[Introduction to Data Mining]

[Rahmad Mahendra, M.Sc.]

Data Mining for Big Data

Pusat Ilmu Komputer Universitas Indonesia

16 – 20 Juli 2018

Course Objectives

- To know the range of problems that can be solved with data mining.
- To understand what the data is.
- To learn the variation of data mining approaches.

Agenda

- Data Mining and Related Subjects
- Data
- Data Mining Approaches

Data Mining

- **Data Mining:** the **process** of discovering **hidden** and **actionable** patterns from data
- It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems
- Extracting/“mining” knowledge from large-scale data (big data)
- Data-driven discovery and modeling of hidden patterns in big data
- Extracting information/knowledge from data that is
  - implicit,
  - previously unknown,
  - unexpected, and
  - potentially useful

Data Mining vs Database

- **Data mining** is the *process* of extracting hidden and actionable patterns from data

- **Database systems** store and manage data
  - Queries return part of stored data
  - Queries do not extract hidden patterns

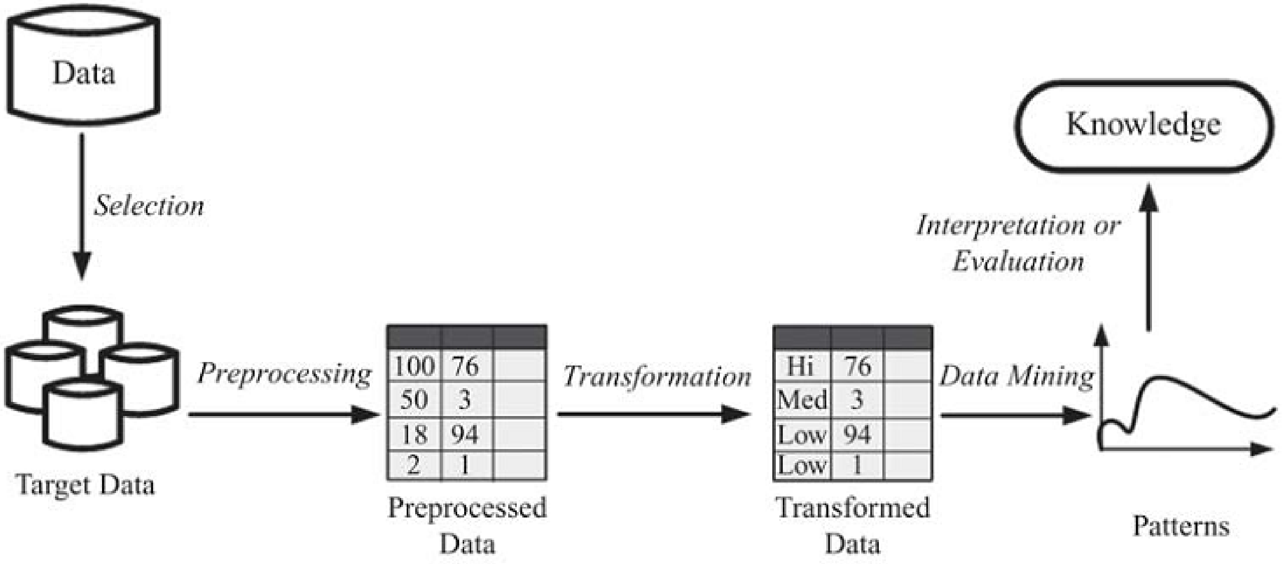
- Examples of querying databases
  - Find all employees with income more than $250K
  - Find top spending customers in last month
  - Find all students from *engineering college* with GPA more than average

Examples of Data Mining Applications

- **Fraud/Spam Detections:** Identifying fraudulent transactions of a credit card or spam emails
  - You are given a user's purchase history and a new transaction, identify whether the transaction is fraud or not;
  - Determine whether a given email is spam or not
- **Frequent Patterns:** Extracting purchase patterns from existing records
  - beer ⇒ dippers (80%)
- **Forecasting:** Forecasting future sales and needs according to some given samples
- **Finding Like-Minded Individuals:** Extracting groups of like-minded people in a given network

Knowledge Discovery

- The process of extracting **useful patterns** from raw data is known as **Knowledge** discovery in databases (KDD)

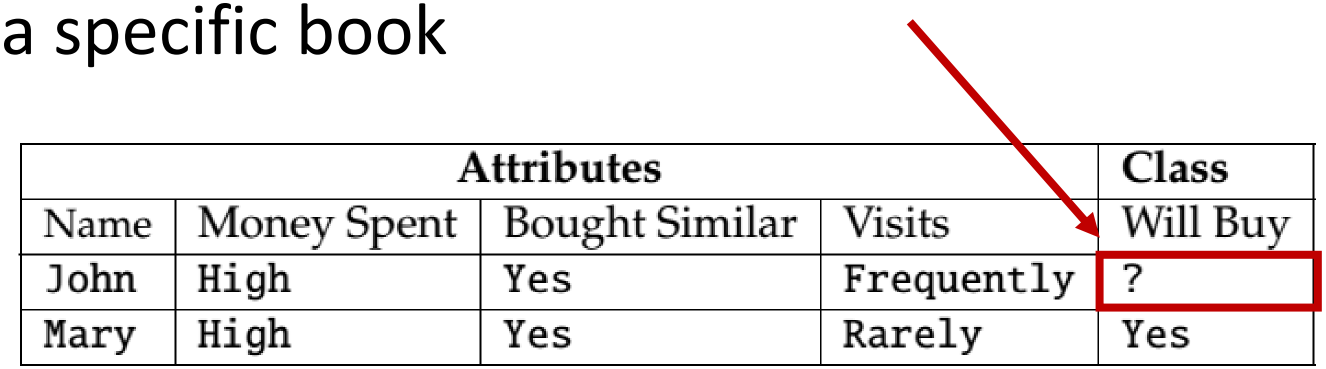


100	76	
50	3	
18	94	
2	1	

Hi	76	
Med	3	
Low	94	
Low	1	

Data

- In the KDD process, data is represented in a **tabular** format.
- Consider the example of predicting whether an individual who visits an online book seller is going to buy a specific book

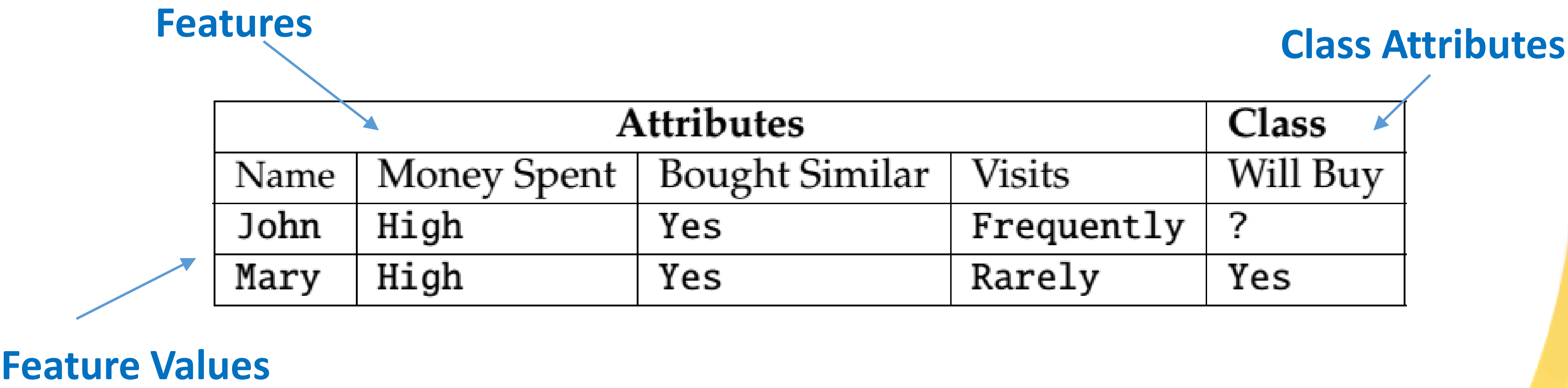


Attributes				Class
Name	Money Spent	Bought Similar	Visits	Will Buy
John	High	Yes	Frequently	?
Mary	High	Yes	Rarely	Yes

John is an example of an **instance**.
A **dataset** consists of one or more instances.

Features

- A dataset is represented using a set of **features (or attributes)**.
- an instance is represented using values assigned to these features.



Features

Class Attributes

Attributes				Class
Name	Money Spent	Bought Similar	Visits	Will Buy
John	High	Yes	Frequently	?
Mary	High	Yes	Rarely	Yes

Feature Values

Type of Feature Values (levels of measurement)

- Nominal (categorical)
  - For instance, a customer’s **name** is a nominal feature
- Ordinal
  - the feature values have an intrinsic order to them
  - In our example, **Money Spent** is an ordinal feature because a High value for Money Spent is more than a Low one
- Interval
  - differences are meaningful whereas ratios are meaningless
  - Example: **time** (e.g. 6:16, 6:45, etc)
- Ratio
  - add the additional properties of multiplication and division

Sample Data – Twitter User

Activity	*Date Joined*	*Number of Followers*	*Verified Account?*	*Has Profile Picture?*
High	2015	50	FALSE	no
High	2013	300	TRUE	no
Average	2011	860000	FALSE	yes
Low	2012	96	FALSE	yes
High	2008	8,000	FALSE	yes
Average	2009	5	TRUE	no
Very High	2010	650,000	TRUE	yes
Low	2010	95	FALSE	no
Average	2011	70	FALSE	yes
Very High	2013	80,000	FALSE	yes
Low	2014	70	TRUE	yes
Average	2013	900	TRUE	yes
High	2011	7500	FALSE	yes
Low	2010	910	TRUE	no

If the data is not Tabular?

- In social media, individuals generate many types of non-tabular data, such as **text, voice, or video**.

- These types of data are first converted to tabular data and then processed using data mining algorithms.
  - voice can be converted to feature values using approximation techniques such as the **fast Fourier transform (FFT)**.
  - To convert text into the tabular format, we can use a process denoted as **vectorization**.

Types of Algorithms

- Supervised Learning
  - We have labeled dataset, that is, instances in this set are tuples in the format **(x, y)**, where **x** is a feature vector and **y** is the class attribute.
  - We learn a mapping **f(.)**, such that **f(x) = y**
  - Example: *Classification* and *Regression*

- Unsupervised Learning
  - the dataset has **NO** class attribute, and our task is to find similar instances in the dataset and group them.
  - Example: Clustering

Thank You