



DATA MINING

By

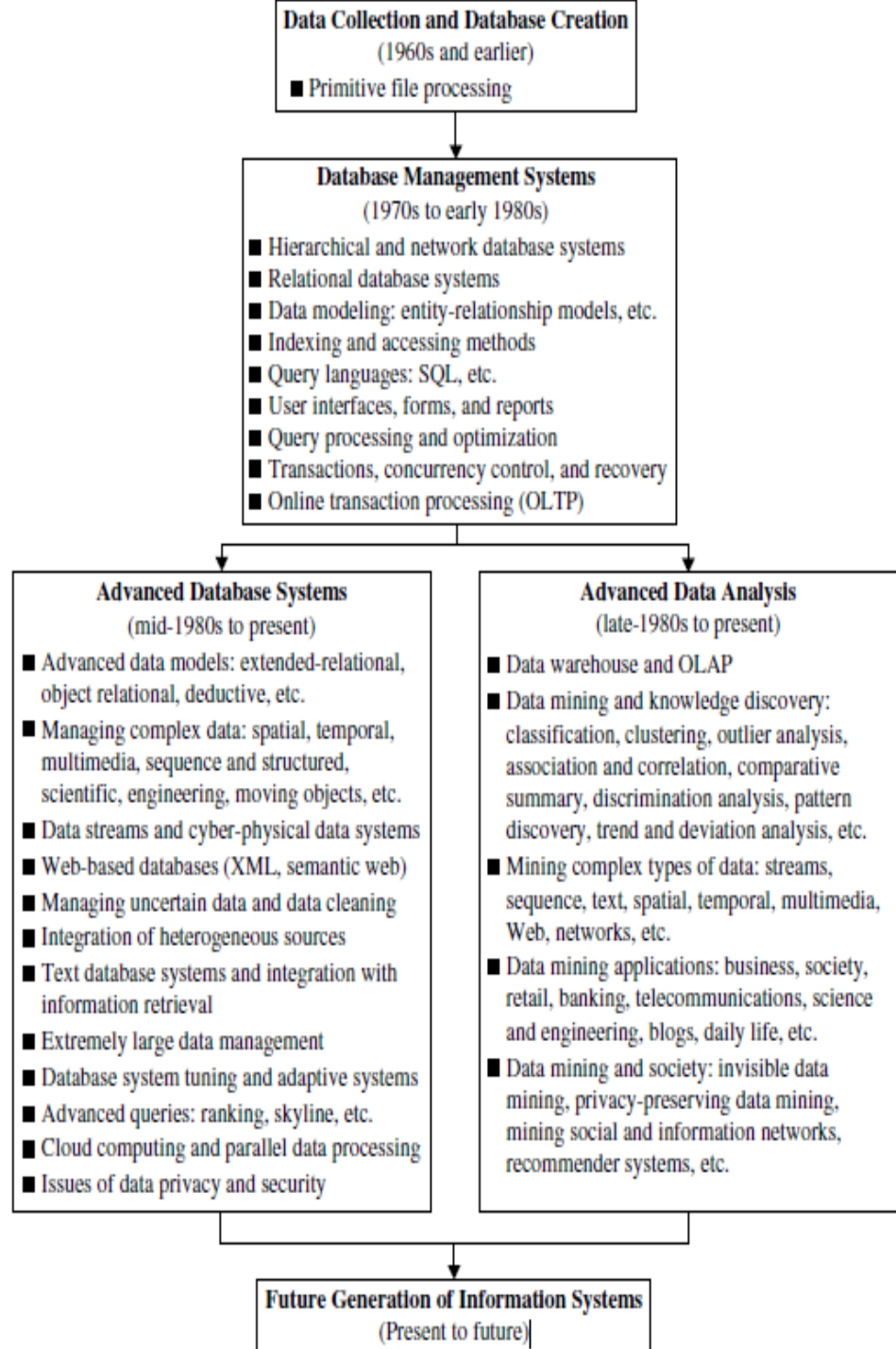
R. Siva Narayana

RGUKT Nuzvid

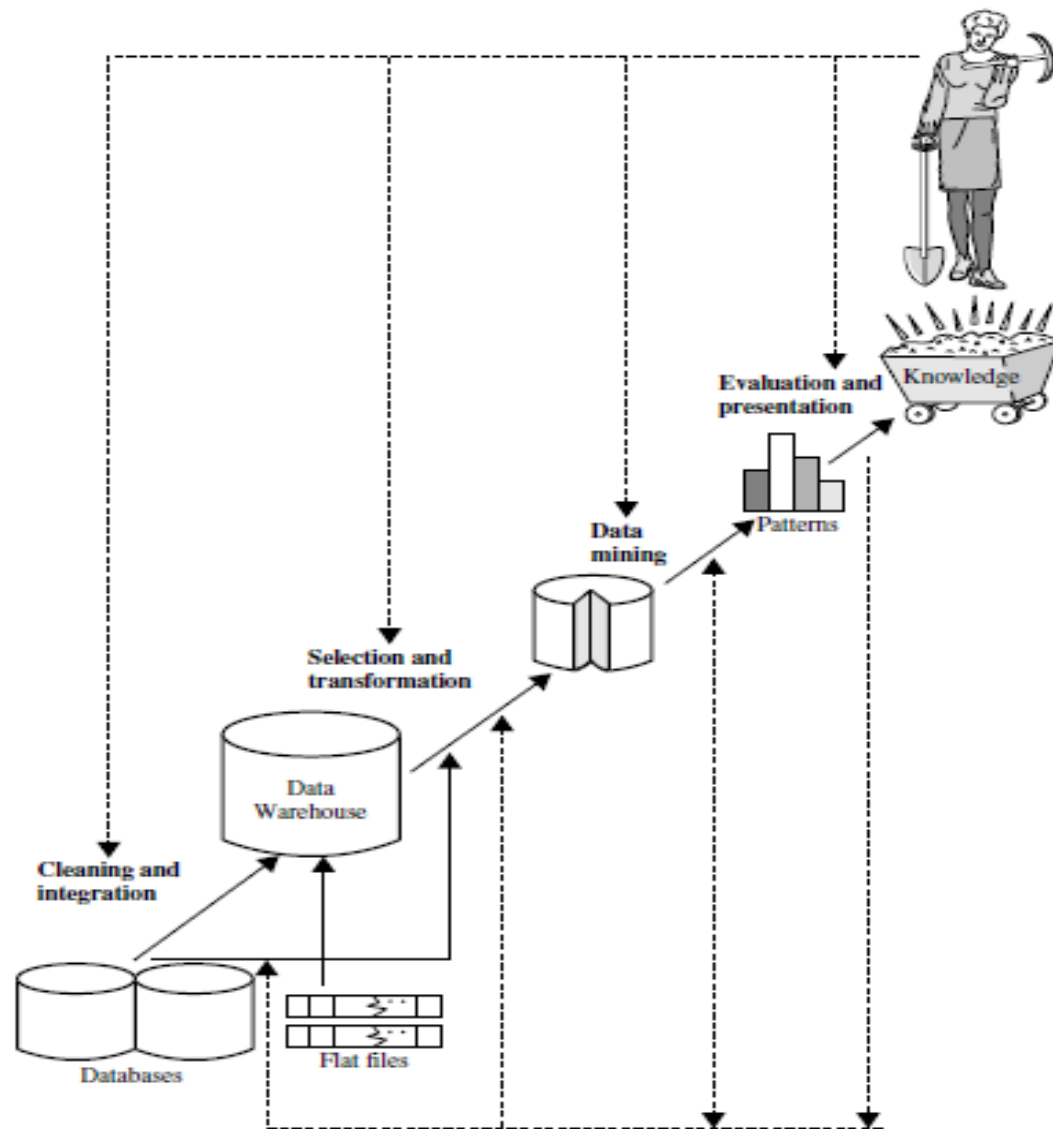
Data Mining

- **Data mining** is the *process* of discovering interesting patterns and knowledge from *large* amounts of data.
 - The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Data Mining Evolution



KDD



What kinds of data can be mined?

- Database data
- Data warehouses
- Transactional data
- Other kinds of data
 - Time-related or sequence data(stock-exchange)
 - Data-streams(Video surveillance and sensor data)
 - Spatial data(maps)
 - Hypertext and multimedia

Data Objects and Attribute Types

- Datasets are made up of data objects
- Data objects are typically describes by attributes
- Attribute is a field, representing a characteristic or feature of data object
- Attribute, dimension, feature and variable
- Type of an attribute can be determined by the set of possible values. We have
 - Nominal
 - Binary
 - Ordinal
 - Numeric

Basic Statistical Description of Data

- Used to **identify properties** of the data
- Identify which data values should be treated as **Noise or Outliers**
- **Measuring Central Tendency** – Measures the location of the middle or center of a data distribution
- **Measuring the Dispersion of Data** – How data are spread out? Used to identify outliers.
- **Graphic Displays** – Visually inspect our data.

Measuring of central tendency (Mean, Median, Mode and Midrange)

- Attribute $X(\text{salary})$, which has been recorded for a set of objects.
- $x_1, x_2, x_3, \dots, x_N$ be the set of N observations for X .
