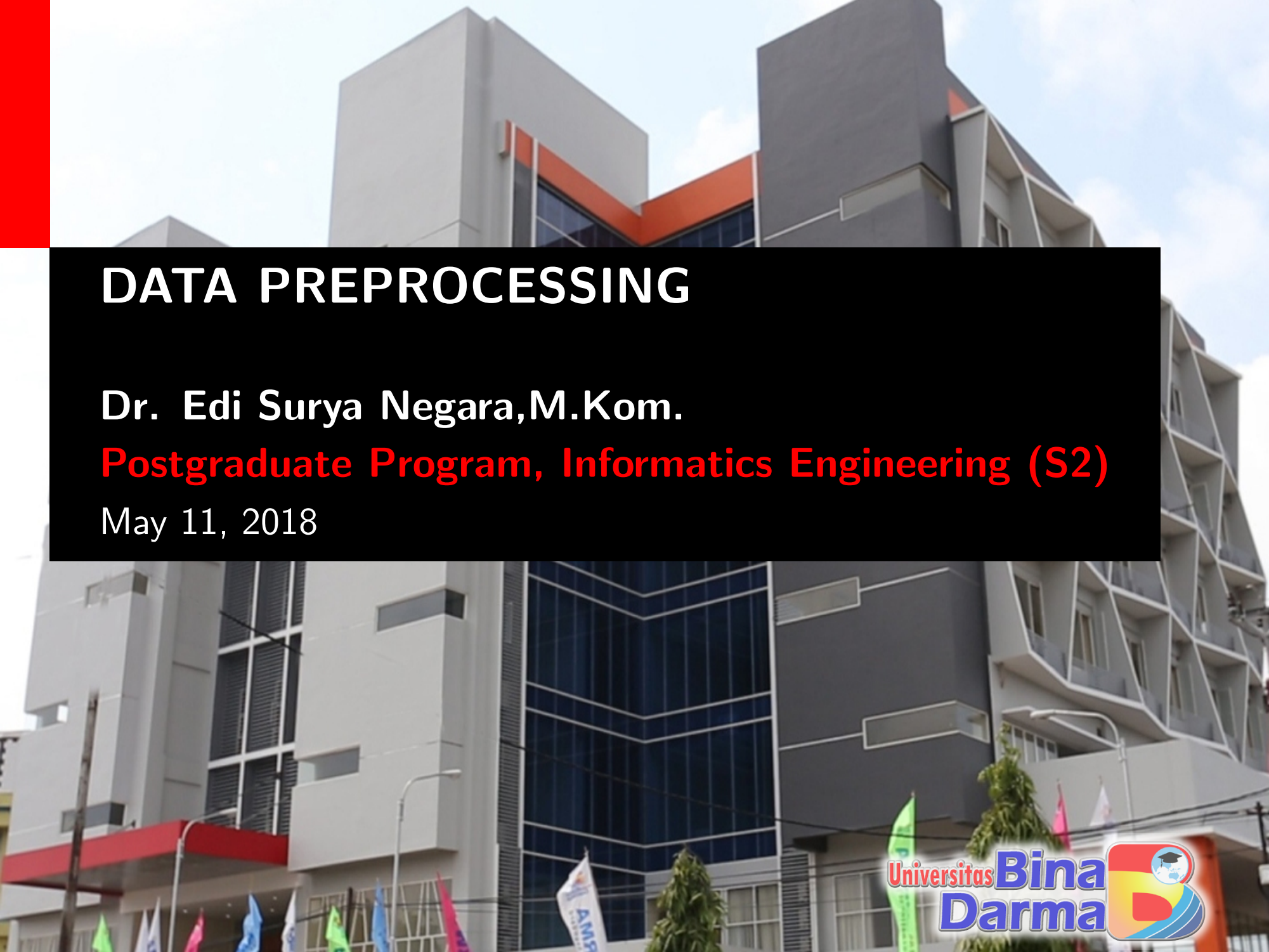


DATA PREPROCESSING

Dr. Edi Surya Negara,M.Kom.

Postgraduate Program, Informatics Engineering (S2)

May 11, 2018

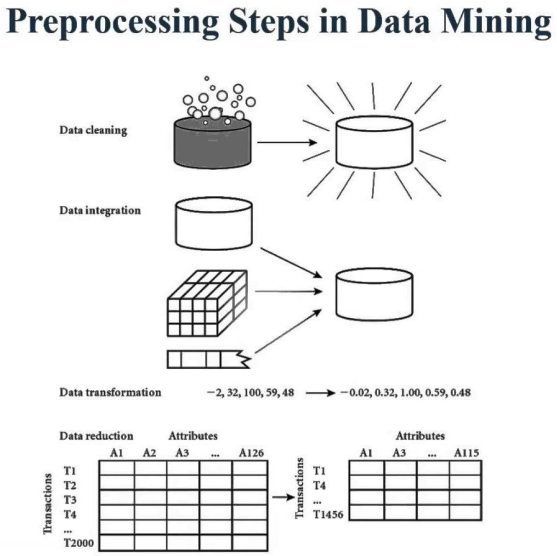
Chapter 1: Objectives

After completing this chapter, students will be able to:

- Explain how to data cleaning.
- Explain how to data integration.
- Explain how to data transformation.

Reference: Tan et al: Introduction to data mining. Some slides are adopted from Tan et al.

Objectives and Benefits Data Preprocessing



- Reduction of computing time.
- Can make the data value becomes smaller without changing the information it contains.

Data Quality?

Data quality is measured from several aspects:

- **Accuracy**. The degree of conformity of a measure to a standard or a true value.
- **Completeness**. The degree to which all required measures are known.
- **Consistency**. The degree to which a set of measures are equivalent in across systems.
- **Timeliness**. Whether the data is uptudate.
- **Trusted**. The degree to which the measures conform to defined business rules or constraints.
- **Interpreted**. Whether the data can be understood.

What is data cleaning?

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
Data cleaning includes:

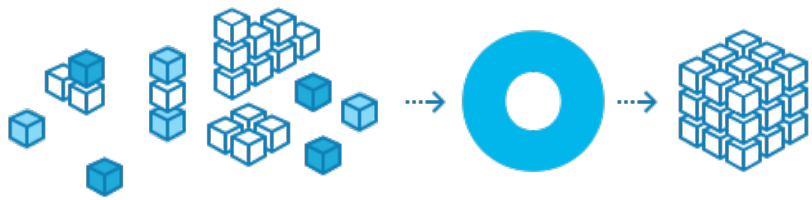
- Fulfillment missing values.
- Smooth data noisy.
- Identify outliers.
- Eliminating inconsistencies.

Attribute value error could be due

- Error instrument data collection
- Data entry error
- Data delivery error
- Technological limitations
- Inconsistent naming conventions
- Records are repeated
- Incomplete data
- Inconsistent data

Data Transformation

Data Transformation Goals:



- Overcome challenge of variable underlying data set structures through:
  - Creating a uniform, integrated data set that allows for timely and easily accessible reports
  - Integrating data needs according to a central schema

Several ways to data transformation

- Centering
- Normalization z-core
- Scaling

$$X' = \begin{bmatrix} X_{11}, X_{21}, \ldots, X_{n1} \\ X_{12}, X_{22}, \ldots, X_{n2} \\ \ldots, \ldots, \ldots, \ldots \\ X_{1p}, X_{2p}, \ldots, X_{np} \end{bmatrix}$$

Centering

- Based on fitting a distribution to the data
- Distance function between distributions

Centering formulation:

$$\widehat{X} = X - \overline{X} \tag{1}$$

$\widehat{X}$= vector results after centering
X= vector of the original column
$\overline{X}$=average of the columns in question.

Operation Centering

Scatter Matrix is the distance between attributes X_1 with X_2, or X_1 with X_3 and so on.

$$Scatter = \widehat{X}^T \widehat{X} \tag{2}$$

Kovarian Matrix is the distance between attributes X_1 with X_2, or X_1 with X_3 and so o, but divided by each entry in the matrix by the amount of data $m - 1$.

$$kovarian = \frac{\widehat{X}^T \widehat{X}}{m - 1} \tag{3}$$

Normalization z-core

Normalization dividing each data that has been cast with the standard deviation of the attribute in question.

$$\widehat{X} = \frac{X - \overline{X}}{\sigma_x} \tag{4}$$

Data after this standardization will have a mean of $= 0$ and variance $= 1$

Scaling

Scaling procedure to change the data so that it is in a certain scale.

$$\widehat{X} = \frac{X - X_{min}}{X_{max} - X_{min}} * (BA - BB) + BB \tag{5}$$

Data Transformation: Example

E:\TEACHING\Data Mining dan Big Data Analytics\Data-Mining-and-Big-Data-Analytics-Book-master\Data-Mining-and-Big-Data-Analytics-Book-master\3.4 Transformasi_Data.py - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

client1.py | server1.py | chat_server.py | plot_stock_market.py | 3.4 Transformasi_Data.py

```
1  #import numpy
2  import numpy as np
3
4  #mendefinisikan matrik x, bebas, untuk contoh saja
5  x = np.matrix([[-20,23,5],[4,-8,15]])
6
7  #centering (data asli dikurangi nilai reratanya)
8  #sesuai Persamaan 1
9  x_centering = x - x.mean()
10 print("x_centering: ", x_centering, "\n")
11
12 #standarisasi sesuai Persamaan 4
13 x_standarisasi = (x-x.mean()) / x.std()
14 print("x_standarisasi: ", x_standarisasi)
15 print("mean x_standarisasi: ", x_standarisasi.mean())
16 print("varian x_standarisasi: ", x_standarisasi.var(), "\n")
17
18 #scaling ke range 0-1, sesuai Persamaan 5
19 BA = 1; BB = 0 #BA=batas atas, BB = batas bawah
20 x_scaling = (x - x.min()) / (x.max()-x.min()) * (BA-BB) + BB
21 print("x_scaling: ", x_scaling)
```

Data Transformation: Output

Command Prompt

```
E:\TEACHING\2018\Advanced Data Based\SLIDE\koding>python Transformasi_Data.py
('x_centering: ', matrix([[-23.16666667,  19.83333333,   1.83333333],
        [  0.83333333, -11.16666667,  11.83333333]]), '\n')
('x_standarisasi: ', matrix([[-1.6389276 ,  1.40311068,  0.12969931],
        [ 0.05895423, -0.78998669,  0.83715007]]))
('mean x_standarisasi: ', -1.8503717077085941e-17)
('varian x_standarisasi: ', 1.0000000000000002, '\n')
('x_scaling: ', matrix([[0, 1, 0],
        [0, 0, 0]]))

E:\TEACHING\2018\Advanced Data Based\SLIDE\koding>
```

Universitas Bina Darma

Road to 2025

IT Based International Standard