Clustering

CS273P

Topics

- Clustering
- K-Means clustering
- Agglomerative Clustering
- Gaussian Mixtures and Expectation-Maximization (EM)

Supervised learning

label

$label_1$

$label_3$

$label_4$

$label_5$

model/
predictor

Supervised learning: given labeled examples

Unsupervised learning





Unsupervised learning: given data, i.e. examples, but no labels

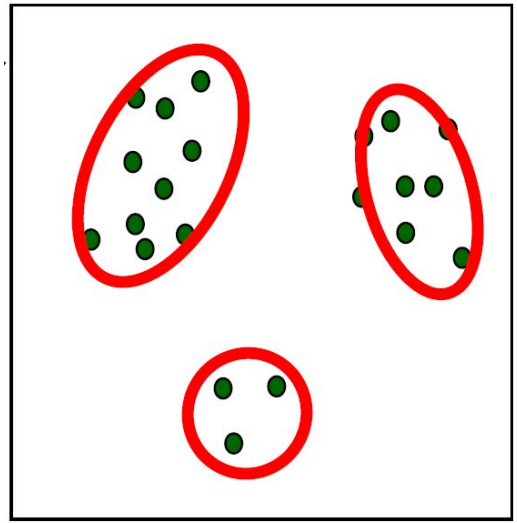
Unsupervised learning

Supervised learning

- Predict target value (“y”) given features (“x”)

Unsupervised learning

- Understand patterns of data (just “x”)
- Useful for many reasons
  - **Data mining (“explain”)**
  - **Missing data values (“impute”)**
  - **Representation (feature generation or selection)**



One example: clustering

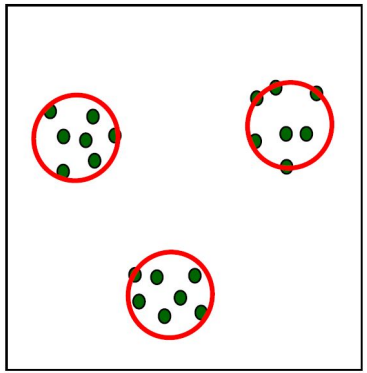
- Describe data by discrete “groups” with some characteristics

Clustering

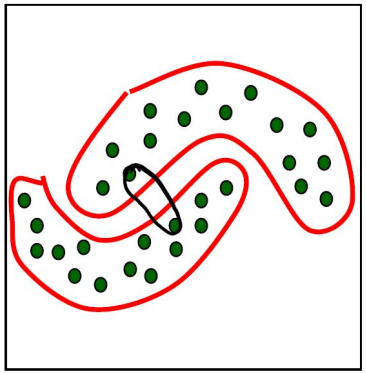
Clustering describes data by “groups”

The meaning of “groups” may vary by data!

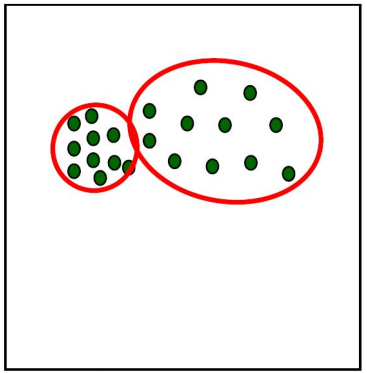
Examples



Location

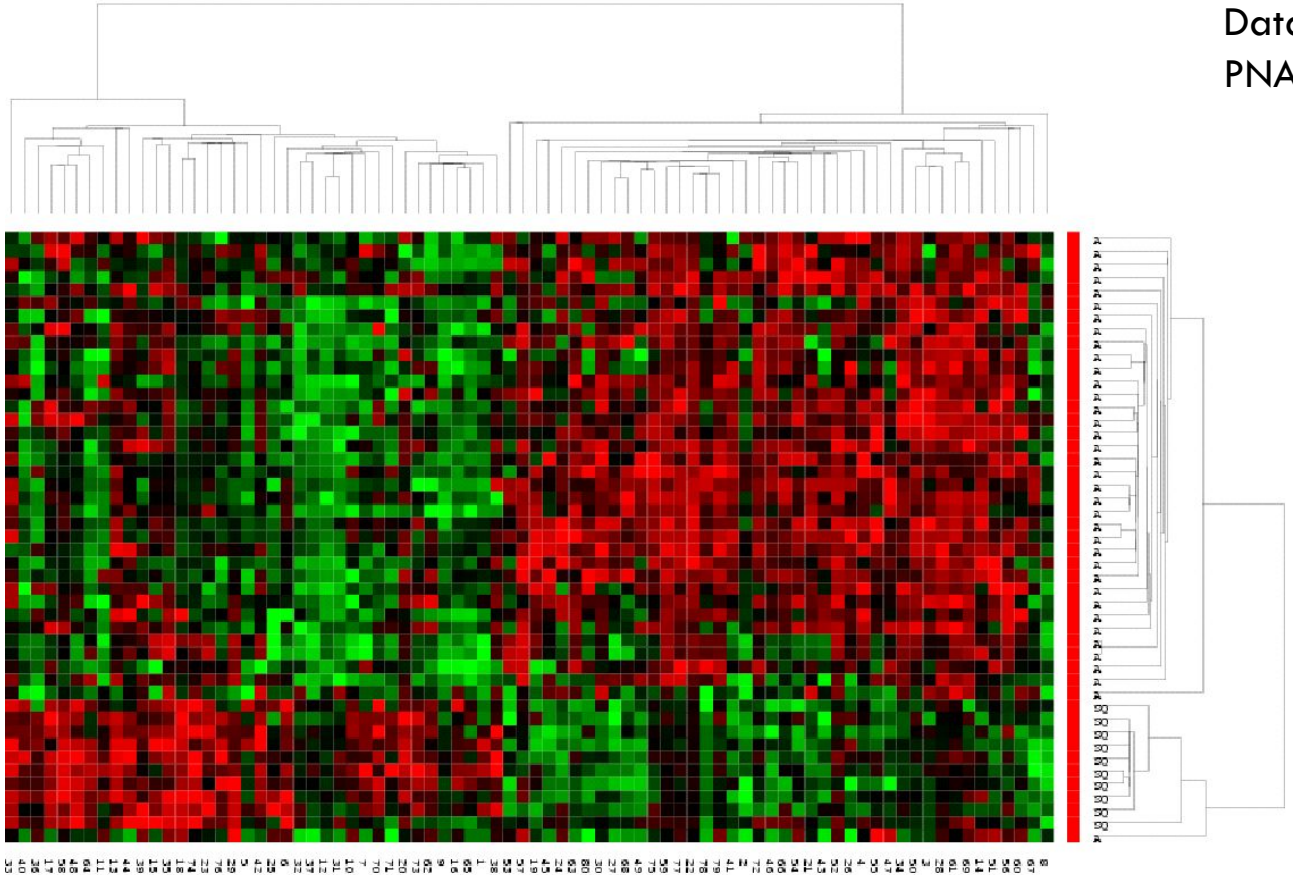


Shape



Density

Gene expression data



Data from Garber et al.
PNAS (98), 2001.

Face clustering

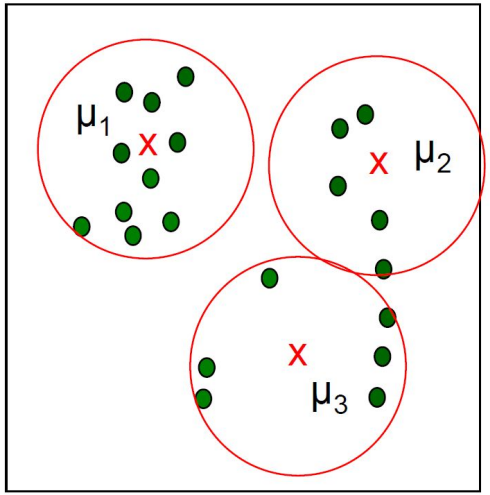
K-Means clustering

K-Means Clustering

A simple clustering algorithm

Iterate between

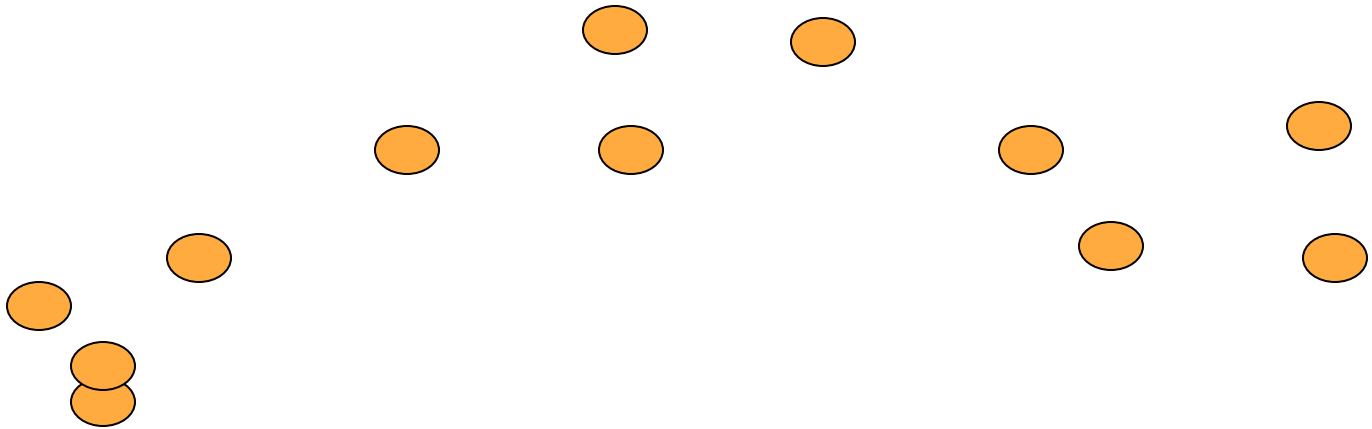
- Updating the assignment of data to clusters
- Updating the cluster's summarization



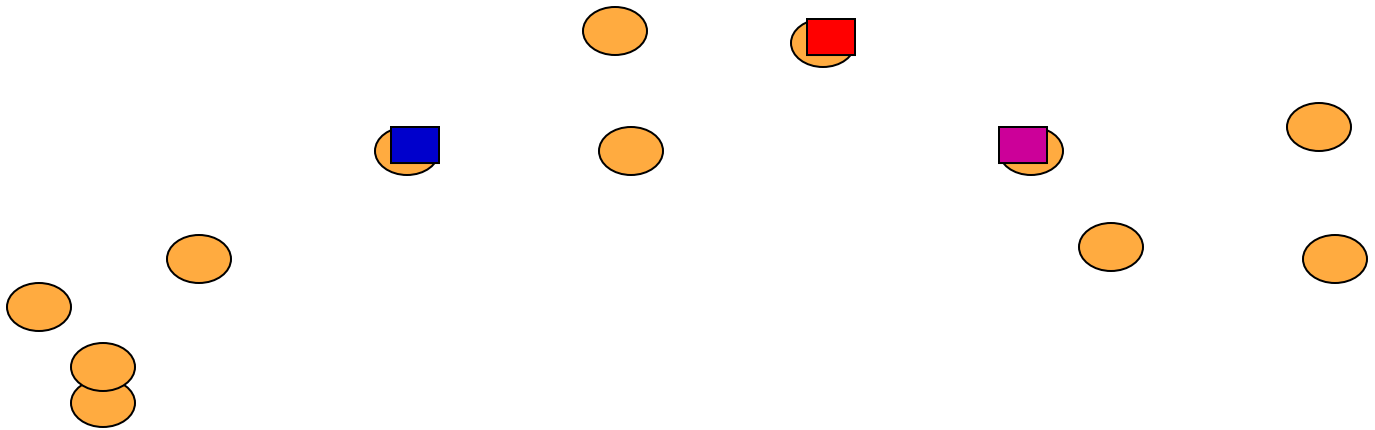
Notation:

1. Data example i has features x_i
2. Assume K clusters, (K=3)
3. Each cluster c "described" by a center μ_c
4. Each cluster will "claim" a set of nearby points

