Data Mining
Unit I

B.Tech (CSE) VI Semester
2020-21

Vision and Mission of the Department

Vision

To evolve as a centre of **academic and research excellence** in the area of Computer Science and Engineering

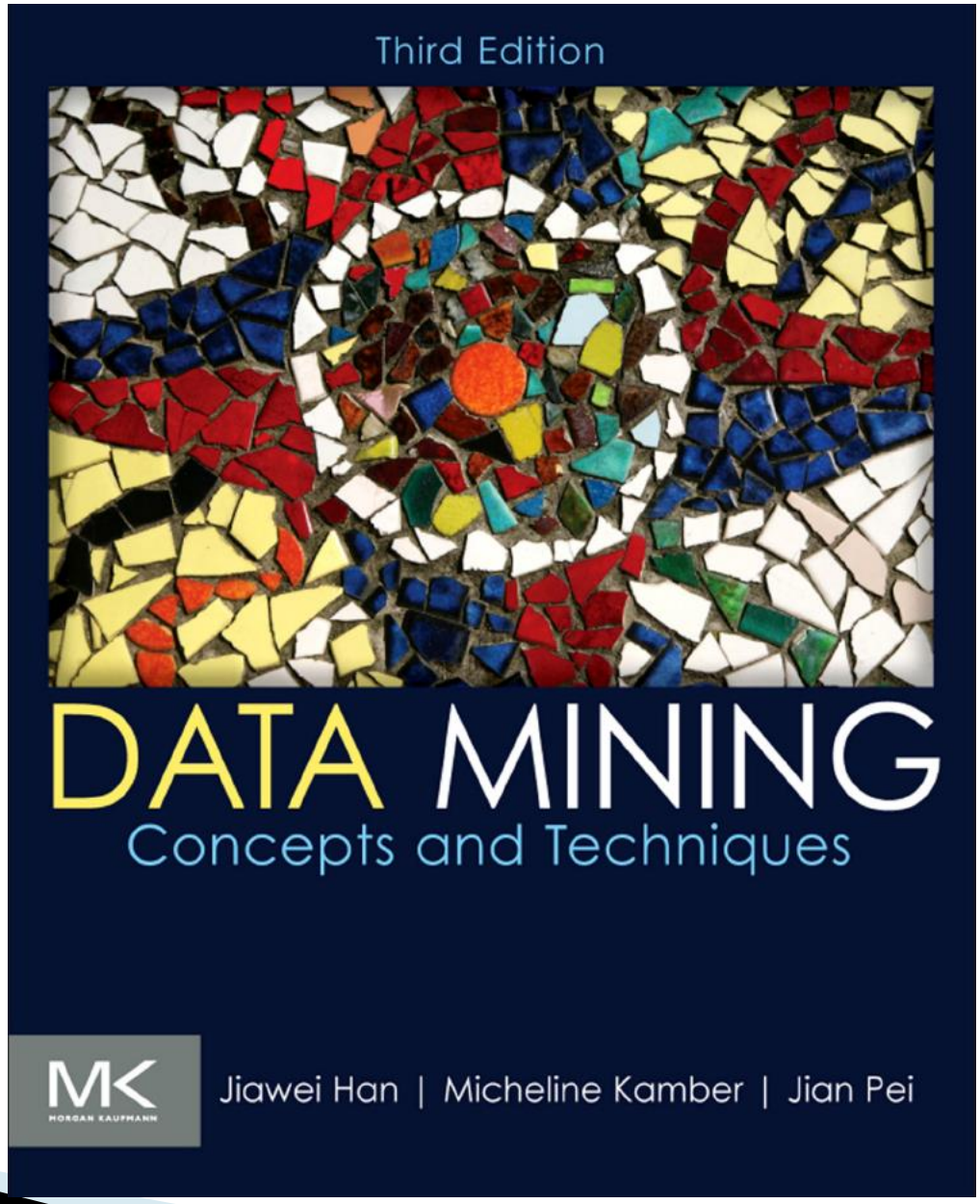
Mission

The Computer Science & Engineering Department is committed to

1. Utilize **innovative learning methods** for academic improvement

2. Encourage **higher studies and research** to meet the futuristic requirements of Computer Science and Engineering

3. Inculcate **ethics and human values** for developing students with good character


Third Edition
DATA MINING
Concepts and Techniques
MK
MORGAN KAUFMANN
Jiawei Han | Micheline Kamber | Jian Pei

Course Outcomes

After Successful completion of the Course, the student will be able to:

CO1: Explain the concept of Data Mining and its functionalities (K2)

CO2: Discuss various Data Preprocessing Techniques (K2)

CO3: Demonstrate Association Analysis Techniques (K3)

CO4: Illustrate various Classification Techniques (K3)

CO5: Demonstrate Alternative techniques for Classification (K3)

CO6: Use different Clustering techniques to cluster data (K3)

CO1: Topics to be covered

- Introduction: Need for data mining
- Knowledge Discovery from Data: Data Mining
- Kinds of Data Mined
- Kinds of Patterns Mined: Data mining functionalities
- Technologies used
- Kinds of Applications targeted
- Major issues in data mining
- Data Objects
- Attribute types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity

Introduction: Need for data mining

- The Explosive Growth of Data: from terabytes to petabytes
  - 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

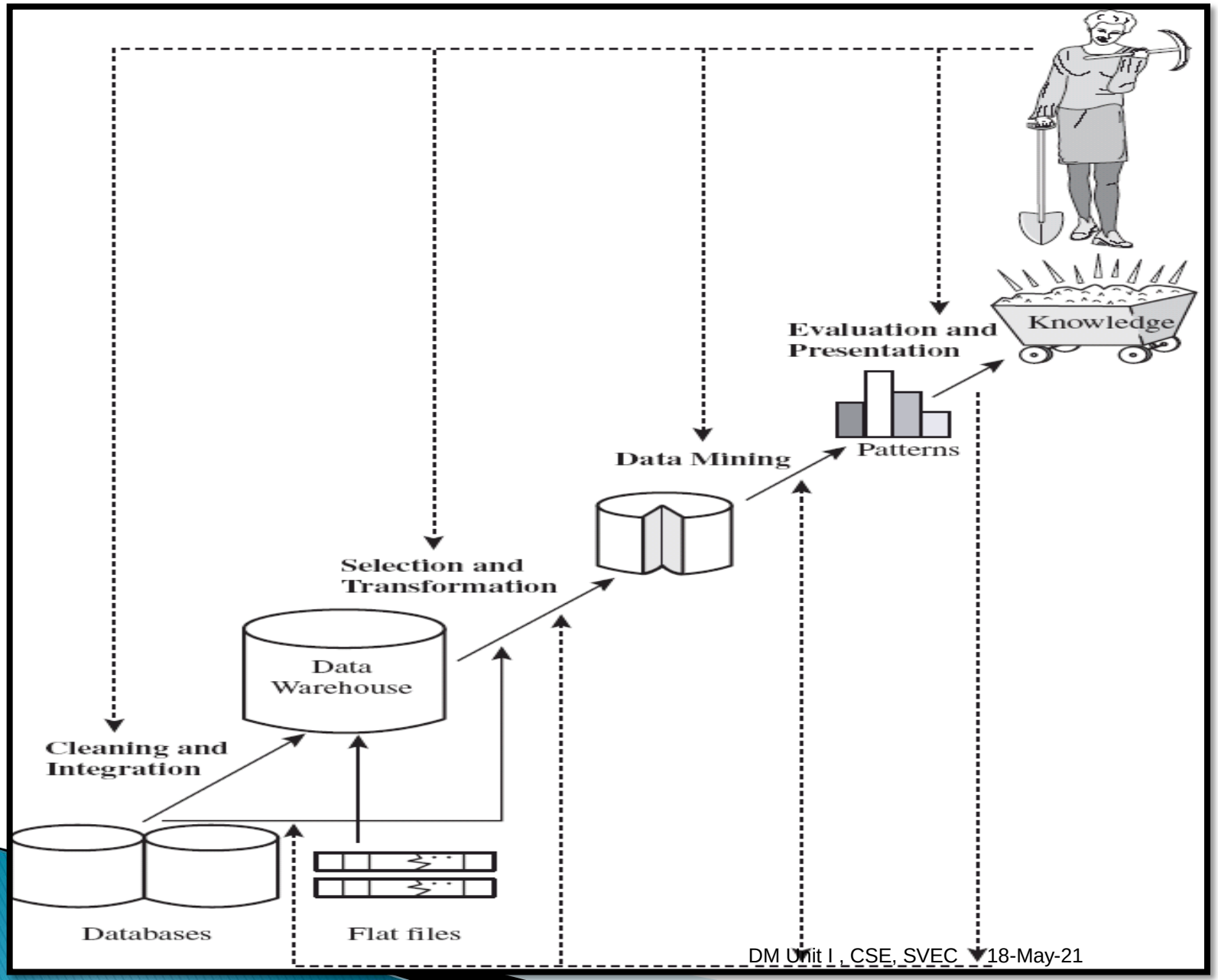
Knowledge Discovery from Data: Data Mining

- 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
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Knowledge Discovery from Data: Data Mining[contd.]

1. **Data cleaning** - remove noise and inconsistent data
2. **Data integration** - multiple data sources may be combined
3. **Data selection** - data relevant to the analysis task are retrieved from database
4. **Data transformation** - data transformed or consolidated into forms appropriate for mining
5. **Data mining** - an essential process where intelligent methods are applied to extract data patterns
6. **Pattern evaluation** -identify the truly interesting patterns
7. **Knowledge presentation** - mined knowledge is presented to the user with visualization or representation techniques

Knowledge Discovery from Data: Data Mining [contd.]



Kinds of Data Mined

1. Database Data

- DBMS – database management system, contains a collection of interrelated databases
  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 columns and rows, with columns as attributes of data and rows as records.
- Tables can be used to represent the relationships between or among multiple tables.

1. Database Data : *A Relational Database for 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

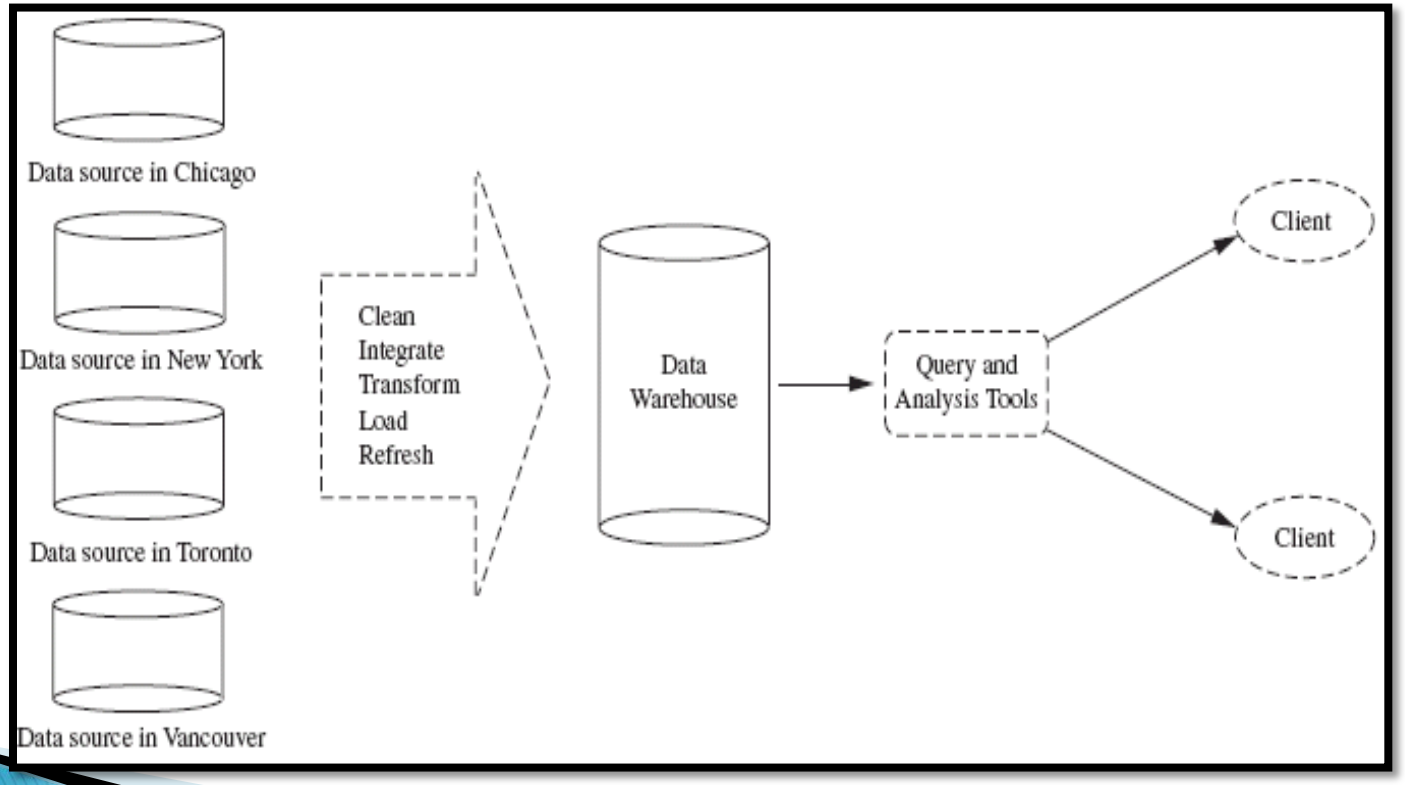
empl_ID	branch_ID
E55	B1
...	...

1. Database Data [contd.]

- 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.

2. 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.

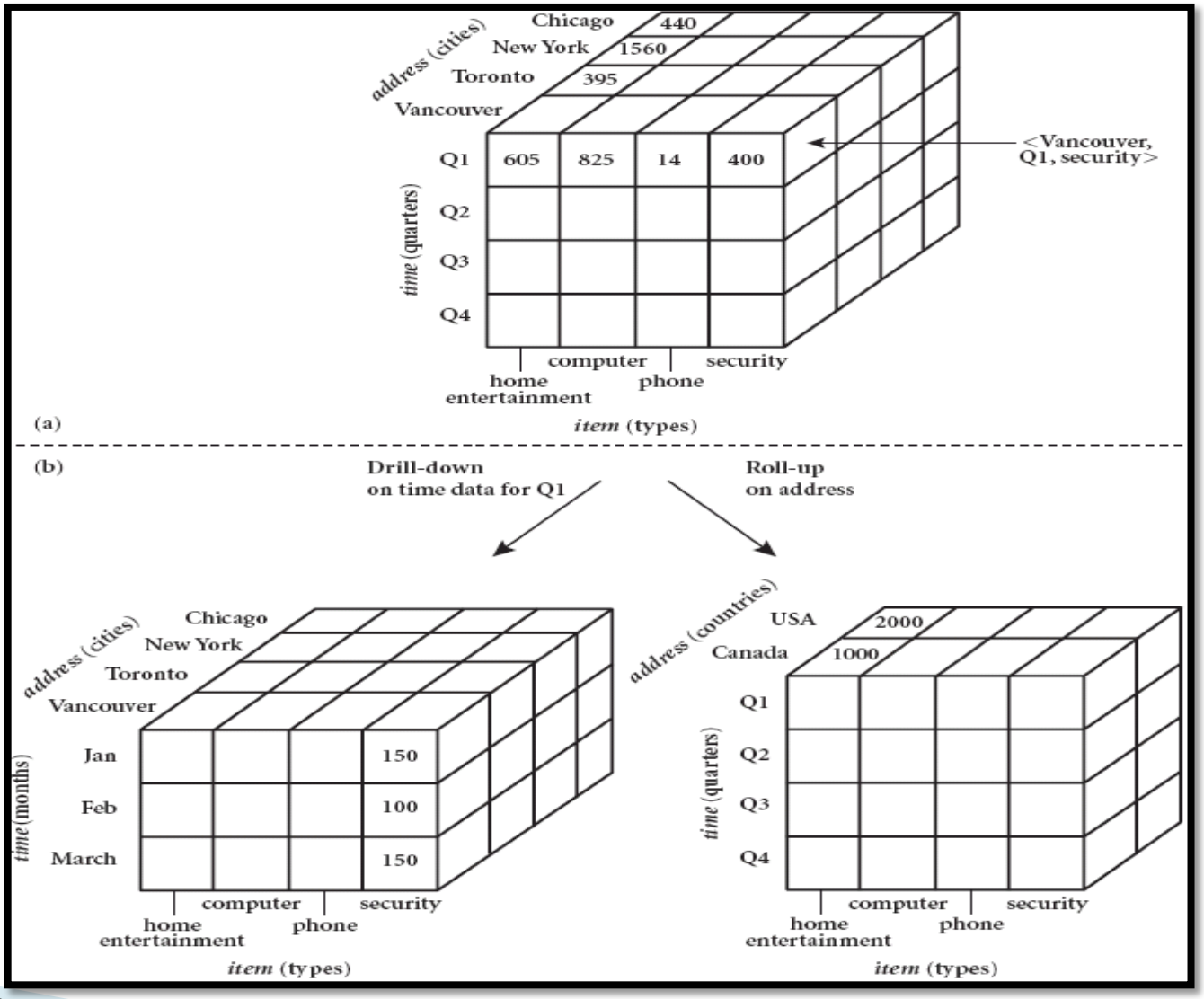


2. Data Warehouses [contd.]

- 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

2. Data Warehouses [contd.]



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
- Easy to identify items that are frequently sold together

Others kinds of Data

- **Time-related (temporal) or sequence data :** *eg. Historical records, stock exchange data, time-series and biological sequence data*
- **Data Streams :** *eg. video surveillance and sensor data*
- **Spatial Data :** *eg. maps*
- **Engineering design data :** *eg. the design of buildings, system components, or integrated circuits*
- **Hypertext and multimedia data :** *eg. including text, image, video, and audio data*
- **Graph and networked data :** *eg. social and information networks*
- **The World-Wide Web :** *eg. characterize and classify web pages and uncover web dynamics*

Kinds of Patterns Mined: Data mining functionalities

1. 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.

Kinds of Patterns Mined: Data mining functionalities[contd.]

- **Data discrimination** – comparing the target class with one or a set of comparative classes
  - E.g. Compare the general features of software products whose sales increase by 10% in the last year with those whose sales decrease by 30% during the same period
- **Or both of the above**

2. 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)
- **Frequent subsequence**: a pattern that customers tend to purchase product A, followed by a purchase of product B

Kinds of Patterns Mined: Data mining functionalities[contd.]

- **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.

- **Correlation Analysis**: additional analysis to find statistical correlations between associated pairs

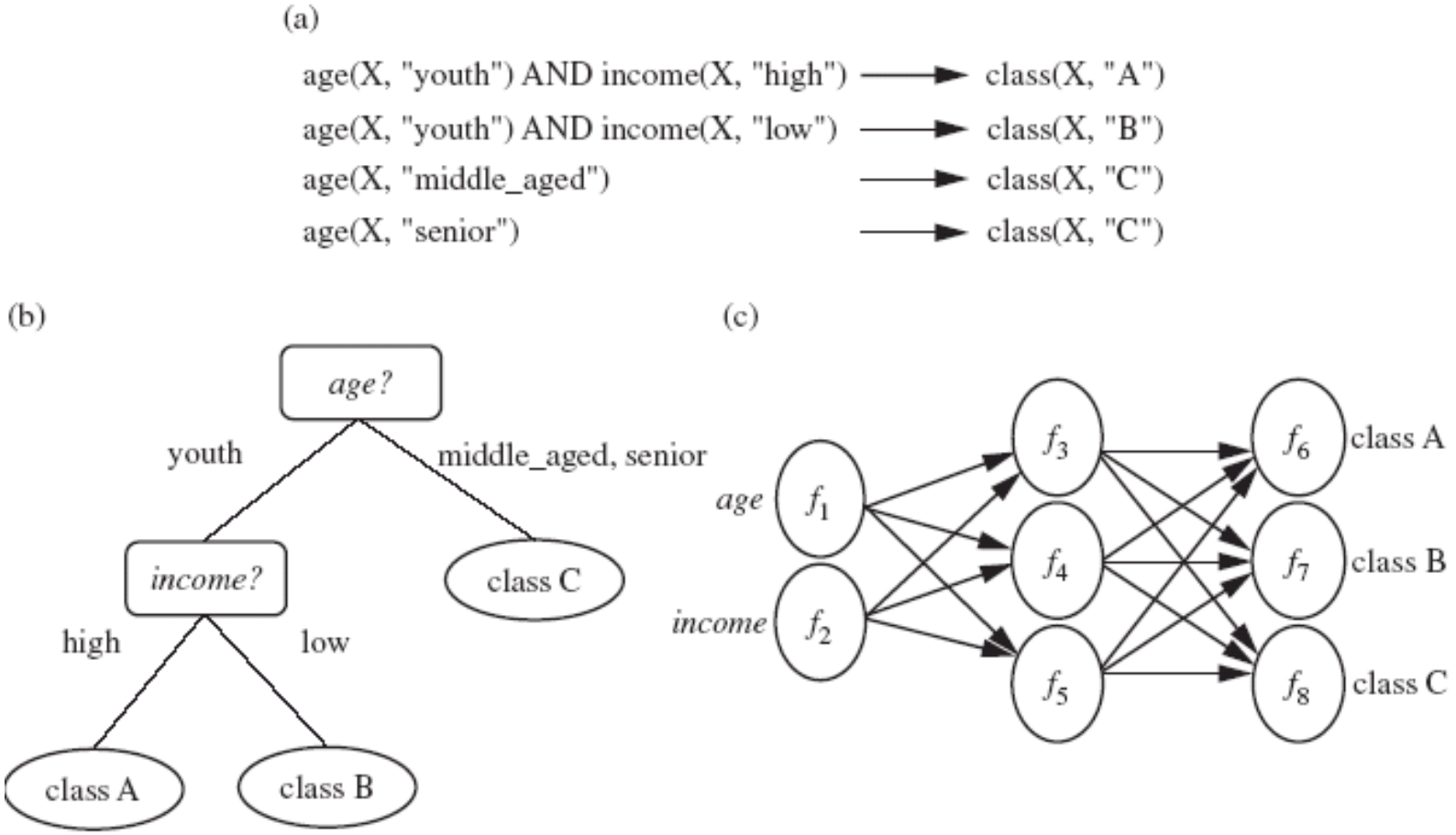
Kinds of Patterns Mined: Data mining functionalities[contd.]

3. Classification and Regression

- **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.

Kinds of Patterns Mined: Data mining functionalities[contd.]

3. Classification and Regression



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

Kinds of Patterns Mined: Data mining functionalities[contd.]

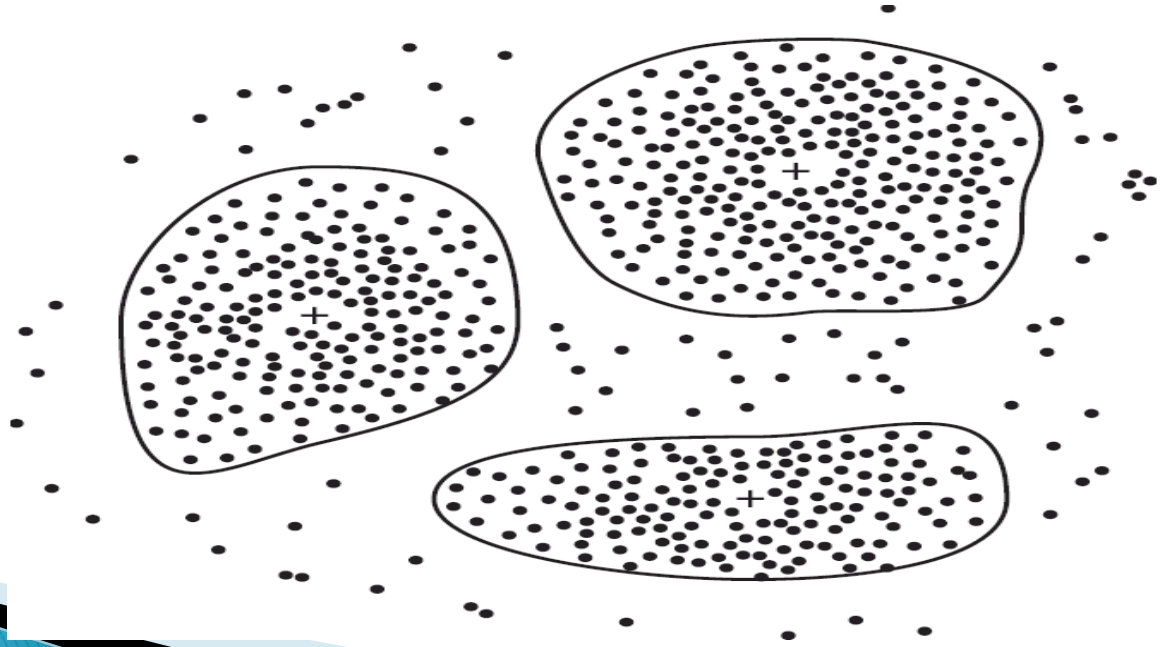
3. Classification and Regression

- **Regression**
  - Predict missing or unavailable numerical data values
  - For eg. Suppose instead, that rather than predicting categorical response labels for each store item, you would like to predict the amount of revenue that each item will generate during an upcoming sale at *AllElectronics*, based on the previous sales data. This is an example of regression analysis because the regression model constructed will predict a continuous function (or ordered value.)

Kinds of Patterns Mined: Data mining **functionalities**[contd.]

4. Cluster Analysis

- 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.



Kinds of Patterns Mined: Data mining functionalities[contd.]

5. Outlier Analysis

- Data that do no comply with the general behavior or model.
- Outliers are usually discarded as noise or exceptions.
- Useful for fraud detection.
  - E.g. Detect purchases of extremely large amounts

Kinds of Patterns Mined: Data mining functionalities [contd.]

Are All of the Patterns Interesting?

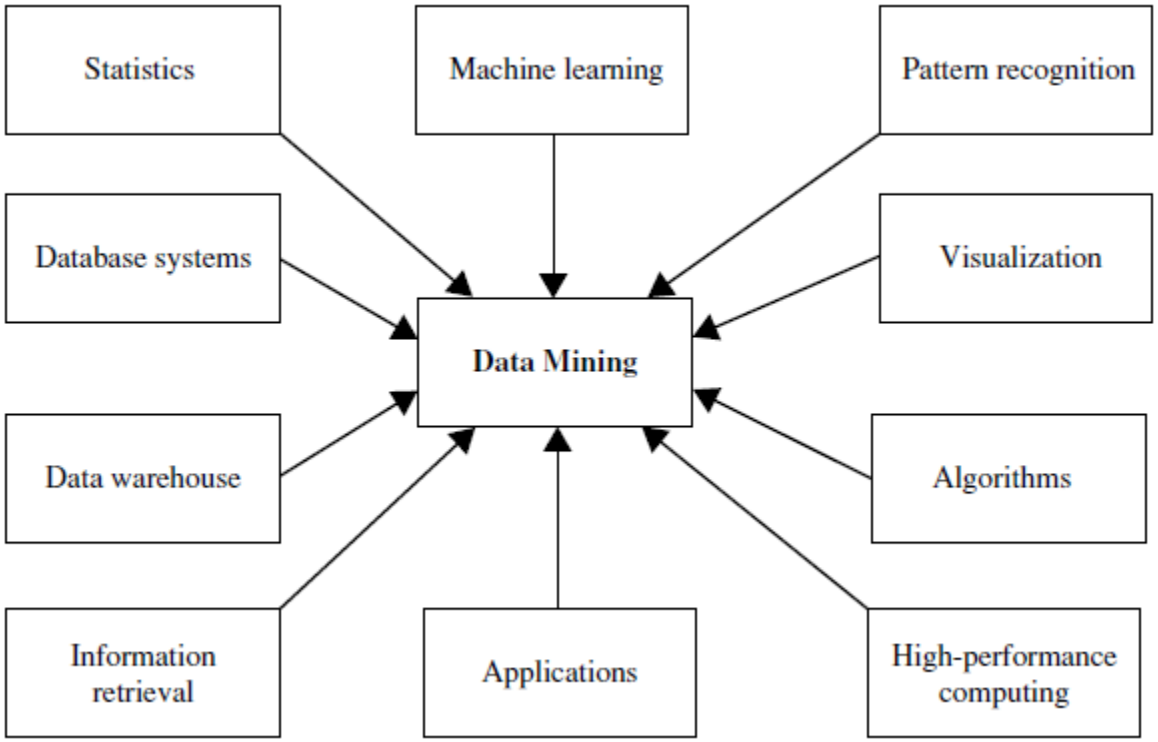
- Data mining may generate thousands of patterns: Not all of them are interesting
- A pattern is interesting if it is
  - easily understood by humans
  - valid on new or test data with some degree of certainty,
  - potentially useful
  - novel
  - validates some hypothesis that a user seeks to confirm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

Kinds of Patterns Mined: Data mining functionalities [contd.]

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.)
- 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.

Technologies used



Data mining adopts techniques from many domains

Technologies used [contd.]

1. Statistics

- **Statistics** studies the collection, analysis, interpretation or explanation, and presentation of data.
- Data mining has an inherent connection with statistics.
- A **statistical model** is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions.
- Statistical models are widely used to model data and data classes.

Technologies used [contd.]

2. Machine Learning

- Machine learning investigates how computers can learn (or improve their performance) 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
- Classic problems in machine learning that are highly related to data mining:
  - **Supervised learning** is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set

Technologies used [contd.]

2. Machine Learning

- **Unsupervised learning** is essentially a synonym for clustering
- **Semi-supervised learning** is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model
- **Active learning** is a machine learning approach that lets users play an active role in the learning process

Technologies used [contd.]

3. Database Systems and Data Warehouses

- Database systems research focuses on the creation, maintenance, and use of databases
- A data warehouse integrates data originating from multiple sources and various timeframes. It consolidates data in multidimensional space to form partially materialized data cubes. The data cube model not only facilitates Online Analytical Processing (OLAP) in multidimensional databases but also promotes multidimensional data mining

Technologies used [contd.]

4. Information Retrieval

- Information retrieval (IR) is the science of searching for documents or information in documents.
- Documents can be text or multimedia, and may reside on the Web.
- Information retrieval assumes that:

  (1) the data under search are unstructured; and

  (2) the queries are formed mainly by keywords, which do not have complex structures

Kinds of Applications targeted

1. Business Intelligence(BI)

- BI technologies provide historical, current, and predictive views of business operations. Examples include reporting, online analytical processing, business performance management, competitive intelligence, benchmarking, and predictive analytics.
- Data mining is the core of BI. Online analytical processing tools in business intelligence rely on data warehousing and multidimensional data mining.
- Classification and prediction techniques are the core of predictive analytics in BI, for which there are many applications in analyzing markets, supplies, and sales.

Kinds of Applications targeted [contd.]

2. Web Search Engines

- A **Web search engine** is a specialized computer server that searches for information on the Web.
- The search results of a user query are often returned as a list (sometimes called *hits*). The hits may consist of web pages, images, and other types of files.
- Some search engines also search and return data available in public databases or open directories.
- Web search engines are essentially very large data mining applications. Search engines pose grand challenges to data mining.

1) They have to handle a huge and ever-growing amount of data.

2) Web search engines often have to deal with online data.

3) Web search engines often have to deal with queries that are asked only a very small number of times.

Major Issues in Data Mining

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Ad hoc data mining and data mining query languages
  - Presentation and visualization of data mining results

Major Issues in Data Mining [contd.]

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attribute Types

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal Attributes :** categories, states, or “names of things”
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary Attributes**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal Attributes**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Attribute Types [contd.]

- **Numeric Attributes** - A **numeric attribute** is q*uantitative*; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be *interval-scaled* or *ratio-scaled*.
  - **Interval**
    - **No true zero-point**
    - **Interval-scaled attributes** are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values.
    - A temperature attribute is interval-scaled. Suppose that we have the outdoor temperature value for a number of different days, where each day is an object. By ordering the values, we obtain a ranking of the objects with respect to temperature. In addition, we can quantify the difference between values.
    - For example, a temperature of 20ºC is five degrees higher than a temperature of 15ºC. Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart. Temperatures in Celsius and Fahrenheit do not have a true zero-point, that is, neither 0ºC nor 0ºF indicates “no temperature.”
    - Although we can compute the *difference* between temperature values, we cannot talk of one temperature value as being a *multiple* of another. Without a true zero, we cannot say, for instance, that 10ºC is twice as warm as 5ºC. That is, we cannot speak of the values in terms of ratios. Similarly, there is no true zero-point for calendar dates.

Attribute Types [contd.]

- **Ratio**
  - **Inherent zero-point**
  - A **ratio-scaled attribute** is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value. In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.
  - Unlike temperatures in Celsius and Fahrenheit, the Kelvin (K) temperature scale has what is considered a true zero-point (0ºK = - 273.15ºC)
  - It is the point at which the particles that comprise matter have zero kinetic energy.
  - Other examples of ratio-scaled attributes include count attributes such as years of experience (e.g., the objects are employees) and number of words (e.g., the objects are documents)

Attribute Types [contd.]

- **Discrete versus Continuous Attributes**
- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Basic Statistical Descriptions of Data

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  Where *N* is number of observations

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Trimmed mean: chopping extreme values

Basic Statistical Descriptions of Data

Measuring the Central Tendency

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

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

  - where L_1 is lower boundary of median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of frequencies of all the intervals that are lower than the median interval, $freq_{median}$ is the frequency of median interval and width is the width of median interval

Basic Statistical Descriptions of Data [contd.]

Measuring the Central Tendency

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

$$mean - mode = 3 \times (mean - median)$$

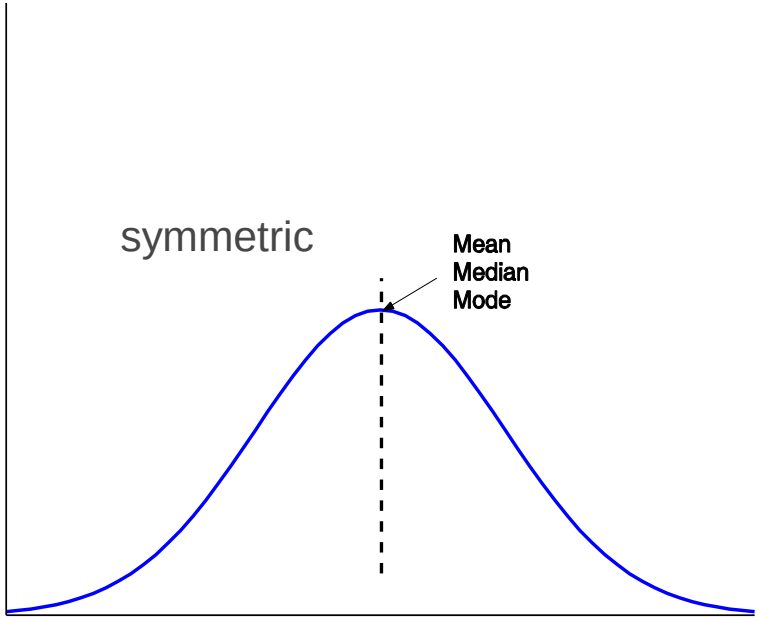
- Midrange:

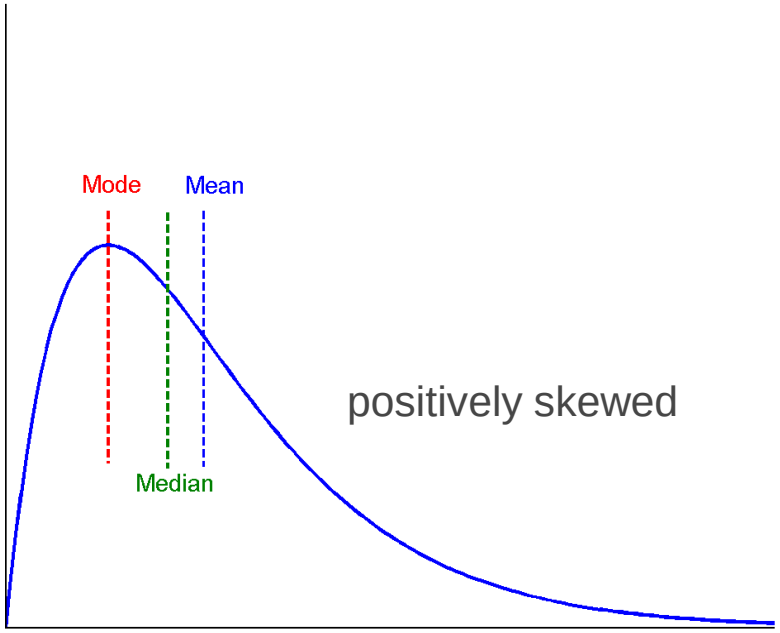
$$Midrange = \frac{\min + \max}{2}$$

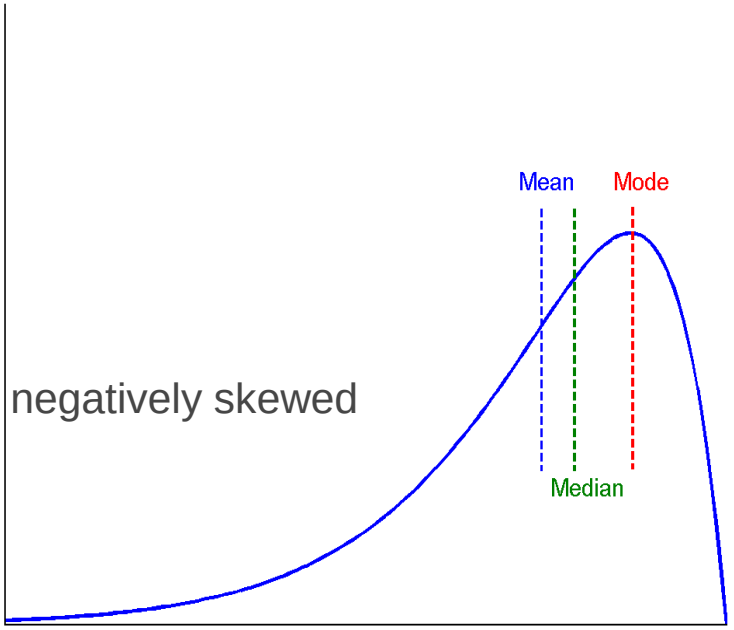
Basic Statistical Descriptions of Data [contd.]

Measuring the Central Tendency

- **Symmetric vs. Skewed Data**
  - Median, mean and mode of symmetric, positively and negatively skewed data







Basic Statistical Descriptions of Data [contd.]

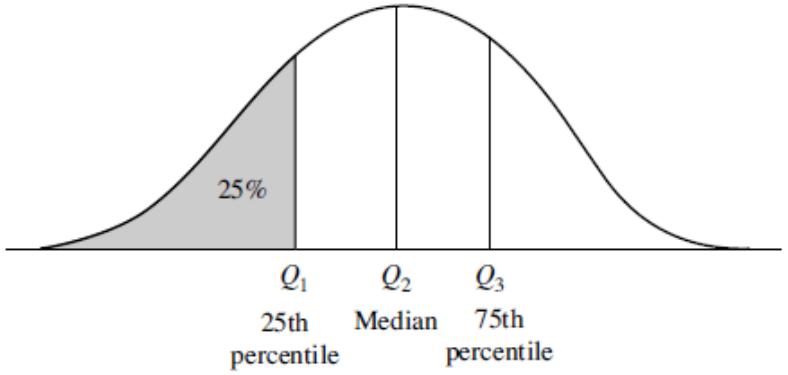
Measuring the Dispersion of Data

- **Range, Quartiles and Interquartile Range**
  - Let $x_1, x_2, : : : , x_N$ be a set of observations for some numeric attribute, X.
  - **Range**: The range of the set is the difference between the largest (max()) and smallest (min()) values.
  - **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.
  - **Quartiles**: The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as quartiles.
  - For eg. Q_1 (25^{th} percentile), Q_2 (50^{th} percentile) Q_3 (75^{th} percentile)
  - The 100-quantiles are more commonly referred to as **percentiles**
  - **Inter-quartile range**: $IQR = Q_3 - Q_1$

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Range, Quartiles and Interquartile Range**

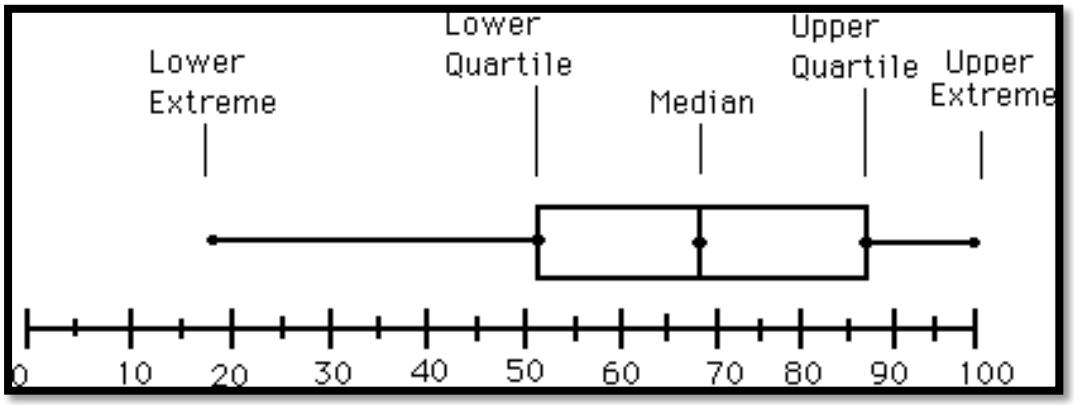


A plot of the data distribution for some attribute X.

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Five number summary, Boxplot and Outliers**
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: points beyond a specified outlier threshold, plotted individually usually, a value higher/lower than 1.5 x IQR

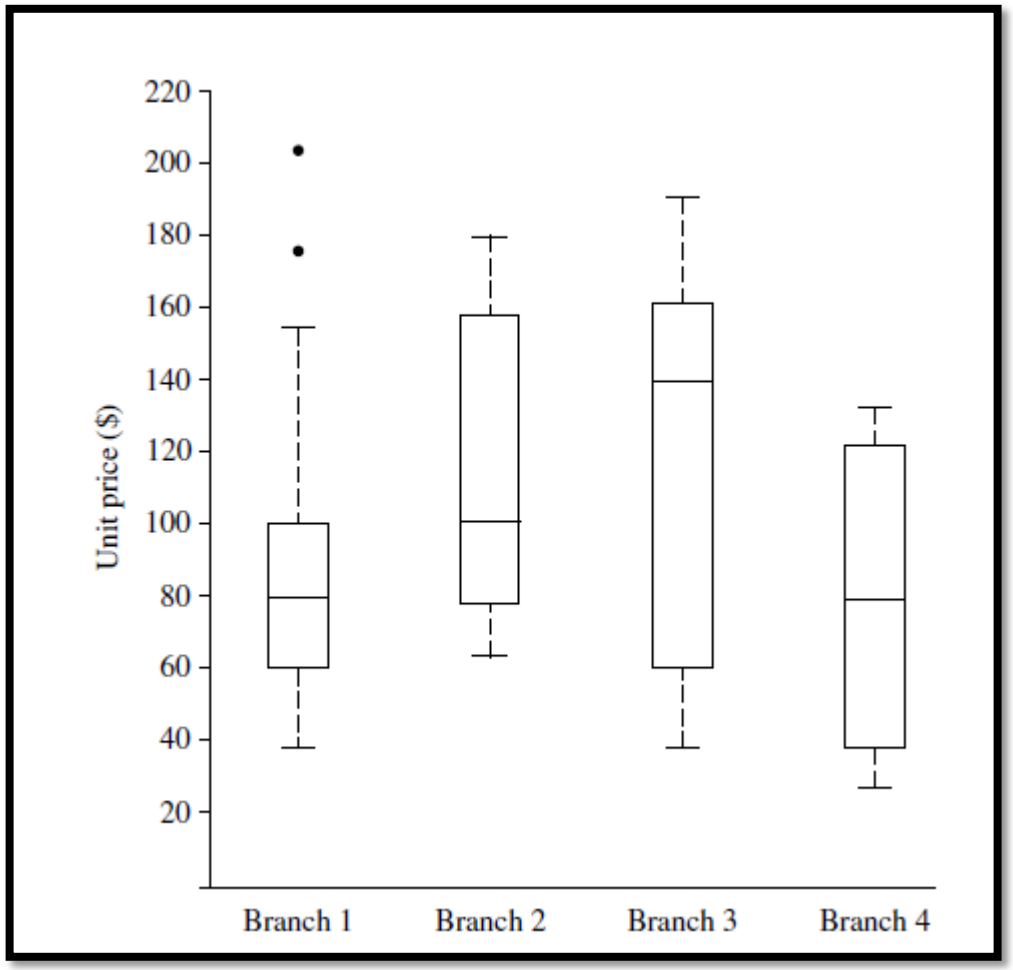


Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Boxplot example**

Boxplot for the unit price data for items sold at four branches of AllElectronics during a given time period



Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Variance and standard deviation**
- Variance and standard deviation indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.
  - **Variance**: (algebraic, scalable computation)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{n}x_i^2\right) - \mu^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Basic Statistical Descriptions of Data [contd.]

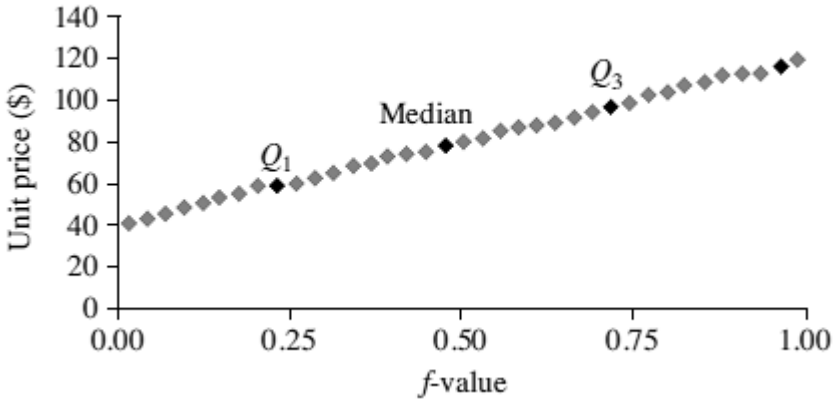
Graphic Displays of Basic Statistical Descriptions

- **Quantile plot**: It is a simple and effective way to have a first look at a univariate data distribution.
- It displays all of the data for the given attribute

$$f_i = \frac{i - 0.5}{N}.$$

- It plots quantile information
- Each observation x_i is paired with a percentage f_i , indicating that approximately f_i *100 % of data are below the value x_i

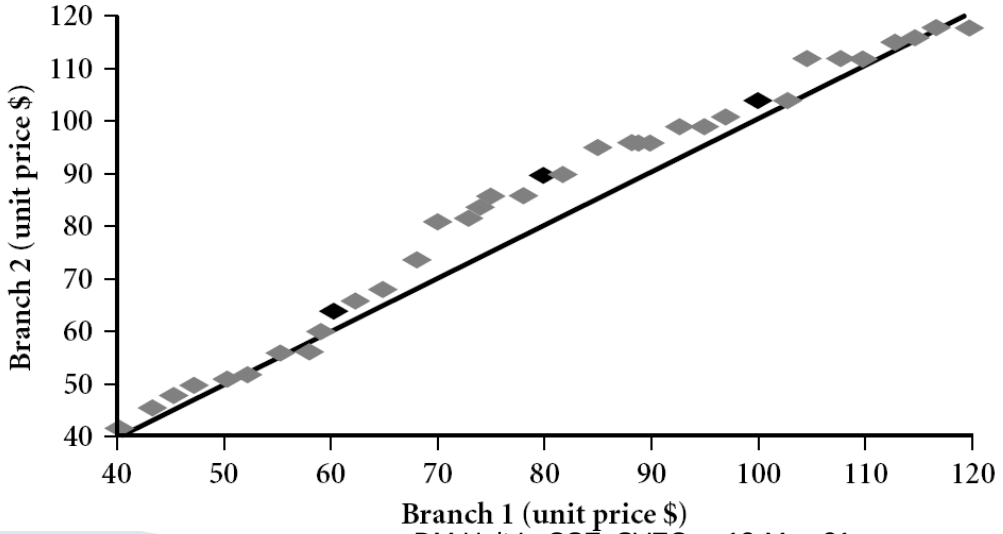
A Quantile plot for the unit price data



Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

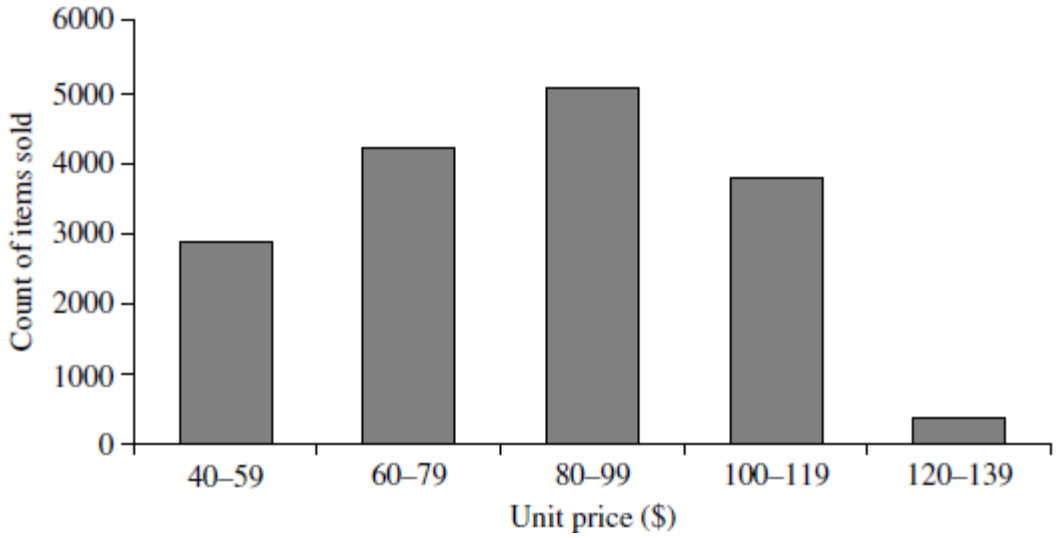
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

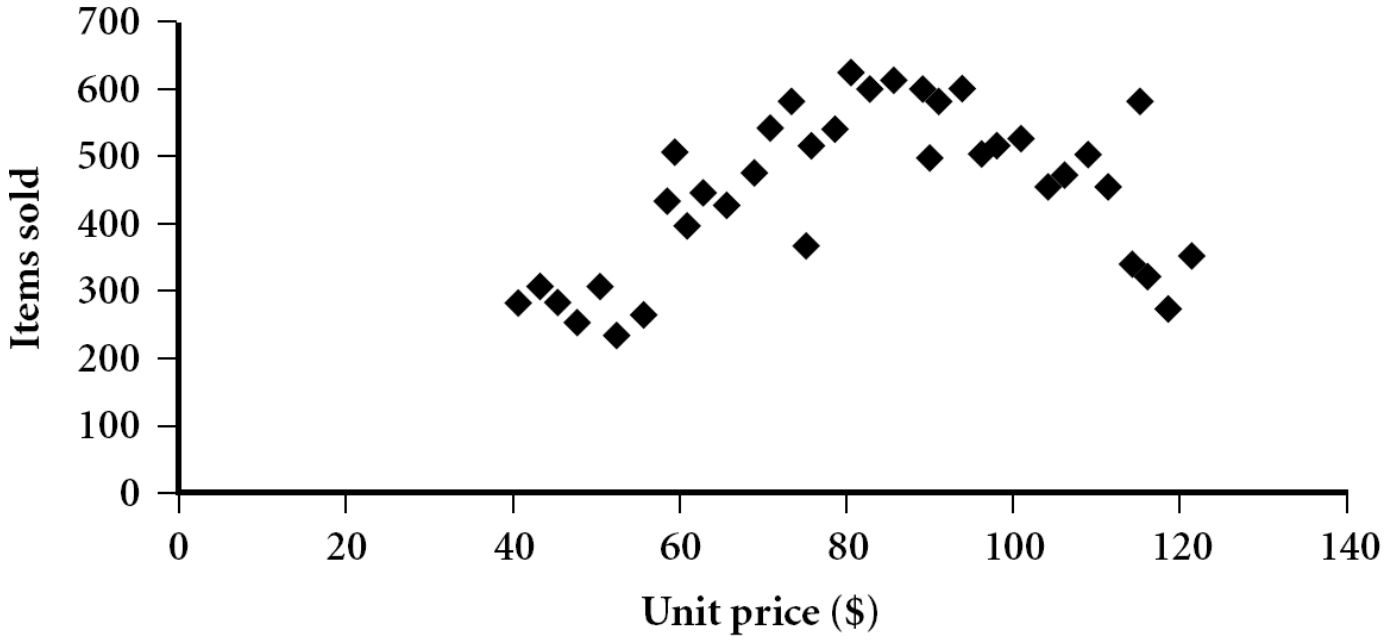
- **Histograms**: Graph display of tabulated frequencies, shown as bars
- It summarizes the distribution of a given attribute
- The range of values for *X (numeric)* is partitioned into disjoint consecutive subranges.
- The subranges, referred to as *buckets* or *bins*, are disjoint subsets of the data distribution for *X*. The range of a bucket is known as the **width**.



Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

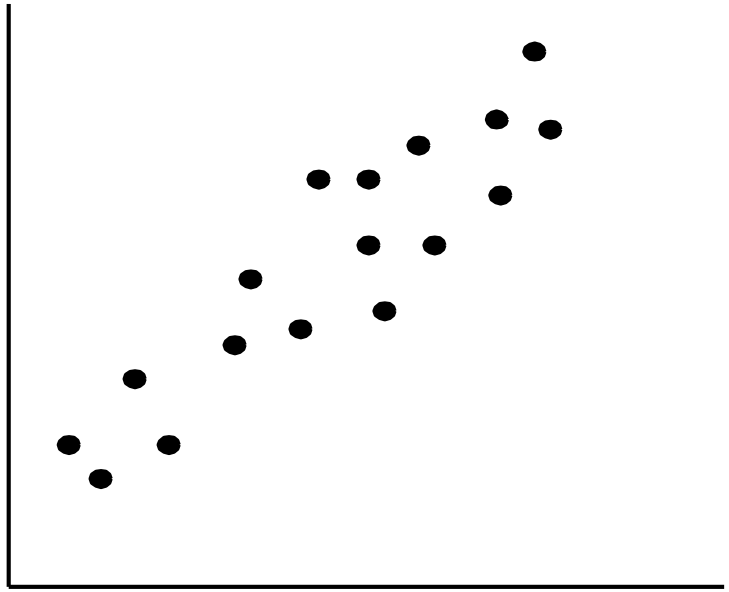
- **Scatter plots and Data Correlation :** Scatter plot provides a first look at bivariate data to see clusters of points and outliers, or to explore possibility of correlation
- Most effective graphical method for determining if there appears to be a relationship, pattern or trend between two numeric attributes



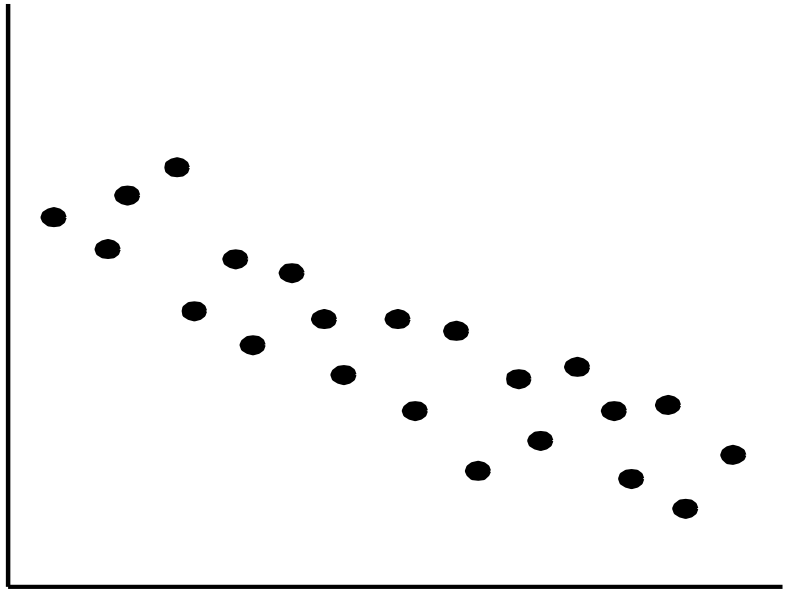
Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

Scatter plots and Data Correlation



Positively correlated data

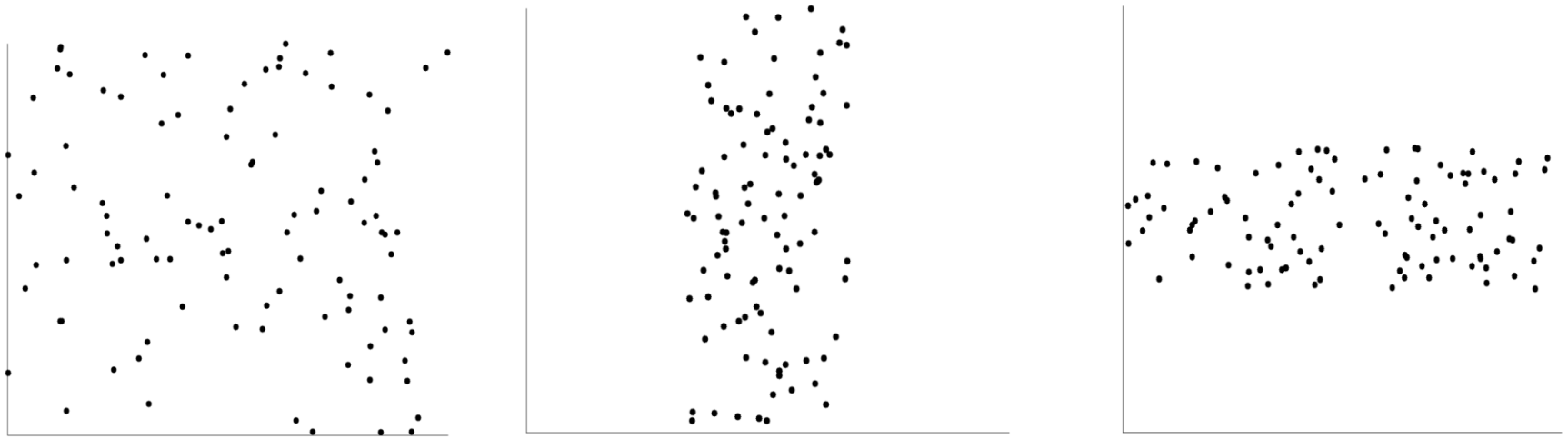


Negative correlated data

Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

Scatter plots and Data Correlation



Uncorrelated Data

Measuring Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Measuring Similarity and Dissimilarity [contd.]

- Data matrix (n x p)
  - n data points with p dimensions
  - Two modes

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix (n x n)
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Measuring Similarity and Dissimilarity [contd.]

- A data matrix is made up of two entities or “things,” namely rows (for objects) and columns (for attributes). Therefore, the data matrix is often called a **two-mode** matrix
- The dissimilarity matrix contains one kind of entity (dissimilarities) and so is called a **one-mode** matrix.
- Measure of similarity can be expressed as a function of measures of dissimilarity

  For example, for nominal data, $sim(i, j)= 1-d(i, j)$

Measuring Similarity and Dissimilarity [contd.]

Proximity Measure for Nominal Attributes

- A nominal attribute can take on two or more states. For example, *map color* is a nominal attribute that may have, say, five states: *red, yellow, green, pink*,and *blue*
- Let the number of states of a nominal attribute be *M*. The states can be denoted by letters, symbols, or a set of integers, such as 1, 2, … , *M*
- Dissimilarity

$$d(i, j) = \frac{p - m}{p}$$

where m: no. of matches, p: total no. of nominal attributes

Measuring Similarity and Dissimilarity [contd.]

Proximity Measure for Nominal Attributes

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

- Here p=1, d(i, j) evaluates to 0 if objects i and j match, and 1 if the objects differ

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Measuring Similarity and Dissimilarity [contd.]

Proximity Measures for Binary Attributes

- A contingency table for binary data

Contingency Table for Binary Attributes

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- Distance measure for symmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s+t}.$$

- Distance measure for asymmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s}.$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j).$$

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

  - Gender is a symmetric attribute
  - The remaining attributes are asymmetric binary
  - Let the values Y and P be 1, and the value N be 0

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75$$

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity of Numeric Data : Minkowski Distance

- The most popular distance measure is **Euclidean distance**

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ be two objects described by p numeric attributes

- Another well-known measure is the **Manhattan (or city block) distance**

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity of Numeric Data : Minkowski Distance

- Both the Euclidean and the Manhattan distance satisfy the following mathematical properties:
- **Non-negativity**: $d(i, j) \geq 0$ if $i \neq j$, Distance is a non-negative number
- **Identity of indiscernibles:** $d(i, i) = 0$, distance of an object to itself is 0
- **Symmetry**: $d(i, j) = d(j, i)$, Distance is a symmetric function
- **Triangle Inequality**: $d(i, j) \leq d(i, k) + d(k, j)$, Going directly from object i to object j in space is no more than making a detour over any other object k.
- A measure that satisfies these conditions is known as a metric

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity of Numeric Data : Minkowski Distance

- **Minkowski distance** is a generalization of the Euclidean and Manhattan distances

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects,

h is a real number such that $h \geq 1$ (the distance so defined is also called L-h norm)

Measuring Similarity and Dissimilarity [contd.]

Special Cases of Minkowski Distance

- $h = 1$: Manhattan (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

- $h = 2$: (L_2 norm) Euclidean distance

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

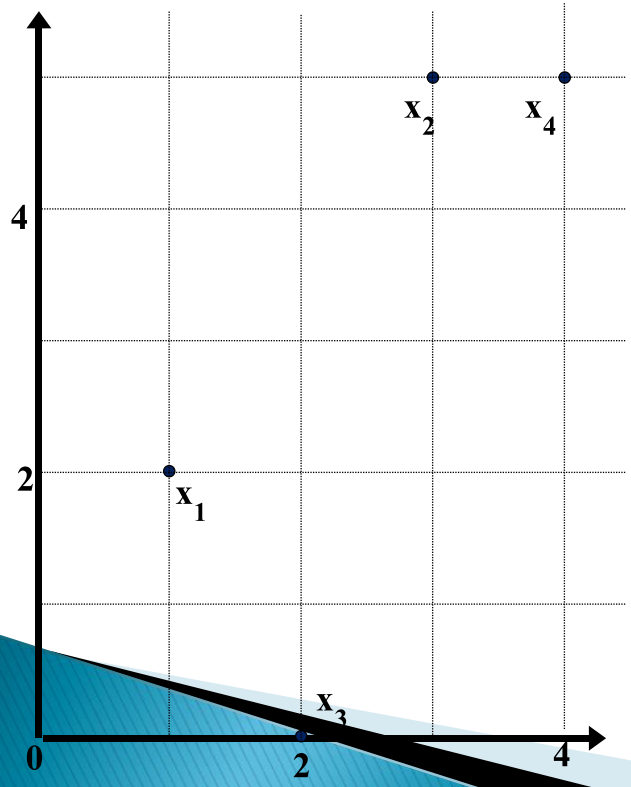
- $h \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|.$$

Measuring Similarity and Dissimilarity [contd.]

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Measuring Similarity and Dissimilarity [contd.]

Proximity Measures for Ordinal Variables

- The values of an ordinal attribute have a meaningful order or ranking about them,yet the magnitude between successive values is unknown
- An example includes the sequence *small*, *medium*, *large* for a *size* attribute
- Let *M* represent the number of possible states that an ordinal attribute can have. These ordered states define the ranking 1, … , M_f .
- Dissimilarity can be calculated in following steps:

1. The value of *f* for the i[th] object is x_{if} , and *f* has M_f ordered states, representing the ranking 1, …, M_f , Replace each x_{if} by its corresponding rank

$$r_{if} \in \{1,\ldots,M_f\}$$

2. Map the range of each attribute onto [0.0, 1.0] so that each attribute has equal weight, using

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

3. Dissimilarity can then be computed using any of the distance measures

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity for Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

$\delta_{ij}^{(f)} = 0$ if either (1) x_{if} or x_{jf} is missing, or (2) $x_{if} = x_{jf} = 0$ and attribute *f* is asymmetric binary;

otherwise, $\delta_{ij}^{(f)} = 1$

- *If f* is numeric, $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$ where h runs over all non-missing objects for attribute f

- *If f* is nominal or binary:

  $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise

- *If* f is ordinal

  Compute ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$ and treat z_{if} as numeric

Measuring Similarity and Dissimilarity [contd.]

Cosine Similarity

- It is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words.
- A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as a keyword) or phrase in the document
- **Cosine measure**: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

where $\bullet$ indicates vector dot product, $||d||$: the length of vector d

Measuring Similarity and Dissimilarity [contd.]

Example: Cosine Similarity

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Ex: Find the **similarity** between documents 1 and 2.

d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)

d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$

$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$

$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| = 25/(6.481*4.12) = 0.94$

Reference:

- Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei,3rd Edition, Morgan Kaufmann Publishers