Foundations of Data

Data Science

SAMPLER

Foundations of Data

Data Science

Foundations of Data:
Data Science

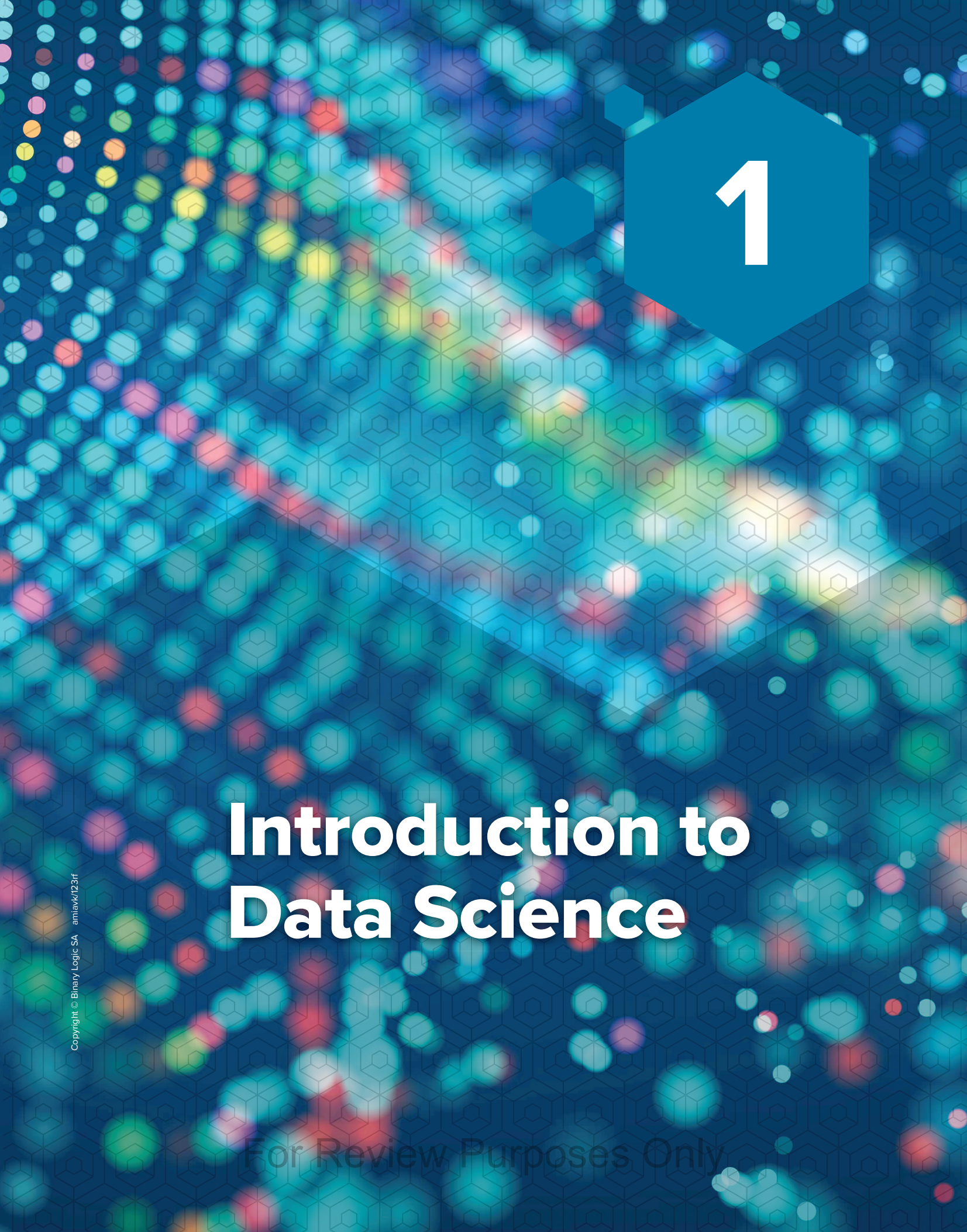
Printed and distributed by McGraw Hill in association with Binary Logic SA.

Cover Credit: © peshkova/123rf

 MHID: 1265796556 ISBN: 9781265796556

mheducation.com **binarylogic.net**

Contents



1

Introduction to Data Science

1

Introduction to Data Science

INTRODUCTION

Data Science plays a big role in understanding and solving important real-world problems by helping us make sense of the vast amounts of data around us. This unit covers the basics of Data Science, including the differences between data, information, and knowledge, and explore key concepts like the Data Science Life Cycle, Big Data, data governance, and career paths in this exciting field.

LEARNING OBJECTIVES

In this unit, you will:

> explain what Data Science is.

> distinguish the difference between data, information, and knowledge.

> recognize the differences between Data Science and Business Intelligence.

> examine the convergence of Data Science and Artificial Intelligence.

> identify the stages of the Data Science Life Cycle.

> describe what Big Data is.

> identify the characteristics of Big Data.

> categorize Big Data technologies.

> define what data governance is.

> identify data governance principles.

> discuss the skills and tools Data Science requires.

> identify professions related to Data Science.

> understand the importance of Data Science online communities.

LESSON 1

Data, Information, and Knowledge

Data Science

The importance of **Data Science** lies in the fact that data has become an essential part of industry, because companies require data to function, grow, and improve their businesses. Data helps companies in making proper decisions through data-driven approaches that analyze a large amount of data to derive meaningful insights.

Data Science application areas:

- Commercial and industrial applications.
- Healthcare, bioinformatics, and natural sciences.
- Digital economy, social media, and social networks analysis.
- Smart homes, smart cities, and smart transportation.
- Education, e-learning, and entertainment.
- Energy, sustainability, and climate.

Data Science

Data Science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unrevealed patterns, derive meaningful information, and make business decisions.

Data and Information

We are surrounded daily by data. We receive information from television, newspapers, books, and the Web. But what is the difference between data and information?

Data is a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process. For example, below is a collection of a student's personal data.

When data is processed, organized, structured or presented in a given context so as to make it useful, it is called **information**.

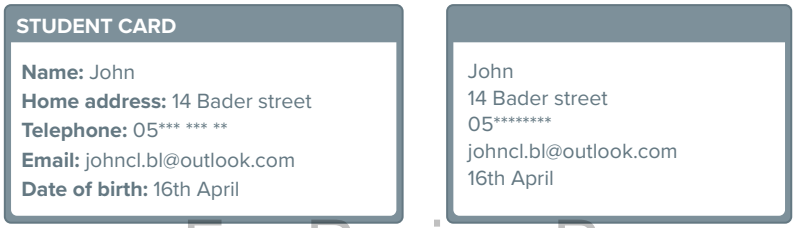
For example, the student card provides organized information about a student. On this student card, you can notice information such as the name, home address, telephone, email, and date of birth.

Data

The representation of facts or ideas in a suitable format for storage, processing, or transmission.

Information

A set of processed, organized, and structured data that provides context and enables decision making processes.



Raw Data and Information

Raw data is data that has just been collected from various sources and has not yet been processed for use. Data usually refers to raw data. Once the data has been analyzed, it is considered information. Let's think about some examples:

- The number "8122001" is considered raw data because it is a value with no contextual meaning. Now, if this value is presented as "8/12/2001, your date of birth", then this is information as it provides knowledge about a certain matter.
- Each student's test score is one piece of data. The average score of a class or of the entire school is information that can be derived from the data.

Information for further processing

Data or information from different sources can also be combined together to create more powerful datasets. This process is called data blending. For example, you can combine information from the marketing and sales departments to understand which marketing campaigns were more successful and profitable for a group of products.

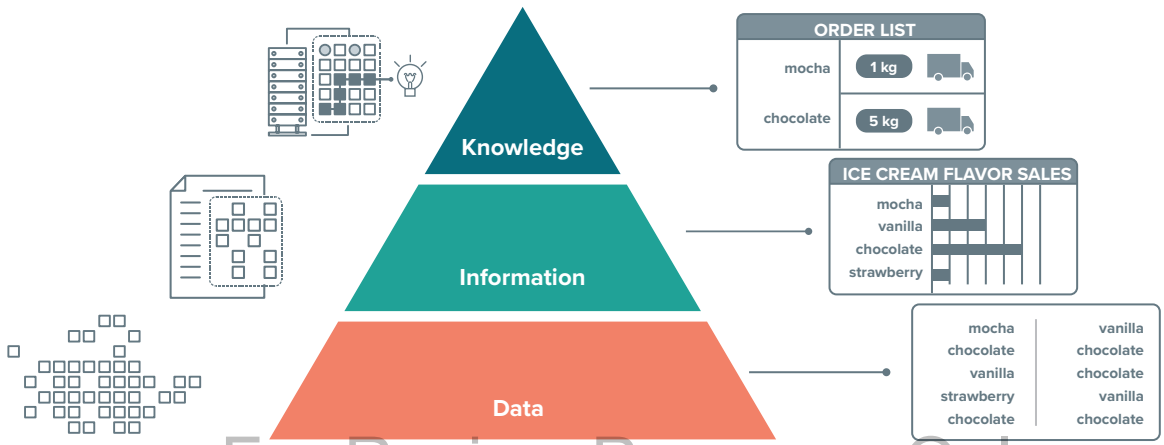
Differences between data and information

Data	Information
Unstructured	Has a logical structure.
Presented in the form of numbers, figures, or statistics.	Presented through reports, graphs, or plots.
No dependencies	Dependent on data.
Derived from user or computer system inputs.	Derived from data processing.

Knowledge

Knowledge is our understanding of the world. In other words, it is the appropriate collection of information in a way that makes it useful. We can say that when a person understands some information about something, then they have knowledge about it. Information becomes knowledge when critical thinking, evaluation, structure, or organization is applied.

Let's explore the example of the pyramid below: The data you can notice at the bottom is a list of words having no context.



Now, if we organize this data, we can provide information. Let's suppose that this is a list of the sales of ice cream flavors from yesterday. A bit of analysis is useful to glean more information. For example, the most popular flavor of ice cream sold yesterday was chocolate.

The knowledge is that the shop manager can notice that chocolate is the most popular ice cream flavor. The next time he places an order, he will ask for five times as much chocolate ice cream as mocha ice cream.

Differences between information and knowledge

	Information	Knowledge
Meaning	A refined form of processed data.	Relevant information that leads to conclusions.
Predictability	Not sufficient to make predictions.	Provides the ability to predict or make decisions.
Transfer	Can be transferred easily through verbal, written, or electronic means.	Requires learning of the subject.
Outcome	The outcome is understanding.	The outcome is comprehension.
Objective	Answers the questions of who, when, what, and where.	Answers the questions of how and why.

Data Science versus Business Intelligence

Data is everywhere around us, and it is used, processed, and analyzed in every field today. At the same time, data is constantly evolving and is used in several business applications, like Business Intelligence. **Business Intelligence** is a technology-driven process that analyzes data, providing important information that helps executives and managers make careful business decisions. While both Data Science and Business Intelligence involve data, they are different from one another.

Business Intelligence

A data-driven system that incorporates data collection, data storage, data analysis, and data visualization to support decision making.

Data Science is much more complex compared to Business Intelligence. The scope of Business Intelligence is limited to the business domain. In Business Intelligence, past data is analyzed by developing dashboards, creating business insights, organizing data, and extracting information that would help the businesses to grow, with the final goal being the understanding of the current trends of the business. However, in Data Science, we use data to make future predictions and forecast the growth of the business, using a wide array of complex statistical algorithms and predictive models.

Additionally, Business Intelligence tools are limited to analyzing organizational information and setting up business strategies. On the other hand, the tools of a **data scientist** involve complex algorithmic models, data processing, and even **big data** tools.

Differences between Data Science and Business Intelligence

	Data Science	Business Intelligence
Scope	Data is used to make future forecasts for the development of the business.	Past data is analyzed to understand the current trends of the business.
Tools	It includes complex algorithmic models, data processing, and even big data tools.	The tools are limited to analyzing management information and overseeing business strategies.
Data types	It works with structured data, but mainly deals with unstructured and semi-structured data.	It works with structured data that is typically data warehoused or stored in data silos.
Complexity	It has more complexity compared to business intelligence.	It is much simpler compared to data science.
Flexibility	It is much more flexible as data sources can be added as required.	It is less flexible as data sources must be pre-designed.

Data Science and Artificial Intelligence

Data science has already been defined, and you are aware that **Artificial Intelligence** (**AI**) is another field that deals with massive amounts of data. These two technologies can be used independently to solve difficult challenges and they can also converge and complement one another.

Artificial Intelligence (AI)

A computer science field that focuses on building systems capable of performing tasks that usually require human intelligence, such as learning, reasoning, problem-solving, language, and perception.

Data science processes historical data using computational tools to describe situations (descriptive analysis), predict results (predictive analysis), and provide recommended solutions to problems (prescriptive analysis). The most commonly used tools are statistical and management tools, which enable the analysis of historical data. On the other hand, AI employs a variety of techniques to mimic the way people think, decide, and solve problems.

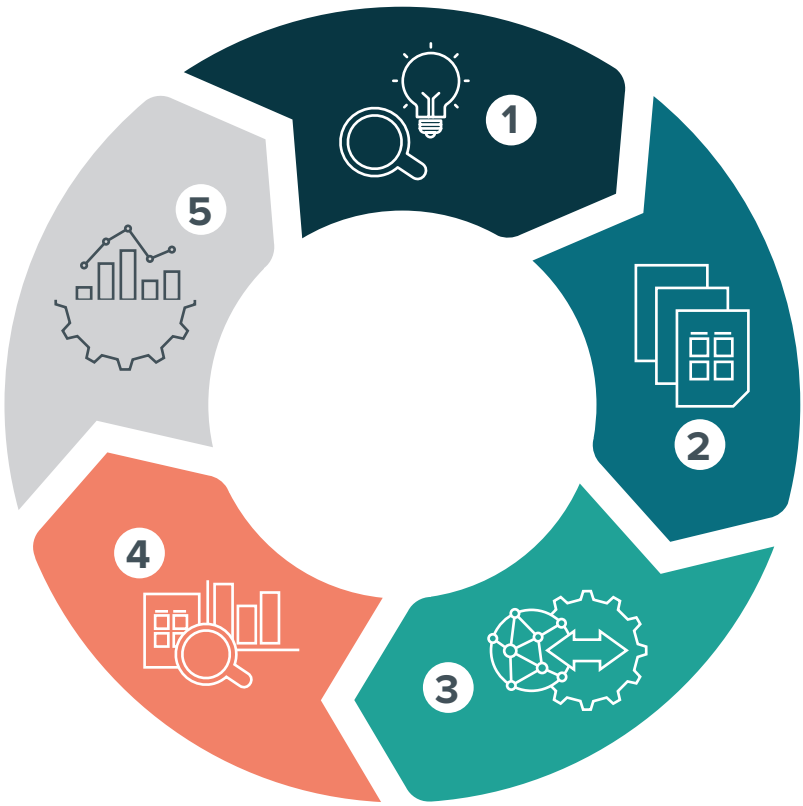
Rather than focusing on computation, the emphasis when working with AI tools is on knowledge and intelligence as critical elements for solving problems. Additionally, AI is concerned with cognitive computing. This distinction is less obvious in practice because sophisticated data science projects often include machine learning (an AI discipline) to facilitate **data analysis** in both prediction and prescription.

Data science and machine learning provide significant contributions to many organizations when used independently. However, traditional data analysis techniques are unsuitable when working with incomplete or inaccurate data, or the business or scientific contexts are changing so quickly that accurate data becomes obsolete very quickly. Similarly, machine learning technologies require a relatively significant amount of data.

Therefore, the next generation of data science tools and business intelligence platforms use machine learning to conduct, for example, pattern recognition to discover hidden patterns and visualize crucial insights. In addition, machine learning and deep learning support data science with more accurate predictions. The availability of large datasets and the reduced cost of processing on the cloud empower machine learning with capabilities not possible in the past. When data science and AI are combined, they create synergies that provide significantly superior results and lead to better and faster decisions.

Data Science Life Cycle

Through their experience working in data science projects, data scientists and data professionals follow specific steps to implement each new project successfully. This process, called the **Data Science Life Cycle**, has five distinct stages. This model has numerous variations that extend the stages to cover special projects, such as AI and machine learning projects, or to represent the internal processes of specific organizations.



- Problem definition and formulation
- Data collection
- Data preparation and cleaning
- Exploratory data analysis
- Data visualization

Problem definition and formulation

In order to design and create a solution for a Data Science problem, we first need to understand what the problem itself is. A thorough analysis of the problem, its environment and the variables that affect it are crucial for developing the solution. The understanding that we have of a problem can greatly improve or hinder the development of its solution because it directly correlates with our approach to that solution. The next objective is to define the goal we want from that solution. A dataset always contains the same data, but the answers we want to derive can vary.

Problem definition and formulation

Understanding the objectives and requirements of a business or scientific problem and converting this knowledge into a data analysis problem.

The most common types of data analysis

Type	Description
Regression analysis	Get the quantities or qualities that exist in the dataset.
Classification analysis	Organize the data into categories.
Clustering analysis	Organize the data into groupings.
Anomaly detection analysis	Find oddities or rarities in the data.
Recommendation engines	Give an informed decision on a specific question.

Data collection

Once the objectives are set, we need the dataset itself. Besides manual entry of data, the most common way is data mining or data gathering. In this stage, enough data must be collected for further processing. The data itself can come from a variety of sources. Environmental sensors or mobile applications and web platforms continuously generate data. This data is automatically stored in databases.

Data Collection

The process of gathering and measuring data, including data acquisition, data labeling, and data improvement.

The most common data storage formats

Type	Format
Formatted Files	JSON, XML, CSV, Spreadsheet XLS/XLSX
Relational Databases	Microsoft SQL Server, Oracle Database, MySQL
Non-Relational (NoSQL) Databases	MongoDB, Azure Cosmos DB, AWS DynamoDB
Graph Databases	Neo4j, AWS Neptune, Dgraph
Time-Series Databases	InfluxDB, AWS Timescale

Data preparation and cleaning

Data cleaning, or data wrangling, is one of the most important stages in the Data Science Life Cycle. The data scientist must clean and prepare the collected data from the **data mining** stage to ensure they are suitable for the subsequent analysis stage. When we combine multiple data sources, there are many chances for data to be duplicated or mixed up, and these issues will need to be fixed. If there are corrupted or incorrectly formatted data, duplicate or false data, or just incomplete data, the insights derived in the analysis stage will be false, and it will be very difficult to deduct whether the problem with the false insights originates from errors in the analysis steps or uncleaned data.

Data Cleaning

The multistage process of reviewing and correcting data to ensure it is in a standardized format, including handling missing values, smoothing noisy data, and resolving inconsistencies and duplicates.

This is why taking the time and the effort to clean and validate the data thoroughly before analyzing it is highly important for the entire process.

Exploratory data analysis

We have collected and thoroughly cleaned our data, and now it is time to analyze the dataset we have gathered and derive the desired answers to our questions. Data analysis is performed with data analysis tools or programming code and the relevant code libraries. It can start with a relatively simple analysis of one or more variables and expand to more sophisticated processes involving advanced statistics.

Exploratory data analysis

The approach to analyzing datasets to summarize their main characteristics, often using visual methods.

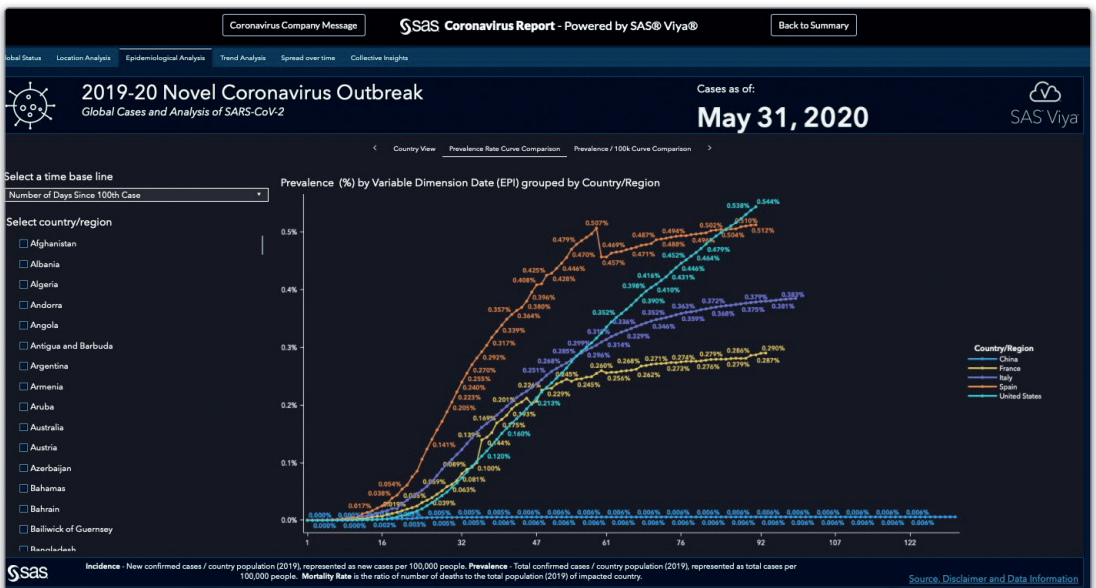
In today's world, the most prominent method of analyzing a dataset is Machine Learning. To analyze data with Machine Learning, we need to follow specific steps. We first need to define the Machine Learning (ML) model. We do this by first specifying what the input and output values are. The next step is to construct the analysis algorithm itself. This is a complicated process, and specialist data scientists and machine learning engineers are sometimes used solely for this task. After the algorithm is completed, it is time to train and test the model. When the training and testing phases are completed, we can then use the production data and finally generate the answers we want.

Data visualization

The analyzed data are usually tables of new data that are useful in the experienced perspective of **data analysts**. Working with a visual representation of the analysis helps to derive better insights. Graphs, plots and charts, or even maps, along with formatted reports, provide an efficient way to notice and understand trends and patterns in data. When working with massive amounts of information, visualization of the results is essential to make data-driven decisions.

Data visualization

A graphical representation of information that highlights patterns and trends in data and aids the reader in gaining quick insights.



EXERCISES

1 Choose the correct answer.

1. A representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process it is called:
   - **A.** Data
   - **B.** Knowledge
   - **C.** Information
   - **D.** Raw data
2. When data is presented in a given context to make it useful, it is called:
   - **A.** Data
   - **B.** Knowledge
   - **C.** Information
   - **D.** Graph
3. Data that has just been collected from various sources and has not yet been processed for use are called:
   - **A.** Data
   - **B.** Knowledge
   - **C.** Information
   - **D.** Raw data

2 Mention three basic differences between Data Science and Artificial Intelligence. Justify your answers providing examples.

3 How effective is the convergence of Data Science and Artificial Intelligence? Search the Internet and find two successful examples.

4 Explain what Data Science is and identify three applications in everyday life for health, business, and entertainment. Why is Data Science so important for these applications?

5 Search on the Internet for Data Science life cycle models that describe the key stages mentioned in this lesson in more detail. Select one of them, identify the additional stages and briefly explain them.

6 Compare and contrast sets of unprocessed and processed data that describe the annual grades and performance of a student. What insights can you get from datasets like this? Can you predict the academic performance of the student at university?

LESSON 2

Working with Data

What is Big Data?

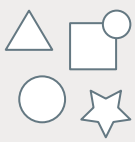
The term **Big Data** refers to data that is either too large or complex to process using typical methods. Due to the fact that this amount of data is too large for typical computing systems to manage, the storing and the processing of these huge datasets is considered a challenge. Furthermore, **data collection** might be so rapid that storage requirements are extremely high.

Big Data

A large dataset that requires scalable technologies for storage, processing, management, and analysis due to its characteristics of volume, variety, velocity, veracity, and value.

Characteristics of Big Data

There are five key concepts that help us to classify any data under the term of Big Data: the **Variety**, the **Value**, the **Volume**, the **Veracity,** and the **Velocity**. Data is considered "Big" when it comes in large volumes, at a very fast rate, with great Variety, and is accurate and useful. Data must fulfill all these "Vs" in order to be considered "Big Data".



Variety

Variety refers to the many different types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semi-structured data types (such as text, audio, and video) require additional preprocessing to derive meaning and support metadata information. Without the metadata, it will be impossible to know what is stored and how it can be processed.

Value

Just because we collected lots of data, this does not mean it is of any value, we have to garner some insights out of it. Value refers to how useful the data is in decision-making. We need to extract the value of the big data using proper analytics.

Volume

Because large volumes of low-density, unstructured data must be handled, the amount of data is a critical aspect in big data. This can be unvalued data like clickstreams on a website or mobile app, or sensor-enabled IoT devices. It might be tens of terabytes of data at times, and hundreds of petabytes at other times.

Veracity

Data veracity has to do with how accurate or truthful a dataset may be. It's not just the quality of the data itself but how trustworthy the data source, type, and processing is.

Velocity

The rate at which data is captured and stored is referred to as velocity. Most of the Internet-connected smart devices (IoT devices) and mobile devices work in real-time or near real-time, requiring instant data collection, transmission, and storage.

Technologies that Enable the Management of Big Data

Businesses use computer systems and databases to keep records of transactions such as order processing, payments, customer tracking, and cost management. A company will require a reporting system to provide information that will help it run more efficiently and help executives make more informed and, hopefully, better decisions.

For example, an e-shop will need to enhance the buying experience and ensure that website visitors convert into customers or that existing customers return for future purchases. By analyzing the data captured during e-shop browsing on the web or through a mobile app, the company can identify where visitors place their cursors, which parts of the website they focus on the most, and how long they hover over a product before clicking for more information or making an actual purchase. Tiny details are becoming vast amounts of data waiting to be analyzed and turned into valuable insights. This information will drive changes in the website layout, price adjustments, and product campaigns on social media to influence buying behavior.

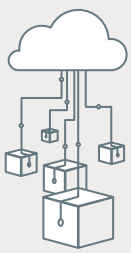
Companies require new technologies and tools to manage and analyze big data to extract business value. The required data must be gathered from internal sources such as sales, manufacturing, and accounting, and external sources such as demographic and competition data to extract concise, reliable information about the company's current state and market dynamics. Modern infrastructure for business intelligence has an array of tools to store and process data to obtain useful information from big data. These technologies include **data warehouses**, **data lakes,** and **in-memory computing**.

Data warehouse

As the most traditional tool to analyze corporate data, a data warehouse refers to the database that stores current and historical data originating from many core operational transaction systems (sales, customer support, manufacturing) and makes data available to a company's decision makers. This data is combined with data from external sources to transform incomplete data to structured data before being stored in the data warehouse. A data warehouse system also provides a range of ad hoc and standardized query analysis and graphical reporting tools.

Data lake

A data lake is a repository, usually in the cloud, to store huge amounts of raw and unprocessed data. It uses a flat URL structure to support both structured data (such as databases) and unstructured data (such as emails and documents).



In-memory computing

This is a way of facilitating big data analysis, because it relies primarily on the computer's main memory (RAM) for data storage. Users access data stored in system primary memory, thereby eliminating bottlenecks from retrieving and reading data that are present in a traditional, disk-based database, and dramatically shortening query response times. Very large quantities of RAM on cloud servers facilitate this method.

The distinction between these three technologies is important because they serve different purposes and require different handling to be properly optimized. They do not work together, depending on the type of company, one of the three is chosen. A data lake may work well for one company, while a data warehouse will be a better fit for another.

Mining Big Data

Big data is being continuously collected by sensors and by applications in our environments and applications that we use personally. But collecting the data is only the first step in the process referred to as Knowledge Discovery. Knowledge discovery refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. **Data mining** is the application of specific algorithms for extracting patterns from data and identifying relationships. The additional steps in the knowledge discovery process, such as data cleaning, data integration, data transformation, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data.

Data mining

Analysis of large pools of data to find patterns and rules that can be used to guide decision-making and predict future behavior.

Some of the main tasks accomplished by data mining are:

- Analyzing data to discover patterns and trends.
- Formulating predictions for different dataset inputs.
- Classifying, clustering, or forecasting the different values of the dataset.
- Facilitating decision recommendations.

Steps of Knowledge Discovery

Steps	Description
Data cleaning	Clean corrupt data, irregularities, false data types, etc.
Data integration	Data mining occurs from data that originates from multiple sources. These data sources need to be merged into a single dataset.
Data selection	Selecting the part of the dataset that we want to use for the data mining process. It is important to select the dataset that is most representative of your goals because data mining is a time-consuming task.
Data transformation	Preparing and formatting raw datasets is necessary because data mining processes need their inputs to have a specific format in order to analyze them.
Data mining	The actual process of analyzing the data and extracting the desired results from the analysis through patterns.
Pattern evaluation	Decoding the patterns that were generated by the data mining steps and deciding which are beneficial for each specific goal.
Knowledge representation	Visualizing the generated results with clear and concise reports, graphs, and plots.

Big Data and Cloud Storage

There are two options when storing big data: **cloud storage** and on-premises storage. In the beginning, the development of big data applications usually required keeping data in on-premises storage, which means inside expensive, local data warehouses with complex software installed. However, subsequent developments spelled the end of this approach, introducing cloud storage as the optimal solution for big data storage:

- The widespread availability of high-speed broadband which facilitates the movement of data from one place to another. Data produced locally no longer needs to be analyzed locally. It can be moved to the cloud for analysis.
- In today's world, the majority of applications are cloud-based, meaning that more data is being produced and stored in the cloud. Increasing numbers of entrepreneurs are building new big data analytics to help companies analyze cloud-based data such as e-commerce transactions and web application performance data.

The biggest benefit of the cloud is versatility. Cloud-based storage services include big data storage and backup systems.

For big data storage, there are a lot of options available offered by service providers such as Amazon, Microsoft, and Google. All of them provide data security and privacy as well as scalability and cost efficiency.

By using cloud backup for big data, enterprises can utilize services from data centers that span multiple geographic locations, ensuring high availability and easy data recovery. Using the cloud, backed up data can be replicated over multiple data centers in different regions of the world. This way, the backups aren't kept at a single location. There is another layer of security to the backup. Service providers ensure that the data being backed up to the cloud is protected via advanced encryption techniques before, after, and during transit.

As mentioned earlier, big data handling requires storage capacity and processing power. In terms of storage capacity, the cloud fulfills this role. Enterprises can acquire storage services that facilitate simplified scalability. And these services are also capable of meeting the computation requirements of big data. Actually, experts recommend cloud powered data analytics for big data analysis based on the almost infinite computing capabilities of the cloud.

Pros and Cons of Big Data Cloud Storage

The combination of big data analytics and cloud computing can generate opportunities not feasible before. Apart from the advantages, the data scientist needs to be aware of the challenges.

Big data cloud storage advantages and disadvantages

Advantages	Disadvantages
Large volumes of structured and unstructured data require increased transmission bandwidth and storage. The cloud provides readily available infrastructure and the ability to scale up to handle any amount of data traffic and storage requirements.	Less direct control over data security. Data breaches could lead to serious penalties under the latest data privacy regulations.
Storing big data in the cloud eliminates the need to maintain expensive on-promises hardware, software, and specialized staff. The pay-as-you-go cloud computing model is more cost-efficient, reducing the waste of resources.	The cloud service provider can raise the rates of their cloud infrastructure anytime. The company that consumes these services may become locked in a business relationship that is not cost-efficient.
The company focuses on the analytics process rather than the infrastructure management, reflecting positively on the business's culture, performance, and competitive advantage.	Storing big data in the cloud means data availability depends on network connectivity. Also, the issue of latency in the cloud environment spills over into the speed of capturing, processing, and storing data.

Data Governance and Policies

The policies, processes, and organizational structures define the decision rights and accountabilities that support data management. Data governance includes internal policies and procedures that control the management of data.

Data governance helps private enterprises or state and non-profit organizations in working with high-quality data management processes through all data life cycle phases. These effective policies and procedures lead to improved business or organizational outcomes. Enterprises and organizations currently collect vast amounts of internal and external data, and data governance is necessary to use that data effectively, manage risks, and reduce costs.

Data governance ensures that data is:

- Secure
- Trustworthy
- Documented
- Managed
- Audited

The importance of data governance

Data inconsistencies in various systems within an organization may not be resolved without proper data governance. In sales and customer service systems, for example, customer names may be listed differently. This could make data integration more challenging and affect the accuracy of business intelligence and reporting. Furthermore, data errors may go undetected and uncorrected, compromising the integrity of data.

More importantly, organizations that must comply with new data privacy and protection legislation, such as the European Union's GDPR and the California Consumer Privacy Act (CCPA), may encounter difficulties or even penalties as a result of poor data governance.

Data governance framework components

The policies, guidelines, processes, organizational structures, and technology implemented as part of a governance program make up a data governance framework. The framework also specifies the program's mission, goals, how success will be measured, and accountability for the functions that will be included in the program. The governance framework of an organization should be established and disseminated internally to explain how the program will work so that everyone engaged has a clear understanding from the start.

There are special types of data, such as financial or health data, that require careful handling. Health data is usually well governed from the time of data collection up to reporting and dissemination of information. All stakeholders fully understand the privacy risk and the constraints set by legislation, therefore a well-defined data governance framework, in a hospital, for example, is valuable.

Data governance standards

ISO, the International Standards Organization, has developed a standard, ISO/IEC 38505, to apply IT governance principles to the data governance requirements.

The six data governance principles

Principles	Actions
Responsibility	Assign to personnel.
Strategy	Align with the mission and vision of the organization.
Acquisition	Align with the organizational requirements.
Conformance	Ensure compliance with legislation, internal policies, and business ethics.
Performance	Meet the requirements of the organization.
Human behavior	Encourage people to get involved.

General data protection regulation

The General Data Protection Regulation (GDPR) is a regulation the European Union (EU) adopted in 2016 and enacted in 2018. The GDPR is a comprehensive data protection law that applies to all organizations that process the personal data of EU citizens, regardless of where those organizations are located.

The GDPR establishes several new rights for EU citizens, including the right to be informed about the collection and use of their data, the right to access their data, the right to have their data corrected, the right to have their data deleted, and the right to restrict or object to the processing of their data. The GDPR also requires organizations to obtain explicit consent from individuals before collecting and processing their data.

The GDPR also imposes strict requirements on organizations regarding handling personal data, including requirements for data security, data breach notifications, and the appointment of a Data Protection Officer. Organizations that fail to comply with the GDPR can be fined up to 4% of their annual global revenue or €20 million, whichever is greater.

California Consumer Privacy Act

The California Consumer Privacy Act (CCPA) is a privacy law that was enacted in California, United States, and went into effect in 2020. The CCPA is designed to give California residents more control over their personal information and regulate businesses' collection and sale of personal information.

Under the CCPA, California residents have the right to know what personal information is being collected about them, request that their personal information be deleted, and opt out of the sale of their personal information. Businesses that collect or sell the personal information of California residents must provide certain disclosures about their data collection and sharing practices, as well as honor requests from California residents to access, delete, or opt out of the sale of their data.

Data governance versus data management

It is critical to recognize that data governance is a component of overall **data management**. Data governance without actual implementation is just paperwork. Data governance establishes all policies and processes, whereas data management implements them to compile data and use it for decision-making. To draw an analogy, data governance is designing the plan of a new building, whereas data management is the act of building it. Furthermore, while you could build a house without a plan, it would be less efficient and effective, with a high risk of structural failures.

Data management

Data management is the creation and implementation of architectures, policies, and procedures that manage an organization's full data life cycle needs.

Data governance challenges

Cloud data and big data are two common data governance concerns that organizations encounter. Cloud services and big data systems introduce new governance requirements. Traditionally, data governance programs have focused on structured data stored in the data center. They now have to cope with the usual mix of structured, unstructured, and semi-structured data found in big data environments, as well as the privacy threats associated with cloud data platforms.

Who is responsible?

The data governance process involves a variety of people in most organizations. End-users familiar with relevant data in an organization's systems are included, as are business executives, data management specialists, and IT personnel. The key individuals are the Chief Information Officer (CIO) or Chief Data Officer (CDO) and the Data Governance Manager (DGM).

Data owner

An individual or people who are accountable for particular data.

The CIO is usually a senior executive in charge of the data governance program. The CIO's responsibilities include obtaining approval, funding and staffing for the program, taking the lead in its establishment, evaluating its development, and acting as its internal advocate.

Depending on the organization's size, a dedicated DGM may be appointed to lead and coordinate the process, hold meetings and training sessions, track KPIs (key performance indicators), and manage internal communications for the initiative. The DGM works with **data owners** and **data stewards**, who ensure that the data governance policies and rules are enforced, and that end-users follow them.

Data steward

A data management role that includes implementing and maintaining data governance policies within an organization.

EXERCISES

1 Choose the correct answer.

1. A What does "Big Data" refer to?

A. Small, manageable datasets

B. Data that is too large or complex for typical processing methods.

C. Data that can only be stored in physical formats.

D. Data used only for artificial intelligence.

2. Data selection in knowledge discovery involves:

A. Selecting all the data from every source.

B. Choosing the part of the dataset for the knowledge discovery process.

C. Deleting unnecessary data.

D. Combining unrelated datasets.

3. What is one of the primary benefits of cloud storage for big data?

A. Limited scalability

B. High cost of storage

C. Versatility, including data storage and backup options.

D. Requirement for on-premises servers.

2 Mention three examples of how big data can help businesses.

3 Search the Internet in order to find today's most popular cloud computing service providers in the global market, which are used to store and process big data.

4 Explain in a few sentences how the cloud helps us deal with the problem of storing the huge amount of data that big data represents.

5 Big data is a recent development in the history of computing. Can you identify two factors that enabled this sudden growth of data collection?

6 Compare the three big data storage technologies. If you developed an application that requires very fast access to data, which one would you choose?

7 Describe the purpose of data governance? Is data governance a synonym of data management?

LESSON 3

Data Science Fundamentals

Mathematics Needed to Become a Data Scientist

Data science algorithms, as well as implementing analyses and discovering insights from data, require mathematical knowledge. While mathematics isn't the only tool required for a data scientist, it is one of the most significant. One of the most critical elements in a data science project workflow is identifying and comprehending business challenges and turning them into mathematical ones.

Linear algebra

Linear algebra is concerned with matrix and vector operations. This is very important because in data science models and algorithms, all the numbers and information are converted into matrices. Another technique linear algebra is used for is dimensionality reduction which is necessary for processing large datasets. Computer Vision and Natural Language Processing (NLP) are also data science fields that rely heavily on linear algebra. All the numbers and information are converted into matrices in data science models and algorithms.

Discrete mathematics

Discrete mathematics specializes in logic and deduction methods, which are paramount aspects of algorithm design and are the basis for data science. Another very important field of discrete mathematics is graph theory. Graphs are used for modeling very complex networks such as recommendation systems in platforms related to music/movies. Their study in data science is valuable for the advancement of fields such as precision medicine, systems biology, and many more.

Probability and statistics

When the data from an analysis gets generated, a data scientist needs practical statistical and probabilistic knowledge to be able to understand and interpret that data. Measures such as the variance, correlation, and standard deviation are used extensively by data scientists to gather insight into the underlying relationships of the features of a dataset.

Calculus

Visualizing the results from a data analysis is critical to provide insightful information through the generation of plots and graphs. Calculus is an integral part of the algorithms used for the complex arithmetic operations required in this process. Concepts such as partial derivatives, linear regression, and gradient descent are used extensively in optimization and loss calculation.

Python for Data Science

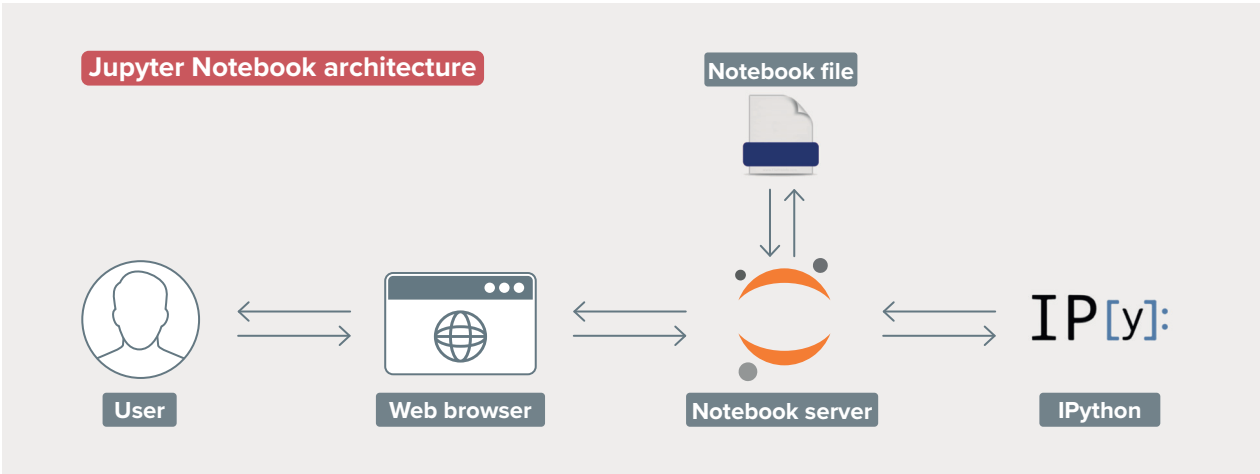
Data Science professionals generally prefer using **Python** for their Data Science projects. It is a high-level, object-oriented programming language that has an easy learning curve. It is easy to begin working on a project, as you can start by writing simple structured code or design and implement a solution with Object Oriented Programming (OOP) principles.

> **Python**
>
> A high-level and general-purpose programming language which has gained increasing popularity in data science and machine learning.

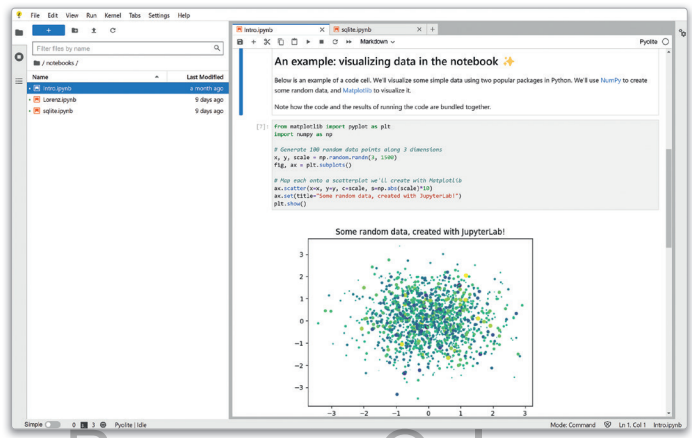
The use of Application Programming Interfaces (APIs) and library modules provides access to powerful functionalities that are easy to use. There are numerous Python libraries that are used by professionals in various enterprises covering a wide variety of needs: data mining, data preparation and analysis, data processing, predictive modeling data visualization and reporting. Going beyond traditional data science applications, Python libraries support machine learning and advanced artificial intelligence requirements.

Introduction to Jupyter

Python scripts can be written in an Integrated Development Environment (IDE) such as Visual Studio Code or JetBrains PyCharm, or they can be written in Jupyter Notebook. Jupyter Notebook is an open-source web application which is used to develop and present data science projects with Python. The interactive environment enables data scientists to create "notebooks". A notebook integrates Python code and its output into a single document that combines visualizations, narrative text, mathematical equations, and other data visualizations. After Jupyter is installed, it runs in a web browser either online or on a personal computer.



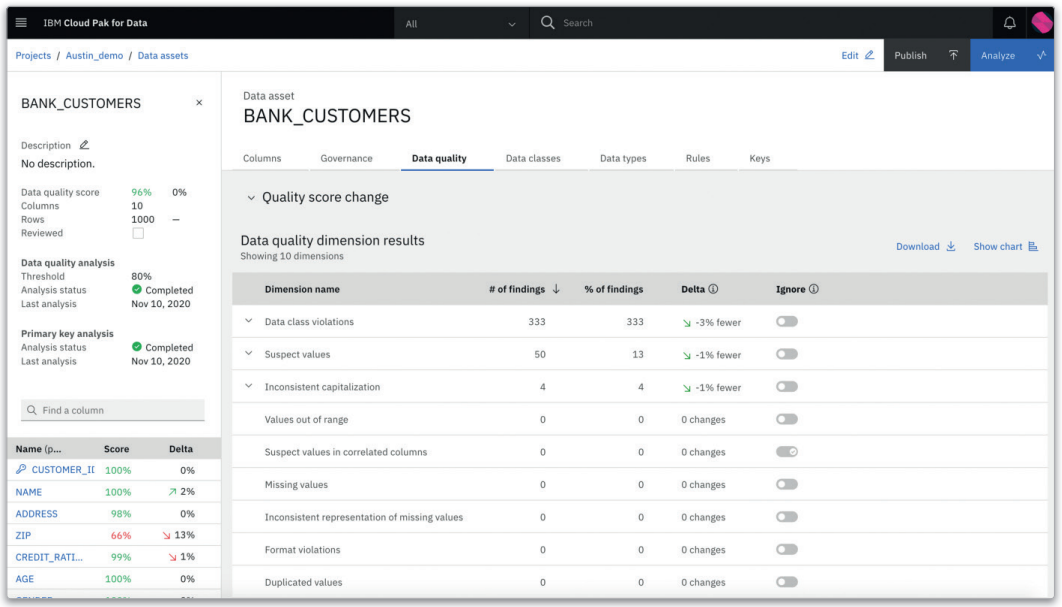
Besides Python, Jupyter Notebook supports over 100 programming languages (called "kernels" in the Jupyter ecosystem) including Java, R, Julia, MATLAB, Octave, Scheme, Processing, Scala, and many more. Out of the box, Jupyter will only run the IPython kernel, but additional kernels may be installed.



We will use Jupyter Notebook for Exploratory Data Analysis later in this book. The latest web-based application for Jupyter is JupyterLab, and all notebook documents work the same in both environments.

Tools for Data Science

Data science is a complex process which requires a lot of steps in order to create a data science solution. For each step of the process there exist numerous tools for accomplishing the desired task. Below are the most popular tools for each data science step:



Popular tools for data science steps

	Purpose	Software tools
Data Storage	The databases where the data is stored.	MySQL, SQL Server, MongoDB, Neo4j
Data Transformation	Tools that query the data that we want to analyze.	Pandas, NumPy, Apache Spark
Modeling	Converting the queried data into models that are appropriate for analysis.	Tensorflow, PyTorch, Scikit-learn, IBM Watson. AWS Sagemaker
Analysis	The process that generates the desired insights.	Pandas, R, SciPy, Excel
Visualization	Visualizing the results in the optimal format.	Matplotlib, Seaborn, D3.js, ggplot, plotly, Tableau, Power BI

Data Science Jobs

Data science has been one of the fastest-growing and most in-demand fields in recent years. The rise of big data and the increasing availability of tools and technologies for analyzing and interpreting large datasets have created a growing demand for skilled data scientists who can transform data and turn it into insights and actionable recommendations. This demand for data scientists is expected to continue in the coming years as more companies and industries seek to harness the power of data to drive business decisions and improve outcomes. Beyond business, data science is also critical in fields like healthcare, where data scientists analyze medical data to improve patient care and outcomes. As a result, there are many opportunities for people with the skills and knowledge to work in this exciting and rapidly evolving field.

Professions related to Data Science

Professions	Description
Data scientist	Their job is to find, process, and analyze data for companies and organizations. They take raw and unprocessed data and extract insights and patterns from the data that help companies and organizations analyze their performance and make mission critical decisions.
Machine Learning engineer	They are responsible for implementing Machine Learning (ML) solutions and systems for the appropriate applications. They need to be knowledgeable in software engineering and statistics in order to be able to test their solutions and judge the correctness of the produced ML models.
Machine Learning specialist	While ML engineers are concerned with the application of ML models, ML specialists focus on the mathematics of the specific algorithms that produce the models that engineers are then able to utilize.
Applications architect	They design the information systems for organizations and companies.
Enterprise architect	They combine business and technical knowledge, and they are in constant communication between stakeholders and technical departments. They are tasked with translating business and organization data needs into technological specifications and solutions which they forward to the technical teams.
Data architect	They are responsible for the storage and flow of information in a company or organization. They work with data scientists and engineers to build the appropriate data pipelines for dataset input, analysis, and results output.
Data engineer	They assist the data architects in building the digital framework for data capture, storage and processing, which both data scientists and analysts will use for their work.

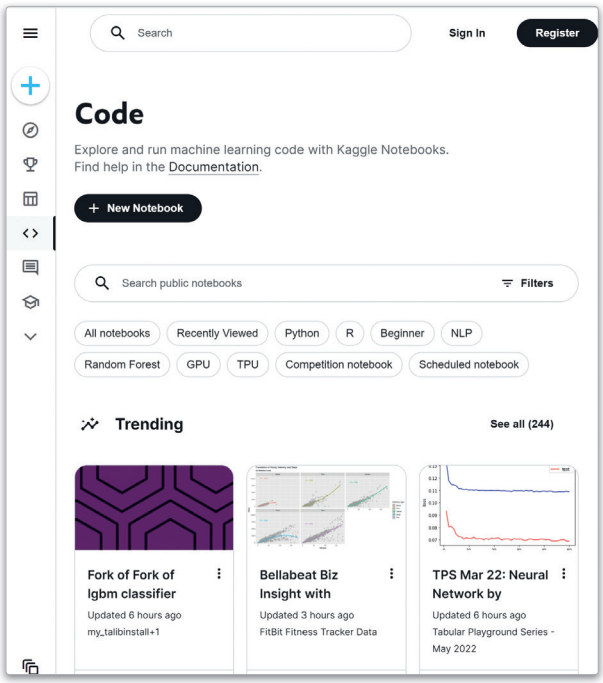
Data Science Online Communities

Data scientists want to stay in touch with their peers in the field or in similar professions to learn new ideas and approaches because Data Science methodologies and technologies are always changing. Only online resources can aid data scientists in keeping up the pace. The need for a community of Data Science experts to support this work has sparked a variety of online fora and groups. Data scientists can connect and efficiently evolve the field by participating in Data Science online communities. The most prominent communities are mentioned below but this is an area where new communities may emerge and become successful.

Kaggle

Kaggle, a Google subsidiary, is the largest data science community with millions of active members and a wide range of resources. Data scientists can find public datasets, educational resources and cloud-based workbenches to support their data analysis work.

https://www.kaggle.com



IBM Data Community

IBM Data Community is an online forum with blogs dedicated to data science. It hosts research papers, webcasts, and presentations that are updated as the field evolves.

https://community.ibm.com/community/user/ai-datascience

There are more online communities, some of them supported by governments, some run by volunteers. Some are more focused on the community side with face-to-face meetings, while others are focused on the code required for data science projects.

Online communities

Communities	URLs
Data Science Central	https://www.datasciencecentral.com
Stack Exchange	https://datascience.stackexchange.com
Data Science Society	https://dssberkeley.com
Driven Data	https://www.drivendata.org
Data Community DC	https://www.datacommunitydc.org
Reddit	https://www.reddit.com/r/datascience

Remember to always check the online reputation of the content contributor before using a dataset, code, or tools. Check for the permissions of use for each dataset and try to download software tools directly from their developers' repositories.

EXERCISES

1 Choose the correct answer.

1. What type of knowledge is essential for a data scientist to understand and interpret data generated from an analysis?
   - **A.** Practical statistical and probabilistic knowledge
   - **B** Knowledge of quantum mechanics
   - **C.** Extensive knowledge of physics
   - **D.** Advanced programming skills only
2. What is one of the primary responsibilities of a Data Scientist?
   - **A.** Managing cloud infrastructure.
   - **B.** Creating financial projections.
   - **C.** Generating reports, visualizations, and analytics aligned with data science project objectives.
   - **D.** Writing code for mobile applications.

2 Mention the most important tools for Data Science. How exactly do they contribute to each Data Science step?

3 Describe why understanding statistics is a fundamental skill for a data scientist. Can you mention of an example involving data analysis?

4 You are learning to become a data scientist and have mastered Python coding. What other tools will you need for your data science toolkit?

5 Compare and contrast an IDE and Jupyter Notebook. What are the main ways they differ?

PROJECT

Social networks and Privacy

Social networks accumulate vast amounts of information every day.

1. Identify three daily routines that produce private data useful to these organizations.

2. More specifically, you have to answer questions like:
   • What types of data are collected?
   • Is all this data available to the public?

3. Prepare a presentation about the privacy concerns related to social networks and how a user can be protected. What are the best practices to avoid your data becoming useful information that can be exploited by others?

1 WRAP UP

THIS UNIT COVERED HOW TO:

> define Data Science.

> differentiate between data, information, and knowledge.

> differentiate Data Science from Business Intelligence and Artificial Intelligence.

> describe the stages of the Data Science Life Cycle.

> define big data.

> explain how Python or other tools can contribute to Data Science.

KEY TERMS

- Artificial Intelligence
- Big Data
- Business Intelligence
- Cloud Storage
- Data
- Data Analysis
- Data Analyst
- Data Cleaning
- Data Collection
- Data Lake
- Data Mining
- Data Preparation
- Data Science
- Data Science Life Cycle
- Data Scientist
- Data Visualization
- Data Warehouse
- Exploratory Data Analysis
- Information
- In-memory Computing
- Knowledge
- Raw Data
- Value
- Variety
- Velocity
- Veracity
- Volume

Foundations of Data

Data Science

Design, Simulate, and Innovate

Imagine exploring the world of data and using its power to uncover patterns, make predictions, and solve complex real-world problems. What if you could learn to collect, validate, and analyze data, then create models that forecast future outcomes? This course takes you through the essential steps of data science, from gathering raw data to developing predictive models.

Foundations of Data Science guides you through the fundamentals of data analysis, gets hands-on with Python libraries, and discovers the power of data visualization to communicate insights. Master the art of predictive data modeling and learn techniques for optimization and forecasting that will give you the edge in data-driven decision-making.

By the end of this course, you'll have the expertise to collect, analyze, and model data with confidence, empowering you to innovate and transform data into valuable solutions across industries.

Mc Graw Hill
binarylogic

ISBN-13: 978-1-2657-9655-6
ISBN-10: 1-2657-9655-6
9 781265 796556