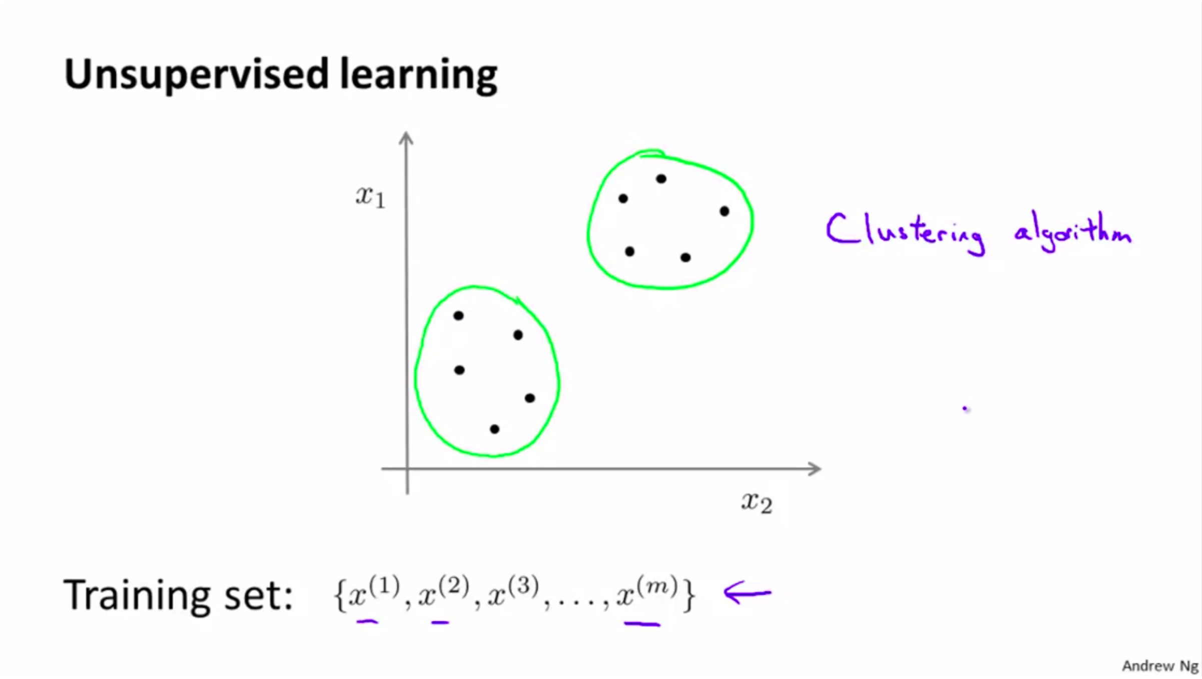
Week 8 - Lecture 1

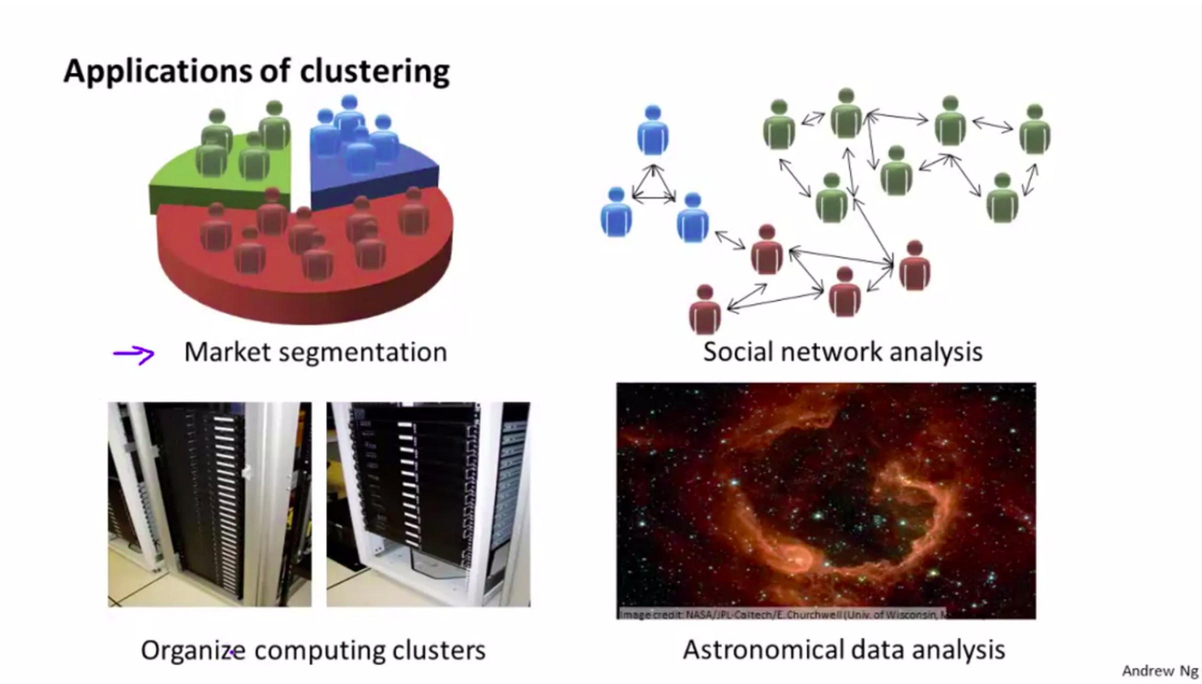
Unsupervised Learning

In an unsupervised learning problem we have a dataset with unlabeled (unstructured) data, that may look like the following:



and we let the algorithm find some structure on the data for us.

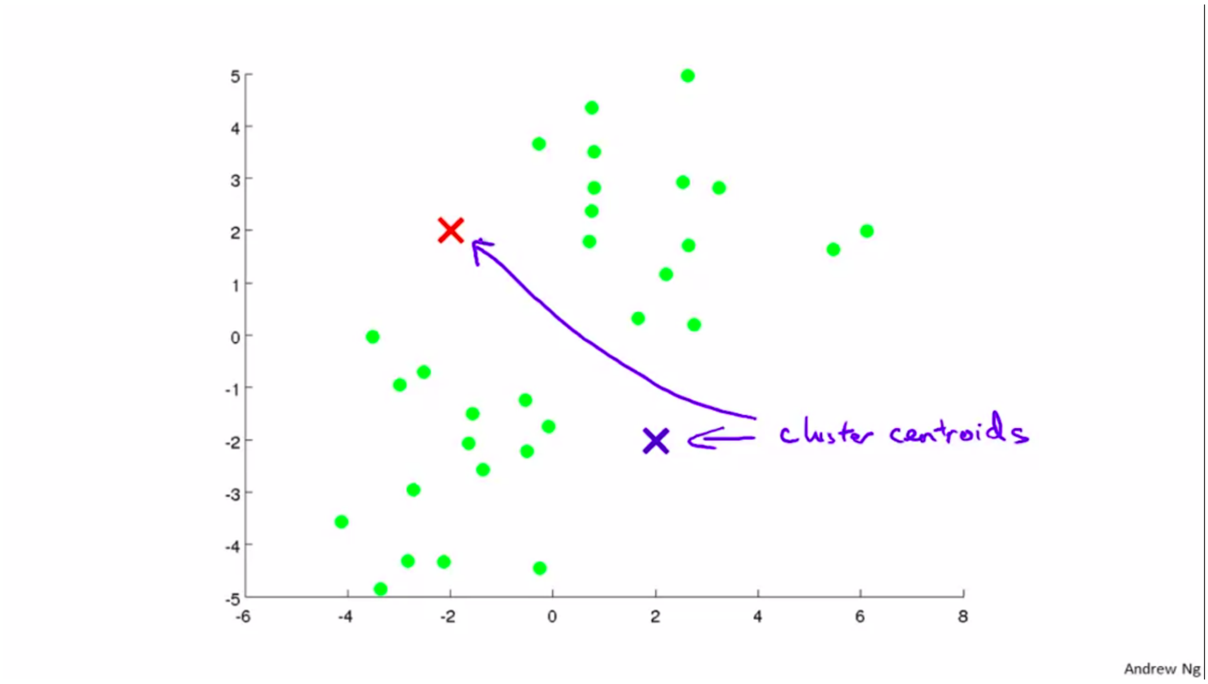
Use cases of clustering:



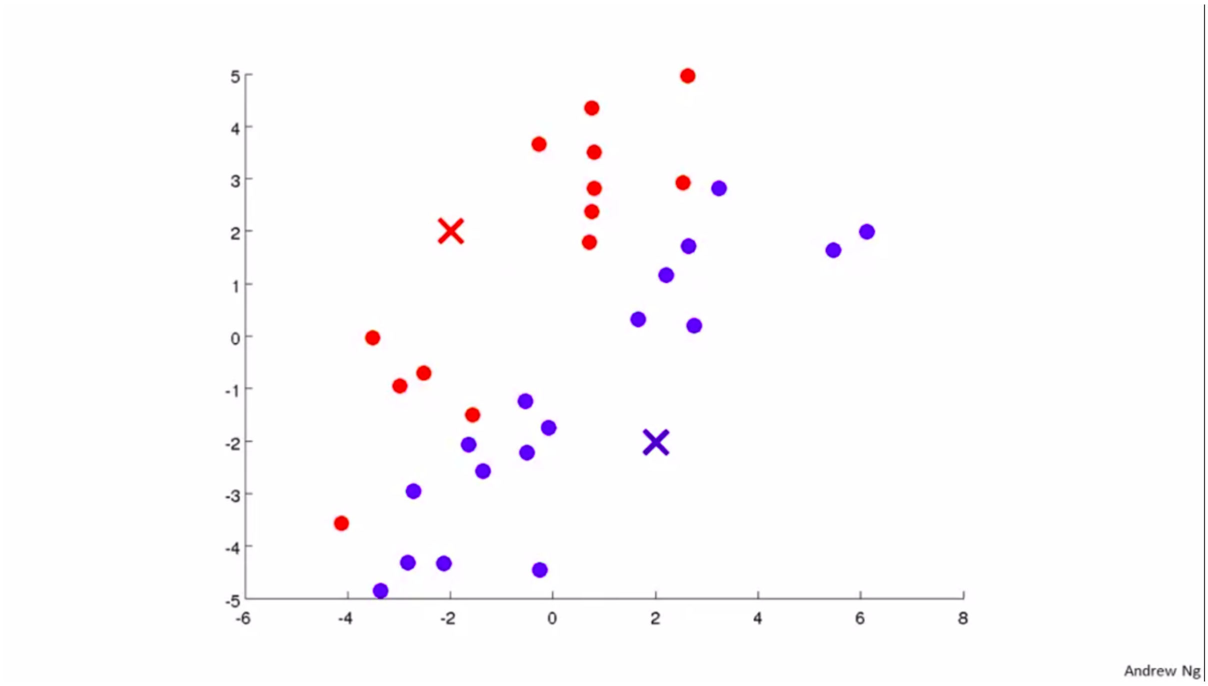
K-Means Algorithm

In the clustering problem, we are given a dataset and we would like to find some structure in it, i.e., to divide it into some coherent groups. K-Means is one clustering algorithm.

