PENNSTATE

Chapter 3: Exploring Data

Presentation extended from the slides of the textbook, Introduction to Data Mining by Tan et al. and supplementary material

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)*
  - Seminal book is Exploratory Data Analysis by J. 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 introduction of data exploration, we discuss
  - *Summary statistics*
  - *Visualization*
  - *Online Analytical Processing (OLAP)*

Iris Sample Data Set

- Discussion of many 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

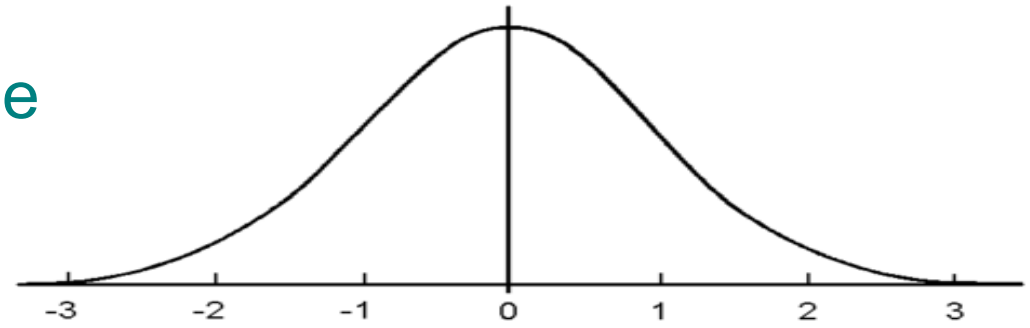
Iris Setosa

Iris Virginica

Iris Versicolour

Summary Statistics

- Summary statistics are quantities (i.e., numbers) that summarize properties of the data
  - Summarized properties include *frequency*, *location* and *spread*
  - For example



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 times 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 an attribute is the *most frequent attribute value*
- The notions of frequency and mode are typically used with categorical data

Percentiles

- *Percentile* is more useful for continuous data.
- 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 $x_{50\%}$.
- There is no standard way of calculating percentile
- *Nearest Rank* Approach
  - N is the total number of observations
  - n is the rank
  - Use the rank to find the value of percentile

$$n = \left\lceil \frac{P}{100} \times N \right\rceil$$

Nearest Rank: Examples

- Consider the ordered list {15, 20, 35, 70, 80}, which contains 5 data values
- 5th percentile? 15 (rank=1)

$$\left\lceil \frac{5}{100} \times 5 \right\rceil = \lceil 0.25 \rceil = 1$$

- 40th percentile? 20 (rank=2)

$$\left\lceil \frac{40}{100} \times 5 \right\rceil = \lceil 2.0 \rceil = 2$$

- 100th percentile? 80 (rank=5)

$$\left\lceil \frac{100}{100} \times 5 \right\rceil = \lceil 5 \rceil = 5$$

- 0th percentile?

Measures of Location: Mean and Median

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

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

- However, *the mean is very sensitive to outliers*.
- Thus, the median or a trimmed mean is also commonly used.

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

Measures of Spread: Range and Variance

极差

方差

- *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.

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

Average Absolute Deviation

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

Median Absolute Deviation

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

Visualization

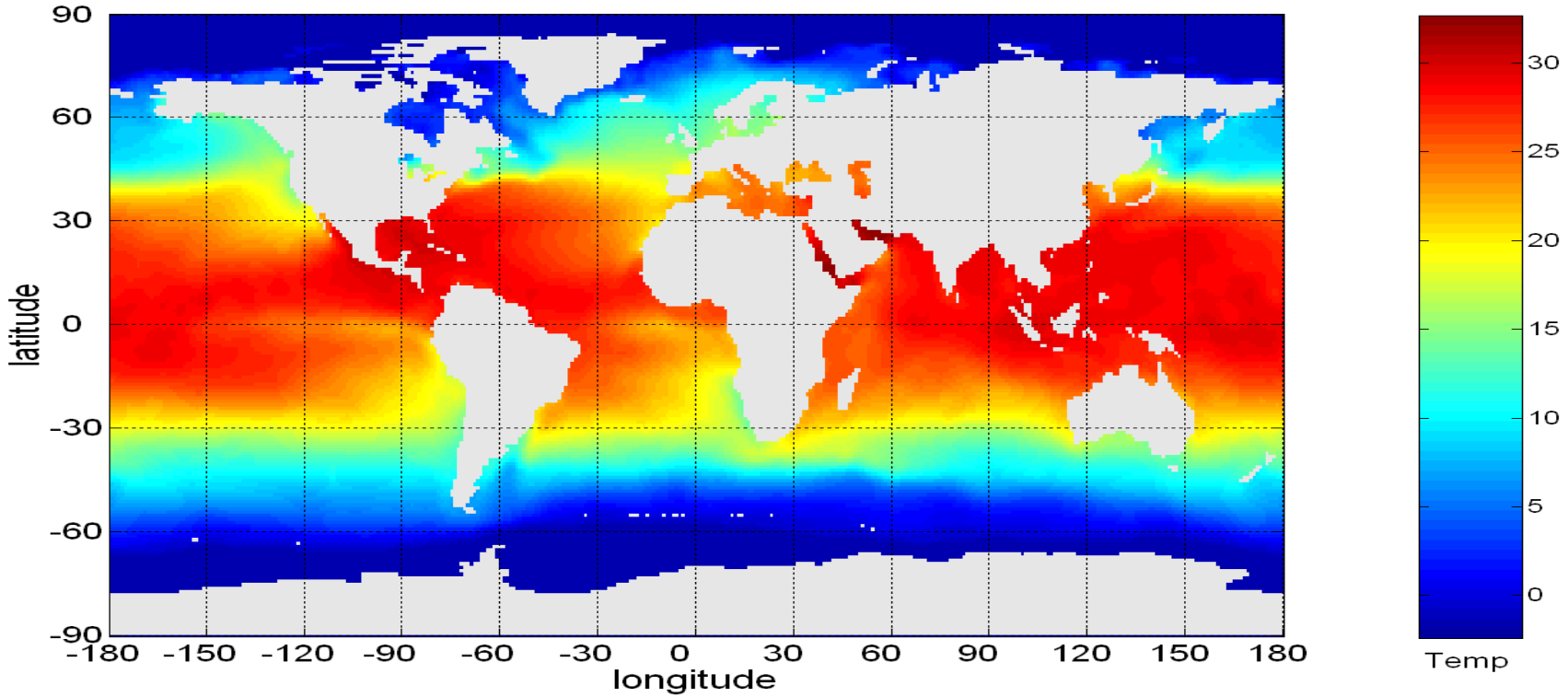
- Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- 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

- https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Example: Sea Surface Temperature

- The following figure shows the Sea Surface Temperature (SST) for July 1982
  - Information from ~250,000 data points are summarized and can be easily comprehended and interpreted



Representation

- Key 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:
  - Data 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. Otherwise, the relationship can also be explicitly represented by lines.

Arrangement

- Arrangement, i.e., the placement of visual elements within a display, is crucial to good visualization
- It can make a big difference in how easy the data to be understood.
- For example: (what do you observe?)

	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

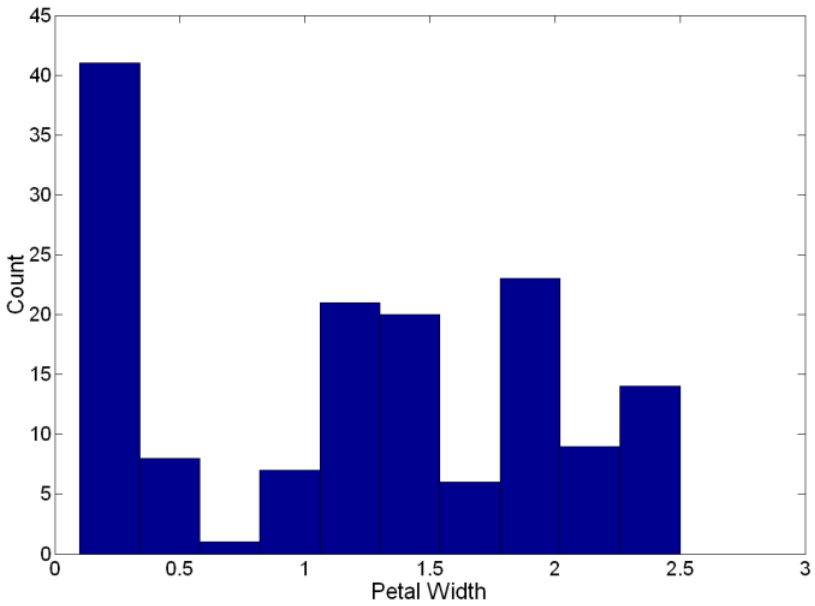
	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

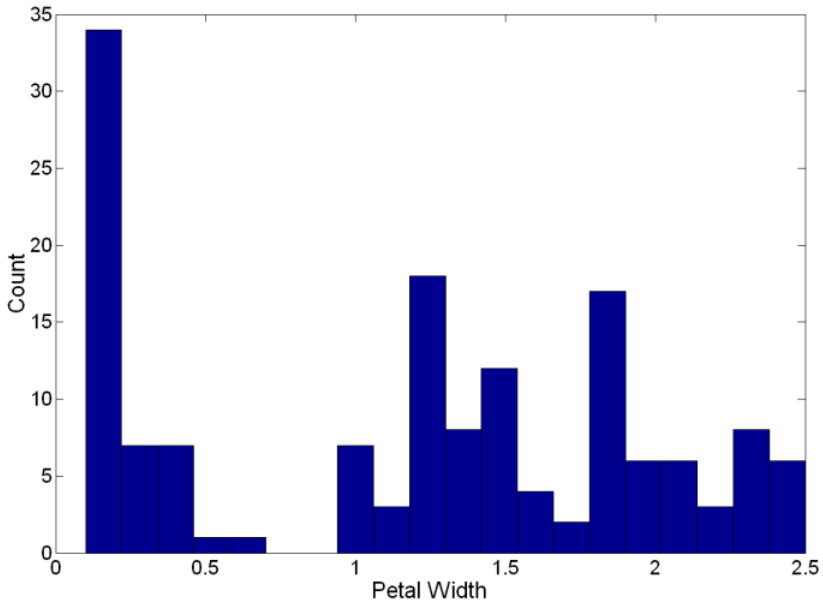
Selection

- To eliminate or de-emphasize certain objects and attributes
- Choosing a subset of attributes for display
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, display pairs of attributes
- 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
- Usually show in one or a series of 2-dimensional plots for easy viewing

Visualization Tech.: Histograms

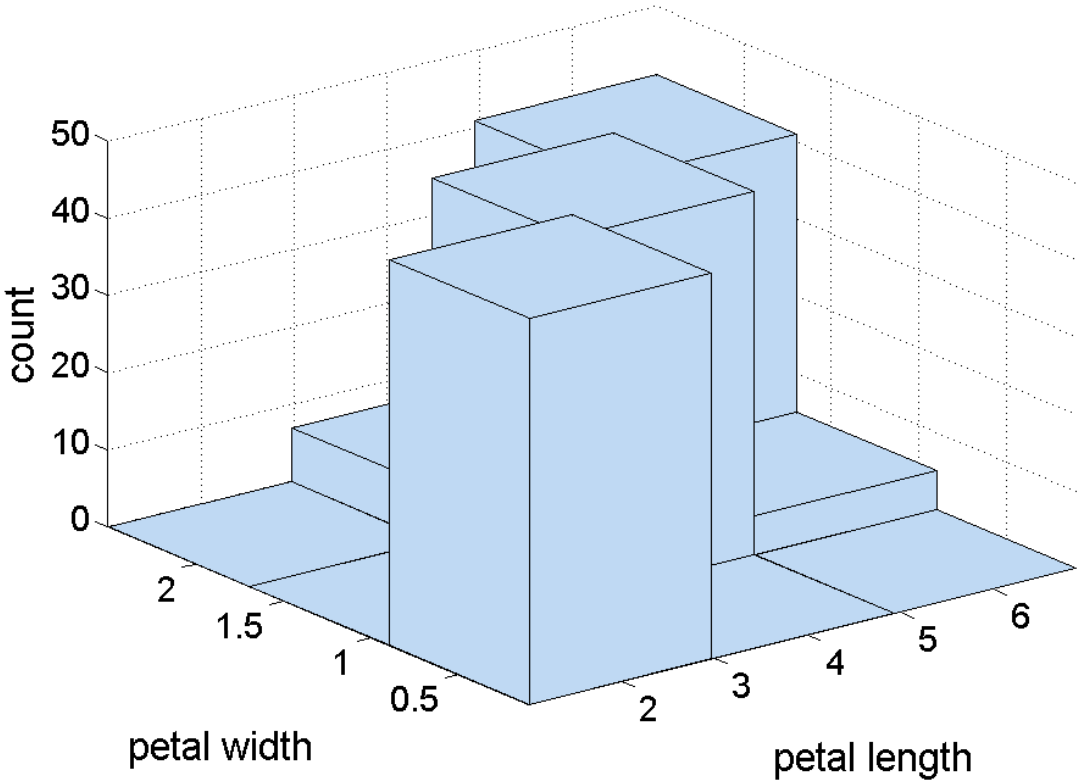
- Shows *distribution of values* in variables
- Divide the values into bins and show a bar plot
- Height of each bar indicates number or frequency of objects
- Shape of histogram depends on number of bins
  - Example: Petal Width (10 and 20 bins, respectively)





Two-Dimensional Histograms

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



Pie Chart

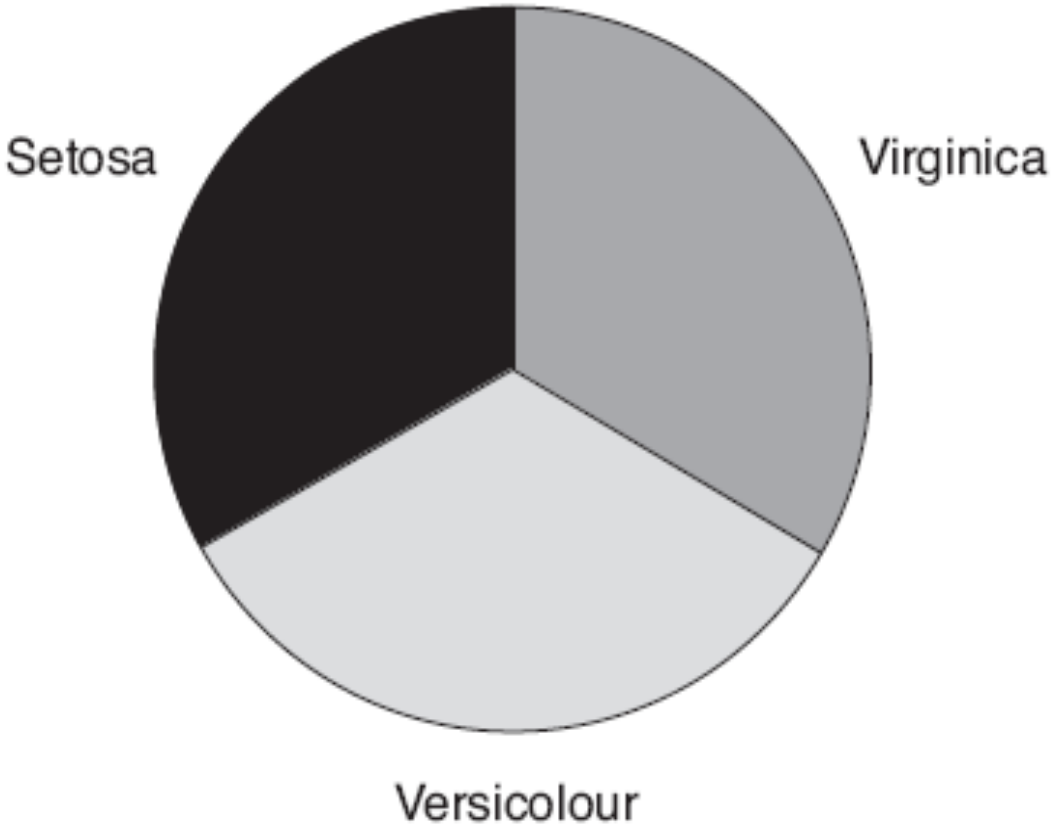
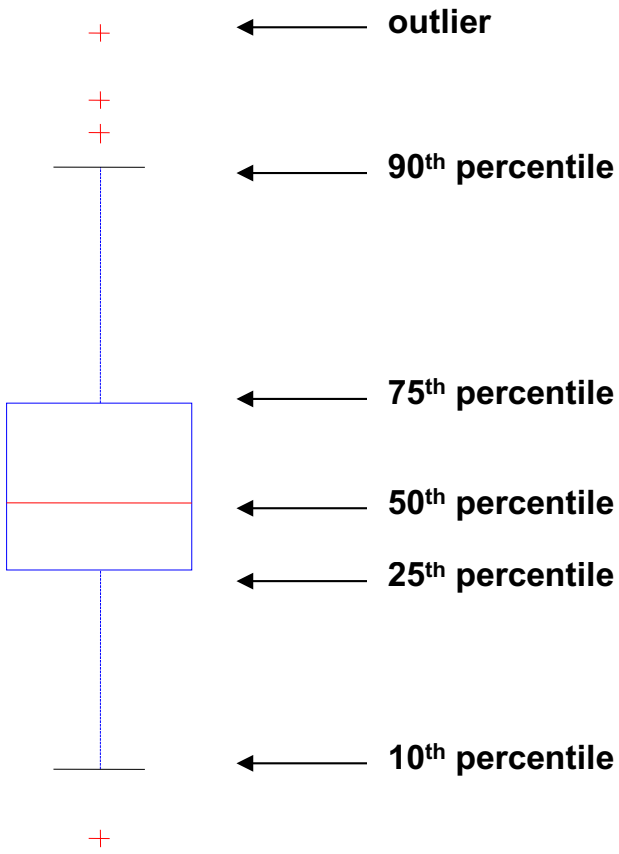


Figure 3.13. Distribution of the types of Iris flowers.

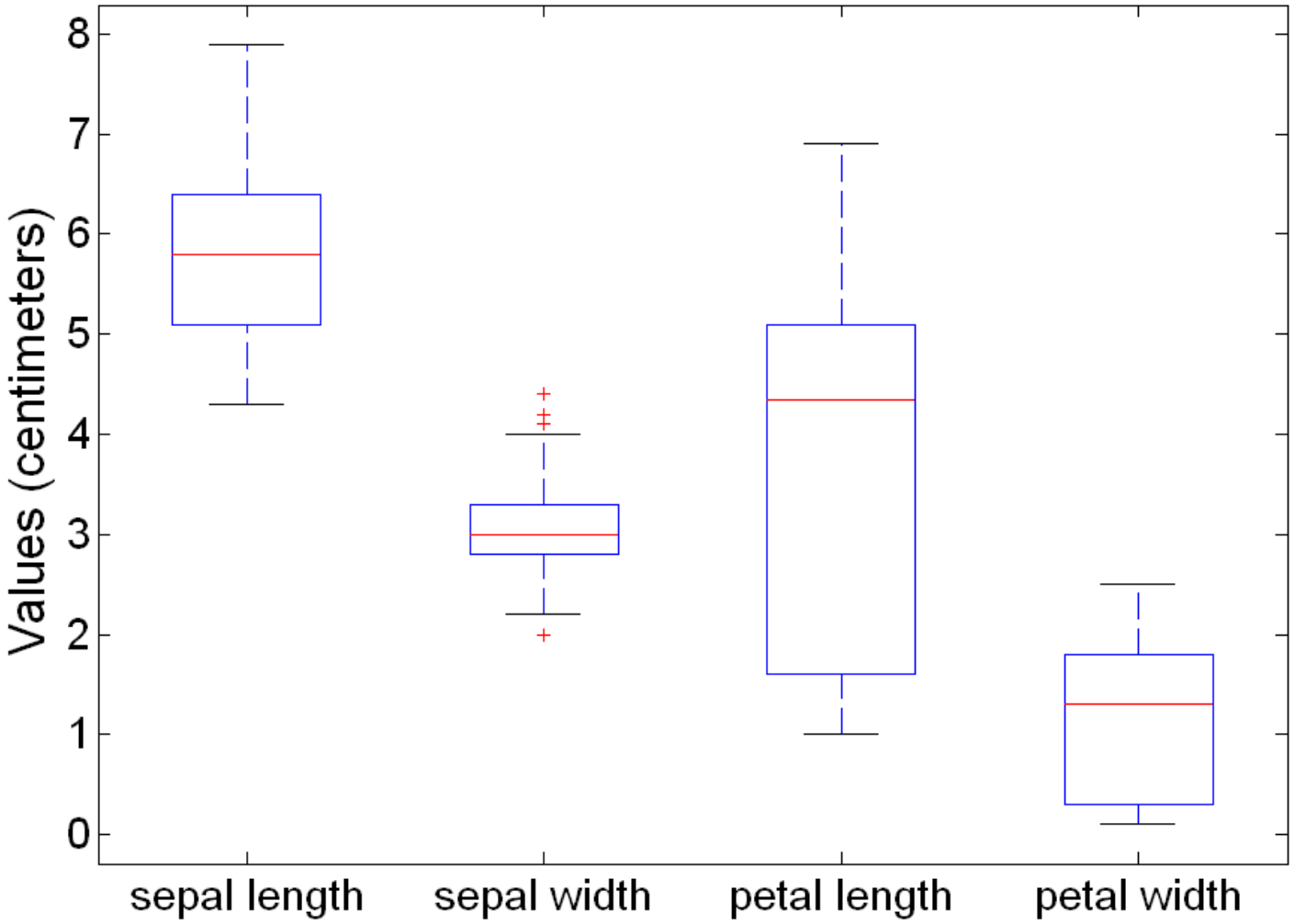
Visualization Tech.: Box Plots

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



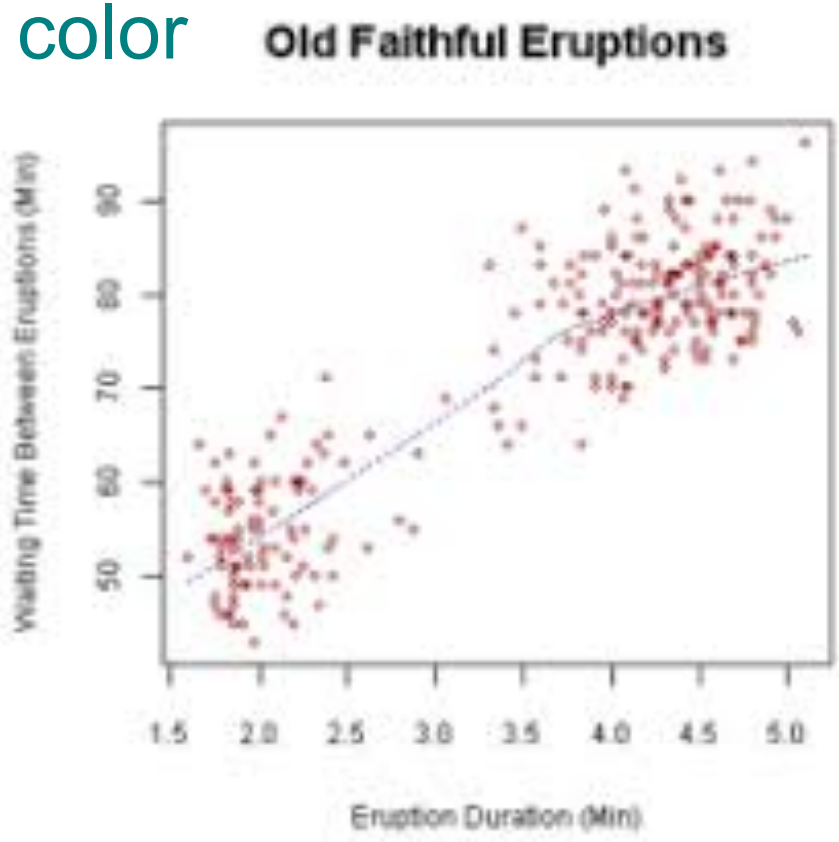
Example of Box Plots

- Box plots can be used to compare attributes

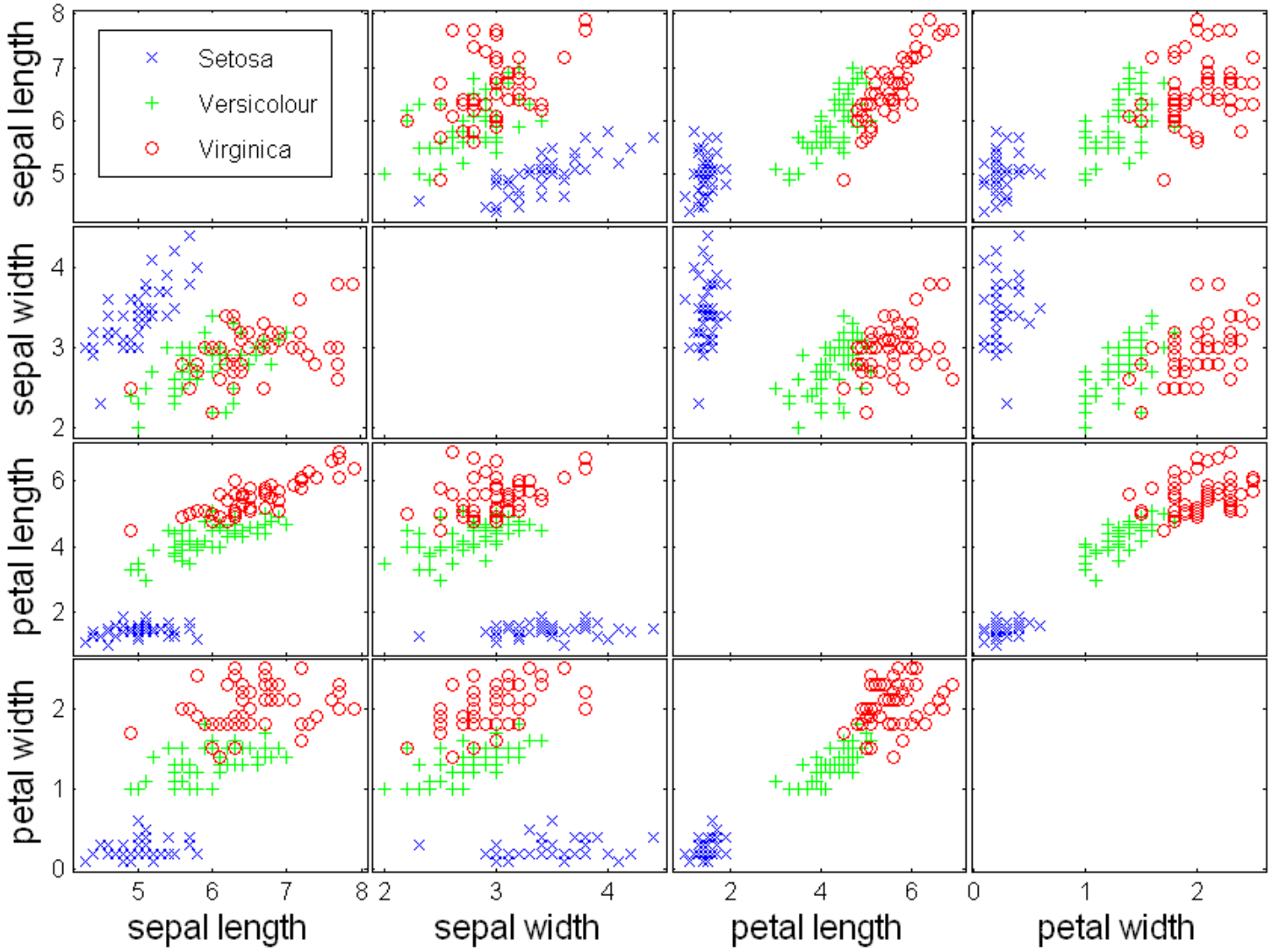


Visualization Tech: Scatter Plots

- 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 to compactly summarize the relationships of several pairs of attributes



Scatter Plot Array of Iris Attributes



3D Scatter Plot

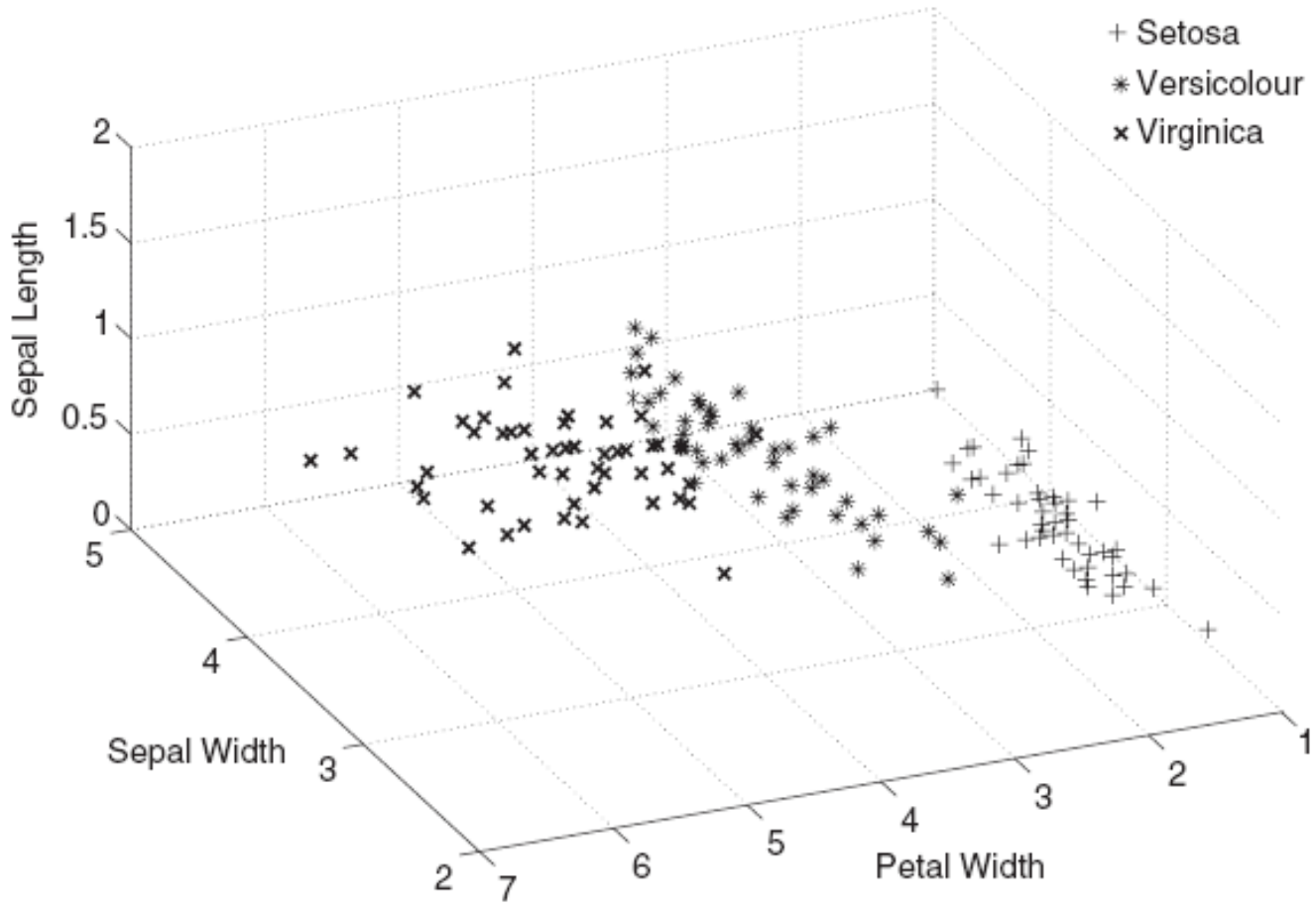
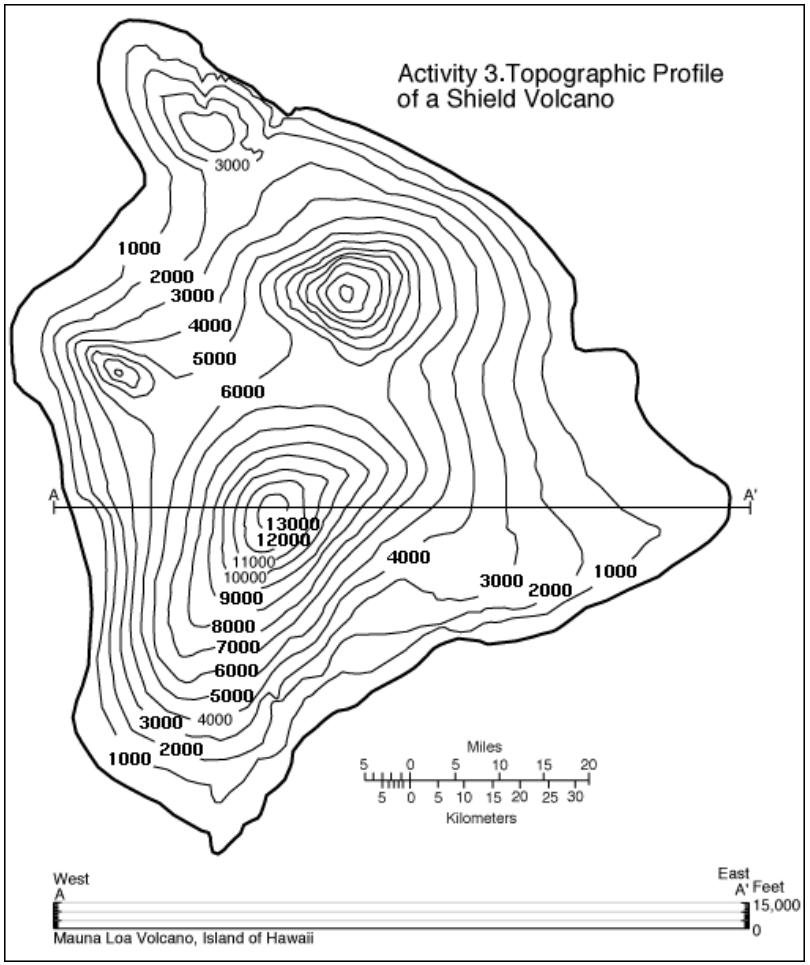


Figure 3.17. Three-dimensional scatter plot of sepal width, sepal length, and petal width.

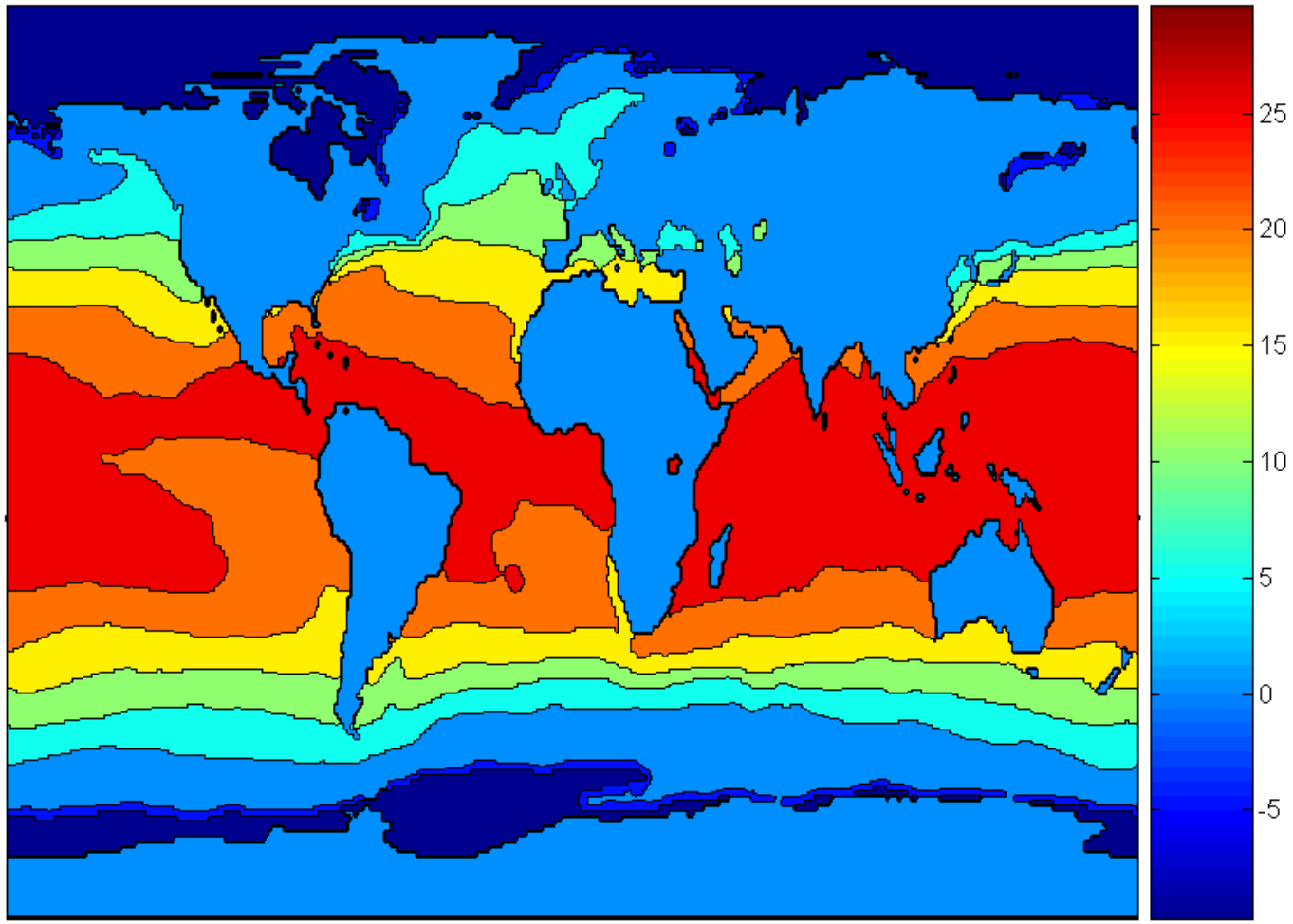
Visualization Tech.: Contour Plots

- Useful when a continuous attribute is measured on a spatial grid
- Partition the plane into regions of similar values
- The contour lines that form the boundaries of these regions connect points with equal values
- A common example is contour maps of elevation



Contour Plot Example

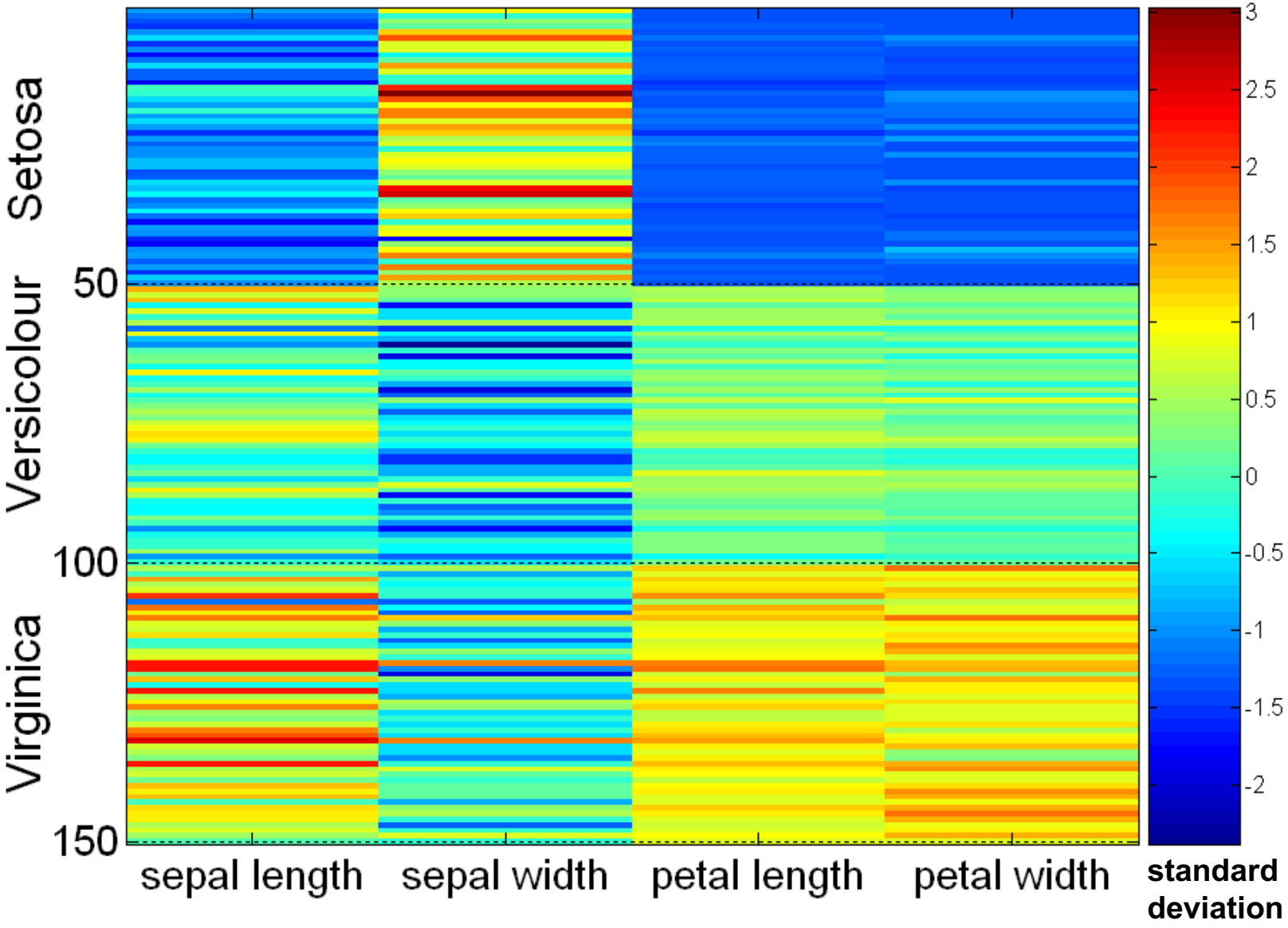
- Can also display temperature, rainfall, air pressure, etc.



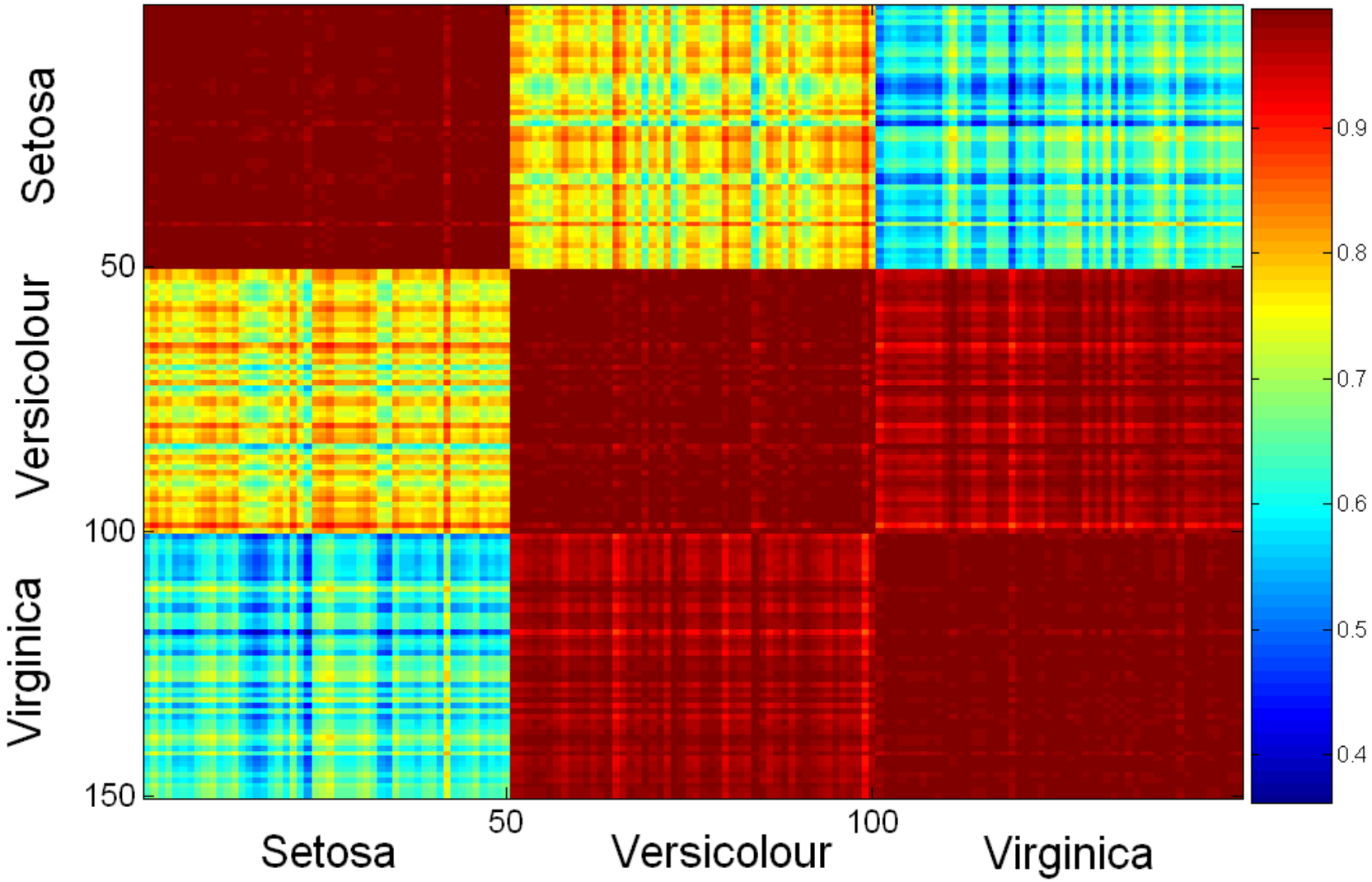
Visualization Tech.: Matrix Plots

- 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 (50 flowers in each species)



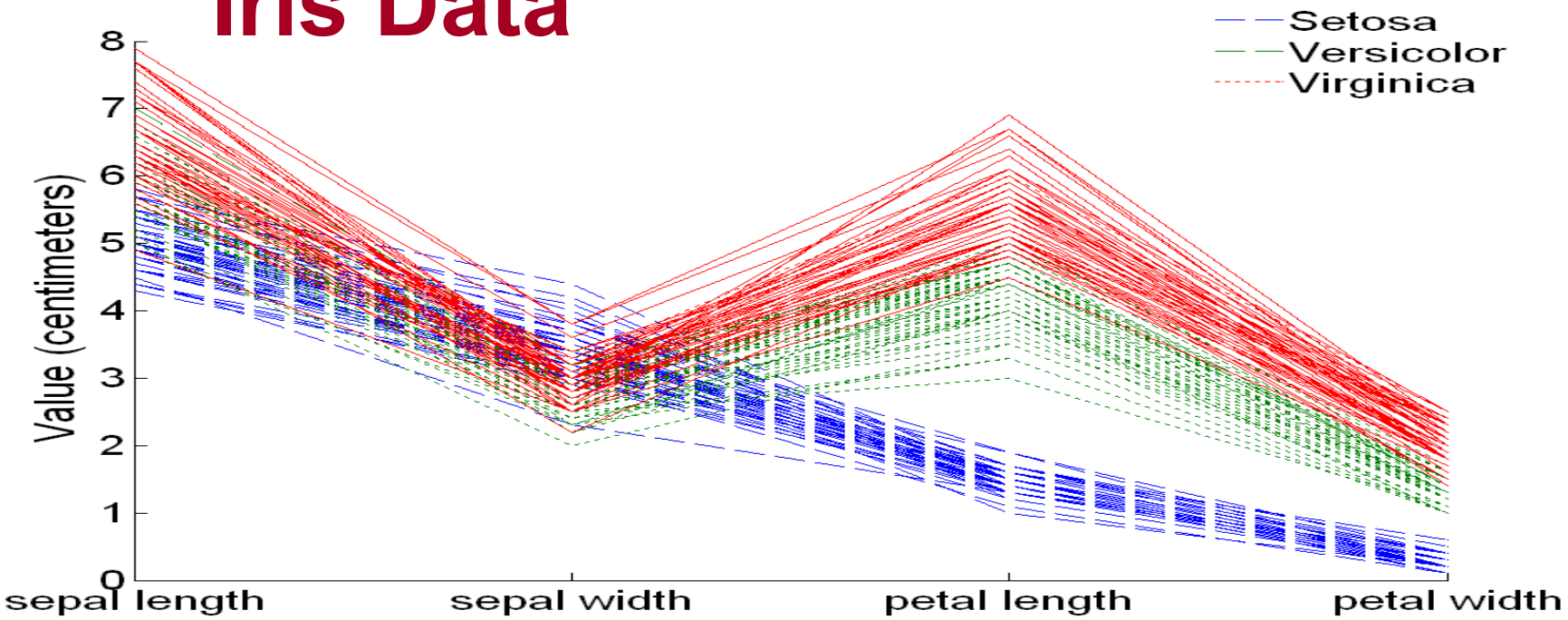
Visualization of the Iris Correlation Matrix (50 flowers in each species)

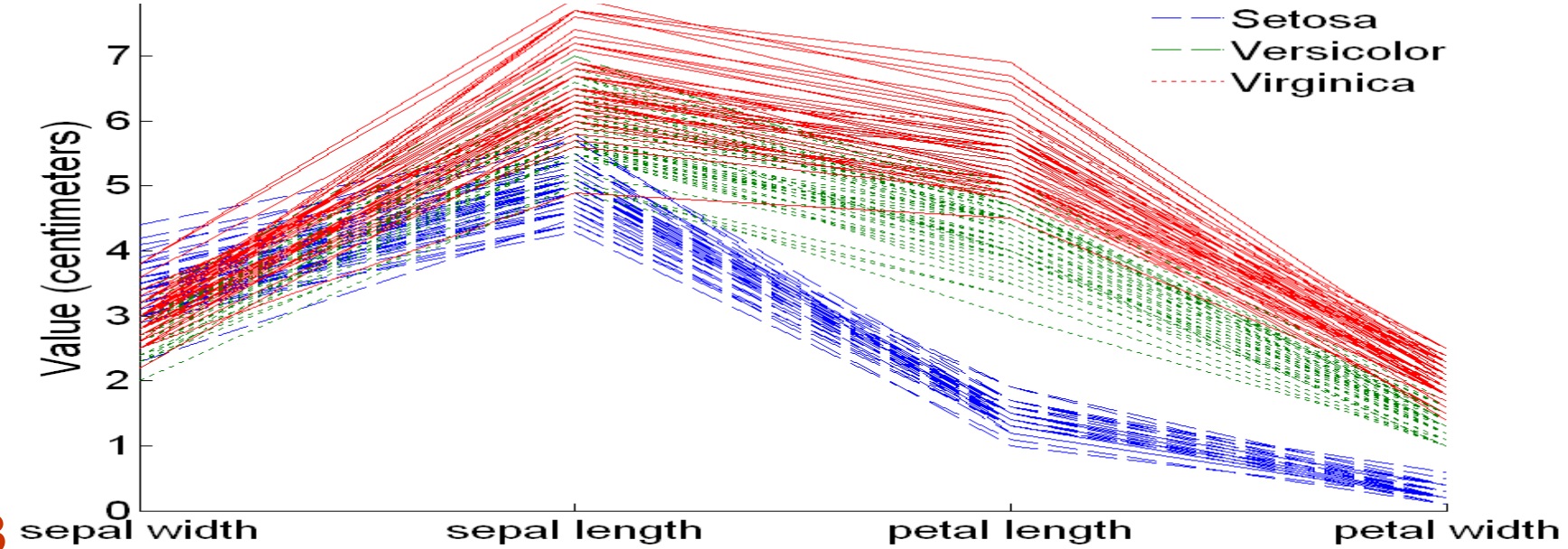


Visualization Techniques: Parallel Coordinates

- 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
- 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



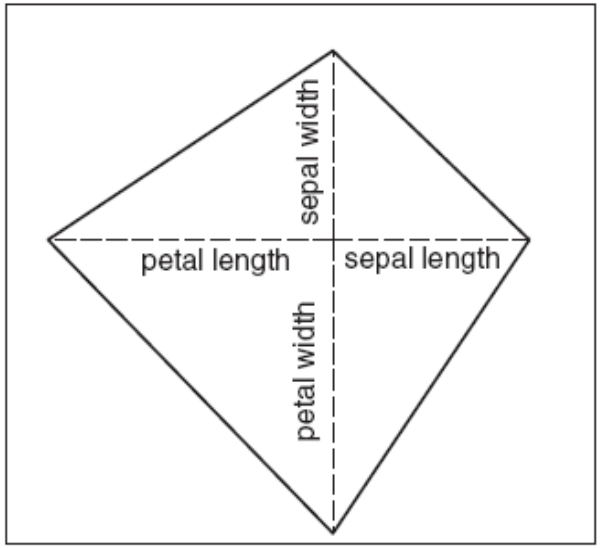


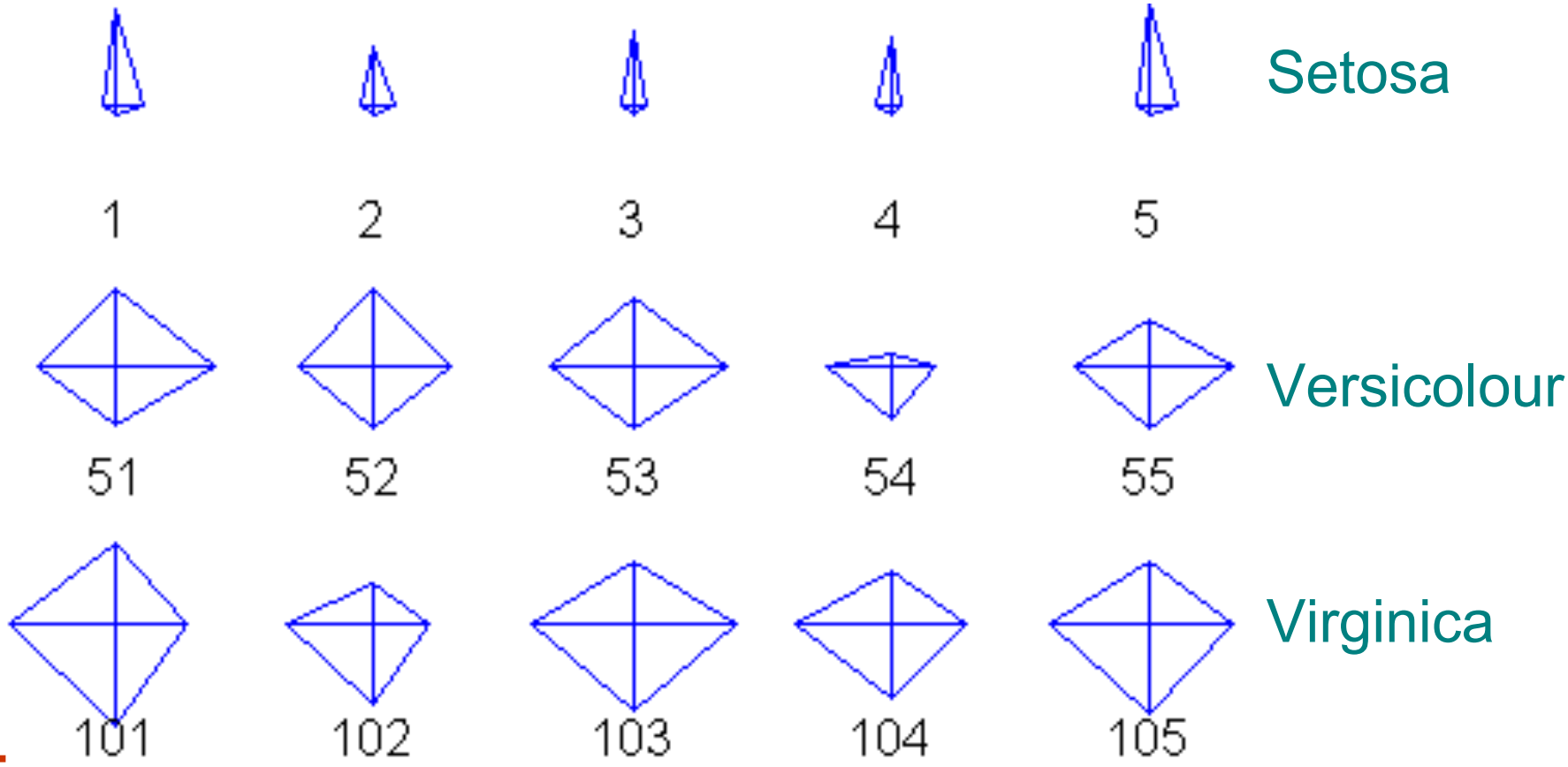
Other Visualization Techniques

- 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

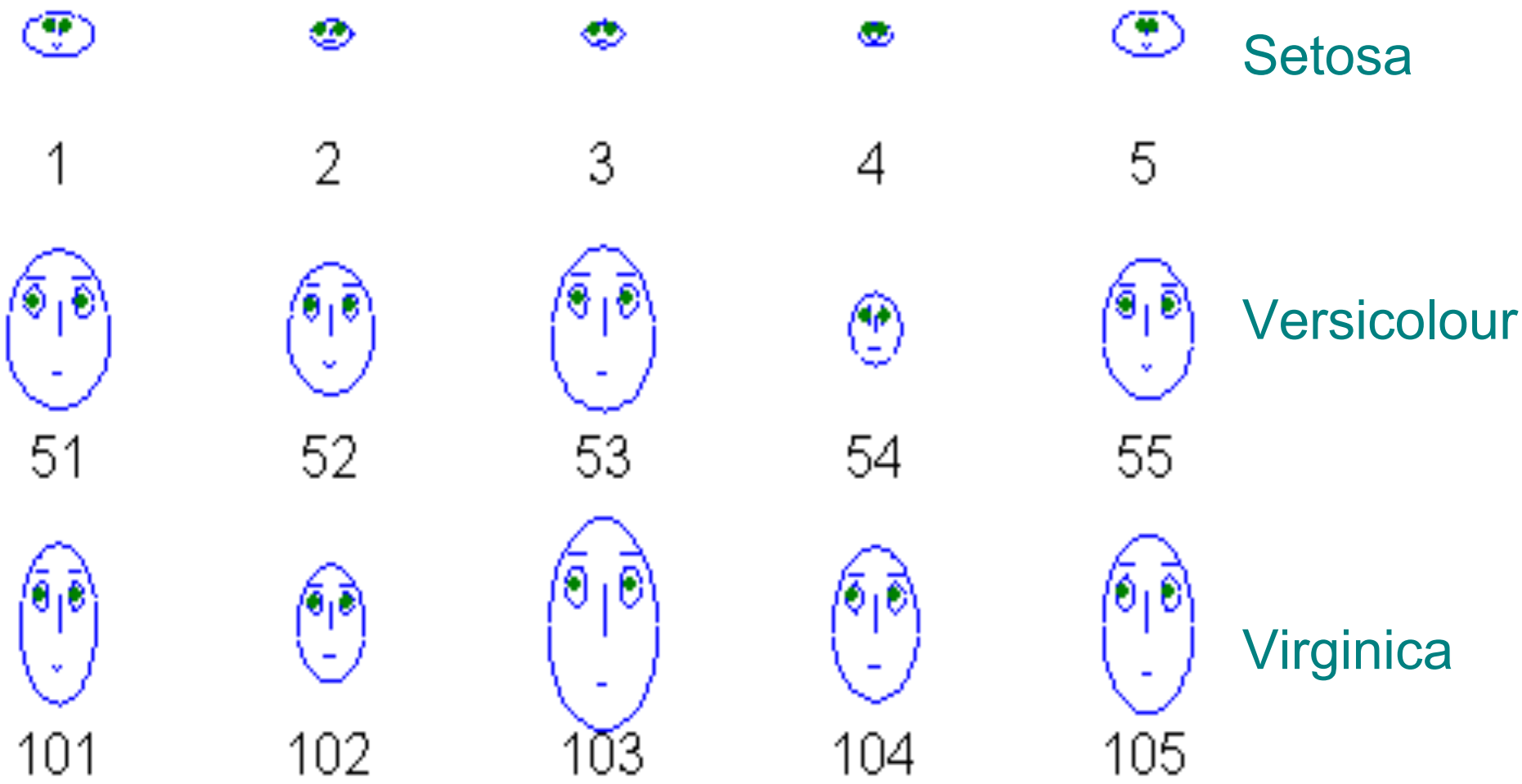
- Each object is mapped into a polygon
- Attributes normalized to [0,1]





