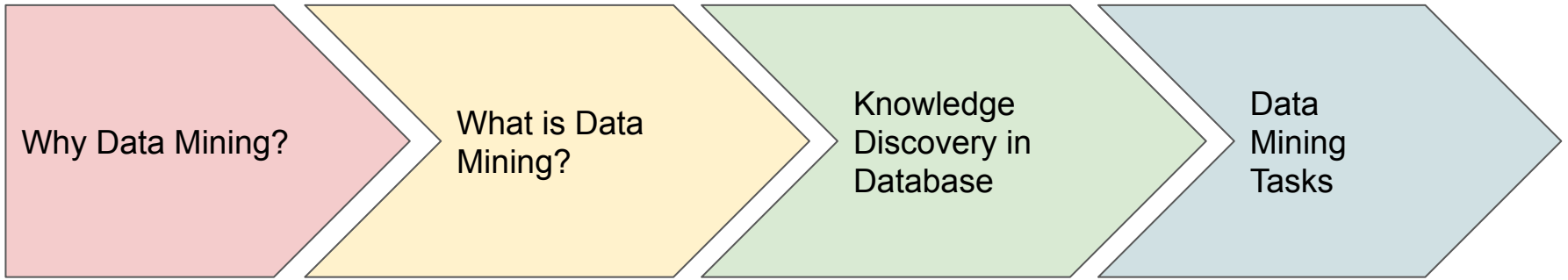
Lesson 2

Core concepts of Data Mining

Agenda


Why Data Mining?
What is Data Mining?
Knowledge Discovery in Database
Data Mining Tasks

Why the need for data mining?

Data is everywhere and it is expanding exponentially

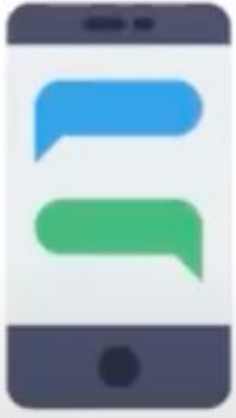
Data is being generated from multiple sources and in multiple formats

“Over 2.5 quintillion bytes of data are created every single day, and it's only going to grow from there. By 2020, it is estimated that 1.7MB of data will be created every second for every person on earth”

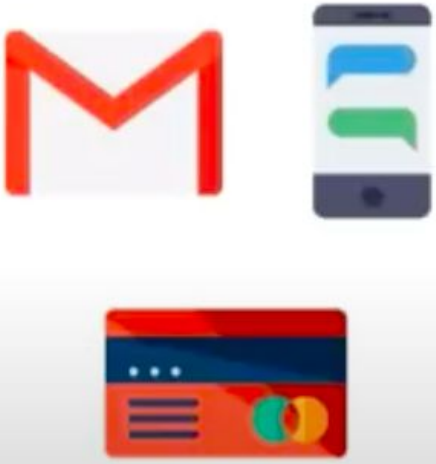

Structured data
Semi-structured data
Unstructured data
Databases
XML / JSON data
Email
Web pages
Audio
Video
Image data
Natural language
Documents

I have this financial data with me, I need to find out if any of the transactions are fraudulent.

I have this email data with me, I need to check how many of the mails are spam.

