Applied Data Science with Python

Introduction to Data Science

Learning Objectives

By the end of this lesson, you will be able to:

- Explain the basics of data science and its application
- List the steps of the data science process
- Explore the Python packages for data science
- Describe the types of plots available for visualization

Introduction

Data Science

It is a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract meaningful insights and knowledge from structured and unstructured data.

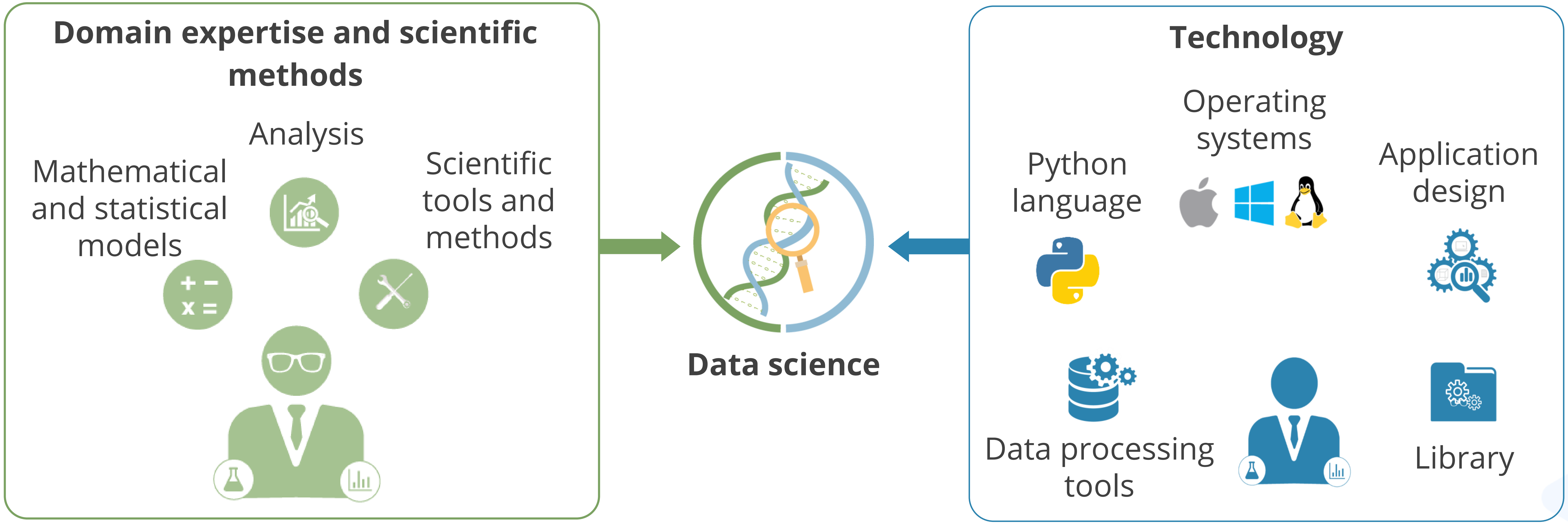
Example

Using a search engine or buying from Amazon gives valuable inputs to data-science-based software systems working in the background.

Data on interactions with online platforms is gathered to understand user preferences and suggest search results or items to buy.

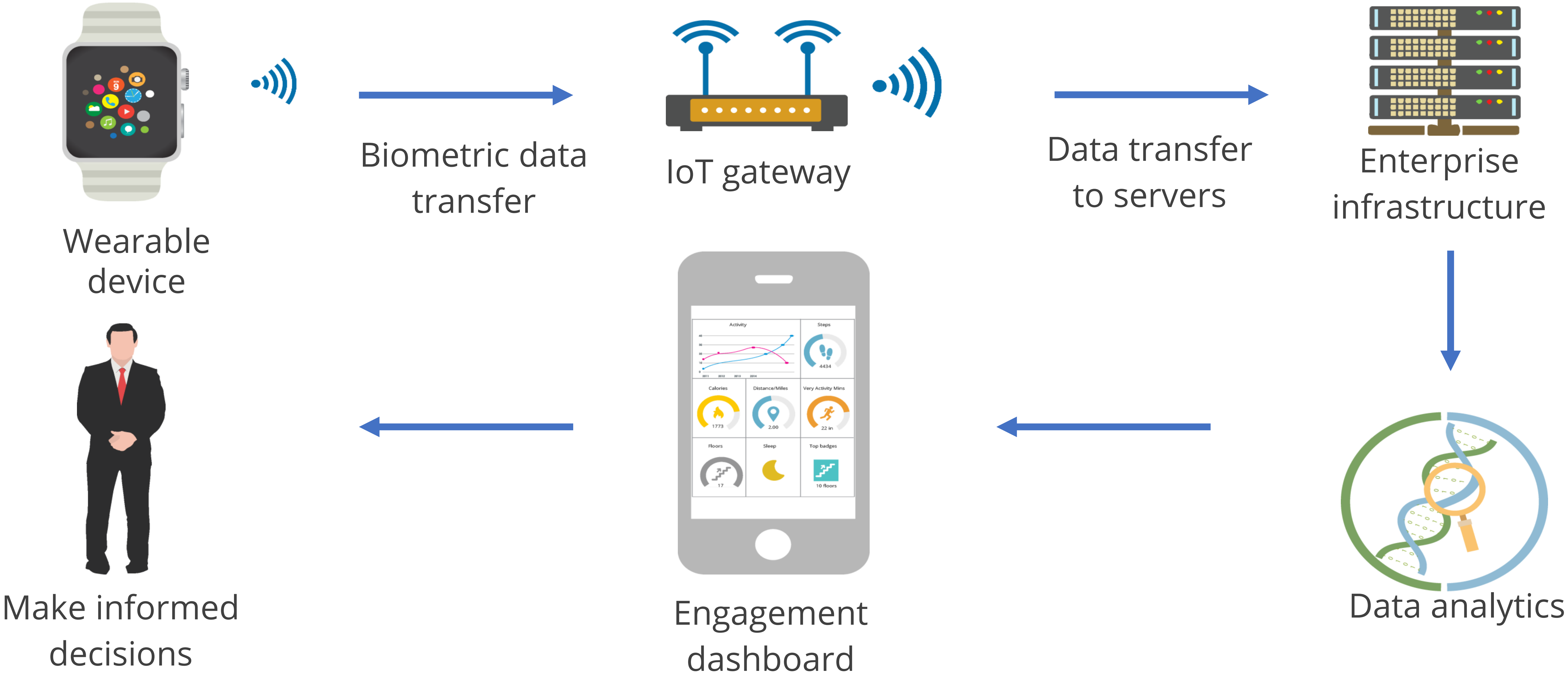
Data Science

Data science is created when subject expertise and scientific methodologies are combined with technology.



Application of Data Science: Healthcare

Wearable devices use data science to analyze data gathered by their biometric sensors.



Wearable device

Biometric data transfer

IoT gateway

Data transfer to servers

Enterprise infrastructure

Data analytics

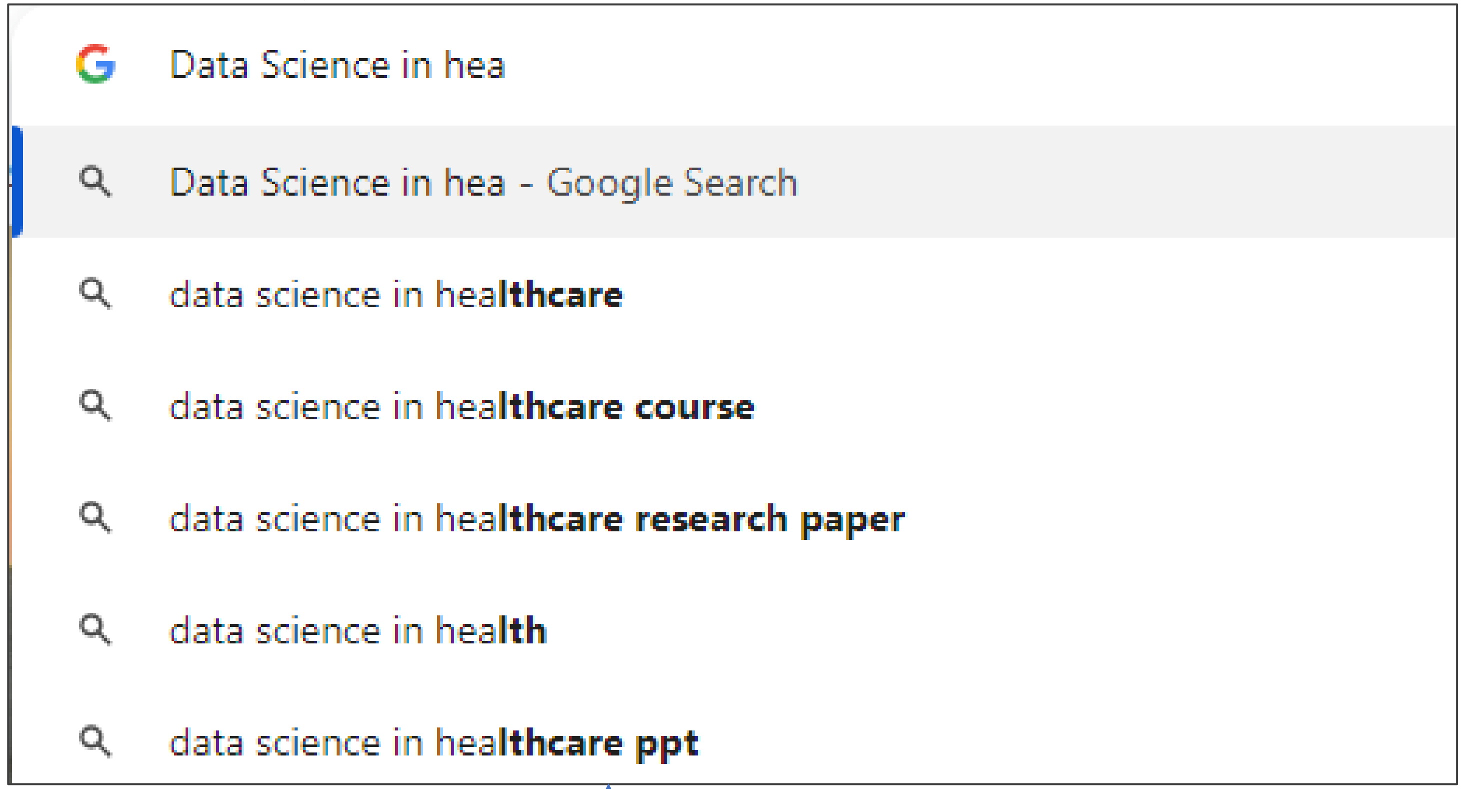
Engagement dashboard

Make informed decisions

Application of Data Science: Search Engines

Google uses data science to provide relevant search recommendations as the user types a query.

