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 to derive meaningful insights
- List the steps of the data science process to solve problems systematically
- Explore the Python packages for data science to efficiently perform data analysis, data manipulation, and data visualization
- Describe the types of plots available for visualization to communicate data insights and trends effectively

Introduction

Data Science

Data science is a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to derive meaningful insights from structured and unstructured data.

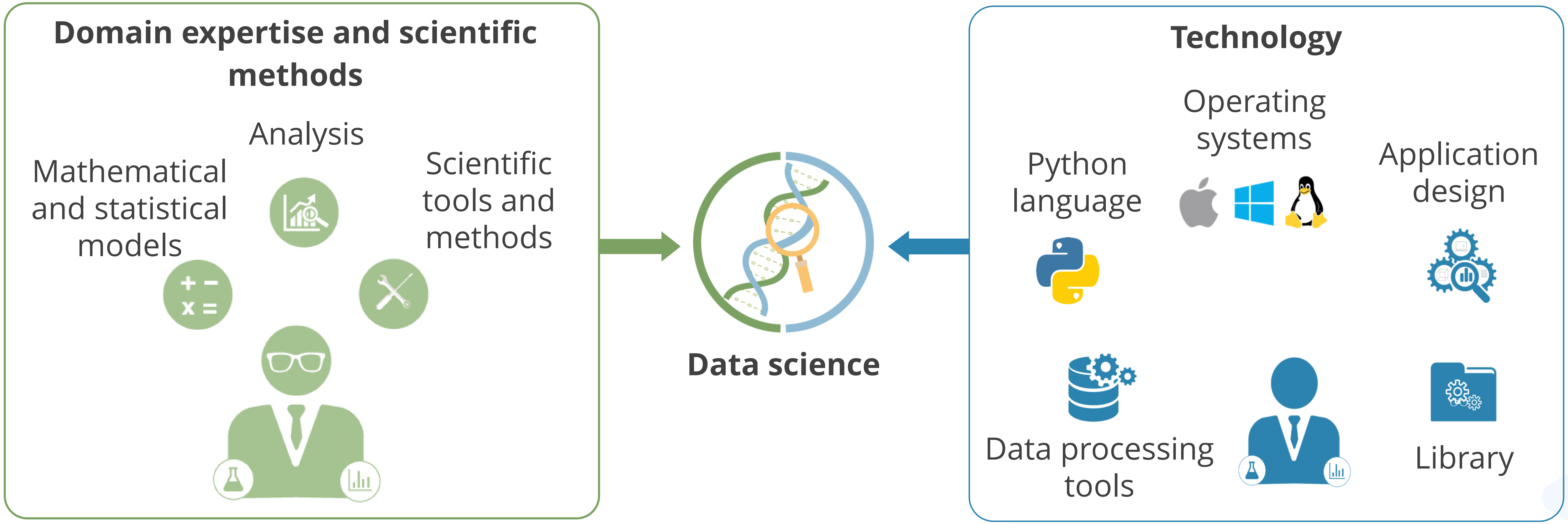
Example

Using a search engine or making a purchase on Amazon provides valuable data to data science-driven software systems operating
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 emerges from the combination of subject expertise, scientific methodologies, and technology.



