What is data science?

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI) and 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. 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), 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 models. This stage includes cleaning data, deduplicating, transforming and combining the data using 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, 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.

What data scientists do

Data scientists are experts at extracting industry-specific insights and answers from data. They possess computer science and pure science skills beyond those of a typical business analyst or data analyst, as well as a deep understanding of the specifics of the industry or business discipline in which they work (such as automobile manufacturing, eCommerce or healthcare).

A data scientist must be able to:

Know enough about the business to ask pertinent questions and identify business pain points.

Apply statistics and computer science, along with business acumen, to data analysis.

Use a wide range of tools and techniques for preparing and extracting data—everything from databases and SQL to data mining to data integration methods.

Extract insights from big data using predictive analytics and artificial intelligence (AI), including machine learning models, natural language processing, and deep learning.

Write programs and algorithms that automate data processing and calculations.

Tell—and illustrate—stories that clearly convey the meaning of results to decision-makers and stakeholders at every level of technical understanding.

Explain how the results can be used to solve business problems.

Collaborate with other data science team members, such as data and business analysts, IT architects, data engineers, and application developers.

These skills are in high demand, and as a result, many individuals that are breaking into a data science career, explore a variety of data science programs, such as certification programs, data science courses, and degree programs offered by educational institutions.

Data scientists are not necessarily directly responsible for all the processes involved in the data science lifecycle. For example, data pipelines are typically handled by data engineers—but the data scientist may make recommendations about what sort of data is useful or required. While data scientists can build machine learning models, scaling these efforts at a larger level requires more software engineering skills to optimize a program to run more quickly. As a result, it’s common for a data scientist to partner with machine learning engineers to scale machine learning models.

Data scientist responsibilities can commonly overlap with a data analyst, particularly with exploratory data analysis and data visualization. However, a data scientist’s skillset is typically broader than the average data analyst. Comparatively speaking, data scientist leverage common programming languages, such as R and Python, to conduct more statistical inference and data visualization.

Data science versus business intelligence

It may be easy to confuse the terms “data science” and “business intelligence” (BI) because they both relate to an organization’s data and analysis of that data, but they do differ in focus.

Business intelligence (BI) is typically an umbrella term for the technology that enables data preparation, data mining, data management, and data visualization. Business intelligence tools and processes allow end users to identify actionable information from raw data, facilitating data-driven decision-making within organizations across various industries. While data science tools overlap in much of this regard, business intelligence focuses more on data from the past, and the insights from BI tools are more descriptive in nature. It uses data to understand what happened before to inform a course of action. BI is geared toward static (unchanging) data that is usually structured. While data science uses descriptive data, it typically utilizes it to determine predictive variables, which are then used to categorize data or to make forecasts.

Data science and BI are not mutually exclusive—digitally savvy organizations use both to fully understand and extract value from their data.

Data science tools

Data scientists rely on popular programming languages to conduct exploratory data analysis and statistical regression. These open source tools support pre-built statistical modeling, machine learning, and graphics capabilities. These languages include the following (read more at "Python vs. R: What's the Difference?"):

R Studio: An open source programming language and environment for developing statistical computing and graphics.
Python: It is a dynamic and flexible programming language. The Python includes numerous libraries, such as NumPy, Pandas, Matplotlib, for analyzing data quickly.

To facilitate sharing code and other information, data scientists may use GitHub and Jupyter notebooks.

Some data scientists may prefer a user interface, and two common enterprise tools for statistical analysis include:

SAS: A comprehensive tool suite, including visualizations and interactive dashboards, for analyzing, reporting, data mining, and predictive modeling.
IBM SPSS: Offers advanced statistical analysis, a large library of machine learning algorithms, text analysis, open source extensibility, integration with big data, and seamless deployment into applications.

Data scientists also gain proficiency in using big data processing platforms, such as Apache Spark, the open source framework Apache Hadoop, and NoSQL databases. They are also skilled with a wide range of data visualization tools, including simple graphics tools included with business presentation and spreadsheet applications (like Microsoft Excel), built-for-purpose commercial visualization tools like Tableau and IBM Cognos, and open source tools like D3.js (a JavaScript library for creating interactive data visualizations) and RAW Graphs. For building machine learning models, data scientists frequently turn to several frameworks like PyTorch, TensorFlow, MXNet, and Spark MLib.

Given the steep learning curve in data science, many companies are seeking to accelerate their return on investment for AI projects; they often struggle to hire the talent needed to realize data science project's full potential. To address this gap, they are turning to multipersona data science and machine learning (DSML) platforms, giving rise to the role of "citizen data scientist."

Multipersona DSML platforms use automation, self-service portals, and low-code/no-code user interfaces so that people with little or no background in digital technology or expert data science can create business value using data science and machine learning. These platforms also support expert data scientists by also offering a more technical interface. Using a multipersona DSML platform encourages collaboration across the enterprise.

Data science and cloud computing

Cloud computing scales data science by providing access to additional processing power, storage, and other tools required for data science projects.

Since data science frequently leverages large data sets, tools that can scale with the size of the data is incredibly important, particularly for time-sensitive projects. Cloud storage solutions, such as data lakes, provide access to storage infrastructure, which are capable of ingesting and processing large volumes of data with ease. These storage systems provide flexibility to end users, allowing them to spin up large clusters as needed. They can also add incremental compute nodes to expedite data processing jobs, allowing the business to make short-term tradeoffs for a larger long-term outcome. Cloud platforms typically have different pricing models, such a per-use or subscriptions, to meet the needs of their end user—whether they are a large enterprise or a small startup.

Open source technologies are widely used in data science tool sets. When they're hosted in the cloud, teams don't need to install, configure, maintain, or update them locally. Several cloud providers, including IBM Cloud®, also offer prepackaged tool kits that enable data scientists to build models without coding, further democratizing access to technology innovations and data insights.

Data science use cases

Enterprises can unlock numerous benefits from data science. Common use cases include process optimization through intelligent automation and enhanced targeting and personalization to improve the customer experience (CX). However, more specific examples include:

Here are a few representative use cases for data science and artificial intelligence:

> An international bank delivers faster loan services with a mobile app using machine learning-powered credit risk models and a hybrid cloud computing architecture that is both powerful and secure.
>
> An electronics firm is developing ultra-powerful 3D-printed sensors to guide tomorrow's driverless vehicles. The solution relies on data science and analytics tools to enhance its real-time object detection capabilities.
>
> A robotic process automation (RPA) solution provider developed a cognitive business process mining solution that reduces incident handling times between 15% and 95% for its client companies. The solution is trained to understand the content and sentiment of customer emails, directing service teams to prioritize those that are most relevant and urgent.

A digital media technology company created an audience analytics platform that enables its clients to see what’s engaging TV audiences as they’re offered a growing range of digital channels. The solution employs deep analytics and machine learning to gather real-time insights into viewer behavior.

An urban police department created statistical incident analysis tools to help officers understand when and where to deploy resources in order to prevent crime. The data-driven solution creates reports and dashboards to augment situational awareness for field officers.

Shanghai Changjiang Science and Technology Development used IBM® Watson® technology to build an AI-based medical assessment platform that can analyze existing medical records to categorize patients based on their risk of experiencing a stroke and that can predict the success rate of different treatment plans.