Data Science Lifecycle

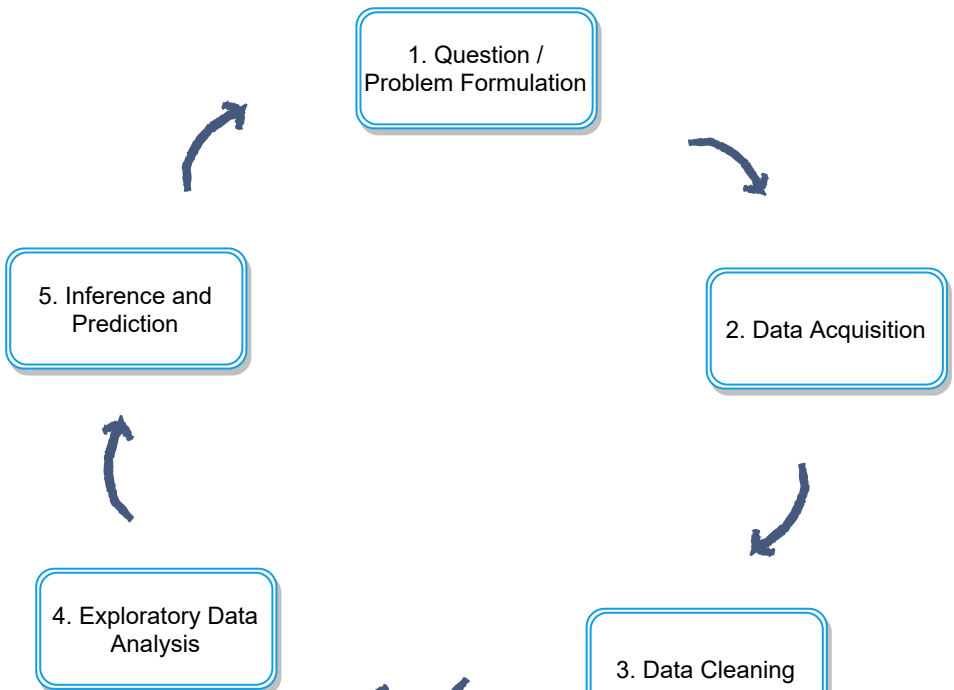
This lesson will introduce learners to the Data Science lifecycle, i.e., the process of extracting insights from data.

WE'LL COVER THE FOLLOWING

We know that the objective of Data Science as a discipline is to extract insights and meaning from data. To achieve this goal, data scientists follow a process that is known as the **Data Science Lifecycle.**

The Data Science lifecycle

The **Data Science lifecycle** involves 5 steps as shown in the following figure:



1. Problem formulation

The lifecycle starts with a question or a problem that we face. This can be a business question or a genuine curiosity of finding the relationships between different events. For instance, Data Science has been previously used for:

- Predicting and catching fraud
- Matching organ donors to patients
- Optimal staff scheduling
- Churn prediction
- Analyzing the performance in sports
- Increasing sales for businesses.

2. Data acquisition

Once the problem is identified, the next step is to gather data. This requires answering some of these questions:

- What kind of data do we need for our problem?
- Do we have any data already?
- From what sources we will collect data?
- How will we manage data during and after gathering?

3. Data cleaning

This is a crucial step in the lifecycle. Almost all the data that we gather is untidy (contains heterogeneous values, missing values, or large errors) and full of inconsistencies. Or we may have unnecessary data that we do not need. This step takes a lot of time in the lifecycle.

4. Exploratory data analysis

This step is where we really get to know our data. During exploratory data analysis we find the relationships and biases in the data. This includes visualizations as well. Visualizations involve producing images that communicate relationships among the represented data.

5. Inference and prediction

This is where all of the statistics and machine learning comes into play. We infer from the data and make predictive models that help us in decision making.

These steps keep repeating since it is a lifecycle. All of the lifecycle from step 2 is done using different tools like *Excel*, *R*, and *Python*. In the next lesson, we will look at which tool is best for Data Science.