Clustering

Clustering

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to objects in other clusters

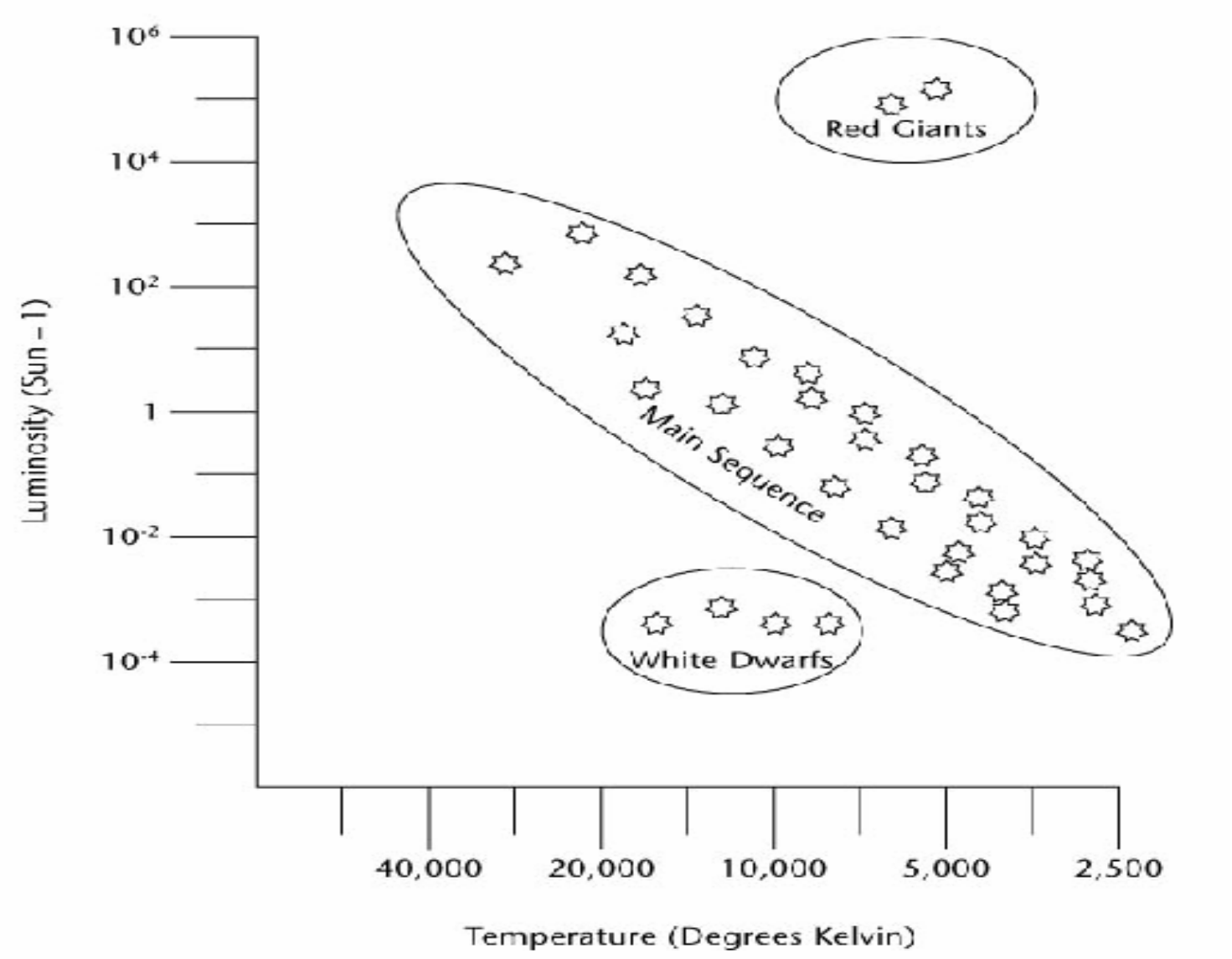
- Cluster analysis
  - Grouping a set of data objects into clusters

- Unsupervised classification: no predefined classes

- Typical applications
  - Making sense of structure of complex data
    Break up large data into meaningful subsets
  - Customer segments
  - Prototypical cases, outliers

Hertzsprung-Russell diagram
Star clusters by temp. and brightness

Clusters represents stars at different phases in stellar life-cycle



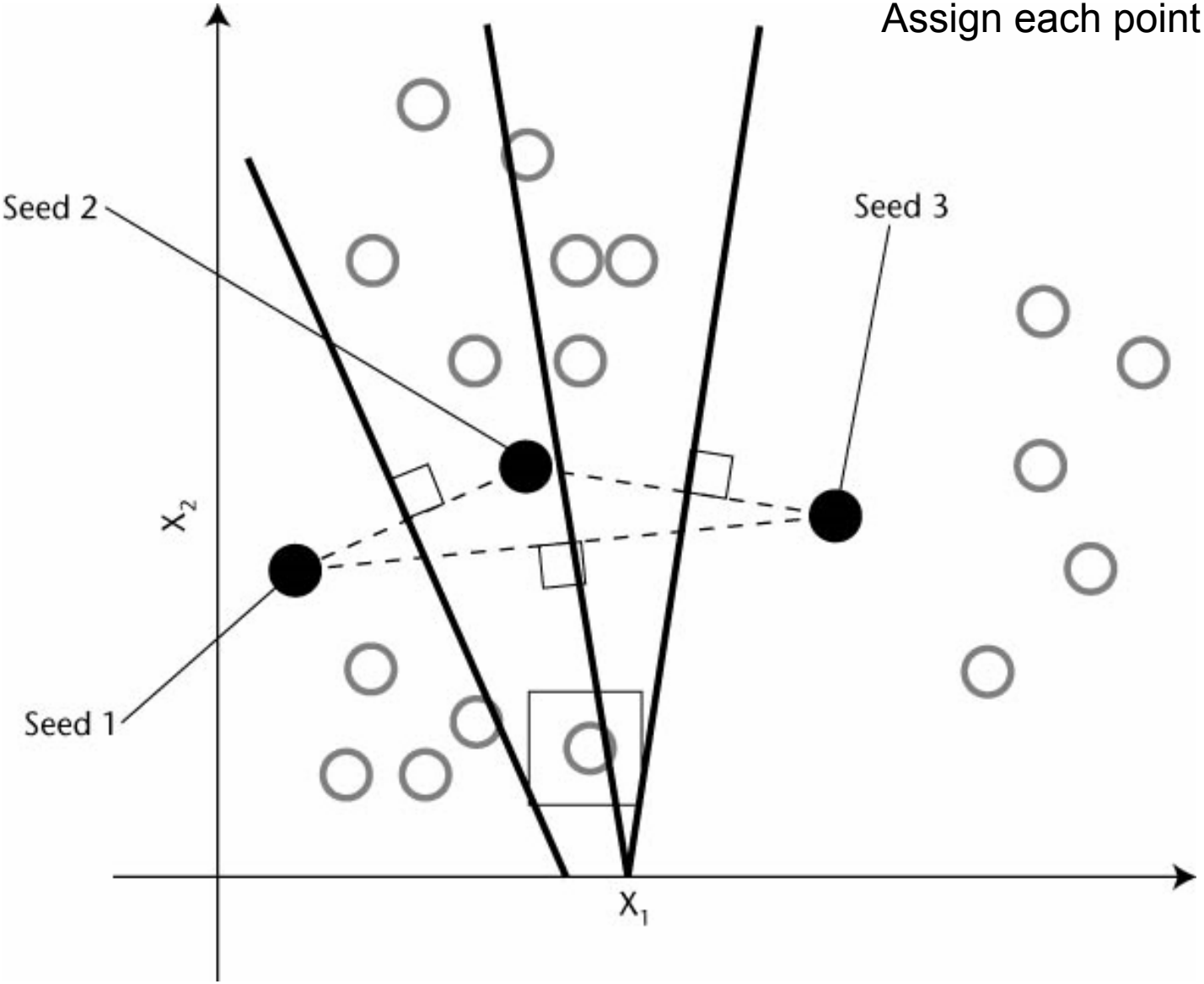
K-Means Clustering

Partitioning approach
Data as points in n-dimensional space

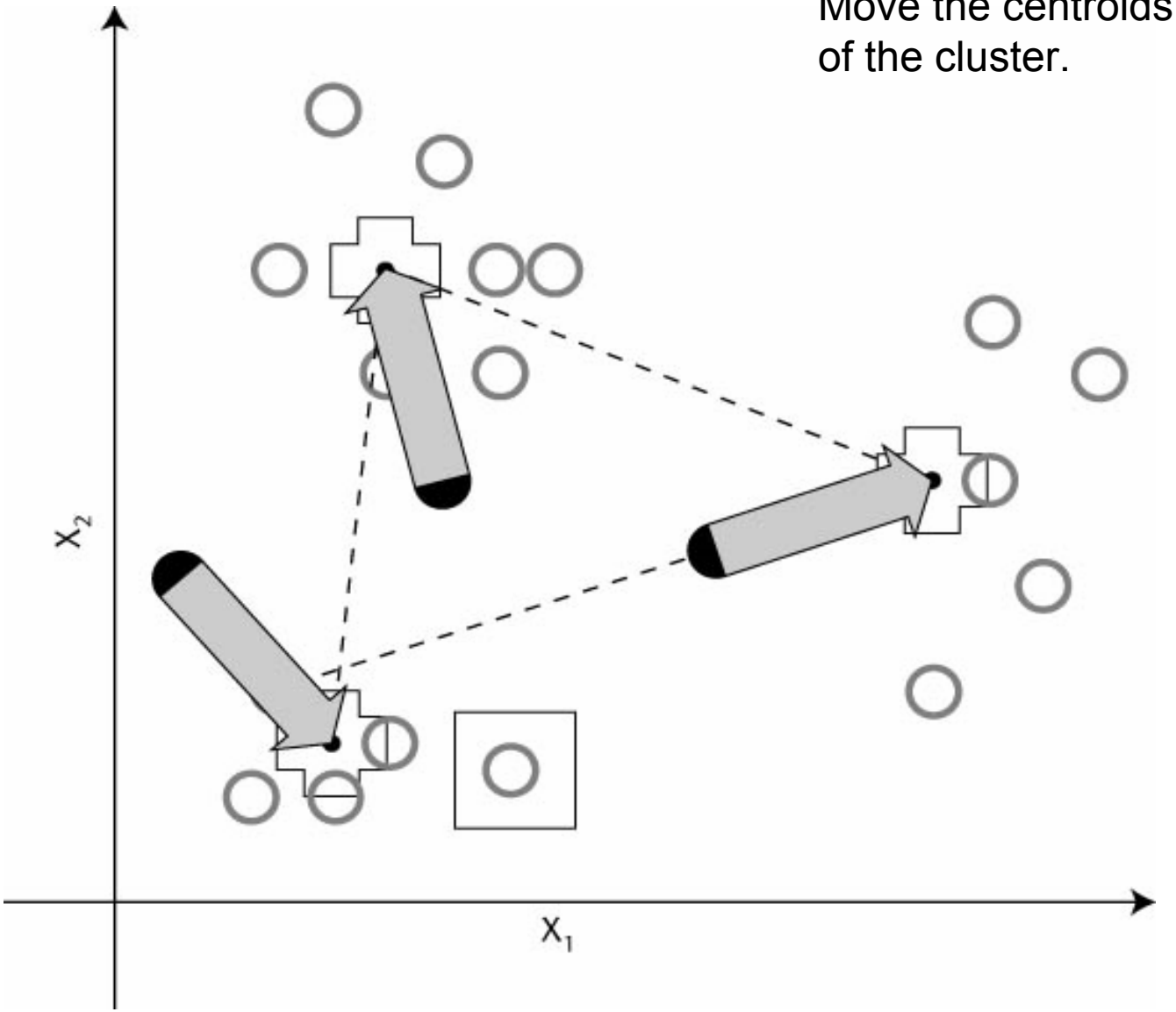
1. Start with k seeds as cluster centroids
   (Normally taken as k data points)
2. Calculate distance of all points from cluster centers
   Assign each data point to closest cluster
3. Re-calculate cluster centroids from all data point in the cluster (averages in each cluster)

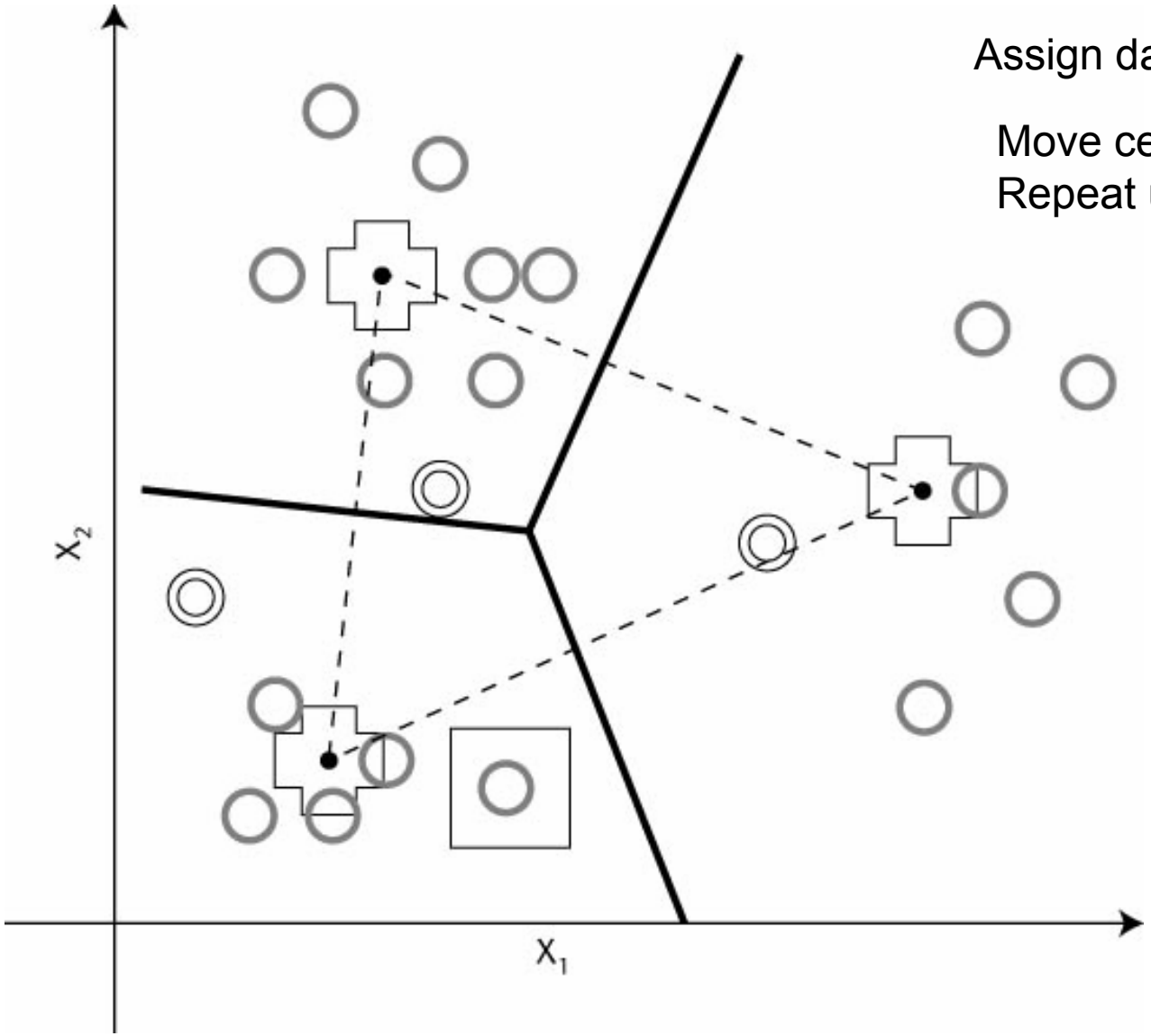
Repeat from 2.

Select k initial seeds
Assign each point to a centroid



Move the centroids to the averages of the cluster.



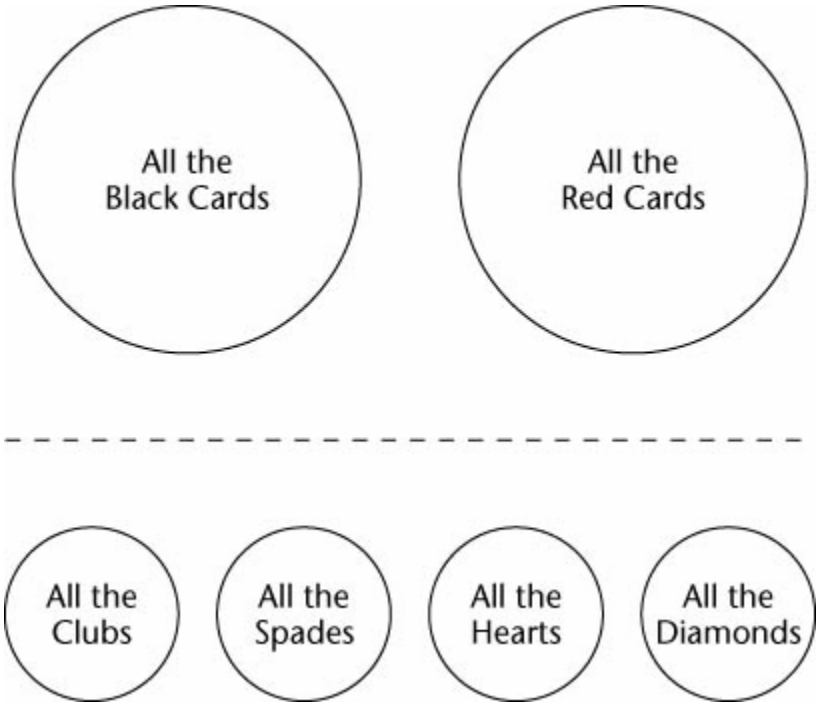


Assign data to new clusters

Move centroids

Repeat until centroids are stable

How many clusters (k)?



- What are we looking for?
- Try with different values of k
- Select one that yields best clusters
  - Low variance in cluster
  - Large distance between clusters

Drawbacks of K-means clustering

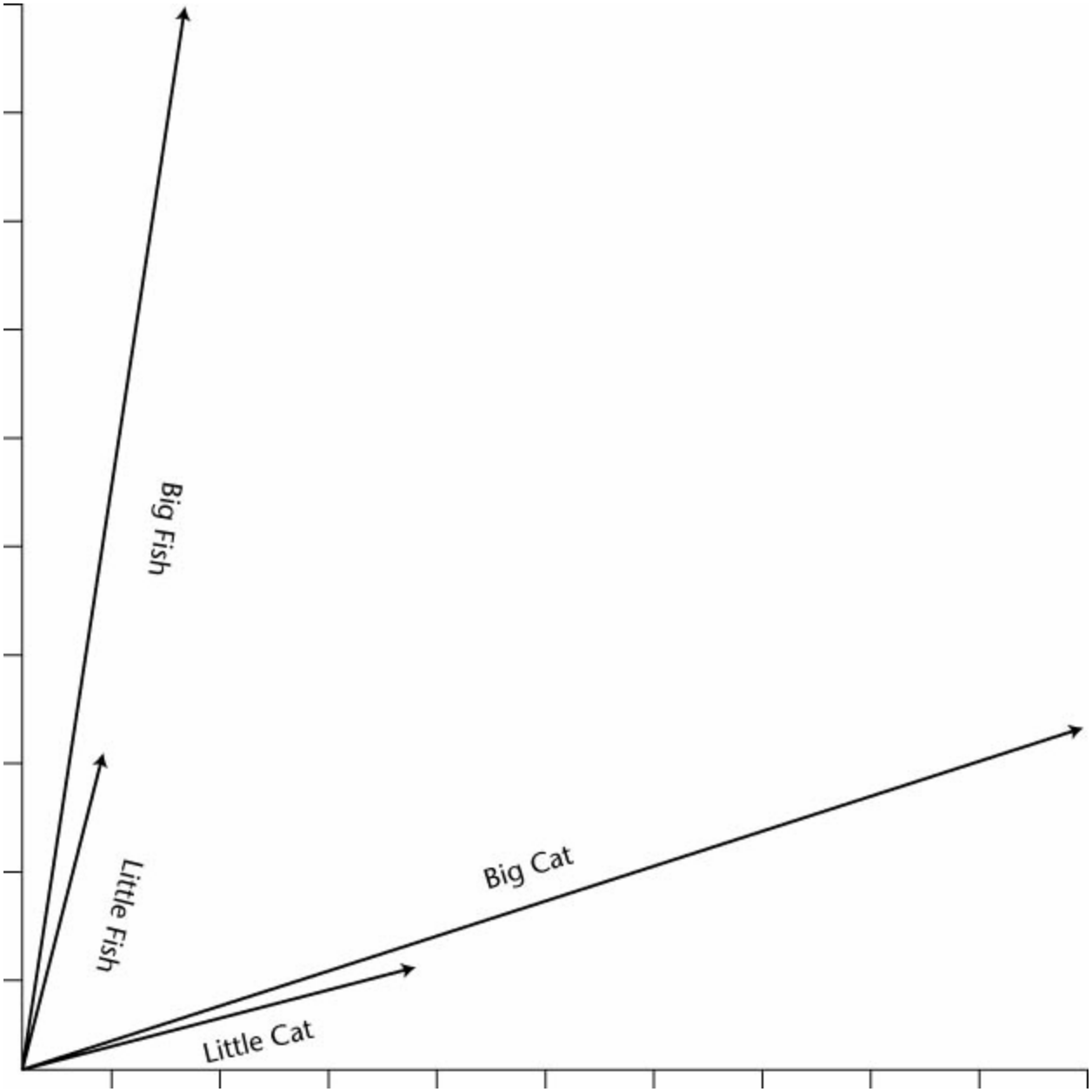
- Does not work well with overlapping clusters
- Sensitive to outliers, noise
- Each data point is either in a cluster or not – some form of membership score can be more reasonable
- Not suitable for discovering clusters with non-convex shapes
- Based on calculation of means
  - issues with using categorical data

Measuring distance

- Euclidean distance
- Manhattan distance
- Normalized sum of standardized values

- Categorical values
  Ratio of matching to non-matching fields

- Angle between vectors representing the data points
  Useful where within record similarities are important


Big Fish
Little Fish
Big Cat
Little Cat

Gaussian Mixture Models

- Gaussian distribution
  - Generalizes normal distribution to many variables
  - Often assumed for high-dimensional data

- Distribution of points is described by K different density functions
- Each Gaussian has *responsibility* for each data point
  - Strong responsibility of close points, low responsibility for distant points

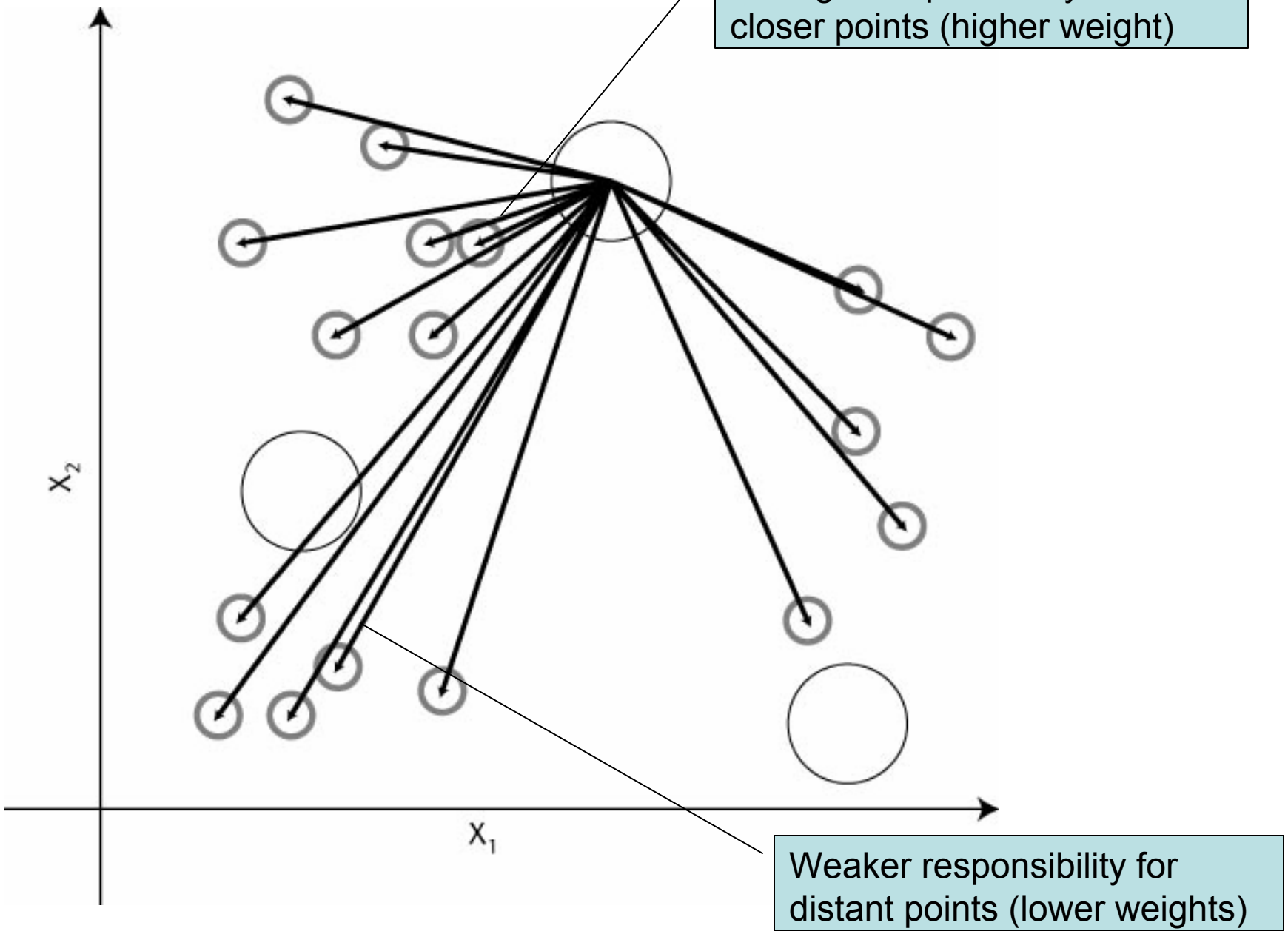
Gaussian mixture models

1. K seeds (considered as means of Gaussian distributions)
2. Estimation step: Calculate responsibility of each Gaussian for each data point
3. Maximization step: Using responsibility as weights, move the mean of the centroid towards the weighted average of all points

   Repeat 2 and 3 until the Gaussians no longer move.

Gaussians move and also change shape.

Each Gaussian is constrained – high responsibility for close points imply sharp drop off in responsibilities
(values must integrate to unity)
Larger Gaussians are weaker


Stronger responsibility for closer points (higher weight)
x_2
x_1
Weaker responsibility for distant points (lower weights)

Mixture models – soft clustering

Mixture model: probability at each data point is the sum of a mixture of many distributions

- Each point is tied to different distributions with different probabilities
- Soft clustering: points are not assigned to single cluster
- Data point can be assigned to single cluster that has strongest responsibility

Agglomerative clustering

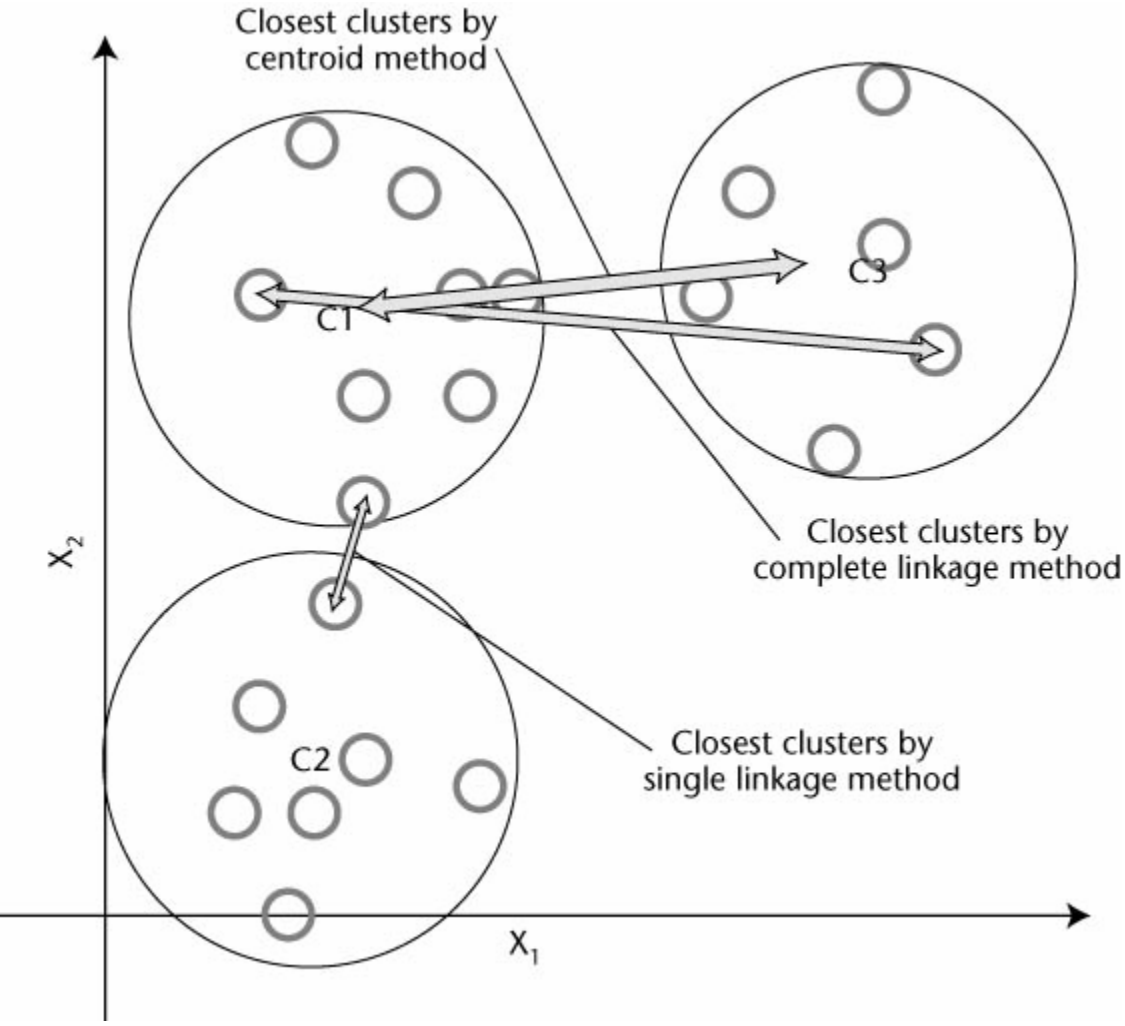
1. Begin with n clusters (n data points)
2. Create similarity matrix – pair-wise distances between points
3. Find two most similar clusters and merge them
4. Update the smaller similarity matrix

   Repeat until there is only one large cluster including all data points

Each step yields a candidate clustering

Clustering people by age
Distance: age difference

Measuring distance between clusters



Single linkage: distance between closest members
Every point in a cluster is closer to at least one point in the cluster than to any point outside the cluster.

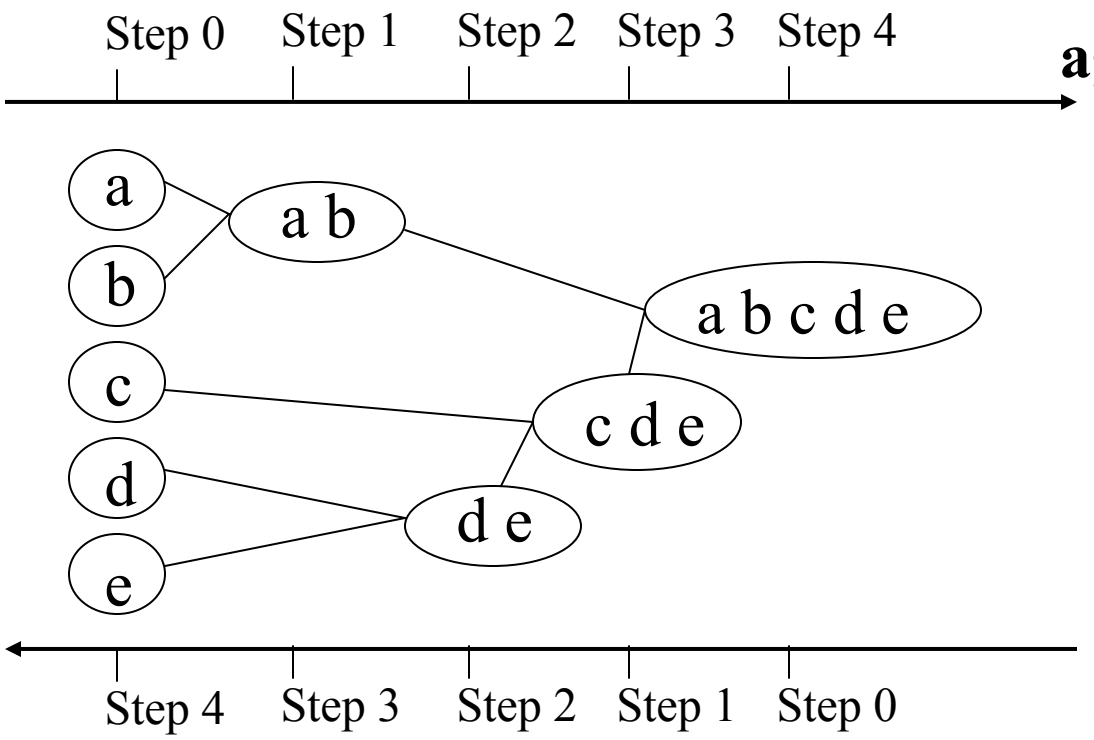
Complete linkage: distance between most distant members
All members of a cluster are within some maximum distance of one another.

