EE226 Big Data Mining Lecture 2

Data Mining Fundamentals

Liyao Xiang
http://xiangliyao.cn/
Shanghai Jiao Tong University

http://jhc.sjtu.edu.cn/public/courses/EE226/

- Please check https://oc.sjtu.edu.cn/login/ canvas for slides, announcement, assignment, grades, etc.

Reference and Acknowledgement

- Most of the slides are credited to Prof. Jiawei Han’s book “Data Mining: Concepts and Techniques.”

Outline

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

Outline

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

Data Objects

- Data sets are made up of data objects.
- A data object represents an entity. Also called samples, examples, instances, data points.
  - e.g., sales database: customers, store items, sales
  - e.g., medical database: patients, treatments
  - e.g., university database: students, professors, courses
- In a database, objects are stored as data tuples (rows). Attributes correspond to columns.

Attributes

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
- Values for a given attribute is called observations.
  - e.g., customer_ID, name, address
- Types:
  - Nominal (categorical, no meaningful order)
    - e.g., hair_color = {black, brown, blond, auburn, grey, white}
    - could use numeric values to represent
    - most commonly occurring value

Attributes

- Types:
  - Binary: a nominal attribute with 0 or 1 states (Boolean if the states are true or false)
    - e.g., smoker = {0: not smoke, 1: smokes} for patient
    - symmetric binary: both states are equally important
      - e.g., gender = {0: male, 1: female}
    - asymmetric binary: states are not equally important
      - e.g., HIV test result = {0: negative, 1: positive}
  - Ordinal: an attribute with values that have a meaningful order but magnitude between successive values is unknown
    - e.g., grades = {A+, A, A-, B+, …}

Attributes

- Nominal, binary, and ordinal attributes are qualitative. Their values are typically words (or codes) representing categories.
- Numeric attributes are quantitative: represented in integer or real values, including:
    - interval-scaled: measured on a scale of equal-size units. Allow to compare and quantify the difference between values.
        - e.g., temperature (no true zero, no ratios)
    - ratio-scaled: a numeric attribute with an inherent zero-point.
        - e.g., years_of_experience, number_of_words, weight, height, monetary quantities (you are 100 times richer with $100 than with $1)

Discrete vs Continuous Attributes

- Another way to organize attribute types
- Discrete attribute: has a finite or countably infinite set of values
  - e.g., hair_color, smoker, medical_test, binary attribute …
  - e.g., customer_ID (one-to-one correspondence with natural numbers)
- Continuous attribute: typically represented as floating-point variables.
  - often used interchangeably with numeric attribute

Summary

- Data objects
- Attributes
  - Nominal
    - Binary
    - Ordinal
    - Numeric
      - interval-scaled
      - ratio-scaled
  - Discrete
  - Continuous

Question

- In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling the problem.

Answer

- Ignoring the tuple: not effective unless the tuple contains several attributes with missing values
- Manually filling in the missing value: not reasonable when the value to be filled in is not easily determined
- Using a global constant to fill in the missing value: “unknown,” “-∞.” But may form an interesting concept
- Using the global attribute mean for quantitative values or global attribute mode for categorical values
- Using the class-wise attribute mean for quantitative values or class-wise attribute mode for categorical values
- Using the most probable value to fill in

Outline

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

Basic Statistical Descriptions of Data

- Measures of central tendency: measure the location of the middle or centre of a data distribution. Where do most of its values fall?
  - e.g., mean, median, mode, midrange
- Dispersion of data: How are the data spread out?
  - e.g., range, quartiles, interquartile range, five-number summary, boxplots, variance, standard deviation
- Graphic display
  - e.g., bar charts, pie charts, line graphs, quantile plots, quantile-quantile plots, histograms, scatter plots

Measuring the Central Tendency

- mean:

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

- weighted average:

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

- Problem: a small number of extreme values can corrupt the mean
- trimmed mean: the mean obtained after chopping off values at the high and low extremes.
  - e.g., sort the values observed for salary and remove the top and bottom 2% before computing the mean

Measuring the Central Tendency

- Median: the middle value in a set of ordered data values
  - a better measure of the centre of skewed (asymmetric) data
  - e.g., N values of ordinal data. If N is odd, median is the middle value. If N is even, median is the two middlemost values and any value in between.
  - expensive to compute if we have a large number of observations
  - approximation: assuming data are grouped in intervals and the frequency of each interval is known. Compute median frequency and let the interval contains median frequency be median interval.

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

width of median interval

lower boundary of the median interval

number of values in dataset

frequency of median interval

sum of frequencies of all intervals lower than median interval

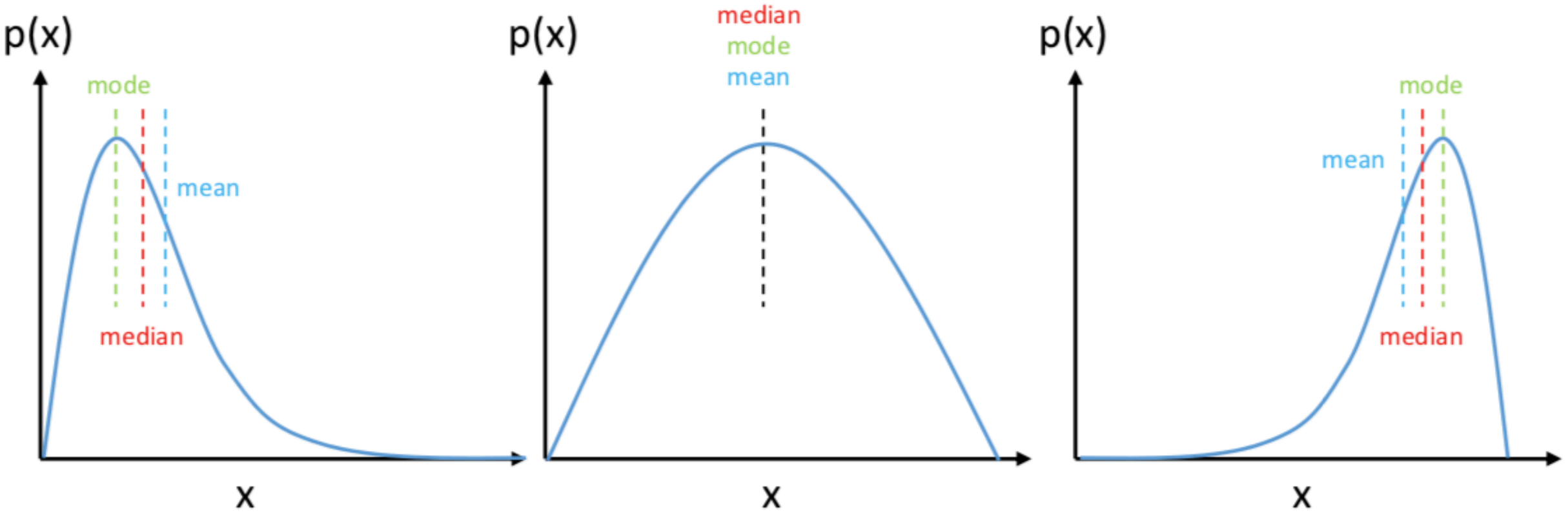
Measuring the Central Tendency

- Mode: the value that occurs most frequently in the set
  - can be determined for qualitative and quantitative attributes
  - e.g., unimodal, bimodal, trimodal, multimodal
  - no mode if data value occurs only once
  - approximate mode for unimodal data that are moderately skewed

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

- Midrange: the average of the largest and smallest values

Unimodal Frequency Curve



Positively skewed data

mode < median

Symmetric data

mode = median

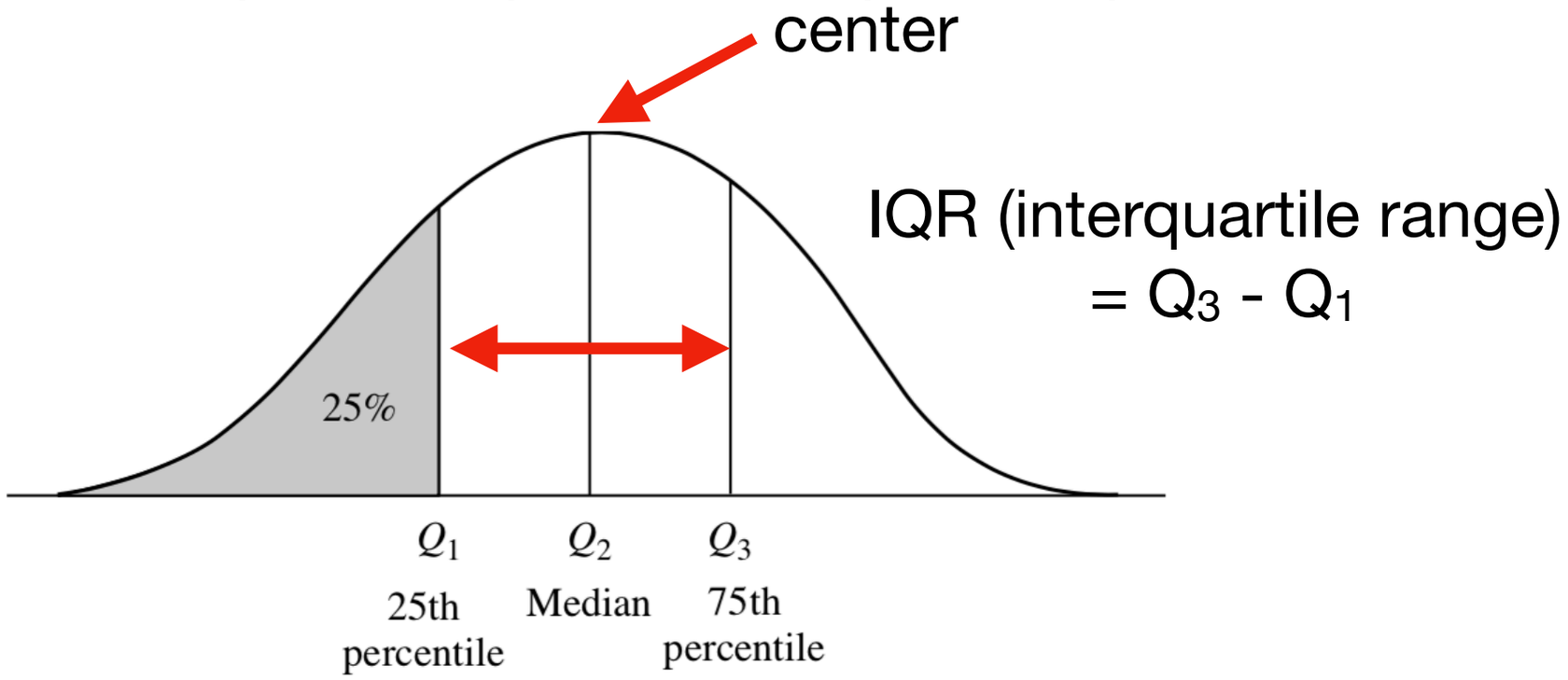
Negatively skewed data

mode > median

Slide credit: Weinan Zhang

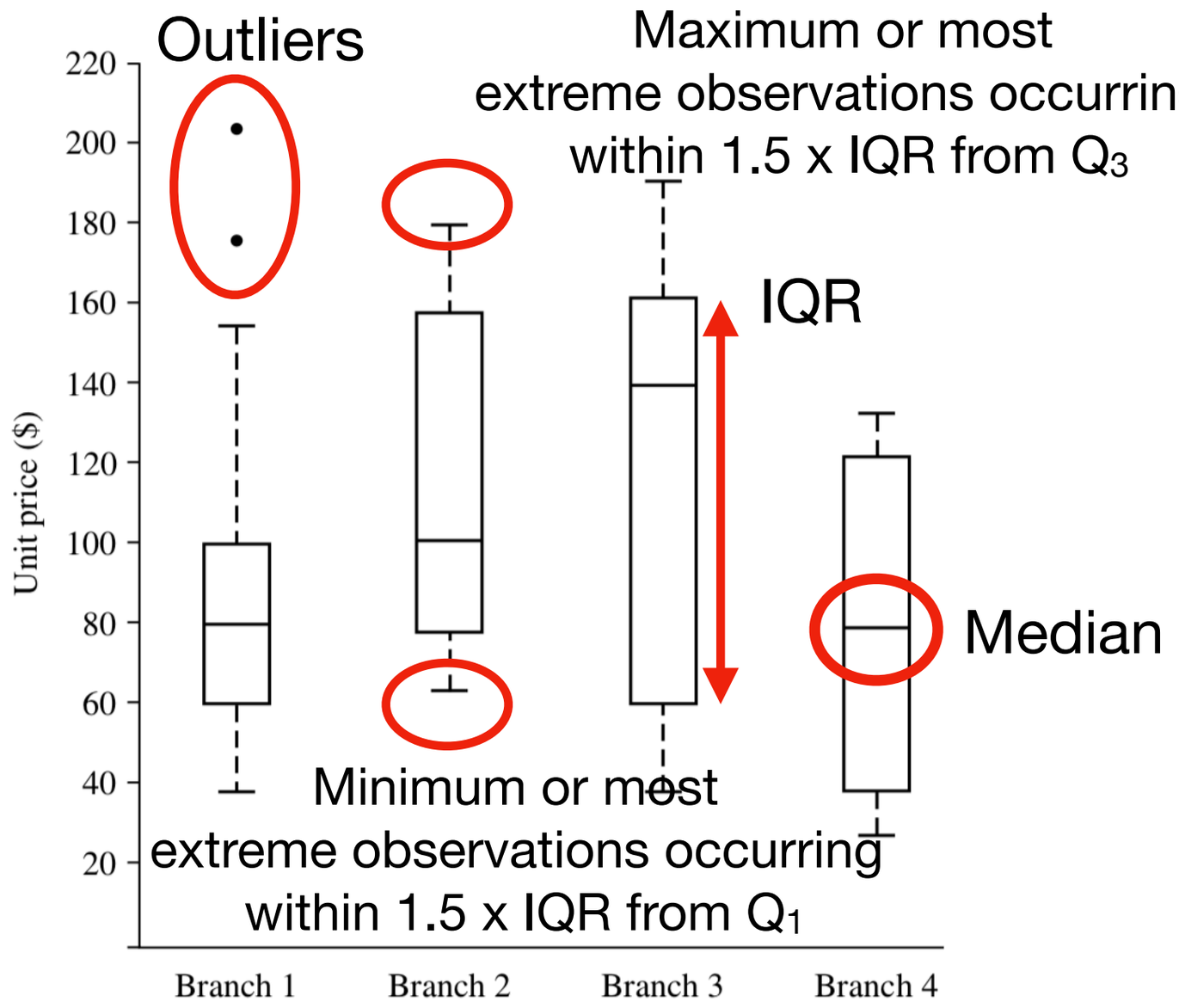
Measuring the Dispersion of Data

- Range: the difference between the largest and smallest values
- Quantile: the data points that split the data distribution into equal-size consecutive sets
  - e.g., kth q-quantile is the value x s.t. k/q of the data $< x$, and $(q - k)/q$ of the data are more than x.
- e.g., median = 2-quantile, quartile = 4-quantile, percentile = 100-quantile



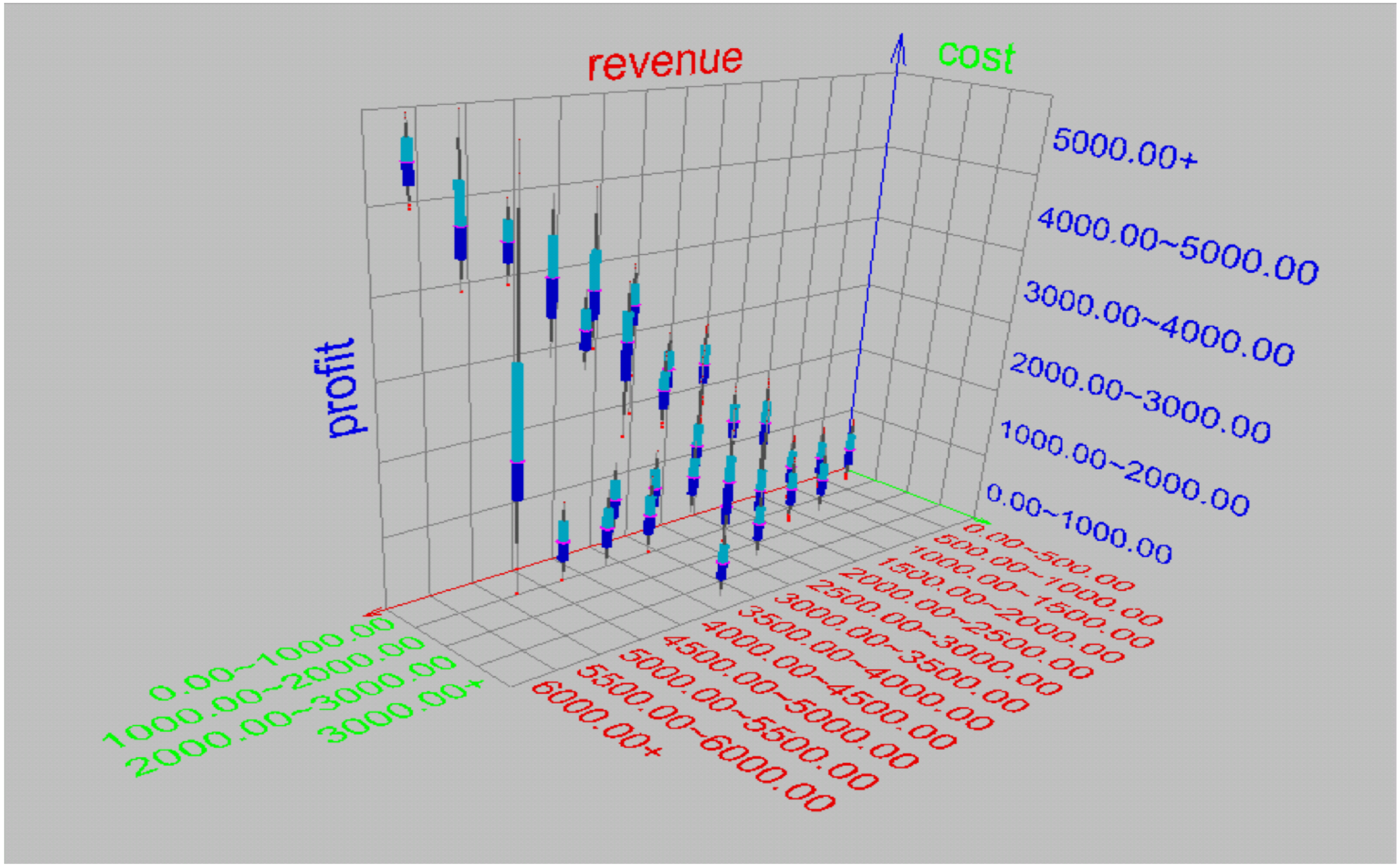
Measuring the Dispersion of Data

- Outliers: values falling at least 1.5 x IQR above Q_3 or below Q_1
- Five-Number Summary: Minimum, Q_1, Median, Q_3, Maximum
- Boxplots:



Measuring the Dispersion of Data

- An example of 3-D Boxplots:



Question

- What is the time complexity for computing boxplots? How about approximating the boxplots?

Answer

- $O(n \log n)$. Sorting algorithm. Approximation takes linear or sublinear time.

Measuring the Dispersion of Data

- Variance and Standard Deviation (STD)
  - low STD indicates observations are close to the mean, otherwise the data are spread out over a large range
  - variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2$
  - STD: σ
  - At least $\left((1 - \frac{1}{k^2}) \times 100\right)\%$ of the data are within $k\sigma$ from the mean

Why?

By Chebyshev's inequality:

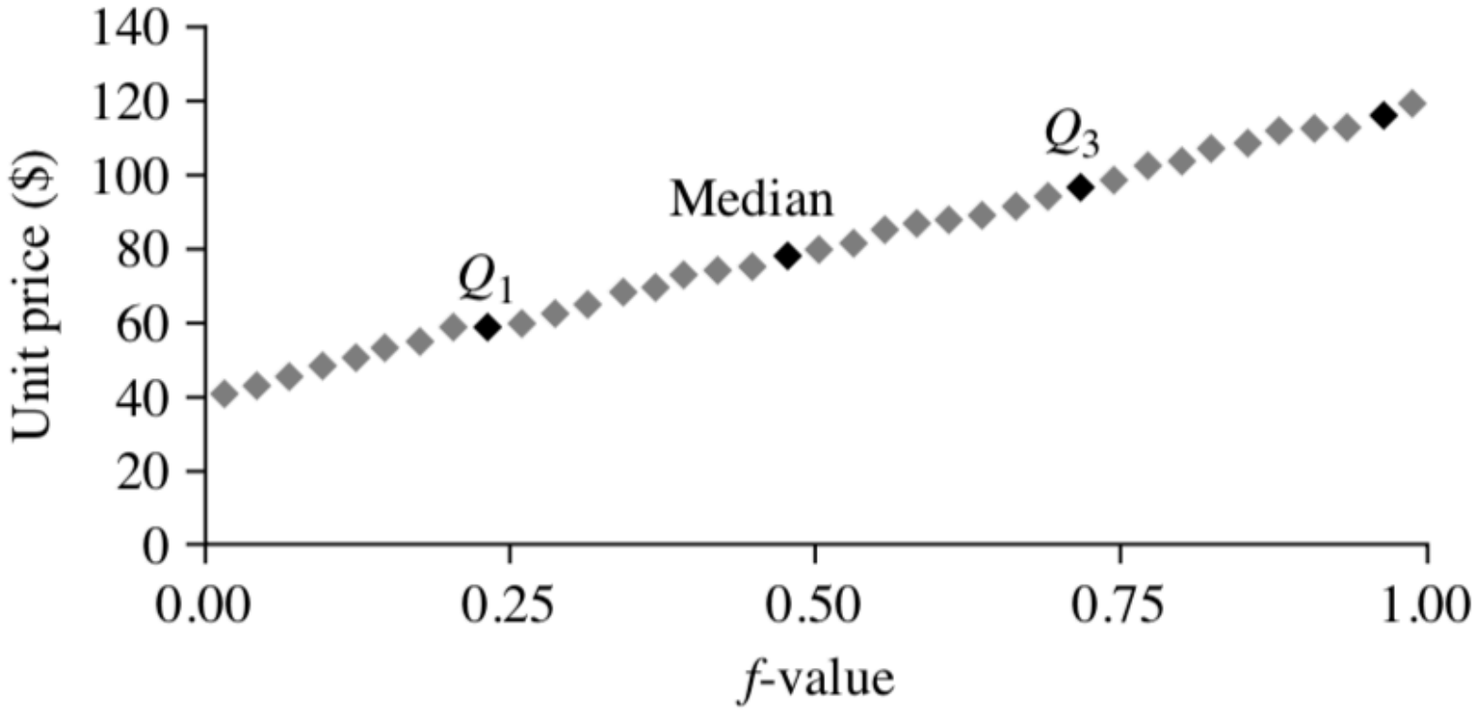
$$\Pr(|x - \bar{x}| \geq k\sigma) \leq \frac{1}{k^2}$$

Graphic Displays of Basic Statistical Descriptions (univariate distributions)

- Quantile plot: Each value x_i is paired with f_i indicating approximately $(100\, f_i)\%$ of the data are $\leq x_i$
  - Sort data in increasing order.
  - Compute $f_i = (i - 0.5) / N$

A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

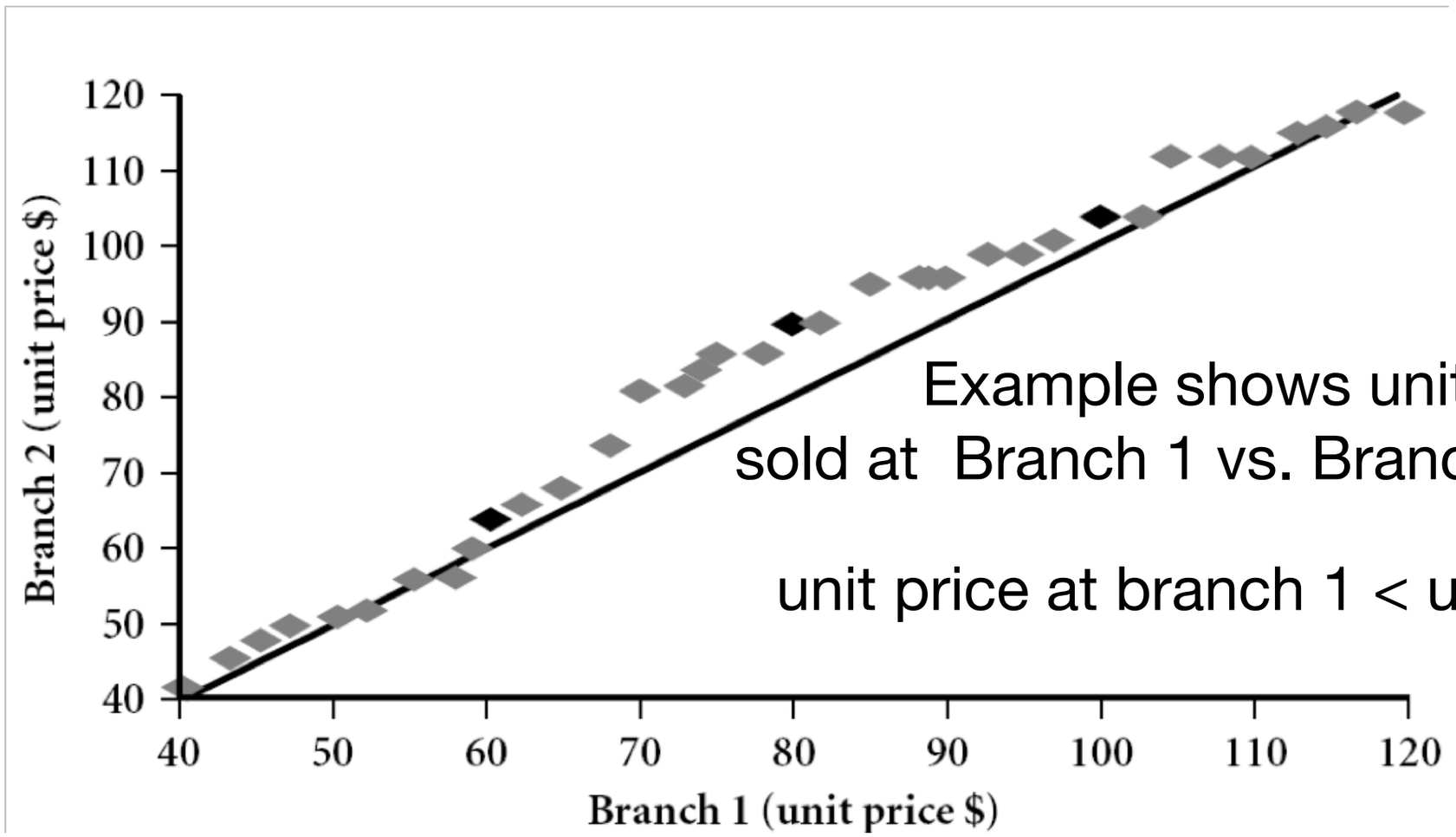
Unit price ($)	*Count of items sold*
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350



Graphic Displays of Basic Statistical Descriptions (univariate distributions)

- Quantile-Quantile Plot: graphs the quantiles of one univariate distribution against the corresponding quantiles of another

- Is there 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.

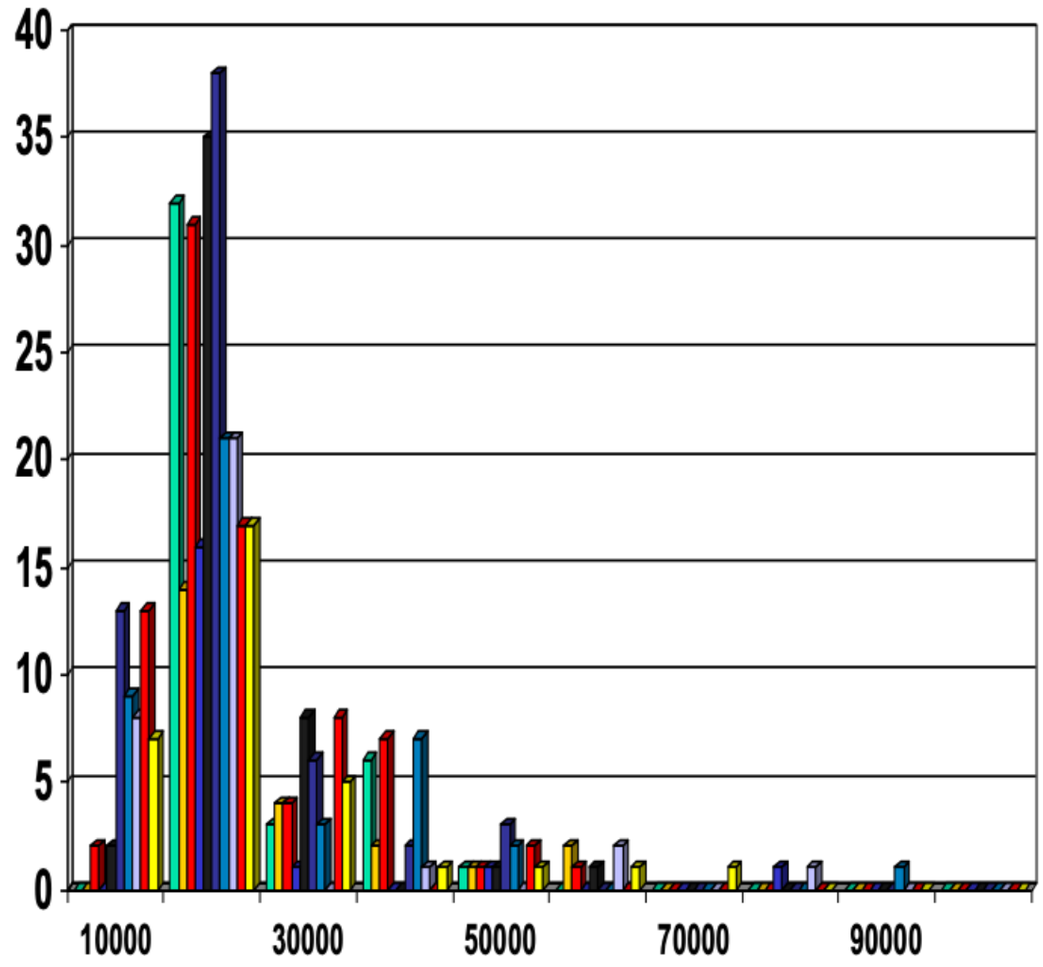
unit price at branch 1 < unit price at branch 2

Graphic Displays of Basic Statistical Descriptions (univariate distributions)

- Histograms: a chart of bars of which the height indicates frequency

pole chart

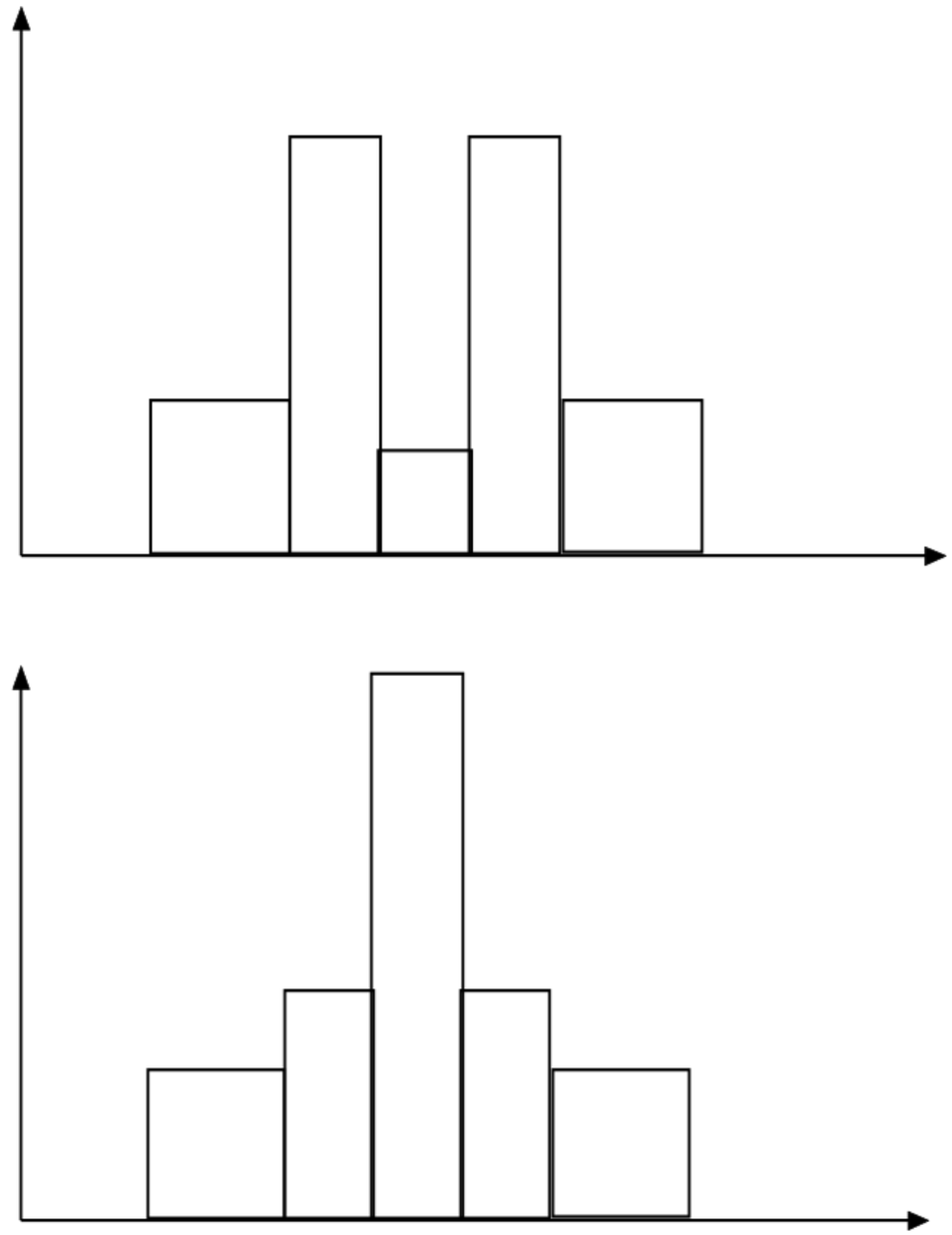
  - The range of values is partitioned into disjoint consecutive subranges (buckets or bins).
  - The range of a bucket is known as the width.
  - The bar height represents the total count of items within the subrange.



Histograms often Tells More than Boxplots

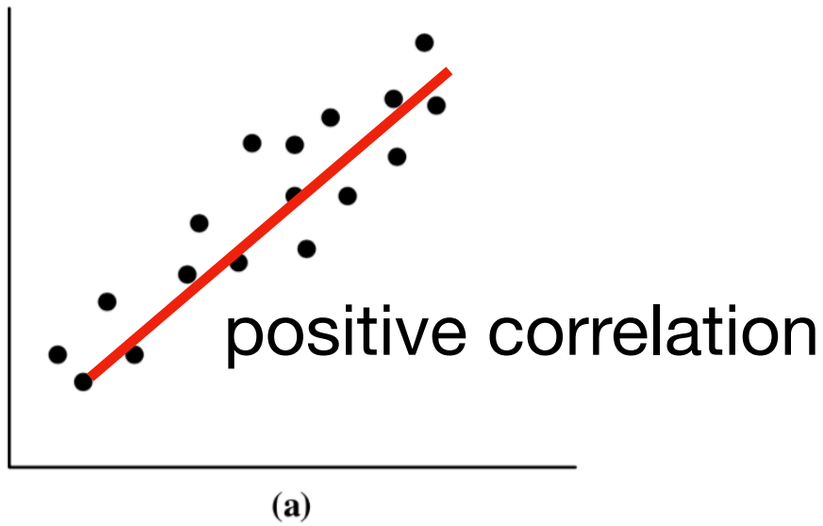
- The two histograms may have the same boxplot representation: min, Q_1, median, Q_3, max

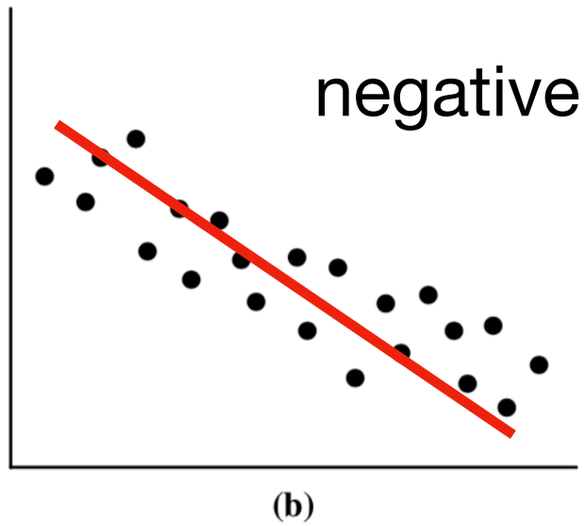
- But they have rather different distributions



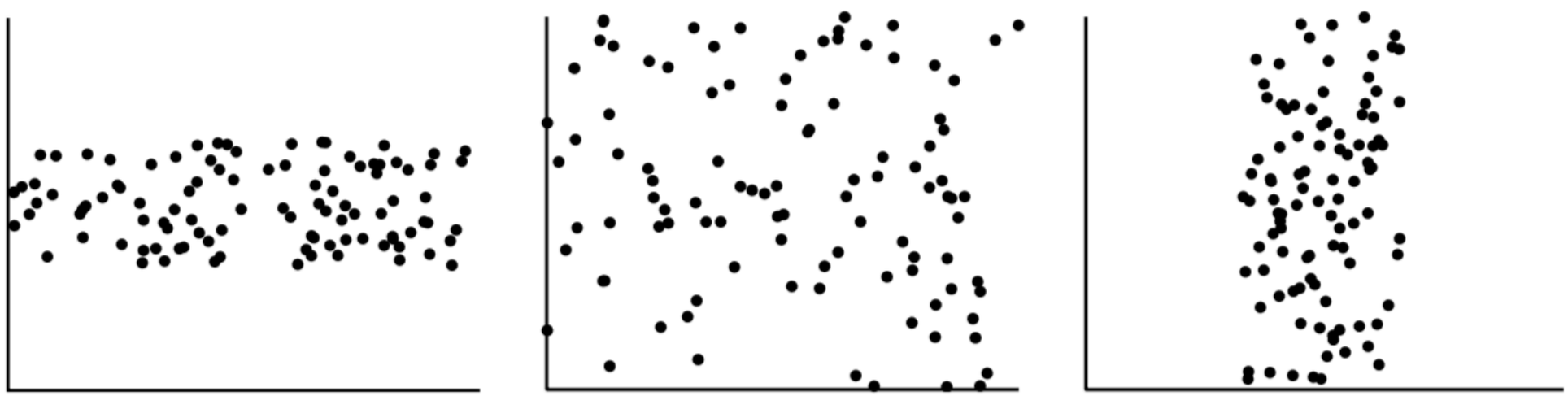
Graphic Displays of Basic Statistical Descriptions (bivariate distributions)

- Scatter plot: provides a first look at bivariate data to see clusters of points, outliers, or the correlation relationships.
  - X and Y are correlated if one implies the other.





uncorrelated



Summary

- Basic Statistical Descriptions of Data
  - Measures of central tendency: mean, median, mode, midrange
  - Measures of dispersion: range, quartiles, interquartile range, five-number summary, boxplots, variance, standard deviation
  - Graphic display: quantile plot, quantile-quantile plot, histogram, scatter plot

Question

- Give three statistical measures not illustrated yet for the characterization of data dispersion, and discuss how they can be computed efficiently in large databases.

Answer

- mean deviation = $\frac{\sum_{i=1}^{n} |x - \bar{x}|}{n}$ (absolute deviations from means)
- measure of skewness = $\frac{\bar{x} - \text{mode}}{\text{STD}}$ (how far, in STD, the mean is from the mode)
- coefficient of variation = $\frac{\text{STD}}{\bar{x}} \times 100$ (STD expressed as a percentage of the mean)

Can be efficiently calculated by partitioning the database, computing the values for each partition, and then merging these values into an equation to calculate the value for the entire database.

Outline

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

Data Visualization

- Why visualization?
  - Gain insight by mapping data onto graphical primitives
  - Provide qualitative overview of large datasets
  - Search for patterns, trends, structures, irregularities, relationships
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of knowledge representations
- Categories of visualization techniques:
  - Pixel-oriented
  - Geometric projection
  - Icon-based
  - Hierarchical
  - Visualizing non-numeric data

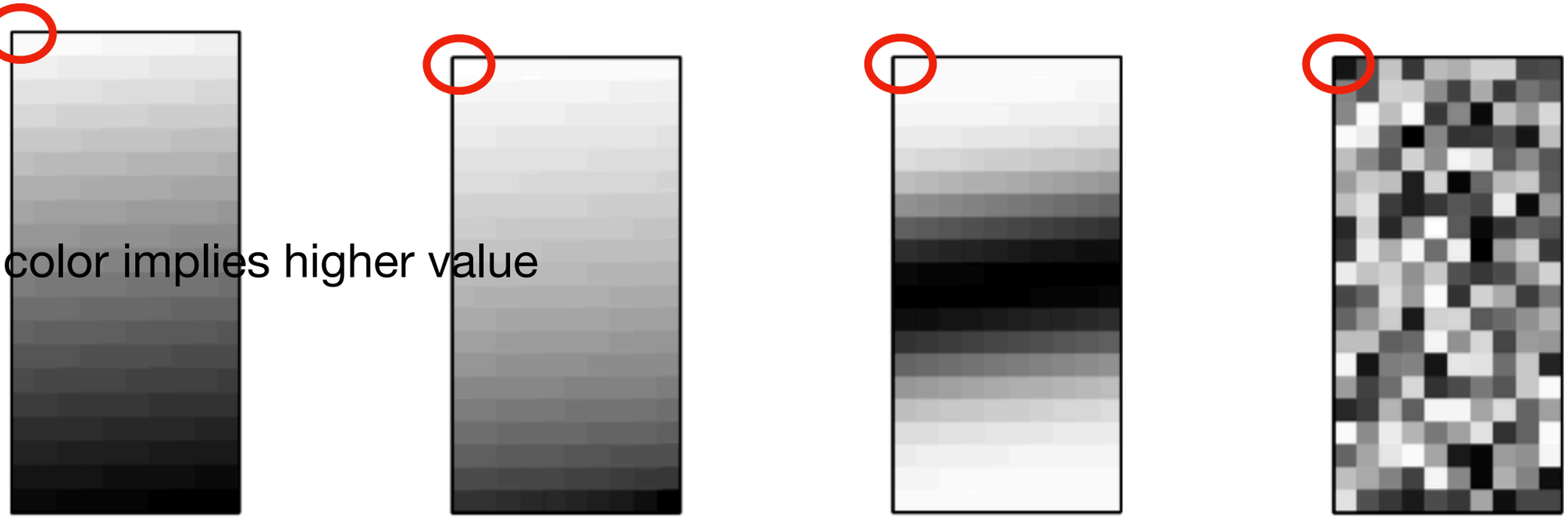
Pixel-Oriented Visualization

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

customer 1 by the income order

m dimension values for customer 1

darker color implies higher value

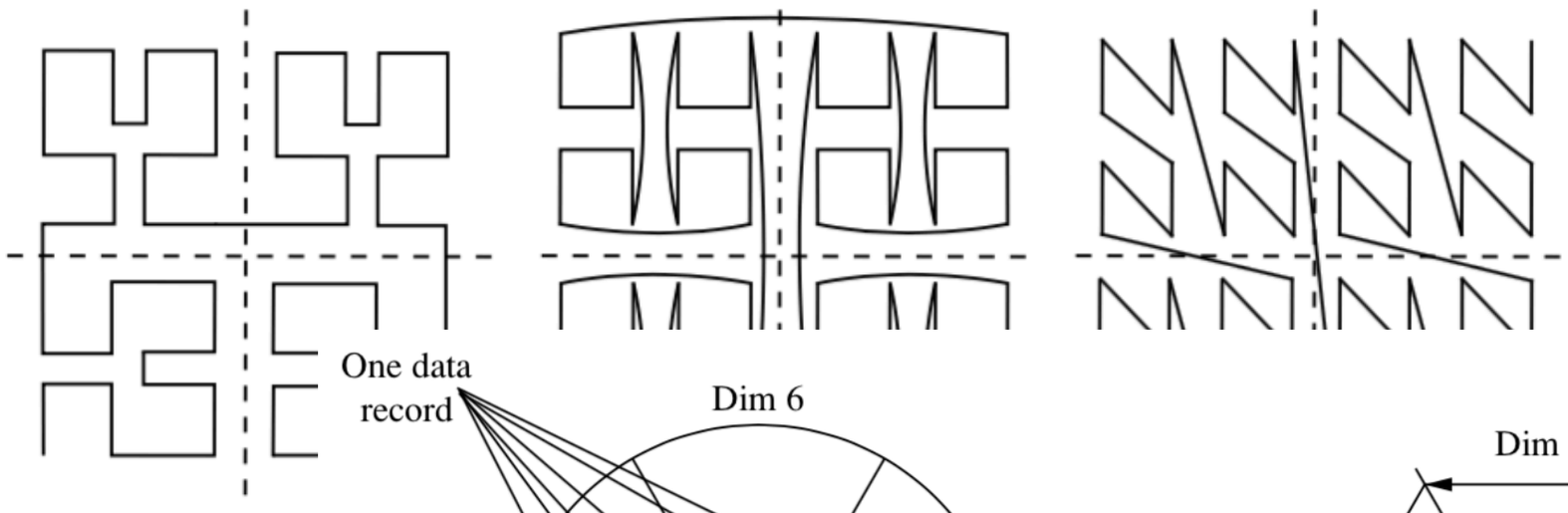


(a) *income* (b) *credit_limit* (c) *transaction_volume* (d) *age*

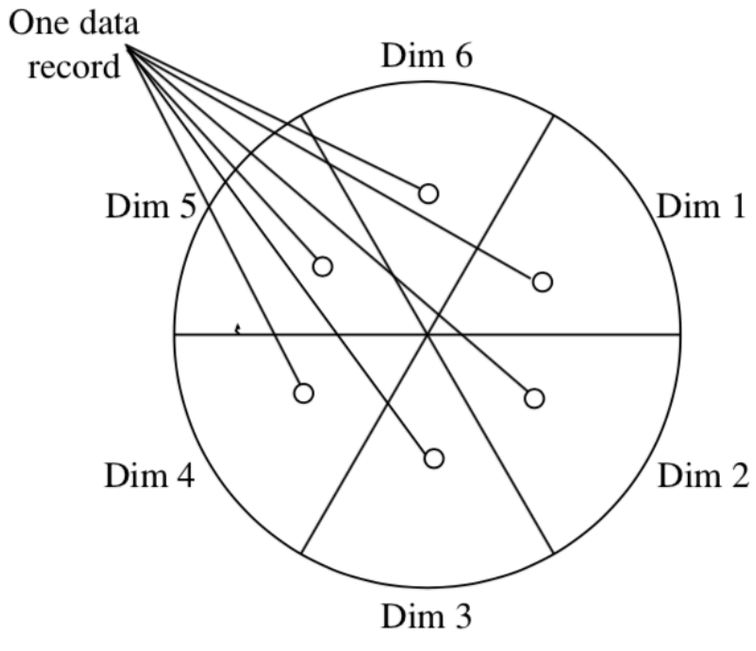
credit_limit increases with the income; customers with medium income are more likely to purchase; no clear correlation between income and age

Pixel-Oriented Visualization

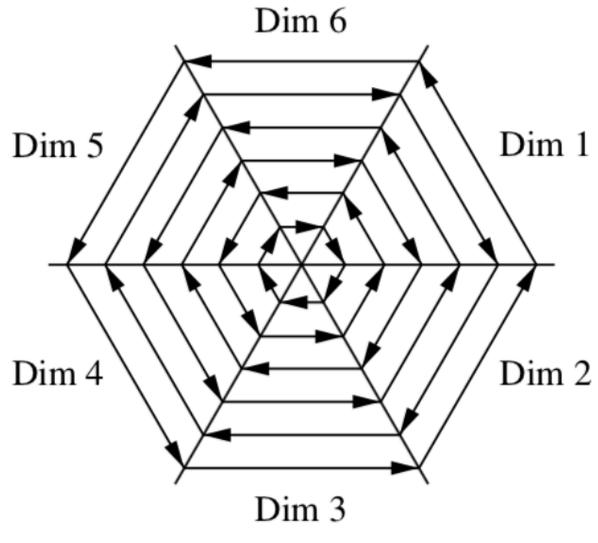
- Problems: pixels separated far apart should be next to each other in the global order, or the other around
- Solutions: 1. arrange data along the space-filling curves; 2. circle segment



(a) Hilbert c



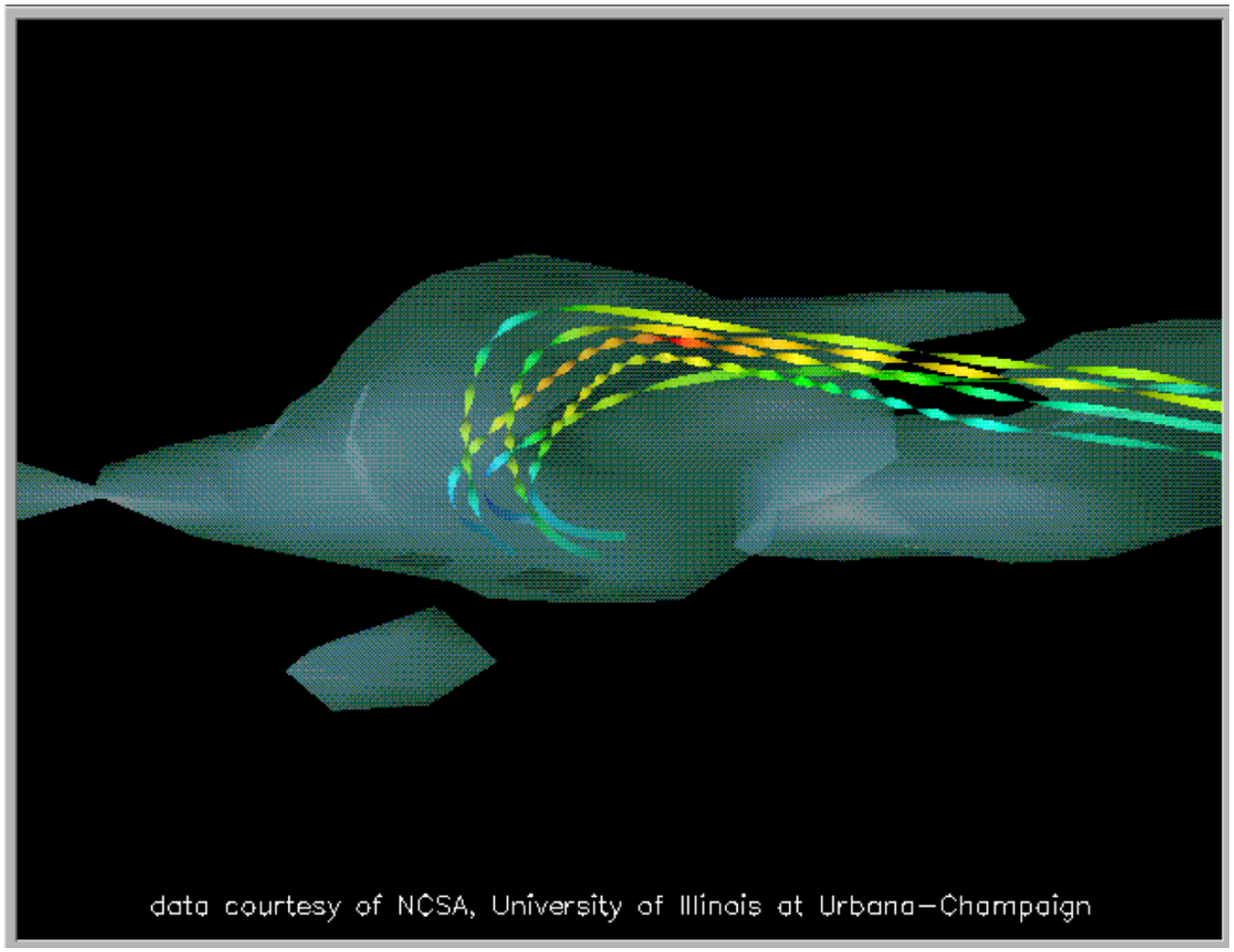
(a)



(b)

Geometric Projection Visualization

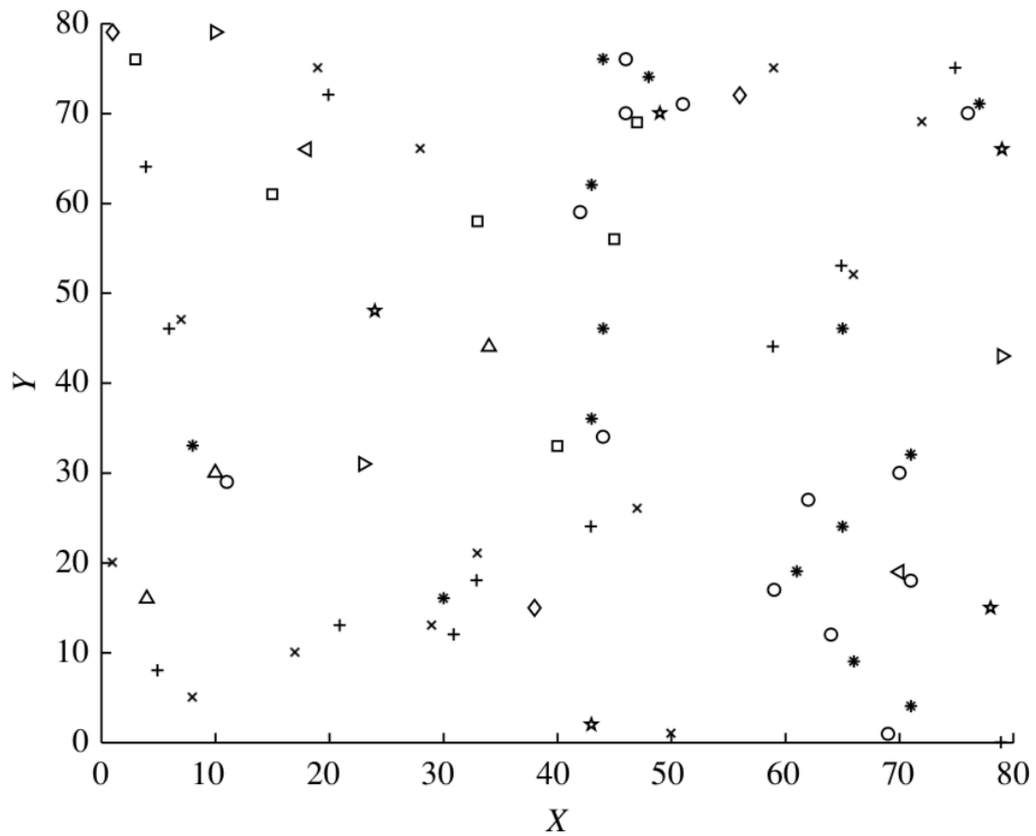
- How to visualize a high-dimensional space on a 2-D display?
- Methods: direct data visualization, scatter plot, scatter-plot matrices, landscapes, parallel coordinates
- Direct data visualization



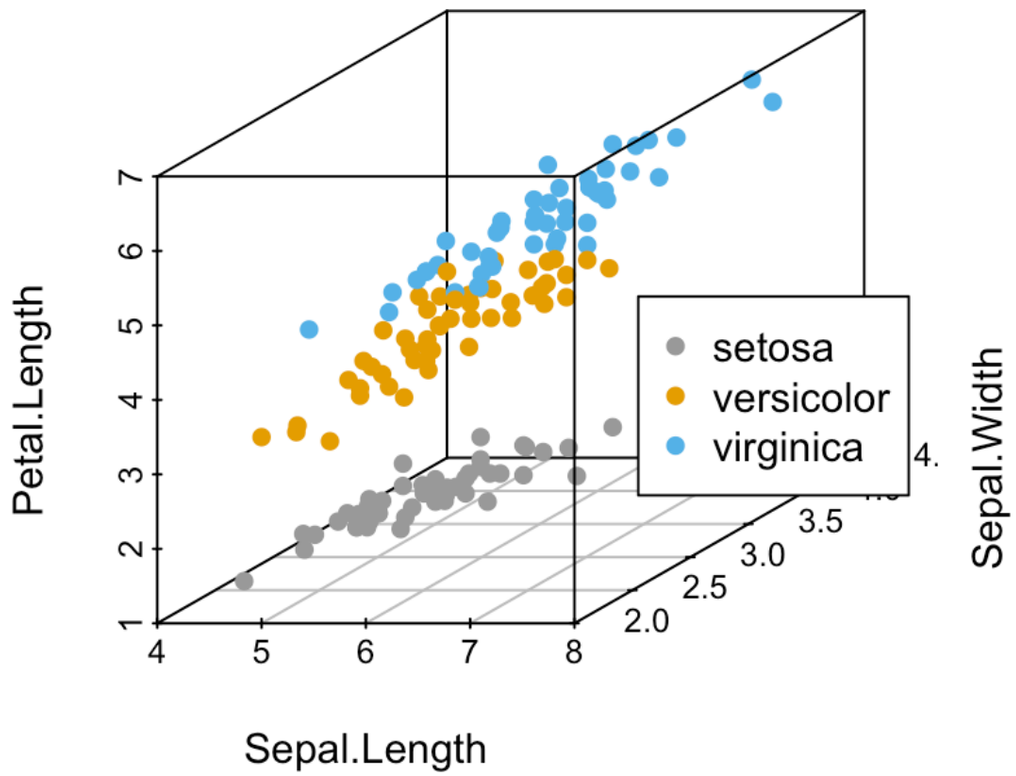
Ribbons with Twists
Based on Vorticity

Geometric Projection Visualization

- Scatter plot



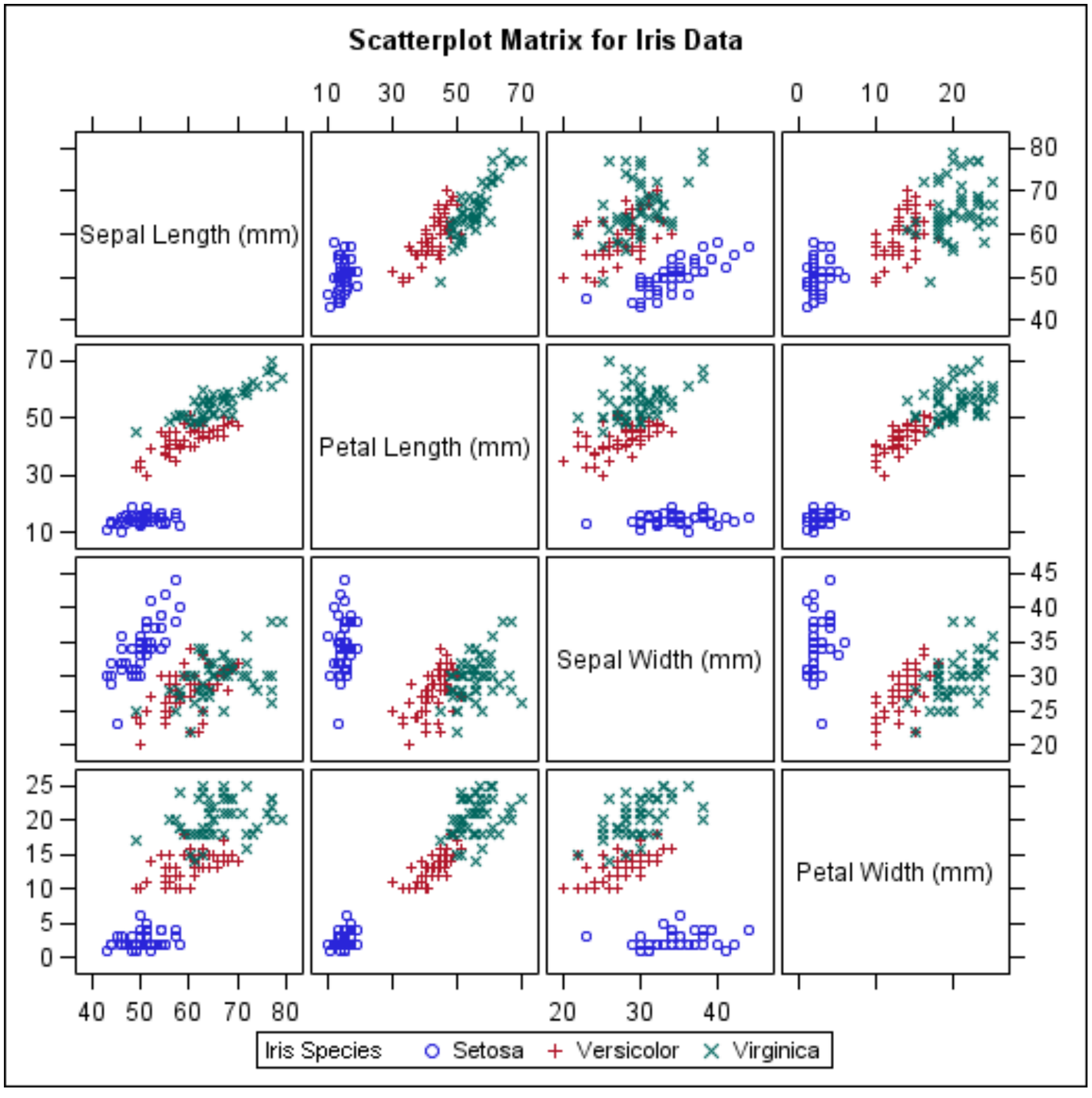
- 3-dimensional data (X and Y are two attributes, the 3rd one is represented by different shapes)



- 4-dimensional data

Geometric Projection Visualization

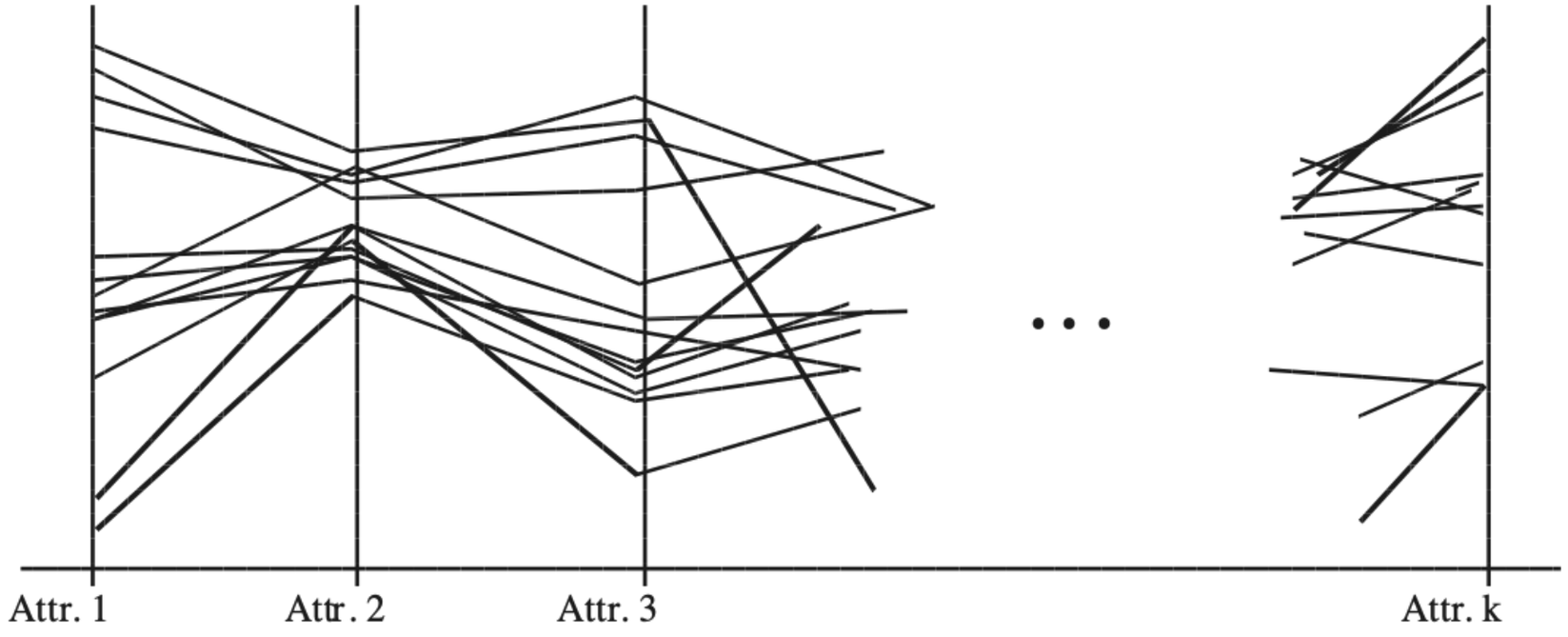
- Scatter-plot Matrices



For an n-dimensional dataset, a scatter-plot matrix is an n^2 grid of 2-D scatter plots

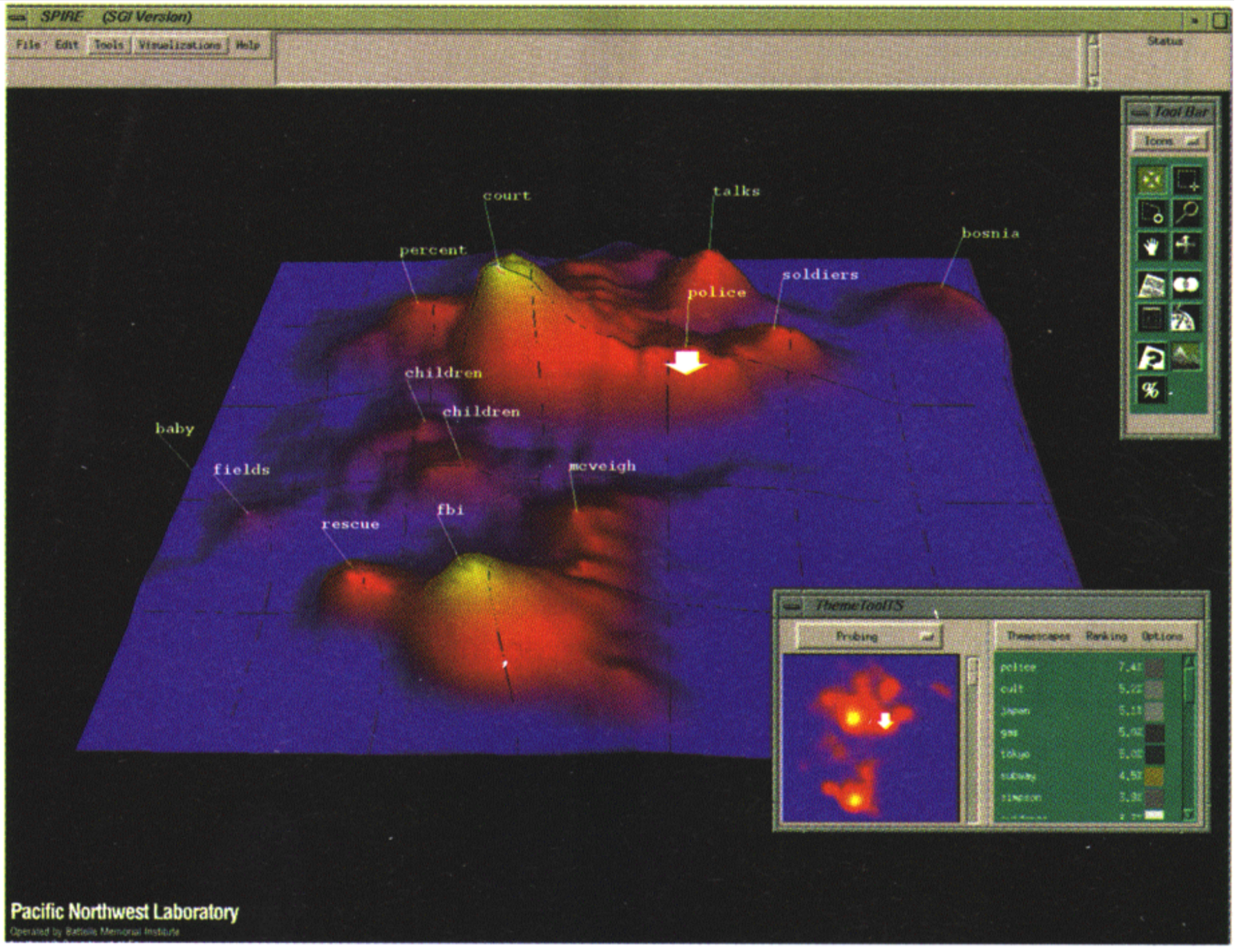
Geometric Projection Visualization

- Parallel coordinates: can handle higher dimensionality
  - k equally spaced axes, one for each dimension
  - A data record is represented by a polygonal line that intersects each axis at the corresponding value



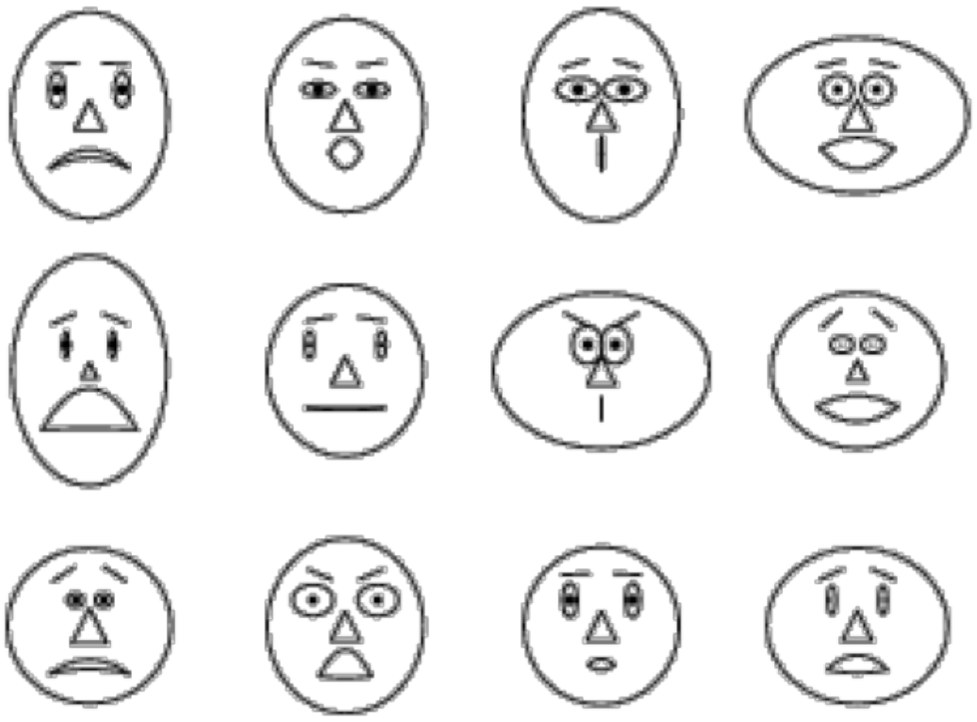
Geometric Projection Visualization

- Landscapes:

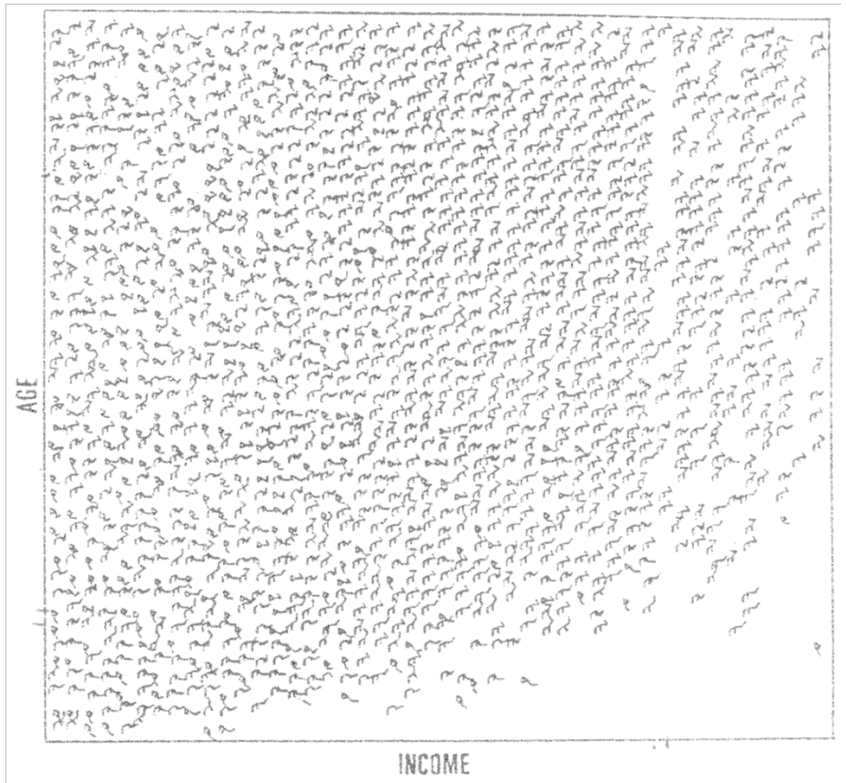


Icon-Based Visualization

- Visualization of data values using features of icons
- Visualization techniques:
  - Chernoff faces: dimensions are mapped to facial characteristics; viewing many facial characteristics at once



  - Stick figures: map dimensions to the angle and/or length of the limbs; texture patterns -> data trends

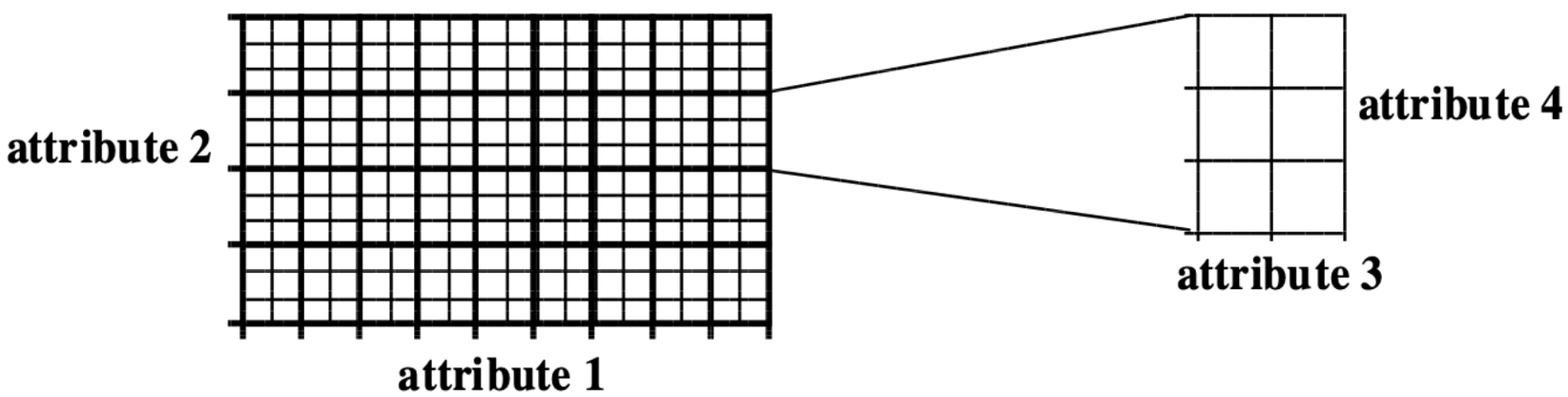


Hierarchical Visualization

- Visualization of data using a hierarchical partitioning into subspaces
- Techniques:
  - Dimensional Stacking
  - Worlds-within-Worlds
  - Tree-Map

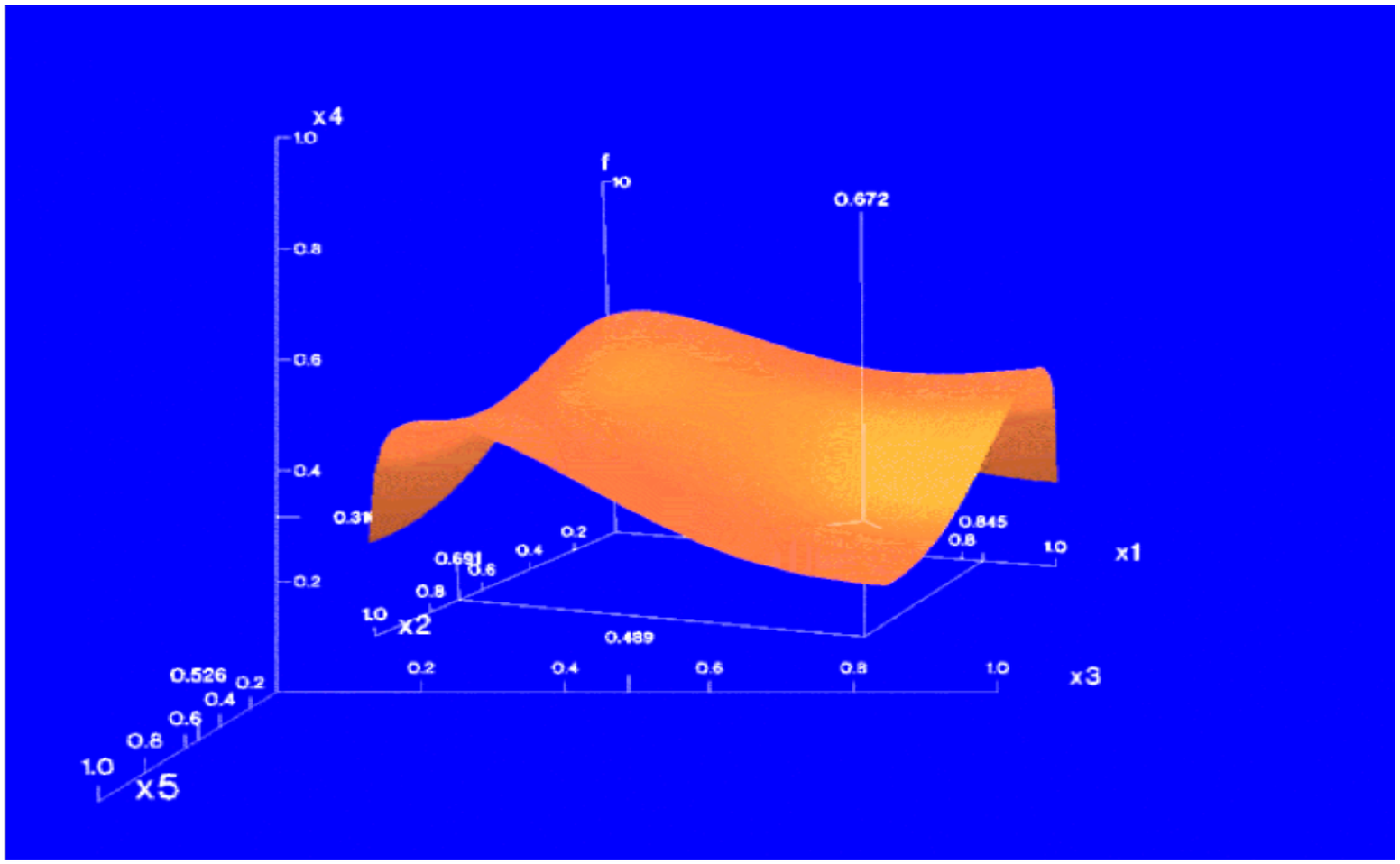
Hierarchical Visualization

- Dimensional Stacking: Partitioning of the n-dimensional attribute space in 2-D subspaces, which are ‘stacked’ into each other.
  - Discretizing the ranges of each dimension
  - Each dimension is assigned an ordering
  - 2 lowest ordering are used to divide a virtual screen into sections, with cardinality to determine the number of sections
- Each section is used to define the virtual screen for the next 2 dimensions



Hierarchical Visualization

- Worlds-within-Worlds:
  - Suppose we want to visualize a 6-D dataset with dimensions f, x_1, x_2… x_5. We first fix the values of dimensions x_3, x_4, x_5 to be c_3, c_4, c_5.
  - Visualize f, x_1, x_2 using 3-D plot, called ‘inner world,’ of which the origin is at (c_3, c_4, c_5) in the outer world
  - A user can change the origin of the inner world in the outer world and view the resulting changes.



Hierarchical Visualization

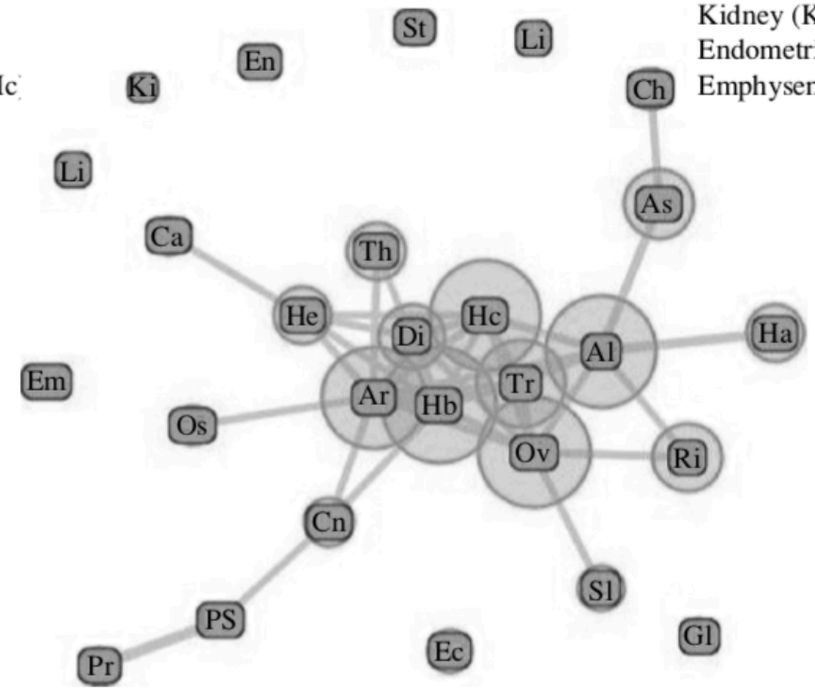
- Tree-maps:
  - e.g., disk space usage.

Examples from: https://www.jam-software.com/treesize_free/tree_map.shtml

Visualizing Non-Numeric Data

- Tag cloud:
  - visualize text or social network data
  - importance of a tag is indicated by font size or colour (node size or link width)

High blood pressure (Hb)
Allergies (Al)
Overweight (Ov)
High cholesterol level (Hc)
Arthritis (Ar)
Trouble seeing (Tr)
Risk of diabetes (Ri)
Asthma (As)
Diabetes (Di)
Hayfever (Ha)
Thyroid problem (Th)
Heart disease (He)
Cancer (Cn)
Sleep disorder (Sl)
Eczema (Ec)
Chronic bronchitis (Ch)
Osteoporosis (Os)
Prostate (Pr)
Cardiovascular (Ca)
Glaucoma (Gl)
Stroke (St)
Liver condition (Li)
PSA test abnormal (PS)
Kidney (Ki)
Endometriosis (En)
Emphysema (Em)



Summary

- Categories of visualization techniques:
  - Pixel-oriented
  - Geometric projection: direct data visualization, scatter plot, scatter-plot matrices, landscapes, parallel coordinates
  - Icon-based: Chernoff faces, stick figures
  - Hierarchical: Dimensional Stacking, Worlds-within-Worlds, Tree-Map
  - Visualizing non-numeric data

Outline

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

Measuring Data Similarity & Dissimilarity

- Measures of proximity
- Cases where we need to access how alike and unalike objects are in comparison with each other: clustering, outlier analysis, and nearest-neighbour classification
- Data structures: data matrix (store data objects) and dissimilarity matrix (store dissimilarity values)

- Data matrix

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

the value for object x_i of f^{th} attribute

Measuring Data Similarity & Dissimilarity

- Dissimilarity matrix

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

the measured dissimilarity between object 3 and 2

  - d(i, j) becomes larger when object i and j differ
  - Measures of similarity: sim(i, j) = 1 - d(i, j)

- How to compute proximity measures for different attributes?

Proximity Measures for Nominal Data

- d(i, j) = (p - m) / p, where m: number of matched attributes, p: total number of attributes
  - e.g., A data table containing mixed-type attributes

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

d(2, 1) = (1 - 0) / 1 = 1

d(4, 1) = (1 - 1) / 1 = 0

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Proximity Measures for Binary Data

- Contingency Table for Binary Attributes

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

- Symmetric binary dissimilarity: d(i, j) = (r + s) / (q + r + s + t)
- Asymmetric binary dissimilarity: d(i, j) = (r + s) / (q + r + s)

omitting negative matches

  - e.g., Relational Table of Symptoms

name	***gender***	***fever***	***cough***	***test-1***	***test-2***	***test-3***	***test-4***
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

d(Jack, Mary)
= 1 / (2 + 1) = 0.33

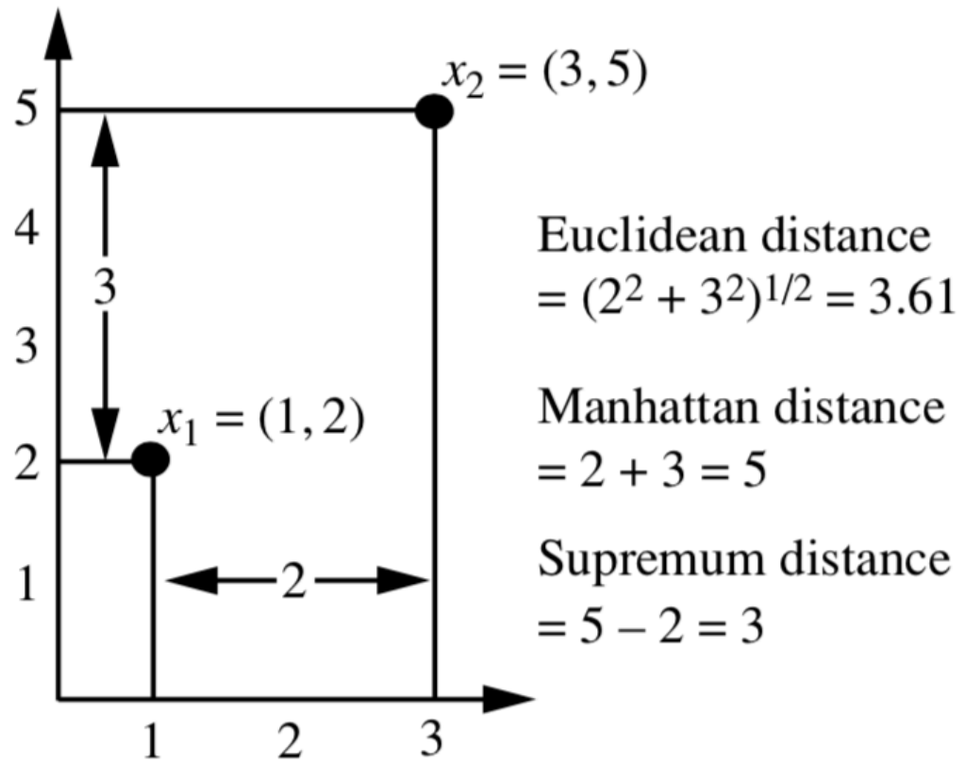
d(Jim, Mary)
= 3 / (1 + 1 + 2) = 0.75

Proximity Measure for Numeric Data

- L_p norm (Minkowski distance) between object i and j:

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

- L_1 norm = Manhattan (or city block) distance
- L_2 norm = Euclidean distance
- L_∞ norm (called infinity norm) = supremum (or Chebyshev) distance



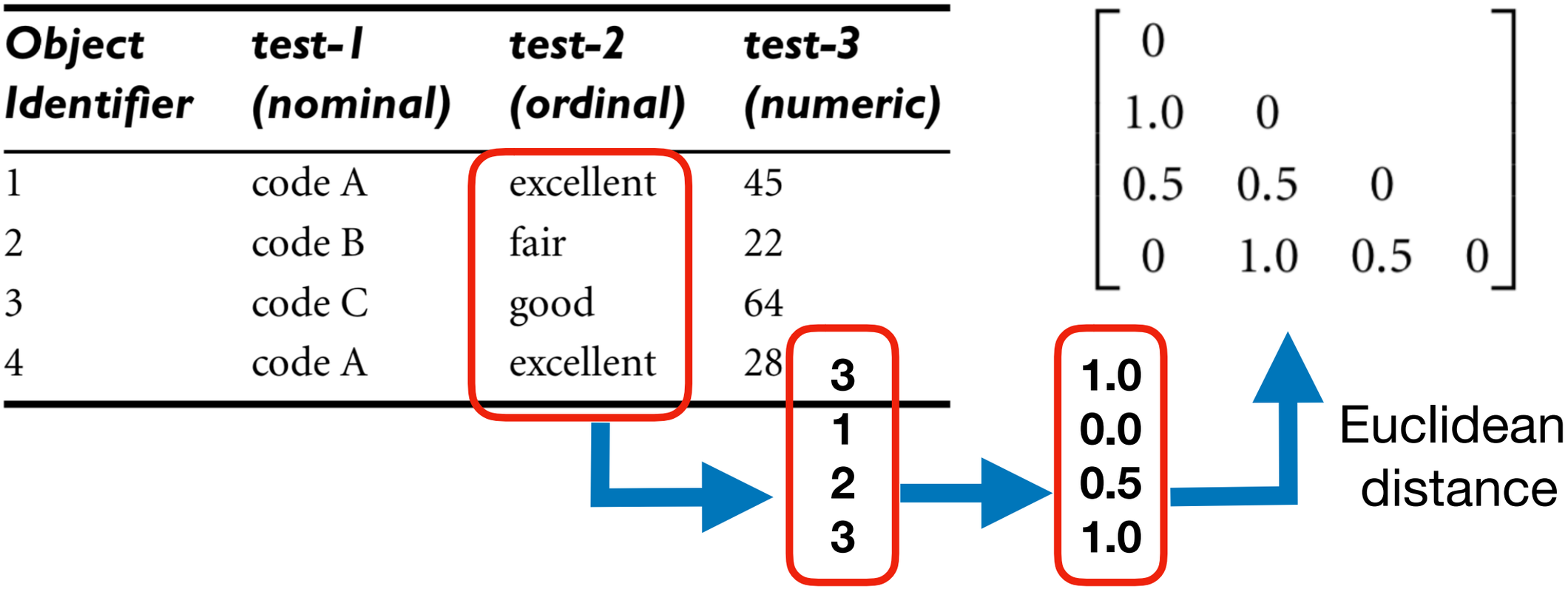
Proximity Measure for Ordinal Data

- Attribute f for object i has M_f ordered states: $1, \ldots, M_f$
- Ranking $r_{if} \in \{1, \ldots, M_f\}$. Map the ranking onto [0.0, 1.0] by

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

- Dissimilarity is computed by $\mathrm{L_p}$ norm
- e.g., A data table containing mixed-type attributes

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



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

Proximity Measure for Mixed-Typed Data

- Dissimilarity between object i and j:

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

$\delta_{ij}^{(f)} = 0$ if x_{if} or x_{jf} is missing or they are negative matches when f is asymmetric binary

$\delta_{ij}^{(f)} = 1$ otherwise

normalize numeric data by $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$

Proximity Measure for Mixed-Typed Data

- e.g., A data table containing mixed-type attributes

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

dissimilarity matrix of the data described by the three attributes of mixed types:

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

d(3, 1) = (1 + 0.5 + 0.45) / 3 = 0.65

Measuring Data Similarity & Dissimilarity

- Metric: a measure with the following properties
  - Non-negativity $d(i, j) \geq 0$
  - Identity of indiscernible $d(i, i) = 0$
  - Symmetry $d(i, j) = d(j, i)$
  - Triangle inequality $d(i, j) \leq d(i, k) + d(k, j)$

All proximity measures up to now are metrics.

- Non-metric measure: cosine similarity — cosine of the angle between vectors $\mathbf{x}$ and $\mathbf{y}$

$$\text{sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

Summary

- Measuring data similarity and dissimilarity
- data structures: data matrix, dissimilarity matrix
  - dissimilarity matrix for different types of attributes:
    - nominal data: d(i, j) = (p - m) / p
    - binary data (symmetric vs. asymmetric)
    - numeric data: L_p norm (Minkowski distance)
    - ordinal data: normalized ranking
    - mixed type
- non-metric measure

Overview of the Lecture

- Attribute Types: qualitative — nominal, binary, ordinal; quantitative — numeric; continuous, discrete
- Basic Statistical Descriptions of Data: measures of central tendency — mean, median, mode, midrange; dispersion — range, quartiles, interquartile range, five-number summary, boxplots, variance, standard deviation; graph display — quantile plot, quantile-quantile plot, histogram, scatter plot
- Data Visualization: pixel-oriented, geometric projection, icon-based, hierarchical
- Measuring Data Similarity and Dissimilarity: metric, non-metric measures