# What is DATA SCIENCE?

Have you ever wonder from where this question arises?

***"Everybody loves a data scientist"*** wrote Simon Rogers (2012) in the *Guardian*. Mr. Rogers also traced the newfound love for number crunching to a quote Google's Hal Varian, who declared that *"the sexy job in the next ten years will be statisticians."* And it was widely believed that what he really meant were data scientists. This raises several important questions:

- What is data Science?
- How does it differs from statistics?
- What make someone a data scientist?

Even the prestigious *Harvard Business Review* called data science ***"the sexiest job of 21st century."***

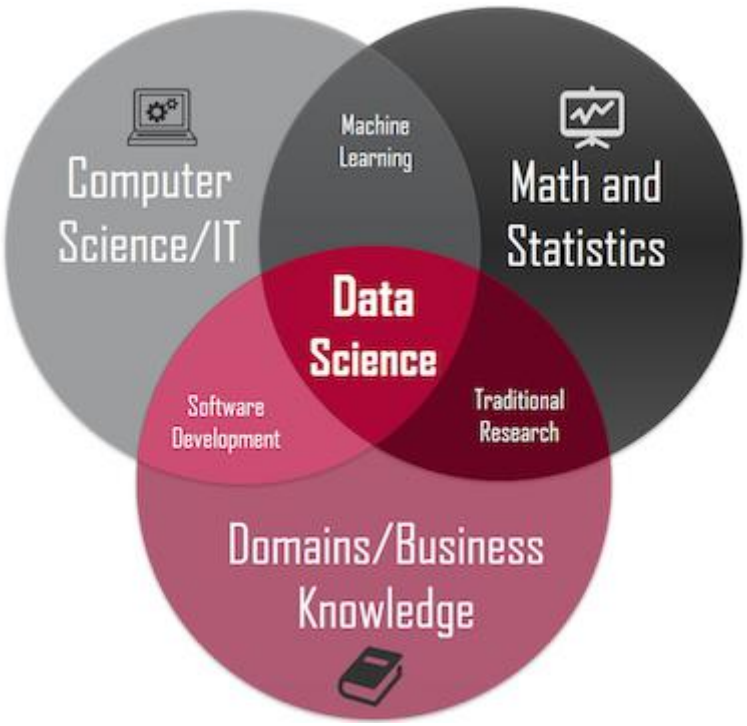
So, I have to define data science in the simplest language then I say that ***"Data Science is something that Data Scientists do."***

Everyone has a different view and perspective that who data Scientists are and let's have a look on some of them:

- Dr Vincent Granville defined a data scientist as one who can ***"easily process 50 million row data set in a couple of hours"*** and who distrusts (statistical) models. He distinguishes data science from statistics. Yet he lists algebra, calculus and training in probability and statistics as necessary background to understand data science.
- Dr Rachael Schutt defined a data scientist as someone who is part computer scientist, part software engineer and part statistician. But that's the definition of an average data scientist. ***"The Best"*** she contended, ***"tend to be really curious people, thinkers who ask good questions and are O.K. dealing with unstructured situations and trying to find structure in them."***
- Dr Patil told the *Guardian* newspaper in 2012 that a ***"data scientist is that unique blend of skills that can both unlock the insights of data and tell a fantastic story via the data."*** What is admirable about Dr Patil's definition is that it is inclusive of individuals of various academic backgrounds and training, and does not restrict the definition of a data science to a particular tool or subject it to a certain arbitrary threshold of data size.

We have heard differing views on it. So, what actually is Data Science?

The term *Data Science* has been used interchangeably for statistics, analytics, business analytics, and business intelligence. **Data Science** is simply the study of data. It is a multi-disciplinary field that aims to extract knowledge and insights from structured and unstructured data. It makes use of statistical and mathematical methods, computer science tools, and knowledge from related fields in order to draw conclusions from the data and make decisions and predictions based on these conclusions.




**Data Science vs. Business Intelligence**

Data Science and Business Intelligence (BI) are often confused with each other. While both focus on gaining insights from the data, they differ in the kind of questions they answer. BI helps monitor the current state of business data to understand the historical performance of a business. BI is mainly used for reporting or descriptive analysis while Data Science is used for both descriptive and predictive analysis.

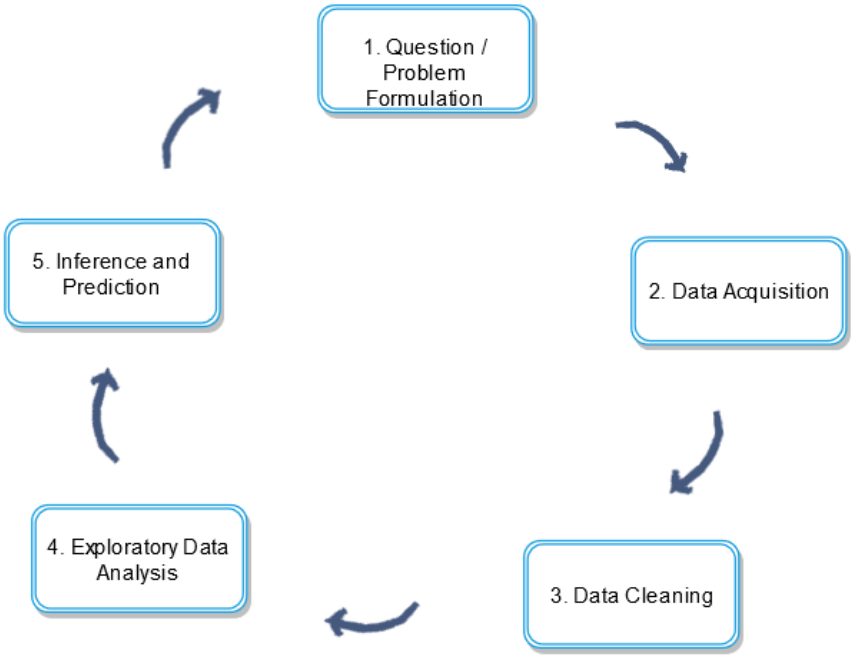
BI provides new information on previously known things, using some formula that is available. Data Science works with the unknown, answering data questions that nobody has answered before, without formula in hand.

## Why everyone should learn Data Science?

For professionals who are not statisticians and programmers, they have the domain/business knowledge, but it can become a little hard for them to tap into the data and generate meaningful insights from it. Therefore, everyone should learn how to effectively use the vast amounts of data that we have for solving our business problems.

# Data Science Lifecycle

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.** Which involves the following steps:




## 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.