Chernoff Faces for Iris Data

- Each attribute is associated with a feature of face



OLAP

- On-Line Analytical Processing (OLAP) was proposed by E. F. Codd, the father of the relational database.
- Relational databases put data into tables, while OLAP uses a *multidimensional array* representation.
  - Such representations of data previously existed in statistics and other fields
- There are a number of data analysis and data exploration operations that are easier with such a data representation.

Creating a Multidimensional Array

- Two key steps in converting tabular data into a multidimensional array.
  - First, identify which attributes are to be the *dimensions* and which attribute is to be the target attribute whose values appear as *entries* in the multidimensional array.
    - The attributes used as dimensions must have discrete values
    - The target value is typically a count or continuous value, e.g., the cost of an item
    - Can have no target variable at all except the count of objects that have the same set of attribute values
  - Second, find the value of each entry in the multidimensional array by summing the values (of the target attribute) or count of all objects that have the attribute values corresponding to that entry.

Example: Iris data

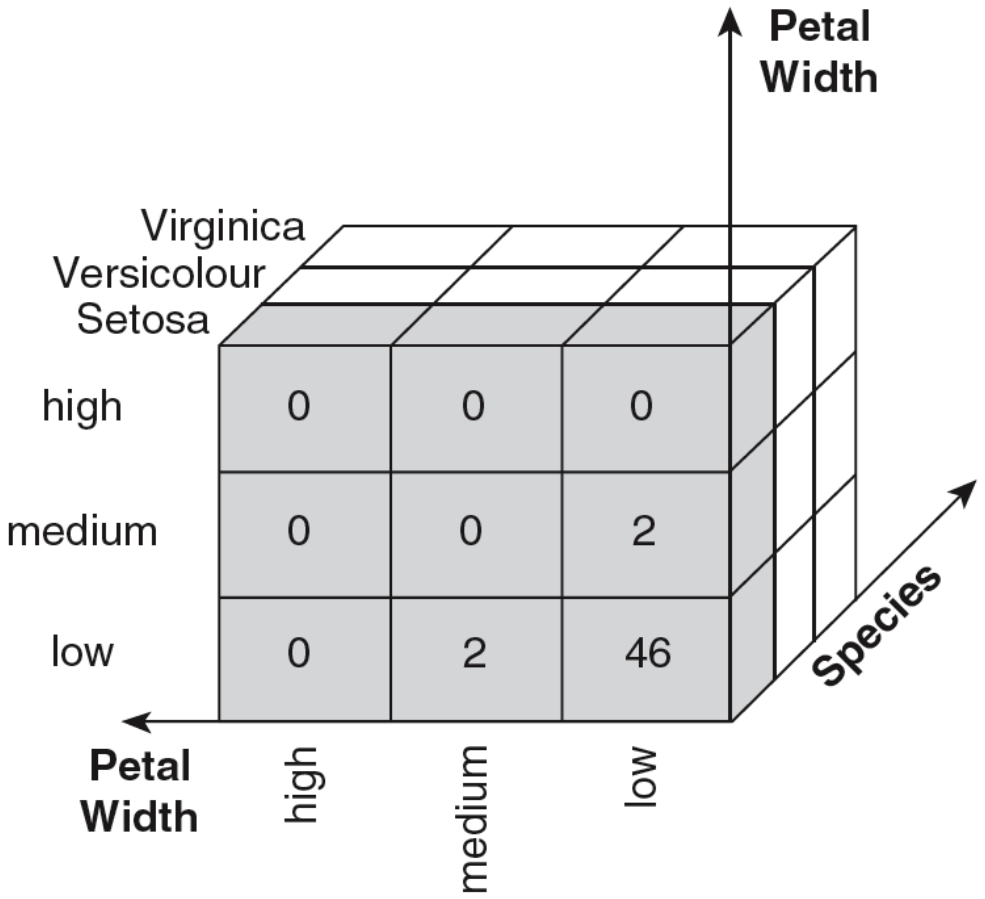
- We show how the attributes, petal length, petal width, and species type can be converted to a multidimensional array
  - First, we discretized the petal width and length to have categorical (discrete) values: *low*, *medium*, and *high*
  - We get the following table - note the count attribute

Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Target

Example: Iris data (continued)

- Each unique tuple of petal width, petal length, and species type identifies one element of the array.
- This element is assigned the corresponding count value.
- The figure illustrates the result.
- All non-specified tuples are 0.



Example: Iris data (continued)

- Slices of the multidimensional array are shown by the following cross-tabulations
- What do these tables tell us?

Setosa

Length \ Width	low	medium	high
low	46	2	0
medium	2	0	0
high	0	0	0

Versicolour

Length \ Width	low	medium	high
low	0	0	0
medium	0	43	3
high	0	2	2

Virginica

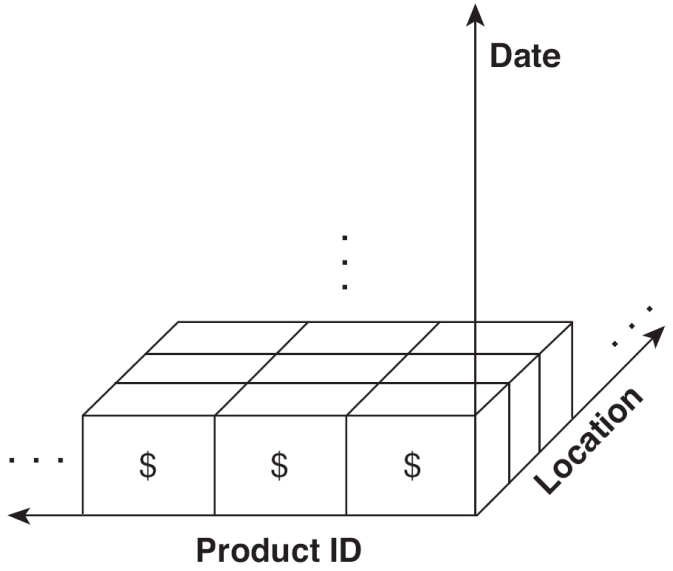
Length \ Width	low	medium	high
low	0	0	0
medium	0	0	3
high	0	3	44

OLAP Operations: Data Cube

- The key operation of a OLAP is the formation of a data cube
- A data cube is a *multidimensional representation of data, together with all possible aggregates*.
- By all possible aggregates, we mean the aggregates that result by selecting a proper subset of the dimensions and summing over all remaining dimensions.
- For example, if we choose the species type dimension of the Iris data and sum over all other dimensions, the result will be a one-dimensional entry with three entries, each of which gives the number of flowers of each type.

Data Cube Example

- Consider a data set that records the sales of products at a number of company stores at various dates.
- This data can be represented as a 3 dimensional array
- There are 3 two-dimensional aggregates (3 choose 2), 3 one-dimensional aggregates and 1 zero-dimensional aggregate (the overall total)



Data Cube Example (continued)

- The following figure table shows one of the two dimensional aggregates, i.e., (Product ID, Date), along with two of the one-dimensional aggregates, i.e., Product ID and Date, and one three-dimensional aggregate, i.e., overall total

product ID \ date	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127