Domain expertise and scientific methods

Analysis

Mathematical and statistical models

Scientific tools and methods

Data science

Technology

Operating systems

Python language

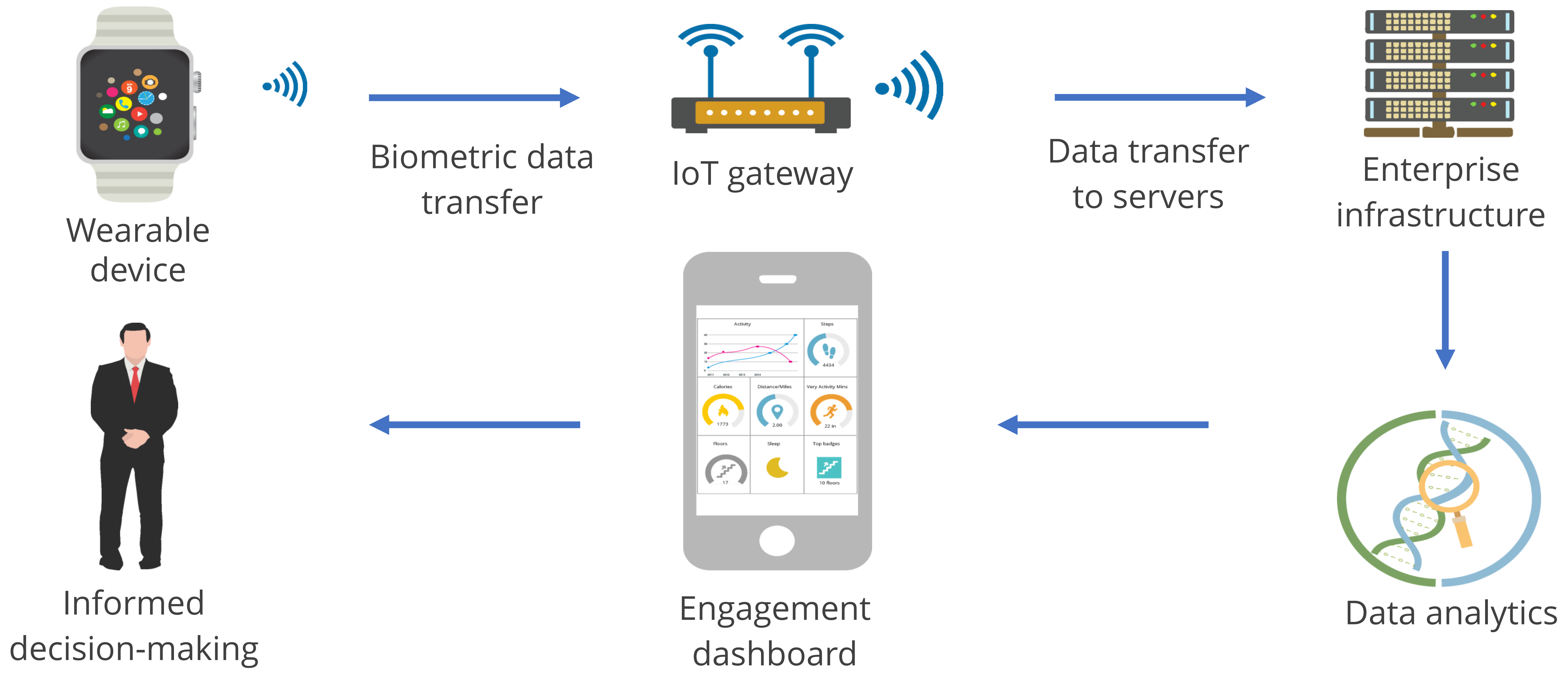
Application design

Data processing tools

Library

Application of Data Science: Healthcare

Wearable devices use data science to analyze data from their biometric sensors.



Wearable device

Biometric data transfer

IoT gateway

Data transfer to servers

Enterprise infrastructure

Data analytics

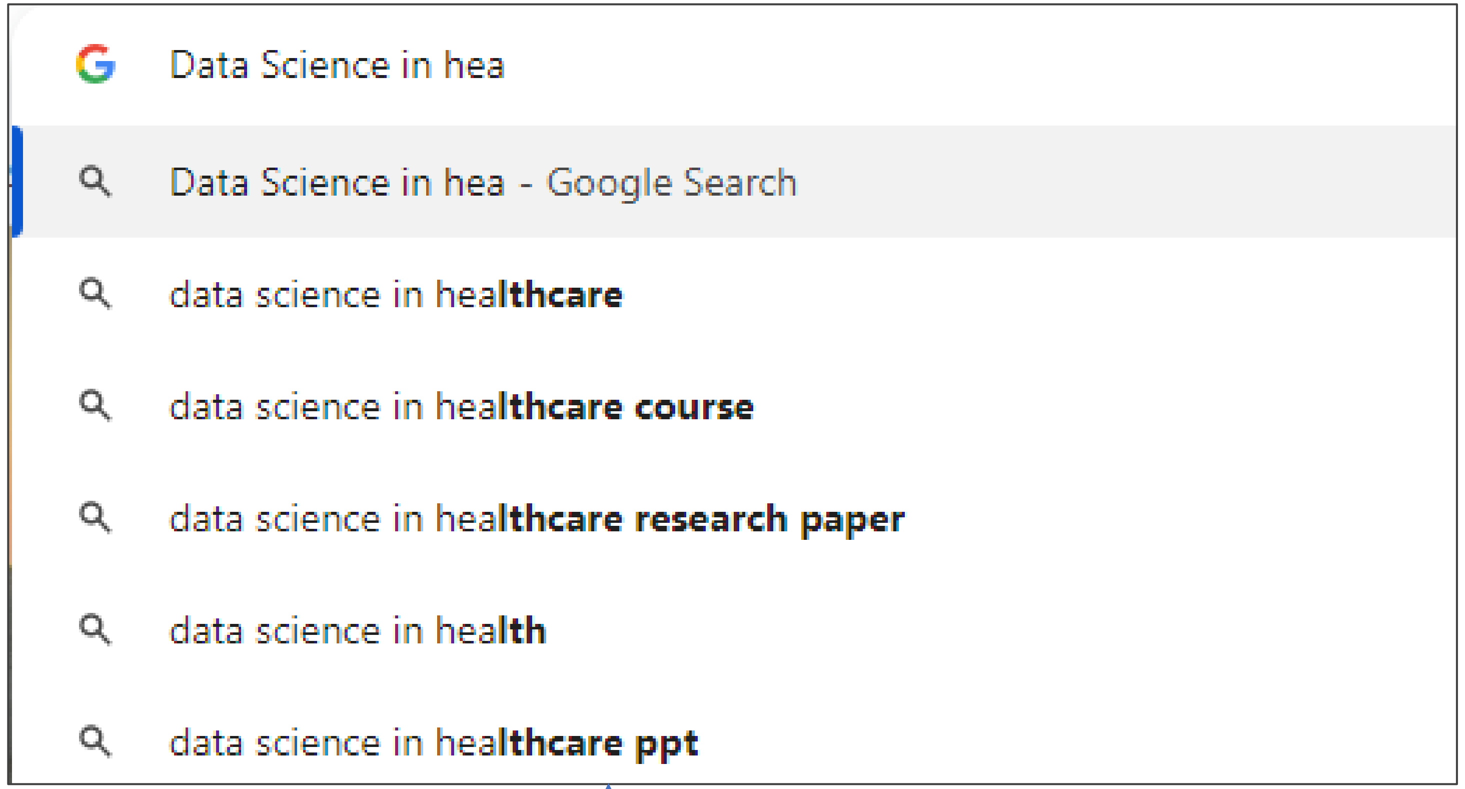
Engagement dashboard

Informed decision-making

Application of Data Science: Search Engines

Google employs data science to offer relevant search recommendations as users type in their queries.

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**

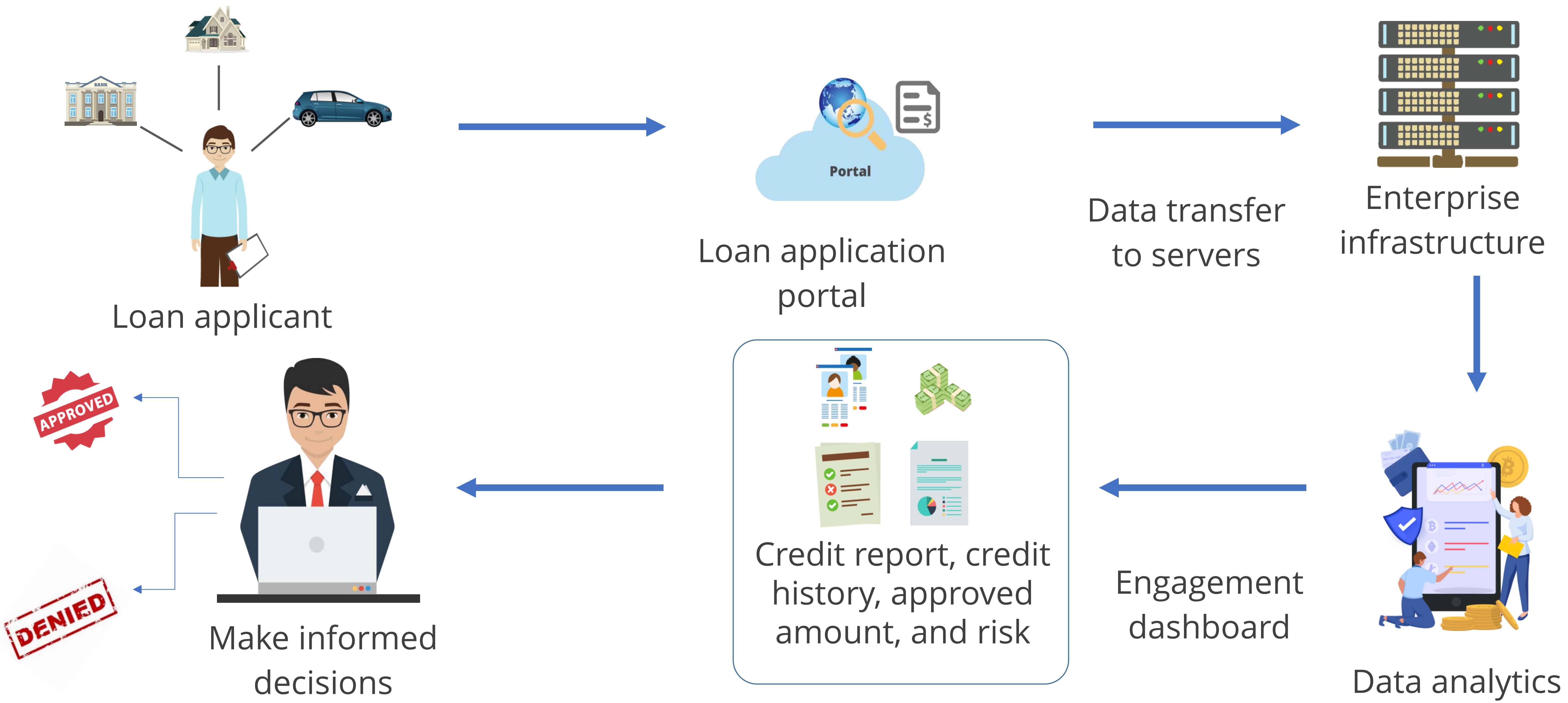
data science in hea**lth**

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

Using data science, a loan manager can easily access and sift through an applicant's financial details.



Loan applicant

Loan application portal

Data transfer to servers

Enterprise infrastructure

Data analytics

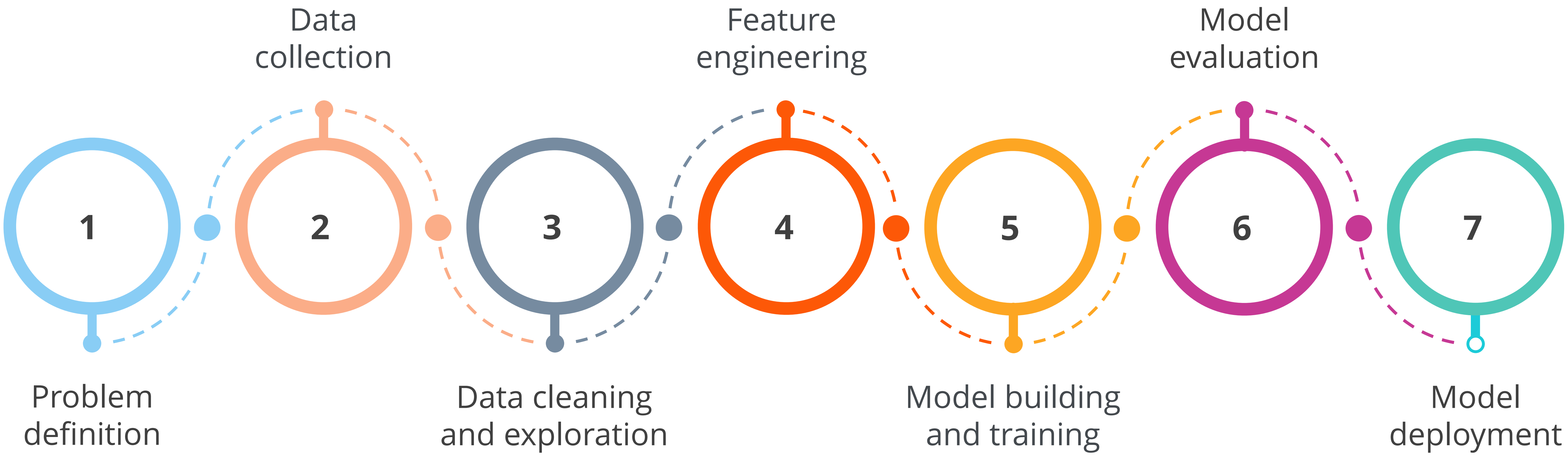
Engagement dashboard

Credit report, credit history, approved amount, and risk

Make informed decisions

Data Science Process

Data Science Process



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 Using Python for Data Science

- Open-source, interpreted, high-level language that supports 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 is a Python library for scientific computing that supports large multi-dimensional arrays and matrices and includes a comprehensive mathematical library.

Use NumPy to perform complex mathematical operations on large datasets, such as linear algebra calculations, statistical analysis, and Fourier transform.

Financial analysts use NumPy for quantitative analysis, such as calculating the mean return and volatility of stocks to inform investment decisions.

Pandas

Pandas is a library for efficient storage and manipulation of structured data, such as time series and tables.

pandas

E-commerce companies use Pandas to analyze customer purchase history to recommend products and personalize shopping experiences.

Researchers use Pandas to manage and analyze large datasets of health records to identify trends in disease outbreaks.

SciPy

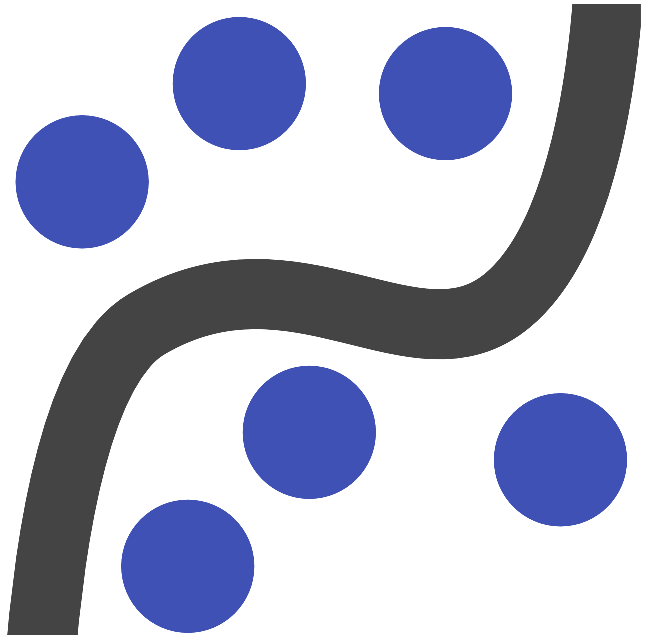
SciPy (Scientific Python) is an open-source library built on top of NumPy and is used for implementing scientific formulas like PV=nRT and V=IR.

It is tailored for scientific and engineering applications such as weather forecasting and drug discovery.

Engineers use SciPy to solve complex mathematical problems in structural engineering, such as stress and strain analysis in materials.

Statsmodels

Statsmodels is a Python module that provides classes and functions for estimating many different statistical models and conducting statistical data exploration.



Market researchers use Statsmodels to perform regression analysis to understand how different factors, such as advertising spend, affect sales.

Policy analysts use Statsmodels to evaluate the impact of public policies on social and economic outcomes through statistical testing and analysis.

Scikit-Learn

Scikit-learn is a widely-used open-source machine learning library for Python, known for its simplicity, ease of use, and versatility in handling various machine learning tasks.

It identifies objects in images for autonomous vehicles and facial recognition systems.

It detects fraudulent transactions in banking and e-commerce platforms.

It analyzes customer reviews for sentiment classification in marketing and social media analysis.

Matplotlib

Matplotlib library is a comprehensive tool for building static, animated, and interactive visualizations, including line plots, scatter plots, bar charts, histograms, pie charts, and more.

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, like histograms, box plots, violin plots, heatmaps, and error bars.
- It simplifies the process of creating aesthetically pleasing and informative plots, especially for statistical and categorical data.

Plotly

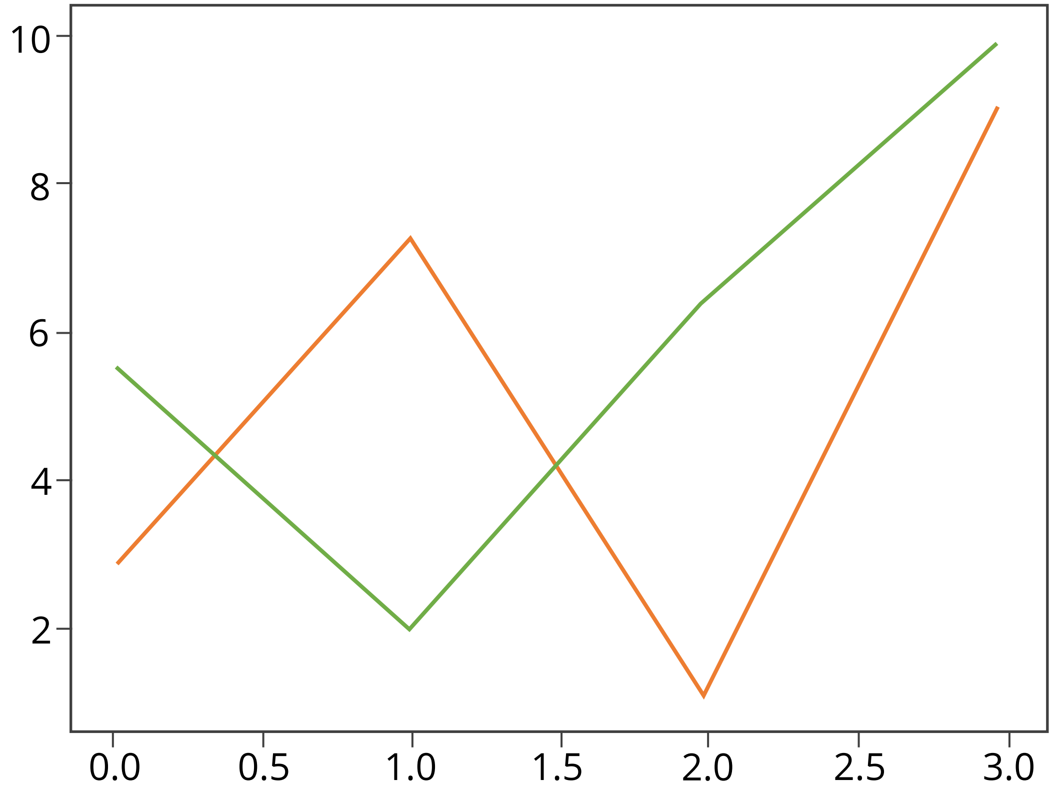
Plotly is a Python library for creating interactive, publication-quality graphs and visualizations. It is suitable for web-based applications.

- It enables Python users to create interactive and customizable visualizations for data analysis.
- It supports various plot types, such as line plots, scatter plots, and histograms, which enhance data exploration.

Types of Plots with Examples

Types of Plots: Line Plot

A line plot displays data points connected by straight lines, often used to visualize trends or relationships between two variables over time or other continuous intervals.

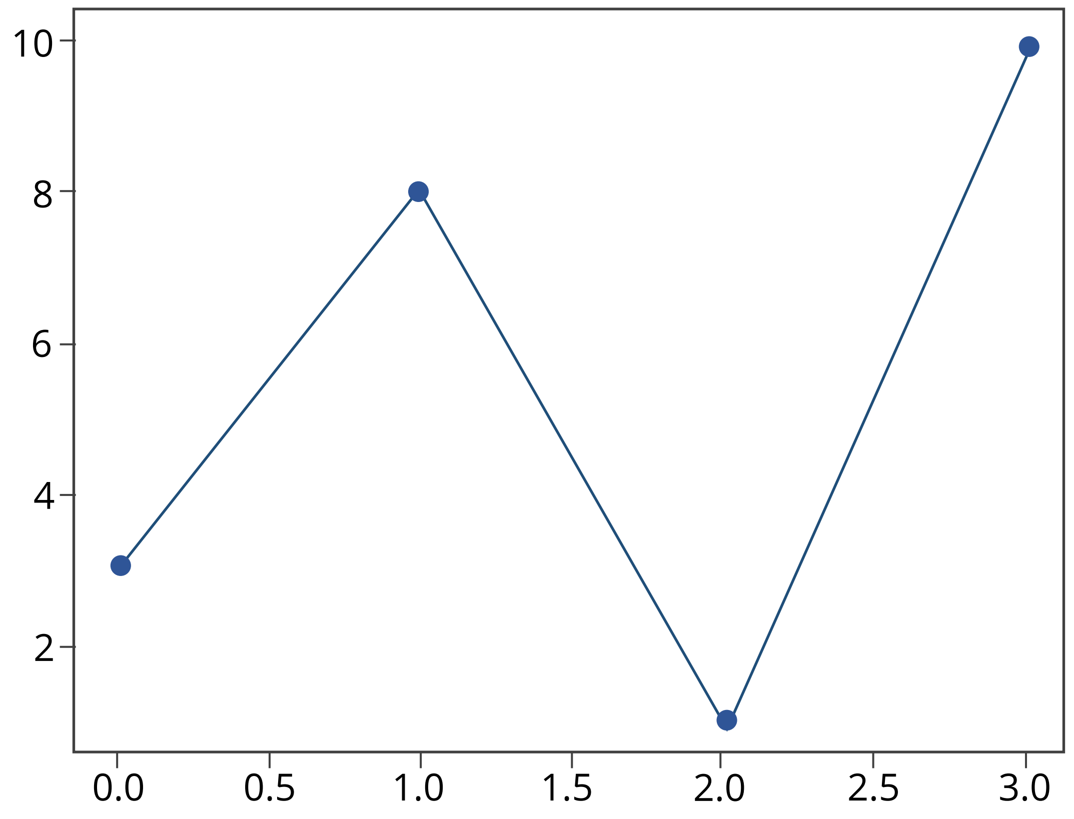


- A line chart visualizes stock prices over time, tracking trends and making investment decisions based on historical data.

- It displays temperature variations throughout the year to analyze seasonal patterns and plan agricultural activities effectively.

Types of Plots: Marker Plots

A marker plot displays data points with markers, useful for scatter plots and visualizing individual data observations.

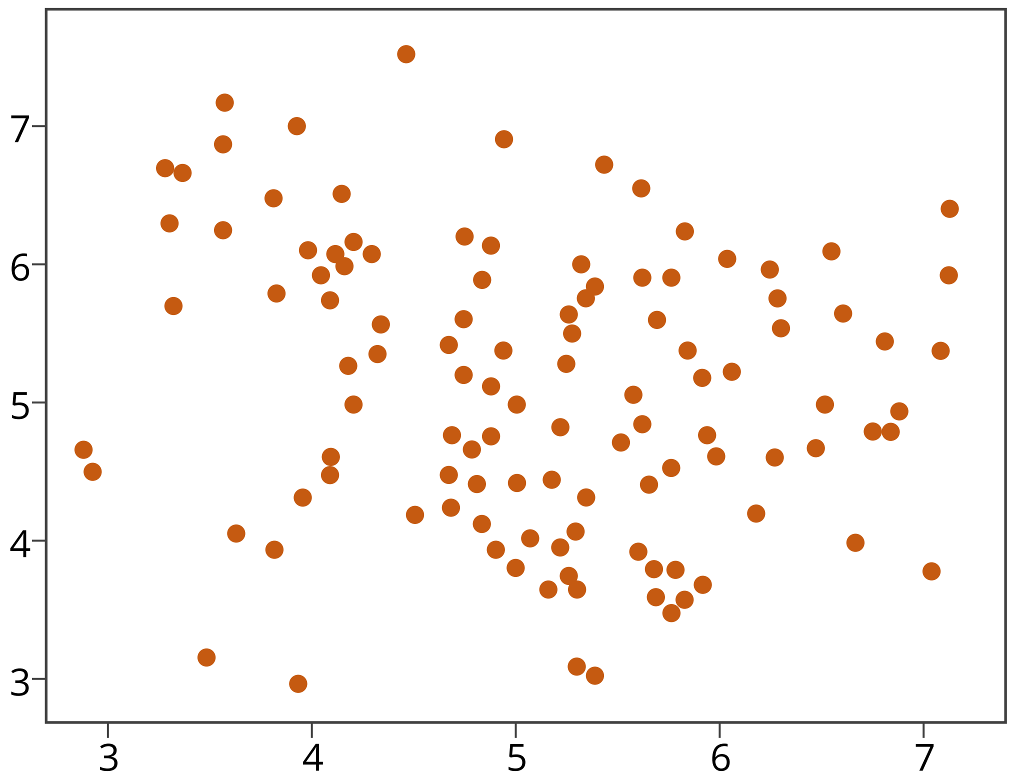


- A marker plot is used to display individual data points on a map, such as marking specific locations for a survey.

- It is employed to plot stock prices over time, with markers indicating specific events like buy or sell signals.

Types of Plots: Scatter Plots

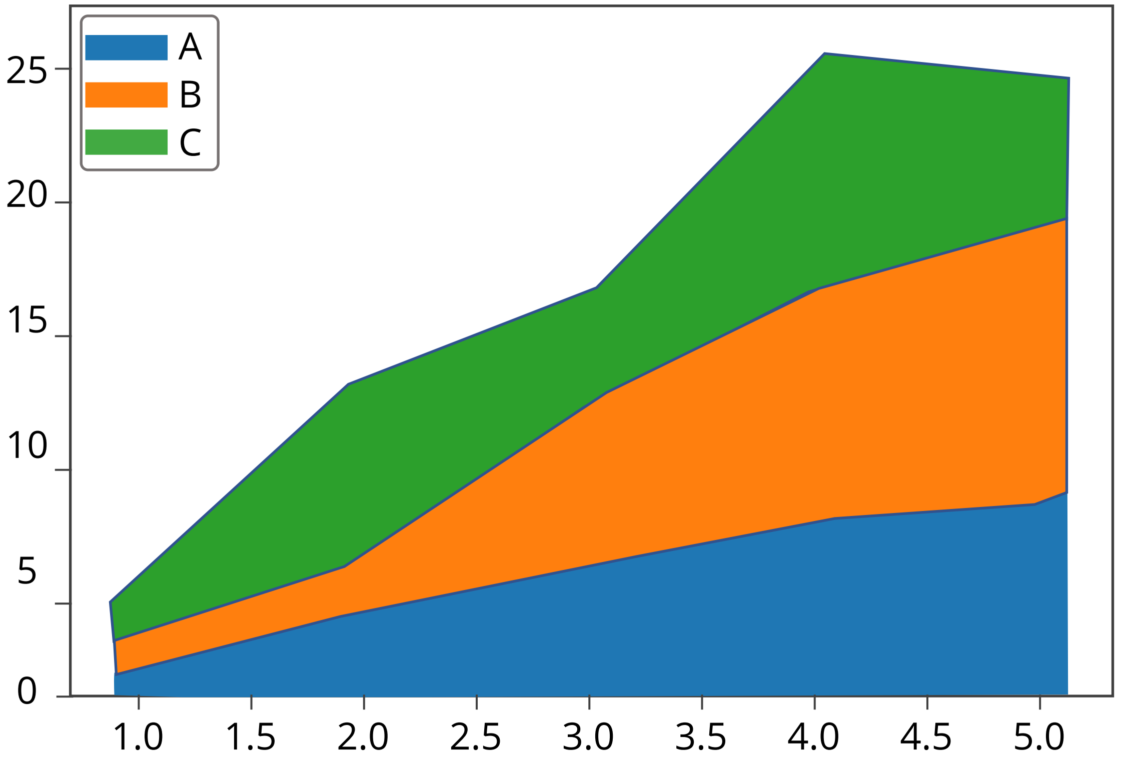
The scatter plot is a collection of points plotted on two axes, horizontal and vertical.



- The scatter plot analyzes relationships between two variables, like comparing height and weight in a population.
- It helps visualize data clusters, such as grouping students based on exam scores.

Types of Plots: Area Plots

An area plot represents data with shaded areas useful for showing cumulative totals or proportions over time.

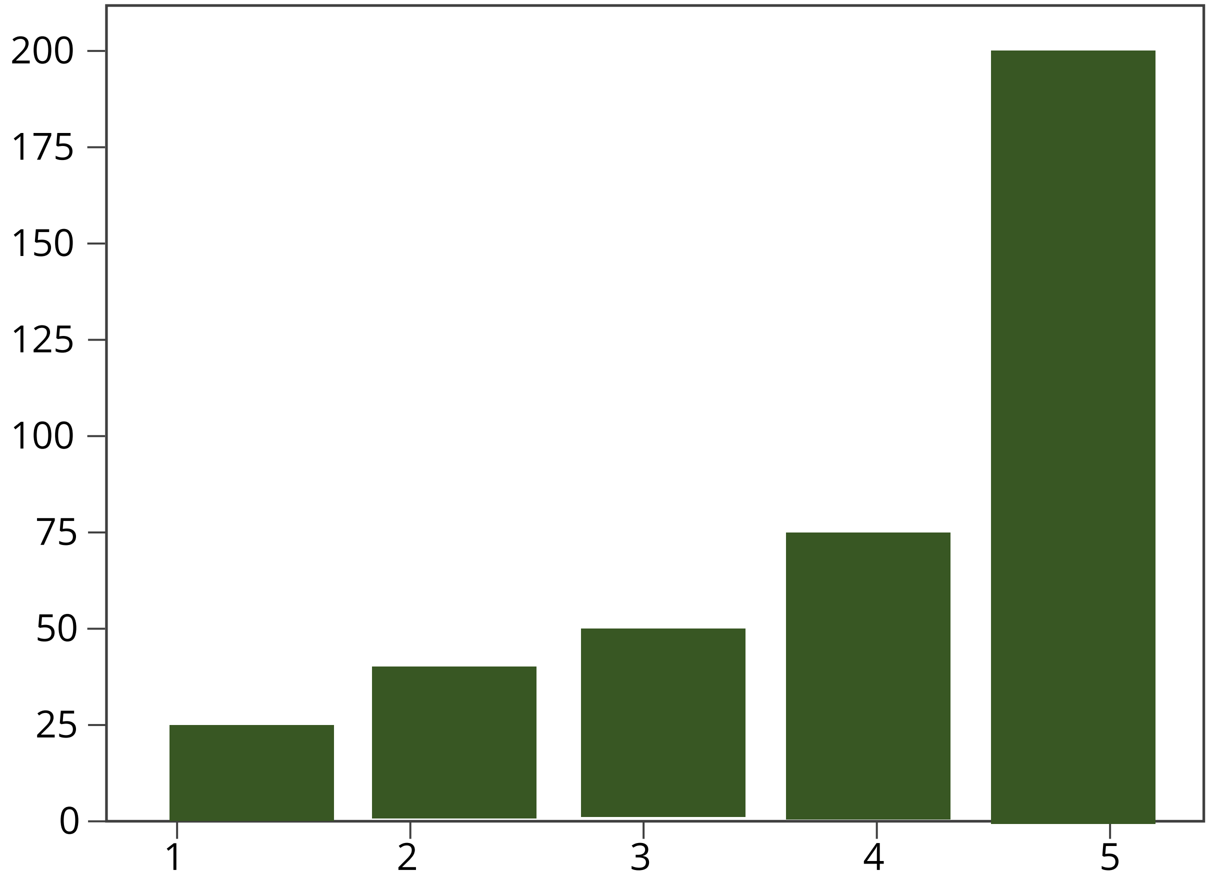


- An area plot visualizes cumulative data changes over time, such as tracking total sales revenue over successive quarters.

- It illustrates the distribution of different categories' contributions to a whole, like displaying market share evolution.

Types of Plots: Bar Plots

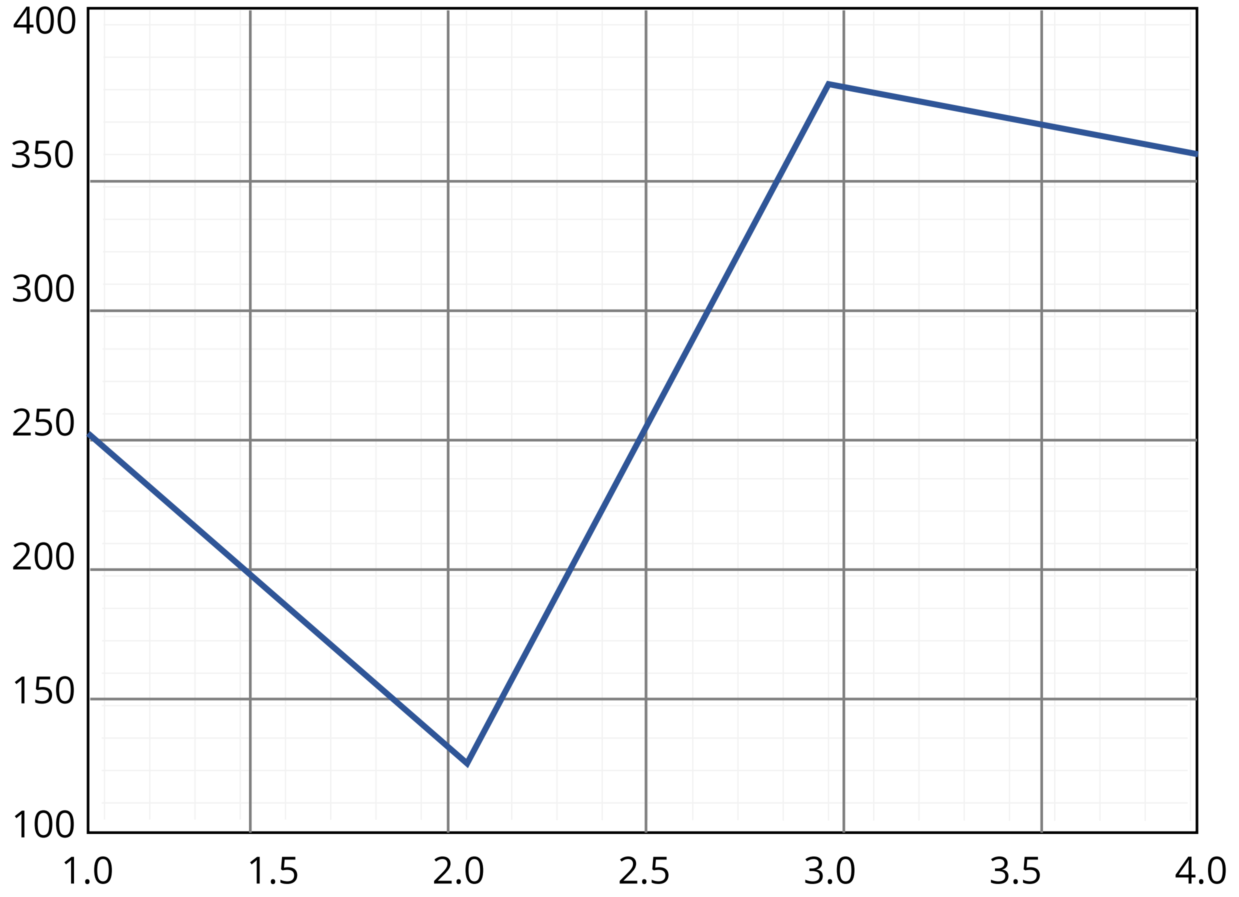
Bar plots are rectangular graphs that show vertical and horizontal data comparisons based on another axis (usually the X-axis).



- A bar plot compares sales of different products over a month.
- It also displays student grades in different subjects.

Types of Plots: Grid Plots

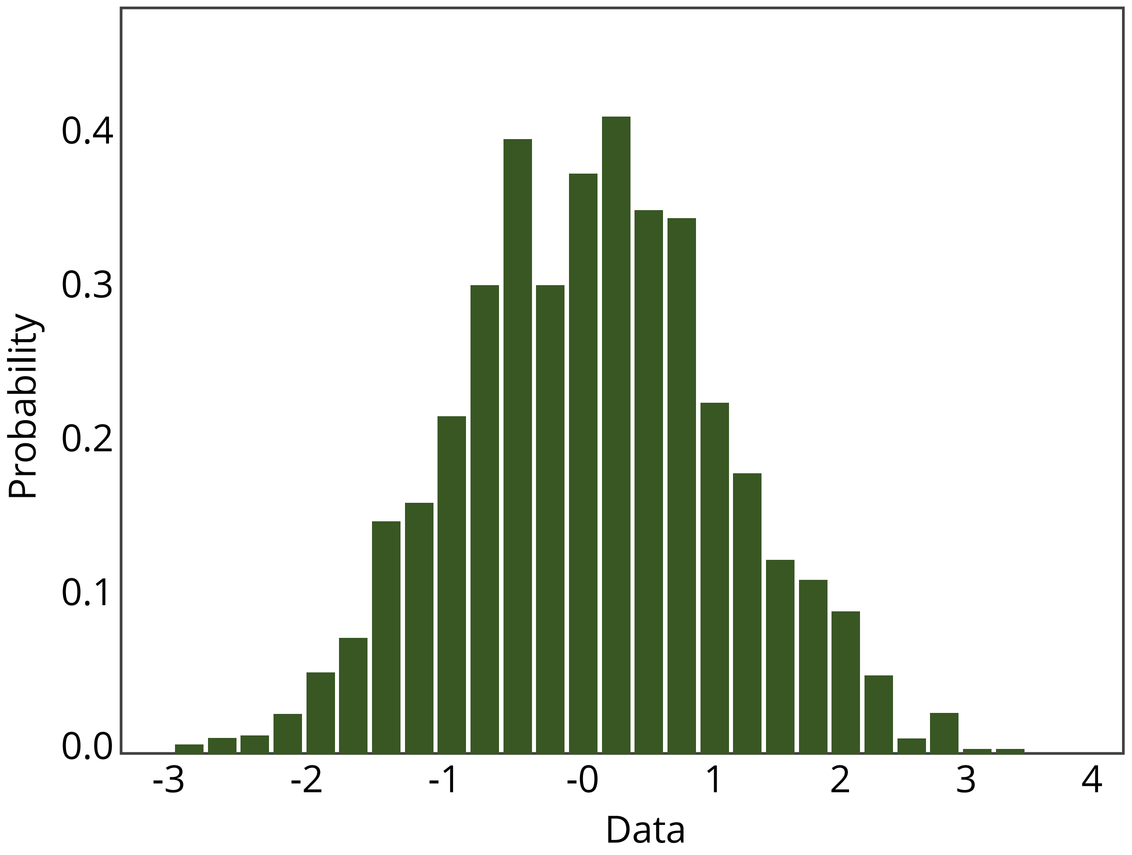
Grid plots assist chart viewers in determining what value an unlabelled data point represents.



- Grid plots enable side-by-side comparison of multiple plots, enhancing visual analysis.
- They enhance presentation clarity by organizing complex information systematically.

Types of Plots: Histogram Chart

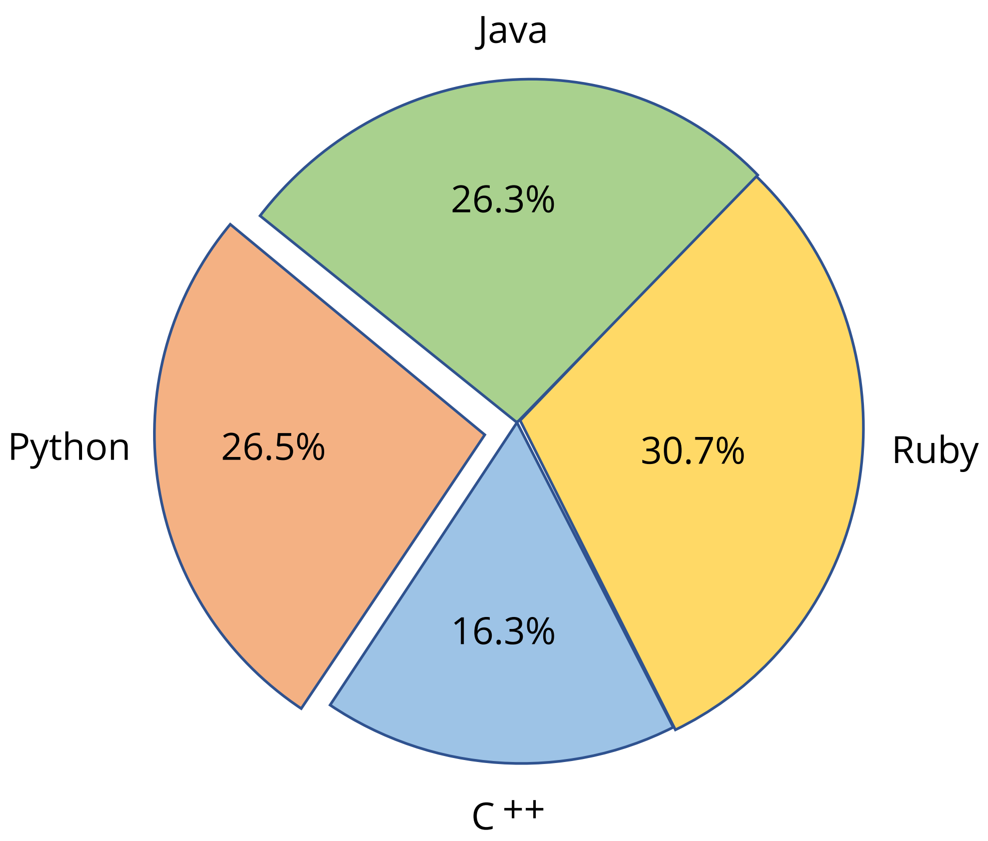
A histogram visually displays the distribution of a dataset by dividing values into bins and representing the frequency of each bin with bars.



- Histograms visualize the distribution of numerical data, like income levels or exam scores.
- They help make inferences about data characteristics and underlying patterns, guiding decision-making processes.

Types of Plots: Pie Chart

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



- Pie plots show the proportions of a whole, like market shares or survey responses.

- They simplify complex data by representing it in an easily understandable format.

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 data for analysis by addressing missing values, outliers, and 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