
Basic Visualization Techniques: Charts, Graphs and Maps

CS598 Information Visualization
Spring 2010

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 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.

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

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

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

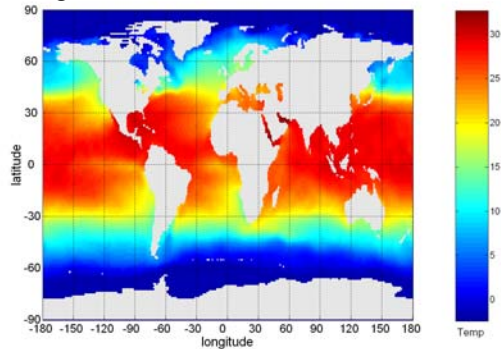
Visualization

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.

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

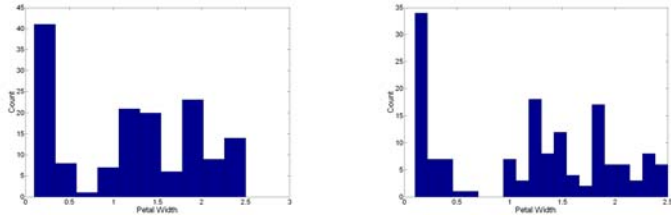
	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

Selection

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

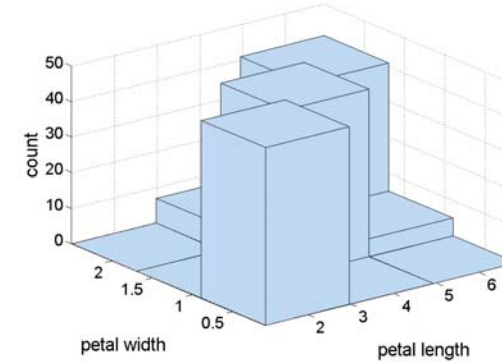
Visualization Techniques: Bar Charts

- Bar Chart (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)



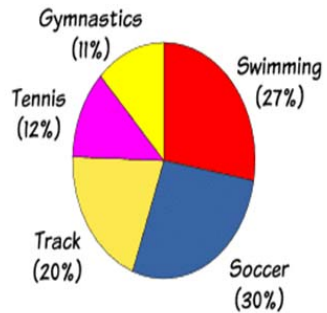
Two-Dimensional Bar Charts

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



Visualization Techniques: Pie Charts

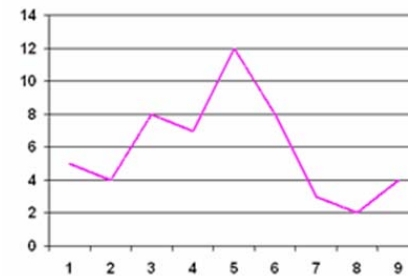
Pie Chart / Pie Graph



A pie chart is a circular chart divided into sectors, each sector shows the relative size of each value.

Visualization Techniques: Line Graphs

Line Graph

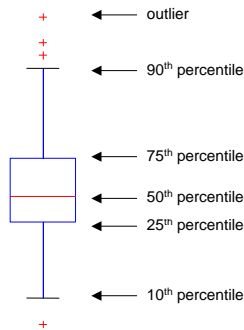


A graph that uses points connected by lines to show how something changes in value (as time goes by, or as something else happens).

Visualization Techniques: Box Plots

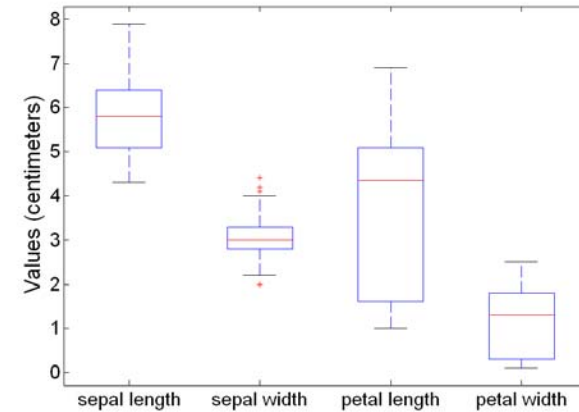
• Box Plots

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



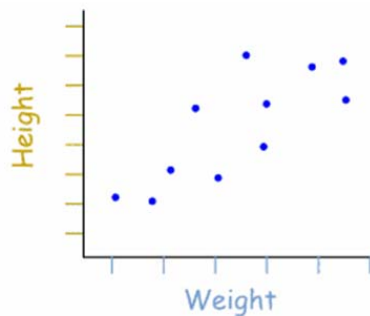
Example of Box Plots

- Box plots can be used to compare attributes



Visualization Techniques: Scatter Plots

Scatter Plot



A graph of plotted points that show the relationship between two sets of data.

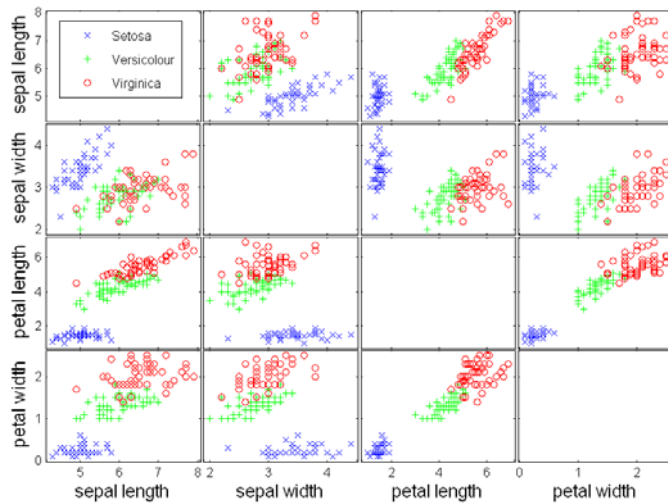
In this example, each dot represents one person's weight versus their height.

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: Maps

Map



A diagram of a place.

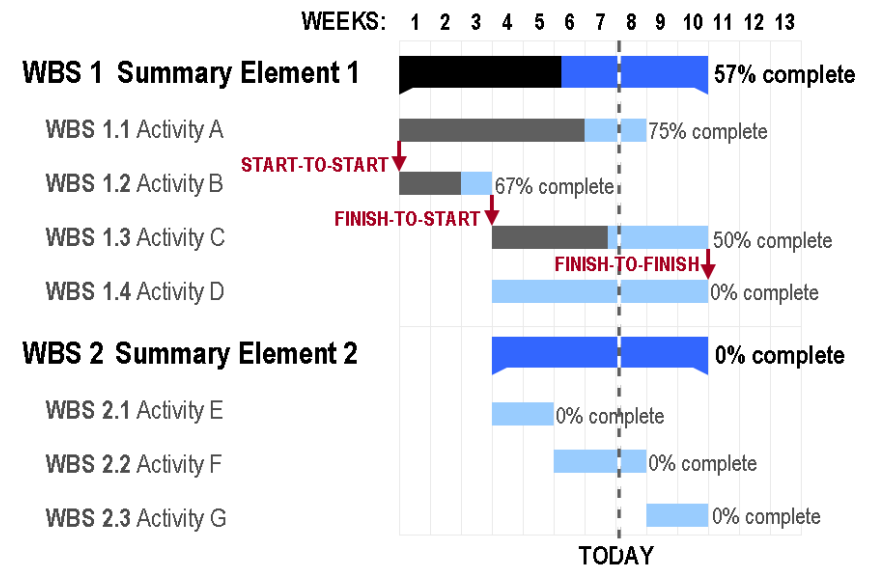
Usually showing features like roads, rivers and cities.

This is a map of Switzerland.

