Connect, Message, Like,Follow&Share, 100% Free Counselling➔ThankYou

Life Cycle of Data Science projects

To understand how the life cycle of data science projects works so that it's easier for you to implement your individual projects in a similar pattern. The step-by-step implementation process of any data science project in a real-world scenario.

What is a Data Science Project Lifecycle?

In simple terms, a data science life cycle is nothing but a repetitive set of steps that you need to take to complete and deliver a project/product to your client. Although the data science projects and the teams involved in deploying and developing the model will be different, every data science life cycle will be slightly different in every other company. However, most of the data science projects happen to follow a somewhat similar process.

In order to start and complete a data science-based project, we need to understand the various roles and responsibilities of the people involved in building, developing the project. Let us take a look at those employees who are involved in a typical data science project:

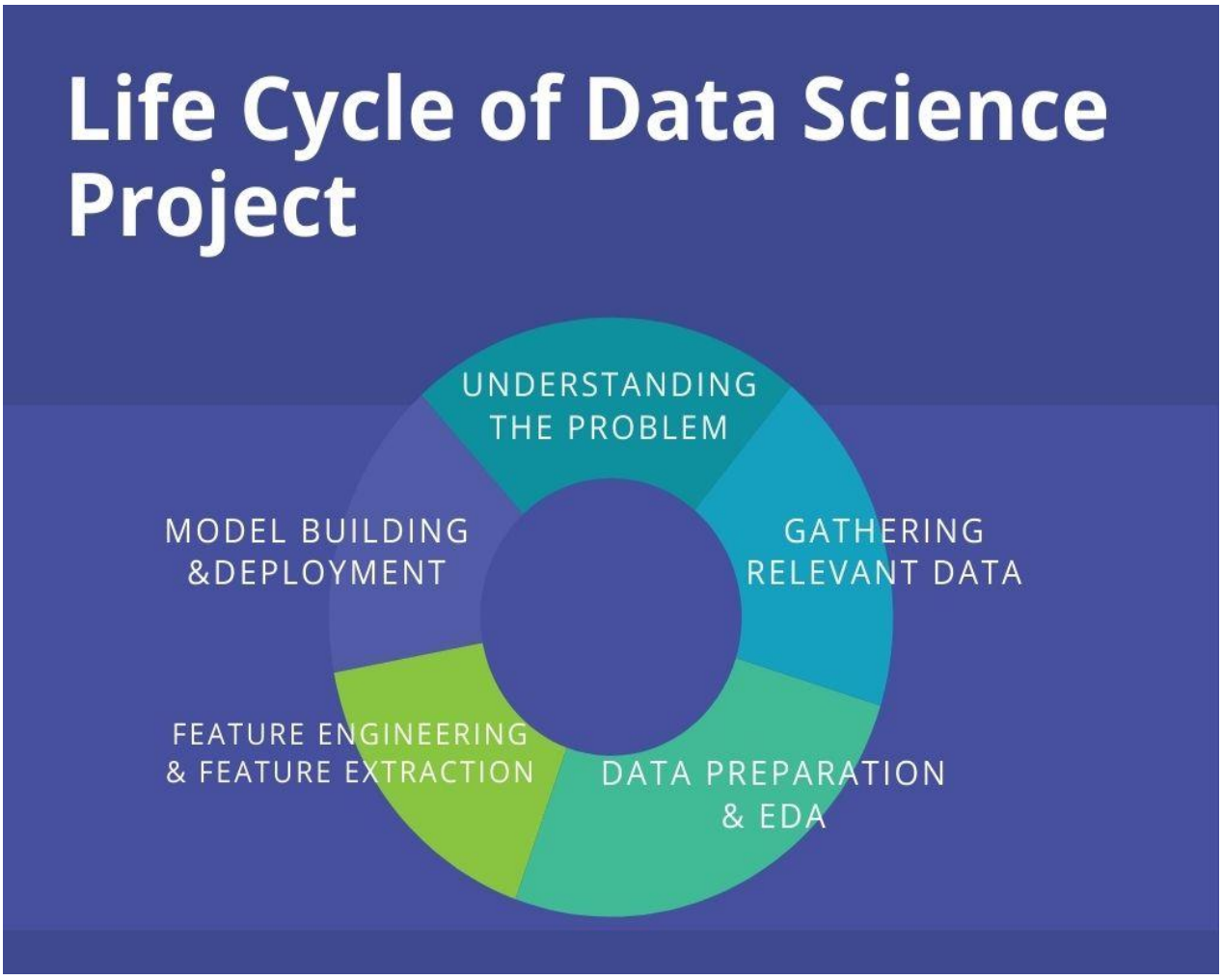
Who Are Involved in The Projects:

1. **Business Analyst**
2. **Data Analyst**

3. **Data Scientists**
4. **Data Engineer**
5. **Data Architect**
6. **Machine Learning Engineer**

Now that we have an idea of who all are involved in a typical business project, let's understand what is a data science project and how do we define the life cycle of the data science project in a real-world scenario like a fake news identifier.

Why do we need to define the Life Cycle of a data science project?



In a normal case, a Data Science project contains data as its main element. Without any data, we won't be able to do any analysis or predict any outcome

as we are looking at something unknown. Hence, before starting any data science project that we have got from either our clients or stakeholder first we need to understand the underlying problem statement presented by them. Once we understand the business problem, we have to gather the relevant data that will help us in solving the use case. However, for beginners many questions arise like:

In what format do we need the data?

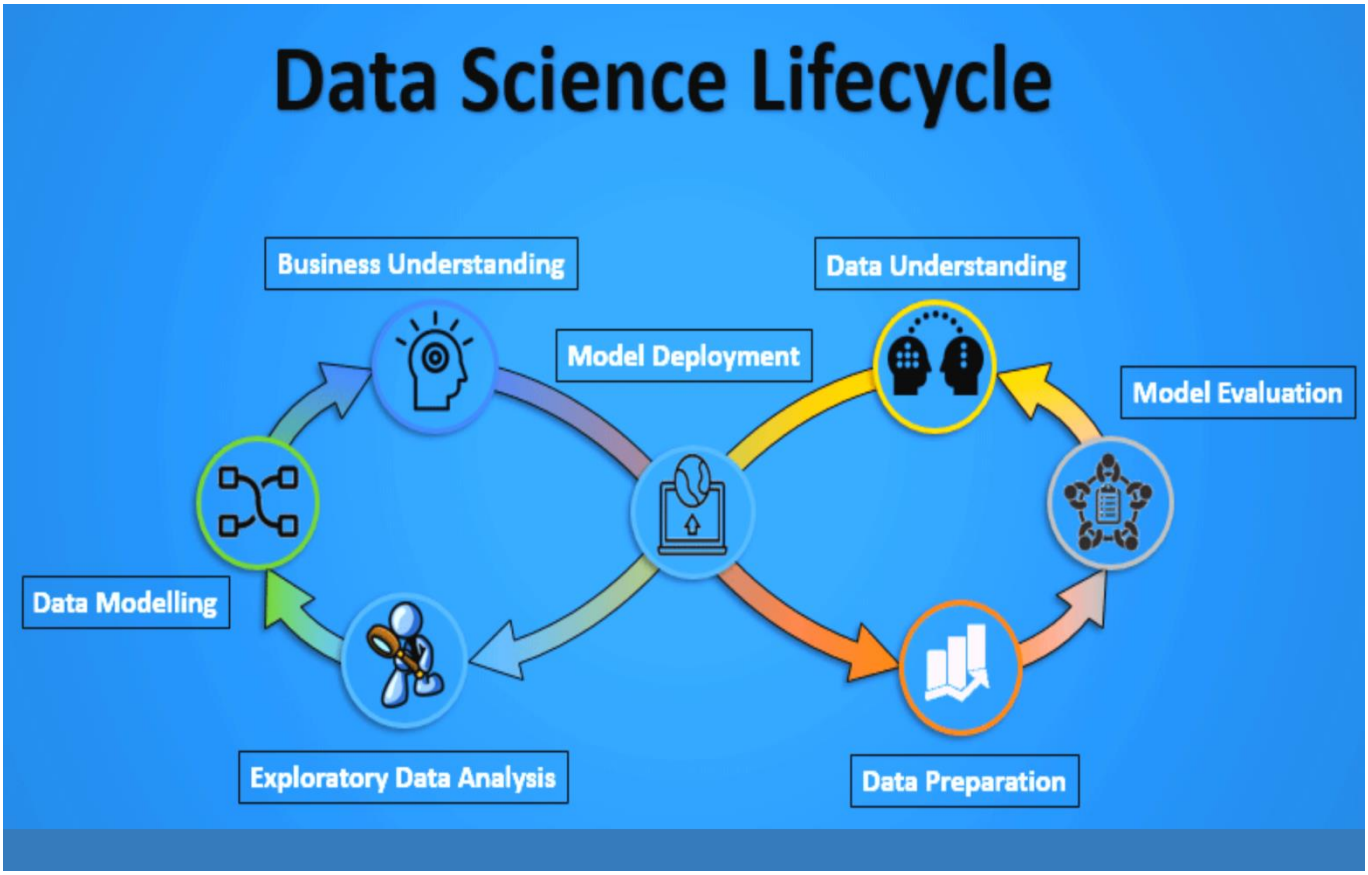
How to get the data?

What do we need to do with data?

So many questions yet answers might vary from person to person. Hence in order to address all these concerns right away, we do have a pre-defined flow that is termed as Data Science Project Life Cycle. The process is fairly simple wherein the company has to first gather data, perform data cleaning, perform EDA to extract relevant features, preparing the data by performing feature engineering and feature scaling. In the second phase, the model is built and deployed after a proper evaluation. This entire lifecycle is not a one man's job, for this, you need the entire team to work together to get the work done by achieving the required amount of efficiency for the project

The globally accepted structure in resolving any sort of analytical problem is popularly known as Cross Industry Standard Process for Data Mining or abbreviated as CRISP-DM framework.

Life Cycle of a Typical Data Science Project Explained:



1) Understanding the Business Problem:

In order to build a successful business model, its very important to first understand the business problem that the client is facing. Suppose he wants to predict the customer churn rate of his retail business. You may first want to understand his business, his requirements and what he is actually wanting to achieve from the prediction. In such cases, it is important to take consultation from domain experts and finally understand the underlying problems that are present in the system. A Business Analyst is generally responsible for gathering the required details from the client and forwarding the data to the data scientist team for further speculation. Even a minute error in defining the problem and understanding the requirement may be very crucial for the project hence it is to be done with maximum precision.

After asking the required questions to the company stakeholders or clients, we move to the next process which is known as data collection.

2) Data Collection

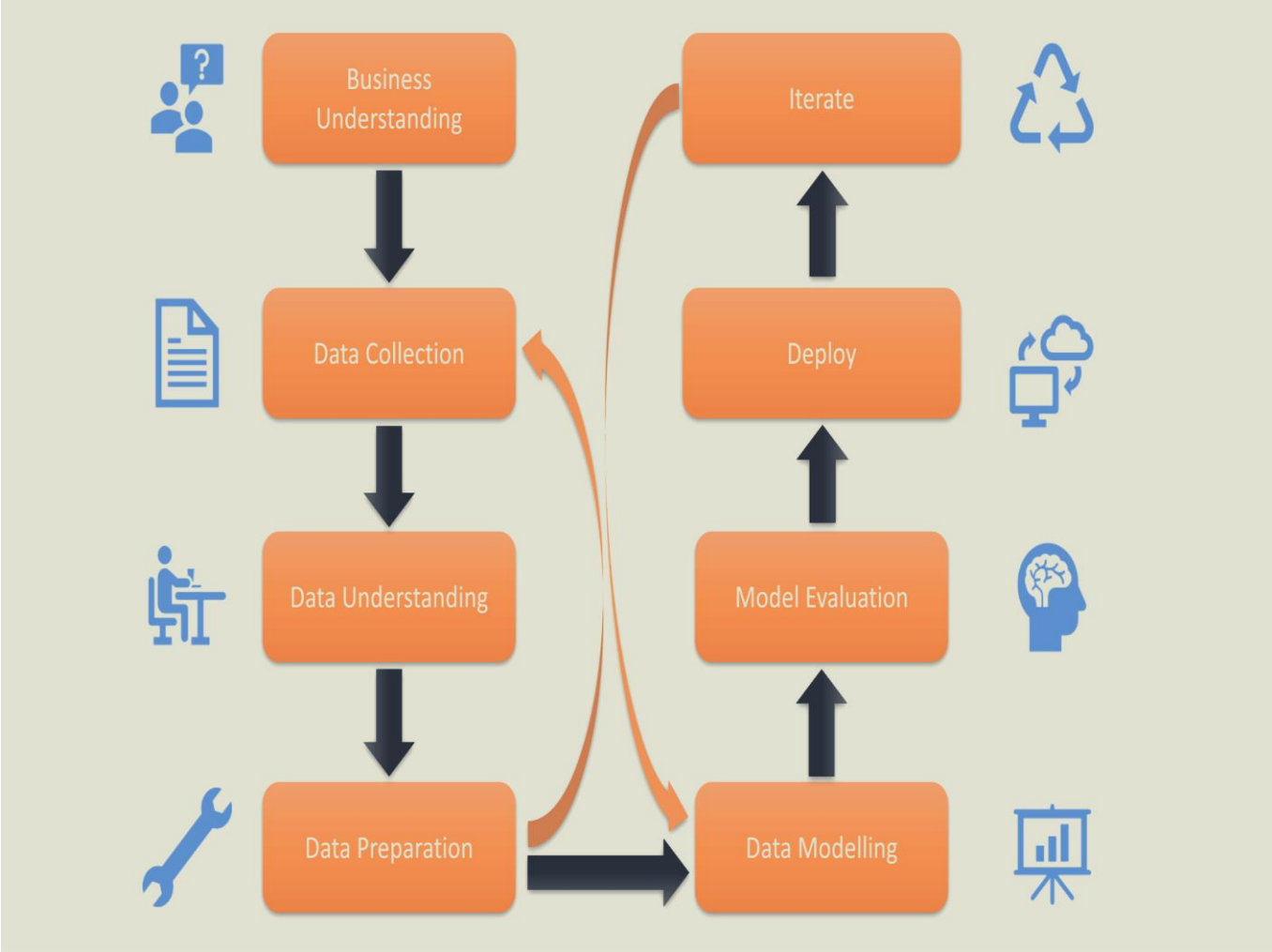
After gaining clarity on the problem statement, we need to collect relevant data to break the problem into small components.

The data science project starts with the identification of various data sources, which may include web server logs, social media posts, data from digital libraries such as the US Census datasets, data accessed through sources on the internet via APIs, web scraping, or information that is already present in an excel spreadsheet. Data collection entails obtaining information from both known internal and external sources that can assist in addressing the business issue.

Normally, the data analyst team is responsible for gathering the data. They need to figure out proper ways to source data and collect the same to get the desired results.

There are two ways to source the data:

1. Through web scraping with Python
2. Extracting Data with the use of third party APIs



3) Data Preparation

After gathering the data from relevant sources we need to move forward to data preparation. This stage helps us gain a better understanding of the data and prepares it for further evaluation.

Additionally, this stage is referred to as Data Cleaning or Data Wrangling. It entails steps such as selecting relevant data, combining it by mixing data sets, cleaning it, dealing with missing values by either removing them or imputing them with relevant data, dealing with incorrect data by removing it, and also checking for and dealing with outliers. By using feature engineering, you can create new data and extract new features from existing ones. Format the data

according to the desired structure and delete any unnecessary columns or functions. Data preparation is the most time-consuming process, accounting for up to 90% of the total project duration, and this is the most crucial step throughout the entire life cycle.

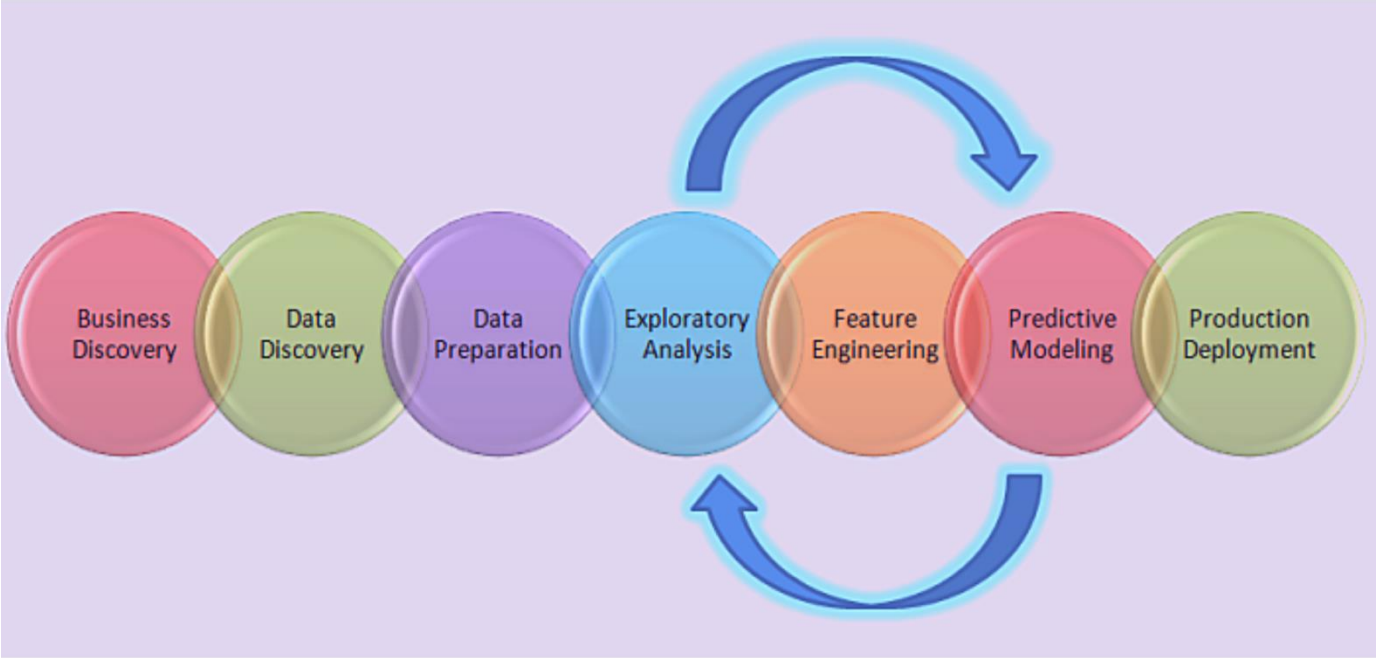
Exploratory Data Analysis (EDA) is critical at this point because summarising clean data enables the identification of the data's structure, outliers, anomalies, and trends. These insights can aid in identifying the optimal set of features, an algorithm to use for model creation, and model construction.

4) Data Modeling

Throughout most cases of data analysis, data modeling is regarded as the core process. In this process of data modeling, we take the prepared data as the input and with this, we try to prepare the desired output.

We first tend to select the appropriate type of model that would be implemented to acquire results, whether the problem is a regression problem or classification, or a clustering-based problem. Depending on the type of data received we happen to choose the appropriate machine learning algorithm that is best suited for the model. Once this is done, we ought to tune the hyperparameters of the chosen models to get a favorable outcome.

Finally, we tend to evaluate the model by testing the accuracy and relevance. In addition to this project, we need to make sure there is a correct balance between specificity and generalizability, which is the created model must be unbiased.



5) Model Deployment

Before the model is deployed, we need to ensure that we have picked the right solution after a rigorous evaluation has been. Later on, it is then deployed in the desired channel and format. This is naturally the last step in the life cycle of data science projects. Please take extra caution before executing each step in the life cycle to avoid unwanted errors. For example, if you choose the wrong machine learning algorithm for data modeling then you will not achieve the desired accuracy and it will be difficult in getting approval for the project from the stakeholders. If your data is not cleaned properly, you will have to handle missing values or the noise present in the dataset later on. Hence, in order to make sure that the model is deployed properly and accepted in the real world as an optimal use case, you will have to do rigorous testing in every step.

All the steps mentioned above are equally applicable for beginners as well as seasoned data science practitioners. As a beginner, your job is to learn the process first, then you need to practice and deploy smaller projects like fake news detector, titanic dataset, etc. You can refer to portals like, kaggle.com, hackerearth.com to get the dataset and start working on it.

Luckily for beginners, these portals have already cleaned most of the data, and hence proceeding with the next steps will be fairly easy. However, in the real world, you have to acquire not just any data set but the data that might meet the requirements of your data science project. Hence, initially, your job is to first proceed with all the steps of the data science life cycle very sincerely, and once you are thorough with the process and deployment you are ready to take the next step towards a career in this field. Python and R are the two languages that are most widely used in data science use cases.

Nowadays, even Julia is becoming one of the preferred languages for deploying the model. However, along with the clarity in the process, you should be comfortable in coding via such languages. From process understanding to proficiency in the programming language, you need to be adept with all.