More Clustering Algorithms

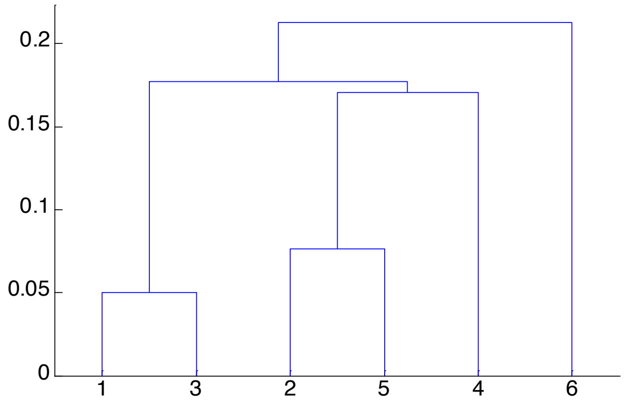
Blink Sakulkueakulsuk

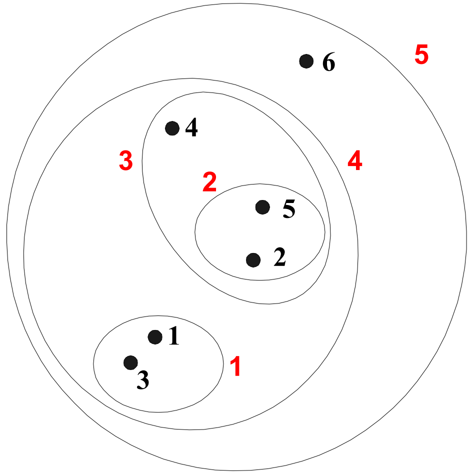
Hierarchical Clustering

Produces a set of nested clusters organized as a hierarchical tree

Can be visualized as a dendrogram

- A tree like diagram that records the sequences of merges or splits





Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Strengths of Hierarchical Clustering

Do not have to assume any particular numbers of clusters

- Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level

They may correspond to meaningful taxonomies

- Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Hierarchical Clustering

Two main types of hierarchical clustering

- Agglomerative:
  - Start with the points as individual clusters
  - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left

- Divisive:
  - Start with one, all-inclusive cluster
  - At each step, split a cluster until each cluster contains an individual point (or there are k clusters)

Traditional hierarchical algorithms use a similarity or distance matrix

- Merge or split one cluster at a time

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Agglomerative Clustering Algorithm

Key Idea: Successively merge closest clusters

Basic algorithm

1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

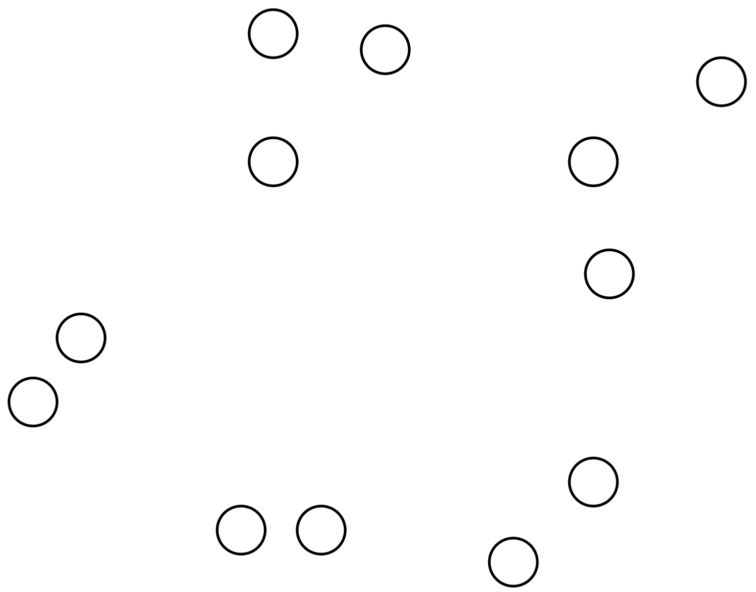
Key operation is the computation of the proximity of two clusters

- Different approaches to defining the distance between clusters distinguish the different algorithms

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

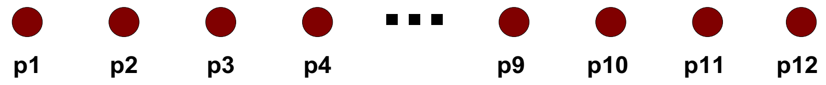
Steps 1 and 2

Start with clusters of individual points and a proximity matrix



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

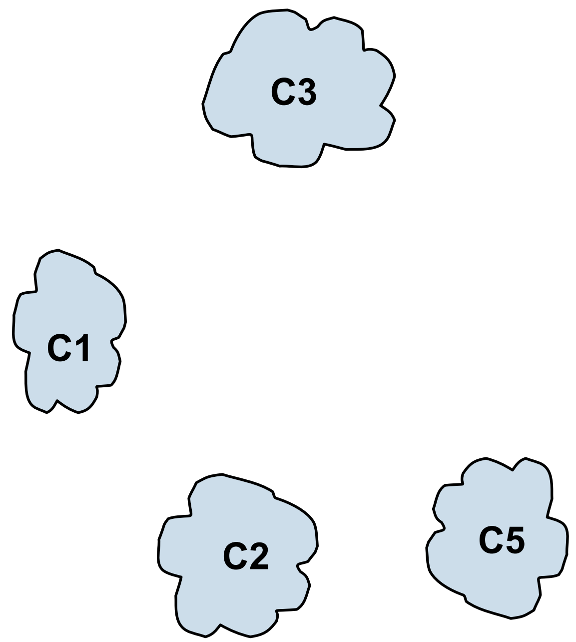
Proximity Matrix

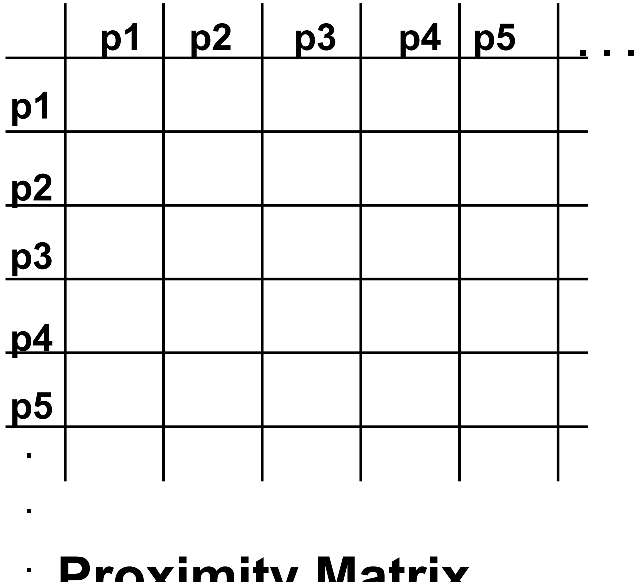


Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

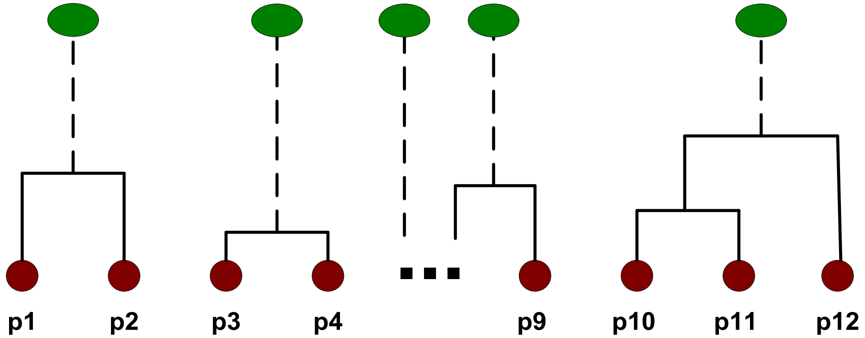
Intermediate Situation

After some merging steps, we have some clusters





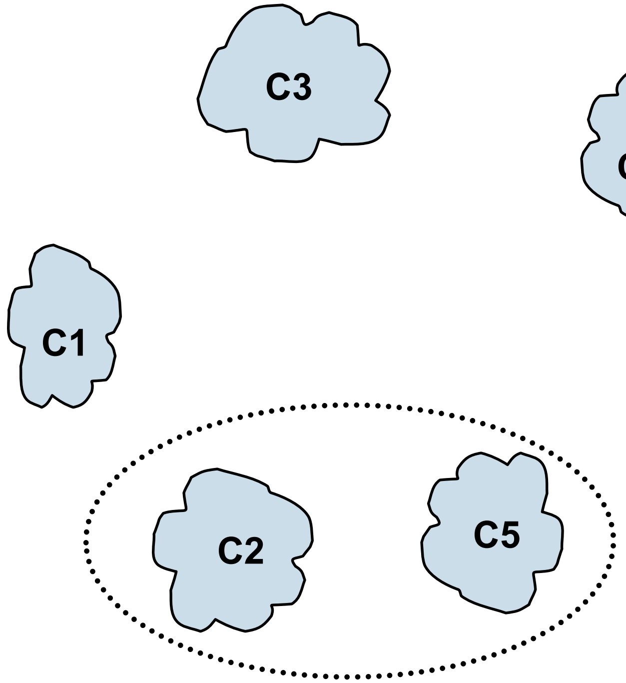
Proximity Matrix

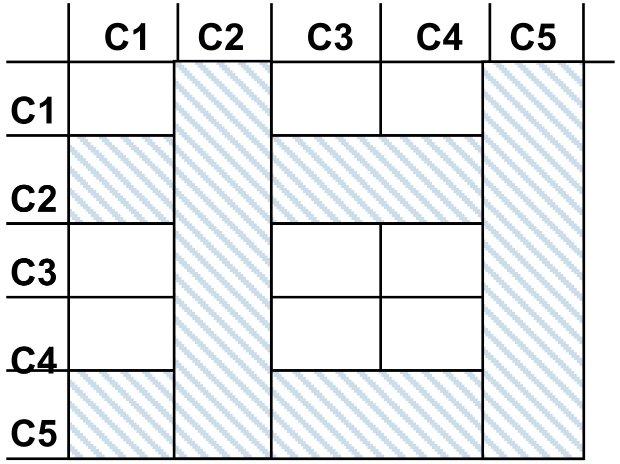


Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

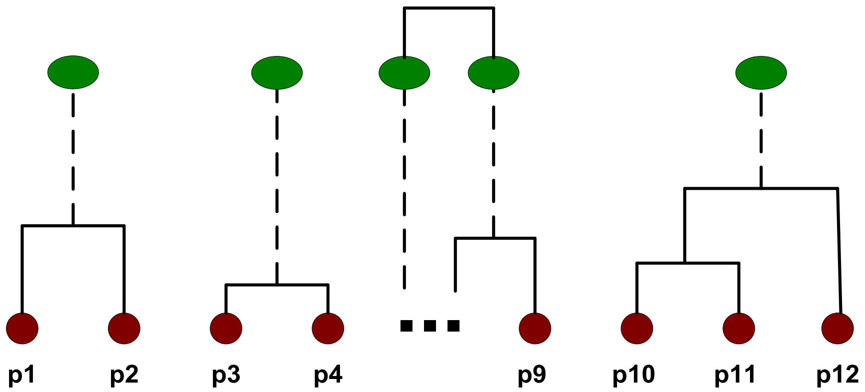
Step 4

We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.





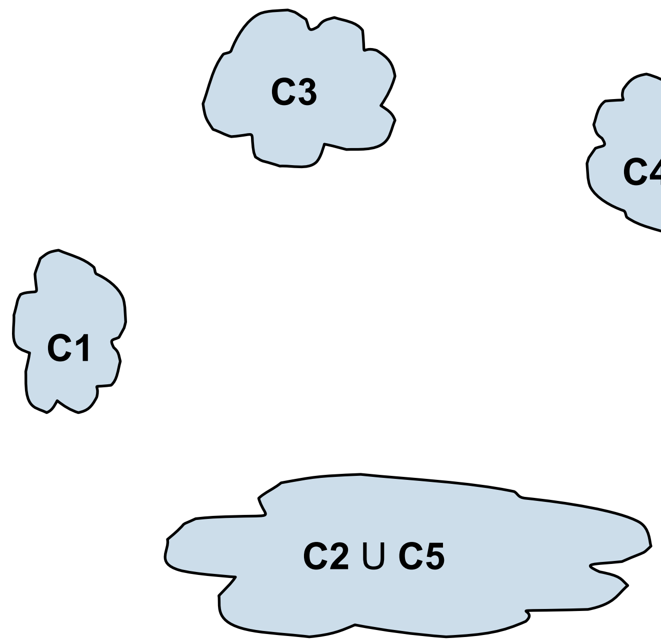
Proximity Matrix



Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

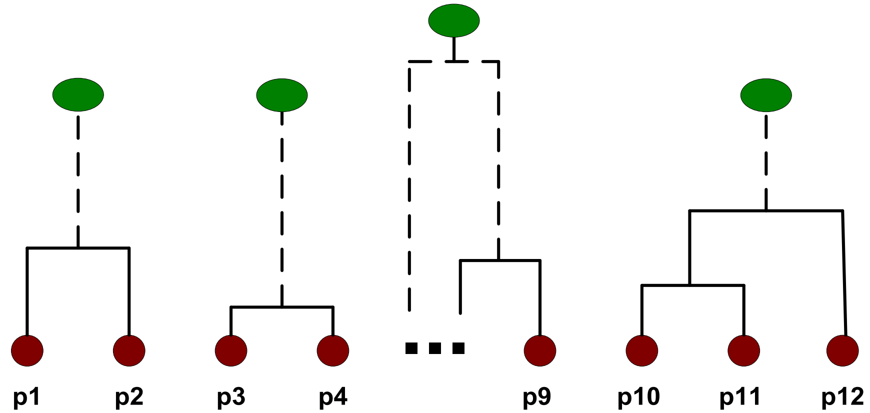
Step 5

The question is “How do we update the proximity matrix?”



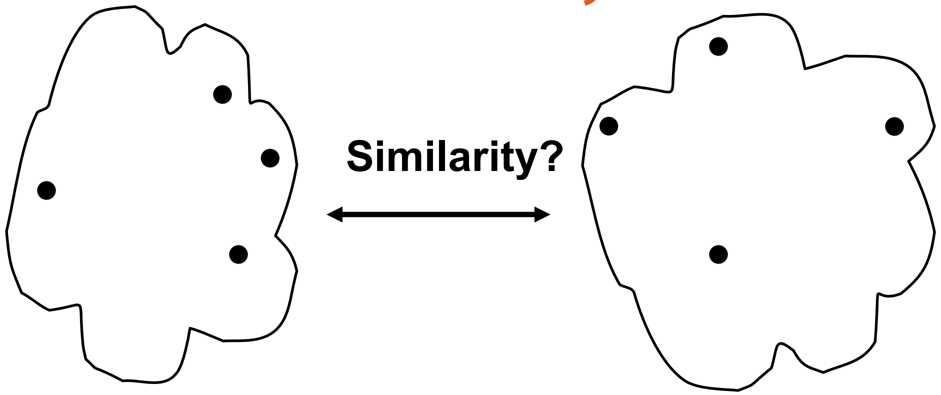
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Defining Inter-Cluster Similarity



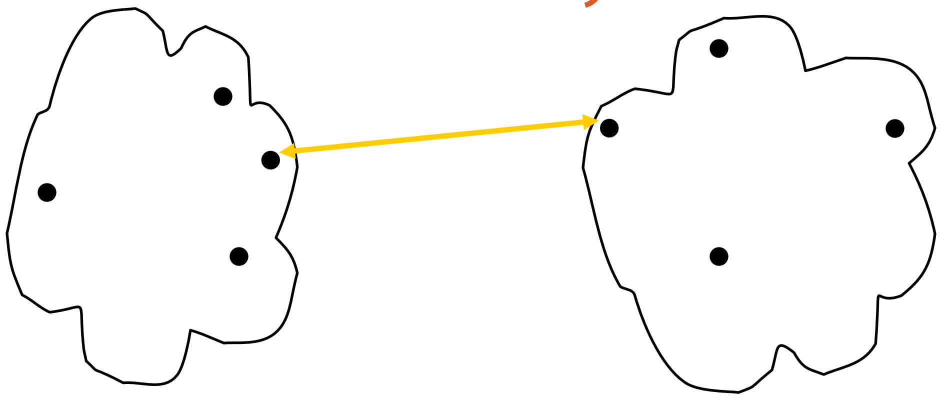
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Defining Inter-Cluster Similarity



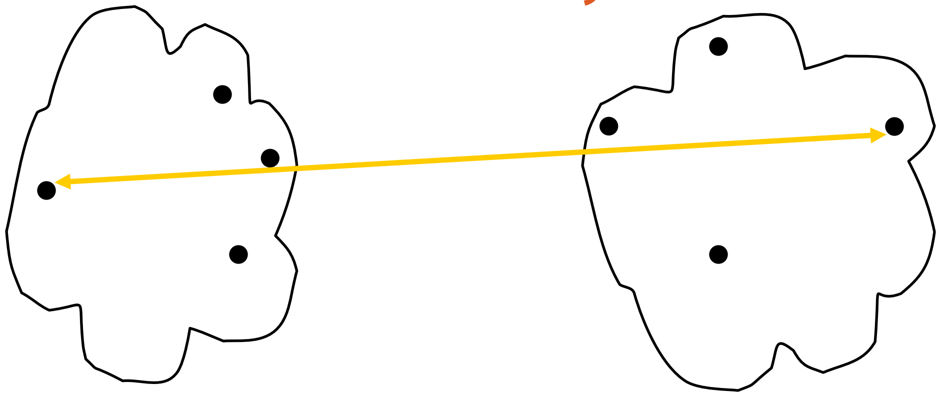
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Defining Inter-Cluster Similarity



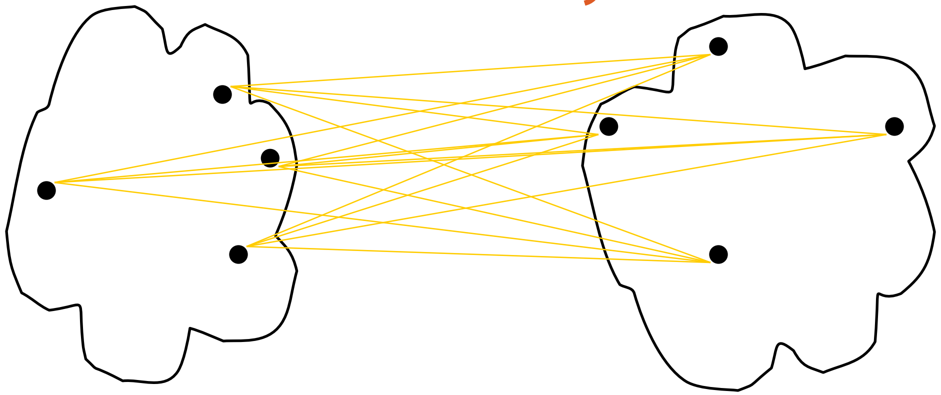
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Defining Inter-Cluster Similarity

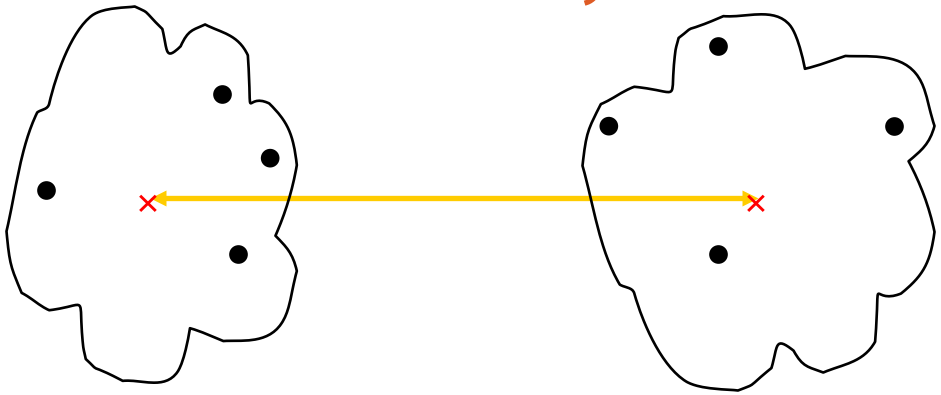


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Defining Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

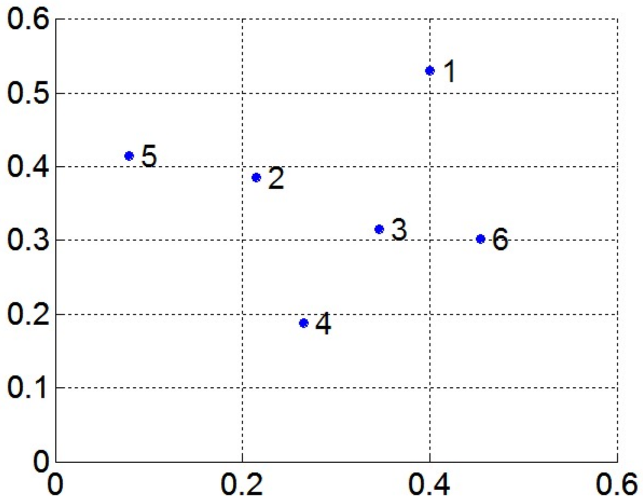
Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

MIN or Single Link

Proximity of two clusters is based on the two closest points in the different clusters

- Determined by one pair of points, i.e., by one link in the proximity graph

Example:

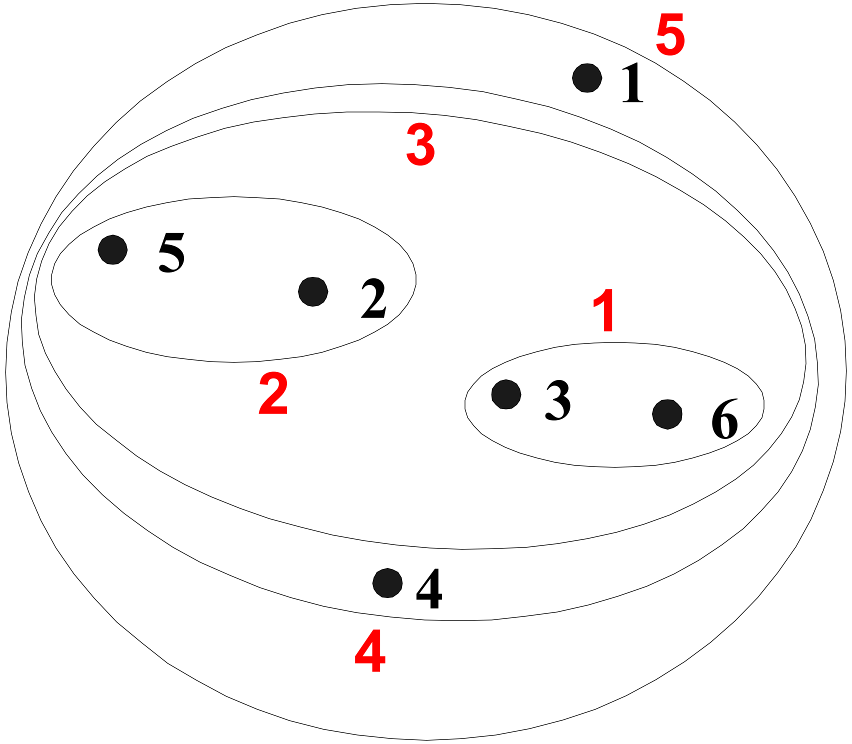


Distance Matrix:

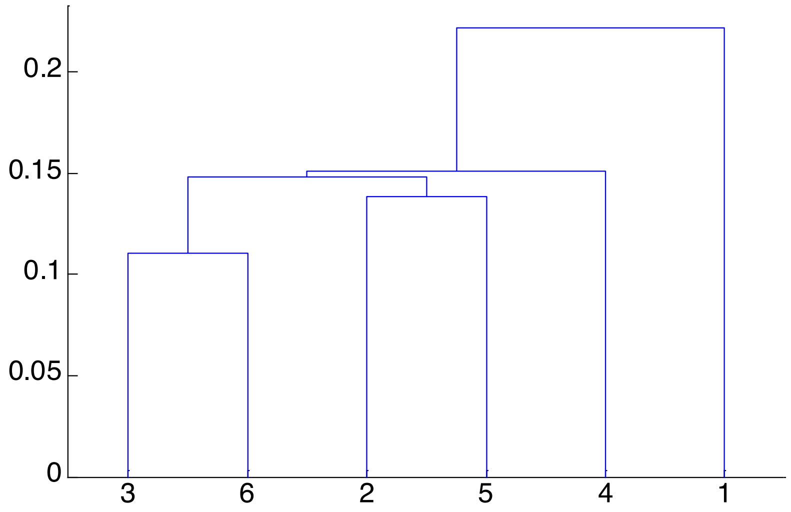
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

