MODULE 4

5.2 DATA MINING CONCEPTS AND APPLICATIONS

Definitions

- Data mining is a term used to describe discovering or "mining" knowledge from large amounts of data.
- The term "data mining" is a misnomer; analogous to "gold mining," the focus is on extracting valuable knowledge rather than the irrelevant parts (rock or dirt).
- Other names for data mining include knowledge extraction, pattern analysis, data archaeology, information harvesting, pattern searching, and data dredging.
- Technically, data mining is a process that uses statistical, mathematical, and artificial intelligence techniques to extract and identify useful information and subsequent knowledge (or patterns) from large datasets.
- Patterns discovered can include business rules, affinities, correlations, trends, and prediction models.

Definitions by Literature

- Fayyad et al. (1996): Data mining is "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases."

Explanation of Key Terms

- **Process**: Data mining involves iterative steps, not a single action.
- **Nontrivial**: Involves experimentation and inference; not straightforward computation.
- **Valid**: Patterns must hold true on new data with sufficient certainty.
- **Novel**: Patterns should not already be known in the given system context.
- **Potentially Useful**: Patterns must offer benefits to the user or task.
- **Ultimately Understandable**: Patterns should make logical sense and provide insights to users.

Position of Data Mining

- Data mining sits at the intersection of several disciplines, including statistics, artificial intelligence, machine learning, management science, information systems, and databases.
- It leverages advancements in these fields to extract useful knowledge from large datasets.

Characteristics and Objectives of Data Mining

1. **Data Characteristics**:
    - Data are often deeply embedded in large databases, sometimes spanning several years.
    - Such data are usually cleansed and consolidated into a data warehouse, presented in various formats.
2. **Technological Environment**:
    - Data mining operates in a client/server architecture or a Web-based information system architecture.
3. **Tools and Techniques**:
    - Advanced visualization tools help uncover hidden information in corporate files and public records.
    - Mining often involves processing soft data (unstructured text) from intranets, Lotus Notes, or text files.
4. **End-User Empowerment**:
    - Data miners are often end users who use data drills and query tools to pose ad hoc questions and quickly obtain answers.
    - Discovering unexpected results requires creative thinking during analysis and interpretation.
5. **Integration with Other Tools**:
    - Data mining tools integrate easily with spreadsheets and software tools, allowing rapid analysis and deployment of findings.
6. **Parallel Processing**:
    - Due to the large volume of data and search effort, parallel processing is sometimes necessary.

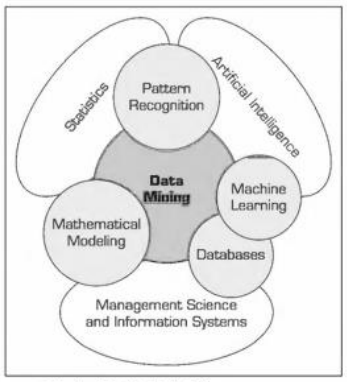


FIGURE 5.1 Data Mining as a Blend of Multiple Disciplines.

Benefits of Data Mining

- A company that effectively uses data mining tools can achieve and maintain strategic competitive advantages.
- Data mining provides organizations with decision-enhancing environments to transform data into strategic assets.
- Organizations can exploit new opportunities by converting raw data into actionable strategies.

5.3 DATA MINING APPLICATIONS

Overview

- Data mining addresses complex business problems and opportunities effectively.
- It aims to solve urgent problems or explore new opportunities for competitive advantages.
- Applications span various industries and domains, as detailed below.

Customer Relationship Management (CRM)

- CRM extends traditional marketing by focusing on **one-on-one relationships** with customers.
- Through various interactions (e.g., sales, service, product reviews, social media), businesses gather rich customer data.
- **Applications of Data Mining in CRM**:
    1. Identify **most likely responders or buyers** for new products and services (customer profiling).
    2. Understand causes of **customer attrition** and improve retention (churn analysis).
    3. Discover time-based product/service associations to maximize customer value.
    4. Identify **profitable customers** and their needs to build loyalty and boost sales.

Banking

- Data mining enhances banking operations by:
    1. **Automating loan applications** and predicting likely defaulters.

2. Detecting fraudulent activities in **credit card and online banking transactions**.
3. Identifying product/service preferences to improve **customer value maximization**.
4. Forecasting cash flow needs for ATMs and branches to optimize returns.

Retailing and Logistics

- Retailers use data mining to:
    1. Predict sales volumes at specific locations to optimize **inventory levels**.
    2. Perform **market-basket analysis** to find sales relationships between products, improving store layouts and promotions.
    3. Forecast product consumption by **seasonal and environmental conditions** to optimize logistics.
    4. Discover product movement patterns in supply chains, especially for **perishable and time-sensitive products** using sensory and RFID data.

Manufacturing and Production

- Data mining supports manufacturers by:
    1. Predicting **machinery failures** using sensory data, enabling **condition-based maintenance**.
    2. Identifying anomalies and commonalities to optimize **manufacturing capacity**.
    3. Discovering patterns to improve **product quality**.