Centroid distance: distance between centroids of clusters

Divisive clustering

- Divides data set into clusters of lower within-group variance
- Similar to decision trees
  Similarity metric as measure of node purity

Hierarchical Clustering
Agglomerative vs. Divisive



agglomerative (AGNES)

(Agglomerative Nesting)

divisive (DIANA)

(Divisive Analysis

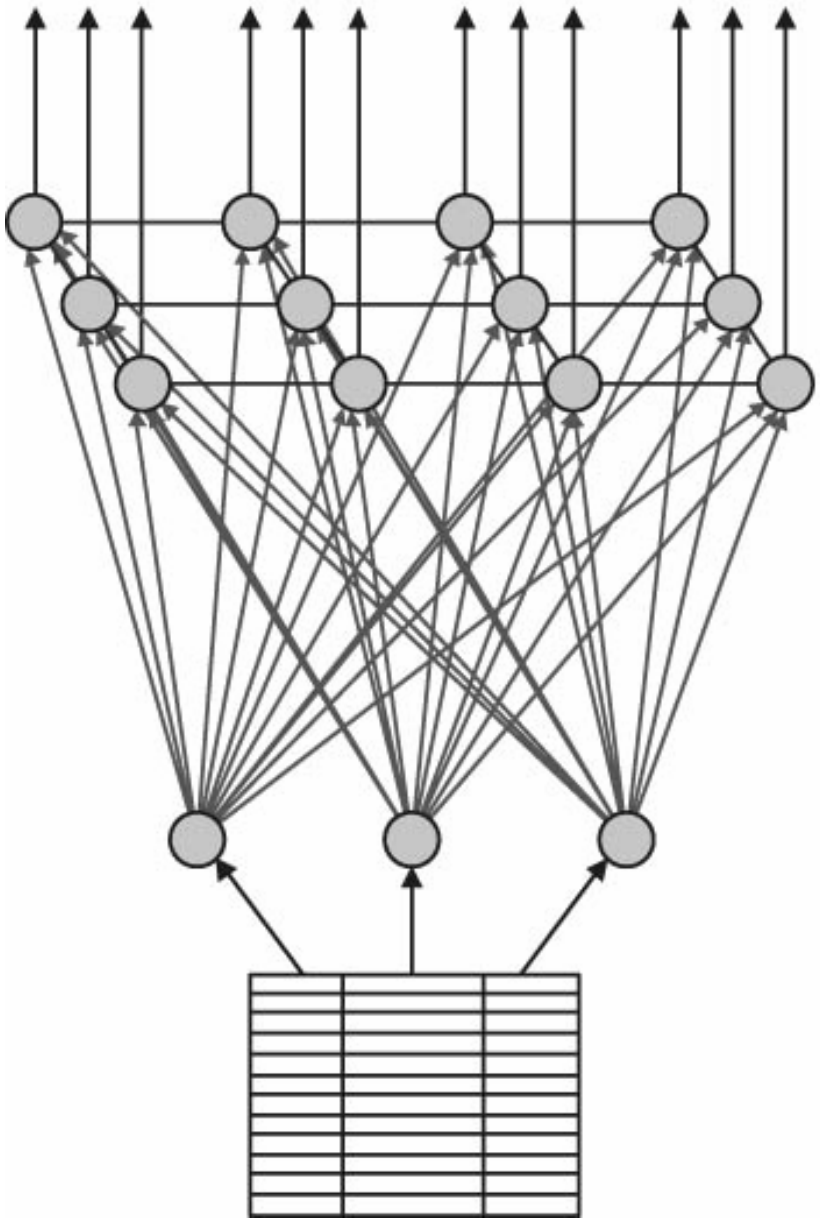
Evaluating clusters

- High within-cluster similarity
  Low Variance (sum of squares of distances from mean)
  Average variance (variance / clusterSize)

- Understanding clusters
  - Means of each variable calculated over points within a cluster, compared with over all means, or means in different clusters
  - Use decision tree to obtain rules describing different clusters
- One or two strong clusters with other weak clusters
  - Remove data points corresponding to strong clusters and apply clustering again on other data

- Single cluster can also be useful
  Distance from center can indicate rare cases (fraud, defects, etc)

Kohonen Nets



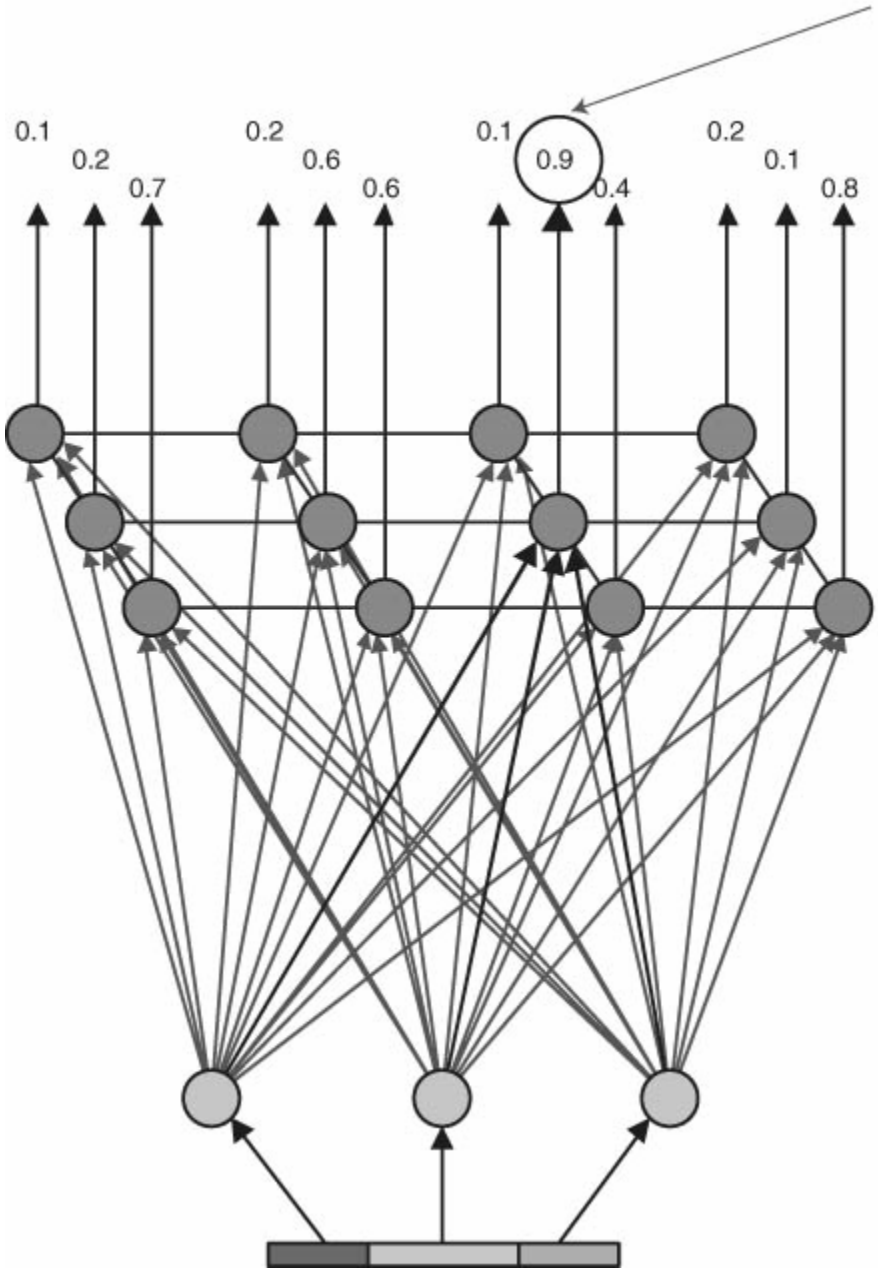
The output units compete with each other for the output of the network.

The output layer is laid out like a grid. Each unit is connected to all the input units, but not to each other.

The input layer is connected to the inputs.

Kohonen Nets (Self Organizing Maps)

- Topology preserving map
  - Topological structure on nodes (neurons)
- Competition in learning
  - Only the highest activation neuron is allowed to output (winner takes all)
  - Only winner and its *neighbors* update weights during training
- Feature map
  - Consider two input vectors x1 and x2, and let n1 and n2 be the neurons that ‘fire’ on these two inputs respectively. If x1 is similar to x2, then n1 and n2 should be close to each other



Single neuron with highest output ‘fires’
Paths to winner neuron strengthened
Paths to neighbors in output layer grid are also strengthened

Group of output neurons may represent a cluster.

Kohonen weight update

- On input x, let n_i be the winning neuron
  Weight update for neuron n_k:

  $W_k(new) = w_k(old) + \eta\, q(I, k)\, (x - w_k)$

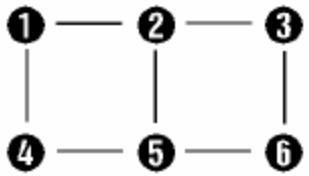
  q (i, k) is a neighborhood function
  = 1 for i = k
  Value decreases with increasing distance between n_i and n_k

  Example: $q(i, k) = \exp(- dist(i, k)^2 / 2\sigma^2)$

  where σ is a width parameter that decreases over time.

