What is data science?

Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

Why is data science important?

Data science is important because it combines tools, methods, and technology to generate meaning from data. Modern organizations are inundated with data; there is a proliferation of devices that can automatically collect and store information. Online systems and payment portals capture more data in the fields of e-commerce, medicine, finance, and every other aspect of human life. We have text, audio, video, and image data available in vast quantities.

History of data science

While the term data science is not new, the meanings and connotations have changed over time. The word first appeared in the ’60s as an alternative name for statistics. In the late ’90s, computer science professionals formalized the term. A proposed definition for data science saw it as a separate field with three aspects: data design, collection, and analysis. It still took another decade for the term to be used outside of academia.

Future of data science

Artificial intelligence and machine learning innovations have made data processing faster and more efficient. Industry demand has created an ecosystem of courses, degrees, and job positions within the field of data science. Because of the cross-functional skillset and expertise required, data science shows strong projected growth over the coming decades.

What is data science used for?

Data science is used to study data in four main ways:

data like the number of tickets booked each day. Descriptive analysis will reveal booking spikes, booking slumps, and high-performing months for this service.

2. Diagnostic analysis

Diagnostic analysis is a deep-dive or detailed data examination to understand why something happened. It is characterized by techniques such as drill-down, data discovery, data mining, and correlations. Multiple data operations and transformations may be performed on a given data set to discover unique patterns in each of these techniques.For example, the flight service might drill down on a particularly high-performing month to better understand the booking spike. This may lead to the discovery that many customers visit a particular city to attend a monthly sporting event.

3. Predictive analysis

Predictive analysis uses historical data to make accurate forecasts about data patterns that may occur in the future. It is characterized by techniques such as machine learning, forecasting, pattern matching, and predictive modeling. In each of these techniques, computers are trained to reverse engineer causality connections in the data.For example, the flight service team might use data science to predict flight booking patterns for the coming year at the start of each year. The computer program or algorithm may look at past data and predict booking spikes for certain destinations in May. Having anticipated their customer's future travel requirements, the company could start targeted advertising for those cities from February.

4. Prescriptive analysis

Prescriptive analytics takes predictive data to the next level. It not only predicts what is likely to happen but also suggests an optimum response to that outcome. It can analyze the potential implications of different choices and recommend the best course of action. It uses graph analysis, simulation, complex event processing, neural networks, and recommendation engines from machine learning.

Back to the flight booking example, prescriptive analysis could look at historical marketing campaigns to maximize the advantage of the upcoming booking spike. A data scientist could project booking outcomes for different levels of marketing spend on various marketing channels. These data forecasts would give the flight booking company greater confidence in their marketing decisions.

What are the benefits of data science for business?

Data science allows businesses to uncover new patterns and relationships that have the potential to transform the organization. It can reveal low-cost changes to resource management for maximum impact on profit margins.For example, an e-commerce company uses data science to discover that too many customer queries are being generated after business hours. Investigations reveal that customers are more likely to purchase if they receive a prompt response instead of an answer the next business day. By implementing 24/7 customer service, the business grows its revenue by 30%.

Innovate new products and solutions

Data science can reveal gaps and problems that would otherwise go unnoticed. Greater insight about purchase decisions, customer feedback, and business processes can drive innovation in internal operations and external solutions.For example, an online payment solution uses data science to collate and analyze customer comments about the company on social media. Analysis reveals that customers forget passwords during peak purchase periods and are unhappy with the current password retrieval system. The company can innovate a better solution and see a significant increase in customer satisfaction.

Real-time optimization

It's very challenging for businesses, especially large-scale enterprises, to respond to changing conditions in real-time. This can cause significant losses or disruptions in business activity. Data science can help companies predict change and react optimally to different circumstances.For example, a truck-based shipping company uses data science to reduce downtime when trucks break down. They identify the routes and shift patterns that lead to faster breakdowns and tweak truck schedules. They also set up an inventory of common spare parts that need frequent replacement so trucks can be repaired faster.

What is the data science process?

A business problem typically initiates the data science process. A data scientist will work with business stakeholders to understand what business needs. Once the problem has been defined, the data scientist may solve it using the OSEMN data science process:

O – Obtain data

Data can be pre-existing, newly acquired, or a data repository downloadable from the internet. Data scientists can extract data from internal or external databases, company CRM software, web server logs, social media or purchase it from trusted third-party sources.

- Changing all date values to a common standard format.·
- Fixing spelling mistakes or additional spaces.·
- Fixing mathematical inaccuracies or removing commas from large numbers.

E – Explore data

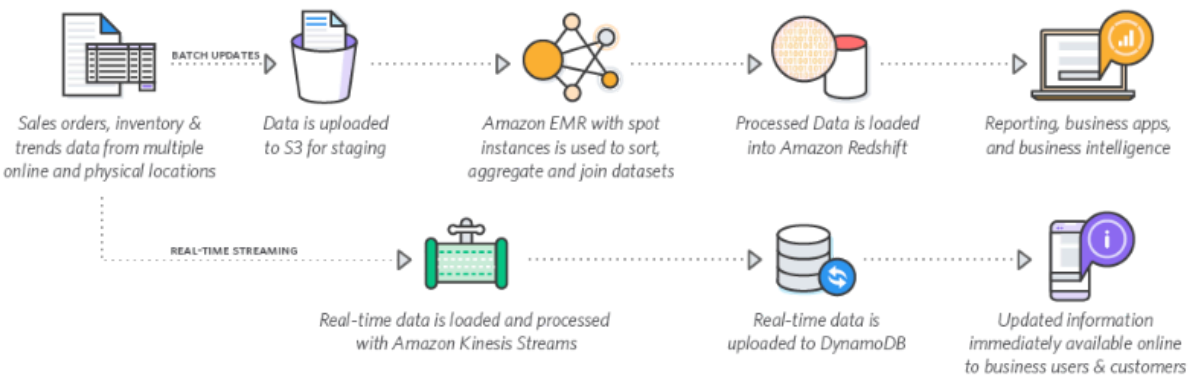
Data exploration is preliminary data analysis that is used for planning further data modeling strategies. Data scientists gain an initial understanding of the data using descriptive statistics and data visualization tools. Then they explore the data to identify interesting patterns that can be studied or actioned.

M – Model data

Software and machine learning algorithms are used to gain deeper insights, predict outcomes, and prescribe the best course of action. Machine learning techniques like association, classification, and clustering are applied to the training data set. The model might be tested against predetermined test data to assess result accuracy. The data model can be fine-tuned many times to improve result outcomes.

N – Interpret results

Data scientists work together with analysts and businesses to convert data insights into action. They make diagrams, graphs, and charts to represent trends and predictions. Data summarization helps stakeholders understand and implement results effectively.



What are the data science techniques?

Classification is the sorting of data into specific groups or categories. Computers are trained to identify and sort data. Known data sets are used to build decision algorithms in a computer that quickly processes and categorizes the data. For example:·

- Sort products as popular or not popular·
- Sort insurance applications as high risk or low risk·
- Sort social media comments into positive, negative, or neutral.

Data science professionals use computing systems to follow the data science process.

Regression

Regression is the method of finding a relationship between two seemingly unrelated data points. The connection is usually modeled around a mathematical formula and represented as a graph or curves. When the value of one data point is known, regression is used to predict the other data point. For example:·

- The rate of spread of air-borne diseases.·
- The relationship between customer satisfaction and the number of employees.·
- The relationship between the number of fire stations and the number of injuries due to fire in a particular location.

Clustering

Clustering is the method of grouping closely related data together to look for patterns and anomalies. Clustering is different from sorting because the data cannot be accurately classified into fixed categories. Hence the data is grouped into most likely relationships. New patterns and relationships can be discovered with clustering. For example: ·

- Group customers with similar purchase behavior for improved customer service.·
- Group network traffic to identify daily usage patterns and identify a network attack faster.
- Cluster articles into multiple different news categories and use this information to find fake news content.

The basic principle behind data science techniques

While the details vary, the underlying principles behind these techniques are:

- Teach a machine how to sort data based on a known data set. For example, sample keywords are given to the computer with their sort value. “Happy” is positive, while “Hate” is negative.

What are different data science technologies?

Data science practitioners work with complex technologies such as:

1. **Artificial intelligence:** Machine learning models and related software are used for predictive and prescriptive analysis.
2. **Cloud computing:** Cloud technologies have given data scientists the flexibility and processing power required for advanced data analytics.
3. **Internet of things:** IoT refers to various devices that can automatically connect to the internet. These devices collect data for data science initiatives. They generate massive data which can be used for data mining and data extraction.
4. **Quantum computing:** Quantum computers can perform complex calculations at high speed. Skilled data scientists use them for building complex quantitative algorithms.

How does data science compare to other related data fields?

Data science is an all-encompassing term for other data-related roles and fields. Let's look at some of them here:

What is the difference between data science and data analytics?

While the terms may be used interchangeably, data analytics is a subset of data science. Data science is an umbrella term for all aspects of data processing—from the collection to modeling to insights. On the other hand, data analytics is mainly concerned with statistics, mathematics, and statistical analysis. It focuses on only data analysis, while data science is related to the bigger picture around organizational data.In most workplaces, data scientists and data analysts work together towards common business goals. A data analyst may spend more time on routine analysis, providing regular reports. A data scientist may design the way data is stored, manipulated, and analyzed. Simply put, a data analyst makes sense out of existing data, whereas a data scientist creates new methods and tools to process data for use by analysts.

What is the difference between data science and business analytics?

While there is an overlap between data science and business analytics, the key difference is the use of technology in each field. Data scientists work more closely with data technology than business analysts.Business analysts bridge the gap between business and IT. They define business cases, collect information from stakeholders, or validate solutions. Data scientists, on the other hand, use technology to work with business data. They may write programs, apply

What is the difference between data science and data engineering?

Data engineers build and maintain the systems that allow data scientists to access and interpret data. They work more closely with underlying technology than a data scientist. The role generally involves creating data models, building data pipelines, and overseeing extract, transform, load (ETL). Depending on organization setup and size, the data engineer may also manage related infrastructure like big-data storage, streaming, and processing platforms like Amazon S3.Data scientists use the data that data engineers have processed to build and train predictive models. Data scientists may then hand over the results to the analysts for further decision making.

What is the difference between data science and machine learning?

learning?Machine learning is the science of training machines to analyze and learn from data the way humans do. It is one of the methods used in data science projects to gain automated insights from data. Machine learning engineers specialize in computing, algorithms, and coding skills specific to machine learning methods. Data scientists might use machine learning methods as a tool or work closely with other machine learning engineers to process data.

What is the difference between data science and statistics?

Statistics is a mathematically-based field that seeks to collect and interpret quantitative data. In contrast, data science is a multidisciplinary field that uses scientific methods, processes, and systems to extract knowledge from data in various forms. Data scientists use methods from many disciplines, including statistics. However, the fields differ in their processes and the problems they study.

What are different data science tools?

AWS has a range of tools to support data scientists around the globe:

Data storage

For data warehousing, Amazon Redshift can run complex queries against structured or unstructured data. Analysts and data scientists can use AWS Glue to manage and search for data. AWS Glue automatically creates a unified catalog of all data in the data lake, with metadata attached to make it discoverable.

Machine learning

- Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 or Glacier. It is fast, serverless, and works using standard SQL queries.
- Amazon Elastic MapReduce (EMR) processes big data using servers like Spark and Hadoop.
- Amazon Kinesis allows aggregation and processing of streaming data in real-time. It uses website clickstreams, application logs, and telemetry data from IoT devices.
- Amazon OpenSearch allows search, analysis, and visualization of petabytes of data.

What does a data scientist do?

A data scientist can use a range of different techniques, tools, and technologies as part of the data science process. Based on the problem, they pick the best combinations for faster and more accurate results.

A data scientist's role and day-to-day work vary depending on the size and requirements of the organization. While they typically follow the data science process, the details may vary. In larger data science teams, a data scientist may work with other analysts, engineers, machine learning experts, and statisticians to ensure the data science process is followed end-to-end and business goals are achieved.

However, in smaller teams, a data scientist may wear several hats. Based on experience, skills, and educational background, they may perform multiple roles or overlapping roles. In this case, their daily responsibilities might include engineering, analysis, and machine learning along with core data science methodologies.

What are the challenges faced by data scientists?

Multiple data sources

Different types of apps and tools generate data in various formats. Data scientists have to clean and prepare data to make it consistent. This can be tedious and time-consuming.

Understanding the business problem

Data scientists have to work with multiple stakeholders and business managers to define the problem to be solved. This can be challenging—especially in large companies with multiple teams that have varying requirements.

data from middle-aged individuals, it may be less accurate when making predictions involving younger and older people. The field of machine learning provides an opportunity to address biases by detecting them and measuring them in the data and model.

How to become a data scientist?

There are usually three steps to becoming a data scientist:

1. Earn a bachelor's degree in IT, computer science, math, physics, or another related field.
2. Earn a master's degree in data science or related field.
3. Gain experience in a field of interest

Data science next steps

Check out additional product-related resources

Learn more about datalakes and analytics »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »