Data Mining Functionalities

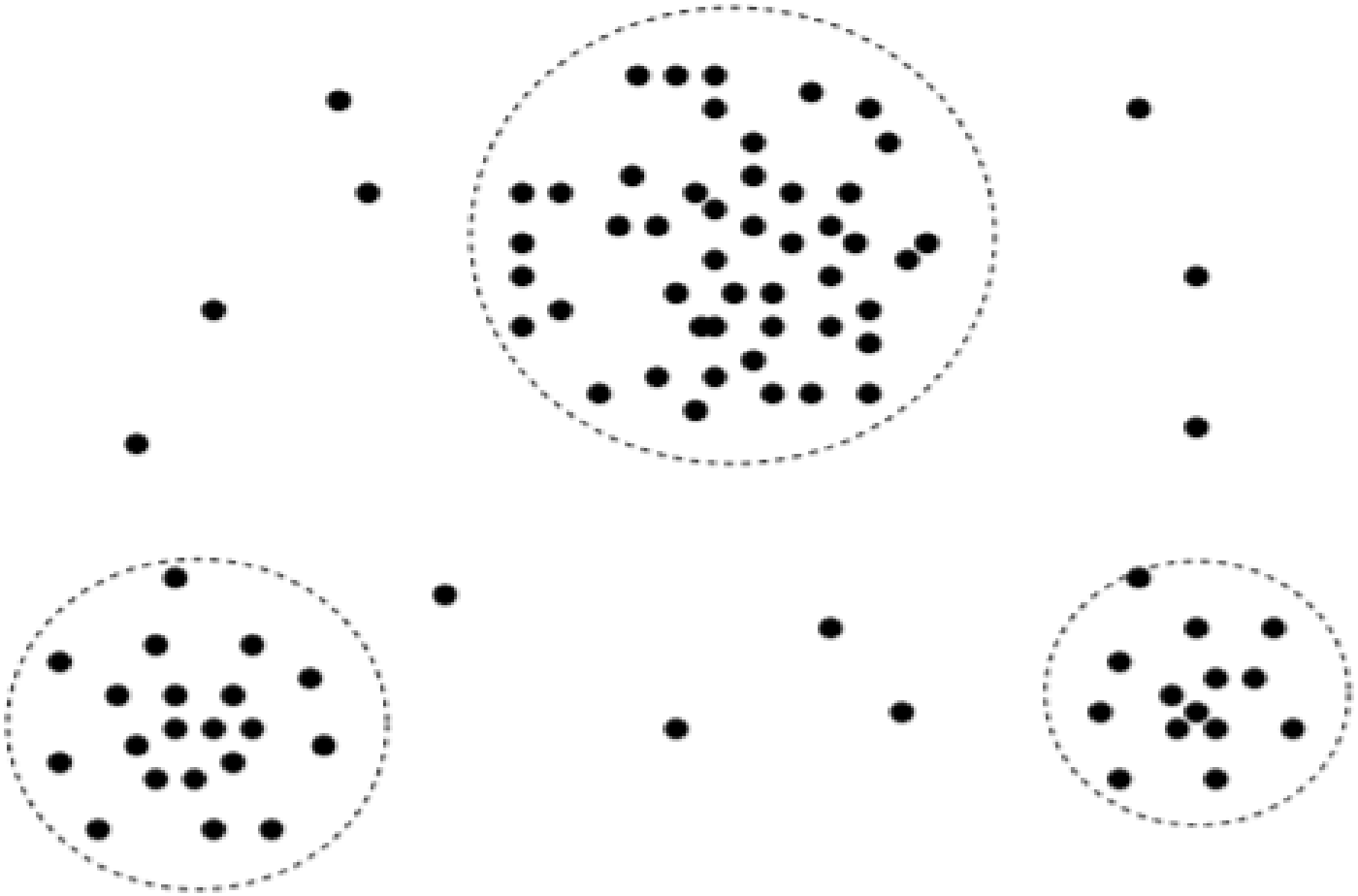
Cluster Analysis

Clustering analyzes data objects without consulting class labels.

"The objects are clustered or grouped based on the principle of maximizing the intraclass(or cluster) similarity and minimizing the interclass (or cluster) similarity"

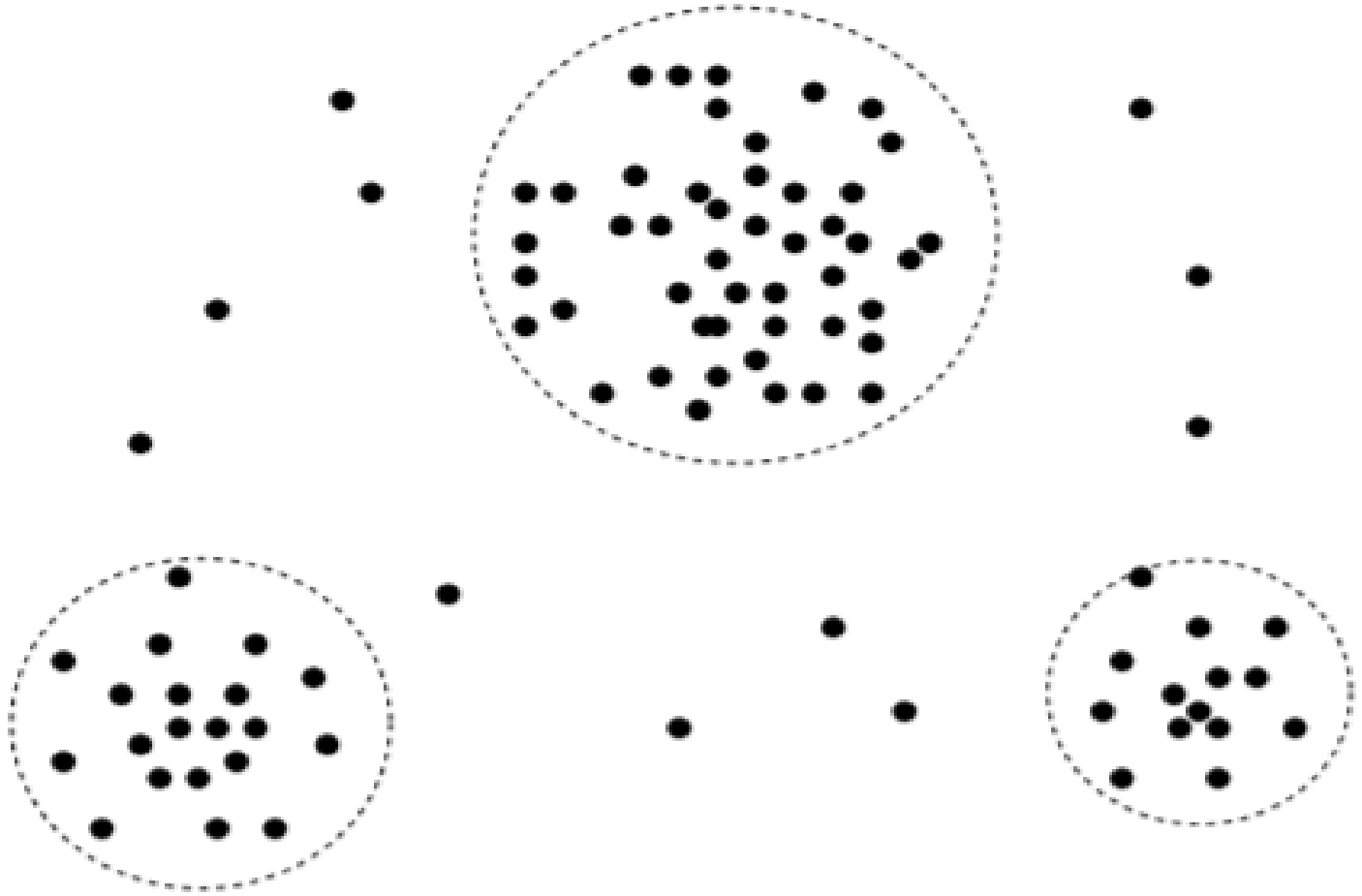
Data Mining Functionalities

Cluster Analysis



Data Mining Functionalities

Outlier Analysis : Objects do not comply with the general behavior of the system or existing clusters- those objects are referred to as outlier



Data Mining Functionalities

Example of Cluster Analysis:

Input: Data Matrix
Finding: Either Dissimilarity or Similarity with Euclidean Distance

$$D = \begin{matrix} 80 & 80 & 80 \\ 70 & 70 & 70 \\ 60 & 60 & 60 \\ 90 & 80 & 90 \\ 75 & 70 & 60 \\ 60 & 62 & 61 \end{matrix}$$

Data Mining Functionalities

Procedure for Cluster Analysis:

