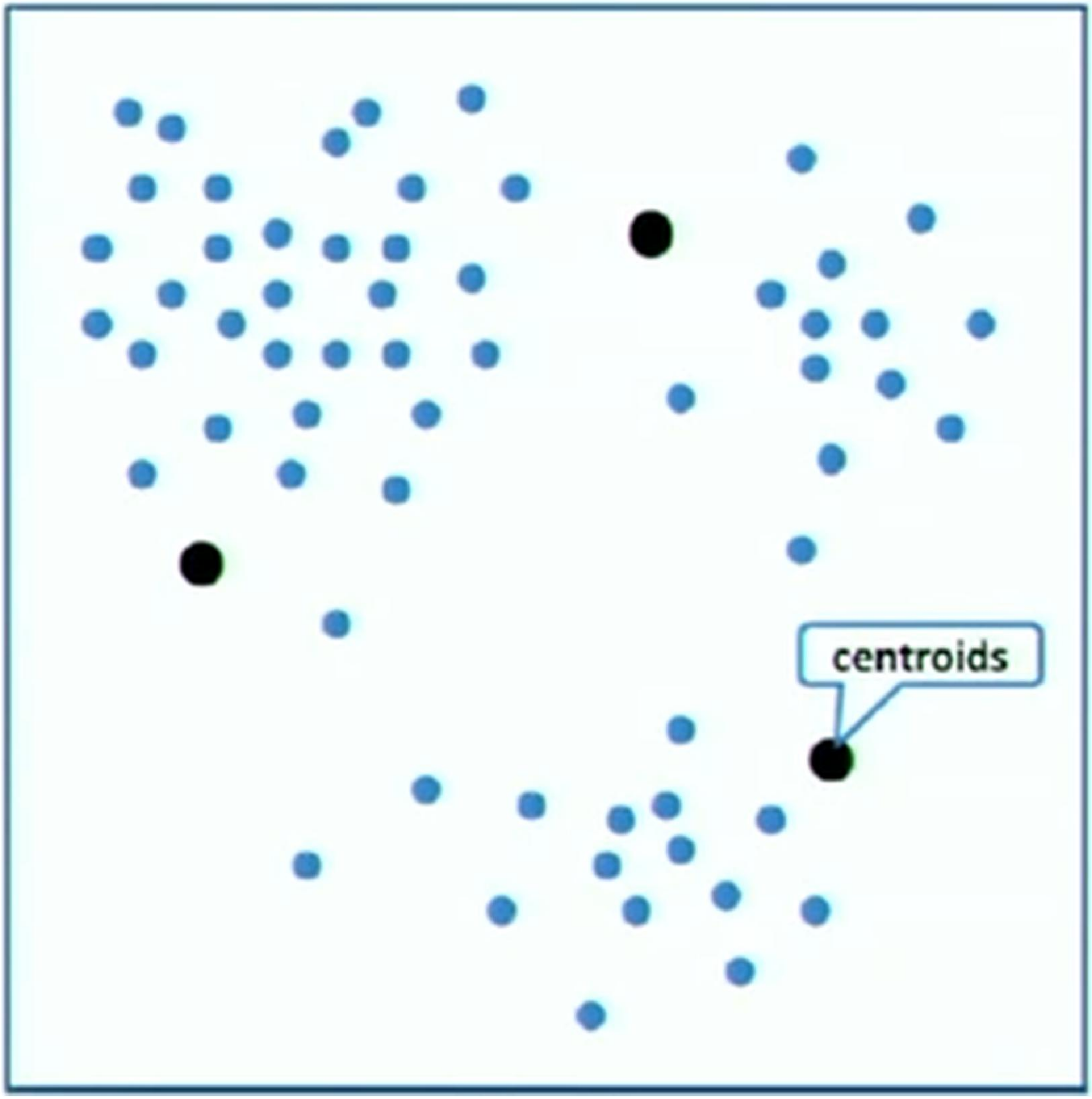


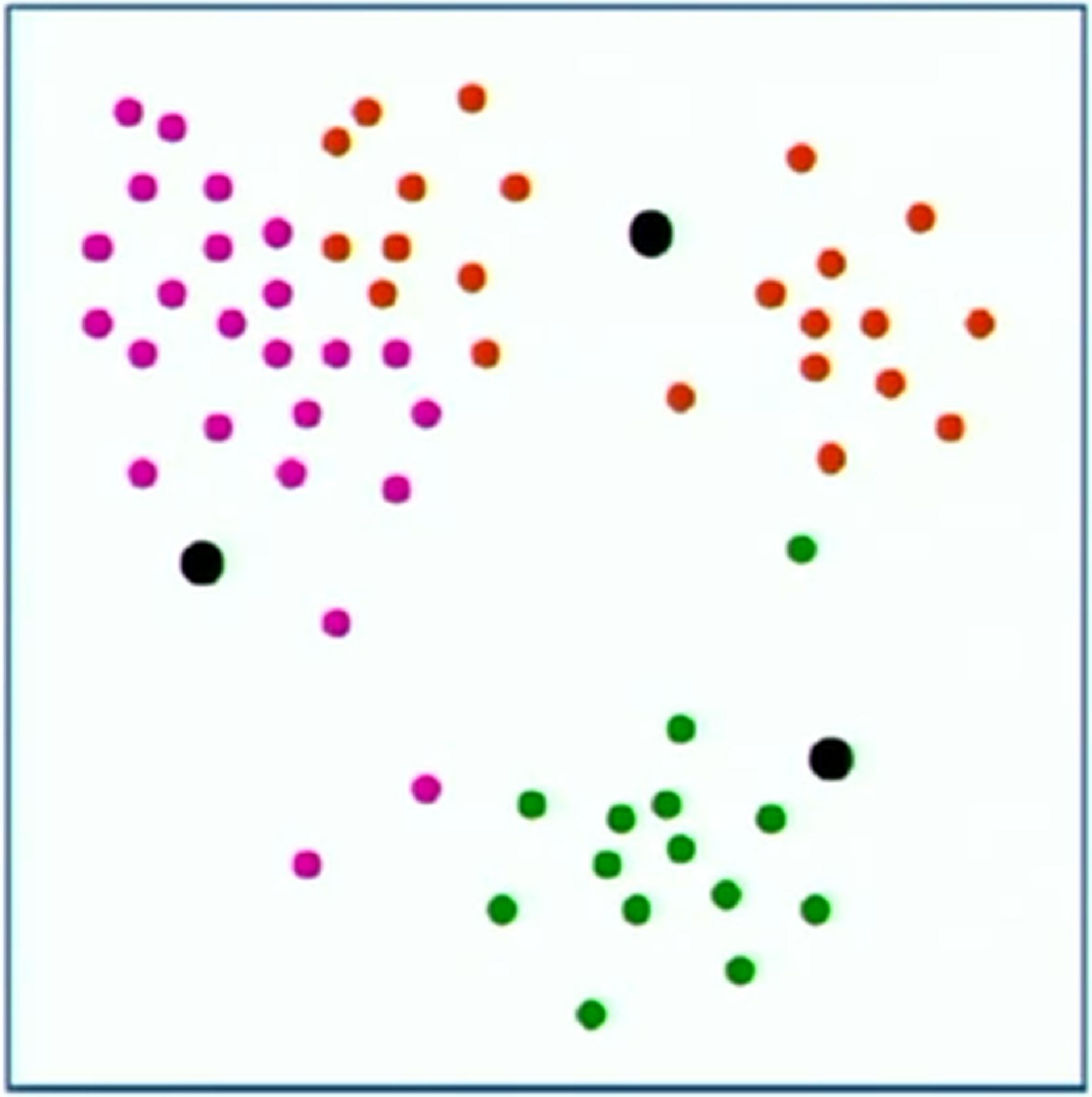


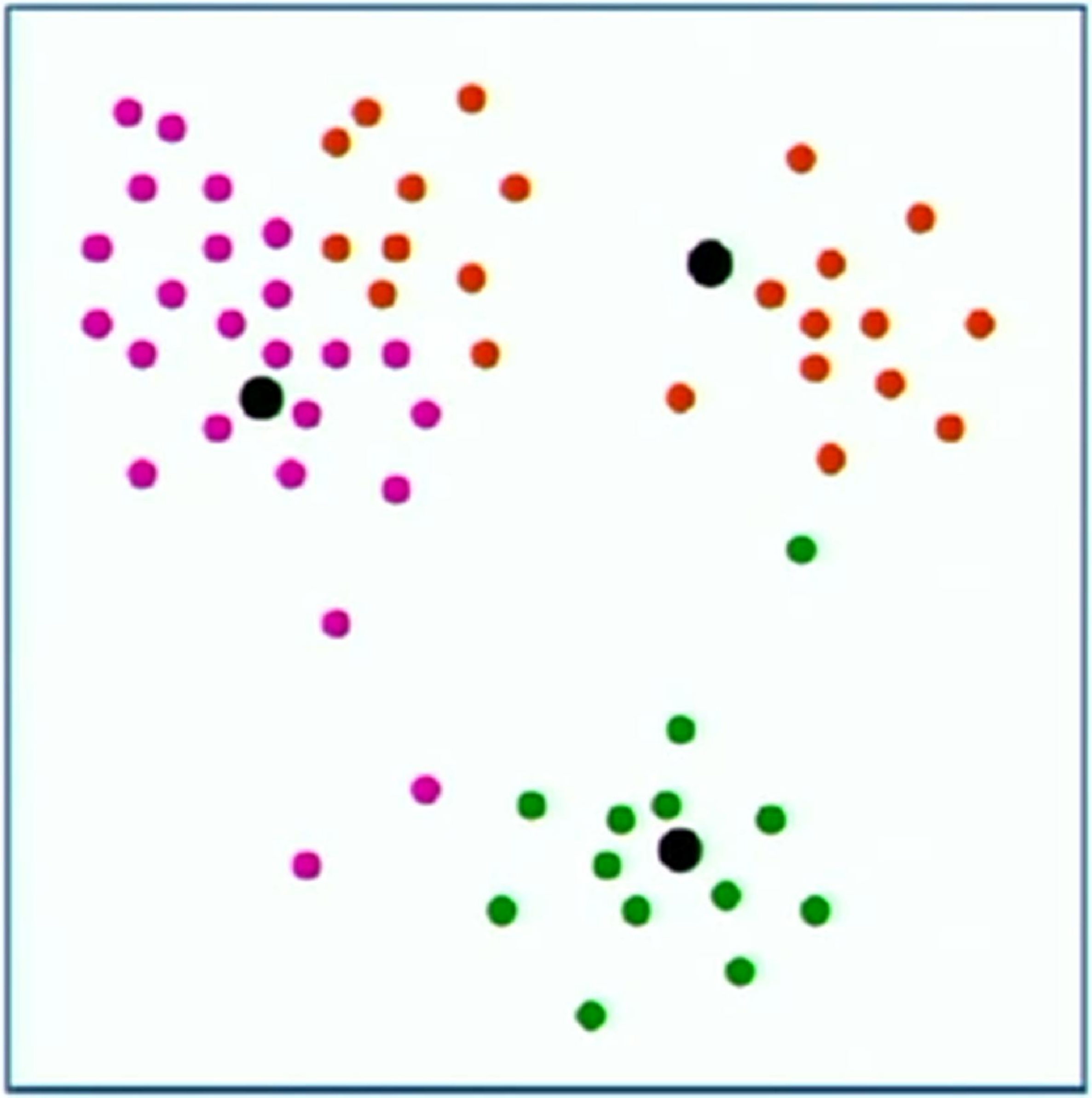
Forgy Initialization

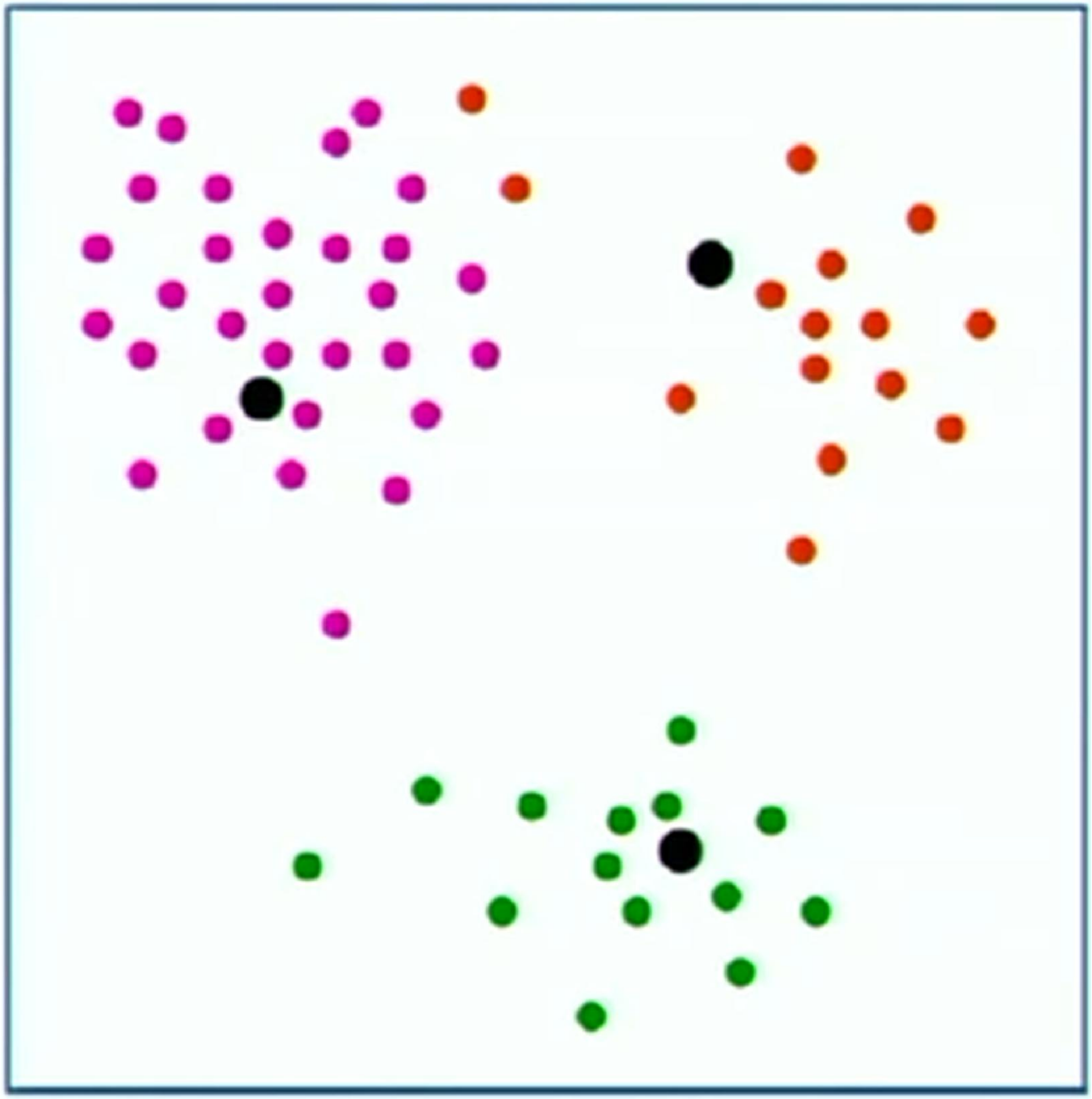
Second Example (Erroneous!)