Visualization Techniques: Gantt Charts

- A **Gantt chart** is a type of *bar chart* that illustrates a project schedule.
- Gantt charts illustrate the *start* and *finish* dates of the *terminal* elements and *summary* elements of a project. Terminal elements and summary elements comprise the work breakdown structure of the project.
- Some Gantt charts also show the *dependency* (i.e., precedence network) relationships between activities.
- Gantt charts can be used to show current schedule status using percent-complete shadings and a vertical "TODAY" line.

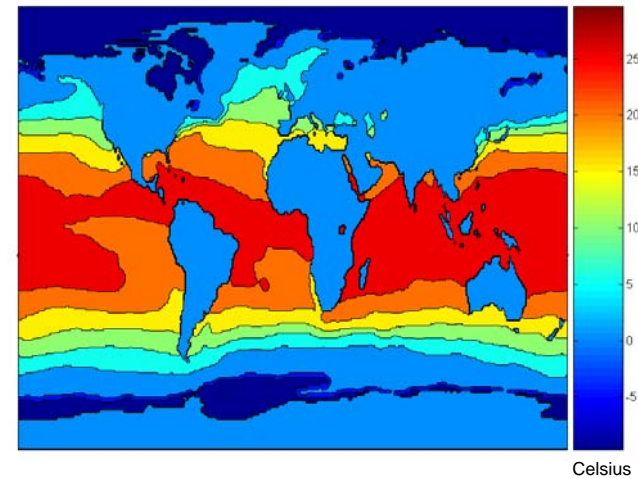
Gantt Chart Anatomy



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

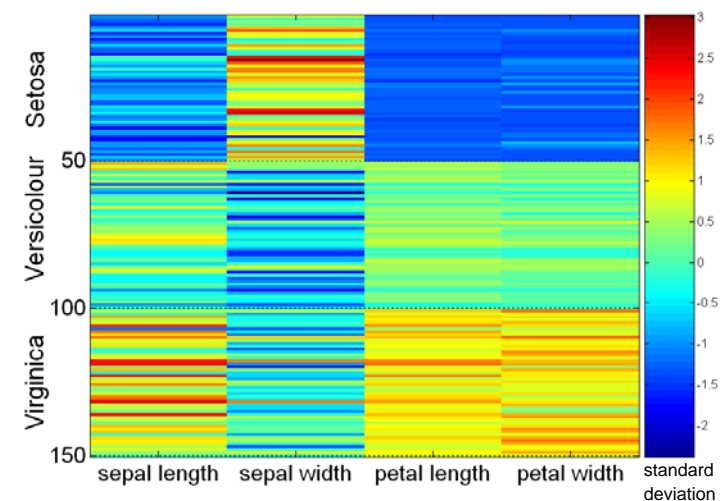
Contour Plot Example: SST Dec, 1998



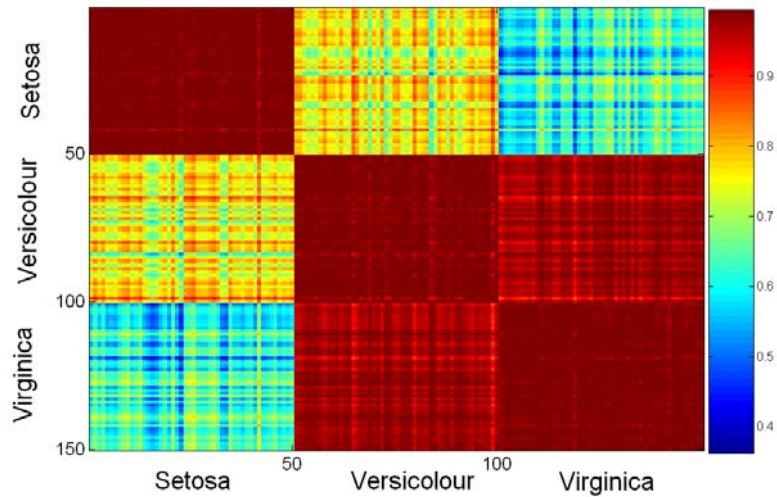
Visualization Techniques: Matrix Plots

- 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



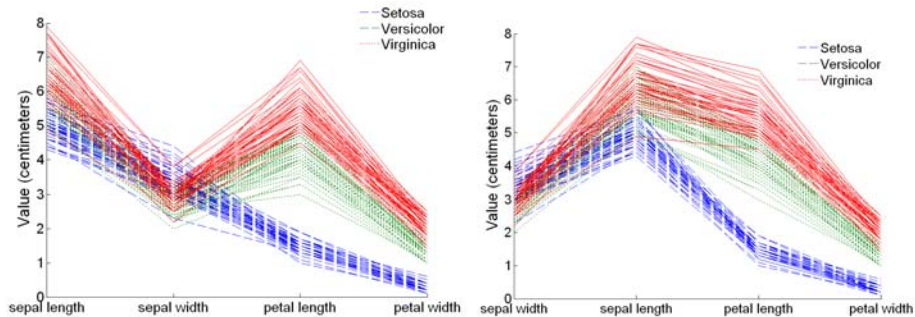
Visualization of the Iris Correlation Matrix



Visualization Techniques: Parallel Coordinates

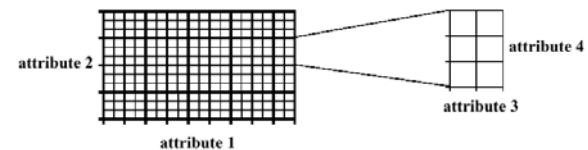
- Parallel Coordinates
 - Used to plot the attribute values of high-dimensional data
 - Instead of using perpendicular axes, use a set of parallel axes
 - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
 - Thus, each 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

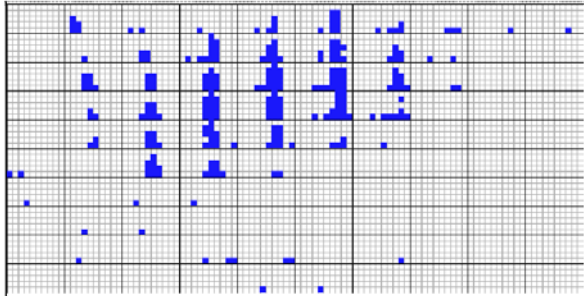


Dimensional Stacking

- Partitioning of the n-dimensional attribute space in 2-dimensional subspaces which are 'stacked' into each other
- Partitioning of the attribute value ranges into classes
- The important attributes should be used on the outer levels
- Adequate especially for data with ordinal attributes of low cardinality



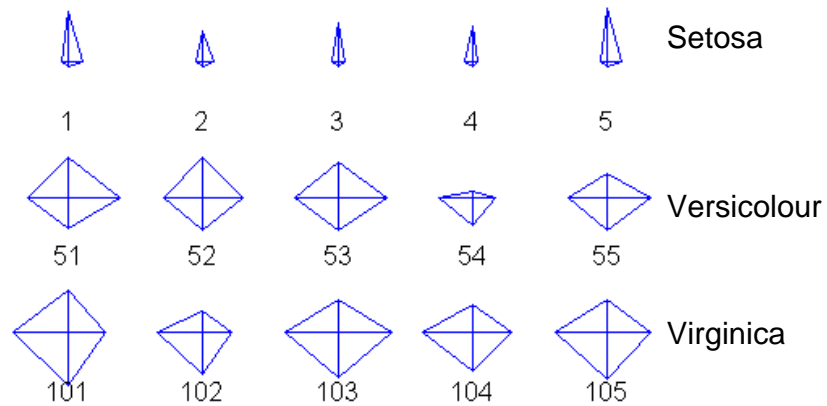
Example of Dimensional Stacking



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

