Pre-processing:
Data Cleansing

DSA 8102: Data mining, data storage and retrieval

Strathmore UNIVERSITY

D. Owuor (Ph.D)

Objectives

- Describe data pre-processing and the concept of data quality.
- Differentiate data exploration from EDA.
- Discuss ways to fix missing value problems.
- Determine what an outlier is and ways to detect and remove it.

© Strathmore University, 2021

Strathmore
UNIVERSITY

Data Quality

Data Quality

- **Definition (standard-based perspective):**
  - "degree to which the inherent characteristics of data fulfills requirements" (how well does it represent the real-world construct to which it refers).
  - "measure of the condition of data (i.e. accuracy, completeness, consistency, reliability, up-to-date etc.) for its application" (fit for its intended use).
- **Data governance:**
  - used to form agreed upon definitions and standards for data quality.
  - to achieve this: data pre-processing among other techniques are used.

© Strathmore University, 2021

Strathmore
UNIVERSITY

Data Exploration

Data Exploration

Strathmore UNIVERSITY

- Originally known as **Exploratory Data Analysis (EDA)**, created by statistician John Tukey.
  - Currently EDA techniques (outlier detection, clustering) have evolved into independent areas of research.
- **Description:**
  - visualization and computation to better understand characteristics of data.
- **Objectives:**
  - aims at creating a mental understanding of the data (relationship between variables)
  - quickly identify faulty points in data (errors, missing values etc.)

© Strathmore University, 2021

Data Exploration

Strathmore UNIVERSITY

- **Motivations:**
  - Help select right tool for preprocessing
  - Make use of human's ability to recognize patterns (e.g., it takes a lot of work for an algorithm to recognize faces than humans)
- **Technique & Tools:**
  - Statistical graphics and data visualizations (i.e., Box plot, Histogram, Run chart, Scatter plot etc.)

© Strathmore University, 2021

Class Exercise

1. Differentiate data exploration from EDA.

© Strathmore University, 2021

Strathmore
UNIVERSITY

Processing vs Pre-processing

Data Processing

- **Definition:**
  - task of collecting data and translating it into usable information.
- **Stages:**
  - **Data extraction:** pulling from numerous sources.
  - **Data pre-processing:** preparation stage.
  - **Data input:** clean data is uploaded into its destination (i.e., data warehouse like Redshift).
  - **Processing:** actual processing for interpretation/results.
  - **Data output:** Interpretations simplified for non-data scientists (i.e., charts, graphs etc.).
  - **Data storage:** for future use.

© Strathmore University, 2021

Data Pre-processing

- **Definition 1:**
  - technique used to convert raw data into a clean data set, preparing the data set for modelling (i.e., ML, DM).
- **Definition 2:**
  - a technique used to transform raw data into a useful and efficient format.
- **Definition 3:**
  - step in which data gets transformed/encoded, in order to bring it to a state where its features can be easily interpreted by an algorithm.

© Strathmore University, 2021

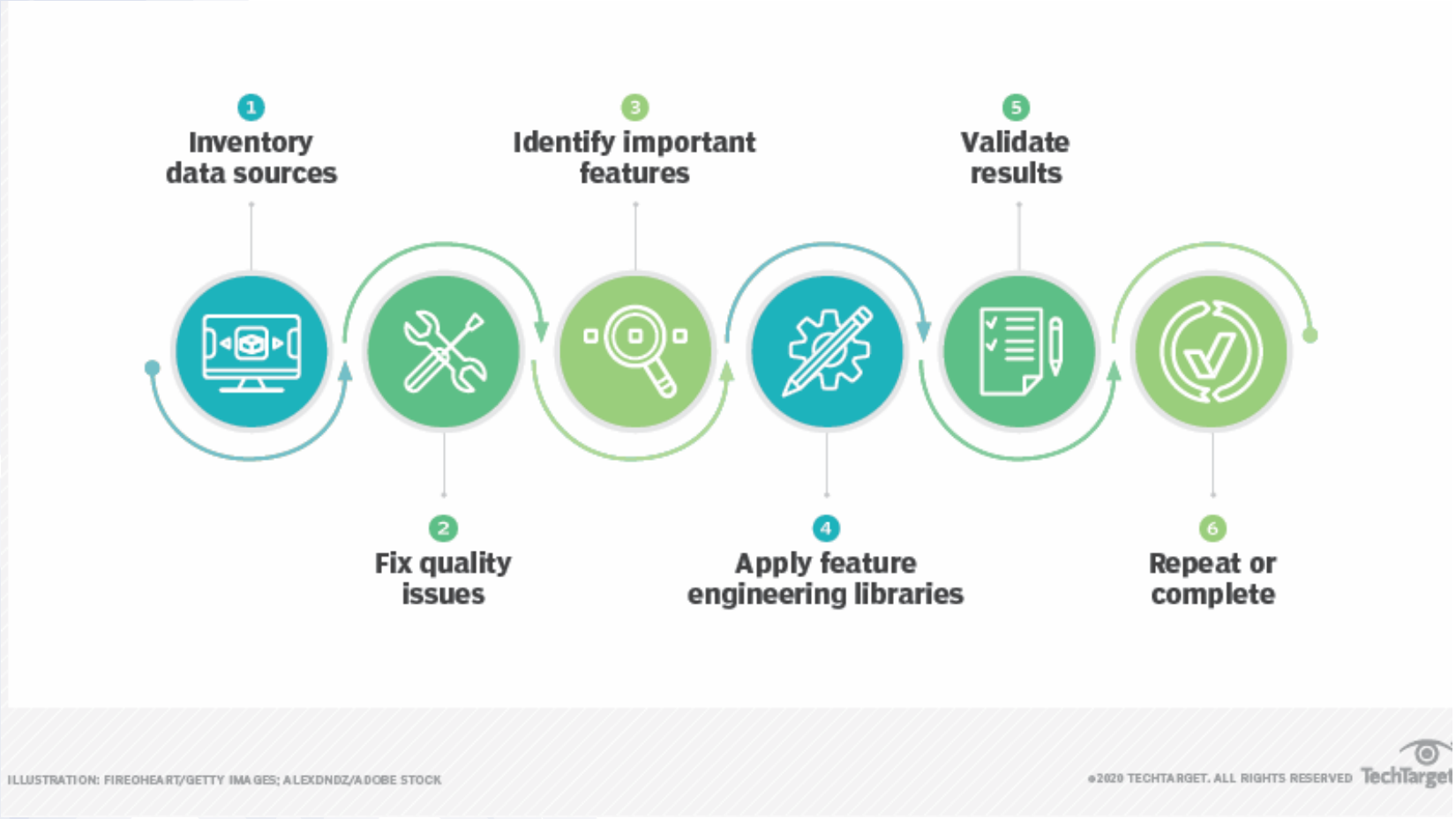
Pre-processing Significance

Strathmore UNIVERSITY

- **Analogy:** data is like crude oil
  - cannot be used directly from its source; so,
  - must be processed before being used for diff. purposes.
- **Real-world data issues:**
  - comes from different sources with different structures and data types.
  - Often incomplete, inconsistent, noisy (i.e., errors, outliers) and/or lacking in certain behaviors.
- **Pre-processing:**
  - used to improve data quality (purify data).
  - prepare data set for meaningful analysis.
  - product is a clean data/training set.

© Strathmore University, 2021

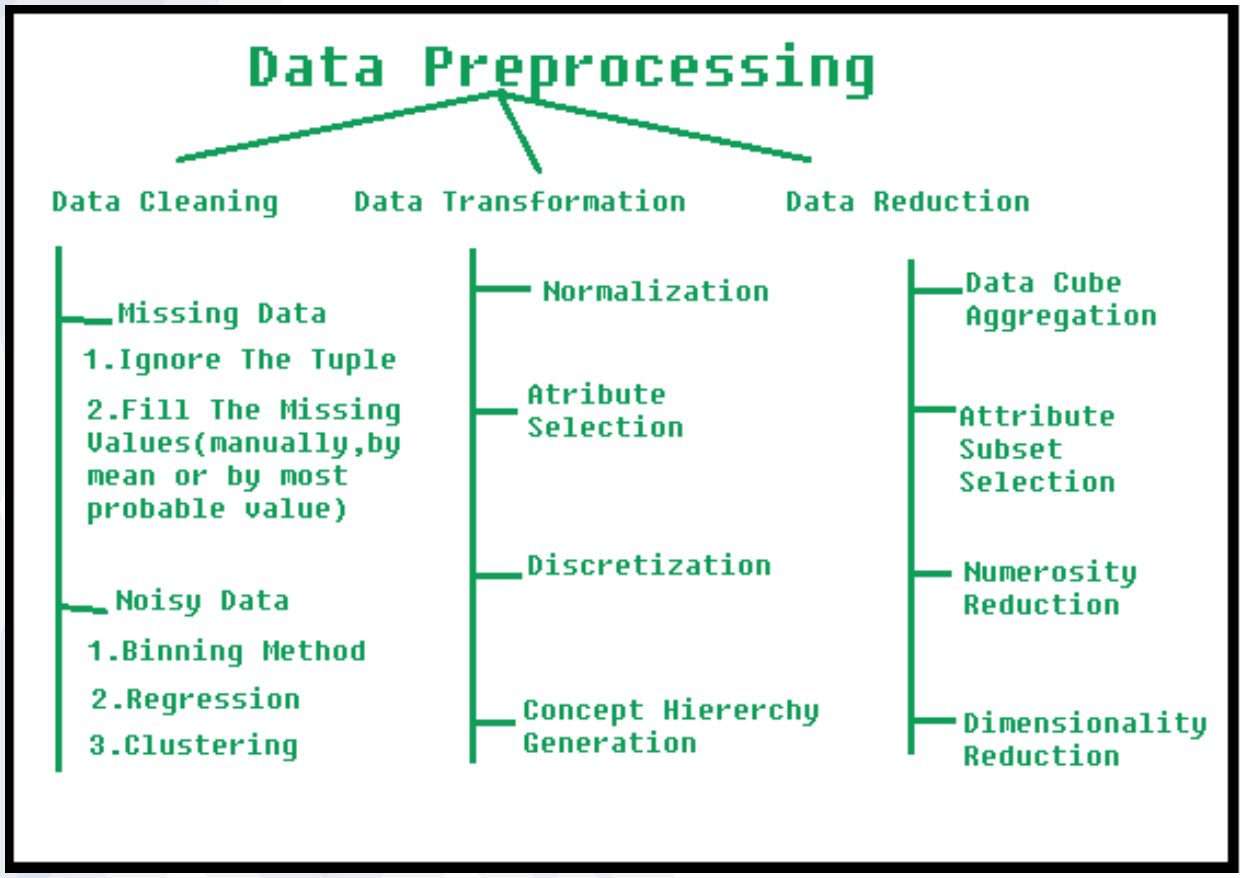
Pre-processing steps



Source: https://searchsqlserver.techtarget.com/definition/data-preprocessing

© Strathmore University, 2021

Pre-processing Tasks



Source: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

© Strathmore University, 2021

Class Exercise

1. What could be the possible reasons for data to be inconsistent?
2. Why is data quality important? Or what could be the benefits of pre-processing data before using it?

© Strathmore University, 2021

Pre-processing steps

1. Inventory (data sources)
2. Data Cleansing
3. Feature Selection
4. Feature Transformation
5. Validation (2, 3, 4)

© Strathmore University, 2021

Strathmore
UNIVERSITY

Pre-processing (step 2):
data cleansing

Strathmore
UNIVERSITY

Data cleansing (or cleaning)

© Strathmore University, 2021

Data Cleansing

- **Definition:**
  - process of detecting and correcting the corrupt parts of a data set.
- **Benefits:** (purify data)
  - correct errors;
  - detect and analyze outliers.
- **Data cleansing tasks (examples):**
  - cleaning missing values
  - reducing noise (i.e., class and attribute noise)

© Strathmore University, 2021

Strathmore
UNIVERSITY

Fixing missing values

© Strathmore University, 2021

Missing Values

Strathmore UNIVERSITY

- **Description:**
  - also known as **data holes**
  - incomplete data point (s) in records
  - most data mining and machine learning techniques do not support data with missing values
- **Possible Reasons:**
  - No information is provided;
  - Incorrect data types for attributes during extraction;
  - Integrating data from multiple sources
  - List goes on ...

© Strathmore University, 2021

Dealing with Missing Values

Strathmore UNIVERSITY

- **Ignore incomplete records**
  - Not practical for data sets that contain large portions of missing values (poor rep. of initial pop.)
- **Update with descriptive/sensible values**
  - e.g., mean, mode, median etc.
  - Introduces inconsistencies
- **Investigate an identify patterns of missing values. E.g.**
  - Data is truly unavailable
  - Duplicate records (one updates another)
  - Simple human errors (like wrong input types)
  - List goes on ...

© Strathmore University, 2021

Strathmore
UNIVERSITY

Detecting and handling outliers

© Strathmore University, 2021

Noise

Strathmore UNIVERSITY

- **Description:**
  - Refers to erroneous values (i.e., outliers)
  - Random error or abnormal variance in a measured variable
  - Adversely influences model/data mining performance
  - Requires detection and removal
- **Types:**
  - Class noise (categorical variables): erroneous class labels
  - Attribute noise (numeric variables): corruptive values and outliers
- **Possible Reasons:**
  - Data entry errors (i.e., human mistakes)
  - Faulty data collection (i.e., automated record collection)
  - List goes on ...

© Strathmore University, 2021

Handling Outliers

- **Binning**
  - Segmenting data into small bins (or buckets) and smoothing (or replacing) bin values by mean, mode etc.
- **Clustering**
  - Group (numeric) data by similarity {detects outliers}
- **Classification**
  - Arrange (categorical) data according to shared classes {erroneous classes have very few members}
- **Regression**
  - Fit data into regression functions (i.e., linear) {remove random variance}

$y = mx + b$: where y is predicted value, x is actual value

© Strathmore University, 2021

Class Exercise

1. Which collection errors may lead to:
   a. missing values
   b. noisy data

© Strathmore University, 2021

Research Perspectives

- **Pre-processing**
  - ReTallón-Ballesteros A.J., Riquelme J.C. (2015) Data Cleansing Meets Feature Selection: A Supervised Machine Learning Approach.
- **Missing Values:**
  - Shah, Faaiz, Arnaud Castelltort, and Anne Laurent. "Handling missing values for mining gradual patterns from NoSQL graph databases." *Future Generation Computer Systems* 111 (2020): 523-538.
  - Li, Zhenghui, et al. "Grid-Constrained Data Cleansing Method for Enhanced Busload Forecasting." *IEEE Transactions on Instrumentation and Measurement* 70 (2021): 1-10.
  - Dakka, M. A., et al. "Automated Detection of Poor-Quality Data: Case Studies in Healthcare." (2021).
- **Outlier Detection:**
  - Tushar, Deoras Tejas. "Binary Priority Outlier Classifier Based Outlier Elimination." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.3 (2021): 4261-4266.

© Strathmore University, 2021

References

1. Talavera, Luis. "Feature selection as a preprocessing step for hierarchical clustering." *ICML*. Vol. 99. 1999.
2. https://medium.com/@divyagera2402/data-journey-from-acquiring-it-to-feeding-it-in-a-model-steps-from-data-cleaning-handling-eda-4eacf9d316f1
3. https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/
4. https://www.talend.com/resources/what-is-data-processing/
5. https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825
6. https://en.wikipedia.org/wiki/Data_quality
7. https://searchdatamanagement.techtarget.com/definition/data-quality.
8. https://www.youtube.com/watch?v=oskDpBfxBEI
9. https://en.wikipedia.org/wiki/Data_exploration

© Strathmore University, 2021

Strathmore
UNIVERSITY

Thank you!

Any Questions?