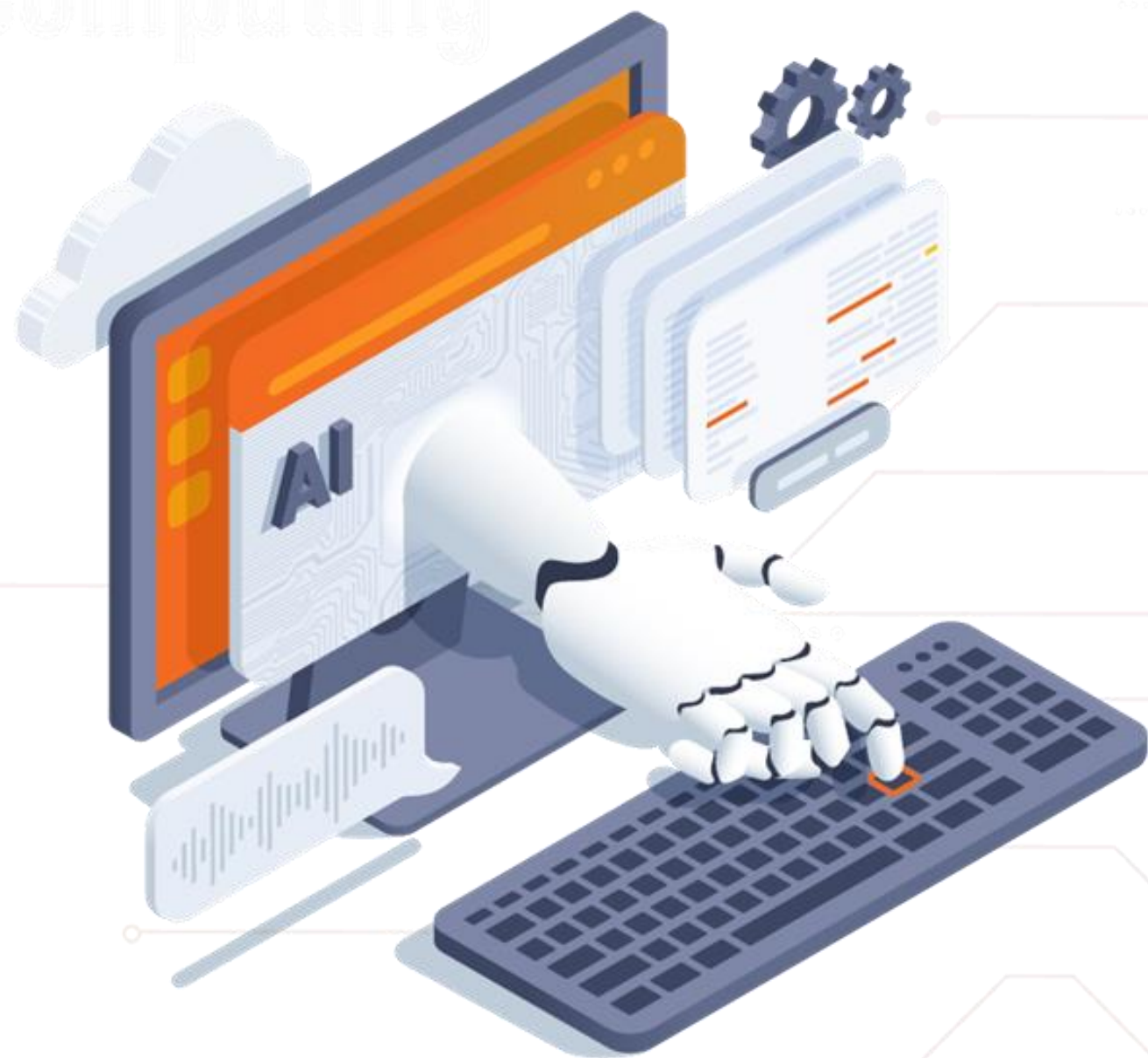


DATA AND
ARTIFICIAL INTELLIGENCE
Computing



Caltech

**Center for Technology &
Management Education**

Introduction to Data Science

Learning Objectives

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

- 👁 Define data science
- 👁 Discuss the roles and responsibilities of a data scientist
- 👁 List various applications of data science
- 👁 Explain data science and its importance



Data Science

What Is Data Science?

Some common definitions of Data Science are:

A powerful new approach to make discoveries from data

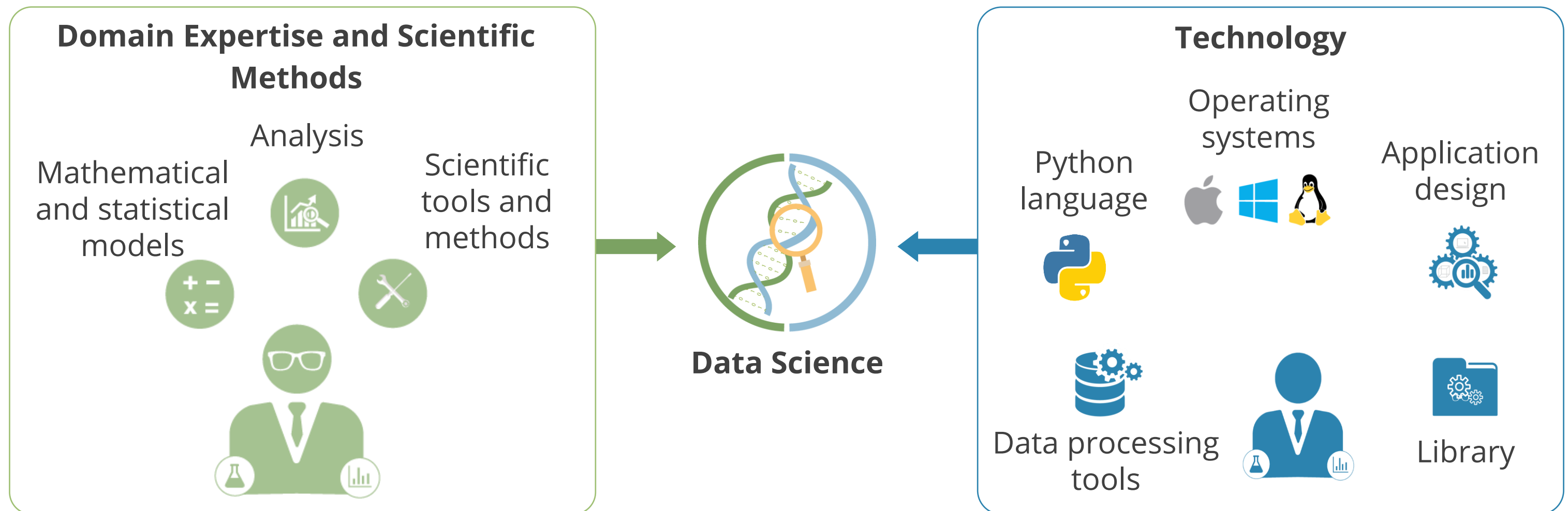


An automated way to analyze enormous amounts of data and extract information

A new discipline that combines the aspects of statistics, mathematics, programming, and visualization to turn data into information

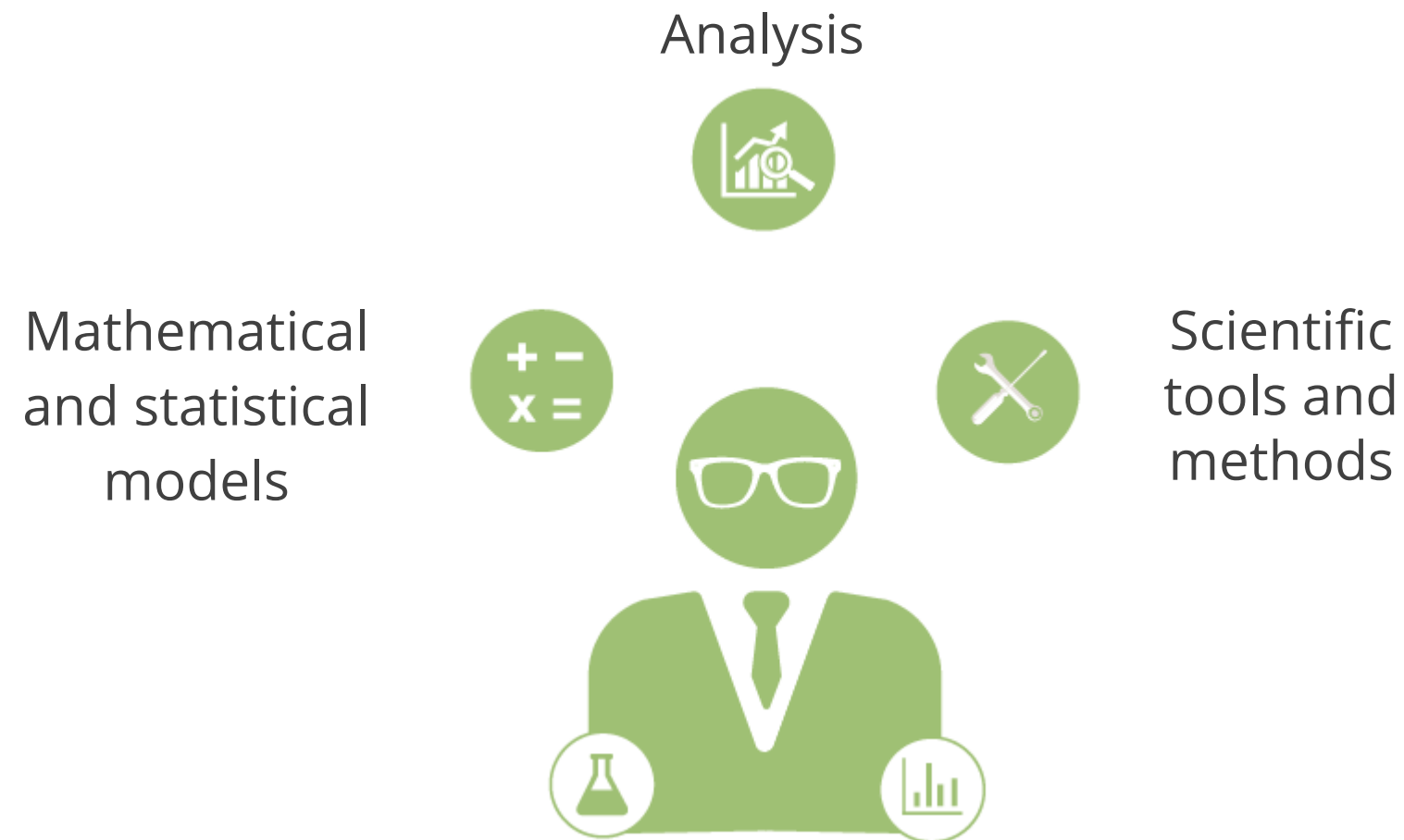
Components of Data Science

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



Domain Expertise and Scientific Methods

Data scientists collect, explore, analyze, and visualize data. They apply mathematical and statistical models to find patterns and solutions in the data.



Domain Expertise and Scientific Methods

Data analysis helps to extract insights from data to make better business decisions.

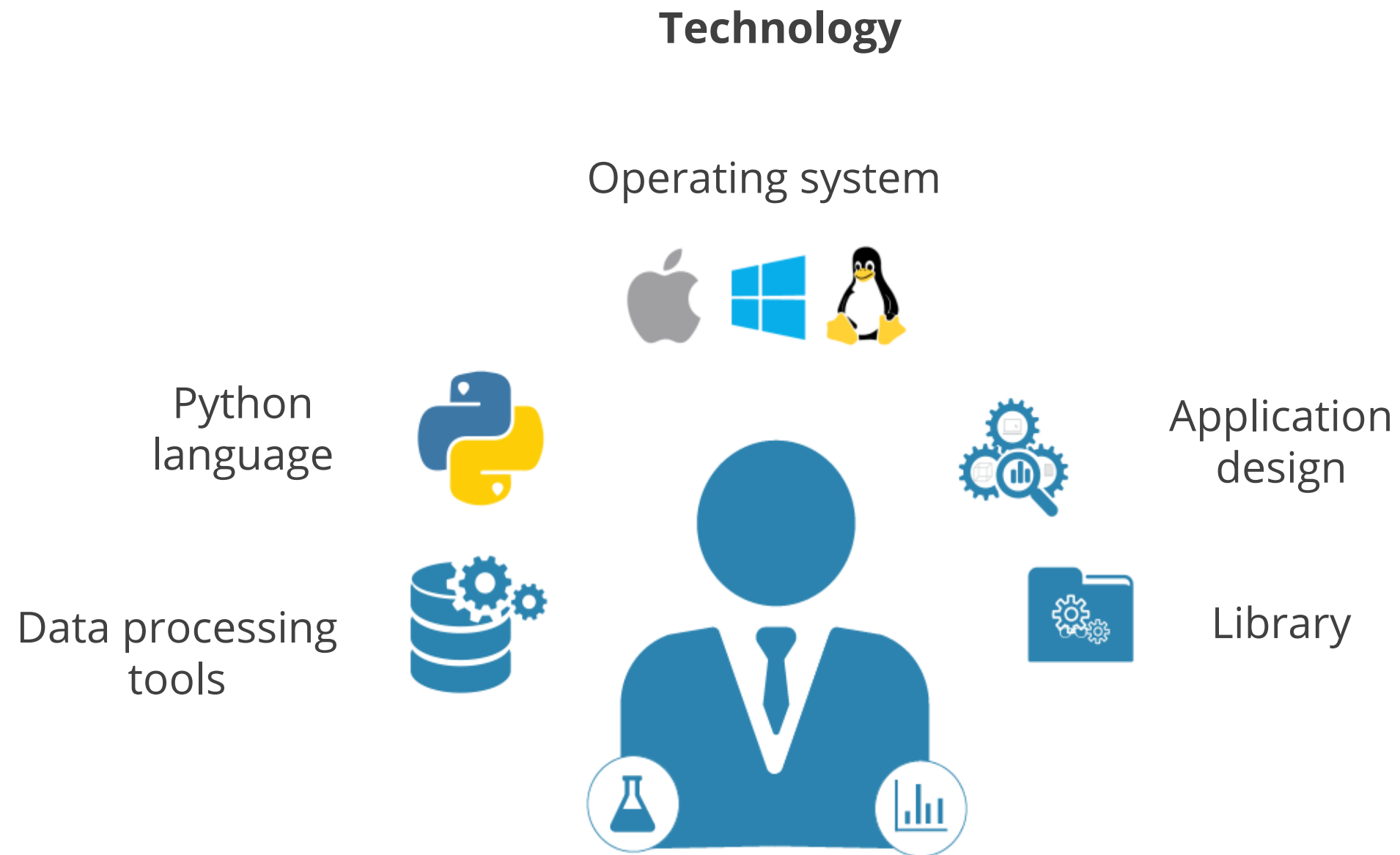


Data analysis can be:

- Descriptive: Study a dataset to decipher the details
- Predictive: Create a model based on existing information to predict outcome and behavior
- Prescriptive: Suggest actions for a given situation using the collected information

Data Processing and Analytics

Modern tools and technologies have made data processing and analytics faster and more efficient.



Data Processing and Analytics

These technologies help data scientists to:



- Build and train machine learning models
- Manipulate data with technology
- Build data tools, applications, and services
- Extract information from data

Applications of Data Science

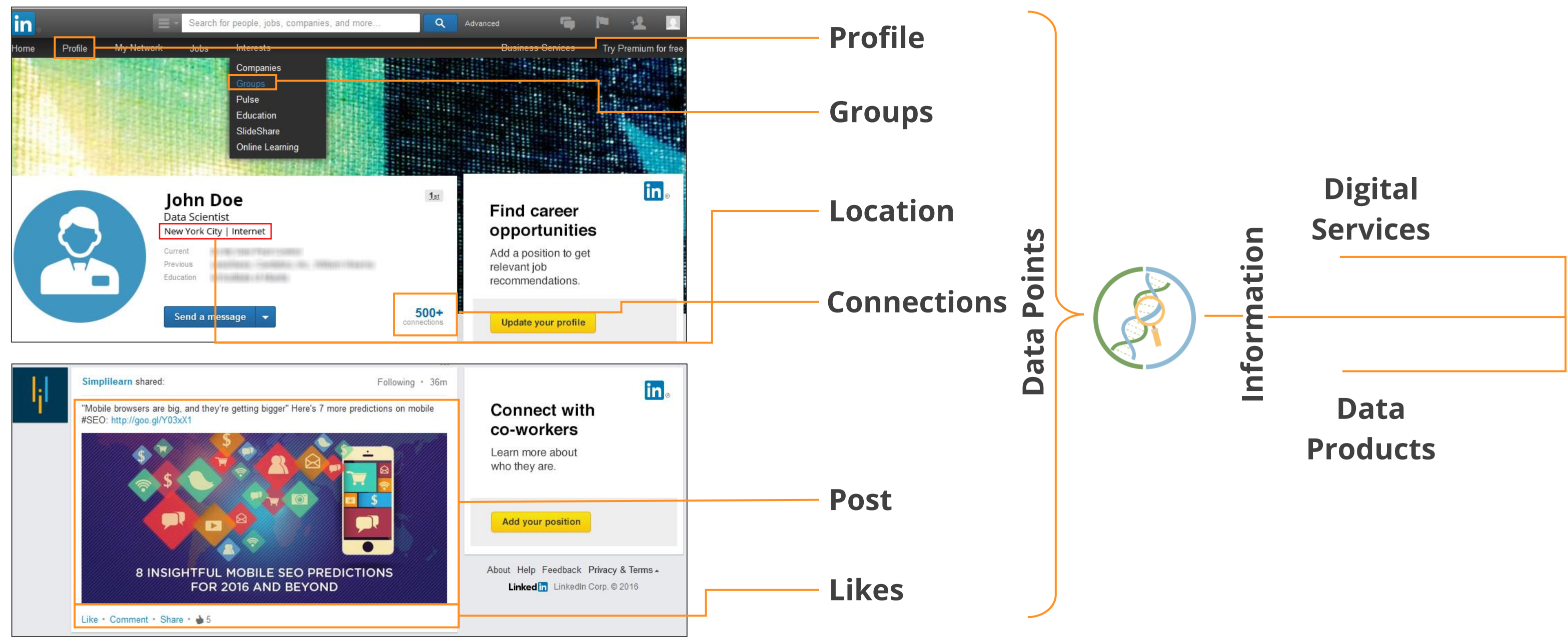
Different Sectors Using Data Science

Various sectors use data science to extract the information they need to create different services and products.



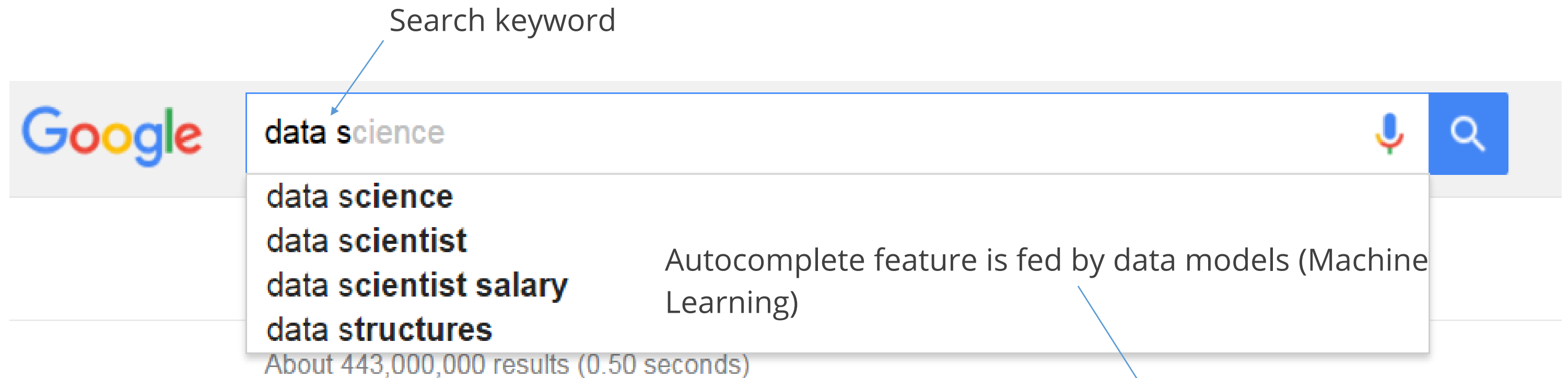
Using Data Science: Social Network Platforms

LinkedIn uses data points from its users to provide relevant digital services and data products.



Using Data Science: Search Engines

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



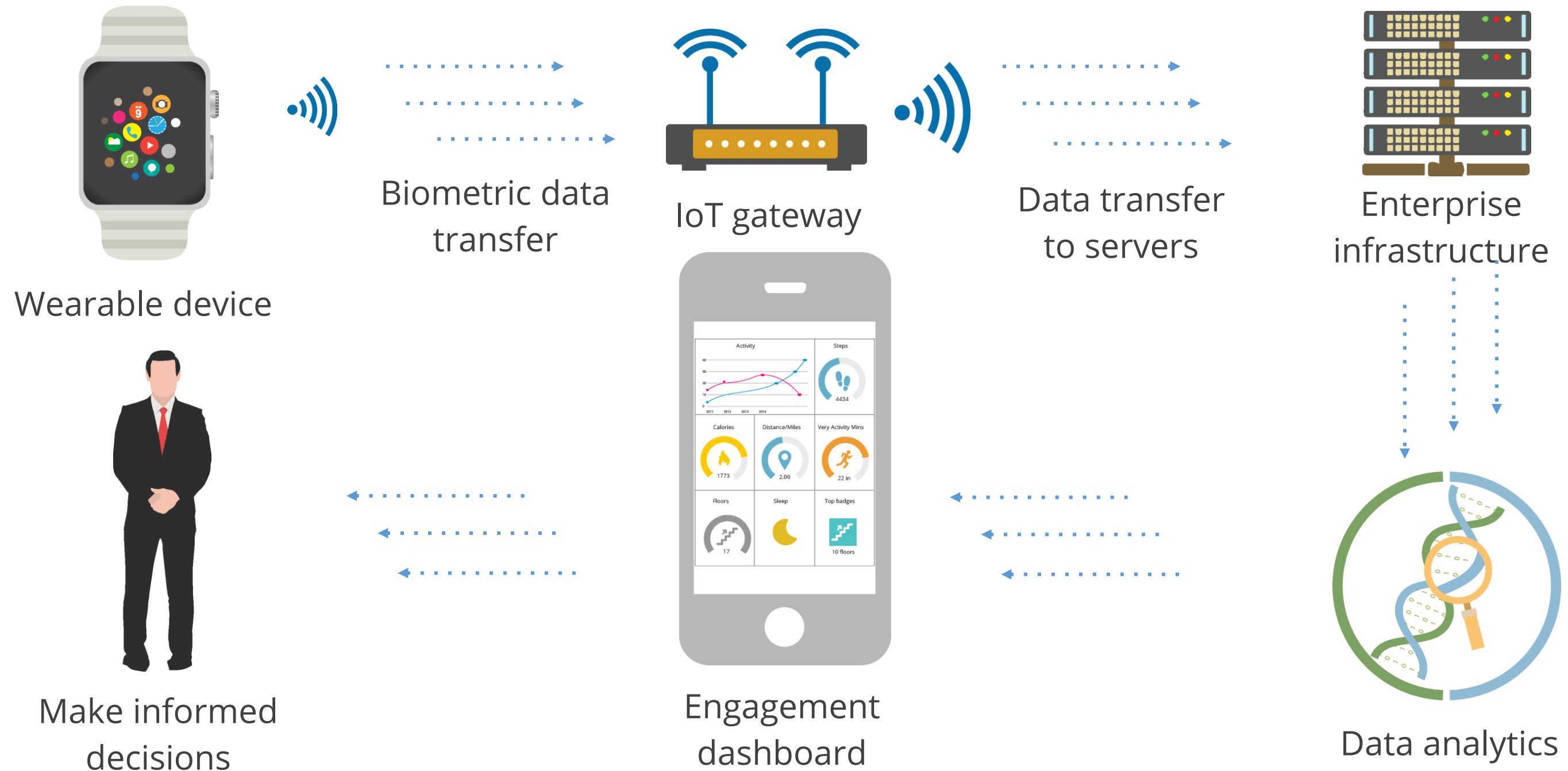
Fast and real-time analytics is made possible by modern and advanced infrastructure, tools, and technologies

Influencing Factors

- Query volume – Unique and verifiable users
- Geographical locations
- Keyword or phrase matches on the web
- Scrubbing for inappropriate content

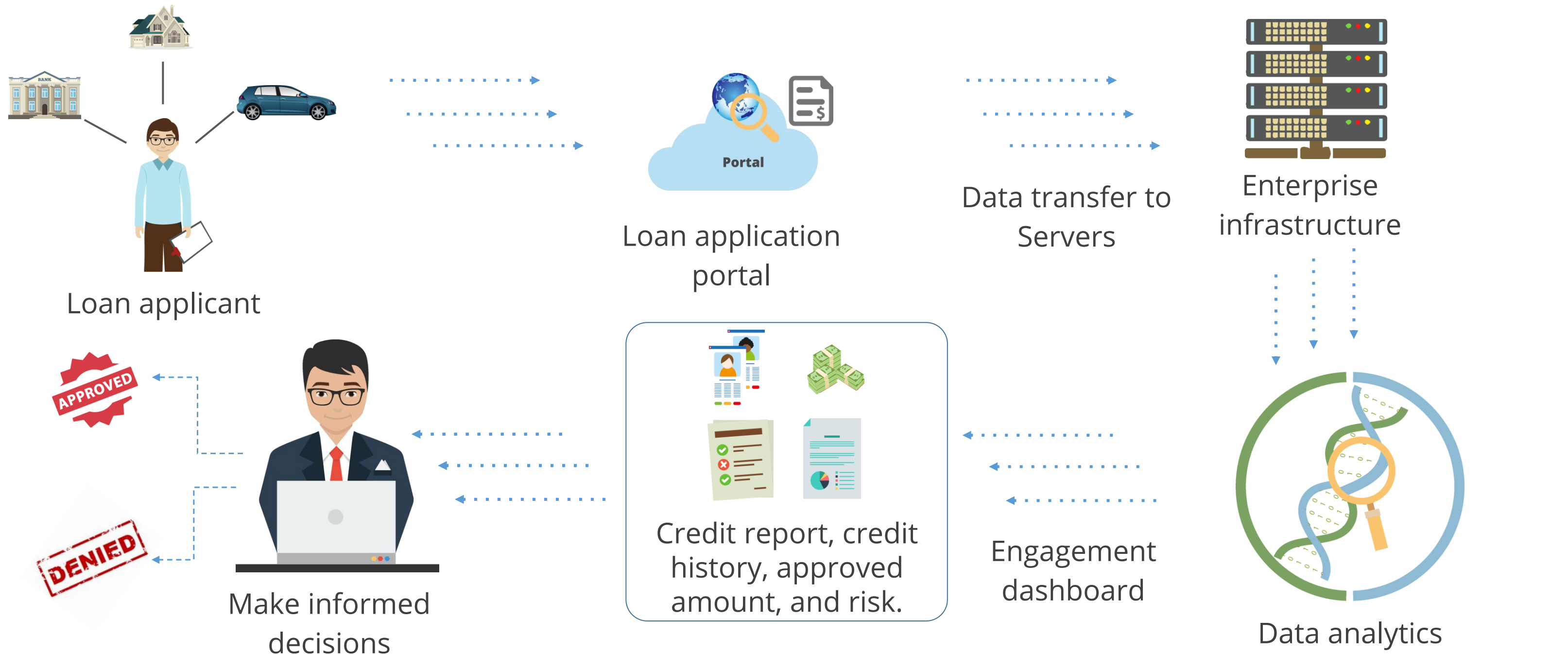
Using Data Science: Healthcare

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



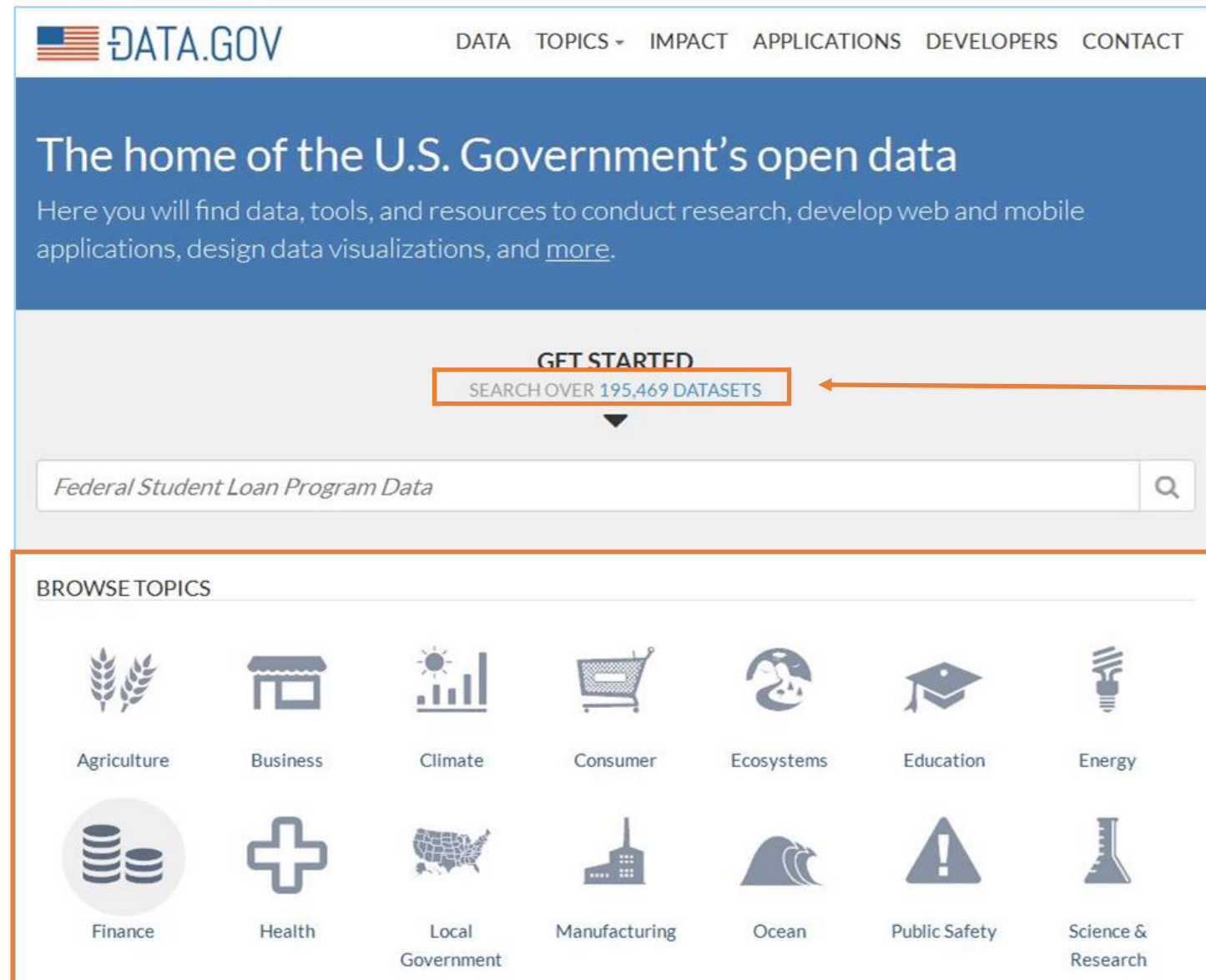
Using Data Science: Finance

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



Using Data Science: Public Sector

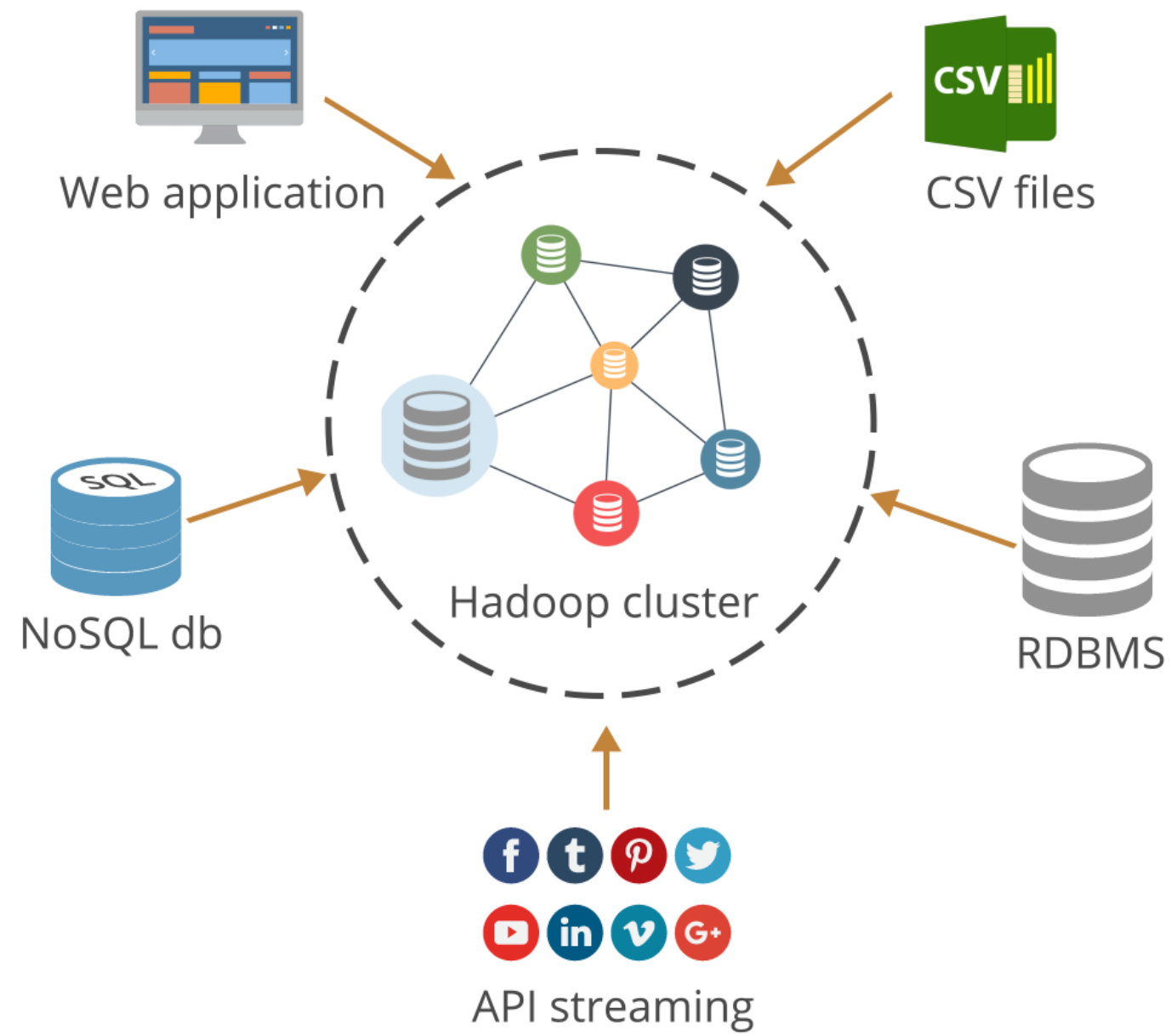
Governments in different countries share large datasets from various domains with the public. Data.gov is a website hosted and maintained by the U.S. government.



Large collection of datasets

Sectors or domains

The Real Challenge



Some of the challenges data scientists face in the real world are:

- Data quality doesn't conform to the set standards.
- Data integration is a complex task.
- Data is distributed into large clusters in HDFS, which is difficult to integrate and analyze.
- Unstructured and semi-structured data are harder to analyze.

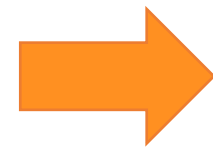
Python

Data Analytics and Python

Python deals with each stage of data analytics efficiently by applying different libraries and packages.



Data analytics



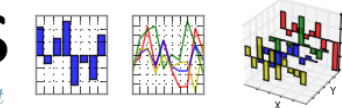
Acquire



Wrangle

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Explore

matplotlib

Model



Visualize

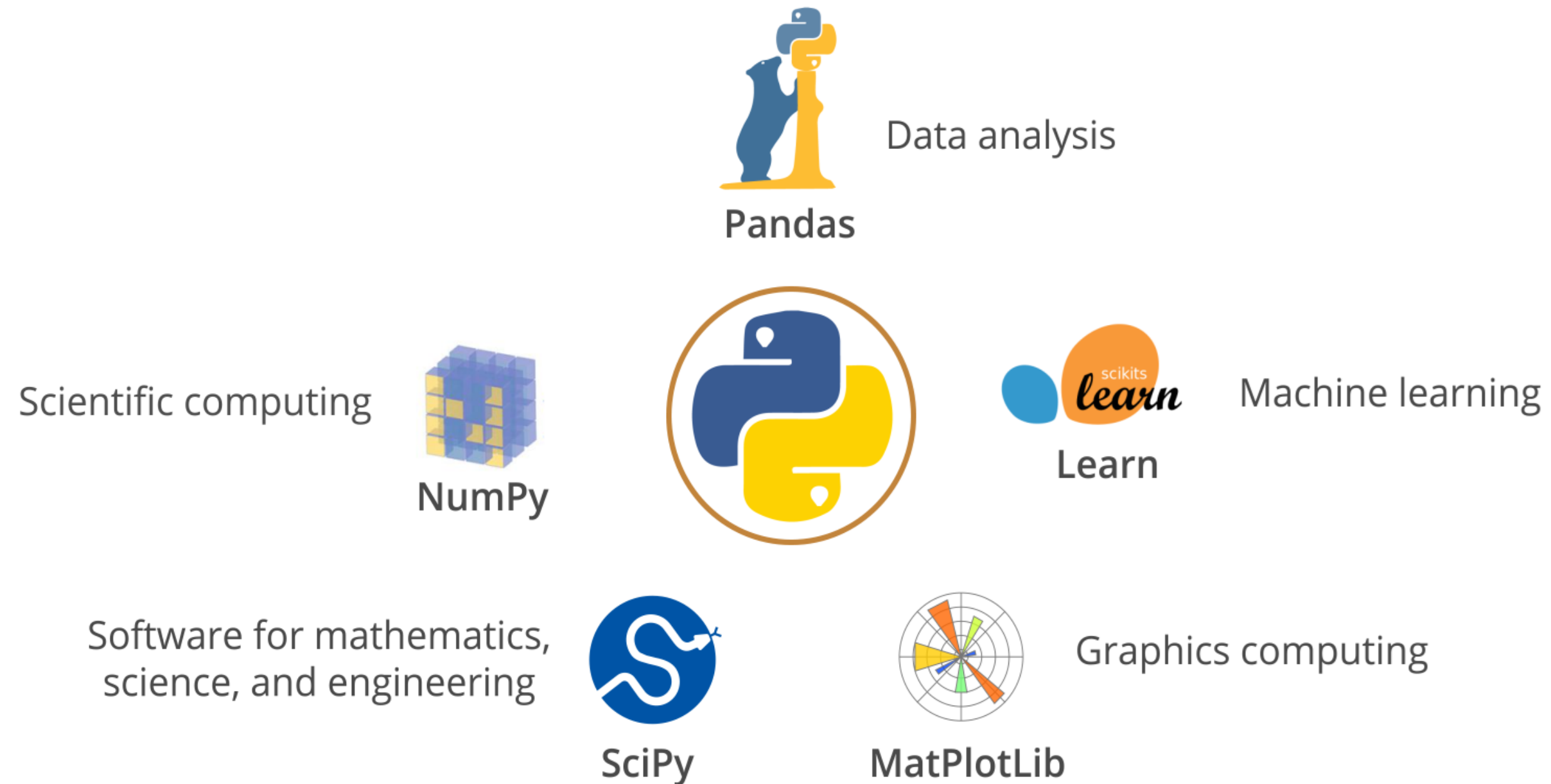


Bokeh

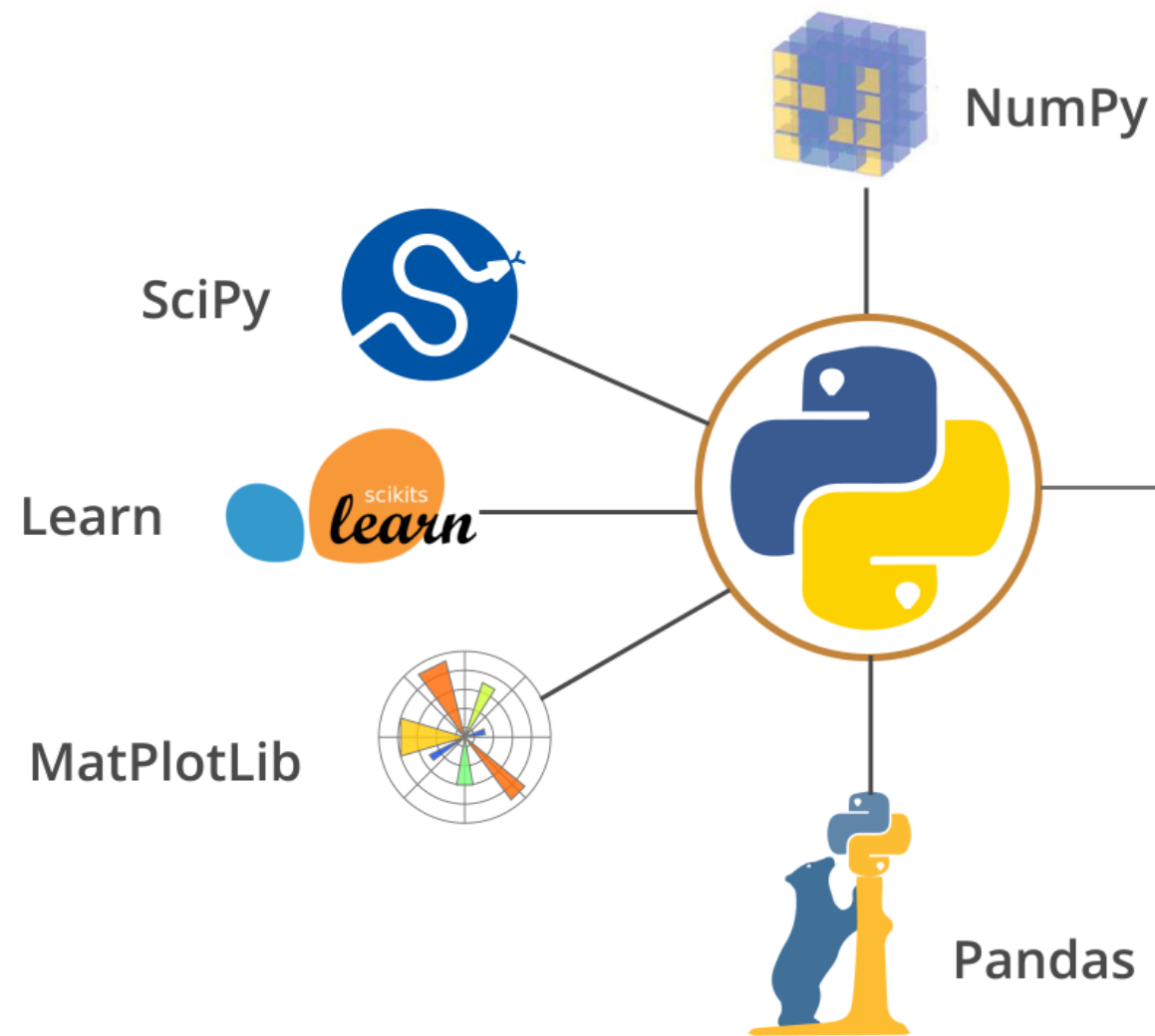


Python Tools and Technologies

Python is a general-purpose, open-source programming language that lets you work quickly and integrate systems more effectively.



Benefits of Python



Easy to learn

Open source

Efficient and multi-platform support

Huge collection of libraries, functions, and modules

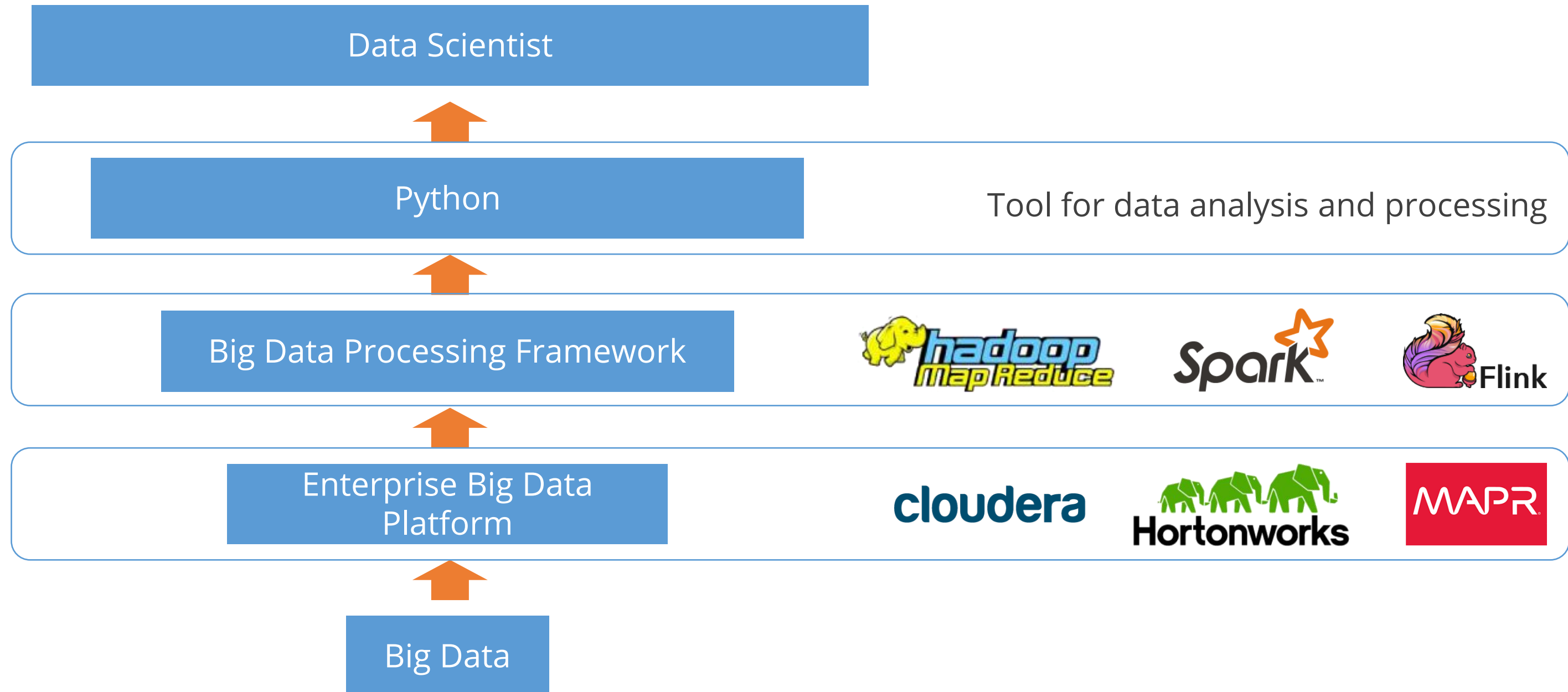
Big open-source community

Integrates well with enterprise apps and systems

Great vendor and product support

Big Data Platforms and Processing Frameworks for Python

Python is supported by well-established data platforms and processing frameworks that help analyze data in a simple and efficient way.





Knowledge Check

Knowledge Check

1

A data scientist ____.

- A. Asks the right questions
- B. Acquires data
- C. Performs data wrangling and data visualization
- D. All of the above



Knowledge Check

1

A data scientist ____.

- A. Asks the right questions
- B. Acquires data
- C. Performs data wrangling and data visualization
- D. All of the above



The correct answer is **D**

A data scientist asks the right questions to the stakeholders, acquires data from various sources and data points, performs data wrangling that makes the data available for analysis, and creates reports and plots for data visualization.

Knowledge Check

2

The search engine's autocomplete feature identifies unique and verifiable users who search for a particular keyword or phrase to____. Select all that apply.

- A. Scrub inappropriate content
- B. Build a query volume
- C. Tag the location to a query
- D. Find similar instances on the web



Knowledge Check

2

The search engine's autocomplete feature identifies unique and verifiable users who search for a particular keyword or phrase to _____. Select all that apply.

- A. Scrub inappropriate content
- B. Build a query volume
- C. Tag the location to a query
- D. Find similar instances on the web



The correct answer is **B, C**

The search engine's autocomplete feature identifies unique and verifiable users who search for a particular keyword or phrase to build a query volume. It also helps identify the users' locations and tags them with the query, enabling it to be location-specific.

Knowledge Check

3

What is the sequential flow of analysis in data science?

- A. Data wrangling, exploration, modeling, acquisition, and visualization
- B. Data exploration, acquisition, modeling, wrangling, and visualization
- C. Data acquisition, wrangling, exploration, modeling, and visualization
- D. Data modeling, acquisition, exploration, wrangling, and visualization



Knowledge Check

3

What is the sequential flow of analysis in data science?

- A. Data wrangling, exploration, modeling, acquisition, and visualization
- B. Data exploration, acquisition, modeling, wrangling, and visualization
- C. Data acquisition, wrangling, exploration, modeling, and visualization
- D. Data modeling, acquisition, exploration, wrangling, and visualization



The correct answer is **C**

In data science, the data is acquired from various sources and is then wrangled to ease its analysis. This is followed by data exploration and data modeling. The final stage is data visualization, where the data is presented and the patterns are identified.

Key Takeaways

- 👁 Data science is an approach that combines the aspects of statistics, mathematics, programming, and visualization for analyzing data to derive information and insights from it that are used in a variety of application domains.
- 👁 Data scientists use data analysis and analytics to perform tasks such as extracting information from data, building and training machine learning models, and developing data tools and applications.
- 👁 Data science is used in a variety of fields, including analysis, mathematical and statistical modeling, and financial planning.
- 👁 Python is a versatile programming language with a wide number of libraries and packages that can help with each stage of data analytics and can be integrated with a variety of systems.

