



UNIVERSITAS
GADJAH MADA

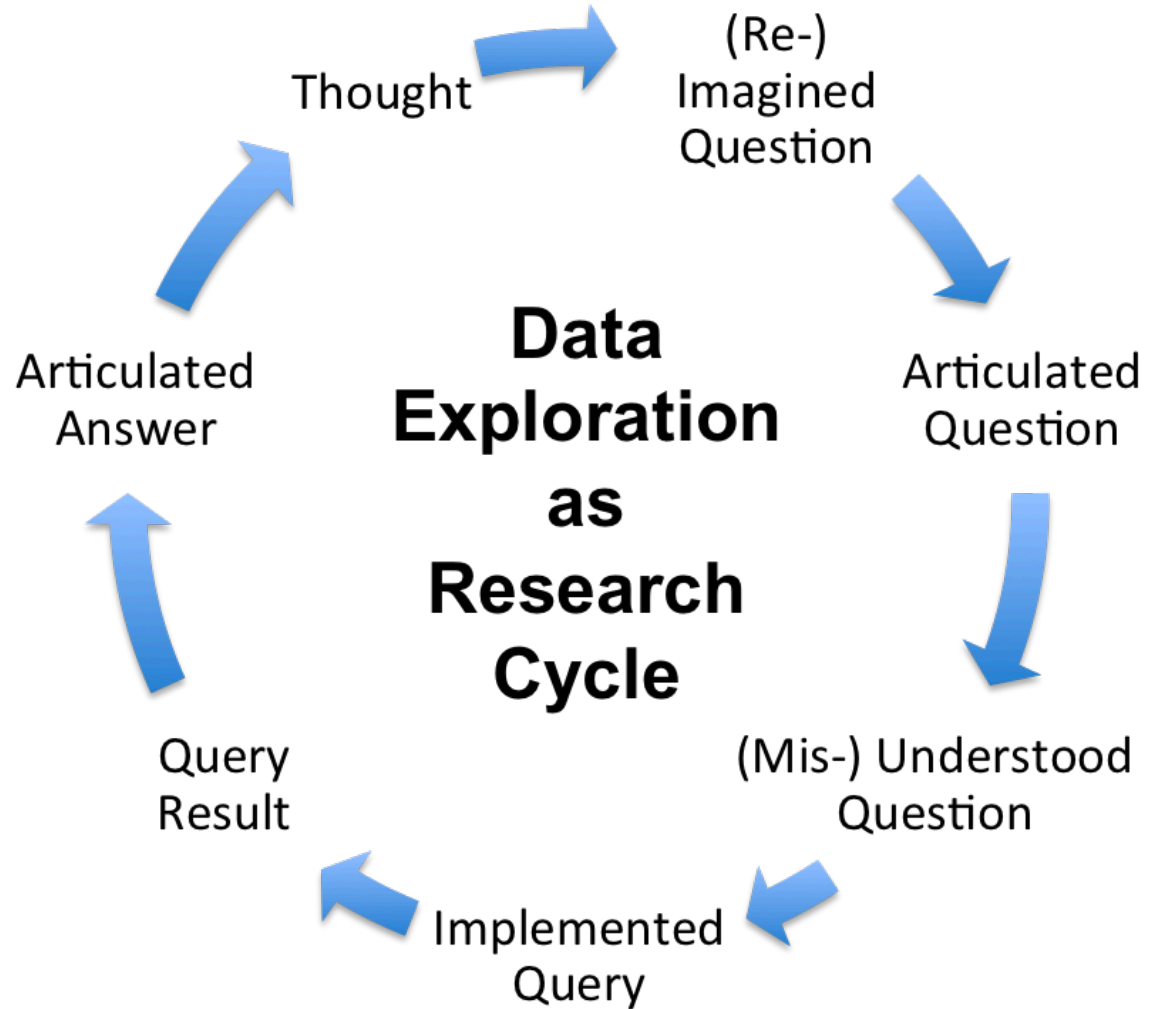
Descriptive & Prescriptive Analytic

Dr. Mardhani Riasetiawan

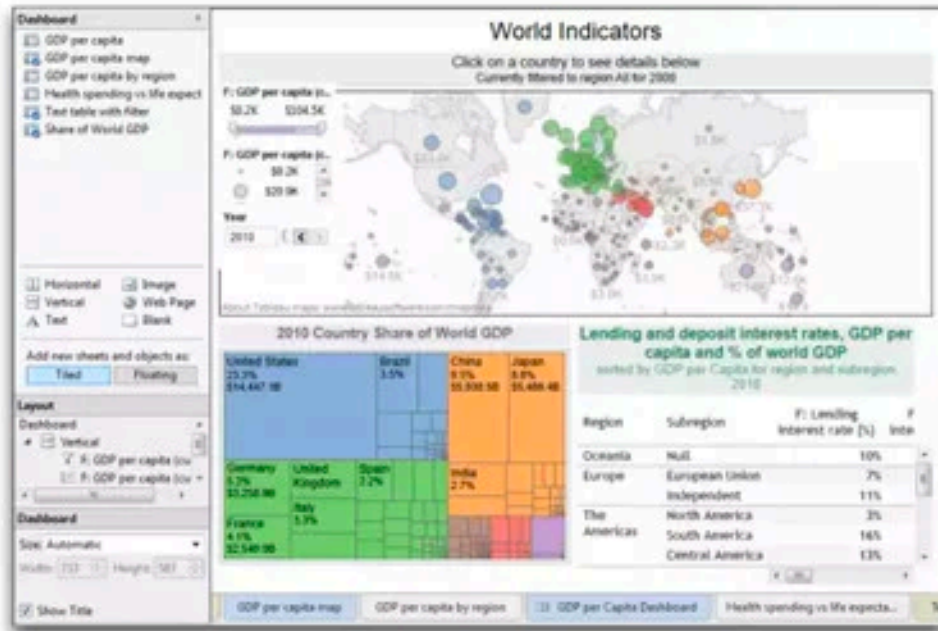
mardhani@ugm.ac.id | mardhani.staff.ugm.ac.id

Laboratorium Sistem Komputer dan Jaringan
Departemen Ilmu Komputer dan Elektronika
Fakultas Matematika dan Ilmu Pengetahuan Alam

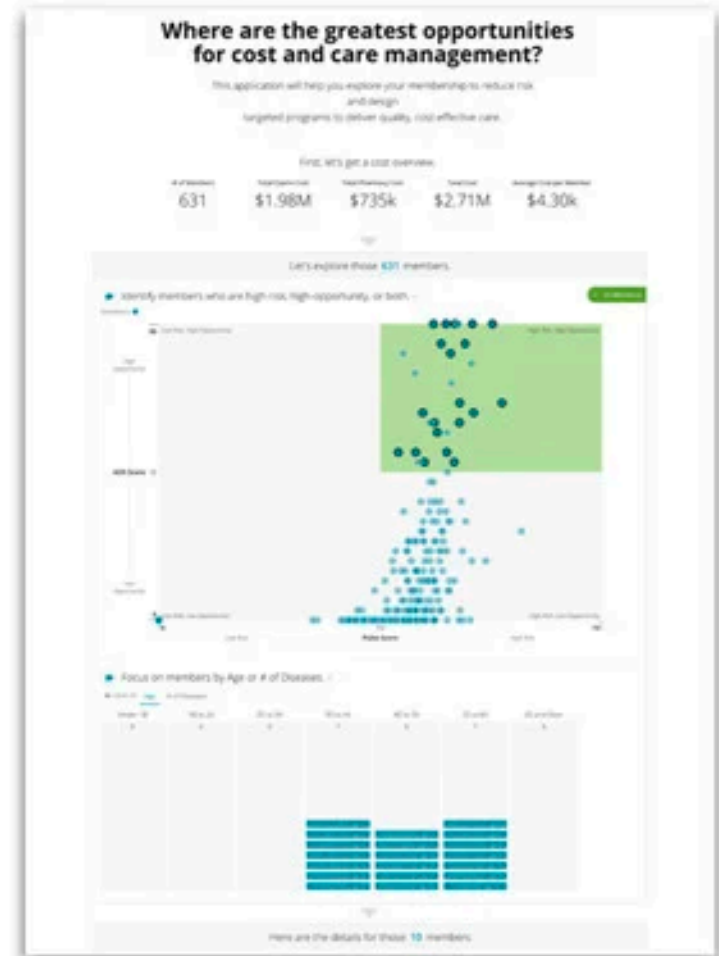
Universitas Gadjah Mada



data exploration



data presentation





UNIVERSITAS
GADJAH MADA

Data Quality



Data Quality includes examining data accuracy, consistency, completeness, and relevance.



Business Questions

Unknown

**Expanding,
Understanding
and Investigating**

**Exploration and
Discovery**

Known

**Foundational
Core**

**Establishing
Value**

Known

Unknown

Data

Source: Gartner (October 2017)



- Rejection or Confirmation of Hypothesis
- Effect Size & Significance

Decisions and Conclusions

- Confidence & Sampling Errors
- Correlations & Causations

Interpretations

- Mean, Median, Mode
- Dispersion, Skewness
- Correlation Coefficients

Quantitative Outputs

- Data Manipulations
- Descriptive Statistics
- Inferential Statistics

Quantitative Analysis

- SQL/NOSQL Database
- Big Data Technologies
- Optimal Queries

Data Management

Data Collection

- Random Sample of Subjects
- Collection of Observations

Hypotheses Formulation

- Hypothesizes Questions
- Null Hypothesis
- Alternative Hypothesis

Research Questions

- Formulates Research Problem
- Identifies Variables

Research Problem

- Purpose Statement
- Problem Statement

Data Science: An Iterative Process

Big Data Exploration

The big data landscape for most enterprises is a vast wilderness. It is a growing and complex ecosystem of different data types from multiple sources, including new data from social media and raw data collected from sources like sensors. Only after effectively exploring and navigating this terrain can businesses begin to mine and refine their data resources to extract value—using trusted information to pave the roads to new insights and smarter decision making.

Raw Data

Enterprise Data

External Data

WHAT DO YOU NEED TO SUCCEED?

Bird's Eye View

Visualize and understand all available data across systems and silos—both inside and outside the enterprise.



Useful Functionality

Establish the ability to access and use big data to support decision making and day-to-day operations.



Mine Valuable Information

Get the most value from your data by using big data tools to separate the most useful content from the rest.



Fool's Gold?

Ensure the veracity of all data and support confident decision making with enterprise-level big data management.



Elephant Crossing

Combine data stored in Hadoop with data in enterprise and external systems to drive better insights.



Discover Hidden Insights

Find insights in new and unstructured data while adding important context to raw data, advancing analytics and structured operational data.



Dynamic Directions

Focus employees' productivity with dynamic contextual views of the most relevant enterprise and external content. Include important analytics in context to help drive day-to-day decisions.



Avoid Pitfalls

Identify areas of risk by understanding the full scope of all data and protecting confidential and strategic information.

Tame Wild Big Data

- Reduce operational & IT costs
- Achieve greater process efficiencies
- Eliminate redundant systems
- Leverage existing knowledge
- Reduce risk & improve governance

THE RESULTS

Utilize Your Resources

- Improve decision making
- Develop new business models
- Gain market presence & revenue
- Accelerate innovation
- Improve customer knowledge
- Increase employee performance



UNIVERSITAS
GADJAH MADA

Exploratory Data Analytic



1. 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

0, 2, 3, 7, 8

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

11.

5

$$\text{standard_deviation}(x) = s_x$$

3.

- However, this is also sensitive to outliers, so that³other measures are often used.

2.8

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

(Mean Absolute Deviation) [Han]
(Absolute Average Deviation) [Tan]

1

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

(Median Absolute Deviation)

5

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



2. Visualization

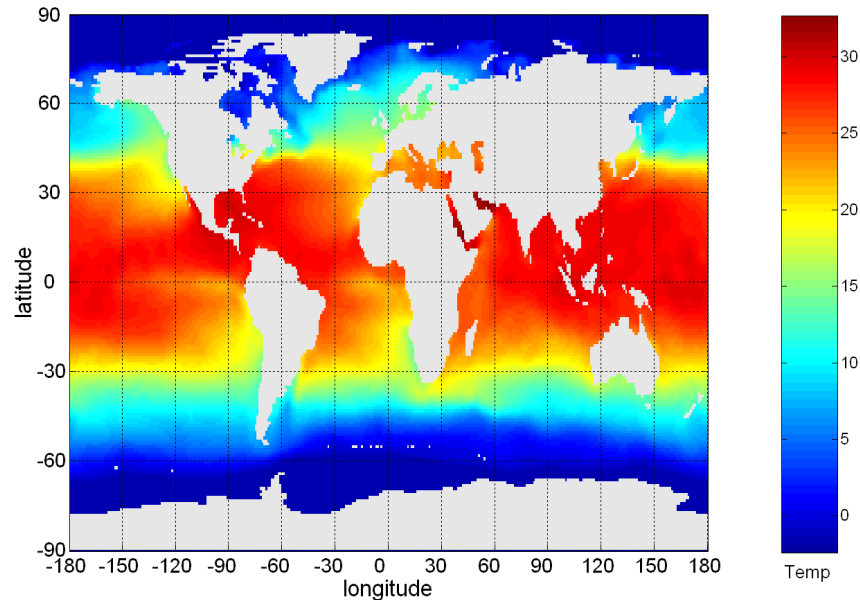
Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- 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.



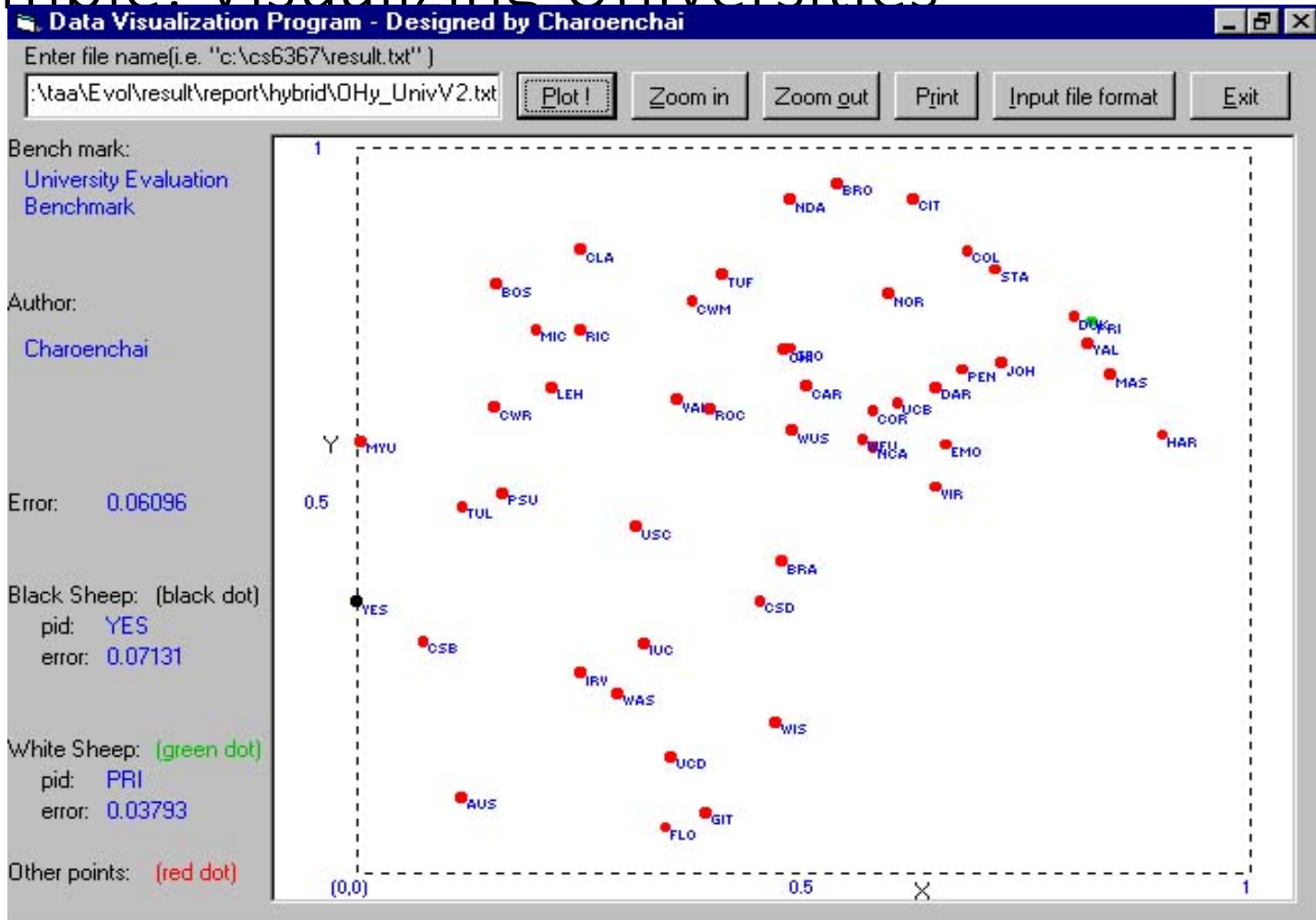
Arrangement

- 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

Example: Visualizing Universities





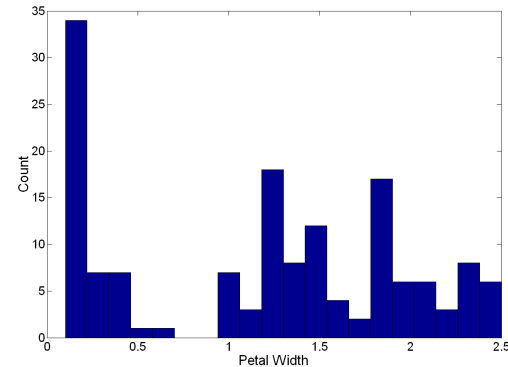
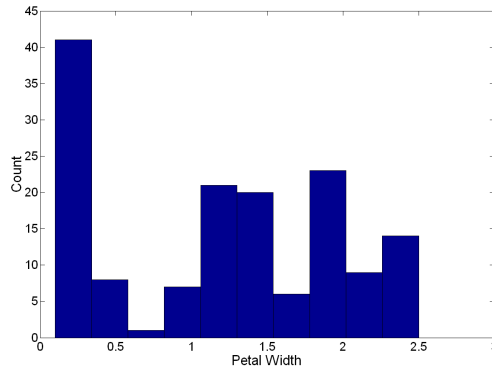
Selection

- 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

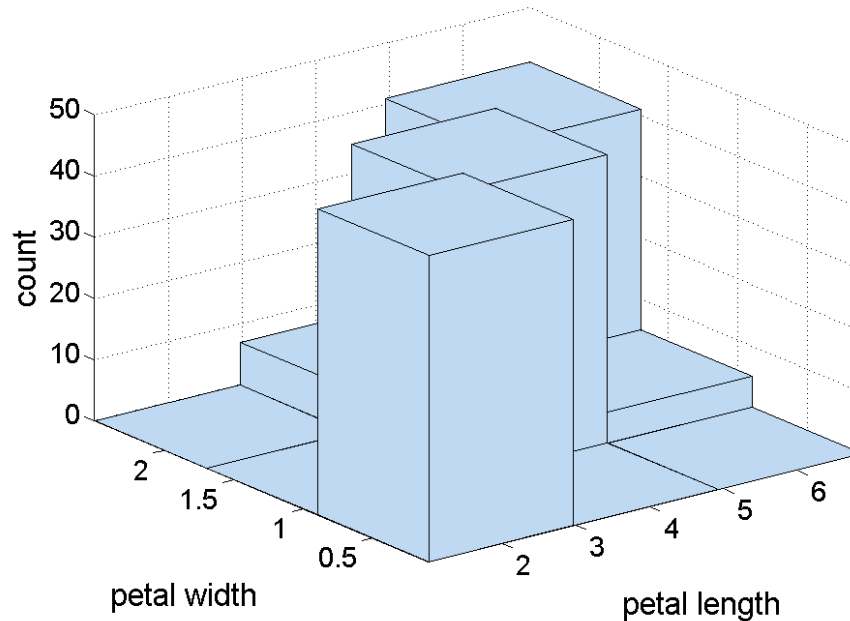
- Histogram
 - Usually shows the distribution of values of a single variable
 - Divide the values into bins and show a bar plot of the number of objects in each bin.
 - The height of each bar indicates the number of objects
 - Shape of histogram depends on the number of bins
- Example: Petal Width (10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





Two-Dimensional Histograms

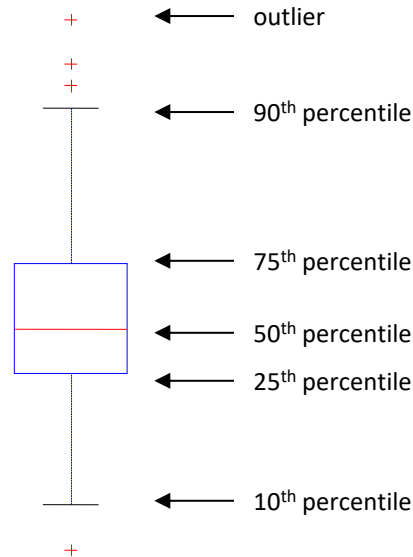
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
 - What does this tell us?





Visualization Techniques: Box Plots

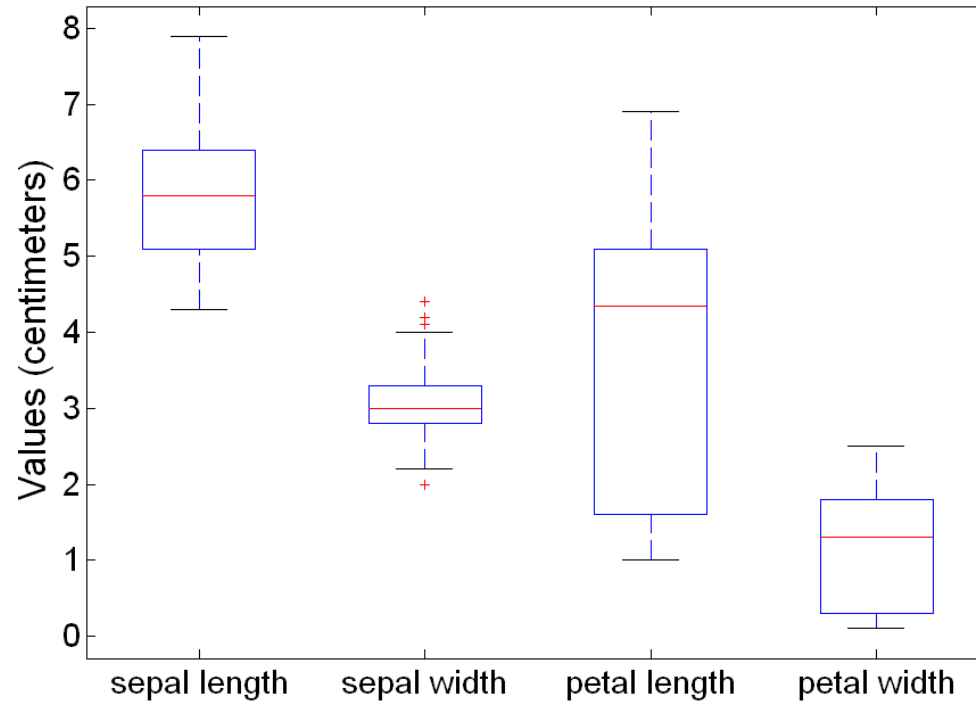
- Box Plots
 - Invented by J. Tukey
 - Another way of displaying the distribution of data
 - Following figure shows the basic part of a box plot





