**AIFE: UNIT-2**
***MS. SHWETA TIWARI***
***Published: NOVEMBER, 2021***

# "Artificial Intelligence for Engineering/Engineers (KMC-201)"_

# UNIT-2: DATA And ALGORITHMS

FALL SEMESTER, YEAR (2ND, 1ST)

**FALL SESSION (2021-22)**
***(AIFE)***
***MS. SHWETA TIWARI***
***Published: NOVEMBER, 2021***

## Rajkiya Engineering College | Ambedkar Nagar, UP, India

**Faculty Name: Ms. Shweta Tiwari, Subject: Artificial Intelligence for Engineering (KMC201) Year- 1st Year, Semester- 2nd Sem, Branch- CE, Session- Even Semester (2021-22)**

PREPARED FOR
Engineering Students
All Engineering College

PREPARED BY
SHWETA TIWARI
Guest Faculty

November, 2021

**Notes Part-2**
MS. SHWETA TIWARI

# UNIT-2

# “Artificial Intelligence for Engineering/Engineers (KMC-201)”

**By SHWETA TIWARI**

Artificial Intelligence for Engineering/Engineers (KMC-201)”

## ***UNIT-2: DATA And ALGORITHMS***

November, 2021

**Notes Part-2**

MS. SHWETA TIWARI

# UNIT-2: DATA And ALGORITHMS

November, 2021

**Notes Part-2**

MS. SHWETA TIWARI

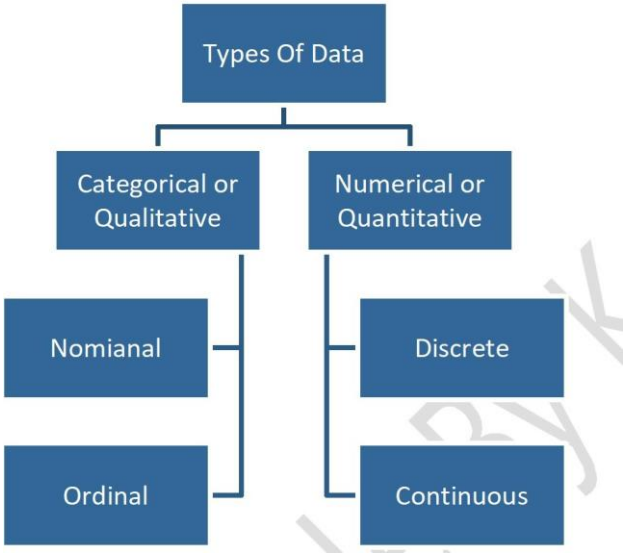
# Unit-2: 'AI Unit 2 - Data & Algorithms'

## 2.1 Data & it's Types

In Computers, Data is information processed (used) or stored by a computer. This information may be in the form of text documents, images, audio clips, software programs, or other types of data.

Data is a bunch of ones and zeros, known as binary data. So. it can be created, processed, saved, and stored digitally.

- **Structured data** is highly-organized and formatted in a way so it's easily searchable in relational databases.
- **Unstructured data** has no pre-defined format or organization, making it much more difficult to collect, process, and analyze.




**A) <u>Qualitative Data Type</u>**

This type of data can't be counted or measured easily using numbers.

1. **Nominal:** These are the set of values that don't possess a natural ordering.
   Example- Colours, we can't compare one color with others. It is not possible to state that 'Red' is greater than 'Blue'.

2. **Ordinal:** These types of values have a natural ordering while maintaining their class of values. Example- Size of Shirt, we can easily sort them according to their name tag in the order of small < medium < large.

**B) Quantitative Data Type**

This data type tries to quantify things and it does by considering numerical values that make it countable in nature.

1. **Discrete:** The numerical values which fall under are integers or whole numbers are placed under this category.

2. **Continuous:** The fractional numbers are considered as continuous values.

## 2.1 Algorithms

An algorithm is a set of instructions designed to perform a specific task or solving a problem.

Big data is a huge data that it does not fit in the main memory of a single machine, and the need to process big data by efficient algorithms.

## 2.2 History of Data

Data feeds the growth of exciting technology like artificial intelligence (AI). It helps personalise our online experiences, and it helps us move forward in the quest for more knowledge.

**1600s: Early Interpretations**

In 1640s, The word 'data' first saw English use. **John Graunt** conducted one of the earliest recorded instances of data analysis.

So, Graunt is generally considered to be the founder.

**Late 1800s: Data Processing Problems**

In 1880 US census. The issue was that there was now more data of populance than collectors could analyse.

**Hollerith**'s machine made it possible to process and analyse large amounts of data.

**1900s: A Question of Storage**

In 1928, **Fritz Pfleumer** invented magnetic tape for recording purposes (to collect & store data). The hard disc drives, floppy discs and tapes that would follow toward the end of the century were all enabled by magnetic data storage.

In 1960s, the conception (but not the creation) of cloud data storage by **Dr Joseph** is given. This idea forms the basis of cloud computing.

In 1970s, E.F. Codd presented a framework for a relational model of database management.

**1990s: The Internet**

Naturally, the most noteworthy 1990s event in the history of data is the invention of the internet.

**Sir Tim Berners Lee** created hyperlinks and hypertext, enabling data sharing worldwide.

In mid 1990s, The first instance of all web-based storage was launched by AT&T.

In 1997, The Google Search Engine was launched. This put data very much in the hands of anyone with computer access.

## 2.3 Data Storage & it's Types

Data storage means that files and documents are recorded digitally and saved in a storage system for future use.
Data storage can occur on physical hard drives, disk drives, USB drives or virtually in the cloud.

**Types of Data Storage:**
There are two types of data Storage.

1. **Direct Attached Storage (DAS)**
   It is a types of data storage that are physically connected to your computer.
   This storage is generally accessible to only a single machine.
   Example: Hard Drive, SSD, CD/DVD, Pendrives,etc.

2. **Network Attached Storage (NAS)**
   It allows multiple machines to share storage over a network. Data can be easily shared among connected machines.

**Types of Data Storage Devices:**
There are lot of Data Storage device availble in market, some of them are old and slow. Example -

1. Hard Drive Disks (HDD)
2. Floppy Disks
3. Tapes
4. Compact Discs (CDs)
5. DVD and Blu-ray Discs
6. USB Flash Drives
7. Secure Digital Cards (SD Cards)
8. Solid-State Drives (SSDs)
9. Cloud Storage, etc.

## 2.4 Importance of Data

Below are the various reasons, why the data is important.

**1. Improve people's lives**
Data will help people to improve quality of life: Improving quality is first and foremost among the reasons why organizations should be using data.

**2. Make informed decisions**
Data = Knowledge. Good data provides indisputable evidence, while casually collected data might lead to wasted resources due to taking action based on an incorrect conclusion.

**3. Find solutions to problems**
Data allows organizations to more effectively determine the cause of problems then find the appropriate solution based on the data.

**4. Be strategic in your approaches**
Data increases efficiency. Effective data collection and analysis will allow you to make a good strategy.

**5. Get The Results You Want**
Data allows organizations to measure the effectiveness of a given strategy. Collecting data will allow you to determine how well your solution is performing, and whether or not your approach needs to be tweaked or changed over the long-term.

**6. Stop The Guessing Game**
Data will help you explain (both good and bad) decisions to your stakeholders. You can be confident that you developed your approach based not upon guesses, but good solid data.

**7. Know What You Are Doing Well**
Data analysis will support you to identify high-performing programs, service areas, and people. Once you identify your high-performers, you can study them in order to develop strategies to assist programs, service areas and people that are low-performing.

**8. Keep Track Of It All**
Good data allows organizations to establish baselines, benchmarks, and goals to keep moving forward. Because data allows you to measure, you will be able to establish baselines, find benchmarks and set performance goals.

**9. Stand out from the crowd**
By knowing and comparing your business against others, you can find out what you offer that others don't, and use this to your advantage.

## 2.5 Data Acquisition (DAQ/DAS)

**Data acquisition** is the process of digitizing data from the world around us. So that it can be displayed, analyzed and stored in a computer.
Example - The process of measuring temperature in a room as a digital value using a sensor.

**Data Acquisition Systems**: The systems, used for data acquisition are known as data acquisition systems.
These data acquisition systems will perform the tasks such as conversion of data, storage of data, transmission of data and processing of data.

**Types of Data Acquisition Systems:**
Data acquisition systems can be classified into the following two types.

1. **Analog Data Acquisition Systems :** It is a systems, which can be operated with analog signals are known as analog data acquisition systems. Following are the blocks of analog data acquisition systems.
   - **Transducer** − It converts physical quantities into electrical signals.
   - **Signal conditioner** − It performs the functions like amplification and selection of desired portion of the signal.
   - **Display device** − It displays the input signals for monitoring purpose.
   - **Graphic recording instruments** − These can be used to make the record of input data permanently.
   - **Magnetic tape instrumentation** − It is used for acquiring, storing & reproducing of input data.

2. **Digital Data Acquisition Systems :** It is a systems, which can be operated with digital signals are known as digital data acquisition systems. So, they use digital components for storing or displaying the information. Following are the blocks of Digital data acquisition systems.
   - **Transducer** − It converts physical quantities into electrical signals.
   - **Signal conditioner** − It performs the functions like amplification and selection of desired portion of the signal.
   - **Multiplexer** − connects one of the multiple inputs to output. So, it acts as parallel to serial converter.
   - **Analog to Digital Converter** − It converts the analog input into its equivalent digital output.
   - **Display device** − It displays the data in digital format.
   - **Digital Recorder** − It is used to record the data in digital format.

## 2.6 Data Processing

It is a process of converting raw facts or data into a meaningful information. Below are the stages of Data processing.

1. **Collection**
   - Collection of data refers to gathering of data.
   - The data gathered should be defined and accurate.
2. **Preparation**
   Preparation is a process of constructing a dataset of data from different sources for future use.
3. **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.
4. **Processing**
   - The process refers to concept of an actual execution of instructions.
   - In this stage, raw facts or data is converted to meaningful information.
5. **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.
6. **Storage**
   In this process, we can store data, instruction and information in permanent memory for future reference.

## 2.7 Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps.

It enables decision makers to see analytics presented visually.

It helps users in analyzing a large amount of data in a simpler way because it converts large and small data sets into visuals, which is easy to understand and process for humans.

**Why Use Data Visualization?**

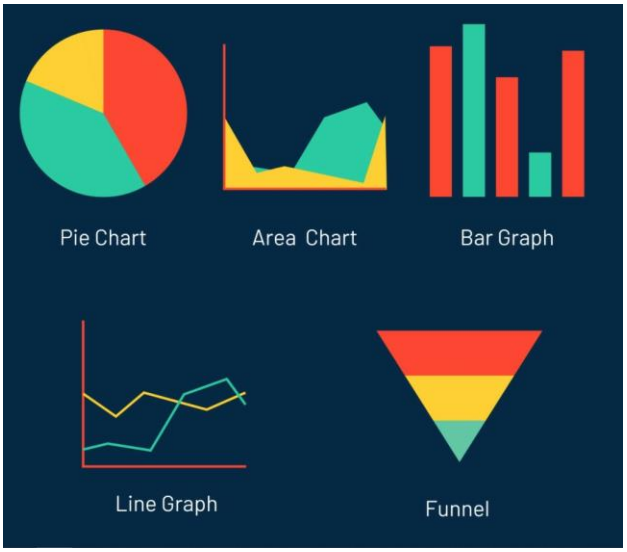
✓ To make easier in understand and remember.
✓ To discover unknown facts, outliers, and trends.
✓ To ask a better question and make better decisions.
✓ To competitive analyze.
✓ To improve insights.

**Pros of Data Visualization**

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

**General types of data visualization:**

- Charts
- Tables
- Graphs
- Maps
- Histograms
- Heat maps
- Tree maps
- Pie Chart
- Venn Diagram




## 2.8 Regression

Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of Market Trends, prediction of House prices, etc.

The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).

**Example** - For weather forecasting, we will use the Regression algorithm. The model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

**Types of Regression Algorithm:**

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

## 2.9 Classification

Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters.

In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes.

The task of the classification algorithm is to find the mapping function to map the input(x) to the discrete output(y).

**Example** - In Email Spam Detection, The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.

**Types of ML Classification Algorithms:**

- Logistic Regression
- K-Nearest Neighbours
- Support Vector Machines
- Kernel SVM
- Naïve Bayes
- Decision Tree Classification
- Random Forest Classification

## 2.10 Regression V/s Classification

| Regression Algorithm | Classification Algorithm |
|---|---|
| The output variable must be of continuous nature or real value. | The output variable must be a discrete value. |
| The task is to map the input value (x) with the continuous output variable(y). | The task is to map the input value(x) with the discrete output variable(y). |
| Used with continuous data. | Used with discrete data. |
| In Regression, we try to find the best fit line, to predict the output more accurately. | In Classification, we try to find the decision boundary, to divide the dataset into different classes. |
| Used to solve such as Weather Prediction, House price prediction, etc. | Used to solve such as Identification of spam emails, Identification of cancer cells, etc. |
| Divided into Linear and Non-linear Regression. | Divided into Binary Classifier and Multi-class Classifier. |

## 2.11 Prediction

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome.

It is a type of guess. However, a prediction is an estimation made from observation

Example - You observe that everytime the wind blow, flower petals fall from the tree. So, you can predict that if the wind blows, petals will fall from the tree.

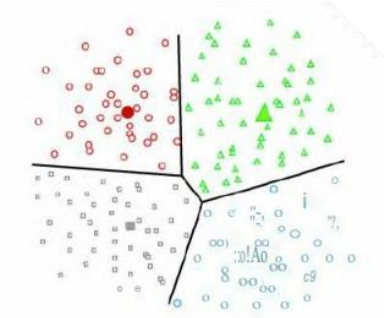
## 2.12 Clustering

Clustering methods are one of the most useful unsupervised machine learning 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.

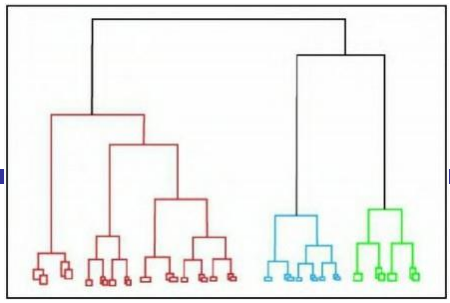
In Simple words, Cluster is a group of objects that belongs to the same class. Which means, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

The following are the most important and useful clustering algorithms:

1. **K-means Clustering:** This clustering algorithm computes the centroids and iterates until it finds optimal centroid. It assumes that the numbers 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.



2. **Hierarchical Clustering:** It is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics.



**Applications of Clustering:**

- **In Identification of Cancer Cells:** The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.
- **In Biology:** It is used in the biology stream to classify different species of plants and animals using the image recognition technique.
- **Customer Segmentation:** It is used in market research to segment the customers based on their choice and preferences.
- **Libraries:** Clustering books based on Topics & Information.

## 2.13 Recommender Systems

Recommender systems are the systems that are designed to recommend things to the user based on many different factors. These systems predict the most likely product that the users are most likely to purchase and are of interest to. Companies like Netflix, Amazon, etc. use recommender systems.

**Benefits of Recommender systems**

- Benefits users in finding items of their interest.
- Help item providers in delivering their items to the right user.
- Identity products that are most relevant to users.
- Personalized content.
- Help websites to improve user engagement.

**Methods for Building Recommender systems**

1. **Content-Based Recommendation:** The goal of this content based recommendation is to predict the scores for unrated items of the users. The basic idea behind content filtering is that each items have some feature x. For example, a movie has a high score for a feature $x_1$ but a low score for feature $x_2$

2. **Collaborative filtering:** The disadvantage of the content filtering is that it need the side information for each item. The collaborative filtering are done based on the user's behaviour. History of the user plays an important role. It is of two types.

   I. **User-User collaborative filtering:** In this, user vector include all the item purchased by the user and the rating given for each particular product.

   II. **Item-Item collaborative filtering:** In this, rather

than considering similar users, similar items are considered.