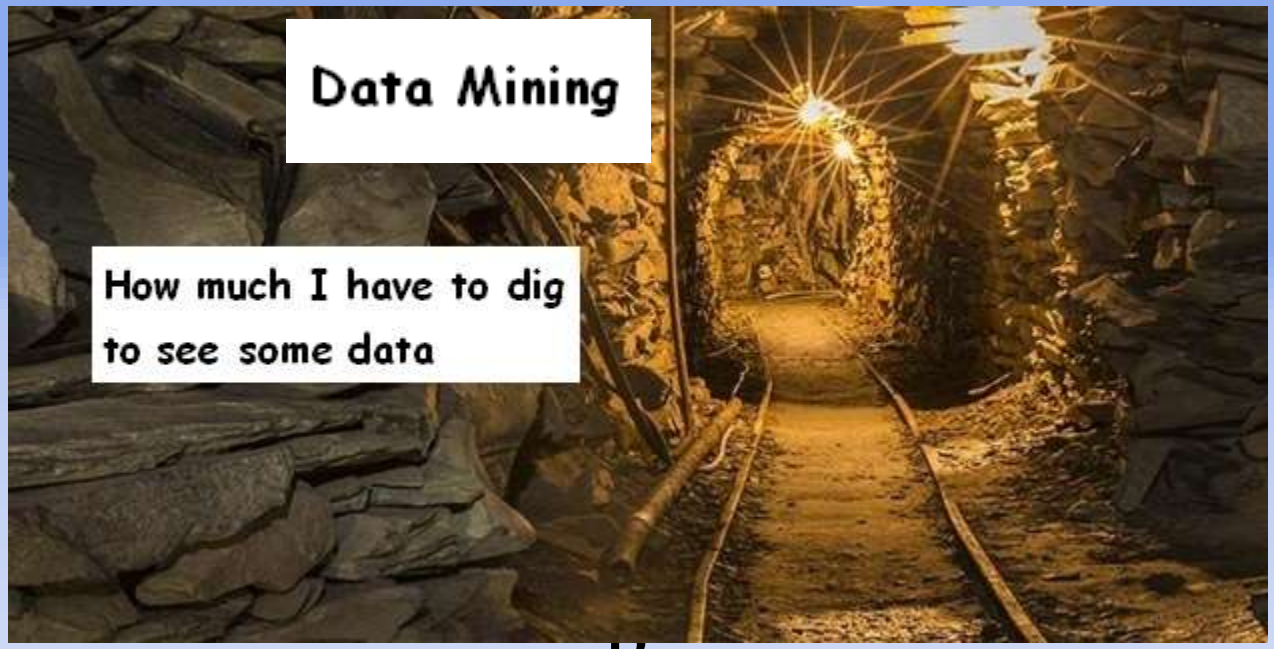
DMDW – Module-2

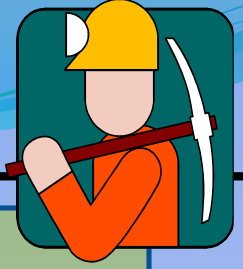


By
Dr. Pulak Sahoo
Associate Professor
Dept of CSE, SIT, BBSR

Module-2 Syllabus

Data Mining Basics: Introduction, Application areas in data mining, KDD process; Getting to know your data: Data Objects and attributes types; Data Pre-processing: Why pre-process data? Data cleaning, Data integration, Data transformation and reduction.

What Is Data Mining?



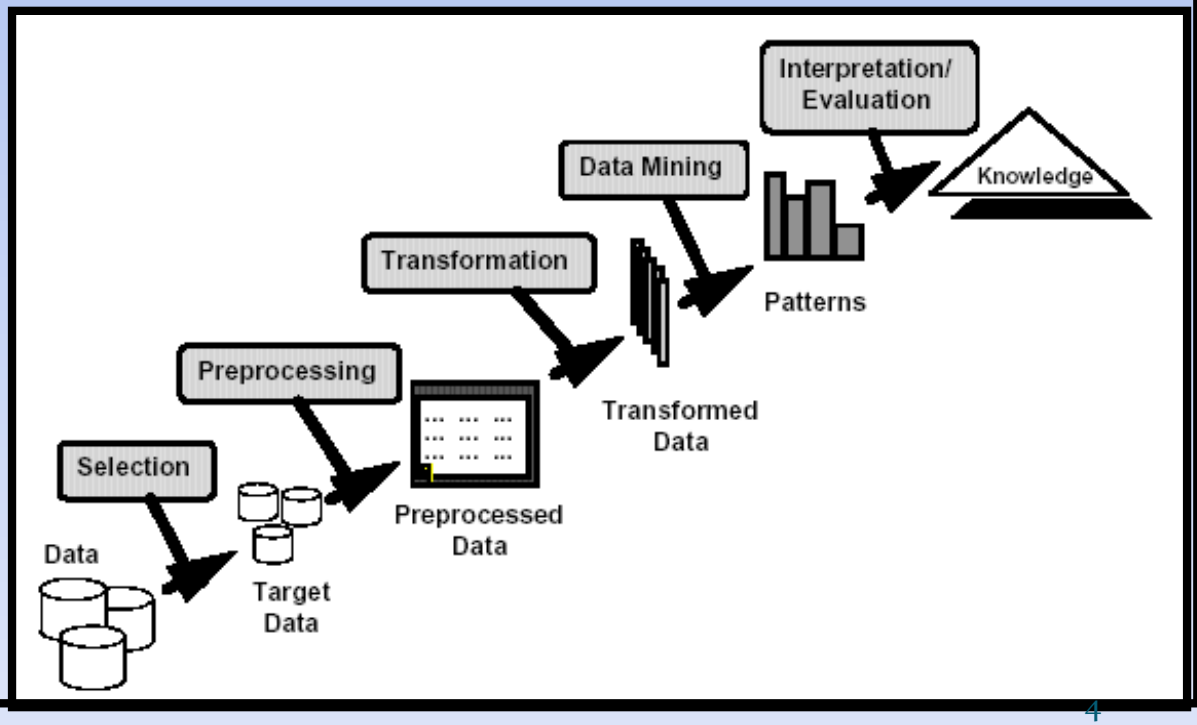
- **Data mining (knowledge discovery from data)**
  - Extraction of interesting (implicit, previously unknown & potentially useful) patterns or knowledge from huge amount of data

- **Alternative names**
  - **Knowledge discovery in databases (KDD)**, knowledge extraction, pattern analysis, info. harvesting, **Business Intelligence (BI)** etc.

- **Process of analyzing large databases to find patterns that are:**
  - **valid**: holds with certainty
  - **novel**: non-obvious to the system
  - **useful**: possible to act based on it
  - **understandable**: should be able to interpret the pattern

What is Data Mining?

- **Many Definitions**
  - Exploration & analysis of large quantities of Data, by automatic or semi-automatic means, to discover meaningful patterns



DATA MINING

- **Simplest Definition**
  - Data mining is the process of discovering interesting patterns & knowledge from large amounts of data
- The **data sources** can include DBs, DWHs, the Web, other info. repositories or Data streaming into system (stock, elections, traffic..)

What does DM Do?

Explores Your Data

Finds Patterns

Performs Predictions

Data Mining: On What Kind of Data?

- **Database** data
- **DWHs**
- **Transactional data** (OLTP – Ex: PoS, Ticket booking….)
- **Advanced DB & info. repositories**
  - **Object-oriented** databases
  - **Spatial** databases (Ex: maps, pics, graphs, traffic…)
  - **Time-series** & **temporal** data (Ex: sensor reading, env vars, medical)
  - **Text** databases & **multimedia** databases (Ex: Books, journal, newspapers, video clippings, satellite images….)
  - **Heterogeneous** & **legacy** databases (Ex: People, cells, quality, census data….)
  - **WWW**

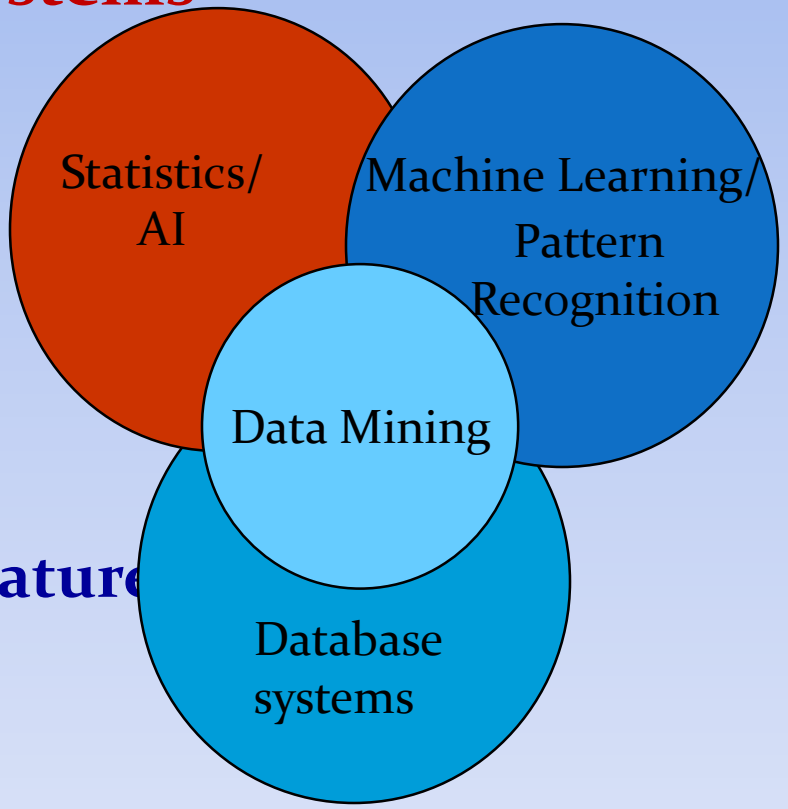
Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Relational, DWH, transactional, stream, obj-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, WWW
- **Knowledge to be mined**
  - Characterization, **classification**, **clustering**, trend, deviation, outlier analysis, etc.
- **Techniques utilized**
  - DWH (OLAP), Machine Learning, statistics etc.
- **Applications adapted**
  - *Retail, telecom, banking, fraud analysis, stock market analysis, text/web mining etc.*

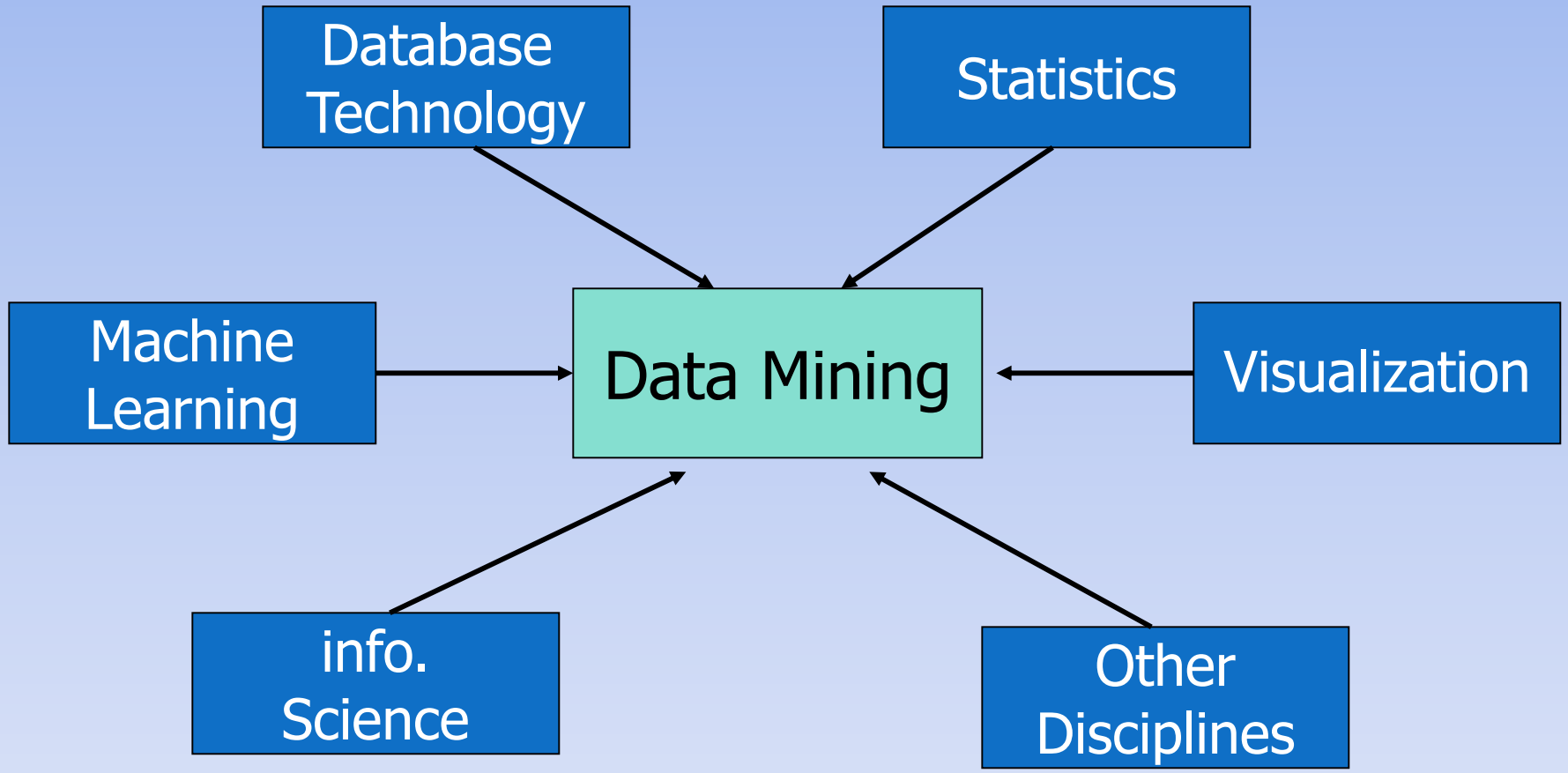
Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics & DB systems**

- **Traditional Techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



Data Mining: Confluence of Multiple Disciplines

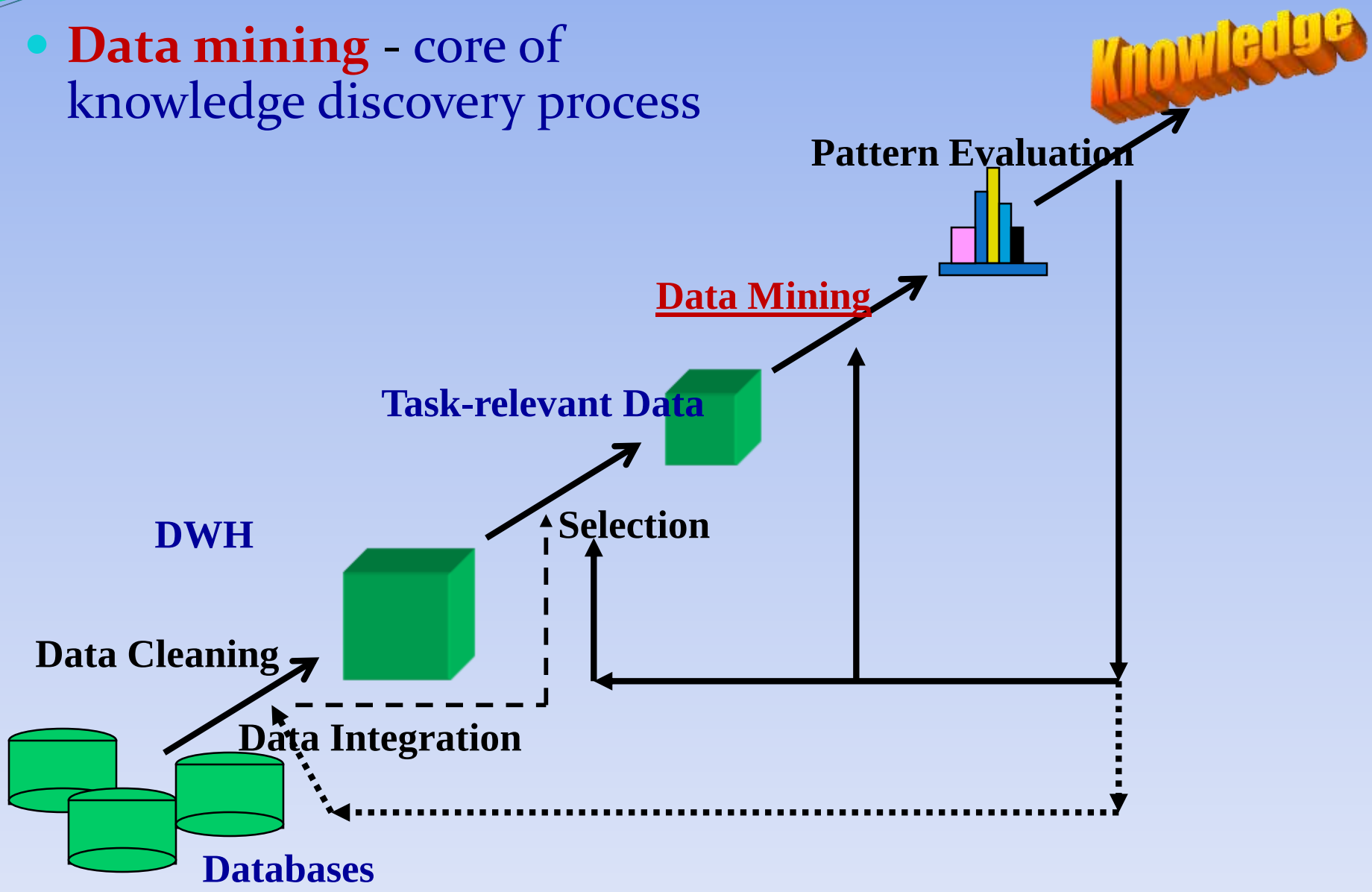


DBMS, OLAP, & DM

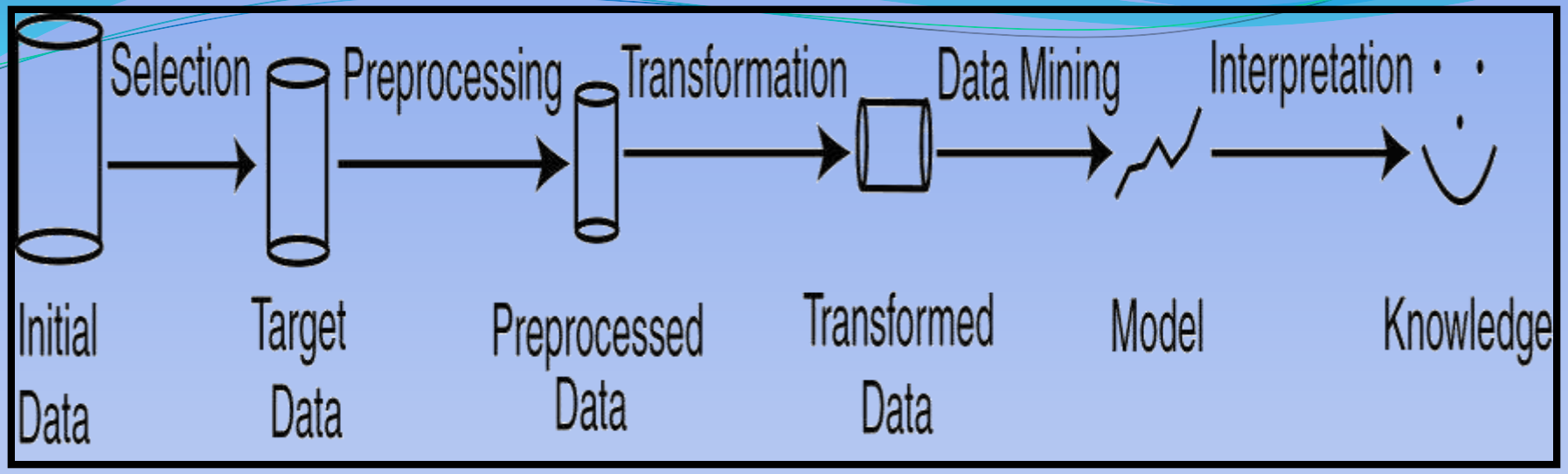
	DBMS	OLAP	Data Mining
Task	Extraction of detailed & summary data	Summaries, trends & forecasts	Knowledge discovery of hidden patterns & insights
Type of result	info.	Analysis	Insight & Prediction
Method	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
Example question	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months & why?

Knowledge Discovery (KDD) Process

- **Data mining** - core of knowledge discovery process



KDD Process



Description of each phase

- **Data Cleaning**: To remove noise & inconsistent data
- **Data Integration**: Multiple data sources may be combined
- **Data selection**: Data relevant to the analysis task are retrieved from the db
- **Data transformation**: Where data are transformed & consolidated in to forms appropriate for mining by performing summary or aggregation operations

Contd…

- **Data Mining**: An essential process where intelligent methods are applied to extract data patterns
- **Pattern Evaluation**: To identify the truly interesting patterns representing knowledge
- **Knowledge presentation**: Where visualization & knowledge representation techniques are used to present mined knowledge to users

Architecture of DM system

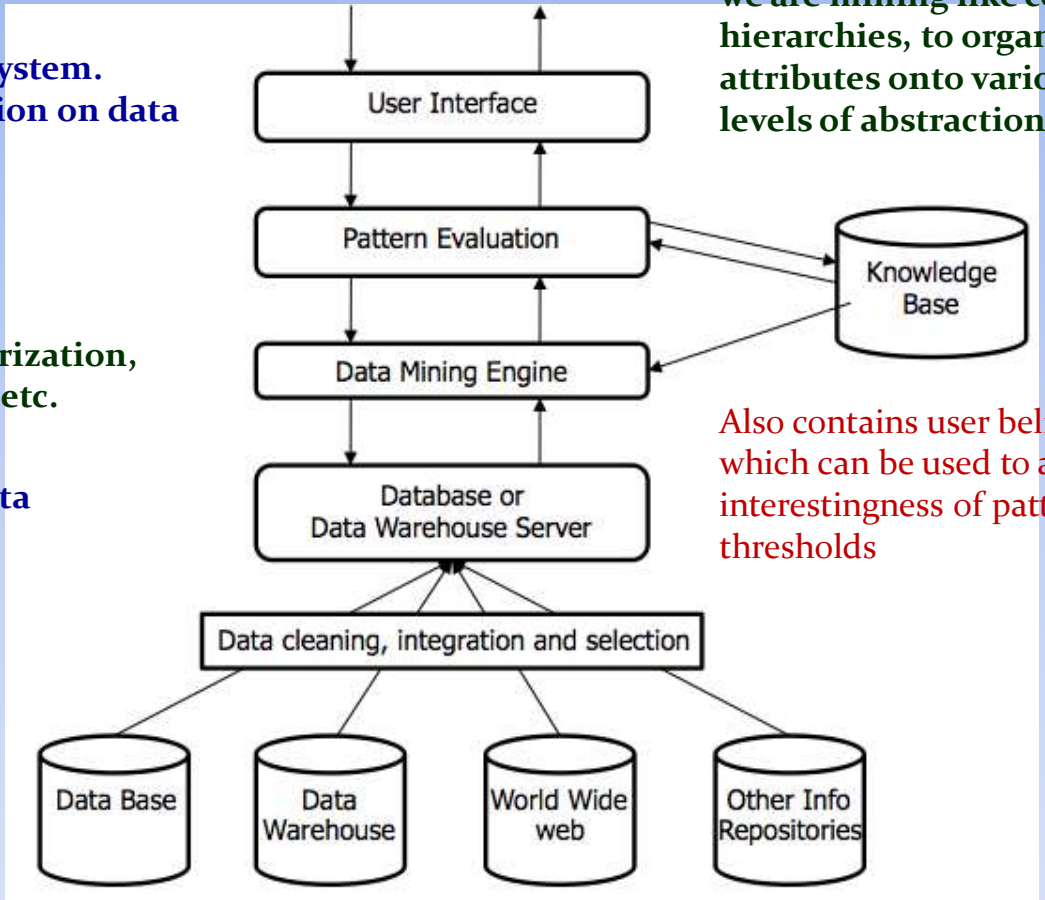
Communicates between users & DM system.
Visualizes results or perform exploration on data & schemas.

Tests for interestingness of a pattern

Performs functionalities like characterization, association, classification, prediction etc.

Is responsible for fetching relevant data based on user request

This is usually the source of data.
The data may require cleaning & integration.



This is the info. of domain we are mining like concept hierarchies, to organize attributes onto various levels of abstraction

Also contains user beliefs, which can be used to access interestingness of pattern or thresholds

Architecture of DM system

Applications of Data Mining

- **Banking**: loan/credit card approval –predict good customers based on old customers data
- **Customer relationship management**: –identify those who are likely to leave for a competitor
- **Targeted marketing**: –identify likely responders to promotions
- **Fraud detection**: télécoms, Financial transactions from an online stream of event identify fraudulent events
- **Manufacturing & production**: –automatically adjust knobs when process parameter changes
- **Medicine**: disease outcome, effectiveness of treatments - analyze patient disease history: find relationship between diseases
- **Molecular/Pharmaceutical** : identify new drugs
- **Scientific data analysis** : identify new galaxies by searching for sub clusters
- **Web site/store design & promotion**: find affinity of visitor to pages & modify layout

Potential Applications

- **Database analysis & decision support**
  - **Market analysis & management** : target marketing, customer relation mgmt, market basket analysis, cross selling, market segmentation
  - **Risk analysis & management :** Forecasting, customer retention, quality control, competitive analysis
  - Fraud detection & management
- **Other Applications**
  - Text mining (email, documents) & Web analysis
  - Intelligent query answering

Other Objective interestingness measures

- **Accuracy**: Accuracy tells us the % of data that are correctly classified by a rule
- **Coverage**: It gives the % of data to which a rule applies

Major Issues in Data Mining

Major issues in DM are partitioned into **five groups**

- **Mining Methodology**
- **User Interaction**
- **Efficiency & Scalability**
- **Diversity of data types**
- **Data Mining & Society**

Major Issues in DM Cont…

Issues related to Mining Methodology:

- Investigation of new kinds of knowledge
- Mining knowledge in multidimensional space
- Handling uncertainty, noise & incompleteness of data
- Pattern evaluation & pattern or constraint guided mining

Issues related to User Interaction:

- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages & ad-hoc data mining
- Expression & visualization of DM results

Major Issues in DM Cont…

Issues related to Efficiency & Scalability :

- Efficiency & scalability of DM algorithms
- Parallel, distributed & incremental mining methods

Issues related to Diversity of data types :

- Handling complex types of data
- Mining dynamic, networked & global data repositories

Issues related to DM & Society :

- Social impacts of data mining
- Privacy-Preserving data mining

Data Objects & Attribute Types

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**

What is an Attribute?

- An **attribute** (dimension, feature & variable) is a data field that represents a characteristic or feature of a data object
- **Ex: customer object** has *cust_ID, name, address* attributes
- Observed values for a given attribute are called observations
- A set of attributes used to describe a given object is called an **attribute vector** or **feature vector**
- The **type of an attribute** is determined by set of possible values
- **Types:** Nominal, Binary, Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Nominal Attributes

- **Nominal** means "relating to names"
- The values of nominal attribute are symbols/names of things
- Each value represents some kind of category, code or state
- Nominal attributes are also referred to as categorical attribute
- The values do not have any meaningful order

Example:

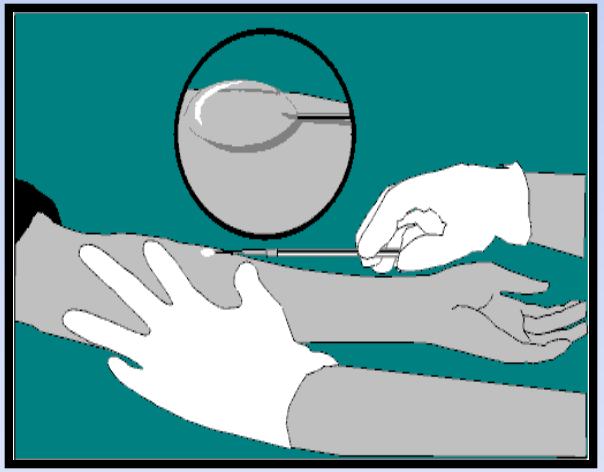
- ***hair_color:*** *black, brown, gray, & white*
- ***marital_status:*** *single, married, divorced, widowed*
- ***occupation:*** *teacher, dentist, farmer, programmer*

Binary Attributes

- A **binary attribute** is a nominal attribute with only two categories or states : 0 or 1, where 0 means that the attribute is absent & 1 means that it is present
- **Binary attributes** are referred to as Boolean if the 2 states correspond to true & false
- **Example:** *Attribute **medical test** is binary where a value of 1 means the result is positive, 0 means the result is negative*

Binary Attribute Types

- **Binary Variables:** attribute with only 2 states (0 & 1)
  - However, it can be symmetric or asymmetric
  - ***Symmetric binary***: both outcomes are equally important
    - Ex: gender
  - ***Asymmetric binary***: outcomes are not equally important.
    - Ex: outcome of medical test (positive vs. negative)
    - **Convention:** assign 1 to most important outcome (e.g., HIV positive)



Ordinal Attributes

- Values have a meaningful order but magnitude between successive values is not known

- **Example**:

  Size = {small, medium, large}

  Grade = { A+, A, A-, B+, & so on }

  Professional_rank = { assistant, associate, professor }

- The values have a meaningful sequence. They describe a feature of an object without giving an actual size or quantity

Numeric Attribute

- Nominal, Binary & Ordinal attributes are **qualitative**
- They describe a feature of an object without giving an actual quantity. The values of qualitative attributes are typically words representing categories
- A numeric attribute is a measurable **quantity** represented in integer or real values
- **Numeric attributes** can be **interval-scaled** or **ratio-scaled**

Interval-Scaled Attributes

- Interval-scaled attributes are measured on a scale of equal-size units
- The values of interval-scaled attributes have order & can be +ve / -ve
- They provide a ranking of values, allow us to compare & quantify the difference between values
- A **temp.** attribute is interval-scaled because we can quantify the diff. between 2 temp. values

  Ex: a temp of 20C is 5 degrees higher than a temp of 15C

  calendar dates (year 2002 & 2010 are 8 years apart)

- **An interval-scaled attribute doesn't have an inherent zero-point.**

  Ex: temp & calendar dates do not have a 0-point

Ratio-Scaled Attributes

- A **ratio-Scaled attribute** is a numeric attribute with an inherent zero-point
- If a measurement is ratio-scaled, we can speak of a value as being a multiple or ratio of another value
- The values are ordered & diff. between values can be computed
- **Ex:** Temp in Kelvin scale has a true zero-point.

  count attributes such as years_of_exp (employees objects) & no_of_words (documents objects)

Discrete versus Continuous Attributes

- **A discrete attribute has a finite set of value**s, which may or may not be represented as integers
- The attributes ***hair_color*** (black, white, brown, gray), ***smoker*** (yes,no) , ***medical_test*** (positive, negative), ***size*** (large, medium, small) each have a **finite no. of values &** are discrete
- An attribute is countably infinite if the set of possible values is infinite but the values can be put in a one-to-one correspondence with natural no.s. **Ex**: the attribute ***customer_ID*** is **countably infinite**
- If an attribute is not discrete, it is **continuous**
- **Continuous** attributes are represented as floating-point variables
  - **EX:** temp, height or weight

Basic Statistical Descriptions of Data

Basic Statistical Descriptions of Data

- To **understand the data better**, we need to find it's: **(1) central tendency, (2) variation and (3) spread**
- **Also need to understand below statistical figures:**
- **Data set characteristics**
  - ***median, max, min, quantiles, outliers, variance,*** *etc.*
- **Numerical dimensions**
  - **Data dispersion**: analysis with multiple granularities of precision
    - **Boxplot** or **quantile** analysis

Measuring of central tendency

- **Measures of central tendency** - measure the location of the center of a data distribution in order to describe the set of data
- Also called measures of central location
- Measures of central tendency : **mean, median, mode**

MEAN

- The most common & effective numeric measure of the "**center**" of a set of data is ***(arithmetic) mean***
- *Let **x1,x2,… ,xN** be a set of **N** numeric values (Ex: Salary, Age, Length) .*
  *The **Mean** =*

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

- *Let each value **xi** in a set may be associated with a weight (importance, Freq.) **wi** . The **Weighted Avg /Mean** =*

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

- **Trimmed mean:** Excluding the extreme values (outliers)

- **Mean: Consider following values for *salary (in thousands of dollars)*:** 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
- Mean =(30+36+47+50+52+52+56+60+63+70+70+110) = 696/12 = 58.
- **Mean** is the most useful quantity for describing a data set
- Problem with the mean: **sensitivity to extreme (e.g., outlier) values**
- Even a **small number of extreme values can corrupt the mean**
- An **outlier** is a value that differs significantly from the others

Participant	1	2	3	4	5
Reaction time (milliseconds)	832	345	365	298	380

$$\sum x = 832 + 345 + 365 + 298 + 380 = 2220$$

$$\text{Mean } (\bar{x}) = \sum x/n = 2220/5 = 444$$

Due to the outlier, the mean becomes much higher, even though all the other numbers in the data set stay the same.

