

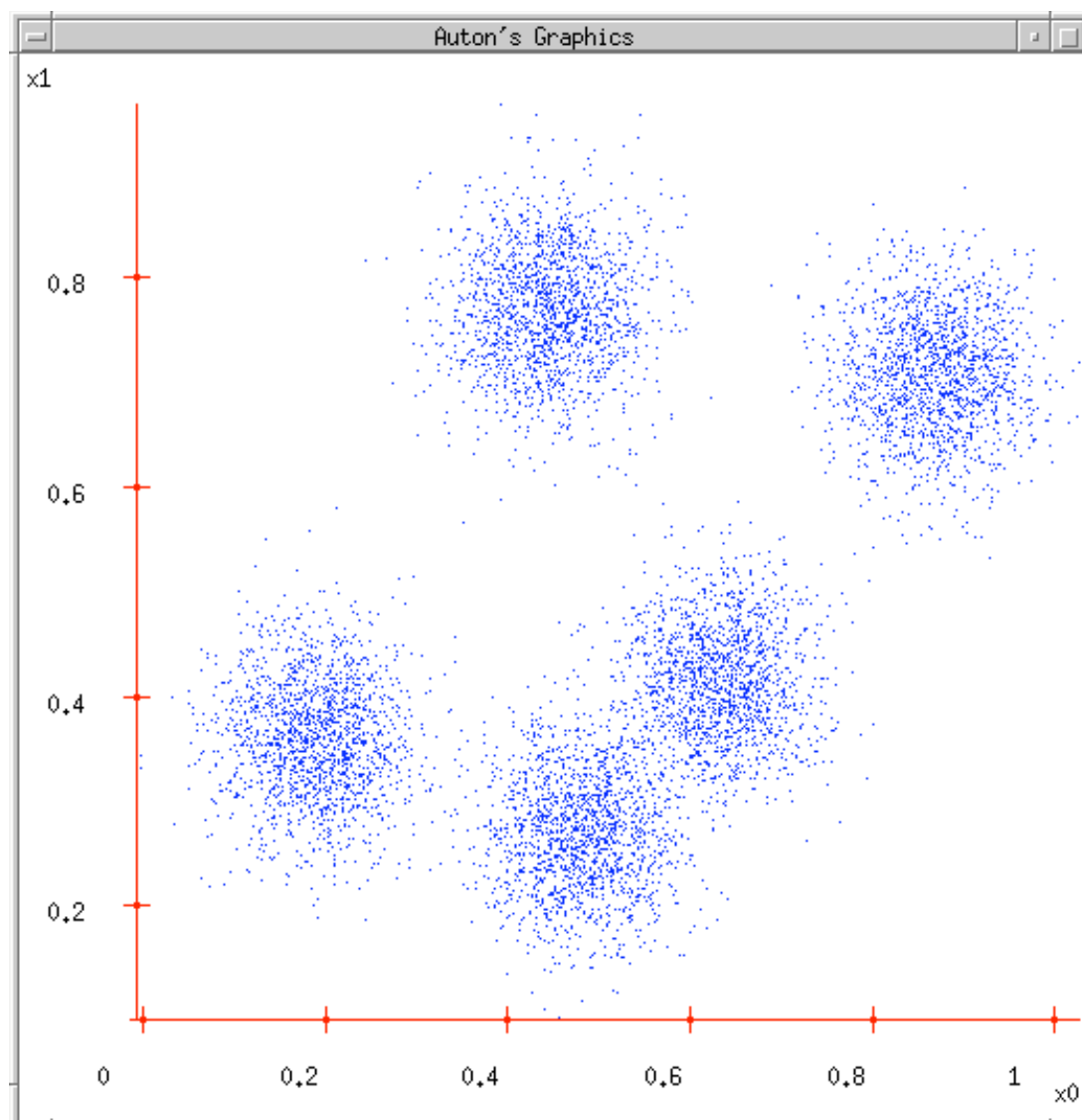
Unsupervised Learning: K-Means Clustering

Some material adapted from slides by Andrew Moore, CMU.