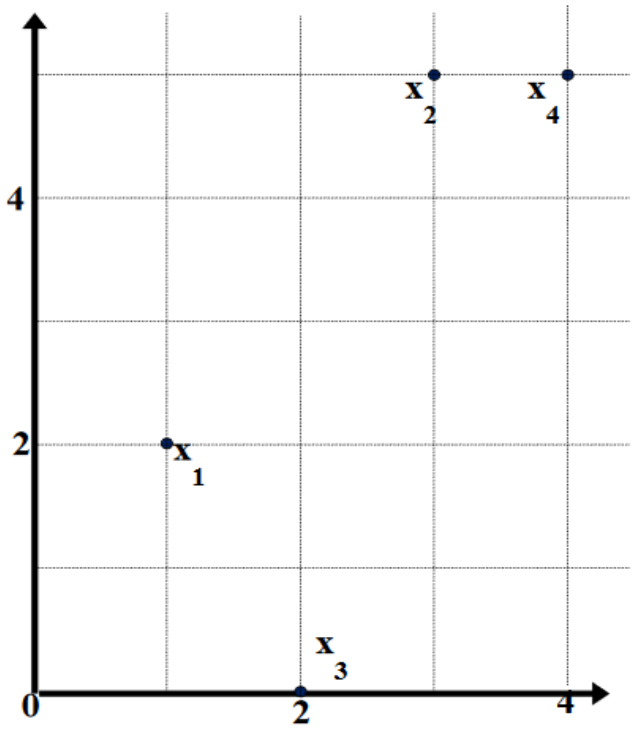
$$\text{Dissimilarity Matrix DM} = \begin{bmatrix} d(1,1) & d(1,2) & d(1,3) & d(1,4) & d(1,5) & d(1,6) \\ d(2,1) & d(2,2) & d(2,3) & d(2,4) & d(2,5) & d(2,6) \\ d(3,1) & d(3,2) & d(3,3) & d(3,4) & d(3,5) & d(3,6) \\ d(4,1) & d(4,2) & d(4,3) & d(4,4) & d(4,5) & d(4,6) \\ d(5,1) & d(5,2) & d(5,3) & d(5,4) & d(5,5) & d(5,6) \\ d(6,1) & d(6,2) & d(6,3) & d(6,4) & d(6,5) & d(6,6) \end{bmatrix}$$

Normalize the DM

Find Similarity Matrix

Similarity Matrix = 1- Dissimilarity Matrix

Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix

(with Euclidean Distance)

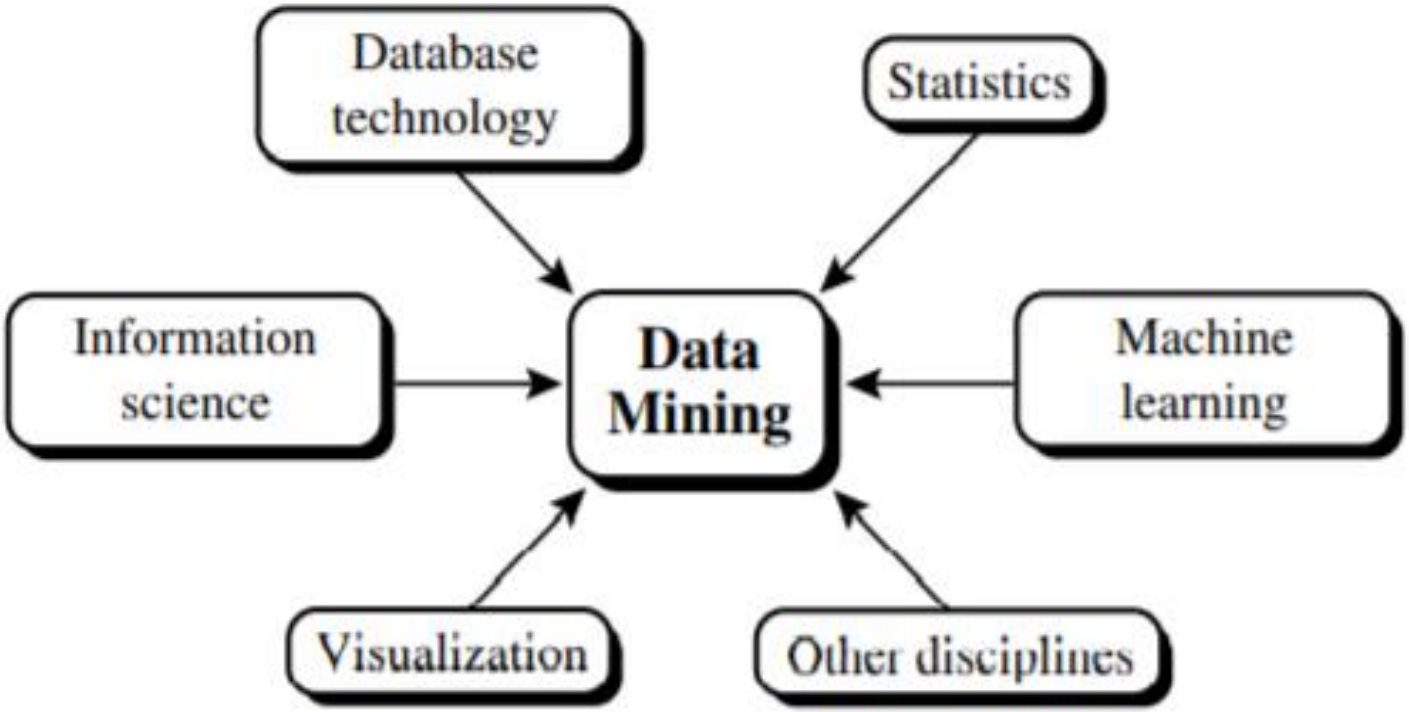
	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Possible Questions

1. Define classification.
2. Distinguish between classification and regression.
3. What are the key steps of classification process.
4. Write the sample examples of classification
5. Write an example of regression
6. Define clustering process
7. What are the important principles of clustering?
8. How the clustering process is different from classification process?
9. Describe the clustering process with an example.
10. Describe outlier analysis

Classification of Data Mining Systems

Data mining is interdisciplinary field



Data mining as a confluence of multiple disciplines.

Classification of Data Mining Systems

Classification according to the kinds of databases mined

- database systems are classified according to different criteria
- data mining systems can therefore be classified accordingly

Classification according to the kinds of knowledge mined

- Data mining functionalities
- Based on granularity or abstraction of the knowledge mined

Classification according to the kinds of techniques utilized

- Degree of user interaction involved
- Methods of data analysis employed

Classification according to the applications adapted

- Data mining systems can also be categorized according to the applications they adapt

Data Mining Task Primitives

Data Mining Query is defined interms of data mining task primitives.

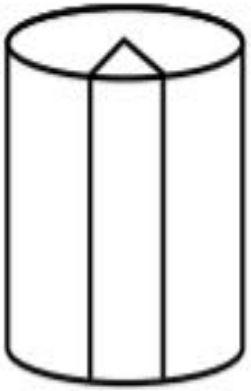
- **allow the user**
- **communicate with data mining system**
- **examine the findings**

1. **Task Relevant Data to be Mined**
2. **Kind of Knowledge to be Mined**
3. **Background Knowledge to be used in the Discovery Process**
4. **Interestingness Measures and Thresholds for Pattern Evaluation**
5. **The Expected Representation for Visualizing the Discovered Patterns**

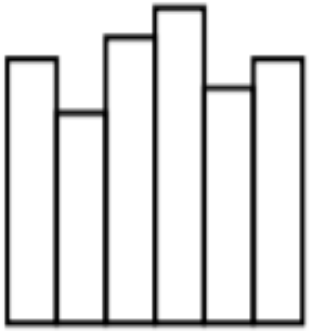
Data Mining Task Primitives

1. Task Relevant Data to be Mined
2. Kind of Knowledge to be Mined
3. Background Knowledge to be used in the Discovery Process
4. Interestingness Measures and Thresholds for Pattern Evaluation
5. The Expected Representation for Visualizing the Discovered Patterns

Task Relevant Data to be Mined & Kind of Knowledge to be Mined

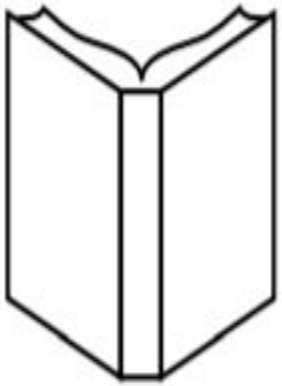


Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria



Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering

Background Knowledge & Interestingness Measures



Background knowledge
Concept hierarchies
User beliefs about relationships in the data

Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty

Visualizing the Discovered Patterns

Visualization of discovered patterns
Rules, tables, reports, charts, graphs, decision trees, and cubes
Drill-down and roll-up

Thank You