

# DATA MINING 1

## Clustering

---

Dino Pedreschi, Riccardo Guidotti

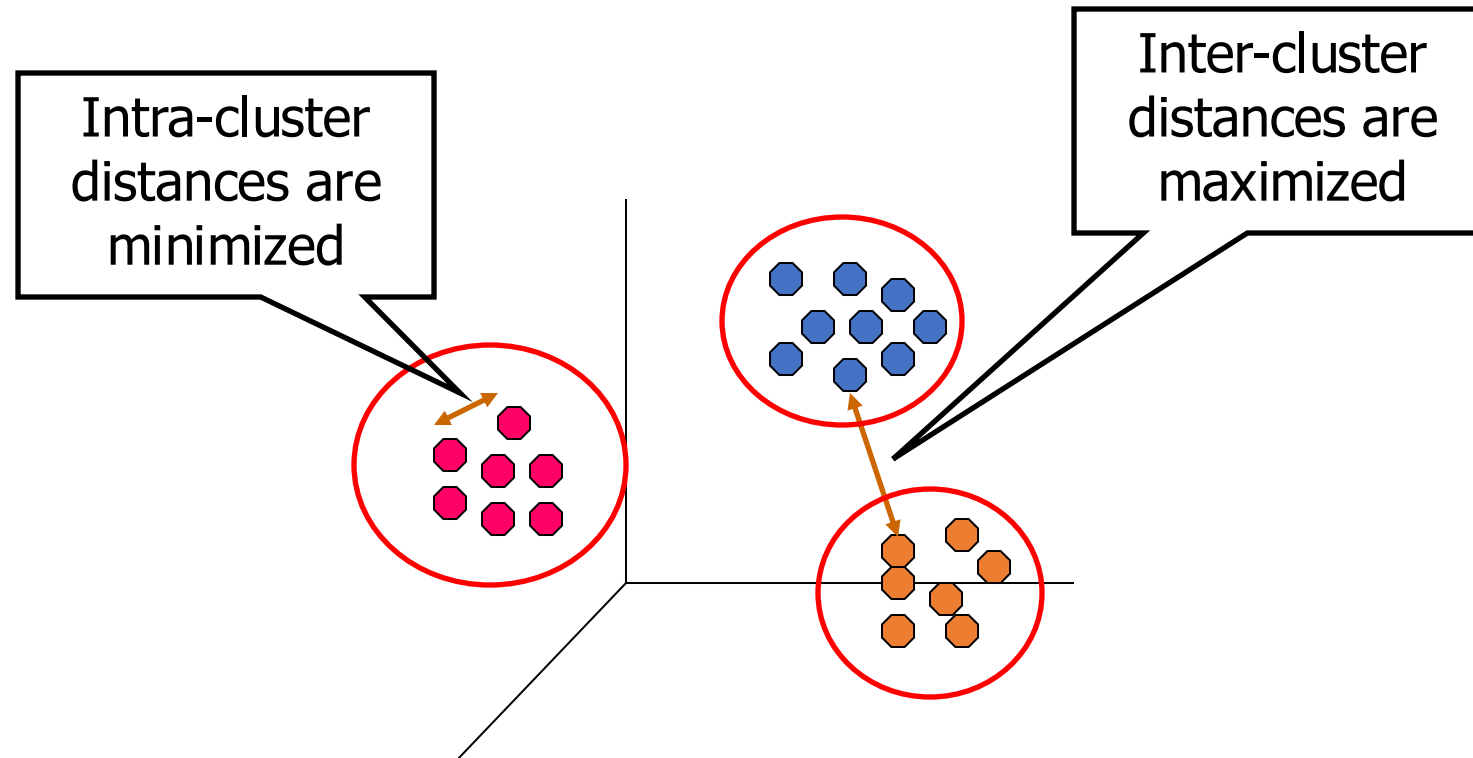
Revisited slides from Lecture Notes for Chapter 7 “Introduction to Data Mining”, 2nd  
Edition by Tan, Steinbach, Karpatne, Kumar



UNIVERSITÀ  
DI PISA

# What is Cluster Analysis?

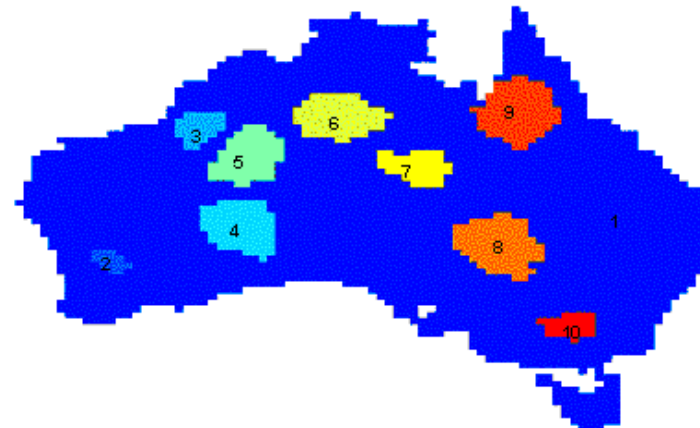
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.



# Applications of Cluster Analysis

- **Understanding**
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets

	<i>Discovered Clusters</i>	<i>Industry Group</i>
<b>1</b>	Applied-Matl-DOWN, Bay-Network-DOWN, 3-COM-DOWN, Cabletron-Sys-DOWN, CISCO-DOWN, HP-DOWN, DSC-Comm-DOWN, INTEL-DOWN, LSI-Logic-DOWN, Micron-Tech-DOWN, Texas-Inst-DOWN, Tellabs-Inc-DOWN, Natl-Semiconduct-DOWN, Oracl-DOWN, SGI-DOWN, Sun-DOWN	Technology1-DOWN
<b>2</b>	Apple-Comp-DOWN, Autodesk-DOWN, DEC-DOWN, ADV-Micro-Device-DOWN, Andrew-Corp-DOWN, Computer-Assoc-DOWN, Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN, Microsoft-DOWN, Scientific-Atl-DOWN	Technology2-DOWN
<b>3</b>	Fannie-Mae-DOWN, Fed-Home-Loan-DOWN, MBNA-Corp-DOWN, Morgan-Stanley-DOWN	Financial-DOWN
<b>4</b>	Baker-Hughes-UP, Dresser-Inds-UP, Halliburton-HLD-UP, Louisiana-Land-UP, Phillips-Petro-UP, Unocal-UP, Schlumberger-UP	Oil-UP



Clustering precipitation  
in Australia

# What is not Cluster Analysis?

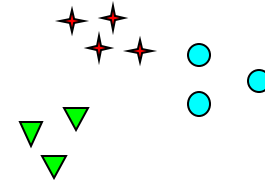
---

- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
  - Clustering is a grouping of objects based on the data
- Supervised classification
  - Have class label information
- Association Analysis
  - Local vs. global connections

# Notion of a Cluster can be Ambiguous



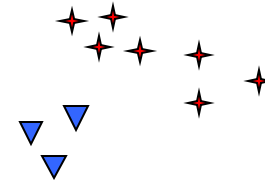
How many clusters?



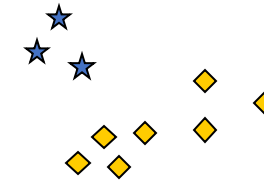
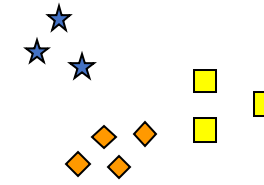
Six Clusters



Two Clusters



Four Clusters



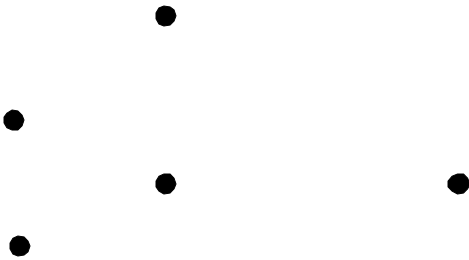
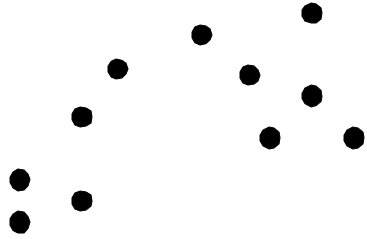
# Types of Clusterings

---

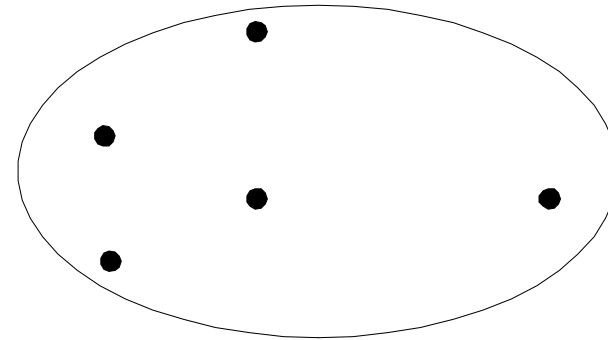
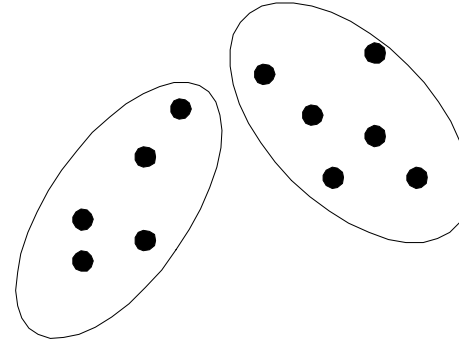
- A **clustering** is a set of clusters
- Important distinction between **hierarchical** and **partitional** sets of clusters
- Partitional Clustering
  - A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

# Partitional Clustering

---



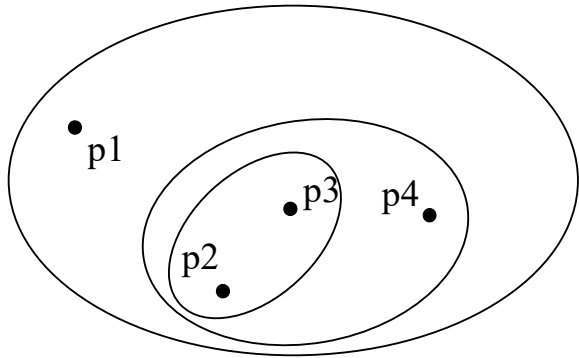
**Original Points**



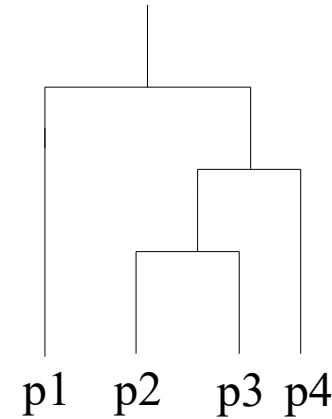
**A Partitional Clustering**

# Hierarchical Clustering

---



**Traditional Hierarchical Clustering**



**Traditional Dendrogram**



# Other Distinctions Between Sets of Clusters

---

- Exclusive versus non-exclusive
  - In non-exclusive clusterings, **points may belong to multiple clusters.**
  - Can represent multiple classes or '**border**' points
- Fuzzy versus non-fuzzy
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - **Probabilistic** clustering has similar characteristics
- Partial versus complete
  - In some cases, we only want to cluster some of the data
- Heterogeneous versus homogeneous
  - Clusters of widely different sizes, shapes, and densities

# Types of Clusters

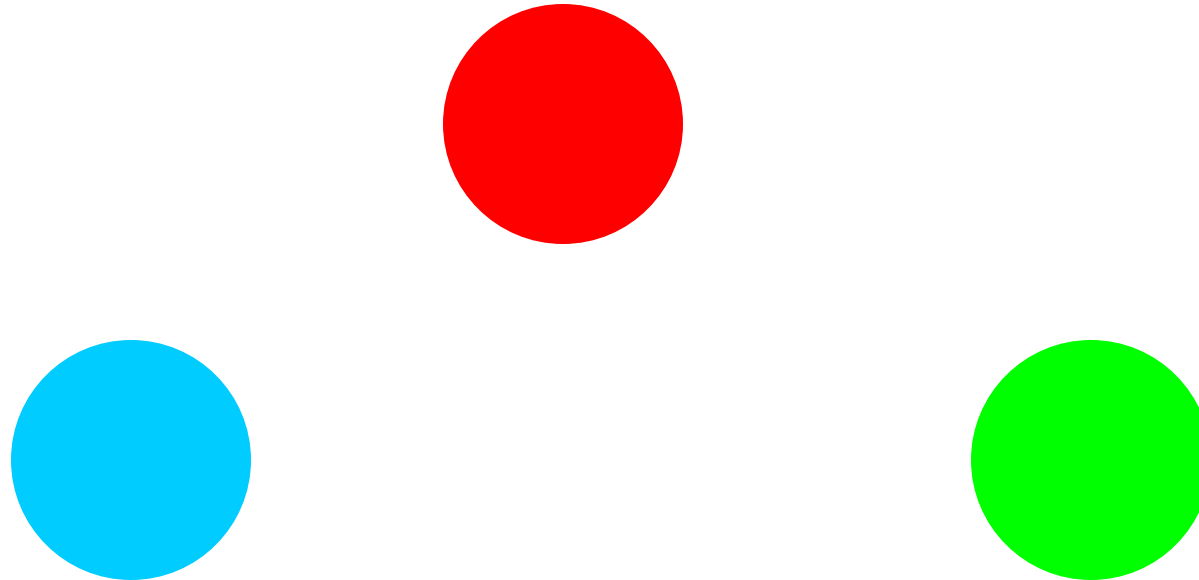
---

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

# Types of Clusters: Well-Separated

---

- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

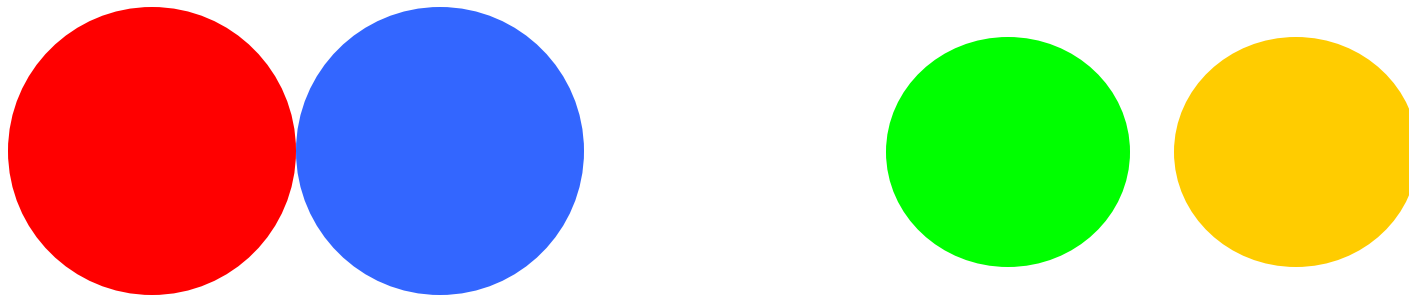


**3 well-separated clusters**

# Types of Clusters: Center-Based

---

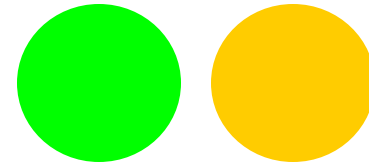
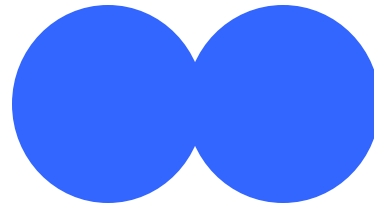
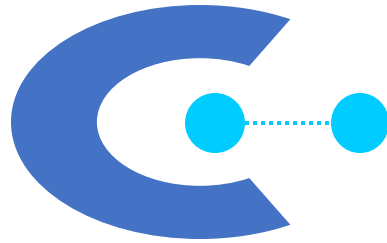
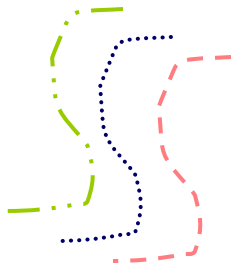
- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
  - The center of a cluster is often a **centroid**, the average of all the points in the cluster, or a **medoid**, the most “representative” point of a cluster



**4 center-based clusters**

# Types of Clusters: Contiguity-Based

- Contiguous Cluster (Nearest neighbor or Transitive)
  - Each point is closer to at least one point in its cluster than to any point in another cluster.
  - Graph based clustering

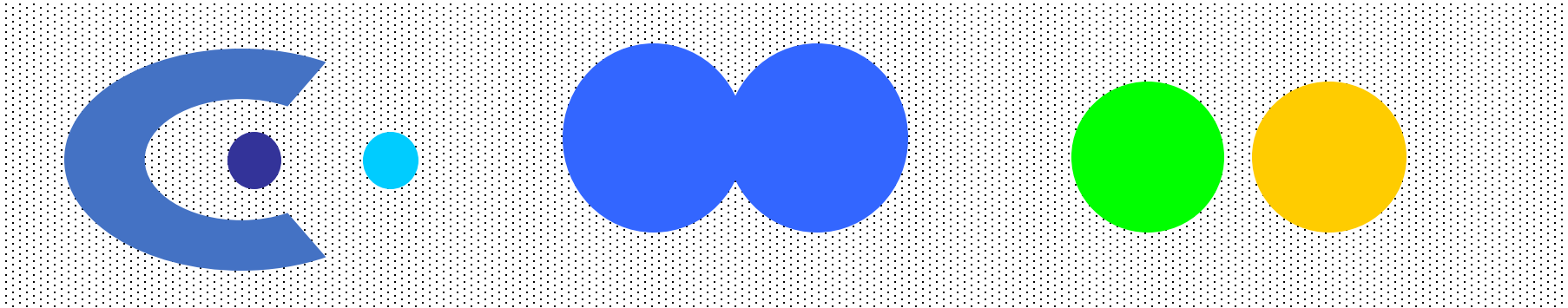


- This approach can **have trouble when noise is present** since a small bridge of points can **merge two distinct clusters**

8 contiguous clusters

# Types of Clusters: Density-Based

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



**6 density-based clusters**

# Types of Clusters: Objective Function

---

- Clusters Defined by an Objective Function
  - Finds clusters that **minimize** or **maximize an objective function**.
  - Enumerate all possible ways of dividing the points into clusters and evaluate the 'goodness' of each potential set of clusters by using the given objective function. (NP Hard)
  - Can have global or local objectives.
    - Hierarchical clustering algorithms typically have local objectives
    - Partitional algorithms typically have global objectives

# Characteristics of the Input Data Are Important

---

- Type of proximity or density measure
  - Central to clustering
  - Depends on data and application
- Data characteristics that affect proximity and/or density are
  - Dimensionality and Sparseness
  - Attribute type
  - Special relationships in the data (e.g., autocorrelation)
  - Distribution of the data
- Noise and Outliers
  - Often interfere with the operation of the clustering algorithm



# Cluster Validity

---

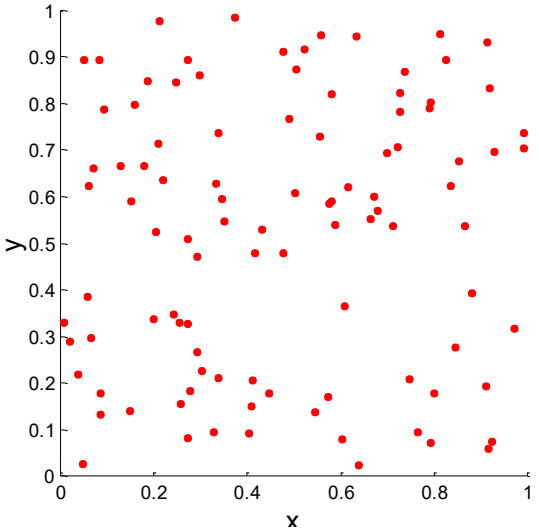
# Cluster Validity

---

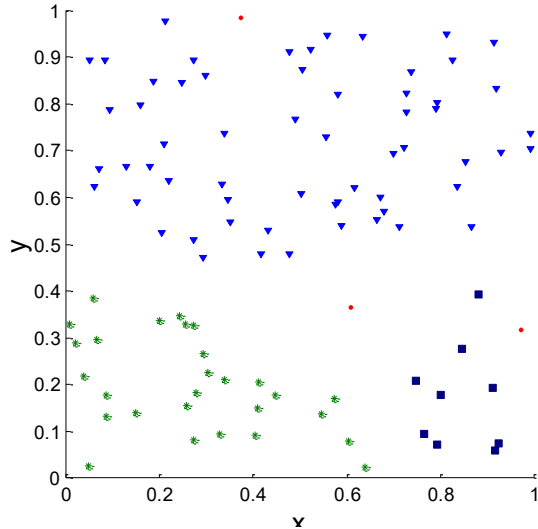
- How can we evaluate the “goodness” of the resulting clusters?
- But “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters

# Clusters found in Random Data

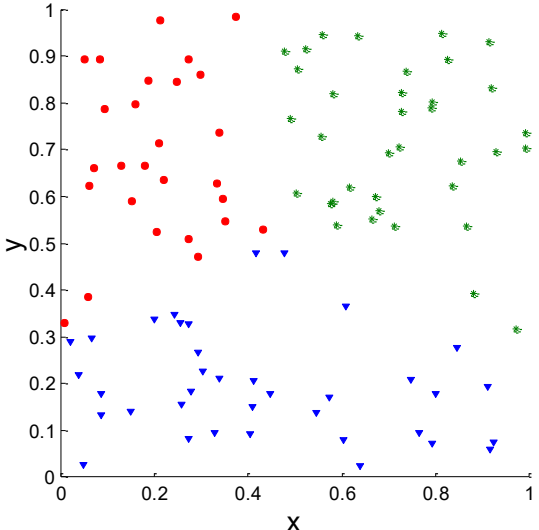
Random  
Points



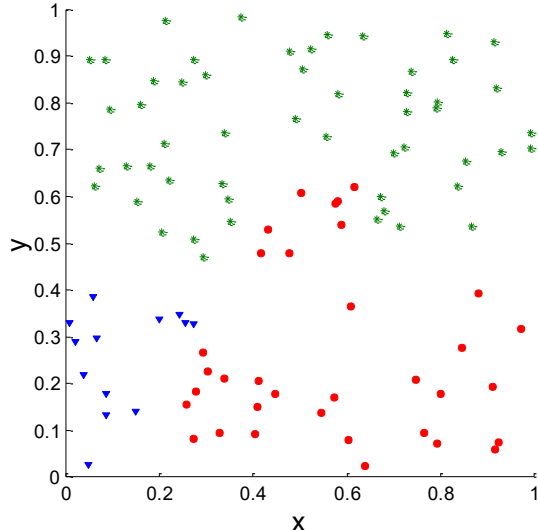
DBSCAN



K-means



Complete  
Link



# Different Aspects of Cluster Validation

---

1. Determining the **clustering tendency** of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information (Use only the data).
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the 'correct' number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

# Measures of Cluster Validity

---

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - **External Index:** Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - **Internal Index:** Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
  - **Relative Index:** Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as **criteria** instead of **indices**
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

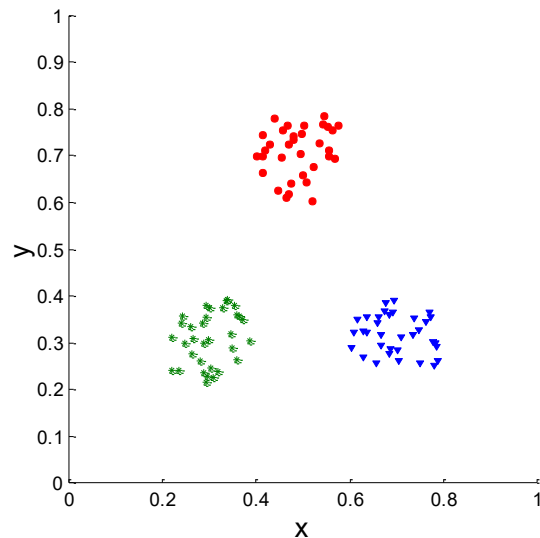
# Measuring Cluster Validity Via Correlation

---

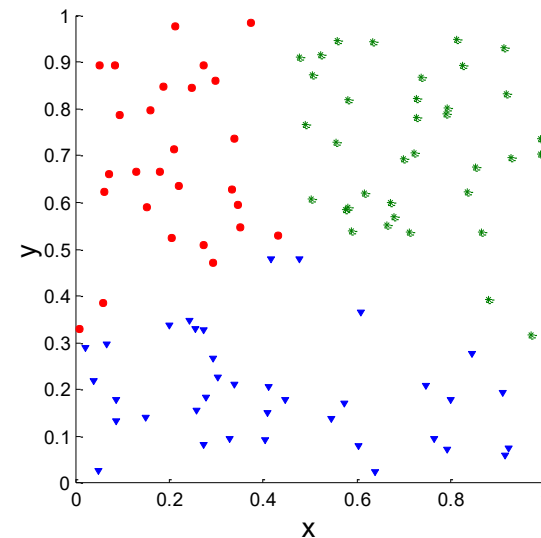
- Two matrices
  - A: Distance/Similarity Matrix
  - B: Ideal Similarity Matrix
    - One row and one column for each data point
    - An entry is 1 if the associated pair of points belong to the same cluster
    - An entry is 0 if the associated pair of points belongs to different clusters
- Compute the correlation between the two matrices A and B
  - Since the matrices are symmetric, only the correlation between  $n(n-1) / 2$  entries needs to be calculated.
- High correlation indicates that points that belong to the same cluster are close to each other.
- 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 of the following two data sets.



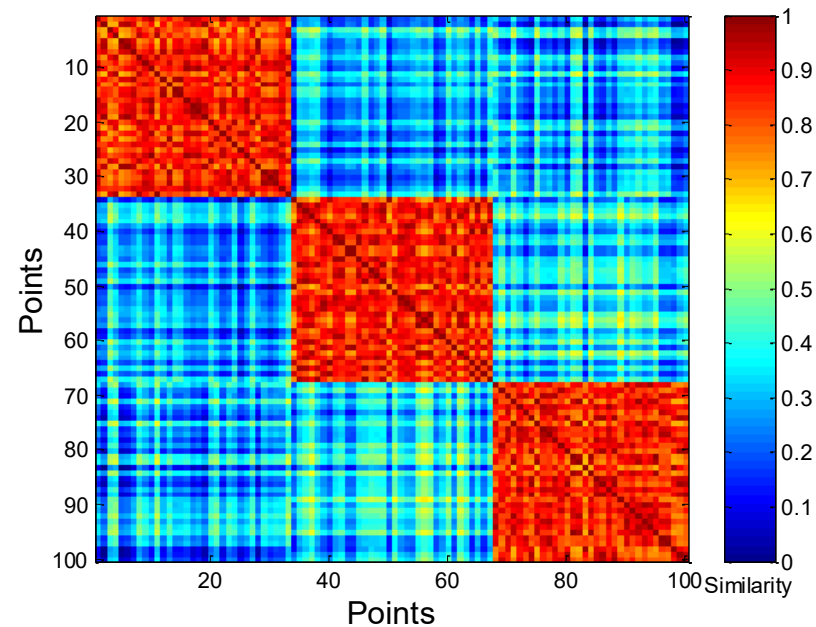
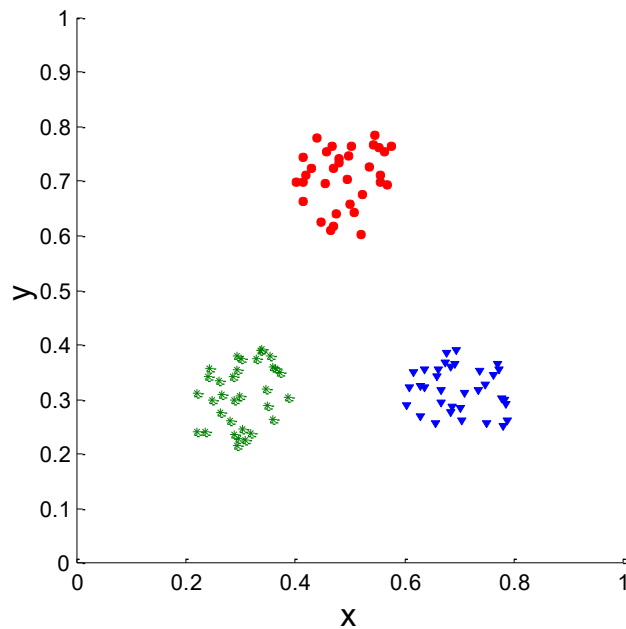
**Corr = -0.9235**



**Corr = -0.5810**

# Using Similarity Matrix for Cluster Validation

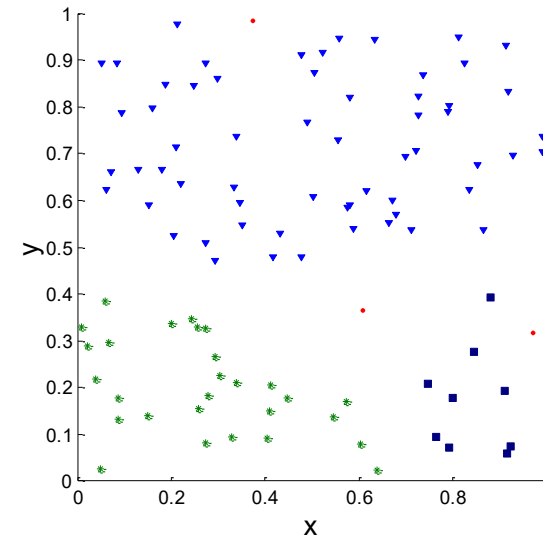
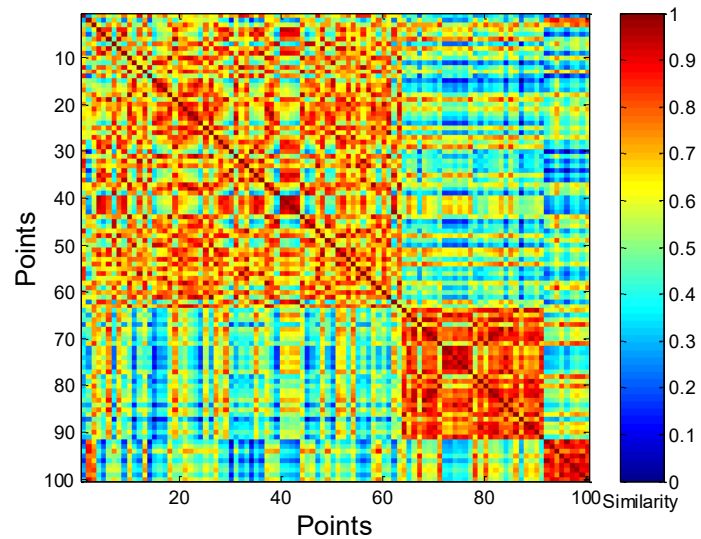
- Order the distance matrix with respect to cluster labels and inspect visually.





# Using Similarity Matrix for Cluster Validation

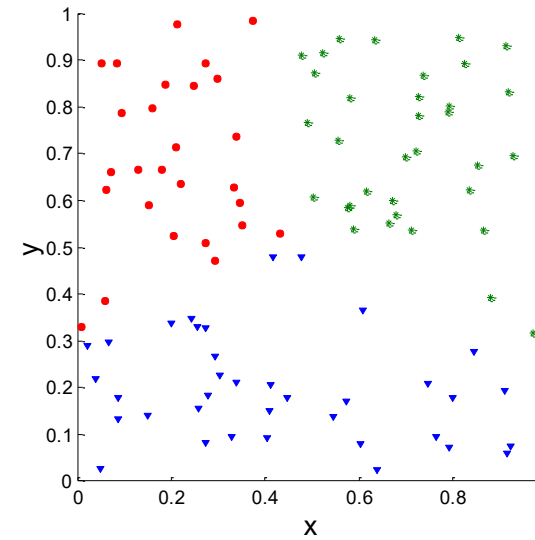
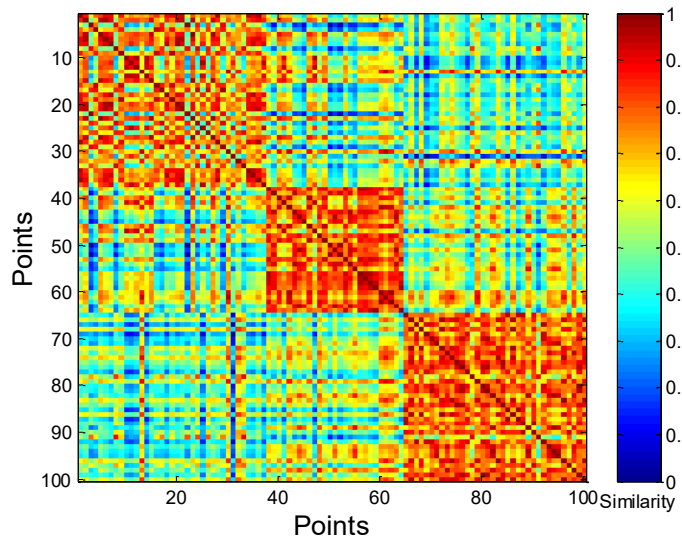
- Clusters in random data are not so crisp



**DBSCAN**

# Using Similarity Matrix for Cluster Validation

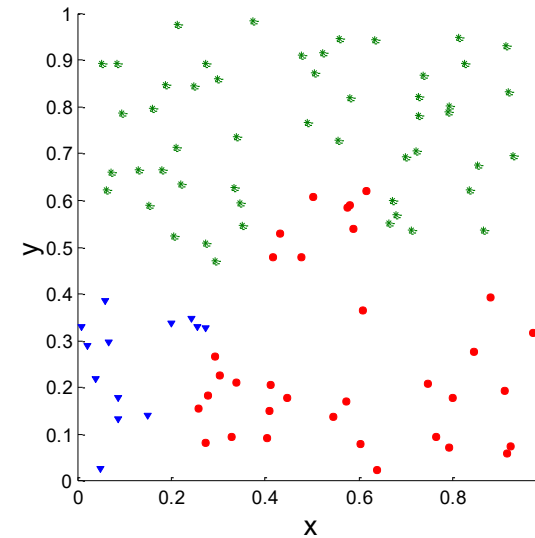
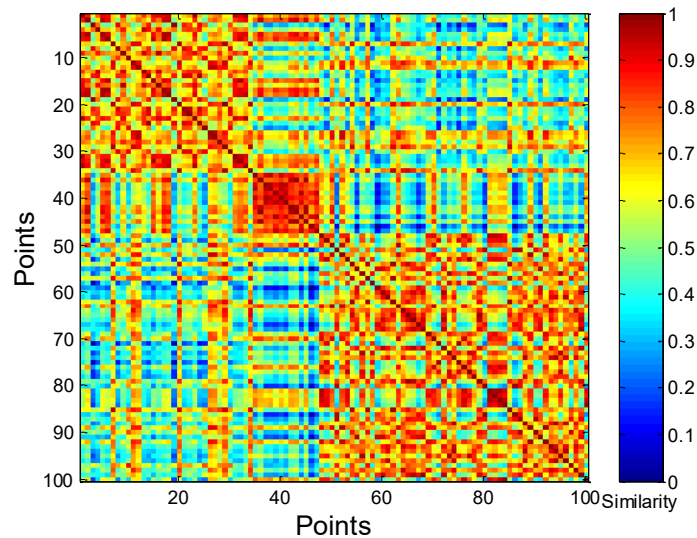
- Clusters in random data are not so crisp



**K-means**

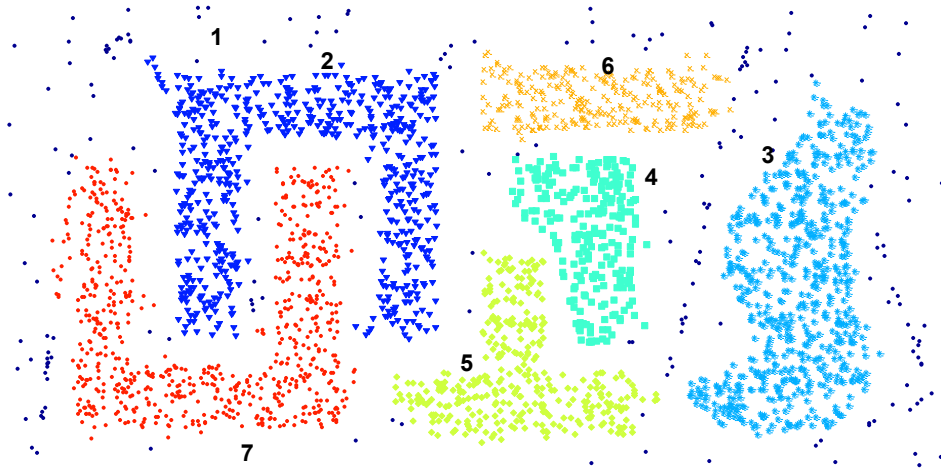
# Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

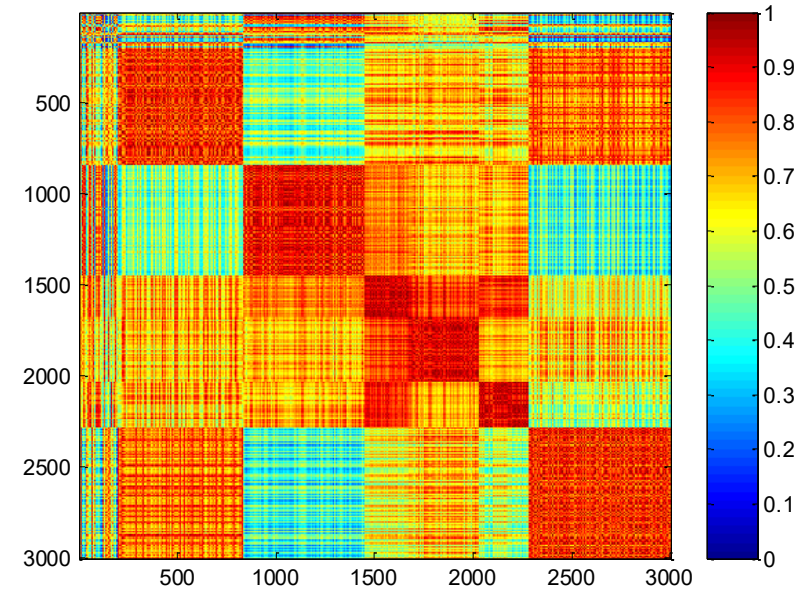


**Complete Link**

# Using Similarity Matrix for Cluster Validation

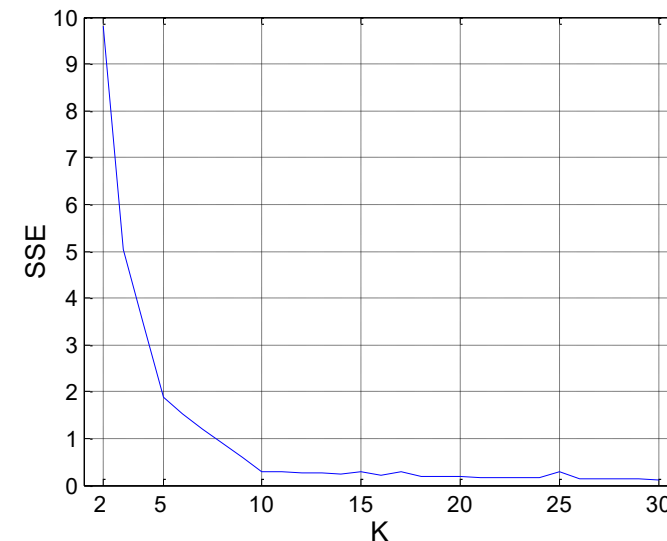
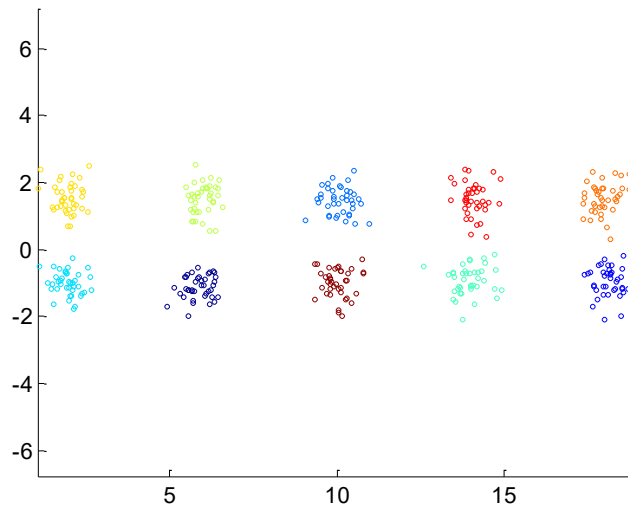


**DBSCAN**



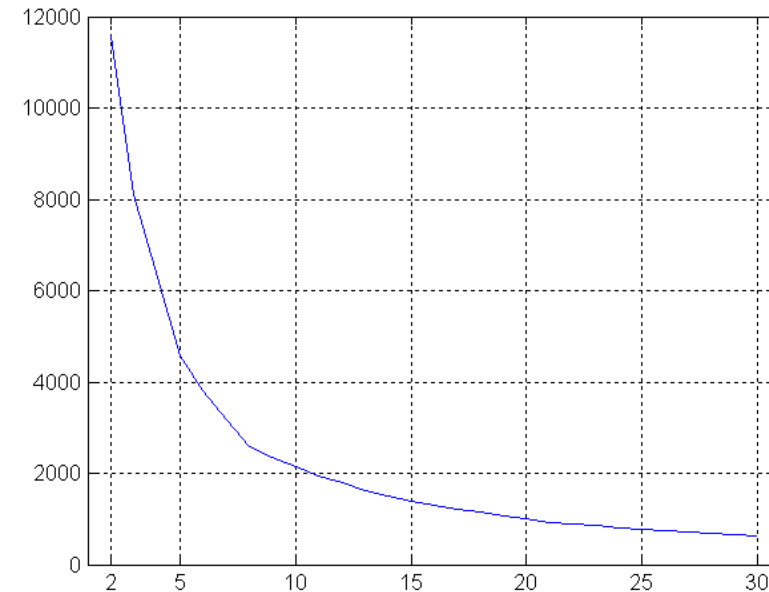
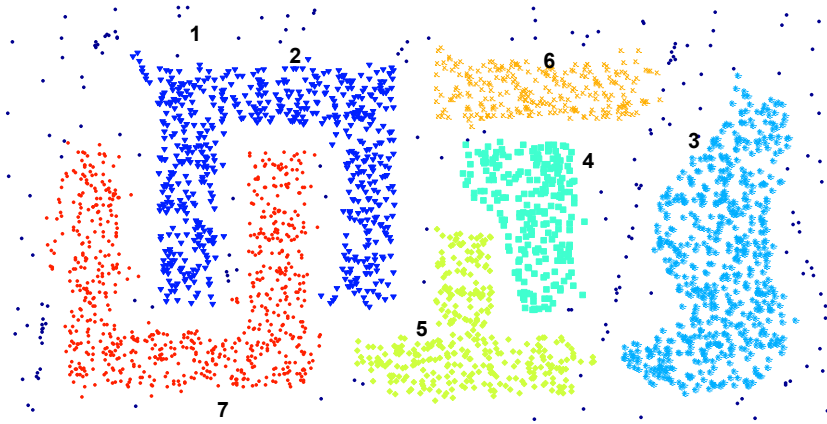
# Internal Measures: SSE

- Clusters in more complicated figures aren't well separated
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information
  - SSE
- SSE is good for comparing two clusterings or two clusters (average SSE).
- Can also be used to estimate the number of clusters



# Internal Measures: SSE

- SSE curve for a more complicated data set



**SSE of clusters found using K-means**

# Internal Measures: Cohesion and Separation

---

- **Cluster Cohesion**: Measures how closely related are objects in a cluster
  - Example: SSE
- **Cluster Separation**: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE)

$$SSE = WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

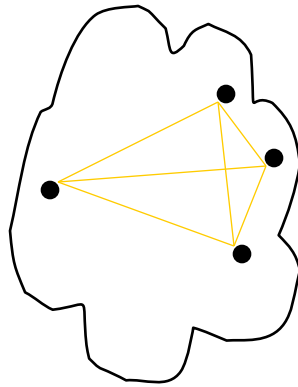
- Separation is measured by the between cluster sum of squares

$$BSS = \sum_i |C_i| (m - m_i)^2$$

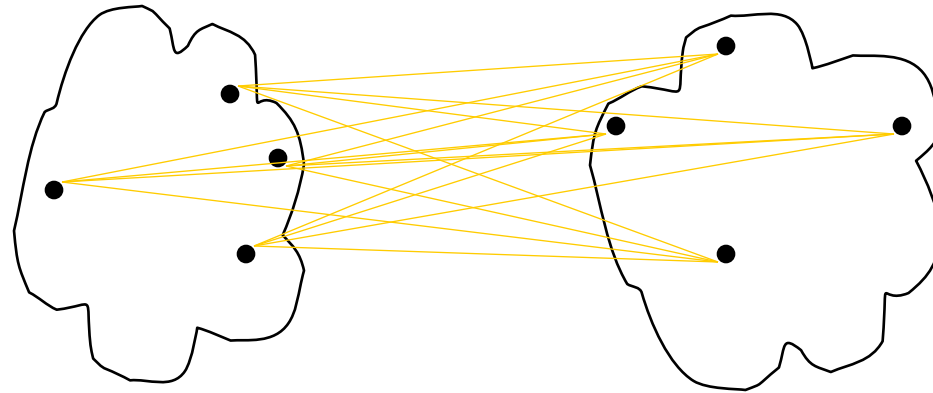
- Where  $|C_i|$  is the size of cluster  $i$

# Internal 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

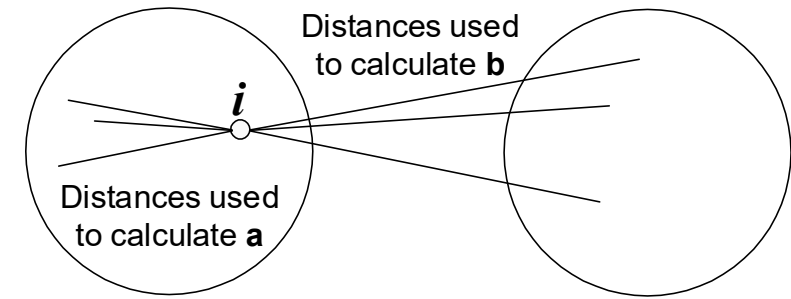


separation



# Internal 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,  $i$ 
  - Calculate  $a$  = average distance of  $i$  to the points in its cluster
  - Calculate  $b$  = min (average distance of  $i$  to points in another cluster)
  - The silhouette coefficient for a point is then given by
$$s = (b - a) / \max(a, b)$$
  - Typically between 0 and 1.
  - The closer to 1 the better.



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

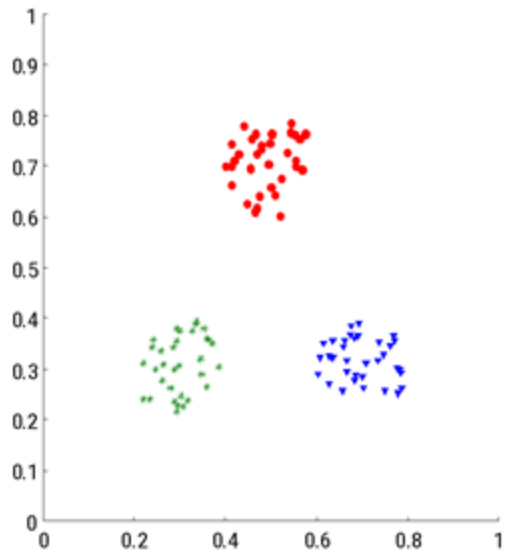
# Assessing the significance of Cluster Validity Measures

---

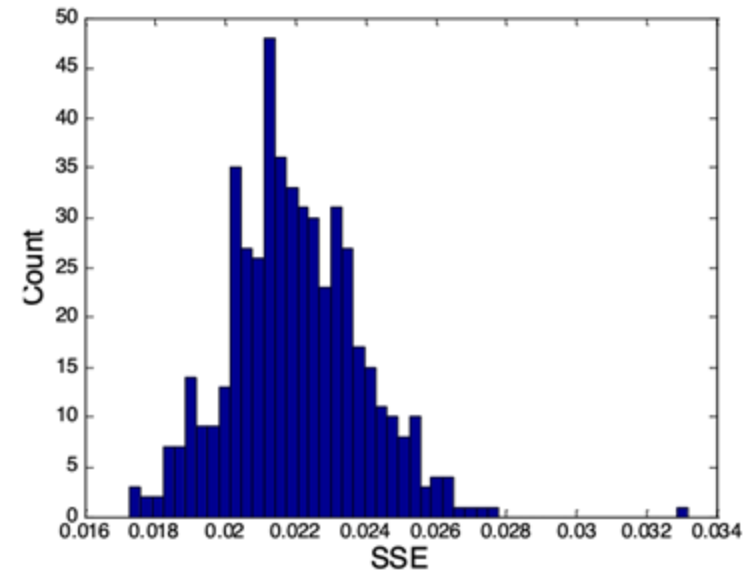
- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more “atypical” a clustering result is, the more likely it represents valid structure in the data
  - 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



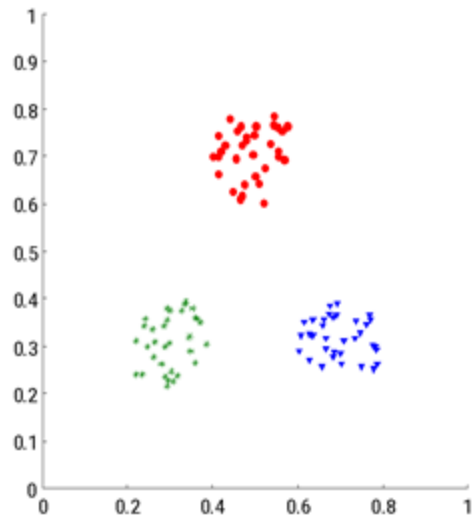
**SSE = 0.005**



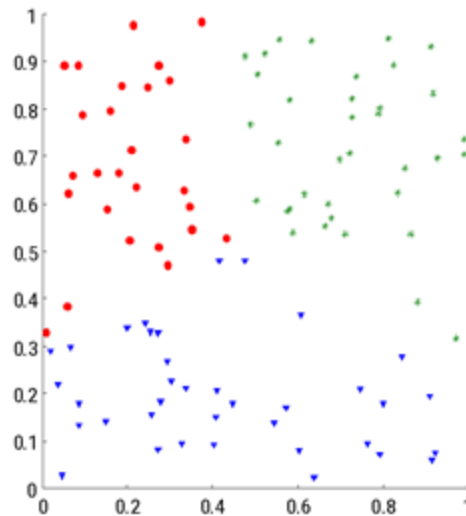
Histogram shows SSE of three clusters in 500 sets of random points of size 100 distributed over the range 0.2 - 0.8 for x and y values.

# Statistical Framework for Correlation

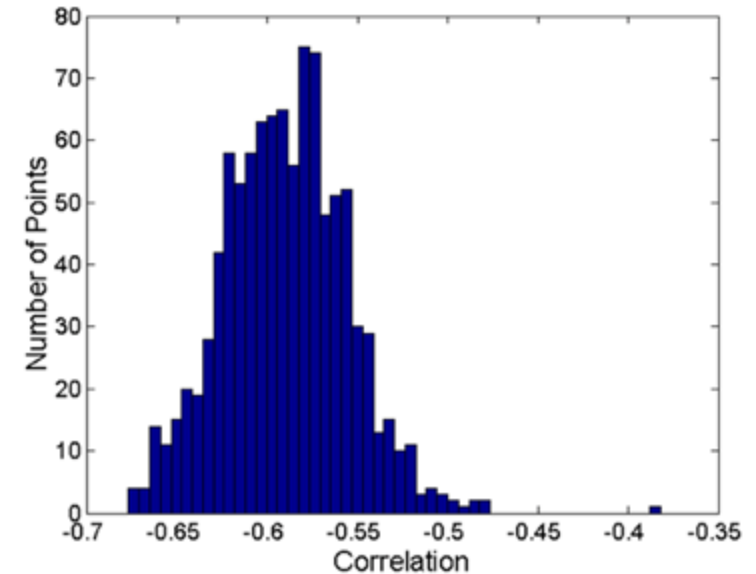
- Correlation of ideal similarity and proximity matrices for the K-means clustering of the following two datasets.



**Corr = -0.9235**



**Corr = -0.5810**



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

# Final Comment on Cluster Validity

---

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

*Algorithms for Clustering Data*, Jain and Dubes

# References

---

- Clustering. Chapter 7. Introduction to Data Mining.

