Unsupervised Learning: Clustering

Some material adapted from slides by Andrew Moore, CMU. See http://www.autonlab.org/tutorials/ for a repository of Data Mining tutorials

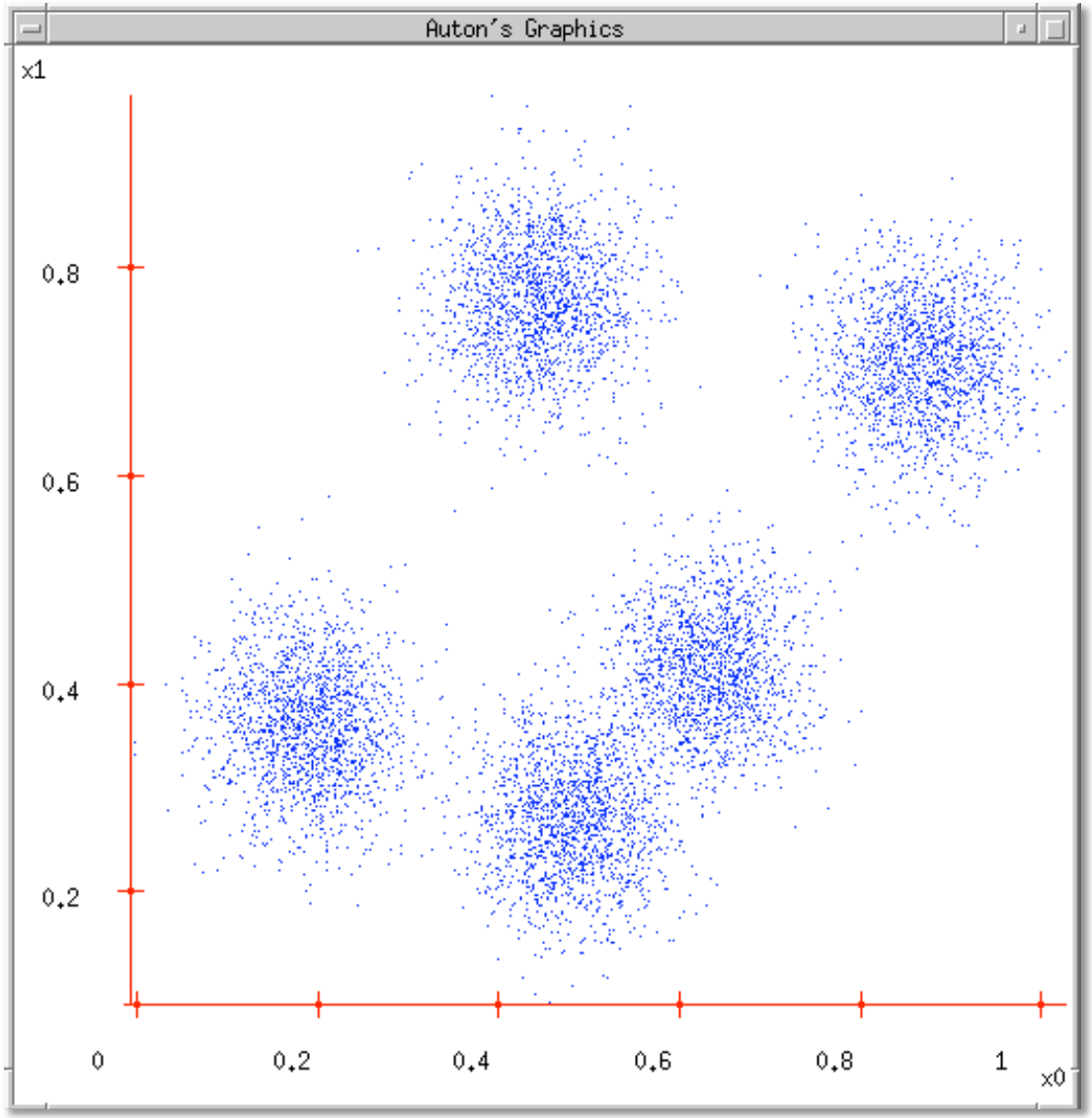
Unsupervised Learning

- Supervised learning used labeled data pairs (x, y) to learn a function f : X→Y.
- But, what if we don't have labels?
- No labels = **unsupervised learning**
- Only some points are labeled = **semi-supervised learning**
  - Getting labels may be expensive, so we only get a few
- **Clustering** is the unsupervised grouping of data points. It can be used for **knowledge discovery**.

Top-down vs. Bottom Up

- Clustering is typically done using a distance measure defined between instances
- The distance is defined in the instance feature space
- Agglomerative approach works bottom up:
  - Treat each instance as a cluster
  - Merge the two closest clusters
  - Repeat until the stop condition is met
- Top-down approach starts a cluster with all instances
  - Find a cluster to split into two or more smaller clusters
  - Repeat until stop condition met

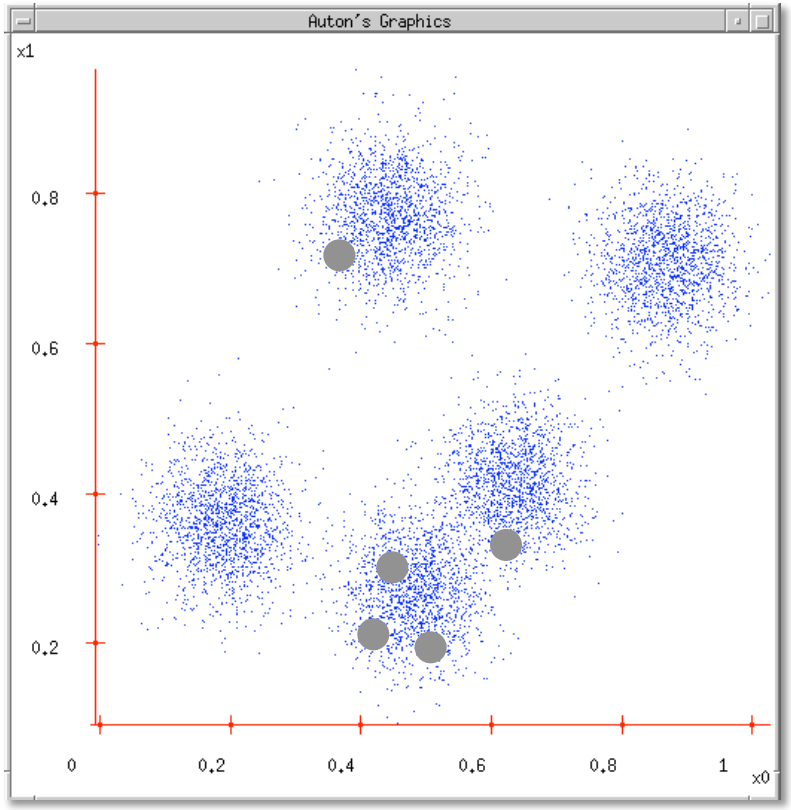
Clustering Data



K-Means Clustering

K-Means (k , data)

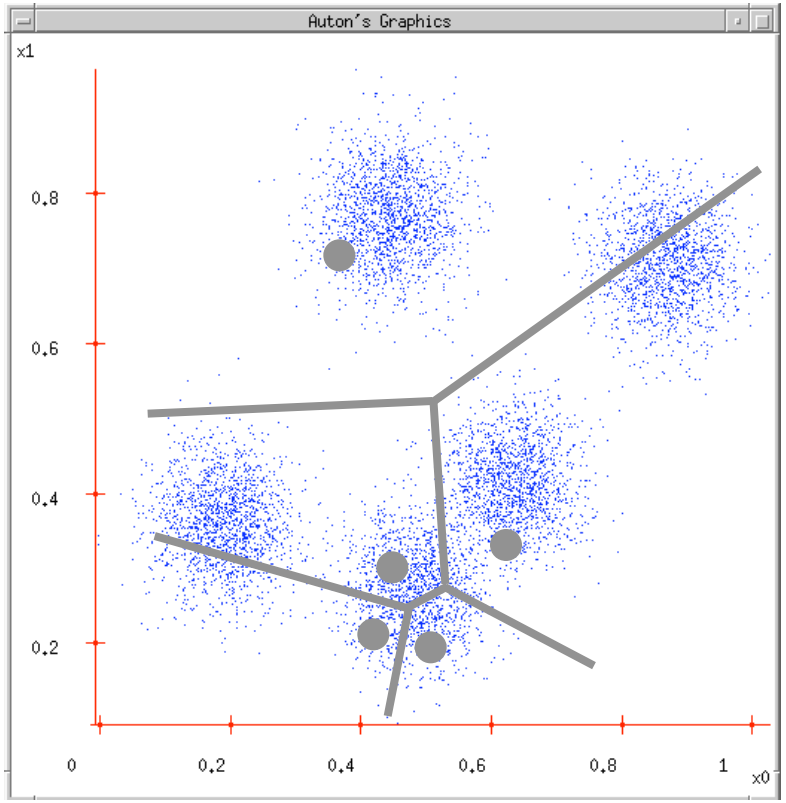
- Randomly choose k cluster center locations (centroids).
- Loop until convergence
  - Assign each point to the cluster of the closest centroid.
  - Reestimate the cluster centroids based on the data assigned to each.



K-Means Clustering

K-Means (k , data)

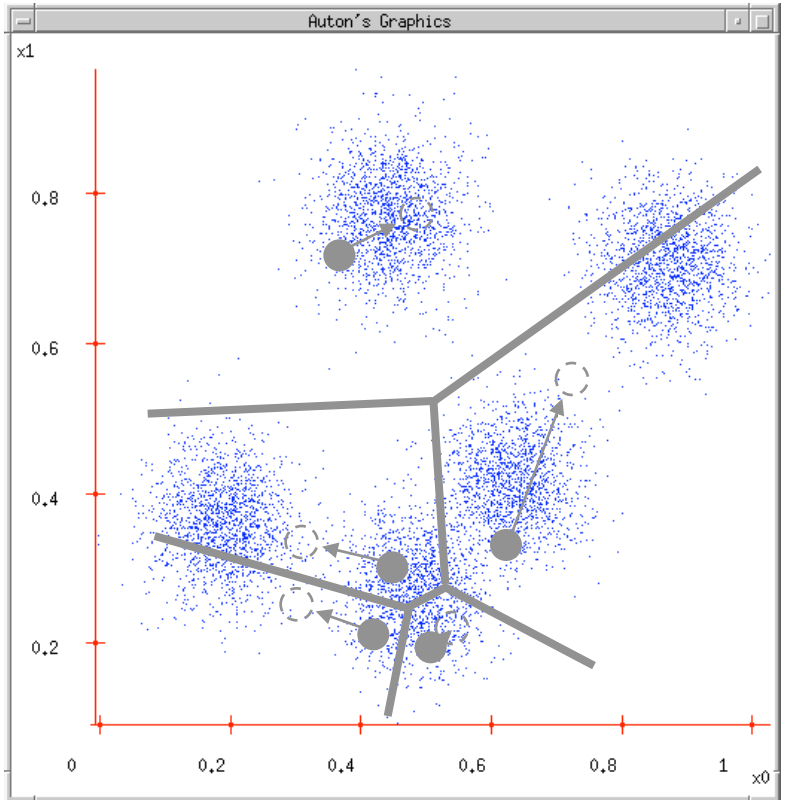
- Randomly choose k cluster center locations (centroids).
- Loop until convergence
  - Assign each point to the cluster of the closest centroid.
  - Reestimate the cluster centroids based on the data assigned to each.



K-Means Clustering

K-Means (k , data)

- Randomly choose k cluster center locations (centroids).
- Loop until convergence
  - Assign each point to the cluster of the closest centroid.
  - Reestimate the cluster centroids based on the data assigned to each.



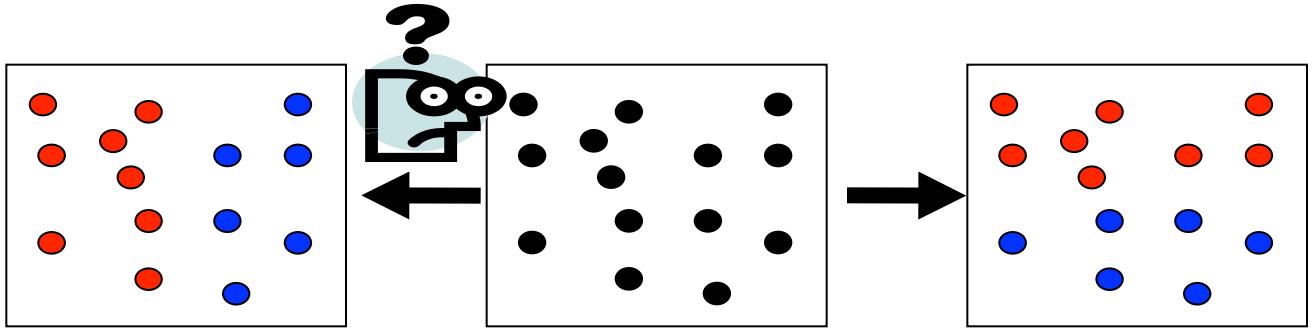
Problems with K-Means

- ***Very*** sensitive to the initial points
  - Do many runs of k-Means, each with different initial centroids.
  - Seed the centroids using a better method than random. (e.g. Farthest-first sampling)

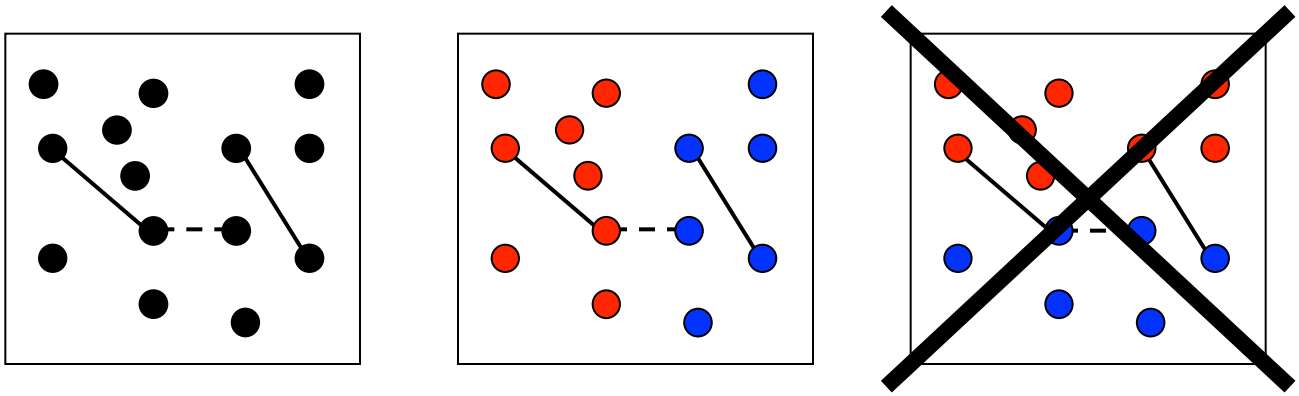
- Must manually choose k
  - Learn the optimal k for the clustering. (Note that this requires a performance measure.)

Problems with K-Means

- How do you tell it which clustering you want?



– Constrained clustering techniques



Same-cluster constraint (must-link)

Different-cluster constraint (cannot-link)