OntarioTech Science

Kourosh Davoudi
kourosh@uoit.ca

Week 2: Data Exploration

CSCI 4150U: Data Mining

Data Mining: Data Exploration

Outline

- Motivation of Data Exploration
- Summary Statistics
- Visualization Techniques
- Data Warehouse and OLAP

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/SEMATECH e-Handbook of Statistical Methods 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
  - Online Analytical Processing (OLAP)

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.

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 one/two passes through the data

A. True
B. False

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 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 50'th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.

What is the $x_{30\%}$ in the following data?

7, 2, 8, 16, 9, 12, 6, 17, 21, 12

A. 7

B. 8

C. 9

D. 12

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.

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

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

Measures of Spread: Range and Variance

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

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

- Because of outliers, other measures are often used.

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

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

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

What is the Median Absolute Deviation in the following data?

7, 2, 8, 16, 9, 12, 6, 17, 21, 12

A. 3

B. 3.5

C. 4

D. 4.5

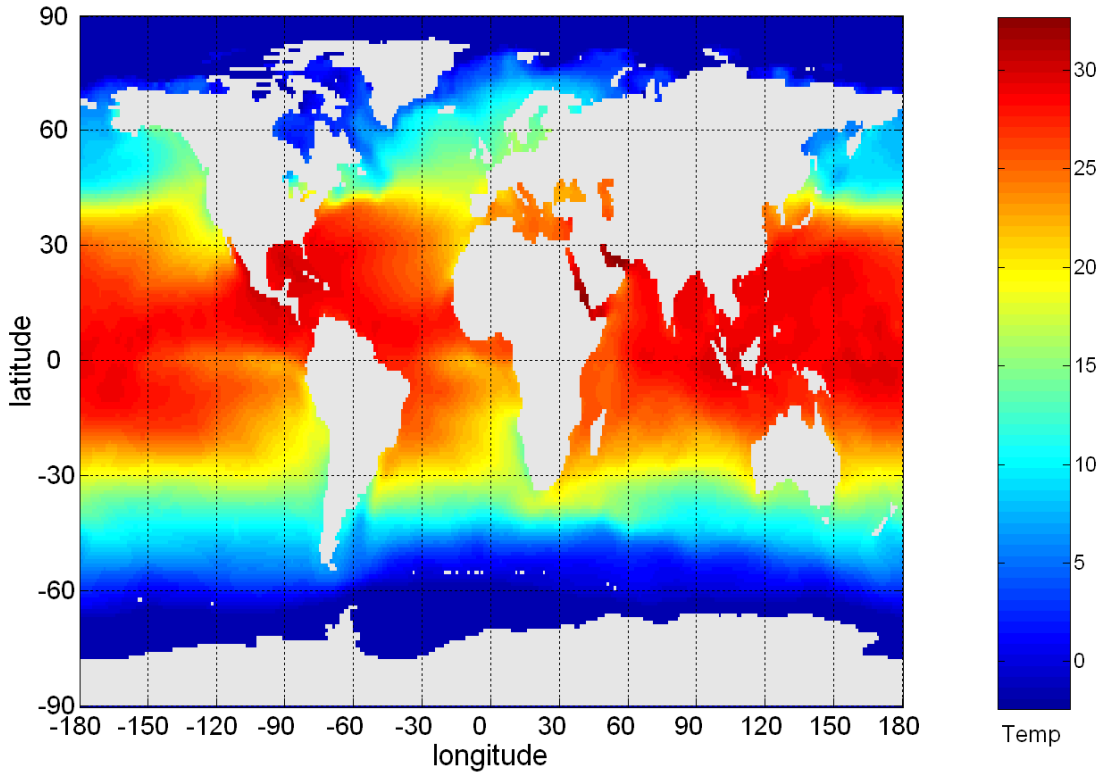
Visualization

- 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
  - 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

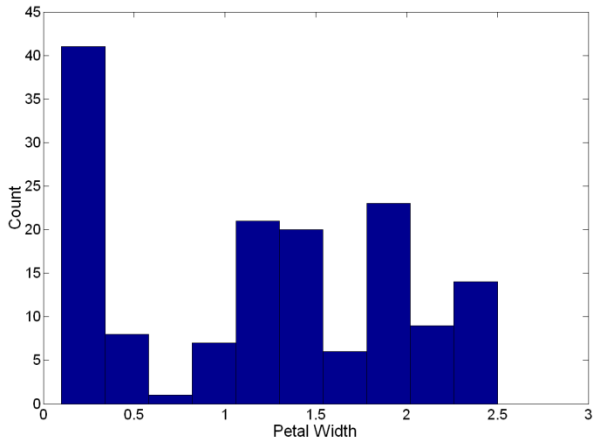
	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

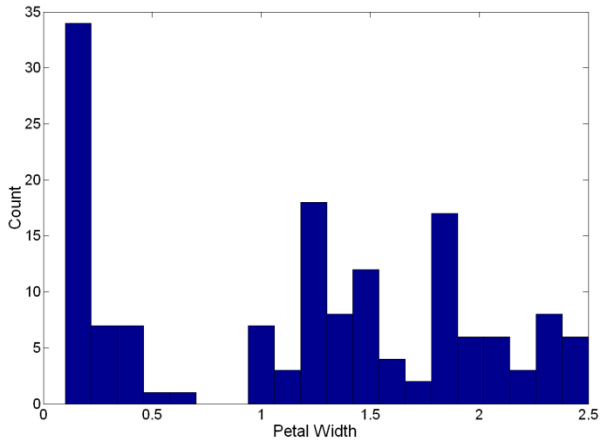
Selection for Visualization

- Is the elimination or the de-emphasis of certain objects and attributes
- Selection may involve the choosing 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

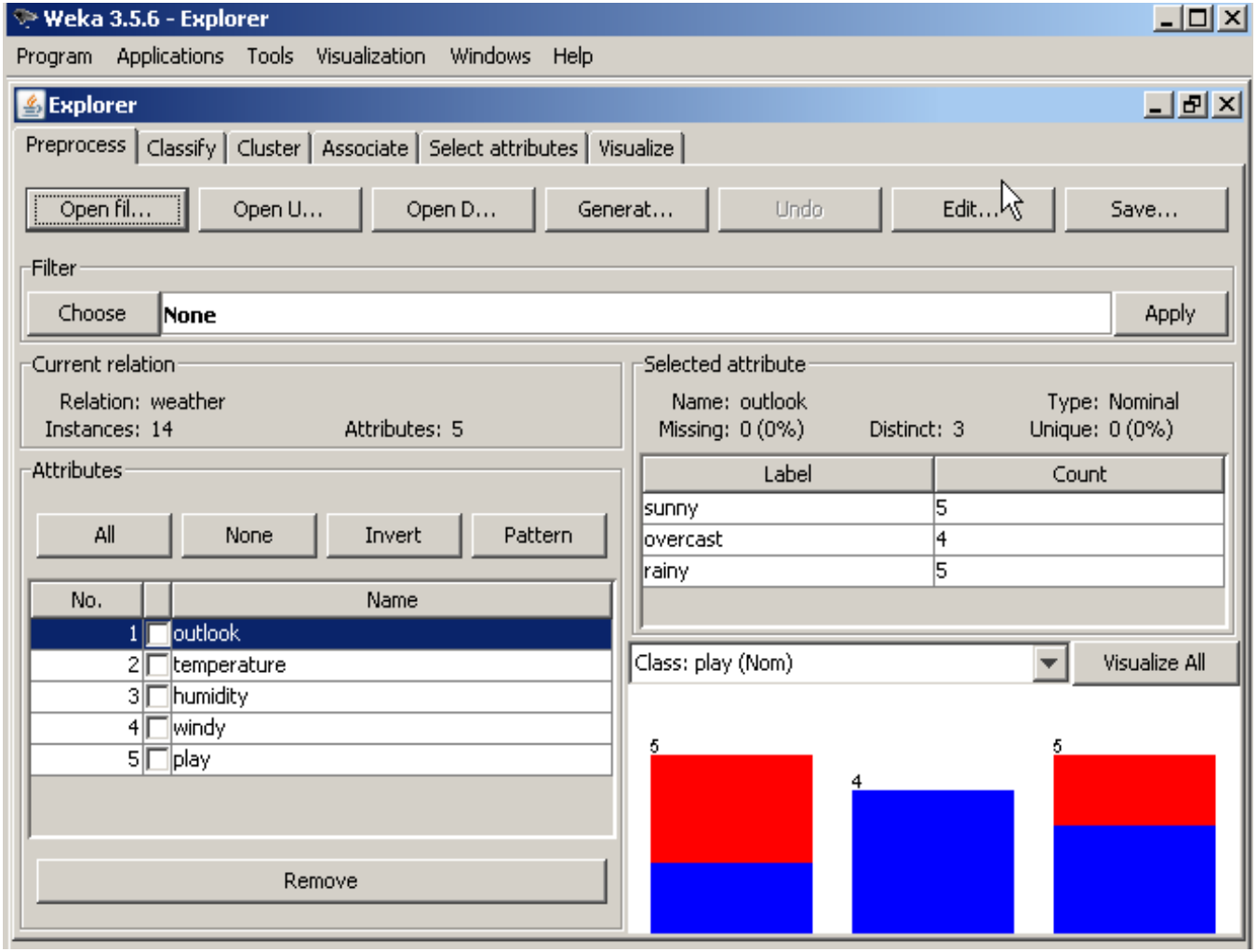
Visualization Techniques: Histograms

- 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)





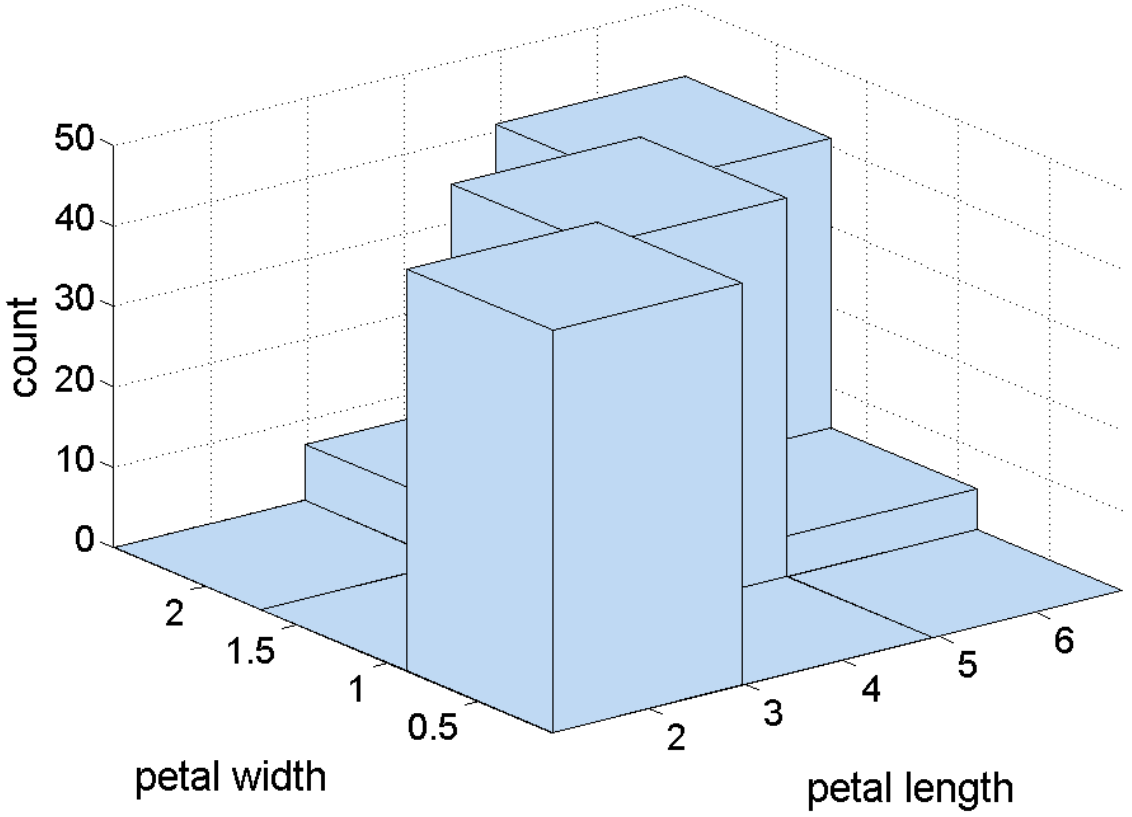
Histogram from Weka



https://www.cs.waikato.ac.nz/ml/weka/

Two-Dimensional Histograms

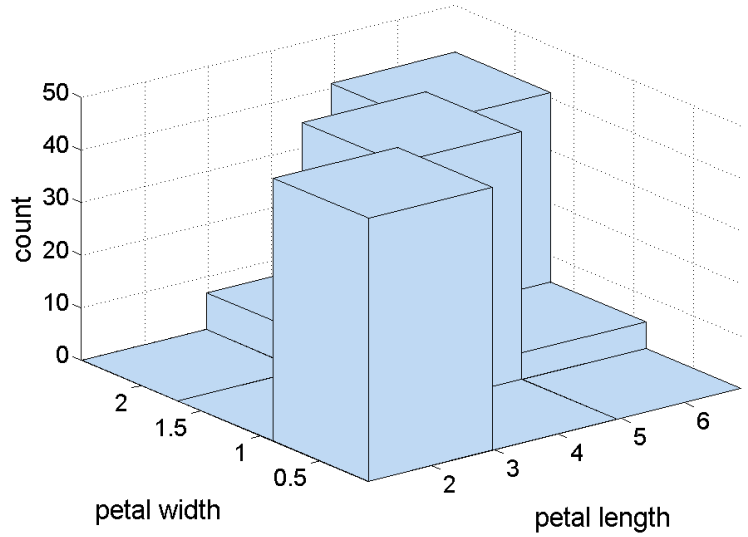
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length



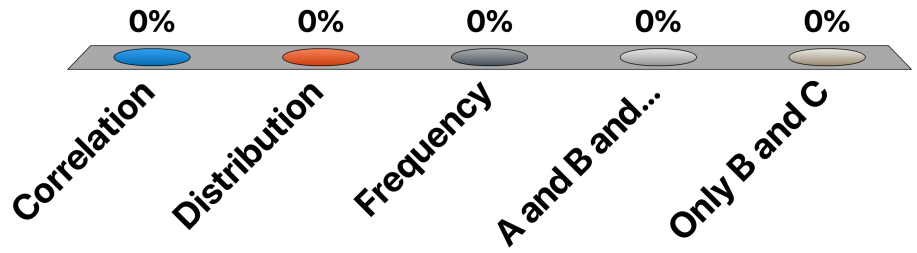
The key issue for three dimensional plots is how to display information so that as little information is obscured as possible.

A. True
B. False

What information this graph gives you?

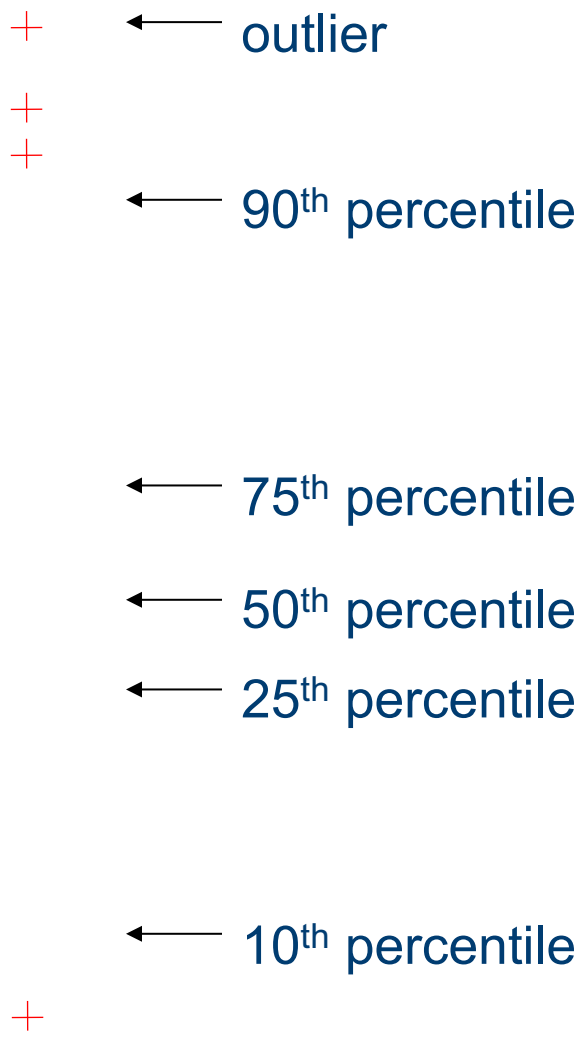


A. Correlation
B. Distribution
C. Frequency
D. A and B and C
E. Only B and C



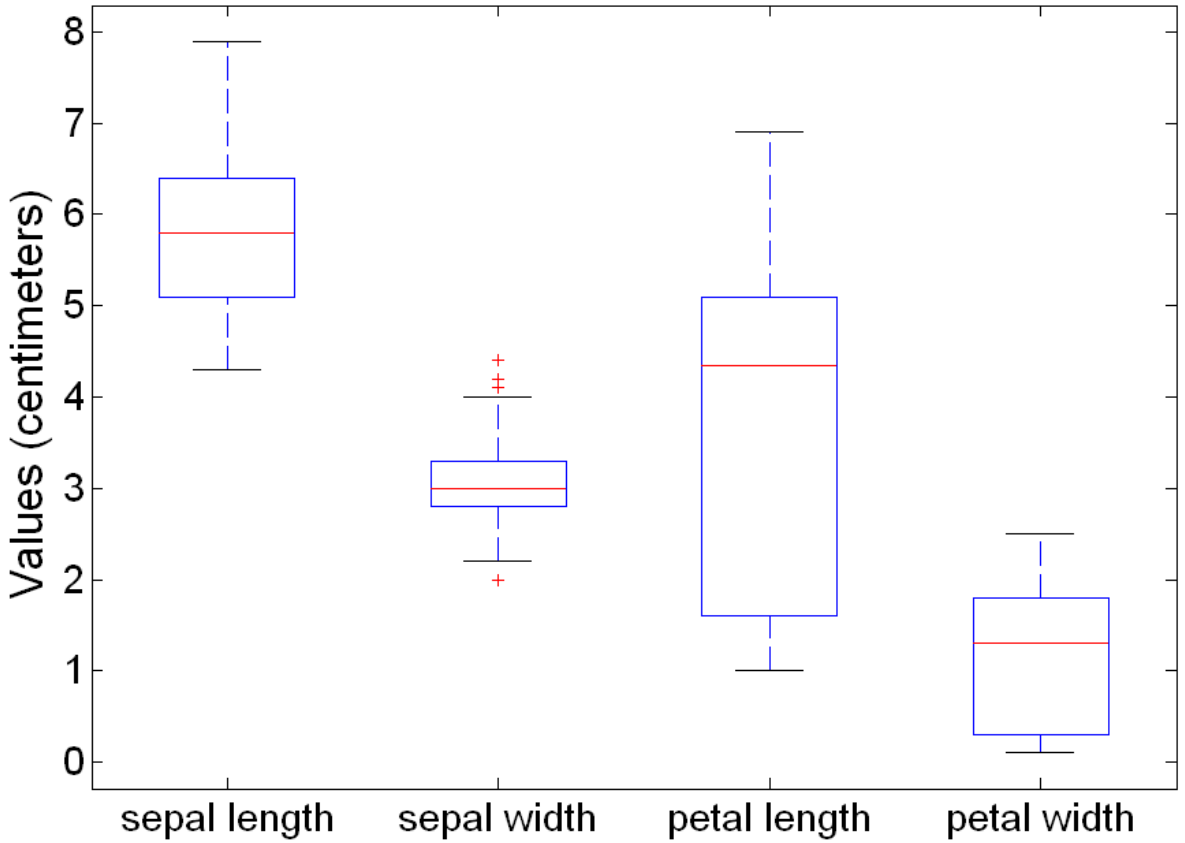
Visualization Techniques: Box Plots

- 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

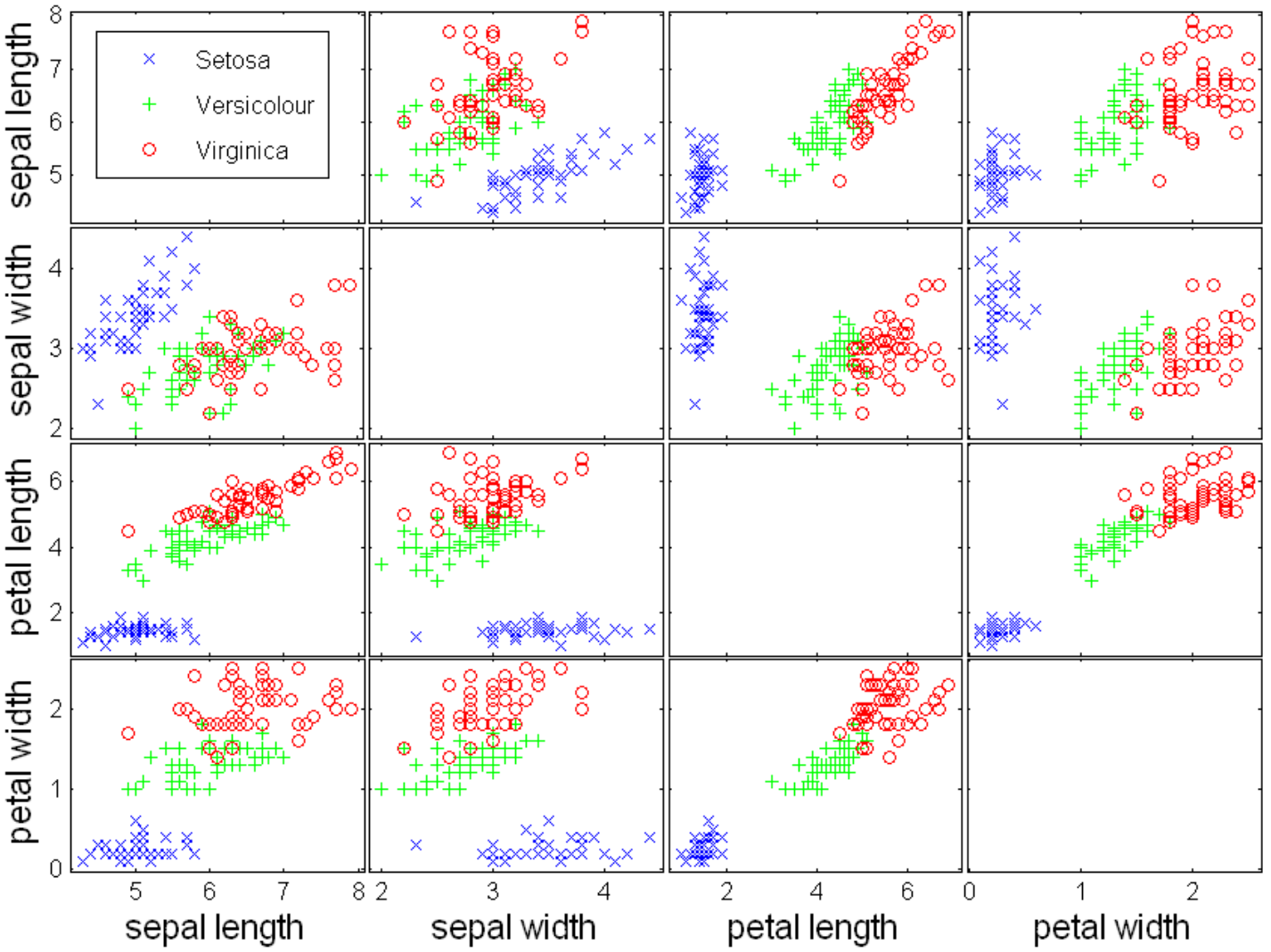
- Box plots can be used to compare attributes



Visualization Techniques: Scatter Plots

- 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

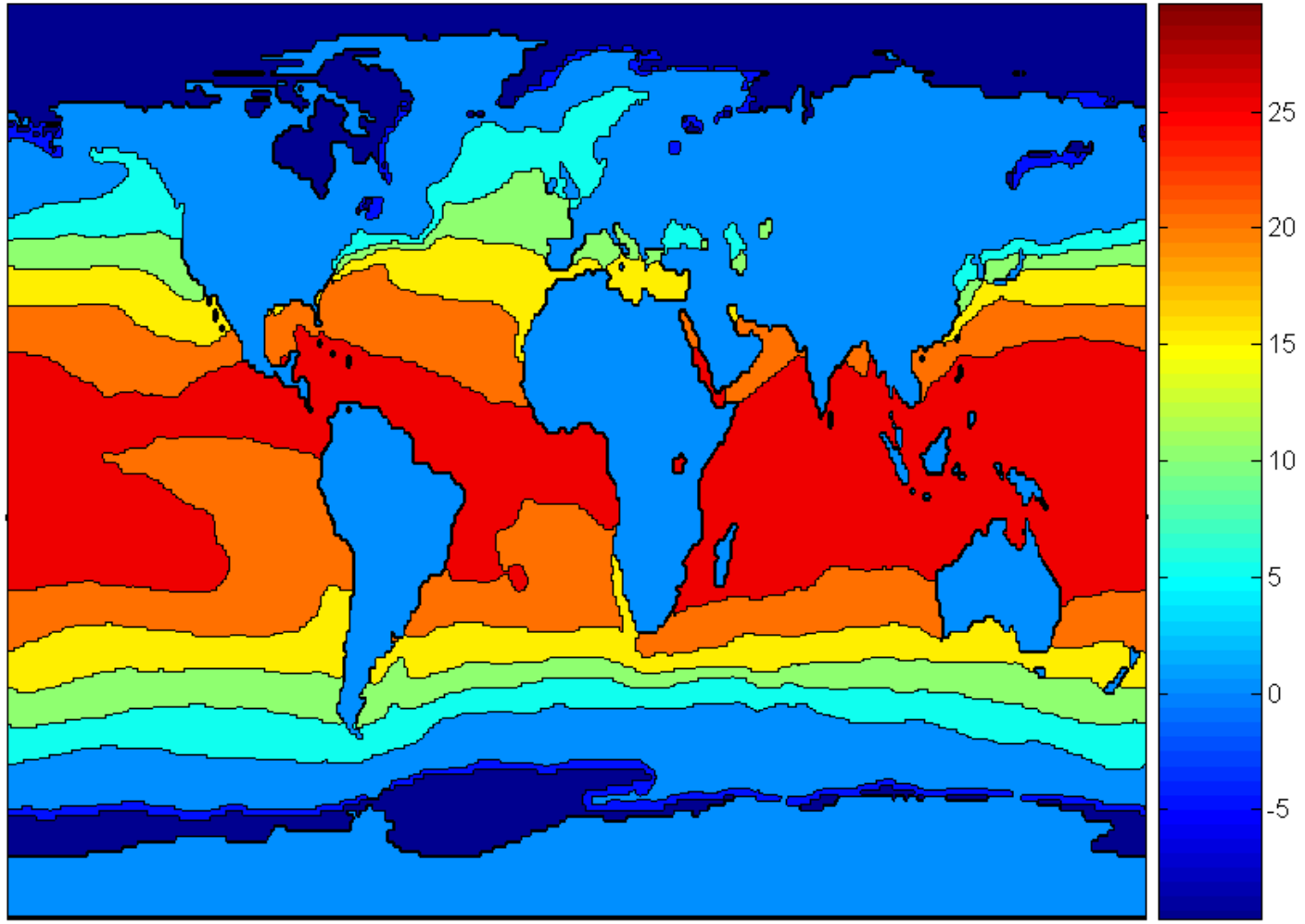
Scatter Plot Array of Iris Attributes



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

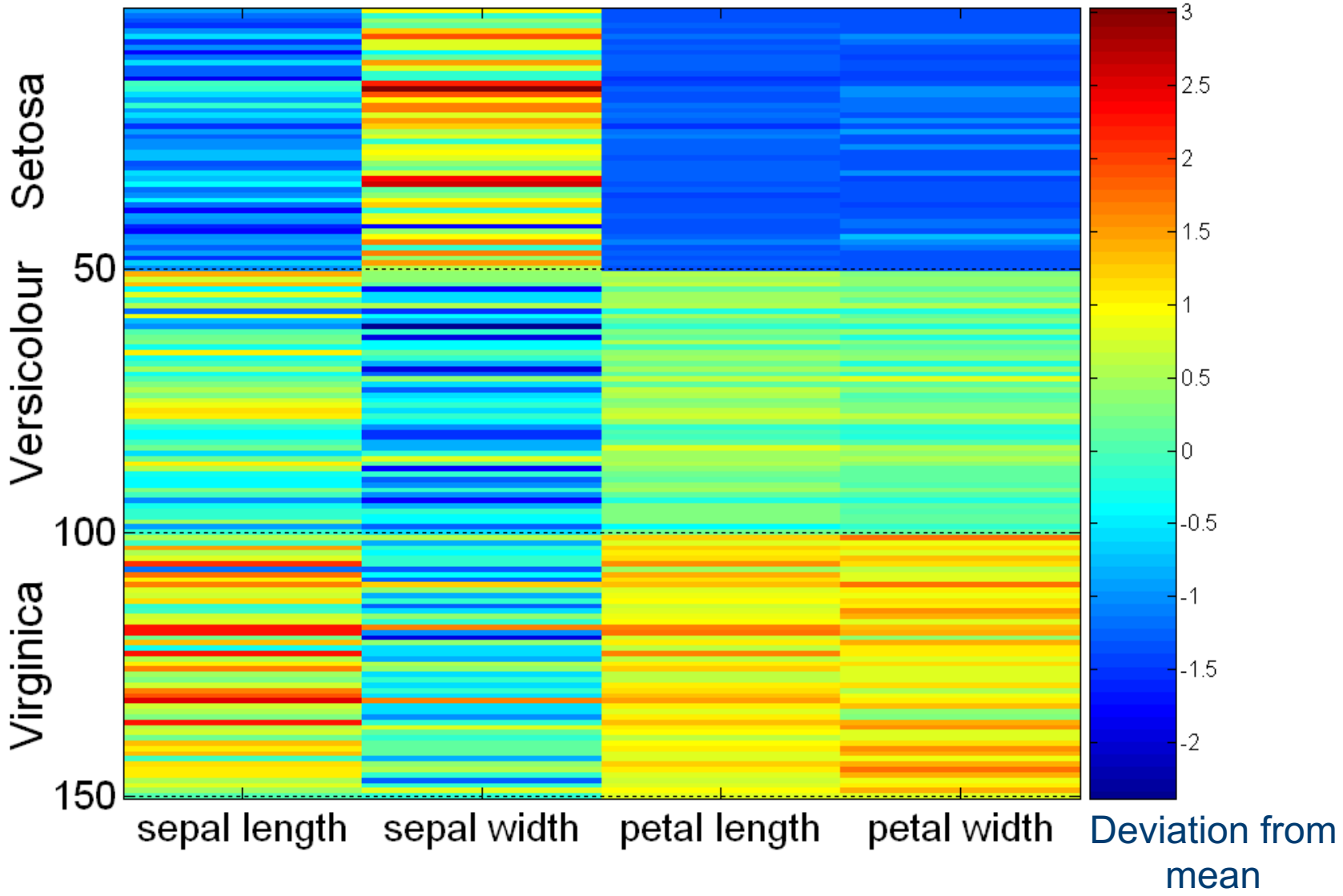
Contour Plot Example: SST Dec, 1998



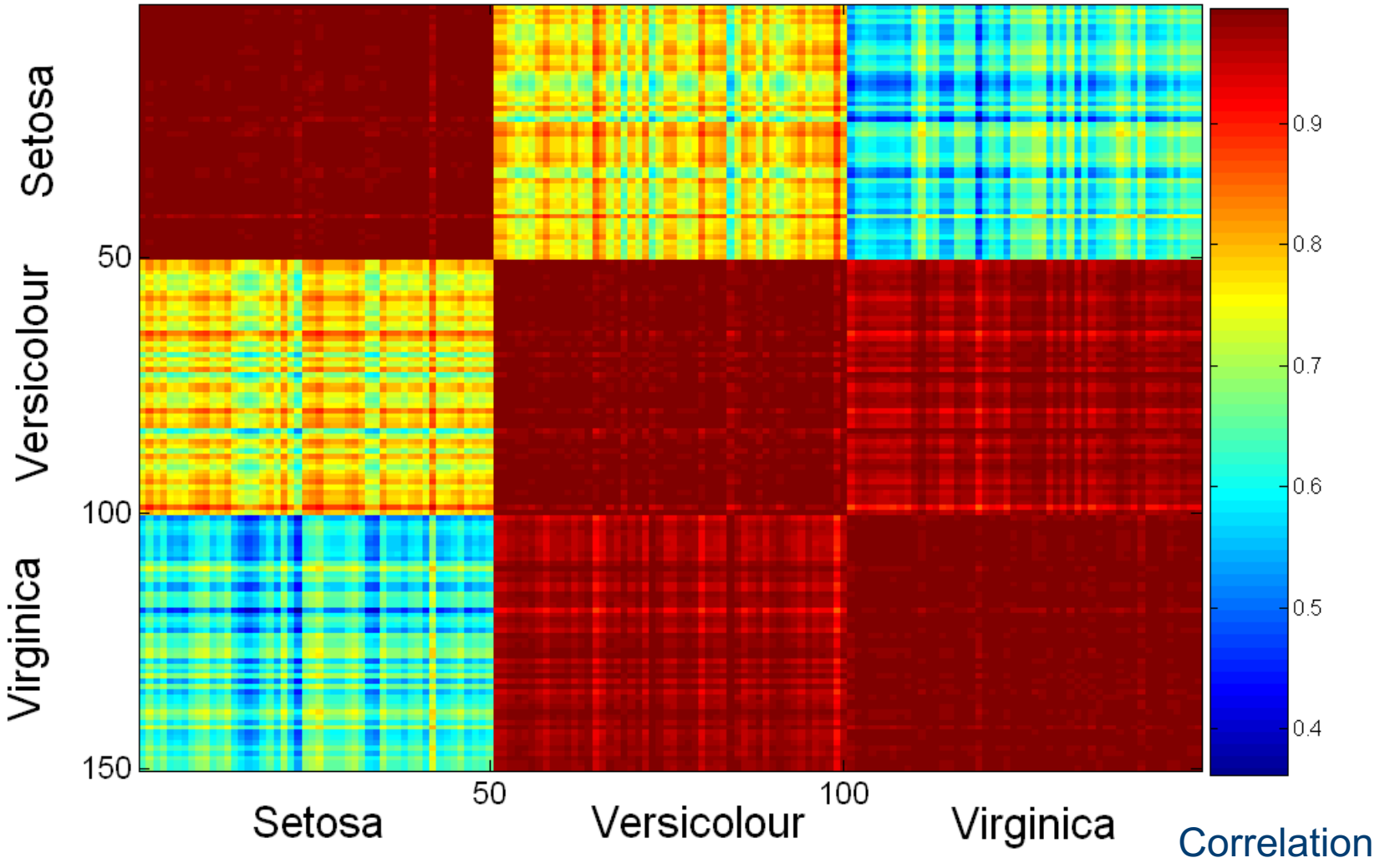
Visualization Techniques: Matrix Plots

- 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
  - Examples of matrix plots are presented on the next two slides

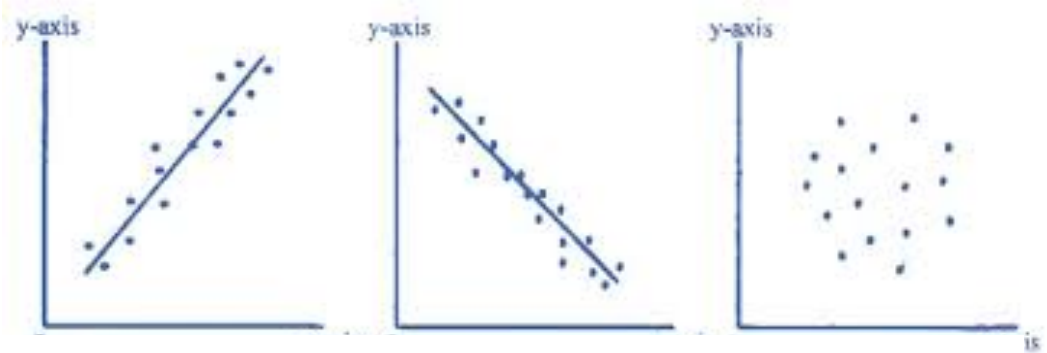
Visualization of the Iris Data Matrix



Visualization of the Iris Correlation Matrix



Which one figure demonstrates negative correlation?



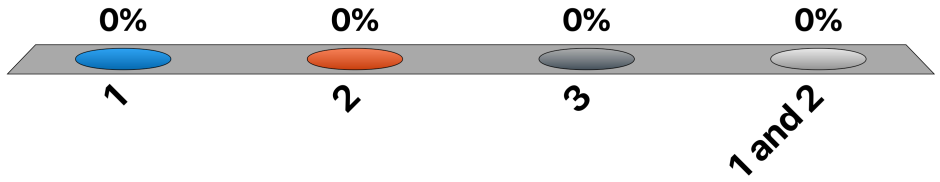
(1) (2) (3)

A. 1

B. 2

C. 3

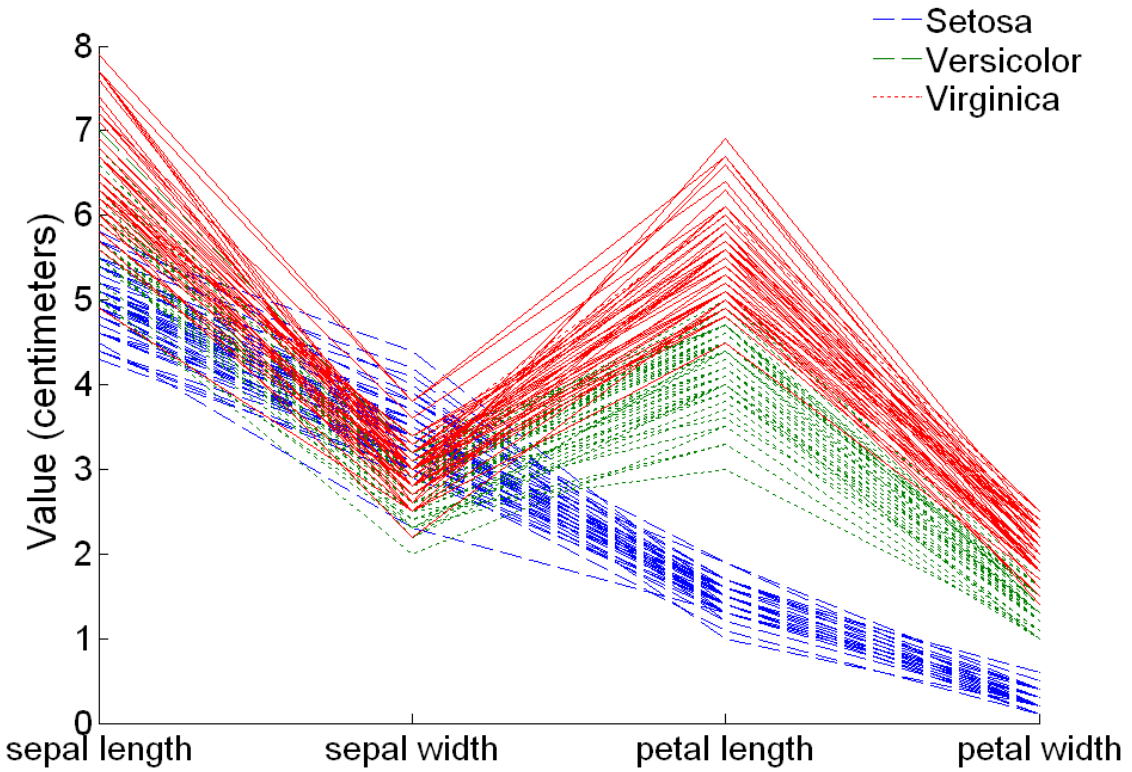
D. 1 and 2

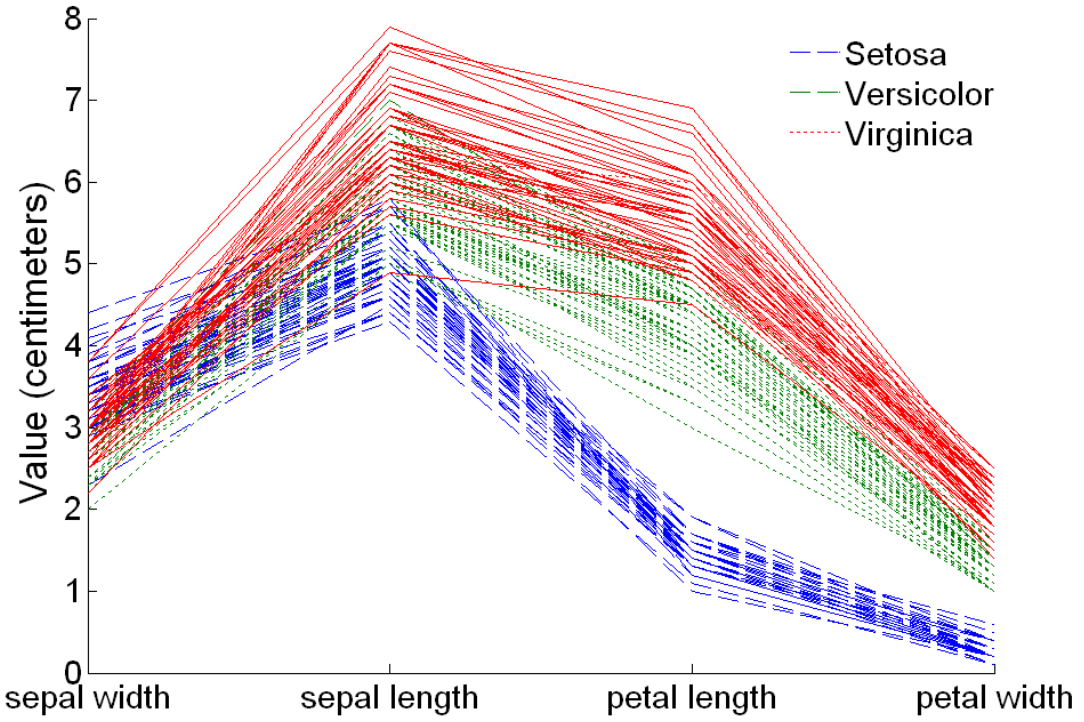


Visualization Techniques: Parallel Coordinates

- 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
  - Ordering of attributes is important in seeing such groupings

