Applied Data Science with python-CSE3038

Module 1

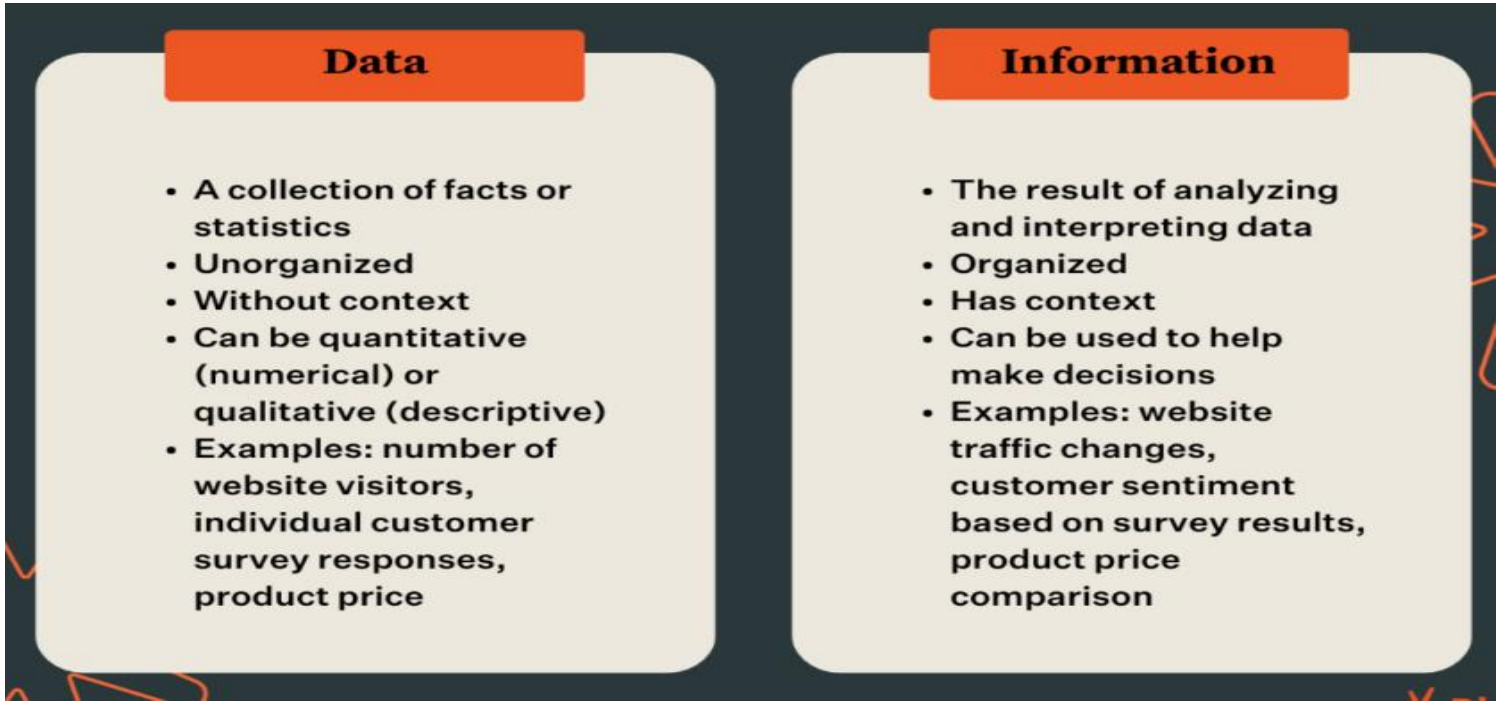
Dr. Srabana Pramanik
Assistant professor
Dept. of SCSE
Presidency University

What is Data?

- Data is defined as a collection of individual facts or statistics.
- There are two main types of data:
- **Quantitative data** is provided in numerical form, like the weight, volume, or cost of an item.
- **Qualitative data** is descriptive, but non-numerical, like the name, Gender, or eye color of a person.

What is information ?

- Information is defined as knowledge gained through study, communication, research, or instruction.



