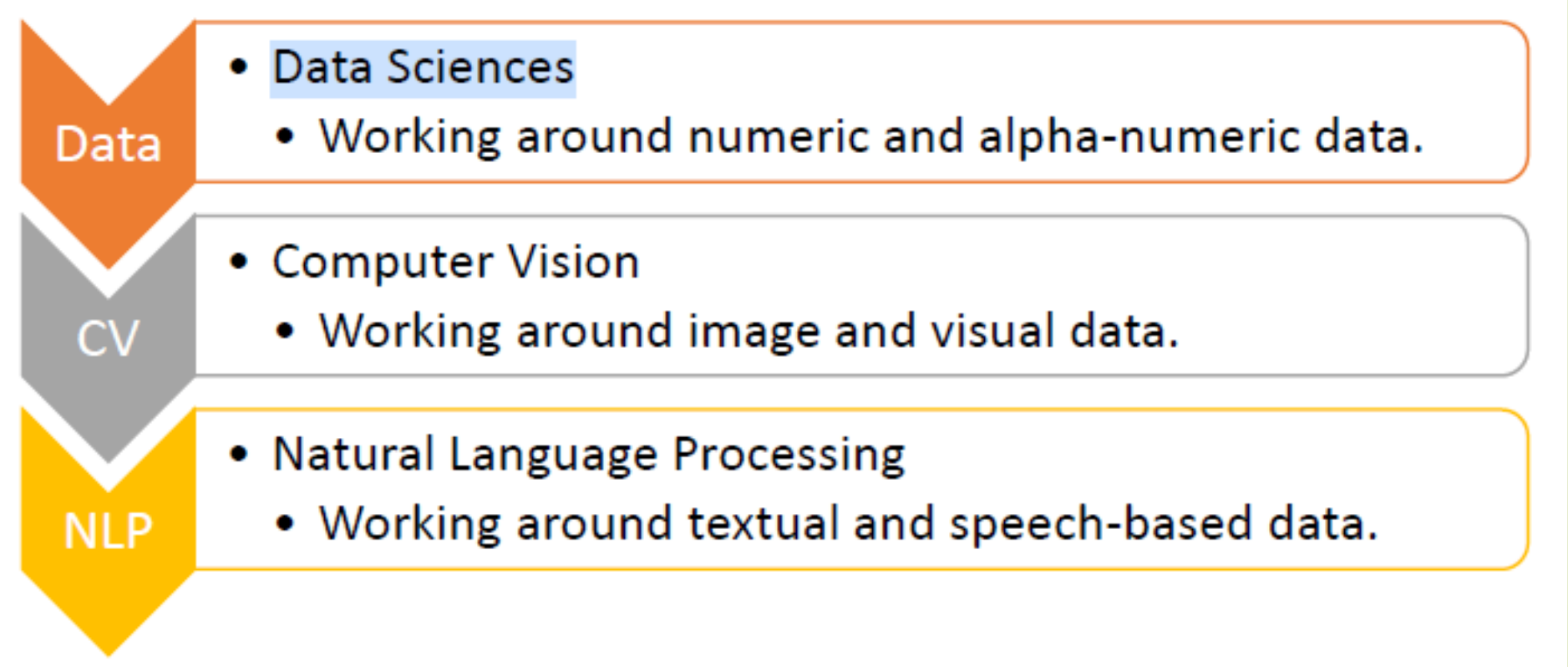
Chapter 3
Data science

INDIAN SCHOOL AL WADI AL KABIR

CLASS 10 ARTIFICIAL INTELLIGENCE

Introduction

- Artificial Intelligence is a technology which completely depends on data. It is the data which is fed into the machine which makes it intelligent.
- And depending upon the type of data we have; AI can be classified into three broad domains:



DATA SCIENCE

- **Data Sciences is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyse actual phenomena with data.**
- **It employs techniques and theories drawn from many fields within the context of Mathematics, Statistics, Computer Science, and Information Science.**
- **Rock, Paper & Scissors: https://www.afiniti.com/corporate/rock-paper-scissors**

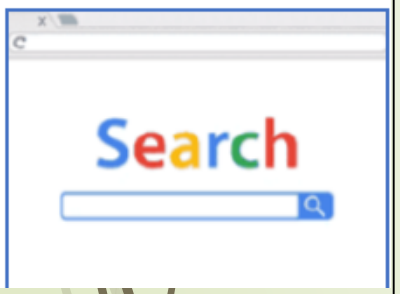
Applications of Data Sciences

- Data Science is not a new field. Data Sciences majorly work around analysing the data and when it comes to AI, the analysis helps in making the machine intelligent enough to perform tasks by itself. There exist various applications of Data Science in today's world. Some of them are:

1. Fraud and Risk Detection*: **The earliest applications of data science were in Finance. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paperwork while sanctioning loans. They decided to bring in data scientists in order to rescue them from losses.**

Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyse the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

2. **Genetics & Genomics*:** Data Science applications also enable an advanced level of treatment personalization through research in genetics and genomics. The goal is to understand the impact of the DNA on our health and find individual biological connections between genetics, diseases, and drug response. Data science techniques allow integration of different kinds of data with genomic data in disease research, which provides a deeper understanding of genetic issues in reactions to particular drugs and diseases. As soon as we acquire reliable personal genome data, we will achieve a deeper understanding of the human DNA. The advanced genetic risk prediction will be a major step towards more individual care.



3. Internet Search*: When we talk about search engines, we think 'Google'. Right? But there are many other search engines like Yahoo, Bing, Ask, AOL, and so on. All these search engines (including Google) make use of data science algorithms to deliver the best result for our searched query in the fraction of a second. Considering the fact that Google processes more than 20 petabytes of data every day, had there been no data science, Google wouldn't have been the 'Google' we know today.


Bytes
Megabyte
1,000,000
Gigabyte
1,000,000,000
Terabyte
1,000,000,000,000
Petabyte
1,000,000,000,000,000
Exabyte
1,000,000,000,000,000,000
Zettabyte
1,000,000,000,000,000,000,000
Yottabyte
1,000,000,000,000,000,000,000,000

4. Targeted Advertising*: **If you thought Search would have been the biggest of all data science applications, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital billboards at the airports – almost all of them are decided by using data science algorithms. This is the reason why digital ads have been able to get a much higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user's past behavior**

5. Website Recommendations:* **Aren't we all used to the suggestions about similar products on Amazon? They not only help us find relevant products from billions of products available with them but also add a lot to the user experience.**
A lot of companies have fervidly used this engine to promote their products in accordance with the user's interest and relevance of information. Internet giants like Amazon, Twitter, Google Play, Netflix, LinkedIn, IMDB and many more use this system to improve the user experience. The recommendations are made based on previous search results for a user.

*CTR is the number of clicks that your ad receives divided by the number of times your ad is shown:

6. Airline Route Planning*: **The Airline Industry across the world is known to bear heavy losses. Except for a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air-fuel prices and the need to offer heavy discounts to customers, the situation has got worse. It wasn't long before airline companies started using Data Science to identify the strategic areas of improvements. Now, while using Data Science, the airline companies can:**

- **Decide which class of airplanes to buy**
- **Whether to directly land at the destination or take a halt in between (For example, A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)**
- **Effectively drive customer loyalty programs**

Data Collection

Data collection is nothing new which has come up in our lives. It has been in our society since ages. Even when people did not have fair knowledge of calculations, records were still maintained in some way or the other to keep an account of relevant things. Data collection is an exercise which does not require even a tiny bit of technological knowledge. But when it comes to analysing the data, it becomes a tedious process for humans as it is all about numbers and alpha-numerical data. That is where Data Science comes into the picture. It not only gives us a clearer idea around the dataset, but also adds value to it by providing deeper and clearer analyses around it. And as AI gets incorporated in the process, predictions and suggestions by the machine become possible on the same.

Now that we have gone through an example of a Data Science based project, we have a bit of clarity regarding the type of data that can be used to develop a Data Science related project. For the data domain-based projects, majorly the type of data used is in numerical or alpha-numerical format and such datasets are curated in the form of tables. Such databases are very commonly found in any institution for record maintenance and other purposes. Some examples of datasets which you must already be aware of are:

Banks	Databases of loans issued, account holder, locker owners, employee registrations, bank visitors, etc.
ATM Machines	Usage details per day, cash denominations transaction details, visitor details, etc.
Movie Theatres	Movie details, tickets sold offline, tickets sold online, refreshment purchases, etc.

Now look around you and find out what are the different types of databases which are maintained in the places mentioned below. Try surveying people who are responsible for the designated places to get a better idea.

Your classroom

Your school

Your city

Sources of Data

- There exist various sources of data from where we can collect any type of data required and the data collection process can be categorised in two ways: Offline and Online.

Offline Data Collection	Online Data Collection
Sensors	Open-sourced Government Portals
Surveys	Reliable Websites (Kaggle)
Interviews	World Organisations' open-sourced statistical websites
Observations	

While accessing data from any of the data sources, following points should be kept in mind:

1. Data which is available for public usage only should be taken up.
2. Personal datasets should only be used with the consent of the owner.
3. One should never breach someone's privacy to collect data.
4. Data should only be taken form reliable sources as the data collected from random sources can be wrong or unusable.
5. Reliable sources of data ensure the authenticity of data which helps in proper training of the AI model.

Types of Data

For Data Science, usually the data is collected in the form of tables. These tabular datasets can be stored in different formats. Some of the commonly used formats are:

1. CSV: CSV stands for comma separated values. It is a simple file format used to store tabular data. Each line of this file is a data record and reach record consists of one or more fields which are separated by commas. Since the values of records are separated by a comma, hence they are known as CSV files.
2. Spreadsheet: A Spreadsheet is a piece of paper or a computer program which is used for accounting and recording data using rows and columns into which information can be entered. Microsoft excel is a program which helps in creating spreadsheets.
3. SQL: SQL is a programming language also known as Structured Query Language. It is a domain-specific language used in programming and is designed for managing data held in different kinds of DBMS (Database Management System) It is particularly useful in handling structured data.

Data Access

After collecting the data, to be able to use it for programming purposes, we should know how to access the same in a Python code. To make our lives easier, there exist various Python packages which help us in accessing structured data (in tabular form) inside the code. Let us take a look at some of these packages:

1. NumPy

NumPy, which stands for Numerical Python, is the fundamental package for Mathematical and logical operations on arrays in Python. It is a commonly used package when it comes to working around numbers. NumPy gives a wide range of arithmetic operations around numbers giving us an easier approach in working with them. NumPy also works with arrays, which is nothing but a homogenous collection of Data.

```
import numpy
A=numpy.array([1,2,3,4,5,6,7,8,9,0])
```

2. Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. The name is derived from the term "panel data",

Pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labelled at all to be placed into a Pandas data structure

3. Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They're typically instruments for reasoning about quantitative information. Some types of graphs that we can make with this package are listed below: