TEAM CSGO

BIG DATA & CONNECTIONIST AI

A Comprehensive Study Guide
Book for Beginners

Authors: Thomas, Kevin, Richard, Nathan, Russell, An

Purpose

The goal of this study guide is clear and meaningful: to help beginners understand the key ideas of Big Data and Connectionist AI, two forces shaping our technology today. These fields not only serve as the foundation of computer science but also change industries like finance, healthcare, and education. Today, processing large amounts of data is essential for good decision-making while Connectionist AI uses neural networks to solve many real-world problems and drives innovation.

These subjects can seem intimidating to beginners because of their complex theories and math concepts. But it is essential to learn these two topics, especially with their growing relevance in modern technology. Don't worry though, as this guide aims to simplify difficult ideas into understandable concepts, encouraging curiosity and further learning.

While both fields are changing quickly, their basic principles remain unchanged. This guide focuses on the main concepts, technologies, and structures behind Big Data and connectionist AI. It provides readers with a solid foundation for future study.

Writing this guide has been a rewarding challenge. At first, we had little experience or knowledge about AI or computer science when we started at BIT. However, our curiosity and understanding of how important these technologies were becoming motivated us to learn. We gained knowledge through extensive research, including reading scholarly articles, studying textbooks, and searching online. Each team member focused on a different area, and we shared our findings to create a clear and understandable guide.

Understanding complex concepts was one of our biggest challenges. Since we taught ourselves, grasping advanced theories was tough, especially since we had never encountered them before. Additionally, we had to simplify without losing accuracy, ensuring the material remained clear and precise. We frequently asked ourselves how we would want these ideas explained if we were hearing them for the first time. As a result, we used relatable analogies, simple language, and real-world examples. Another challenge was managing our time effectively. This required discipline, teamwork, and careful planning to complete each chapter on schedule.

Learning Suggestions

To get the most out of this study guide, here are some learning strategies we suggest:

Focus on basics and going slow

We've made sure to create this study guide clearly and easily understandable for everyone, including beginners; it's not a must to have prior strong knowledge in mathematics and data before reading this guide.

However, in many cases, we've seen readers jumping straight into the complex topics before diving into the basic concepts first. In which, as a result, readers are struck with the never-before-seen terminologies and difficult theories, causing discouragement.

Thus, it is best to lower the pace and get one thing at a time. By having a clear grasp of the basics, it will make it much easier to understand harder topics later on.

Linking theoeretical knowledge with practical use

Relating theoretical knowledge to practical applications is essential for deepening understanding and developing the ability to transfer concepts to real-world scenarios. Readers are encouraged to examine case studies, research reports, or documented examples to observe how abstract principles are implemented in practice.

For Big Data, this may include analyzing how organizations leverage large-scale datasets to improve operational efficiency, optimize marketing strategies, enhance customer experiences, or inform strategic decision-making.

Understanding the specific tools and frameworks used, such as Hadoop, Spark, or cloud-based analytics platforms, can further contextualize the theoretical concepts.

Do active learning

Reinforcing theoretical understanding through practical exercises and experimentation is critical for consolidating knowledge and developing applied skills. Passive reading alone is often insufficient to fully grasp complex topics in Big Data and Connectionist AI. Learners are encouraged to actively interact with datasets, algorithms, and models to observe the underlying principles in action.

Do reviews at the end of each section

At the conclusion of each section, it is highly recommended that learners summarize key concepts in concise notes or structured outlines. This practice facilitates active reflection on the material, allowing learners to identify which areas have been fully understood and which may require further study. Summarization can include definitions of technical terms, descriptions of key frameworks, diagrams illustrating system architectures, and brief explanations of practical examples.

Learning Materials

Books:

- Big Data: A Tutorial-Based Approach by Nasir Raheem
- Big Data Analytics: A Guide to Data Science Practitioners (Ulrich Matter)
- Connectionism: A Hands-on Approach
- Big Data Concepts Technology and Architecture- Wiley (2021)
- Neural Networks (Simon O. Haykin)

YouTube Videos:

- "Big Data In 5 Minutes | What Is Big Data?"
- "Big Data & Hadoop Full Course in 12 Hours [2024]"
- "How Does Connectionism Work in AI?"

Websites:

- GeeksforGeeks
- Tutorials Point – Big Data & Analytics Tutorials
- DataFlair – Big Data Tutorials Home

Table of Contents

PART I BIG DATA

PART I
BIG DATA

CH1 Foundations of Big Data

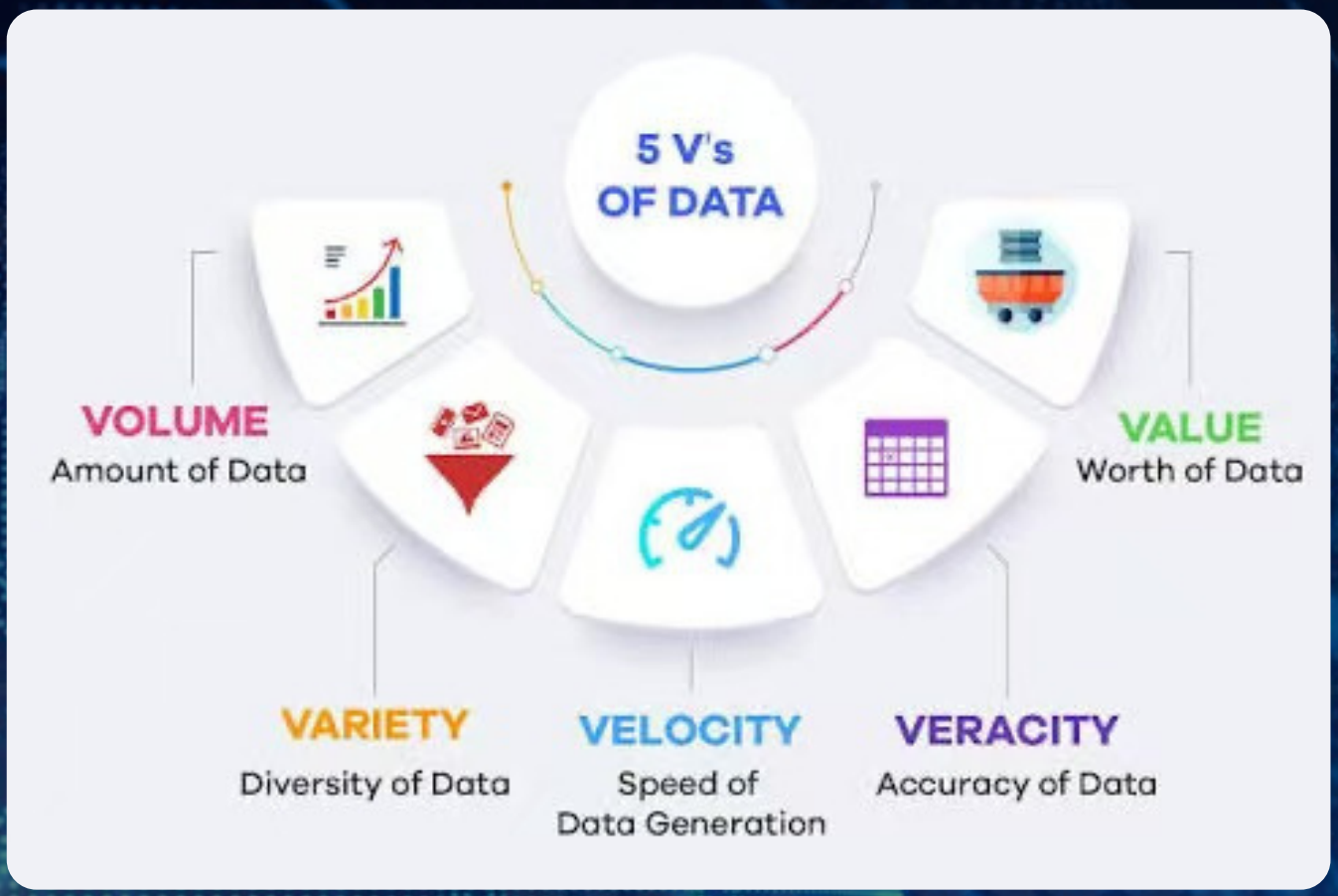
1.1 Foundations of Big Data

Big data refers to vast and complex datasets that are too large or intricate to be effectively managed, processed, and analyzed using traditional data management systems or techniques. Unlike smaller datasets, which can be handled with standard database tools and software, big data requires advanced tools, algorithms, and infrastructures to store, process, and derive meaningful insights. The rapid growth in the volume and complexity of data has led to the emergence of big data as a critical resource for organizations across industries.

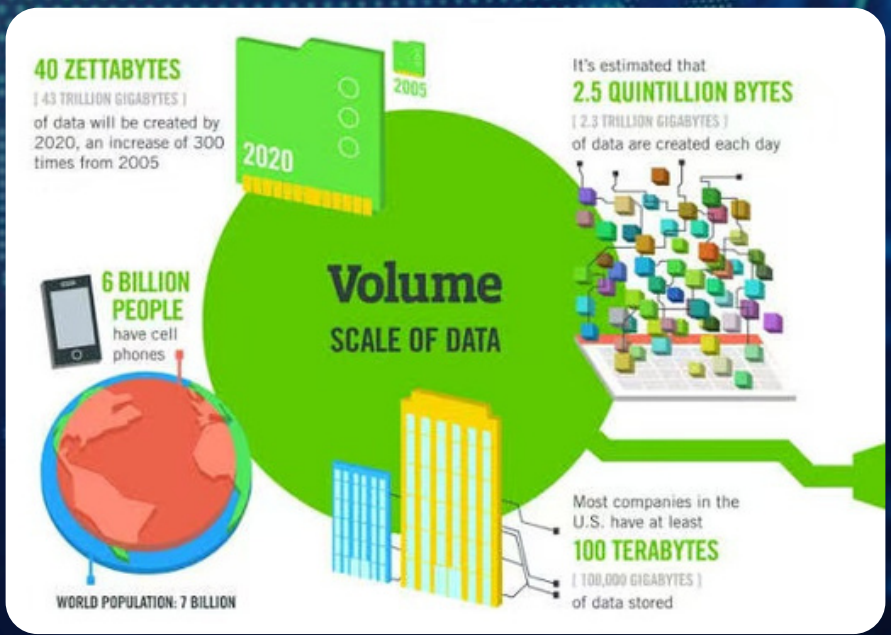
The foundations of big data are built upon several key concepts and technologies that enable the efficient processing, storage, and extraction of value from these massive datasets. These technologies include distributed computing, cloud storage, data mining techniques, and machine learning algorithms, all of which allow organizations to process data far beyond the limits of traditional data management systems.

By leveraging these tools, companies can uncover patterns, correlations, and trends that would otherwise remain hidden, providing them with the ability to make data driven decisions, optimize operations, and predict future outcomes.

1.2 The 5Vs in Big Data



The characteristics of big data are often characterized in the 5 Vs each of which plays a crucial role in defining what constitutes big data and how it is utilized.

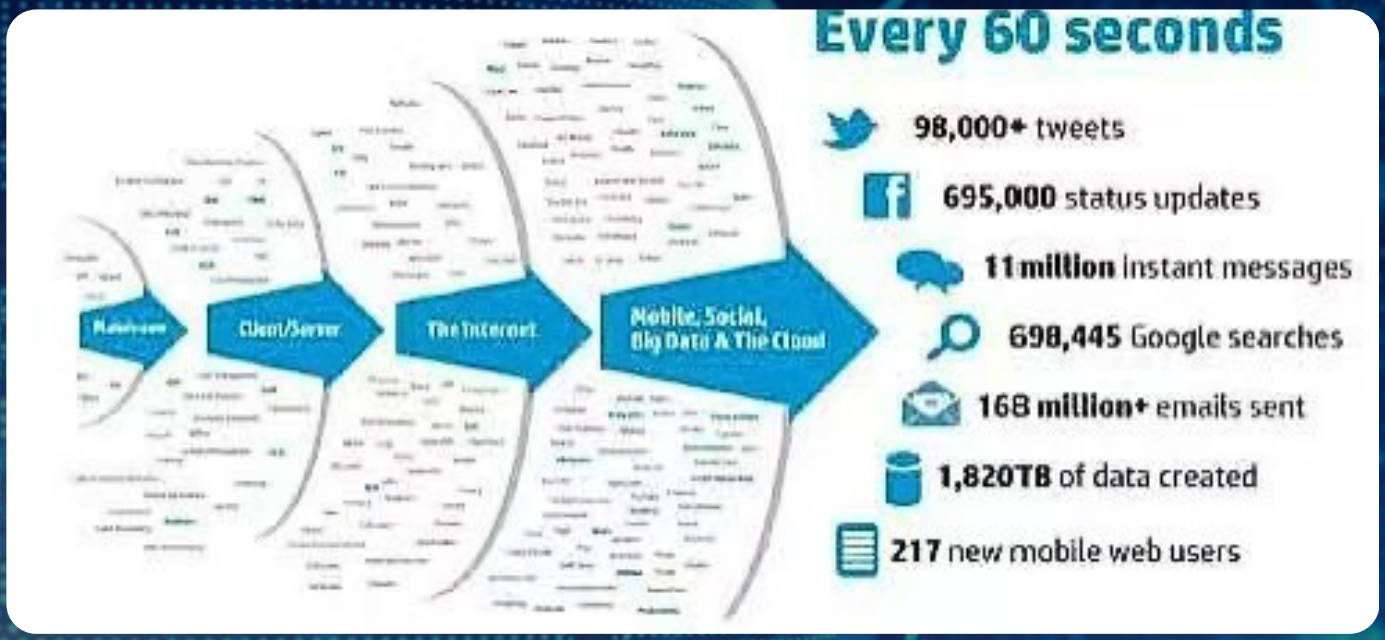


1) Volume

This refers to the sheer **amount of data** being generated, often measured in terabytes or petabytes.

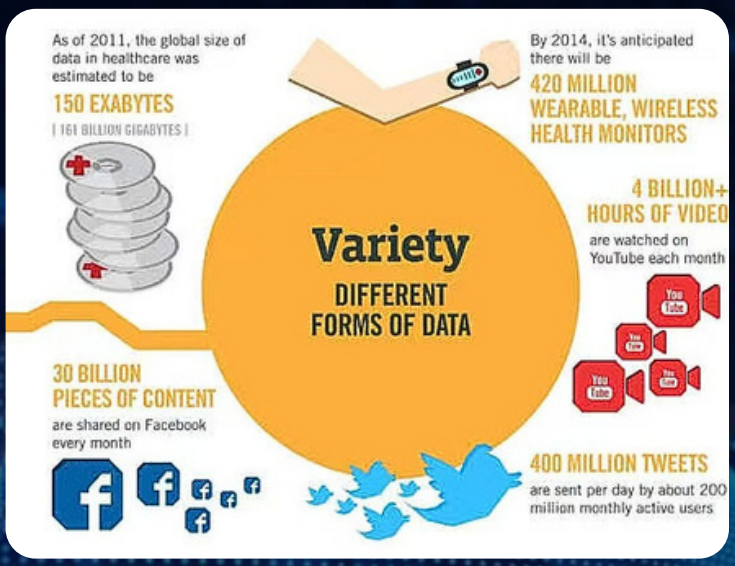
As organizations continue to collect data from various sources such as social media, sensors, and transaction logs, the volume of data grows exponentially, requiring scalable storage solutions to accommodate it.

2) Velocity



Velocity indicates the **speed at which data is generated, collected, and processed**.

With real-time data streams coming from sources like financial markets, social media feeds, and sensors in devices, organizations must be able to analyze data at high speed to make timely decisions and react to changing conditions quickly.



3. Variety

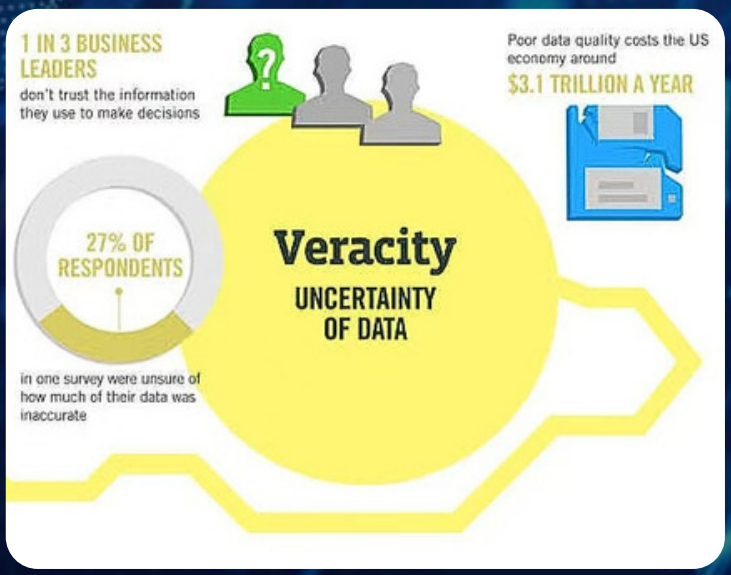
Data comes in many **different formats**, from structured data in traditional databases to unstructured data like videos, images, social media posts, and text documents.

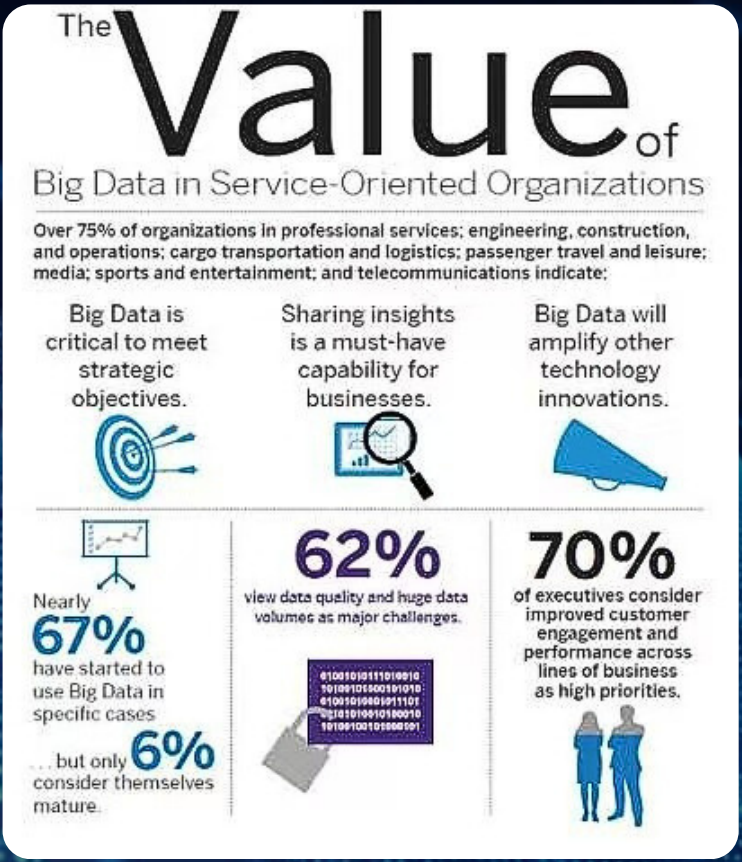
This adds complexity to the ways in which they are processed and analyzed, requiring specialized techniques to handle them.

4. Veracity

Veracity refers to the trustworthiness or **quality of the data**.

In the big data world, not all data is accurate or reliable. Managing data veracity involves ensuring that the data is clean, consistent, and free from errors or biases. With the vast amounts of data available, distinguishing useful, accurate data from misleading or irrelevant data is a significant challenge.





5. Value

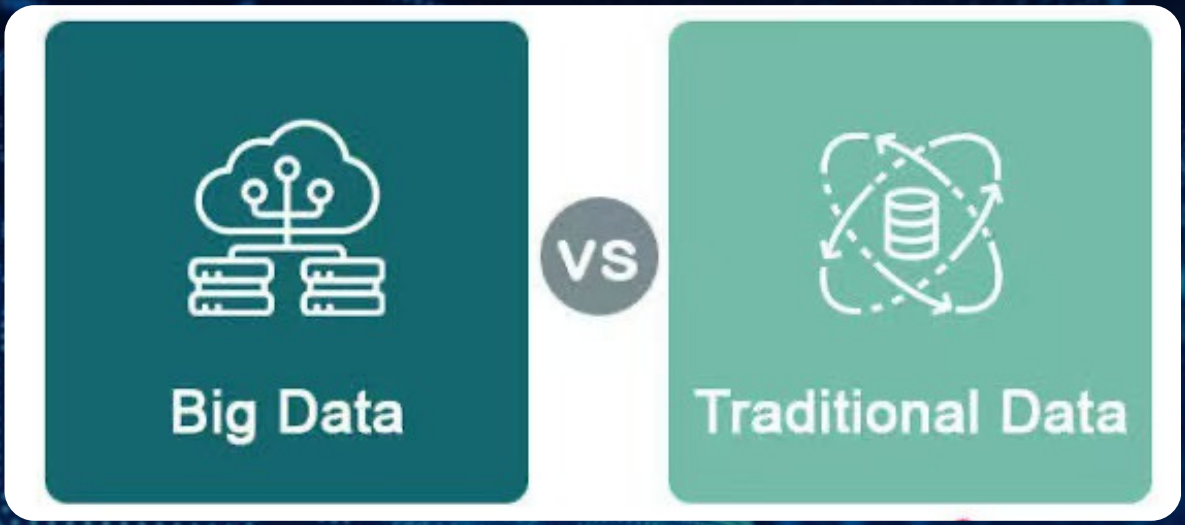
Value pertains to the **usefulness of the data** and the insights that can be derived from it.

Big data may be vast and plentiful, but its true value lies in its ability to generate actionable insights that can drive business decisions, improve efficiencies, or innovate new products and services.

The 5Vs in Big Data

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.

1.3 Big Data vs. Traditional Data



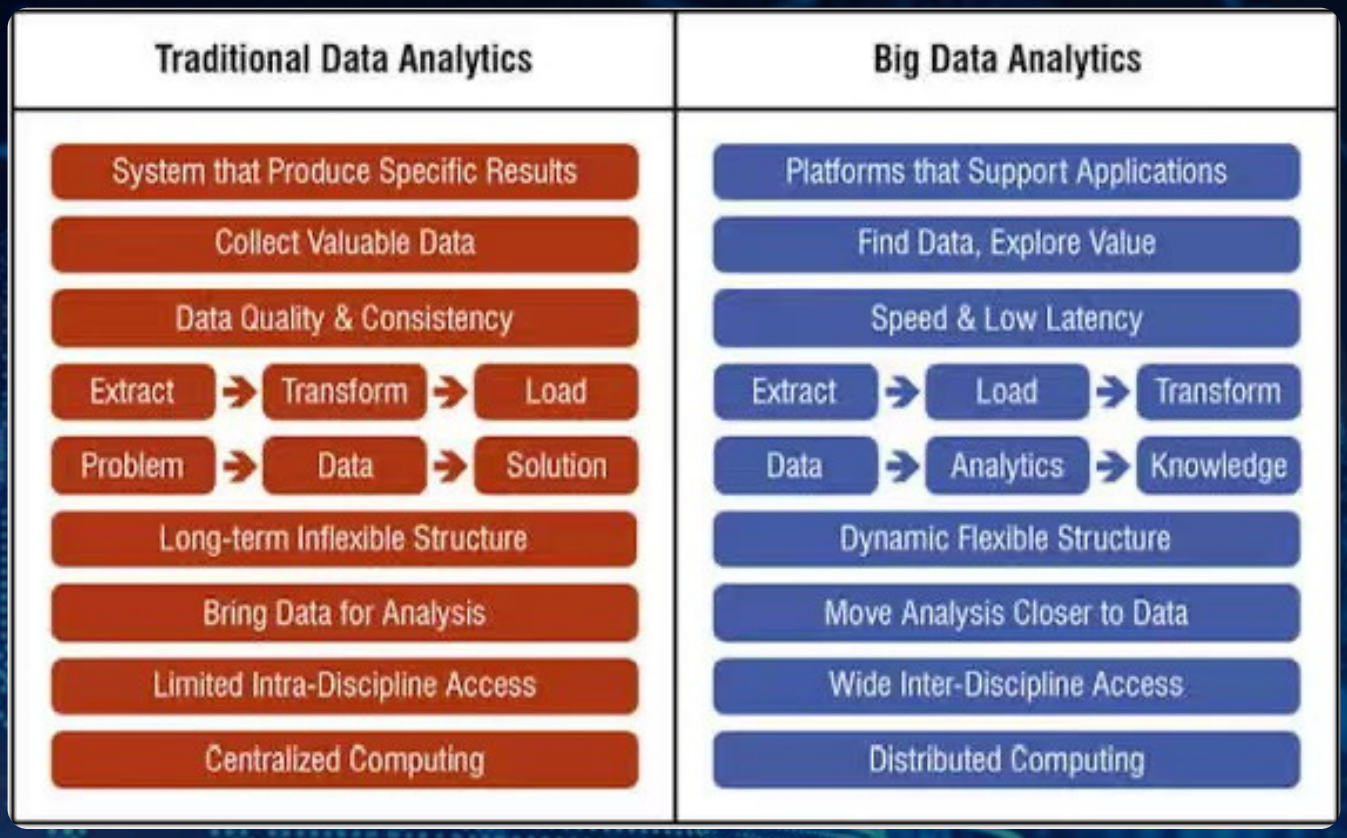
- **Volume:** Traditional data is small and easy to manage, often stored in simple systems like spreadsheets or relational databases. In contrast, big data is massive, typically measured in terabytes or petabytes, and requires specialized systems and technologies to store and process.

- **Variety:** Traditional data is structured, organized in rows and columns, and easy to analyze using basic tools. Big data, however, includes a mix of structured, semi-structured, and unstructured formats, such as social media posts, images, videos, and sensor data, making it more complex to process and analyze.

- **Velocity**: Traditional data is updated periodically, such as daily or weekly. Big data, on the other hand, is generated in real-time, requiring tools that can process and analyze it instantly, such as stream processing systems for things like website activity or real-time sensor data.

- **Complexity**: Traditional data is relatively simple to manage and analyze using standard tools. Big data requires advanced technologies like distributed computing, cloud storage, and machine learning to handle the volume, variety, and speed of data.

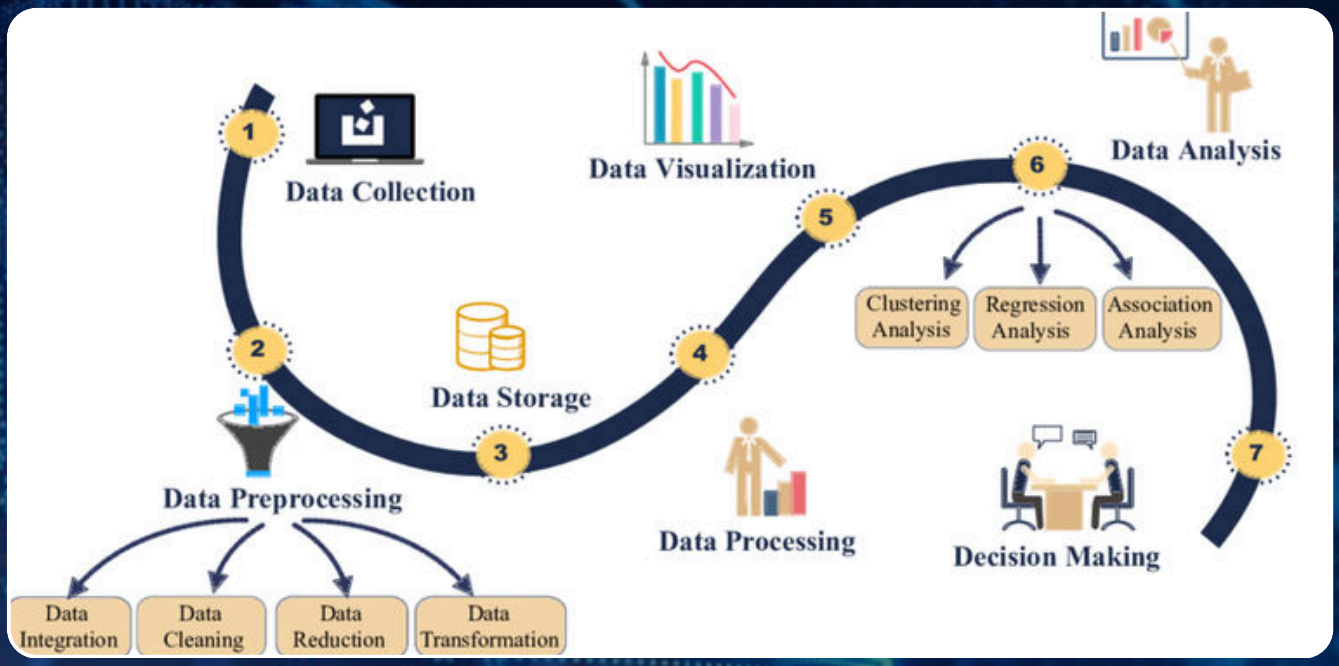
- **Value**: Traditional data provides limited insights, usually answering specific, structured questions. Big data offers far greater potential by uncovering complex patterns and trends that drive deeper, more valuable insights for business decision-making.



In conclusion, traditional data is easy to organize and store in systems like spreadsheets or small databases. It includes structured information such as customer records, sales transactions, or inventory levels, which can be analyzed using basic tools.

In contrast, big data is larger, more diverse, and more complex. It comes from various sources like social media, sensors, and online activity, and is often unstructured, requiring more advanced tools for storage and analysis. Big data can provide deeper insights and more significant business value, but it demands specialized infrastructure and techniques, such as cloud computing and machine learning, to manage and analyze effectively.

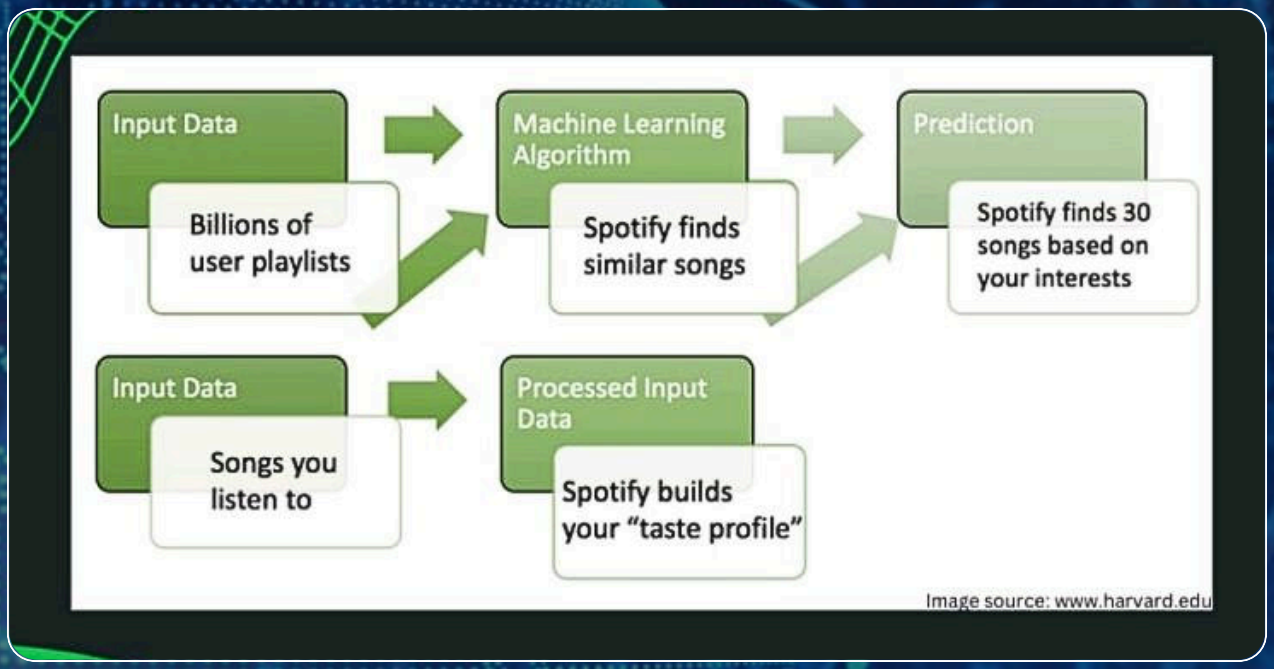
1.4 Importance of Big Data in Modern Computing



In modern computing, big data analytics begins by collecting vast amounts of structured and unstructured data from sources like social media, sensors, devices, transactions, etc. This data can range from neatly organized information like customer records to unstructured data such as social media posts, images, and videos.

Once collected, the data is processed, cleaned, and transformed into a usable format. Technologies like Hadoop, Apache Spark, and NoSQL databases help handle large volumes of data by distributing processing tasks across multiple servers. These tools ensure that the data is ready for analysis by removing errors and standardizing formats.

Next, advanced analytics techniques are applied to detect patterns, trends, and anomalies in the data. Machine learning algorithms and artificial intelligence (AI) can identify correlations and make predictions. For example, businesses can uncover trends in customer behavior or identify potential fraud by analyzing historical data.



The insights gained from big data analytics are used to drive informed decisions. For instance, Spotify leverages big data to offer personalized song recommendations, keeping users engaged. Similarly, Amazon uses analytics to suggest products based on customer behavior, enhancing shopping experiences and boosting sales.

In short, big data analytics enables organizations to unlock value from massive datasets, ultimately improving decision-making, driving innovation, and delivering personalized services to consumers.

Exercise

A. Multiple Choice Questions

1. What does the "Volume" characteristic of Big Data refer to?

A) The diversity of data types

B) The quality and accuracy of data

C) The amount of data collected

D) The speed at which data is processed

2. Which of the following is an example of "Velocity" in Big Data?

A) Customer purchase data for future predictions

B) Stock market transactions

C) Combining structured and unstructured data

D) Product reviews from websites

3. What challenge does the "Variety" of Big Data address?

A) Standardizing and distributing data from various sources

B) The ability to process data in real-time

C) Managing data quality and accuracy

D) The speed at which data is processed

4. Which of the following is an example of "Veracity" in Big Data?

A) Data from sensors in machines

B) Customer data used for purchase predictions

C) Data that may contain errors, discrepancies, or biases

D) Social media posts combined with customer data

5. What does the "Value" characteristic of Big Data primarily refer to?
A) The accuracy of the data
B) The size of the dataset
C) The speed at which the data is created
D) The benefits and insights that can be derived from the data

6. Which technology is mentioned for processing Big Data?
A) Excel spreadsheets
B) Hadoop, Spark, and NoSQL databases
C) SQL databases
D) Cloud storage systems

B. True or False

1. Big data can only be processed with structured data formats.
2. Traditional data updates periodically, while big data updates in real-time.
3. Big data includes only structured data, such as numerical tables.
4. Big data's "Veracity" refers to the trustworthiness and accuracy of the data.

Answers

Multiple Choice Questions

1. C
2. B
3. A
4. C
5. D
6. B

True or False

1. F
2. T
3. F
4. T

CH2 Big Data Storage

What is big data storage?

Big data storage refers to a system or method used to store and manage extremely large amounts of data, mostly generated from sources such as social media, sensors, transactions, IoT, and other sources that traditional storage systems struggle to handle efficiently.

Big data storage was designed to handle:

1. Large volumes (terabytes to petabytes)
2. High velocity (data generating and processing speed)
3. Great variety (different formats of data – structured, semi-structured, unstructured)

The purposes of big data storage are:

1. Store and manage massive amounts of data efficiently for ease of access and analysis;
2. Handle various formats of data (structured, semi-structured, and unstructured);
3. Expand itself to handle increasing volumes of data by adding more servers or storage nodes;
4. Ensure that data is always available for access, secured, and protected from loss during system failures or cyberattacks;
5. Provide information and insights for big data analyzing, machine learning, and business intelligence.

Examples of big data storage are as such.

HDFS

NoSQL

Cloud Storage

Data Lake

Data Warehouse

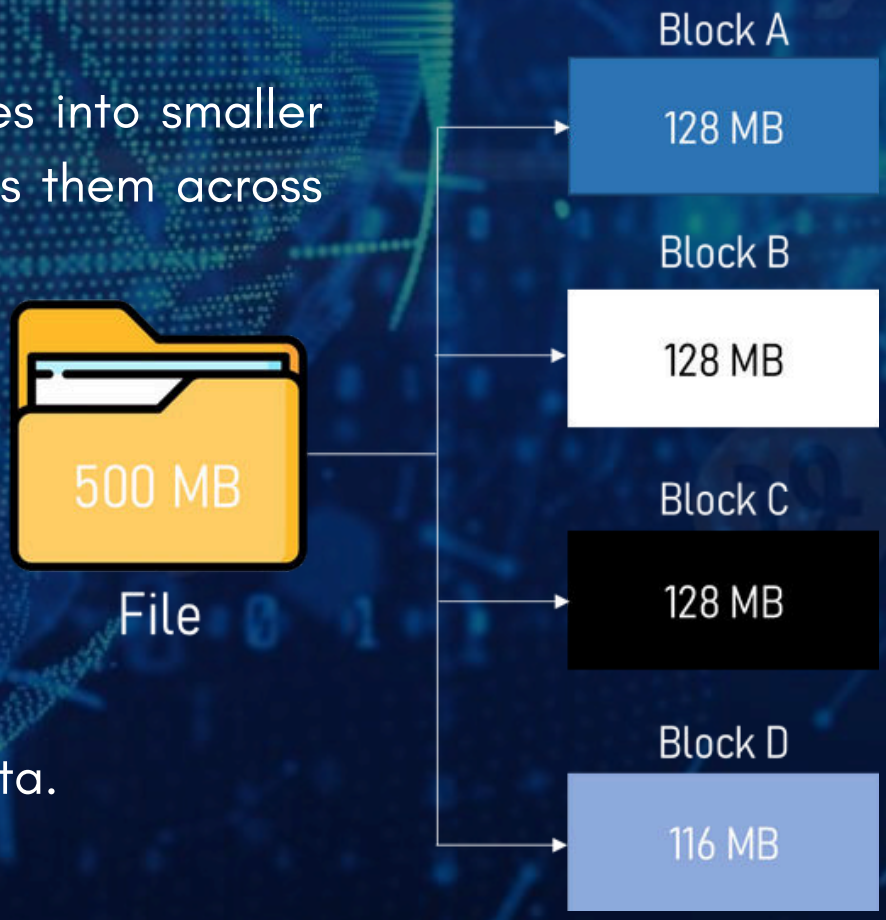
2.1 Hadoop Distributed File System

What is HDFS?

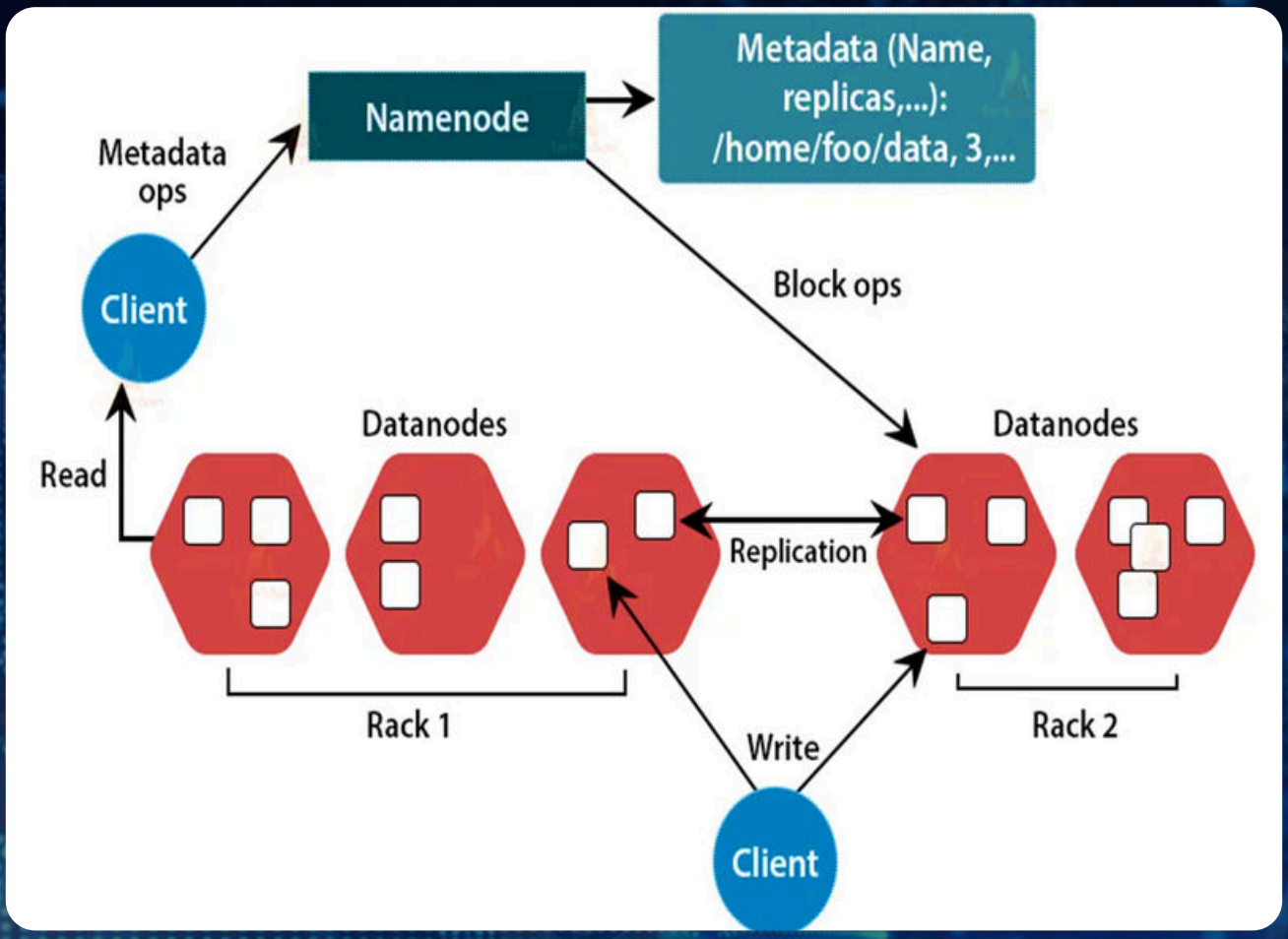
HDFS is the primary storage system for Hadoop applications that stores and manages massive datasets across clusters of computers.

The key components of HDFS are the namenode (master) that manages the file system and all metadata and the datanode (slave) which stores the data blocks on the nodes, perform read and write requests, and manage block creation, deletion, and replication as instructed by the namenode.

HDFS splits large files into smaller blocks and replicates them across different nodes, providing high availability for big data processing. If one datanode fails, HDFS still has other copies to recover data.



How does HDFS work?



a. Replication

Each block is replicated multiple times, then stored in different datanodes to ensure that those data are always available in case of failure.

b. Metadata management

The namenode keeps track of the metadata (file names, directories, locations of each data block, etc).

c. Reading operation

When a client wants to read a file, it first contacts the namenode for the metadata. The namenode then directs the client to the appropriate datanodes to read the blocks. To minimize network latency, the client is directed

to the closest replica of the blocks.

d. Writing operation

When the client wants to write a file, the client writes data by sending it to the datanodes, starting with the nearest datanode. Each data block is first written in the primary datanode, which then forwards it to the next datanode in the replication chain.

2.2 NoSQL

What is NoSQL?

The NoSQL (Not Only SQL) database is a type of database system that was designed to store, manage, and access large volumes of unstructured and semi-structured data efficiently.

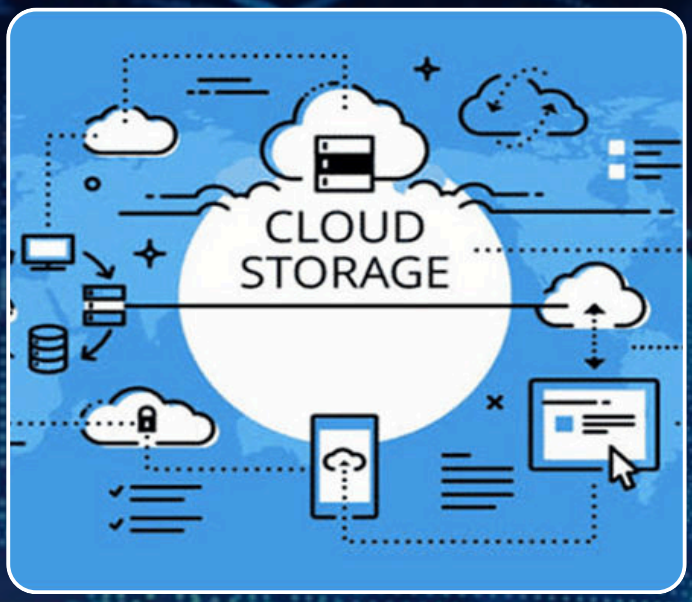
NoSQL databases use flexible data models and are not limited to only tables, rows, and columns. Just like big data storage, NoSQL was built to handle the 3Vs (massive volume, velocity, and variety of data).

There are four main types of NoSQL. Each type uses a different data model to handle specific challenges and use cases.

Type	Use	Example
Document-based	Uses documents (JSON, BSON, XML documents) to store data in the database.	MongoDB, Couchbase
Wide-column	Uses columns instead of rows to store data in the database.	Apache Cassandra, HBase
Key-Value	Uses key-value pairs to store data in the database. The data can be retrieved using a unique key allotted to each element in the database.	Redis, Amazon DynamoDB
Graph-based	Stores data in the form of nodes in the database.	Neo4j, Amazon Neptune

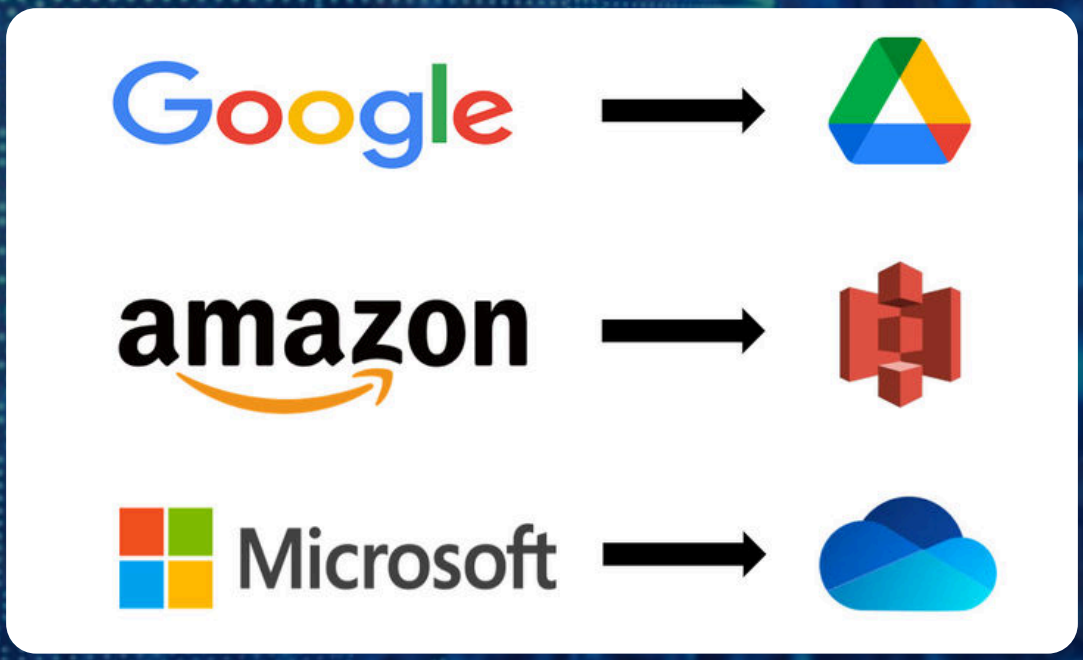
2.3 Cloud Storage

What is a Cloud Storage?



Cloud storage is an online data storage service that allows you to store, manage, and access data over the internet instead of keeping it locally on your physical storage device.

Cloud storages are usually provided by companies like Google, Amazon, and Microsoft.



Files in cloud storages are kept on remote servers called data centers. With cloud storages, you can access them anytime and anywhere as long as you are connected to the internet.

How does Cloud Storage work?

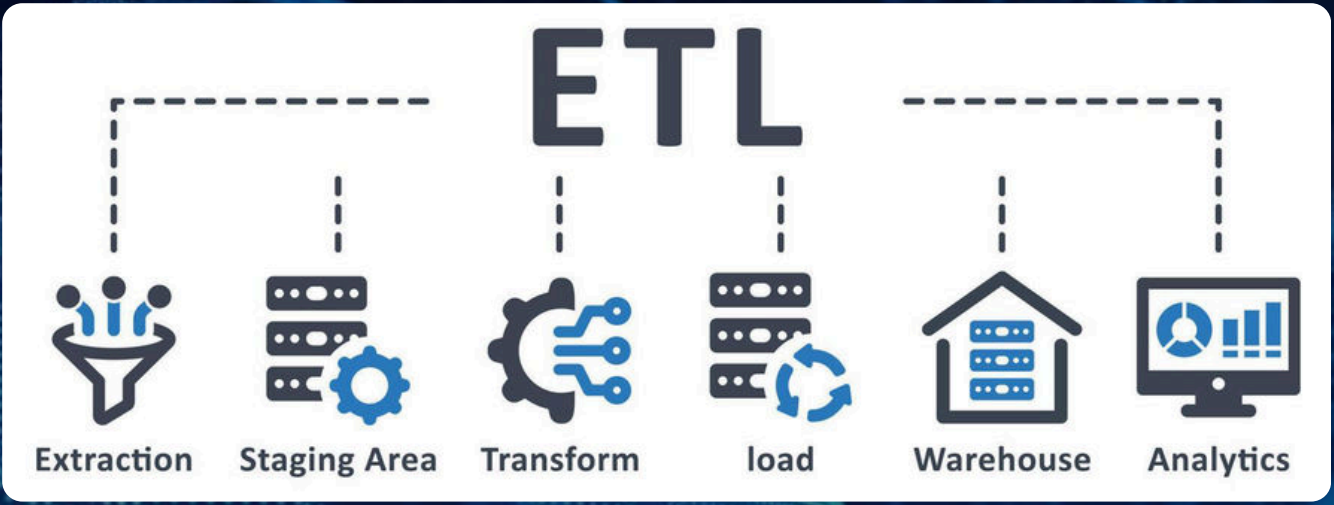
The user uploads files (documents, photos, etc.) via the cloud storage application on the internet. The file's data is then stored in virtualized servers managed by a cloud provider. When a user needs the data that they previously stored, they can retrieve said data through the web as long as they are connected to the internet.

2.4 Data Warehouse

What is a Data Warehouse?

Data warehouse is an organized repository and information system that stores organized data. Data warehouse is mostly used for analytical processing (OLAP), a processing system created to analyze huge amounts of data.

In a data warehouse, data is extracted from multiple sources, transformed into a consistent format, and loaded into a central data warehouse (also known as ETL).



The ETL process is a process used to prepare data for storage, analysis, and reporting in a data warehouse. This process involves three major phases to process raw data into a structured and usable form: Extraction, Transformation, and Loading.

a. Extraction

Raw data is collected from various sources. The data collected from those sources can include structured, semi-structured, and unstructured. The main goal of this phase is to gather data without changing its format so that it can be processed in the next phase.

b. Transformation

Extracted data in the previous phase is cleaned and formatted. This phase ensures that the data meets quality standards for analysis. Common transformations include data filtering, data sorting, data aggregating, and other

complex according to the organizational needs.

c. Loading

Clean and transformed data is loaded into a data warehouse. Depending on the use case, there are two types of loading methods, full load and incremental load.

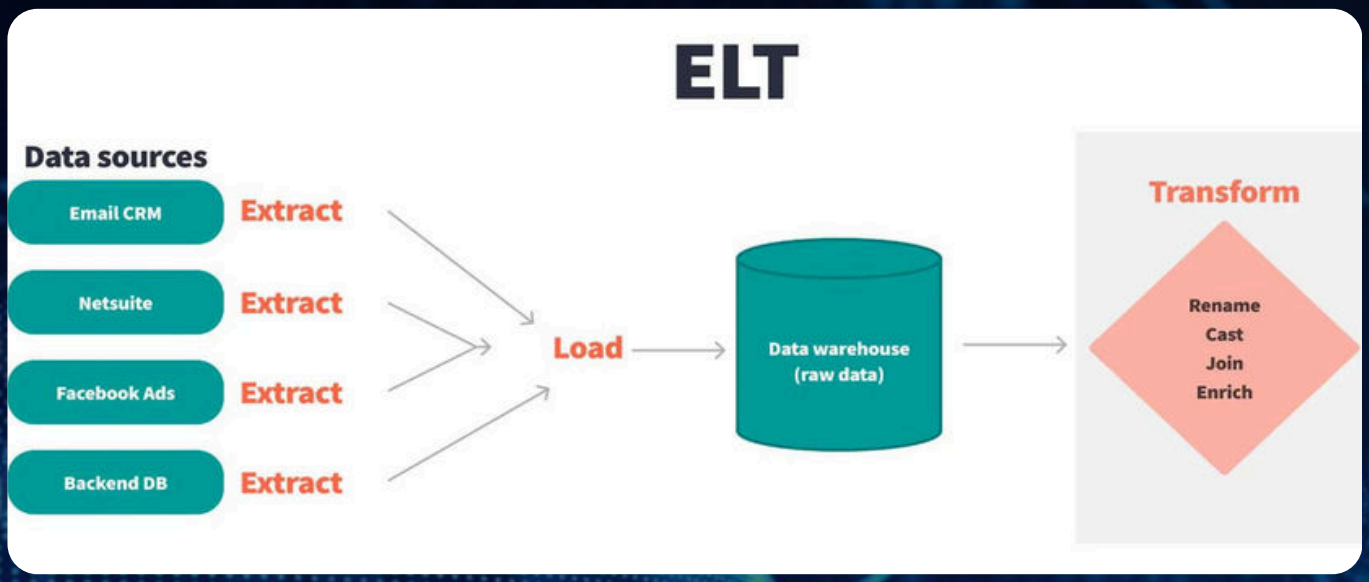
2.5 Data Lake

What is a Data Lake?

Data lake is a storage repository designed to capture and store a large amount of all types of raw data (be it a video, an audio, an image, a graph, a document, or anything else). The data can be structured, semi-structured, or even unstructured.

Once a data is in a data lake, the data can be used for machine learning and AI algorithms or for analytical purposes. Data in a data lake can also be organized and put into a data warehouse.

Data is extracted from multiple sources, loaded into the target system, and then structured into a format suitable for analysis (also known as ELT).



The ELT process is a process used to prepare and store data in a data lake for machine learning or further analytical purposes. This process involves three major phases: Extraction, Loading, and Transformation.

a. Extraction

The extraction phase in ELT has the same concept as in ETL. Data is gathered from various sources. The data gathered from those sources can include structured, semi-structured, and unstructured data.

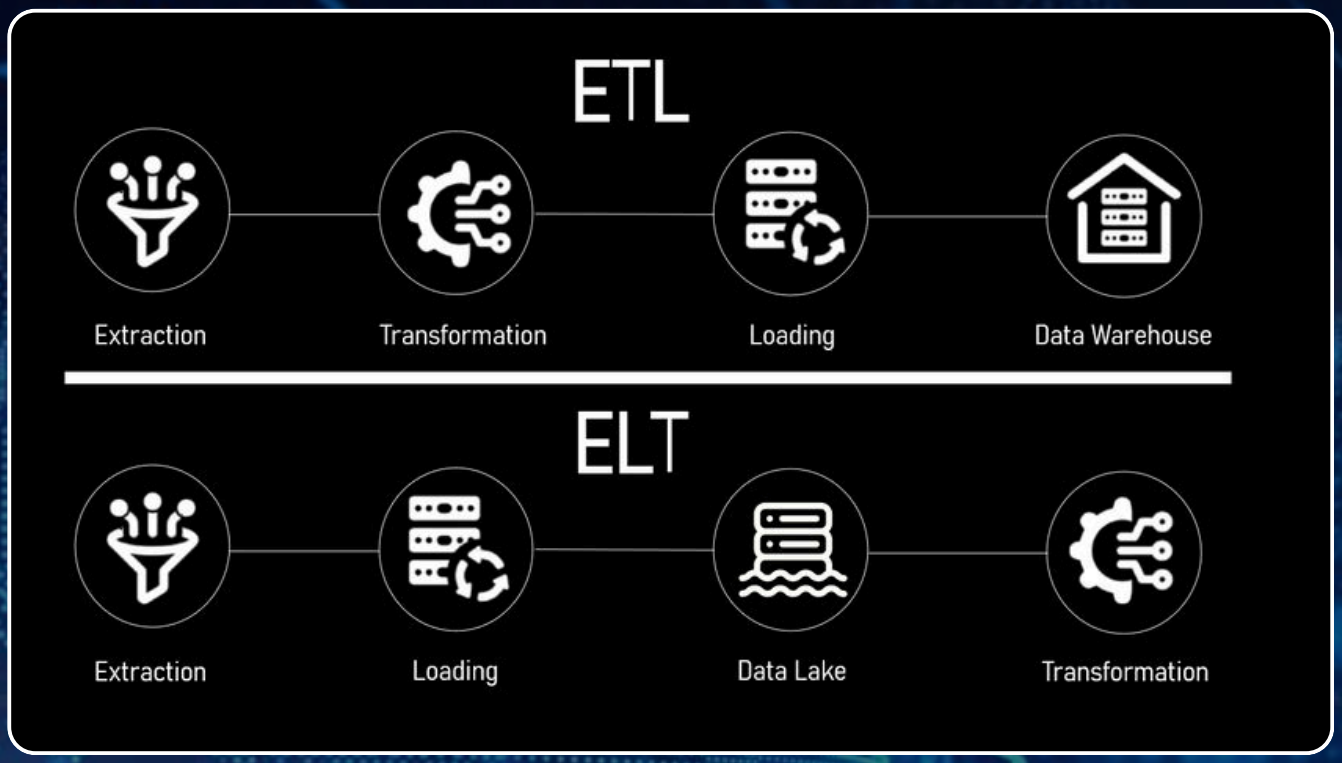
b. Loading

Extracted data is loaded in its original format straight into the target system. The general target for the loading phase is a data lake cloud that can handle large amounts of structured, semi-structured, or unstructured data.

c. Transformation

After raw data is loaded in the target system, it is cleansed and structured into a format suitable for analysis or machine learning.

Difference Between ETL and ELT



Aspect	ETL	ELT
Order	Extract, transform, load	Extract, load, transform
Transformation	In a seperate server or tool	In the target system
Suited For	Data Warehouse	Data Lake
Storage Requirements	Needs transformed data	Raw data can be stored
Data Type	Structured	Structured, semi-structured, unstructured

2.6 Use Cases

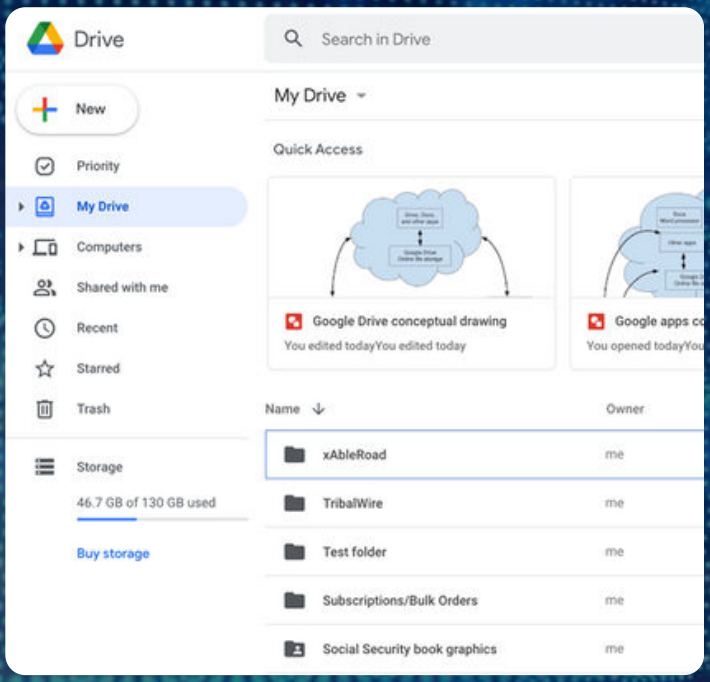
Large e-commerce platforms (like Amazon or Shopee) use HDFS to store massive web logs of daily user actions like clicks, searches, purchases, and browsing sessions, which quickly reach terabytes or petabytes in size.

How HDFS is used:

1. Stores massive website log files from user activity
2. The log files are split into blocks and replicated across multiple nodes
3. Hadoop's MapReduce processes the distributed data to identify user behavior trends, such as popular products or abandoned carts
4. This helps the platform improve product recommendations and marketing

Retailers (like Walmart) often need to update product information like prices, stock, images, and reviews. These updates need to happen in real time across multiple regions.

How NoSQL is used:
- Document-based databases (like MongoDB) store each product as a flexible JSON document, making it easy to update or add fields (e.g., adding "discount" or "color" later).
- Wide-column stores (like Cassandra) distribute product data across many nodes for high availability and fast access.
- Real-time applications can quickly query products without rigid schemas like SQL.



Companies and apps need to store and access data globally from user uploads to backups and documents without managing physical servers.

How Cloud Storage is used:
1. Files are uploaded to cloud platforms such as AWS S3, Google Drive, or Dropbox.
2. Data is automatically replicated across multiple geographic regions for durability.
3. Users or applications can retrieve files via the internet from any location.

Organizations collect data from multiple systems like sales, marketing, finance, etc. They need to merge this into one consistent source for reporting and analytics.

How Data Warehouse is used:

1. Data is collected from various databases using ETL.
2. Stored data is structured and optimized for querying.
3. The stored data is used for OLAP.

Organizations generate huge amounts of raw data like images, videos, text logs, IoT data, that might be used for future AI or ML analysis.

How Data Lake is used:

1. All raw data is stored without pre-processing in systems like AWS S3 Data Lake or Azure Data Lake.
2. The ELT process is used.
3. Data scientists access the data using Python, Spark, or ML frameworks to train models.

Exercise

A. Multiple Choice Questions

1. Which statement best describes big data storage?
A. A system used to store only small structured data sets
B. A method designed to manage and store extremely large, diverse, and fast-growing data efficiently
C. A local database used for personal computing
D. A tool to delete old or unnecessary data automatically

2. What is the main role of the NameNode in HDFS?
A. To store raw data in blocks
B. To manage metadata and file system structure
C. To replicate data blocks
D. To perform machine learning tasks

3. Which of the following pairs correctly matches a NoSQL type with its example?
A. Document-based – MongoDB
B. Graph-based – Redis
C. Key-value – HBase
D. Wide-column – Neo4j

4. What distinguishes a data lake from a data warehouse?
A. Data lakes store only structured data
B. Data warehouses store raw, unprocessed data

C. Data lakes can store all types of raw data while data warehouses store processed and structured data
D. There is no difference between them

B. Fill in the Blanks

1. Big data storage is designed to handle large _____, high _____, and great _____ of data.
2. In HDFS, the _____ manages metadata and directory information, while the _____ stores data blocks.
3. NoSQL databases are designed to store and manage large volumes of _____ and _____ data.
4. The ETL process in data warehousing stands for _____, _____, and _____.

C. True or False

1. Big data storage systems are only capable of storing structured data.
2. Cloud storage allows users to store and access files through the internet rather than on local devices.
3. A data warehouse stores raw, unprocessed data for machine learning models.
4. HDFS replicates data blocks across multiple nodes to ensure data availability even if one node fails.

Answers

Multiple Choice Questions

1. B
2. B
3. A
4. C

Fill in the Blanks

1. Volume, velocity, variety
2. Namenode, datanode
3. Semi-structured, unstructured
4. Extract, transform, load

True or False

1. F
2. T
3. F
4. T

CH3 Big Data Processing Frameworks

3.1 Introduction to Big Data Processing Frameworks

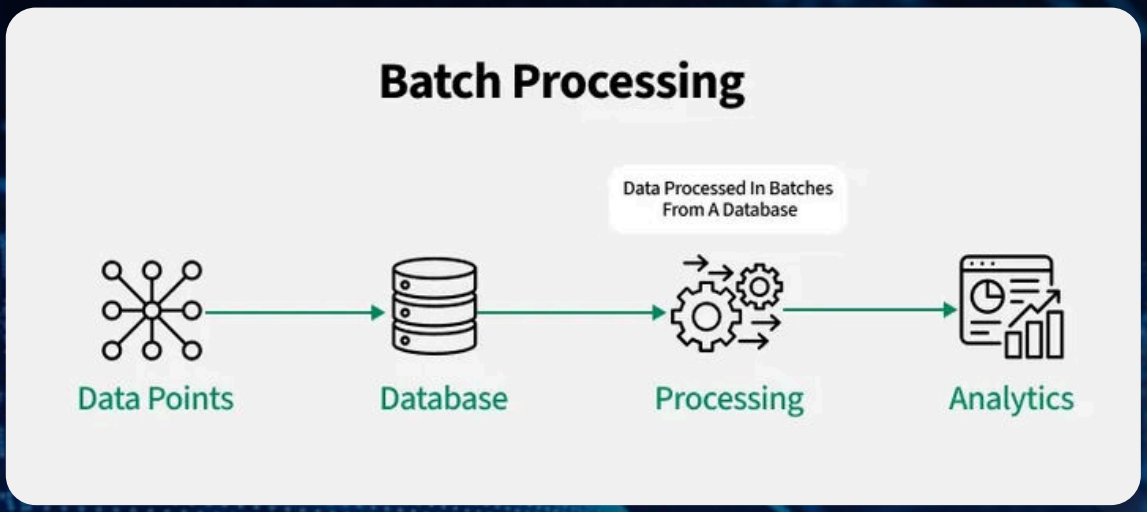
Big Data Processing Frameworks are specialized software tools designed to handle the immense volume, velocity, and variety of data that traditional systems cannot process efficiently.

They provide the architecture and utilities to store, process, and analyze massive datasets, enabling organizations to extract valuable insights. A core concept in this domain is understanding the two primary data processing paradigms: Batch Processing and Stream Processing.

3.2 Batch Processing & Stream Processing

3.2.1 Batch Processing

Batch processing is a method of data processing where computers handle high-volume,repetitive tasks by grouping data into "batches." The system collects data over a period, and then the entire batch is processed at once as a single job. This approach is not designed for tasks that require instant or real-time results.



Key Characteristics

· Data Nature: Processes finite, static, and predetermined chunks of data.

· Processing Time: Operates on a set schedule (e.g., hourly, daily, nightly).

· Latency: High latency; insights are only available after the entire batch job is completed.

Advantages

· Efficiency for Large Datasets: Highly optimized for processing terabytes or petabytes of data in a single, efficient job.

· Cost-Effective Resource Usage: Can be scheduled for off-peak hours, leveraging idle computing resources.

· Streamlined Workflows: Well-defined start and end points make workflow management straightforward.

· Error Handling: Robust error handling can be implemented for the entire batch.

· Scalability: Easily scaled by adding more data to a batch or increasing compute resources for batch jobs.

Disadvantages

· Delayed Outcomes: The primary drawback; no insights are available until the batch job completes.

· Resource Spikes: Can cause significant spikes in resource consumption (CPU, memory) when a job starts.

· Inflexibility: The data is static; new data arriving after the job starts is not included.

· Error Propagation: An error in the data or code can invalidate the entire batch's output, requiring a full re-run.

· Higher Upfront Costs: Designing and building large-scale batch pipelines can be complex and costly.

3.2.2 Stream Processing

Stream processing is a method that continuously ingests, processes, and analyzes data records as they are generated. Instead of waiting for data to accumulate, it processes data in motion, enabling immediate responses to changing conditions. This is critical for use cases requiring fast, event-driven decisions.

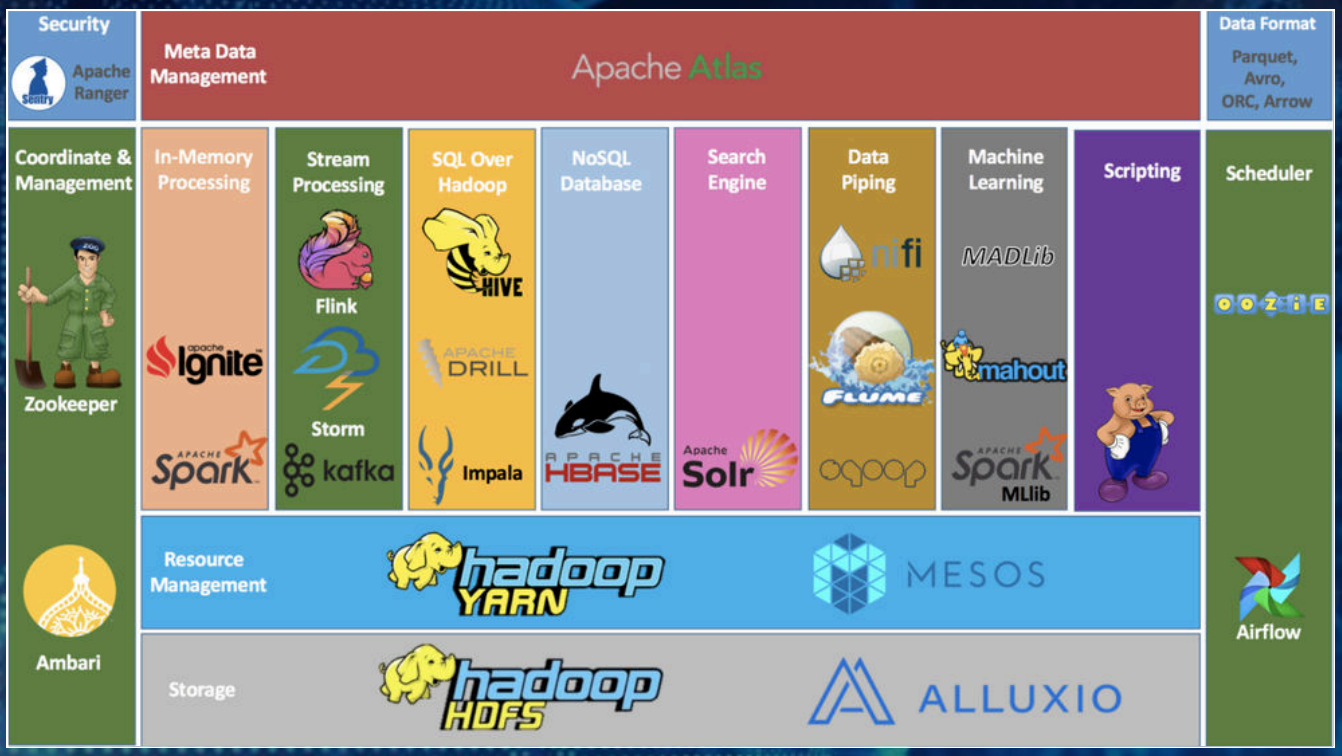
Disadvantages

- Increased Complexity: Designing, implementing, and managing low-latency, stateful streaming applications is challenging.
- Higher Operational Costs: Requires systems to be always-on and monitored, leading to higher infrastructure and maintenance costs.
- Data Accuracy Challenges: Handling late data, out-of-order events, and ensuring exactly-once processing is complex.
- Monitoring and Maintenance: Requires constant vigilance to ensure data is flowing and being processed correctly.
- Limited Historical Context: By default, focuses on the most recent data, though this can be mitigated with stateful processing.

