# Data Exploration

Dr. Meng Qu
Rutgers University

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

# 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

## Napoleon's March to Moscow    The War of 1812

Charles Joseph Minard

This classic of Charles Joseph Minard (1751–1870), the French engineer, shows the terrible fate of Napoleon's army in Russia. Described by E. J. Marey as seeming to defy the pen of the historian by its brutal eloquence, this combination of data map and time-series, drawn in 1869, portrays the devastating losses suffered in Napoleon's Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen River, the thick band shows the size of the army (422,000 men) as it invaded Russia in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn.

Edward R. Tufte, *The Visual Display of Quantitative Information*    Graphics Press, Box 430 Cheshire, Connecticut 06410

# 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.

# Iris Flower



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 'sex' 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 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.

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

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

# Harmonic Mean

- It is proper for situations when the average of rates is desired

The harmonic mean $H$ of the positive real numbers $x_1, x_2, ..., x_n > 0$ is defined to be

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n \cdot \prod_{j=1}^{n} x_j}{\sum_{i=1}^{n} \frac{\prod_{j=1}^{n} x_j}{x_i}}.$$

- The harmonic mean of the precision and the recall is often used as an aggregated performance score for the evaluation of algorithms and systems: the F-score (or F-measure). (gives equal weight to each data point)

- Arithmetic mean >= Geometric Mean >= Harmonic Mean

# Geometric Mean

- Indicates the central tendency or typical value of a set of numbers.

More generally, if the numbers are $x_1, \ldots, x_n$, the geometric mean $G$ satisfies

$$G = \sqrt[n]{x_1 x_2 \cdots x_n},$$

and hence

$$\ln G = \frac{1}{n} \sum_{i=1}^{n} \ln x_i.$$

# Differences

- Arithmetic Mean simply computed by adding the three heights together and divided by three.
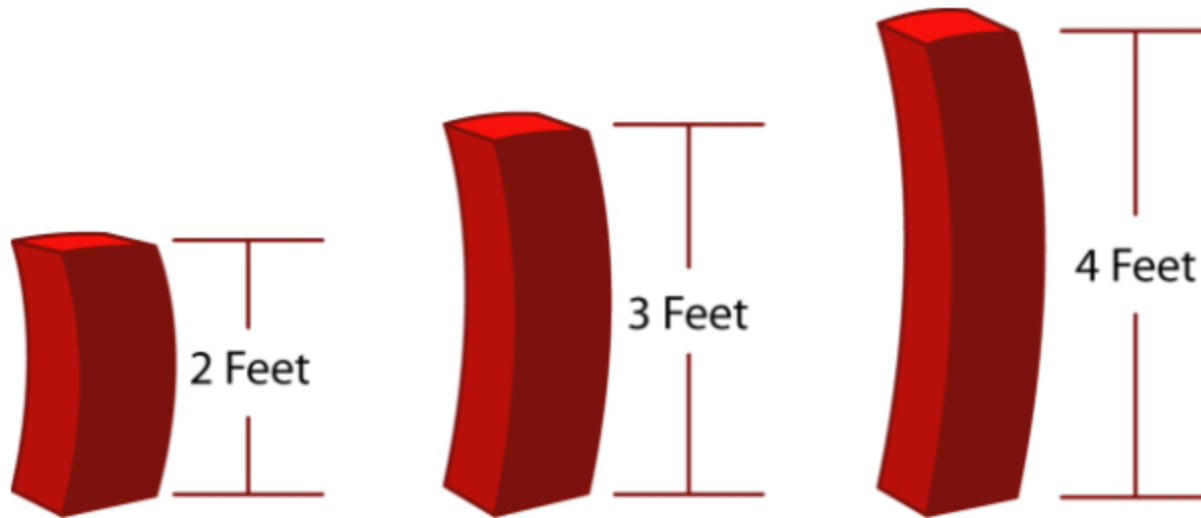


- Arithmetic mean is best used in situations where:
  – The data are not skewed
  – The individual data points are not dependent on each other

# Differences

- The harmonic average is less biased due to a small number of outliers.
    - Arithmetic mean=47.7
    - Harmonic mean=41



35    48    35    40    50    35    35    40    150    35    40    35    45    45
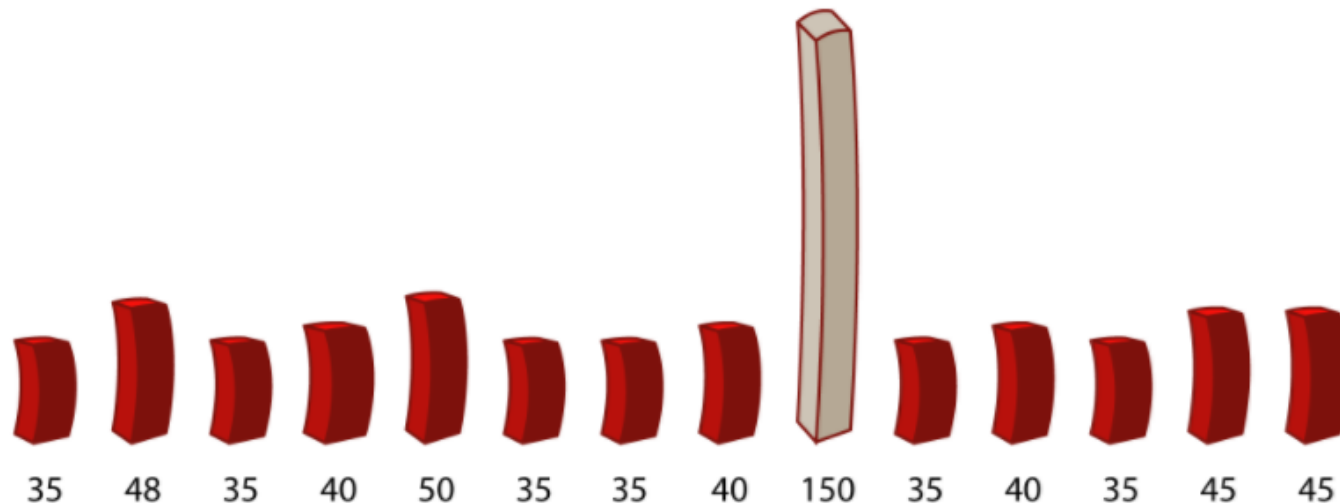
- Harmonic mean is best used in situations where:
    - A large population where the majority of the values are distributed uniformly but where there are a few outliers with significantly higher values.

# Differences

- Geometric mean should be used when ever the data are inter-related. Ex: returns on investment, interest rates.

