Artificial Intelligence for Engineering

Unit 2nd

Unit 2 Data & Algorithms

Syllabus : History Of Data , Data Storage And Importance of Data and its Acquisition , The Stages of data processing , Data Visualization , Regression, Prediction & Classification Clustering & Recommender Systems

1. What is Data

In general, **data** is any set of characters that is gathered and translated for some purpose, usually analysis. If data is not put into context, it doesn't do anything to a human or computer

Data is a collection of unorganized facts & figures and does not provide any further information regarding patterns, context, etc. Hence data means "unstructured facts and figures".

Information is a structured data i.e. organized meaningful and processed data. To process the data and convert into information, a computer is used.

Data can be defined as a representation of facts, concepts, or instructions in a formalized manner, which should be suitable for communication, interpretation, or processing by human or electronic machine.

Data is represented with the help of characters such as alphabets (A-Z, a-z), digits (0-9) or special characters (+,-,/,*,<,>,= etc.)

Data refers to distinct pieces of information, usually formatted and stored in a way that is concordant with a specific purpose. Data can exist in various forms: as numbers or text recorded on paper, as bits or bytes stored in electronic memory, or as facts living in a person's mind.

Data – a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process

Whichever industry you work in, or whatever your interests, you will almost certainly have come across a story about how “data” is changing the face of our world. It might be part of a study helping to cure a disease, boost a company’s revenue, make a building more efficient or be responsible for those targeted ads you keep seeing.

In general, data is simply another word for information. But in computing and business (most of what you read about in the news when it comes to data – especially if it's about Big Data), data refers to information that is machine-readable as opposed to human-readable.

Machine-readable vs. human-readable data

Human-readable (also known as unstructured data) refers to information that only humans can interpret and study, such as an image or the meaning of a block of text. If it requires a person to interpret it, that information is human-readable.

Machine-readable (or structured data) refers to information that computer programs can process. A program is a set of instructions for manipulating data. And when we take data and apply a set of programs, we get software. In order for a program to perform instructions on data, that data must have some kind of uniform structure.

For example, US Naval Officer Matthew Maury, turned years of old hand-written shipping logs (human-readable) into a large collection of coordinate routes (machine-readable). He was then able to process these routes en masse to reduce the average Naval journey by 33%.

All data can be categorized as machine-readable, human-readable, or both. Human-readable data utilizes natural language formats (such as a text file containing ASCII codes or PDF document), whereas machine-readable data uses formally structured computer languages (Parquet, Avro, etc.) to be read by computer systems or software. Some data is readable by both machines and humans, as in the case of CSV, HTML, or JSON. The line between machine- and human-readable data is becoming increasingly blurred because so many formats that are prevalent today are accessible enough to be navigated by a human yet structured enough to be processed by a machine. This is largely the result of artificial intelligence, machine learning, and automation, which streamlines tasks and workflows so manual data entry and analysis is done by a machine rather than a human.

3. Data acquisition

Data acquisition is the process of measuring physical world conditions and phenomena such as electricity, sound, temperature and pressure. This is done through the use of various sensors which sample the environment's analog signals and transform them to digital signals using an analog-to-digital converter. The resulting digital numeric values can then be directly manipulated by a computer, allowing for the analysis, storage and presentation of these data

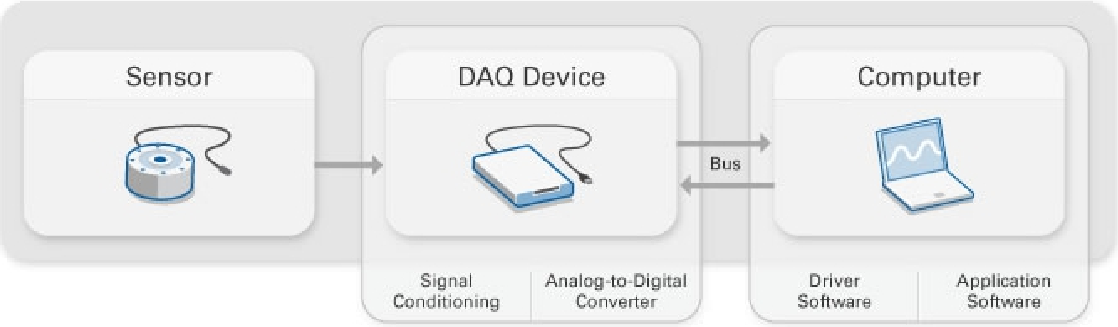
Data acquisition is primarily done using a combination of instruments and tools that form a data acquisition system (DAQ or DAS). The DAS samples the environmental signals and transforms these to machine-readable signals, while the software processes the acquired data for storage or presentation.

There are three components required for data acquisition:

- Sensors that are able to capture environmental analog signals like temperature, pressure, light or sound
- Signal-conditioning circuitry that normalizes captured signals; noise reducers and amplifiers are good examples
- Analog-to-digital converter that converts the conditioned signals into digital data

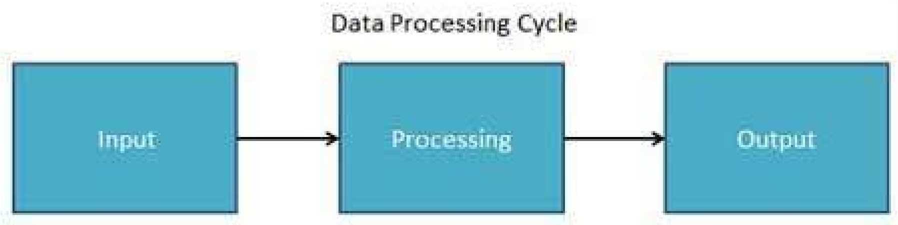
Specific DAQs are often created for specific physical properties. For example, there are dedicated systems for measuring just temperature or just pressure, but smaller dedicated data acquisition systems can be integrated into a bigger system via software by simply taking the data gathered by those individual systems and presenting them to the user.

