دانشگاه آزاد اسلامی واحد تبریز

نام درس: داده کاوی

بخش: مفاهیم و تکنیک ها در داده کاوی

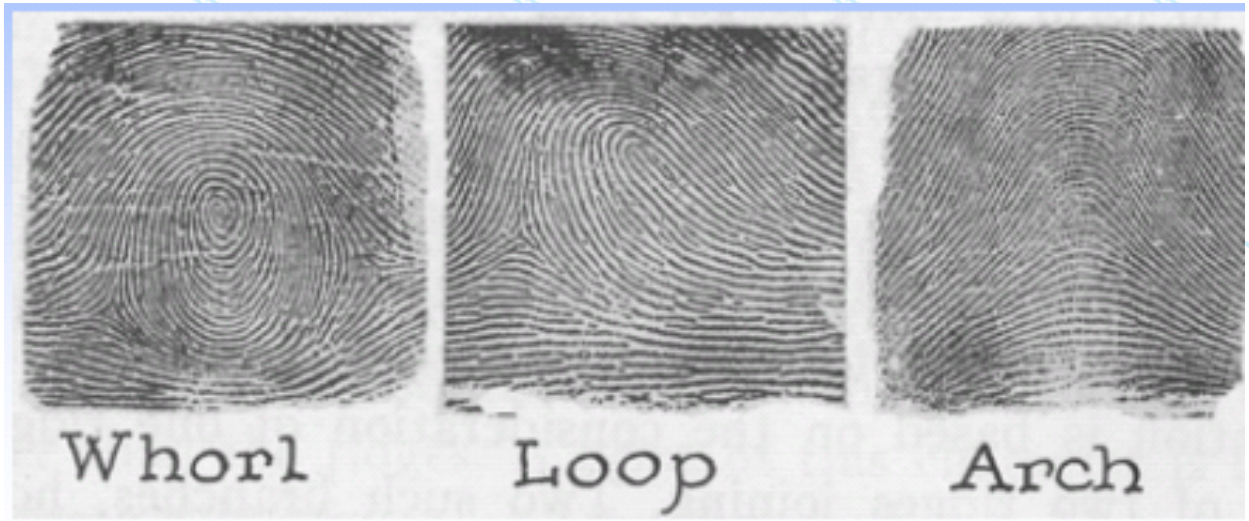
نام استاد: دکتر مسعود کارگر

Roadmap

- Data Preprocessing: Why?
- Preprocessing Tasks
  - Cleaning/Exploratory analysis/data reduction/conversion/feature selection or extraction
- Preprocessing Techniques
  - Weka filters
  - Summary Statistics
  - Visualization
  - Feature Selection
  - Dimension Reduction

Fingerprint Recognition Case

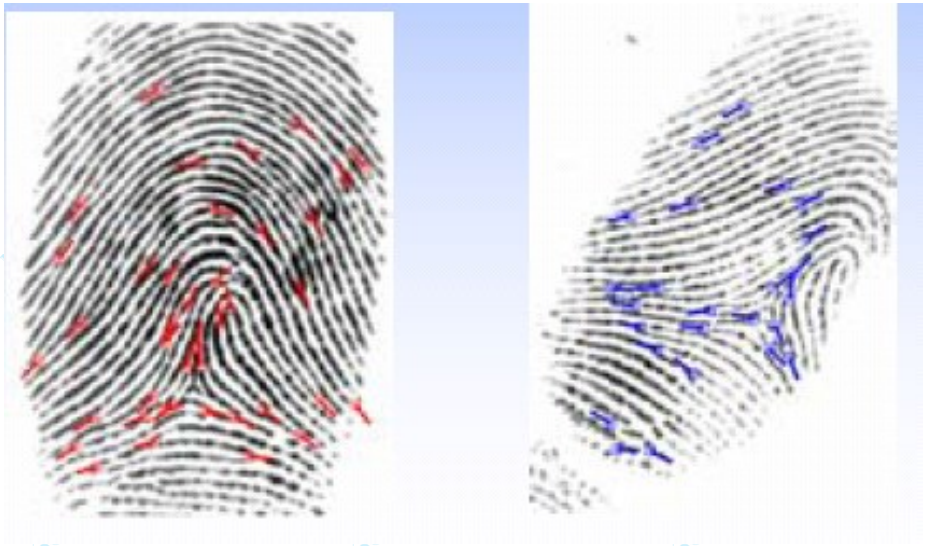
- Fingerprint identification at the gym

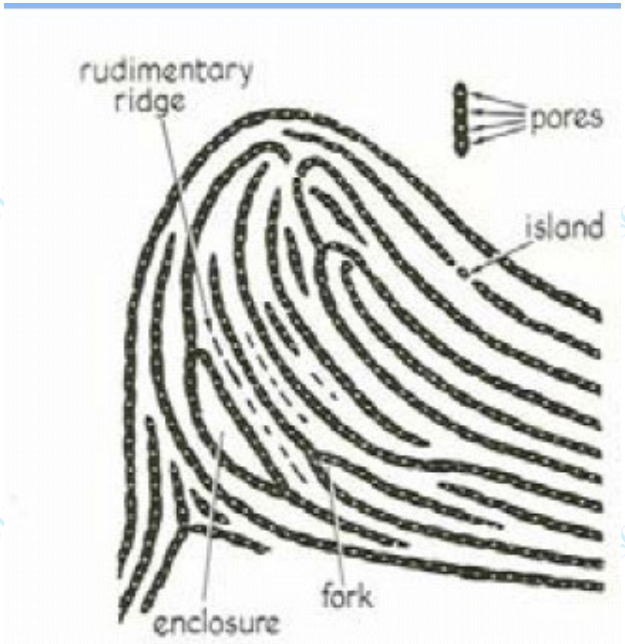


HOW?

Feature vector: 10.2, 0.23, 0.34, 0.34, 20, …

Feature Extraction in Fingerprint Recognition





It is not the points, but what is in between the points that matters... Edward German

•Identifying/extracting a good feature set is the most challenging part of data mining.

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
  - Redundant: including everything, some of which are irrelevant to our task

Why Is Data Dirty?

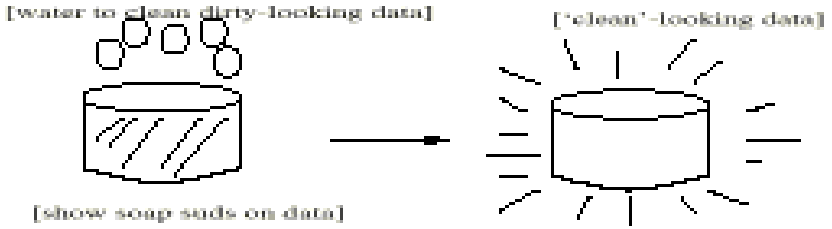
- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - 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 data 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
    - At least 70% of effort in a good data mining project is devoted to preprocessing
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

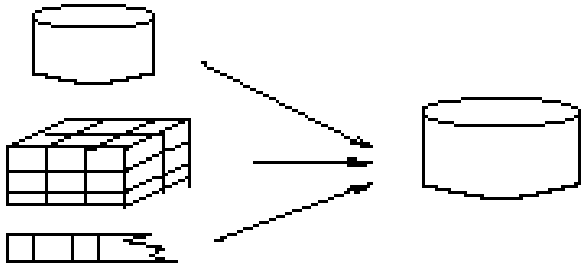
Forms of Data Preprocessing

Data cleaning



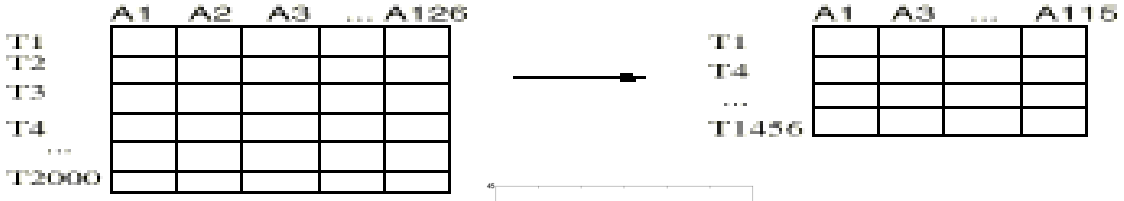
Data Integration

Data transformation

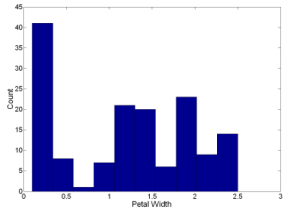


-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data reduction



Data Exploratory Analysis



Preprocessing in weka

- Weka Filters (normalization, etc)
  - **Discretization**
  - Attribute selection
  - Normalize
  - Standardize N(0,1)
  - Replace missing values
  - Numeric2binary
  - Random projection to reduce dimension

What is Data Exploration?

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

- Key motivations 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

Exploratory Data Analysis Techniques

- Summary Statistics
- Visualization
- Feature Selection (big topic)
- Dimension Reduction (big topic)

Roadmap

- Data Preprocessing: Why?
- Preprocessing Tasks
  - Cleaning/Exploratory analysis/data reduction/conversion/feature selection or extraction
- Preprocessing Techniques
  - Summary Statistics
  - Visualization
  - Feature Selection
  - Dimension Reduction

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.