Brokerage and Securities Trading

- Applications include:
    1. Predicting changes in **bond prices**.
    2. Forecasting the range and direction of **stock price fluctuations**.
    3. Assessing market movements based on issues/events.
    4. Identifying and preventing **fraudulent activities** in securities trading.

Insurance

- Insurance firms apply data mining to:
    1. Forecast claims for better **business planning**.
    2. Develop **optimal rate plans** by analyzing customer and claims data.
    3. Predict which customers might buy **new policies** with special features.
    4. Identify and prevent **fraudulent claims**.

Computer Hardware and Software

- Data mining helps in:
    1. Predicting **disk drive failures**.
    2. Filtering **unwanted web content and spam emails**.
    3. Detecting and preventing **network security breaches**.
    4. Identifying **unsecure software products**.

Government and Defense

- Military and government use cases:
    1. Forecasting costs of moving **personnel and equipment**.
    2. Predicting adversary strategies for **better military planning**.
    3. Estimating **resource consumption** for budgeting.
    4. Sharing knowledge from military operations to improve future outcomes.

Travel Industry

- Applications in airlines, hotels, and rental car companies:
    1. Predicting service demand (e.g., seat types, room types) to optimize pricing (**yield management**).
    2. Forecasting demand at different locations to allocate resources efficiently.
    3. Identifying **profitable customers** for personalized services.
    4. Reducing employee attrition by understanding its causes.

Healthcare

- Applications in healthcare include:
    1. Identifying people without health insurance and understanding underlying factors.
    2. Discovering **cost-benefit relationships** between treatments.
    3. Forecasting demand at service locations for better resource allocation.
    4. Understanding reasons for **customer and employee attrition**.

Medicine

- Data mining complements clinical and biological research by:
    1. Identifying patterns to improve **cancer survivability**.
    2. Predicting **success rates of organ transplants** to develop better matching policies.
    3. Understanding gene functions (**genomics**).
    4. Exploring relationships between symptoms, illnesses, and treatments for better decision-making.

Entertainment Industry

- Applications include:
    1. Analyzing viewer data to optimize programming and **advertisement placements**.
    2. Predicting **financial success of movies** to guide investments.
    3. Forecasting demand for entertainment events for better scheduling and resource allocation.
    4. Developing **pricing policies** to maximize revenues.

Homeland Security and Law Enforcement

- Applications involve:
    1. Identifying patterns of **terrorist behaviors**.
    2. Discovering **crime patterns** to aid investigations.
    3. Predicting and preventing **biological/chemical threats** to infrastructure.
    4. Stopping **malicious cyberattacks** on information systems.

Sports

- Applications in professional and college sports include:
    1. NBA and MLB teams use data mining for **performance optimization** and resource utilization.
    2. Predicting outcomes of games (e.g., NCAA Bowl Games, March Madness).
    3. Helping teams maximize their **chances of success** with predictive models.

5.4 DATA MINING PROCESS

- Data mining projects follow standardized processes to ensure success.
- **CRISP-DM (Cross-Industry Standard Process for Data Mining)** is the most popular methodology.
- The CRISP-DM process has six iterative steps, beginning with understanding the business and ending with solution deployment.
- Backtracking is common, making the process time-consuming and iterative.

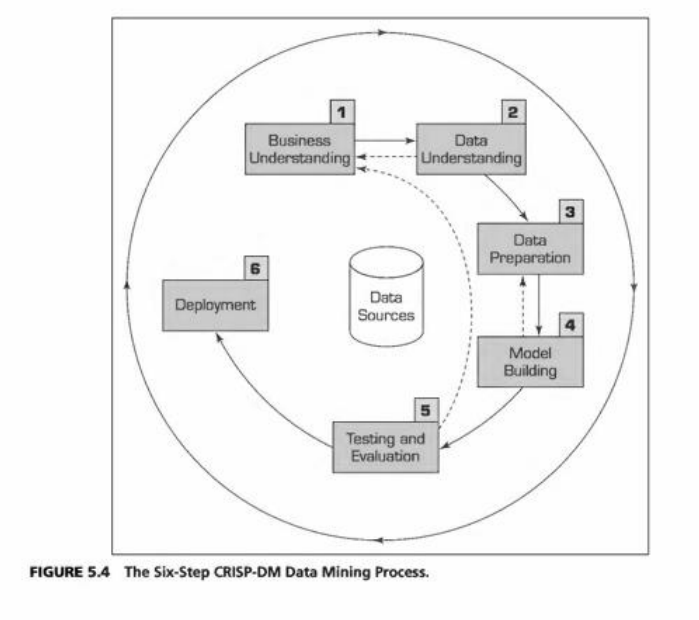


FIGURE 5.4 The Six-Step CRISP-DM Data Mining Process.

Overview

- Data mining follows a step-by-step process to ensure systematic execution.
- **CRISP-DM (Cross-Industry Standard Process for Data Mining)** is the most popular methodology.
- The process has six steps, starting with understanding the business and ending with deploying the solution.

