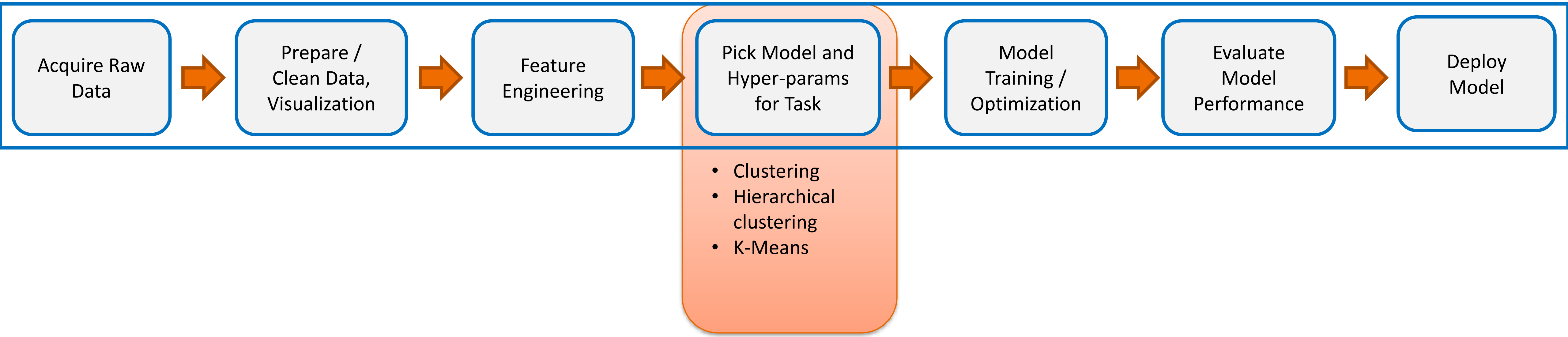
Focus for this lecture



Unsupervised Learning

Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

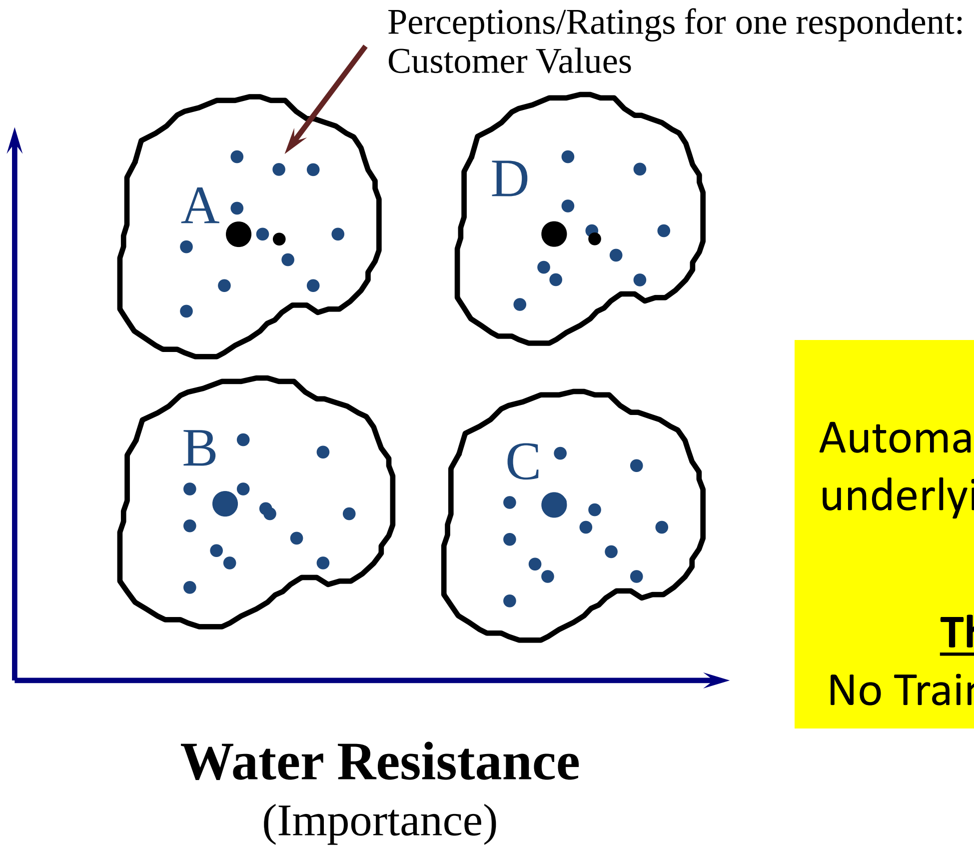
- Clustering

Clustering

Identifying Similar Patterns

Market Segmentation Problem – Carpet Fiber

Strength
(Importance)



A,B,C,D:
Location of
segment centers.

Typical Members:

A: Schools
B: Light commercial
C: Indoor/outdoor Carpeting
D: Health clubs

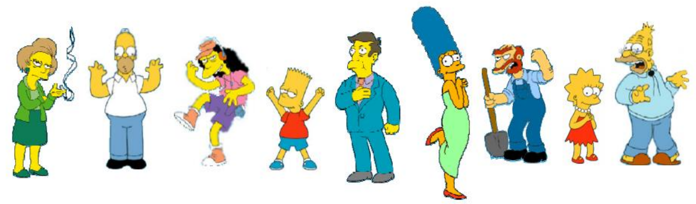
<u>The Task</u>
Automatically Discover the underlying structure in the data

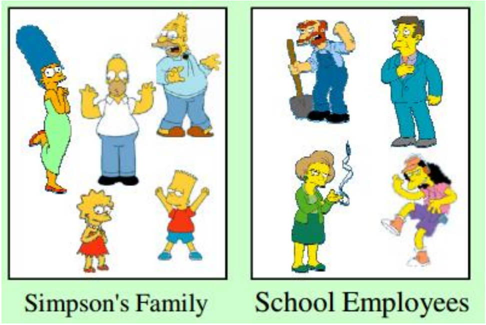
<u>The Challenge</u>
No Training Data available

What is Clustering?

- Clustering – Grouping objects together based on similarity
- Unsupervised Learning – No predefined classes

What is a natural grouping among these objects?





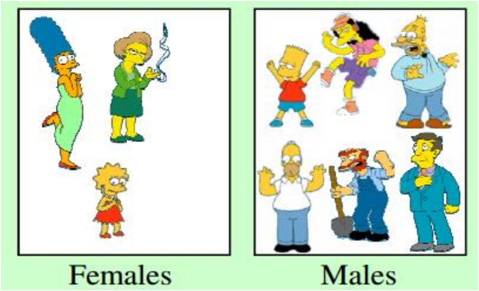
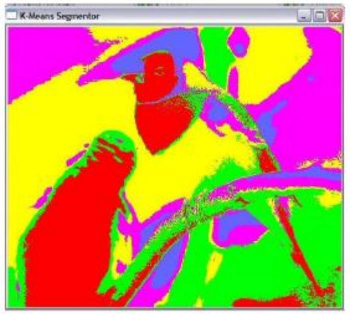


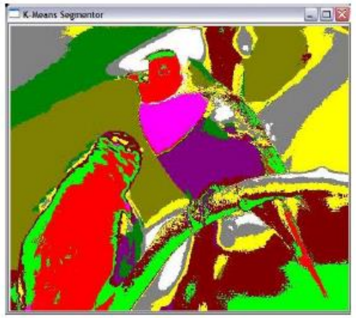
Image Segmentation

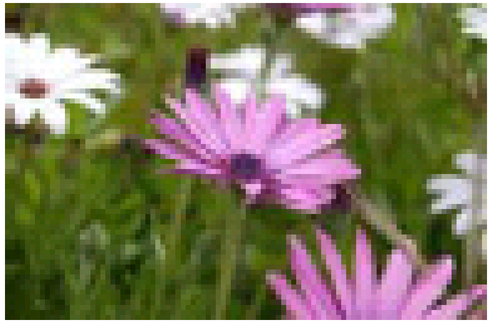
Image Segmentation

K=5, RGB space

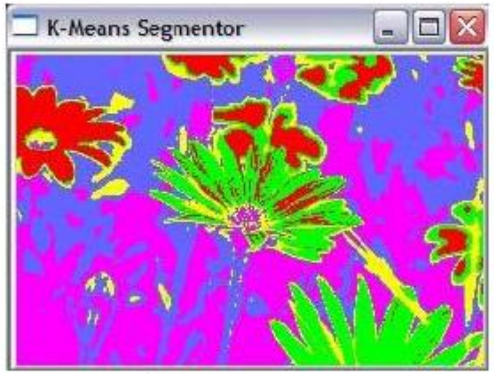


K=10, RGB space

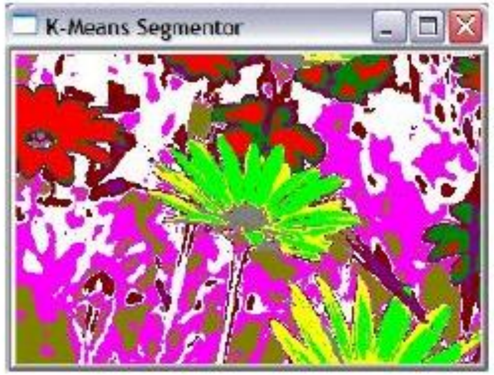


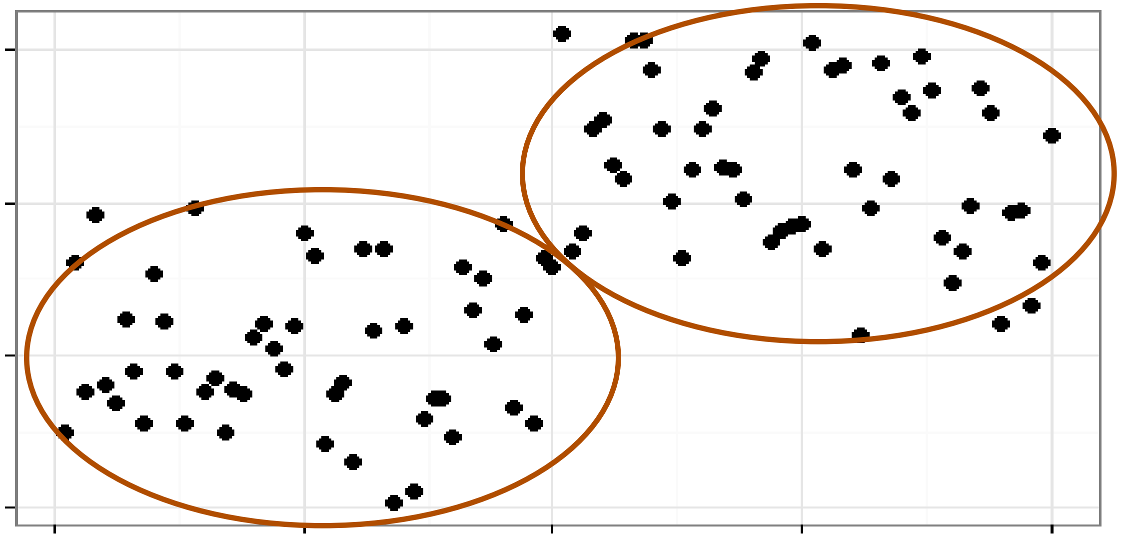


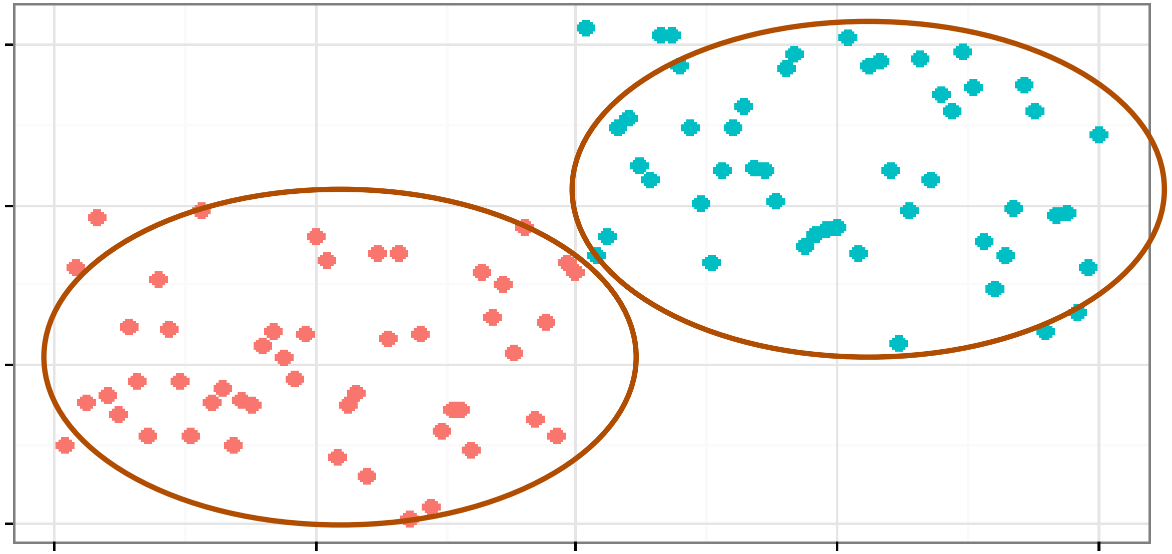
K=5, RGB space



K=10, RGB space







Clustering

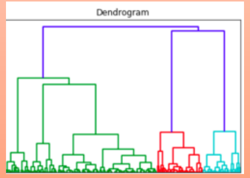
- Finding similarity groups in data, called **clusters**. I.e.,
  - data instances that are similar to (near) each other are in the same cluster
  - data instances that are very different (far away) from each other fall in different clusters.

Algorithms

- **Hierarchical** approach: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- **Partitioning** approach: Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors (K-means)

Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- ✓ Clustering
- Hierarchical clustering



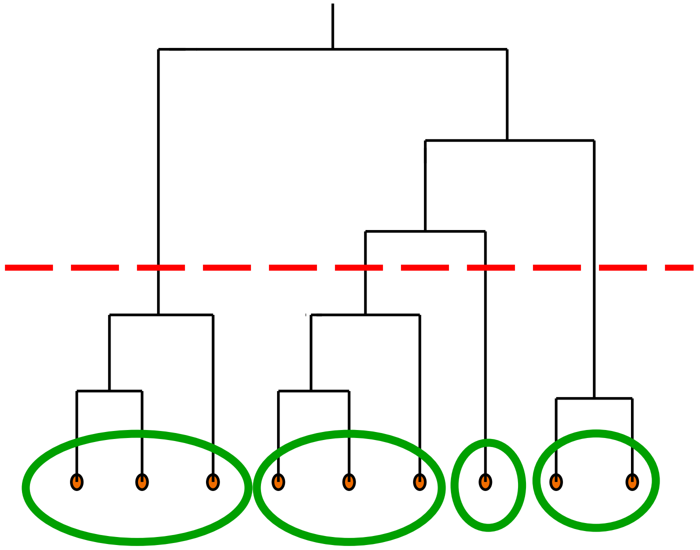
Hierarchical Clustering

Hierarchical (Agglomerative) clustering

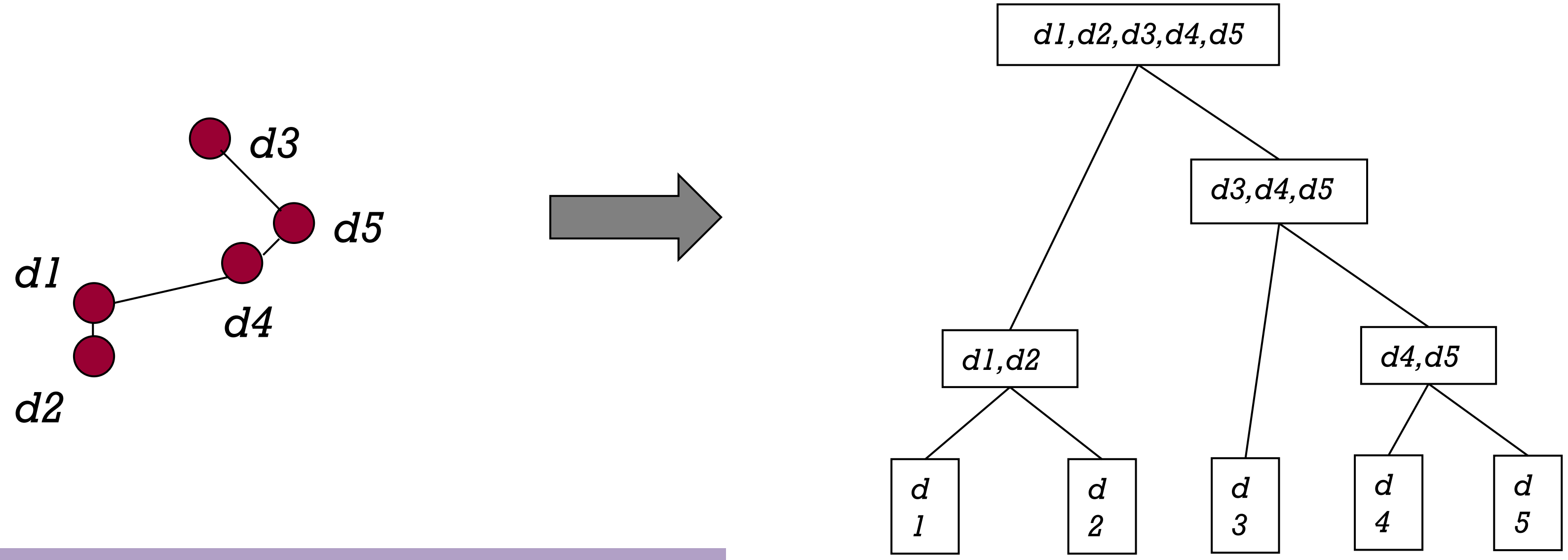
Hierarchical (Agglomerative) clustering

Hierarchical Clustering

- Many applications require hierarchical clustering of data
  - Clustering web documents into topical hierarchy (Yahoo!, Dmoz directories)
- The hierarchy obtained during the clustering process is called "***Dendrogram***"
- A specific clustering is obtained by cutting-off the dendrogram at desired level
  - The connected components form the clusters
- No need to know the number of clusters a-priori



Hierarchical Agglomerative Clustering (HAC)



Key Point

How to define inter-cluster similarities?

Example of agglomerative clustering

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	223	1308	802	2815	2934	2786	1771
DC	429	223	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

At each iteration, pick two data points that have least distance between them. Add the points into a cluster.

https://jameskle.com/writes/clustering

	BOS/NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY	0	223	1308	802	2815	2934	2786	1771
DC	223	0	1075	671	2684	2799	2631	1616
MIA	1308	1075	0	1329	3273	3053	2687	2037
CHI	802	671	1329	0	2013	2142	2054	996
SEA	2815	2684	3273	2013	0	808	1131	1307
SF	2934	2799	3053	2142	808	0	379	1235
LA	2786	2631	2687	2054	1131	379	0	1059
DEN	1771	1616	2037	996	1307	1235	1059	0

Note how we update distances between other clusters. The lower distance is picked. Distance between BOS to DC was 429, now set to 223. Distance from BOS to MIA was 1504, now set to 1308.

Averaging may also be used instead of taking distance to the closest point.

	BOS/NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	379	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0

	BOS/NY/DC	MIA	CHI	SEA	SF/LA	DEN
BOS/NY/DC	0	1075	671	2684	2631	1616
MIA	1075	0	1329	3273	2687	2037
CHI	671	1329	0	2013	2054	996
SEA	2684	3273	2013	0	808	1307
SF/LA	2631	2687	2054	808	0	1059
DEN	1616	2037	996	1307	1059	0

Note how creation of SF/LA cluster has changed distances

	BOS/NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	379	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0

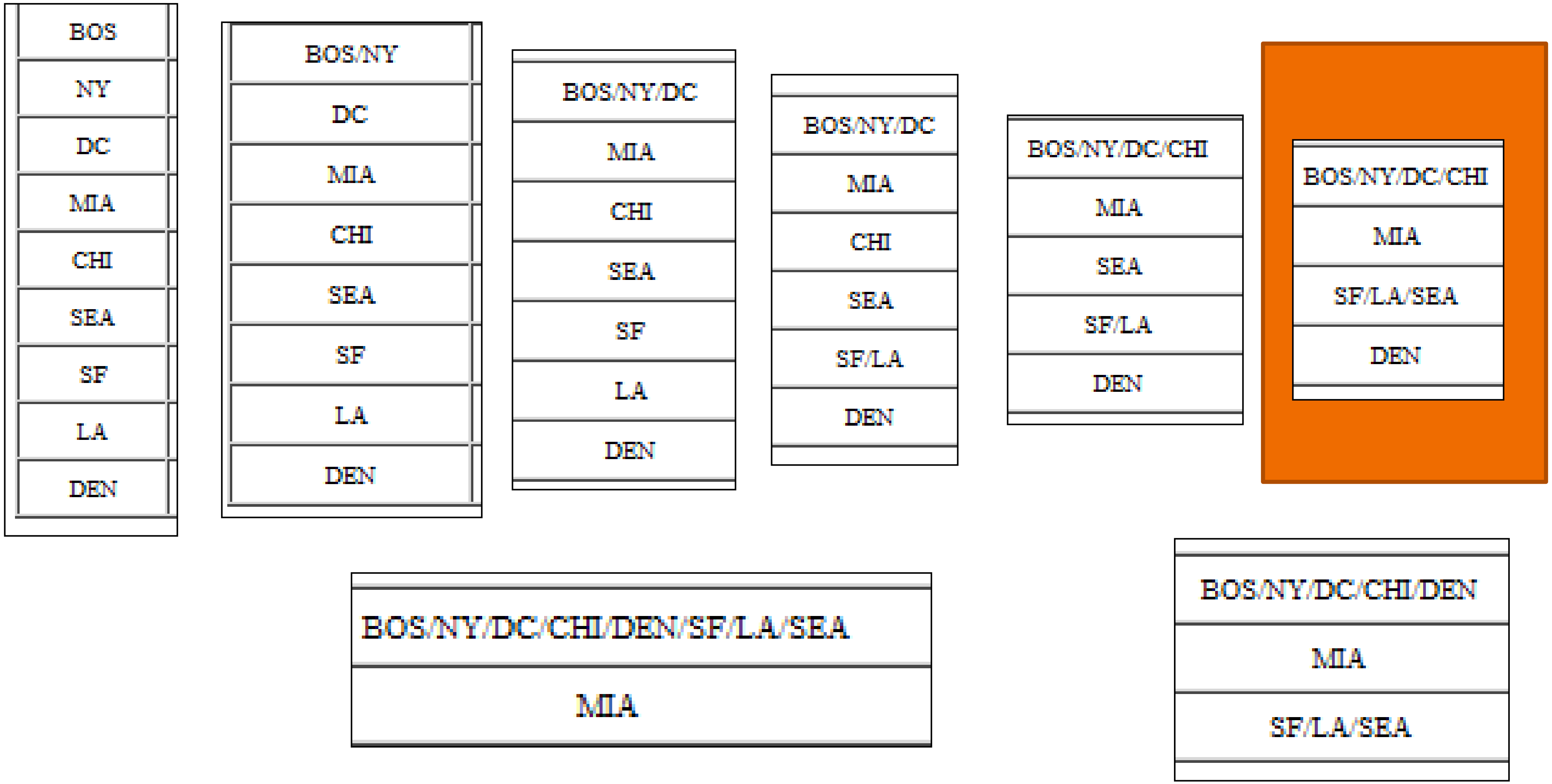
	BOS/NY/DC/ CHI	MIA	SEA	SF/LA	DEN
BOS/NY/DC/CHI	0	1075	2013	2054	996
MIA	1075	0	3273	2687	2037
SEA	2013	3273	0	808	1307
SF/LA	2054	2687	808	0	1059
DEN	996	2037	1307	1059	0

	BOS/NY/DC/CHI	MIA	SF/LA/SEA	DEN
BOS/NY/DC/CHI	0	1075	2013	996
MIA	1075	0	2687	2037
SF/LA/SEA	2054	2687	0	1059
DEN	996	2037	1059	0

	BOS/NY /DC/CHI/DEN	MIA	SF/LA/SEA
BOS/NY/DC/CHI/DEN	0	1075	1059
MIA	1075	0	2687
SF/LA/SEA	1059	2687	0

	BOS/NY /DC/CHI /DEN/SF /LA/SEA	MIA
BOS/NY/DC/CHI/DEN/SF/LA/SEA	0	1075
MIA	1075	0

Agglomerative clustering (Hierarchical)

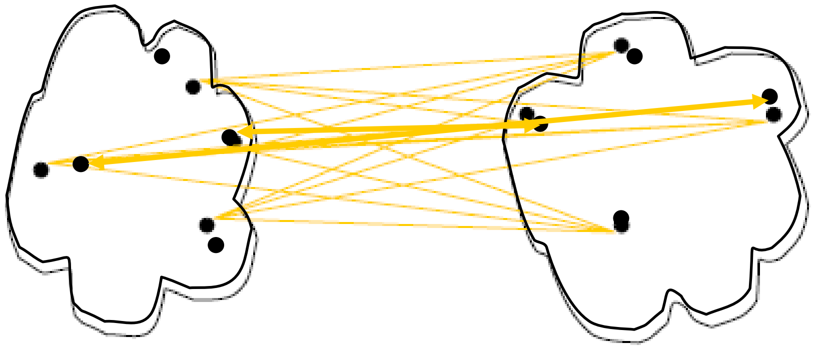


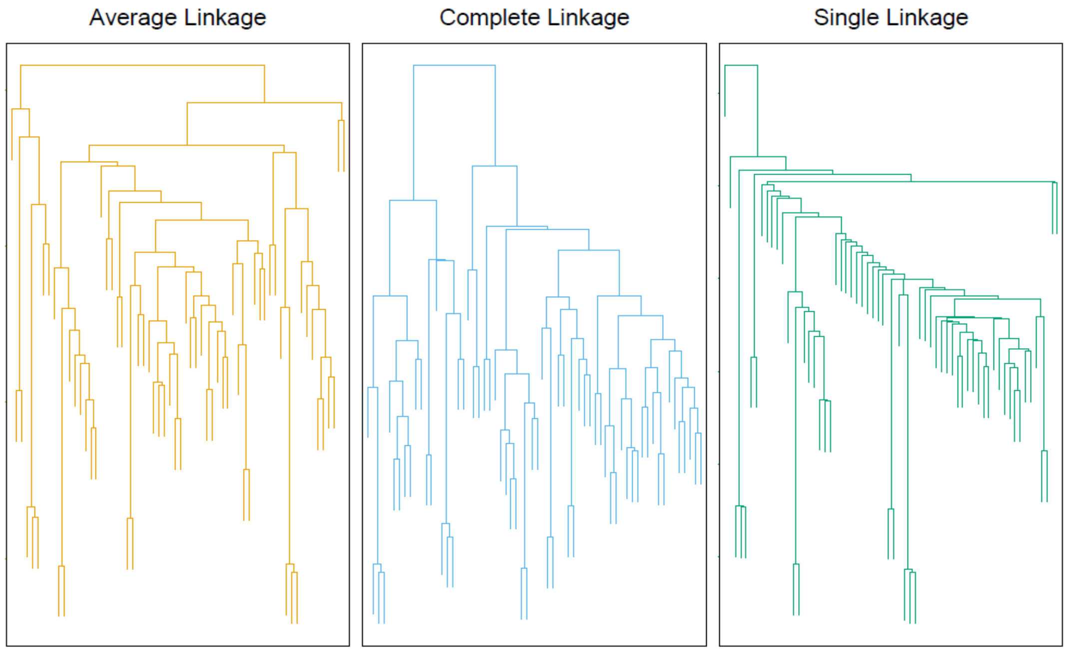
Typically, a particular "level" of the hierarchy is selected to be your clustering result

Highlighted clusters divide airports into North-East, Central, South and Pacific areas

Inter-Cluster Distance Functions

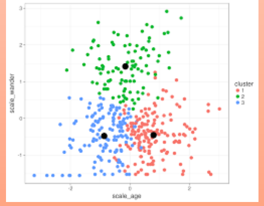
- Single-linkage (MIN)
  - ***Minimum distance*** between any two points across clusters
- Complete-linkage (MAX)
  - ***Maximum distance*** between any two points across clusters
- Average-linkage (AVG)
  - ***Average distance*** between the points across clusters





Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

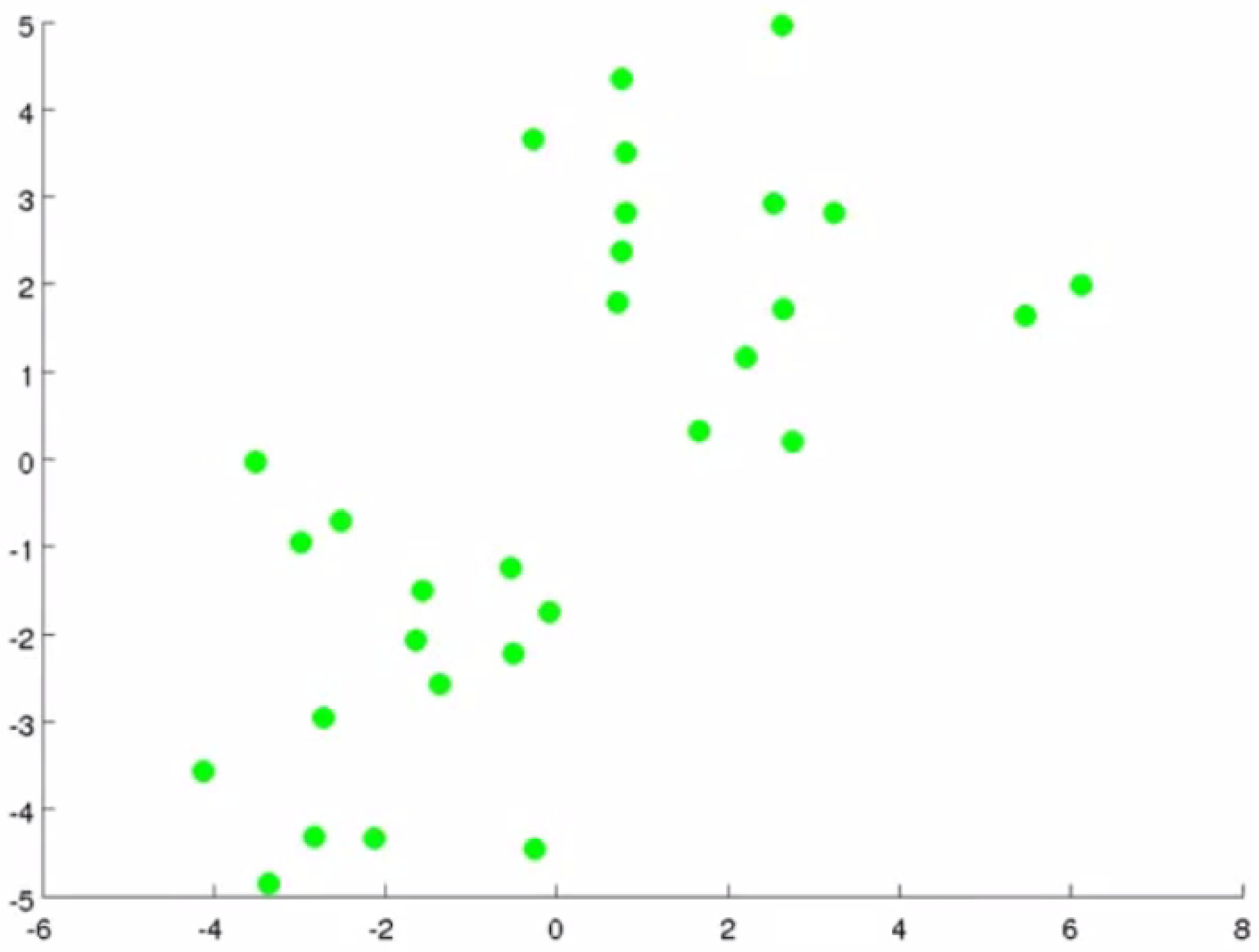
- ✓ Clustering
- ✓ Hierarchical clustering
- • KMeans



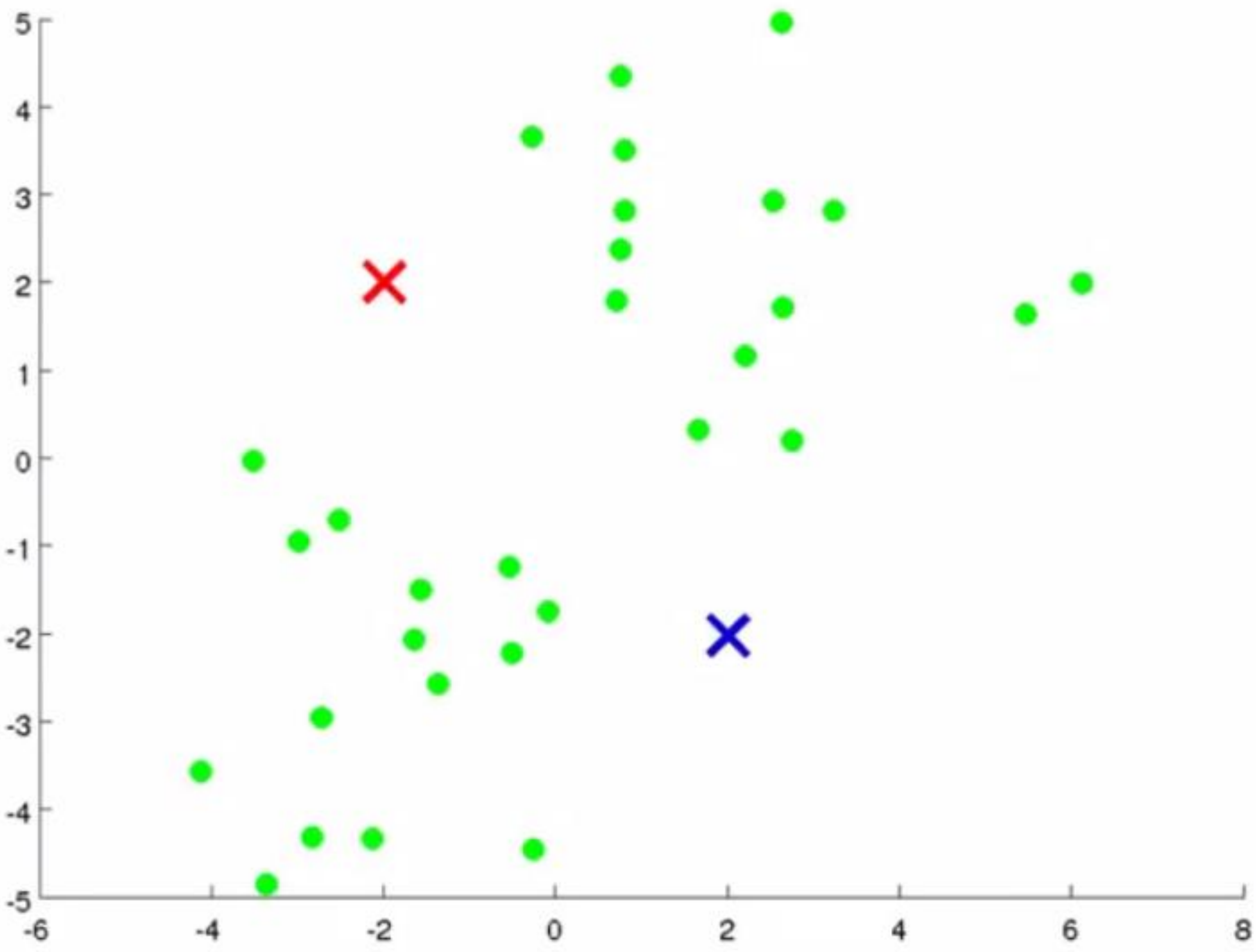
KMeans

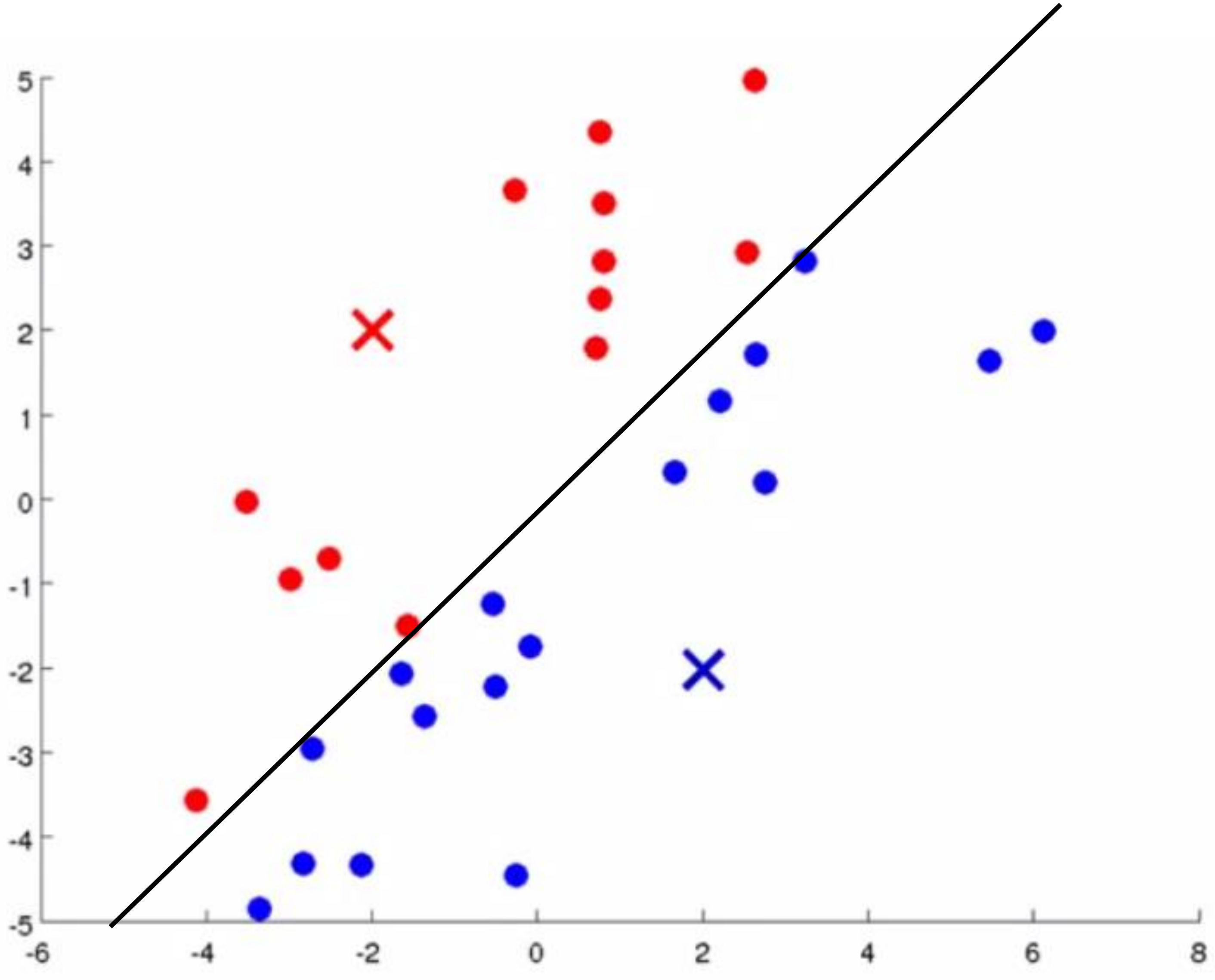
K-means clustering

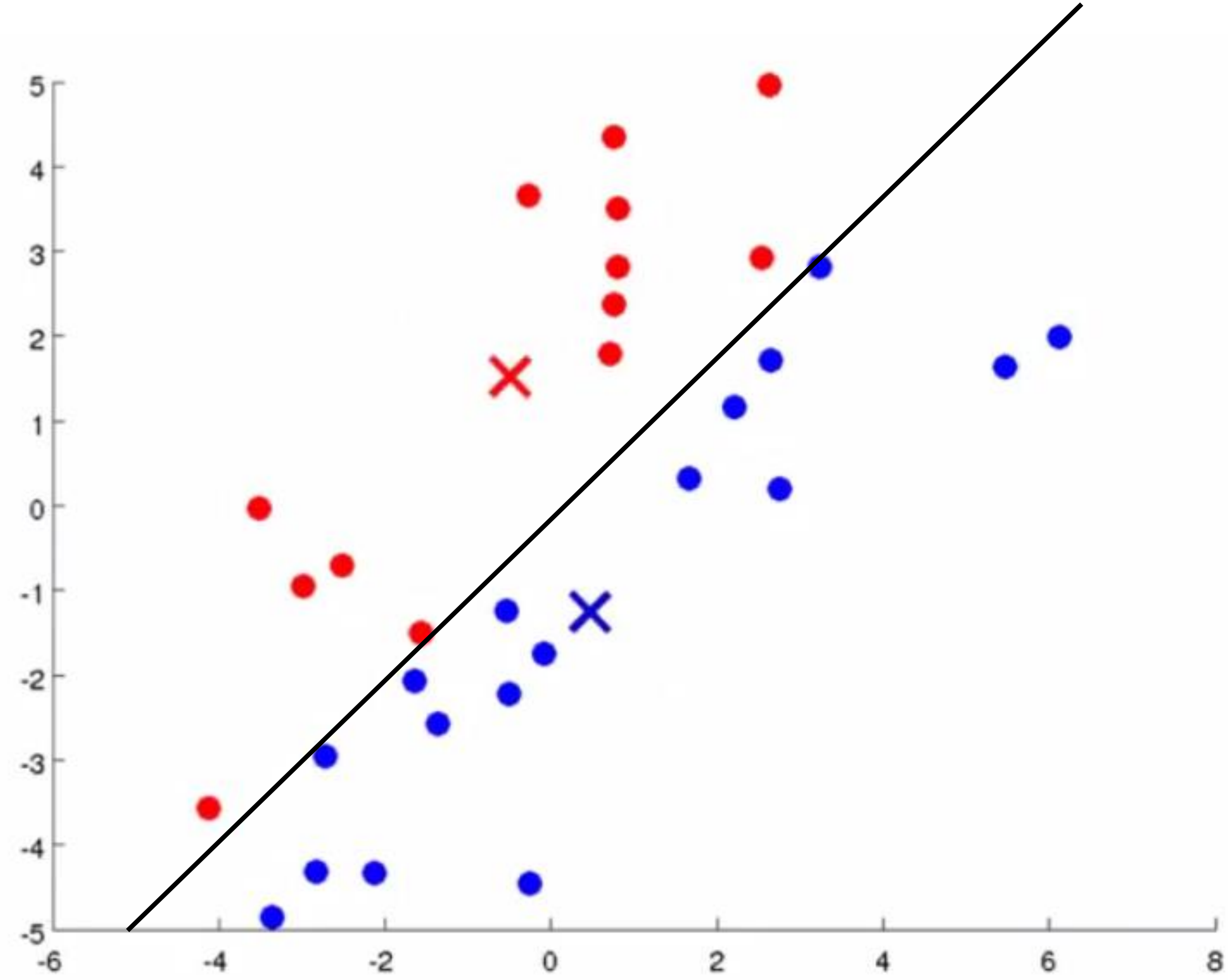
- K-means is a partitional clustering algorithm as it partitions the given data into k clusters.
  - Each cluster has a cluster center, called centroid.
  - k is specified by the user

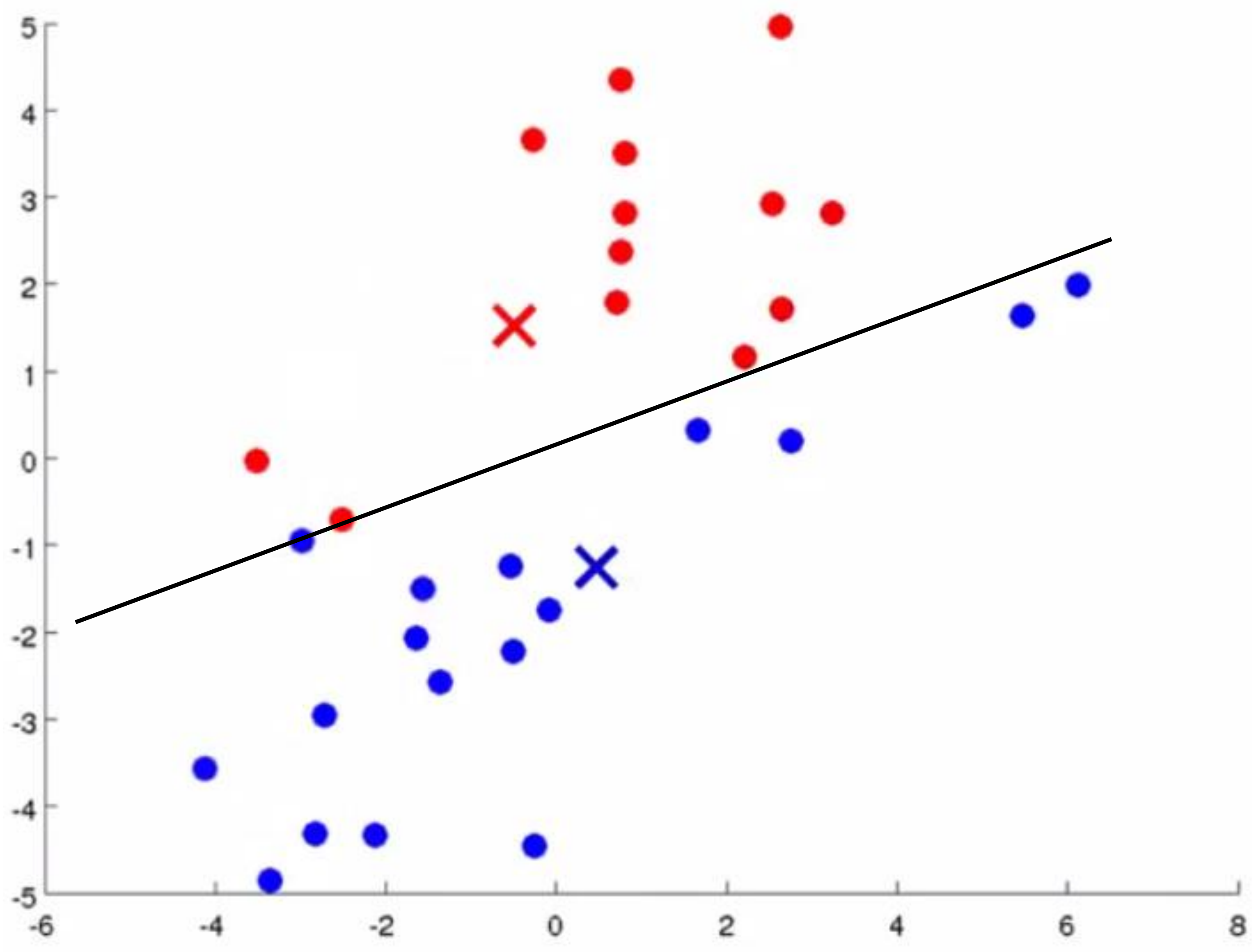


Pictures courtesy Andrew Ngs course on Coursera

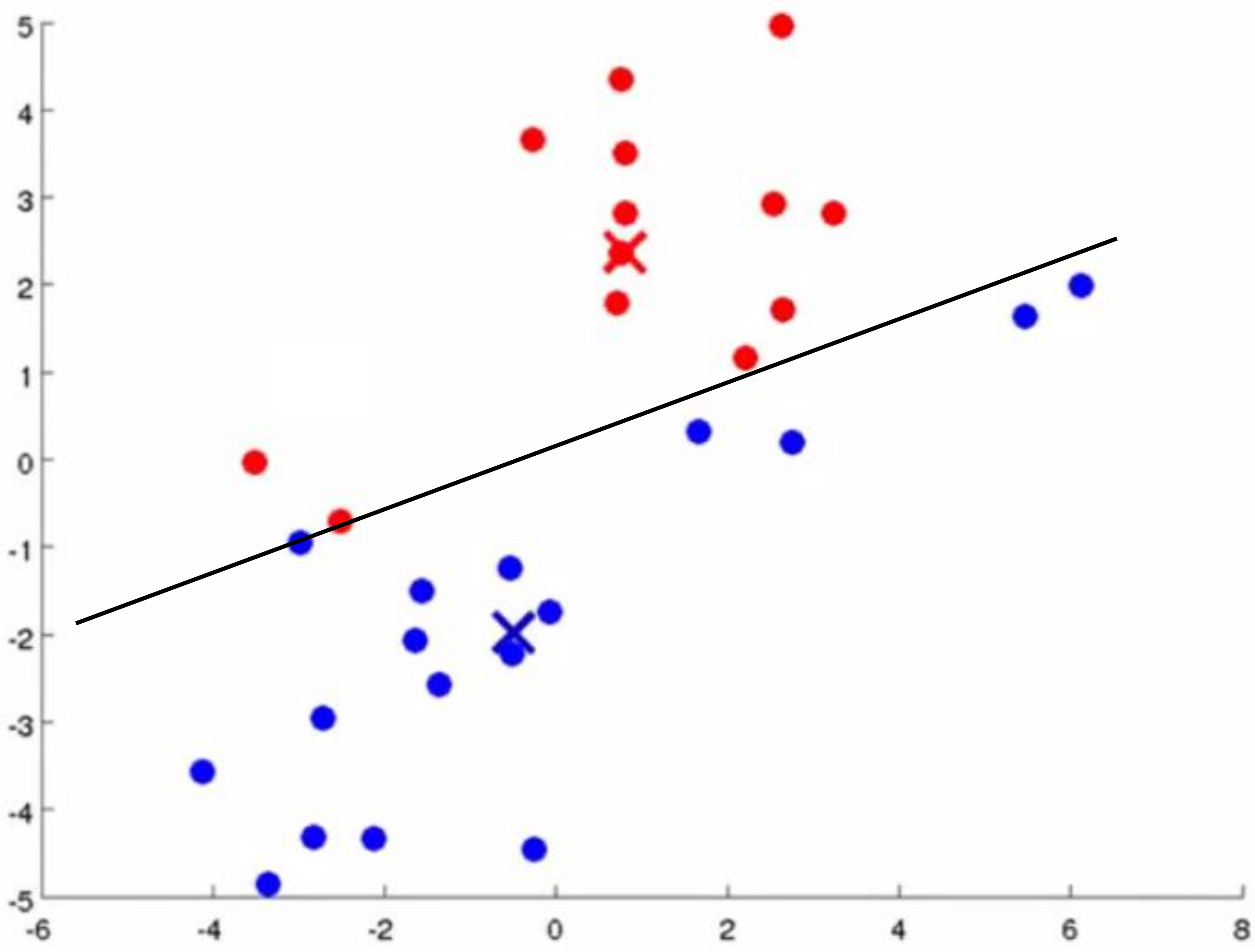

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

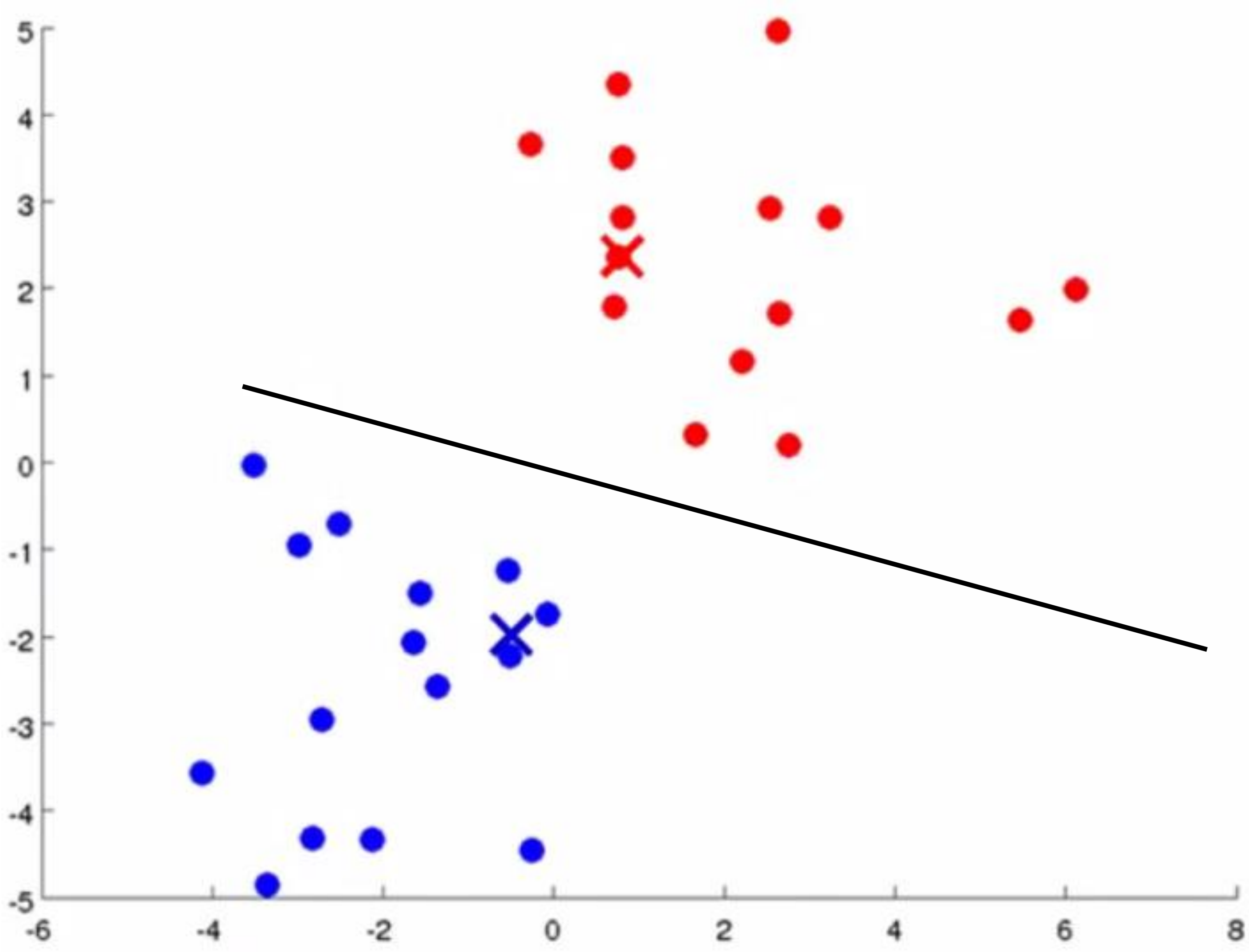

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

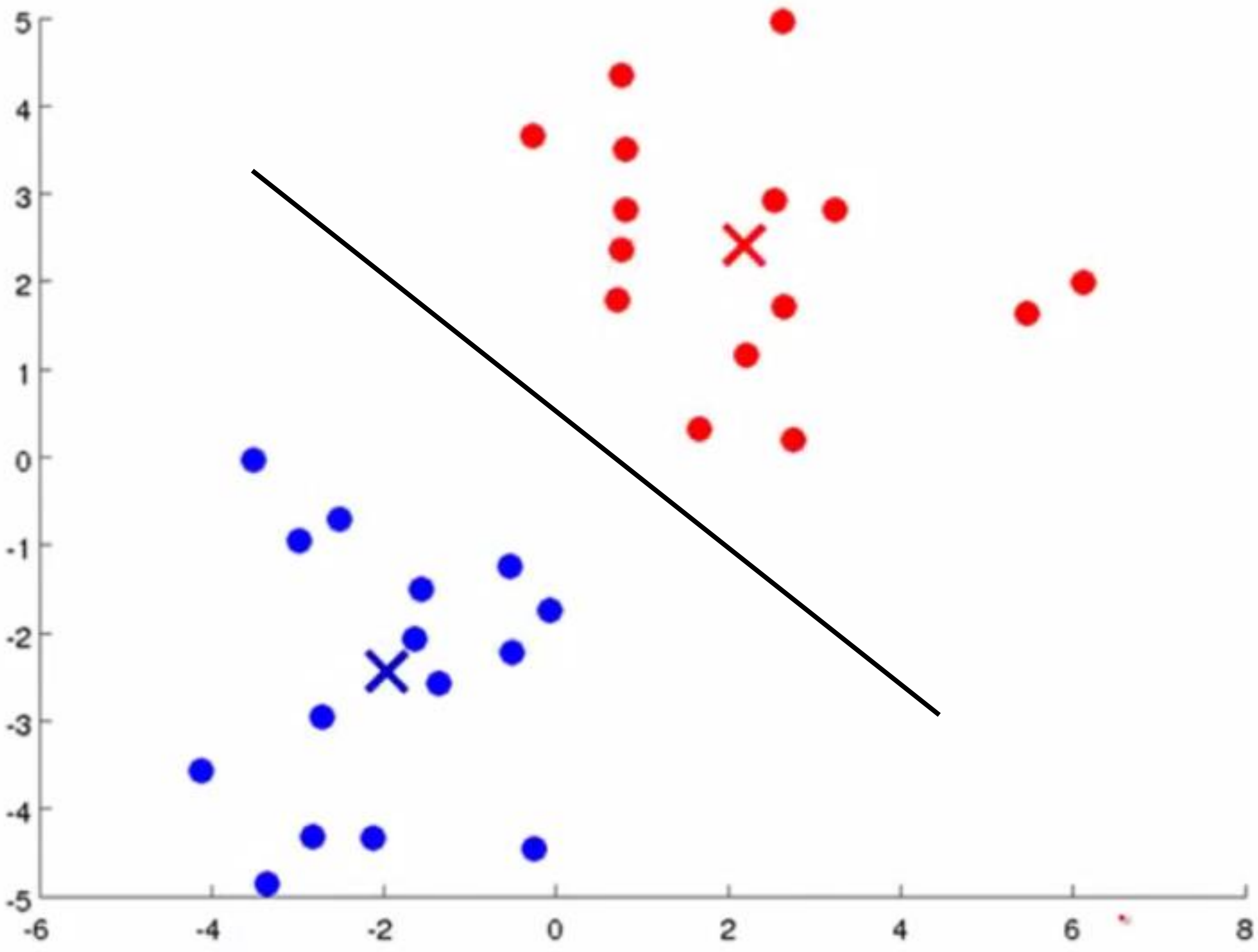

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8



5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

K-means algorithm

- Given k, the k-means algorithm works as follows:
  1. Randomly choose k data points (seeds) to be the initial centroids, cluster centers
  2. Assign each data point to the closest centroid
  3. Re-compute the centroids using the current cluster memberships.
  4. If a convergence criterion is not met, go to 2.

Stopping/convergence criterion

1. no (or minimum) re-assignments of data points to different clusters,
2. no (or minimum) change of centroids, or
3. minimum decrease in the sum of squared error (SSE),

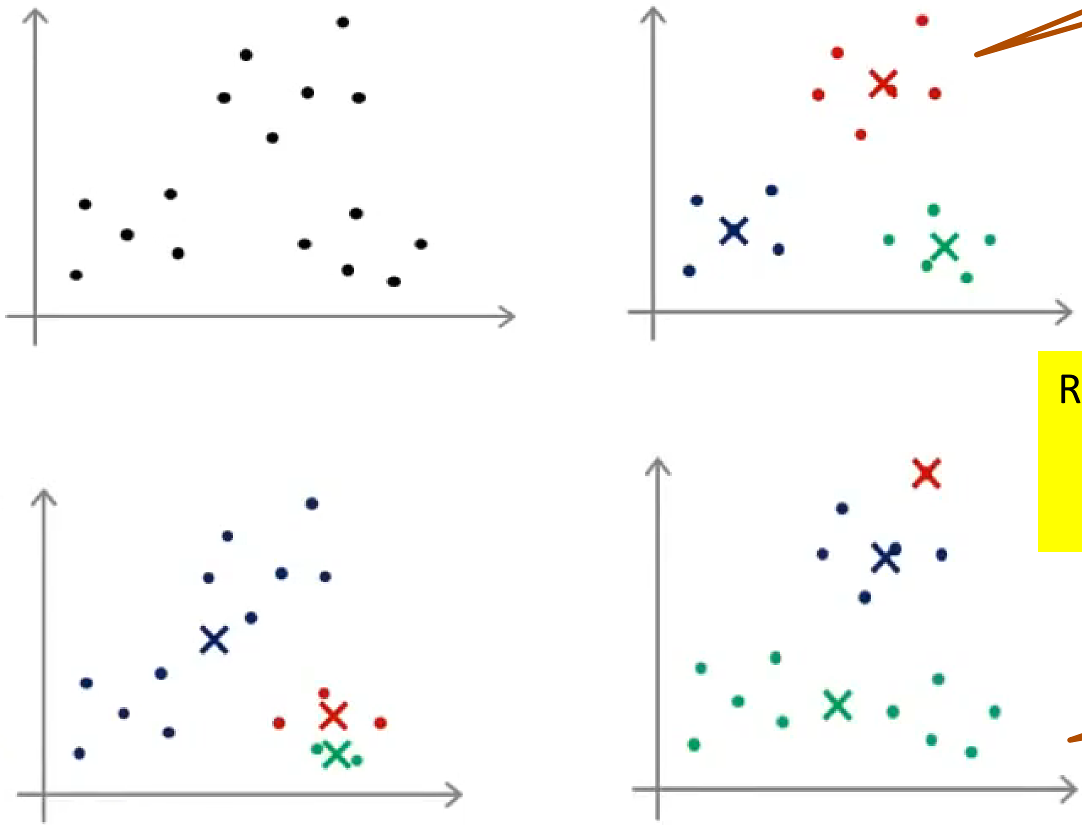
$$SSE = \sum_{j=1\ldots k} \sum_{x \epsilon C_j} dist(x, m_j)^2 \qquad (1)$$

- C_j is the jth cluster, $\mathbf{m}_j$ is the centroid (mean) of cluster C_j

Check SCREE plots.

http://stackoverflow.com/questions/15376075/cluster-analysis-in-r-determine-the-optimal-number-of-clusters

Preventing Local Optima



Good cluster

Randomly initialize K-means.

Run K-means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.

Compute cost function (distortion)

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

Repeat above multiple times (usually 50-1000 times)

Choose the clustering assignment with $\min J$

Sub-optimal clusters

References/Resources

- Classic and Modern Data Clustering
  - http://learning.stat.purdue.edu/mlss/_media/mlss/meila.pdf
- AutoLab Tutorial on Gaussian Mixture Models
  - http://www.autonlab.org/tutorials/gmm14.pdf
- Andrew Ng's Lecture on CourseEra
  - https://class.coursera.org/ml-003/lecture/preview
- The Elements of Statistical Learning – Data Mining, Inference, and Prediction
  - http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html
  - http://www.math.unipd.it/~dulli/corso04/ng94efficient.pdf
  - https://anuradhasrinivas.files.wordpress.com/2013/04/lesson8-clustering.pdf
  - http://www.vlfeat.org/overview/kmeans.html
  - http://repository.cmu.edu/cgi/viewcontent.cgi?article=2397&context=compsci
  - http://www.cs.ucsb.edu/~veronika/MAE/Global_Kernel_K-Means.pdf

Summary

Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- ✓ Clustering
- ✓ Hierarchical clustering
- ✓ K-means

Thanks!!

Questions?