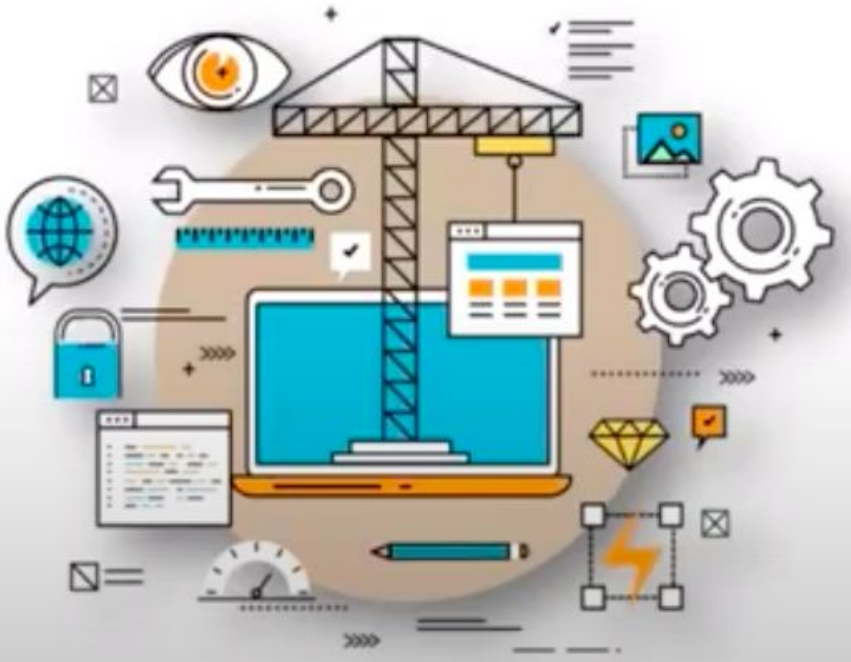


I have this telecom data with me, I need to find out how many of the customers will churn out.


How do I obtain Knowledge from this data?
Hey, you can use data mining techniques to find interesting insights from the data.

What is Data Mining?

Data mining is the computing process of discovering patterns in large datasets involving methods at the intersection of *machine learning*, *statistics*, and *database systems*.



The extracted information should give **new** patterns , relationships among the data entities

New

Correct

Potentially useful

As everything that glitters is not gold, similarly, all the mined information might not be correct/valid. The mined information needs to be evaluated for its correctness before it will used for any other purpose

New

Correct

Potentially useful

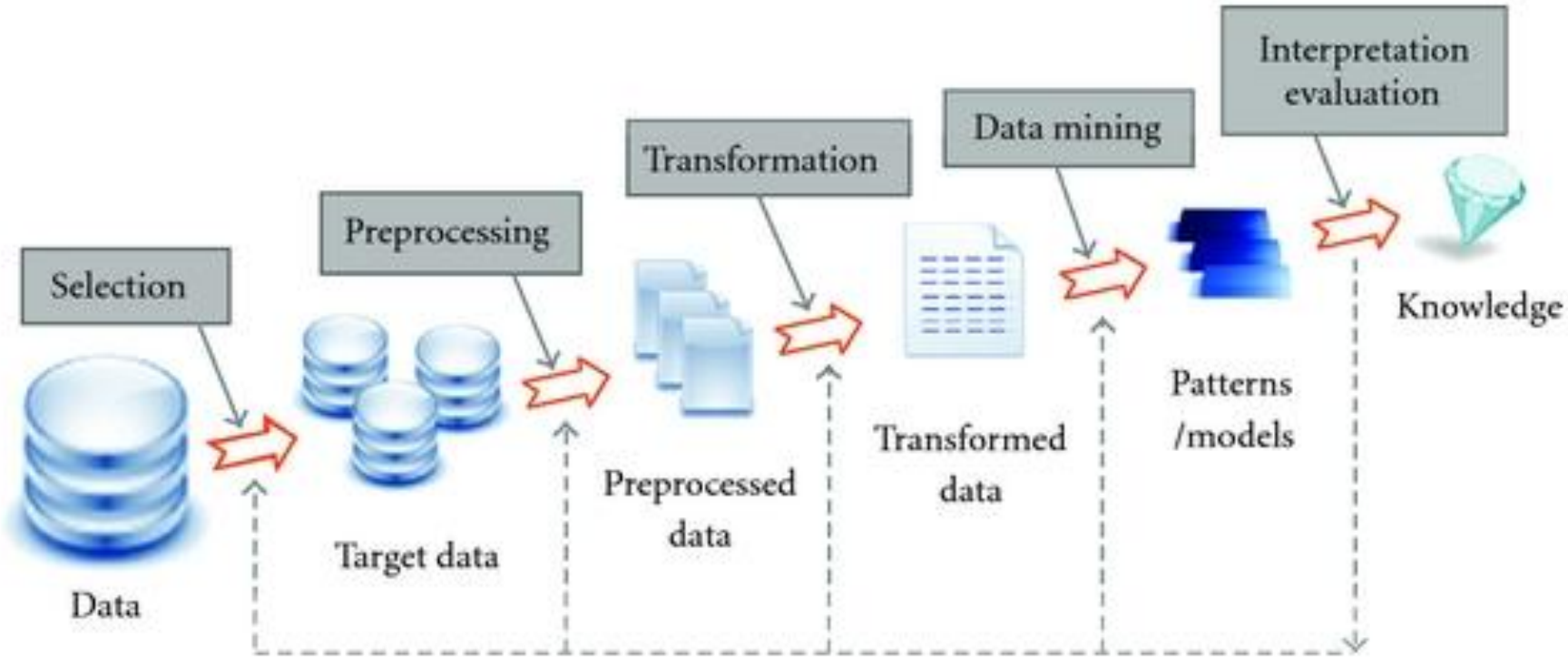
As we extract useful products such as petrol, diesel, etc from crude oil, similarly, the mined information from raw data should be useful and relevant

