CS 412 Intro. to Data Mining

Chapter 2. Getting to Know Your Data

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

Chapter 2. Getting to Know Your Data

คุณสมบัติของ Data (เป็นคอลัม)

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

บอกคุณลักษณะของ Data ในรูปตาราง *

แบบ ①

Types of Data Sets: (1) Record Data

- Relational records
  - Relational tables, highly structured
- Data matrix, e.g., numerical matrix, crosstabs

มาจาก data base

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

— no relation

แยกตาราง

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

แบบ relational

* อย่างแรก เก็บข้อมูลเป็นตาราง

ตาราง matrix

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

สต็อก | ราคา

- Transaction data

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

* เป็น Transaction ที่เรียงตามเวลา เช่น วันนี้ โอนเท่าไร พรุ่งนี้ โอนเท่านี้

	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

ตาราง Document

* สร้าง Documet 1,2,3/บทความ เพื่อดูว่าในแต่ละ Documet มีคำไหนบ้าง โดยตารางจะโชว์จำนวนข้อคำว่า ซ้ำกี่ครั้ง เพื่อดูว่าเป็นบทความเกี่ยวกับอะไร

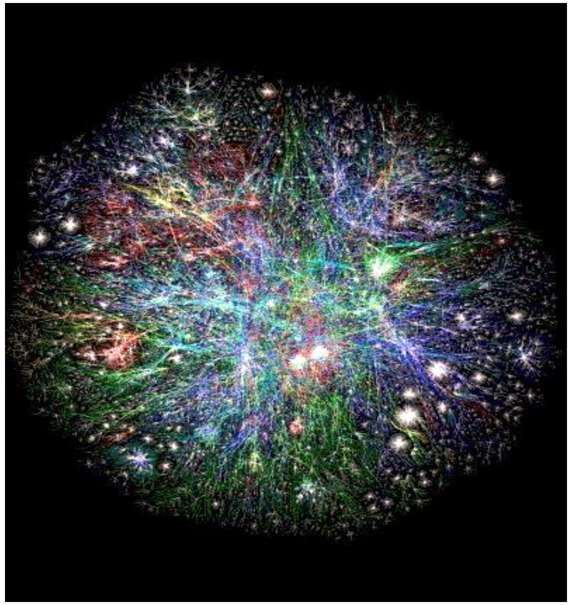
- Document data: Term-frequency vector (matrix) of text documents

แบบ 2

Types of Data Sets: (2) Graphs and Networks

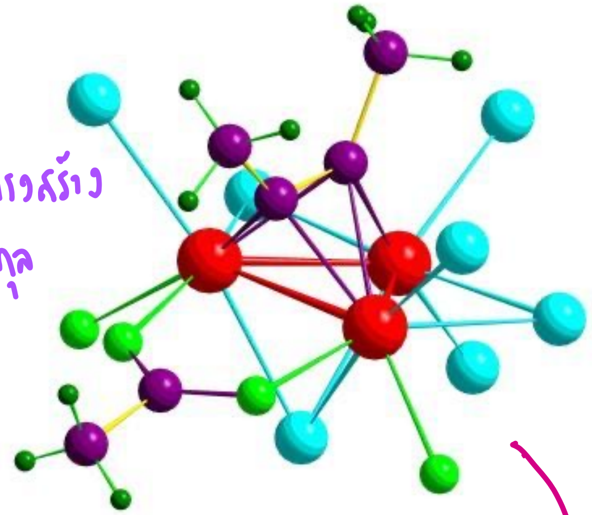
- ❑ Transportation network
- ❑ World Wide Web

World Wide Web

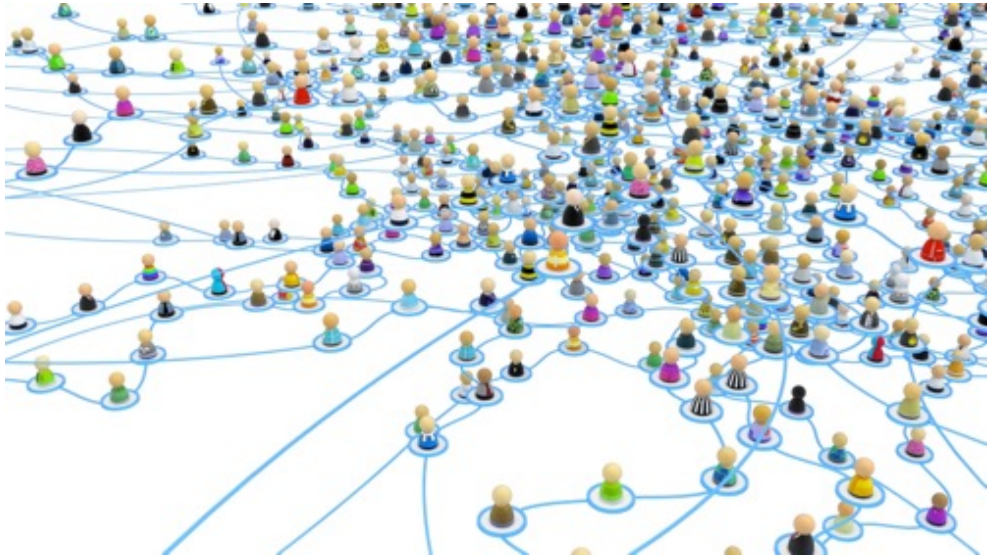


* เกี่ยวกับการลิงก์ของเว็บ บ่งบอกคุณลักษณะของเว็บ โดยที่ได้ต้องเข้าไปดูเนื้อหา

* หาโครงสร้างของโมเลกุล



- ❑ Molecular Structures
- ❑ Social or information networks



* เชื่อมโยงของกลุ่มคนที่เหมือนกัน ลักษณะคล้ายๆกัน

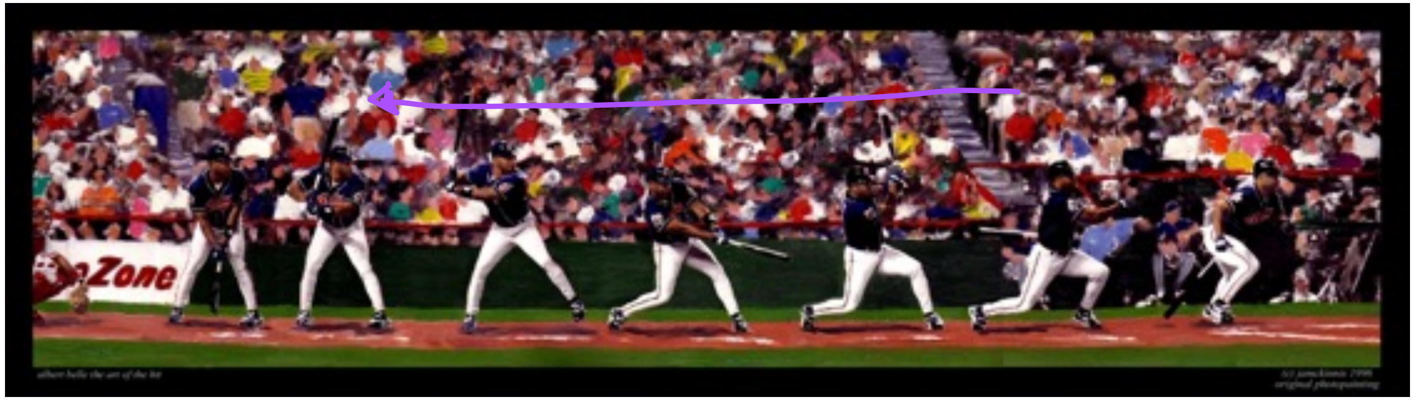
แบบที่ 3
ดูลำดับเป็นหลัก

Types of Data Sets: (3) Ordered Data

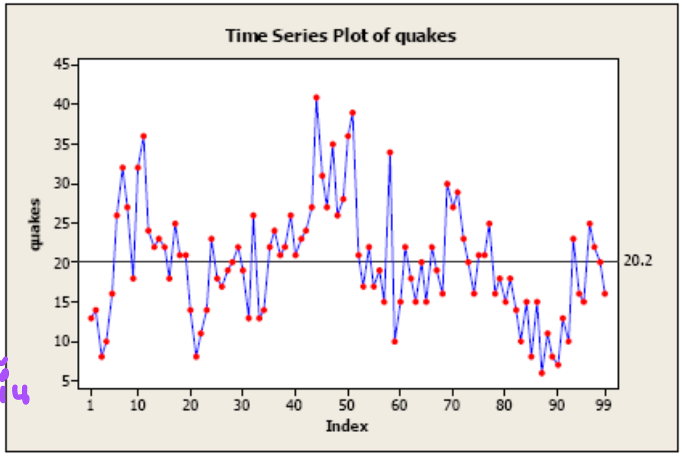
* สนใจ เลยดูว่า แพทเทิร์น ต่างกัน

* ภาพย้อนกลับ

- ❑ Video data: sequence of images



- ❑ Temporal data: time-series



* ดูกราฟ
เช่นนั้น
* ดูแพทเทิร์น
ว่ามีการซ้ำ ?
เพื่อดูว่ามันจะพุ่งขึ้น
มั้ย?

* บอกลำดับเพื่อดูว่า
คุณลักษณะไหนที่ส่งผลให้
เกิด ?

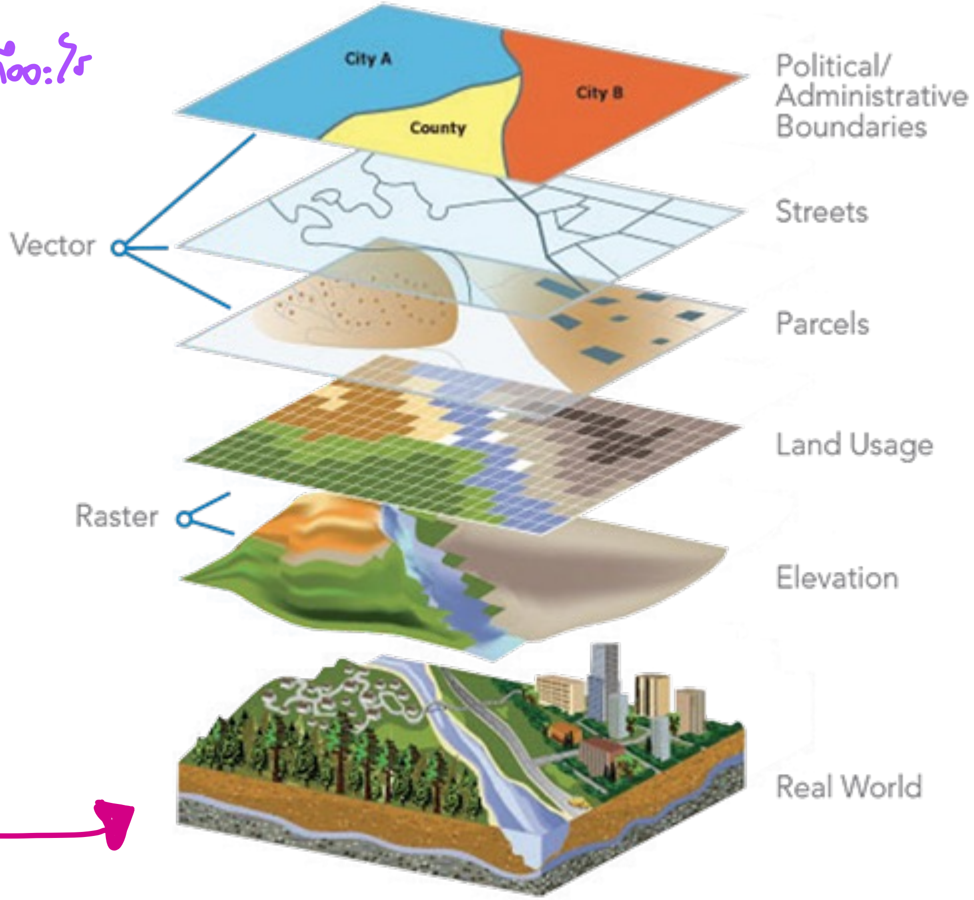
- ❑ Sequential Data: transaction sequences

- ❑ Genetic sequence data

Types of Data Sets: (4) Spatial, image and multimedia Data

แบบที่ 4

- ❑ Spatial data: maps * เพื่อดูสีของรูปแล้วบ่งบอกว่า รูปนี้คืออะไร



- ❑ Image data:
- ❑ Video data:

* Spatio-temporal

* ลักษณะของข้อมูลที่เป็นตาราง

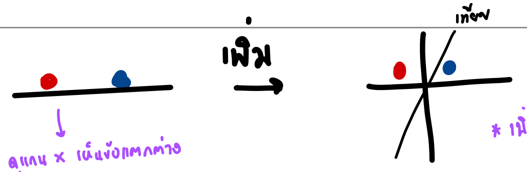
Important Characteristics of Structured Data

เริ่มแรก

มีทั้งหมดกี่มิติ

1. - Dimensionality มิติ (จำนวนคอลัมน์)

ex



ดูแกน x เห็นข้อแตกต่าง

* เพิ่ม Dimen ทำให้ดูความแตกต่างยาก

  - Curse of dimensionality * ยิ่งมีมิติมาก ยิ่งไม่ดี
- Sparsity ความว่างเปล่า (ค่า 0 / ค่าว่าง)
  - Only presence counts
- Resolution ความละเอียดของข้อมูล
  - Patterns depend on the scale เช่นข้อมูประเทศ → จังหวัด → อำเภอ → ละเอียดมากสุด
- Distribution การกระจายข้อม ของ Dimen ว่า Distribution เป็นแบบไหนยังไง
  - Centrality and dispersion

Data Objects

ประกอบด้วย Data หลายๆอัน

- Data sets are made up of data objects
- A **data object** represents an entity

  data 1 จุด / data point

- Examples:

  rerational   treansation

  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples*

  ดูอธิบายด้วย

- Data objects are described by **attributes**
- Database rows → data objects; columns → attributes

Attributes

* ใช้อธิบาย Data point

ค.หมาย

แมชชีนเลิร์นนิ่ง

- **Attribute** (or **dimensions, features, variables**)
  - A data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal (e.g., red, blue) เป็นชื่อ ไม่สามารถเอามาเปรียบเทียบ
  - Binary (e.g., {true, false}) * ลักษณะข้อมูลมี 2 ค่า ex. No , Yes
  - Ordinal (e.g., {freshman, sophomore, junior, senior}) * คล้าย Nominal แต่ Ordinal เรียงค่า (ปี 1 ปี 2 ปี 3)
  - Numeric: quantitative
    - Interval-scaled: 100°C is interval scales
    - Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50 °K
- Q1: Is student ID a nominal, ordinal, or interval-scaled data?
- Q2: What about eye color? Or color in the color spectrum of physics?

Attribute Types

- **Nominal:** categories, states, or "names of things" * ไม่สามารถแปลงเป็นตัวเลขได้
  - *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)
  - <u>Symmetric binary</u>: both outcomes equally important
    - e.g., gender * โอกาสที่ ใช่ / ไม่ใช่ = โอกาสเท่ากัน
  - <u>Asymmetric binary</u>: outcomes not equally important. ex. การเกิดโรค
    - e.g., medical test (positive vs. negative) * สนใจด้านที่เป็น + มากกว่า -
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal** แปลงเป็นตัวเลขได้
  - 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** มี 0 แท้
  - 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

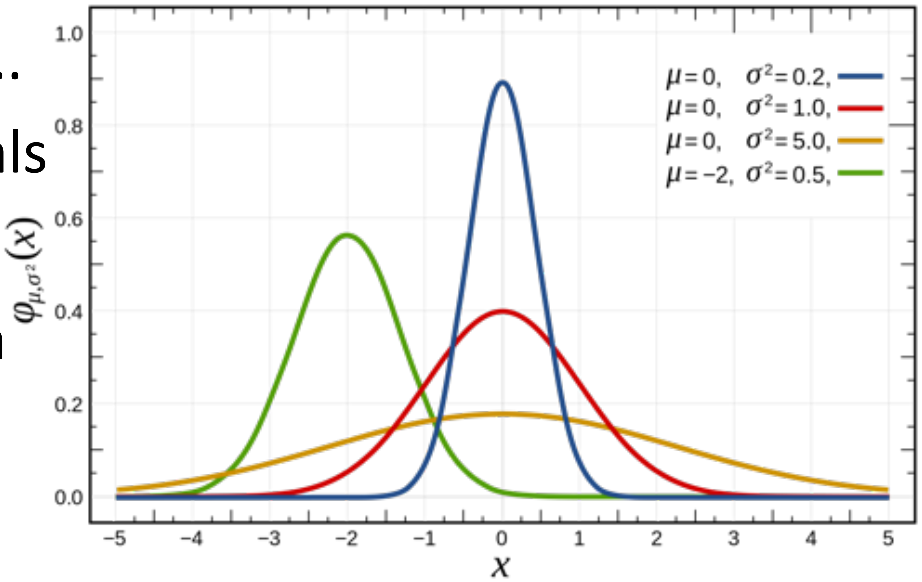
Chapter 2. Getting to Know Your Data

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

ส่วนที่สนใจ

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, ...
- 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: (1) Mean

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

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

sample

population

- Weighted arithmetic mean:

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

- Trimmed mean:
  - Chopping extreme values (e.g., Olympics gymnastics score computation)

Measuring the Central Tendency: (2) Median

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

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

Approximate median

Sum before the median interval

Interval width ($L_2 - L_1$)

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

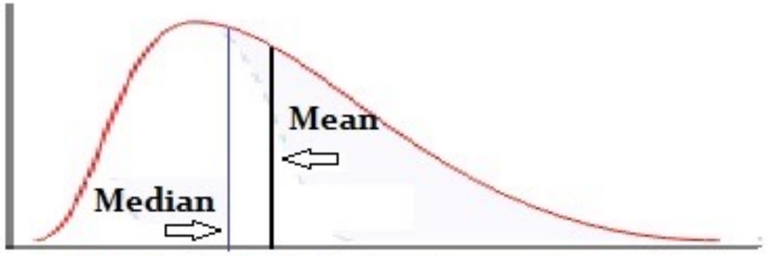
Low interval limit

Measuring the Central Tendency: (3) Mode

- Mode: Value that occurs most frequently in the data ค่าที่มากที่สุด

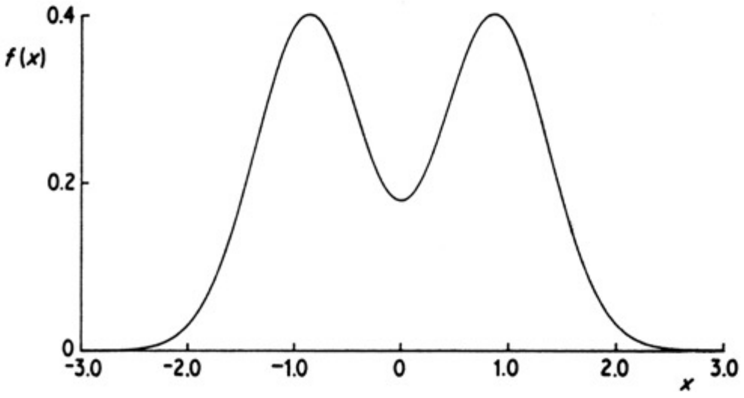
- Unimodal
  - Empirical formula:

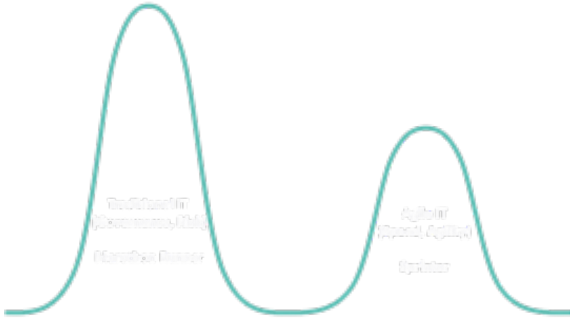
$$mean - mode = 3 \times (mean - median)$$

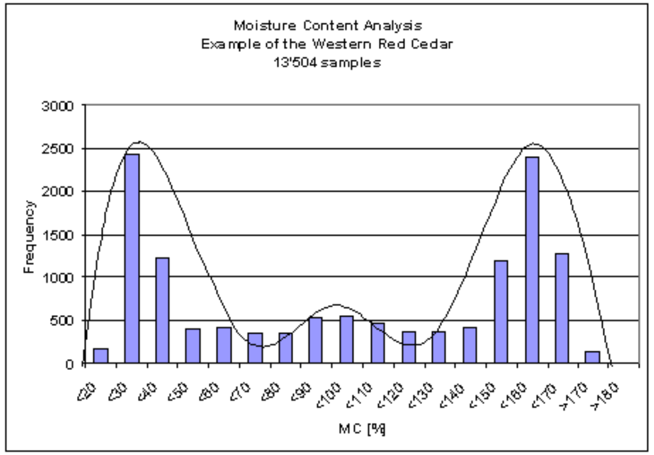


Right skewed distribution: Mean is to the right

- Multi-modal
  - Bimodal
  - Trimodal

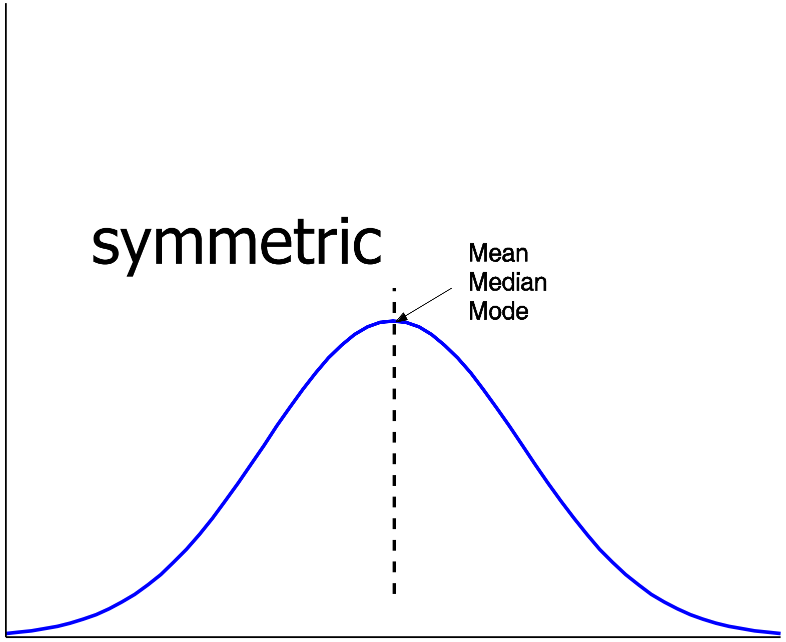


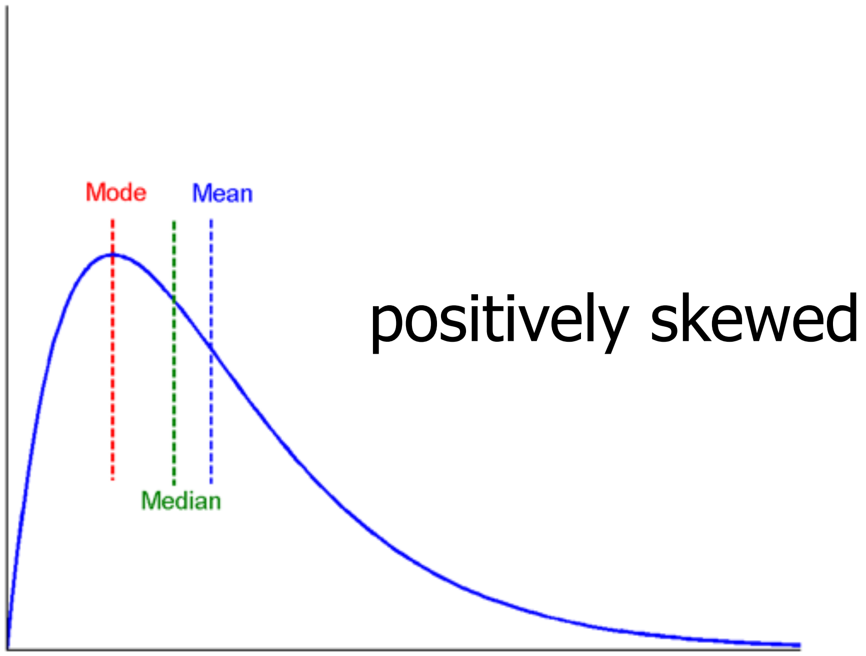


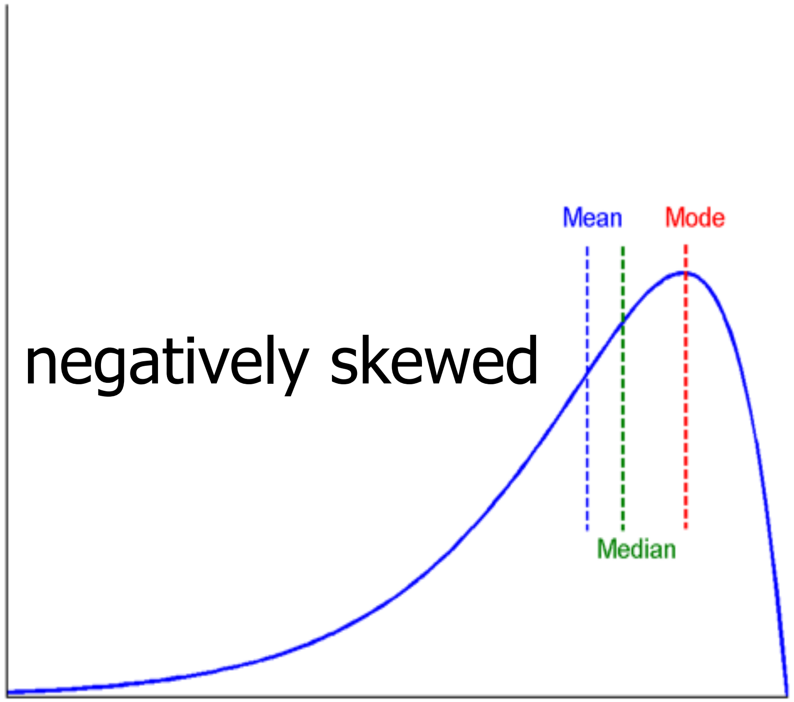


Symmetric vs. Skewed Data

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

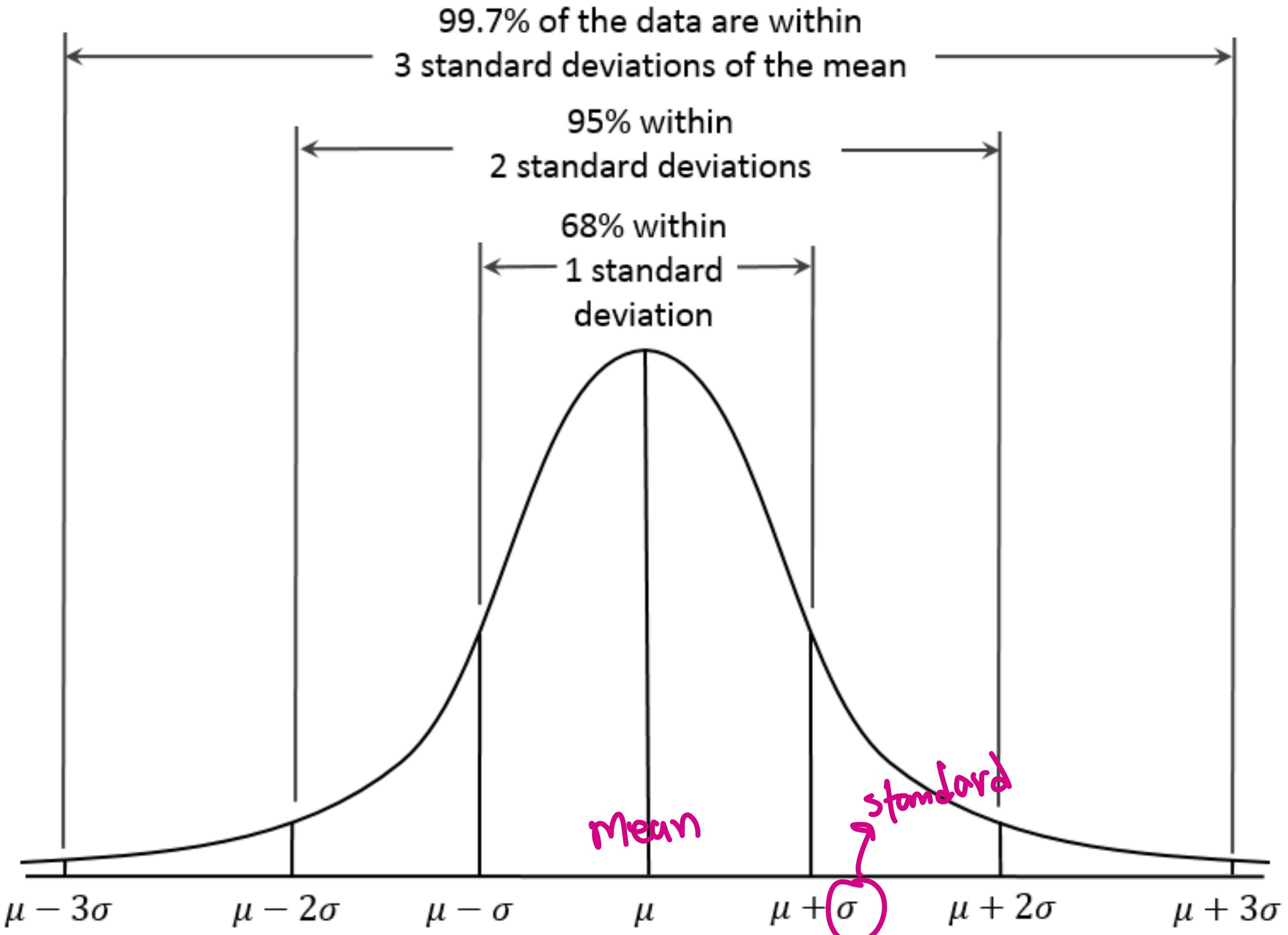






Properties of Normal Distribution Curve

← — ————Represent data dispersion, spread — ————→



Represent central tendency

Measures Data Distribution: Variance and Standard Deviation

- Variance and standard deviation (*sample: s, population: σ*)
  - **Variance**: (algebraic, scalable computation)
    - Q: Can you compute it incrementally and efficiently?

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

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

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

Graphic Displays of Basic Statistical Descriptions

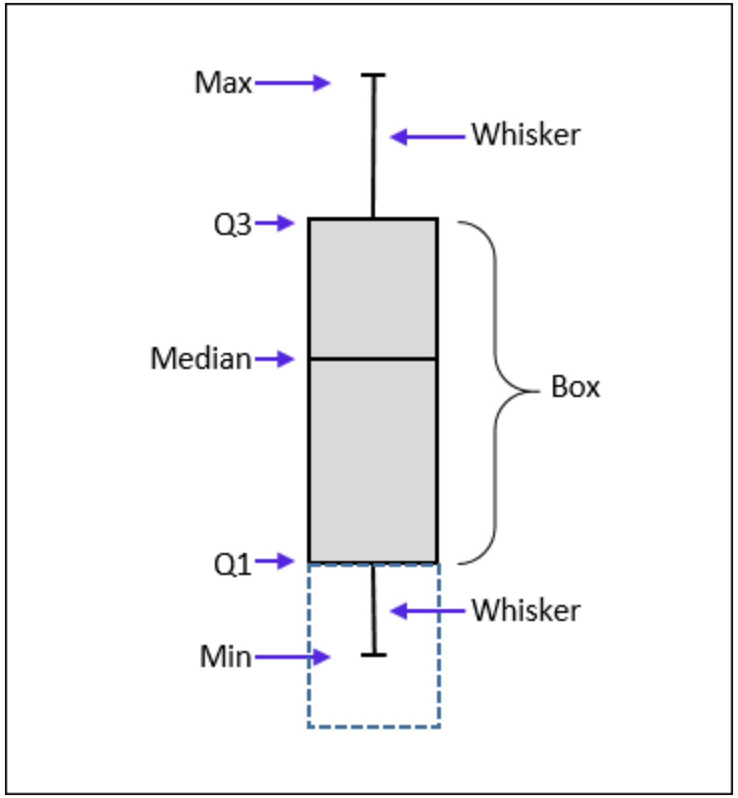
ใช้

- **Boxplot**: graphic display of five-number summary

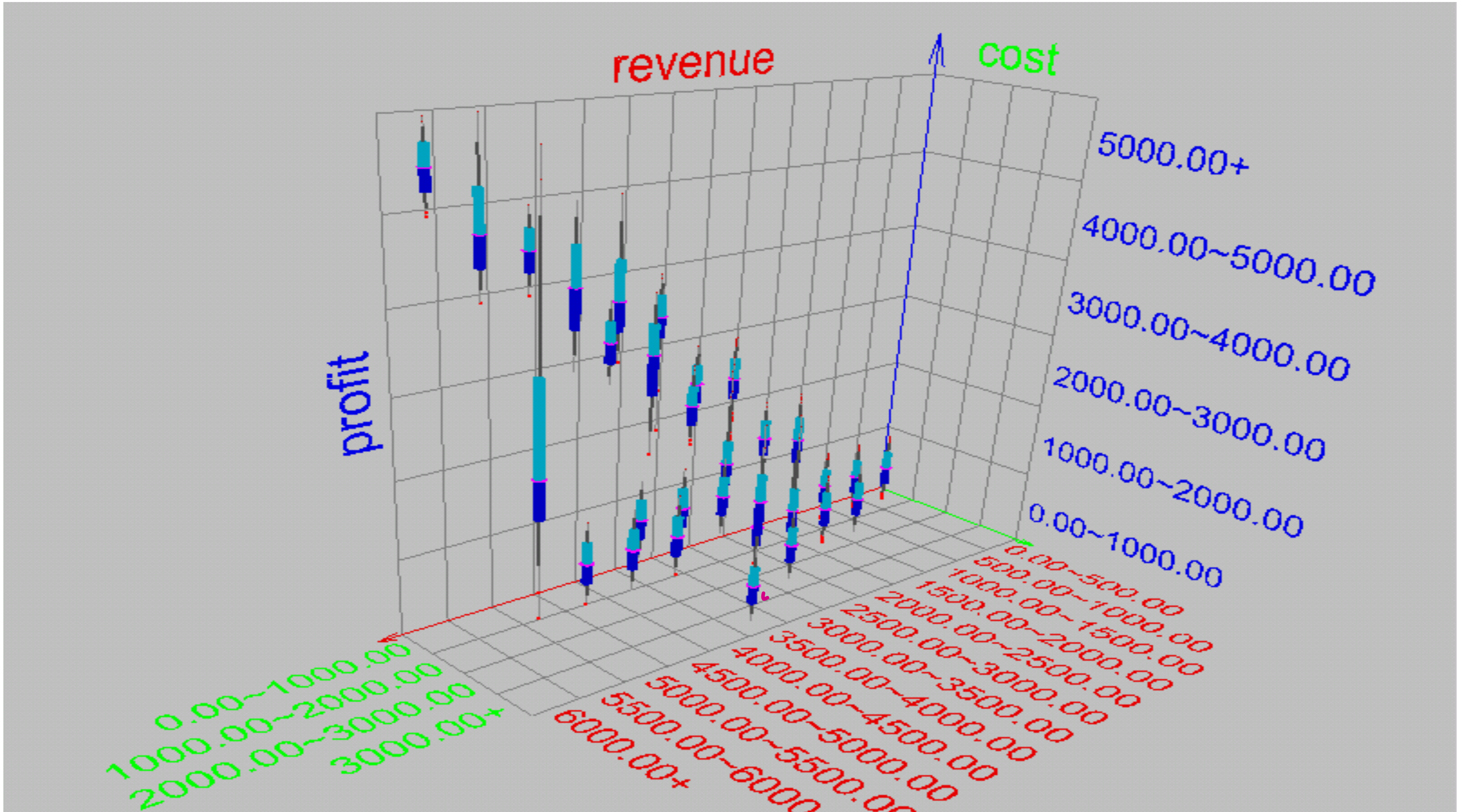
  ค.ถี่
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles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

Measuring the Dispersion of Data: Quartiles & 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**: Data is represented with a box
  - Q_1, Q_3, IQR: The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - Median (Q_2) 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
    - **Outlier**: usually, a value higher/lower than 1.5 x IQR



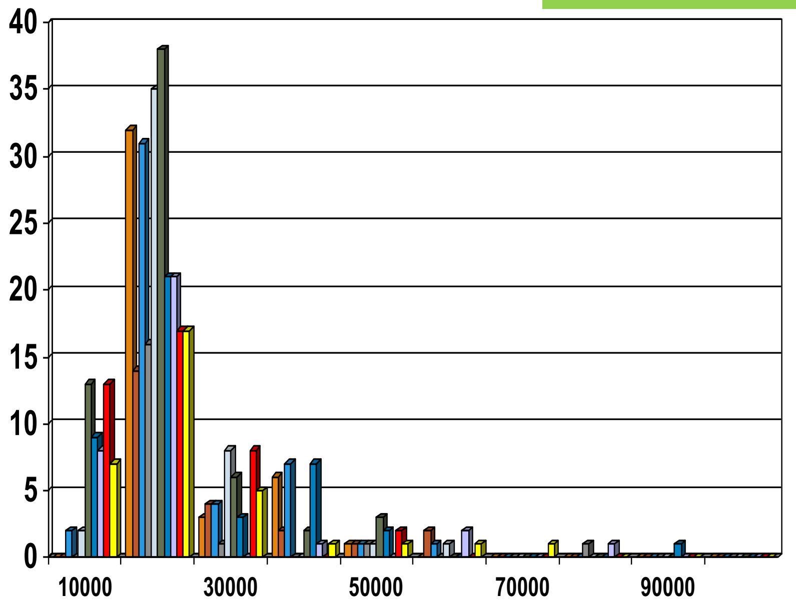
Visualization of Data Dispersion: 3-D Boxplots

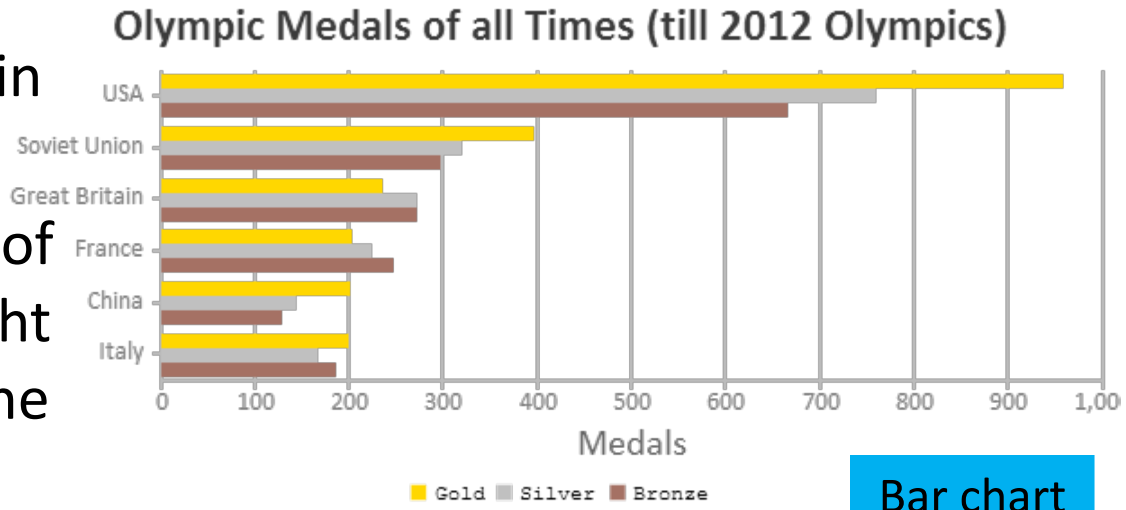


Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- Differences between histograms and bar charts
  - Histograms are used to show distributions of variables while bar charts are used to compare variables
  - Histograms plot binned quantitative data while bar charts plot categorical data
  - Bars can be reordered in bar charts but not in histograms
  - 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

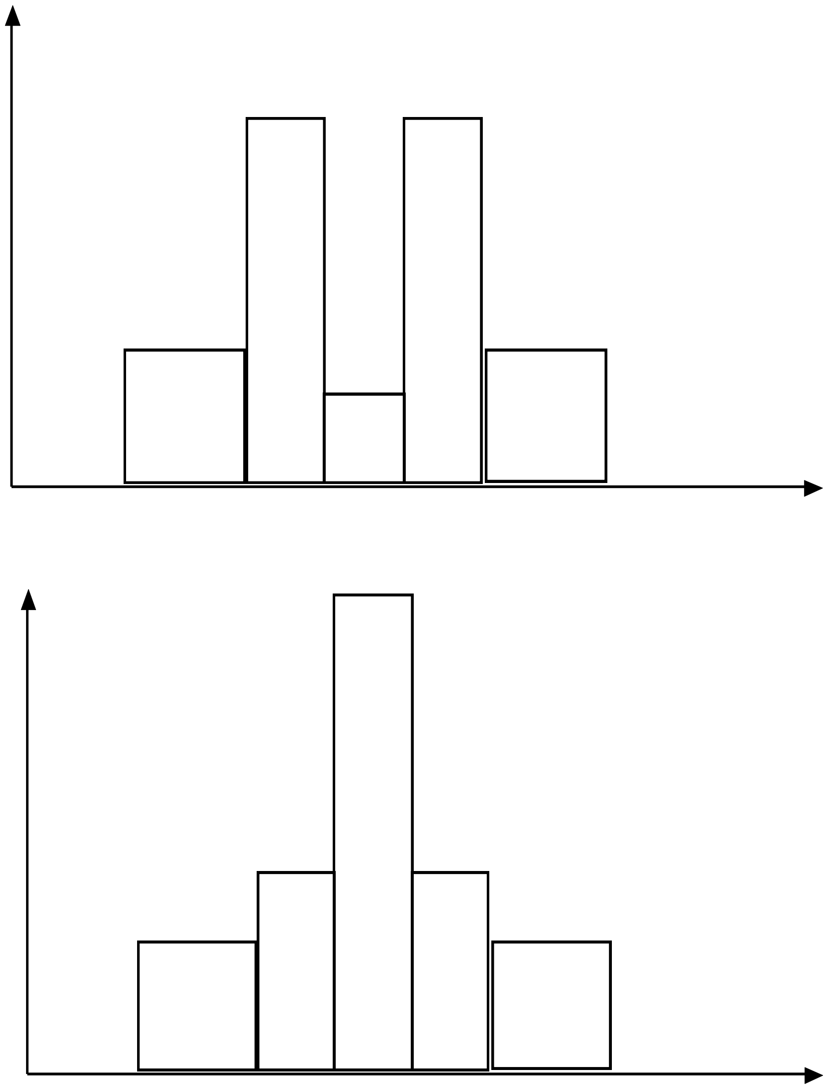
Histogram





Bar chart

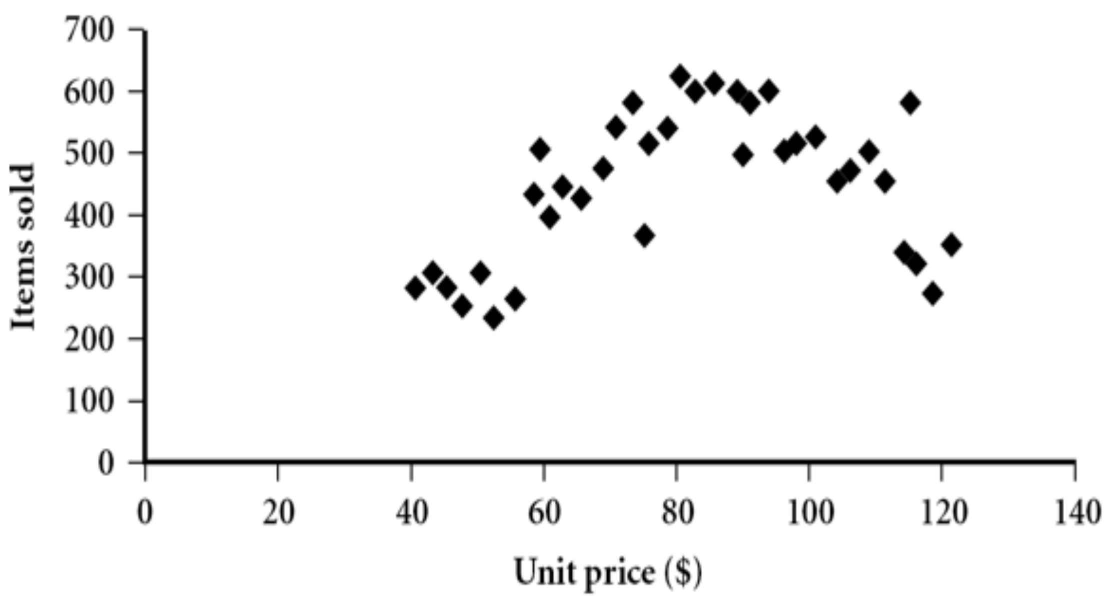
Histograms Often Tell More than Boxplots

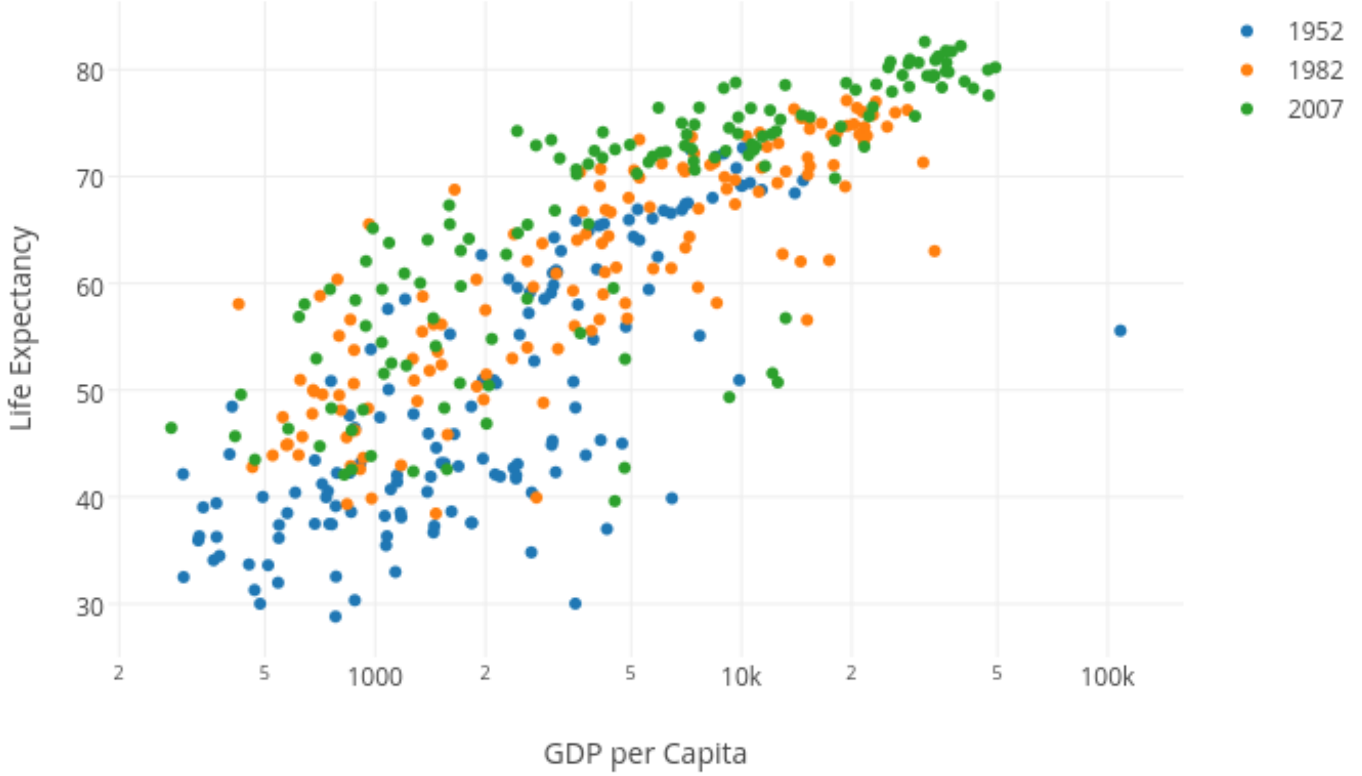


- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

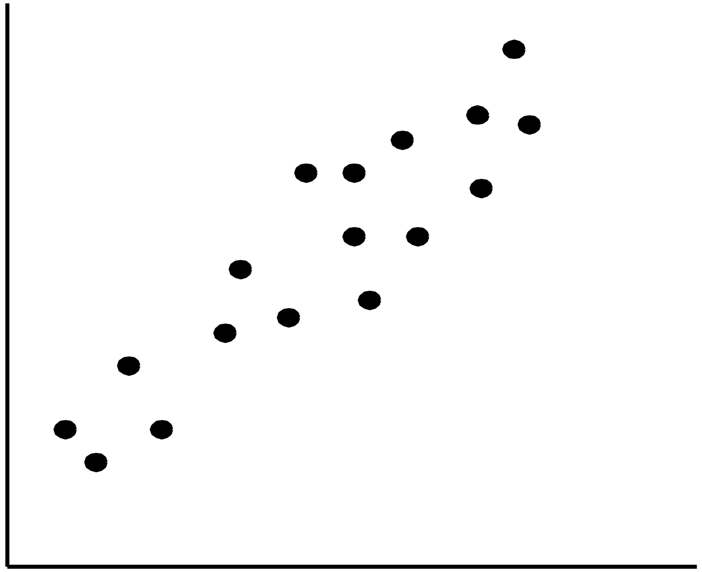
Scatter plot

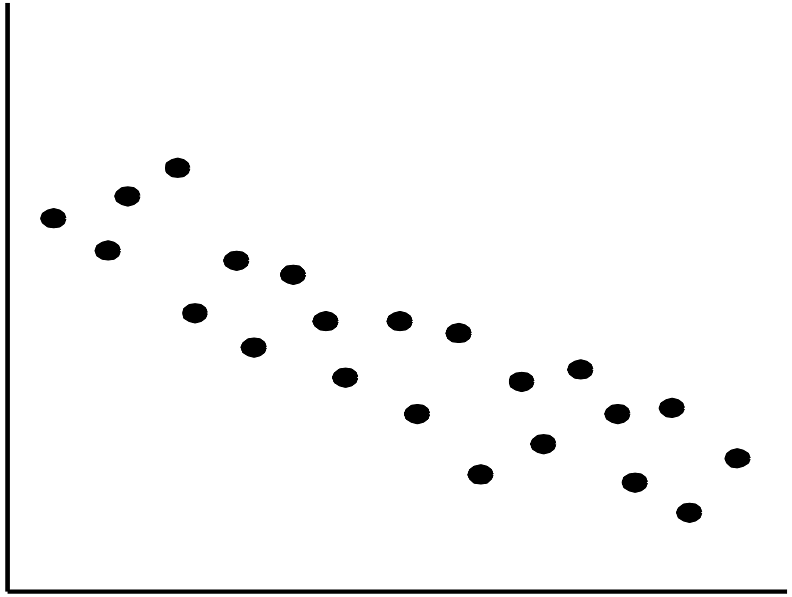
- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

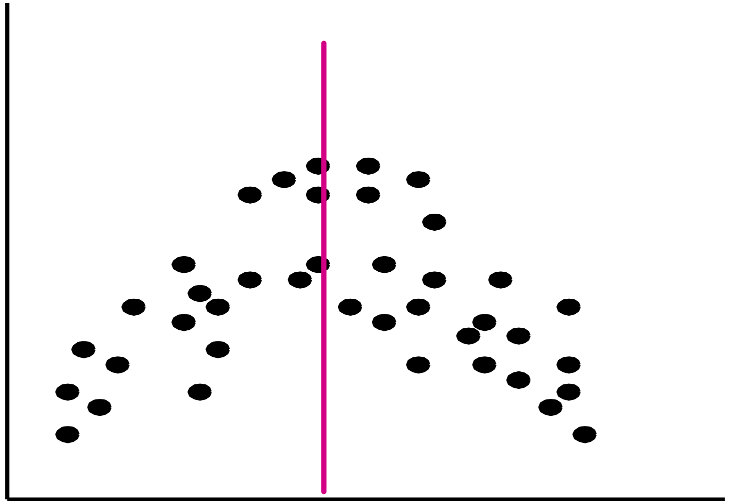




Positively and Negatively Correlated Data

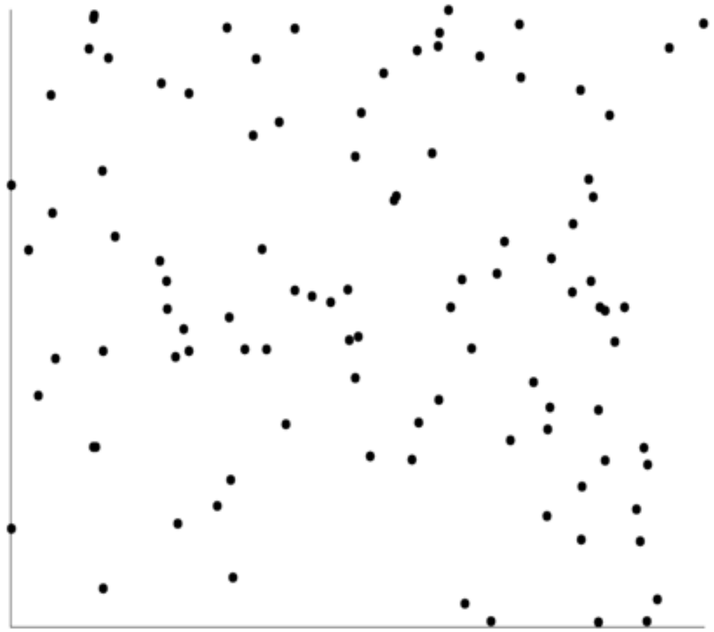


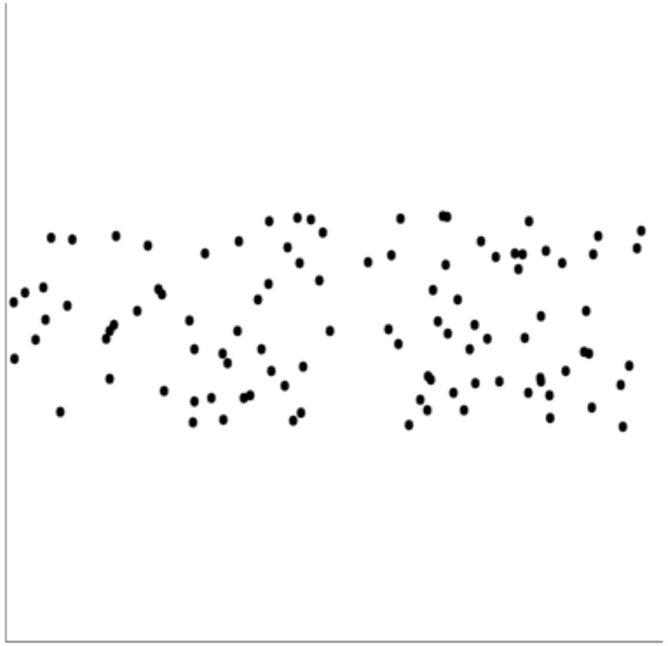


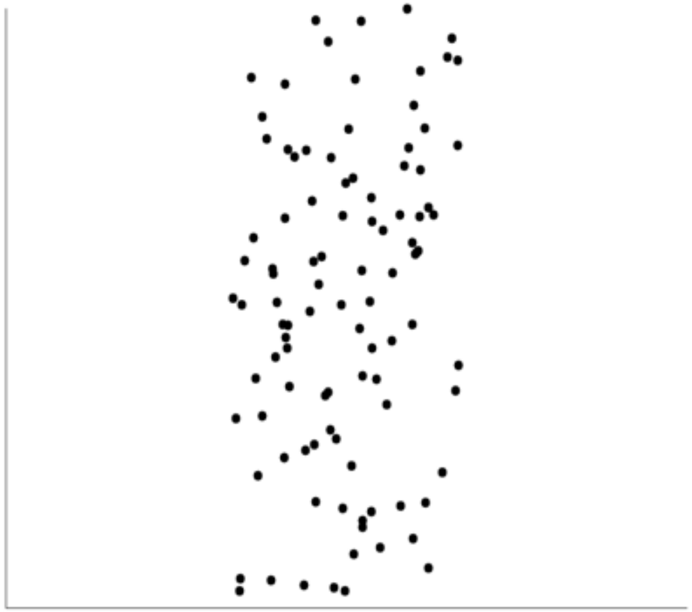


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data ไม่มีค.สัมพันธ์กัน







Chapter 2. Getting to Know Your Data

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

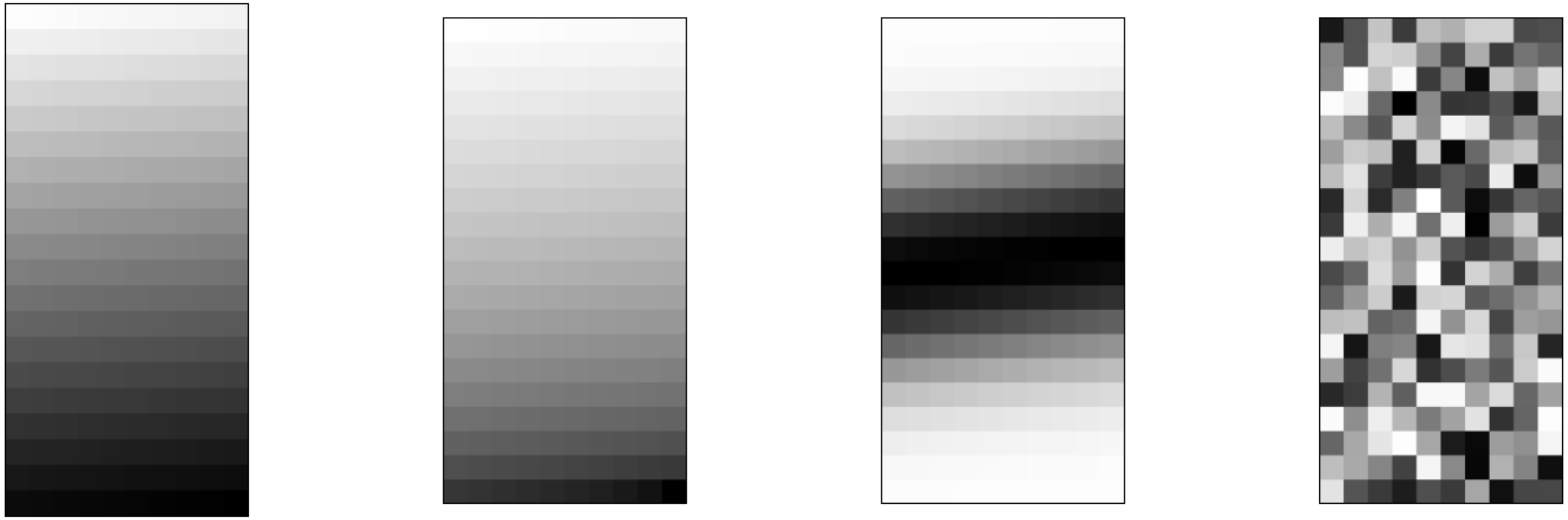
Data Visualization

- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

- For a data set of *m* dimensions, create *m* windows on the screen, one for each dimension
- The *m* dimension values of a record are mapped to *m* pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values

* plot data ด้วยสี



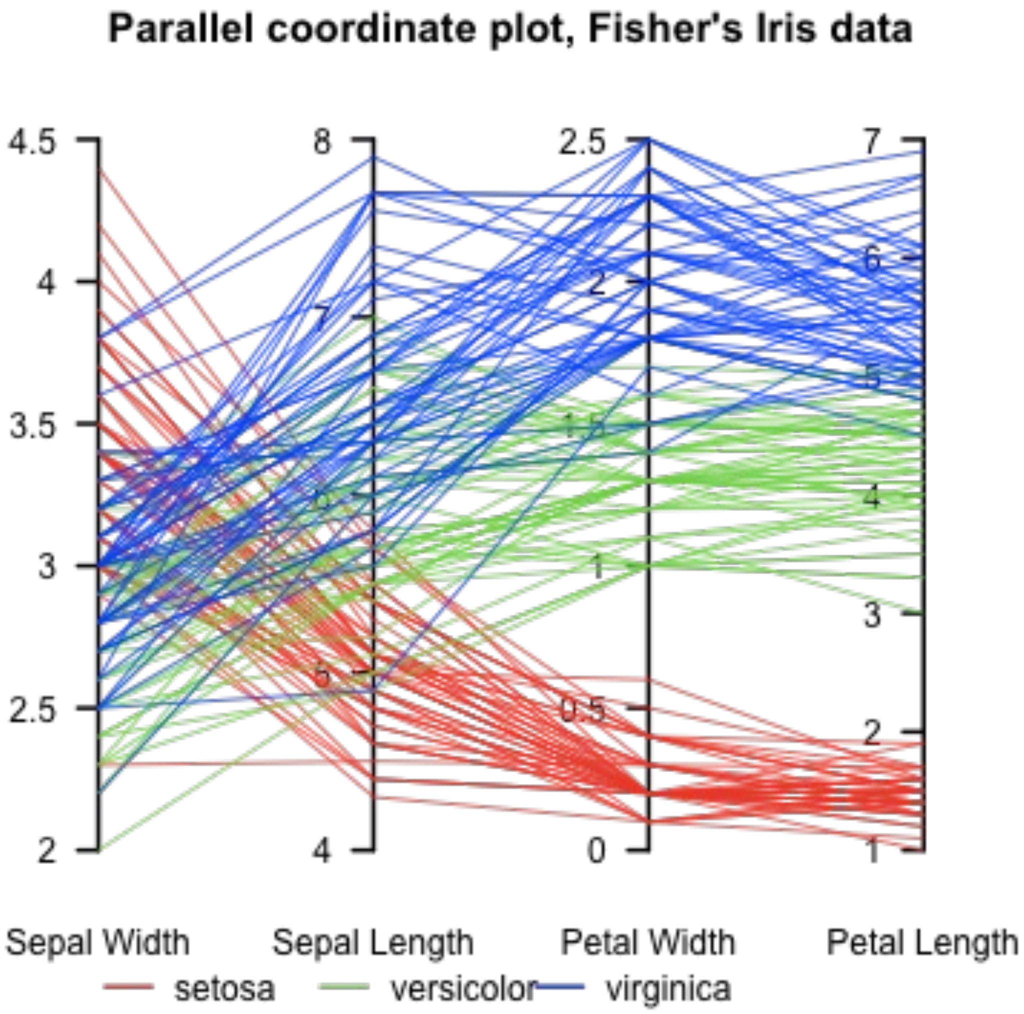
(a) Income (b) Credit Limit (c) transaction volume (d) age

Parallel Coordinates

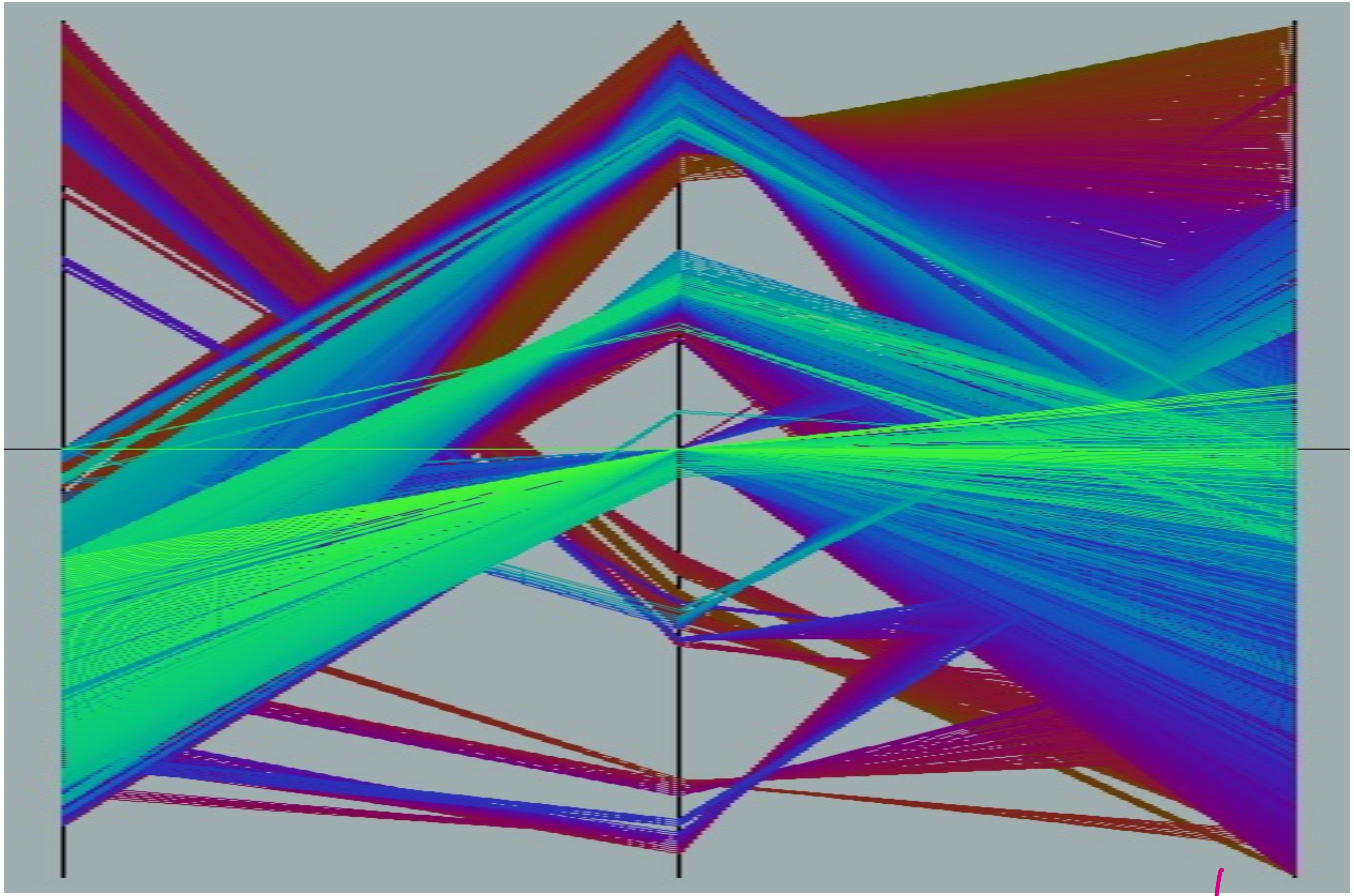
อธิบายคุณลักษณะดอกไม้

* แต่ละเส้นคือ ดอกไม้

- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute



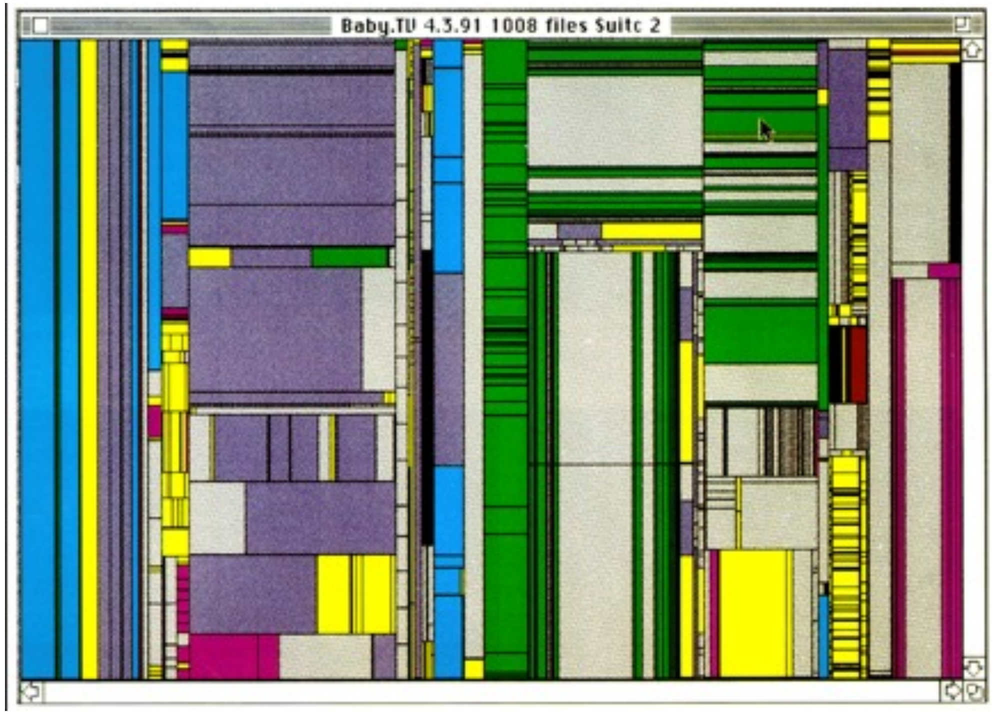
Parallel Coordinates of a Data Set



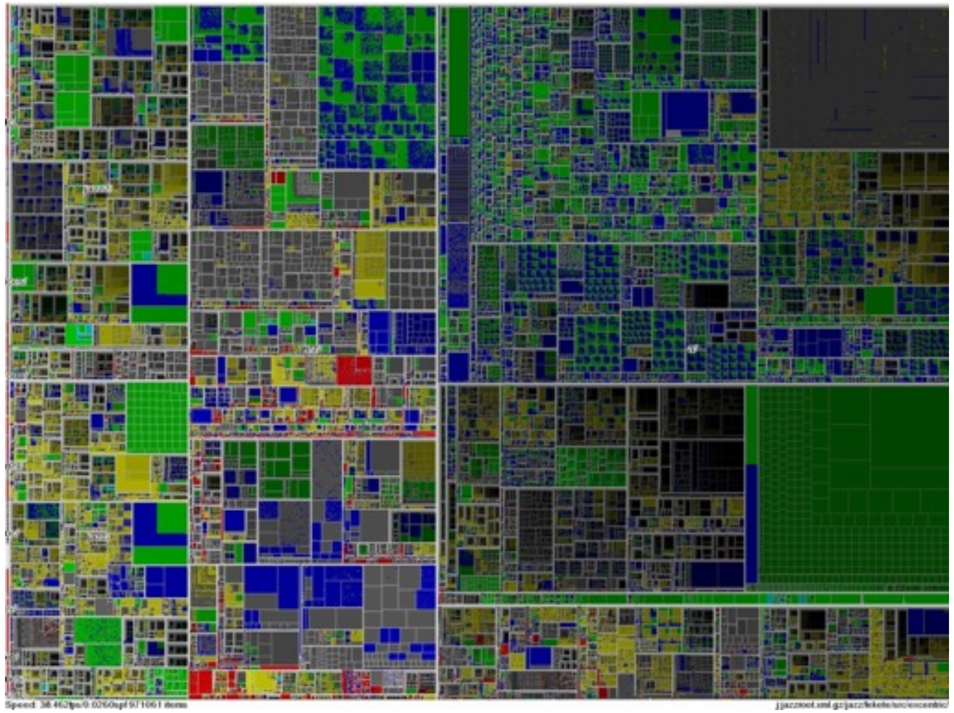
Tree-Map

แบ่งพื้นที่เป็น 100% แล้วแบ่งคลาส แล้วบอกแต่ละ
เซ็ตในคลาสนั้น แล้วบอกปริมาณด้วยขนาดของ
พื้นที่

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)



Schneiderman@UMD: Tree-Map of a File System



Schneiderman@UMD: Tree-Map to support large data sets of a million items

Visualizing Complex Data and Relations: Tag Cloud

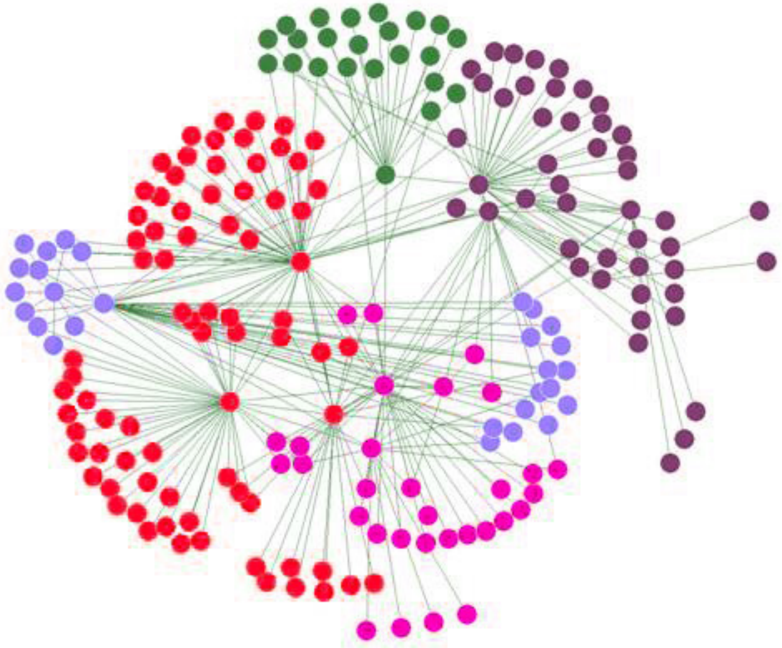
- Tag cloud: Visualizing user-generated tags
  - The importance of tag is represented by font size/color
  - Popularly used to visualize word/phrase distributions

KDD 2013 Research Paper Title Tag Cloud

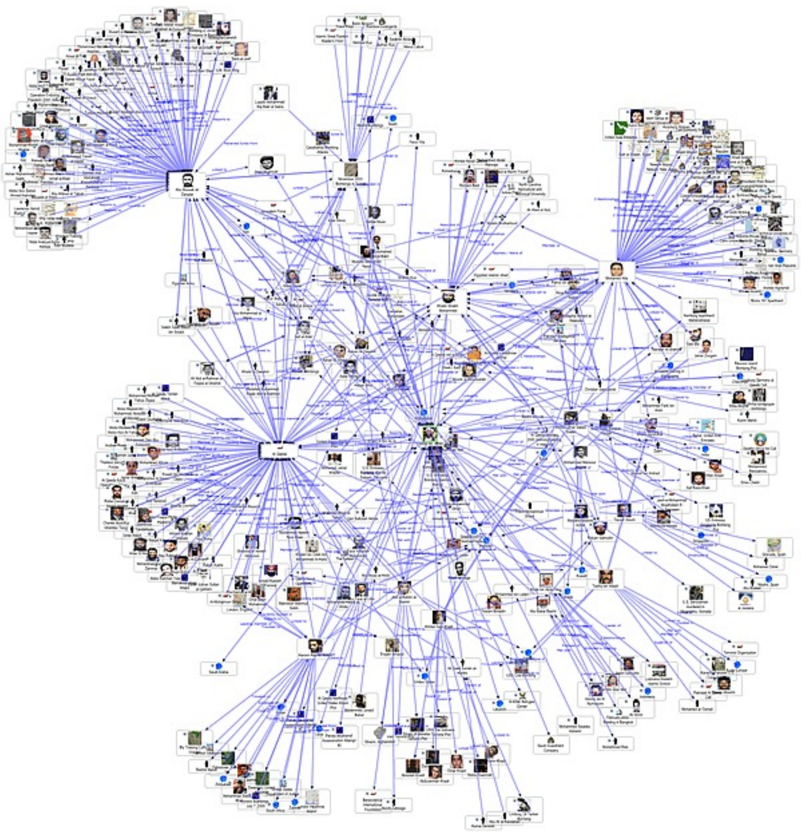
Newsmap: Google News Stories in 2005

Visualizing Complex Data and Relations: Social Networks

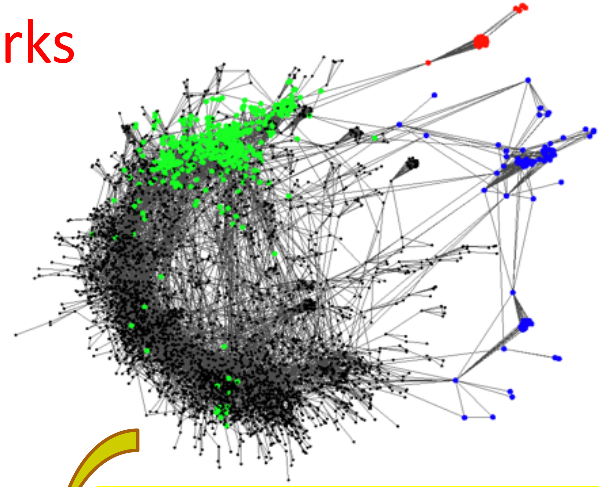
- Visualizing non-numerical data: social and information networks



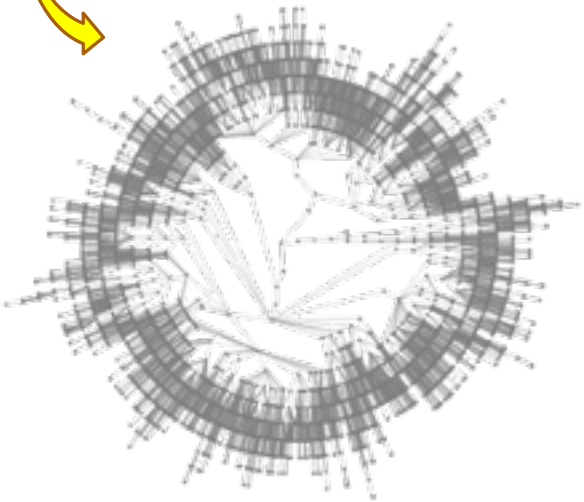
A typical network structure



A social network



organizing information networks



Data Mining

Chapter 2. Getting to Know Your Data

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

Similarity, Dissimilarity, and Proximity

- **Similarity measure** or **similarity function**
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike ดูค่าความเหมือน ยิ่งเหมือน ค่ายิ่งมาก
  - Often falls in the range [0,1]: 0: no similarity; 1: completely similar

สนใจ > Dissimilarity (คำนวณค่า ค.ห่าง)

- **Dissimilarity** (or **distance**) measure
  - Numerical measure of how different two data objects are
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)
  - Range [0, 1] or [0, ∞) , depending on the definition
- **Proximity** usually refers to either similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix
  - A data matrix of n data points with *l* dimensions

$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$$

- Dissimilarity (distance) matrix
  - n data points, but registers only the distance *d*(*i, j*) (typically metric)
  - Usually symmetric, thus a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal, ratio, and vector variables
  - Weights can be associated with different variables based on applications and data semantics

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

Standardizing Numeric Data

- Z-score: $$z = \frac{x - \mu}{\sigma}$$
  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, "+" when above
- An alternative way: Calculate the mean absolute deviation

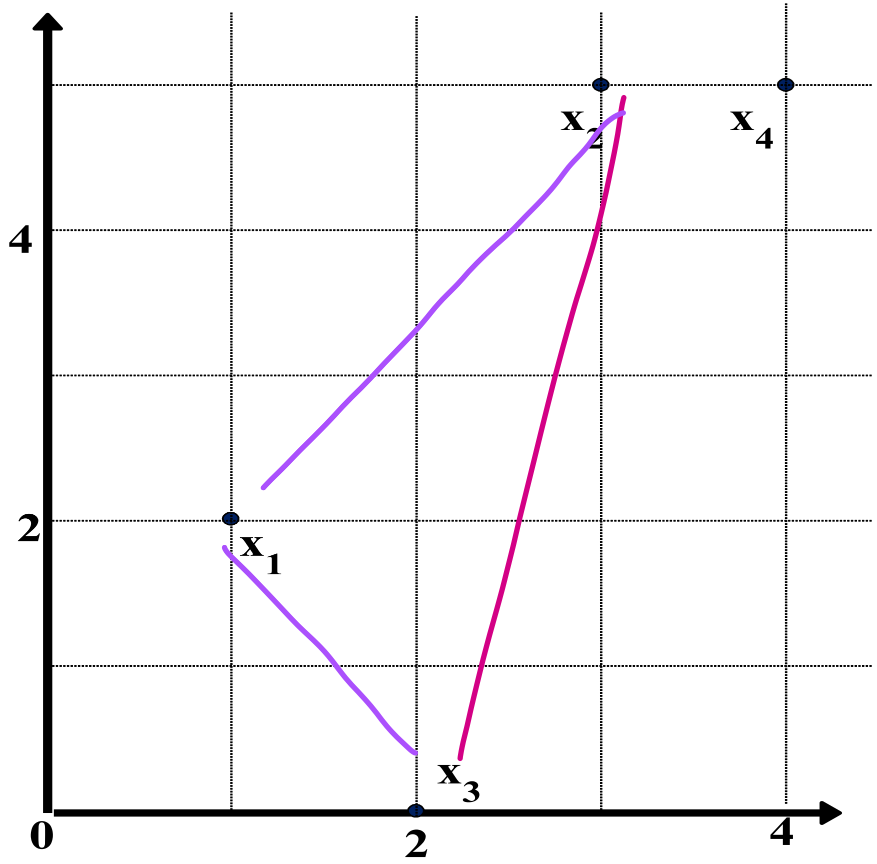
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

  - standardized measure (*z-score*): $$z_{if} = \frac{x_{if} - m_f}{s_f}$$
- Using mean absolute deviation is more robust than using standard deviation

Example: Data Matrix and Dissimilarity Matrix



$X_1 \rightarrow X_2, X_3$ ยาวกว่า $X_2 \rightarrow X_3$

Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

$z = \sqrt{x^2 + y^2}$ = หาความห่าง

Dissimilarity Matrix (by Euclidean Distance)

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

Distance on Numeric Data: Minkowski Distance

- Minkowski distance: A popular distance measure

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

  where $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ are two l-dimensional data objects, and p is the order (the distance so defined is also called L-p norm)

- Properties
  - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positivity)
  - $d(i, j) = d(j, i)$ (Symmetry)
  - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a metric
- Note: There are nonmetric dissimilarities, e.g., set differences

Special Cases of Minkowski Distance

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

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{il} - x_{jl}|$$

- p = 2: (L_2 norm) Euclidean distance หาง่ายสุด ?

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$$

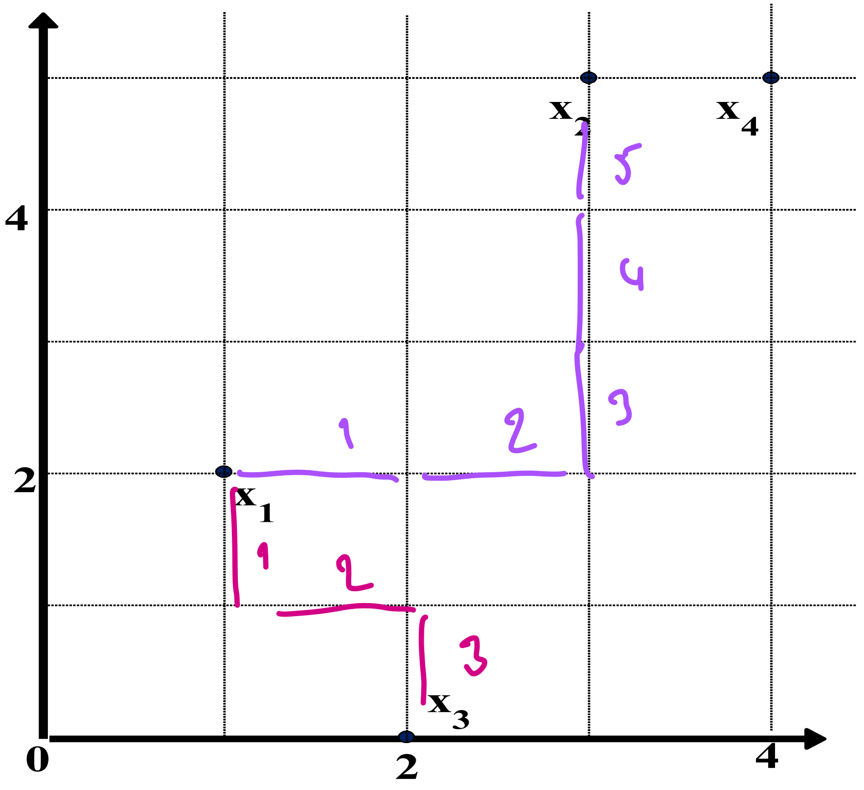
- $p \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{p\to\infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l} |x_{if} - x_{jf}|$$

Example: Minkowski Distance at Special Cases

Data

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



Manhattan (L_1)

มองความห่างเป็น box

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

Euclidean (L_2)

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

Supremum (L_∞)

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

Proximity Measure for Binary Attributes

- A contingency table for binary data

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

- Distance measure for symmetric binary variables $$d(i, j) = \frac{r+s}{q+r+s+t}$$

โอกาสเกิ ดุ่

- Distance measure for asymmetric binary variables: $$d(i, j) = \frac{r+s}{q+r+s}$$
- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables): $$sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$$
- Note: Jaccard coefficient is the same as

(a concept discussed in Pattern Discovery)

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q+r)+(q+s)-q}$$

Example: Dissimilarity between Asymmetric Binary Variables

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 (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
- ex Distance: $d(i, j) = \frac{r + s}{q + r + s}$

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

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

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

Jack (rows) vs. Mary (columns):

	1	0	Σ_{row}
1	2	0	2
0	1	3	4
Σ_{col}	3	3	6

Jack (rows) vs. Jim (columns):

	1	0	Σ_{row}
1	1	1	2
0	1	3	4
Σ_{col}	2	4	6

Jim (rows) vs. Mary (columns):

	1	0	Σ_{row}
1	1	1	2
0	2	2	4
Σ_{col}	3	3	6

จัดกลุ่ม

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes
  - Example: Color (red, yellow, blue, green), profession, etc.

วิธี 1

- 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

Dummy Variable

one-hot-Encode

ex

C	N	R	B
black	Thai ✓	B ✓	O
blue	Thai ✓	B ✓	A

$$d(i,j) = \frac{p-m}{p} \qquad m = 2 \qquad p = 4$$

$$= \frac{4-2}{4}$$

$$= 0.5 \ \#$$

Ordinal Variables

C: [black, blue , green)

C

-black → C-black | C-blue

-blue

C-black	C-blue
1	0
0	1

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
  - Replace *an ordinal variable value* by its rank: $r_{if} \in \{1, ..., M_f\}$
  - Map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
    - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1

ที่เป็นได้ด้าน

    - Then distance: d(freshman, senior) = 1 (มากสุด), d(junior, senior) = 1/3
  - Compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

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

ค่าเฉลี่ยแบบถ่วงน้ำหนัก

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

  - If f is numeric: Use the normalized distance
  - If f is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; or $d_{ij}^{(f)} = 1$ otherwise
  - If f is ordinal
    - Compute ranks z_{if} (where $z_{if} = \frac{r_{if} - 1}{M_f - 1}$)
    - Treat z_{if} as interval-scaled

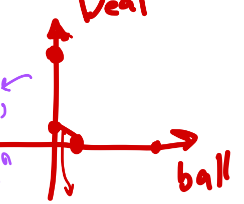
Cosine Similarity of Two Vectors

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document

ตารางนับคำในบทความ

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

*ปัญหา



- Other vector objects: Gene features in micro-arrays

ใช้แบ่งสัตว์, แบ่งพืช

- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

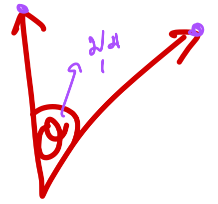
เวกเตอร์ d

ขนาดของเวกเตอร์

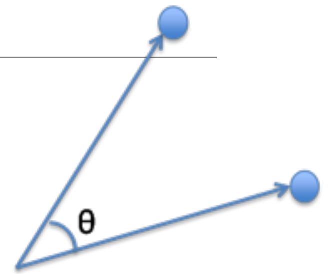
*Cosine Sim

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

วัดความห่างผ่านมุม



Example: Calculating Cosine Similarity



- Calculating Cosine Similarity:

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

where • indicates vector dot product, ||*d*||: the length of vector *d*

Fuctor Vactor

- 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)

  - First, calculate vector dot product

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

  - Then, calculate $||d_1||$ and $||d_2||$

$$\| d_1 \| = \sqrt{5\times5+0\times0+3\times3+0\times0+2\times2+0\times0+0\times0+2\times2+0\times0+0\times0} = 6.481$$

$$\| d_2 \| = \sqrt{3\times3+0\times0+2\times2+0\times0+1\times1+1\times1+0\times0+1\times1+0\times0+1\times1} = 4.12$$

หาระยะห่างของ 2 เวกเตอร์

  - Calculate cosine similarity: cos(d_1, d_2) = 25/ (6.481 X 4.12) = 0.94

Chapter 2. Getting to Know Your Data

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

การวัดระยะห่าง ของข้อมูล (สำคัญ)

Summary

ตัวแปร data ↗ เรียงลำดับได้

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled

ความหลากหลายของ data

- 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
  - Data visualization: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing
- Many methods have been developed but still an active area of research

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu, et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009