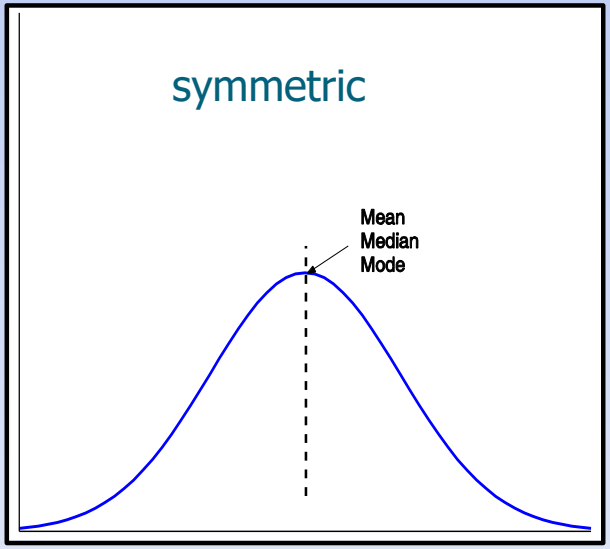
Median

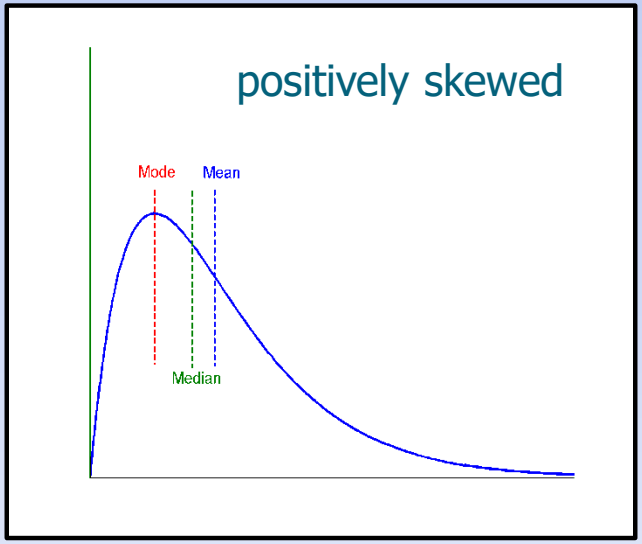
- For **skewed (asymmetric) data**, a better measure of the center of data is the **median.** It is the **middle value in a set of ordered data values**
- One middle value if odd no. of values or Avg. of the middle two values otherwise. For grouped data, Median is calculated by interpolation

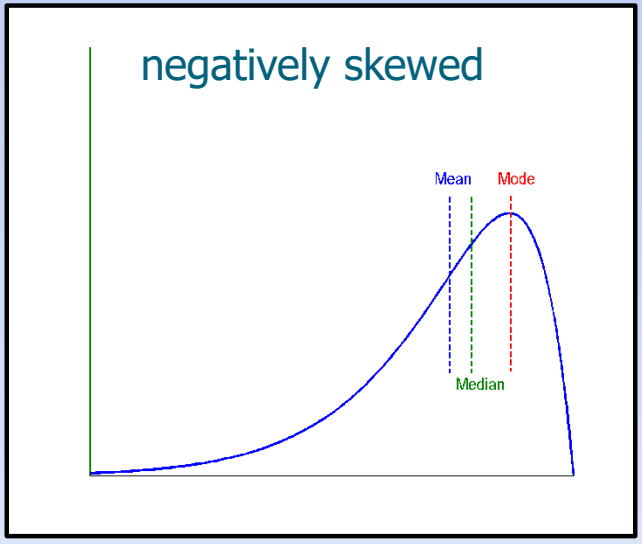
$$\text{Median} = \ell + h\left(\frac{\frac{N}{2} - C}{f}\right)$$

Symmetric vs. Skewed Data

Median, mean & mode of symmetric, positively & negatively skewed data







Skewed data set

- A data is called as **skewed** when curve appears distorted either to the left or to the right, in a statistical distribution
- In a normal distribution, the graph appears **symmetry** meaning that there are about as many data values on the left side of the median as on the right side

Effects of skewness

- If there are too much skewness in the data, then many statistical model don't work
- skewed data, the tail region may act as an outlier for the statistical model we know that outliers adversely affect model's performance especially on regression-based models

- **Suppose we have the following values for *salary (in thousands of dollars), shown*** in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110
- Median = the avg. of the two middlemost values
- That is =(52+56)/2=108/2=54
- Suppose that we had only the first 11 values in the list
- From the odd no. of values, the median = middlemost value
- That is 6th value = 52

The following data represents the heights (in cm) of 51 girls of Class x. Find the median height.

Height (in cm)	Number of Girls
Less than 140	4
Less than 145	11
Less than 150	29
Less than 155	40
Less than 160	46
Less than 165	51

Working rule to find median

Step 1: Prepare a table containing less than type cumulative frequency with the help of given frequencies.

Step 2 : Find out the cumulative frequency to which $\frac{N}{2}$ belongs. Class-interval of this cumulative frequency is the median class-interval.

Step 3 : Find out the frequency f and lower limit l of this median class.

Step 4 : Find the width h of the median class interval

Wages (in Rs)	No. of labourers
200 – 300	3
300 – 400	5
400 – 500	20
500 – 600	10
600 – 700	6

Wages (in Rs)	No. of labours	Less than type cumulative frequency
200 – 300	3	3
300 – 400	5	8 = C
400 – 500	20 = f	28
500 – 600	10	38
600 – 700	6	44 = N

Here, the median class is 400 – 500 as $\frac{44}{2}$ i.e. 22 belongs to the cumulative frequency of this class interval.

Lower limit of the median class = ℓ = 400

width of the class interval = h = 100

Cumulative frequency preceding median class frequency = C = 8

Frequency of Median class = f = 20

$$\text{Median} = \ell + h\left(\frac{\frac{N}{2} - C}{f}\right)$$

$$\text{Median} = \ell + h\left(\frac{\frac{N}{2} - C}{f}\right) = 400 + 100\left(\frac{\frac{44}{2} - 8}{20}\right)$$

$$= 400 + 100\left(\frac{22-8}{20}\right) = 400 + 100\left(\frac{14}{20}\right)$$

$$= 400 + 70 = 470$$

Hence, the median of the given frequency distribution is 470.

Exercise

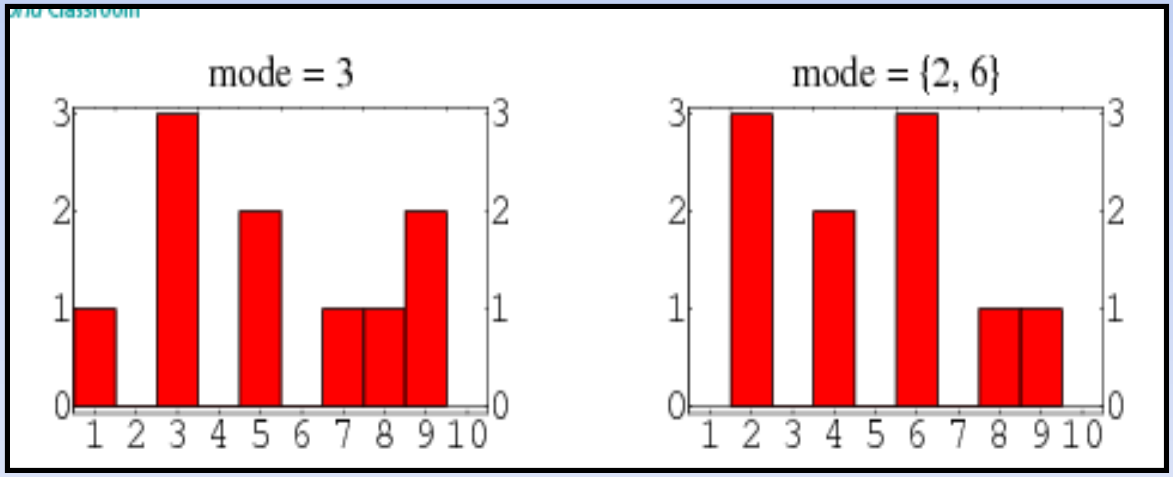
2.3 Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Compute an *approximate median* value for the data.

Measuring "Mode"

- **Mode**
  - **Value that occurs most frequently in the data**
    - **Types of data: *Unimodal, bimodal, trimodal, multimodal***
  - **Example:**
  - For a data set (3, 7, 3, 9, 9, 3, 5, 1, 8, 5), the unique mode is 3.



- Similarly, for a data set (2, 4, 9, 6, 4, 6, 6, 2, 8, 2) there are two modes: 2 and 6 (bimodal)

- **Example:**
- Suppose we have the following values for *salary (in thousands $) shown* in increasing order: 30, 36, 47, 50, **52**, 52, **56**, 60, 63, **70**, **70**, 110. What type of modal it is ?
- **Ans**-Bimodal (modes: 52 & 70)

- The Relationship between the sample mean, statistical median, & mode which appears to hold for unimodal curves of moderate asymmetry is given by

$$\text{mean} - \text{mode} \approx 3\,(\text{mean} - \text{median})$$

Exercise

- Suppose that the data for analysis includes the **attribute *age***. *The values for the data tuples are* 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36,40, 45, 46, 52, 70.

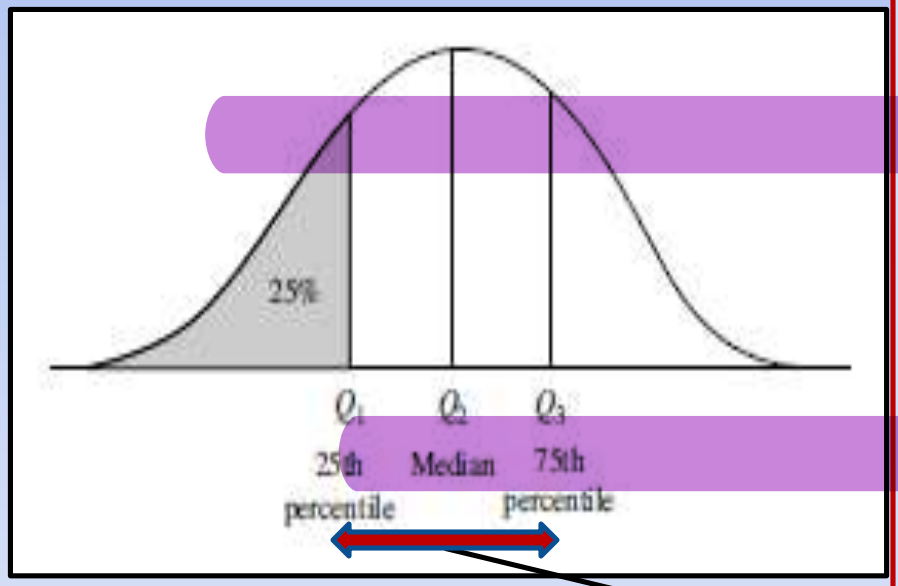
- (a) What is the ***mean*** *of the data? What is the* ***median****?*
- *30,25*

- (b) What is the ***mode*** *of the data? Comment on the* ***data's modality*** *(i.e., bimodal, trimodal, etc.).*
- *25,35*

Measuring the scattering of Data: Range, Quartiles, Variance, Standard Deviation & Inter-quartile Range

- Are the different measures to assess the **distribution** or **spread** of numeric data
- Let ***x1,x2,… ,xN*** *be a set of observations for some numeric* ***attribute X***
- *The* ***range*** of the set – It is the difference between the largest & the smallest value
- Suppose that the data for attribute *X are sorted in increasing numeric order &* split into equal-size sets
- Division points that split data into 4 equal parts are called *Q1* (1st quartile), *Q2* (2nd quartile or median), & *Q3* (3rd quartile)
- These data points are called ***Quantiles*** – Points taken at regular intervals of a data distribution, dividing it into equal size consecutive sets

- **2-quantile:** the data point dividing the lower & upper halves of data distribution (same as the median)
- **4-quantiles:** the three data points that split the data distribution into four equal parts, each part represents one-4th of the data distribution
  - Also referred to as quartiles
- **100-quantiles** are more commonly referred to as percentiles; divide the data distribution into 100 equal-sized consecutive sets



The quartiles give an indication of a distribution's center, spread & shape

The **1st quartile** ***Q1, is the 25th percentile. It cuts off the lowest 25% of the data***

The **3rd quartile** ***Q3, is the 75th percentile—it cuts off the lowest 75%*** (*or* highest 25%) of the data

The **2nd quartile** is the **50th percentile** (median), gives the center of the data distribution

The distance between 1st & 3rd quartiles is a measure of spread (range covered by the middle half of data. This distance is called the **interquartile range (IQR)**

- Suppose we have the following values for *salary (in thousands of dollars), shown* in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
- The quartiles are the 3 values that split the data set into 4 equal parts.
- The data contain 12 observations, already sorted in increasing order.

Measures Data Variability

- **Variability** (spread or dispersion) refers to how spread a set of data is
- Variability gives us a way to describe how much data sets vary
- This allows us to use statistics to compare our data to other data sets

- **The 4 main ways to describe variability in a data set are:**
  - **Range**
  - **Interquartile range**
  - **Variance**
  - **Standard deviation**

- **Range** is the amount between your smallest & largest item in the set

- **Ex**: If you earned \$250 1st week, \$30 the 2nd week and \$800 the 3rd week. The range for your pay is \$30 to \$800 (800-30=770)

- **Inter-quartile range** is a measure of where the "**middle fifty**" is in a data set

- It is a **measure of where the bulk of the values lie**

- The **inter-quartile range** is the 1st quartile subtracted from the 3rd quartile: **$IQR = Q_3 - Q_1$**

Example – Odd number

- Step 1: **Put the numbers in order**
  1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- Step 2: **Find the median**
  1, 2, 5, 6, 7, **9**, 12, 15, 18, 19, 27.
- Step 3: **Place parentheses around the numbers above & below the median.** It makes Q1 and Q3 easier to spot.
  (1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
- Step 4: **Find Q1 and Q3**
  Find Q1 as a median in the lower half of the data & Q3 as a median for the upper half.
  (1, 2, **5**, 6, 7), **9**, (12, 15, **18**, 19, 27). Q1 = 5 and Q3 = 18.
- Step 5: Find **Q3-Q1 the inter-quartile range**
  18 – 5 = 13.

Example – Even number

- Find IQR for the following data set: 3, 5, 7, 8, 9, 11, 15, 16, 20, 21.
- Step 1: **Put the numbers in order**.
  3, 5, 7, 8, 9, 11, 15, 16, 20, 21.
- Step 2: **Make a mark in the center of the data**:
  3, 5, 7, 8, 9, | 11, 15, 16, 20, 21.
- Step 3: **Place parentheses around the numbers above & below the mark, making Q1 and Q3 easier to spot.**
  (3, 5, 7, 8, 9), | (11, 15, 16, 20, 21).
- Step 4: **Find Q1 and Q3**
  Q1 is the median (the middle) of the lower half of the data, and Q3 is the median (the middle) of the upper half of the data.
  (3, 5, **7**, 8, 9), | (11, 15, **16,** 20, 21). Q1 = 7 and Q3 = 16.
- Step 5: **Subtract Q1 from Q3**.
  $IQR = 16 - 7 = 9$.

Variance & Standard Deviation

- **Variance** & **Standard deviation** are measures of data distribution
- They indicate **how spread out a data distribution is**
- Low standard deviation => data observations are close to the **mean**
- High standard deviation => data are spread over a large range of values
- The **variance of *N observations, x1,x2, … ,xN, for a numeric attribute X is***

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2, \qquad (2.6)$$

where $\bar{x}$ is the mean value of the observations, as defined in Eq. (2.1). The **standard deviation**, σ, of the observations is the square root of the variance, σ^2.

Example – Find Standard Deviation σ

$$\sigma^2 = \frac{1}{12}(30^2 + 36^2 + 47^2 \ldots + 110^2) - 58^2$$

$$\approx 379.17$$

$$\sigma \approx \sqrt{379.17} \approx 19.47.$$

Measuring the Dispersion of Data

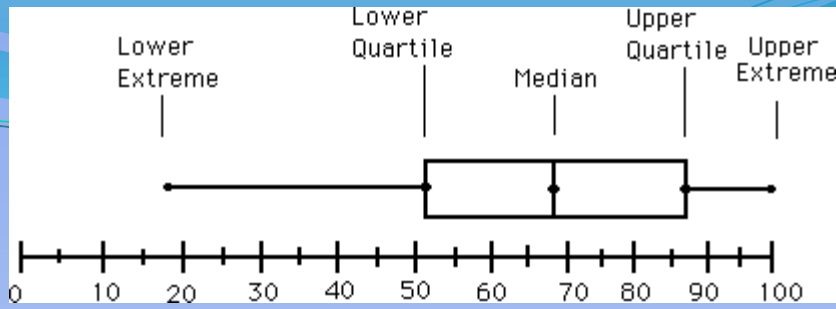
- **Quartiles, Outliers & Boxplots**
- **Quartiles**: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
- **Inter-quartile range**: IQR = $Q_3 - Q_1$
- **Five number summary**: **min, Q_1, median, Q_3, max**
- **Boxplot**: ends of the box are the **quartiles**; **median** is marked; add **whiskers**, and plot **outliers** individually
- **Outlier**: usually, a value higher/lower than 1.5 x IQR
- **Variance & standard deviation**
  - **Variance**:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

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

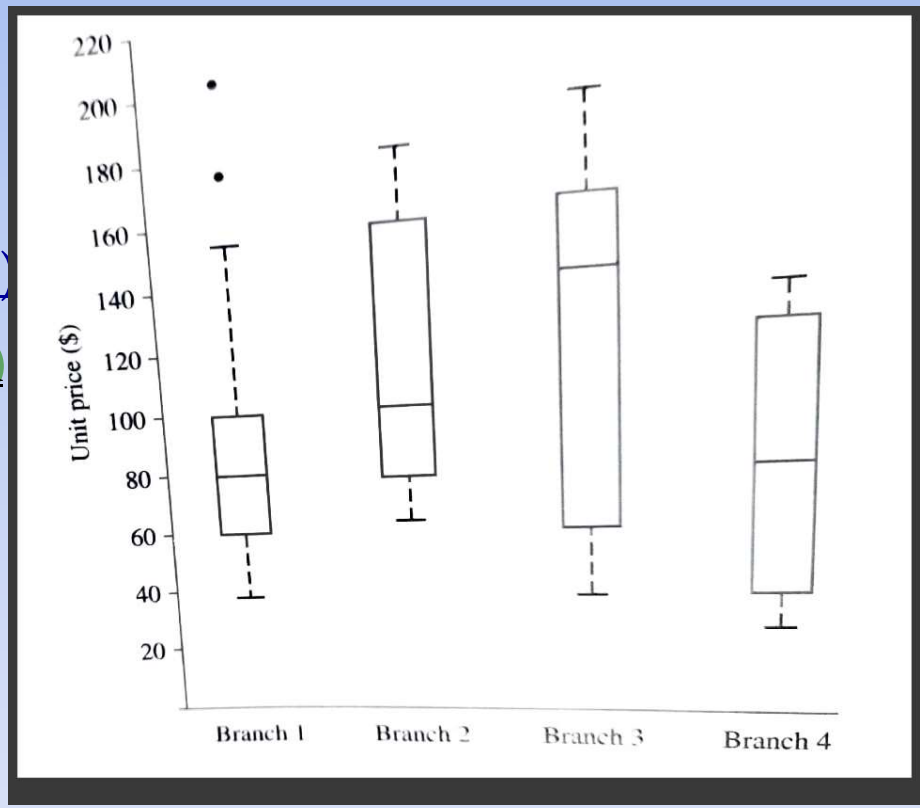
  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Boxplot Analysis

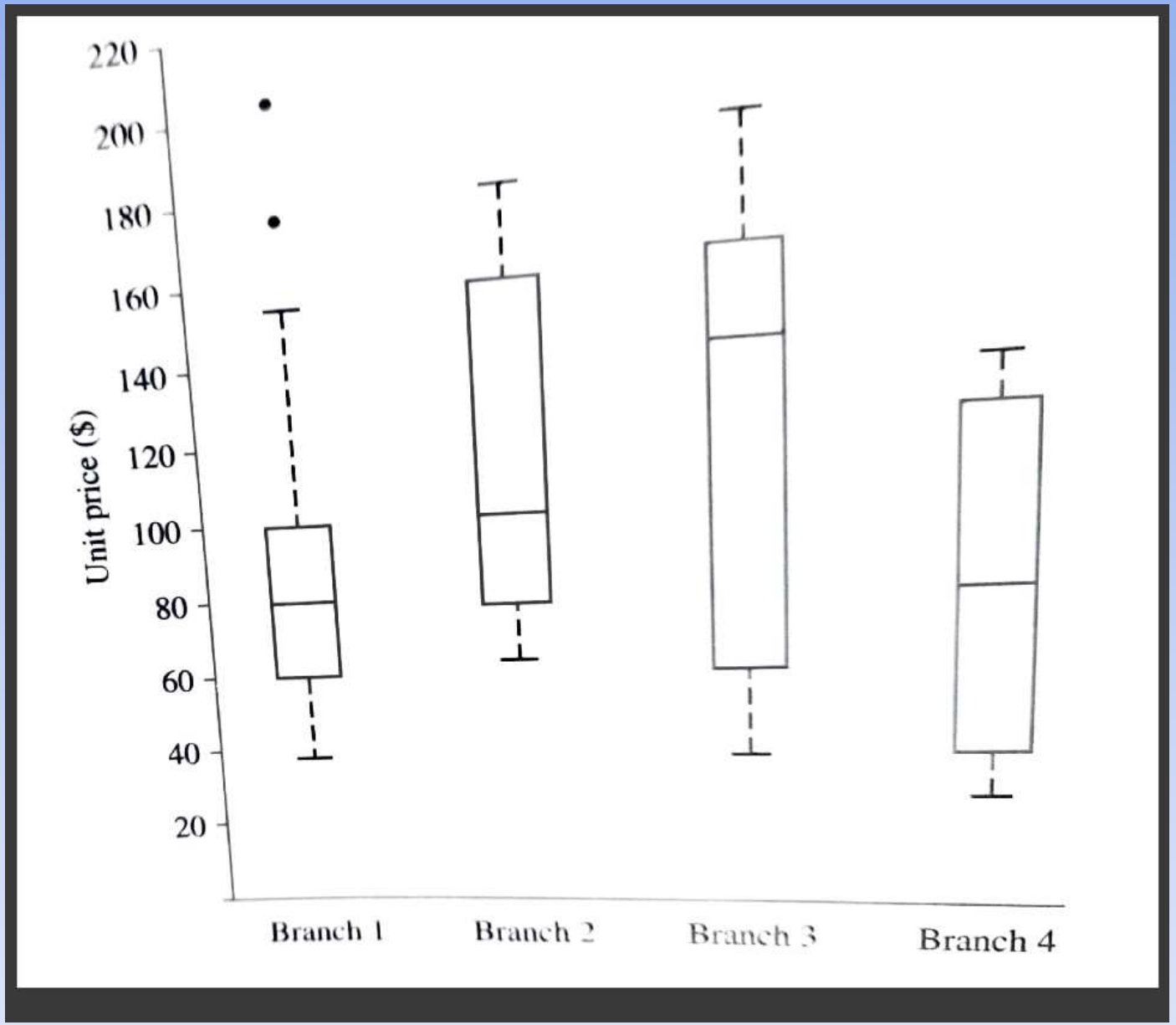


- **Five-number summary** of a distribution
  - ***Minimum, Q1, Median, Q3, Maximum***

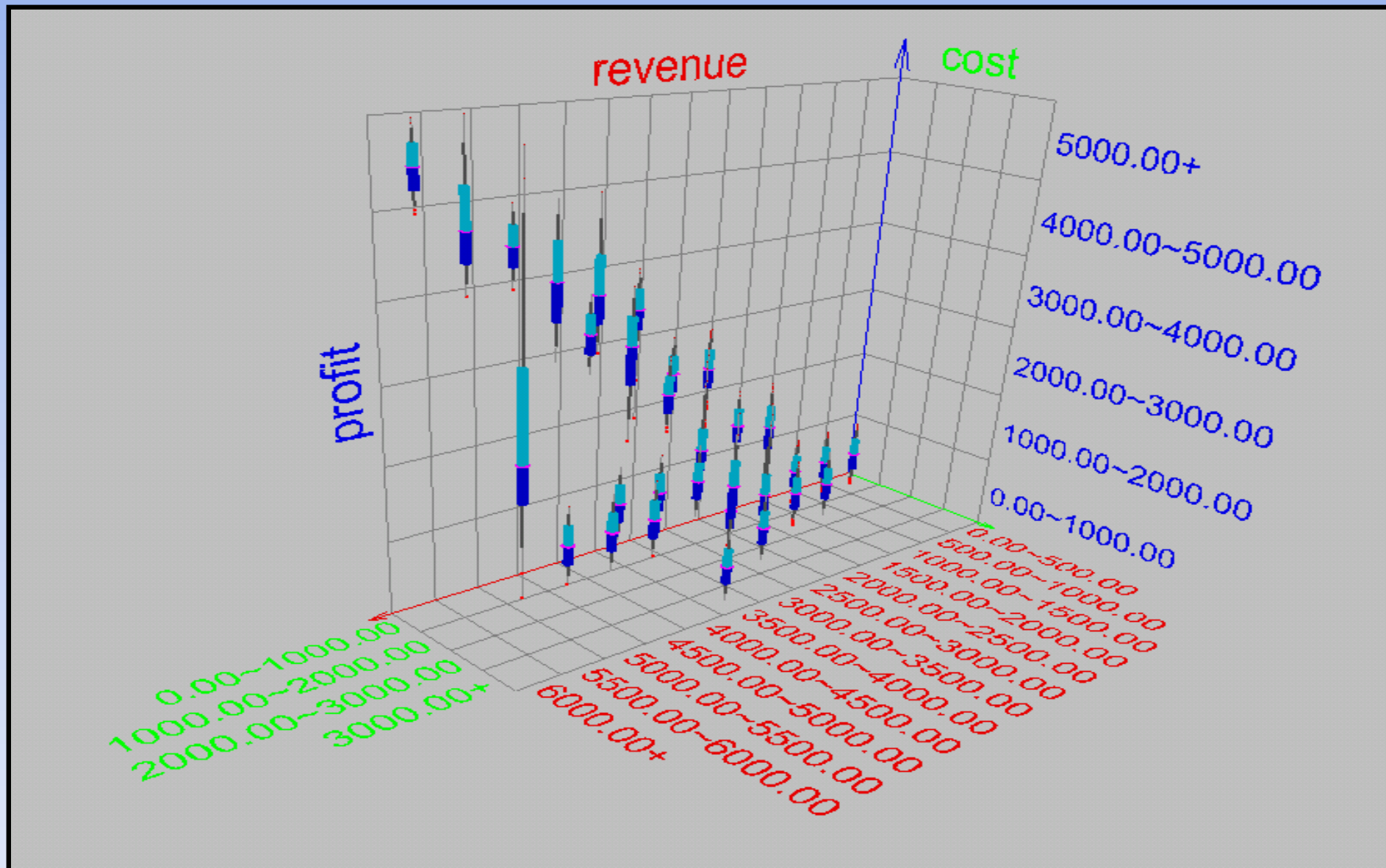
- **Boxplot**
- Data is represented with a **box**
- The ends of the box are at the **1st & 3rd quartiles** (the height of box is **IQR**)
- The **median** is marked by a line within the box
- **Whiskers**: two lines outside the box extended to Minimum & Maximum
- **Outliers**: Points beyond a specified outlier threshold, plotted individually



BoxPlot

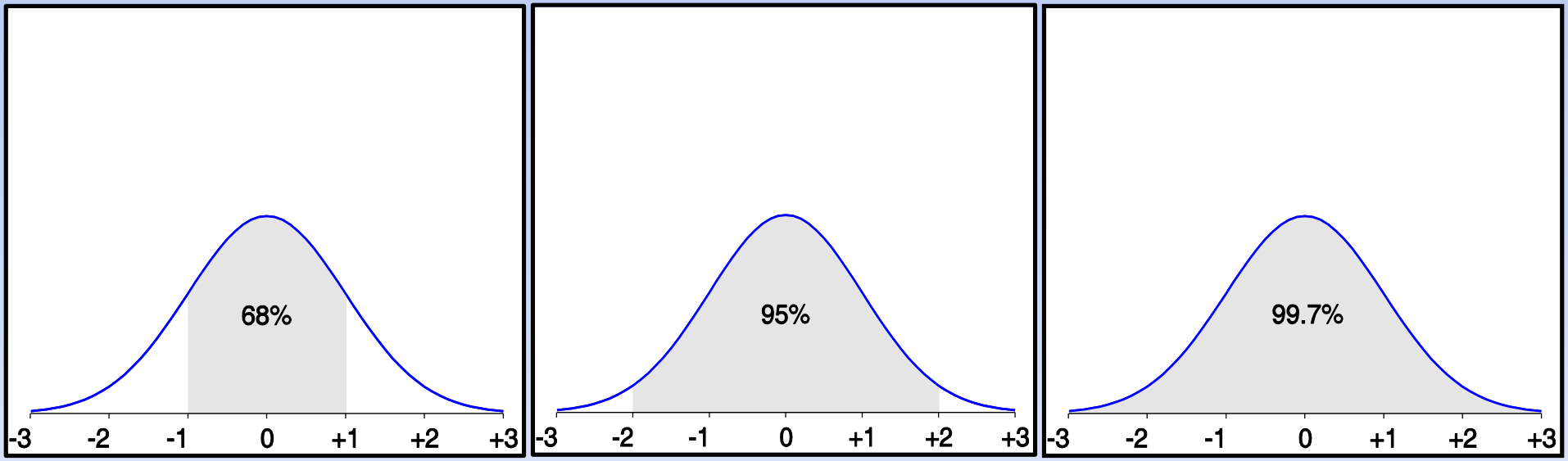

220
200
180
160
140
120
100
80
60
40
20
Unit price ($)
Branch 1
Branch 2
Branch 3
Branch 4

Visualization of Data Dispersion: 3-D Boxplots



Properties of Normal Distribution Curve

- **The normal (distribution) curve**
  - **From μ–σ to μ+σ**: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - **From μ–2σ to μ+2σ:** contains about 95% of it
  - **From μ–3σ to μ+3σ**: contains about 99.7% of it

