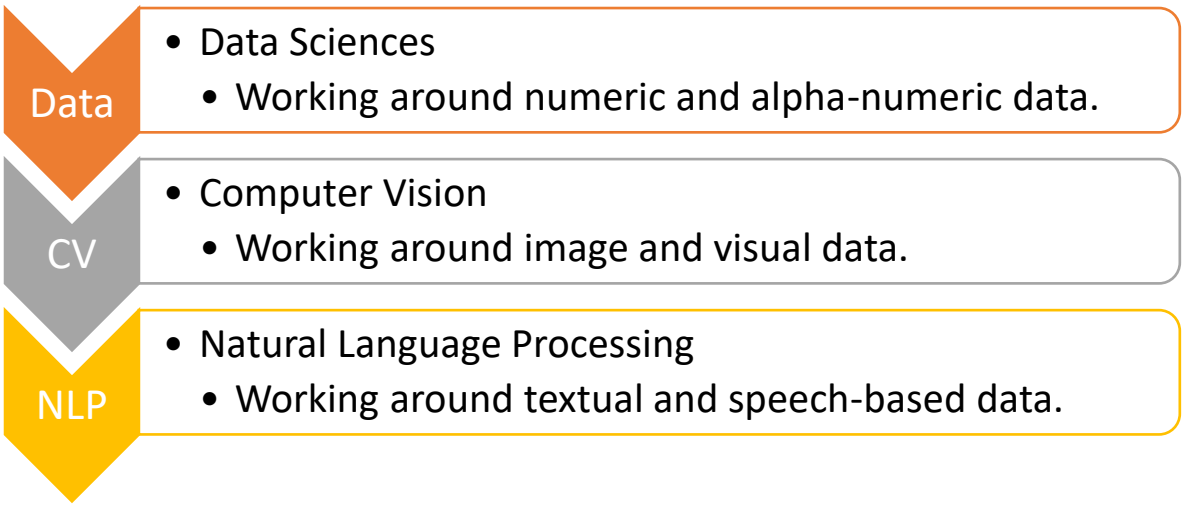
Data Sciences

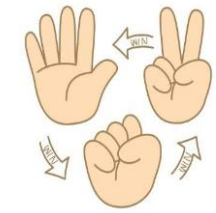
Introduction

As we have discussed earlier in class 9, 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:



Each domain has its own type of data which gets fed into the machine and hence has its own way of working around it. Talking about Data Sciences, it 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.

Now before we get into the concepts of Data Sciences, let us experience this domain with the help of the following game:



*** Rock, Paper & Scissors: https://www.afiniti.com/corporate/rock-paper-scissors**

Go to this link and try to play the game of Rock, Paper Scissors against an AI model. The challenge here is to win 20 games against AI before AI wins them against you.

Did you manage to win?

_____

_____

What was the strategy that you applied to win this game against the AI machine?

_____
_____
_____
_____

Was it different playing Rock, Paper & Scissors with an AI machine as compared to a human?

_____
_____
_____
_____

What approach was the machine following while playing against you?

_____
_____
_____
_____

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:

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.

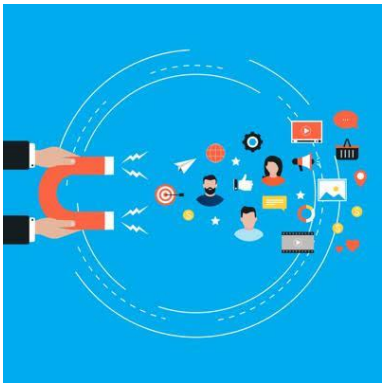
Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyse the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

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.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

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.

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 banilrs 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 behaviour.



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.

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:

* Images shown here are the property of individual organisations and are used here for reference purpose only.

- Predict flight delay
- 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

Getting Started

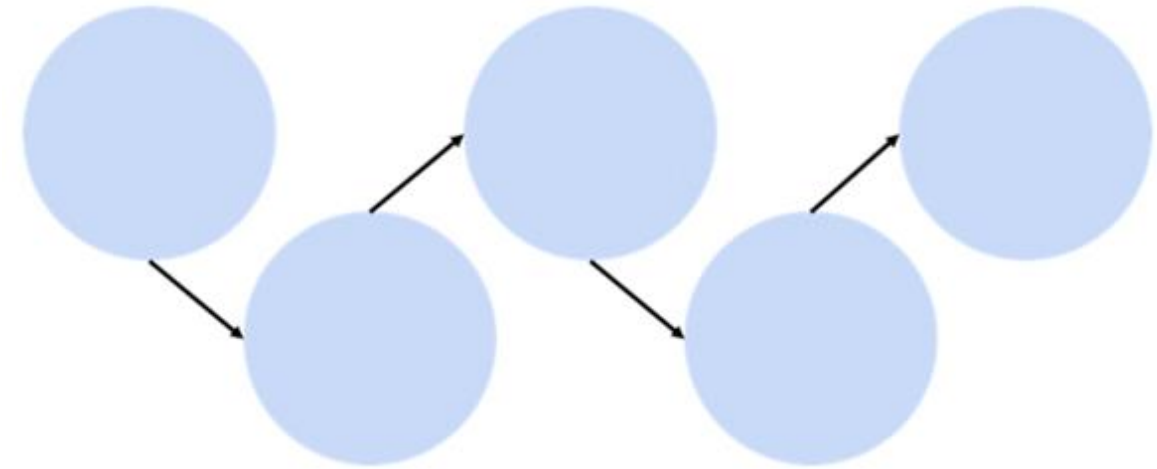
Data Sciences is a combination of Python and Mathematical concepts like Statistics, Data Analysis, probability, etc. Concepts of Data Science can be used in developing applications around AI as it gives a strong base for data analysis in Python.

Revisiting AI Project Cycle

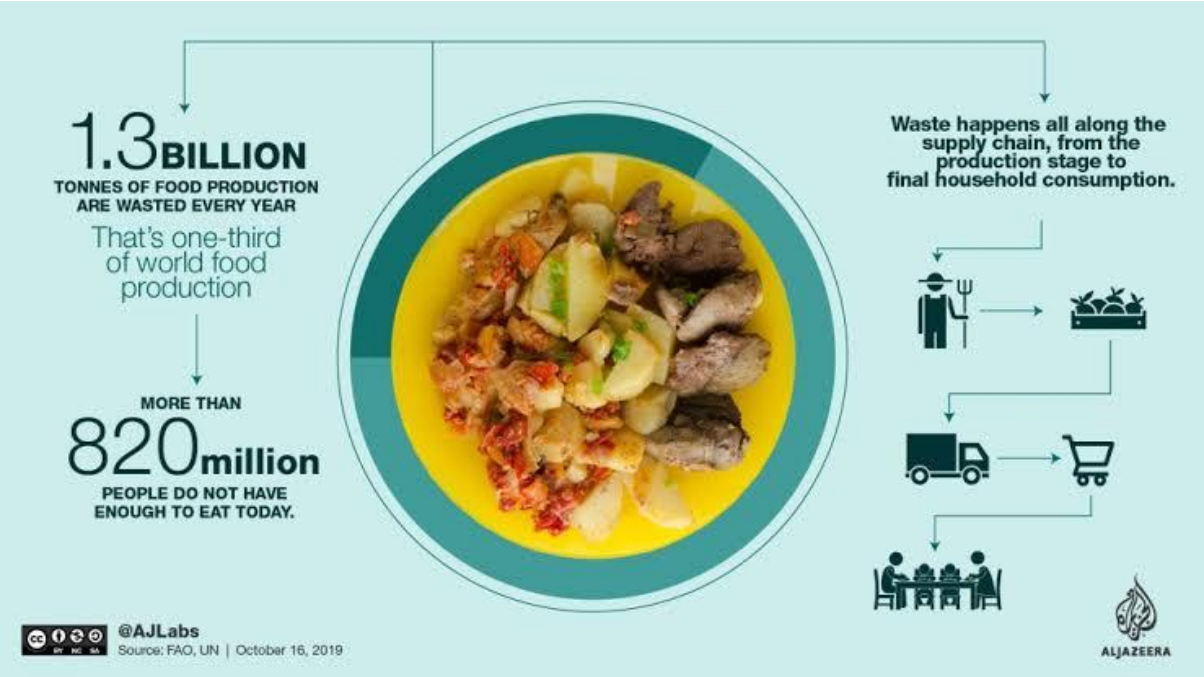
But, before we get deeper into data analysis, let us recall how Data Sciences can be leveraged to solve some of the pressing problems around us. For this, let us understand the AI project cycle framework around Data Sciences with the help of an example.

Do you remember the AI Project Cycle?

Fill in all the stages of the cycle here:



The Scenario*



Humans are social animals. We tend to organise and/or participate in various kinds of social gatherings all the time. We love eating out with friends and family because of which we can find restaurants almost everywhere and out of these, many of the restaurants arrange for buffets to offer a variety of food items to their customers. Be it small shops or big outlets, every restaurant prepares food in bulk as they expect a good crowd to come and enjoy their food. But in most cases, after the day ends, a lot of food is left which becomes unusable for the restaurant as they do not wish to serve stale food to their customers the next day. So, every day, they prepare food in large quantities keeping in mind the probable number of customers walking into their outlet. But if the expectations are not met, a good amount of food gets wasted which eventually becomes a loss for the restaurant as they either have to dump it or give it to hungry people for free. And if this daily loss is taken into account for a year, it becomes quite a big amount.

Problem Scoping

Now that we have understood the scenario well, let us take a deeper look into the problem to find out more about various factors around it. Let us fill up the 4Ws problem canvas to find out.

Who Canvas – Who is having the problem?

Who are the stakeholders?	○ Restaurants offering buffets ○ Restaurant Chefs
What do we know about them?	○ Restaurants cook food in bulk every day for their buffets to meet their customer needs. ○ They estimate the number of customers that would walk into their restaurant every day.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

What Canvas – What is the nature of their problem?

What is the problem?	○ Quite a large amount of food is leftover everyday unconsumed at the restaurant which is either thrown away or given for free to needy people. ○ Restaurants have to bear everyday losses for the unconsumed food.
How do you know it is a problem?	○ Restaurant Surveys have shown that restaurants face this problem of food waste.

Where Canvas – Where does the problem arise?

What is the context/situation in which the stakeholders experience this problem?	○ Restaurants which serve buffet food ○ At the end of the day, when no further food consumption is possible

Why? – Why do you think it is a problem worth solving?

What would be of key value to the stakeholders?	○ If the restaurant has a proper estimate of the quantity of food to be prepared every day, the food waste can be reduced.
How would it improve their situation?	○ Less or no food would be left unconsumed. ○ Losses due to unconsumed food would reduce considerably.

Now that we have noted down all the factors around our problem, let us fill up the problem statement template.

Our	**Restaurant Owners**	**Who?**
Have a problem of	Losses due to food wastage	**What?**
While	The food is left unconsumed due to improper estimation	**Where?**
An ideal solution would	Be to be able to predict the amount of food to be prepared for every day consumption	**Why**

The Problem statement template leads us towards the goal of our project which can now be stated as:

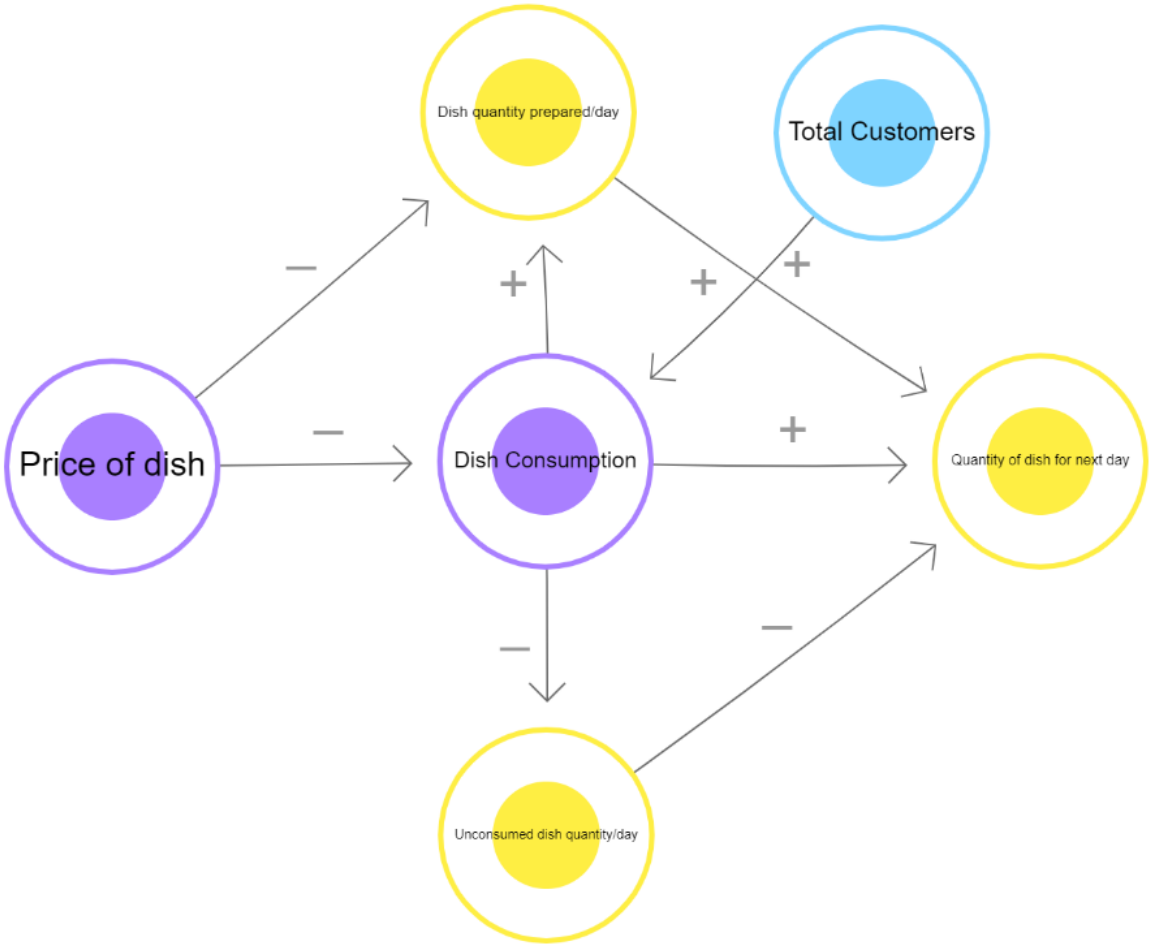
"To be able to predict the quantity of food dishes to be prepared for everyday consumption in restaurant buffets."

Data Acquisition

After finalising the goal of our project, let us now move towards looking at various data features which affect the problem in some way or the other. Since any AI-based project requires data for testing and training, we need to understand what kind of data is to be collected to work towards the goal. In our scenario, various factors that would affect the quantity of food to be prepared for the next day consumption in buffets would be:

Total Number of Customers	Quantity of dish prepared per day	Dish consumption
Unconsumed dish quantity per day	Price of dish	Quantity of dish for the next day

Now let us understand how these factors are related to our problem statement. For this, we can use the System Maps tool to figure out the relationship of elements with the project's goal. Here is the System map for our problem statement.



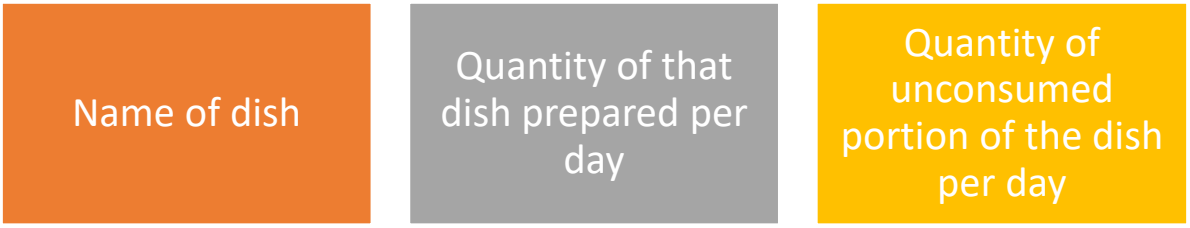
In this system map, you can see how the relationship of each element is defined with the goal of our project. Recall that the positive arrows determine a direct relationship of elements while the negative ones show an inverse relationship of elements.

After looking at the factors affecting our problem statement, now it's time to take a look at the data which is to be acquired for the goal. For this problem, a dataset covering all the elements mentioned above is made for each dish prepared by the restaurant over a period of 30 days. This data is collected offline in the form of a regular survey since this is a personalised dataset created just for one restaurant's needs.

Specifically, the data collected comes under the following categories: Name of the dish, Price of the dish, Quantity of dish produced per day, Quantity of dish left unconsumed per day, Total number of customers per day, Fixed customers per day, etc.

Data Exploration

After creating the database, we now need to look at the data collected and understand what is required out of it. In this case, since the goal of our project is to be able to predict the quantity of food to be prepared for the next day, we need to have the following data:



Thus, we extract the required information from the curated dataset and clean it up in such a way that there exist no errors or missing elements in it.

Modelling

Once the dataset is ready, we train our model on it. In this case, a regression model is chosen in which the dataset is fed as a dataframe and is trained accordingly. Regression is a Supervised Learning model which takes in continuous values of data over a period of time. Since in our case the data which we have is a continuous data of 30 days, we can use the regression model so that it predicts the next values to it in a similar manilr. In this case, the dataset of 30 days is divided in a ratio of 2:1 for training and testing respectively. In this case, the model is first trained on the 20-day data and then gets evaluated for the rest of the 10 days.

Evaluation

Once the model has been trained on the training dataset of 20 days, it is now time to see if the model is working properly or not. Let us see how the model works and how is it tested.

Step 1: The trained model is fed data regards the name of the dish and the quantity produced for the same.

Step 2: It is then fed data regards the quantity of food left unconsumed for the same dish on previous occasions.

Step 3: The model then works upon the entries according to the training it got at the modelling stage.

Step 4: The Model predicts the quantity of food to be prepared for the next day.

Step 5: The prediction is compared to the testing dataset value. From the testing dataset, ideally, we can say that the quantity of food to be produced for next day's consumption should be the total quantity minus the unconsumed quantity.

Step 6: The model is tested for 10 testing datasets kept aside while training.

Step 7: Prediction values of testing dataset is compared to the actual values.

Step 8: If the prediction value is same or almost similar to the actual values, the model is said to be accurate. Otherwise, either the model selection is changed or the model is trained on more data for better accuracy.

Once the model is able to achieve optimum efficiency, it is ready to be deployed in the restaurant for real-time usage.

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

As you can see, all the type of data which has been mentioned above is in the form of tables. Tables which contain numeric or alpha-numeric data. But this leads to a very critical dilemma: are these datasets accessible to all? Should these databases be accessible to all? What are the various sources of data from which we can gather such databases? Let's find out!

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.

A lot of other formats of databases also exist, you can explore them online!

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: