Unsupervised Learning: K-Means Clustering

Some material adapted from slides by Andrew Moore, CMU.

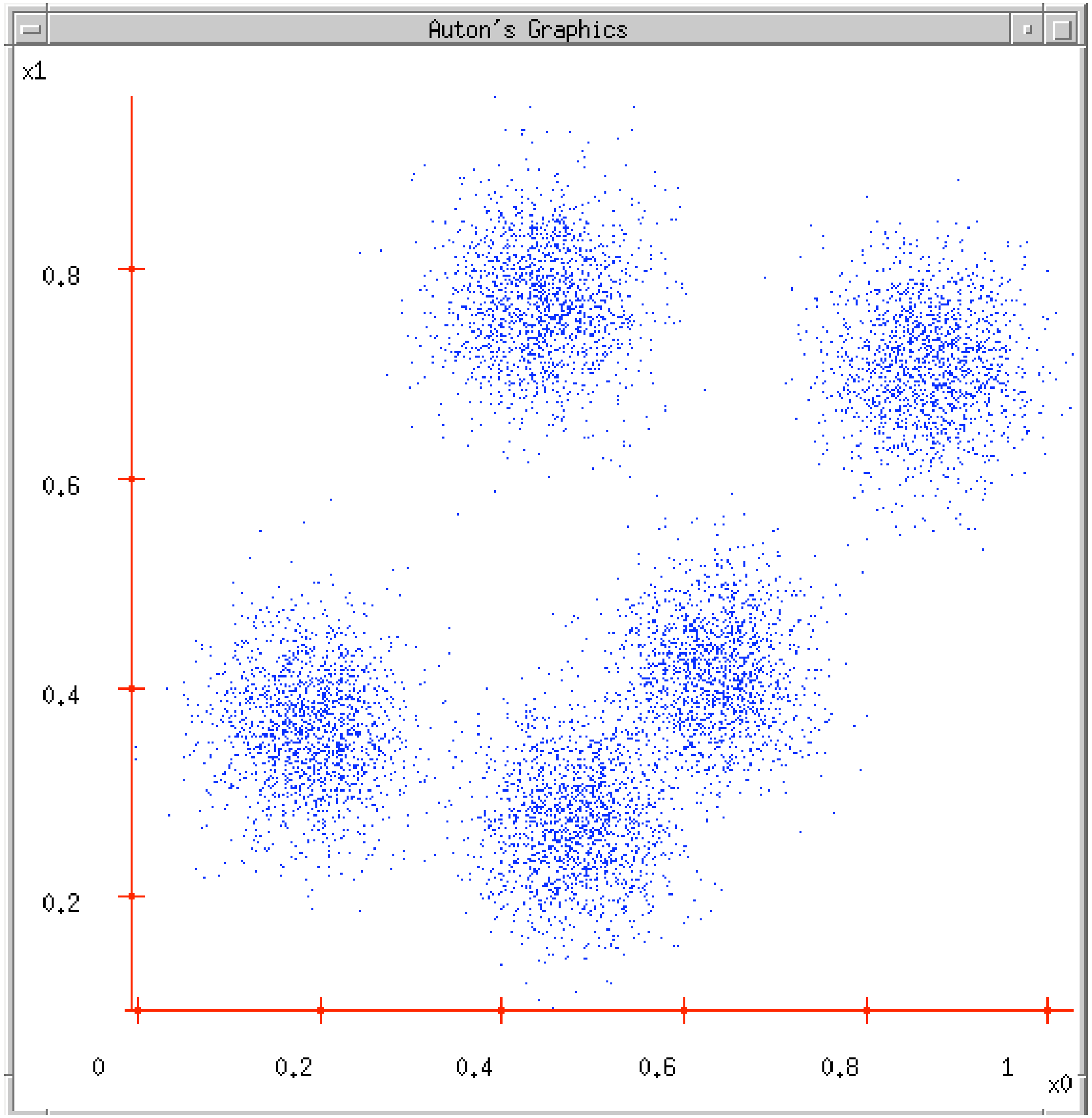
Unsupervised Learning

case of when you dont have a target to reach a ground truth, unlike regression or classifier. no ground truth in unsupervised, purpose is other things than predictions, one pupose is clustering. how to ground together based on similarirty

- 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**
  – 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**.

classification vs clustering, classified already in past to learn. clustering just have random points in spce, no name no categories, need to categoried outselves.

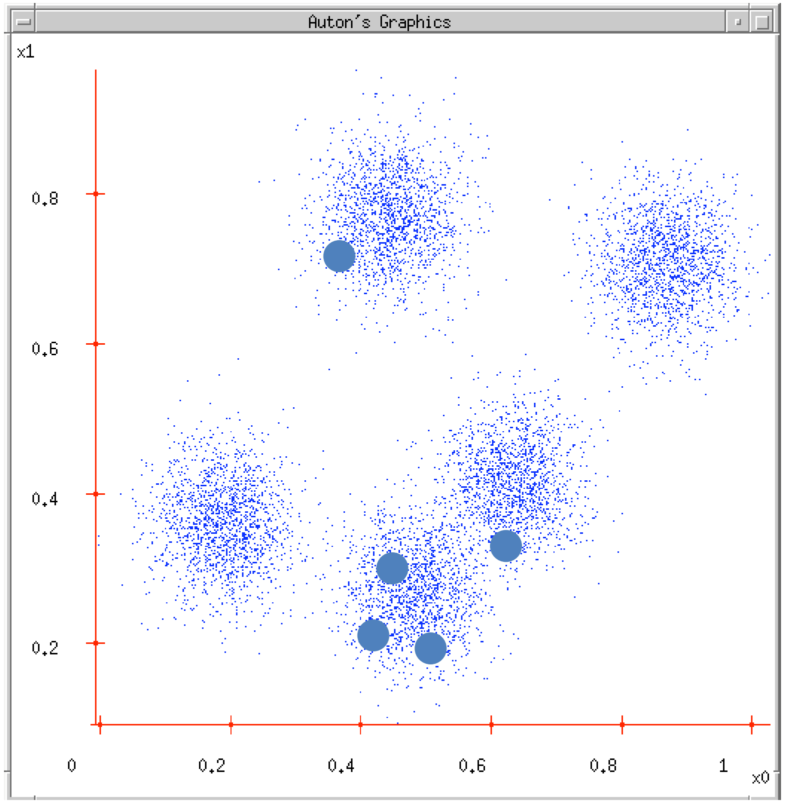
Clustering Data



K-Means Clustering

K-Means (k , data)

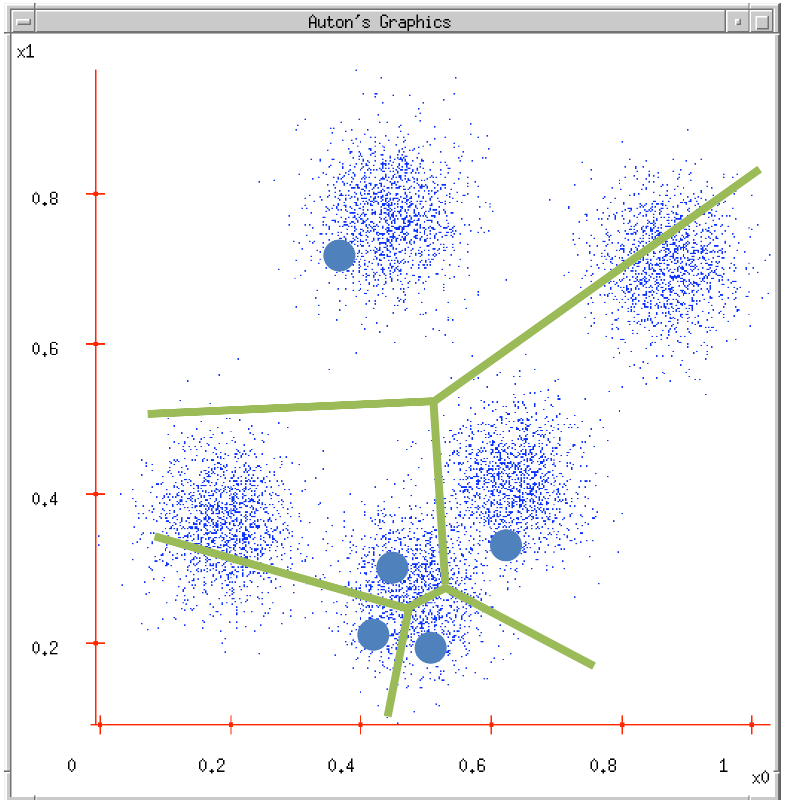
- Randomly choose k cluster center locations (centroids).
- Loop until convergence
  - Assign each point to the cluster of the closest centroid.
  - Reestimate the cluster centroids based on the data assigned to each.



K-Means Clustering

K-Means (k , data)

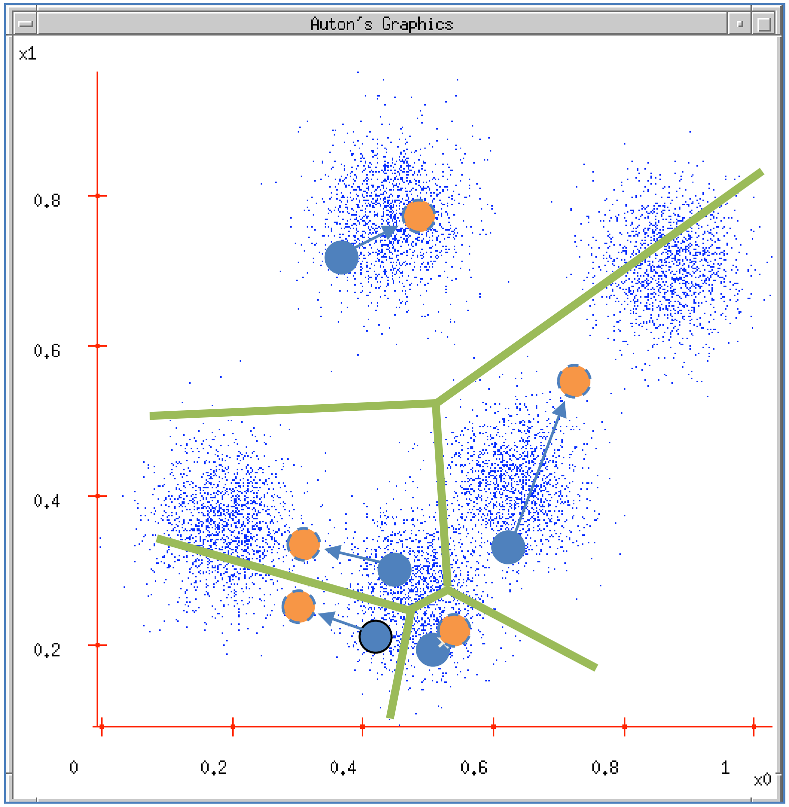
- Randomly choose k cluster center locations (centroids).
- Loop until convergence
  - Assign each point to the cluster of the closest centroid.
  - Reestimate the cluster centroids based on the data assigned to each.



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:

$$C(i) = \arg\min_{1 \le k \le K} \left\| x_i - m_k \right\|^2, \; i = 1, \ldots, N$$

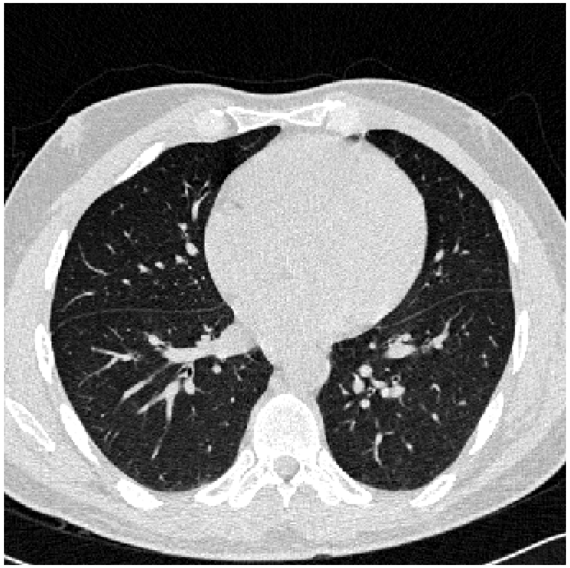
- For a given assignment *C*, compute the cluster means m_k:

$$m_k = \frac{\sum\limits_{i:C(i)=k} x_i}{N_k}, \; k = 1, \ldots, K.$$

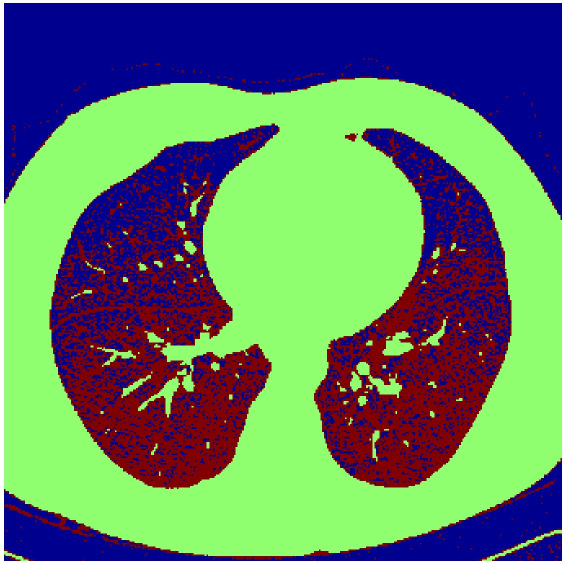
- Iterate above two steps until convergence

Text

Image Segmentation Results



An image (*I*)

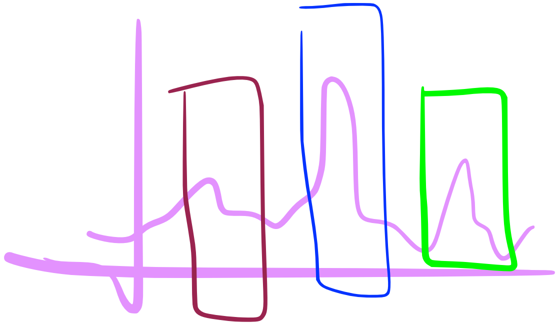


Three-cluster image (*J*) on gray values of *I*

Matlab code:

```
I = double(imread( '…'));

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



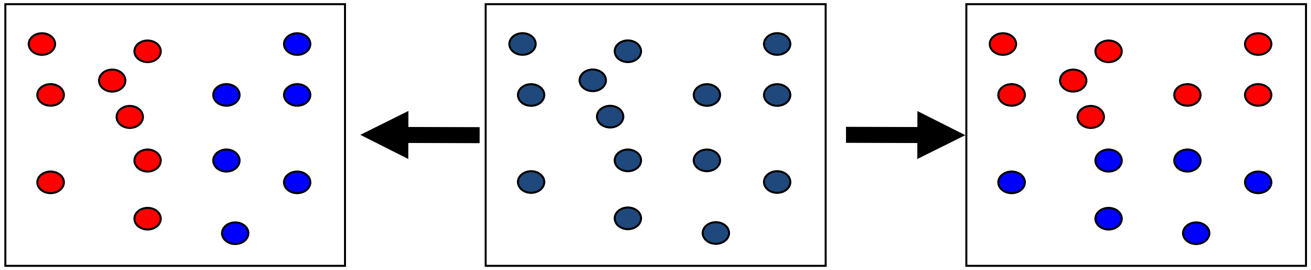
Problems with K-Means

- ***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)

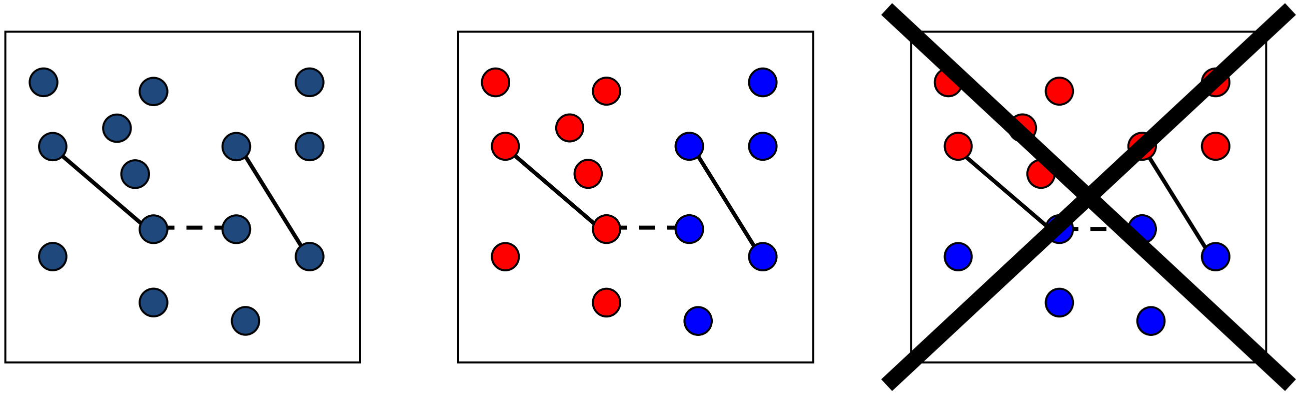
- Must manually choose k.
  - Learn the optimal k for the clustering (meta-learning). (Note that this requires a performance measure.)

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)