## What is data science?

Data science combines math and statistics, specialized programming, advanced [analytics](https://www.ibm.com/consulting/analytics), [artificial intelligence (AI)](https://www.ibm.com/think/topics/artificial-intelligence) and [machine learning](https://www.ibm.com/think/topics/machine-learning) with specific subject matter expertise to uncover actionable insights hidden in an organization’s data. These insights can be used to guide decision making and strategic planning.

The accelerating volume of data sources, and subsequently data, has made data science is one of the fastest growing field across every industry. As a result, it is no surprise that the role of the data scientist was dubbed the “sexiest job of the 21st century” by [Harvard Business Review](https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century" \t "https://www.ibm.com/think/topics/_blank). Organizations are increasingly reliant on them to interpret data and provide actionable recommendations to improve business outcomes.

The data science lifecycle involves various roles, tools, and processes, which enables analysts to glean actionable insights. Typically, a data science project undergoes the following stages:

* ****Data ingestion****: The lifecycle begins with the data collection—both raw structured and unstructured data from all relevant sources using a variety of methods. These methods can include manual entry, web scraping, and real-time streaming data from systems and devices. Data sources can include structured data, such as customer data, along with unstructured data like log files, video, audio, pictures, the [Internet of Things (IoT)](https://www.ibm.com/think/topics/internet-of-things), social media, and more.
* ****Data storage and data processing****: Since data can have different formats and structures, companies need to consider different storage systems based on the type of data that needs to be captured. Data management teams help to set standards around data storage and structure, which facilitate workflows around analytics, machine learning and [deep learning](https://www.ibm.com/think/topics/deep-learning) models. This stage includes cleaning data, deduplicating, transforming and combining the data using [ETL](https://www.ibm.com/think/topics/etl) (extract, transform, load) jobs or other data integration technologies. This data preparation is essential for promoting data quality before loading into a [data warehouse](https://www.ibm.com/think/topics/data-warehouse), [data lake](https://www.ibm.com/think/topics/data-lake), or other repository.
* ****Data analysis****: Here, data scientists conduct an exploratory data analysis to examine biases, patterns, ranges, and distributions of values within the data. This data analytics exploration drives hypothesis generation for a/b testing. It also allows analysts to determine the data’s relevance for use within modeling efforts for predictive analytics, machine learning, and/or deep learning. Depending on a model’s accuracy, organizations can become reliant on these insights for business decision making, allowing them to drive more scalability.
* ****Communicate****: Finally, insights are presented as reports and other data visualizations that make the insights—and their impact on business—easier for business analysts and other decision-makers to understand. A data science programming language such as R or Python includes components for generating visualizations; alternately, data scientists can use dedicated visualization tools.