Introduction To Big Data Analytics INSY 8413

ADVENTIST UNIVERSITY OF CENTRAL AFRICA

Instructor:

- Eric Maniraguha | eric.maniraguha@auca.ac.rw | LinkedIn Profile

6h00 pm – 8h50 pm
- Monday A-G104
- Tuesday E-G108
- Wednesday A-G104
- Thursday E-G108

2h30 pm – 8h50 pm
- Sunday B- G205

June 2025

Reference reading

- Machine Learning Tutorial
- Types of Machine Learning
- 5 Different Types of Neural Networks
- A review of deep learning-based detection methods for COVID-19

Lecture 04 – Introduction to Machine Learning

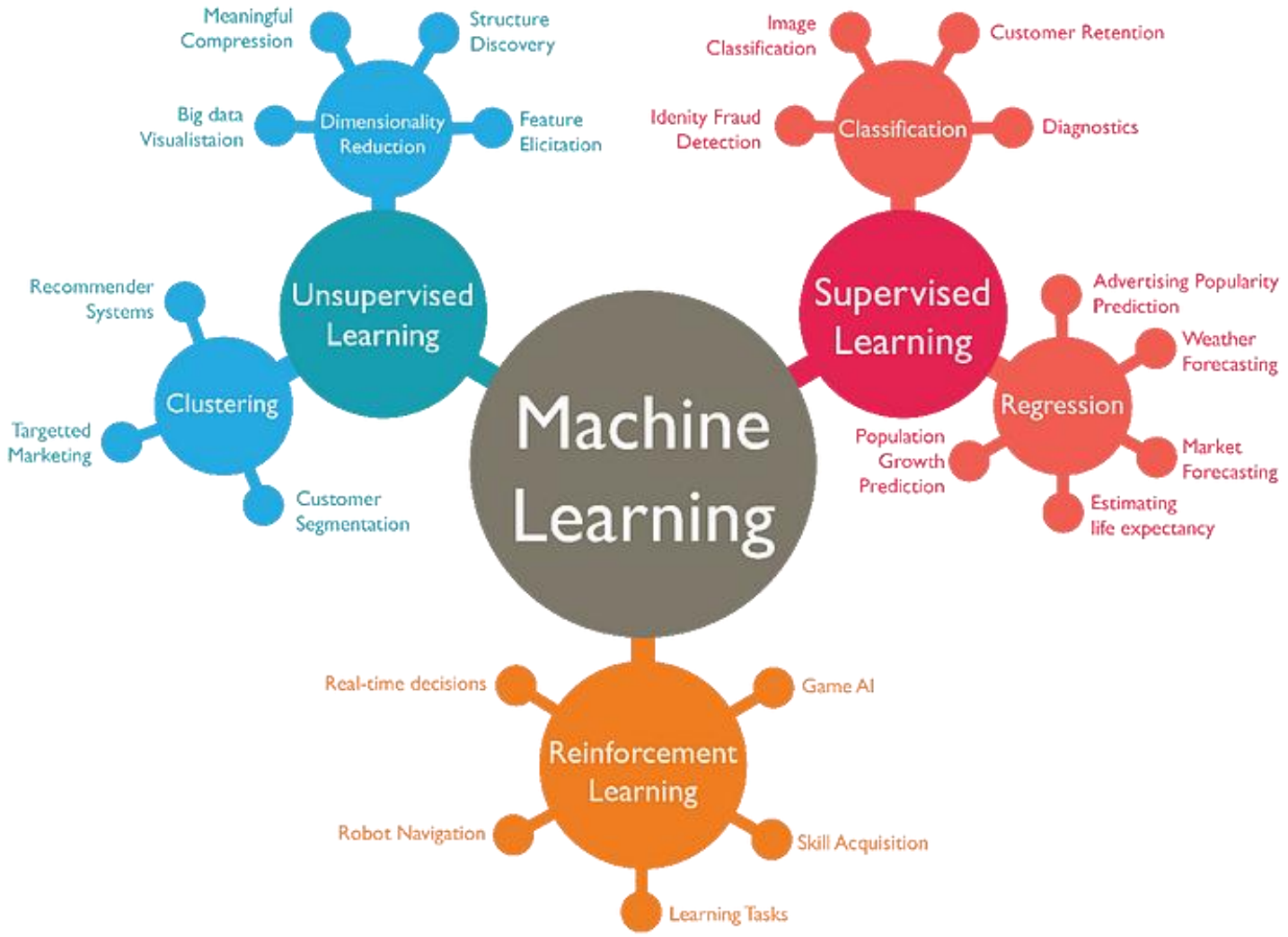
Lecture Objective

To provide students with a **foundational understanding of Machine Learning concepts, techniques, and applications**, enabling them to analyze data, build simple models, and understand the role of ML in Artificial Intelligence systems.

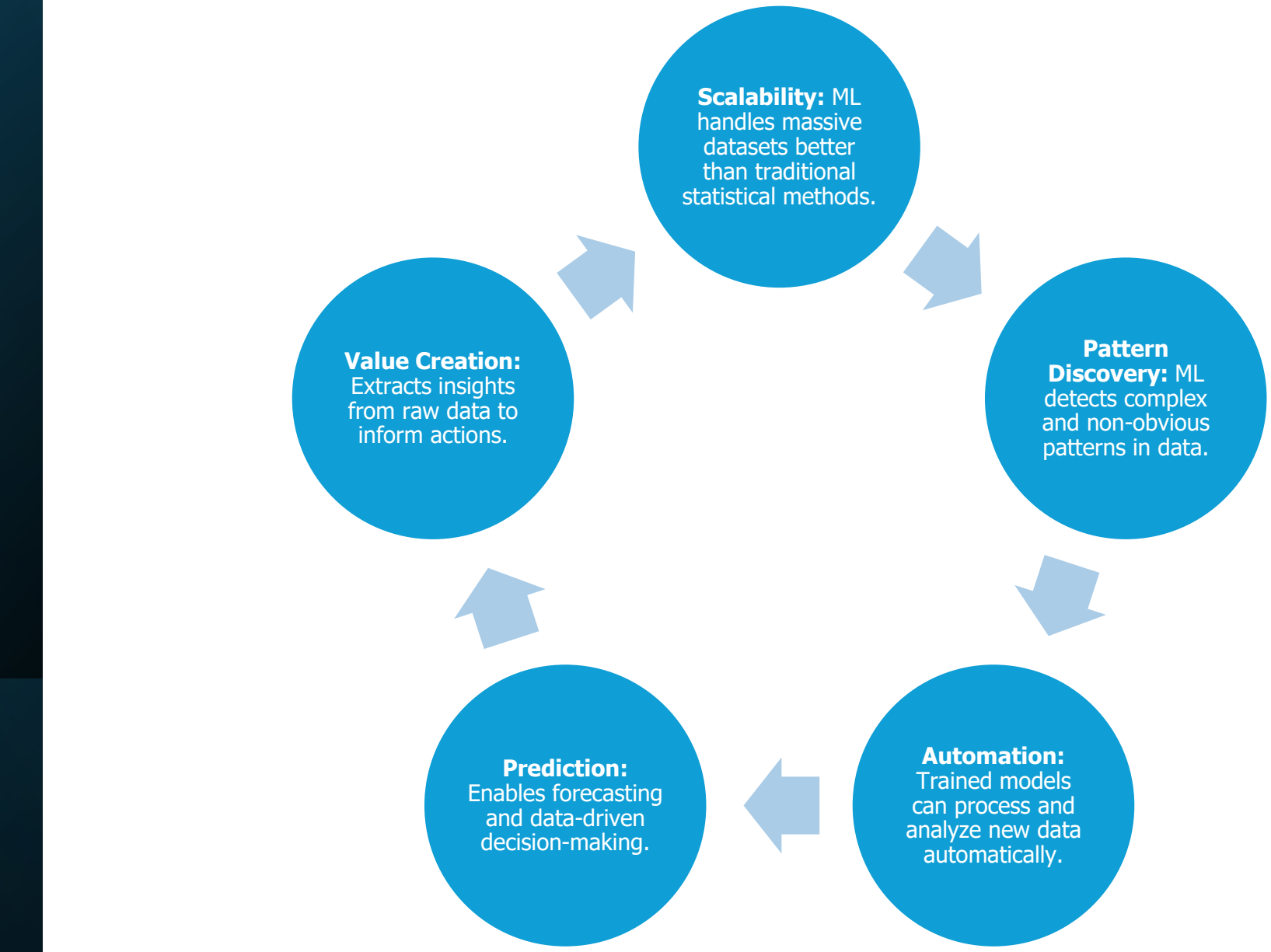
What is Machine Learning?

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that **enables computers to learn from data and make decisions or predictions without being explicitly programmed for every scenario**. Rather than writing fixed rules for all possibilities, ML uses algorithms that identify patterns in data and improve over time.

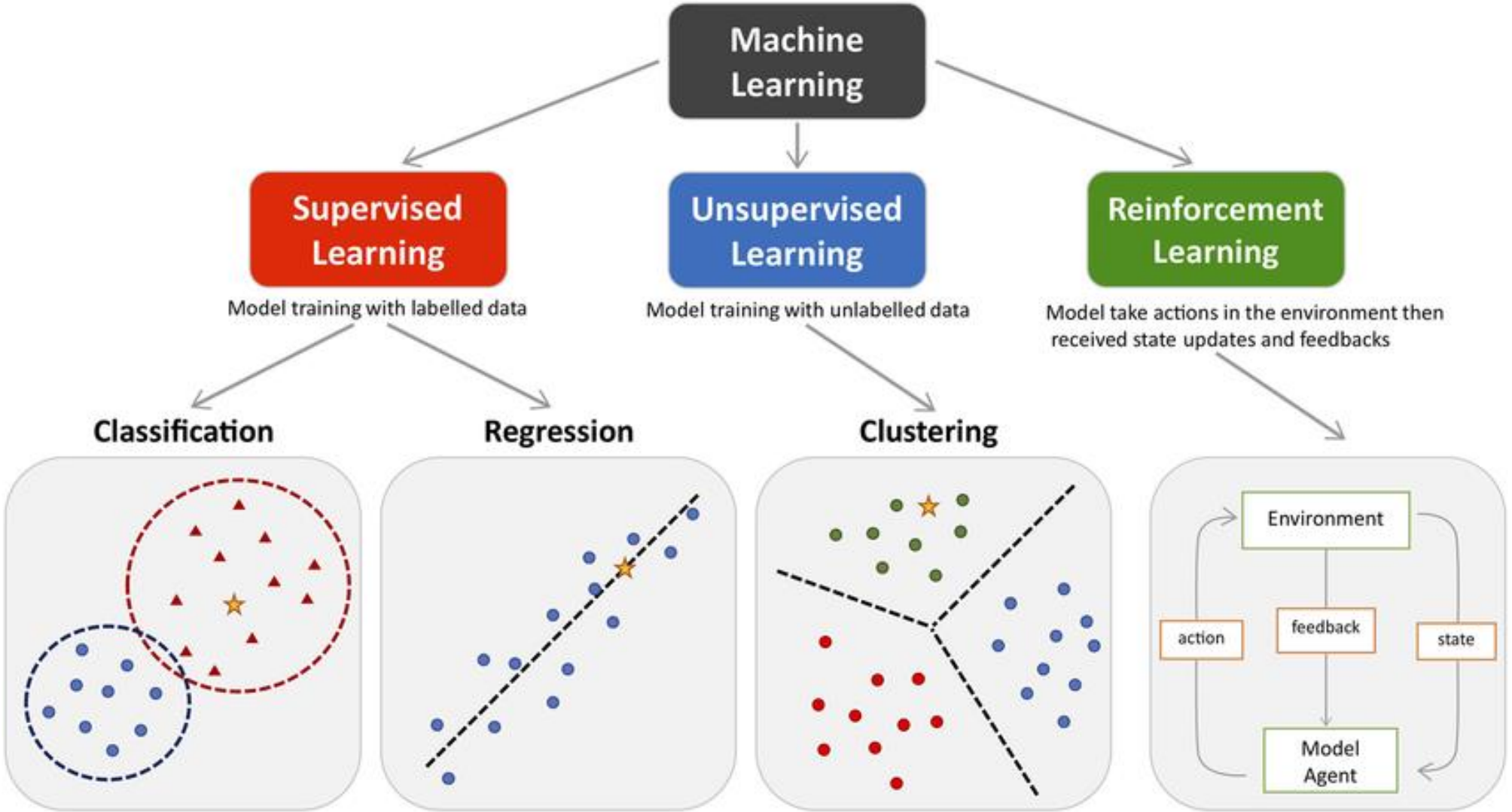
Key Concept: ML systems learn and improve their performance on tasks through experience with data.



Why Use Machine Learning in Big Data?



Types of Machine Learning



Types of Machine Learning

Machine Learning (ML) is typically **classified based on the type of learning signal or feedback** available to a learning system.

Supervised
(Learning by examples)

Classification
- Disease detection
- Email spam detection
- Image classification
- Bank loan prediction

Regression
- House price prediction
- Stock price prediction

Unsupervised
(Learning by observation)

Clustering
- Search engines
- Face recognition
- Targeted marketing
- Recommender system

Association Rule Mining
- Market basket analysis
- Medical diagnosis
- Census data

Semi-Supervised
(Learning from few labeled and many unlabeled examples)

Example Applications
- Speech analysis
- Web content classification
- Text document categorization
- Protein sequence classification
- Large-scale image recognition

Reinforcement
(Learning from mistakes)

Example
- Self-driving car
- Gaming AI
- Robot navigation
- Inventory management
- Finance sector

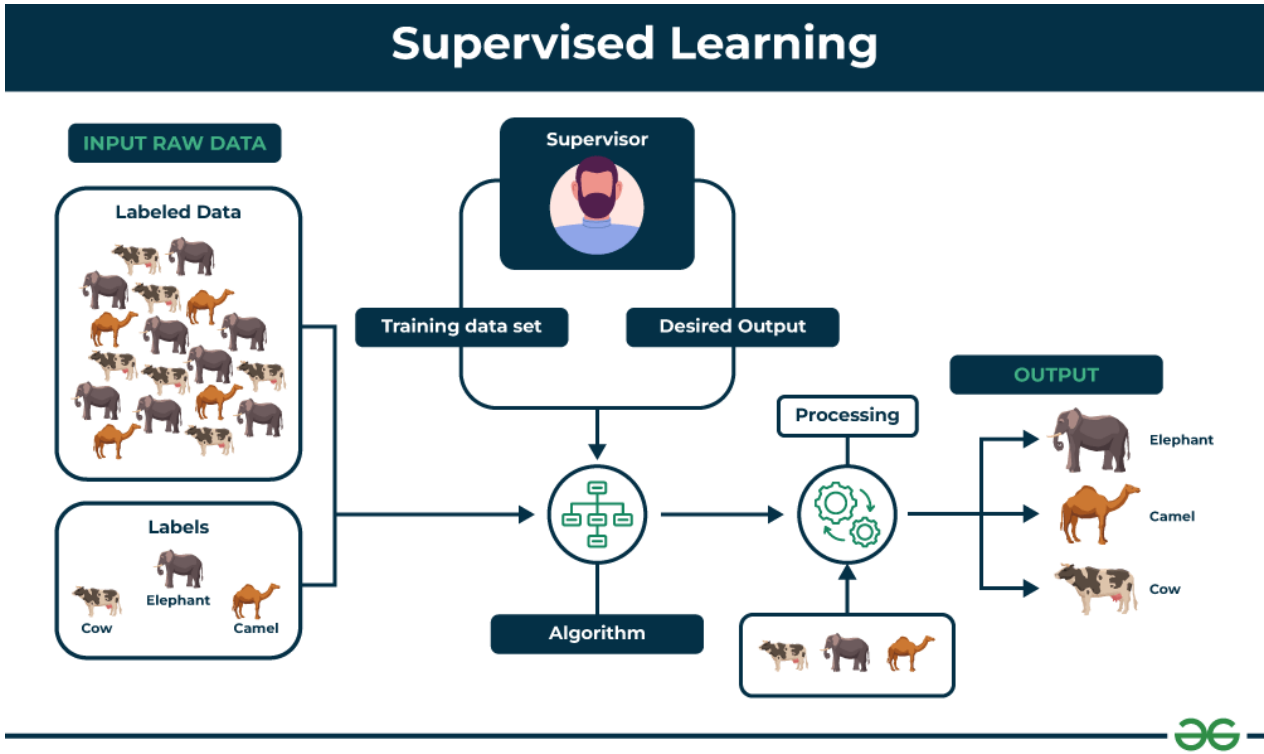
Types of Machine Learning - Supervised Learning

Supervised Learning: In supervised learning, the algorithm is trained on **a labeled dataset**, meaning each **data point is associated with a known output or target variable.** The goal is to learn a mapping function that can predict the output for new, unseen data.

Goal: Learn a mapping from inputs to outputs.

Examples include:

- **Classification:** Predicting a categorical output (e.g., Email spam detection
- or not spam, cat or dog).
- **Regression:** Predicting a continuous output (e.g., house price, temperature).
- Image classification
- Loan default prediction



Source Image: https://www.geeksforgeeks.org/machine-learning/machine-learning/

Types of Supervised Learning

1. Classification
Classification assigns input data into predefined categories. Popular algorithms include:

- Linear Classifiers
- Support Vector Machines (SVM)
- Decision Trees
- k-Nearest Neighbors (k-NN)
- Random Forests
- Boosting methods

Deep learning models, especially neural networks, excel at complex classification tasks, leveraging layers of interconnected "neurons" that transform data through weights, biases, and activation functions.

2. Regression
Regression is used to predict **continuous values**. For example, predicting house prices, temperature, or stock market trends.

Evaluation Metrics for Supervised Learning
For Regression:

- **Mean Squared Error (MSE)**: Average squared difference between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: Square root of MSE; penalizes large errors.
- **Mean Absolute Error (MAE)**: Average absolute difference between predictions and actual values.
- **R-squared (R^2)**: Proportion of variance in the dependent variable explained by the model.

For Classification:

- **Accuracy**: Percentage of correctly predicted instances.
- **Precision**: Ratio of true positives to total predicted positives.
- **Recall**: Ratio of true positives to total actual positives.
- **F1 Score**: Harmonic mean of precision and recall, useful for imbalanced datasets.
- **Confusion Matrix**: Visual representation of model performance across classes.

Real-World Applications

- **Image classification & object detection**: Classifying animals in images or detecting vehicles in autonomous driving.
- **Customer churn prediction & recommendations**: Anticipating customer exit or offering product suggestions.
- **Predictive modelling & regression**: Forecasting sales, stock prices, or housing values.

Types of Machine Learning - Unsupervised Learning

Unsupervised Learning: In unsupervised learning, the algorithm is trained on an **unlabeled dataset**, meaning there are no pre-defined output variables. The goal is to discover hidden patterns, structures, or relationships within the data.

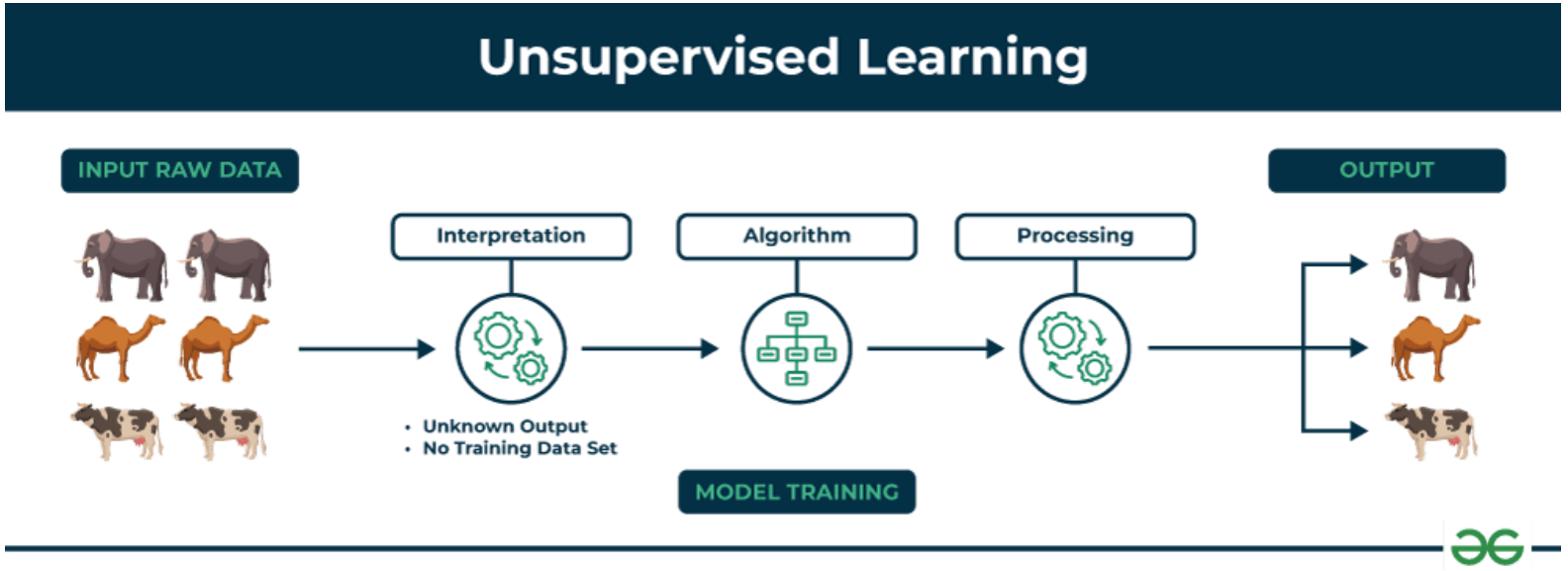
Examples include:

- **Clustering:** Grouping similar data points together (e.g., customer segmentation).
- **Dimensionality Reduction:** Reducing the number of variables while preserving important information (e.g., feature extraction).
- **Anomaly Detection:** Identifying unusual or outlier data points (e.g., fraud detection).

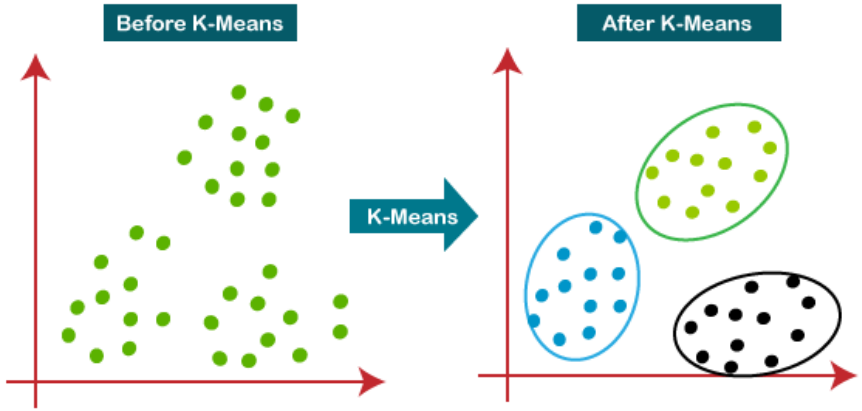
Goal: Discover patterns, groupings, or structure in data.

Real-World Applications

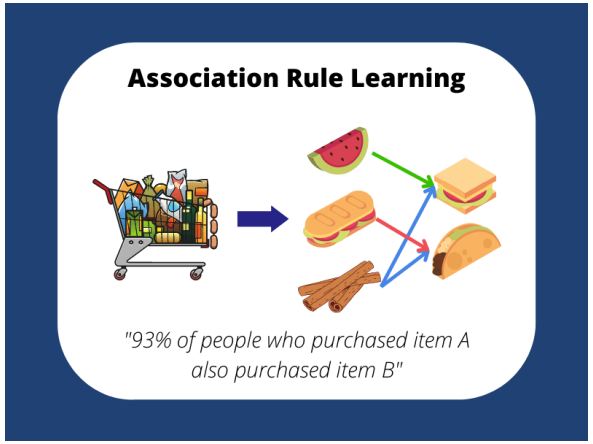
- **Image segmentation & anomaly detection**: Spotting unusual patterns like fraud or cyber threats.
- **Customer segmentation**: Grouping users for targeted marketing.
- **Exploratory Data Analysis (EDA)**: Uncovering insights in complex datasets.
- **Dimensionality reduction**: Simplifying high-dimensional data for visualization or model efficiency.



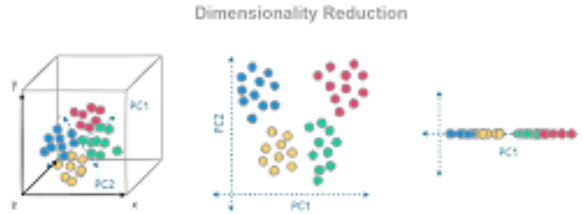
Types of Unsupervised Learning



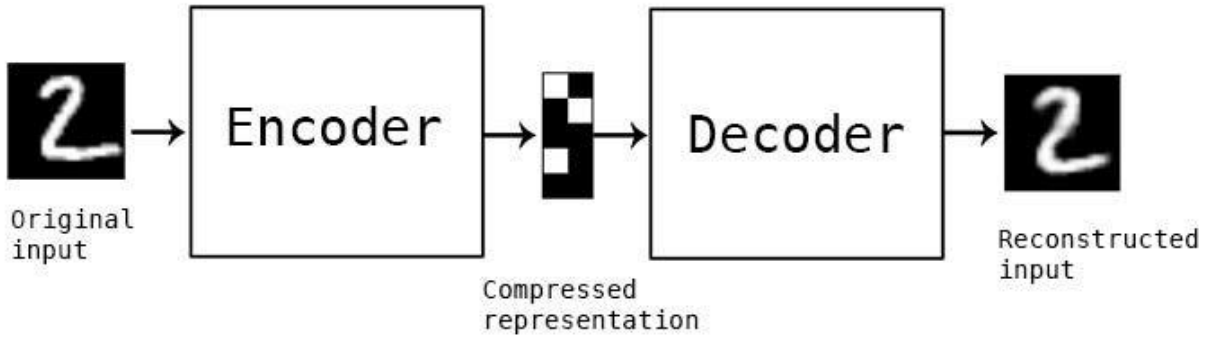
1. Clustering
Groups data points based on similarity. Clustering helps identify natural groupings in datasets, especially large and complex ones.



2. Association Rule Learning
Discovers interesting relationships between variables, such as market basket analysis.



3. Dimensionality Reduction
Techniques like PCA reduce the number of features while retaining essential information.



4. Autoencoders
Neural networks that learn efficient representations (encodings) of data, often used in anomaly detection or data compression.

Evaluation Metrics for Unsupervised Learning

- **Elbow Method**: Identifies the optimal number of clusters by plotting variance vs. number of clusters.
- **Silhouette Score**: Measures how similar an object is to its own cluster vs. others; ranges from -1 to 1.
- **Calinski-Harabasz Index**: Ratio of between-cluster dispersion to within-cluster dispersion.
- **Davies-Bouldin Index**: Evaluates average similarity between each cluster and its most similar cluster; lower is better.
- **Visual Inspection**: Tools like PCA or t-SNE project high-dimensional data into 2D/3D to visually assess cluster quality.

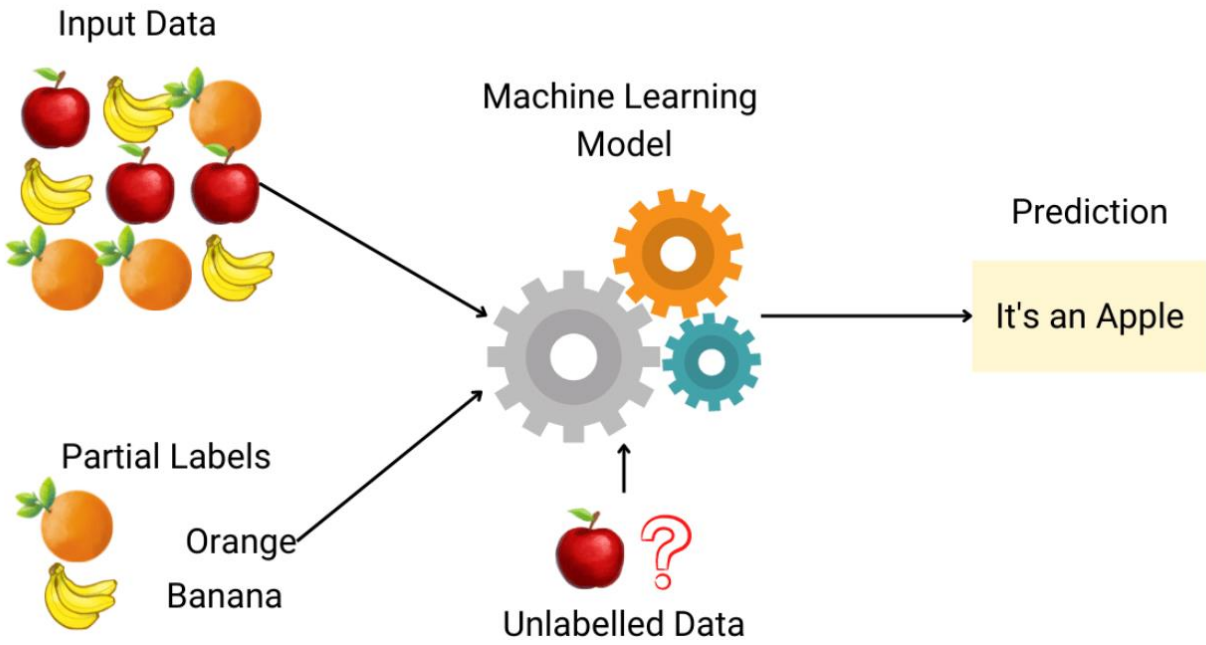
Semi-Supervised Learning

Definition: The model learns from a small amount **of labeled data and a large amount of unlabeled data**.

Goal: Improve learning accuracy using both labeled and unlabeled data.

Examples:

- Text classification with few labeled examples
- Medical image classification
- Web content categorization

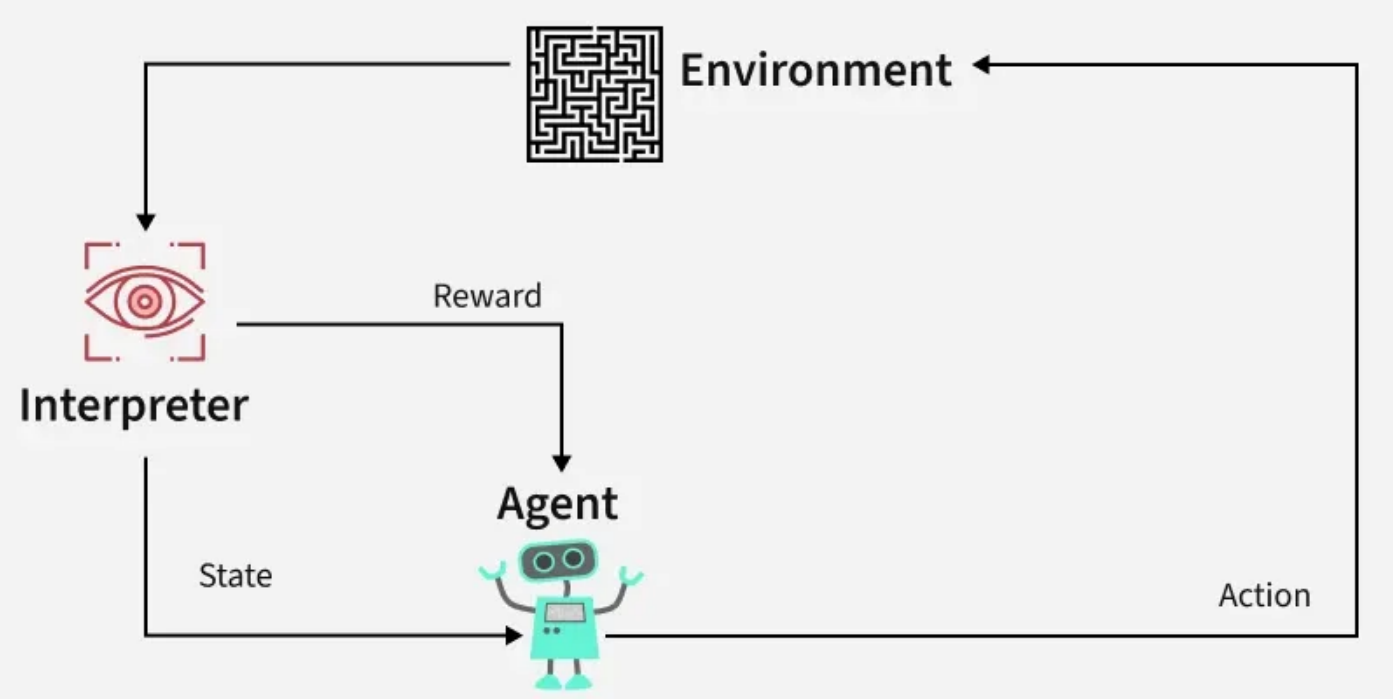


Source Image: https://medium.com/@gayatri_sharma/a-gentle-introduction-to-semi-supervised-learning-7afa5539beea

Types of Reinforcement Learning

Reinforcement Learning: In reinforcement learning, the algorithm learns to make decisions in an environment to maximize a reward. The algorithm (agent) interacts with the environment, receives feedback (reward or punishment), and learns to optimize its actions over time. Examples include:

- **Game playing:** Training an AI to play games like chess or Go.
- **Robotics:** Training a robot to perform tasks in a physical environment.
- **Resource management:** Optimizing the allocation of resources in a system.



Reinforcement Learning revolves around the idea that an agent (the learner or decision-maker) interacts with an environment to achieve a goal. The agent performs actions and receives feedback to optimize its decision-making over time.

- **Agent**: The decision-maker that performs actions.
- **Environment:** The world or system in which the agent operates.
- **State:** The situation or condition the agent is currently in.
- **Action:** The possible moves or decisions the agent can make.
- **Reward:** The feedback or result from the environment based on the agent's action.

Here's a breakdown of RL components:

- **Policy:** A strategy that the agent uses to determine the next action based on the current state.
- **Reward Function:** A function that provides feedback on the actions taken, guiding the agent towards its goal.
- **Value Function:** Estimates the future cumulative rewards the agent will receive from a given state.
- **Model of the Environment:** A representation of the environment that predicts future states and rewards, aiding in planning.

Basics of scikit learn

Note: Scikit-learn Basics & Hands-on: Simple ML demo – To Be Covered Later

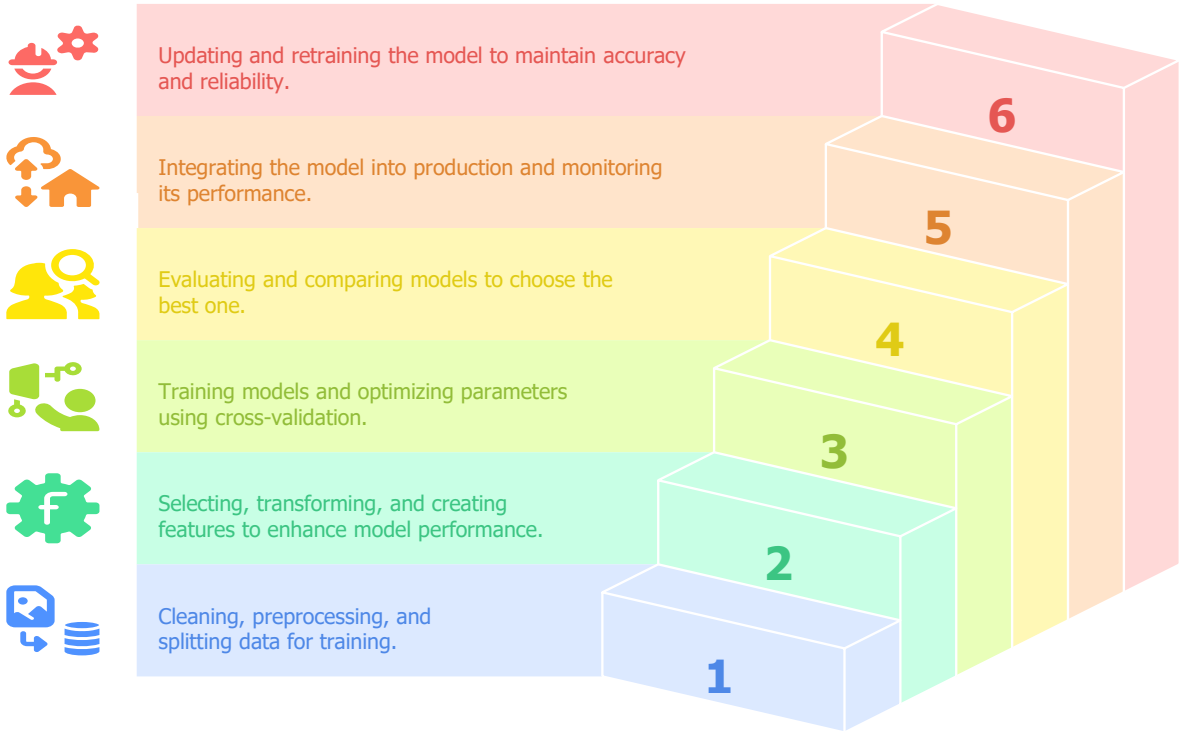
This section on Scikit-learn will be introduced only after learners have a solid understanding of Python basics (variables, loops, functions, lists, dictionaries, etc.).

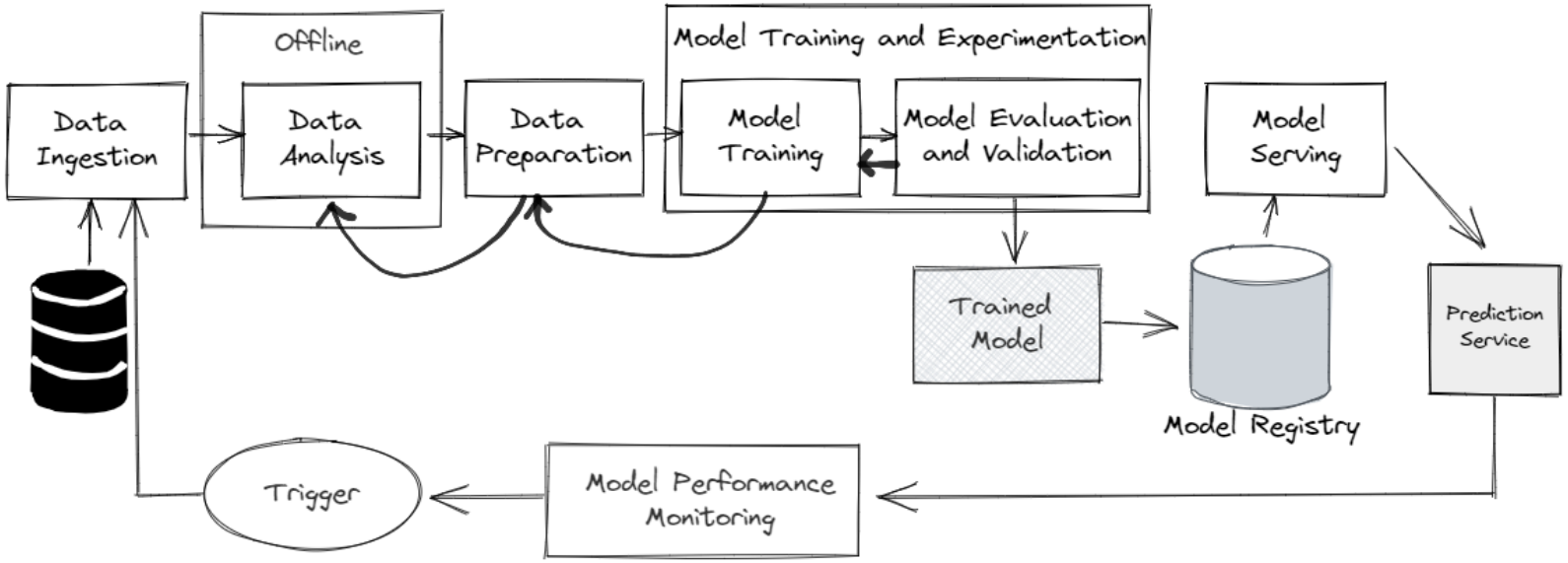
Once Python fundamentals are clear, we will explore:

- Loading datasets
- Splitting data
- Training models
- Evaluating performance
- Popular algorithms in Scikit-learn

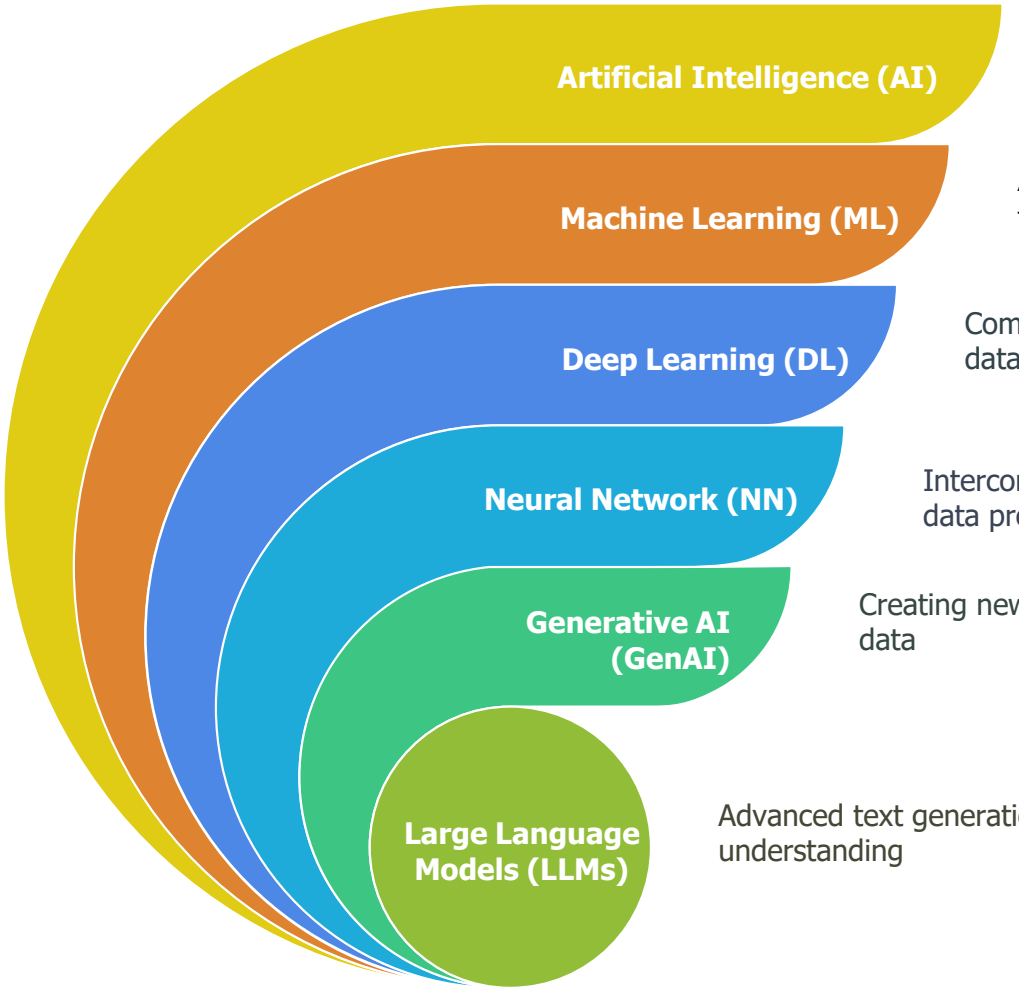
Planned Module: *Machine Learning with Scikit-learn (After Python Basics)*

The Supervised Machine Learning Pipeline





Hierarchy of AI Technologies



Creating machines that can think and act like human beings.

Algorithms, often based on statistics, that learn patterns from data to make predictions or decisions.

Complex neural networks for data analysis

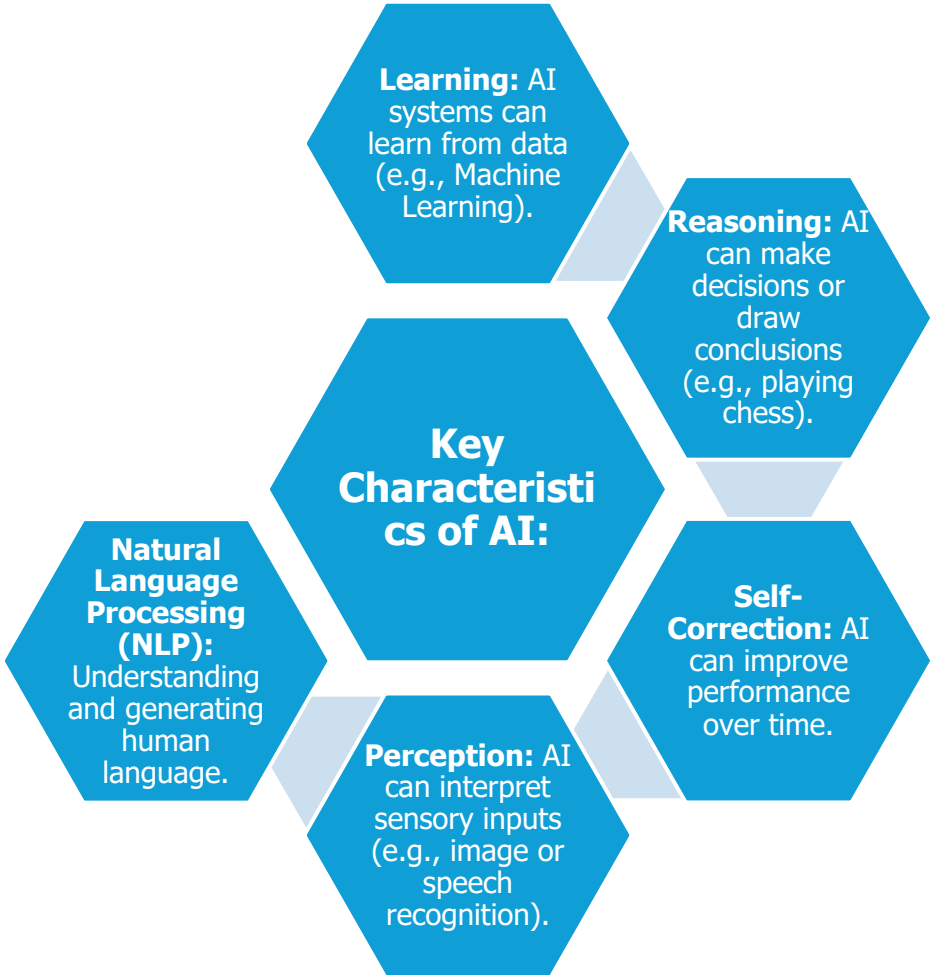
Interconnected nodes for data processing

Creating new content from data

Advanced text generation and understanding

What is Artificial Intelligence (AI)?

Artificial Intelligence (AI) refers to the ability of machines or computer systems **to simulate human intelligence to perform tasks such as learning, reasoning, problem-solving, perception, and language understanding.**



Example: AI in Healthcare – Medical Diagnosis Systems
AI System: IBM Watson for Oncology
Function: Assists doctors by analyzing patient data and medical literature to suggest personalized cancer treatment options.

What is Deep Learning?

Deep Learning is a subset of **machine learning** that uses algorithms inspired by the **structure and function of the human brain**, called **artificial neural networks** (ANNs).
It focuses on learning patterns from **large amounts of data** through multiple layers of computation.

Key Characteristics of Deep Learning:

- **Multiple Layers (Depth):**
  Involves many layers of neurons — input layer, hidden layers, and output layer — allowing the system to learn complex features.
- **Feature Learning:**
  Automatically extracts relevant features from raw data without the need for manual feature engineering.
- **Large Data Requirement:**
  Performs best with big datasets (e.g., images, videos, text).
- **High Computational Power:**
  Often needs GPUs/TPUs for training deep models due to complexity.
- **Backpropagation:**
  Uses backpropagation algorithm to update weights based on error/loss.

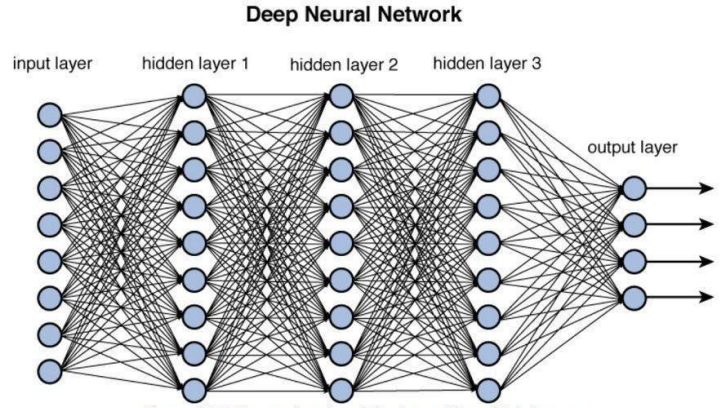


Figure 12.2 Deep network architecture with multiple layers.

How It Works (Simplified):

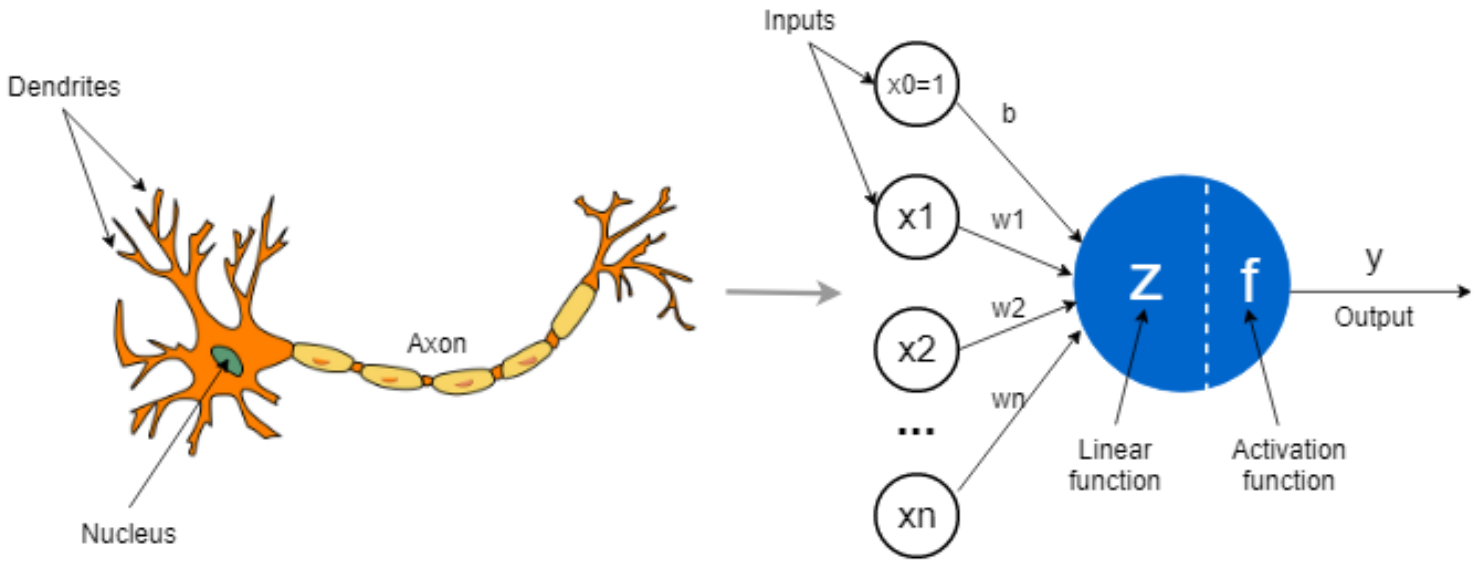
1. Input data is fed into the **input layer**.
2. Passes through **multiple hidden layers** with activation functions.
3. Final result is given at the **output layer**.
4. Model compares the prediction to actual result → computes error.
5. Error is **propagated back** to update the network's weights.

Examples of Deep Learning Applications:

- Self-driving cars (e.g., **Tesla Autopilot**)
- Voice assistants (e.g., **Siri, Google Assistant**)
- Facial recognition (e.g., **Facebook tagging**)
- Medical diagnosis (e.g., **detecting cancer in X-rays**)
- Language translation (e.g., **Google Translate**)

Neural Networks (NN) as an Example of Deep Learning

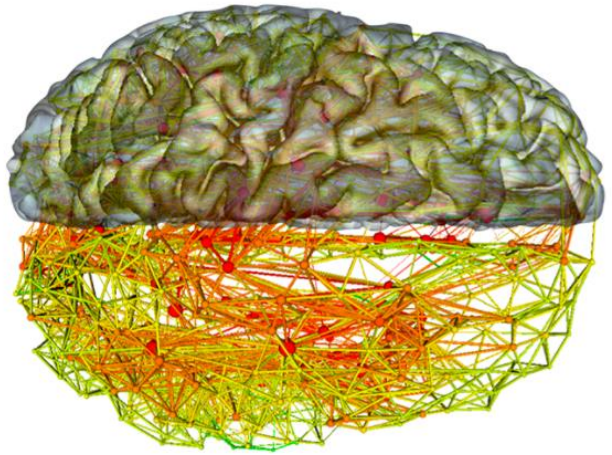
A **Neural Network** is a subset of **Machine Learning** inspired by the structure and functioning of the **human brain**. It consists of layers of nodes (called *neurons*) that process input data and identify patterns to make decisions or predictions.



Source Image: https://towardsdatascience.com/the-concept-of-artificial-neurons-perceptrons-in-neural-networks-fab22249cbfc/

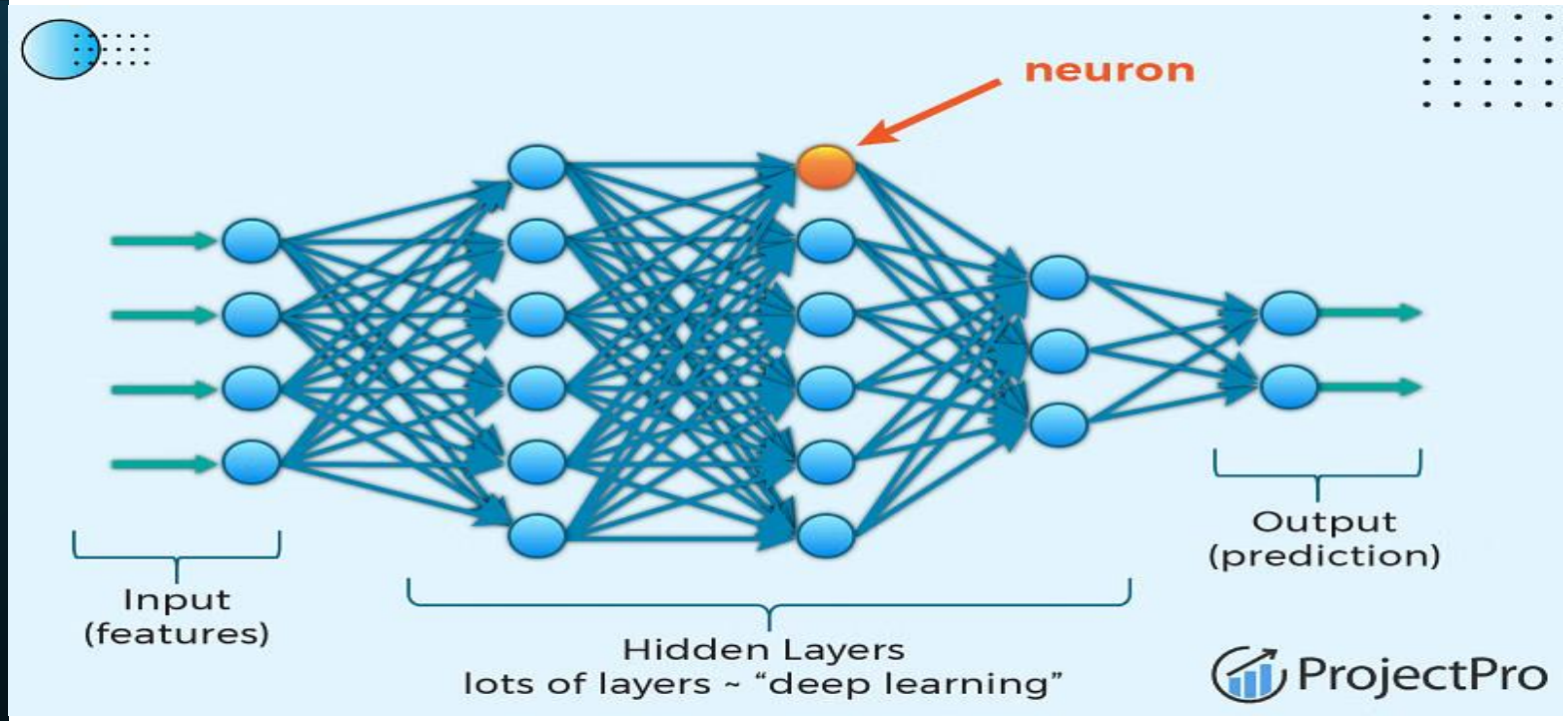
cture of a Neural Network:

- **Input Layer:** Receives raw data (e.g., image pixels, text, numbers).
- **Hidden Layers:** Perform computations and extract features.
- **Output Layer:** Produces the final result (e.g., class label, prediction).



Source Image: https://towardsdatascience.com/the-concept-of-artificial-neurons-perceptrons-in-neural-networks-fab22249cbfc/

Neural Networks (NN)



Source Image: https://www.projectpro.io/article/deep-learning-architectures/996

Example: Handwritten Digit Recognition
System: MNIST Digit Classifier
Task: Recognize digits (0–9) from images of handwritten numbers.
How it works:

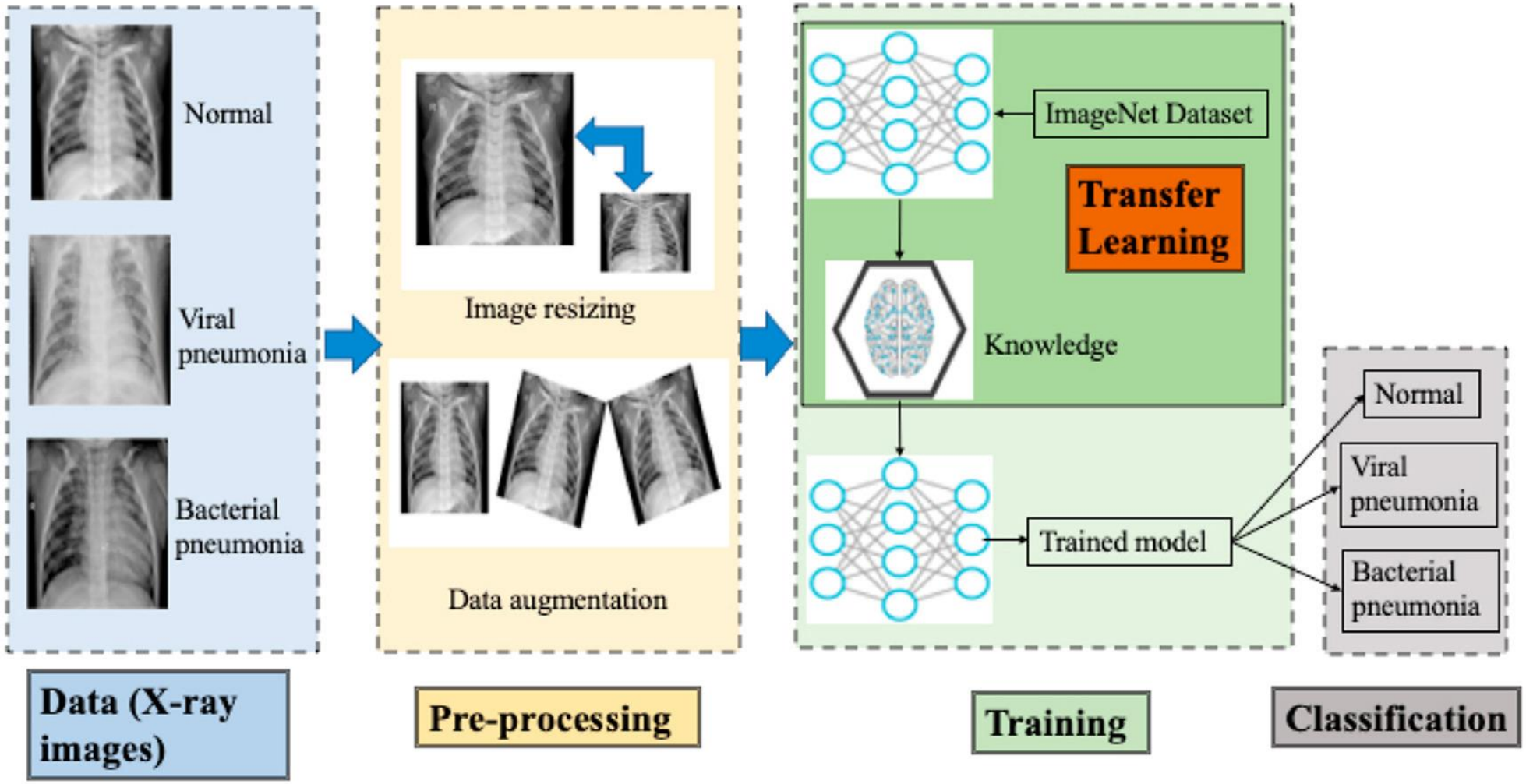
- Input: Image of a handwritten digit (28x28 pixels).
- NN processes the image through multiple layers.
- Output: Predicted digit (e.g., "5").

Key Characteristics of Neural Networks:

- **Pattern Recognition:** Learns features like shapes and curves in images.
- **Adaptability:** Improves as more data is fed during training.
- **Non-linear Modeling:** Can solve complex problems traditional models can't.

Image Processing

COVID-19 detectionDL-Based COVID-19 detectionLung image classificationCoronavirus pandemicMedical image processing



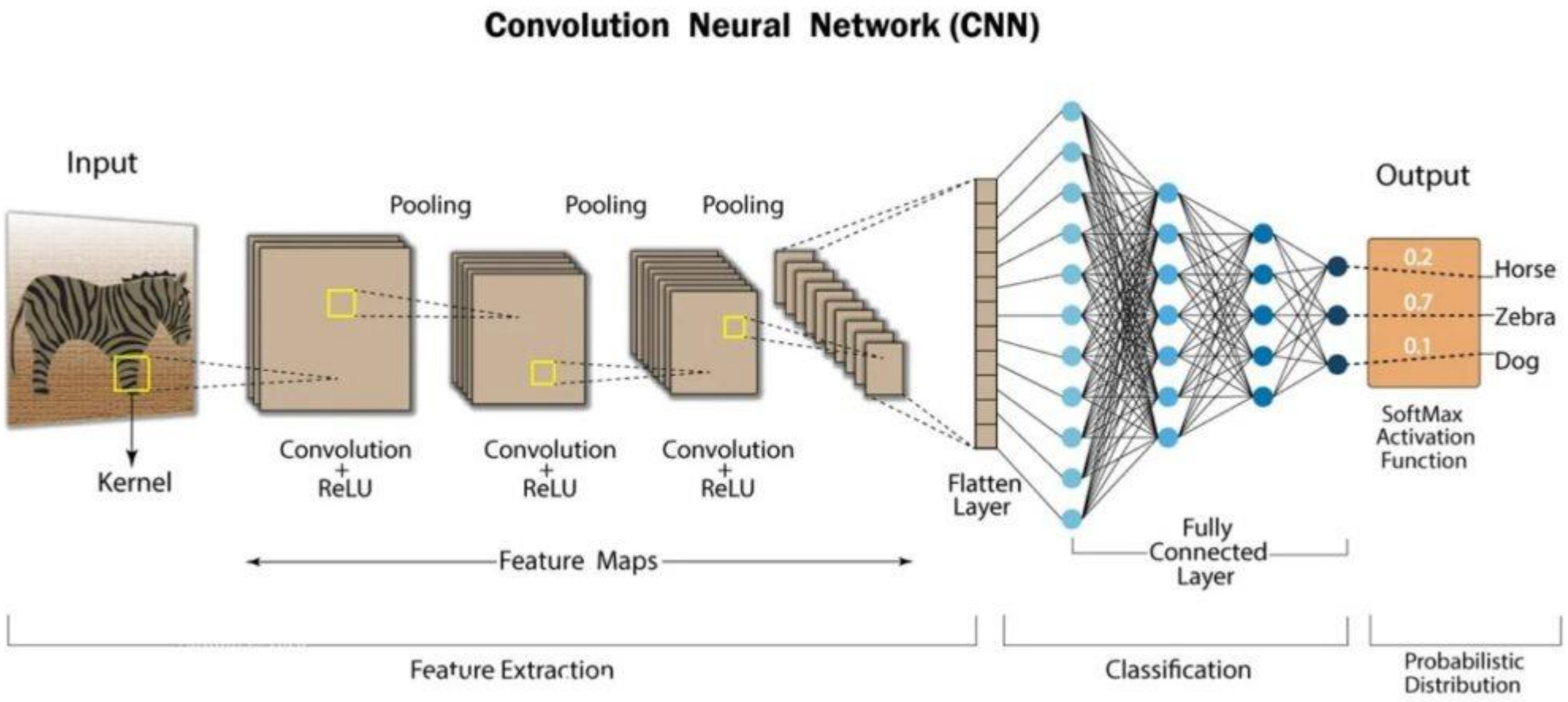
Source Image: https://www.sciencedirect.com/science/article/pii/S0010482522000257

Detect Objects

Generating YOLO-NAS Predictions

Source Imagehttps: https://voxel51.com/blog/state-of-the-art-object-detection-with-yolo-nas-fiftyone

Convolutional Operation



Source Image: https://svitla.com/blog/cnn-for-image-processing/

Generative AI (GenAI)

A type of Artificial Intelligence that can **create new content** (text, images, audio, video, code, etc.) by learning patterns from existing data.

Key Characteristics:

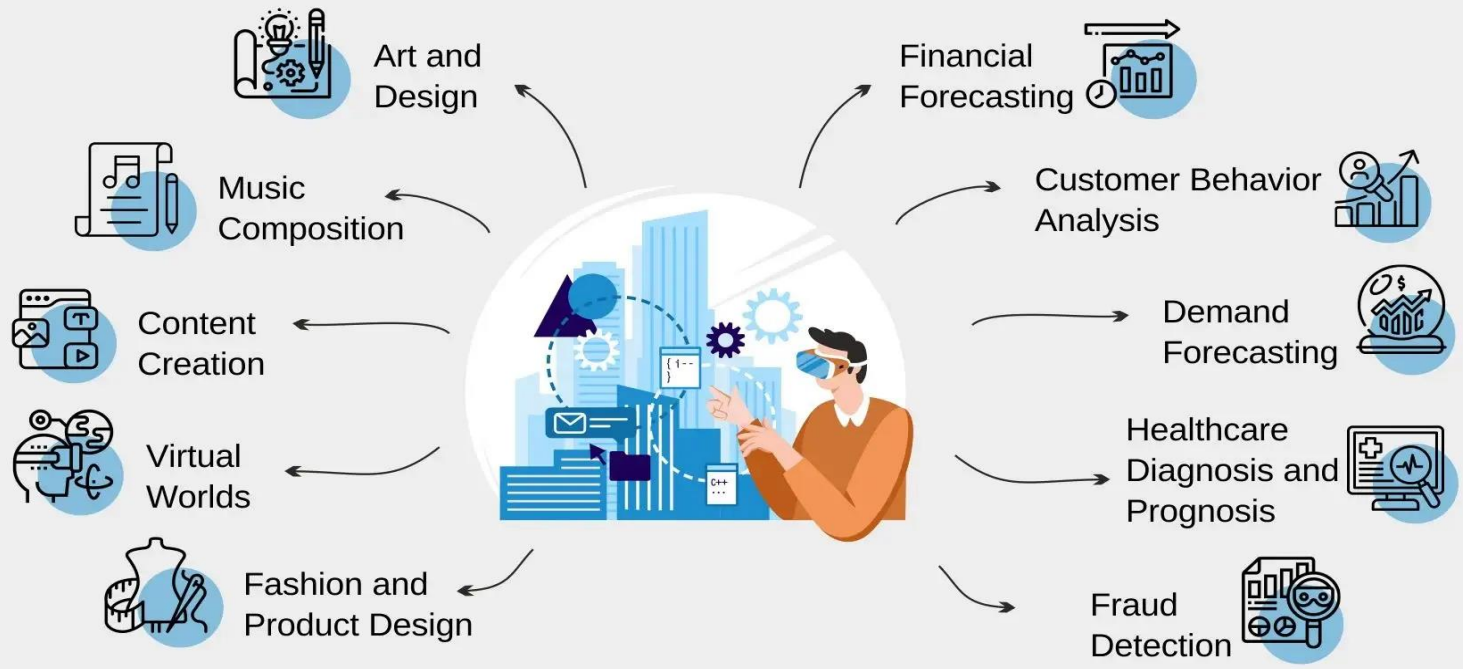
- Learns from large datasets
- Can generate **realistic and creative outputs**
- Often powered by **deep learning models** like **transformers**

Examples:

- **ChatGPT** – generates human-like text
- **DALL·E / MidJourney** – creates images from text prompts
- **Sora by OpenAI** – generates videos from text
- **Jukebox** – generates music
- **GitHub Copilot** – writes code suggestions

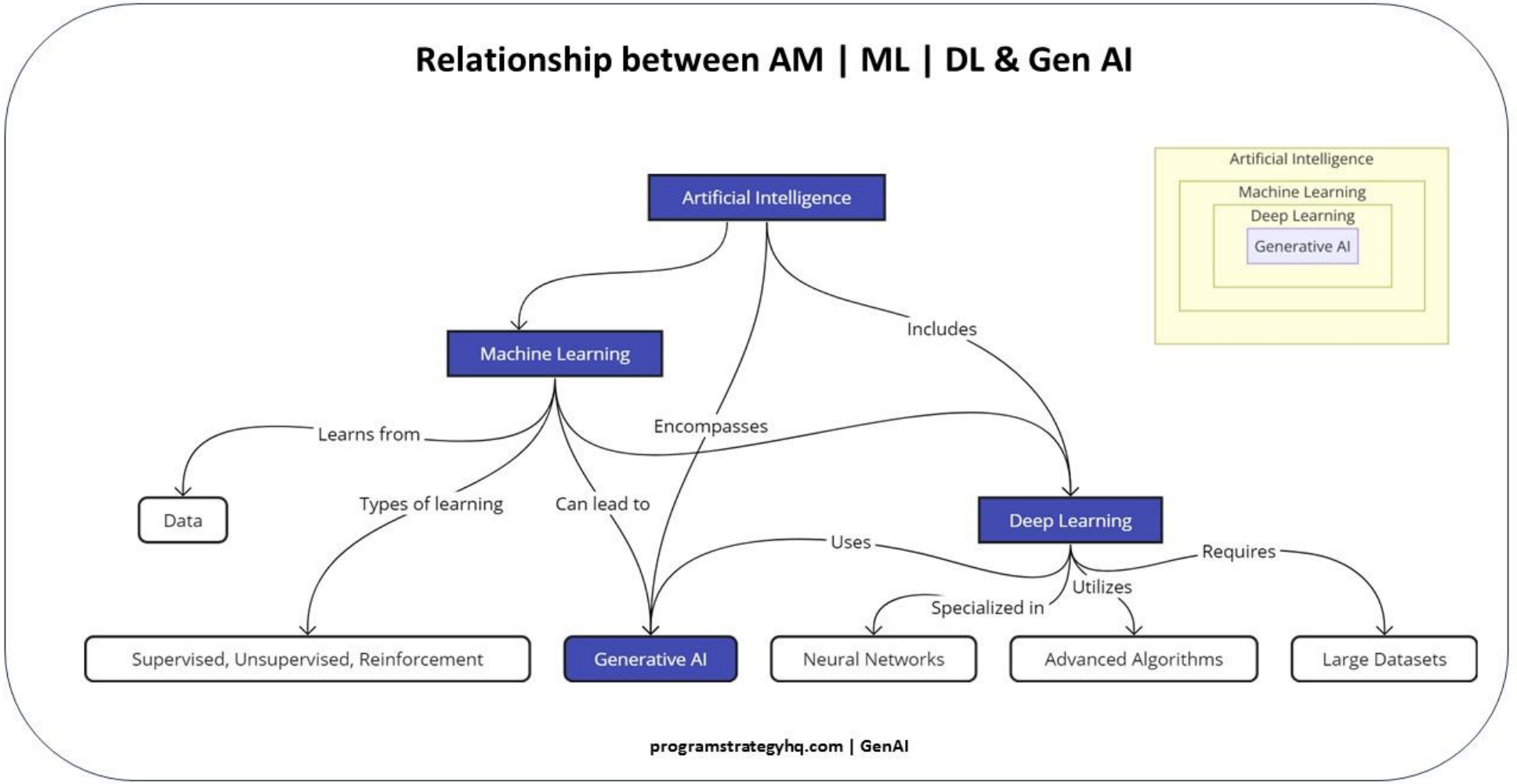
Neebal Technologies
www.neebal.com

Generative AI Applications



Source Image: https://botpenguin.com/blogs/understanding-the-basics-of-generative-ai-development

GenAI Overview | AI ML Basics



Source Image: https://www.programstrategyhq.com/post/ai-ml-basics

Too Lazy to Read? Just Click and Watch!

- Machine Learning Explained in 100 Seconds
- What's the Difference Between AI, Machine Learning, and Deep Learning?
- Generative AI explained in 2 minutes
- Explained In A Minute: Neural Networks
- LLM Explained | What is LLM
- How to evaluate ML models | Evaluation metrics for machine learning

NLP vs LLM – What's the Difference?

Aspect	NLP (Natural Language Processing)	LLM (Large Language Model)
Definition	A **field of AI** that focuses on understanding and generating human language.	A **type of NLP model** trained on massive text data to understand and generate language.
Scope	Broad: Includes tasks like translation, sentiment analysis, question answering, etc.	Specific: A very powerful model that can perform many NLP tasks with minimal tuning.
Examples	SpaCy, NLTK, TextBlob, basic BERT models	GPT-4, ChatGPT, LLaMA, Claude, Gemini
Size	Can be small and task-specific	Typically very large (billions of parameters)
Training Data	Often smaller and task-focused	Trained on huge datasets (e.g., the whole internet)
Flexibility	Often needs task-specific design	Can handle multiple tasks with minimal adjustment

NLP vs LLM

- **NLP is like a Swiss Army knife** — lots of little tools for specific tasks.
- **LLM is like Iron Man** — one powerful suit that can do it all (but needs a lot of energy ⚡).

Alphabet Soup of AI: What All Those Acronyms Mean

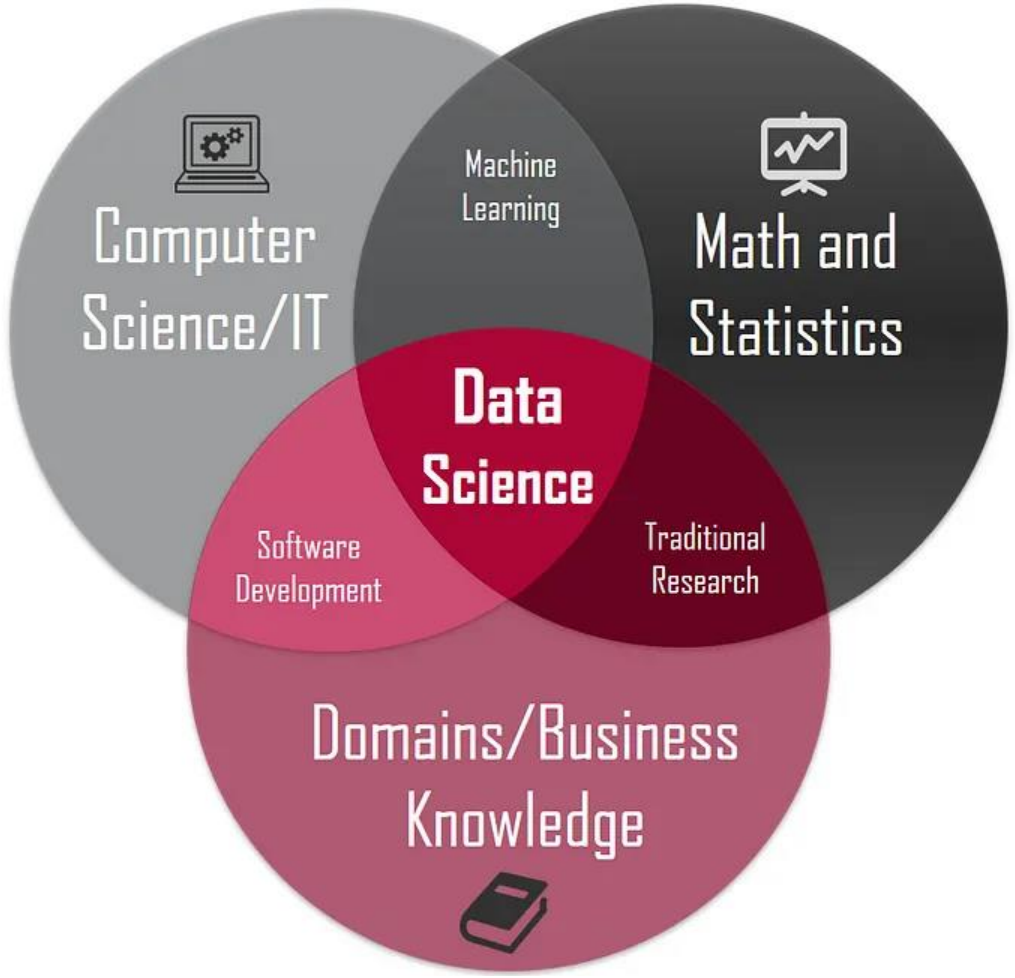
Term	Description	Example
AI (Artificial Intelligence)	The science of building machines that can simulate human intelligence.	Siri, Alexa, website chatbots
ML (Machine Learning)	A subfield of AI where systems learn patterns from data to make decisions without explicit programming.	Gmail Smart Reply
NN (Neural Networks)	Algorithms inspired by the human brain, made up of layers of nodes (neurons).	Pattern recognition, signal processing
DL (Deep Learning)	Neural networks with many (3+) hidden layers capable of solving complex tasks.	Facial recognition, Google Translate
GenAI (Generative AI)	A form of DL that generates new content like images, music, or text from learned patterns.	DALL·E for image generation, ChatGPT for text
LLM (Large Language Model)	A specialized GenAI model trained on large text datasets to generate human-like language.	ChatGPT, Google Gemini
NLP (Natural Language Processing)	A branch of AI that helps machines understand, interpret, and respond to human language.	Translation apps, sentiment analysis, spam filters

Data Science

Data Science is a **multidisciplinary field** that combines:

- **AI and ML**
- **Statistics**
- **Mathematics (including algebra and linear algebra)**
- **Domain knowledge**

To **extract insights from data** and **build intelligent applications**.



Key Takeaways: From AI to Deep Learning

1. Artificial Intelligence (AI) Overview
AI is the broad field focused on creating machines that can simulate human intelligence.
Machine Learning (ML) is a subset of AI that enables systems to learn from data and make predictions or decisions without being explicitly programmed.

2. Machine Learning (ML) Breakdown
Supervised Learning: Learns from labeled data
    Examples:
        Classification – Predicts categories (e.g., spam detection)
        Regression – Predicts continuous values (e.g., house prices)
Unsupervised Learning: Finds hidden patterns in unlabeled data
    Example: **Clustering** – Groups similar data (e.g., customer segments)
Reinforcement Learning: Learns optimal actions through trial and error using feedback (rewards and penalties)
    Example: Robot navigation, game playing (e.g., AlphaGo)

3. Core ML Concepts & Algorithms
K-Nearest Neighbors (KNN): A simple algorithm used for both classification and regression based on similarity to neighbors.
ML Success Depends on Data: Clean, labeled, and relevant data is essential for training effective models.

4. Neural Networks (NN) and Deep Learning (DL)
Neural Networks are the foundation of Deep Learning, consisting of layers of interconnected "neurons" that detect patterns.
Activation Functions like ReLU, Sigmoid, and Tanh enable networks to learn non-linear relationships.
Bias and Weights are internal parameters adjusted during training to improve accuracy.
Deep Learning involves neural networks with many layers, enabling state-of-the-art performance in:
    Image recognition
    Natural Language Processing (NLP)
    Speech recognition

5. Generative AI (GenAI) and LLMs
Generative AI creates new content (text, images, music) using learned patterns.
    Examples: ChatGPT, DALL·E
Large Language Models (LLMs) are powerful GenAI models trained on vast text data to generate human-like responses.
Natural Language Processing (NLP) helps machines understand, process, and generate human language.
    Examples: Language translation, sentiment analysis, spam filters

Thank you!