
DATA MINING EXAM SO FUNNY
I FORGOT TO LAUGH
memegenerator.net

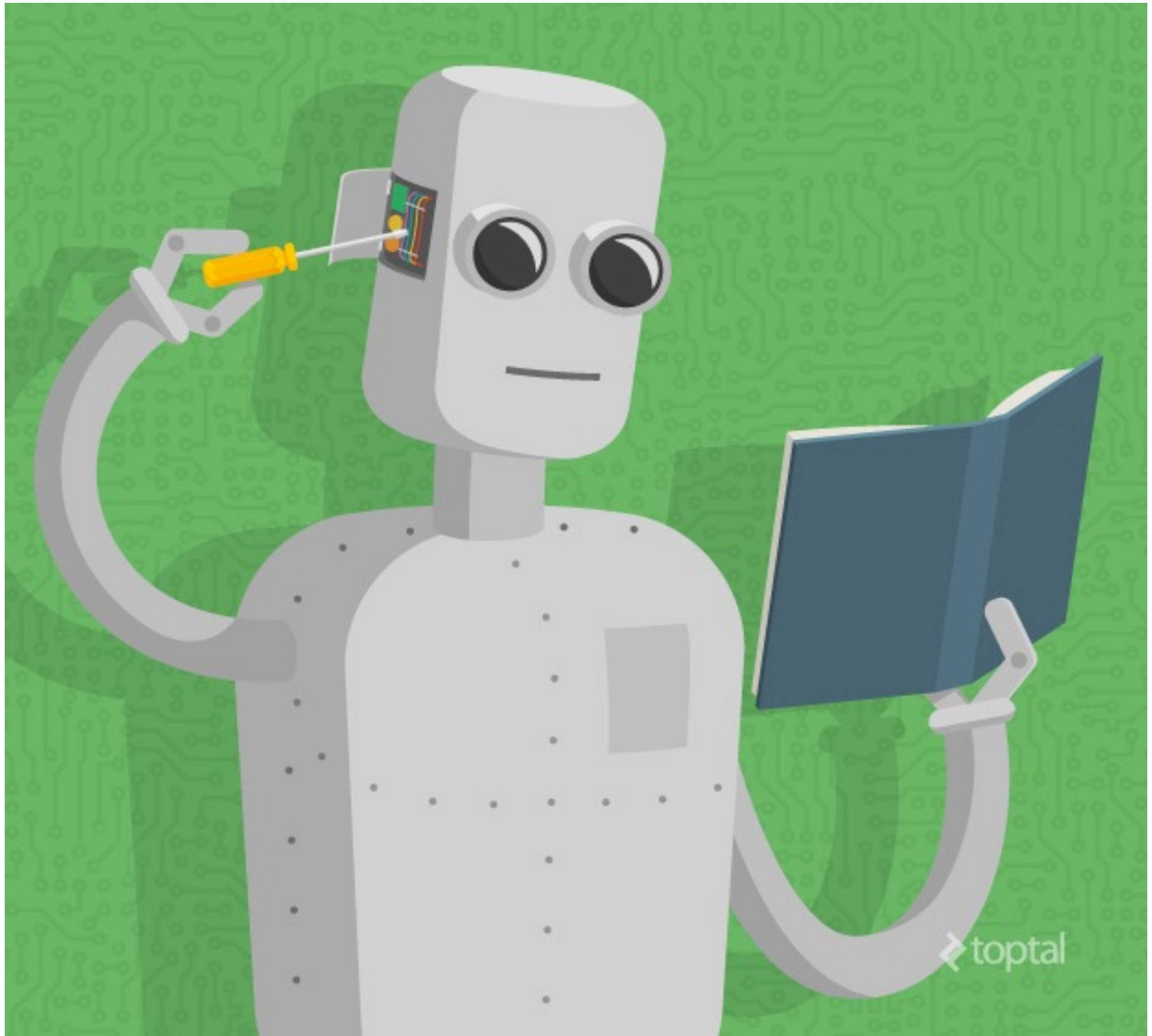
Target knows your secrets…

"Just wait. We'll be sending you coupons for things you want before you even know you want them." –Andrew Pole, Target statistician

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#3a157e826668

Clustering

Unsupervised Machine Learning



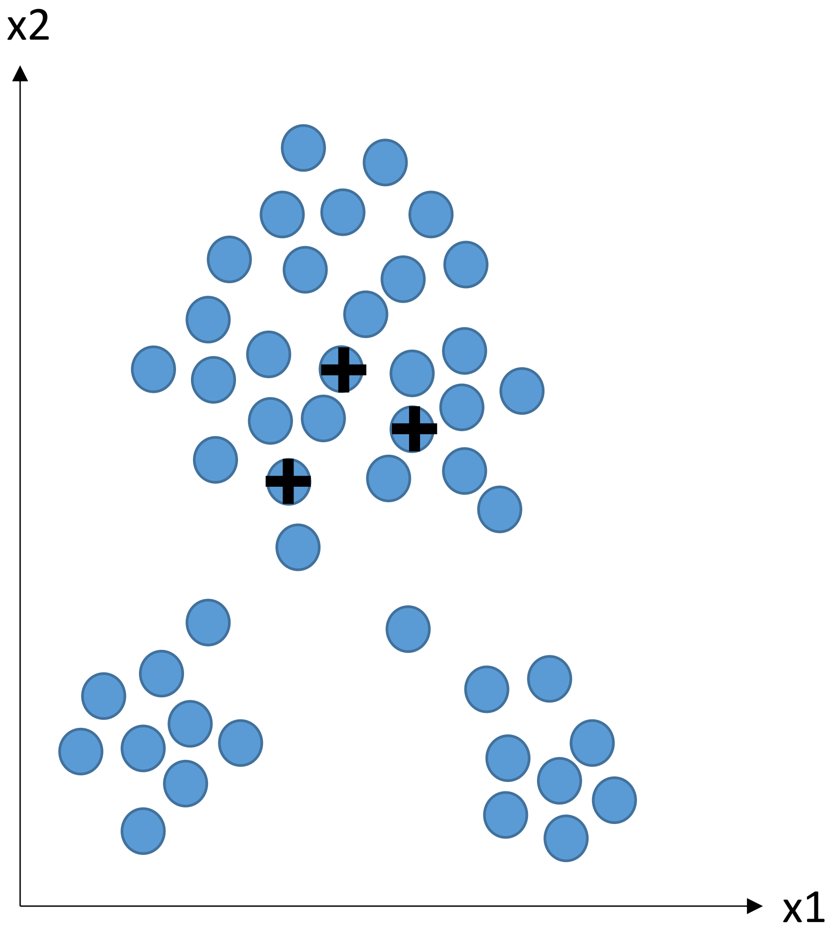
Clustering

- Cluster analysis divides data into groups (clusters) that are meaningful and/or useful
- Clusters are potential classes and cluster analysis finds them in unlabeled data

- Items in a cluster should be similar to each other, but different from those outside of their cluster

