

Statistical Learning and Data Mining

CS 363D/ SSC 358

Lecture: Exploratory Data Analysis

Pradeep Ravikumar
pradeepr@cs.utexas.edu

Adapted From: Pang-Ning Tan, Steinbach, Kumar

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

Summary Statistics

- Summary statistics 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 p th 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 $x_{50\%}$.

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.

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

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

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.

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

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

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

$$\text{MAD}(x) = \text{median}\left(\{|x_1 - \bar{x}|, \dots, |x_m - \bar{x}|\}\right)$$

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

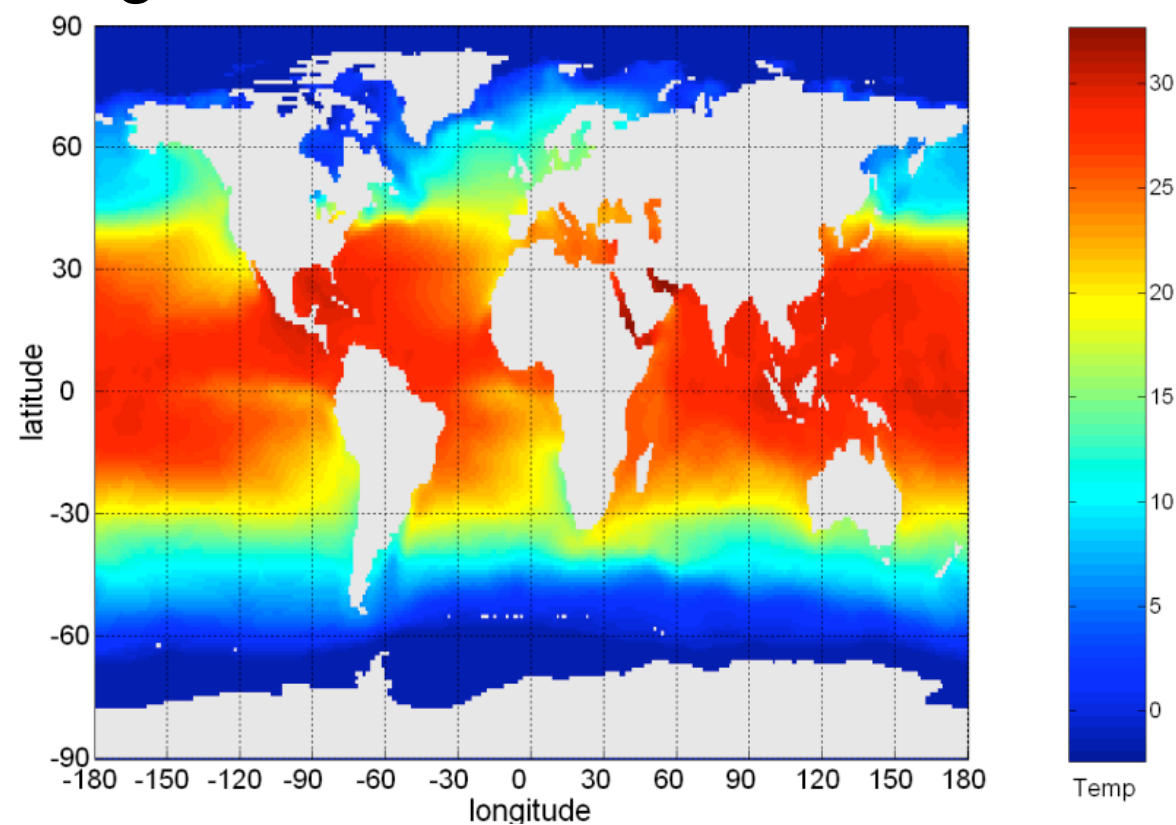
Visualization

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 (SST) 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.

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
 - Can be obtained from the UCI Machine Learning Repository <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



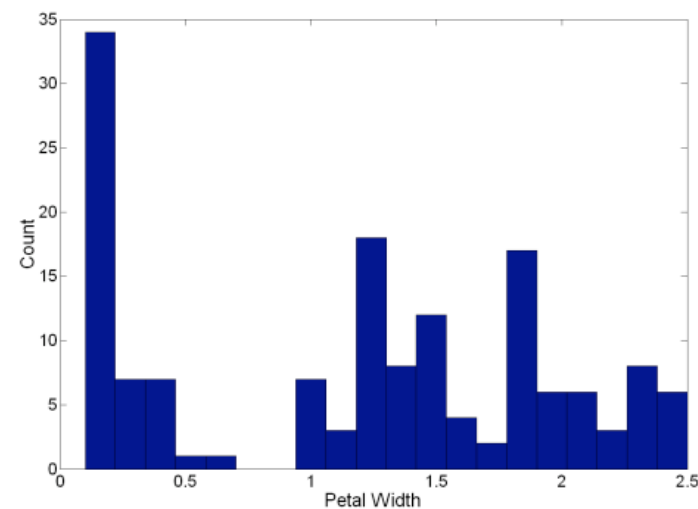
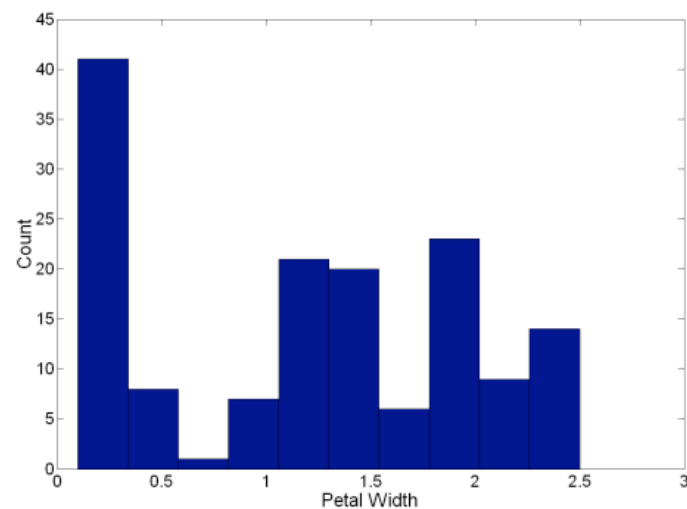
Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Visualization Techniques: Histograms

- Histogram

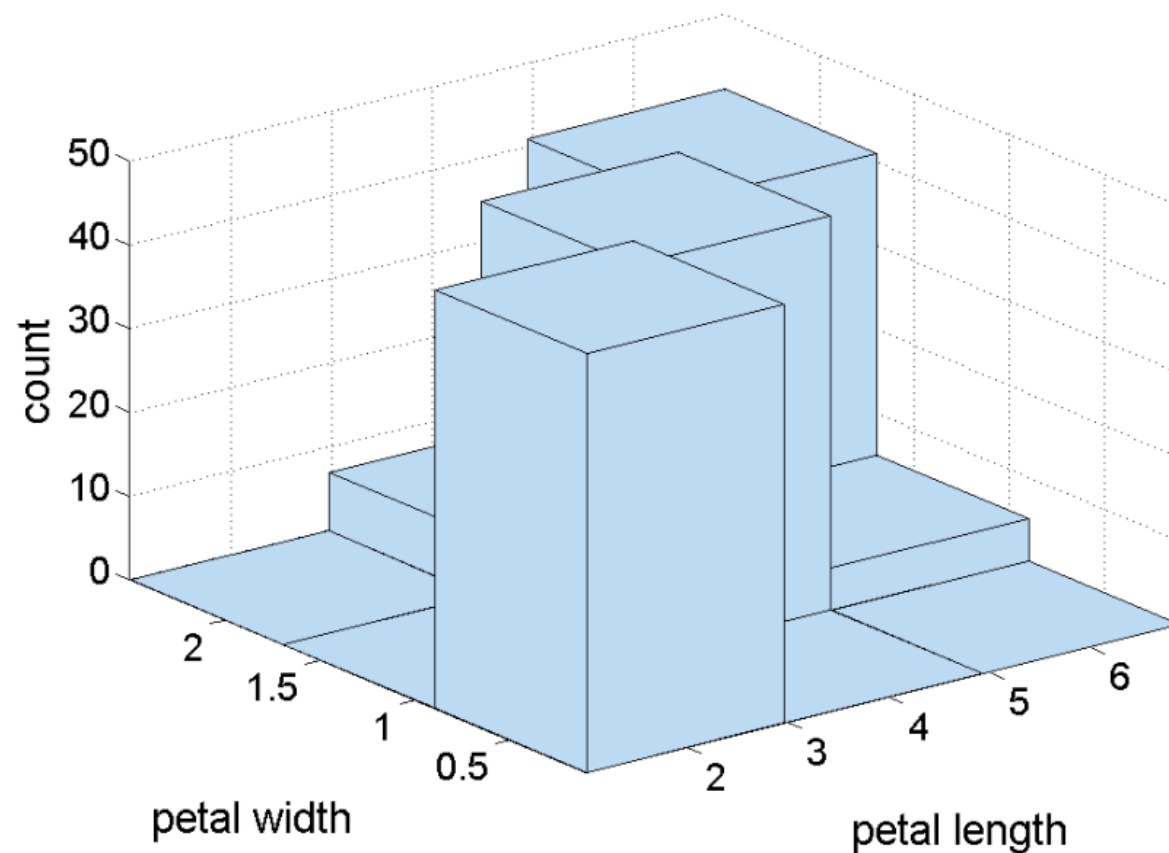
- Usually shows the distribution of values of a single variable
- Divide the values into bins and show a bar plot of the number of objects in each bin.
- The height of each bar indicates the number of objects
- Shape of histogram depends on the number of bins

- Example: Petal Width (10 and 20 bins, respectively)



Two-Dimensional Histograms

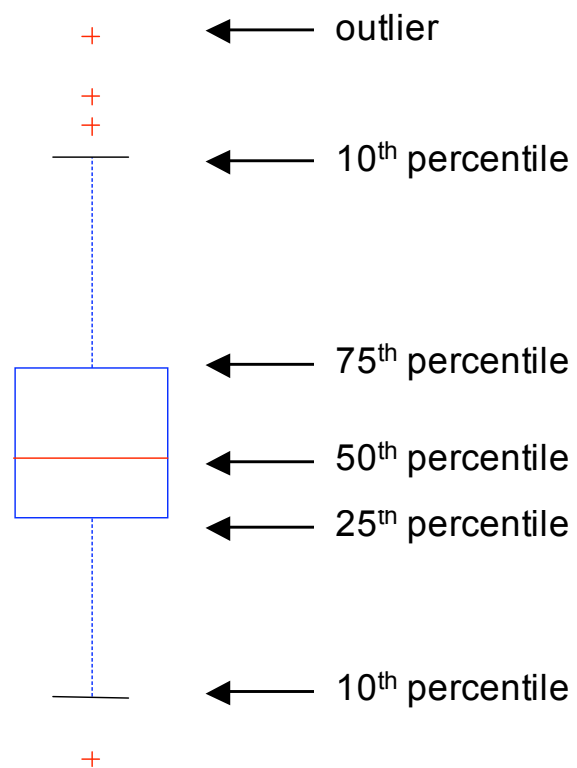
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
 - What does this tell us?



Visualization Techniques: Box Plots

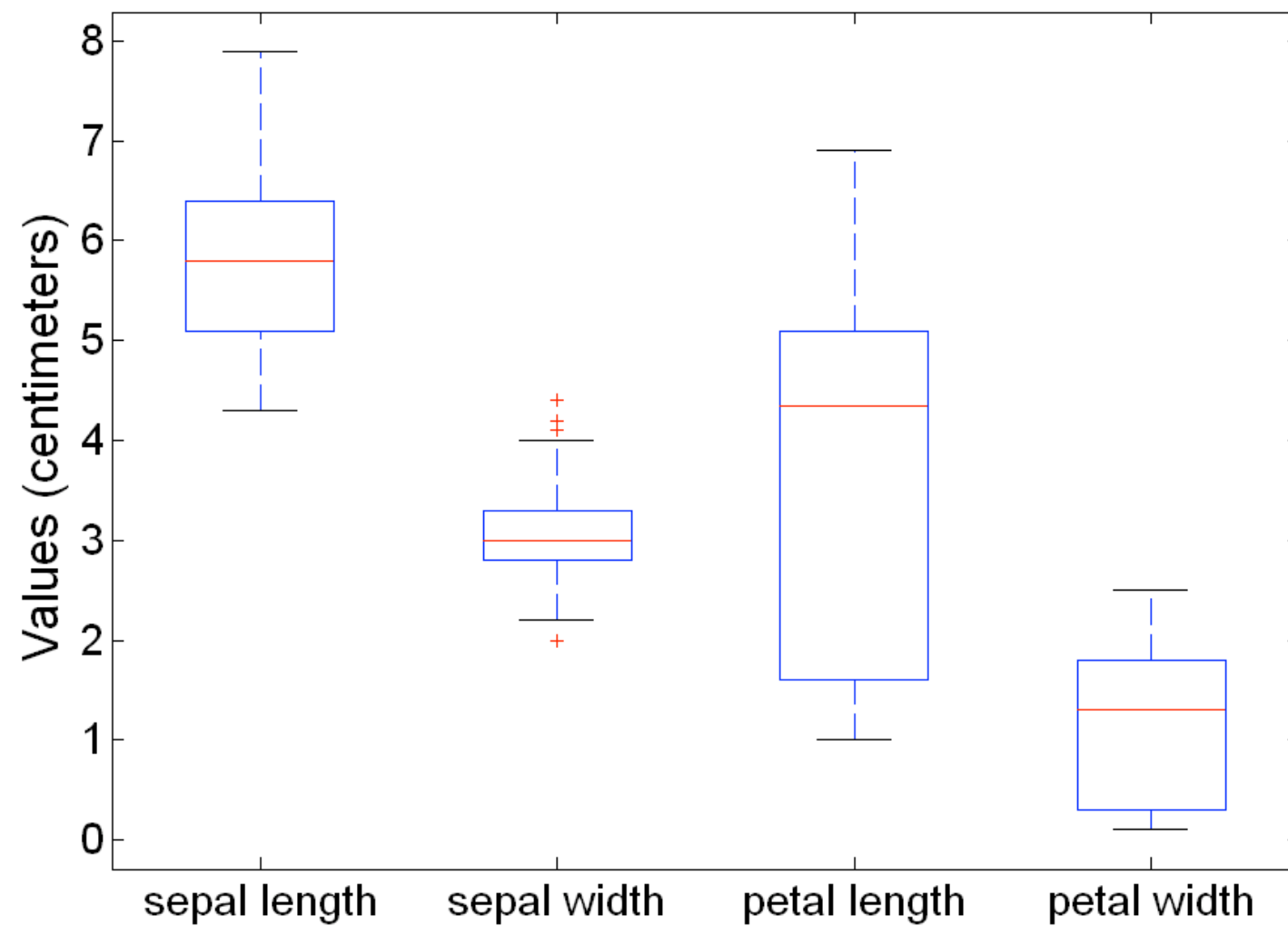
● Box Plots

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



Box Plots

- Box plots can be used to compare attributes

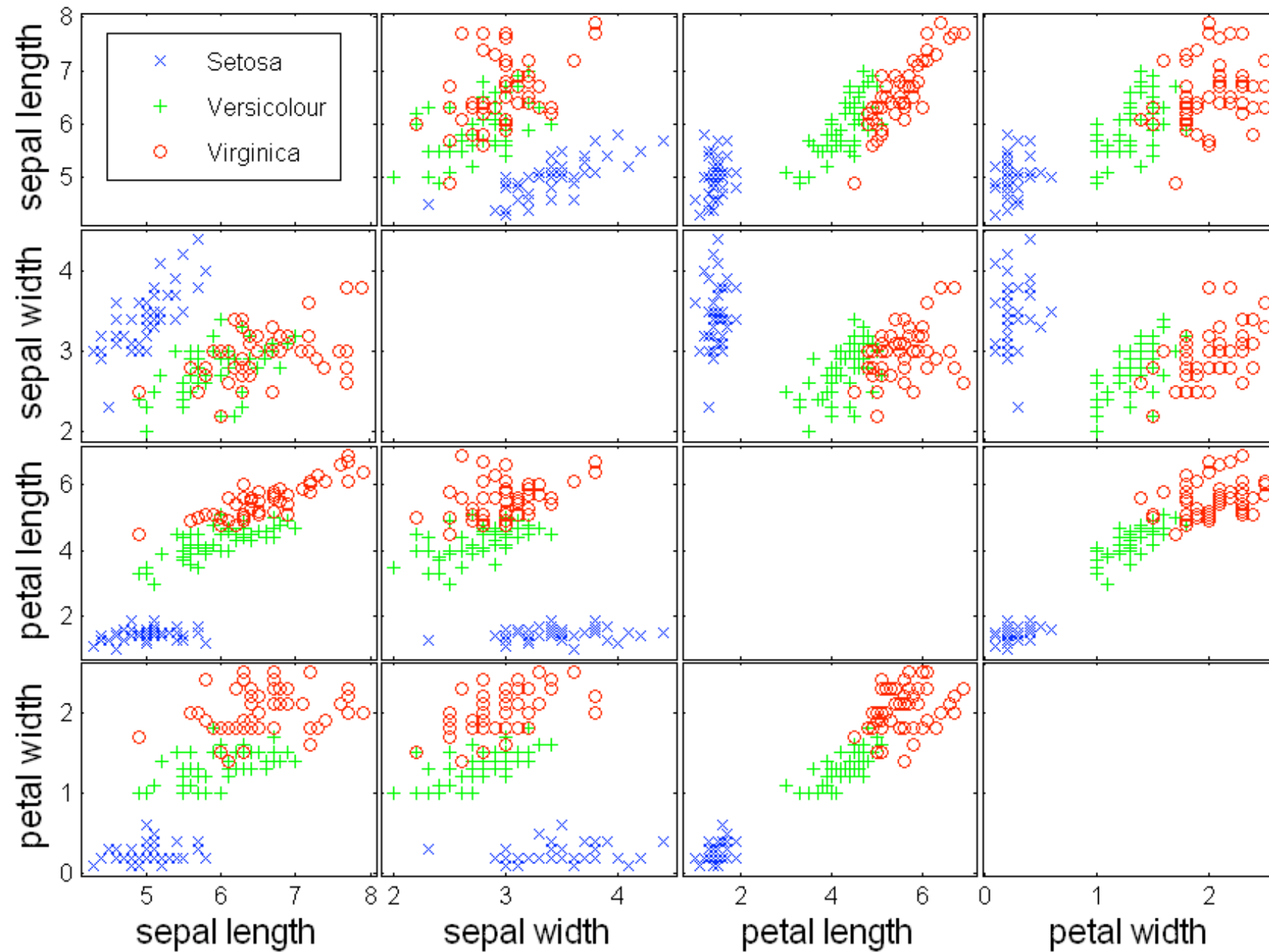


Visualization Techniques: Scatter Plots

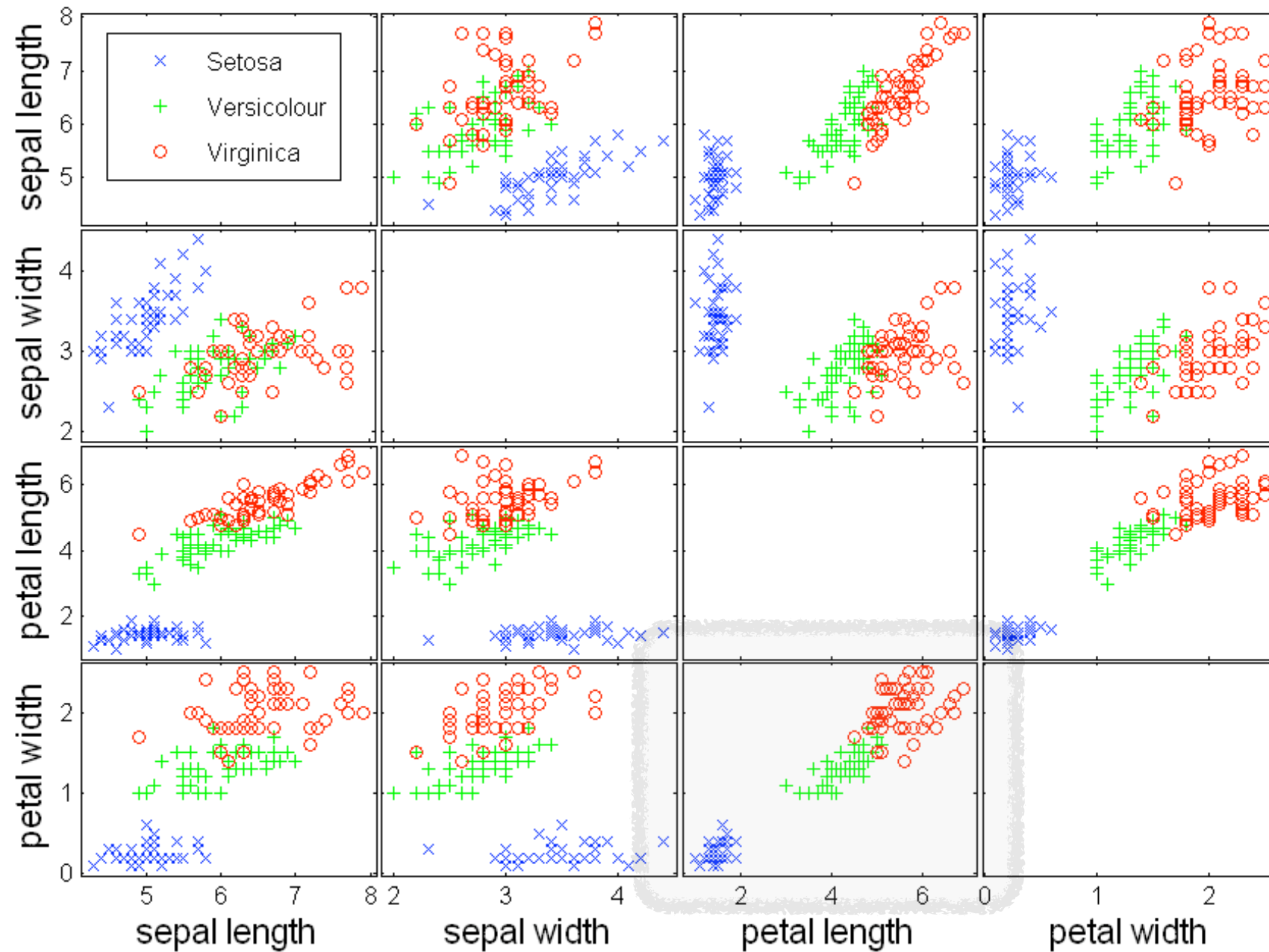
- Scatter plots

- 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
 - ◆ See example on the next slide

Scatter Plot Array of Iris Attributes



Scatter Plot Array of Iris Attributes

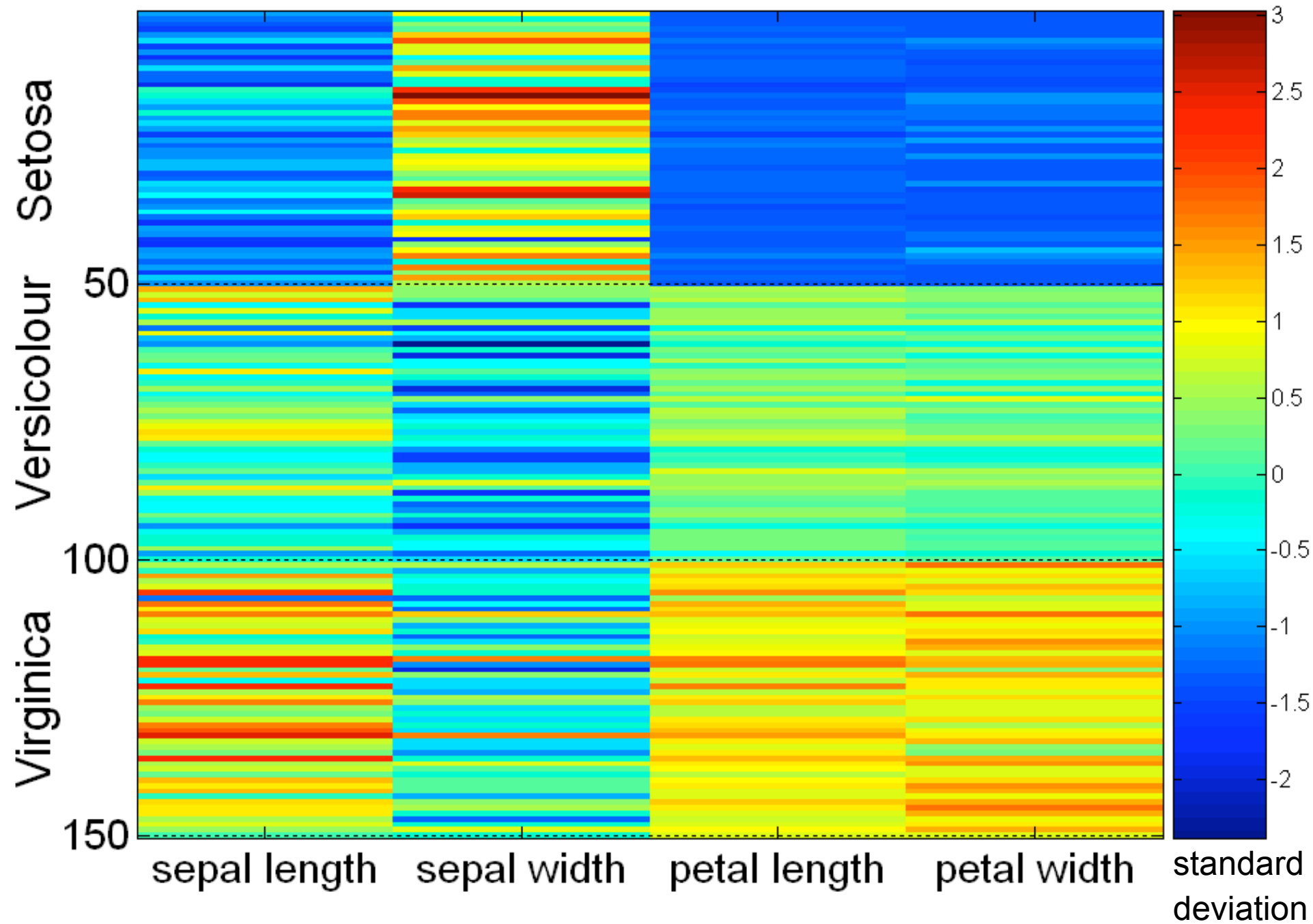


Visualization Techniques: Matrix Plots

- 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

Visualization of the Iris Data Matrix

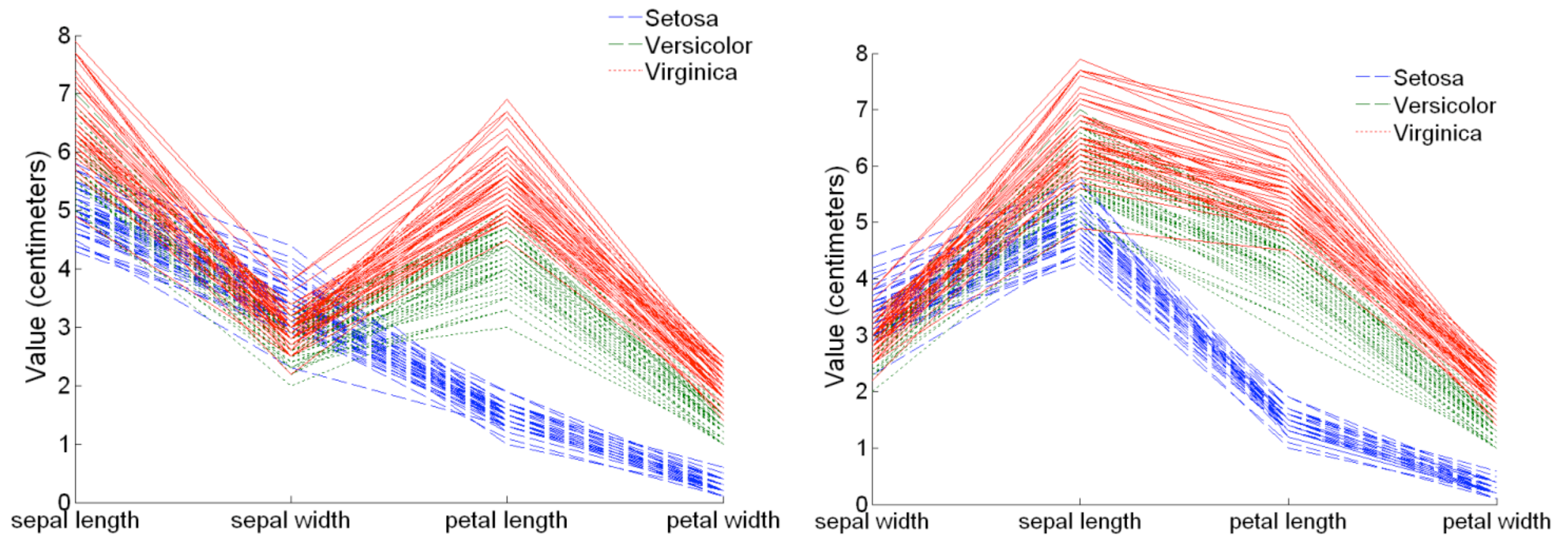


Visualization Techniques: Parallel Coordinates

● 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



Visualization Techniques: Contour Plots

- 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.
 - ◆ An example for Sea Surface Temperature (SST) is provided on the next slide

Contour Plot: SST Dec 1998