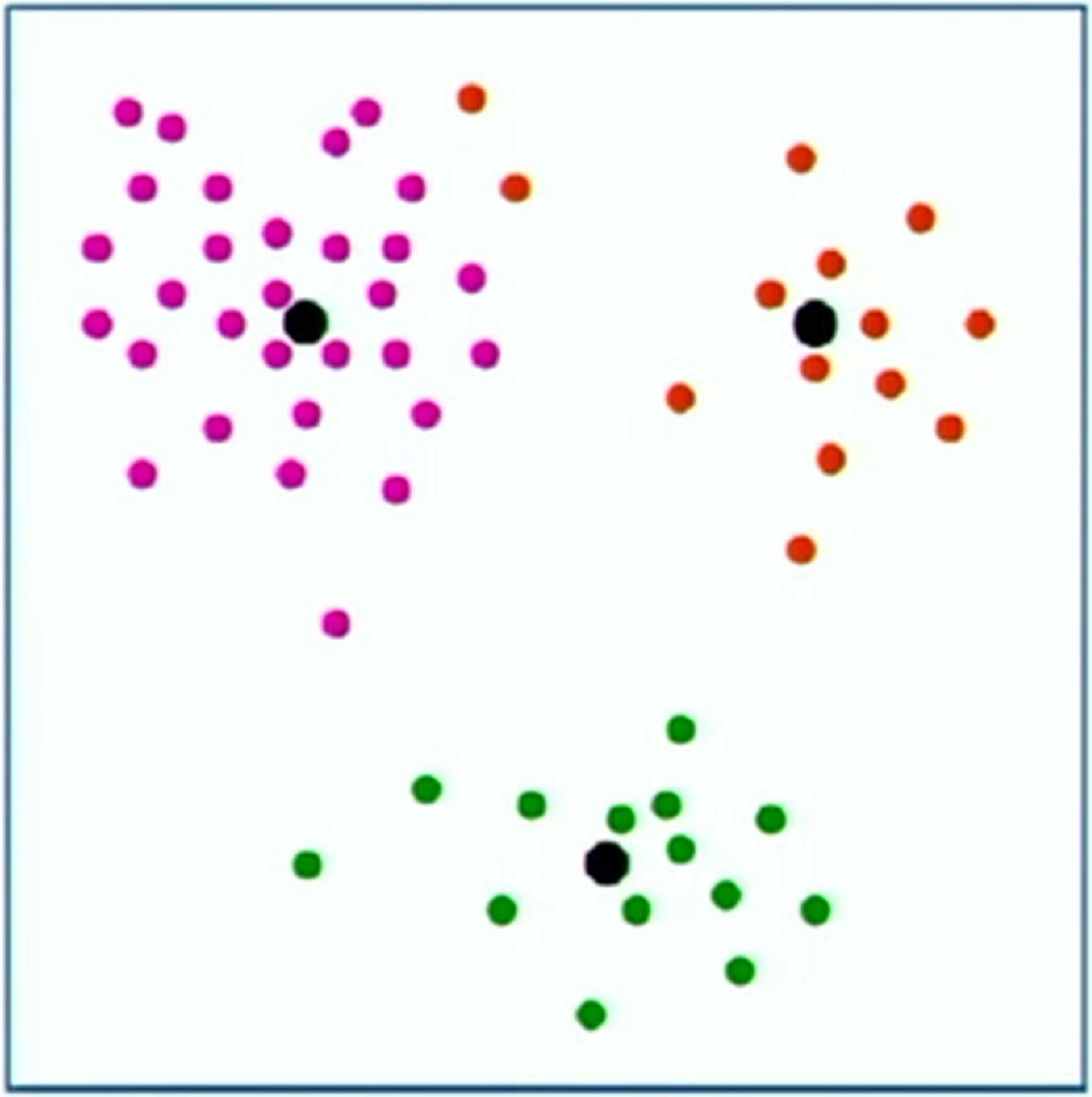


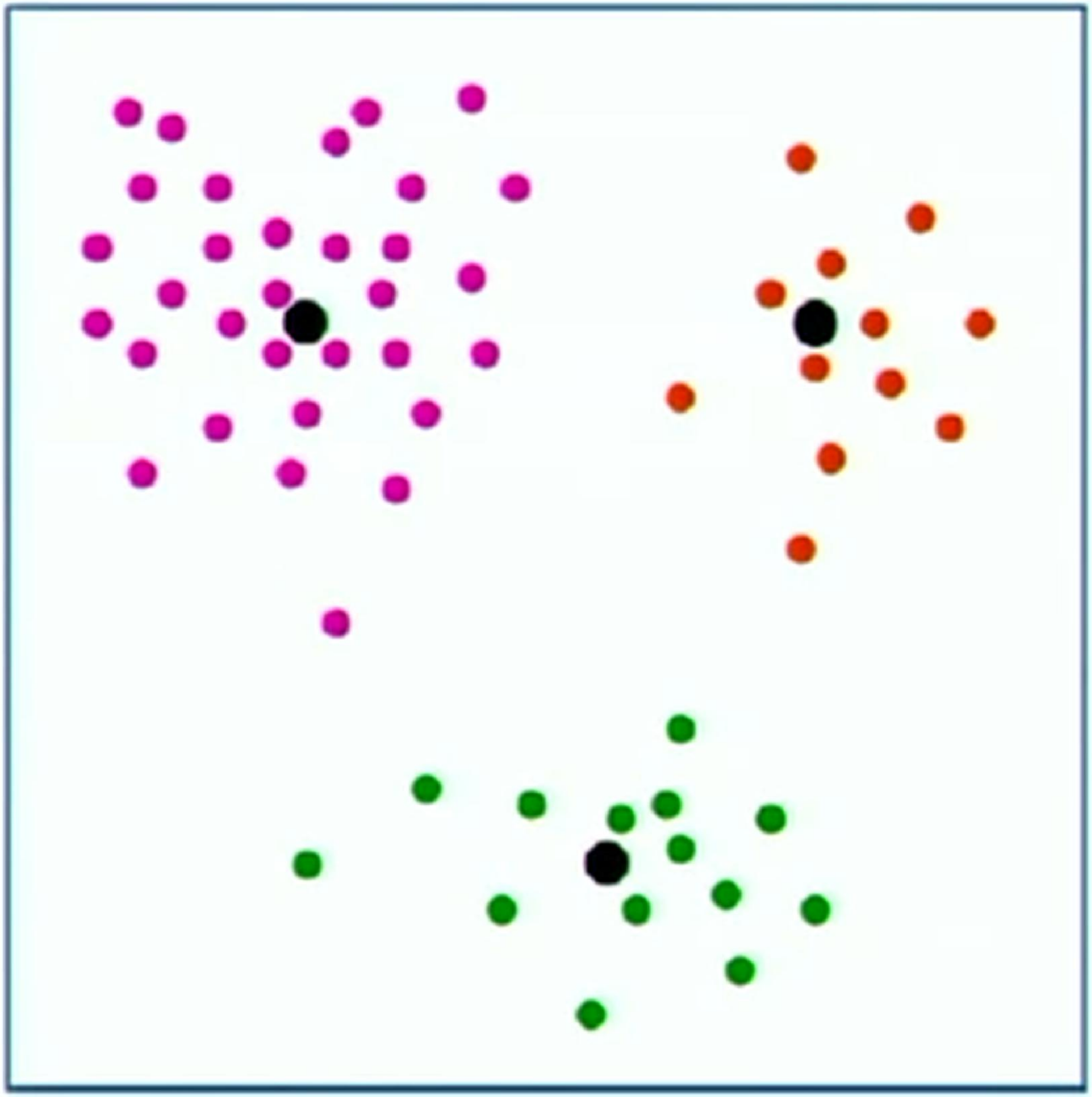

centroids