2.1, 2.5, 1.4, 5.0 → Setosa (S)
1.0, 1.2, 3.0, 4.0 → Virginica (V)
1.0, 2.4, 5.0, 1.0 → Versicolour (R)

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

Central Tendency: Mean and Median

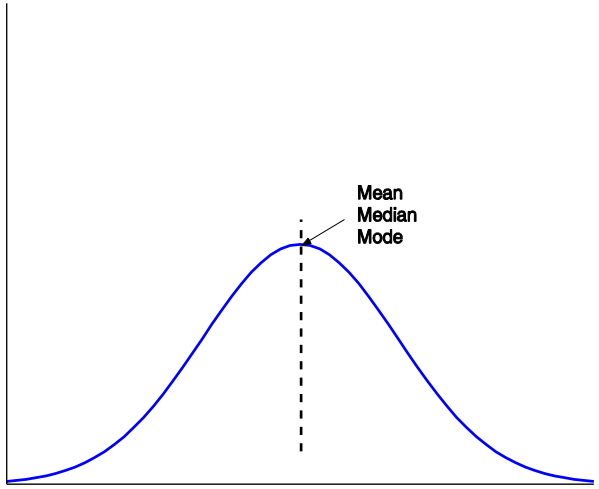
- The mean/average is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

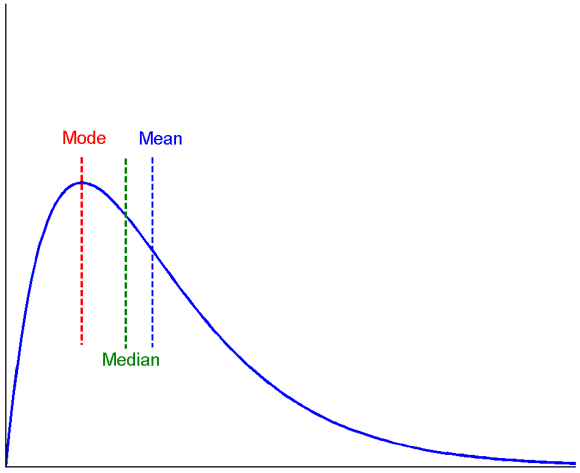
$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

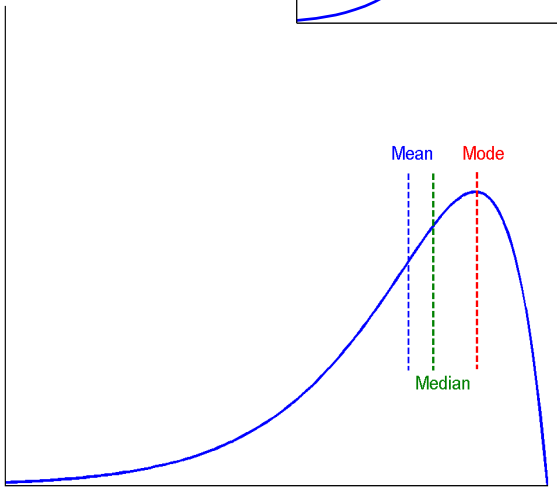
$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed 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 $x_{50\%}$.

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

  - Standard deviation *s (or σ)* is the square root of variance *s^2 (or σ^2)*

Measures of Spread: Additional Statistics

- Range, variance or standard deviation are sensitive to outliers, so that other measures are often used.

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

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

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

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

(male, female)
(red, green, blue)

Roadmap

- Data Preprocessing: Why?
- Preprocessing Tasks
  - Cleaning/Exploratory analysis/data reduction/conversion/feature selection or extraction
- Preprocessing Techniques
  - Summary Statistics
  - Visualization
  - Feature Selection
  - Dimension Reduction

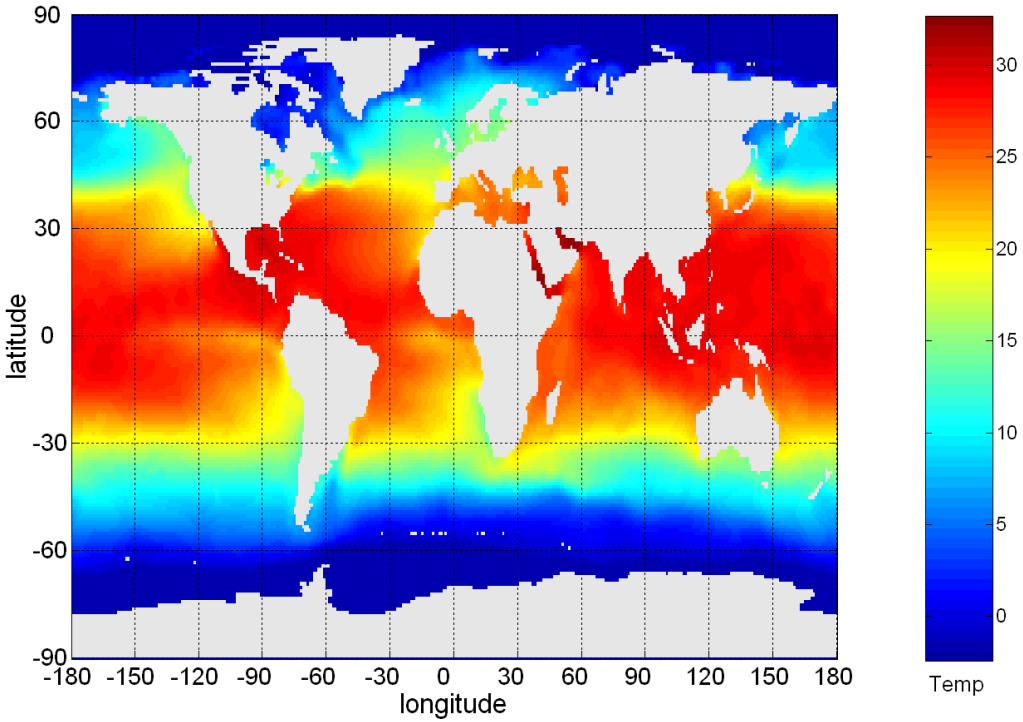
Visualization

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.

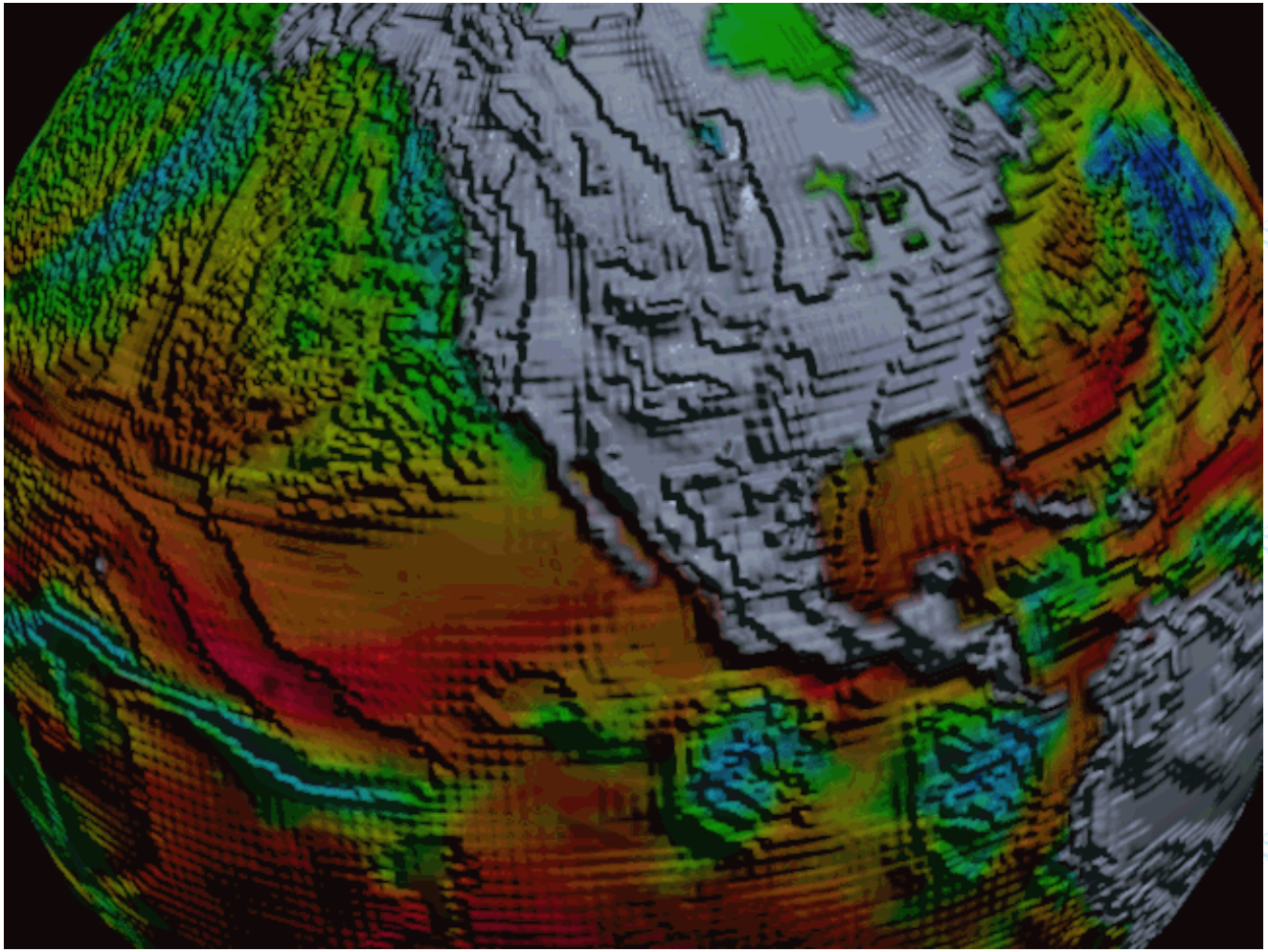
- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed amazing ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

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



Visualization of Wind Speed and Heat



Representation in visualization

- 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 in visualization

- 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

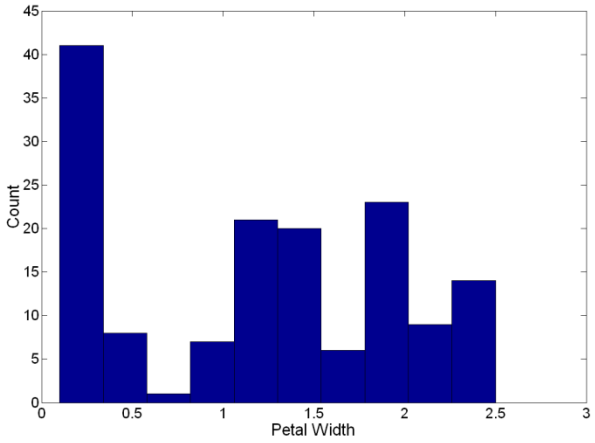
	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

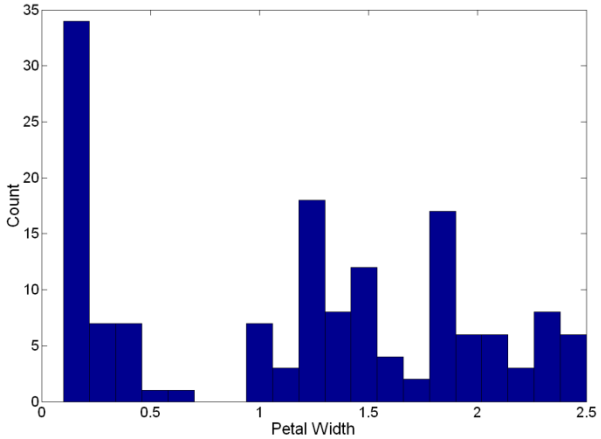
Feature Selection in Visualization

- 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

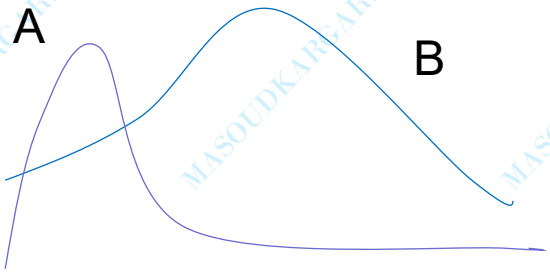
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width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





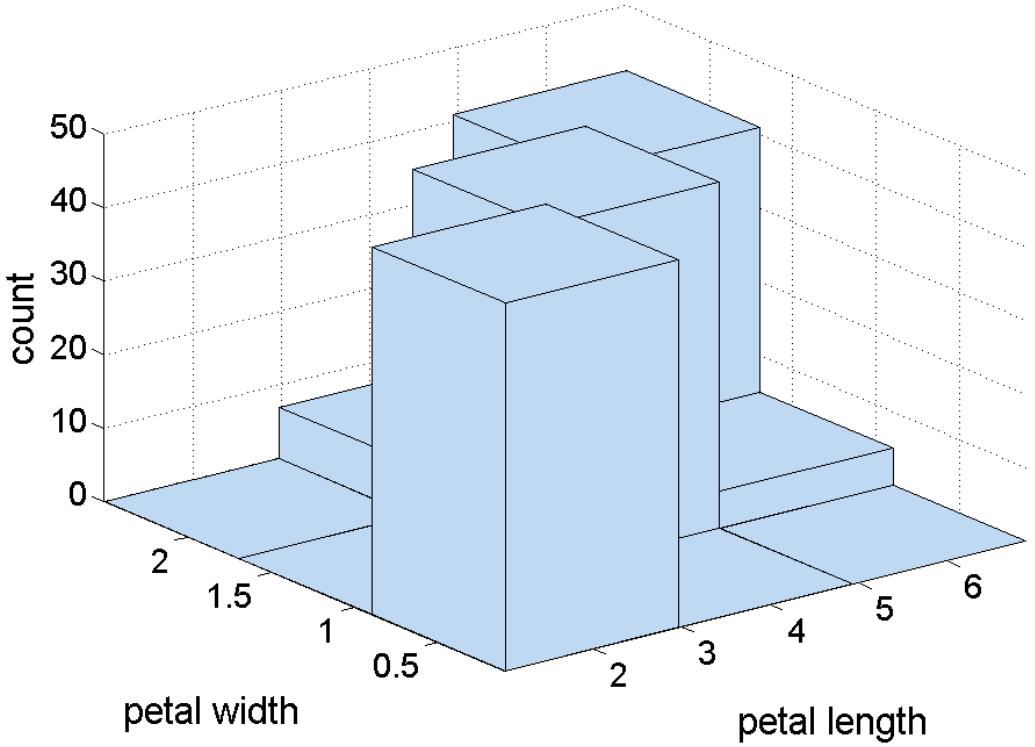
Applying Histogram to Feature Selection

- Suspect detection problem
- How to find features that can distinguish Suspects from Non-suspects?
  - From the training data, collect all the values of feature k for both the Label=1 and label=-1
  - Plot the histogram and check the two distributions
  - If they are very different, then this feature is relevant to the classification



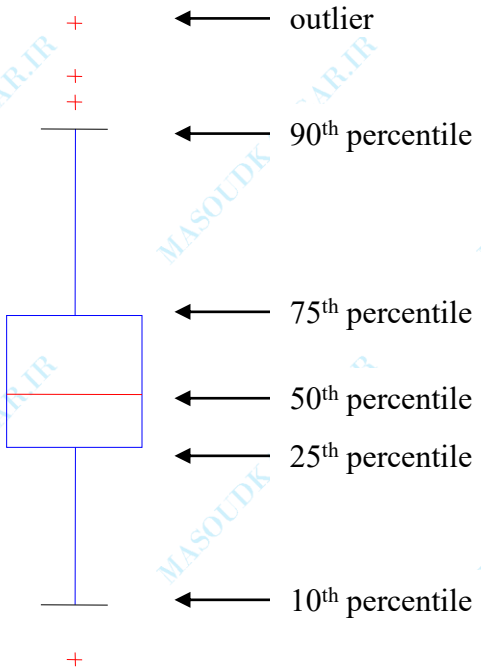
Two-Dimensional Histograms

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



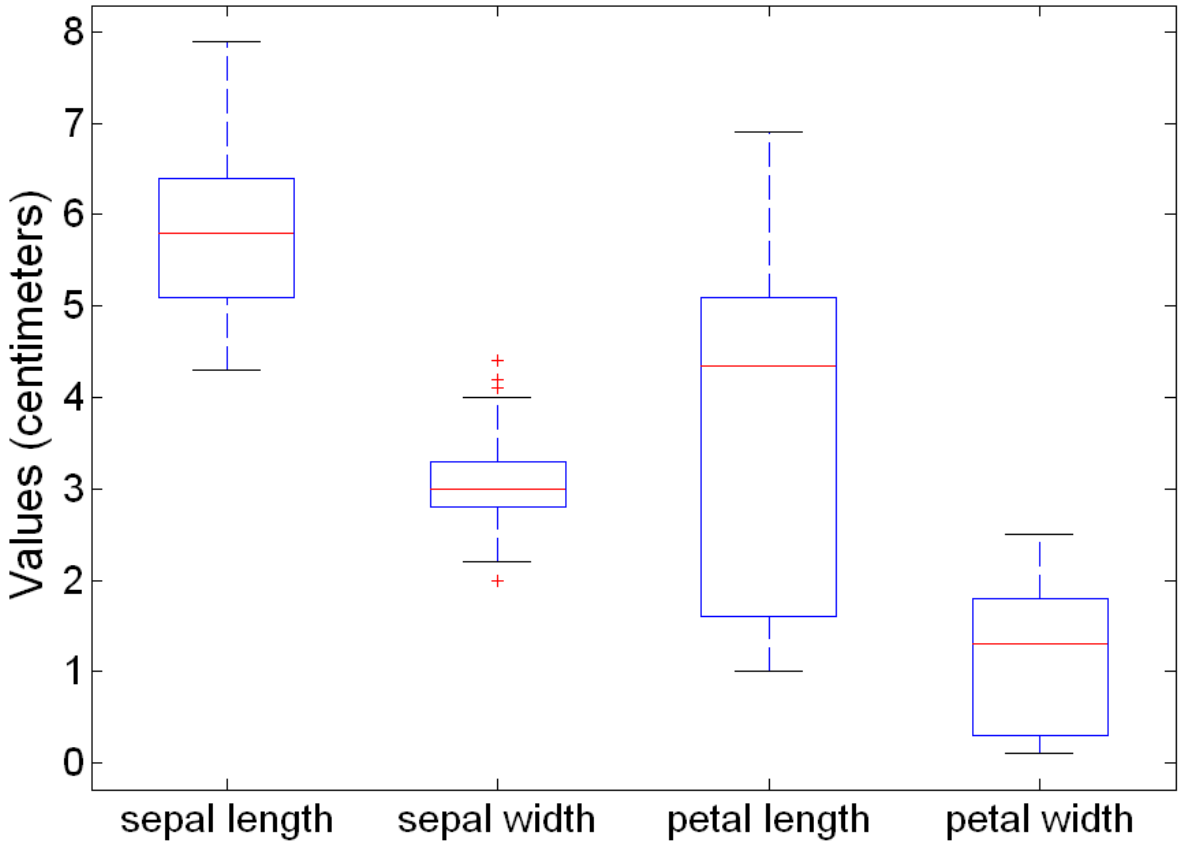
Visualization Techniques: Box Plots

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

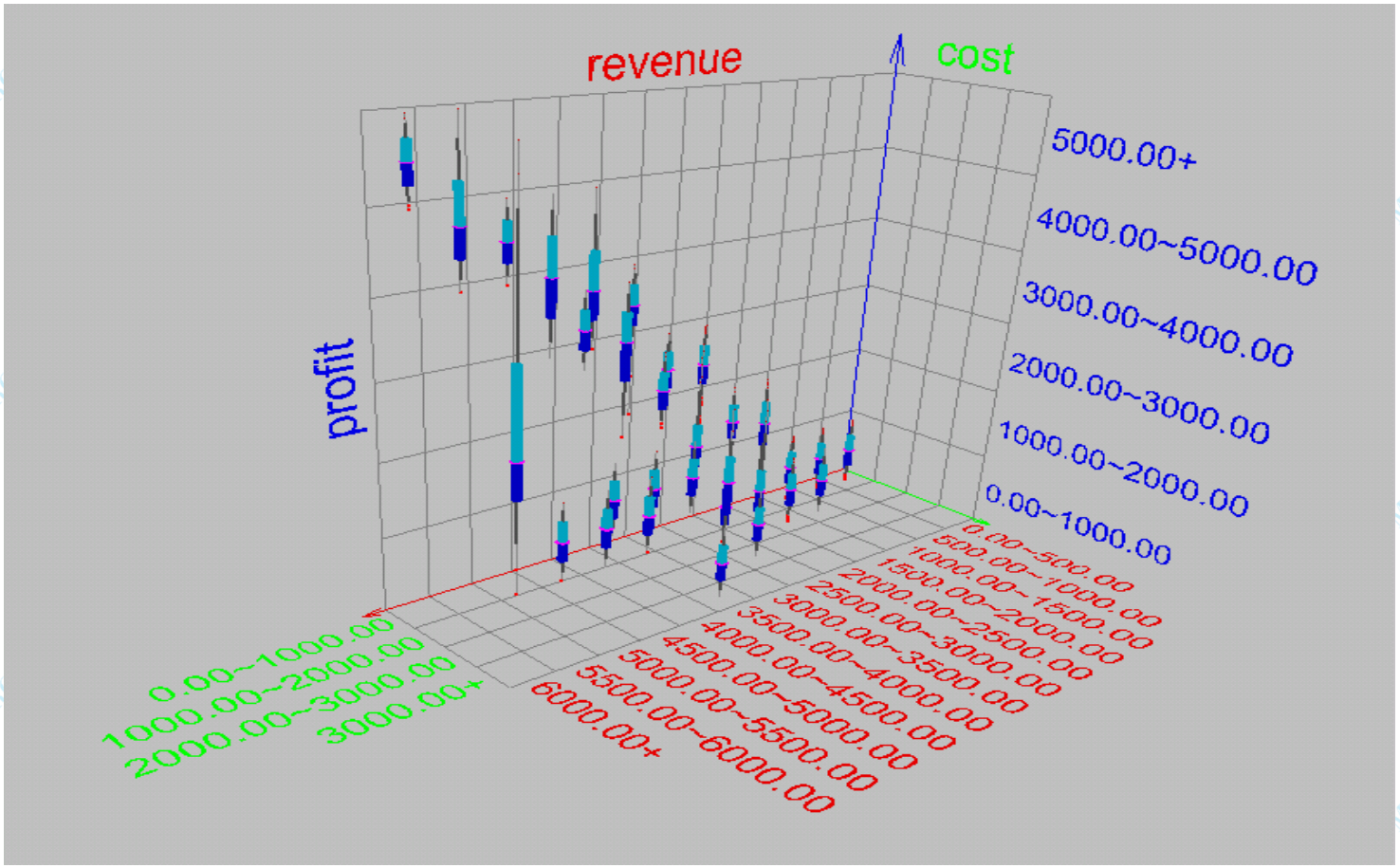


Example of Box Plots

- Box plots can be used to compare attributes

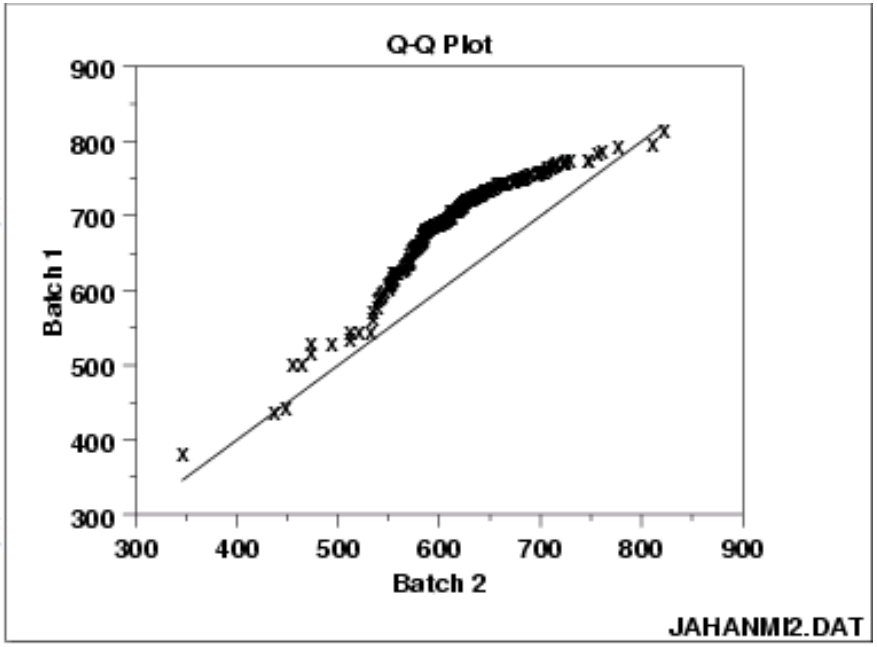


Multi-dimensional Boxplot Analysis



Quantile-Quantile (Q-Q) Plot

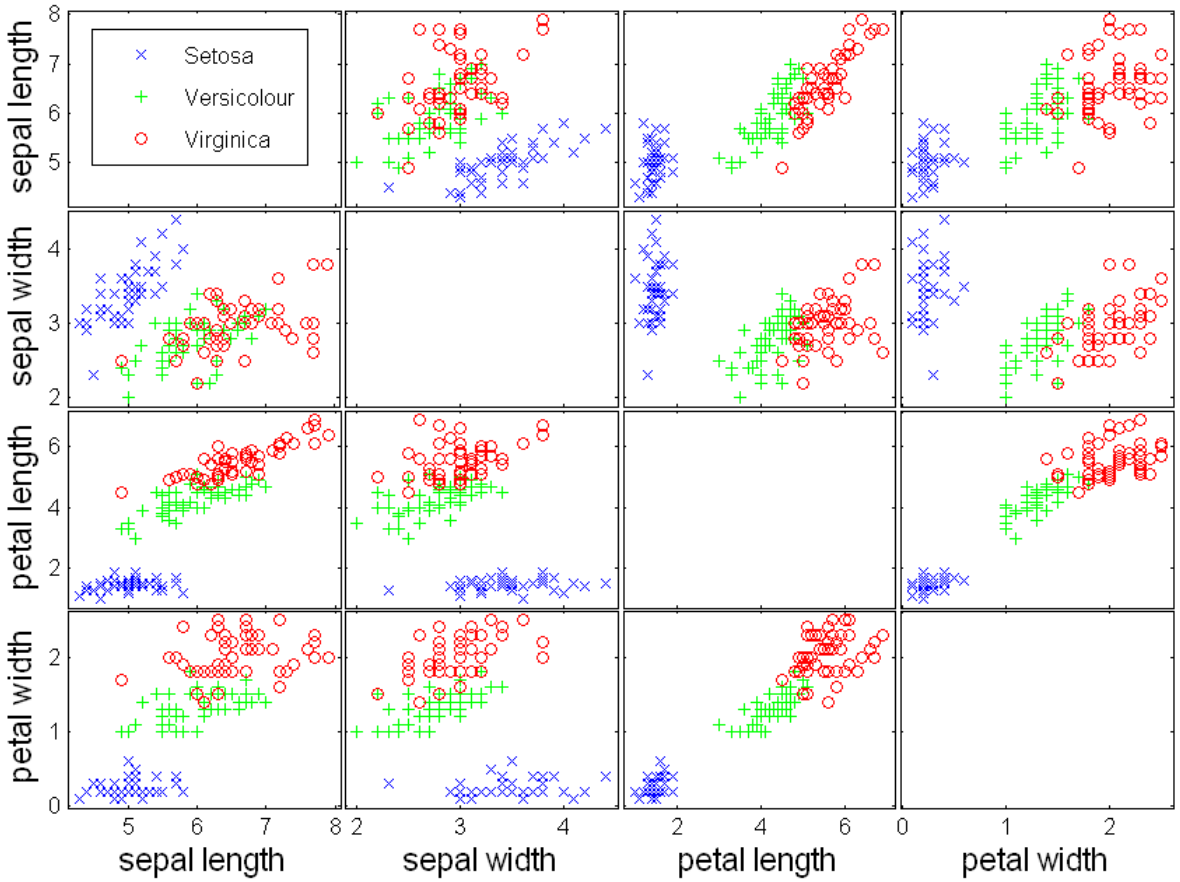
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another



Visualization Techniques: Scatter Plots

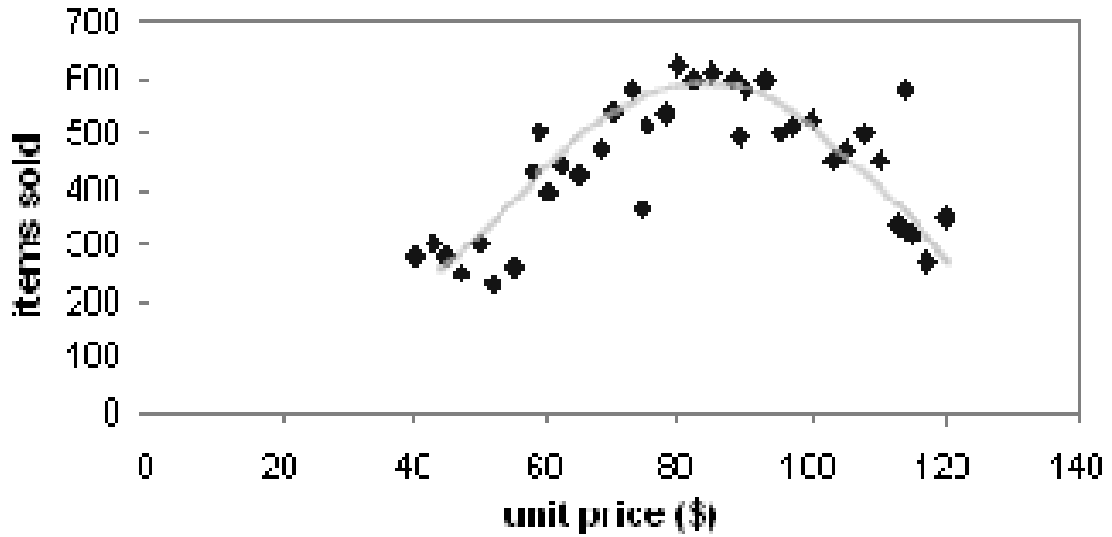
- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have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

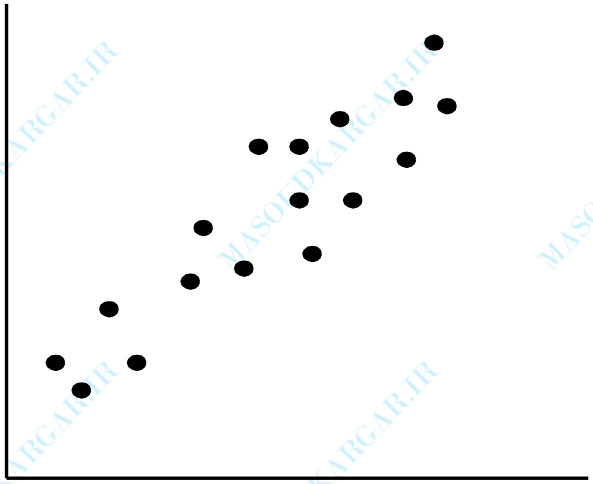


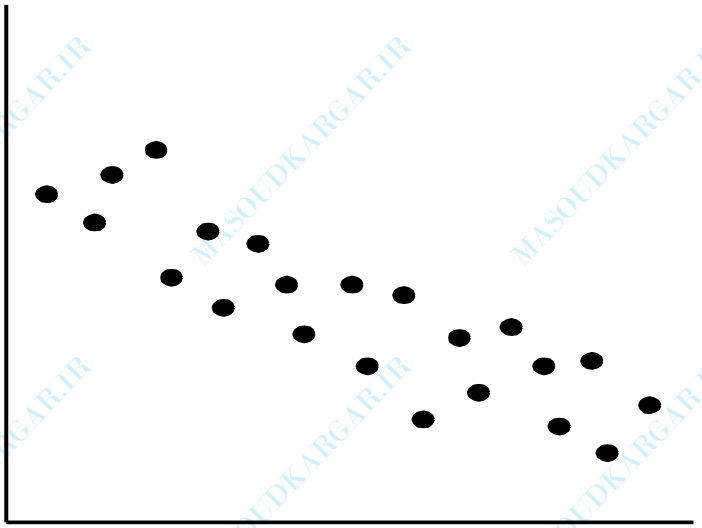
Loess Curve

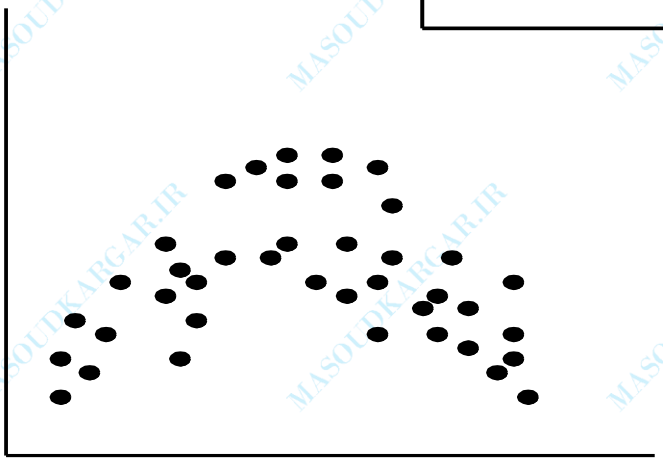
- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression



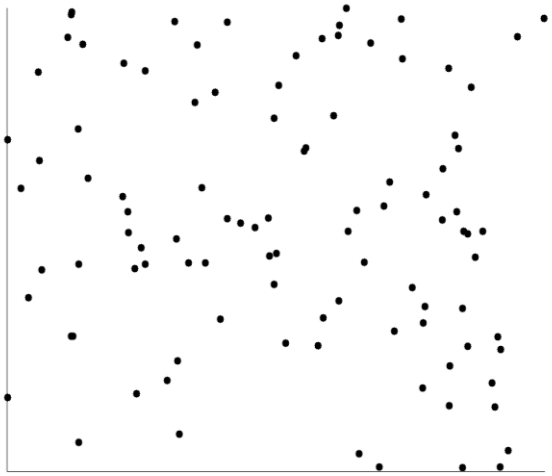
Detecting Positively and Negatively Correlated Data

