CHAPTER 8

What Is Data Science?

Data science encompasses the tools and techniques for extracting information from data. Data science techniques draw extensively from the field of mathematics, statistics, and computation. However, data science is now encapsulated into software packages and libraries, thus making them easily accessible and consumable by the software development and engineering communities. This is a major factor to the rise of intelligence capabilities now integrated as a major staple in software products across all sorts of domains.

This chapter will discuss broadly on the opportunities for data science and big data analytics integration as part of the transformation portfolio of businesses and institutions and give an overview on the data science process as a reusable template for fulfilling data science projects.

The Challenge of Big Data

Due to the expansion of data at the turn of the twenty-first century epitomized by the so-called 3Vs of big data, which are volume, velocity, and variety. Volume refers to the increasing size of data, velocity the speed at which data is acquired, and variety the diverse types of data that are available. For others, this becomes 5Vs with the inclusion of value and veracity to mean the usefulness of data and the truthfulness of data, respectively. We have observed data volume blowout from the megabyte (MB) to the terabyte (TB) scale and now exploding past the petabyte (PB). We have to find new and improved means of storing and processing this ever-increasing dataset. Initially, this challenge of storage and data processing was addressed by the Hadoop ecosystem and other supporting frameworks, but even these have become expensive to manage and scale, and this is why there is a pivot to cloud-managed, elastic, secure, and high-availability data storage and processing capabilities.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_8

On the other hand, for most applications and business use cases, there is a need to carry out real-time analysis on data due to the vast amount of data created and available at a given moment. Previously, getting insights from data and unlocking value had been down to traditional analysis on batch data workloads using statistical tools such as Excel, Minitab, or SPSS. But in the era of big data, this is changing, as more and more businesses and institutions want to understand the information in their data at a real-time or at worst near real-time pace.

Another vertical to the big data conundrum is that of variety. Formerly, a pre-defined structure had to be imposed on data in order to easily store them as well as make it easy for data analysis. However, a wide diversity of datasets are now collected and stored such as spatial maps, image data, video data, audio data, text data from emails and other documents, and sensor data. As a matter of fact, a far larger amount of datasets in the wild are unstructured. This led to the development of unstructured or semi-structured databases such as Elasticsearch, Solr, HBase, Cassandra, and MongoDB, to mention just a few.

The Data Science Opportunity

In the new age, where data has inevitably and irreversibly become the new gold, the greatest needs of organizations are the skills required for data governance and analytics to unlock intelligence and value from data as well as the expertise to develop and productionize enterprise data products. This has led to new roles within the data science umbrella such as

- Data analysts/scientist who specialize in mining intelligence from data using statistical techniques and computational tools by understanding the business use case
- Data engineers/architects who specialize in architecting and managing the infrastructure for efficient big data pipelines by ensuring that the data platform is redundant, scalable, secure, and highly available
- Machine learning engineers who specialize in designing and developing machine learning algorithms as well as incorporating them into production systems for online or batch prediction services