Data Mining: Unveiling Knowledge from Data

Module No.	Topic	No. of Lectures
1	Introduction	1
2	Data, measurements, and data preprocessing	3
3	Data warehousing and online analytical processing	3
4	Pattern mining: basic concepts and methods	1
5	Pattern mining: advanced methods	3
6	Classification: basic concepts and methods	3
7	Classification: advanced methods	3
8	Cluster analysis: basic concepts and methods	2
9	Cluster analysis: advanced methods	2
10	Deep learning	3
11	Outlier detection	2
12	Data mining trends and research frontiers	2

Course Details

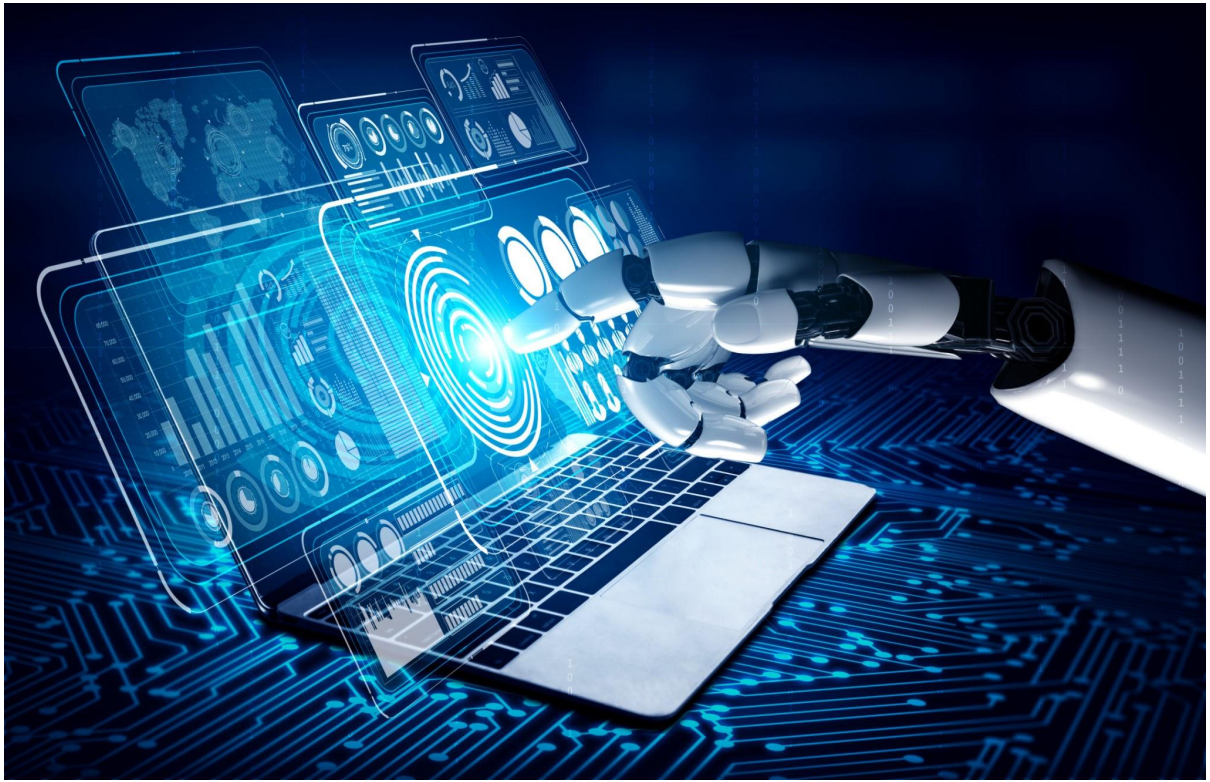
L-T-P: 2-0-3

Weightage:

- Mid Sem: 20
- End Sem: 50
- Lab: 20
- Quiz: 10

TextBook: Data Mining Concepts and Techniques 4th edition Han, Pei, and Tong

Content



Part 01　Introduction & Objectives

Data Mining Overview

Benefits of Data Mining

Data mining enables uncovering hidden trends; improves decision-making through insights;supports predictive analysis for future trends and helps optimize processes and resource allocation across various industries.

Data Mining Definition

Data mining is the automated or semi-automated process of discovering meaningful patterns and knowledge from large volumes of data and it goes beyond simple retrieval or querying, involving inference, pattern discovery, and predictive modeling.

Learning Objectives

Objective 1: Defining Data Mining

Define data mining and its role in knowledge discovery; understand how it transforms raw data into actionable intelligence; and know its relationship with machine learning.

Objective 2: Identifying Data Types

Identify various data types and knowledge mined in data mining: recognizing the diversity of data forms, from structured tables to unstructured streams; and exploring the different types of knowledge extraction.

Objective 3: Understanding Interdisciplinary Foundations

Understand data mining's dependency on statistics, machine learning, and database technology; show how these disciplines contribute to methodologies; and highlight interdisciplinary collaboration for innovative solutions.

Objective 4: Recognizing Applications and Societal Implications

Recognize applications and societal implications, including real-world examples; consider its impact on privacy; and promote responsible practices to mitigate potential adverse effects.

Part 02 What is Data Mining?

Definition of Data Mining

The Essence of Data Mining

Data mining is the automated or semi-automated process of extracting previously unknown, valid, and actionable patterns from large datasets.

Analogy: Mining for Gold

Like mining gold from ore data mining sifts through vast amounts of raw data to find valuable nuggets of information; it demands a systematic approach to unearth patterns effectively.

Core Functionalities

It goes beyond simpler retrieval or querying; it involves inference, pattern discovery, and predictive modeling; with these core functionalities, data mining facilitates enhanced decision making.

Example 1.1. Data mining turns a large collection of data into knowledge. A search engine (e.g., Google) receives billions of queries every day. What novel and useful knowledge can a search engine learn from such a huge collection of queries collected from users over time? Interestingly, some patterns found in user search queries can disclose invaluable knowledge that cannot be obtained by reading individual data items alone. For example, Google's *Flu Trends* uses specific search terms as indicators of flu activity. It found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms. A pattern emerges when all of the search queries related to flu are aggregated. Using aggregated Google search data, *Flu Trends* can estimate flu activity up to two weeks faster than what traditional systems can.[1] This example shows how data mining can turn a large collection of data into knowledge that can help meet a current global challenge. □

Example 1.2. Association analysis. Suppose that, a webstore manager wants to know which items are frequently purchased together (i.e., in the same transaction). An example of such a rule, mined from the transactional database, is

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}webcam\text{”}) \quad [support = 1\%, confidence = 50\%],$$

where X is a variable representing a customer. A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy webcam as well. A 1% **support** means

that 1% of all the transactions under analysis show that computer and webcam are purchased together. This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as “*computer* $\Rightarrow$ *webcam* [1%, 50%].”

Suppose, mining the same database generates another association rule:

$$age(X, \text{“}20..29\text{”}) \wedge income(X, \text{“}40K..49K\text{”}) \Rightarrow buys(X, \text{“laptop”})$$
$$[support = 0.5\%, confidence = 60\%].$$

The rule indicates that of all its customers under study, 0.5% are 20 to 29 years old with an income of $40,000 to $49,000 and have purchased a laptop (computer). There is a 60% probability that a customer in this age and income group will purchase a laptop. Note that this is an association involving more than one attribute or predicate (i.e., *age, income*, and *buys*). Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule**. □

Part 03 Data Mining in the Knowledge Discovery Process

Knowledge Discovery in Databases (KDD)

KDD Process Steps

Data mining is one step in the broader KDD process, encompassing data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation.

Importance of Preprocessing

Without proper preprocessing, data mining can produce misleading results; thus, preprocessing is a basic step for enhancing data quality and validity.

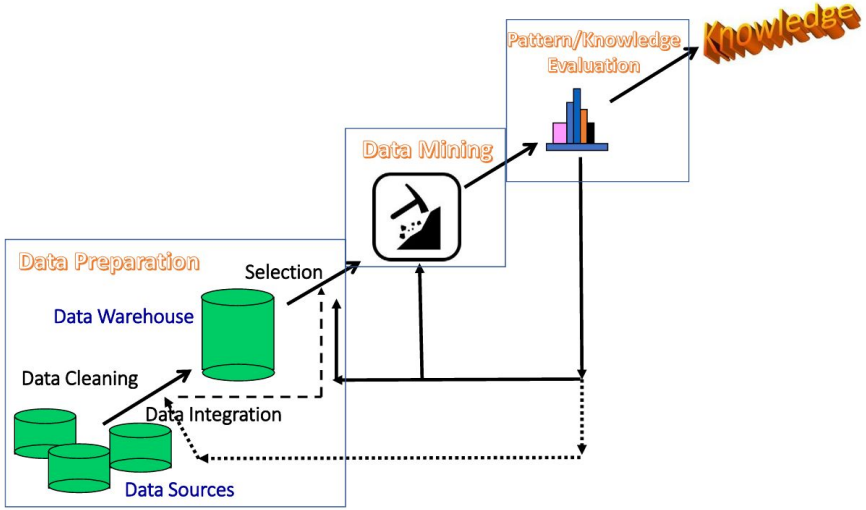


FIGURE 1.1

Data mining: An essential step in the process of knowledge discovery.

Part 04 Diversity of Data Types

Definition and Types of Structured Data

Tables, databases, and spreadsheets characterize structured data; it can be easily organized and analyzed using traditional methods.

Definition and Examples of Semi-structured Data

XML, JSON, and emails are semi-structured data, exhibiting some organizational properties; its flexible nature requires flexible parsing for effective analysis.

Definition and Examples of Unstructured data

Text, images, audio, and video exemplify unstructured data, lacking predefined formats; advanced methods like NLP and image processing are essential for meaningful interpretation.

Definition and Key Characteristics of Big Data

Volume, velocity, variety, and veracity characterize big data; its scale and complexity call for distributed computing and advanced algorithms.

Definition and Relevance of Spatial and Temporal Data

GPS and time-series fall under spatial and temporal data; their dependencies on location and time necessitate specialized methods.

Definition and Applications of Streaming Data

Real-time sensor or social media feeds constitute streaming data, dynamically arriving and requiring immediate processing; its velocity and volume mandate efficient, real-time analytical approaches.

Part 05 Types of Knowledge Mined

Multidimensional Data Summarization

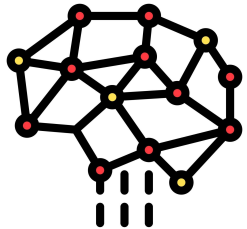
The Function of Data Summarization

Summarizing data along dimensions (e.g., sales by region, time, product) enables understanding; often uses OLAP (Online Analytical Processing) and aggregation for data summarization.

Mining Frequent Patterns, Associations, and Correlations

Uncovering Relationships

Example: Market basket analysis (e.g., "Customers who buy bread often buy milk."); Metrics: Support, confidence, lift are key for Mining Frequent Patterns.



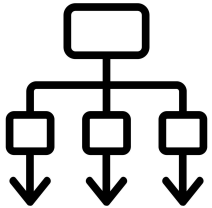
Apriori algorithm

Apriori algorithm is a classic method and association rule mining, enabling efficient identification of frequent itemsets.

Classification and Regression for Predictive Analysis

Classification

Classification predicts categorical labels (e.g., spam vs. not spam); regression predicts numerical values (e.g., house prices).



Types of Algorithms

Algorithms: Decision trees, logistic regression, and neural networks each have unique strengths in the predictive analysis.

age(X, "youth") AND income(X, "high") ⟶ *class(X, "A")*

age(X, "youth") AND income(X, "low") ⟶ *class(X, "B")*

age(X, "middle_aged") ⟶ *class(X, "C")*

age(X, "senior") ⟶ *class(X, "C")*

(a)

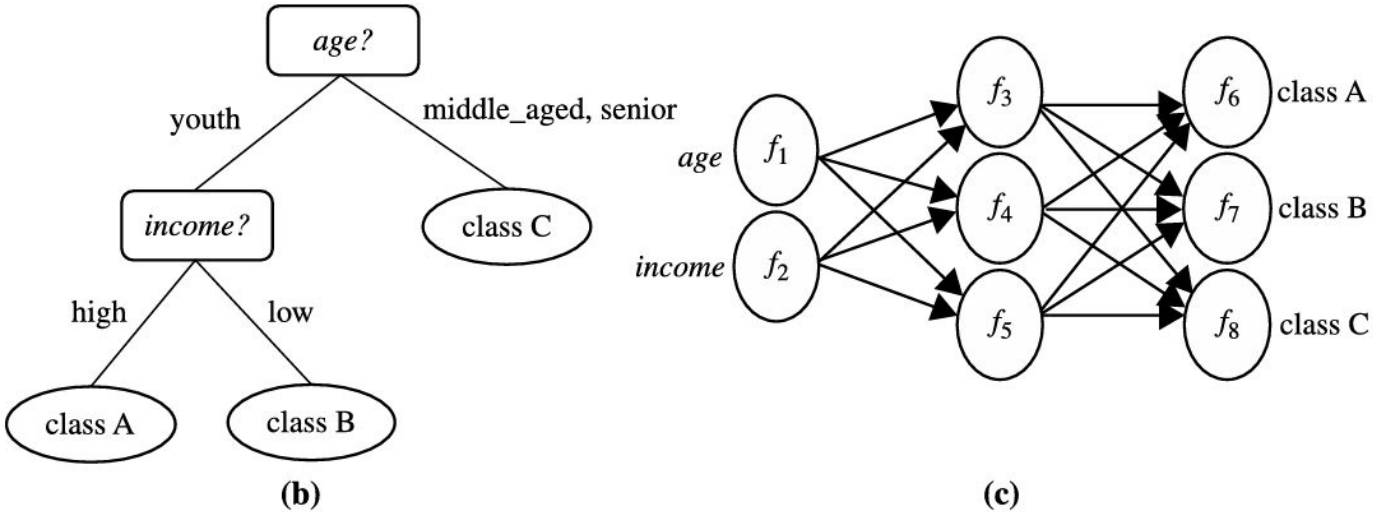


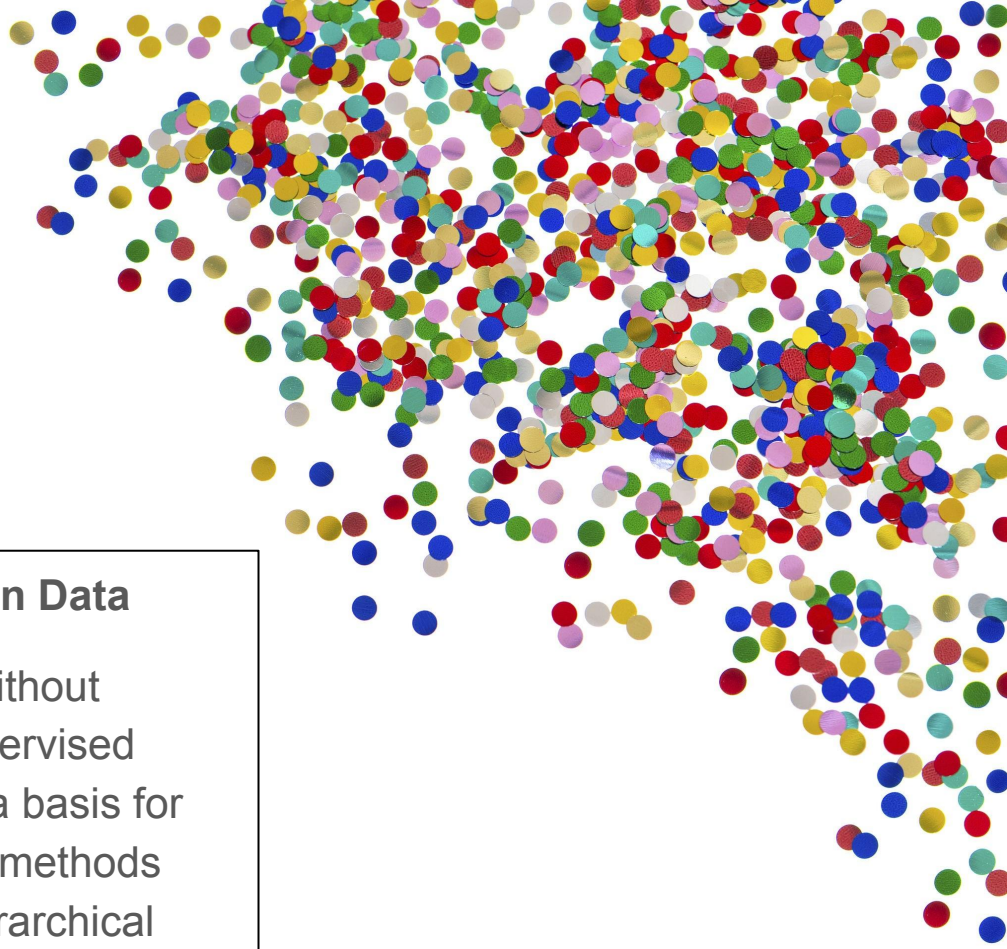
FIGURE 1.2

A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Example 1.3. Classification and regression. Suppose a webstore sales manager wants to classify a large set of items in the store, based on three kinds of responses to a sales campaign: *good response*, *mild response*, and *no response*. You want to derive a model for each of these three classes based on the descriptive features of the items, such as *price*, *brand*, *place_made*, *type*, and *category*. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

Suppose that the resulting classification is expressed as a decision tree. The decision tree, for instance, may identify *price* as being the first important factor that best distinguishes the three classes. Other features that help further distinguish objects of each class from one another include *brand* and *place_made*. Such a decision tree may help the manager understand the impact of the given sales campaign and design a more effective campaign in the future.

Suppose instead, that rather than predicting categorical response labels for each store item, you would like to predict the amount of revenue that each item will generate during an upcoming sale, based on the previous sales data. This is an example of regression analysis because the regression model constructed will predict a continuous function (or ordered value.) □



Cluster Analysis

Discovering Structure in Data

Groups similar objects without predefined labels (unsupervised learning) and serves as a basis for customer segmentation; methods include k-means and hierarchical clustering.

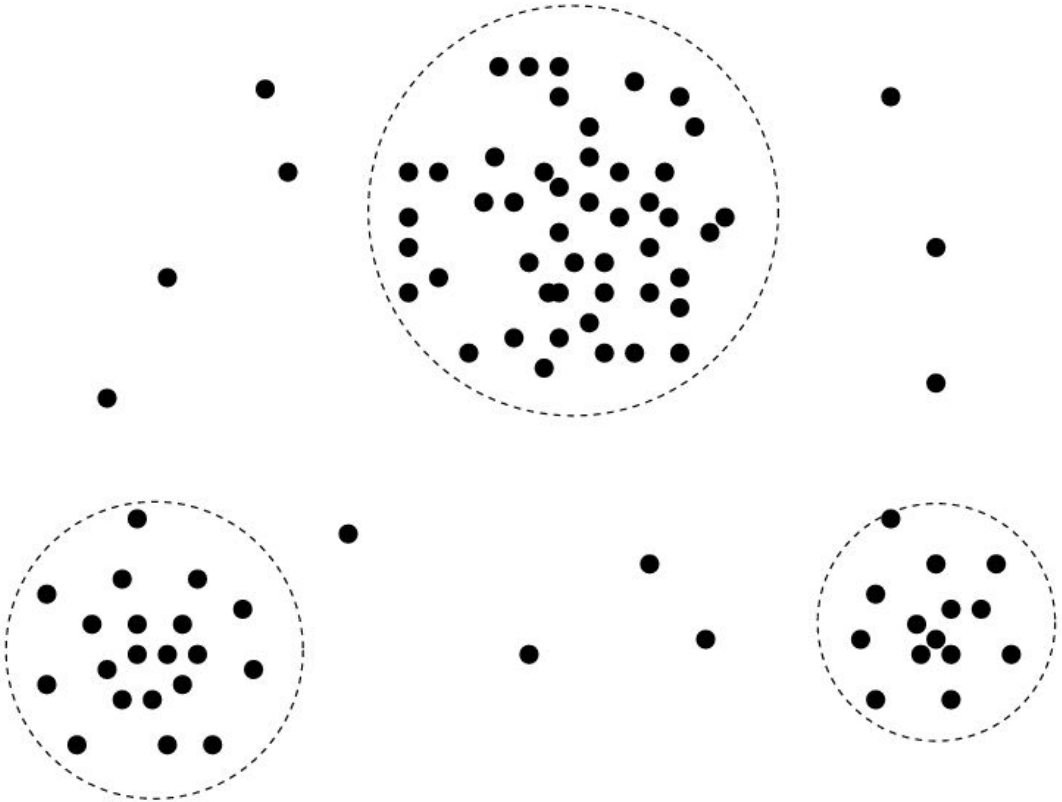


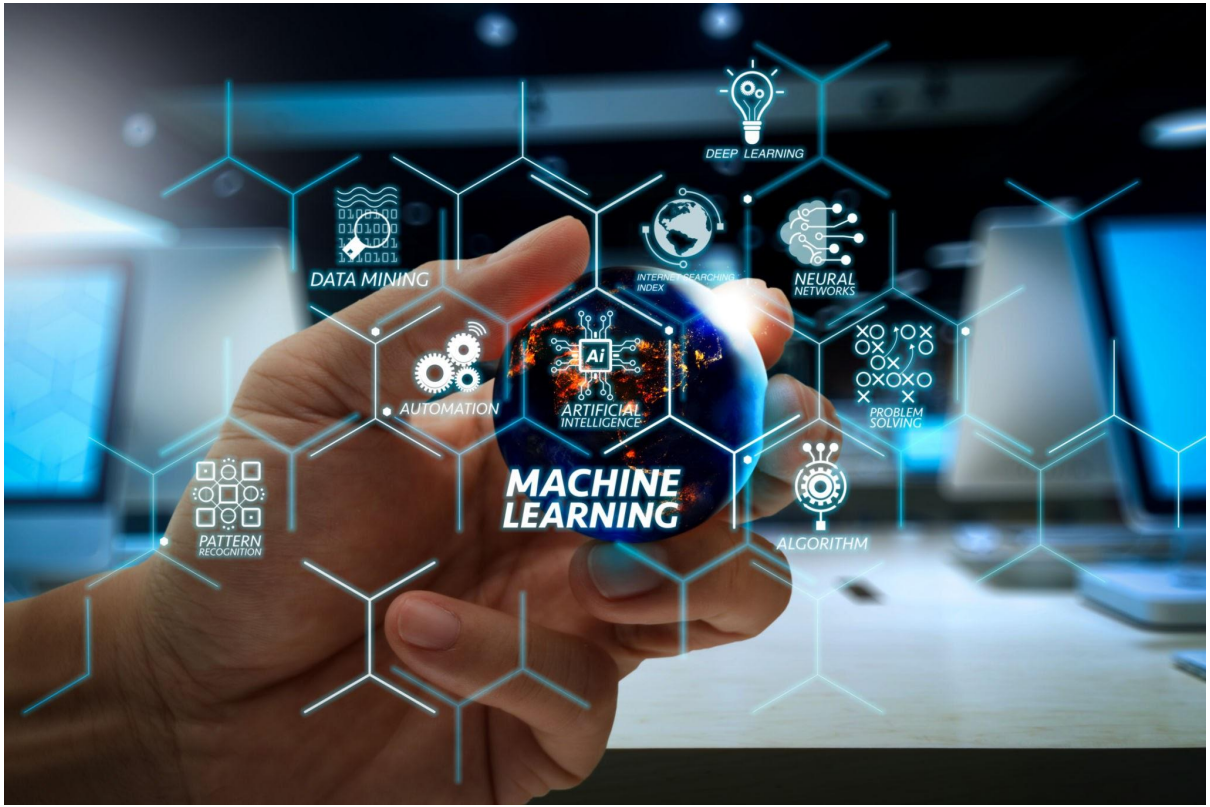
FIGURE 1.3

A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

Deep Learning

Revolutionizing Automated Feature Extraction

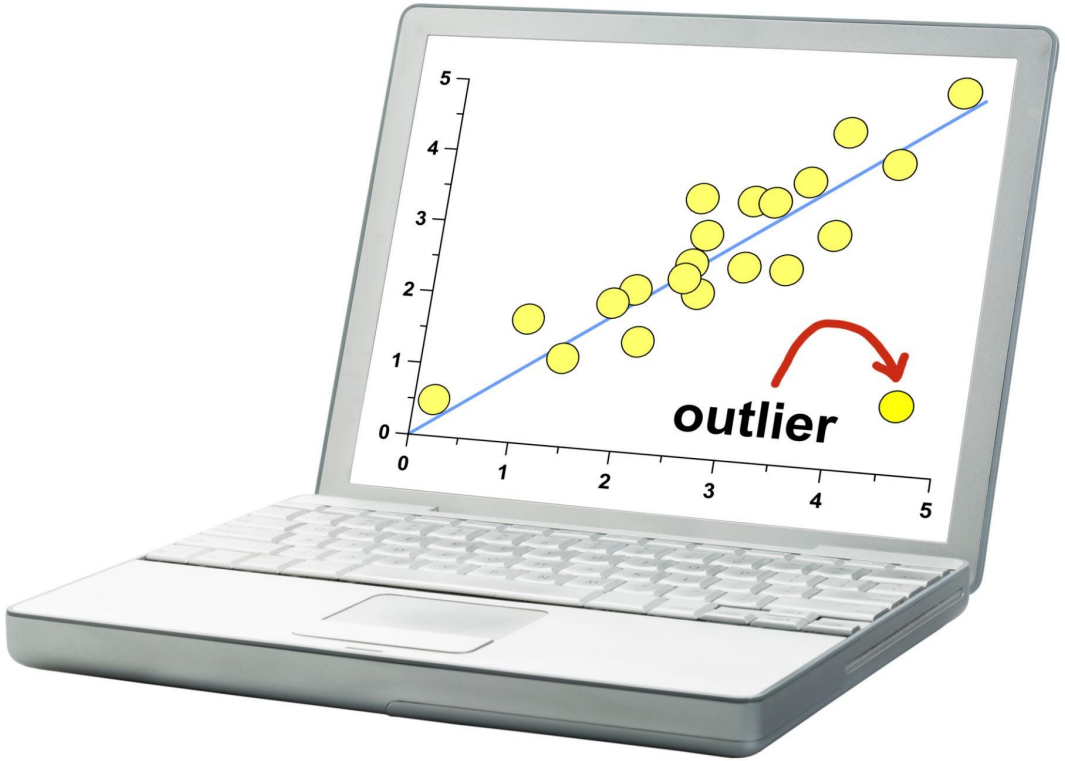
A subset of machine learning uses multi-layered neural networks; excels in image, speech, and text data; it highlights its growing role in automated feature extraction.



Outlier Analysis

Detecting Anomalies

Detecting anomalies or unusual patterns (e.g., fraud detection, network intrusion) is critical for security and quality control and important to prevent and solve issues.



Example 1.5. Outlier analysis. Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency. □

Are All Mining Results Interesting?

No, patterns must be valid, novel, useful, and understandable;
Interestingness measures: Objective (statistical significance) and subjective (user relevance).

Part 06 Confluence of Multiple Disciplines

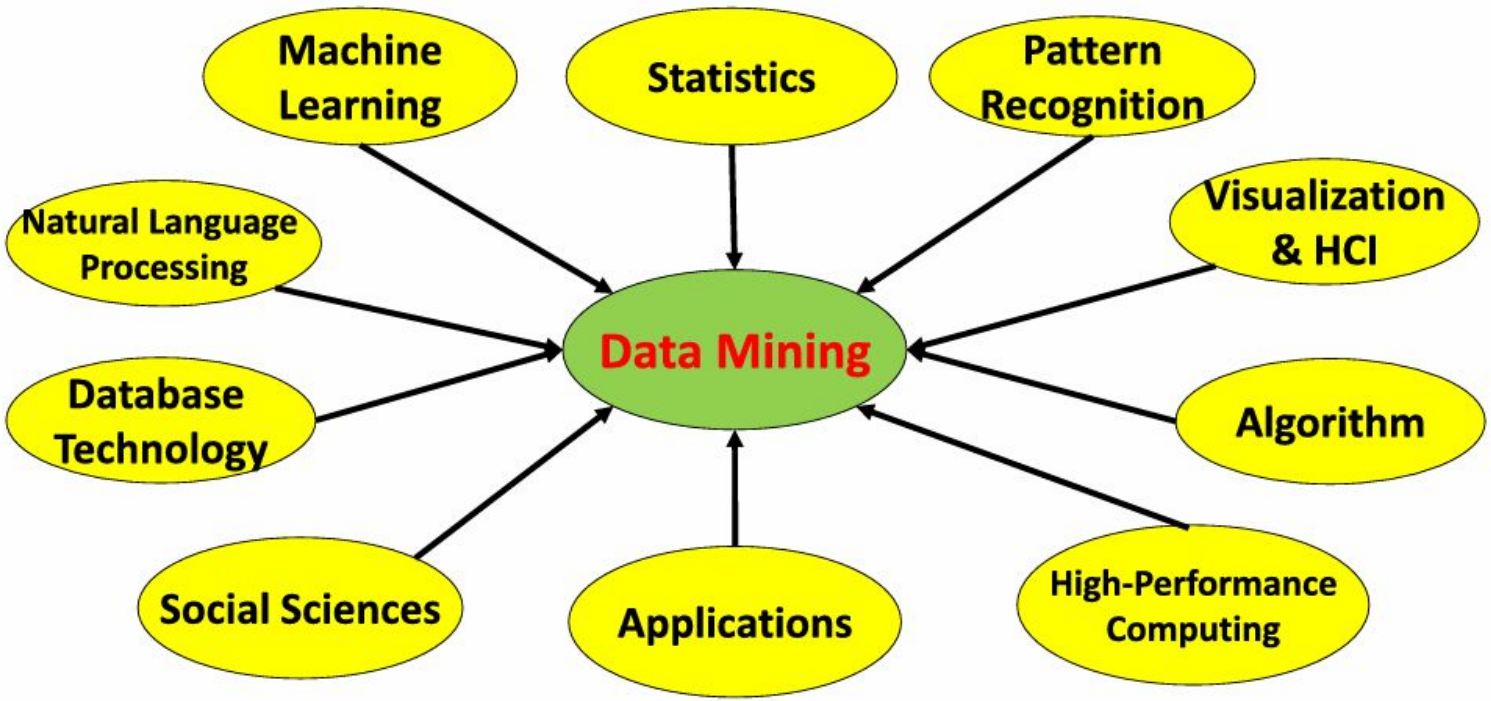


FIGURE 1.4

Data mining: Confluence of multiple disciplines.

Statistics and Data Mining

Statistical foundations

Provides foundational techniques: regression, hypothesis testing, distributions; fundamental to ensure the validity and reliability of findings.