Search keyword



Data Science in hea

Data Science in hea - Google Search

data science in hea**lthcare**

data science in hea**lthcare course**

data science in hea**lthcare research paper**

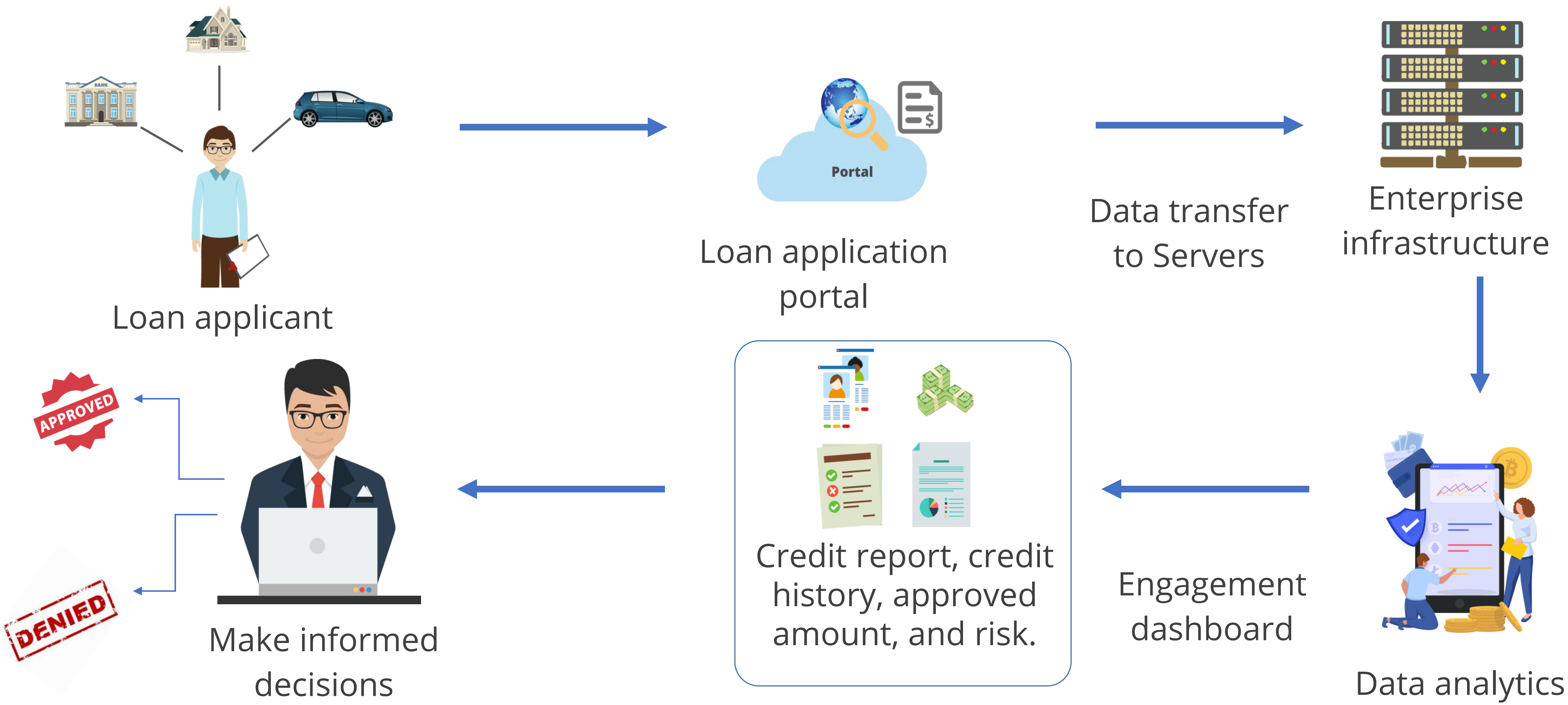
data science in hea**lth**

data science in hea**lthcare ppt**

Fast and real-time analytics is made possible by modern and advanced infrastructure, tools, and technologies.

Application of Data Science: Finance

A loan manager can easily access and sift through a loan applicant’s financial details using data science.



Loan applicant

Loan application portal

Data transfer to Servers

Enterprise infrastructure

Data analytics

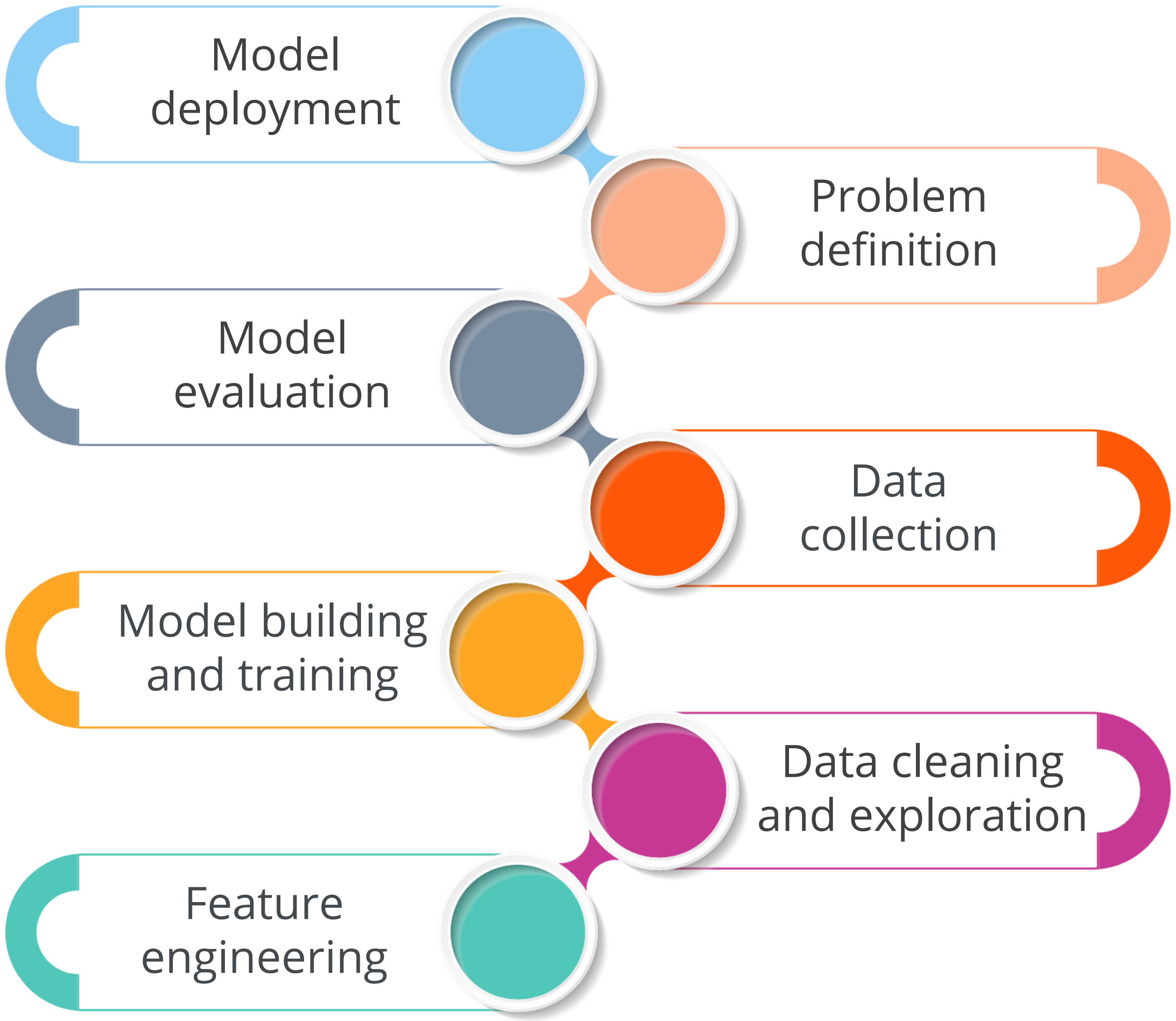
Engagement dashboard

Credit report, credit history, approved amount, and risk.

Make informed decisions

Data Science Process

Data Science Process



Model deployment

Problem definition

Model evaluation

Data collection

Model building and training

Data cleaning and exploration

Feature engineering

Data Science Process

Problem definition:

Clearly define the goal or question to be addressed through data analysis, forming the foundation for subsequent steps.

Data collection:

Gather relevant datasets or information sources necessary to address the defined problem.

Data cleaning and exploration:

Preprocess the data by handling missing values, outliers, and other inconsistencies, and explore the dataset to gain insights and identify patterns.