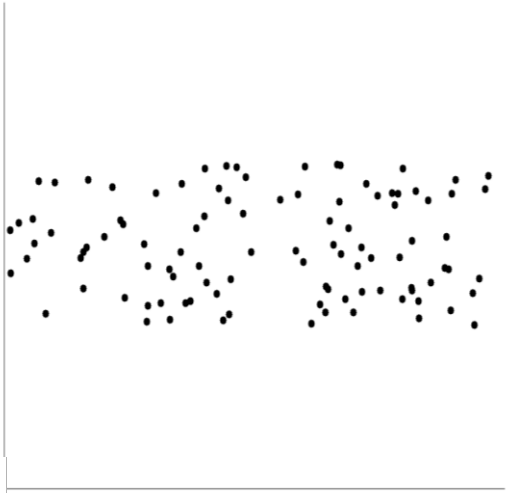


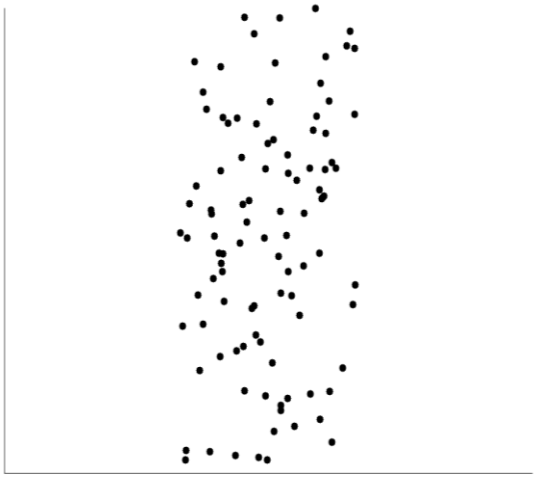




Not Correlated Data



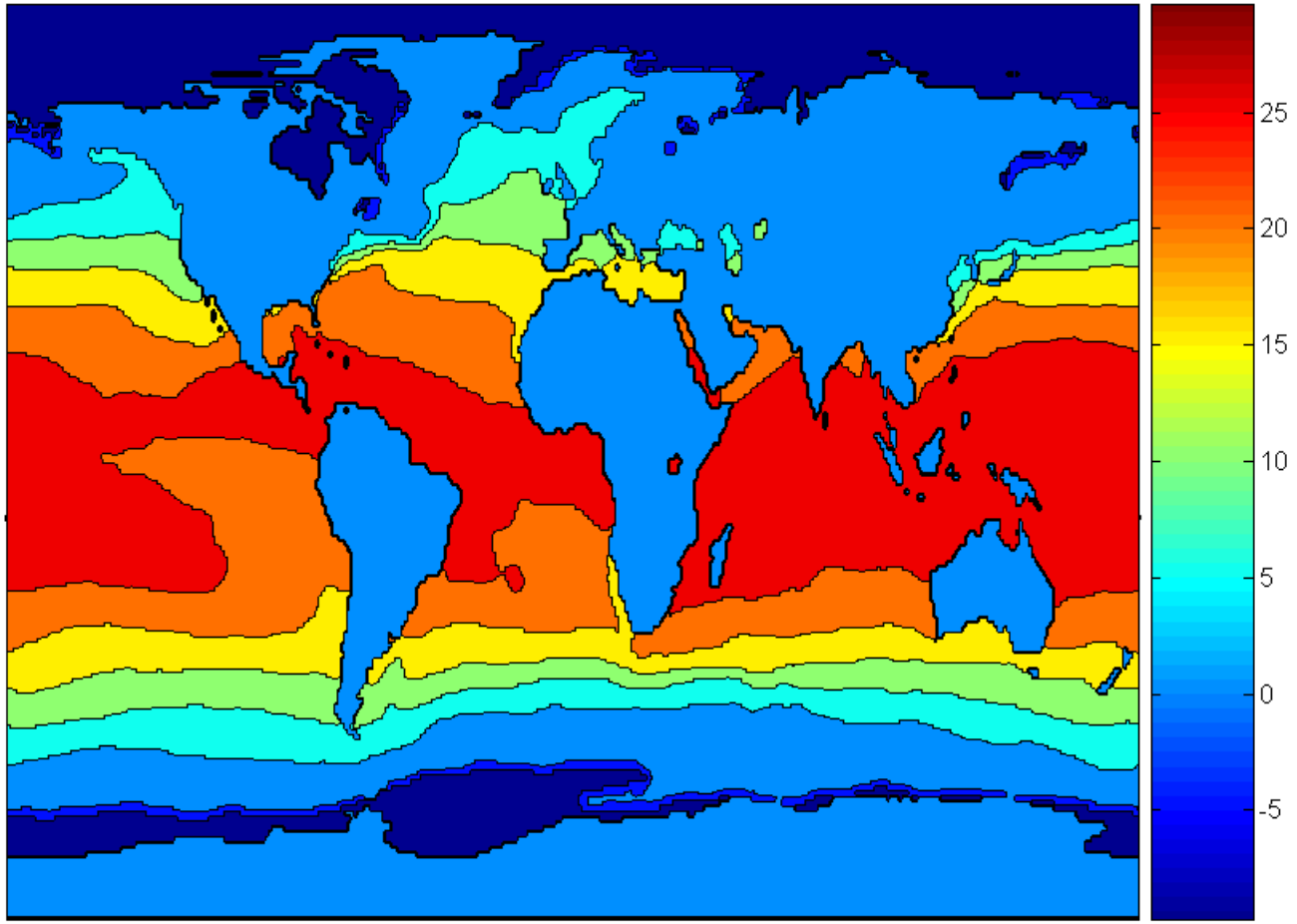




Visualization Techniques: Contour Plots

- 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

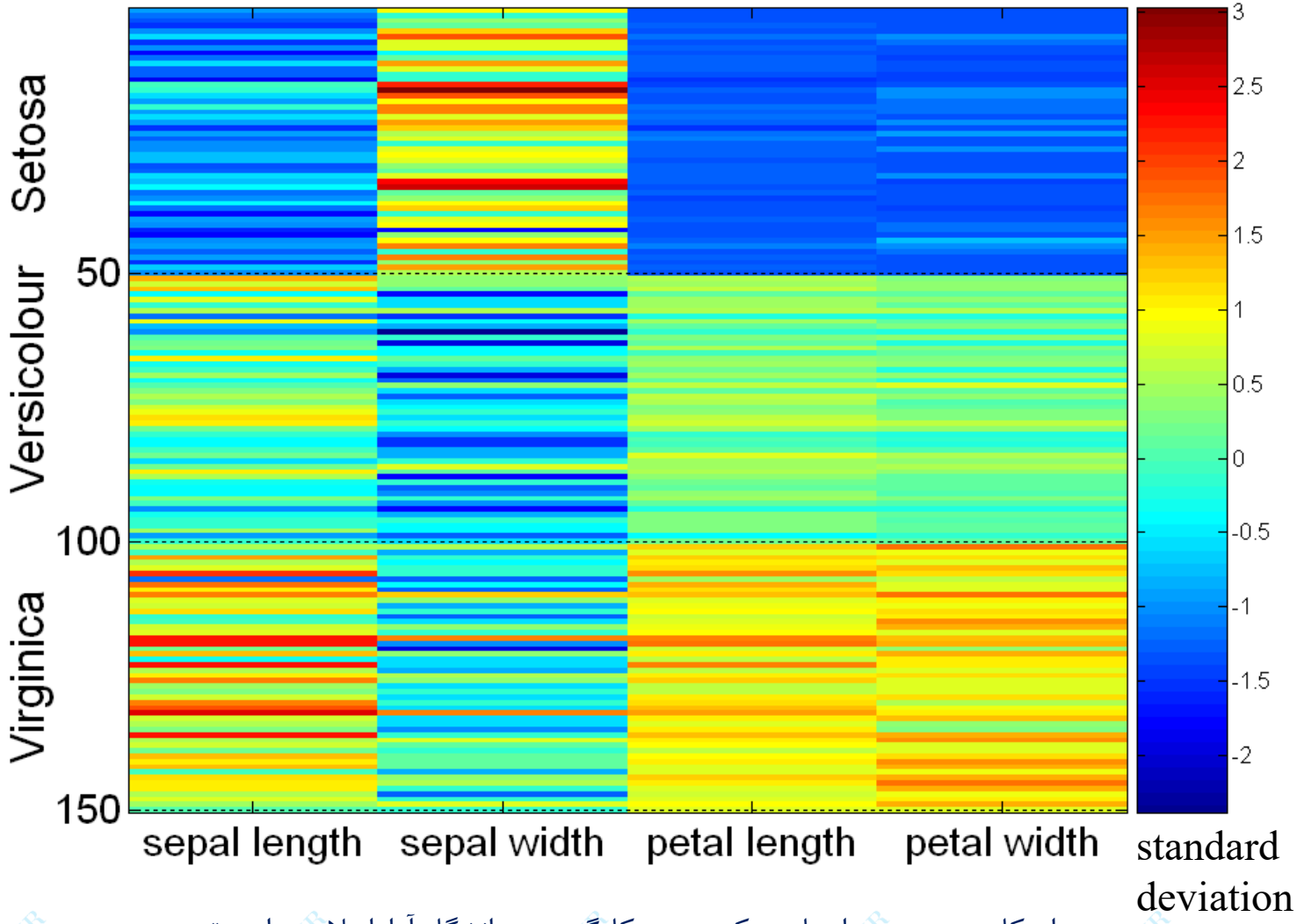
Contour Plot Example: SST Dec, 1998



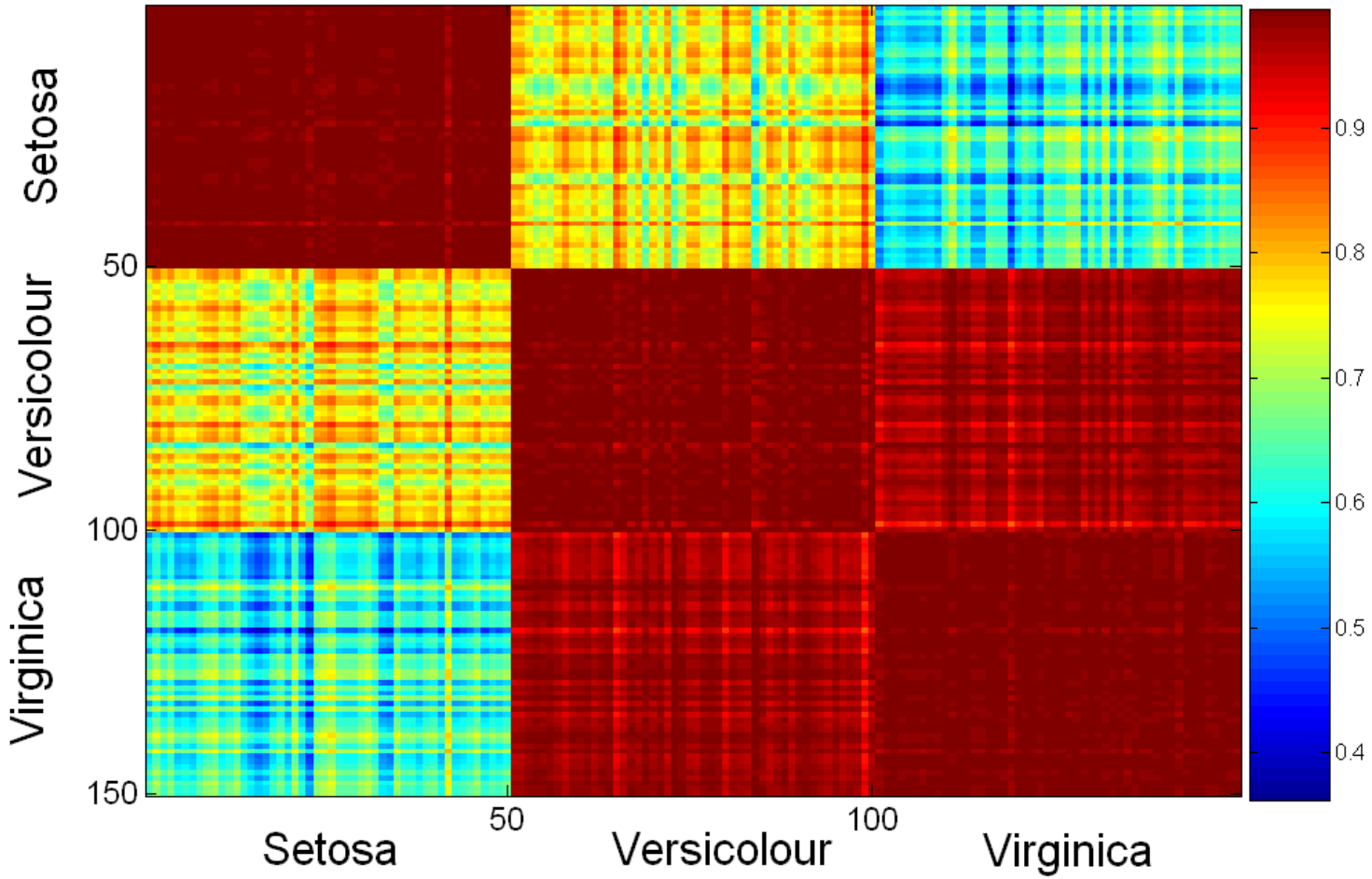
Visualization Techniques: Matrix Plots

- 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

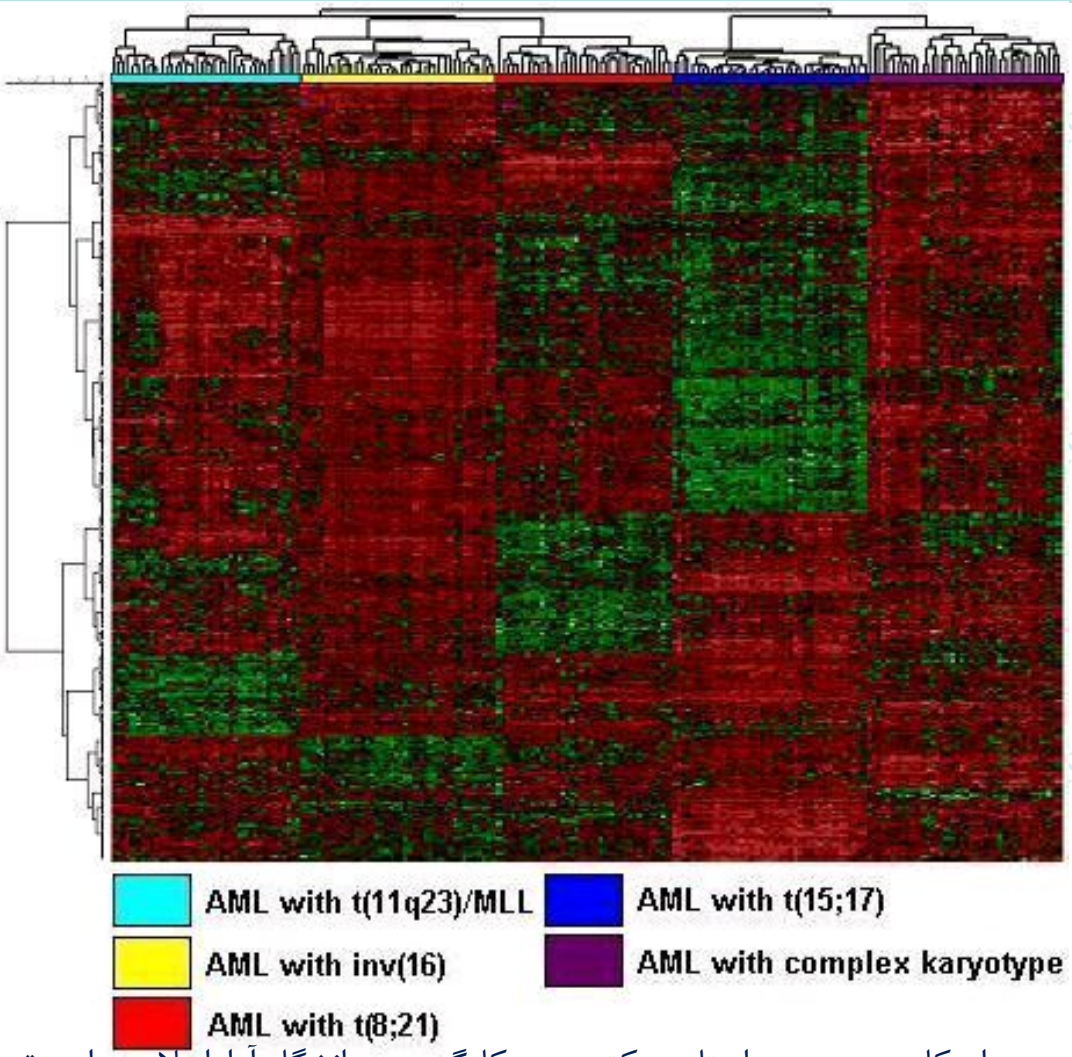
Visualization of the Iris Data Matrix



Visualization of the Iris Correlation Matrix



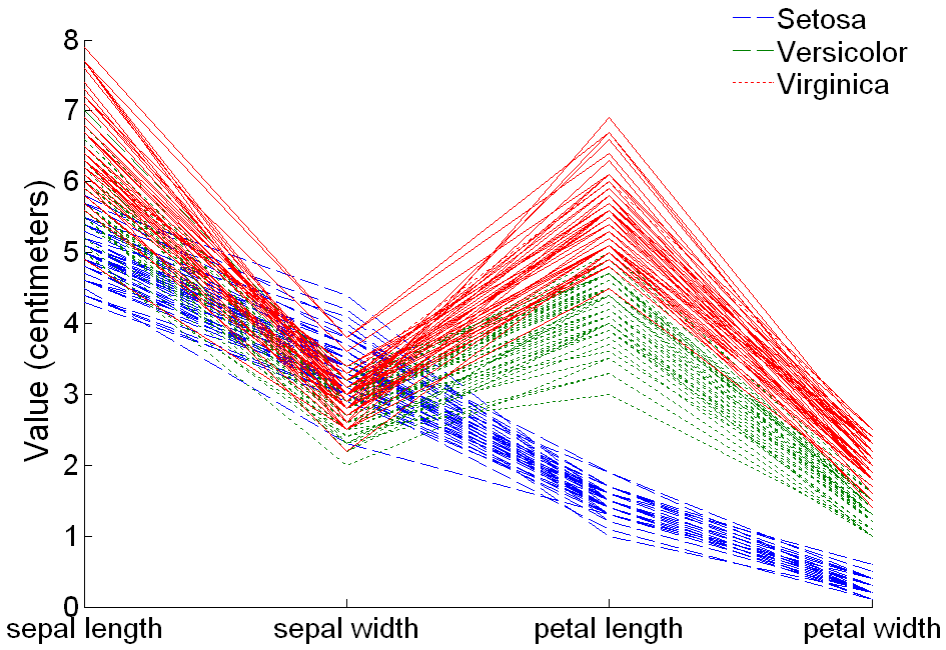
Heatmap of Microarray dataset

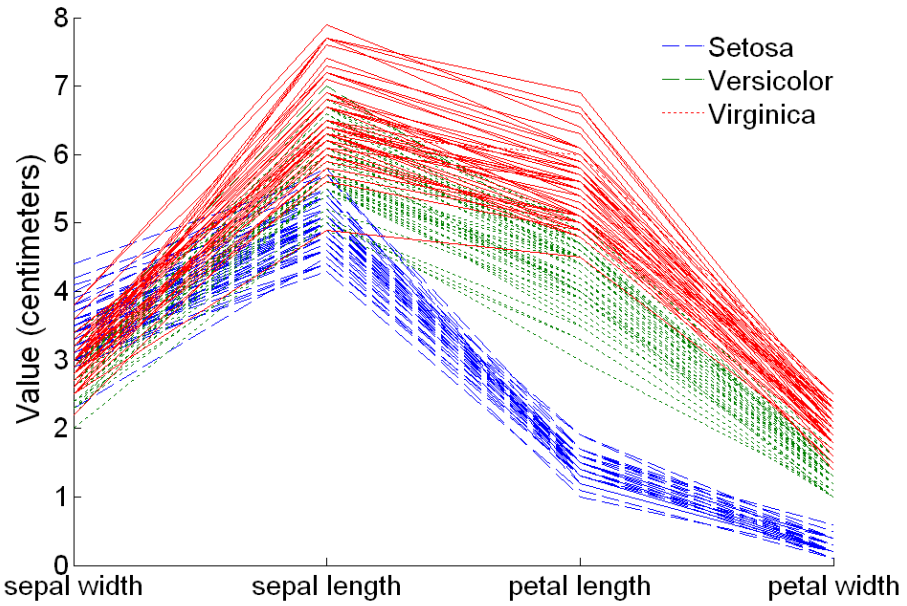


