CS 60050
Machine Learning

Unsupervised Learning

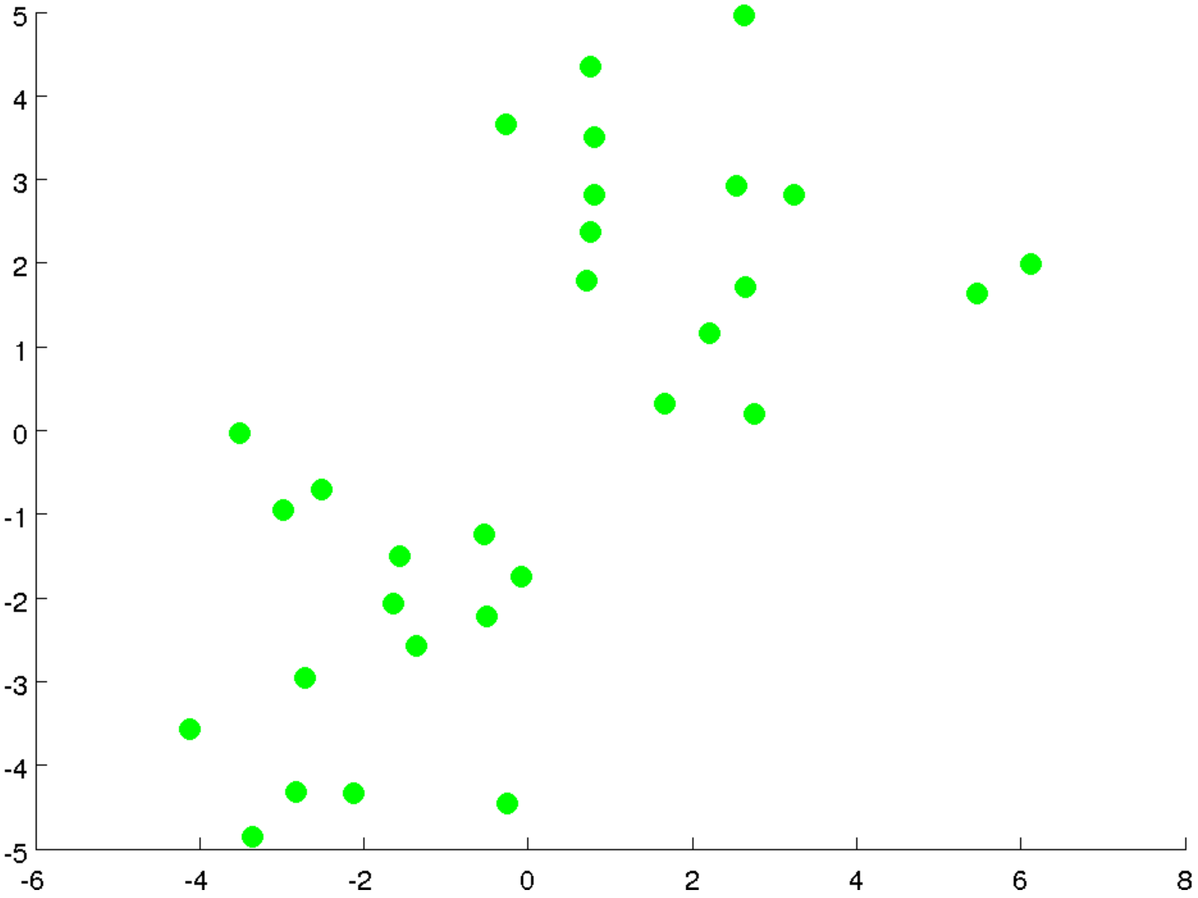
Some slides taken from course materials of Andrew Ng

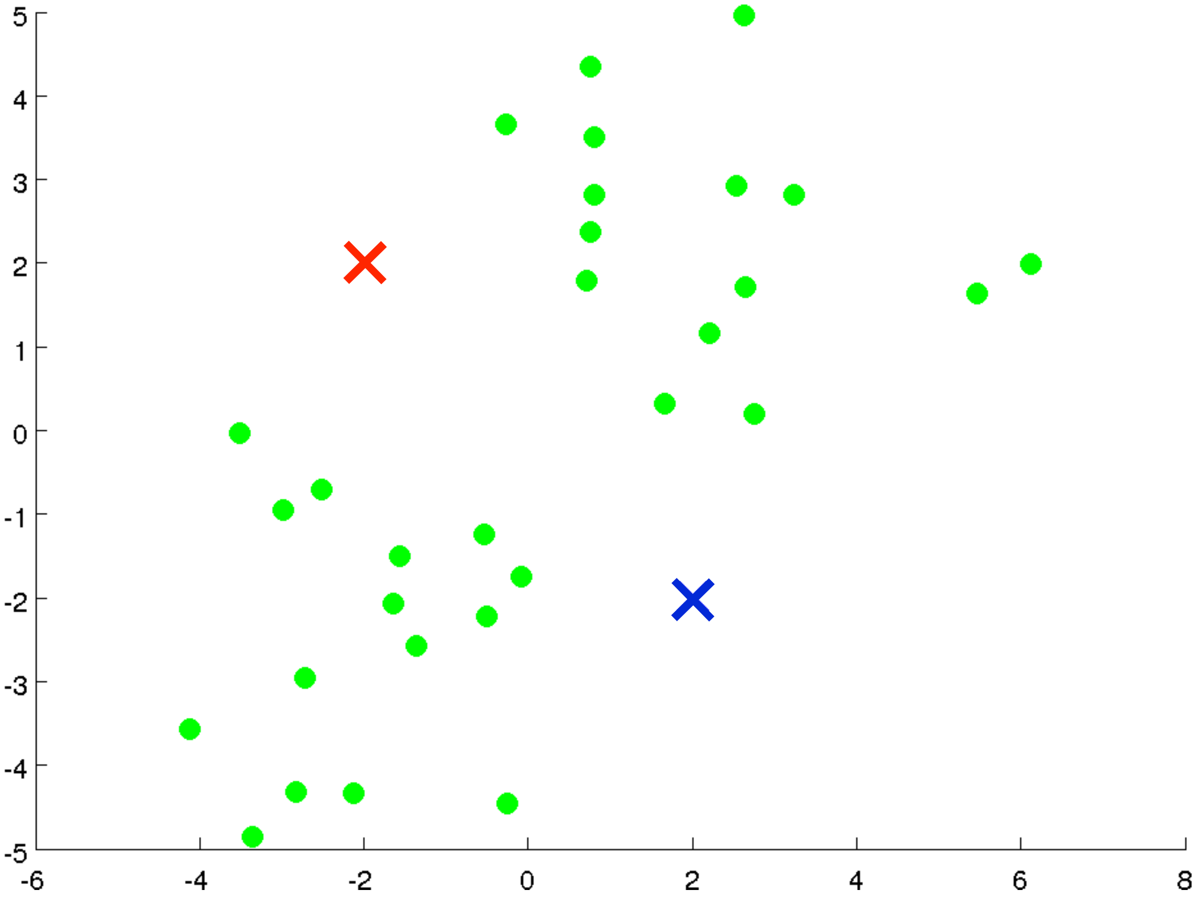
Unsupervised learning

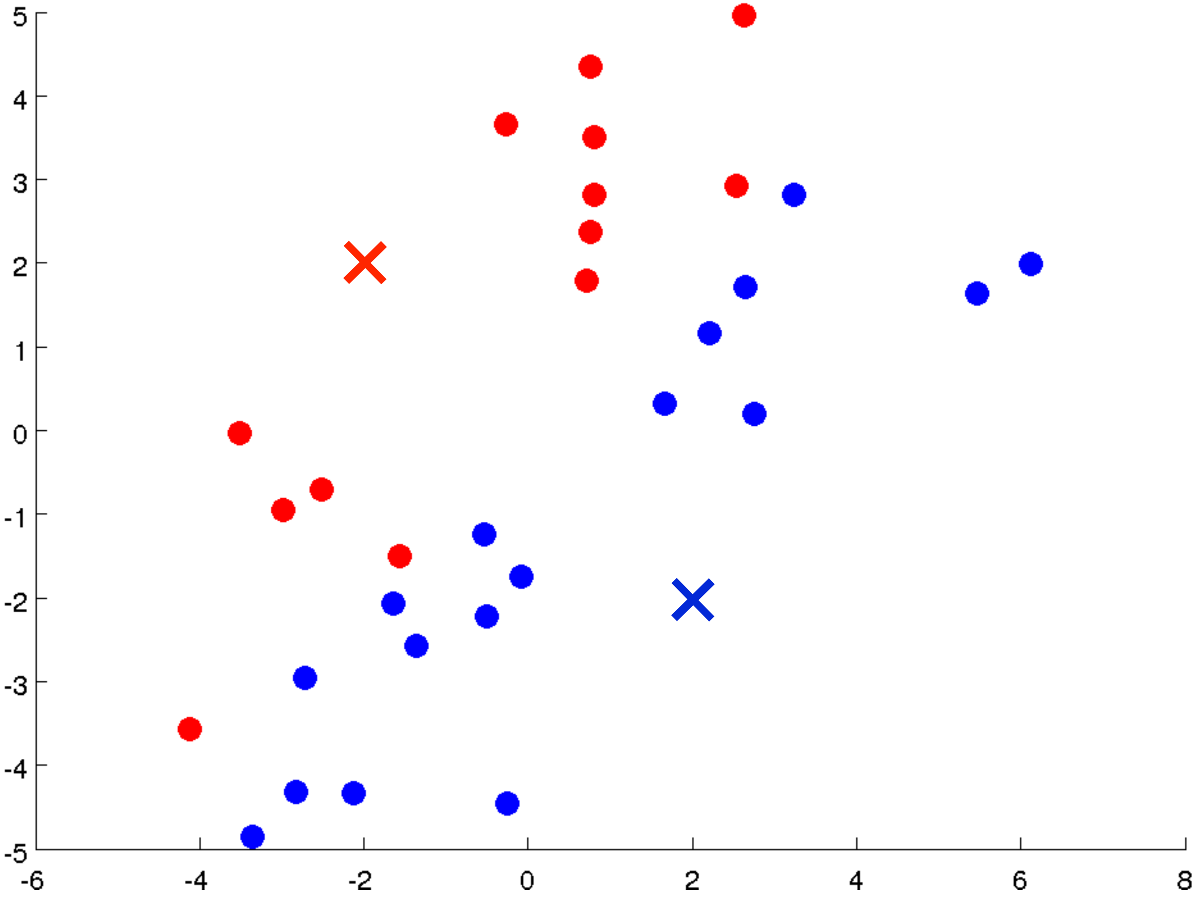
- Given a set of unlabeled data points
- Find patterns or structure in the data

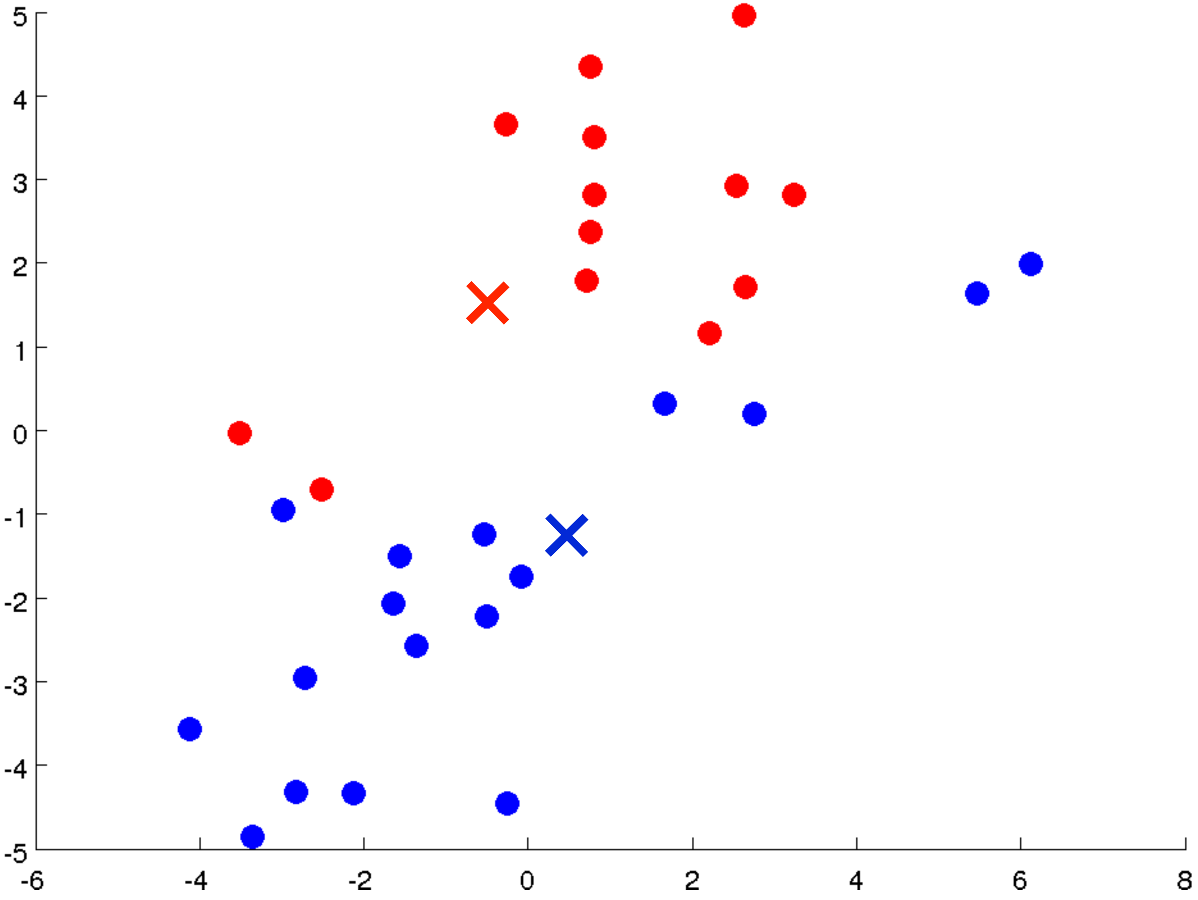
- Clustering: automatically group the data into groups of 'similar' points
- A popular algorithm: K-means clustering

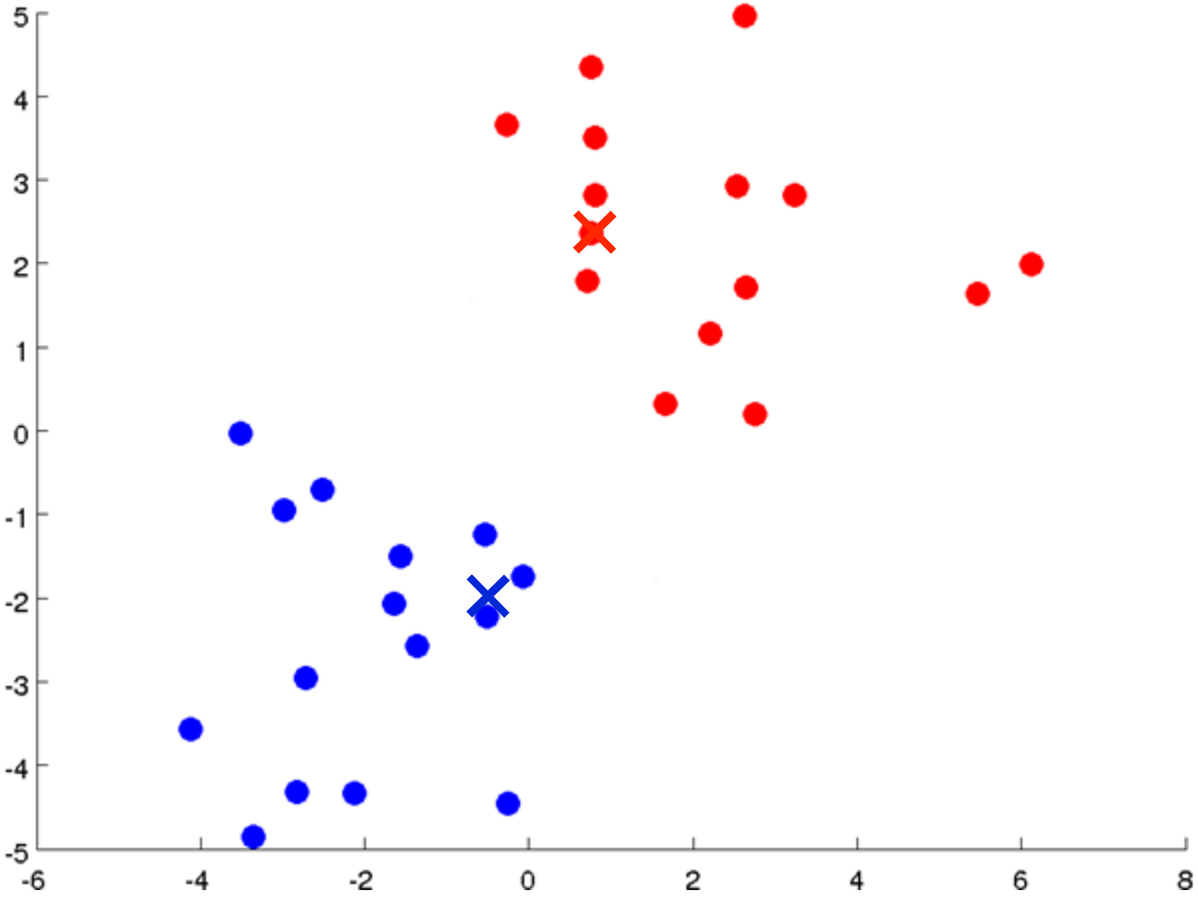
K-means clustering

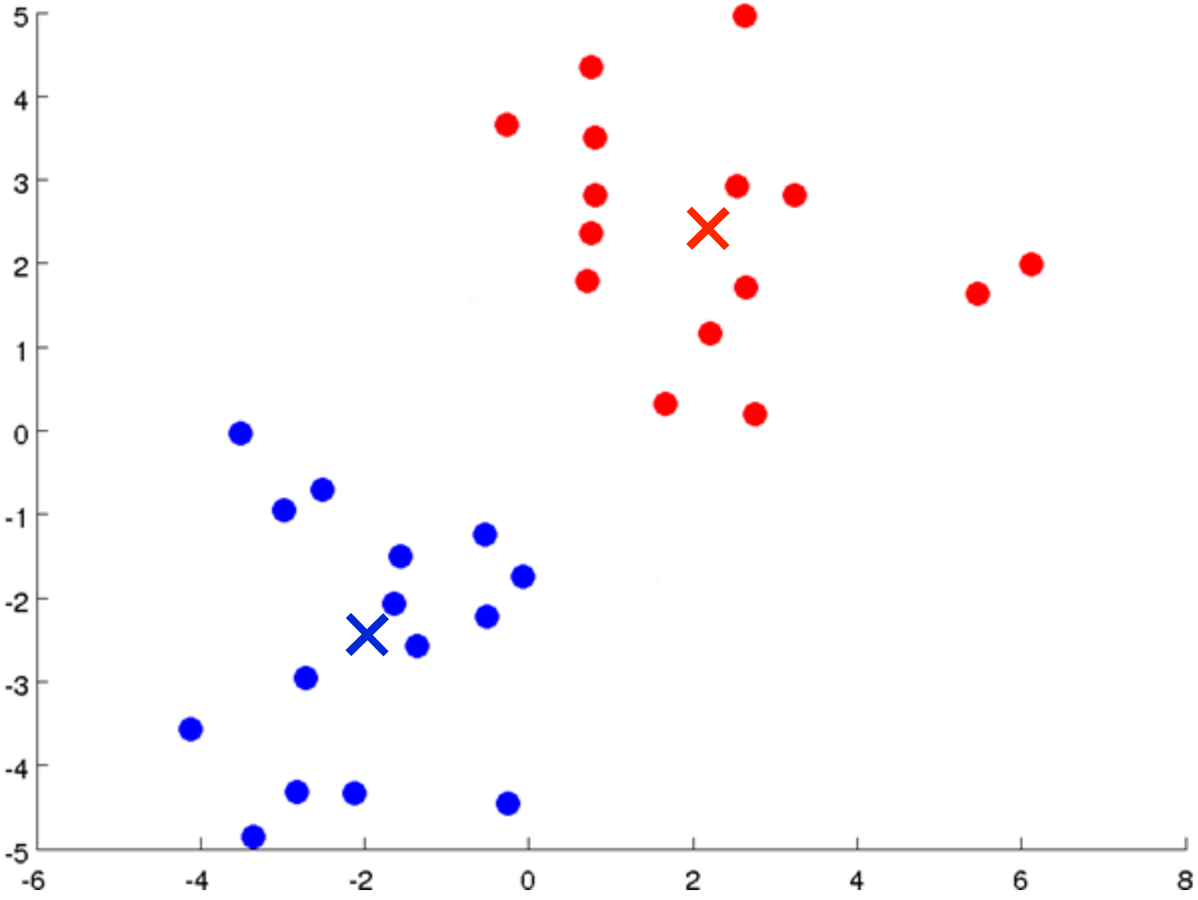

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


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

```
Randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$
Repeat {
             for $i$ = 1 to $m$
Cluster          $c^{(i)}$ := index (from 1 to $K$) of cluster centroid
assignment                closest to $x^{(i)}$
             for $k$ = 1 to $K$
Move             $\mu_k$ := average (mean) of points assigned to cluster $k$
centroid



             }
```

Choosing value of K

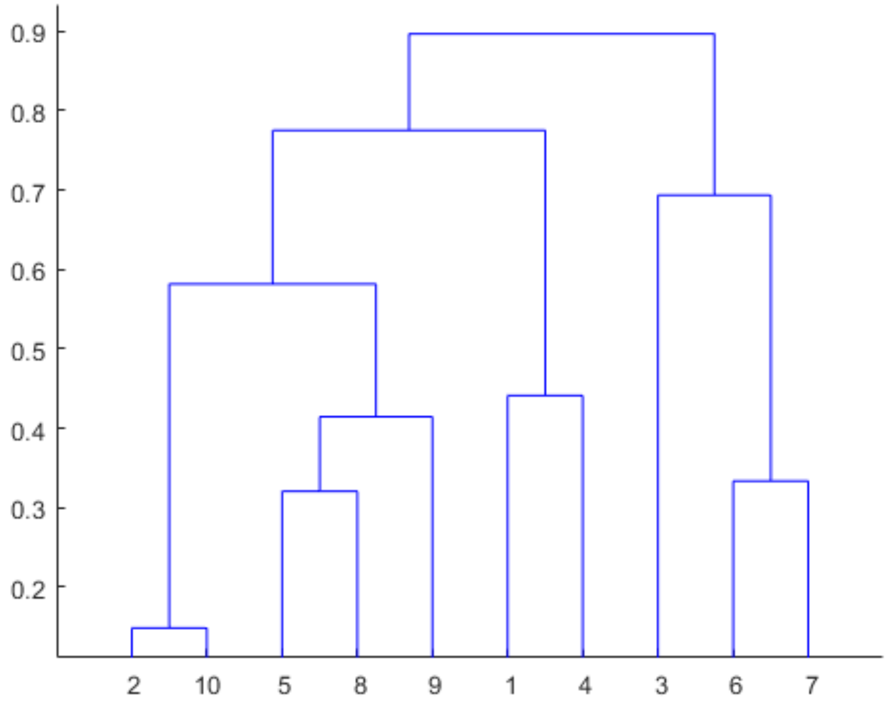
- Based on some measure of cluster quality (how good the clusters are)

- Based on domain knowledge about suitable number of clusters for a particular problem domain

Other types of clustering

- Hierarchical clustering
  - Treat each data point as a singleton cluster
  - Successively merge clusters until all points have been merged into a single remaining cluster.
  - Often represented as a dendrogram

Dendrogram



Types of hierarchical clustering

- Complete linkage
  - merge in each step the two clusters with the smallest **maximum** pairwise distance
- Single linkage
  - merge in each step the two clusters with the smallest **minimum** pairwise distance.
- Refer https://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-clustering-1.html

Other unsupervised learning problems

- Principal Component Analysis (PCA) – can be used to reduce number of features, by selecting few

- Topic modeling, e.g., Latent Dirichlet Allocation (LDA) – discover abstract "topics" that occur in a collection of documents