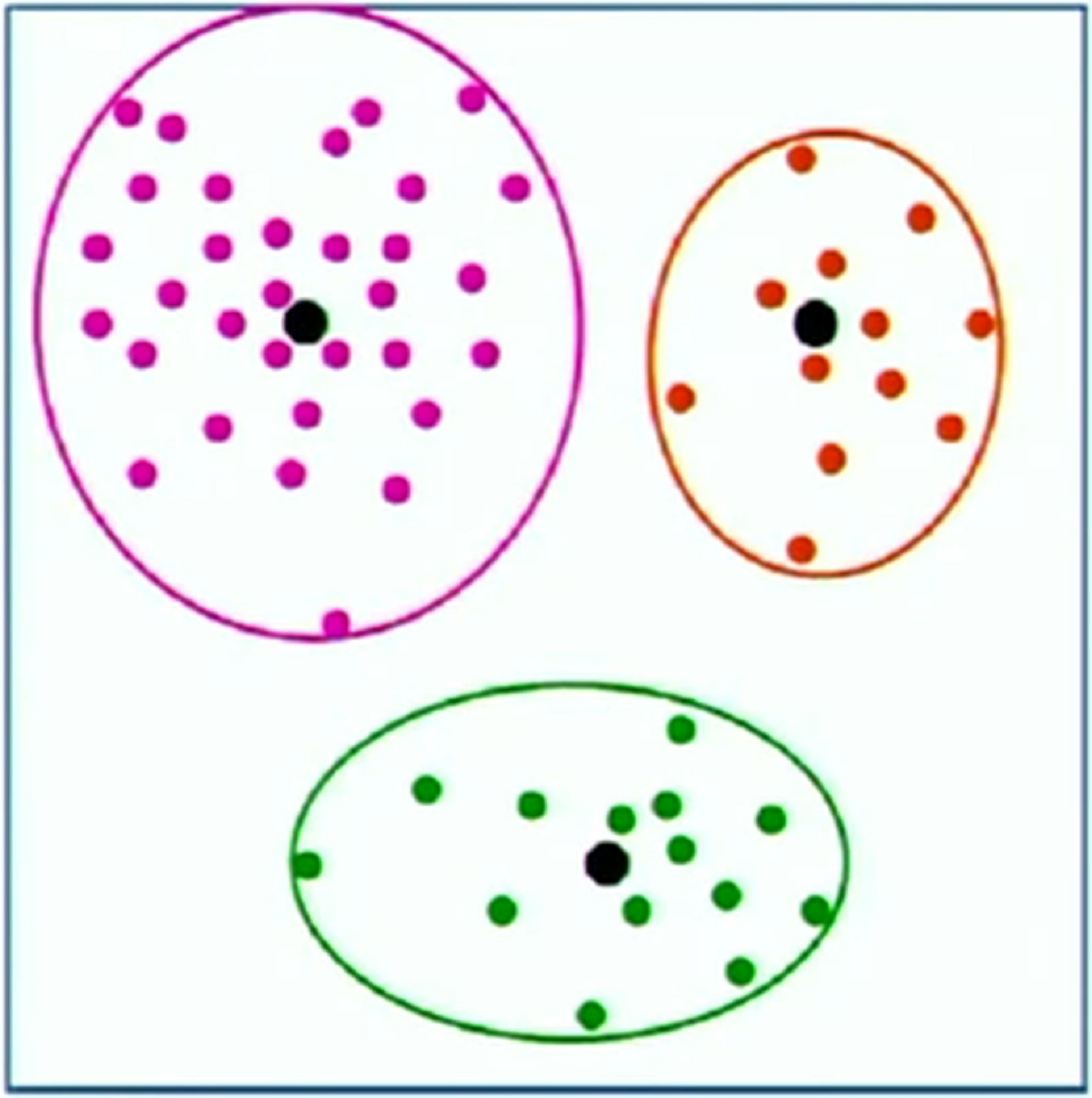






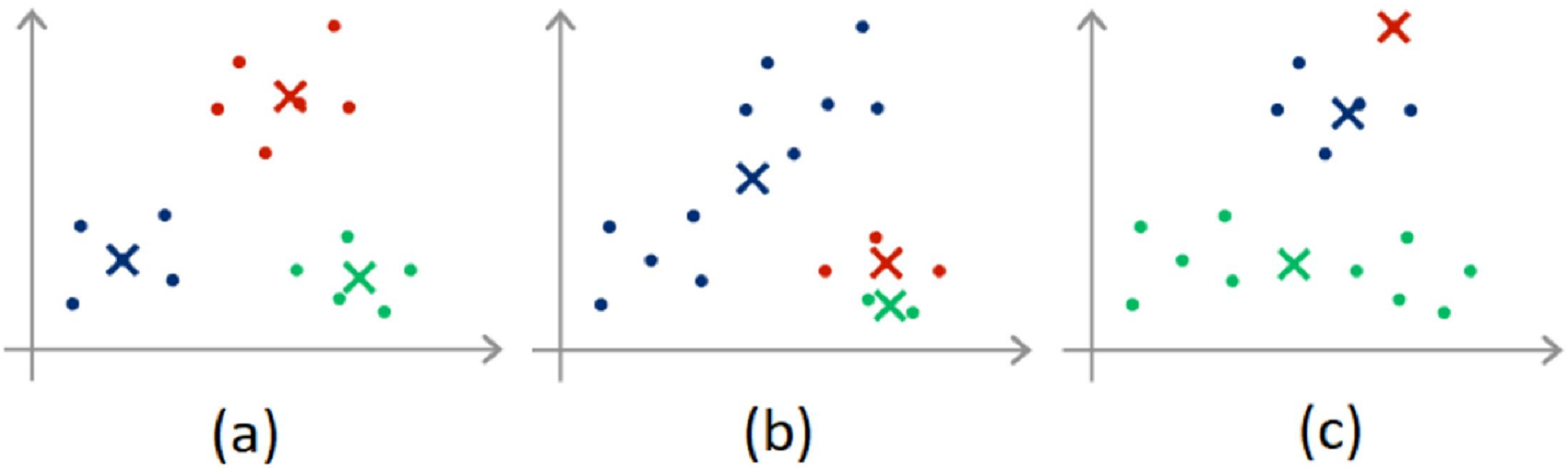






K-means Properties (cont.)

