UNIVERSITY OF LIVERPOOL

# Exploring Data

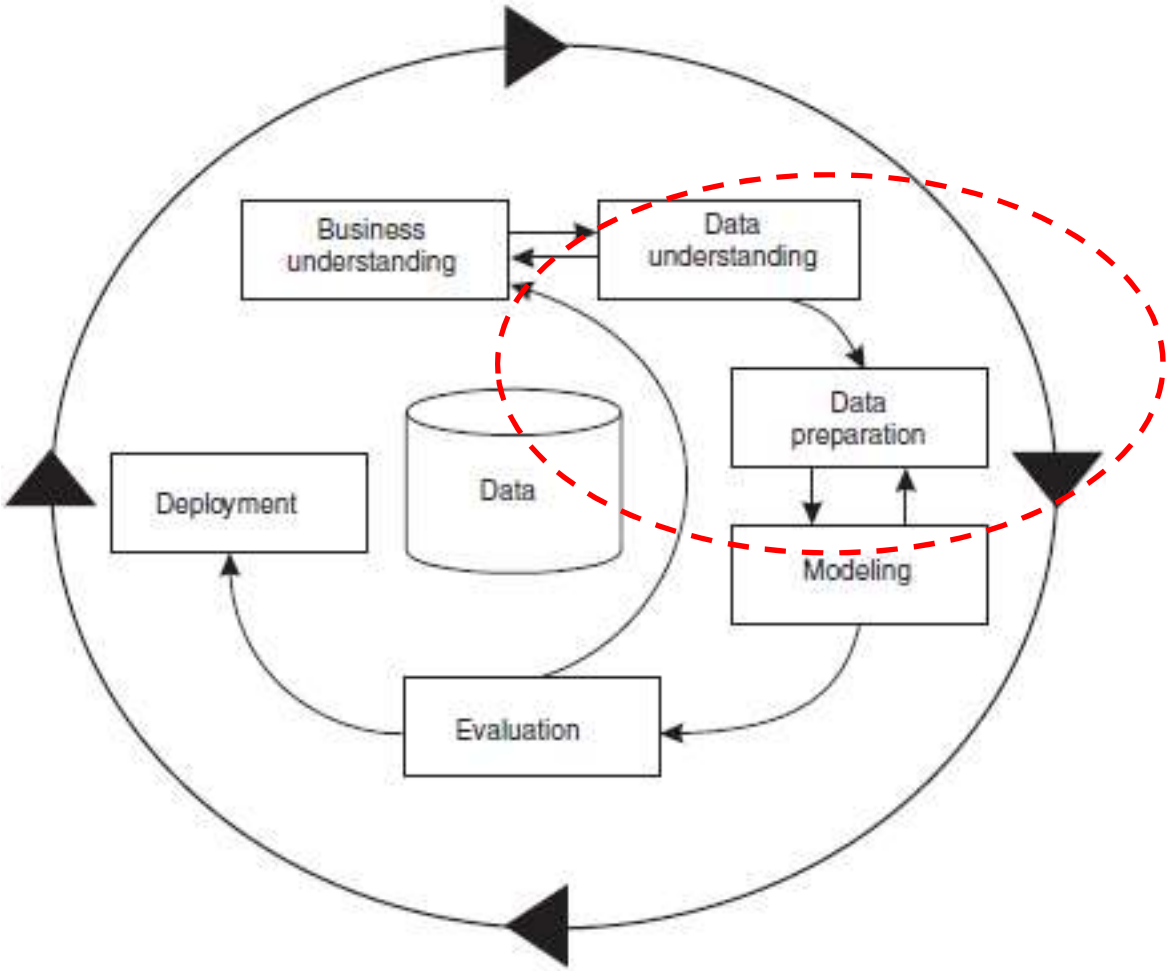
Prof. Dongping Song

University of Liverpool Management School

Email: Dongping.song@liv.ac.uk

Mainly based on Ch3 of Tan, P.N., Steinbach, M., and Kumar, V. (2014) Introduction to Data Mining, Pearson.

# The CRISP-DM Process Model




CRISP=Cross-Industry Standard Process

Mariscal et al (2010). A survey of data mining and knowledge discovery process models and methodologies, Knowledge Engineering Review, 25, 137-166.

# Learning Outcomes

- Data exploration
- Techniques used in data exploration
- Summary statistics
- Visualization:
  - Representation, arrangement, selection
  - Visualization techniques
- Online analytical processing (OLAP)

# What Is Data Exploration?

**A preliminary investigation of the data to better understand its specific 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 in 1970s
  - A nice online introduction of EDA can be found in the NIST Engineering Statistics Handbook at:

  http://www.itl.nist.gov/div898/handbook/index.htm

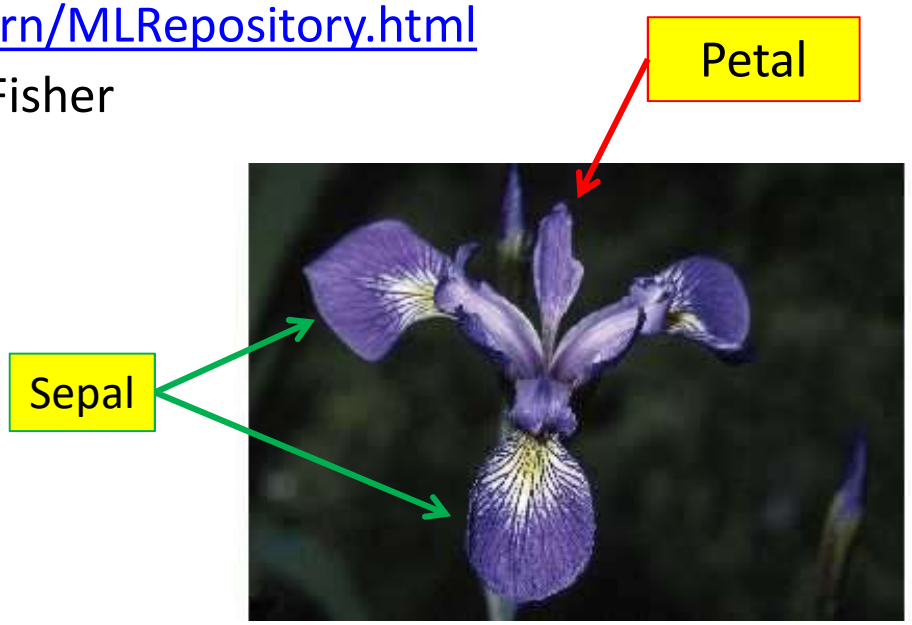
# Techniques Used In Data Exploration

- **EDA is an approach for data analysis that employs a variety of techniques**
  - 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




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 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 a categorical attribute is the value that has the highest frequency
- The notions of frequency and mode are typically used with categorical data

frequency(vi) = number of objects with attribute value vi / total num of objects.

Mode = argmax{frequency(vi)};

# Example

| Class | Size | Frequency |
|---|---|---|
| Freshman | 200 | 0.33 |
| Sophomore | 160 | 0.27 |
| Junior | 130 | 0.22 |
| senior | 110 | 0.18 |

What is the mode of the class attribute?

Categorical attributes may have a small number of values and consequently, the mode and frequencies of these values can be interesting and useful.

# 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.

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

- Can the mean interpreted as the middle of a set of values?

- The **median** is defined as

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

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

# Measures of Spread: Range and Variance

- The **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.

Absolute average difference
Median absolute difference

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

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

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

# Example: Containership at Southampton

| ShipID | Speed | GT | Carrier |
|---|---|---|---|
| 1 | 19.4 | 161635 | O3 |
| 2 | 20.3 | 166269 | O3 |
| 3 | 21.1 | 173022 | O3 |
| 4 | 19.9 | 152991 | O3 |
| 6 | 21.0 | 170269 | O3 |
| **Range** | | | |
| **Mean** | 20.2 | 163479 | |
| **Median** | | | |
| **StdDev** | 0.72 | 7892 | |