- The steps are iterative and may involve backtracking to refine outcomes.

Step 1: Business Understanding

1. Clearly define the purpose of the study (e.g., identifying customer behavior or improving retention).
2. Specify business objectives with clear questions like:
    - "What are the traits of customers leaving for competitors?"
    - "What profiles of customers provide the most value?"
3. Create a project plan:
    - Assign roles for data collection, analysis, and reporting.
    - Include a high-level budget estimate to support the project.
4. A clear understanding of the business problem is essential to guide the study in the right direction.

Step 2: Data Understanding

1. Identify and select relevant data based on the task at hand.
    - Example: Retail projects might focus on demographics, credit card transactions, and purchase behaviors.
2. Understand:
    - **Data Sources**: Location, format, collection methods (manual or automated), and update frequency.
    - **Variables**: Ensure they are relevant, independent, and free from overlap or conflicts.
3. Use statistical and graphical techniques for data exploration:
    - Compute averages, medians, and standard deviations for numeric data.
    - Create modes and frequency tables for categorical data.
    - Visualize data using scatter plots, histograms, and box plots.
4. Types of data:
    - **Quantitative Data**: Numeric data that can be discrete (integers) or continuous (real numbers).
    - **Qualitative Data**:
        - **Nominal**: Non-ordered categories like gender.

        - **Ordinal**: Ordered categories like credit ratings (excellent, fair, bad).
5. Once relevant data is identified, proceed to data preparation.

Step 3: Data Preparation (Data Preprocessing)

- This is the **most time-consuming step** (about 80% of the project time).
- Real-world data often has issues like incompleteness, noise, and inconsistencies.

Phases of Data Preparation

1. **Data Collection and Integration**:
    - Collect data from identified sources.
    - Filter unnecessary sections and integrate multiple sources.
    - Handle synonyms and homonyms to avoid confusion.
2. **Data Cleaning (Data Scrubbing)**:
    - Manage missing values:
        - Impute (fill) missing values with probable estimates.
        - Ignore missing values when they are irrelevant (e.g., income left blank by high-income earners).
    - Identify and smooth noisy values (outliers).
    - Fix inconsistencies using expert knowledge.
3. **Data Transformation**:
    - Normalize variables to a common scale to avoid bias.
    - Discretize numeric variables into categories (e.g., low, medium, high).
    - Aggregate nominal values into broader groups (e.g., states grouped into regions).
    - Create new variables to simplify and enhance the data (e.g., a single variable for blood type match).
4. **Data Reduction**:
    - **Dimensional Reduction**: Reduce the number of variables using techniques like principal component analysis (PCA).
    - **Record Sampling**: Use subsets of data when datasets are too large.
        - Ensure subsets reflect the entire dataset (random sampling or stratified sampling).
        - Balance skewed data by oversampling underrepresented classes.

Step 4: Model Building

1. Choose appropriate modeling techniques to address the business goal.
2. Compare different models to find the best-performing one.
3. Types of data mining tasks:
    - **Prediction**: Classification or regression models.
    - **Association**: Discover relationships between variables.
    - **Clustering**: Group similar data points together.
4. Commonly used algorithms:
    - Decision Trees for classification.
    - K-means for clustering.
    - Apriori algorithm for association rule mining.
5. Fine-tune model parameters for optimal results.
6. Return to data preparation if reformatting is needed for the model.

Step 5: Testing and Evaluation

1. Assess models for accuracy and generalizability.
2. Ensure the model meets business objectives.
3. Test models in real-world scenarios if resources allow.
4. Evaluate business value from the discovered patterns:
    - Use visualization tools like pivot tables, pie charts, scatter plots, and histograms.
5. Collaboration between data analysts, business analysts, and decision-makers is essential for interpreting patterns.

Step 6: Deployment

1. Organize and present knowledge gained in a user-friendly format:
    - Deployment can be as simple as a report or as complex as enterprise-wide implementation.
2. Prepare the customer to carry out deployment actions.
3. Include a plan for monitoring and maintaining deployed models:
    - Monitor changes in business conditions and update models accordingly.

- Avoid long periods of incorrect usage of outdated models.

4. Develop a monitoring plan to ensure continued effectiveness of the data mining results.

5.7 DATA MINING PRIVACY ISSUES AND MYTHS

Data Mining Privacy Issues

Privacy Concerns in Data Mining

1. **Sensitive Data Involvement**:
   - Includes personal information like identification details, demographics, financial data, purchase history, and personal events.
   - Data often accessed through third-party providers, raising privacy risks.
2. **De-Identification**:
   - Anonymizing records to prevent tracing data back to individuals.
   - Widely used in public datasets like CDC and SEER.
3. **Ethical and Legal Obligations**:
   - Data mining professionals must ensure privacy and protect individuals' rights.
   - Users of de-identified data sources are often required to pledge not to re-identify individuals.

Examples of Privacy Breaches

1. **JetBlue Airlines Case (2003)**:
   - Shared over a million passenger records with a government contractor, Torch Concepts, for developing terrorist profiles.
   - Torch augmented data with family size and Social Security numbers from Acxiom, a data broker.
   - Actions were conducted without passenger consent, leading to lawsuits and government investigations.
2. **Social Media Privacy Issues**:
   - Allegations against social networks for selling customer-specific data to third parties for targeted marketing.
3. **Target Case (2012)**:
   - Predicted pregnancy statuses using non-private data.
   - Despite no laws being broken, the ethical concerns drew public attention.

Data Mining Myths

1. **Myth: Data Mining is a Fully Automated Process**
    - **Reality**: Data mining requires human expertise to set objectives, preprocess data, select algorithms, and interpret results.
2. **Myth: Data Mining Provides Instant and Precise Answers**
    - **Reality**: Results depend on data quality and may require refinement and iterative testing.
3. **Myth: Anyone Can Perform Data Mining**
    - **Reality**: Data mining requires domain expertise, statistical knowledge, and understanding of algorithms.
4. **Myth: Data Mining is Only for Large Companies**
    - **Reality**: Small and medium enterprises can also benefit from data mining with proper tools and scaled-down solutions.
5. **Myth: Data Mining Violates Privacy by Default**
    - **Reality**: Proper safeguards, ethical practices, and legal compliance minimize privacy risks.

MODULE 4 (remaining)

A Simple Taxonomy of Data in Data Mining.

- Data refers to a collection of facts usually obtained as the result of experiences, observations, or experiments.
- Data may consist of numbers, letters, words, images, voice recordings, and so on as measurements of a set of variables.
- Data are often viewed as the lowest level of abstraction from which information and then knowledge is derived. At the highest level of abstraction, one can classify data as structured and unstructured (or semistructured).
- Unstructured/semistructured data is composed of any combination of textual, image, voice, and Web content. Structured data is what data mining algorithms use, and can be classified as categorical or numeric.
- The categorical data can be subdivided into nominal or ordinal data, whereas numeric data can be subdivided into interval or ratio. Figure 5.2 shows a simple taxonomy of data.

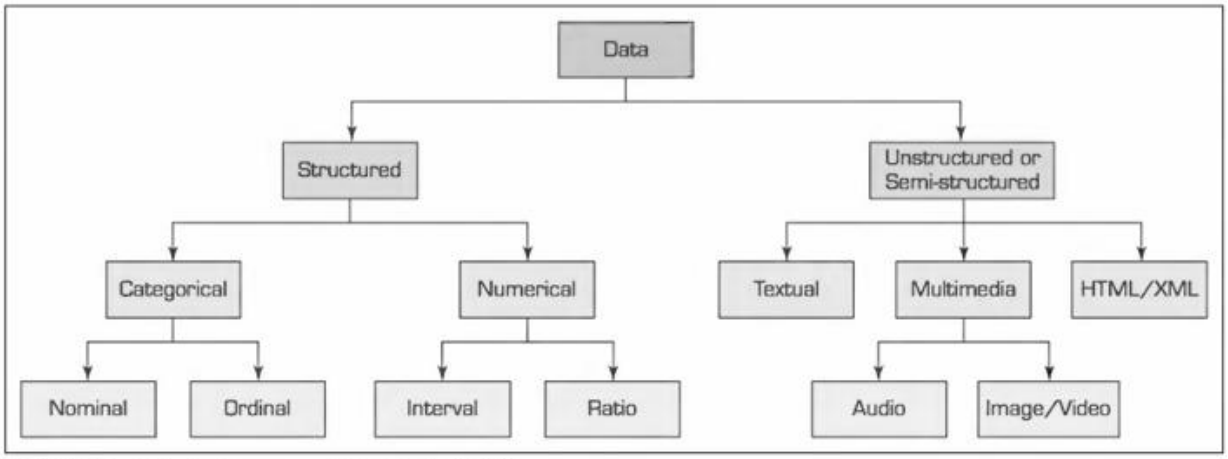


FIGURE 5.2 A Simple Taxonomy of Data in Data Mining.

• **Categorical data** represent the labels of multiple classes used to divide a variable into specific groups. Examples of categorical variables include race, sex, age group, and educational level. The categorical data may also be called discrete data, implying that it represents a finite number of values with no continuum between them. Even if the values used for the categorical (or discrete) variables are numeric, these numbers are nothing more than symbols and do not imply the possibility of calculating fractional values.

• **Nominal data** contain measurements of simple codes assigned to objects as labels, which are not measurements. For example, the variable marital status can be generally categorized as (1) single, (2) married, and (3) divorced. Nominal data can be represented with binomial values having two possible values (e.g., yes/no, true/ false, good/bad), or multinomial values having three or more possible values (e.g., brown/green/ blue, white/ black/Latino/ Asian, single/ married/divorced).

• **Ordinal data** contain codes assigned to objects or events as labels that also represent the rank order among them. For example, the variable credit score can be generally categorized as (1) low, (2) medium, or (3) high. Similar ordered relationships can be seen in variables such as age group (i.e., child, young, middle-aged, elderly) and educational level (i.e., high school, college, graduate school

• **Numeric data** represent the numeric values of specific variables. Examples of numerically valued variables include age, number of children, total household income (in U.S. dollars), travel distance (in miles), and temperature (in Fahrenheit degrees). Numeric values representing a variable can be integer (taking only whole numbers) or real (taking also the fractional number). The numeric data may also be called continuous data, implying that the variable contains continuous measures on a specific scale that allows insertion of interim values. Unlike a discrete variable, which represents finite, countable data, a continuous variable represents scalable measurements, and it is possible for the data to contain an infinite number of fractional values.

• **Interval data** are variables that can be measured on interval scales. A common example of interval scale measurement is temperature on the Celsius scale. In this particular scale, the unit of measurement is 1/ 100 of the difference between the melting temperature and the boiling temperature of water in atmospheric pressure; that is, there is not an absolute zero value.

• **Ratio data** include measurement variables commonly found in the physical sciences and engineering. Mass, length, time, plane angle, energy, and electric charge are examples of physical measures that are ratio scales. The scale type takes its name from the fact that measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. Informally, the distinguishing feature of a ratio scale is the possession of a nonarbitrary zero value.

Other data types, including textual, spatial, imagery, and voice, need to be converted into some form of categorical or numeric representation before they can be processed by data mining algorithms.

How Data Mining Works

- Using existing and relevant data, data mining builds models to identify patterns among the attributes presented in the data set.
- Models are mathematical representations (simple linear relationships and/or complex highly nonlinear relationships) that identify patterns among the attributes of the objects (e.g., customers) described in the data set.
- Some patterns are explanatory (explaining the interrelationships and affinities among the attributes), whereas others are predictive (foretelling future values of certain attributes).
- Data mining seeks to identify four major types of patterns:
    1. **Associations**:
        - Find commonly co-occurring groupings of things, such as beer and diapers going together in market-basket analysis.
    2. **Predictions**:
        - Tell the nature of future occurrences of certain events based on what has happened in the past, such as predicting the winner of the Super Bowl or forecasting the absolute temperature of a particular day.
    3. **Clusters**:
        - Identify natural groupings of things based on their known characteristics, such as assigning customers to different segments based on their demographics and past purchase behaviors.
    4. **Sequential Relationships**:
        - Discover time-ordered events, such as predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.
- These patterns have been manually extracted from data by humans for centuries.
- The increasing volume of data in modern times has created a need for automatic approaches.

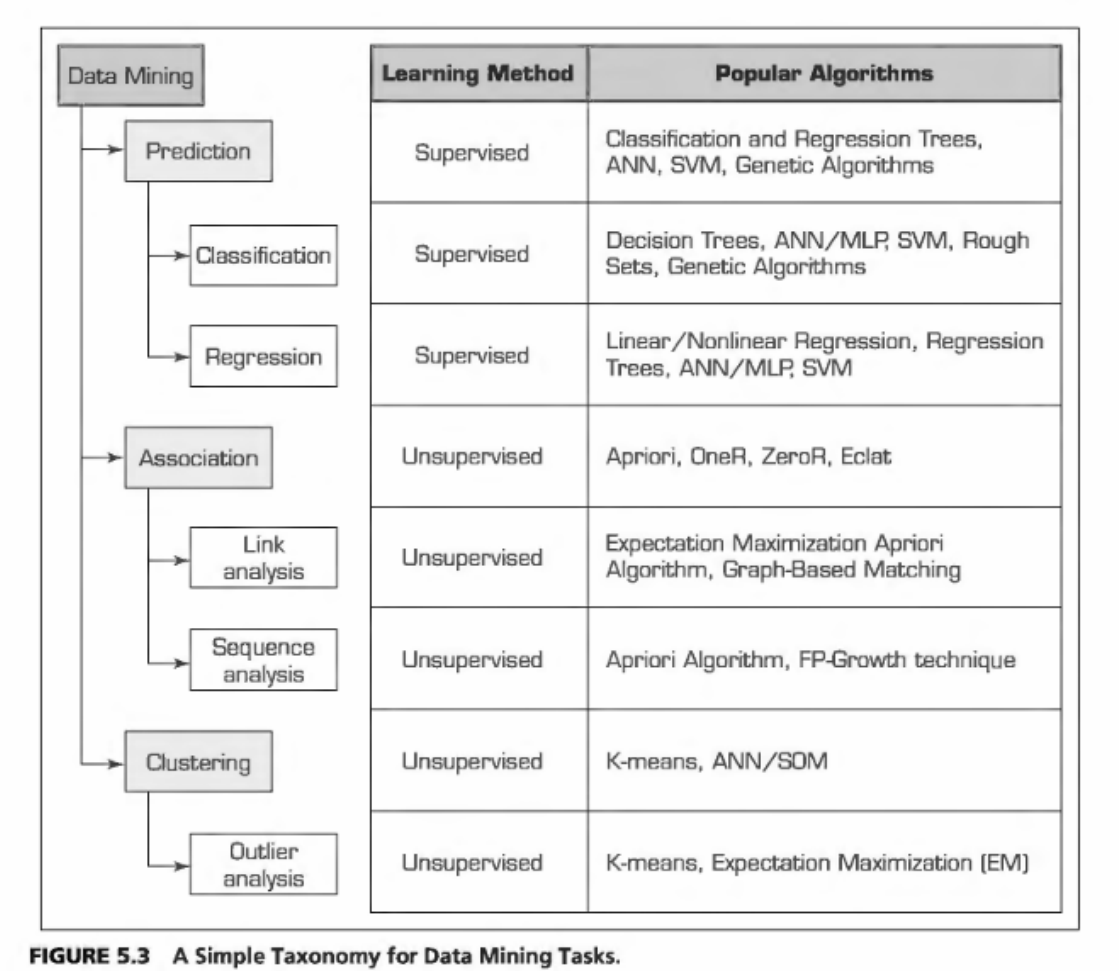
- As data sets have grown in size and complexity, direct manual data analysis has increasingly been augmented with indirect, automatic data processing tools that use sophisticated methodologies, methods, and algorithms.
- This evolution of automated and semi-automated processing of large data sets is now referred to as data mining.