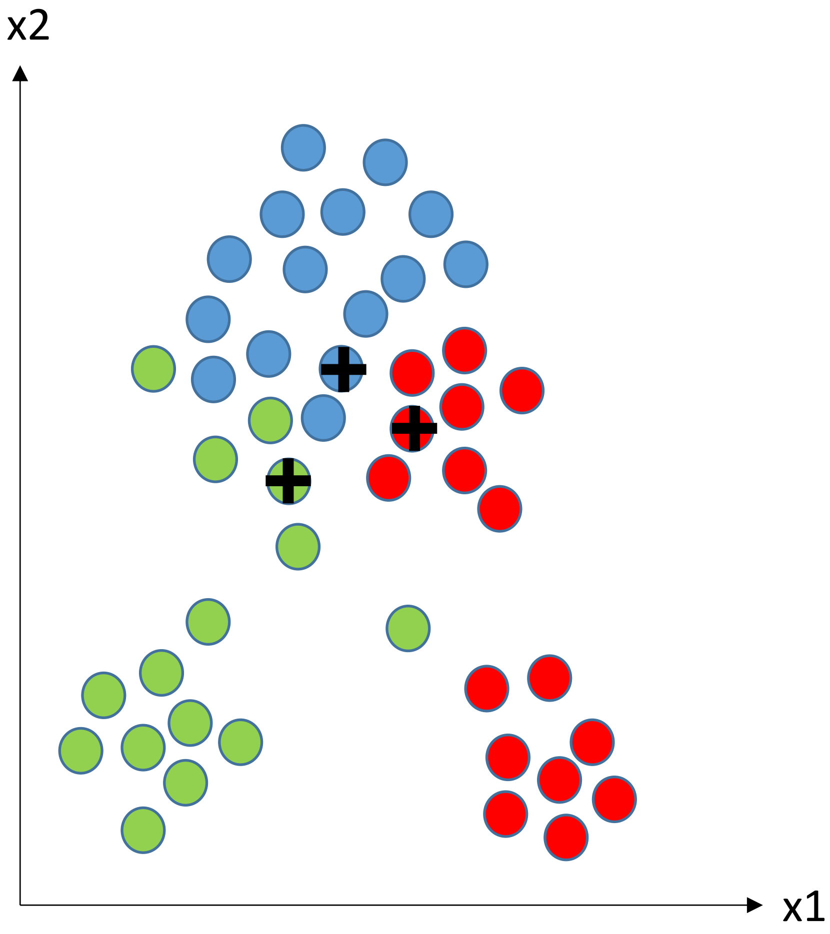
K-means algorithm

- Choose K random data points from the training data to be initial centroids
- Each data point is assigned to the closest centroid, to form clusters



