
Data Mining and Analysis

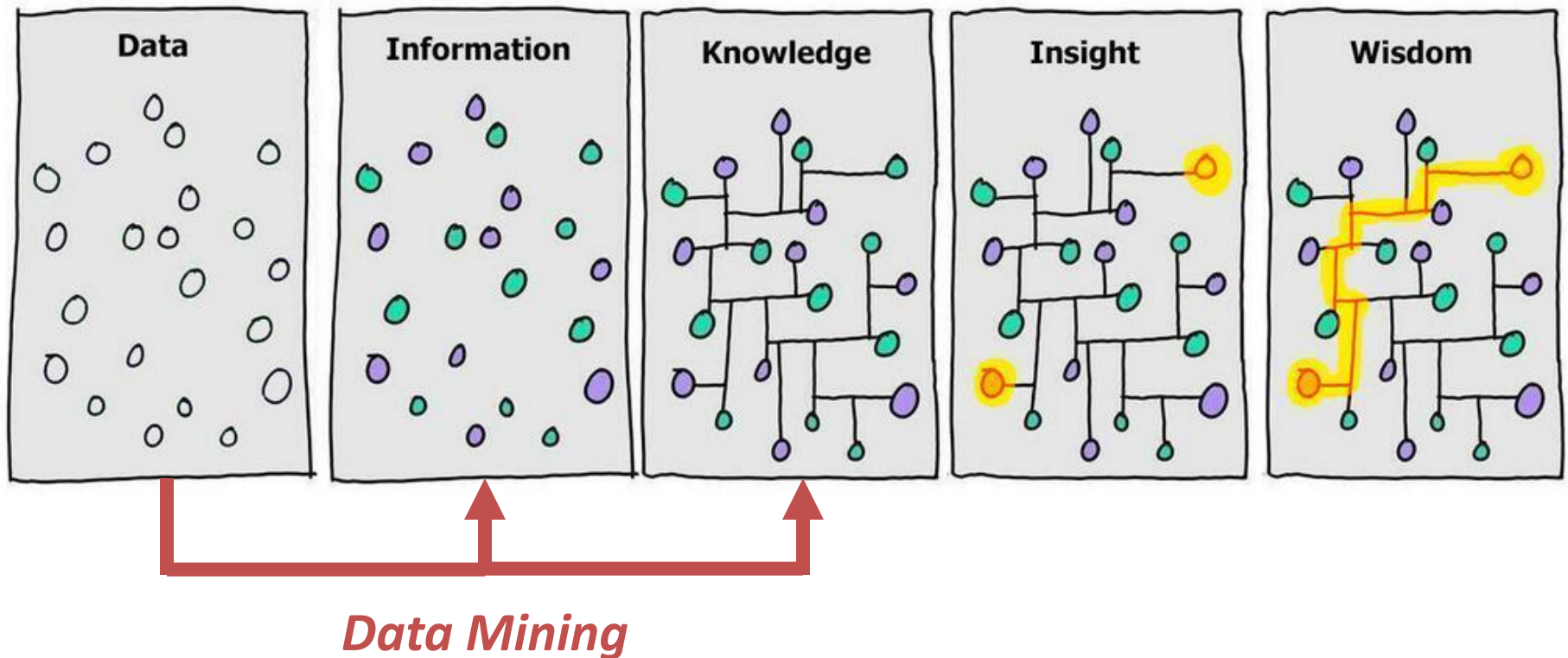
2023 Spring

Instructor: **Ki Yong Lee, Ph.D.**

Professor

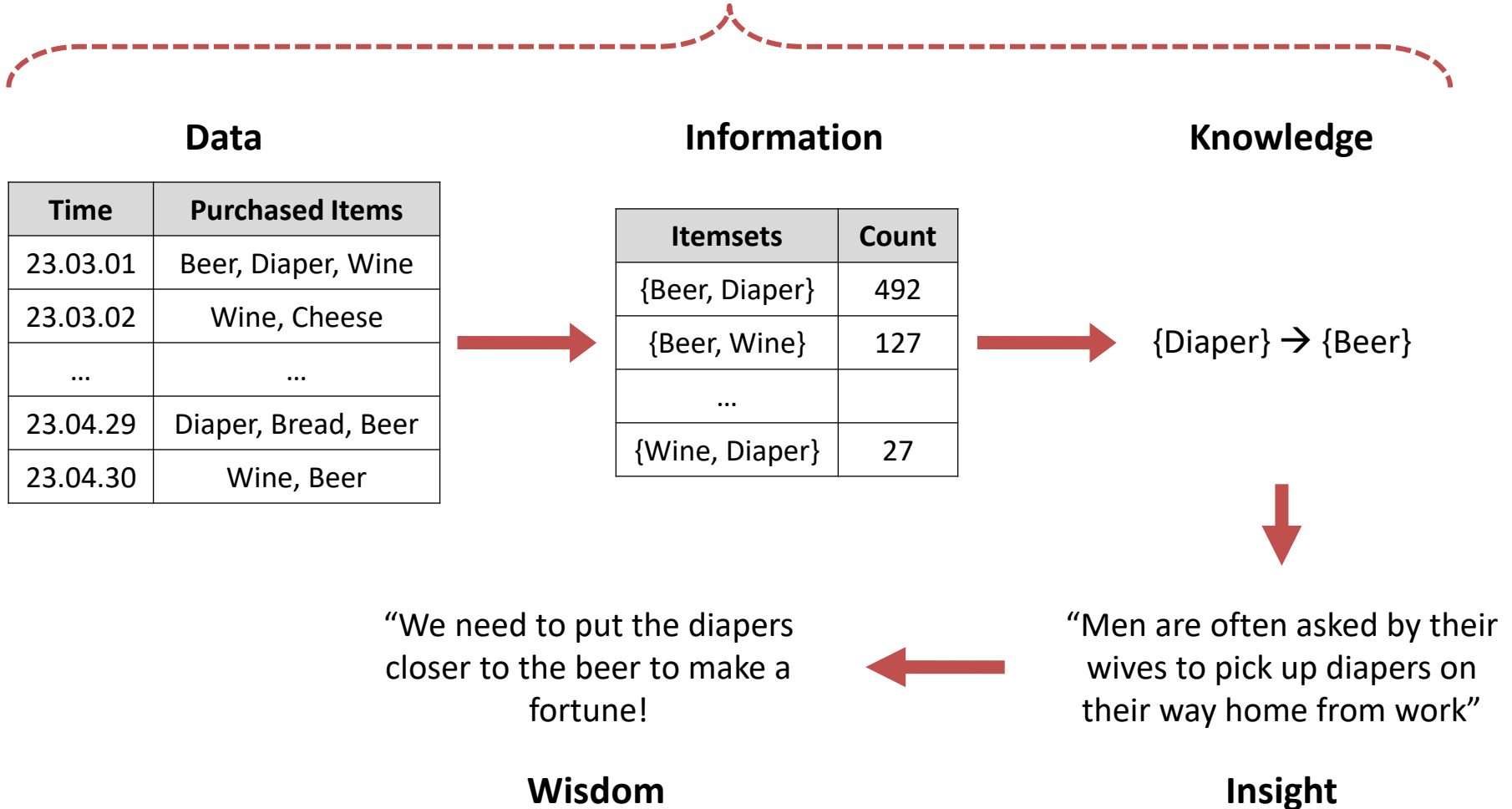
Division of Computer Science
Sookmyung Women's University

Before the Lecture...



Before the Lecture...

Data Mining



Before the Lecture...

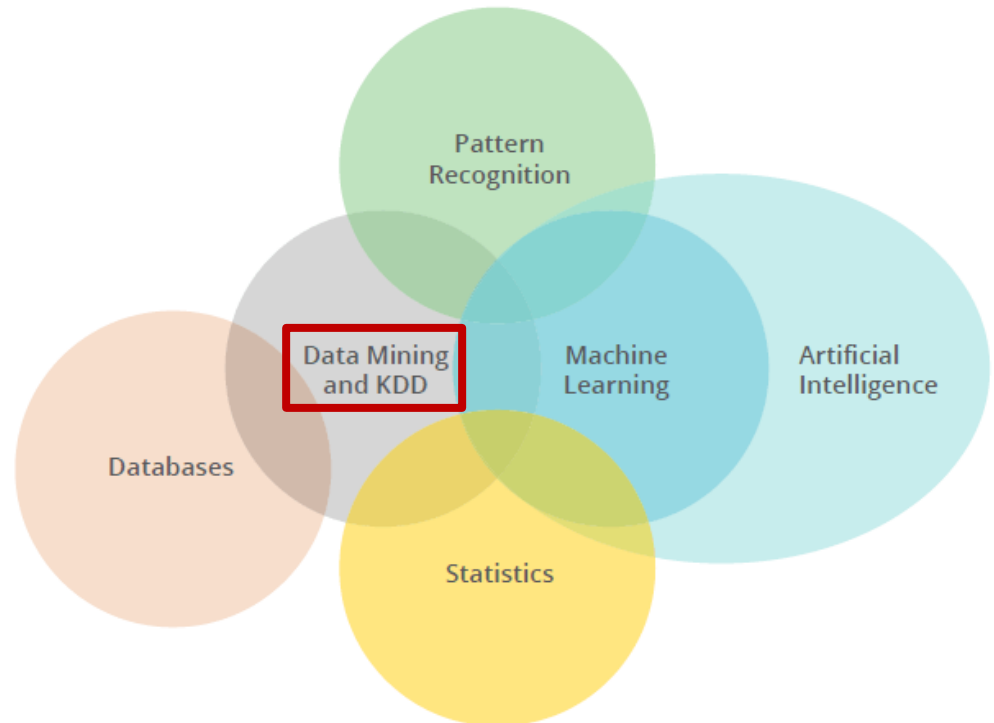
“ Data mining is the practice of searching through large amounts of computerized data to find useful patterns or trends.

Merriam-Webster Dictionary



Before the Lecture...

- What is “data mining”?
 - The process of discovering *hidden patterns* or *knowledge* from *large data*
 - Involves methods from various fields
 - Computer science (*esp.* databases), statistics, machine learning, ...



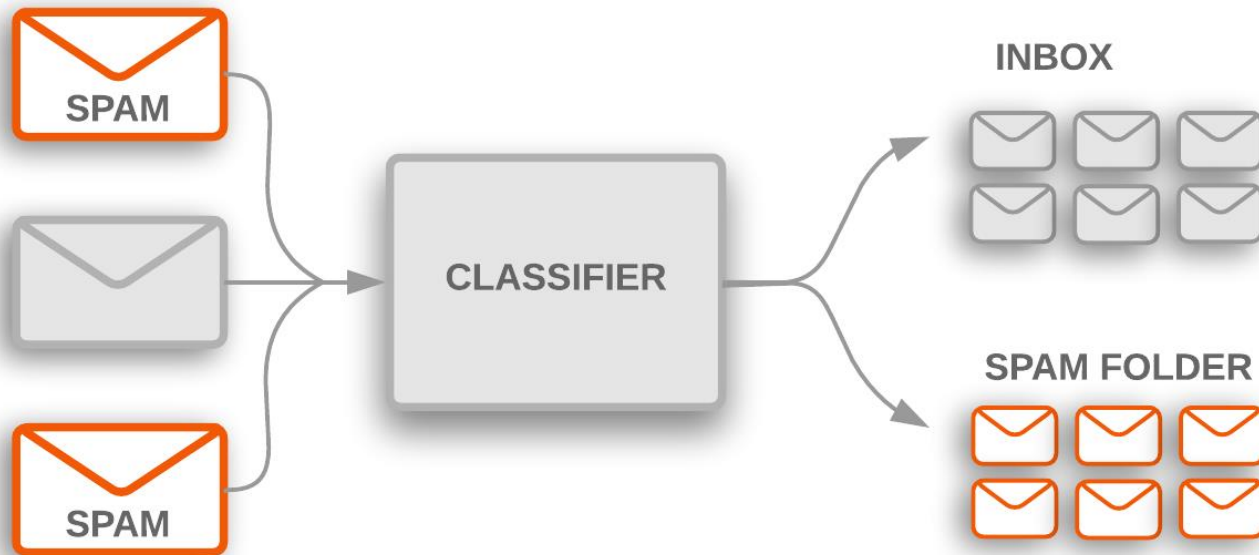
Before the Lecture...

- Four main tasks of data mining
 - ***Classification***
 - Identifying to which category an unseen data belongs (e.g., *spam* or *non-spam*)
 - ***Association analysis***
 - Discovering interesting relations between items (e.g., {*diaper*} → {*beer*})
 - ***Clustering***
 - Grouping similar objects into groups (e.g., finding similar news or customers)
 - ***Anomaly detection***
 - Identifying unusual or rare data (e.g., fraud detection, fault detection)

Before the Lecture...

- Classification

- Identifying to which category an unseen data belongs (*spam* or *non-spam*)









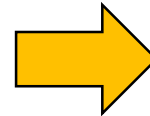
- Algorithms: ***decision tree, naïve Bayes, logistic regression, artificial neural network, ensemble methods, ...***

Before the Lecture...

- Association analysis

- Discovering interesting relations between items ($\{diaper\} \rightarrow \{beer\}$)

Transaction 1	
Transaction 2	
Transaction 3	
Transaction 4	
Transaction 5	
Transaction 6	



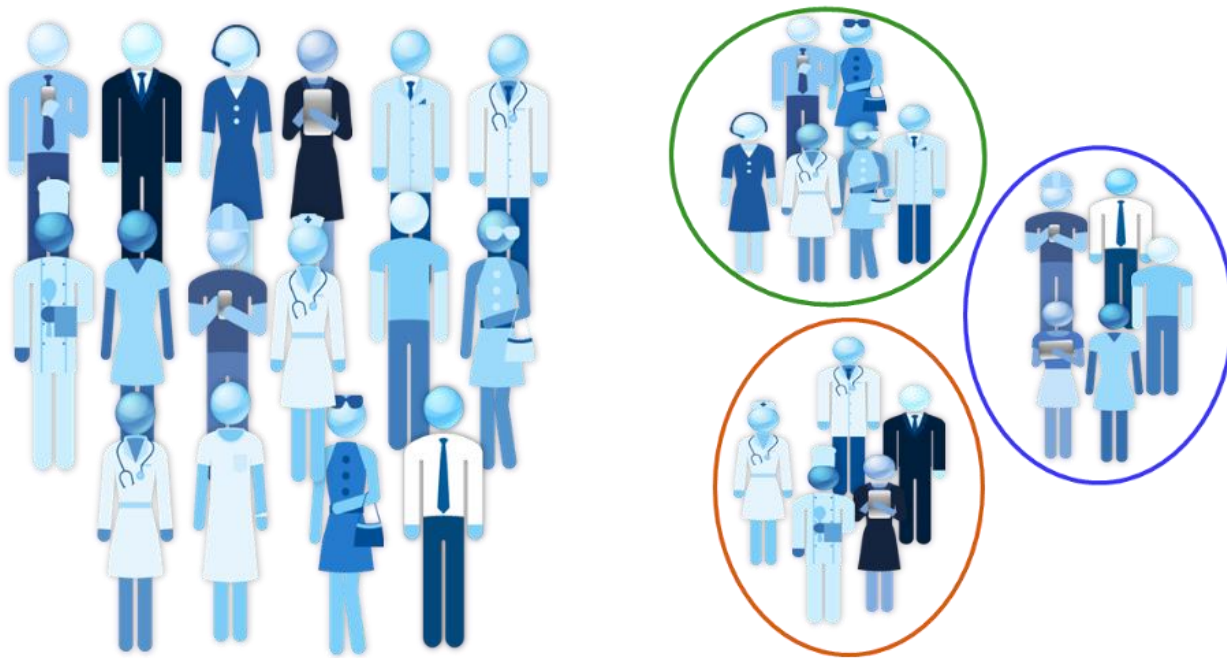
$\{\text{Bread, Butter}\} \rightarrow \{\text{Jam}\}$
(Support = 33.3%)
(Confidence = 66.7%)

- Algorithms: *Apriori*, *FP-growth*, *sequential patterns*, ...

Before the Lecture...

- Clustering

- Grouping similar objects into groups (finding similar news or customers)

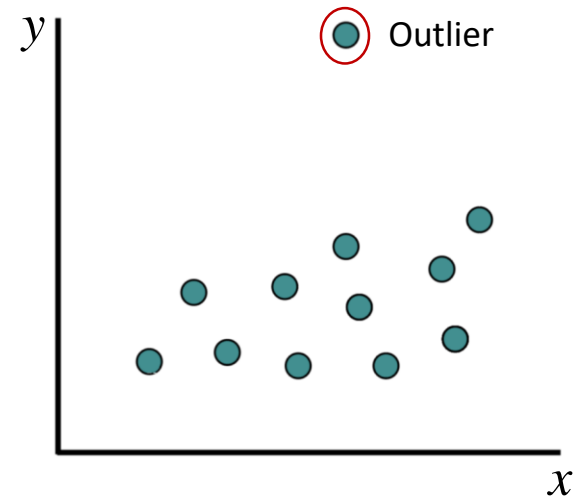
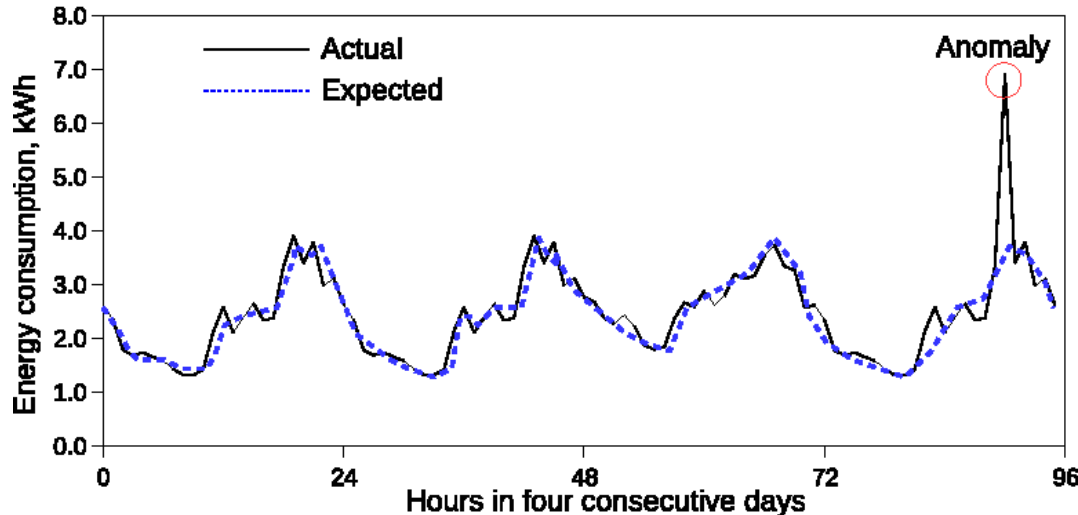


- Algorithms: *k-means*, *hierarchical clustering*, *DBSCAN*, ...

Before the Lecture...

- Anomaly detection

- Identifying unusual or rare data (fraud detection, fault detection)



- Algorithms: ***statistical approaches, proximity-based approaches, clustering-based Approaches, reconstruction-based Approaches, ...***

Two Objectives of This Course

1. Understanding *existing* data mining algorithms
 - Classification, association analysis, clustering, anomaly detection
2. Having the ability to develop *new* algorithms based on the understanding of the existing algorithms
 - For a new problem you are facing



Course Information (1/3)

■ Instructor

- Ki Yong Lee (Professor, Division of Computer Science)
 - Office: Saehim Hall 406
 - Phone: 02-2077-7583, 010-... (upon request)
 - Email: kiyonglee@sookmyung.ac.kr (*most preferred*)
 - Homepage: <http://cs.sookmyung.ac.kr/~kylee>
 - Office hour: You are *always* welcome! (but prior appointment is recommended)

■ Course homepage

- Snowboard → 데이터마이닝및분석 (001)

■ Other communications

- Slack → Please refer to the invitation link in the Snowboard

Course Information (2/3)

- Main topics

- ***Data***

- Types of data, data preprocessing, measures of similarity

- ***Classification***

- Decision tree, k -NN classifier, naïve Bayes, logistic regression, artificial neural network, ensemble methods, etc.
 - Model selection, model evaluation

- ***Association analysis***

- Apriori algorithm, FP-growth algorithm, sequential patterns, etc.

- ***Clustering***

- K-means, hierarchical clustering, DBSCAN, etc.

- ***Anomaly detection***

- Statistical, proximity-based, clustering-based, reconstruction-based, etc.

Course Information (3/3)

■ Textbook

- Pan-Ning Tan et al, “Introduction to Data Mining,” 2/E, Pearson, 2019
 - (Translation) 용환승 외, “데이터 마이닝 (2판),” 휴먼사이언스, 2020

■ Grading policy

- Mid-term Exam : 35%
 - Final Exam : 35%
 - Homework #1 : 10% (a simple(?) data mining program)
 - Homework #2 : 10% (a simple(?) data mining program)
 - Homework #3 : 10% (a simple(?) data mining program)
-
- Total : 100%

Homework

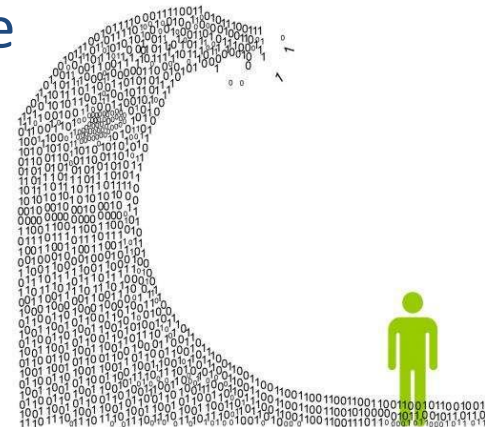
- Goal
 - Implement a simple data analysis program that uses data mining algorithms covered in the class
 - You can use ***Python*** or ***Java***
- Deliverables
 - A program
 - A brief report
- I will guide your progress as much as possible
- ***Details will be announced in the class***