New

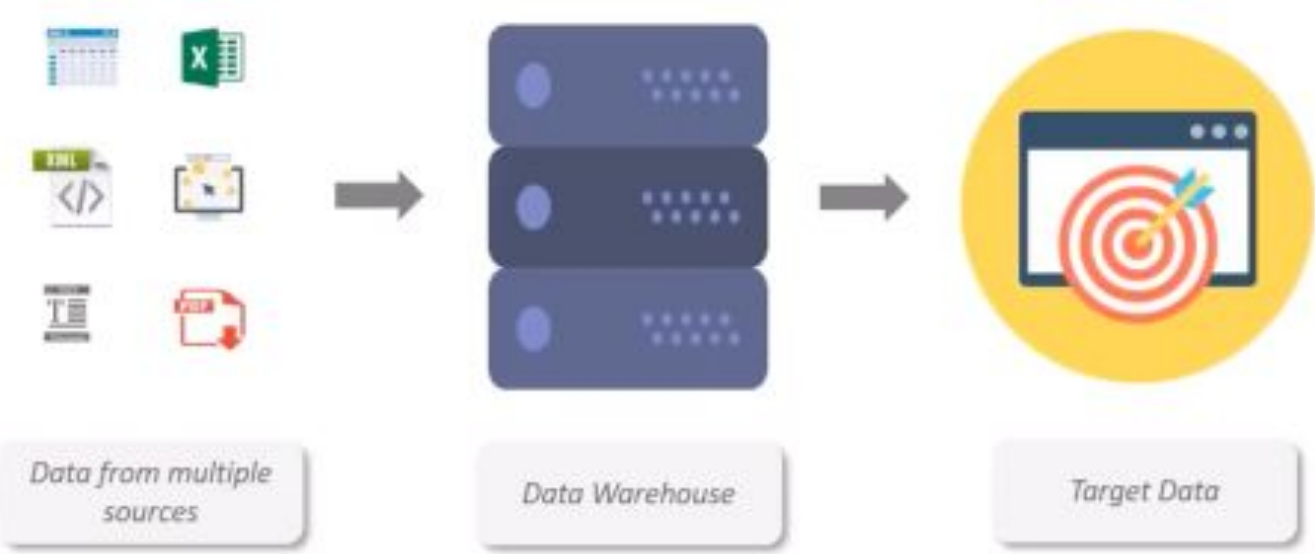
Correct

Potentially useful

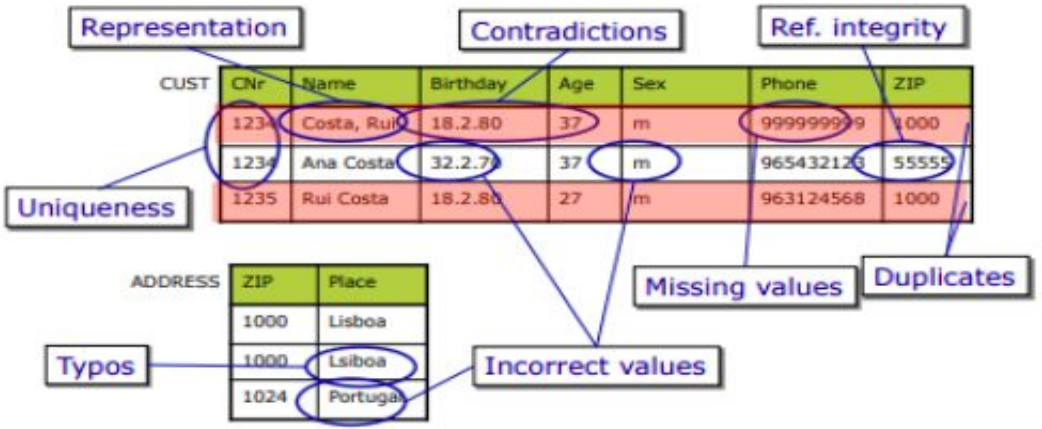
Knowledge Discovery in Databases (KDD)



- Selection
- Pre-processing
- Transformation
- Data Mining
- Evaluation



- Selection
- Pre-processing
- Transformation
- Data Mining
- Evaluation



- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

- Selection
- Pre-processing
- Transformation
- Data Mining
- Evaluation

-2,32,100,59,48 → -0.02,0.32,1.00,0.59,0.48

- Selection
- Pre-processing
- Transformation
- **Data Mining**
- Evaluation

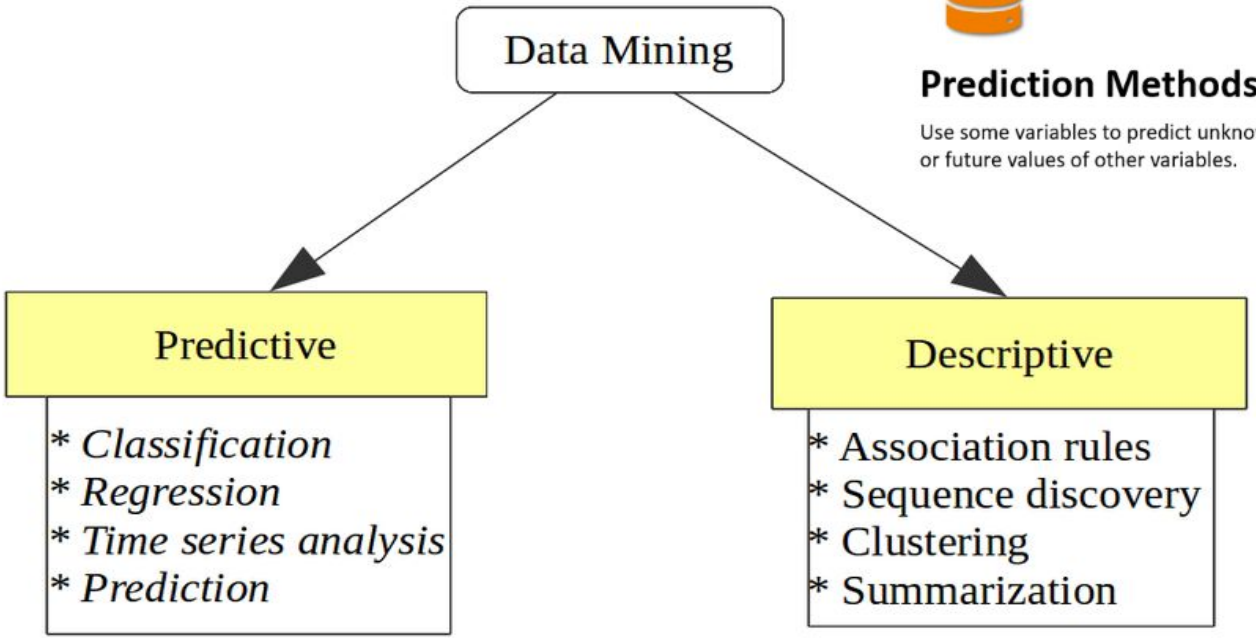
This is the most important step in the KDD process

Intelligent operation such as clustering, classification, regression are applied in order to extract patterns

- Selection
- Pre-processing
- Transformation
- Data Mining
- Evaluation

Once the data mining techniques have been applied, the obtained results need to be evaluated for accuracy

Data Mining Task

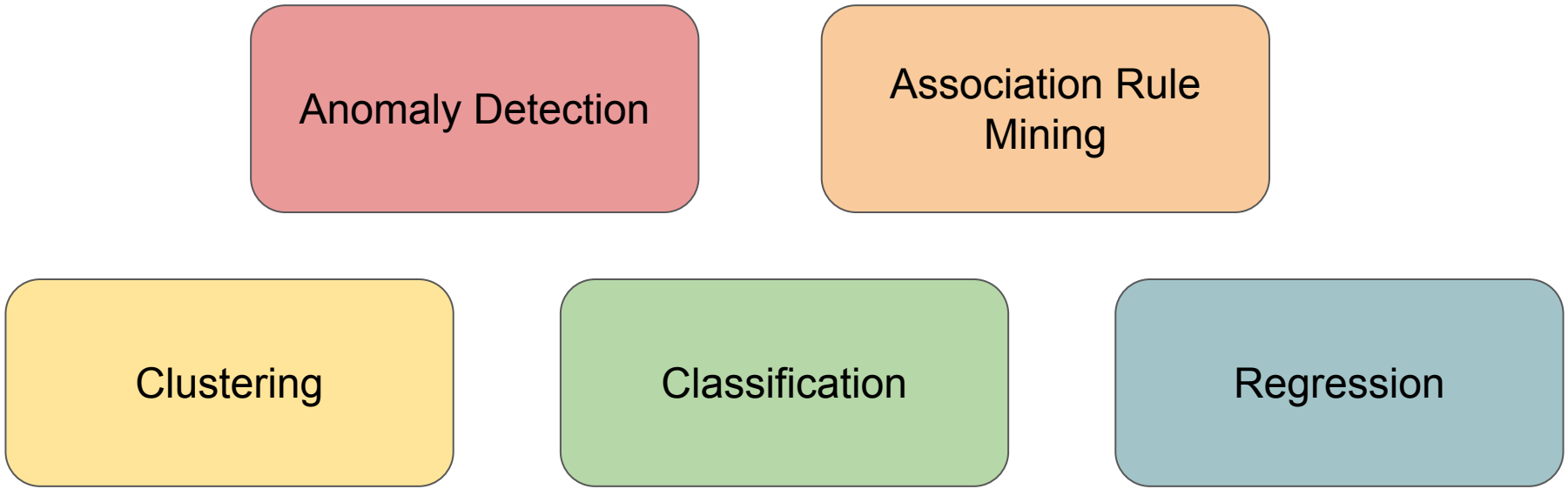


Prediction Methods

Use some variables to predict unknown or future values of other variables.

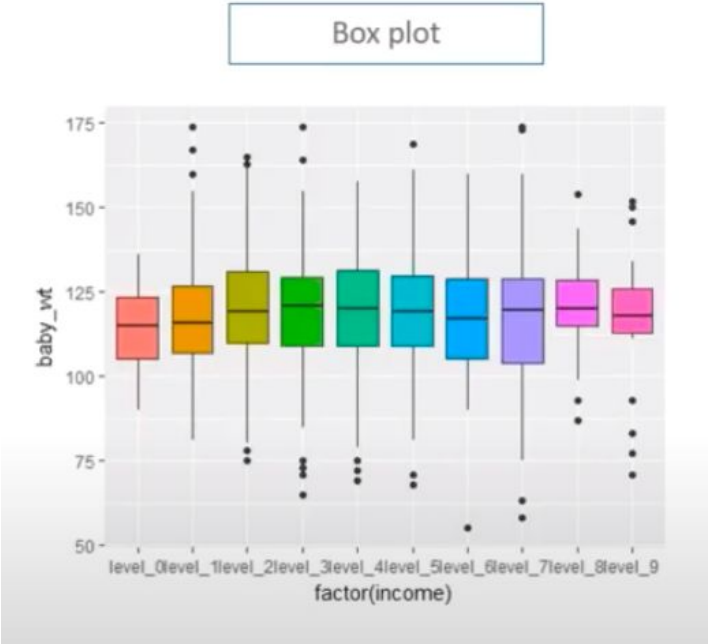
Description Methods

Find human-interpretable patterns that describe the data.


Anomaly Detection
Association Rule Mining
Clustering
Classification
Regression

- Anomaly Detection
- Association Rule Mining
- Clustering
- Classification
- Regression

Identification of unusual patterns or outliers, which help us in understanding the variation in data



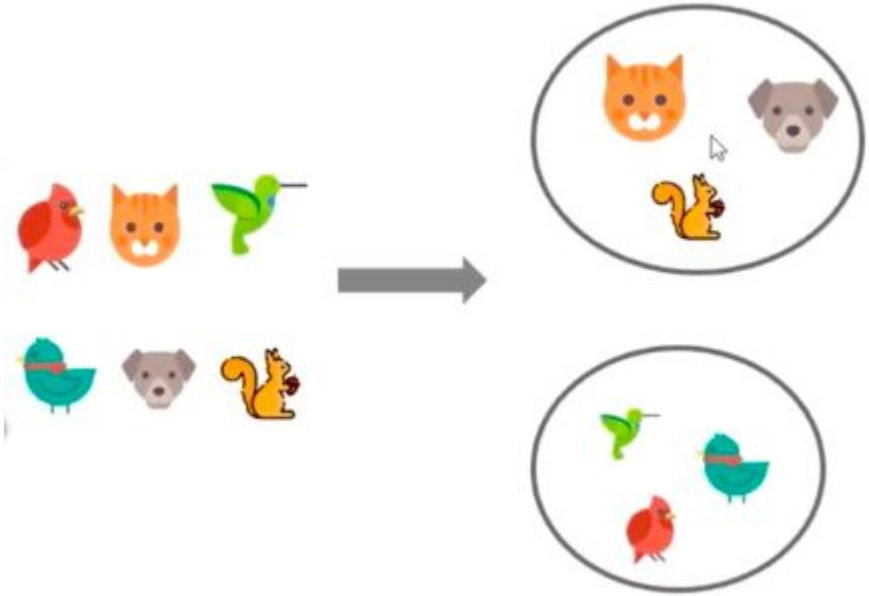
- Anomaly Detection
- Association Rule Mining
- Clustering
- Classification
- Regression

Also referred to as market basket analysis. This method is used for discovering interesting association patterns among variables

- Anomaly Detection
- Association Rule Mining
- Clustering
- Classification
- Regression

Identifying groups/classes in data which are similar to each other

The similarity inside the cluster is high, and between the clusters is low



- Anomaly Detection
- Association Rule Mining
- Clustering
- Classification
- Regression

Classification is the process of identifying, to which category does an observation belong.

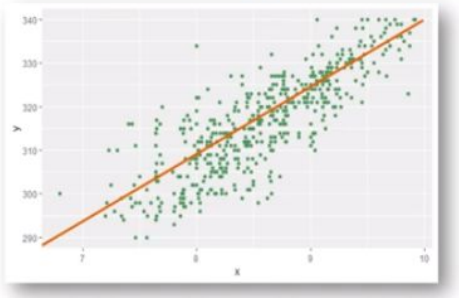
- Anomaly Detection
- Association Rule Mining
- Clustering
- Classification
- **Regression**

With regression, we can identify the extent of relationship among variables

Understanding how the “dependent” variable varies with respect to the variation in “independent” variable

Dependent Variable ← → Independent Variable

$Y=f(x)$



Linear Regression

Database

- Find all credit applicants with last name of Montenegro
- Identify customers who have purchased more than Php 10,000.00 in the last month
- Find all customers who have purchased milk

Data mining

- Find all credit applicants who are poor credit risk (classification)
- Identify customers with similar buying habits (clustering)
- Find all items which are frequently purchased with milk (association rules)

Questions

1) Differentiate Data Mining and Big Data.
2) Explain the concept of data mining and its purpose in extracting valuable patterns, insights, and knowledge from large datasets.
3) Discuss the ways in which data has grown in size in recent years.