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

The Data Science Process

The data science process involves components for data ingestion and serving of data models. However, we will discuss briefly on the steps for carrying out data analytics in lieu of data prediction modeling.

These steps consist of

1. Data summaries: The vital statistical summaries of the datasets' variables or features. This includes information such as the number of variables, their data types, the number of observations, and the count/percentage of missing data.
2. Data visualization: This involves employing univariate and multivariate data visualization methods to get a better intuition on the properties of the data variables and their relationship with each other. This includes metrics such as histograms, box and whisker plots, and correlation plots.
3. Data cleaning/preprocessing: This process involves sanitizing the data to make it amenable for modeling. Data rarely comes clean with each row representing an observation and each column an entity. In this phase of a data science effort, the tasks involved may include removing duplicate entries, choosing a strategy for dealing with missing data, as well as converting data features into numeric data types of encoded categories. This phase may also involve carrying out statistical transformation on the data features to normalize and/or standardize the data elements. Data features of wildly differing scales can lead to poor model results as they become more difficult for the learning algorithm to converge to the global minimum.
4. Feature engineering: This practice involves systematically pruning the data feature space to only select those features relevant to the modeling problem as part of the model task. Good feature engineering is often the difference between an average and high performant model.