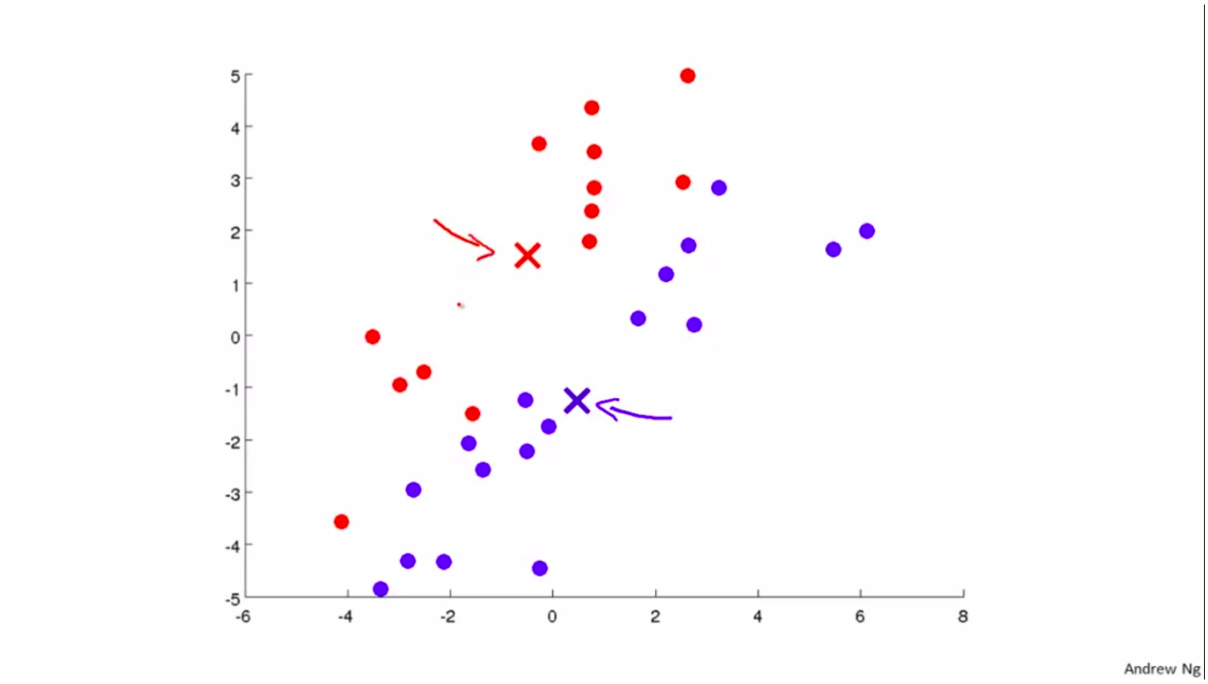
- **Step 1:** Initialization. The algorithm randomly chooses k centroids.



- **Step 2:** (Cluster assignment step) The algorithm assigns a label to each data point depending on which centroid is closer.



- **Step 3:** (Move centroids step) The algorithm moves the centroids and then repeats itself from step 2.



> **Note:** the algorithm repeates steps 2 and 3 until it converges.

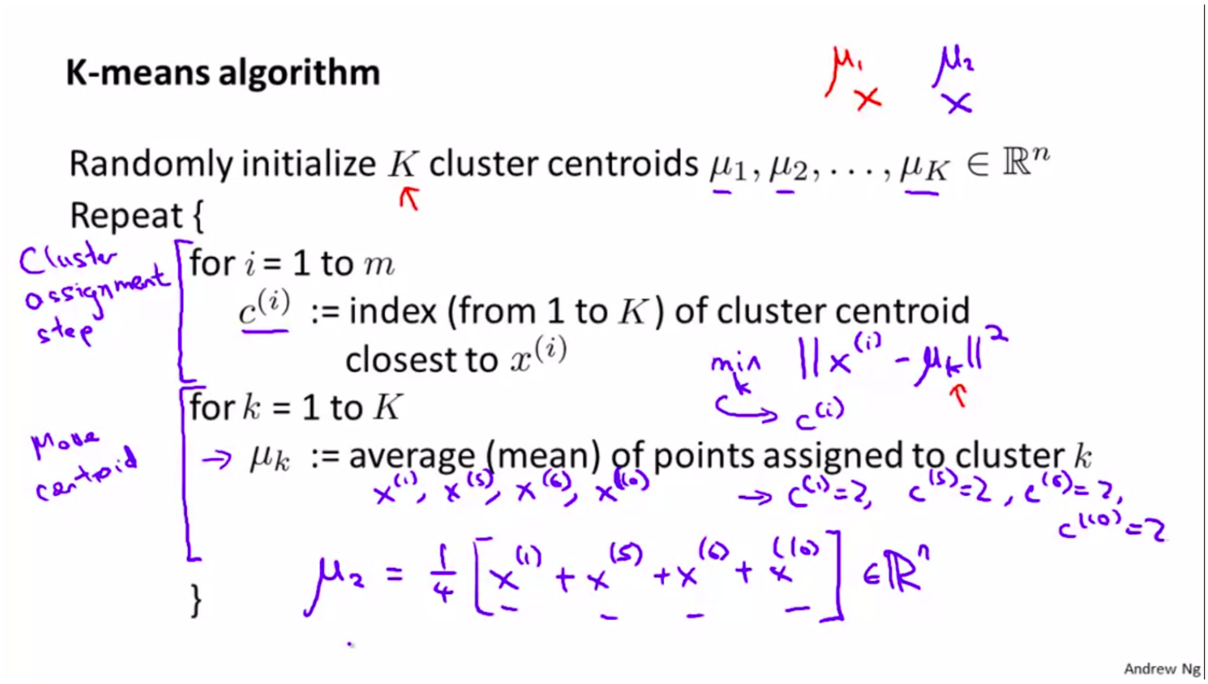
K-means algorithm

Input:

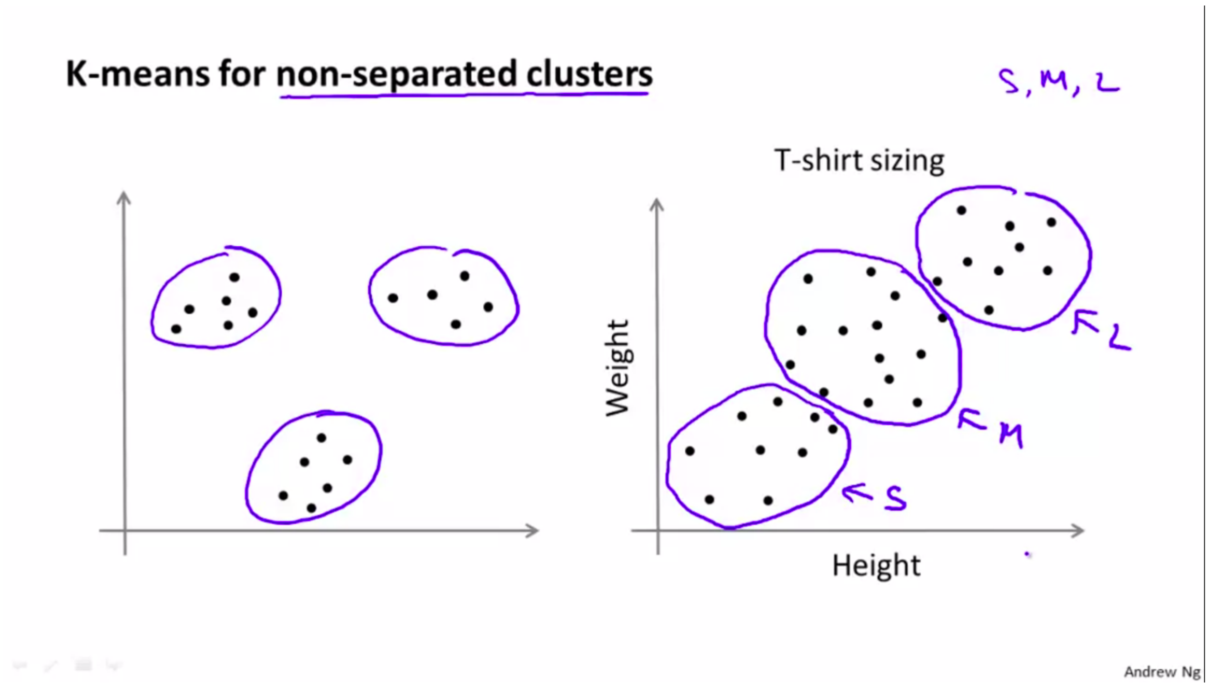
- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

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

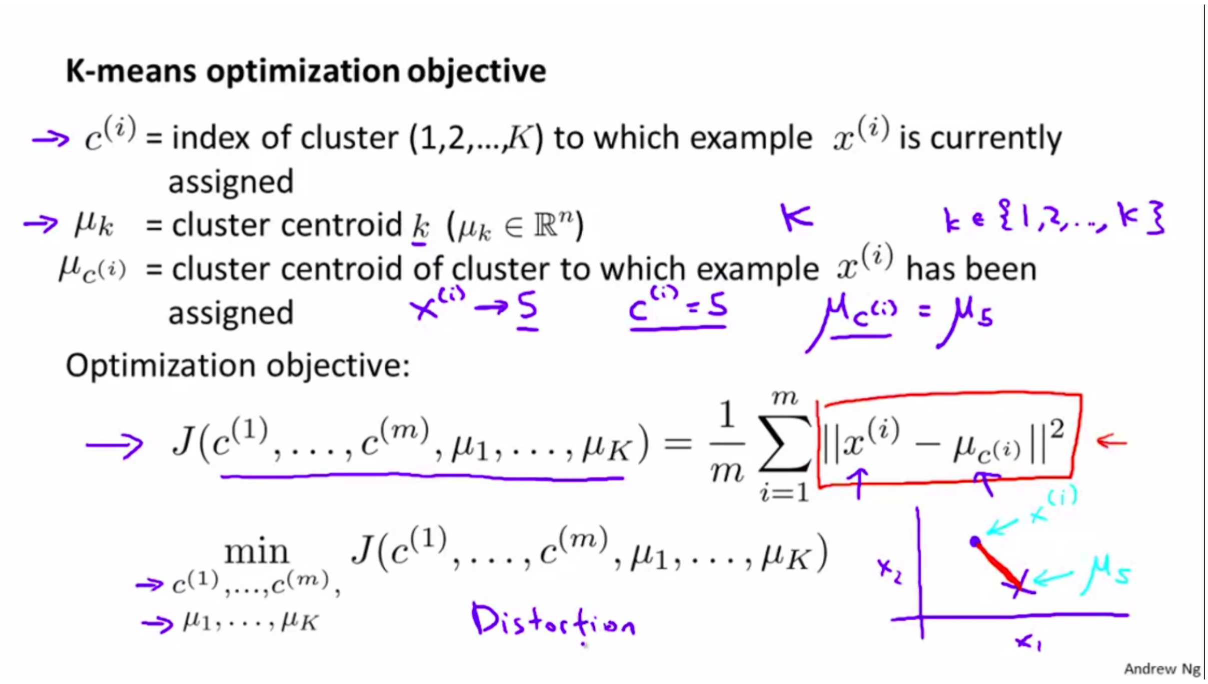
Andrew Ng



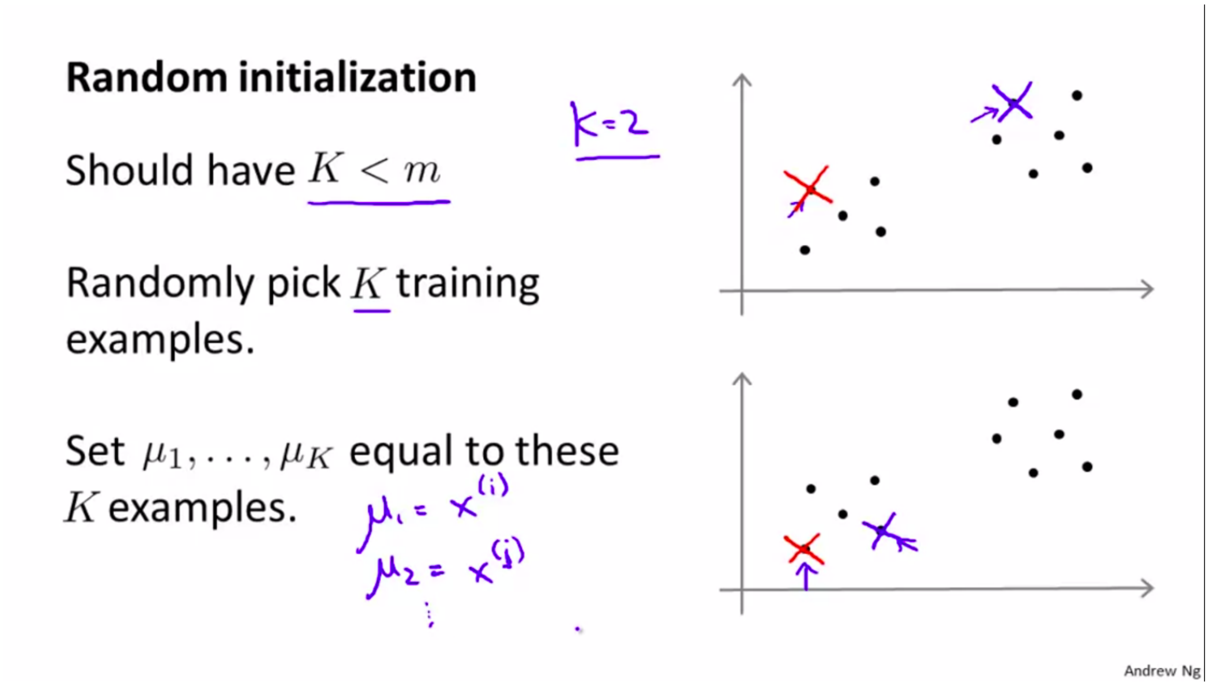
Example:



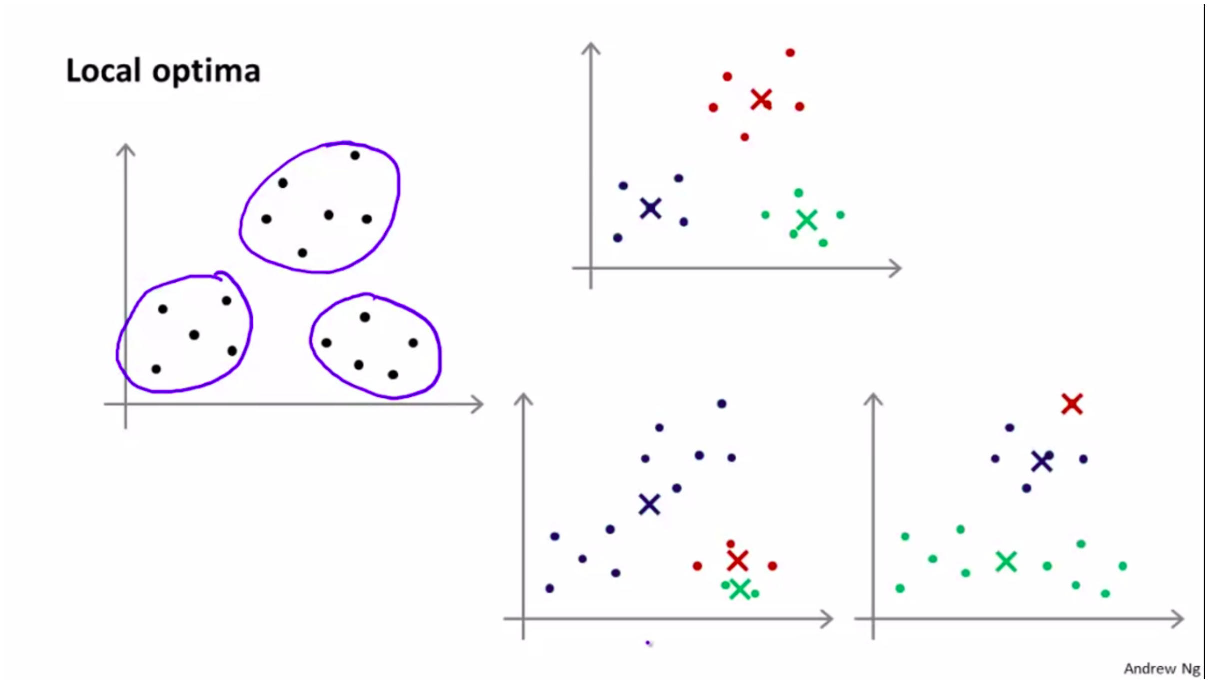
Optimization Objective



Random Initialization

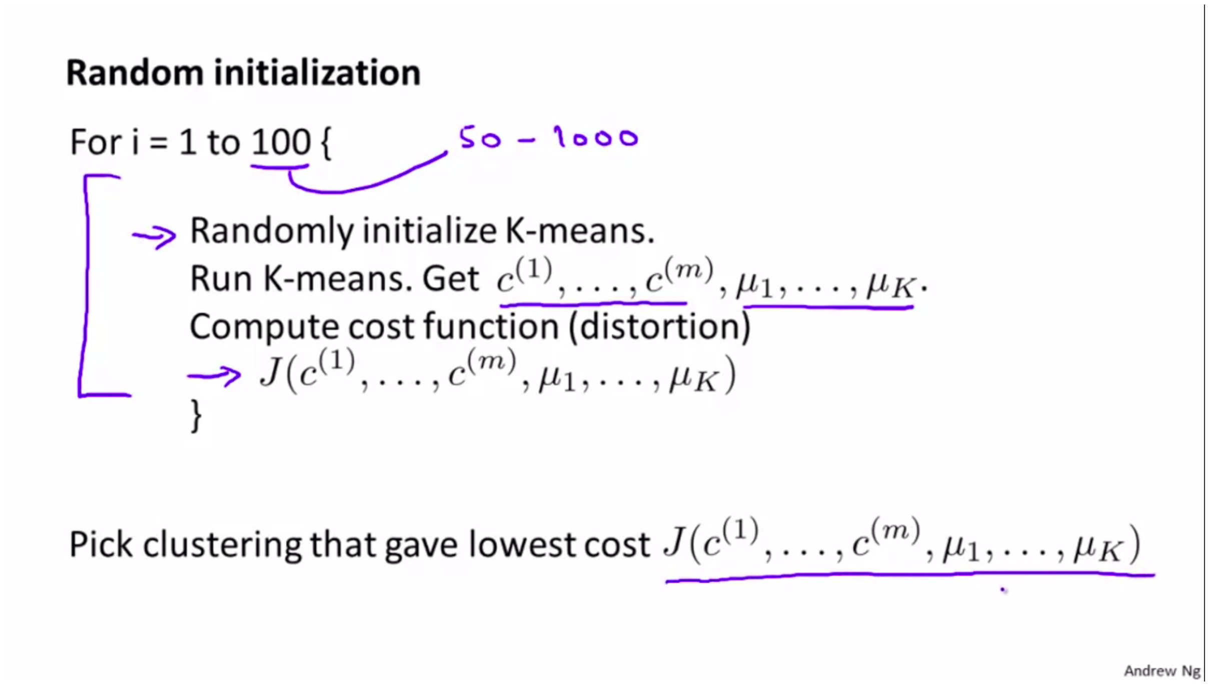


Depending on which initialization centroids we choose, K-means might end up in a local optima.



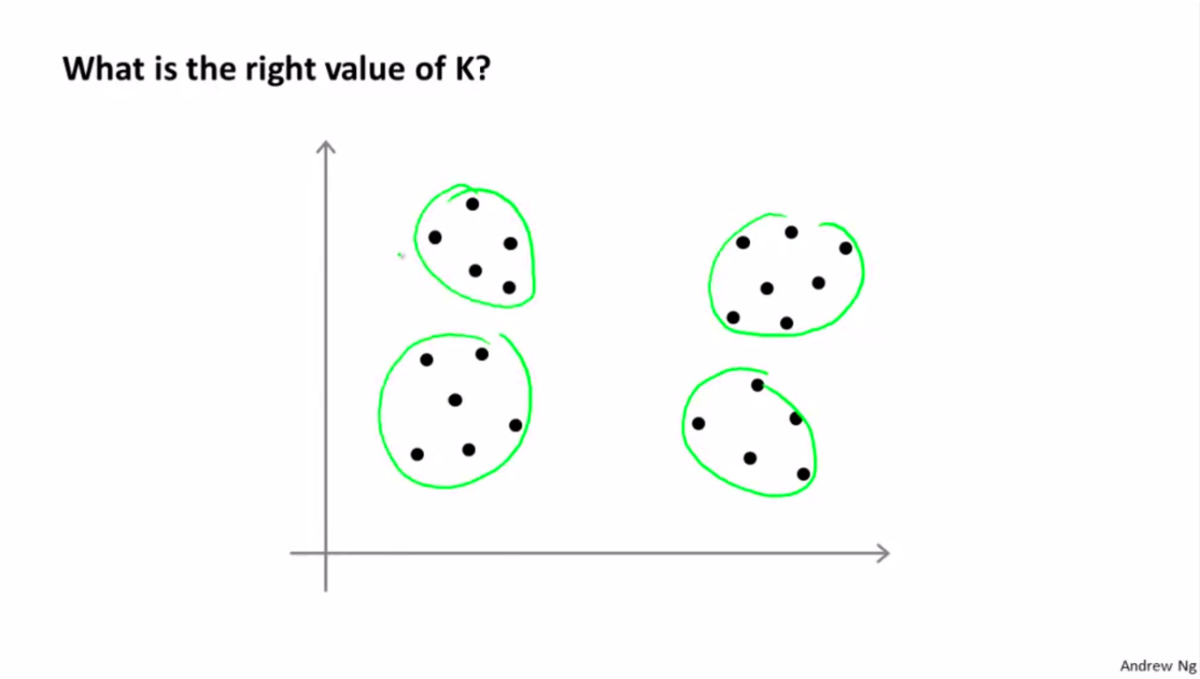
In the figure above, the upper case might be the global optima and the two on the bottom might be the local optima (subotimal solutions)

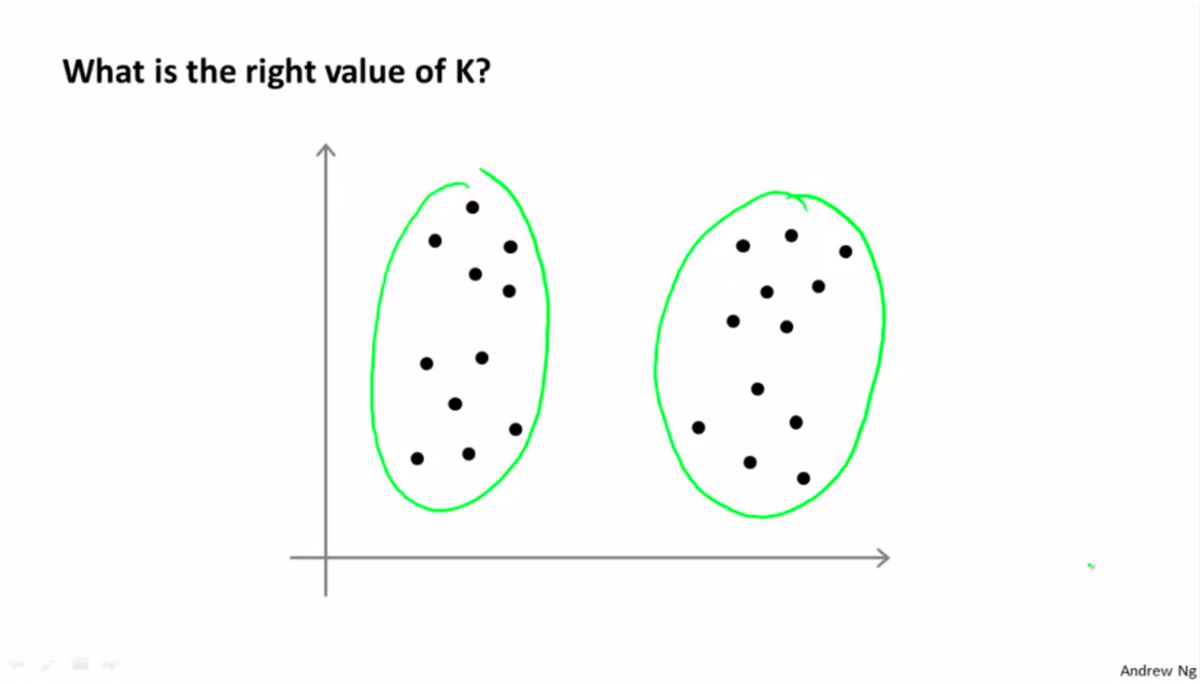
Good practice to avoid this:



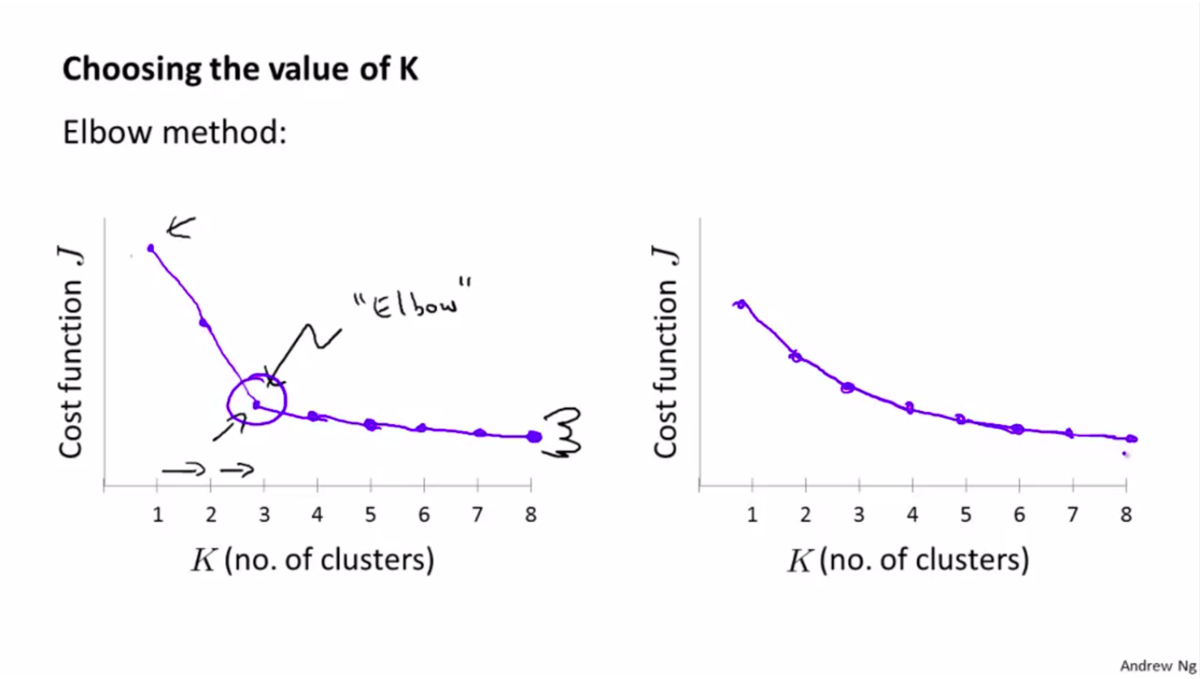
Choosing the Number of Clusters

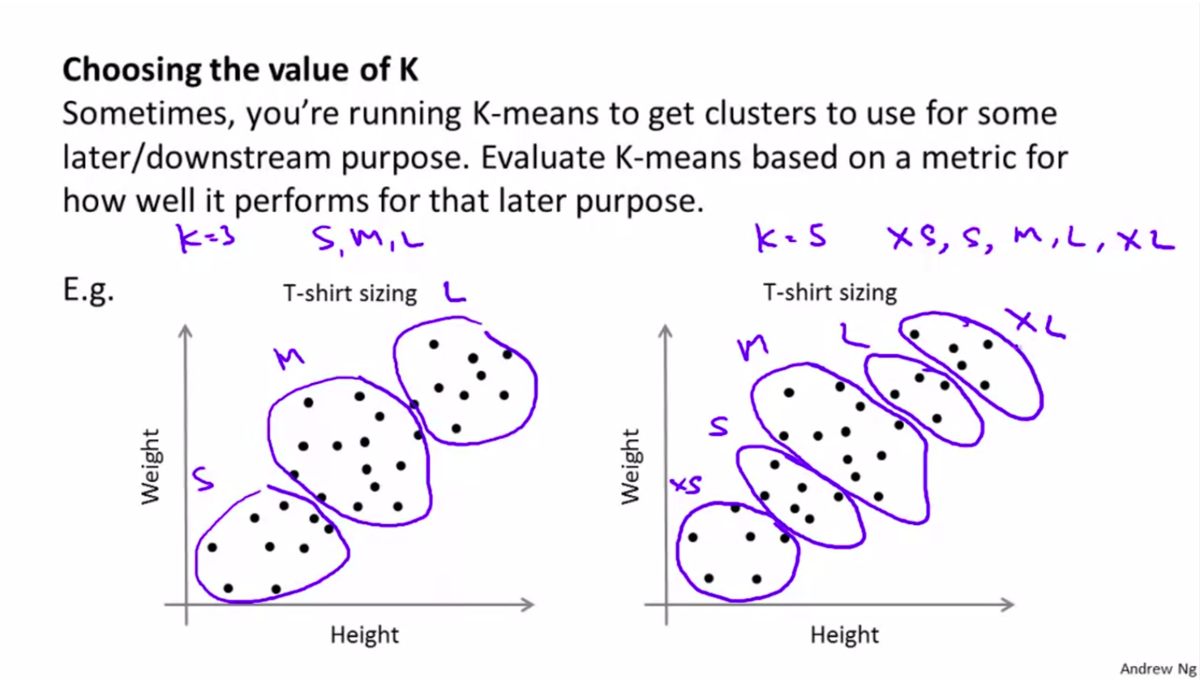
The most common way of choosing the number K of clusters is by doing it manually, looking at visualizations or other kind of intuition. So this is a somewhat intuitive step and, to help develop this intuition, some examples are presented below.





Although, some methods might help with this choice:





References

[1] Machine Learning - Stanford University (https://www.coursera.org/learn/machine-learning)