Data Mining and Business Intelligence

ITA5007

PROF. E.P.EPHZIBAH

What is Data Mining?

1. Extracting useful information from large data sets. (Hand et al., 2001)
2. Another definition from the Gartner Group, the information technology research firm: Data Mining is the process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques.

What is Data Mining?

The term data mining indicates the process of exploration and analysis of a dataset, usually of large size, in order to find regular patterns, to extract relevant knowledge and to obtain meaningful recurring rules.

The data mining process is based on inductive learning methods, whose main purpose is to derive general rules starting from a set of available examples, consisting of past observations recorded in one or more databases.

Data mining is often defined as finding hidden information in a database. Alternatively, it has been called exploratory data analysis, data driven discovery, and deductive learning.

Origin and rapid growth of Data mining

1960s: Data collection, database creation, Information Management System(IMS) and network DBMS

1970s: Relational data model, relational DBMS implementation

1980s: RDBMS, advanced data models (extended-relational, OO, deductive, etc.) , Application-oriented DBMS (spatial, scientific, engineering, etc.)

1990s: Data mining, data warehousing, multimedia databases, and Web databases

2000s: Stream data management and mining, Data mining and its applications, Web technology (XML, data integration) and global information systems

Origin and rapid growth of Data mining

Before 1600, **empirical science (based on observation and experience)**

1600-1950s, **theoretical science**

- Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.

1950s-1990s, **computational science**

- Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
- Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.

1990-now, **data science**

- The flood of data from new scientific instruments and simulations
- The ability to economically store and manage petabytes of data online
- The Internet and computing Grid that makes all these archives universally accessible
- Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!

Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Data mining models and tasks

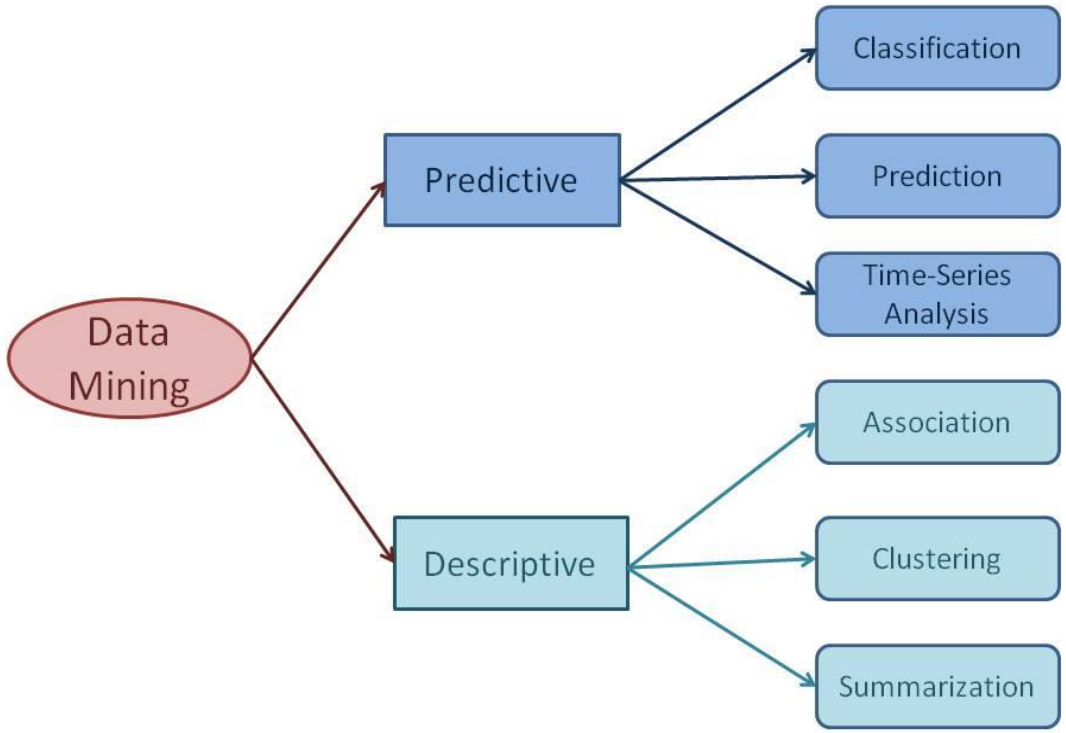
Data mining involves many different algorithms to accomplish different tasks. All of these algorithms attempt to fit a model to the data.

The algorithms examine the data and determine a model that is closest to the characteristics of the data being examined.

Depending on the data there are two broad categories of data mining tasks. They are

1. Predictive and

2. Descriptive



Data mining tasks & models

A predictive model makes a prediction about values of data using known results found from different data. Predictive modeling may be made based on the use of other historical data.

For example, a credit card use might be refused not because of the user's own credit history, but because the current purchase is similar to earlier purchases that were subsequently found to be made with stolen cards. Predictive model data mining tasks include classification, regression, time series analysis, and prediction.

A descriptive model identifies patterns or relationships in data. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties. Clustering, summarization, association rules, and sequence discovery are usually viewed as descriptive in nature.

Core Ideas in Data Mining

1. Classification
2. Prediction
3. Association rules
4. Predictive analytics
5. Data reduction
6. Data exploration
7. Data visualization

Core Ideas in Data Mining- classification

Classification maps data into predefined groups or classes.

It is often referred to as supervised learning because the classes are determined before examining the data.

Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to the classes.

Pattern recognition is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes.

Core Ideas in Data Mining -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).

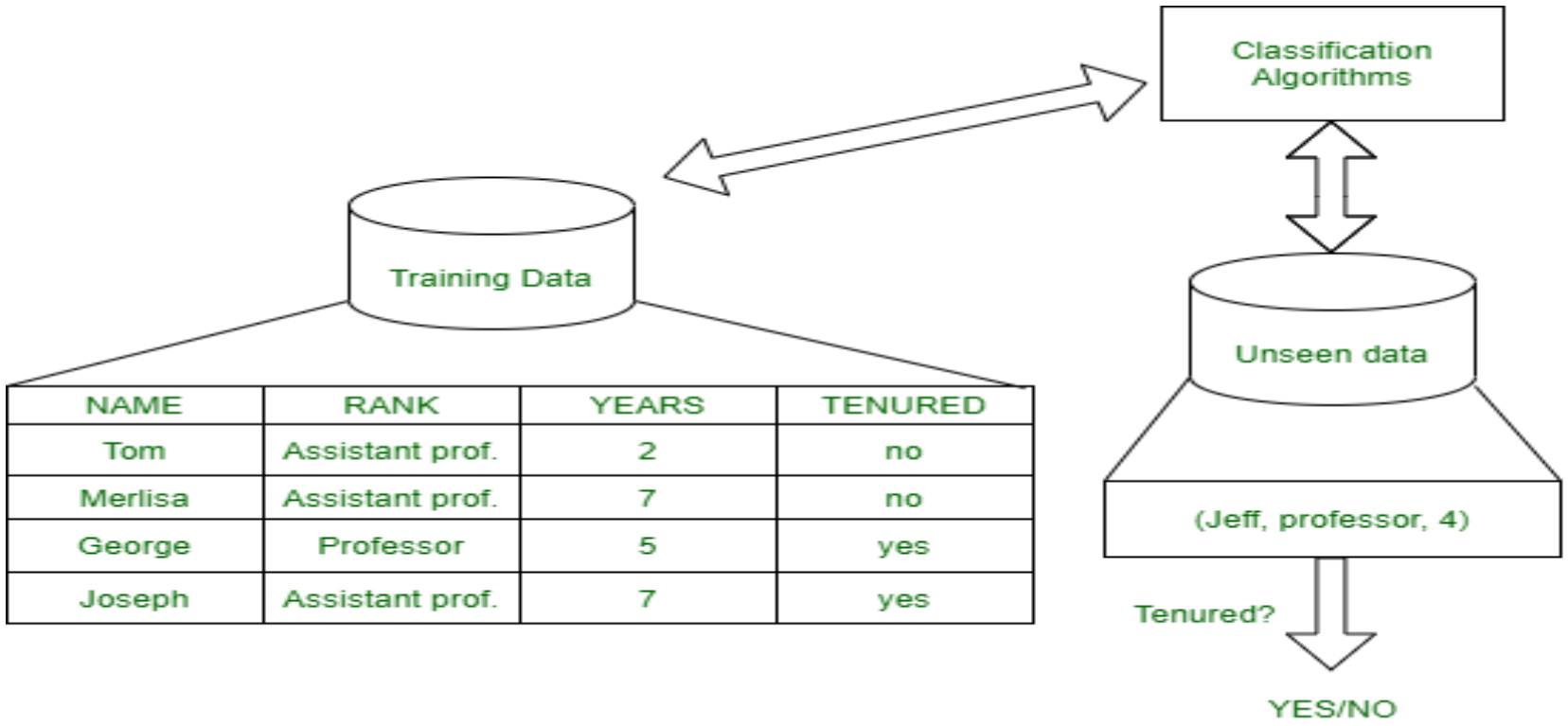
The model is used to predict the class label of objects for which the class label is unknown.

How is the derived model presented?

The derived model may be represented in various forms, such as classification rules (i.e., IF-THEN rules).

Core Ideas in Data Mining -classification

Example:



NAME	RANK	YEARS	TENURED
Tom	Assistant prof.	2	no
Merlisa	Assistant prof.	7	no
George	Professor	5	yes
Joseph	Assistant prof.	7	yes

Core Ideas in Data Mining -classification

Patient	Attributes			
	Headache	Vomiting	Temperature	Viral illness
#1	No	Yes	High	Yes
#2	Yes	No	High	Yes
#3	Yes	Yes	Very high	Yes
#4	No	Yes	Normal	No
#5	Yes	No	High	No
#6	No	Yes	Very high	Yes

If (Headache=No AND Vomiting = Yes AND Temperature = High)
THEN Viral illness = Yes

Core Ideas in Data Mining -prediction

Prediction is similar to classification, except that we are trying to predict the value of a numerical variable (e.g., amount of purchase) rather than a class (e.g., purchaser or nonpurchaser).

The term *prediction* refers to the prediction of the value of a continuous variable. (Eg: Rainfall level prediction, House Price Prediction, Salary prediction)

(Sometimes in the data mining literature, the term *estimation* is used to refer to the prediction of the value of a continuous variable, and *prediction* may be used for both continuous and categorical data.)

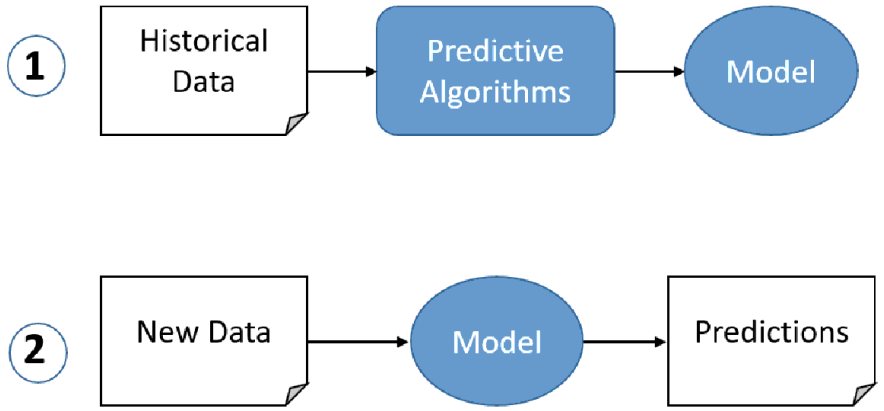
Core Ideas in Data Mining -prediction

Classification and prediction techniques are the core of predictive analytics in business intelligence, for which there are many applications in analyzing markets, supplies, and sales.

Many real-world data mining applications can be seen as predicting future data states

based on past and current data. Prediction can be viewed as a type of classification.

Example:



Association Rules

Large databases of customer transactions lend themselves naturally to the analysis of associations among items purchased.

Association rules, or affinity analysis, can then be used in a variety of ways.

- grocery stores can use such information after a customer's purchases, have all been scanned to print discount coupons.
- Online merchants such as Amazon.com and Netflix.com use these methods as the heart of a "recommender" system that suggests new purchases to customers.

Association rules: example

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Predictive Analytics

Classification, prediction, and to some extent, affinity analysis constitute the analytical methods employed in predictive analytics.

A significant proportion of the models used in business intelligence systems, such as optimization models, require input data concerned with future events.

The predictive analysis includes different statistical techniques such as data mining, predictive modeling, and machine learning.

All these techniques interpret historical and present situations so that future analyses can be made about the data.

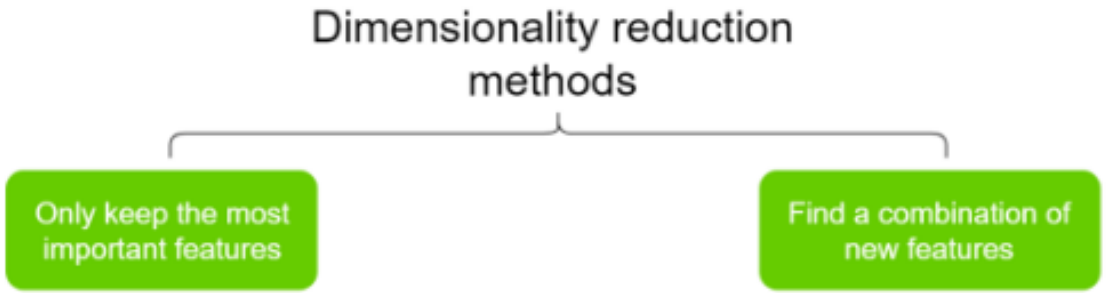
This creates predictions about the data for future or other unknown events. Using historical data, mathematical trends can be made to understand data. Some examples include health, retail, weather, insurance, and risk assessment. This analysis makes useful information from the data and helps the business grow. It transforms the raw data to provide more information and insights.

Data reduction

Sensible data analysis often requires distillation of complex data into simpler data.

Rather than dealing with thousands of product types, an analyst might wish to group them into a smaller number of groups.

This process of consolidating a large number of variables (or cases) into a smaller set is termed data reduction.



Data Exploration

Unless our data project is very narrowly focused on answering a specific question determined in advance (in which case it has drifted more into the realm of statistical analysis than of data mining), an essential part of the job is to review and examine the data to see what messages they hold full understanding of the data may require a reduction in its scale or dimension to allow us to see the forest without getting lost in the trees

Data exploration is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more.

Data exploration is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems.

Data Visualization

Another technique for exploring data to see what information they hold is through graphical analysis.

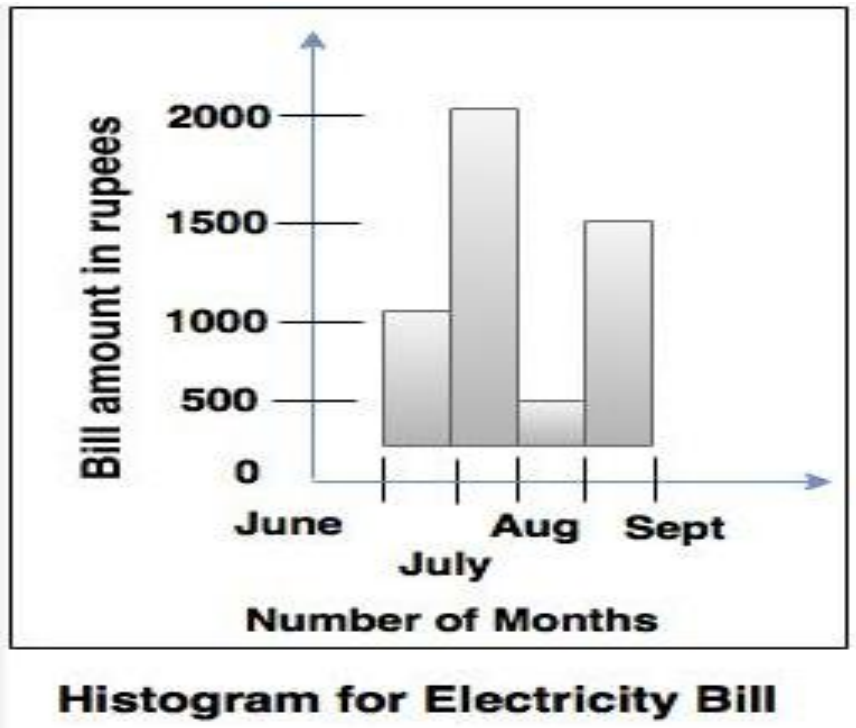
This includes looking at each variable separately as well as looking at relationships between variables.

For numerical variables, we use histograms and boxplots to learn about the distribution of their values, to detect outliers (extreme observations), and to find other information that is relevant to the analysis task.

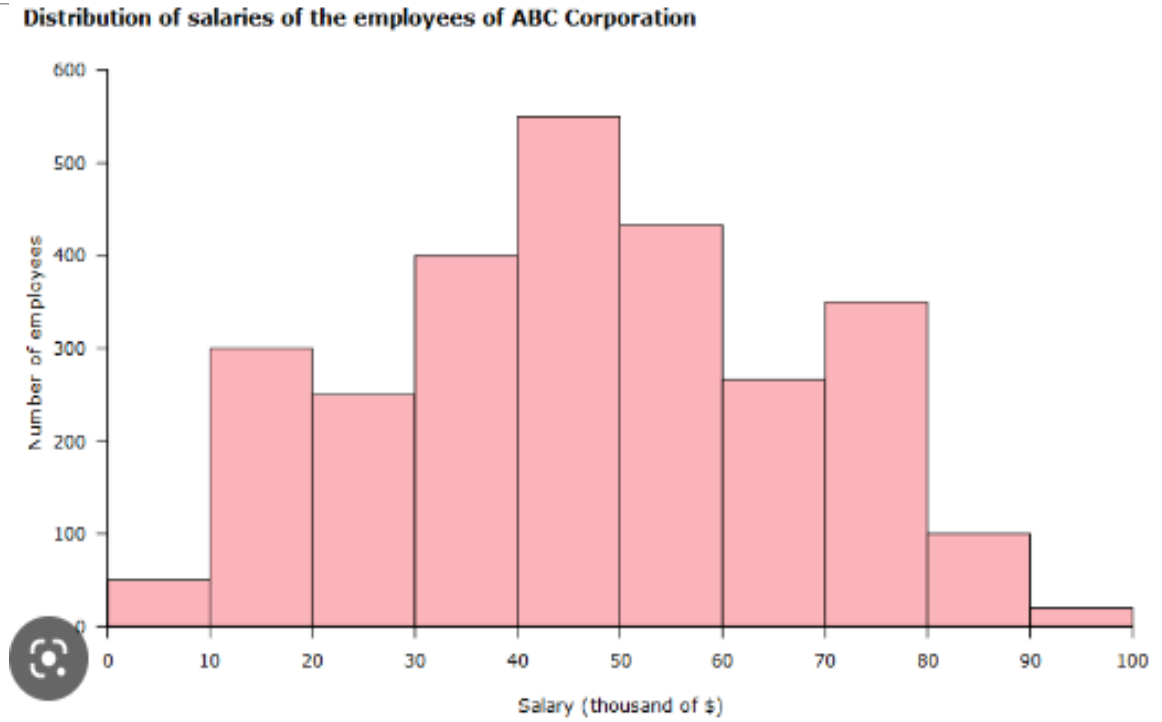
For categorical variables, we use bar charts. We can also look at scatterplots of pairs of numerical variables to learn about possible relationships, and the type of relationship, and again, to detect outliers.

Visualization can be greatly enhanced by adding features such as color, zooming, and interactive navigation.

Histogram



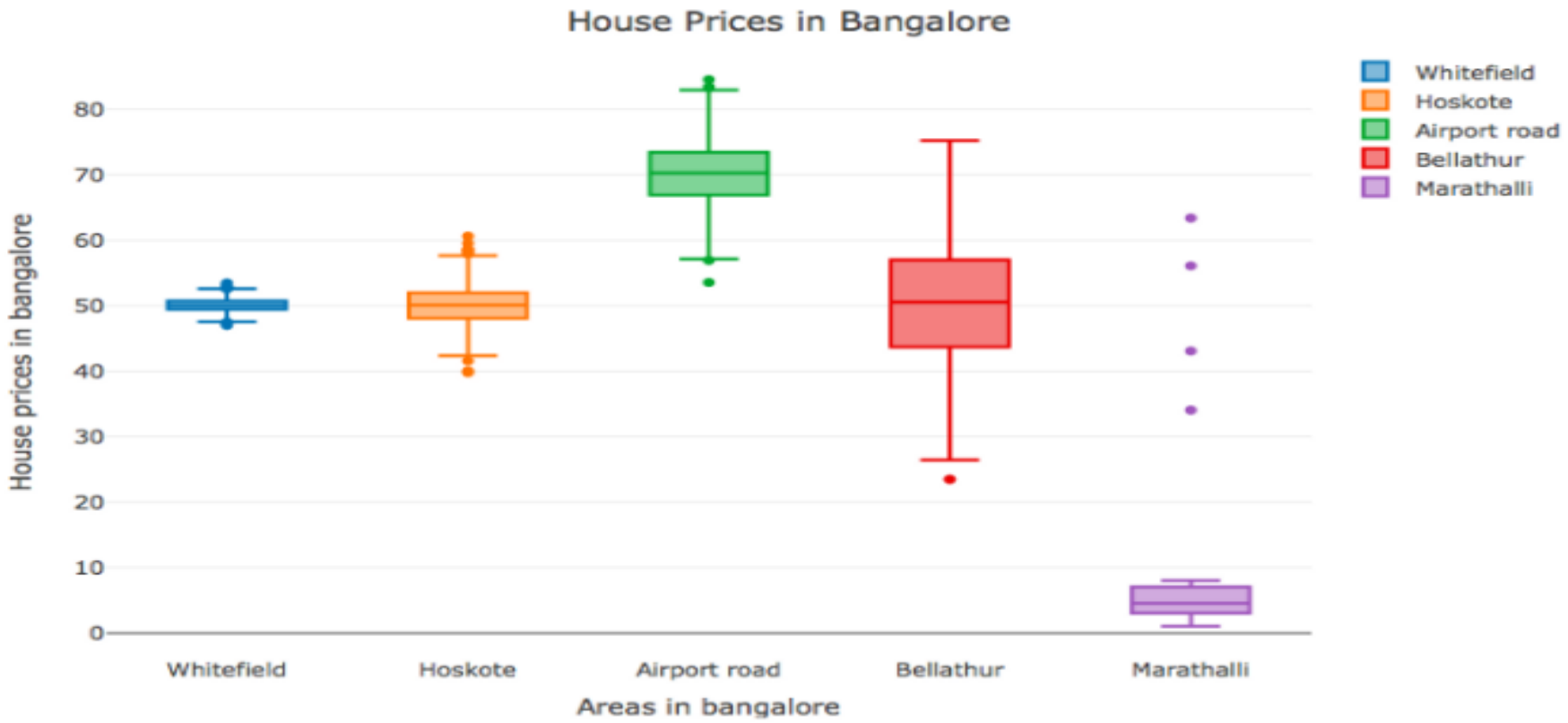
Histogram for Electricity Bill



https://www.tutorialride.com/data-mining/knowledge-representation-in-data-mining.htm

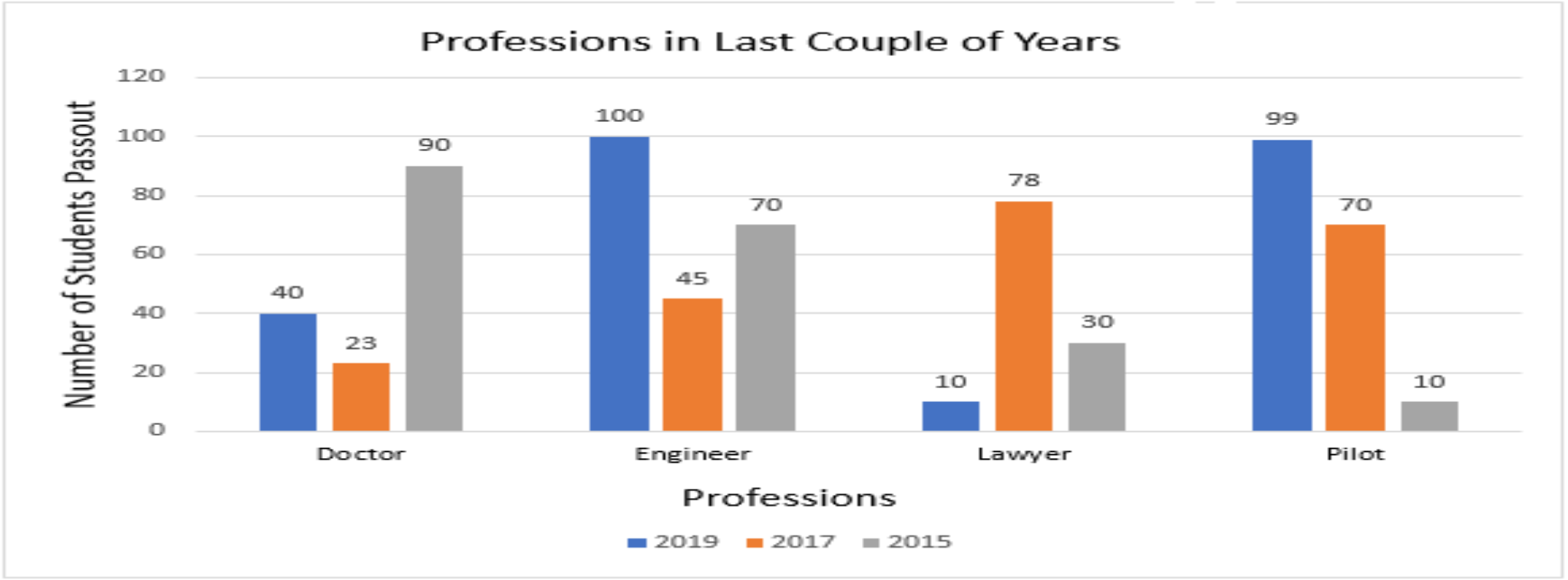
https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/histo/5214822-eng.htm

BOX Plot



https://dimensionless.in/what-is-a-box-plot/

Bar Chart



https://www.javatpoint.com/bar-graph

Question:1

Credit card companies must determine whether to authorize credit card purchases. Suppose that based on past historical information about purchases, each purchase is placed into one of four classes: (1) authorize, (2) ask for further identification before authorization, (3) do not authorize, and (4) do not authorize but contact police. The data mining functions here are twofold. First the historical data must be examined to determine how the data fit into the four classes. Then the problem is to apply this model to each new purchase.

Give your suggestions about the scenario.

Question:2

An airport security screening station is used to determine: if passengers are potential terrorists or criminals. To do this, the face of each passenger is scanned and its basic pattern (distance between eyes, size and shape of mouth, shape of head, etc.) is identified. This pattern is compared to entries in a database to see if it matches any patterns that are associated with known offenders.

Which data mining task would be more appropriate for this scenario?

Question:3

Predicting flooding is a difficult problem. One approach uses monitors placed at various points in the river. These monitors collect data relevant to flood prediction: water level, rain amount, time, humidity, and so on. Then the water level at a potential flooding point in the river can be predicted based on the data collected by the sensors upriver from this point. The prediction must be made with respect to the time the data were collected.

Give the data mining approach you would suggest as a solution to this problem.

Question: 4

Data mining methodologies can be applied to a variety of domains, from marketing and manufacturing process control to the study of risk factors in medical diagnosis, from the evaluation of the effectiveness of new drugs to fraud detection. With suitable examples, write about the applications of data mining.