INTRODUCTION TO KDD PROCESS AND DATA MINING

STQD6414 PERLOMBONGAN DATA

Assoc. Prof. Dr. Nurulkamal Masseran

Depatment of Mathematical Sciences,

Universiti Kebangsaan Malaysia

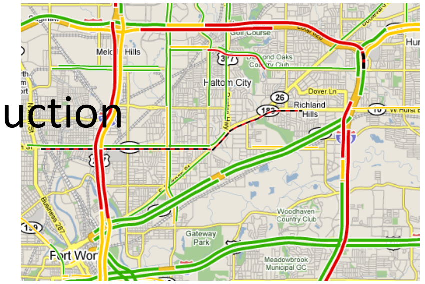
INTRODUCTION:

- In this era, data is everywhere.
- Data is readily available in large quantities in line with the technological development of the 4th industrial revolution.
- Data also comes from a variety of different sources.
- Big data is hard to understand explicitly.
- If not analysed, the data is meaningless.
- Analysis should be conducted to unearth useful information and answer any questions.
- This is the job of statisticians, data scientists, data analysts, data engineers and etc.
- **Article:** Which Jobs Earn The Highest Salaries In Malaysia?

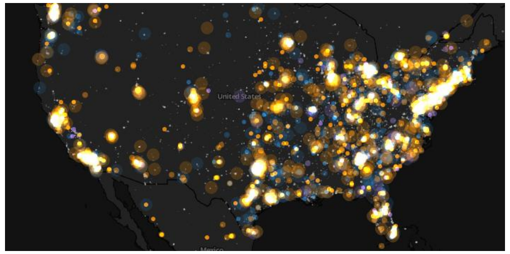
Keselamatan Cyber

E-Dagang

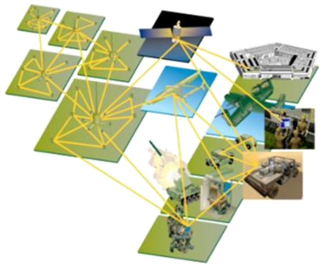
Introduction



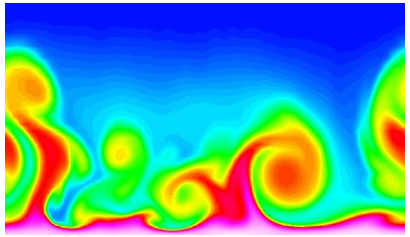
Corak Trafik



Jaringan Sosial: (Facebook, Twitter, dll)



Jaringan Sensor



Simulasi Komputer

DATA MINING: INDUSTRY PERFECTIVE

- In industry, a large amount of data is stored in a data warehouse (data warehoused).
- Example: Website data (Web data)
  - Google has a Tera Bytes of web data.
  - Facebook has billions of active subscribers.
- Example: Purchasing data (retail, ecommerce)
  - Millions of purchase transactions in the Supermarket every day.
  - Millions of customers make online purchases at Amazon.com and Shoppe every day.
- Example: Bank Data and Credit Card transactions.
- Example: Health data of Malaysians (MySejahtera system)

- Competition between industries becoming more challenging.
  - Need to run better service and management.
  - Understand the sentiment of market demand.
  - Produce better quality and cheaper products.
  - All of these require information from data & statistical analysis.
  - Fortunately, computers today are more powerful and cheaper to adapt into this scenario.

DATA MINING: SCIENTIFIC PERFECTIVE

- On a scientific point of view as well, data is observed and collected very quickly.

- Example: Satellite Data.

- NASA EOSDIS collects more than a dozen earth science - related data each year.

- Example: Astronomical telescope.

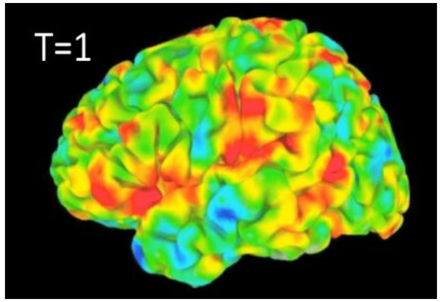
- Sky survey data.

- Example: High intensity biological data.

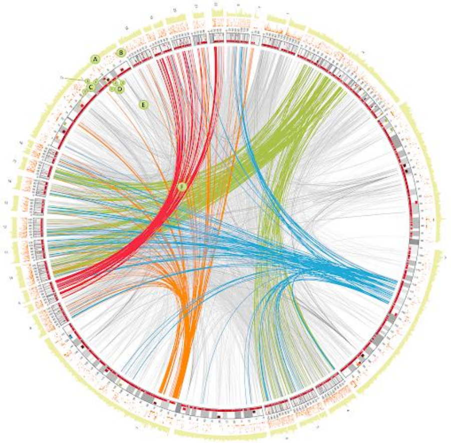
- DNA sequence data.
- Gene Expression Data
- MRI data
- Genome Data

- Example: Demographic Data.

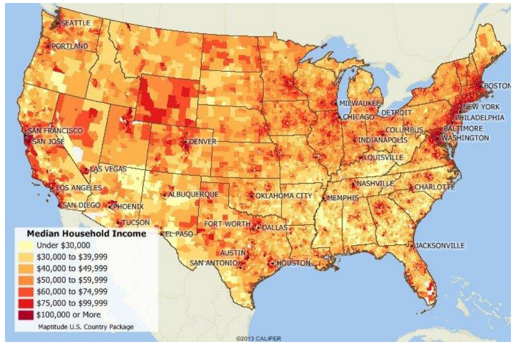
- Income data.
- Population Profile Data



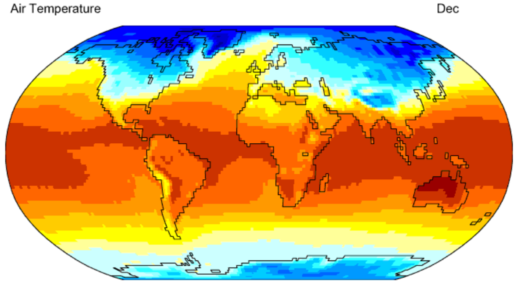
MRI data from human brain



Genome Data



Demographic Data



Earth's surface temperature data

DATA MINING:

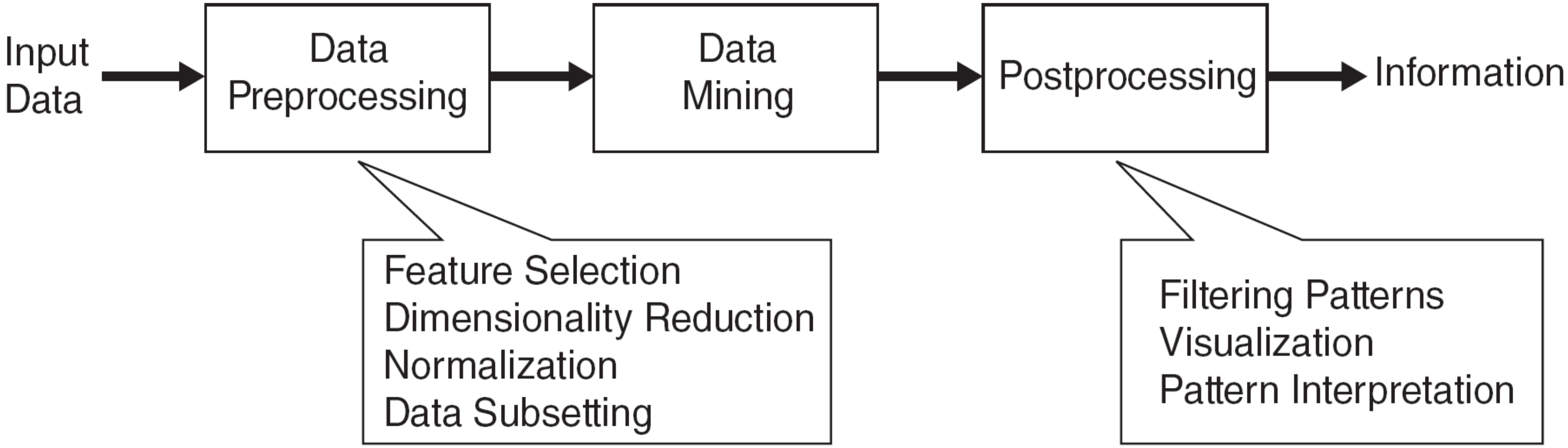
- Data mining is the methods of analysis which used in the process of "knowledge discovery in databases" or KDD.

- Specifically, it aims to:

i) Extract information & answer any related questions.

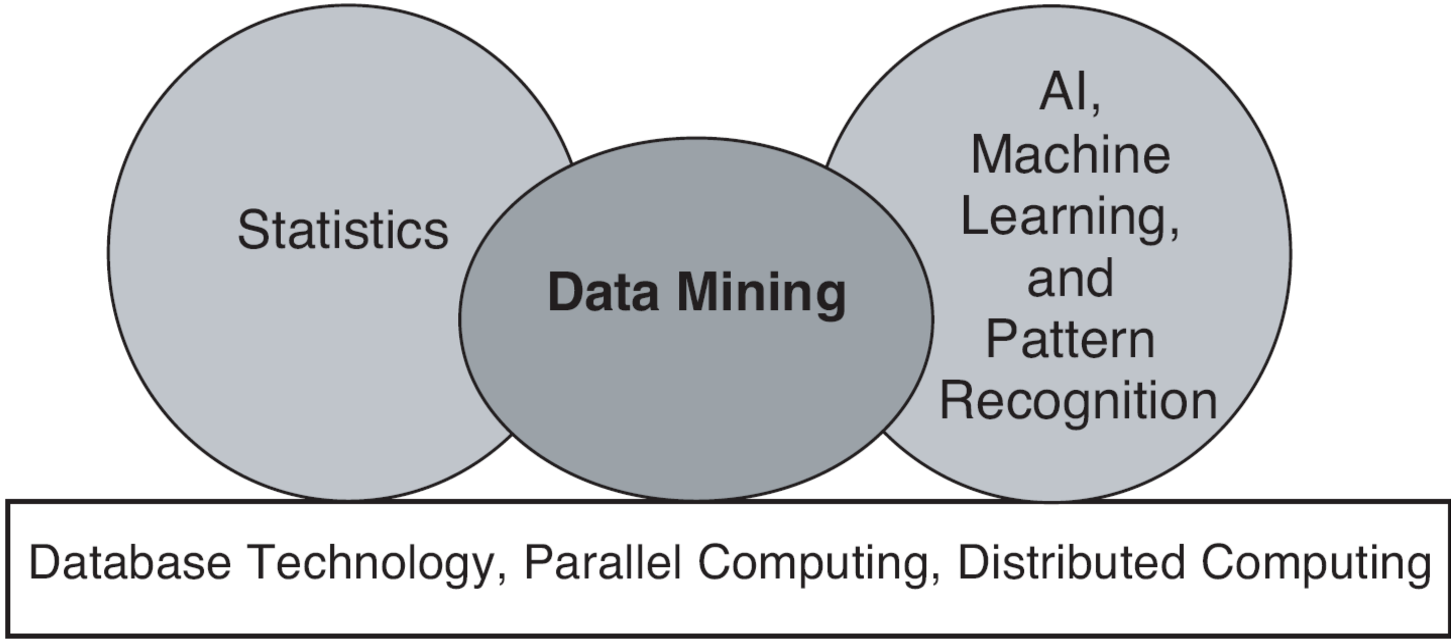
ii) Model the data & predict future values of random variables.

iii) Identify the patterns & trends in data



DATA MINING:

- Most of the knowledge and techniques of data mining come from the fields of Statistics, artificial intelligence machine learning, pattern recognition and database systems.


Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

EXAMPLES OF APPLICATION:

- Banking (loan/credit card approval):
- Predict 'good' or 'bad' customer categories based on old customer records.

- Customer relationship management:
- Identify potential customers to move out of service (example: customers transition from celcom to maxis).

- Target marketing:
- Identify promotional targets to specific groups.

- Fraud detection:
- Telecommunications, financial transactions.

- Manufacturing and production:
- Adjusts the system automatically when process parameters change.

- Medical:
- Analyse the patient's disease history, look for relationships between diseases.
- Identify the nature of the disease, effectiveness of treatment.

KDD PROCESS:

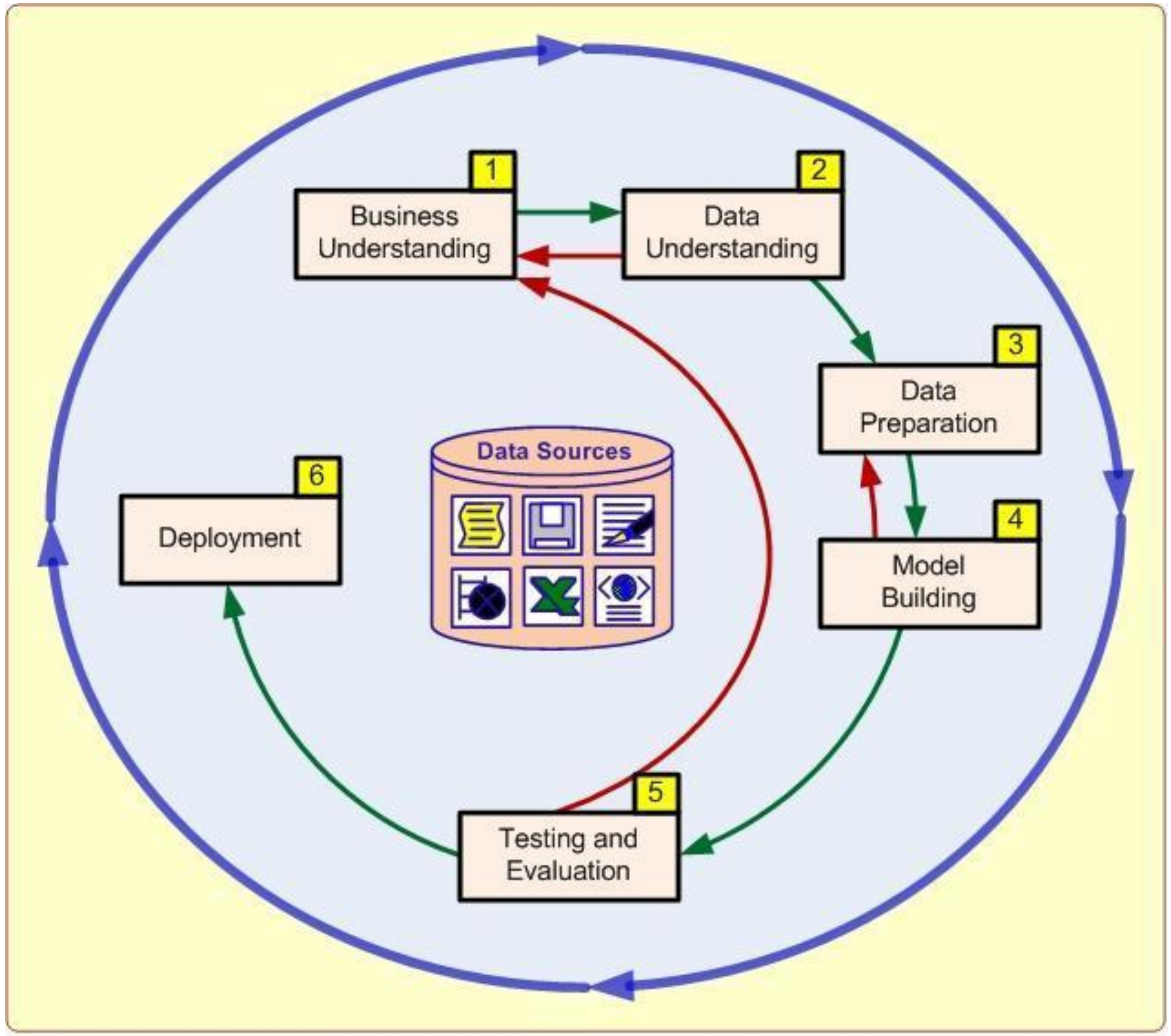
- The KDD process is a repetitive process involving several steps:

1. Problem Formulation.
2. Data Collection & Understanding.
3. Data Pre-Processing:

- Data Cleaning.
- Data Integration.
- Data Transformation.
- Data Reduction.

4. Select appropriate statistical methods or models and perform data mining analysis.
5. Outcome Evaluation and Visualization.
6. Deployment.

KDD PROCESS:

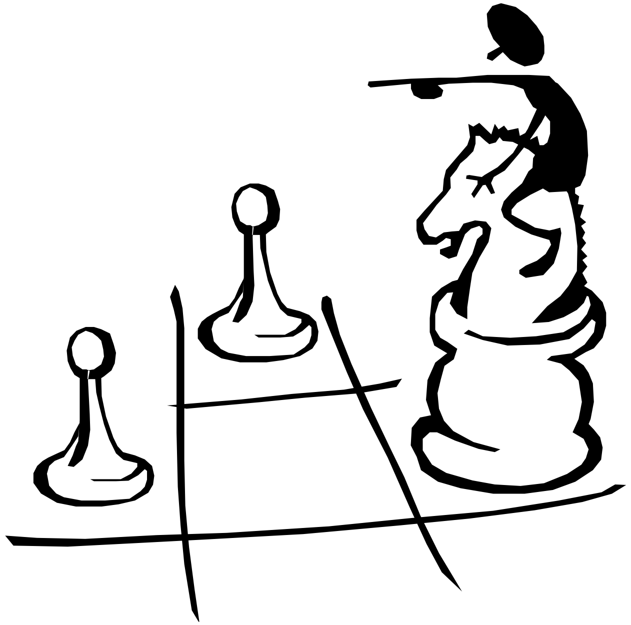


Source: Adapted from CRISP-DM.org.

DATA WAREHOUSE & DATA MINING:

- Data Warehouse give institutions the ability to store a lot of information (memory).

- While, Data Mining helps institutions to make a decisions based on information from data (intelligence)



GENERAL METHODS IN DATA MINING:

- Descriptive Method:
  - Identify patterns that can explain the data.

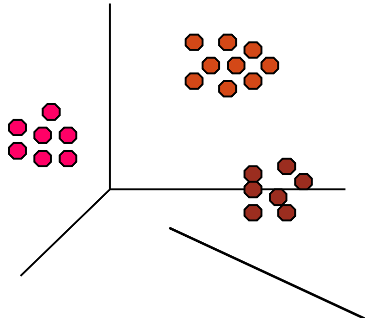
- Forecasting Method:
  - Use some variables to predict future values for other variables.

STATISTICAL SOFTWARE:

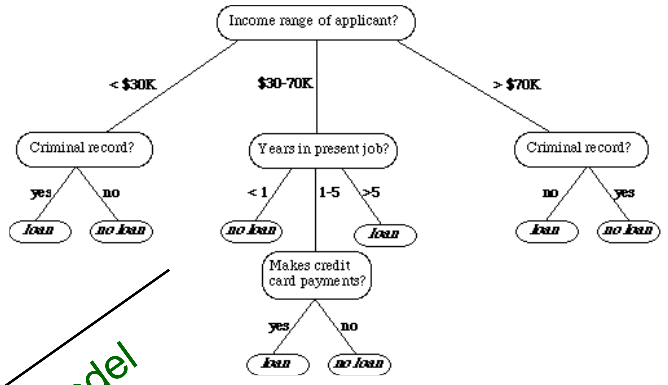
- R programming:

- Phyton:

Examples of Data Mining Techniques:



Clustering



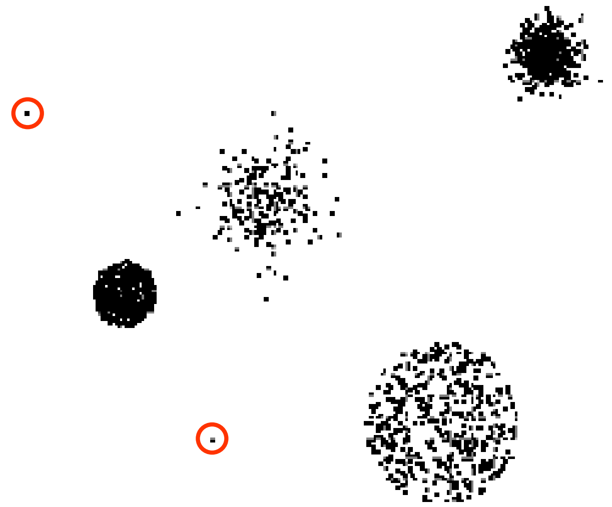
Predictive Model

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association Rules

Anomaly Detection



Milk
Pampers

Basic of R Programming

STQD6414 PERLOMBONGAN DATA

Assoc. Prof. Dr. Nurukamal Masseran

Department of Mathematical Sciences,

Universiti Kebangsaan Malaysia

WHY SHOULD WE USE R SOFTWARE?

- It's free. Can be downloaded from the internet.
- One of the best statistical software at the moment.
- Has access to more than 8000 packages to conduct various statistical and data mining analysis.
- There are many forums/workshops/short courses to help you learn about R software.
- Skills that are much needed by companies related to analytical data

HOW TO INSTALL R:

1. Go to website: https://www.r-project.org/

- Or search "R programing download" in google search.

2. Download R installer.

3. Install R software in your computer.

INTERFACE IN R PROGRAMMING:

1. R console:

- This section shows the output for the executed code.
- You can also write your own code directly on the console.

2. R Script:

- Section for writing programming codes in more flexibly.
- To run the codes, just highlight the codes and run it in the R console.

3. Graphic Output:

- This section shows the graphs or plots constructed while conducting the data analysis.

BASIC COMPUTATION IN R:

- Please type the following mathematical calculations in your R console:

- 2 + 3
  > 5

- 6 / 3
  > 2

- (3*8)/(2*3)
  > 4

- log(12)
  > 1.07

- sqrt (121)
  > 11

- The use of variables is very important in code writing.

- Suppose the variable x represents the sum of 7 and 8. This can be written as:

- x <- 8 + 7
  x
  > 15

- Several variables:

- y<- 10/2

- z<- x*y

- z<- 75

TYPES OF VARIABLES IN R:

1. Qualitative Random Variable: also called a categorical variable, is a variable that isn’t numerical.

1.1 Nominal Random Variable:

- Nominal means "name -related.“
- The nominal variable take the value of symbol or name of a category.

Example:

- Suppose hair color and marital status are variables are used to describe individual data in area A.
- Each individual's hair color change takes the following values: **black, brown, white, gray**.
- P/change marital status takes the following values: **single, married, divorced, single mother.**

1.1.1 Binary/Boolean Variable:

- A nominal variable that has only two categories (0, or 1).

Example: Smoking status of heart patients. Take a value of **0 = no smoking, 1 = smoking**.

1.2 Ordinal Variable: Variables that take categorical values that can be ordered or ranked.

Example:

- Student grade: **A+, A-, A, B+, B, B-, C+, C, C-, D, D-, E.**
- Customer Satisfaction: **0 = very dissatisfied, 1=unsatisfactory, 2=moderate, 3=satisfactory, 4=very satisfactory.**

2. Quantitative Variable : Quantitative variables take numerical values.

- Divided into either discrete variable & continuous variable.

2.1 Discrete Random Variable: Variables that take a finite or infinite value that can be counted (countable infinite).

- In term of integer form: ***0, 1, 2.....n***

Example: Number of children, number of cars, age, and etc.

2.2 Continuous Random Variable: A variable that takes an infinite value.

- In the form of any real number. Can take any value in the interval, for example: $40<X<70$

Example: total income, height of malaysians, room temperature, wind speed and etc.

R CLASSES:

- Everything encoded in R is known as an object.

- Objects in R consist of main 5 classes:

i) Character

ii) Numeric/Real Numbers

iii) Integer

iv) Logikal (True/ False)

v) Compleks Number ($a+bi$)

DATA IN R:

- In general, R has 6 forms of data, namely:

i) Scalar

ii) Vector

iii) Matrix

iv) Data frame

v) List

vi) Array

- **Skalar:**

- Vector with one element.

- **Vektor:**

- Data for a single variable is stored in vector form.
- All elements in the vector are in the same class.

- **Matrix:**

- The combination of several vectors will form a matrix.
- A matrix is a presentation of a 2 -dimensional data structure.
- It is indicated by a set number of rows and columns.
- However, elements in a matrix can only contain real numbers or integers.

- **Data Frame:**

- A Data Frame is a presentation of a 2 -dimensional data structure (similar to a matrix).
- However, the elements in a data frame can consist of different classes.

- **List:**

- A list is a combination of several vectors, matrix, data frame and etc.
- Elements in the list can consist of different classes.

- **Array:**

- A matrix or data frame that has more than 2 dimensions.

OPERATOR DALAM R:

i) Arithmetic Operators:

- Operator that located between two operands.

Operator	Description	Operator	Description
+	Additions	^ atau **	Exponent
-	Subtraction	x %% y	modulus (x mod y) Example: 5%%2 is equal to 1 (remaining)
*	Multiplication	x %/% y	Integer Division Example: 5%/%2 is 2
/	Division	x %*% y	Matrix Multiplication

ii) Relational Operator :

- operators that used to perform comparisons between two variables.

Operator	Description
<	Less than
<=	Less than or equal to
>	Greater than
>=	Greater than or equal to
==	Equal to

iii) Logic Operator :

- Logical operators are used to carry
- Operators that used to combine multiple relational operators.

Operator	Description
!=	Not Equal to
!x	Not x
x | y	x OR y
x & y	x AND y
isTRUE(x)	Test if X is TRUE

THE FUNCTIONS THAT ARE ALREADY AVAILABLE ARE IN R:

Fungsi Matematik:

```
abs(x)     # The absolute value of "x"
log(x),logb(),log10(),log2(),exp(),expm1(),log1p(),sqrt()   #Fairly obvious
cos(),sin(),tan(),acos(),asin(),atan(),atan2()       # Usual stuff
cosh(),sinh(),tanh(),acosh(),asinh(),atanh()         # Hyperbolic functions
union(),intersect(),setdiff(),setequal()             # Set operations
+,-,*,/,^,%%,%/%                                     # Arithmetic operators
<,>,<=,>=,==,!=                                      # Comparison operators

eigen()      # Computes eigenvalues and eigenvectors
deriv()      # Symbolic and algorithmic derivatives of simple expressions
integrate()  # Adaptive quadrature over a finite or infinite interval.
sqrt(),sum()
```

Fungsi Statistik:

```
cor.test()            # Perform correlation test
cumsum(); cumprod(); cummin(); cummax()# Cumulative functions
density(x)            # Compute kernel density estimates
ks.test()             # Performs one or two sample Kolmogorov-Smirnov tests
loess(), lowess()     # Scatter plot smoothing
mad()                 # Calculate median absolute deviation
mean(x), weighted.mean(x), median(x), min(x), max(x), quantile(x)
rnorm(), runif()  # Generate random data with Gaussian/uniform distribution
splinefun()           # Perform spline interpolation
smooth.spline()       # Fits a cubic smoothing spline
sd()                  # Calculate standard deviation
summary(x)            # Returns a summary of x: mean, min, max etc.
t.test()              # Student's t-test
var()                 # Calculate variance
sample()              # Random samples & permutations
ecdf()                # Empirical Cumulative Distribution Function
qqplot()              # quantile-quantile plot
lm              # Fit liner model
glm             # Fit generalised linear model
nls             # non-linear (weighted) least-squares fitting
lqs             # "library(MASS)" resistant regression
optim           # general-purpose optimisation
optimize        # 1-dimensional optimisation
constrOptim     # Constrained optimisation
nlm             # Non-linear minimisation
nlminb          # More robust (non-)constrained non-linear minimisation
```

BASIC PLOTS IN R:

1. Histogram and Density plot.
2. Boxplot.
3. Scatter Plot.
4. Q-Q plot.
5. Pai Chart

And Many More!!

R PACKAGES:

- The packages in R contain a collection of functions, data, code specific to a particular analysis.
- The directory where the package is stored is called the library.
- Special packages in R can be downloaded for free.

- **install.packages**(“*package*")

- The use of R packages will make data mining analysis easier.
- In fact, a variety of more complex statistical analysis and data mining techniques can be carried out.
- **library(***package***)**

REFERENCES:

- Aggarwal, C.C. 2015. *Data Mining: The Textbook*. New York: Springer.
- Bramer, M. 2020. *Principles of Data Mining. (4th Ed.)* Springer.
- Garcia, S., Luengo, J. & Herrera, F. 2014. *Data Preprocessing in Data Mining.* Heidelberg: Springer.
- Han, J., Kamber, M. & Pei, J. 2012. *Data Mining: Concepts and Techniques*. *(3rd Edition).* Massachusetts: Elsevier.
- Jamsa, K. 2020. *Introduction to Data Mining and Analytics.* Jones & Bartlett Learning.

REFERENCES:

- Larose, D.T. & Larose, C.D. 2014. *Discovering Knowledge in Data: An Introduction to Data Mining. (2nd Edition).* New Jersey: John Wiley & Sons, Inc.
- North, M. 2018. *Data Mining for the Masses: With Implementations in RapidMiner and R. (3rd Ed.).* CreateSpace Independent Publishing Platform.
- Olson, D. L. & Lauhoff, G. 2019. *Descriptive Data Mining.* Springer.
- Zaki, M. J. & Meira Jr, W. 2020. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms. (2nd Ed.).* Cambridge University Press.

NEXT TOPIC:

Basic Techniques of Data Exploration Using R