Unsupervised learning

* Only observe features
* Goal to discovery things about measurements
  + Informative way to visualize date
  + Discover subgroups among variables or observations
* Two methods:
  + Principal components analysis:
    - Tool used for data visualization or data pre-processing before supervised techniques are applied
  + Clustering:
    - Broad class of methods for discovering unknown subgroups
* Challenge:
  + More subjective, as there is no simple goal like prediction
  + Easier to obtain unlabeled data than labeled data, which can require intervention

Principal Components Analysis:

* Low-dimension representation of data set
  + Finds sequence of linear combination of variables that have maximal variance and are mutually uncorrelated
  + Tool for data reduction
  + Also tool for data visualization
* First PC loading vector:
  + Defines line in p-dimensional space that is closest to n observation

Clustering

* Very broad set of techniques for finding subgroups
* Seek partition of data into groups so that observations within each group are similar
* Must define what it means for observations to be similar
* Looks for homogenous subgroups among observations
* K-means
  + Partition observations into pre-specified number of cluster
    - Each observation belongs to 1 cluster
    - Clusters are non-overlapping
  + Goal is to minimize with-in cluster varation summed over all clusters
* Hierarhical clustering
  + End up with tree-like visual representation of observations called dendogram that allows us to view clusterings obtained for each possible number of clusters