Data Mining & Knowledge Discovery

Introduction to Data Mining

Md. Biplob Hosen

Lecturer, IIT-JU

Email: biplob.hosen@juniv.edu

Contents

- Data Mining Overview
- Data Mining Applications
- Origins of Data Mining
- Data Mining Tasks
- Data Mining Challenges

Reference Books

- Introduction to Data Mining, 2nd Edition by Tan, Steinbach, Karpatne, Kumar.

Data Mining Overview

Large-scale Data is Everywhere!

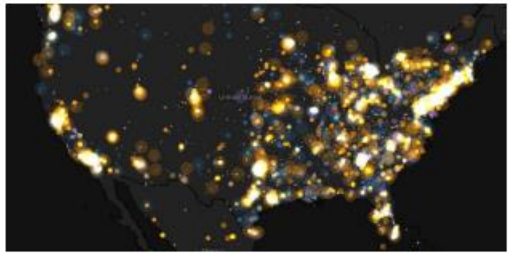
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- Gather whatever data you can whenever and wherever possible.
- Expectations: Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Cyber Security

[NEWSFACTOR NETWORK]

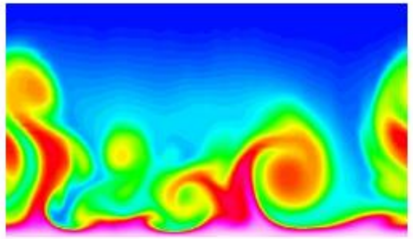
E-Commerce

Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

Data Mining Overview

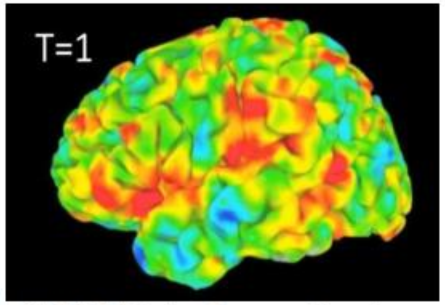
Why Mine Data (Commercial Viewpoint)?

- Lots of data is being collected and warehoused.
- Web data
  - Google has Peta Bytes of web data
  - Facebook has billions of active users
- Purchases at department/ grocery stores, e-commerce
  - Amazon handles millions of visits/day
- Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Data Mining Overview

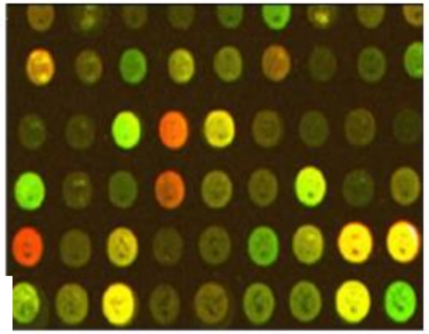
Why Mine Data (Scientific Viewpoint)?

- Data collected and stored at enormous speeds.
- Remote sensors on a satellite: NASA EOSDIS archives over petabytes of earth science data per year.
- Telescopes scanning the skies: Sky survey data.
- High-throughput biological data
- Scientific simulations: Terabytes of data generated in a few hours.
- Data mining helps scientists:
  - In automated analysis of massive datasets
  - In hypothesis formation

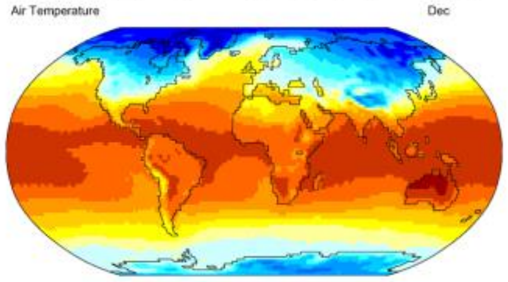


fMRI Data from Brain

Sky Survey Data



Gene Expression Data



Surface Temperature of Earth

Data Mining Overview

What is Data Mining?

- The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data.
- Exploration & analysis, by automatic or semi-automatic means of large quantities of data in order to discover meaningful patterns.

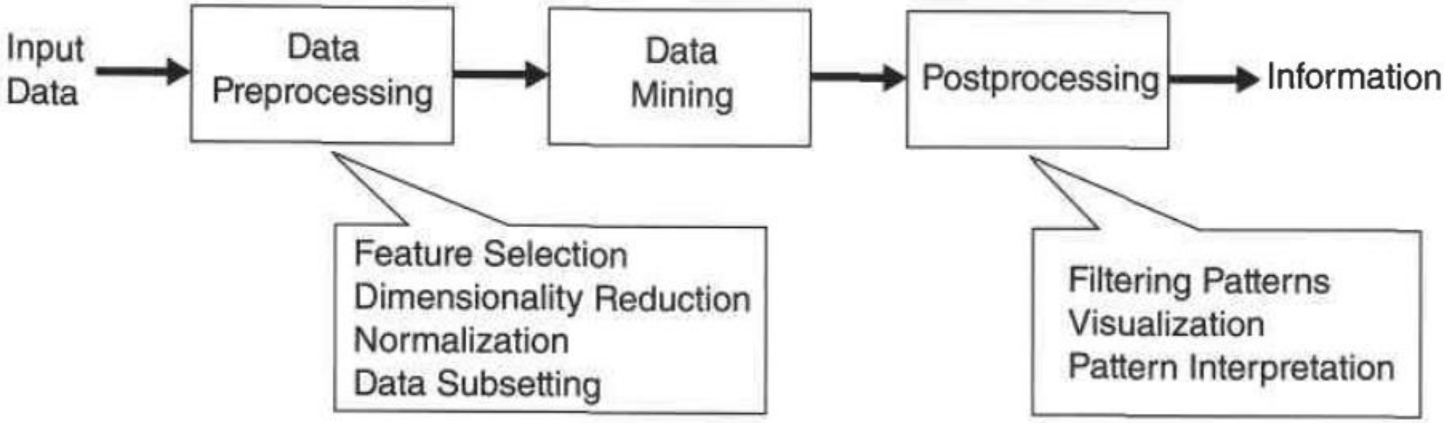
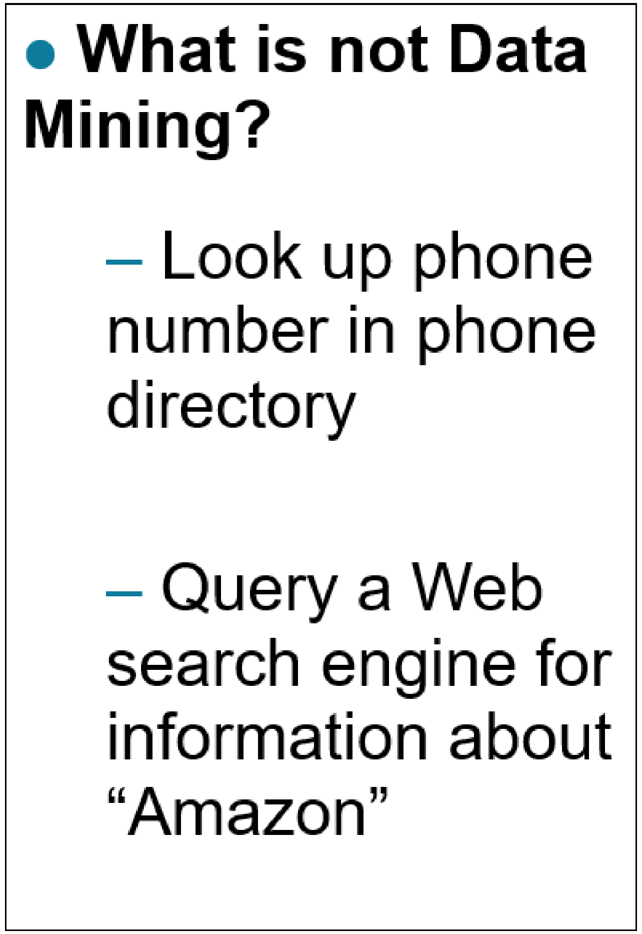


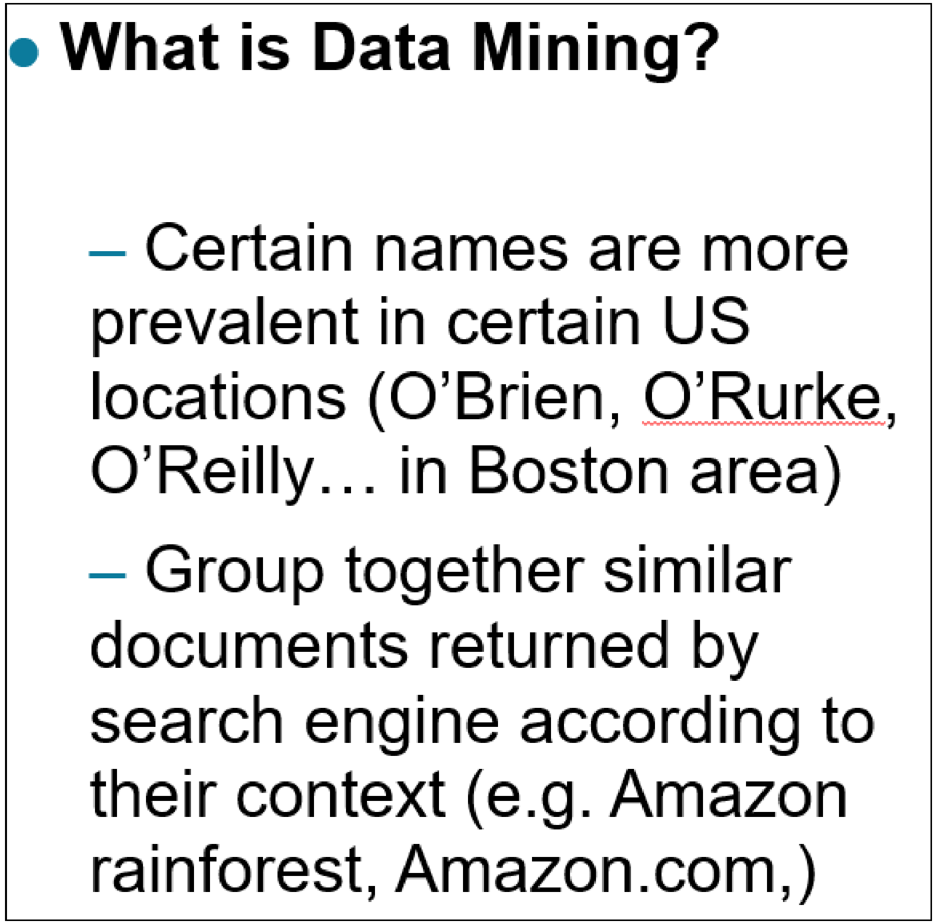
Figure 1.1. The process of knowledge discovery in databases (KDD).

Data Mining Overview

What is (not) Data Mining?

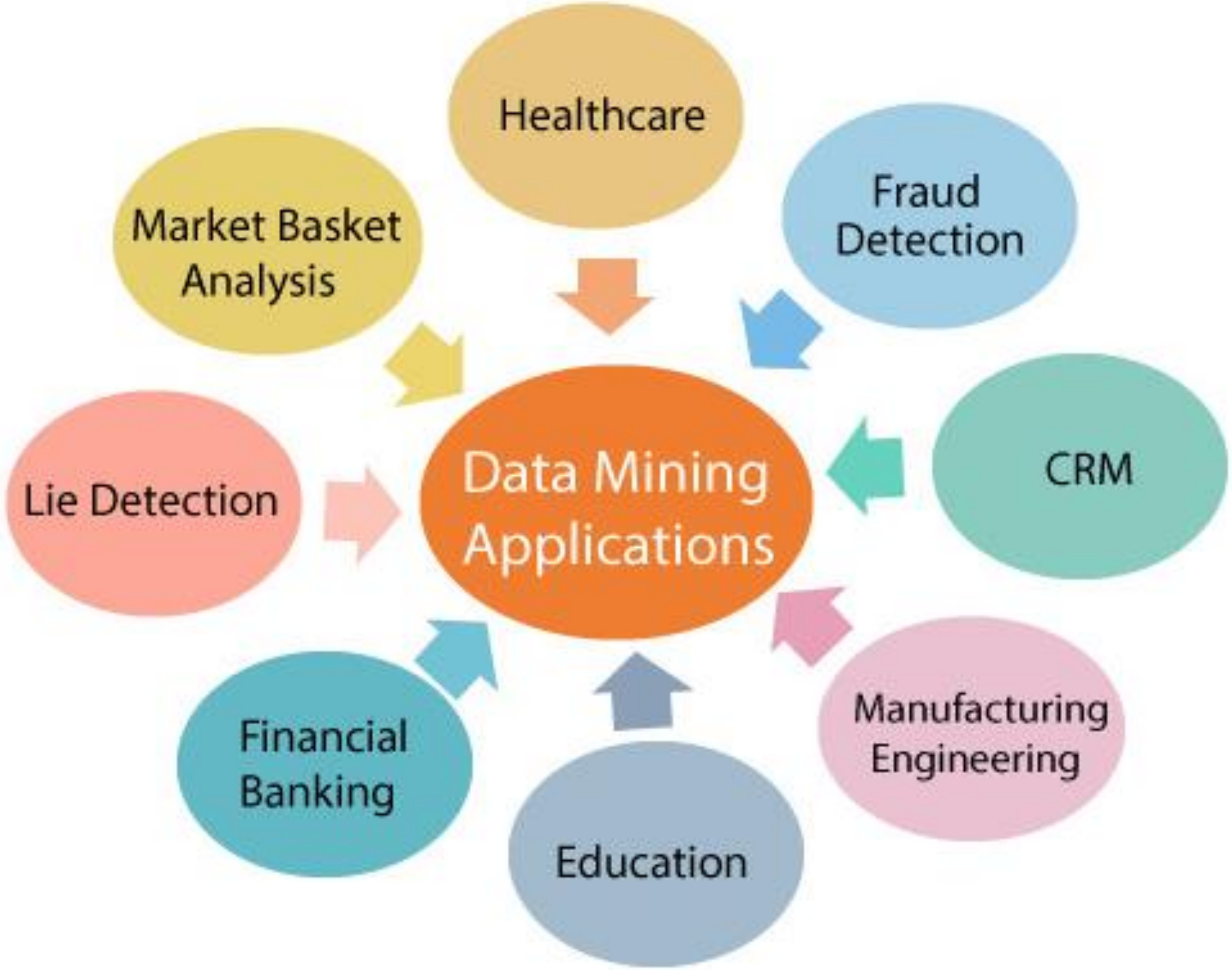


- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"



- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Data Mining Applications



Data Mining Applications

Data Mining in Healthcare:

- Data mining in healthcare has excellent potential to improve the health system.
- It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs.
- Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics.
- Data Mining can be used to forecast patients in each category.
- The procedures ensure that the patients get intensive care at the right place and at the right time.
- Data mining also enables healthcare insurers to recognize fraud and abuse.

Data Mining in Market Basket Analysis:

- Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products.
- This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly.
- Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

Data Mining Applications

Data mining in Education:

- Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments.
- EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science.
- An organization can use data mining to make precise decisions and also to predict the results of the student.
- With the results, the institution can concentrate on what to teach and how to teach.

Data Mining in Manufacturing Engineering:

- Data mining tools can be beneficial to find patterns in a complex manufacturing process.
- Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers.
- It can also be used to forecast the product development period, cost, and expectations among the other tasks.

Data Mining Applications

Data Mining in CRM (Customer Relationship Management):

- Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies.
- To get a decent relationship with the customer, a business organization needs to collect data and analyze the data.
- With data mining technologies, the collected data can be used for analytics.

Data Mining in Fraud detection:

- Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated.
- Data mining provides meaningful patterns and turning data into information.
- An ideal fraud detection system should protect the data of all the users.
- Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent.
- A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

Data Mining Applications

Data Mining in Lie Detection:

- Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task.
- Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc.
- This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text.
- The information collected from the previous investigations is compared, and a model for lie detection is constructed.

Data Mining Financial Banking:

- The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction.
- The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts.
- The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

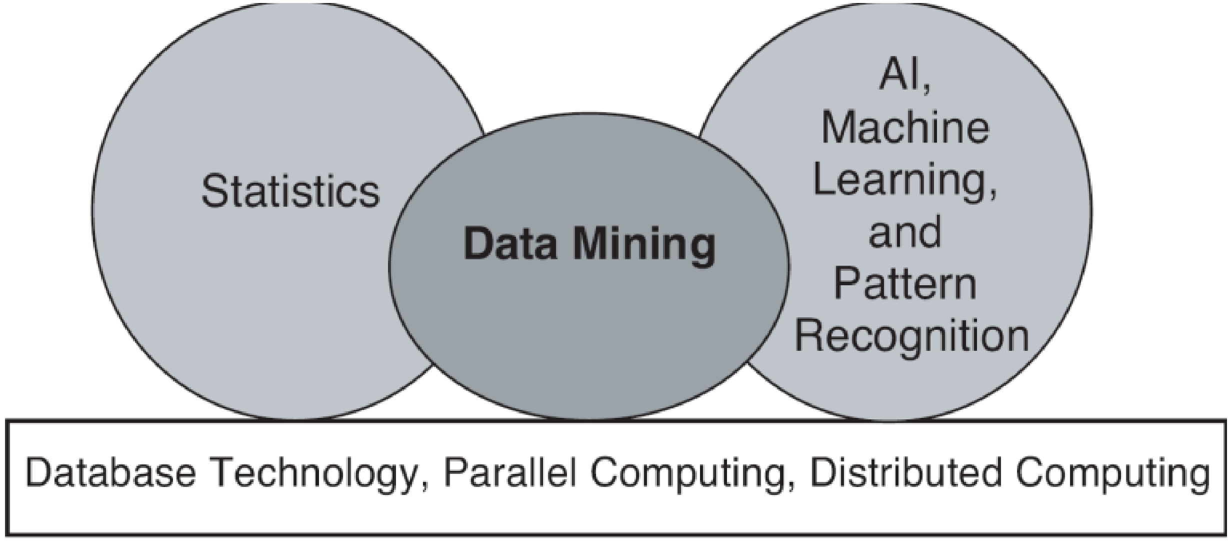
Origins of Data Mining

- Data mining draws upon ideas, such as:

(1) Sampling, estimation, and hypothesis testing from statistics.

(2) Search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning.

(3) Data mining has also been quick to adopt ideas from other areas, including optimization, evolutionary computing, information theory, signal processing, visualization, and information retrieval.

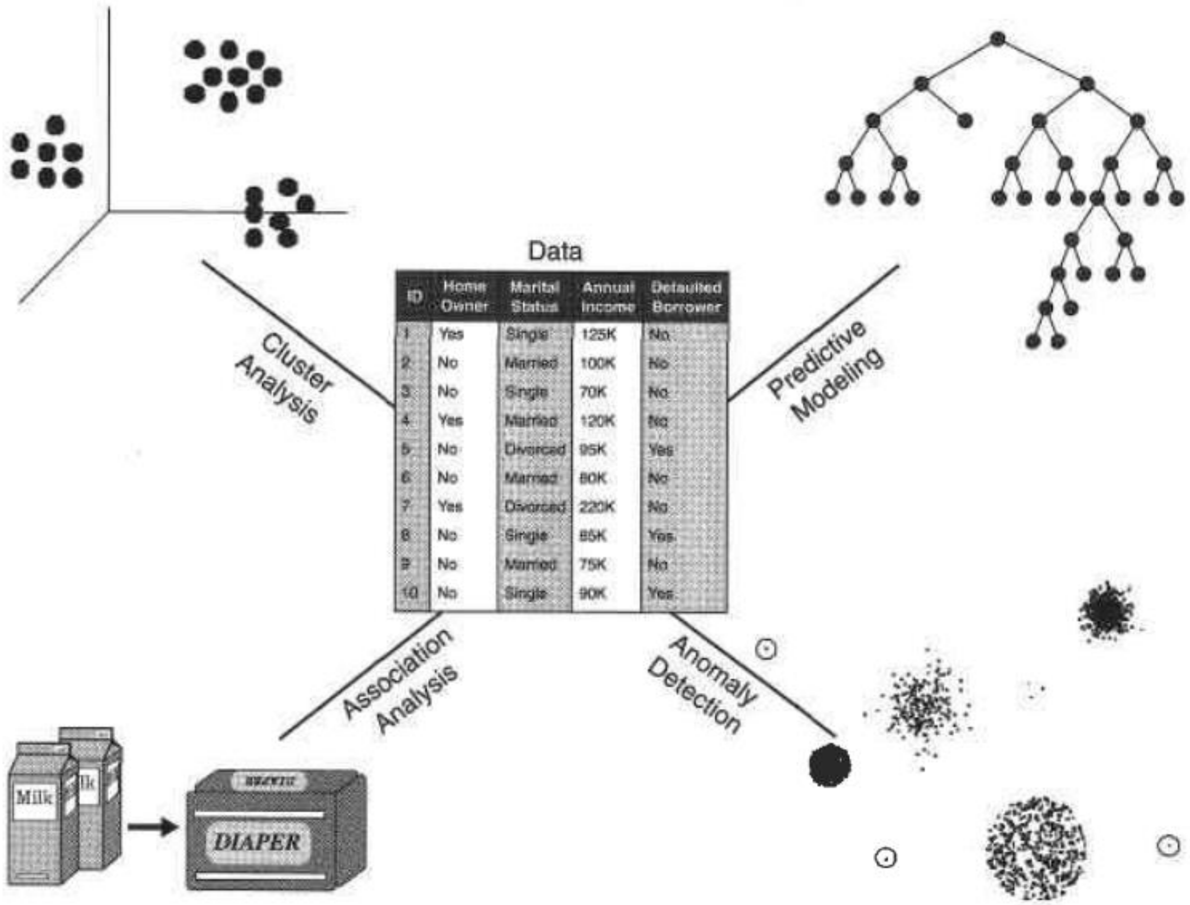

Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

Data Mining Tasks

- Prediction Methods: Use some variables to predict unknown or future values of other variables.
- Description Methods: Find human-interpretable patterns that describe the data.
- Examples:

1. Classification [Predictive]
2. Clustering [Descriptive]
3. Association Rule Discovery [Descriptive]
4. Sequential Pattern Discovery [Descriptive]
5. Regression [Predictive]
6. Deviation Detection [Predictive]

Data Mining Tasks



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 1.3. Four of the core data mining tasks.

Predictive Tasks

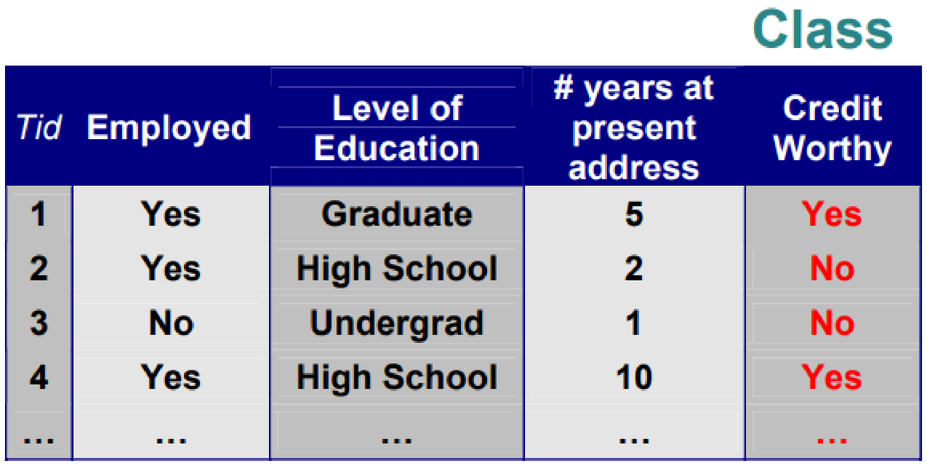
- Predictive modeling refers to the task of building a model for the target variable as a function of the explanatory variables.
- There are two types of predictive modeling tasks: classification, which is used for discrete target variables, and regression, which is used for continuous target variables.
- For example, predicting whether a web user will make a purchase at an online bookstore is a classification task because the target variable is binary-valued.
- On the other hand, forecasting the future price of a stock is a regression task because price is a continuous-valued attribute.
- The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable.
- Predictive modeling can be used to identify customers that will respond to a marketing campaign, predict disturbances in the Earth's ecosystem, or judge whether a patient has a particular disease based on the results of medical tests.

Predictive Tasks (Classification)

- Given a collection of records (training set).
- Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
- A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Predictive Tasks (Classification)

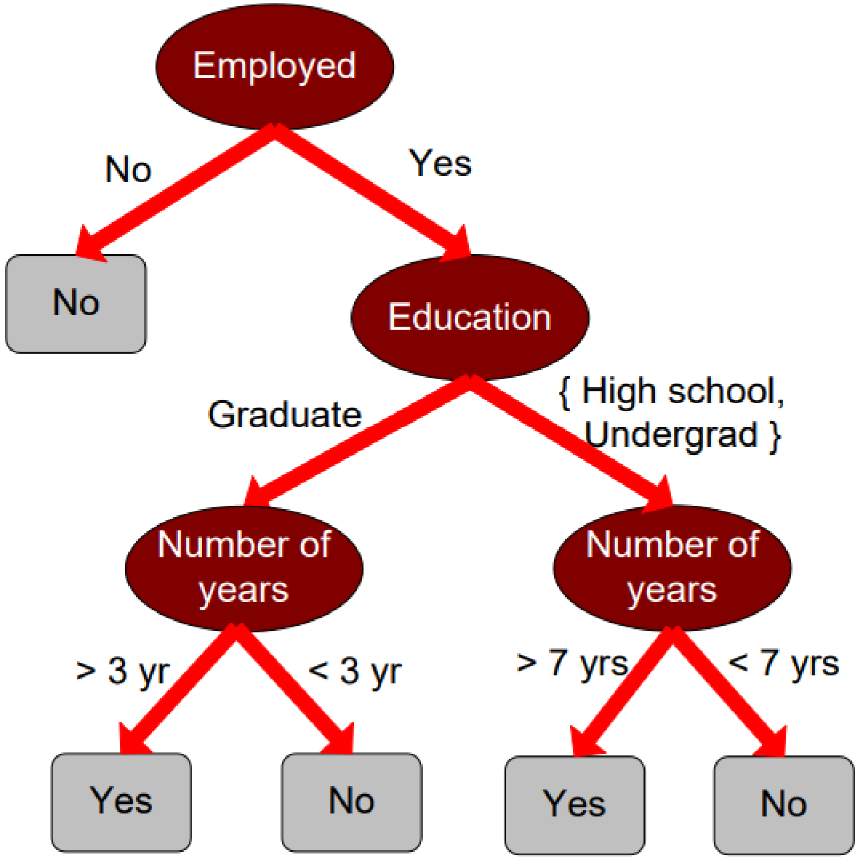
- Find a model for class attribute as a function of the values of other attributes



Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

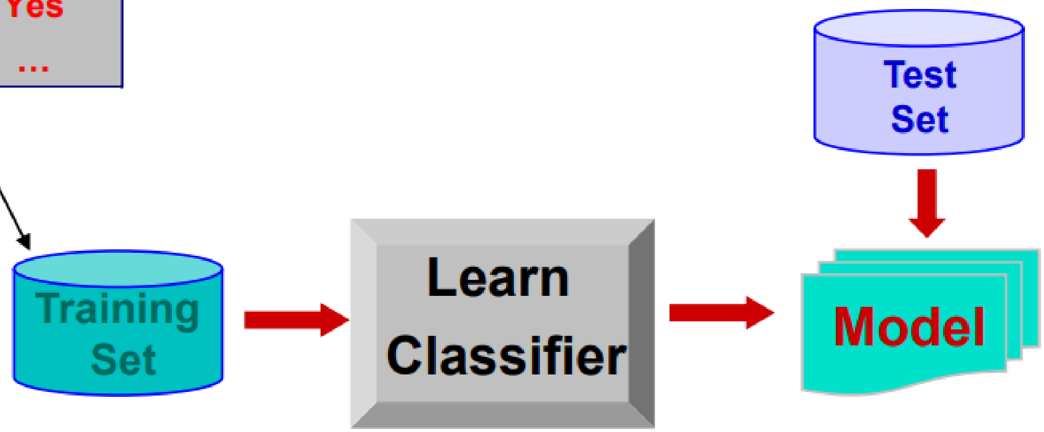
Model for predicting credit worthiness



Predictive Tasks (Classification)

	categorical	categorical	quantitative	class
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Classification: Application 1

Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions.
- Approach:

1. Use credit card transactions and the information on its account-holder as attributes.
    a. When does a customer buy, what does he buy, how often he pays on time, etc.
2. Label past transactions as fraud or fair transactions. This forms the class attribute.
3. Learn a model for the class of the transactions.
4. Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 2

Direct Marketing

- Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.
- Approach:

1. Use the data for a similar product introduced before.
2. We know which customers decided to buy and which decided otherwise. This **{buy, don’t buy}** decision forms the class attribute.
3. Collect various demographic, lifestyle, and company-interaction related information about all such customers.
4. Type of business, where they stay, how much they earn, etc.
5. Use this information as input attributes to learn a classifier model.

Classification: Application 3

Customer Attrition/Churn:

- Goal: To predict whether a customer is likely to be lost to a competitor.
- Approach:

1. Use detailed record of transactions with each of the past and present customers, to find attributes.
2. How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
3. Label the customers as loyal or disloyal.
4. Find a model for loyalty.

Classification: Application 4

Sky Survey Cataloging:

- Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images.
- Approach:

1. Segment the image.
2. Measure image attributes (features).
3. Model the class based on these features.
4. Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classification: Application - Practice

- Write short note on following classification applications in terms of goals and approaches:

1. Loan Default Prediction
2. Brain MRI Analysis for Dementia Detection

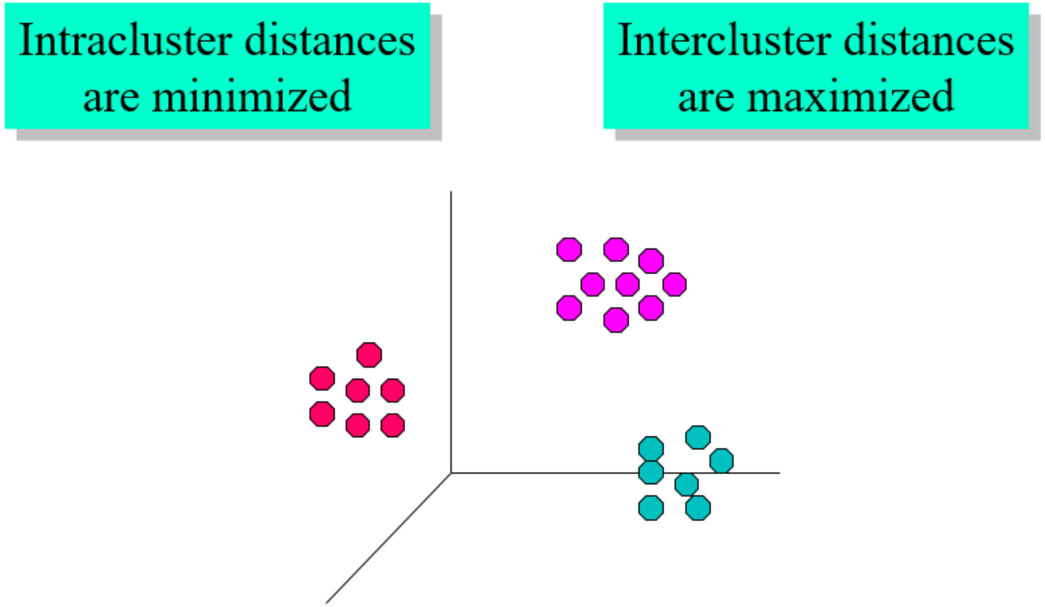
Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:

1. Predicting sales amounts of new product based on advertising expenditure.
2. Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
3. Time series prediction of stock market indices.

Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.



Intracluster distances are minimized

Intercluster distances are maximized

Clustering: Application 1

Market Segmentation:

- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:

1. Collect different attributes of customers based on their geographical and lifestyle related information.
2. Find clusters of similar customers.
3. Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

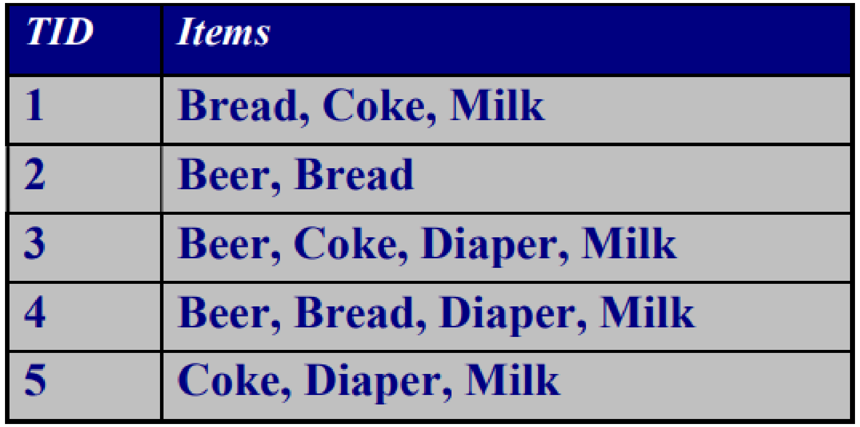
Clustering: Application 2

Document Clustering:

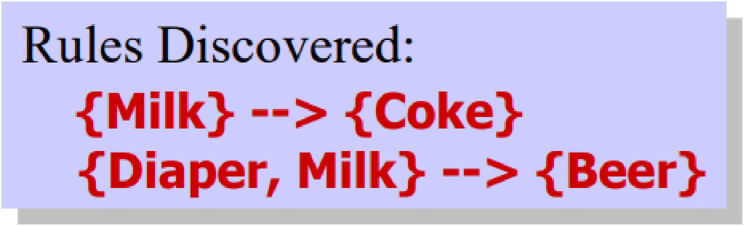
- Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
- Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
- Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Association Rule

- Given a set of records each of which contain some number of items from a given collection:
  – Produce dependency rules which will predict occurrence of an item based on occurrences of other items.



TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association Analysis: Applications

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Association Analysis: Application (1)

Supermarket shelf management:

- Goal: To identify items that are bought together by sufficiently many customers.
- Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
- A classic rule --
  - If a customer buys diaper and milk, then he is very likely to buy beer.
  - So, don’t be surprised if you find six-packs stacked next to diapers!

Association Analysis: Application (2)

Inventory Management:

- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
- Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Anomaly Detection

- Anomaly detection is the task of identifying observations whose characteristics are significantly different from the rest of the data.
- Such observations are known as anomalies or outliers. The goal of an anomaly detection algorithm is to discover the real anomalies and avoid falsely labeling normal objects as anomalous.
- In other words, a good anomaly detector must have a high detection rate and a low false alarm rate.
- Applications of anomaly detection include the detection of fraud, network intrusions, unusual patterns of disease, and ecosystem disturbances.

Example 1.4 (Credit Card Fraud Detection)-

- A credit card company records the transactions made by every credit card holder, along with personal information such as credit limit, age, annual income, and address. since the number of fraudulent cases is relatively small compared to the number of legitimate transactions, anomaly detection techniques can be applied to build a profile of legitimate transactions for the users.
- When a new transaction arrives, it is compared against the profile of the user. If the characteristics of the transaction are very different from the previously created profile, then the transaction is flagged as potentially fraudulent.

Data Mining Challenges

- Data Quality
- Scalability
- High Dimensionality
- Complex and Heterogeneous Data
- Privacy Preservation
- Data Ownership and Distribution

Data Mining Challenges

Data Quality:

- The accuracy, completeness, and consistency of the data affect the accuracy of the results obtained.
- The data may contain errors, omissions, duplications, or inconsistencies, which may lead to inaccurate results.
- Moreover, the data may be incomplete, meaning that some attributes or values are missing, making it challenging to obtain a complete understanding.
- Data quality issues arise due to a variety of reasons, including data entry errors, data storage issues, data integration problems, and data transmission errors.
- To address these challenges, data mining practitioners must apply data cleaning and data preprocessing techniques to improve the quality of the data.
- Data cleaning involves detecting and correcting errors, while data preprocessing involves transforming the data to make it suitable for data mining.

Data Mining Challenges

<u>Scalability:</u>

- Data mining algorithms must be scalable to handle large datasets efficiently.
- As the size of the dataset increases, the time and computational resources required to perform data mining operations also increase.
- Moreover, the algorithms must be able to handle streaming data, which is generated continuously and must be processed in real-time.
- To address this challenge, data mining practitioners use distributed computing frameworks such as Hadoop and Spark.
- These frameworks distribute the data and processing across multiple nodes, making it possible to process large datasets quickly and efficiently.

Data Mining Challenges

High Dimensionality:

- High-dimensional data are defined as data in which the number of features (variables observed) p , are close to or larger than the number of observations (or data points), n.
- Analyses of high-dimensional data require consideration of potential problems that come from having more features than observations.
- High-dimensional data have become more common in many scientific fields as new automated data collection techniques have been developed.
- Datasets in which p>=n are becoming more common. Such datasets pose a challenge for data analysis as standard methods of analysis, such as linear regression, are no longer appropriate.
- High-dimensional datasets are common in the biological sciences. Subjects like genomics and medical sciences often use both tall (in terms of n) and wide (in terms of p) datasets that can be difficult to analyze or visualize using standard statistical tools.
- An example of high-dimensional data in biological sciences may include data collected from hospital patients recording symptoms, blood test results, behaviors, and general health, resulting in datasets with large numbers of features.

Data Mining Challenges

Complex and Heterogeneous Data:

- Data complexity refers to the vast amounts of data generated by various sources, such as sensors, social media, and the internet of things (IoT).
- The complexity of the data may make it challenging to process, analyze, and understand.
- In addition, the data may be in different formats, making it challenging to integrate into a single dataset.
- To address this challenge, data mining practitioners use advanced techniques such as clustering, classification, and association rule mining.
- These techniques help to identify patterns and relationships in the data, which can then be used to gain insights and make predictions.

Data Mining Challenges

Privacy Preservation:

- Data privacy and security is another significant challenge in data mining.
- As more data is collected, stored, and analyzed, the risk of data breaches and cyber-attacks increases.
- The data may contain personal, sensitive, or confidential information that must be protected.
- Moreover, data privacy regulations such as GDPR, CCPA, and HIPAA impose strict rules on how data can be collected, used, and shared.
- To address this challenge, data mining practitioners must apply data anonymization and data encryption techniques to protect the privacy and security of the data.
- Data anonymization involves removing personally identifiable information (PII) from the data, while data encryption involves using algorithms to encode the data to make it unreadable to unauthorized users.

Practice Tasks

- From the book-
    - **Exercises:** 1 & 2

Thank You!