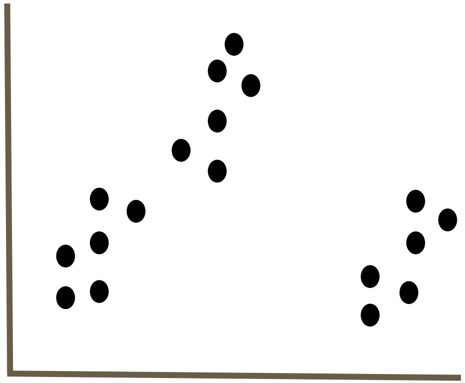
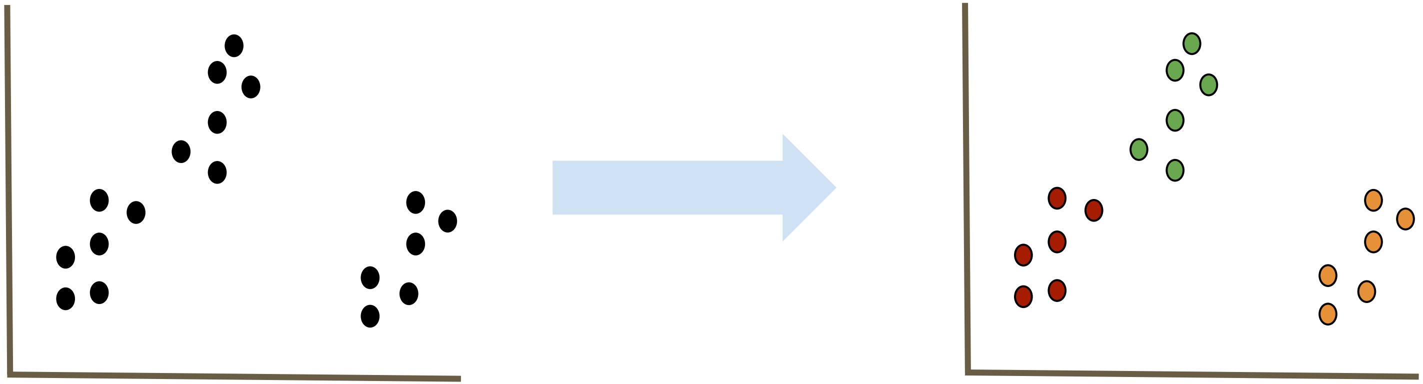
Clustering - Kmeans

Boston University CS 506 - Lance Galletti

What is a Clustering



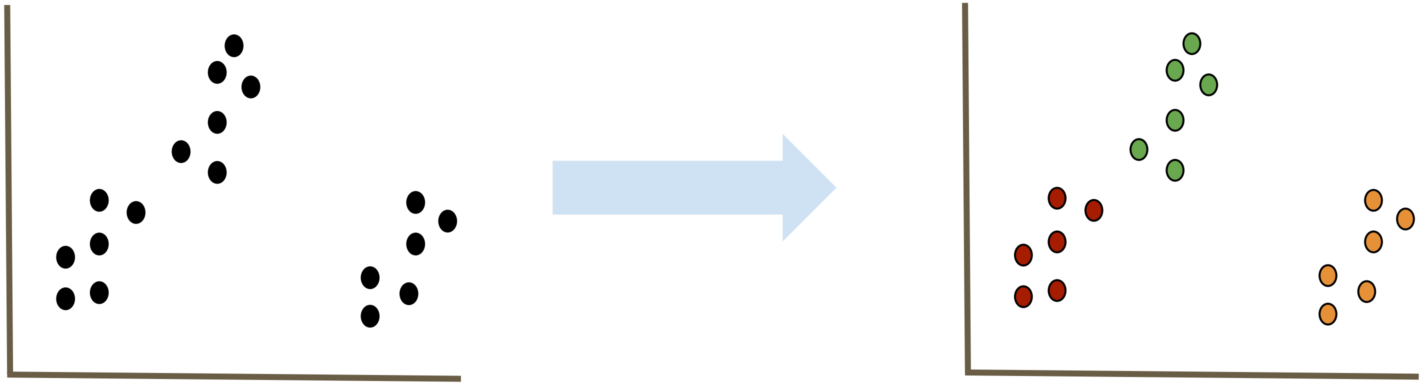
What is a Clustering



What is a Clustering

A clustering is a grouping / assignment of objects (data points) such that objects in the same group / cluster are:

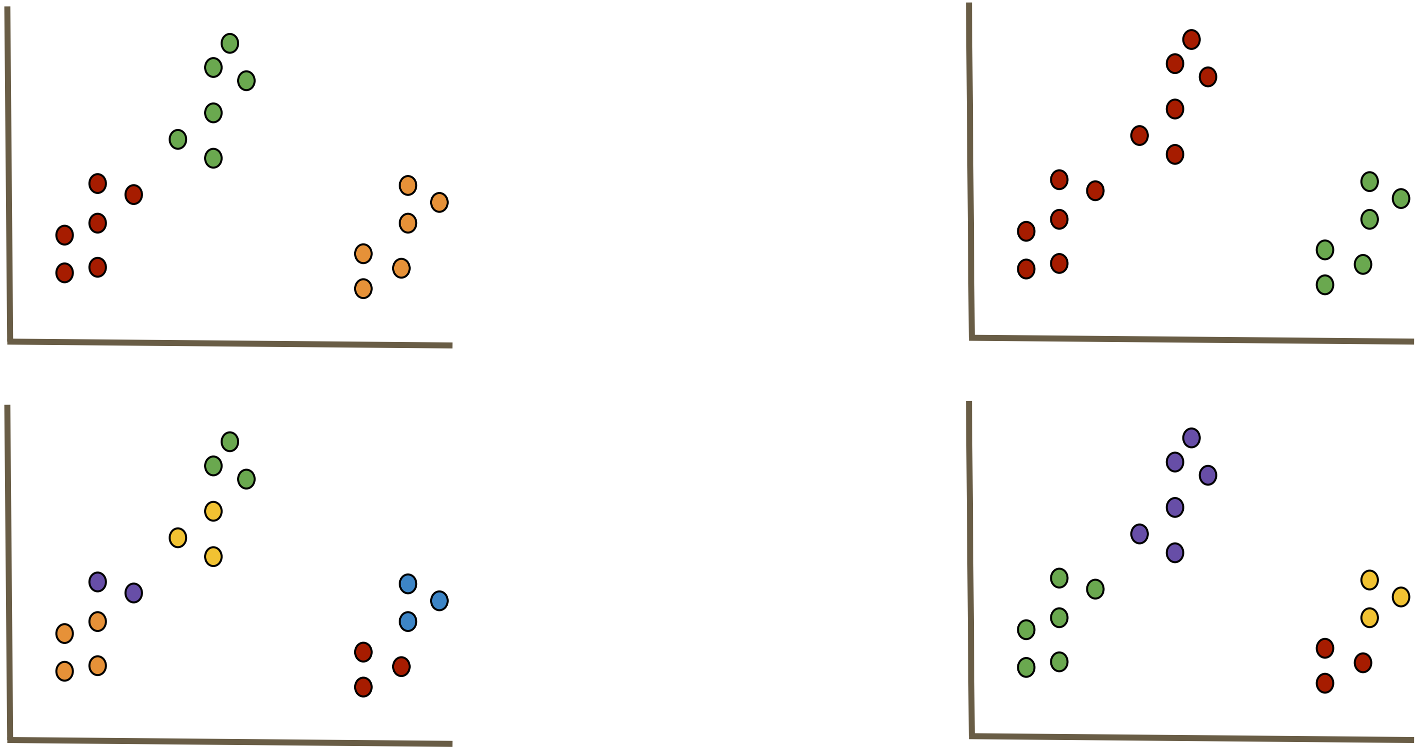
- similar to one another
- dissimilar to objects in other groups



Applications

- Outlier detection / anomaly detection
  - Data Cleaning / Processing
  - Credit card fraud, spam filter etc.
- Feature Extraction
- Filling Gaps in your data
  - Using the same marketing strategy for similar people
  - Infer probable values for gaps in the data (similar users could have similar hobbies, likes / dislikes etc.)

Clusters can be Ambiguous



Types of Clusterings

Partitional

Each object belongs to exactly one cluster

Hierarchical

A set of nested clusters organized in a tree

Density-Based

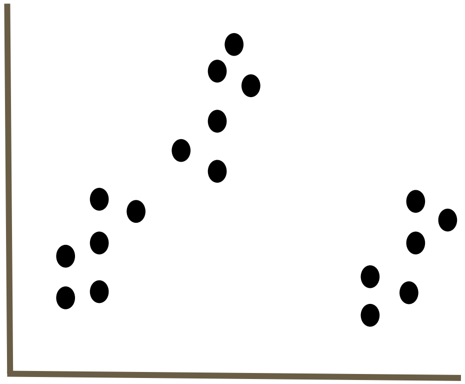
Defined based on the local density of points

Soft Clustering

Each point is assigned to every cluster with a certain probability

Partitional Clustering

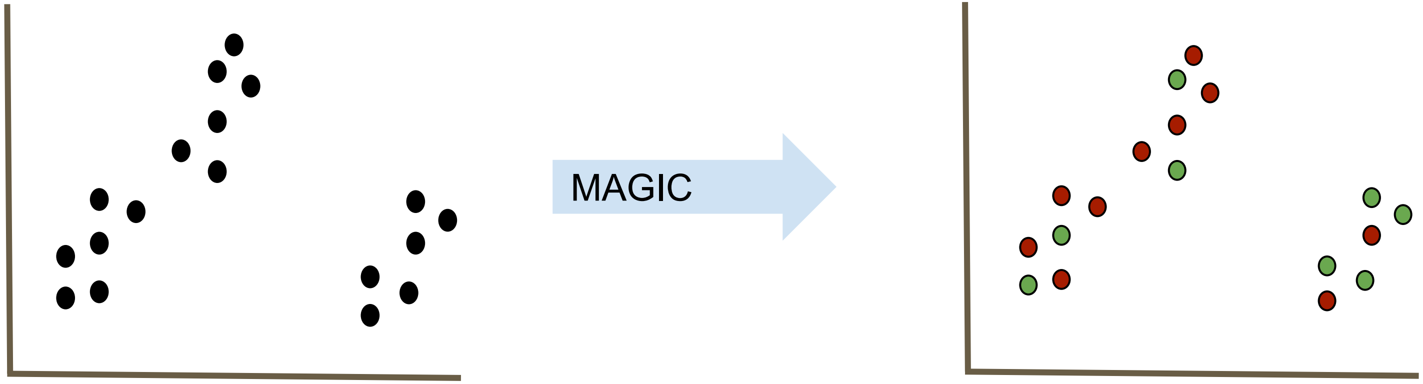
Partitional Clustering



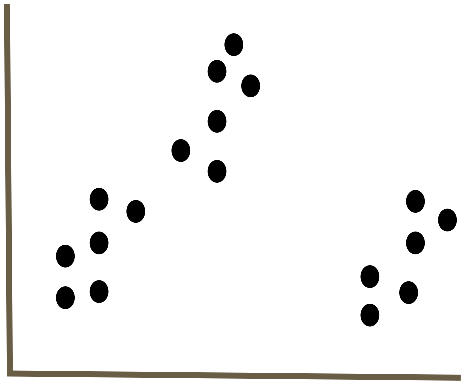
Goal: partition dataset into k partitions

Partitional Clustering

Goal: partition dataset into k partitions



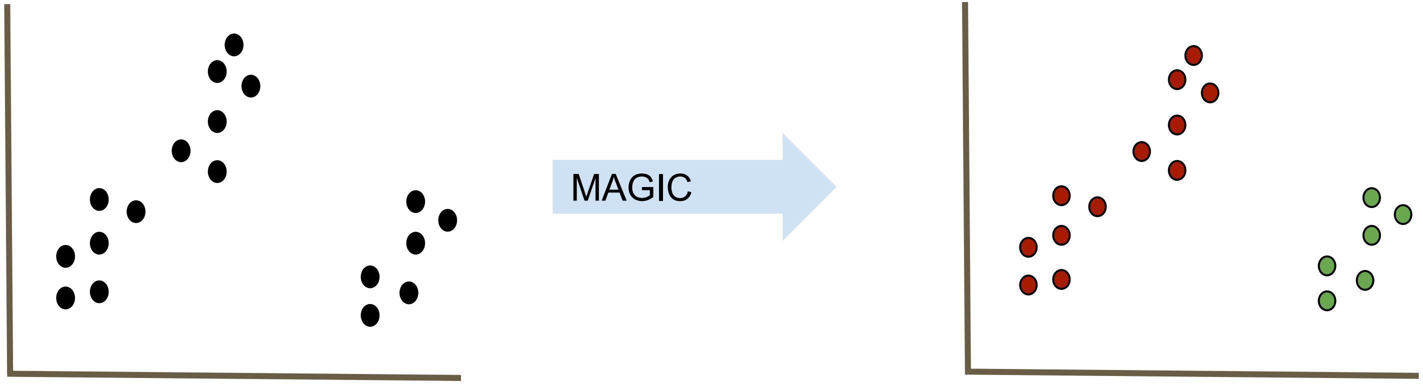
Partitional Clustering



Goal: partition dataset into k partitions

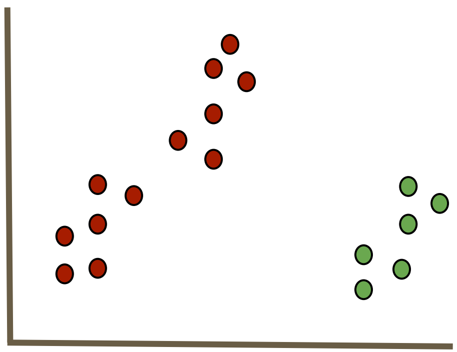
Partitional Clustering

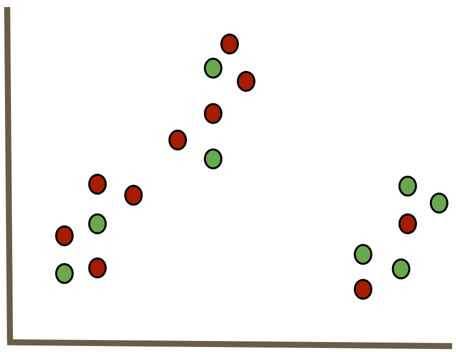
Goal: partition dataset into k partitions



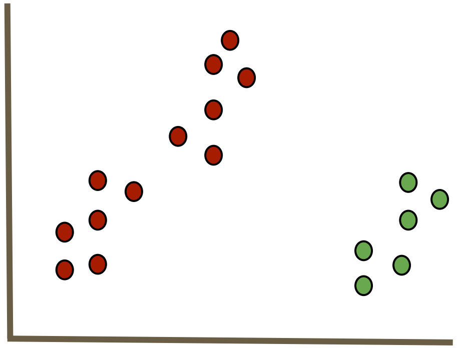
MAGIC

Partitional Clustering

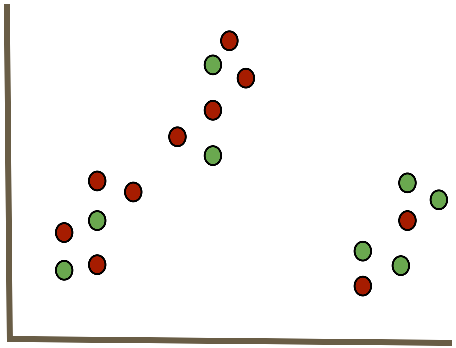




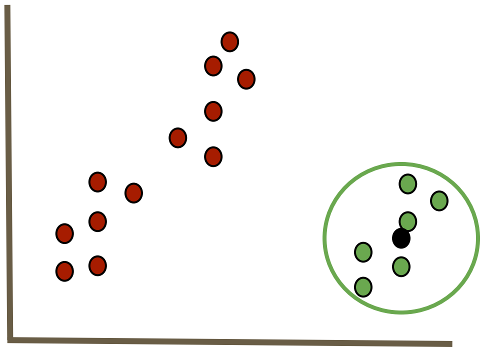
Partitional Clustering



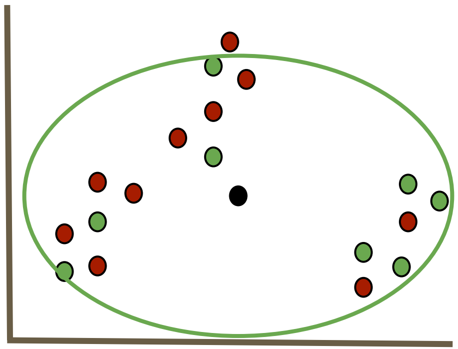
VS



Example

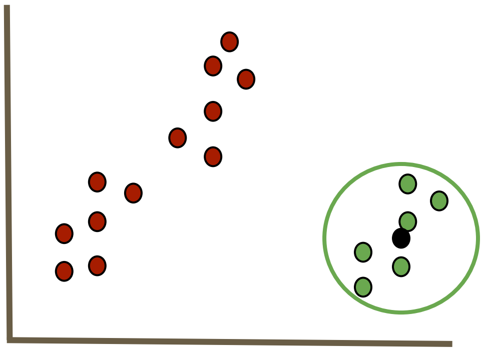


VS

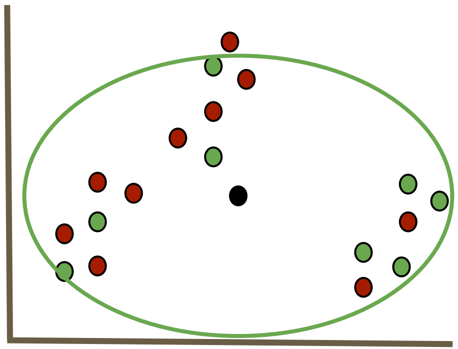


Given a distance function **d**, we can find points (not necessarily part of our dataset) for each cluster called **centroids** that are at the center of each cluster.

Example



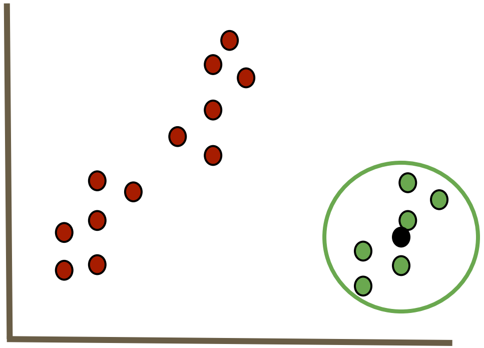
VS



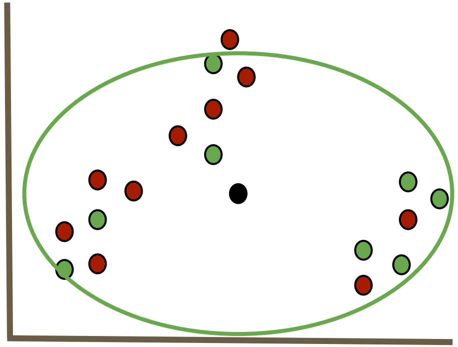
Q: When **d** is Euclidean, what is the **centroid** (also called **center of mass**) of **m** points $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$?

A: The mean / average of the points

Example



VS



Looking at the sum of the distances of points in a cluster to its centroid also captures the “spread” (variance) of a cluster

$$\sum_{i}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

Mean of cluster i

Cluster i

Cost Function

- Way to evaluate and compare solutions
- Hope: can find some algorithm that find solutions that make the cost small

Q: Can you suggest a cost function to use for partitional clustering?

$$\sum_{i}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

K-means

Given **X = {x_1, … , x_n}** our dataset and **k**

Find **k** points **{μ_1, … , μ_k}** that minimize the **cost function**:

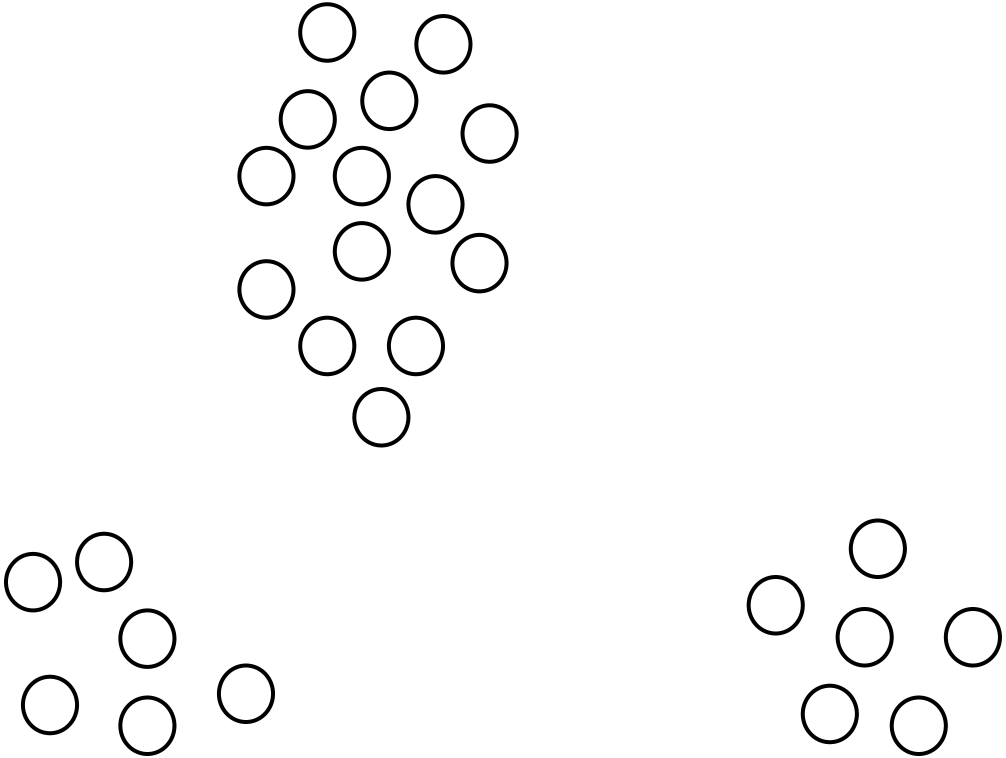
$$\sum_{i}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

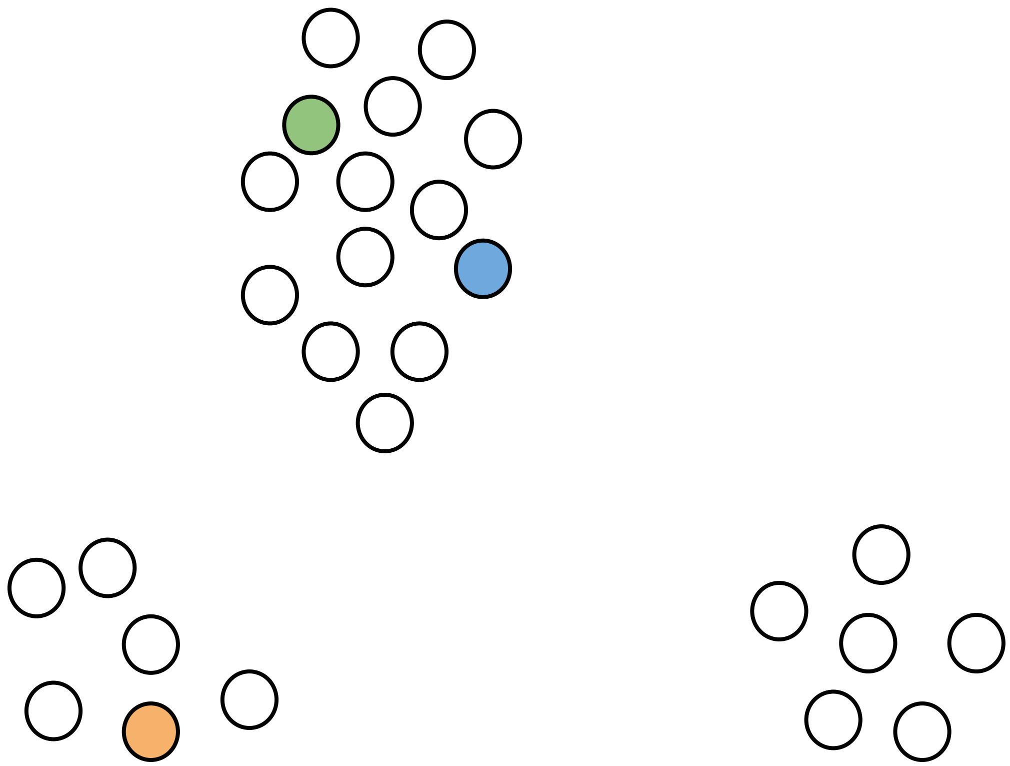
When **k=1** and **k=n** this is easy. Why?

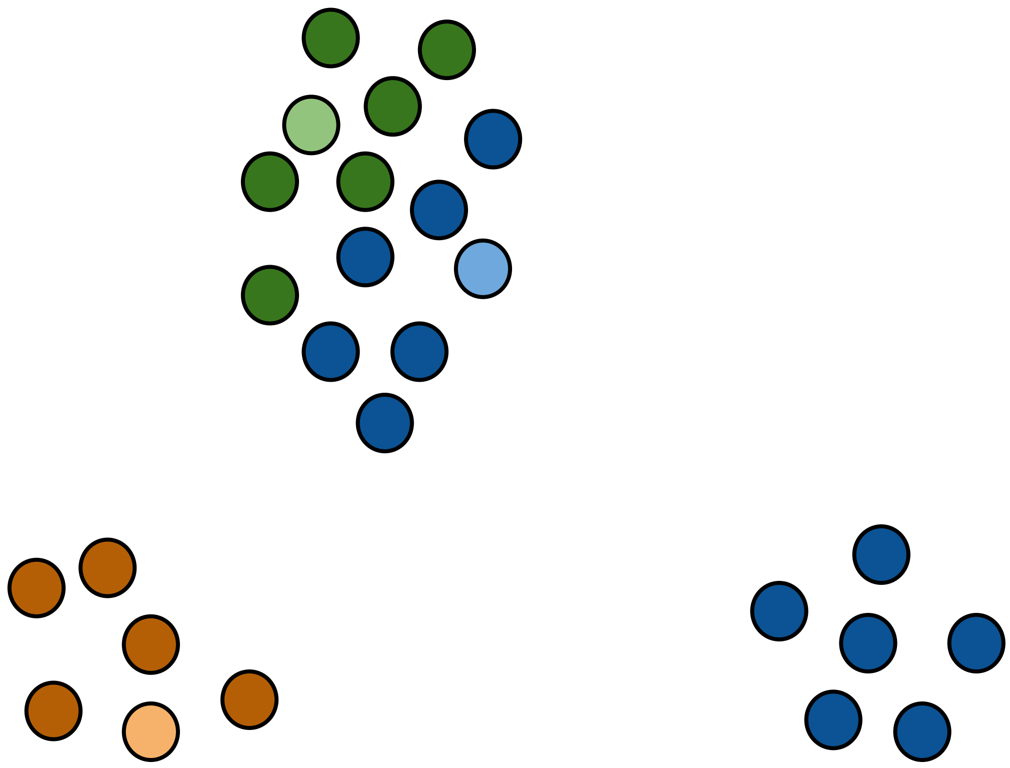
When **x_i** lives in more than 2 dimensions, this is a very difficult (**NP-hard**) problem

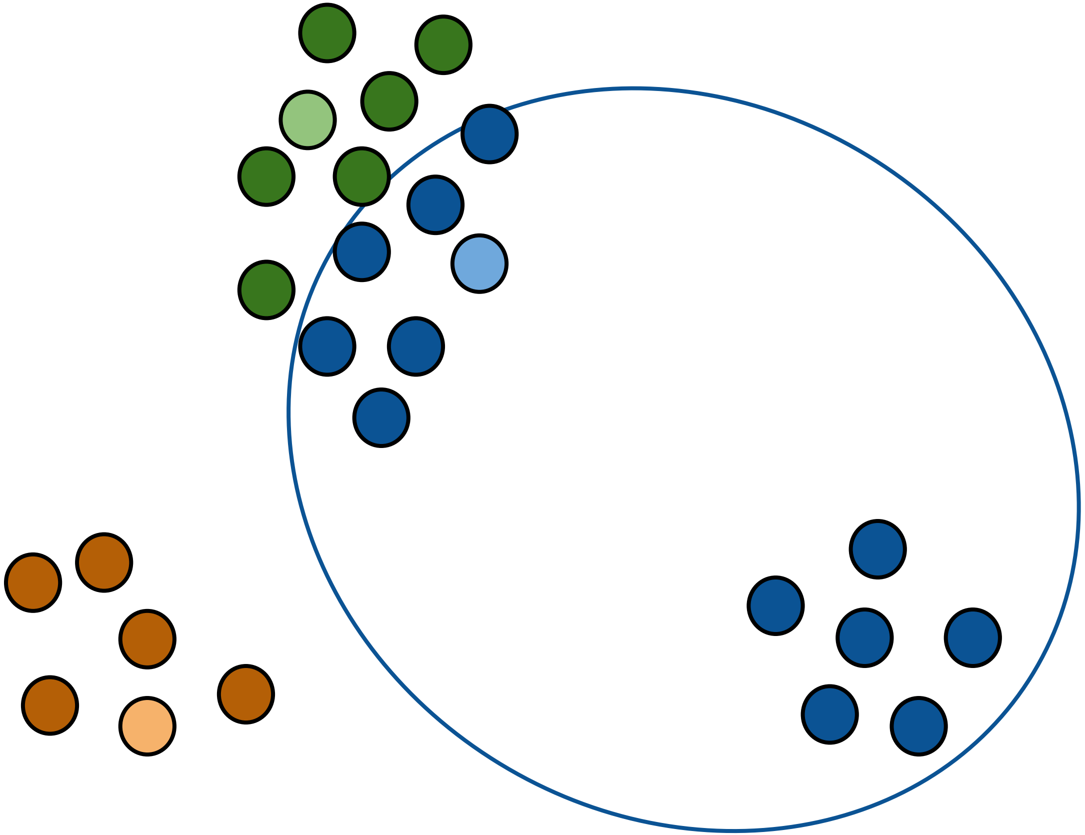
K-means - Lloyd's Algorithm

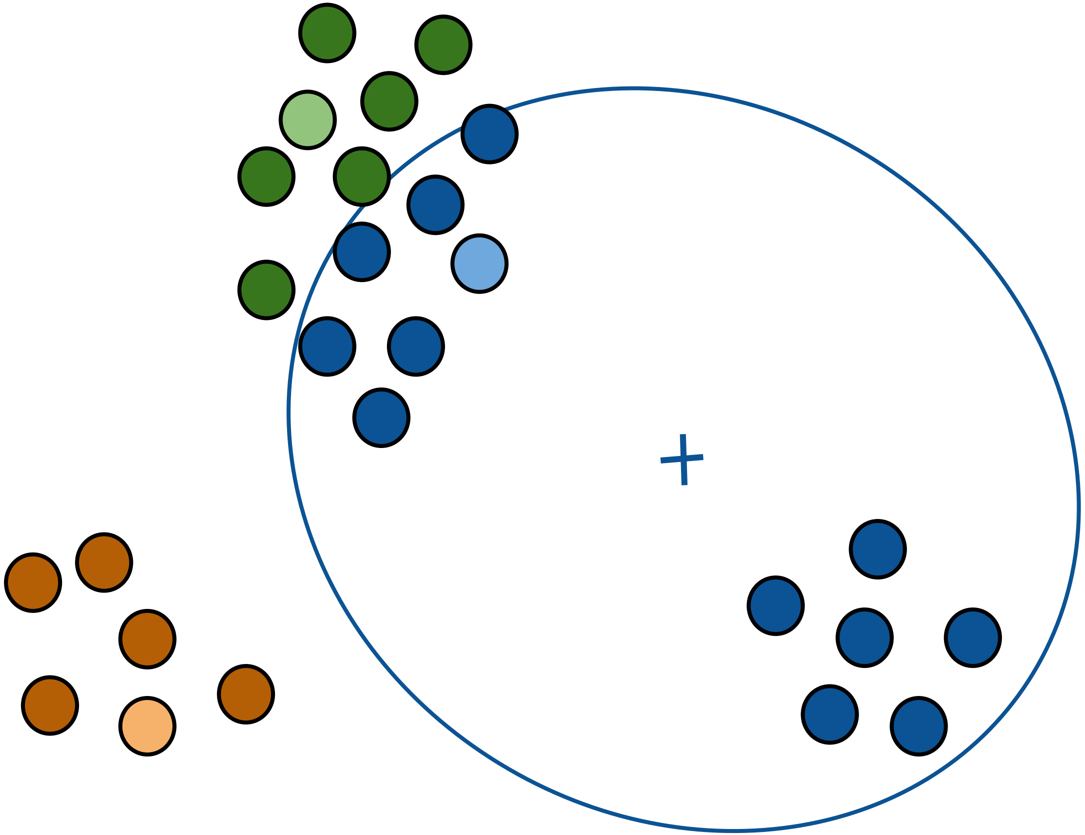
1. Randomly pick **k** centers $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k\}$
2. Assign each point in the dataset to its closest center
3. Compute the new centers as the means of each cluster
4. Repeat 2 & 3 until convergence

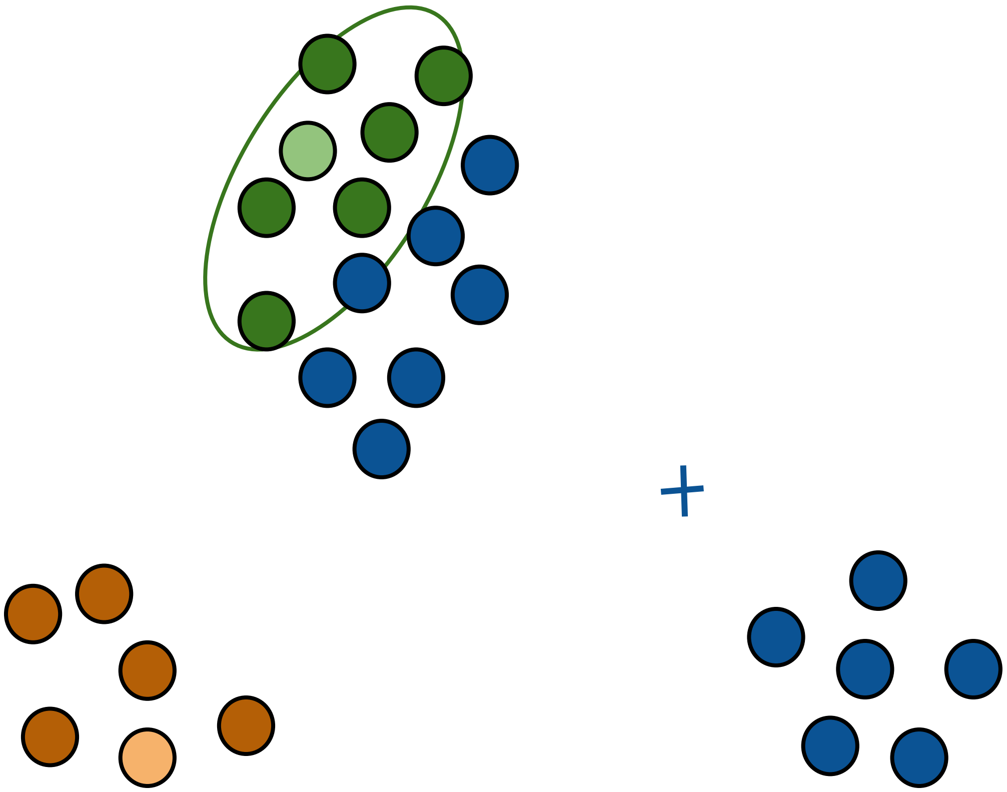


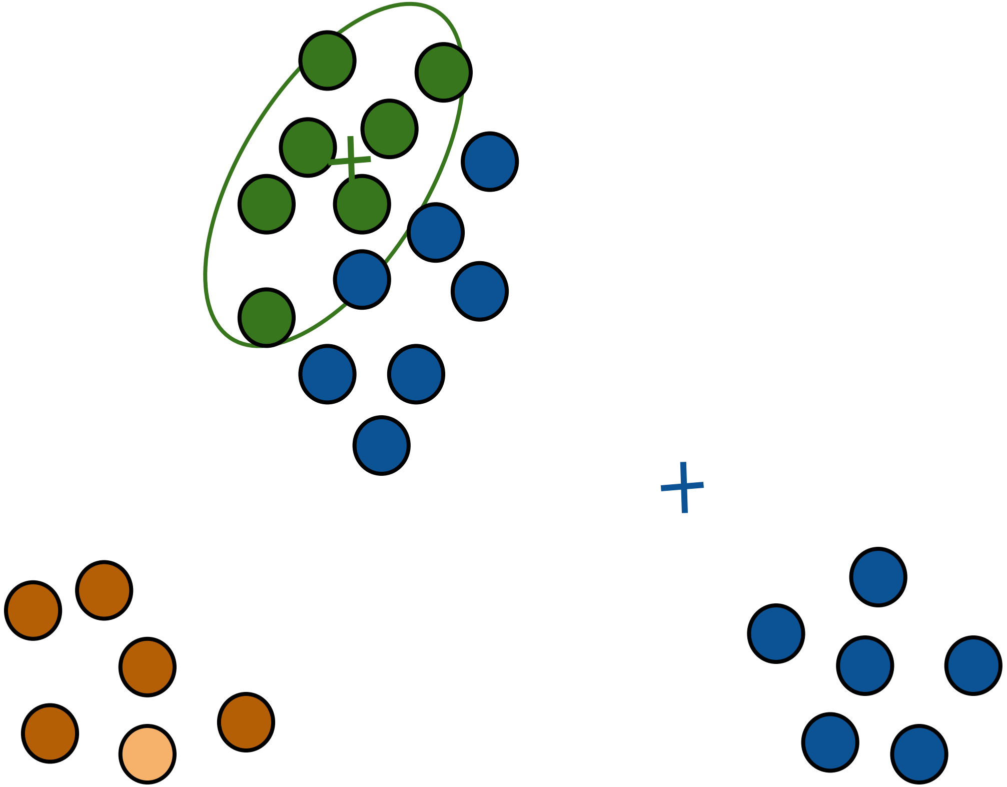


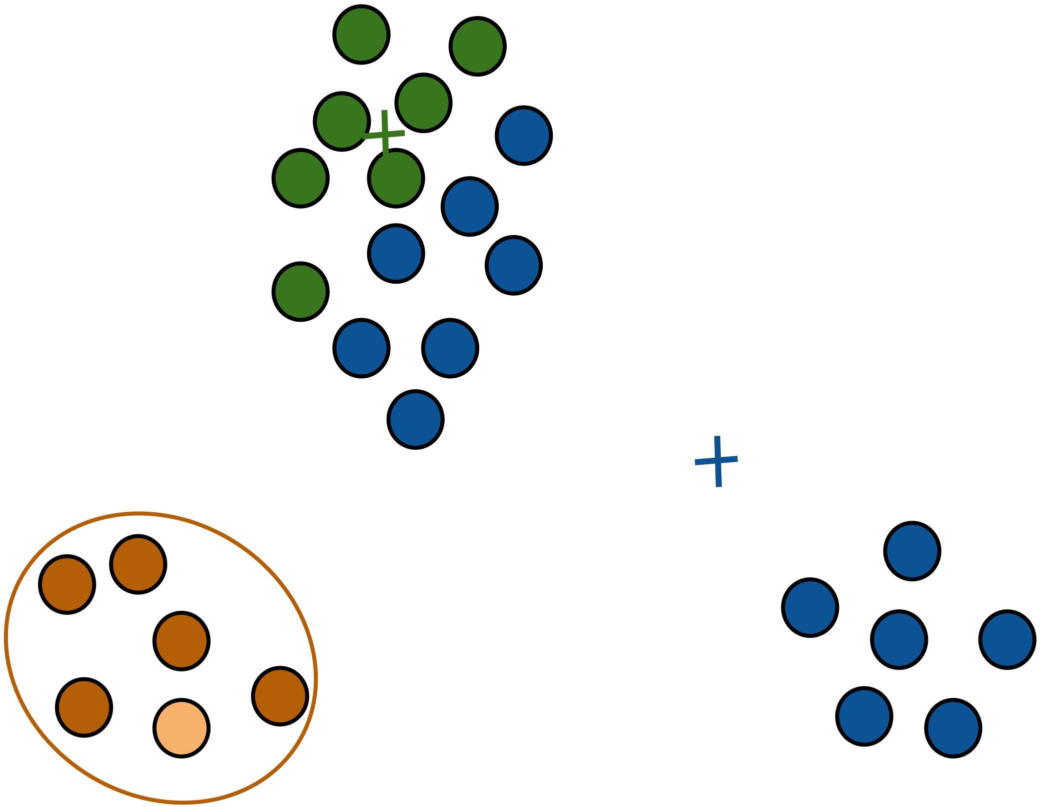


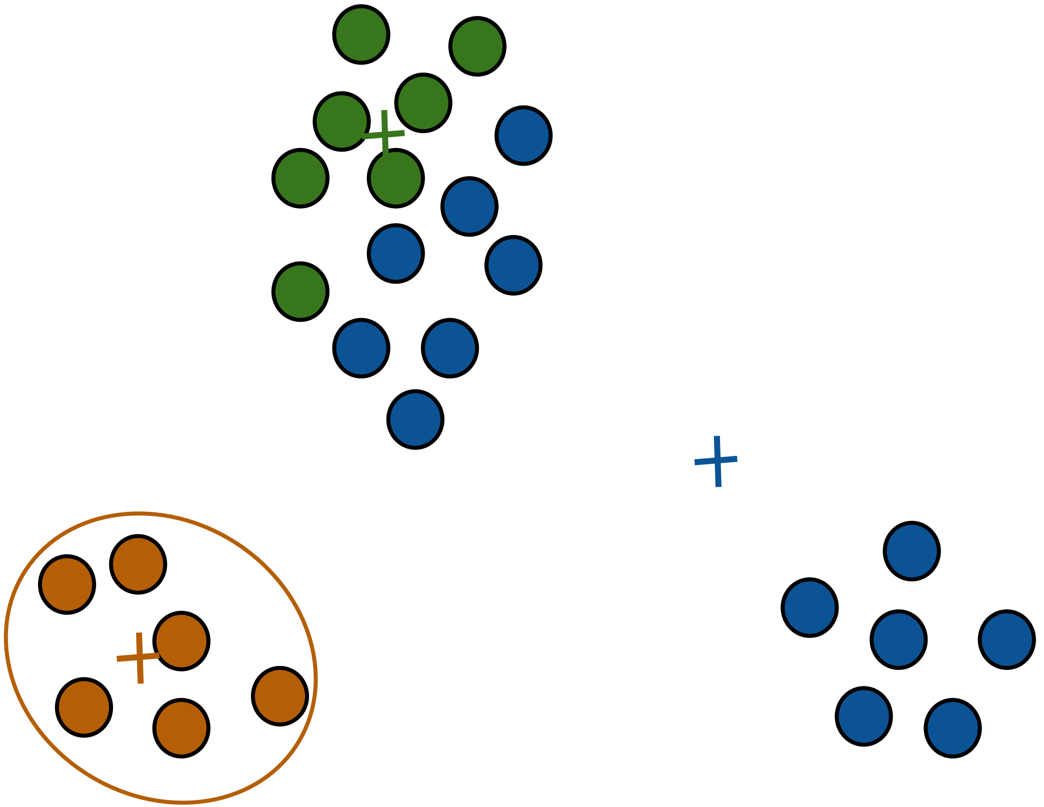


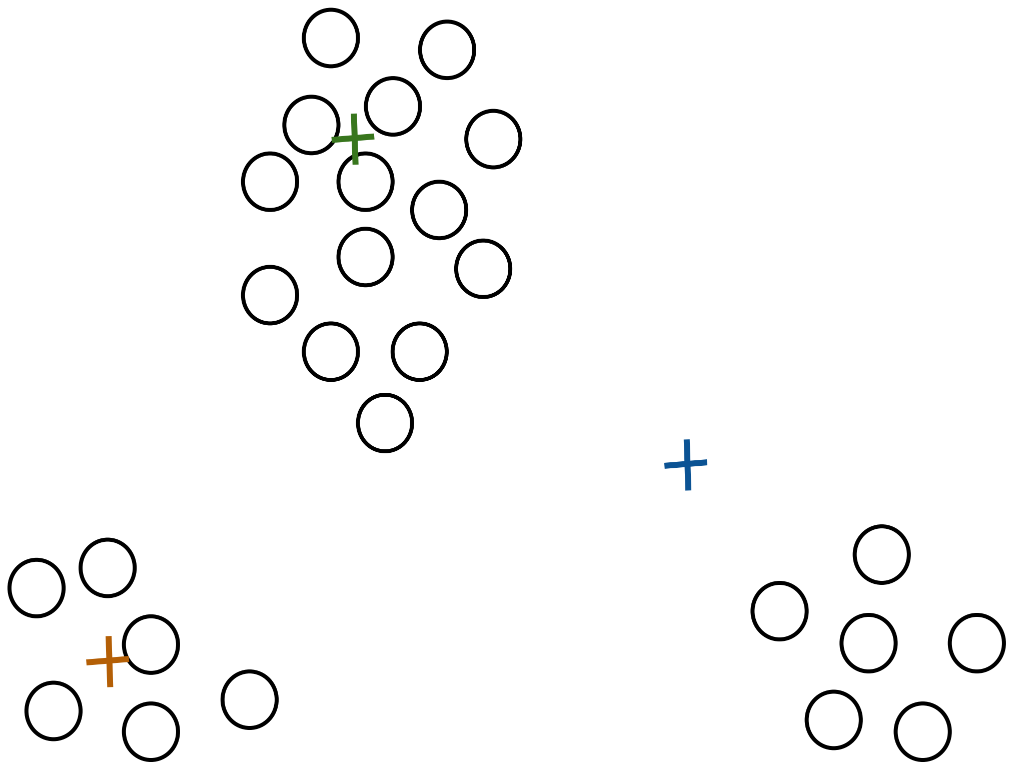


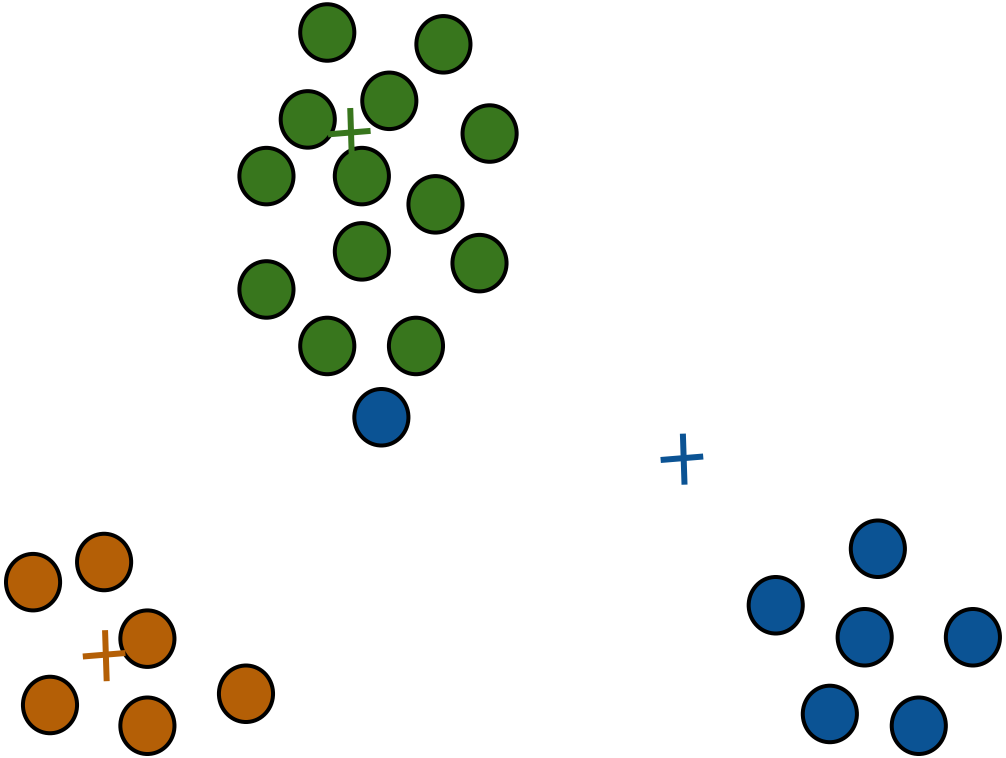


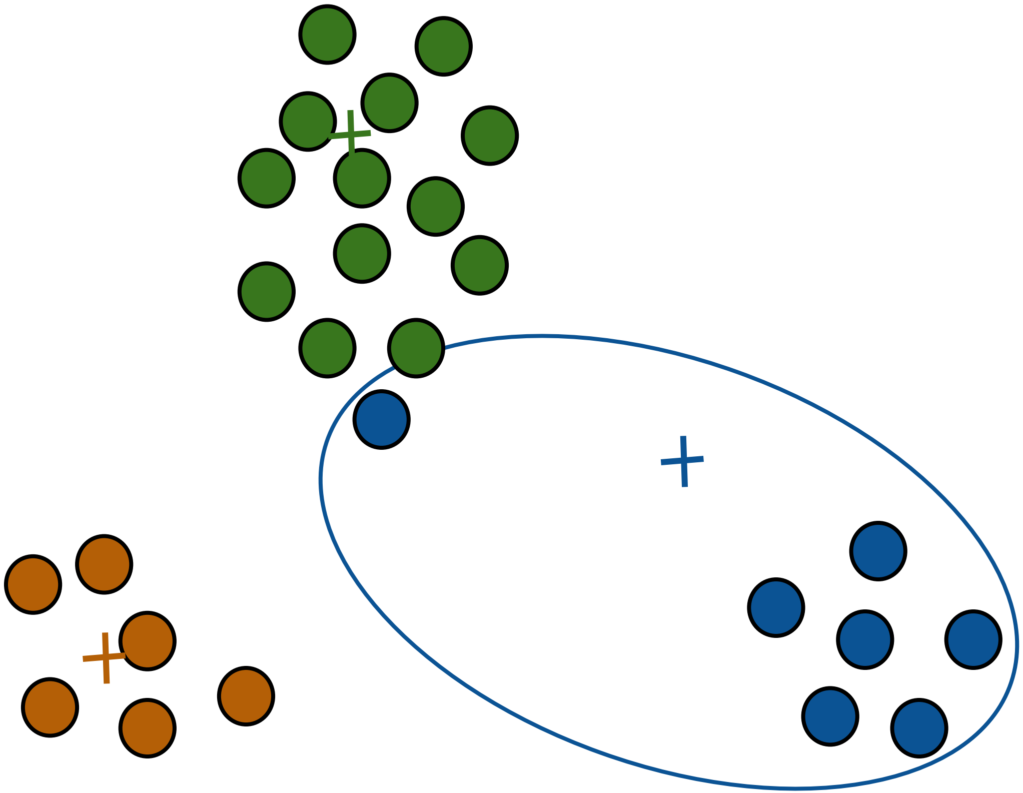


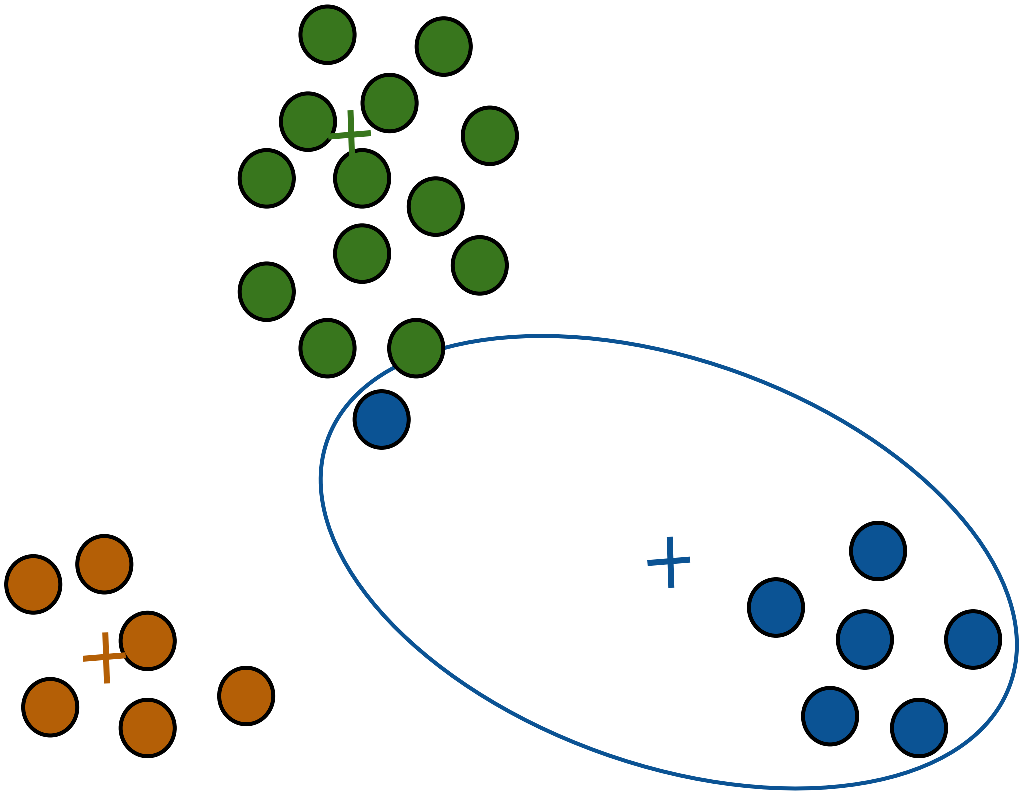


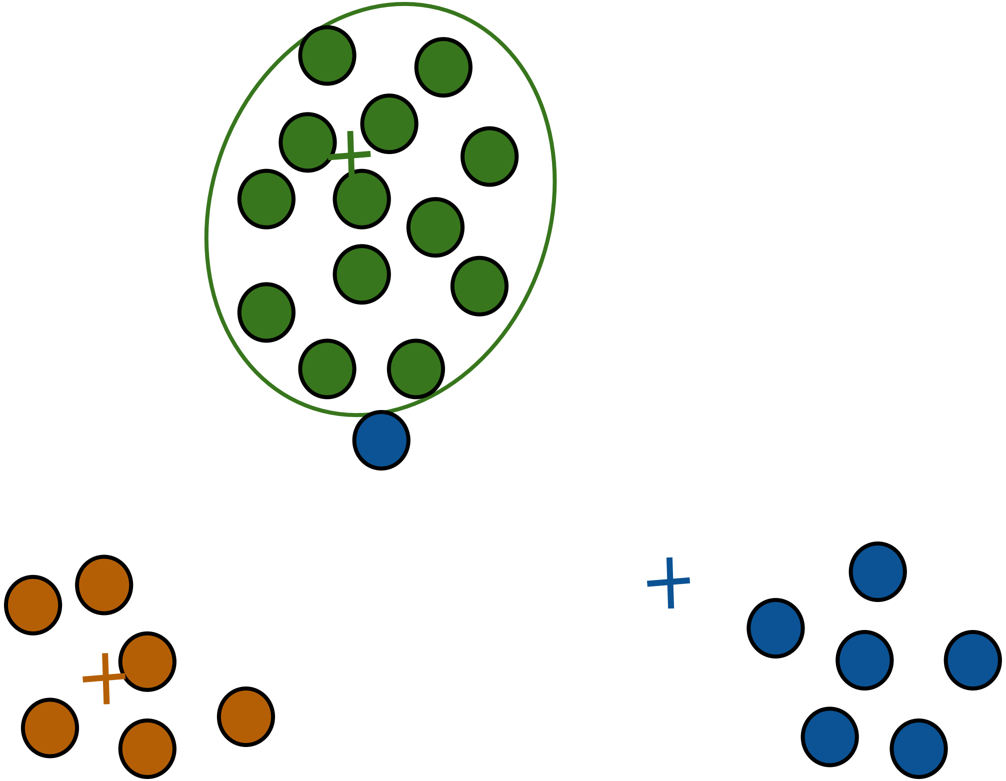


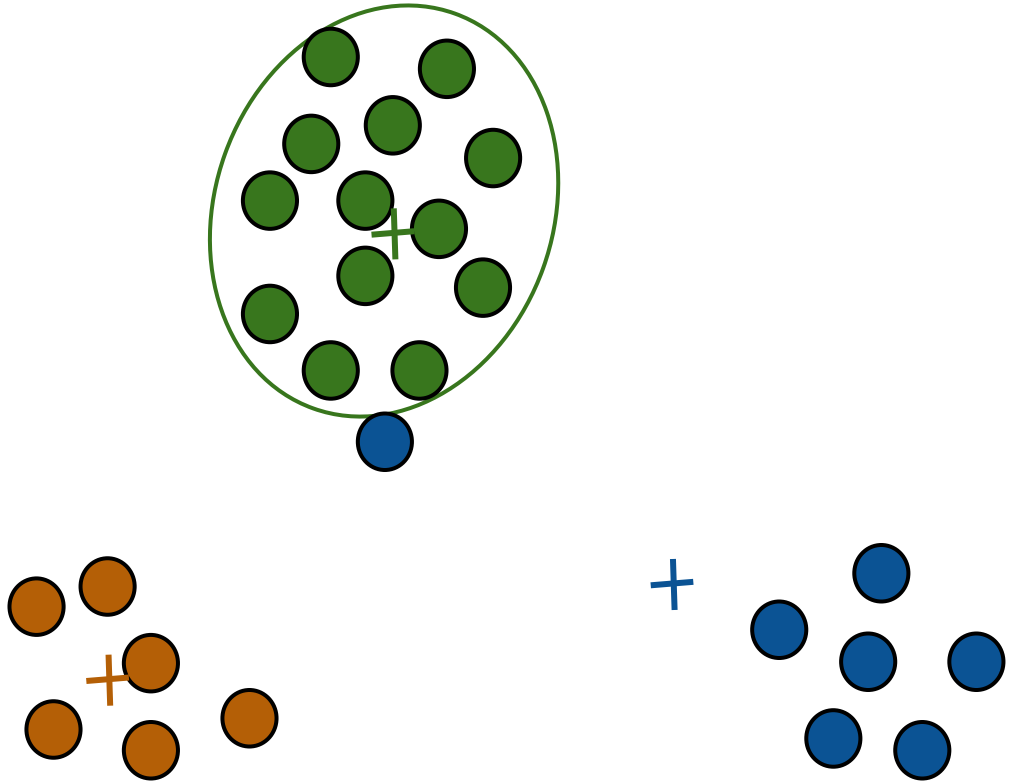


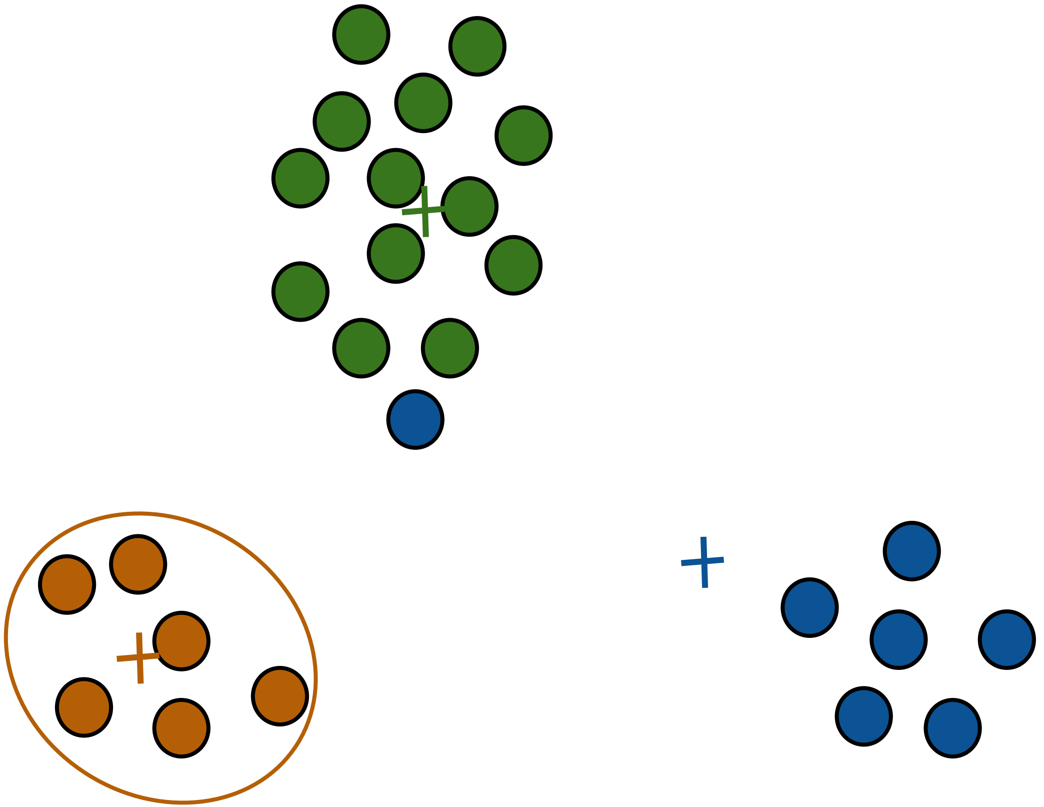


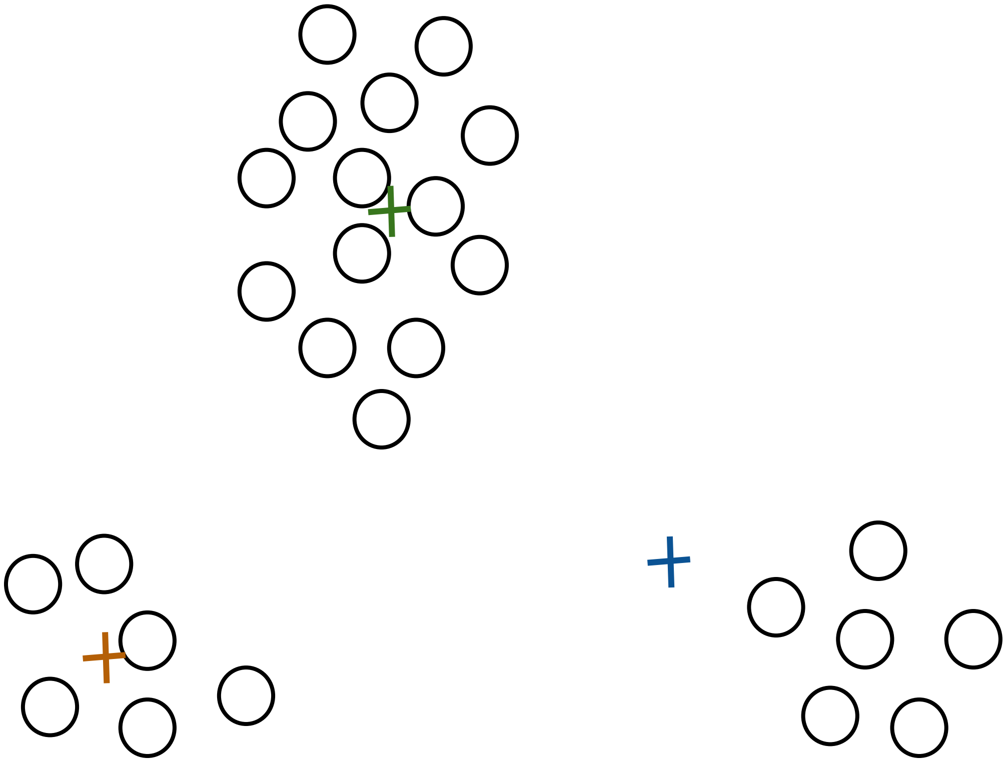