Data acquisition applications are usually controlled by software programs developed using various general purpose programming languages such as Assembly, BASIC, C, C++, C#, Fortran, Java, LabVIEW, Lisp, Pascal, etc. Stand-alone data acquisition systems are often called data loggers.



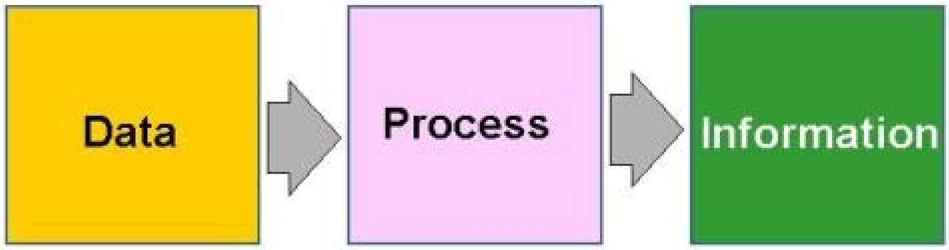
4. Data Processing Cycle

Data processing is the re-structuring or re-ordering of data by people or machine to increase their usefulness and add values for a particular purpose. Data processing consists of the following basic steps - input, processing, and output. These three steps constitute the data processing cycle.



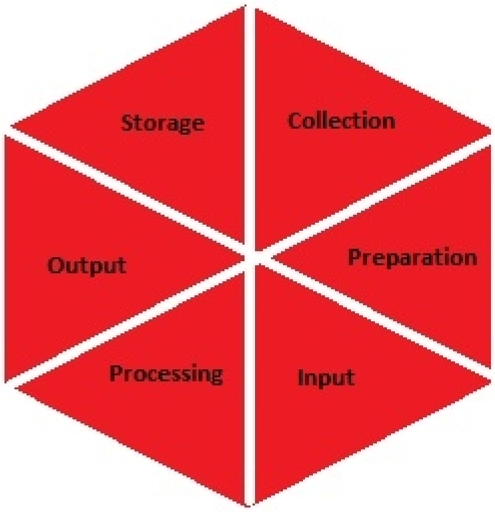
- **Input** − In this step, the input data is prepared in some convenient form for processing. The form will depend on the processing machine. For example, when electronic computers are used, the input data can be recorded on any one of the several types of input medium, such as magnetic disks, tapes, and so on.

- **Processing** − In this step, the input data is changed to produce data in a more useful form. For example, pay-checks can be calculated from the time cards, or a summary of sales for the month can be calculated from the sales orders.
- **Output** − At this stage, the result of the proceeding processing step is collected. The particular form of the output data depends on the use of the data. For example, output data may be pay-checks for employees.



Stages of Data Processing

Data processing consists of following 6 stages −



Collection

Collection of data refers to gathering of data. The data gathered should be defined and accurate.

Preparation

Preparation is a process of constructing a dataset of data from different sources for future use in processing step of cycle.

Input

Input refers to supply of data for processing. It can be fed into computer through any of input devices like keyboard, scanner, mouse, etc.

Processing

The process refers to concept of an actual execution of instructions. In this stage, raw facts or data is converted to meaningful information.

Output and Interpretation

In this process, output will be displayed to user in form of text, audio, video, etc. Interpretation of output provides meaningful information to user.

Storage

In this process, we can store data, instruction and information in permanent memory for future reference.

Collection, manipulation, and processing collected data for the required use is known as data processing. It is a technique normally performed by a computer; the process includes retrieving, transforming, or classification of information.

However, the processing of data largely depends on the following −

- The volume of data that need to be processed
- The complexity of data processing operations
- Capacity and inbuilt technology of respective computer system
- Technical skills
- Time constraints

5. Data Visualization

- Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.
- Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.
- Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.
- In the world of Big Data, the data visualization tools and technologies are required to analyze vast amounts of information.
- Data visualizations are common in your everyday life, but they always appear in the form of graphs and charts. The combination of multiple visualizations and bits of information are still referred to as Infographics.
- Data visualizations are used to discover unknown facts and trends. You can see visualizations in the form of line charts to display change over time. Bar and column charts are useful for observing relationships and making comparisons. A pie chart is a great way to show parts-of-a-whole. And maps are the best way to share geographical data visually.

Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts. It helps users in analyzing a large amount of data in a simpler way. It makes complex data more accessible, understandable, and usable.

Tables are used where users need to see the pattern of a specific parameter, while charts are used to show patterns or relationships in the data for one or more parameters.

Tips to follow while representing data visually –

- Number all diagrams
- Label all diagrams
- Ensure that units of measurement on axes are clearly labelled
- Place any explanatory information in footnotes below the visual
- Check layouts to ensure maximum clarity

Pro and Cons of Data Visualization

Here are some pros and cons to representing data visually –

Pros

- It can be accessed quickly by a wider audience.
- It conveys a lot of information in a small space.
- It makes your report more visually appealing.
- .

Cons

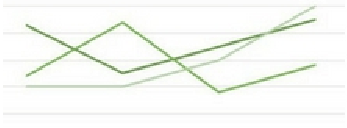
- It can misrepresent information – if an incorrect visual representation is made.
- It can be distracting – if the visual data is distorted or excessively used.

Data visualization techniques

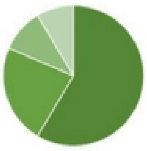
- **Charts**

The easiest way to show the development of one or several data sets is a chart.

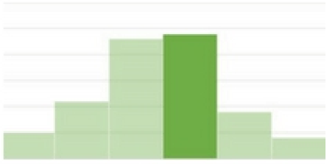
Charts vary from bar and line charts that show the relationship between elements over time to pie charts that demonstrate the components or proportions between the elements of one whole.



Line



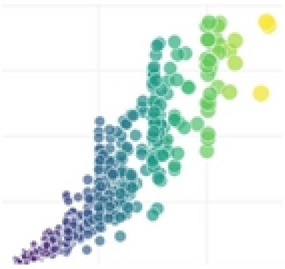
Pie



Bar

- **Plots**

Plots allow to distribute two or more data sets over a 2D or even 3D space to show the relationship between these sets and the parameters on the plot.



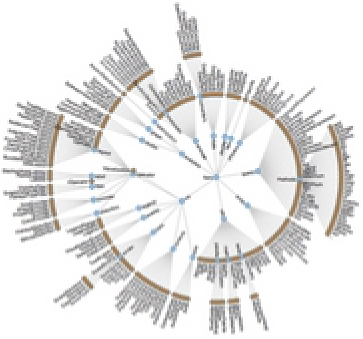
Bubble

Scatter

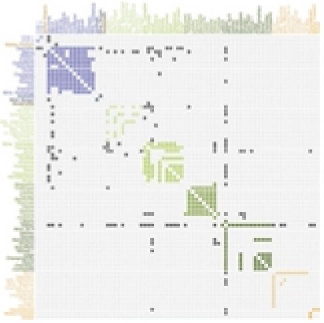
- **Maps**

Maps are popular ways to visualize data used in different industries. They allow to locate elements on relevant objects and areas — geographical maps, building plans, website layouts, etc.

- **Diagrams and matrices**
- Diagrams are usually used to demonstrate complex data relationships and links and include various types of data on one visualization. They can be hierarchical, multidimensional, tree-like.
- Matrix is one of the advanced data visualization techniques that help determine the correlation between multiple constantly updating (steaming) data sets**.**



Tree



Matrix

-

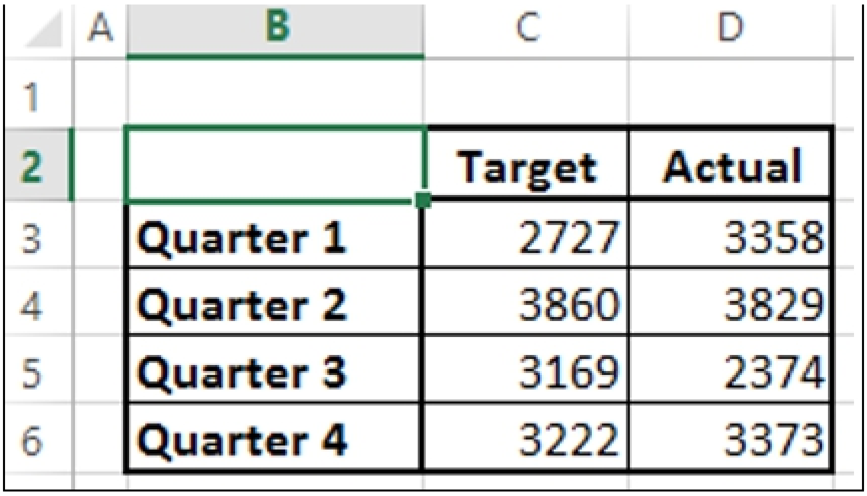
Data visualization tools

Data visualization allows you to interact with data. **Google**, **Apple**, **Facebook**, and **Twitter** all ask better a better question of their data and make a better business decision by using data visualization.

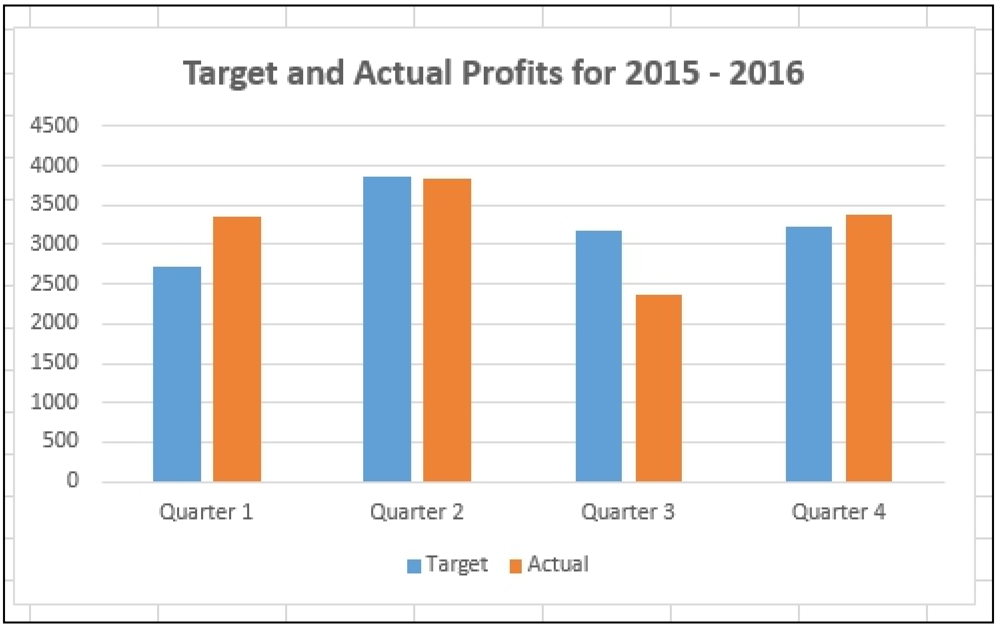
Here are the some data visualization tools that help you to visualize the data:

1. MS Excel

- You can display your data analysis reports in a number of ways in Excel. However, if your data analysis results can be visualized as charts that highlight the notable points in the data, your audience can quickly grasp what you want to project in the data. It also leaves a good impact on your presentation style
- In Excel, charts are used to make a graphical representation of any set of data. A chart is a visual representation of the data, in which the data is represented by symbols such as bars in a Bar Chart or lines in a Line Chart. Excel provides you with many chart types and you can choose one that suits your data or you can use the Excel Recommended Charts option to view charts customized to your data and select one of those.



	A	B	C	D
1				
2			Target	Actual
3		Quarter 1	2727	3358
4		Quarter 2	3860	3829
5		Quarter 3	3169	2374
6		Quarter 4	3222	3373



2. Tableau

Tableau is a data visualization tool. You can create graphs, charts, maps, and many other graphics.

3 Infogram

Infogram is also a data visualization tool. It has some simple steps to process that:

1. First, you choose among many templates, personalize them with additional visualizations like maps, charts, videos, and images.
2. Then you are ready to share your visualization

4 Plotly

Plotly will help you to create a slick and sharp chart in just a few minutes or in a very short time. It also starts from a simple spreadsheet.

5 Chartblocks

Chartblocks is an easy way to use online tool which required no coding and builds visualization from databases, spreadsheets, and live feeds.

Data visualization have some more specialties such as:

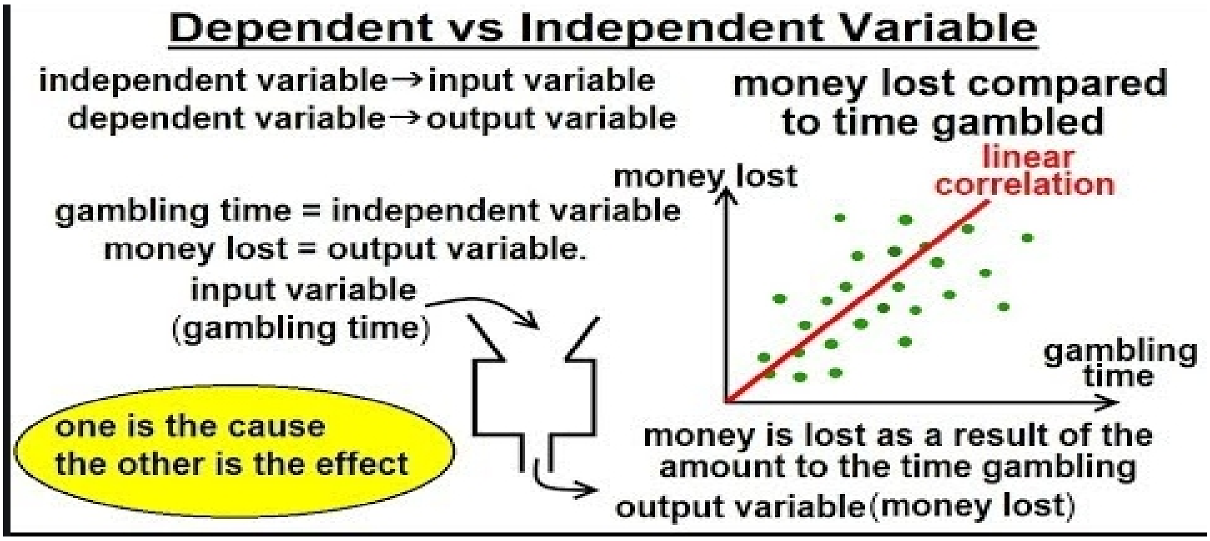
- Data visualization can identify areas that need improvement or modifications.
- Data visualization can clarify which factor influence customer behavior.
- Data visualization helps you to understand which products to place where.
- Data visualization can predict sales volumes.

6. Regression

Regression is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.

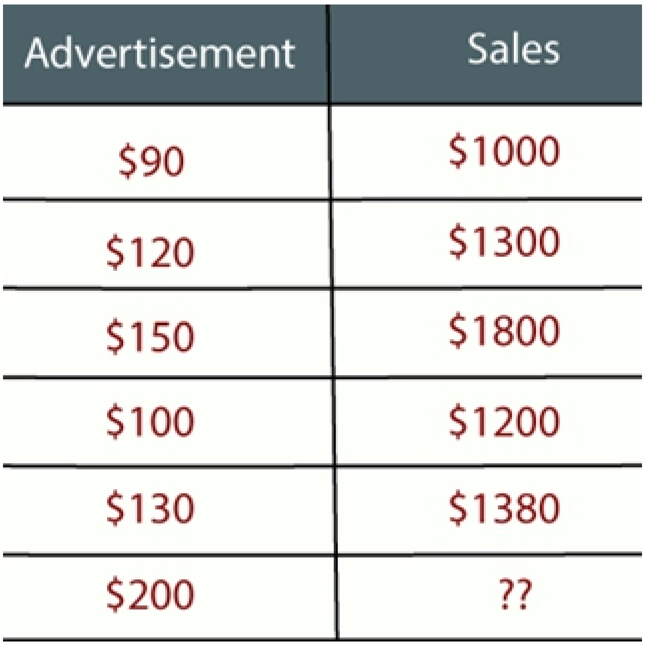
More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables)


Dependent vs Independent Variable
independent variable→input variable
dependent variable→output variable
money lost compared to time gambled
gambling time = independent variable
money lost = output variable.
input variable (gambling time)
money lost
linear correlation
gambling time
one is the cause
the other is the effect
money is lost as a result of the amount to the time gambling
output variable(money lost)

We can understand the concept of regression analysis using the below example:

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:



Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Now, the company wants to do the advertisement of $200 in the year 2019 **and wants to know the prediction about the sales for this year**. So to solve such type of prediction problems in machine learning, we need regression analysis.

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, ***"Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum."*** The distance between datapoints and line tells whether a model has captured a strong relationship or not.

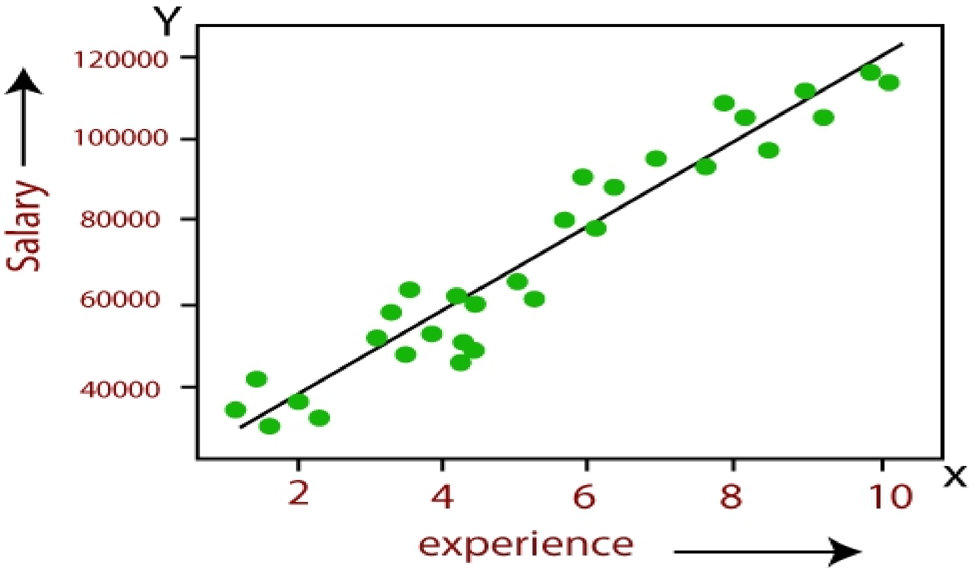
Some examples of regression can be as:

- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

Linear Regression:

- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.

- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



- Below is the mathematical equation for Linear regression:

1. Y= aX+b

Here, Y = dependent variables (target variables),
X= Independent variables (predictor variables),
a and b are the linear coefficients

Some popular applications of linear regression are:

- **Analyzing trends and sales estimates**
- **Salary forecasting**
- **Real estate prediction**
- **Arriving at ETAs in traffic.**

7. Classification & Prediction

There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follows −

- Classification
- Prediction

Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

Classification is the process of identifying the category or class label of the new observation to which it belongs. **Predication** is the process of identifying the missing or unavailable numerical data for a new observation. That is the key difference between classification and prediction. The predication does not concern about the class label like in classification.

- **1.Prediction** is like saying something which may going to be happened in future.Prediction may be a kind of classification
- 2.**Prediction** is mostly based on our future assumptions
- whereas
- 1.**Classification** is categorization of the things or data that we already have with us.This categorization can be based on any kind of technique or algorithms
- 2.**Classification** is mostly based on our current or past assumptions

What is classification?

Following are the examples of cases where the data analysis task is Classification −

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

What is prediction?

Following are the examples of cases where the data analysis task is Prediction −

Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

Note − Regression analysis is a statistical methodology that is most often used for numeric prediction.

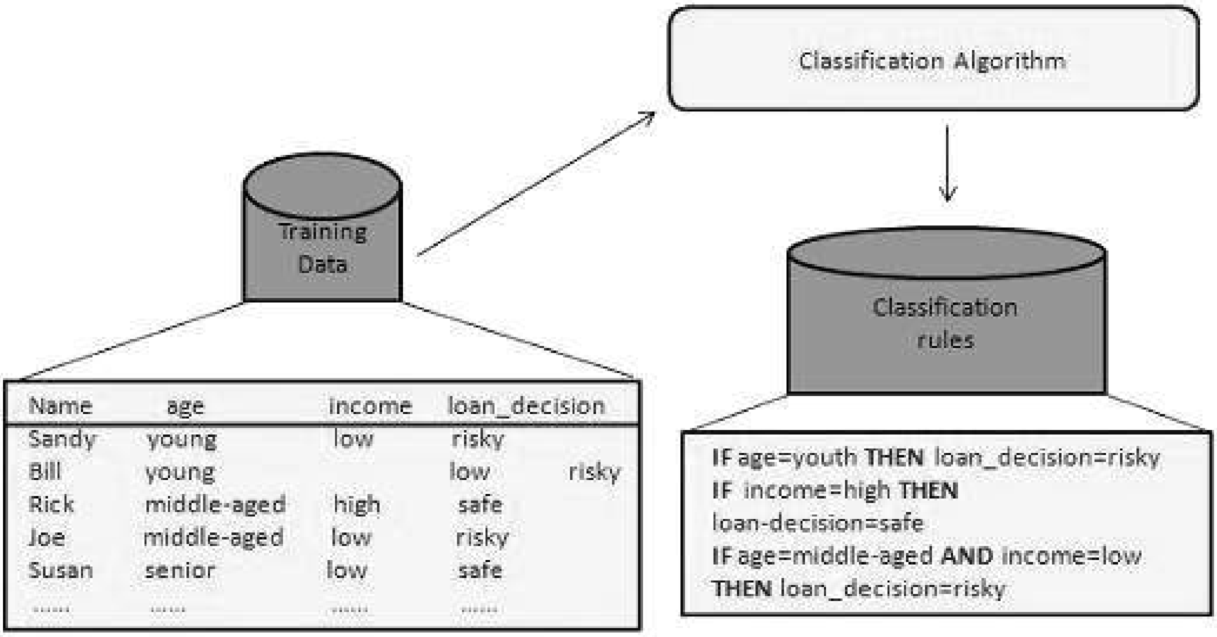
How Does Classification Works?

With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification process includes two steps −

- Building the Classifier or Model
- Using Classifier for Classification

Building the Classifier or Model

- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

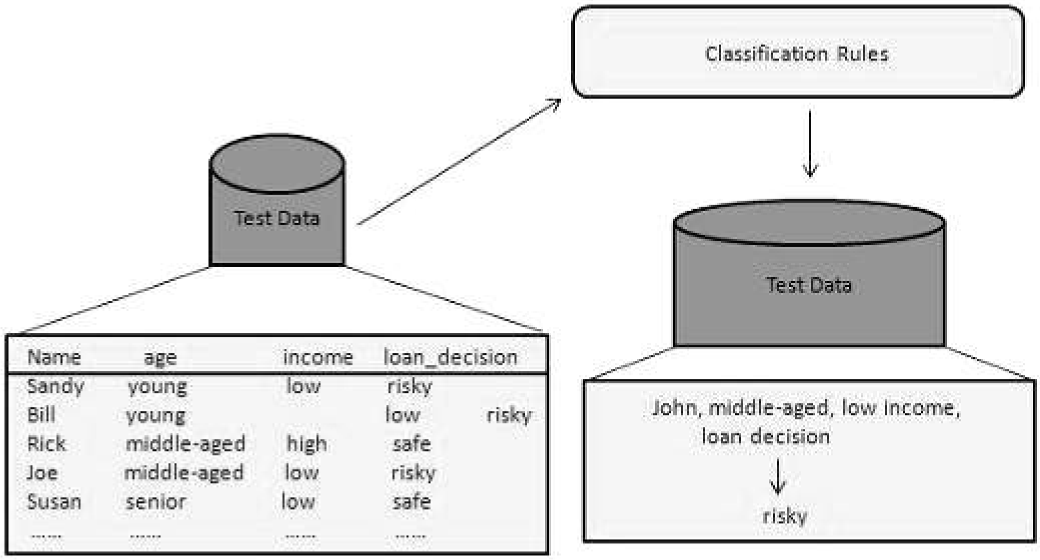


Name	age	income	loan_decision	
Sandy	young	low	risky	
Bill	young		low	risky
Rick	middle-aged	high	safe	
Joe	middle-aged	low	risky	
Susan	senior	low	safe	
......	......	......	......	

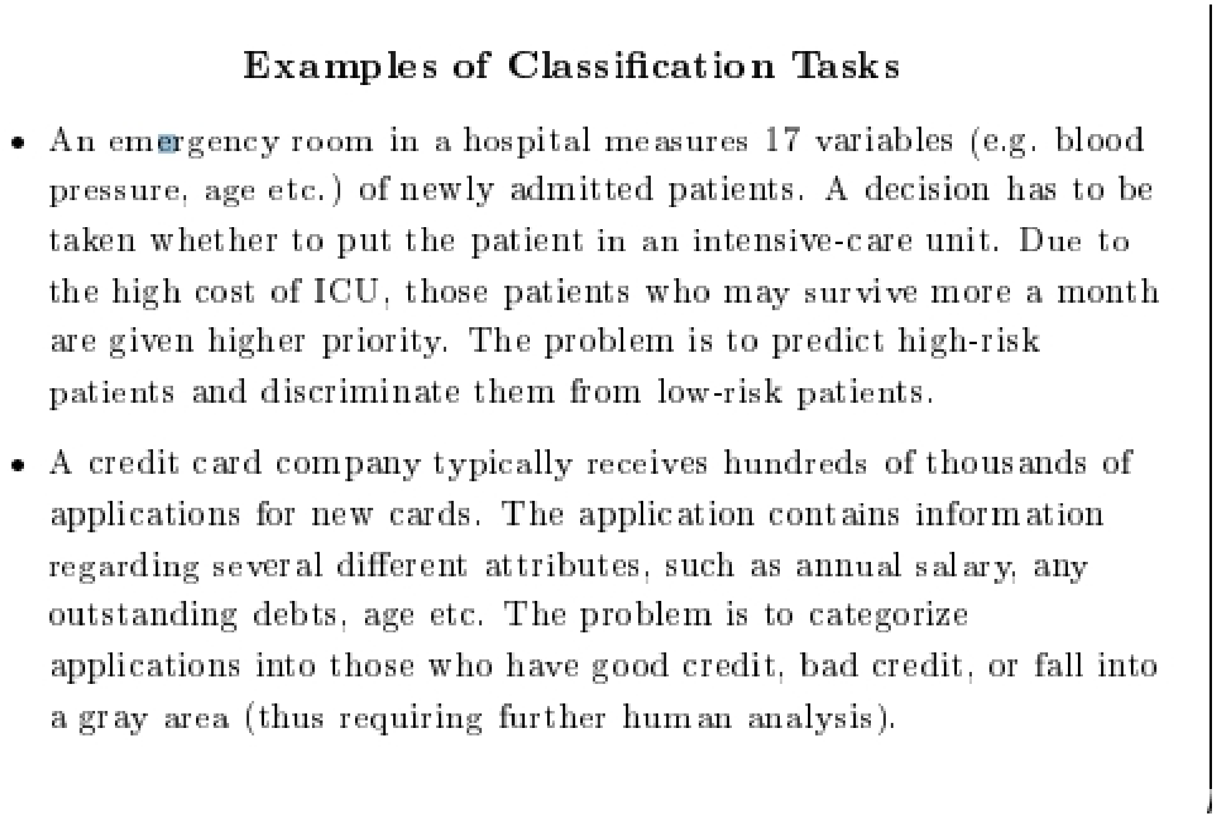
IF age=youth THEN loan_decision=risky
IF income=high THEN
loan-decision=safe
IF age=middle-aged AND income=low
THEN loan_decision=risky

Using Classifier for Classification

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



Name	age	income	loan_decision	
Sandy	young	low	risky	
Bill	young		low	risky
Rick	middle-aged	high	safe	
Joe	middle-aged	low	risky	
Susan	senior	low	safe	
……	……	……	……	



Examples of Classification Tasks

- An emergency room in a hospital measures 17 variables (e.g. blood pressure, age etc.) of newly admitted patients. A decision has to be taken whether to put the patient in an intensive-care unit. Due to the high cost of ICU, those patients who may survive more a month are given higher priority. The problem is to predict high-risk patients and discriminate them from low-risk patients.
- A credit card company typically receives hundreds of thousands of applications for new cards. The application contains information regarding several different attributes, such as annual salary, any outstanding debts, age etc. The problem is to categorize applications into those who have good credit, bad credit, or fall into a gray area (thus requiring further human analysis).

8. What is Clustering?

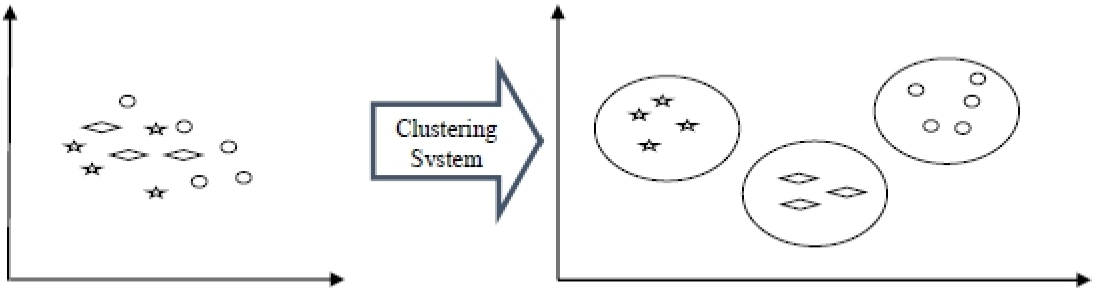
Clustering is the process of making a group of abstract objects into classes of similar objects.

Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.
- Clustering methods are one of the most useful unsupervised ML methods. These methods are used to find similarity as well as the relationship patterns among data samples and then cluster those samples into groups having similarity based on features.
- Clustering is important because it determines the intrinsic grouping among the present unlabeled data. They basically make some assumptions about data points to constitute their similarity. Each assumption will construct different but equally valid clusters.
- For example, below is the diagram which shows clustering system grouped together the similar kind of data in different clusters −



-

Types of ML Clustering Algorithms

The following are the most important and useful ML clustering algorithms −

K-means Clustering

This clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called **flat clustering** algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.

Mean-Shift Algorithm

It is another powerful clustering algorithm used in unsupervised learning. Unlike K-means clustering, it does not make any assumptions hence it is a non-parametric algorithm.

Hierarchical Clustering

It is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics.

We will be discussing all these algorithms in detail in the upcoming chapters.

be used in biological data analysis.

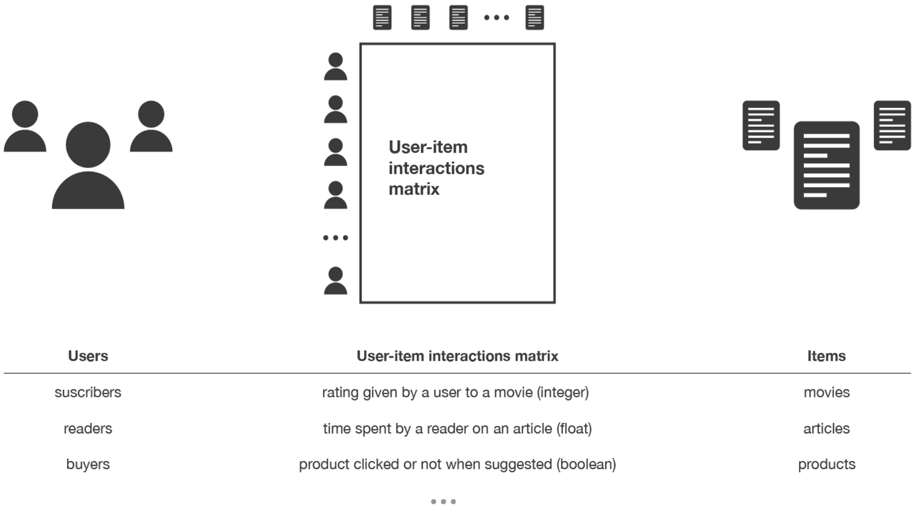
Recommender Systems

During the last few decades, with the rise of Youtube, Amazon, Netflix and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys.

In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy or anything else depending on industries).

Collaborative filtering methods

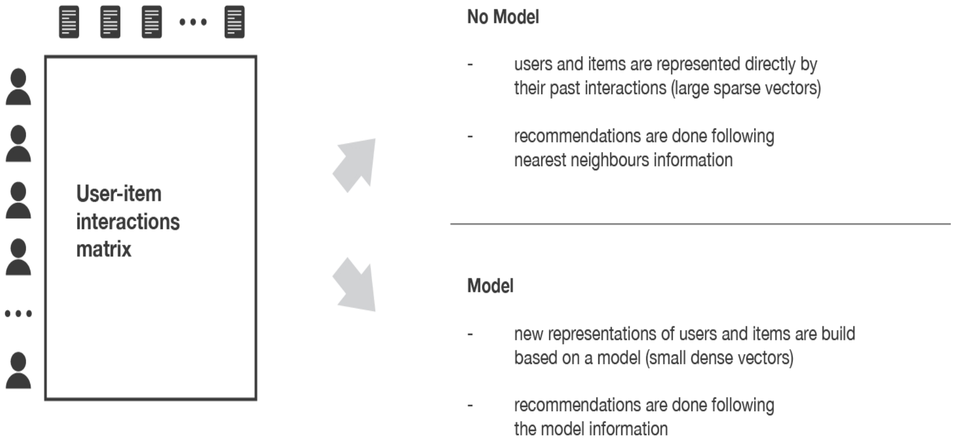
Collaborative methods for recommender systems are methods that are based solely on the past interactions recorded between users and items in order to produce new recommendations. These interactions are stored in the so-called “user-item interactions matrix”.



Users	User-item interactions matrix	Items
suscribers	rating given by a user to a movie (integer)	movies
readers	time spent by a reader on an article (float)	articles
buyers	product clicked or not when suggested (boolean)	products
	•••	

Illustration of the user-item interactions matrix.

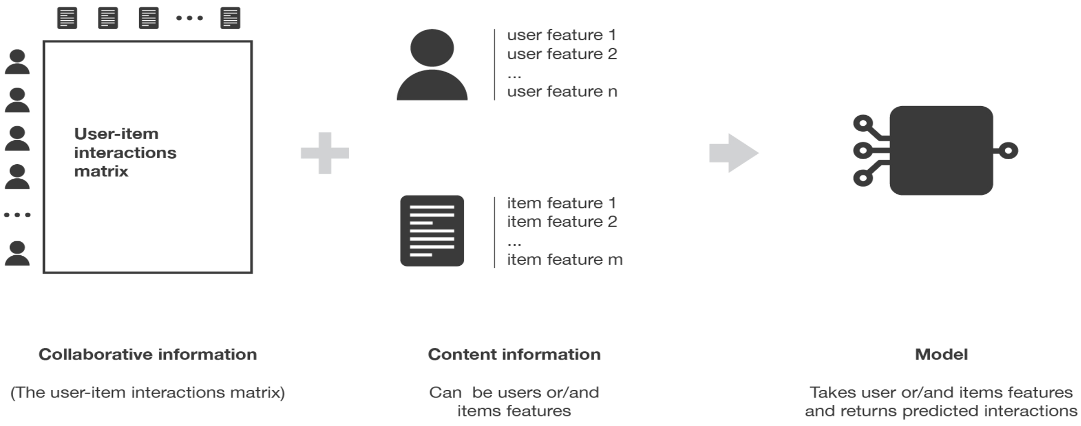
Then, the main idea that rules collaborative methods is that these past user-item interactions are sufficient to detect similar users and/or similar items and make predictions based on these estimated proximities.

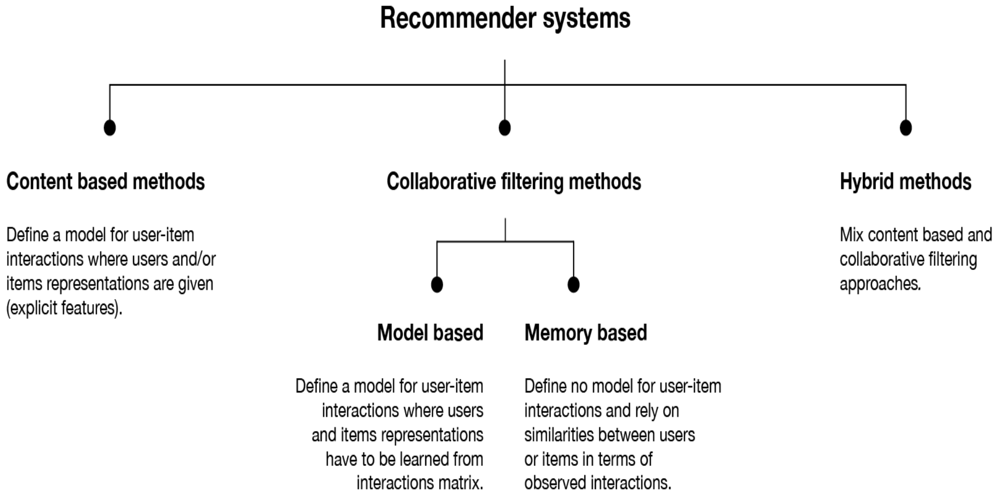

User-item interactions matrix
No Model
- users and items are represented directly by their past interactions (large sparse vectors)
- recommendations are done following nearest neighbours information
Model
- new representations of users and items are build based on a model (small dense vectors)
- recommendations are done following the model information

Overview of the collaborative filtering methods paradigm.

Content based methods

Unlike collaborative methods that only rely on the user-item interactions, content based approaches use additional information about users and/or items. If we consider the example of a movies recommender system, this additional information can be, for example, the age, the gender, the job or any other personal information for users as well as the category, the main actors, the duration or other characteristics for the movies (items).


User-item interactions matrix
user feature 1
user feature 2
...
user feature n
item feature 1
item feature 2
...
item feature m
Collaborative information
(The user-item interactions matrix)
Content information
Can be users or/and items features
Model
Takes user or/and items features and returns predicted interactions



Recommender systems

Content based methods

Define a model for user-item interactions where users and/or items representations are given (explicit features).

Collaborative filtering methods

Model based

Define a model for user-item interactions where users and items representations have to be learned from interactions matrix.

Memory based

Define no model for user-item interactions and rely on similarities between users or items in terms of observed interactions.

Hybrid methods

Mix content based and collaborative filtering approaches.