  – An investment has the following returns over the next five years:

| 2010 | 2011 | 2012 | 2013 | 2014 |
|------|------|------|------|------|
| 5%   | 20%  | 25%  | -10% | 20%  |

  – If use arithmetic mean, the amount at the end of 2014=$17623.42

  – This is incorrect. The table below illustrates the actual cash flows the investment above.

| 2010   | 2011   | 2012   | 2013   | 2014   |
|--------|--------|--------|--------|--------|
| 5%     | 20%    | 25%    | -10%   | 20%    |
| 10,500 | 12,600 | 15,750 | 14,175 | 17,010 |

# 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.

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

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

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

# Coefficient of Variation (CV)

| LA1 | 375 | 487 | 301 | 905 | 248 | 759 | CV = 0.516 |
|-----|-----|-----|-----|-----|-----|-----|------------|

**CV: Coefficient of Variation**

for $x = \{x_0, x_1, ..., x_n\}$

$$CV = \frac{S_x}{\bar{x}}$$

Where $\bar{x}$ is the average of $x$, and
$S_x$ is the standard deviation of $x$

**CV is a measure of the dispersion. The larger the CV value; the greater the variability in the data.**

# 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
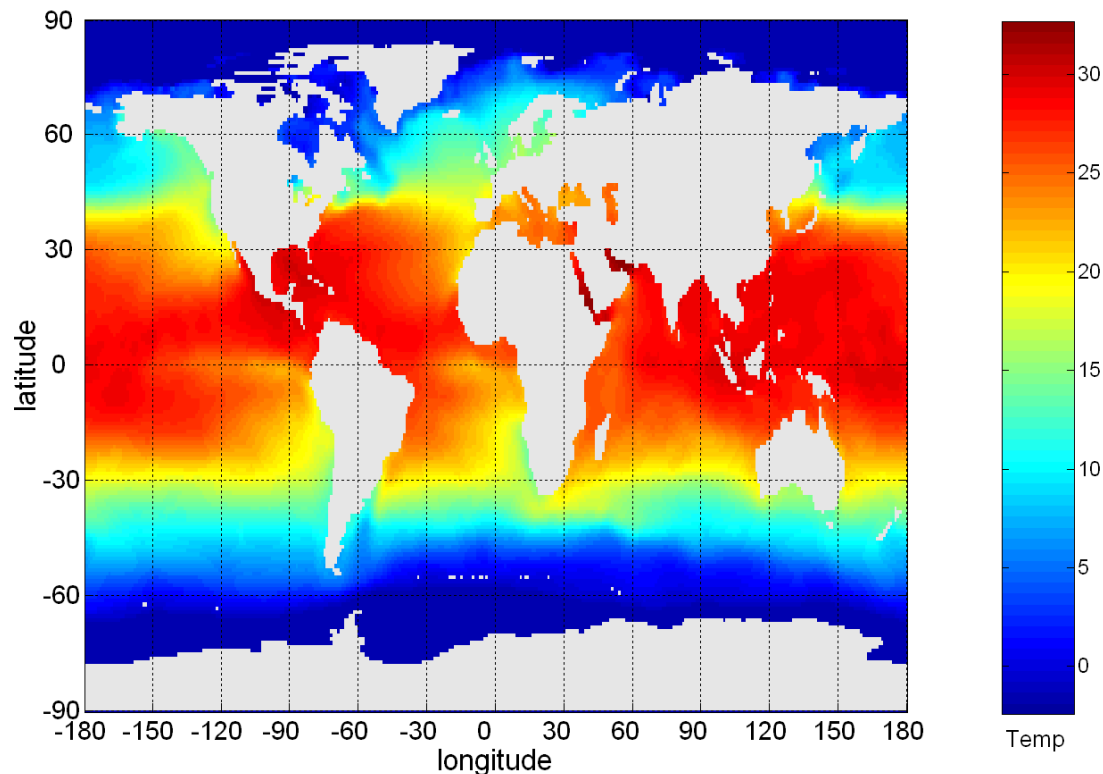
# Representation

- 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

- 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)
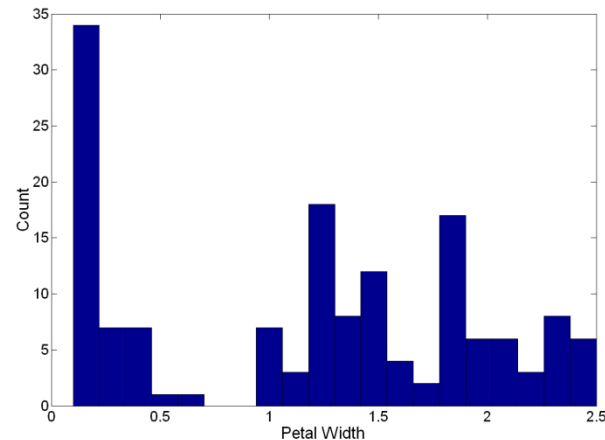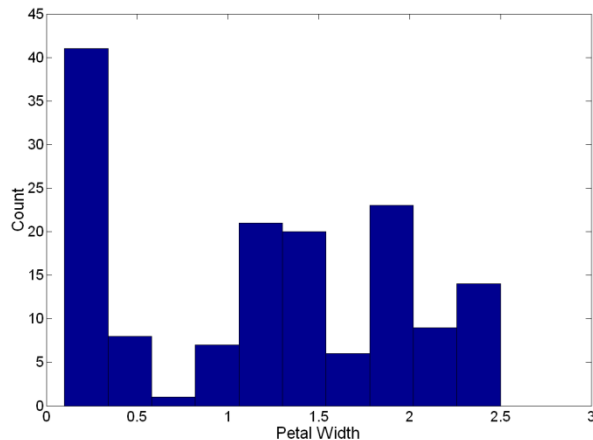
# Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes

- Example: petal width and petal length
  - What does this tell us?
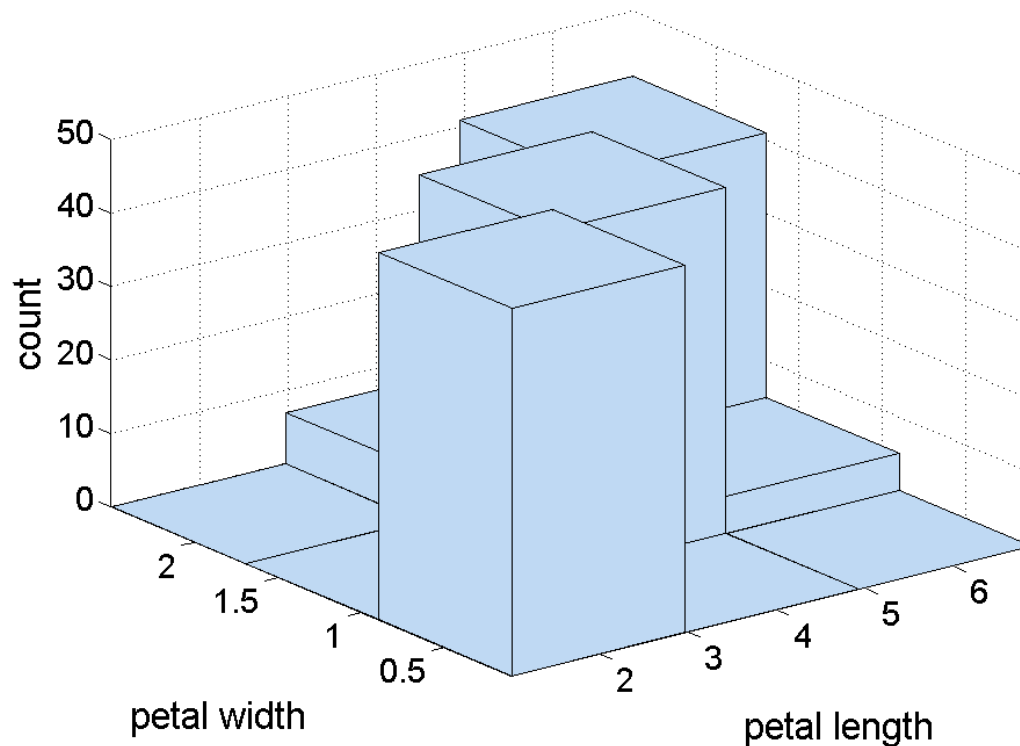
# Visualization Techniques: Box Plots

● Box Plots

- – Invented by J. Tukey

- – Another way of displaying the distribution of data

- – Following figure shows the basic part of a box plot

outlier ←

90th percentile ←

75th percentile ←

50th percentile ←

25th percentile ←

10th percentile ←
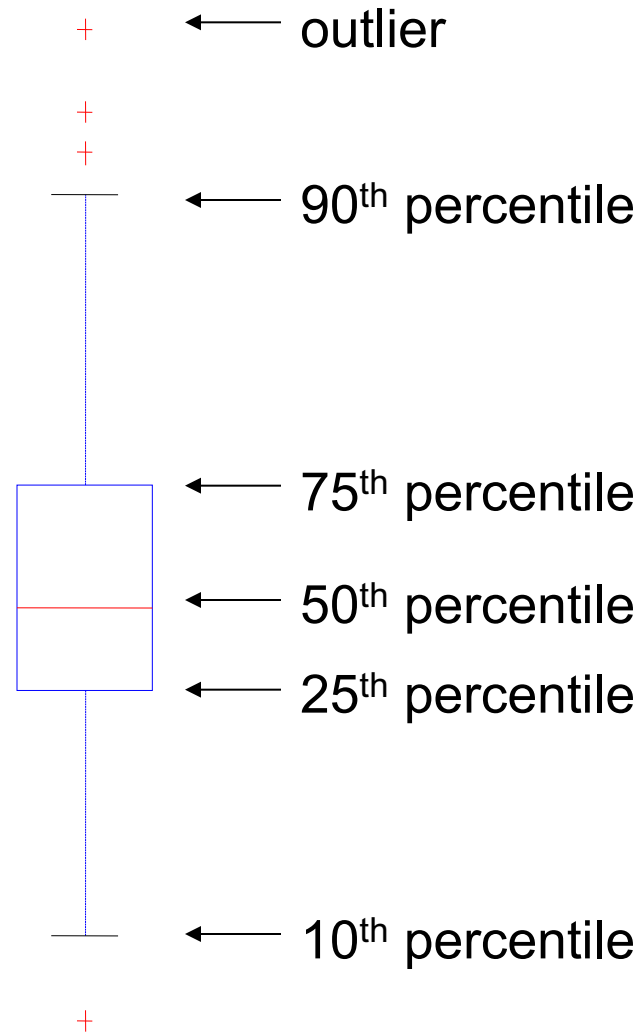
# Example of Box Plots

- Box plots can be used to compare attributes
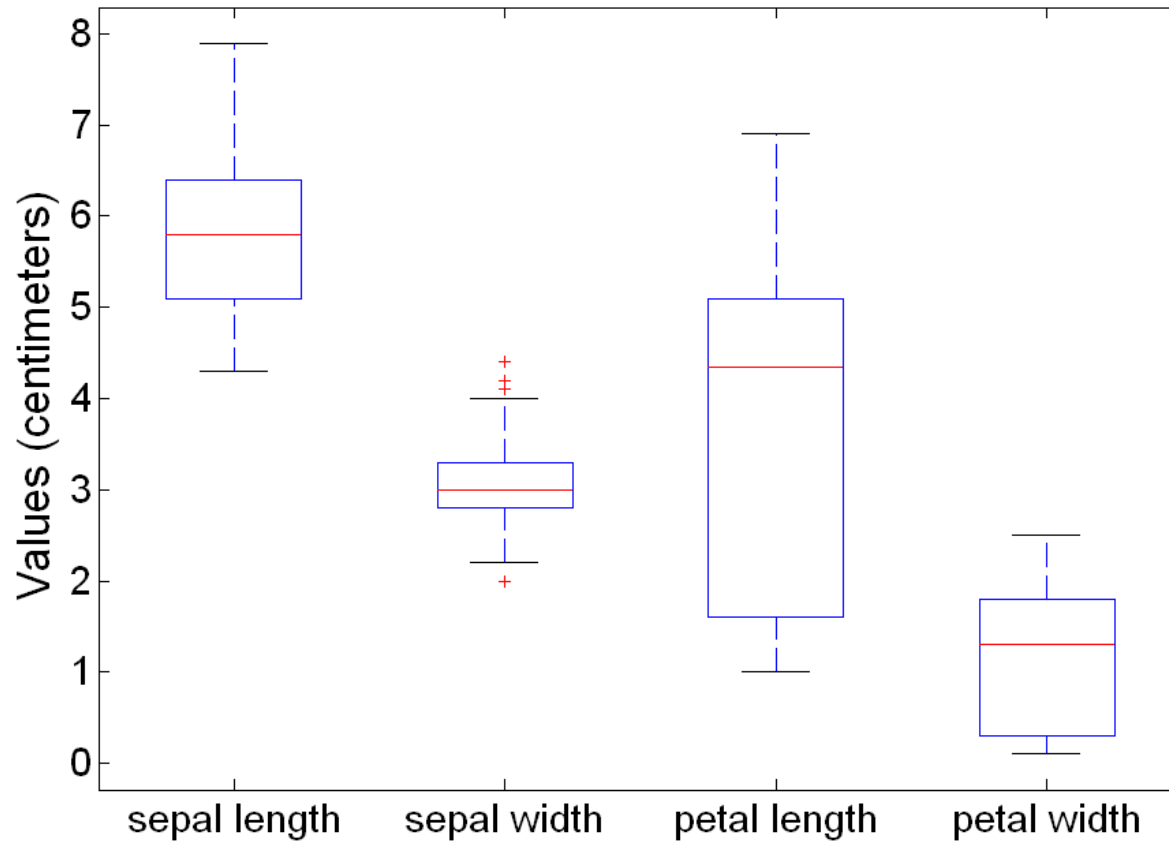
# Visualization Techniques: Scatter Plots

● 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
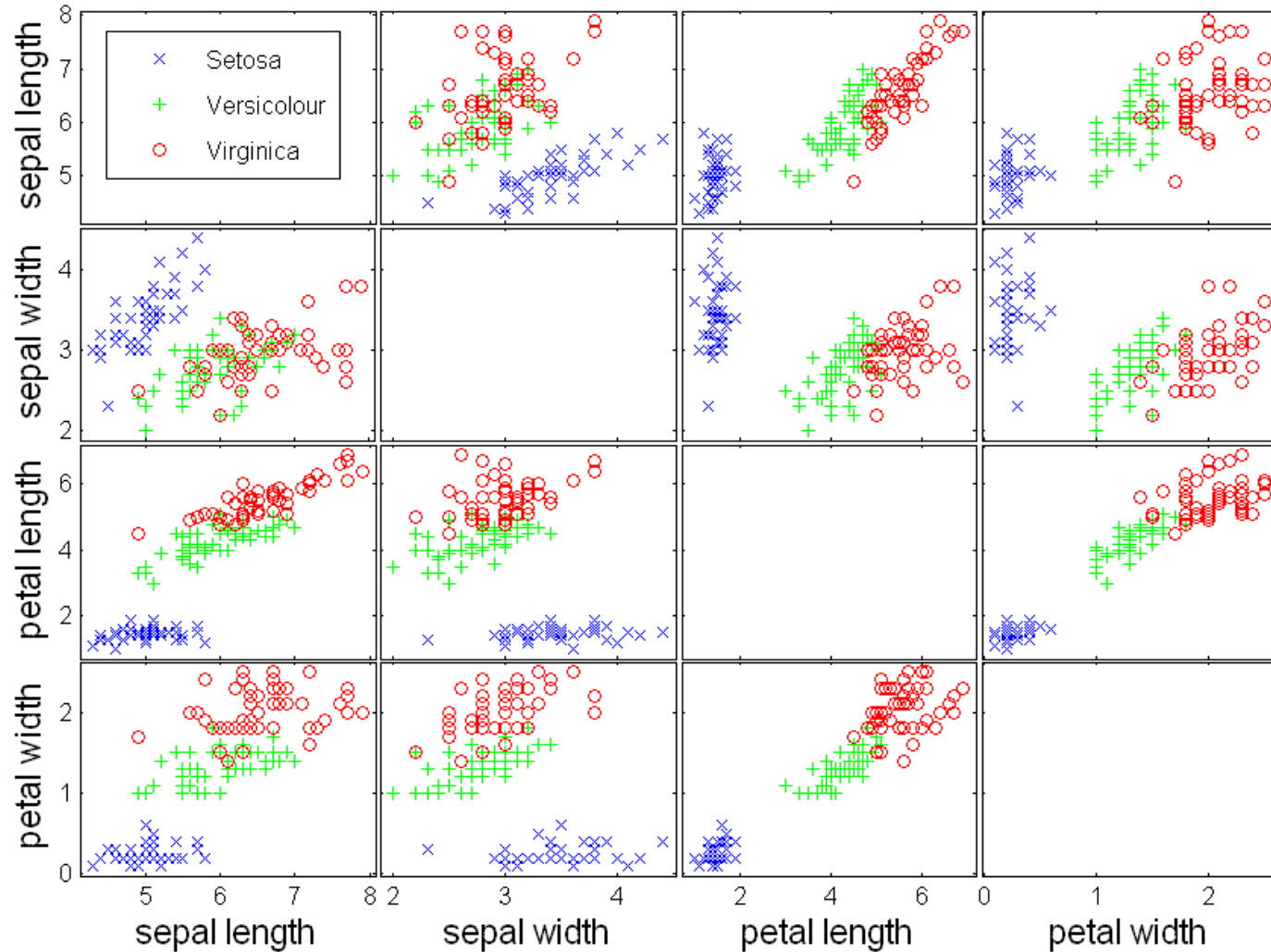
# Visualization Techniques: Contour Plots

● 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



Celsius
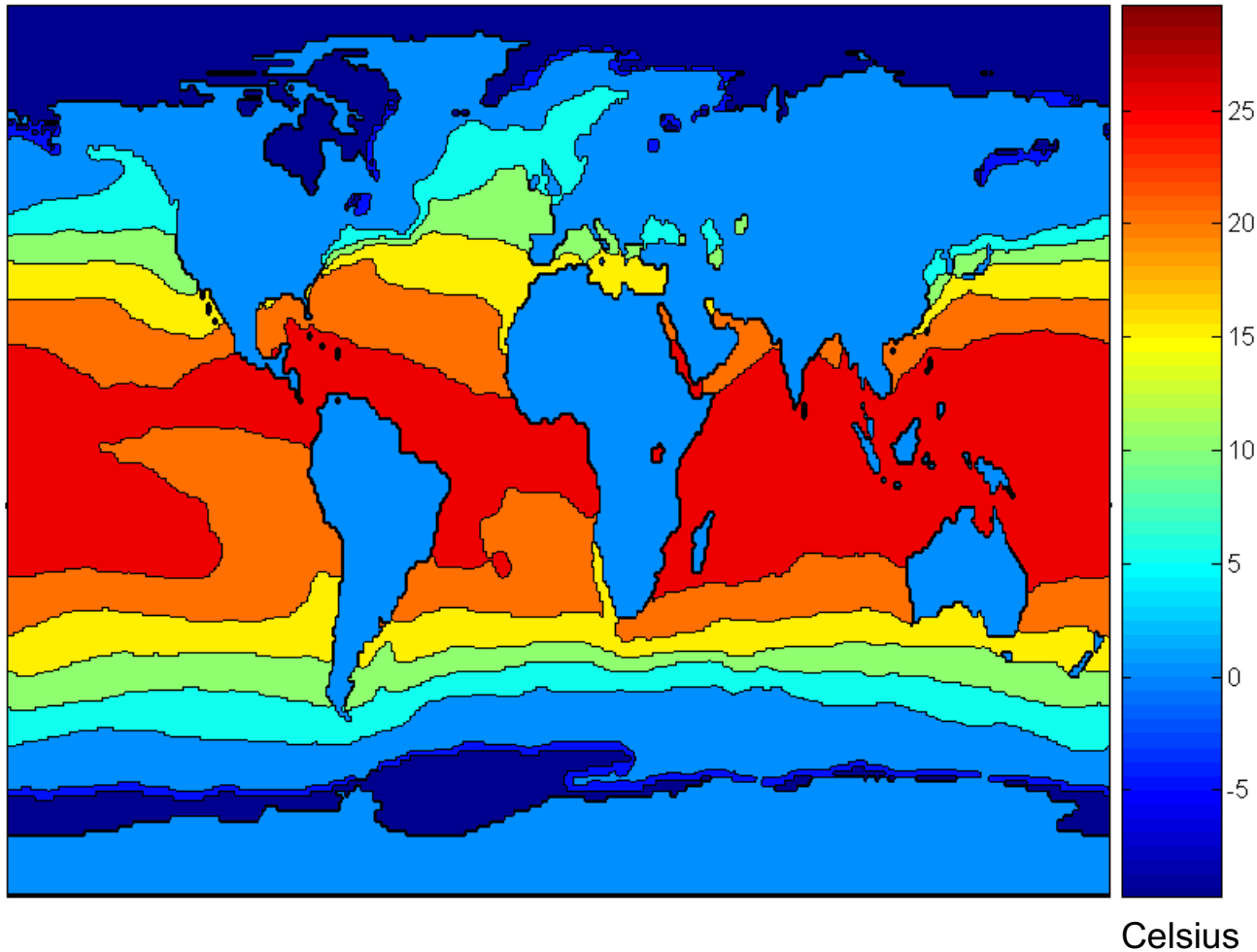
# Visualization Techniques: Matrix Plots

- 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
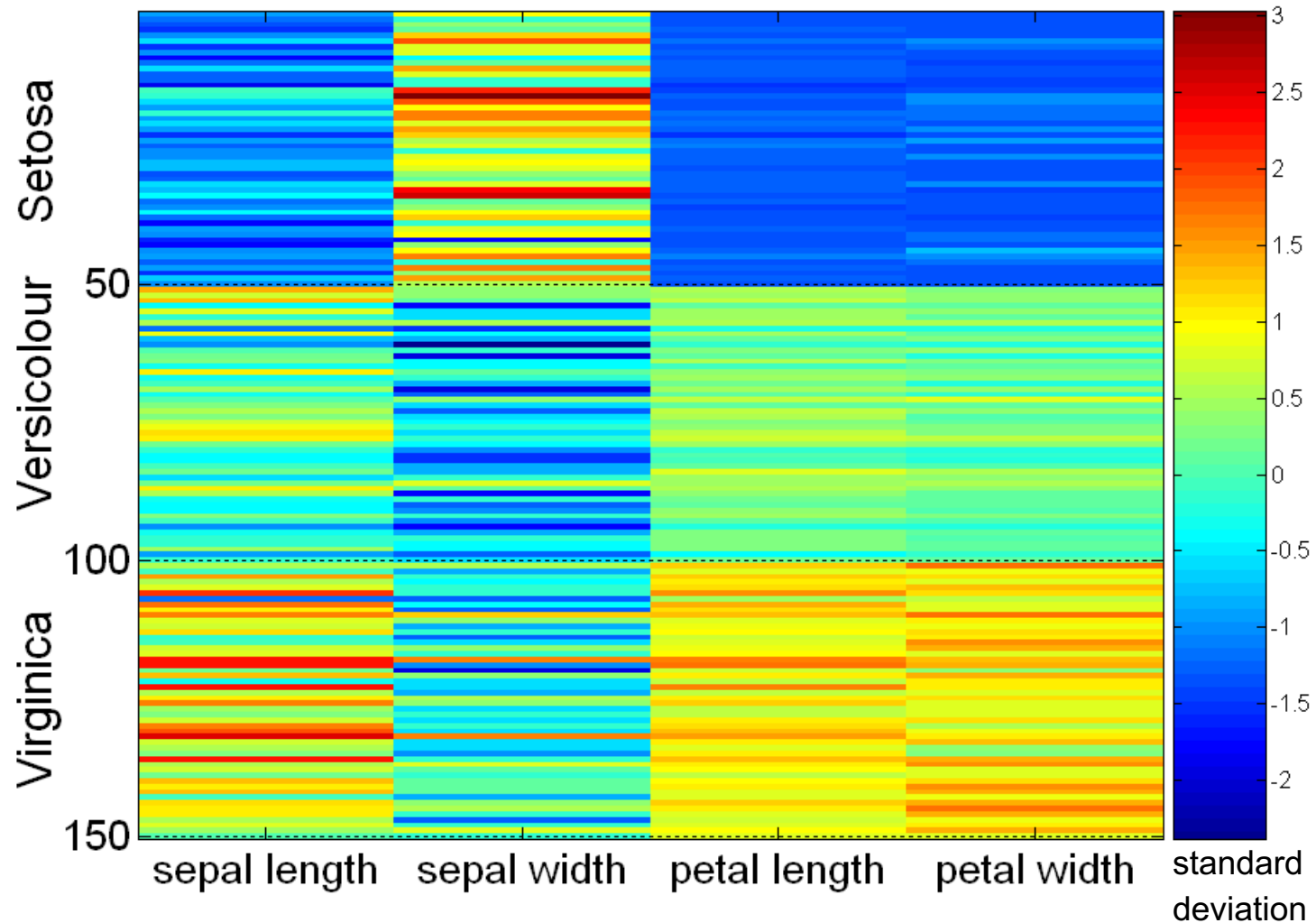
# Visualization of the Iris Correlation Matrix
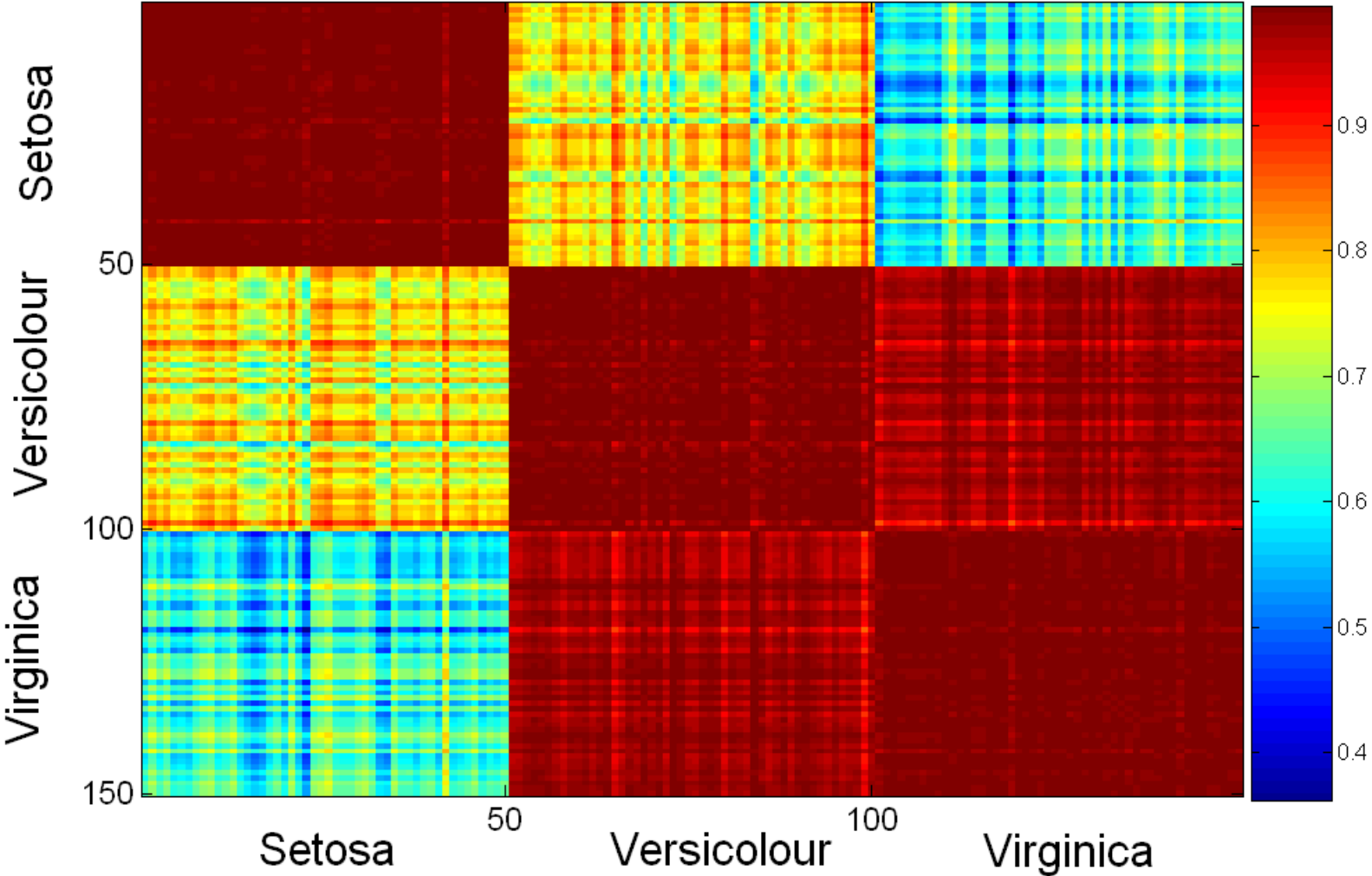
# Visualization Techniques: Parallel Coordinates

- ● 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
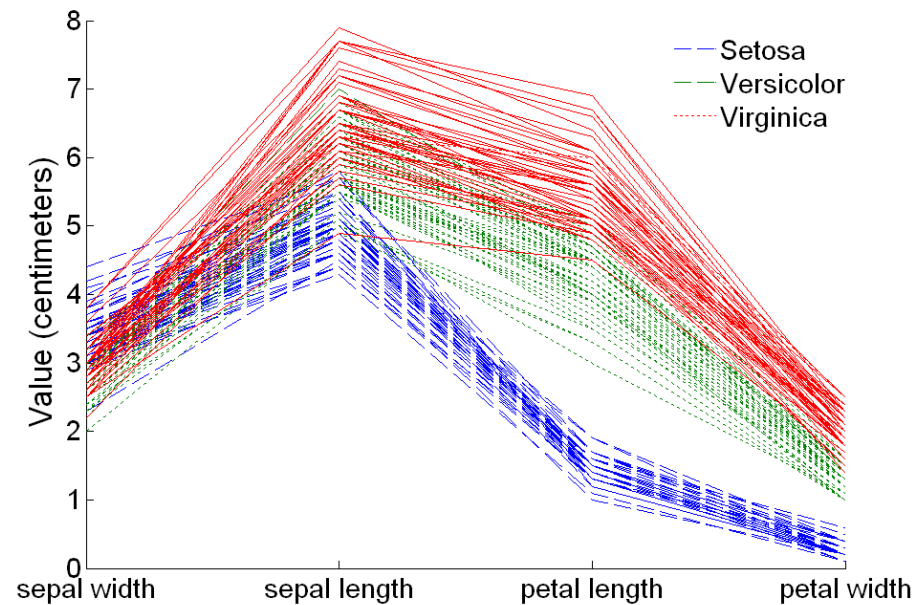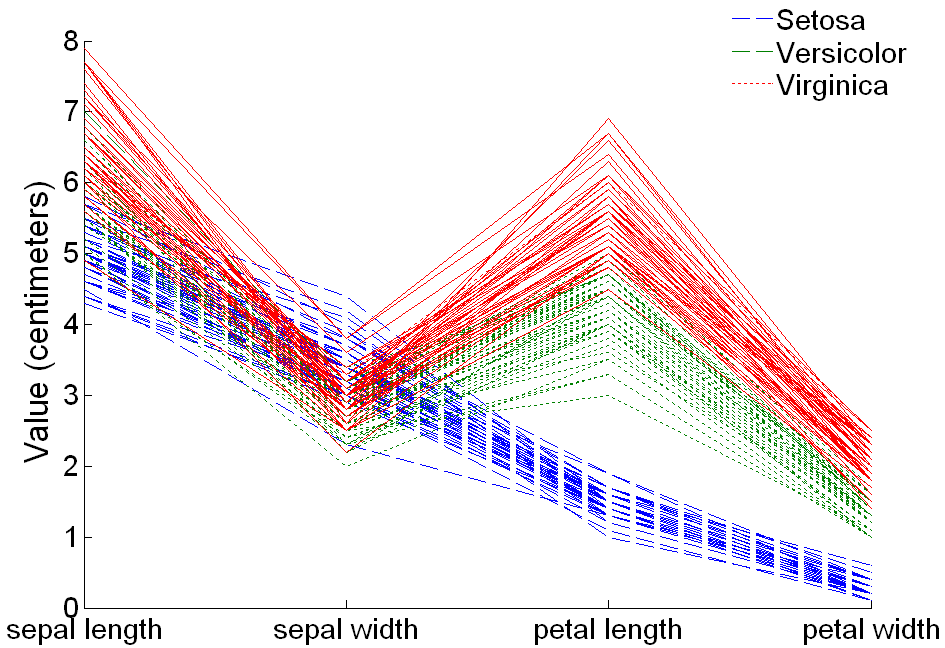
# Other Visualization Techniques

- ## 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
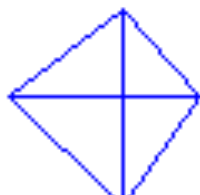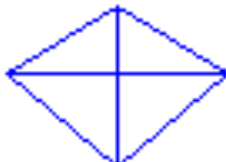
# Star Plots for Iris Data
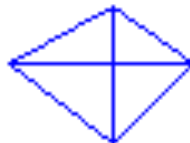
# Chernoff Faces for Iris Data

Setosa

1  2  3  4  5

Versicolour

51  52  53  54  55

Virginica

101  102  103  104  105
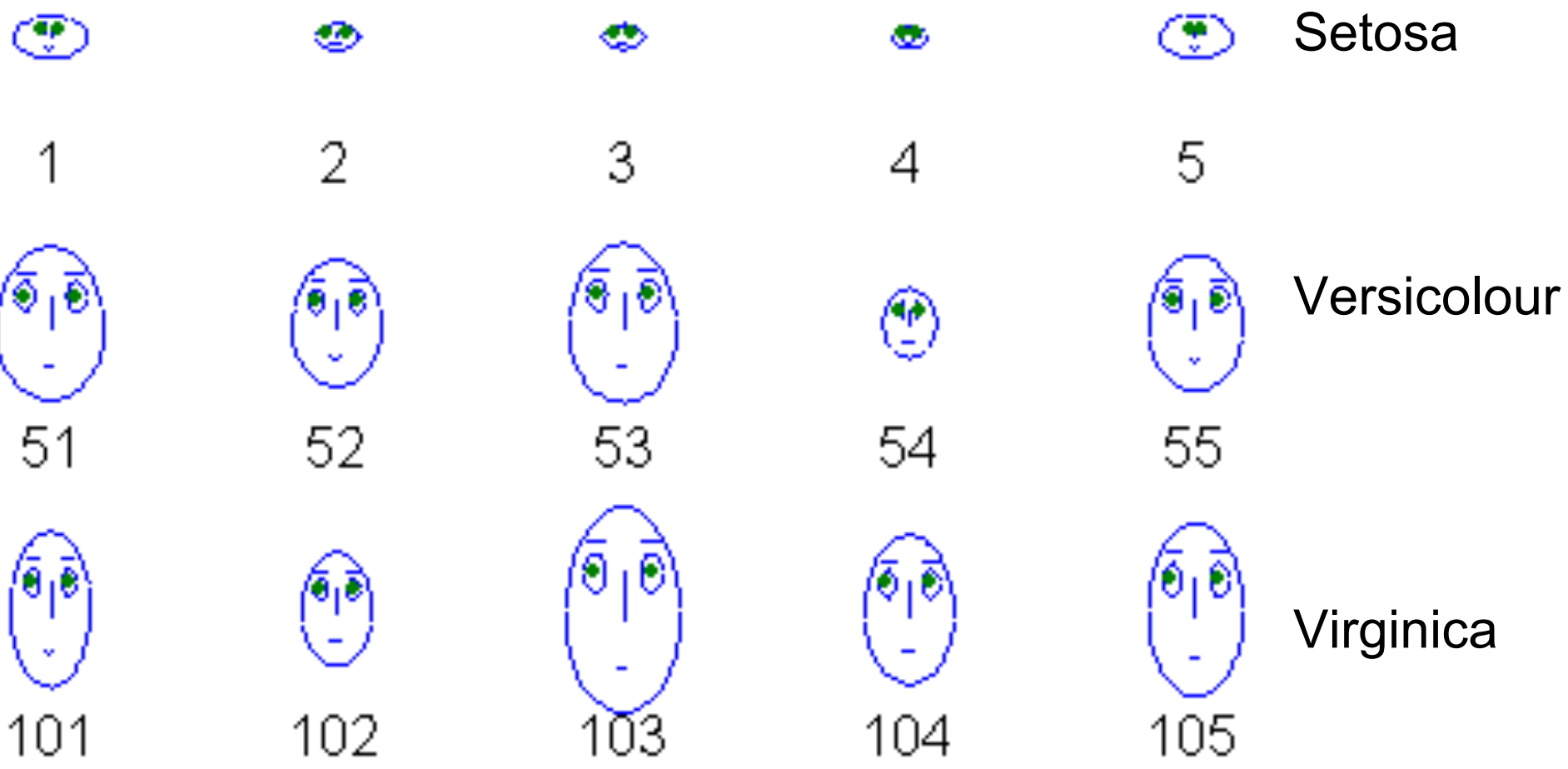
# Chernoff Faces for Iris Data

- The face mapped the features to the four features listed blow. Other feathers of the face, such as width between the eyes and length of the mouth, are given default values.

| Data Feature | Facial Feature |
|---|---|
| Sepal length | Size of Face |
| Sepal width | Forehead/jaw relative arc length |
| Petal length | Shape of forehead |
| Petal width | Shape of jaw |

# 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

- Converting tabular data into a multidimensional array:
  - Identify which attributes are to be the dimensions and which attribute is to be the target attribute
    - Attributes used as dimensions must have discrete values
    - Values of target variable appear as entries in the array
    - The target value is typically a count or continuous value
    - Can have no target variable at all except the count of objects that have the same set of attribute values
  - Find the value of each entry in the multidimensional array by summing the values (of the target attribute) or the count of all objects that have the attribute values corresponding to that entry.
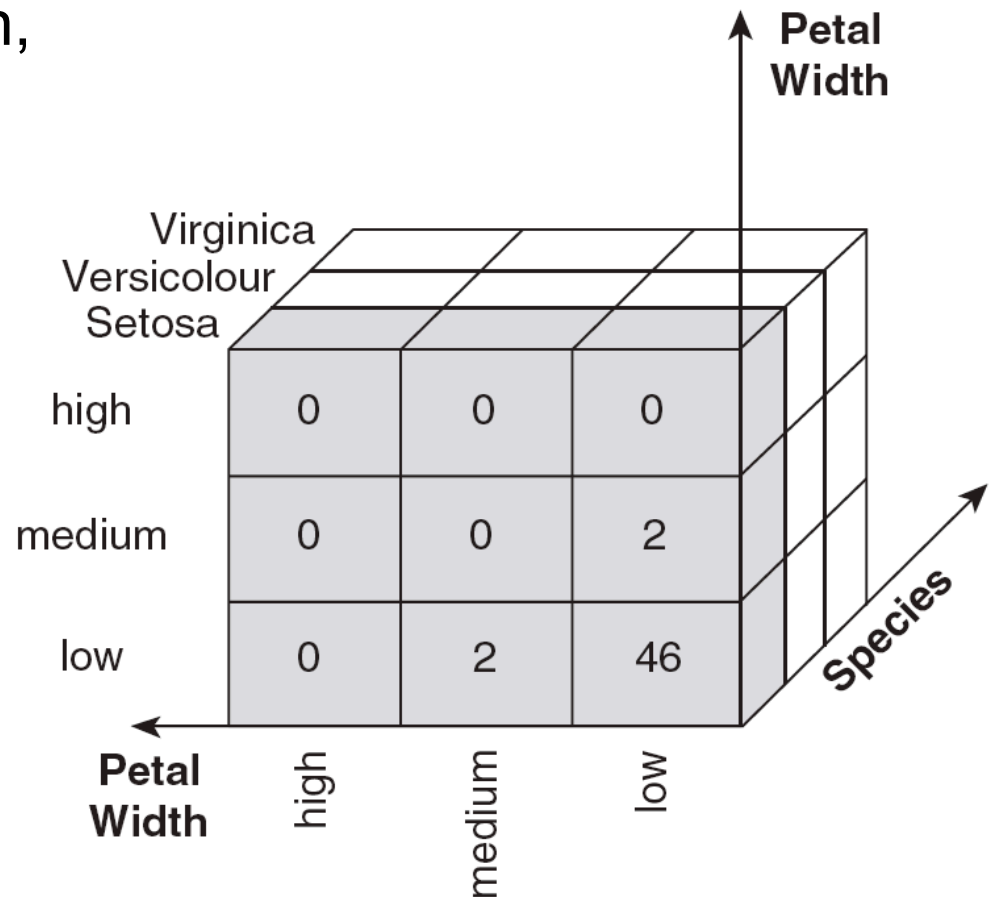
# Example: Iris data

● 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*

| Petal Length | Petal Width | Species Type | Coun |
|:---:|:---:|:---:|:---:|
| 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.

# Example: Iris data (continued)

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

|        | Width |        |      |
|--------|-------|--------|------|
|        | low   | medium | high |
| **Length** low | 46 | 2 | 0 |
| medium | 2 | 0 | 0 |
| high   | 0 | 0 | 0 |

|        | Width |        |      |
|--------|-------|--------|------|
|        | low   | medium | high |
| **Length** low | 0 | 0 | 0 |
| medium | 0 | 43 | 3 |
| high   | 0 | 2 | 2 |

|        | Width |        |      |
|--------|-------|--------|------|
|        | low   | medium | high |
| **Length** 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)

Date

Location

$ $ $
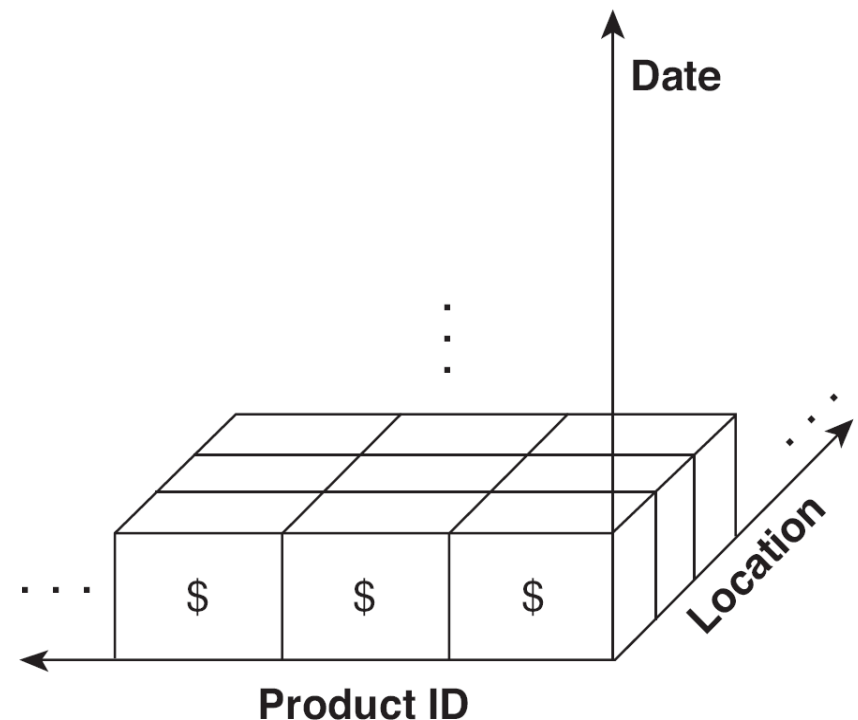
Product ID

# Data Cube Example (continued)

- The following figure table shows one of the two dimensional aggregates, along with two of the one-dimensional aggregates, and the overall total

|  | date | | | |
| --- | --- | --- | --- | --- |
| product ID | 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 |

# OLAP Operations: Slicing and Dicing

- Slicing performs a selection on one dimension of the given data cube.

- Dicing defines a subcube by performing a selection on two or more dimensions.

  - (location = "Toronto" or "Vacouver") and (time ="fall" or "spring).

- In practice, both operations can also be accompanied by aggregation over some dimensions.

# OLAP Operations: Roll-up and Drill-down

- Attribute values often have a hierarchical structure.

  - Each date is associated with a year, month, and week.

  - A location is associated with a continent, country, state (province, etc.), and city.

  - Products can be divided into various categories, such as clothing, electronics, and furniture.

- Note that these categories often nest and form a tree or lattice

  - A year contains months which contains day

  - A country contains a state which contains a city

# OLAP Operations: Roll-up and Drill-down

- This hierarchical structure gives rise to the roll-up and drill-down operations.

    – For sales data, we can aggregate (roll up) the sales across all the dates in a month.

    – Conversely, given a view of the data where the time dimension is broken into months, we could split the monthly sales totals (drill down) into daily sales totals.

    – Likewise, we can drill down or roll up on the location or product ID attributes.