Iteration 1

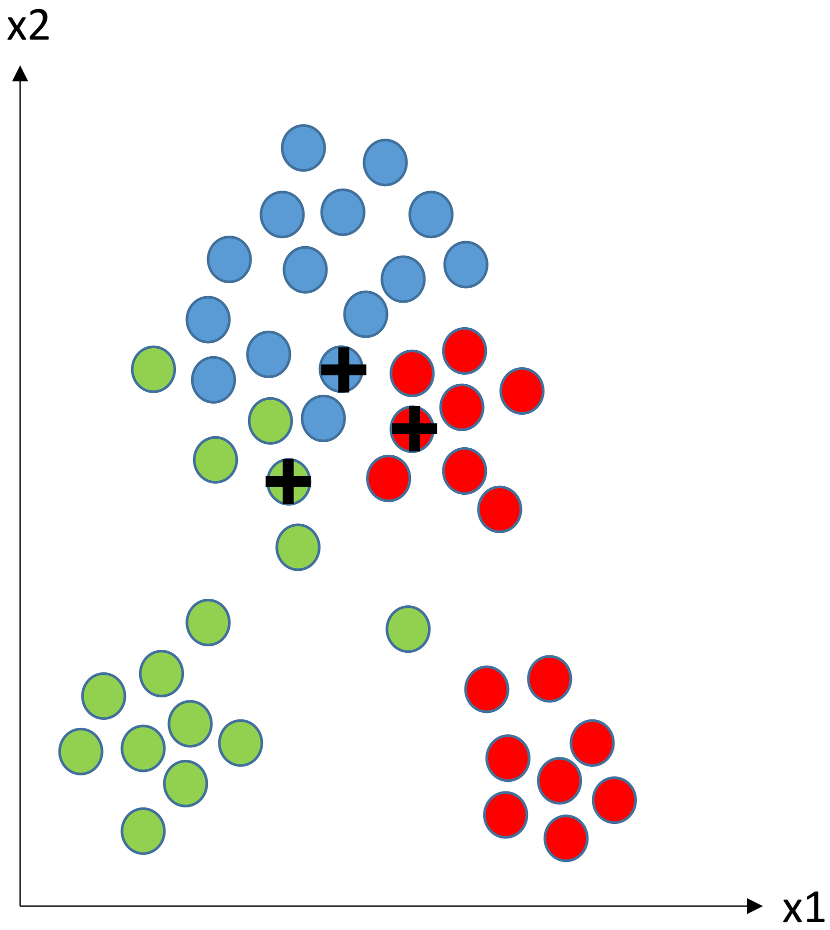
K=3



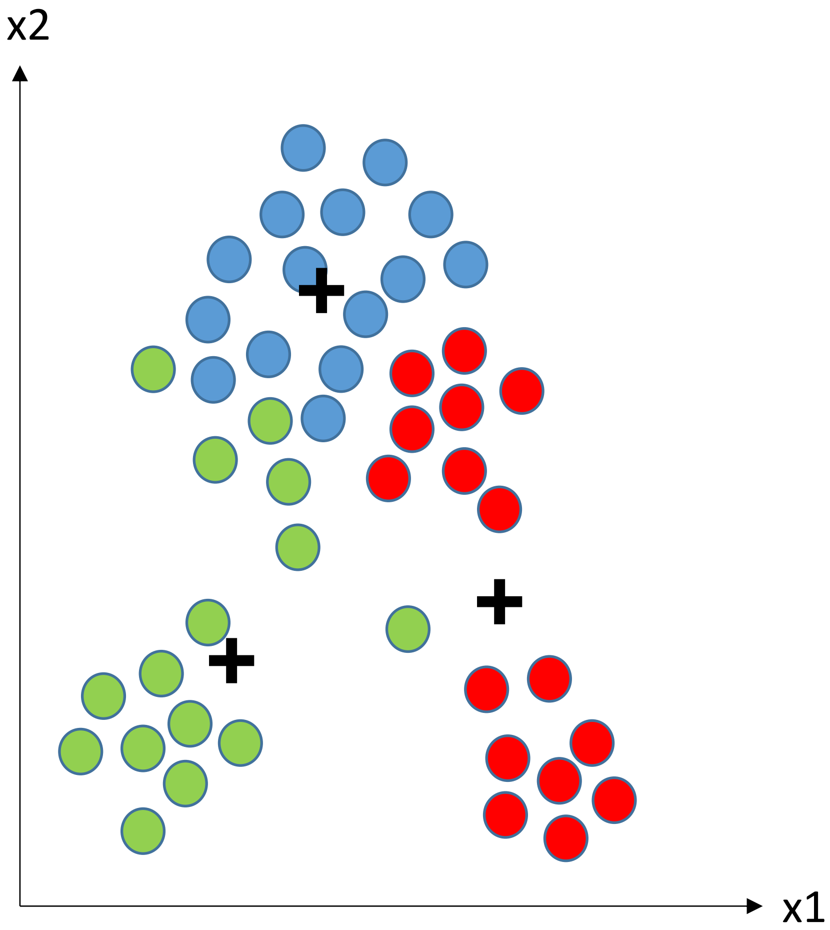
Iteration 1

K-means algorithm

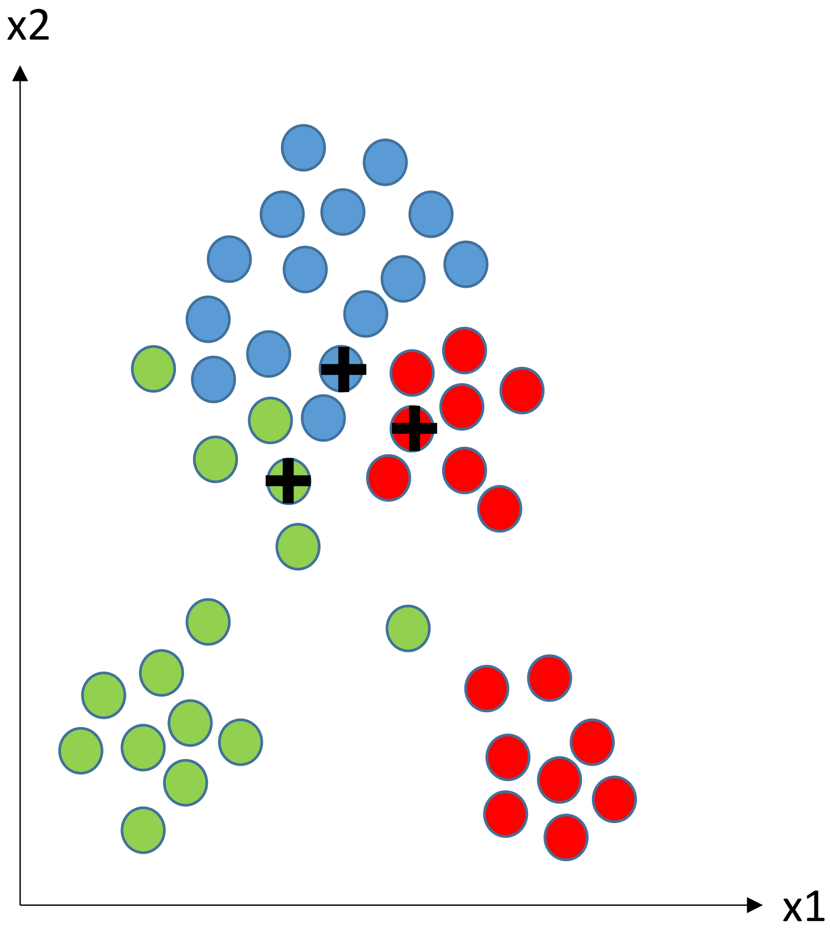
- Choose K random data points from the training data to be initial centroids
- Each data point is assigned to the closest centroid, to form clusters
- Update the centroid of each cluster based on the mean of the points assigned to that cluster
- Re-assign points to their closest centroid



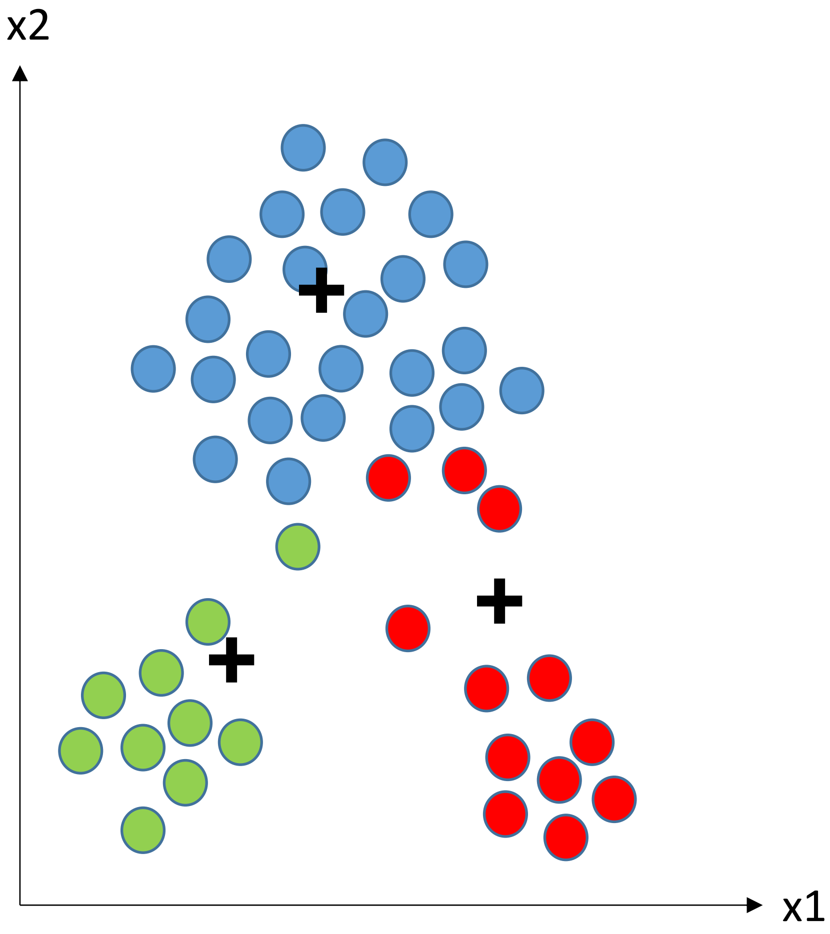
Iteration 1



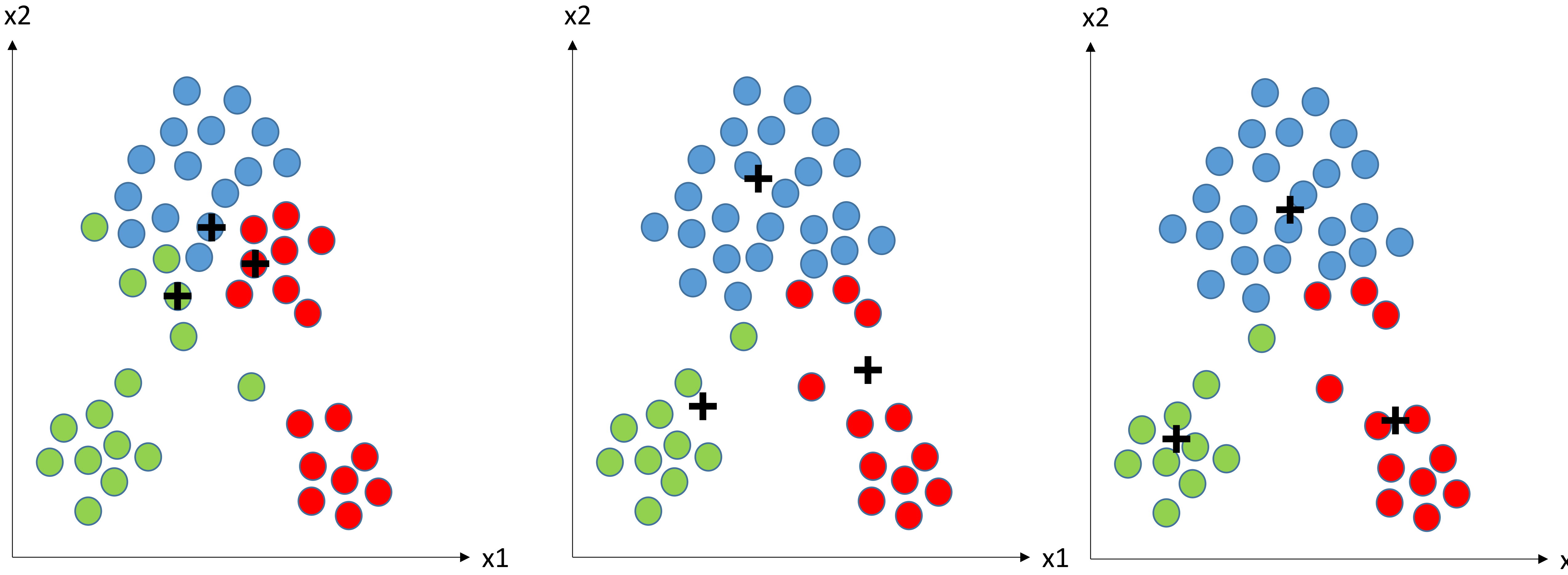
Iteration 2



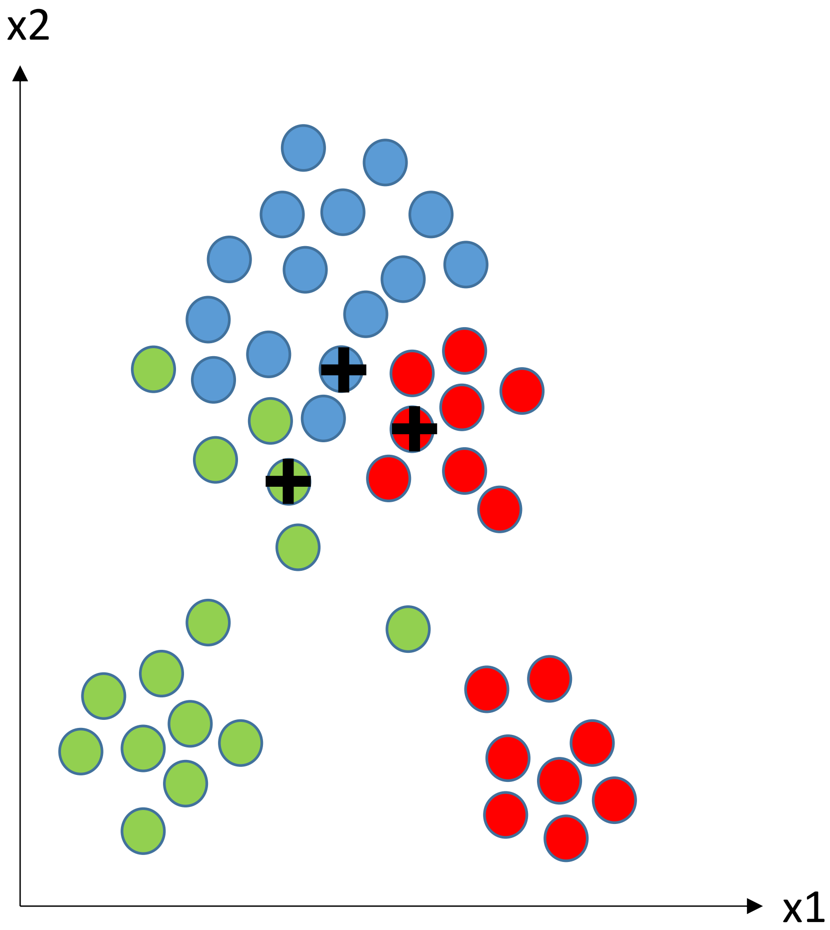
Iteration 1



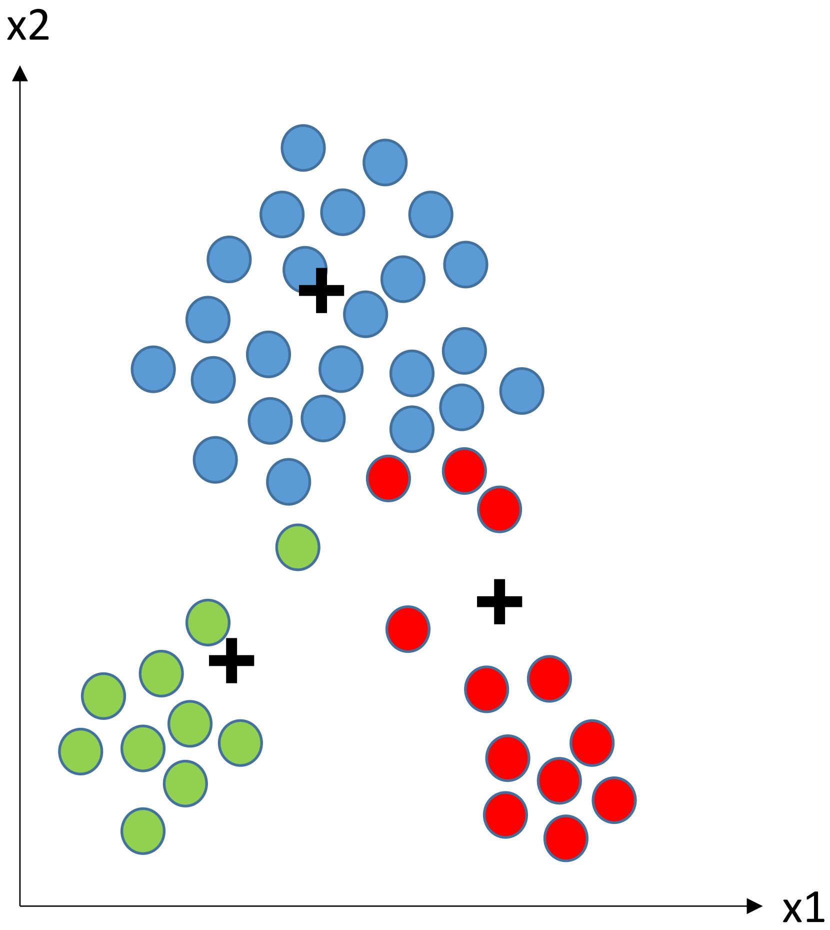
Iteration 2



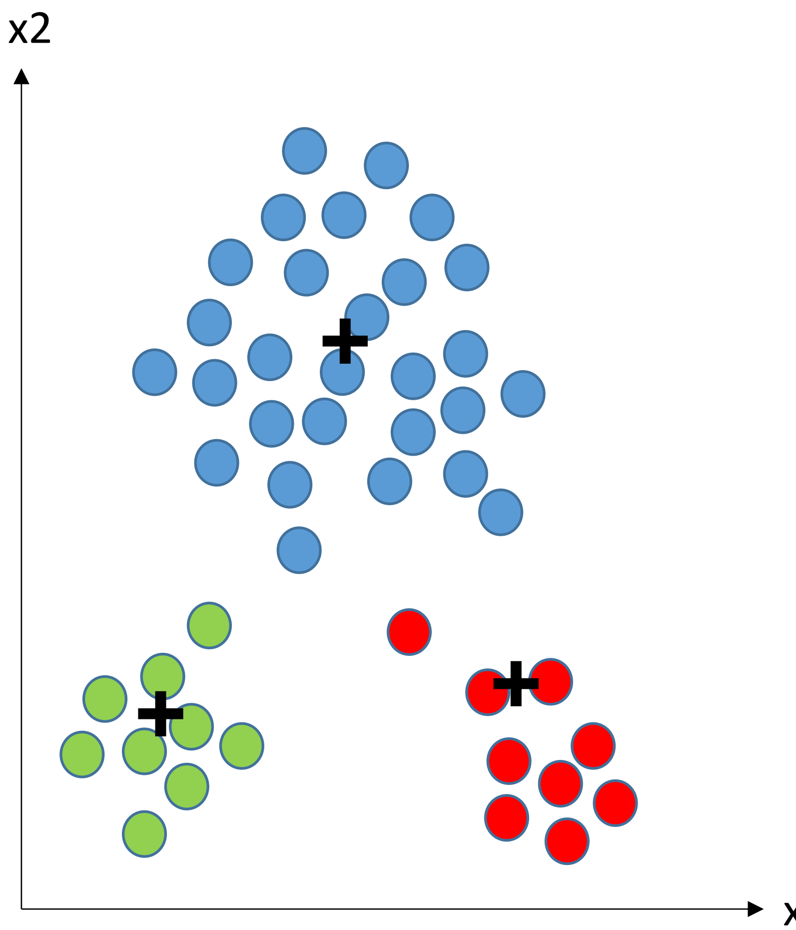
Iteration 1 Iteration 2 Iteration 3



Iteration 1



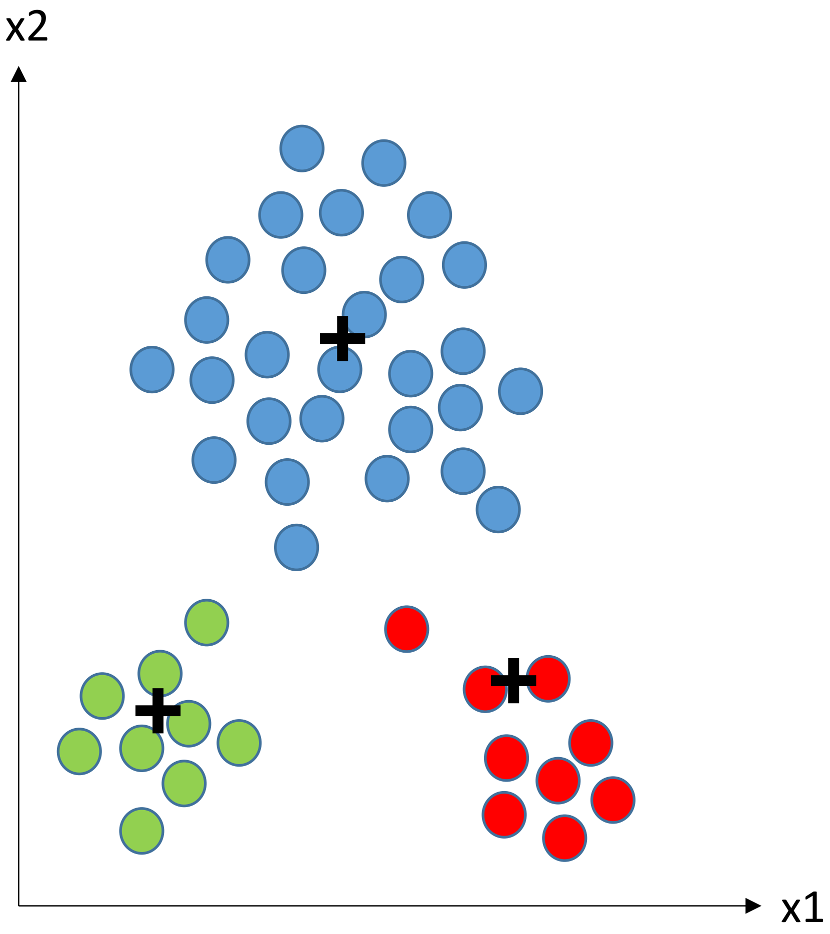
Iteration 2



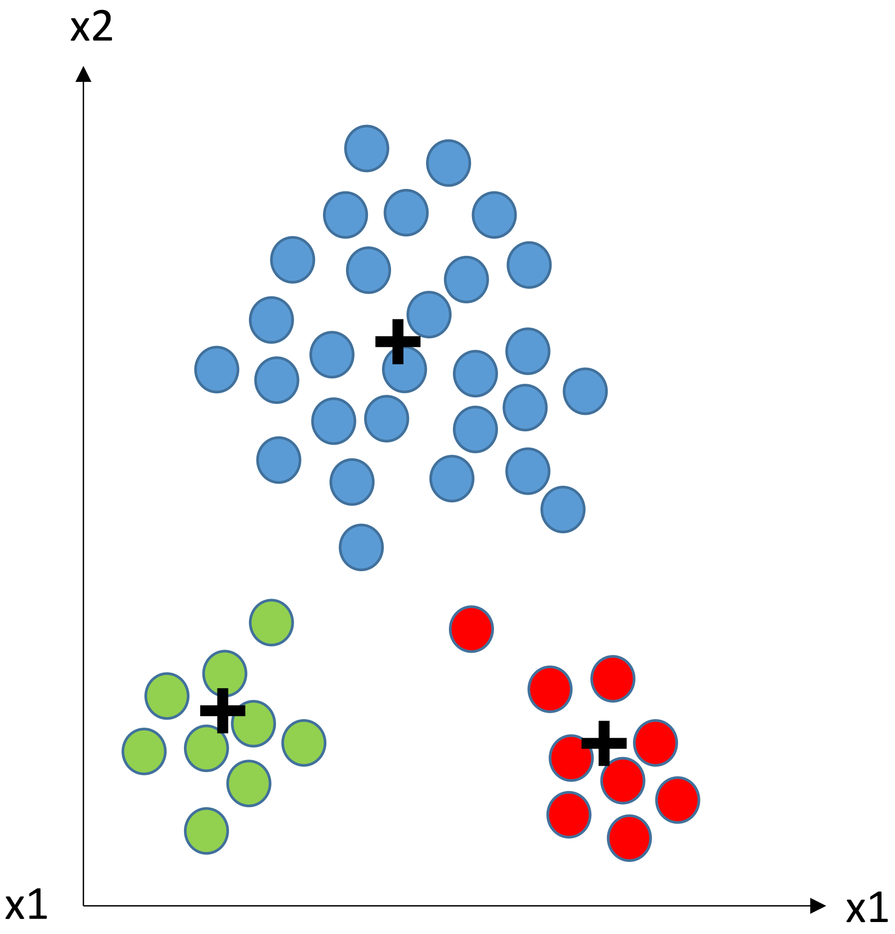
Iteration 3

K-means algorithm

- Choose K random data points from the training data to be initial centroids
- Each data point is assigned to the closest centroid, to form clusters
- Update the centroid of each cluster based on the points assigned to that cluster
- Re-assign points to their closest centroid
- Repeat until no point changes cluster (or equivalently, no centroid changes)



Iteration 3



Iteration 4

Objective Function

- The goal of clustering is usually specified by an objective function and a proximity measure

- With a proximity measure of Euclidian Distance, the objective is to minimize the **sum of the squared error (SSE)**, or **scatter**

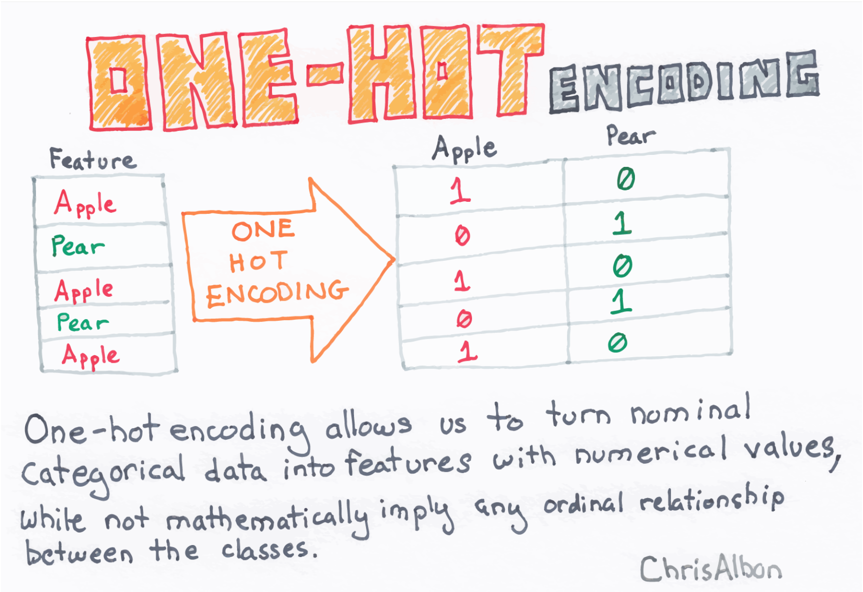
$$\boldsymbol{SSE} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{K}} \sum_{\boldsymbol{x} \in \boldsymbol{C}_{\boldsymbol{i}}} \boldsymbol{dist}(\boldsymbol{c}_{\boldsymbol{i}}, \boldsymbol{x})^{\boldsymbol{2}}$$

Where K is the number of clusters
C_i is the i^{th} cluster
c_i is the centroid of cluster C_i
x is a data point

- Proximity measure, centroid definition, and objective function differ depending on the data and the task.

Handling categorical data:

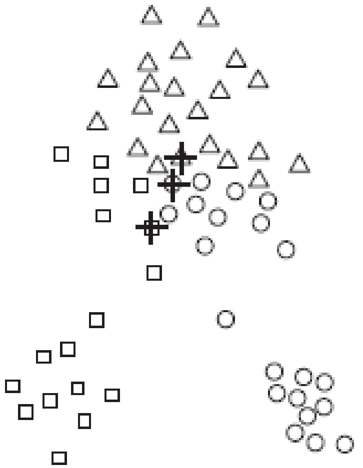
- One-hot-encoding



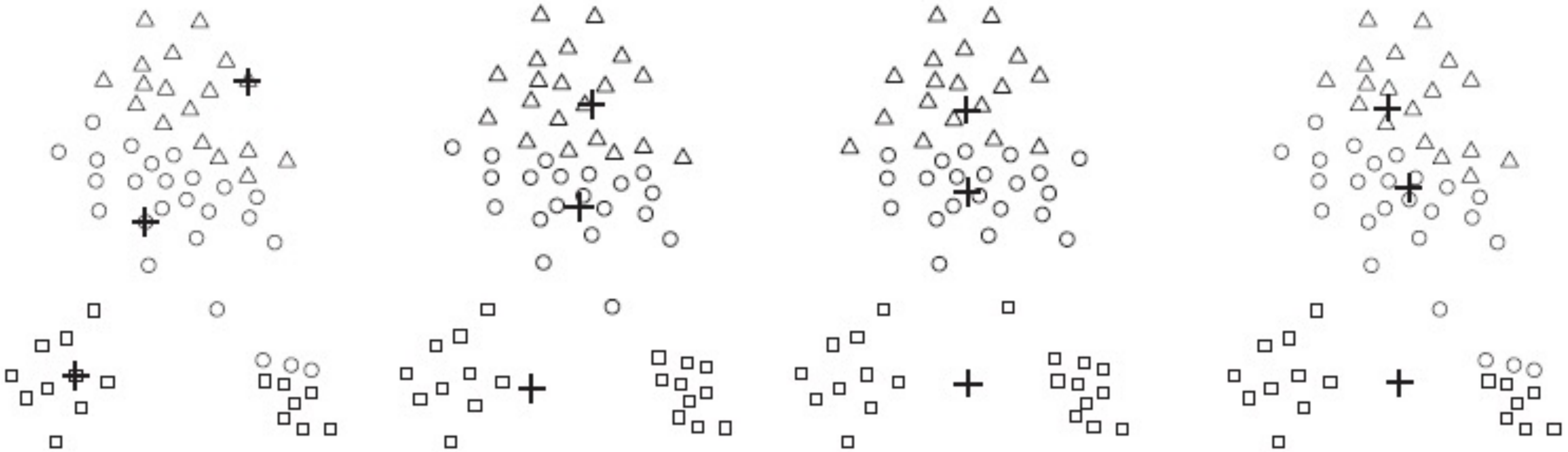
Handling empty clusters:

- If no points are assigned to a centroid, choose a new centroid, either:
  - Randomly
  - The point that is farthest away from any current centroid. This eliminates the point that currently contributes most to SSE.
  - Choose a replacement from the cluster that has the highest SSE. This will split the cluster and typically reduce overall SSE.

Centroid Initialization



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Non-optimal solution (local minimum)

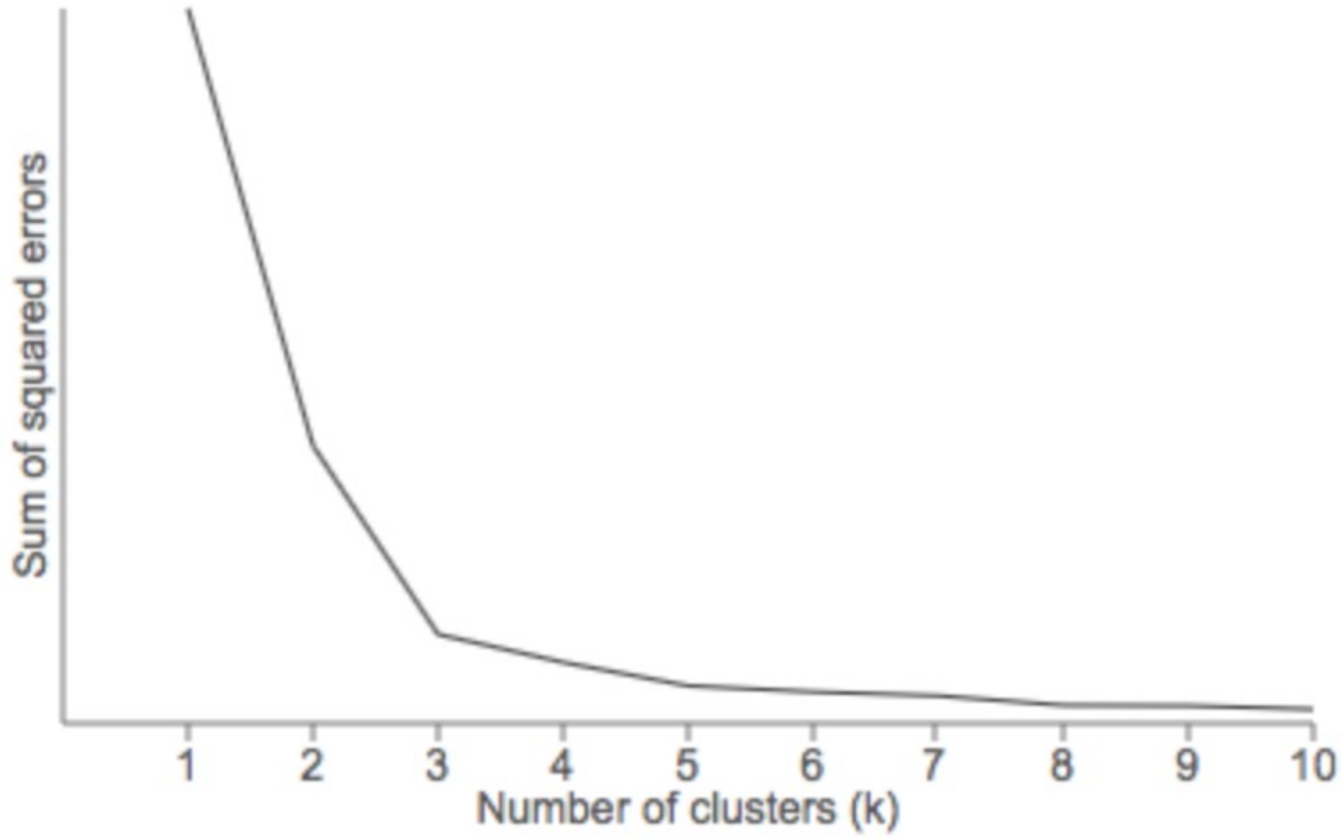
K-means++

- Choose a random centroid
- Repeat until there are K centroids:
  - Compute distance from every point to its nearest centroid
  - Use squared-distance as a probability distribution for choosing the next centroid (farther are more likely to be selected)

Choosing the Right K

- Prior knowledge of how many clusters are in the data
- How many clusters are desired for the application
- Let the data tell you how many clusters it naturally has

Choosing the Right K – Elbow Method



select the value of k at the “elbow” i.e. the point after which the distortion/inertia start decreasing in a linear fashion

Bisecting K-means

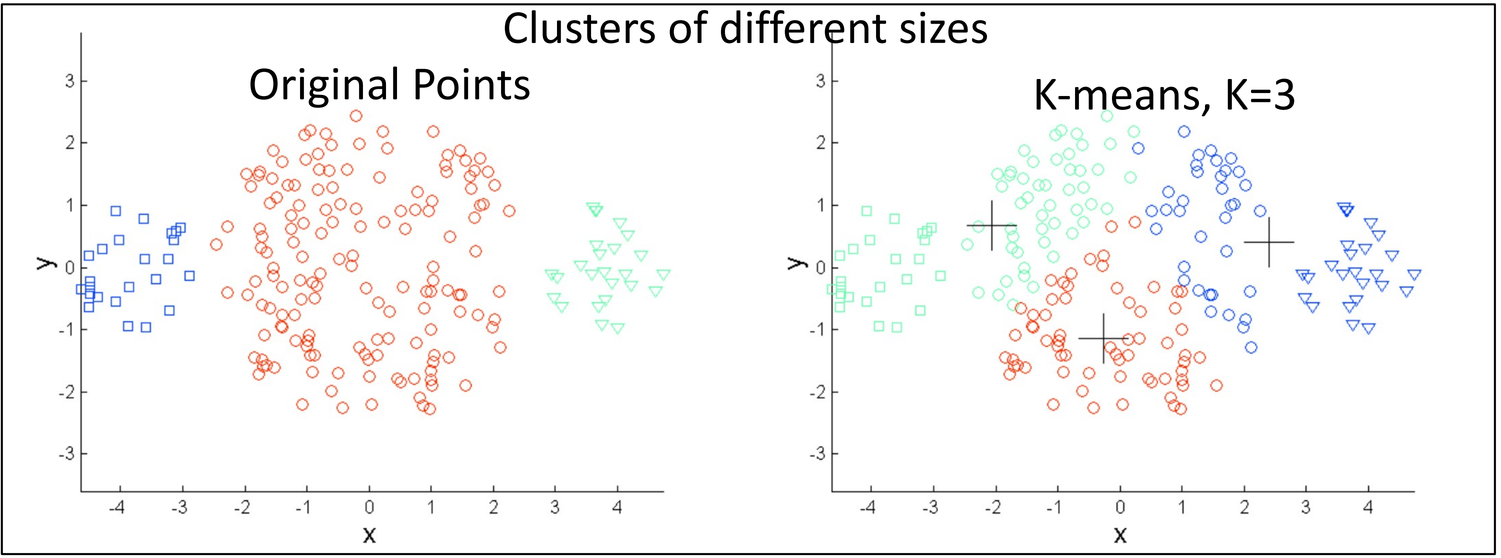
- To obtain K clusters, split the set of all data points into 2 clusters using K-means with K=2
- Choose one of the clusters to split
- Split the chosen cluster using K-means with K=2
- Continue until you have K clusters

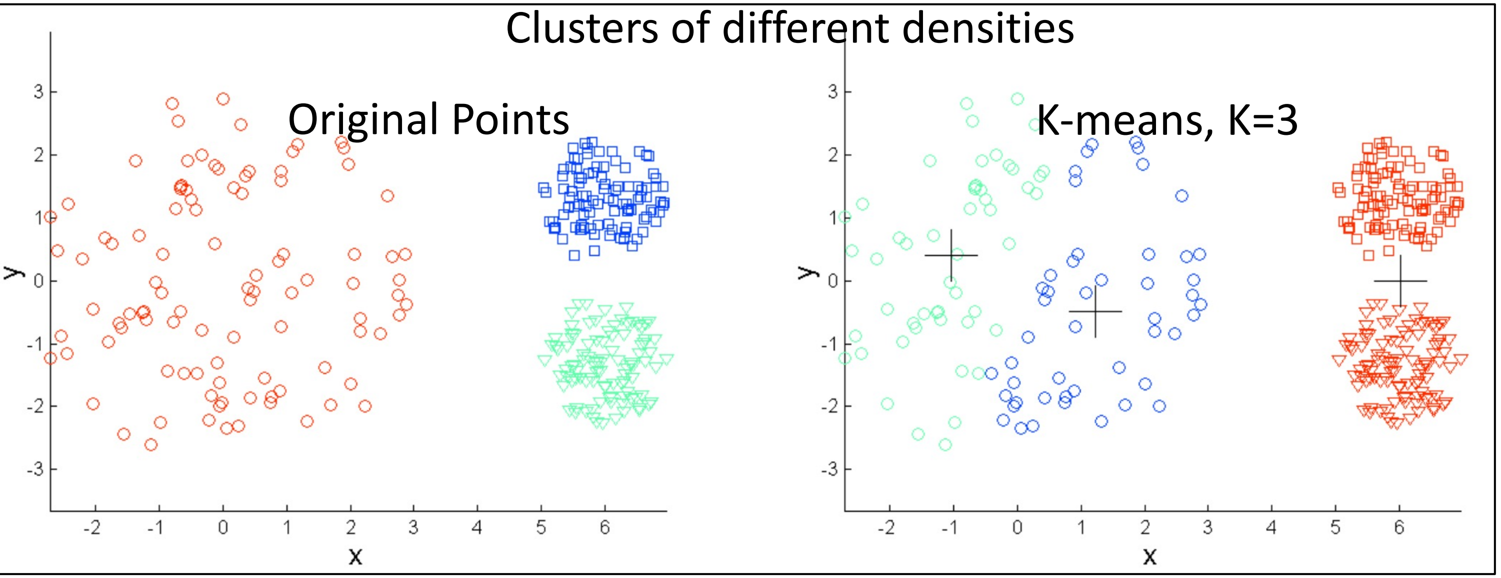
- Less susceptible to initialization problems than K-means
- Shown to converge on better clusters, better overcoming local minimums

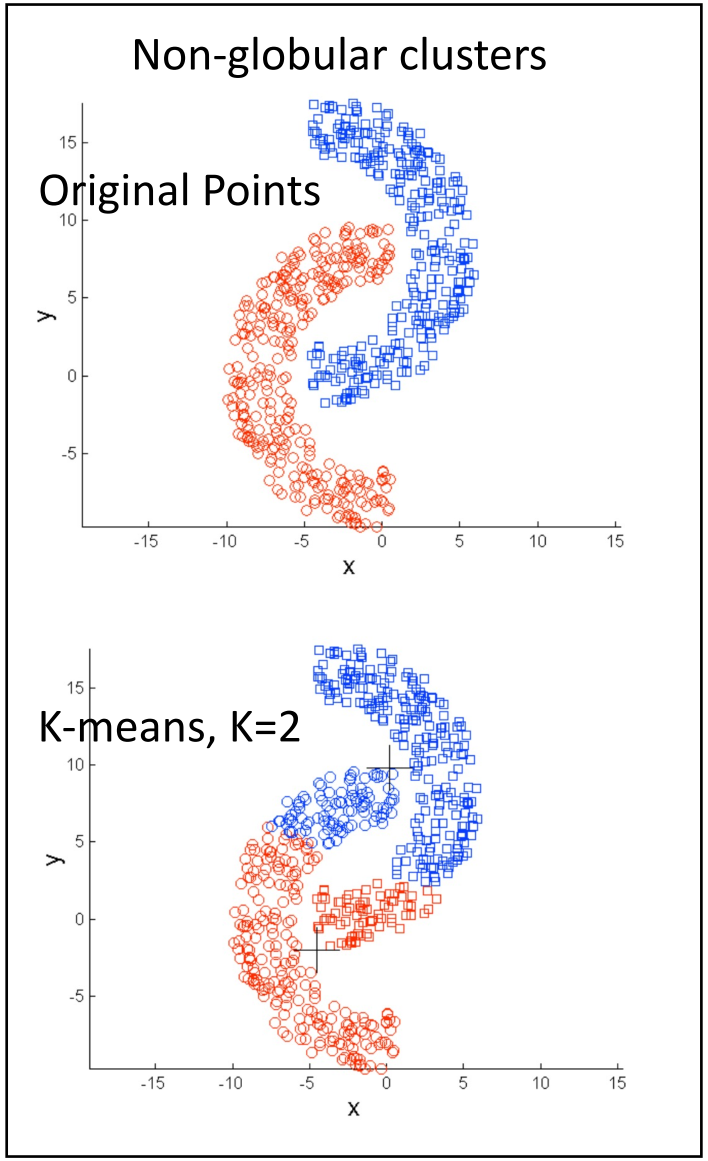
Characteristics of K-means

- Simple and can be used for a wide variety of data types
- Quite efficient, even when run multiple times
- There are variations (like Bisecting K-means) that are even more efficient and less susceptible to initialization problems
- Outliers can alter results, but outlier detection and removal can be done before clustering
- Curse of Dimensionality: As dimensionality increases, distance and similarity between points lose meaning
- Difficulties with non-globular data, clusters of different sizes, clusters of different densities

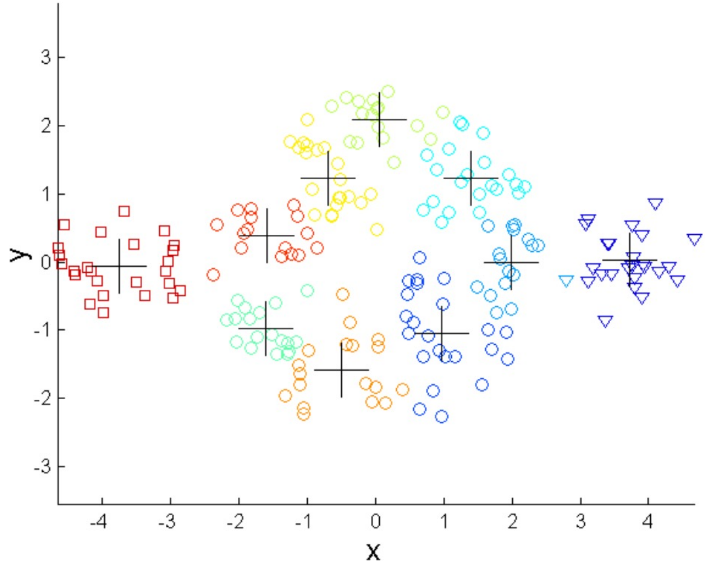
Weaknesses of K-means

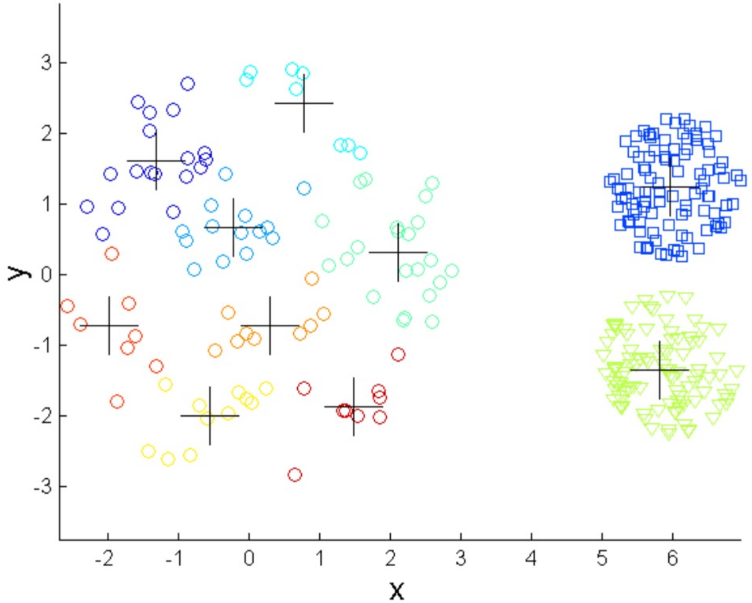


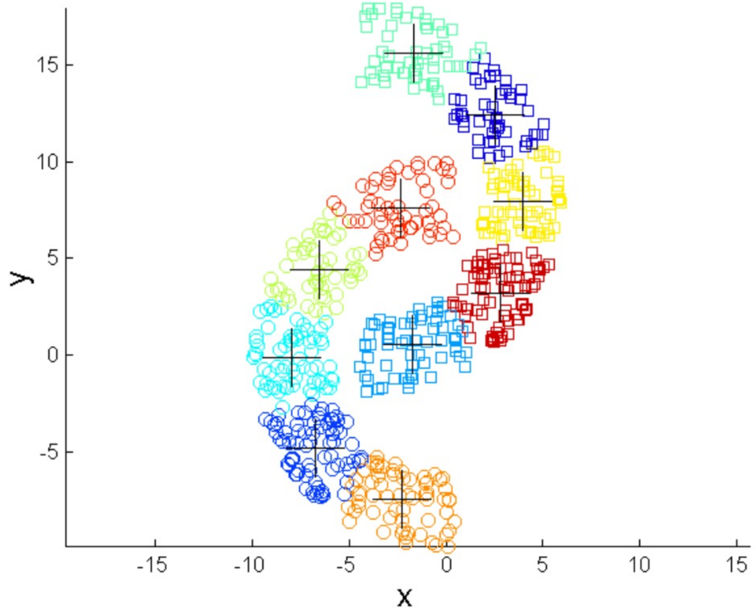




Increasing K to Overcome Weaknesses







(Backup) Terminology

- **Partitional Clustering**: division of dataset into non-overlapping subsets, such that each data object is in exactly one cluster
- **Hierarchical Clustering**: Clusters can have subclusters. Nested clusters are organized into a tree. Each node in the tree (except leaf nodes) is the union of its children.

- **Exclusive Clustering**: each object is assigned only to a single cluster
- **Overlapping/Non-exclusive clustering**: an object can simultaneously belong to more than one cluster
- **Fuzzy/Probabilistic Clustering**: every object belongs to every cluster with a membership weight from 0 to 1

- **Complete Clustering**: assigns every object to a cluster
- **Partial Clustering**: not every object is necessarily assigned to a cluster

(Backup) K-means Example

Given this dataset with one attribute, cluster it using k-means with k=2.
Use data points A and B as the initial centroids.

	A	B	C	D	E
X1	1	6	8	20	30