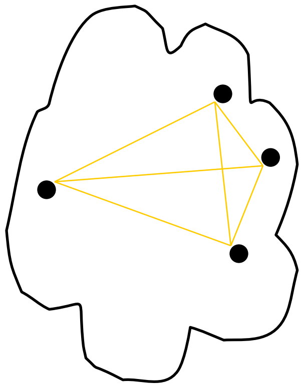
CLUSTERING

CLUSTERING ALGORITHMS

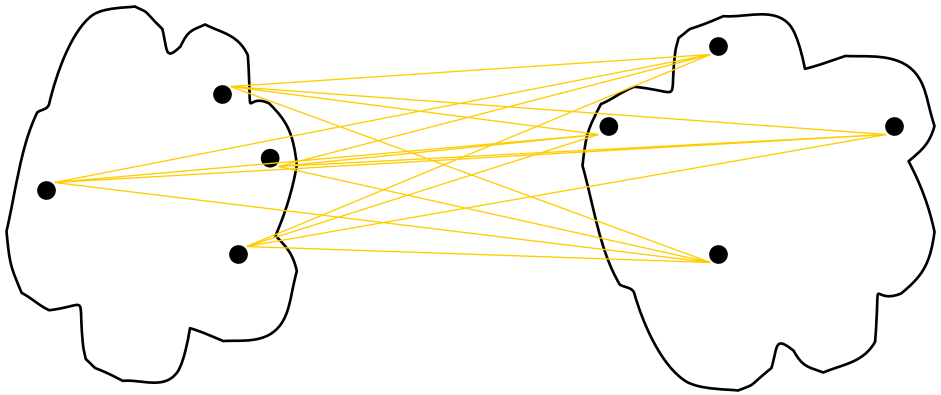
- K-means and its variants
- Hierarchical clustering
- Density-based clustering

UNSUPERVISED MEASURES: COHESION AND SEPARATION

- A proximity graph-based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion



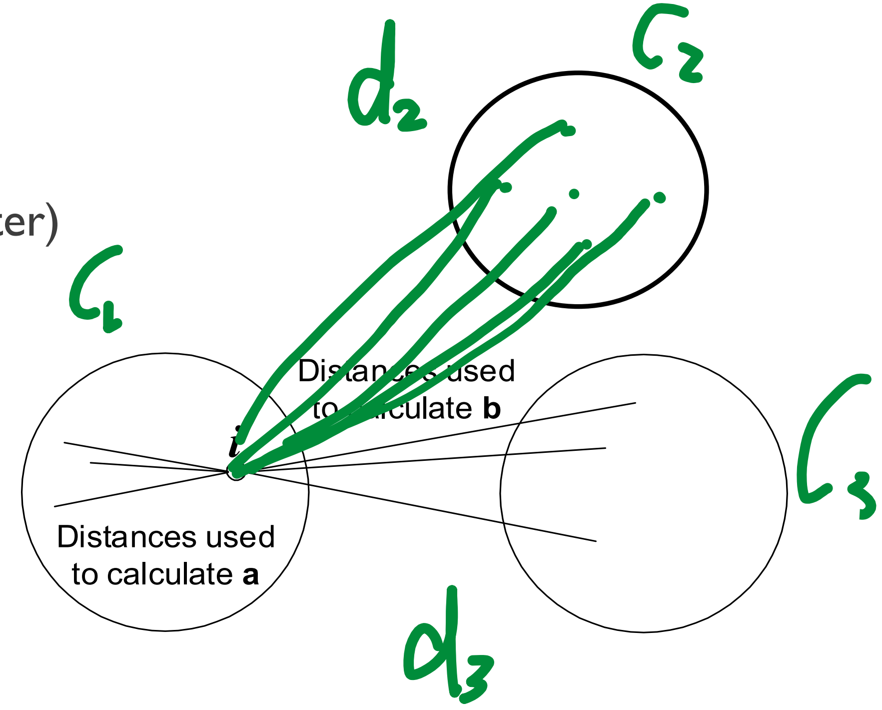
separation

UNSUPERVISED MEASURES: SILHOUETTE COEFFICIENT

- Silhouette coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, $\boldsymbol{i}$
  - Calculate $\boldsymbol{a}$ = average distance of $\boldsymbol{i}$ to the points in its cluster
  - Calculate $\boldsymbol{b}$ = min (average distance of $\boldsymbol{i}$ to points in another cluster)
  - The silhouette coefficient for a point is then given by

    s = (b – a) / max(a,b) → Intra-d inter-d

  - Value can vary between -1 and 1
  - Typically ranges between 0 and 1.
  - The closer to 1 the better.



b = min(d2, d3)

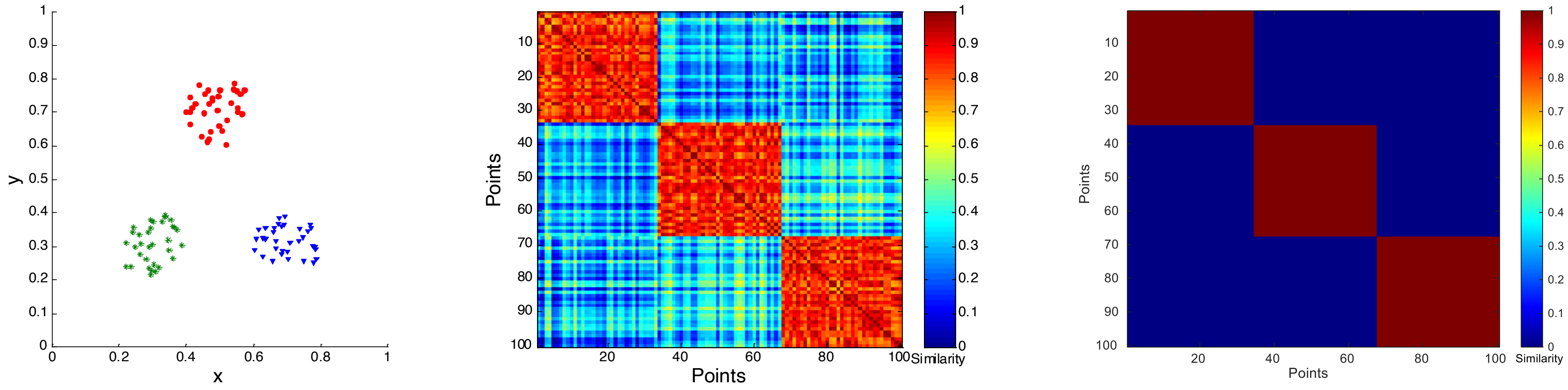
- Can calculate the average silhouette coefficient for a cluster or a clustering

MEASURING CLUSTER VALIDITY VIA CORRELATION

- Two matrices
  - Proximity Matrix
  - Ideal Similarity Matrix
    - One row and one column for each data point
    - An entry is 1 if the associated pair of points belong to the same cluster
    - An entry is 0 if the associated pair of points belongs to different clusters
- Compute the correlation between the two matrices
  - Since the matrices are symmetric, only the correlation between $n(n-1) / 2$ entries needs to be calculated.
- High magnitude of correlation indicates that points that belong to the same cluster are close to each other.
  - Correlation may be positive or negative depending on whether the similarity matrix is a similarity or dissimilarity matrix
- Not a good measure for some density or contiguity based clusters.

MEASURING CLUSTER VALIDITY VIA CORRELATION

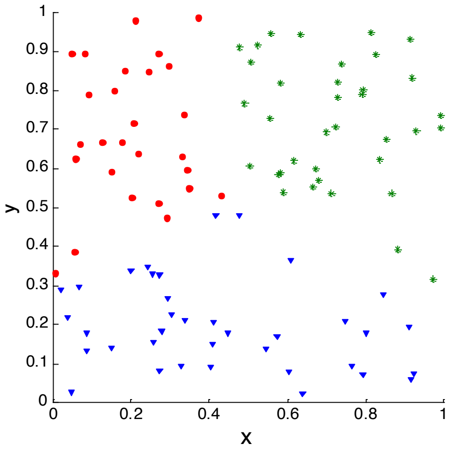
- Correlation of ideal similarity and proximity matrices for the K-means clusterings (partition cluster) of the following well-clustered data set.