Example of Box Plots

- Box plots can be used to compare attributes





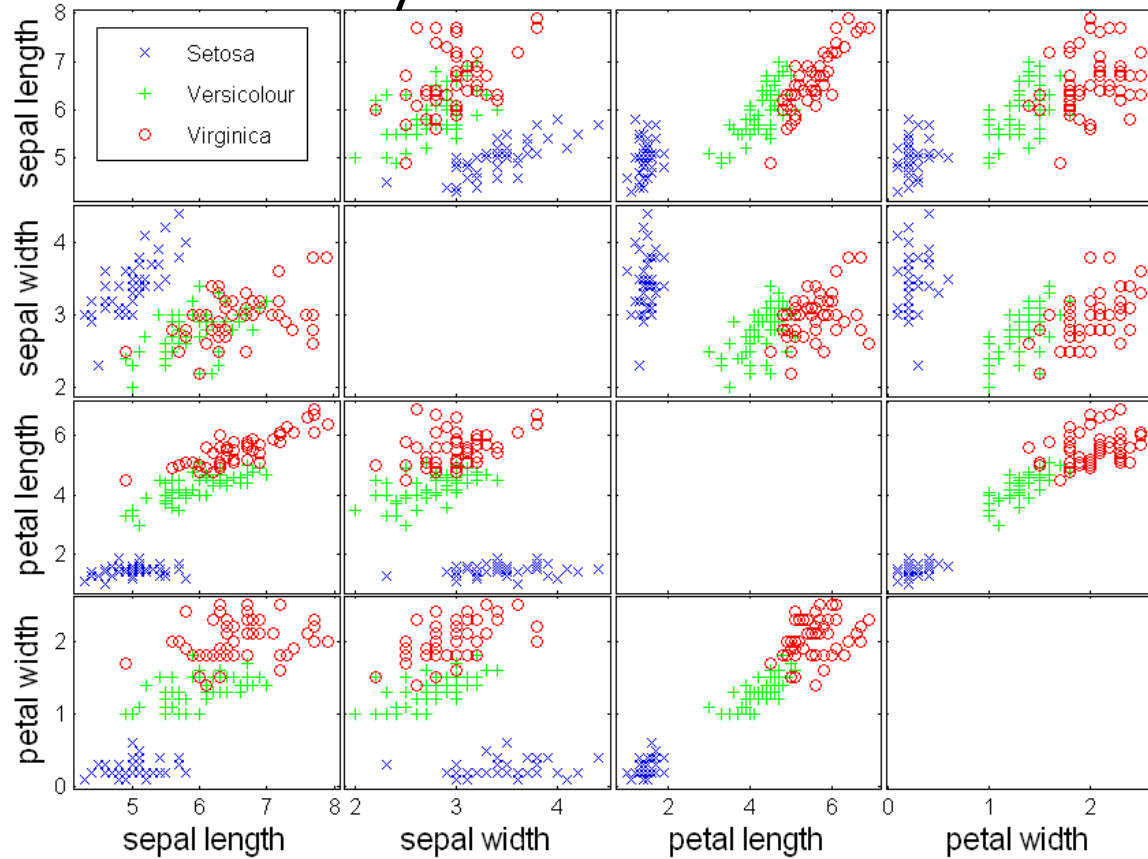
Visualization Techniques: Scatter Plots

- 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

GA





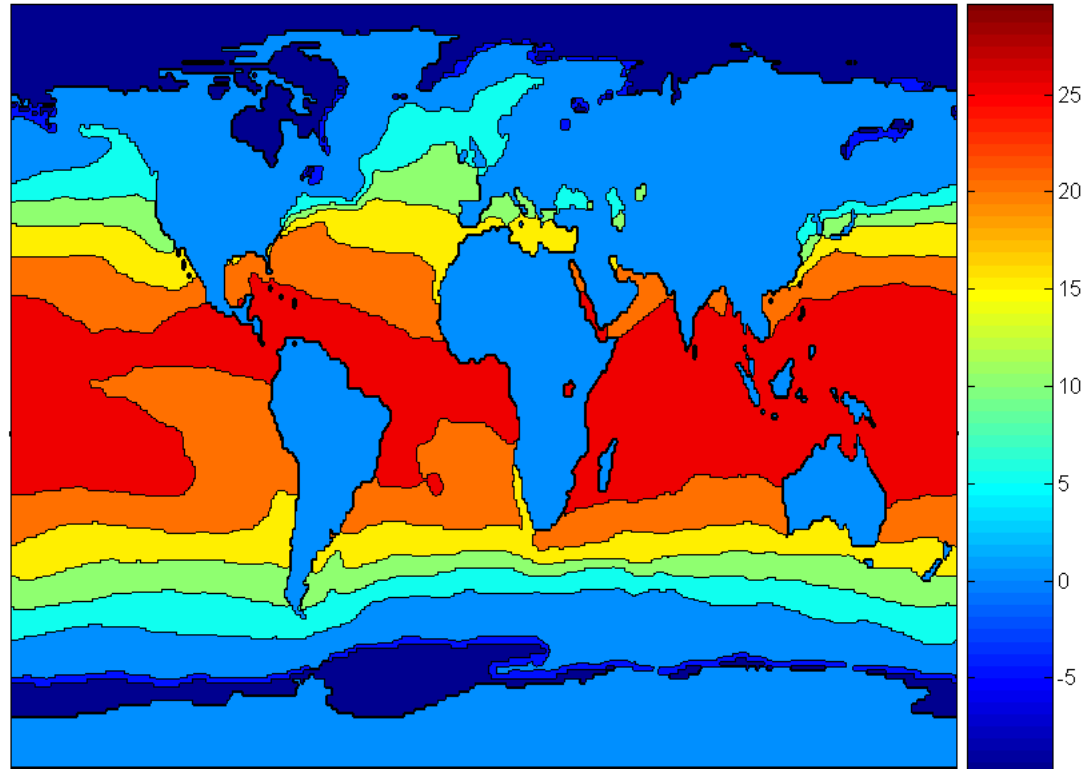
Visualization Techniques: Contour Plots

- 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



UNIVERSITAS
GADJAH MADA

Contour Plot Example: SST Dec, 1998



Celsius



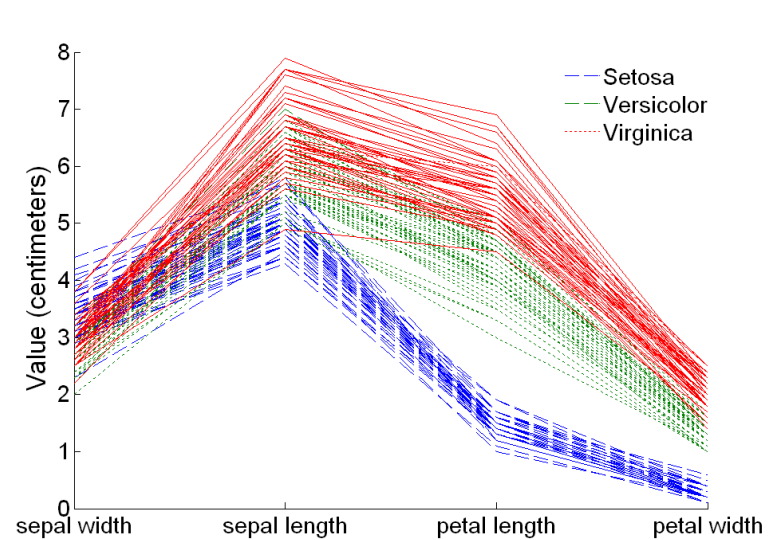
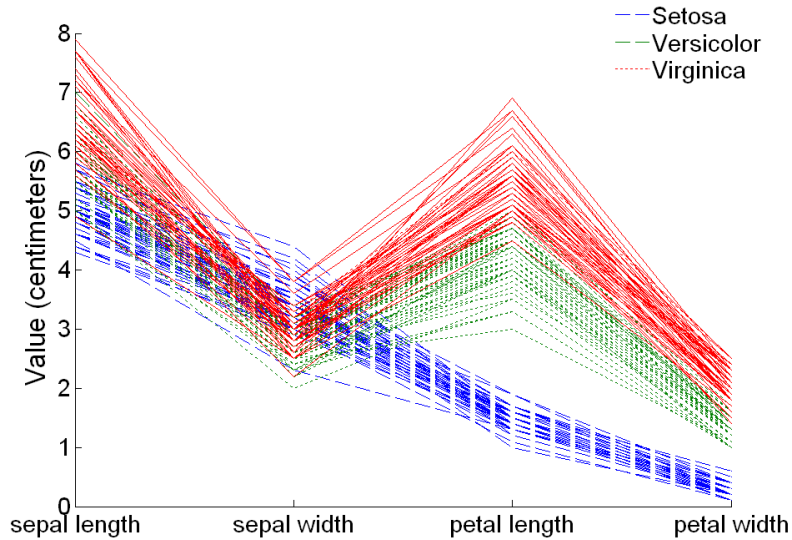
Visualization Techniques: Parallel Coordinates

- 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



UNIVERSITAS
GADJAH MADA

Parallel Coordinates Plots for Iris Data





Other Visualization Techniques

- Star Coordinate 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
 - <http://people.cs.uchicago.edu/~wiseman/chernoff/>
 - <http://kspark.kaist.ac.kr/Human%20Engineering.files/Chernoff/Chernoff%20Faces.htm#>