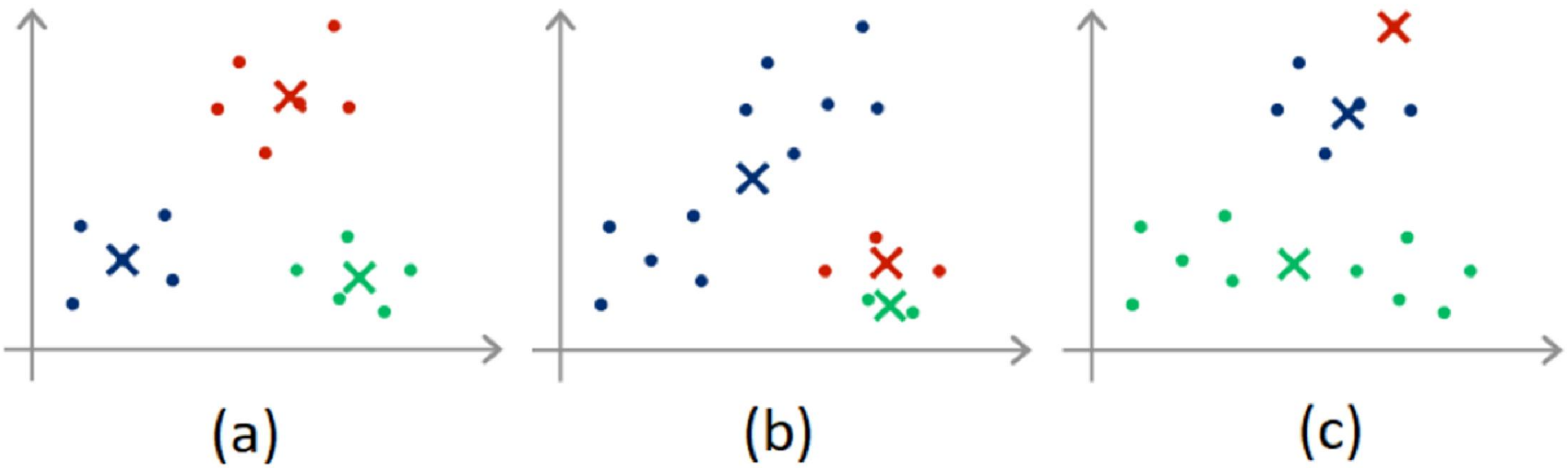
- Poor initialization may lead to poor clustering



- Solution?

K-means Properties (cont.)

- Poor initialization may lead to poor clustering



- Solution?
  - Multiple Initializations → randomness
  - K-means++, Intelligent K-means, Genetic K-means

K-means Properties (cont.)

- Distance metrics
  - l_1 norm (Manhattan distance)
  - l_2 norm (Euclidean distance)
  - Cosine similarity

- Centroids
  - Mean
  - Median
  - Medoid
  - …

K-means Properties (cont.)

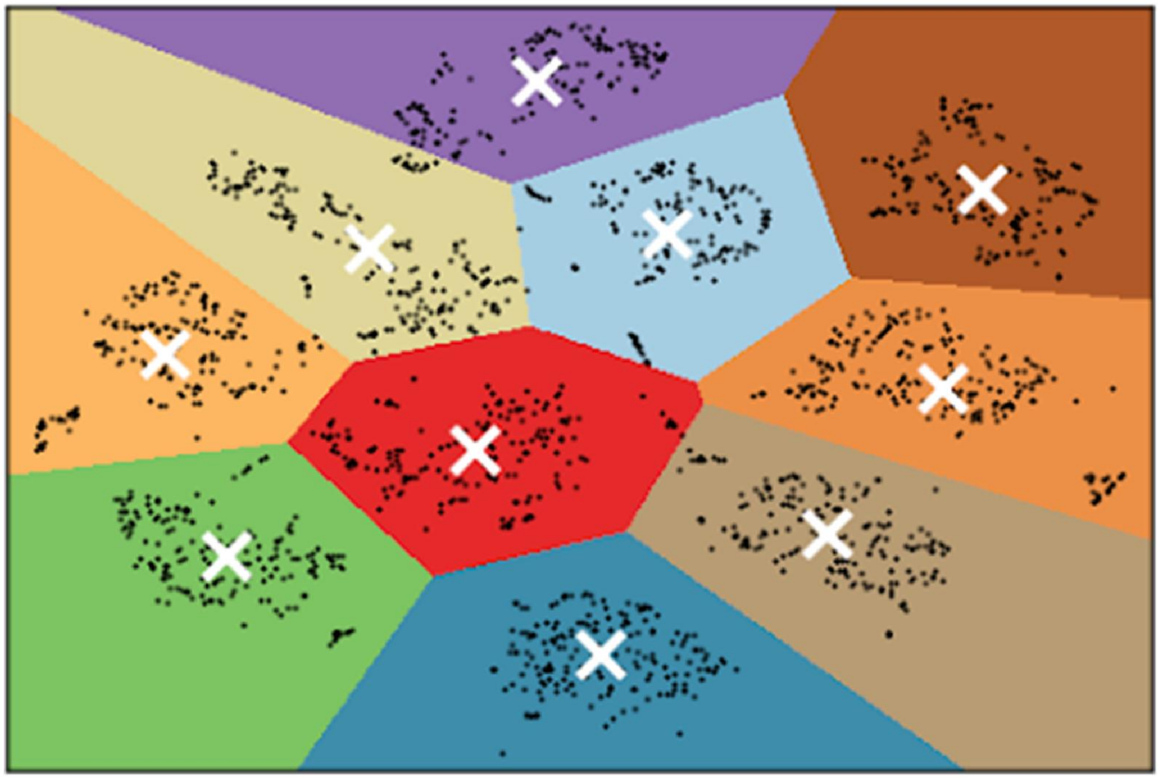
- Distance metrics
  - l_1 norm (Manhattan distance)
  - l_2 norm (Euclidean distance)
  - Cosine similarity
- Centroids
  - Mean
  - Median → Outliers?
  - Medoid
    - Most commonly used on data when a mean or centroid cannot be defined, such as graphs.
  - …

K-means Properties

- Unsupervised
- Instance-based
- Time complexity: $O(tkn)$
- Non-parametric
- Linearly separable data

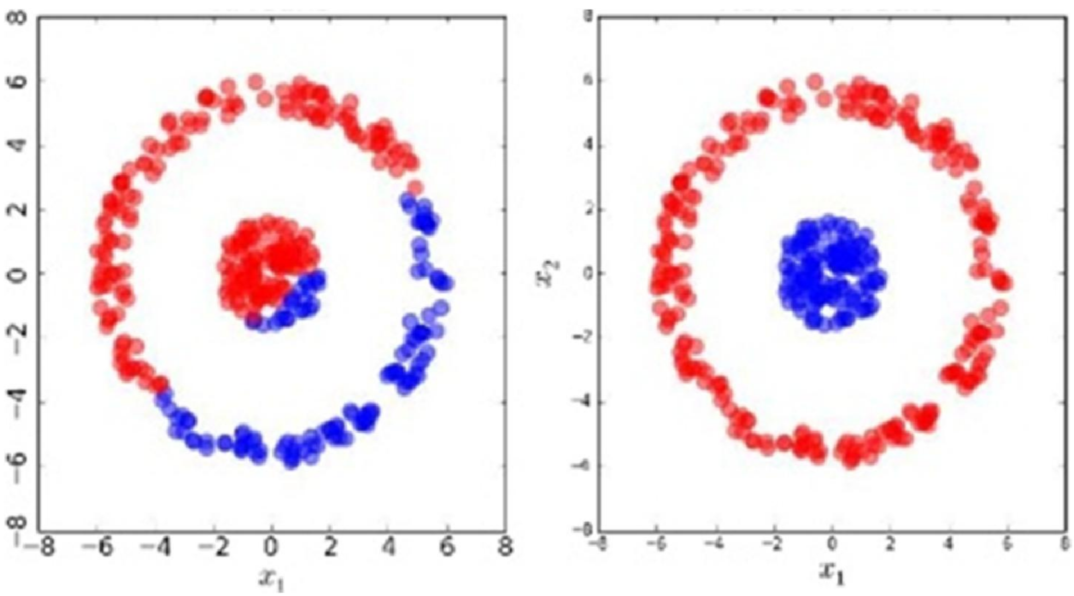
K-means Properties (cont.)

- Linearly separable data



K-means Variations

- Kernel K-means



- Fuzzy C-means

k-means Clustering

Video

Sum of Square Error

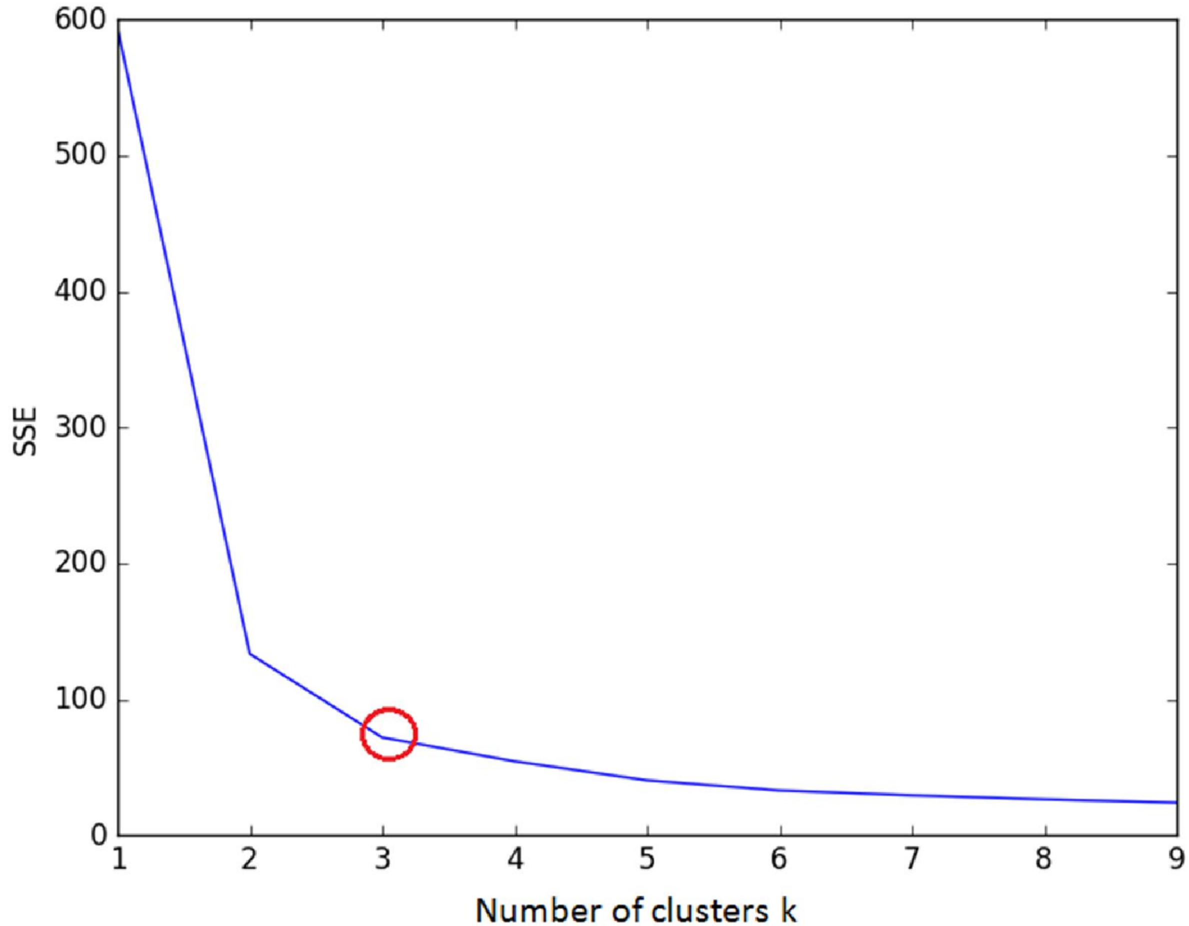
- Sum of Square Error (SSE)

$$SSE = \sum_k \sum_{\bar{x}_i \in C_k} ||\bar{x}_i - C_k||^2$$

  - Goal: minimizing within-cluster distance

Optimal number of Clusters

- Elbow method



Applications

- Document classification (news, …)
- Sentiment analysis (customer reviews, …)
- Anomaly detection
- Fraud detection
- Trend analysis
- …