

# **Data Mining:**

---

## **Concepts and Techniques**

### **— Chapter 2 —**

Jianjun Cheng

School of Information Science & Engineering

Lanzhou University

©2011 Han, Kamber, and Pei. All rights reserved.

# 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

- Record

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

- 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:

<i><b>TID</b></i>	<i><b>Items</b></i>
<b>1</b>	<b>Bread, Coke, Milk</b>
<b>2</b>	<b>Beer, Bread</b>
<b>3</b>	<b>Beer, Coke, Diaper, Milk</b>
<b>4</b>	<b>Beer, Bread, Diaper, Milk</b>
<b>5</b>	<b>Coke, Diaper, Milk</b>

# Data Objects

---

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples, examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns -> attributes.

# Important Characteristics of Structured Data

---

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

# 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

- Binary

- ordinal

- Numeric: quantitative

- Interval-scaled

- Ratio-scaled

4.8	3	1.4	0.1	Iris-setosa
4.3	3	1.1	0.1	Iris-setosa
5.8	4	1.2	0.2	Iris-setosa
...				
7	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.5	2.3	4	1.3	Iris-versicolor
...				
6.3	3.3	6	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
7.1	3	5.9	2.1	Iris-virginica
6.3	2.9	5.6	1.8	Iris-virginica

# Attribute Types

- **Nominal(categorical):** categories, states, or "names of things"
  - *Hair\_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

# Numeric Attribute Types

---

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



# Discrete vs. Continuous Attributes

---

## ■ Discrete Attribute

- Has only a finite or countably infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

## ■ Continuous Attribute

- Has real numbers as attribute values
  - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables

# 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



# Basic Statistical Descriptions of Data

---

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

# Measuring the Central Tendency

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

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

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

  - Weighted arithmetic mean:
 
$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$
  - Trimmed mean: chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):
- Mode

$$median = L_1 + \left( \frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:  $mean - mode = 3 \times (mean - median)$

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

# Example for median

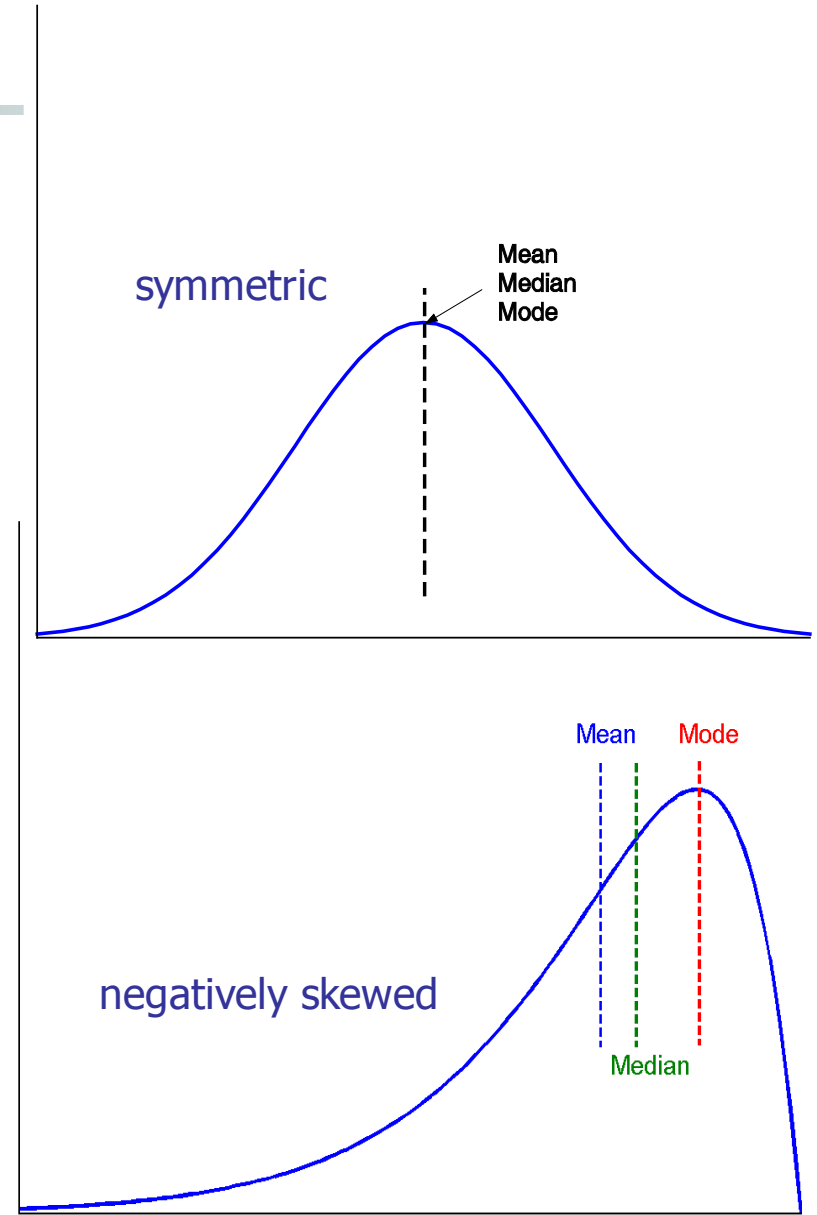
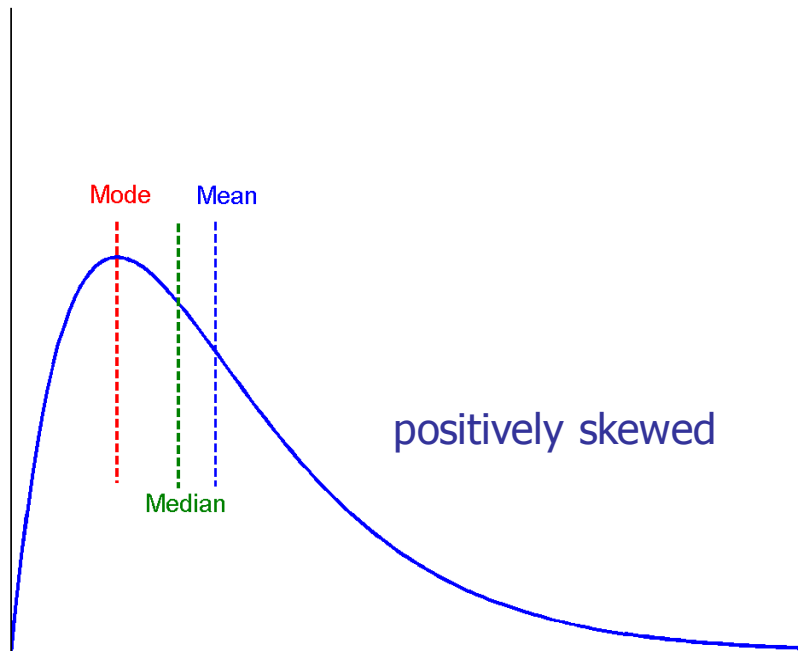
Group	Frequency	Accu. Freq.
0-5	2	2
5-15	6	8
15-25	20	28
25-35	15	43
35-45	8	51
>45	4	55

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

- $N/2=27.5$ , median is in interval 15-25
- $median=15+(27.5-8)/20*10=24.75$

# Symmetric vs. Skewed Data

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



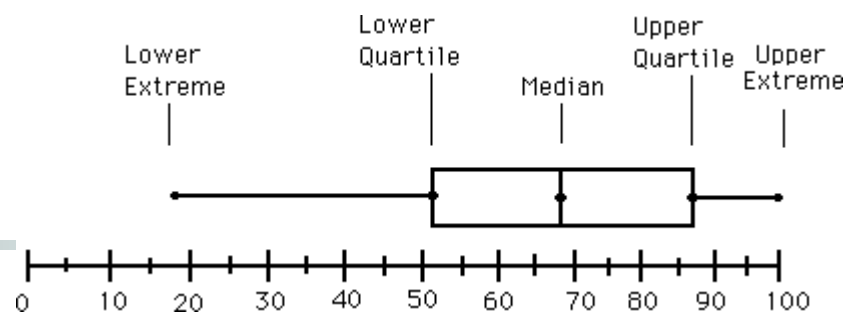
# Measuring the Dispersion of Data

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

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

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

# Boxplot Analysis



- **Five-number summary** of a distribution

- Minimum,  $Q_1$ , Median,  $Q_3$ , Maximum

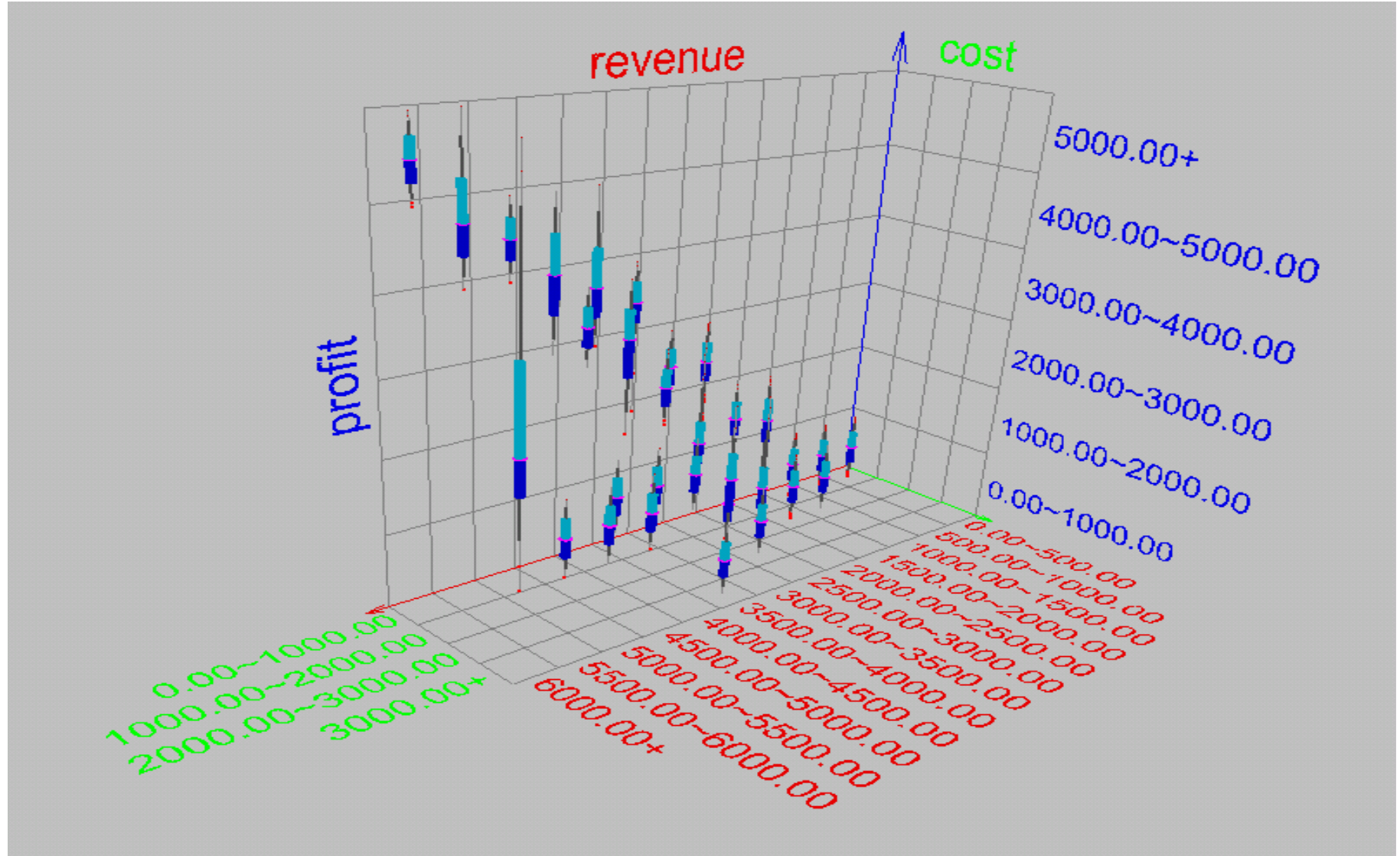
- **Boxplot**

- 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 IQR
- The median is marked by a line within the box
- Whiskers: two lines outside the box extended to Minimum and Maximum
- Outliers: points beyond a specified outlier threshold, plotted individually





# Visualization of Data Dispersion: 3-D Boxplots



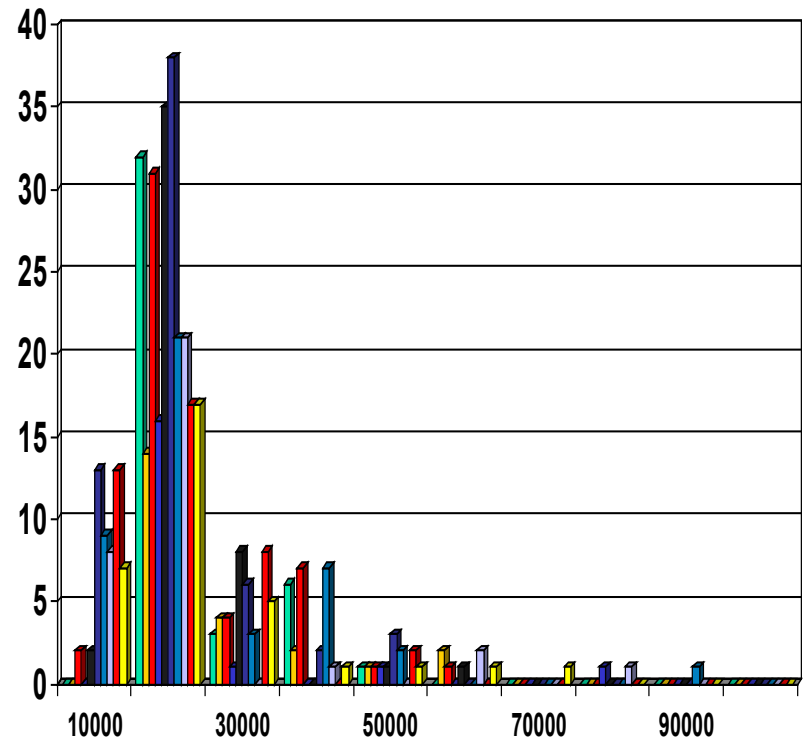
# Graphic Displays of Basic Statistical Descriptions

---

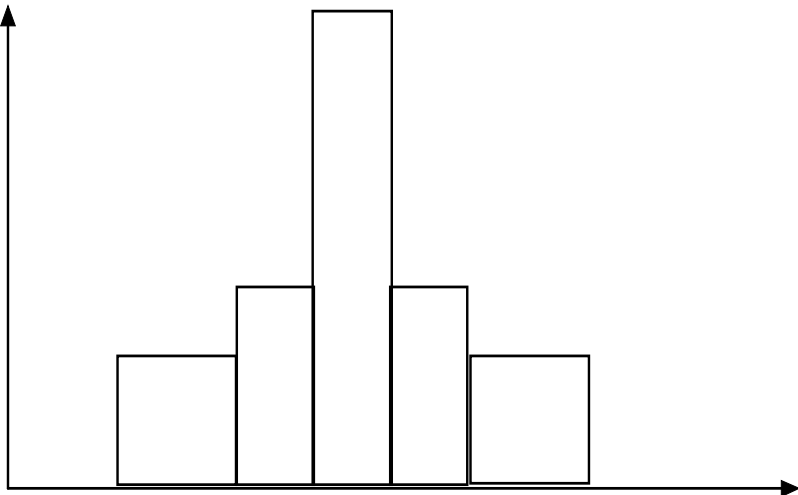
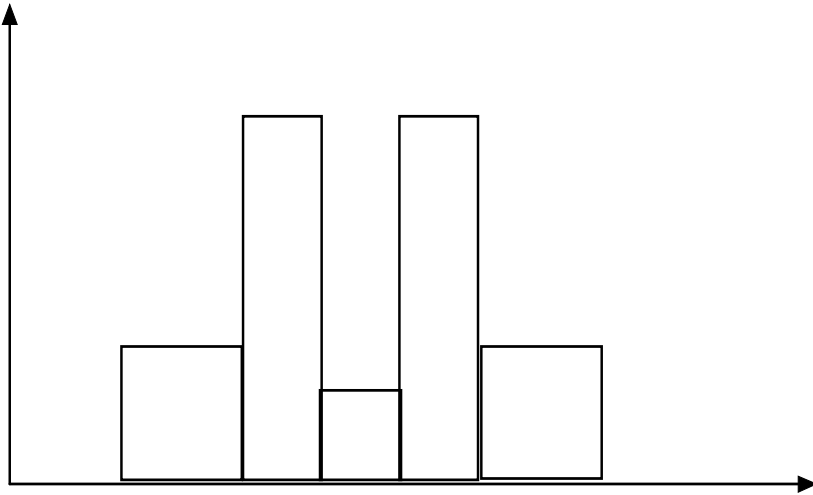
- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis repres. frequencies
- **Quantile plot:** each value  $x_i$  is paired with  $f_i$  indicating that approximately  $100 f_i \%$  of data are  $\leq x_i$
- **Quantile-quantile (q-q) plot:** graphs the quantiles 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



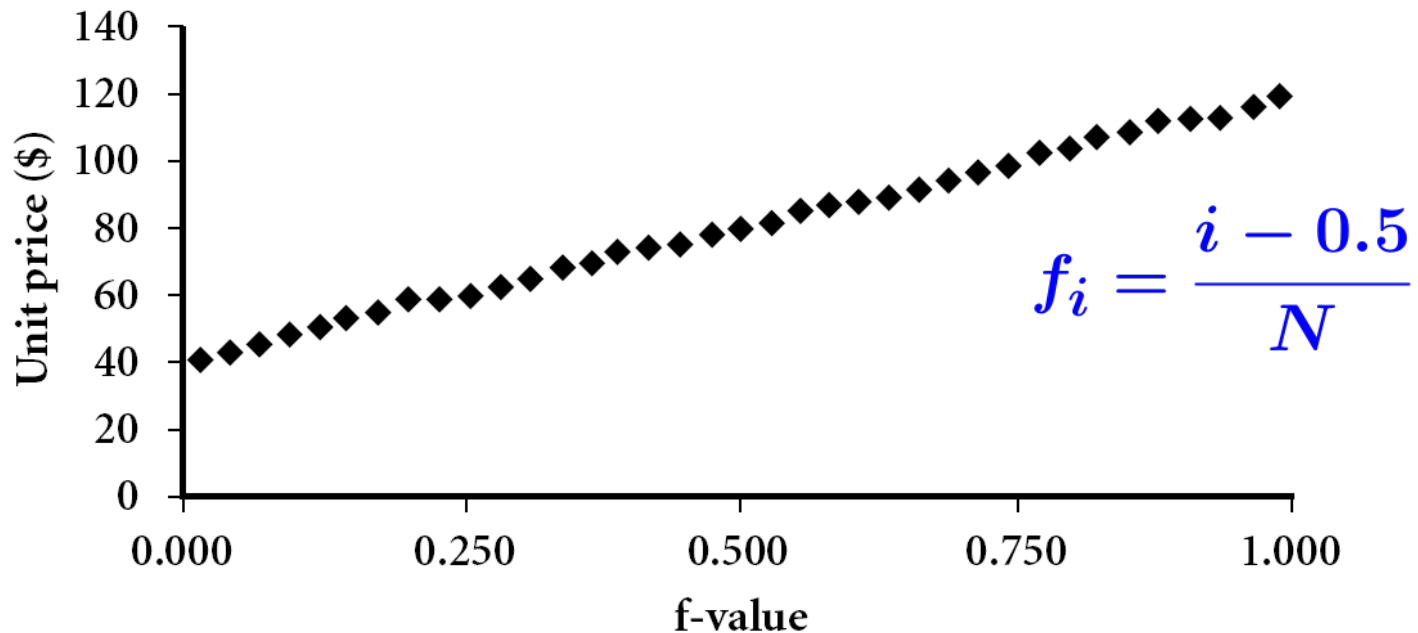
# Histograms Often Tell More than Boxplots



- 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

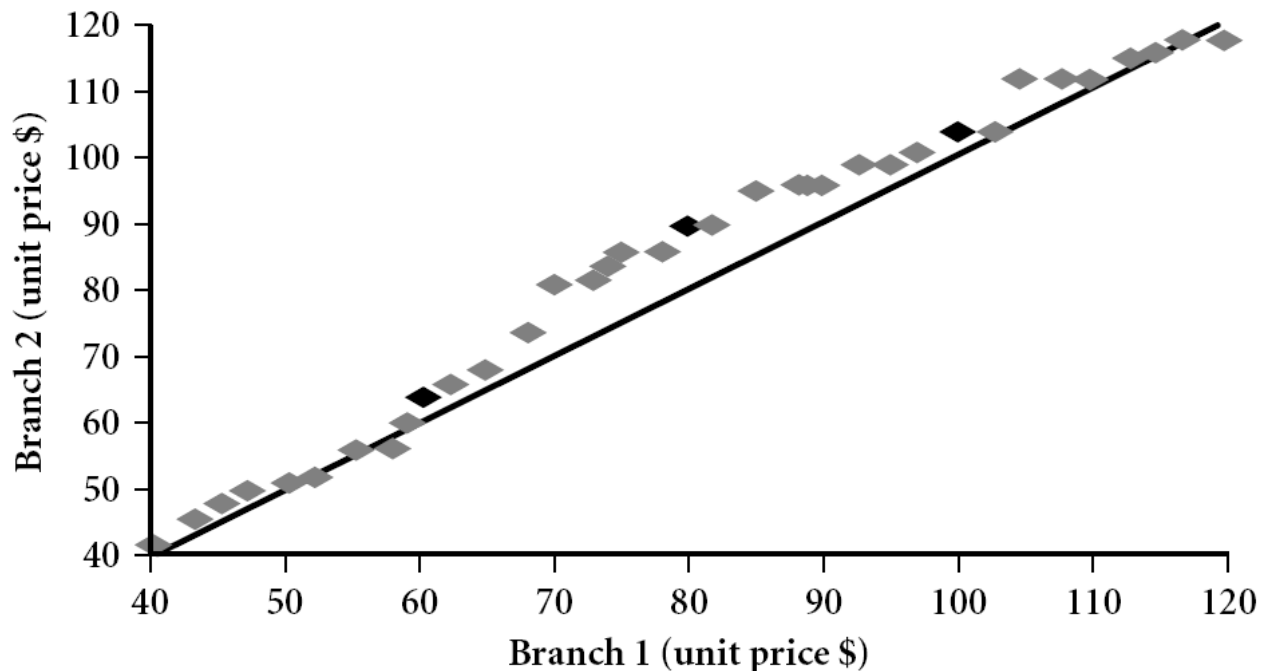
# Quantile Plot

- 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  $100 \times f_i\%$  of the data are below or equal to the value  $x_i$



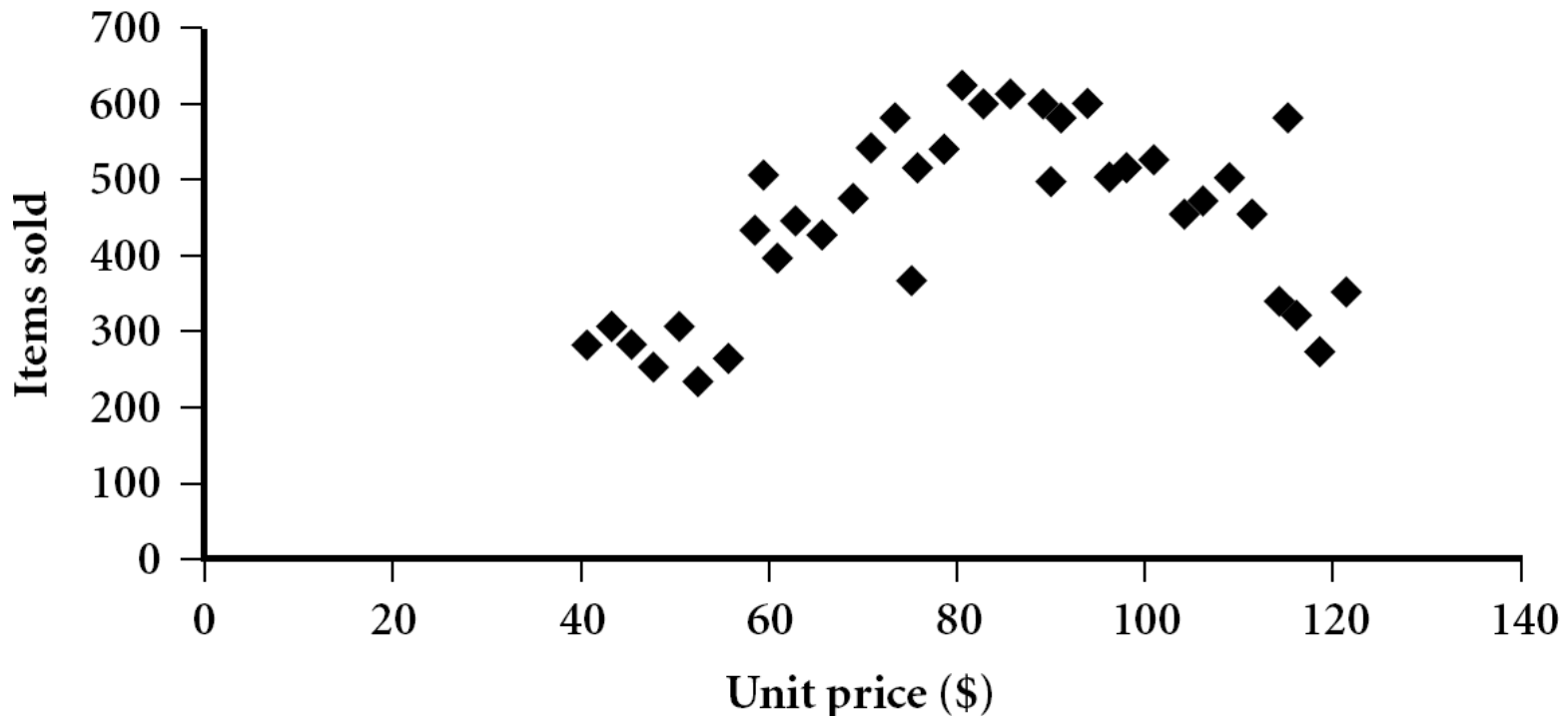
# Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

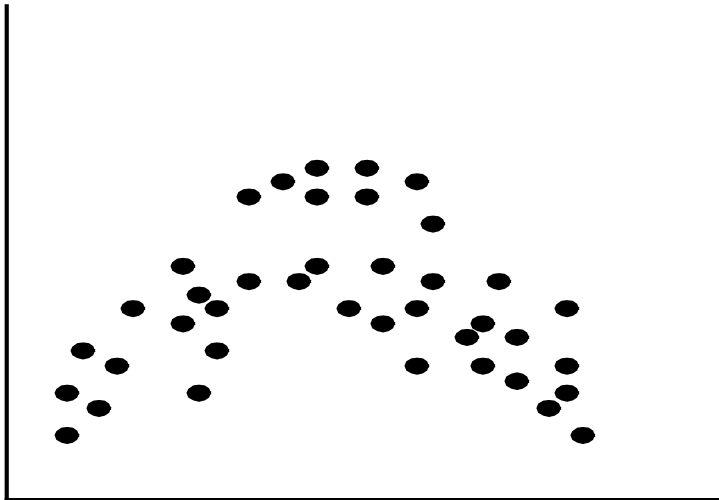
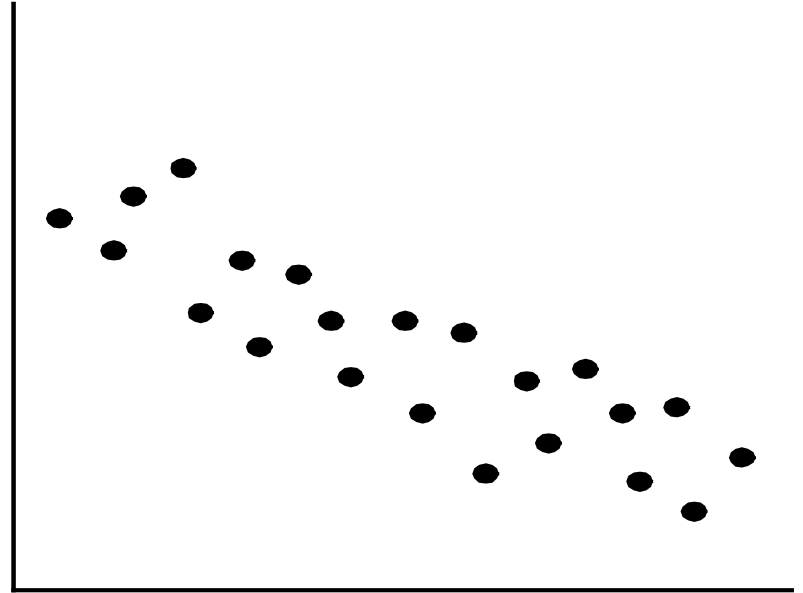
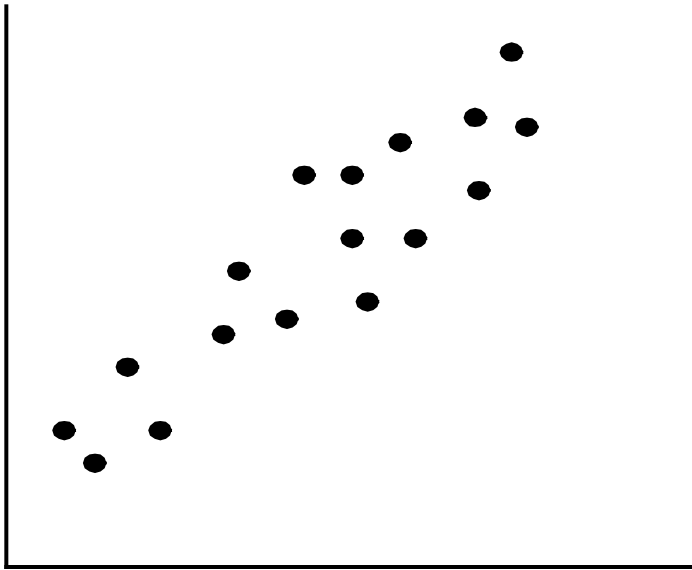


# Scatter plot

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



# Positively and Negatively Correlated Data

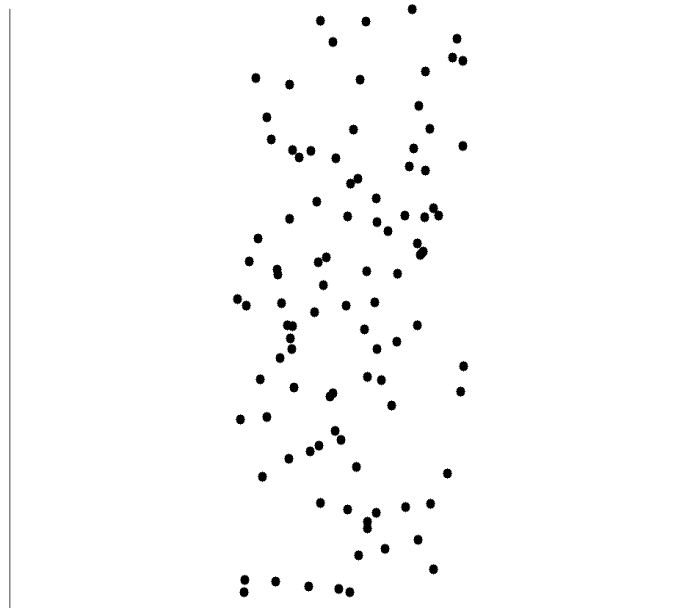
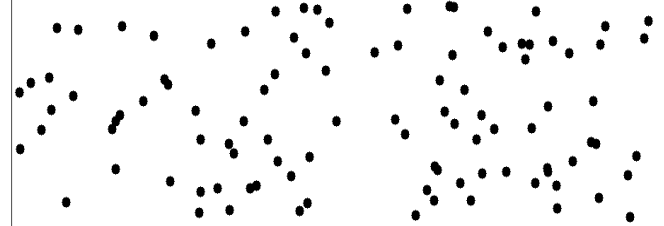
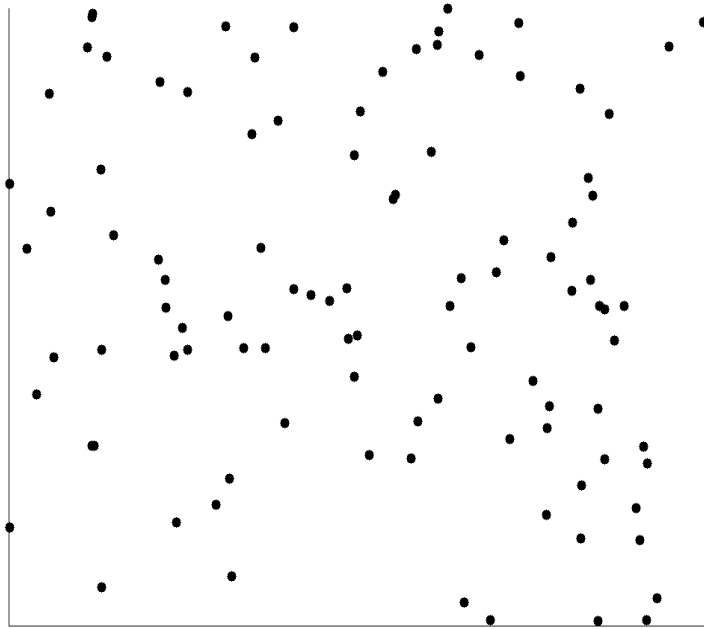


- The left half fragment is positively correlated
- The right half is negative correlated



# Uncorrelated Data

---



# 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



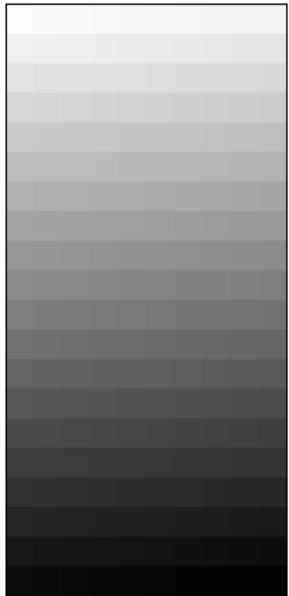
# Data Visualization

---

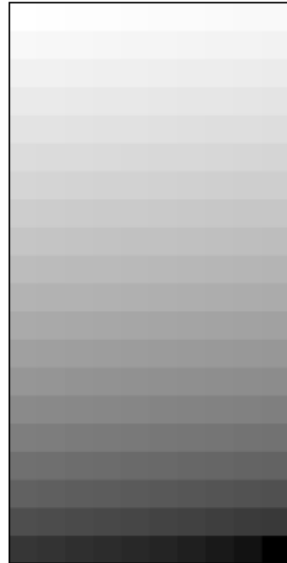
- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters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



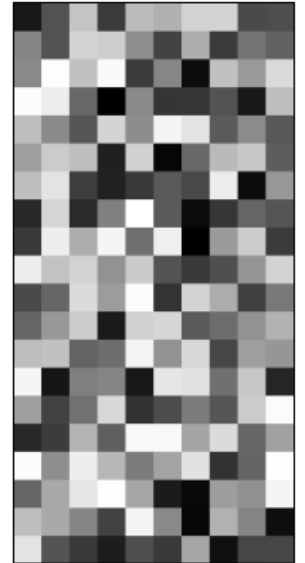
(a) Income



(b) Credit Limit



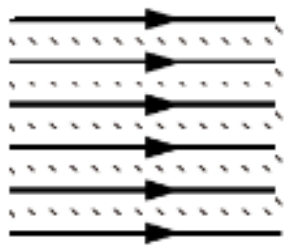
(c) transaction volume



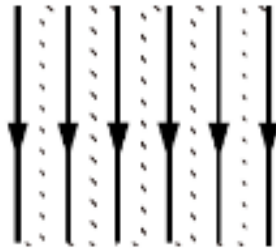
(d) age

# Pixel-Oriented Visualization Techniques

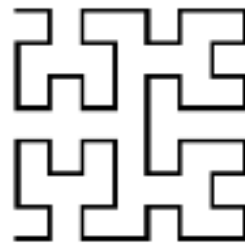
- To solve problems resulted from filling windows in a linear way, we can lay out the data records in a space-filling curve to fill the windows
- A space-filling curve is a curve with a range that covers the entire n-dimensional unit hypercube



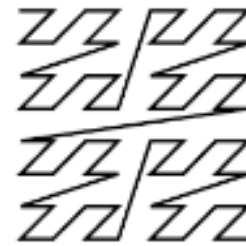
a. line-by-line



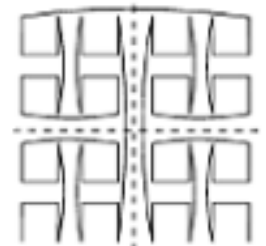
b. column-by-column



c. Peano-Hilbert



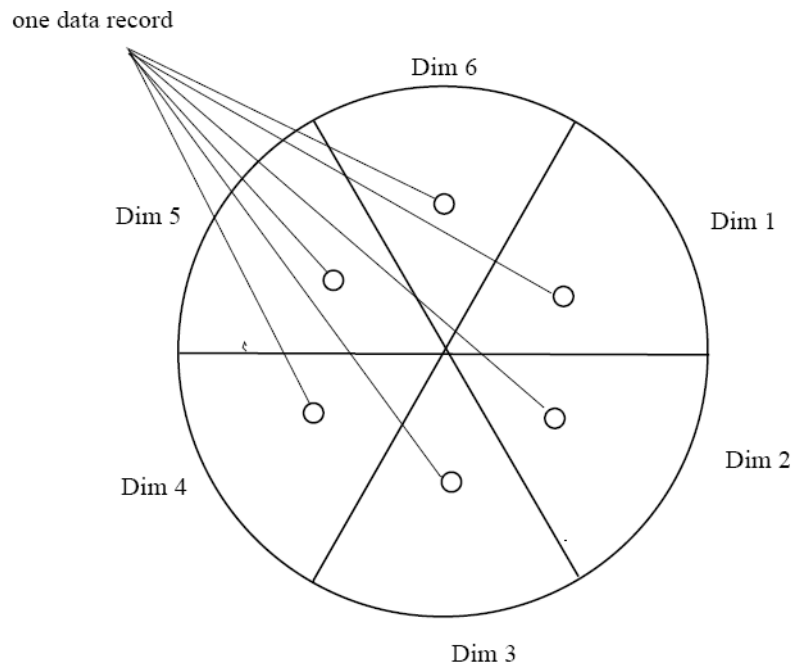
d. Morton



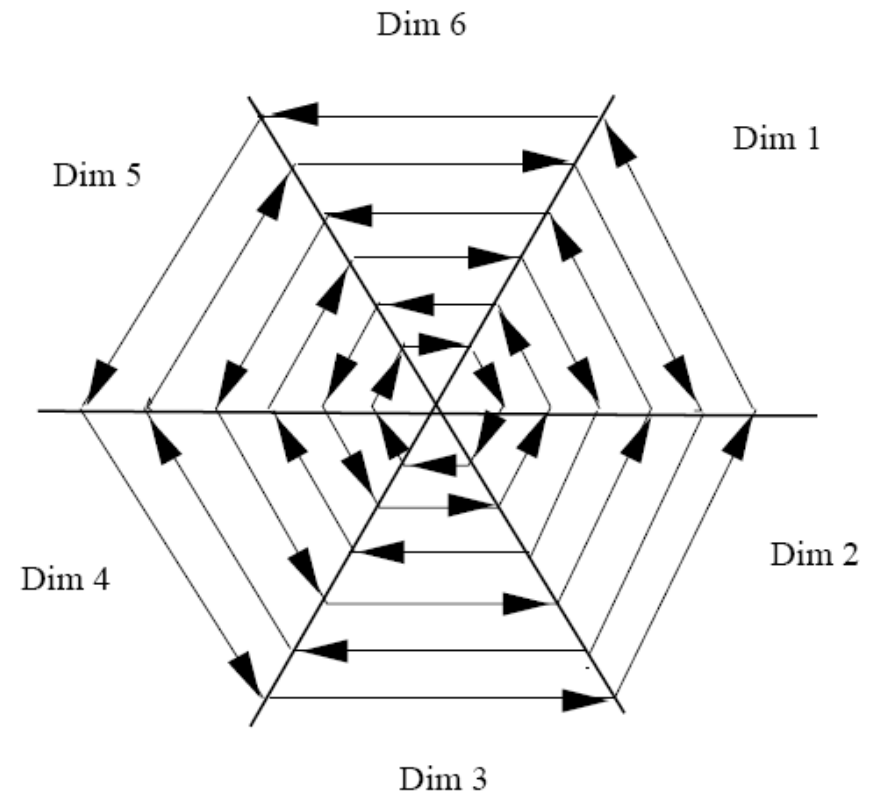
e. Gray code

# Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment



(a) Representing a data record in circle segment



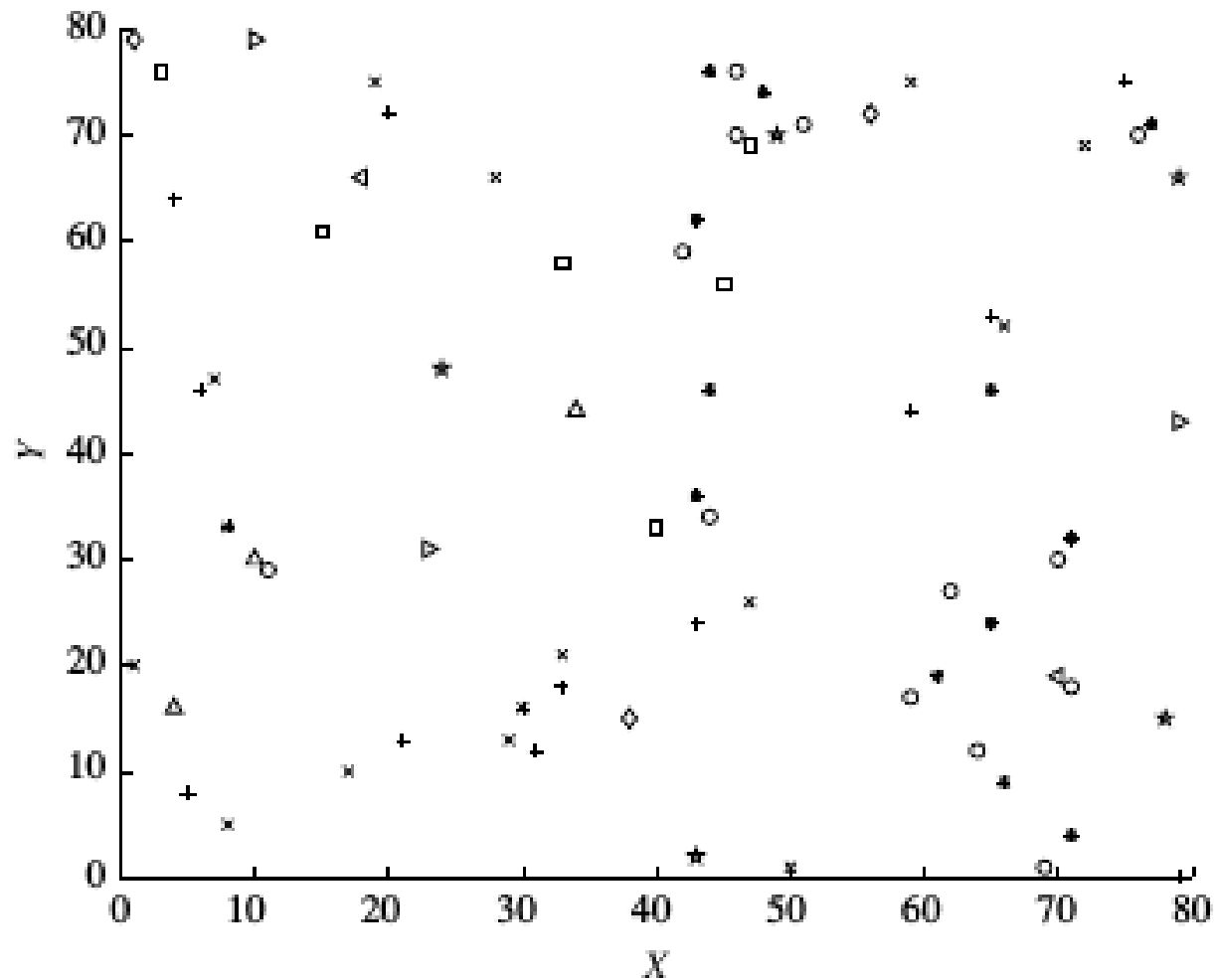
(b) Laying out pixels in circle segment

# Geometric Projection Visualization Techniques

---

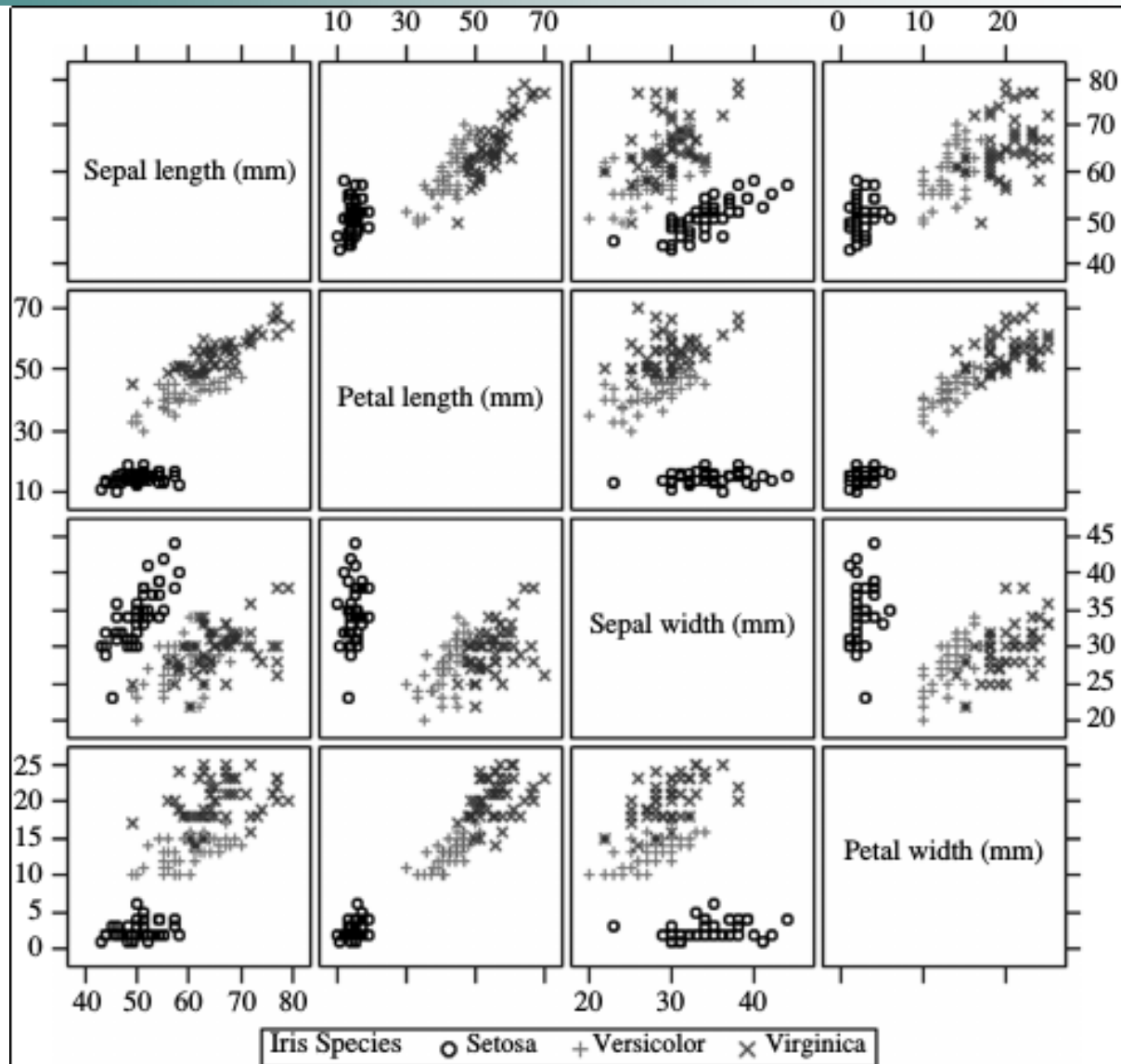
- Visualization of geometric transformations and projections of the data
- Help users find interesting projections of multidimensional data sets
- problem: how to visualize a high-dim. space on 2-D display
- Methods
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Hyperslice
  - Parallel coordinates

# Scatterplot



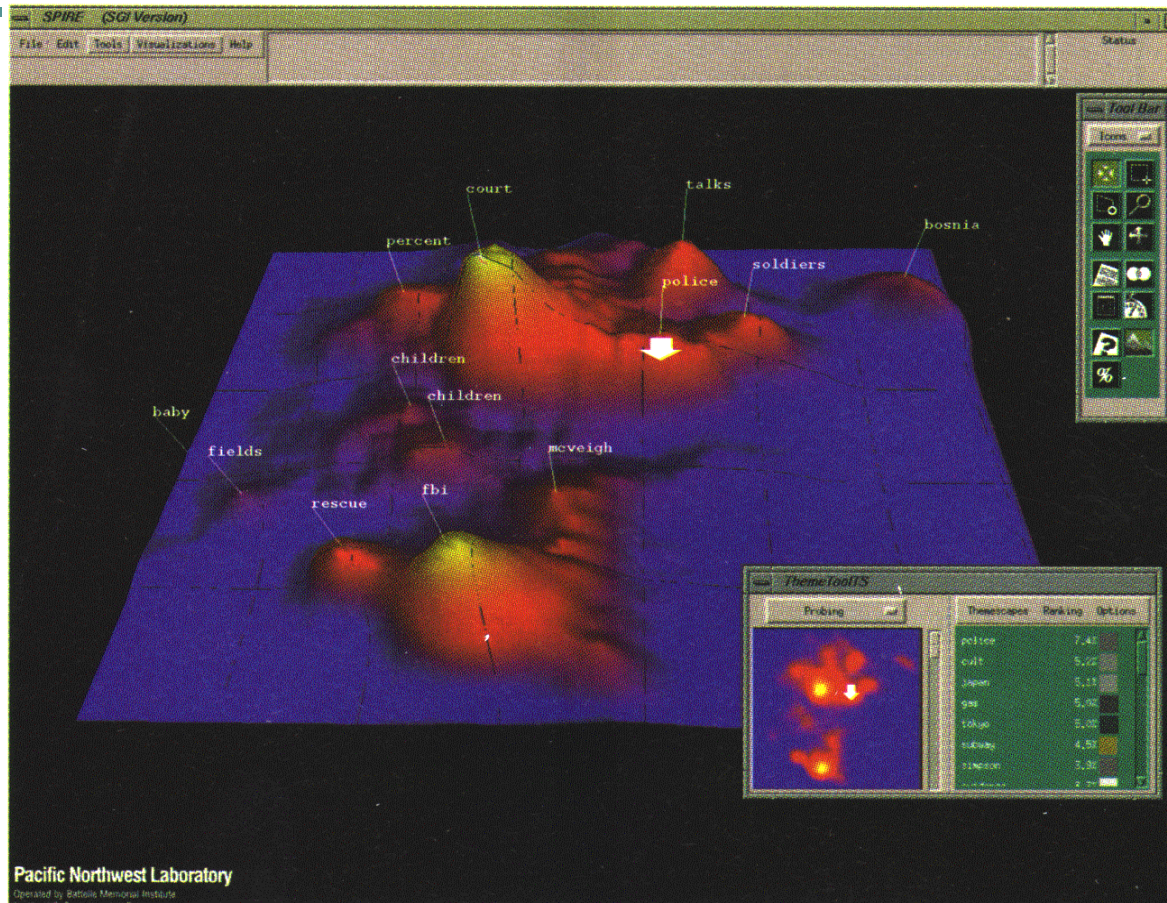


# Scatterplot Matrices



# Landscapes

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

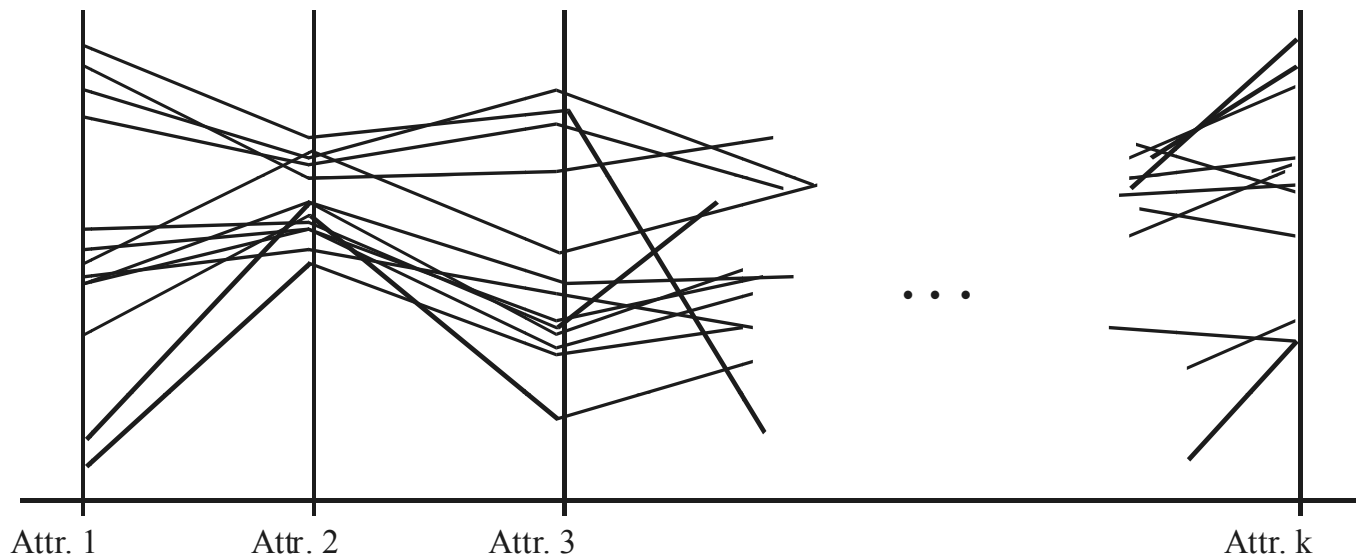


news articles  
visualized as  
a landscape

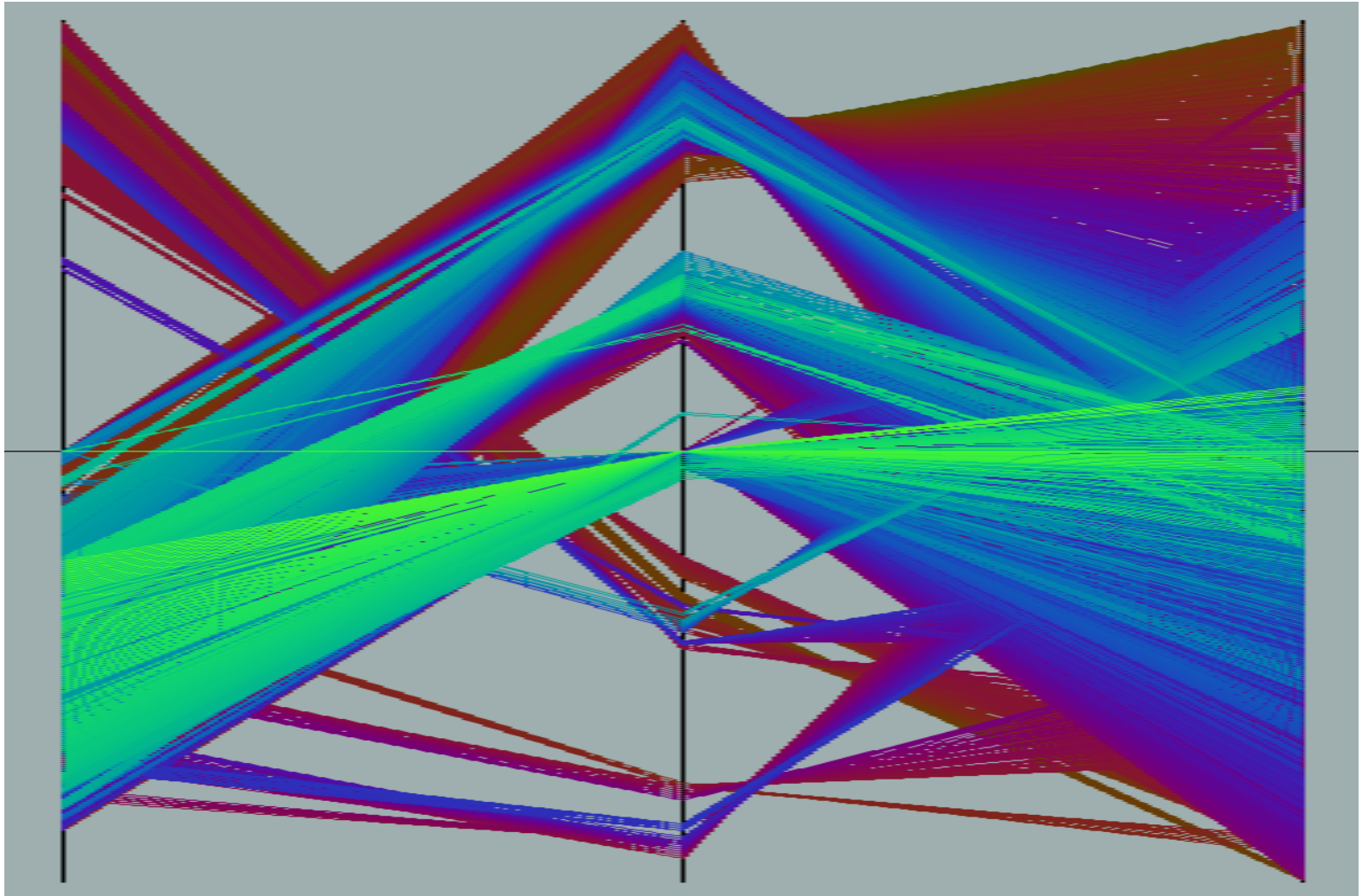
- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

# 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





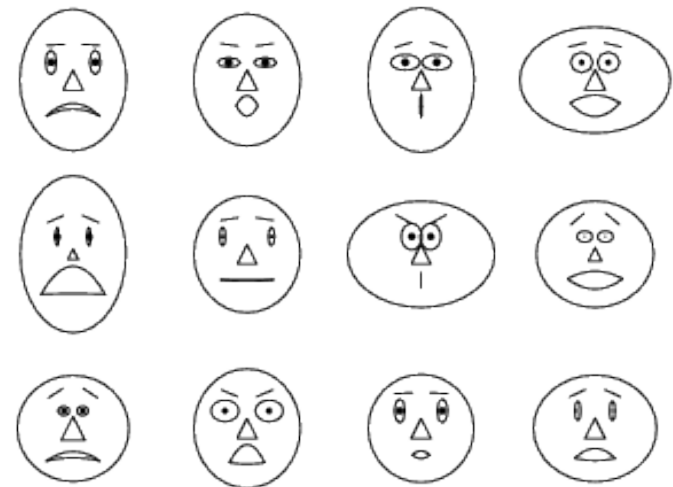
# Icon-Based Visualization Techniques

---

- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Stick Figures
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

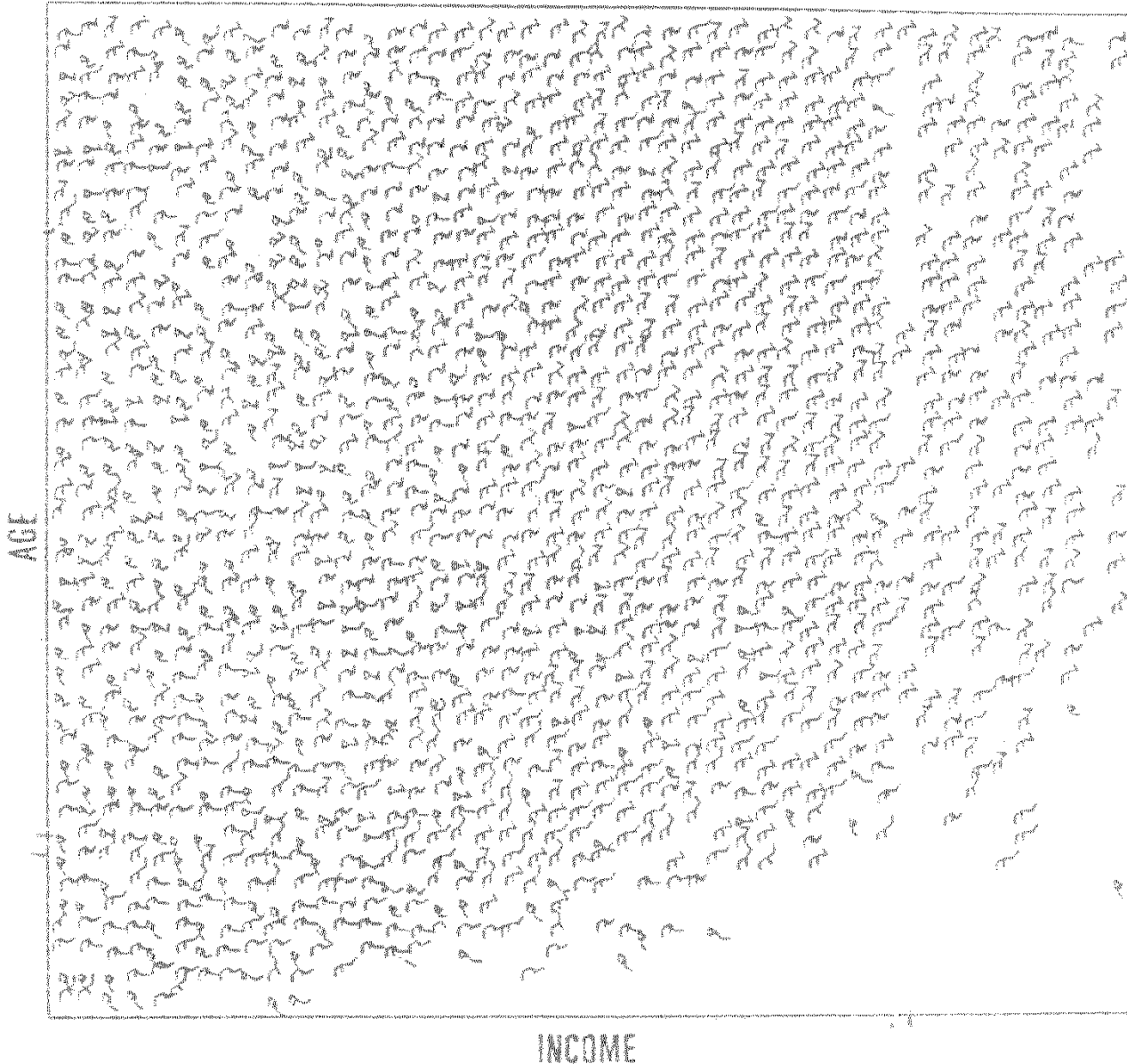
# Chernoff Faces

- A way to display variables on a two-dimensional surface, e.g., let  $x$  be eyebrow slant,  $y$  be eye size,  $z$  be nose length, etc.
- The figure shows faces produced using 10 characteristics--head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using [Mathematica](#) (S. Dickson)
- REFERENCE: Gonick, L. and Smith, W. [\*The Cartoon Guide to Statistics\*](#). New York: Harper Perennial, p. 212, 1993
- Weisstein, Eric W. "Chernoff Face." From *MathWorld*--A Wolfram Web Resource. [mathworld.wolfram.com/ChernoffFace.html](http://mathworld.wolfram.com/ChernoffFace.html)



# Stick Figure

used by permission of G. Grinstein, University of Massachusetts at Lowell



A census data figure showing age, income, gender, education, etc.

A 5-piece stick figure (1 body and 4 limbs w. different angle/length)

Two attributes mapped to axes, remaining attributes mapped to angle or length of limbs". Look at texture pattern

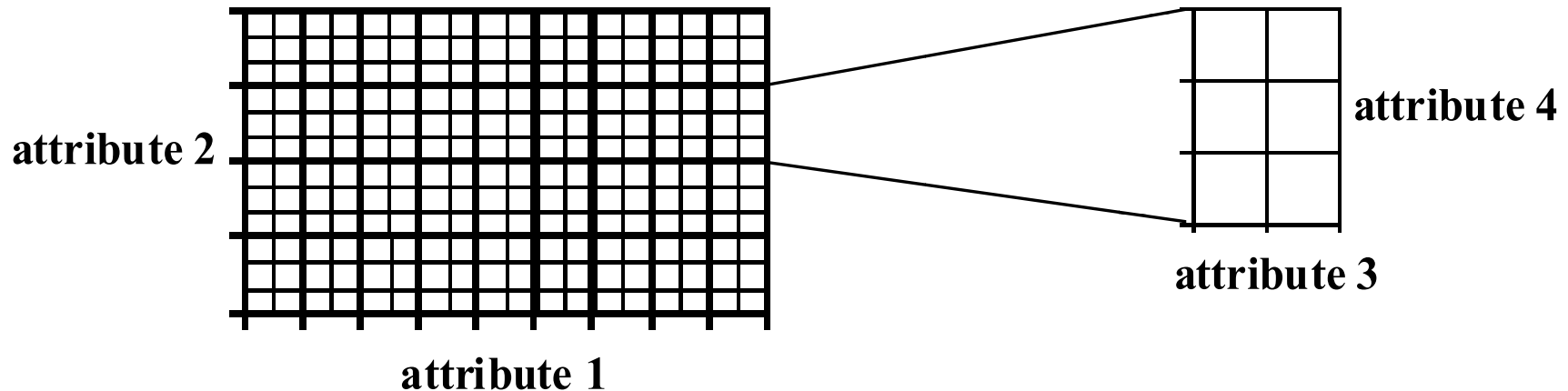
# Hierarchical Visualization Techniques

---

- Visualization of the data using a hierarchical partitioning into subspaces
- Methods
  - Dimensional Stacking
  - Worlds-within-Worlds
  - Tree-Map
  - Cone Trees
  - InfoCube



# Dimensional Stacking



- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately

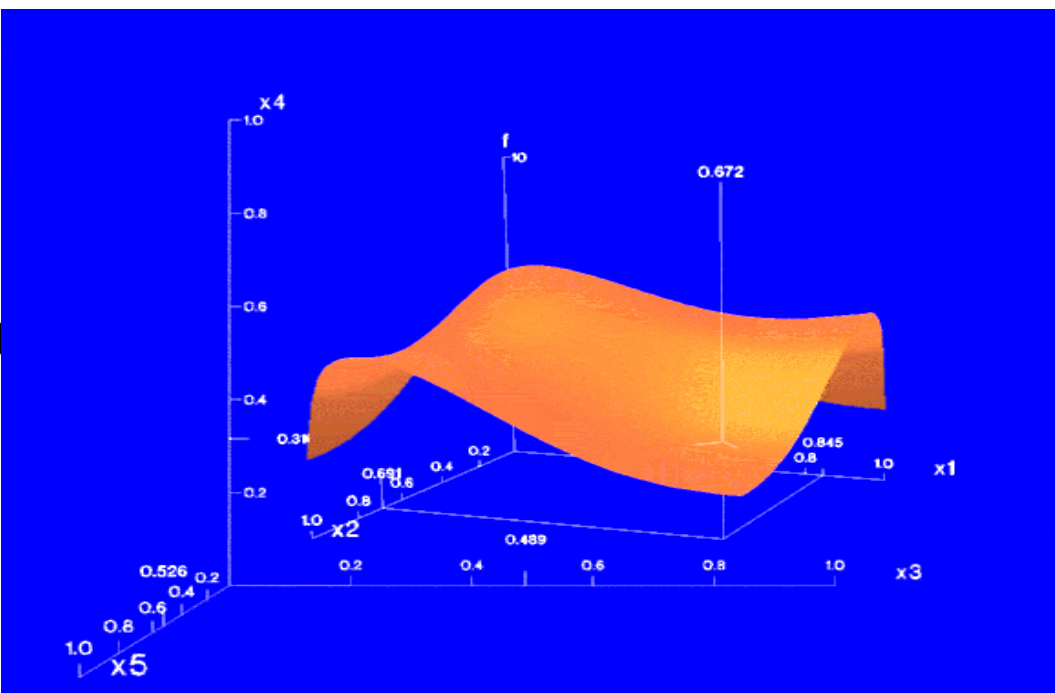
# Dimensional Stacking

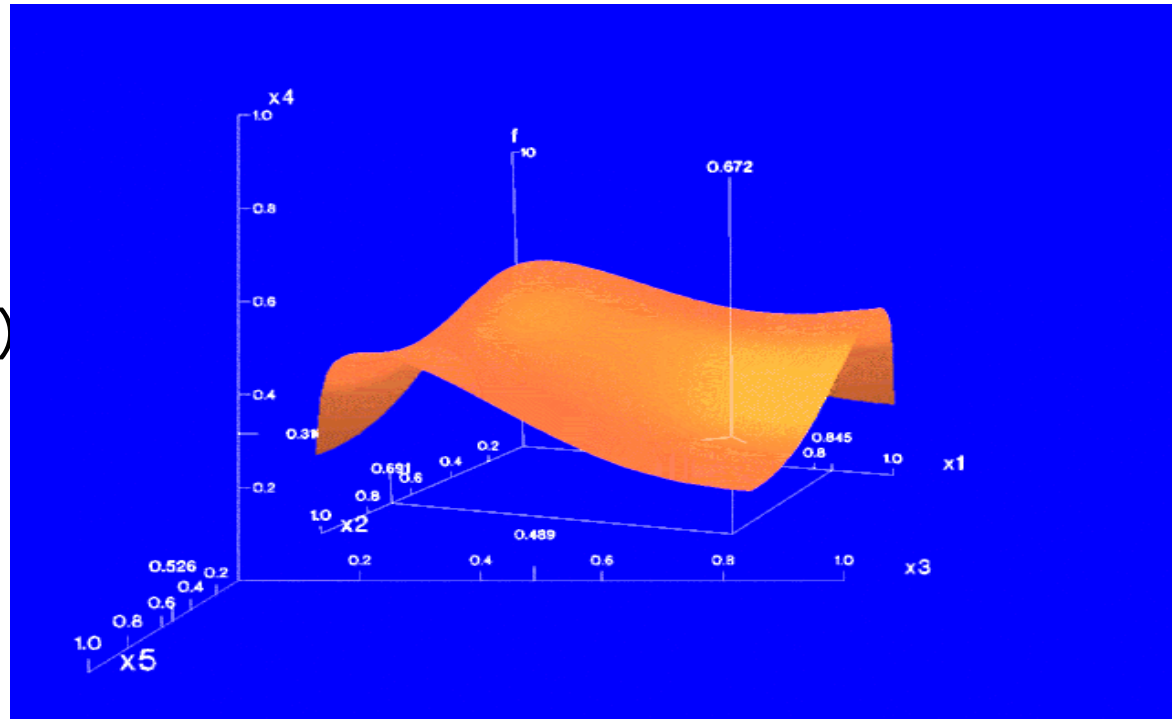
var	min	max
postgraduate	0	10
teacher	0	20
tutor	0	10
research assistant	0	10



The highlighted cell means postgraduate (2.5-5.0, horiz. outer), teacher (0-5, vert. outer), tutor (0-2.5, horiz. inner), research assistant (2.5-5.0, vert. inner)

# Worlds-within-Worlds

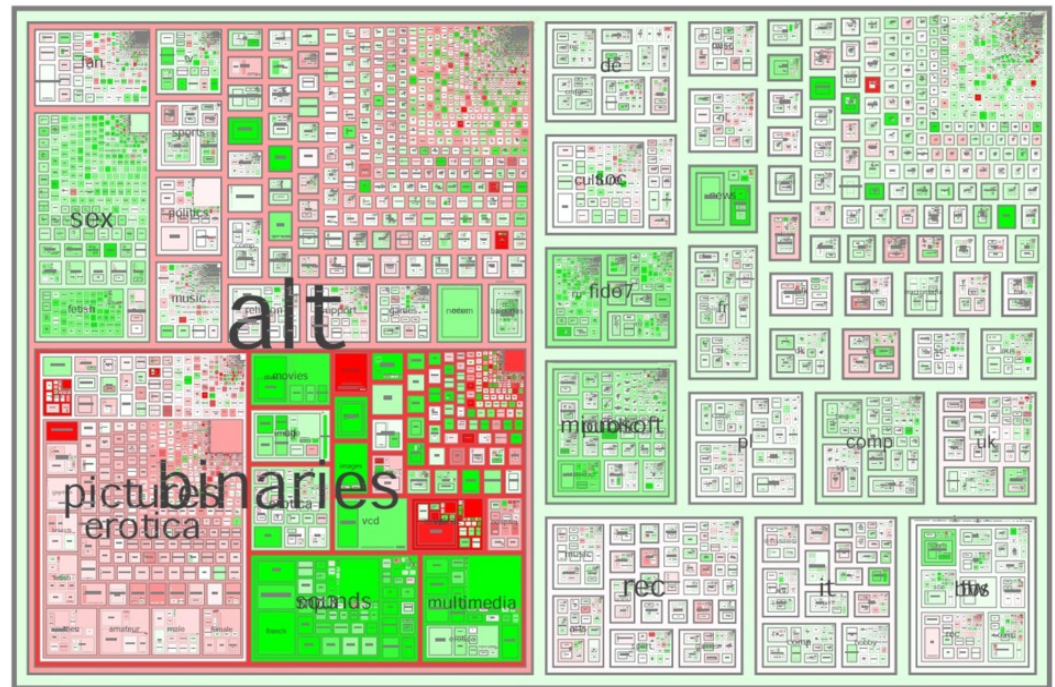
- Assign the function and two most important parameters to innermost world
  - Fix all other parameters at constant values - draw other (1 or 2 or 3 dimensional worlds choosing these as the axes)
  - Software that uses this paradigm
    - N-vision: Dynamic interaction through data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
    - Auto Visual: Static interaction by means of queries
- 



# Tree-Map

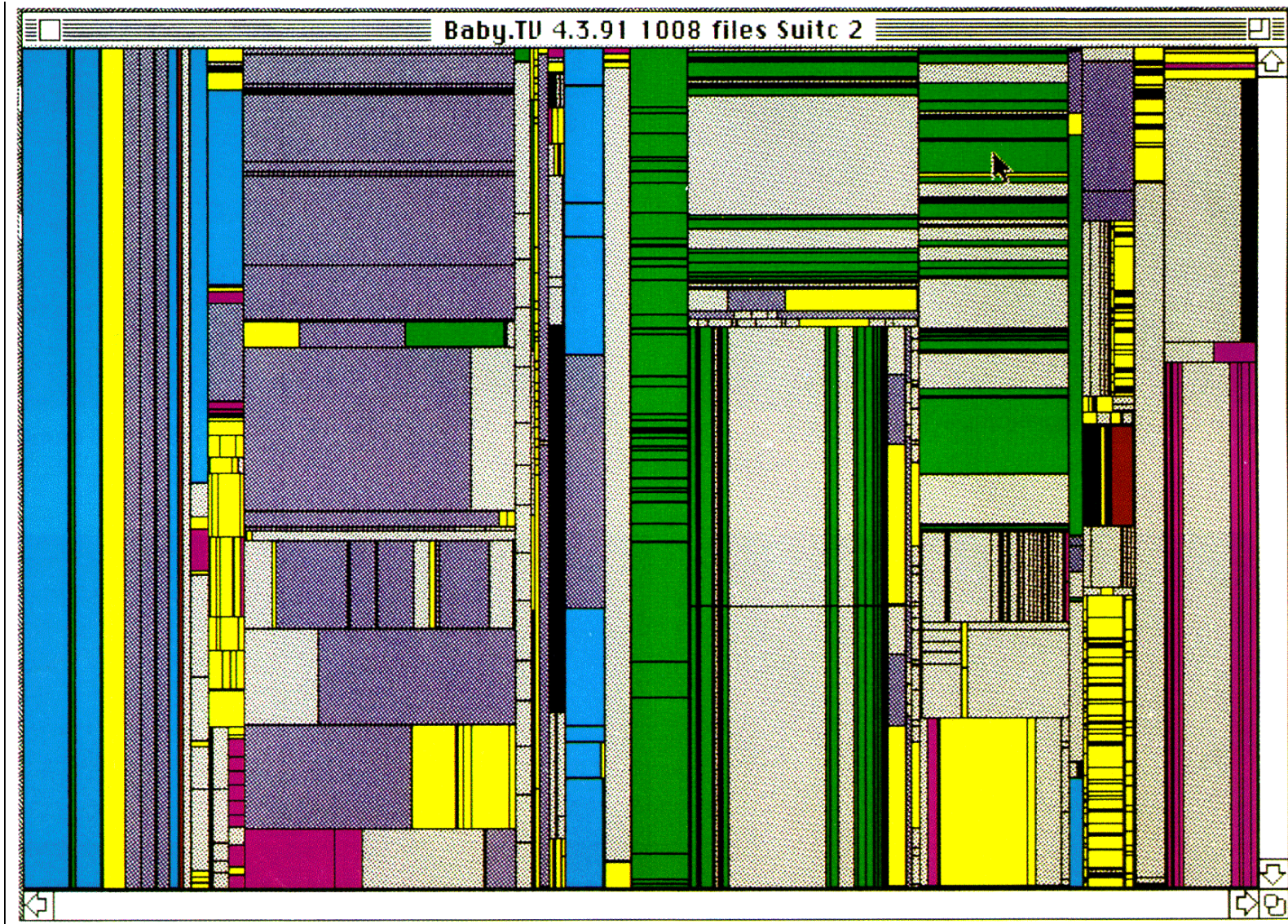
- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)

MSR Netscan Image



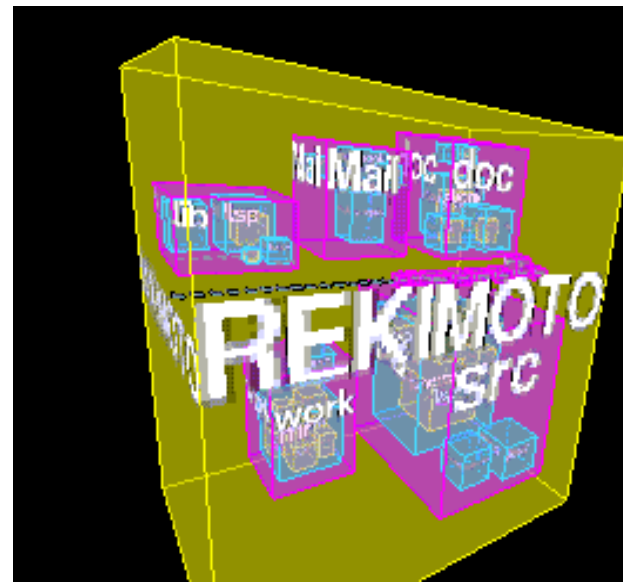


# Tree-Map of a File System (Schneiderman)



# InfoCube

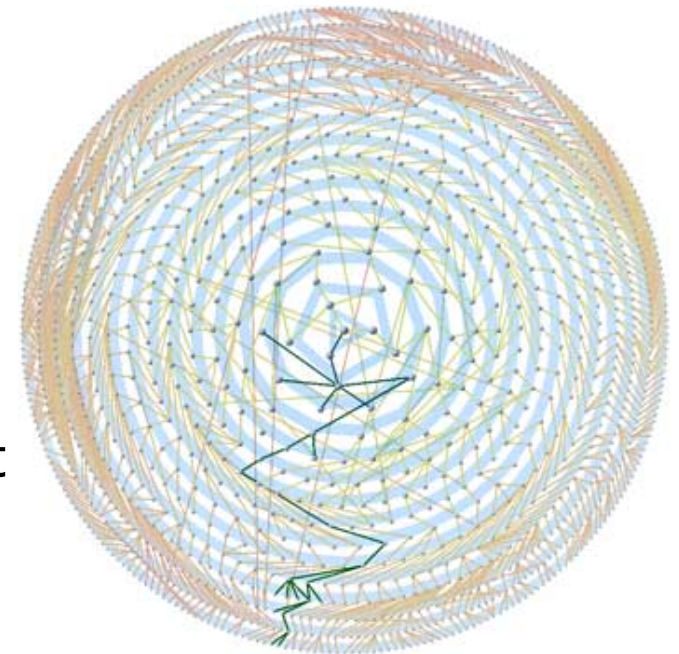
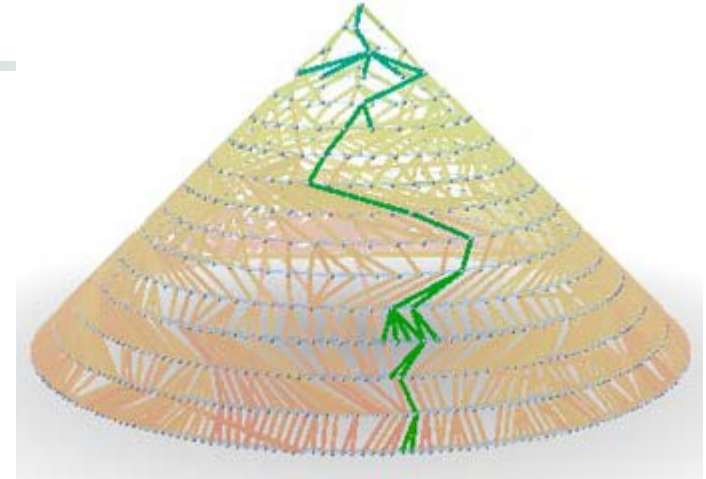
- A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes
- The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on





# Three-D Cone Trees

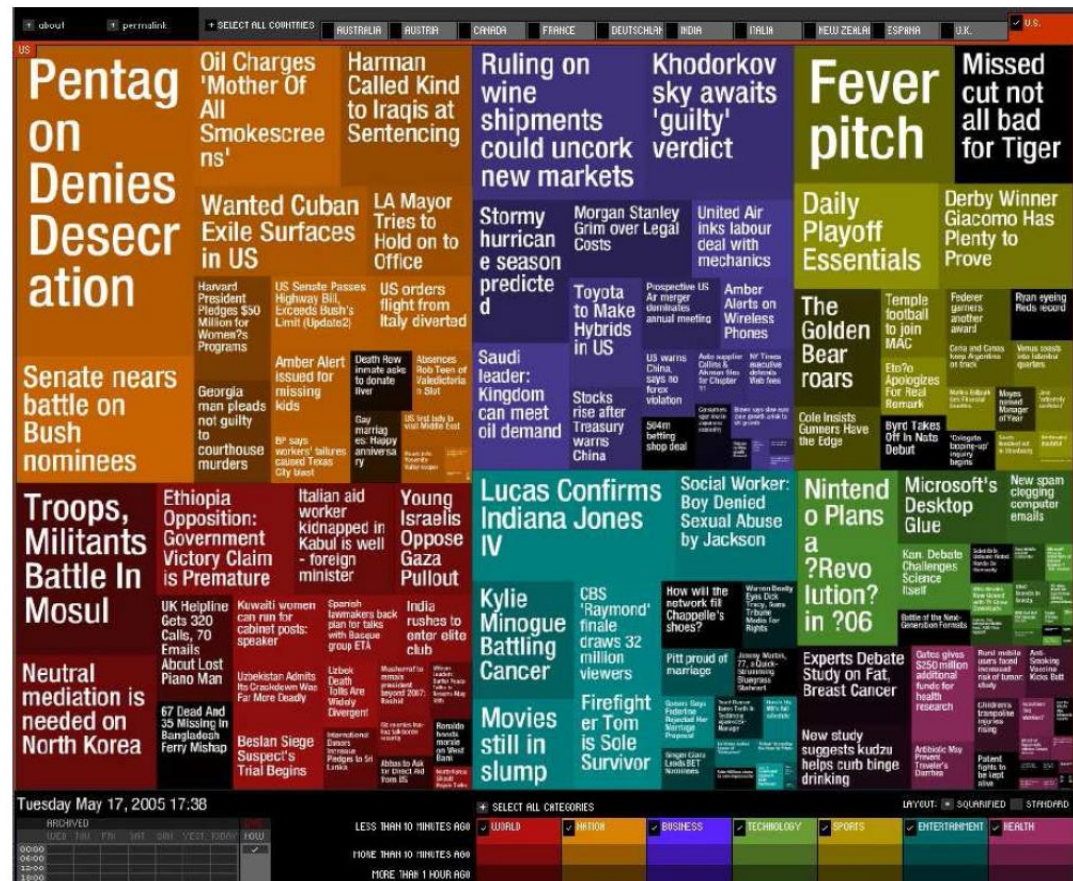
- *3D cone tree* visualization technique works well for up to a thousand nodes or so
- First build a *2D circle tree* that arranges its nodes in concentric circles centered on the root node
- Cannot avoid overlaps when projected to 2D
- G. Robertson, J. Mackinlay, S. Card. "Cone Trees: Animated 3D Visualizations of Hierarchical Information", *ACM SIGCHI'91*
- Graph from Nadeau Software Consulting website: Visualize a social network data set that models the way an infection spreads from one person to the next



# Visualizing Complex Data and Relations

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags

- The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing social networks



Newsmap: Google News Stories in 2005



# 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 and Dissimilarity

---

- **Similarity**

- Numerical measure of how alike two data objects are
- Value is higher when objects are more alike
- Often falls in the range  $[0,1]$

- **Dissimilarity** (e.g., distance)

- Numerical measure of how different two data objects are
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

- **Proximity** refers to a similarity or dissimilarity

# Data Matrix and Dissimilarity Matrix

## ■ Data matrix

- n data points with p dimensions
- Two modes

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

## ■ Dissimilarity matrix

- n data points, but registers only the distance
- A triangular matrix
- Single mode

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

# 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:

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

- 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

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

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

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


# 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

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

- Example

Object ID	test-1
1	code-A
2	code-B
3	code-C
4	code-A


$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

# Proximity Measure for Nominal Attributes

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

Object ID	code-A	Code-B	Code-C
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0

# Distance on Numeric Data: Minkowski Distance

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

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

where  $i = (x_{i1}, x_{i2}, \dots, x_{ip})$  and  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  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  $i \neq j$ , and  $d(i, i) = 0$  (Positive definiteness)
  - $d(i, j) = d(j, i)$  (Symmetry)
  - $d(i, j) \leq d(i, k) + d(k, j)$  (Triangle Inequality)
- A distance that satisfies these properties is a **metric**



# 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

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

- $h = 2$ : ( $L_2$  norm) **Euclidean** distance

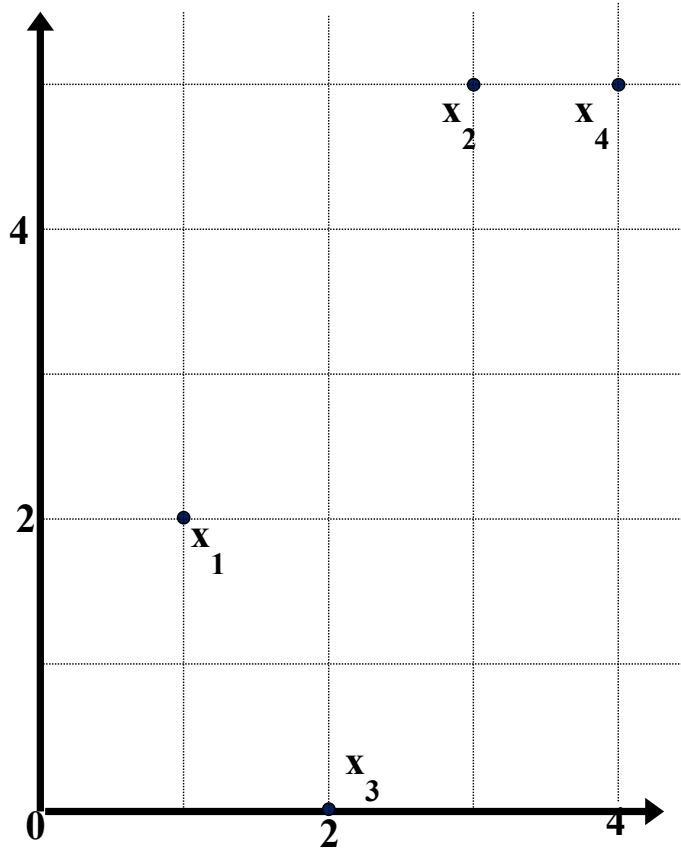
$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $h \rightarrow \infty$ . **“supremum”** ( $L_{\max}$  norm,  $L_{\infty}$  norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

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

# 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		
$x3$	5.1	5.1	0	
$x4$	4.24	1	5.39	0

# Example: Minkowski Distance

## Dissimilarity Matrices

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

### Manhattan ( $L_1$ )

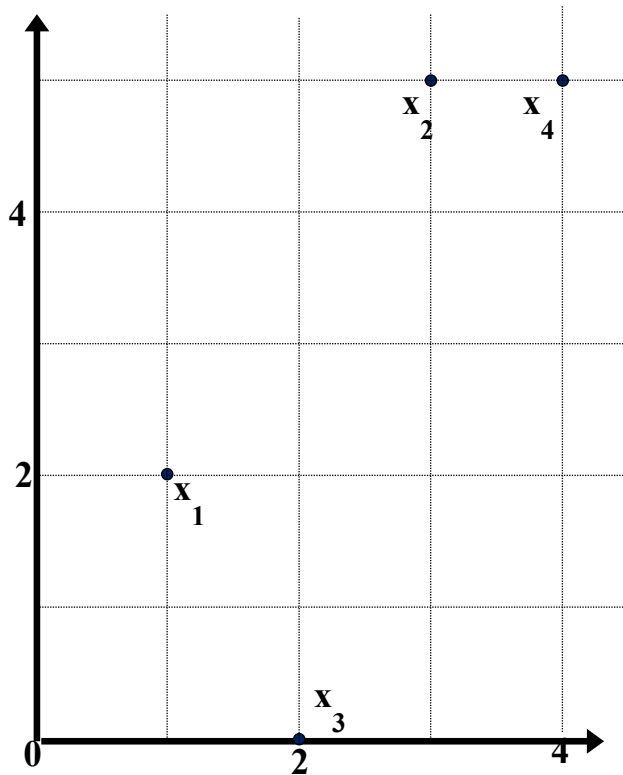
L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

### 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

$L_\infty$	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  $r_{if} \in \{1, \dots, M_f\}$
  - map the range of each variable onto  $[0, 1]$  by replacing  $i$ -th object in the  $f$ -th variable by

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

- compute the dissimilarity using methods for numeric variables

# Ordinal Variables

- Example

Object ID	test-2
1	excellent
2	fair
3	good
4	excellent



Object ID	test-2
1	3
2	1
3	2
4	3



$$\begin{bmatrix} 0 \\ 1 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

# Attributes of Mixed Type

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

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

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

# 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	teamcoach		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

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / (||d_1|| ||d_2||),$$

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

# Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\|$  ,  
where  $\cdot$  indicates vector dot product,  $\|d\|$ : the length of vector  $d$
- Ex: Find the **similarity** between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d_1 \cdot d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$$

$$\|d_1\| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$$

$$\cos(d_1, d_2) = 0.94$$



# 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