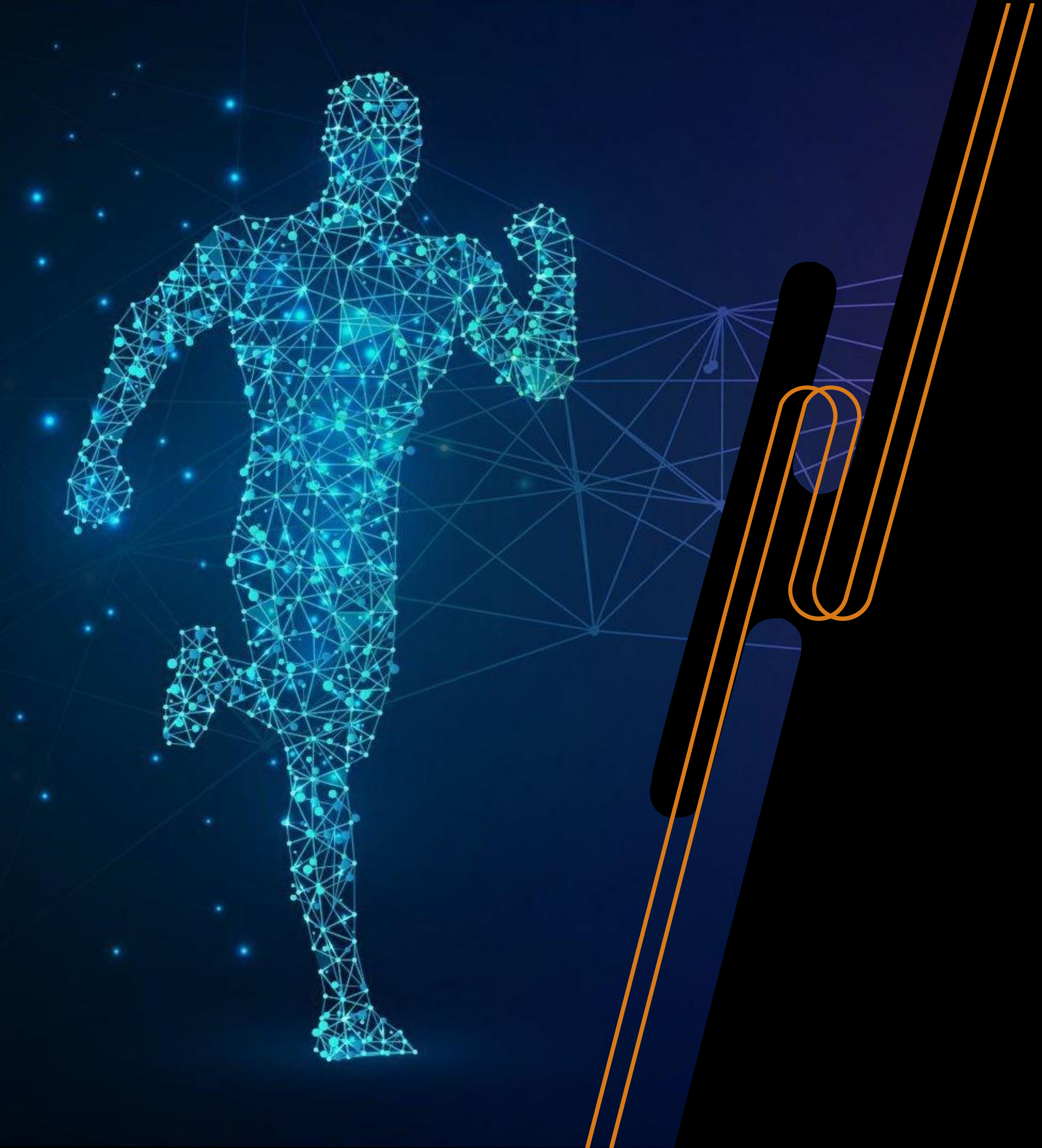
Introduction to Data Science

BY

VIJAYA NANDINI



Introduction

Topics covered

- What is AI, DS, ML, DL and GenAI?
- Scope of Data Science
- Data Science Workflow and Lifecycle
- Applications and use cases
- Tools and Frameworks

What is Artificial Intelligence?

Artificial Intelligence or AI refers to ability of a computer systems to perform tasks that typically require human intelligence.

- The tasks include visual perception, speech recognition, decision-making, and language translation.
- AI systems are designed to adapt and improve their performance over time based on their experiences and interactions with the environment.

What is Data Science?

Data science or DS is a multidisciplinary field that involves using various techniques, algorithms, and systems to extract knowledge and insights from data in various forms.

- Data scientists analyze, interpret, and visualize data to uncover patterns, trends, and insights that can be used to make informed decisions.

What is Machine Learning?

Machine Learning or ML is the subset of AI that enable systems to learn from data and improve their performance over time without being explicitly programmed to do so.

- Machine learning algorithms are designed to identify patterns and relationships in data, allowing them to make predictions or decisions without being explicitly programmed for every possible scenario.
- Common applications of machine learning include image and speech recognition, recommendation systems, predictive analytics, and natural language processing.

What is Deep Learning?

Deep learning or DL is a subset of ML that uses artificial neural networks to model and solve complex problems. Deep learning algorithms are inspired by the structure and function of the human brain.

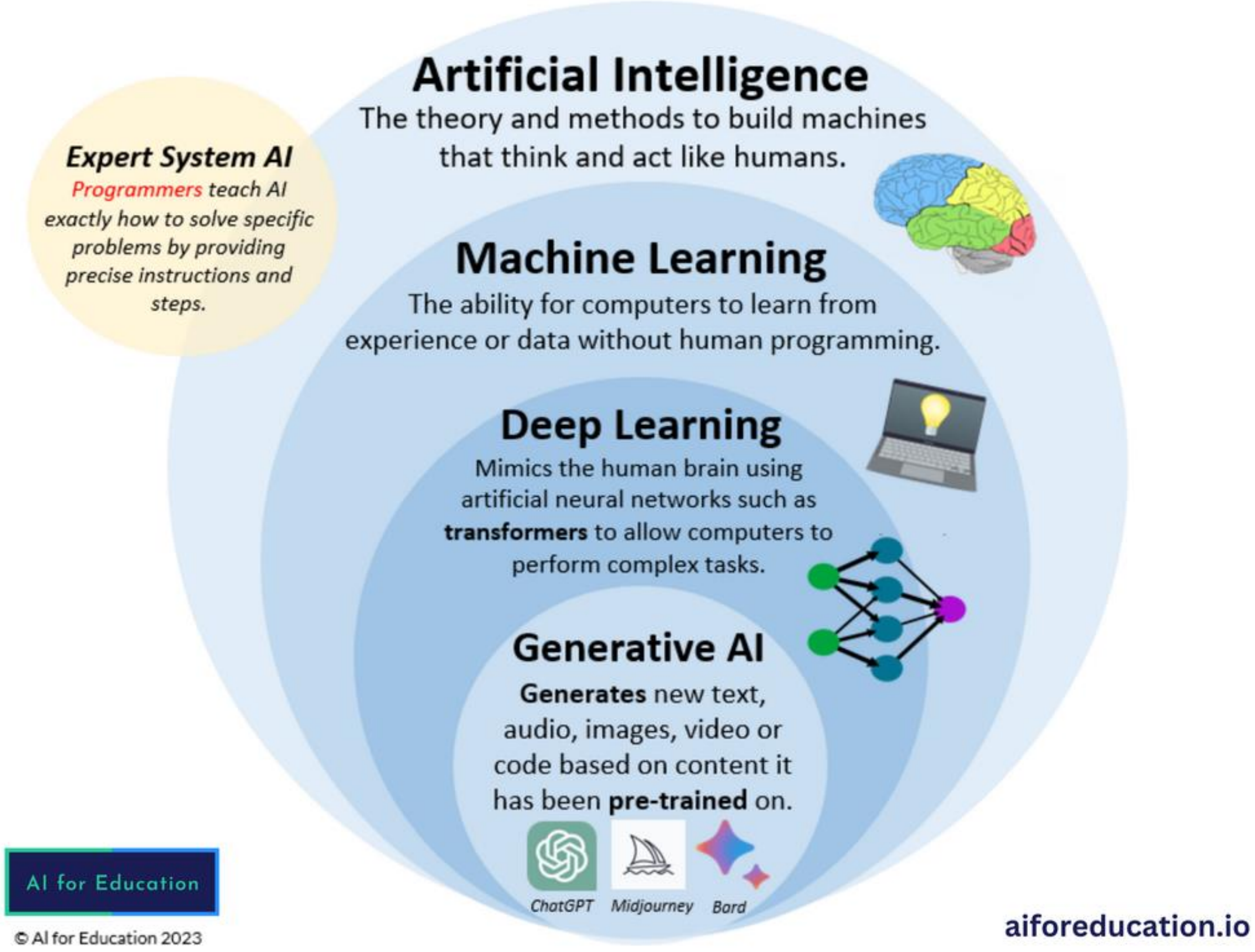
- Deep learning has gained popularity in recent years due to its ability to automatically learn representations from data, significantly outperforming traditional machine learning methods in various applications.
- Training process of DL models is computationally intensive and typically requires large amounts of data, but the results can be highly accurate and robust across different domains.

What is Generative AI?

Generative AI or GenAI refers to a type of artificial intelligence that can create new content, such as text, images, or music, by learning patterns and structures from existing data.

- It is often used in fields like creative arts, content creation, and even language translation.
- Generative AI models, such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), are designed to generate original and novel content based on the patterns they have learned from large datasets during their training process.

Relation between AI, ML, DL & GenAI



Expert System AI

Programmers teach AI exactly how to solve specific problems by providing precise instructions and steps.

Artificial Intelligence

The theory and methods to build machines that think and act like humans.

Machine Learning

The ability for computers to learn from experience or data without human programming.

Deep Learning

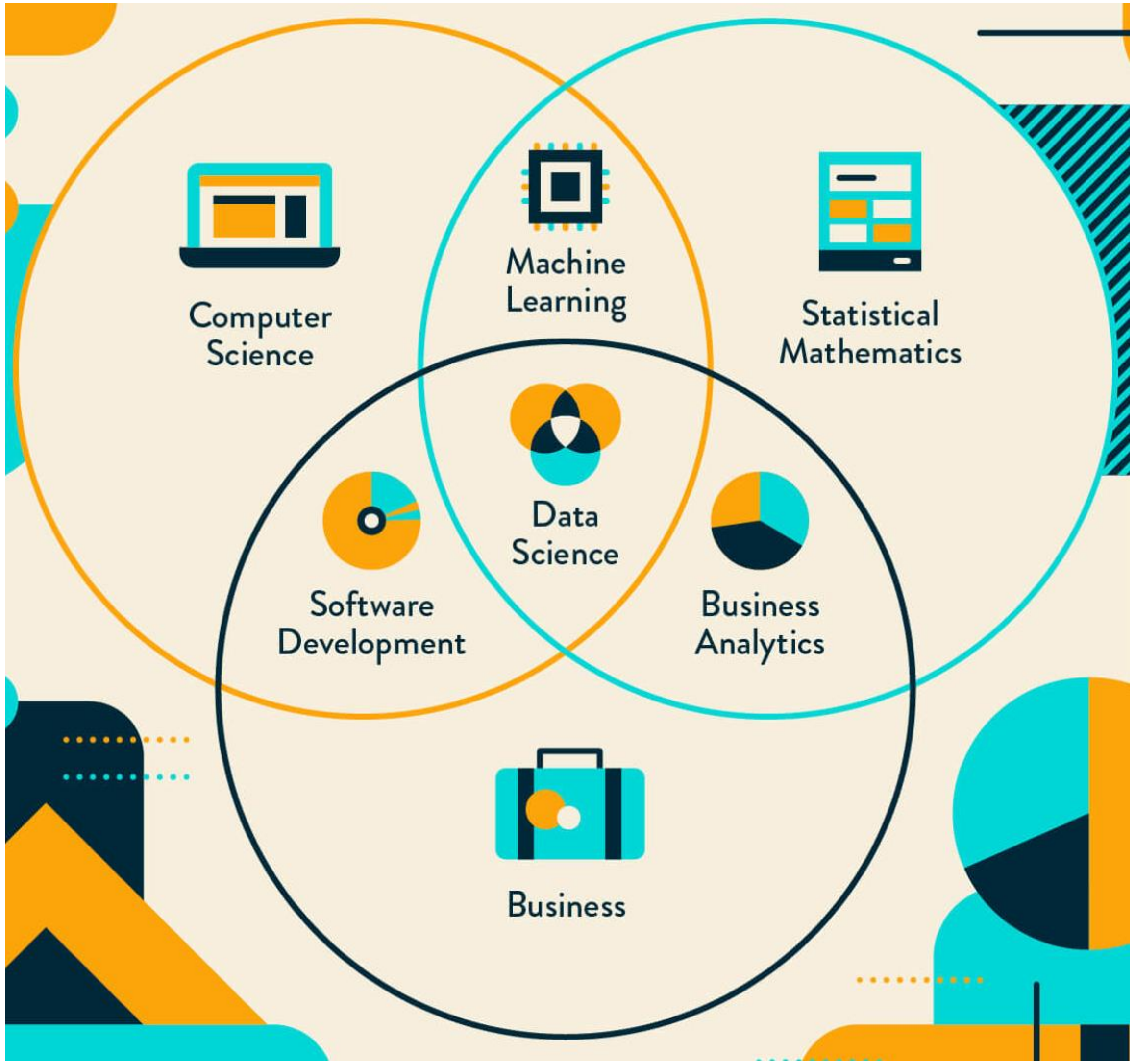
Mimics the human brain using artificial neural networks such as **transformers** to allow computers to perform complex tasks.

Generative AI

Generates new text, audio, images, video or code based on content it has been **pre-trained** on.

aiforeducation.io

Scope of Data Science and ML



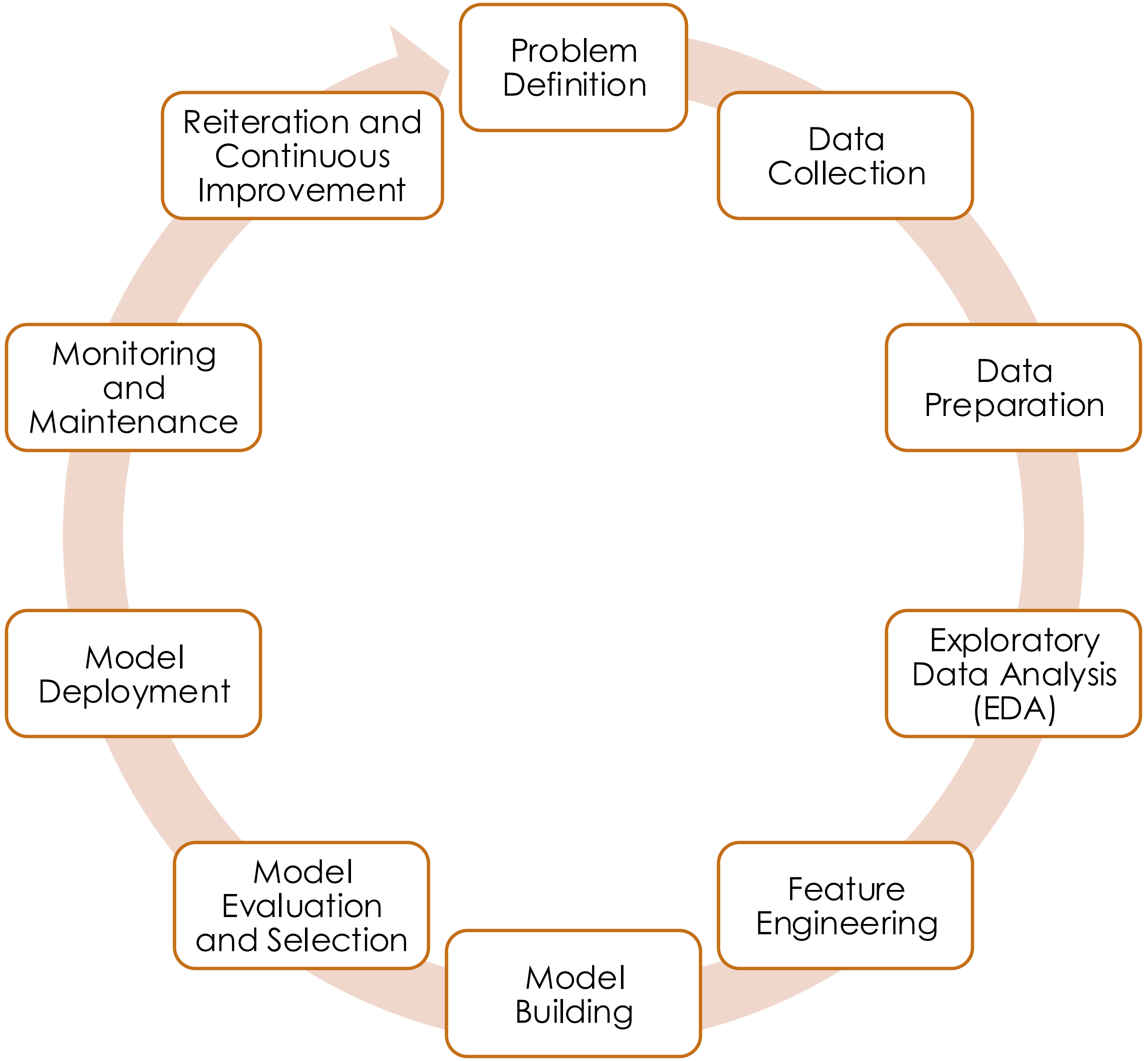
Scope of Data Science

- Analyzing, interpreting, and deriving meaningful patterns and trends from large volumes of data.
- Helping businesses make informed decisions by uncovering hidden patterns and correlations within data.
- Applying techniques from statistics, mathematics, computer science, and domain-specific knowledge.
- Collecting, cleaning, and processing data to derive actionable insights.
- Using machine learning algorithms and statistical models for data analysis.

Skills Required for Data Science Fresher

- Proficiency in programming languages such as Python or R.
- Strong foundation in statistics.
- Knowledge of machine learning techniques.
- Familiarity with data visualization tools.
- Curious mindset to explore and analyze data effectively.

Data Science Workflow and Lifecycle


Problem Definition
Data Collection
Data Preparation
Exploratory Data Analysis (EDA)
Feature Engineering
Model Building
Model Evaluation and Selection
Model Deployment
Monitoring and Maintenance
Reiteration and Continuous Improvement

Data Science Workflow and Lifecycle

Problem Definition

Involves understanding business problem that data science aims to address.
It is essential to clearly define the goals and objectives of the project.

Data Collection

Relevant data is collected from various sources, such as databases, APIs, or external datasets.
Data quality and quantity are crucial factors at this stage.

Data Preparation

Involves cleaning and preprocessing data to make it suitable for analysis.
Tasks include handling missing values, removing duplicates, standardizing formats, and transforming variables.

Exploratory Data Analysis (EDA)

Involves exploring data to identify patterns, trends, and relationships.
Data visualization techniques are often used to gain insights into the data.

Feature Engineering

Focuses on creating new features or variables from existing data to improve the performance of machine learning models.
Requires domain knowledge and creativity.

Data Science Workflow and Lifecycle

Model Building

Various machine learning algorithms are applied to the data to build predictive models.

Models are trained using historical data and validated using different techniques.

Model Evaluation and Selection

Involves evaluating the performance of the models using metrics such as accuracy, precision, recall, or F1 score.

The best-performing model is selected for deployment.

Model Deployment

Model is deployed into production where it can make predictions on new data.

Involves integrating the model into existing systems or apps.

Monitoring and Maintenance

Model's performance is monitored to ensure that it continues to perform well over time.

Monitoring involves tracking model accuracy, detecting drift, and retraining the model when necessary.

Reiteration and Continuous Improvement

Improves models and processes based on feedback and new data.

Continuous improvement involves refining models, updating data collection strategies, and adapting to changing business requirements.

Data Science key applications & use cases

Data science has a wide range of applications and use cases across various industries.

- Predictive analytics: To predict future trends and outcomes based on historical data. This is widely used in areas such as finance, healthcare, marketing, and e-commerce.
- Customer segmentation and targeting: To segment customers based on behavior and preferences, allowing for targeted marketing campaigns and personalized customer experiences.
- Fraud detection: To detect fraudulent activities in industries such as banking, insurance, and e-commerce by analyzing patterns and anomalies in transaction data.

Data Science key applications & use cases

- Recommendation systems: To suggest products, movies, music, or content based on user preferences and behavior. This is commonly used in e-commerce, streaming platforms, and social media.
- Supply chain optimization: To optimize supply chain operations by predicting demand, identifying bottlenecks, and improving inventory management through data-driven decision-making.
- Sentiment analysis: To analyze and interpret customer feedback, reviews, and social media data to understand sentiment and extract insights.

Data Science key applications & use cases

- Healthcare analytics: To improve patient outcomes, optimize treatment plans, predict disease outbreaks, and personalize medicine through analysis of electronic health records and medical imaging data.
- Image and video recognition: For object and pattern recognition in images and videos, enabling applications such as facial recognition, autonomous vehicles, and quality control in manufacturing.
- Financial risk management: To assess and manage risks in financial markets by analyzing market data, credit scores, and other factors to make informed investment decisions.

Types of Machine Learning

Machine learning can be broadly categorized into three main types:

1. Supervised learning
2. Unsupervised learning, and
3. Reinforcement learning.

Supervised Learning

- Model is trained on labeled data, where each data point is paired with the corresponding target or label. The goal is to learn the mapping between input features and output labels to make predictions on unseen data.
- Supervised learning tasks include classification (predicting a discrete label), regression (predicting a continuous value), and recommendation systems.
- Examples of supervised learning algorithms include Linear Regression, Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, K-Nearest Neighbors (KNN), and Neural Networks.
- Supervised learning involves two main phases: training, where the model learns from labeled data, and testing, where the model is evaluated on unseen data to assess its performance.

Unsupervised Learning

- Model is trained on unlabeled data, where the goal is to find hidden patterns or structures in the data without explicit labels.
- Unsupervised learning tasks include clustering (grouping similar data points together), dimensionality reduction (reducing the number of features), and anomaly detection (identifying outliers).
- Examples of unsupervised learning algorithms include K-Means Clustering, Hierarchical Clustering, Principal Component Analysis (PCA), t-SNE, and Gaussian Mixture Models.
- Unsupervised learning is used for exploratory data analysis, data preprocessing, and generating insights from unstructured data.

Reinforcement Learning

- Model learns to make sequential decisions through trial-and-error interactions with an environment. The model learns to maximize a reward signal by taking specific actions in a given state.
- Reinforcement learning tasks include game playing, robotics control, autonomous driving, and recommendation systems.
- The model interacts with the environment, observes the rewards or penalties for its actions, and adjusts its strategy to achieve long-term goals.
- Examples of reinforcement learning algorithms include Q-Learning, Deep Q Network (DQN), Policy Gradient methods, and Actor-Critic algorithms.
- Reinforcement learning is used in dynamic and complex environments where the model needs to learn a policy through exploration and exploitation.

Tools and frameworks for Data Science

- Programming Language: Python, R
- Python packages:
  - NumPy - Numerical computation library
  - Pandas - Data manipulation tools
  - Matplotlib and Seaborn - Data visualization library
  - Scikit-learn - Machine learning library
  - TensorFlow and PyTorch - Deep learning frameworks
  - Keras - High-level deep learning
- IDEs: Jupyter Notebook, Visual Studio Code, R Studio
- SQL
- Git, and GitHub
- Cloud Technologies – AWS, GCP

Summary

What we've learned so far

Artificial Intelligence

Makes machines smart like humans. It helps them see, hear, understand language, and make decisions.

Data Science

Helps people find valuable information in data. It involves analyzing and visualizing data to see patterns and make better decisions.

Machine Learning

Teaches computers to learn from data and make predictions without being explicitly programmed. It helps machines get better at tasks over time.

Deep Learning

A type of Machine Learning inspired by the human brain. It uses neural networks to learn complex patterns in data, like recognizing images or understanding speech.

Summary

What we've learned so far

Generative AI

A smart machine that can create new things, such as writing stories, drawing pictures, or composing music. It learns from existing examples and data to produce its own original creations.

Data Science Lifecycle

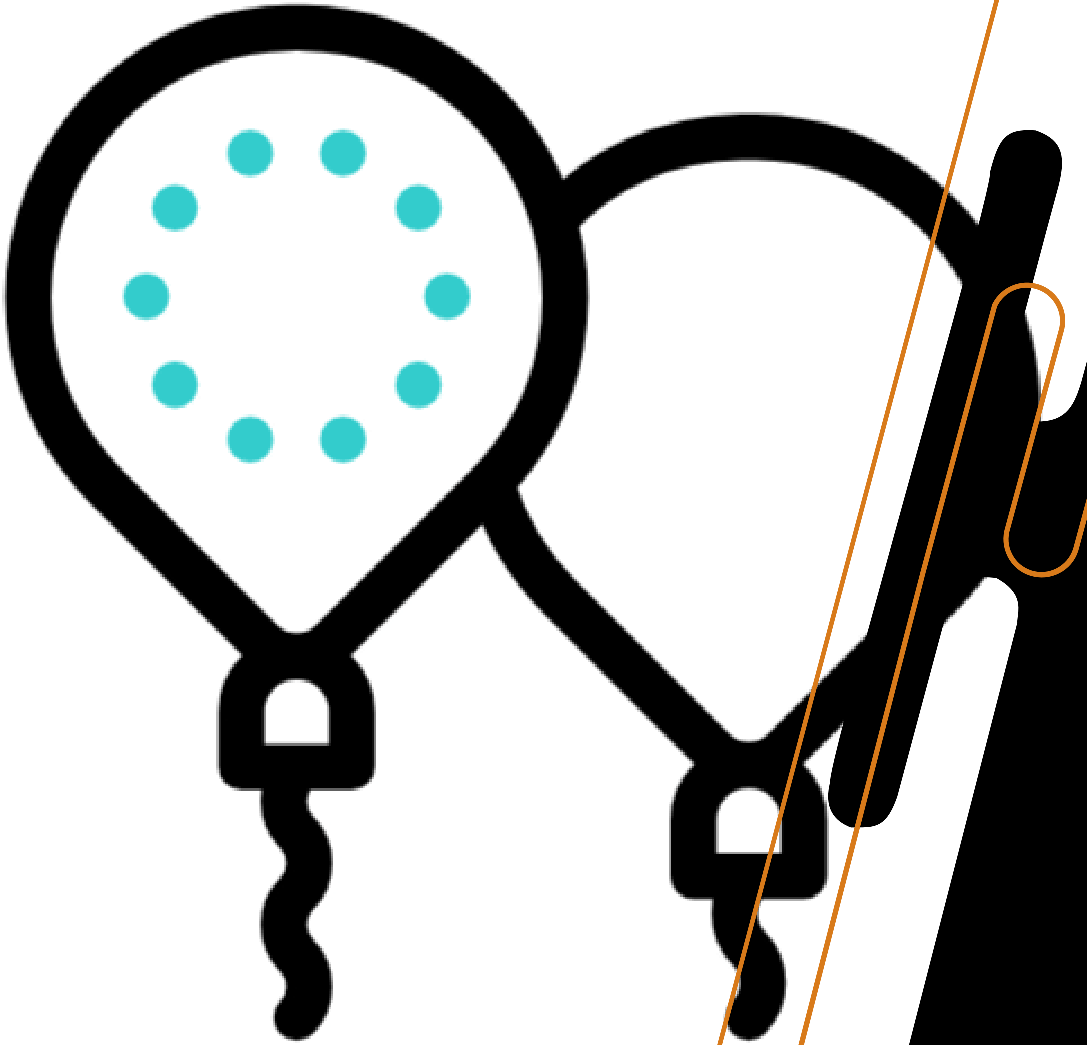
Problem definition, Data preprocessing, Model development, evaluation, selection and deployment, Monitoring and Continuous Improvement

Applications of Data Science

Fraud detection, Sentiment Analysis, Recommendation systems, Customer Segmentation, etc.,

Types and Machine Learning

Supervised, unsupervised and Reinforcement Learnings



Thank you

Please send all your questions to:
nandini@datavalley.ai