Data Science Process

Feature engineering:

Create or transform new features to enhance the dataset's information and improve model performance.

Model building and training:

Develop a predictive or descriptive model using machine learning algorithms and train it on the prepared dataset.

Model evaluation and deployment:

Evaluate, optimize, and fine-tune the model for peak performance, then deploy it into a production environment for real-world use.

Python for Data Science

Python for Data Science

Python is the preferred programming language for data science projects across industries.

It has multiple open-source packages like NumPy and Pandas for data cleaning, exploration, and visualization.

Advantages of Python for Data Science

- Open-source, interpreted, high-level language that’s great for object-oriented programming.
- Ease of use and simple syntax
- Scalability when compared to R
- Availability of a wide variety of data science libraries and packages
- Compatibility with all major operating systems
- Creation of new data science libraries daily by a vast number of online user communities.
- Powerful visualization libraries

Python Packages for Data Science

NumPy

NumPy (Numerical Python) is an open-source library, predominantly used when working with arrays.

It enables most of the operations required in linear algebra.

It uses arrays instead of typical Python lists, which makes it more computationally efficient.

It is used with SciPy and Matplotlib and has replaced Matlab as the industry standard for technical and engineering calculation.

Pandas

Pandas is an open-source library built on top of NumPy and is used for data manipulation.

The word Pandas is derived from panel data, a term from econometrics.

It can be used with NumPy to analyze and manipulate data.

It allows working with tabular data, time series data, and matrix data.

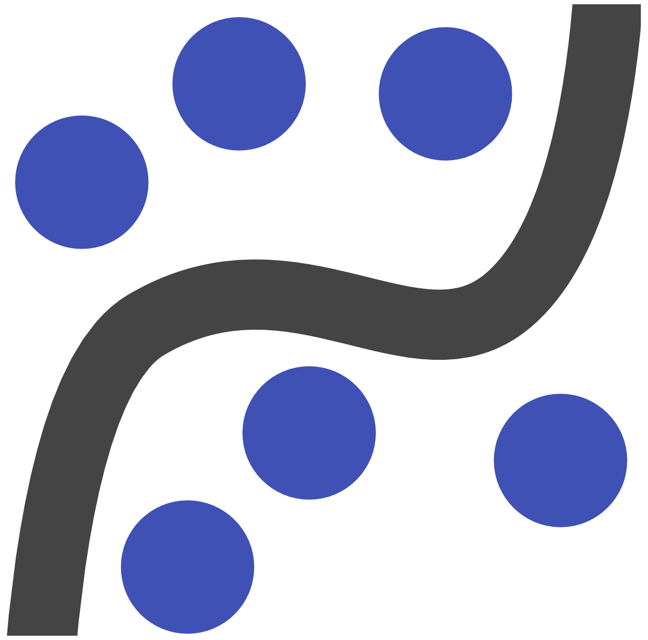
SciPy

SciPy (Scientific Python) is an open-source library built on top of NumPy and is used for implementing scientific formulas.

It is tailored for scientific and engineering applications.

Statsmodels

It is an important statistical analysis library that:



Allows the estimation of statistical models and performs statistical tests

Covers descriptive statistics, statistical tests, plotting functions, and so on.

Is capable of handling deep statistical research projects

Scikit-Learn

Scikit-learn is a popular open-source machine-learning library for Python. It's known for its simplicity and ease of use, as well as its broad applicability for various machine learning tasks.

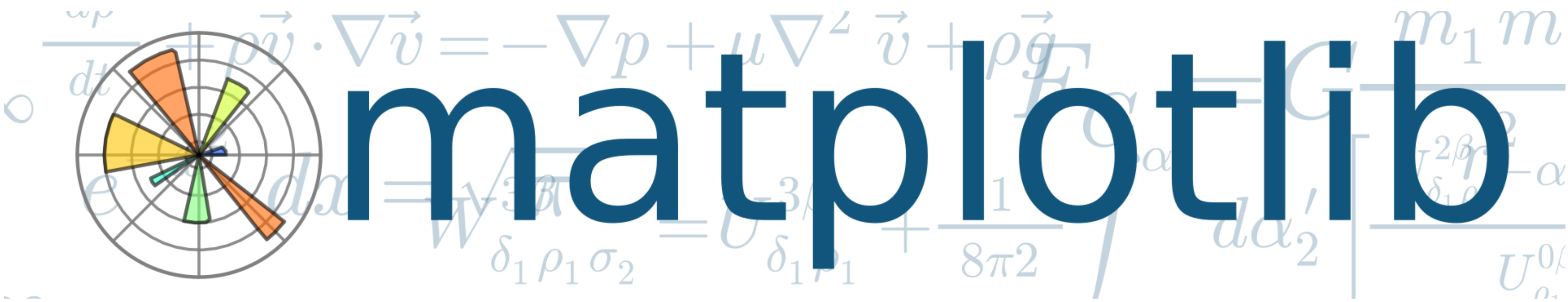
Allows many tools to identify, organize, and solve real-life problems

Provides a collection of free downloadable datasets

Consists of many libraries to learn and predict

Matplotlib

Matplotlib library is a comprehensive tool for building static, animated, and interactive visualizations.



Matplotlib is an open-source library and can be used freely.

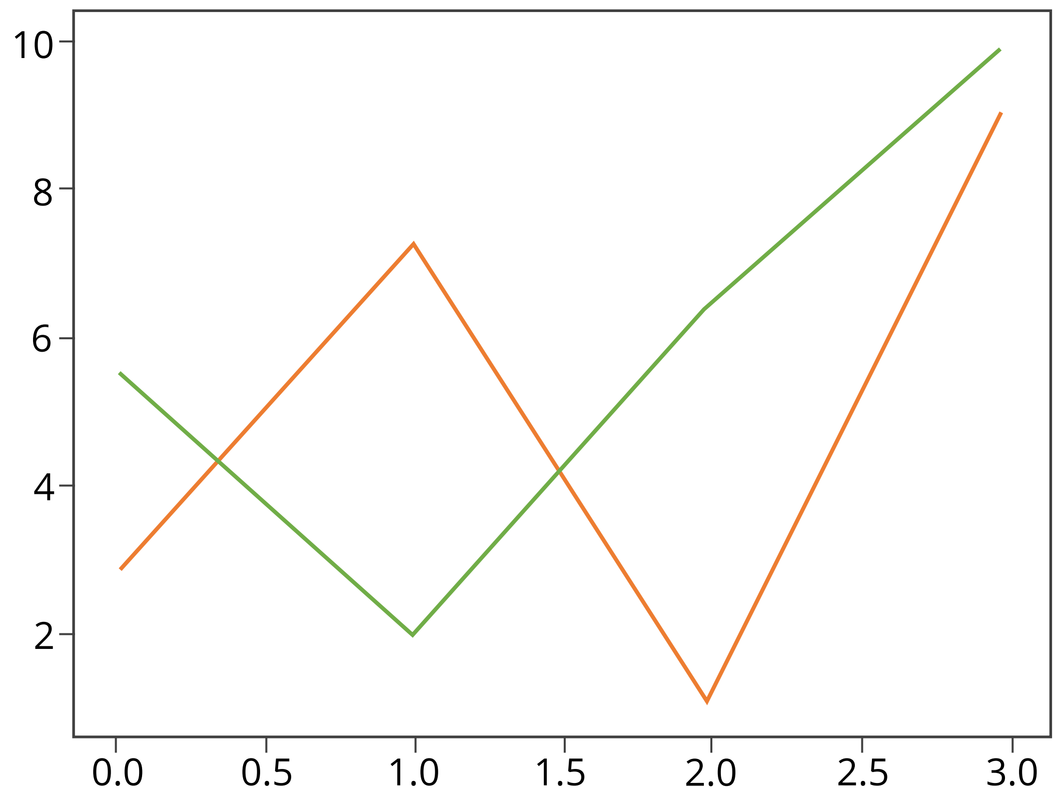
Seaborn

Seaborn is a data visualization library in Python that is built on top of Matplotlib.

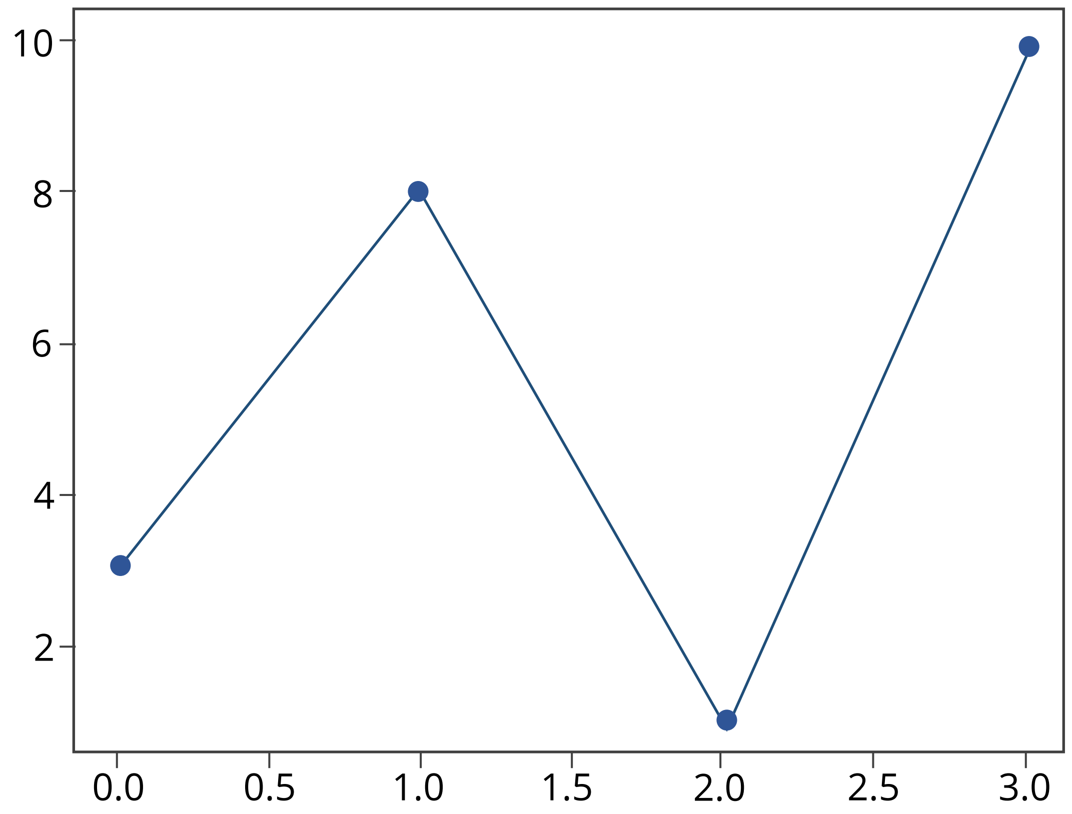
- It provides a high-level interface for creating attractive and informative statistical graphics.
- It simplifies the process of creating aesthetically pleasing and informative plots, especially for statistical and categorical data.

Types of Plots with Examples

Types of Plots: Line and Marker Plots

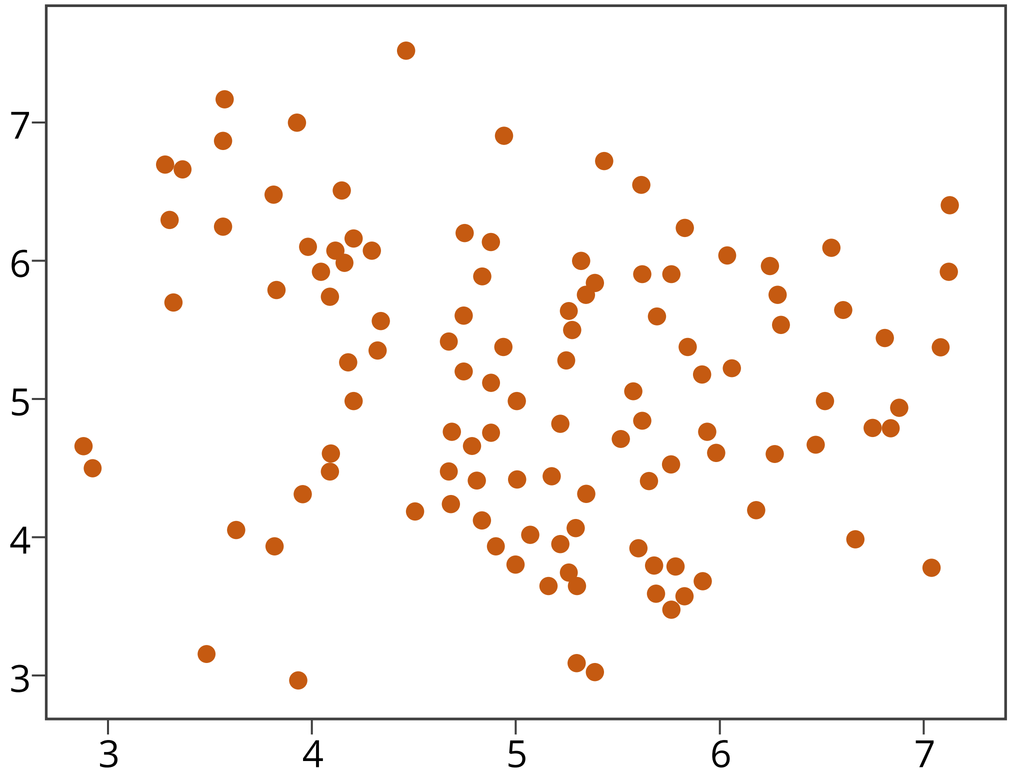


Line plots are created by connecting data points with straight lines where the x and y-axis values overlap.

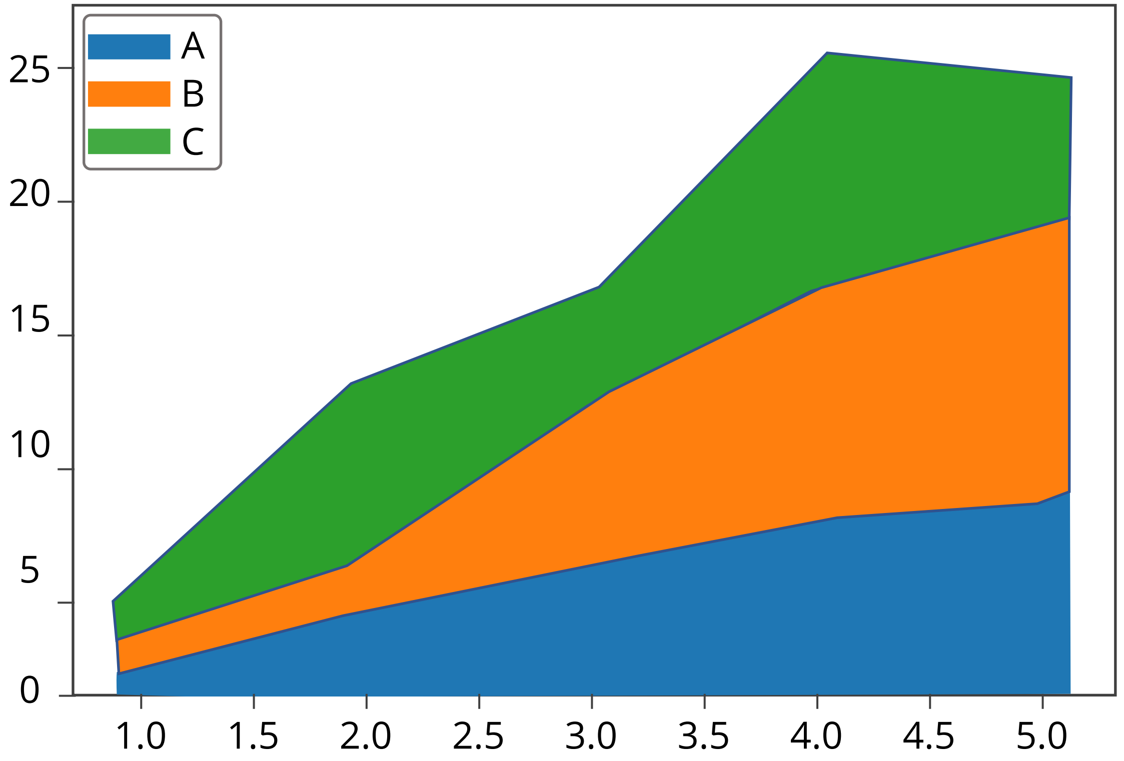


Marker plots are used in the Matplotlib library to simply enhance the visual of line size in a plot.

Types of Plots: Scatter and Area Plots

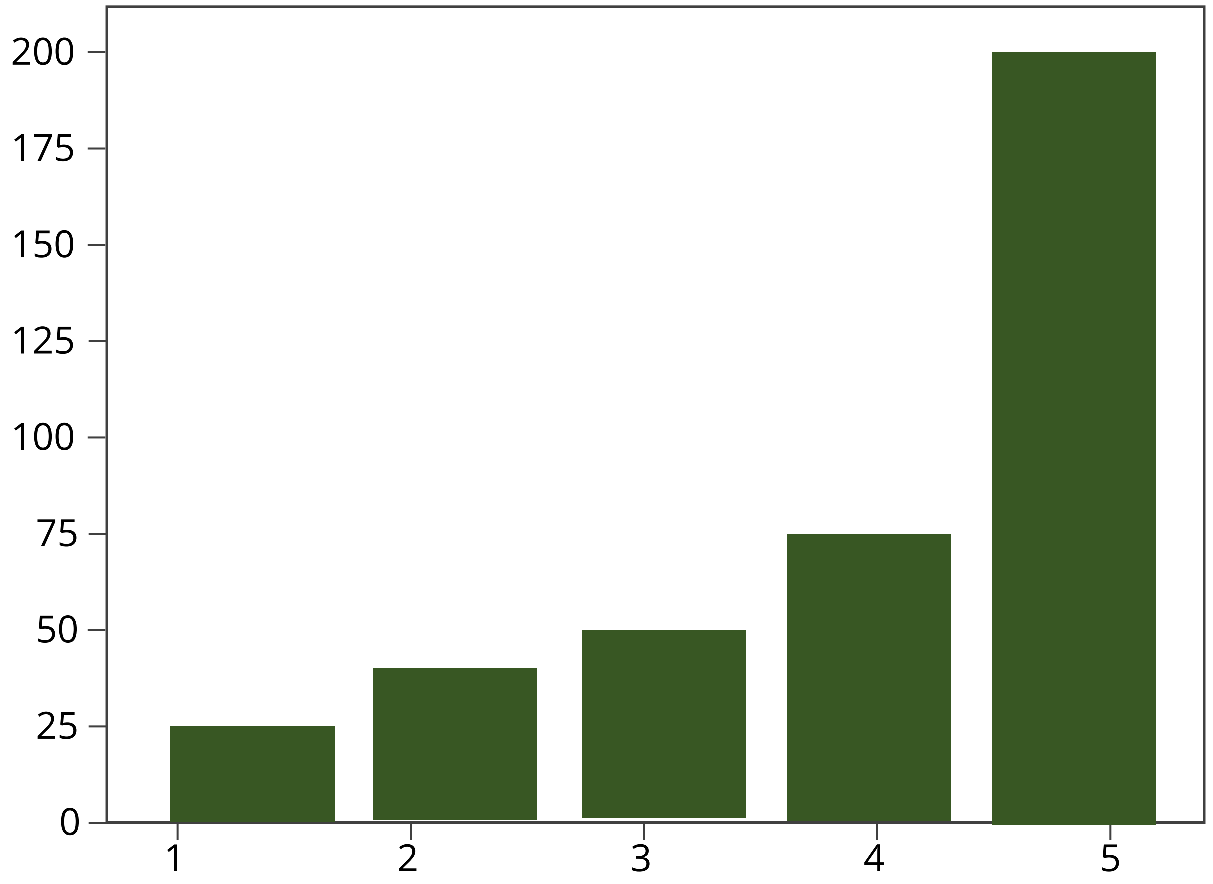


The scatter plot is a collection of points plotted on two axes, horizontal and vertical.

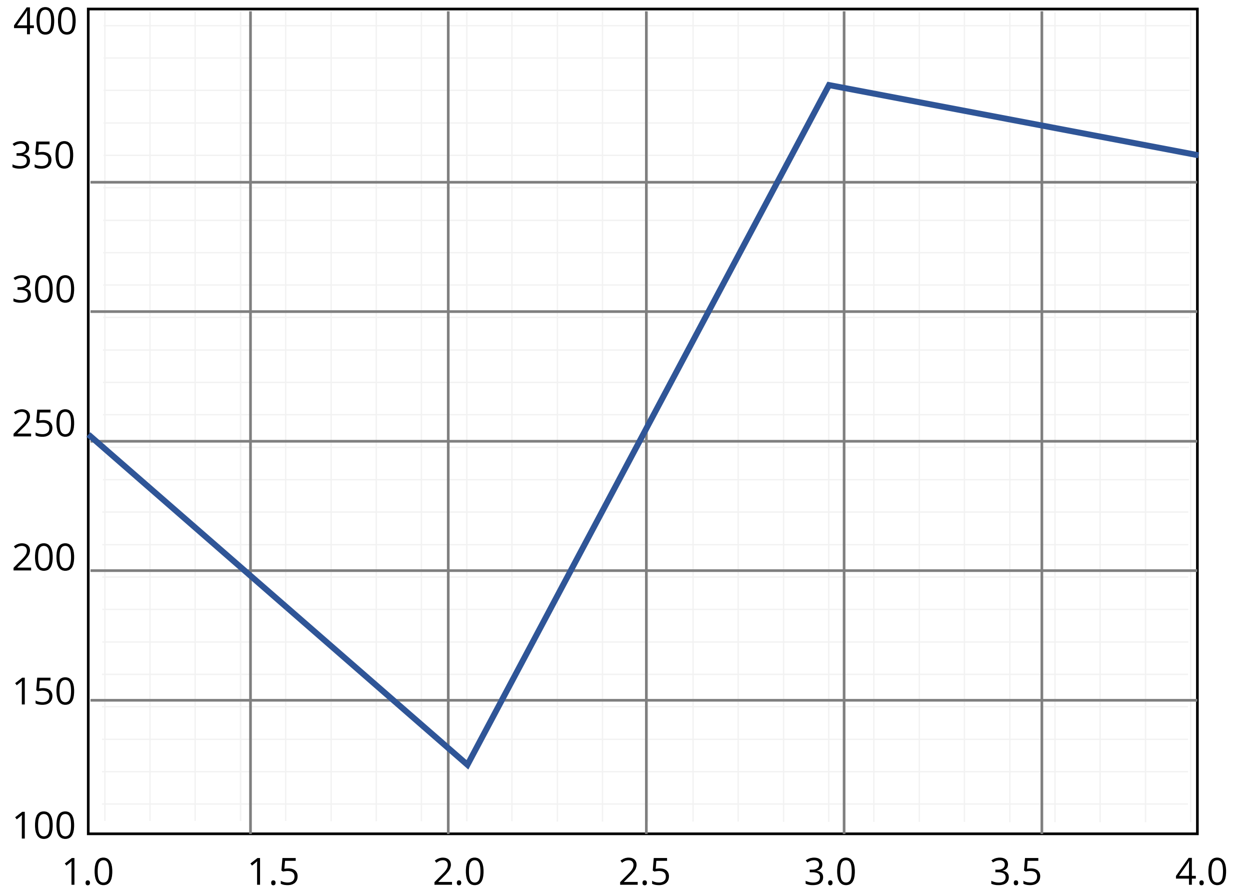


Area plots, also known as stack plots, are dispersed throughout certain areas with bumps and drops (highs and lows).

Types of Plots: Bar and Grid Plots

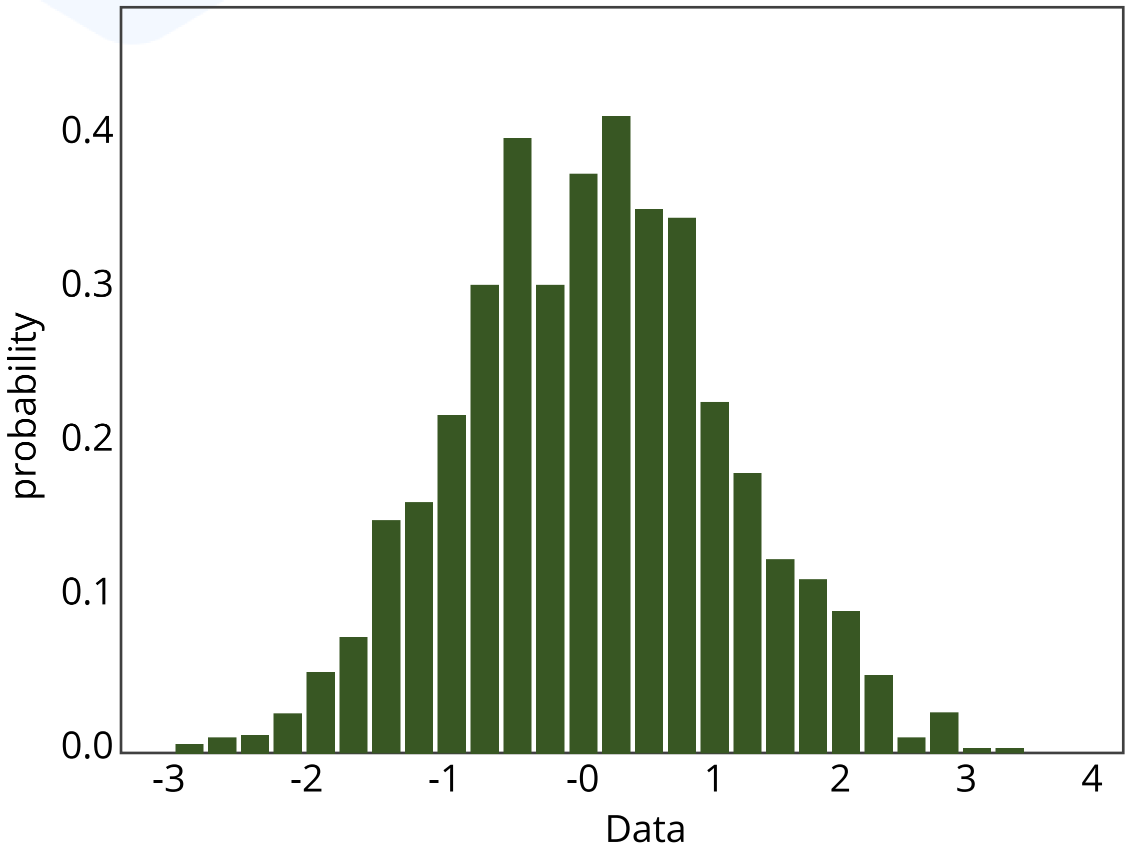


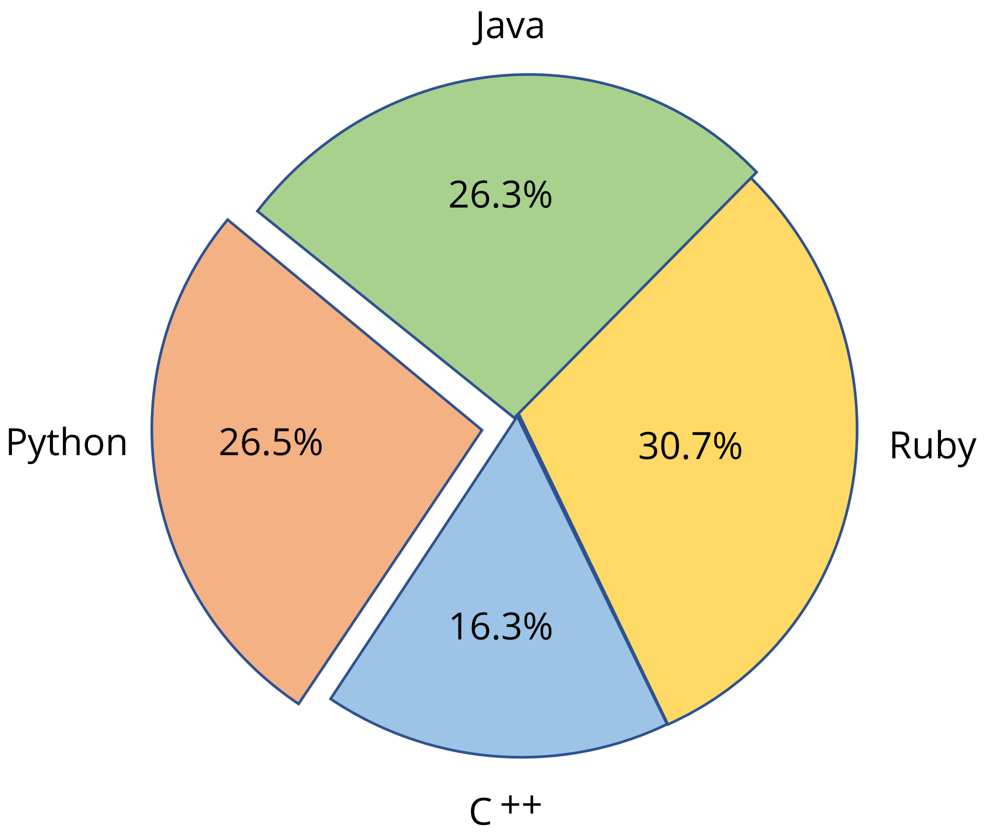
Bar plots are rectangular graphs that show vertical and horizontal data comparisons based on another axis (usually the X-axis).



Grid plots assist chart viewers in determining what value an unlabeled data point represents.

Types of Plots: Histogram and Pie Chart





A histogram visually displays the distribution of a dataset by dividing values into bins and representing the frequency of each bin with bars.

Pie charts are circular graphs in which data are plotted within components or segments of the pie.

Note: Examples of these plots are provided in the **Data Visualization** lesson, accompanied by detailed explanations and Python code.

Key Takeaways

- Data science involves the analysis and interpretation of data to generate actionable insights.
- NumPy (Numerical Python) is an open-source library, predominantly used when working with arrays.
- Seaborn is a data visualization library in Python that is built on top of Matplotlib.
- Python is the preferred programming language for data science projects across industries.

Knowledge Check

Knowledge Check

1

Which stage in the data science process involves preparing the data for modeling by addressing missing values, outliers, and data formatting?

A. Data collection

B. Data cleaning and exploration

C. Model building and training

D. Model evaluation

Knowledge Check

1

Which stage in the data science process involves preparing the data for modeling by addressing missing values, outliers, and data formatting?

A. Data collection

B. Data cleaning and exploration

C. Model building and training

D. Model evaluation

The correct answer is **B**

Data cleaning and exploration involves preparing the data for analysis by handling missing values, outliers, and ensuring proper data format.

Knowledge Check

2

What is the purpose of data cleaning and preparation in the data science process?

A. To increase the size of the dataset

B. To decrease the accuracy of the model

C. To ensure greater accuracy while building the model

D. To reduce the amount of data required for the model

Knowledge Check

2

What is the purpose of data cleaning and preparation in the data science process?

A. To increase the size of the dataset

B. To decrease the accuracy of the model

C. To ensure greater accuracy while building the model

D. To reduce the amount of data required for the model

The correct answer is **C**

Data cleaning and preparation are important steps in the data science process to ensure greater accuracy while building the model.

Knowledge Check 3	**What is another term for stack plots, featuring dispersed areas indicating highs and lows?**

A. Line plot

B. Scatter plot

C. Area plot

D. Box plot

Knowledge Check

3

What is another term for stack plots, featuring dispersed areas indicating highs and lows?

A. Line plot

B. Scatter plot

C. Area plot

D. Box plot

The correct answer is **C**

Area plots, also known as stack plots, are dispersed throughout certain areas with bumps and drops (highs and lows).

Thank You