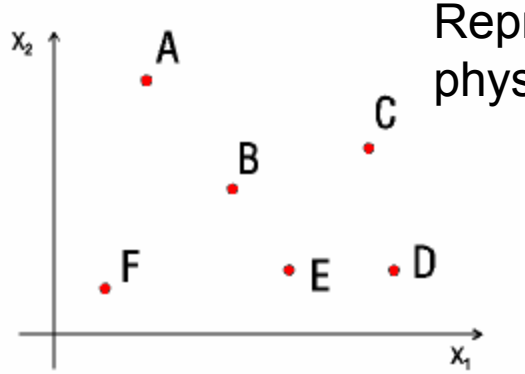
Simple example

2 input neurons, 6 output neurons in a grid

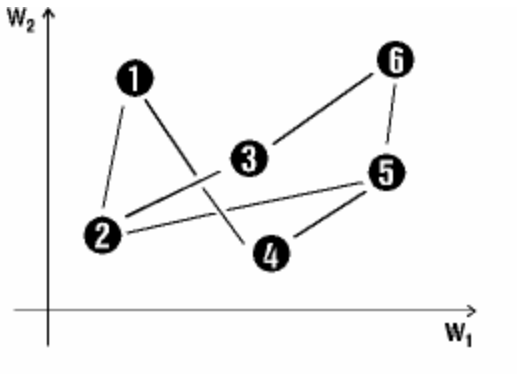


6 inputs presented to the net

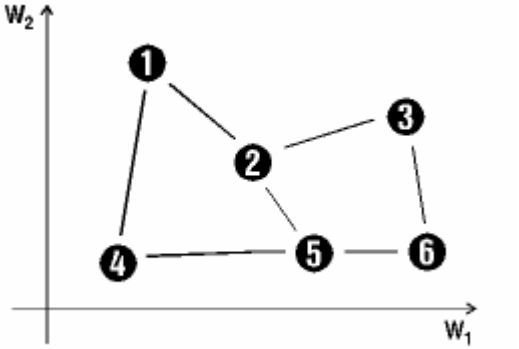
Representation in physical space



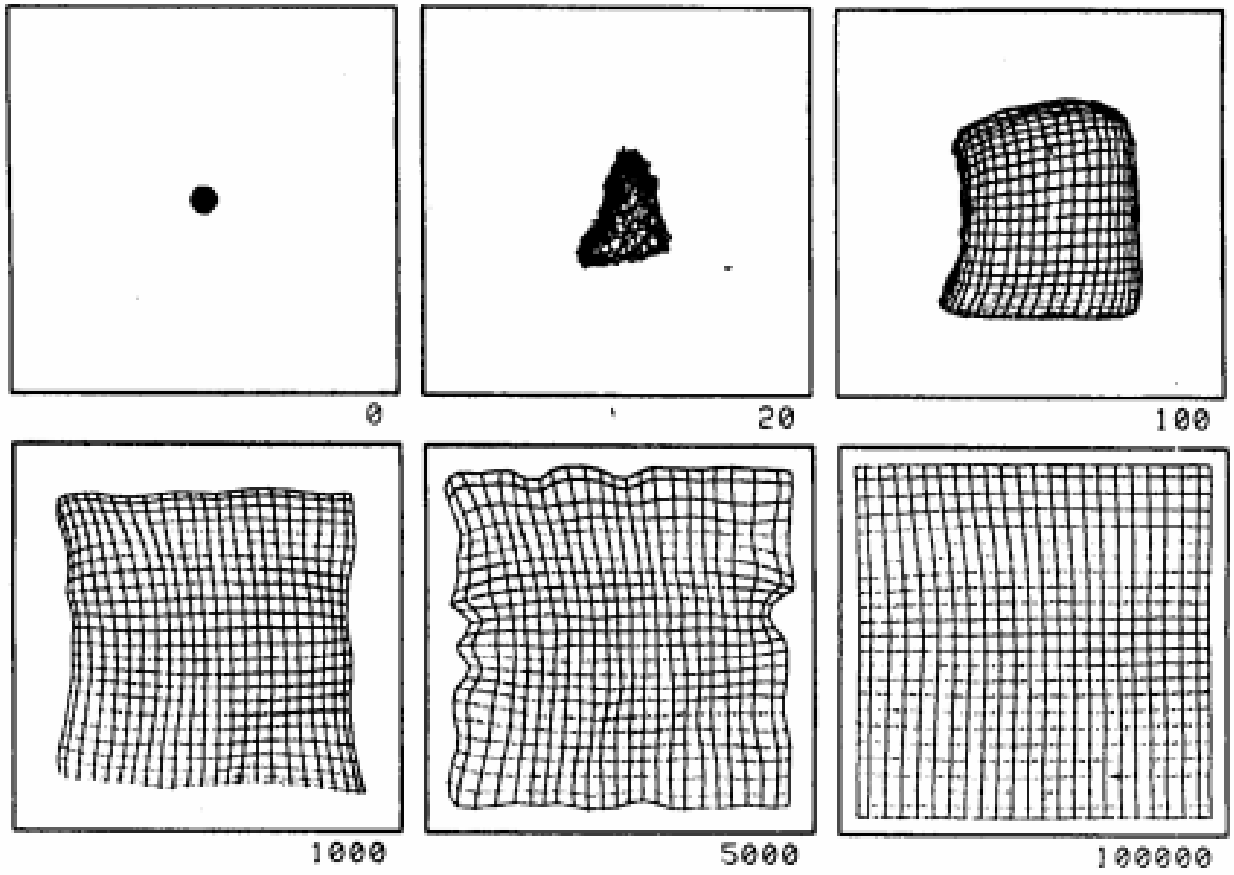
Initial random weights to the 6 output neurons



Weight space representation after training



5. Self-Organizing Feature Maps

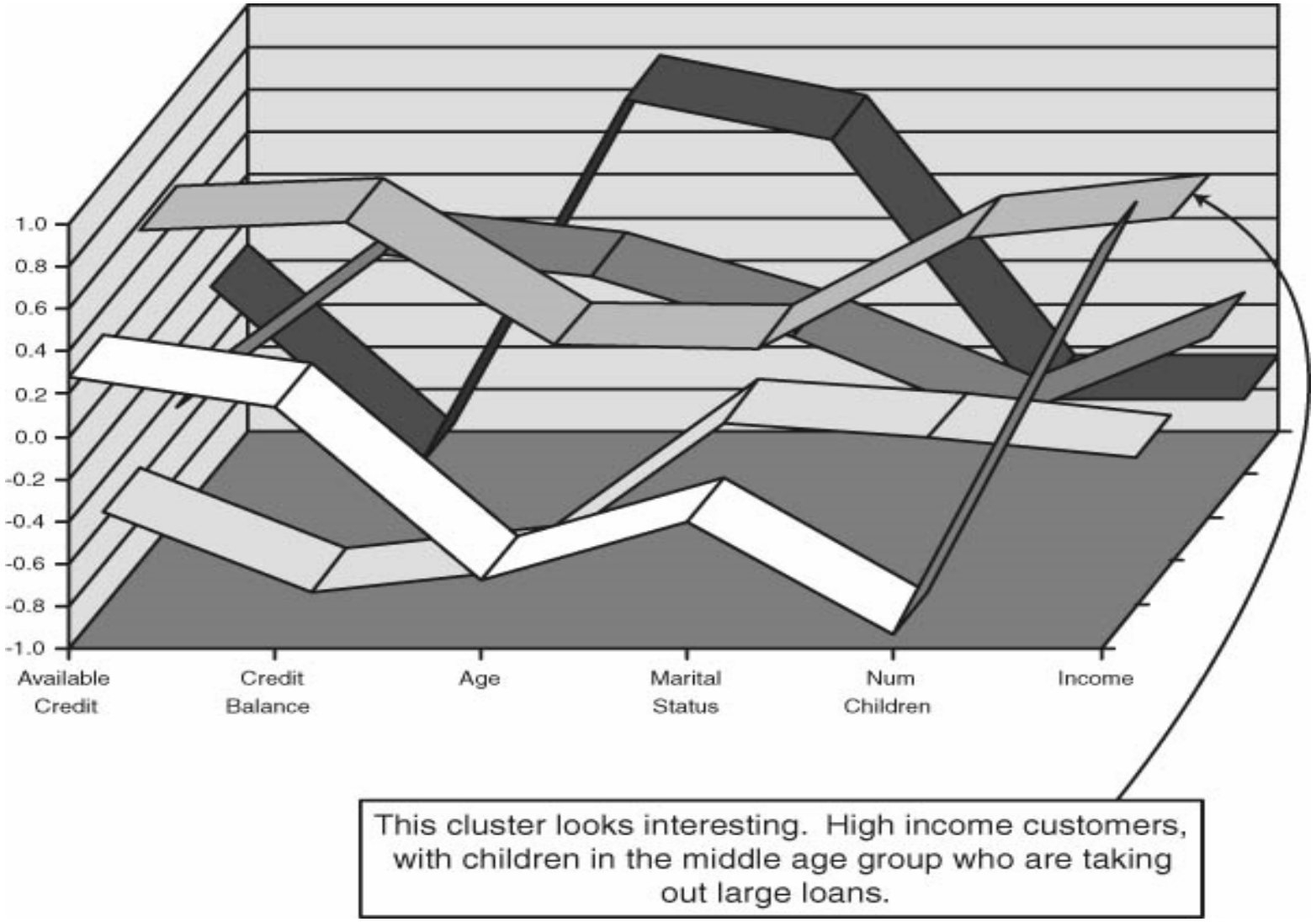


Identifying clusters with a Kohonen net

- Large bank is interested in increasing the number of home-equity loans that it sells. Bank wants to understand customers who currently have home-equity loans, to help determine the best strategy for increasing its market share

- Data on 5000 customers with home equity loans and 5000 customers w/o home equity loans
  - Appraised value of property
  - Amount of available credit
  - Amount of credit granted
  - Age
  - Marital status
  - Number of children
  - Household income

- Kohonen net identifies 5 clusters
- What do the clusters mean?



Children's age? – in their late teens

Home equity loans to fund college education?

- Disappointing results with marketing campaign designed for college tuition

- Include additional data (all accounts, credit data, etc.)
  - Cluster of customers with college age children
  - These customers have business as well as personal accounts

  Parents starting new business when children leave home.