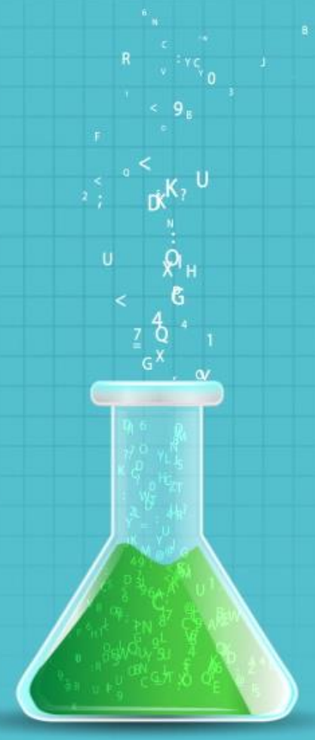
Data Science with Python

Lesson 1—Data Science Overview

DATA SCIENCE

What You'll Learn

- Know what Data Science is
- Discuss the roles and responsibilities of a Data Scientist
- List various applications of Data Science
- Understand how Data Science and Big Data work together
- Explore Data Science as a discipline
- Understand how and why Data Science is gaining importance
- Understand what Python is and what problems it resolves



What is Data Science

Some common definitions of Data Science are as follows:

A powerful new approach to make discoveries from data

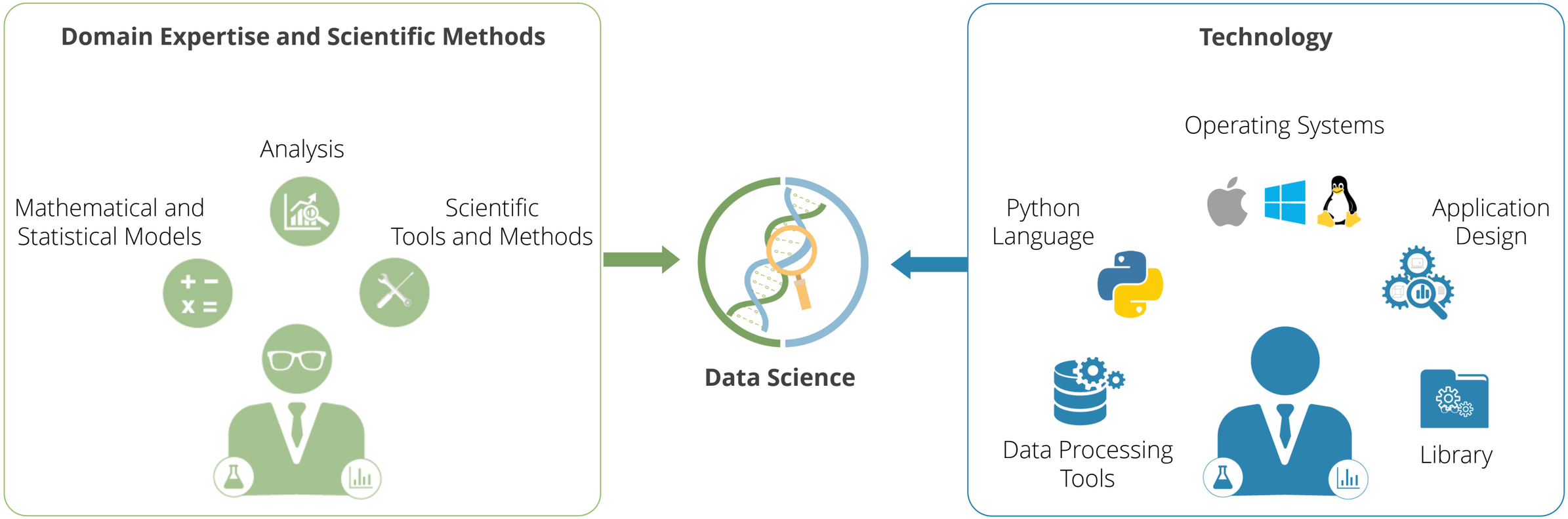
An automated way to analyze enormous amounts of data and extract information

Data Science

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

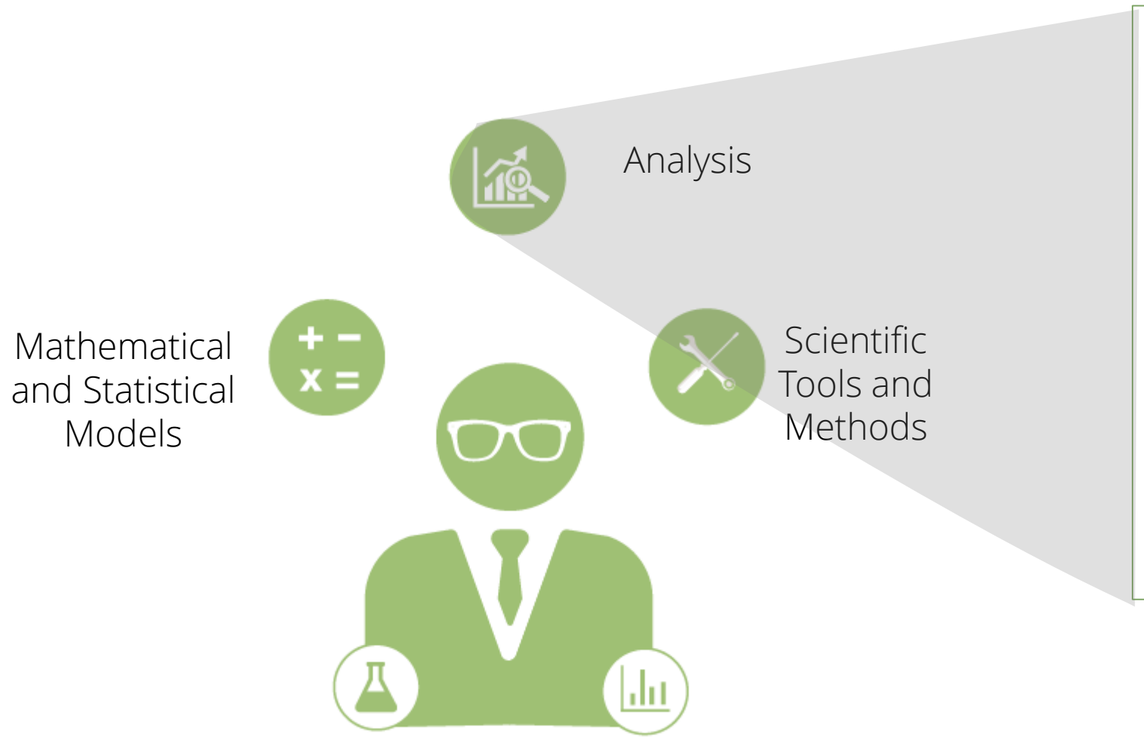
The Components of Data Science

When we combine domain expertise and scientific methods with technology, we get Data Science.



Domain Expertise and Scientific Methods

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



Data analysis can be:

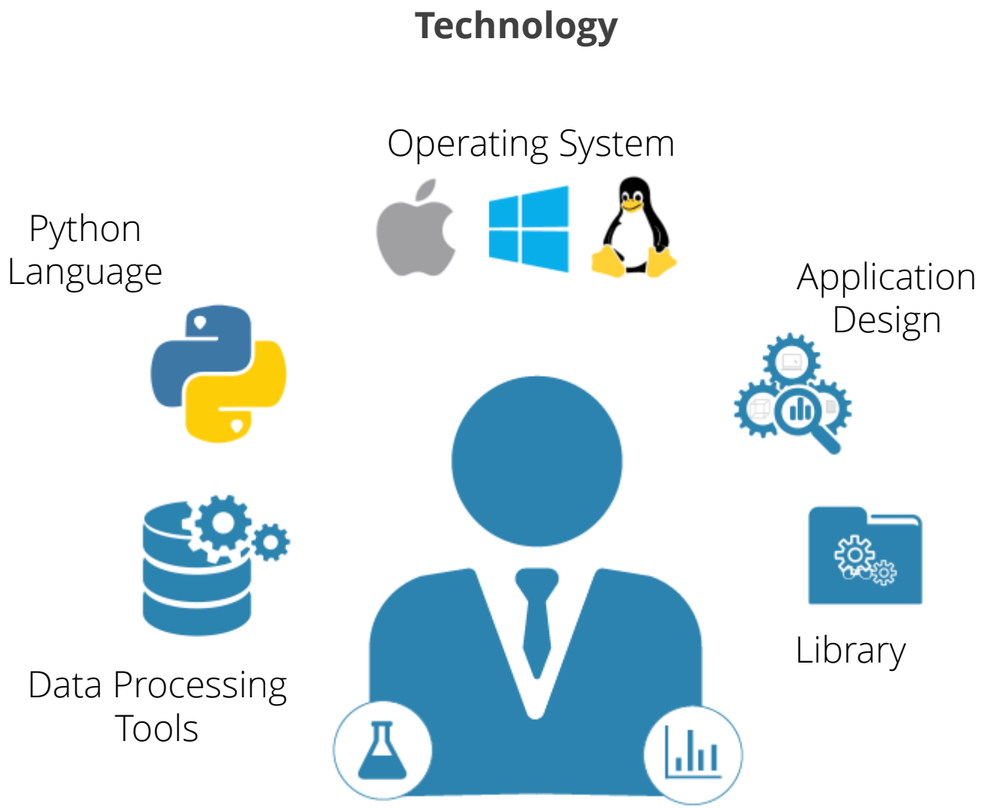
- Descriptive: Study a dataset to decipher the details
- Predictive: Create a model based on existing information to predict outcome and behavior
- Prescriptive: Suggest actions for a given situation using the collected information

Data Processing and Analytics

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

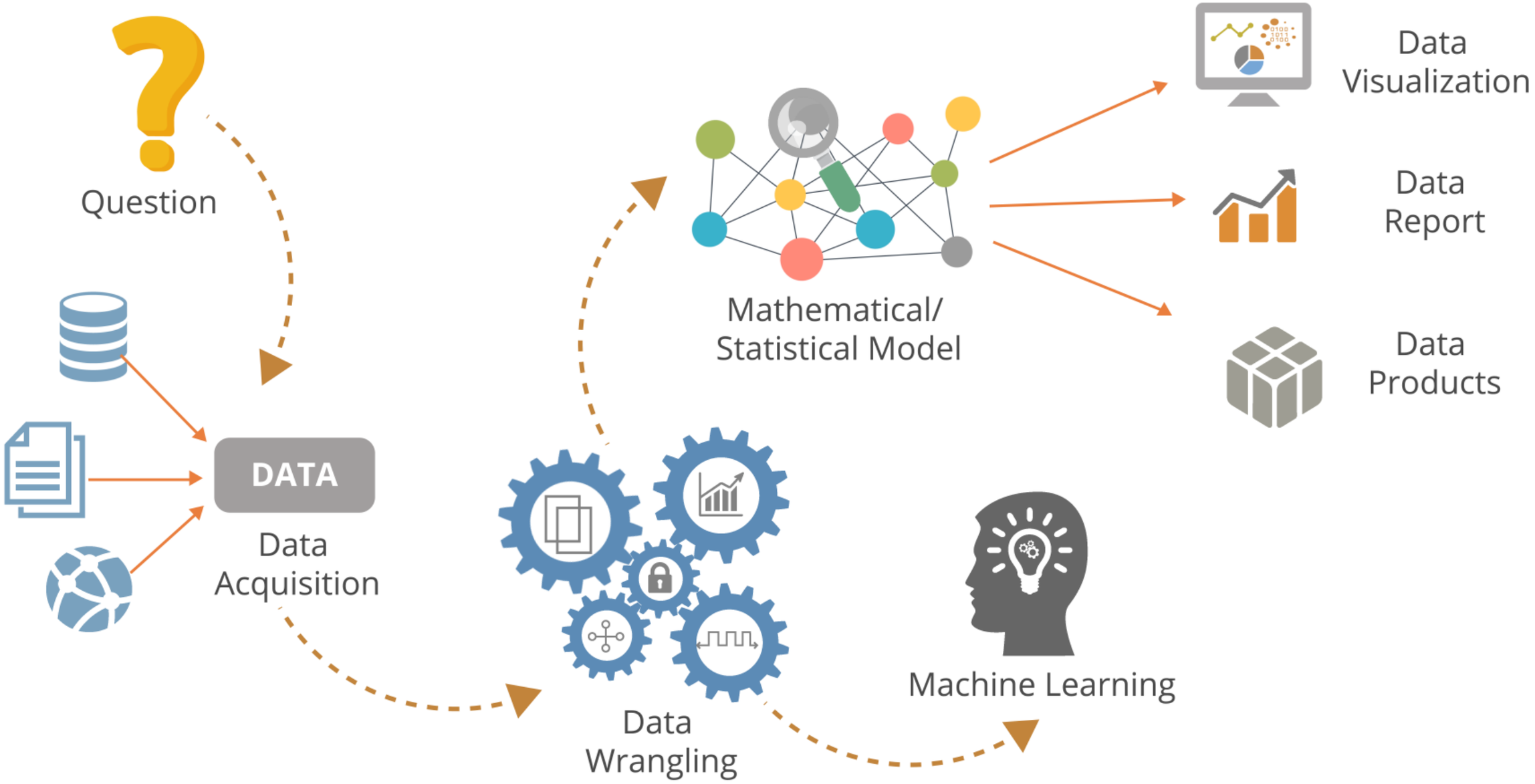
These technologies help Data Scientists to

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



Data analysis that uses only technology and domain knowledge without mathematical and statistical knowledge often leads to incorrect patterns and wrong interpretations. This can cause serious damage to businesses.

A Day in a Data Scientist's Life



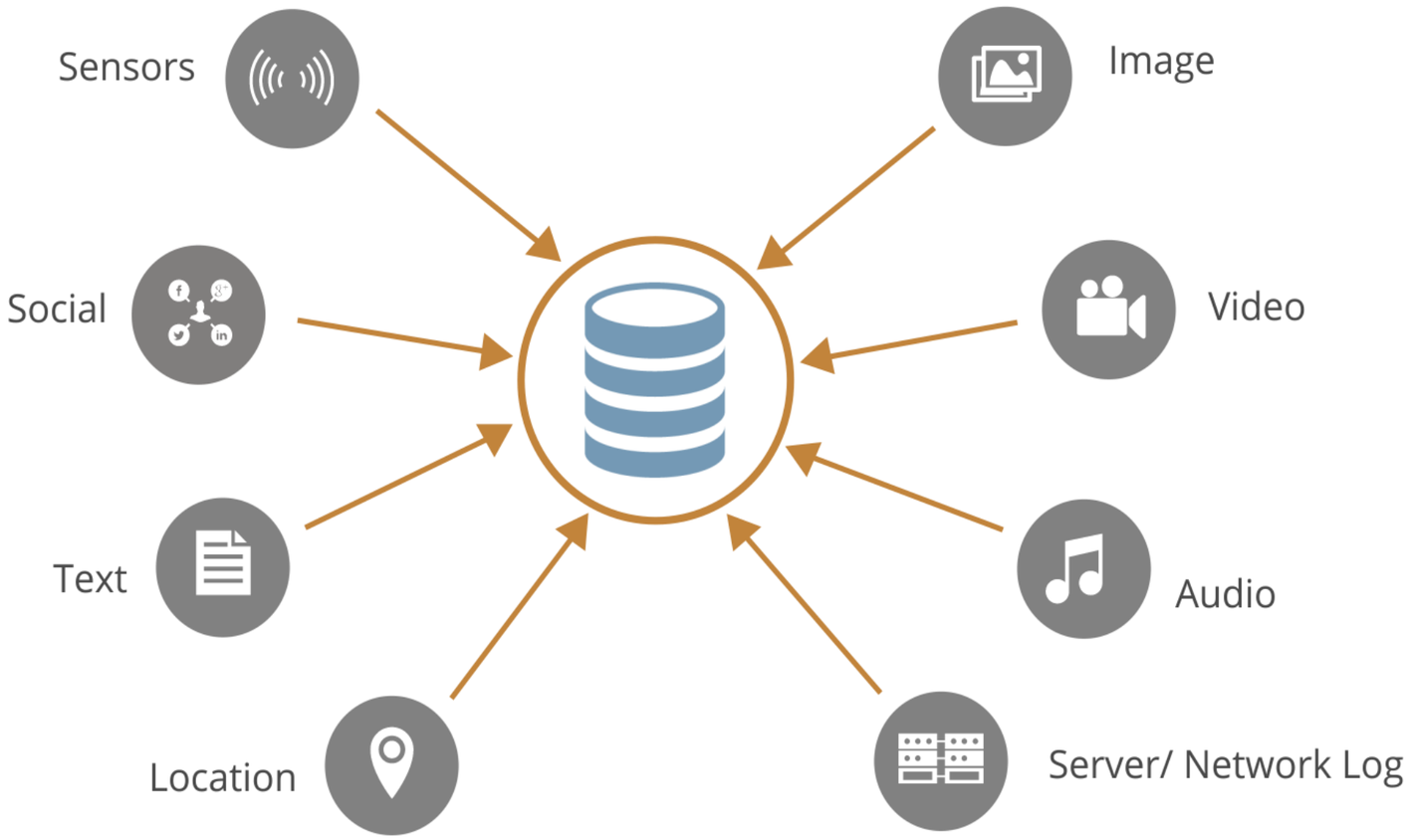
Basic Skills of a Data Scientist

A Data Scientist should be able to

- Ask the right questions
- Understand data structure
- Interpret and wrangle data
- Apply statistical and mathematical methods
- Visualize data and communicate with stakeholders
- Work as a team player

Sources of Big Data

Data Scientists work with different types of datasets for various purposes. Now that Big Data is generated every second through different media, the role of Data Science has become more important.



The 3 Vs of Big Data

Big Data is characterized by the following:

Volume	Enormous amount of data generated from various sources
Velocity	Large amount of data streaming in at great speeds, which requires quick data processing
Variety	Different formats of data: Structured, Semi-structured, and Unstructured

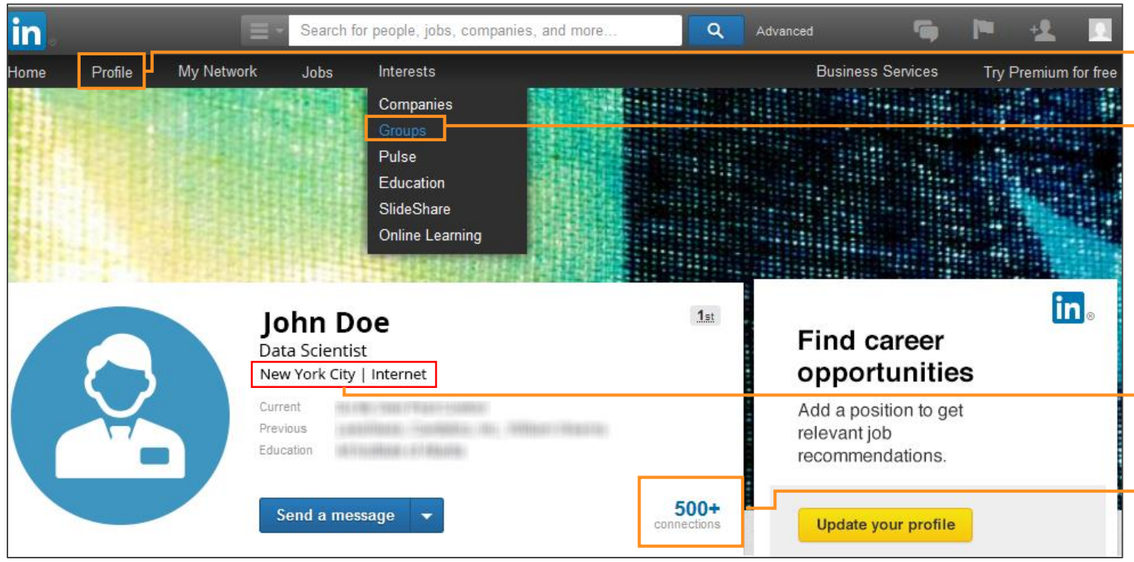
Big Data is a huge collection of data stored on distributed systems/machines popularly referred to as Hadoop clusters. Data Science helps extract information from the Data and build information-driven enterprises.

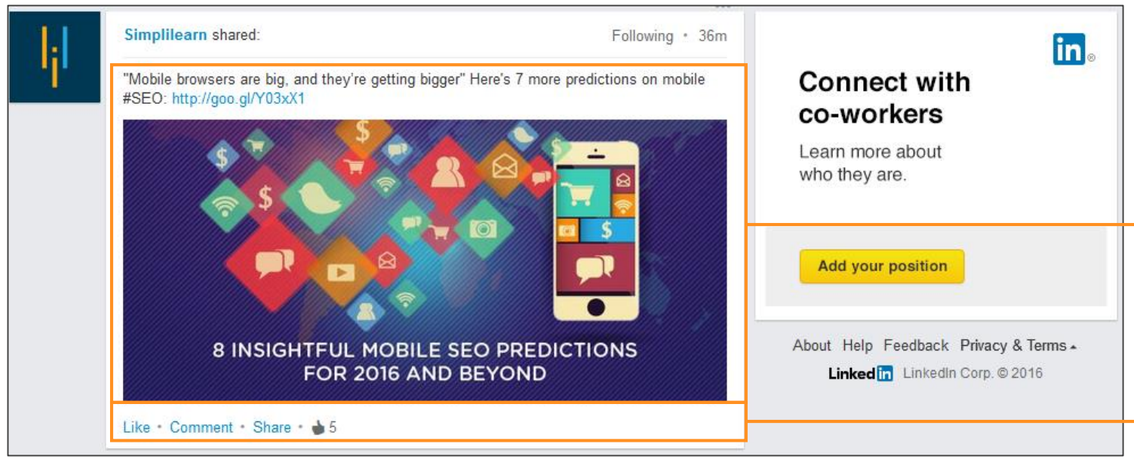
Different Sectors Using Data Science

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 them with relevant digital services and data products.





Profile

Groups

Location

Connections

Post

Likes

Data Points

Information

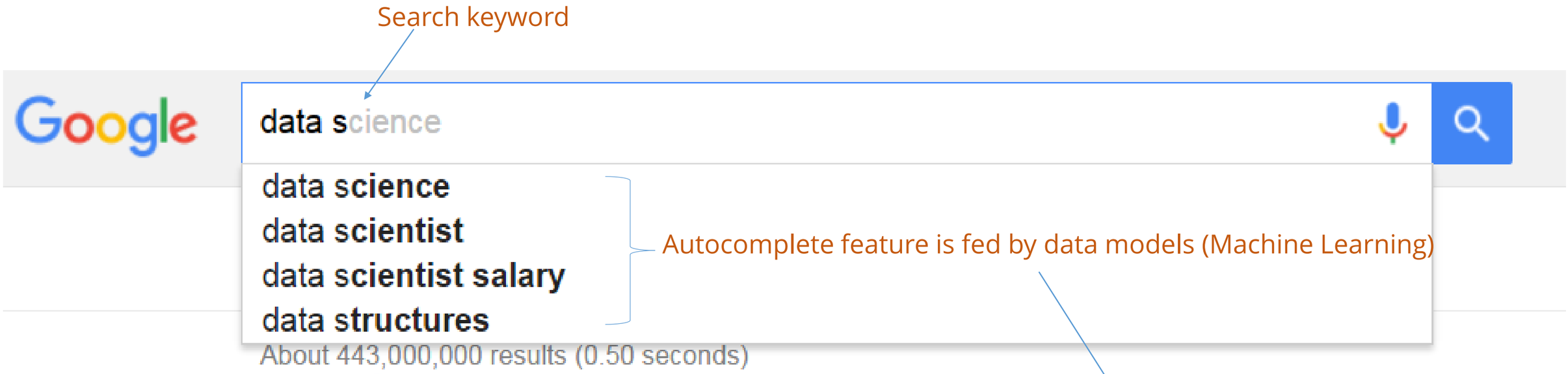
Digital Services

Data Products

simplilearn

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

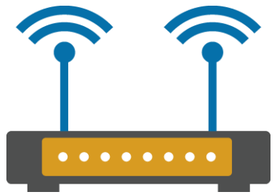
1. Query Volume – Unique and verifiable users
2. Geographical locations
3. Keyword/phrase matches on the web
4. Some scrubbing for inappropriate content