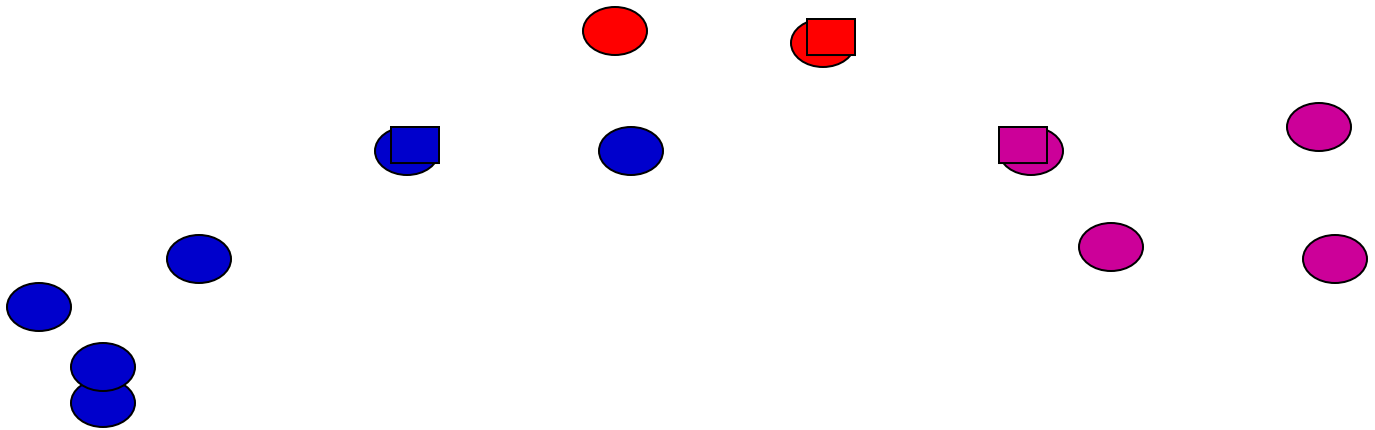
K-means: an example



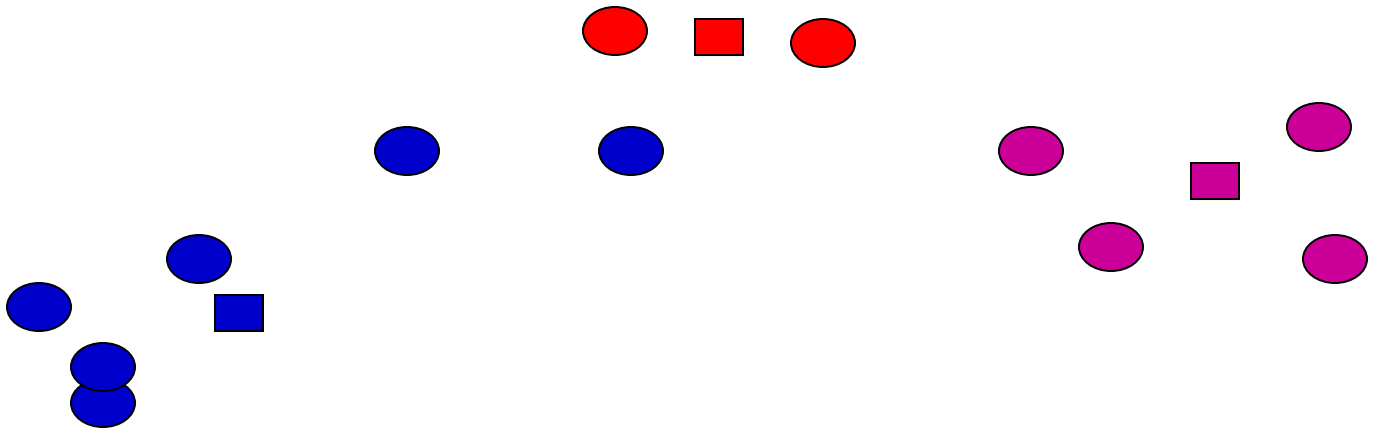
K-means: Initialize centers randomly



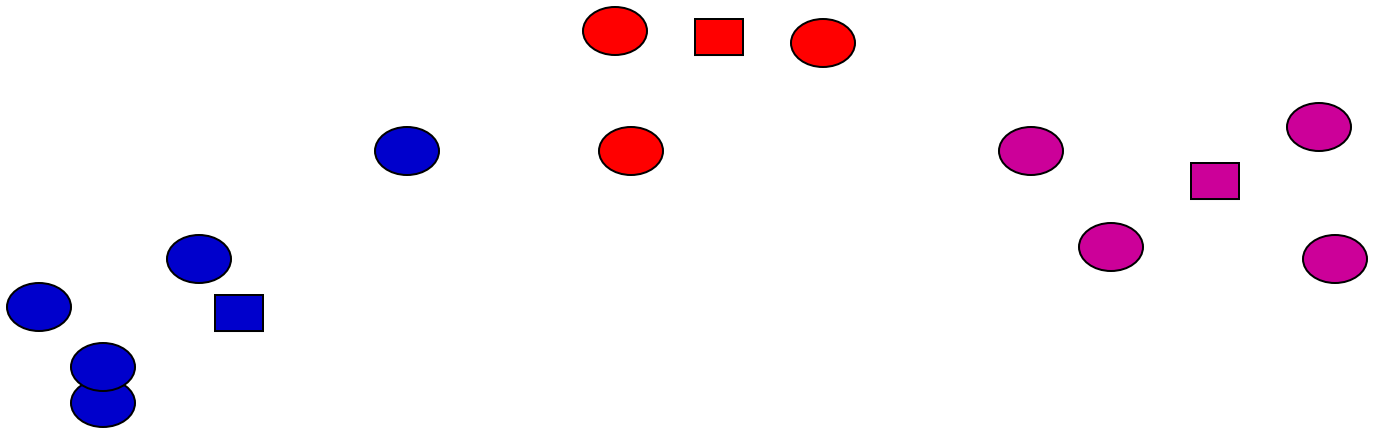
K-means: assign points to nearest center



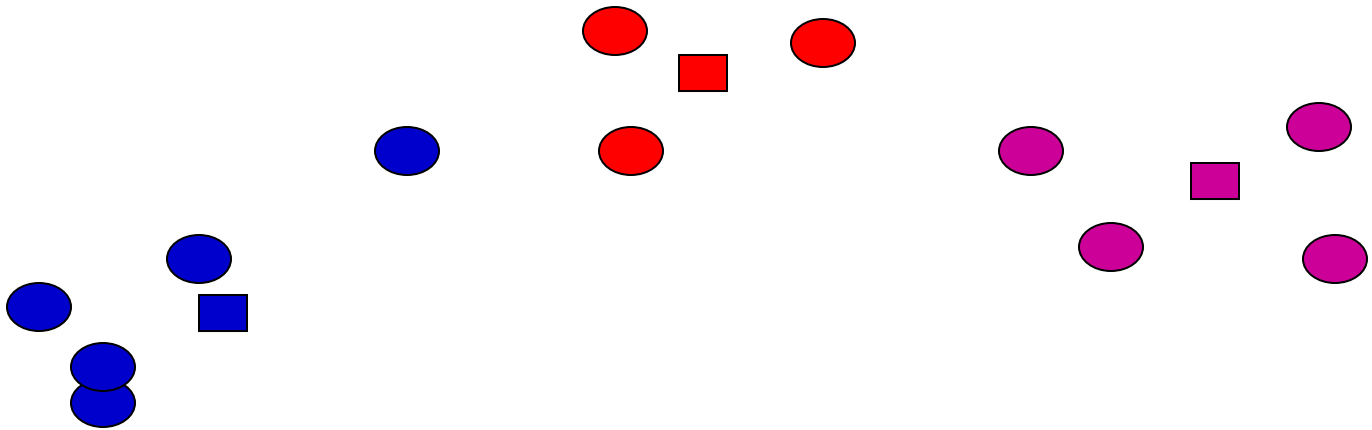
K-means: readjust centers



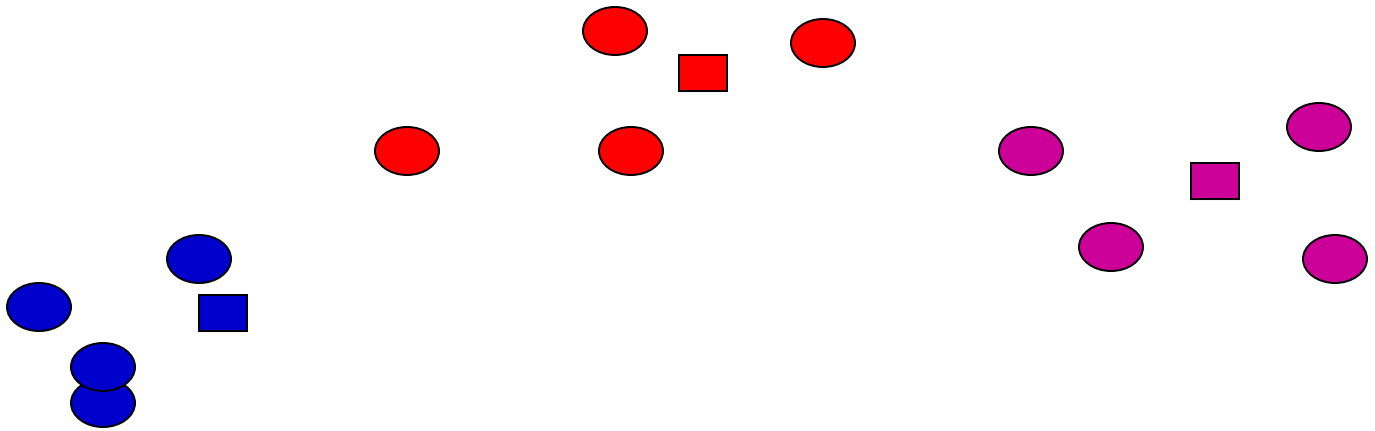
K-means: assign points to nearest center



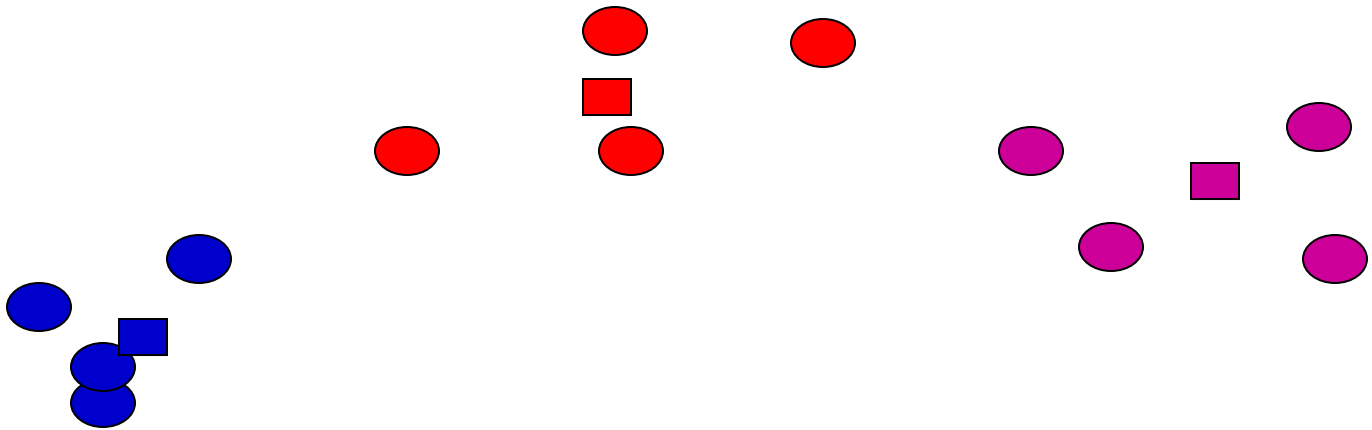
K-means: readjust centers



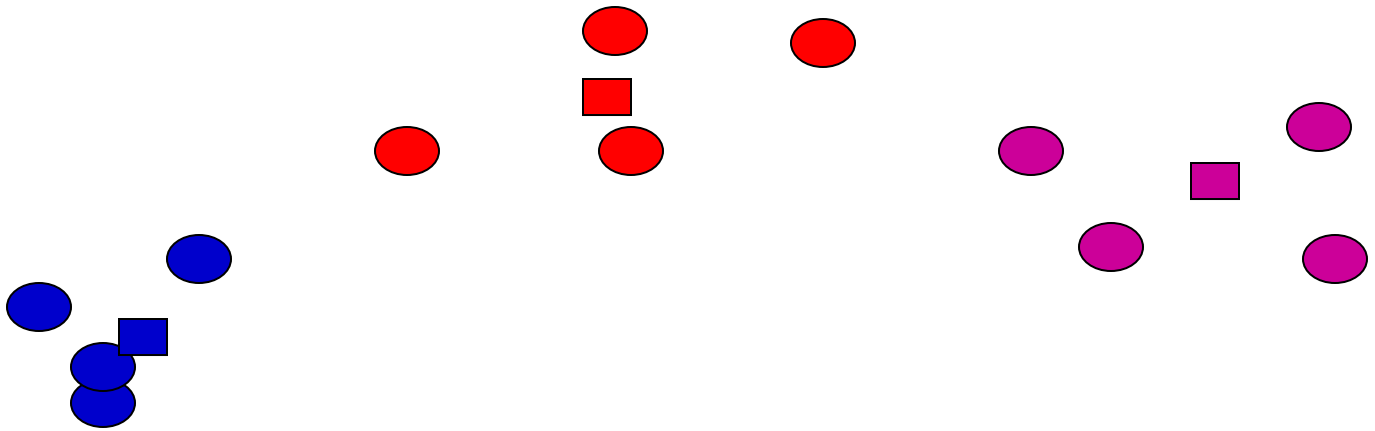
K-means: assign points to nearest center



K-means: readjust centers



K-means: assign points to nearest center



No changes: Done

K-Means Clustering

Iterate until convergence:

(A) For each datum, find the closest cluster: (z_i denotes cluster membership)

$$z_i = \arg\min_c \|x_i - \mu_c\|^2 \qquad \forall i$$

(B) Set each cluster to the mean of all assigned, for each cluster c:

$$\mu_c = \frac{1}{m_c} \sum_{i \in S_c} x_i \qquad S_c = \{i : z_i = c\},\ m_c = |S_c|$$

K-Means clustering

Optimizing the cost function:

$$C(\underline{z}, \underline{\mu}) = \sum_i \|x_i - \mu_{z_i}\|^2$$

Coordinate descent:

Over the cluster assignments (fixed μ)

- Only one term in sum depends on z_i
- Minimized by selecting closest μ_c

Over the cluster centers (fixed z)

- Cluster c only depends on x_i with z_i =c
- Minimized by selecting the mean

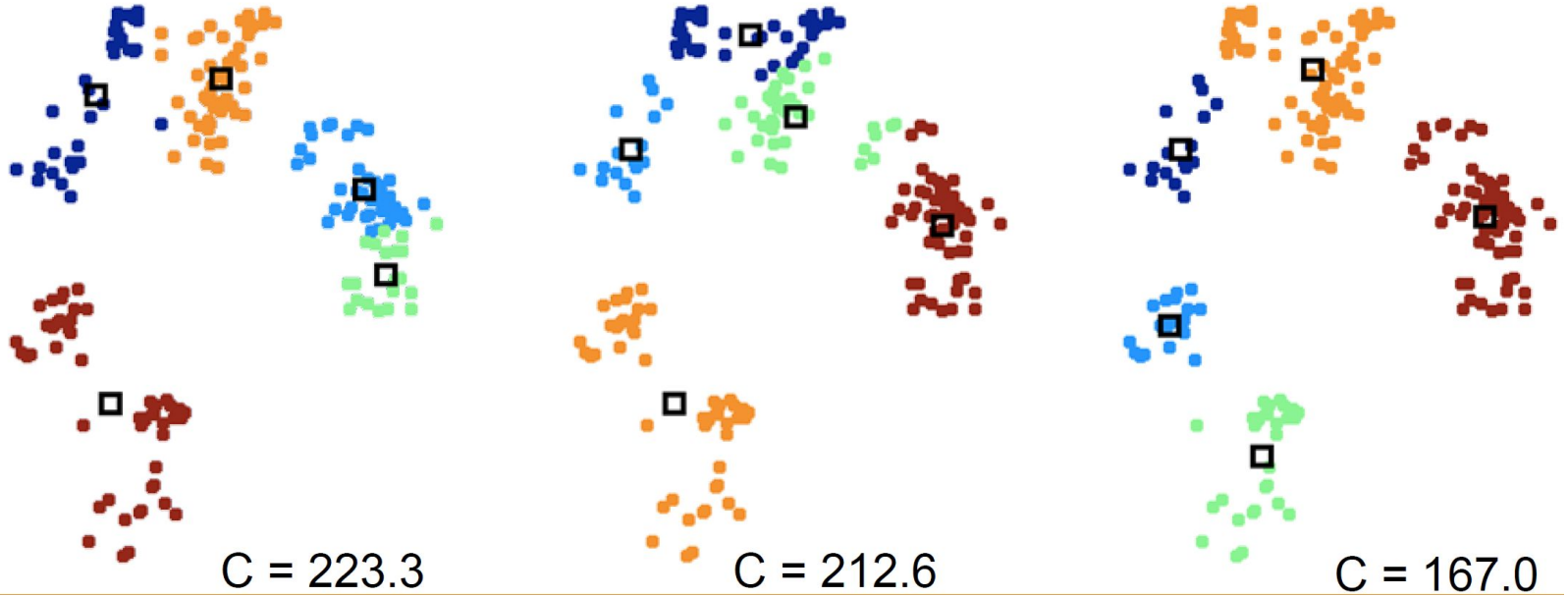
Guaranteed to converge after finite number of steps!

Initialization

Multiple local optima, depending on initialization

Try different (randomized) initializations

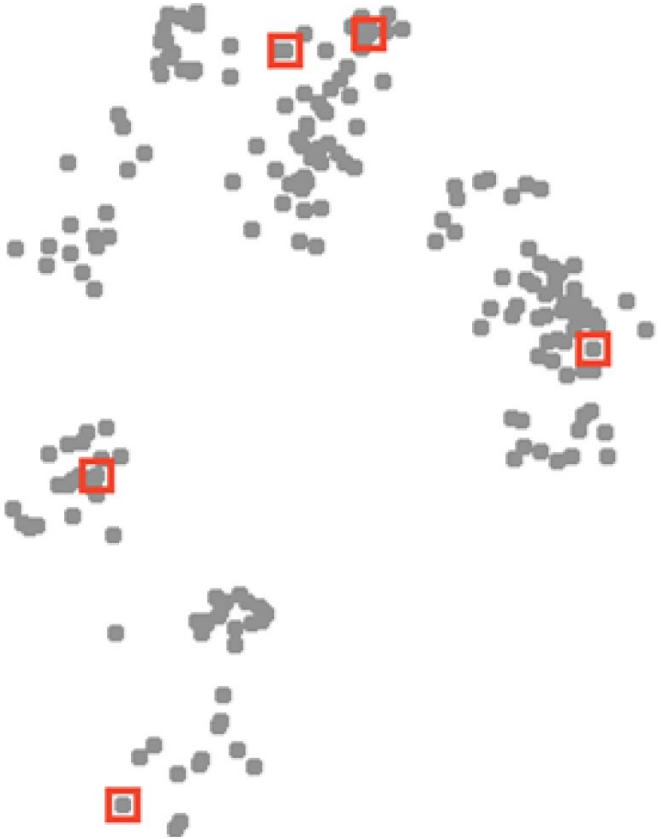
Can use cost C to decide which we prefer



Initialization methods

Random

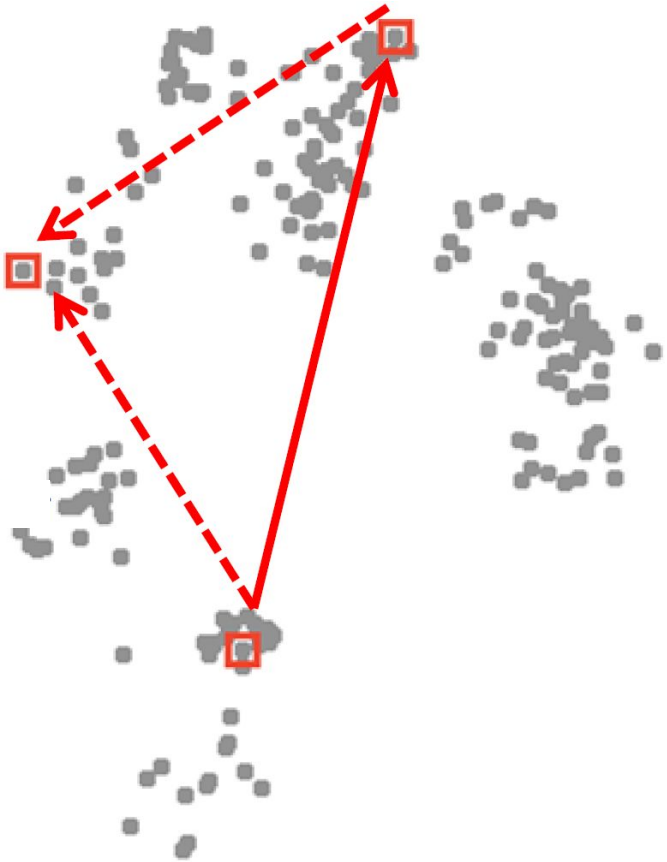
- Usually, choose random data index
- Ensures centers are near some data
- Issue: may choose nearby points



Initialization methods

Distance based

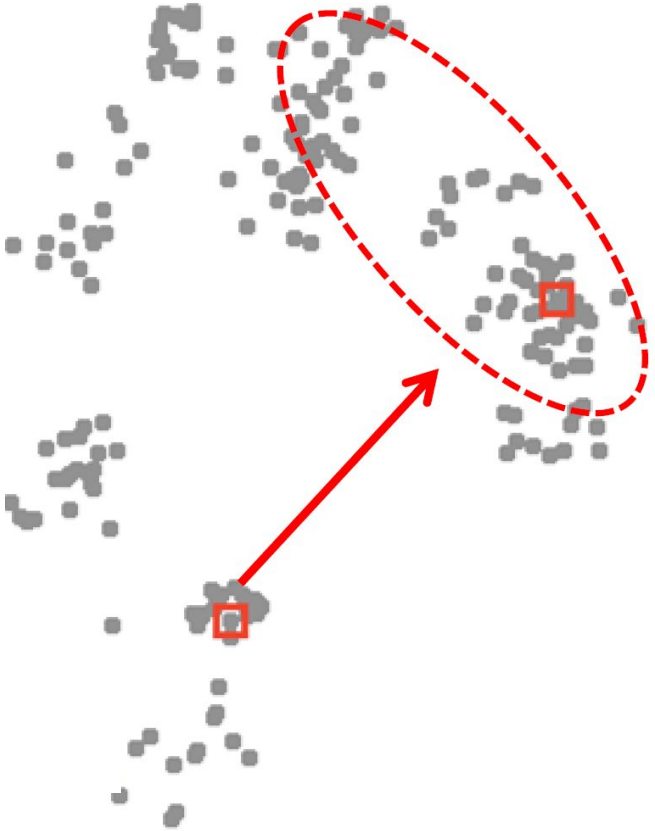
- Start with one random data point
- Find the point farthest from the clusters chosen so far
- Issue: may choose outliers



Initialization methods

Random + distance (“kmeans++”)

- Choose next points “far but randomly”
  - p(x) ~ squared distance from x to current centers
- Likely to put a cluster far away, in a region with lots of data



Choosing Number of Clusters

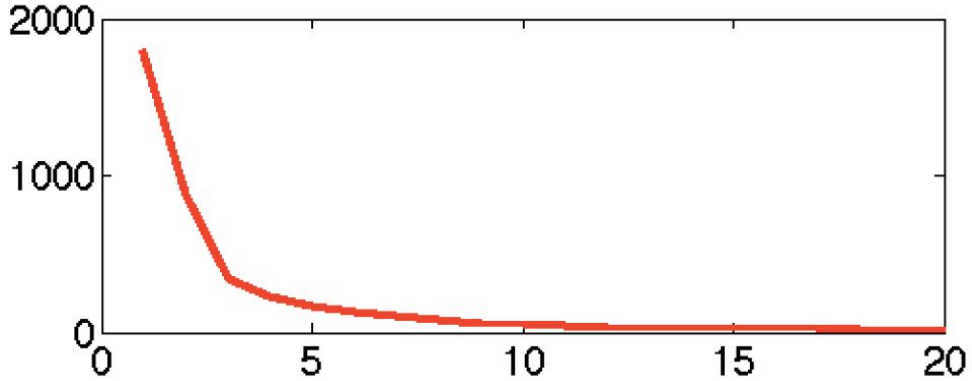
With cost function

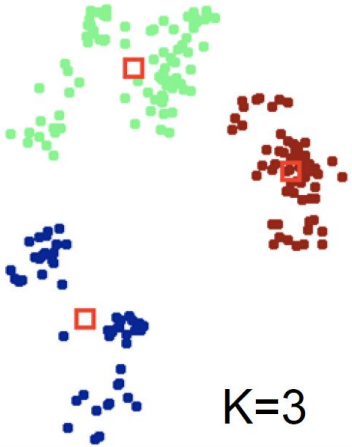
$$C(\underline{z}, \underline{\mu}) = \sum_i \|x_i - \mu_{z_i}\|^2$$

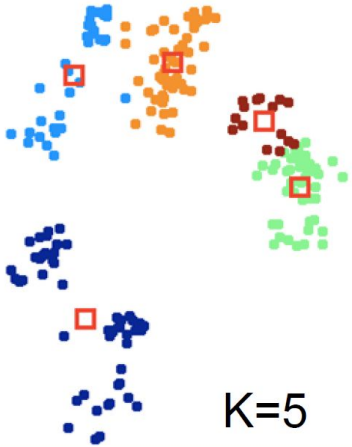
what is the optimal value of k?

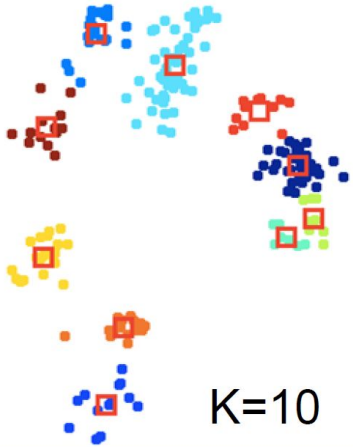
Cost always decreases with k!

A model complexity issue…







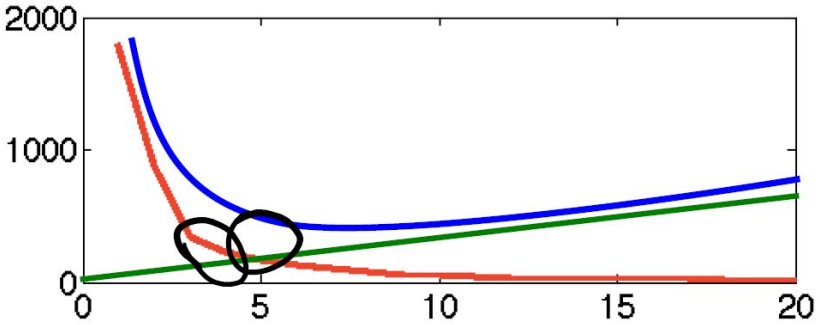


Choosing the number of clusters

One solution is to penalize for complexity

Add penalty: Total = Error + Complexity

Now more clusters can increase cost, if they don't help "enough"



Ex: simplified BIC penalty

$$J(\underline{z}, \underline{\mu}) = \log\left[\frac{1}{m\,d}\sum_{i}\|x_i - \mu_{z_i}\|^2\right] + k\frac{\log m}{m}$$

More precise version: see e.g. "X means" (Pelleg & Moore, 2000)

Summary

K-Means clustering

- Clusters described as locations (“centers”) in feature space

Procedure

- Initialize cluster centers
- Iterate:
  - assign each data point to its closest cluster center
  - move cluster centers to minimize mean squared error

Properties

- Coordinate descent on MSE criterion
- Prone to local optima; initialization important

Choosing the # of clusters , K

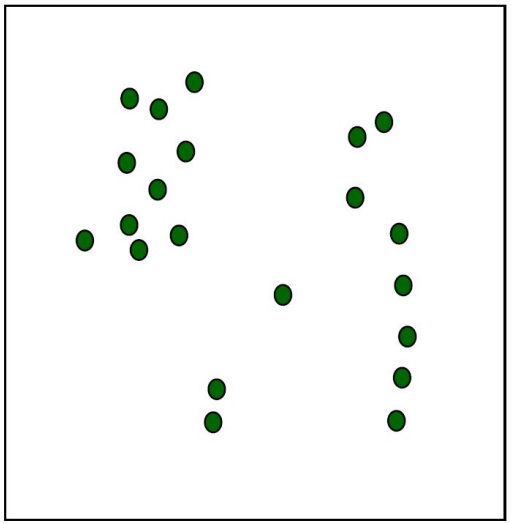
- Model selection problem; penalize for complexity (BIC, etc.)

Agglomerative Clustering

Hierarchical Agglomerative Clustering

Initially, every datum is a cluster

Data:

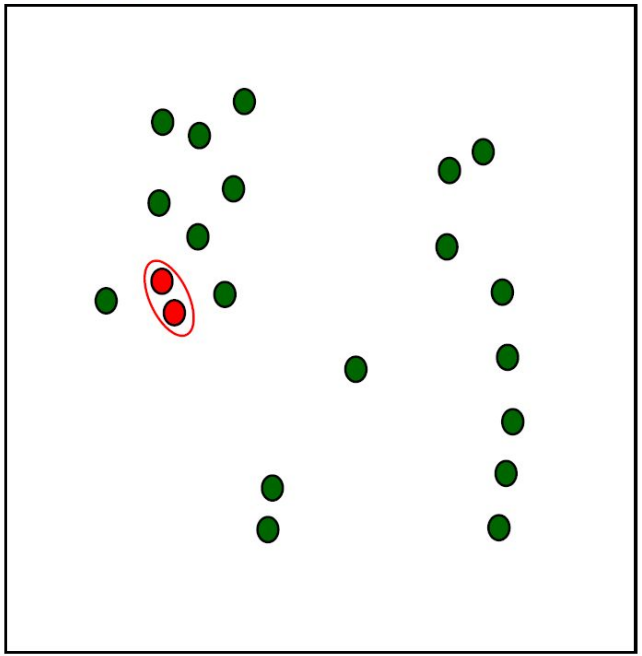


- A simple clustering algorithm
- Define a distance (or dissimilarity) between clusters (we'll return to this)
- Initialize: every example is a cluster
- Iterate:
  - Compute distances between all clusters (store for efficiency)
  - Merge two closest clusters
- Save both clustering and sequence of cluster operations
- Dendrogram

Iteration 1

Builds up a sequence of clusters (“hierarchical”)

Data:



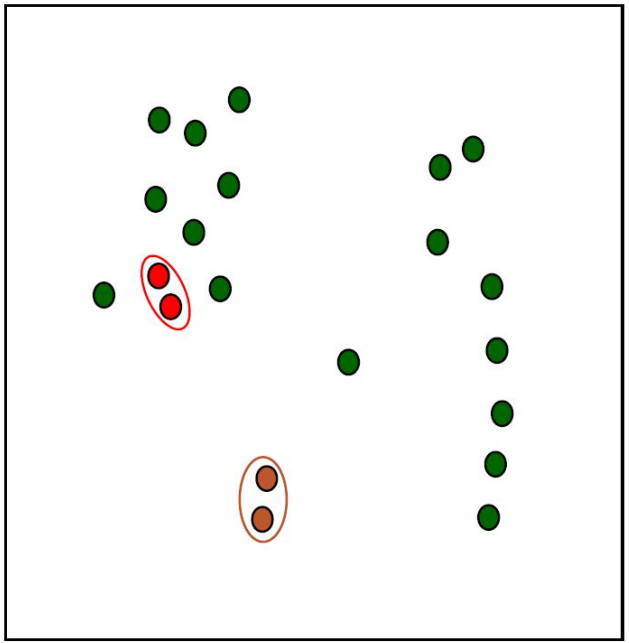
Dendrogram:

Height of the join
indicates dissimilarity

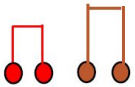
Iteration 2

Builds up a sequence of clusters (“hierarchical”)

Data:



Dendrogram:

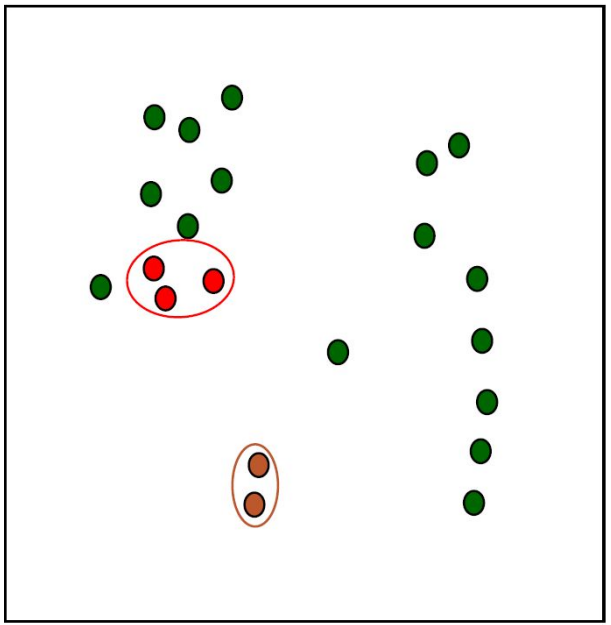


Height of the join
indicates dissimilarity

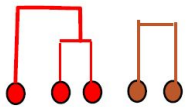
Iteration 3

Builds up a sequence of clusters (“hierarchical”)

Data:



Dendrogram:

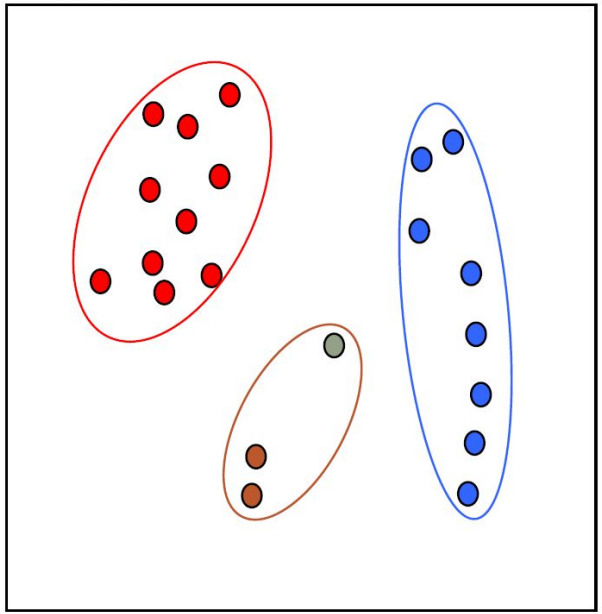


Height of the join
indicates dissimilarity

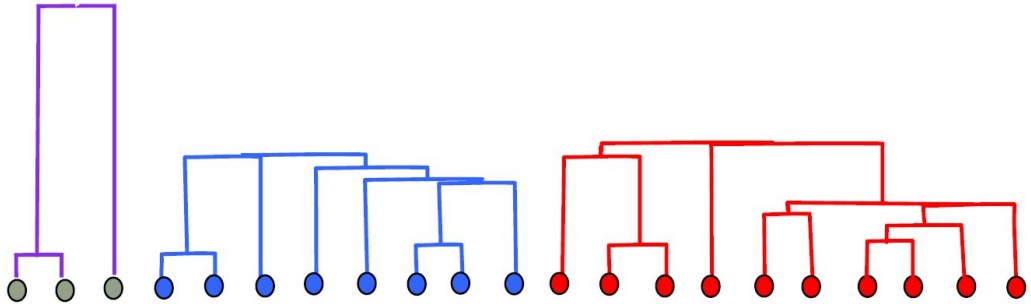
Iteration m-3

Builds up a sequence of clusters (“hierarchical”)

Data:



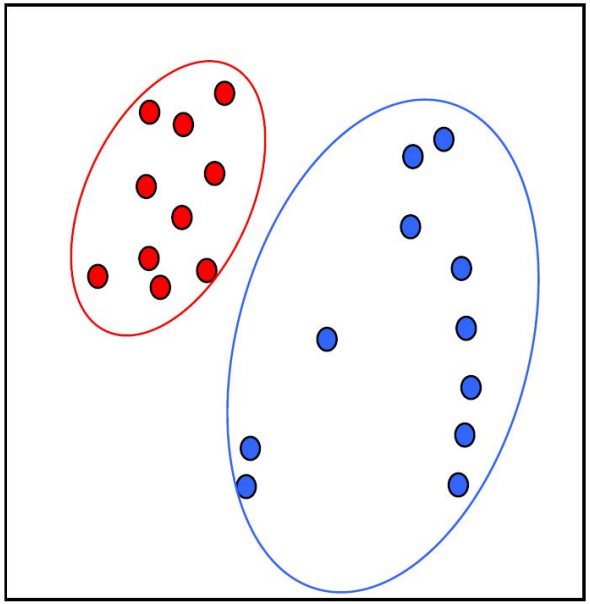
Dendrogram:



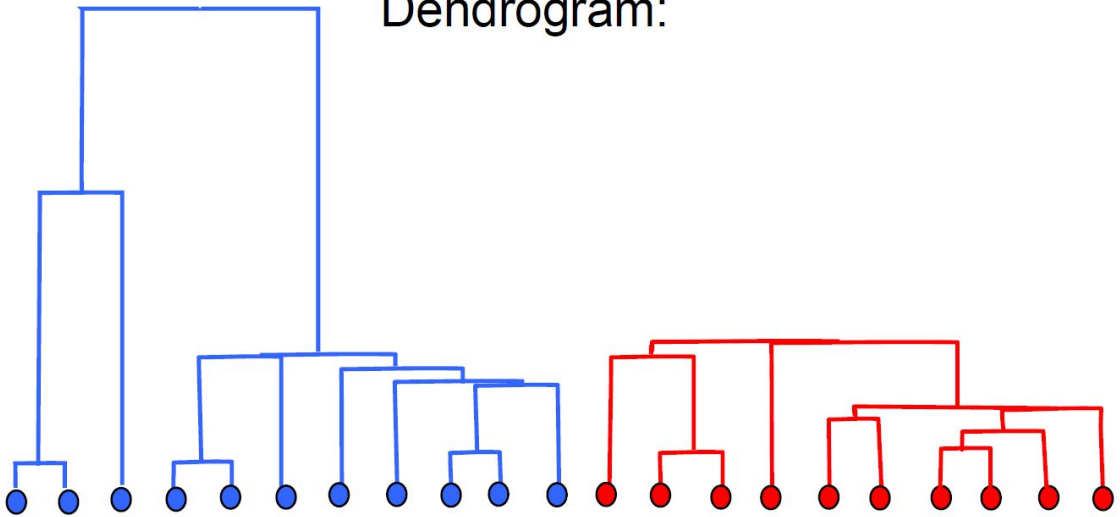
Iteration m-2

Builds up a sequence of clusters (“hierarchical”)

Data:



Dendrogram:

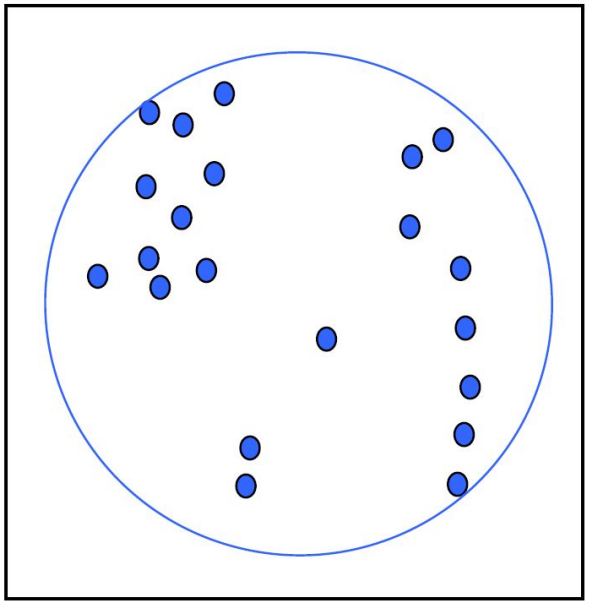


In mltools: “agglomerative”

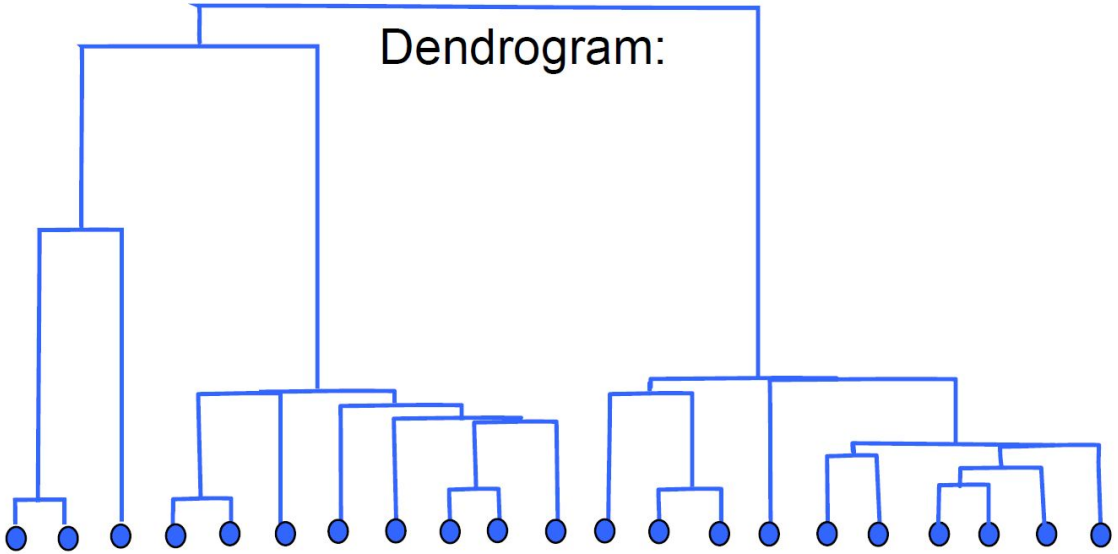
Iteration m-1

Builds up a sequence of clusters (“hierarchical”)

Data:

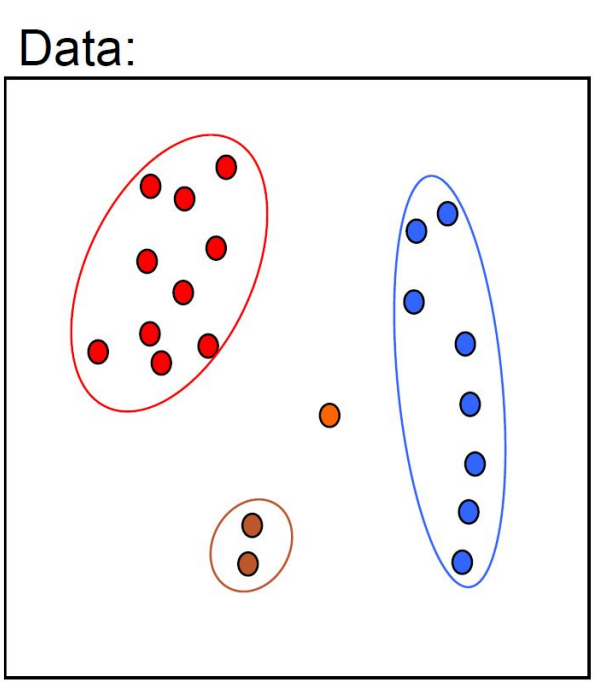


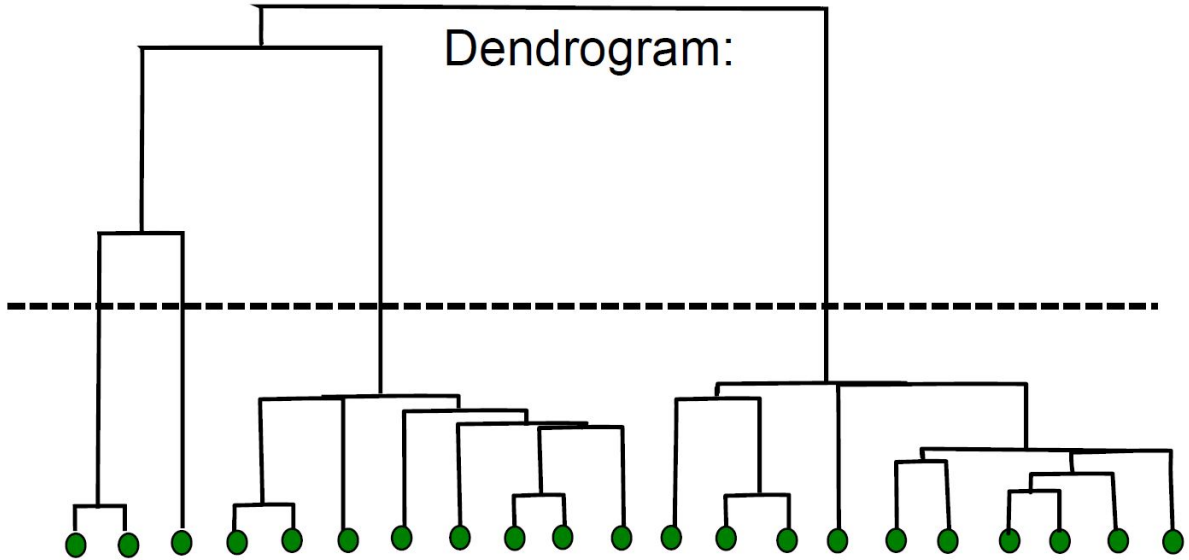
Dendrogram:



From dentrogram to clusters

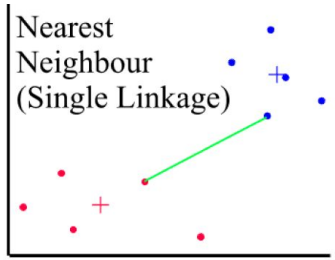
Given the sequence, can select a number of clusters or a dissimilarity threshold:





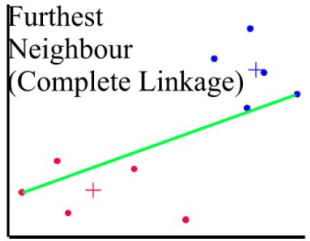
Cluster distances

$$D_{\min}(C_i, C_j) = \min_{x \in C_i,\ y \in C_j} \|x - y\|^2$$



produces minimal spanning tree.

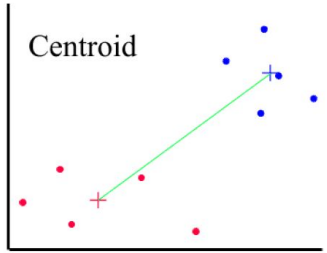
$$D_{\max}(C_i, C_j) = \max_{x \in C_i,\ y \in C_j} \|x - y\|^2$$



avoids elongated clusters.

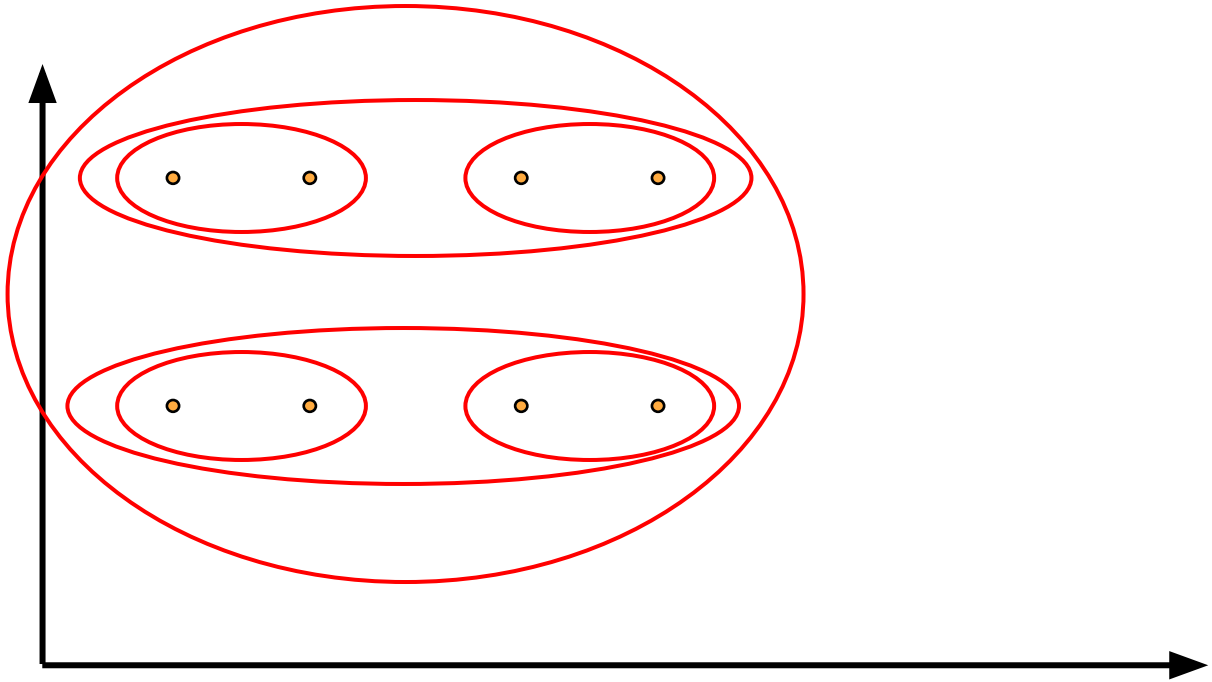
$$D_{\text{avg}}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i,\ y \in C_j} \|x - y\|^2$$

$$D_{\text{means}}(C_i, C_j) = \|\mu_i - \mu_j\|^2$$

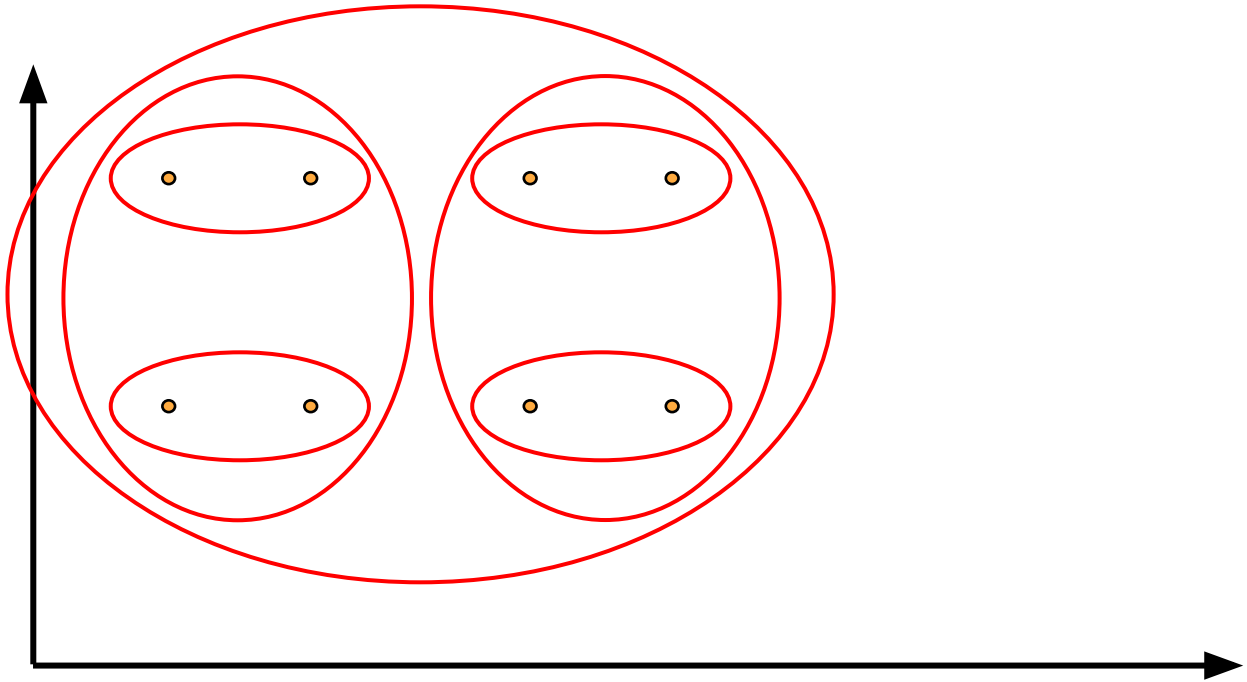


Constant time: D(A,C) → D(A+B,C)
D(B,C) →

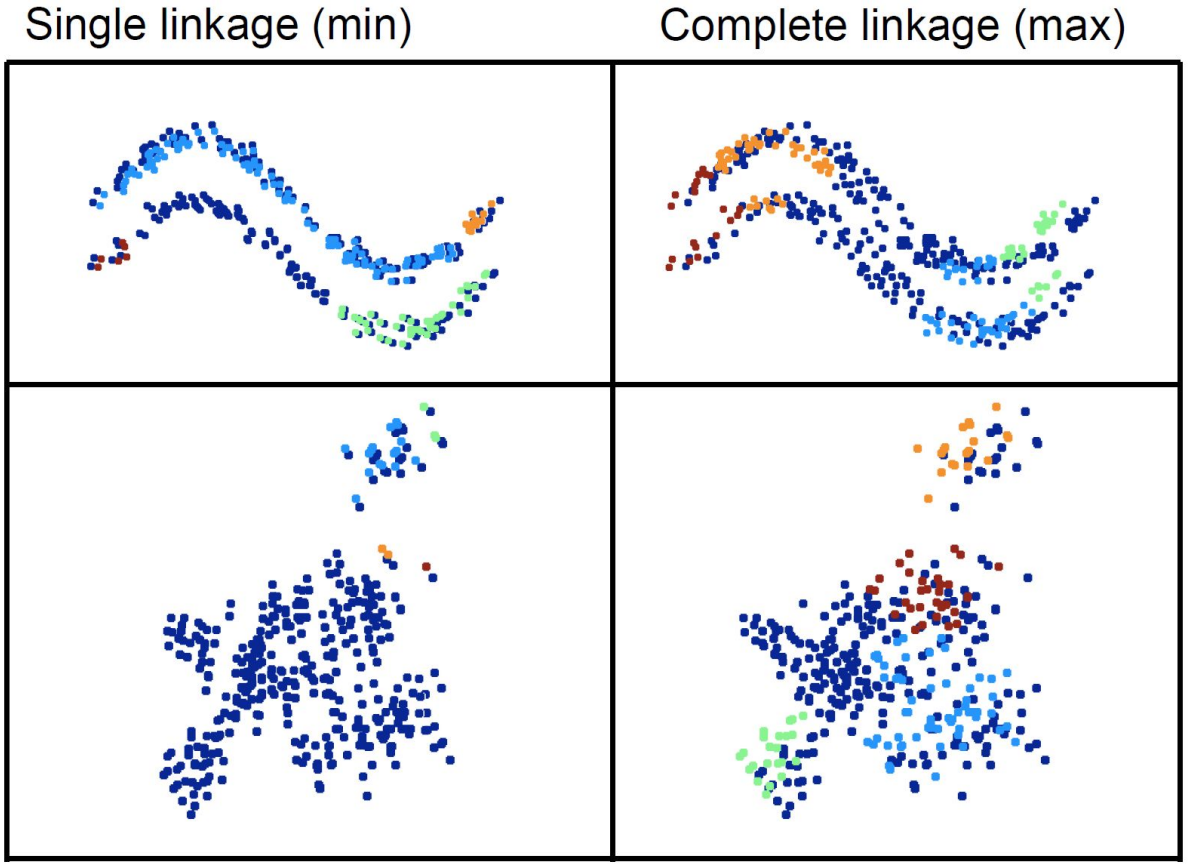
Single Link (min distance) Example



Complete Link (max distance) Example



Cluster distances - difference choices will affect clusters created



Example: gene expression clustering

Measure gene expression

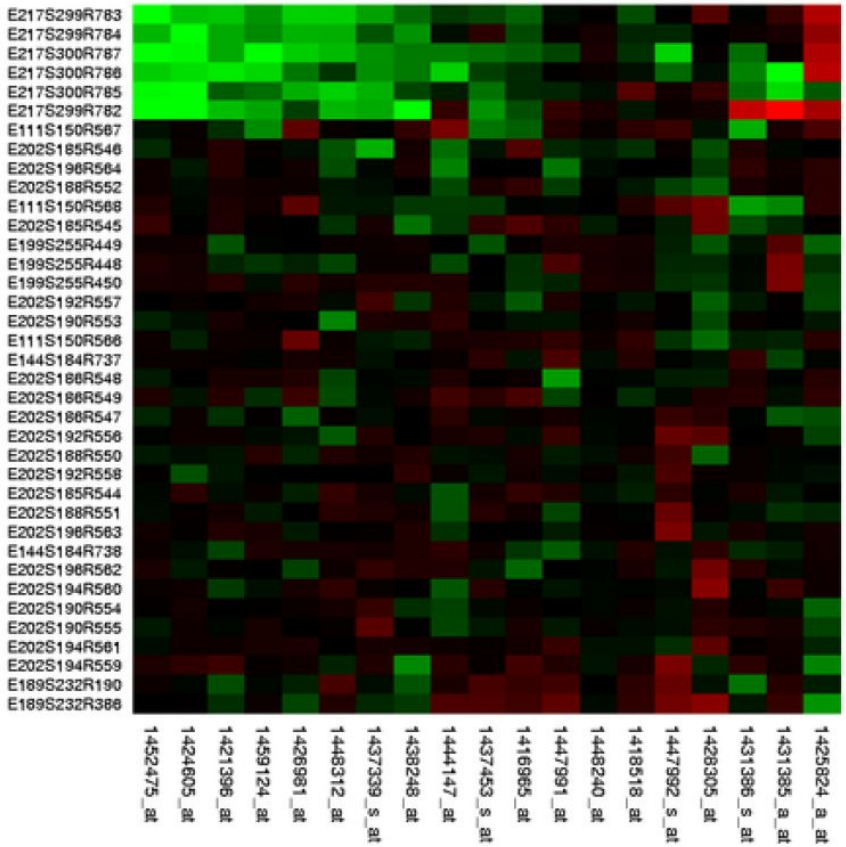
Various experimental conditions

- Disease vs. normal
- Time
- Subjects

Explore similarities

- What genes change together?
- What conditions are similar?

Cluster on both genes and conditions



Summary

Agglomerative clustering

- Choose a cluster distance / dissimilarity scoring method
- Successively merge closest pair of clusters
- “Dendrogram” shows sequence of merges & distances

Agglomerative clusters depend critically on dissimilarity

- Choice determines characteristics of “found” clusters

“Clustergram ” for understanding data matrix

- Build clusters on rows (data) and columns (features)
- Reorder data & features to expose behavior across groups

Gaussian Mixtures and EM

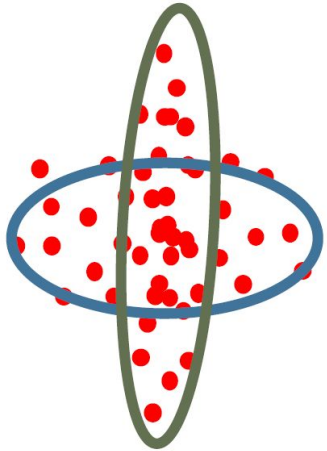
Mixtures of Gaussians

K-means algorithm

- Assigned each example to exactly one cluster
- What if clusters are overlapping?
  - Hard to tell which cluster is right
  - Maybe we should try to remain uncertain
- Used Euclidean distance
- What if cluster has a non circular shape?

Gaussian mixture models

- Clusters modeled as Gaussians
  - Not just by their mean
- EM algorithm: assign data to cluster with some probability
- Gives probability model of x! (“generative”)



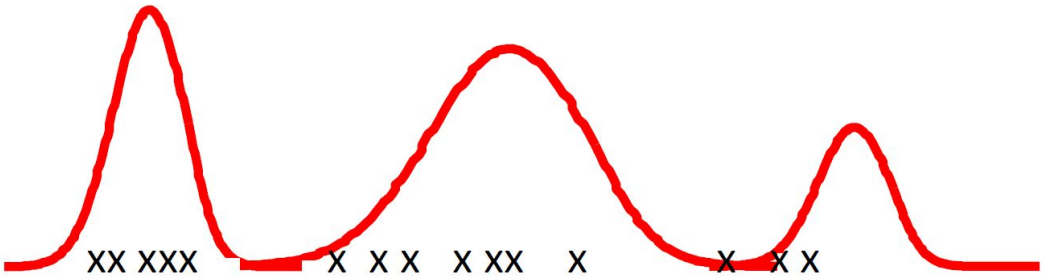
Mixtures of Gaussians

Start with parameters describing each cluster

Mean μ_c , variance Σ_c , “size” π_c

Probability distribution:

$$p(x) = \sum_c \pi_c \, \mathcal{N}(x \; ; \; \mu_c, \sigma_c)$$



Mixtures of Gaussians

Start with parameters describing each cluster

Mean μ_c , variance Σ_c , "size" π_c

Probability distribution:

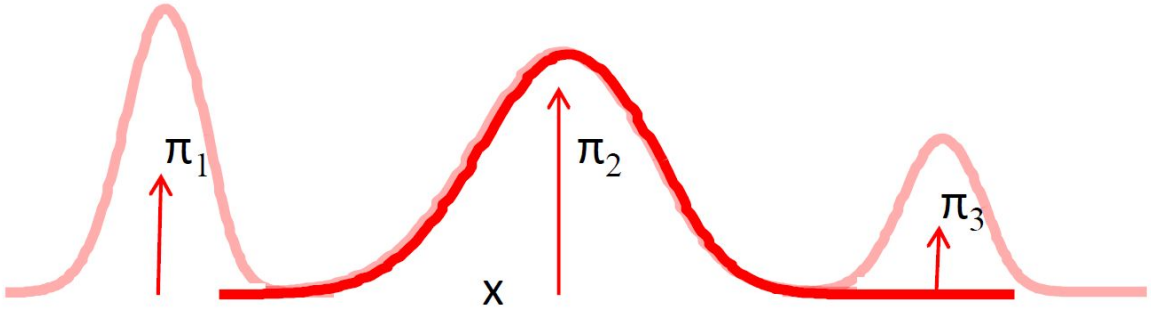
$$p(x) = \sum_c \pi_c \, \mathcal{N}(x \; ; \; \mu_c, \sigma_c)$$

Equivalent "latent variable" form:

- Select a mixture component with probability π_c $\quad p(z = c) = \pi_c$
- Sample from that component's Gaussian $\quad p(x|z = c) = \mathcal{N}(x \; ; \; \mu_c, \sigma_c)$

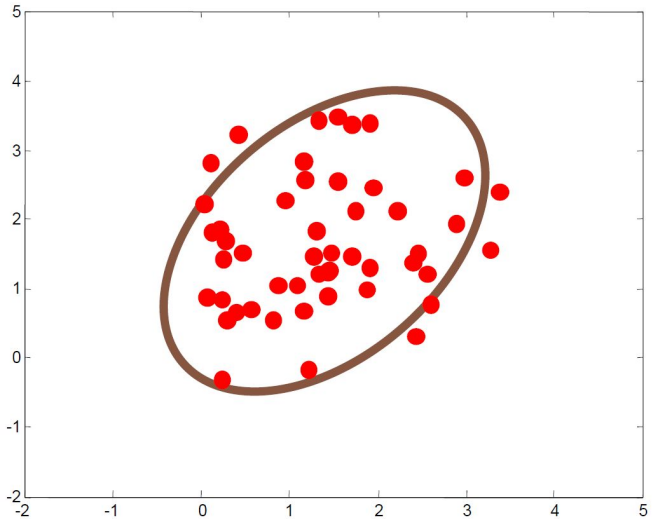
"Latent assignment" z:
we observe x, but z is hidden

p(x) = marginal over x



Multivariate Gaussian Models

$$\mathcal{N}(\underline{x}\ ;\ \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$



Maximum Likelihood estimates

$$\hat{\mu} = \frac{1}{m} \sum_i x^{(i)}$$

$$\hat{\Sigma} = \frac{1}{m} \sum_i (x^{(i)} - \hat{\mu})^T (x^{(i)} - \hat{\mu})$$

We'll model each cluster using one of these Gaussian "bells"…

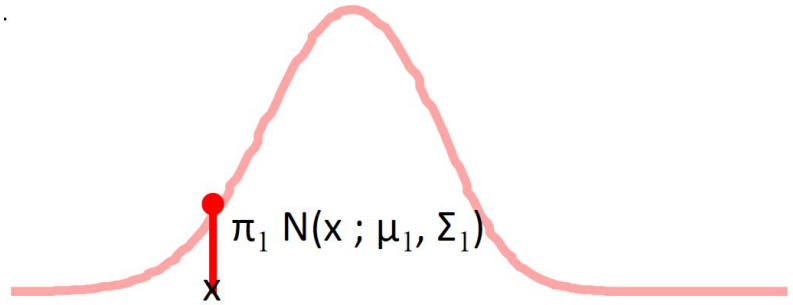
EM Algorithm: E-step

Start with clusters: Mean μ_c , Covariance Σ_c , size π_c

E-step (“Expectation”)

- For each datum (example) x_i ,
- Compute r_{ic}, the probability that it belongs to cluster c
  - Compute its probability under model c
  - Normalize to sum to one (over clusters c)

$$r_{ic} = \frac{\pi_c \mathcal{N}(x_i \; ; \; \mu_c, \Sigma_c)}{\sum_{c'} \pi_{c'} \mathcal{N}(x_i \; ; \; \mu_{c'}, \Sigma_{c'})}$$



EM Algorithm: E-step

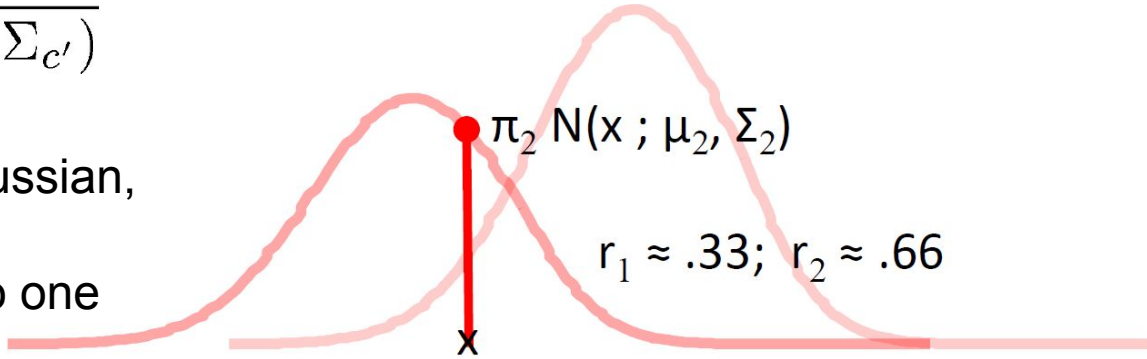
Start with clusters: Mean μ_c , Covariance Σ_c , size π_c

E-step ("Expectation")

- For each datum (example) x_i ,
- Compute r_{ic}, the probability that it belongs to cluster c
  - Compute its probability under model c
  - Normalize to sum to one (over clusters c)

$$r_{ic} = \frac{\pi_c \mathcal{N}(x_i \; ; \; \mu_c, \Sigma_c)}{\sum_{c'} \pi_{c'} \mathcal{N}(x_i \; ; \; \mu_{c'}, \Sigma_{c'})}$$

If x_i is very likely under the c-th Gaussian, it gets high weight
Denominator just makes r' s sum to one



EM Algorithm: M-Step

Start with assignment probabilities r_{ic}

Update parameters: mean μ_c , Covariance Σ_c , size π_c

M-step (Maximization)

- For each cluster (Gaussian) z = c,
- Update its parameters using the (weighted) data points

$$m_c = \sum_i r_{ic}$$

Total responsibility allocated to cluster c

$$\pi_c = \frac{m_c}{m}$$

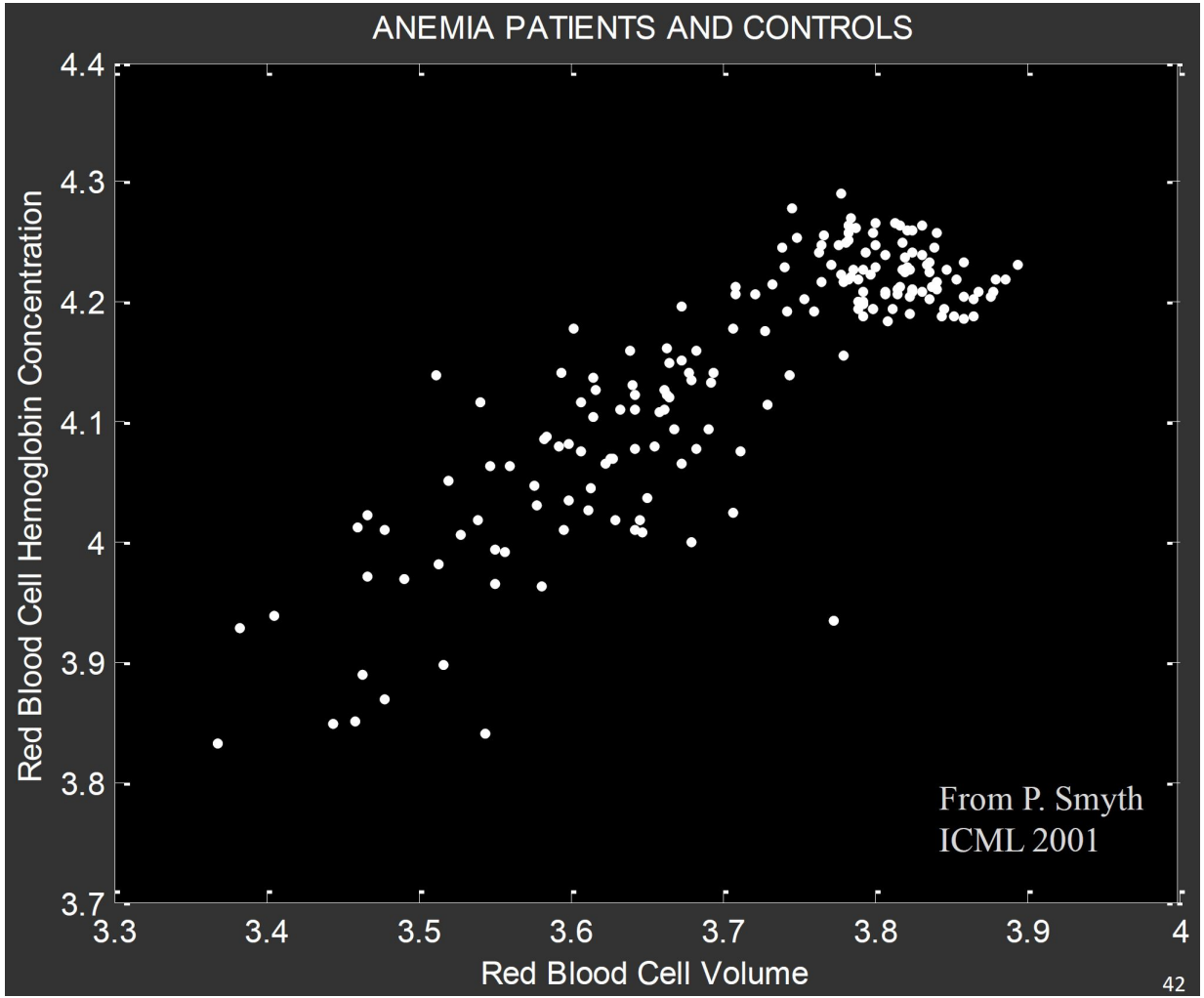
Fraction of total assigned to cluster c

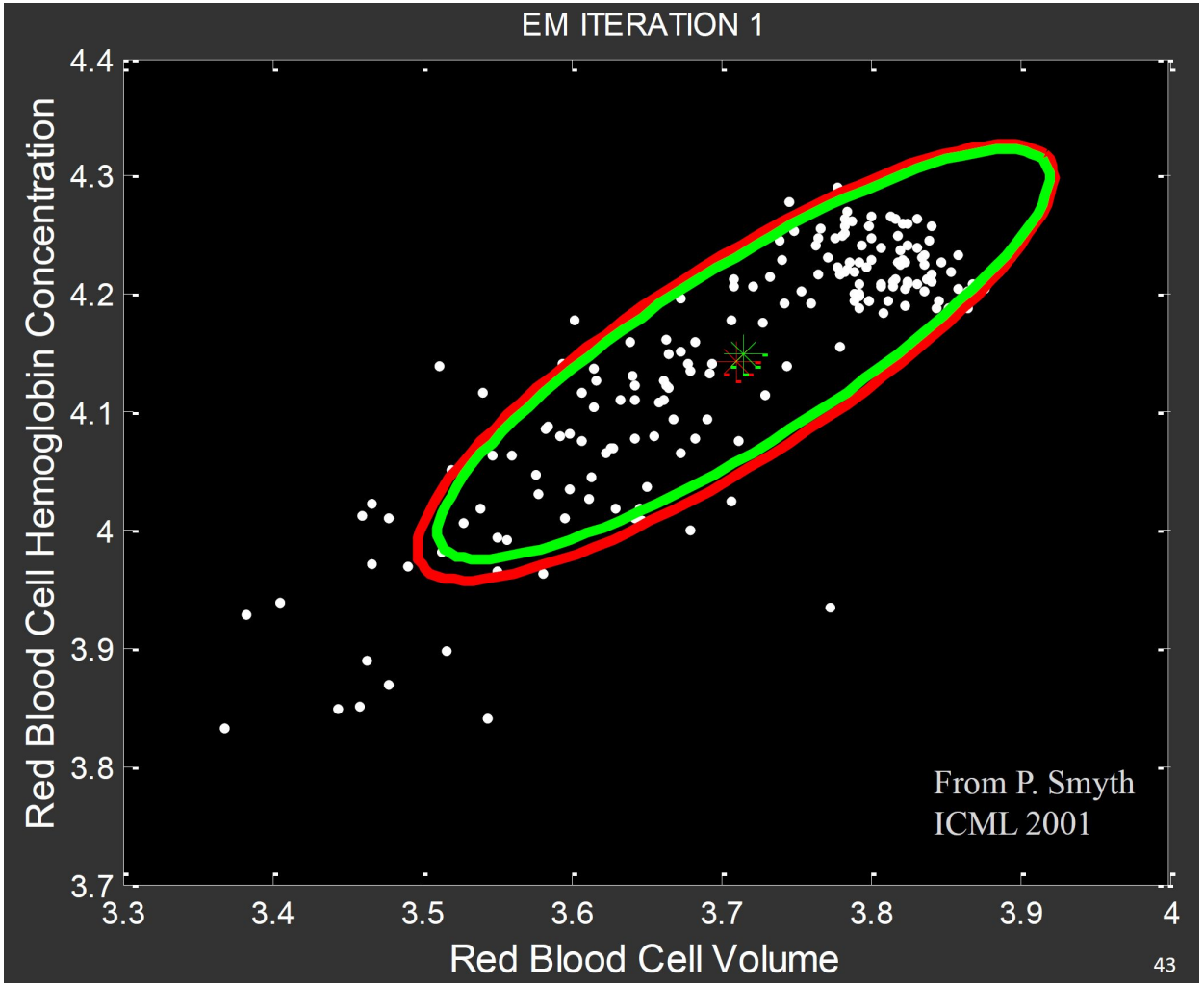
$$\mu_c = \frac{1}{m_c} \sum_i r_{ic} x^{(i)}$$

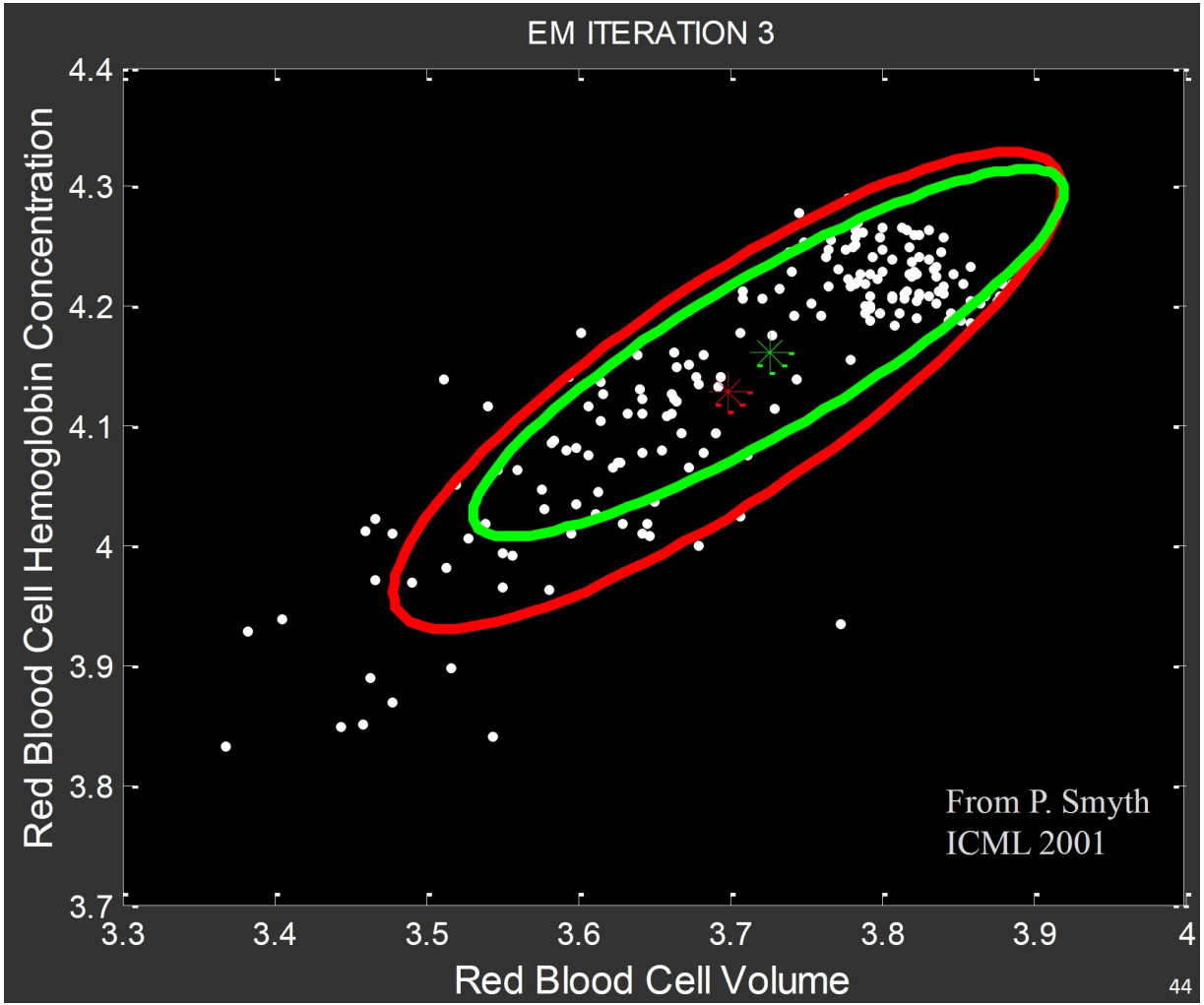
Weighted mean of assigned data

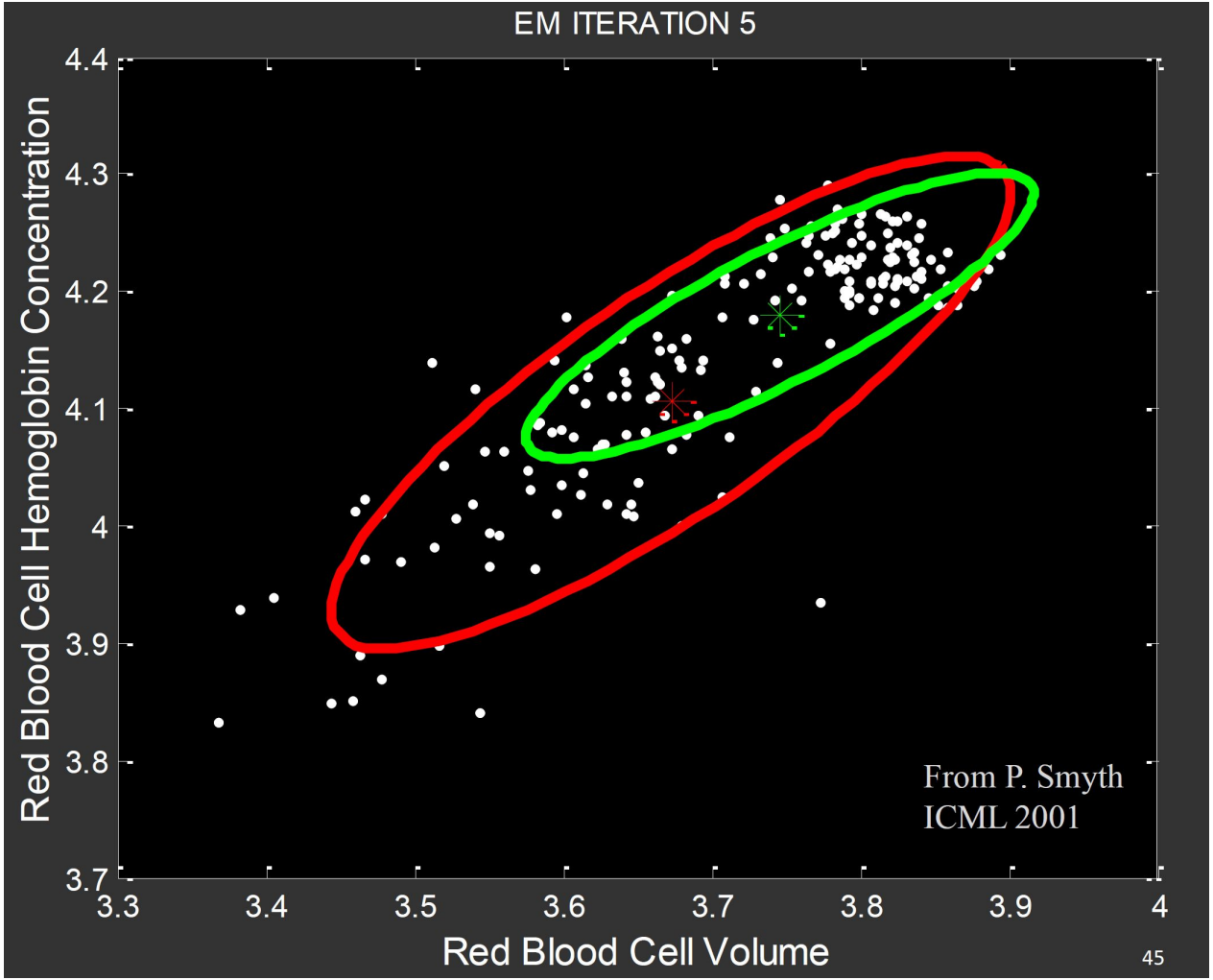
$$\Sigma_c = \frac{1}{m_c} \sum_i r_{ic} (x^{(i)} - \mu_c)^T (x^{(i)} - \mu_c)$$

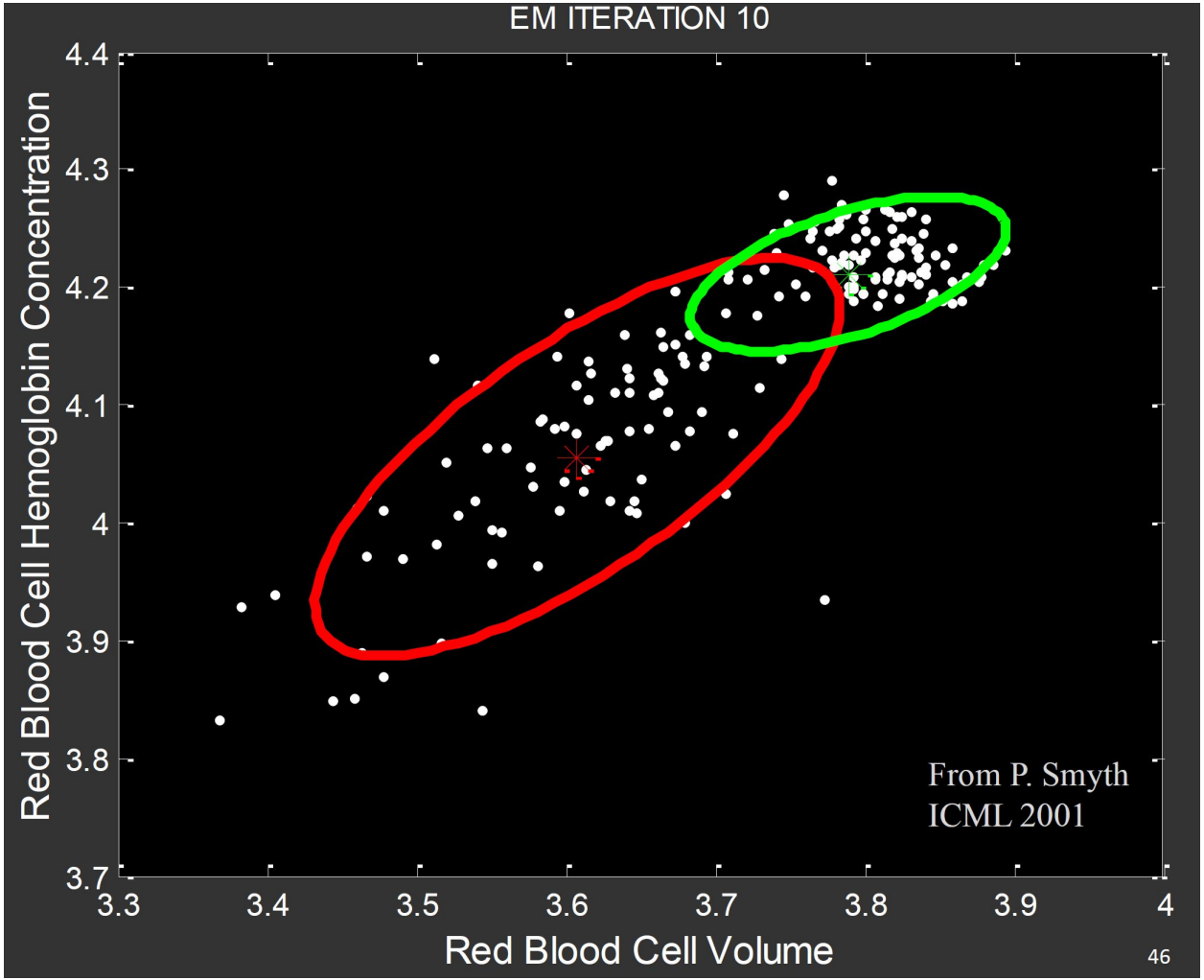
Weighted covariance of assigned data

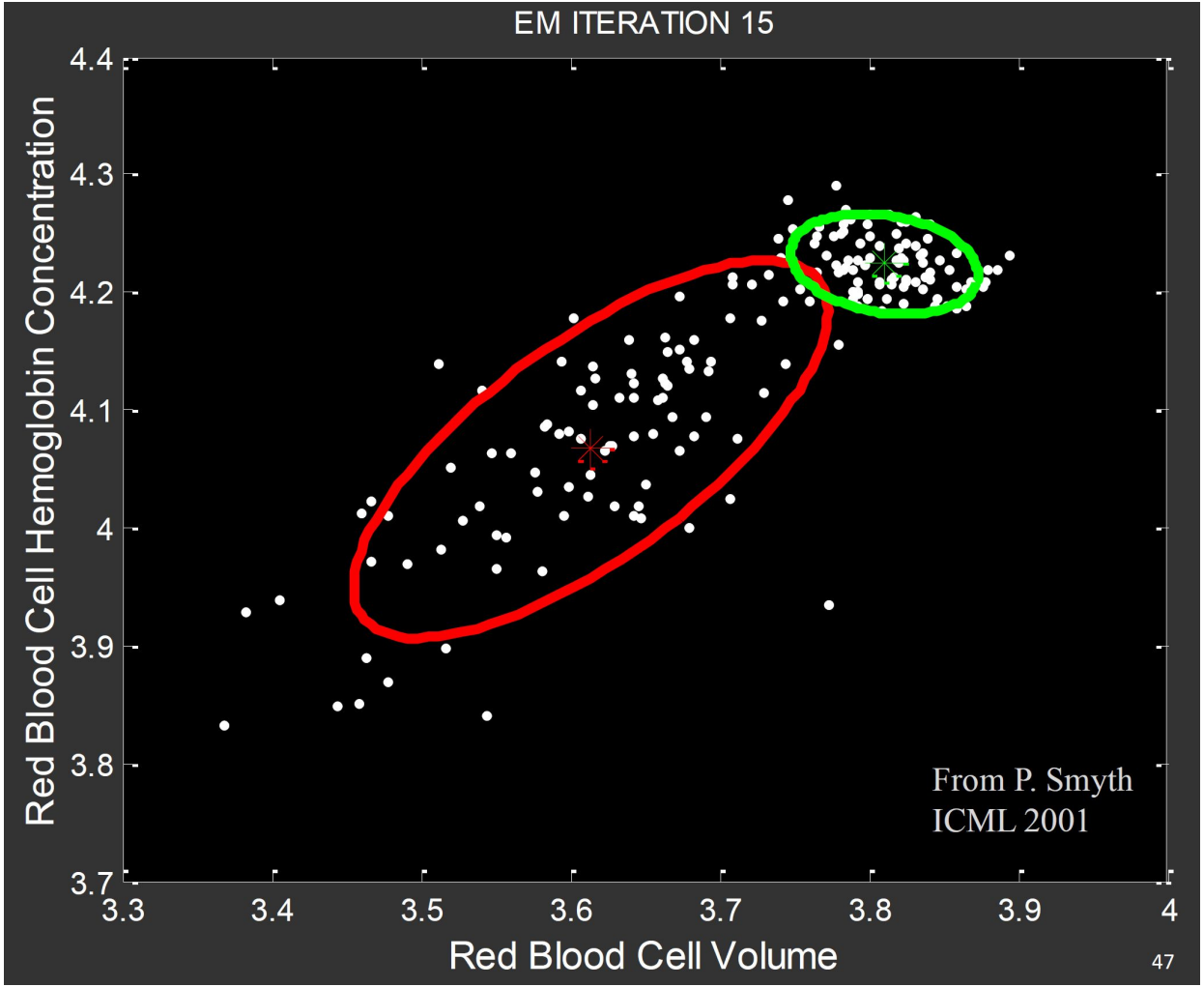

ANEMIA PATIENTS AND CONTROLS
Red Blood Cell Hemoglobin Concentration
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
Red Blood Cell Volume
From P. Smyth
ICML 2001

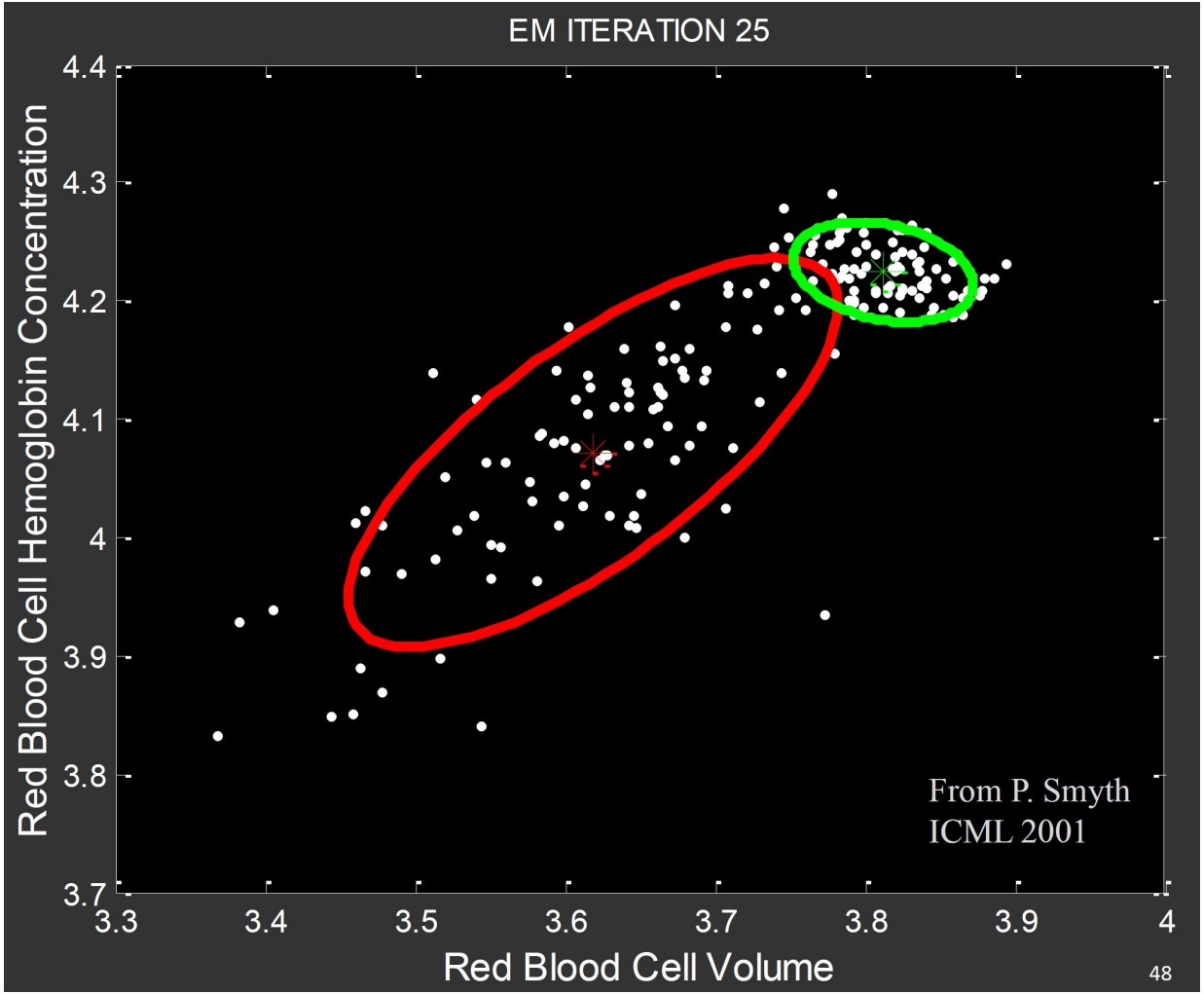

EM ITERATION 1
Red Blood Cell Hemoglobin Concentration
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
Red Blood Cell Volume
From P. Smyth
ICML 2001
43

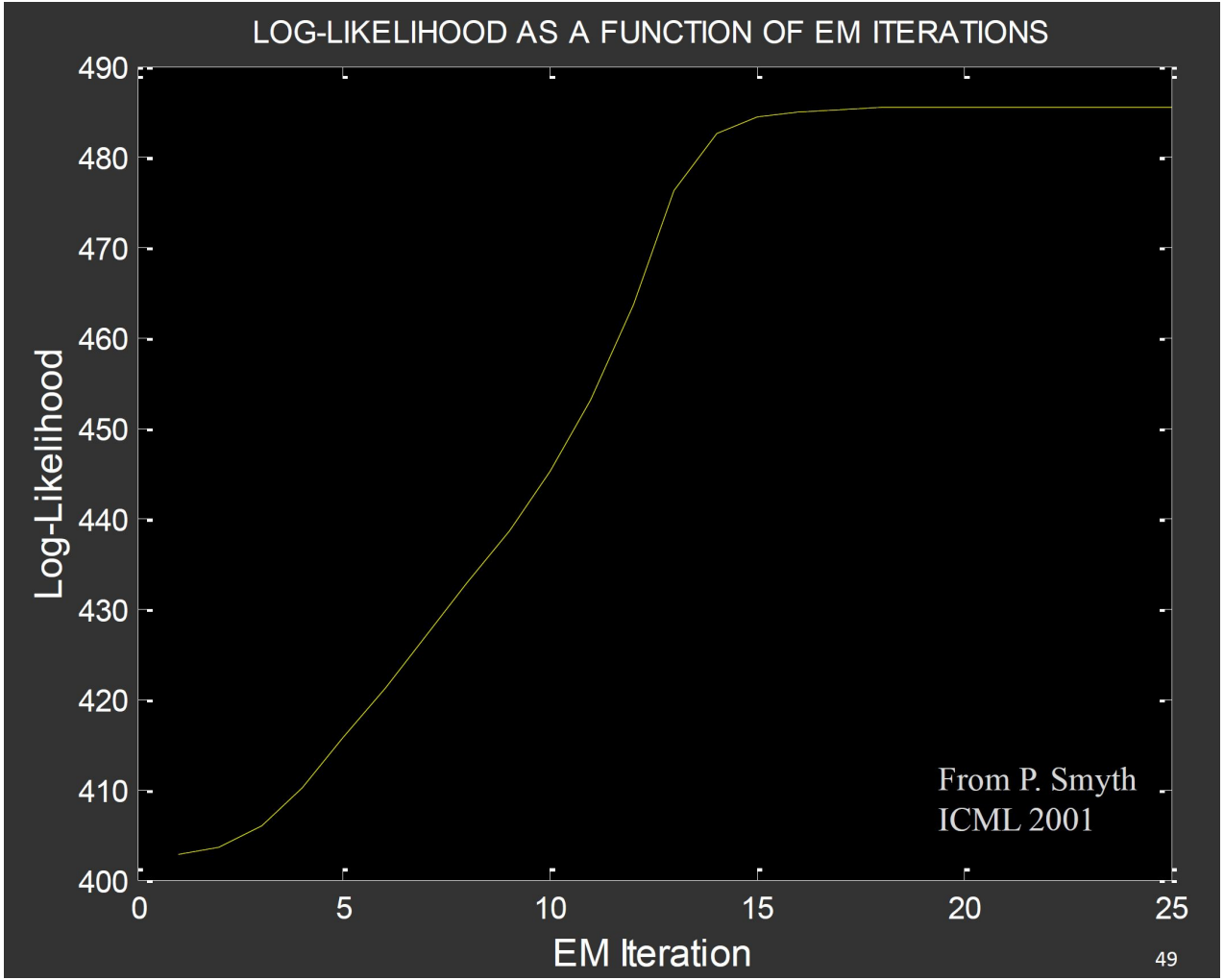

EM ITERATION 3
Red Blood Cell Hemoglobin Concentration
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
Red Blood Cell Volume
From P. Smyth
ICML 2001


EM ITERATION 5
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001


EM ITERATION 10
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001


EM ITERATION 15
Red Blood Cell Hemoglobin Concentration
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
Red Blood Cell Volume
From P. Smyth
ICML 2001
47


EM ITERATION 25
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001
48


LOG-LIKELIHOOD AS A FUNCTION OF EM ITERATIONS
Log-Likelihood
EM Iteration
490
480
470
460
450
440
430
420
410
400
0
5
10
15
20
25
From P. Smyth
ICML 2001

Demo Time

https://lukapopijac.github.io/gaussian-mixture-model/