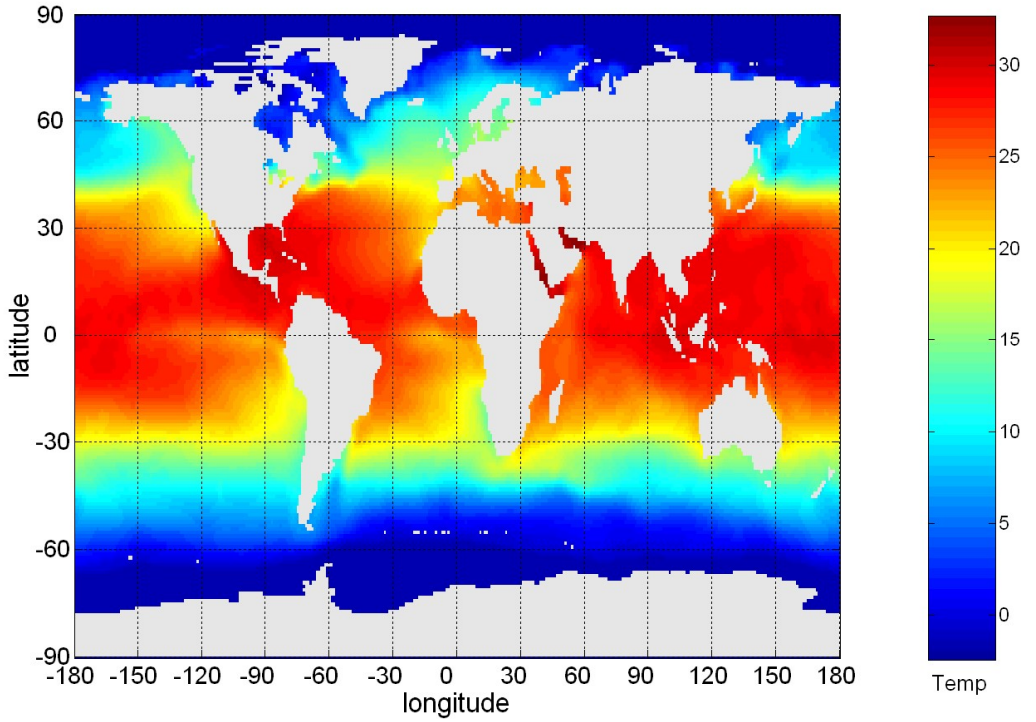
# Visualization

- 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
  - Make use of the domain knowledge that is locked up in people's heads
  - 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

- 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

- Arrangement is the **placement** of visual elements within a display
- It can make a large difference in how easy it is to understand the data.
- Example: 9 objects with 6 binary attributes,

| | 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 |

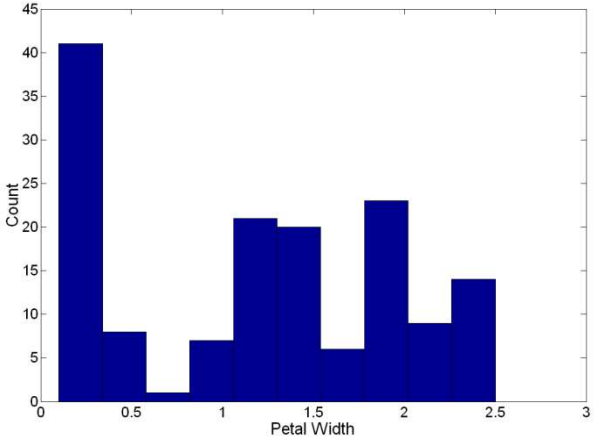
| | 6 | 1 | 3 | 2 | 5 | 4 |
|---|---|---|---|---|---|---|
| 4 | 1 | 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 |
| 6 | 1 | 1 | 1 | 0 | 0 | 0 |
| 8 | 1 | 1 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 1 | 1 |
| 3 | 0 | 0 | 0 | 1 | 1 | 1 |
| 9 | 0 | 0 | 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| 7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

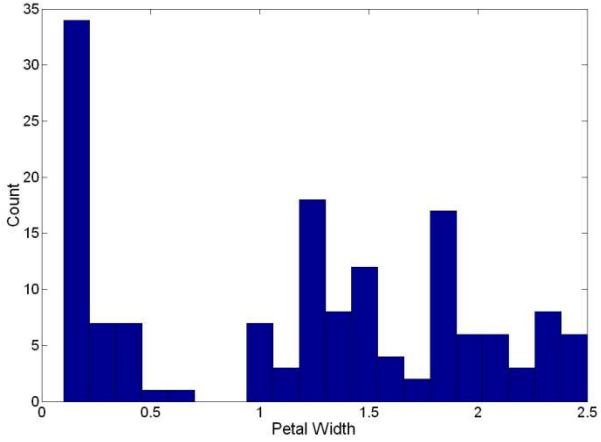
# Selection

- 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: Histograms

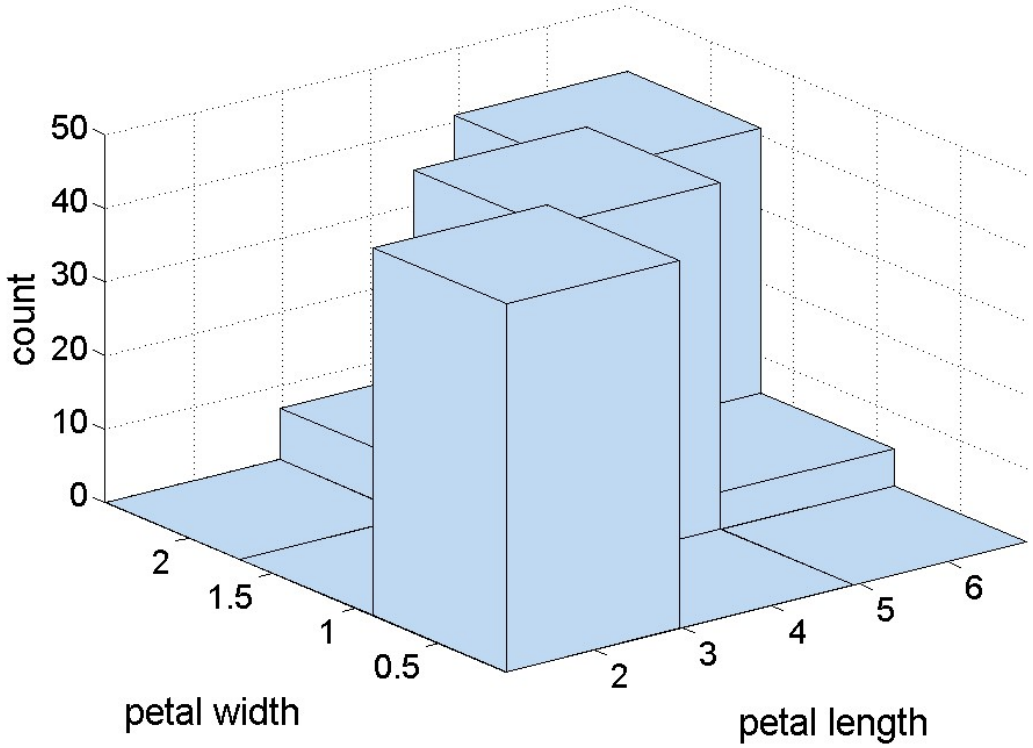
- Histogram
  - Usually shows the **distribution of values** of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal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



# Two-Dimensional Histograms

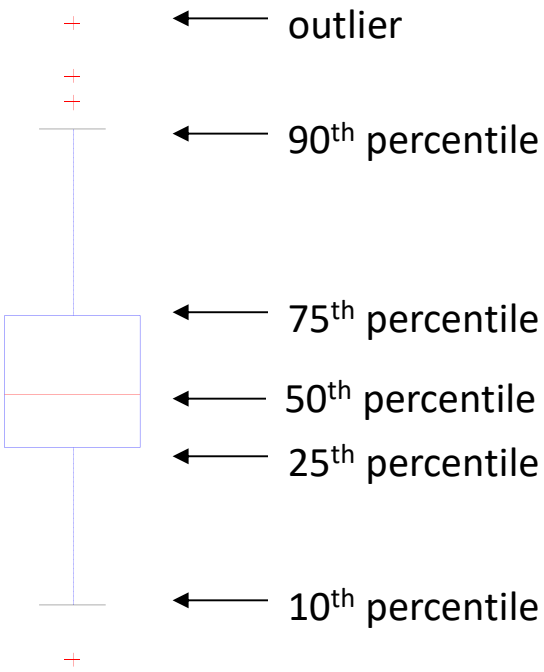
- Show the **joint distribution** of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



# Visualization Techniques: Box Plots

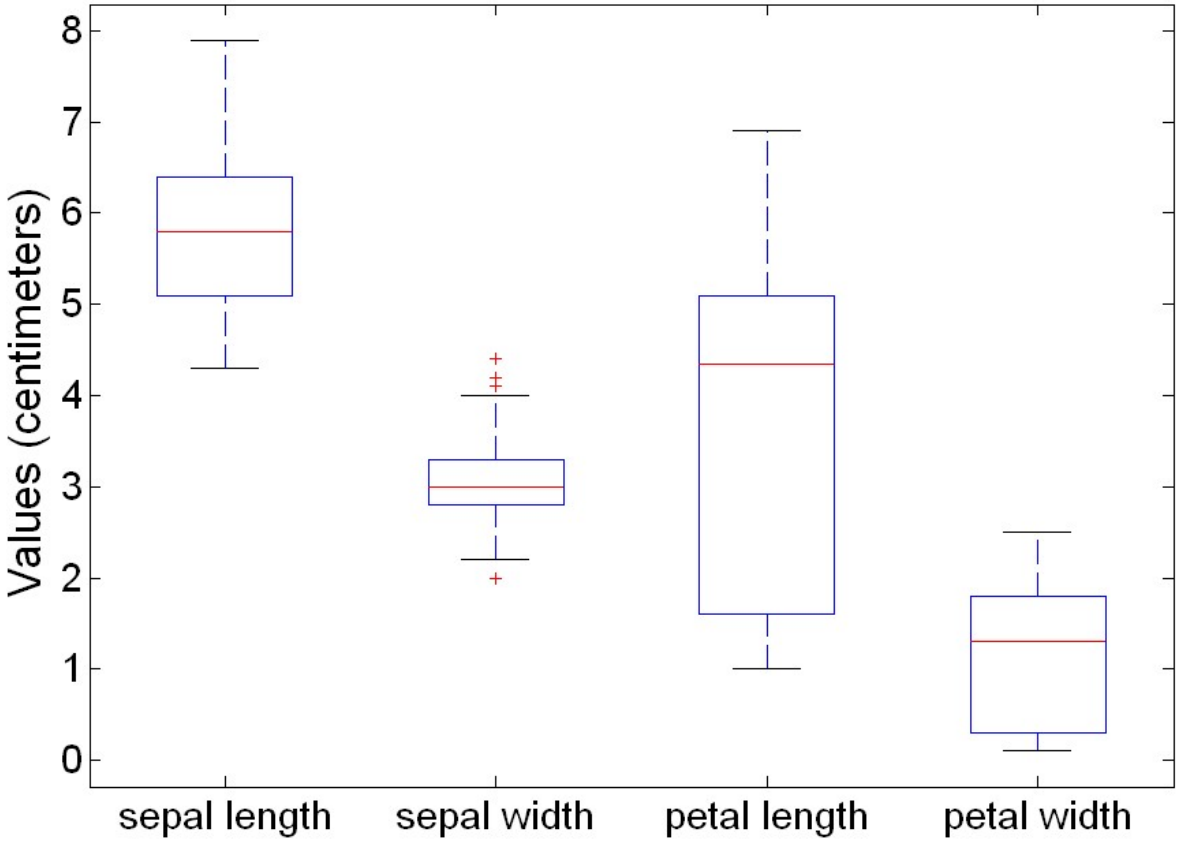
- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the **distribution of data**
  - Following figure shows the basic part of a box plot

A labeled box plot for sepal length



# Discuss: Box Plots

- Box plots can be used to compare attributes



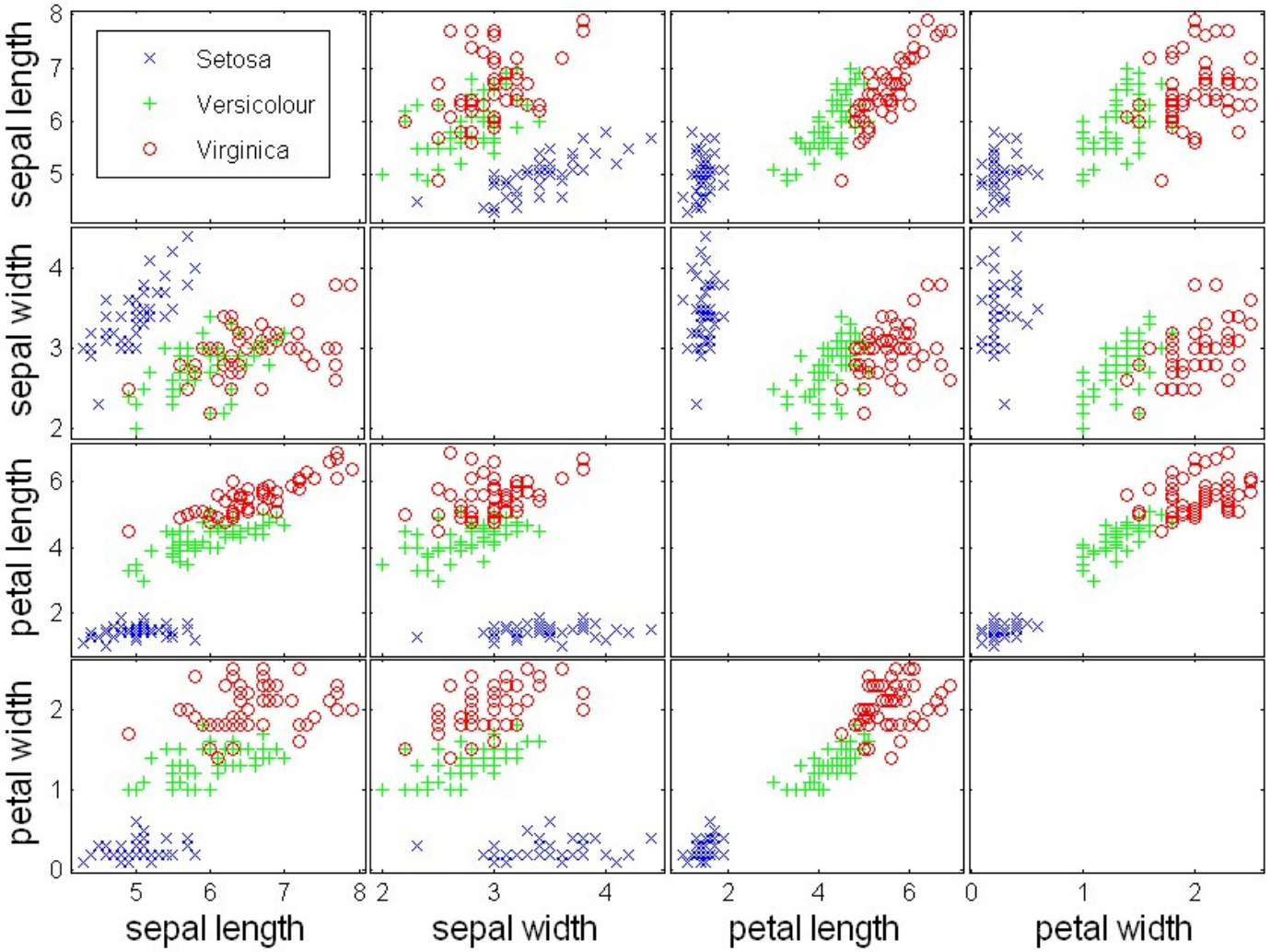

Box plot for Iris attributes

How a box plot can give information about symmetry?

# 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

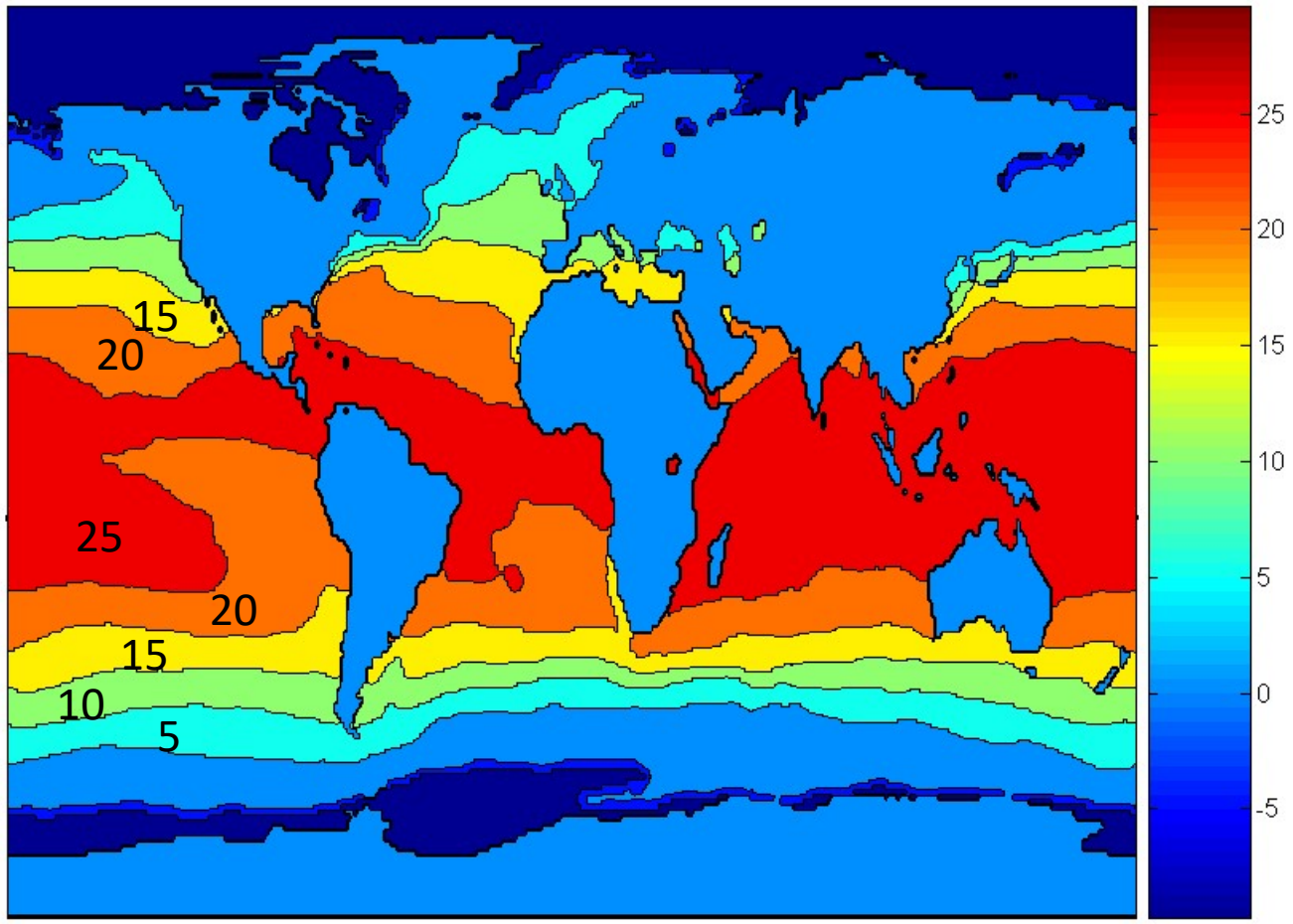
# Scatter Plot Array of Iris Attributes



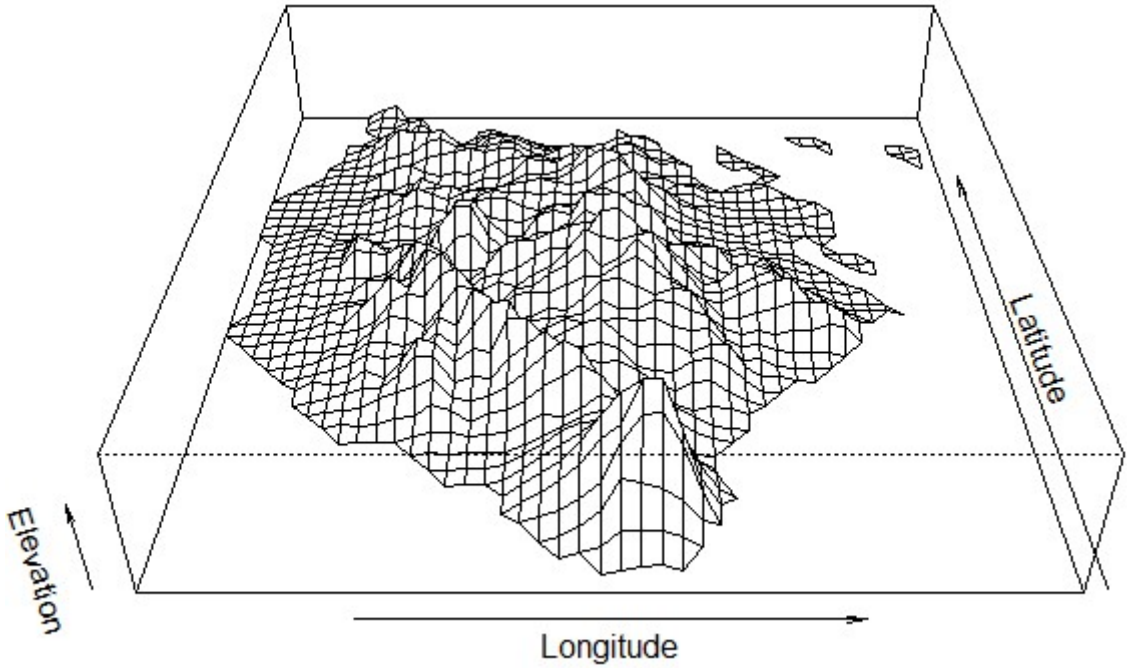
# Visualization Techniques: Contour Plots

- 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.

# Contour Plot Example: SST Dec, 1998



# Visualization Techniques: Surface Plots




3D view of the elevation in Marinduque, Philippines

Often used to describe mathematical functions or physical surfaces

# Example: Containership at Southampton-- Visualisation

| Speed | GT | arrDelay | depDelay |
|---|---|---|---|
| 19.4 | 161635 | 0.30 | 0.34 |
| 20.3 | 166269 | 0.17 | 0.17 |
| 21.1 | 173022 | 0.04 | 0.14 |
| 19.9 | 152991 | 0.27 | 0.08 |
| 21.0 | 170269 | 0.06 | 0.05 |
| 20.0 | 165343 | 0.34 | 0.59 |
| 21.9 | 177991 | 0.08 | 0.24 |
| 19.5 | 140259 | 0.34 | 0.56 |
| 15.2 | 131332 | 2.08 | 1.74 |
| 21.1 | 173022 | 0.05 | 0.18 |
| 21.4 | 175343 | 0.05 | -0.04 |
| 21.0 | 170269 | 0.04 | 0.07 |
| 15.0 | 125688 | 2.08 | 2.06 |
| … | … | … | … |

# On-Line Analytical Processing

- On-Line Analytical Processing (OLAP) was proposed by E. F. Codd in 1993, 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.

**Typical applications of OLAP** include business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting.

# 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.

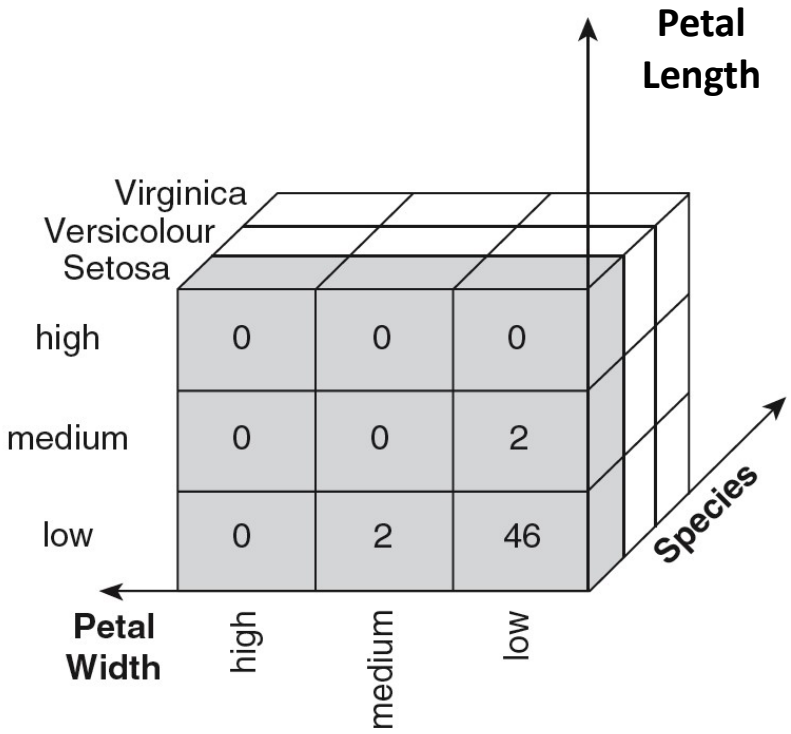
# Example: Iris data

- 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 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 |

# 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.



# Case: Online Brand Management

# Summary

- Data exploration
- Techniques used in data exploration
- Summary statistics
- Visualization:
  - Representation, arrangement, selection
  - Visualization techniques

    Histogram; Box; Scatter; Contour; Surface plot
- Online analytical processing (OLAP)