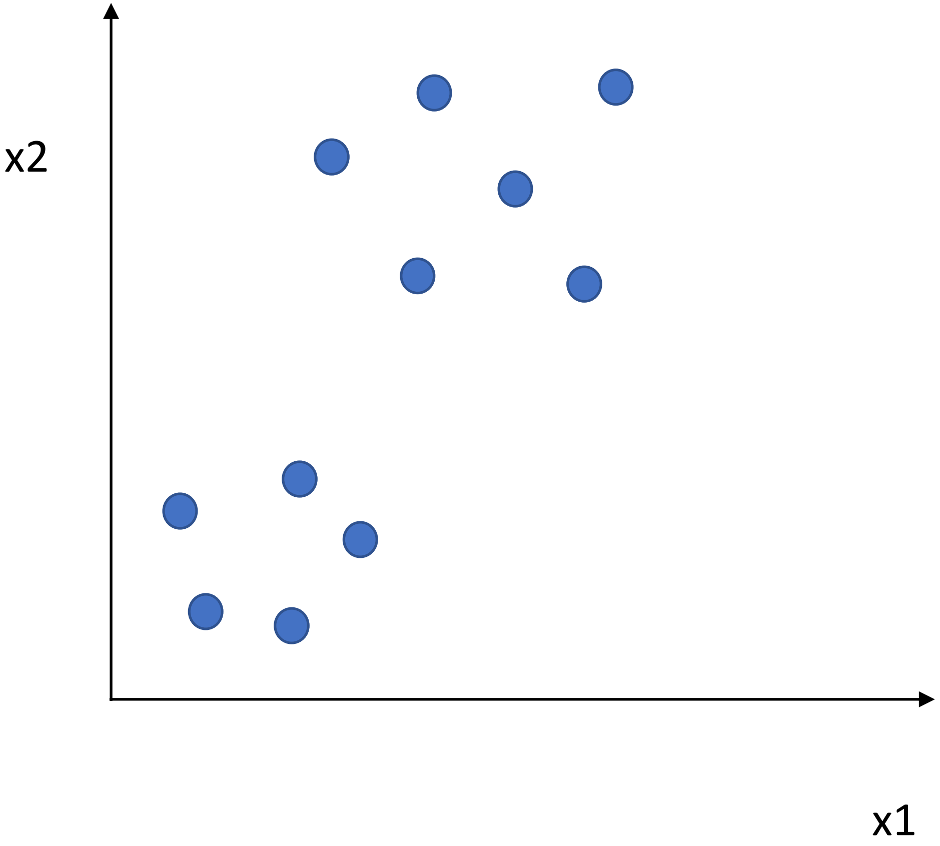
Clustering

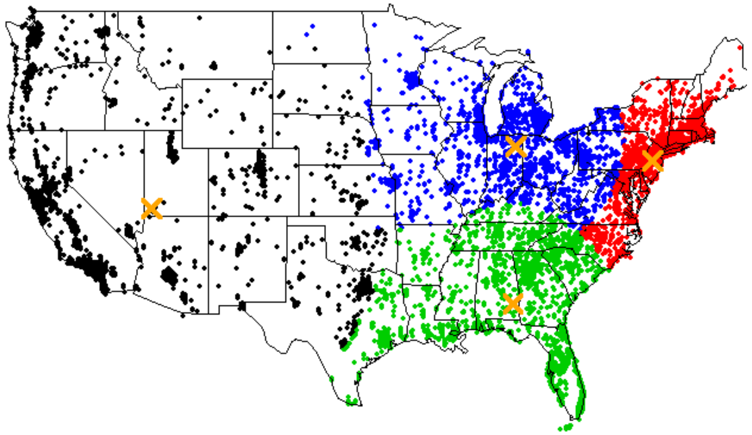
PhD Abay Nussipbekov

Unsupervised learning



Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$

Applications of clustering

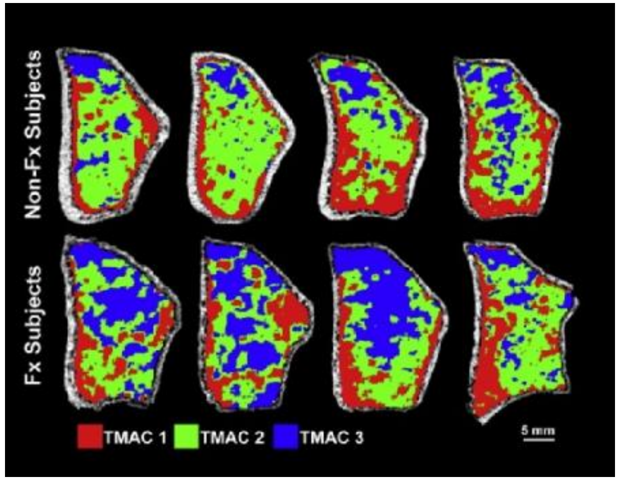


Maps

Market segmentation

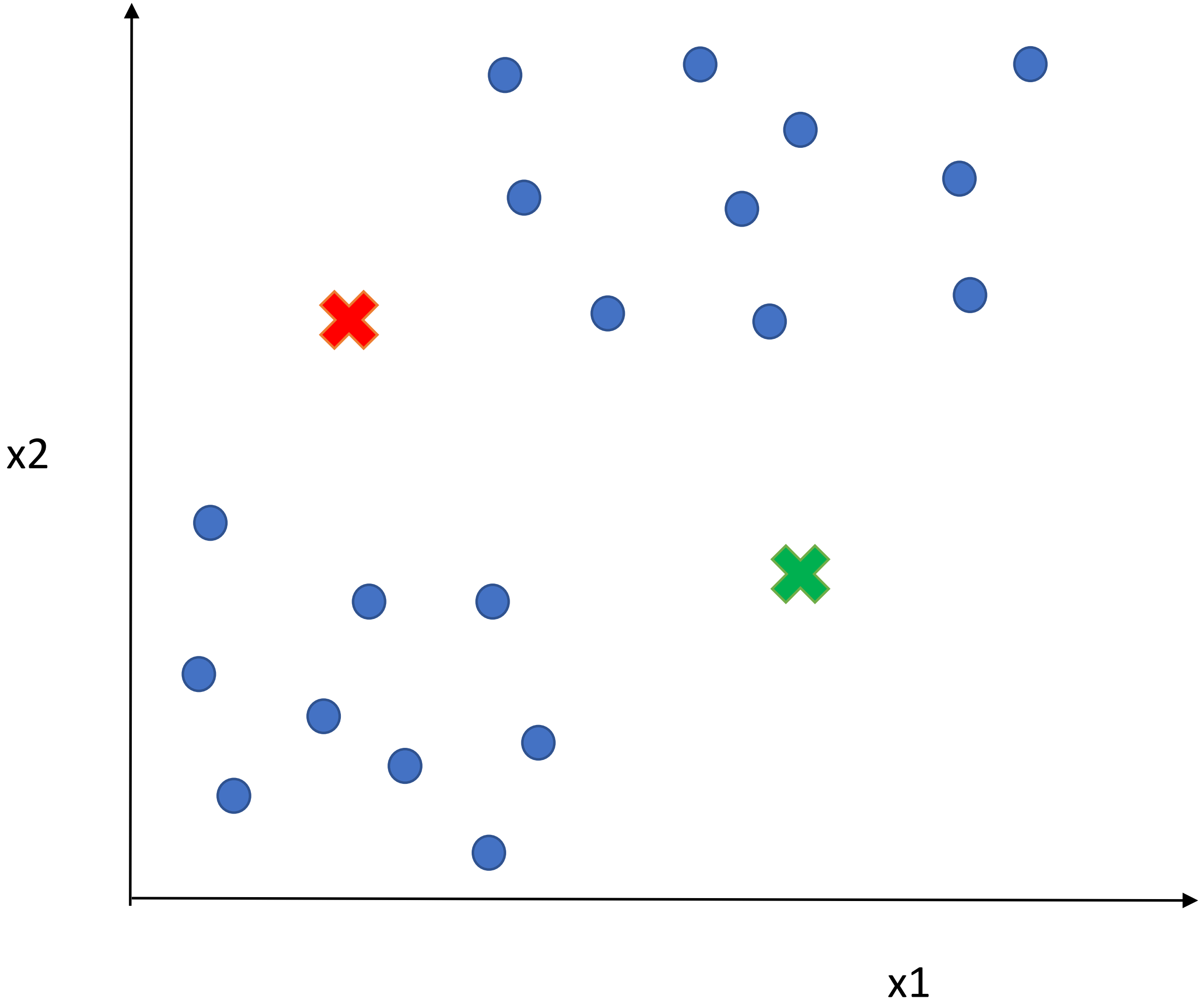
powered by
TouchGraph

Social network analysis

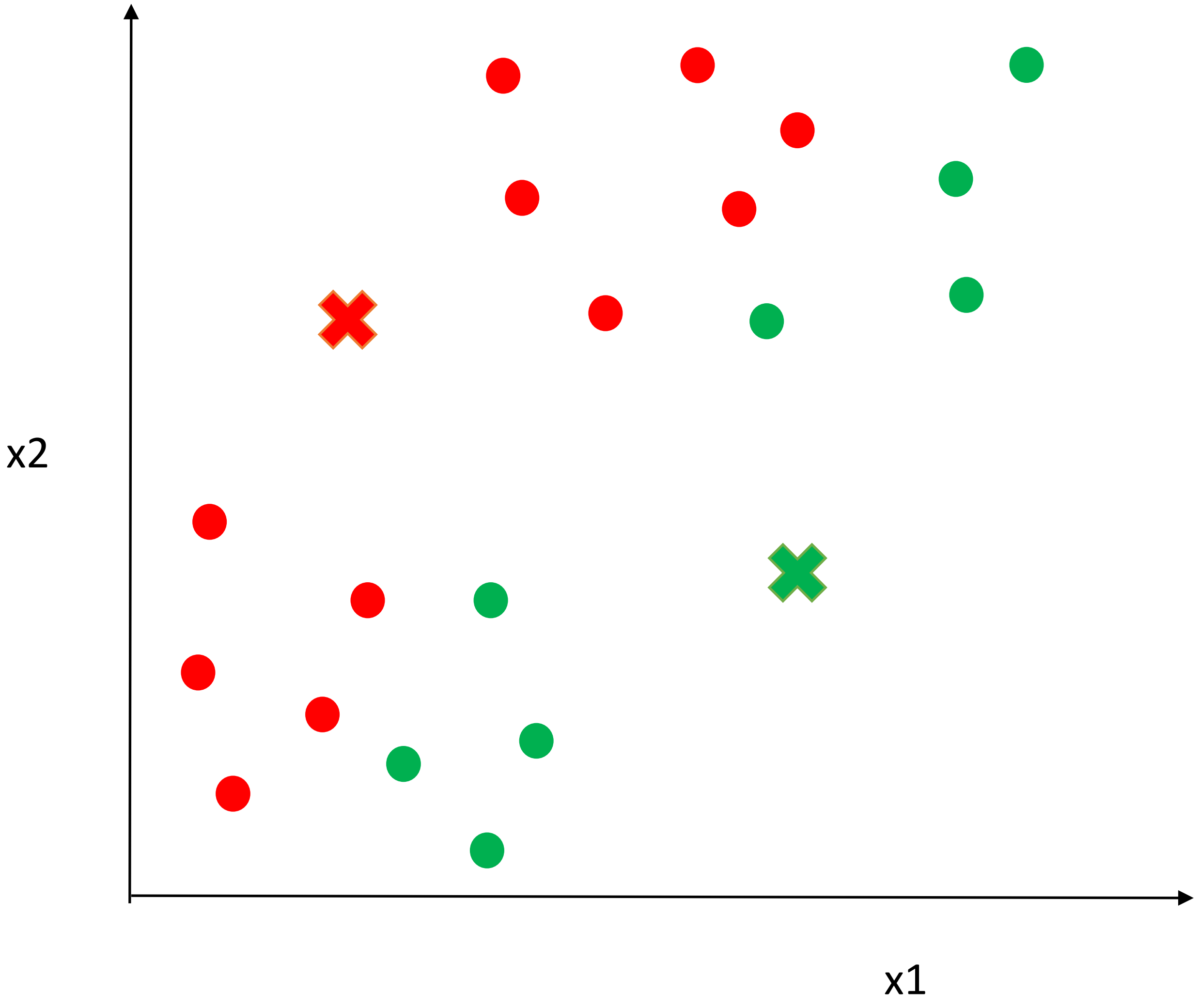


Medical imaging

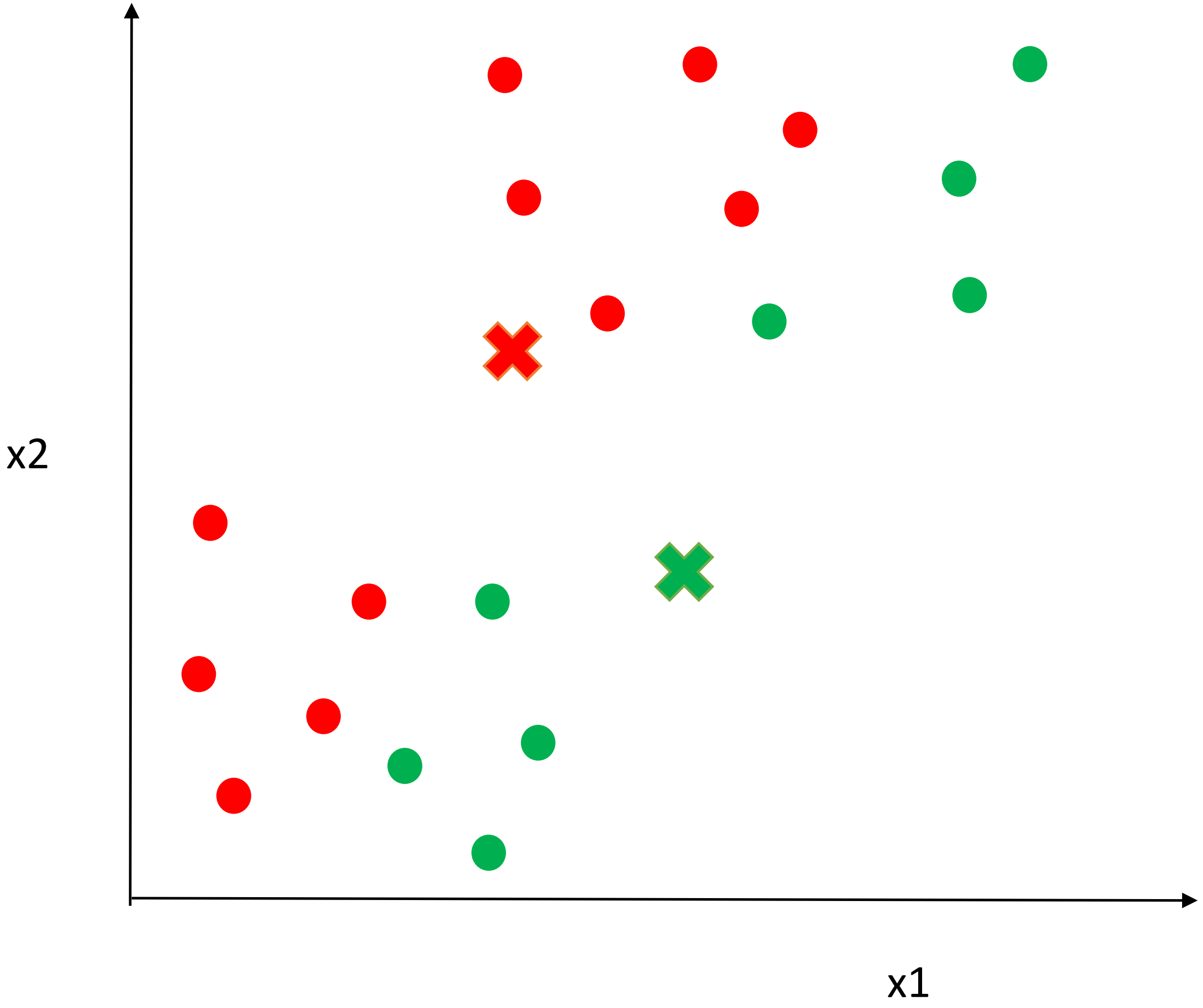
K-means clustering algorithm



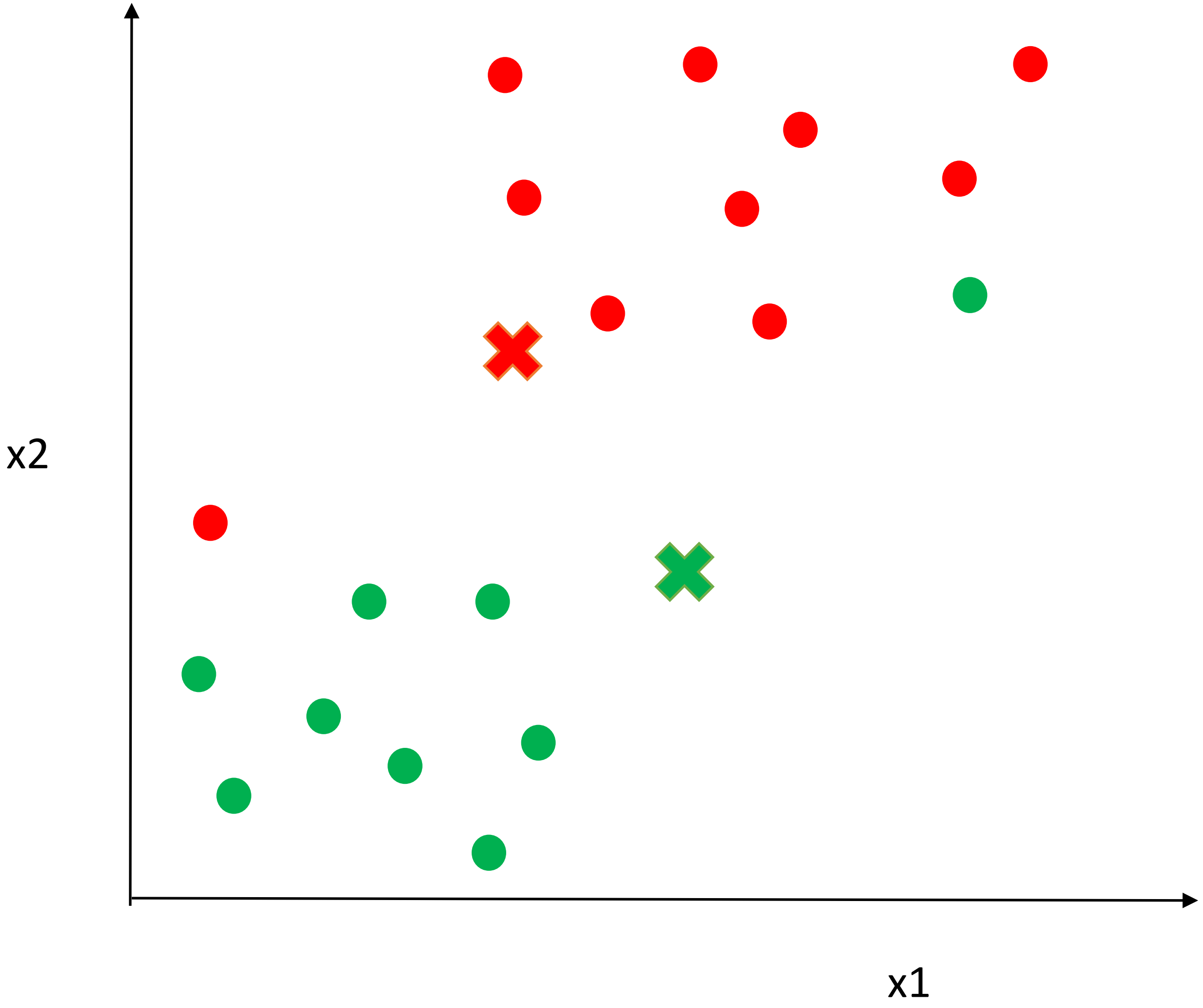
K-means clustering algorithm



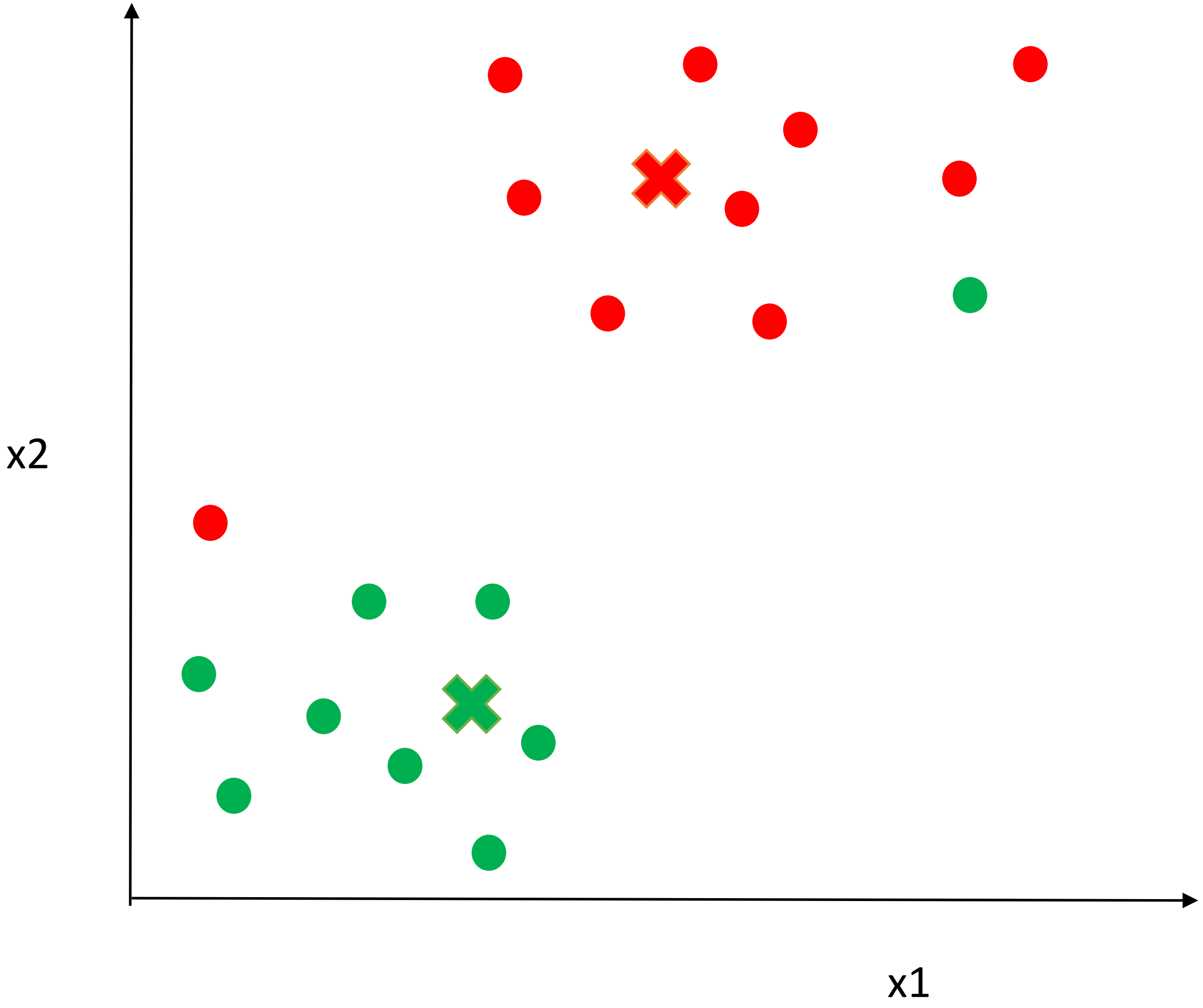
K-means clustering algorithm



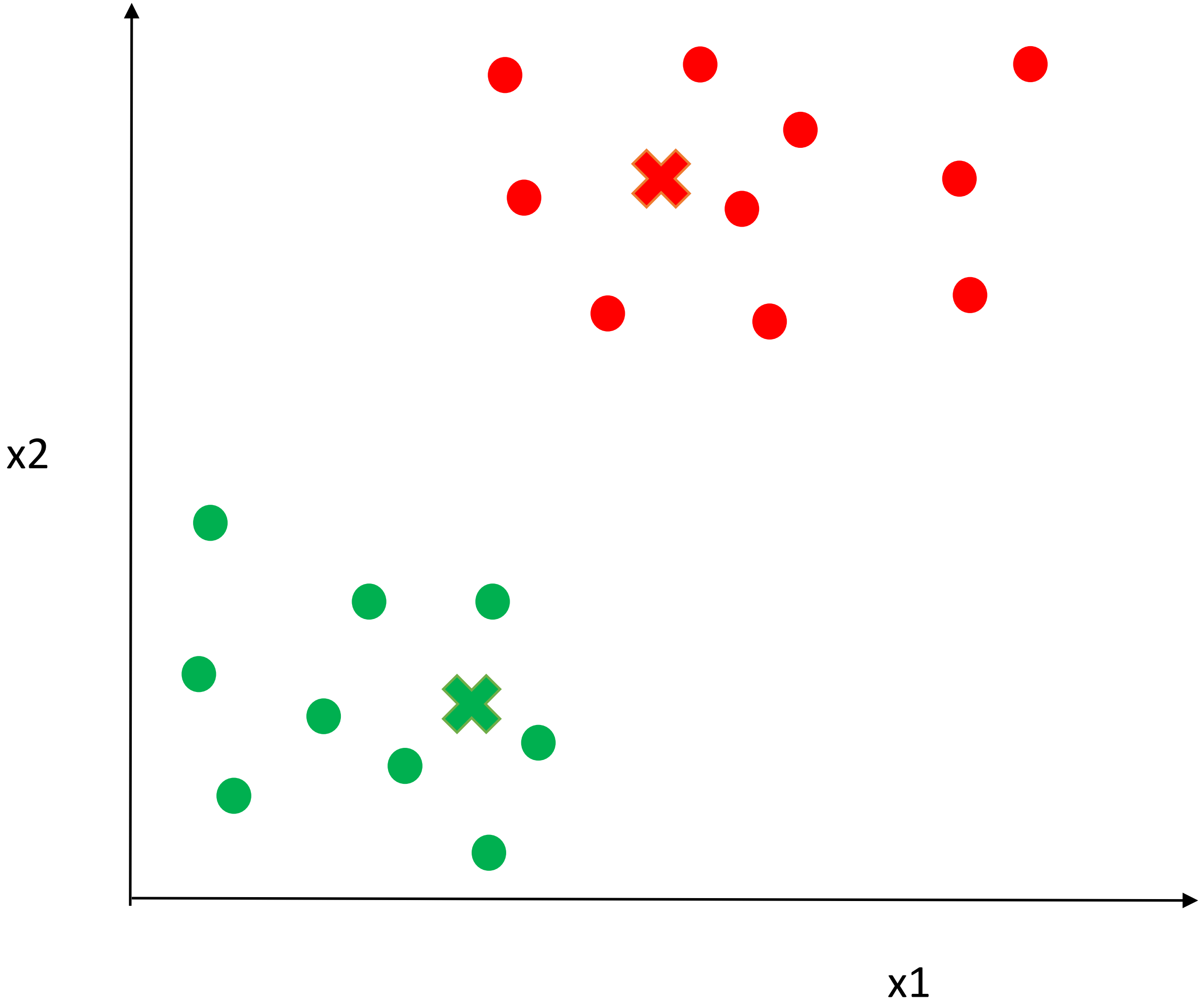
K-means clustering algorithm



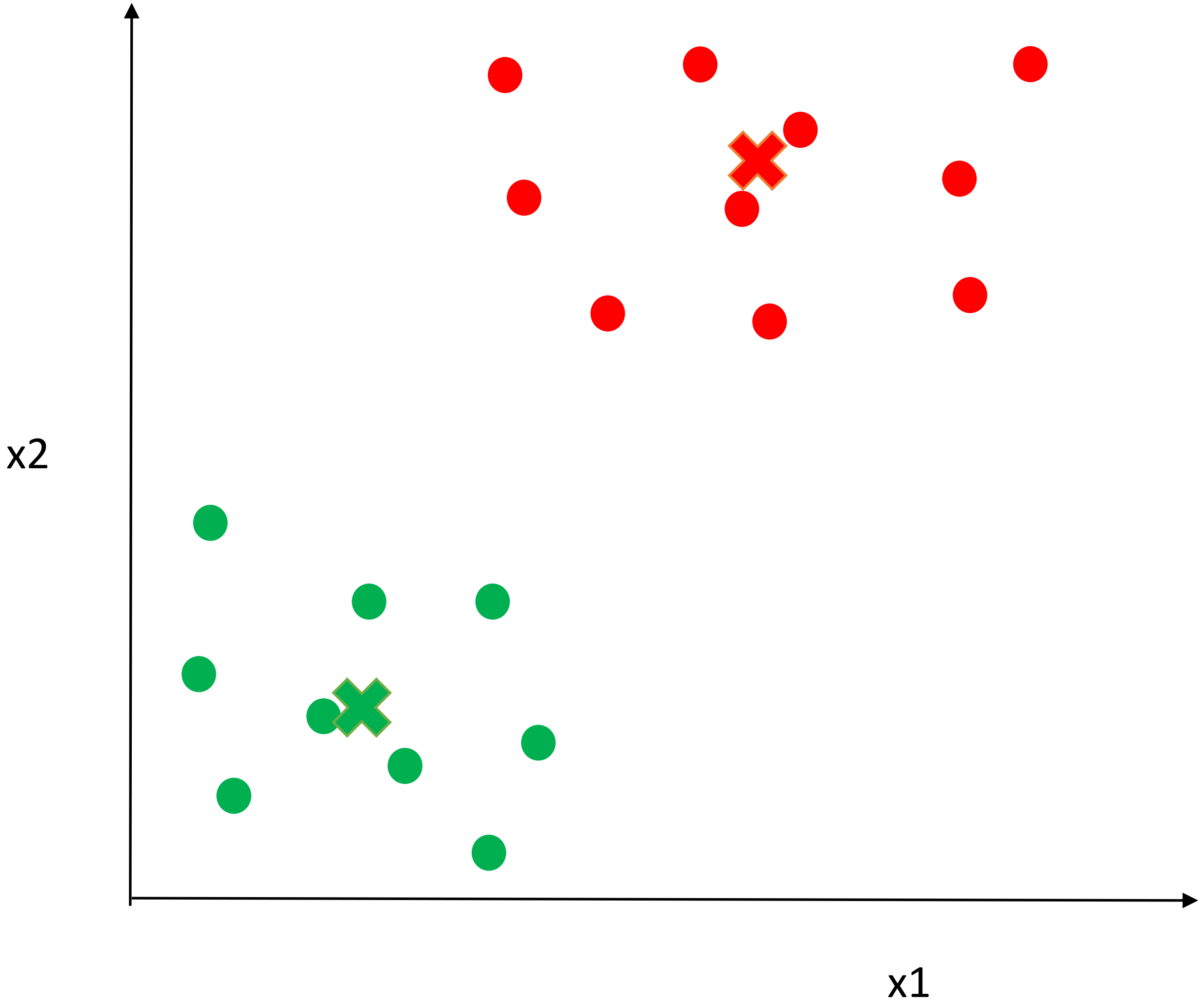
K-means clustering algorithm



K-means clustering algorithm



K-means clustering algorithm



K-means algorithm

- Input:
  - K (number of clasters)
  - Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

$x^{(i)} \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)

K-means algorithm

```
Randomly initialize K cluster centroids $\mu_0, \mu_1, \dots, \mu_K \in \mathbb{R}^n$
Repeat {
        for i=1 to m
                $c^{(i)}$ = index (from 1 to K) of cluster centroid
                        closest to $x^{(i)}$
        for k=1 to K
                $\mu_k$ = average (mean) of points assigned to cluster k
        }
```

Optimization objective

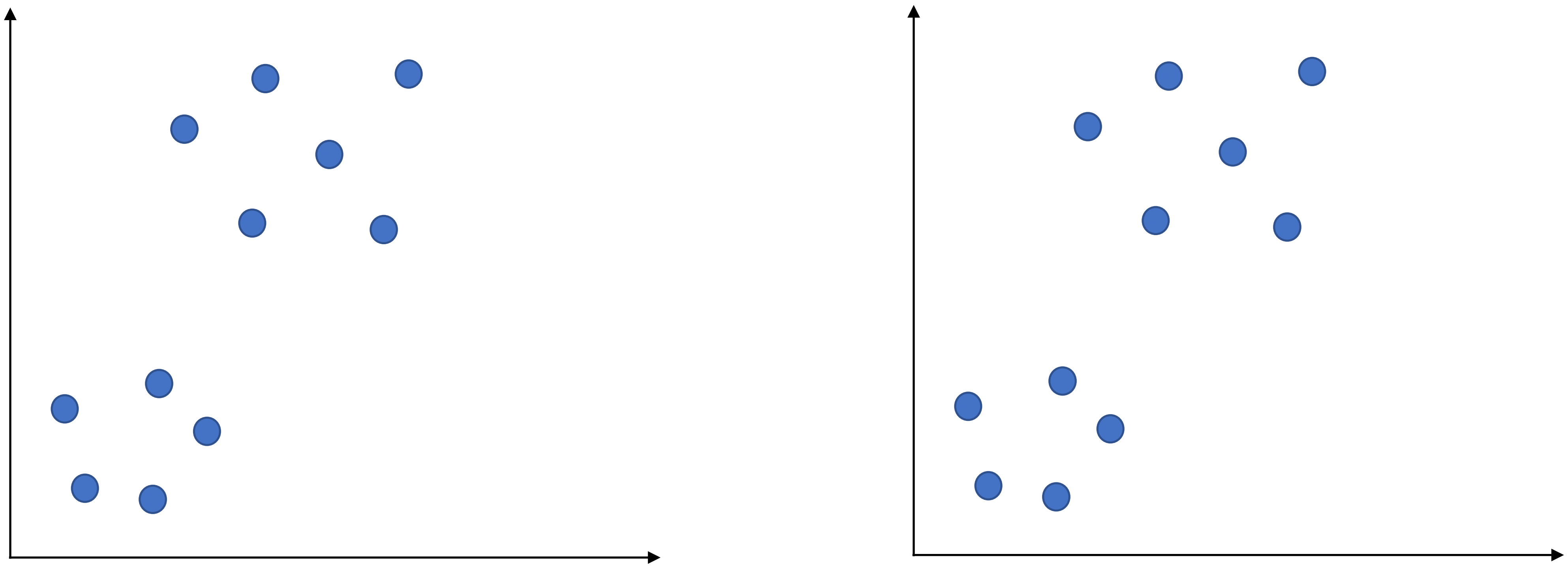
- $c^{(i)}$ = index of cluster (1, 2, …, K) to which example $x^{(i)}$ is currently assigned
- μ_k = cluster centroid k $(\mu_k \ \in \ \mathbb{R}^n)$
- $\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

Optimization objective:

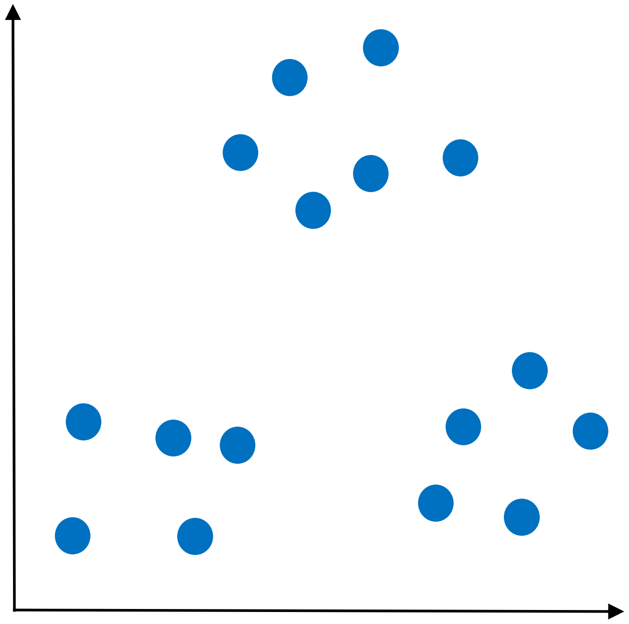
$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$$

Random initialization

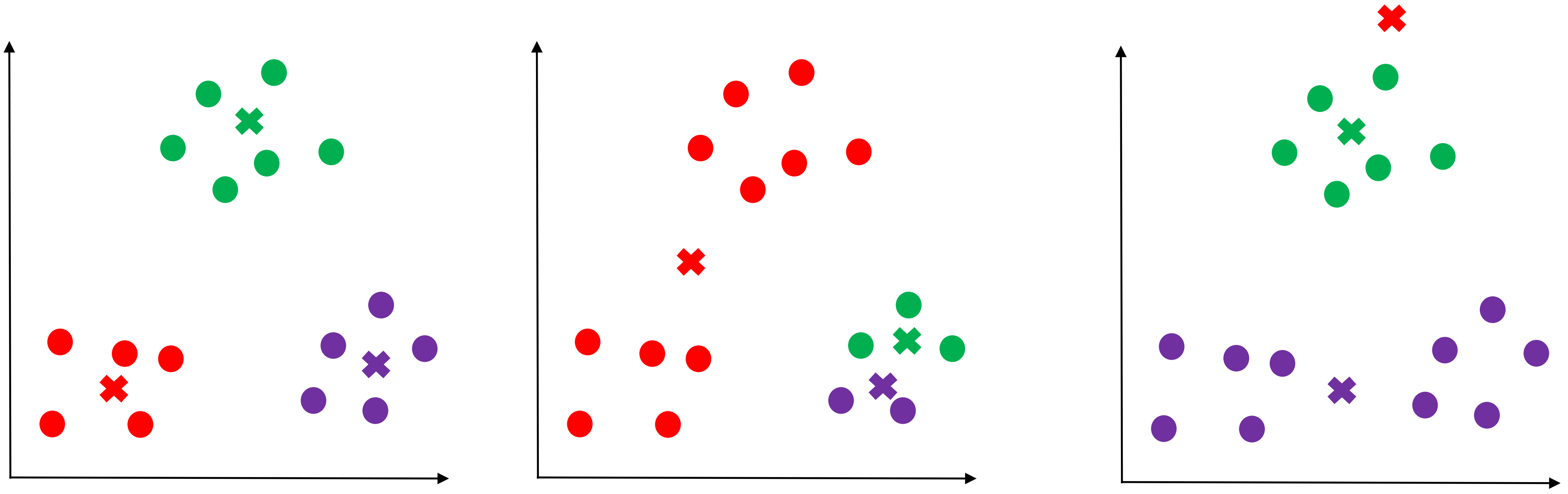
- Randomly select K training examples
- Set $\mu_1, \dots, \mu_k$ equal to these K examples



Global and local optima



Global and local optima



How to find best parameters?

```
for i to 100 {
        randomly initialize k-means
        run k-means. get $c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K$
        compute cost function $J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$
}
```

Pick clustering that gave lowest cost $J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$

How to define number of clusters?

- Elbow curves
- Task dependent