1

Introduction

IT 326: Data Mining

First Semester 2021-2022

Chapter 1, "Data Mining: Concepts and Techniques" (3rd ed.)

Outline

- Why Data Mining?
- What Is Data Mining?
- Knowledge Discovery (KDD) Process
- Data Mining Tasks
- Data Mining Functions
- Summary

Why Data Mining? ... ?

to get knowledge ✓

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability → a lot of data
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of data:
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, cameras, YouTube
- We are drowning in data, but starving for knowledge!
- Data mining → Automated analysis of massive data sets

↑ طرائق to get knowledge

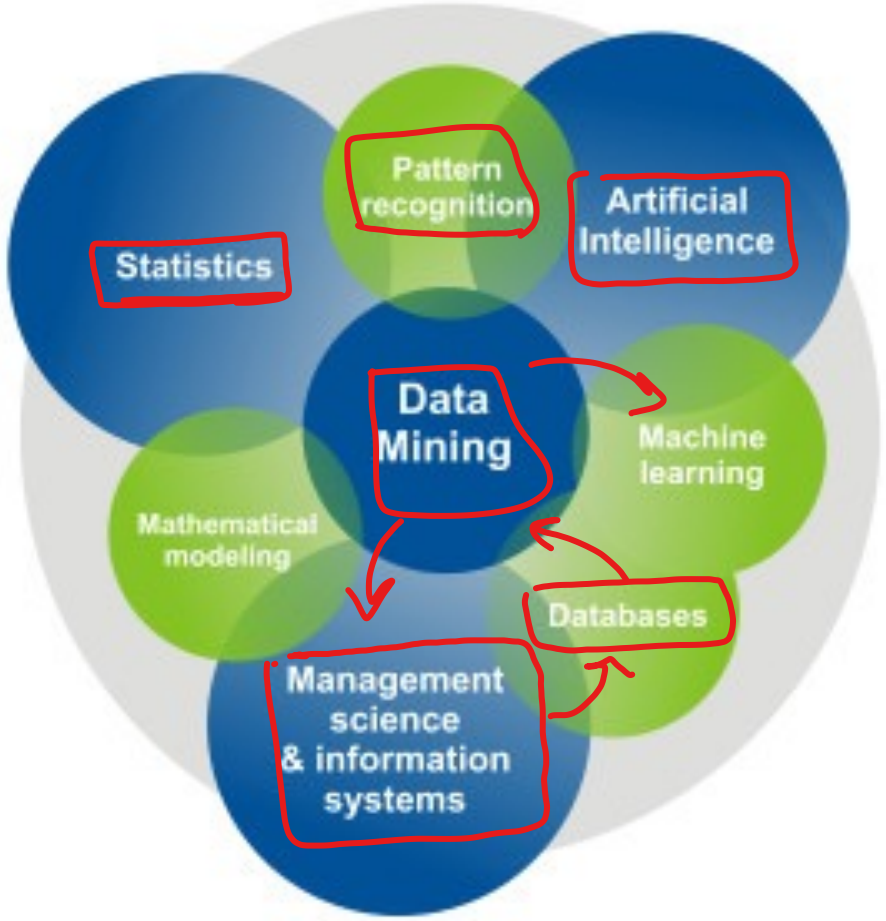
What is Data Mining?

- **Data mining** (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.
- Alternative names:
  - Knowledge discovery (mining) from data (KDD), knowledge extraction, data/pattern analysis, business intelligence, etc...

Data Mining – interdisciplinary

تربط أكثر من علم مع بعض

يستزم أكثر من علم



What is Data Mining?

What is the difference between data mining and database query?

Difference

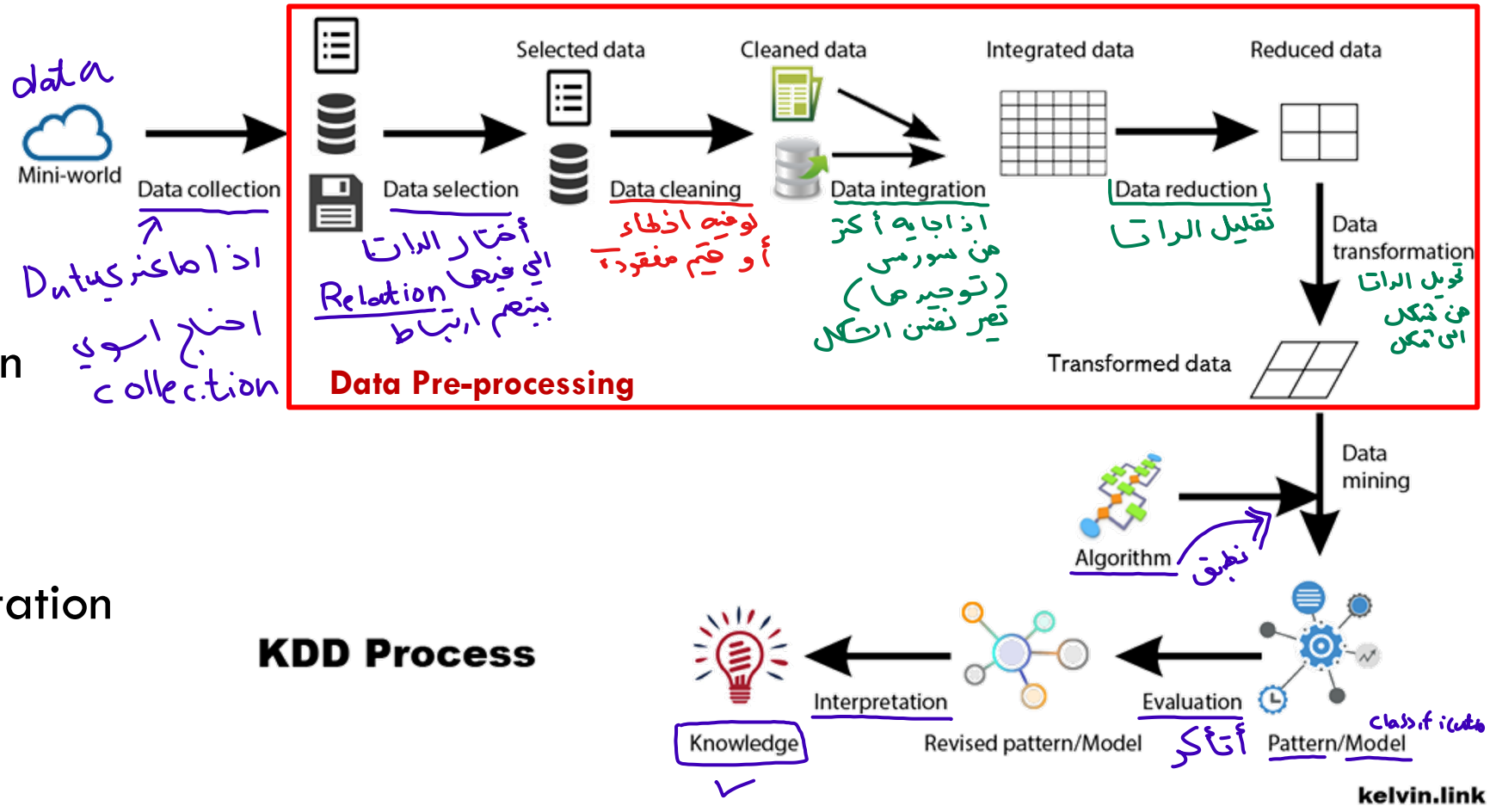
Database Query	Data Mining Tasks
Find all credit applicants with last name of Smith.	Find all credit applicants who are high credit risks. (classification)
Identify customers who have purchased more than $10,000 in the last month.	Identify customers with similar buying habits. (Clustering)
Find all customers who have purchased milk	Find all items which are frequently purchased with milk. (association rules)

Knowledge Discovery (KDD) Process

- Data mining plays an essential role in the knowledge discovery process:

