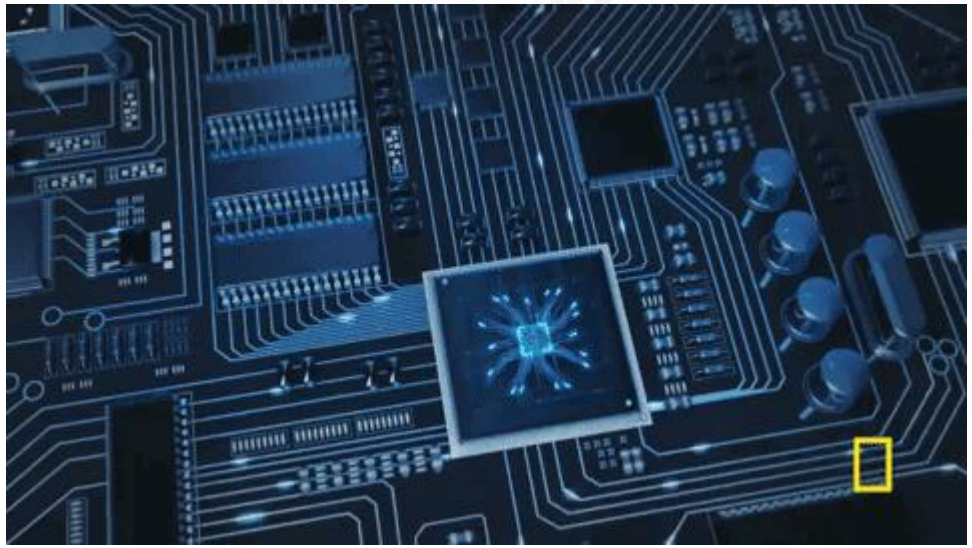
ATME College of Engineering

Department of EEE
Emitting Elite Energy

BEE714D
Big Data Analytics in Power Systems
Module-1: Introduction



Presented by,

Mr.Shreeshayana R

Assistant Professor

Electrical and Electronics Engineering

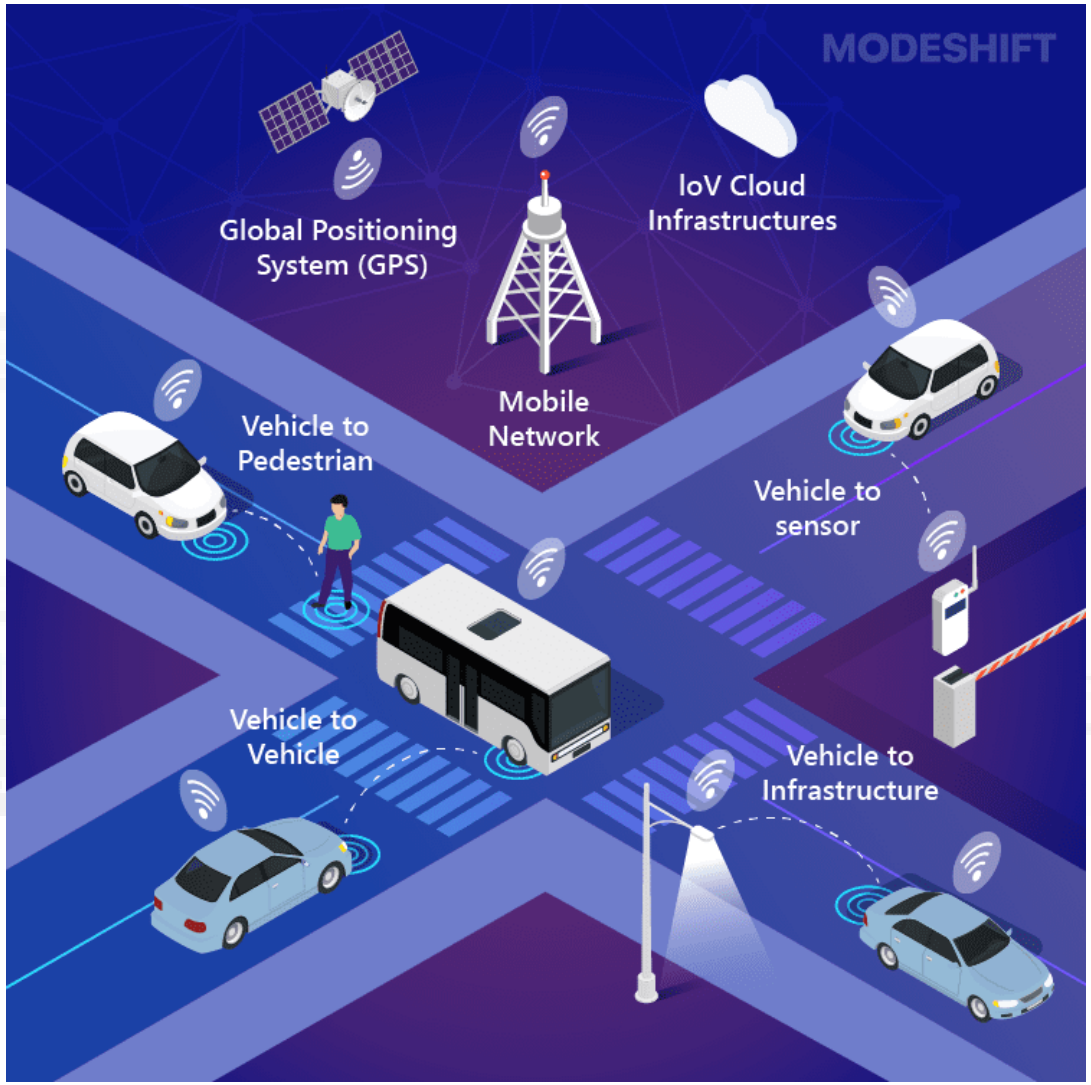
ATME College of Engineering, Mysuru

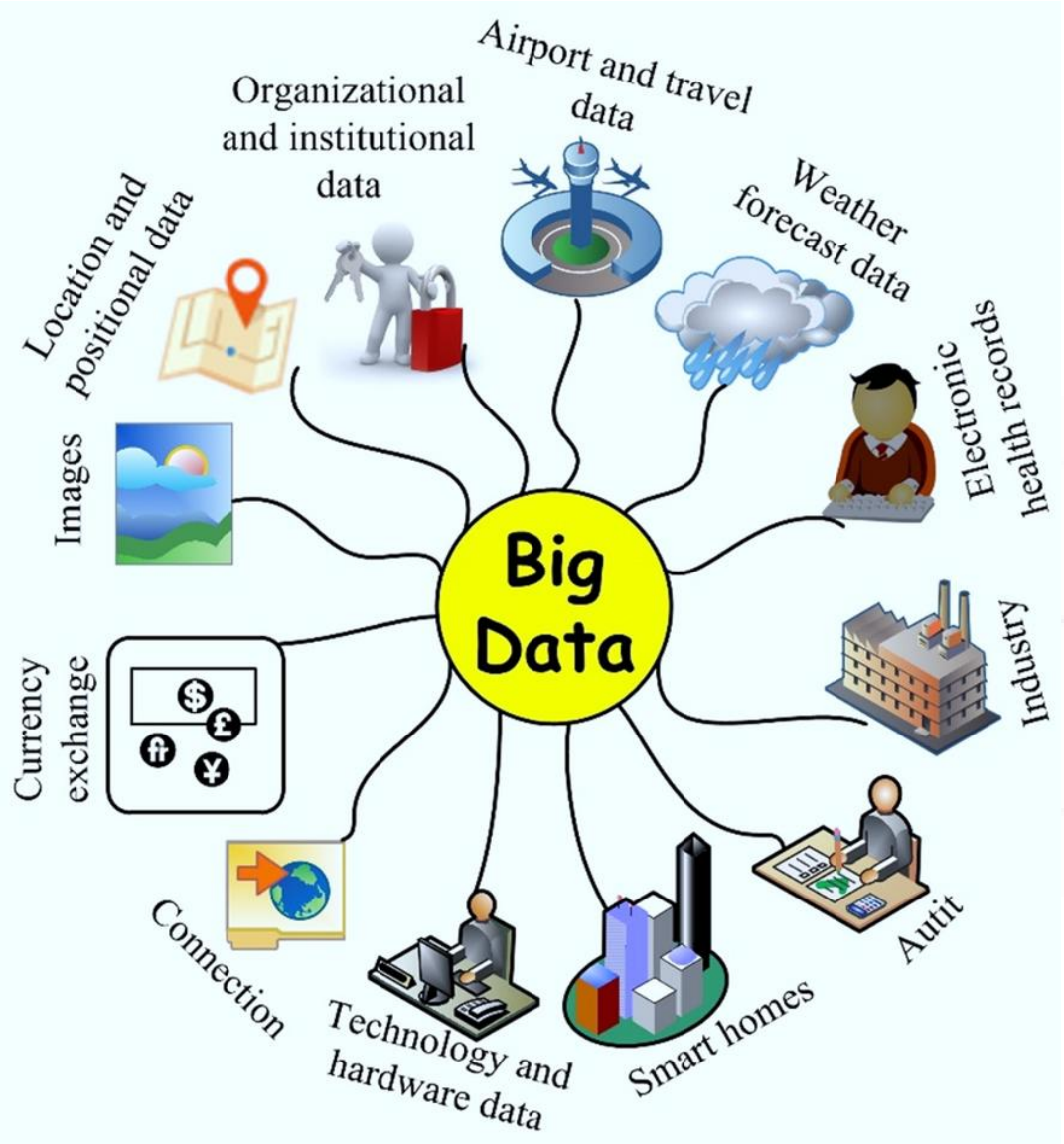
SESSION-1

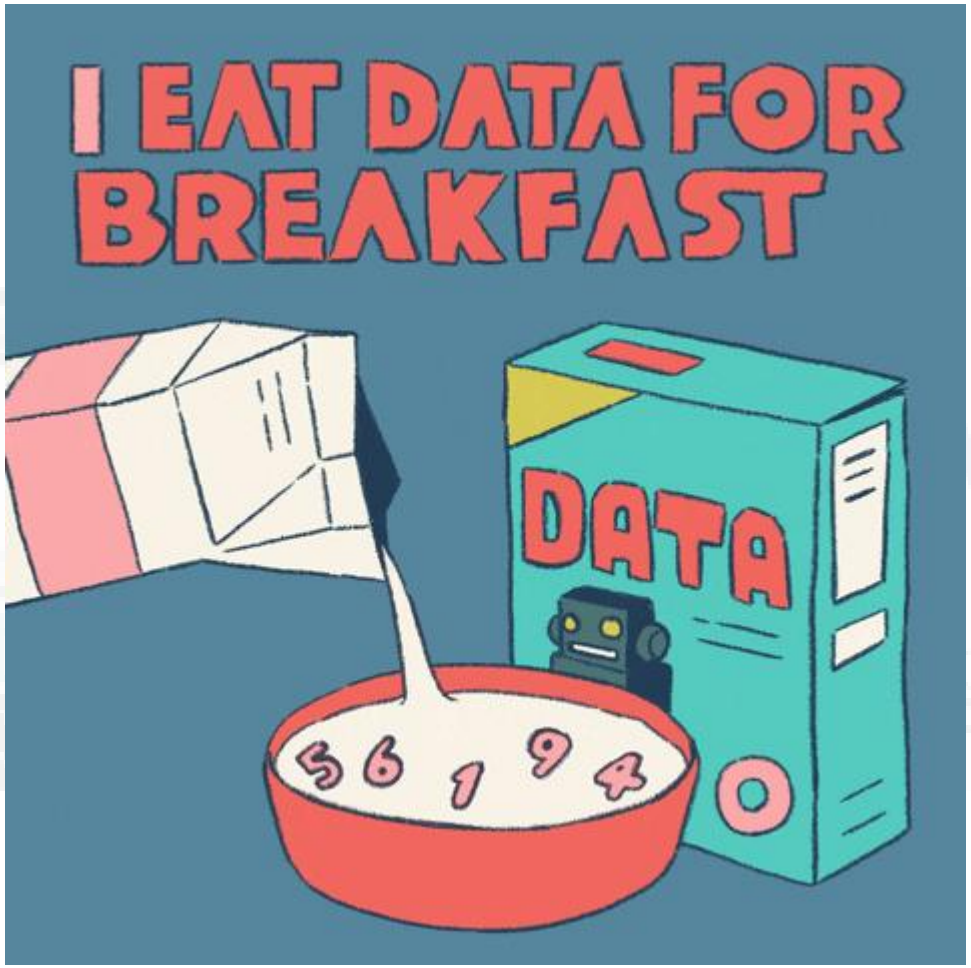
Big Data??

Instagram
YouTube
Data
Predictions

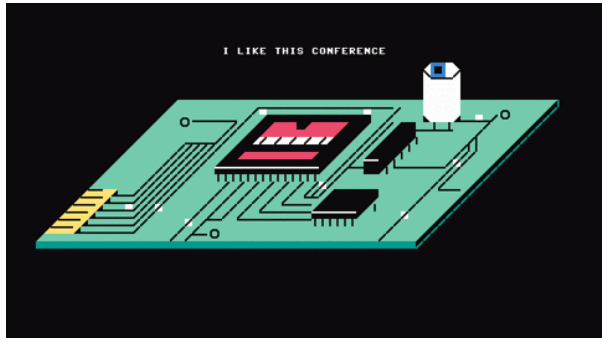
SMART CITY
Ubermetrics


MODESHIFT
Global Positioning System (GPS)
Mobile Network
IoV Cloud Infrastructures
Vehicle to Pedestrian
Vehicle to sensor
Vehicle to Vehicle
Vehicle to Infrastructure


Big Data
Airport and travel data
Organizational and institutional data
Location and positional data
Weather forecast data
Electronic health records
Images
Industry
Currency exchange
Connection
Technology and hardware data
Smart homes
Autit


I EAT DATA FOR BREAKFAST
DATA
5 6 1 9 4 0

ATME College of Engineering



Course Overview

- **Course Code**: BEE714D
- **Course Title**: Big Data Analytics in Power Systems
- **Type:** Professional Elective
- **Prerequisite**: Power System Analysis-I
- **Contact Hours**: 40 Hours

Course Objectives

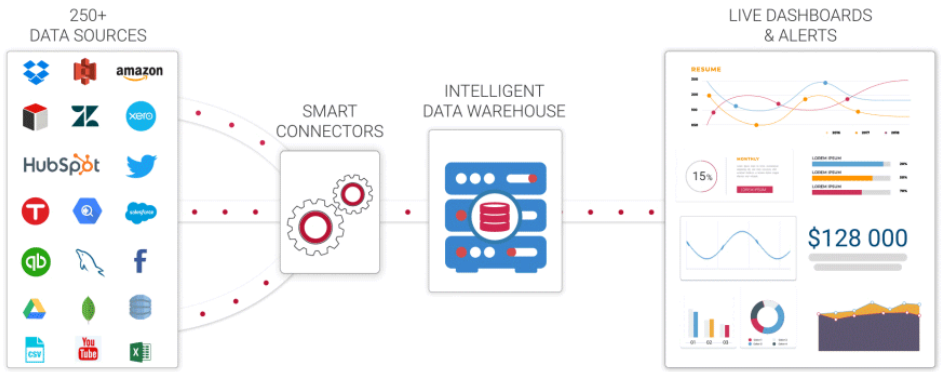
1. To define big data and to explain big data application and analytics to power systems.
