华东师范大学计算机科学与技术学院
School of Computer Science and Technology

数据分析实践

第2课. 了解数据

兰 曼

计算机科学与技术学院

华东师范大学

内容

1 Data Objects and Attribute Types

2 Basic Statistical Descriptions of Data

3 Data Visualization

4 Measuring Data Similarity and Dissimilarity

1 Data Objects and Attribute Types

Types Of Data Sets

- **Record**
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- **Graph and network**
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- **Ordered**
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- **Spatial, image and multimedia:**
  - Spatial data: maps
  - Image data:
  - Video data:

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

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

Data Objects

- Data sets are made up of data objects.
- A data object represents an entity.
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
  - Also called *samples , examples, instances, data points, objects, tuples, records, rows.*
- Data objects are described by attributes.
  - Representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
  - Also called *fields, columns, dimensions, features, variables.*
- Feature vector (or attribute vector) is a group of attributes

Attribute Types

- **Nominal：** categories, states, enumeration, 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**

Mining Data Descriptive Characteristics

- **Motivation**
  - To better understand the data： central tendency， variation and spread
- **Measures of central tendency**
  - mean, median, mode, midrange, etc.
- **Data dispersion characteristics**
  - median, max, min, quantiles, outliers, variance, etc.

Measuring the Central Tendency

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

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

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

- **Weighted arithmetic mean:**

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

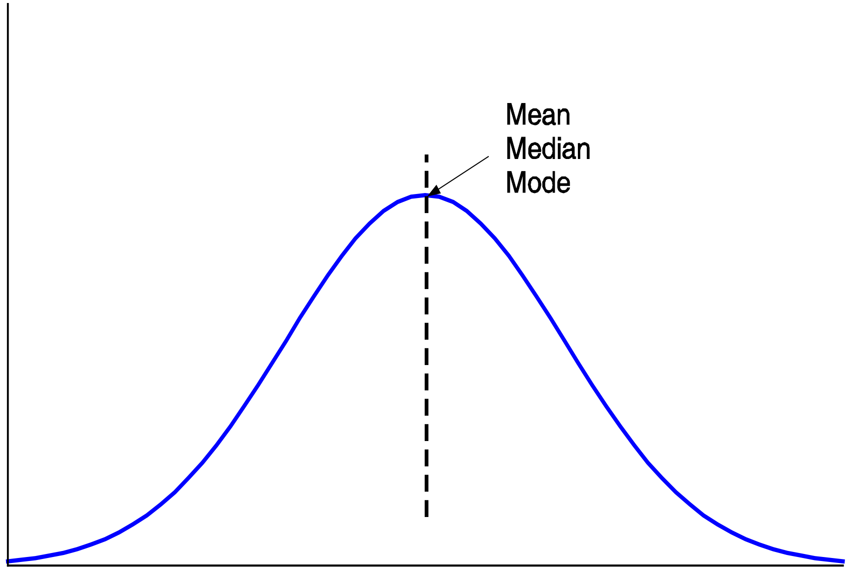
- **Trimmed mean: chopping extreme values**

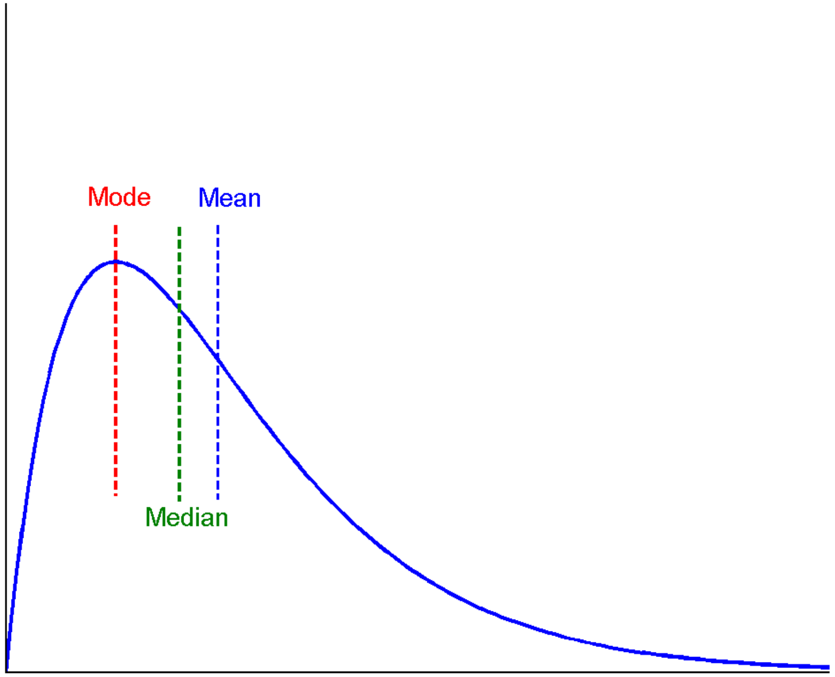
Measuring the Central Tendency

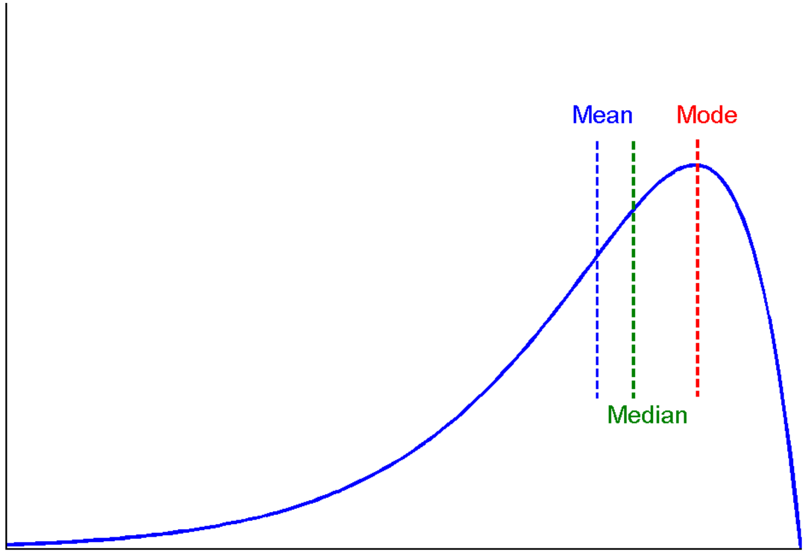
- **Median: A holistic measure**
  - **Middle value if odd number of values, or average of the middle two values otherwise**
- **Mode**
  - **Value that occurs most frequently in the data**
  - **Unimodal, bimodal, trimodal, multimodal**
- **Midrange**
  - **Average of min and max (*min()* and *max()* in SQL)**

Symmetric vs. Skewed Data

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







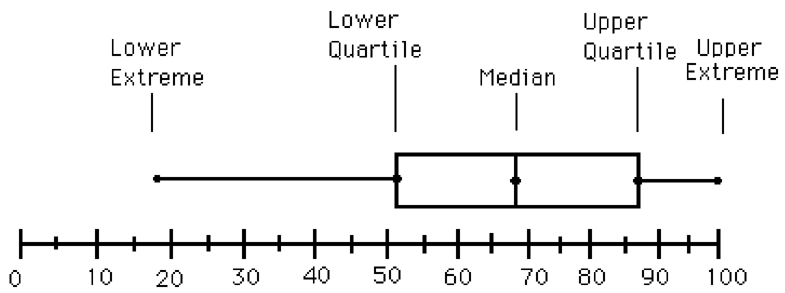
Measuring the Dispersion of Data

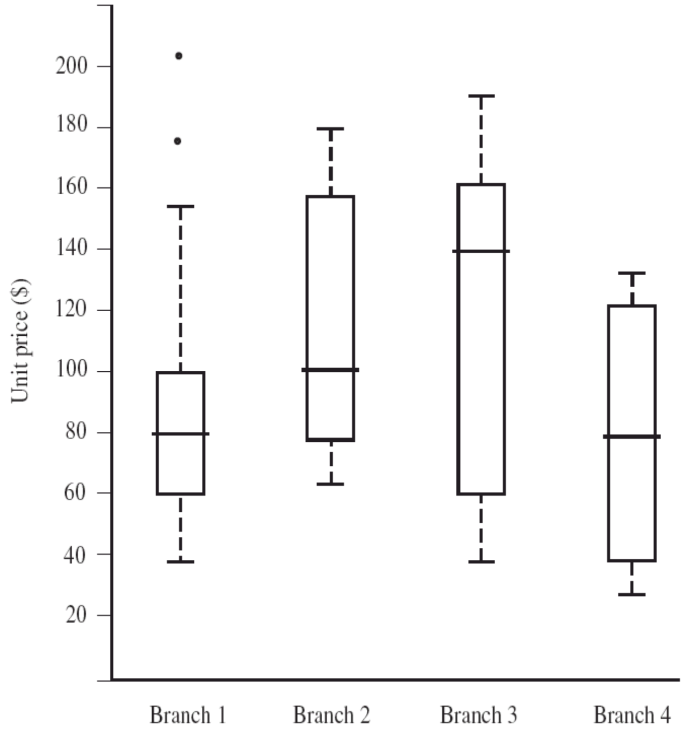
Quartiles, outliers and boxplots

- Range : = max –min
- Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
- Inter-quartile range: IQR = $Q_3 – Q_1$
- Five number summary: min, Q_1, M, Q_3, max
- Outlier: usually, a value higher/lower than 1.5 x IQR
- Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually

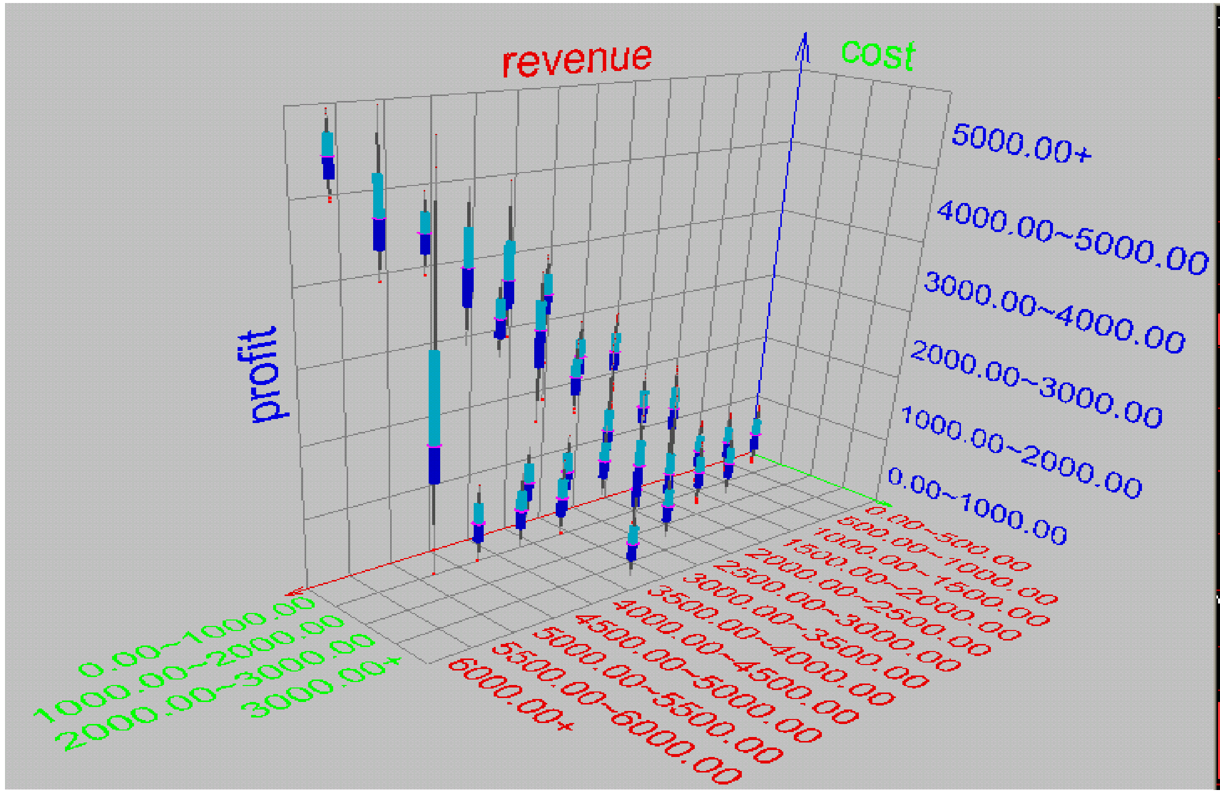
Boxplot Analysis

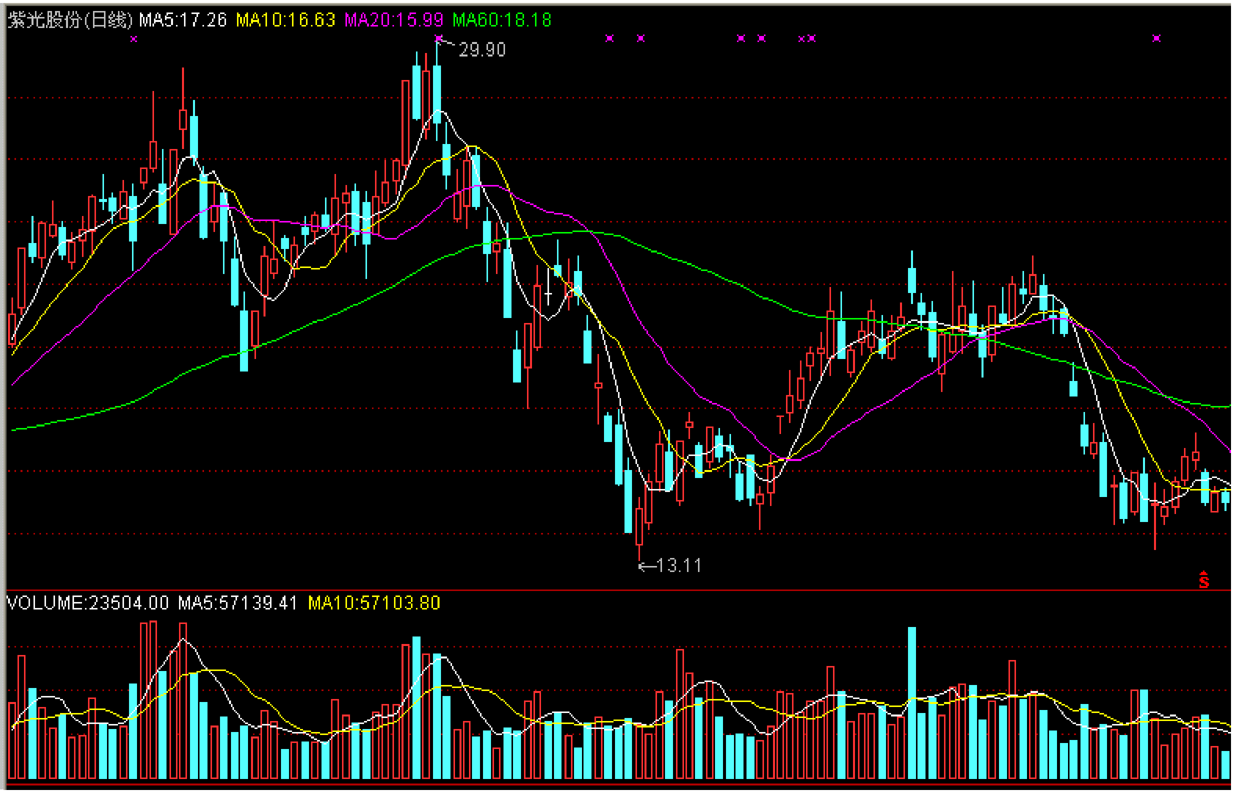
- **Five-number summary of a distribution:**
  - Minimum, Q_1, M, Q_3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually





Visualization of Data Dispersion: 3-D Boxplots





Measuring the Dispersion of Data

➢Variance and standard deviation (*sample: s, population: σ*)

- **Variance:** **(algebraic, scalable computation)**

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

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

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

Properties of Normal Distribution Curve

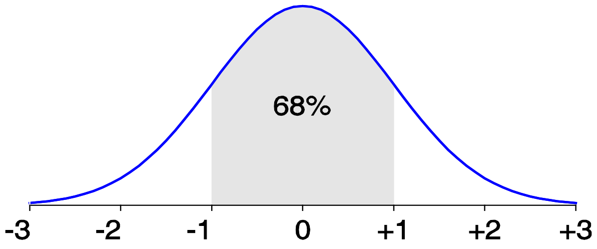
➢The normal (distribution) curve

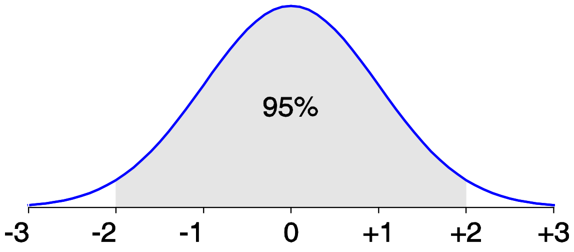
- From $\mu–\sigma$ to $\mu+\sigma$: contains about 68% of the measurements

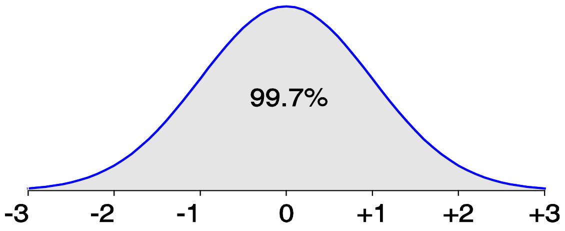
  (μ: mean, σ: standard deviation)

- From $\mu–2\sigma$ to $\mu+2\sigma$: contains about 95% of it

- From $\mu–3\sigma$ to $\mu+3\sigma$: contains about 99.7% of it







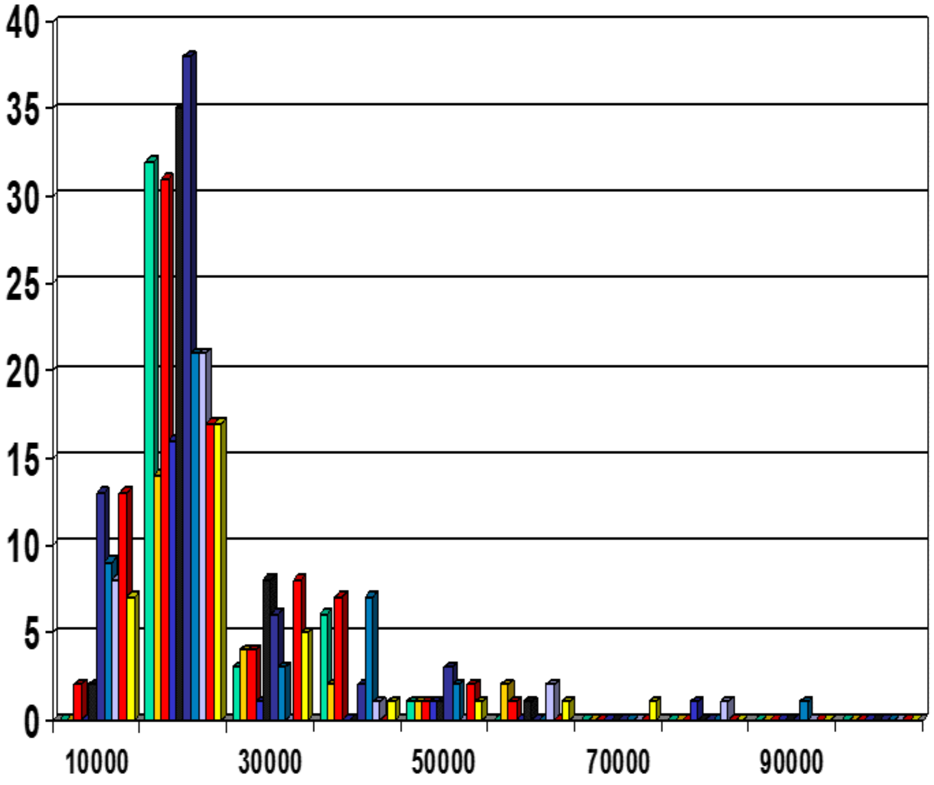
Graphic Displays of Basic Statistical Descriptions

- **Boxplot:** graphic display of five-number summary (shown before)
- **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

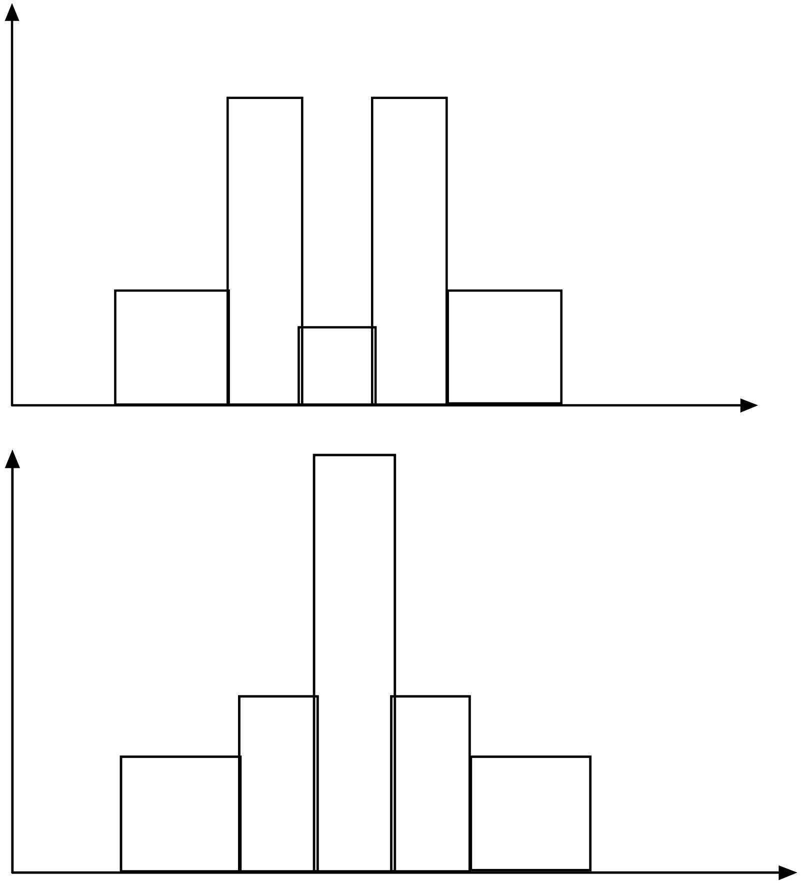
Histogram Analysis

➢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
- A univariate graphical method



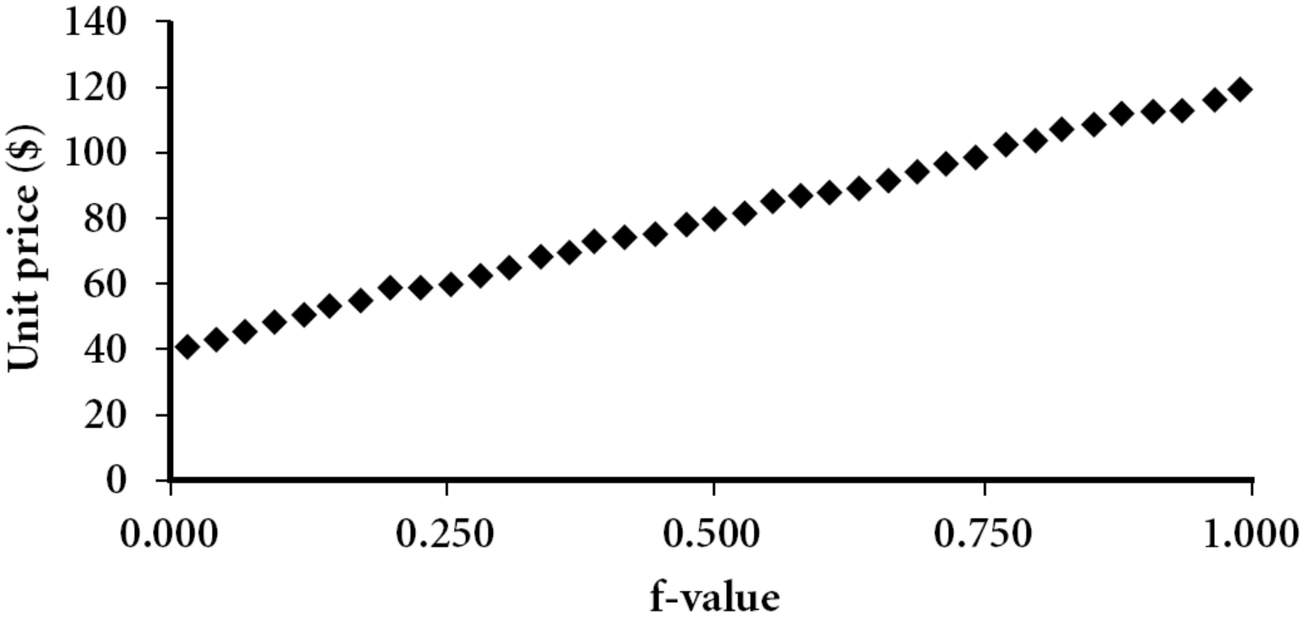
Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q_1, median, Q_3, max
- But they have rather different data distributions

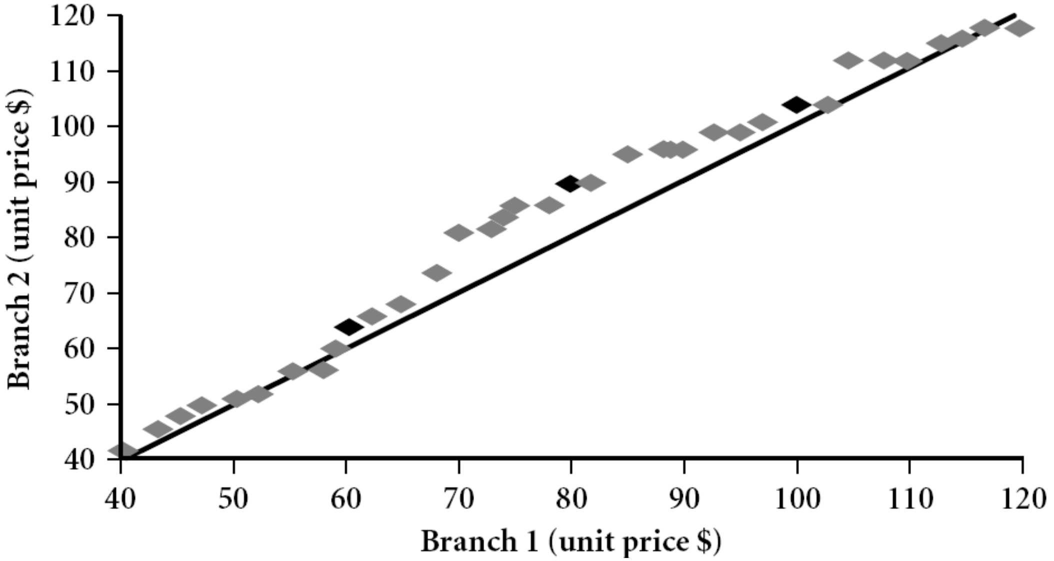
Quantile Plot

- 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 f_i % of the data are below or equal to the value x_i



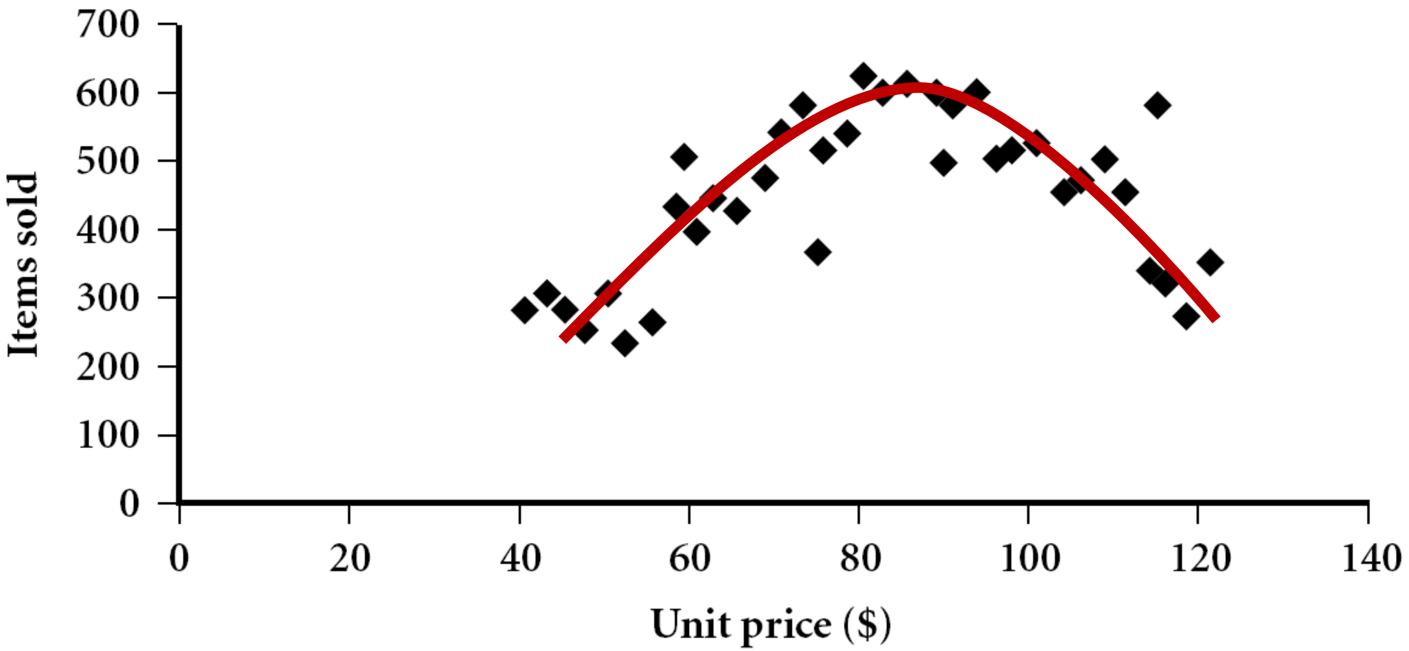
Quantile-Quantile (Q-Q) Plot

- **Graphs the quantiles of one univariate distribution against the corresponding quantiles of another**
- **Allows the user to view: whether 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.**

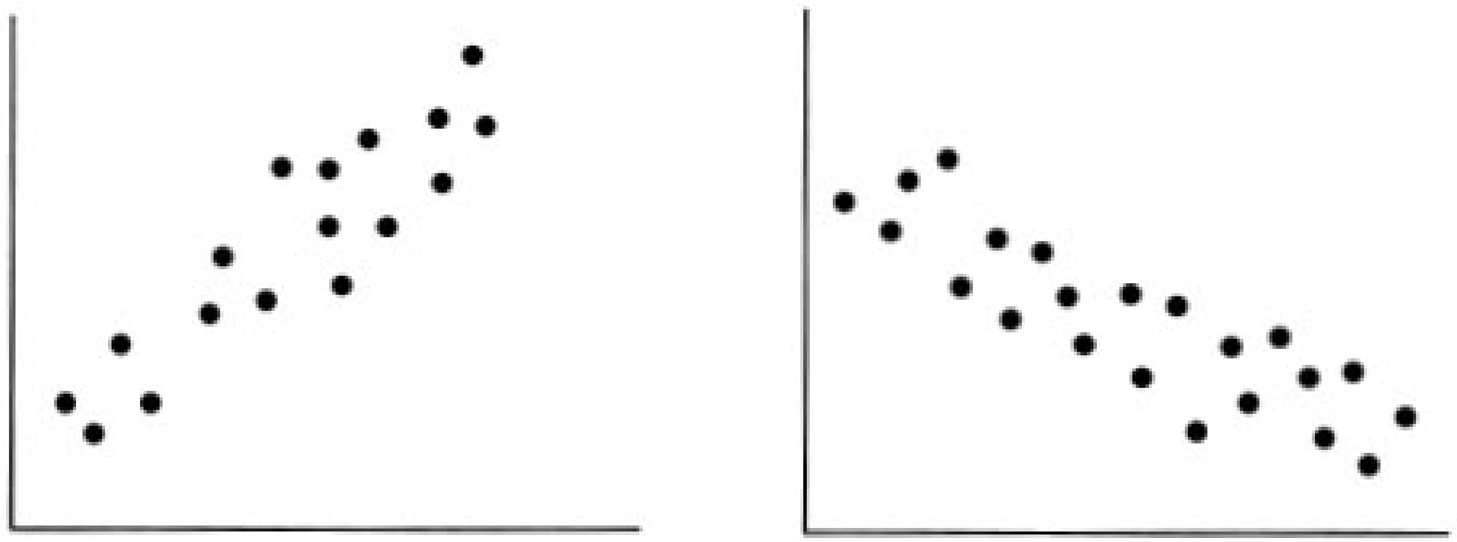


Scatter plot

- 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

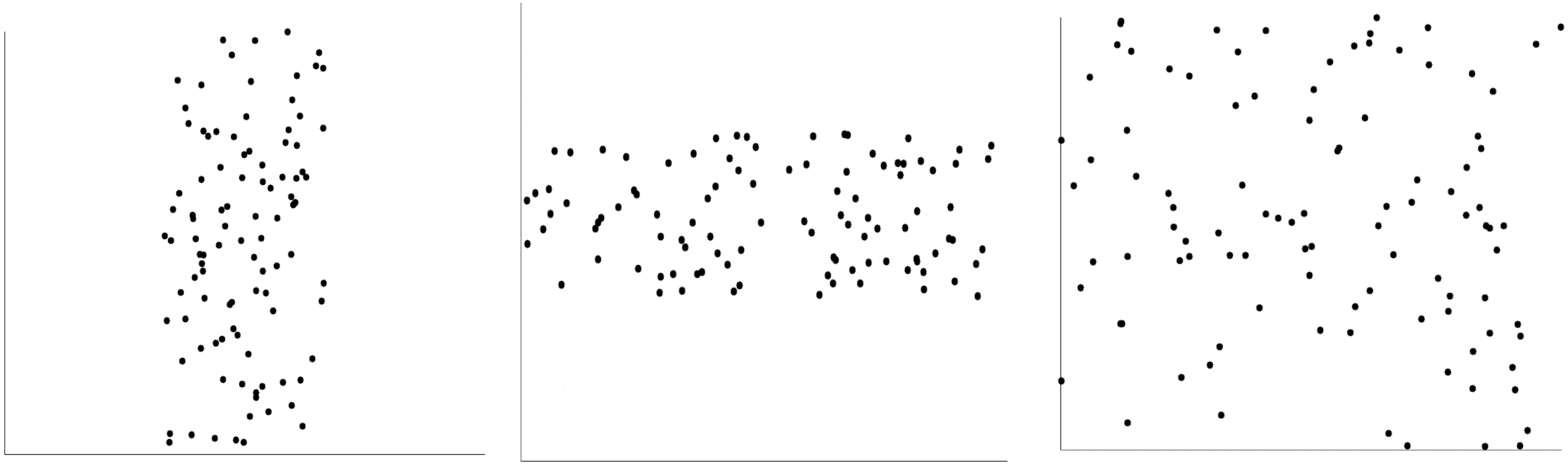


Positively and Negatively Correlated Data



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

Not Correlated Data



3 Data Visualization

Why data visualization?

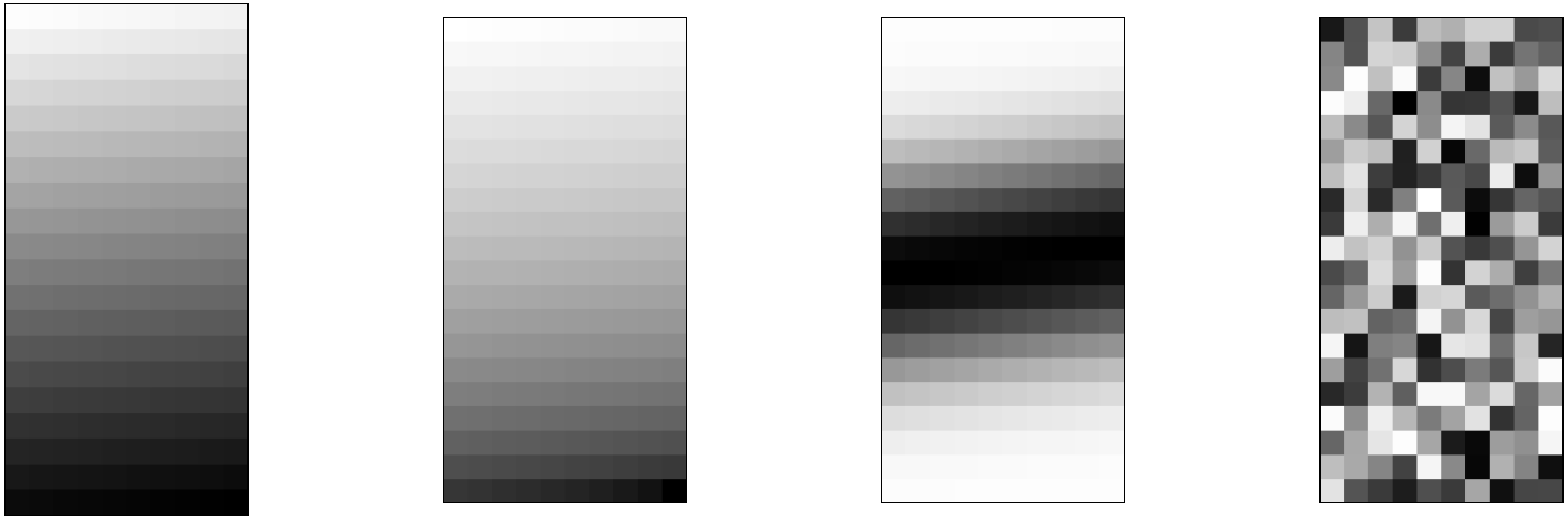
- Gain insight into an information space by mapping data onto graphical primitives
- Provide qualitative overview of large data sets
- Search for patterns, trends, structure, irregularities, relationships among data
- Help find interesting regions and suitable parameters for further quantitative analysis
- 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

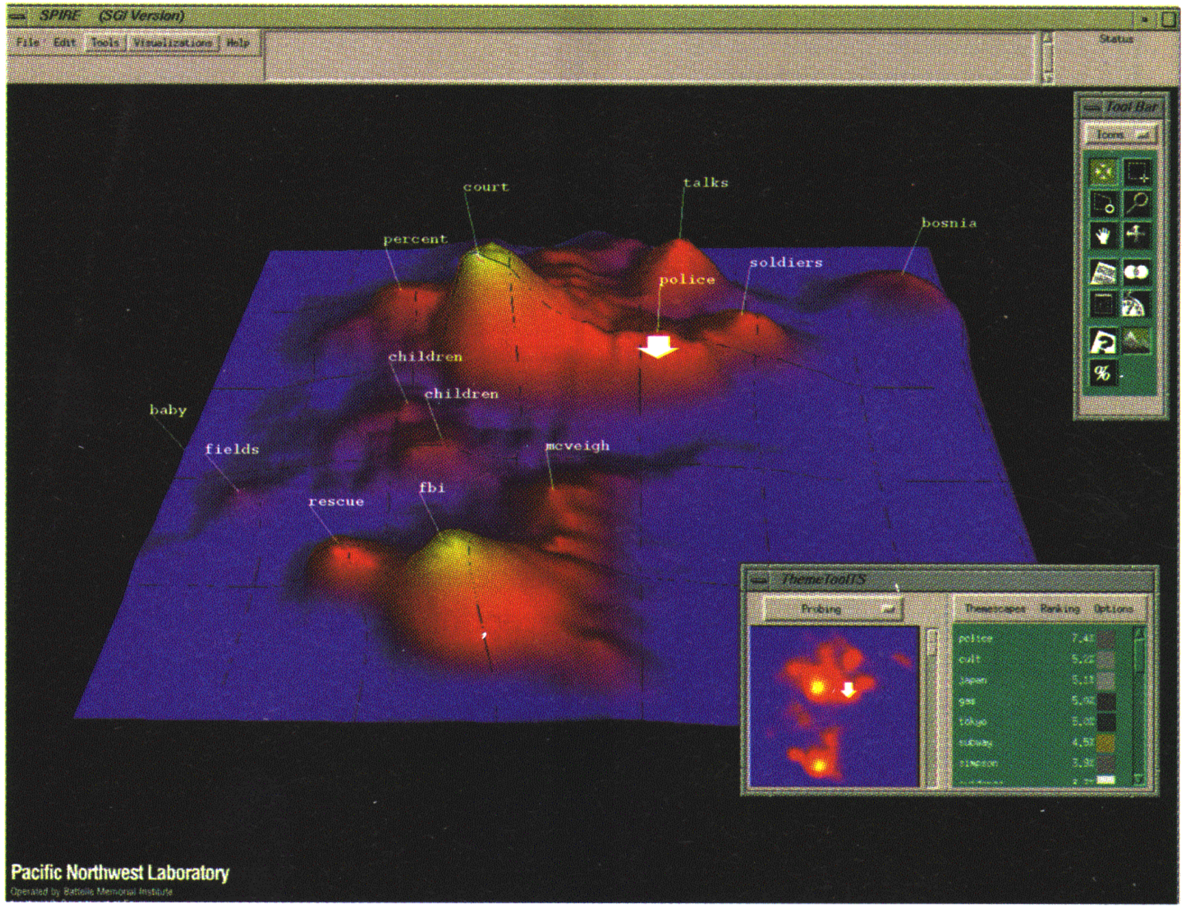
- For a data set of *m* dimensions, create *m* windows on the screen, one for each dimension
- The *m* dimension values of a record are mapped to *m* pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



(a) Income (b) Credit Limit (c) Transaction Volume (d) Age

Landscapes

Used by permission of B. Wright, Visible Decisions Inc.

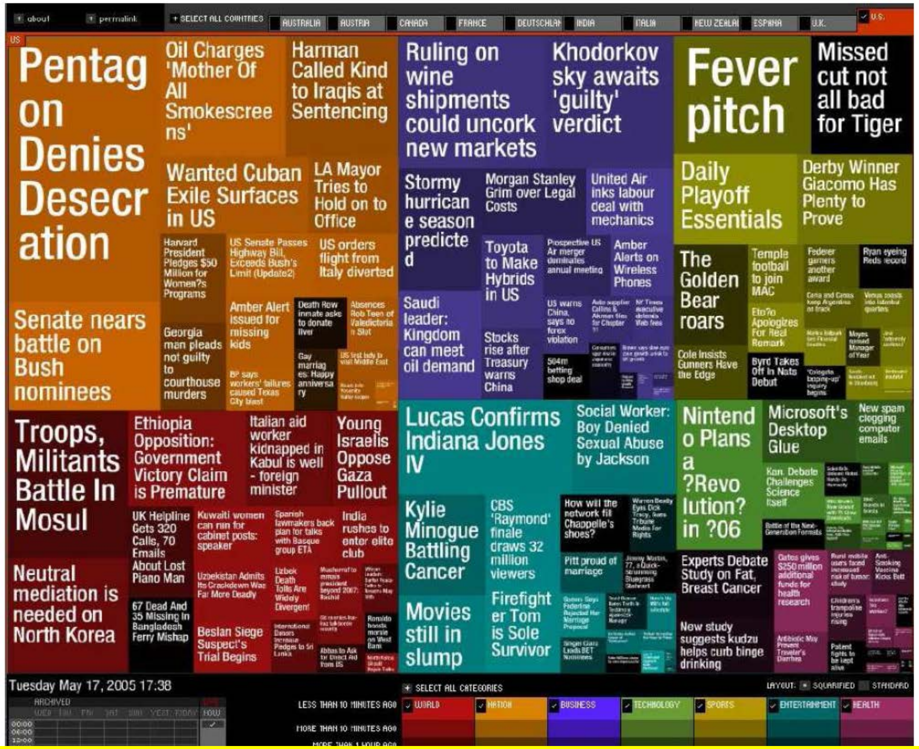


news articles
visualized as
a landscape

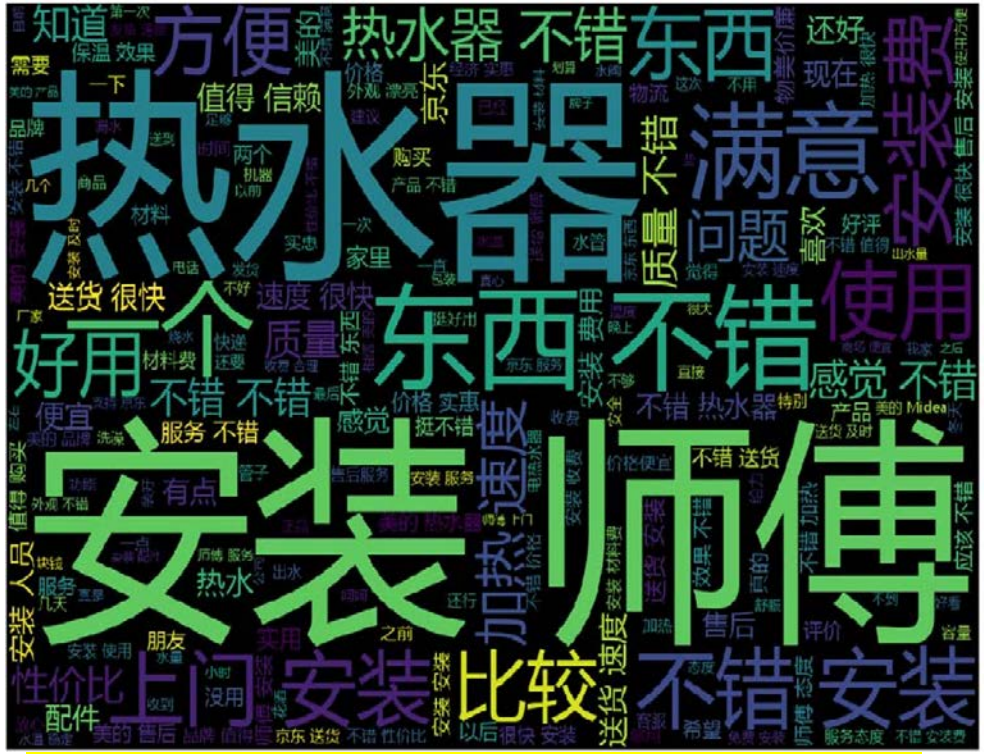
- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

Visualizing Complex Data

- ➢ Visualizing non-numerical data: text and social networks
- ➢ Tag cloud: visualizing user-generated tags
  - o The importance of tag is represented by font size/color


about
permalink
SELECT ALL COUNTRIES
AUSTRALIA
AUSTRIA
CANADA
FRANCE
DEUTSCHLAND
INDIA
ITALIA
NEW ZEALAND
ESPAÑA
U.K.
U.S.
Pentagon Denies Desecration
Oil Charges 'Mother Of All Smokescreens'
Harman Called Kind to Iraqis at Sentencing
Ruling on wine shipments could uncork new markets
Khodorkov sky awaits 'guilty' verdict
Fever pitch
Missed cut not all bad for Tiger
Wanted Cuban Exile Surfaces in US
LA Mayor Tries to Hold on to Office
Stormy hurrican e season predicted
Morgan Stanley Grim over Legal Costs
United Air inks labour deal with mechanics
Daily Playoff Essentials
Derby Winner Giacomo Has Plenty to Prove
Senate nears battle on Bush nominees
Toyota to Make Hybrids in US
The Golden Bear roars
Saudi leader: Kingdom can meet oil demand
Troops, Militants Battle In Mosul
Ethiopia Opposition: Government Victory Claim is Premature
Italian aid worker kidnapped in Kabul is well - foreign minister
Young Israelis Oppose Gaza Pullout
Lucas Confirms Indiana Jones IV
Social Worker: Boy Denied Sexual Abuse by Jackson
Nintendo Plans a ?Revolution? in ?06
Microsoft's Desktop Glue
Kylie Minogue Battling Cancer
CBS 'Raymond' finale draws 32 million viewers
Neutral mediation is needed on North Korea
Movies still in slump
Firefighter Tom is Sole Survivor
Experts Debate Study on Fat, Breast Cancer
New study suggests kudzu helps curb binge drinking
Tuesday May 17, 2005 17:38
SELECT ALL CATEGORIES
WORLD
NATION
BUSINESS
TECHNOLOGY
SPORTS
ENTERTAINMENT
HEALTH
LAYOUT: SQUARIFIED STANDARD
LESS THAN 10 MINUTES AGO
MORE THAN 10 MINUTES AGO
MORE THAN 1 HOUR AGO

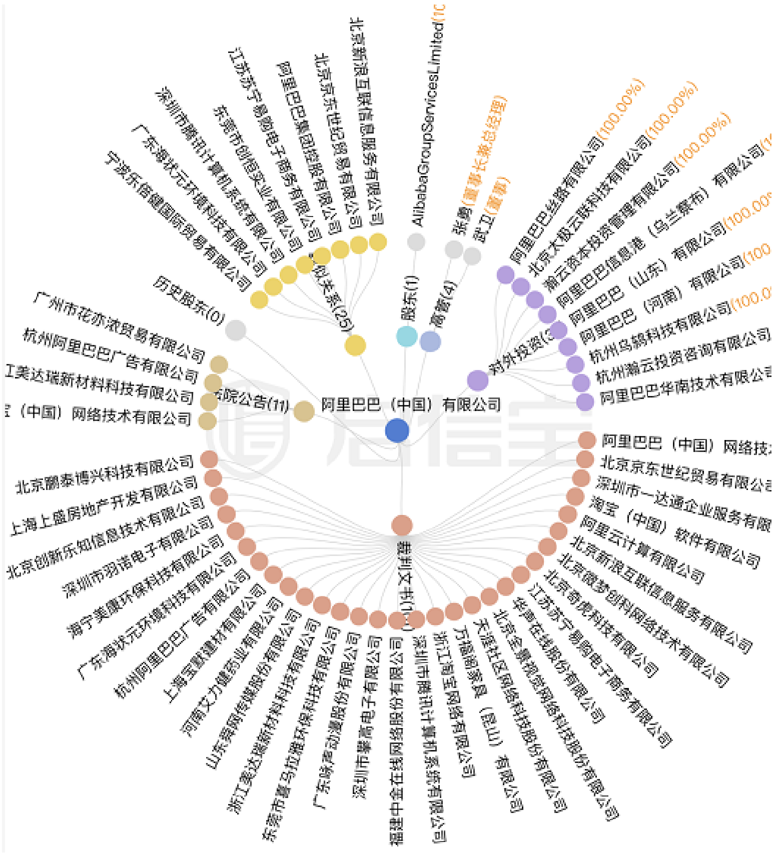
Newsmap: Google News Stories in 2005

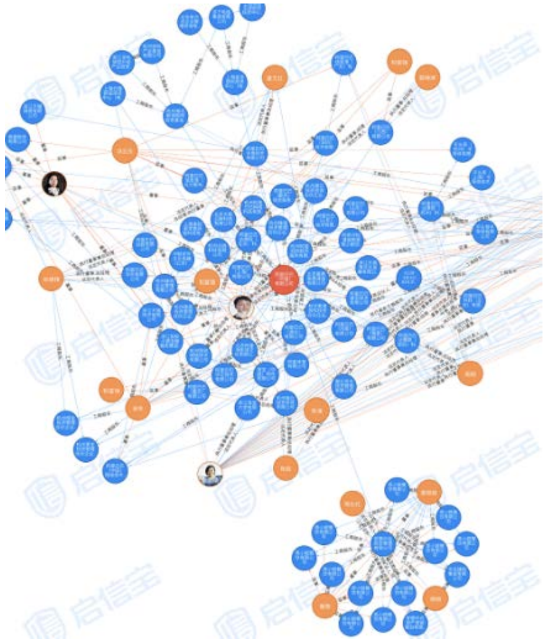


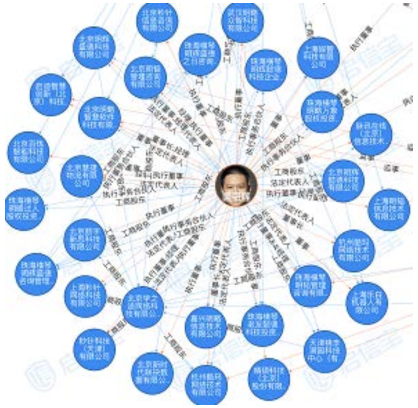
Reviews: JD Product Reviews in 2017

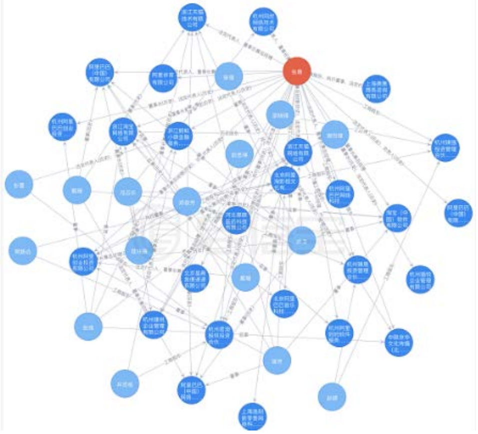
Visualizing Relations

- Visualizing relationships: entity relations & social networks
- Network Graph or Network Diagram: visualizing Relations, Knowledge Graph
  - The importance of tag is represented by font size/color of nodes and edges


阿里巴巴（中国）有限公司
历史股东(0)
广州市花亦浓贸易有限公司
杭州阿里巴巴广告有限公司
江美达瑞新材料科技有限公司
宝（中国）网络技术有限公司
法院公告(11)
宁波乐倍健国际贸易有限公司
广东海状元环境科技有限公司
深圳市腾讯计算机系统有限公司
东莞市创信实业有限公司
江苏苏宁易购电子商务有限公司
阿里巴巴集团控股有限公司
北京京东世纪贸易有限公司
北京新浪互联信息服务有限公司
AlibabaGroupServicesLimited(1(
股东(1)
张勇(董事长兼总经理)
武卫(董事)
高管(4)
对外投资(3
阿里巴巴丝路有限公司(100.00%)
北京大极云联科技有限公司(100.00%)
瀚云资本投资管理有限公司(100.00%)
阿里巴巴信息港（乌兰察布）有限公司(1
阿里巴巴（山东）有限公司(100.00
阿里巴巴（河南）有限公司(100
杭州乌鸫科技有限公司(100.0
杭州瀚云投资咨询有限公司
阿里巴巴华南技术有限公司
裁判文书(1
北京鹏泰博兴科技有限公司
上海上盛房地产开发有限公司
北京创新乐知信息技术有限公司
深圳市羽诺电子有限公司
海宁美康环保科技有限公司
广东海状元环境科技有限公司
杭州阿里巴巴广告有限公司
上海宝獻建材有限公司
河南艾力健药业有限公司
山东舞鹏传媒股份有限公司
浙江美达瑞新材料科技有限公司
东莞市喜马拉雅环保科技有限公司
广东咏声动漫股份有限公司
深圳市攀高电子有限公司
福建中金在线网络股份有限公司
深圳市腾讯计算机系统有限公司
浙江淘宝网络有限公司
万福阁家具（昆山）有限公司
天涯社区网络科技股份有限公司
北京全景视觉网络科技股份有限公司
华声在线股份有限公司
江苏苏宁易购电子商务有限公司
北京奇虎科技有限公司
北京微梦创科网络技术有限公司
北京新浪互联信息服务有限公司
阿里云计算有限公司
淘宝（中国）软件有限公司
深圳市一达通企业服务有限
北京京东世纪贸易有限公司
阿里巴巴（中国）网络技







- **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**
  - o *n* data points with *p* dimensions
  - o Two modes
- **Dissimilarity matrix**
  - o *n* data points, but registers only the distance
  - o A triangular matrix
  - o Single mode

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

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

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

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

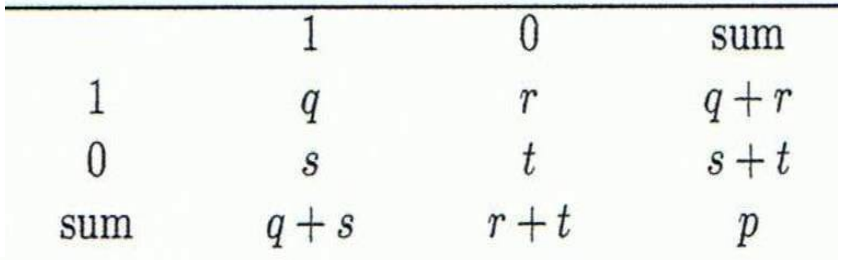
- **Method 2:** Use a large number of binary attributes
  - creating a new binary attribute for each of the M nominal states

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:**

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

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

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

✓ **Note: Jaccard coefficient is the same as “coherence”:**

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

Dissimilarity between Binary Variables

Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

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

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

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

Standardizing Numeric Data

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

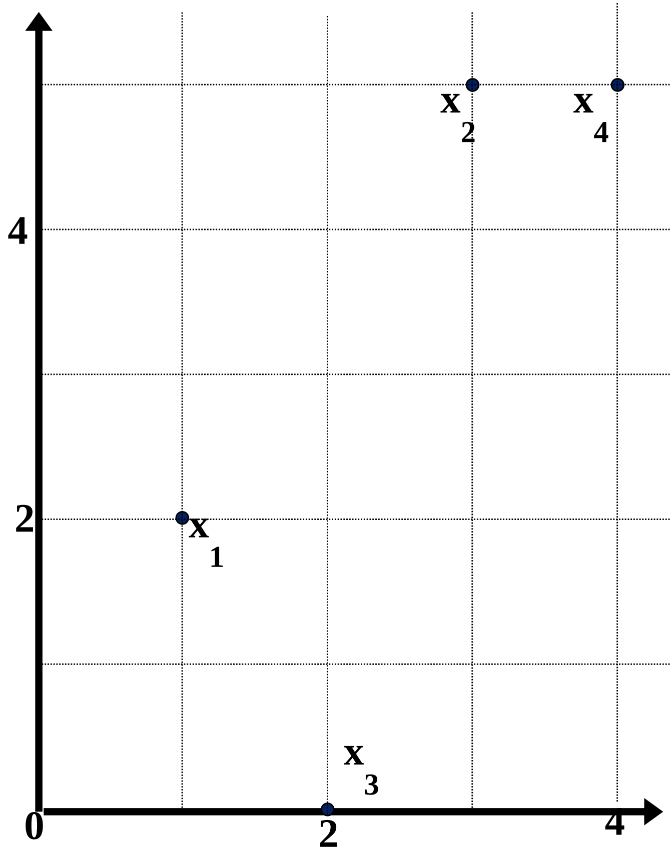
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where

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

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

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

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

Dissimilarity Matrix

(with 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**

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

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

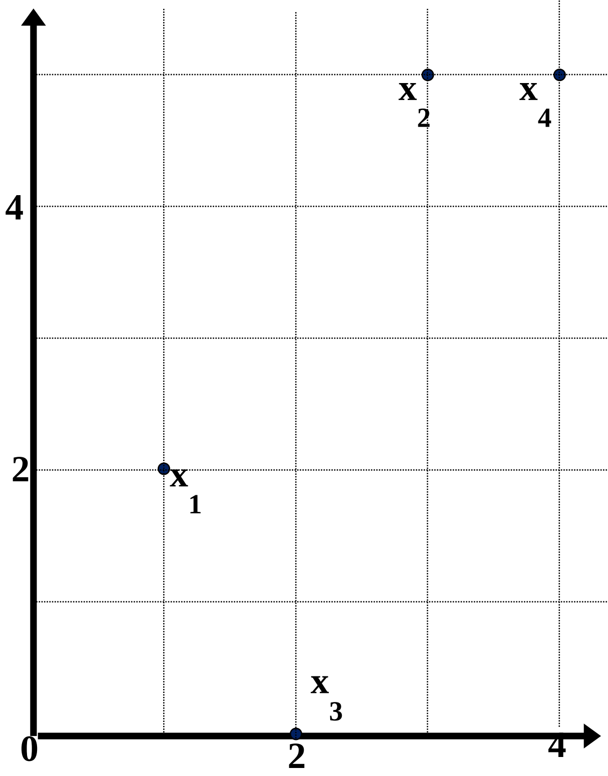
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$. **"Supremum" (L$_{max}$ norm, L$_\infty$ 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 (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_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Ordinal Variables

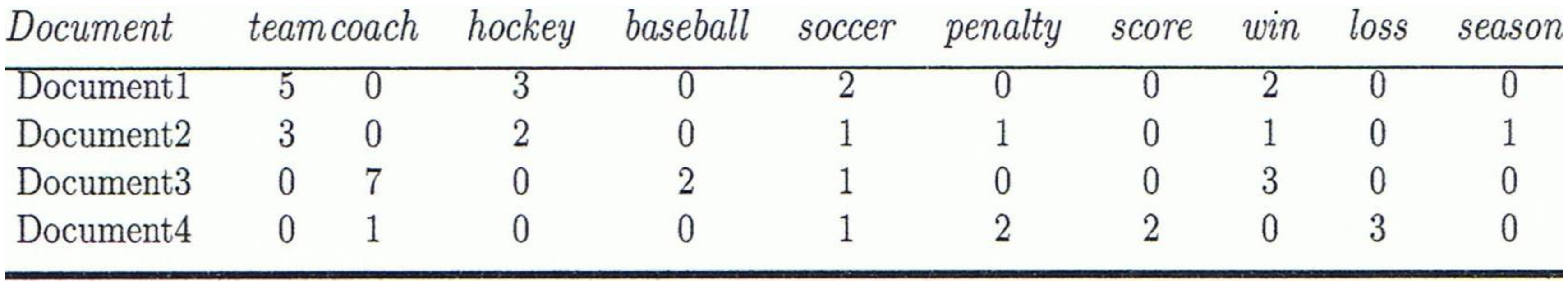
- **An ordinal variable can be discrete or continuous**
- **Order is important, e.g., rank**
- **Can be treated like interval-scaled**
  - **replace x_{if} by their rank** $r_{if} \in \{1, \ldots, M_f\}$
  - **map the range of each variable onto [0, 1] by replacing *i-th* object in the *f-th* variable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

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

  - **compute the dissimilarity using methods for interval-scaled variables**

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

Example: Cosine Similarity

- cos(d_1, d_2) = (d_1 • d_2) /||d_1|| ||d_2|| ,
  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)

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

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**

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