DATA AND
ARTIFICIAL INTELLIGENCE

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:

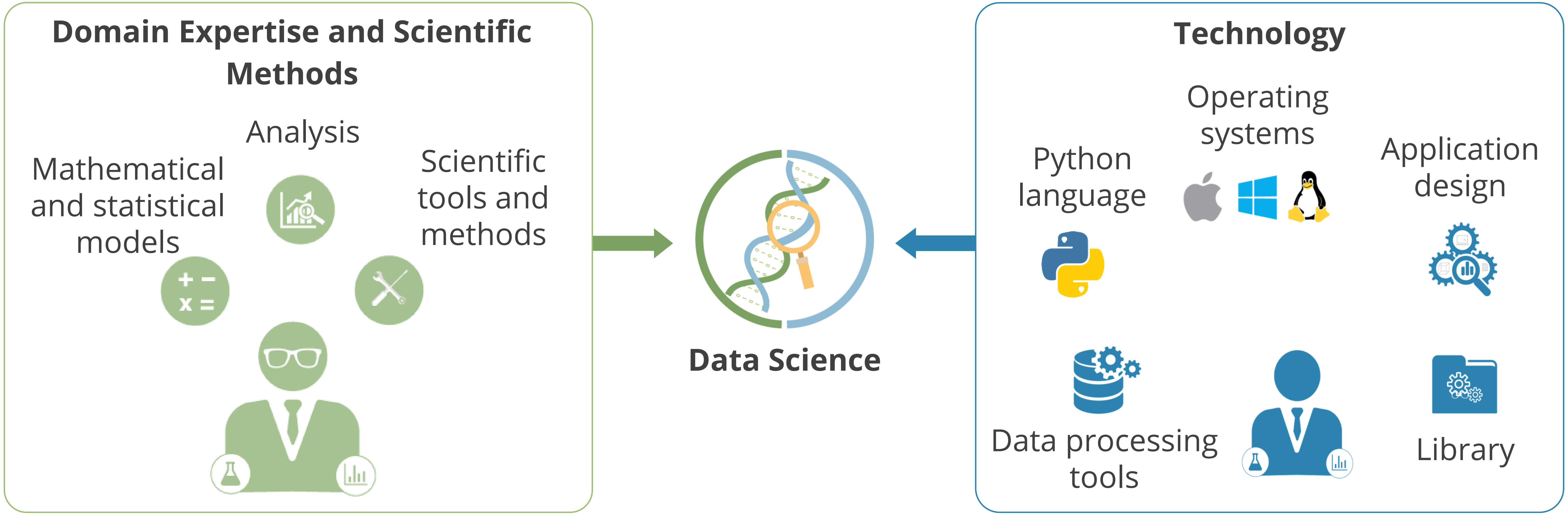
A powerful new approach to make discoveries from data

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

Analysis

Mathematical and statistical models

Scientific tools and methods

Data Science

Technology

Python language

Operating systems

Application design

Data processing tools

Library

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.

Analysis

Mathematical and statistical models

Scientific tools and methods

Domain Expertise and Scientific Methods

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

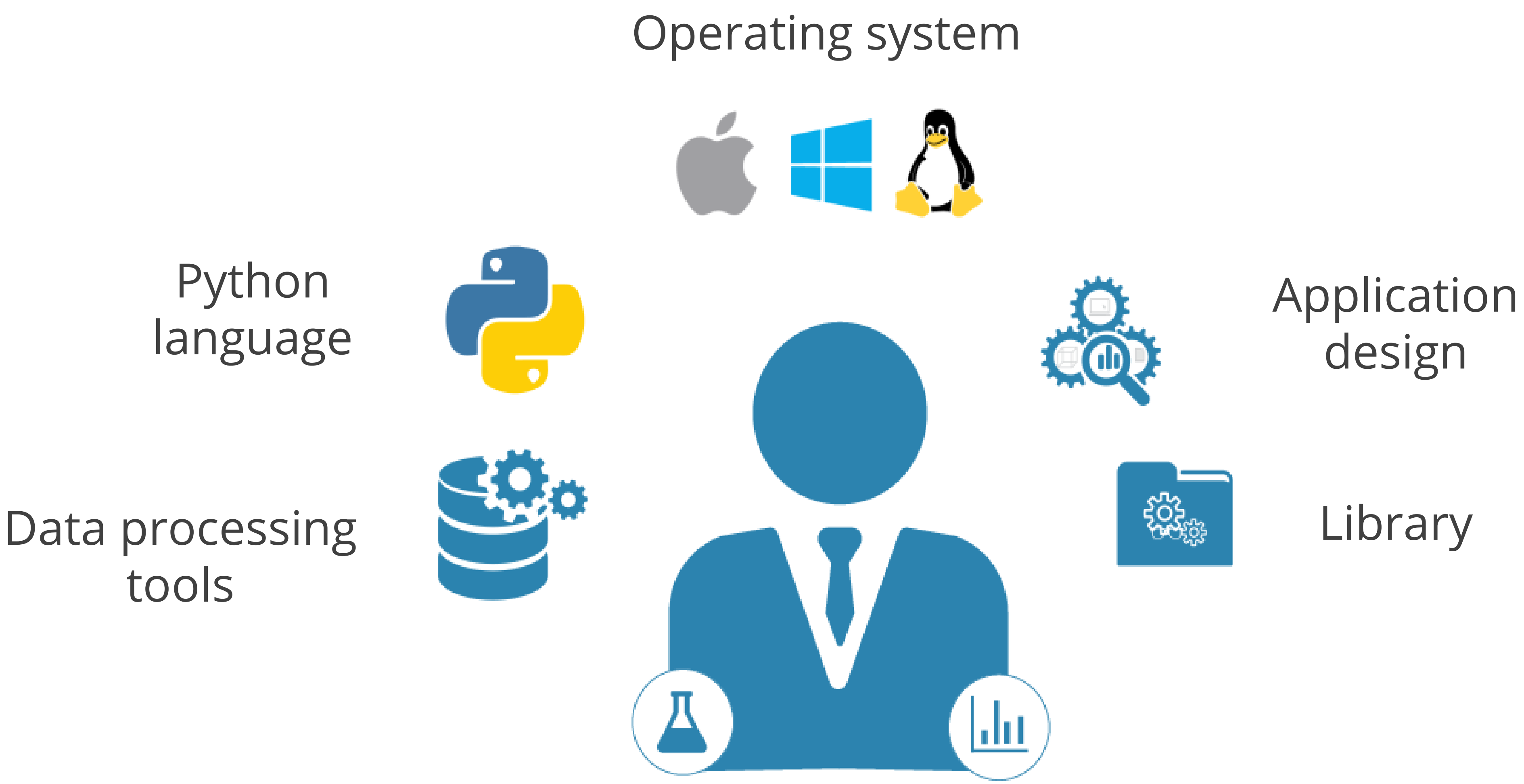
Data analysis can be:

- 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.

Technology



Operating system

Python language

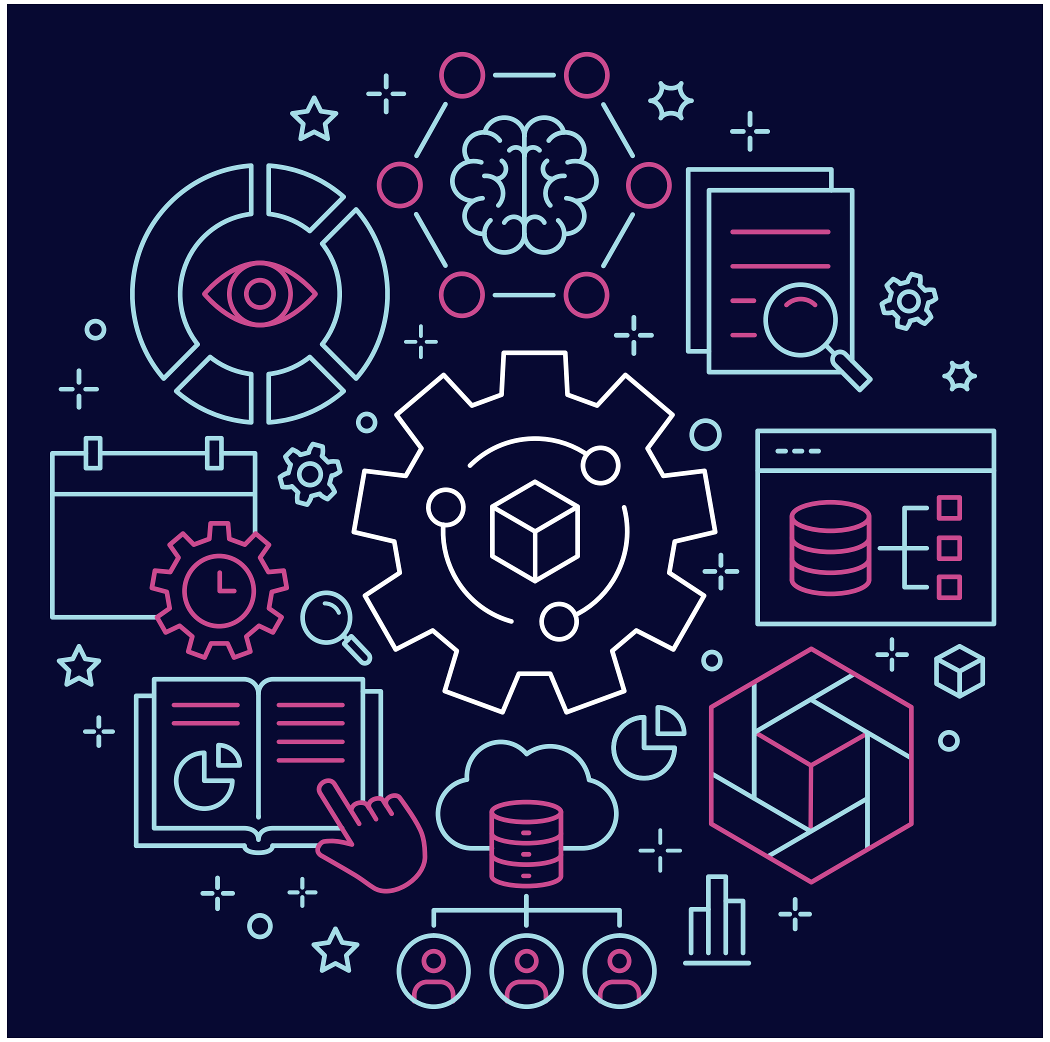
Application design

Data processing tools

Library

Data Processing and Analytics

These technologies help data scientists to:

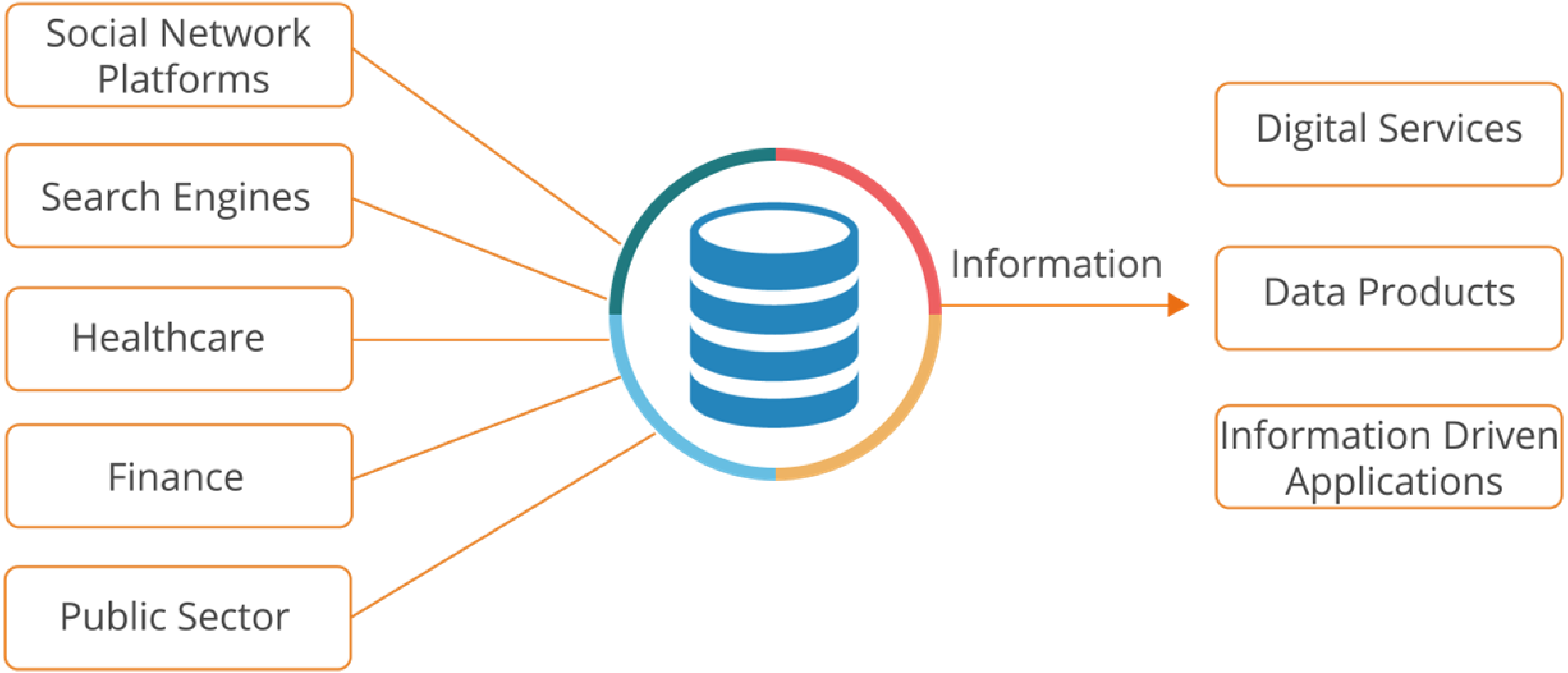


- Build and train machine learning models
- Manipulate data with technology
- Build data tools, applications, and services
- Extract information from data

Applications of Data Science

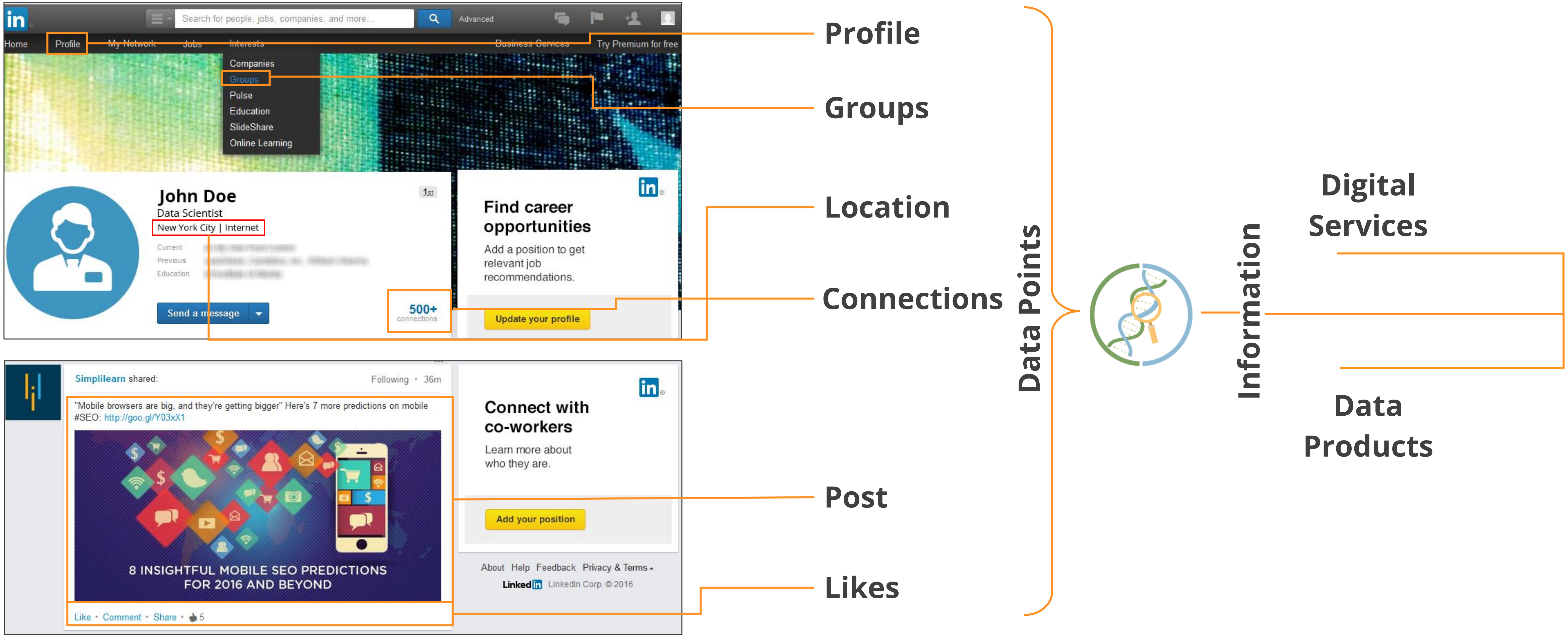
Different Sectors Using Data Science

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.



Search for people, jobs, companies, and more... | Advanced

Home | Profile | My Network | Jobs | Interests | Business Services | Try Premium for free

Companies
Groups
Pulse
Education
SlideShare
Online Learning

John Doe 1st
Data Scientist
New York City | Internet
Current [redacted]
Previous [redacted]
Education [redacted]

Send a message

500+ connections

Find career opportunities
Add a position to get relevant job recommendations.
Update your profile

Simplilearn shared: Following • 36m

"Mobile browsers are big, and they're getting bigger" Here's 7 more predictions on mobile #SEO: http://goo.gl/Y03xX1

8 INSIGHTFUL MOBILE SEO PREDICTIONS FOR 2016 AND BEYOND

Like • Comment • Share • 5

Connect with co-workers
Learn more about who they are.
Add your position

About Help Feedback Privacy & Terms
Linked in LinkedIn Corp. © 2016

Profile

Groups

Location

Connections

Post

Likes

Data Points

Information

Digital Services

Data Products

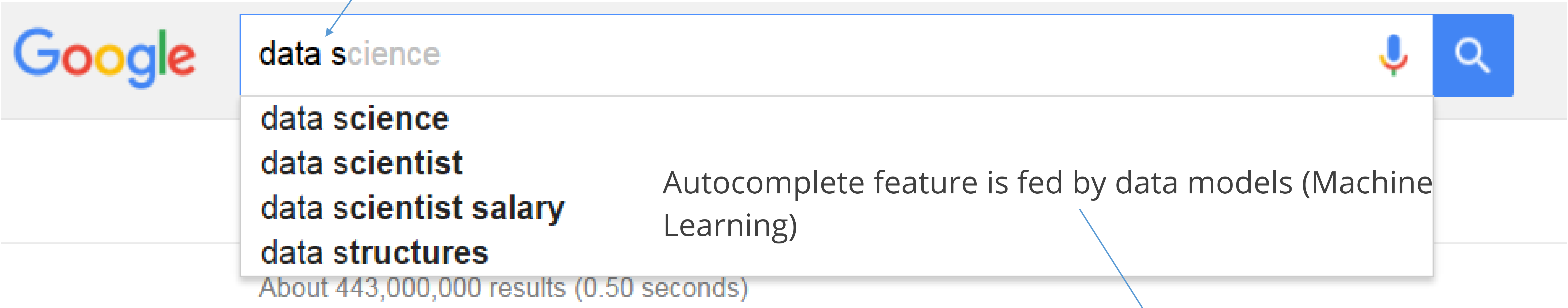
Powered by simplilearn

PURDUE UNIVERSITY

Using Data Science: Search Engines

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

Search keyword



data science

data science

data scientist

data scientist salary

data structures

About 443,000,000 results (0.50 seconds)

Autocomplete feature is fed by data models (Machine Learning)

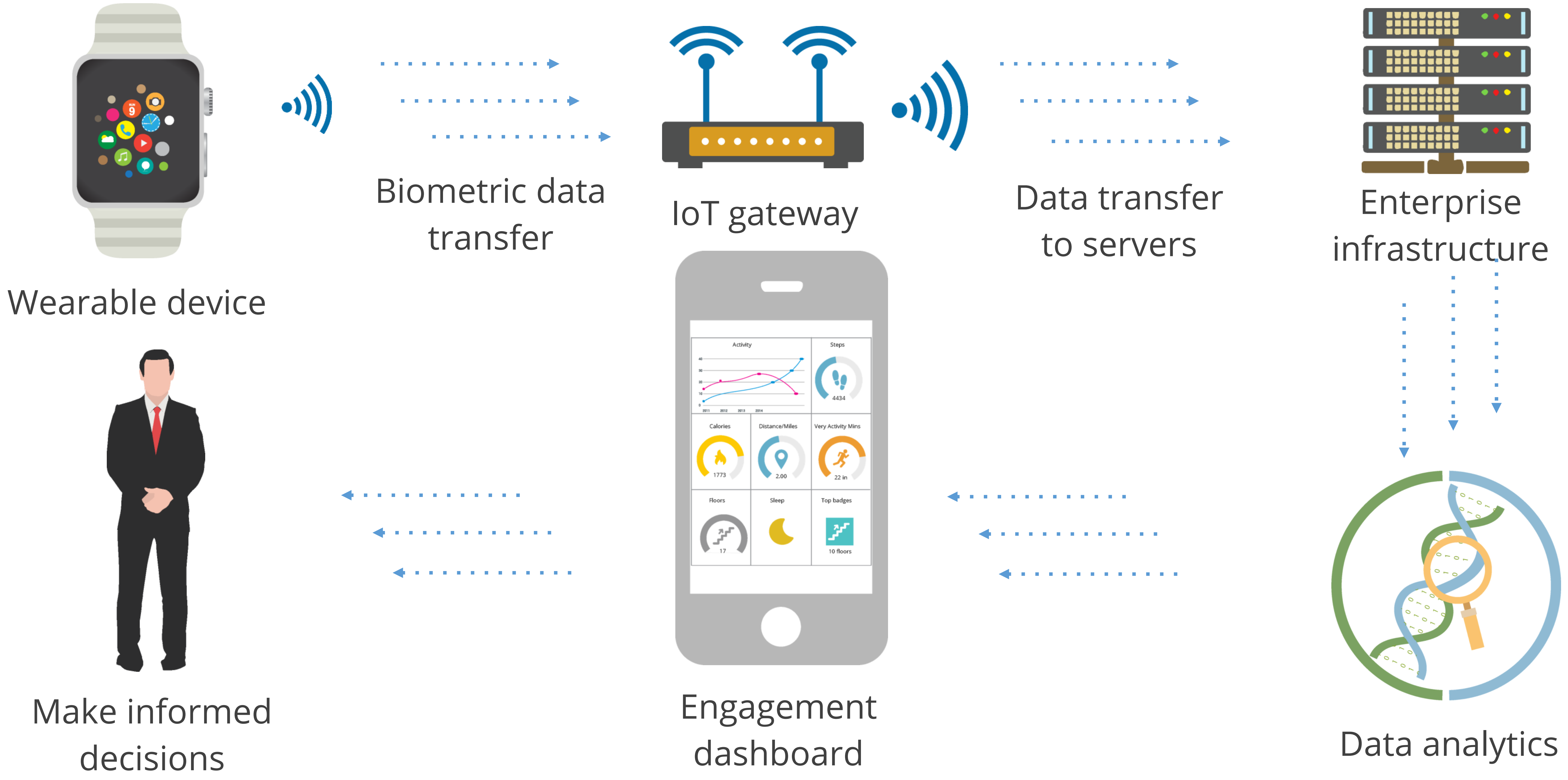
Fast and real-time analytics is made possible by modern and advanced infrastructure, tools, and technologies

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.



Wearable device

Biometric data transfer

IoT gateway

Data transfer to servers

Enterprise infrastructure

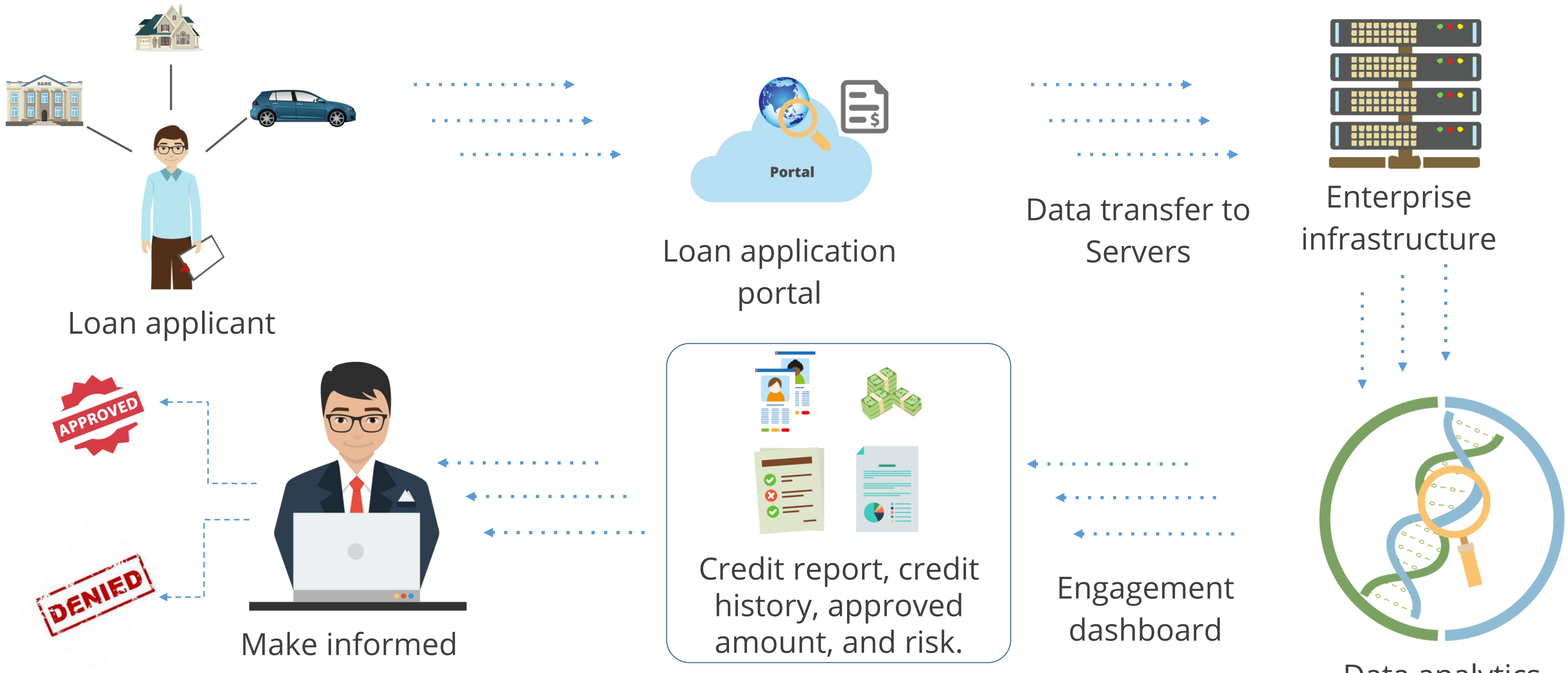
Data analytics

Engagement dashboard

Make informed decisions

Using 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

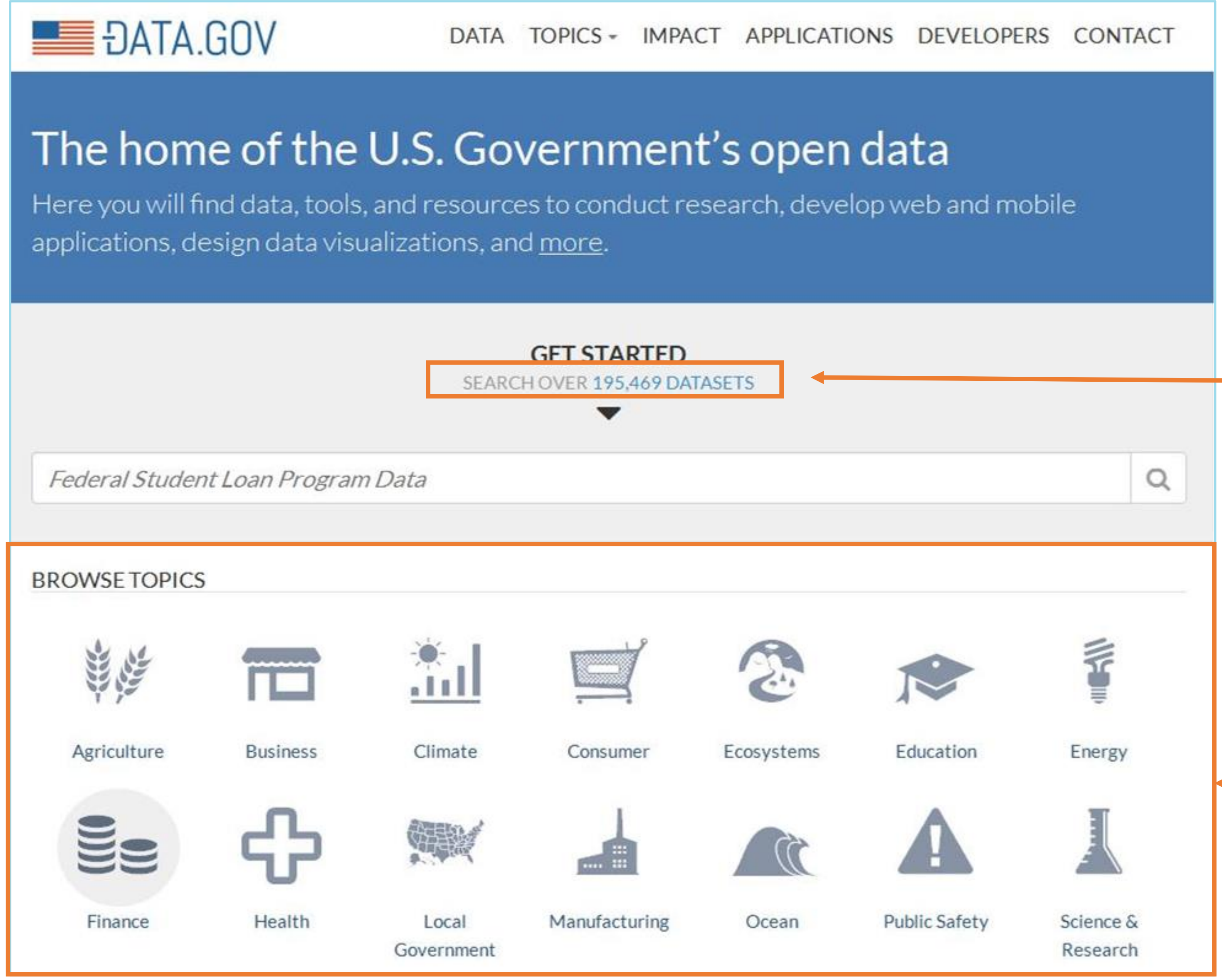
Engagement dashboard

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

Make informed decisions

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.



DATA.GOV

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER 195,469 DATASETS

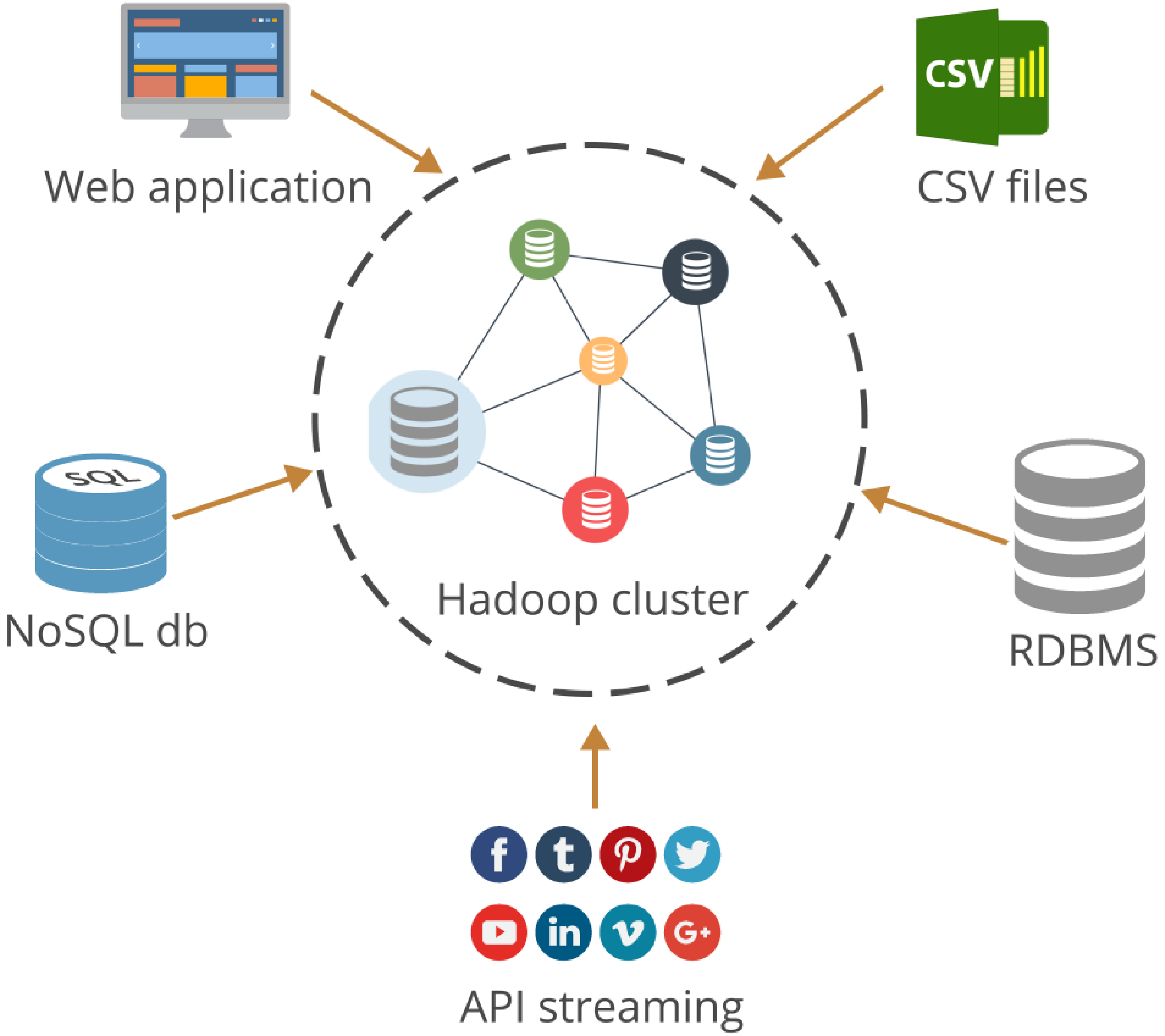
Federal Student Loan Program Data

BROWSE TOPICS

Large collection of datasets

Sectors or domains

The Real Challenge



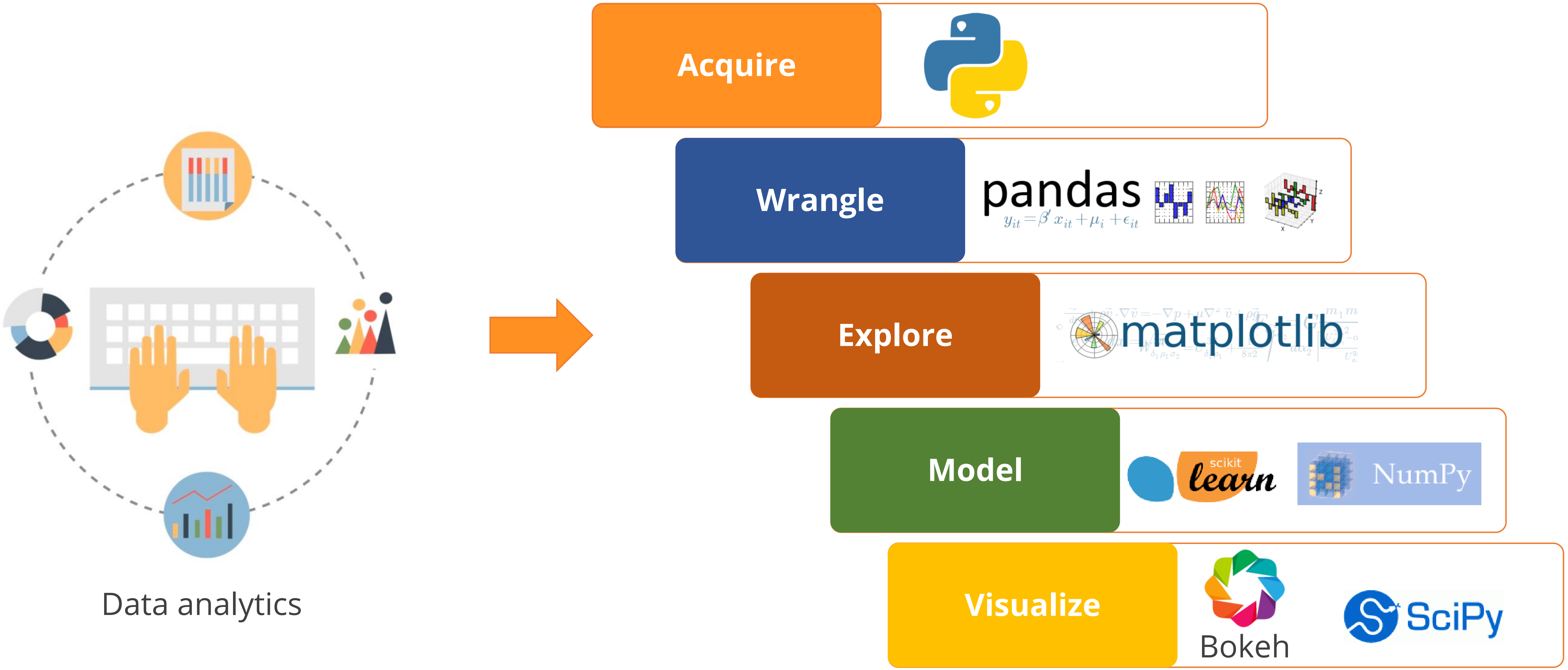
Some of the challenges data scientists face in the real world are:

- 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.

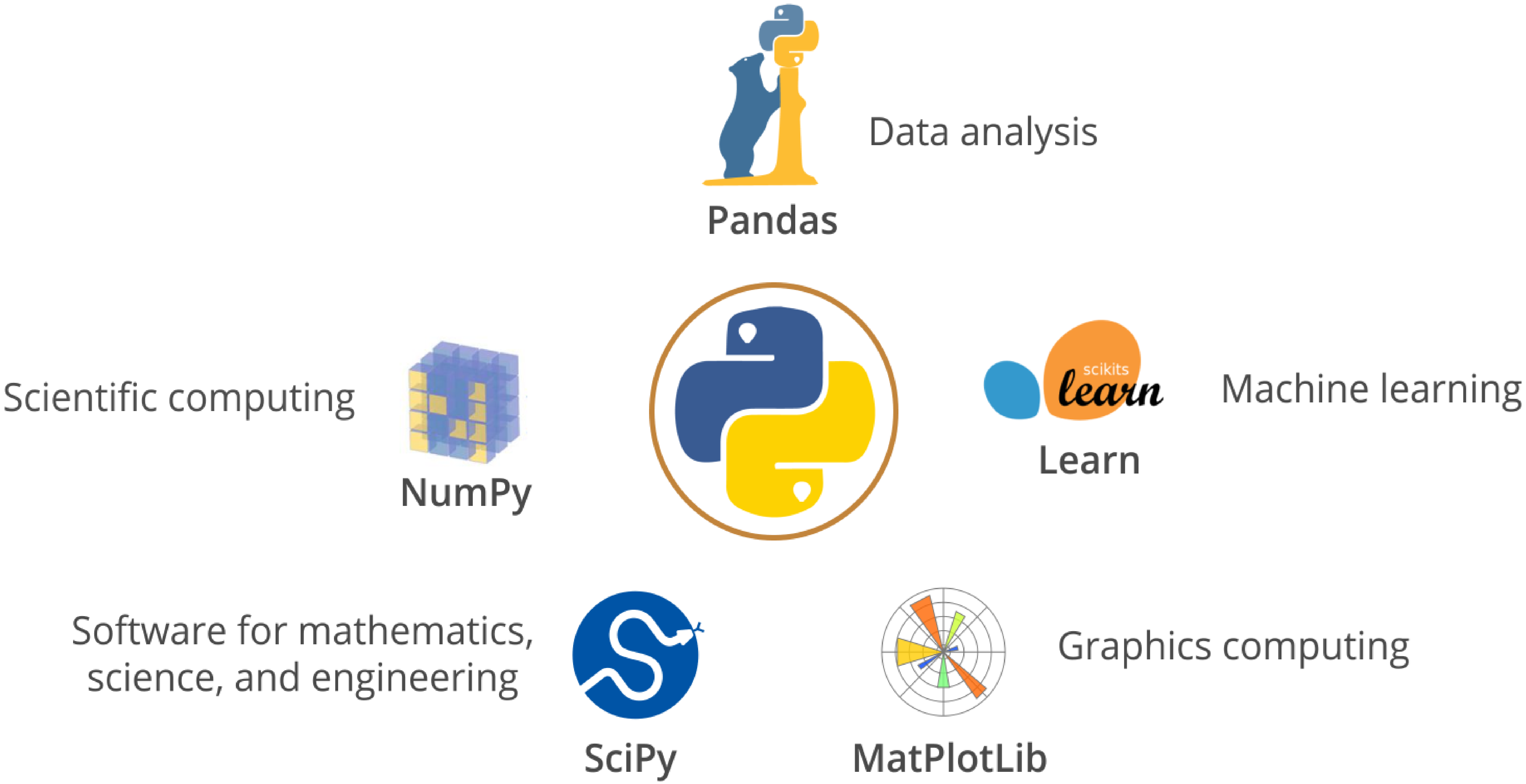


Data analytics

- Acquire – Python
- Wrangle – pandas
- Explore – matplotlib
- Model – scikit learn, NumPy
- Visualize – Bokeh, SciPy

Python Tools and Technologies

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



Pandas

Data analysis

NumPy

Scientific computing

Learn

Machine learning

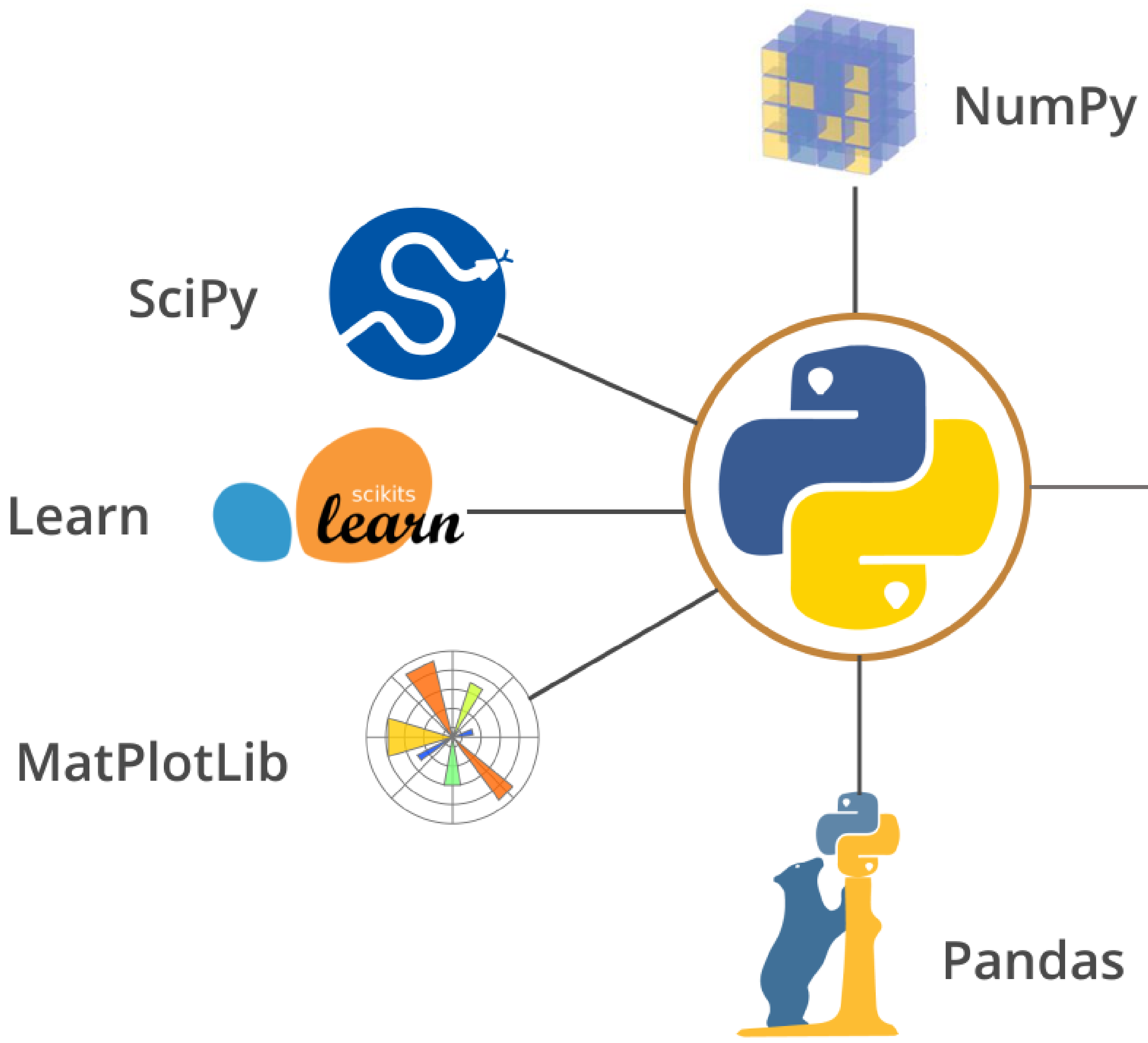
SciPy

Software for mathematics, science, and engineering

MatPlotLib

Graphics computing

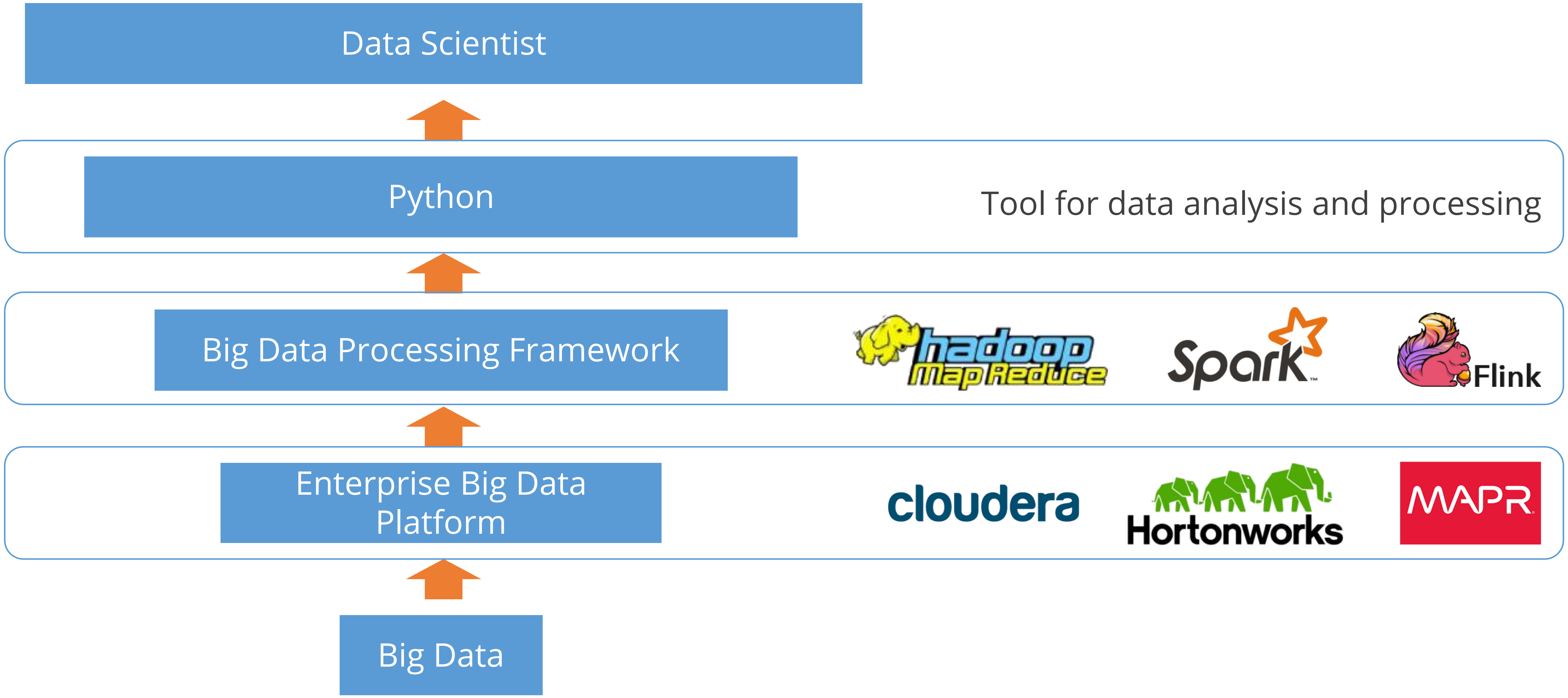
Benefits of Python



- 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.



Data Scientist

Python

Tool for data analysis and processing

Big Data Processing Framework

Enterprise Big Data Platform

Big Data

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 exploration, acquisition, modeling, wrangling, 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 exploration, acquisition, modeling, wrangling, 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.