Unsupervised Learning

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

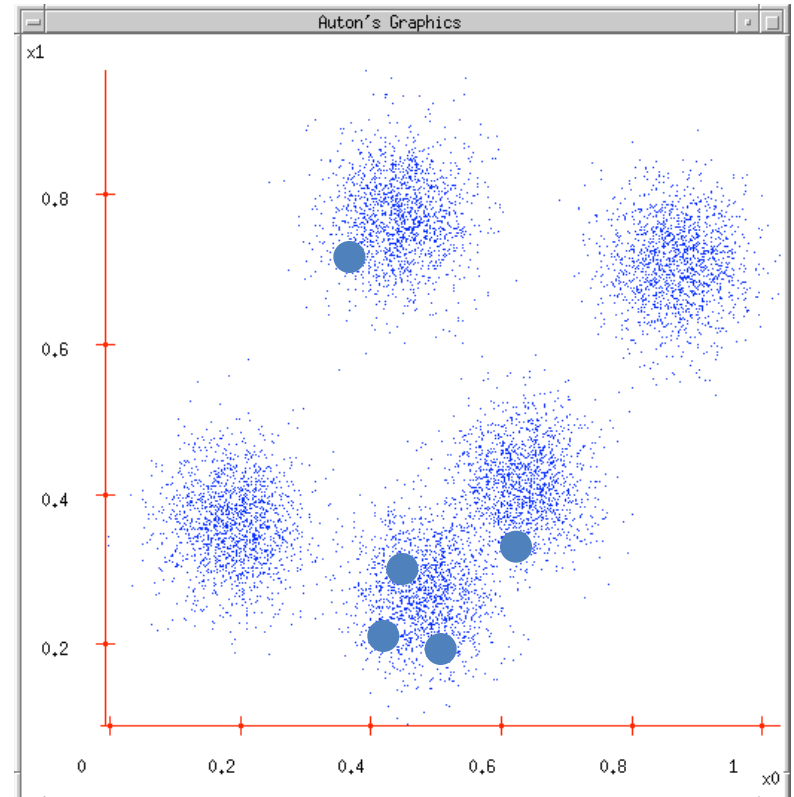
Clustering Data



K-Means Clustering

K-Means (k , data)

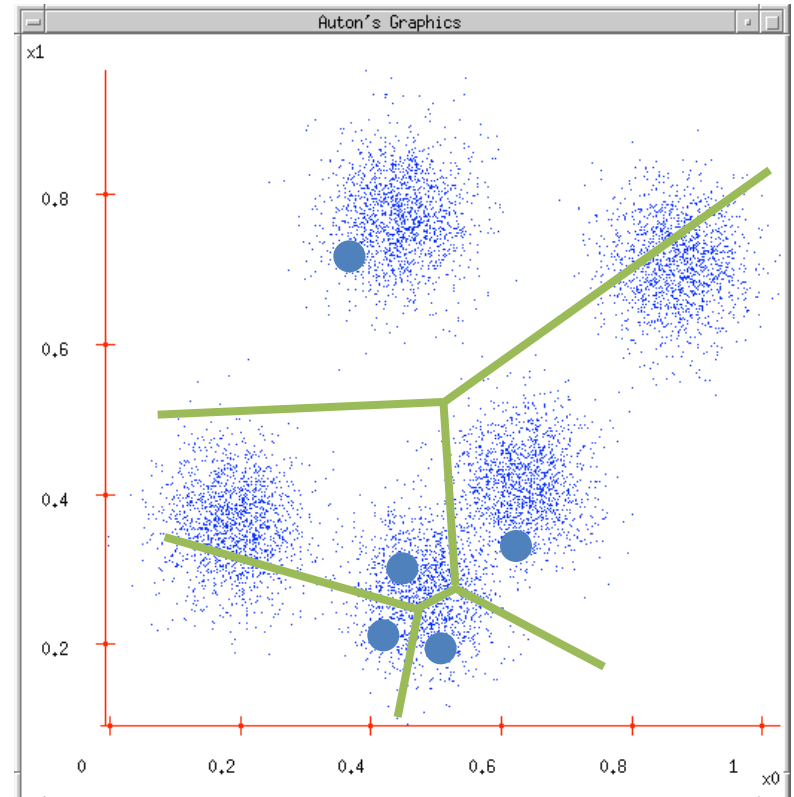
- Randomly choose k cluster center locations (centroids).
- Loop until convergence
 - Assign each point to the cluster of the closest centroid.
 - Reestimate the cluster centroids based on the data assigned to each.



K-Means Clustering

K-Means (k , data)

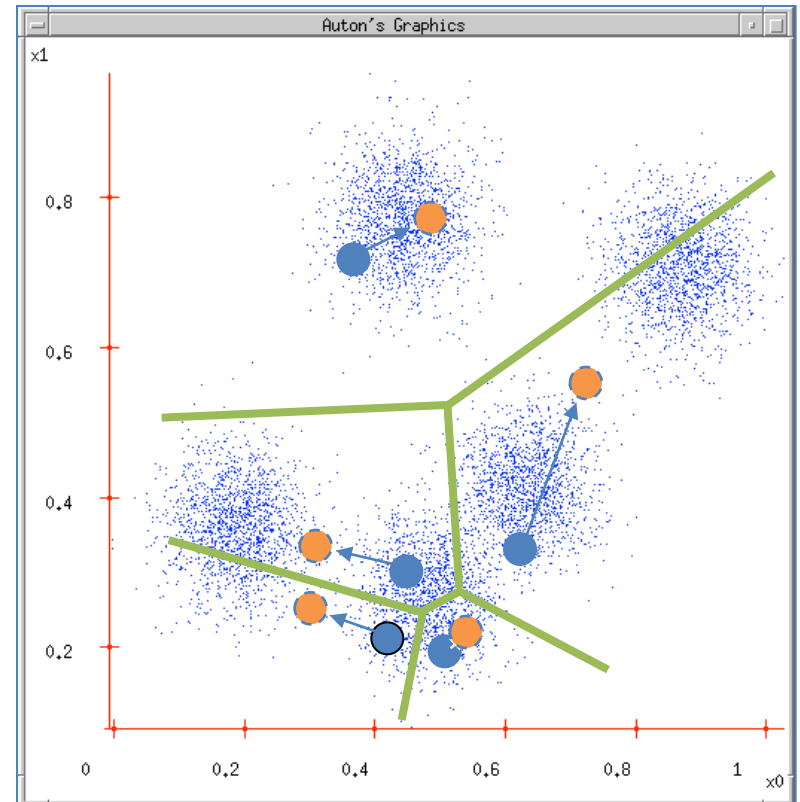
- Randomly choose k cluster center locations (centroids).
- Loop until convergence
 - Assign each point to the cluster of the closest centroid.
 - Reestimate the cluster centroids based on the data assigned to each.



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.



K-means Algorithm

- For a current set of cluster means, assign each observation as:

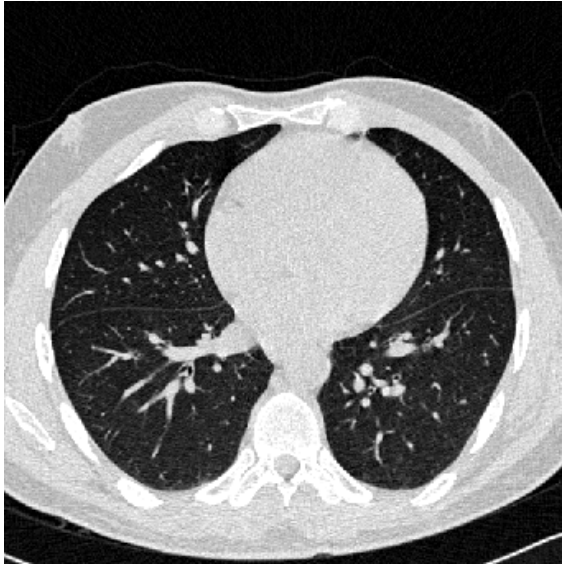
$$C(i) = \arg \min_{1 \leq k \leq K} \|x_i - m_k\|^2, \quad i = 1, \dots, N$$