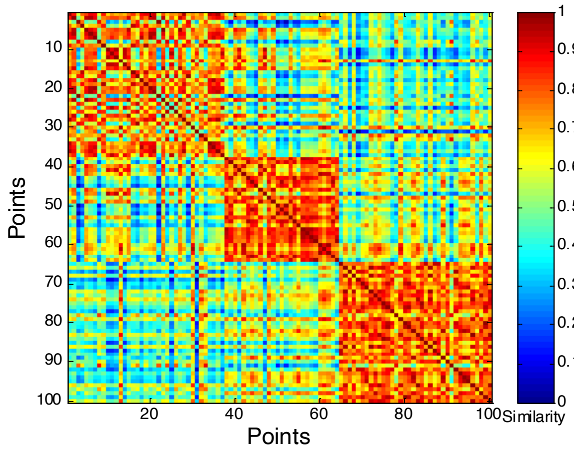


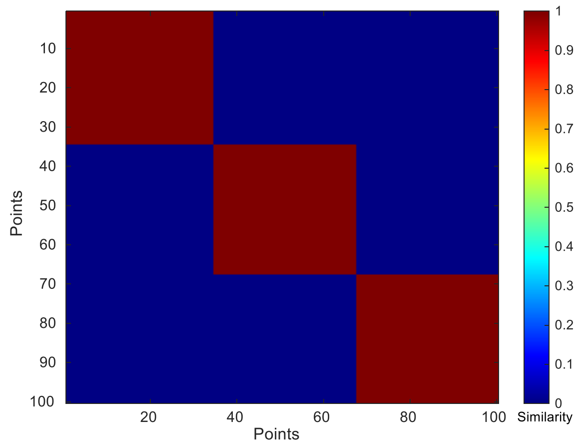
Corr = 0.9235

MEASURING CLUSTER VALIDITY VIA CORRELATION

- Correlation of ideal similarity and proximity matrices for the K-means clusterings of the following random data set.





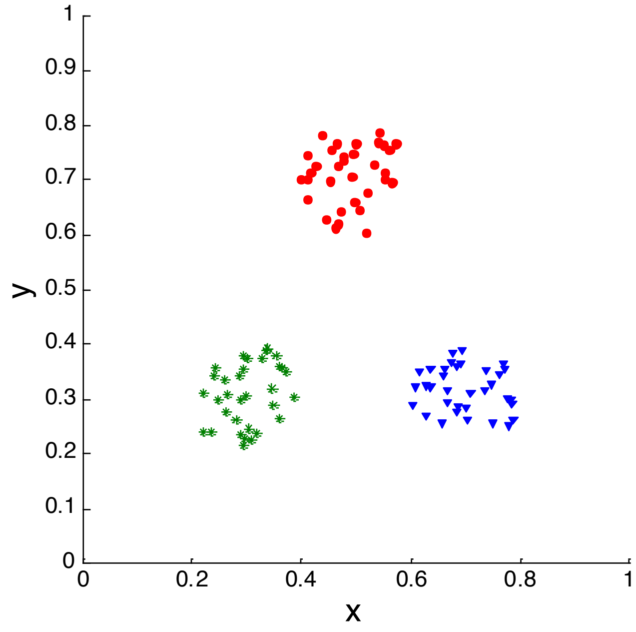


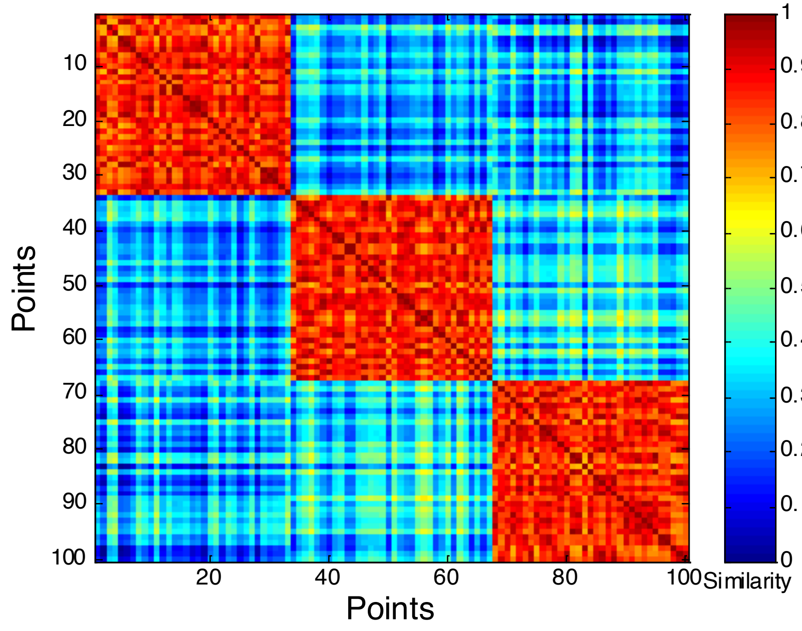
K-means

Corr = 0.5810

JUDGING A CLUSTERING VISUALLY BY ITS SIMILARITY MATRIX

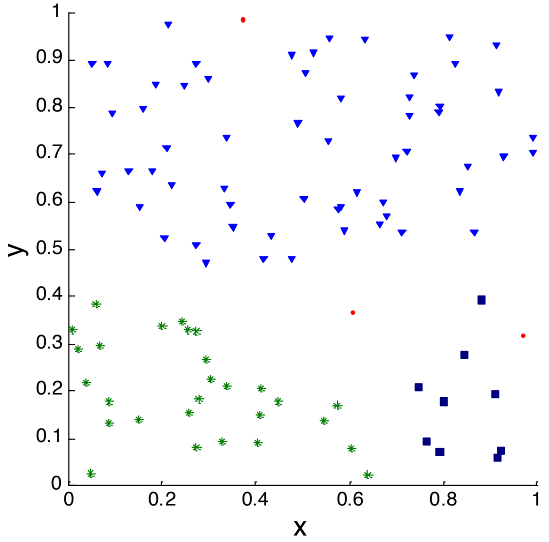
- Order the similarity matrix with respect to cluster labels and inspect visually.

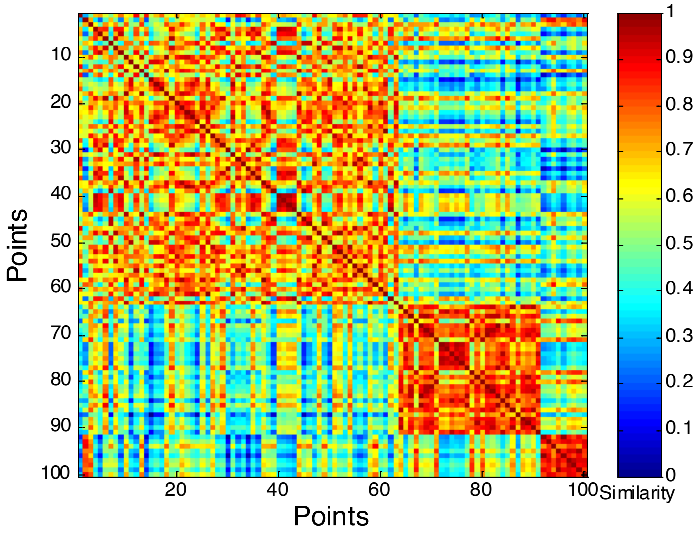




JUDGING A CLUSTERING VISUALLY BY ITS SIMILARITY MATRIX

- Clusters in random data are not so crisp

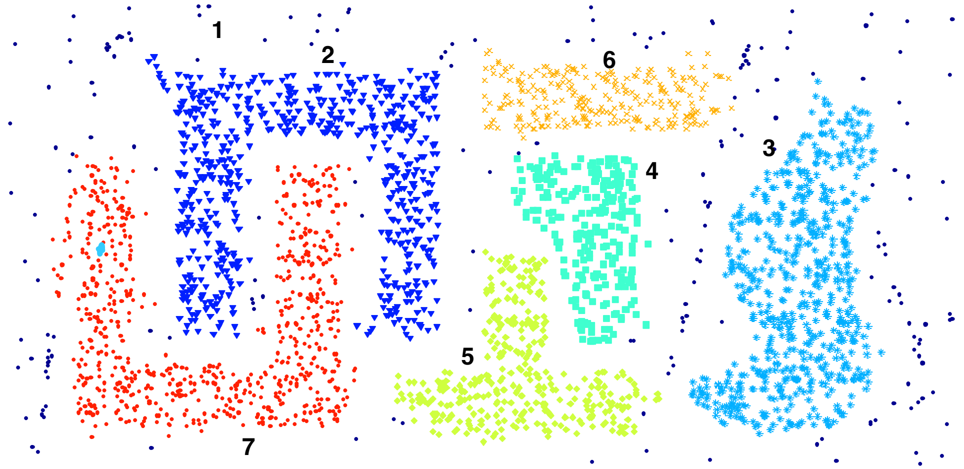


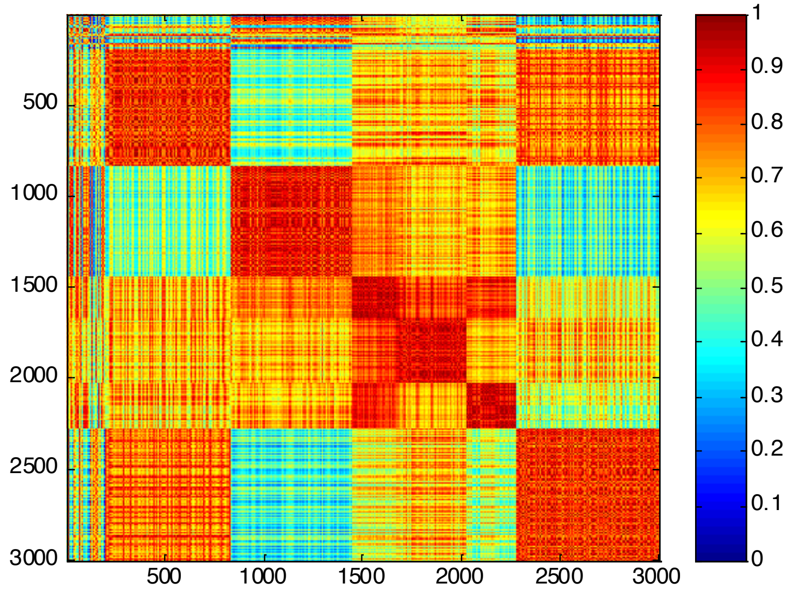


DBSCAN (density based clustering)

Correlation may be not a good measure for some density-based clusters.

JUDGING A CLUSTERING VISUALLY BY ITS SIMILARITY MATRIX





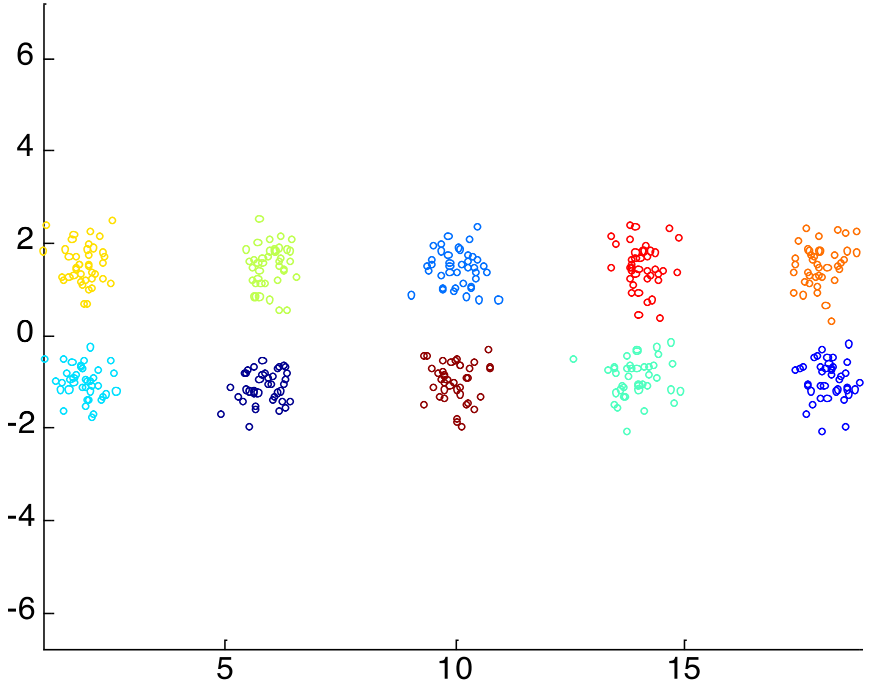
DBSCAN

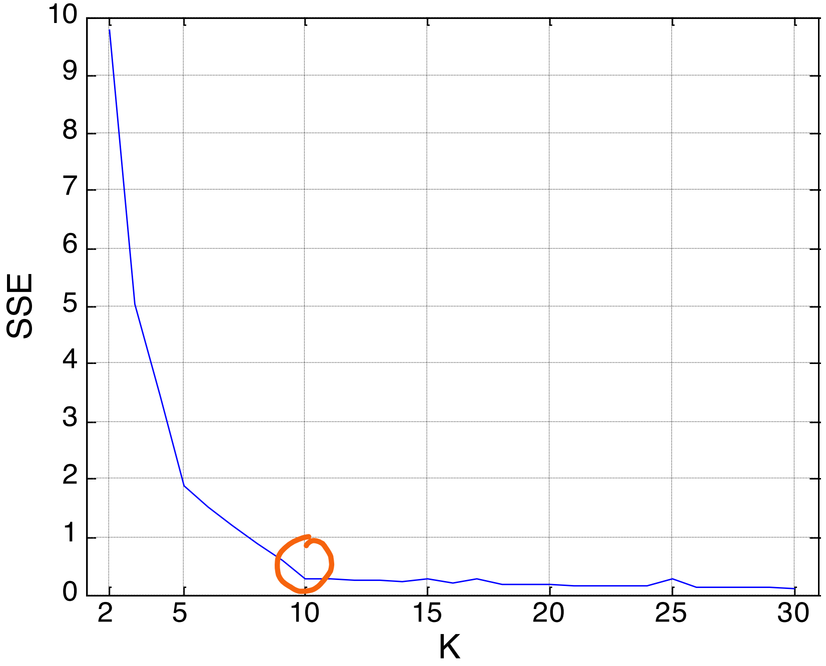
DETERMINING THE CORRECT NUMBER OF CLUSTERS

- SSE is good for comparing two clusterings or two clusters
- SSE can also be used to estimate the number of clusters

Elbow: after that point, the values of s
Do not change dramatically

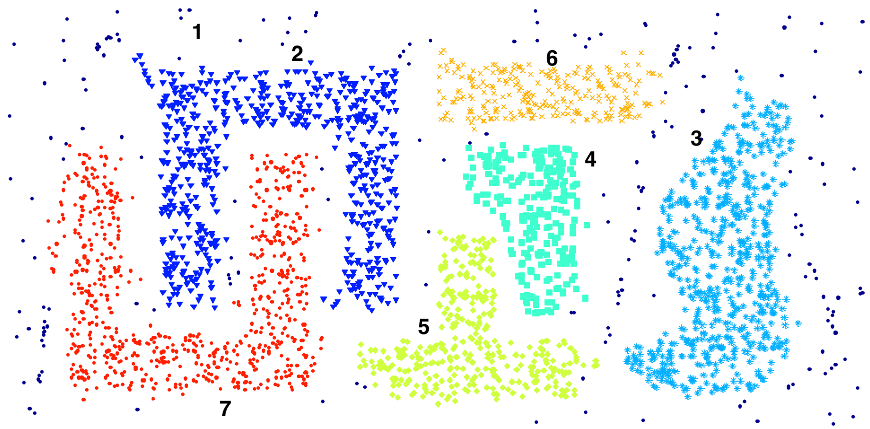
"elbow curve"



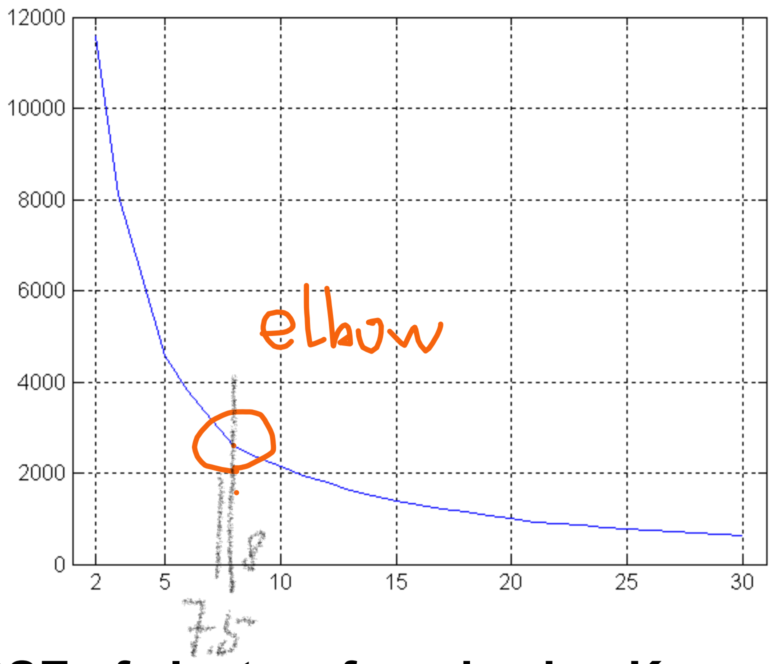


DETERMINING THE CORRECT NUMBER OF CLUSTERS

- SSE curve for a more complicated data set



K=8
K=7



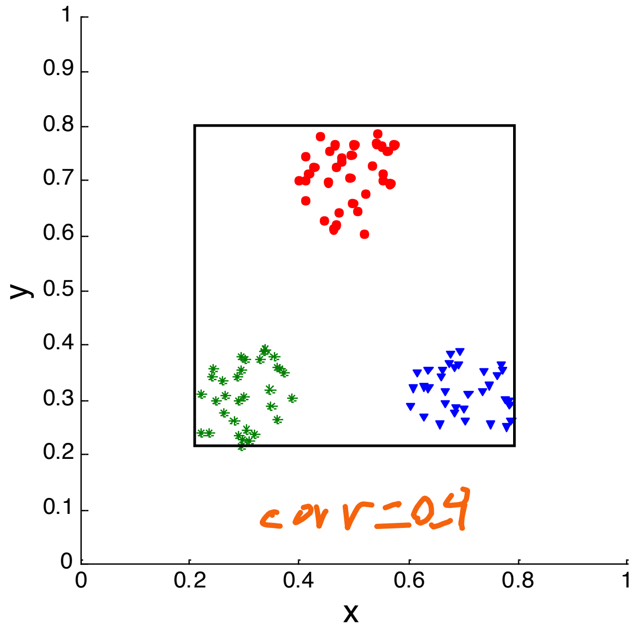
SSE of clusters found using K-means

ASSESSING THE SIGNIFICANCE OF CLUSTER VALIDITY MEASURES

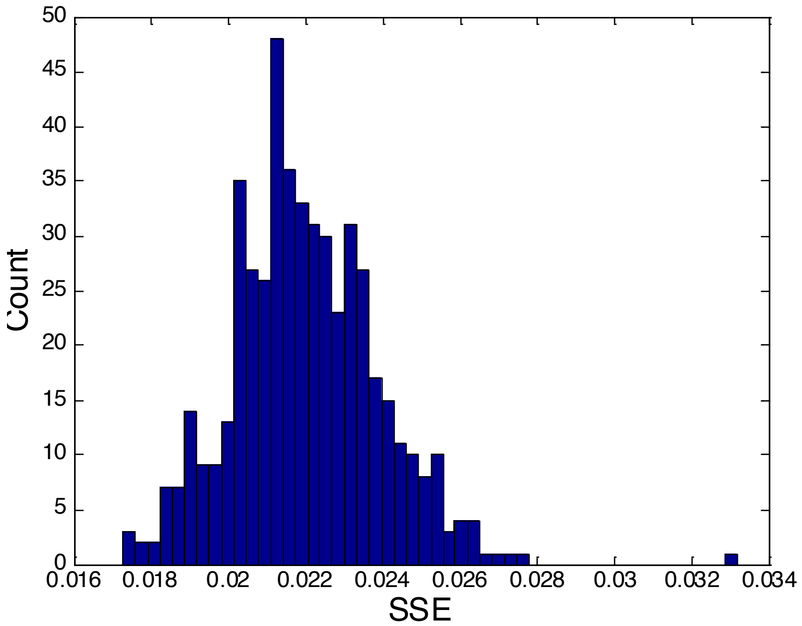
- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more "atypical" a clustering result is, the more likely it represents valid structure in the data
  - Compare the value of an index obtained from the given data with those resulting from random data.
    - If the value of the index is unlikely, then the cluster results are valid

STATISTICAL FRAMEWORK FOR SSE

- Example
  - Compare SSE of three cohesive clusters against three clusters in random data



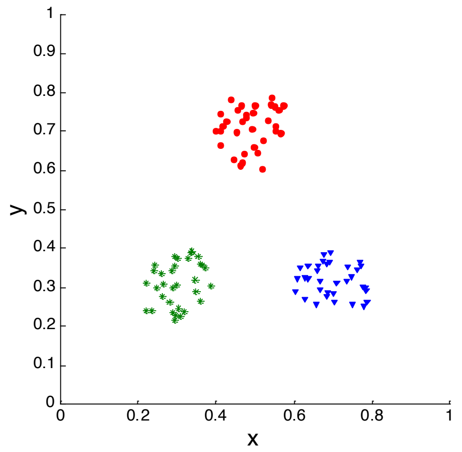
SSE = 0.005

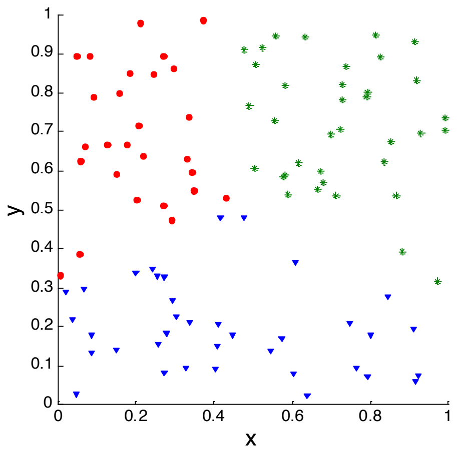


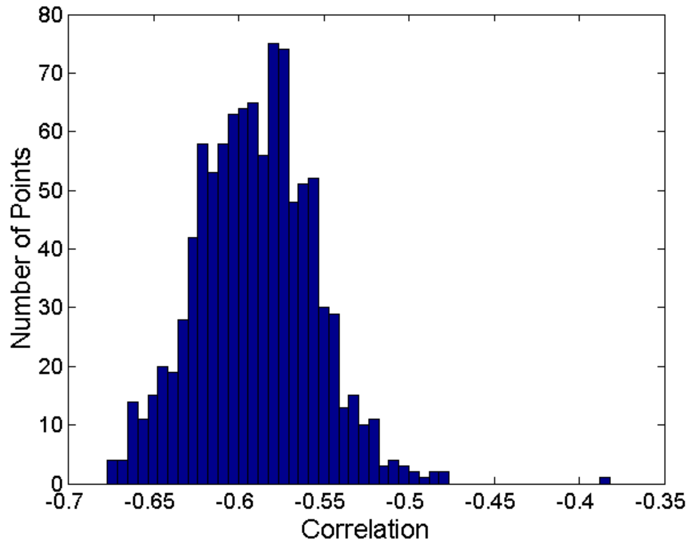
Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values

STATISTICAL FRAMEWORK FOR CORRELATION

Baseline







Corr = |-0.9235|

Corr = |-0.5810|

Correlation is negative because it is calculated between a distance matrix and the ideal similarity matrix. Higher magnitude is better.

Histogram of correlation for 500 random data sets of size 100 with *x* and *y* values of points between 0.2 and 0.8.

OTHER CLUSTER METHODS

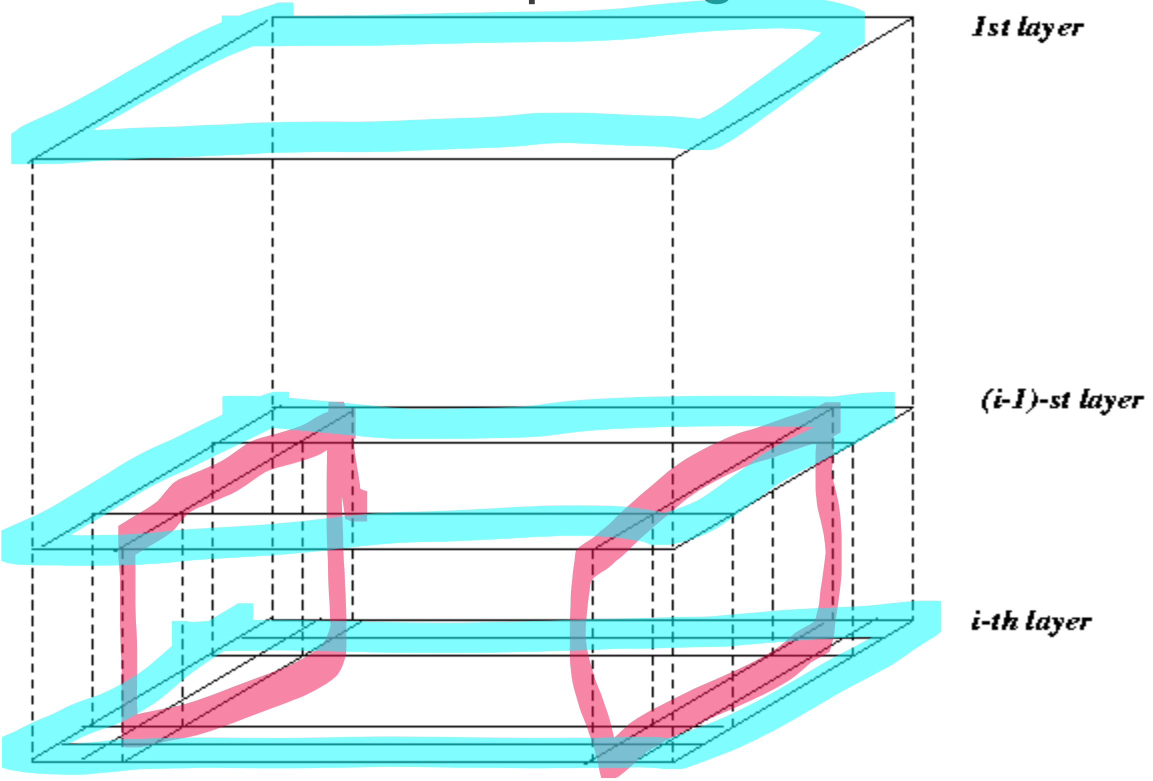
1. Partitioning Methods k-means
2. Hierarchical Methods single / min complete / max
3. Density-Based Methods DBSCAN
4. Grid-Based Methods
5. Model-Based Methods
6. Clustering High-Dimensional Data
7. Constraint-Based Clustering
8. Outlier Analysis

SUMMARY

- Clustering Algorithm
  - Partitioning-Based
    1. K-means
    2. K-medoids
    3. K-modes
    4. PAM
    5. CLARANS
    6. CLARA
    7. FCM
  - Hierarchical-Based
    1. BIRCH
    2. CURE
    3. ROCK
    4. Chameleon
    5. Echidna
  - Density-Based
    1. DBSCAN
    2. OPTICS
    3. DBCLASD
    4. DENCLUE
  - Grid-Based
    1. Wave-Cluster
    2. STING
    3. CLIQUE
    4. OptiGrid
  - Model-Based
    1. EM
    2. COBWEB
    3. CLASSIT
    4. SOMs

STING: A STATISTICAL INFORMATION GRID APPROACH

- Wang, Yang and Muntz (VLDB'97)
- The spatial area aea is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution

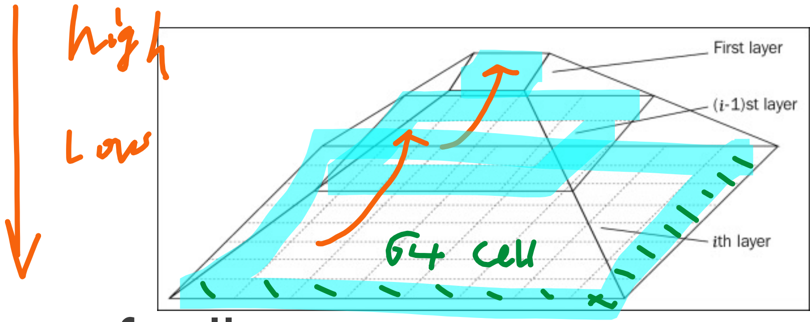


THE STING CLUSTERING METHOD

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
- Parameters of higher level cells can be easily calculated from parameters of lower level cell

  count, *mean*, *s*, *min*, *max*

  type of distribution—normal, *uniform*, etc.



- Use a top-down approach to answer spatial data queries
- Start from a pre-selected layer—typically with a small number of cells
- For each cell in the current level compute the confidence interval

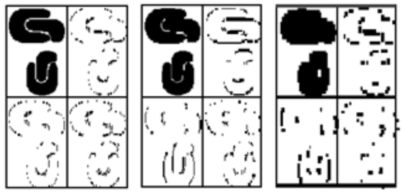
COMMENTS ON STING

- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level
- Repeat this process until the bottom layer is reached
- Advantages:
  - Query-independent, easy to parallelize, incremental update
  - $O(K)$, where K is the number of grid cells at the lowest level
- Disadvantages:
  - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

WAVE CLUSTER: CLUSTERING BY WAVELET ANALYSIS

- Sheikholeslami, Chatterjee, and Zhang
- A multi-resolution clustering approach which applies wavelet transform to the feature space
- How to apply wavelet transform to find clusters
  - Summarizes the data by imposing a multidimensional grid structure onto data space
  - These multidimensional spatial data objects are represented in a n-dimensional feature space
  - Apply wavelet transform on feature space to find the dense regions in the feature space
  - Apply wavelet transform multiple times which result in clusters at different scales from fine

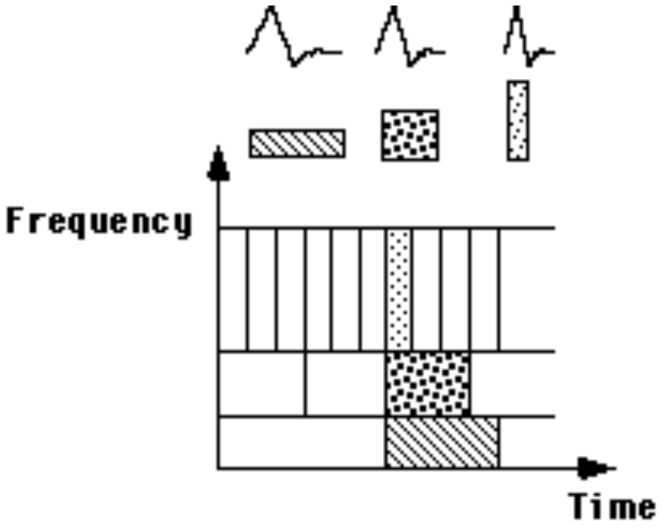
Figure 7.16 A sample of two-dimensional feature space. From [SCZ98].

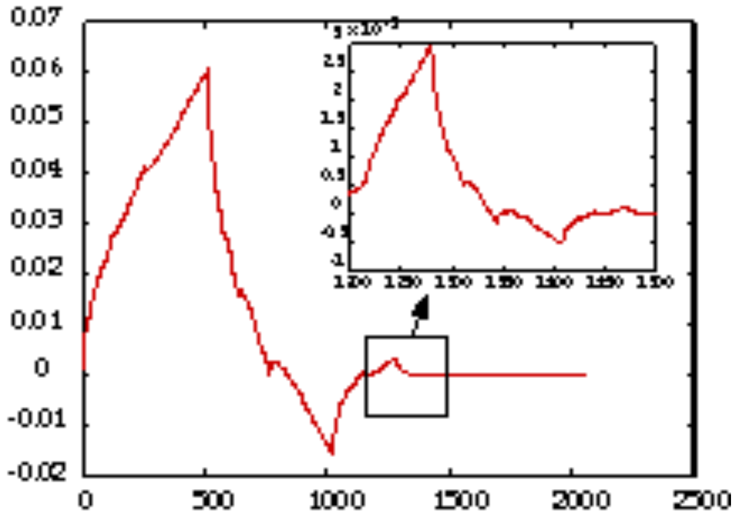


(a) (b) (c)

WAVELET TRANSFORM

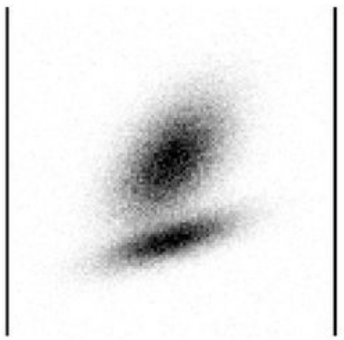
- Wavelet transform: A signal processing technique that decomposes a signal into different frequency sub-band (can be applied to n-dimensional signals)
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allows natural clusters to become more distinguishable



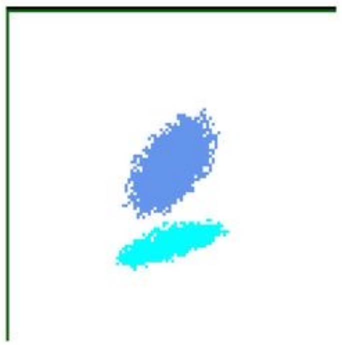


THE WAVECLUSTER ALGORITHM

- Input parameters
  - # of grid cells for each dimension
  - the wavelet, and the # of applications of wavelet transform
- Why is wavelet transformation useful for clustering?
  - Use hat-shape filters to emphasize region where points cluster, but simultaneously suppress weaker information in their boundary
  - Effective removal of outliers, multi-resolution, cost effective
- Major features:
  - Complexity O(N)
  - Detect arbitrary shaped clusters at different scales
  - Not sensitive to noise, not sensitive to input order
  - Only applicable to low dimensional data
- Both grid-based and density-based



a)



b)

QUANTIZATION & TRANSFORMATION

- First, quantize data into m-D grid structure, then wavelet transform
  - a) scale 1: high resolution
  - b) scale 2: medium resolution
  - c) scale 3: low resolution

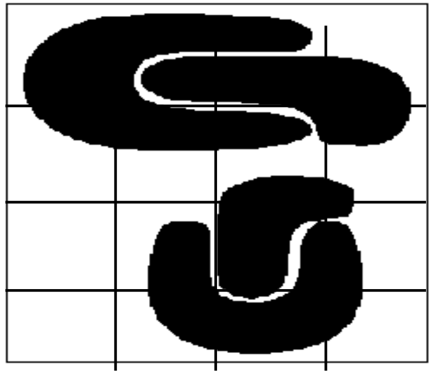


Figure 1: A sample 2-dimensional feature space.

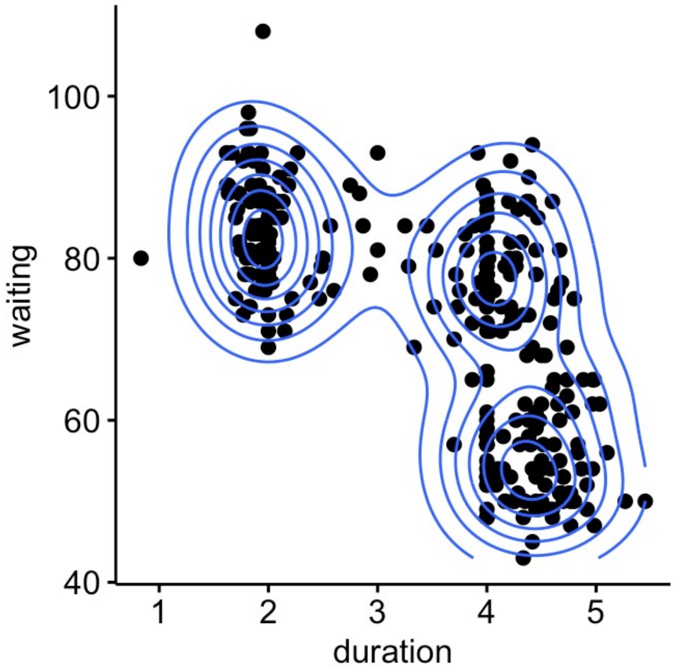
a) b) c)

OTHER CLUSTER METHODS

1. Partitioning Methods (k-means)
2. Hierarchical Methods
3. Density-Based Methods (non-elliptical shape)
4. Grid-Based Methods
5. Model-Based Methods
6. Clustering High-Dimensional Data
7. Constraint-Based Clustering
8. Outlier Analysis

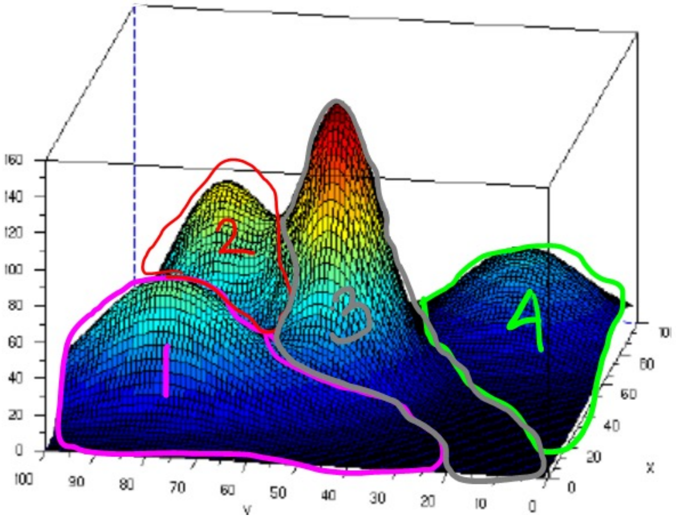
MODEL-BASED CLUSTERING

- What is model-based clustering?
  - Attempt to optimize the fit between the given data and some mathematical model
  - Based on the assumption: Data are generated by a mixture of underlying probability distribution
- Typical methods
  - Statistical approach
    - EM (Expectation maximization), AutoClass
  - Machine learning approach
    - COBWEB, CLASSIT
  - Neural network approach
    - SOM (Self-Organizing Feature Map)

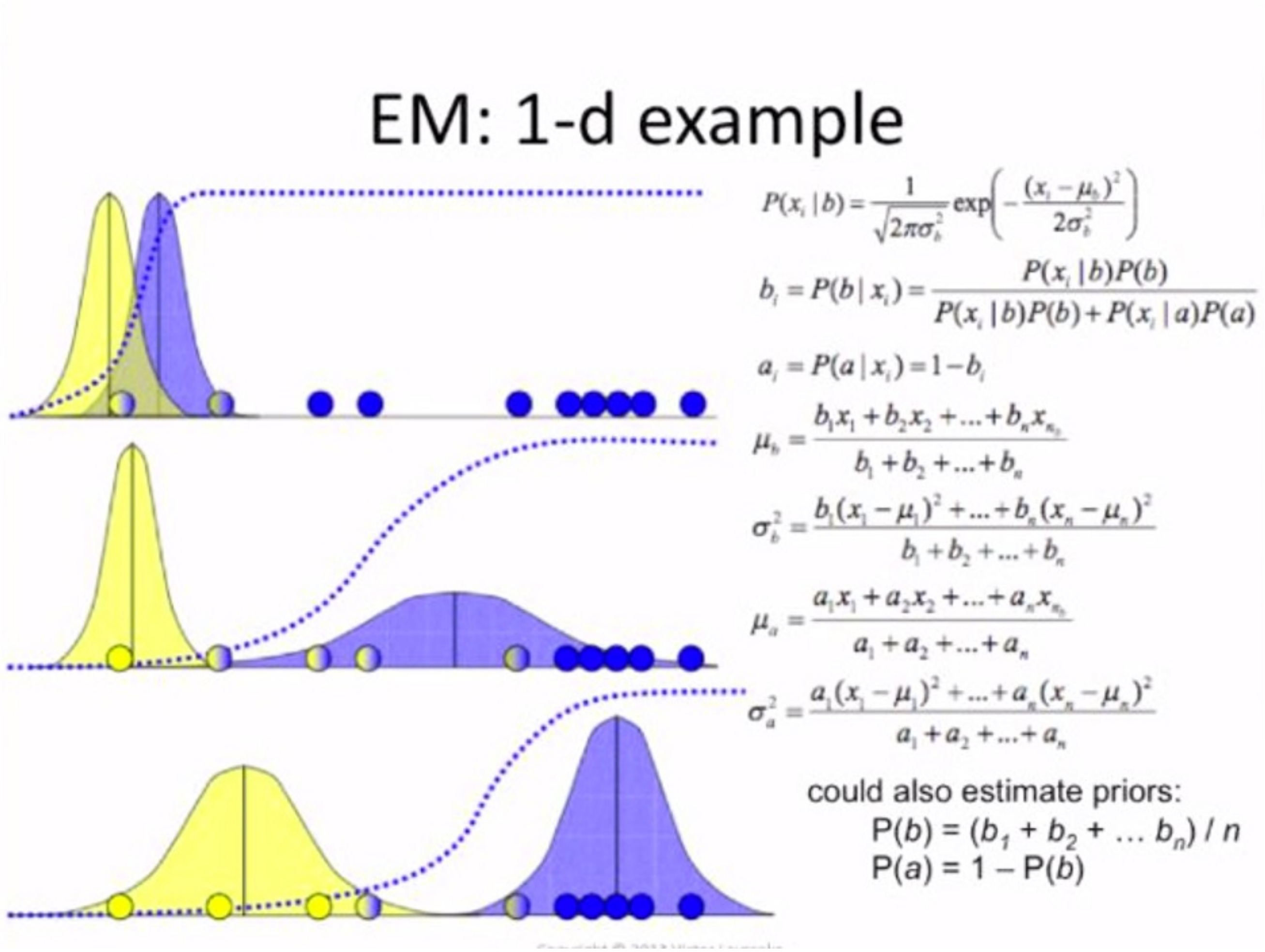


EM — EXPECTATION MAXIMIZATION

- EM — A popular iterative refinement algorithm
- An extension to k-means
  - Assign each object to a cluster according to a weight (prob. distribution)
  - New means are computed based on weighted measures
- General idea
  - Starts with an initial estimate of the parameter vector
  - Iteratively rescores the patterns against the mixture density produced by the parameter vector
  - The rescored patterns are used to update the parameter updates
  - Patterns belonging to the same cluster, if they are placed by their scores in a particular component
- Algorithm converges fast but may not be in global optima



EM — EXPECTATION MAXIMIZATION



For each data point, EM calculates a vector of Probabilities.

Each probability will refer to each cluster.

Group the point to the cluster.

THE EM (EXPECTATION MAXIMIZATION) ALGORITHM

- Initially, randomly assign k cluster centers
- Iteratively refine the clusters based on two steps
  - Expectation step: assign each data point X_i to cluster C_i with the following probability

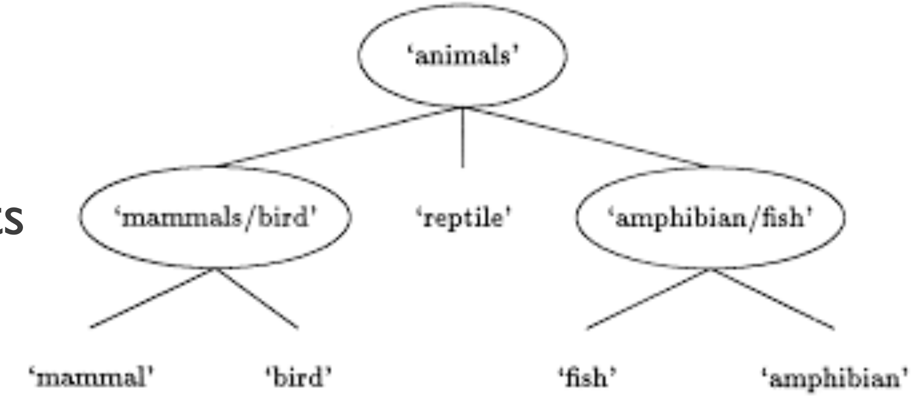
$$P(X_i \in C_k) = p(C_k|X_i) = \frac{p(C_k)p(X_i|C_k)}{p(X_i)},$$

  - Maximization step:
    - Estimation of model parameters

$$m_k = \frac{1}{N}\Sigma_{i=1}^{N}\frac{X_i P(X_i \in C_k)}{\Sigma_j P(X_i \in C_j)}.$$

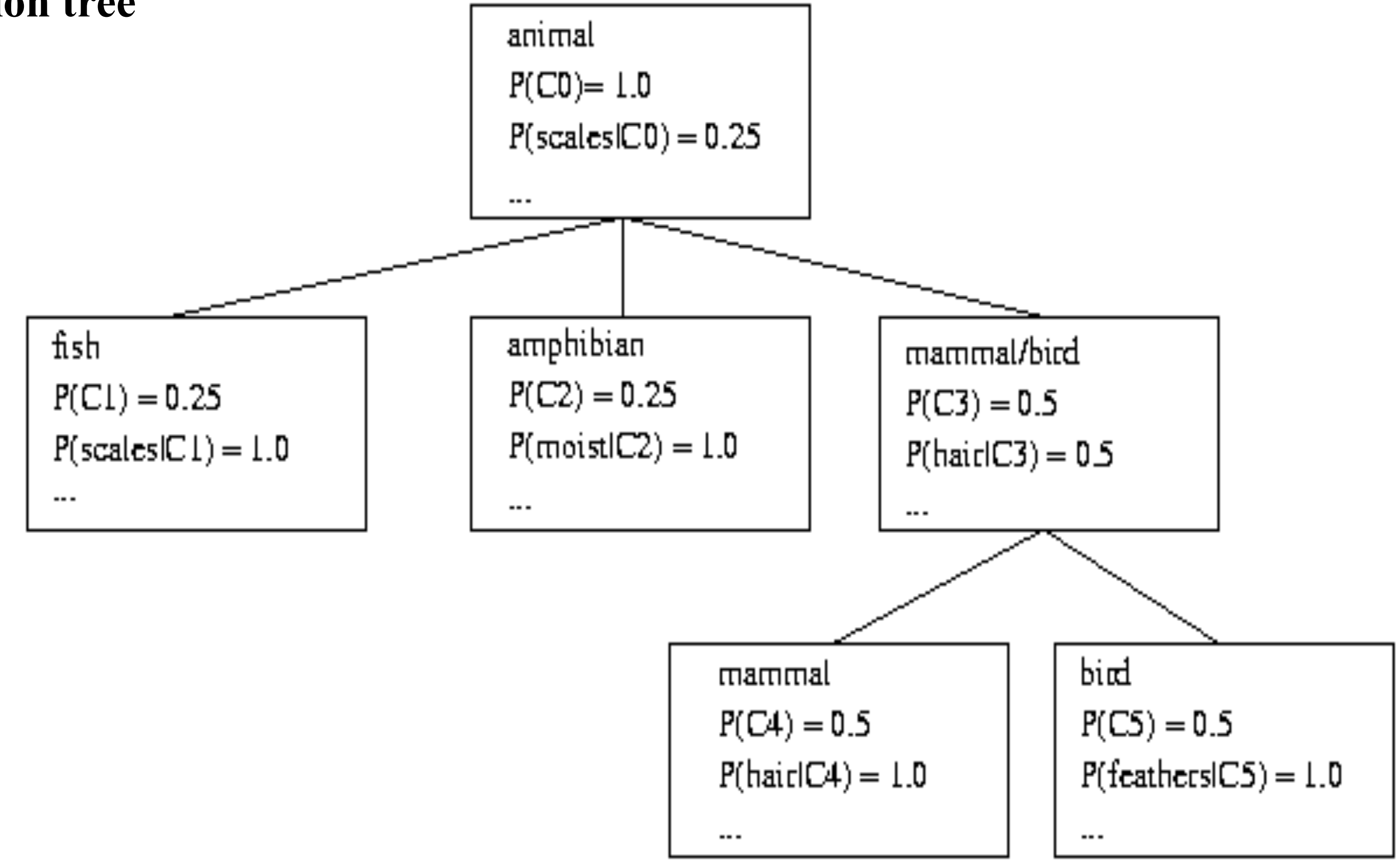
CONCEPTUAL CLUSTERING

- Conceptual clustering
  - A form of clustering in machine learning
  - Produces a classification scheme for a set of unlabeled objects
  - Finds characteristic description for each concept (class)
- COBWEB
  - A popular a simple method of incremental conceptual learning
  - Creates a hierarchical clustering in the form of a classification tree
  - Each node refers to a concept and contains a probabilistic description of that concept



COBWEB CLUSTERING METHOD

A classification tree



MORE ON CONCEPTUAL CLUSTERING

- Limitations of COBWEB
  - The assumption that the attributes are independent of each other is often too strong because correlation may exist
  - Not suitable for clustering large database data – skewed tree and expensive probability distributions
- CLASSIT
  - an extension of COBWEB for incremental clustering of continuous data
  - suffers similar problems as COBWEB
- AutoClass (Cheeseman and Stutz, 1996)
  - Uses Bayesian statistical analysis to estimate the number of clusters
  - Popular in industry

NEURAL NETWORK APPROACH

- Neural network approaches
  - Represent each cluster as an exemplar, acting as a "prototype" of the cluster
  - New objects are distributed to the cluster whose exemplar is the most similar according to some distance measure

large features + small samples

over-fitting

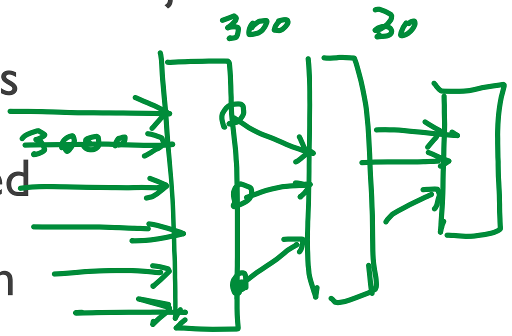
- Typical methods
  - SOM (Soft-Organizing feature Map)
  - Competitive learning
    - Involves a hierarchical architecture of several units (neurons)
    - Neurons compete in a "winner-takes-all" fashion for the object currently being presented

high dimen ⇒ low features

thousand genes

SELF-ORGANIZING FEATURE MAP (SOM)

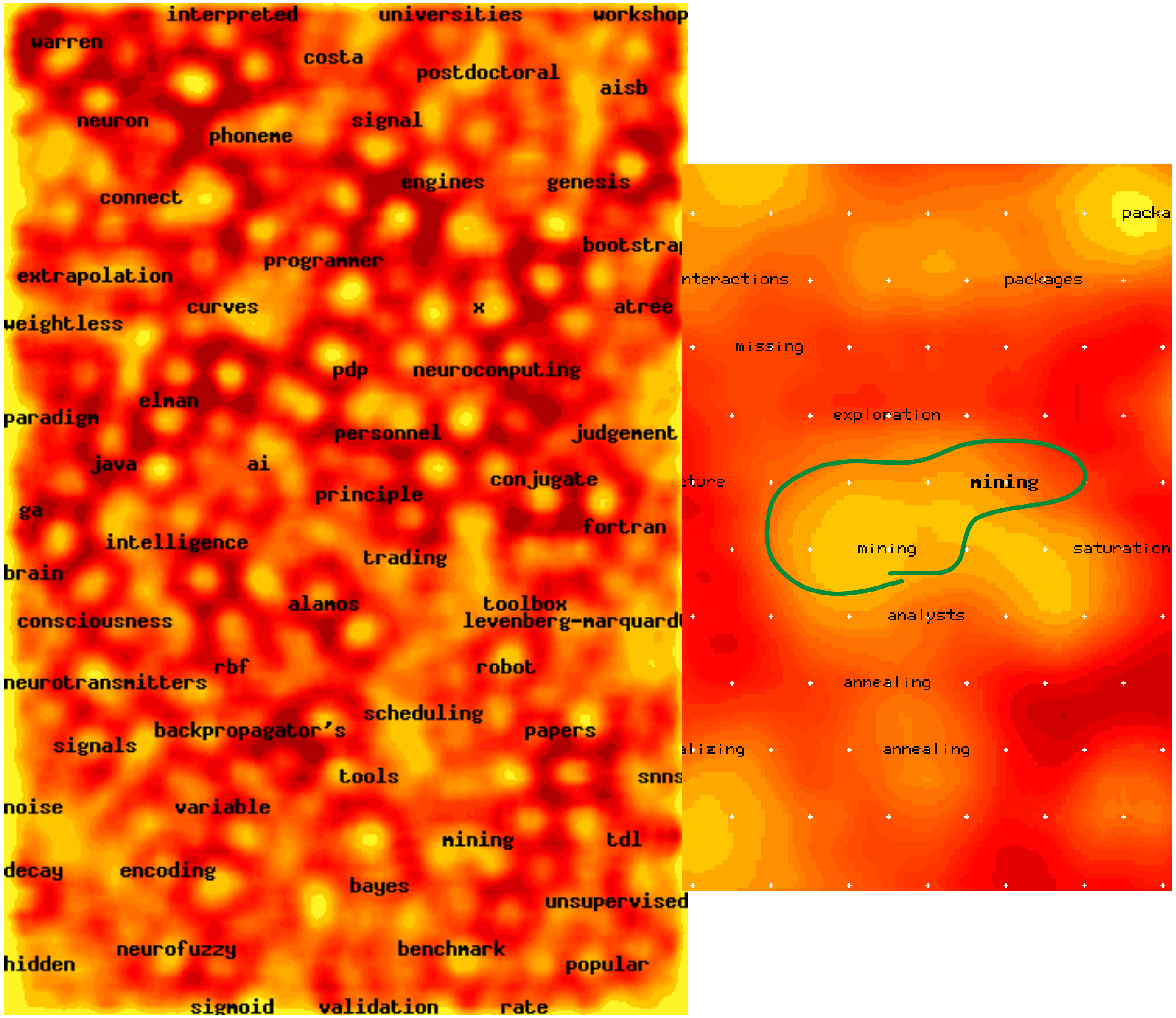
- SOMs, also called topological ordered maps, or Kohonen Self-Organizing Feature Map (KSOMs)
- It maps all the points in a high-dimensional source space into a 2 to 3-d target space, s.t., the distance and proximity relationship (i.e., topology) are preserved as much as possible
- Similar to k-means: cluster centers tend to lie in a low-dimensional manifold in the feature space
- Clustering is performed by having several units competing for the current object
  - The unit whose weight vector is closest to the current object wins
  - The winner and its neighbors learn by having their weights adjusted
- SOMs are believed to resemble processing that can occur in the brain
- Useful for visualizing high-dimensional data in 2- or 3-D space



WEB DOCUMENT CLUSTERING USING SOM

- The result of SOM clustering of 12088 Web articles (www)
- The picture on the right: drilling down on the keyword "mining"
- Based on websom.hut.fi Web page

tensorflow



SUMMARY

Clustering Algorithm

Partitioning-Based

1. K-means
2. K-medoids
3. K-modes
4. PAM
5. CLARANS
6. CLARA
7. FCM

Hierarchical-Based

tree ⟷ graph

1. BIRCH
2. CURE
3. ROCK
4. Chameleon
5. Echidna

Density-Based

1. DBSCAN
2. OPTICS
3. DBCLASD
4. DENCLUE

Grid-Based

1. Wave-Cluster
2. STING
3. CLIQUE
4. OptiGrid

Vis

Model-Based

prob distri

1. EM
2. COBWEB (prob)
3. CLASSIT
4. SOMs

traffic Math