- Mean: Most common effective numeric measure for measuring the CENTER of the data
- Built in aggregate function, average (`avg()` in SQL) in RDBMS

Weighted arithmetic mean or Weighted average

- Sometimes, each value in a set may be associated with a weight for $i=1, \dots, N$
- The weight reflect the significance, importance, or occurrence frequency attached to their respective values.

Problems and Solution

- Problem with mean is its sensitivity to extreme values either large or small (Outliers)
 - **Ex:** Mean score of a class in an exam is decreased by a few very low scores.
- To offset the effect, we use **Trimmed Mean**
- Trimmed Mean is obtained after chopping off (2%) values at the high and low extremes.

Median

- For skewed(Asymmetric) data, median is the better measure for center of data
- Median is the middle value in a set of ordered data values.
- It separates the data set into two halves.
- If N is odd, then the median is the middle value of the ordered set
- If N is even, then the median is the average of middle two values.

Cont.,

- If data are grouped in the intervals and the frequency is known, then

$$\text{median} = L_1 + \left(\frac{N/2 - (\sum \text{freq})l}{\text{freq}_{\text{median}}} \right) \text{width}$$

- L_1 is the lower boundary of the median interval
- N is the number of values in the entire dataset
- $(\sum \text{freq})l$ is the sum of frequencies of all the intervals that are lower than the median interval
- $\text{freq}_{\text{median}}$ is the frequency of the median interval
- Width is the width of the median interval

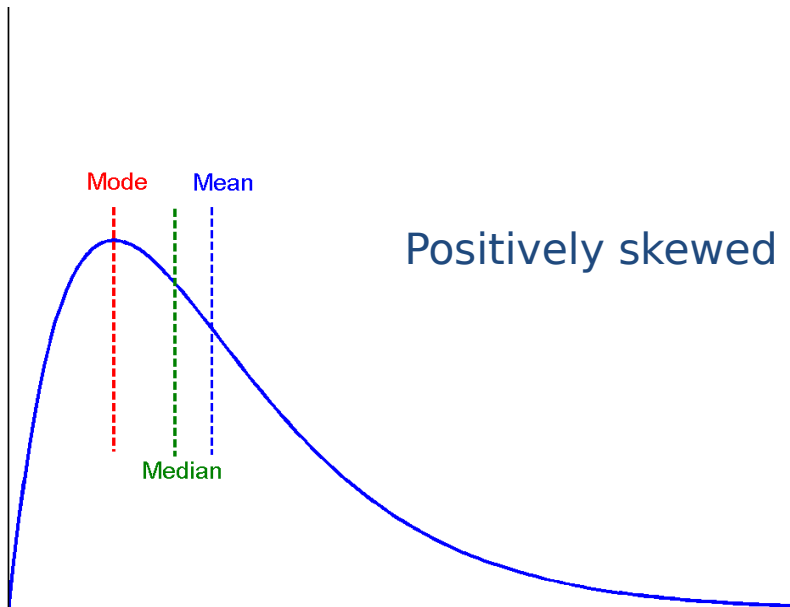
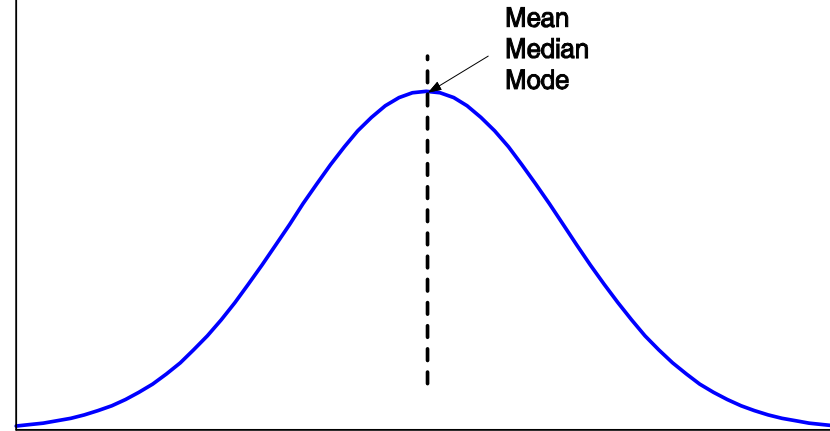
Mode and Midrange

- Most frequency occurred value in the set
- Determined for both qualitative and quantitative attributes
- Data set with one, two or more modes called Unimode, Bimode and Trimode
- $\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$
- Unimode empirical formula
- Midrange is the avg of Min + Max

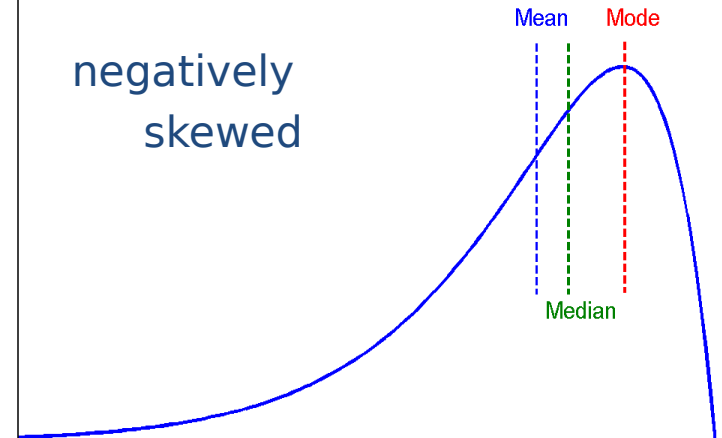
Symmetric vs Skewed data

- Symmetric data distribution- mean, median and mode all are the same center value or nearly to the center.

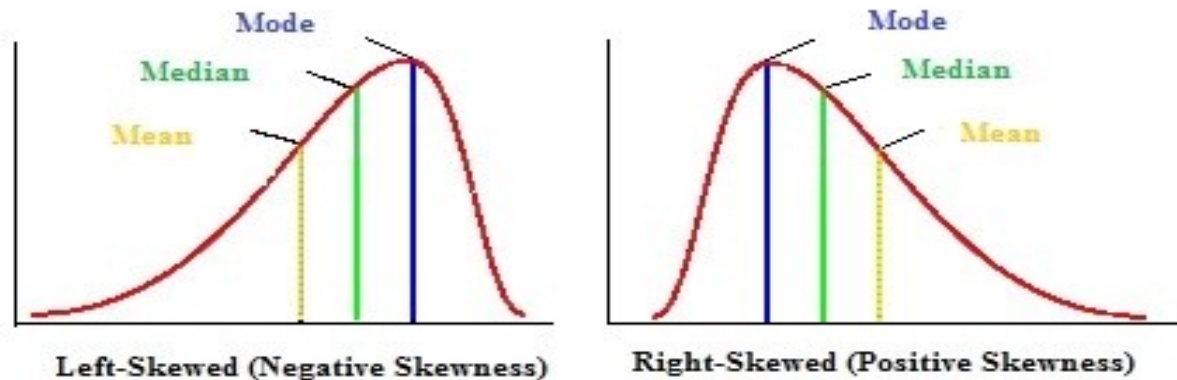
Symmetric data



negatively skewed



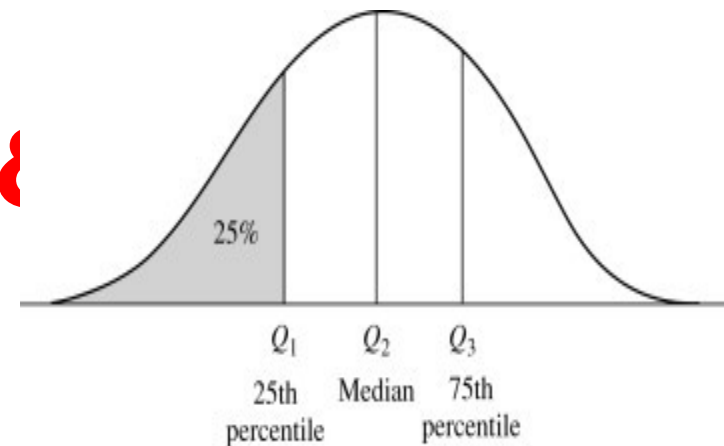
- Positively skewed- mode occurs at a value that is smaller than the median
- Negatively skewed- mode occurs at a value greater than the median



Dispersion of Data

- Spread of numeric data distribution.
- Range, Quartiles and Interquartile Range
- Five-Number summary, Boxplots and Outliers
- Variance and Standard Deviation

Range, Quartiles &



- Range: Difference between the largest and smallest values.
- Quartiles: Data points that split the data distribution into equal-size consecutive sets.
- 1-quartile is the data point dividing the lower and upper halves of the data distribution
- 4-quartiles are the 3 data points that split the data distribution into 4 equal parts
- Each equal part commonly referred to as quartiles.
- 100 quartiles commonly referred to as Percentiles
- Interquartile range (IQR)= $Q_3 - Q_1$

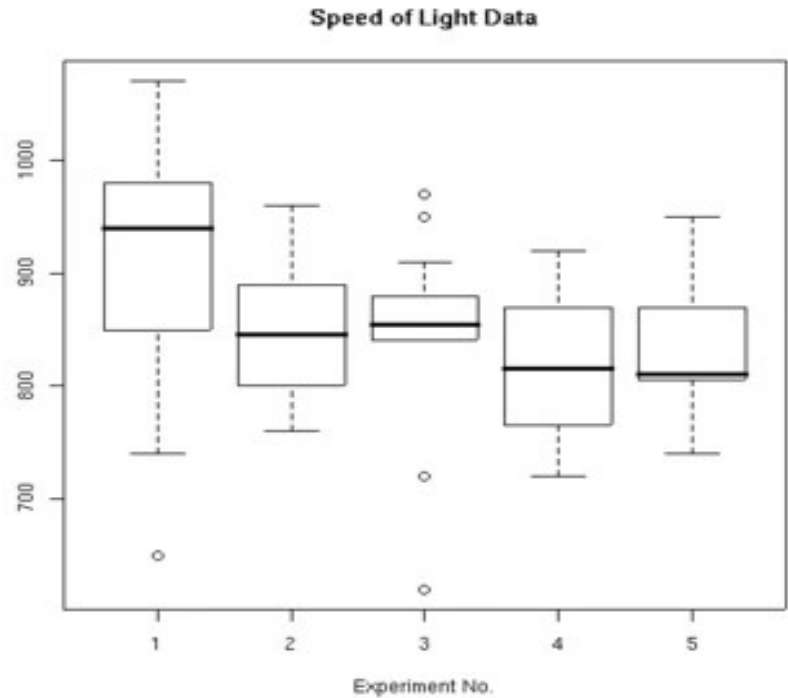
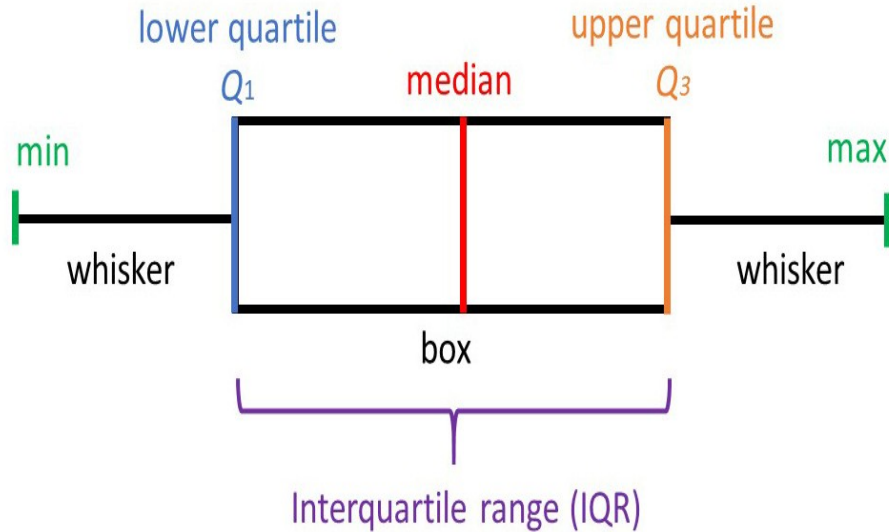
Five-Number Summary and Boxplots

- In Symmetric distribution, the median splits the data into equal-size halves.
- In Skewed distribution, it having two quartiles Q1 and Q3 along with median distribution.
- To **identifying suspected OUTLIERS** is to finding the values falling at least $1.5 \times IQR$ above the third quartile or below the first quartile.

Five-Number Summary and Boxplots, Cont.,

- Five-number summary of distribution consists
 - Minimum
 - Quartile Q1(25th Quantile)
 - Median Q2
 - Quartile Q3(75th quantile)
 - Maximum
- Boxplots are popular way of visualizing a distribution. It incorporates the five-number summary.