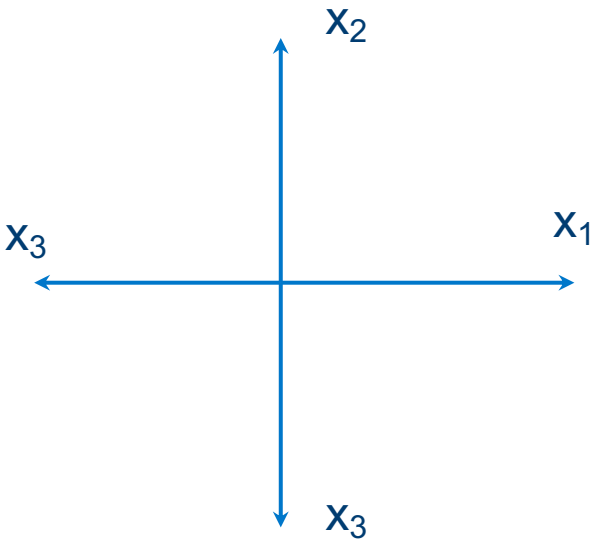
Parallel Coordinates Plots for Iris Data

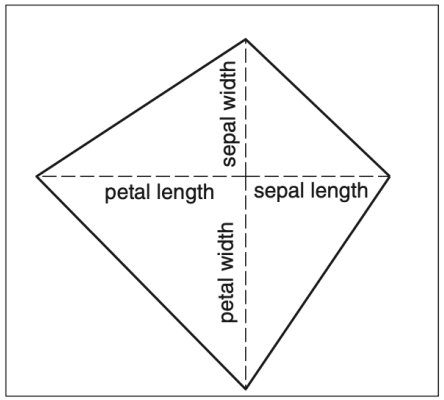


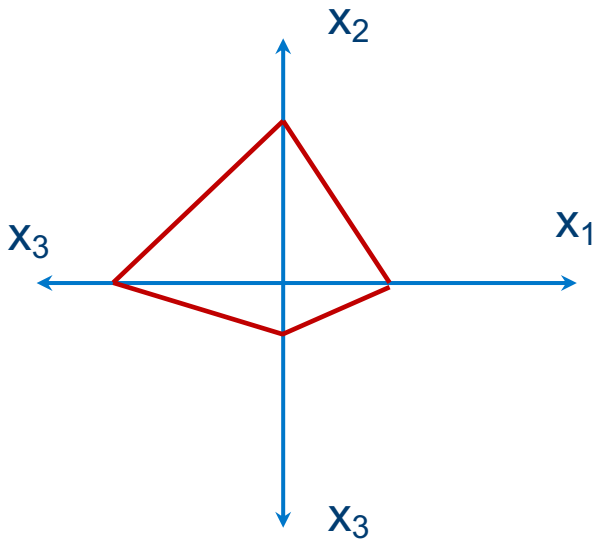


Other Visualization Techniques

- 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

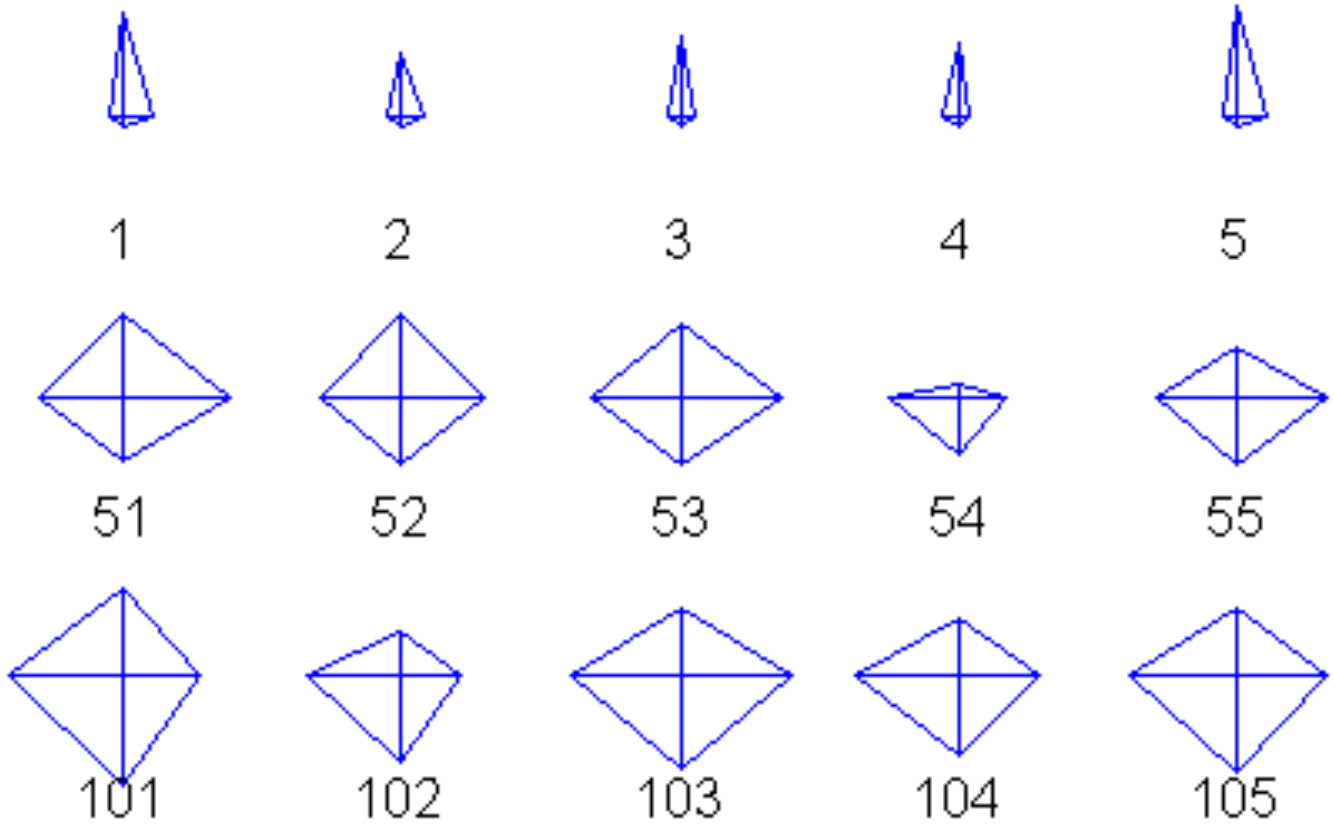






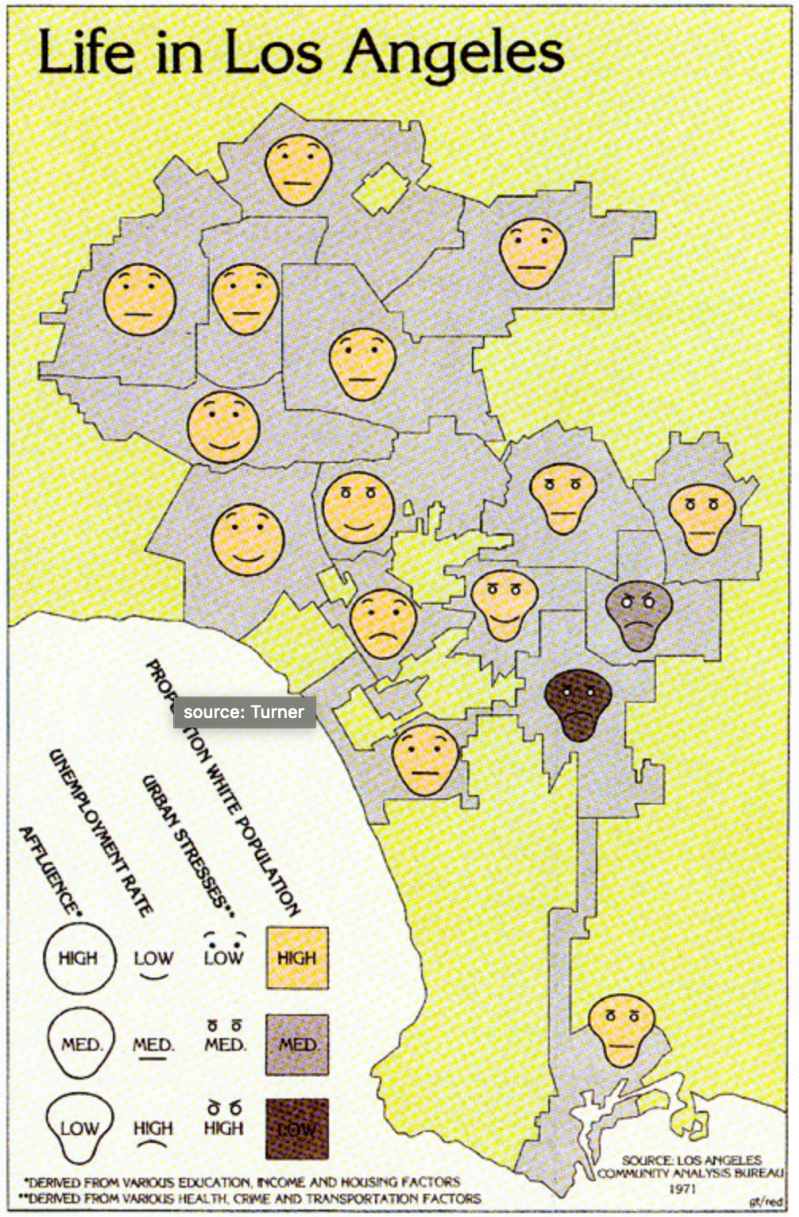
Star Plots for Iris Data

- Setosa
- Versicolour
- Virginica



Other Visualization Techniques

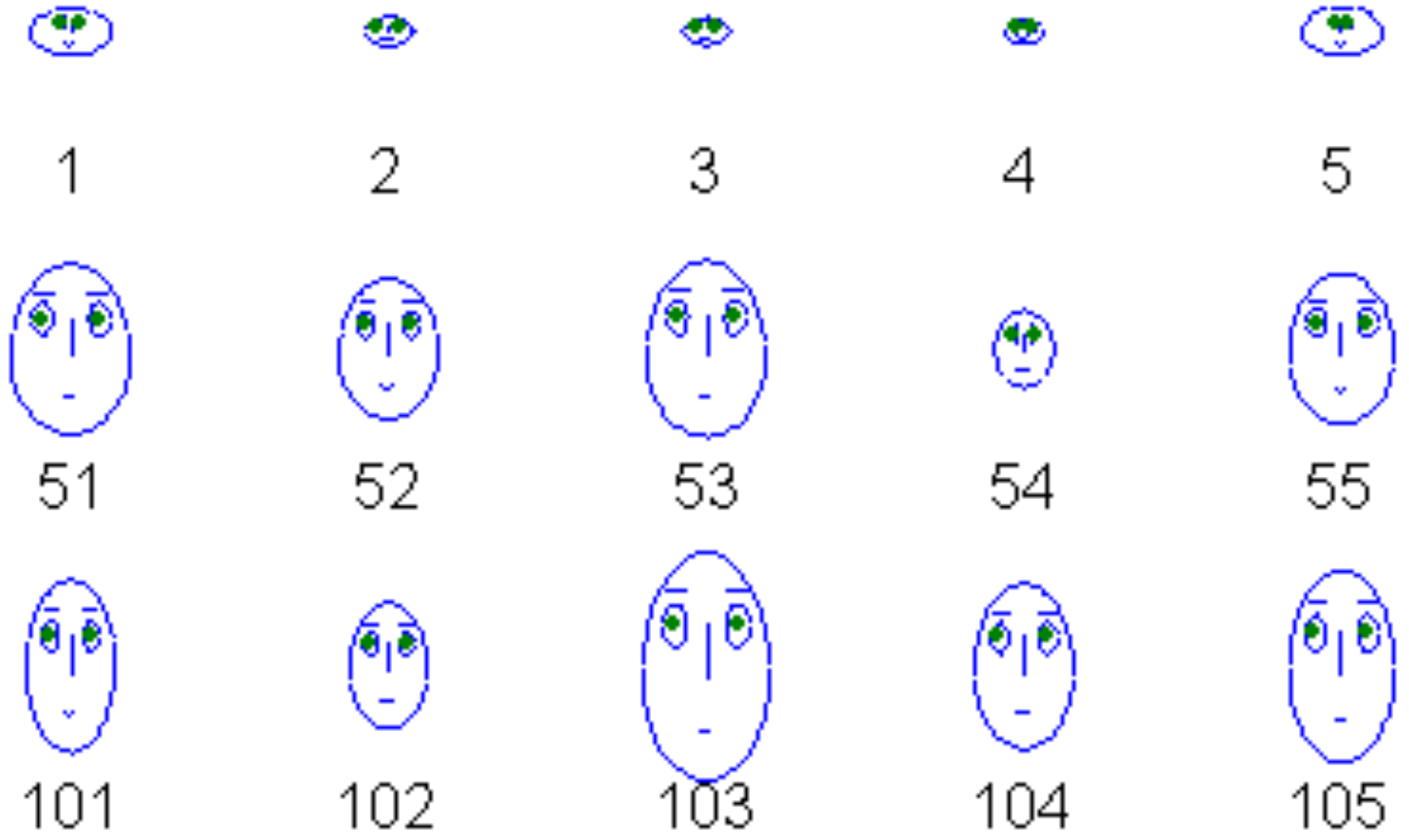
- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human's ability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



http://ceadserv1.nku.edu/longa/classes/2008fall/mat115-002/days/day34/ChernoffFaces.html

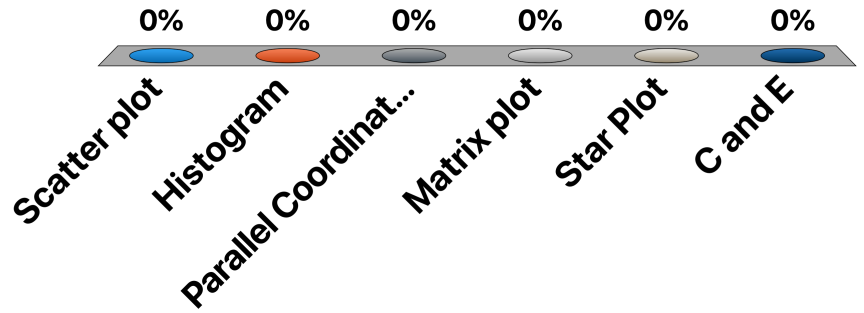
Chernoff Faces for Iris Data

- Setosa
- Versicolour
- Virginica



Which visualization method is appropriate for high-dimensional data?

A. Scatter plot
B. Histogram
C. Parallel Coordinates
D. Matrix plot
E. Star Plot
F. C and E



Today Participant Leaders

Points	Participant	Points	Participant