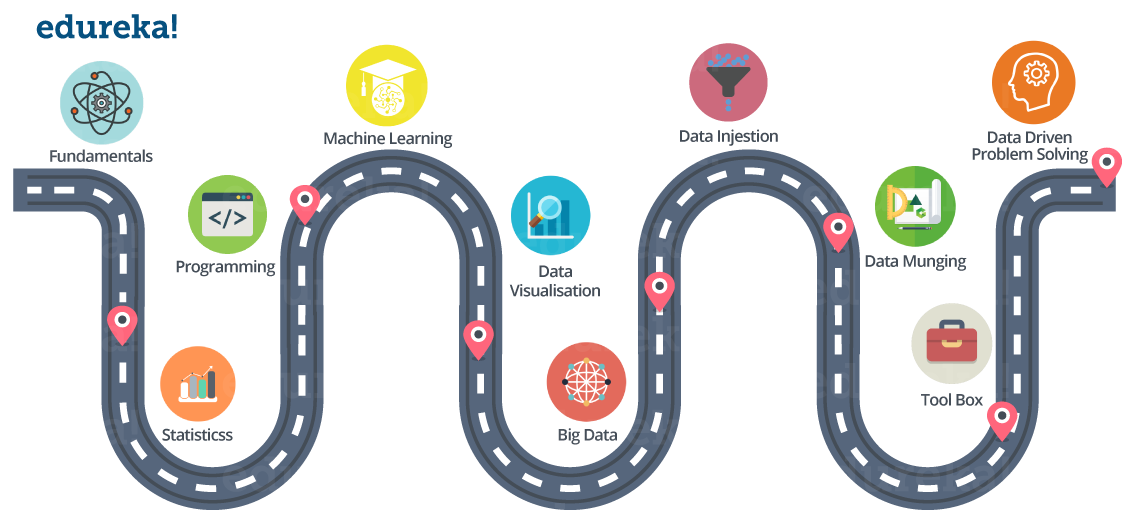
**Requirement for Data Scientist**

Skills required for Data Scientist:

1. Fundamentals
2. Statistics
3. Programming
4. Machine Learning and Advanced Machine Learning (Deep Learning)
5. Data Visualization
6. Big Data
7. Data Ingestion
8. Data Munging
9. Tool Box
10. Data-Driven Problem Solving



## Fundamentals:

This includes:

* Matrices and Linear Algebra Functions
* Hash Functions and Binary Tree
* Relational Algebra, Database Basics
* ETL ( Extract Transform Load )
* Reporting VS BI (Business Intelligence) VS Analytics

## Statistics:

This includes:

* Descriptive Statistics (Mean, Median, Range, Standard Deviation, Variance)
* Exploratory Data Analysis
* Percentiles and Outliers
* Probability Theory
* Bayes Theorem
* Random Variables
* Cumulative Distribution function (CDF)
* Skewness
* Other Statistics fundamentals

I would suggest you to pick a dataset from UCI repo. and start right now!

## Programming:

Expertise in any one programming language, I would suggest ‘**R**’ or ‘**Python**.

## Machine Learning and Advanced Machine Learning (Deep Learning):

Understand what is Machine learning and how it works.

Understand different types of Machine Learning techniques:

* Supervised Learning
* Unsupervised Learning
* Reinforcement Learning

Good knowledge on various Supervised and Unsupervised learning algorithms is required such as:

* Linear Regression
* Logistic Regression
* Decision Tree
* Random Forest
* K Nearest Neighbor
* Clustering (for example K-means)

Nowadays everyone is talking about Deep Learning, as it solved a lot of limitations of traditional Machine Learning approaches. Have to understand how Deep Learning works. I have listed down few Deep Learning concepts that you should be familiar with:

* Fundamentals of Neural Networks
* Any one library used for creating Deep Learning models, such as Tensorflow or Keras.
* Understand how Convolutional Neural Networks, Recurrent Neural Networks and RBM and Autoencoders work.

## Data Visualization:

Data visualization is a very important part of Data life-cycle.

Good hands-on knowledge is required on various visualization tools. Even, you can use a programming language for that purpose.

Below are few visualization tools:

* Tableau
* Kibana
* Google Charts
* Datawrapper

## Big Data:

[Big Data](https://www.edureka.co/blog/big-data-tutorial) is everywhere and there is almost an urgent need to collect and preserve whatever data is being generated, for the fear of missing out on something important.

There is a huge amount of data floating around. What we do with it is all that matters right now. This is why Big Data Analytics is in the frontiers of IT. Big Data Analytics has become crucial as it aids in improving business, decision makings and providing the biggest edge over the competitors. This applies for organizations as well as professionals in the Analytics domain.

As a Data Scientist it is very important to have knowledge about frameworks that can process Big Data. Two of the most famous ones are ‘Hadoop’ and ‘Spark’.

## Data Ingestion:

The process of importing , transferring , loading and processing data for later use or storage in a database is called Data Ingestion. This involves loading data from a variety of sources.

Below are few Data Ingestion tools:

* Apache Flume
* Apache Sqoop

## Data Munging:

If you have ever performed data analysis, you might have come across feature selection before you apply your Analytical model to the data.

So, in general, all the activity that you do on the raw data to make it “clean” enough to input to your analytical algorithm is data munging.

You can use ‘R’ and ‘Python’ packages for that.

It is one of the most important part of the data life-cycle.

As a Data Scientist you should be able to understand what all features are important in the dataset and what all features can be removed. You should also be able to identify your dependent variable or label.

Obviously, you have to remove inconsistency in the dataset.

All of these things are part of Data Munging (Data Wrangling).

## Tool Box:

You might find this section pretty redundant, but I think it is very very important to have good knowledge on certain tools like:

* MS Excel
* Python or R
* [Hadoop](https://www.edureka.co/blog/hadoop-tutorial/)
* [Spark](https://www.edureka.co/blog/spark-tutorial/)
* Tableau

## Data-Driven Problem Solving:

All the things we have discussed so far, includes tools and technologies that you can learn. But, Data-Driven problem solving approach is something that you need to develop. It will only come with experience.

A Data Scientist needs to know how to productively approach a problem.

This means identifying a situation’s

* salient features,
* figuring out how to frame a question that will yield the desired answer,
* deciding what approximations make sense, and
* consulting the right co-workers at the appropriate junctures of the analytic process.