Boxplot representation



Variance and Standard Deviation

- Both indicate how data spread or distributed.
- Low standard deviation means that the data observations tend to very close to mean
- High standard deviation indicates that the data are spread out over a large range of values.
- Variance:

Population Variance Formula



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

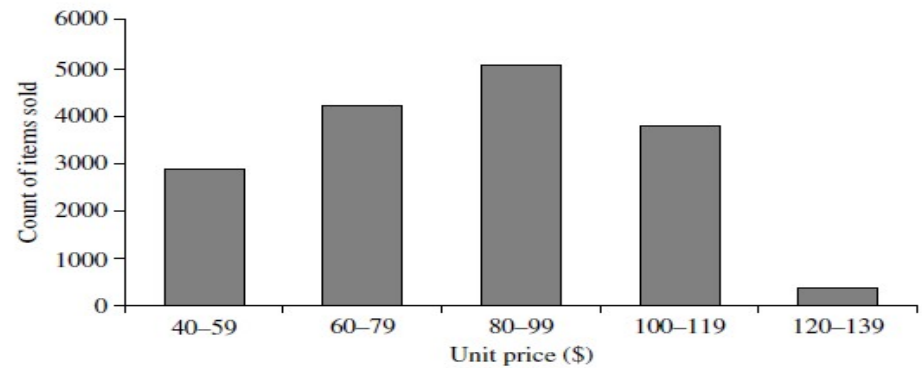


- Standard Deviation:

Graphic Displays

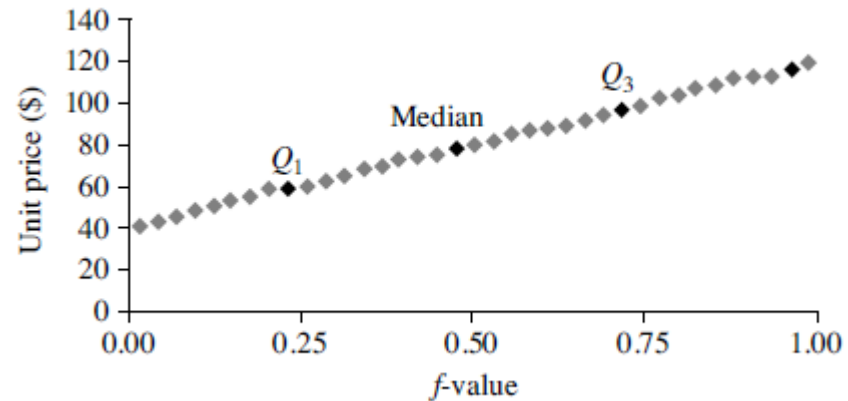
- Graphs are helpful to visual description of data, which is useful in data preprocessing by identifying the noise and outliers
 - Histograms
 - Quantile Plot
 - Quantile-Quantile plot
 - Scatter Plots

Histograms



- Histo means pole, gram means chart, so histogram is a chart of poles
- Histogram is used to summarizing the given attribute X
- Height of the bar indicates the frequency(count) of the attribute X value
- Numeric attributes are preferred for plot the histograms
- The range of values for X is partitioned into disjoint consecutive equal subranges called buckets or bins
- Range of bin is known as width

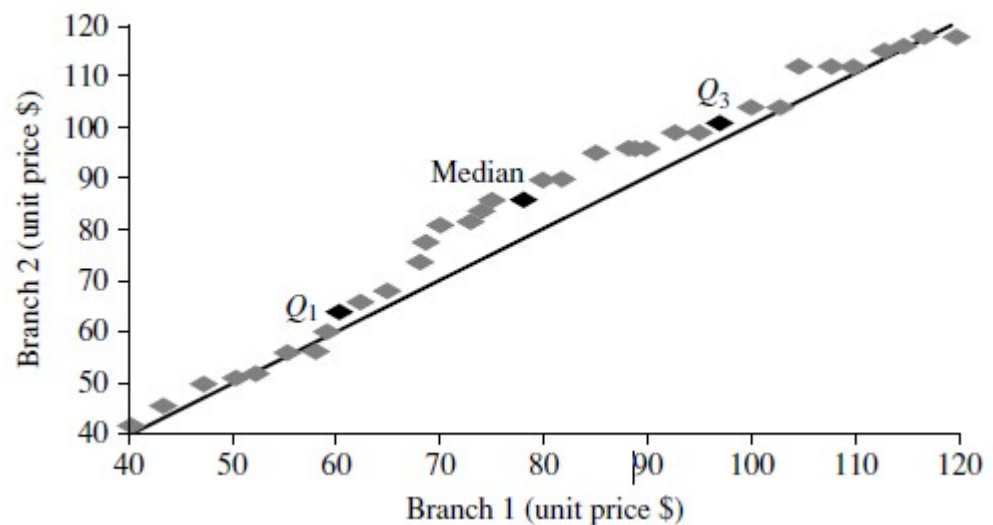
Quantile Plot



- A simple and effective way to have a first look at a one variable data distribution.
- It displays all of the data for the given attribute
- It plots quantile information
- Each observation, is paired with a percentage which indicates that approximately % of the data are below the value .
- =

q-q plot

- It graphs the quantiles of one univariate distribution against another.
- If $M=N$, then plot the quantile plot
- If $M < N$, then q plot



Scatter Plot

- To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.
- Provides first look at bivariate data to see clusters and outliers or to explore the possibility of correlations.

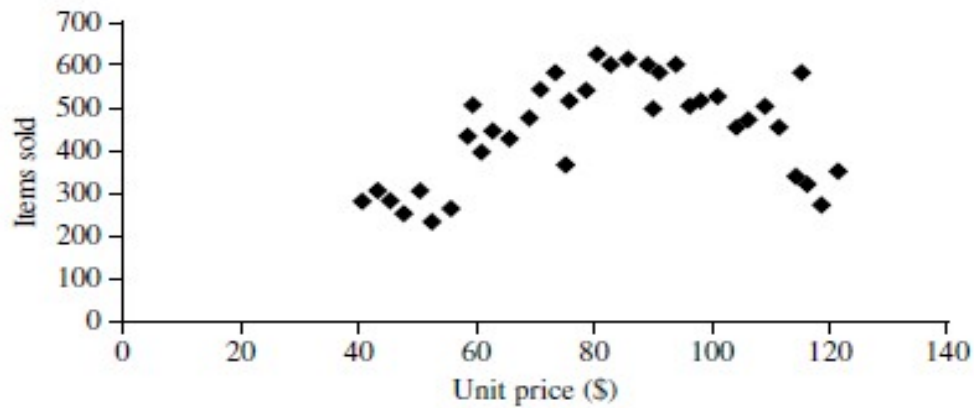


Figure 2.7 A scatter plot for the Table 2.1 data set.

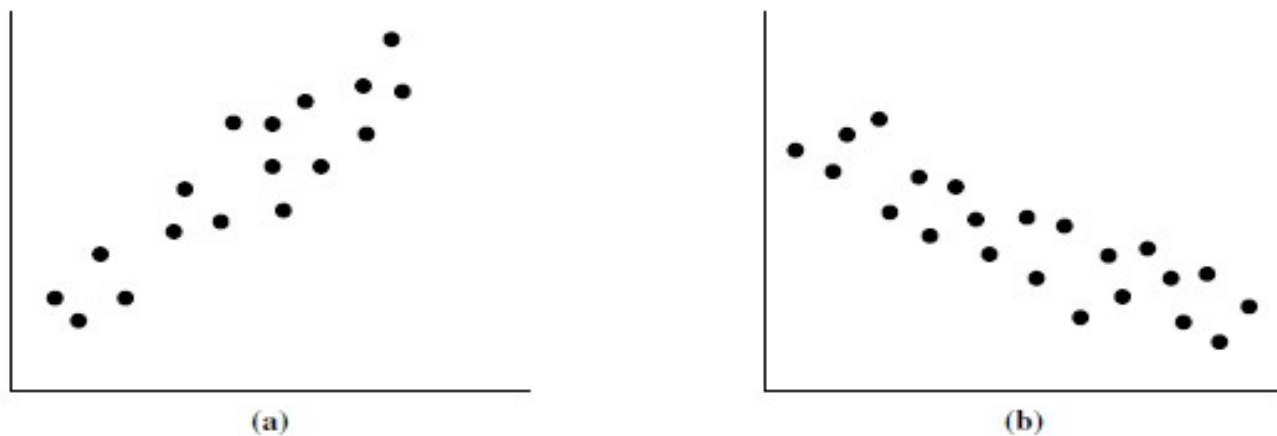


Figure 2.8 Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

Types of Datasets

- General characteristics of datasets
 - Dimensionality (Preprocessing require dimensionality reduction)
 - Sparsity (Saves the storage and computational time)
 - Resolution (seems to be different with corresponding difference)

Types of Datasets

- Record data
 - Transaction or market basket data
 - Data matrix or pattern matrix
 - Sparse data matrix or Document term matrix
- Graph based data
- Ordered data
 - Sequential data
 - Sequence data
 - Time series data
 - Spatial data

Record Data

- Dataset is a collection of records (Data objects) each of which consists of a fixed set of data fields (attributes)
- There is no explicit relationship among records or data fields and every record has same set of attributes.
- Record data usually stored either in flat-files or in relational DBs or DB Servers.

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

Transaction or Market Basket data

- It is a special type of record data, where each record(transaction) includes a set of items.
- Transaction data is a collection of set of items, but it can be viewed as a set of records whose fields are asymmetric attributes.
- The attributes can be
 - Discrete: No. of items purchased
 - Continuous: Amount spent

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Data Matrix or Pattern Matrix

- Set of data objects can be interpreted as an $m \times n$ matrix is called Data matrix.
- Data matrix is a variation of record data, but we can apply standard matrix operations to transform.
- It is the standard data format for most statistical data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

Sparse Data Matrix

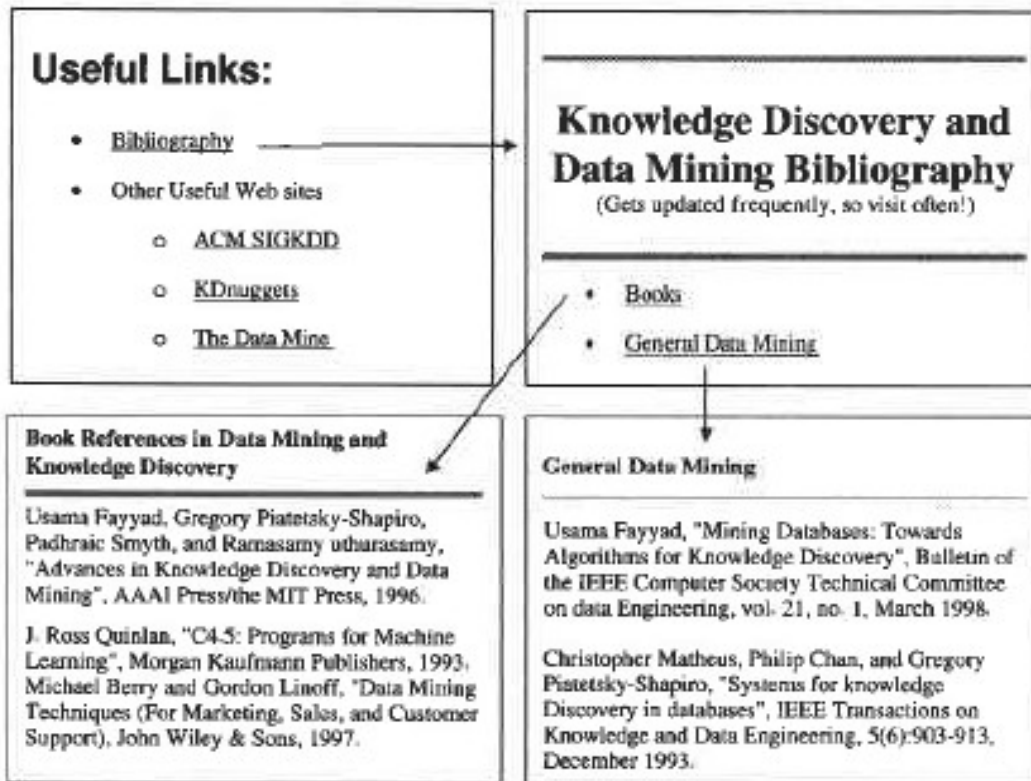
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- It is a spl.case of Data matrix in which the attributes are of the same type and are asymmetric(Only non-zero values are imp)
- Transaction data is an example of a Sparse matrix has only 0-1 entries.
- Ex: Document data
- Document can be a term vector, where each term is a component(attribute) of the vector and the values of each component is the no.of times that corresponding term occurs.

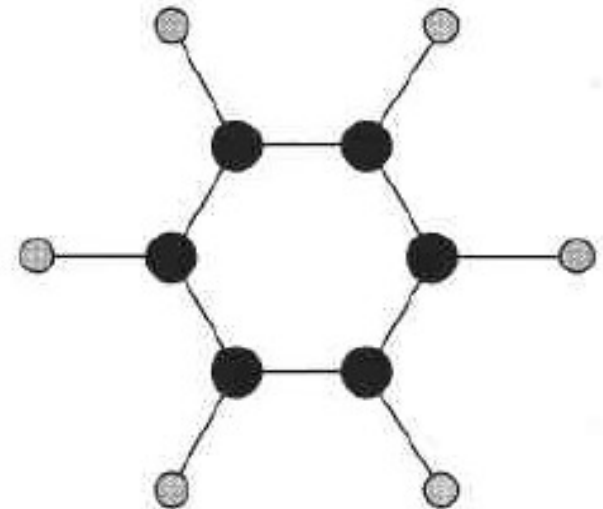
Graph-Based Data

- Two cases
 - The graph captures relationships among data objects
 - The data objects themselves are represented as graphs.

Different variations of Graph Data



(a) Linked Web pages.



(b) Benzene molecule.

Ordered Data

- Different types of ordered data are:
 - Sequential Data(Transaction)
 - Sequence data
 - Time series data
 - Spatial data

Sequential Data

- **Sequential data:**
Extension of record data, where each record has a time associated with it.
 - Ex: Retail transaction Dataset

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

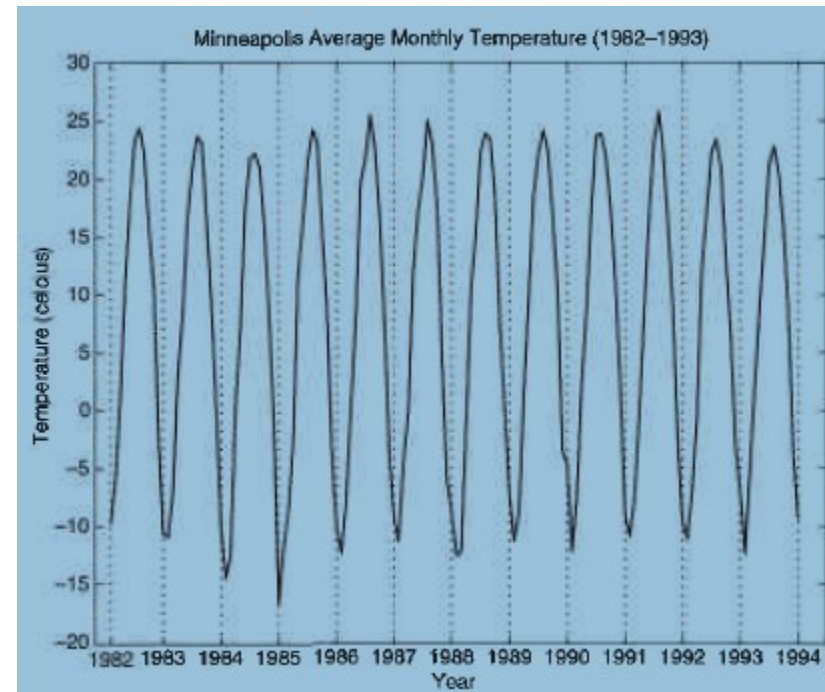
Sequence Data

- contains a sequence of individual entities, such as seq.of letters or words.
 - **Ex:** Genetic sequence data to predict similarities in the structure and functions of genes

