Data Mining:
Concepts and Techniques

— Unit1 —

— Introduction —

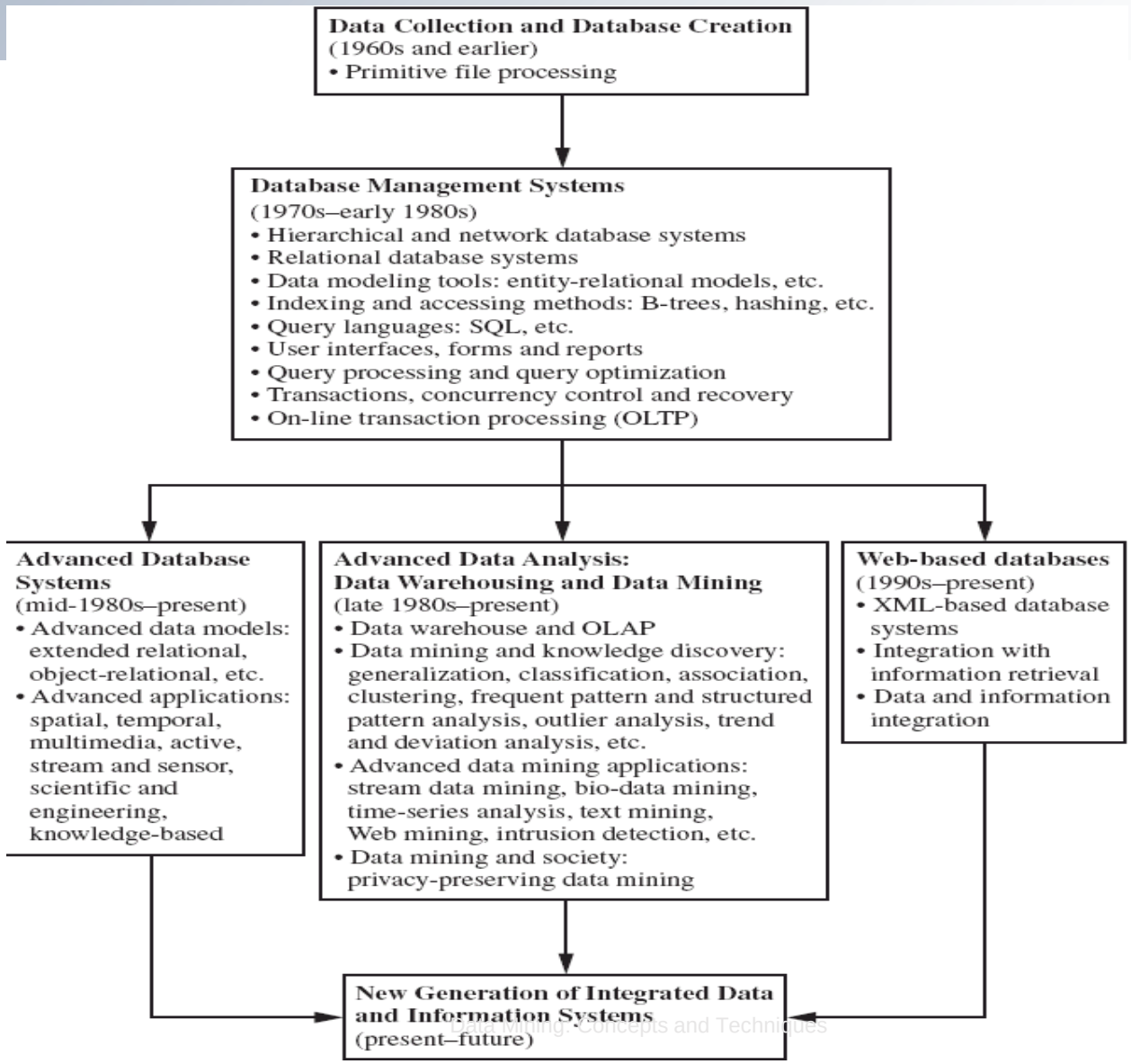
Outline

- 1.1 Motivation: Why data mining?
- 1.2 What is data mining?
- 1.3 Data Mining: On what kind of data?
- 1.4 Data mining functionality: What kinds of Patterns Can Be Mined?
- 1.5 Are all the patterns interesting?
- 1.6 Classification of data mining systems
- 1.7 Data Mining Task Primitives
- 1.8 Integration of data mining system with a DB and DW System
- 1.9 Major issues in data mining

1.1 Why Data Mining?

- The Explosive Growth of Data: from terabytes(1000^4) to yottabytes(1000^8)
  - Data collection and data availability
    - Automated data collection tools, database systems, web
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: bioinformatics, scientific simulation, medical research …
    - Society and everyone: news, digital cameras, …
- Data rich but information poor!
  - What does those data mean?
  - How to analyze data?
- Data mining — Automated analysis of massive data sets

Evolution of Database Technology



1.2 What Is Data Mining?

- Data mining (knowledge discovery from data)
  - 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.

Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

Ex.: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards,

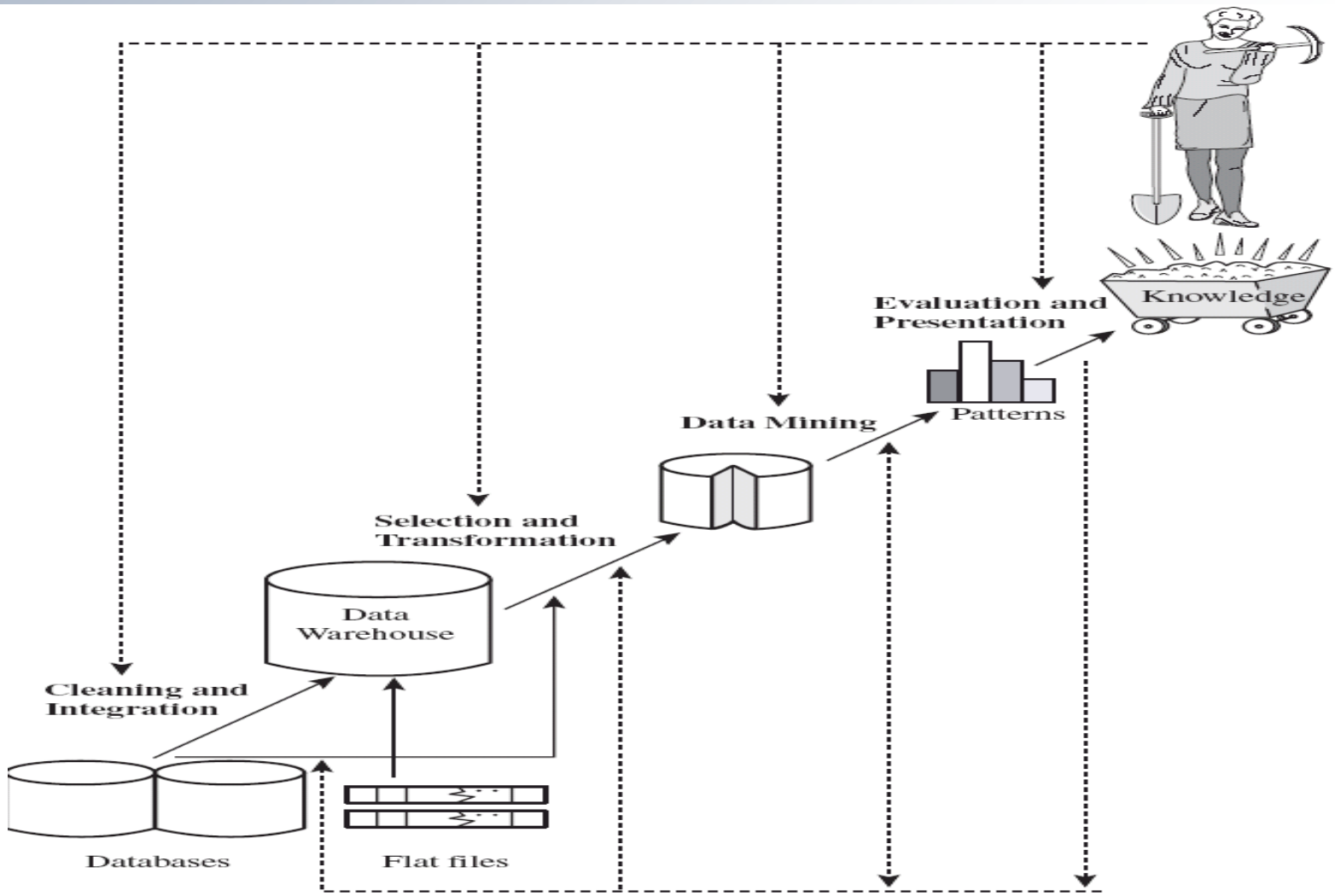
  discount coupons, customer complaint calls, surveys …
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest,

    income level, spending habits, etc.,
    - E.g. Most customers with income level 60k – 80k with food expenses $600 - $800 a month live in that area
  - Determine customer purchasing patterns over time
    - E.g. Customers who are between 20 and 29 years old, with income of 20k – 29k usually buy this type of CD player
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
  - E.g. Customers who buy computer A usually buy software B

Ex.: Market Analysis and Management (2)

- Customer requirement analysis
  - Identify the best products for different customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
    - E.g. Summarize all transactions of the first quarter from three different branches
      Summarize all transactions of last year from a particular branch
      Summarize all transactions of a particular product
  - Statistical summary information
    - E.g. What is the average age for customers who buy product A?
- Fraud detection
  - Find outliers of unusual transactions
- Financial planning
  - Summarize and compare the resources and spending

Knowledge Discovery (KDD) Process

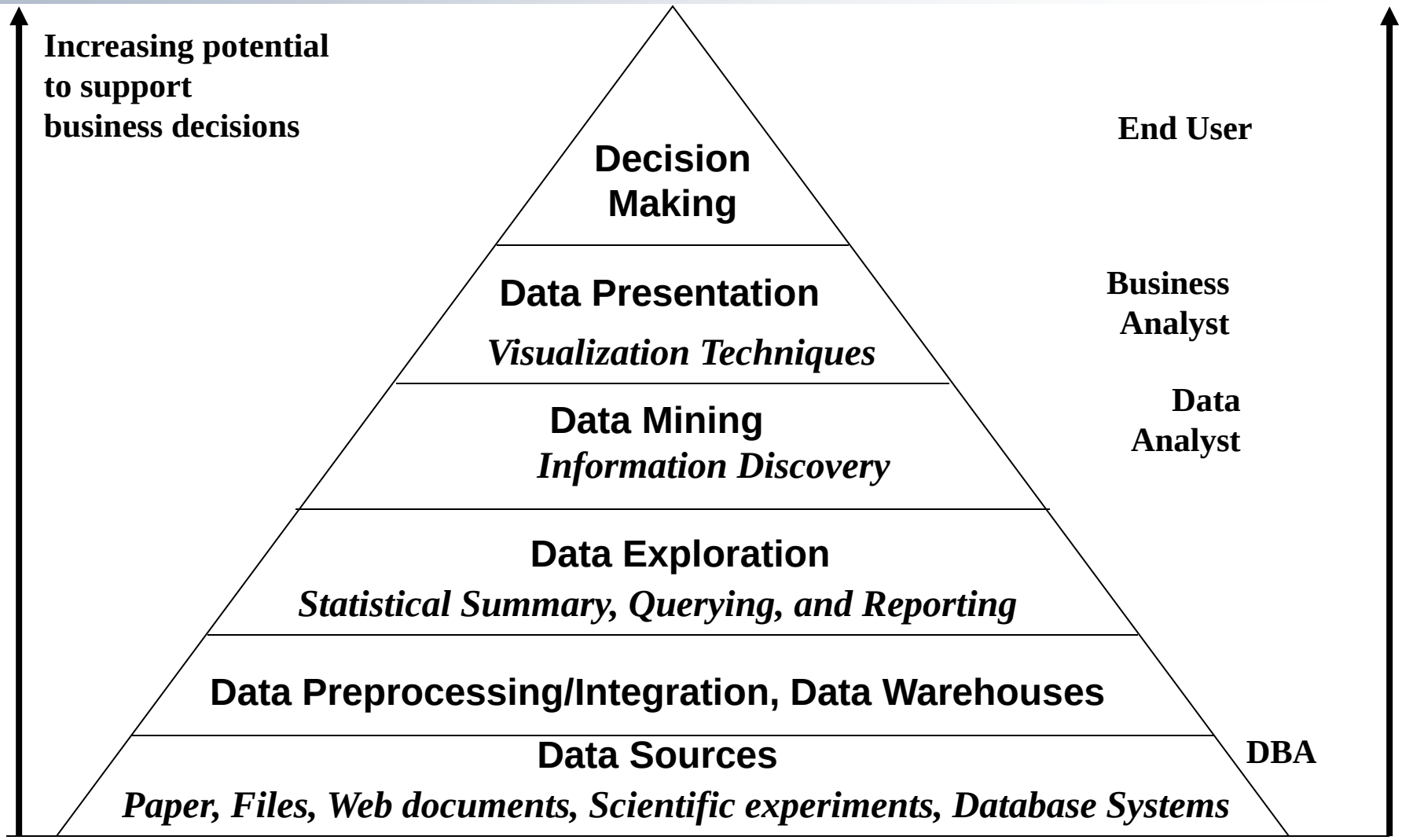


KDD Process: Several Key Steps

- Learning the application domain
  - relevant prior knowledge and goals of application
- Identifying a target data set: data selection
- Data processing
  - **Data cleaning** (remove noise and inconsistent data)
  - **Data integration** (multiple data sources maybe combined)
  - **Data selection** (data relevant to the analysis task are retrieved from database)
  - **Data transformation** (data transformed or consolidated into forms appropriate for mining)

    (Done with data preprocessing)
  - **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
  - **Pattern evaluation** (indentify the truly interesting patterns)
  - **Knowledge presentation** (mined knowledge is presented to the user with visualization or representation techniques)
- Use of discovered knowledge

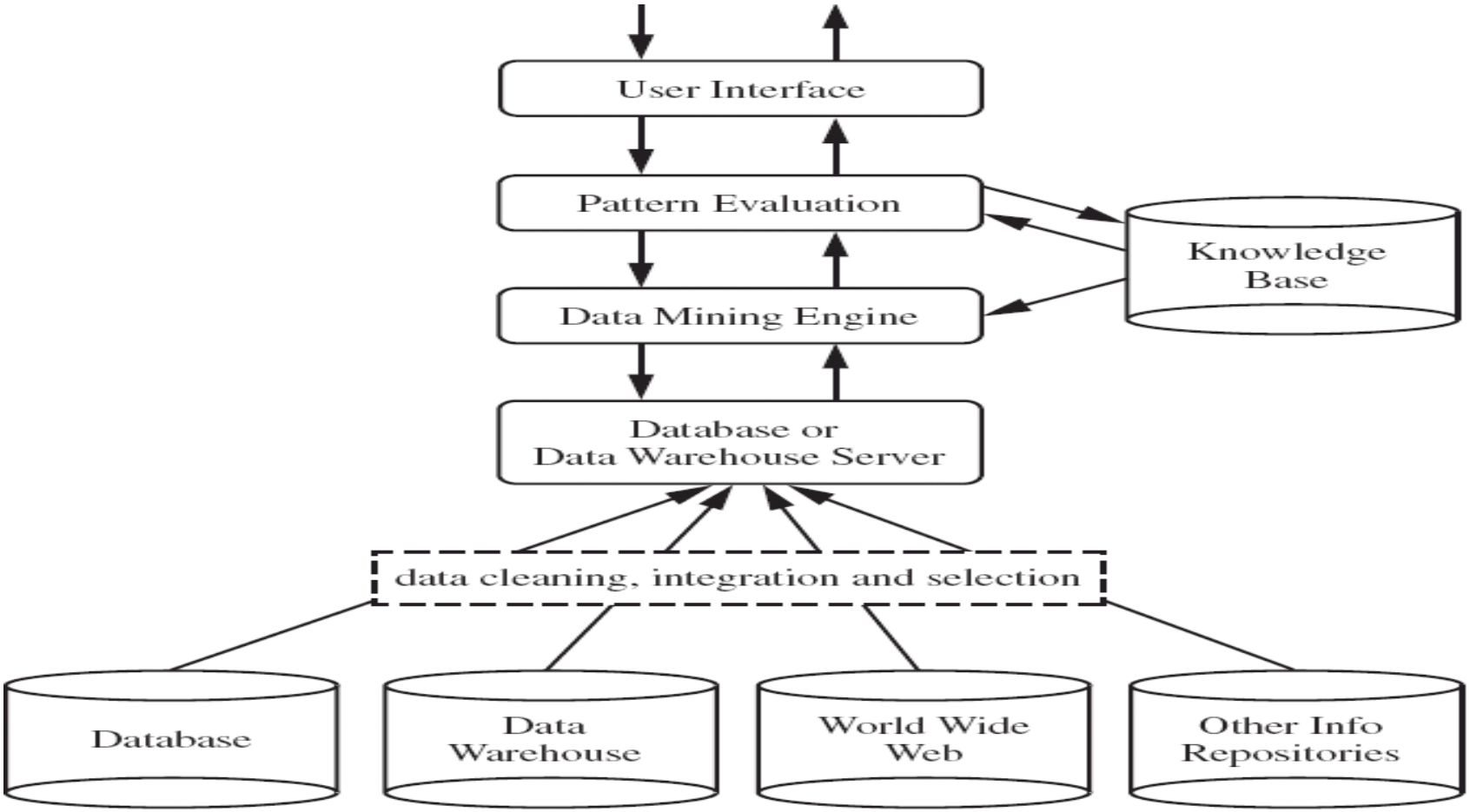
Data Mining and Business Intelligence



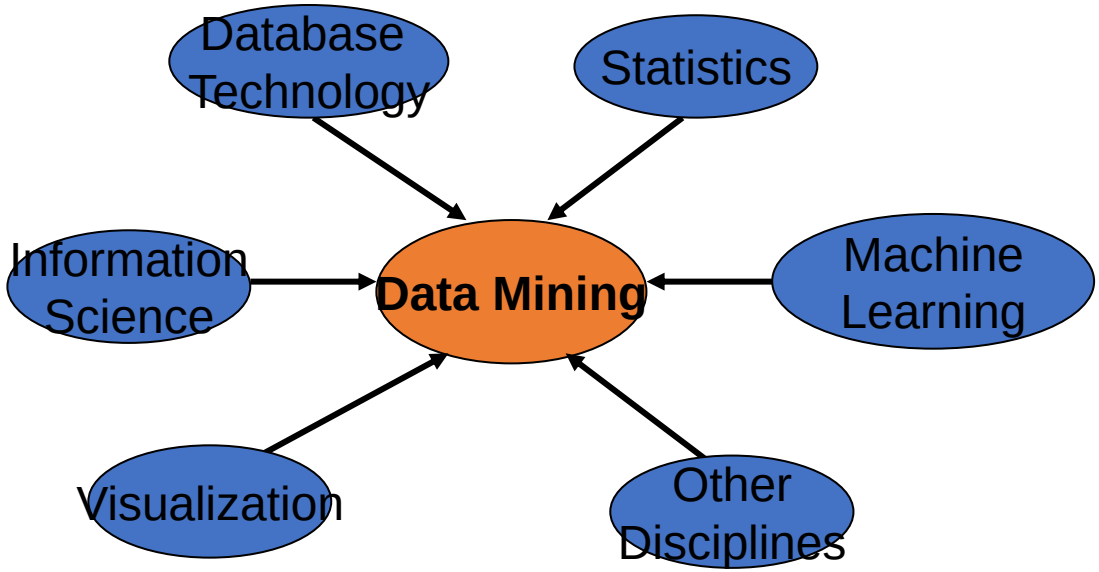
A typical DM System Architecture

- Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases,data warehouses or other important repositories.
- Database, data warehouse, WWW or other information repository (store data)
- Database or data warehouse server (fetch and combine data)
- Knowledge base (turn data into meaningful groups according to domain knowledge)
- Data mining engine (perform mining tasks)(like Characterization, association, correlation analysis, classification, prediction, cluster analysis, outlier analysis and evolution analysis.
- Pattern evaluation module (find interesting patterns)(integrated/interacts with Data Mining module)
- User interface (interact with the user)

A typical DM System Architecture (2)



Confluence of Multiple Disciplines



- Efficient and scalable Data Mining techniques must be used.
- An algorithm to be scalable, its running time should grow approximately linearly in proportion to the size of the data.
- Not all "Data Mining System" performs true data mining
  - machine learning system, statistical analysis (small amount of data)
  - Database system (information retrieval, deductive querying…)

1.3 On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Object-Relational Databases
  - Temporal Databases, Sequence Databases, Time-Series databases
  - Spatial Databases and Spatiotemporal Databases
  - Text databases and Multimedia databases
  - Heterogeneous Databases and Legacy Databases
  - Data Streams
  - The World-Wide Web

Relational Databases

- DBMS – database management system, contains a collection of interrelated data's from databases.

Set of software program manage and access data.

e.g. Faculty database, student database, publications database

- Each database contains a collection of tables and functions to manage and access the data.

  e.g. student_bio, student_graduation, student_parking
- Each table contains set of attributes(columns) and set of tuples(rows), with columns as attributes of data and rows as records.
- Tables can be used to represent the relationships between or among multiple tables.
- Each tuple in the relational table represents an object identified by a unique key and described by a set of attribute values.
- Relational data can be accessed by database queries (SQL).

Relational Databases (2) – *AllElectronics* store

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
...	...	...	...	...	...	...