1. Data collection
2. Data selection
3. Data cleaning
4. Data integration
5. Data transformation
6. Data mining
7. Pattern evaluation
8. Knowledge presentation

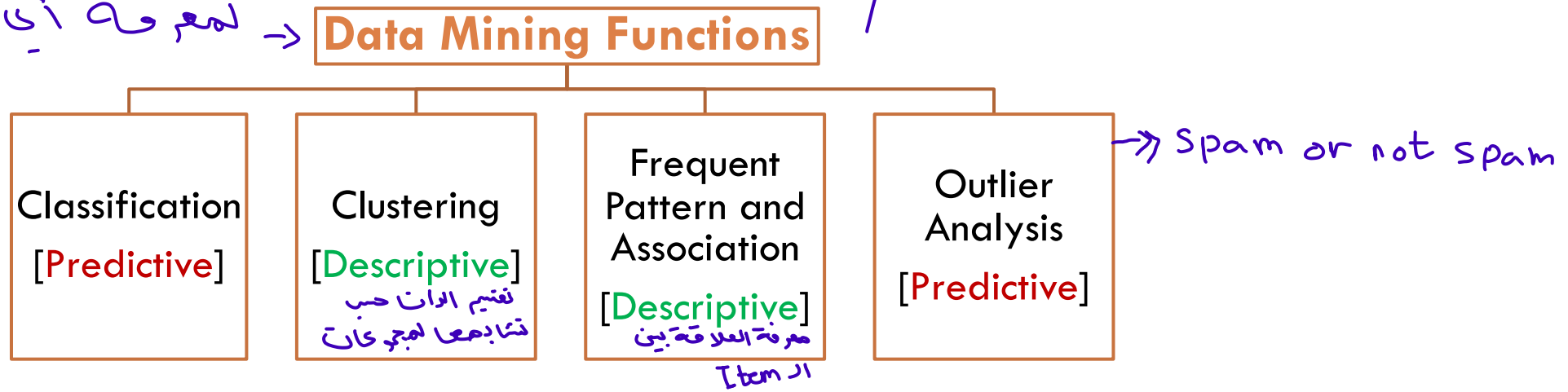
Data Pre-processing



Data Mining Tasks

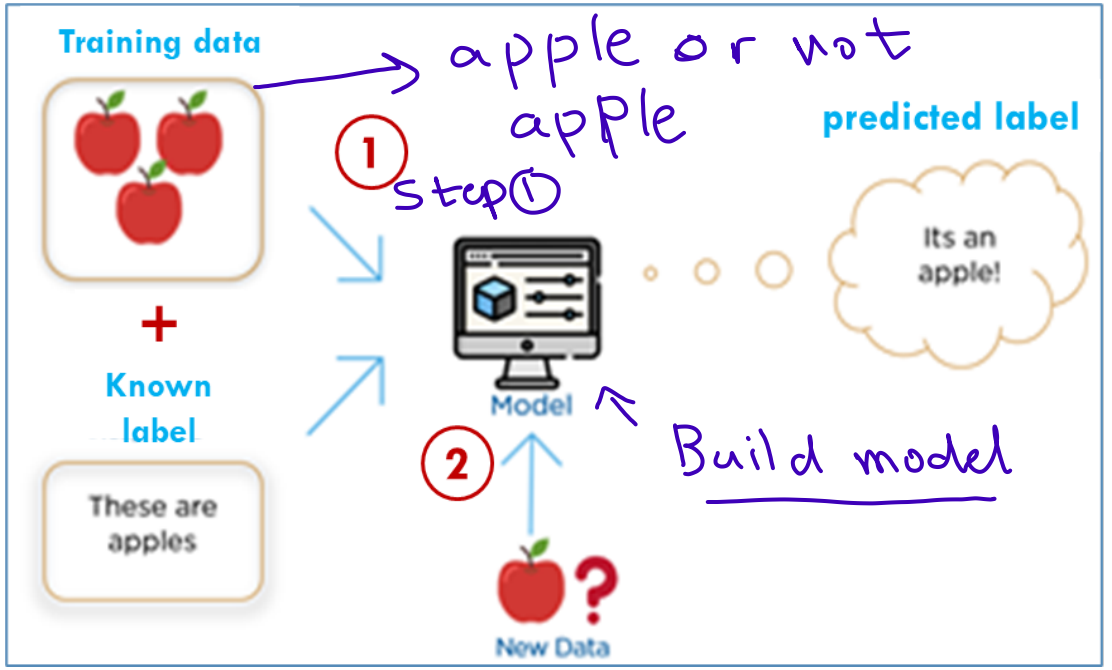
- Two main types of data mining tasks: ⇒ نوعين
  1. **Descriptive** mining tasks characterize properties of the data in a target data set.
  2. **Predictive** mining tasks perform induction on the current data in order to predict values of new data. بناء على داتا سابقة
- Data mining functions are used to specify the kinds of patterns to be found in data mining tasks.

لمعرفة أي نوع داتا ماينق ←



Classification

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.
- The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are **known**).
- The model is used to predict the class label of objects for which the class label is **unknown**.

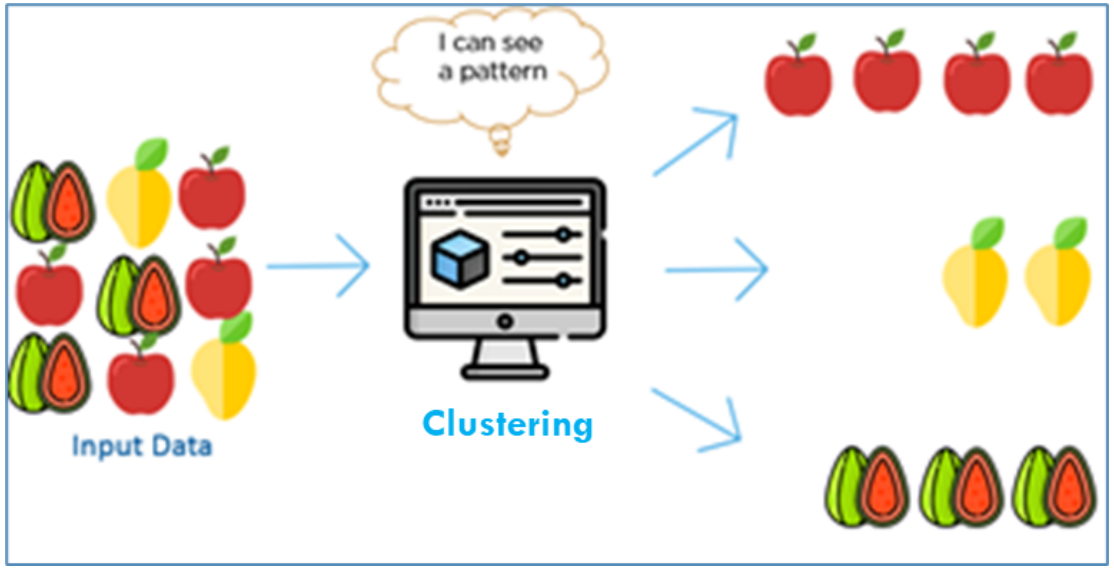


Classification: Application

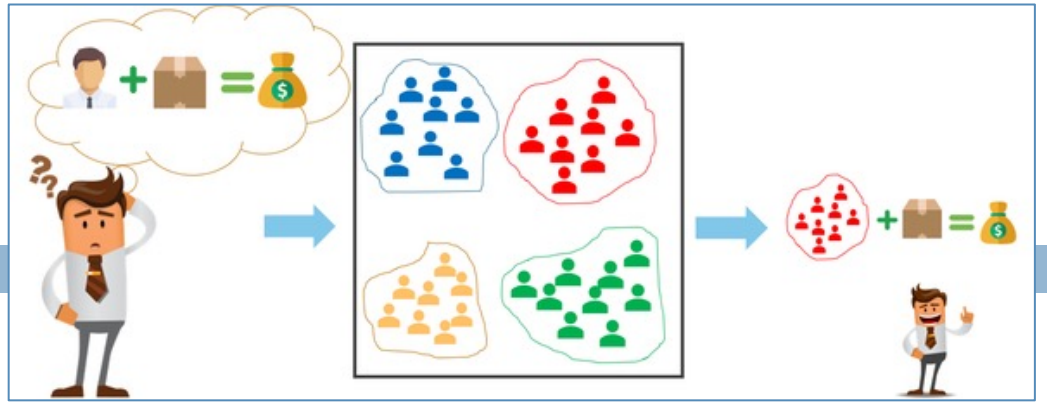
- Direct Marketing:
  - Goal: Reduce cost of mailing/advertising by targeting a set of consumers likely to buy a new product.
  - Approach:
    - Use the data for a similar product introduced before. We know which customers decided to buy and which decided otherwise. This {**buy, don't buy**} decision forms the **class label**.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc…
    - Use this information as input attributes to train a classifier model.

Clustering

- Clustering analyzes data objects without labels.
  - can be used to generate class labels for a group of objects.
- The objects are clustered or grouped based on the similarity.
  - That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.



Clustering: Application

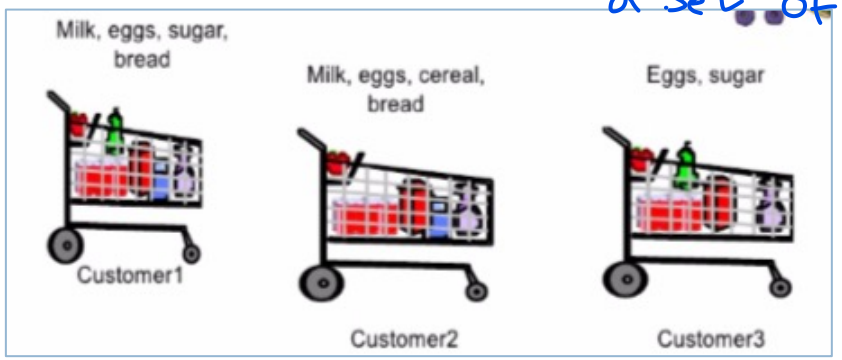


- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix. ⇒
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers. ✓
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Frequent patterns, Association and Correlation Analysis

- Frequent patterns are patterns that occur frequently in data.
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.
- Example: What items are frequently purchased together in the supermarket?

Transactional Data → Data has a set of item. (Set of input)



Association rule: → output

Bread → Milk [50%, 75%]

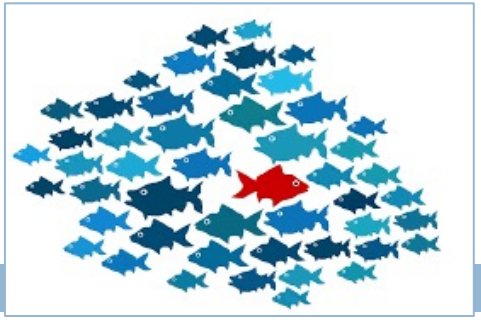
Support= 50% (milk + bread / total)

(means that 50% of all the transactions under analysis show that bread and milk are purchased together).

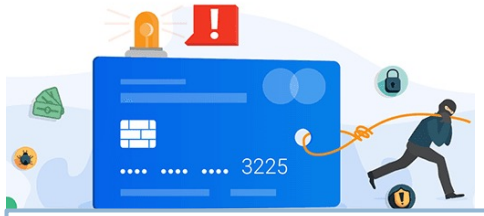
Confidence= 75% (milk + bread / bread)

(means that if a customer buys Bread, there is a 75% chance that she will buy Milk as well.)

Outlier Analysis



- Outlier: A data object that does not comply with the general behavior of the data.
- Useful in fraud detection, and rare events analysis.
- Example: "Find unusual activity in a client's banking transactions"?



unusual activity

Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of **unusually large amounts** for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

Summary

- Data mining is the discovery of interesting patterns and knowledge from massive amount of data.
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation.
- Mining can be performed in a variety of data.
- Data mining functions: classification, clustering, association and outlier analysis.