Week 10

Clustering (Part II)

Seokho Chi
Assistant Professor I Ph.D.
SNU Construction Innovation Lab

Source: Tan, Kumar, Steinback (2006)

Grid-Based Clustering

- Efficient way to organize a set of data, at least in low dimensions
- Algorithm:
  1. Define a set of grid cells
  2. Assign objects to the appropriate cells and compute the density of each cell
  3. Eliminate cells having a density below a specified threshold
  4. Form clusters from contiguous (adjacent groups of dense cells

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

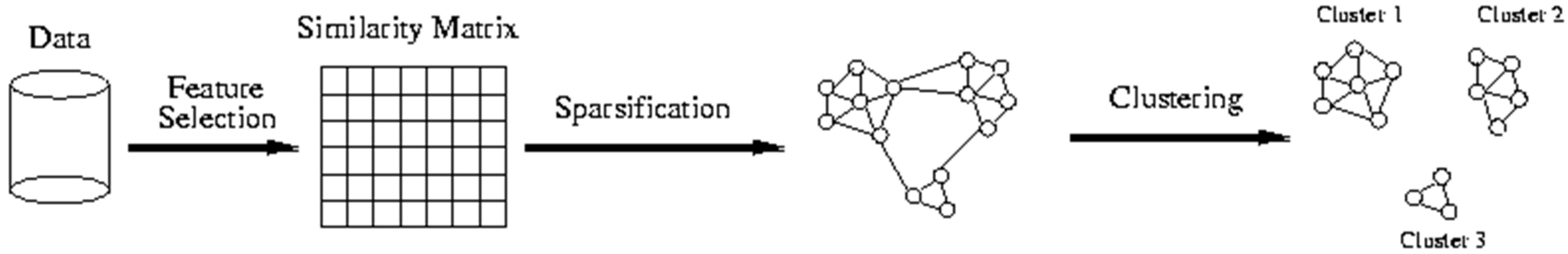
Graph-Based Clustering

- Graph-Based clustering uses the proximity graph
  – Start with the proximity matrix
  – Consider each point as a node in a graph
  – Each edge between two nodes has a weight which is the proximity between the two points
  – Initially the proximity graph is fully connected

- In the simplest case, clusters are connected components in the graph.

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Sparsification in the Clustering Process



Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Graph-Based Clustering: Sparsification

- The amount of data that needs to be processed is drastically reduced
  - Setting many of low similarity values to 0
  - Break all links that have a similarity below a specified threshold or keep only links to the k nearest neighbors of point
  - Sparsification can eliminate more than 99% of the entries in a proximity matrix
  - The amount of time required to cluster the data is drastically reduced
  - The size of the problems that can be handled is increased

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Graph-Based Clustering: Sparsification …

- Clustering may work better
  - Sparsification techniques keep the connections to the most similar (nearest) neighbors of a point while breaking the connections to less similar points.
  - The nearest neighbors of a point tend to belong to the same class as the point itself.
  - This reduces the impact of noise and outliers and sharpens the distinction between clusters.

- Sparsification facilitates the use of graph partitioning algorithms (or algorithms based on graph partitioning algorithms).
  - Chameleon and Hypergraph-based Clustering

Adapted from:

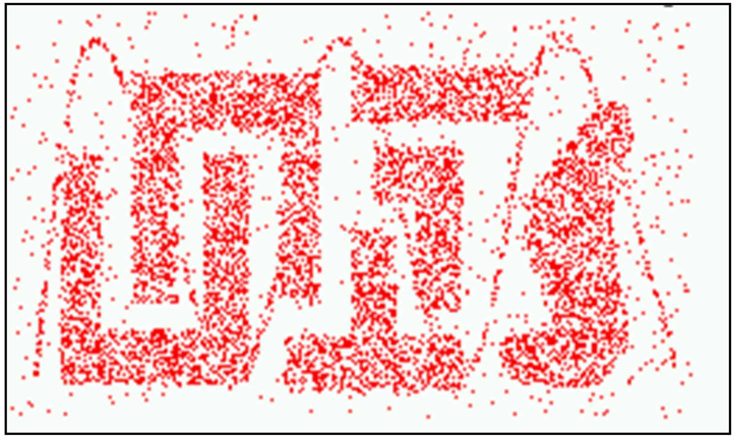
Tan,Steinbach, Kumar - Introduction to Data Mining

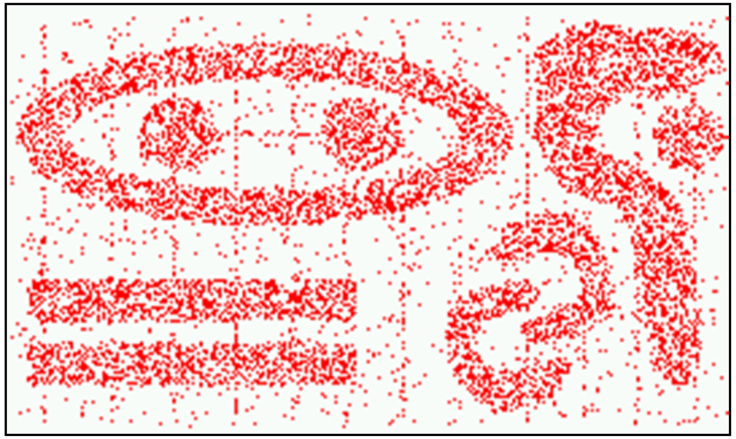
Han, Kamber - Data Mining: Concepts and Techniques

Characteristics of Spatial Data Sets

- **Clusters are defined as densely populated regions of the space**
- **Clusters have arbitrary shapes, orientation, and non-uniform sizes**
- **Difference in densities across clusters and variation in density within clusters**
- **Existence of special artifacts (*streaks*) and noise**

The clustering algorithm must address the above characteristics and also require minimal supervision.





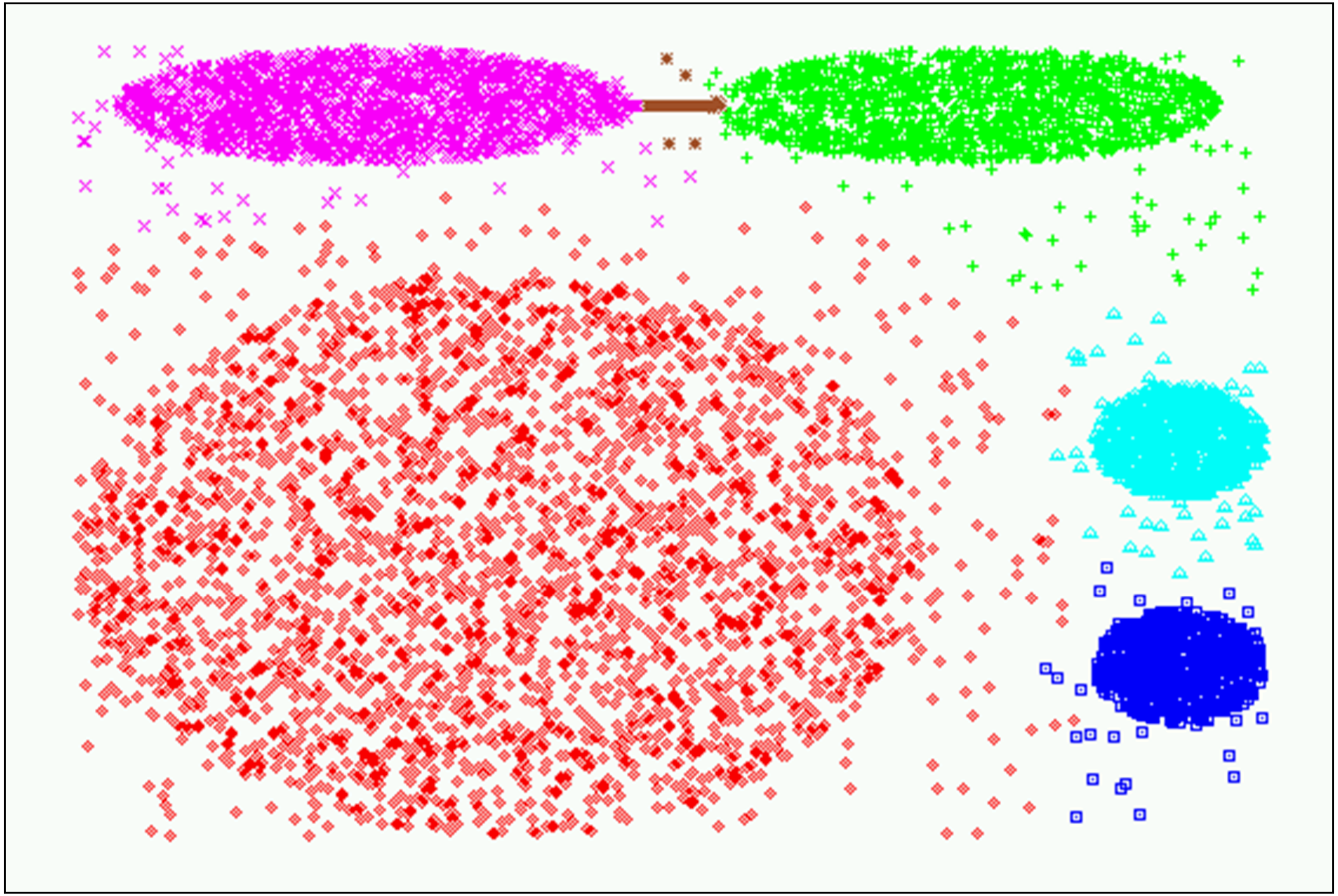
Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Experimental Results: **CHAMELEON**

An agglomerative clustering algorithm with dynamic modeling using an efficient graph partitioning algorithm (considers density, shapes, closeness, interconnectivity, etc.)

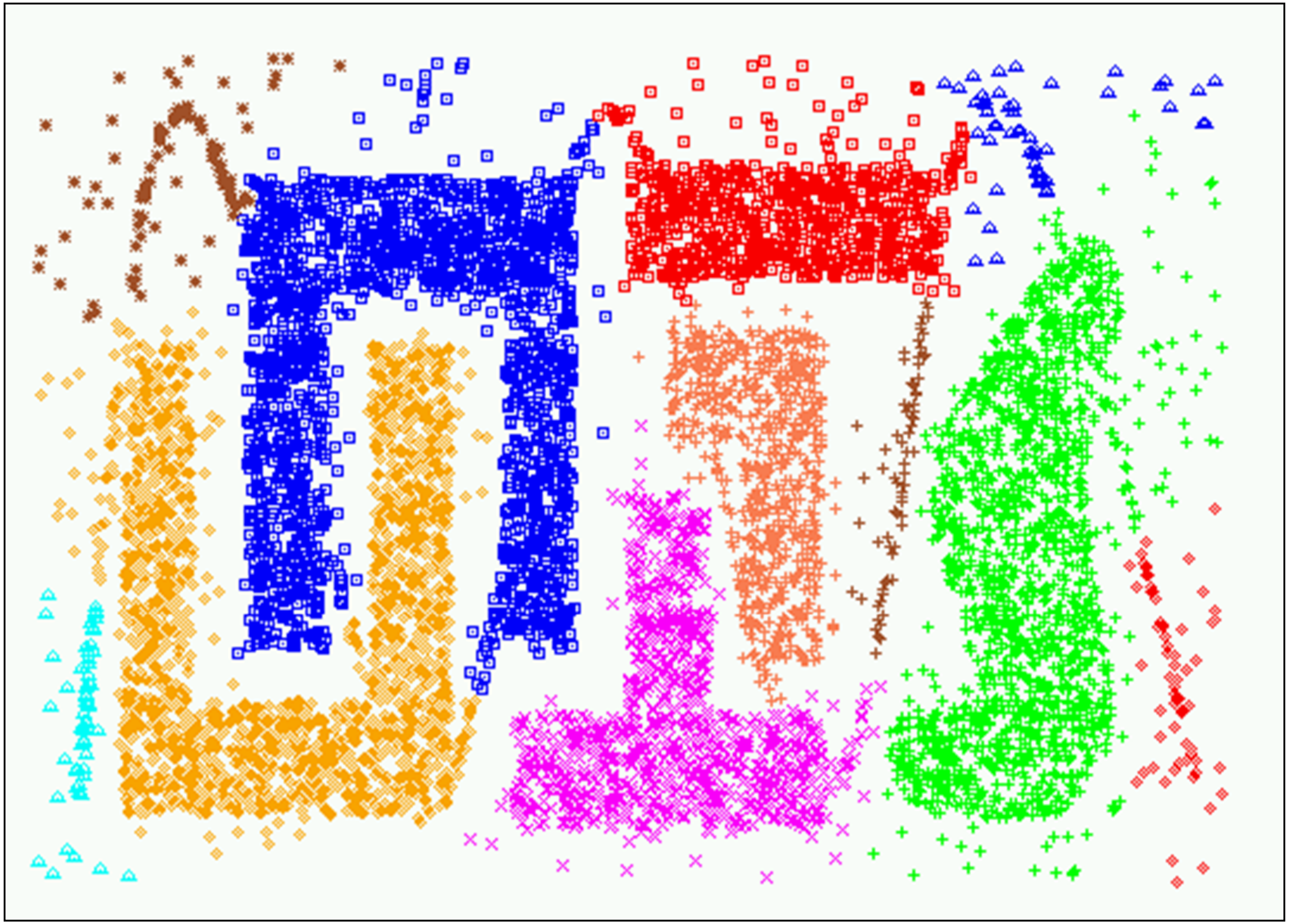


Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Experimental Results: **CHAMELEON**



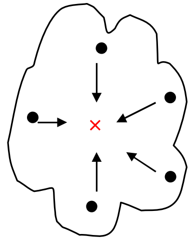
Adapted from:

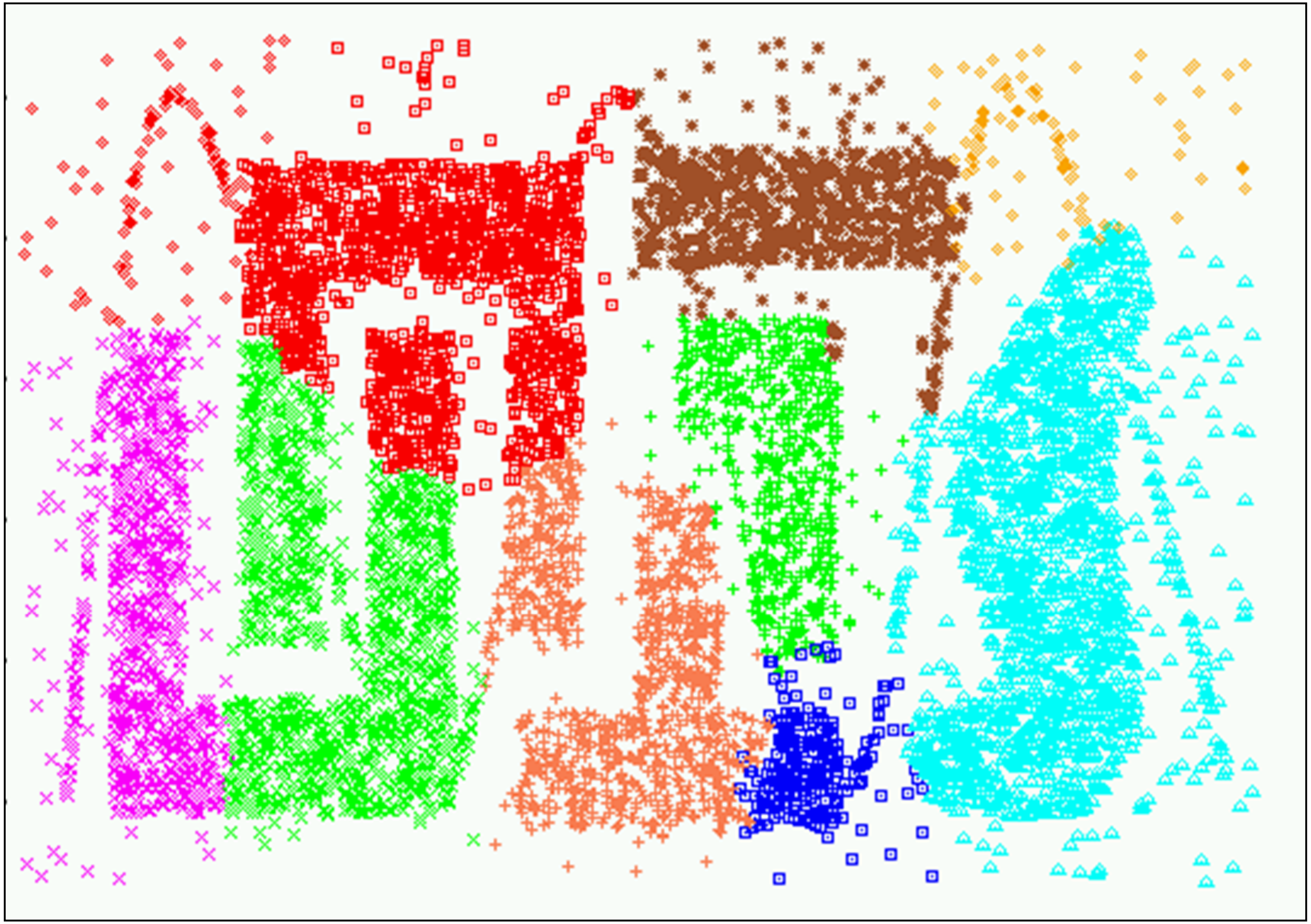
Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Experimental Results: **CURE (*10 clusters*)**

Clustering Using REpresentatives: (1) Equally partitions, (2) uses a constant number of points to represent a cluster which capture the geometry and shape of the cluster and (3) then shrinks them toward the center of the cluster by a factor, alpha



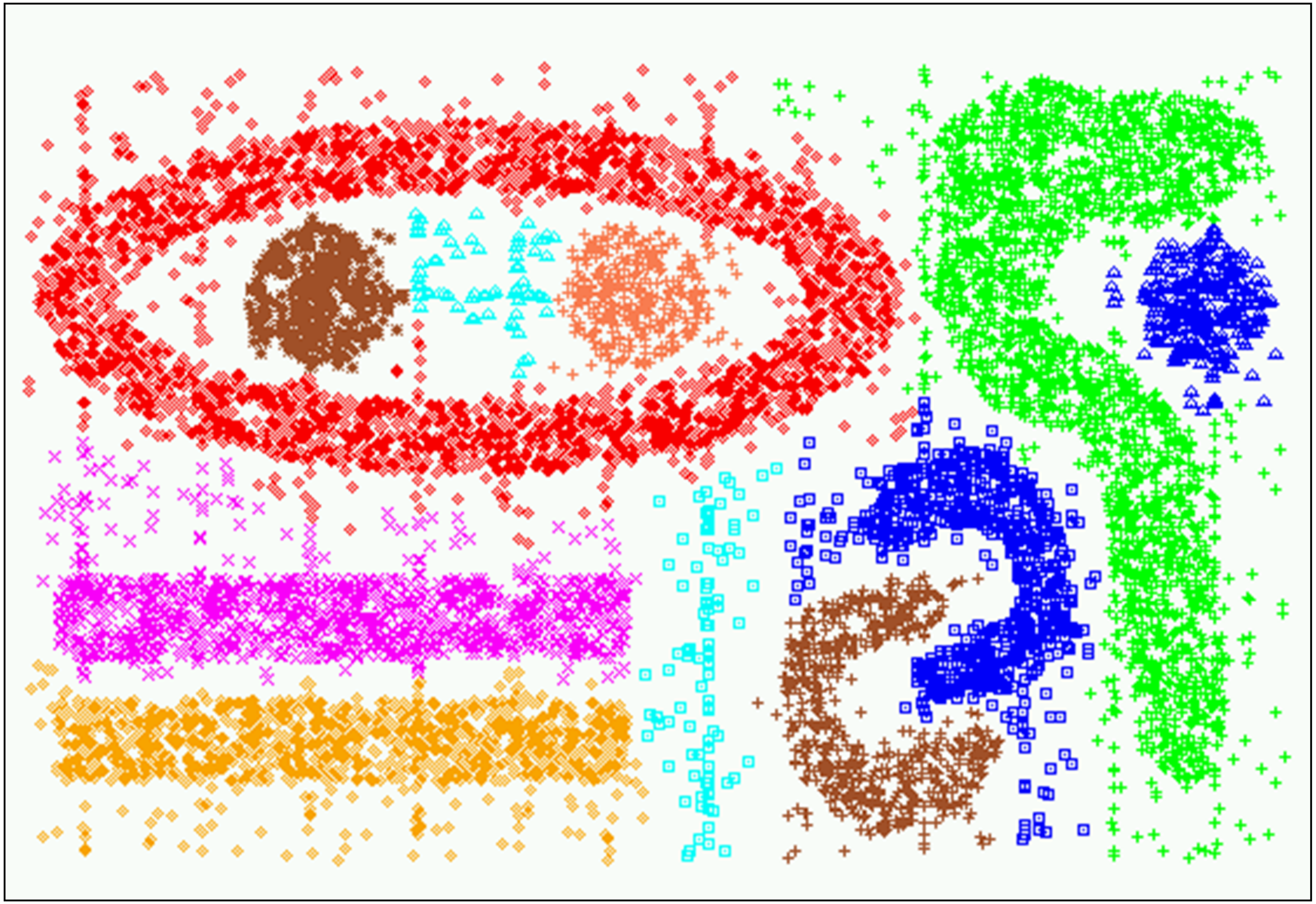


Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Experimental Results: **CHAMELEON**

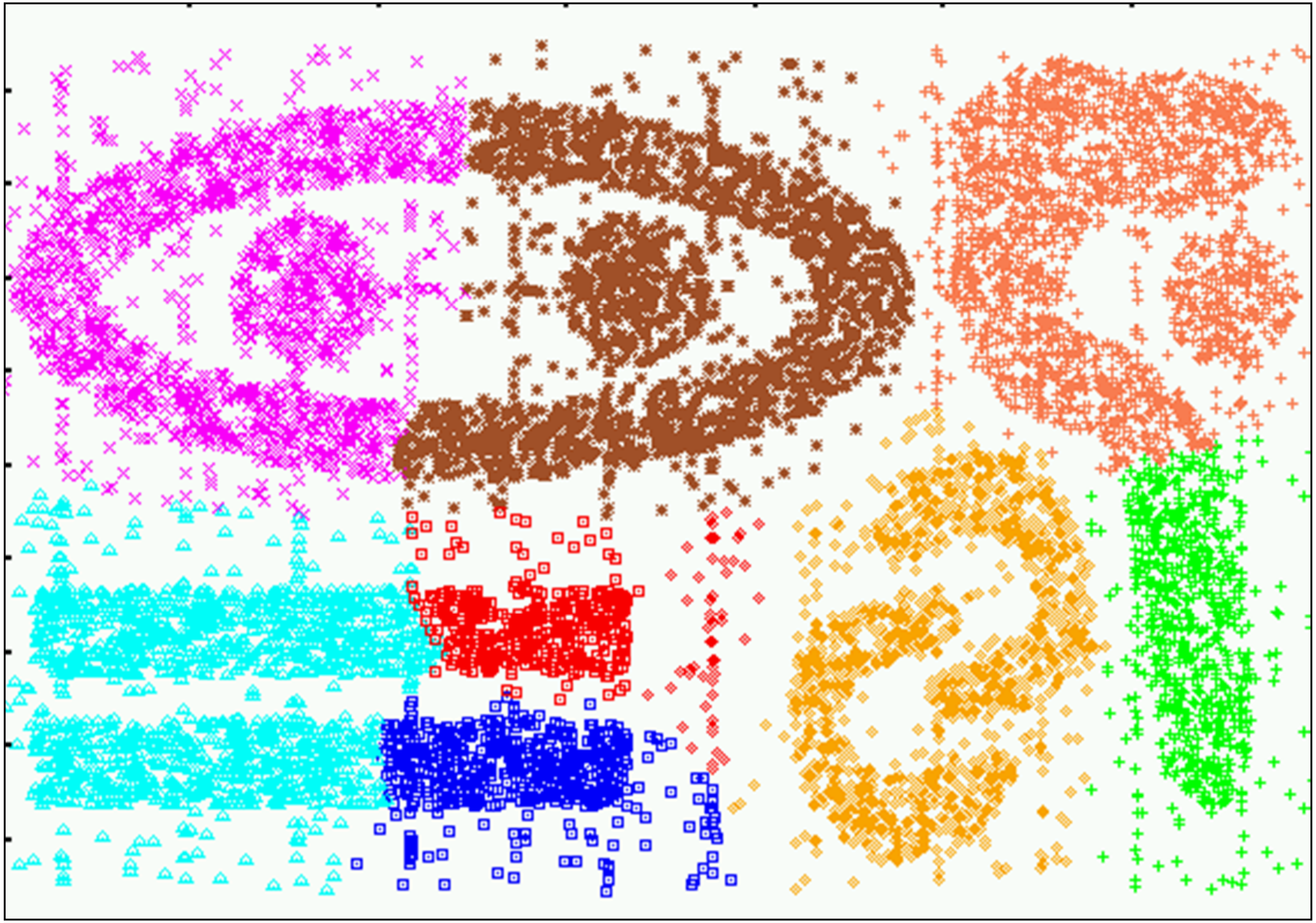


Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Experimental Results: **CURE (*9 clusters*)**



Adapted from:

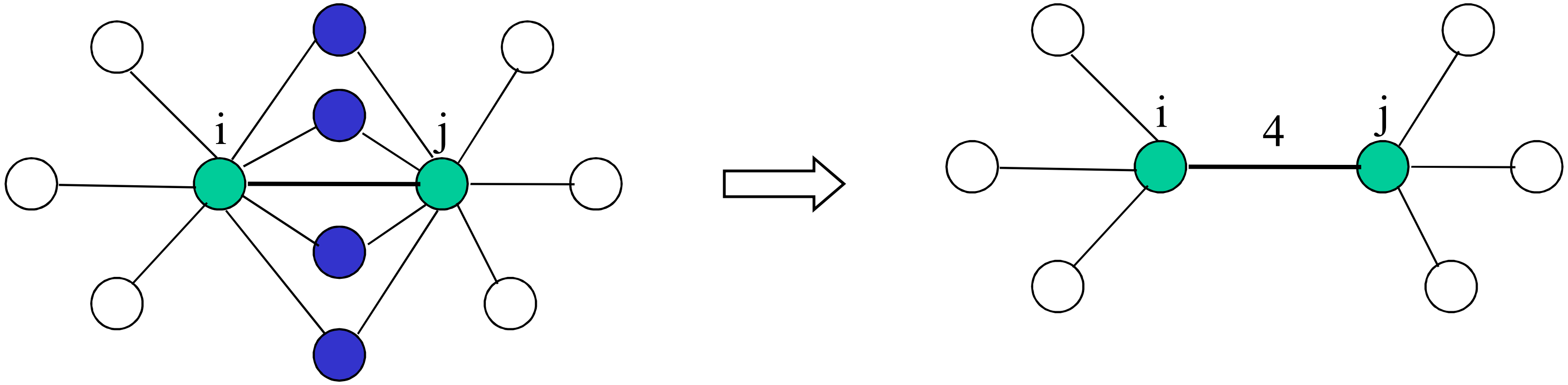
Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Shared Nearest Neighbor (SNN) Similarity

If two points are similar to many of the same points, then they are similar to one another, even if a direct measurement of similarity does not indicate this.

SNN graph: the weight of an edge is the number of shared neighbors between vertices given that the vertices are connected



Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

SNN Clustering Algorithm

1. **Compute the similarity matrix**
   This corresponds to a similarity graph with data points for nodes and edges whose weights are the similarities between data points
2. **Sparsify the similarity matrix by keeping only the *k* most similar neighbors**
   This corresponds to only keeping the *k* strongest links of the similarity graph
3. **Construct the shared nearest neighbor graph from the sparsified similarity matrix.**
   At this point, we could apply a similarity threshold and find the connected components to obtain the clusters (Jarvis-Patrick algorithm)
4. **Find the SNN density of each Point.**
   Using a user specified parameters, *Eps*, find the number points that have an SNN similarity of *Eps* or greater to each point. This is the SNN density of the point

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

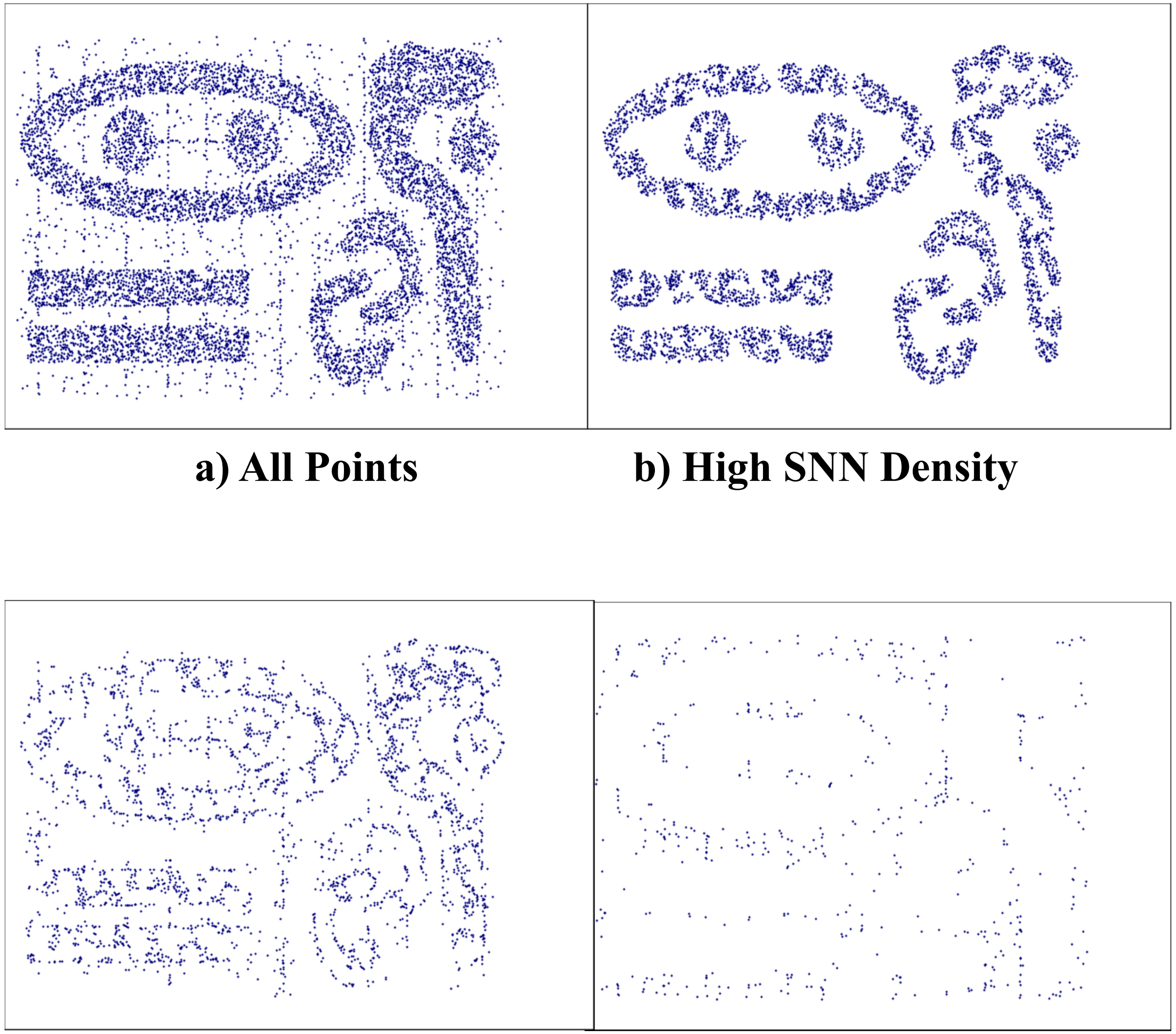
Han, Kamber - Data Mining: Concepts and Techniques

SNN Clustering Algorithm …

5. **Find the core points**
   Using a user specified parameter, *MinPts*, find the core points, i.e., all points that have an SNN density greater than *MinPts*
6. **Form clusters from the core points**
   If two core points are within a radius, *Eps*, of each other they are place in the same cluster
7. **Discard all noise points**
   All non-core points that are not within a radius of *Eps* of a core point are discarded
8. **Assign all non-noise, non-core points to clusters**
   This can be done by assigning such points to the nearest core point

   (Note that steps 4-8 are DBSCAN)

SNN Density



a) All Points b) High SNN Density

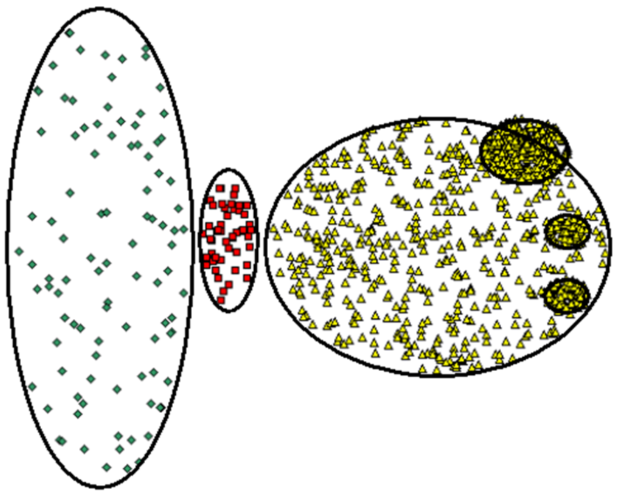
c) Medium SNN Density d) Low SNN Density

Adapted from:

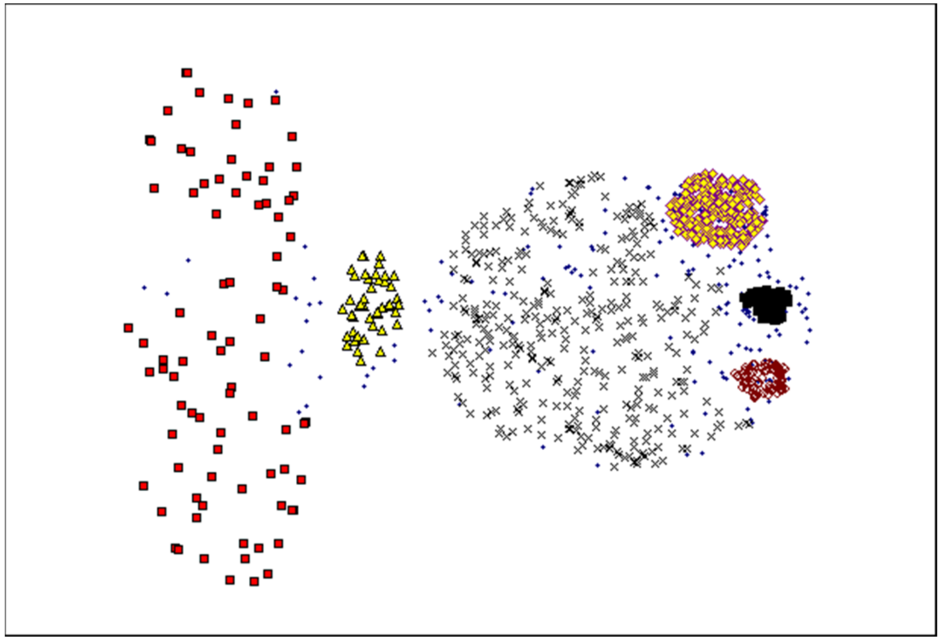
Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

SNN Clustering Can Handle Differing Densities



Original Points



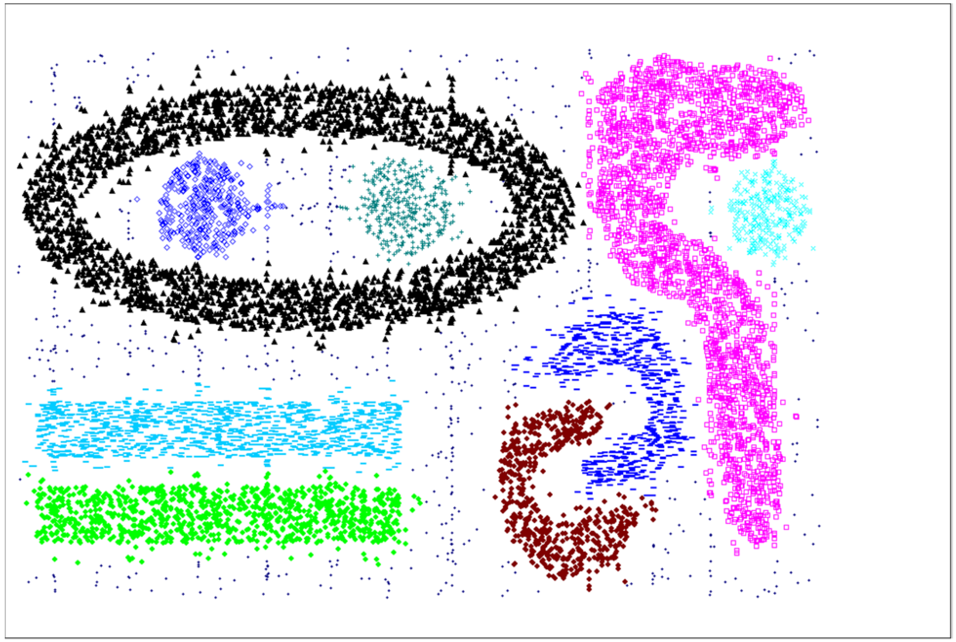
SNN Clustering

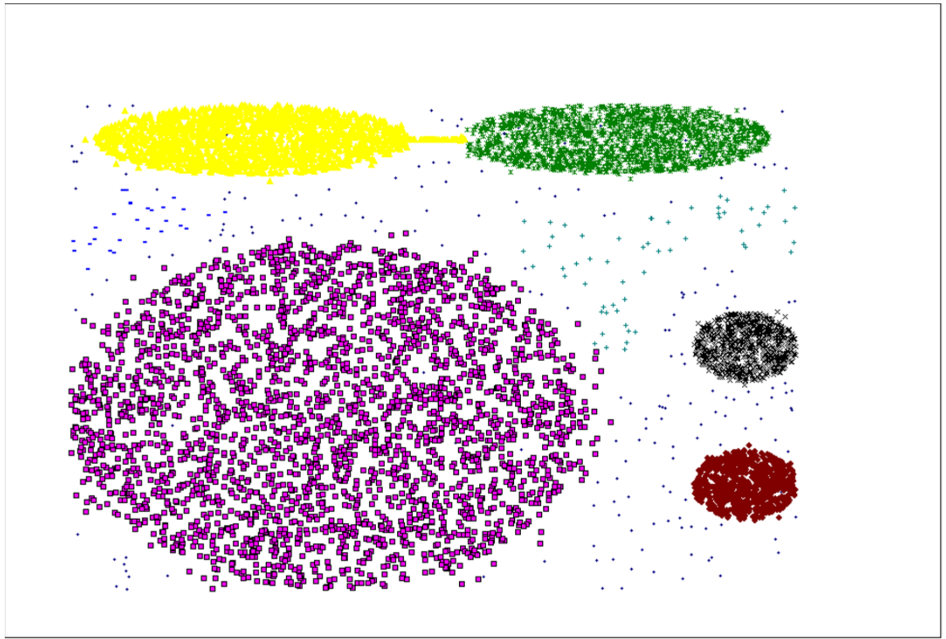
Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