Data

- A collection of facts or statistics
- Unorganized
- Without context
- Can be quantitative (numerical) or qualitative (descriptive)
- Examples: number of website visitors, individual customer survey responses, product price

Information

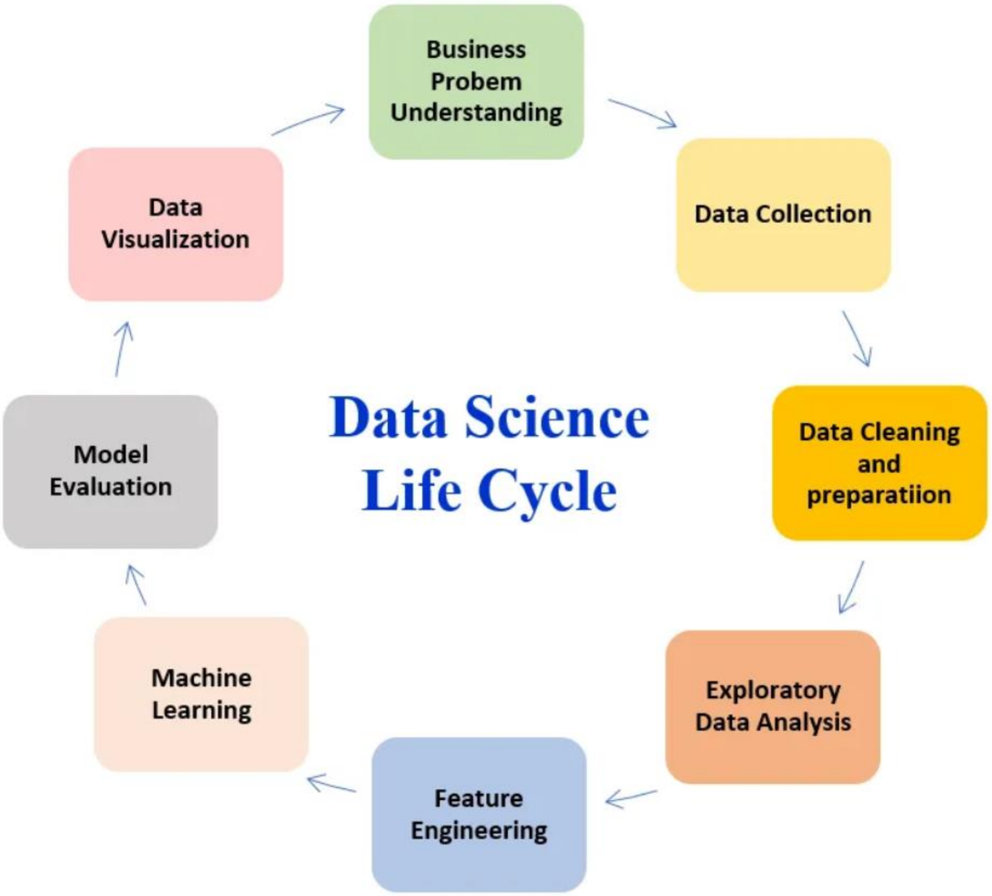
- The result of analyzing and interpreting data
- Organized
- Has context
- Can be used to help make decisions
- Examples: website traffic changes, customer sentiment based on survey results, product price comparison

Definition of data science

- Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it.
- Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization’s data.
- These insights can be used to guide decision making and strategic planning.

The data science life cycle

- The data science lifecycle involves various roles, tools, and processes, which enables analysts to glean actionable insights. Typically, a data science project undergoes the following stages:
- **Data Collection**: The lifecycle begins with the data collection--both raw structured and unstructured data from all relevant sources using a variety of methods.
- These methods can include manual entry, web scraping, and real-time streaming data from systems and devices.
- Data sources can include structured data, such as customer data, along with unstructured data like log files, video, audio, pictures, the Internet of Things (IoT), social media, and more.
- **Data storage and data processing:** Since data can have different formats and structures, companies need to consider different storage systems based on the type of data that needs to be captured.
- Data management teams help to set standards around data storage and structure, which facilitate workflows around analytics, machine learning and deep learning models.
- This stage includes cleaning data, deduplicating, transforming and combining the data using ETL (extract, transform, load) jobs or other data integration technologies.
- This data preparation is essential for promoting data quality before loading into a data warehouse, data lake, or other repository.



- **Data analysis:** Here, data scientists conduct an exploratory data analysis to examine biases, patterns, ranges, and distributions of values within the data.
- This data analytics exploration drives hypothesis generation for a/b testing. It also allows analysts to determine the data’s relevance for use within modeling efforts for predictive analytics, machine learning, and/or deep learning.
- Depending on a model’s accuracy, organizations can become reliant on these insights for business decision making, allowing them to drive more scalability.
- **Communicate:** Finally, insights are presented as reports and other data visualizations that make the insights—and their impact on business—easier for business analysts and other decision-makers to understand.
- A data science programming language such as R or Python includes components for generating visualizations; alternately, data scientists can use dedicated visualization tools.

Data universe

- The term "data universe" refers to the entirety of data that is relevant to a particular context, problem, or domain.
- It encompasses all the data available or potentially accessible for analysis, exploration, and decision-making within a specific scope.
- In a broader sense, the data universe can encompass a wide range of data types, sources, and formats.
- Which include
- Structured data (such as databases and spreadsheets),
- Semi-structured data (like JSON or XML files), and even
- Unstructured data-(like text documents and images ,Video files, audio files, etc.)
- real-time streaming data.
- By comprehensively exploring the data universe, data scientists and analysts can derive insights, patterns, and trends that lead to informed decision-making and valuable outcomes.

Sources of Data

- Data can be gathered from various places, both digital and physical.
- The source of data depends on the type of information you're seeking and the context of your analysis.
- Here are some common sources of data:
- **Databases**: Structured data can be sourced from relational databases, data warehouses, and NoSQL databases.
- These can include customer records, sales transactions, inventory data, and more.
- **Files**: Data can be obtained from various file formats, such as CSV, Excel spreadsheets, JSON, XML, and more.
- These files might contain survey responses, logs, or other structured data.
- **Web APIs**: Many online platforms and services offer APIs (Application Programming Interfaces) that allow you to access and retrieve data.
- Examples include social media platforms, weather data providers, and financial data services.
- **Web Scraping**: You can extract data from websites using web scraping techniques. This is useful for collecting data that might not be available through APIs.
- **Sensor Data**: In fields like IoT (Internet of Things), data can be collected from sensors embedded in devices, vehicles, machinery, and other physical objects.

- **Textual Data**: Text data can come from sources like articles, books, social media posts, emails, and more. Natural language processing (NLP) techniques are often used to analyze and extract insights from textual data.
- **Images and Videos**: Visual data, such as images and videos, can provide valuable information. Computer vision techniques are used to analyze and interpret this type of data.
- **Surveys and Questionnaires**: Data can be collected through surveys, questionnaires, and feedback forms. This can provide insights into user preferences, opinions, and behaviors.
- **Public Data Repositories**: Various organizations and governments provide publicly available datasets for research and analysis.
- Examples include data.gov, Kaggle, and UCI Machine Learning Repository.
- **Social Media**: Social media platforms are a rich source of user-generated content, including text, images, videos, and interactions.
- **Historical Records**: Archives, historical documents, and records can provide insights into the past and support historical research.
- **Physical Measurements**: Scientific experiments, research studies, and industrial processes often generate data through physical measurements and observations.
- **Transaction Logs**: E-commerce platforms, financial institutions, and online services often maintain transaction logs that can be used for analysis.
- **Private Data**: Organizations may have internal databases and records that contain proprietary or sensitive information.

It's important to note that data collection should always adhere to ethical and legal considerations, such as privacy regulations and intellectual property rights.

Additionally, the quality and reliability of the data source are crucial factors for ensuring accurate and meaningful analyses.

Application of Data Science

Data science has a wide range of applications across various industries and fields. Here are some notable examples of how data science is applied:

- **Business and Marketing:**
  - Customer Segmentation: Data science helps businesses identify distinct customer groups based on behavior, preferences, and demographics for targeted marketing.
  - Market Analysis: Analyzing market trends, competitor data, and consumer sentiment to inform business strategies and product development.
  - Recommender Systems: Creating personalized recommendations for products, services, or content based on user behavior and preferences.
  - Pricing Optimization: Utilizing data to set optimal prices for products or services based on market demand and competition.
- **Healthcare and Medicine:**
  - Disease Prediction and Diagnosis: Applying data science to medical records, genetic data, and other health data for early disease detection and accurate diagnosis.
  - Drug Discovery: Using data analysis and machine learning to identify potential drug candidates and predict their effectiveness.
  - Medical Imaging Analysis: Enhancing medical image interpretation through image recognition and analysis algorithms for improved diagnosis.
  - Health Monitoring: Developing wearable devices and apps that collect and analyze health data for monitoring and early intervention.

- **Manufacturing and Supply Chain:**
  - Demand Forecasting: Predicting future demand for products to optimize inventory management and production planning.
  - Quality Control: Analyzing sensor data and production metrics to identify defects and improve product quality.
  - Supply Chain Optimization: Optimizing the flow of goods, reducing costs, and improving efficiency through data-driven insights.
  - Preventive Maintenance: Using data to predict equipment failures and schedule maintenance to minimize downtime.
- **Energy and Utilities:**
  - Energy Consumption Analysis: Monitoring and analyzing energy consumption patterns to identify opportunities for energy efficiency.
  - Grid Management: Optimizing energy distribution and load balancing in smart grids using real-time data.
  - Predictive Maintenance: Anticipating equipment failures in power plants and infrastructure to minimize disruptions.

- **Finance and Banking:**
  - Fraud Detection: Identifying unusual patterns in financial transactions to detect and prevent fraudulent activities.
  - Risk Assessment: Evaluating credit risk, investment opportunities, and market trends to make informed financial decisions.
  - Algorithmic Trading: Using data science to develop trading algorithms that leverage historical data and market signals.
  - Personalized Financial Advice: Providing customized financial recommendations and investment strategies based on individual goals and risk tolerance.
- **Transportation and Logistics:**
  - Route Optimization: Utilizing data to find the most efficient routes for delivery vehicles, reducing fuel consumption and delivery times.
  - Traffic Management: Analyzing traffic patterns to improve urban planning and alleviate congestion.
  - Ride-Sharing and Mobility Services: Matching riders with drivers, optimizing routes, and managing vehicle fleets.
- **Social Sciences and Public Policy:**
  - Social Media Analysis: Extracting insights from social media data to understand public sentiment, trends, and opinions.
  - Crime Analysis: Using data to predict crime hotspots and allocate law enforcement resources effectively.
  - Education Analytics: Analyzing student performance data to improve teaching methods and educational outcomes.
- These are just a few examples of how data science is applied in various domains. The field continues to evolve, and new applications are constantly being explored as data becomes more integral to decision-making and problem-solving in virtually every industry.

Information Commons

- Definitions for “information commons” can vary, but a generally accepted meaning has been “a specific location designated to deliver electronic resources for research and production that is maintained by technically proficient staff” (Cowgill et al., 2001).
- The term "Information Commons" typically refers to a physical or virtual space where people have access to various forms of information, resources, and technologies for learning, research, collaboration, and knowledge sharing.
- It's often associated with libraries, academic institutions, and other public spaces that provide access to a wide range of information and tools to support education and research.

Geographical
Transport
Cultural
Natural
Scientific
Types of Data
Meteorological
Financial
Statistical

Data - Wikipedia

- Information Commons can be found in various settings, including universities, public libraries, research institutions, and community centers.
- The concept highlights the importance of providing equitable access to information and technology, fostering collaborative learning environments, and supporting lifelong learning and research.
- As technology continues to advance, the role of Information Commons in facilitating information sharing and knowledge creation remains significant.

Data Science Project Life Cycle: OSEMN Framework

- The OSEMN framework is a data science methodology used for the end-to-end process of working with data to extract valuable insights and make informed decisions. Each letter in the acronym "OSEMN" stands for a key step in the data science workflow:

1. **Obtain**: In this phase, you gather the data needed for your analysis. This could involve data collection from various sources, such as databases, APIs, files, or even web scraping.
2. **Scrub (or Scrubbing)**: Here, the collected data is cleaned, preprocessed, and transformed. This step involves handling missing values, removing duplicates, and dealing with outliers. The goal is to ensure the data is accurate and ready for analysis.
3. **Explore**: In the exploration phase, you perform exploratory data analysis (EDA) to understand the data's characteristics, relationships, and patterns. Visualization and summary statistics are often used to gain insights into the data's structure and potential trends.

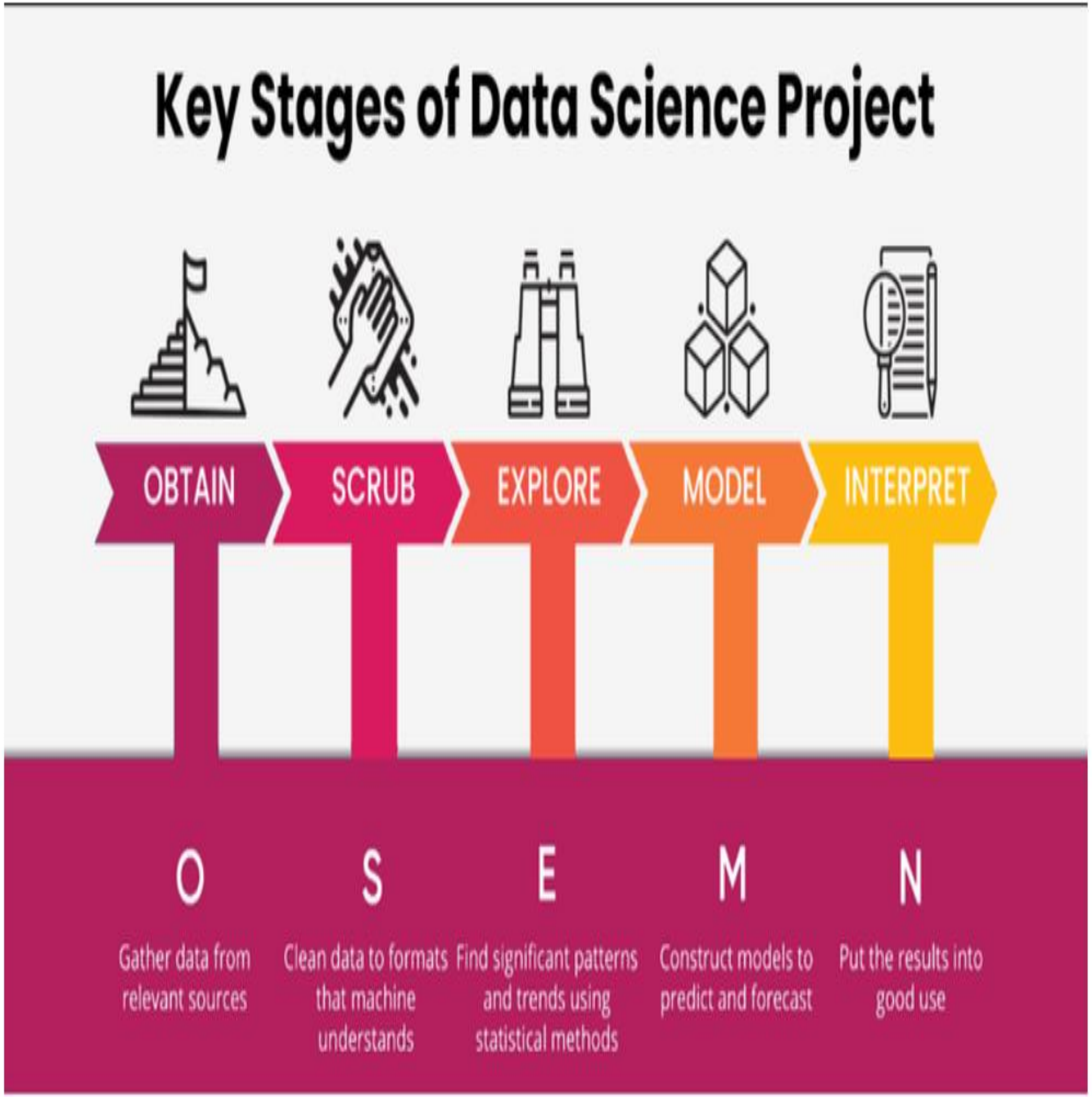
OSEMN framework

- **Model:** In this step, you build and train machine learning models or statistical models depending on the analysis goals. This involves selecting appropriate algorithms, model training, and evaluation.

- **Interpret (or Interpretation):** After obtaining model results, you interpret them to draw meaningful insights and conclusions. This step helps answer the original questions or objectives of the analysis. It may also involve refining models, exploring feature importance, and understanding how well the models perform.

- **Communicate (or Communication):** The final step involves communicating the findings and insights to stakeholders. This could be in the form of reports, dashboards, presentations, or any other means that effectively convey the results of the analysis. Clear communication is crucial for making informed decisions based on the analysis.

The OSEMN framework provides a structured approach to tackling data science projects, ensuring that important steps like data preparation and model evaluation are not overlooked. It helps data scientists and analysts manage the complexities of working with data and extracting meaningful information from it.



Key Stages of Data Science Project

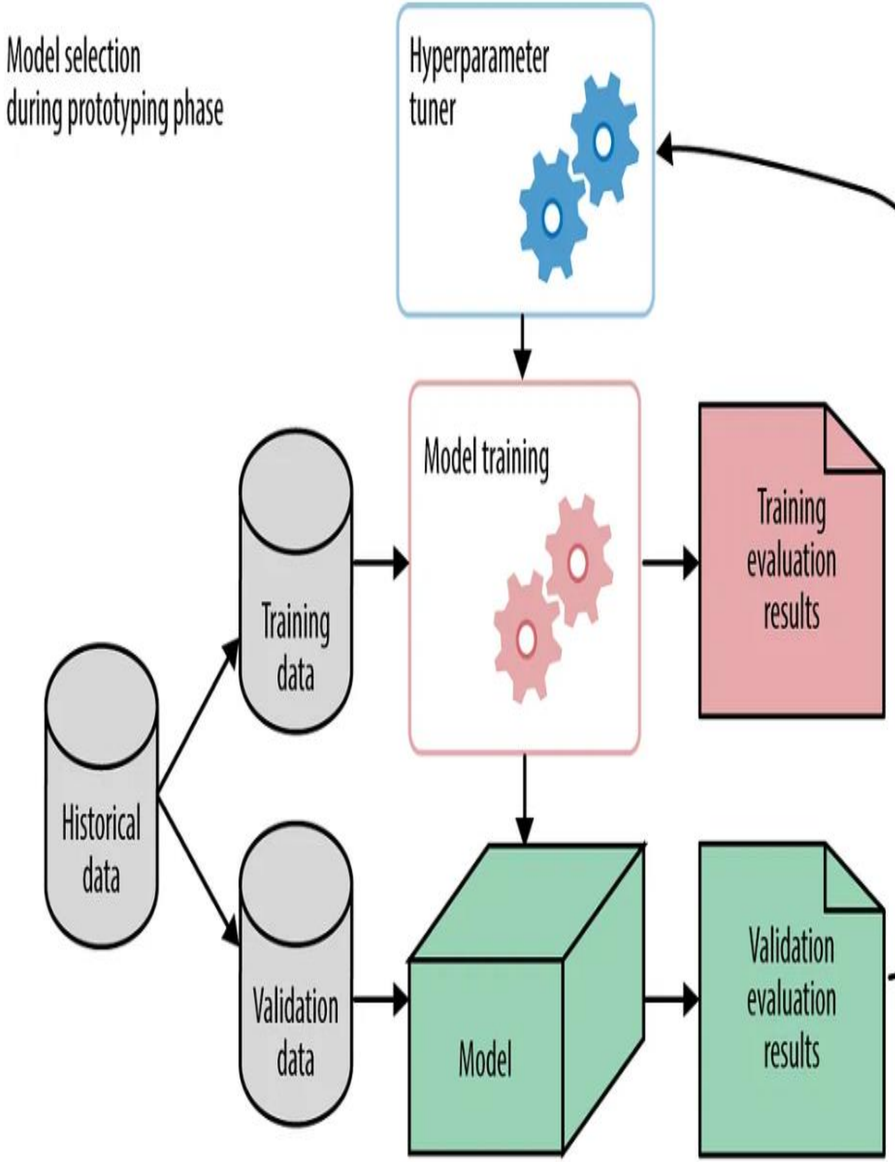
OBTAIN — O — Gather data from relevant sources

SCRUB — S — Clean data to formats that machine understands

EXPLORE — E — Find significant patterns and trends using statistical methods

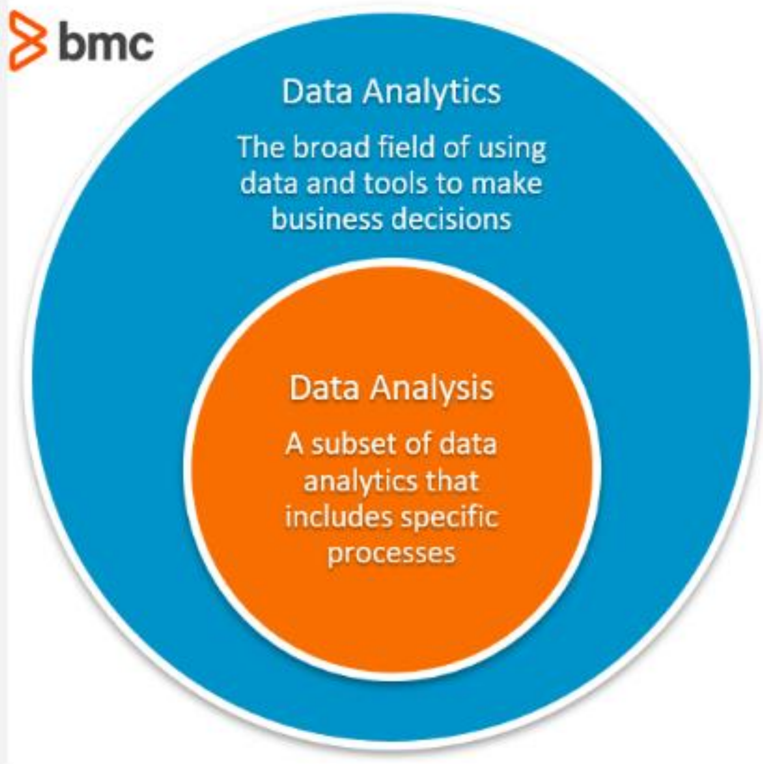
MODEL — M — Construct models to predict and forecast

INTERPRET — N — Put the results into good use



Difference between data analysis and data analytics.

- Data analysis and data analytics are related terms that often overlap, but they have distinct differences in their scope, focus, and methodologies.
- Let's explore the key differences between the two:
- **Data Analysis:** Data analysis involves the process of inspecting, cleaning, transforming, and organizing raw data to extract meaningful insights, discover patterns, and draw conclusions.
- It is a fundamental step in understanding and interpreting data.
- Data analysis is typically more focused on examining historical data and identifying trends, correlations, and relationships within the data set.
- It often includes descriptive statistics and visualization techniques to communicate findings effectively.
- Data analysis is important for making informed decisions, but it might not always involve advanced statistical or predictive techniques.


bmc
Data Analytics
The broad field of using data and tools to make business decisions
Data Analysis
A subset of data analytics that includes specific processes

- **Data Analytics:** Data analytics encompasses a broader range of activities that go beyond simple data analysis.
- It involves the application of various techniques, tools, and algorithms to explore data, uncover hidden patterns, and derive actionable insights.
- Data analytics includes not only descriptive analysis but also diagnostic, predictive, and prescriptive analysis.
- The goal of data analytics is to answer specific business questions, make predictions about future trends, and guide decision-making.
- It often involves more sophisticated statistical modeling, machine learning, and data mining techniques to provide deeper insights and predictive capabilities.
- Data analysis is a subset of data analytics.

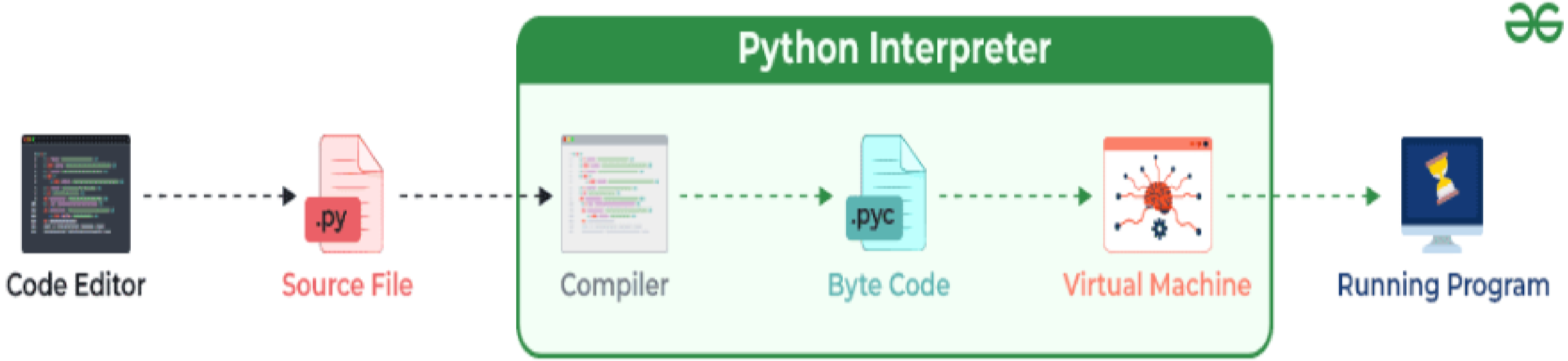
Python Programming language

- Python was developed by "Guido Van Rossum" in late 1980's in National Research Institute for Mathematics and Computer Science in Netherlands.

What is Python ?

- 1.Python is powerful programming language.
- 2.Specially used in data science, machine learning to solve daily life real-time problems and providing the solutions.
- 3.Python is a high-level, intrepeted, interactive and object oriented scripting language
- **Why python?**
- 1.Easy to understand
- 2.Easy syntax
- 3.Beginner friendly
- 4.Supports multiple libraries
- 5.Supports OOP's

INTERNAL WORKING OF PYTHON



Python doesn’t convert its code into machine code, something that hardware can understand. It converts it into something called byte code. So within Python, compilation happens, but it’s just not in a machine language. It is into byte code **(.pyc or .pyo)** and this byte code can’t be understood by the CPU. So we need an interpreter called the Python virtual machine to execute the byte codes.

PYTHON VARIABLES:

- 1.Variables are nothing but reserved memory locations to store values.
- 2.When we create a varaible ,automatically some space in memory is reserved.
- 3.Based on the data types the interpreter allocates the memory.
- 4.Varaible can take any data type such as integer,float numbers,strings etc

Python Data types

- 1.Integer data type
- 2.Float point numbers
- 3.Strings data type
- 4.Boolean data type

Python Numpy

- 1.NUMPY-Numpy stands for “Numeric Python” or “Numerical python.
- 2.Numpy is a package that contains several classes, functions, variables etc. to deal with scientific calculations in Python.
- 3.Numpy is useful to create and process single and multi-dimensional arrays.
- 4.In addition, numpy contains a large library of mathematics like linear algebra functions and Fourier transformations.
- NOTE: The arrays which are created using numpy are called n dimensional arrays where n can be any integer.
- If n = 1 it represent a one dimensional array.
- If n= 2, it is a two dimensional array etc.
- Numpy array can accept only one type of elements.
- We cannot store different data types into same arrays

Feature of NumPy