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

works_at

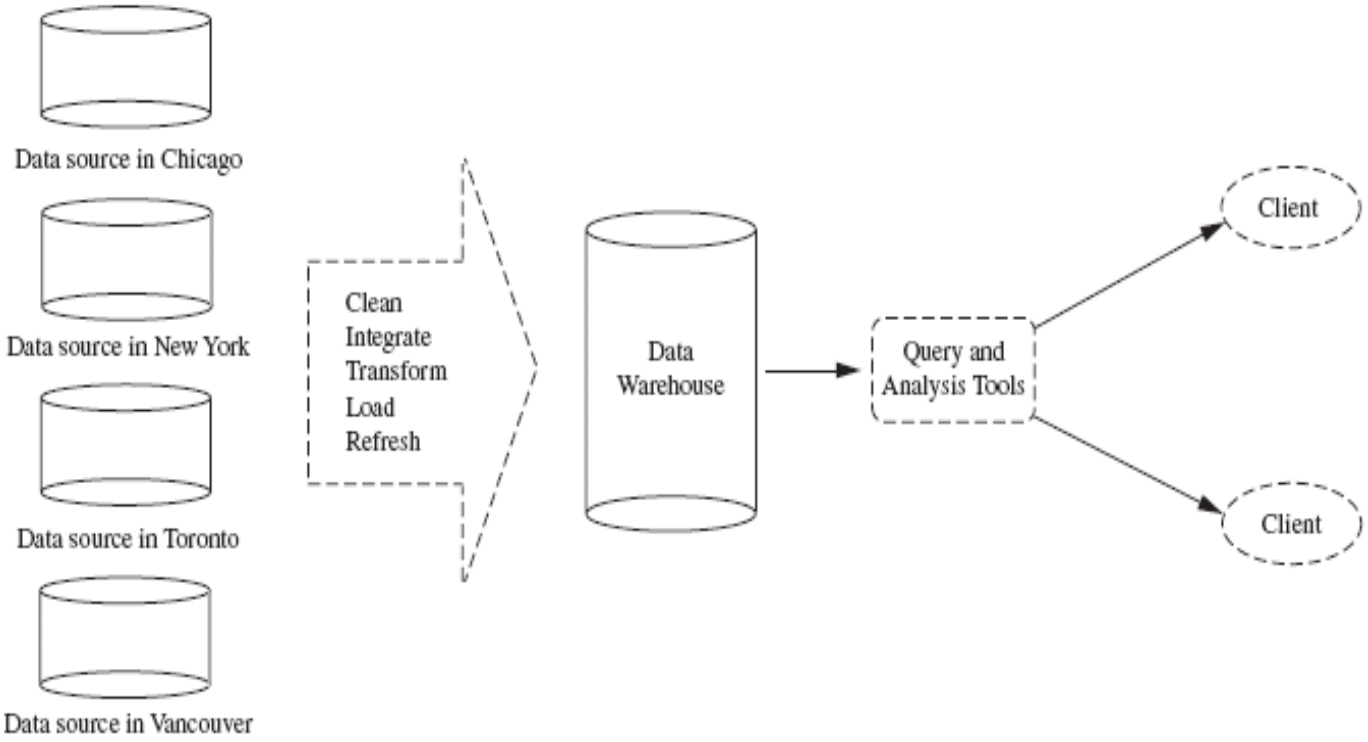
empl_ID	*branch_ID*
E55	B1
...	...

Relational Databases (3)

- With a relational query language, e.g. SQL, we will be able to find answers to questions such as:
  - How many items were sold last year?
  - Who has earned commissions higher than 10%?
  - What is the total sales of last month for Dell laptops?
- When data mining is applied to relational databases, we can search for trends or data patterns.
- Relational databases are one of the most commonly available and rich information repositories, and thus are a major data form in our study.

Data Warehouses

- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.

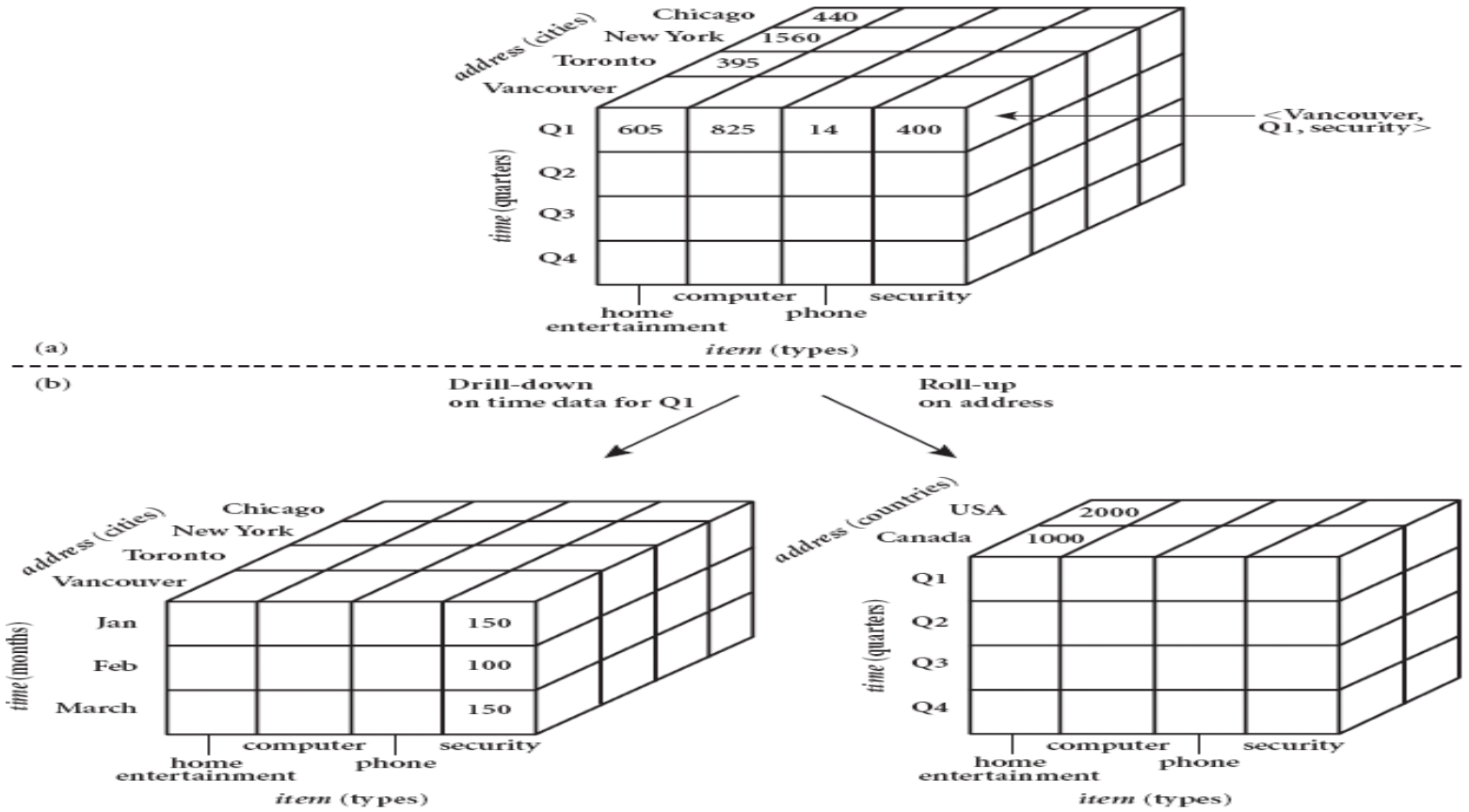


Data Warehouses (2)

- To facilitate decision making, Data are organized around major subjects, e.g. customer, item, supplier and activity.
- Provide information from a historical perspective (e.g. from the past 5 – 10 years)
- Typically summarized to a higher level (e.g. a summary of the

  transactions per item type for each store)
- User can perform drill-down or roll-up operation to view the data at different degrees of summarization.(**Drill-down and Roll-up. ... Drill-down refers to the process of viewing data at a level of increased detail, while roll-up refers to the process of viewing data with decreasing detail.)**
- A data warehouse is usually modelled by a multidimensional database structure,
- where 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 *sales amount*.
- The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube.

Data Warehouses (3)



Transactional Databases

- Consists of a file where each record represents a transaction.
- A transaction typically includes a unique transaction ID and a list of the items making up the transaction.

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
. . .	. . .

- Either stored in a flat file or unfolded into relational tables.(**It allows the user to specify data attributes, such as columns and data types table by table, and stores those attributes separate from applications)**
- Easy to identify items that are frequently sold together.
- Market Basket Analysis enable you to bundle groups of items together. Eg: printers are commonly purchased together with computers.

Object-Relational Databases

- Object-relational databases are constructed based on an object-relational data model.
- This model extends the relational model by providing a rich data type for handling complex objects and object orientation.
- inherits the essential concepts of object-oriented databases.
- entity is considered as an object.
- *AllElectronics* example, objects can be individual employees, customers, or items.
- Data and code relating to an object are *encapsulated* into a single unit.
- Objects that share a common set of properties can be grouped into an object class.

- A **set of variables** that describe the objects. These correspond to attributes in the entity-relationship and relational models.
- A **set of messages** that the object can use to communicate with other objects, or with the rest of the database system.
- A **set of methods**, where each method holds the code to implement a message. Upon receiving a message, the method returns a value in response.
- For instance, the method for the message *get photo*(*employee*) will retrieve and return a photo of the given employee object.

Temporal Databases, Sequence Databases, and Time-Series Databases

- A **temporal database** typically stores relational data that include time-related attributes.
- These attributes may involve several timestamps, each having different semantics.
- A **sequence database** stores sequences of ordered events, with or without a concrete notion of time.
- Eg: customer shopping sequences.
- A **time-series database** stores sequences of values or events obtained over repeated measurements of time (e.g., hourly, daily, weekly).
- Eg: data collected from the stock exchange, inventory control.
- Data mining techniques can be used to find the characteristics of object evolution, or the trend of changes for objects in the database.

Spatial Databases and Spatiotemporal Databases

- Spatial databases contain spatial-related information.
- Eg: geographic(map) databases commonly used in vehicle navigation and dispatching systems.
- Spatial data may be represented in raster format, consisting of *n*-dimensional bit maps or pixel maps.
- ***What kind of data mining can be performed on spatial databases?***
- Data mining may uncover patterns describing the characteristics .
- Eg: Characteristics of houses located near a specified kind of location, such as a park.
- Clusters and outliers can be identified by spatial cluster analysis.

- spatial classification can be performed to construct models for prediction based on the relevant set of features of the spatial objects.
- Furthermore, “spatial data cubes” may be constructed to organize data into multidimensional structures and hierarchies, on which OLAP operations (such as drill-down and roll-up) can be performed.
- A spatial database that stores spatial objects that change with time is called a **spatiotemporal database.**

Text Databases and Multimedia Databases

- **Text databases** are databases that contain word descriptions for objects. These word descriptions are usually not simple keywords but rather long sentences or paragraphs.
- Eg: product specifications, error or bug reports, warning messages etc.
- Text databases may be highly unstructured.
- *What can data mining on text databases uncover?*
- By mining text data, one may uncover general and concise descriptions of the text documents, keyword or content associations.
- **Multimedia databases** store image, audio, and video data. They are used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems etc.

Heterogeneous Databases and Legacy Databases

- A **heterogeneous database** consists of a set of interconnected, autonomous component databases.
- The components communicate in order to exchange information and answer queries.
- A **legacy database** is a group of *heterogeneous databases* that combines different kinds of data systems.
- Eg: relational or object-oriented databases, hierarchical databases, network databases, spread sheets, multimedia databases, or file systems.

Data Streams

- Many applications involve the generation and analysis of a new kind of data, called **stream data**.
- where data flow in and out of an observation platform (or window) dynamically.
- Such data streams have the following **unique features**:
- *huge or possibly infinite volume.*
- *dynamically changing.*
- *flowing in and out in a fixed order.*
- *allowing only one or a small number of scans.*
- *demanding fast (often real-time) response time*.
- Eg: scientific and engineering data, time-series data, stock exchange, video surveillance, network traffic, weather or environment monitoring.

The World Wide Web

- Capturing user access patterns in such distributed information environments is called **Web usage mining (or Weblog mining).**
- **Authoritative Web page** analysis based on linkages among Web pages can help rank Web pages based on their importance, influence, and topics.
- **Automated Web page clustering and classification** help group and arrange Web pages in a multidimensional manner based on their contents.
- **Web community analysis** helps identify hidden Web social networks and communities and observe their evolution.

1.4 Data Mining Functionalities - What kinds of patterns can be mined?

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
- In general, data mining tasks can be classified into two categories: descriptive and predictive.
- **Descriptive mining** tasks characterize the general properties of the data in the database. (Identifying web pages that are accessed together.(human interpretable pattern))
- **Predictive mining** tasks perform inference on the current data in order to make predictions.(Judge if a patient has specific disease based on his/her medical tests results.)

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- Concept/Class Description: Characterization and Discrimination
  - Data can be associated with classes or concepts.
    - E.g. classes of items – computers, printers, …

      concepts of customers – bigSpenders, budgetSpenders, …
    - How to describe these items or concepts?
  - Descriptions can be derived via
    - Data characterization – summarizing the general characteristics of a target class of data.
      - E.g. summarizing the characteristics of customers who spend more than $1,000 a year at *AllElectronics*. Result can be a general profile of the customers, such as 40 – 50 years old, employed, have excellent credit ratings.

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

comparative classes

- E.g. Compare the general features of software products whole sales increase by 10% in the last year with those whose sales decrease by 30% during the same period

- Or both of the above

- Mining Frequent Patterns, Associations and Correlations
  - Frequent itemset: a set of items that frequently appear together in a transactional data set (e.g. milk and bread)

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- Association Analysis: find frequent patterns
  - E.g. a sample analysis result – an association rule:

    buys(X, “computer”) => buys(X, “software”) [support = 1%, confidence = 50%]

    (if a customer buys a computer, there is a 50% chance that she will buy software.

    1% of all of the transactions under analysis showed that computer and software

    are purchased together.)
  - Associations rules are discarded as uninteresting if they do not satisfy both

    a minimum support threshold and a minimum confidence threshold.
- A data mining system may find association rules like

  age(X, “20:::29”)^income(X, “20K:::29K”)) => buys(X, “CD player”)
- [support = 2%, confidence = 60%].
- multidimensional association rule.
- Correlation Analysis: additional analysis to find statistical correlations between

  associated pairs

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- Classification and Prediction
  - Classification
    - The process of finding a model that describes and distinguishes the data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
    - The derived model is based on the analysis of a set of training data (data objects whose class label is known).
    - The model can be represented in *classification (IF-THEN) rules,* decision trees, *neural networks,* etc.
  - Prediction
    - Predict missing or unavailable numerical data values

1.4 Data Mining Functionalities

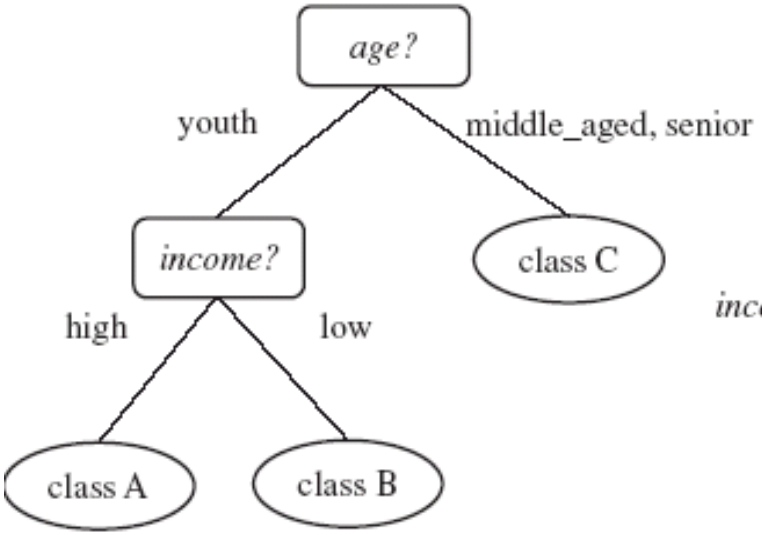
- What kinds of patterns can be mined?

Classification rules can be represented various forms such as i) if then rules ii)decision tree (or) iii)neural networks

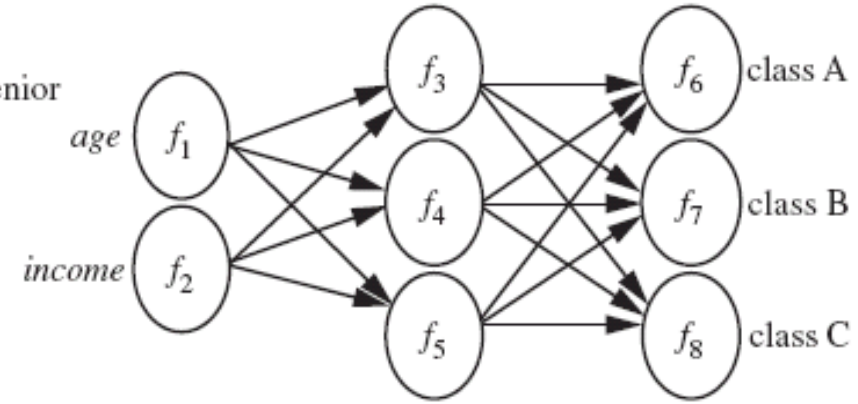
(a)

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")

(b)

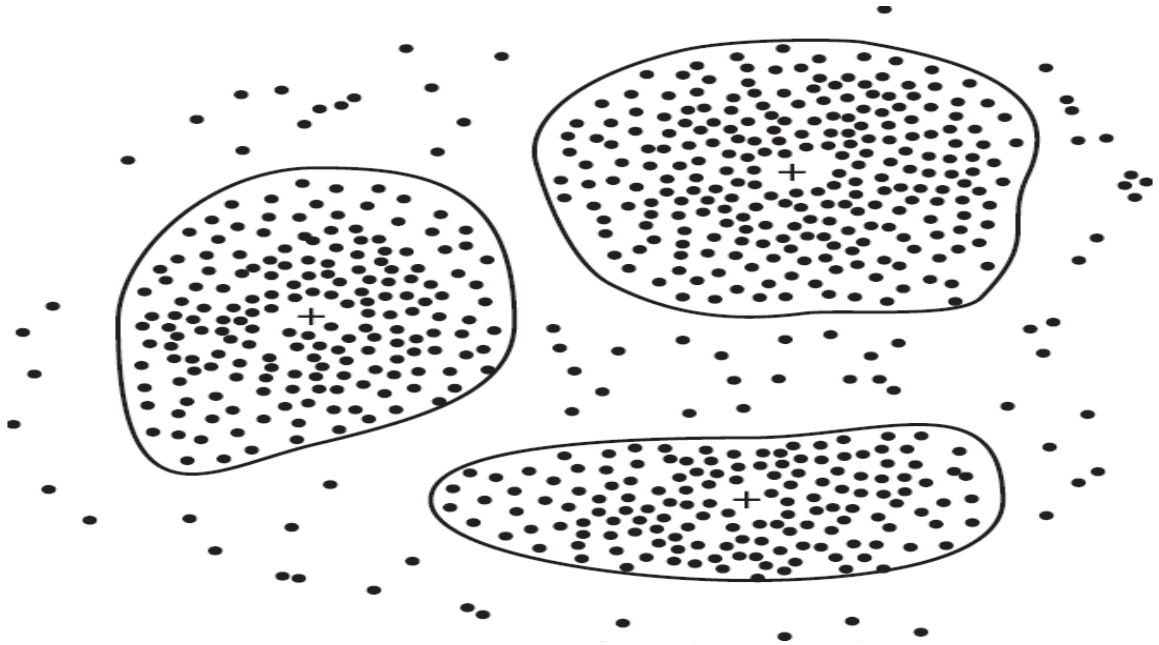


(c)



Data Mining Functionalities (2)

- Cluster Analysis
- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Class label is unknown: group data to form new classes.
  - Clusters of objects are formed based on the principle of *maximizing intra-class similarity & minimizing interclass similarity*
    - E.g. Identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.
- A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster "center" is marked with a "+".
- In general, the class labels are not present in the training data simply because they are



- **Centralized** - each cluster is represented by a single vector mean, and a object value is compared to these mean values
  - Distributed – the cluster is built using statistical distributions
  - Connectivity – he connectivity on these models is based on a distance function between elements
  - Group – algorithms have only group information
  - Graph – cluster organization and relationship between members is defined by a graph linked structure
  - Density – members of the cluster are grouped by regions where observations are dense and similar

1.5 Are All of the Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
- A pattern is interesting if it is
  - easily understood by humans
  - valid on new or test data with some degree of certainty,
  - potentially useful
  - novel
  - validates some hypothesis that a user seeks to confirm
- An interesting measure represents *knowledge* !

1.5 Are All of the Patterns Interesting?

- Objective measures
  - Based on statistics and structures of patterns, e.g., support, confidence, etc. (Rules that do not satisfy a threshold are considered uninteresting.)
  - Support(X =>Y) = P(XUY) - transactions contains both X and Y
  - Confidence(X =>Y) = P(X|Y) – Probability that the transaction containing X also Contains Y.
- Subjective measures
  - Reflect the needs and interests of a particular user.
    - E.g. A marketing manager is only interested in characteristics of customers who shop frequently.
  - Based on user's belief in the data.
    - e.g., Patterns are interesting if they are unexpected, or can be used for strategic planning, etc
- Objective and subjective measures need to be combined.

1.6 Classification of data mining systems

- Classification according to the kinds of Database mined:
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Classification according to the kinds of** Knowledge mined: based on mining functionalities such as
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- data mining systems can be distinguished based on the granularity or levels of abstraction of the knowledge mined.
- primitive-level knowledge (at a raw data level).
- Knowledge at multiple levels (several levels of abstraction).

Data mining systems can also be categorized as

- those that mine data regularities (commonly occurring patterns)
- those that mine data irregularities (such as exceptions, or outliers).

1.6 Classification of data mining systems

- **Classification according to the kinds of Techniques utilized:**
  - **Degree of user interaction** involved (e.g., autonomous systems, interactive exploratory systems, query-driven systems).
  - The **methods of data analysis** employed (e.g., database-oriented or data warehouse–oriented techniques, machine learning, statistics etc..).
- **Classification according to the applications adapted:**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

1.7 Data Mining Task Primitives

- A data mining task can be specified in the form of a data mining query, which is input to the data mining system.
- A data mining query is defined in terms of data mining task primitives.
- These primitives allow the user to *interactively* communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.
- The set of *task-relevant data* to be mined
- The *kind of knowledge* to be mined
- The *background knowledge* to be used in the discovery process
- The *interestingness measures and thresholds* for pattern evaluation
- The expected *representation for visualizing* the discovered patterns

1.7 Data Mining Task Primitives

- The primitives specify:

  (1) **The set of task-relevant data to be mined** – which portion of the database or set of data to be used.

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

1.7 Data Mining Task Primitives

- The primitives specify:

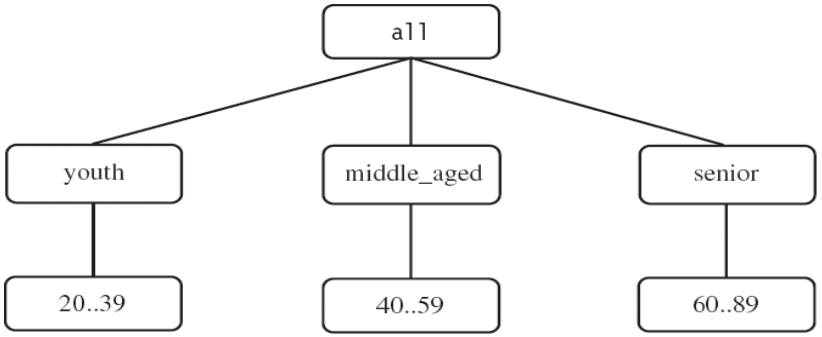
  (2) **The kind of knowledge to be mined** – what DB functions to be performed

  - Characterization
  - Discrimination
  - Association
  - Classification/prediction
  - Clustering
  - Outlier analysis
  - Other data mining tasks

1.7 Data Mining Task Primitives

(3) The background knowledge to be used in the discovery process –

what domain knowledge, concept hierarchies, etc.

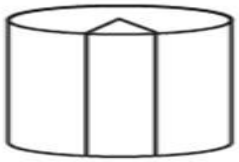


(4) Interestingness measures and thresholds – support, confidence, etc.

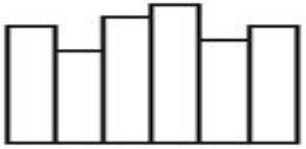
(5) Visualization methods – what form to display the result, e.g. rules,

tables, charts, graphs, …

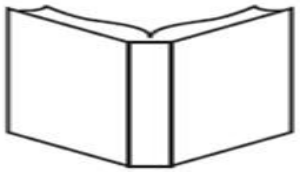
Primitives for specifying a data mining task.



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

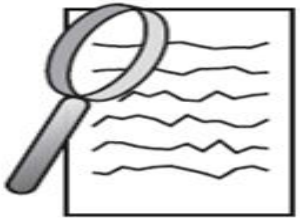


Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering



Background knowledge
Concept hierarchies
User beliefs about relationships in the data

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



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

1.7 Data Mining Task Primitives

- DMQL – Data Mining Query Language
  - Designed to incorporate these primitives
  - Allow user to interact with DM systems
  - Providing a standardized language like SQL

Why Data Mining Query Language?

- Automated vs. query-driven?
  - Finding all the patterns autonomously in a database?—unrealistic because the patterns could be too many but uninteresting
- Users must be provided with a set of primitives to be used to communicate with the data mining system
- Incorporating these primitives in a data mining query language
  - More flexible user interaction
  - Foundation for design of graphical user interface
  - Standardization of data mining industry and practice

1.8 Integration of Data Mining and Data Warehousing

- No coupling
  - Flat file processing, DM system will not utilize any functions of a DB/DW system (In this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms. The data mining result is stored in another file.)
  - Not recommended
- Loose coupling
  - Fetching data from DB/DW(the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory managed by these systems and performs data mining on that data.)
  - Does not explore data structures and query optimization methods provided by DB/DW system
  - Difficult to achieve high scalability and good performance with large data sets

1.8 Integration of Data Mining and Data Warehousing

- Semi-tight-(In this scheme, the data mining system is linked with a database or a data warehouse system and in addition to that, efficient implementations of a few data mining primitives can be provided in the database)
  - Efficient implementations of a few essential data mining primitives in a DB/DW system are provided, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some stat functions.

  Some frequently used intermediate mining results can be precomputed And stored in the DB/DW system.
  - Enhanced DM systems performance
- Tight -(the data mining system is smoothly integrated into the database or data warehouse system)
  - DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query analysis, data structures, indexing, query processing methods of a DB/DW system.
  - A uniform information processing environment, highly desirable
  - High system Performance.

1.9 Major Issues in Data Mining

- Mining methodology and User interaction
  - Mining different kinds of knowledge
    - DM should cover a wide spectrum of data analysis and knowledge discovery tasks
    - Enable to use the database in different ways
    - Require the development of numerous data mining techniques
  - Interactive mining of knowledge at multiple levels of abstraction
    - Difficult to know exactly what will be discovered
    - Allow users to focus the search, refine data mining requests.
    - Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
  - Incorporation of background knowledge

Information regarding the domain under study, may be used to guide the discovery process.

- Allow discovered patterns to be expressed in concise terms and different levels of abstraction.

1.9 Major Issues in Data Mining

- Data mining query languages and ad hoc data mining
  - High-level query languages need to be developed
  - Should be integrated with a DB/DW query language
- Presentation and visualization of results
  - Knowledge should be easily understood and directly usable.
  - High level languages, visual representations or other expressive forms.
  - Require the DM system to adopt the above techniques.
- Handling noisy or incomplete data
  - Require data cleaning methods and data analysis methods that can handle noise.
- Pattern evaluation – the interestingness problem
- There is a huge amount of **data** available in the Information Industry.Extraction of information is not the only process we need to perform; **data mining** also involves other processes such as **Data** Cleaning,**Data** Integration, **Data** Transformation, **Data Mining**,**Pattern Evaluation** and **Data** Presentation.
  - How to develop techniques to access the interestingness of discovered patterns, especially with subjective measures bases on user beliefs or expectations.

1.9 Major Issues in Data Mining

- Performance Issues
  - Efficiency and scalability.
  - To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.
    - Running time must be predictable and acceptable
  - Parallel, distributed and incremental mining algorithms
    - Divide the data into partitions and processed in parallel
    - Incorporate database updates without having to mine the entire data again from scratch.
- Issues relating to Diversity of Database Types
  - Other database that contain complex data objects, multimedia data,

    spatial data, etc.

  Mining information from heterogeneous databases and global information systems:
- Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases.
- The discoveryof knowledge from different sources of structured, semistructured, or unstructured data with diverse data semantics poses great challenges to data mining.

Why Data Preprocessing?

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogenous sources.

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? How can the data be preprocessed so as to improve the efficiency and ease of the mining process?

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility

Major Tasks in Data Preprocessing

- Data cleaning

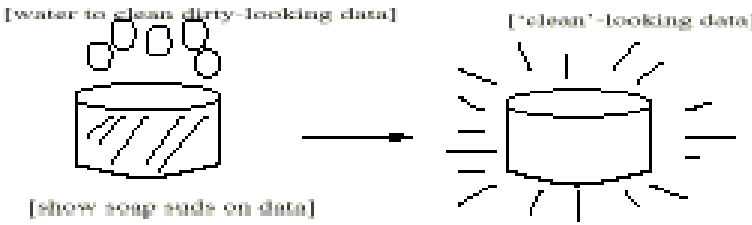
-Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

- Data integration
  - Integration of multiple databases or files
  - Having a large amount of redundant data may slow down or confuse the knowledge discovery process. Clearly, in addition to data cleaning, steps must be taken to help avoid redundancies during data integration.
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization

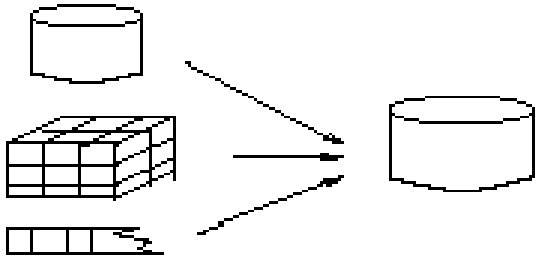
Part of data reduction but with particular importance, especially for numerical data.(**e.g., removing irrelevant attributes through correlation analysis)**

Forms of Data Preprocessing

Data Cleaning



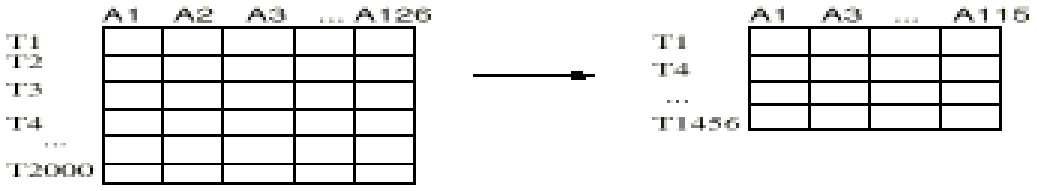
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Descriptive Data Summarization
Measuring the Central Tendency

- For many data preprocessing tasks, users would like to learn about data characteristics regarding both central tendency and dispersion of the data. Measures of central tendency include *mean, median, mode*, and *midrange*, while measures of data dispersion include *quartiles, interquartile range (IQR)*, and *variance.*
- The most common and most effective numerical measure of the "center" of a set of data is the *(arithmetic) mean*. Let *x*1;*x*2; : : : ;*xN* be a set of *N* values.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

- A **distributive measure** is a measure (i.e., function) that can be computed for a given data set by partitioning the data into smaller subsets, computing the measure for each subset, and then merging the results.eg: sum(),count().
- An **algebraic measure** is a measure that can be computed by applying an algebraic function to one or more distributive measures.eg: avg(),mean().

- A **holistic measure** is a measure that must be computed on the entire data set as a whole.
- Another Central Tendency Measure is **Median**

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width,$$

- Another Central Tendency Measure is **Mode**
- The Mode is the number which appears most often .

Example 1 ▶ Comparing Measures of Central Tendency

On an interview for a job, the interviewer tells you that the average annual income of the company's 25 employees is $60,849. The actual annual incomes of the 25 employees are shown below. What are the mean, median, and mode of the incomes? Was the person telling you the truth?

$17,305,	$478,320,	$45,678,	$18,980,	$17,408,
$25,676,	$28,906,	$12,500,	$24,540,	$33,450,
$12,500,	$33,855,	$37,450,	$20,432,	$28,956,
$34,983,	$36,540,	$250,921,	$36,853,	$16,430,
$32,654,	$98,213,	$48,980,	$94,024,	$35,671

Solution

The mean of the incomes is

$$\text{Mean} = \frac{17{,}305 + 478{,}320 + 45{,}678 + 18{,}980 + \cdots + 35{,}671}{25}$$

$$= \frac{1{,}521{,}225}{25} = \$60{,}849.$$

Car Rental A car rental company kept the following record of the numbers of miles a rental car was driven. What are the mean, median, and mode of this data?

Monday	410	Tuesday	260
Wednesday	320	Thursday	320
Friday	460	Saturday	150

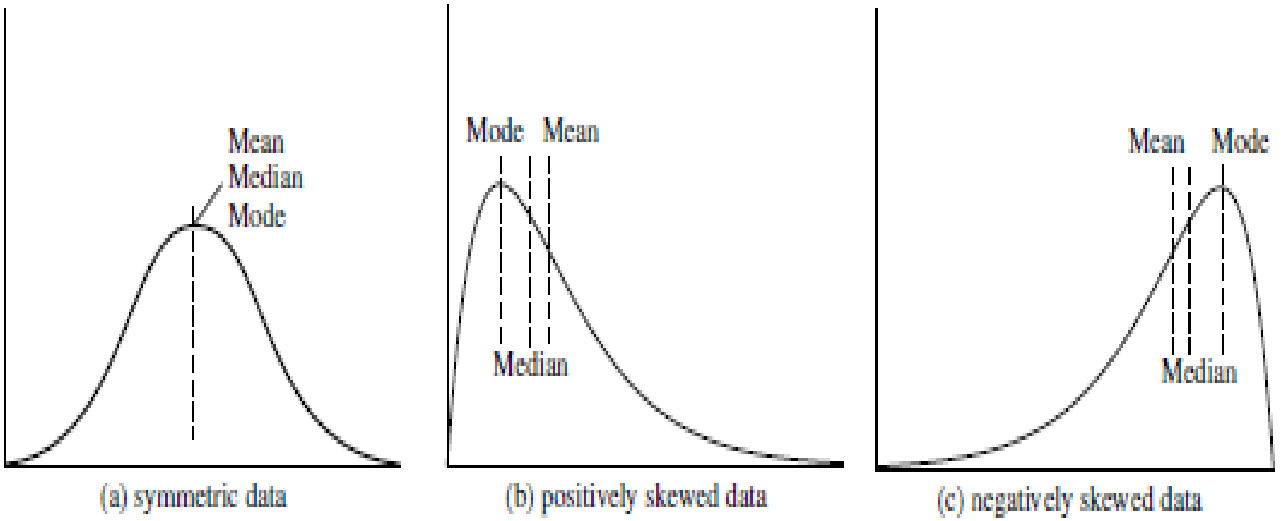


Figure 2.2 Mean, median, and mode of symmetric versus positively and negatively skewed data.

Range, Quartiles, Outliers and Boxplots

- The **range** of the set is the difference between the largest (max()) and smallest (min()) values.
- The ***k*th percentile** of a set of data in numerical order is the value *xi* having the property that *k* percent of the data entries lie at or below *xi*.
- The most commonly used percentiles other than the median are **quartiles.**
- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data.
- This distance is called the **interquartile range (*IQR*)** and is defined as

$$IQR = Q3 - Q1.$$

- A common rule of thumb for identifying suspected **outliers** is to single out values falling at least **1.5X*IQR*** above the third quartile or below the first quartile.
- five-number summary of a distribution consists of the median, the quartiles*Q*1 and*Q*3, and the smallest and largest individual observations, written in the order
- ***Minimum*; *Q*1; *Median*; *Q*3; *Maximum***
- **Boxplots** are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary
  1. Typically, the ends of the box are at the quartiles, so that the box length is the interquartile range, *IQR*.
  2. The median is marked by a line within the box.
  3. Two lines (called *whiskers*) outside the box extend to the smallest (*Minimum*) and largest (*Maximum*) observations.

Example 7 ▶ Finding Quartiles of a Set

Find the lower and upper quartiles for the set.

34, 14, 24, 16, 12, 18, 20, 24, 16, 26, 13, 27

Solution

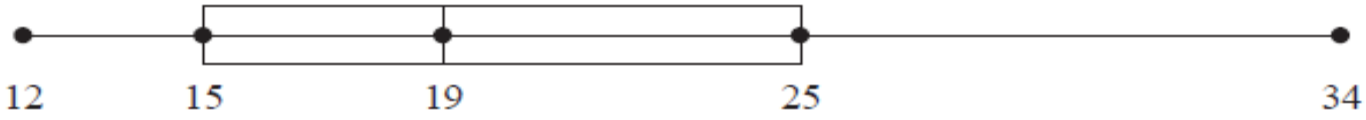
Begin by ordering the set.

12, 13, 14, 16, 16, 18, 20, 24, 24, 26, 27, 34

1st 25% 2nd 25% 3rd 25% 4th 25%

The median of the entire set is 19. The median of the six numbers that are less than 19 is 15. So, the lower quartile is 15. The median of the six numbers that are greater than 19 is 25. So, the upper quartile is 25.

Quartiles are represented graphically by a **box-and-whisker plot,** as shown in Figure A.6. In the plot, notice that five numbers are listed: the smallest number, the lower quartile, the median, the upper quartile, and the largest number. Also notice that the numbers are spaced proportionally, as though they were on a real number line.



Example 8 ▶ Sketching Box-and-Whisker Plots

Sketch a box-and-whisker plot for each set.

a. 27, 28, 30, 42, 45, 50, 50, 61, 62, 64, 66

b. 82, 82, 83, 85, 87, 89, 90, 94, 95, 95, 96, 98, 99

c. 11, 13, 13, 15, 17, 18, 20, 24, 24, 27

Variance and Standard Deviation

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}\left(\sum x_i\right)^2\right],$$

The **standard deviation**, σ, of the observations is the square root of the variance, σ^2.

The basic properties of the standard deviation, σ, as a measure of spread are

σ measures spread about the mean and should be used only when the mean is chosen as the measure of center.

$\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise $\sigma > 0$.

Graphic Displays

- There are many types of graphs for the display of data summaries and distributions, such as:
  - Bar charts
  - Pie charts
  - Line graphs
  - Boxplot
  - Histograms
  - Quantile plots
  - Scatter plots
  - Loess curves

Histogram Analysis

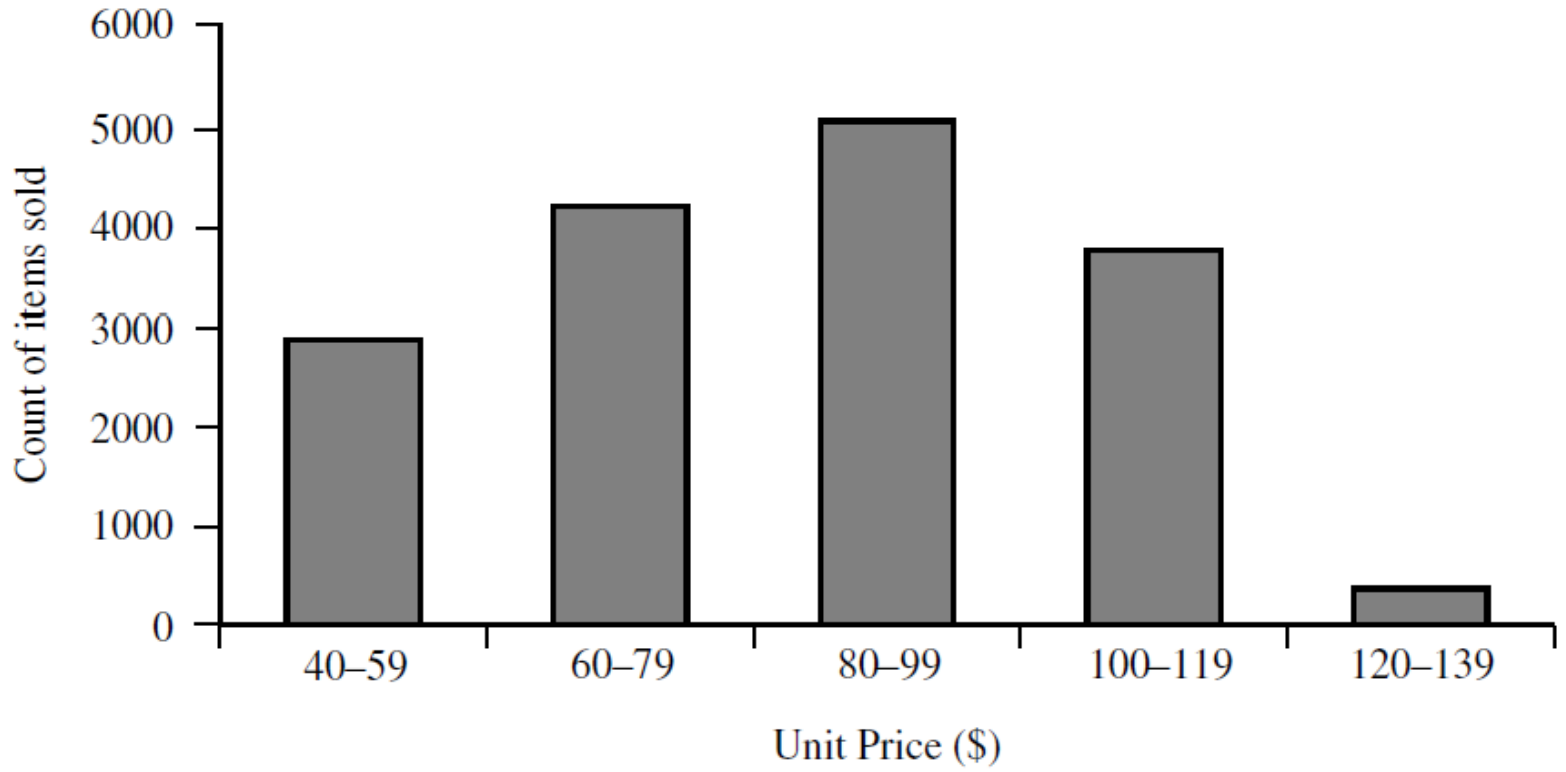
- Histograms or frequency histograms
  - A univariate graphical method
  - Consists of a set of **rectangles** that reflect the counts or frequencies of the classes present in the given data
  - If the attribute is categorical, then one rectangle is drawn for each known value of A, and the resulting graph is more commonly referred to as a bar chart.
  - If the attribute is numeric, the term histogram is preferred.

Histogram Analysis

- **Example:** A set of unit price data for items sold at a branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350

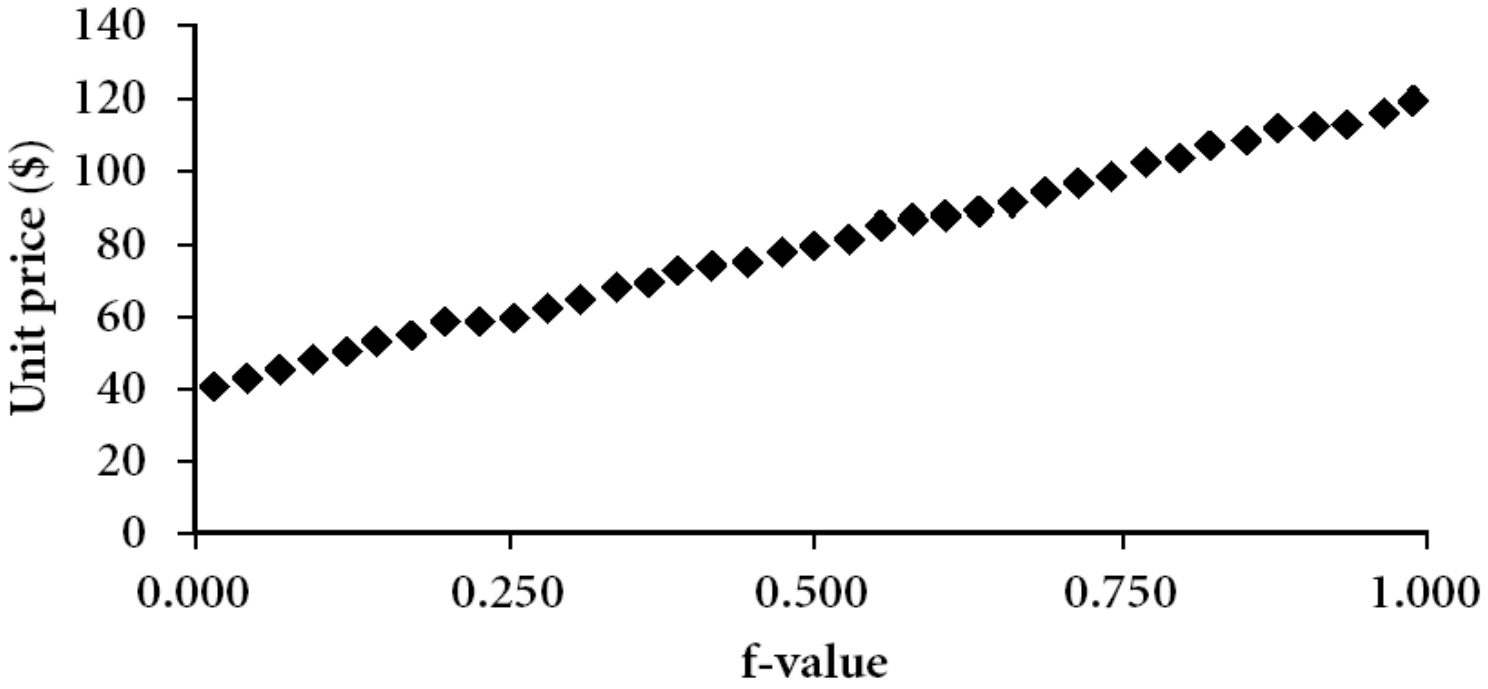
- **Example: A histogram**



Quantile Plot

- A quantile plot is a simple and effective way to have a first look at a **univariate** data distribution.
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i
- Note that
  - the 0.25 quantile corresponds to quartile Q1,
  - the 0.50 quantile is the median, and
  - the 0.75 quantile is Q3.

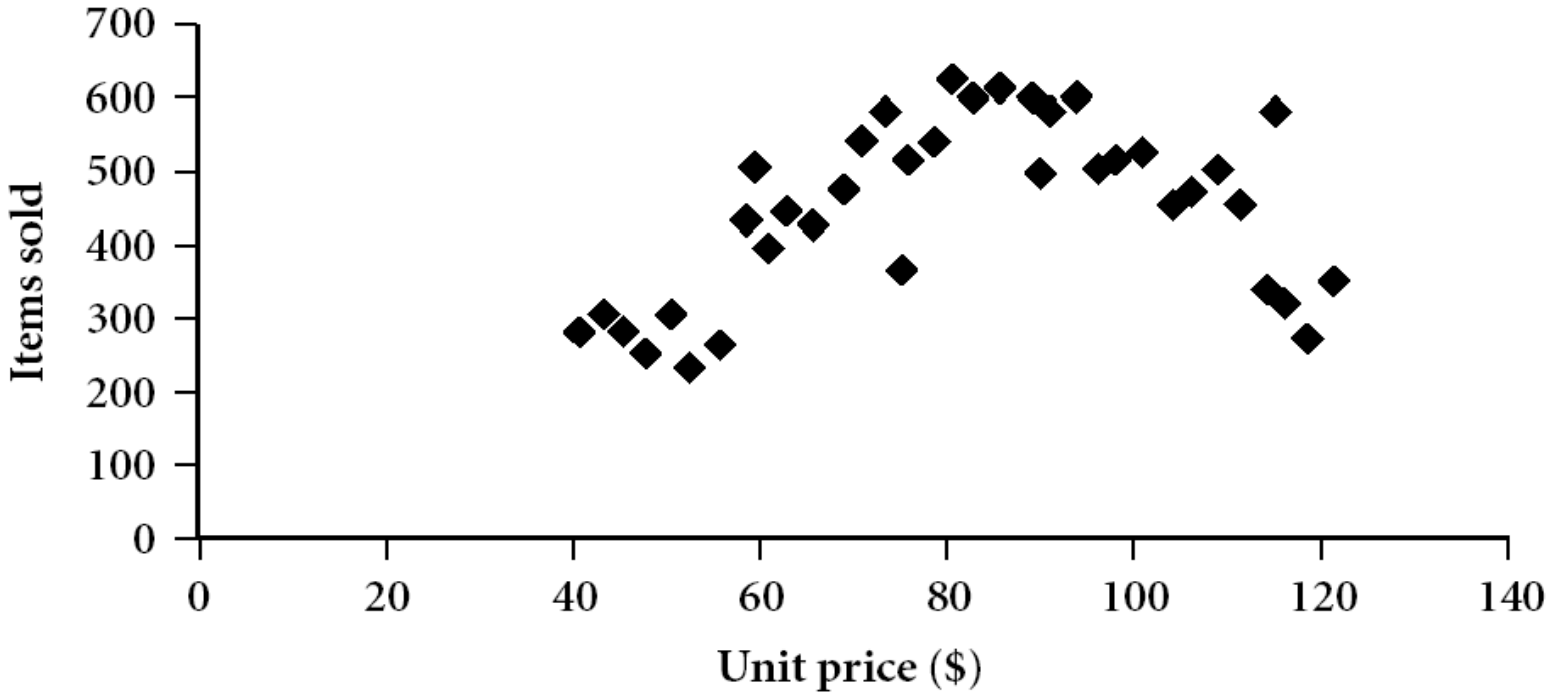
- A quantile plot for the unit price data of AllElectronics.



Scatter plot

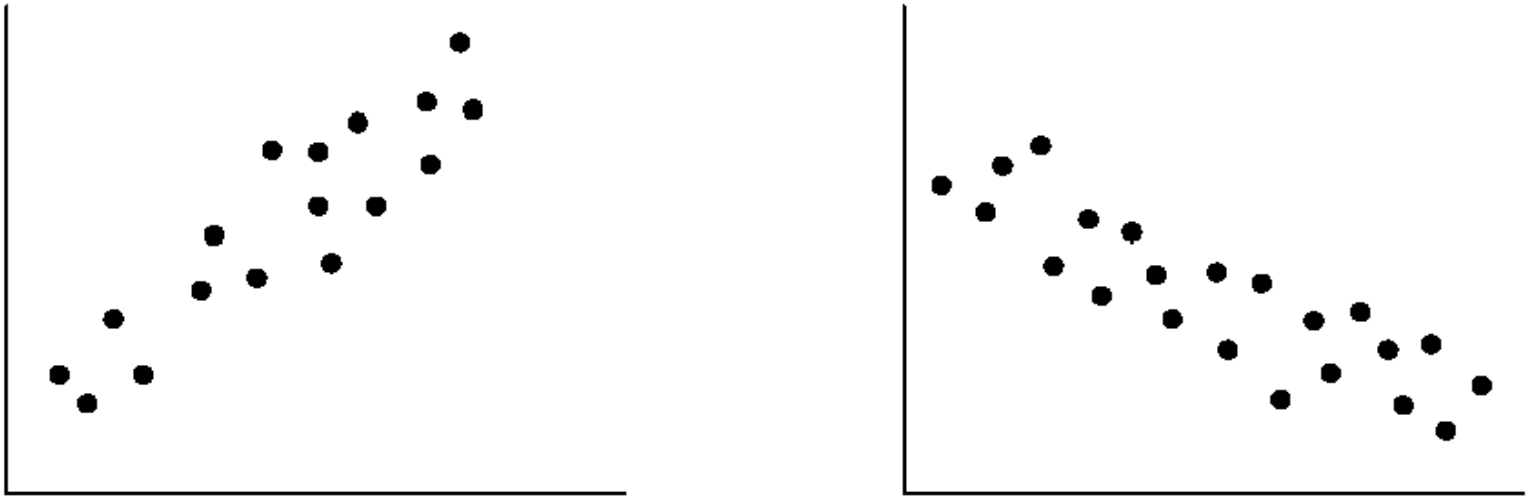
Scatter plot is one of the most effective graphical methods for determining if there appears to be a **relationship**, **clusters of points**, or **outliers** between two numerical attributes.
Each pair of values is treated as a pair of coordinates and plotted as points in the plane.

- A scatter plot for the data set of AllElectronics.



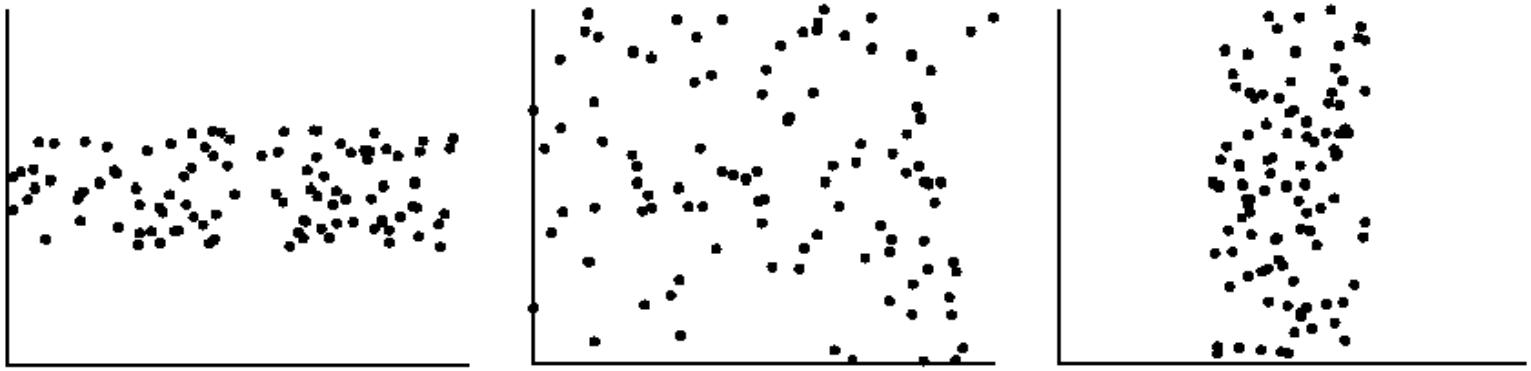
Scatter plot

- Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.



Scatter plot

- Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.



Data Cleaning

- Importance
  - garbage in garbage out principle (GIGO)

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification).
- Fill in the missing value manually
- Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like *"Unknown".*
- Use the attribute mean to fill in the missing value: For example, suppose that then average income of *All Electronics* customers is $56,000. Use this value to replace the missing value for *income*.
- Use the attribute mean for all samples belonging to the same class as the given tuple.
- Use the most probable value to fill in the missing value.

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Class label noise is hard to deal with
  - sometimes we don’t know whether the class label is correct or it is simply unexpected
- Noise demands robustness in training algorithms, that is, training should not be sensitive to noise

How to Handle Noisy Data?

- Binning
- Statistical **data binning** is a way to group a number of more or less continuous values into a smaller number of "bins". For example, if you have **data**about a group of people, you might want to arrange their ages into a smaller number of age intervals.
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression(**Regression** is a **data mining** function that predicts a number. Age, weight, distance, temperature, income, or sales could all be predicted using **regression** techniques.)
  - smoothing by fitting the data into regression functions.
  - ***Linear regression*** involves finding the “best” line to fit two attributes (or variables), so that one attribute can be used to predict the other.
  - ***Multiple linear regression*** is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.
- Clustering(**Clustering** is a process of partitioning a set of **data** (or objects) into a set of meaningful sub-classes, called **clusters)**
  - detect and remove outliers.
  - similar values are organized into groups, or “clusters.”
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of data points
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

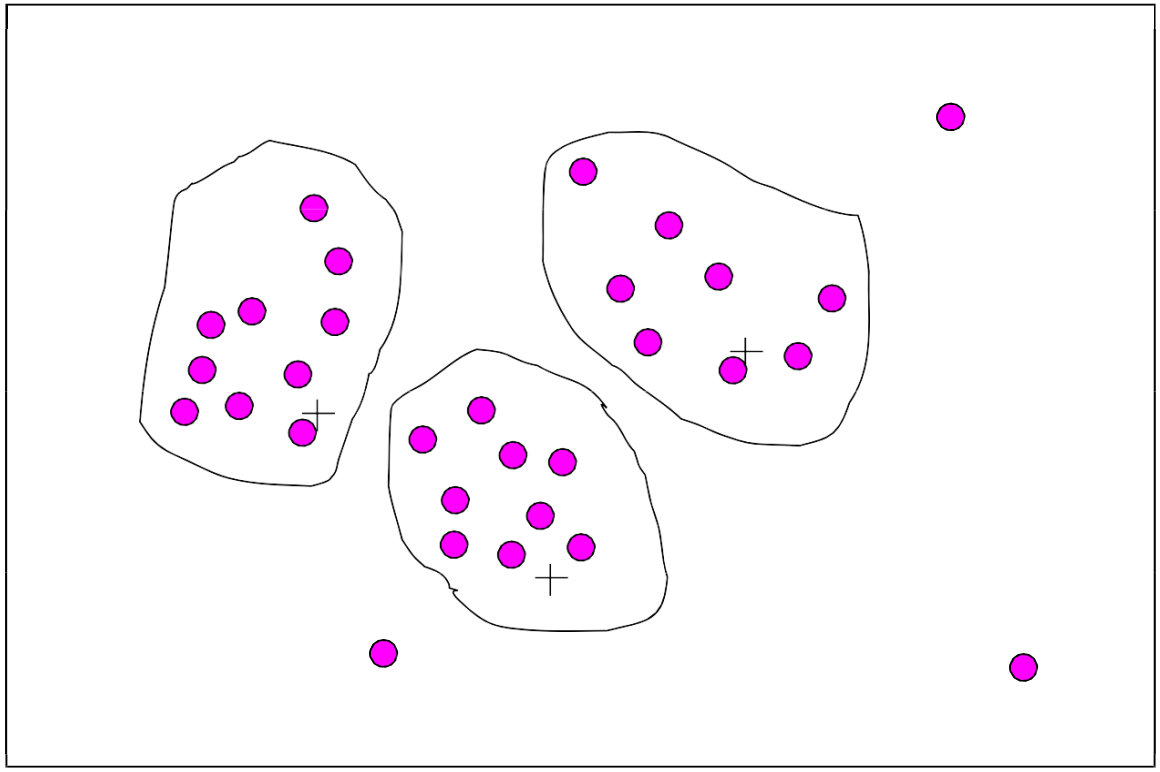
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Binning Methods for Data Smoothing

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Cluster Analysis as Binning



Data Cleaning as process

- Missing values, noise, and inconsistencies contribute to inaccurate data.
- The first step in data cleaning as a process is ***discrepancy detection*.**
- Discrepancies can be caused by several factors, including **poorly designed data entry forms** that have many optional fields, **human error in data entry**, **deliberate errors** (e.g., respondents not wanting to divulge information about themselves), and **data decay** (e.g., outdated addresses).
- A **unique rule** says that each value of the given attribute must be different from all other values for that attribute.
- A **consecutive rule** says that there can be no missing values between the lowest and highest values for the attribute, and that all values must also be unique (e.g., as in check numbers).
- A **null rule** specifies the use of blanks, question marks, special characters, or other strings that may indicate the null condition.

- **Field overloading** is another source of errors that typically results when developers squeeze new attribute definitions into unused (bit) portions of already defined attributes.

Discrepancy detection tools

- **Data scrubbing tools** use simple domain knowledge (e.g., knowledge of postal addresses, and spell-checking) to detect errors and make corrections in the data.
- These tools rely on **parsing and fuzzy matching techniques** when cleaning data from multiple sources.
- **Data auditing tools** find discrepancies by analyzing the data to discover rules and relationships, and detecting data that violate such conditions.

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Integration

- For **numerical attributes**, we can evaluate the correlation between two attributes, *A* and *B*, by computing the correlation coefficient.

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B},$$

- If rA,B > 0, A and B are positively correlated (A's values
- increase as B's). The higher, the stronger correlation.
- rA,B = 0: uncorrelated;
- rA,B < 0: negatively correlated.

Chi-square test

- For **categorical (discrete) data**, a correlation relationship between two attributes, *A* and *B*, can be discovered by a c2 (chi-square) test.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

- where *oi j* is the *observed frequency* (i.e., actual count) of the joint event (*Ai*;*Bj*) and *ei j* is the *expected frequency* of (*Ai*;*Bj*), which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N},$$

Chi-Square Calculation: An Example

	Male	Female	Total (row)
Like science fiction	250	200	450
Not like science fiction	50	1000	1050
Total (col.)	300	1200	1500

$$e_{11} = \frac{count(male) \times count(fiction)}{N} = \frac{300 \times 450}{1500} = 90,$$

Chi-Square Calculation: An Example

	male	female	Total (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Total (col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and male are correlated in the group

Data Transformation

- **Smoothing:** which works to remove noise from the data. Such techniques include binning, regression, and clustering.

- **Aggregation**: where summary or aggregation operations are applied to the data.
- **Eg:** the daily sales data may be aggregated so as to compute monthly and annual total amounts.

- **Generalization** of the data, where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
- Eg: **categorical attributes**, like *street*, can be generalized to higher-level concepts, like *city* or *country*.
- **Numerical attributes**, like *age*, may be mapped to higher-level concepts, like *youth, middle-aged*, and *senior*.

- **Normalization:** where the attribute data are scaled so as to fall within a small specified range, such as - 1.0 to 1.0, or 0.0 to 1.0.

- **Attribute construction** (or *feature construction*),where new attributes are constructed and added from the given set of attributes to help the mining process.

- There are many methods for data normalization.
- ***min-max normalization***
- ***z-score normalization***
- ***normalization by decimal scaling***.

Min-max normalization

- performs a linear transformation on the original data.
- Min-max normalization preserves the relationships among the original data values.
- It will encounter an “out-of-bounds” error if a future input case for normalization falls outside of the original data range for *A*.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

- Suppose that the minimum and maximum values for the attribute *income* are $12,000 and $98,000, respectively. We would like to map *income* to the range [0.0;1.0]. By min-max normalization, a value of $73,600 for *income* is transformed

$$73600 - 12000/98000 - 12000\ (1.0 - 0.0) + 0$$
$$= 0.716$$

z-score normalization (or *zero-mean normalization*)

The values for an attribute, *A*, are normalized based on the mean and standard deviation of *A*.

where $\bar{A}$ and σA a ard deviation, respectively, of a

$$v' = \frac{v - \bar{A}}{\sigma_A},$$

- Suppose that the mean and standard deviation of the values for the attribute *income* are $54,000 and $16,000, respectively. With z-score normalization, a value of $73,600 for *income* is transformed

$$73600 - 54000 / 16000$$

$$= 1.225$$

Normalization by decimal scaling

- Normalizes by moving the decimal point of values of attribute *A*. The number of decimal points moved depends on the maximum absolute value of *A*.

$$v' = \frac{v}{10^j},$$

where j is the smallest integer such that $Max(|v'|) < 1$.

- Suppose that the recorded values of *A* range from - 986 to 917. The maximum absolute value of *A* is 986.

- To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., *j* = 3) so that - 986 normalizes to – 0.986 and 917 normalizes to 0.917.

Data Reduction

- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume.
- closely maintains the integrity of the original data.

Strategies for data reduction

1. Data cube aggregation: where aggregation operations are applied to the data in the construction of a data cube.

2. Attribute subset selection: where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

3. Dimensionality reduction: where encoding mechanisms are used to reduce the data set size.

4. Numerosity reduction: where the data are replaced or estimated by alternative.

5. Discretization and concept hierarchy generation: where raw data values for attributes are replaced by ranges or higher conceptual levels.

- http://www.yourarticlelibrary.com/education/statistics/central-tendency-meaning-uses-and-measures/64944/