2. To explain the role of big data in smart grid communications and optimization of big data in electric power systems.
3. To explain security methods for the infrastructure communication and data mining methods for theft detection in power systems.
4. To explain the application of unit commitment method in the control of smart grid.

- To explain protection algorithm for transformer based on data pattern recognition

Module 1: Introduction

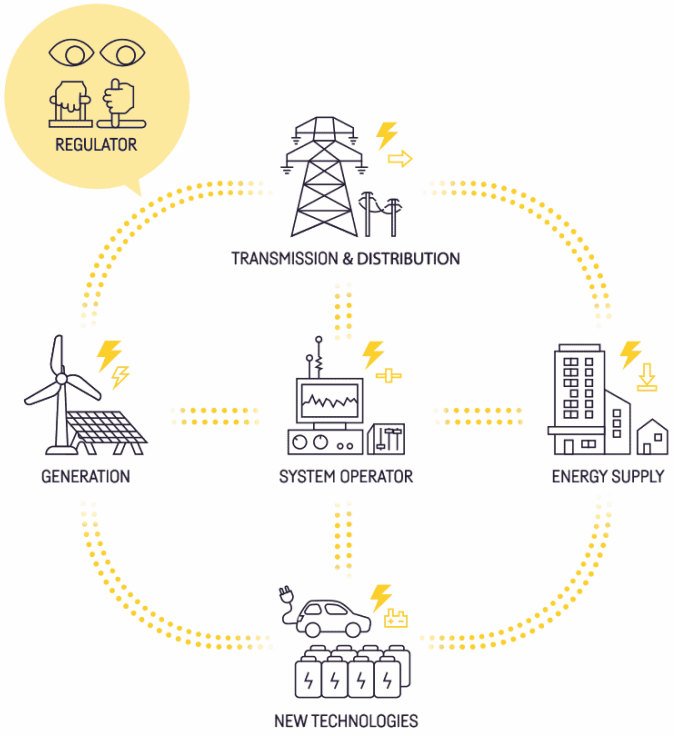
- **Introduction:** Big Data, Future Power Systems.
- **Big Data Application and Analytics in a Large - Scale Power System**: Introduction, General Applications of Big Data, Algorithms for Processing Big Data, Application of Big Data in Power Systems.

Delivery Plan: Week-1 to Week-3



Module 2: Role of Big Data in Smart Grid Communications

- **Role of Big Data in Smart Grid Communications:** Introduction, The Grid Modernization, The Grid Interconnection with the Internet of Things, Data Traffic Pattern in a Smart Grid Environment, The Massive Flow of Information in a Smart Scenario ,The Volume of Generated Data in a Smart Distribution System: A Case of Study.
- **Big Data Optimization in Electric Power Systems:** Introduction, Background, Scientometric Analysis of Big Data, Big Data and Power Systems, Optimization Techniques Used in the Big Data Analysis.



Delivery Plan: Week-4 to Week-5

IA

IA	Module	COs
IA-1	Module-1	CO-1
	Module-2 a	CO-1
IA-2	Module-3	CO-3
	Module-2 b	CO-2
IA-3	Module-4	CO-4
	Module-5	CO-4

Module 3:

- **Security Methods for Critical Infrastructure Communications:** Introduction, Effects of Successful Communication System Threats, General Communication System Operations, Industrial Control Networks and Operations, High-Level Communication System Threats, **Cyber Threats and Security.**
- **Data - Mining Methods for Electricity Theft Detection:** Introduction, Transmission and Distribution System Losses, Electricity Theft Methods, Data Mining and Electricity Theft, Issues and Directions in Electricity Theft-Related Data-Mining Research.

Module 4:

- **Unit Commitment Control of Smart Grids:** Introduction, Renewable Energy Resources, The Unit Commitment Problem, A Multi-agent Architecture, Illustrative Example

Module 5

Transformer Differential Protection Algorithm Based on Data Pattern Recognition: Big Data and Power System Protection, Methods for Differential Protection Blocking, Principal Component Analysis, Curvilinear Component Analysis (CCA), PCA Applied to Discriminate Between Inrush and Fault, Currents in Transformers, Application of the CCA as a Base for a Differential Protection System Under Study, Results.

Text Books

Big Data Analytics in Future Power Systems, Ahmed F. Zobaa and Trevor J. Bihl, CRC Press 2019. 2019.

List of Additional Reference Books/URLs, Text Books, Notes, Multimedia Content, etc

1. **Big Data Analytics for Power Systems – Big Data Analytics in Power Systems**
2. **Application of Big-Data Analytics in Power System Protection-Lec-37: Application of Big-Data Analytics in Power System Protection**

Activity-Based Learning

Activity Based Learning (Suggested Activities in Class)/ Practical Based learning:

<u>Activity Assignment 1: Case Study on Big Data Applications in Power Systems</u>

Objective:

To enable students to **identify and understand** real-world applications of Big Data in large-scale power systems.

1. 17 Groups (4 members in one group). Each group will **select a case study** on Big Data applications in the following areas:

- Smart grid monitoring and optimization
- Renewable energy integration
- Power demand forecasting
- Fault detection and predictive maintenance

1. Prepare a **3–4 page report** including:

- **Problem addressed** in the power system
- **Big Data tools/techniques** used
- **Key outcomes** and **benefits**

1. Submit the report and **present in 5 minutes per group**.

Bloom's Level: L1 – Remembering, L2 – Understanding

Expected Outcome:

Students will **relate theoretical concepts** to **practical applications** in real-world power systems.

Activity Assignment 2: Data Analysis & Algorithm Simulation for Smart Grid

Objective:

To **apply Big Data algorithms** to understand data flow and optimization in smart grid systems.

Instructions:

1. 17 Groups
2. Provide **sample power system datasets** (load demand, voltage, current, and weather data).
3. Each group will:

- Identify **data traffic patterns** and **massive data flow** in a smart grid.
- Use **Python/MATLAB/Excel** to demonstrate a **simple algorithm** such as:
    - Load forecasting using **moving average**
    - Fault pattern detection using **correlation or clustering**

1. Submit:

- **Code/Excel sheet with analysis**
- **One-page summary of findings**
- Submit the report and **present in 5 minutes per group**.

Bloom's Level: L2 – Understanding, L3 – Applying

Expected Outcome:

Students will **visualize data flow**, **analyze trends**, and **apply algorithms** for power system optimization.

Activity Assignment 3: Security Threat & Electricity Theft Detection Analysis

Objective:

To **evaluate communication threats and electricity theft detection** using data-mining approaches.

Instructions:

1. 17 Groups
2. Each group will perform:

- **Research on one cyber threat** to critical infrastructure communications.
- **Study an electricity theft case** and identify **data-mining methods** used for detection (Decision Tree, Clustering, etc.).

1. Prepare a **Poster or Infographic** including:

- Type of cyber threat / theft method
- Impact on the grid
- Suggested **preventive measures**

1. Conduct a **class exhibition** where each group explains their poster in **3 minutes**.
2. Submit the report and **present in 5 minutes per group**.

Bloom's Level: L2 – Understanding, L3 – Applying

Expected Outcome:

Students will **analyze vulnerabilities**, **propose mitigation methods**, and **connect theory to practical security challenges** in power systems.

Course Outcomes & Bloom's Taxonomy

CO-1: Interpret the role of big data and machine-learning methods applicable to power systems and in particular to Smart Grid communications. [L2]

CO-2: Apply optimization methods which are suitable for big data models in power systems. [L3]

CO-3: Identify various cyber security issues, electricity theft detection and mitigation that exist in IoT-enable future power systems. [L3]

CO-4: Identify renewable energy planning concerns associated with planned future power systems that have high renewable penetration. [L3]

Assessment Structure (CIE & SEE)

- CIE: 50%
- SEE: 50% (University Exam, 3-hour duration)
- Minimum passing marks: 40% (40 out of 100)

Course Code:	BEE714D	TITLE: Big Data Analytics in Power Systems								Faculty Member: SHREESHAYANA R		
List of Course Outcomes	Program Outcomes											
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO-1	2	-	-	-	2	-	-	-	2	-	-	2
CO-2	2	2	-	-	2	-	-	-	2	-	-	2
CO-3	2	2	-	-	2	-	-	-	2	-	-	2
CO-4	2	2	-	-	2	-	-	-	2	-	-	2

Course Code:	BEE714D	TITLE: Big Data Analytics in Power Systems	Faculty Member: SHREESHAYANA R
List of Course Outcomes	Program Specific Outcomes		
	PSO1		PSO2
CO-1	2		
CO-2	2		-
CO-3	2		-
CO-4	2		-

Module-1
Introduction

SYLLABUS

1.1 Introduction: Big Data,

1.2 Future Power Systems.

1.3 **Big Data Application and Analytics in a Large - Scale Power System**: Introduction

1.4 General Applications of Big Data

1.5 Algorithms for Processing Big Data

1.6 Application of Big Data in Power Systems.

1.1 Introduction: Big Data

•**Definition** – **Big Data** refers to extremely **large datasets** that cannot be processed efficiently using traditional data processing techniques.

•**Characteristics (5 Vs)** – Volume (size), Velocity (speed), Variety (different types), Veracity (quality/reliability), and Value (usefulness).

•**Sources** – Generated from social media, IoT devices, sensors, business transactions, healthcare, and industrial machines.

•**Challenges** – Data storage, processing, real-time analysis, and maintaining data security.

1.1 Introduction: Big Data

- **Technologies Used** – Hadoop, Spark, NoSQL databases, cloud computing platforms.

- **Applications** – Power systems, healthcare, finance, e-commerce, weather forecasting, and smart cities.

Question 1:

Which of the following is **NOT** a characteristic of Big Data?

a) Volume

b) Velocity

c) Variety

d) Voltage

Question 2:

Which of the following is a **major source of Big Data**?

a) Social media and IoT devices

b) Chalkboard writings

c) Handwritten notes

d) Telephone directories

Question 3:

Which technology is commonly used to **store and process Big Data**?

a) MS Excel

b) Hadoop

c) PowerPoint

d) Photoshop

Question 4:

Which of the following describes the "**Velocity**" characteristic of Big Data?

a) Data comes in different types and formats

b) Data is generated at high speed

c) Data needs high electrical energy

d) Data comes only from video sources

General Information

Installed Capacity & Generation Mix

•As of **June 30, 2025**, India's total installed power capacity reached approximately **484.8 GW**, with **242 GW** from fossil (coal, gas, diesel) and **184.6 GW** from renewables + hydro (~50%) .

•By **March 31, 2025**, installed capacity stood at **475.2 GW**, corroborating ~49% non-fossil share (226.9 GW renewables + 8.8 GW nuclear) AffairsCloud.

•As of January 2025, installed capacity breakdown included:

- Coal: 220.49 GW (≈47.3%)
- Solar: 100.33 GW (21.5%)
- Wind: 48.37 GW (10.4%)
- Hydro (large): 46.97 GW (10.1%)
- Nuclear: 8.18 GW (1.8%)

General Information

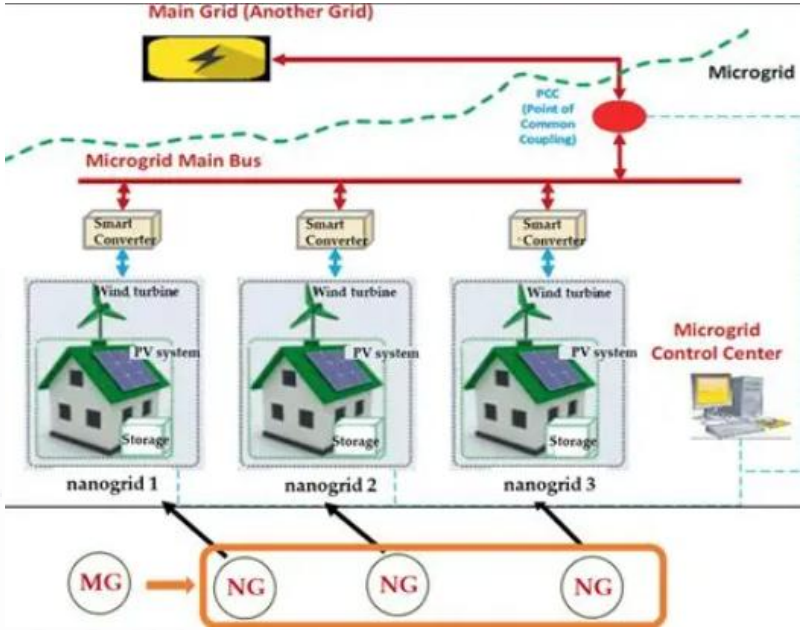
Grid Infrastructure (Transmission & Distribution)

•Transmission lines: about **494,000 circuit-km** as of March 2025, with substation capacity ~**1.337 million MVA**

India's power grid is undergoing rapid transformation—**with near-50%** of installed capacity now **non-fossil, large renewable additions in 2025**, and expanding grid infrastructure.

However, critical challenges remain: achieving grid stability amid high intermittent sources, addressing stranded RE projects, and accelerating energy storage to meet ambitious transition targets while balancing energy security needs.

1.2 Future Power Systems.

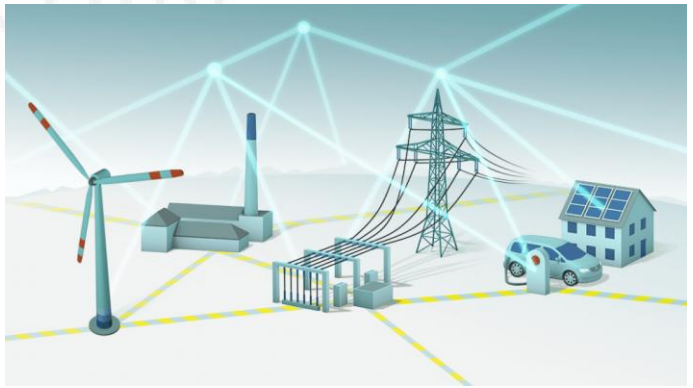


Characteristics of Future Power Systems

- Increased **decentralization** with microgrids and nanogrids.
- **Expanded communication and monitoring** capabilities.
- **Wider variety of energy sources**, including more renewables.

Research Thrusts in Future Power Systems

- Expansion of the **Smart Grid** to improve monitoring and control.
- **Internet of Things (IoT)** integration for device-level communication.
- Increased **renewable energy penetration**.
- **Microgrid and nanogrid deployment** for local resiliency.



What is a microgrid?

Nanogrid Electricity in India | World Stories

Smart Grid Implications

•Enables **data collection** for theft detection and grid optimization.
•Generates **big data challenges** due to the volume and velocity of collected data

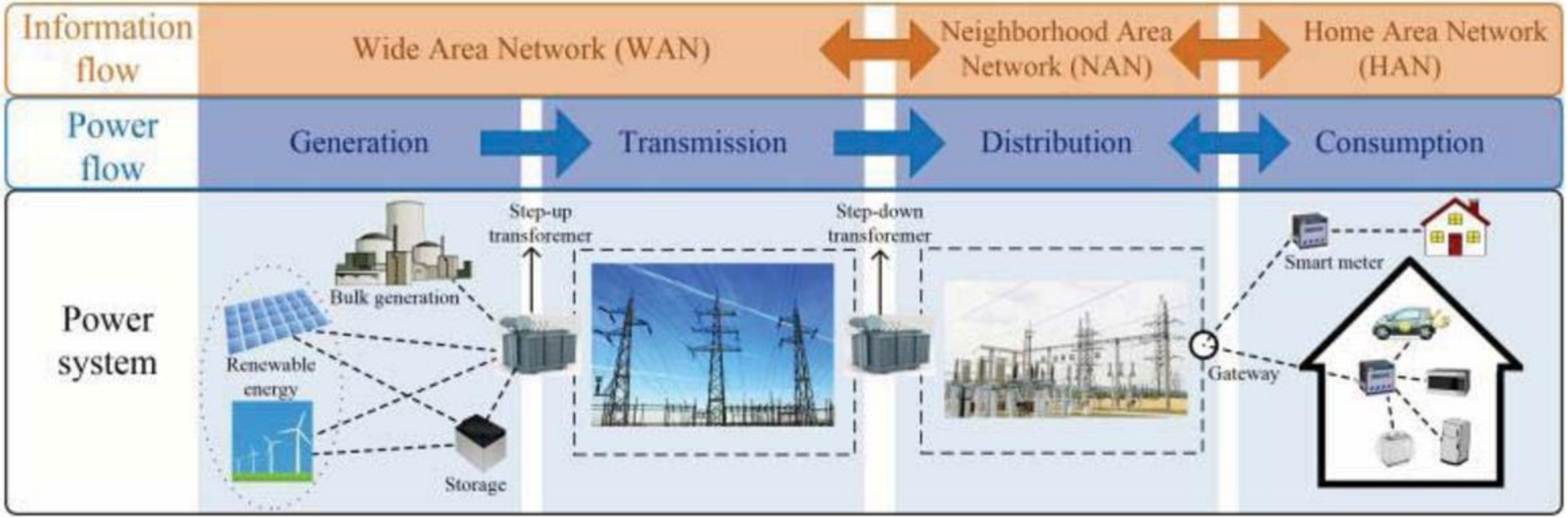
IoT-enabled Power Grid

•Allows **communication with all devices** in the network.
•Provides **real-time monitoring of critical infrastructure (CI)**.
•Introduces **big data and cybersecurity challenges**.

Microgrids and Renewables

•Offer **local resiliency** but create **uncertainty for larger grid planning**.
•Renewable energy sources add **availability and reliability challenges**.

Transforming Solar Nanogrid Can Power Homes (And Charge a Tesla)


Information flow
Wide Area Network (WAN)
Neighborhood Area Network (NAN)
Home Area Network (HAN)
Power flow
Generation
Transmission
Distribution
Consumption
Power system
Bulk generation
Step-up transforemer
Step-down transforemer
Renewable energy
Storage
Gateway
Smart meter

1. Which feature is a key characteristic of future power systems?

A) Reduced communication capabilities

B) Centralized power generation only

C) Increased decentralization with microgrids and nanogrids

D) Exclusive reliance on fossil fuels

2. What is a primary research thrust in future power systems?

A) Limiting renewable energy use

B) Reducing smart grid communication

C) Expanding the smart grid for better monitoring and control

D) Removing IoT from the power grid

3. What is one major implication of implementing a smart grid?

A) Reduced data generation

B) Improved theft detection and grid optimization

C) Decreased need for monitoring

D) Elimination of cybersecurity risks

4. What challenge arises from integrating microgrids and renewable energy sources?

A) Guaranteed power stability at all times

B) Simplified large grid planning

C) Increased uncertainty in availability and reliability

D) Elimination of the need for local resiliency

1.3 **Big Data Application and Analytics in a Large - Scale Power System**: Introduction

Big Data Analysis Methods

•Relies heavily on **machine learning techniques**.

•Machine learning is associated with **pattern recognition, statistics, and data mining**.

Emergence of Advanced Techniques

•**Deep learning** (large-scale neural networks) is increasingly used.

•Capable of handling the **size and complexity** of big data.

Applications in Power Systems

•Initially successful in **image recognition** tasks.

•Now being **applied to big data analysis in power systems** for enhanced insights and decision-making.

Big Data Analysis Methods in Power Systems

1. Role of Machine Learning (ML)

1. ML algorithms are essential for handling large-scale, complex datasets generated by modern power grids.
2. ML is associated with **pattern recognition, statistics, and data mining**, enabling automatic detection of patterns and anomalies in grid operations.

2. Applications in Power Systems

•**Load Forecasting:** Predicting electricity demand for better grid management.
•**Fault Detection and Diagnosis:** Recognizing abnormal patterns in sensor data to prevent outages.
•**Power Theft Detection:** Identifying irregular consumption patterns from smart meter data.
•**Renewable Energy Integration:** Forecasting solar and wind power output to optimize generation and storage.

3. Advanced Methods

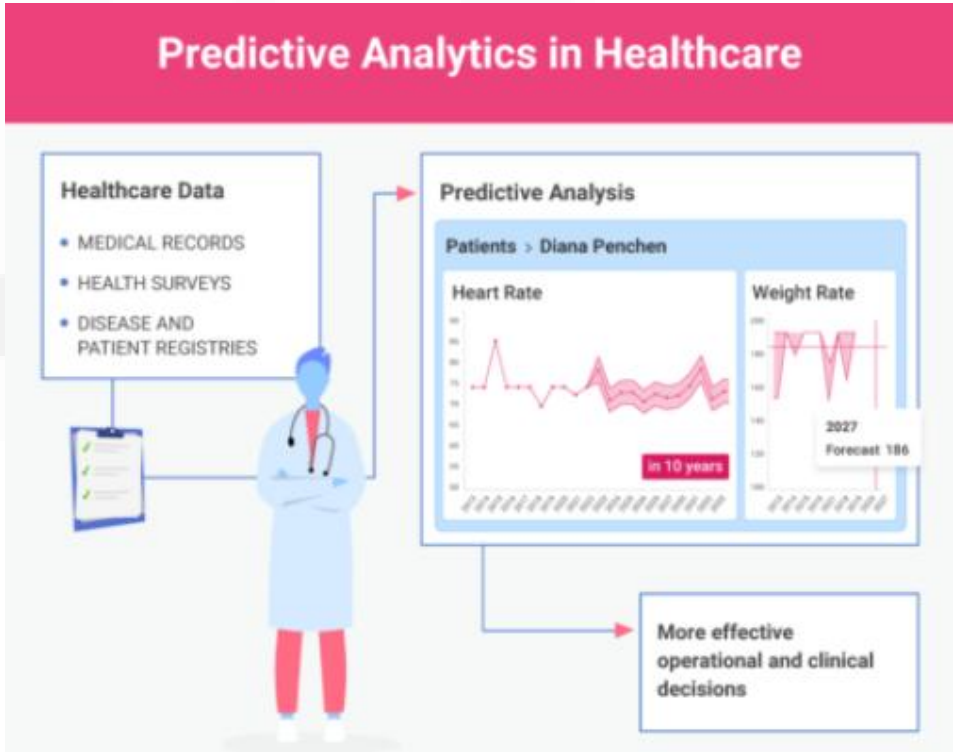
1. **Deep Learning (DL):** Large-scale neural networks analyze high-dimensional, real-time data such as PMU (Phasor Measurement Unit) streams.
2. **Predictive Analytics:** Uses **historical and live data** to make proactive grid management decisions.
3. **Data Mining:** Extracts **hidden trends from vast power system databases** for planning and optimization.

1.4 General Applications of Big Data

Sector	Key Use Cases
Healthcare	Personalized treatment, predictive risk, diagnostics
Marketing	Targeting, personalization, campaign optimization
Transportation	Smart traffic, route optimization, congestion control
Finance	Fraud detection, risk analytics, compliance
Retail	Recommendations, dynamic pricing, inventory management
Government	Public safety, fraud control, social services
Manufacturing	Predictive maintenance, efficiency in operations
Education	Student performance analysis, learning insights
Agriculture	Crop monitoring, satellite insights, yield boost
Cybersecurity	Behavioral threat detection, anomaly identification

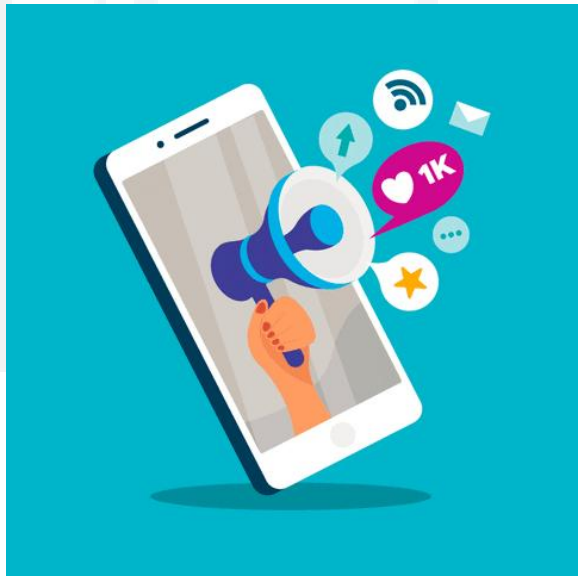
Key Applications of Big Data

•**Healthcare**: Enables personalized medicine, predictive analytics for **clinical risk, automated reporting, and improved diagnostics** using vast and varied medical data



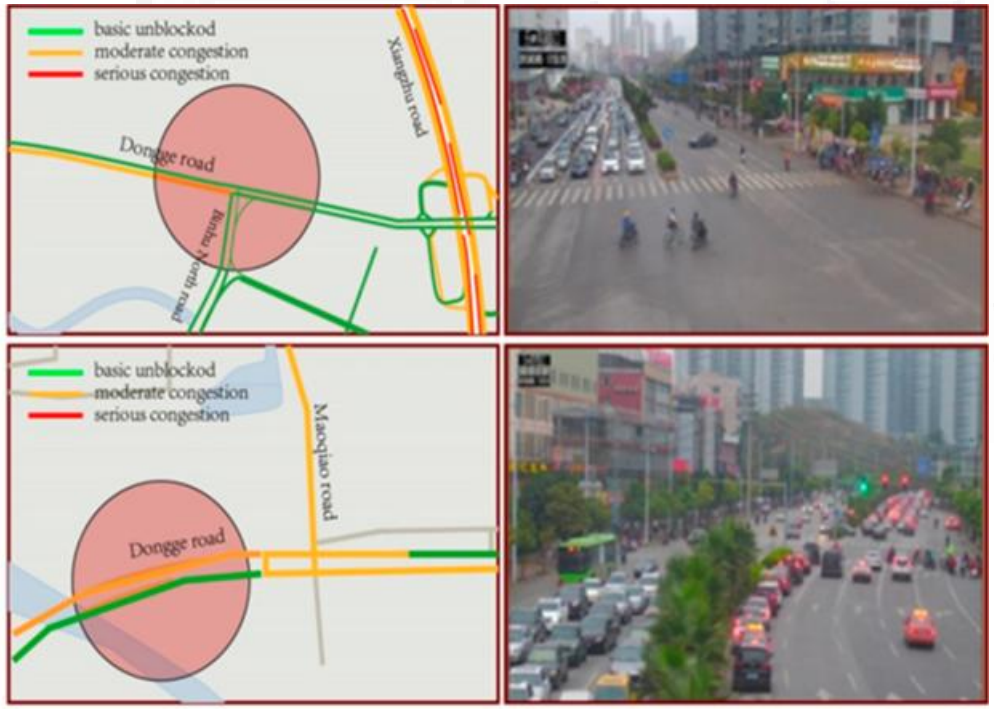
Key Applications of Big Data

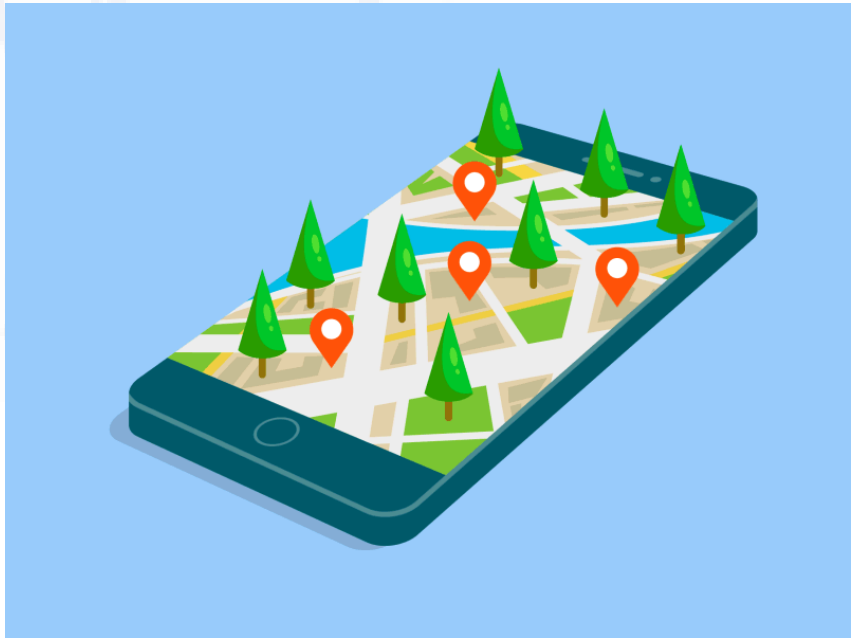
•**Marketing & Advertising**: Supports precise customer targeting and **<u>personalized messaging</u>** by analyzing behavior and preferences at scale.



Key Applications of Big Data

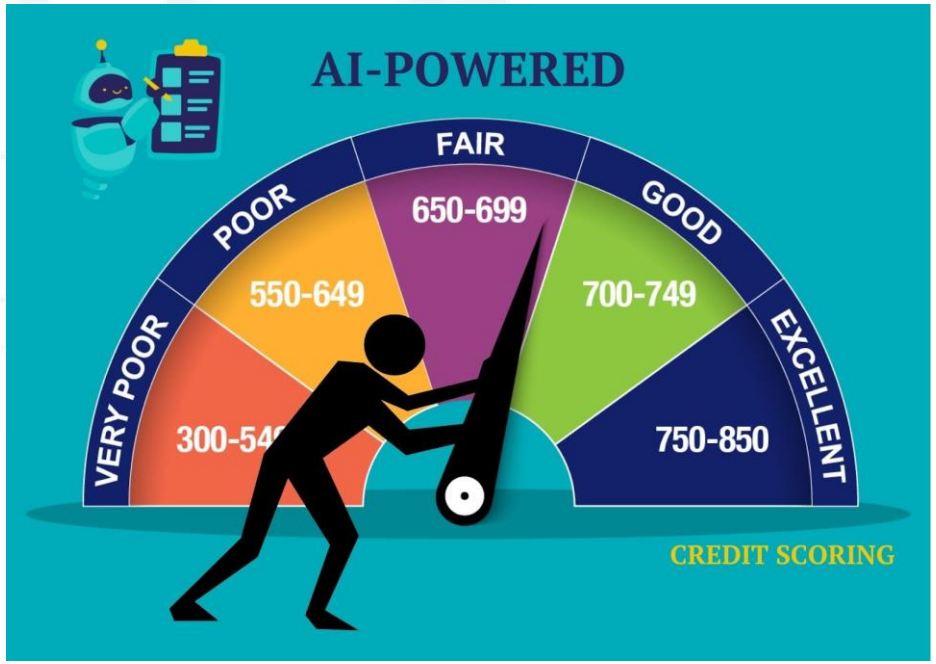
•**Transportation & Smart Cities**: Powers intelligent traffic systems, real-time route planning, congestion prediction, and urban computing solutions like GPS-based path estimation





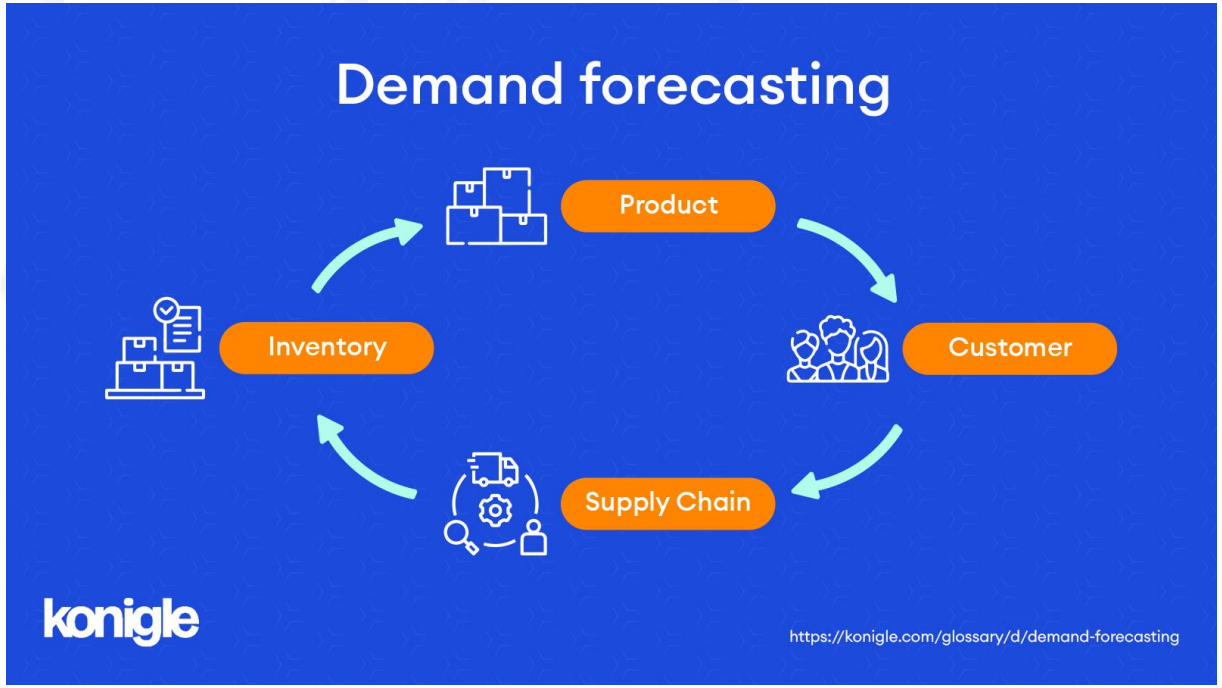
Key Applications of Big Data

•**Finance & Banking:** Used for fraud detection, credit scoring, risk management, portfolio optimization, and regulatory compliance



Key Applications of Big Data

•**Retail & E-commerce**: Enhances customer experience through recommendation systems, inventory optimization, dynamic pricing, and demand forecasting.

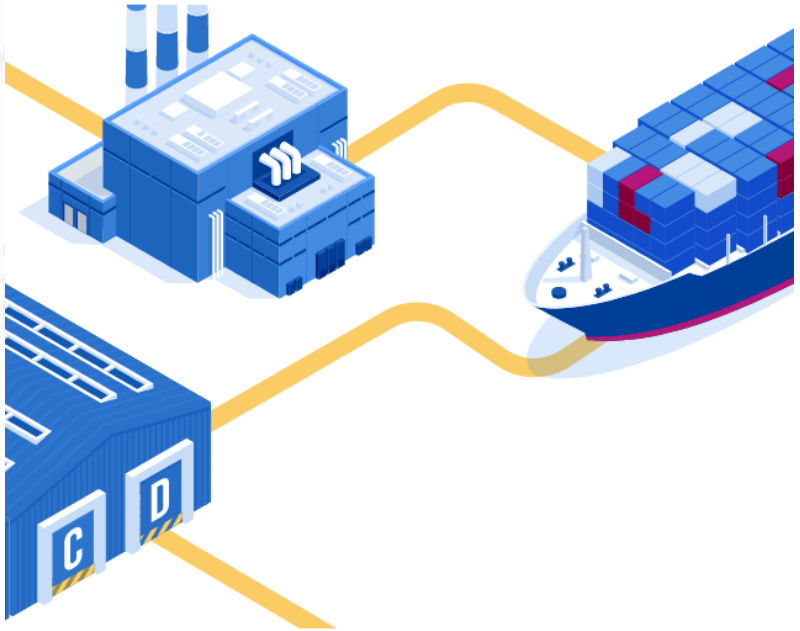


Key Applications of Big Data

•**Government & Public Sector**: Facilitates fraud detection, public health tracking, environmental protection, and efficient public service delivery.

Key Applications of Big Data

•**Manufacturing & Industrial**: Enables predictive maintenance, smart factory operations, real-time analytics, and supply chain visibility.

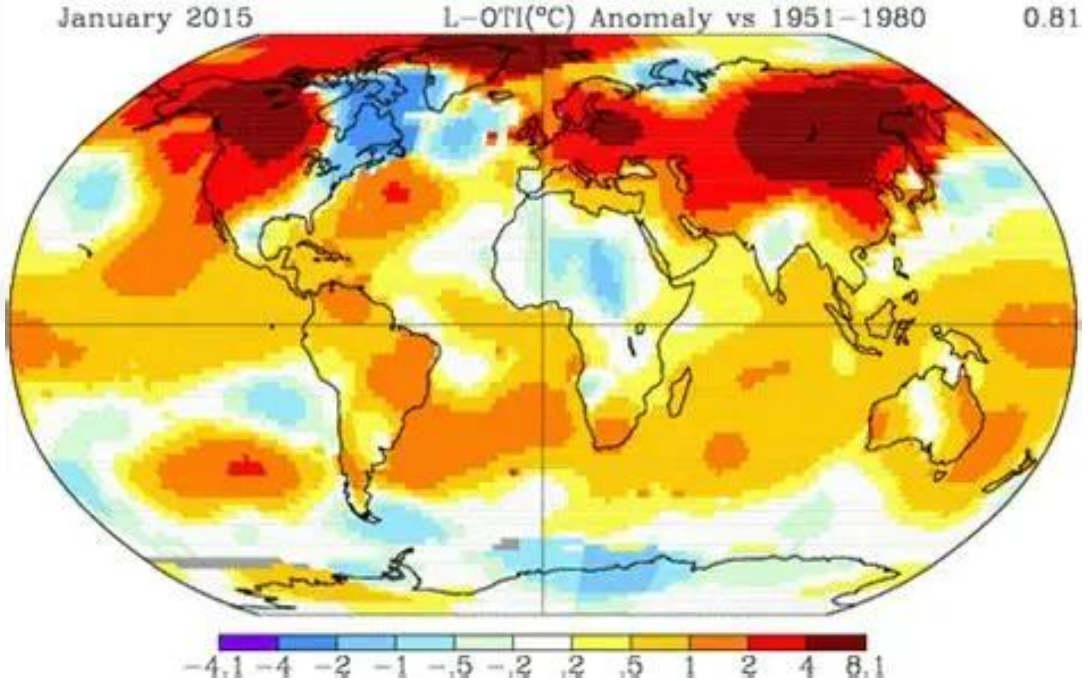


Key Applications of Big Data

•**Education & Learning Analytics**: Helps track student engagement, performance trends, and improve teaching effectiveness

Key Applications of Big Data

•**Climate, Agriculture & Earth Science**: Supports satellite data analysis for precision agriculture, yield forecasting, and climate trend monitoring.



Key Applications of Big Data

•**Cybersecurity & Behavioral Analytics**: Detects anomalies, predicts potential fraud or threats, and analyzes user behavior to secure systems.

Content Beyond Syllabus

Algorithms for Big Data Processing

- **MapReduce** – Splits tasks into smaller chunks (Map), processes them in parallel, and combines results (Reduce).
- **K-Means Clustering** – Groups large datasets into clusters based on similarity.
- **Apriori Algorithm** – Finds frequent itemsets and association rules in large transactional data.
- **PageRank** – Ranks web pages based on link structure, widely used in search engines.
- **Gradient Descent** – Optimizes parameters for large-scale machine learning models.

- **FP-Growth** – Fast pattern mining for identifying frequent itemsets without candidate generation.
- **Support Vector Machines (SVM)** – Classifies data efficiently, especially in high-dimensional spaces.
- **Naive Bayes** – Simple probabilistic classifier for large text datasets and spam filtering.
- **Streaming Algorithms** – Process data on-the-fly without storing the entire dataset (e.g., Count-Min Sketch).
- **Random Forests** – Ensemble method that handles large datasets with high accuracy.

- **MapReduce** – Counting the frequency of each word in 10 million news articles.
- **K-Means Clustering** – Segmenting 5 lakh retail customers into groups based on purchasing behavior.
- **Apriori** – Finding that "Customers who buy bread and milk also buy butter" in a supermarket transaction log.
- **FP-Growth** – Discovering that "Tea and sugar" are bought together in 65% of 1 crore grocery bills.
- **Count-Min Sketch** – Tracking trending hashtags like **#Olympics2024** in real-time from Twitter without storing all tweets.

1. MapReduce

Purpose: Process large datasets in a distributed manner.

Steps:

1. **Input Splitting:** Break the dataset into smaller chunks.
2. **Mapping:** Apply a function to each chunk in parallel.
3. **Shuffling & Sorting:** Group intermediate results by key.
4. **Reducing:** Aggregate values for each key.
5. **Output:** Store final processed results.

Example:

Dataset: List of words in documents.

- **Map:** Count each word occurrence.
- **Reduce:** Sum counts for each word.

Output: Word frequency list.

MapReduce Example in Power Systems: Regional Energy Consumption Analysis

Objective:

Calculate the **total energy consumed** by each household over a day using smart meter data.

Scenario:

Dataset: Smart meter logs from thousands of homes.Each record:Household_ID, Timestamp, Energy_Consumed (kWh)

Step-by-Step Process Using MapReduce

1. Input Splitting

•The large dataset (millions of records) is split into smaller chunks for distributed processing.

2. Map PhaseInput to Mapper:

One line per reading,

e.g.H001, 2025-08-25 01:00, 0.5Mapper

Output (Key, Value):
Key = Household_ID,
Value = Energy_Consumed→

Output:
(H001, 0.5)
(H001, 0.6)
(H002, 0.4)

3. Shuffle & Sort Phase: The framework groups all values with the same Household_ID:(H001, [0.5, 0.6, 0.8, ...])
(H002, [0.4, 0.3, 0.5, ...])

4. Reduce Phase:
The reducer adds all energy values for each household:For H001:Total = 0.5 + 0.6 + 0.8 + ...
For H002:
Total = 0.4 + 0.3 + 0.5 + ...

Final Output

A summary showing **total daily energy consumption** per household, which can be used for:

- Billing
- Load forecasting
- Demand response analysis
- Peak consumption tracking

2. K-Means Clustering

Purpose: Group similar data points into K clusters.

Steps:

1.Select K random centroids from the dataset.

2.Assign each point to the nearest centroid.

3.Calculate new centroids as the mean of points in each cluster.

4.Repeat steps 2–3 until centroids don't change.

Example:

Customer data with age & spending score.

•Algorithm groups customers into 3 segments: low spenders, moderate spenders, high spenders.

3. Apriori Algorithm

Purpose: Discover frequent itemsets and association rules in large transaction datasets.

Steps:

1.Identify all items with support above a threshold.

2.Generate candidate itemsets of size k from frequent itemsets of size k−1.

3.Count support for each candidate.

4.Prune itemsets below the threshold.

5.Repeat until no new itemsets appear.

Example:

Transactions in a supermarket.

•Rule: If {Milk, Bread} is bought, {Butter} is also bought with 70% confidence.

4. FP-Growth Algorithm

Purpose: Faster frequent pattern mining without generating candidates.

Steps:

1.Scan database for frequent items and order them by frequency.

2.Build an FP-tree representing transactions.

3.Extract conditional pattern bases.

4.Recursively mine the FP-tree to find frequent itemsets.

Example:

Transaction dataset shows frequent buying patterns: {Tea, Sugar} appears in 65% of sales.

5. Streaming Algorithm (Count-Min Sketch)

Purpose: Approximate counts in data streams with low memory usage.

Steps:

1. Initialize a 2D array (d × w) with zeros.
2. For each incoming element, hash it using d hash functions.
3. Increment counters in corresponding positions.
4. Estimate the count as the minimum value from all hash outputs.

Example:

Real-time Twitter hashtag counting without storing all tweets.

Big Data Applications in Power Systems

1. Smart Grid Management

•**Real-Time Monitoring:** Collect and analyze data from sensors, smart meters, and substations.

•**Fault Detection & Prediction:** Predict and localize faults in real-time using data patterns.

•**Grid Stability:** Analyze data to maintain voltage and frequency stability.

Example: Using real-time data from PMUs (Phasor Measurement Units) to detect oscillations and prevent blackouts.

Big Data Applications in Power Systems

2. Load Forecasting

•**Short-Term Forecasting:** Uses real-time and historical data to predict demand for hours or days.

•**Long-Term Forecasting:** Helps in planning infrastructure upgrades and generation capacity.

Example: Machine learning models using weather and historical consumption data to predict peak loads.

Big Data Applications in Power Systems

3. Renewable Energy Integration

•**Variability Management:** Analyze weather and solar/wind generation data to improve grid integration.
•**Predictive Maintenance:** Detect equipment failure in wind turbines or solar panels early.

Example: Using satellite data and IoT sensors to forecast solar PV output and adjust grid operations accordingly.

Big Data Applications in Power Systems

4. Energy Theft Detection

•**Anomaly Detection:** Analyze consumption patterns to detect illegal connections or meter tampering.

Example: Comparing smart meter data to expected usage models to identify unusual drops or spikes.

Big Data Applications in Power Systems

5. Asset Management & Predictive Maintenance

•**Condition-Based Maintenance:** Use sensor data from transformers, breakers, etc., to anticipate failures.

•**Asset Lifecycle Management:** Optimize the maintenance schedule and replacement strategies.

Example: Using vibration and temperature data from transformers to predict failure risks.

Big Data Applications in Power Systems

6. Demand Response Optimization

•**Consumer Behavior Analysis:** Analyze usage data to optimize load shedding and time-of-use pricing.

•**Dynamic Pricing Models:** Create real-time pricing based on demand patterns.

Example: Incentivizing users to reduce load during peak hours through automated demand response.

Big Data Applications in Power Systems

7. Market Operation & Energy Trading

•**Price Forecasting:** Use historical market data to forecast electricity prices.

•**Optimization Algorithms:** Maximize profit in bidding and scheduling in deregulated markets.

Example: Using Big Data to model and simulate real-time market scenarios for strategic bidding.

Big Data Applications in Power Systems

8. Electric Vehicle (EV) Integration

- **Charging Pattern Analysis:** Monitor and predict charging behavior to avoid grid stress.
- **V2G (Vehicle-to-Grid):** Manage bi-directional energy flow using real-time data from EVs.

Example: Coordinating EV charging during off-peak hours to balance the grid.

Big Data Applications in Power Systems

Tools & Technologies Used:

- **Data Sources:** Smart Meters, PMUs, SCADA, IoT sensors, Weather Stations
- **Analytics Tools:** Hadoop, Spark, TensorFlow, Python, MATLAB
- **Visualization:** Power BI, Tableau, Grafana
- **Storage:** Cloud (AWS, Azure), NoSQL (MongoDB, Cassandra)

Algorithms as per Prescribed Text Book for Processing Big Data

Algorithms for Processing Big Data

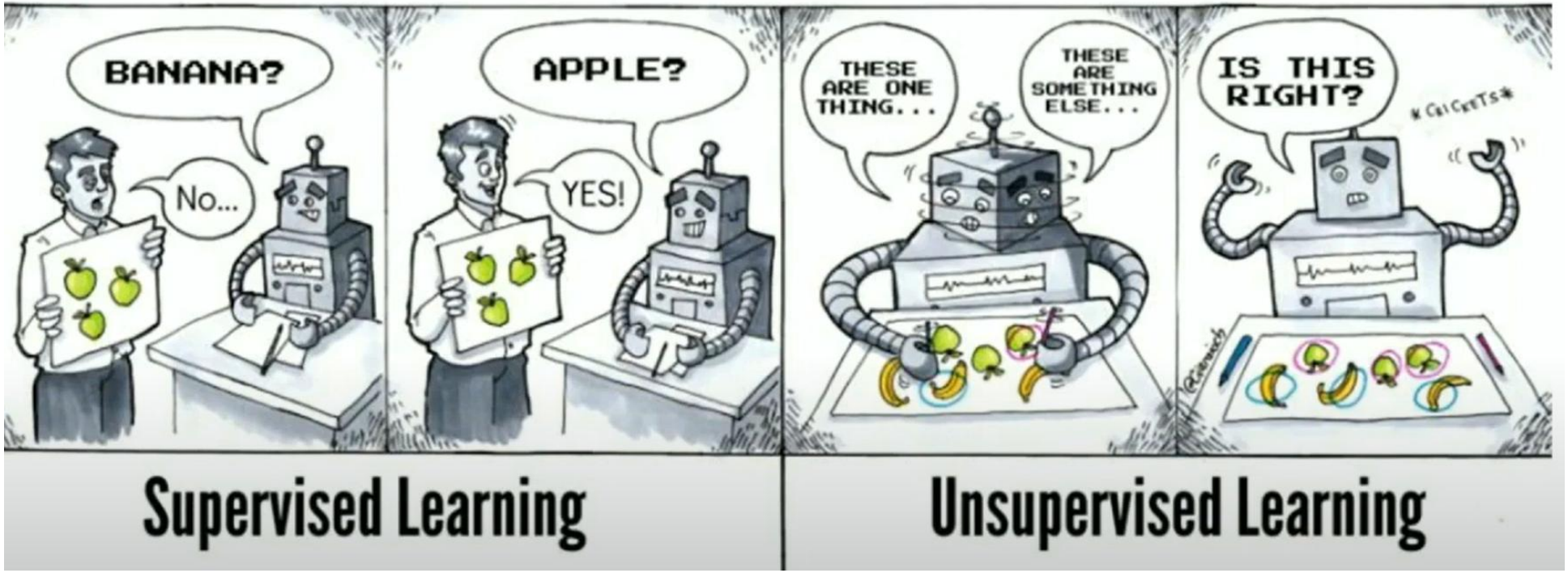
- Algorithms for big data analysis include **machine learning** and **deep learning.**
- **Deep learning** is a **subset of machine learning** but uses advanced methods and architectures.

Machine Learning and Deep Learning Generalities

- **Machine Learning (ML): Uses training data to classify unseen data.**
- **Algorithms define weights for models and update them until convergence.**
- **Two types:**

1. **Supervised Learning:** Uses labelled data (input + output known). Learns mapping rules.
2. **Unsupervised Learning:** Uses unlabelled data. Aims to find structure/patterns in data.

Supervised and Unsupervised Learning



1.5.1 Artificial Neural Network (ANN) Model

- **Neural networks** are effective at understanding complex data patterns.
- They can detect **patterns too complex for humans** or **traditional computer methods.**
- ANN consists of layers of interconnected processing nodes (neurons).
- The **Multi-layer Perceptron (MLP)** is the most commonly used ANN architecture.
- **A typical ANN has three layers:**
  1. **Input Layer – receives data from the environment.**
  2. **Hidden Layer – processes data and connects input to output.**
  3. **Output Layer – delivers the final response or classification.**
- Neurons are represented as circular nodes; connections (with weights) represent data flow between them.
- Weighted connections link the output of one neuron to the input of another.
- Hidden layers help extract complex information for classification.

Real-World Example: Predicting House Prices with ANN

🏠 **Scenario:**

A real estate company wants to **predict house prices** based on features such as:

- Size of the house (in square meters)
- Number of bedrooms
- Location score
- Age of the house

These features are inputs to an **Artificial Neural Network**, and the **output** is the **predicted house price**.

🧠 How the ANN Works in This Example:

- **Input Layer:** Takes the features:
  - x_1: size
  - x_2: number of bedrooms
  - x_3: location score
  - x_4: age
- **Hidden Layers:** Process the data through weighted connections and activation functions.
- **Output Layer:** Produces a **single predicted value** $\hat{y}$, which is the **estimated price** of the house.

❌ Error (Cost) Function in Use:

The **real price** of the house (known from past data) is y, and the **predicted price** is $\hat{y}$.

To train the model, we need to measure how far off our prediction is. A common choice:

▶️ Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

📃 Example Data:

House	Actual Price ($)	Predicted Price ($)
1	300,000	310,000
2	450,000	460,000
3	500,000	480,000

$$\text{MSE} = \frac{1}{3}[(300000 - 310000)^2 + (450000 - 460000)^2 + (500000 - 480000)^2]$$

$$= \frac{1}{3}[(-10000)^2 + (-10000)^2 + (20000)^2]$$

$$= \frac{1}{3}[100000000 + 100000000 + 400000000] = \frac{600000000}{3} = 200,000,000$$

📉 The model uses this **MSE value (200 million)** to understand how far off it is, and adjusts weights to minimize this during training.

www.atme.in

Summary:

- **Real-world problem:** Predict house prices.
- **Model:** ANN with features like size, rooms, location, etc.
- **Error function used:** Mean Squared Error (MSE).
- **Goal:** Minimize MSE so predictions get closer to real prices.

Support Vector Machine (SVM)

- **SVM** is an algorithm like ANN used for classification and regression tasks.
- It identifies a **function mapping** that splits input data into separate classes.
- The goal is to:
  - Map data into a **new space** where it becomes **linearly separable**.
  - Perform **linear classification** in this new space.

Imagine you want to classify emails as Spam or Not Spam based on the number of certain keywords.

- **Step 1: Data** — You have emails with features like the count of suspicious words, and labels (Spam or Not Spam).
- **Step 2: Map Features** — Sometimes the data can't be separated by a simple line, so features are transformed to a new space.
- **Step 3: Find the best boundary** — SVM finds the line (or hyperplane) that best separates Spam emails from Not Spam emails.
- **Step 4: Margin Maximization** — It chooses the boundary that keeps the biggest gap (margin) between the two groups.

Imagine you want to classify emails as Spam or Not Spam based on the number of certain keywords.

- **Step 5: Handle errors** — Some emails might be misclassified to keep the margin large enough (soft margin).
- **Step 6: Kernel Trick** — If emails aren't separable in original features, SVM uses a kernel (like RBF) to separate them in a higher-dimensional space.
- **Step 7: Final Model** — The decision boundary is formed mainly by a few important emails called "support vectors."

Decision-Tree Classifier

1. Overview

- A Decision Tree is used for multi-stage decision-making.
- It follows a recursive top-down approach.

2. Structure

- Root node: The starting point of the tree (no incoming edges).
- Internal nodes: Perform recursive partitioning of input space using decision rules.
- Leaf nodes: Terminal nodes representing the final class labels (no outgoing edges).

3. Working

- The input space is partitioned into **two or more subclasses** at each internal node.
- **Recursion continues** until all data points are classified into appropriate leaf nodes.
- Each **leaf node** is assigned one class that best fits the data reaching that point.

CART Algorithm

- **CART**: Binary decision tree algorithm for **classification** and **regression**.
- **Splitting**: Chooses features that maximize **purity** (e.g., Gini Index, MSE).
- **Classification**: Uses **Gini Impurity** or **Entropy**.
- **Regression**: Uses **variance reduction** or **Mean Squared Error (MSE)**.
- **Tree growth**: Built **recursively**.
- **Pruning**: Applied to reduce **overfitting**.

Example: Classify Animals as Cat or Dog based on Tail Length

Animal	Tail Length	Label
1	Short	Cat
2	Short	Cat
3	Long	Dog
4	Long	Dog
5	Short	Dog

Step 1: Calculate Gini Index for whole data

- Total animals = 5
- Cats = 2, Dogs = 3

$$p_{Cat} = \frac{2}{5} = 0.4, \quad p_{Dog} = \frac{3}{5} = 0.6$$

$$Gini = 1 - (0.4^2 + 0.6^2) = 1 - (0.16 + 0.36) = 0.48$$

Step 2: Split by Tail Length (Short vs Long)

- **Short Tail:** Animals 1, 2, 5

  Cats = 2, Dogs = 1

$$p_{Cat} = \frac{2}{3} \approx 0.67, \quad p_{Dog} = \frac{1}{3} \approx 0.33$$

$$Gini = 1 - (0.67^2 + 0.33^2) = 1 - (0.45 + 0.11) = 0.44$$

- **Long Tail:** Animals 3, 4

  Cats = 0, Dogs = 2

$$p_{Dog} = 1, \quad Gini = 0 \text{ (pure node)}$$

Step 3: Calculate weighted Gini after split

$$\text{Weighted Gini} = \frac{3}{5} \times 0.44 + \frac{2}{5} \times 0 = 0.264$$

Step 4: Calculate Gini Gain

$$GiniGain = \text{Original Gini} - \text{Weighted Gini} = 0.48 - 0.264 = 0.216$$

Conclusion:

- Splitting by tail length reduces impurity by 0.216 (improves classification).
- Decision Tree will split first on **Tail Length**.

Simple takeaway:

Gini Index measures impurity, Gini Gain shows how well a split separates classes, and the tree picks splits with highest gain.

Deep Learning Models

1. Deep learning analyses big data by using complex data representations and abstractions.
2. It is effective for decision-making, information retrieval, and classification, especially with unsupervised data.
3. Deep learning models use stacked layers where each layer performs nonlinear transformations.
4. Data flows hierarchically through these layers, automating feature extraction at each step.
5. More layers enable deeper, more complex feature extraction and richer data representation.
6. The final model output is a highly nonlinear function of the original input data.
7. Deep learning works with both supervised (labelled) and unsupervised (unlabelled) data.
8. **Convolutional Neural Networks (CNNs)** are a type of deep learning model with convolutional and pooling layers.

9. Convolutional layers apply learnable filters to extract features; pooling layers reduce data size via averaging or max-pooling.

10. Lower CNN layers capture simple features, while higher layers capture more abstract, complex features, with layer depth depending on problem complexity.

Example: Diagnosing Pneumonia from Chest X-Ray Images

1. **Input:** Chest X-ray images of patients' lungs.
2. **Convolutional Layers:** The CNN applies filters to detect edges, textures, and patterns indicating lung abnormalities.
3. **Pooling Layers:** These reduce the size of the image data while preserving important features like spots or shadows.
4. **Deeper Layers:** Later layers combine these basic features to identify complex signs of pneumonia, such as specific shapes or textures in lung tissue.
5. **Output Layer:** The model outputs a prediction—whether the patient has pneumonia or not.

Why CNN works well here:

- Automatically extracts important features from raw images.
- Handles the complexity and variability in medical images.
- Reduces the need for manual feature engineering by experts.

Thank You