Introduction to Data Mining

DS515

Shahira Azazy

Contact Info.

- Shahiraazazy@gmail.com

- https://fssr.gnomio.com/

Required Software

- Google Colab
  - https://colab.research.google.com/

- Download Anaconda
  - https://www.anaconda.com/

Introduction to Data Mining

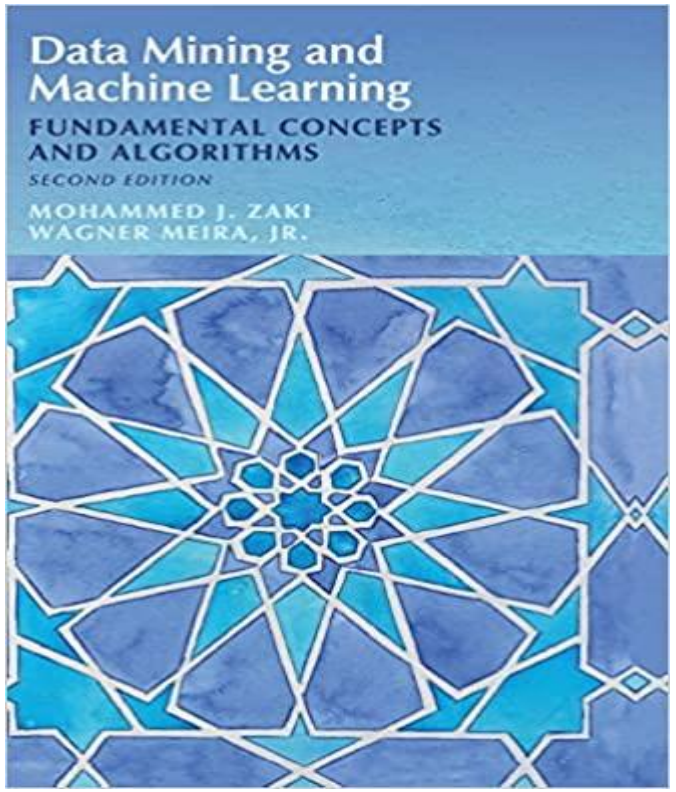
- **Data Warehousing and OLAP**
- **Exploratory data analysis????**
- **Data Preprocessing**
- Data Mining Knowledge Representation
- Attribute-Oriented Analysis
- Data Mining Algorithms
- Association Rules
- Classification
- Prediction
- Evaluating What's Been Learned
- Mining Real Data
- Clustering Advanced Techniques
- Data MiningSoftware
- Applications.

References (1)

Data Mining and Machine Learning: Fundamental Concepts and Algorithms (2nd Edition), Mohammed J. Zaki and Wagner Meira, Jr, Cambridge University Press, 2020.

Main Page | Data Mining and Machine Learning (dataminingbook.info)

CSCI4390-6390 Data Mining | Zaki Home Page (rpi.edu)

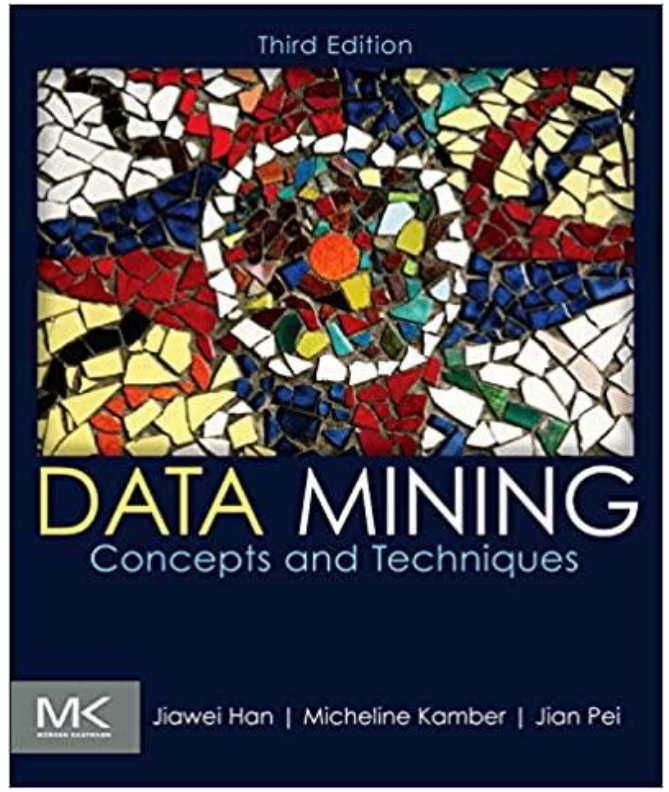


References(2)

Data Mining: Concepts and Techniques (3rd Edition), Jiawei Han , Micheline Kamber ,and Jian Pei, The Morgan Kaufmann Series in Data Management Systems.

cs412slides - Jiawei Han

https://www.coursera.org/specializations/data-mining



Why Data Mining?

- The amount of raw data stored in corporate databases is exploding. From trillions of point-of-sale transactions and credit card purchases to pixel-by-pixel images of galaxies, databases are now measured in gigabytes, terabytes, petabytes, exabytes and zettabytes. (One terabyte = one trillion bytes. A terabyte is equivalent to about 2 million books!)
- According to market intelligence company IDC, the 'Global Datasphere' in 2018 reached 18 zettabytes, , IDC predicts the world's data will grow to 175 zettabytes in 2025.
- For instance, every day, Wal-Mart uploads 20 million point-of-sale transactions to an A&T massively parallel system with 483 processors running a centralized database. Raw data by itself, however, does not provide much information.
- Another Example, social media usage in 2018. In just one minute:
  - Twitter users sent 473,400 tweets
  - Snapchat users shared 2 million photos
  - Instagram users posted 49,380 pictures
  - LinkedIn gained 120 new users.

Data Mining (utexas.edu)

Why Data Mining?

- Another mind-blowing data stats include:
  - Google processes more than 40,000 searches every second, or 3.5 billion searches a day.
  - 1.5 billion people are active on Facebook every day. That's one-fifth of the world's population.
- Companies need to rapidly turn these terabytes of raw data into significant insights into their customers and markets to guide their marketing, investment, , and management strategies.

(14) How Much Data Is There In The World? | LinkedIn

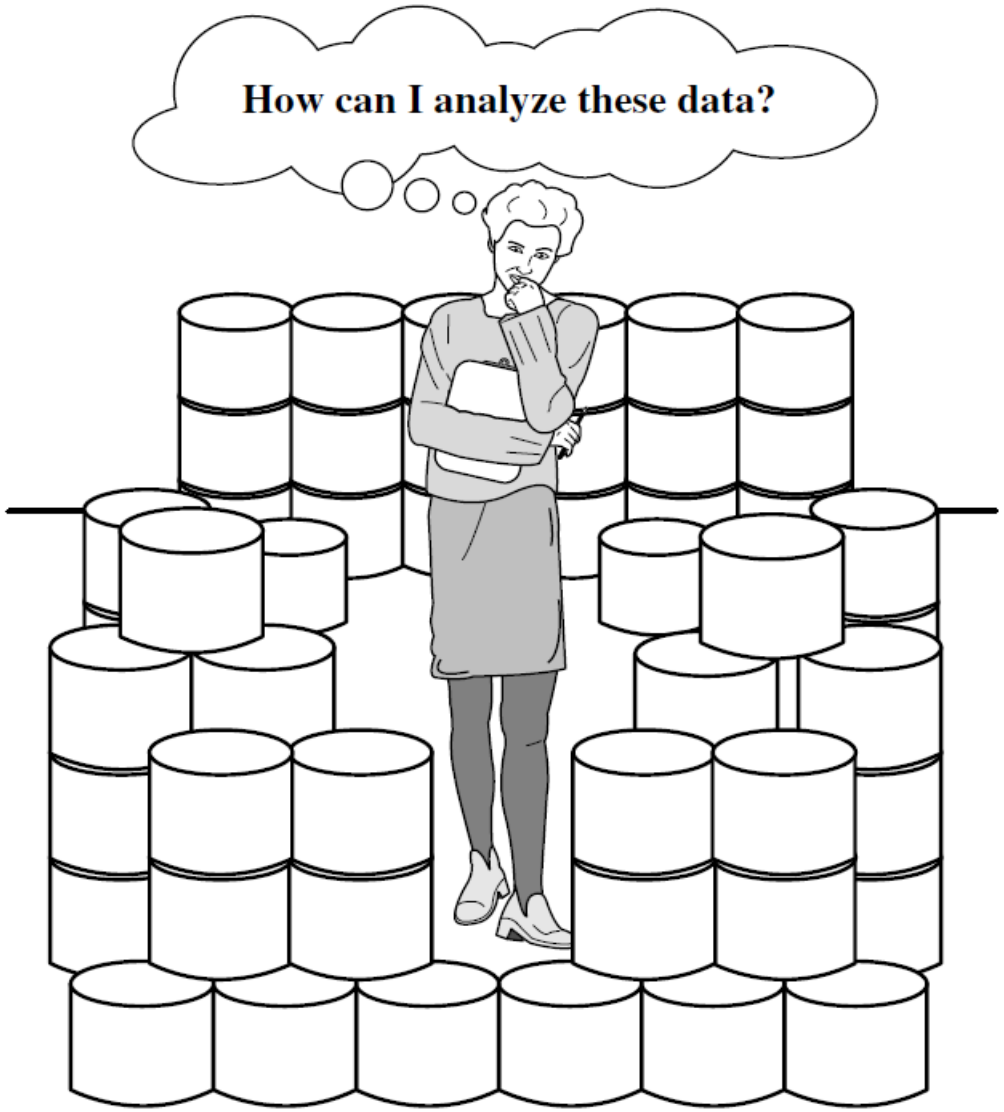
Why Data Mining?

- Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
- Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- Needed Automated analysis of massive data.

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 traditional systems can.

The world is data rich but information poor



What is Data Mining?

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Data mining provides a core set of technologies that help organizations anticipate future outcomes, discover new opportunities and improve business performance. It can be applied to a variety of customer issues in any industry – from customer segmentation and targeting, to fraud detection and credit risk scoring, to identifying adverse drug effects during clinical trials

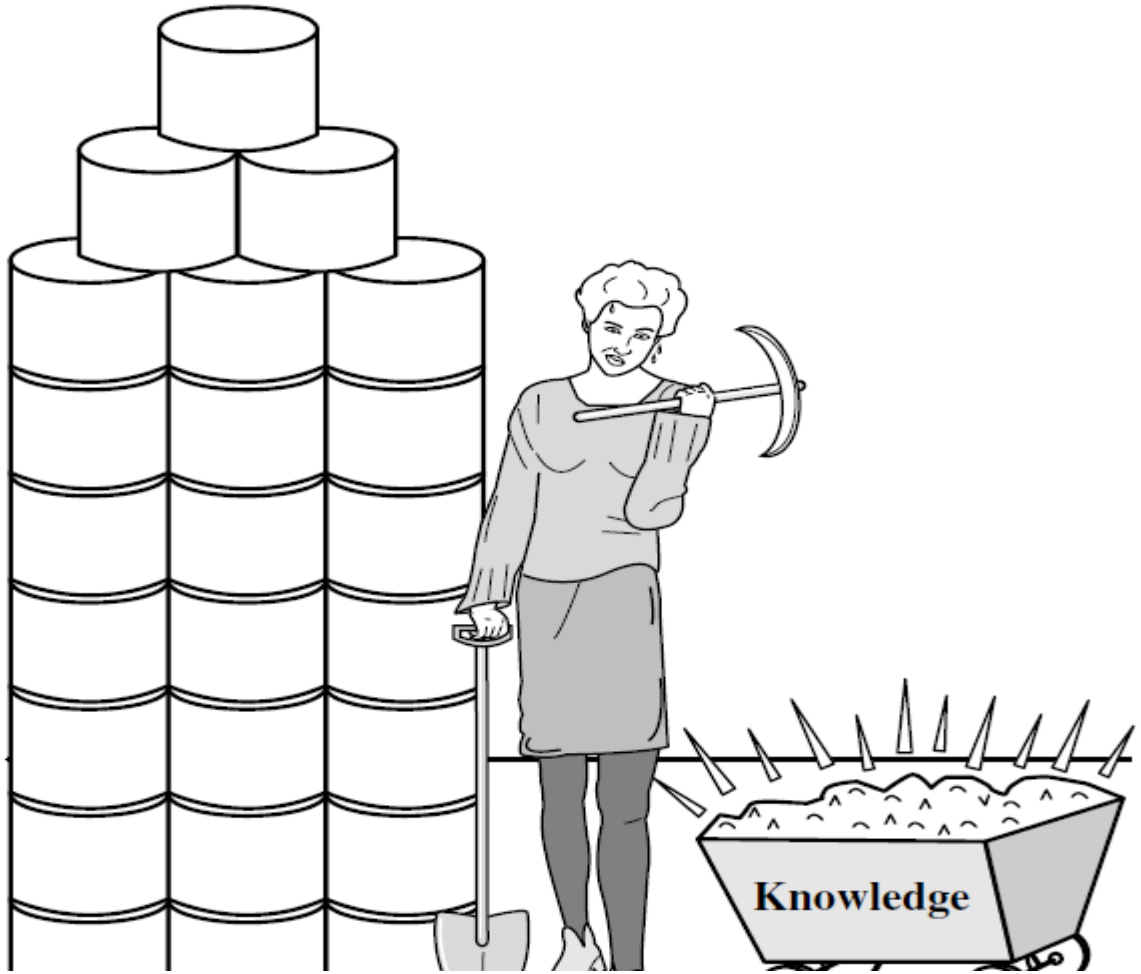
Data Mining From A to Z (sas.com)

What are patterns?

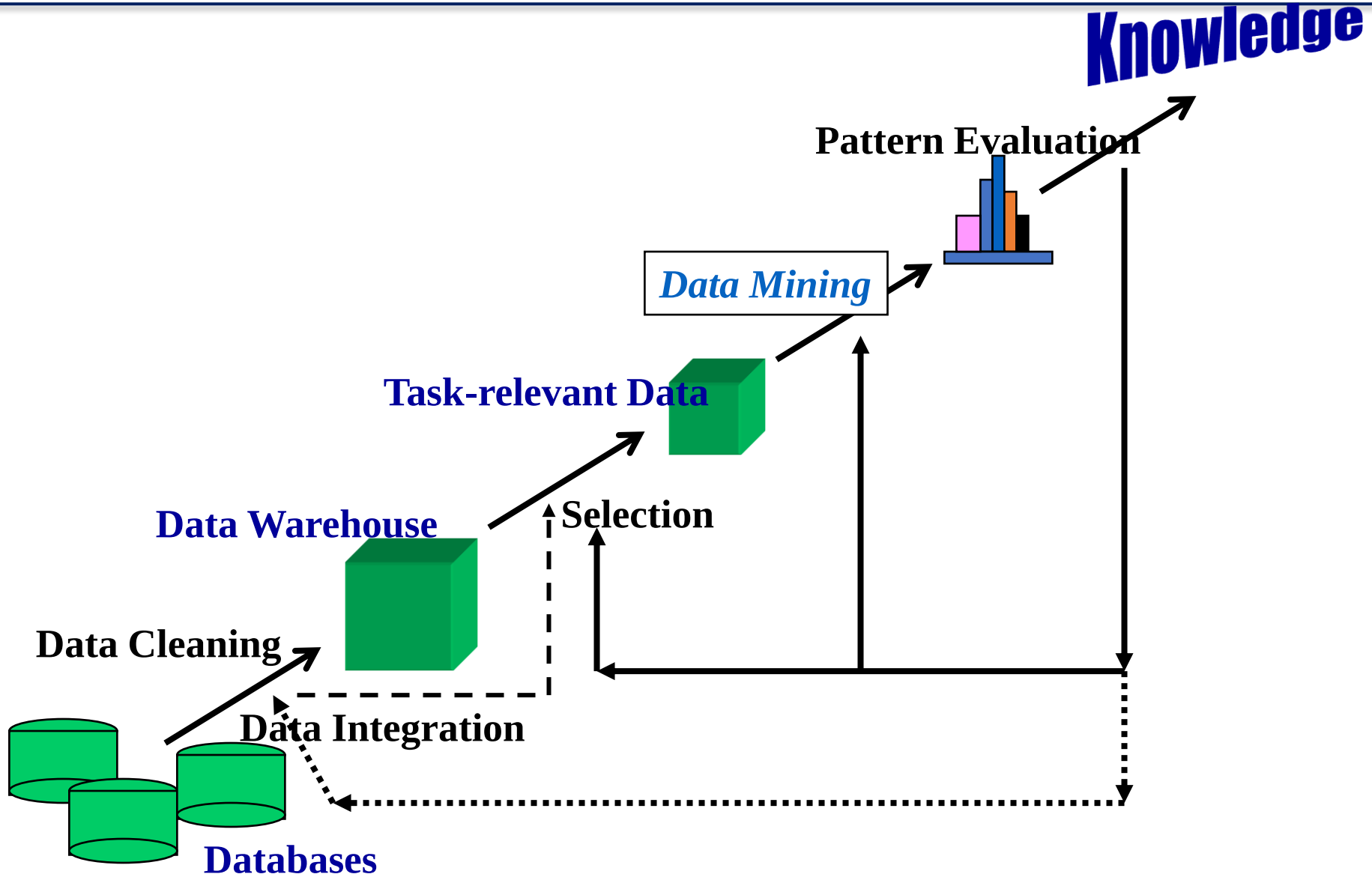
- Patterns: A set of items, subsequences, or substructures that occur frequently together (or strongly correlated) in a data set
- Patterns represent intrinsic and important properties of datasets
- Pattern discovery: Uncovering patterns from massive data sets

Goal of Data Mining

- Searching for knowledge (interesting patterns) in data.



Steps of knowledge Discovery from Data



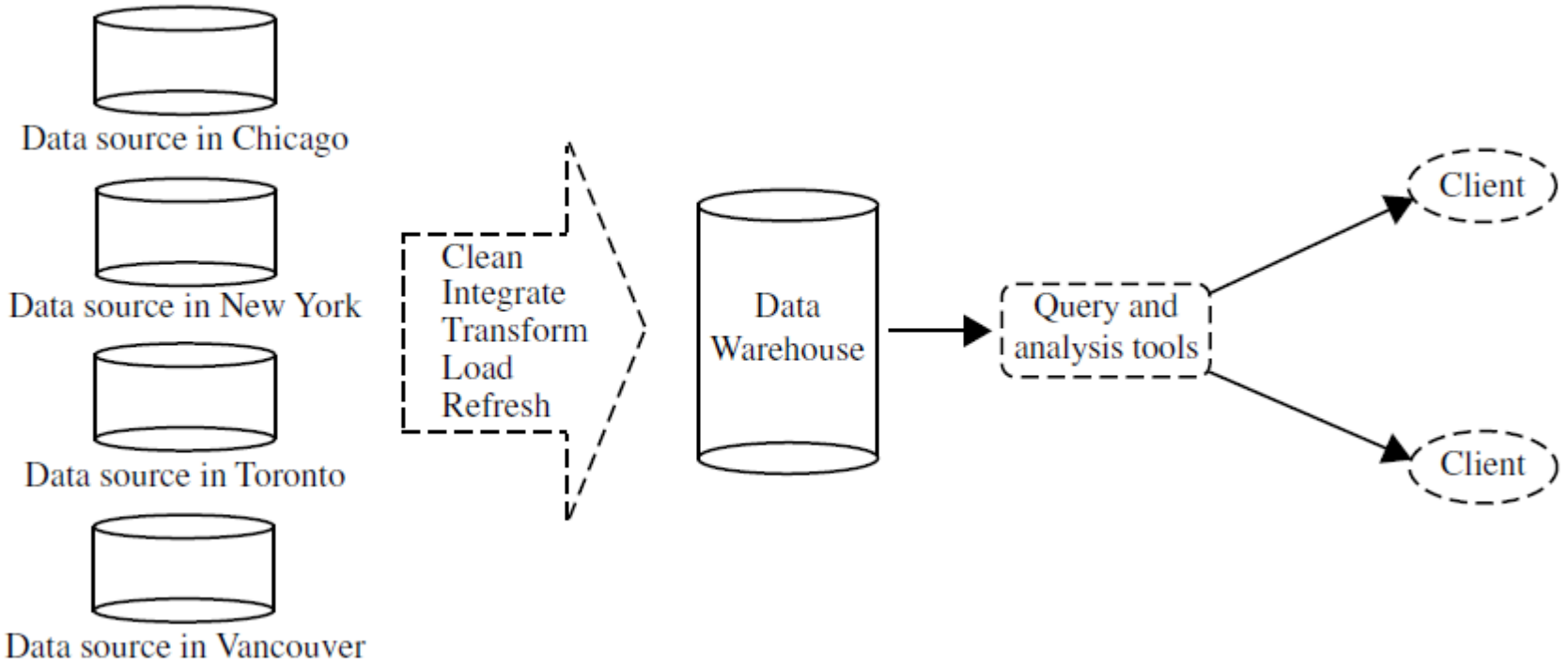
Steps of knowledge Discovery from Data

- Data cleaning (to remove noise and inconsistent data)
- Data integration (where multiple data sources may be combined).
- Data selection (where data relevant to the analysis task are retrieved from the database)
- Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
- Data mining extract data patterns
- Pattern evaluation
- Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

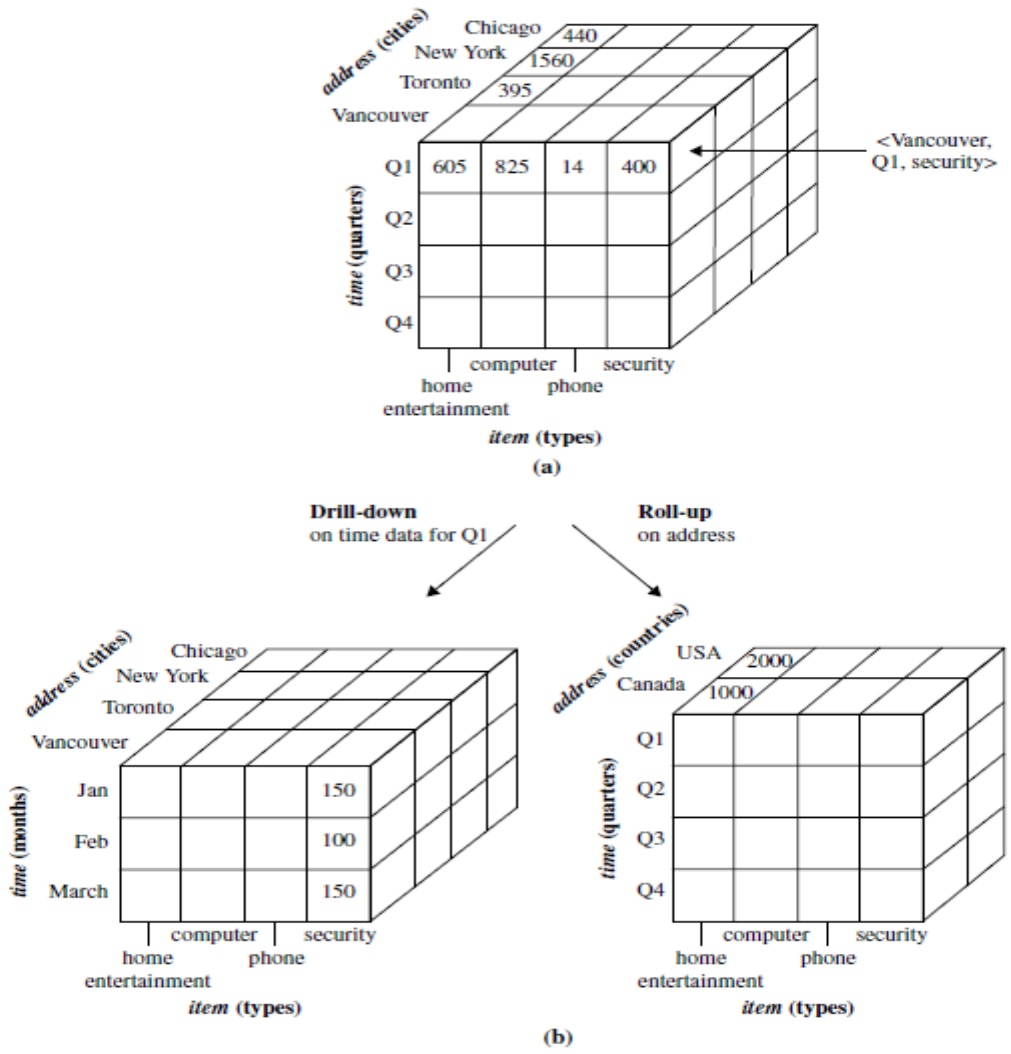
Data Warehouse

- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
- To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity). The data are stored to provide information froma historical perspective, such as in the past 6 to 12 months, and are typically summarized.
- For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per itemtype for each store or, summarized to a higher level, for each sales region.
- A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum sales amount. A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

Data Warehouse



Data Warehouse



Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Data Mining Function: (1) Generalization

- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics.
    - (1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms
    - (2) data discrimination, by comparison to the target class with one or a set of comparative classes (often called the contrasting classes).
  - Example
    - Data characterization, Summarize the characteristics of customers who spend more than $5000 a year The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.
    - Data discrimination, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patternsare patterns that occur frequently in data.
- There are many kinds of frequent patterns, including
  - frequent itemsets → a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together
  - frequent subsequences (also known as sequential patterns) A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card.
  - frequent substructures. A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences.
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data

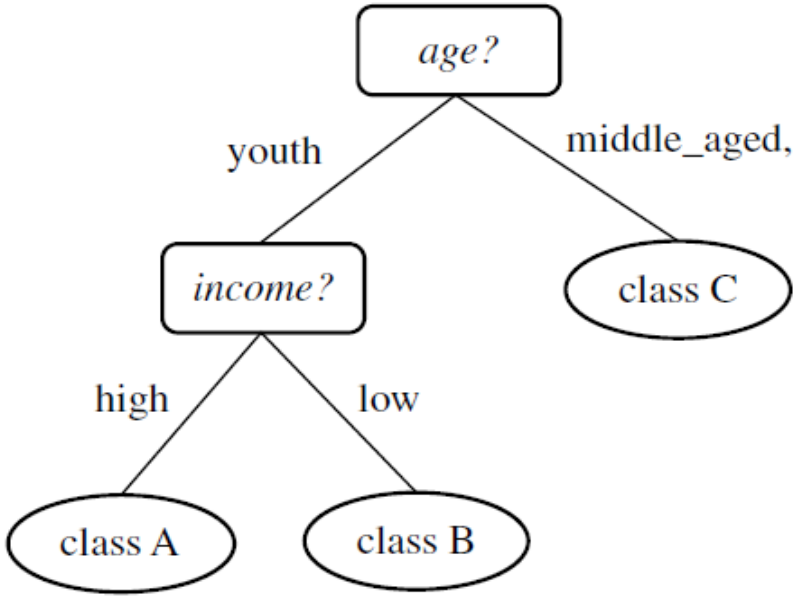
Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …
- The derived model may be represented in various forms, such as classification rules (i.e., IF-THEN rules), decision trees, or neural networks.

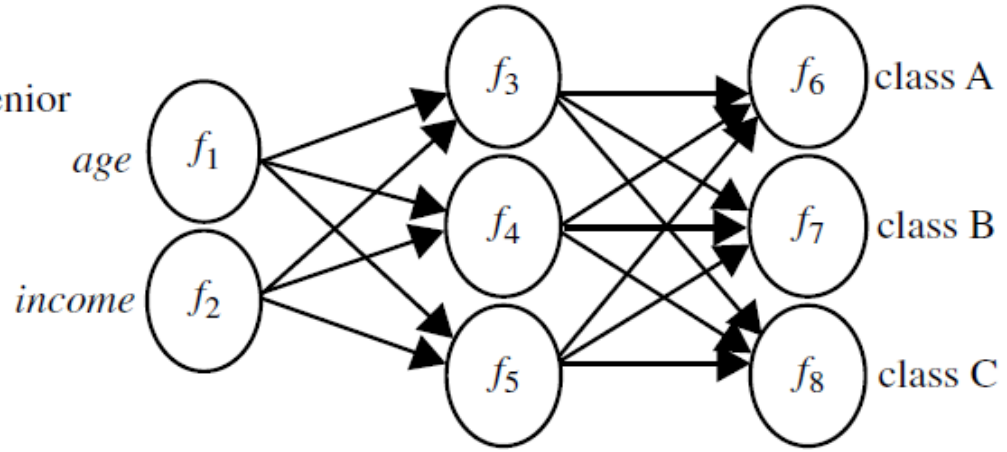
Data Mining Function: (3) Classification

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)



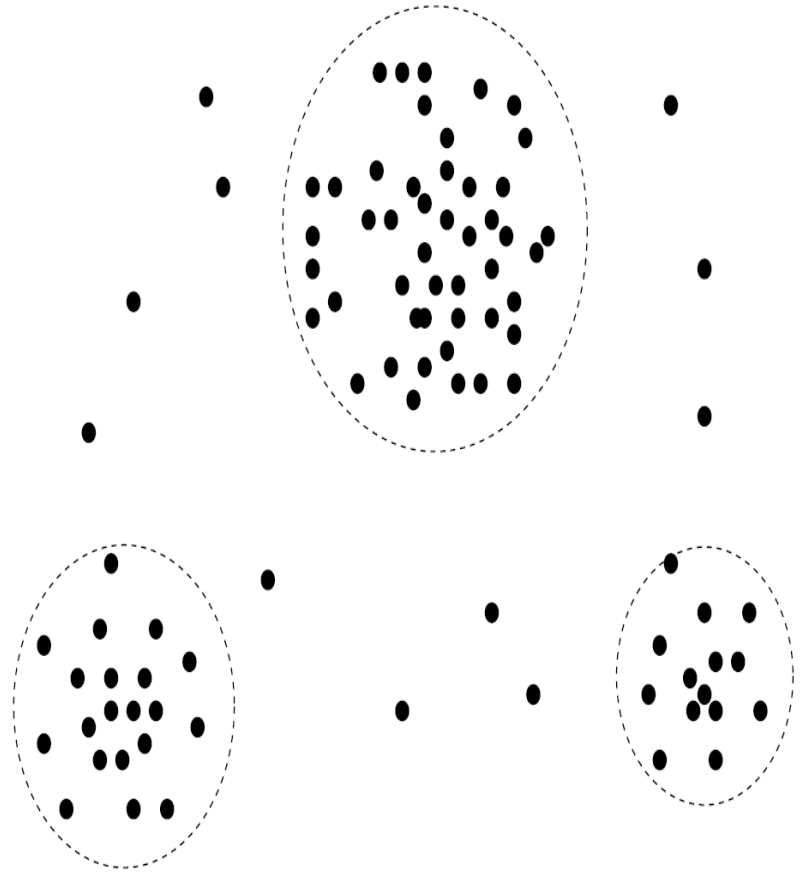
(b)



(c)

Data Mining Function: (4) Cluster Analysis

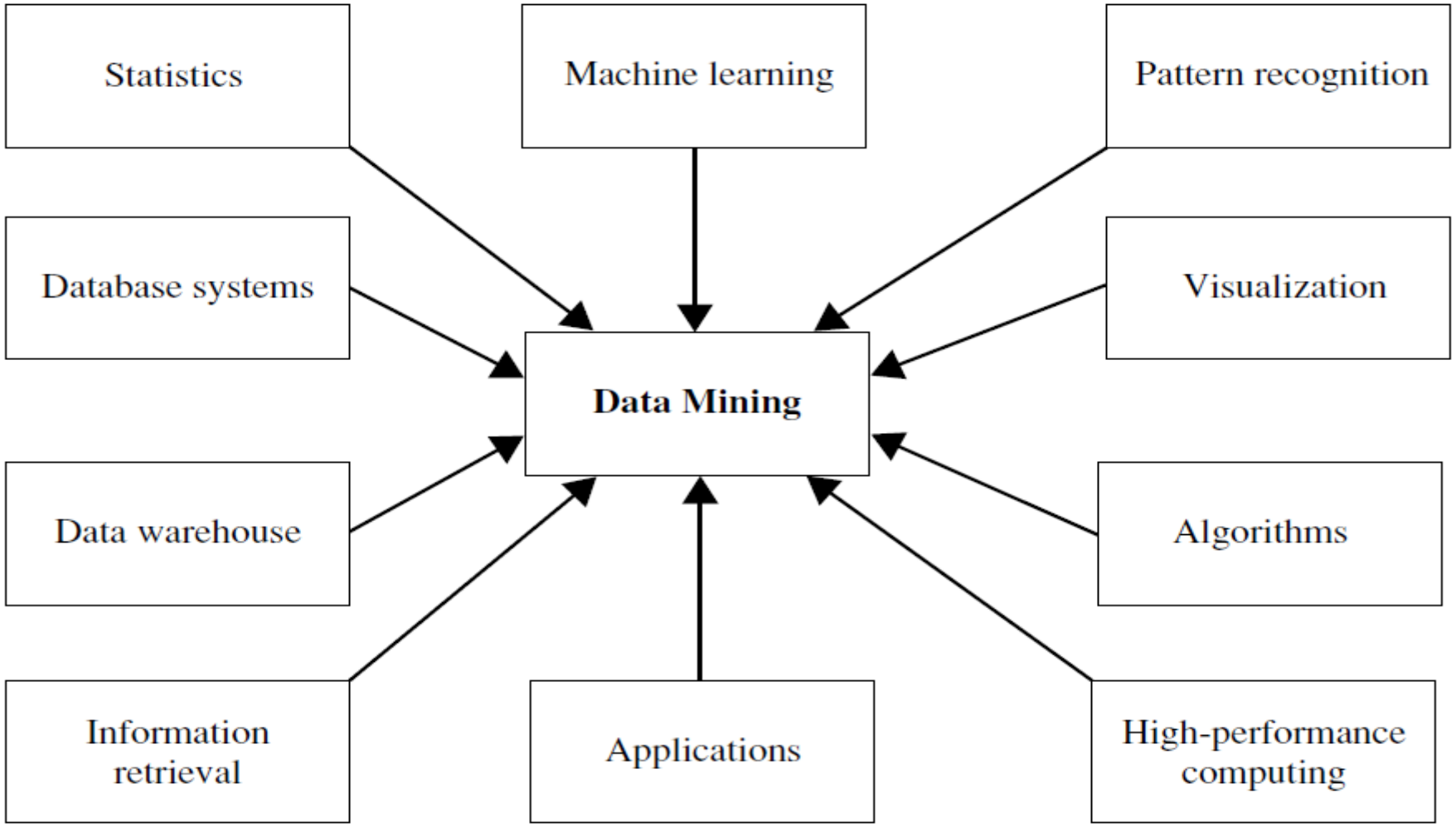
- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Example identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.



Data Mining Function: (5) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

Which Technologies Are Used?



Machine Learning

- Machine learning investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data. For example, a typical machine learning problem is to program a computer so that it can automatically recognize handwritten postal codes on mail after learning from a set of examples.

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Data Mining Phases