Artificial Intelligence

Lecture 4. Data Acquisition & Preprocessing

I. Know your data

Spring 2022

Prof. Jonghoon Chun, Ph.D.

E-mail : jchun@mju.ac.kr
Lecture Note : http://lms.mju.ac.kr

Data Acquisition & Preprocessing

- Data (Know your data)

- Data Acquisition

- Data Preprocessing

Data Acquisition & Preprocessing

- Data (Know your data)

- Data Acquisition

- Data Preprocessing

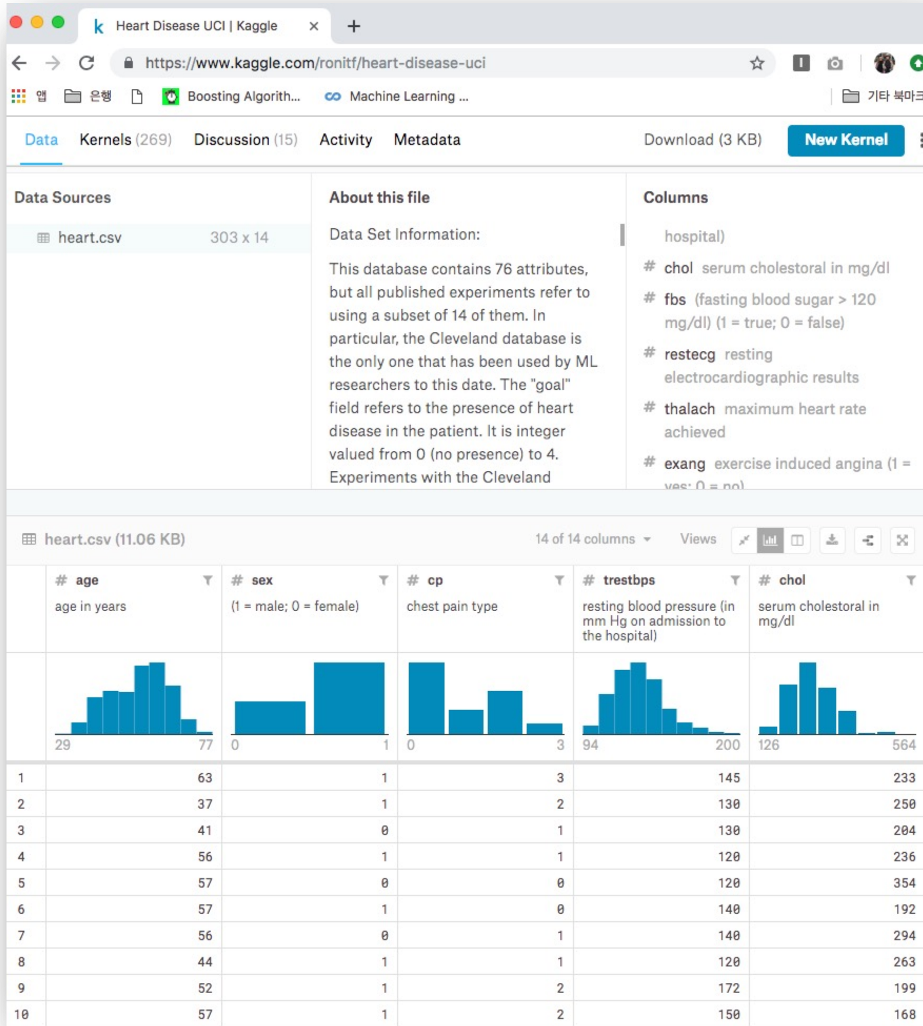
Know your data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Summary

DATA OBJECTS AND ATTRIBUTE TYPES

Where to look for data?

- Kaggle datasets (https://www.kaggle.com/datasets)



- Datahub.io (https://datahub.io/)
- Data.gov (https://www.data.gov/)
- Datausa.io (https://datausa.io/)
- European data portal (https://www.europeandataportal.eu/en)

Where to look for data?

- 공공데이터 개방
  - 공공데이터포털(data.go.kr), 지자체別 데이터개방(서울시 등)
  - 빅데이터 플랫폼 및 센터(bigdata-map.kr)
  - 인공지능 학습 데이터(aihub.or.kr)
  - 언론진흥재단(bigkinds.or.kr)

- Internet
  - 블로그와 SNS
    - Facebook, Twitter
  - 전자상거래 데이터
    - 네이버 웹 API (developers.naver.com/products/intro/plan/)
  - 금융정보
    - 주식, 환율, 금값
  - 이미지 데이터
  - 위키피디아

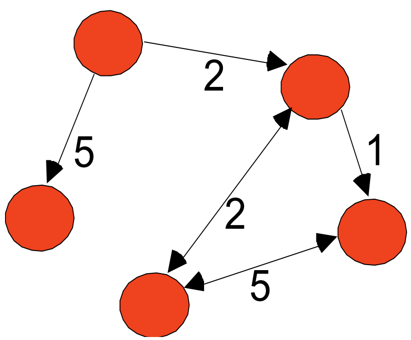
Types of Data Sets

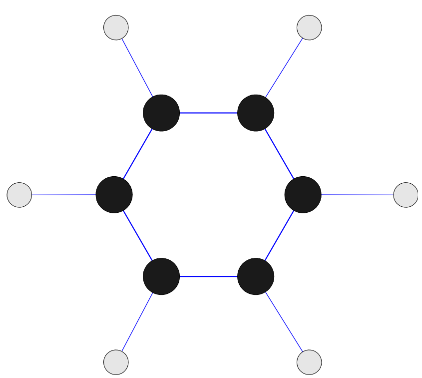
- Record
  - Relational database tuples(records)
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data

	team	coach	pla y	ball	score	game	wi n	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

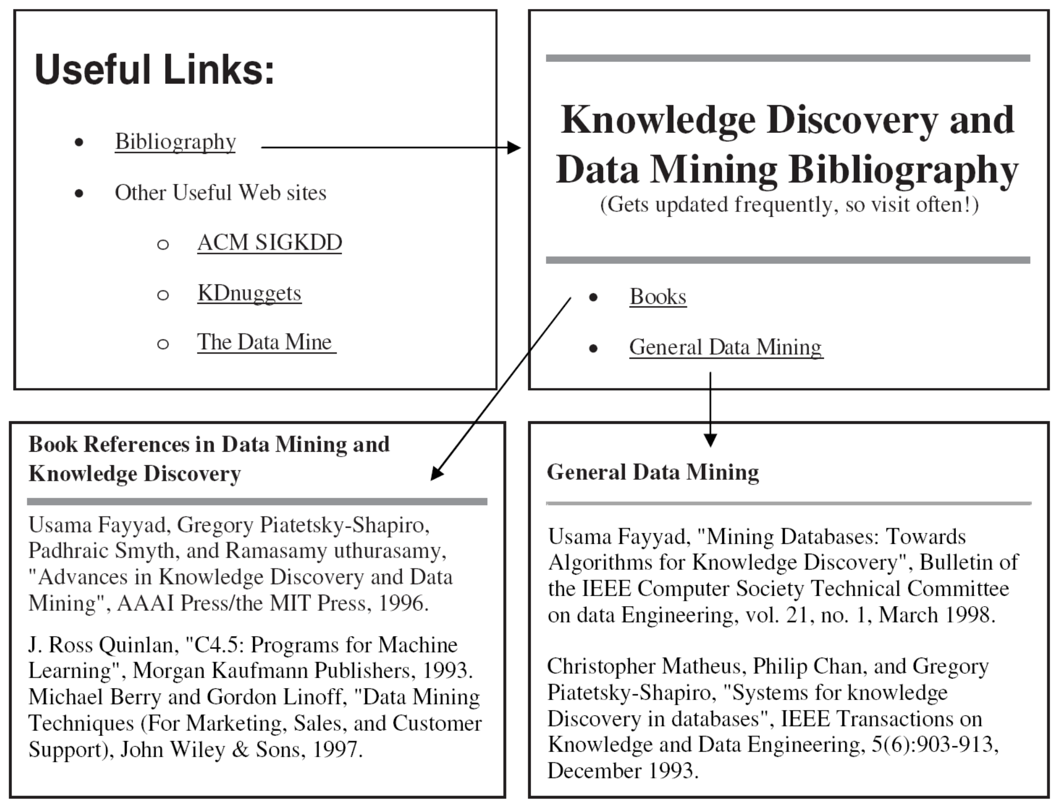
Types of Data Sets

- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures





Benzene Molecule: C6H6



Types of Data Sets

- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

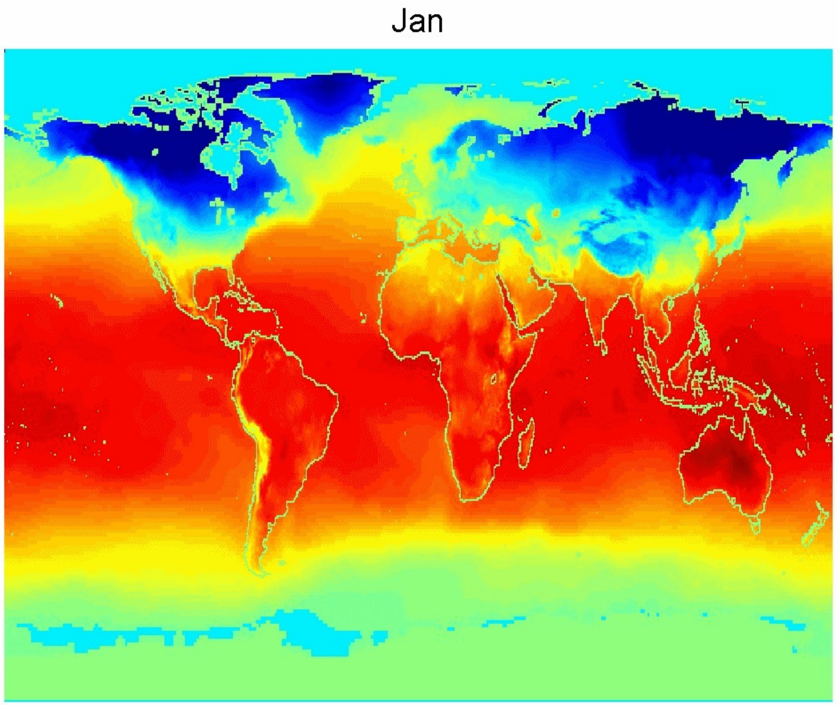
Sequence of transactions

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

Genetic sequence data

Types of Data Sets

- Spatial, image and multimedia
  - Spatial data: maps
  - Image data
  - Video data



Spatio-temporal data: Average Monthly Temperature of land and ocean

Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

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.

Attributes

- **Attribute (**or **dimensions, features, variables**)
  - a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*

- Attribute Types
  - Nominal: e.g., ID numbers, eye color, zip codes
  - Ordinal: e.g., rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Numeric: quantitative
    - Interval-scaled: e.g., calendar dates, temperatures in Celsius or Fahrenheit.
    - Ratio-scaled: e.g., length, time, counts

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - 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**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent zero-point
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., temperature in Kelvin, length, counts, monetary quantities

Discrete vs. 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

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \mu = \frac{\sum x}{N}$

  Note: n is sample size and N is population size.

  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

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

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

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*):

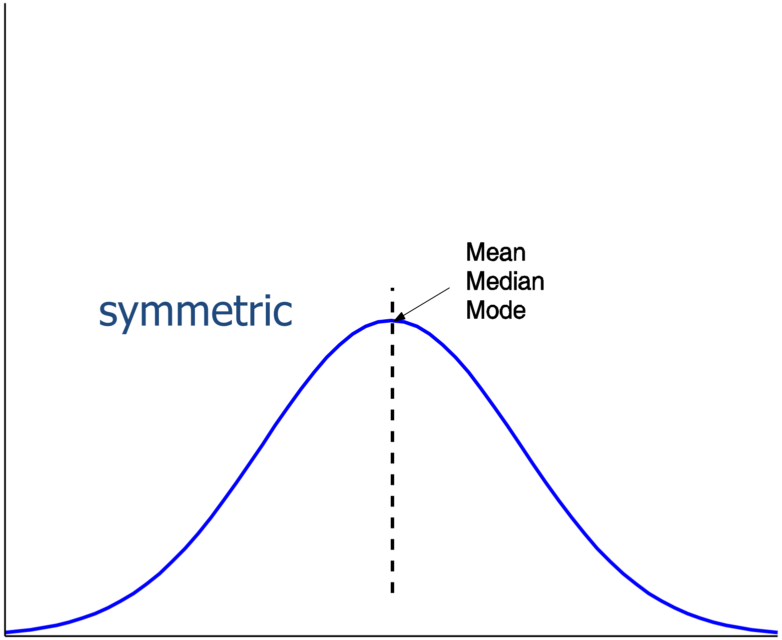
$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}}\right) width$$

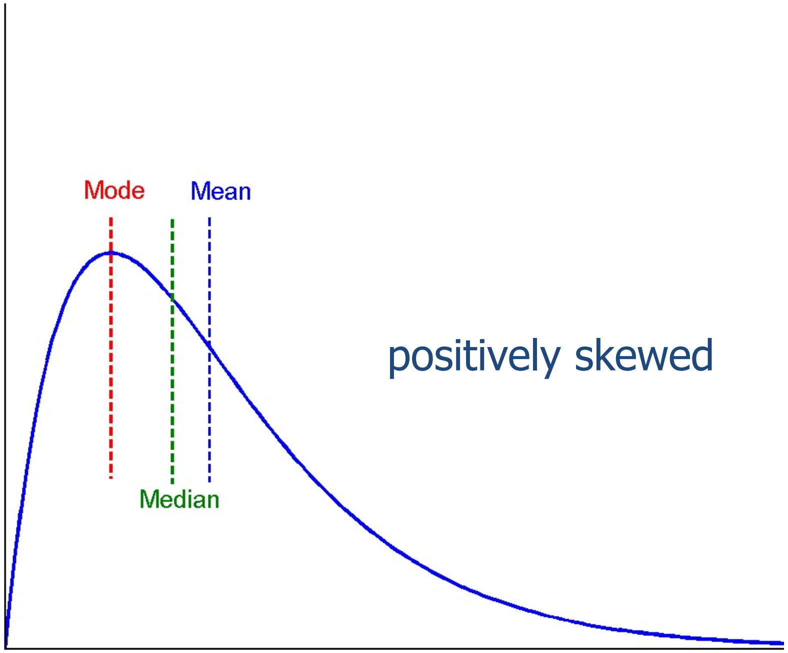
where L_1 is the lower boundary of the median interval, n is the number of values in the entire data set, $(\Sigma freq)l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and *width* is the width of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

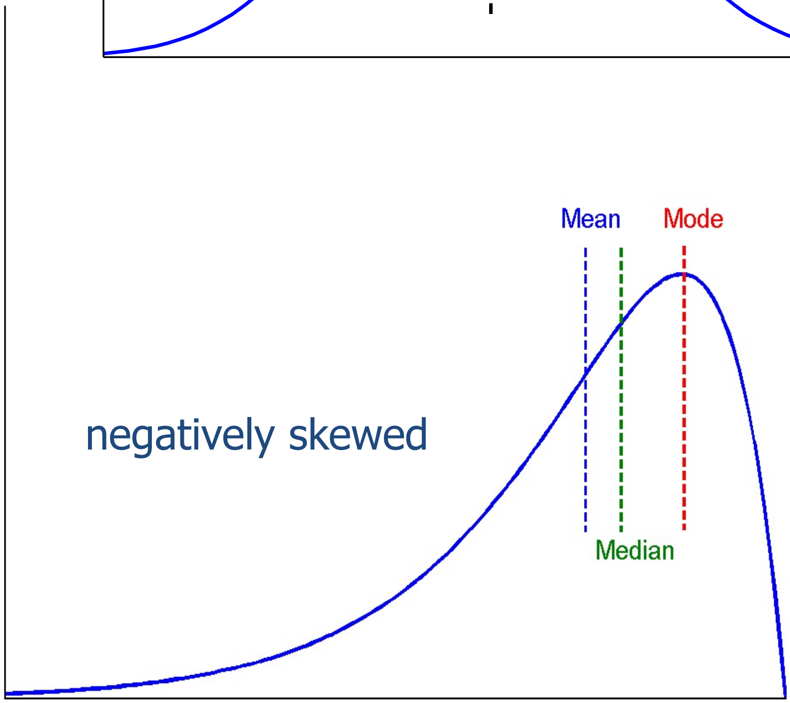
Age	Frequency
1-5	20
6-10	35
11-16	150
16-20	300
21-50	1500
51-80	700
81-110	44

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Measuring the Dispersion of Data

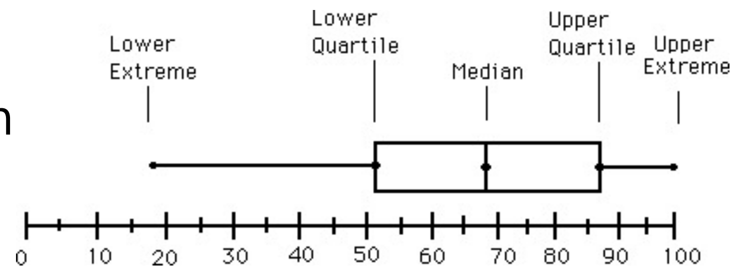
- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

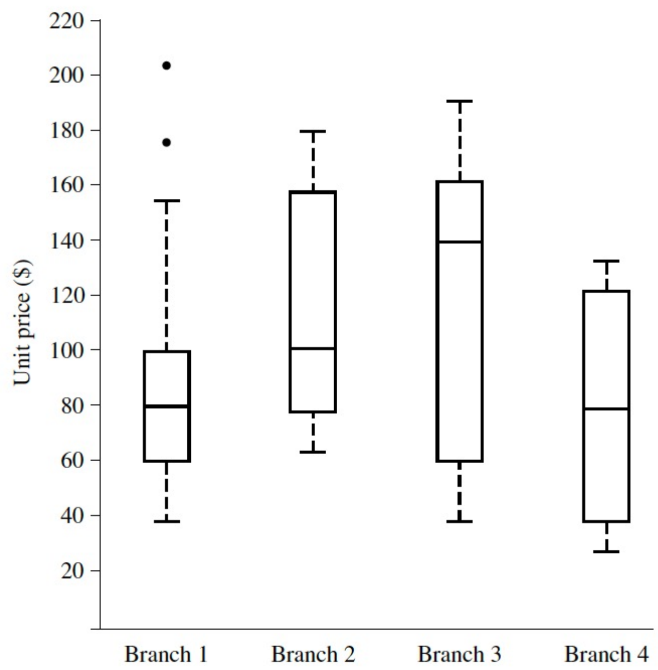
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^{\ 2} - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2] \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\ 2} - \mu^2$$

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

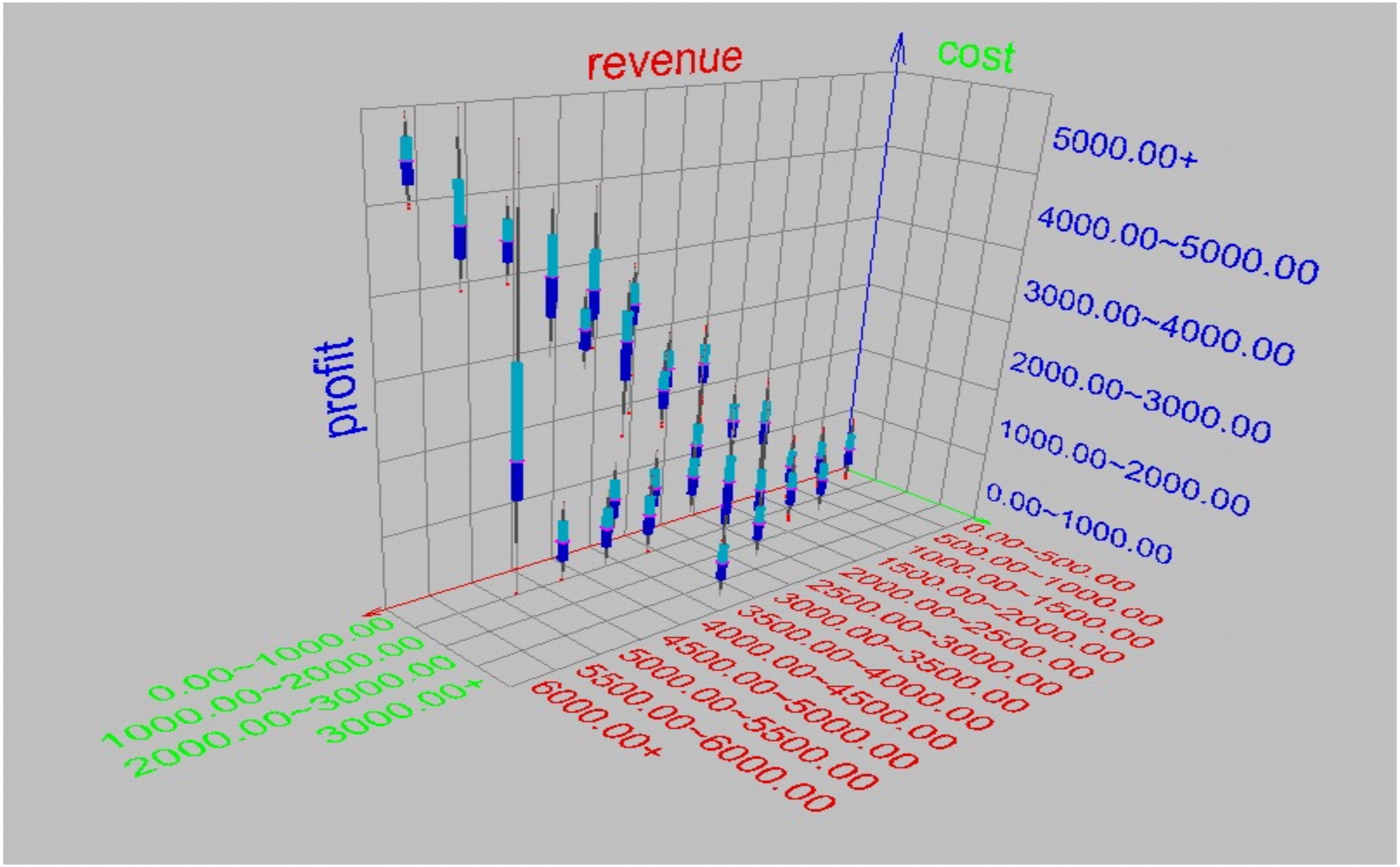
Boxplot Analysis

- **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 first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually





Visualization of Data Dispersion: 3-D Boxplots

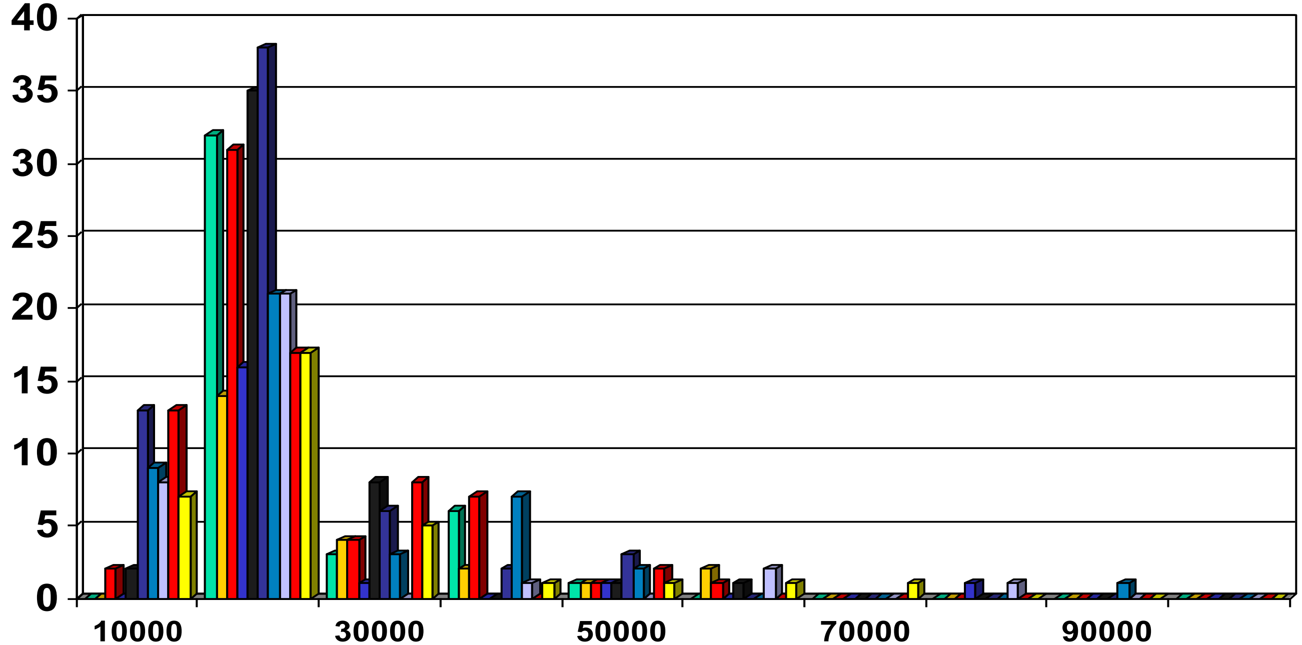


Graphic Displays of Basic Statistical Descriptions

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

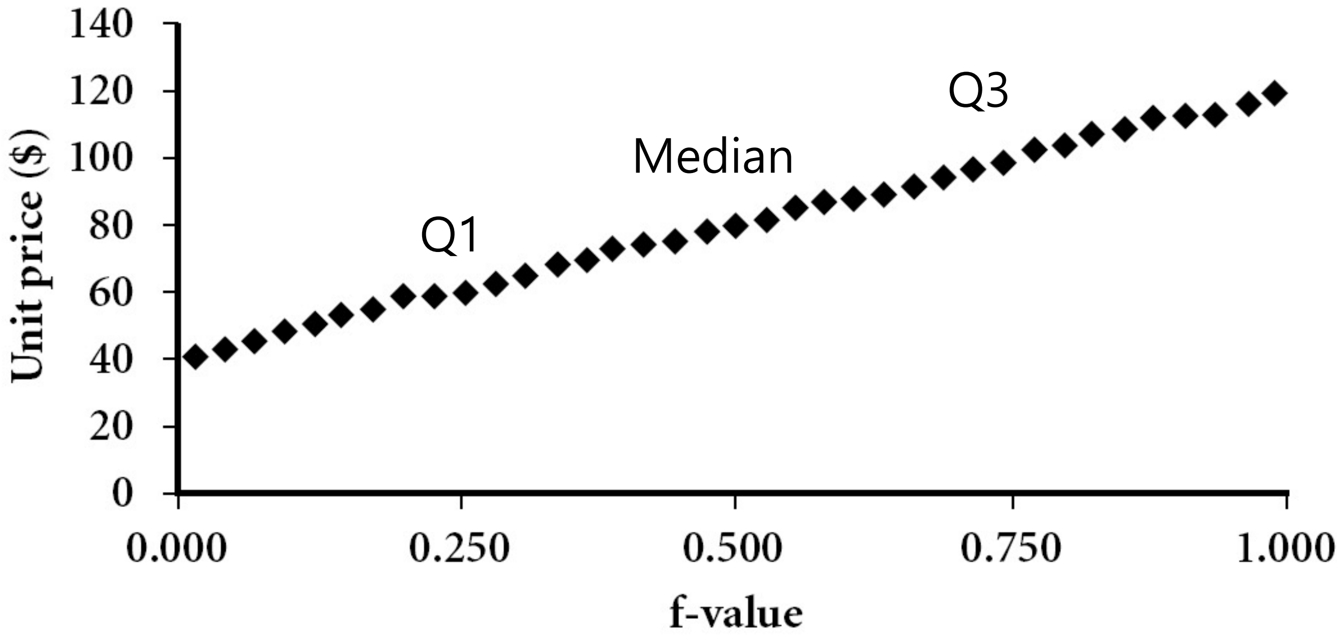
Histogram Analysis

- 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, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



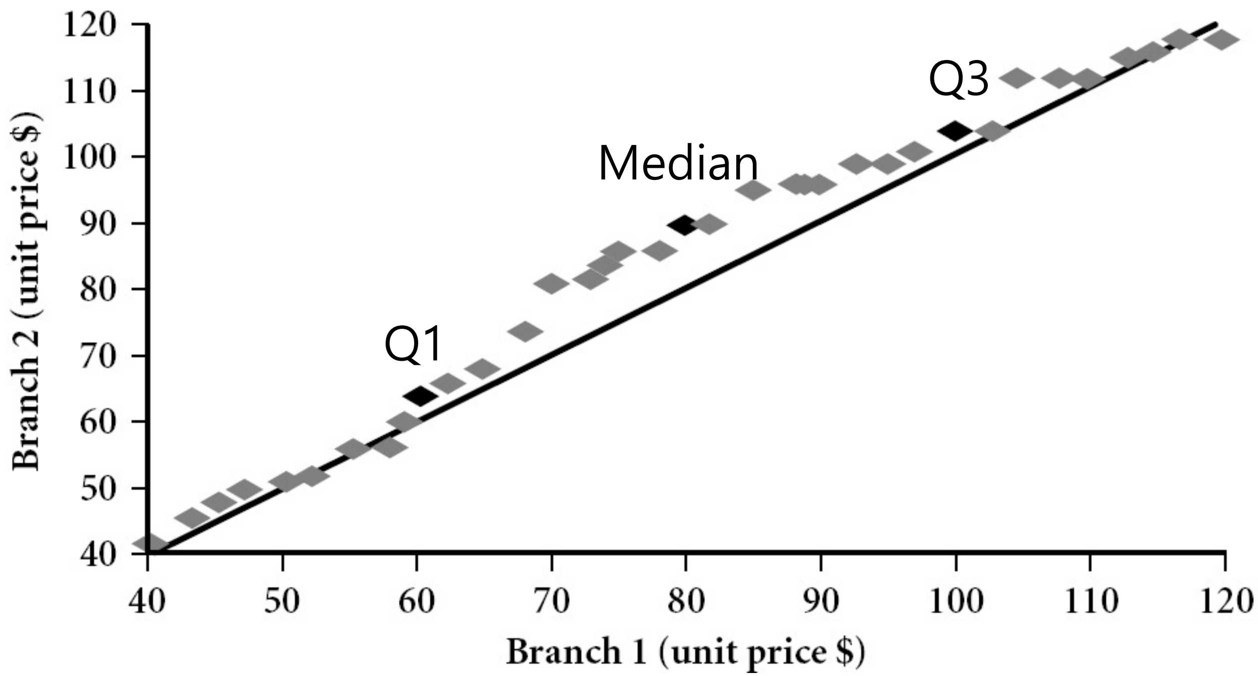
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 X f_i% of the data are below or equal to the value x_i



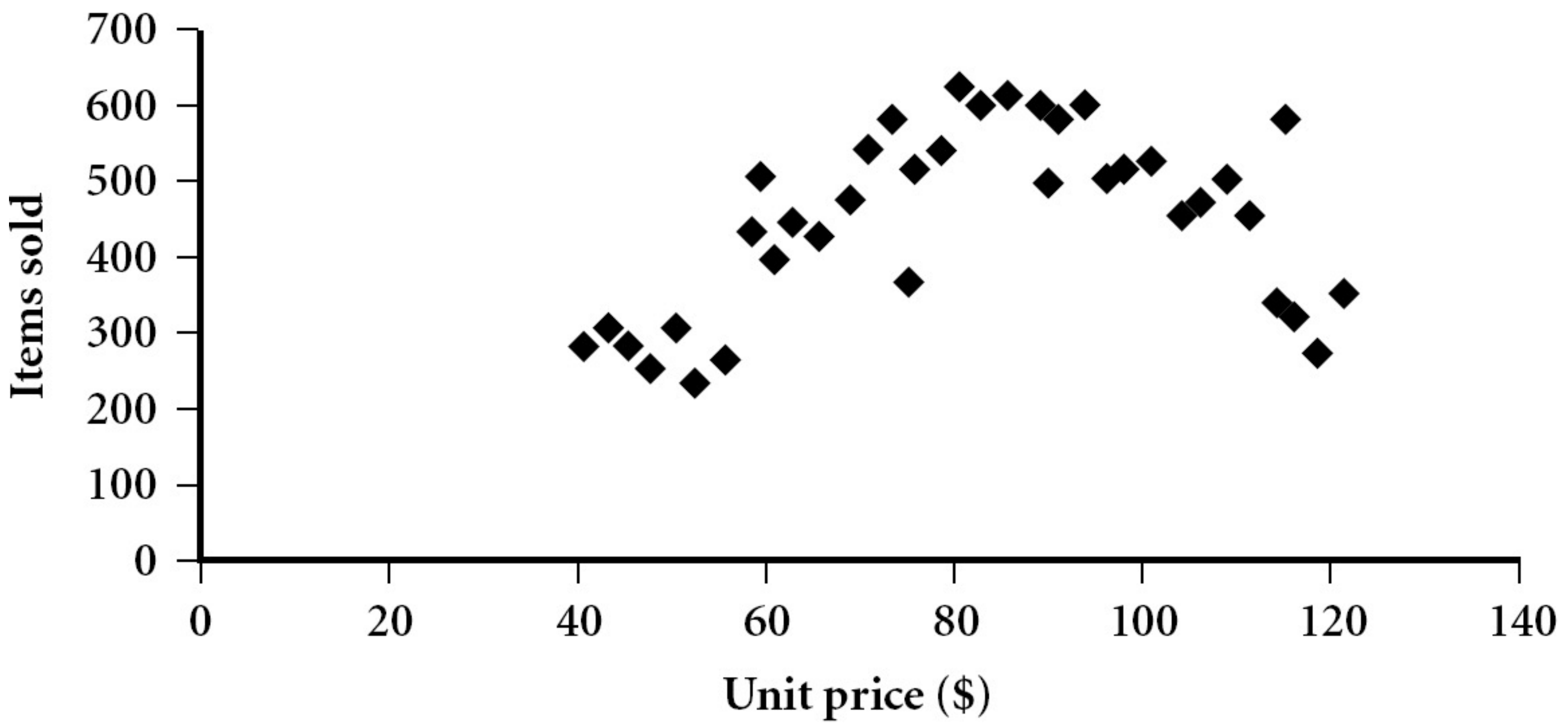
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is 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.

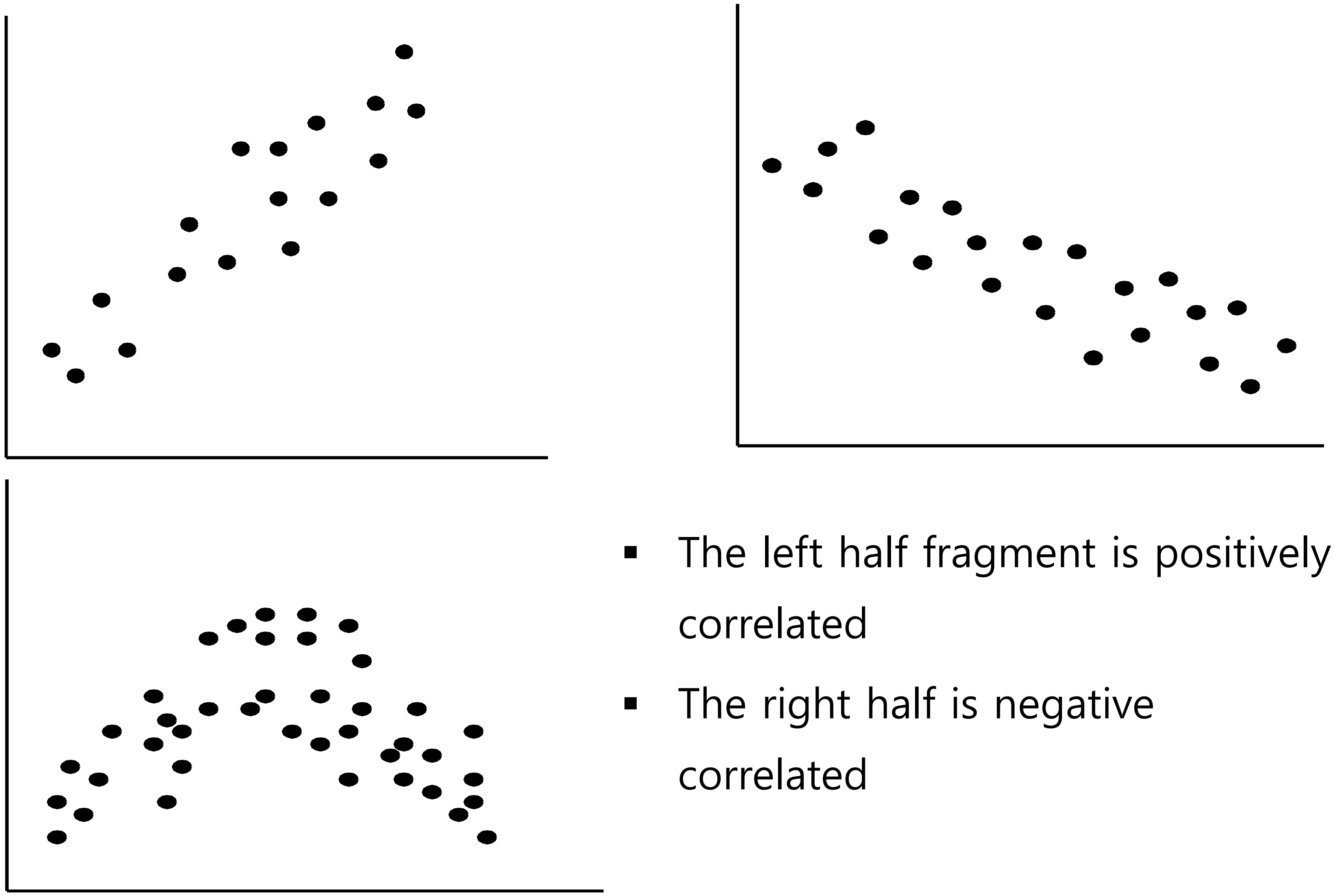


Scatter plot

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

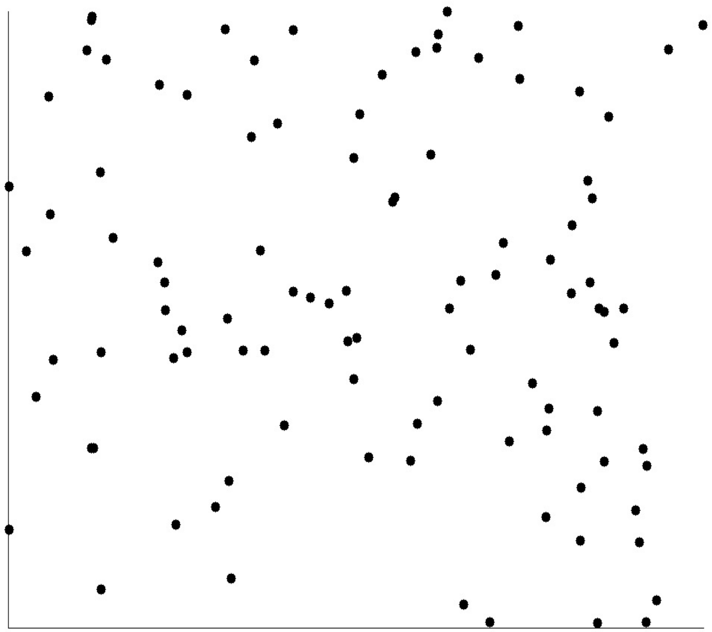


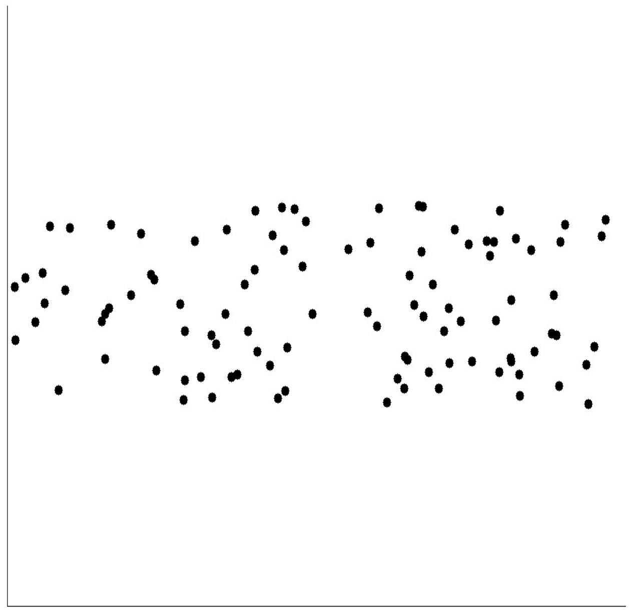
Positively and Negatively Correlated Data

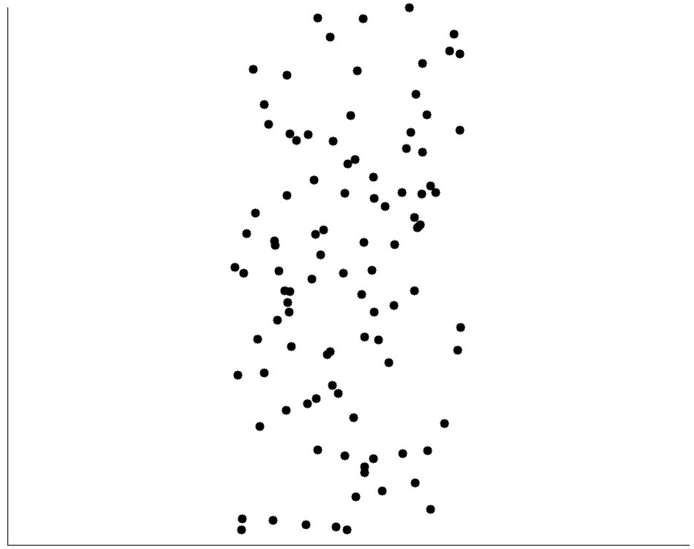


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data







MEASURING DATA SIMILARITY AND DISSIMILARITY

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

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 registers only the distance
  - A triangular matrix
  - Single mode

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

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)

- Method 1: Simple matching
  - m: # of matches, p: total # of variables

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

- Method 2: Use a large number of binary attributes
  - creating a new binary attribute for each of the M nominal states
  - E.g., For {red, yellow, blue, green}, yellow = 0100, blue = 0010

Proximity Measure for Binary Attributes

- A contingency table for binary data

Object *j*

Object *i*		1	0	sum
	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:

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

*Recall that for symmetric binary variables each state is equally valuable.

Proximity Measure for Binary Attributes

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

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

- Example: gender는 symmetric, 나머지는 asymmetric. Y=Yes, N=Negative, P=Positive. 유사한 질병을 가질 확률이 있는 사람은 누구와 누구인가?

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

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 $i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $j = (x_{j1}, x_{j2}, ..., x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)

- Properties
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)
- A distance that satisfies these properties is a metric

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

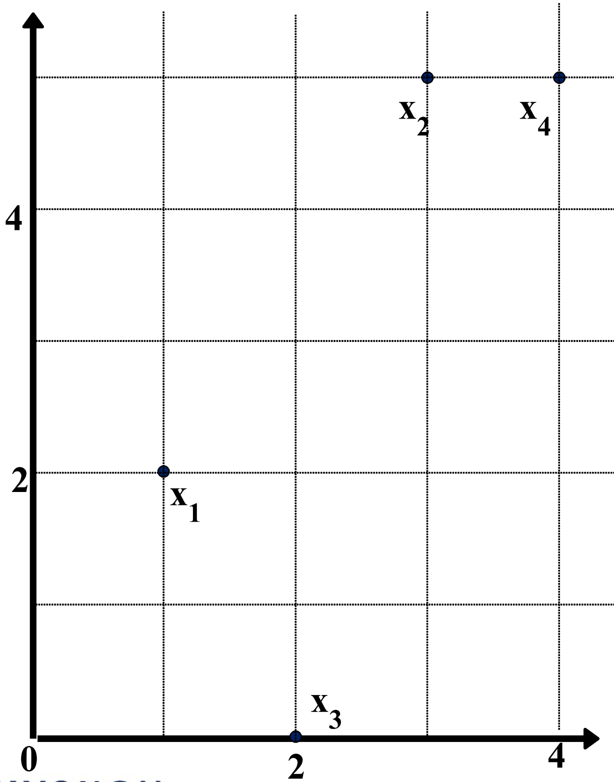
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

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

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

Example: Minkowski Distance

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



Dissimilarity Matrices

Manhattan (L1)

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

Euclidean (L2)

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

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by

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

  - compute the dissimilarity using methods for interval-scaled variables

Example

- Sample data

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

- Rank = {fair:1, good:2, excellent:3}
- Normalization rank 1 → 0, rank 2 → 0.5, rank 3 → 1
- Then use Euclidean distance to measure the dissimilarity

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

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

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

  - f 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} and
    - Treat z_{if} as interval-scaled $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

Cosine Similarity

- A document can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

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

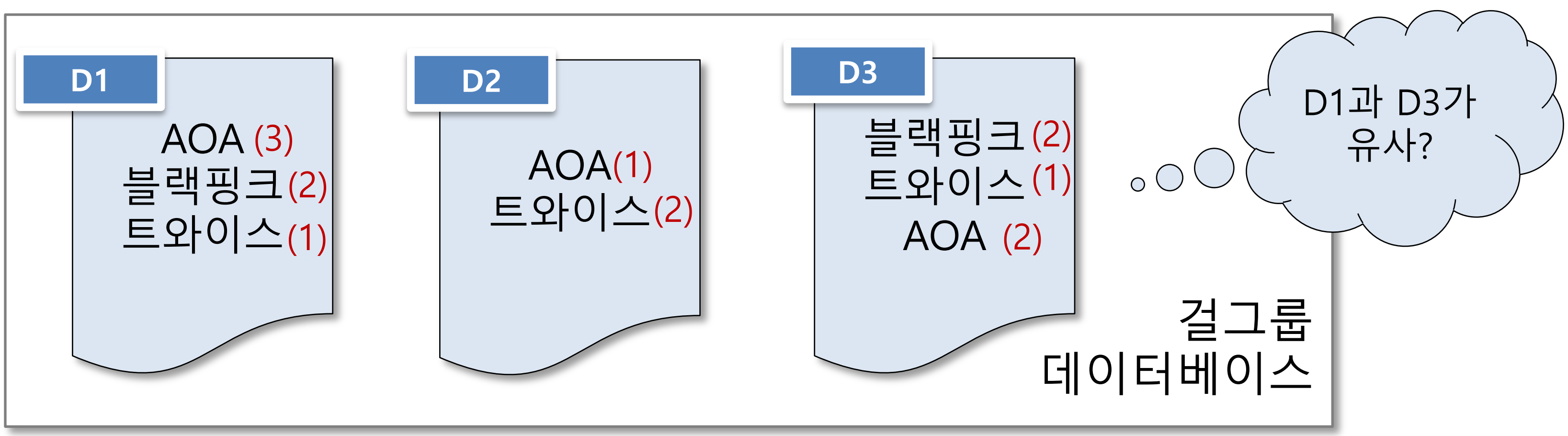
- Other vector objects: gene features in micro-arrays, …

- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...

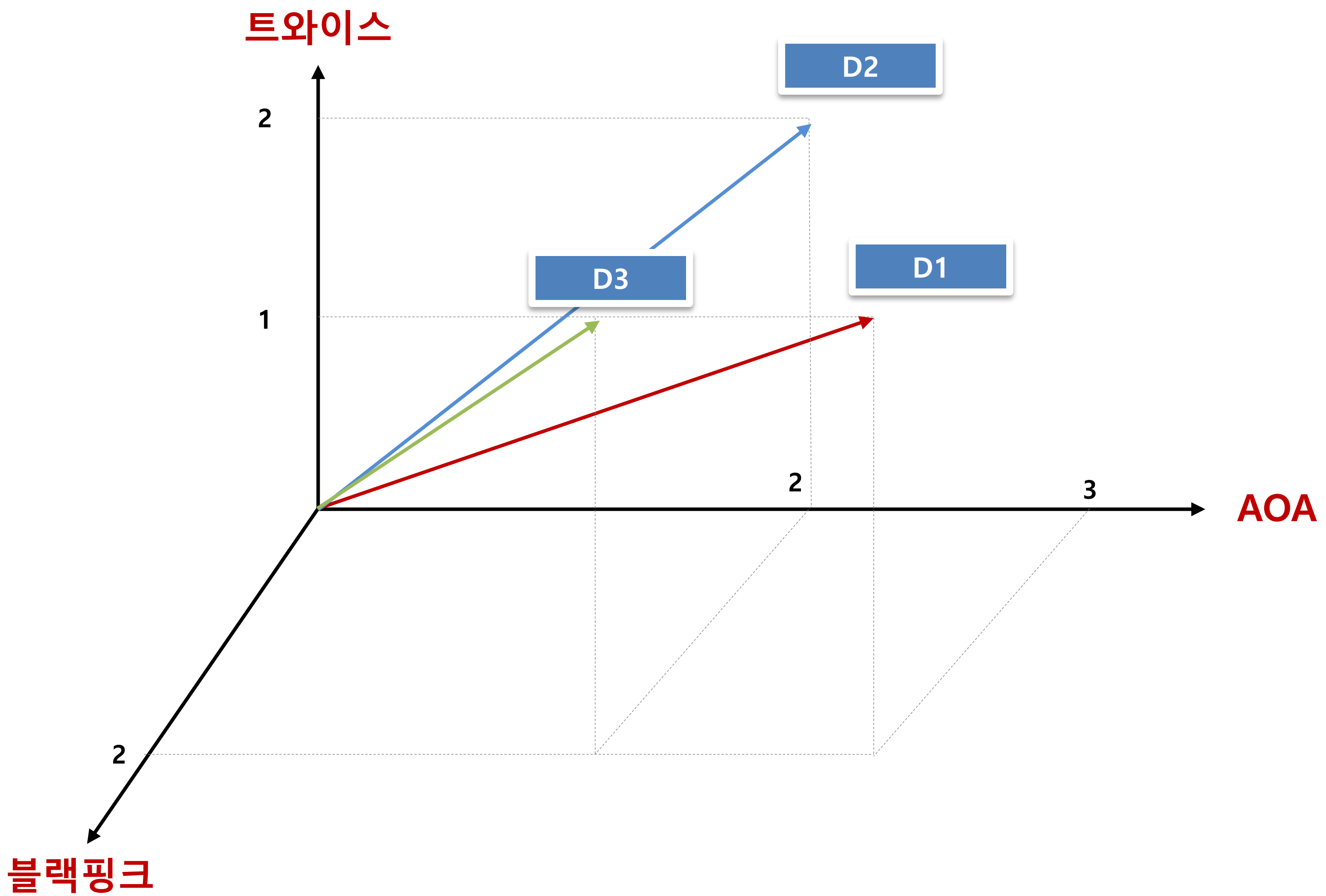
Cosine Similarity

- Vector Space Model
  - 문서(document)가 벡터?
  - 단어 하나 하나가 벡터의 차원이 됨!

- Assumption
  - 가정 1: 단어는 단 3개만 존재한다고 가정(AOA, 블랙핑크, 트와이스)
  - 가정 2: Database에 단 3건의 문서만 저장



Vector Space Model



Vector Space Model

- Similarity를 pair-wise로 구해보자
  - D1과 D2의 유사도는? D1 vector와 D2 vector의 유사도는?
  - D1과 D3의 유사도는? D1 vector와 D3 vector의 유사도는?
  - D2와 DS3의 유사도는? D2 vector와 D3 vector의 유사도는?
  - 방향성이 없으므로 N(N-1)/2 의 유사도를 측정

- 벡터간의 유사도를 측정
  - 2개의 벡터간에 벌어진 각도를 측정 → cosine measure
  - $x \bullet y = x_1 y_1 + x_2 y_2$ (2차원인 경우)
  - $x \bullet y = x_1 y_1 + x_2 y_2 + x_3 y_3$ (3차원인 경우 → AOA, 블랙핑크, 트와이스)
  - $x \bullet y = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$ (n차원인 경우)
    $= \|x\| \, \|y\| \cos\theta$

$$x \bullet y = \|x\| \, \|y\| \cos\theta$$
$$\cos\theta = (x \bullet y) \; / \; \|x\| \, \|y\|$$

Cosine Similarity

- 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

- Two vectors are orthogonal => the value of Cosine is 0 (하나도 유사하지 않은 경우)
- Two vectors are identical => value of Cosine is 1 (완전히 동일한 경우)

- 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)

  $d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$
  $||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
  $||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$
  $\cos(d_1, d_2) = 0.94$

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.

END