SNN Clustering Can Handle Other Difficult Situations



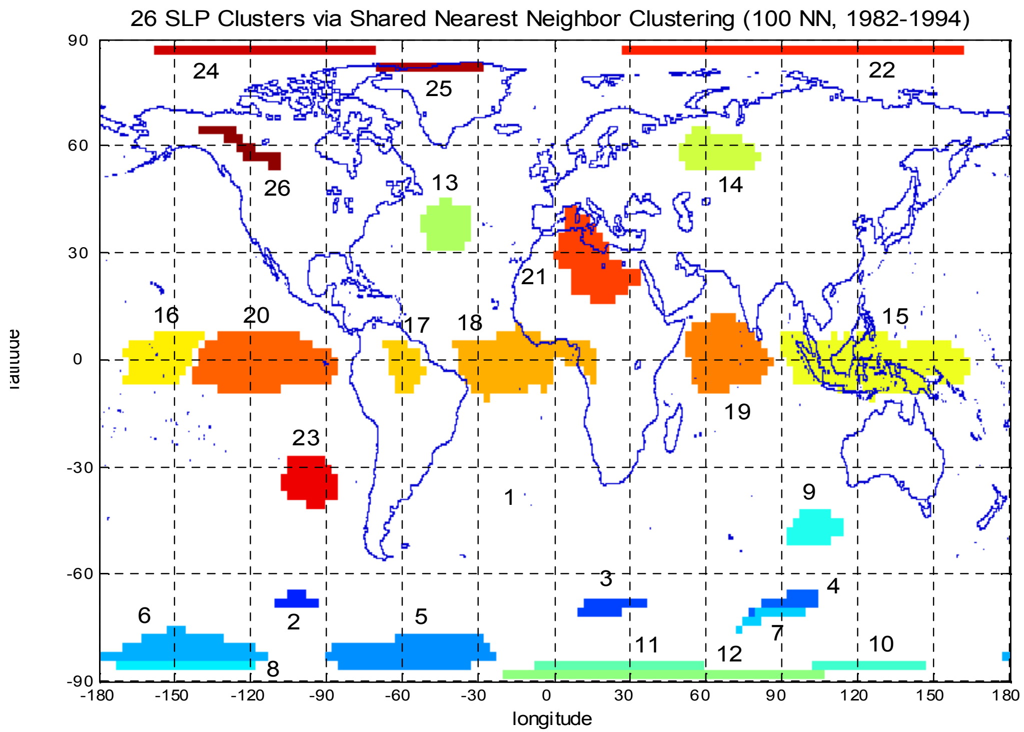


Adapted from:

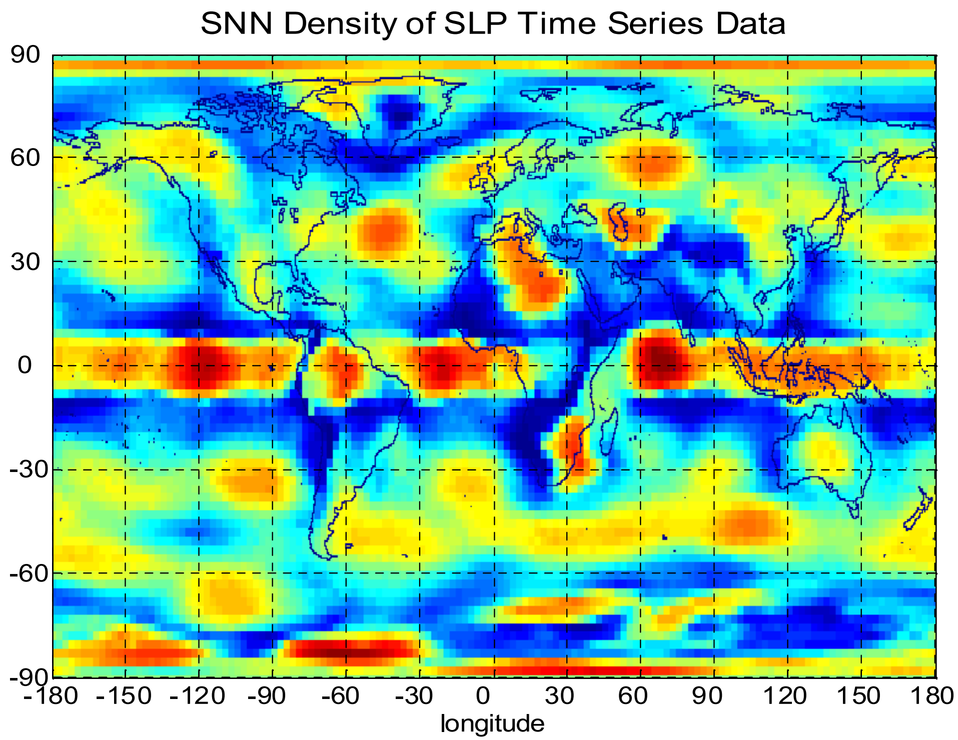
Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Finding Clusters of Time Series In Spatio-Temporal Data



SNN Clusters of Pressure



SNN Density of Pressure

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques