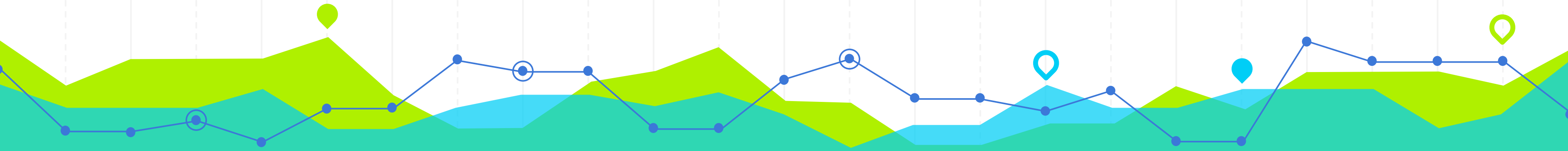
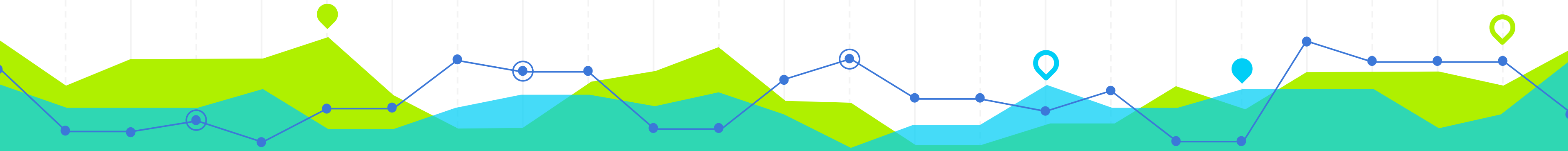


The Complete Guide to Data : Science, Analysis, Engineering & Roadmap



DATA SCIENCE & DATA ANALYTICS

Their uses, processes and difference

1

WHAT IS DATA SCIENCE?

- Data science is an interdisciplinary field that involves the extraction of insights and knowledge from data using a combination of statistical and computational techniques. It combines various disciplines such as mathematics, statistics, computer science, and domain-specific knowledge to analyze and interpret complex data sets.
- Data science involves a series of processes, including data collection, data preparation, data analysis, and data visualization. It also includes the use of machine learning and other advanced analytical techniques to create predictive models and make data-driven decisions.
- Data science is used in a variety of industries, including finance, healthcare, retail, and transportation. It has also become increasingly important in the field of scientific research, as it enables researchers to extract insights and knowledge from large and complex data sets.
- The goal of data science is to extract valuable insights and knowledge from data that can be used to inform decision-making and drive innovation. With the rapid growth of data in recent years, data science has become a crucial field for businesses and organizations looking to gain a competitive advantage and drive growth

USES OF DATA SCIENCE

Business and Finance: used to optimize business processes, improve customer engagement, detect fraud, and predict market trends.

1.**Healthcare:** used to improve patient outcomes, reduce costs, and enhance patient experience. It is used for patient monitoring, personalized medicine, disease diagnosis, and drug development.

1.**Retail:** used to optimize pricing, predict demand, and improve customer experience. It is used for customer segmentation, product recommendation, supply chain management, and inventory optimization.

USES OF DATA SCIENCE

Transportation: used to optimize routes, reduce congestion, and improve safety. It is used for traffic management, predictive maintenance, autonomous vehicles, and logistics optimization.

1. **Scientific research**: used to analyze large and complex data sets to gain insights into various phenomena. It is used in fields such as astrophysics, genomics, environmental science, and social sciences.

1. **Social Media and Marketing**: used to analyze customer behavior, develop marketing strategies, and measure the effectiveness of marketing campaigns. It is used for sentiment analysis, social network analysis, and customer segmentation.

HOW IS DATA SCIENCE EVOLVING?

Some of the ways in which data science is changing include:

1.**Increasing use of automation**: With the development of more advanced tools and algorithms, many routine data science tasks are becoming automated. This frees up data scientists to focus on more complex problems.

2.**More focus on explainability**: As machine learning models become more sophisticated, there is a growing need for them to be transparent and explainable. This is particularly important in fields such as healthcare and finance where decisions based on data can have significant consequences.

3.**Greater emphasis on ethical considerations**: Data scientists are increasingly aware of the ethical implications of their work, particularly in areas such as privacy and bias. There is a growing focus on developing ethical frameworks and guidelines to ensure that data science is used responsibly.

HOW IS DATA SCIENCE EVOLVING?

4.Integration of data science with other disciplines: Data science is increasingly being used in combination with other disciplines such as biology, economics, and psychology to address complex problems that require a multidisciplinary approach.

5.Growing importance of real-time data: As more and more devices are connected to the internet, there is an increasing demand for real-time data analysis. This requires the development of new techniques and tools to handle large volumes of data in real-time.

Overall, data science is changing rapidly and is likely to continue to do so in the coming years as new technologies and techniques are developed.

WHAT IS DATA ANALYTICS?

- The process of inspecting, cleaning, transforming, and modeling data in order to extract meaningful insights and information. It involves the use of statistical and computational techniques to identify patterns, trends, and relationships within data sets.

- It can be performed on different types of data, such as numerical, categorical, or text data, and can involve different levels of complexity, from simple descriptive statistics to advanced machine learning models.

- Data analytics is used in a wide range of applications, including business intelligence, scientific research, and social science.

STEPS OF DATA ANALYTICS

Data Collection: The first step in data analysis is to collect relevant data. This can involve various methods, including surveys, experiments, or observations.

1. **Data Cleaning**: Once the data has been collected, it is important to clean it by removing any irrelevant or duplicate data points, dealing with missing data, and correcting errors.

1. **Data Transformation**: In order to make the data suitable for analysis, it may need to be transformed into a different format or structure. This can involve tasks such as data normalization, feature extraction, or data aggregation.

STEPS OF DATA ANALYTICS

Data Modeling: The next step is to create a model that can be used to analyze the data. This can involve statistical techniques such as regression analysis or machine learning algorithms such as decision trees or neural networks.

1. **Data Visualization**: Once the analysis has been performed, it is important to present the results in a clear and meaningful way. This can involve creating graphs, charts, or other visualizations that can help to communicate the insights that have been gained.

DATA SCIENCE VS DATA ANALYTICS

Data Science:

A broader field that involves the use of statistical and computational techniques to extract insights and knowledge from data.

It includes tasks such as data collection, data cleaning, data transformation, data modeling, and data visualization.

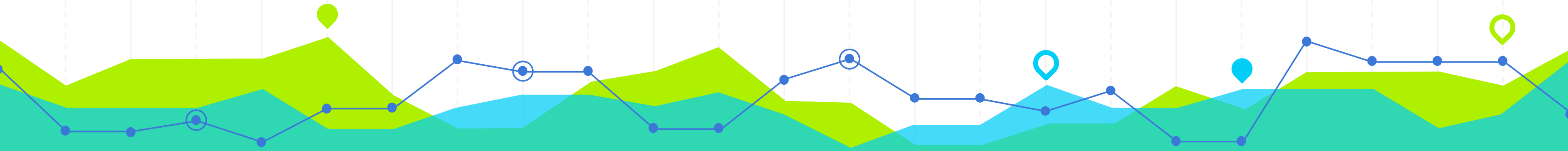
Data science often involves the use of advanced analytical techniques such as machine learning to create predictive models and drive decision-making.

Data Analytics:

A more specialized field that focuses on the analysis of data to extract insights and knowledge.

It typically involves tasks such as data cleaning, data modeling, and data visualization, but may not involve tasks such as data collection or data transformation.

Data analytics often involves the use of descriptive statistical techniques to summarize and visualize data, as well as predictive techniques such as regression analysis to make predictions based on historical data.



DATA SCIENCTIST

ROLES AND RESPONSIBILITIES

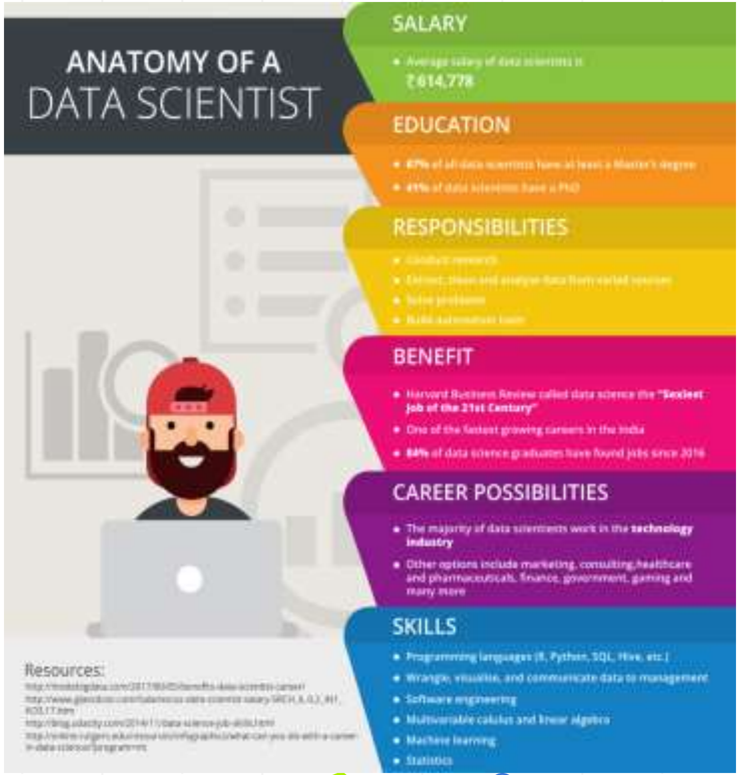
2

WHO IS A DATA SCIENCTIST?

- A data scientist is a professional who is skilled in the field of data science, which involves the use of statistical and computational techniques to extract insights and knowledge from data. Data scientists are responsible for designing, developing, and implementing algorithms and models to analyze data, as well as interpreting and communicating the results of their analyses to stakeholders.
- Data scientists typically have expertise in a variety of areas, including statistics, mathematics, computer science, and domain-specific knowledge in fields such as business, healthcare, or social science. They are also proficient in programming languages such as Python or R, and are familiar with various tools and technologies used in data science, such as data visualization software and machine learning libraries.

REQUIRED SKILLS OF A DATA SCIENTIST

- PYTHON/R
- MACHINE LEARNING



ANATOMY OF A DATA SCIENTIST

SALARY

- Average salary of data scientists is ₹614,778

EDUCATION

- **87%** of all data scientists have at least a Master's degree
- **41%** of data scientists have a PhD

RESPONSIBILITIES

- Conduct research
- Extract, clean and analyse data from varied sources
- Solve problems
- Build automation tools

BENEFIT

- Harvard Business Review called data science the **"Sexiest Job of the 21st Century"**
- One of the fastest growing careers in the India
- **84%** of data science graduates have found jobs since 2016

CAREER POSSIBILITIES

- The majority of data scientists work in the **technology industry**
- Other options include marketing, consulting,healthcare and pharmaceuticals, finance, government, gaming and many more

SKILLS

- Programming languages (R, Python, SQL, Hive, etc.)
- Wrangle, visualize, and communicate data to management
- Software engineering
- Multivariable calculus and linear algebra
- Machine learning
- Statistics

Resources:

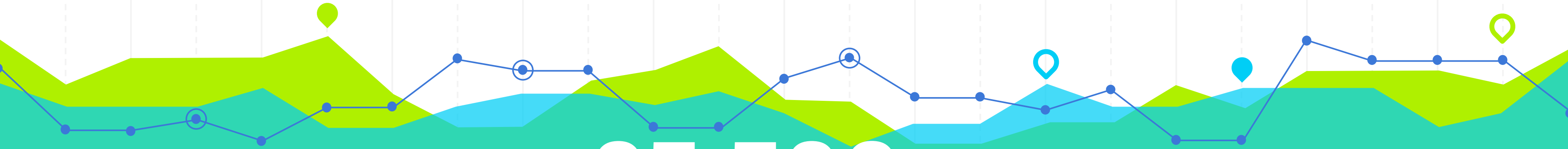
[illegible]

[illegible]

[illegible]

[illegible]

- STATISTICAL MODELLING
- DATA VISUALIZATION



85,590

Whoa! That’s the total number of Data Scientists in the US.

Spread across the following Industries:

1. **Computer Systems**
2. **Management Companies**
3. **Scientific and Technical Consultancies**
4. **Insurance Carriers**
5. **Web Search Portals**

In Depth Analysis of Roles and Their Skills

DATA ANALYST

Extract, clean, and analyze data to uncover insights and trends.

Create reports and dashboards to communicate findings to stakeholders.

SKILLS :

- Data visualization tools
- SQL
- Excel
- Statistical analysis

MACHINE LEARNING ENGINEER

Build and deploy machine learning models at scale.

Optimize and tune models for performance.

Design and implement systems to manage and process large datasets..

SKILLS :

- R/Python
- Machine learning
- Deep learning
- Natural language processing (NLP)

In Depth Analysis of Roles and Their Skills

BUSINESS INTELLIGENCE ANALYST

Analyze business metrics and KPIs to identify trends and opportunities.

Create reports and dashboards to communicate insights to stakeholders.

Work with cross-functional teams to define and track metrics and goals.

SKILLS :

- SQL
- Data visualization
- Data modeling
- Reporting

DATA ENGINEER

Design, build, and maintain data pipelines to ensure data quality and integrity.

Manage and optimize databases and data warehouses.

Work with cross-functional teams to ensure data infrastructure meets business needs.

SKILLS :

- SQL
- ETL tools
- Data warehousing
- Cloud computing

In Depth Analysis of Roles and Their Skills

DATA VISUALIZATION DESIGNER

Create compelling visualizations to communicate complex data and insights.

Work with stakeholders to understand their needs and develop effective visualizations.

Maintain design consistency across multiple visualizations.

SKILLS :

- Data visualization tools
- Graphic design
- Storytelling

In Depth Analysis of Roles and Their Skills

AI DESIGNER

Conducting research on artificial intelligence technologies, developing new algorithms and models.

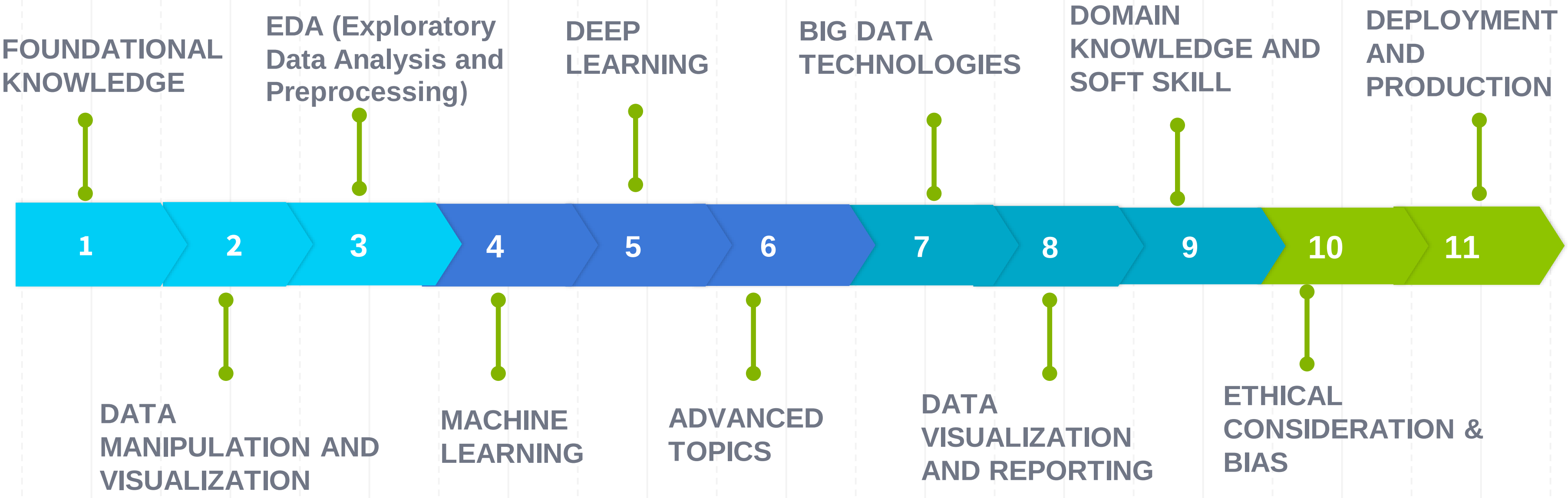
SKILLS :

- Artificial intelligence
- Machine learning
- Deep learning
- Natural language processing

DATA SCIENCTIST ROADMAP

3

ROADMAP



1 FOUNDATIONAL KNOWLEDGE

2 DATA MANIPULATION AND VISUALIZATION

3 EDA (Exploratory Data Analysis and Preprocessing)

4 MACHINE LEARNING

5 DEEP LEARNING

6 ADVANCED TOPICS

7 BIG DATA TECHNOLOGIES

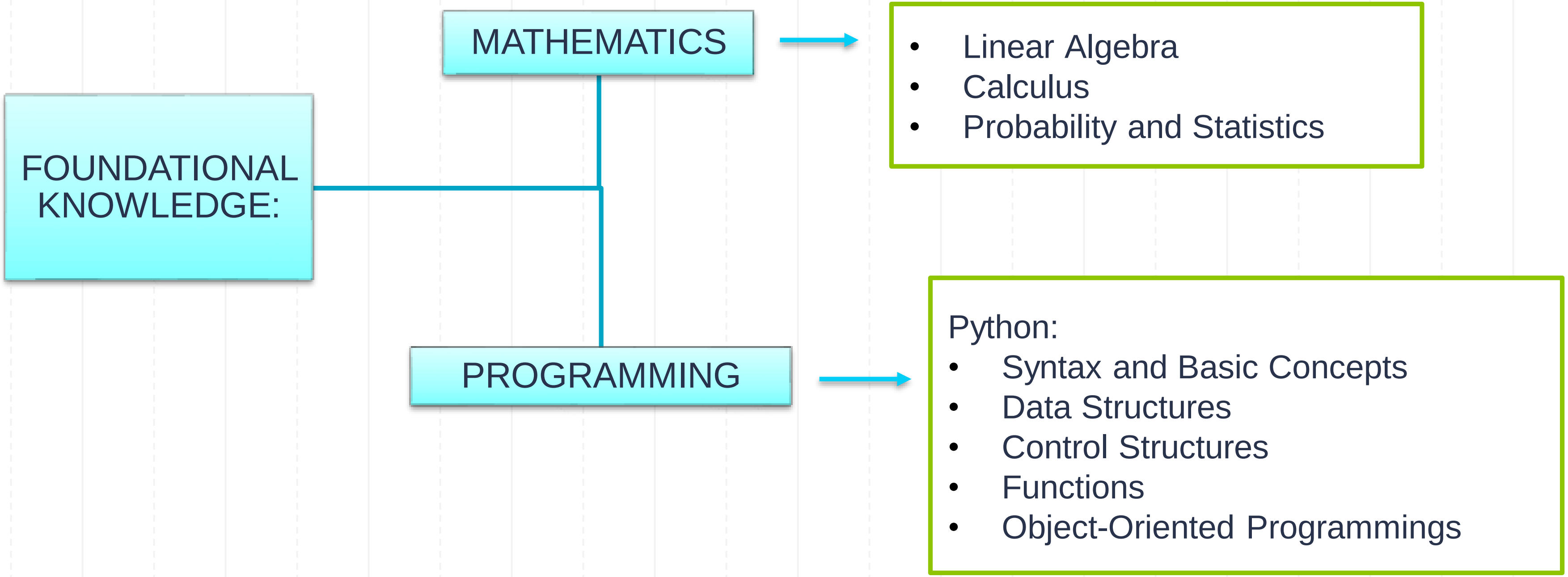
8 DATA VISUALIZATION AND REPORTING

9 DOMAIN KNOWLEDGE AND SOFT SKILL

10 ETHICAL CONSIDERATION & BIAS

11 DEPLOYMENT AND PRODUCTION

STEP 1:



FOUNDATIONAL KNOWLEDGE:

MATHEMATICS

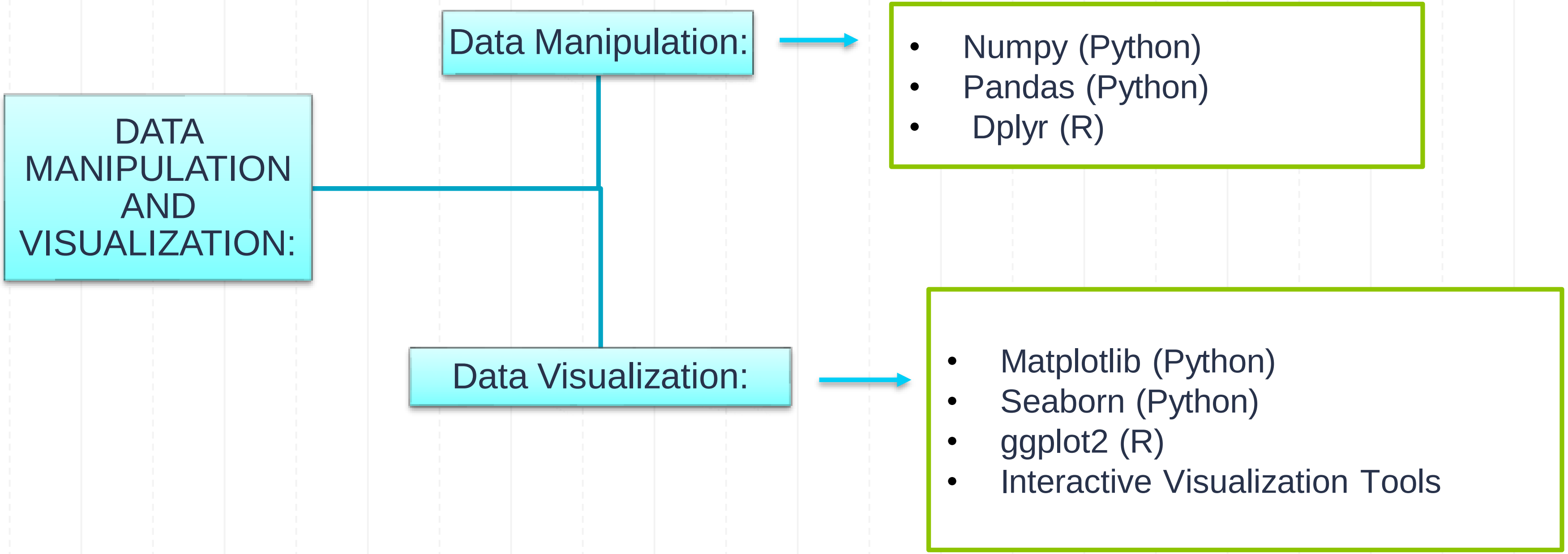
- Linear Algebra
- Calculus
- Probability and Statistics

PROGRAMMING

Python:

- Syntax and Basic Concepts
- Data Structures
- Control Structures
- Functions
- Object-Oriented Programmings

STEP 2:



DATA MANIPULATION AND VISUALIZATION:

Data Manipulation:

- Numpy (Python)
- Pandas (Python)
- Dplyr (R)

Data Visualization:

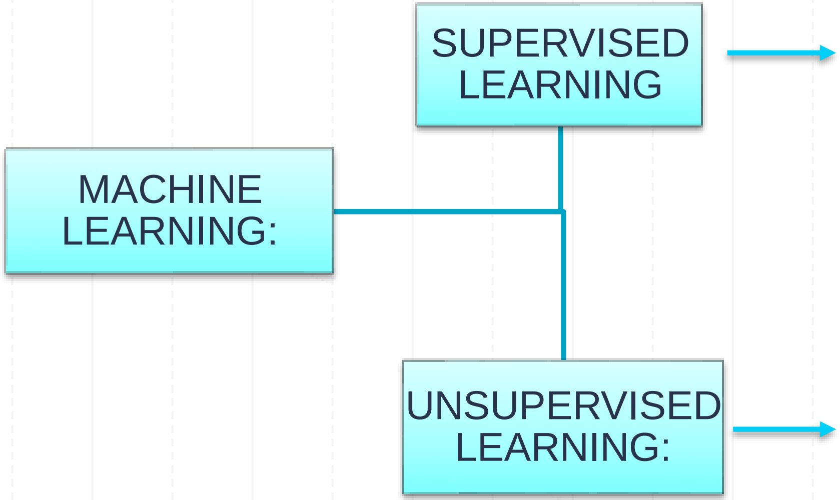
- Matplotlib (Python)
- Seaborn (Python)
- ggplot2 (R)
- Interactive Visualization Tools

STEP 3:

Exploratory Data Analysis (EDA) and Preprocessing:

- Exploratory Data Analysis Techniques
- Feature Engineering
- Data Cleaning
- Handling Missing Data
- Data Scaling and Normalization
- Outlier Detection and Treatment

STEP 4:



MACHINE LEARNING:

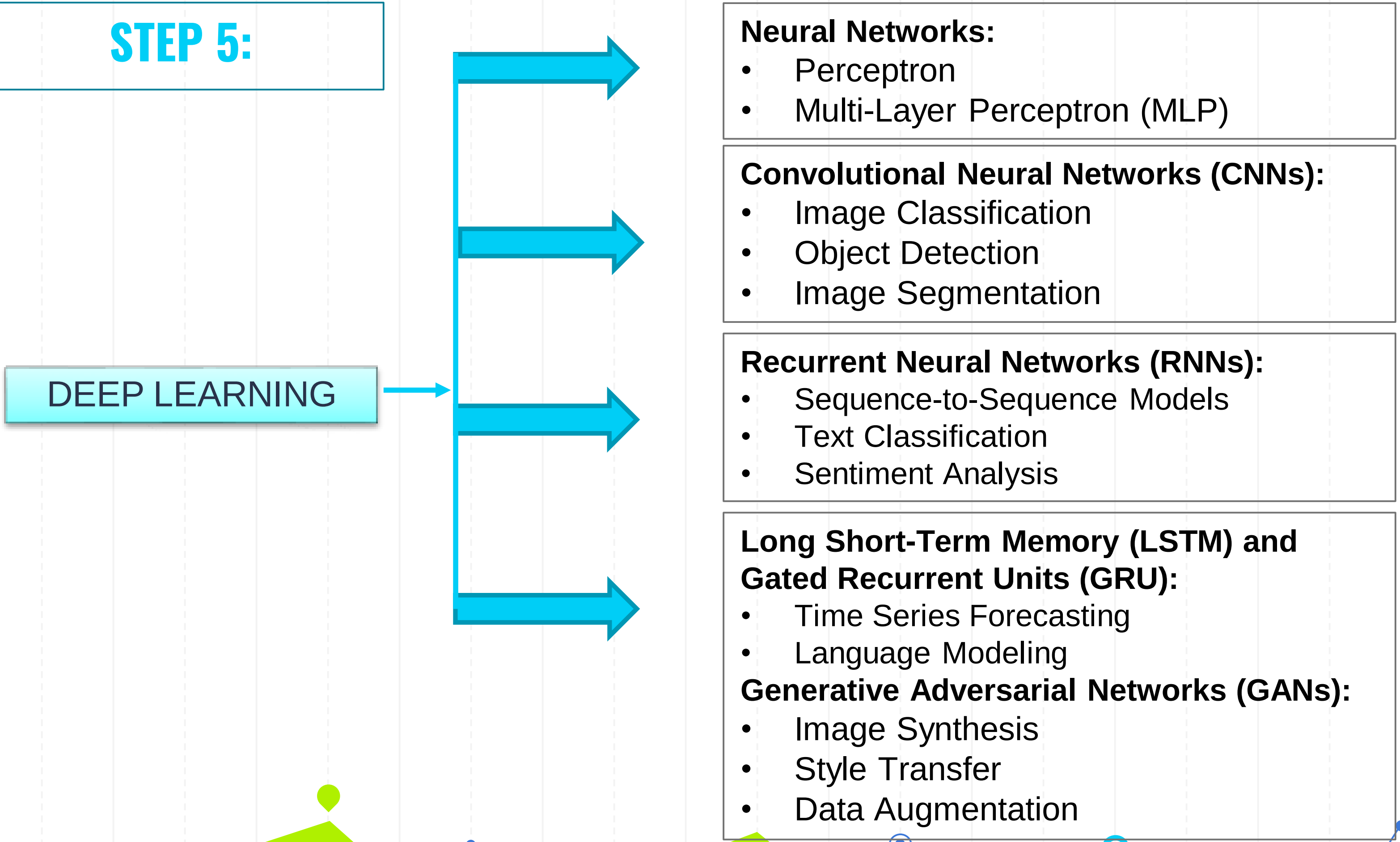
SUPERVISED LEARNING

UNSUPERVISED LEARNING:

- Linear Regression
- Polynomial Regression
- Regularization Techniques
- Classification:
- Logistic Regression
- k-Nearest Neighbors (k-NN)
- Support Vector Machines (SVM)
- Decision Trees
- Random Forest
- Gradient Boosting

- K-means
- DBSCAN
- Hierarchical Clustering
- Dimensionality Reduction:
- Principal Component Analysis (PCA)
- t-Distributed Stochastic Neighbor Embedding
- Linear Discriminant Analysis (LDA)
- Association Rule Learning

STEP 5:



DEEP LEARNING

Neural Networks:

- Perceptron
- Multi-Layer Perceptron (MLP)

Convolutional Neural Networks (CNNs):

- Image Classification
- Object Detection
- Image Segmentation

Recurrent Neural Networks (RNNs):

- Sequence-to-Sequence Models
- Text Classification
- Sentiment Analysis

Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU):

- Time Series Forecasting
- Language Modeling

Generative Adversarial Networks (GANs):

- Image Synthesis
- Style Transfer
- Data Augmentation

STEP 6:

ADVANCED TOPICS

Natural Language Processing (NLP):

- Text Preprocessing
- Word Embeddings (e.g., Word2Vec, GloVe)
- Recurrent Neural Networks for NLP
- Transformer Models (e.g., BERT, GPT)

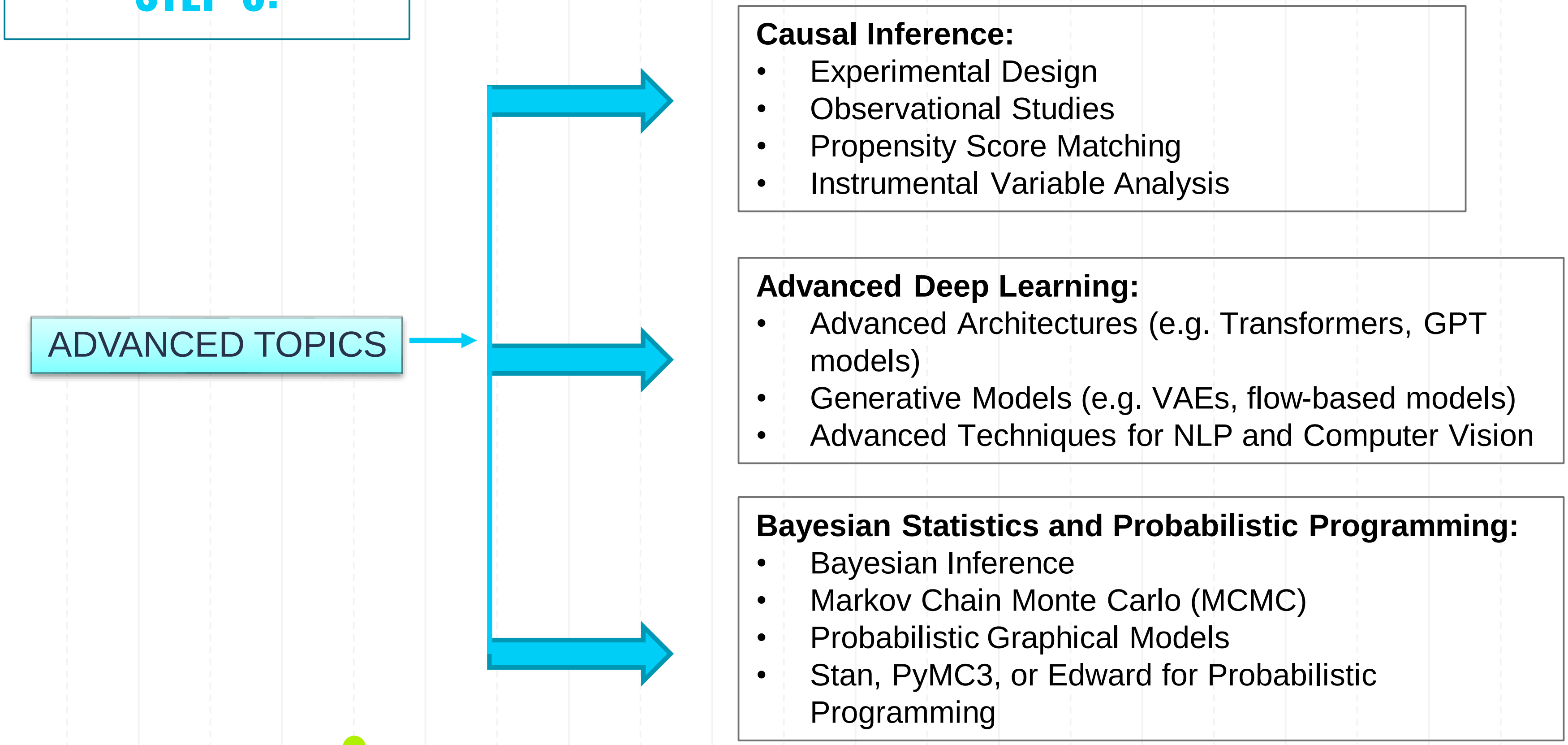
Time Series Analysis:

- Time Series Decomposition
- Autoregressive Integrated Moving Average (ARIMA)
- Seasonal ARIMA (SARIMA)
- Exponential Smoothing Methods
- Prophet

Recommender Systems:

- Collaborative Filtering
- Content-Based Filtering
- Matrix Factorization
- Hybrid Methods

STEP 6:



ADVANCED TOPICS

Causal Inference:

- Experimental Design
- Observational Studies
- Propensity Score Matching
- Instrumental Variable Analysis

Advanced Deep Learning:

- Advanced Architectures (e.g. Transformers, GPT models)
- Generative Models (e.g. VAEs, flow-based models)
- Advanced Techniques for NLP and Computer Vision

Bayesian Statistics and Probabilistic Programming:

- Bayesian Inference
- Markov Chain Monte Carlo (MCMC)
- Probabilistic Graphical Models
- Stan, PyMC3, or Edward for Probabilistic Programming

STEP 7:

BIG DATA TECHNOLOGIES:

- Hadoop:
- HDFS
- MapReduce
- Spark:
- RDDs
- DataFrames
- MLlib
- NoSQL Databases:
- MongoDB
- Cassandra
- HBase
- Couchbase
- Stream Processing Frameworks:
- Apache Kafka
- Apache Flink
- Apache Storm

STEP 8:

DATA VISUALIZATION AND REPORTING:

- Dashboarding Tools:
- Tableau
- Power BI
- Dash (Python)
- Shiny (R)
- Storytelling with Data
- Effective Communication

STEP 9:

DOMAIN KNOWLEDGE AND SOFT SKILLS:

- Industry-specific Knowledge
- Problem-solving
- Communication Skills
- Time Management
- Teamwork

STEP 10:

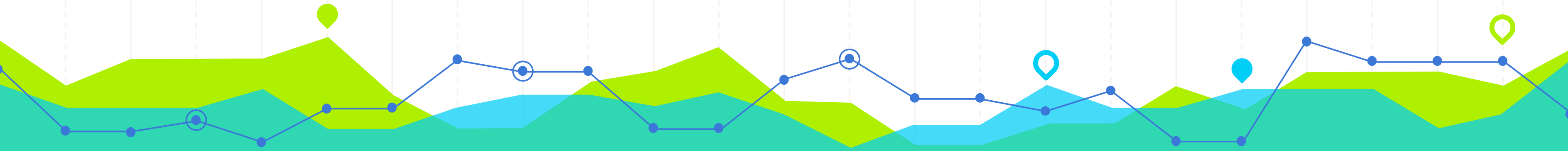
ETHICAL CONSIDERATION AND BIAS:

- Fairness in Machine Learning
- Bias Detection and Mitigation
- Privacy and Data Security

STEP 11:

DEPLOYMENT AND PRODUCTIONIZATION:

- Model Deployment Techniques
- Containerization (e.g., Docker)
- Model Serving and APIs
- Scalability and Performance Optimization

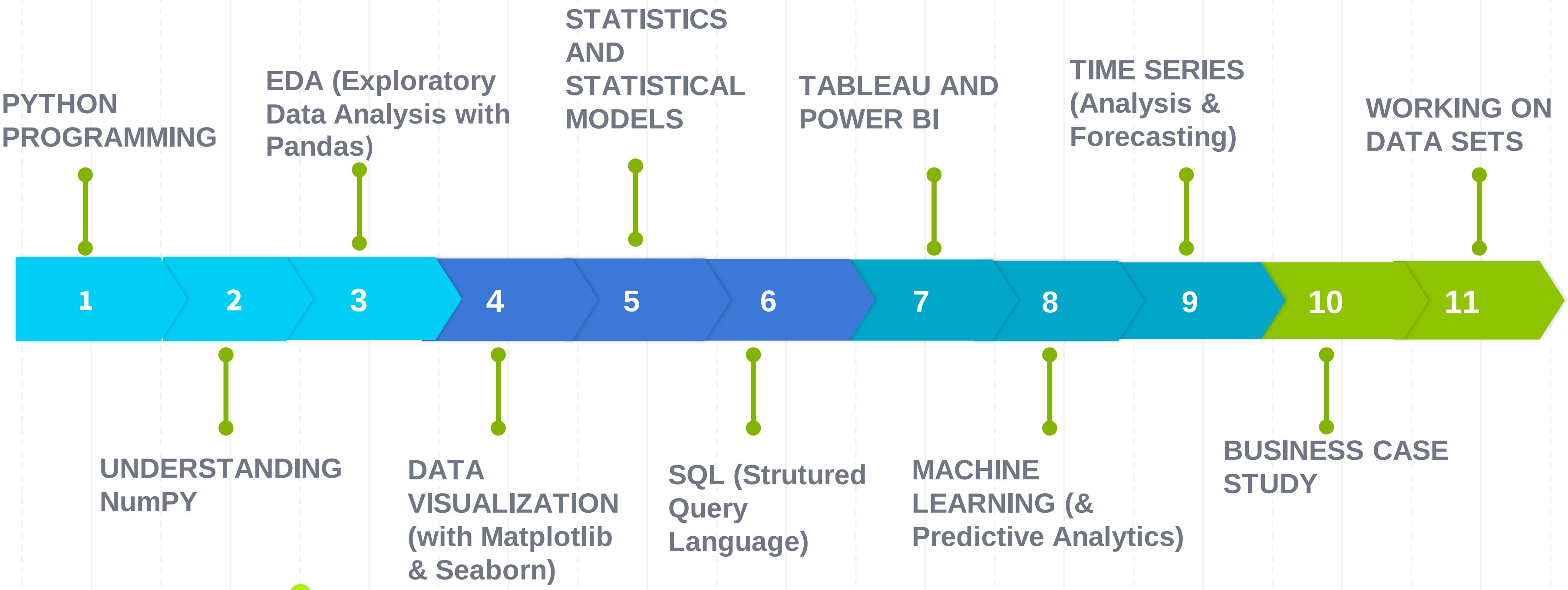


DATA ANALYST ROADMAP

PROFILE AND ROADMAP

4

PROFILE



1 PYTHON PROGRAMMING

2 UNDERSTANDING NumPY

3 EDA (Exploratory Data Analysis with Pandas)

4 DATA VISUALIZATION (with Matplotlib & Seaborn)

5 STATISTICS AND STATISTICAL MODELS

6 SQL (Strutured Query Language)

7 TABLEAU AND POWER BI

8 MACHINE LEARNING (& Predictive Analytics)

9 TIME SERIES (Analysis & Forecasting)

10 BUSINESS CASE STUDY

11 WORKING ON DATA SETS

WHAT IS PYTHON PROGRAMMING?

- Python is a popular high-level programming language that is known for its simplicity, readability, and ease of use. It was first released in 1991 by Guido van Rossum and has since become one of the most widely used programming languages in the world.
- Python is an interpreted language, which means that code written in Python is translated and executed on-the-fly, rather than being compiled beforehand like some other programming languages. This makes Python code easy to write, test, and debug, as changes can be made and tested quickly without the need to recompile the entire program.
- Python is used in a wide range of applications, including web development, scientific computing, data analysis, artificial intelligence, machine learning, and more. It has a large and active community of developers who contribute to its development and offer support through forums, tutorials, and other resources.

Key Elements of PYTHON PROGRAMMING

- While Loops, Lists, Strings
- For Loop, Dictionary, Tuples, Set
- Functions
- Modules, Packages, and PIP
- Virtual Environment, Flask, and Python Web Scrapping

Key Elements of PYTHON PROGRAMMING

- **While Loops, Lists, Strings** :

- **A while loop** is a control structure in Python that allows you to execute a block of code repeatedly as long as a certain condition is true.

- **A list** is a collection of values in Python that can be of any data type. Lists are one of the most commonly used data structures in Python, and are represented by square brackets. Each value in a list is called an element, and elements are separated by commas. You can access elements in a list using their index, which starts at 0 for the first element.

- **Strings** are used to represent text data, and can be manipulated using various string methods. String methods are called using dot notation, where the method is called on the string object and the result is returned as a new string.

Key Elements of PYTHON PROGRAMMING

- A **loop** is a control structure in Python that allows you to execute a block of code repeatedly. There are two types of loops in Python: **for** loops and **while** loops. A **for** loop is used to iterate over a sequence of values, such as a list or a range of numbers, while a **while** loop is used to repeat a block of code as long as a certain condition is true.

- A **dictionary** is a collection of key-value pairs in Python. Dictionaries are used to store and retrieve data based on keys, rather than indexes. Keys are unique and can be of any immutable data type, such as strings or numbers, while values can be of any data type.

- A **tuple** is a collection of values in Python that is similar to a list, but is immutable. Tuples are represented by parentheses **()** and elements are separated by commas. Tuples are often used to represent a group of related values, such as the coordinates of a point or the RGB values of a color.

Key Elements of PYTHON PROGRAMMING

- **Functions**:

- are a fundamental concept in programming and play an important role in Python. A function is a block of reusable code that performs a specific task.

- Functions are designed to be modular, meaning that they can be called multiple times with different input values and produce the same output each time.

- Functions are defined using the **def** keyword, followed by the function name, and then parentheses **()**, which may include any parameters that the function expects to receive.

- The function block is indented and contains the code that will be executed each time the function is called. The **return** keyword is used to specify the value that the function will return when it is called.

Key Elements of PYTHON PROGRAMMING

◉ **Modules**:

➢ In Python, a module is a file containing Python code that can be reused in other Python programs.

➢ A module can contain functions, variables, classes, and other definitions, and can be imported into other modules or programs using the **import** statement.

➢ Python provides a large number of standard modules that are included with the language, such as **math**, **random**, **datetime**, and **os**, among others.

➢ These modules contain pre-written functions and other code that can be used to perform common tasks without having to write the code from scratch.

➢ To use a module in your code, you first need to import it using the **import** statement. Once you have imported a module, you can access its contents using dot notation.

Key Elements of PYTHON PROGRAMMING

- Packages:

➢ A package is a hierarchical module organization structure that contains modules, subpackages, and other packages. A package is simply a directory that contains a special file called **__init__.py**, which can be empty or can contain initialization code for the package.

➢ Packages provide a way to organize related modules and other code into a single namespace, and can help avoid naming conflicts between modules from different sources.

➢ Packages can also be distributed and installed using Python's package management system, **pip**, making it easy to share code with others.

➢ Packages can be nested to create a hierarchy of related functionality.

Key Elements of PYTHON PROGRAMMING

◉ PIP:

- PIP (short for "Pip Installs Packages") is a package manager for Python that allows you to easily install and manage third-party Python packages and libraries. PIP comes installed by default with Python versions 3.4 and above.

- Using PIP, you can install packages from the Python Package Index (PyPI), which is a repository of software packages for Python that are maintained and shared by the Python community.

- PyPI contains thousands of packages that provide additional functionality for Python, such as scientific computing libraries, web frameworks, database drivers, and more.

- To install a package using PIP, you can use the **pip install** command followed by the name of the package you want to install.

Key Elements of PYTHON PROGRAMMING

- Flask is a popular micro web framework for Python that allows you to quickly build web applications with Python. It is designed to be lightweight, modular, and easy to use, and is particularly well-suited for building small to medium-sized web applications.

- Flask provides a range of features for building web applications, including routing, templating, request handling, and more. It also supports extensions that allow you to add additional functionality to your application, such as database support, authentication, and more.

- One of the key features of Flask is its simplicity and ease of use.

- Flask has a small footprint and is easy to learn, making it a popular choice for beginners and experienced developers alike. Flask also provides a flexible and extensible architecture that allows you to customize the framework to suit your needs.

Key Elements of PYTHON PROGRAMMING

- PYTHON WEB SCRAPPING:

- Web scraping is the process of automatically extracting data from websites using software tools or scripts. It involves sending requests to a website and then parsing the HTML content of the response to extract the relevant data.

- Web scraping can be done manually by inspecting the source code of a website and copying the data, but it is much more efficient and scalable to use automated tools.

- Web scraping is used for a variety of purposes, including data mining, price monitoring, market research, content aggregation, and more. Web scraping can be used to extract data such as product information, stock prices, news articles, social media posts, and more.

- However, web scraping can also be a sensitive issue because it can potentially violate website terms of service or copyright laws. It is important to be mindful of ethical considerations when web scraping, and to ensure that you are not infringing on the rights of others or engaging in any illegal activities.

Key Elements of NumPY

- NumPy basics
- Working with Matrix
- Linear Algebra operations
- Descriptive Statistics
- Normal Distribution Operations
- Mean, Variance, and Standard Deviation
- Reshaping arrays

What is NumPY?

- NumPy basics

- NumPy is a popular Python library for numerical computing that provides powerful tools for working with large, multi-dimensional arrays and matrices of numeric data. It is designed to be fast, efficient, and easy to use, and is widely used in scientific computing, data analysis, and machine learning.

- NumPy provides a range of features for working with arrays and matrices, including functions for performing mathematical operations, linear algebra, Fourier transforms, and more.

- It also provides tools for indexing, slicing, and manipulating arrays and matrices, as well as functions for generating random data and working with data stored in different file formats.

- One of the key advantages of NumPy is its performance. NumPy is built on top of highly optimized C and Fortran code, and provides efficient implementations of common numerical operations that can be orders of magnitude faster than equivalent Python code.

What is Matrix?

- In mathematics, a matrix is a rectangular array of numbers or other mathematical objects, arranged in rows and columns.

- Matrices are often used to represent linear transformations or systems of linear equations, and they have applications in a wide range of fields, including physics, engineering, computer graphics, and more.

- A matrix can be denoted by enclosing its entries in brackets or parentheses, with rows separated by commas or semicolons

What is Linear Algebra?

- Linear algebra is a branch of mathematics that deals with the study of linear equations, vectors, matrices, and linear transformations.
- It is a fundamental tool in many areas of science and engineering, including physics, engineering, computer science, and data analysis.
- Linear algebra provides a framework for solving systems of linear equations, which arise in many real-world problems, such as optimizing complex systems, analyzing data, and modeling physical systems.
- Linear algebra is also used to analyze geometric shapes and transformations, such as rotations, translations, and reflections.
- Some of the key concepts in linear algebra include vectors, matrices, determinants, and eigenvectors.

What is Descriptive Statistics?

- Descriptive statistics is a branch of statistics that deals with the analysis and summarization of data. It involves methods for describing, summarizing, and visualizing data sets in order to gain insight into their characteristics and properties.

- Descriptive statistics can be used to summarize data in a variety of ways, such as measures of central tendency (e.g., mean, median, mode), measures of variability (e.g., range, standard deviation, variance), and measures of shape (e.g., skewness, kurtosis).

- These measures can help to provide a concise summary of the data and to identify patterns and trends.
- Descriptive statistics can also be used to create visual representations of data, such as histograms, box plots, and scatterplots.

- These visualizations can provide a quick and intuitive way to understand the distribution of the data and to identify outliers or unusual patterns.

What is Normal Distribution?

- Normal distribution, also known as Gaussian distribution, is a probability distribution that is commonly used in statistics and probability theory.

- It is a bell-shaped distribution that is symmetric around the mean and is characterized by two parameters: the mean (μ) and the standard deviation (σ).

- The normal distribution is a continuous probability distribution that describes the distribution of a random variable that is expected to be influenced by many small and independent factors.

- The normal distribution has several important properties. First, it is unimodal, meaning that it has a single peak at the mean. Second, it is symmetric around the mean, meaning that the probabilities of values above the mean are equal to the probabilities of values below the mean. Third, the standard deviation determines the spread of the distribution, with higher values of σ corresponding to wider distributions.

What is Mean & Standard Deviation?

- The mean, also known as the average, is a measure of the central tendency of a set of data. It is calculated by adding up all the values in the data set and dividing the sum by the total number of values.

- The mean is sensitive to outliers, meaning that extreme values in the data set can have a significant impact on the calculated mean.

- The standard deviation is a measure of the dispersion or spread of a set of data. It is calculated by taking the square root of the variance, which is the average of the squared differences between each value and the mean.

- The standard deviation gives an idea of how much the data deviates from the mean, with higher values indicating a greater degree of variability in the data.

- Together, the mean and standard deviation can provide a useful summary of the central tendency and spread of a data set.

What is Reshaping Arrays?

- In NumPy, reshaping arrays means changing the shape or dimensions of an existing array, without changing the underlying data.

- The process of reshaping involves creating a new array with the same elements as the original array, but with a different shape.

- The reshape() function in NumPy can be used to reshape an array. The function takes a tuple specifying the new shape of the array as its argument.

- The new shape must be compatible with the number of elements in the original array. If the new shape does not have the same number of elements as the original array, a ValueError is raised.

WHAT IS EDA ?

- EDA stands for Exploratory Data Analysis. It is an approach to analyzing and understanding data that involves visually exploring, summarizing, and interpreting the patterns and relationships in the data.
- EDA is typically performed before any formal modeling or statistical testing is done, and it is used to gain insights into the data and develop hypotheses about the relationships between variables.
- EDA involves a range of techniques for summarizing and visualizing data, including histograms, scatterplots, boxplots, and summary statistics such as the mean, median, and standard deviation.
- EDA can also involve identifying and dealing with missing data, outliers, and other data quality issues that may affect the validity of any subsequent analysis.

Key Elements of EDA

- Pandas
- Data Analysis basics
- Dataframe operations
- Working with 2-dimensional data
- Data Cleaning
- Data Grouping
- Working with Datasets

WHAT IS DATA CLEANING AND GROUPING ?

- Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a data set. The goal of data cleaning is to ensure that the data is accurate, complete, and consistent so that it can be used for analysis or other purposes.
- Data grouping is the process of grouping data together based on some common characteristic or feature.
- This can be useful for organizing and summarizing data, as well as for identifying patterns and relationships between variables.
- In many cases, data cleaning and grouping go hand in hand.
- For example, when cleaning data, it may be necessary to group data together based on some common characteristic in order to identify errors or inconsistencies.

COMMON TECHNIQUES IN DATA CLEANING AND GROUPING

- Some common techniques used in data cleaning include:
- Removing duplicate records or observations
- Handling missing or null values
- Correcting errors in data entry or formatting
- Standardizing data formats and units of measurement
- Some common techniques used in data grouping include:
- Creating categories or bins based on ranges of values
- Grouping data based on a categorical variable, such as age group or location
- Aggregating data by summarizing or averaging values within each group
- Creating subgroups based on multiple variables or criteria

What is Data Visualization with Matplotlib and Seaborn?

- Matplotlib and Seaborn are two popular Python libraries for data visualization.

- Matplotlib is a general-purpose plotting library that can be used to create a wide range of visualizations, from simple line and scatter plots to complex heatmaps and 3D plots.

- Seaborn is a more specialized library that is focused on statistical data visualization, and provides a higher-level interface for creating more complex visualizations with less code.

- Some common types of visualizations that can be created with Matplotlib and Seaborn include:
- Line plots and scatter plots: These are basic plots that can be used to visualize the relationship between two variables.

- Bar plots and histograms: These are used to visualize the distribution of a single variable.
- Box plots and violin plots: These are used to compare the distribution of a variable across different categories or groups.
- Heatmaps: These are used to visualize the relationship between two variables across a grid of values.
- 3D plots: These are used to visualize data in three dimensions.

Key Elements of Matplotlib and Seaborn

- Matplotlib
- Plot Basics
- Format Strings
- Label and Legends
- Bar Chart
- Pie Chart

WHAT IS FORMAT STRINGS?

- Format strings are a feature of Python that allow you to create formatted strings that contain values from variables or expressions.

- They are often used to create output that is more human-readable and informative than simply printing the raw values of variables.

- In a format string, you specify the positions of values that will be replaced with expressions or variables by using placeholder brackets **{}**.

- You can also include formatting options within the placeholders to specify how the values should be displayed, such as the number of decimal places for floating-point values or the width of a string.

- Format strings provide a convenient and flexible way to create formatted output in Python.They can be used in a wide range of contexts, from simple print statements to more complex logging and debugging output.

What is Descriptive Statistics?

- Descriptive statistics is a branch of statistics that deals with the collection, analysis, interpretation, and presentation of data.

- It is concerned with summarizing and describing the main features of a dataset, such as the central tendency, variability, and distribution of the data.

- It can be used to gain insights into a dataset, identify patterns and trends, and communicate findings to others. Some common measures of descriptive statistics include:
- Measures of central tendency, such as the mean, median, and mode, which provide an indication of where the center of the data lies.
- Measures of variability, such as the range, variance, and standard deviation, which provide an indication of how spread out the data is.
- Measures of shape, such as skewness and kurtosis, which provide information about the symmetry and peakedness of the distribution of the data.
- Graphical displays, such as histograms, box plots, and scatter plots, which can be used to visualize the distribution and relationships between variables in the data.

Key Elements of Descriptive Statistics

- Measure of Frequency and Central Tendency
- Measure of Dispersion
- Probability Distribution
- Gaussian Normal Distribution
- Skewness and Kurtosis
- Regression Analysis
- Continuous and Discrete Functions
- Goodness of Fit
- ANOVA

What is Inferential Statistics?

- Inferential statistics is a branch of statistics that deals with making inferences or predictions about a population based on a sample of data.

- It involves using statistical techniques to analyze a sample of data and then using the results to make generalizations or predictions about the larger population from which the sample was drawn.

- The main goal of inferential statistics is to draw conclusions about a population based on a sample, while taking into account the uncertainty and variability in the data. Some common techniques used in inferential statistics include hypothesis testing, confidence intervals, and regression analysis.

- Hypothesis testing involves testing a hypothesis about a population based on a sample of data.

- It involves formulating a null hypothesis (which assumes that there is no significant difference between the sample and population) and an alternative hypothesis (which assumes that there is a significant difference between the sample and population).

- By analyzing the data, a statistical test can be used to determine whether to reject or fail to reject the null hypothesis.

Key Elements of Inferential Statistics

- t-Test
- z-Test
- Hypothesis Testing
- Type I and Type II errors
- t-Test and its types
- One way ANOVA
- Two way ANOVA
- Chi-Square Test
- Implementation of continuous and categorical data

Descriptive VS Inferential Statistics

The main differences between them are:

- Purpose: Descriptive statistics is used to describe and summarize data, while inferential statistics is used to make inferences and draw conclusions about a population based on a sample of data.
- Population vs Sample: Descriptive statistics is concerned with summarizing the characteristics of a sample of data, while inferential statistics is concerned with making predictions and drawing conclusions about a population based on a sample of data.
- Generalization: Descriptive statistics does not allow for generalization beyond the data being analyzed, while inferential statistics allows for generalization to the broader population.

Descriptive VS Inferential Statistics

- Variables: Descriptive statistics deals with only one variable at a time, while inferential statistics deals with multiple variables and their relationships.

- Methodology: Descriptive statistics uses measures such as mean, median, mode, standard deviation, etc., to describe the data, while inferential statistics uses methods such as hypothesis testing, regression analysis, and ANOVA to make inferences and draw conclusions.

In summary, descriptive statistics is used to describe and summarize data, while inferential statistics is used to make inferences and draw conclusions about a population based on a sample of data. Descriptive statistics deals with one variable at a time, while inferential statistics deals with multiple variables and their relationships.

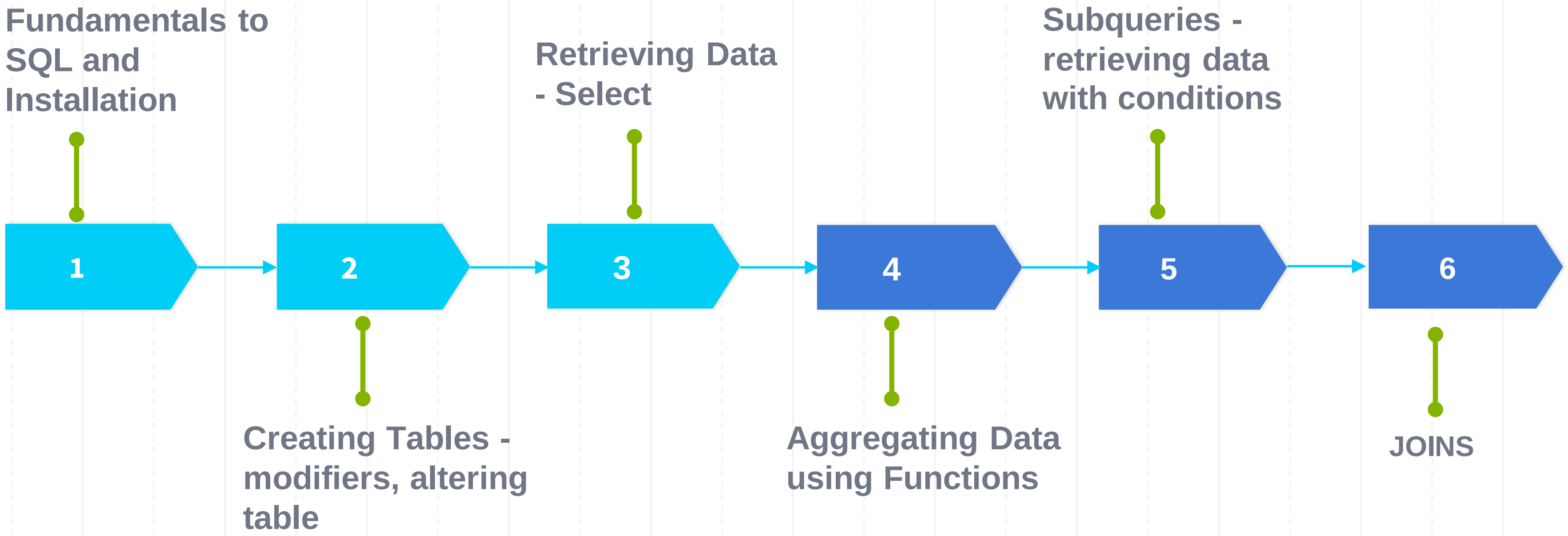
What is SQL?

- SQL (Structured Query Language) is a programming language used to manage and manipulate relational databases.
- It is the standard language for working with relational databases, which are used to store and organize large amounts of structured data.
- SQL allows users to perform various operations on databases, such as creating tables and views, inserting, updating, and deleting data, and retrieving data through queries.
- It provides a set of commands and syntax for interacting with databases and can be used with various relational database management systems (RDBMS) such as MySQL, Oracle, Microsoft SQL Server, PostgreSQL, and SQLite.
- SQL has several components, including data definition language (DDL), data manipulation language (DML), data control language (DCL), and transaction control language (TCL).

What is SQL?

- DDL is used to create and modify the structure of the database, including tables, indexes, and constraints.
- DML is used to insert, update, and delete data in the database. DCL is used to manage the security and permissions for the database objects.
- TCL is used to manage transactions and ensure data consistency.
- SQL is widely used in various fields such as business, finance, healthcare, and government.
- It is an essential tool for data analysts, data scientists, and software developers who need to manage and manipulate large amounts of data stored in databases.

ROADMAP OF SQL



1 Fundamentals to SQL and Installation

2 Creating Tables - modifiers, altering table

3 Retrieving Data - Select

4 Aggregating Data using Functions

5 Subqueries - retrieving data with conditions

6 JOINS

APPLICATION OF SQL

SQL (Structured Query Language) is a widely used language for working with relational databases. Here are some of the main uses of SQL:

1. **Data Management**: SQL is used to create, modify, and delete databases, tables, views, and other database objects. It is also used to insert, update, and delete data in the database.
2. **Data Retrieval**: SQL is used to retrieve data from a database by executing queries on the database. These queries can be simple or complex, and they can involve one or more tables.
3. **Data Analysis**: SQL is used to perform data analysis by aggregating, grouping, and sorting data in various ways. It is also used to join tables together to create more complex queries.
4. **Data Reporting**: SQL is used to generate reports from databases. Reports can be generated in various formats, such as PDF, Excel, HTML, and others.
5. **Business Intelligence**: SQL is used in business intelligence applications to create dashboards, reports, and visualizations that help businesses make informed decisions based on their data.

.

APPLICATION OF SQL

- **E-commerce**: SQL is used in e-commerce applications to manage product catalogs, customer data, and transactions.
- **Web Development**: SQL is used in web development to store and retrieve data from databases. It is often used in conjunction with programming languages such as PHP, Python, and Java.

Overall, SQL is an essential tool for working with relational databases and is used in a wide range of industries and applications.

WHAT IS MACHINE LEARNING?

- Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to learn from data and make predictions or decisions without being explicitly programmed.
- The goal of machine learning is to enable computers to learn from data and improve their performance on a task over time.
- There are several different types of machine learning, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.
- In supervised learning, the computer is given a set of labeled training examples and learns to make predictions based on that data.
- In unsupervised learning, the computer is given unlabeled data and must find patterns or structure in that data.
- Semi-supervised learning is a combination of supervised and unsupervised learning, where the computer is given some labeled and some unlabeled data.

TYPES OF MACHINE LEARNING

There are three main types of machine learning: **supervised learning, unsupervised learning, and reinforcement learning.**

1.Supervised Learning:

- the algorithm is trained on a labeled dataset, which means that the input data is labeled with the correct output.
- The algorithm learns to map the input to the output by minimizing the difference between its predicted output and the true output.
- Supervised learning is used in applications such as image classification, speech recognition, and natural language processing.

2.Unsupervised Learning:

- the algorithm is trained on an unlabeled dataset, which means that the input data is not labeled with the correct output.
- The algorithm learns to find patterns or structure in the input data by grouping similar data points together or identifying outliers.
- Unsupervised learning is used in applications such as clustering, anomaly detection, and dimensionality reduction.

TYPES OF MACHINE LEARNING

3. Reinforcement Learning:

- an agent learns to make decisions based on feedback from its environment.
- The agent interacts with the environment by taking actions, and it receives rewards or penalties based on the outcome of those actions.
- The goal of the agent is to learn a policy that maximizes its cumulative reward over time.
- Reinforcement learning is used in applications such as game playing, robotics, and autonomous driving.

There are also other subtypes and variations of these three main types of machine learning, such as semi-supervised learning, transfer learning, and deep learning. Each type has its own strengths and weaknesses, and the choice of which type to use depends on the specific problem and available data.

APPLICATION OF MACHINE LEARNING

1. Healthcare: Machine learning is used to analyze medical data and images to aid in diagnosis, treatment, and drug development.
2. Finance: Machine learning is used for fraud detection, credit scoring, and stock market analysis.
3. E-commerce: Machine learning is used for product recommendation, customer segmentation, and dynamic pricing.
4. Marketing: Machine learning is used for customer profiling, lead scoring, and campaign optimization.
5. Manufacturing: Machine learning is used for predictive maintenance, quality control, and supply chain optimization.
6. Transportation: Machine learning is used for traffic prediction, route optimization, and autonomous vehicles.
7. Natural Language Processing: Machine learning is used for speech recognition, sentiment analysis, and chatbots.
8. Computer Vision: Machine learning is used for image and video analysis, object recognition, and facial recognition.

SQL VS MACHINE LEARNING

1. SQL is a programming language used for managing and manipulating data stored in relational databases. It is used for tasks such as creating and modifying tables, querying data, and performing database administration tasks. SQL is primarily used for data storage, retrieval, and management.

2. Machine learning, on the other hand, is a branch of artificial intelligence that focuses on developing algorithms and models that can learn from data and make predictions or decisions. Machine learning is used for tasks such as image and speech recognition, natural language processing, fraud detection, and recommendation systems. Machine learning is primarily used for data analysis, prediction, and decision-making.

While SQL and machine learning are used for different purposes, they are often used together in data-driven applications. SQL can be used to extract and prepare data for use in machine learning algorithms, while machine learning can be used to analyze and derive insights from data that has been collected and stored using SQL.

QUICK REFERENCE MATERIAL 5

Data Engineering Quick Reference

- **DATABASE:**

- Relational Database : A database that stores data in tables with a defined schema
- NoSQL Database : A database that does not use the traditional relational database model
- SQL : A language used to interact with relational databases
- MongoDB : A popular NoSQL database that stores data in JSON-like documents
- Cassandra : A popular NoSQL database that is designed for high scalability and availability
- Redis : An in-memory key-value store used for caching and other high-performance use cases
- Amazon RDS : A managed relational database service provided by AWS

Data Engineering Quick Reference

DATA WAREHOUSING:

- Data Warehouse : A large, centralized repository of data from various sources used for business intelligence and decision-making
- OLAP : Online Analytical Processing, used for analyzing data from a data warehouse
- Star Schema : A type of data model used in data warehousing that consists of a central fact table surrounded by dimension tables.
- Snowflake Schema : A variation of the star schema that uses normalized dimension tables
- Slowly Changing Dimensions (SCD) : A technique used for managing changes to dimensional data over time
- ETL : Extract, Transform, Load, the process of moving data from source systems into a data warehouse
- Amazon Redshift : A cloud-based data warehousing service provided by AWS

Data Engineering Quick Reference

- **Big Data Technologies:**
  - Hadoop : An open-source framework for distributed storage and processing of large data sets
  - Spark : An open-source distributed computing system used for big data processing and analytics
  - Hive : A data warehousing system built on top of Hadoop for querying and analysis of large data sets
  - Pig : A high-level platform for creating MapReduce programs used for large-scale data processing
  - MapReduce : A programming model for processing large data sets across clusters of computers

Data Engineering Quick Reference

- **Big Data Technologies:**
  - Impala : A distributed SQL query engine for processing big data sets stored in Hadoop
  - Kafka : A distributed streaming platform used for building real-time data pipelines and streaming applications
  - Amazon EMR : A managed big data processing service provided by AWS

Data Engineering Quick Reference

DATA PROCESSING:

- Data Pipeline : A set of processes used to extract, transform, and load data from various sources into a destination system
- ETL Tools : Tools used to automate the extraction, transformation, and loading of data
- Apache Airflow : An open-source platform used for creating, scheduling, and monitoring data pipelines
- AWS Glue : A fully-managed ETL service provided by AWS
- Talend : A popular open-source ETL tool used for data integration and management
- Data Governance : The process of managing the availability, usability, integrity, and security of data

Data Engineering Quick Reference

DATA STREAMING:

- Data Stream : A continuous flow of data that is processed in real-time
- Apache Kafka : A distributed streaming platform used for building real- time data pipelines and streaming applications
- Kinesis : A fully-managed data streaming service provided by AWS
- Flume : A distributed system for collecting, aggregating, and moving large amounts of log data from different sources to a centralized data store
- Spark Streaming : An extension of the Spark API used for processing real- time data streams
- Flink : An open-source distributed stream processing framework used for real-time data processing

Data Engineering Quick Reference

- **Data Visualization**

- Tableau : A popular data visualization tool used for creating interactive dashboards and reports
- Power BI : A business analytics service provided by Microsoft used for creating interactive visualizations and reports
- D3.js : A JavaScript library used for creating interactive data visualizations in the browser
- ggplot2 : A popular data visualization package for R
- matplotlib : A popular data visualization package for Python

Data Engineering Quick Reference

- **Cloud Technologies**
- AWS : Amazon Web Services, a cloud computing platform provided by Amazon
- Azure : A cloud computing platform provided by Microsoft
- GCP : Google Cloud Platform, a cloud computing platform provided by Google
- Docker : A containerization platform used for packaging and deploying applications
- Kubernetes : An open-source container orchestration platform used for automating the deployment, scaling, and management of containerized applications

Data Engineering Quick Reference

- **Data Governance**
  - Data Security : The process of ensuring data privacy and confidentiality
  - Data Quality : The process of ensuring data accuracy, consistency, and completeness
  - Data Lineage : The process of tracking data from its source to its destination
  - Data Discovery : The process of identifying data assets and their relationships
  - Data Stewardship : The process of managing data assets and their usage

Data Engineering Quick Reference

Data Modeling

- Entity-Relationship Model : A data modeling technique used to represent the relationships between entities in a system
- Dimensional Modeling : A data modeling technique used in data warehousing for creating optimized data structures
- Data Flow Diagrams : A diagrammatic representation of the flow of data through a system
- UML : Unified Modeling Language, a standardized language used for object-oriented modeling
- ERD Tools : Tools used for creating entity-relationship diagrams and other data modeling diagrams

Data Engineering Quick Reference

Data Integration

- Data Federation : The process of combining data from multiple sources into a single virtual view
- Data Replication : The process of copying data from one database to another in near-real time
- Data Synchronization : The process of ensuring that data is consistent across multiple systems
- Extract, Load, Transform (ELT) : A data integration approach where data is extracted from source systems, loaded into a staging area, and transformed before being loaded into a target system
- Change Data Capture (CDC) : A data integration technique where changes in source systems are captured and propagated to target systems in near-real time

Data Engineering Quick Reference

- **Data Architecture**
- Data Lake : A storage repository that holds a vast amount of raw, unstructured data in its native format
- Data Mart : A subset of a data warehouse that is designed for a specific business function or department
- Data Hub : A centralized repository of data that serves as a single source of truth for an organization
- Data Virtualization : A data integration technique that allows data to be accessed and manipulated in real-time without copying or moving it
- Master Data Management (MDM) : The process of creating and maintaining a single, trusted view of key business data

Data Engineering Quick Reference

- **Machine Learning**
- Supervised Learning : A type of machine learning where the algorithm is trained on labeled data
- Unsupervised Learning : A type of machine learning where the algorithm is trained on unlabeled data
- Reinforcement Learning : A type of machine learning where the algorithm learns from feedback in an environment
- Deep Learning : A type of machine learning that uses neural networks to model complex relationships in data
- TensorFlow : An open-source machine learning framework developed by Google
- PyTorch : An open-source machine learning framework developed by Facebook
- Scikit-learn : A popular machine learning library for Python

Data Engineering Quick Reference

Data Science

- Statistical Analysis : The process of analyzing data to uncover relationships and patterns
- Data Exploration : The process of identifying patterns and trends in data
- Predictive Modeling : The process of using data to make predictions aboutfuture events
- Time Series Analysis : The process of analyzing data that is collected over time
- Spatial Analysis : The process of analyzing data that is related to geographic locations
- Data Visualization : The process of representing data graphically
- Data Mining : The process of discovering patterns and relationships in large datasets

Data Engineering Quick Reference

Programming Languages

- Python : A popular programming language used for data engineering and machine learning
- Java : A popular programming language used for building enterprise-level applications and big data technologies
- Scala : A programming language used for building big data technologies and data streaming applications
- SQL : A language used for interacting with relational databases
- R : A programming language used for statistical computing and data analysis

Data Engineering Quick Reference

- **Cloud Computing Services**
  - EC2 : Elastic Compute Cloud, a virtual server provided by AWS
  - S3 : Simple Storage Service, a scalable object storage service provided by AWS
  - Lambda : A serverless compute service provided by AWS
  - CloudFormation : A service provided by AWS for modeling and setting up cloud resources
  - Azure VM : A virtual machine provided by Azure
  - Azure Blob Storage : A scalable object storage service provided by Azure
  - Azure Functions : A serverless compute service provided by Azure
  - Azure Resource Manager : A service provided by Azure for modeling and setting up cloud resources

Data Engineering Quick Reference

- **Cloud Computing Services**
- GCE : Google Compute Engine, a virtual machine provided by GCP
- Cloud Storage : A scalable object storage service provided by GCP
- Cloud Functions : A serverless compute service provided by GCP
- Cloud Deployment Manager : A service provided by GCP for modeling

- USEFUL TECHNOLOGIES:
- Apache Airflow : A platform used for creating, scheduling, and monitoring data pipeline.
- Apache Kafka : A distributed streaming platform used for building real-time data pipelines and streaming applications
- Spark : An open-source distributed computing system used for big data processing and analytics

Important Topics : Q&A

6

Data Engineering Quick Reference

◉ **Cloud Computing Services**

- GCE : Google Compute Engine, a virtual machine provided by GCP
- Cloud Storage : A scalable object storage service provided by GCP
- Cloud Functions : A serverless compute service provided by GCP
- Cloud Deployment Manager : A service provided by GCP for modeling

- USEFUL TECHNOLOGIES:
- Apache Airflow : A platform used for creating, scheduling, and monitoring data pipeline.
- Apache Kafka : A distributed streaming platform used for building real-time data pipelines and streaming applications
- Spark : An open-source distributed computing system used for big data processing and analytics