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

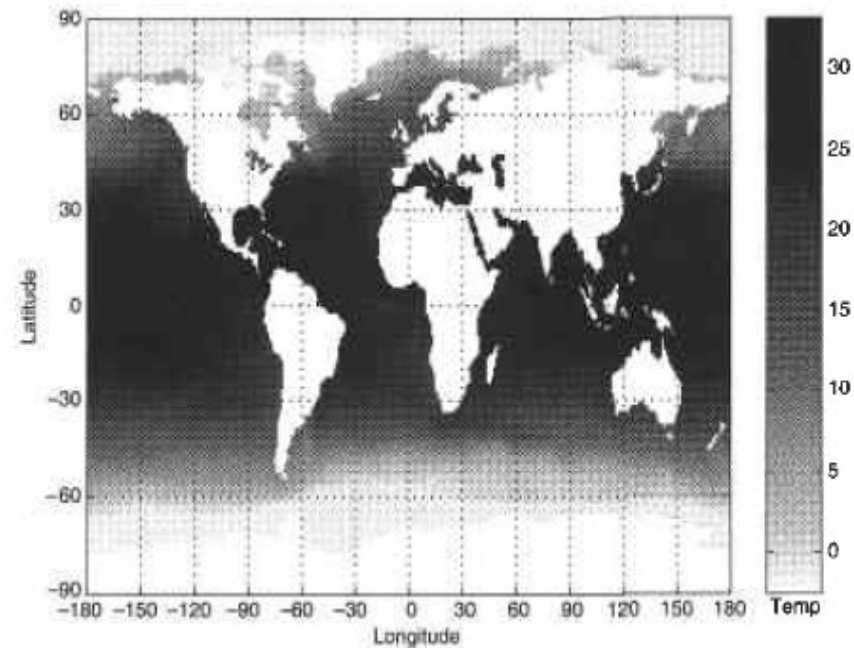
Time Series Data

- Time series data is a special type of sequential data in which each record is a time series over time.
- Ex: Stocks, temperature
- Temporal autocorrelation: When working with temporal data, if two measurements are close in time, then the values of those measurements are often very similar

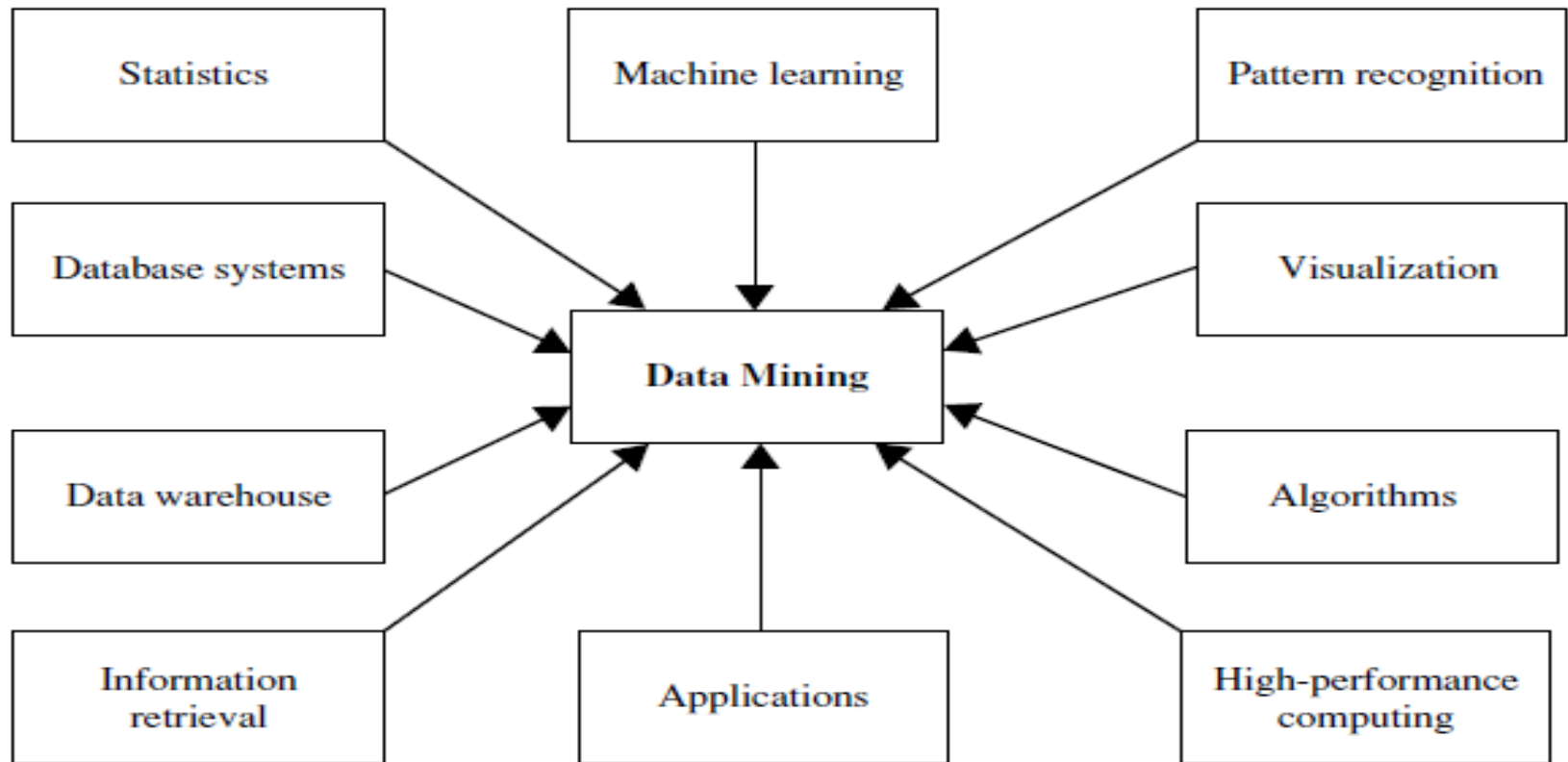


Spatial Data

- Some objects have special attributes, such as position or areas and other types of attributes
- Ex: Weather data collected from various geographical areas.



Technologies adopted by Data Mining



Data mining adopts techniques from many domains.

Technologies adopted by Data Mining

