# Data Intensive Systems (DIS)
# KBH-SW7 E25

## 5. Foundation of Data Mining

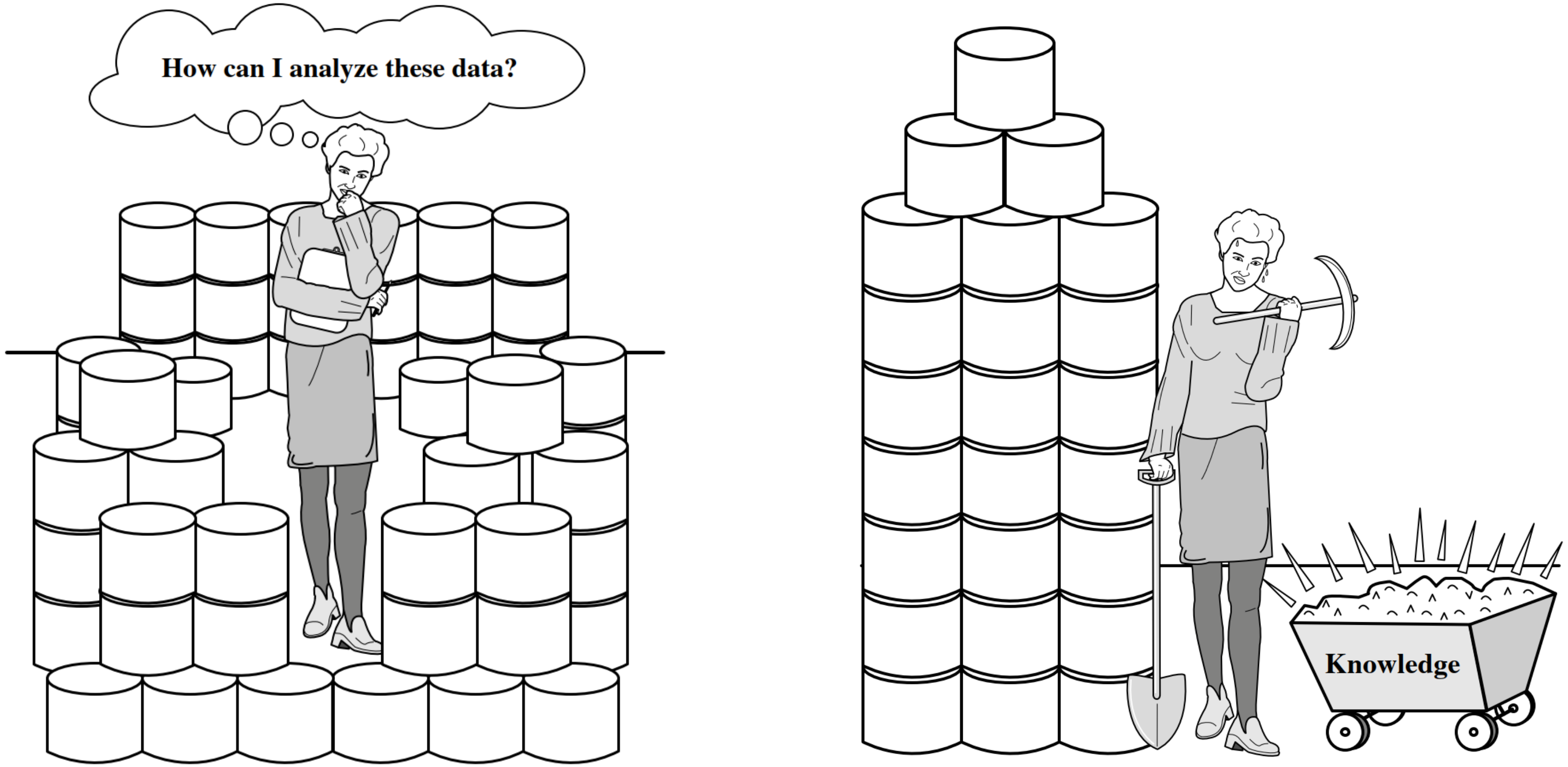
AALBORG
UNIVERSITY

# Agenda

- Introduction to Data Mining
  - Data Mining Process
  - Different Types of Data Mining
  - Technologies used in Data Mining
- Data Preprocessing

# What is Data Mining?

- Data mining turns a large collection of data into knowledge.



# Limitation of First-oder Logic

- First-order logic
  - "All men are mortal. Socrates is a man. Therefore, Socrates is mortal." $(\forall x \,.\, man(x) \rightarrow mortal(x)) \wedge man(Socrates) \rightarrow mortal(Socrates)$
- We can only represent the facts, which are either true or false. First-order logic is not sufficient to represent the complex sentences or natural language statements.
  - $\forall x \,.\, car(x) \wedge black(x) \rightarrow brake(self)$
  - $\forall x \,.\, truck(x) \wedge red(x) \rightarrow brake(self)$?
  - $\forall x \,.\, car(x) \wedge blue(x) \rightarrow brake(self)$?

# Data Mining Process

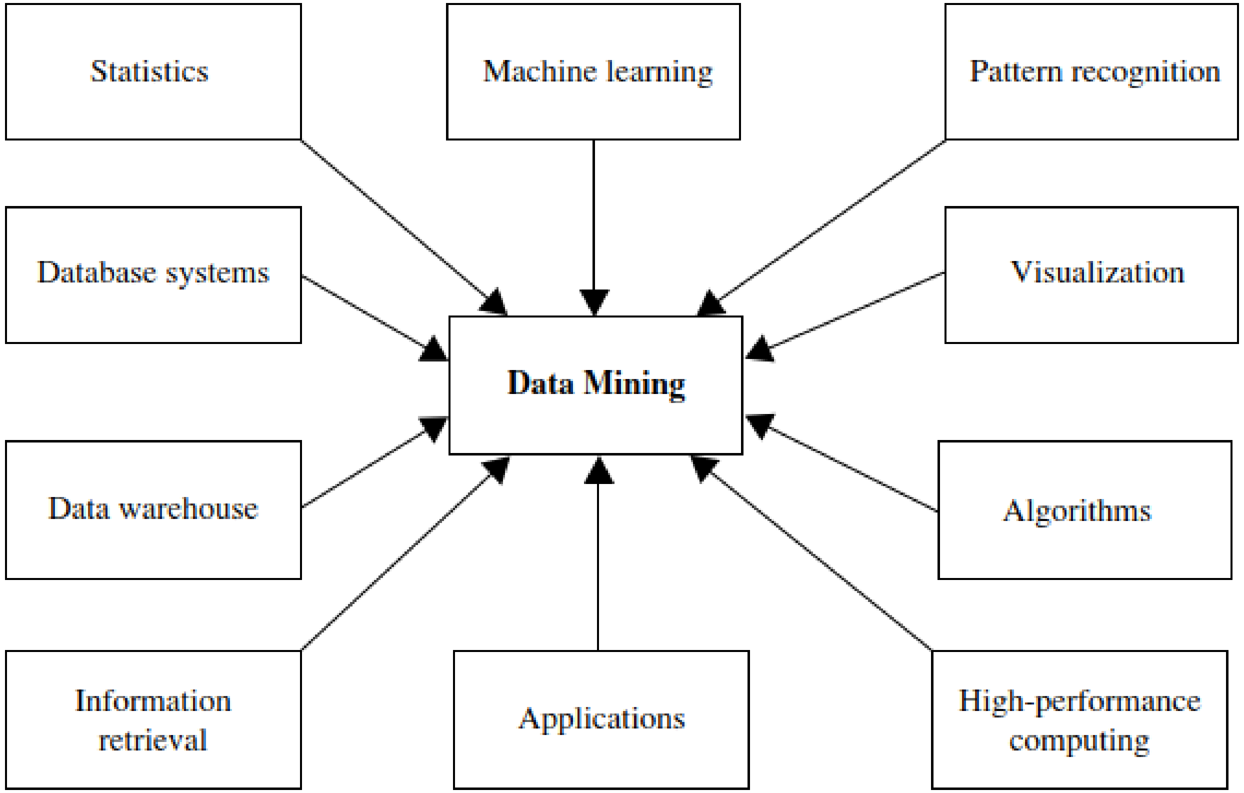
- Data cleaning (to remove noise and inconsistent data)
- Data integration (where multiple data sources may be combined)
- Data selection (where data relevant to the analysis task are retrieved from the database)
- Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
- Data mining (an essential process where intelligent methods are applied to extract data patterns)
- Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
- Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

# Different Types of Data Mining

| Tasks | Description | Algorithms | Examples |
|---|---|---|---|
| Classification | Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known dataset | Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbors | Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups |
| Regression | Predict the numeric target label of a data point. The prediction will be based on learning from a known dataset | Linear regression, logistic regression | Predicting the unemployment rate for the next year Estimating insurance premium |
| Anomaly detection | Predict if a data point is an outlier compared to other data points in the dataset | Distance-based, density-based, LOF | Detecting fraudulent credit card transactions and network intrusion |
| Time series forecasting | Predict the value of the target variable for a future timeframe based on historical values | Exponential smoothing, ARIMA regression | Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated |
| Clustering | Identify natural clusters within the dataset based on inherit properties within the dataset | k-Means, density-based clustering (e.g., DBSCAN) | Finding customer segments in a company based on transaction, web, and customer call data |
| Association analysis | Identify relationships within an item set based on transaction data | FP-growth algorithm, a priori algorithm | Finding cross-selling opportunities for a retailer based on transaction purchase history |
| Recommendation engines | Predict the preference of an item for a user | Collaborative filtering, content-based filtering, hybrid recommenders | Finding the top recommended movies for a user |

# Which Technologies Are Used?

- **Statistics** studies the collection, analysis, interpretation or explanation, and presentation of data. Data mining has an inherent connection with statistics.
- **Machine learning** investigates how computers can automatically learn to recognize complex patterns and make intelligent decisions (or improve their performance) based on data.
- **Database systems** focuses on the creation, maintenance, and use of databases for organizations and end-users.
- **Information retrieval** is the science of searching for documents or information in documents. Documents can be text or multimedia, and may reside on the Web.



# Data Mining Settings

| Criteria | Supervised | Unsupervised | Reinforcement |
|---|---|---|---|
| Definition | Use labeled data | Use unlabeled data without any guidance. | Work on interacting with the environment |
| Type of data | Labeled data | Unlabeled data | No – predefined data |
| Type of problems | Regression and classification | Association and Clustering | Exploitation or Exploration |
| Supervision | Extra supervision | No supervision | No supervision |
| Algorithms | Linear Regression, Logistic Regression, SVM, KNN etc. | K – Means,<br>C – Means, Apriori | Q – Learning,<br>SARSA |
| Aim | Calculate outcomes | Discover underlying patterns | Learn a series of action |
| Application | Risk Evaluation, Forecast Sales | Recommendation System, Anomaly Detection | Self Driving Cars, Gaming, Healthcare |

# Agenda

- Introduction to Data Mining
- Data Preprocessing
    - Motivation of Data Preprocessing
    - Data Cleaning
    - Data Integration
    - Data Reduction
    - Data Transformation

# Prior Knowledge

- Objective: the data mining process starts with a need for analysis, a question, or a business objective.
- Subject area: it is essential to know the subject matter, the context, and the business process generating the data.
- Data: quality of the data, quantity of data, availability of data, gaps in data, labels…

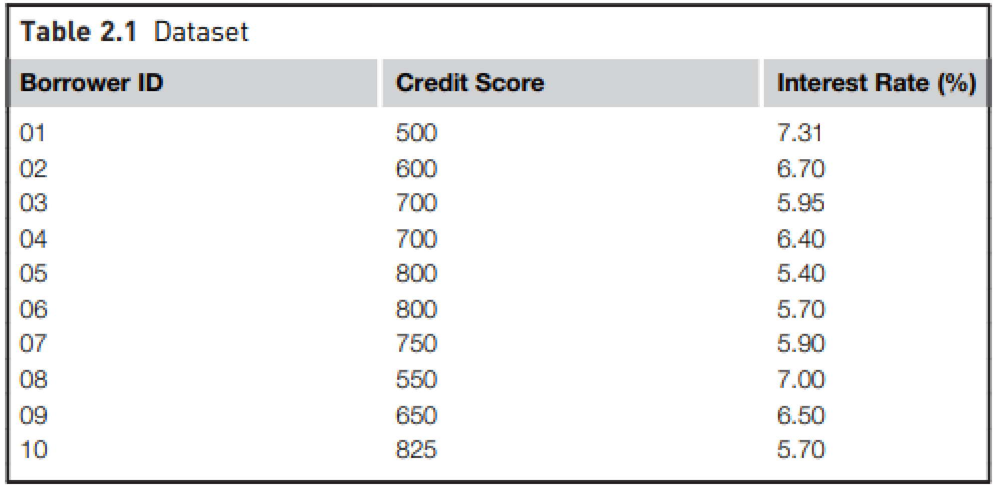


**Table 2.1** Dataset

| Borrower ID | Credit Score | Interest Rate (%) |
|---|---|---|
| 01 | 500 | 7.31 |
| 02 | 600 | 6.70 |
| 03 | 700 | 5.95 |
| 04 | 700 | 6.40 |
| 05 | 800 | 5.40 |
| 06 | 800 | 5.70 |
| 07 | 750 | 5.90 |
| 08 | 550 | 7.00 |
| 09 | 650 | 6.50 |
| 10 | 825 | 5.70 |

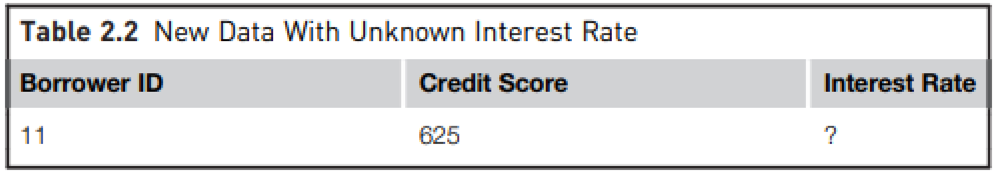


**Table 2.2** New Data With Unknown Interest Rate

| Borrower ID | Credit Score | Interest Rate |
|---|---|---|
| 11 | 625 | ? |

# Data Preprocessing

- Preparing the dataset to suit a data mining task is the most time-consuming part of the process.
- It is extremely rare that datasets are available in the form required by the data mining algorithms.
- Most of the data mining algorithms would require data to be structured.

# Data Preprocessing

- **Data exploration**: in-depth exploration of the data and gaining a better understanding of the dataset.
- **Data quality**: errors in data will impact the representativeness of the model.
- **Missing values**: the most common data quality issues is that some records have missing attribute values.
- **Data type**: different data mining algorithms impose different restrictions on the attribute data types.
- **Transformation**: in some data mining, the input attributes are expected to be numeric and normalized.
- **Outliers**: the presence of outliers needs to be understood and will require special treatments.
- **Feature selection**: a large number of attributes in the dataset significantly increases the complexity of a model and may degrade the performance of the model.
- **Data sampling**: the sample data serve as a representative of the original dataset with similar properties.

# Motivation of Data Preprocessing

- Problem
    - Real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogenous sources.
    - Low-quality data will lead to low-quality mining results.
- Why Preprocess the Data?
    - Incomplete (lacking attribute values or certain attributes of interest, or containing only aggregate data).
    - Inaccurate (containing errors, or values that deviate from the expected).
    - Inconsistent (e.g., containing discrepancies in the department codes used to categorize items).

# Tasks in Data Preprocessing

- Major Tasks in Data Preprocessing
  - Data cleaning.
  - Data integration.
  - Data reduction.
  - Data transformation.

# Data Cleaning

- Missing Values
  - What is missing values? Missing values are the empty values due to the error of recording devices or end-users.
  - What is the problem of missing values? Most of the data mining methods cannot handle missing values.

| ID | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| 1 | 5.1 | 3.4 | 1.4 | 0.2 |
| 2 | 🟥 | 3 | 1.4 | 0.2 |
| 3 | 4.6 | 3.2 | 🟥 | 0.3 |

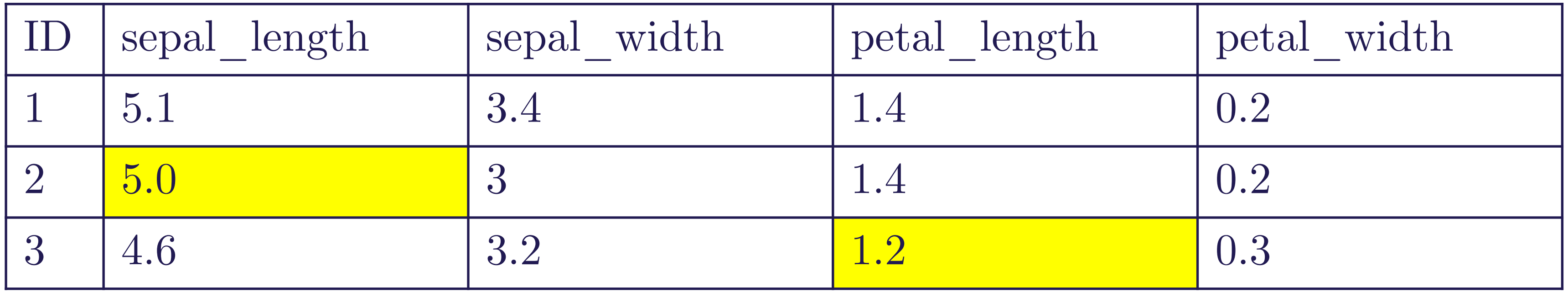
# Data Cleaning

- Missing Values
  - Ignore the tuples.

| ID | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| 1 | 5.1 | 3.4 | 1.4 | 0.2 |

# Data Cleaning

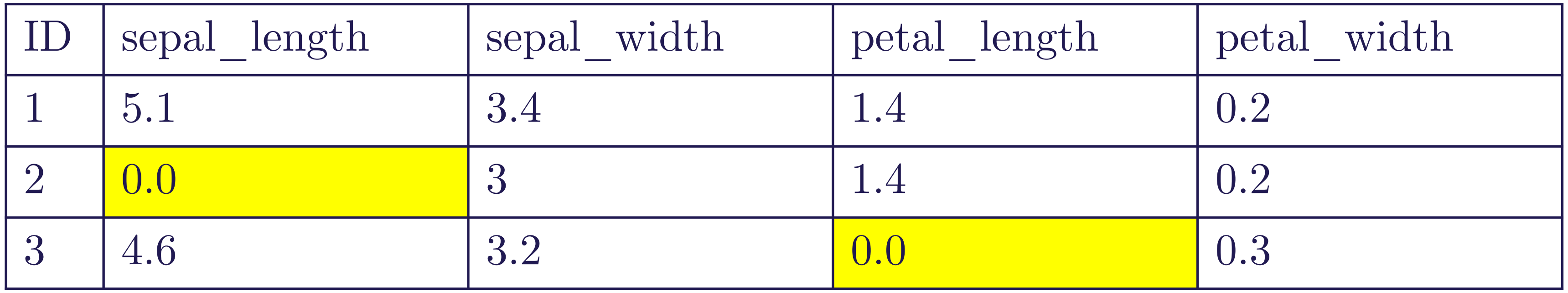
- Missing Values
  - Fill in the missing value manually.



| ID | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| 1 | 5.1 | 3.4 | 1.4 | 0.2 |
| 2 | 5.0 | 3 | 1.4 | 0.2 |
| 3 | 4.6 | 3.2 | 1.2 | 0.3 |

# Data Cleaning

- Missing Values
  - Use a global constant.



| ID | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| 1 | 5.1 | 3.4 | 1.4 | 0.2 |
| 2 | 0.0 | 3 | 1.4 | 0.2 |
| 3 | 4.6 | 3.2 | 0.0 | 0.3 |

# Data Cleaning

- Missing Values
  - Use a measure of central tendency.

$$\text{sepal\_length} = \frac{5.1+4.6}{2}$$

$$\text{petal\_length} = \frac{1.4+1.4}{2}$$

$$x_i = \frac{1}{N}\sum_{1}^{N} x$$

| ID | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| 1 | 5.1 | 3.4 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 4.6 | 3.2 | 1.4 | 0.3 |

# Data Cleaning

- Missing Values
  - Use the most probable value.

$$x = \alpha * y + \beta * z$$

| ID | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| 1 | 5.1 | 3.4 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 4.6 | 3.2 | 1.4 | 0.3 |

# Data Cleaning

- Noisy Data
  - What is noise? Noise is a random error or variance in a measured variable.
  - Sometimes, noise is considered as outliers.

| ID | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| 1 | 5.1 | 3.4 | 1.4 | 0.2 |
| 2 | 10.7 | 3 | 1.4 | 0.2 |
| 3 | 4.6 | 3.2 | 5.3 | 0.3 |

# Data Cleaning

- Noisy Data
  - Smooth data by binning method
  - Predefine bin from minimum values to maximum values
  - Assign the noisy value to a specific bin

| ID | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| 1 | 5.1 => [4.9, 5.1] | 3.4 | 1.4 => [1.2, 1.4] | 0.2 |
| 2 | 10.7 => [5.1, 5.3] | 3 | 1.4 => [1.2, 1.4] | 0.2 |
| 3 | 4.6 => [4.5, 4.7] | 3.2 | 5.3 => [1.6, 1.8] | 0.3 |

# Data Cleaning

- Data Cleaning as a Process
  - Use metadata.
    - Data type
  - Use rules.
    - Unique rule
    - Null rule
    - Consecutive rule

# Data Integration

- Data Integration
  - Data mining often requires data integration—the merging of data from multiple data stores.
  - Careful integration can help reduce and avoid redundancies and inconsistencies.
  - Thus, improve the accuracy and speed of the subsequent data mining process.

# Data Integration

- Entity Identification Problem
  - How can the data analyst or the computer be sure that *customer*_id in one database and *cust_number* in another refer to the same attribute.
  - When matching attributes from one database to another during integration, special attention must be paid to the tructure of the data.

# Data Integration

- Redundancy Problem
  - Redundancy is another important issue in data integration. An attribute (such as annual revenue) may be redundant if it can be “derived” from another attribute or set of attributes.
  - Inconsistencies in attribute or dimension naming can also  cause redundancies in the resulting data set.

# Data Integration

- Tuple Duplication Problem
  - Redundancy is another important issue in data integration. An attribute (such as annual revenue, for instance) may be redundant if it can be “derived” from another attribute or set of attributes.
  - Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

# Data Integration

- Data integration also involves the detection and resolution of data value conflicts. For example, for the same real-world entity, attribute values from different sources may differ.
- This may be due to differences in representation, scaling, or encoding.
- For instance, a weight attribute may be stored in metric units in one system and British imperial units in another.
- For a hotel chain, the price of rooms in different cities may involve not only different currencies but also different services (e.g., free breakfast) and taxes.

# Data Reduction

- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.
- That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

# Data Reduction

- **Dimensionality reduction** is the process of reducing the number of random variables or attributes under consideration.
- Dimensionality reduction methods include wavelet transforms and principal components analysis, which transform or project the original data onto a smaller space.
- Attribute subset selection is a method of dimensionality reduction in which irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed.

# Data Reduction

- **Numerosity reduction** techniques replace the original data volume by alternative, smaller forms of data representation. These techniques may be parametric or non-parametric.
- For parametric methods, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. Regression and log-linear models are examples.
- Nonparametric methods for storing reduced representations of the data include histograms, clustering, sampling, and data cube aggregation

# Data Reduction

- In **data compression**, transformations are applied so as to obtain a reduced or “compressed” representation of the original data.
- If the original data can be reconstructed from the compressed data without any information loss, the data reduction is called lossless. If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called lossy.

# Data Transformation

- Overview
  - In data transformation, the data are transformed or consolidated into forms appropriate for mining so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.
  - Standard data transformation techniques are aggregation, normalization and discretization.

# Data Transformation

- Aggregation
  - Summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.

| ID | name | salary | age |
|---|---|---|---|
| 1 | Jose | 14,5000 | 27 |
| 2 | Remy | 30,000 | 42 |
| 3 | Ben | 7,500 | 35 |
| 4 | Daniel | 24,000 | 31 |
| 5 | John | 16,000 | 22 |
| 6 | Mary | 26,300 | 28 |
| 7 | Walker | 8,700 | 36 |
| 8 |  | (SUM) 127,0008 |  |

# Data Transformation

- Normalization
  - **Min-max Normalization**. Min-max normalization performs a linear transformation on the original data. Suppose that $minA$ and $maxA$ are the minimum and maximum values of an attribute, $A$. Min-max normalization maps a value, $v'_i$, of $A$ to $v'_i$ in the range $[newminA, newmaxA]$
  - $v'_i = \frac{v_i - minA}{maxA - minA}(newmaxA - newminA) + newminA$

# Data Transformation

- Mini-quiz 1
  - Suppose that the minimum and maximum values for the attribute income are \$12,000 and \$98,000, respectively. We would like to map income to the range [0.0,1.0]. By min-max normalization, what is the value of \$73,600?
  - $v'_i = \frac{v_i - minA}{maxA - minA}(newmaxA - newminA) + newminA$
  - $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

# Data Transformation

- Normalization
  - **z-score Normalization**. In z-score normalization, the values for an attribute, $A$, are normalized based on the mean and standard deviation of $A$. $A$ value, $v_i$, of $A$ is normalized to $v'_i$ by computing

$$v'_i = \frac{v_i - A}{\sigma_A}$$

# Data Transformation

- Mini-quiz 2
  - Suppose that the mean and standard deviation of the values for the attribute income are \$54,000 and \$16,000, respectively. With z-score normalization, what is the value of \$73,600?

# Data Transformation

- Mini-quiz 2
  - Suppose that the mean and standard deviation of the values for the attribute income are \$54,000 and \$16,000, respectively. With z-score normalization, what is the value of \$73,600?
  - $v'_i = \frac{v_i - A}{\sigma_A}$
  - $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

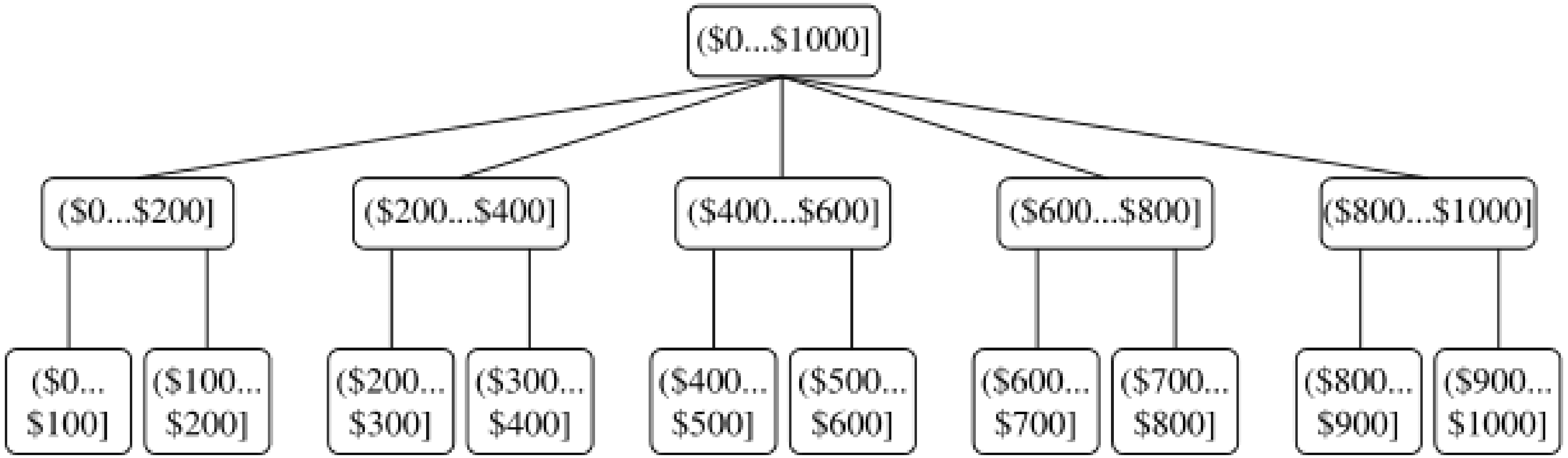
# Data Transformation

- Normalization
  - Decimal Scaling. Decimal scaling normalizes by moving the decimal point of values of attribute $A$. The number of decimal points moved depends on the maximum absolute value of $A$. A value, $v_i$, of $A$ is normalized to $v'_i$ by computing

$$v'_i = \frac{v_i}{10^j}$$

# Data Transformation

- Discretization
  - **Binning**. Binning is a top-down splitting technique based on a specified number of bins. The continuous data is then discretized by putting into appropriate bins



# Data Transformation

- Discretization
  - **Binnning**. Binning is a top-down splitting technique based on a specified number of bins. The continuous data is then discretized by putting into appropriate bins

| ID | name | salary | age |
|---|---|---|---|
| 1 | Jose | 14,500 => [10,000-20,000] | 27 => [20-30] |
| 2 | Remy | 30,000 => [30,000-20,000] | 42 => [40-50] |
| 3 | Ben | 7,500 => [0-10,000] | 35 => [30-40] |
| 4 | Daniel | 24,000 => [20,000-30,000] | 31 => [30-40] |
| 5 | John | 16,000 => [10,000-20,000] | 22 => [20-30] |
| 6 | Mary | 26,300 => [20,000-30,000] | 28 => [20-30] |
| 7 | Walker | 8,700 => [0-10,000] | 36 => [30-40] |

# Data Transformation

- Discretization
  - **Clustering**. A clustering algorithm can be applied to discretize a numeric attribute, A, by partitioning the values of A into clusters or groups so that objects within a cluster are “similar” to one another and “dissimilar” to objects in other clusters.
  - **Decision Tree**. Decision trees can be applied to discretize. Such techniques employ a top-down splitting approach. Unlike the other methods mentioned so far, decision tree approaches to discretize are supervised (use of class label information)..

# Summary

- What is data mining?
- What is data mining process?
- What are data mining types?
- What are data mining settings?
- Why do we need to do the data preprocessing?
- What are data preprocessing tasks?
- How to do data preprocessing?

# References

- Mandatory Reading
  - Jiawei Han, Micheline Kamber, Jian Pei: Data Mining Concepts and Techniques, 3rd Edition, Chapter 1: Introduction
  - Jiawei Han, Micheline Kamber, Jian Pei: Data Mining Concepts and Techniques, 3rd Edition, Chapter 2: Getting to Know Your Data
  - Jiawei Han, Micheline Kamber, Jian Pei: Data Mining Concepts and Techniques, 3rd Edition, Chapter 3: Data Preprocessing

# Exercises

- Download IRIS dataset
  - https://www.kaggle.com/datasets/uciml/iris
  - OR https://scikit-learn.org/1.5/auto_examples/datasets/plot_iris_dataset.html
- Install Python Environment (Conda, WinPython)
- Write code to read IRIS dataset
- Implementing normalization methods and apply them to IRIS