Chapter 1

Introduction

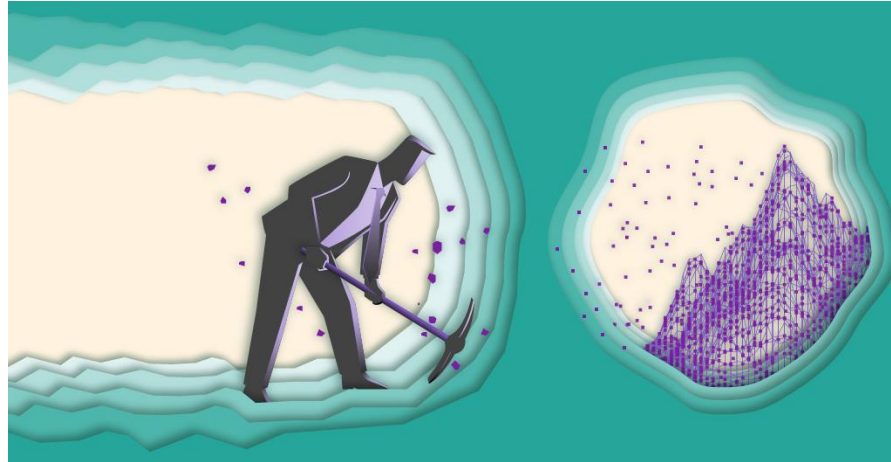
Introduction

- The current age of *big data*
 - The amount of data being collected is growing explosively
 - Triggered by rapid advances in data collection and storage technology
- Deriving *actionable insights* from these large data sets
 - Increasingly important in decision making across almost all areas of society
 - (ex) business, industry, science, engineering, ...
- However, data have become *too great* to analyze
 - 3Vs of big data: volume, velocity, and variety
 - Thus, there is a great need for *methods* and *technology* to extract useful information from these big data



Data Mining

- The process of extracting *useful information* from *large data sets*



- Data mining blends traditional data analysis methods with sophisticated algorithms for processing this abundance of data
 - (ex) traditional statistics + the latest big data technology
- There are many applications that require more advanced techniques for data mining

Applications: Business and Industry (1/2)

- Examples of business intelligence applications

- Point-of-sale data analysis (barcode scanners, RFID, smart cards, etc.)
- Automated buying and selling (e.g., high-speed stock trading)
- Customer profiling
- Targeted marketing
- Store layout
- Fraud detection



- Examples of business questions

- “Who are the most profitable customers?”
- “What products can be cross-sold or up-sold?”
- “What is the revenue outlook of the company for the next year?”

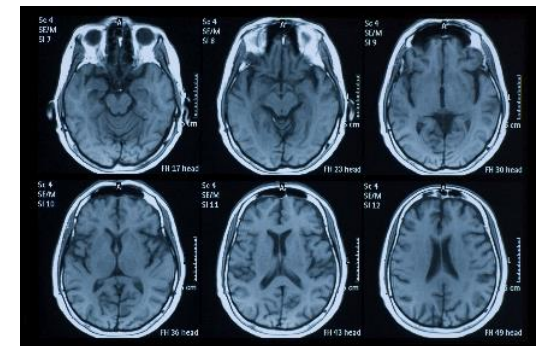
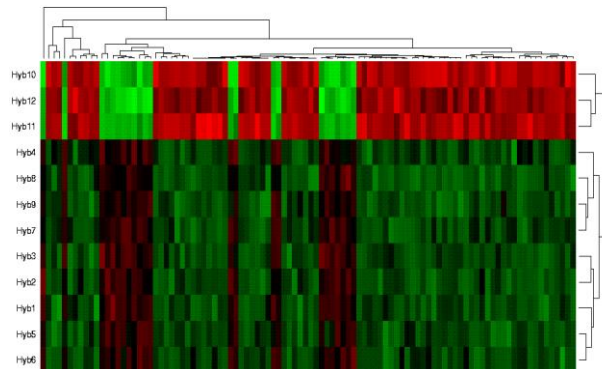
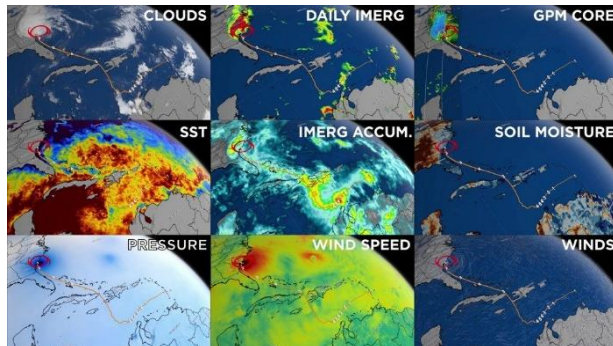
Applications: Business and Industry (2/2)

- Also, there are massive amounts of data on the *Internet*
 - Web browsing, online shopping, messaging, social media postings, ...
 - Can be used for
 - Product recommendation
 - Spam filtering
 - Social connection suggestion
- ***Mobile sensors and devices*** also generate large amounts of data
 - A variety of information about our physical world
 - Collected by smart phones, wearable devices, and physical sensors
 - Can be used for
 - Design of convenient, safe, and energy-efficient home systems (smart home)
 - Urban planning of smart cities



Applications: Science and Engineering

- Examples of large science data
 - Earth's data collected by satellites (land surface, ocean, atmosphere)
 - Can be used to analyze the relationships between observations
 - The large amount of genomic data (microarray data)
 - Can be used to analyze the function of each gene or predict protein structures
 - Electronic health record data (electrocardiograms (ECGs), MRI images)
 - Can be used to provide more personalized patient care



What Is Data Mining?

- Data mining

- The process of ***automatically*** discovering ***useful*** information in ***large*** data
- Find novel and useful patterns that might otherwise remain unknown
 - (ex) Predict the amount a customer will spend at an online store

- ***Not*** all information discovery tasks are data mining

- Simple queries or simple interactions with a database system
- (ex) Find the names of employees in a database whose age is 28

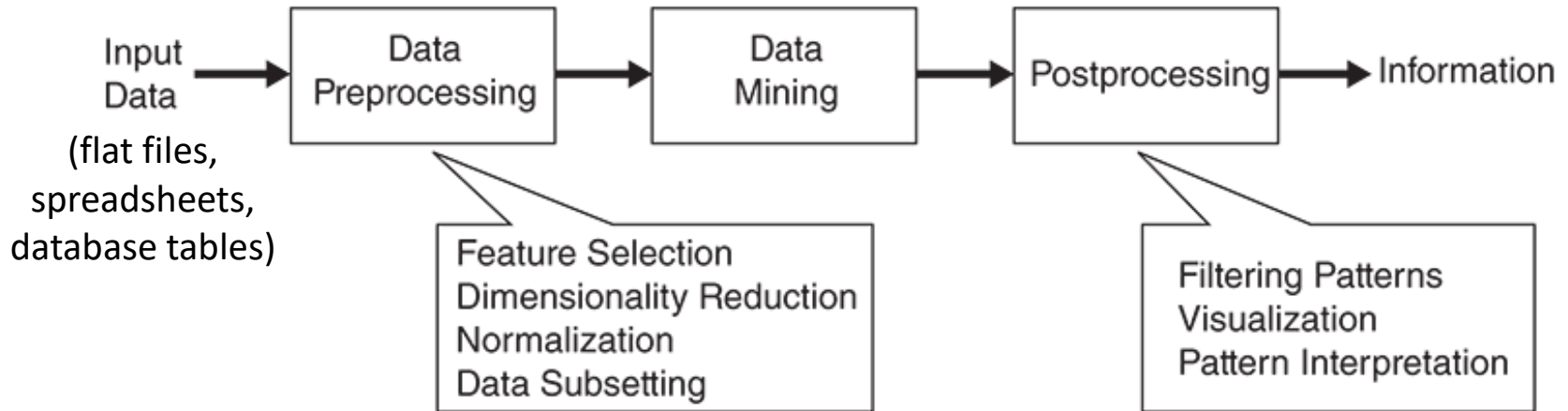
ID	NAME	AGE	CITY	SALARY
1167	Alok Singh	28	New Delhi	36000
1168	Ravi Patel	27	New Delhi	38000
1169	Shubham Shrivastava	29	Noida	48000
1170	Harshit Keshari	25	Chandigarh	38000
1171	Anish Singh	26	Ghaziabad	46500



```
SELECT NAME  
FROM EMPLOYEES  
WHERE AGE = 28;
```

- An integral part of ***knowledge discovery in databases (KDD)***

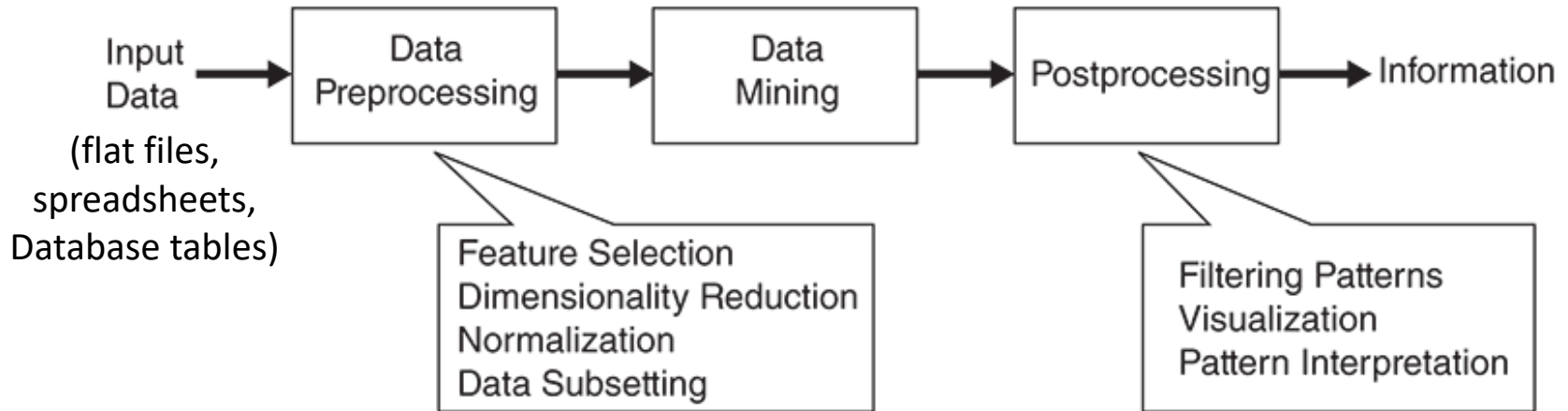
The Process of KDD (1/2)



■ Data preprocessing

- Transforms the raw input data into an appropriate format for analysis
- Examples
 - Fusing data from multiple sources
 - Cleaning data (e.g., remove noise and duplicate observations)
 - Selecting records and features that are relevant to the data mining task
- Perhaps the most laborious and time-consuming step in the overall KDD

The Process of KDD (2/2)



■ Postprocessing

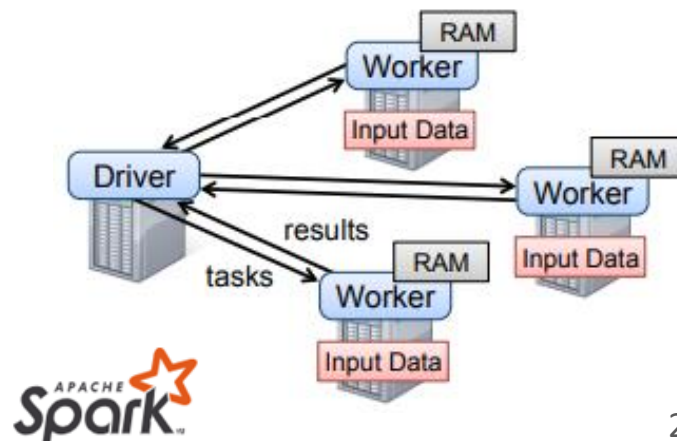
- Ensures that only valid and useful results are incorporated into the decision support system
- Examples
 - Visualization
 - Hypothesis testing (to eliminate spurious data mining results)

Motivating Challenges (1/4)

- Traditional data analysis techniques have practical difficulties in meeting the challenges posed by big data applications

1. Scalability

- Data mining algorithms must handle **massive** data sets (TB, PB, or EB)
- Techniques
 - Special search strategies to handle exponential search problems
 - Novel data structures to access individual records efficiently
 - Out-of-core (disk-based) algorithms
 - Sampling
 - Parallel and distributed algorithms



Motivating Challenges (2/4)

2. High Dimensionality

- It is common for data sets to have ***hundreds*** or ***thousands*** of attributes
 - (ex) In bioinformatics, gene expression data involve thousands of features
- Problems
 - Curse of dimensionality
 - i.e., traditional data analysis techniques don't work well for high-dimensional data
 - The computational complexity increases rapidly

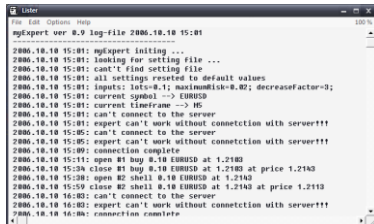
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	wine_type	quality_label
0	7.0	0.17	0.74	12.8	0.045	24.0	126.0	0.99420	3.26	0.38	12.2	8	white	high
1	7.7	0.64	0.21	2.2	0.077	32.0	133.0	0.99560	3.27	0.45	9.9	5	red	low
2	6.8	0.39	0.34	7.4	0.020	38.0	133.0	0.99212	3.18	0.44	12.0	7	white	medium
3	6.3	0.28	0.47	11.2	0.040	61.0	183.0	0.99592	3.12	0.51	9.5	6	white	medium
4	7.4	0.35	0.20	13.9	0.054	63.0	229.0	0.99888	3.11	0.50	8.9	6	white	medium

An example of many attributes (features)

Motivating Challenges (3/4)

3. Heterogeneous data

- Data sets often contain attributes of **different types**
 - (ex) Web and social media data (text, hyperlinks, images, audio, videos)
 - (ex) climate data (temperature, pressure, times, locations, etc.)
- Problems
 - Data mining techniques should consider **complex relationships** in the data



```
File Edit Options Help
ngExpert ver 8.9 log-file 2006.10.10 15:01
2006.10.10 15:01: ngExpert initing ...
2006.10.10 15:01: looking for setting file ...
2006.10.10 15:01: can't find setting file
2006.10.10 15:01: all settings reseted to default values
2006.10.10 15:01: inputs: info=0.1; maximumRisk=0.02; decreaseFactor=0;
2006.10.10 15:01: current symbol -> EURUSD
2006.10.10 15:01: current timeframe -> M5
2006.10.10 15:01: can't connect to the server
2006.10.10 15:01: expert can't work without connection with server!!!
2006.10.10 15:01: can't connect to the server
2006.10.10 15:01: expert can't work without connection with server!!!
2006.10.10 15:01: connection complete
2006.10.10 15:11: open #1 buy 0.10 EURUSD at 1.2143
2006.10.10 15:24: close #1 buy 0.10 EURUSD at 1.2143 at price 1.2143
2006.10.10 15:28: open #2 sell 0.10 EURUSD at 1.2143
2006.10.10 15:59: close #2 sell 0.10 EURUSD at 1.2143 at price 1.2113
2006.10.10 16:03: can't connect to the server
2006.10.10 16:02: expert can't work without connection with server!!!
2006.10.10 16:04: connection complete
```

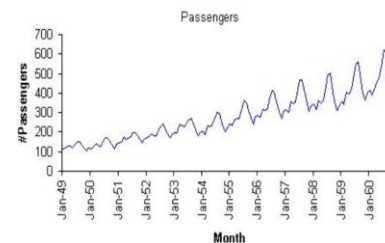
Text data

Transaction ID	Customer ID	Product ID	Purchase date
1112	24221	8977	03-22-2010
1113	24222	8978	03-22-2010
1114	24223	8979	03-22-2010

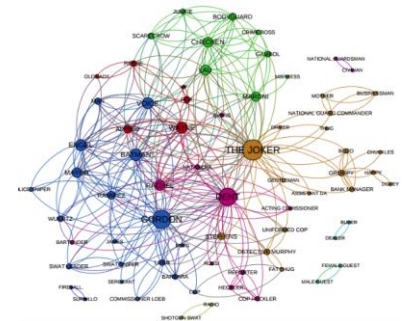
Customer ID	Customer	Address
24221	Bob	123 East street
24222	Alice	223 Main street
24223	Martha	465 North street

Product ID	Name	Price
8977	Banana	.79
8978	TV	400
8979	Watch	50

Table data



Sequence data



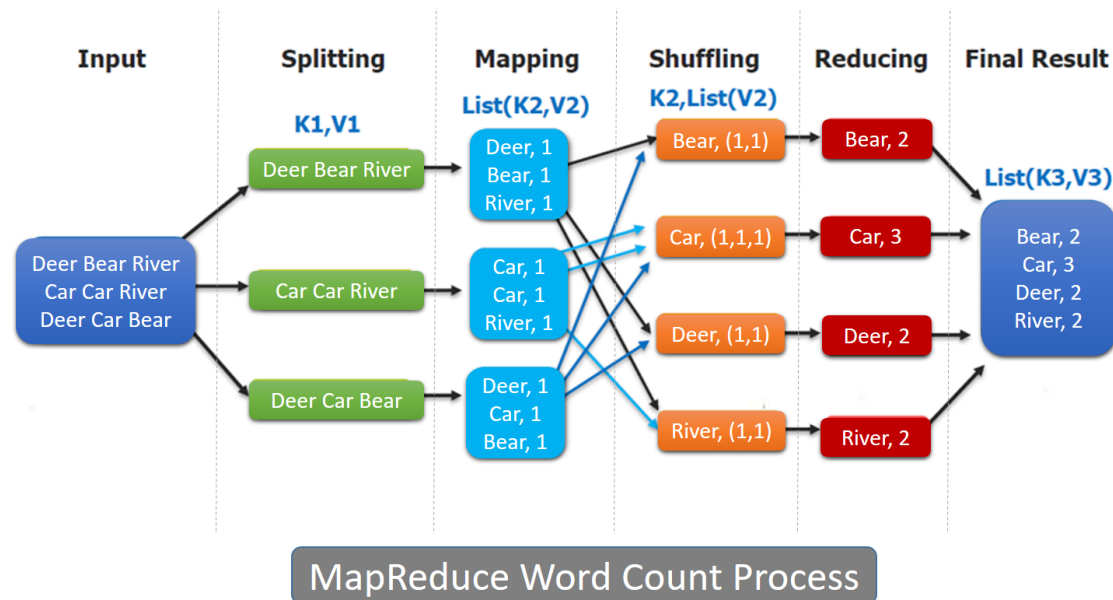
Graph data

Examples of heterogeneous data

Motivating Challenges (4/4)

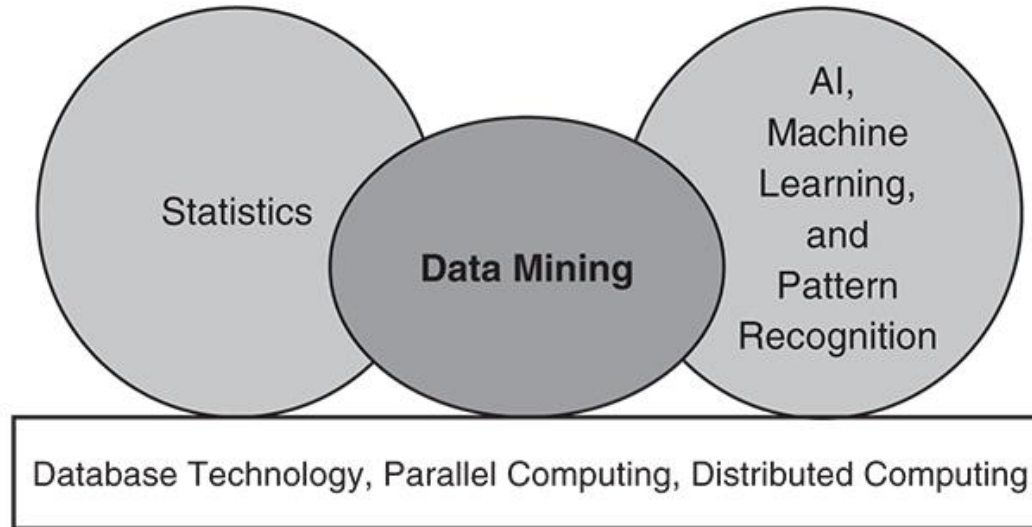
4. Data distribution

- Sometimes, the data is distributed across **multiple** locations
- Key challenges
 - How to reduce the amount of communication needed to perform the computation
 - How to divide the task across multiple locations and merge the partial results obtained from each location



Disciplines Related to Data Mining

- Data mining employs techniques from *many* disciplines

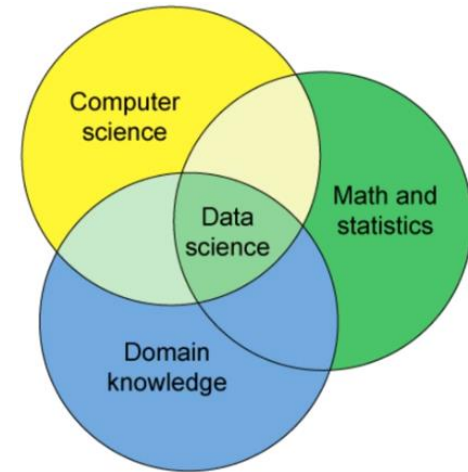


- **Statistics** (sampling, estimation, hypothesis testing, ...)
- **AI** (machine learning, modeling, learning theories, ...)
- **Database** (efficient storage, indexing, query processing, ...)
- **Parallel/distributed computing** (processing data of massive size)
- Others (optimization, information theory, visualization, IR, ...)

Data Science and Data Mining

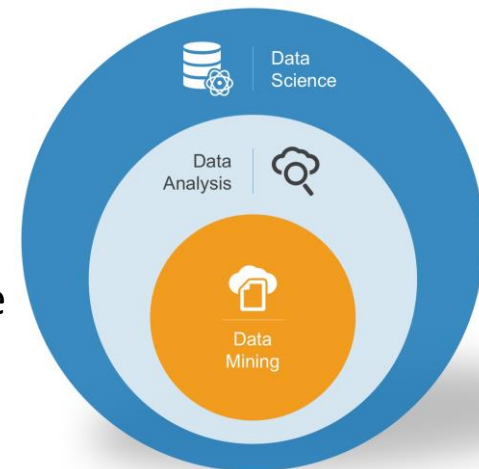
■ Data science

- An interdisciplinary field that studies and applies tools and techniques for deriving useful insights from data
- Emerged as a new field because none of the existing areas provides a complete set of tools for the data analysis tasks
 - Programming skill + math/statistical skill + **domain skill**



■ Data mining

- Emphasizes the direct discovery of patterns and relationships from data, especially in large data sets
 - Often **without** the need for extensive domain knowledge
- Mainly about finding useful information in a dataset



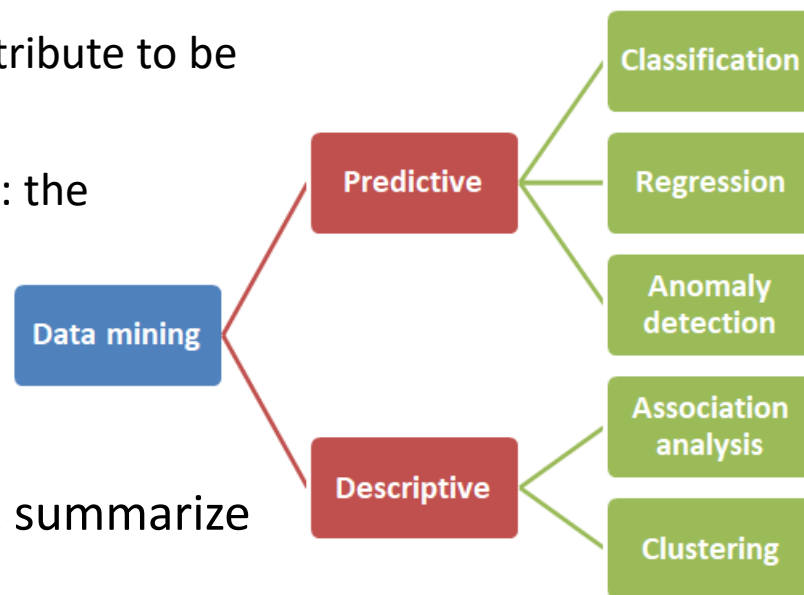
Two Major Categories of Data Mining Tasks

■ Predictive tasks

- The objective is to ***predict*** the value of a particular attribute based on the values of other attributes
 - Target (or dependent) attribute: the attribute to be predicted
 - Explanatory (or independent) attribute: the attributes used for prediction

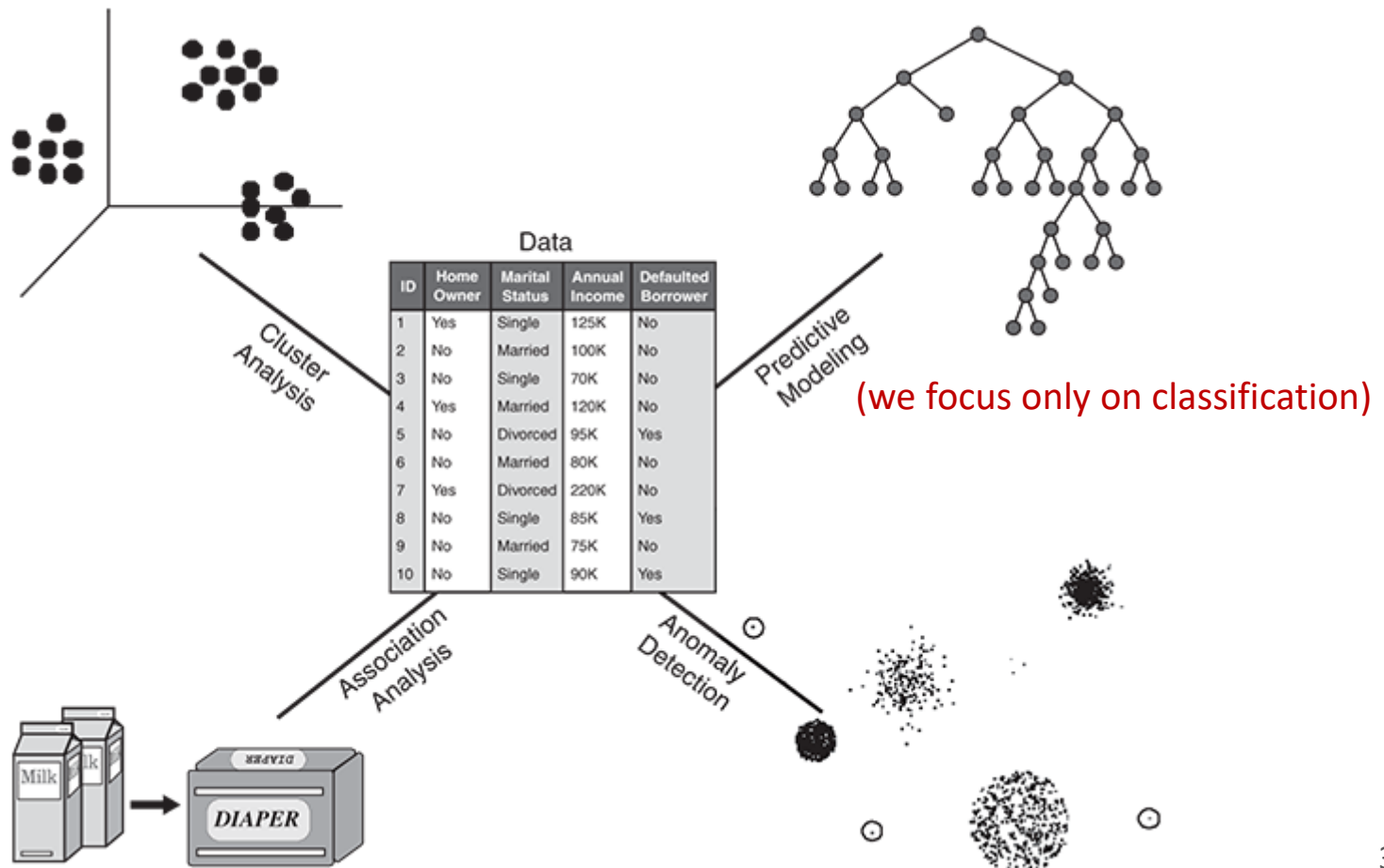
■ Descriptive tasks

- The objective is to ***derive patterns*** that summarize the underlying relationships in data
 - (ex) frequent patterns, correlations, trends, clusters
- Frequently require postprocessing techniques to validate and explain the results



Four Core Data Mining Tasks

- In this course, we consider four of the core data mining tasks
 - Classification, association analysis, clustering, anomaly detection

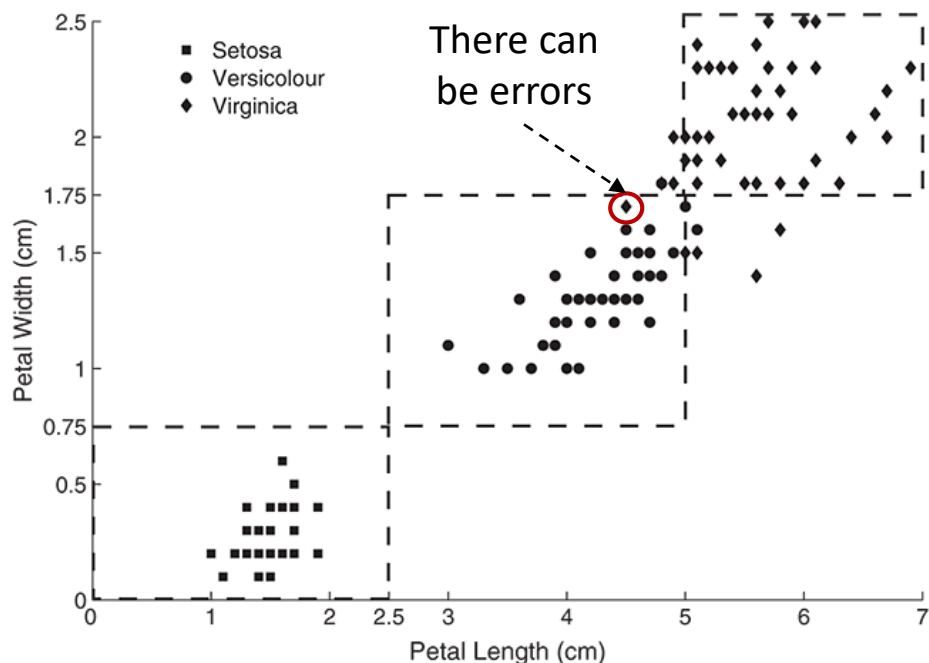


1. Classification

- Predictive modeling: the task of building a **model** $y = f(\mathbf{x})$
 - y : the target variable
 - $\mathbf{x} = (x_1, x_2, \dots, x_n)$: the explanatory variables
- Two types of predictive modeling
 - **Classification**: when y is categorical (e.g., ‘cat’(0), ‘dog’(1), ‘human’(2))
 - (ex) predicting whether a web user will make a purchase ($y = 1$ (yes) or 0 (no))
 - **Regression**: when y is numerical (e.g., any value from $[-20, 50]$)
 - (ex) forecasting the future temperature of an area ($y = 32.7^\circ\text{C}$)
- Goal
 - **Learn a model $f(\mathbf{x})$ that minimizes the error** between the predicted and true values of the target variable

(Ex) Predicting the Type of a Flower

- Consider the task of predicting a species of flower based on the characteristics of the flower
 - Species: $y \in \{\text{'Setosa'}, \text{'Versicolour'}, \text{'Virginica'}\}$
 - Petal length: $x_1 \in [0, 7]$
 - Petal width: $x_2 \in [0, 2.5]$



- We may build a model $f(x_1, x_2)$ as follows:

$$f(x_1, x_2) = \begin{cases} \text{'Setosa'} & \text{if } x_1 \in [0, 2.5) \\ & x_2 \in [0, 0.75) \\ \text{'Versicolour'} & \text{if } x_1 \in [2.5, 5) \\ & x_2 \in [0.75, 1.75) \\ \text{'Virginica'} & \text{if } x_1 \in [5, 7] \\ & x_2 \in [1.75, 2.5] \end{cases}$$

2. Association Analysis

- Discover patterns that describe strongly *associated* items in the data
 - (ex) find groups of products that are bought together
 - (ex) identify web pages that are accessed sequentially
- The discovered patterns are typically represented in the form of implication rules or item subsets
 - (ex) $\{Diapers\} \rightarrow \{Beer\}$ (association rule)
 - (ex) $\{Milk, Ham, Bread\}$ (frequent itemset)
- Goal
 - Extract the most *interesting* patterns in an *efficient* manner
 - Because of the exponential size of its search space

(Ex) Market Basket Analysis

- We are given point-of-sale data collected at the checkout counters of a grocery store
- Association analysis can be applied to find items that are frequently bought together
 - (ex) $\{Diapers\} \rightarrow \{Milk\}$
- This types of rule can be used to identify potential cross-selling opportunities among related items

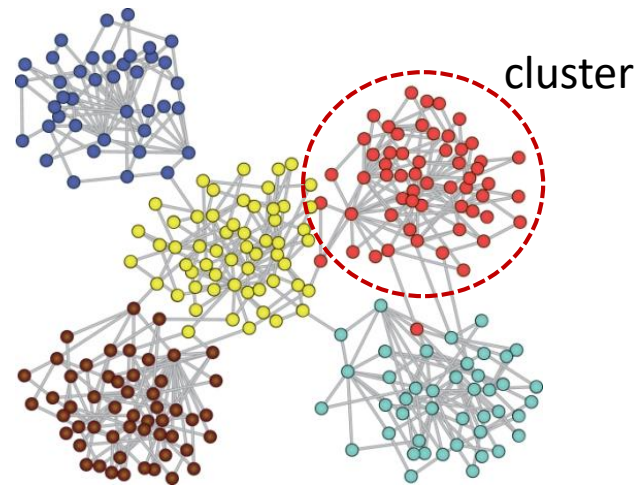
Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

3. Cluster Analysis

- Find groups of *closely related* observations so that
 - Observations in the same cluster are *similar* to each other
 - Observations in different clusters are *dissimilar* to each other
- Application examples
 - Market research: group similar customers (customer segmentation)
 - Social networks: recognize communities within large groups of people



Customer segmentation



Community detection

(Ex) Document Clustering

- We are given a collection of news articles
 - Each article is represented as a set of word-frequency pairs (i.e., $w: c$)
 - w is a word and c is the number of times the word appears in the article

Article	Word-frequency pairs	
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2	Cluster 1 (economy)
2	machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1	
3	job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3	
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2	
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2	Cluster 2 (healthcare)
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3	
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2	
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1	

- Clustering can be applied to group similar new articles
 - A good clustering algorithm should be able to identify the two clusters

4. Anomaly Detection

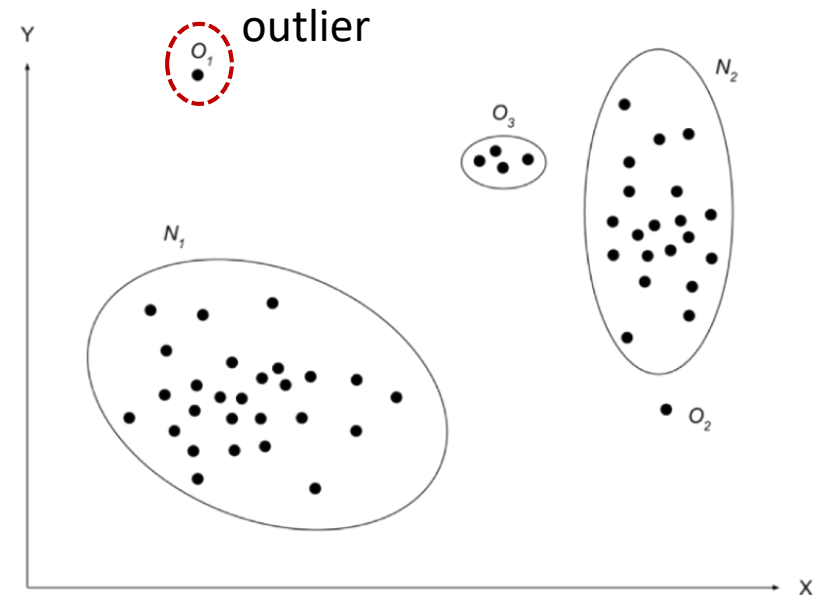
- Identify observations whose characteristics are *significantly different* from the rest of the data
 - Such observations are called **anomalies** or **outliers**

- Goal

- Discover the real anomalies and avoid falsely labeling normal objects as anomalous
 - A high detection rate
 - A low false alarm rate

- Application examples

- The detection of fraud, network intrusions, unusual patterns of disease, and ecosystem disturbances (e.g., droughts, floods, fires, hurricanes)



(Ex) Credit Card Fraud Detection

- A credit card company records the transactions made by every credit card holder
 - Along with personal information (e.g., credit limit, age, income, address)
- Anomaly detection techniques
 - ① Build a profile of legitimate transactions for the users
 - ② When a new transaction arrives, compare it against the user's profile
 - ③ If the characteristics of the transaction are very different from the previously created profile, then flag it as potentially fraudulent