Visualization Techniques: Parallel Coordinates

- 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

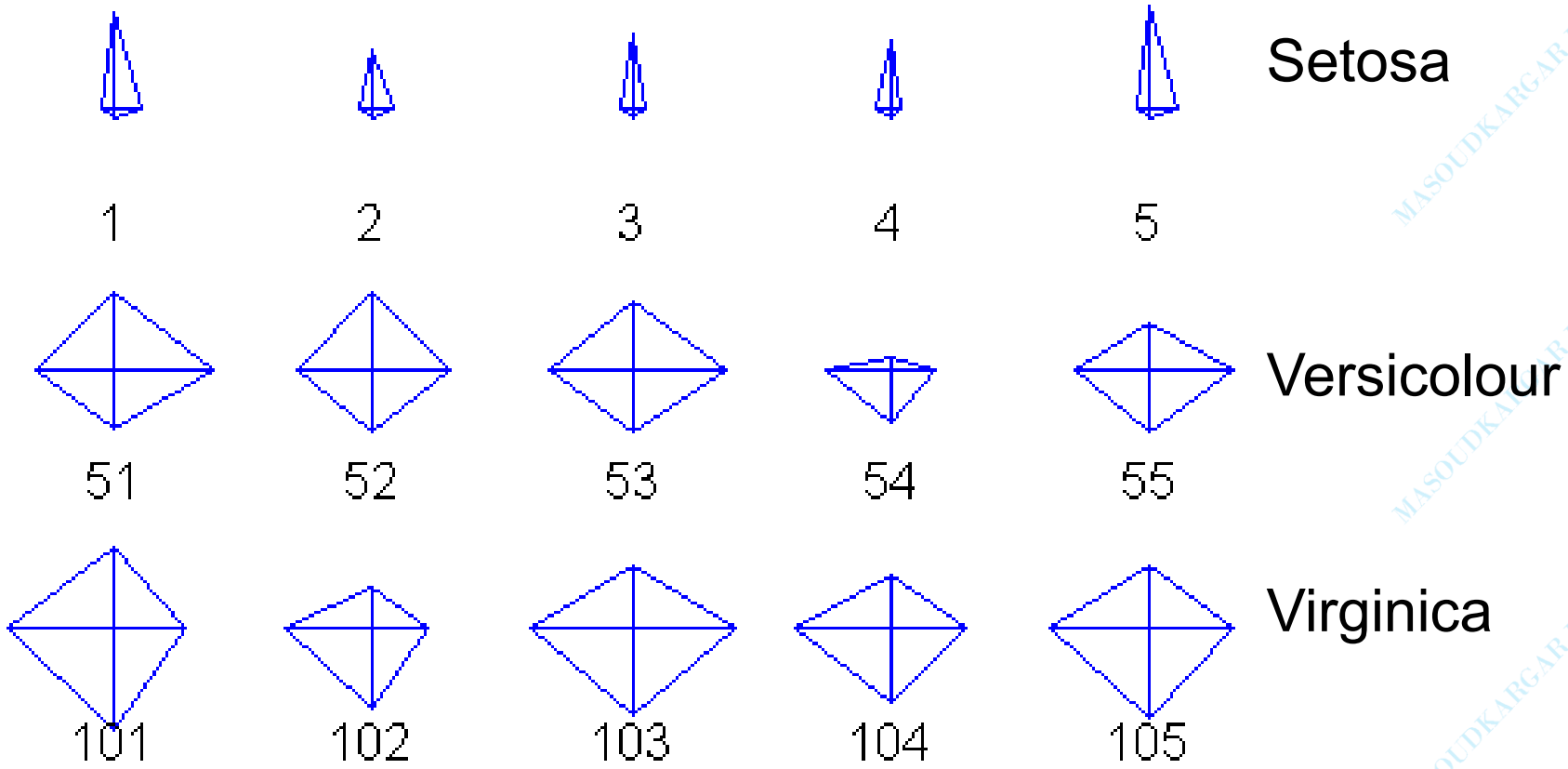




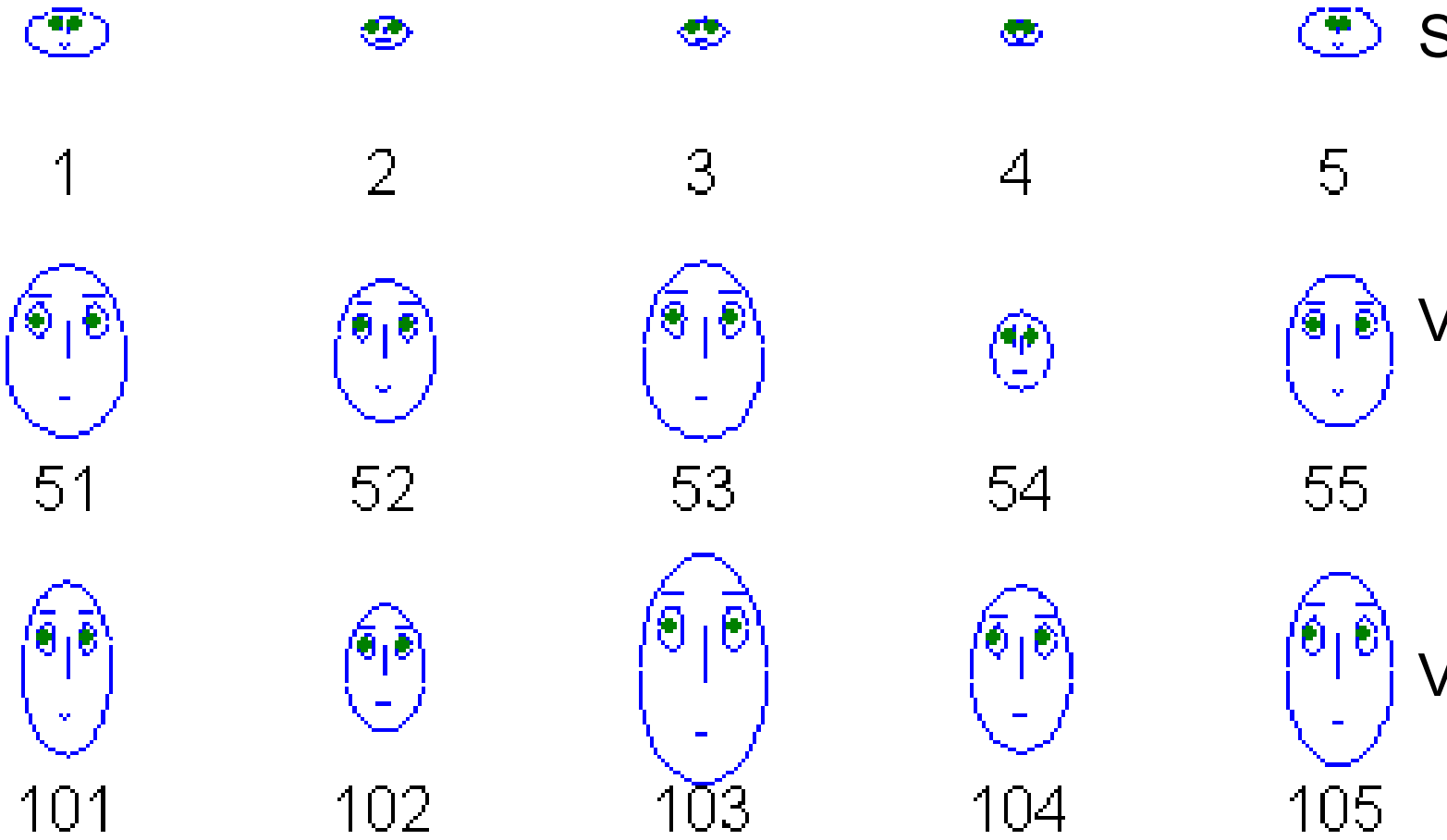
Other Visualization Techniques

- 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



Chernoff Faces for Iris Data



Survey of Concepts

Histogram

Parallel scatter plot

Scatter Plot

Q-Q Plot

Feature selection

Star plot

Data Visualization Toolbox

- How to get hands on?
  - Excel, Matlab, plotting/figure/graphs
  - Data visualization package in Mathematica
  - The Visualization ToolKit (VTK)
  - IBM Manyeyes toolkit
  - GUESS is an exploratory data analysis and visualization tool for graphs and networks
  - http://www.visual-literacy.org/periodic_table/periodic_table.html
  - Try visDB http://www.dbs.informatik.uni-muenchen.de/dbs/projekt/visdb/visdb.html

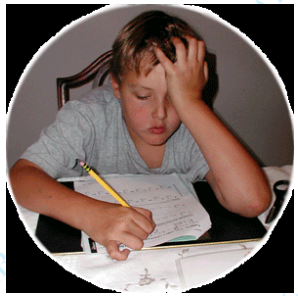
Strategic Learning for CSCE822

Check Topic of the week: data exploration

Read corresponding chapters of the Textbook

Study the textbook slides (Han, Tan)

Discuss, talk, brainstorm
Highlight key points
Obtain big picture



Homework, Assignments
Practice and apply the techniques

Review Slides (Hu)
Refer Textbook

PROJECT LEAD

Accomplish a wonderful final Project to show it out!

قدردانی

- Dr. Jianjun Hu
  http://mleg.cse.sc.edu/edu/csce822/
- University of South Carolina
- Department of Computer Science and Engineering