Cheat Sheet of Data Science

Author, Meghna Pant

A Data Science Foundation Blog

July 2019

----

www.datascience.foundation

The “Big data” is in vogue these days. Most of the people who are aware of the term state that big data is a source of power and can bring about drastic revolutions in the scores of some of the major industrial sectors. However, the tools available that bring about changes in big businesses and small leading to the Big Data Revolution are known to a very few. Here you can get a sneak-peek of the tools that are available and how the tools fit into a broad spectrum of data science. Read on to know more about the Cheat Sheet.

Things you need to know before you get started with Data Science:

The term ‘big data’ mainly refers to data content with high volume, variety, and velocity. The glitch is, traditional database technologies do not go tandem with the handling of big data. Therefore, the introduction of innovative engineered solutions is in huge demand that can handle the big data efficiently.

How can you identify whether the project that you handling can be termed as big data or not? Here are some of the criteria to consider:

- **Volume** of the data should range in between 1 terabytes/year to 10 petabytes/ year.
- **Variety** mainly refers to the combination of data that is structured, unstructured, and semi-structured.
- **Velocity** of the data should range between 30 kilobytes/second to 30 gigabytes/second.

Difference between data engineering and data science:

More often than not when the managers are up for hiring they confuse the data engineers with data scientists. Though it is possible to find an expert who knows both the subjects; each data science and data engineering are vast topics on their own. To find someone with robust skills and impeccable experience in both sectors is just next to impossible.

Therefore, before hiring you must understand your goal and hire the appropriate one to get your work done.

- **Data Scientists** are the ones who use quantitative methods and coding to derive solutions for businesses and complex scientific problems.
- **Data Engineers** make use of their engineering skills to device programs that can derive solutions, handle and manipulate big sets of data.

Difference between business intelligence and data science:

Business analyst and the data scientists who work on solving the problems of complex businesses are considered to be closely related. Though they both use big data to achieve the desired goals of a business, their way of deducing and deriving inferences varies.

- **Data Scientist who is Business-centric:** The data scientists make use of data sets that are provided from the within the organization as well as from external source. The skill-sets and the common tools used in inferring data includes analytical platforms that are cloud-based, mathematical and statistical programming, Python and R language is used in data analysis, and also machine learning is taken into consideration.

From a vast amount of provided data, statistical and mathematical calculations are done generate predictions and analyze the situation of the business.

- **Business Intelligence (BI):** The data sets that are generated from within the organization are taken into consideration by the business intelligence team to deduce and analyze the situation of the business. The technologies and common tools taken into consideration are analytical processing that is done on an online platform, data warehousing, and load and extract transform of data is taken into consideration.

Understanding Machine Learning, the mathematical methods that are used in Data Science, and the Basics of Statistics:

Although we are aware that statistics is the most important tool in deducing inferences from data, we need to understand the difference between a data scientist and a statistician. While Data Science demands from the scientist to have the basic knowledge of statistics, the scope of data science is a lot more than just statistics. Let us dig in deeper to know the core difference.

- **Expertise in subject matter:** The data scientists are experts with a sophisticated degree in whichever field they are using their analytical method in. The data scientists need to be experts in a particular field to understand the critical complications of the matter and understand the applications of the data insights before they deduce solutions.

  The data scientist should be well aware of the subject and should be able to identify the importance of the findings and when necessary take decisions and proceed with their analysis independently.

  On the other hand, the statistician has an advanced degree of knowledge but are very little aware of the subject in which they are to apply the statistical methods and deduce solutions. The statisticians need to consult with experts on the subject to derive a solution.

- **Machine learning and mathematical approach:** We are now aware of the fact that statisticians make use of statistics only while deriving any insights out of a given data set. While the data scientists pull various techniques to reach to a definite solution for the problem. The techniques not only involve statistics but also make use of clustering, mathematics, non-statistical machine learning, and classification approach.

Why is having the basic knowledge in statistics important?

A data scientist does not need to have a sophisticated degree in statistics in order to practice data science. However, some of the basic statistical tools are needed to be used for data analysis. Some of these tools are:

- **Time-series analysis:** Analysis of a collection of data by making use of time attributes is often referred to as time-series analysis. This is used to predict the instances that are to take place in the future based on the incidents that have occurred in the past.
- **Linear regression:** A relationship between a dependent variable and other independent variables can be deduced with the help of linear regression.
- **Monte Carlo simulations:** A simulation technique that is primarily used to generate estimated parameters, predict the outcomes of a particular scenario, test various hypotheses, and validate certain models.
- **Statistics used for spatial data:** The most important aspect of spatial data is –it is not random. It is auto-correlated and spatially dependent. Often when spatial data is being used, random data is avoided with the use of statistical methods.

What are clustering, machine learning, and classification?

Machine learning involves the deduction of patterns by making use of computational algorithms on the raw data that is provided.

Clustering is a special type of unsupervised machine learning, in which computational algorithm is used on unlabeled data and inferential methods are made use of to find out the correlations.

Classification is supervised machine learning, in which computational algorithm is used on labelled data.

Use of mathematical methods in Data science:

Data science certainly involves the use of statistical analysis; however, mathematical methods are often underrated. We are unaware of the fact that mathematics is the root of all the quantitative analysis that is to be done. The two mathematical methods used in data science are:

- **Markov Chains:** In this type of mathematical method a series of randomly generated variables is created, which is used to represent the present state. Further, the present state is used to define how the future states shall be affected due to the events taking place in the present state.
- **Multi-criteria decision making (MCDM):** When you have several criteria or various alternatives, whose evaluation is needed to be done simultaneously, the MCDM mathematical tool is taken into consideration.

How can Data Science make use of Visualization techniques?

If the information that has been deduced cannot be communicated, then the whole process is just a waste of time. Therefore, data scientists should master the skills of communicating their visions and insights to others. As a data scientist, you need to develop visualizations that can be easily understood by your audience. Also, you must keep in mind that these visualizations should be valuable and relevant for the businesses or stakeholders for whom you are doing the work.

Make use of your coding skills:

If you are web-programmer and have your basics about HTML, JavaScript, and CSS brushed up, you can make use of D3.js to build dynamic visualizations that are based on the web. The perfect solution for the problems related to visualization can be adequately solved with the use of D3.js to design interactive web-based visualization of the data inferences.

Making use of Geographical Information Systems in Data Science

The Geographical Information system can be extensively used in data science. When location-based trends are needed to be discovered and calculated, the GIS can be of great use. You can make use of maps to generate spatial visualization of data with the help of GIS. However, there are other forms of advanced data analysis visualization methods that can be generated with the help of GIS software. The two most popular GIS software are –QGIS and ArcGIS for Desktop.

Programming Languages that can be used in Data Science

One of the most important skills that a data scientist needs to master is coding. Though various powerful applications can be used without having a vast knowledge of coding; custom-analysis and visualization, which are the prime parts of data science contexts cannot be dealt with, without possessing adequate knowledge of the programming language. For advanced tasks of analysis, a data scientist needs to code things for themselves with the help of R programming language or Python Programming language.

Using Python Language for Data Science

Python is a programming language which is easy to learn and is readable by humans as well. It can be used for advanced data analysis, munging, and visualization. The software for learning python language can be installed easily and it is rather easier than R language to be learned. The language runs on UNIX, Mac, and Windows as well.

The people who are not very fond of the command line, IPython works well in their favour as it provides a user-friendly atmosphere for coding.

Using R for Data Science

R is one of the very famous programming languages, which is widely used in scientific computing and statistical analysis as well. R Scripting is often the term given to visualization routines and writing analysis done with the help of the R language. Though the language is comparatively difficult than Python to be learned, it has to offer a plentiful of statistical computing packages.

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email: admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation