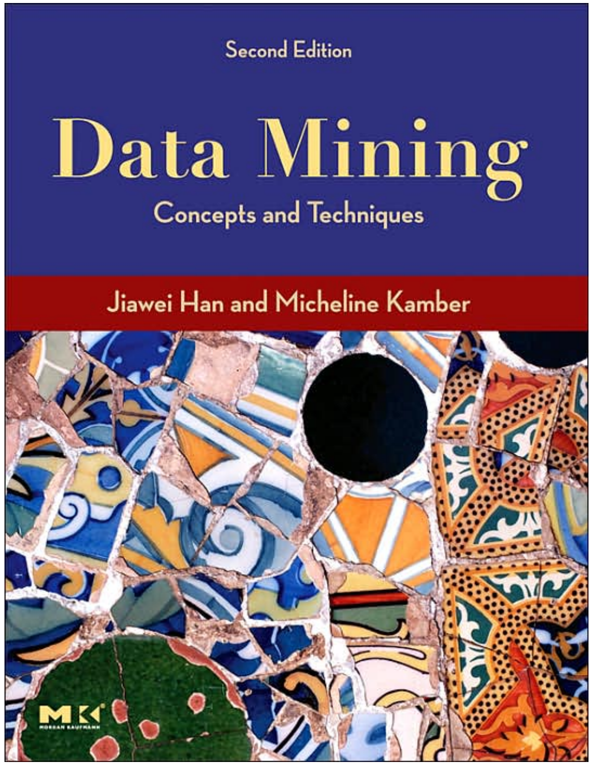
POLITECNICO DI MILANO

Data Exploration and Preprocessing

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

References

- Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems (Second Edition)



Outline

- Data Exploration
  - Descriptive statistics
  - Visualization
- Data Preprocessing
  - Aggregation
  - Sampling
  - Dimensionality Reduction
  - Feature creation
  - Discretization
  - Concept hierarchies

Data Exploration

What is data exploration?

- ❑ A preliminary exploration of the data to better understand its characteristics.

- ❑ Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
  - People can recognize patterns not captured by data analysis tools

- ❑ Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - A nice online introduction can be found in Chapter 1 of the NIST Engineering Statistics Handbook
  - http://www.itl.nist.gov/div898/handbook/index.htm

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory

- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set http://www.ics.uci.edu/~mlearn/MLRepository.html
- From the statistician Douglas Fisher
  - Three flower types (classes): Setosa, Virginica, Versicolour
  - Four (non-class) attributes, sepal width and length, petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Summary Statistics

- They are numbers that summarize properties of the data

- Summarized properties include frequency, location and spread

- Examples
  - Location, mean
  - Spread, standard deviation

- Most summary statistics can be calculated in a single pass through the data

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set

- For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time.

- The mode of a an attribute is the most frequent attribute value

- The notions of frequency and mode are typically used with categorical data

Percentiles

- For continuous data, the notion of a percentile is more useful

- Given an ordinal or continuous attribute x and a number p between 0 and 100, the pth percentile is a value x_p of x such that p% of the observed values of x are less than x_p

- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$

x_p

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points

- However, the mean is very sensitive to outliers

- Thus, the median or a trimmed mean is also commonly used

$$\mathrm{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\mathrm{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- □ Range is the difference between the max and min
- □ The variance or standard deviation is the most common measure of the spread of a set of points

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- □ However, this is also sensitive to outliers, so that other measures are often used.

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

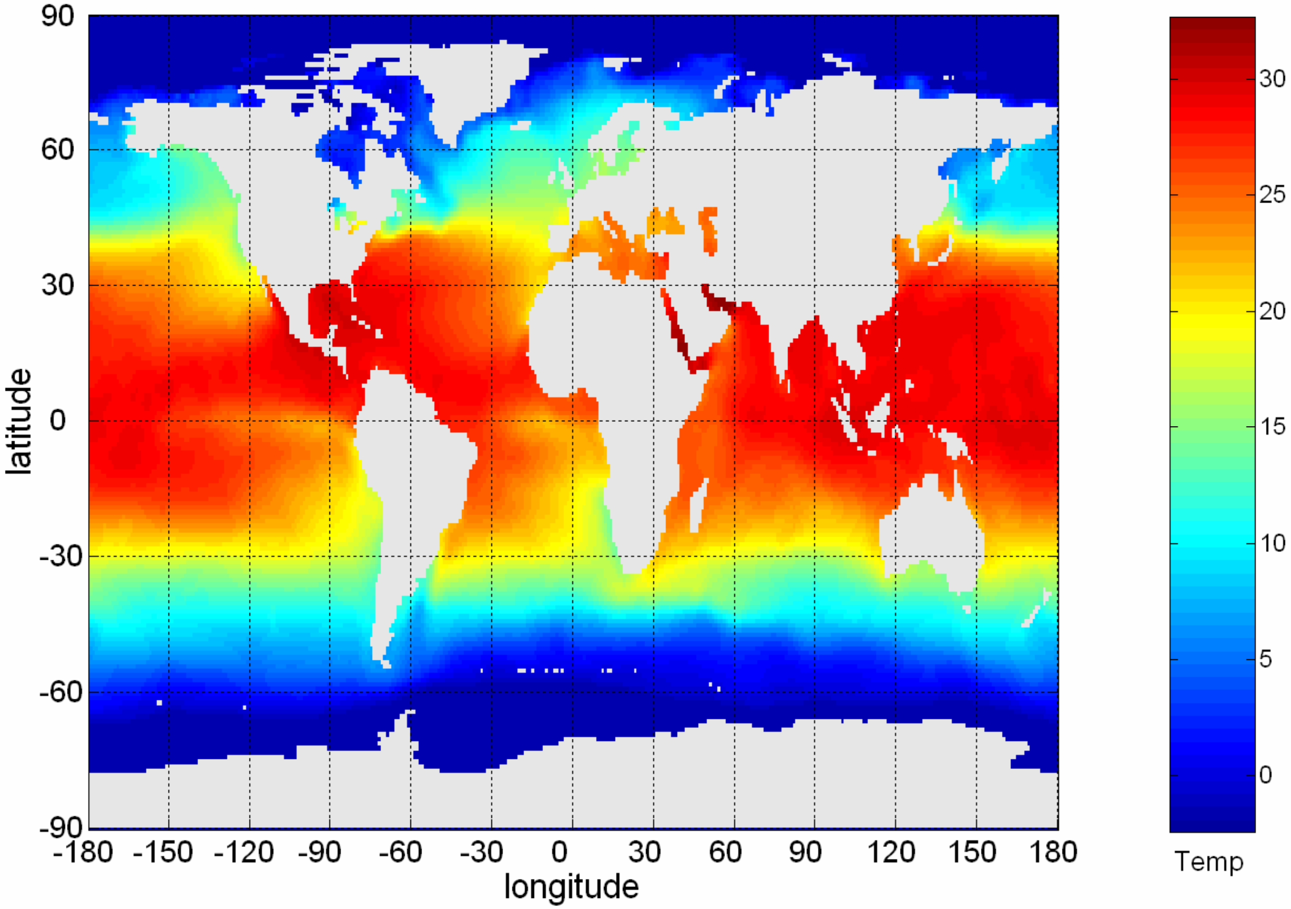
Visualization

- Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported

- Visualization of data is one of the most powerful and appealing techniques for data exploration
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature for July 1982
- Tens of thousands of data points are summarized in a single figure



Representation

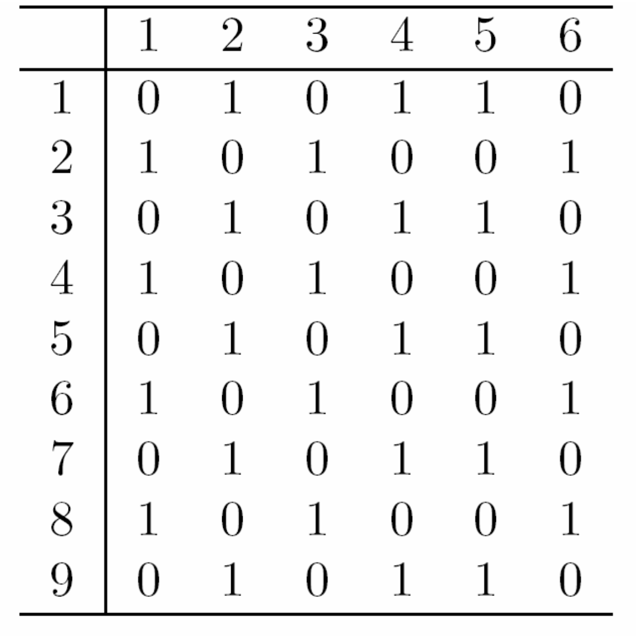
- Is the mapping of information to a visual format

- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors

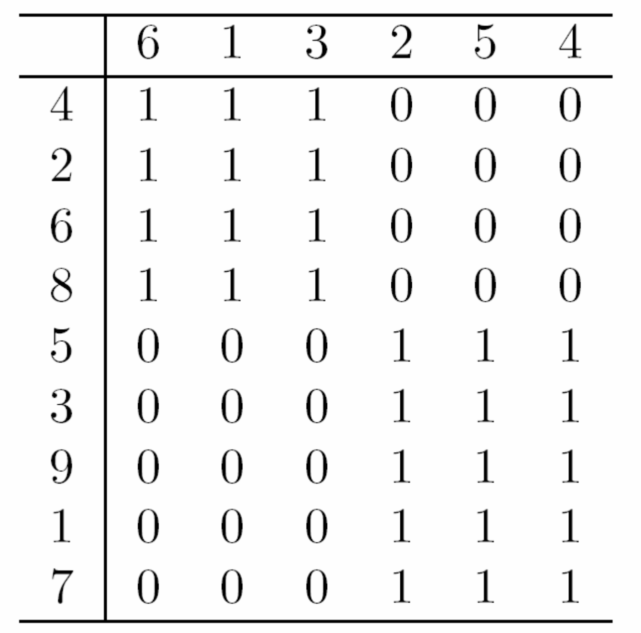
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0



	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

Selection

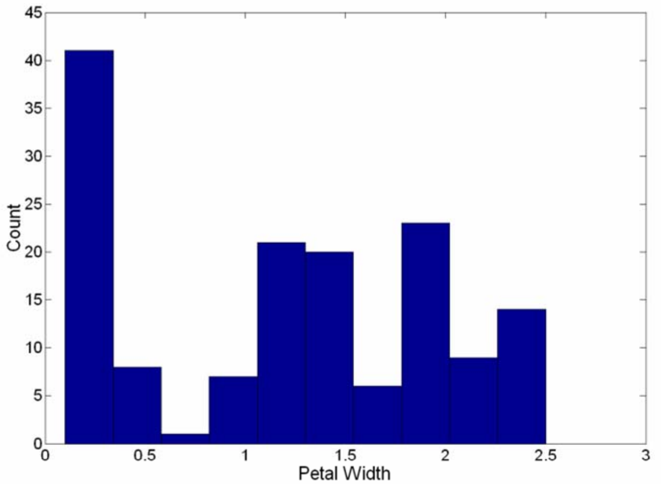
- Is the elimination or the de-emphasis of certain objects and attributes

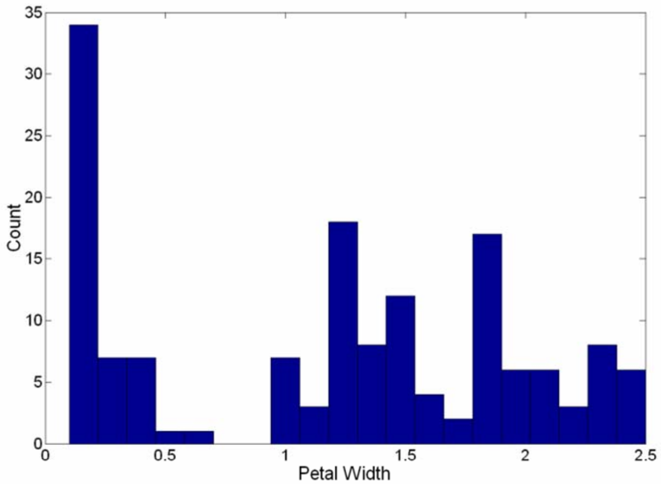
- Selection may involve the chossing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be considered

- Selection may also involve choosing a subset of objects
  - A region of the screen can only show so many points
  - Can sample, but want to preserve points in sparse areas

Visualization Techniques: Histograms

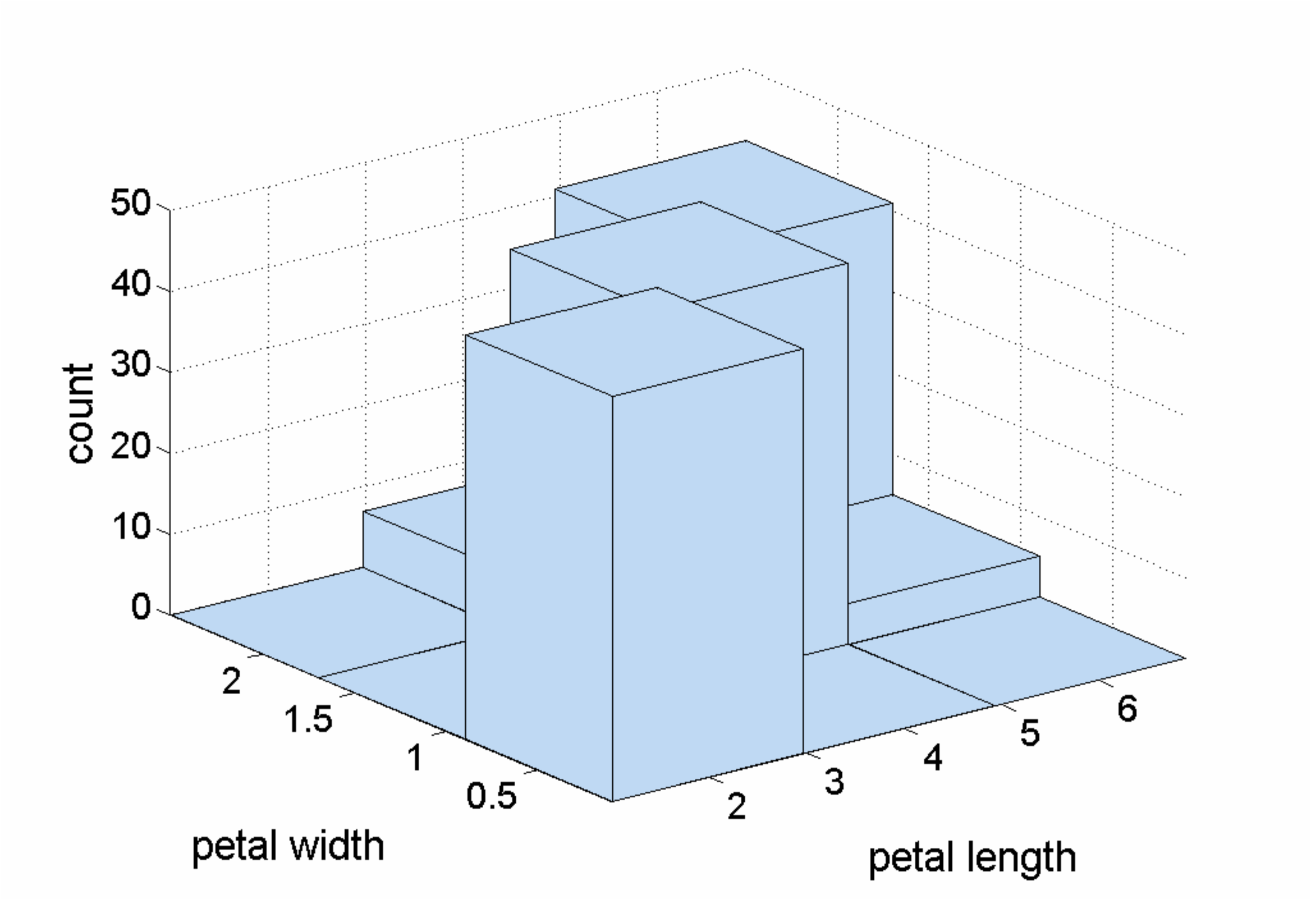
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)





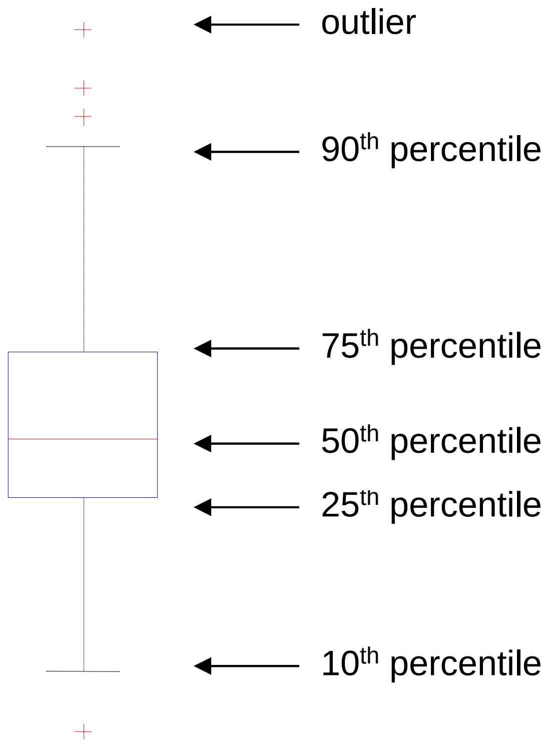
Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



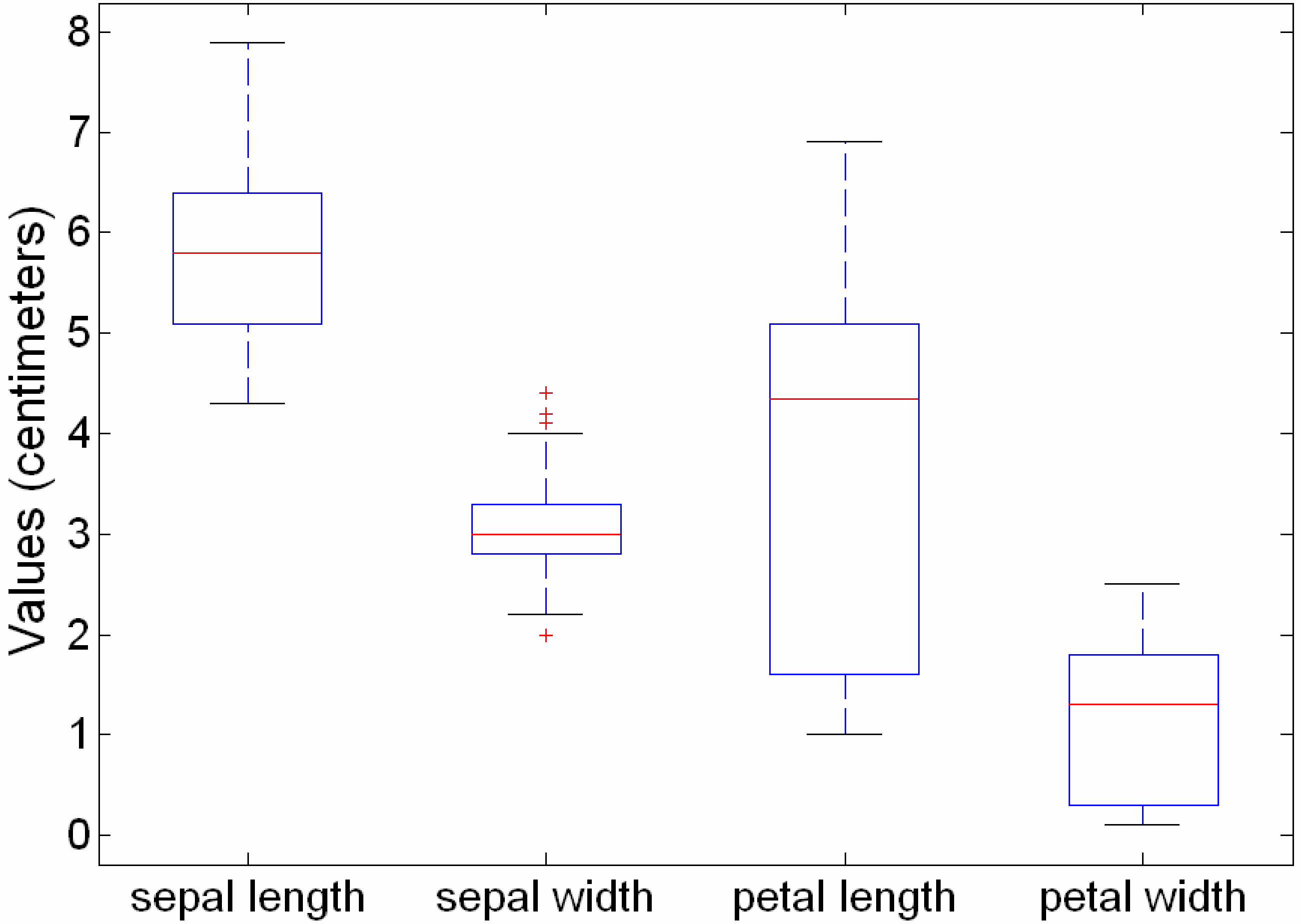
Visualization Techniques: Box Plots

- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



Example of Box Plots

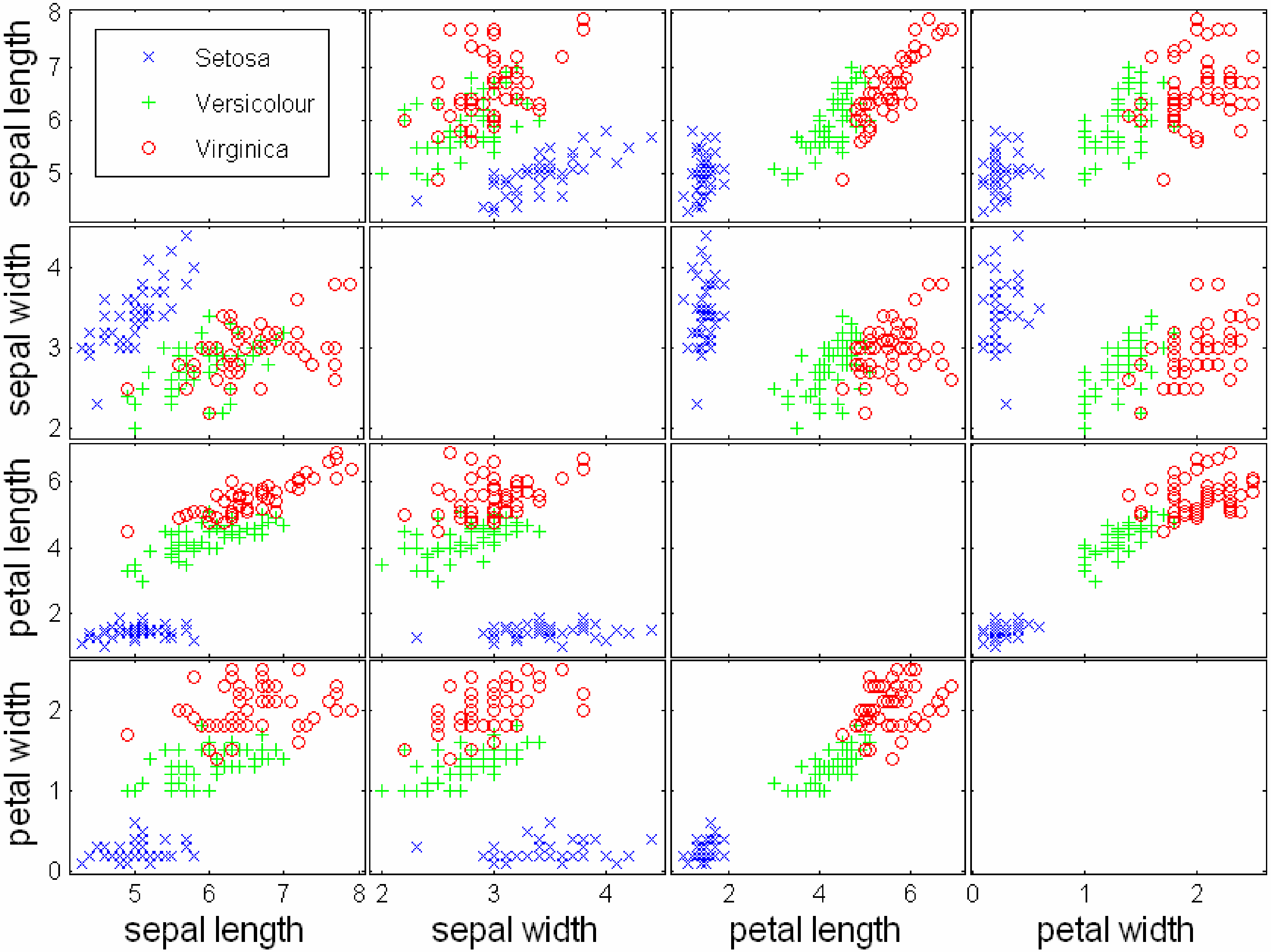
- Box plots can be used to compare attributes



Visualization Techniques: Scatter Plots

- ❑ Attributes values determine the position
- ❑ Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
- ❑ Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
- ❑ It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes

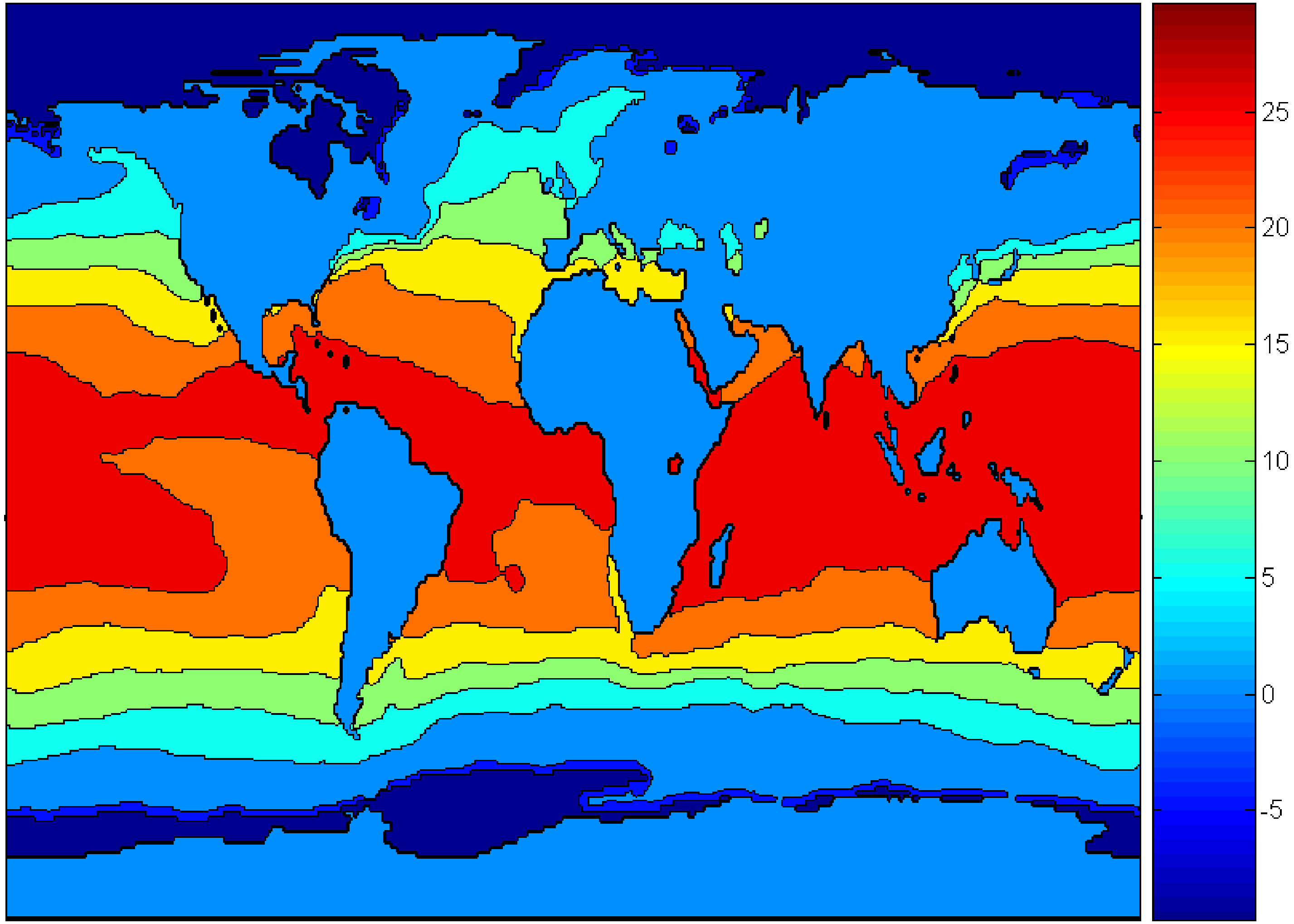
Scatter Plot Array of Iris Attributes



Visualization Techniques: Contour Plots

- ❑ Useful when a continuous attribute is measured on a spatial grid
- ❑ They partition the plane into regions of similar values
- ❑ The contour lines that form the boundaries of these regions connect points with equal values
- ❑ The most common example is contour maps of elevation
- ❑ Can also display temperature, rainfall, air pressure, etc.

Contour Plot Example: SST Dec, 1998

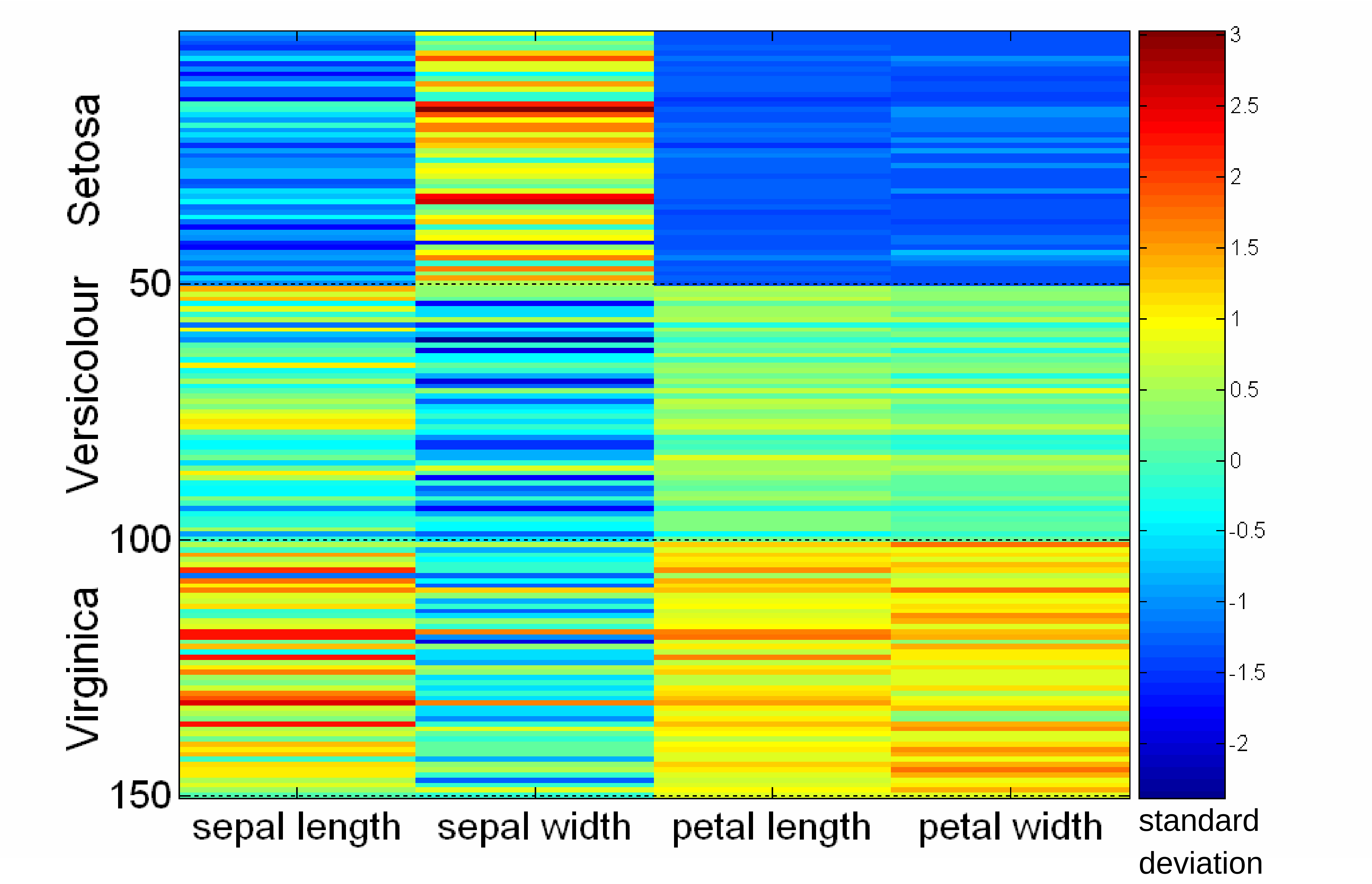


Celsius

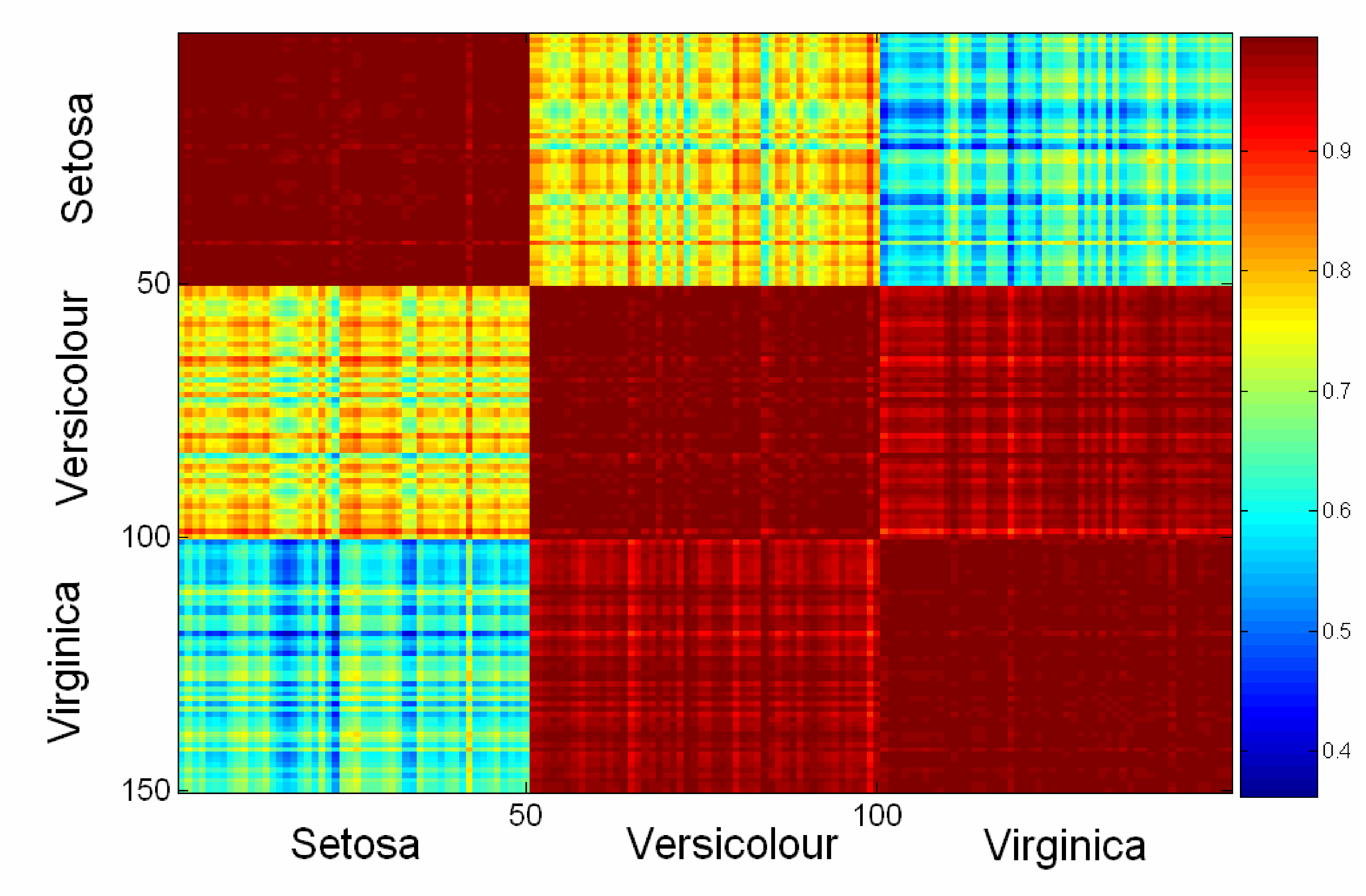
Visualization Techniques: Matrix Plots

- ❑ Can plot the data matrix
- ❑ This can be useful when objects are sorted according to class
- ❑ Typically, the attributes are normalized to prevent one attribute from dominating the plot
- ❑ Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects
- ❑ Examples of matrix plots are presented on the next two slides

Visualization of the Iris Data Matrix



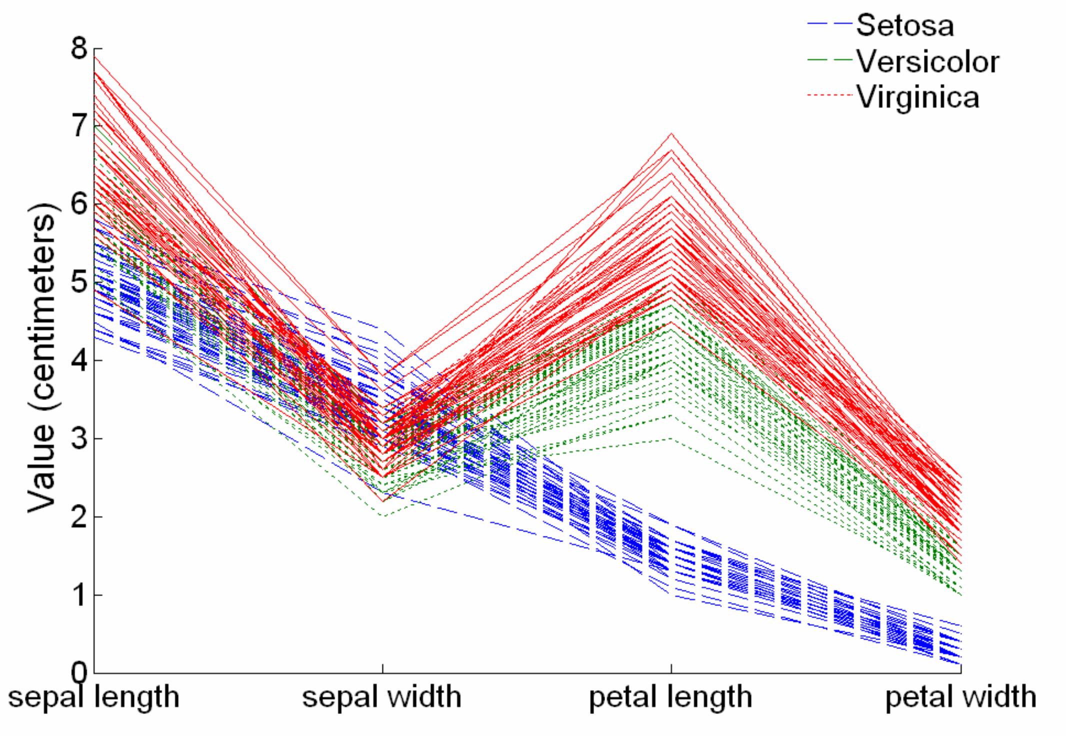
Visualization of the Iris Correlation Matrix

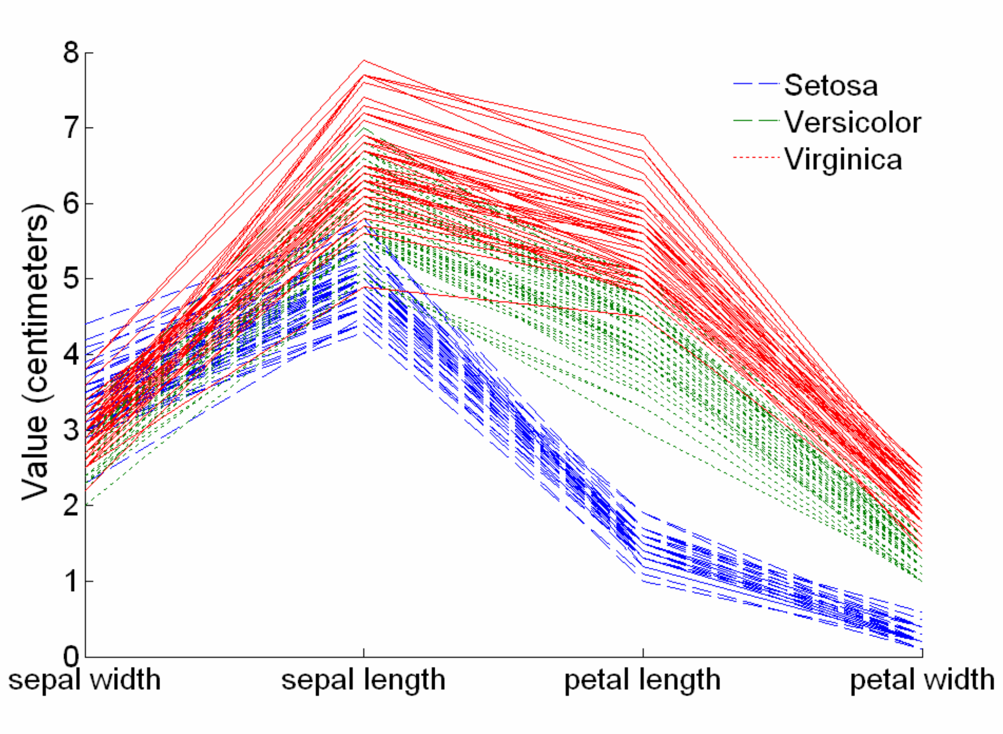


Visualization Techniques: Parallel Coordinates

- ❑ Used to plot the attribute values of high-dimensional data
- ❑ Instead of using perpendicular axes, use a set of parallel axes
- ❑ The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
- ❑ Thus, each object is represented as a line
- ❑ Often, the lines representing a distinct class of objects group together, at least for some attributes
- ❑ Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots for Iris Data



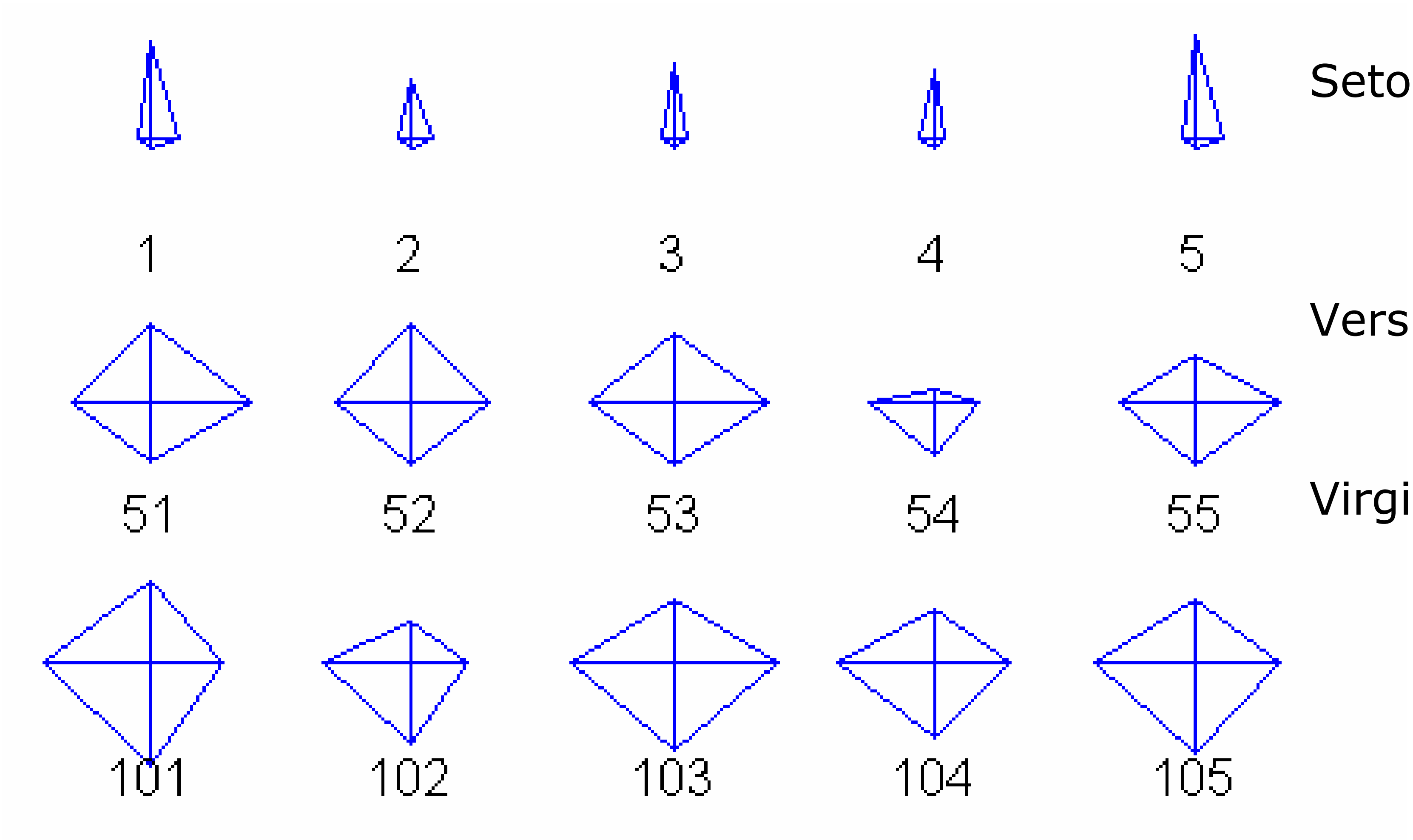


Other Visualization Techniques

- ❑ Star Plots
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon

- ❑ Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human’s ability to distinguish faces

Star Plots for Iris Data

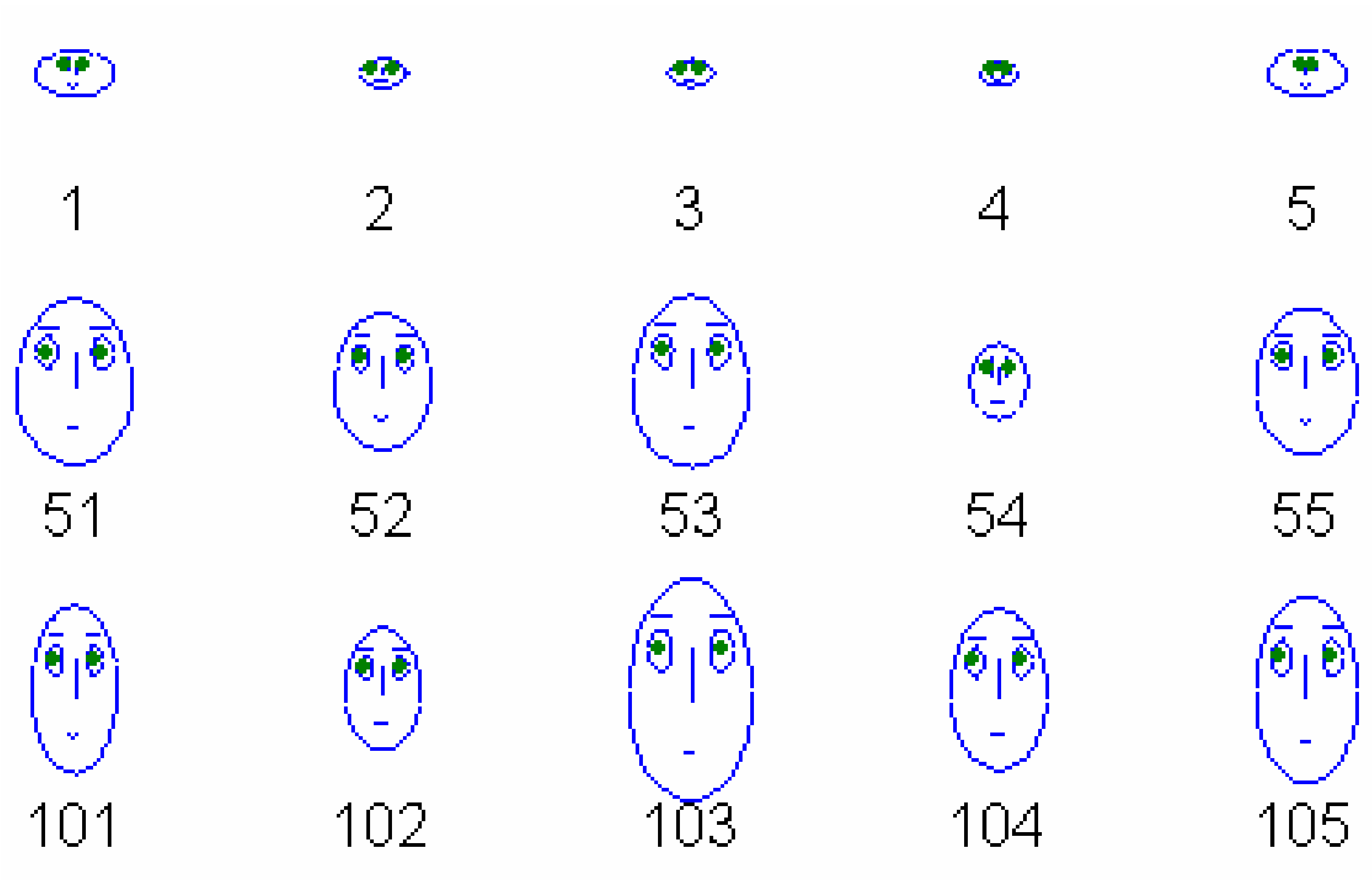


Setosa

Versicolour

Virginica

Chernoff Faces for Iris Data



Setosa

Versicolour

Virginica

Data Preprocessing

Why Data Preprocessing?

- ❑ Data in the real world is dirty

- ❑ Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“ ”
- ❑ Noisy: containing errors or outliers
  - e.g., Salary=“-10”

- ❑ Inconsistent: containing discrepancies in codes or names
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - “Not applicable” data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
- Quality decisions must be based on quality data
  - E.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

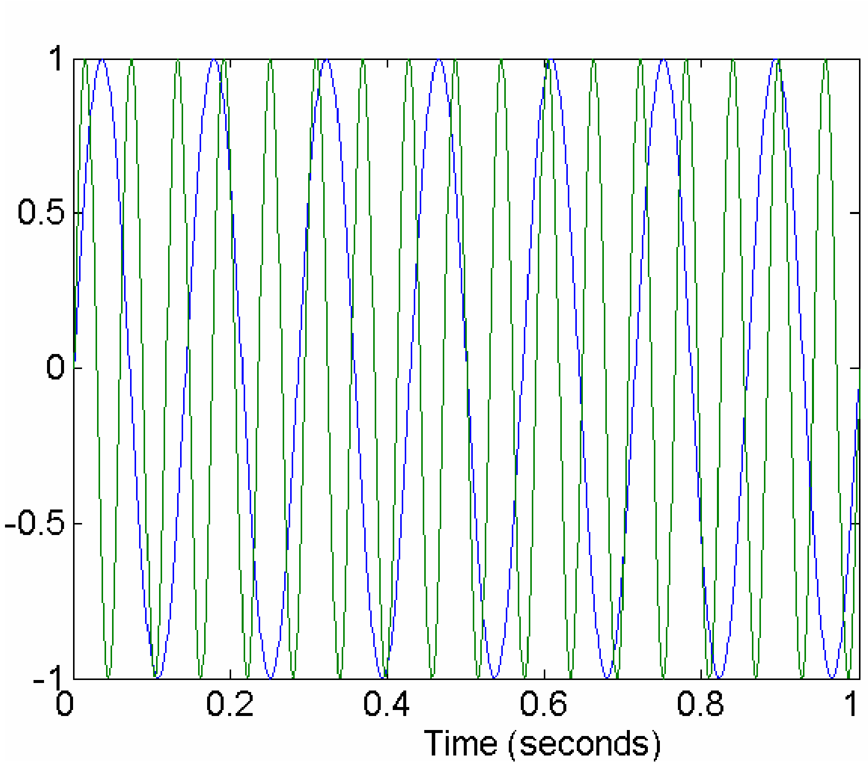
- Examples of data quality problems:
  - noise and outliers
  - missing values
  - duplicate data

Multi-Dimensional Measure of Data Quality

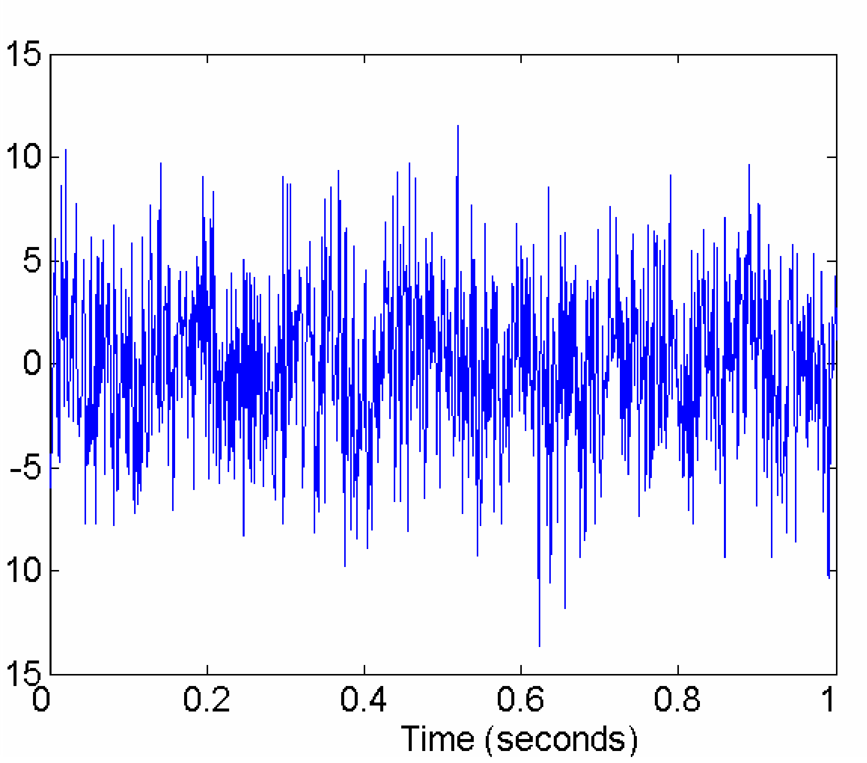
- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility
- Broad categories:
  - Intrinsic, contextual, representational, and accessibility

Noise

- Noise refers to modification of original values
- Examples: distortion of a person’s voice when talking on a poor phone and “snow” on television screen



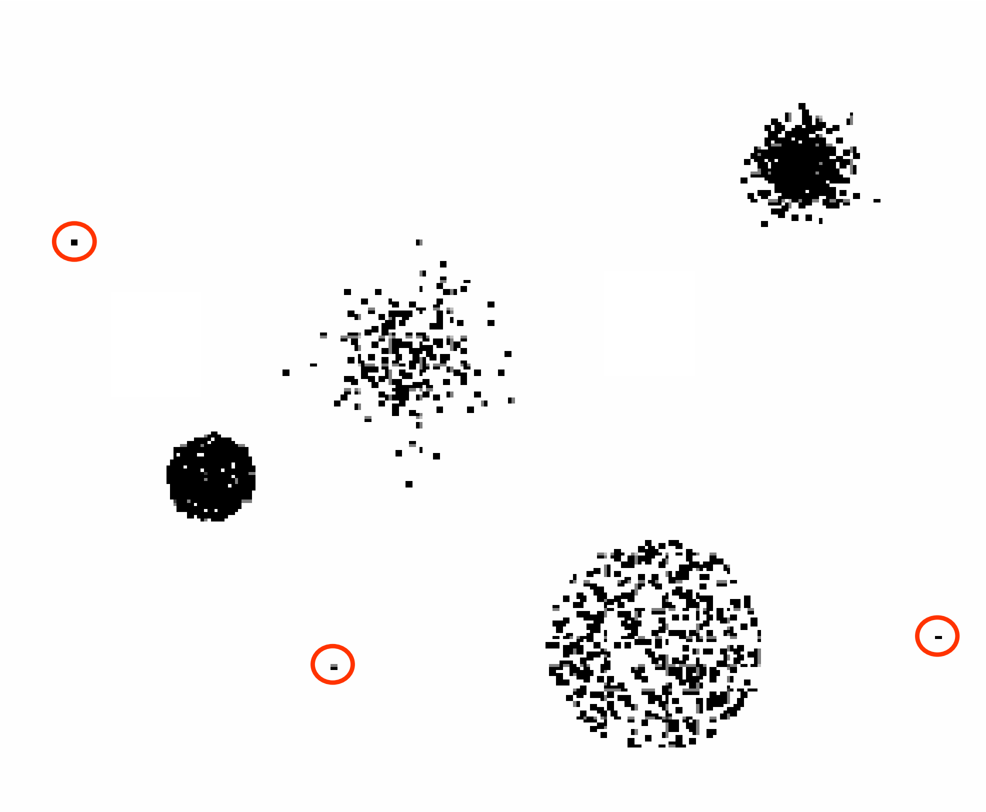
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
- Major issue when merging data from heterogeous sources
- Examples: same person with multiple email addresses
- Data cleaning: process of dealing with duplicate data issues

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter’s Wheels)

Data Preprocessing

- ❑ Aggregation
- ❑ Sampling
- ❑ Dimensionality Reduction
- ❑ Feature subset selection
- ❑ Feature creation
- ❑ Discretization and Binarization
- ❑ Attribute Transformation

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction: reduce the number of attributes or objects
  - Change of scale: cities aggregated into regions, states, countries, etc
  - More “stable” data: aggregated data tends to have less variability

Sampling

- Sampling is the main technique employed for data selection
- It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming

Key principles for Effective Sampling

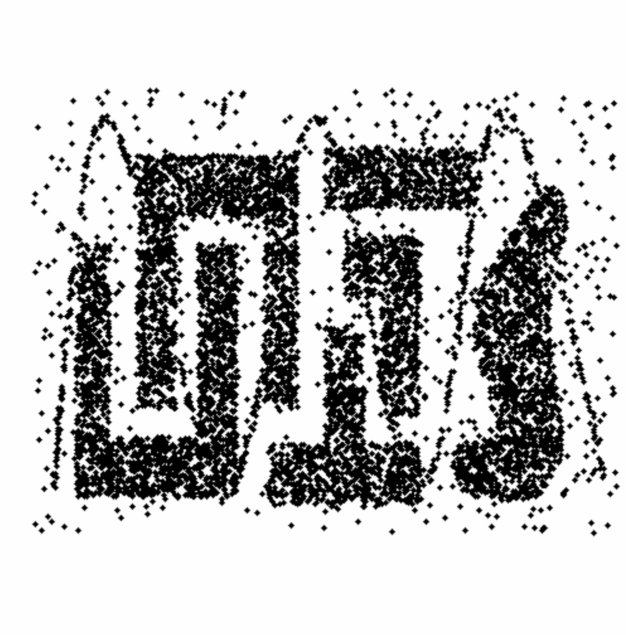
- Using a sample will work almost as well as using the entire data sets, if the sample is representative

- A sample is representative if it has approximately the same property (of interest) as the original set of data

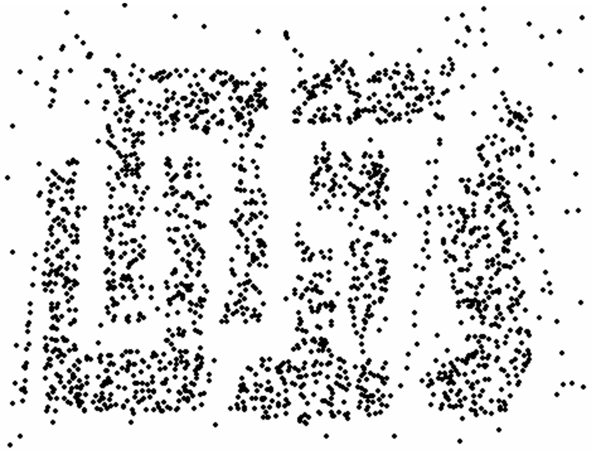
Types of Sampling

- ❑ Simple Random Sampling
  - There is an equal probability of selecting any particular item
- ❑ Sampling without replacement
  - As each item is selected, it is removed from the population
- ❑ Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
  - In sampling with replacement, the same object can be picked up more than once
- ❑ Stratified sampling
  - Split the data into several partitions
  - Then draw random samples from each partition

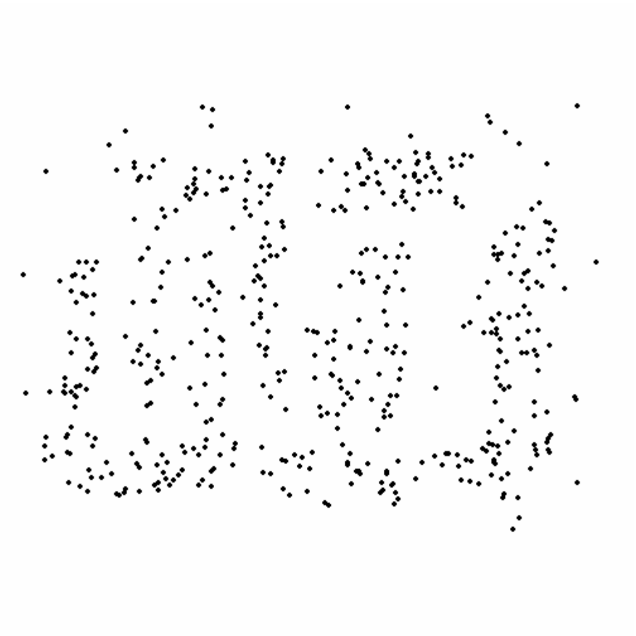
Sample Size



8000 points



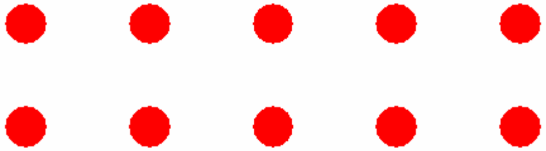
2000 Points

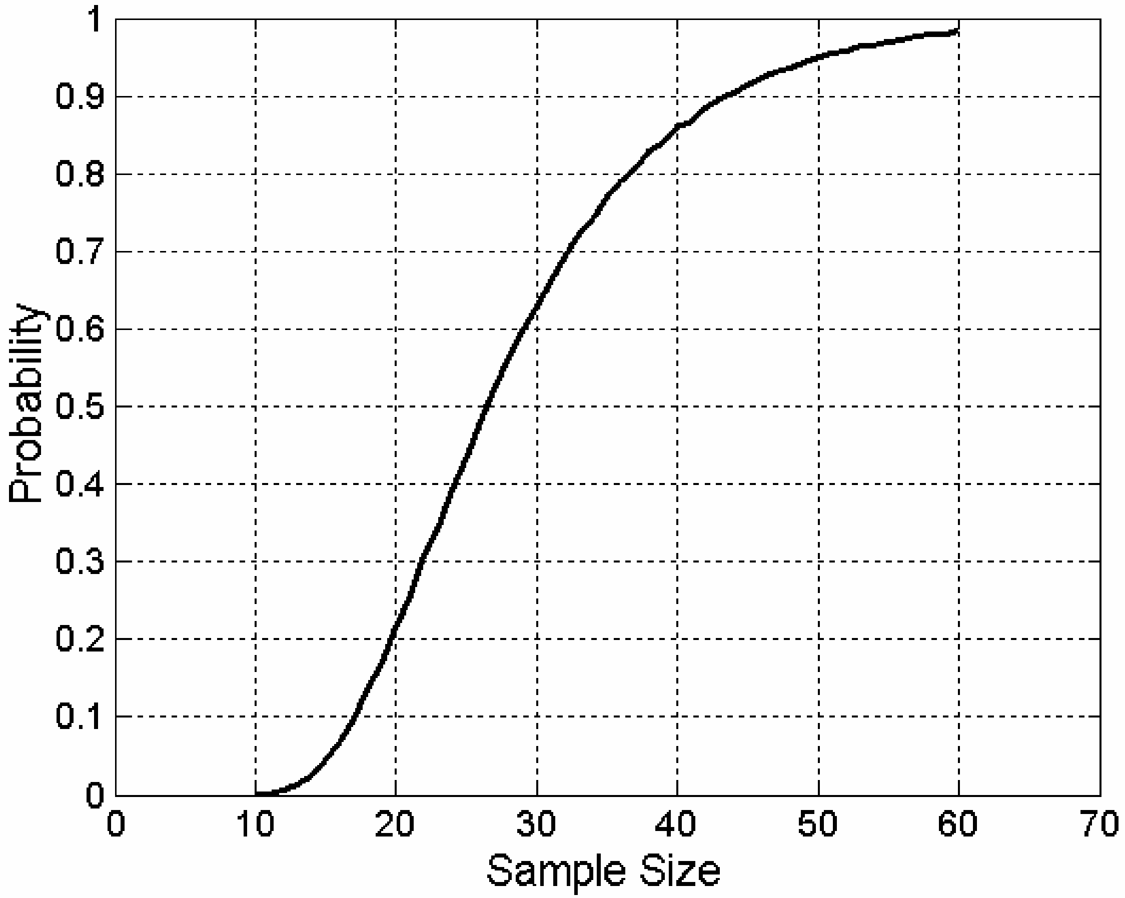


500 Points

Sample Size

- ❑ What sample size is necessary to get at least one object from each of 10 groups.

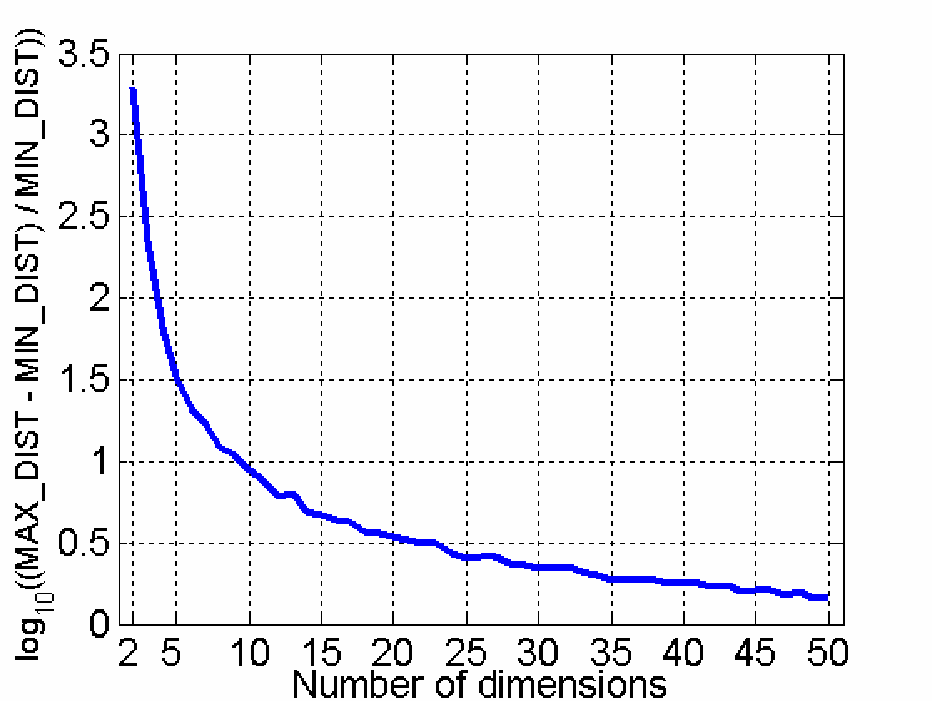




Curse of Dimensionality

❑When dimensionality increases, data becomes increasingly sparse in the space that it occupies

❑Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

Dimensionality Reduction

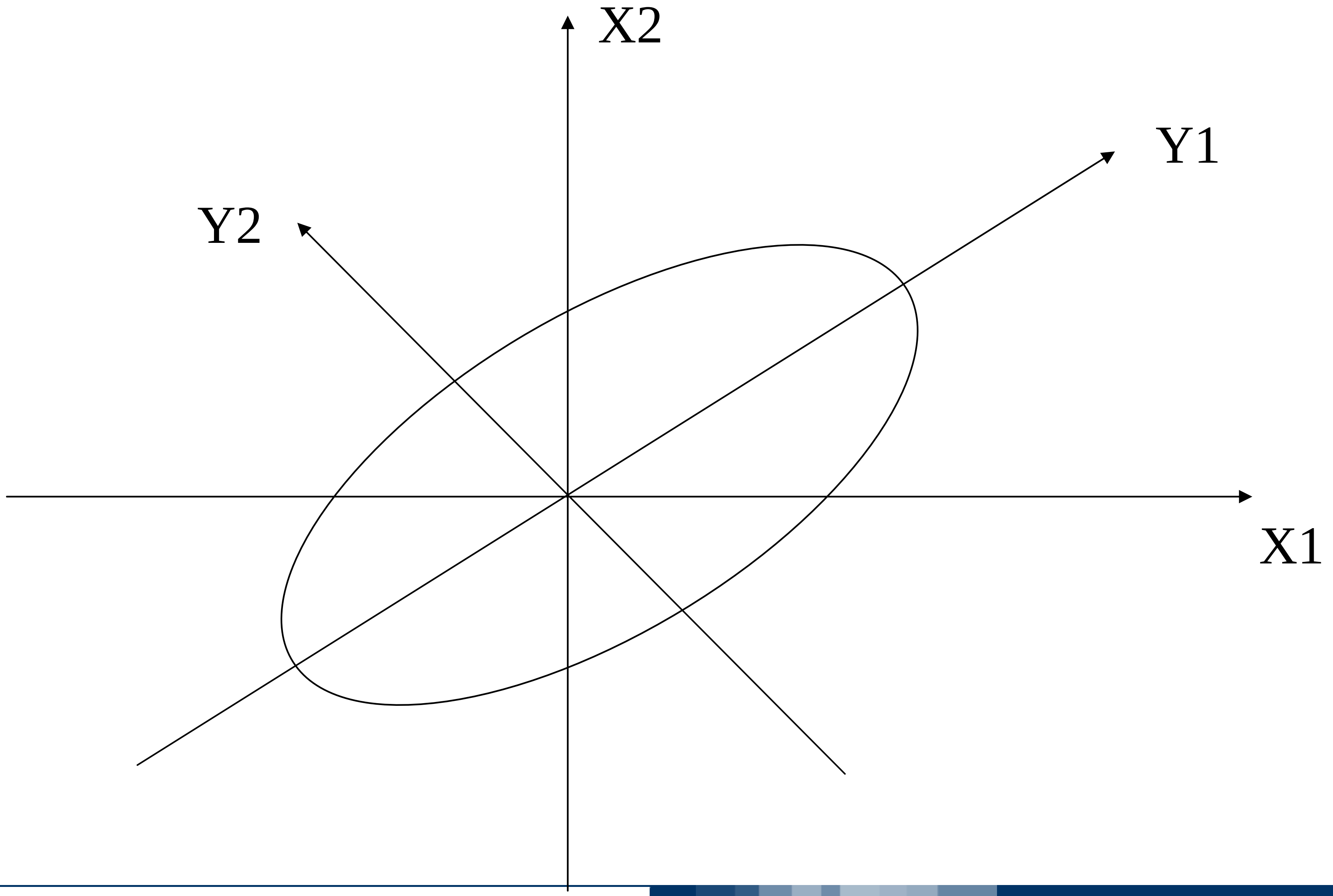
- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Dimensionality Reduction: Principal Component Analysis (PCA)

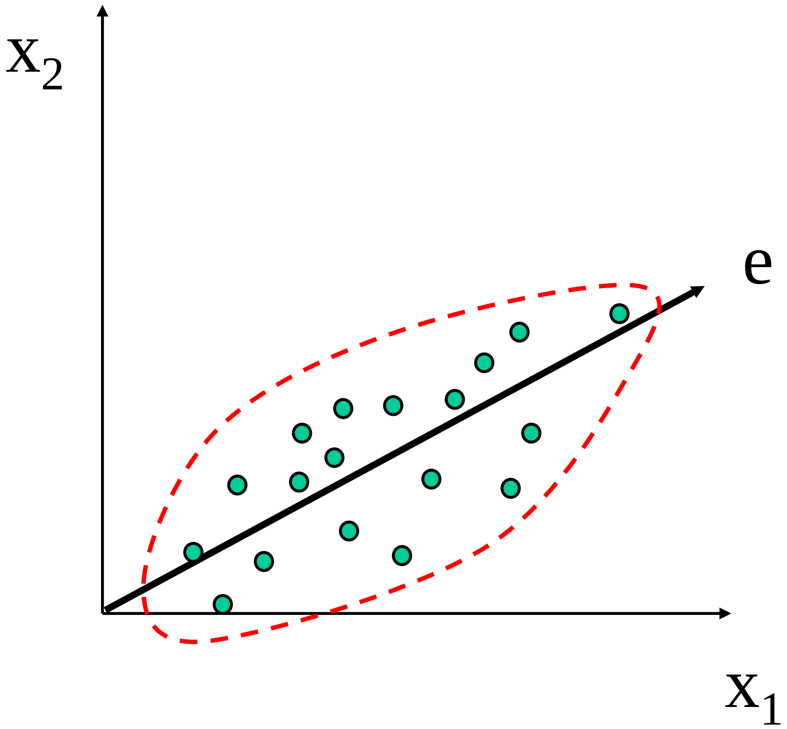
- Given N data vectors from n-dimensions, find k ≤ n orthogonal vectors (principal components) that can be best used to represent data
- Steps
  - Normalize input data: Each attribute falls within the same range
  - Compute k orthonormal (unit) vectors, i.e., principal components
  - Each input data (vector) is a linear combination of the k principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only
- Used when the number of dimensions is large

Principal Component Analysis



Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



Feature Subset Selection

- Another way to reduce dimensionality of data

- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid

- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Subset Selection

- Brute-force approach
  - Try all possible feature subsets as input to data mining algorithm
- Embedded approaches
  - Feature selection occurs naturally as part of the data mining algorithm
- Filter approaches
  - Features are selected using a procedure that is independent from a specific data mining algorithm
  - E.g., attributes are selected based on correlation measures
- Wrapper approaches:
  - Use a data mining algorithm as a black box to find best subset of attributes
  - E.g., apply a genetic algorithm and an algorithm for decision tree to find the best set of features for a decision tree

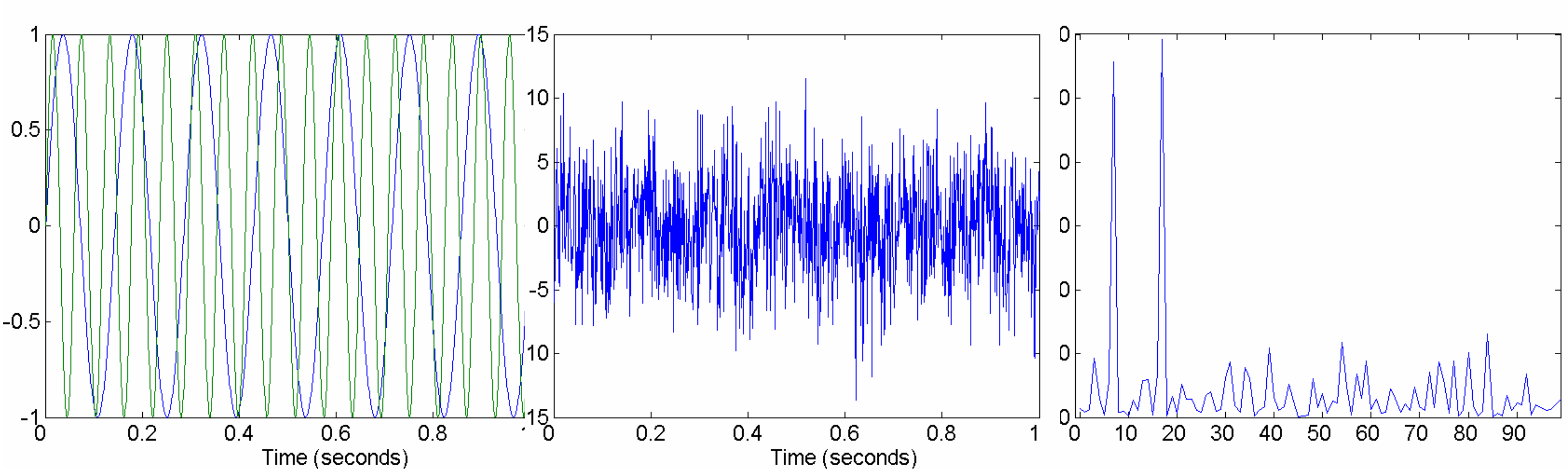
Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- E.g., given the birthday, create the attribute age

- Three general methodologies:
  - Feature Extraction: domain-specific
  - Mapping Data to New Space
  - Feature Construction: combining features

Mapping Data to a New Space

- ❑ Fourier transform
- ❑ Wavelet transform



Two Sine Waves

Two Sine Waves + Noise

Frequency

Discretization

- Three types of attributes:
  - Nominal: values from an unordered set, e.g., color
  - Ordinal: values from an ordered set, e.g., military or academic rank
  - Continuous: real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes
  - Reduce data size by discretization
  - Prepare for further analysis

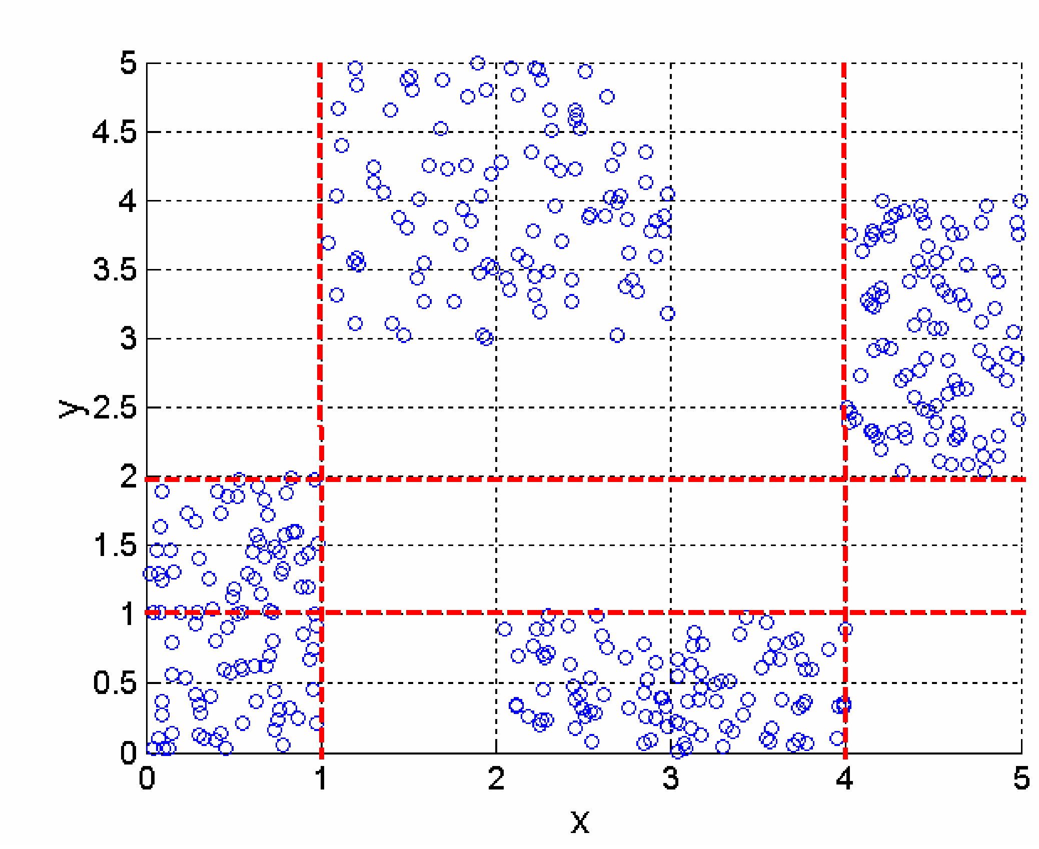
Discretization Approaches

- Supervised
  - Attributes are discretized using the class information
  - Generates intervals that tries to minimize the loss of information about the class

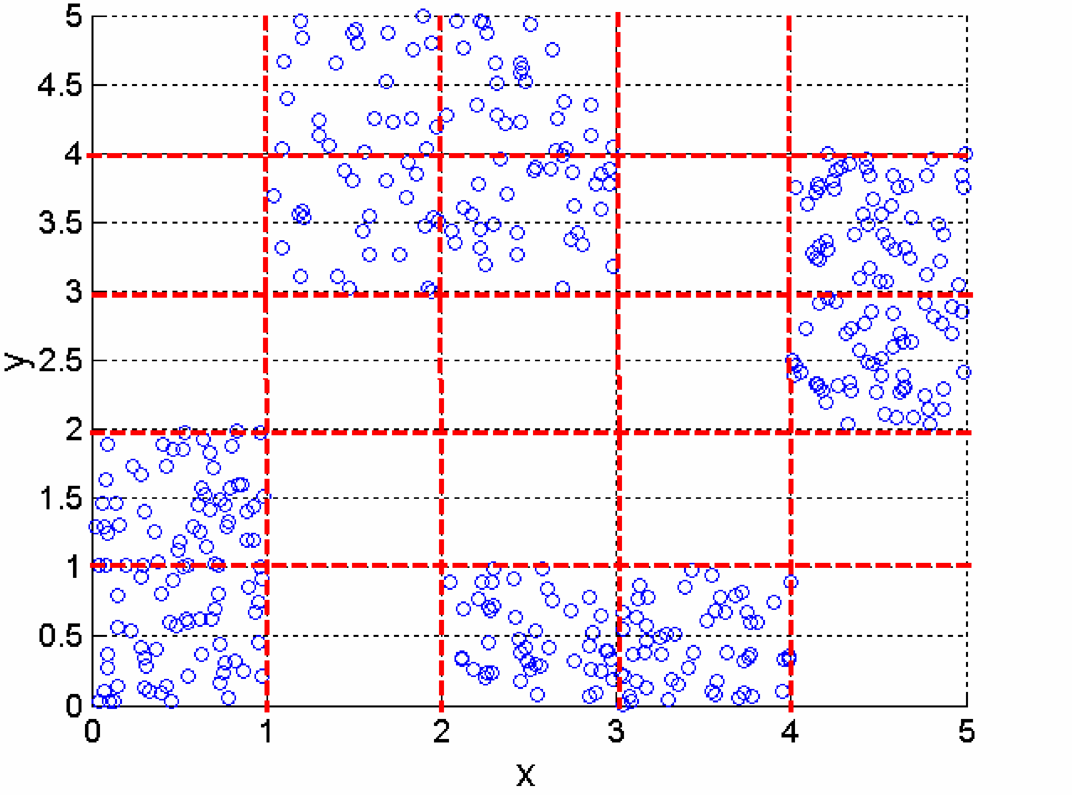
- Unsupervised
  - Attributes are discretized solely based on their values

Discretization Using Class Labels

- Entropy based approach

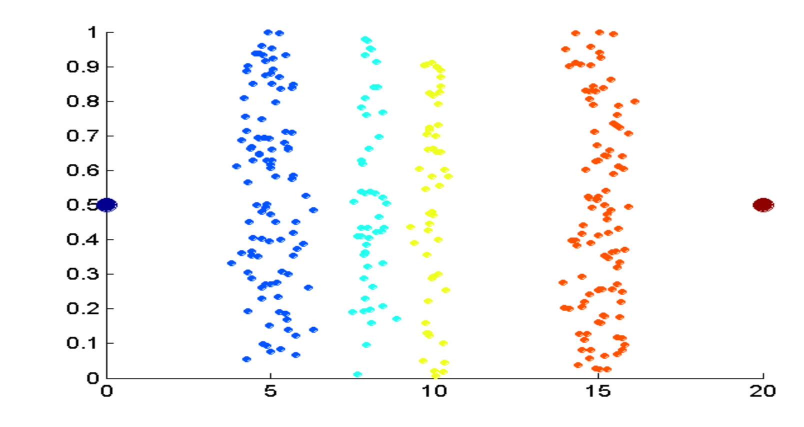


3 categories for both x and y

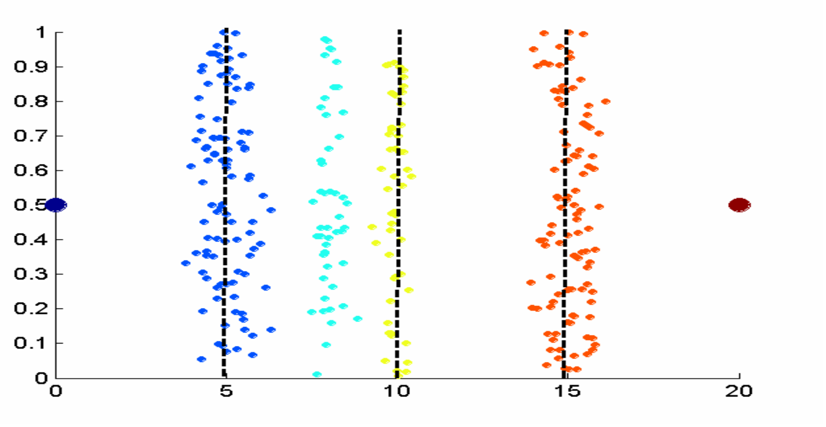


5 categories for both x and y

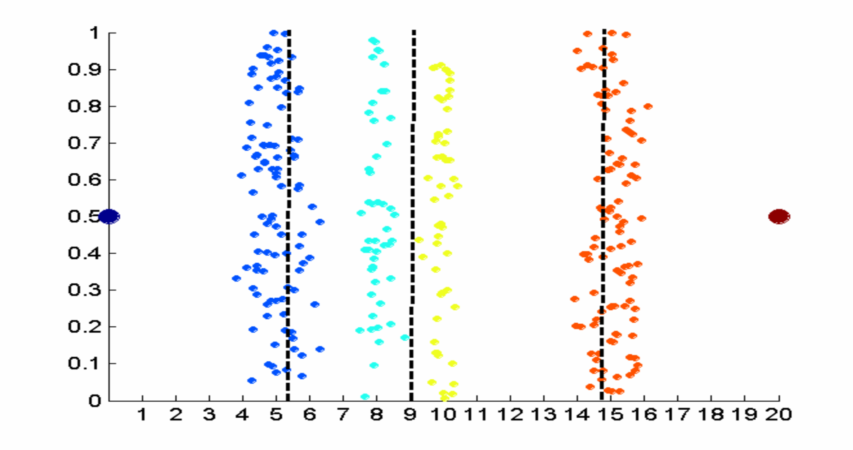
Discretization Without Using Class Labels



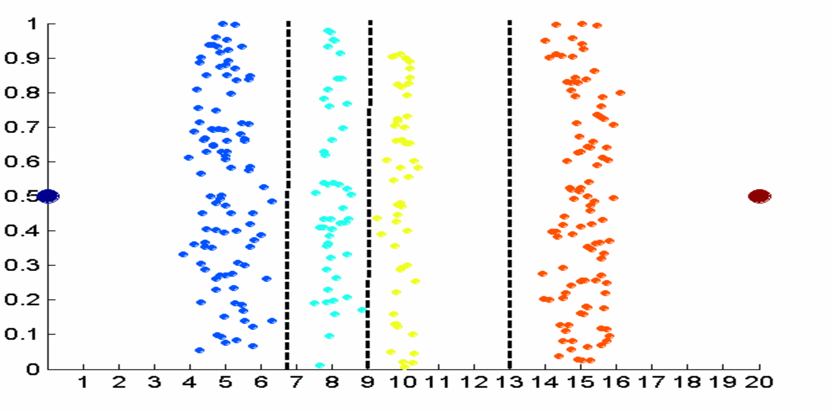
Data



Equal interval width



Equal frequency



K-means

Segmentation by Natural Partitioning

- A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, “natural” intervals.
- If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals
- If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
- If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

Concept Hierarchy Generation for Categorical Data

- ❑ Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- ❑ Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- ❑ Specification of only a partial set of attributes
  - E.g., only street < city, not others
- ❑ Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {street, city, state, country}

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year

Country	15 distinct values
Province	365 distinct values
City	3567 distinct values
Street	674,339 distinct values

Summary

- Data exploration and preparation, or preprocessing, is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but data preprocessing still an active area of research