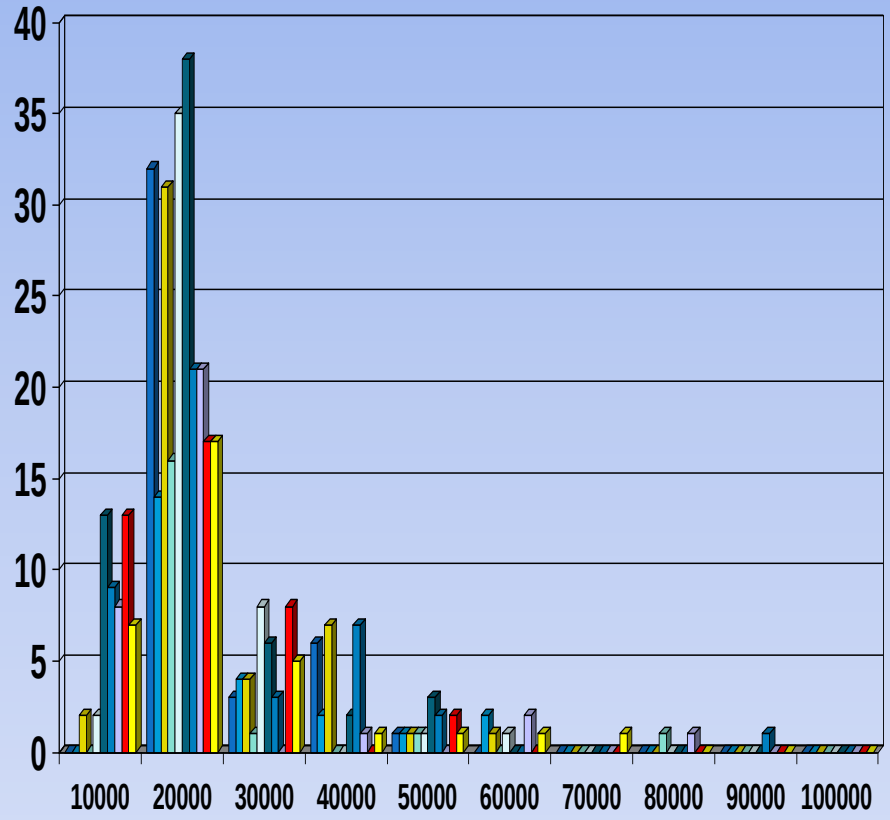


Graphic Displays of Basic Statistical Descriptions

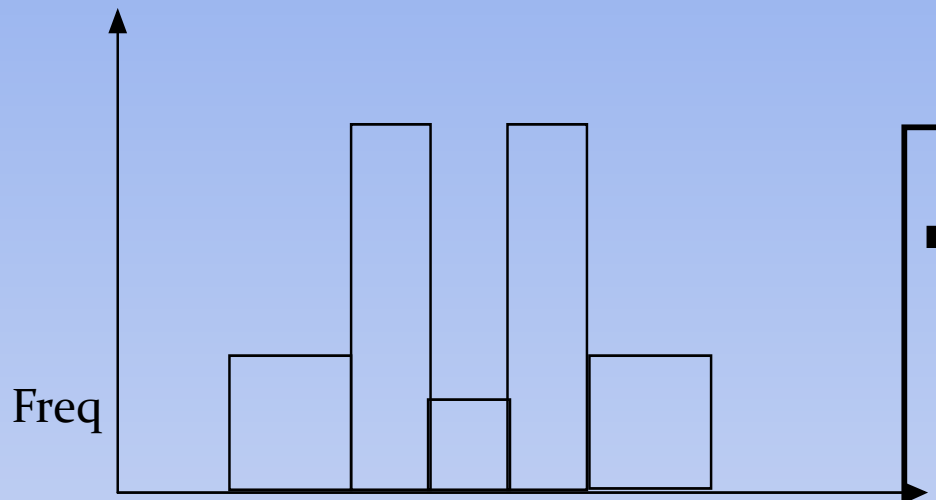
- **Boxplot**: Graphic display of ***five-number summary***
- **Histogram**: **x-axis** are **values**, **y-axis** represents **frequencies**
- **Quantile plot**: Each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the quantiles of another
- **Scatter plot**: Each pair of values is a pair of coordinates and plotted as points in the plane

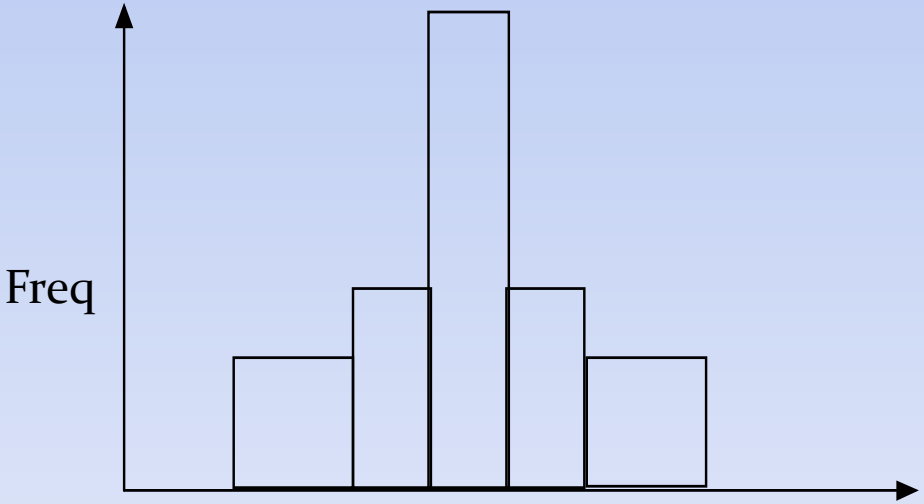
Histogram Analysis

- **Histogram**: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a **bar chart** in that it is the *area* of the bar that denotes the value, not the height as in bar charts
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots

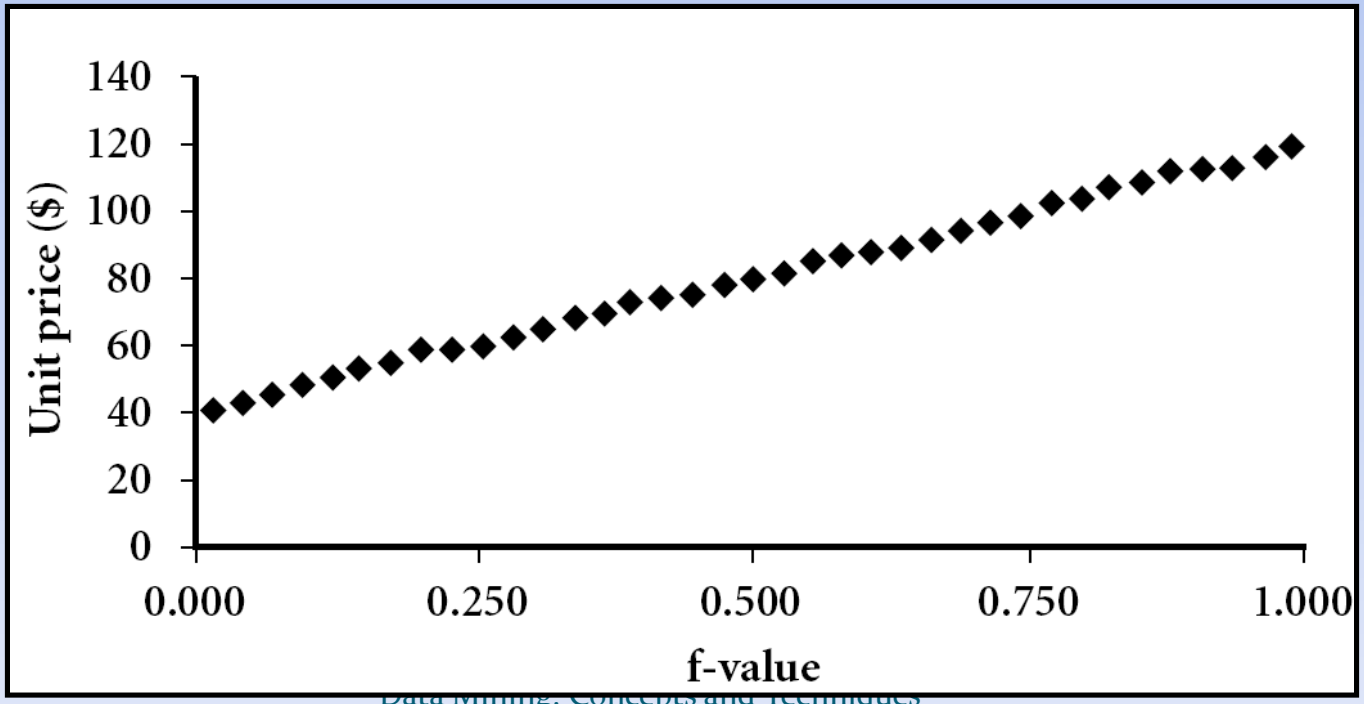




- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have different data distributions

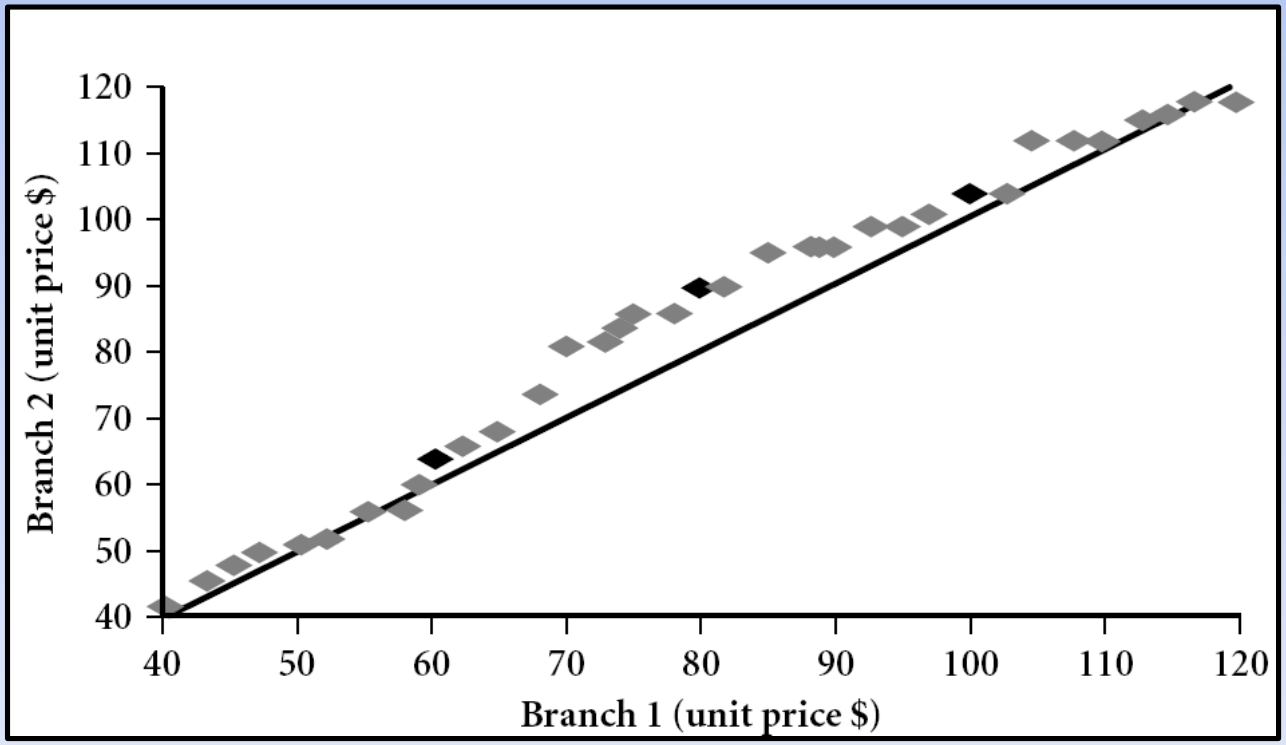
Quantile Plot

- **Displays all of the data** (user has assess to both the overall behavior & unusual occurrences)
- Plots **quantile** information
  - For a data x_i sorted in increasing order
  - f_i indicates that approximately $100\,f_i\%$ of the data are $\Leftarrow$ value x_i



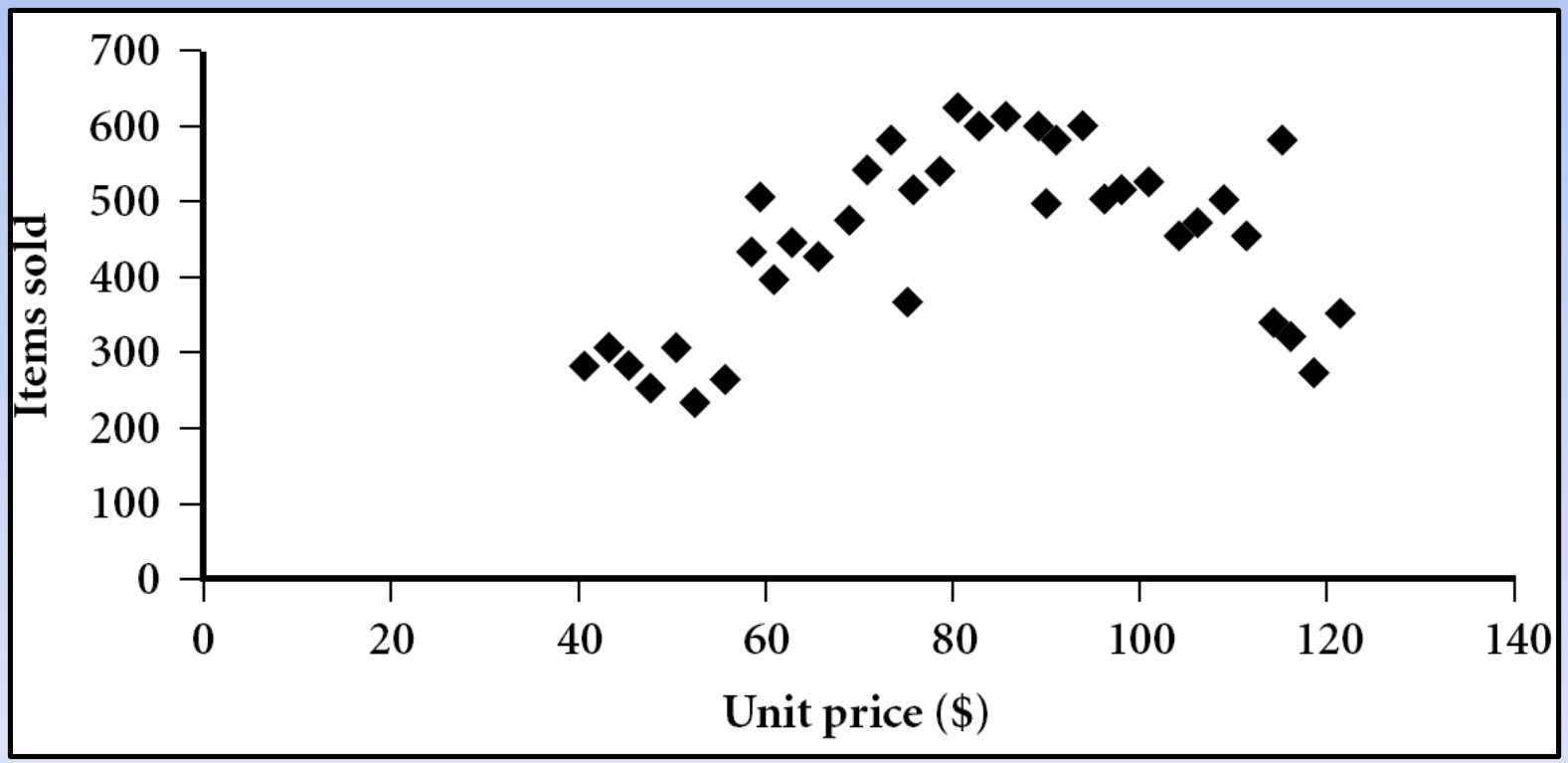
Quantile-Quantile (Q-Q) Plot

- Graphs the **quantiles of one univariate distribution** against **the quantiles of another distribution**
- **View**: Is there a **shift in going from one distribution to another?**
- **Ex:** *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.*

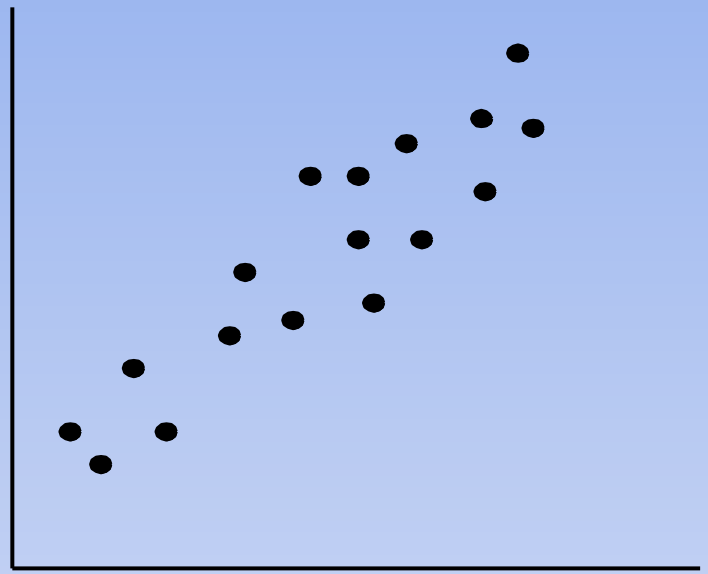


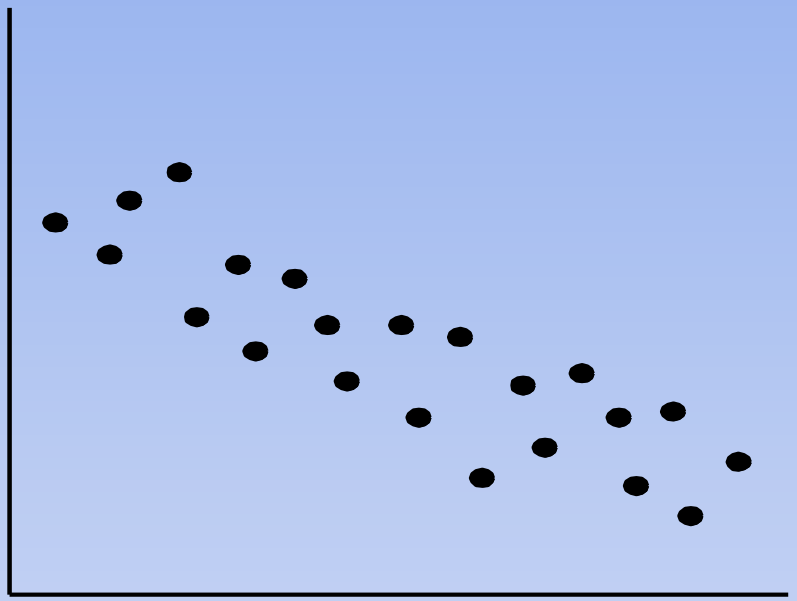
Scatter plot

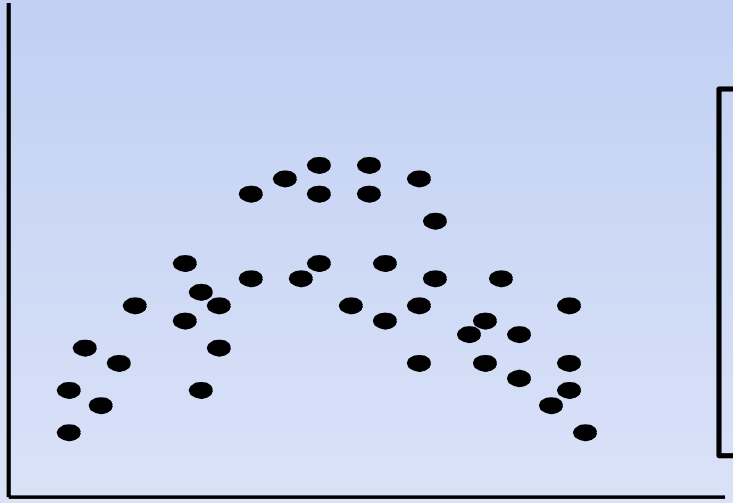
- Provides a look at bivariate data to see clusters of points, outliers etc
- Each pair of values is treated as a pair of coordinates & plotted as points in the plane



Positively and Negatively Correlated Data

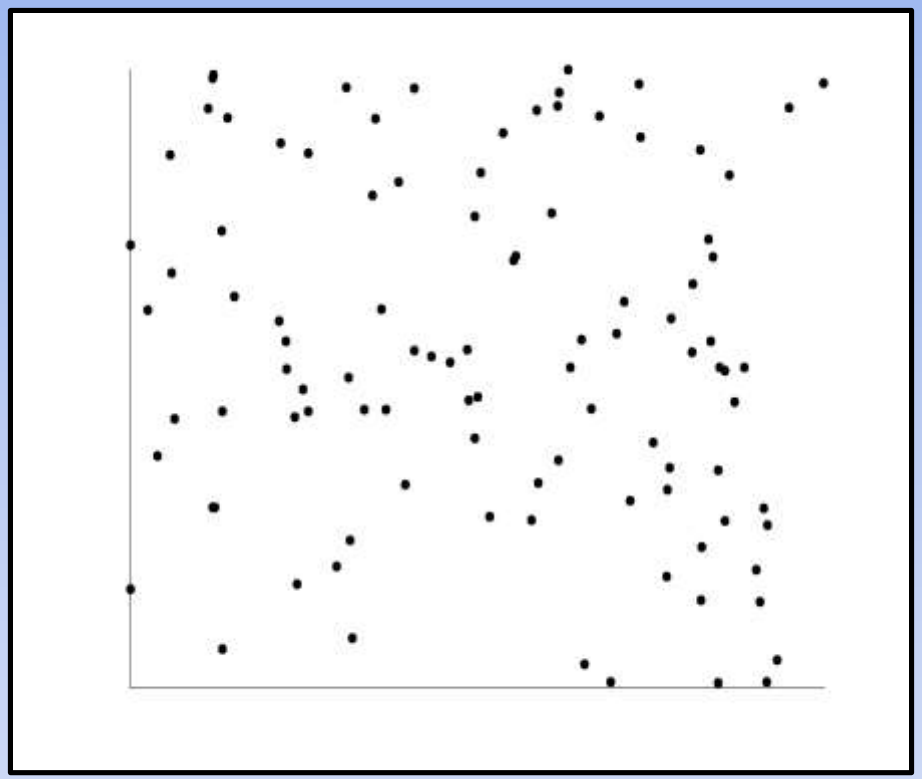


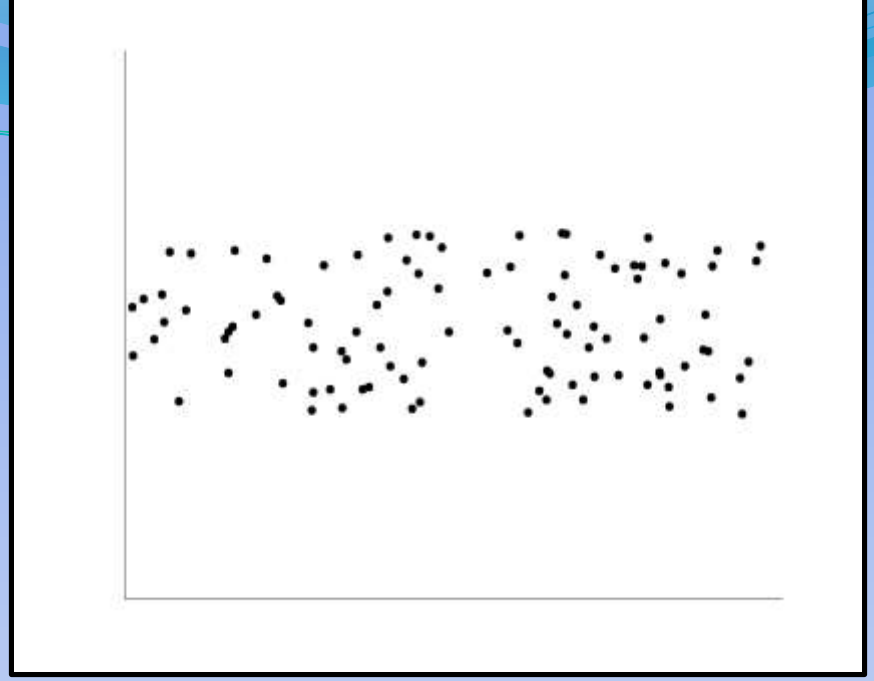


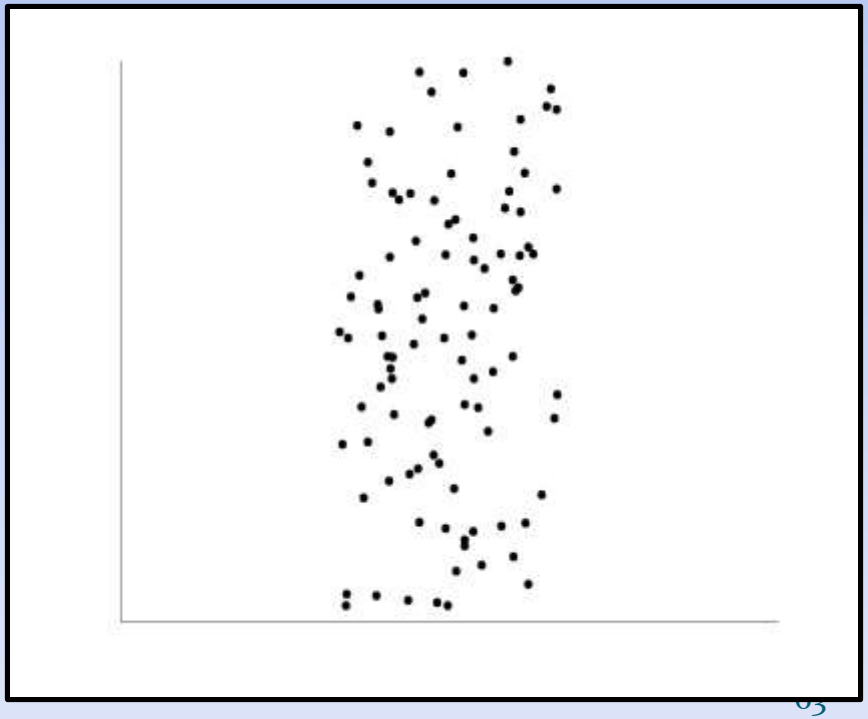


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data







Outlier Analysis

Higher Outline = Q3 + (1.5*IQR)

Lower Outline = Q1 - (1.5*IQR)

Inter-quartile range: IQR = $Q_3 - Q_1$

Example: Find higher and lower outliers of below distribution
8, 10, 11, 13, 14, 17, 20, 22, 24, 27, 29, 35, 49

Solution:

Step-1: (8, 10, **11, 13**, 14, 17), **20**, (22, 24, **27, 29**, 35, 49)

Step-2: Q2 = Median = 20

Step-3: Q1 = (11+13)/2 = 12 & Q3 = (27+29)/2 = 28 & IQR = Q3-Q1 = 28-12 = 16

Step-4: Lower Outlier <= 12 – 1.5*16 = **-12** & Higher Outlier >= 28 + 1.5*16 = **52**

Similarity and Dissimilarity

Similarity and Dissimilarity

- **Similarity**
  - **Numerical measure** of **how alike two data objects are**
  - **Higher value** means **objects are more alike**
  - The **range is [0,1]**

- **Dissimilarity** ***(e.g., distance)***
  - **Numerical measure** of **how different two data objects are**
  - **Lower value** means **objects are more alike**
  - **Minimum dissimilarity** is **0**, Upper limit varies

- **Proximity** refers to a **similarity** or **dissimilarity**

Data Matrix and Dissimilarity Matrix

- **Data matrix**
  - **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 data points**, but displays **only the distance**
  - A **triangular matrix**
  - **Single mode**

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

1. Proximity Measure for Nominal Attributes

- **Can take 2 or more states** *Ex: red, yellow, blue, green*

- **Method 1: Simple matching**
  - ***m*:** # of matches (same value in a column)
  - ***p*:** total # of variables (columns)

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

- Method 2: Use a large number of binary attributes
  - Creating a new binary attribute for each of the *M* nominal states

Nominal attribute

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

A Sample Data Table Containing Attributes of Mixed Type

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

Since here we have one nominal attribute, *test-1*, we set $p = 1$ in Eq. (2.11) so that $d(i, j)$ evaluates to 0 if objects i and j match, and 1 if the objects differ. Thus, we get

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

P = dimensions = No. of columns = 1

M = no. of matches in column 1 = 2

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

From this, we see that all objects are dissimilar except objects 1 and 4 (i.e., $d(4,1) = 0$).

2. Proximity Measure for Binary Attributes

Object *j*

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

- Computing **dissimilarity matrix** between two binary attributes
- Create 2x2 contingency table
- **q** – no. of attributes = 1 for both objects i & j
- **r** - no. of attributes = 1 for objects i & = 0 for object j
- **s** - no. of attributes = 0 for objects i & = 1 for object j
- **t** – no. of attributes = 0 for both objects i & j
- **p** – total no. of attributes = **q+r+s+t**
- **Dissimilarity between i & j**

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

- **Asymmetric binary Dissimilarity :**

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

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

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

Example

Dissimilarity between binary attributes. Suppose that a patient record table (Table 2.4) contains the attributes *name*, *gender*, *fever*, *cough*, *test-1*, *test-2*, *test-3*, and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary.

