

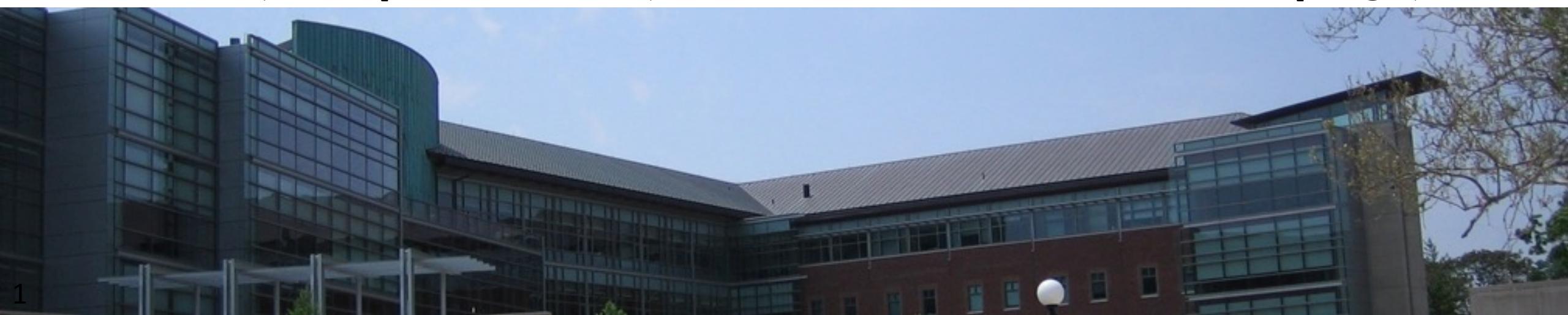


CS 412 Intro. to Data Mining

- ① Data visualization
- ② Mining - many ways to do DATA
- ③ APP 10% of 10%

Chapter 2. Getting to Know Your Data

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





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



Types of Data Sets: (1) Record Data

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

	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

- ## Transaction data

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk 1 ග්‍රෑසු ප්‍රිය
2 - ප්‍රිය සාම්පූහ්‍රේ	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

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

Relational database

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

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

normalize
consistency
① Violation
② Inconsistency
conflict
map violation
inconsistent

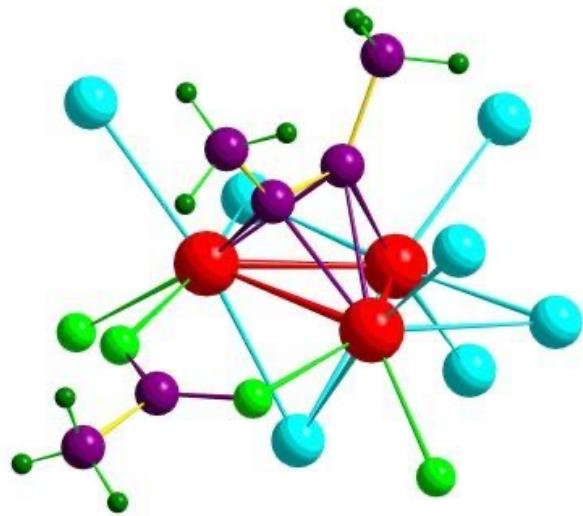
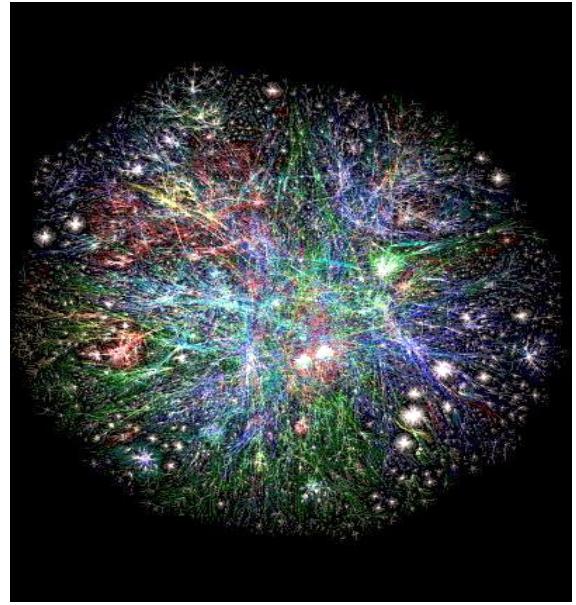
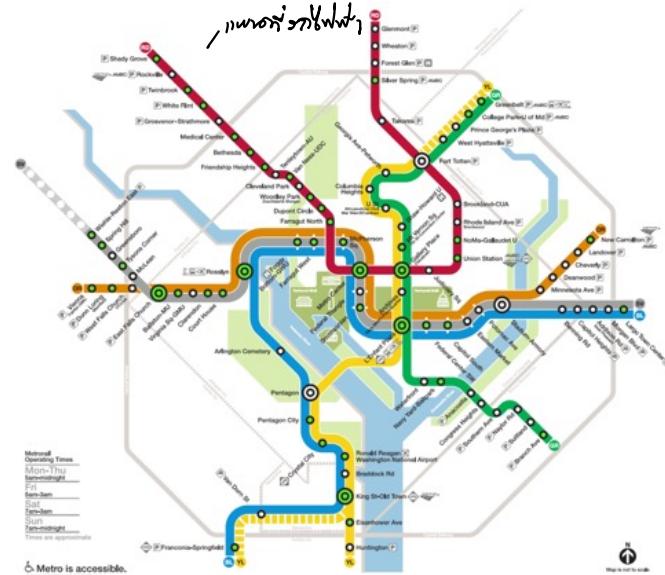
team	coach	play	ball	score	game	lost	win	timeout	season
Document 1	3	0	5	0	2	6	0	2	0
Document 2	0	7	0	2	1	0	0	3	0
Document 3	0	1	0	0	1	2	2	0	3

28 data
Matrix

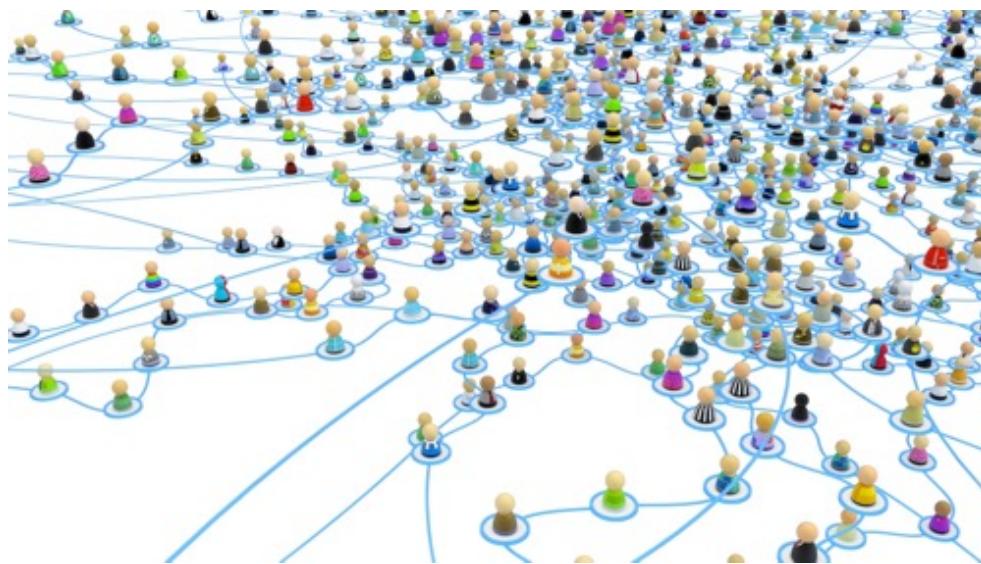
Types of Data Sets: (2) Graphs and Networks

data network

- Transportation network



- Molecular Structures
- Social or information networks



social Network

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

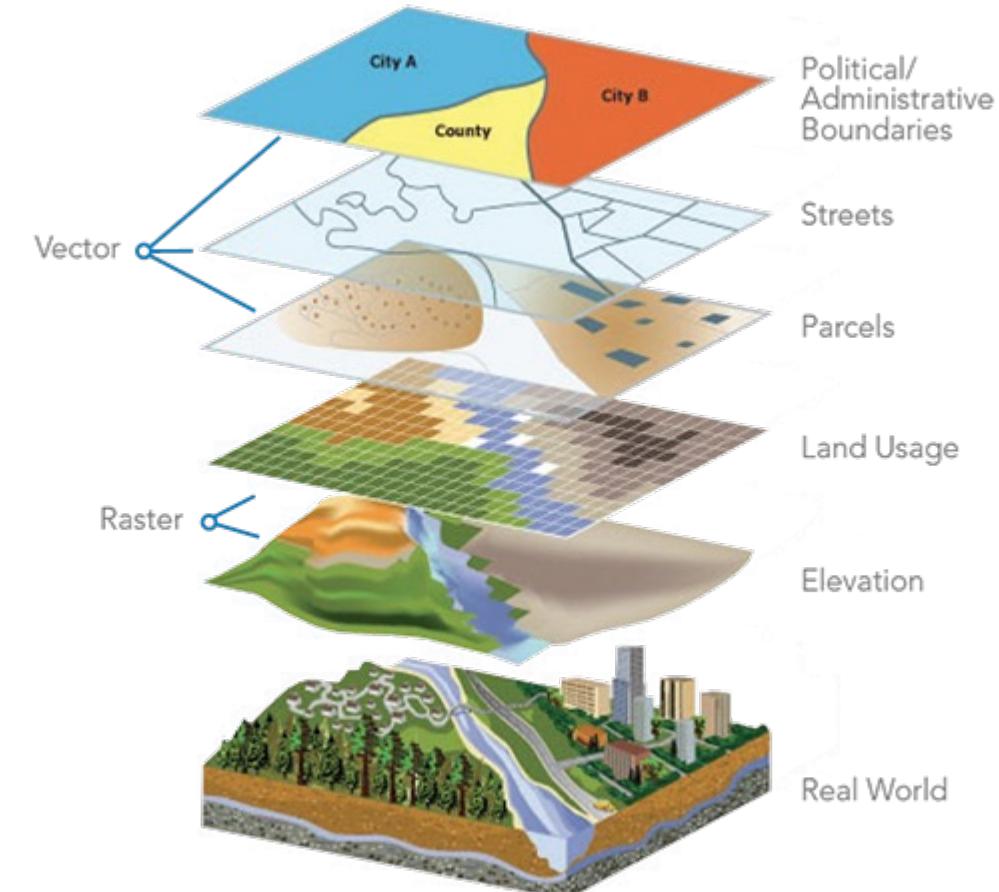
↙ ↗
Data sets → Data sets

Spatial data: maps



Image data:

Video data: Spatial temporal dimension



Important Characteristics of Structured Data

- ✓ Attribute
- Dimensionality 维数
 - Curse of dimensionality 维度诅咒
- Sparsity 稀疏性
 - Only presence counts
- Resolution 尺度
 - $\text{如} \rightarrow \text{厘米 cm, m} \rightarrow \text{模型建模}$
 - Patterns depend on the scale
 - ex. $\text{mm} \rightarrow \text{米 m} \rightarrow \text{千米 km} \rightarrow \text{ML 预测值 取决于 scale}$
- Distribution 分布
 - $\text{如} \rightarrow \text{模型} \rightarrow \text{正态分布} \rightarrow \text{均值 mean, 标准差 standard deviation}$
 - $\text{mean} = \text{均值}$
 - $\text{std} = \text{标准差}$

Data Objects

- Data sets are made up of data objects
 - ↗ data point → තුළ නො ඇති
- A **data object** represents an entity
 - ↗ entity
- Examples:
 - ↗ සාල්සා හෝ ප්‍රතික්‍රියා අංශය
 - ↗ if ML is own Feature
- 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*
- Types:
 - Nominal (e.g., red, blue)
 - Binary (e.g., {true, false})
 - Ordinal (e.g., {freshman, sophomore, junior, senior})
 - 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?

interval scaled
Nominal, Numeric

Attribute Types

A	B	C	D	F	ordinal
ordN	3	7	4	numer	

- **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

Yao-Gin Yiu (Yao-Hui Liu)

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 1.1.5... 2

- 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

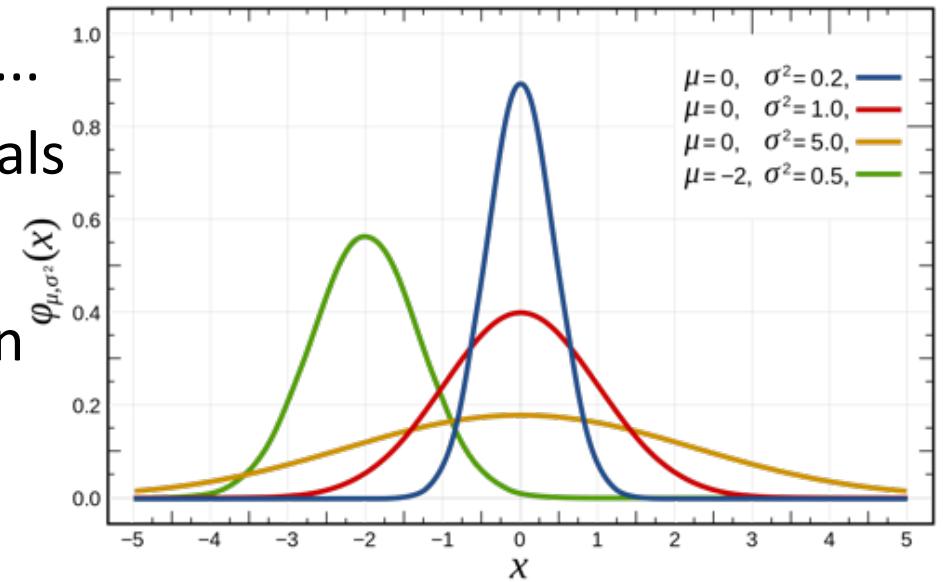
Chapter 2. Getting to Know Your Data

min max mean Q1 Q2 Q3

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

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.

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

- 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 ($L_2 - L_1$)

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

Low interval limit

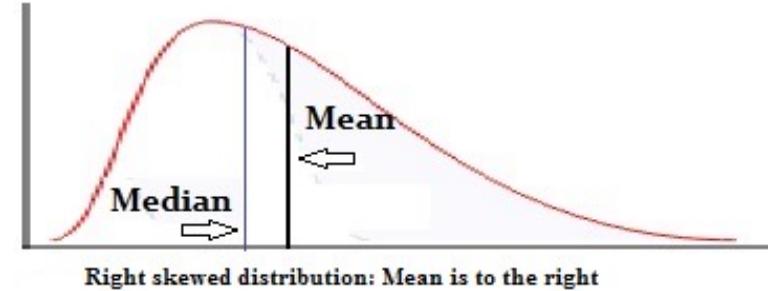
Measuring the Central Tendency: (3) Mode

- Mode: Value that occurs most frequently in the data

- Unimodal

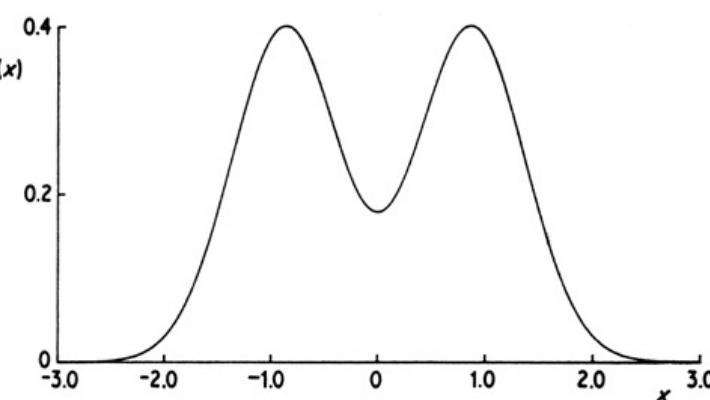
- Empirical formula:

$$\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$$

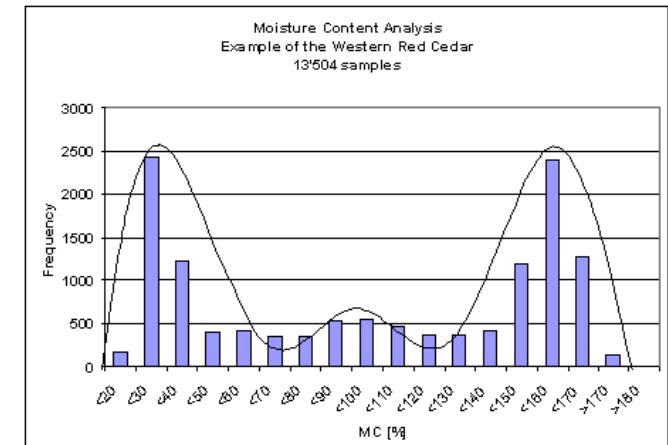
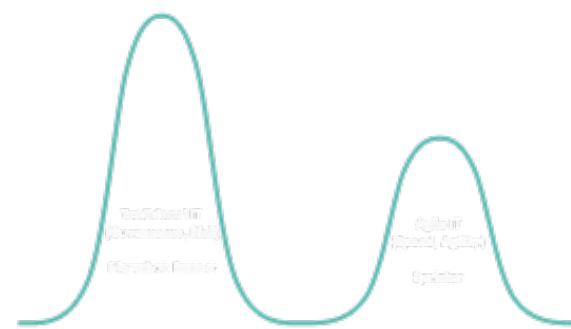


- Multi-modal

- Bimodal



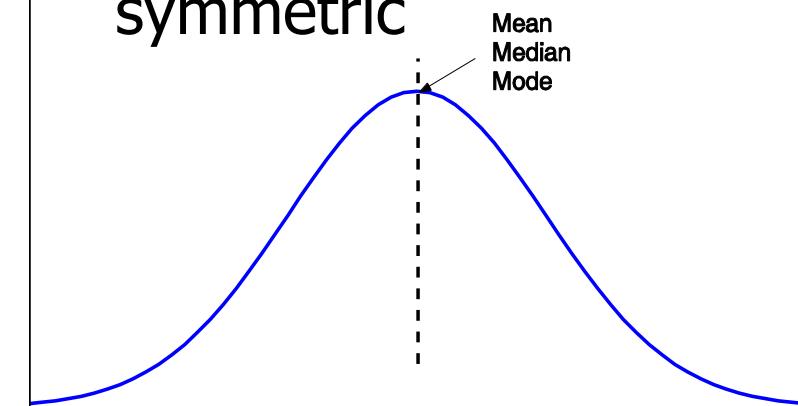
- Trimodal



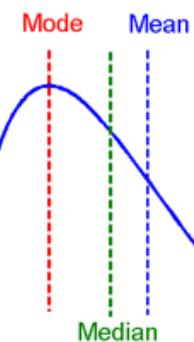
Symmetric vs. Skewed Data

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

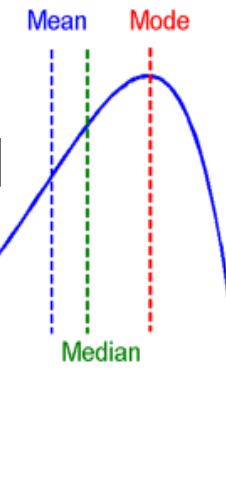
symmetric



positively skewed

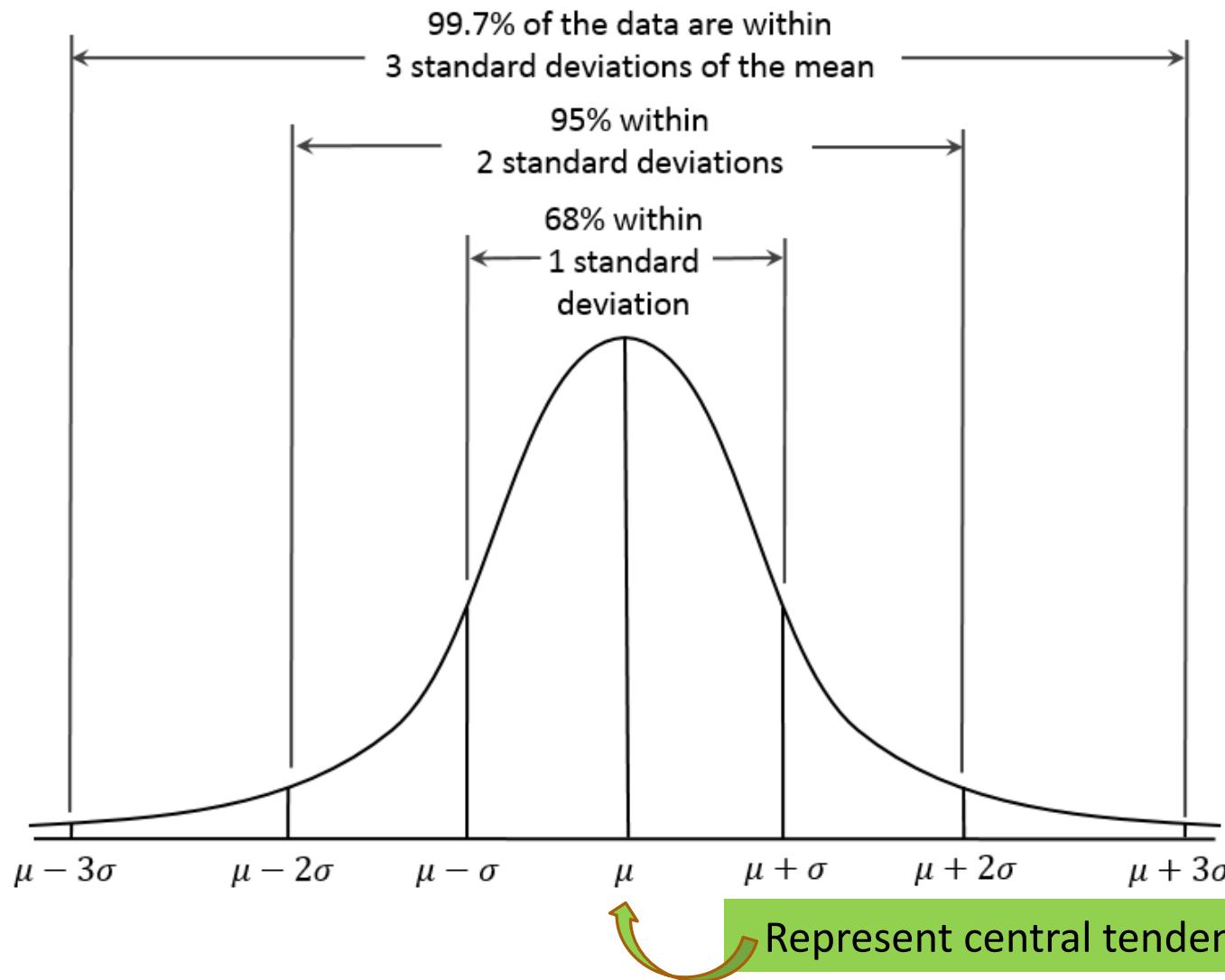


negatively skewed



Properties of Normal Distribution Curve

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



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?

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

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

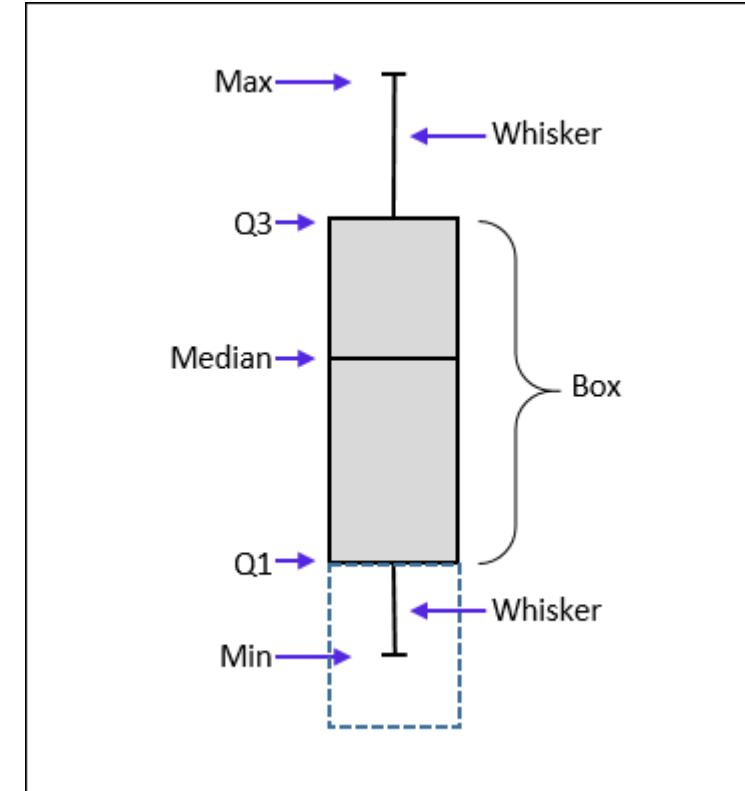
- ❑ **Standard deviation s (or σ)** is the square root of variance s^2 (or σ^2)

Graphic Displays of Basic Statistical Descriptions

- **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 $100 f_i \%$ 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

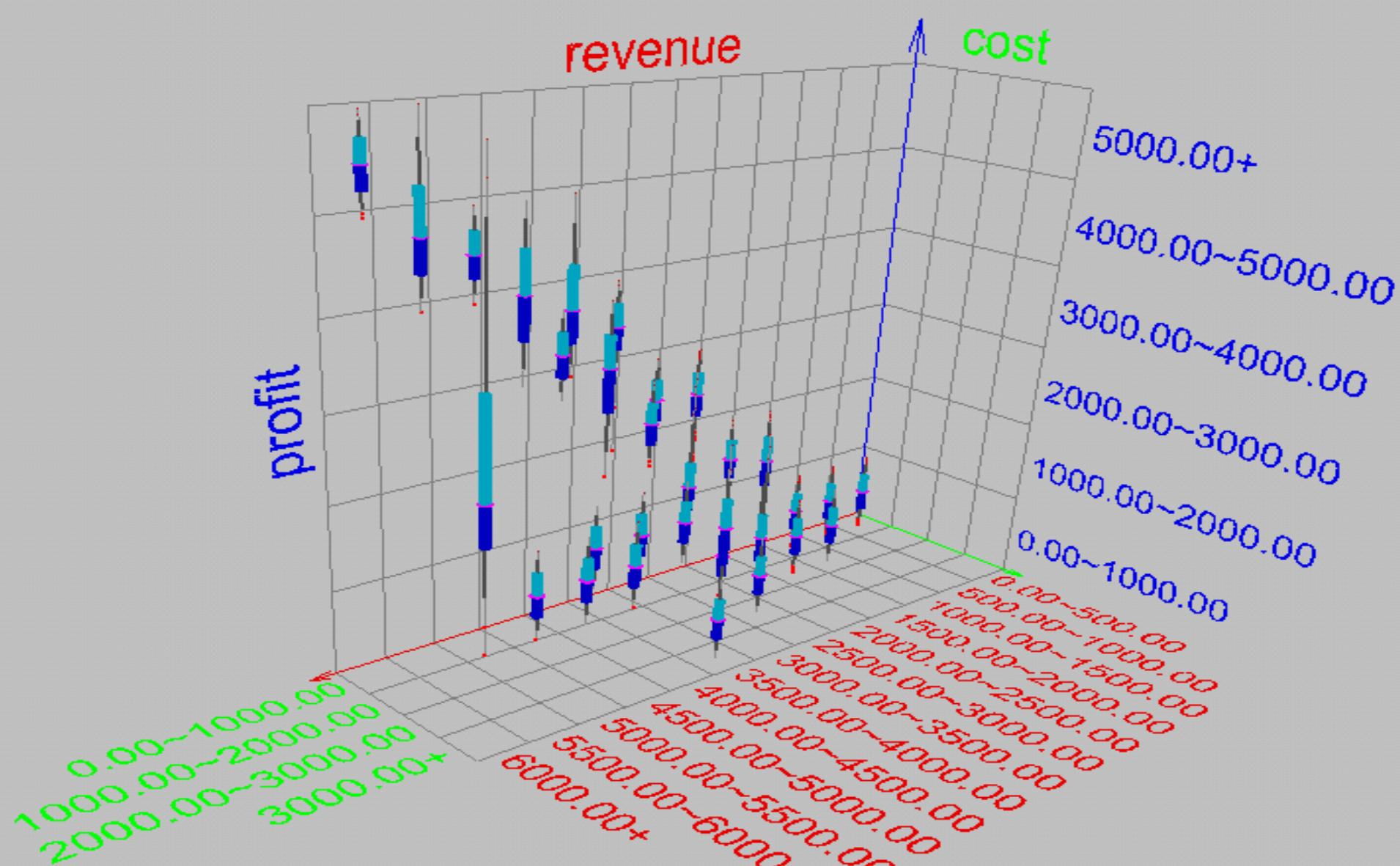
- **Quartiles:** Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
- **Inter-quartile range:** $\text{IQR} = Q_3 - Q_1$
- **Five number summary:** min, Q_1 , median, Q_3 , max
- **Boxplot:** Data is represented with a box
 - Q_1 , Q_3 , IQR: The ends of the box are at the first and third quartiles, 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 $1.5 \times \text{IQR}$



outlier
Q1, Q3

Visualization of Data Dispersion: 3-D Boxplots

Scatterplot 01.15.2022

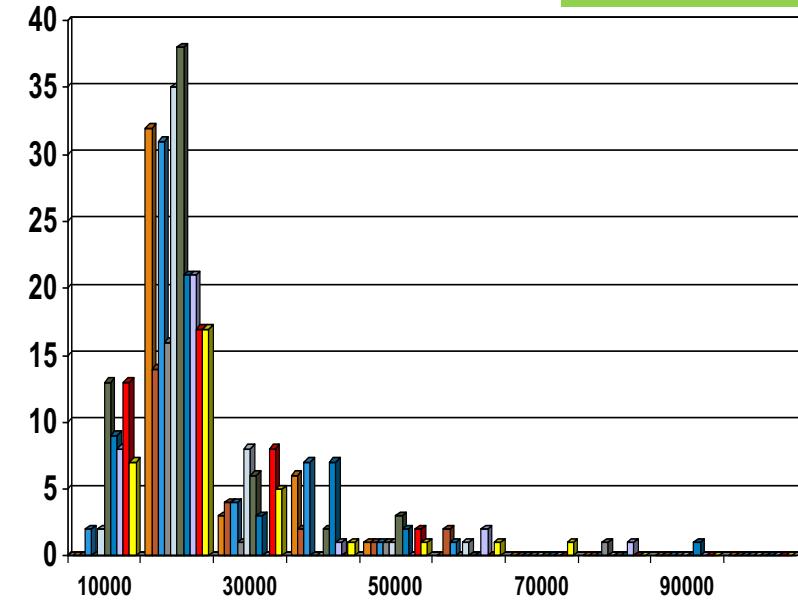


Histogram Analysis

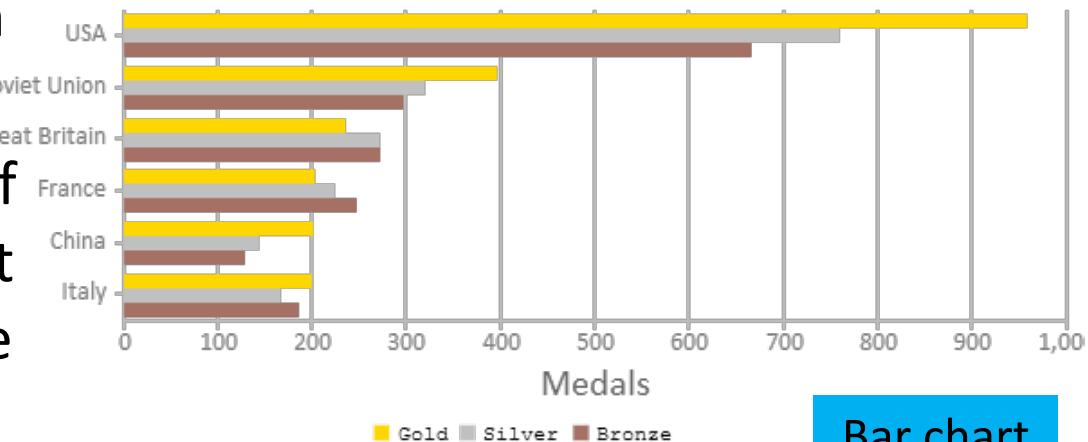
ఫ్రాక్టల్ ట్రైప్

Histogram

- ❑ 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

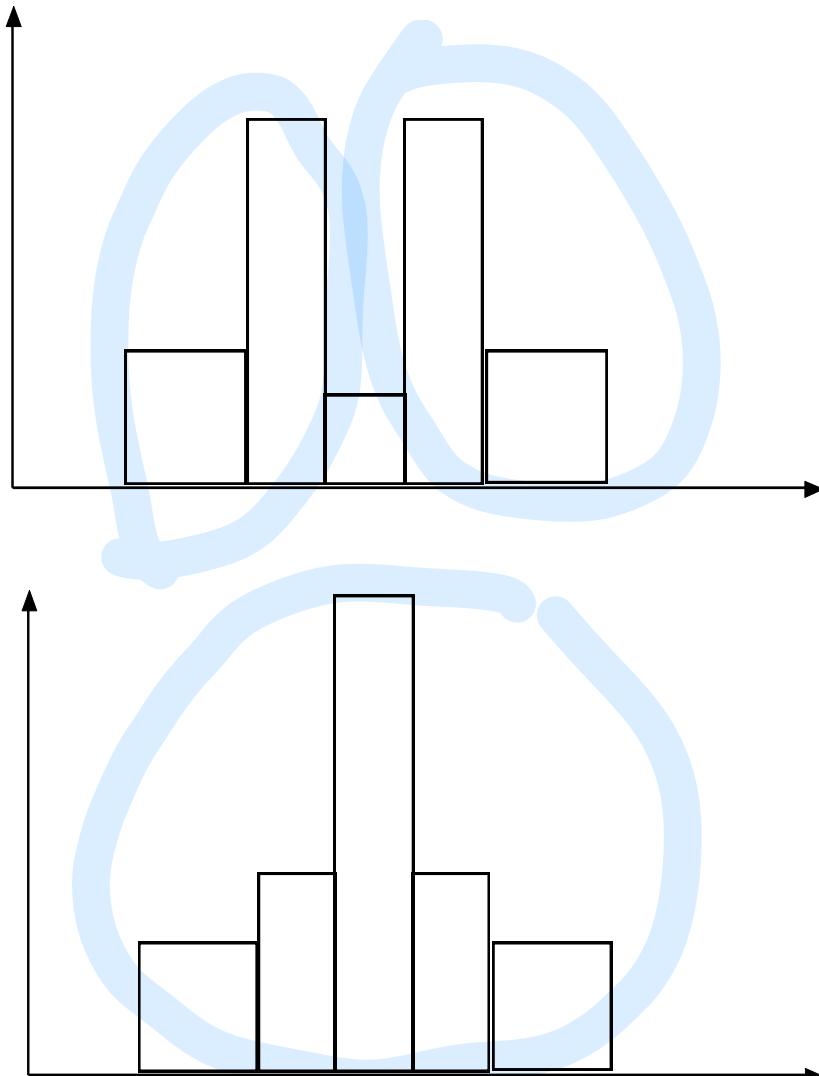


Olympic Medals of all Times (till 2012 Olympics)



Bar chart

Histograms Often Tell More than Boxplots



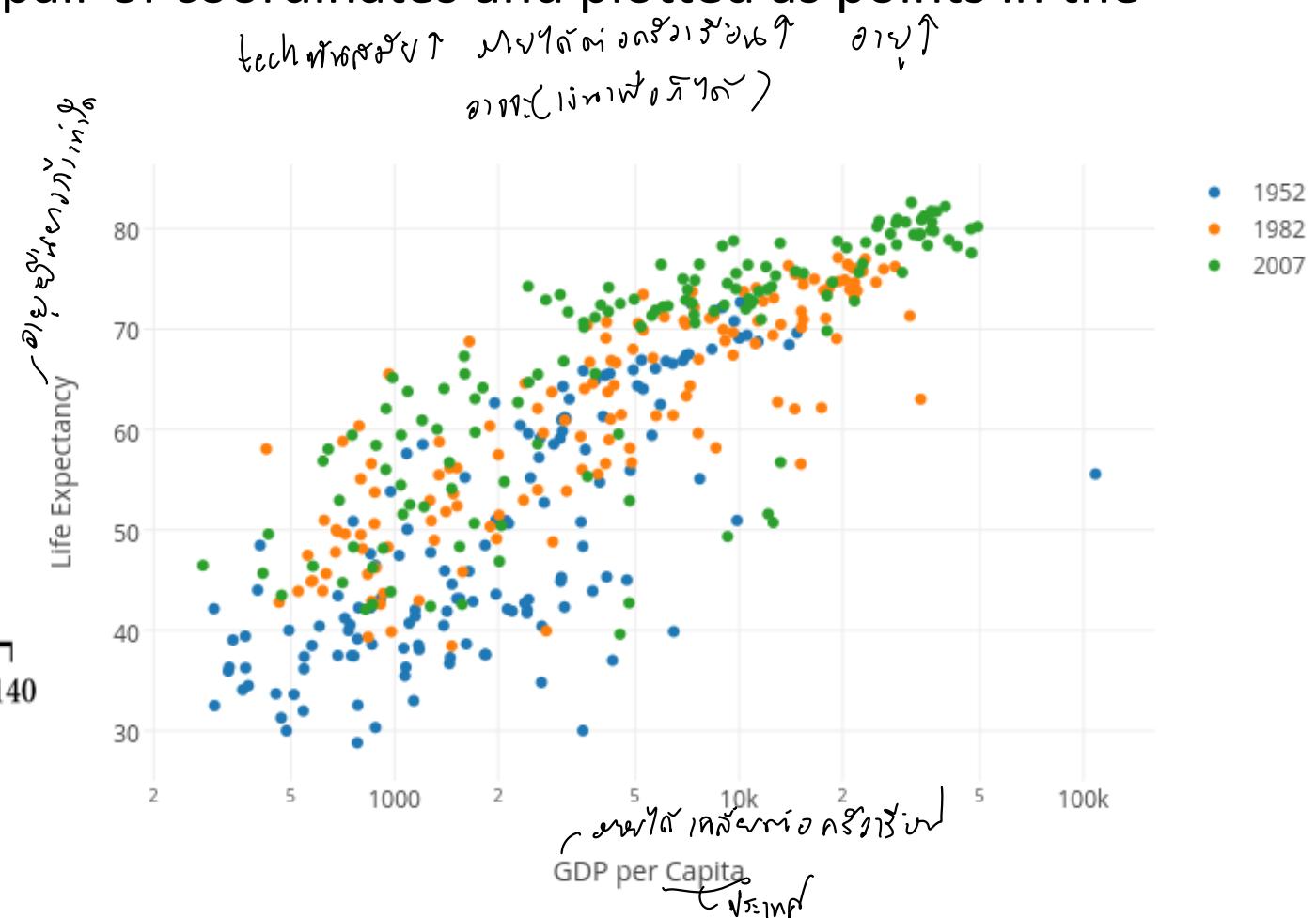
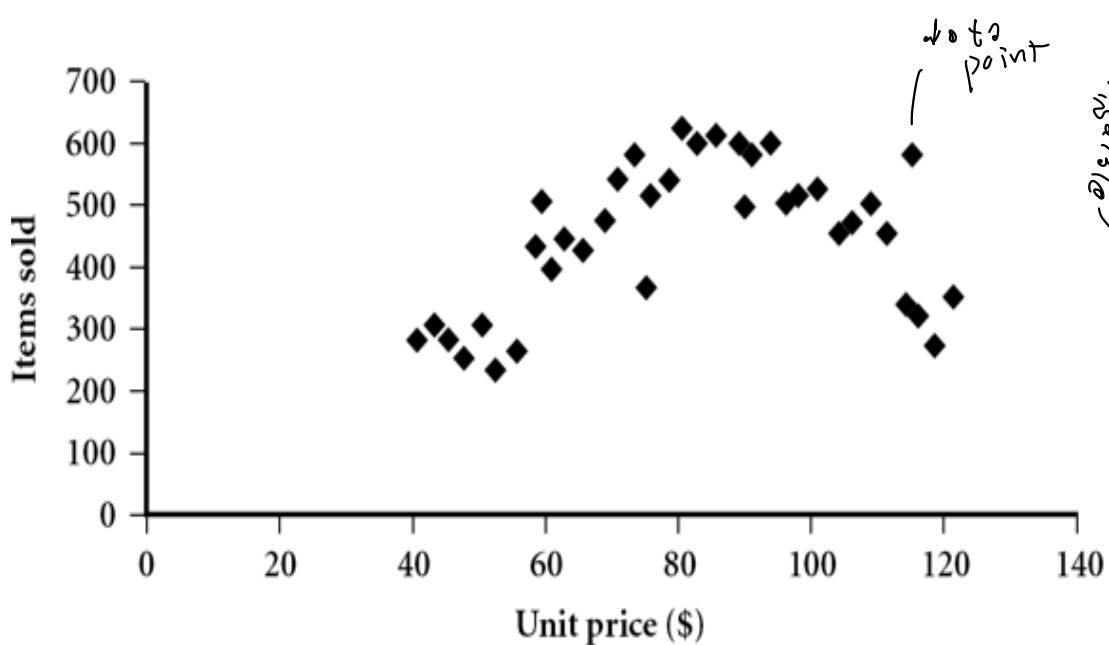
- ❑ The two histograms shown in the left may have the same boxplot representation *υηη.ης.ηιυυυυ, γηγη.ηη.ηι, boxplot*
- ❑ The same values for: min, Q1, median, Q3, max
- ❑ But they have rather different data distributions

if dimension was plot

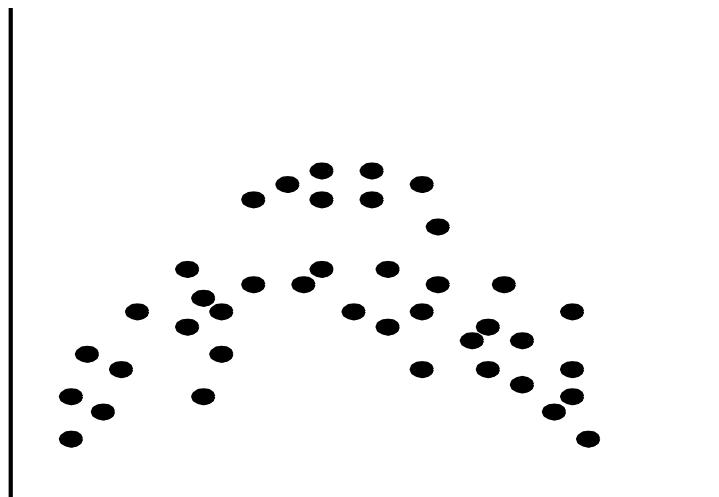
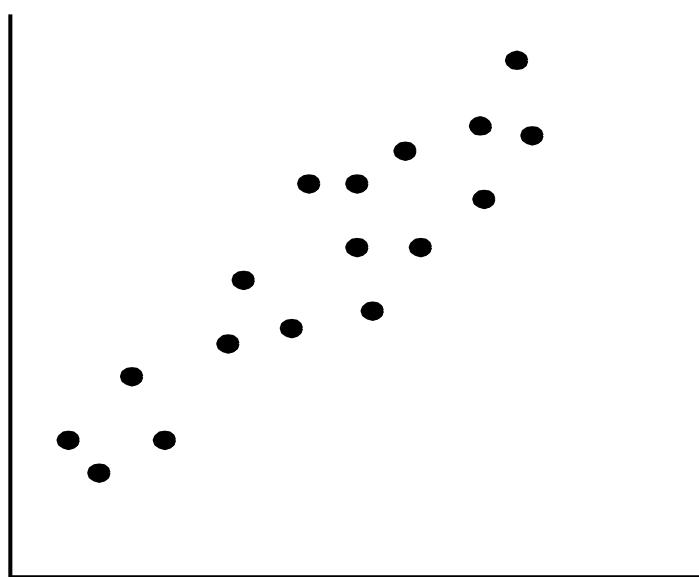
Scatter plot

if dim 1 & 2 — plotting is easy
in plot of $n \times n$.

- 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

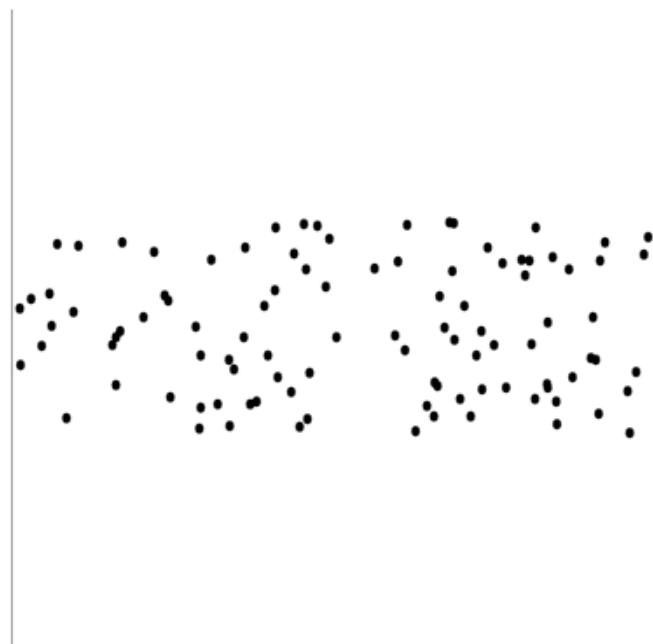


Positively and Negatively Correlated Data



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

Uncorrelated Data

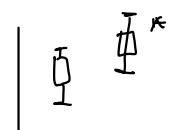


dimension 1, 2, ..., \rightarrow uncorrelated

Chapter 2. Getting to Know Your Data

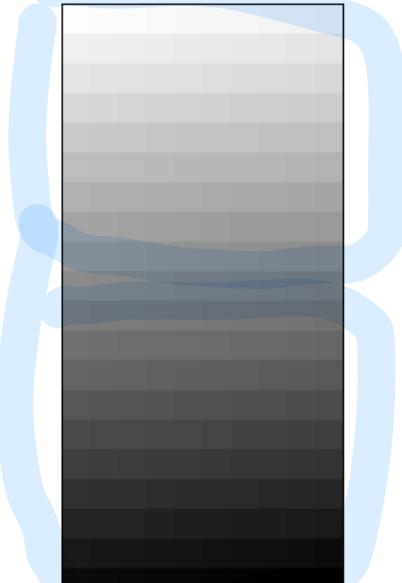
- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization 
- Measuring Data Similarity and Dissimilarity
- Summary

Data Visualization

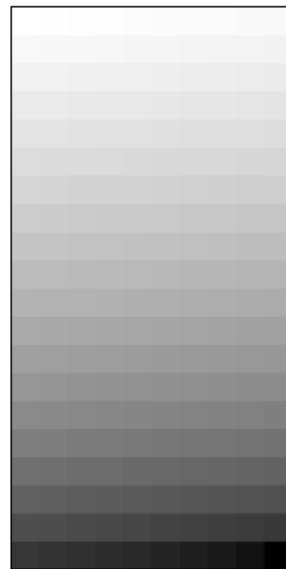
- Why data visualization? ↗ សំណង់ នូវ នៅពី នគរណ៍ → insight
- Gain insight into an information space by mapping data onto graphical primitives
- Provide qualitative overview of large data sets
- Search for patterns, trends, structure, irregularities, relationships among data
- Help find interesting regions and suitable parameters for further quantitative analysis
- Provide a visual proof of computer representations derived ^{និង t-test} $\xrightarrow{\text{proof}}$ box plot 
- Categorization of visualization methods:
 - Pixel-oriented visualization techniques
 - Geometric projection visualization techniques
 - Icon-based visualization techniques
 - Hierarchical ^{សំគាល់សម្រាប់} visualization techniques
 - Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

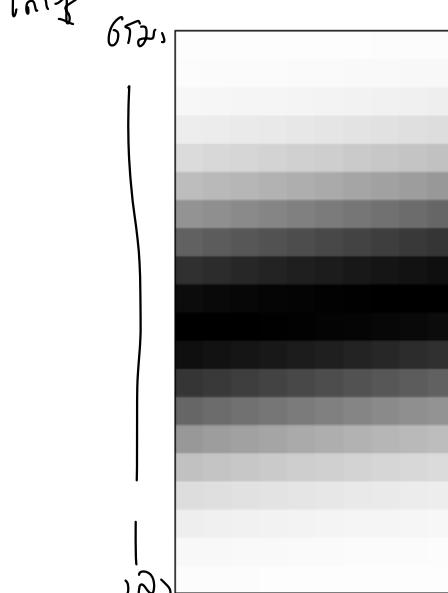
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



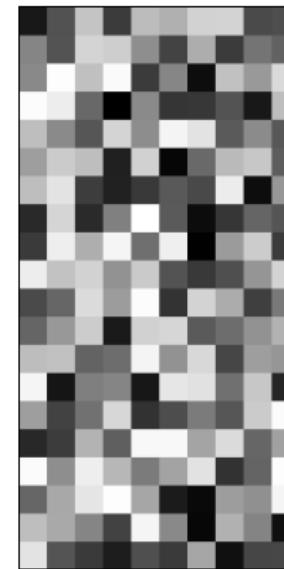
(a) Income



(b) Credit Limit



(c) transaction volume



(d) age

મન સ્થાન વિશેરી

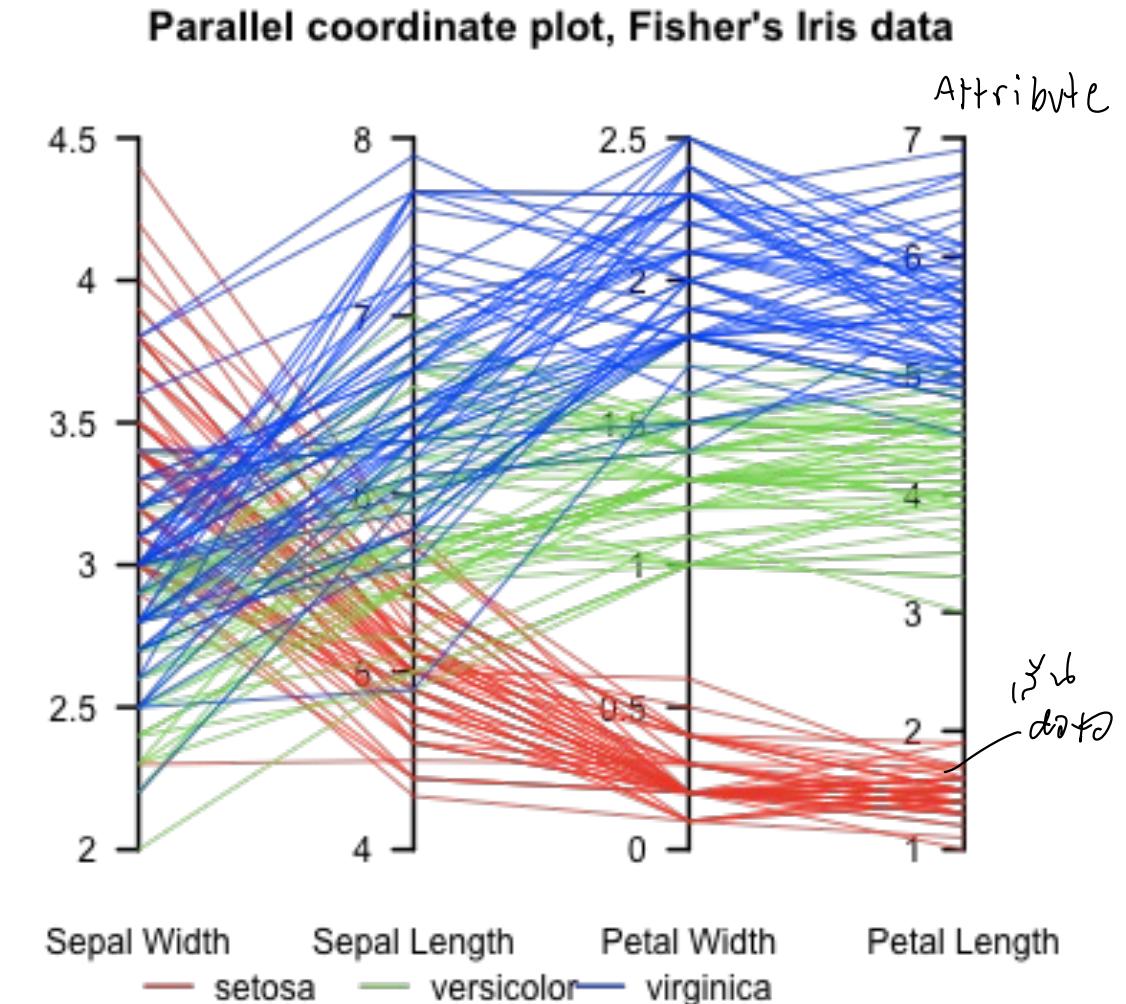
heat map

grid pattern

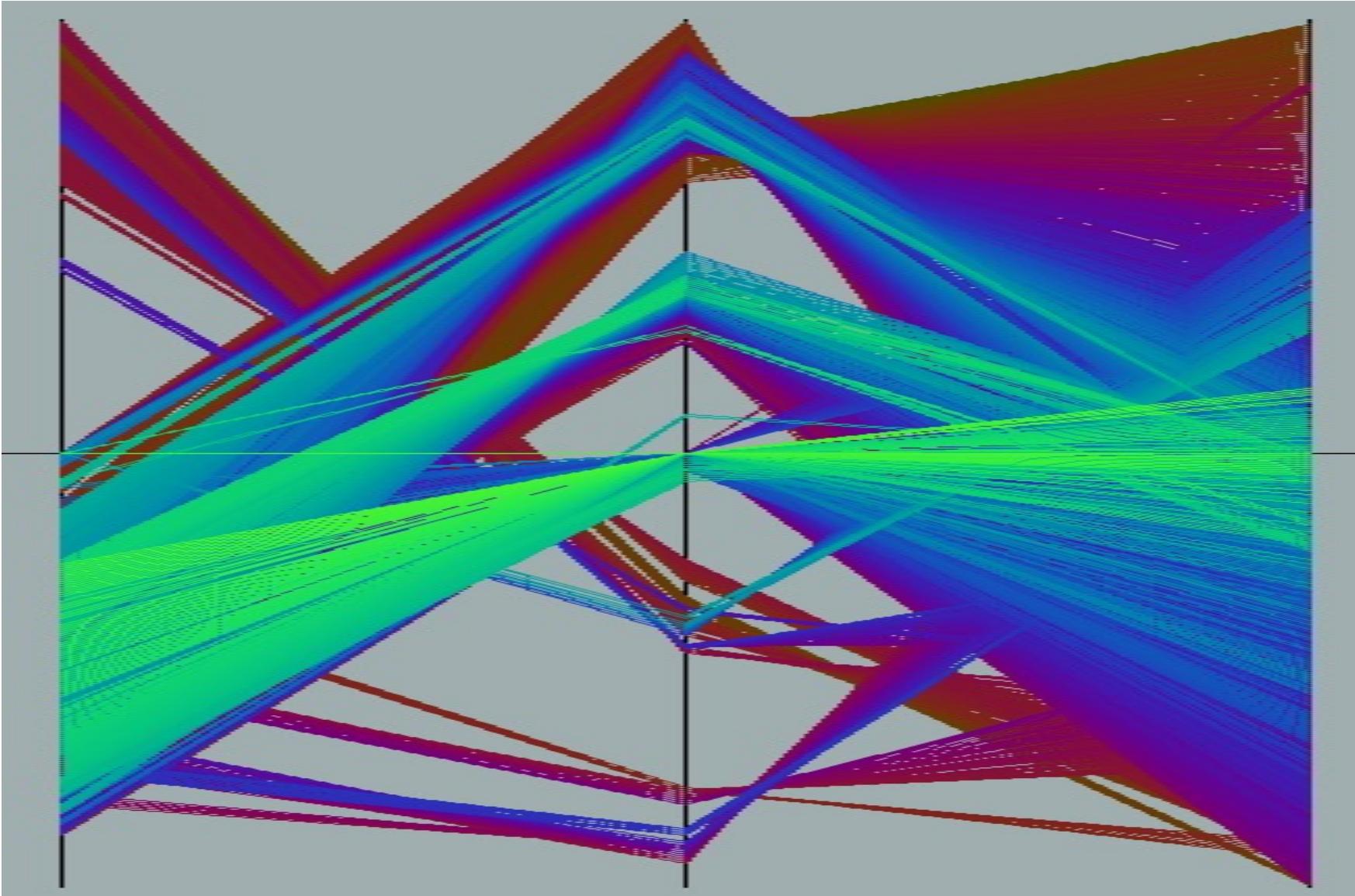
(મન સ્થાન વિશેરી)

Parallel Coordinates

- 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

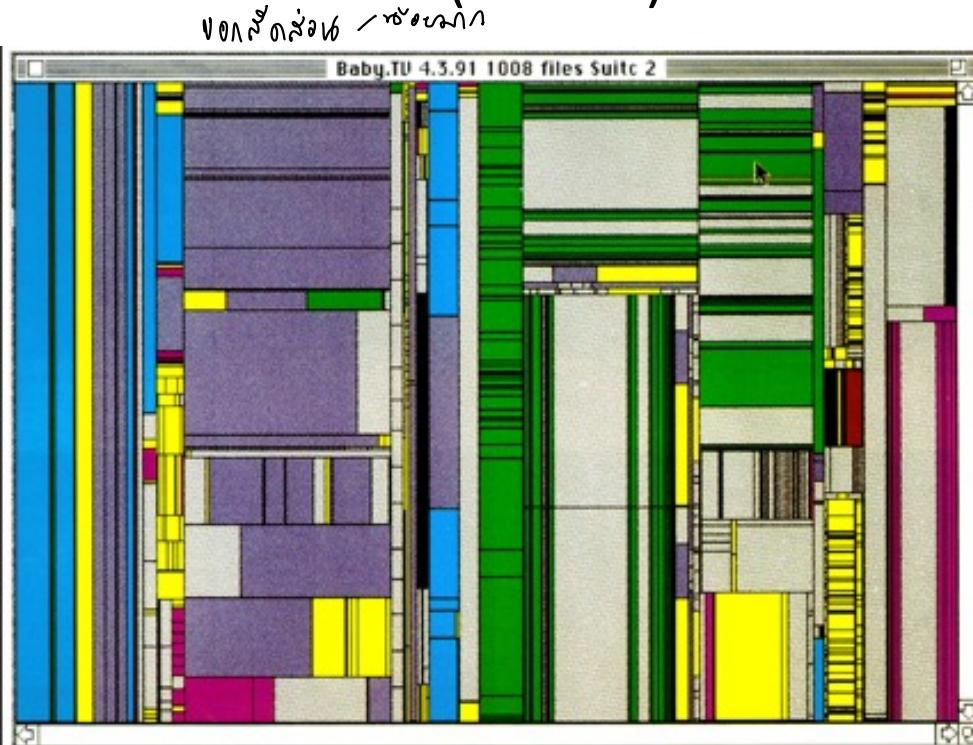


Parallel Coordinates of a Data Set

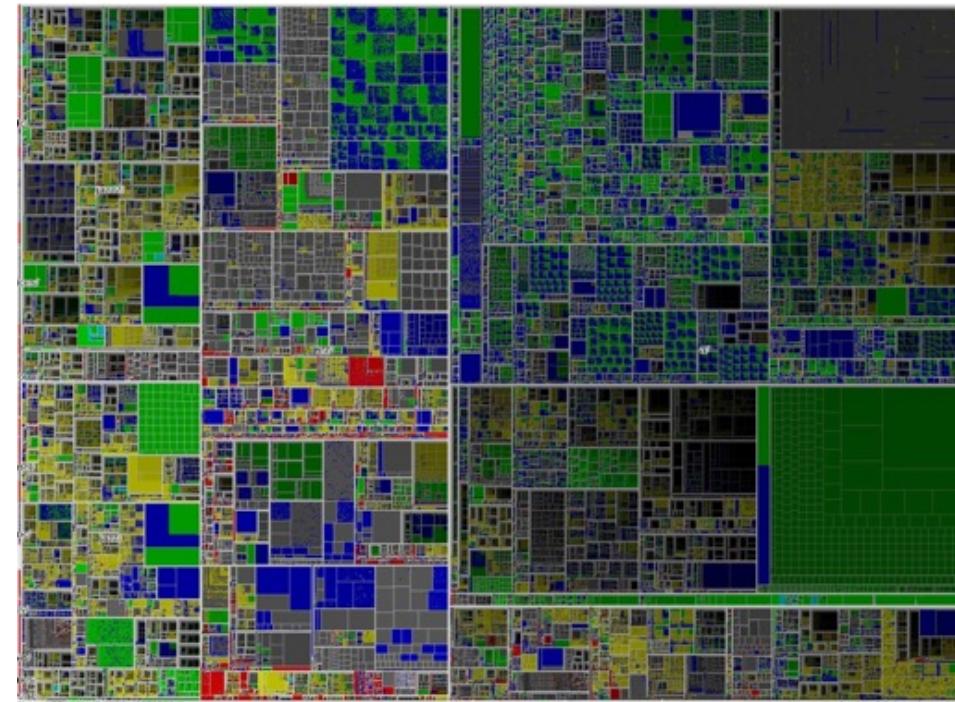


Tree-Map

- 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)



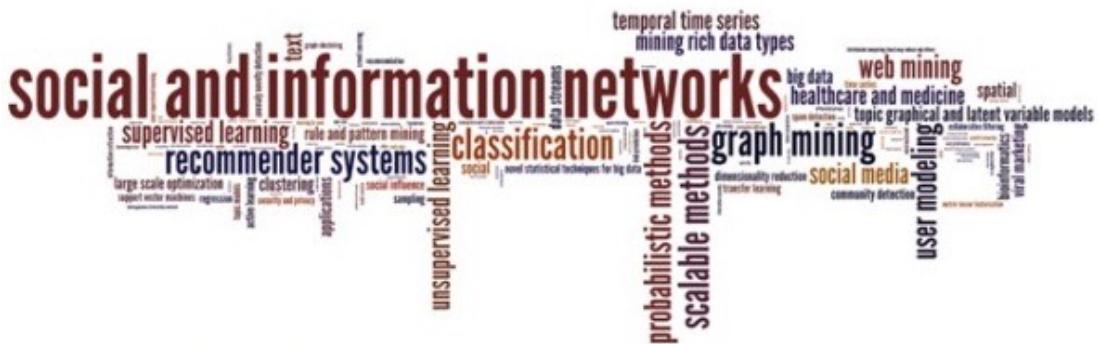
Schneiderman@UMD: Tree-Map of a File System



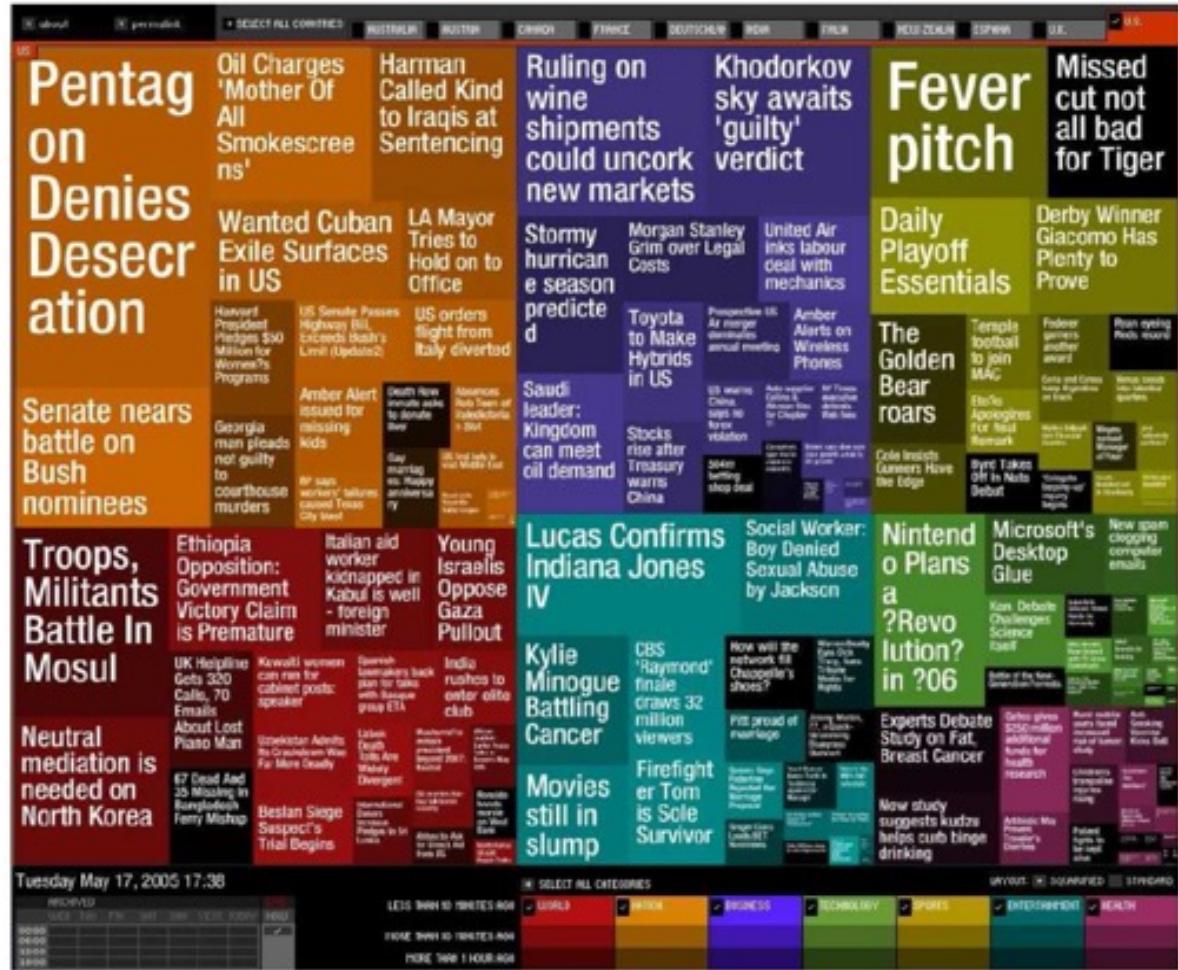
Schneiderman@UMD: Tree-Map to support large data sets of a million items

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



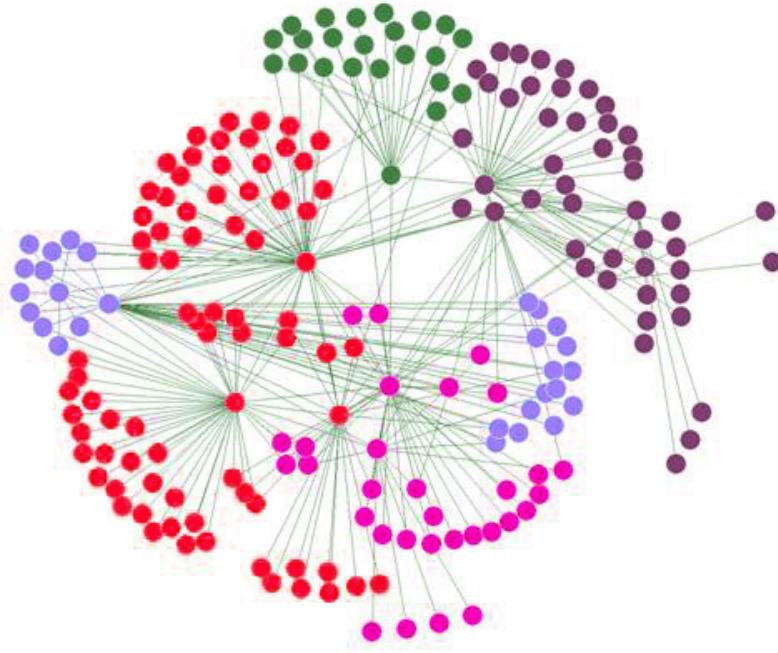
KDD 2013 Research Paper Title Tag Cloud



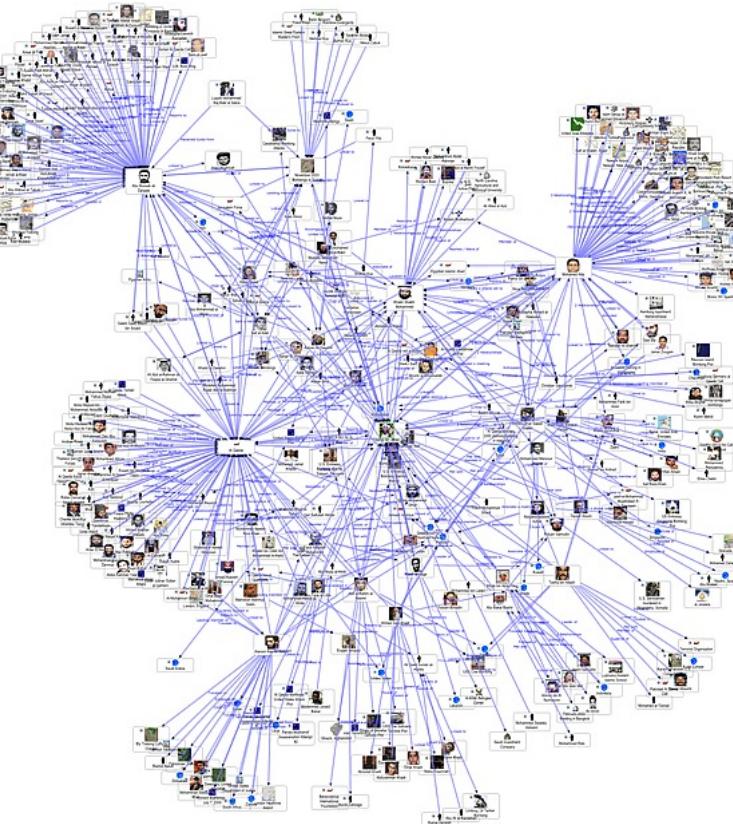
Newsmap: Google News Stories in 2005

Visualizing Complex Data and Relations: Social Networks

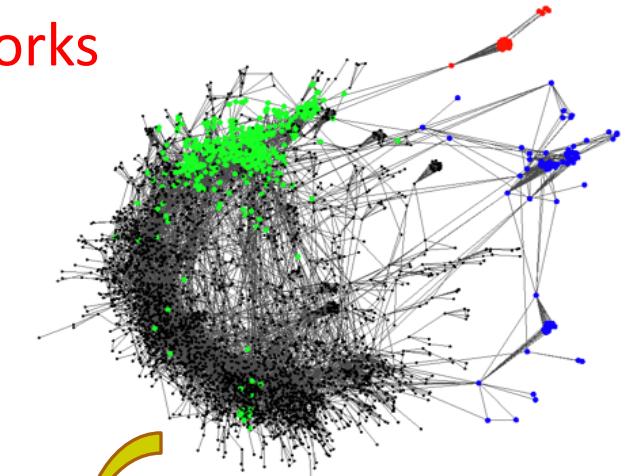
- Visualizing non-numerical data: social and information networks



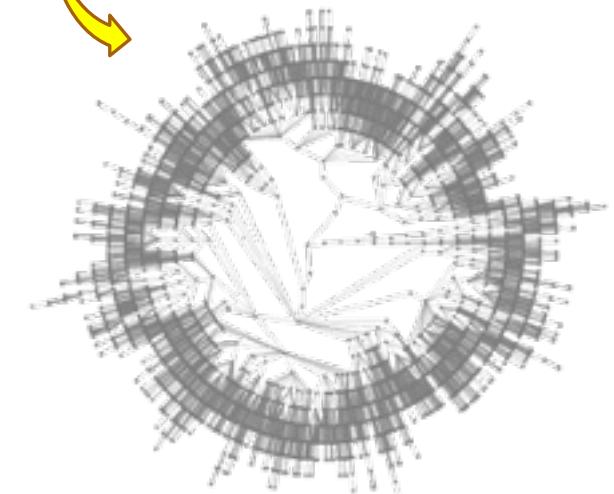
A typical network structure



A social network



organizing
information networks



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



ప్రాథ. ఇంజినీరు / మాన్యమాన్య దాతా

Similarity, Dissimilarity, and Proximity

รากศัพท์ภาษากรีก → Attribute (คุณลักษณะ)

- ❑ **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
 - ❑ Often falls in the range [0,1]: 0: no similarity; 1: completely similar
- ❑ **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

- 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

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

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

Standardizing Numeric Data

1. Standardization

Z-score:

$$z = \frac{x - \mu}{\sigma}$$

where $\mu = 0$ and $\sigma = 1$

- 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

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

Normalization

where

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

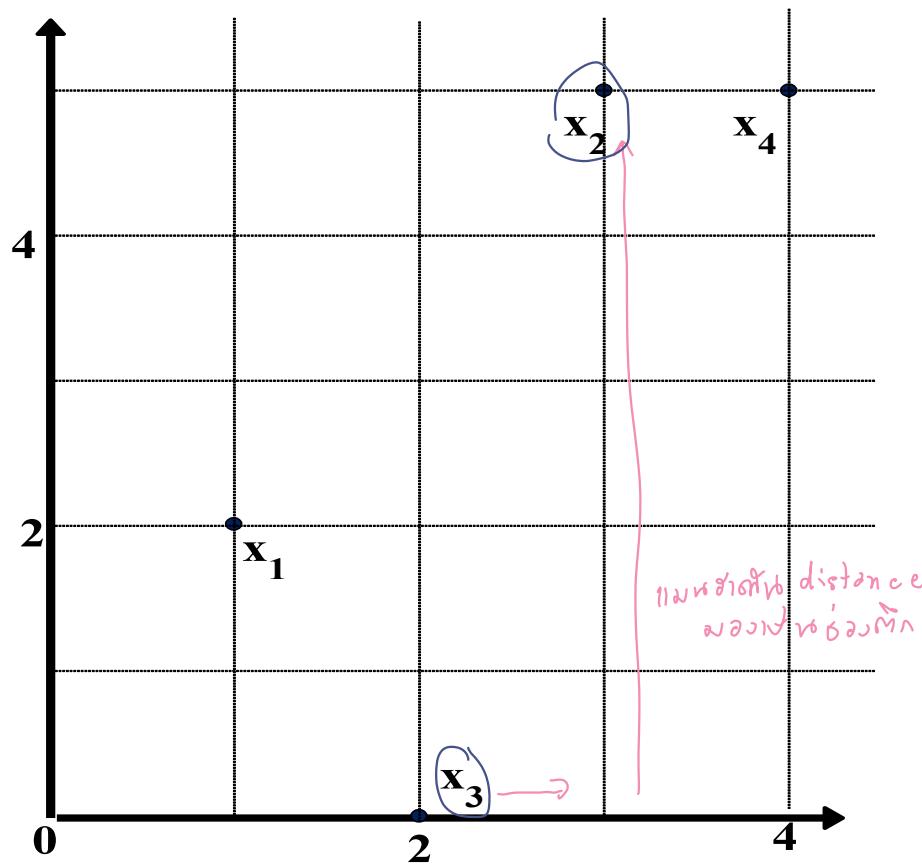
Min Max Scale

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

or x_{scaled}

- 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



จุดตัวอย่าง

Data Matrix

ค่าของแต่ละจุดในแต่ละ特徽

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

data (ตัวอย่าง)

ค่า

if คือ
ตัวอย่าง
ค่าที่ต้องการ
จะได้ค่าเดียวกัน
40,000 คือ
ค่าเดียว
จะได้ค่าเดียวกัน

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

ค่าที่คำนวณได้มาแล้ว

ex. city block distance

Distance on Numeric Data: Minkowski Distance

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

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

where $i = (x_{i1}, x_{i2}, \dots, x_{il})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jl})$ 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

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

- $p = 2$: (L_2 norm) **Euclidean distance** エルツー ディスタンス.

💡

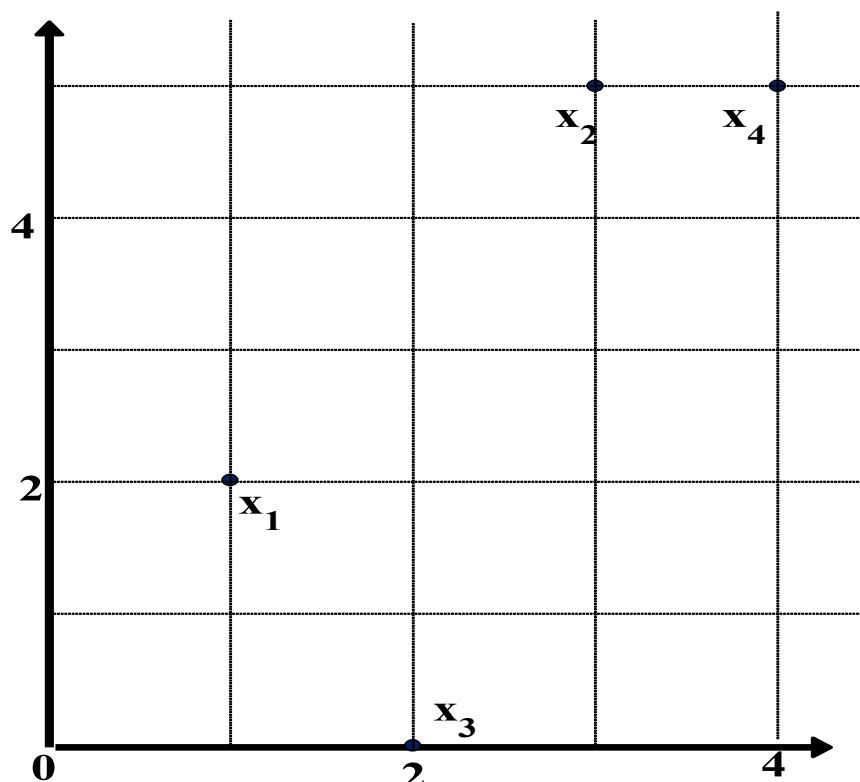
$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{il} - x_{jl}|^2}$$

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

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

Example: Minkowski Distance at Special Cases

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



Manhattan (L_1)

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

Euclidean (L_2)

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

Supremum (L_∞)

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

$$d(x, y) = \lim_{n \rightarrow \infty} \sqrt[n]{|x_1 - y_1|^p + |x_2 - y_2|^p + \dots + |x_n - y_n|^p} = \max_{j=1}^n |x_j - y_j|$$

எனவே L_∞ நாம் $\max_{j=1}^n |x_j - y_j|$ என்று கூறலாம்

b/w 2 & 3 = 5

* L_1 norm

Proximity Measure for Binary Attributes

- A contingency table for binary data

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

- Distance measure for symmetric binary variables

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

- Distance measure for asymmetric binary variables:

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

- Jaccard coefficient (*similarity* measure for

asymmetric binary variables):

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

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

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

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

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

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

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

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

Proximity Measure for Categorical Attributes

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

Ordinal Variables

- 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: $r_{if} \in \{1, \dots, M_f\}$
 - 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
 - Then distance: $d(\text{freshman}, \text{senior}) = 1$, $d(\text{junior}, \text{senior}) = 1/3$
- 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:

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

- 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

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, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

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

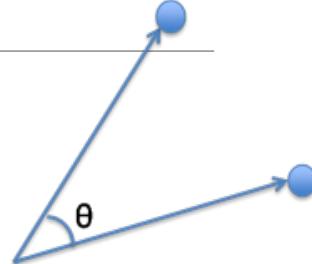
where \bullet indicates vector dot product, $\|d\|$: the length of vector d

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

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

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) \quad d_2 = (3, 0, 2, 0, 1, 1, 1, 0, 1, 0)$$

- First, calculate vector dot product

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

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

$$\|d_1\| = \sqrt{5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.481$$

$$\|d_2\| = \sqrt{3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1} = 4.12$$

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

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



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
 - 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