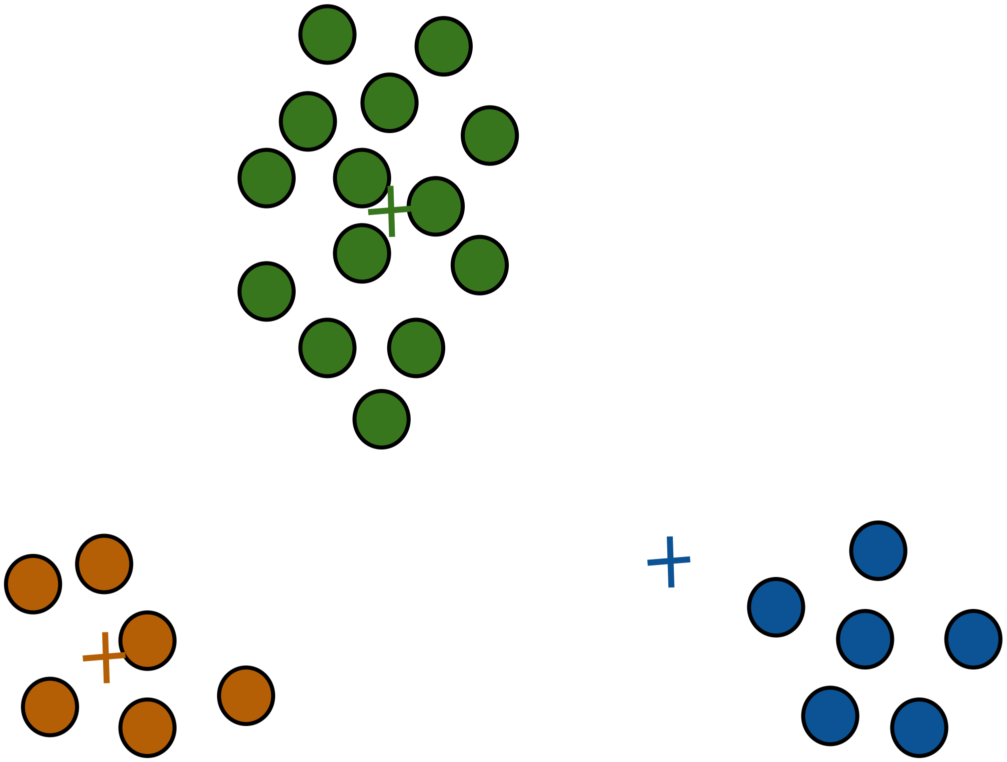


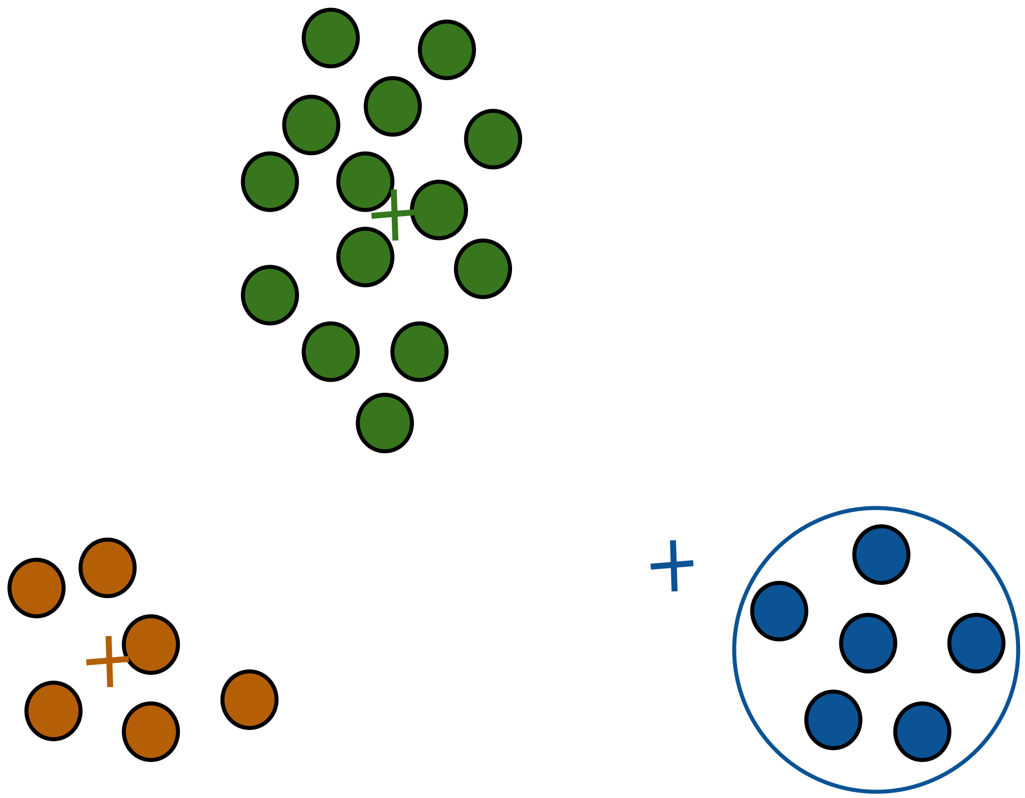


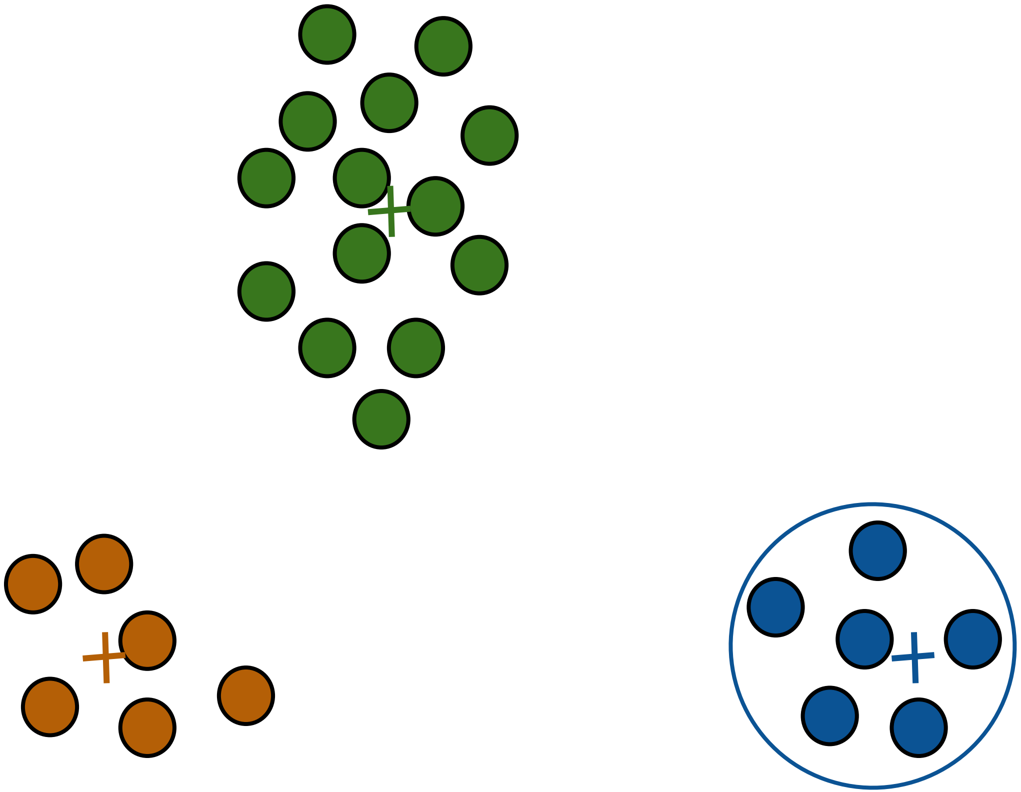


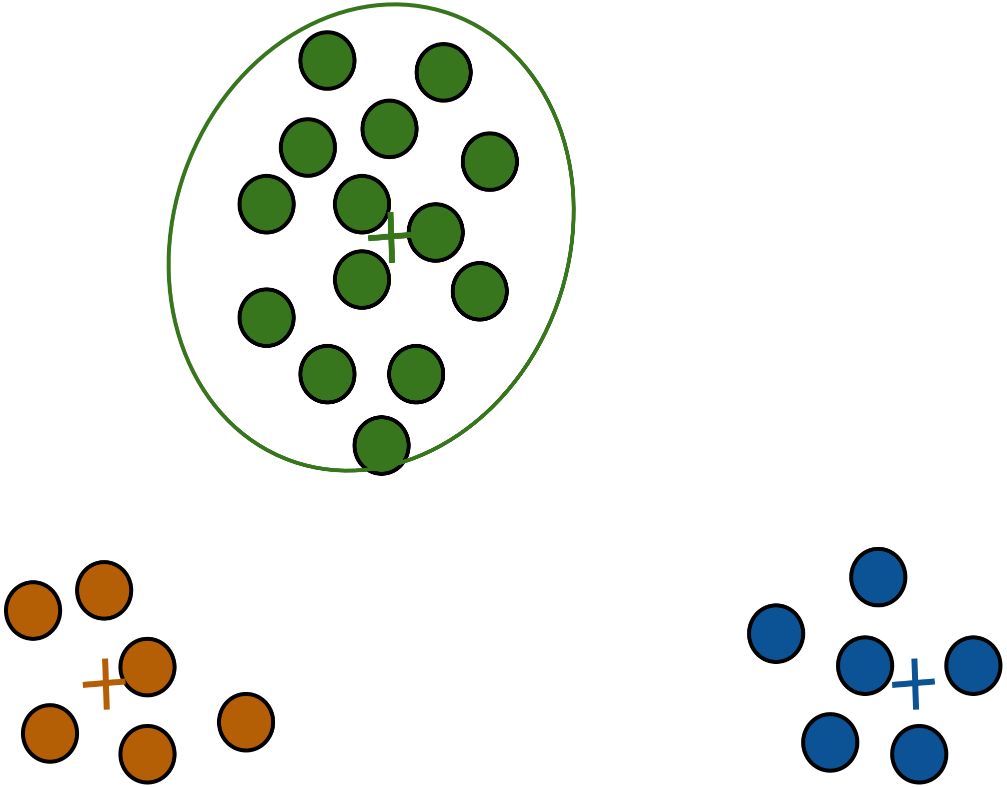


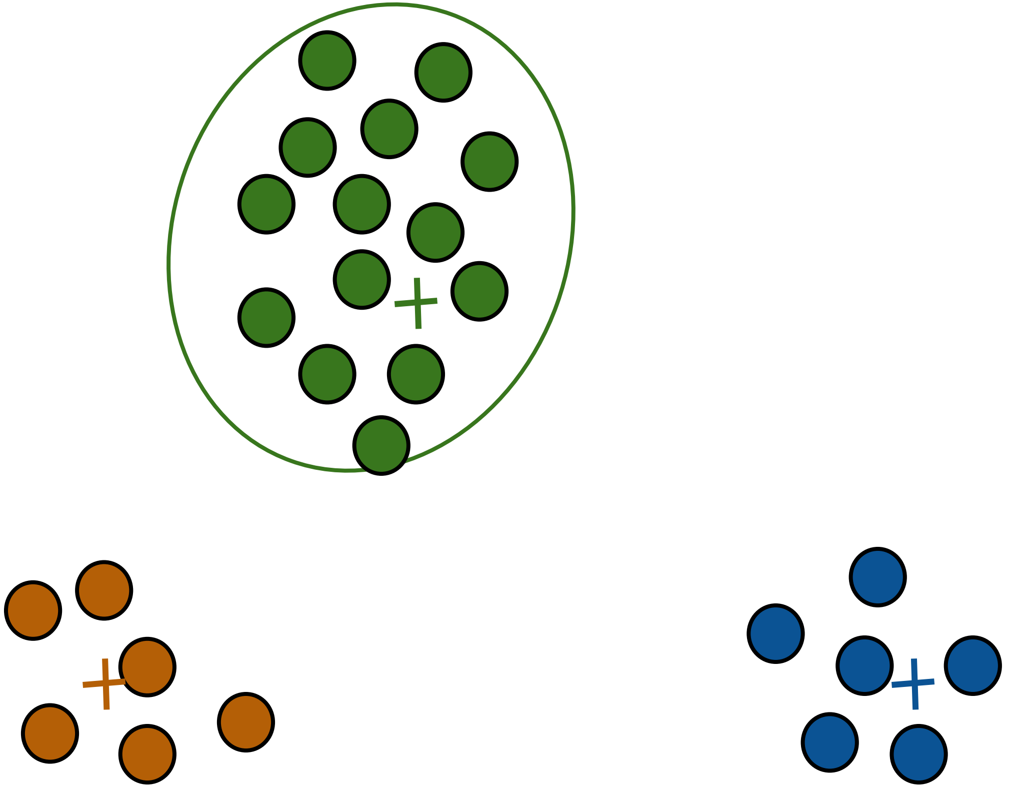


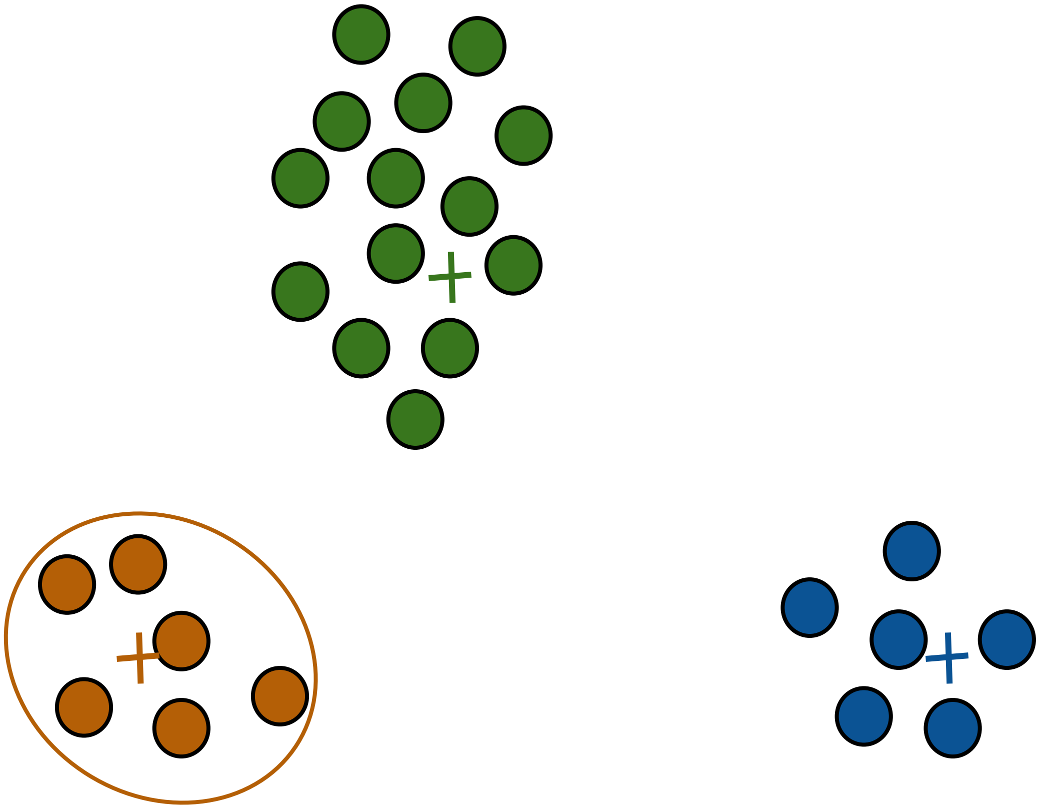


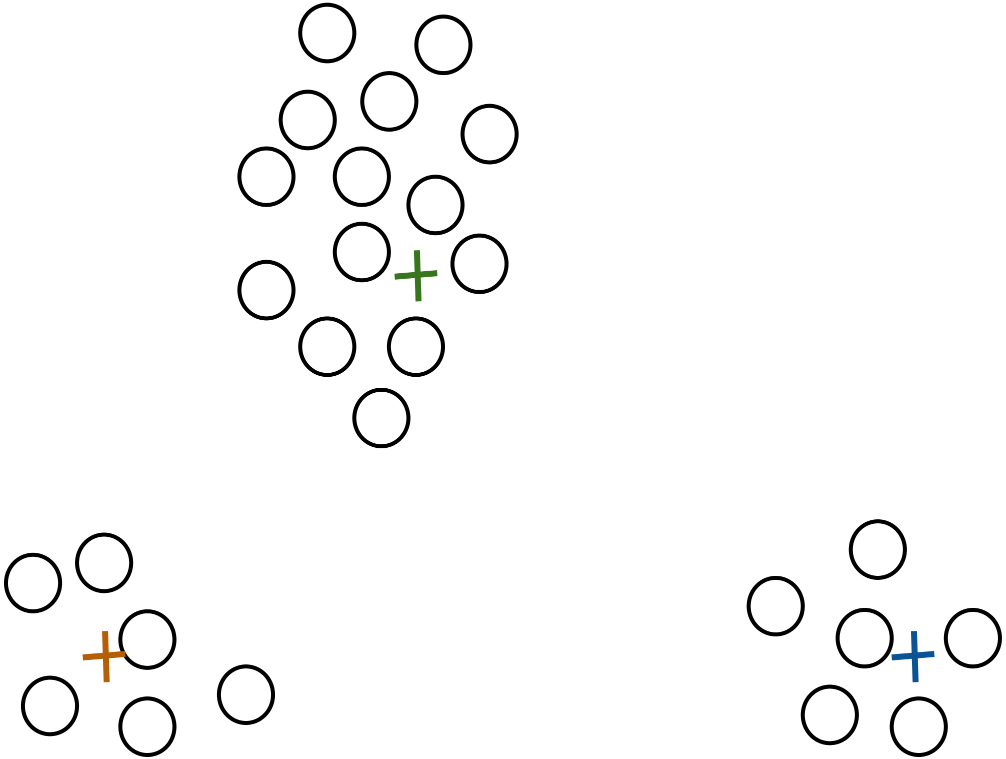


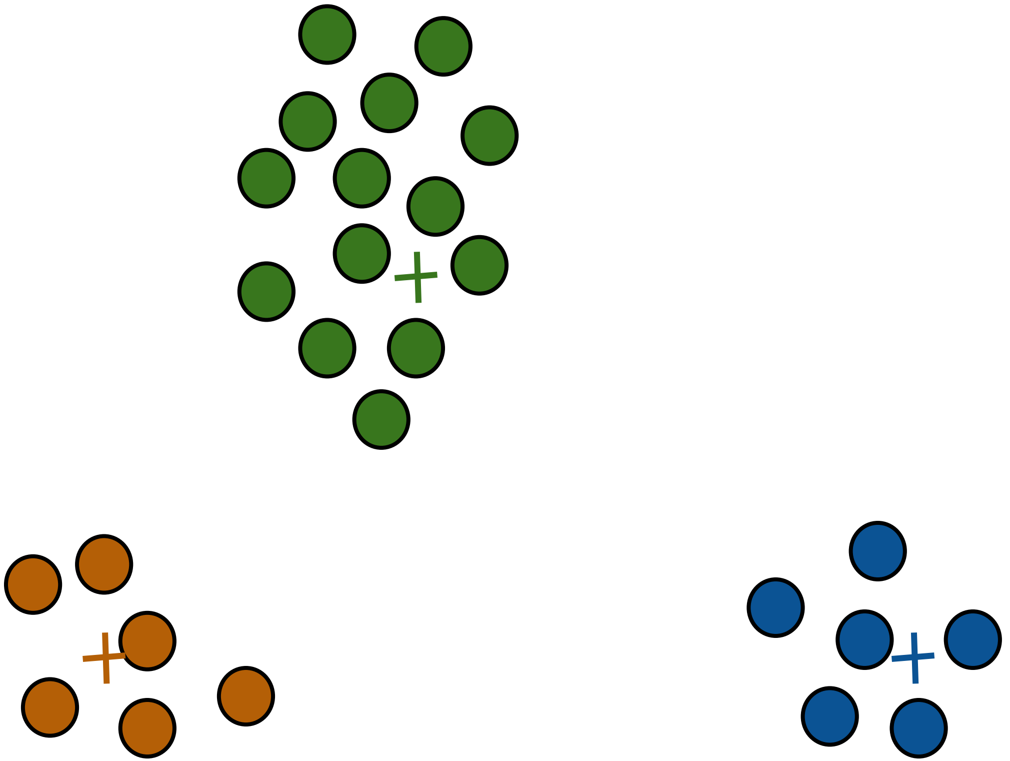


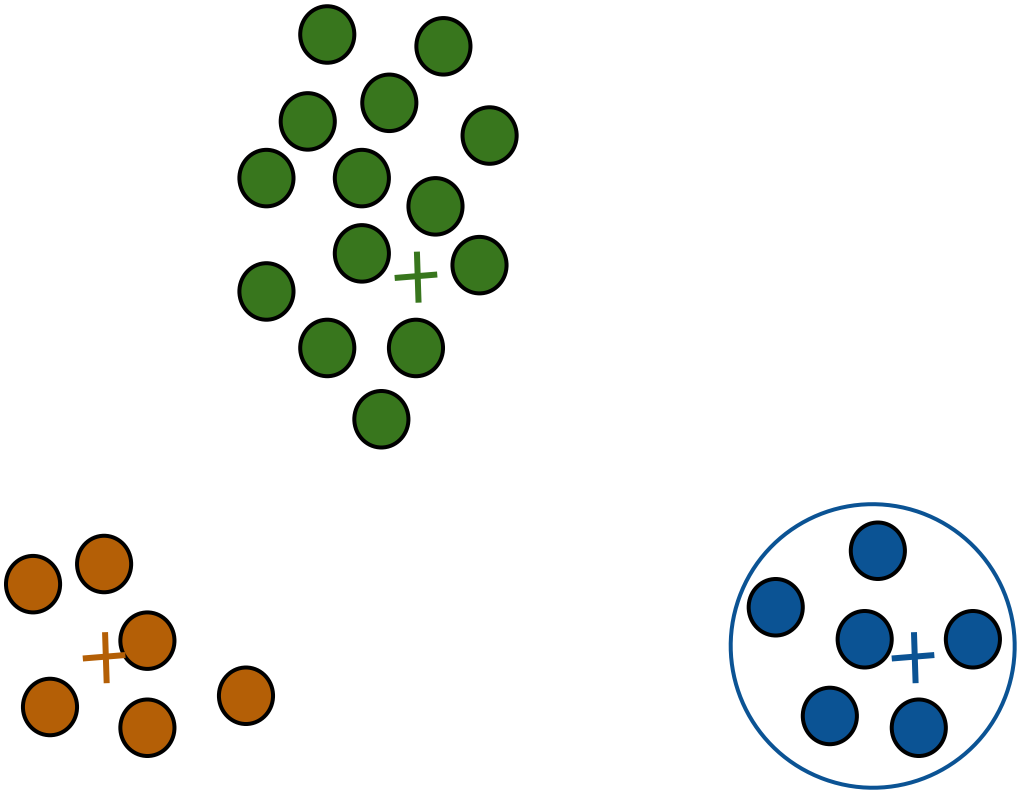


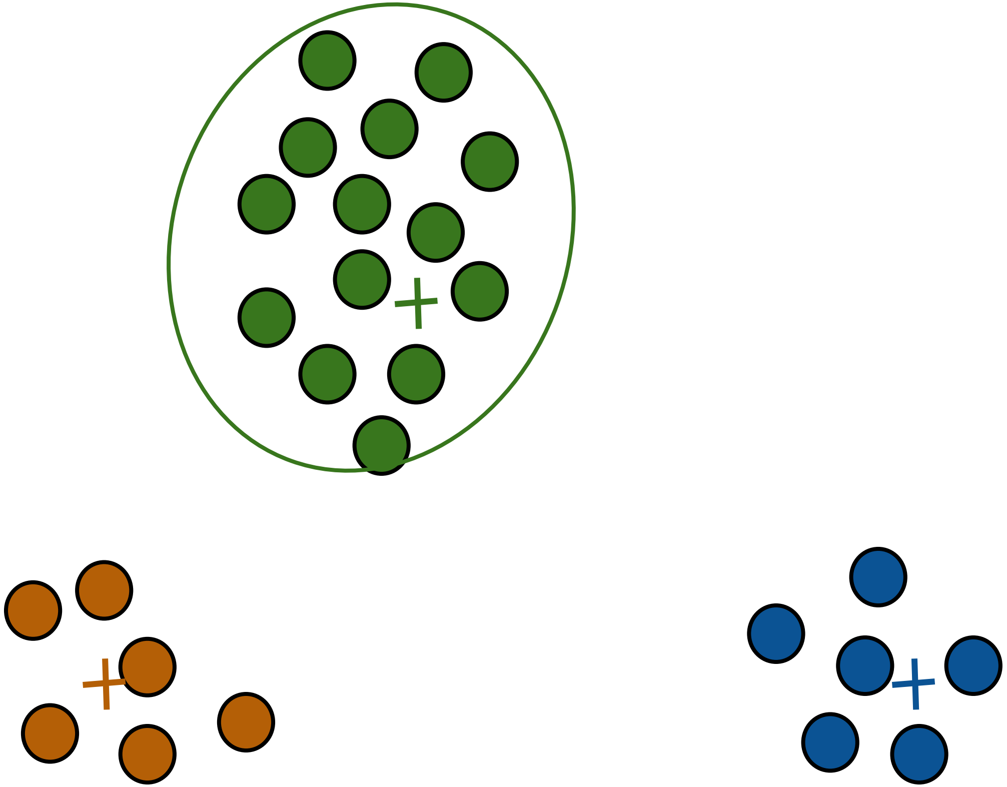


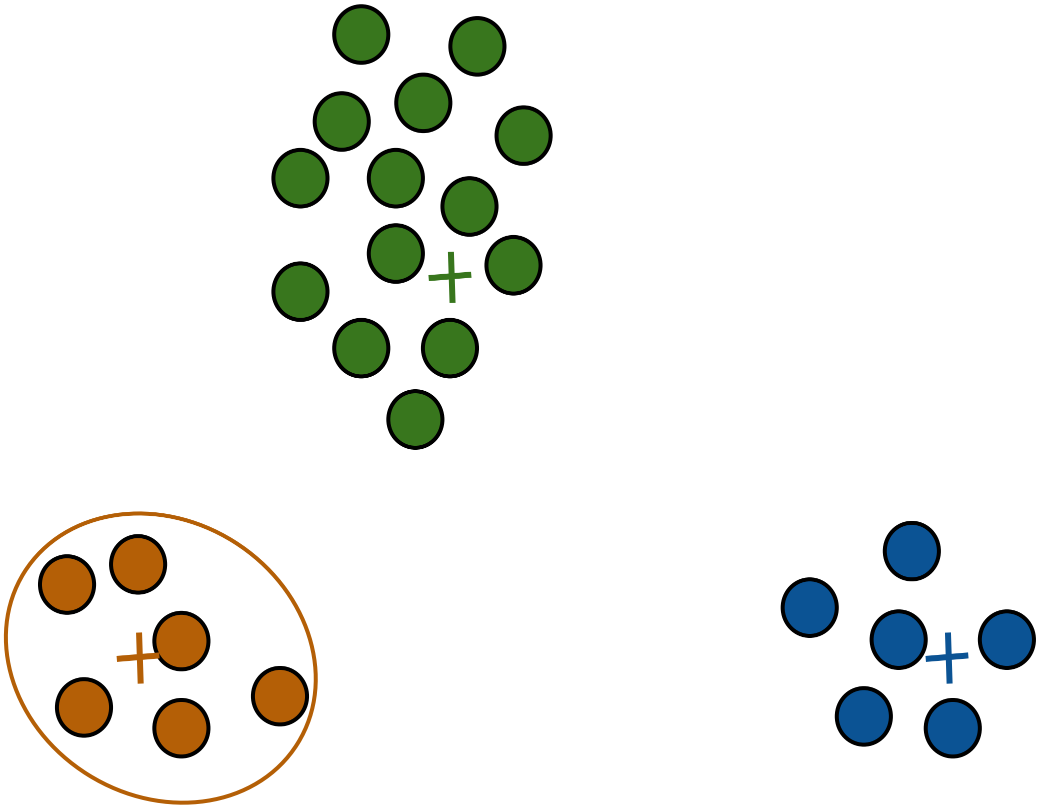


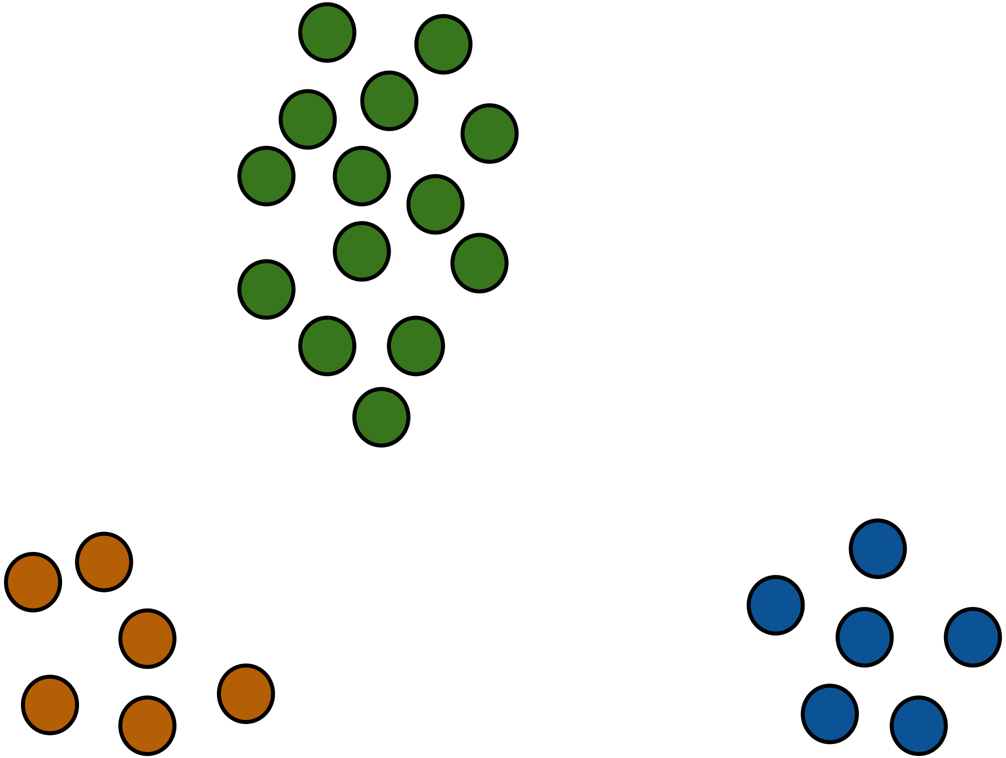




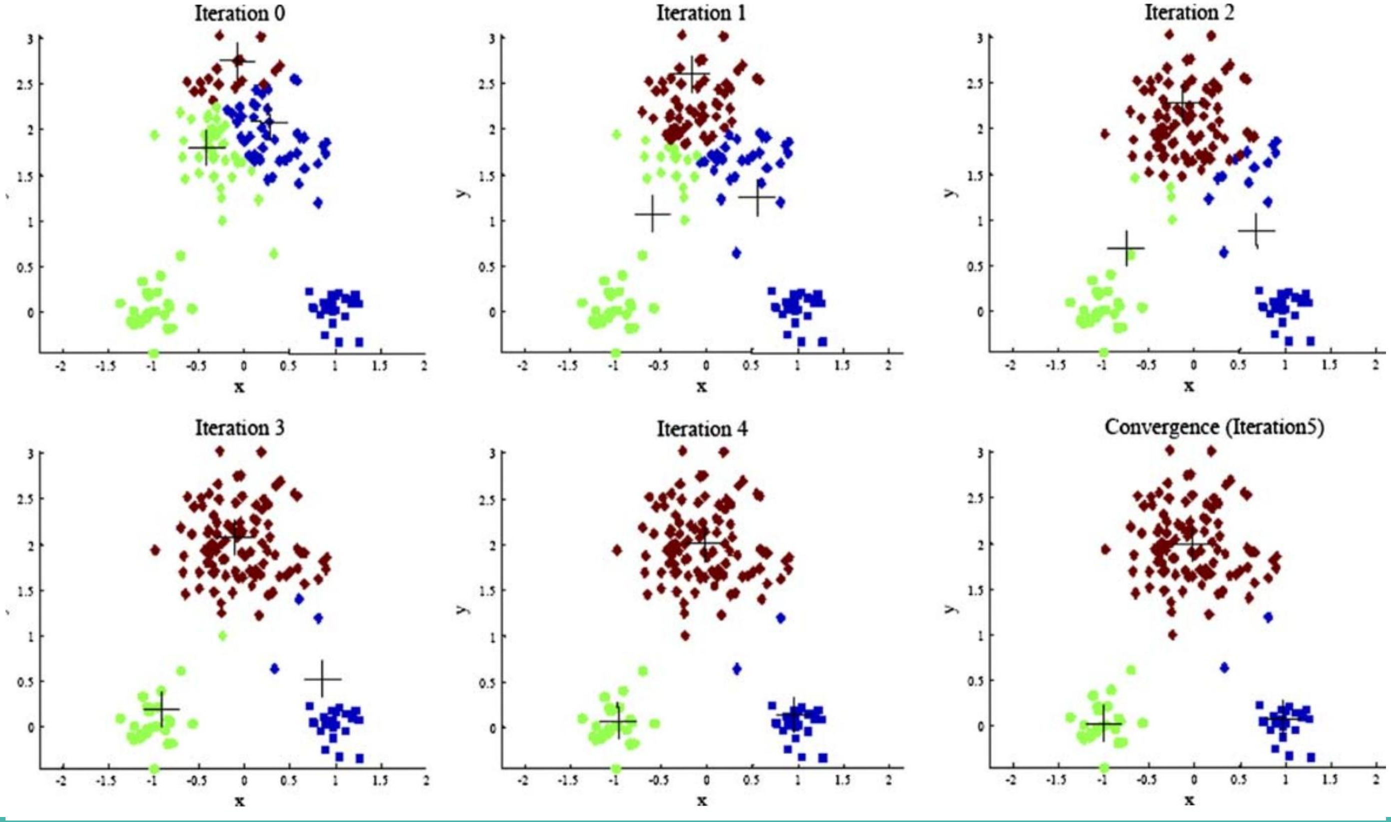








K-means - Lloyd’s Algorithm



Worksheet -5min

Please do a) -> d) of the worksheet with the person sitting next to you.

Worksheet - 5min

Share your answers with the group next to you. Discuss / debate if you have different answers.