Machine Learning and Data Mining

Algorithm Design

ML focuses on algorithm design for learning from data; data mining emphasizes the entire process and scalability and guarantees comprehensive solutions.

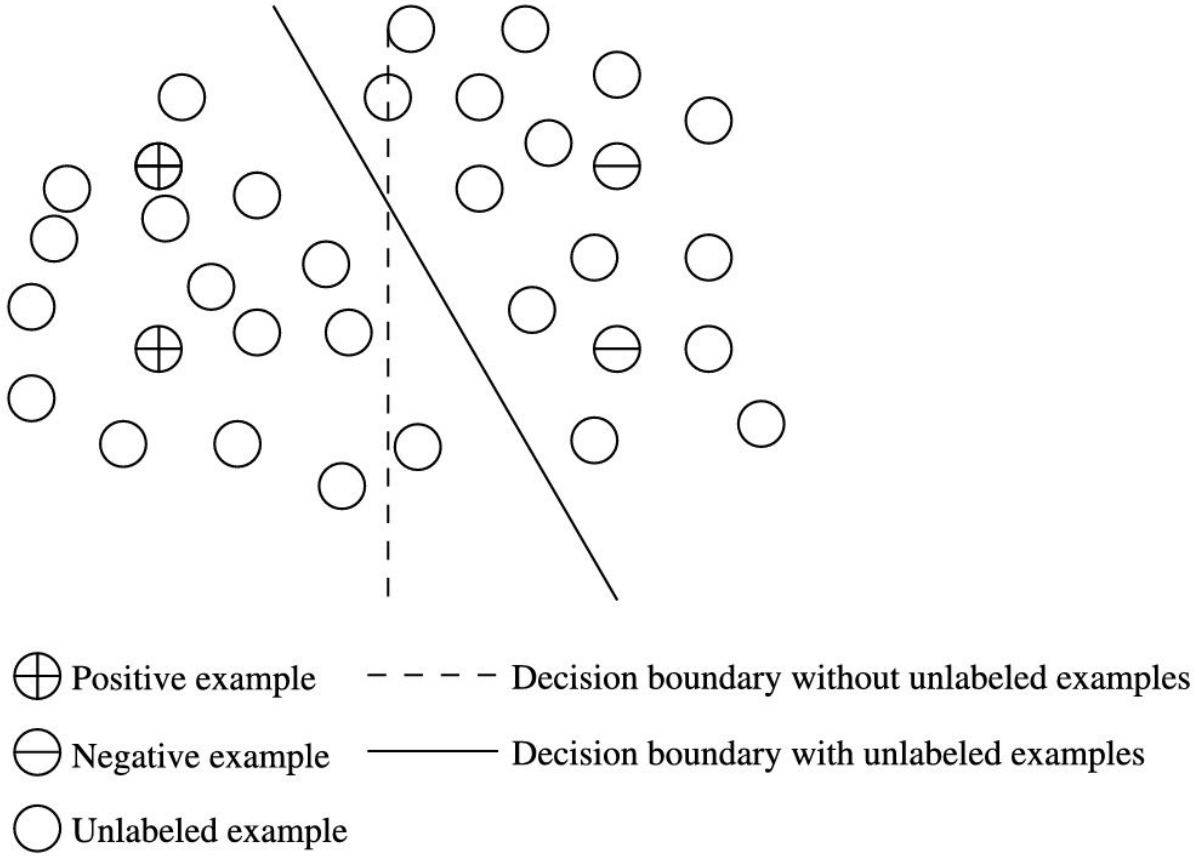


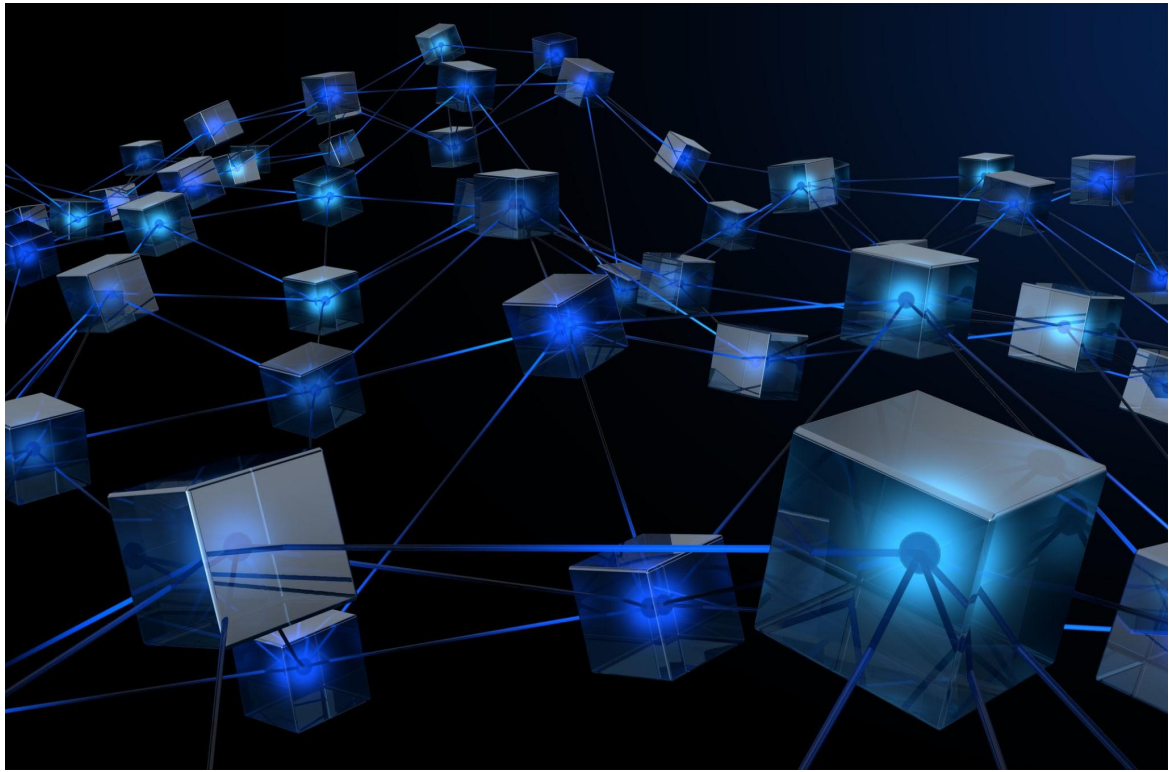
FIGURE 1.5

Semisupervised learning.

Database Technology and Data Mining

Enables Efficient Storage

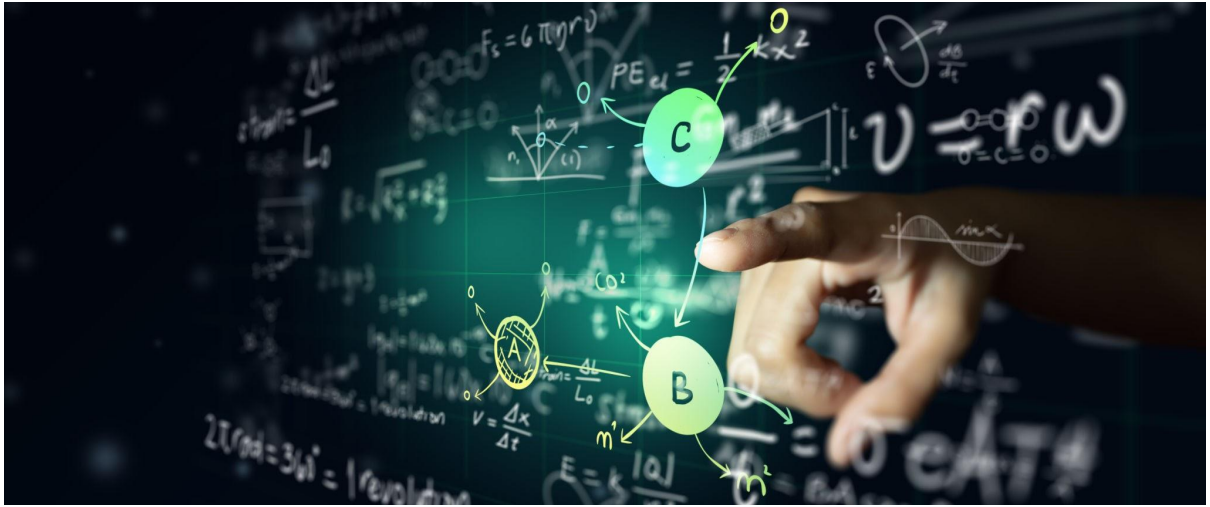
Database systems enable efficient storage, querying, and processing of large datasets (SQL, NoSQL); they enable the data accessing and processing capabilities.



Data Mining and Data Science

Includes Data Engineering

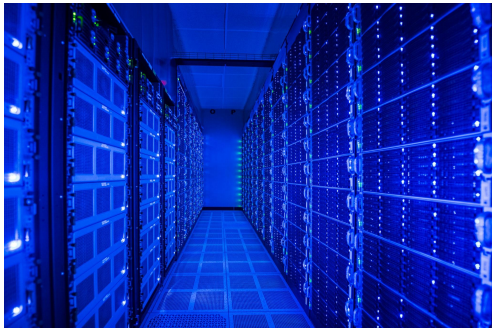
Data science is broader: includes data engineering, visualization, storytelling, and business analysis; data mining is a core analytical component and integral to actionable insights.



Other Disciplines

Optimization

Includes optimization, visualization, high-performance computing, and domain expertise (e.g., bioinformatics, finance); promote comprehensive and detailed insights.



Part 07 Applications and Societal Impact

Data Mining Applications

Business Applications

Customer relationship management, recommendation systems are included in business applications for improving sales, and service; science includes genomics and astronomy.

Data Mining in Science

Data mining is critical for revealing data in genomics and astronomy in order to find and solve hidden patterns in science.

Healthcare and Data Mining

Disease outbreak prediction and patient diagnosis are included; these are basic requirements for healthcare and need to be improved; web and social media includes sentiment analysis and community detection.

Data Mining and Society

Ethical Concerns in Society

Ethical concerns: Privacy, bias in algorithms, and discrimination; responsible data mining needs transparency, fairness, and accountability.

Transparency, Fairness, and Accountability

Regulations: GDPR and HIPAA impact how data is collected and used; promote transparency, fairness, and accountability to minimize risks with transparency, fairness, and accountability.