- For a given assignment C , compute the cluster means m_k :

$$m_k = \frac{\sum_{i:C(i)=k} x_i}{N_k}, \quad k = 1, \dots, K.$$

- Iterate above two steps until convergence

Image Segmentation Results



An image (I)



Three-cluster image (J) on
gray values of I

Matlab code:

```
I = double(imread('...'));
```

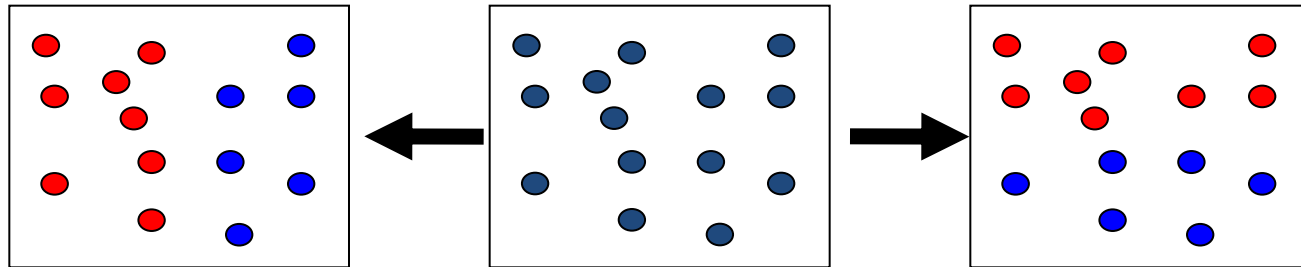
```
J = reshape(kmeans(I(:),3),size(I));
```


Problems with K-Means

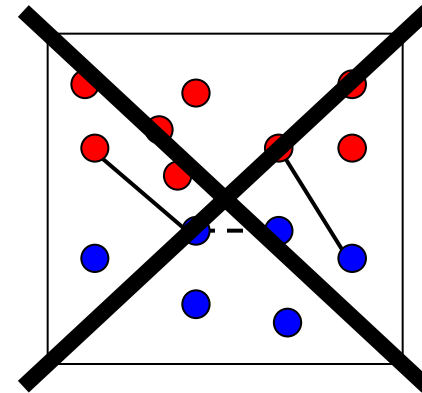
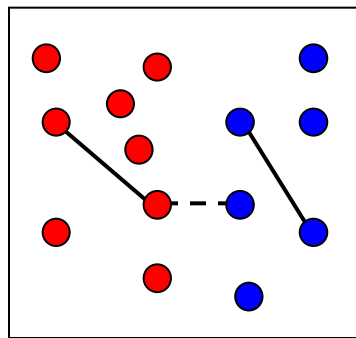
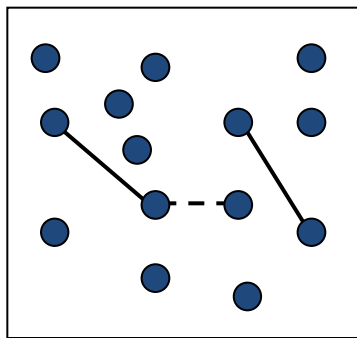
- **Very** sensitive to the initial points.
 - Do many runs of k-Means, each with different initial centroids.
 - Seed the centroids using a better method than random. (e.g. sampling point far apart)
- Must manually choose k.
 - Learn the optimal k for the clustering (meta-learning). (Note that this 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)