MIN or Single Link



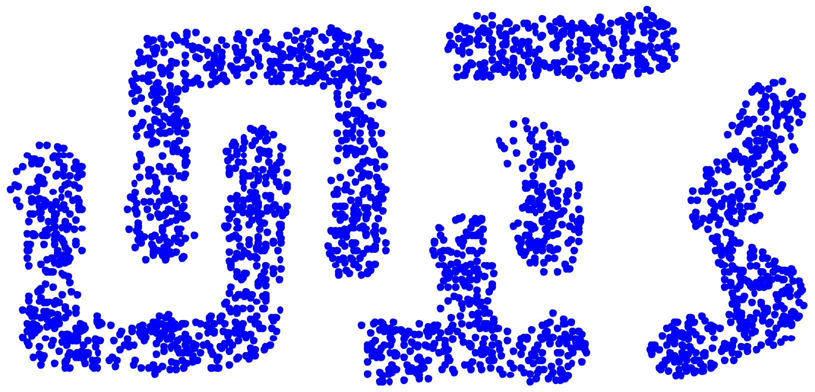
Nested Clusters



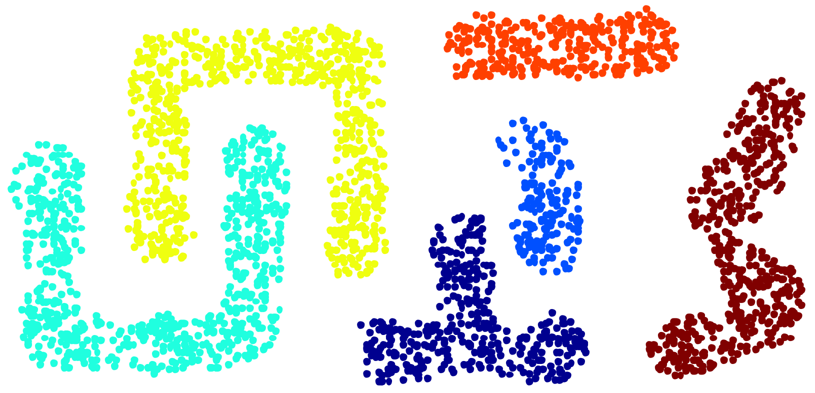
Dendrogram

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Strength of MIN



Original Points

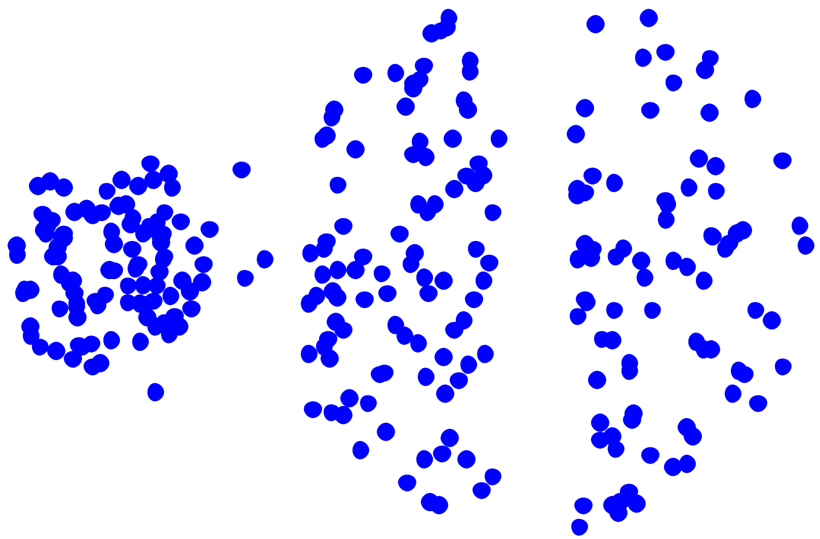


Six Clusters

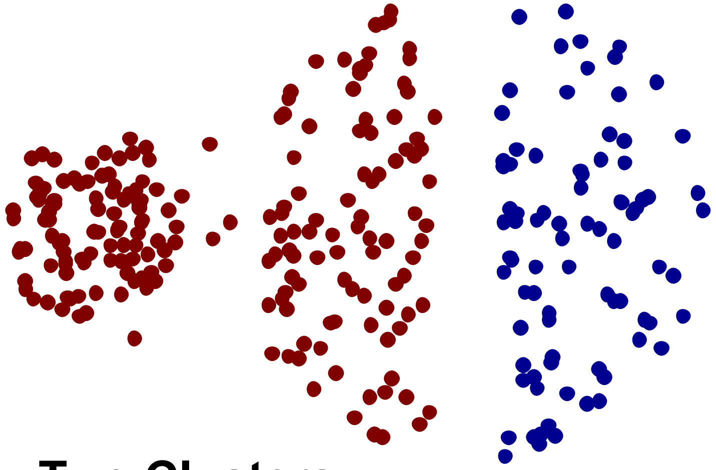
- **Can handle non-elliptical shapes**

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

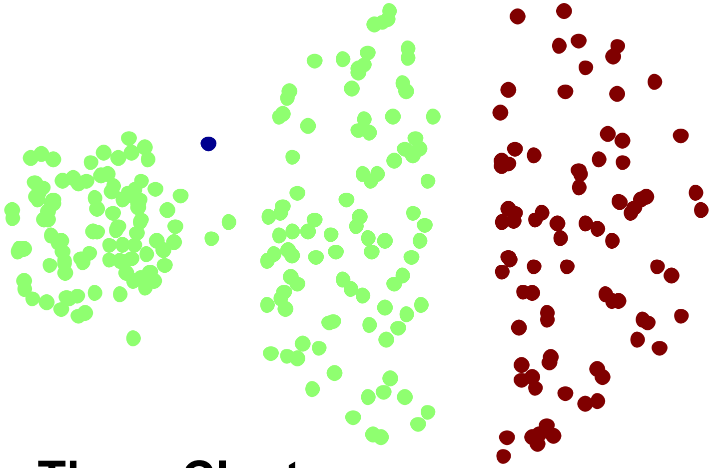
Limitations of MIN



Original Points



Two Clusters



Three Clusters

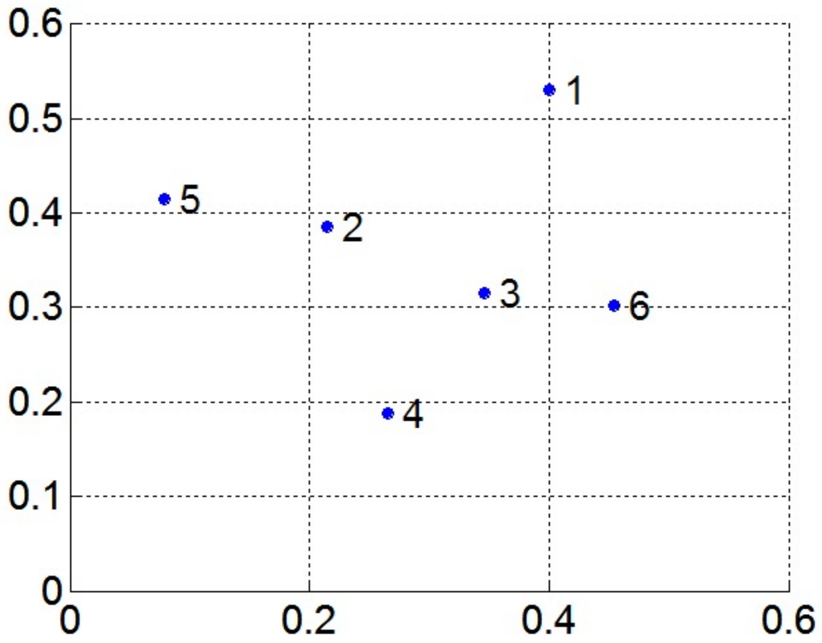
- **Sensitive to noise**

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

MAX or Complete Linkage

Proximity of two clusters is based on the two most distant points in the different clusters

- ints in the two clusters

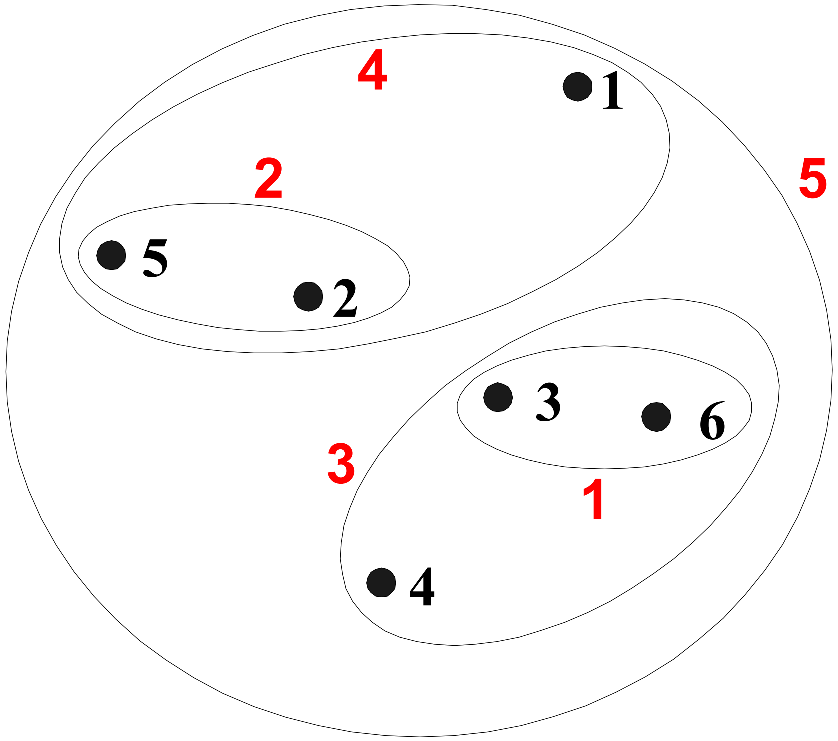


Distance Matrix:

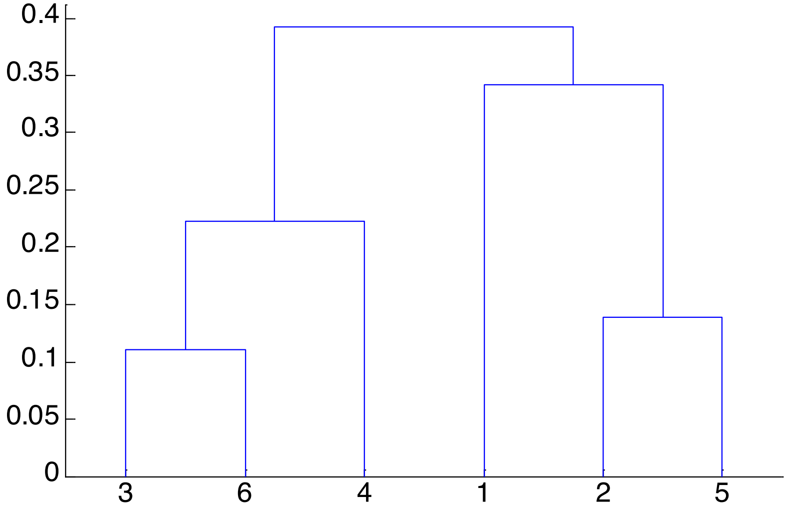
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Hierarchical Clustering: MAX



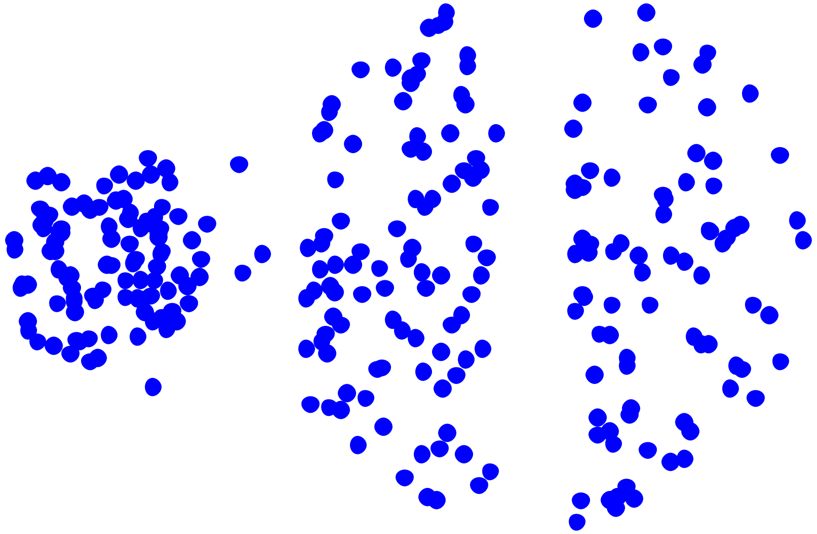
Nested Clusters



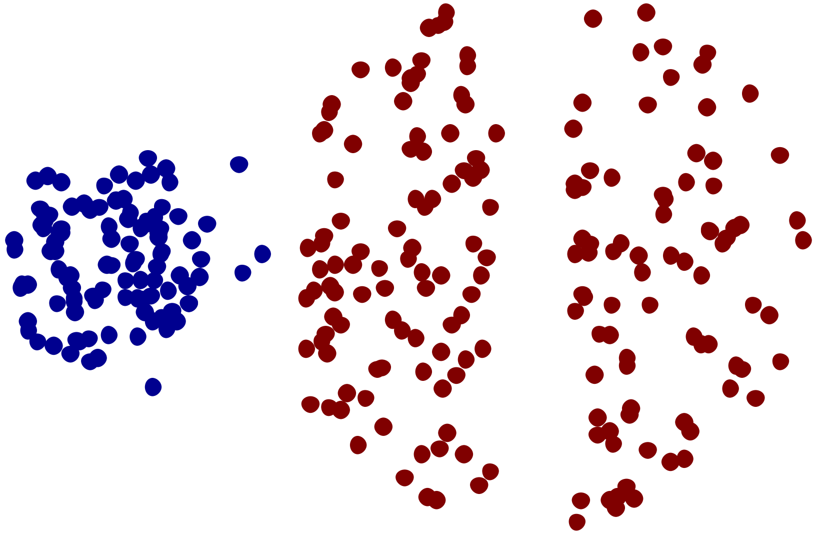
Dendrogram

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Strength of MAX



Original Points

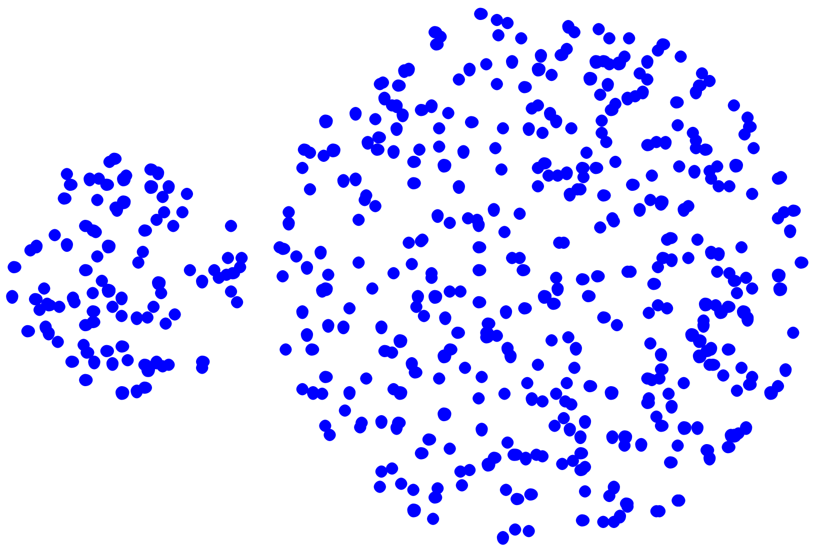


Two Clusters

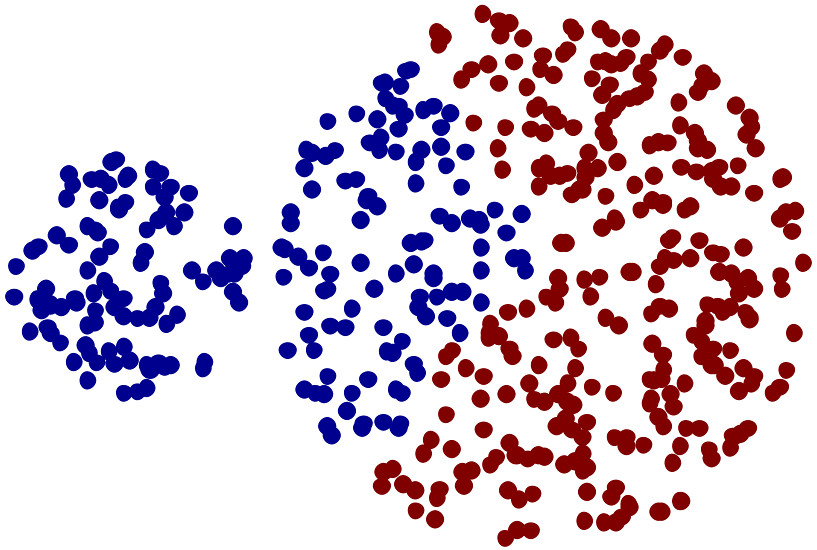
- **Less susceptible to noise**

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Limitations of MAX



Original Points



Two Clusters

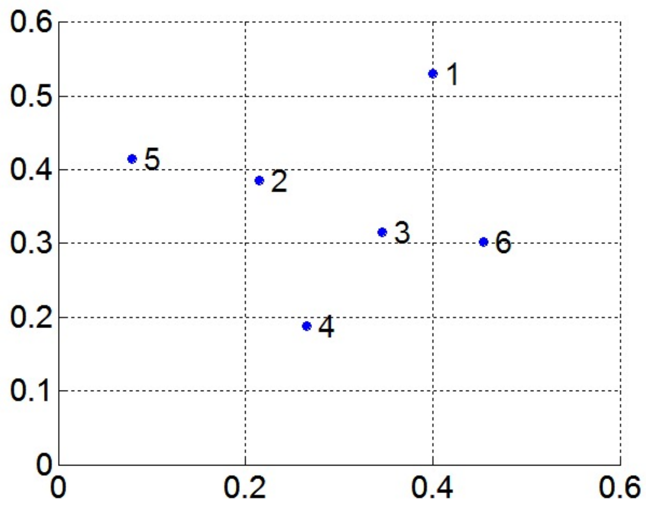
- **Tends to break large clusters**
- **Biased towards globular clusters**

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Group Average

Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\displaystyle\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}$$

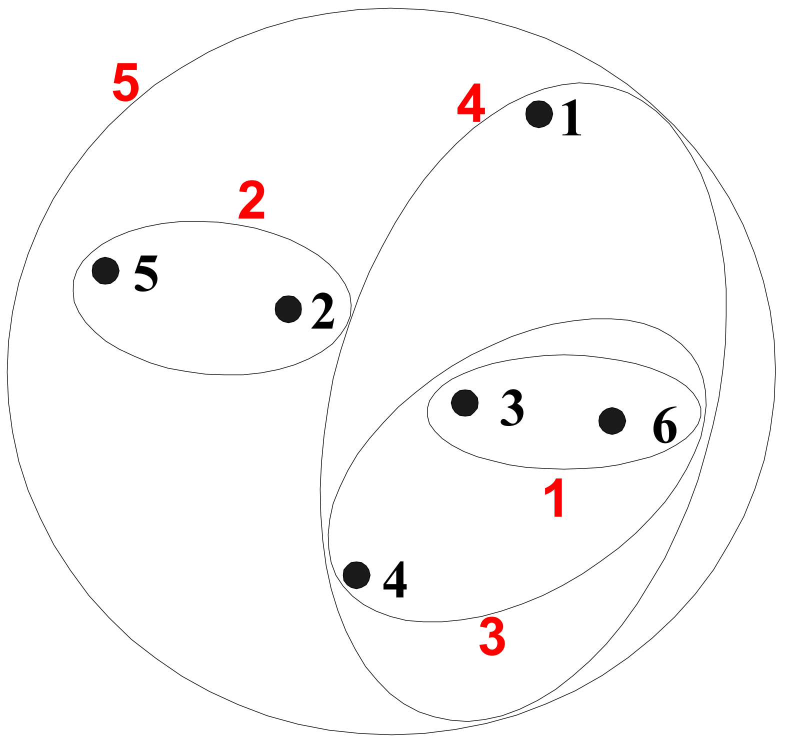


Distance Matrix:

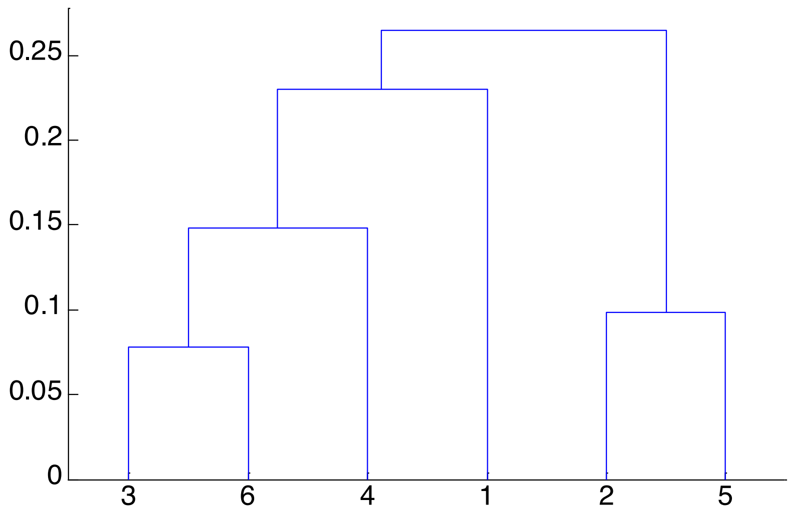
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Hierarchical Clustering: Group Average



Nested Clusters



Dendrogram

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Hierarchical Clustering: Group Average

Compromise between Single and Complete Link

Strengths

- Less susceptible to noise

Limitations

- Biased towards globular clusters

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Cluster Similarity: Ward's Method

Similarity of two clusters is based on the increase in squared error when two clusters are merged

- Similar to group average if distance between points is distance squared

$$\Delta(A,B) = \sum_{i \in A \cup B} \left|\left|\vec{x} - \vec{m}_{A \cup B}\right|\right|^2 - \sum_{i \in A} \left|\left|\vec{x} - \vec{m}_{A}\right|\right|^2 - \sum_{i \in B} \left|\left|\vec{x} - \vec{m}_{B}\right|\right|^2$$

Less susceptible to noise

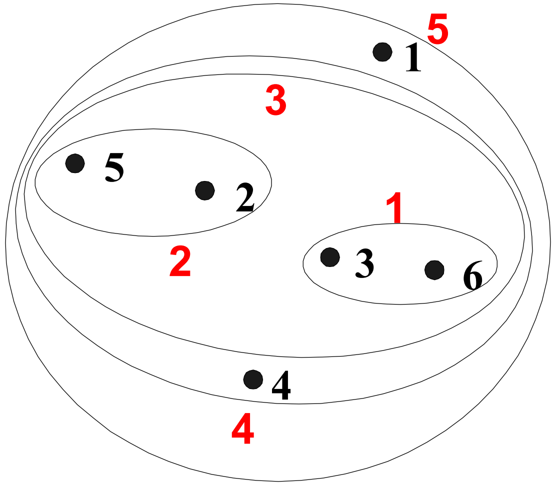
Biased towards globular clusters

Hierarchical analogue of K-means

- Can be used to initialize K-means

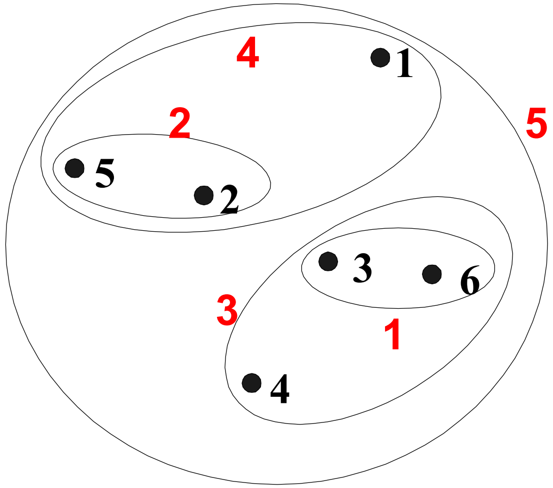
Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

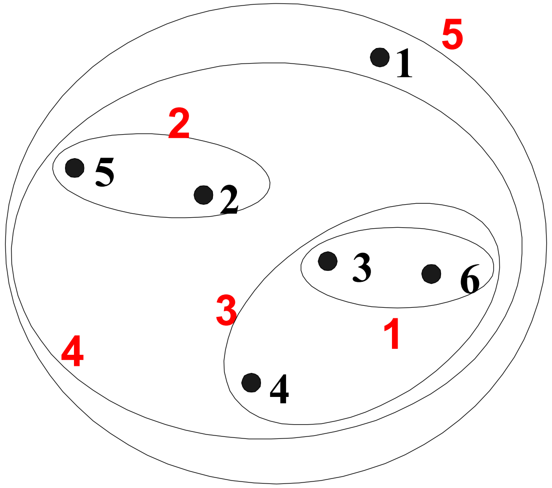
Hierarchical Clustering: Comparison



MIN

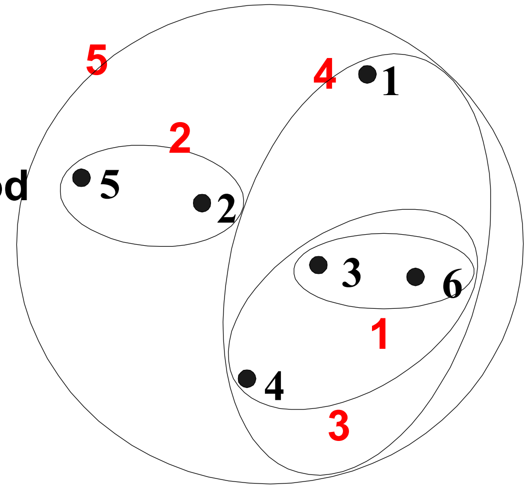
MAX





Group Average

Ward's Method



Hierarchical Clustering: Time and Space requirements

$O(N^2)$ space since it uses the proximity matrix.

- N is the number of points.

$O(N^3)$ time in many cases

- There are N steps and at each step the size, N^2, proximity matrix must be updated and searched
- Complexity can be reduced to $O(N^2 \log(N))$ time with some cleverness

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Hierarchical Clustering: Problems and Limitations

Once a decision is made to combine two clusters, it cannot be undone

No global objective function is directly minimized

Different schemes have problems with one or more of the following:

- Sensitivity to noise
- Difficulty handling clusters of different sizes and non-globular shapes
- Breaking large clusters

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Code Example

Performing Agglomerative Clustering and Drawing Dendrogram.

https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py

Density-Based Clustering

Clusters are regions of high density that are separated from one another by regions on low density.

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

DBSCAN

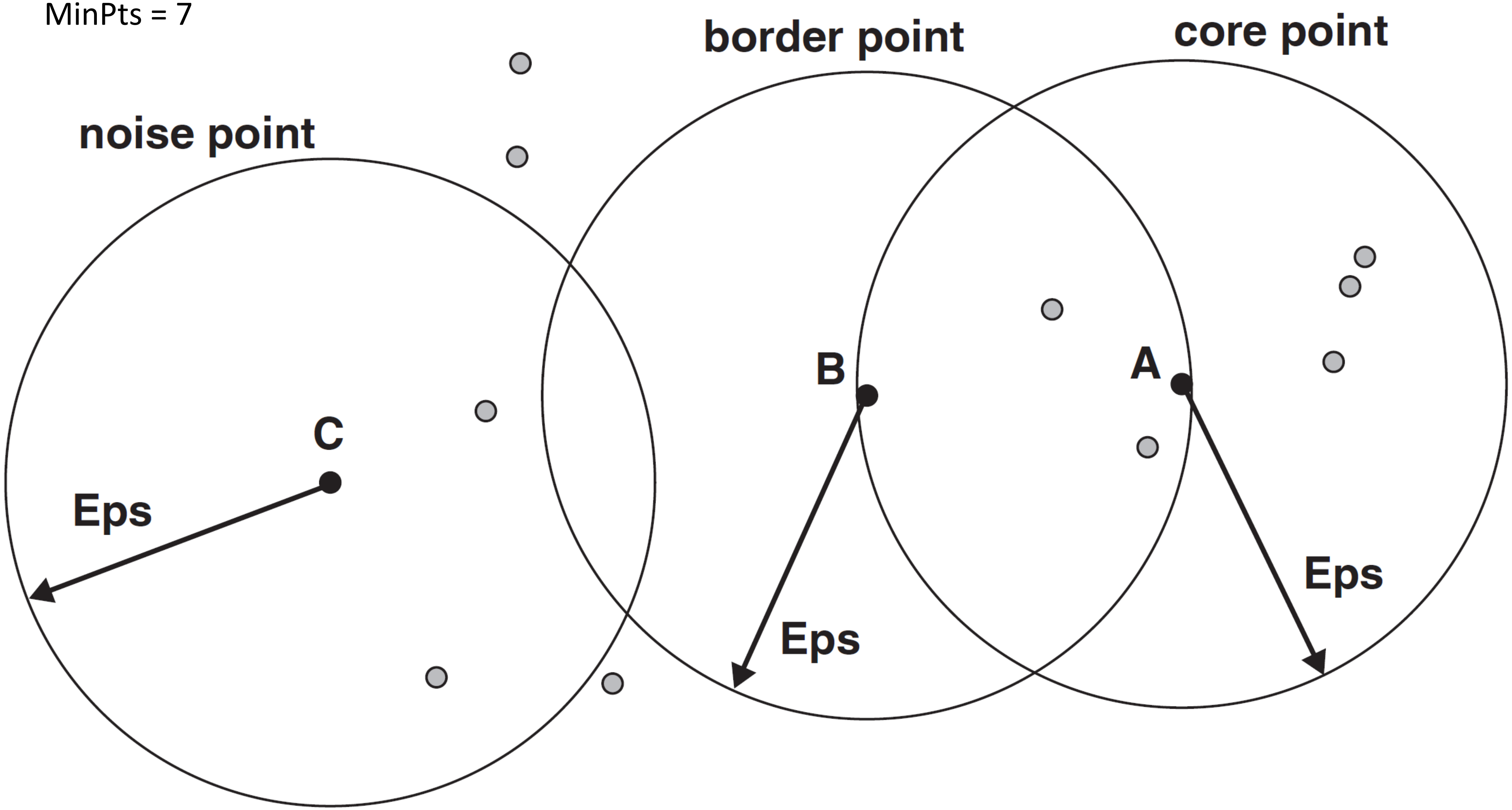
DBSCAN is a density-based algorithm.

- Density = number of points within a specified radius (Eps)
- A point is a core point if it has at least a specified number of points (MinPts) within Eps
  - These are points that are at the interior of a cluster
  - Counts the point itself
- A border point is not a core point, but is in the neighborhood of a core point
- A noise point is any point that is not a core point or a border point

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

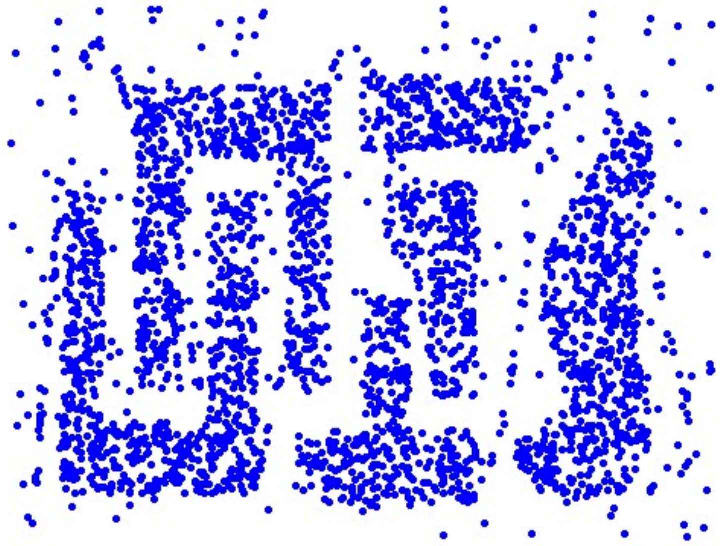
DBSCAN: Core, Border, and Noise Points

MinPts = 7

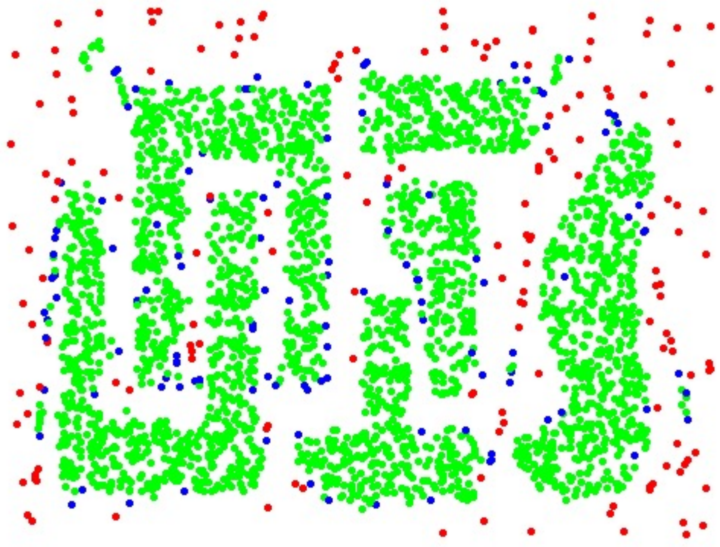


Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

DBSCAN: Core, Border, and Noise Points



Original Points



Point types: core, border and noise

Eps = 10, MinPts = 4

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

DBSCAN Algorithm

Form clusters using core points, and assign border points to one of its neighboring clusters

1: Label all points as core, border, or noise points.

2: Eliminate noise points.

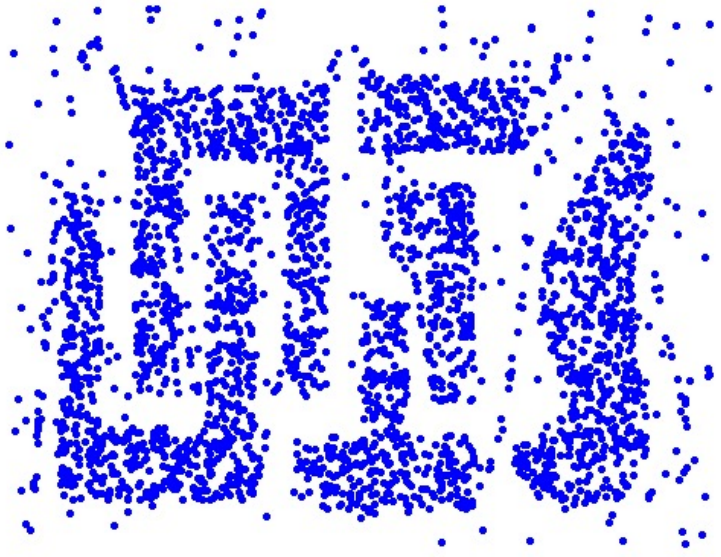
3: Put an edge between all core points within a distance *Eps* of each other.

4: Make each group of connected core points into a separate cluster.

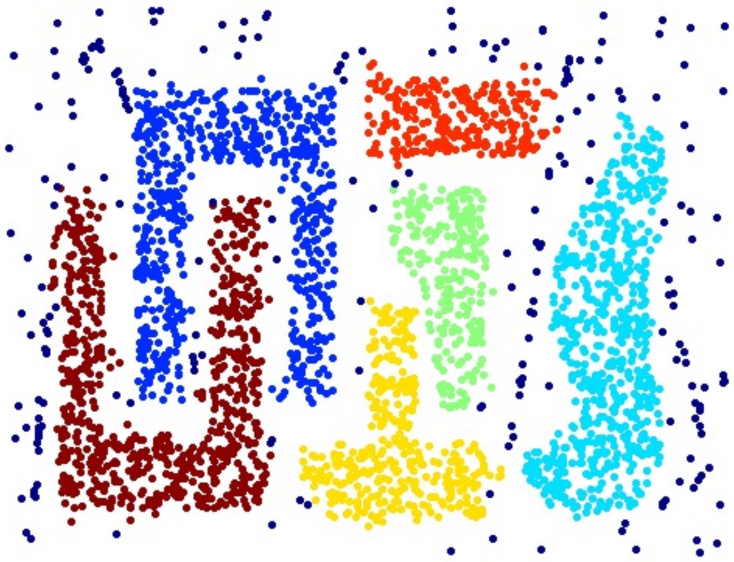
5: Assign each border point to one of the clusters of its associated core points

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

When DBSCAN Works Well



Original Points

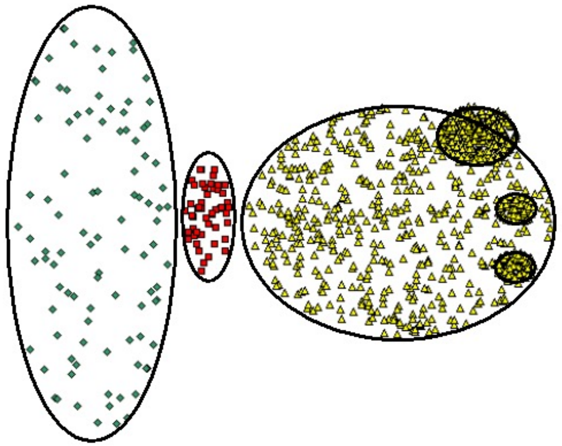


Clusters (dark blue points indicate noise)

- **Can handle clusters of different shapes and sizes**
- **Resistant to noise**

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

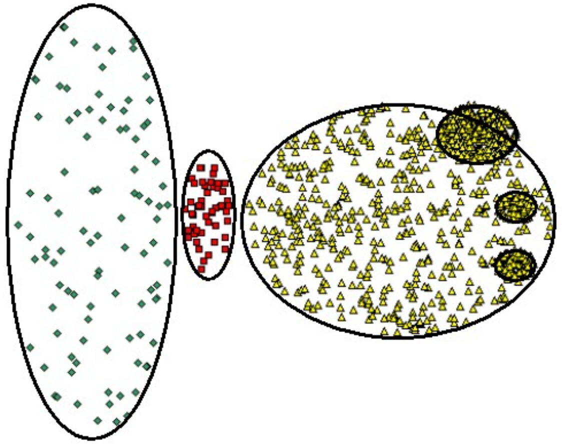
When DBSCAN Does NOT Work Well



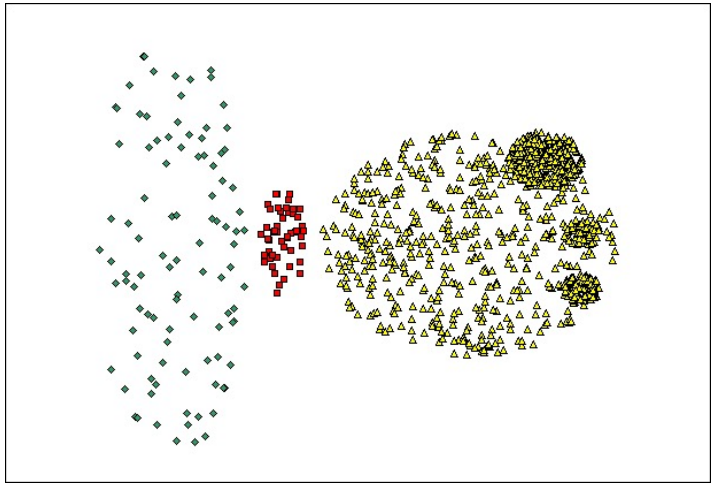
Original Points

Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

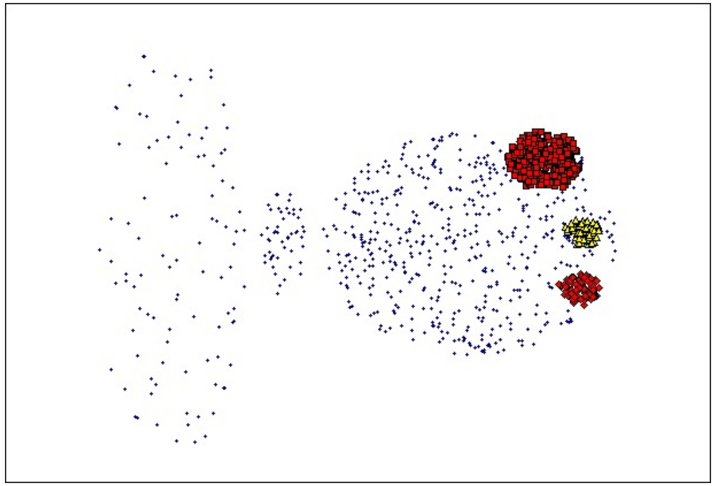
When DBSCAN Does NOT Work Well



Original Points



(MinPts=4, Eps=9.92).



(MinPts=4, Eps=9.75)

- **Varying densities**
- **High-dimensional data**

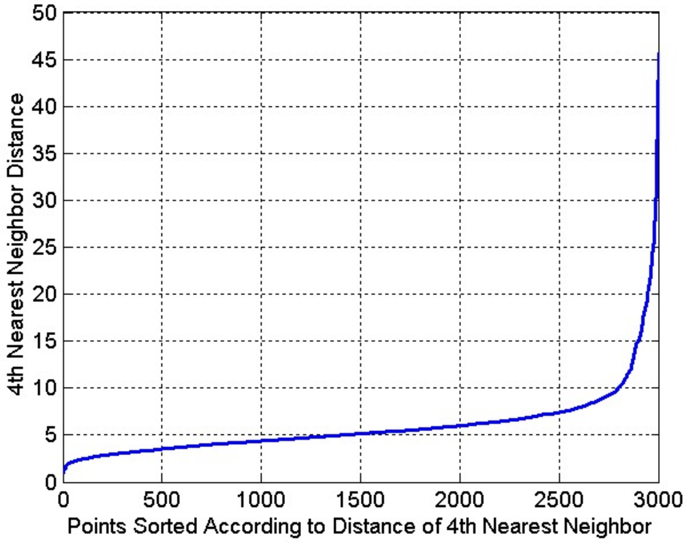
Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

DBSCAN: Determining EPS and MinPts

Idea is that for points in a cluster, their k^{th} nearest neighbors are at close distance

Noise points have the k^{th} nearest neighbor at farther distance

So, plot sorted distance of every point to its k^{th} nearest neighbor



Retrieved from Introduction to Data Mining, 2nd Edution by Tan, Steinbach, Karpatne, Kumar

Code Example

DBSCAN and Coloring the Clusters

https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py