Using Data Science—Healthcare

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

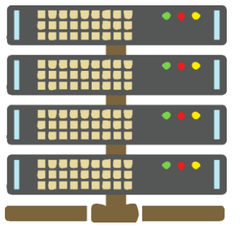
Wearable device

Biometric data transfer



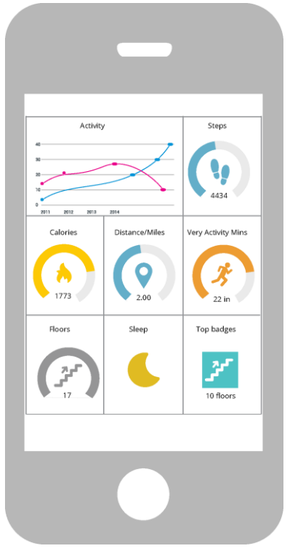
IOT Gateway

Data transfer to Servers



Enterprise Infrastructure

Make informed decisions

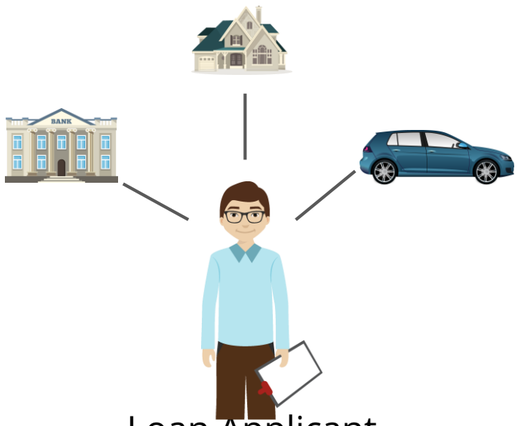


Engagement Dashboard

Data Analytics

Using Data Science—Finance

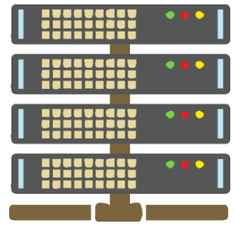
A Loan Manager can easily access and sift through a loan applicant's financial details using Data Science.



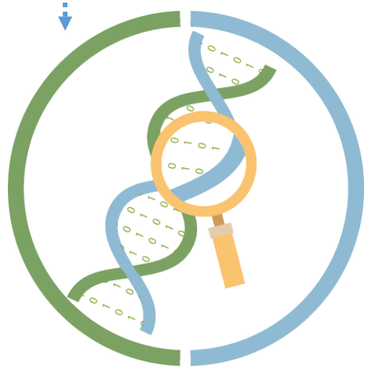
Loan Applicant

Loan Application Portal

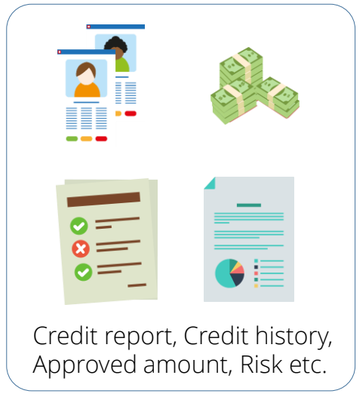
Data transfer to Servers



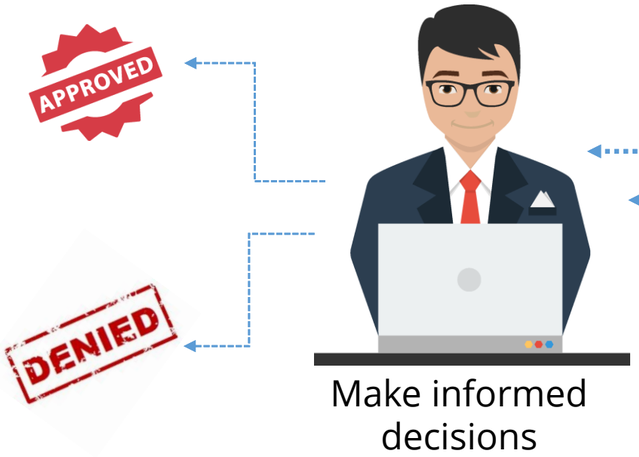
Enterprise Infrastructure



Data Analytics



Engagement Dashboard

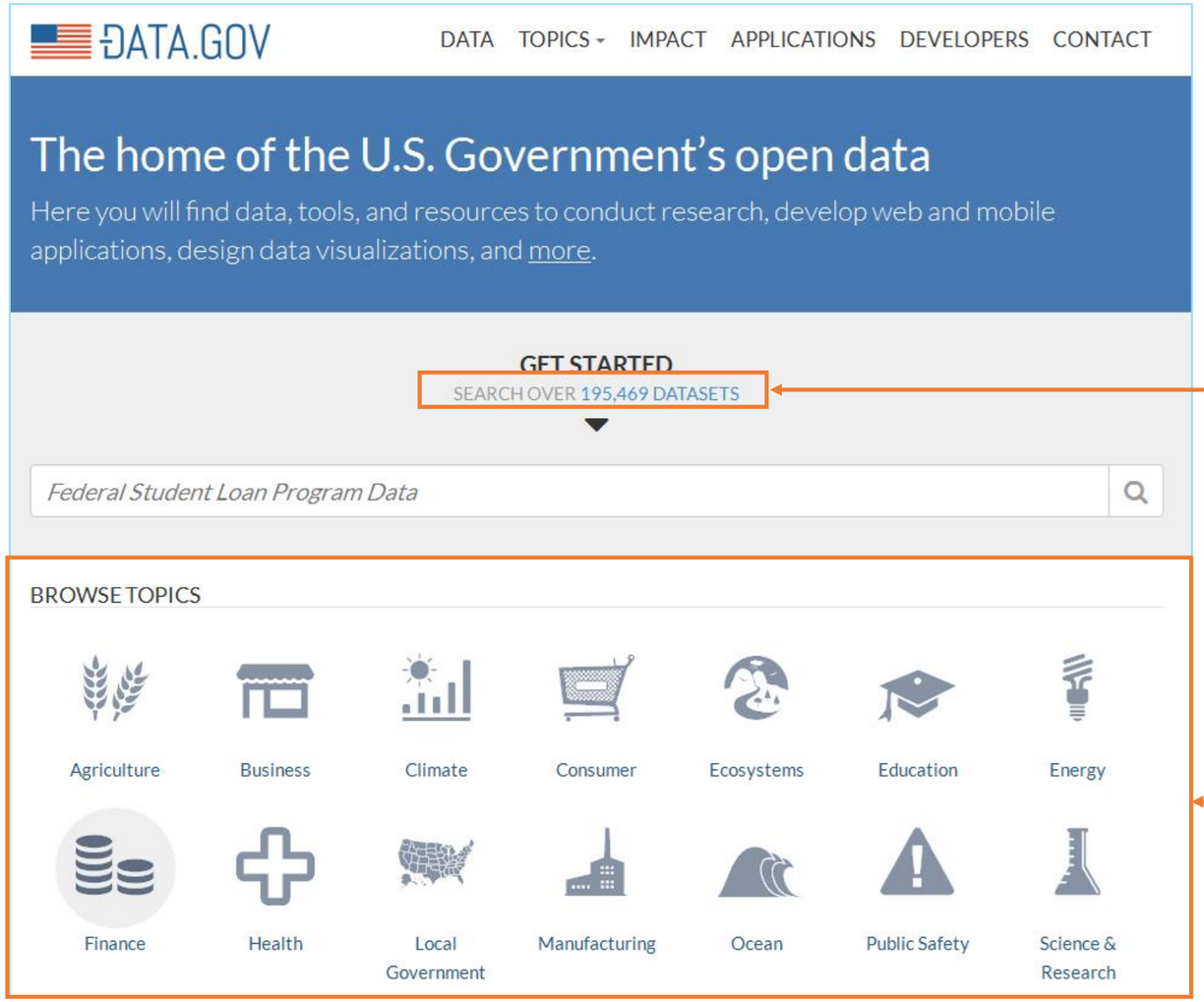


Make informed decisions

Using Data Science—Public Sector

The 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.



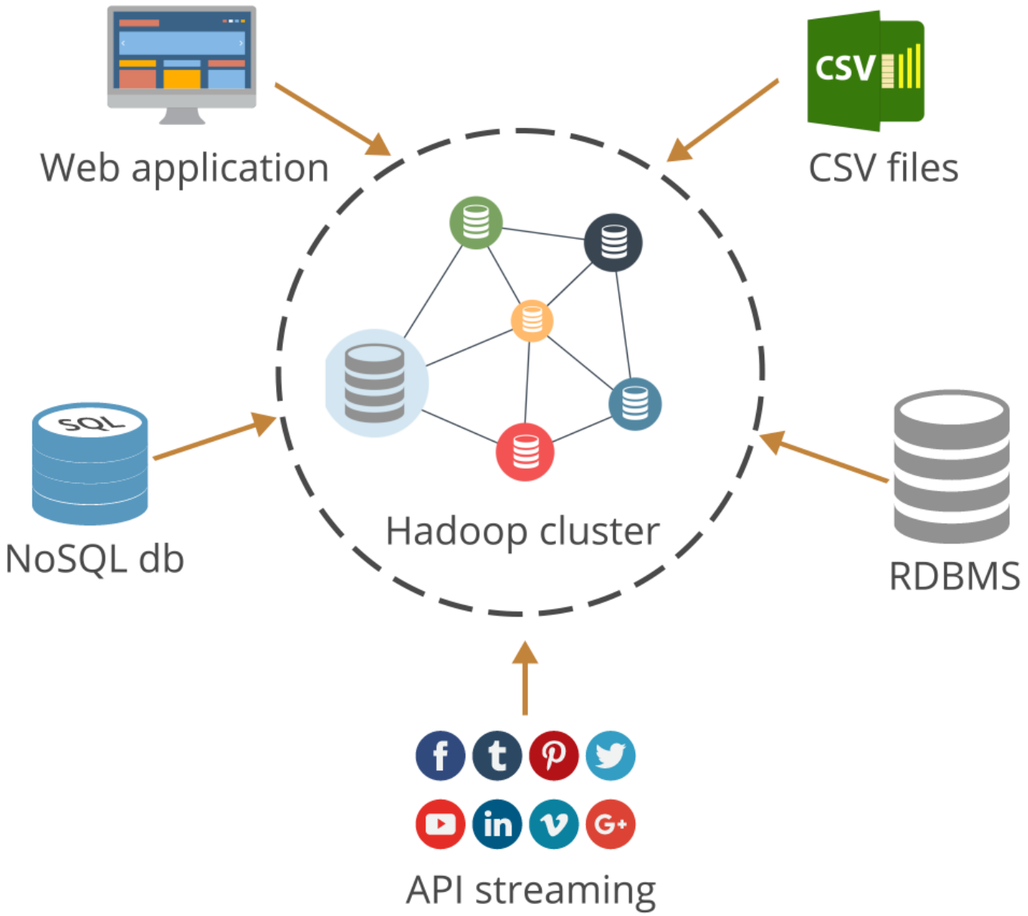
Large collection of datasets

Sectors/Domains

The Real Challenge

Some of the challenges Data Scientists face in the real world are listed here.

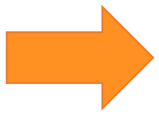
- Data quality doesn't conform to the set standards.
- Data integration is a complex task.
- Data is distributed into large clusters in HDFS, which is difficult to integrate and analyze.
- Unstructured and semi-structured data are harder to analyze.

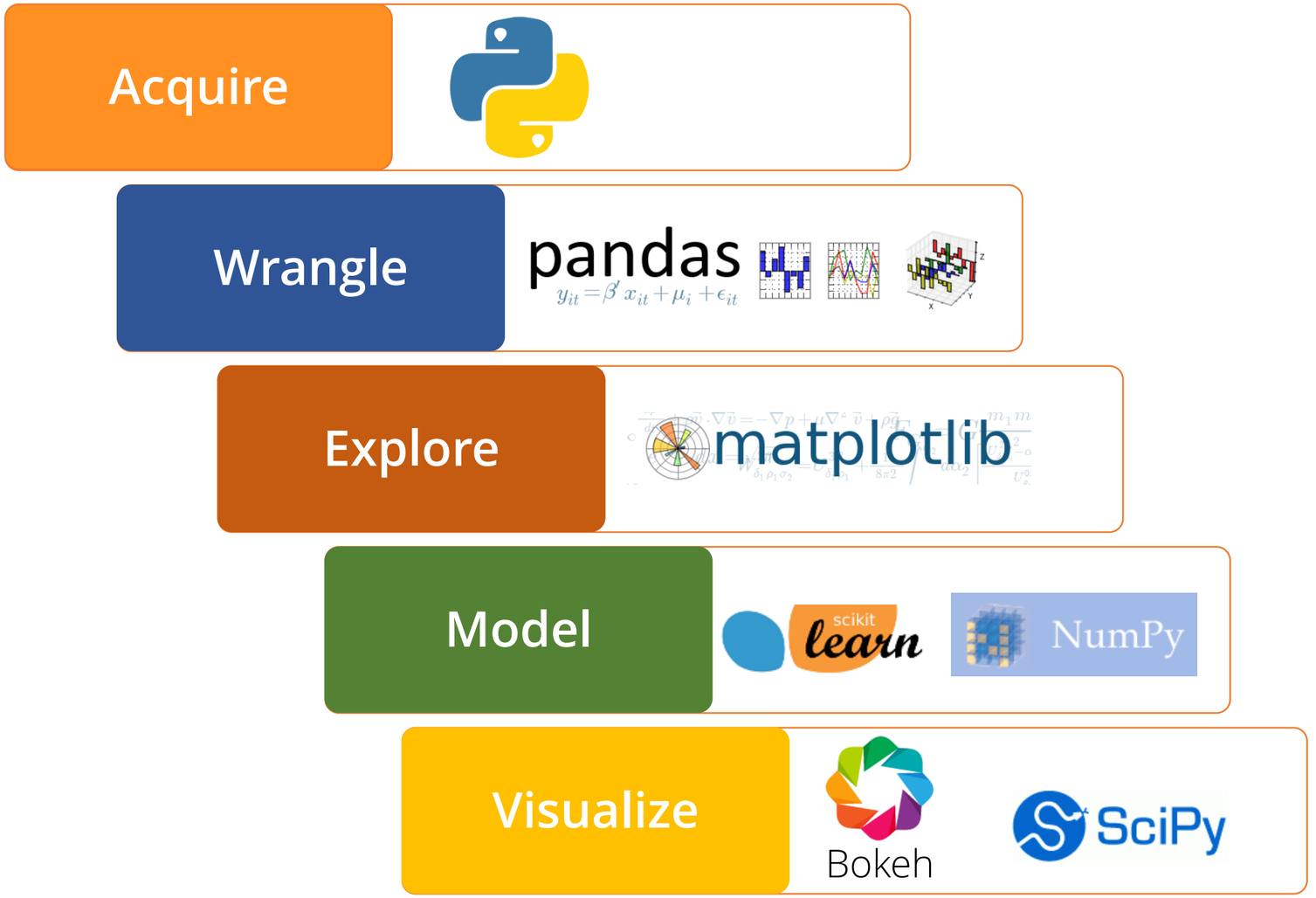


Data Analytics and Python

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

Data Analytics





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

Scientific computing

NumPy

Learn

Machine learning

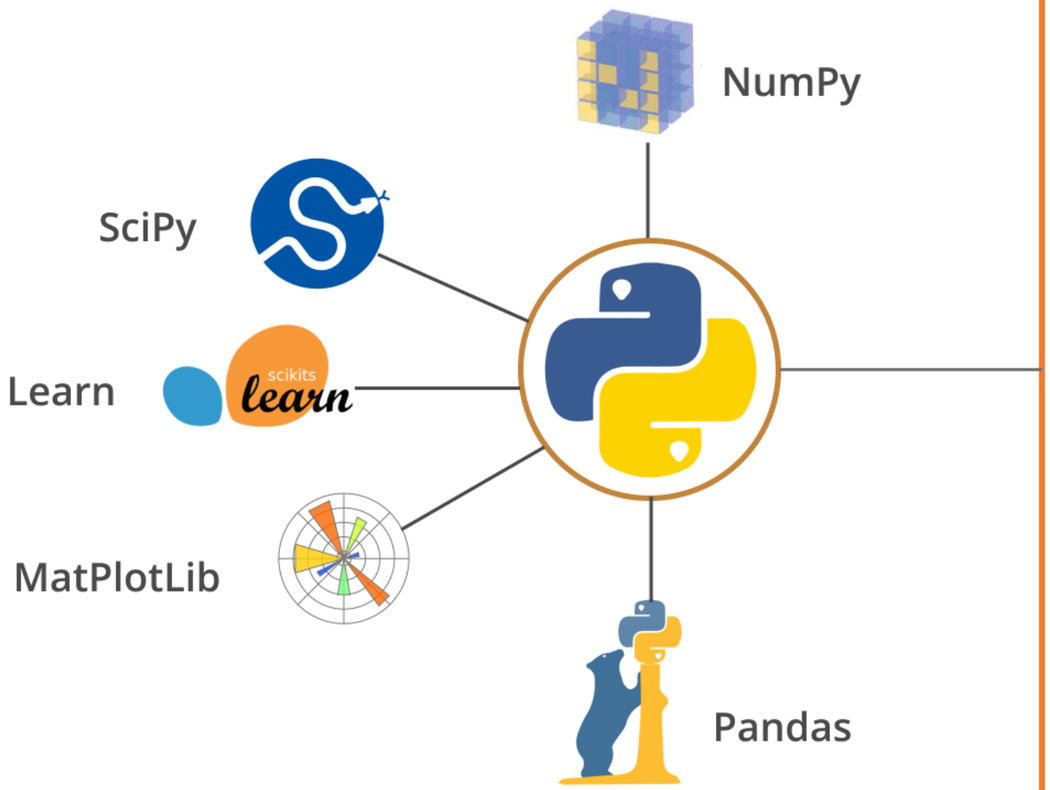
Software for mathematics, science, and engineering

SciPy

MatPlotLib

Graphics computing

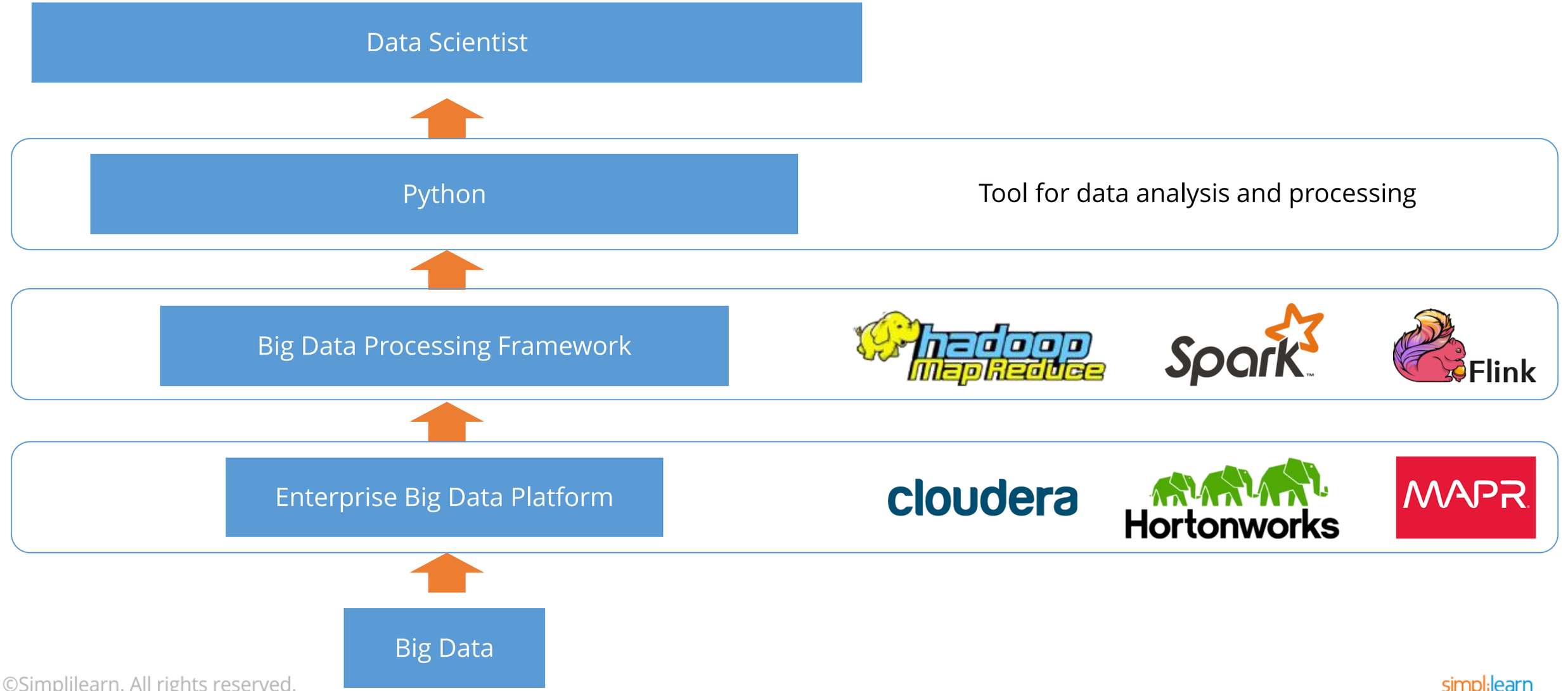
Benefits of Python



- 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 an efficient way.



Key Takeaways

- Data Science is a discipline that combines aspects of statistics, mathematics, programming, and domain expertise.
- Data Scientists solve big problems in public and private sectors.
- A lot of datasets are freely available to apply Data Science and turn them into data services and data products.
- Data Scientists are more in demand with the evolution of Big Data and real-time analytics.
- Python is a powerful language and a preferred tool for Data Science.

Quiz

QUIZ

1

A Data Scientist _____.

a. asks the right questions

b. acquires data

c. performs data wrangling and data visualization

d. All of the above

QUIZ 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**.

Explanation: 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.

QUIZ 2

The Search Engine's Autocomplete feature identifies unique and verifiable users who search for a particular keyword or phrase _____. Select all that apply.

a. to scrub inappropriate content.

b. to build a Query Volume.

c. to tag the location to a query.

d. to find similar instances on the web.

QUIZ 2

The Search Engine's Autocomplete feature identifies unique and verifiable users who search for a particular keyword or phrase _____. Select all that apply.

a. to scrub inappropriate content

b. to build a Query Volume

c. to tag the location to a query

d. to find similar instances on the web

The correct answer is **b, c**.

Explanation: 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 tag them to the query, enabling it to be location-specific.

QUIZ 3

What is the sequential flow of Data Analytics?

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

QUIZ 3

What is the sequential flow of Data Analytics?

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**.

Explanation: In Data Analytics, 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.

This concludes “Data Science Overview.”

The next lesson is “Data Analytics Overview.”