Data Mining Tasks

- Data mining tasks can be classified into three main categories: prediction, association, and clustering.
- Based on the way patterns are extracted from historical data, the learning algorithms of data mining methods can be classified as either supervised or unsupervised:
    - **Supervised Learning Algorithms**:
        - The training data includes both descriptive attributes (independent or decision variables) and the class attribute (output or result variable).
    - **Unsupervised Learning Algorithms**:
        - The training data includes only descriptive attributes.

Prediction

- Prediction is commonly referred to as the act of telling about the future.
- It differs from guessing by considering experiences, opinions, and other relevant information.
- **Forecasting** is a related term, with a subtle difference:
    - Prediction is experience and opinion-based.
    - Forecasting is data and model-based.
- In increasing reliability, the terms can be ordered as guessing, predicting, and forecasting.
- In data mining, prediction and forecasting are used synonymously.
- Depending on the nature of what is being predicted:
    - **Classification**: Predicts a class label, such as "rainy" or "sunny."
    - **Regression**: Predicts a numerical value, such as tomorrow’s temperature (e.g., "65°F").



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision Trees, ANN/MLP, SVM, Rough Sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression Trees, ANN/MLP, SVM
Unsupervised	Apriori, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization Apriori Algorithm, Graph-Based Matching
Unsupervised	Apriori Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

FIGURE 5.3 A Simple Taxonomy for Data Mining Tasks.

Classification

- Classification, or supervised induction, is the most common data mining task.
- The objective is to analyze historical data stored in a database and generate a model to predict future behavior.
- The model generalizes over the records of a training data set to help distinguish predefined classes.
- Common classification tools include:
    - Neural networks and decision trees (from machine learning).
    - Logistic regression and discriminant analysis (from traditional statistics).
    - Emerging tools like rough sets, support vector machines, and genetic algorithms.
- Limitations of some methods:
    - Statistics-based techniques (e.g., logistic regression) assume independence and normality, which may not be realistic.

- Neural networks:
  - Effective for large, complex datasets but require considerable training.
  - Training time increases exponentially with data volume.
  - Difficult to interpret predictions and limited applicability to very large datasets.
- Decision Trees:
  - Classify data into a finite number of classes using hierarchical if-then statements.
  - Faster than neural networks but require discretization of continuous variables.
- Rule Induction:
  - Derives if-then statements directly from the training data without a hierarchical structure.

Clustering

- Clustering partitions a collection of objects into segments whose members share similar characteristics.
- Unlike classification, clustering involves unknown class labels.
- Clusters are established by identifying commonalities in data using heuristic algorithms.
- Results may require expert interpretation and modification before use.
- Common techniques:
  - K-Means (from statistics).
  - Self-Organizing Maps (a neural network architecture by Kohonen, 1982).

Associations

- Associations, or association rule learning, discover interesting relationships among variables in large databases.
- Applications include market-basket analysis, which identifies frequent product groupings (e.g., beer and diapers).
- Subcategories:
  - **Link Analysis**: Identifies linkages among objects, such as web pages or academic publications.

    - **Sequence Mining**: Examines relationships in terms of their order of occurrence.
- Algorithms: Apriori, FP-Growth, OneR, ZeroR, and Eclat.

Visualization and Time-Series Forecasting

- **Visualization**:
    - Used with data mining techniques to understand relationships.
    - Combines analytics and visualization for faster knowledge creation (referred to as visual analytics).
- **Time-Series Forecasting**:
    - Data consists of values captured over time in regular intervals.
    - These values are used to develop forecasting models for future predictions.

Q. What are the main data preprocessing steps? Briefly describe each step with relevant examples.

- The purpose of data preparation (or more commonly called data preprocessing) is to take the data identified in the previous step and prepare it for analysis by data mining methods.
- Figure 5.5 shows the four main steps needed to convert the raw real-world data into minable data sets.
- In the first phase of data preprocessing, the relevant **data is collected** from the identified sources, the necessary records and variables are selected (based on an intimate understanding of the data, the unnecessary sections are filtered out), and the records coming from multiple data sources are integrated (again, using the intimate understanding of the data, the synonyms and homonyms are to be handled properly).
- In the second phase of data preprocessing, the **data is cleaned** (this step is also known as data scrubbing). In this step, the values in the data set are identified and dealt with. In some cases, missing values are an anomaly in the data set, in which case they need to be imputed (filled with a most probable value) or ignored; in other cases, the missing values are a natural part of the data set (e.g., the household income field is often left unanswered by people who are in the top income tier). In this step, the analyst should also identify noisy values in the data (i.e., the outliers)

and smooth them out. Additionally, inconsistencies (unusual values within a variable) in the data should be handled using domain knowledge and/or expert opinion.

- In the third phase of data preprocessing, the **data is transformed** for better processing. For instance, in many cases the data is normalized between a certain minimum and maximum for all variables in order to mitigate the potential bias of one variable (having large numeric values, such as for household income) dominating other variables having smaller values. Another transformation that takes place is discretization and/or aggregation. In some cases, the numeric variables are converted to categorical values (e.g., low, medium, high); in other cases a nominal variable's unique value range is reduced to a smaller set using concept hierarchies (e.g., as opposed to using the individual states with 50 different values, one may choose to use several regions for a variable that shows location) in order to have a data set that is more amenable to computer processing.

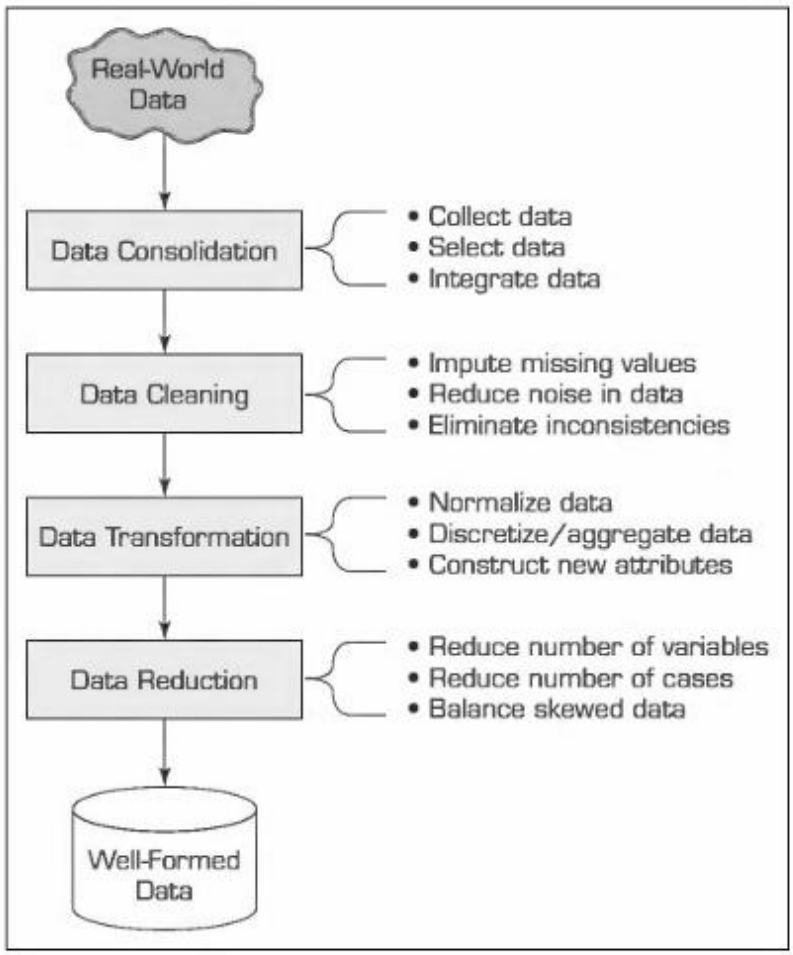


FIGURE 5.5 Data Preprocessing Steps.

- The final phase of data preprocessing is **data reduction.** Even though data miners like to have large data sets, too much data is also a problem. In the simplest sense, one can visualize the data commonly used in data mining projects as a flat file consisting of two dimensions: variables (the number of columns) and cases/records (the number of rows). In some cases (e.g., image processing and genome projects with complex microarray data), the number of variables can be rather large, and the analyst must reduce the number to a manageable size. Because the variables are treated as different dimensions that describe the phenomenon from different perspectives, in data mining this process is commonly called dimensional reduction.

****

TABLE 5.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill-in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as "ML"; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or −1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range or frequency-based binning techniques; for categorical variables reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Principal component analysis, independent component analysis, Chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert–knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

Other Data Mining Standardized Processes and Methodologies

- In order to be applied successfully, a data mining study must be viewed as a process that follows a standardized methodology rather than as a set of automated software tools and techniques.
- In addition to CRISP-DM, there is another well-known methodology developed by the SAS Institute, called SEMMA (2009). The acronym SEMMA stands for "sample, explore, modify, model, and assess." Beginning with a statistically representative sample of the data, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and trans form the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy.
- A pictorial representation of SEMMA is given in Figure 5.6. By assessing the outcome of each stage in the SEMMA process, the model developer can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data; that is, as with CRISP-DM, SEMMA is driven by a highly iterative experimentation cycle.
-

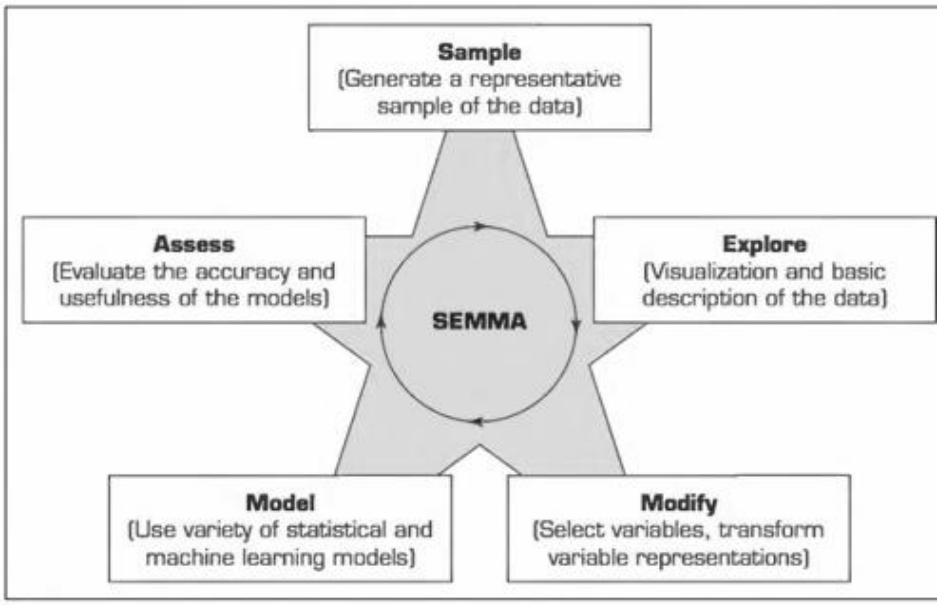


FIGURE 5.6 SEMMA Data Mining Process.

IMPORTANT The **main difference between CRISP-DM and SEMMA** is that CRISP-DM takes a more comprehensive approach-including understanding of the business and the relevant data-to data mining projects, whereas SEMMA implicitly assumes that the data mining

project's goals and objectives along with the appropriate data sources have been identi fied and understood. *

Some practitioners commonly use the term knowledge discovery in databases (KDD) as a synonym for data mining. Fayyad et al. 0996) defined knowledge discovery in databases as a process of using data mining methods to find useful information and patterns in the data, as opposed to data mining, which involves using algorithms to identify patterns in data derived through the KDD process. KDD is a comprehensive process that encompasses data mining. The input to the KDD process consists of organi zational data. The enterprise data warehouse enables KDD to be implemented efficiently because it provides a single source for data to be mined. Dunham (2003) summarized the KDD process as consisting of the following steps: data selection, data preprocess ing, data transformation, data mining, and interpretation/evaluation.

Data Mining Myths and Blunders

Myth	Reality
Data mining provides instant, crystal-ball-like predictions.	Data mining is a multistep process that requires deliberate, proactive design and use.
Data mining is not yet viable for business applications.	The current state-of-the-art is ready to go for almost any business.
Data mining requires a separate, dedicated database.	Because of advances in database technology, a dedicated database is not required, even though it may be desirable.
Only those with advanced degrees can do data mining.	Newer Web-based tools enable managers of all educational levels to do data mining.
Data mining is only for large firms that have lots of customer data.	If the data accurately reflect the business or its customers, a company can use data mining.

The following 10 data mining mistakes are often made in practice

1. Selecting the wrong problem for data mining.

2. Ignoring what your sponsor thinks data mining is and what it really can and cannot do.

3. Leaving insufficient time for data preparation. It takes more effort than is generally understood.

4. Looking only at aggregated results and not at individual records. IBM's DB2 IMS can highlight individual records of interest.

5. Being sloppy about keeping track of the data mining procedure and results.

6. Ignoring suspicious findings and quickly moving on.

7. Running mining algorithms repeatedly and blindly. It is important to think hard about the next stage of data analysis. Data mining is a very hands-on activity.

8. Believing everything you are told about the data.

9. Believing everything you are told about your own data mining analysis.

10. Measuring your results differently from the way your sponsor measures them.