For asymmetric attribute values, let the values *Y* (*yes*) and *P* (*positive*) be set to 1, and the value *N* (*no* or *negative*) be set to 0. Suppose that the distance between objects (patients) is computed based only on the asymmetric attributes. According to Eq. (2.14), the distance between each pair of the three patients—Jack, Mary, and Jim—is

- **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 0**

q – no. of attributes = 1 for both objects i & j

r - no. of attributes = 1 for objects i & = 0 for object j

s - no. of attributes = 0 for objects i & = 1 for object j

t – no. of attributes = 0 for both objects i & j

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

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

3. Proximity Measure for Ordinal Attributes

- **Example**

A Sample Data Table Containing Attributes of Mixed Type

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

- **There are 3 states for *test-2*:** *fair, good, and excellent*, *i.e*, ***Mf = 3.***
- ***Step 1***: *Replace each value for test-2 by its rank*, *the* 4 objects are assigned **ranks 3, 1, 2 & 3** respectively
- ***Step 2*:** **Normalizes** the ranking by mapping rank **1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0.**
- ***Step 3*:** Measure **Manthan distance** as proximity measure

Step-2: Normalization

- Zif=Rf-1/mf-1
- Fair(1)=1-1/3-1=0
- Good(2)= 2-1/2=0.5
- Excellent(3)/3-1/3-1=1

Rf = 1,2,3
Mf = 3

id	test2
1	1
2	0
3	0.5
4	1

Step-3: Manhattan Distance

- **Manhantan distance=|x1-y1|+|x2-y2|**

	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1	0.5	0

4. Proximity Measure for Numeric Attributes

Three approaches:

1. Euclidean distance, 2.Manhattan distance & 3.Supremum distance

Distance on Numeric Data: Minkowski Distance

- ***Minkowski distance***: A popular distance measure

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

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

h=1 (mantattan distince), h=2 (euclidean distance)

h $\rightarrow \infty$ **(supermum distance) -** This is the max diff between any attribute of the objects

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

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

Properties

Non-negativity: $d(i, j) > 0$ if $i \neq j$,

Identity of indiscernibles: $d(i, i) = 0$

Symmetry: $d(i, j) = d(j, i)$ (Symmetry)

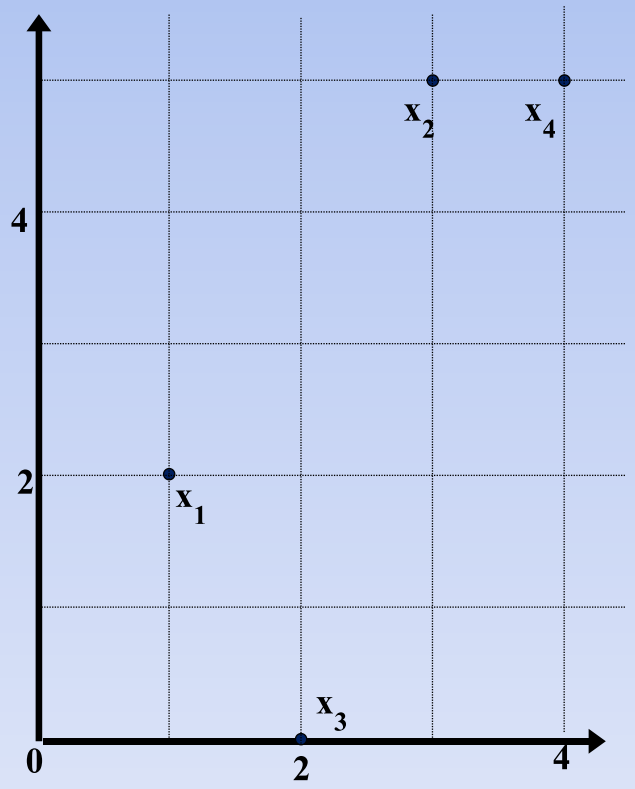
Triangle inequality: $d(i, j) \leq d(i, k) + d(k, j)$

- A distance that satisfies these properties is called a **metric**

Example: Minkowski Distance

Data Matrix

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



Dissimilarity Matrices

Manhattan (L_1) h=1 $\sqrt[h]{|x_{i1}-x_{j1}|^h+|x_{i2}-x_{j2}|^h}$ = |3−1|+|5−2|

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

Euclidean (L_2) h=2 $\sqrt[h]{|x_{i1}-x_{j1}|^h+|x_{i2}-x_{j2}|^h}$

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 h=infinity, p=2 $\lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if}-x_{jf}|^h\right)^{\frac{1}{h}}=\max_{f}^{p}|x_{if}-x_{jf}|$

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

(5) Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, binary, numeric, ordinal
- Then use below weighted formula to combine their effects

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

where the indicator $\delta_{ij}^{(f)} = 0$ if either (1) x_{if} or x_{jf} is missing

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

or = 0

- *f (attribute)* is binary or nominal:
  $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
- *f* is numeric: use the normalized distance
- *f* is ordinal
  - Compute ranks r_{if}
  - Treat z_{if} as numeric

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

Already done

Example: Distance between attributes of Mixed Type

A Sample Data Table Containing Attributes of Mixed Type

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

For Test-1

	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

For Test-2

	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1	0.5	0

Example: Distance between attributes of Mixed Type

For Test-3

A Sample Data Table Containing Attributes of Mixed Type

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

1. Normalize data Between 0 & 1

D(2,1)= |45-22|/64(max)-22(min) = 23/42 = .55

D(3,1)= |64-45|/64-22 = 19/42 = 0.45

0			
0.55	0		
0.45	1.00	0	
0.40	0.14	0.86	0

2. Any missing value put 0, otherwise put 1

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

3. Mixed attribute

D(2,1)=(1*1)+(1*1)+(1*0.55)/1+1+1=0.85

	1	2	3	4
1	0			
2	0.85	0		
3	0.65	0.83	0	
4	0.13	0.71	0.79	0

5. Similarity between binary vectors

- Common situation is that objects, *p* and *q*, have only binary attributes

- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes
  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values
  $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

6. Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the ***frequency*** of a particular word *(such as 'win')*

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

- **Other vector object:** *gene features in micro-arrays*
- **Applications**: info. retrieval, biologic taxonomy, gene feature mapping
- **Cosine measure:** If d_1 & d_2 are two vectors *(Ex: term-frequency vectors)*, then

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

where • indicates vector dot product,

$||d||$: the length of vector d

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$,
  where $\bullet$ indicates vector dot product, $||d|$: the length of vector d

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.

$\boldsymbol{d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)}$

$\boldsymbol{d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)}$

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

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

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

$\boldsymbol{\cos(d_1, d_2) = 25/(6.48*4.12) = 0.94}$

Data Preprocessing

- There are many factors comprising **data quality** such as:

1. **Accuracy**
2. **Completeness**
3. **Consistency**
4. **Timeliness**
5. **Believability**
6. **Interpretability**

Why Preprocess the Data?

- Today's real-world databases are highly susceptible to **noisy, missing & inconsistent data** due to their huge size & likely origin from multiple heterogeneous sources
- **Low quality data** will lead to **low quality mining results**
- **Data Preprocessing techniques** can improve data quality
- **Data in the real world is dirty**
  - **incomplete**: lack attribute values or in aggregated form.
    Ex: occupation=" "
  - **Inaccurate or noisy**: contain errors or outliers i.e. values that deviate from the expected. **Ex: Salary="-10"**
  - **inconsistent**: contain discrepancies in codes or names
    - **Ex: Age="42" Birthday="03/07/1997"**
    - **Ex: Was rating "1,2,3", now rating "A, B, C"**
    - **Ex: Discrepancy between duplicate records**

Why Data Is Dirty?

- **Incomplete data** may come from
  - "Not applicable" data value when collected
  - Attributes of interest may not always be available
  - The time delay between the data collection & data analysis
  - Human/hardware/software problems/ equipment malfunctions
- **Inaccurate or Noisy data** (incorrect attribute values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
  - Users submitted incorrect data values for mandatory fields
  - There may be technology limitations such as limited buffer size

Why Is Data Dirty?

- **Inconsistent data** came from
  - Different data sources
  - Functional dependency violation (Ex: modify some linked data)
- **Duplicate records** also need data cleaning
- **Timeliness** also affects data quality
- Two other factors affecting data quality are: **believability** & **interpretability**. **Believability** reflects how much the data are trusted by users, while **interpretability** reflects how easily the data are understood

Why Is Data Preprocessing Important?

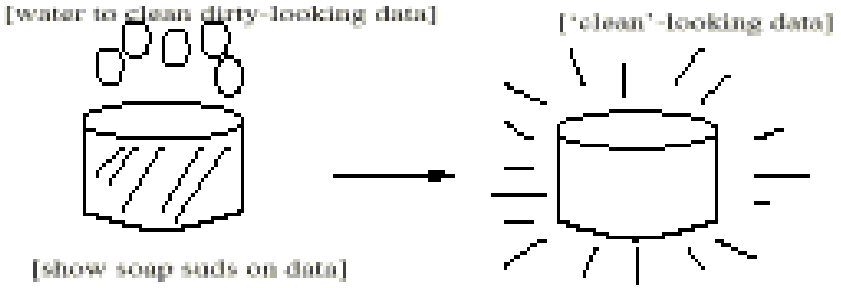
- **No quality data, no quality mining results**
  - Quality decisions must be based on quality data
    - Ex: duplicate or missing data may cause incorrect or misleading statistics
  - DWH needs consistent integration of quality data
- **Data Preprocessing techniques** can improve data quality, thereby helping to improve the accuracy & efficiency of the **mining process**
- **Data Preprocessing** is an important step in the knowledge discovery process

Major Tasks in Data Preprocessing

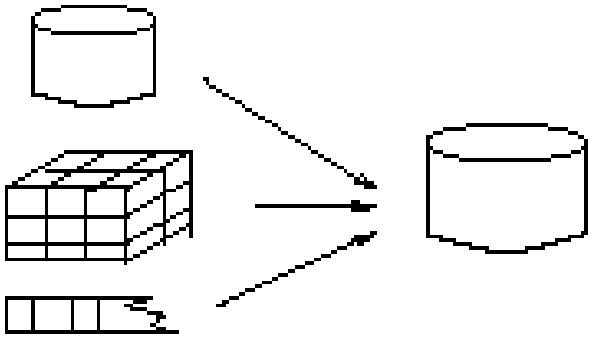
- **Data cleaning**
  - Filling in missing values, smooth noisy data, identify or remove outliers & resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes or files
- **Data reduction**
  - Reduction in volume of the dataset without affecting the analytical results
- **Data transformation**
  - Normalization & aggregation
  - Discretization
  - Hierarchy generation

Forms of Data Preprocessing

Data Cleaning



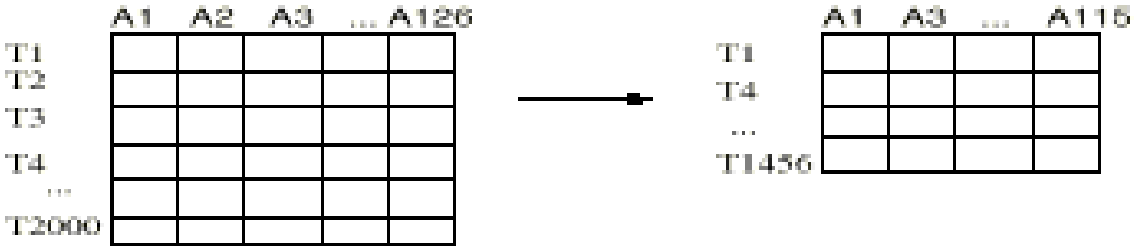
Data Integration



Data Transformation

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

Data Reduction



Data Cleaning

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

Missing values

- 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 & 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

How to handle Missing values?

- ***Ignore the tuple:***

- This is usually done **when class label is missing** (assuming the tasks in classification)
- This method is **not very effective**, unless the tuple contains several attributes with missing values
- It is not effective when the % of missing values per attribute varies considerably
- By ignoring the tuple, we do not make use of the remaining attribute's values in the tuple.

How to handle Missing values Cont…

- ***Fill in the missing value manually:***

➢ This approach is time consuming when the data set is large with many missing values

- ***Use a global constant to fill in the missing value:***

➢ Replace all missing attribute values by a **constant** such as a label like “Unknown” or “ infinite” .

- ***Use a measure of central tendency for the attribute to fill in the missing value:***

➢ For **normal** data distributions, **mean** can be used, while for **skewed** data distribution, **median** can be used.

How to handle Missing values Cont…

- ***Use the attribute mean or median for all samples belonging to the same class***
- ***Use the most probable value to fill in the missing value:***

➢ This is the **most popular strategy.**

➢ This may be determined with **regression**, **Bayesian formalism**, or **decision tree induction**

Noisy Data

- **Noise:** Noise is a random error or variance in a measured variable

- **Incorrect attribute values may be due to :**
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

- **Other data problems which requires data cleaning**
  - incomplete data
  - inconsistent data
  - duplicate records

How to handle noisy data?

- We can "**smooth**" out the data to remove the noise
- **Different data smoothing techniques are:**

➢ ***Binning***

- First **sort data** & **partition into (equal-frequency) bins**
- **Smooth** by **bin means**, **bin median** & **bin boundaries**..

➢ ***Regression***

- Smooth by **fitting the data into regression functions**

➢ ***Clustering***

- Detect & remove **outliers**

Binning

- **Binning methods smooth a sorted data value** by consulting its neighborhood (the values around it)
- The sorted values are distributed into a **set of equal-freq bins**
- In ***smoothing by bin means,*** each bin is replaced by the mean value of the bin
- Similarly in ***smoothing by bin medians,*** each bin is replaced by the bin median
- In ***smoothing by bin boundaries***, the min & max values in a given bin are identified & Each bin value is then replaced by the **closest boundary value**

Example

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

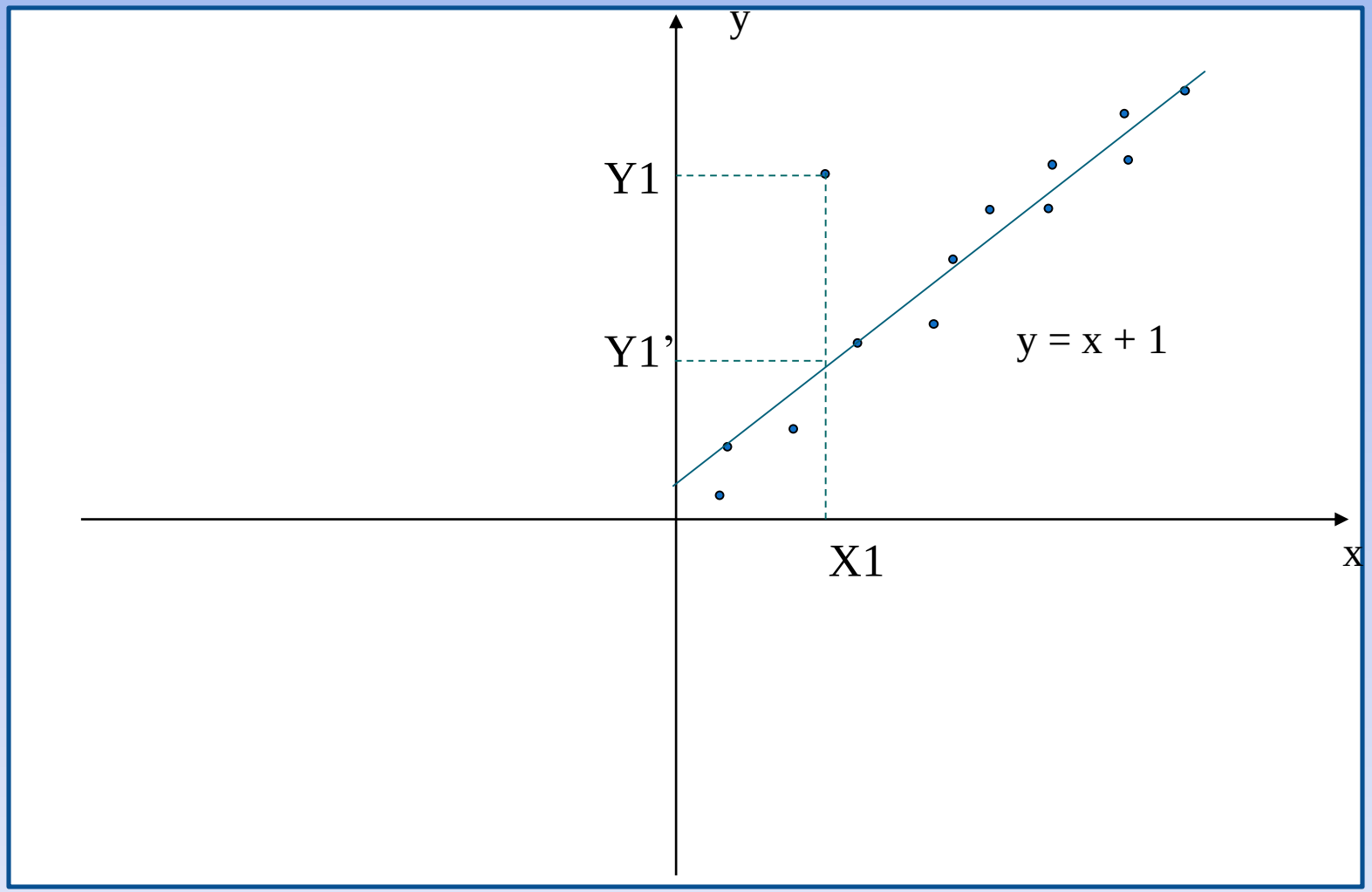
Smoothing by bin boundaries:

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

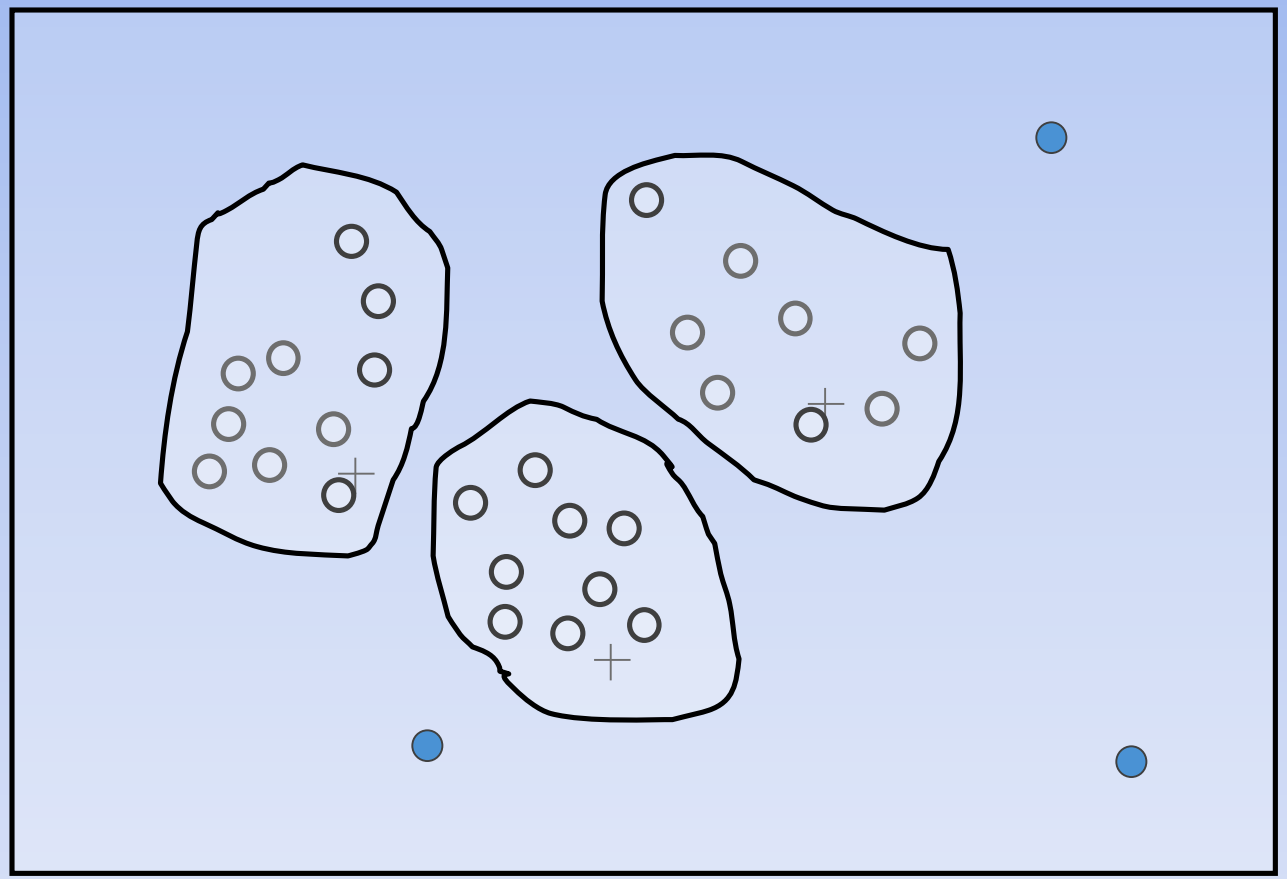
Regression

Linear regression involves finding the "**best**" line to fit two attributes so that one attribute can be used to predict the other.

Multiple linear regression is an extension of linear regression, where more than 2 attributes are involved and the data are fit to a multidimensional surface.



Cluster & Outlier Analysis



Data Cleaning as a Process

- **Data discrepancy detection**
  - Use **metadata** **(Ex: domain, range, dependency, distribution)**
  - Check **field overloading** **(2 attributes combined in one)**
  - Check **uniqueness rule, consecutive rule** & **null rule**
  - Use **commercial tools**
    - **Data scrubbing:** use simple domain knowledge **(Ex: postal code, spell-check)** to detect errors & make corrections
    - **Data auditing**: by analyzing data to discover rules & relationship to detect violators **(Ex: correlation & clustering to find outliers)**
- **Data migration & integration**
  - **Data migration tools:** allow **transformations** to be specified
  - **ETL** (Extraction/Transformation/Loading) tools: allow users to specify **transformations through a GUI**
- **Integration of the two processes**
  - **Iterative** & **interactive**

Data Integration

- Combines data from **multiple sources into a coherent data store**
- **Careful integration** avoids **redundancies** & **inconsistencies** in the resulting data set
- This can help improve accuracy & speed of subsequent DM process
- There are a no. of issues to consider such as **Schema integration & object matching, integration of metadata from different sources**
- Schematic heterogeneity & structure of data are the major challenges in data integration

Entity Identification Problem

- The essence of *entity identification problem* is "***how to match schema & objects from different sources?***"
- *Entity identification problem* addresses the issue of "***How can equivalent entities from multiple data sources be matched up?"***
- **Ex:** *How can the data analyst or computer be sure that* ***customer_id*** *in one database &* ***cust_number*** *in another database refer to the same attribute?*
- **Metadata** can be used to avoid errors in schema integration
- **Metadata** may also be used to **help transform the data**

Entity Identification Problem Cont..

- When matching attributes from one database to another during integration, **attention has to be paid to the structure of data**
- This ensures any **functional dependencies & referential constraints** in the source system match with those in the target system

Redundancy & Correlation Analysis

- **Redundancy** is another important issue in data integration
- **Inconsistencies** in **attribute** can also cause **redundancies in the resulting data set**
- **Redundancies** can be detected by ***correlation analysis***
- **Given 2 attributes. *Correlation analysis* can measure how strongly one attribute implies the other**
- For **nominal data**, $\mathbf{X^2}$ (chi-square) test & for **numeric attributes correlation coefficient** & **covariance** methods are used

Correlation Coefficient for Numeric Data

- For numeric attributes , we can evaluate the **correlation between two attributes, A & B** by computing the ***correlation coefficient*** ***(***also known as ***Pearson's product moment coefficient)***

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

Correlation Coefficient Cont…

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

- Here **N** is the no. of tuples
- a_i **&** b_i are the respective values of **A & B** in **tuple i**
- **Abar & Bbar** are the respective **means** of **A & B,** σ_A & σ_B are the respective **standard deviation** of A & B & $\Sigma a_i b_i$ is the sum of the AB cross-product *(i.e. for each tuple the value for A is multiplied by the value for B in that tuple)*
- If $r_{A,B} > 0$, then A & B are positively correlated (A's values increase as B's). The higher the value, the stronger the correlation
- If $r_{A,B} = 0$: then A & B are independent
- If $r_{A,B} < 0$: then A & B are negatively correlated

Chi-Square Normalization for Nominal data

- For **nominal data**, a correlation relationship between **2 attributes, A and B**, can be found by a **χ 2 (chi-square) test**

- **X^2 (chi-square) test**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The **larger the X^2 value**, the more likely the variables are related

- When actual count is very different from the expected count it contributes the most to the X^2 value

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- **X^2 (chi-square)** calculation *(no. in parenthesis are **expected counts** for both 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 & play_chess are **correlated** in the group

Covariance (Numeric Data)

- **Covariance is similar to correlation**

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

Correlation coefficient:

- where n is the number of tuples, $\bar{A}$ $\bar{B}$ are the respective means, σ_A & σ_B are the respective standard deviation of A & B

- **Positive covariance**: If $\mathbf{Cov_{A,B} > 0}$, then A & B both tend to be **larger** than their expected values

- **Negative covariance**: If $\mathbf{Cov_{A,B} < 0}$ then if **A** is **larger** than its expected value, **B** is **smaller** than its expected value

- **Independence**: $\mathbf{CovA,B = 0}$ A & B are **independent**

Co-Variance: An Example

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose **two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).**
- **Question**: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - $\bar{A}$ = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - $\bar{B}$ = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - **Cov(A,B)** = (2×5+3×8+5×10+4×11+6×14)/5 – 4 × 9.6 = 4
- **Thus, A and B rise together since Cov(A, B) > 0.**

Co-Variance: Example-2

- **Given Table presents a example of stock prices observed at five time points for AllElectronics and HighTech. If the stocks are affected by the same industry trends, will their prices rise or fall together?**

- **Answer**

Stock Prices for *AllElectronics* and *HighTech*

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80$$

$$Cov(AllElectroncis, HighTech) = \frac{6\times 20+5\times 10+4\times 14+3\times 5+2\times 5}{5} - 4\times 10.80$$

$$= 50.2 - 43.2 = 7.$$

$$Cov(A,B) = E(A\cdot B) - \bar{A}\bar{B}$$

- Therefore, given the positive covariance we can say **that stock prices for both companies rise together.**

Covariance Vs. Correlation

- Covariance can take on practically any number while a correlation is limited: -1 to +1.
- Covariance does not tell us the intensity of the co-movement of the variables, only the direction.
- Because covariance's measures variables of different units it has numerical limitations, correlation is more useful for determining **how strong** the relationship is between the two variables.
- Correlation isn't affected by changes in the center (i.e. mean) or scale of the variables.
- We can standardize the covariance however and calculate the correlation coefficient which will tell us not only the direction but provides a scale to estimate the degree to which the variables move together.

Data Transformation

- In **data transformation**, data are **transformed or consolidated** into **forms appropriate for mining**
- **Strategies for data transformation include the followings:**
- **Smoothing** : to **remove noise** from the data
- **Attribute construction** : **New attributes are constructed & added** to help the **mining process**
- **Aggregation** : **summary/aggregation** operations are applied to data
- **Normalization** : the attributes data are **scaled** so as to **fall within a smaller range**, such as [-1.0 to 1.0] or [0.0 to 1.0]
- **Discretization** : the raw values of a numeric attribute (like age) are replaced by interval labels (e.g. 0-10, 11-20 etc.) or conceptual labels (e.g. youth, adult, senior). These labels can recursively be organized into higher level concepts called concept hierarchy

Data Transformation by Normalization

- **Normalization** attempts to give all attributes an equal weight
- Particularly useful for **classification algorithms**
- **Methods are:**
  - **Min-max normalization**
  - **Z-score normalization**
  - **Normalization by decimal scaling**

Min-Max Normalization

- Performs a **linear transformation** on the **original data**
- Suppose that $\mathbf{min_A}$ **&** $\mathbf{max_A}$ *are the minimum & maximum values of an attribute, A*
- **Min-max normalization** maps a value, ***v of A to v'*** *in the range* $[new\_\min_A , new\_\max_A]$ by computing
-

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

Ex:- Let the min & max values of attribute "income" are 12000 & 98000 respectively. We want to map 73600 income to the range [0.0,1.0]

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Z-score Normalization

- Also known as ***zero-mean normalization***
- *The values for an attribute A, are normalized based on the mean & standard deviation of A*
- *A value* ***v*** *of A is* normalized to ***v'*** *by computing*

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

Ex: Suppose that the mean $\bar{A}$ & standard deviation σ_A of the values for the attribute *income are $54,000 & $16,000, respectively.* With z-score normalization, a value of $73,600 for *income is transformed to*

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Decimal Scaling Normalization

- Normalizes by moving the decimal point of values of attribute *A*
- *The number of decimal points moved depends on the maximum absolute* value of ***v'***
- *A value, **v**, of A is normalized to **v'** by computing*

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

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

Decimal Scaling Normalization Cont..

- Suppose that the recorded values of *A range from -986 to 917*
- The maximum absolute value of *A,* ***v'*** *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 & 917 normalizes* to 0.917.

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is smaller in volume but produces the same analytical results
- A **DWH** may store terabytes of data, **Complex data analysis** may take a very long time to run on the complete data set
- **Data reduction strategies:**
  - **Dimensionality reduction**, *Ex: remove unimportant attributes*
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - **Feature subset selection, feature creation**
  - **Numerosity reduction** *(Ex: Data Reduction)*
    - Regression and Log-Linear Models
    - Histograms, **clustering**, **sampling**
    - Data cube **aggregation**
  - **Data compression**

Sampling

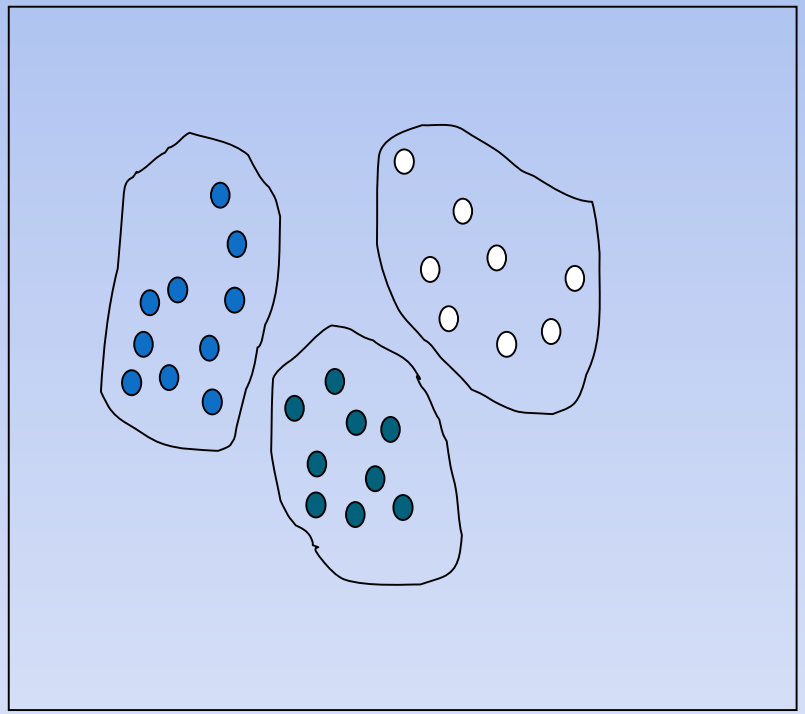
- **Sampling**: obtaining a small sample *s* to represent the whole data set *N*
- Allow a **mining algorithm** to run in low complexity
- **Key principle:** Choose a **representative subset** of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop **adaptive sampling methods**, **Ex:** stratified sampling

Types of Sampling

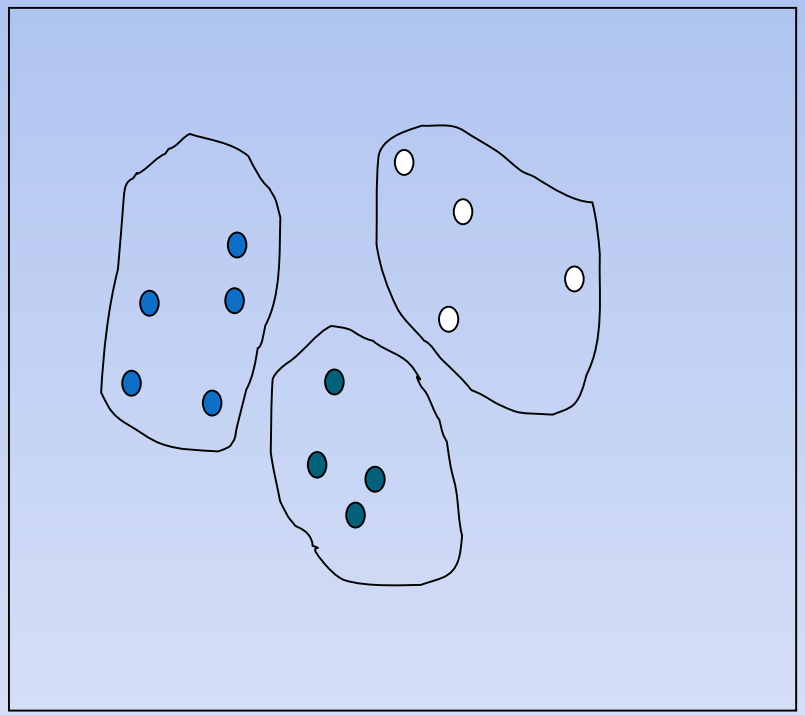
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - **Partition the data set** & draw samples from each partition (proportionally, the same % of the data)
  - Used in conjunction with skewed data

Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample




THAT'S
ALL FOLKS