Criteria	Batch Processing	Stream Processing
The Nature of the Data	Processed gradually in batches.	Processed continuously in a stream.
Processing Time	On a set schedule.	Constant processing.
Complexity	Simple, as it deals with finite and predetermined data chunks.	Complex, as the data flow is constant and may lead to consistency anomalies.
Hardware Requirements	Varies; can be performed by both low-end and high-end systems.	Demanding, requiring the system to be operational at all times.
Throughput	High, optimized for large amounts of data.	Varies depending on the task at hand.
Application	Email campaigns, billing, invoicing, scientific research, image/video processing.	Social media monitoring, fraud detection, healthcare monitoring, network monitoring.
Consistency & Completeness	Data consistency and completeness are usually uncompromised.	Higher potential for computed or out-of-order data.
Error Recognition & Resolution	Errors can only be recognized and resolved after processing is finished.	Errors can be recognized and resolved in real-time.

3.3 The Apache Ecosystem

The Apache Software Foundation provides a suite of robust, open-source tools that form a complete pipeline for big data processing, from ingestion to visualization.



A typical data pipeline using Apache tools follows these stages:

1. Data Ingestion & Streaming (Apache Kafka): Kafka acts as the central nervous system, collecting, buffering, and transmitting real-time data streams from various sources (e.g., logs, IoT sensors, transactions). It decouples data producers from consumers.

2. Data Storage (Apache Hadoop HDFS): The raw data from Kafka can be sent to Hadoop's Distributed File System (HDFS) for cheap, scalable, and long-term storage. HDFS can handle structured, semi-structured, and unstructured data.

3. Data Processing & Analytics (Apache Spark / Apache Flink):
· Apache Spark: Processes large-scale data both in batch and micro-batch (for near-real-time) modes, leveraging in-memory computing for speed. Ideal for ETL, batch analytics, and machine learning.
· Apache Flink: Specializes in true, low-latency stream processing with advanced capabilities like event-time processing and exactly-once semantics. Ideal for complex event processing and real-time analytics.
4. Data Querying & Visualization (Apache Superset): Superset provides a Business Intelligence (BI) dashboard layer. It allows users to run SQL-based queries on the processed data and create interactive visualizations to gain insights.

3.3.1 Apache Kafka

Real-time data streaming and message queuing.

When to use:
· You need a real-time streaming platform and an event-driven architecture.
· You want to decouple microservices in a distributed system.
· You need to ingest high-throughput data from multiple sources (e.g., logs, IoT, financial transactions).
Example: Uber uses Kafka to handle millions of ride events per second.

3.3.2 Apache Hadoop

When to Use:

· You need cheap, reliable, and scalable storage for huge datasets (petabyte-scale).

· You have historical big data that needs to be processed later in batches.

· You are running large-scale batch ETL (Extract, Transform, Load) pipelines.

Example: Facebook stores exabytes of user data in Hadoop for analysis.

3.3.3 Apache Spark

Function: Fast, in-memory data processing for both batch and micro-batch streaming.

When to Use:

· You need fast in-memory batch processing for iterative algorithms (e.g., Machine Learning).

· You want to run complex machine learning models at scale using MLlib.

· You need to perform SQL-based querying (via Spark SQL) over big data.

Example: Netflix uses Spark for its real-time recommendation systems.

3.3.4 Apache Flink

Function: True, low-latency stream processing.

When to Use:

· You need real-time, event-driven processing with very low latency.

· Your use case requires true stream processing with better latency than Spark Streaming's micro-batches.

· You are working with IoT data, financial transactions, or real-time monitoring logs.
Example: Alibaba uses Flink for real-time fraud detection in its financial services.

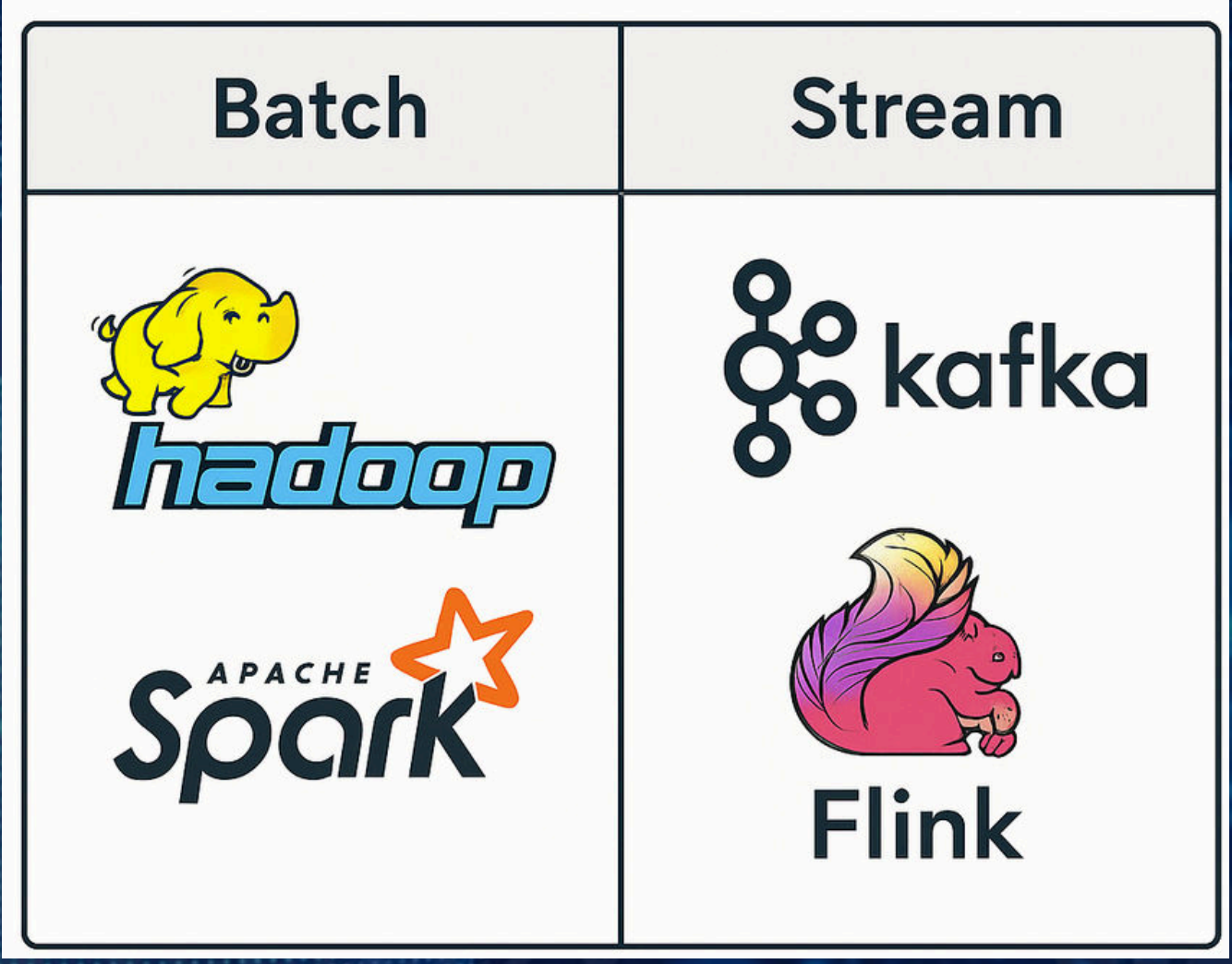
3.3.5 Apache Superset

Function: Data visualization and business intelligence dashboards.
When to Use:
· You want to visualize and explore big data from various sources like Hadoop, Kafka, or Spark.
· You need an open-source, powerful alternative to commercial tools like Tableau or Power BI.
· You need an interactive, SQL-based BI tool for business teams.
Example: Airbnb uses Superset for its internal business analytics and data exploration.

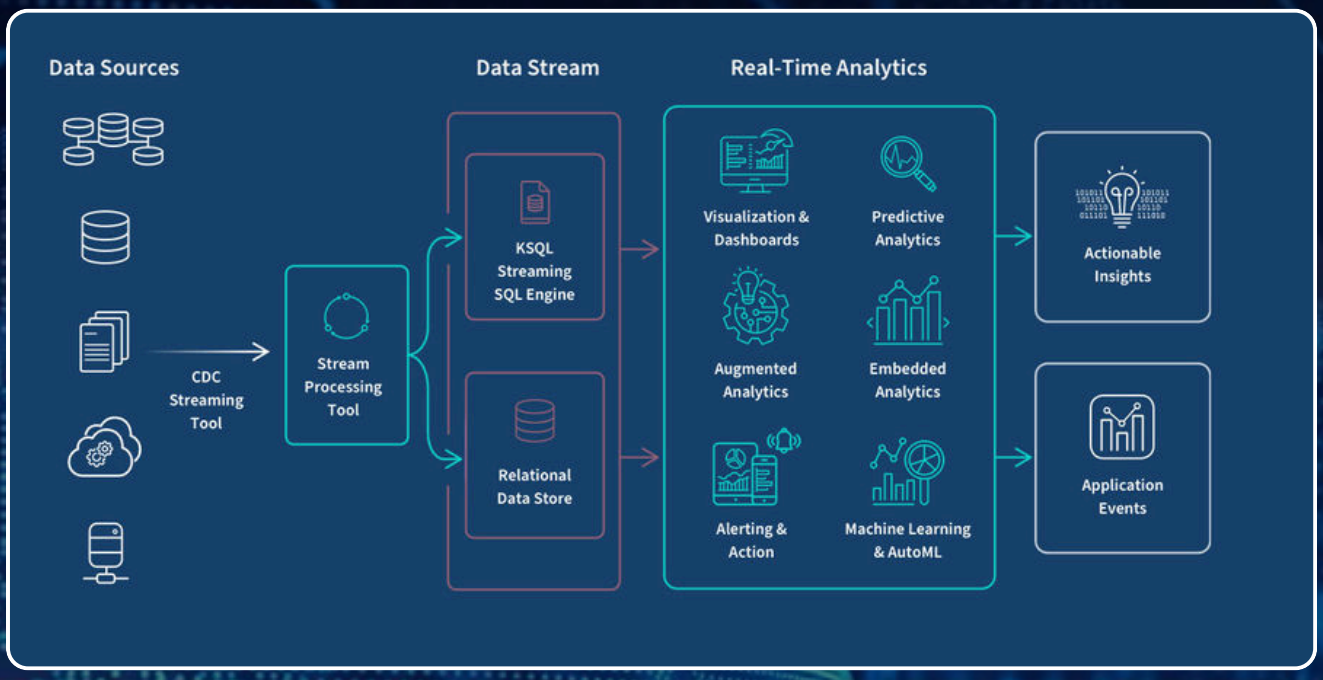
3.3.6 Categorization by Processing Paradigm

· Batch Processing Tools: Apache Hadoop (MapReduce), Apache Hive, Apache Spark (for batch jobs).

· Stream Processing Tools: Apache Kafka (ingestion), Apache Flink, Apache Storm, Kafka Streams.



3.4 Real-Time Analytics



Real-time analytics refers to the practice of analyzing and acting upon data as soon as it is generated. Unlike batch processing, where data is collected, stored, and processed later, real-time analytics systems continuously process incoming data streams with minimal latency—often within milliseconds or seconds. This enables immediate decision-making.

Implementing a real-time analytics solution involves a structured process:

1. Architecture Assessment: Evaluating business requirements and designing a system capable of handling the expected data velocity and volume.
2. Big Data Processing: Selecting and configuring the appropriate processing frameworks (e.g., Flink, Spark Streaming) for the task.
3. Technology Evaluation: Choosing the right mix of technologies from the ecosystem (Kafka, Druid, etc.) that fit the use case.
4. Solution Development & Prototyping: Building, testing, and deploying the end-to-end data pipeline.

3.5 Popular Real-Time Big Data Frameworks

Framework	Type	Key Features	Use Cases
Apache Kafka	Streaming Platform	Pub-sub messaging, durable logs, high throughput.	Event streaming pipelines, data ingestion.
Apache Spark	Micro-batch Processing	Integrates with batch Spark, fault-tolerant.	Real-time analytics where latencies of a few seconds are acceptable.
Apache Flink	Stream Processing	True event-time processing, exactly-once semantics.	Real-time fraud detection, IoT analytics.
Apache Storm	Stream Processing	Low-latency, distributed computation.	Real-time computation (e.g., Twitter analytics).
Kafka Streams	Stream Processing Library	Lightweight, runs inside your application.	Processing Kafka data streams directly.
Apache Samza	Stream Processing	Works with Kafka/YARN, good scalability.	Log processing, real-time monitoring.
Apache Druid	OLAP Datastore	Time-series optimized, very fast queries.	Real-time dashboards, interactive business intelligence.

Exercise

A. Multiple Choice Questions

1. A company needs to analyze terabytes of historical sales data from the previous year to generate its annual report. Which processing paradigm is most suitable for this task?
A. Stream Processing
B. Real-time Processing
C. Batch Processing
D. Event-driven Processing

2. According to the document, which of the following is a key disadvantage of Stream Processing?
A. Delayed Outcomes
B. Inflexibility
C. Data Accuracy Challenges
D. Error Propagation

3. Which Apache framework is primarily described as a "streaming platform" that acts as a message broker to decouple data producers and consumers?
A. Apache Hadoop
B. Apache Spark
C. Apache Flink
D. Apache Kafka

4. A development team is building a financial transaction system and requires the lowest possible latency for processing data streams to detect fraud as it happens. Which framework is specifically noted for having an advantage in low-latency stream processing over Spark Streaming?
A. Apache Hadoop
B. Apache Kafka
C. Apache Flink
D. Apache Superset

5. The document lists several "When to Use" criteria. For which scenario would Apache Hadoop be the most appropriate choice?
A. You need to run complex machine learning models at scale.
B. You need real-time event-driven processing.
C. You need cheap and scalable storage for huge datasets.
D. You want to visualize data with an open-source BI tool.

6. What is a defining characteristic of Real-time Analytics, as described in the document?
A. Data is collected, stored, and processed later on a set schedule.
B. It is best suited for processing finite and predetermined data chunks.
C. It analyzes and acts on data as soon as it's generated, with minimal latency.
D. It has high latency, with insights available only after processing is complete.

B. True or False

1. Batch processing is well-suited for tasks that require instant or real-time results.
2. Apache Spark is capable of both batch processing and real-time data processing.
3. In stream processing, data inputs are static and preset, unlike in batch processing.
4. Apache Superset is primarily used for distributed data storage and batch processing.

Answers

Multiple Choice Questions

1. C) Batch Processing
2. C) Data Accuracy Challenges
3. D) Apache Kafka
4. C) Apache Flink
5. C) You need cheap and scalable storage for huge datasets.
6. C) It analyzes and acts on data as soon as it's generated, with minimal latency.

True or False

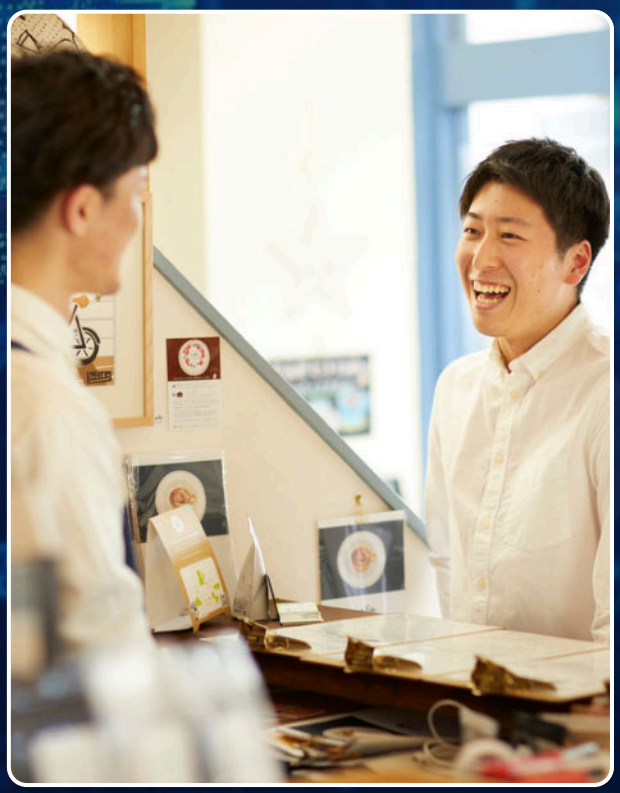
1. False
2. True
3. False
4. False

CH4 Big Data Analysis & Machine Learning

4.1 What is Big Data Analytics?

Big Data Analytics is the science of analyzing large data sets with a variety of data types (unstructured, structured, and semi-structured data), which may be streaming or batch data, to find out the hidden patterns and relationships. This process is vital for businesses, transforming the result into actionable knowledge. Businesses can have better strategies, improve performance and efficiency, and reduce costs by implementing the analytics techniques.

For example, in online retail, analyzing a large volume of transaction data helps retailers learn more about customer behavior and purchasing trends to make business decisions. Also, Facebook analyzes users' posts and likes to determine the suitable advertisement.

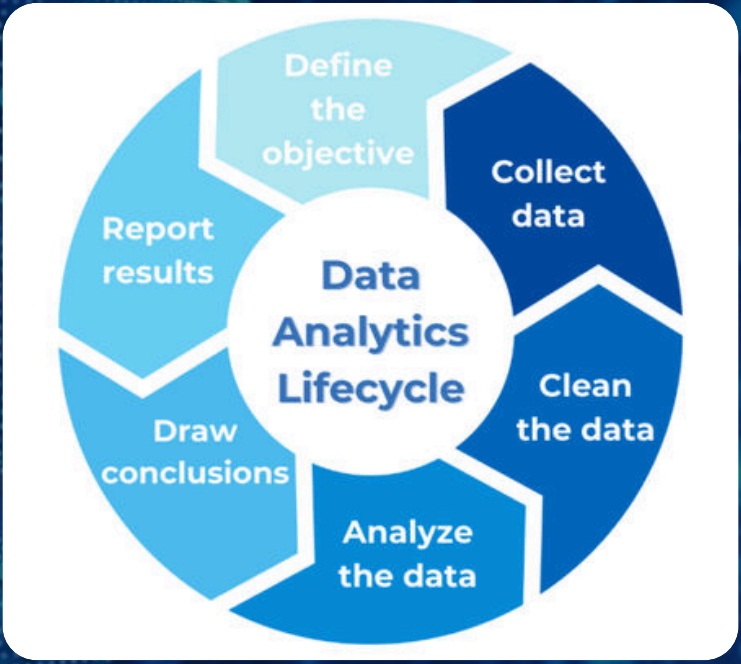


4.2 Data Analytics Life Cycle

Business Case Evaluation and Identifying Source Data

The first step in data analytics is to find the business problem that has to be solved with data analytics. After evaluating the business case, data scientists can have a clear image of the analysis' objective, helping them find the required resources.

To know if the problem is related to big data, it must be connected with the 5Vs of big data, which are volume,velocity, variety, veracity, and value.



The data sources at hand are then evaluated by the data scientists, which may come from within the organization or external providers, to check if the data is sufficient for analysis. If it is not, they may need to collect more data or transform what's available.

Data Preparation and Transformation

Collected data can be inaccurate or inconsistent, making data preprocessing essential for effective analysis. The data needed may be distributed across multiple datasets that must be combined using what's common. For example, a data integration with EmpId field:

EmpId	Name
4567	Maria
4656	John

EmpId	Salary	DOB
4567	$2000	08/10/1990
4656	$3000	06/06/1975

EmpId	Name	Salary	DOB
4567	Maria	$2000	08/10/1990
4656	John	$3000	06/06/1975

But due to the difference in data structure and semantics —different labels with the same meaning, such as "income" and "salary"-- this process becomes complicated. Furthermore, the data have to be cleansed, detecting and removing redundancies. It is done using ETL (extract, transform, and load) in a batch system, but through an in-memory database system in real-time analysis. Data must also be transformed into a format suitable for analysis—normalizing data, aggregating data, and converting formats.

Data Analysis, Visualization, and Applications

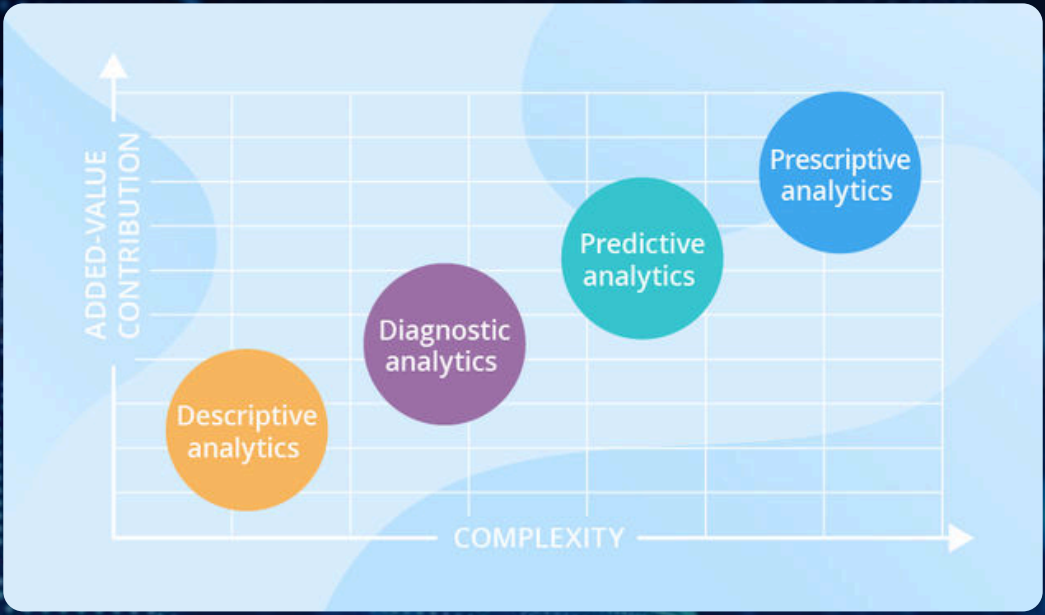
This phase is where the actual analysis on big data is performed. Data analysis can be divided into two categories:

- Confirmatory analysis: deductive—starts with an idea, then tests the hypothesis, whether to confirm or reject it, using the data gathered.

- Exploratory analysis: inductive—the hypothesis is unknown yet, but exploring the data freely, finding the patterns and relationships along the way

Data visualization is a process where results of analysis are converted to be visually presented so that businesses can interpret it easier. This ensures they can put the result into action, evolving business strategies and increasing profits.

4.3 Four Types of Analytics



4.3.1 Descriptive Analysis

Describes the events and gains an understanding into what has happened in the past. In this analysis, past data are mined and summarized to identify patterns, enabling businesses to know their performance over time. Since descriptive analysis involves with processing raw data, it mostly uses statistical methods, such as:

- Measuring the central tendency: finding the mean, median, and mode
- Measuring dispersion: finding how spread out the data is with one another as well as how much the data points differ from the mean value. Common measures include variance and standard deviation

- Trend analysis: Converting raw data into visual tools like line charts or graphs to illustrate trend over time, making it easier to find patterns

4.3.2. Diagnostic Analysis

Diagnostic analytics is an analytics type that helps users comprehend what is occurring and the reason behind it so that, in case something goes wrong, an action can be implemented to solve the issue. Not just identifying the issues, diagnostic analysis uses methods like:

- Regression: how much the independent affect the dependent variable
- Root cause analysis, which uses the correlation between data to answer questions like "why did the sales fall?"

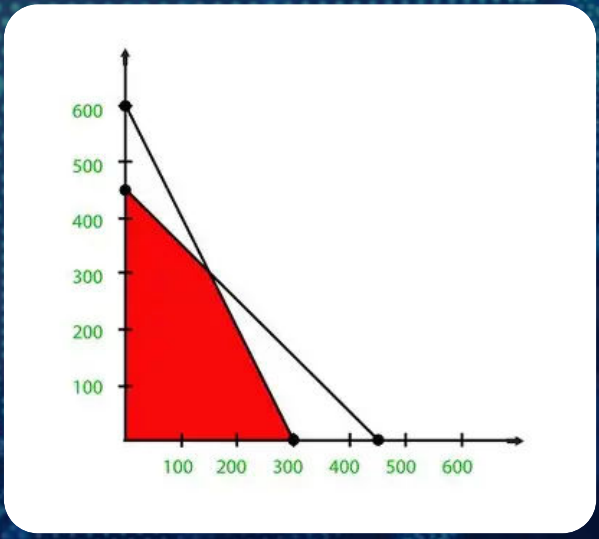
4.3.3 Predictive Analysis

A type of analytics that uses statistical techniques, machine learning algorithms, and historical data to make predictions about future outcomes. It looks at the data to see what might happen in the future. .

As a result, businesses can optimize resource allocation, improving efficiency as well as staying ahead of competition by forecasting demand and market trends

4.3.4 Prescriptive Analysis

Offers decision support so that the analysis's results can be used. Prescriptive analytics, then, proposes ways to extract the benefits and make use of the predictions, going beyond simply evaluating the data and forecasting future events. By choosing among the available options as efficiently as possible, it gives the organizations the best option when dealing with a business situation.



Linear programming

What if

It uses methods like optimization techniques (linear programming to find the best optimal choice) and simulation techniques (testing "what-if" scenarios to see how a decision performs under different situations).

4.4 Big Data Analytics Techniques

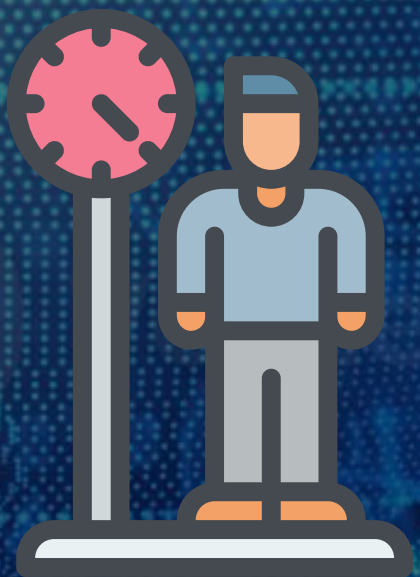
4.4.1. Quantitative Analysis

It is an analysis of quantitative data (dealing with numbers). Its main goal is quantification, meaning results from a sample are to be generalized to a wider population. The different types of quantitative datas:

Discrete data: Can only take certain values; it is always countable. For example, the result of rolling a dice only has the values of 1, 2, 3, 4, 5, and 6. Discrete data includes only values that can only be counted in whole numbers and are separate, which means they can't be broken down into fraction

Continuous data: Can take any value within a range; it is always measurable. Thus the data can be narrowed down to even decimals, as long as it's within the range.

Interval data: Scale where the intervals between values are equal, but there is no true zero. 0°C doesn't mean there is no temperature. While we can make measurements on the value difference, we can't on the proportions. 20°C is hotter than 10°C, but doesn't mean its twice as hot.



Ratio data: Like interval data, but includes a true zero point. And also, we can measure both the difference and the ratio. For example, we can say that a person who weighs 80kg is twice heavier than a person who weighs 40kg. Ratio allows all arithmetic operations whereas interval can only work on addition and substraction

4.4.2. Qualitative Analysis

Deals with non numerical data and focuses more on understanding "why" and "how" behind the events.

Type of qualitative data:

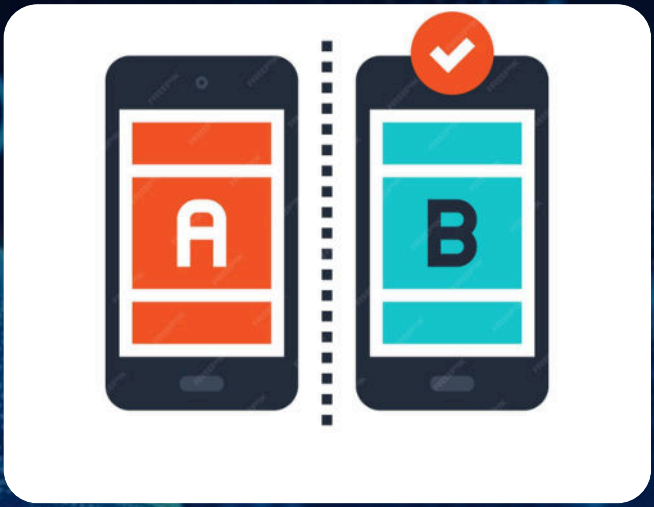
- Ordinal data – Based on ranking or order, not on the exact difference (e.g., ratings, happiness scale).

- Nominal data – Categorical with no numeric meaning; cannot be used in arithmetic (e.g., gender, tall/short).

Type of qualitative analysis:

- Grounded theory – Develops general theories from individual case studies.
- Narrative analysis – Focuses on reinterpreting people's stories or interview data.
- Content analysis – Classifies and summarizes data;

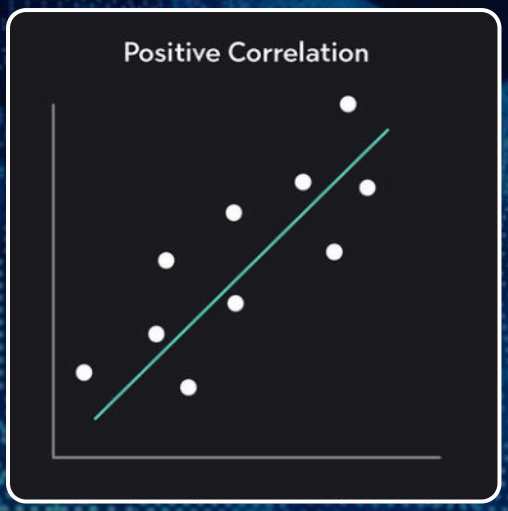
4.4.3. Statistical Analysis



Using statistical methods for analyzing data, comparing two variables to identify relationships. One of the statistical techniques includes A/B testing, which compares two versions of something to see which one is better. Version A is kept as the control version, whereas version B as the modified version (usually a version of the same content with one variable changed). To determine the successful version, both versions will be tested simultaneously. For example when an e-commerce website's versions are compared, a version with more buyers will be considered successful.

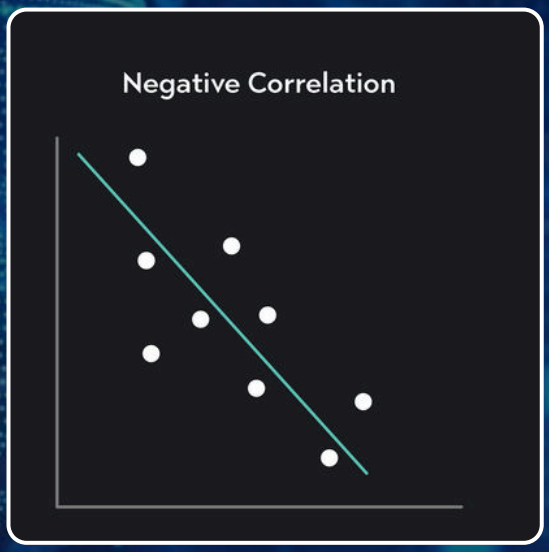
Another statistical technique involves correlation—a method used to determine if there exists a relationship between two variables, usually by constructing a scatterplot. They can be divided into positive, negative, and no correlation. A correlation coefficient like Pearson's (r) can be used to find the strength of correlation

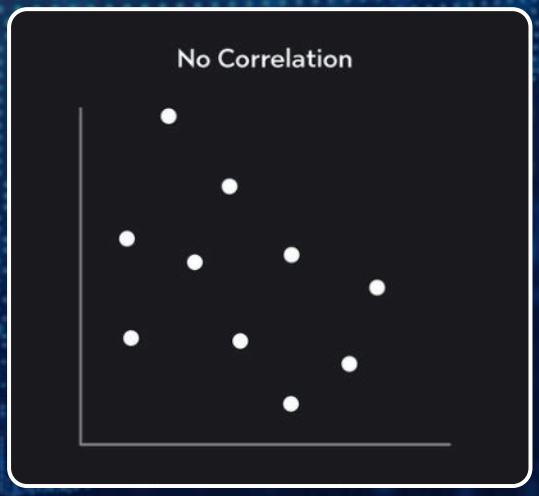
it takes values ranging between -1 to + 1 where values closer to 1 indicate a strong correlation, whereas those close to 0 indicate a weak correlation.



Positive correlation: when two variables move in the same direction—an increase in one causes an increase in the other.

Negative correlation happens when two variables have opposite relationships, so when one variable increases, the other variable decreases.





No correlation is quite self-explanatory - there won't be any change on the other variable even if one increases or decreases.

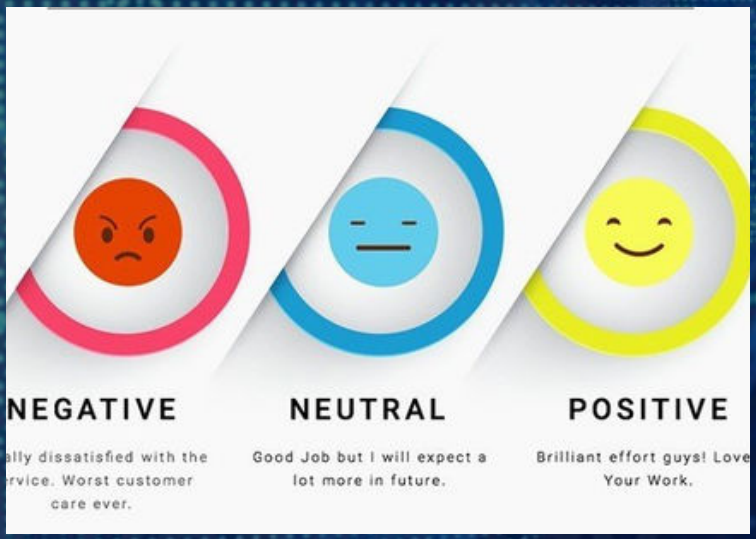
Similar to correlation, a regression technique can be used to find patterns between variables by using a linear equation (Y = a + bX). But unlike correlation, there is a degree of causation in regression; thus, the numerical value of the dependent variable can be determined from the independent.

4.4.4 Semantic Analysis

Semantic analysis involves extracting meaningful information from speech and text data. It seeks to understand how individual words and their combinations convey meaning, which is essential for applications in Natural Language Processing (NLP)—a field of AI that helps computers understand human speech and text. NLP operates through 4 key stages:

- The lexical stage involves breaking down input text data into paragraphs, sentence and words then analyzing their structures.
- The syntatic stage focuses on analyzing grammar and fixing them into the right format.
- The semantic stage checks the text or speech for meaningfulness by interpreting the actual meaning.
- Finally, the pragmatic stage looks deeper on the underlying meaning to interpret what actual message is delivered.

4.4.5. Sentiment Analysis



Sentiment analysis uses the process of looking at texts (reviews) and figuring out whether the attitude is positive, negative, or neutral.

Its goal is to understand what people think about a particular situation or incident and why they feel that way.

4.5 Big Data Business Intelligence

Business intelligence is the process of analyzing data to help businesses in decision-making. From the data analytics, the business may benefit from increase in revenues and productivity. BI data includes both data from the storage (previously stored data) and data that are streaming.

4.5.1 Online transaction processing (OLTP)

Online Transaction Processing (OLTP) refers to a system created to handle and manage transaction-based applications, which are managed in real time rather than in batches. This means the system is used in

situations where immediate responses to users are required. For example, OLTP is used in ATMS, retrieving and displaying sets of records like withdrawals, deposits, and transfers quickly for users.

A single OLTP system can accommodate thousands of users simultaneously, handling both simple and complex transactions. The transactions usually take only a few seconds to complete. With that being said, it's important for OLTP systems to maintain data integrity, especially when multiple users access the system at once.

4.5.2. Online Analytical Processing (OLAP)

Online analytical processing (OLAP) handles and analyzes massive amounts of data, focusing more on complex queries. Unlike OLTP, which manages high volumes of transactions, OLAP systems lean toward in-depth analysis on heavier data collections. They are not analyzed frequently—the data are collected over a period of time, thus in batches. The OLAP system can be summarized as Fast Analysis of Shared Multidimensional Information (FASMI).

- Fast refers to the system's speed at delivering quick responses
- Analysis refers to its ability to create strong analytics without the use of programming

- Shared means the system can be used by multiple users at the same time without issues. While allowing sharing, the system must still protect confidentiality and manage concurrent access — meaning, if several users try to make changes at once, it must prevent data conflicts or corruption.

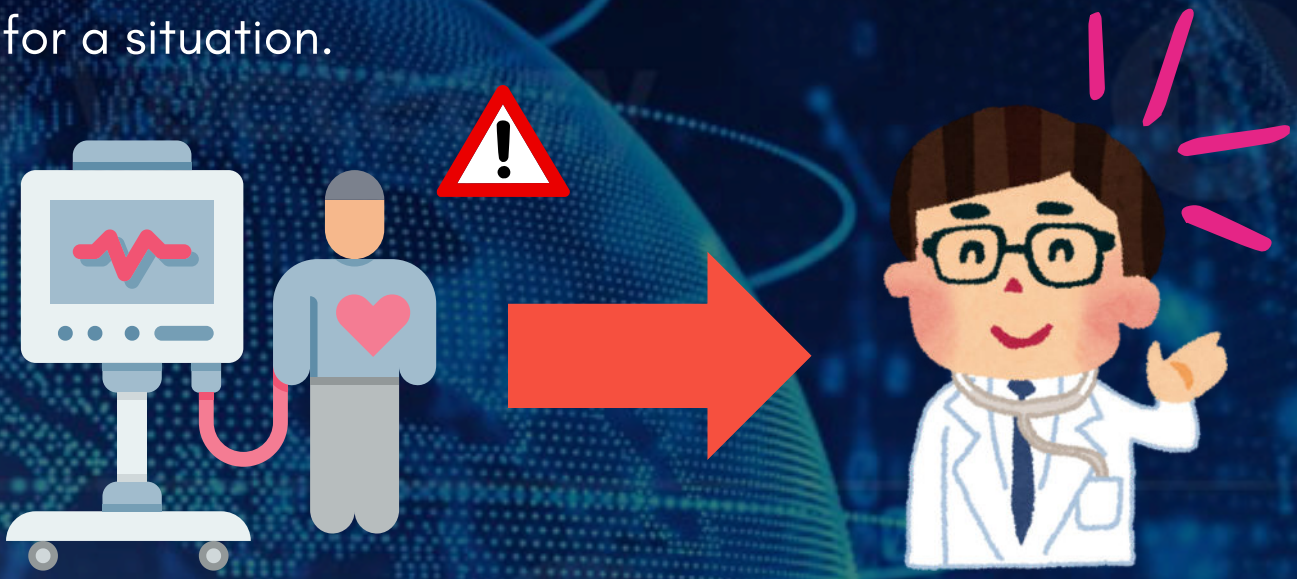
- Multidimensonal means viewing data from multiple perspectives or dimensions (like time, region, product, etc.).

- Information refers to the system's ability to process and manage very large datasets that come from a data warehouse.

4.5.3. Real-Time Analytics Platform (RTAP)

Real-Time Analytics Platform (RTAP) combines the analytical capabilities of OLAP with real-time or near real-time data updates. It can analyze complex queries immediately as they are generated; it is often used when data sources are streaming.

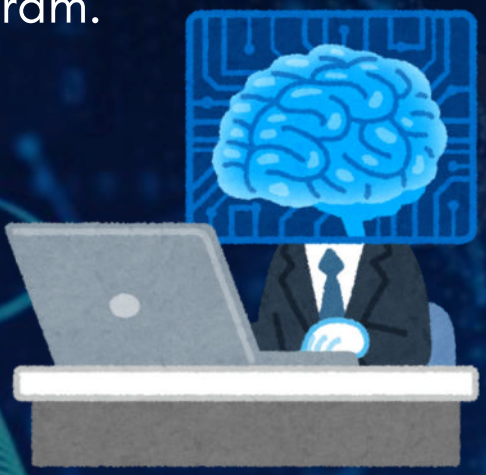
As a result, it can be used in hospital, particularly in monitoring systems. Incoming data like heart rate and oxygen levels are analyzed instantly, and then alerts are sent out to doctors to take a recommended action for a situation.



4.6 Machine Learning

Machine learning is a field that lies at the intersection of artificial intelligence (AI) and statistics and is the ability of a system to learn and improve understanding through experience. As data volumes continue to grow rapidly, it has become essential to create effective machine learning algorithms for a wide range of technological applications.

Today, machine learning has been integrated into our daily activities, ranging from features like video recommendations and product suggestions to listing the friends we may know on Instagram.



Basically, a machine learning algorithm is a program created for pattern recognition and great decision-making. It allows machines to observe data, identifympatterns, and learn over time. By training machines to make decisions in a similar way to how we humans do. They can solve logical problems like differentiating objects and categorizing them.

4.6.1. 4 Stages of Machine Learning

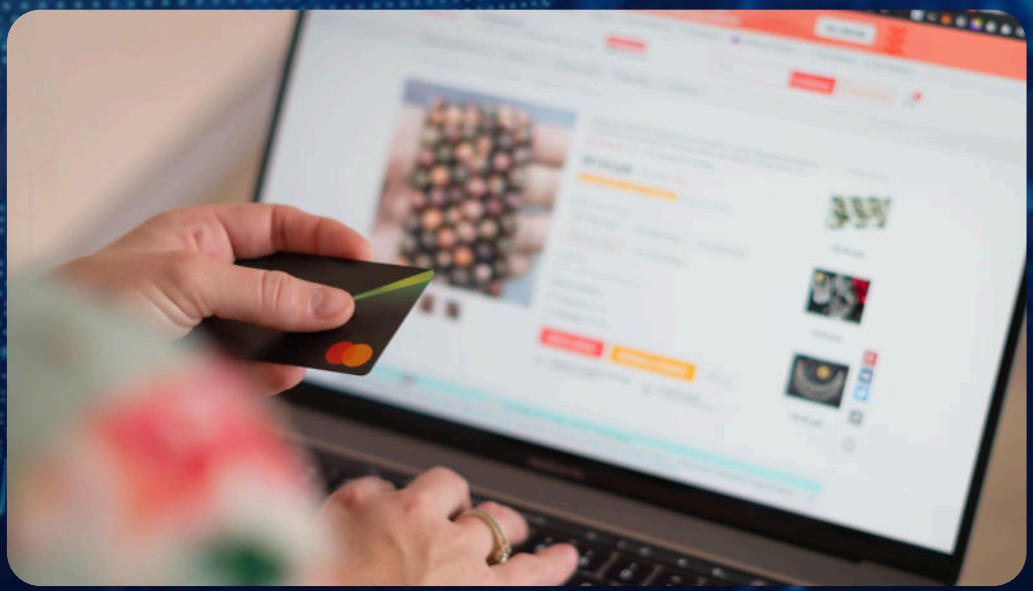
1. Testing phase: A training data set is used to train machines so they can recognize patterns between input and outputs.
2. Validation phase: The validation data set is used to evaluate the performance of the model. A few adjustments will be made to fix errors.

3. Testing phase: the final model is assessed on how it performs on completely new data. By using this new data, predictions can be made, in which the result will be compared to the expected output to measure accuracy.

4. Application phase, the model is used for real-world situations, processing actual data.

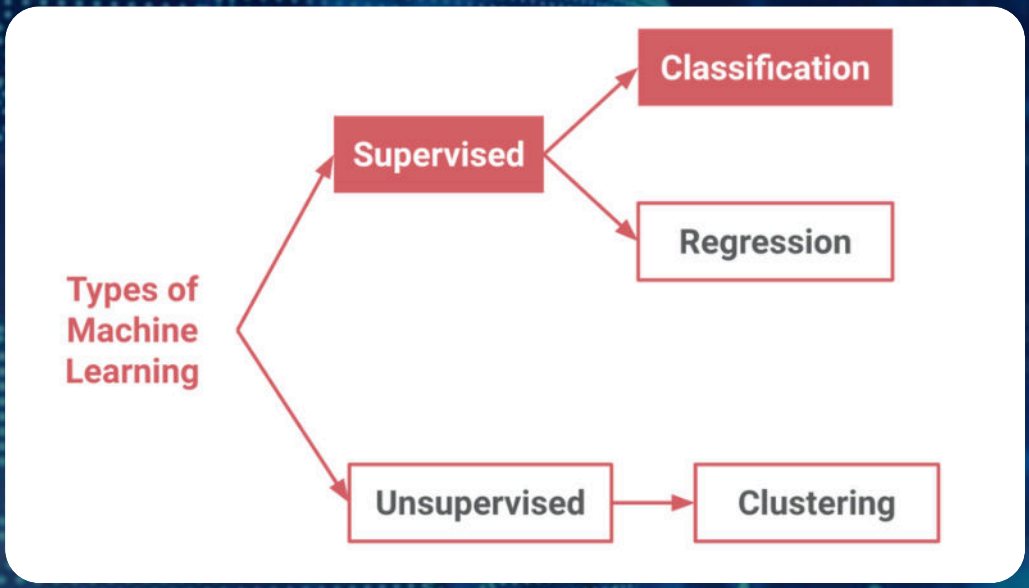
4.7 Machine Learning Use Cases

- Product recommendation: Amazon uses ML to create the recommended product list for consumers. In this type of system, the user behaviors are learned over time from what they searched or bought in the past. Then the products that the users might be interested in are predicted.



- Spam detection: Email service providers use ML algorithms to detect spam. Based on predefined rules, the algorithm can categorize mail as spam or not. If it is, then it's moved to the spam folder instead of the inbox.

4.8 Machine Learning Use Cases



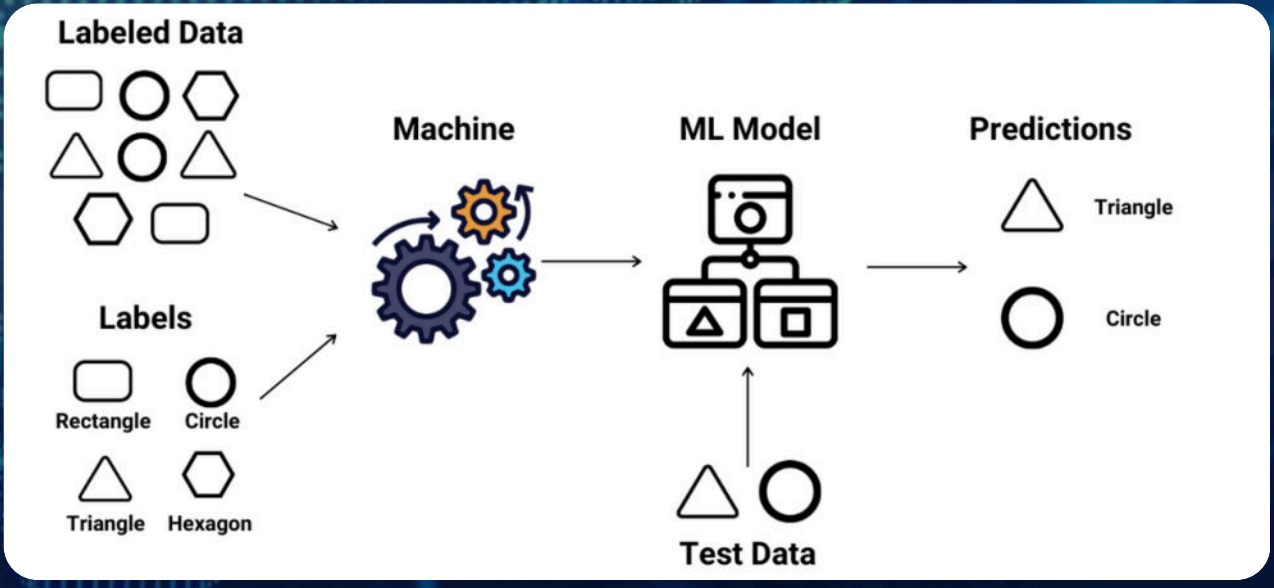
4.8.1. Supervised Machine Learning

In supervised machine learning, the computer learns from a training set, which is built from both inputs and outputs. This model learns the relationship between inputs and outputs so it can predict results from new data. Supervised ML can be divided into two categories:

4.8.1.1. Classification

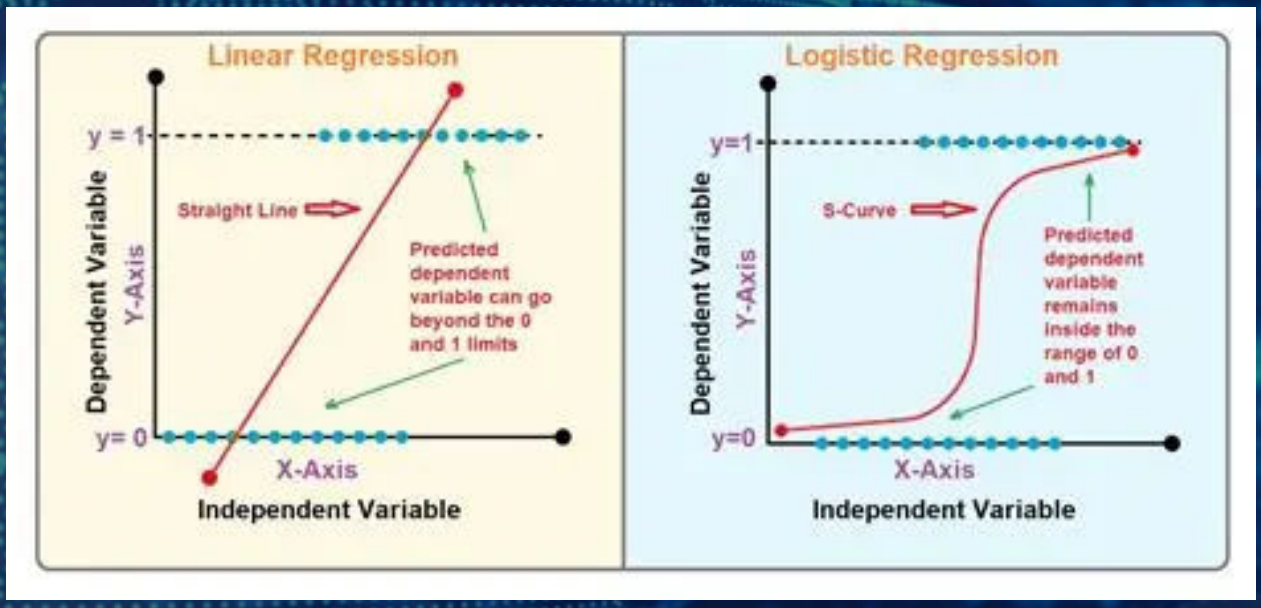
Classification is an ML tool to identify groups based on certain attributes, which is then used to classify them into existing groups. It involves a training data set so that each example is already labeled with the corresponding category. The output is discrete, meaning it falls into one of the predefined classes.

For example: The model is trained using pictures labeled "triangle" or "circle." During the training, it learns the visual characteristics that make them distinguished, such as shape, texture, and edges. Then, when shown a new image, the model analyzes its features and predicts whether it's a triangle or circle.



4.8.1.2. Regression

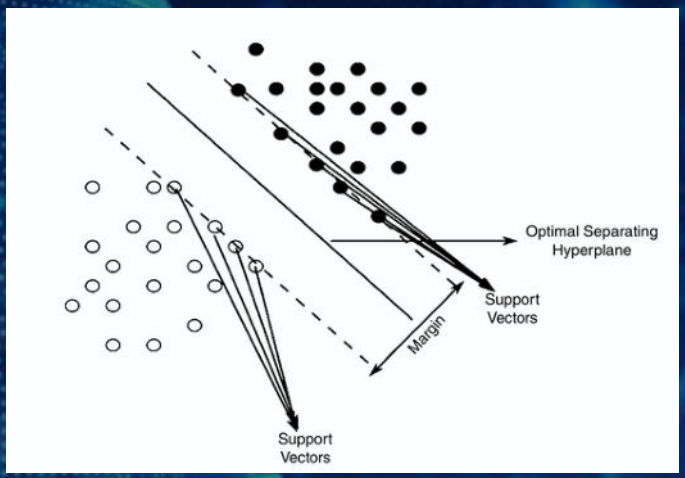
Regression is an ML tool to predict a numeric value for a given input from a continuous set of data. In classification, the model learns from known relationships (patterns learned from labeled data). Whereas in regression, the model is used to find the best mathematical relationship between input and output.



By using a linear regression, a function can be plotted to find the relationship between independent and dependent variables. Another regression that can be used is logistic regression, which aims to find the best-fitting model.

4.8.1.3. Support Vector Machines (SVM)

SVM develops a model with the training data set where the data points that belong to different groups are separated by a gap. The main goal of SVM is to find the optimal "gap" (or called the hyperplane) that best divides the classes.

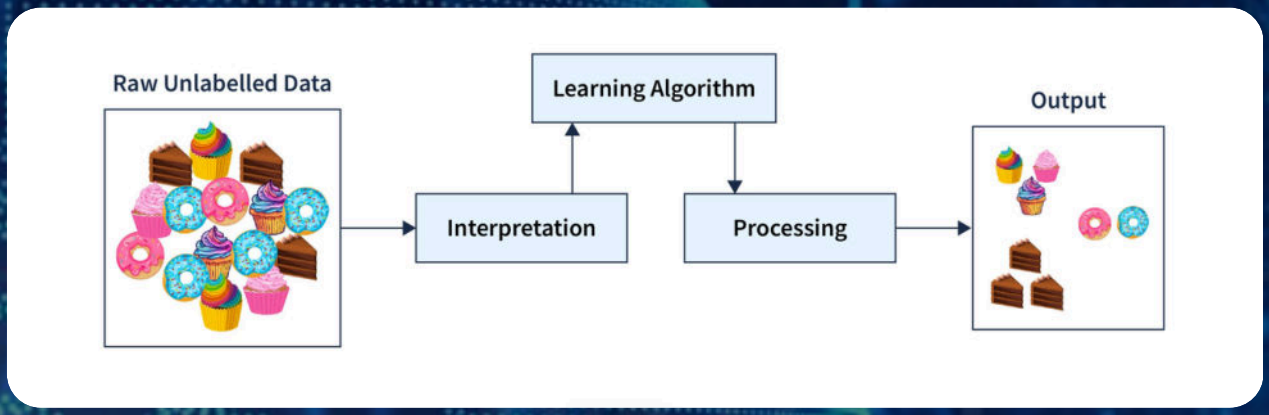


It also tries to maximize the distance between the closest point of each class and the hyperplane (margin). The data points that lie on the margin are the support vectors. The accuracy of the classification increases with the increase in margin, as there is better generalization.

Linear SVM can be separated with a straight line or hyperplane, whereas nonlinear SVM has to be separated using a kernel to map data into a higher-dimensional

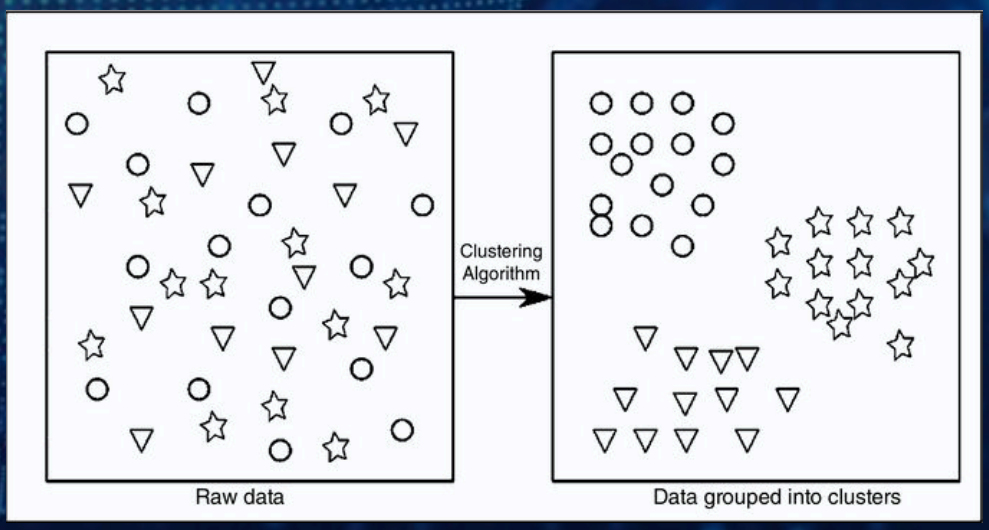
4.8.2. Unsupervised Machine Learning

Technique where input data has no labels, meaning there is no training set to predict output. So the algorithm learns on its own by finding hidden patterns and structures in the data.



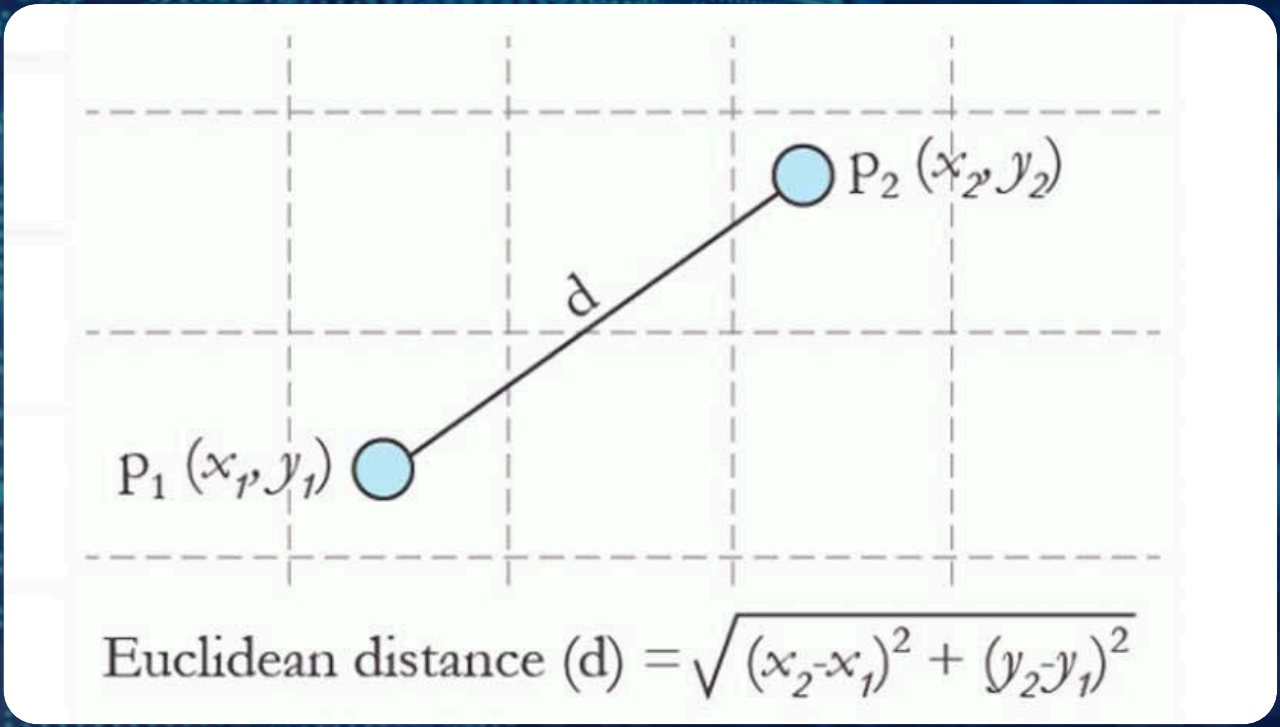
4.8.2.1 Clustering

A common technique is clustering, where similar data are clustered based on similarities in characteristics. There are two ways to measure the similarity, namely:



Correlation: measures how closely related the patterns of data are.

Distance: measures how physically close the data points are to each other. The closer the points, the more similar they are. Ways to calculate the distance may vary; one that is common is the Euclidean distance, which is the straight-line distance between two points. It can be computed using the formula below:



Exercise

Multiple Choice Questions

1. What is the main goal of predictive analysis?
A.Summarizes past data
B. Understanding the reason behind past events
C. Predicting future outcomes
D. Giving recommendations for the best course of action

2. Which statistical method compares two versions of something to determine which one is better?
A. Correlation
B. A/B testing
C. Regression
D. Classification

3. Which analytics technique is used for extracting meaningful informations from words and phrases?
A. Statistical Analysis
B. Quantitative Analysis
C. Qualitative Analysis
D. Semantic Analysis

4. Recording deposits, withdrawals, and transfers in a banking system is an example of _____.
A. Online transaction processing (OLTP)
B. Online analytical processing (OLAP)
C. Real time analytics platform (RTAP)

5. In Support Vector Machines (SVM), what happens if the data is not linearly separable?
A. The model fails instantly and cannot be separated by any other method.
B. SVM uses a kernel trick to map data into a higher-dimensional space.
C. SVM removes data points.

True or False

1. In supervised learning, the model is trained without labeled data.
2. Descriptive analysis summarises past historical data.
3. Natural Language Processing (NLP) is a field of AI that is used for statistical analysis.
4. Correlation implies causation between two variables; the value of dependent variable depends on the independent variable.

Arrange the following steps of the Data Analytics Life Cycle in the correct order:

A. Data Analysis and Visualization
B. Data Preparation
C. Application
D. Business Evaluation and Identification of Sources
E. Data Transformation

Answers

Multiple Choice Questions

1. C
2. B
3. D
4. A
5. B
6. D

Arrange

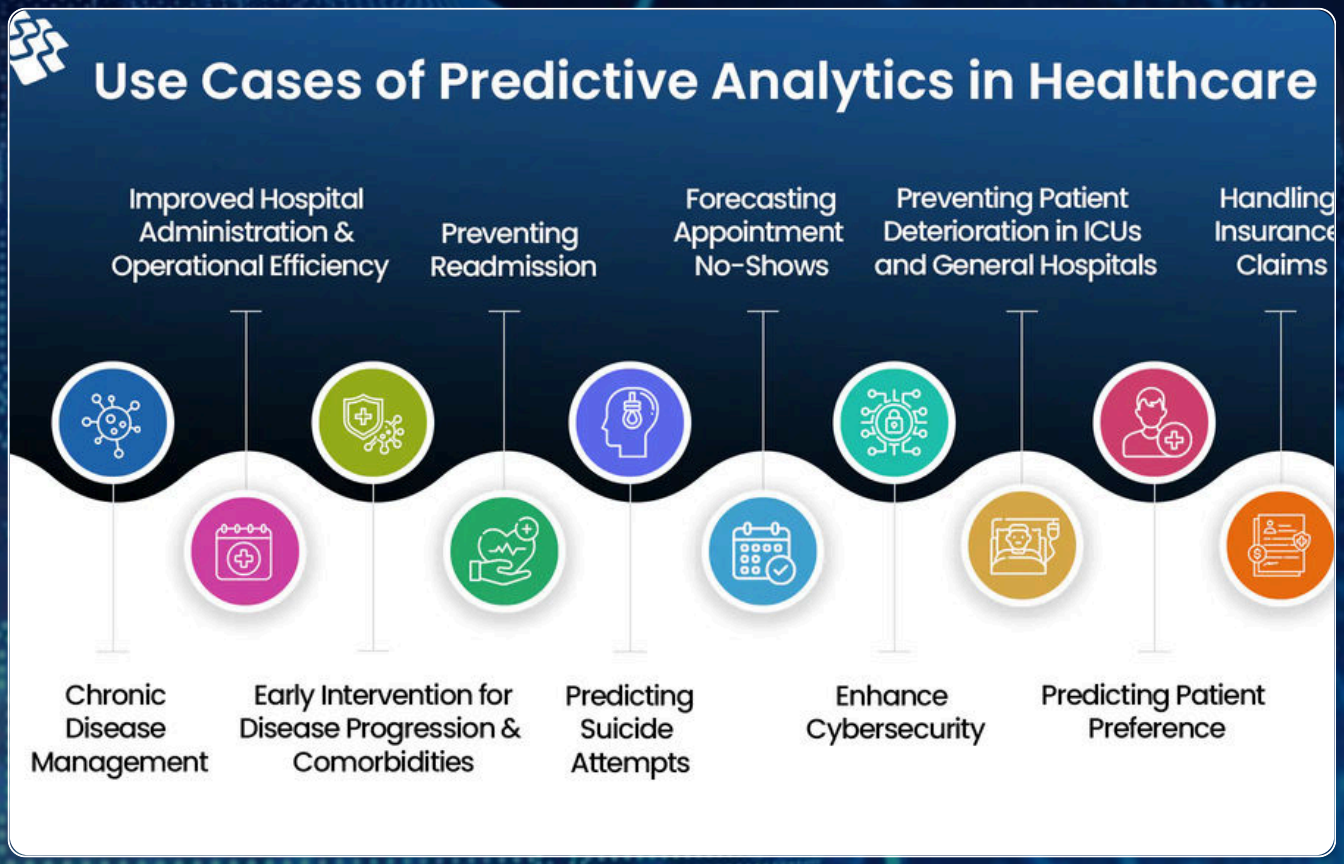
1. D
2. B
3. E
4. A
5. C

True/False

1. False
2. True
3. False
4. False

CH5 Applications Across Industries

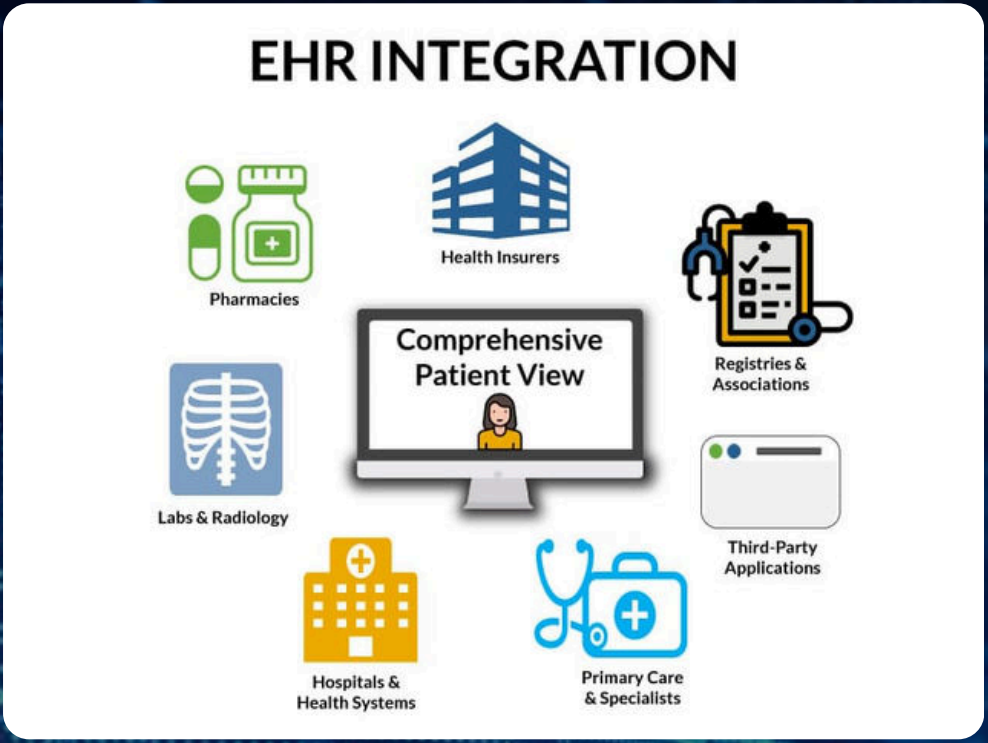
5.1 Healthcare



Big Data in healthcare refers to the massive volume of structured and unstructured health-related data generated from multiple sources such as hospitals, wearable devices, electronic health records (EHR), clinical trials, and insurance claims

Big Data allows for real-time decision-making, predictive modeling, and population-level insights that were previously impossible.

1) Electronic Health Records (EHR) Management



What it does

Organizes and stores comprehensive patient data digitally across systems.

How it works

Big Data platforms aggregate structured (lab results, prescriptions) and unstructured (doctor notes, scans) data into accessible, standardized systems.

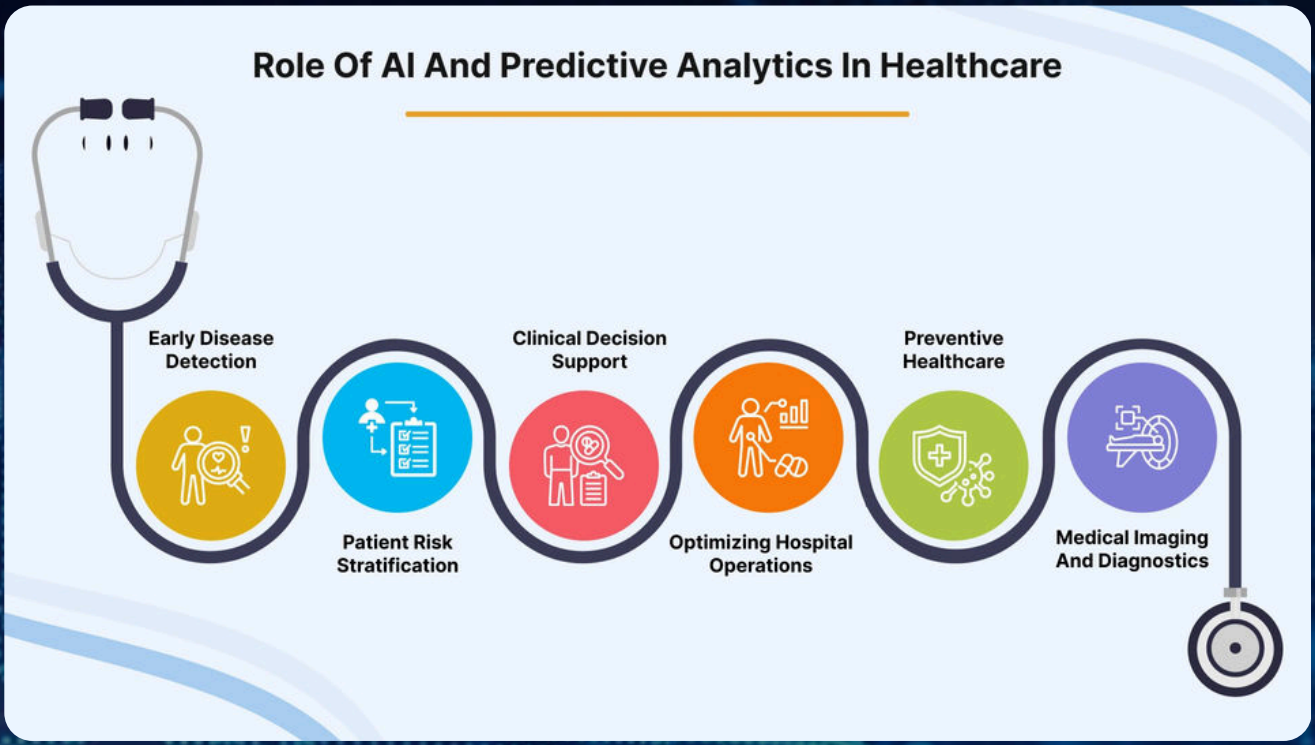
Example

Kaiser Permanente uses Big Data to unify millions of EHRs, enabling care coordination across locations.

Benefits

- Improves access to full patient history
- Reduces repeated tests
- Enables more informed clinical decision

2) Predictive Analytics for Patient Outcomes



What it does

Predicts future medical events like hospital readmission, sepsis risk, or disease onset.

How it works

Machine learning models trained on EHR + real-time patient data identify high-risk patients and trigger alerts.

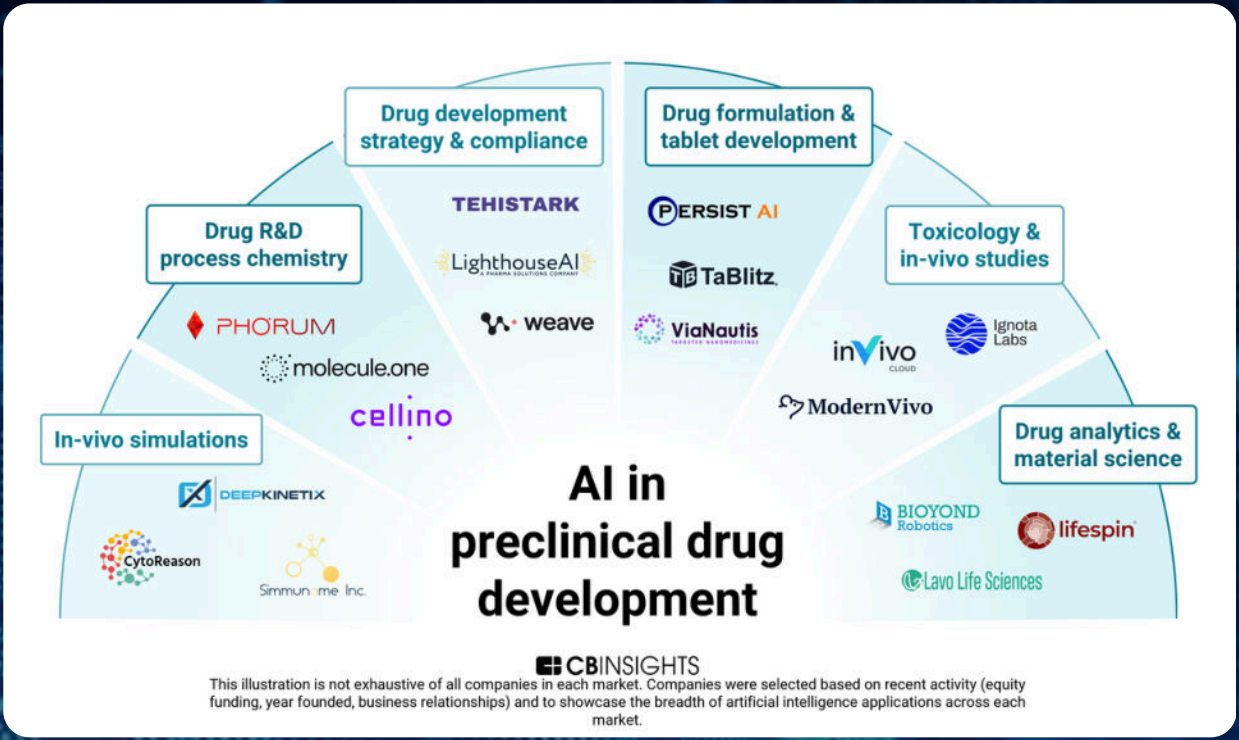
Example

Mount Sinai uses AI to predict sepsis 24–48 hours before symptoms become critical.

Benefits

- Enables early interventions
- Reduces mortality and ICU admissions
- Optimizes hospital resources

3) Drug Discovery & Clinical Research



What it does

Accelerates the identification of potential drugs and treatment targets.

How it works

Big Data combines genomics, patient outcomes, and lab research to discover patterns that traditional methods miss.

Example

The NIH "All of Us" program uses Big Data from over 1 million individuals for biomedical research and drug development.

Benefits

- Reduces time and cost of drug discovery
- Makes trials more inclusive and precise
- Leads to personalized medicine breakthroughs

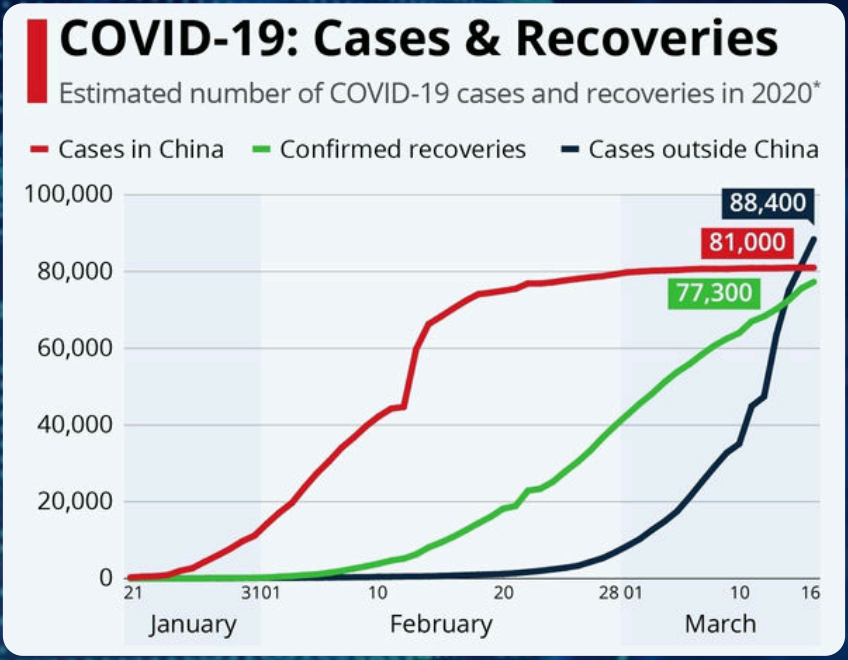
4) Population Health Management

What it does

Monitors trends across large populations to improve public health strategies.

How it works

Data from various sources (hospitals, public health systems, insurance) is analyzed to detect disease outbreaks or monitor vaccination rates.



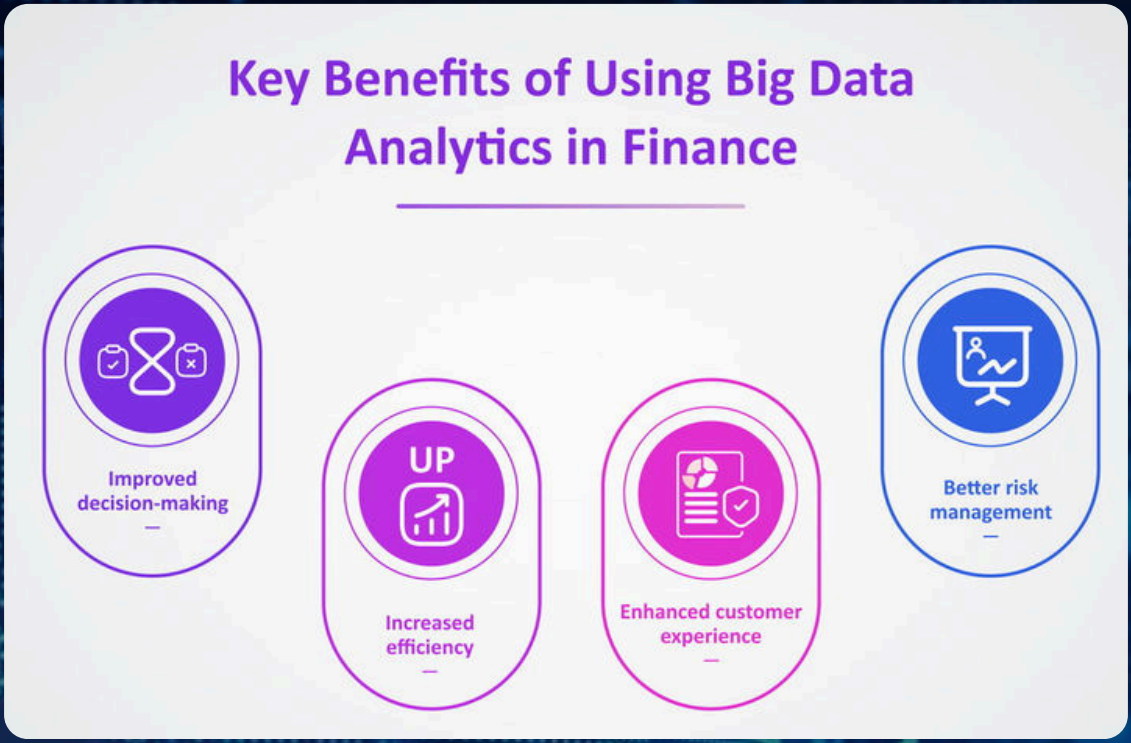
Example

During COVID-19, real-time Big Data dashboards tracked case numbers, ventilator availability, and ICU occupancy.

Benefits

- Informs policy decisions
- Improves healthcare equity
- Enhances emergency response planning

5.2 Finance



Big Data in finance refers to the use of massive, diverse, and fast-updating datasets to analyze market behavior, assess risk, detect fraud, and deliver personalized financial services.

1) Fraud Detection and Prevention

What it does

Identifies suspicious transactions by spotting patterns and anomalies in real-time.

How it works

Big Data systems process millions of transactions per second and apply machine learning to detect behavior that deviates from the norm.

Example

JPMorgan Chase uses Big Data platforms and AI to detect credit card fraud within milliseconds.

Benefits

- Reduces financial loss
- Improves trust and customer safety
- Detects fraud faster than human teams

2) Credit Scoring & Risk Assessment

What it does

Assesses a borrower's likelihood of repaying a loan by analyzing financial behavior and credit history.

How it works

Big Data models use non-traditional data (e.g., phone usage, shopping history) alongside traditional scores to evaluate creditworthiness.

Example

Fintech startups like Lenddo use Big Data to score users without formal credit history.

Benefits

- Expands credit access to underserved groups
- Improves accuracy of loan decisions
- Reduces default rates

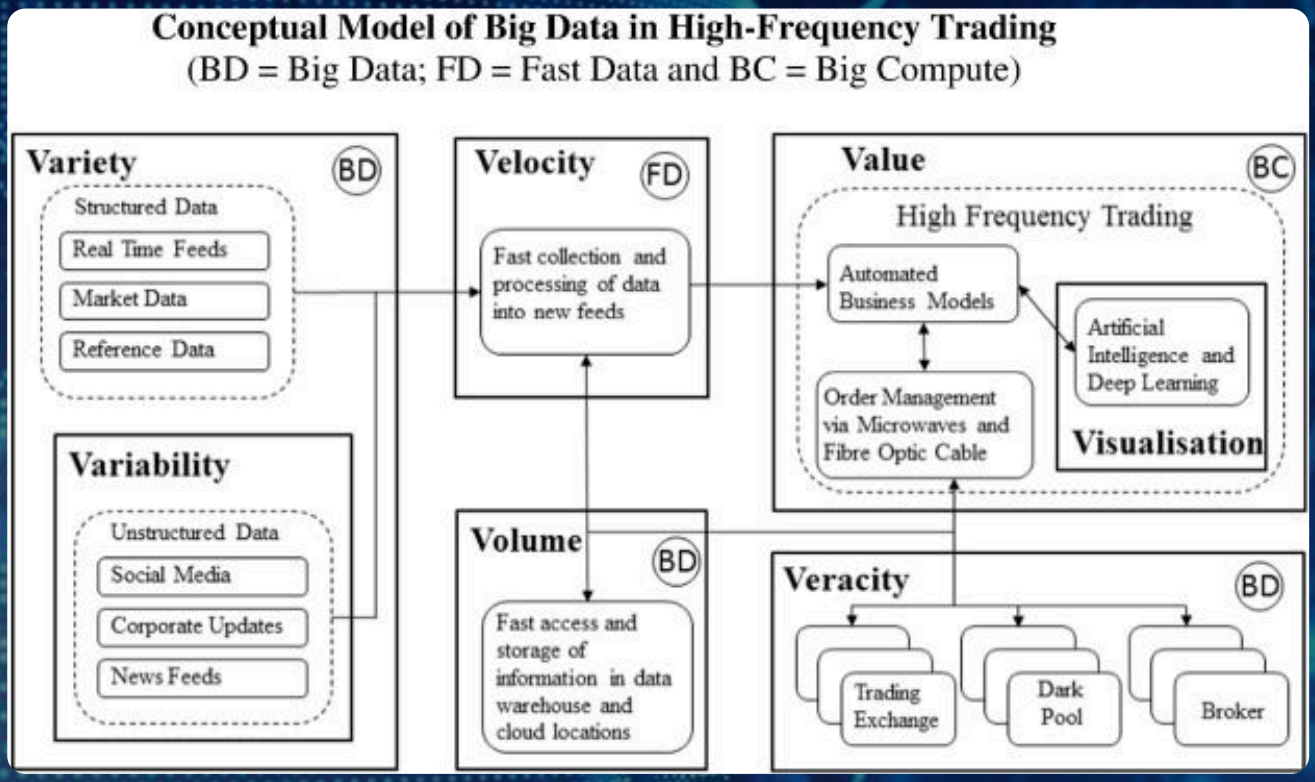
3) Algorithmic Trading

What it does

Uses Big Data to execute financial trades at high speed based on market conditions and predictive models.

How it works

Trading algorithms analyze market trends, news sentiment, social media signals, and financial indicators in real-time.



Example

High-frequency trading (HFT) firms use Big Data to execute thousands of trades per second based on price fluctuations.

Benefits

- Increases trading speed and efficiency
- Identifies market opportunities faster
- Maximizes profits for investors

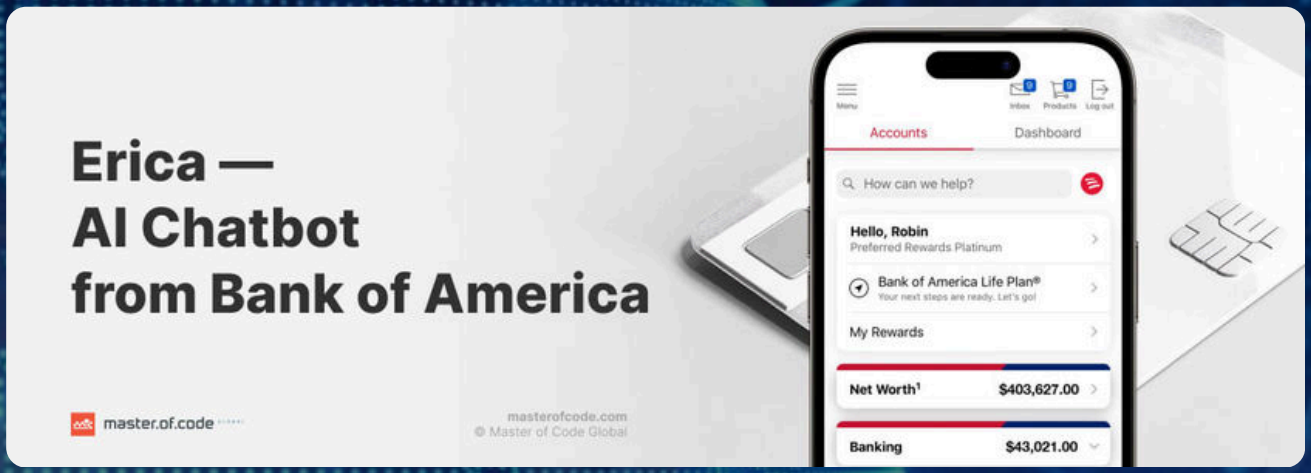
4) Personalized Banking Services

What it does

Delivers custom financial advice, product recommendations, and alerts to individual customers.

How it works

Big Data analyzes user behavior (spending, saving, preferences) to offer tailored experiences.



Example

Bank of America's AI-driven app "Erica" uses customer data to suggest budgeting tips and detect unusual account activity.

Benefits

- Enhances customer experience
- Increases engagement and loyalty
- Offers proactive financial advice

5.3 Retail

Big Data in retail involves collecting and analyzing large volumes of customer, sales, and operational data to optimize marketing, inventory, pricing, and the overall shopping experience.

1) Personalized Product Recommendations

What it does

Suggests products tailored to each customer based on their preferences and behavior.

How it works

Big Data systems analyze browsing history, past purchases, and similar customer behavior to recommend items.

Example

Amazon uses recommendation engines powered by Big Data to suggest products "You may also like."

Benefits

- Increases sales
- Enhances customer experience
- Boosts engagement and conversion rates

2) Inventory Optimization

What it does

Ensures the right products are available at the right time and place, reducing overstock and stockouts.

How it works

Big Data analyzes sales trends, location-specific demand, and supply chain factors to optimize inventory distribution.

Example

Zara uses real-time data to adjust inventory and restock stores based on customer demand.

Benefits

- Reduces inventory costs
- Prevents lost sales due to stockouts
- Improves supply chain efficiency

3) Demand Forecasting

What it does

Predicts what products will be in demand, when, and where.

How it works

AI models trained on historical data, seasonality, weather patterns, and promotional campaigns forecast future demand.

Example

Walmart uses Big Data to forecast item demand for large events like Black Friday.

Benefits

- Better inventory planning
- Minimizes excess stock
- Supports dynamic pricing strategies

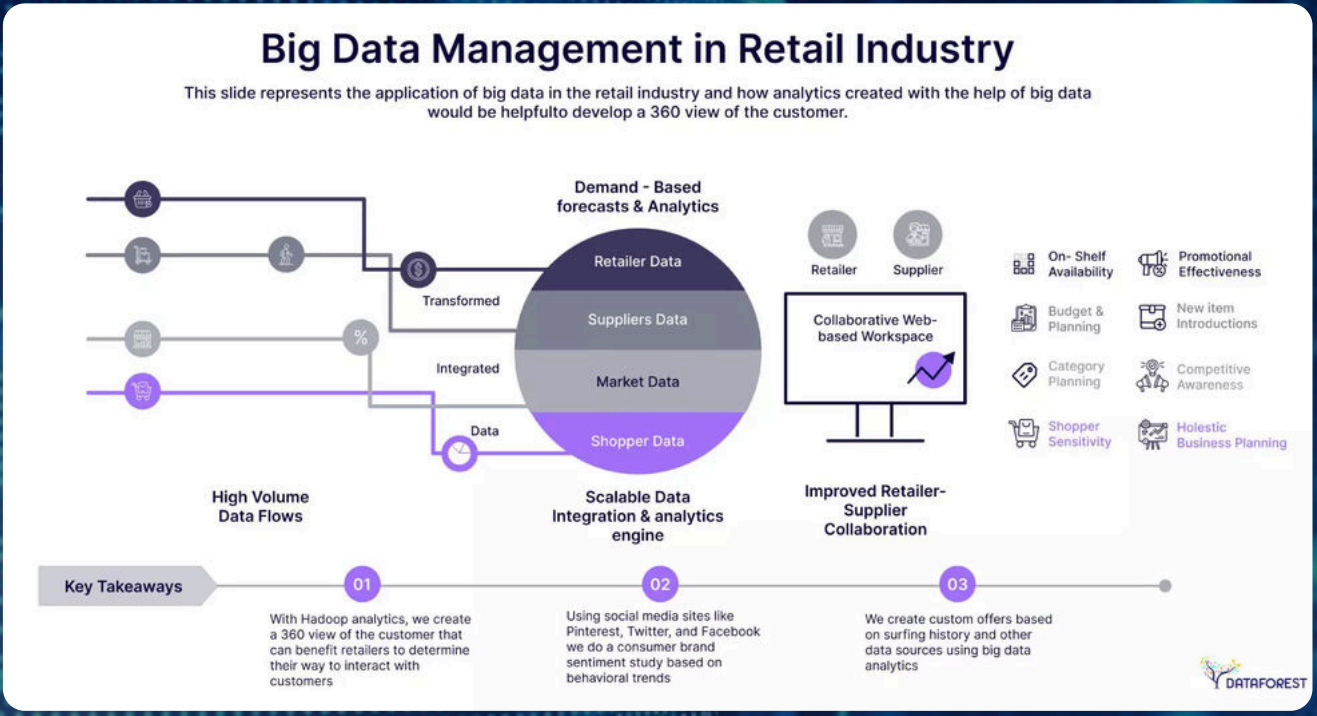
4) Customer Sentiment Analysis

What it does

Extracts insights from customer reviews and social media to understand public perception.

How it works

Natural Language Processing (NLP) models analyze feedback text to determine overall satisfaction or detect complaints.



Example

Retailers use Big Data to detect emerging trends or product issues via online reviews.

Benefits

- Improves product development
- Refines marketing strategies
- Enhances brand reputation management

5.4 Logistics

Big Data in logistics refers to the use of large-scale, real-time data from supply chains, vehicles, sensors, and customer interactions to optimize delivery, warehousing, and transportation processes.

1) Route Optimization

What it does

Identifies the most efficient delivery routes based on real-time conditions.

How it works

Big Data systems process GPS data, traffic updates, weather, and delivery history to continuously improve route planning.

Example

FedEx uses Big Data to reroute drivers during peak traffic to ensure timely deliveries.

Benefits

- Reduces fuel costs
- Shortens delivery time
- Enhances customer satisfaction

2) Predictive Maintenance

What it does

Detects early signs of equipment failure in trucks, aircraft, or warehouse machinery.

How it works

Sensor data is analyzed using predictive models to flag abnormal patterns that signal upcoming malfunctions.

Example

DHL uses predictive analytics to schedule truck servicing before breakdowns occur.

Benefits

- Minimizes downtime
- Avoids costly repairs
- Increases vehicle and machine lifespan

3) Warehouse Management

What it does

Optimizes storage space, picking paths, and inventory placement inside warehouses.

How it works

Big Data tracks item movement, demand frequency, and worker patterns to reorganize layouts and speed up operations.

Example

Amazon's fulfillment centers use Big Data to assign items to the most efficient bin locations.

Benefits

- Increases processing speed
- Reduces picking errors
- Enhances warehouse efficiency

4) Supply Chain Visibility

What it does

Provides real-time tracking and transparency of goods moving through the entire supply chain.

How it works

Big Data integrates IoT sensors, RFID, and shipment logs into centralized dashboards for managers.

Example

Maersk uses Big Data platforms to track global shipments and detect delays early.

Benefits

- Better decision-making
- Improved partner coordination
- Faster response to disruptions

Other Sectors Using Big Data

1) Energy

Forecasts electricity demand, optimizes grid efficiency, and reduces energy waste.

2) Manufacturing

Enables predictive maintenance and real-time process monitoring on factory floors.

3) Agriculture

Uses satellite and sensor data to guide irrigation, fertilizer use, and yield prediction.

4) Entertainment

Recommends content and analyzes viewer behavior for personalized experiences.

5) Public Sector

Tracks urban mobility, public health, and crime patterns to guide policymaking.

1. Which of the following best illustrates Big Data in the logistics sector?
A. Predicting heart disease from medical records
B. Using AI to detect fraud in online banking
C. Tracking global shipments with real-time data
D. Recommending movies on streaming platforms

2. True / False
a. Big Data in healthcare can help doctors predict future patient outcomes based on historical records.
b. In retail, Big Data is mainly used for maintaining IT infrastructure and has no role in product recommendation.

3. Multiple Answers (Select all that apply)
Which of the following are real-world applications of Big Data?
A. Detecting fraudulent credit card transactions
B. Optimizing irrigation and crop yield prediction
C. Managing inventory in hospitals
D. Recommending personalized products online
E. Printing physical bank statements in bulk

4. Short Answer

a. Explain one way Big Data improves warehouse operations.

b. How does Big Data benefit the energy sector?

c. Give one example of how Big Data supports customer experience in retail

5. Fill in the Blanks

Complete the paragraph using the words below:

(energy, fraud, inventory, sentiment, patients)

Big Data helps hospitals monitor _____ in real time, while retailers use it to manage _____ efficiently. In finance, it's used to detect _____. In entertainment, it supports _____ analysis. The energy sector uses Big Data to forecast future _____ usage.

6. Fill in the Blanks

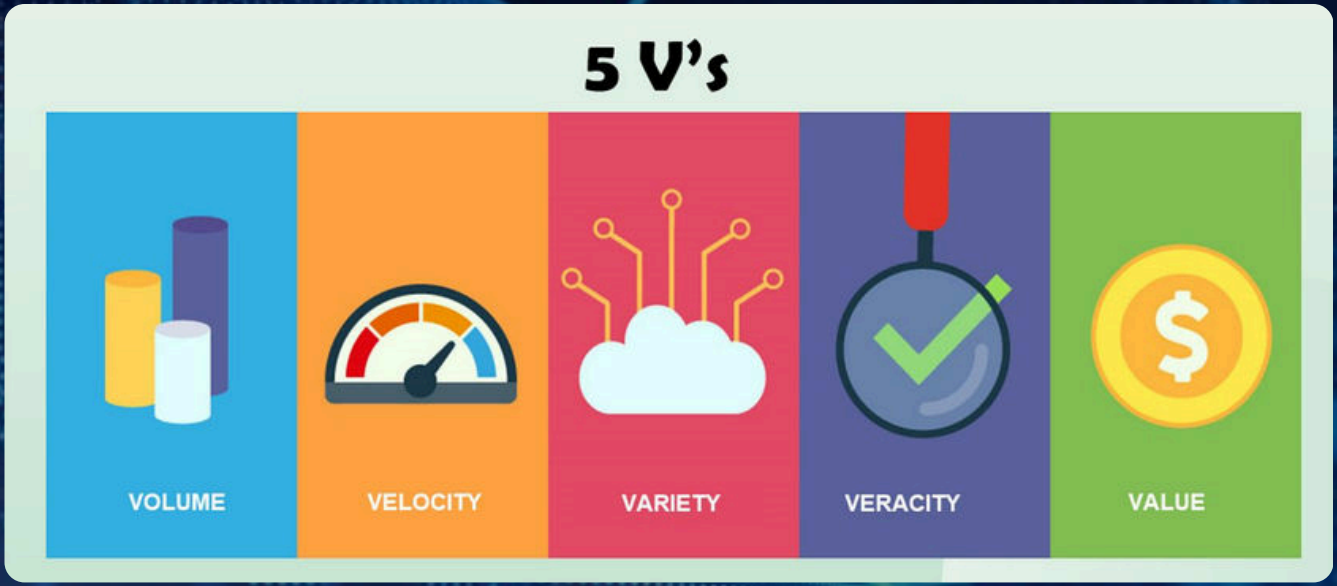
Fill in the blanks using the words below:

(forecasting, automation, supply chain, behavior, accuracy)

Big Data enhances retail by analyzing customer _____, improving recommendation _____, and supporting demand _____. In logistics, it improves _____ in delivery and provides visibility across the entire _____.

CH6 Big Data Challenge, Security and Ethics

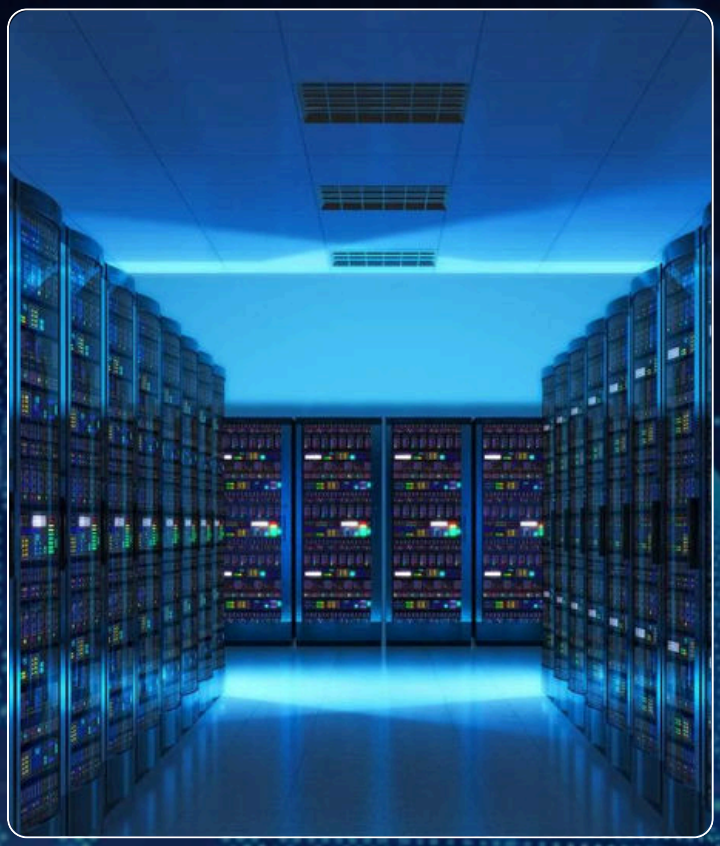
6.1 Big Data 5Vs Challenges



1) Big Data Volume Challenges

Volume in Big Data Challenge refers to the enormous amount of data generated every second from multiple sources including social media, sensors, transactional systems, and more. This massive scale of data creates significant issues in storing, managing, and processing it effectively.

Traditional storage systems often become inadequate as data grows exponentially. For instance, organizations must transition to scalable cloud storage solutions like AWS, Azure, or Google Cloud that offer elastic capacity to accommodate petabytes or even exabytes of data.



Managing and analyzing large volumes require advanced distributed computing frameworks such as Apache Hadoop and Apache Spark. These frameworks allow data to be processed in parallel across clusters, speeding up computation and enabling real-time analytics.

Example

An e-commerce company processing millions of transactions daily needs to efficiently store this data, process it quickly to identify sales trends, maintain its integrity, and keep sensitive customer information secure. Without proper scalable architecture and tools, this volume could overwhelm traditional systems and affect business operations.

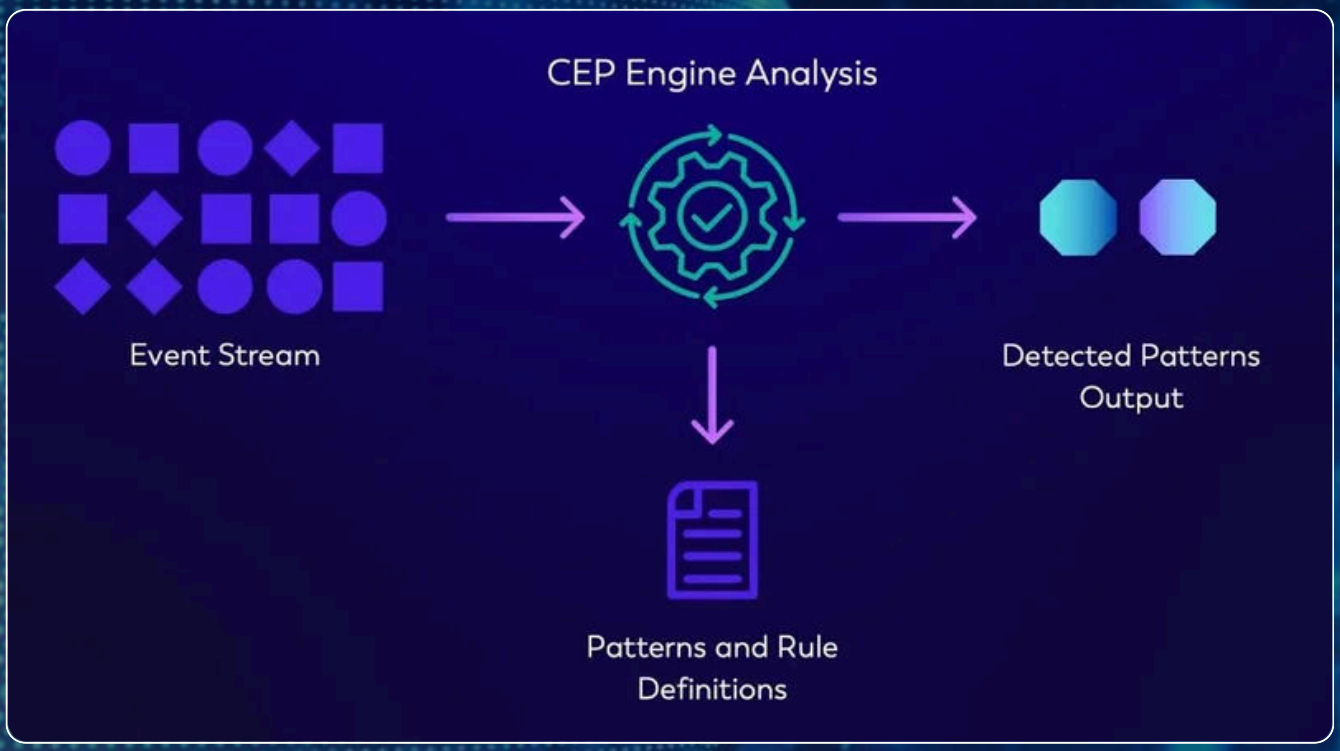
Solutions

adopting cloud-based storage for flexibility, implementing scalable distributed processing frameworks, employing data compression and lifecycle management policies, and designing future-proof architectures that can grow with data demands.

2) Big Data Velocity Challenges

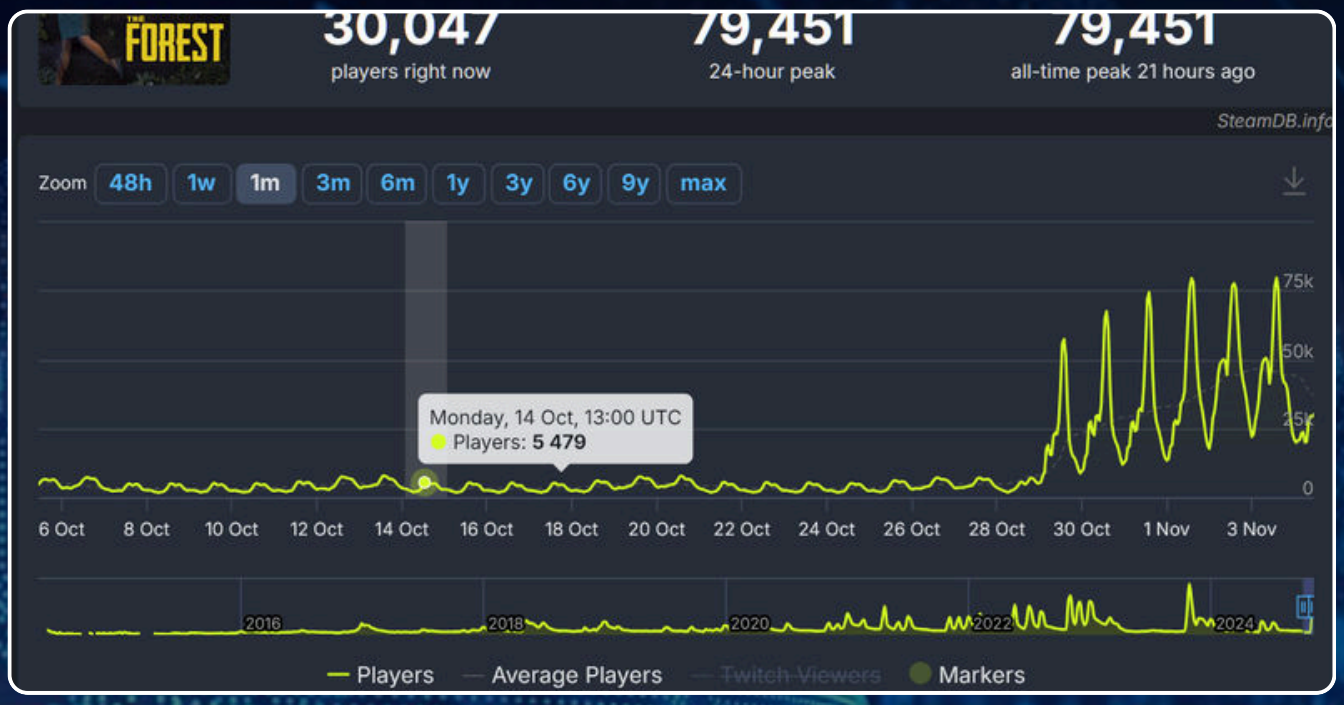
Big Data Velocity refers to the speed at which data is generated, collected, and needs to be processed. This rapid influx of data presents unique challenges for organizations aiming to extract timely insights and make real-time decisions.

a) Complex Event Processing



Rapid data streams often require complex event processing (CEP) to identify patterns, anomalies, or triggers in real time. For instance, IoT devices in smart factories generate sensor data that must be analyzed instantly to prevent equipment failures.

b) Scalability and Flexibility



Systems must scale dynamically to handle spikes in data velocity, such as during major online sales events or viral social media trends.

c) Latency and Throughput

Low latency (minimal delay) and high throughput (ability to process large volumes quickly) are essential. In-memory databases and stream processing frameworks like Apache Kafka, Apache Flink, and Apache Storm are often used to meet these demands.

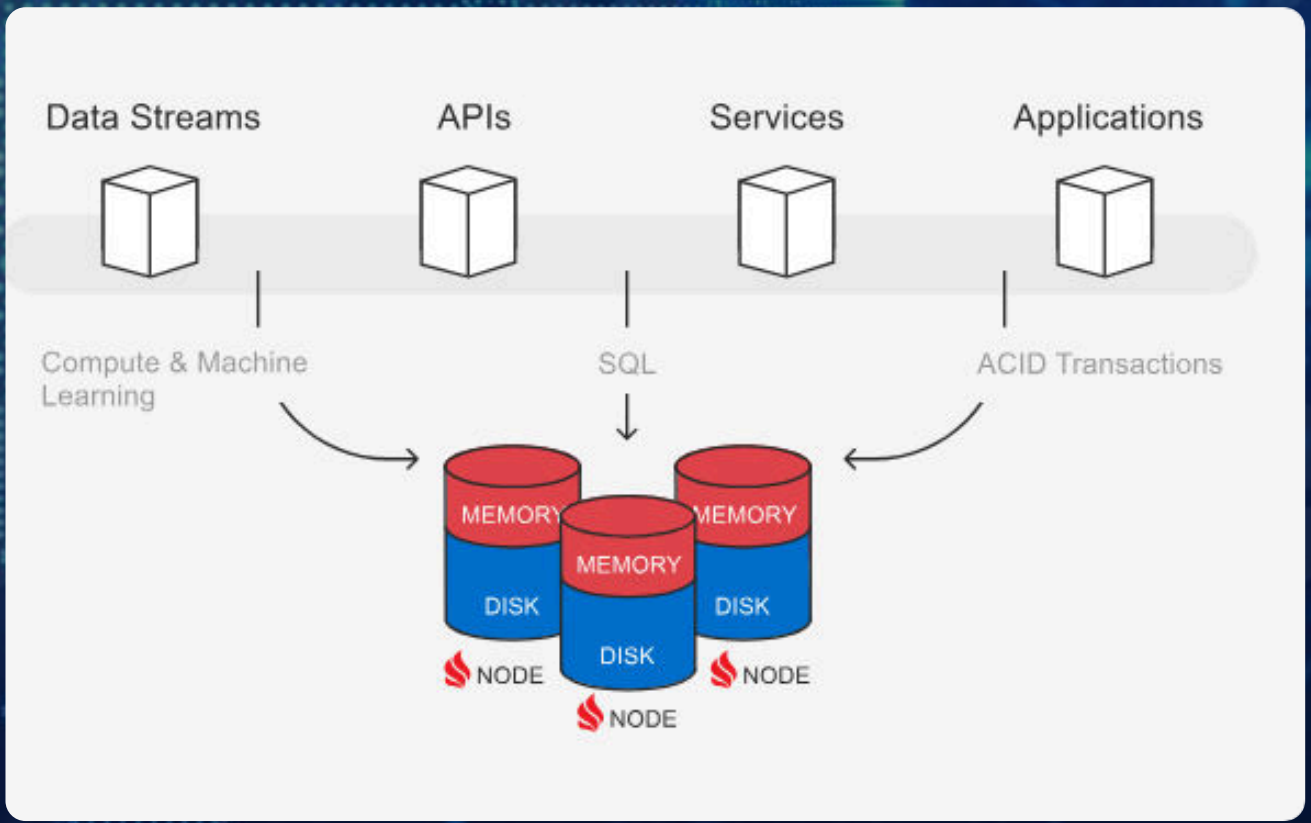
Solutions

1. Stream Processing Technologies

Tools like Apache Kafka, Apache Flink, and Spark Streaming enable real-time data ingestion and analysis.

2. In-Memory Databases

Technologies such as Redis and MemSQL provide fast data access and processing.



3) Big Data Variety Challenges

a) Diverse Data Types

Data comes in structured forms like relational databases, semi-structured formats such as XML or JSON, and unstructured data including text, images, audio, and video. This heterogeneity complicates handling and analysis.

b) Integration Complexity

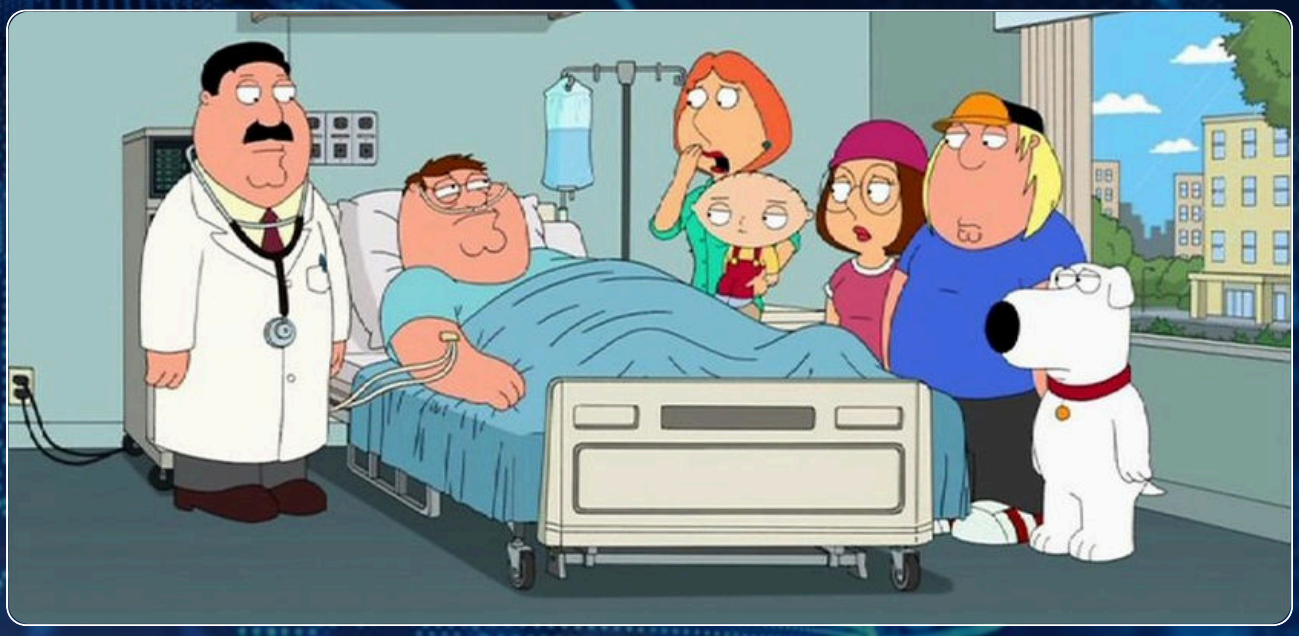
Combining data from different sources and formats into a unified platform requires sophisticated extraction, transformation, and loading (ETL) processes.

c) Storage Requirements

Different data types need different storage solutions—structured data fits well in SQL databases, while unstructured data requires NoSQL databases or object storage systems.

Real Life Examples

Healthcare



Hospitals manage structured patient records, semi-structured lab results, and unstructured doctor's notes and medical images, all needing integrated analysis.

Social Media

Platforms handle structured user profiles and connections alongside massive volumes of unstructured posts, photos, and videos.

Retail

E-commerce businesses merge structured sales data with unstructured customer reviews and semi-structured web clickstreams to understand consumer behavior.

Solutions

1) Use data lakes that can store all types of data in their native formats.

2) Employ ETL tools designed for complex data integration.

3) Adopt NoSQL databases and distributed file systems for unstructured data.

4) Leverage AI and machine learning to analyze unstructured content.

4) Big Data Veracity Challenges

Big Data Veracity refers to the accuracy, quality, and trustworthiness of data. It addresses the challenge of dealing with noisy, biased, incomplete, or inconsistent data that can lead to incorrect insights and flawed decision-making.

Challenges

a) Data Accuracy

Data may contain errors, outliers, duplicates, or missing values that reduce reliability.

b) Bias and Noise

Data sources may introduce bias or irrelevant noise, affecting analysis outcomes.

c) Inconsistent Data

Different systems or departments may use varying definitions or standards, causing conflicts.

d) Data Silos

Fragmented data stored in isolated systems leads to incomplete and inconsistent datasets.

Real World Examples

1. Inaccurate customer information can cause marketing campaigns to target the wrong audience, leading to wasted costs and lost revenue.
2. In healthcare, incomplete or erroneous patient data can jeopardize treatment decisions and patient safety.
3. In financial services, biased or inconsistent data can distort risk assessment models.

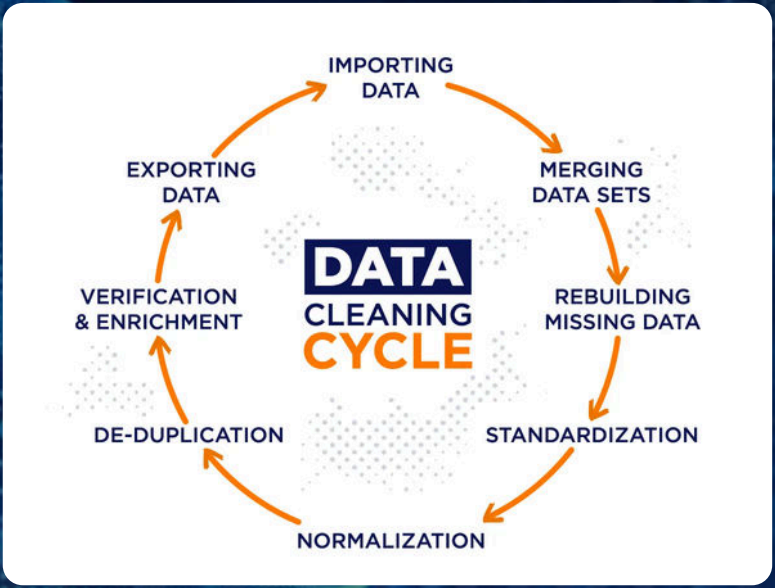
Solutions

1. Data Governance

Establish policies to enforce data quality, ownership, and accountability.

2. Data Cleaning and Validation

Use automated tools to detect and correct errors or inconsistencies.



3. Standardization

Create common data definitions and formats across systems.

4. Continuous Monitoring

Implement tools to track and measure data quality over time.

5. Integration

Break down silos for holistic, consistent datasets.

5) Big Data Value Challenges

a) Identifying Useful Data

Not all collected data holds equal value. Distinguishing data that impacts business outcomes requires clear objectives and domain expertise.

b) Data Overload

Overwhelming amounts of data can obscure relevant insights, leading to analysis paralysis or wasted resources.

c) Interpretation and Analytics

Complex analyses must be accurately interpreted to derive useful conclusions. Poor modeling or misinterpretation can lead to wrong decisions.

Real Life Examples

1. Retailers like Amazon analyze customer purchase behavior to recommend products, increasing sales.
2. Healthcare providers use big data to predict disease outbreaks and improve patient outcomes.
3. Financial institutions detect fraudulent transactions in real time, saving millions.

Solutions

- Define clear use cases and key performance indicators (KPIs) before analysis.
- Use advanced analytics, AI, and machine learning to discover patterns and trends.
- Foster collaboration between data scientists and business units.
- Implement dashboards and real-time reporting to disseminate insights.
- Continuously assess and refine data initiatives based on business impact.

6.2 Big Data Security

A. Big Data Security in the Data Lifecycle

Securing Big Data requires protecting data at every stage of its lifecycle — from creation through destruction — because vulnerabilities can arise at any point.

1. Data Creation / Collection

What happens

Data is generated or collected from sources such as sensors, applications, user input, external partners, or devices.

Security Challenges

- Unauthorized or forged data injection.
- Interception or manipulation of data during capture.
- Lack of data source authentication.

Security Measures

- Strong authentication of data sources.
- Use of secure protocols (TLS/SSL) to protect data in transit.
- Endpoint security on devices collecting data.

2. Data Storage

What happens

Data is stored in databases, data lakes, cloud services, or file systems.

Security Challenges

- Unauthorized access due to weak access controls.
- Data breaches from vulnerabilities or misconfigurations.
- Ransomware or malware attacks encrypting or corrupting stored data.

Security Measures:

- Encryption of data at rest (e.g., AES-256).
- Role-Based Access Control (RBAC) and strict permission management.
- Regular patching and vulnerability management.
- Backup and disaster recovery planning to restore data.

3. Data Usage / Access

What happens

Data is accessed, processed, analyzed, shared across internal or external users, or integrated into applications.

Security Challenges

- Insider threats or privilege abuse.
- Data leakage through reports, APIs, or third-party tools.
- Lack of audit trails to track who accessed or modified data.

Security Measures

- Detailed access logging and monitoring.
- Data masking or tokenization for sensitive fields.
- Enforce least privilege principles and MFA.
- Secure APIs and validate third-party integrations.

4. Data Archival

What happens

Inactive or historical data is moved to long-term storage away from active systems.

Security Challenges

- Reduced security oversight and weaker controls in archives.
- Risks of unauthorized retrieval or data leakage.

Security Measures

- Apply same or heightened encryption and access controls.
- Isolate archives from general network access.
- Maintain integrity checks and auditability.

5. Data Destruction

What happens

Data is permanently deleted when no longer needed, to reduce risk and comply with regulations.

Security Challenges

- Incomplete or improper data erasure leading to data remanence.
- Failure to securely delete copies, backups, or archived versions.

Security Measures

- Use certified data wiping and destruction methods.
- Enforce policies for destroying all redundant copies.

Additional Considerations

- **Data Classification**: Not all data requires the same security level. Sensitive data should receive higher protection.
- **Data Governance**: Continuous policies, training, and awareness programs are crucial.
- **Incident Response**: Prepare plans to quickly respond to and recover from breaches or data loss.
- **Compliance**: Align practices with standards like GDPR, HIPAA, or CCPA.

B. Major Security Threats in Big Data

As big data continues expanding across cloud environments, data lakes, IoT systems, and AI applications, new and evolving security threats have become more prominent. Understanding these threats helps organizations take preventive measures to protect sensitive data and maintain trust.

1. Ransomware Attacks on Data Platforms

Ransomware has evolved from targeting individual systems to advanced attacks that encrypt or exfiltrate data from entire data warehouses, data lakes, and analytics platforms.

Key Risks

- Complete loss of access to critical data and analytics systems.
- Massive downtime impacting business operations.
- Financial losses due to ransom demands and recovery efforts.

Example

In 2025, a financial institution faced a ransomware attack on its Snowflake analytics platform, which resulted in three weeks of downtime and $4.2 million in recovery costs.

Prevention

Regular backups, multi-factor authentication (MFA), network segmentation, and immutable storage to ensure quick recovery without paying ransom.

2. Data Breaches and Unauthorized Access

Data breaches occur when attackers gain unauthorized access to sensitive data due to weak authentication, software vulnerabilities, or misconfigurations.

Key Risks

- Exposure of personal data, financial details, and intellectual property.
- Reputational damage and regulatory penalties.

Example

In June 2025, over 16 billion credentials linked to major tech platforms (Google, Apple, Facebook) were exposed due to infostealer malware, highlighting how easily compromised credentials can impact global systems.

Prevention

Strong encryption for data at rest and in transit, MFA, frequent credential rotation, and zero trust access controls.

3. Insider Threats

Employees or contractors with legitimate access to data can intentionally or accidentally cause breaches by mishandling data or exploiting privileges.

Key Risks

- Theft or misuse of company data.
- Data leaks resulting from carelessness or compromised accounts.

Example

Disgruntled employees leaking proprietary financial data or researchers downloading confidential datasets to personal drives.

Prevention

Implement strict role-based access control (RBAC), conduct background checks, monitor anomalous user behavior, and enforce least-privilege principles.

4. Data Leakage and Misconfigurations

Inadvertent exposure of data through misconfigured cloud storage, poorly secured APIs, or careless data sharing practices.

Key Risks

- Accidental public exposure of sensitive datasets.
- Unprotected APIs providing open access to critical data.

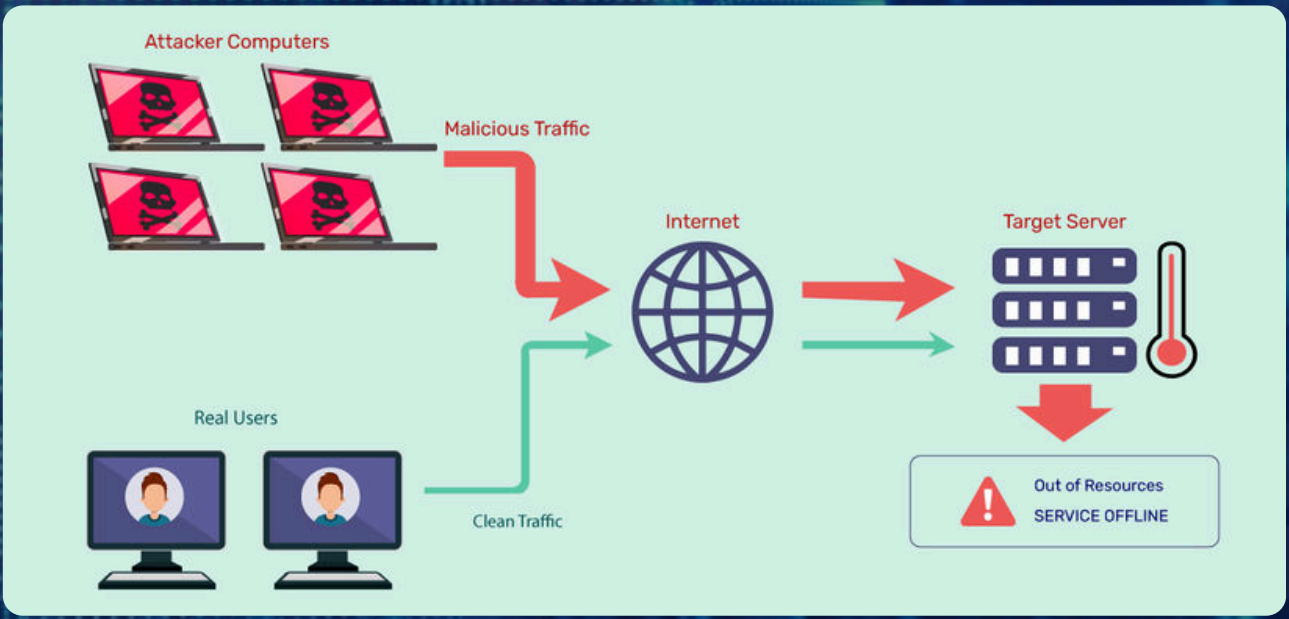
Example

Unsecured cloud buckets have repeatedly exposed millions of customer records due to administrators leaving them publicly accessible.

Prevention

Automated security scans for misconfigurations, least-access policies, encryption, and regular auditing of cloud assets.

5. Distributed Denial-of-Service (DDoS) Attacks



Attackers overwhelm networks or servers with excessive traffic, disrupting services and blocking legitimate access.

Key Risks

- Severe downtime and disruption of services.
- Financial losses due to halted operations.

Example

National infrastructure sites targeted by massive botnet traffic floods, temporarily disabling government and enterprise websites.

Prevention

DDoS mitigation solutions, load balancing, content delivery networks (CDNs), and scalable bandwidth reserves.

C. Key Security Measures for Big Data

Protecting Big Data requires a multi-layered approach that addresses threats across the entire data lifecycle. Here are the most important security measures, with practical details and examples.

1. Encryption

Purpose: Protects data from unauthorized access by converting it into unreadable formats.

Types

- Data at Rest: Encrypt files, databases, and backups using strong algorithms (e.g., AES-256).
- Data in Transit: Use TLS/SSL for secure network communications.

Example: Sensitive customer data stored in a cloud database is encrypted, so even if breached, it remains unreadable without the key.

2. Access Control

Purpose: Ensures only authorized users can view, modify, or delete data.

Types

- Role-Based Access Control (RBAC): Assign permissions based on job roles.
- Least Privilege Principle: Users get only the access needed for their tasks.
- Multi-Factor Authentication (MFA): Adds extra verification steps.

Example: Data analysts can access analytics tools, but not raw data storage; admins have broader access.

2. Access Control

Purpose: Ensures only authorized users can view, modify, or delete data.

Types

- Role-Based Access Control (RBAC): Assign permissions based on job roles.
- Least Privilege Principle: Users get only the access needed for their tasks.
- Multi-Factor Authentication (MFA): Adds extra verification steps.

Example: Data analysts can access analytics tools, but not raw data storage; admins have broader access.

3. Monitoring and Auditing

Purpose: Tracks data access, usage, and system changes to detect suspicious activity.

Tools

- Security Information and Event Management (SIEM): Aggregates logs for real-time analysis.
- User Behavior Analytics (UBA): Detects anomalies in user actions.

Example: Unusual data downloads or login patterns trigger alerts for investigation.

6.3 Big Data Ethics

Big Data Ethics is the discipline of ensuring that data is collected, stored, processed, and shared in ways that respect individual rights, promote fairness, and maintain public trust. As organizations increasingly rely on data analytics and AI, ethical considerations are more important than ever.

1. Transparency

- Organizations must clearly communicate how data is collected, used, and shared.
- Users should be informed about the purposes and methods of data processing.
- Transparent practices build trust and allow individuals to make informed choices.

2. Consent

- Explicit, informed consent must be obtained before collecting or using personal data.
- Individuals should understand what data is being collected and how it will be used.
- Consent should be freely given and revocable at any time.

3. Privacy Protection

- Personal data must be protected from unauthorized access, misuse, or exposure.
- Use privacy-by-design approaches, including anonymization and encryption.
- Limit data collection to only what is necessary for the stated purpose.

4. Data Minimization and Purpose Limitation

- Collect only the data needed for specific, legitimate purposes.
- Avoid repurposing data in ways that conflict with user expectations.

Exercise

1. Why is securing data at every stage of the Big Data lifecycle important?
A. Vulnerabilities can arise at any point
B. Only old data is at risk
C. Security is only needed during storage
D. Security only matters for data destruction

2. What is a common security challenge during the data storage phase in Big Data?
A. Data visualization failures
B. Unauthorized access due to weak controls
C. Slow internet speed
D. Insufficient data labeling

3. Which security measure helps protect data in transit?
A. Unencrypted hard drives
B. TLS/SSL protocols
C. Role-based access only
D. Paper records

4.List two major security threats in Big Data environments.

5. What principle ensures users only have the access necessary for their jobs?
A. Open access policy
B. Least privilege principle
C. Full control configuration
D. API gateway

6. Name one example of an "insider threat" in Big Data.

7. Which measure can prevent employees from emailing confidential files outside the company?
A. Distributed denial-of-service defense
B. Data Loss Prevention tools
C. Role-based marketing
D. Large storage arrays

8. What is a key ethical principle in Big Data concerning how data is collected and used?
A. Unlimited collection without disclosure
B. Transparency
C. Price-fixing
D. Role rotation

9. What should be obtained from users before collecting their personal data?

A. Verbal greetings

B. Informed consent

C. Usage statistics

D. Encryption keys

10. Name two best practices for upholding Big Data ethics.

Answers

1. A. Vulnerabilities can arise at any point
2. B. Unauthorized access due to weak controls
3. B. TLS/SSL protocols
4. Ransomware attacks on data platforms; Data breaches and unauthorized access
5. B. Least privilege principle
6. A disgruntled employee leaking confidential data
7. B. Data Loss Prevention tools
8. B. Transparency
9. B. Informed consent
10. Privacy by design at every project stage; Regular audits for compliance and fairness

Case Studies

Netflix

Netflix is one of the world's largest streaming platforms, with over 250 million subscribers globally. Handling such a vast and diverse audience requires not only strong content but also intelligent systems to deliver the right content to the right viewer.

Big Data lays the foundation of Netflix's operation – used to personalize content, optimize streaming, and guide content creation. Every user interaction on the platform produces valuable data, which Netflix analyzes to enhance user satisfaction.

Data Used

- User behavior (watching patterns, ratings)
- Content data (genres, preferences)
- Device and network performance

Solutions

1. Personalized Recommendations: Content suggestions based on viewing habits.
2. Content Creation: Data helps Netflix decide which shows and movies to produce.
3. Streaming Optimization: Adjusts video quality based on user device and network speed.

Results

- Increased engagement and user retention.
- Successful original content.
- Global growth with region-specific content.

Challenges

Balancing privacy with personalization and refining algorithms.

Overall Test

Multiple Choice Questions

1. Which type of analysis provides decision support and makes use of the prediction so that the analysis' result can be used?
A. Descriptive analysis
B. Diagnostic analysis
C. Predictive analysis
D. Prescriptive analysis

2. Which statistical method is used to determine if there exists a relationship between variables?
A. A/B testing
B. Correlation
C. Regression
D. Quantitative

3. Which of the following best describes the challenge of "Variety" in Big Data?
A. Managing the enormous amount of data being generated
B. Handling different types and formats of data such as text, images
C. Processing data quickly in real time or near real time
D. Ensuring the data is accurate and trustworthy

4. What is a major security threat to big data during data ingress (entry)?
A. Ransomware
B. Data leakage
C. Malicious data injection
D. Data encryption

5. What does the "Velocity" characteristics of big data refer to?
A. The speed at which data is generated and processed in real time.
B. The volume of data collected over time.
C. The diversity of data types and resources.
D. The accuracy and reliability of data.

6. How does big data differ from traditional data in terms of variety?
A. Big data is only structured, while traditional data is unstructured.
B. Traditional data is structured, while big data includes structured, semi-structured, and unstructured data.
C. Both traditional data and big data are structured.
D. Both traditional data and big data are unstructured.

7. A financial institution needs to build a system that can analyze credit card transactions the moment they occur to identify and block fraudulent purchases. Which combination of technologies is most appropriate for the real-time processing and analysis core of this system?
A. Apache Hadoop for data storage and Apache Spark for batch processing the day's transactions.
B. Apache Kafka to ingest the transaction streams and Apache Flink for real-time processing and fraud detection.
C. Apache Superset for visualizing historical transaction data and Apache Hadoop for storing the results.
D. Apache Spark for scheduled batch analytics and Apache Hive for querying large datasets.

8. According to the document, a key disadvantage of batch processing is "Error propagation," where an error can affect an entire dataset, and errors are only recognized after processing is finished. Which characteristic of stream processing directly addresses this specific disadvantage?
A. Higher operational costs
B. Error recognition and resolution in real-time
C. Dynamic scalability
D. Demanding hardware requirements

9. What is one key way Big Data helps improve manufacturing operations?
A. By automatically hiring new employees
B. By printing labels faster
C. By predicting machine failures before they happen
D. By increasing shelf life of products

10. Which statement best describes a data lake?
A. A system that stores only structured data
B. A repository for raw data of all types
C. A database for small-scale applications
D. A backup for local devices

True/False

1. Ratio data has equal intervals between values but lacks a true zero point.
2.Big data always contains accurate and trustworthy information without errors or duplicates.
3. Ethical use of big data includes safeguarding personal information and preventing unauthorized data collection.

4. Apache Spark is considered a suitable tool for a project that requires instantaneous, low-latency insights from a continuous data stream.
5. A major advantage of stream processing is that it provides strong historical context for data analysis because it processes and stores all data before analyzing it.
6. The "Value" characteristics of big data refers to the amount of data collected
7. Big Data can optimize supply chains in retail by analyzing customer purchase patterns.
8. In manufacturing, Big Data has no use in quality control or process monitoring.
9. HDFS can recover data if a DataNode fails.
10. Cloud storage cannot handle multimedia files.

Short Answers

1. In what situation would supervised learning be more appropriate than unsupervised learning?
2. What challenges does the 'Velocity' characteristic of Big Data pose for organizations?
3. How do NoSQL databases help with unstructured data in big data?
4.Give one example of how Big Data is used in the retail industry
5. Give examples of cloud storage services

Word bank

1, Big Data Analytics uses several techniques to make sense of large and complex datasets. One major technique is _____ (1) analysis, which deals mainly with measurable data such as sales or profits. It focuses on analyzing _____ (2) to identify clear _____ (3) and relationships that can guide business decisions. In contrast, _____ (4) analysis focuses on non-numerical data such as customer opinions, reviews, or interviews.

Word banks: {quantitative, qualitative, statistical, numbers, patterns}

2. Big Data presents significant _____ challenges that organizations must address to protect user information. One common threat is _____, where unauthorized access leads to data breaches exposing sensitive personal data. Another concern is _____, which involves injecting malicious or false data to corrupt analytics outcomes. To mitigate these risks, organizations implement _____ strategies, including encryption, access controls, and monitoring. Additionally, _____ considerations require respecting user privacy and ensuring data is collected and used transparently and ethically.

Word banks: malicious data injection, security, ethical, cybersecurity, data leakage

3. A data engineering team is designing a new platform. They plan to use 1. _____ to ingest live user clickstreams from their website. The raw data will then be stored long-term in 2. _____ due to its cost-effectiveness for massive datasets. For their nightly sales aggregation reports, which are not time-sensitive, they will use the 3. _____ paradigm. However, to power their live dashboard that tracks user activity as it happens, they need the 4. _____ paradigm. Finally, business analysts will use 5. _____ to create and share interactive visualizations from the processed data

Word Bank: Batch Processing, Stream Processing, Apache Kafka, Apache Hadoop, Apache Superset

4. Big data is a complex 1._____ that are difficult to analyze using 2._____ data management data and techniques. The foundations of big data are several 3._____ and technologies that enable the processing, and extraction of value from large volumes of data. The 5 Vs (Volume, Velocity, 4._____, Veracity, Value) define the key characteristics of big data.

Word Bank: {variety, traditional, concepts, datasets}

5. Big Data enables _____ maintenance in manufacturing by predicting equipment failure. It also supports _____ control through process data. In retail, it helps in _____ forecasting, customer _____ analysis, and _____ recommendation systems.

Word banks: {quality, product, behavior, predictive, demand}

6. _____ manages metadata and directories in Hadoop. _____ stores raw data in any format for analysis or machine learning. The three Vs of big data are _____, _____, and _____. _____ is a process involving Extraction, Transformation, and Loading. _____ databases handle unstructured and semi-structured data efficiently.

Word Bank: (HDFS, Data Lake, ETL, NoSQL, Cloud Storage, NameNode, DataNode, Volume, Velocity, Variety)

Multiple Answers

1. Which of the following statements about Support Vector Machines (SVMs) are correct?

A. SVM aims to find the hyperplane that best separates different classes in the feature space.

B. SVMs can only be used for linear classification problems.

C. The data points that lie closest to the decision boundary are called support vectors.

D. SVMs can use kernel functions to handle nonlinear data.

E. SVMs are primarily designed for clustering unlabeled data.

2. A company is building a data platform and has chosen Apache Kafka as its central data ingestion hub. According to the provided text, which THREE of the following statements accurately describe the role or function that Kafka will play in this architecture?

A. It will decouple data producers from data consumers, acting as a message broker.

B. It will perform complex, in-memory batch processing on historical data stored in Hadoop.

C. It will collect and buffer real-time data streams from sources like application logs and IoT sensors.

D. It is ideal for implementing an event-driven architecture and handling financial transactions.

3. Which of the following are examples of sources of big data?

A. Social media posts

B. Customer names and phone numbers

C. Website clicks

D. Sales records in spreadsheets

E. Transaction data from sensors

5. Which are types of NoSQL databases?
   A. Document-based
   B. Key-value
   C. Graph-based
   D. Table-based

Answers

Multiple Choice Questions

1. D
2. B
3. B
4. C
5. A
6. B
7. B
8. B
9. C
10. B

True/False

1. False
2. False
3. True
4. False
5. False
6. False
7. True
8. False
9. True
10. False

Short Answers

1. Supervised learning is better when the objective is classification and there are labeled data. So the desired output is already known; the algorithm can create a relationship or function between inputs and the correct output. For example, detecting spam emails or frauds.

2. The "Velocity" challenge refers to the high speed at which data is generated, collected, and needs to be processed. Organizations face difficulties managing this rapid data flow, which can overwhelm traditional data processing systems

3. NoSQL databases like MongoDB store unstructured data (text, images) by using flexible schemas, making it easier to manage diverse data types in big data environments

4. Retailers use Big Data to recommend products based on customers' purchase history and browsing behavior

5. Google Drive / Dropbox / OneDrive / Amazon S3 / iCloud

Word banks

1. Quantitative, numbers, patterns, statistical
2. Security, data leakage, malicious data injection, cybersecurity, ethical
3. Apache Kafka, Apache Hadoop, Batch Processing, Stream Processing, Apache Superset
4. Datasets, traditional, concepts, variety
5. Predictive, quality, demand, behavior, product
6. NameNode; Data Lake; Volume, Velocity, Variety; ETL; NoSQL

PART II
CONNECTIONIST AI

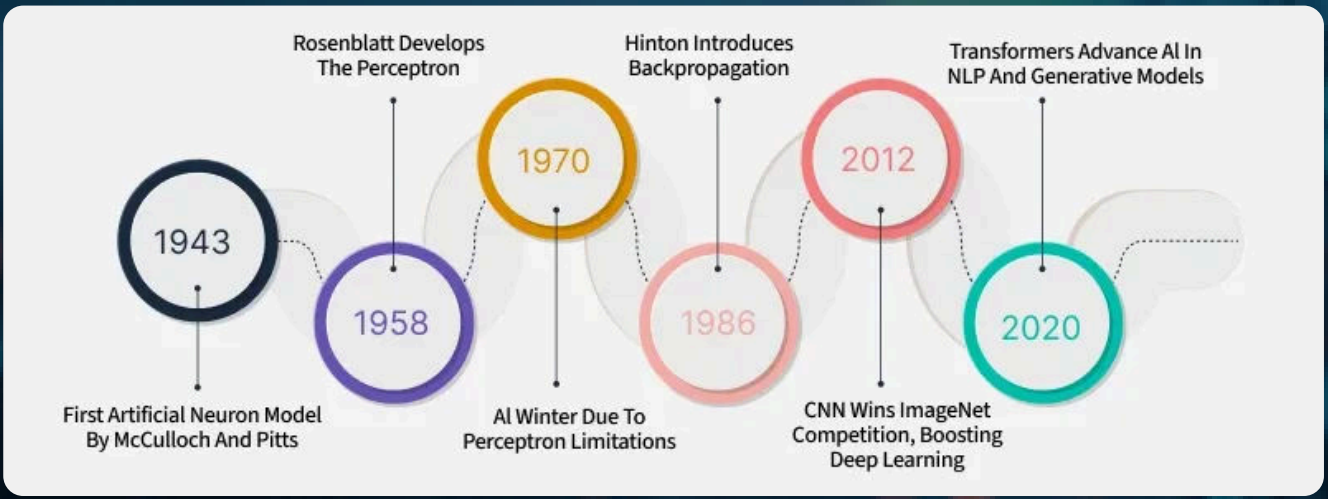
CH1 Foundations of Connectionist AI

1.1 Foundations of Connectionist AI

Connectionist AI is built on the concept of artificial neural networks, which consist of layers of interconnected units that process and transform data through weighted connections. These networks learn by adjusting the weights of connections based on the error between predicted and actual outcomes, typically using algorithms like backpropagation. Through training, neural networks can identify complex patterns in data, which enables them to make predictions or classifications. As the network becomes deeper, with more hidden layers, it gains the ability to learn hierarchical features, making it particularly effective for tasks like image recognition, natural language processing, and reinforcement learning. Despite their power, these models face challenges such as overfitting and generalization, which are addressed with techniques like regularization and transfer learning.

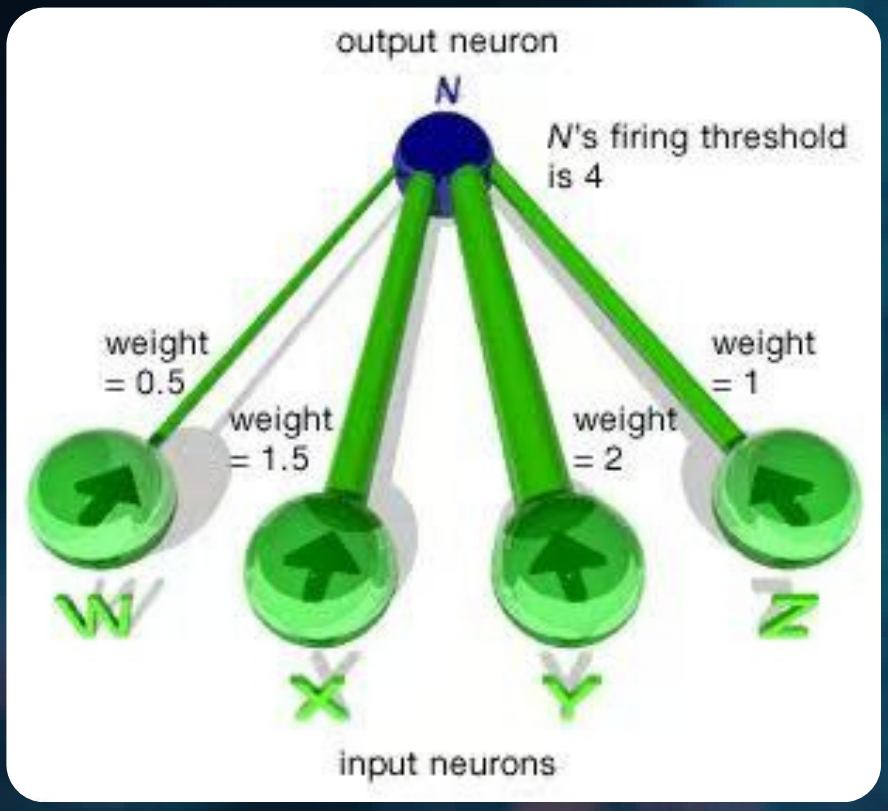
1.2 The Origins of Connectionist AI

Connectionism or neuronlike computing, developed out of attempts to understand how the human brain works at the neural level and how people learn and remember. In 1943, the neurophysiologist Warren McCulloch of the University of Illinois and the mathematician Walter Pitts of the University of Chicago published an influential treatise on neural nets and automatons, according to which each neuron in the brain is a simple digital processor and the brain as a whole is a form of computing machine.



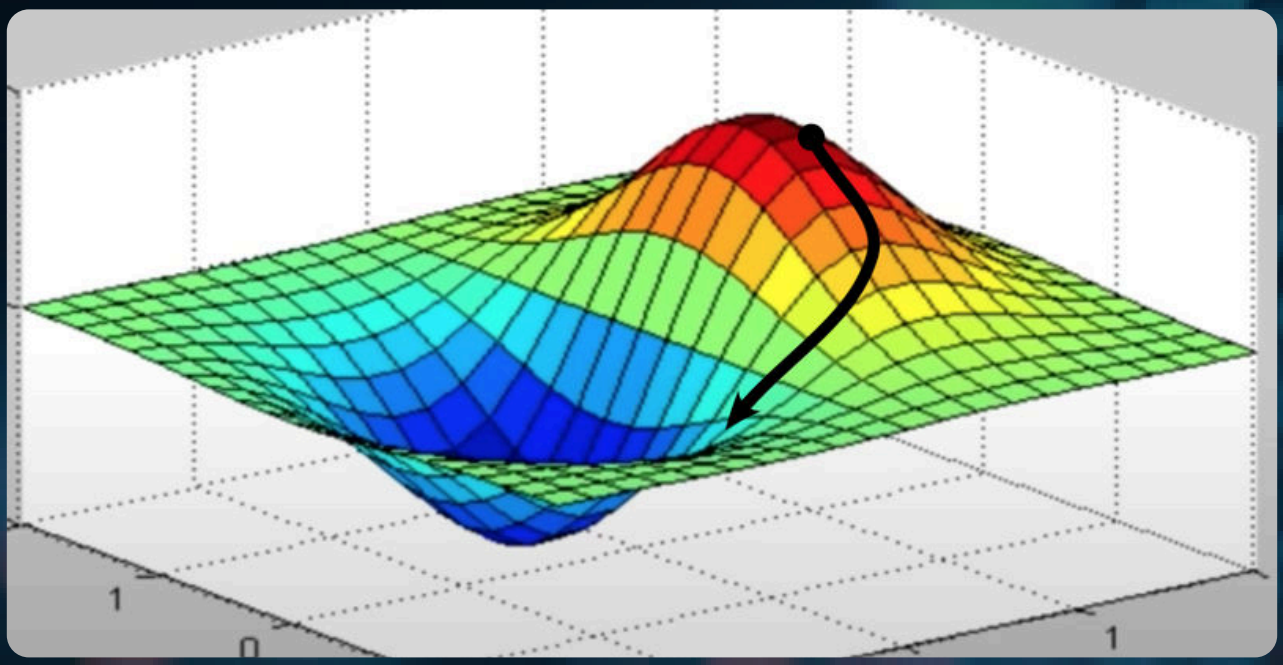
It was not until 1954, that Belmont Farley and Wesley Clark of MIT succeeded in running the first artificial neural network although limited by computer memory to no more than 128 neurons. They were able to train their networks to recognize simple patterns. In addition, they discovered that the random destruction of up to 10 percent of the neurons in a trained network did not affect the network's performance, a feature that is remindful of the brain's ability to tolerate limited damage inflicted by surgery, accident, or disease.

1.3 Knowledge as Patterns of Activation in Networks



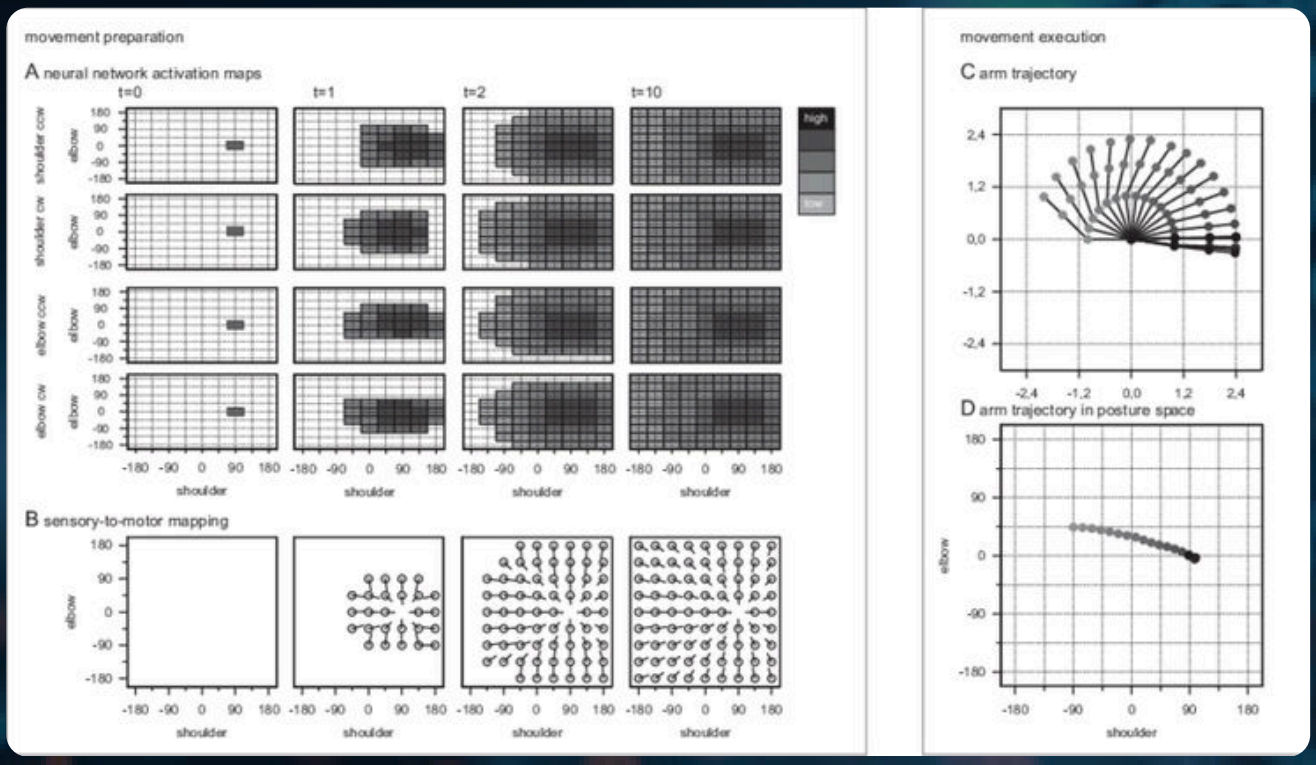
Connectionist AI represents knowledge not through explicit symbols, rules, or logical structures, but as patterns of activation distributed across networks of interconnected units. In this paradigm, each unit contributes to a larger pattern, and meaning arises from the collective state of the network rather than from any single component. The knowledge is embedded in the strengths of the connections known as weights between these units, which are adjusted during training as the model is exposed to data.

Unlike symbolic AI, where knowledge is manipulated through formal, rule-based systems, connectionist models operate through the transformation of activation patterns in response to inputs, enabling them to capture complex statistical regularities and dependencies.



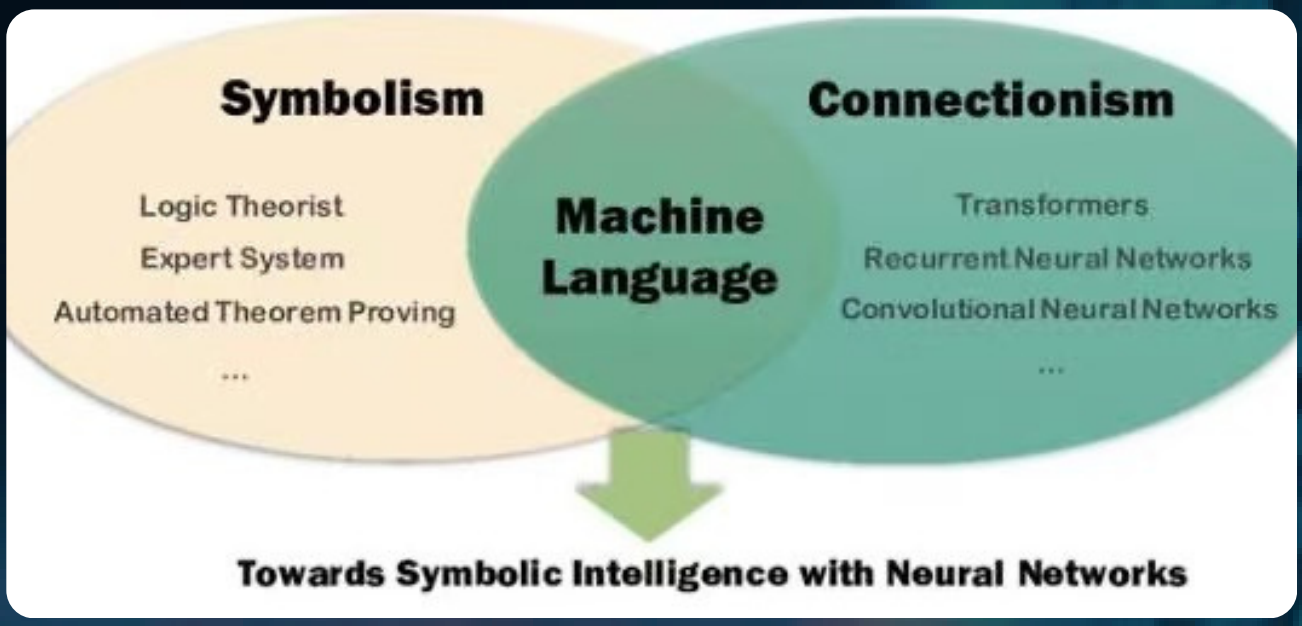
Learning in connectionist systems is typically achieved through gradient-based optimization techniques such as backpropagation, which fine-tune the weights to reduce the discrepancy between the model's output and the expected result. This allows the system to develop internal representations that reflect the structure of the input data, often without any explicit instruction or labeling of concepts. These representations are often distributed, meaning that a single concept or feature is encoded across many units, and each unit participates in representing many different concepts.

Furthermore, the emphasis on activation patterns rather than symbolic manipulation enables connectionist models to handle ambiguity, context dependence, and graded similarity features that are difficult to capture using symbolic approaches.



This makes connectionist AI particularly suitable for tasks like visual recognition, speech processing, and natural language understanding, where inputs are high-dimensional and meanings are fluid. Rather than performing explicit logical reasoning, these models learn to associate inputs with outputs based on learned experience, creating a flexible and adaptive form of intelligence grounded in pattern recognition.

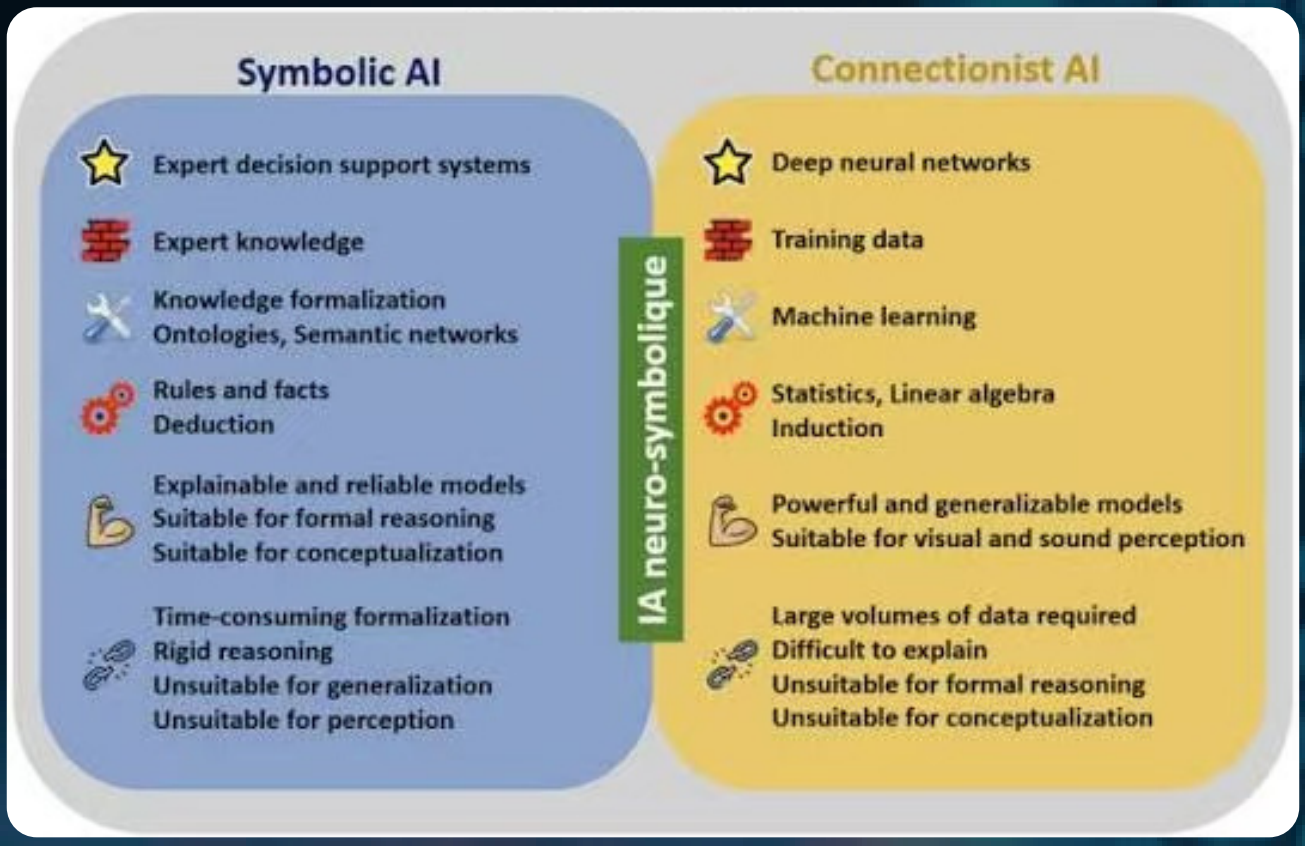
1.4 Difference between Connectionist AI and Symbolic AI



Artificial intelligence has evolved along two distinct paths: Symbolic AI and Connectionist AI.

These approaches represent fundamentally different ways of creating intelligent systems, each with unique strengths and applications. Symbolic AI operates through explicit rules and logical structures.

In contrast, Connectionist AI takes inspiration from the human brain's neural networks. These systems learn patterns from vast amounts of data rather than following predefined rules.



The most striking difference lies in how these systems learn and adapt.

Symbolic AI requires human experts to manually encode rules and knowledge, making it precise but labor-intensive to update.

Connectionist systems learn automatically from examples, allowing them to handle new situations more flexibly, though they may require significant computing resources and training data.

Difference between Symbolic AI and Connectionist AI

Feature	Symbolic AI	Connectionist AI
Knowledge Representation	Easy to read symbols and logical statements	Distributed across networks of artificial neurons
Learning method	Manually encoded rules and knowledge	Automatically learn from examples
Strengths	Precise decision making, logical reasoning, explainability	Pattern recognition, adaptive learning, handling large data
Weaknesses	Labor intensive to update, struggles with ambiguous situations	High computational requirements, unclear decision making
Applications	Medical diagnosis, legal analysis, financial compliance	Image recognition, natural language processing, fraud detection

Exercise

A. Multiple Choice Questions

1. What is the core idea behind Connectionist AI?
A) It uses symbolic logic and rules to represent knowledge.
B) It relies on human experts to encode knowledge manually.
C) It uses artificial neural networks to identify patterns and make predictions.
D) It avoids pattern recognition and focuses on data encryption.

2. Which of the following algorithms is commonly used for training Connectionist AI?
A) Decision Trees
B) Backpropagation
C) K-Nearest Neighbors
D) Genetic Algorithms

3. In Connectionist AI, how is knowledge represented?
A) As explicit symbols and formal rules.
B) As patterns of activation across networks of interconnected units.
C)Through logical deductions and formal proofs.
D) In structured, hierarchical databases.

4. What is one advantage of Connectionist AI over Symbolic AI?
A) It requires manual encoding of knowledge by experts.
B) It is more adaptable and can learn from data automatically.
C) It is highly interpretable and easy to modify.
D) It excels in tasks requiring explicit reasoning.

5. What is a major challenge faced by Connectionist AI models?
A) Overfitting and generalization
B) Lack of data
C) Inability to perform logical reasoning
D) Limited computation power

6. How does Connectionist AI handle ambiguity and context dependence?
A) By performing logical reasoning on explicit rules
B) By creating rigid, fixed representations
C) By relying on activation patterns that are flexible and adaptable
D) By storing information in human-readable symbols

B. True or False

1. Connectionist AI models are designed to use predefined, explicit rules for knowledge representation.
2. Connectionist AI systems are better suited for tasks like pattern recognition, visual recognition, and speech processing.
3. In the early days of Connectionist AI, the first artificial neural network was able to recognize complex patterns and handle large datasets.
4. Symbolic AI excels in tasks requiring adaptive learning and pattern recognition.

Answers

Multiple Choice Questions

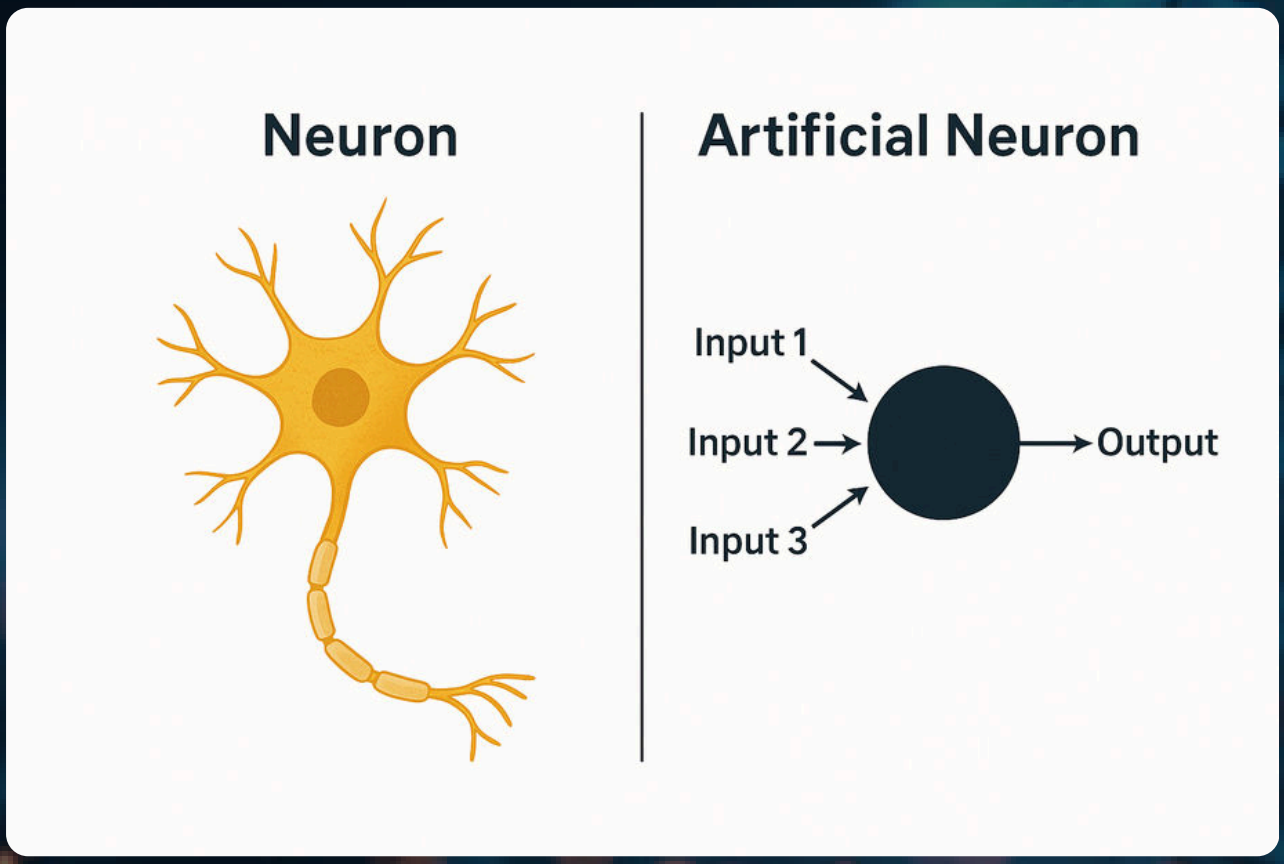
1. C
2. B
3. B
4. B
5. A
6. C

True or False

1. False
2. True
3. False
4. False

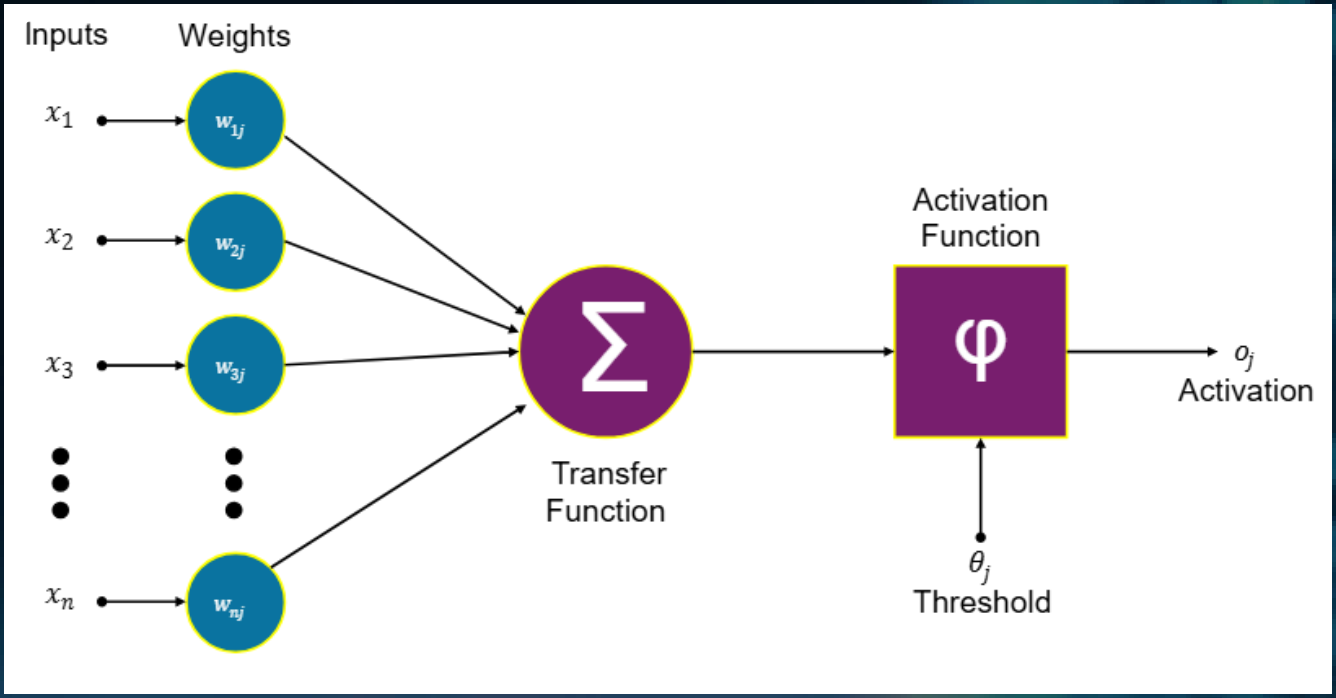
CH2 Artificial Neuron Network

2.1 Definition of an Artificial Neuron



An Artificial Neuron is a mathematical function that serves as the fundamental processing unit of a neural network. Its design is inspired by the biological neurons found in the human brain, mimicking their basic behavior of receiving, processing, and transmitting signals. While a biological neuron uses electrical and chemical signals, an artificial neuron processes numerical data.

2.2 The Basic Function of a Neuron



The operation of a single artificial neuron can be broken down into a three-step process:

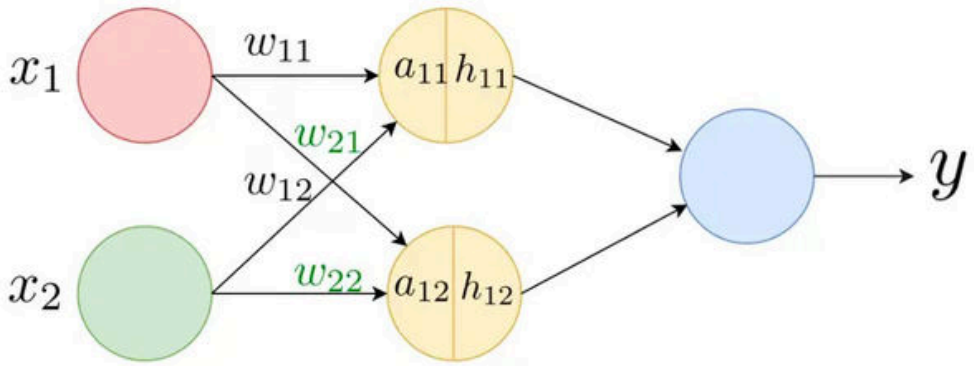
· Step 1: Receive Input Signals: The neuron receives multiple input signals (denoted as x_1, x_2, ..., x_n). These inputs can be raw data or outputs from other neurons in the network.

· Step 2: Combine and Weigh Signals: The neuron combines these input signals, assigning a level of importance to each one. This is done using weights and biases.

· Step 3: Transform into Output: The combined and weighted value is then passed through a special function to produce a single output signal. This output can then be passed on to other neurons or used as the final result.

2.3 Core Components of a Neuron

2.3.1 The Purpose of Weights



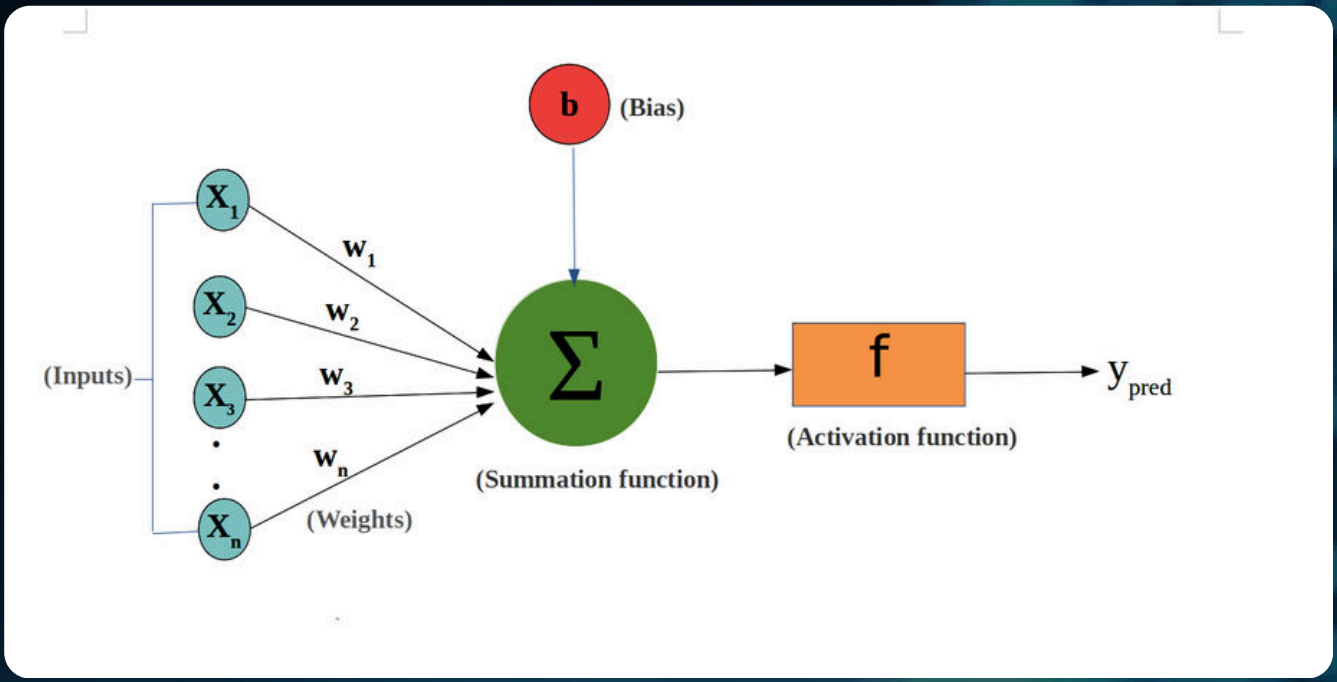
$$a_{11} = w_{11}x_1 + w_{12}x_2$$

$$a_{12} = w_{21}x_1 + w_{22}x_2$$

· Definition: Weights (denoted as w_1, w_2, ..., w_n) are numerical parameters that represent the strength or importance of the connection for each specific input.

· Influence: A higher weight means its corresponding input has a greater influence on the neuron's output. Conversely, a lower weight means the input has less influence. A negative weight means the input inhibits the output.

· Role in Learning: Weights are the primary values that are adjusted during the network's "learning" or "training" process. The algorithm learns by fine-tuning these weights to reduce error and make accurate predictions.



2.3.2 The Purpose of Biases

· Definition: A bias is an additional, trainable parameter that is independent of the input. It is added to the weighted sum of inputs.

· Function: The bias allows the neuron to shift or offset its activation function to the left or right. This flexibility helps the model better fit the data by providing a baseline level of activity, making it easier for the network to learn patterns that are not centered around zero.

2.3.3 The Aggregation and Activation Process

This is the mathematical core of the neuron, where inputs are processed into an output. It occurs in two distinct phases:

2.3.4 Phase One: Aggregation (Summation)

· In this phase, the neuron calculates the weighted sum of all its inputs plus the bias.

· The formula for this aggregation is: $z = (x_1 * w_1) + (x_2 * w_2) + ... + (x_n * w_n) + bias$

· The result, z, is a single value that represents the total combined input to the neuron.

2.3.5 Phase Two: Activation

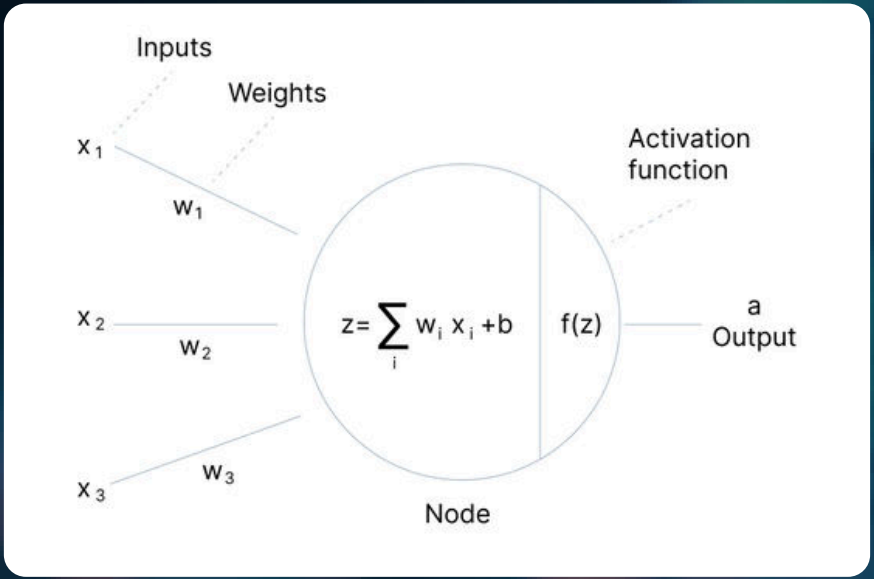
· The aggregated sum z is then passed as an argument to an Activation Function.

· The activation function performs a fixed mathematical transformation on z to produce the neuron's final output.

· The formula for the final output is: y = activation function(z)

2.4 Definition and Purpose of Activation Functions

An activation function is a mathematical function applied to the aggregated sumz of a neuron. It is a critical component for enabling neural networks to learn complex patterns.



Purpose 1: Introduce Non-Linearity

· Without a non-linear activation function, no matter how many layers a neural network has, it would behave just like a single-layer linear regression model. The entire network could be collapsed into a single linear transformation.

· Non-linearity allows the network to learn and model complex, real-world, non-linear patterns and relationships found in data like images, sound, and text. It is what gives deep neural networks their true power.

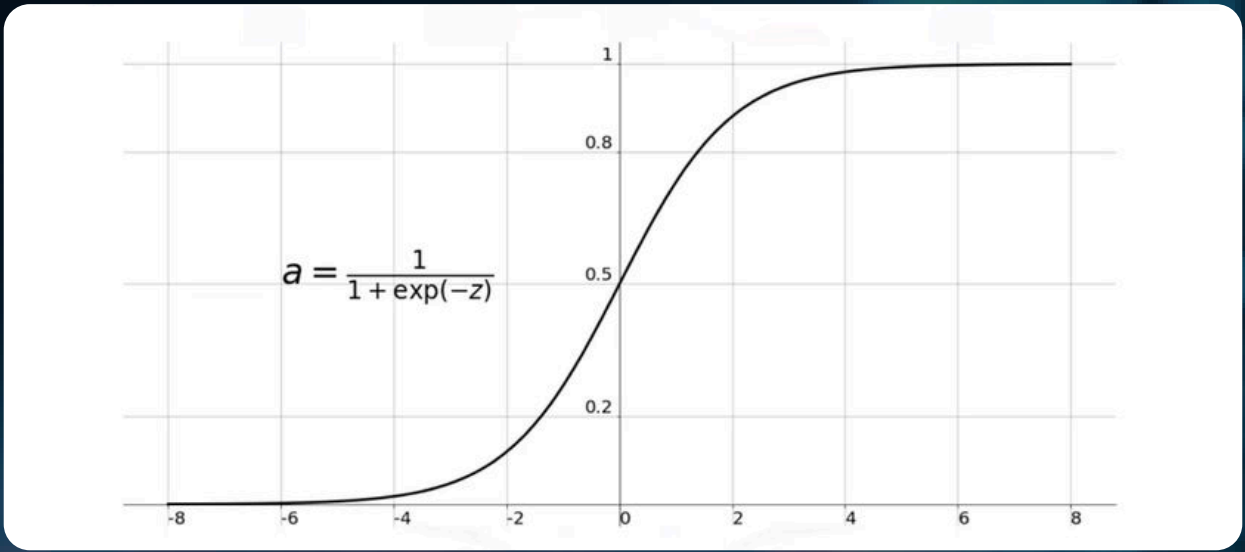
Purpose 2: Map Output to a Desirable Range

· The activation function decides whether a neuron should be "activated" (i.e., "fired" or contribute significantly to the next layer) or not.

· It squashes the potentially unbounded value of z into a fixed, manageable range (e.g., 0 to 1, -1 to 1). This controlled output is necessary for the stability of deep networks.

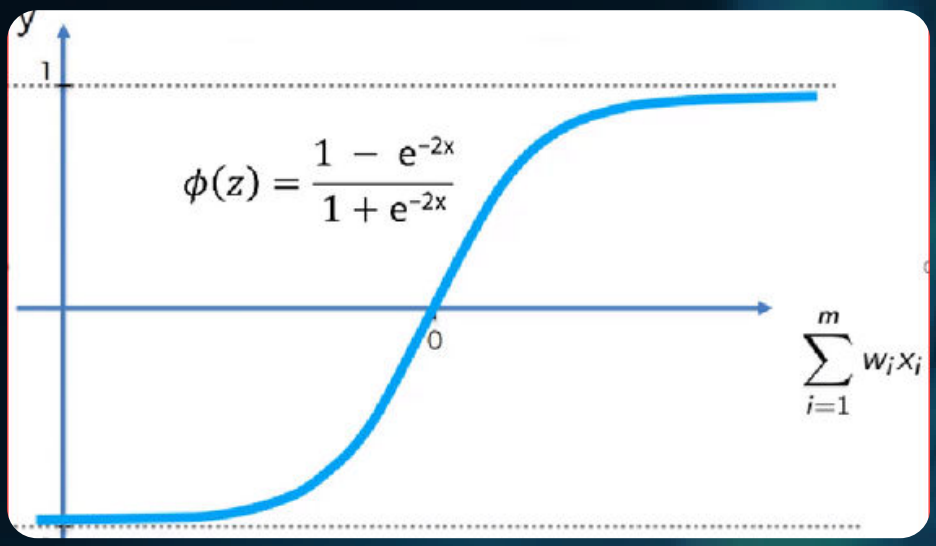
2.5 Common Types of Activation Functions

The Sigmoid Function



· Formula: f(z) = 1 / (1 + e^{-z})
· Output Range: Between 0 and 1.
· Characteristics:
  · It has a smooth, S-shaped curve.
  · Because its output is between 0 and 1, it can be interpreted as a probability. For example, an output of 0.8 can be seen as an 80% chance.
· Common Use: Historically, it was very popular for the output layer in binary classification problems.
· Disadvantage: It suffers severely from the "vanishing gradient" problem. For very high or very low values of z, the function's slope becomes almost zero, which can halt learning in deep networks.

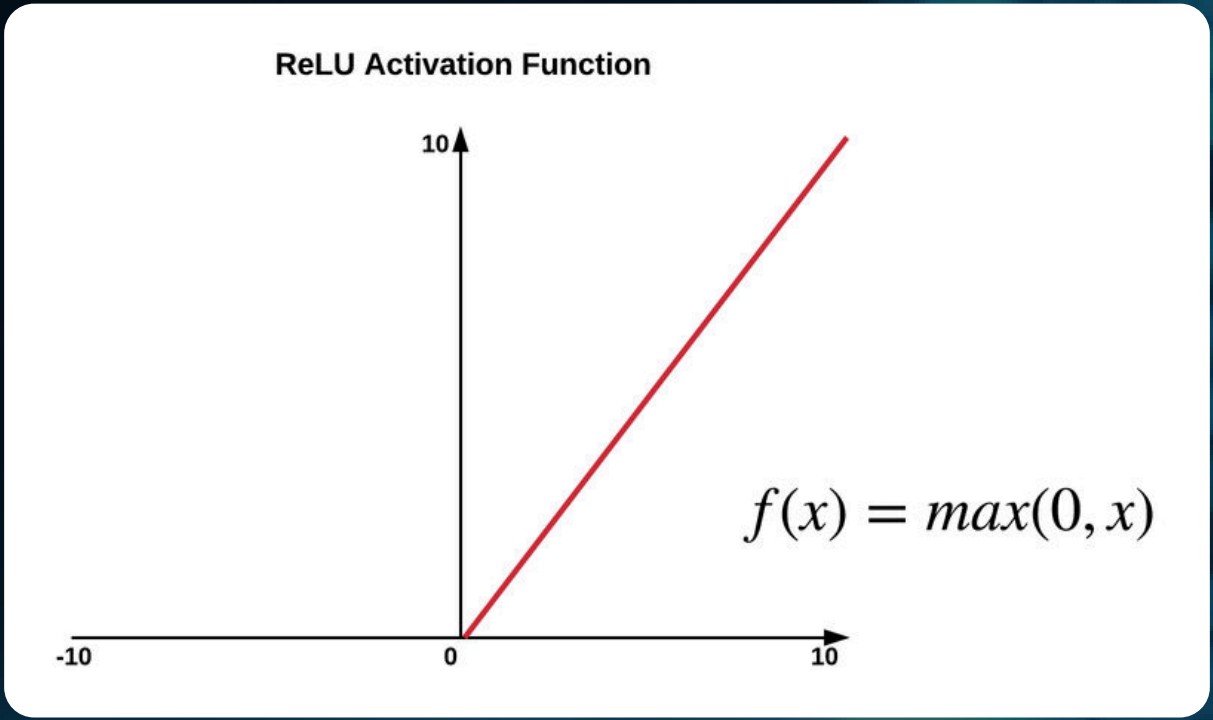
The Tanh Function (Hyperbolic Tangent)



· Formula: f(z) = (e^z - e^{-z}) / (e^z + e^{-z})
· Output Range: Between -1 and 1.
· Characteristics:
· It is also an S-shaped curve, but it is zero-centered (its output range is centered around 0). This property often makes training faster and more effective than with the Sigmoid function.
· Common Use: Often found in the hidden layers of neural networks.
· Disadvantage: Like the Sigmoid, it also suffers from the "vanishing gradient" problem.

The ReLU Function (Rectified Linear Unit)

· Formula: f(z) = max(0, z)
· Output Range: 0 to infinity.
· Characteristics:
· It is very simple and computationally efficient, as it involves only a comparison and no complex exponentials.
· It induces sparse activation: since it outputs zero for all negative inputs, it does not activate all neurons at the same time.



· Common Use: It is the most widely used and default activation function for hidden layers in modern deep learning networks.

· Disadvantage: It can cause the "Dying ReLU" problem. If a neuron's weights adjust in such a way that it always outputs a negative value for all data points, the gradient becomes zero, and the neuron gets "stuck," permanently outputting 0 and no longer contributing to learning.

2.6 The Power of Layers

· A single neuron, also known as a perceptron, is a weak learner with limited capability. It can only solve linearly separable problems.

· The true power of neural networks emerges when many neurons are connected together in layers. This architecture creates a powerful, hierarchical learning model capable of approximating any complex function.

Exercises

A. Multiple Choice Questions

1. What is the primary purpose of an activation function in an artificial neuron?
a) To calculate the weighted sum of the inputs.
b) To initialize the weights and biases to random values.
c) To introduce non-linearity and map the output to a specific range.
d) To receive the raw input data from the previous layer.

2. During the training process of a neural network, what are the primary values that get adjusted?
a) The input data features (x_1, x_2, ...)
b) The activation functions in each layer.
c) The weights and biases of the connections.
d) The learning rate of the output layer.

3. The "Dying ReLU" problem occurs when:
a) The output is always between 0 and 1, causing slow learning.
b) Neurons get stuck and only output 0 for all inputs, ceasing to learn.
c) The function is zero-centered, making it computationally expensive.
d) The gradient becomes too large and causes the model to diverge.

4. What is the key advantage of connecting neurons into layers, as opposed to using single neurons?
a)It reduces the amount of data needed for training.
b)It eliminates the need for any activation functions.
c)It prevents the "vanishing gradient" problem in all cases.
d)It creates a powerful, hierarchical model capable of learning complex patterns.

5. The Tanh activation function is often preferred over the Sigmoid for hidden layers because:
a) Its output range is between 0 and infinity, which is more versatile.
b) It is computationally simpler than the Sigmoid function.
c) It is zero-centered, which can make the training process faster.
d) It completely avoids the issue of vanishing gradients.

6. In the aggregation step of a single neuron, what does the variable 'z' represent?
a) The final output of the neuron after activation.
b) The weighted sum of the inputs plus the bias.
c) The derivative of the activation function.
d) The error between the predicted and actual output.

True or False

1. Without a non-linear activation function, a multi-layer neural network is essentially just a linear regression model.
2. A bias term in a neuron allows it to fit the data better by providing an offset, independent of the input.
3. The Sigmoid function is the most modern and commonly recommended activation function for all hidden layers in deep networks.
4. The main role of the input layer in a neural network is to perform complex feature detection and computation.

Answer

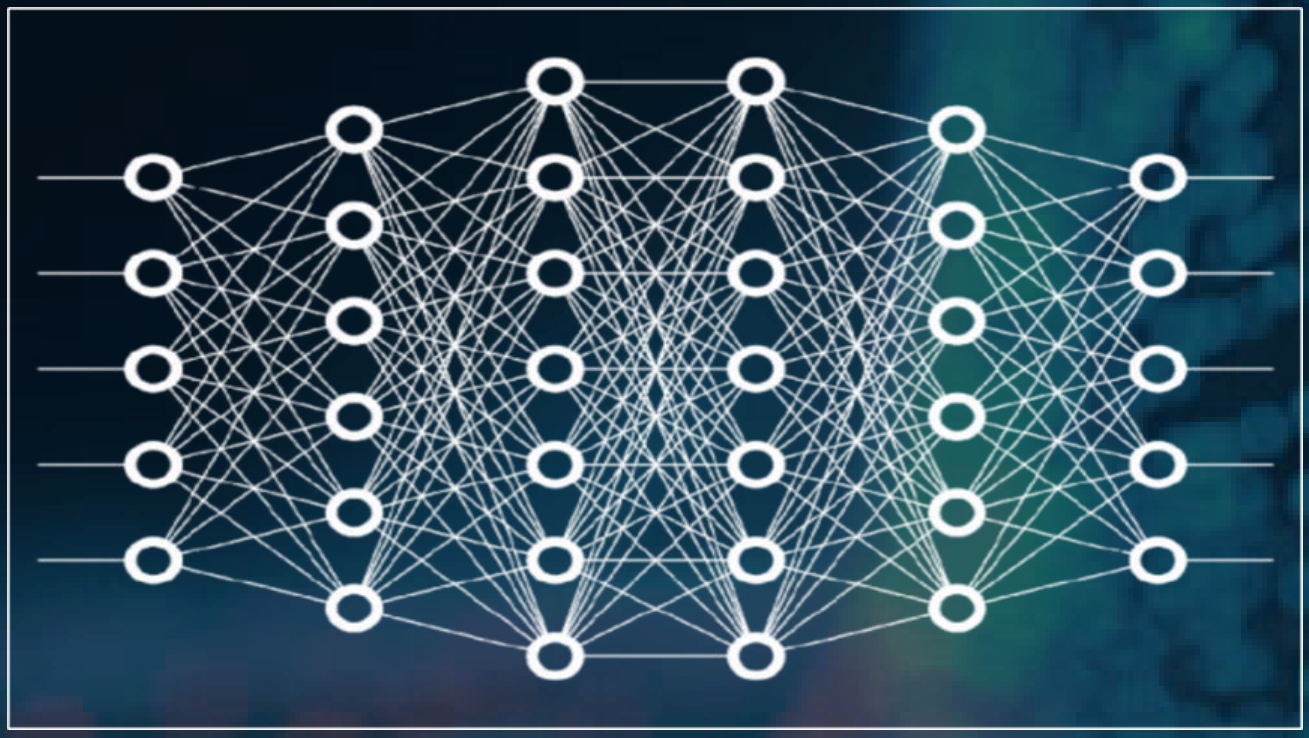
Multiple Choice Answers

1. C
2. C
3. B
4. D
5. C
6. B

True or False

1. True. Without non-linearity, the layers compose into a single linear transform.
2. True. The bias shifts the activation function, providing an offset to improve data fitting.
3. False. ReLU is the most common default for hidden layers. Sigmoid suffers from vanishing gradients and is now typically reserved for output layers in specific cases.
4. False. The input layer only receives raw data. Complex feature detection happens in the hidden layers.

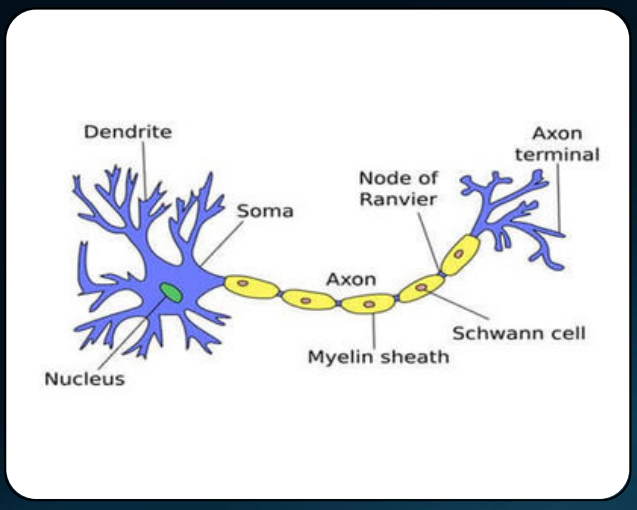
CH3 Learning in Neural Networks



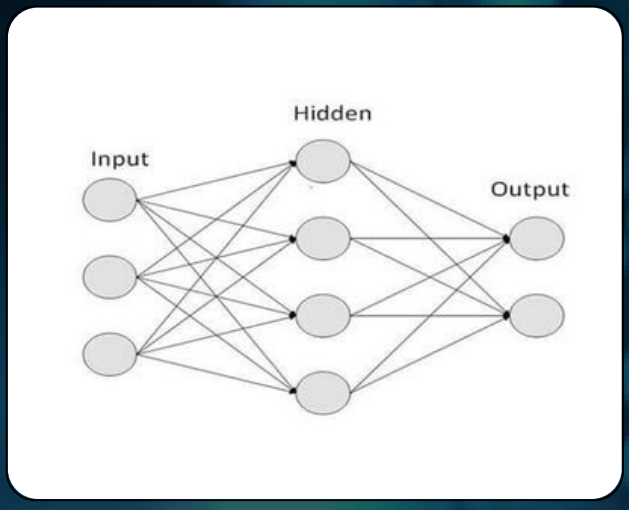
The neural network is a machine learning model inspired by the structure and function of the human brain and how they process data and information. Neural network uses interconnected nodes (neurons) that work together in a layered structure to recognize patterns, make decisions, and predict outcomes.

The neural network architecture is mainly inspired by the human brain. Neurons (human brain cells) form interconnected networks, sending electrical signals to each other to help the brain process data and information. Similarly, the neural network is made of interconnected artificial neurons (nodes) that work-

together, performing mathematical calculations to recognize patterns, make decisions, and predict outcomes.

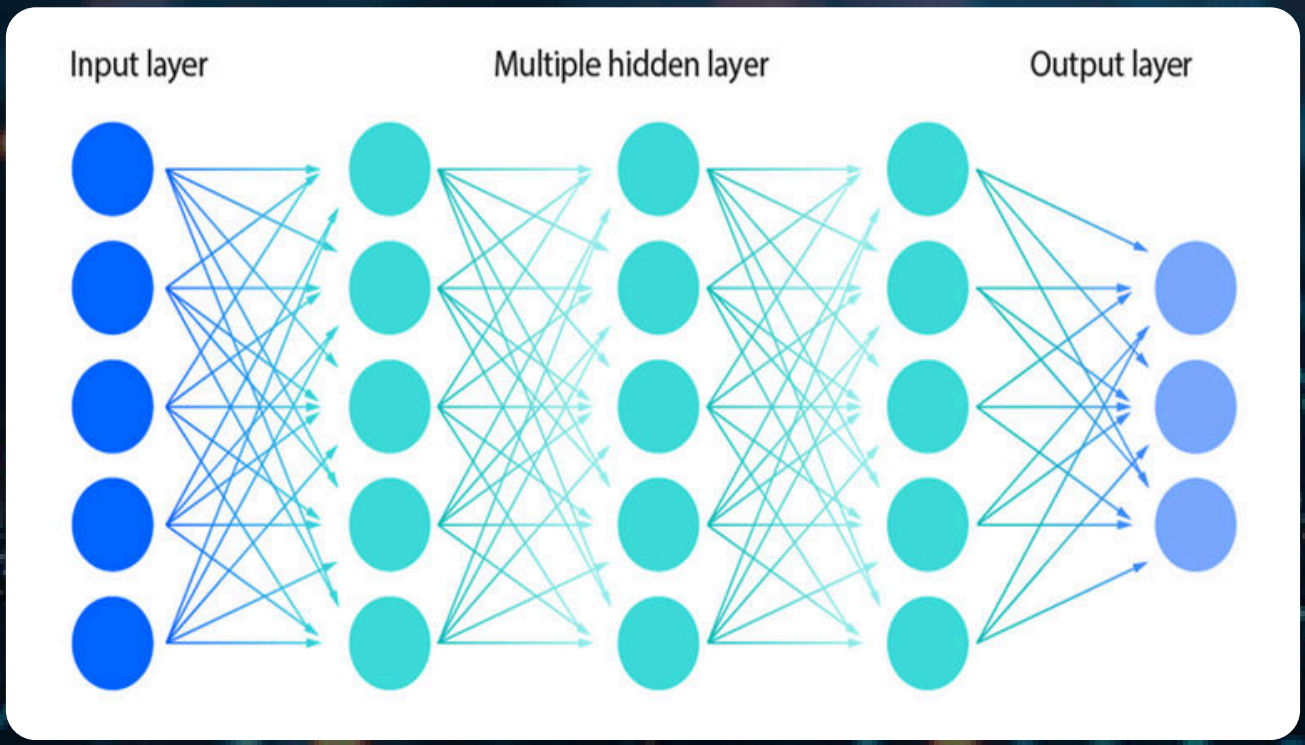


Human Brain Cells



Artificial Neurons

3.1 Neural Network Layers



The Neural Network Layers

Neurons are aggregated into layers that perform different transformations. Each neuron receives input, processes it using weights and an activation function, and passes the result to the next layer. Signals (in the form of real numbers) travel from the first layer (the input layer), pass through multiple intermediate layers (hidden layers) and reach the last layer (the output layer). If a network has at least 2 hidden layers, it is called a deep neural network.

The neural network consists of three types of layers:

a) Input Layer

Information enters the neural network from the input layer. The input nodes process, analyze, or categorize it, and then pass it to the next layer.

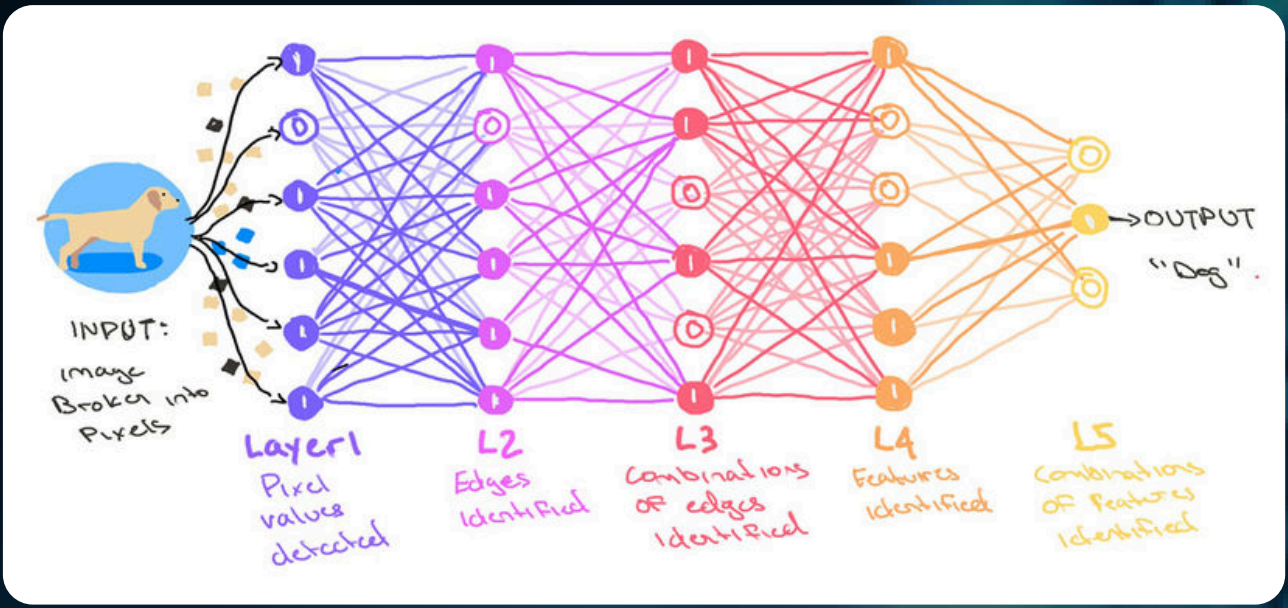
b) Hidden Layer

The hidden layers take input from the input layer or previous hidden layers. Each hidden layer analyzes and processes the output of the previous layer, and then passes it to the next layer.

c) Output Layer

The output layer gives the final result of all the data processing by the neural network.

Here is an example:

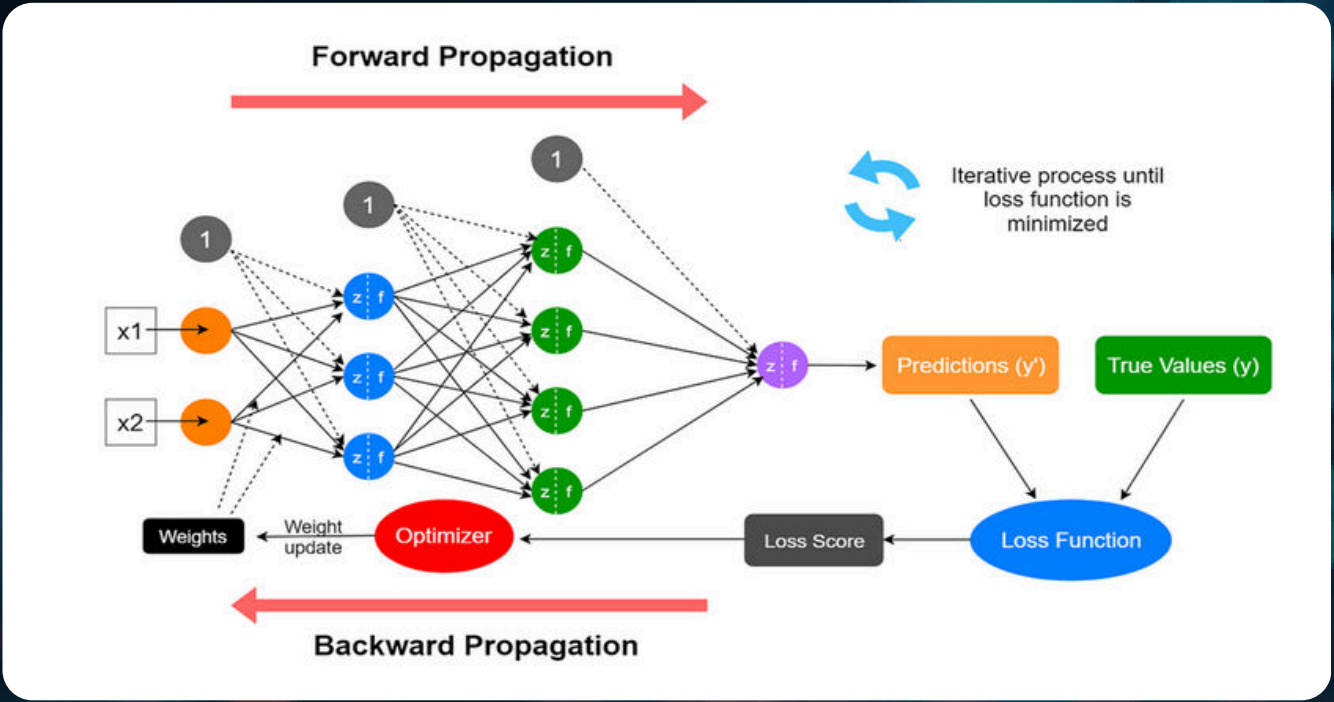


Source: www.notablecap.com

Layer	Type	Explanation
1	Input layer	A picture of a dog enters the network through the input layer.
2	Hidden layer	The first lights up nodes for a section, signaling that a shape is there.
3	Hidden layer	The next layer lights up nodes for what represents the edge of roughly a shape of an animal.
4	Hidden layer	The next layer lights up nodes for what represents the edge of roughly a shape of an animal.
5	Output layer	As it gets to the end, a single output node lights up for “dog”.

3.2 How Neural Networks Learn

Neural networks learn through a process known as training, which involves adjusting their internal parameters (weights and biases) to minimize loss (error).



1. Input

Data is fed into the network.

2. Forward Propagation

Data flows through the network that computes linear combinations, passing through the nonlinear activation function and computes an output prediction.

3. Loss Calculation

The loss function measures the difference (called loss) between the predicted output and the actual answer.

4. Backward Propagation

The loss is propagated backward through the network. The network then adjusts the weights and biases to the direction that reduces the loss, using an optimization method like gradient descent.

5. Iteration

Steps 1-4 are repeated many times until the loss is minimized.

3.3 Neural Network Applications

Neural networks are used across many industries. The examples of applications of neural networks across industries are as such.

a. **Computer Vision**: Face recognition, medical imaging, self-driving cars

b. **Natural Language Processing**: Chatbots, translation, sentiment analysis

c. **Speech Recognition**: Voice assistants (Siri, Alexa)

d. **Finance**: Fraud detection, stock price prediction

e. **Healthcare** : Disease diagnosis, drug discovery

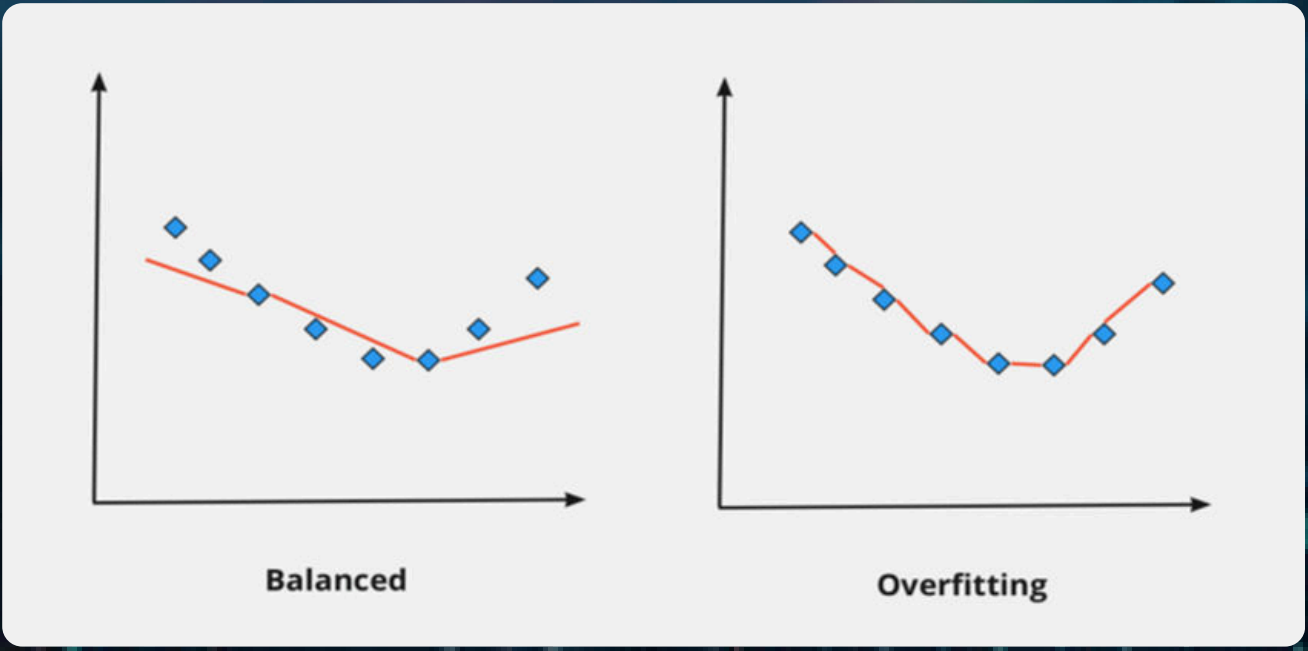
f. **Manufacturing**: Predictive maintenance, quality inspection

g. **Marketing**: Customer behaviour prediction, recommendation systems

3.4 Generalization, Overfitting, & Regularization

In neural networks, there is a term called training data and test data. When neural networks train, they learn parameters (weights and biases) to minimize loss function on the training data and try to predict the test data based on its learning.

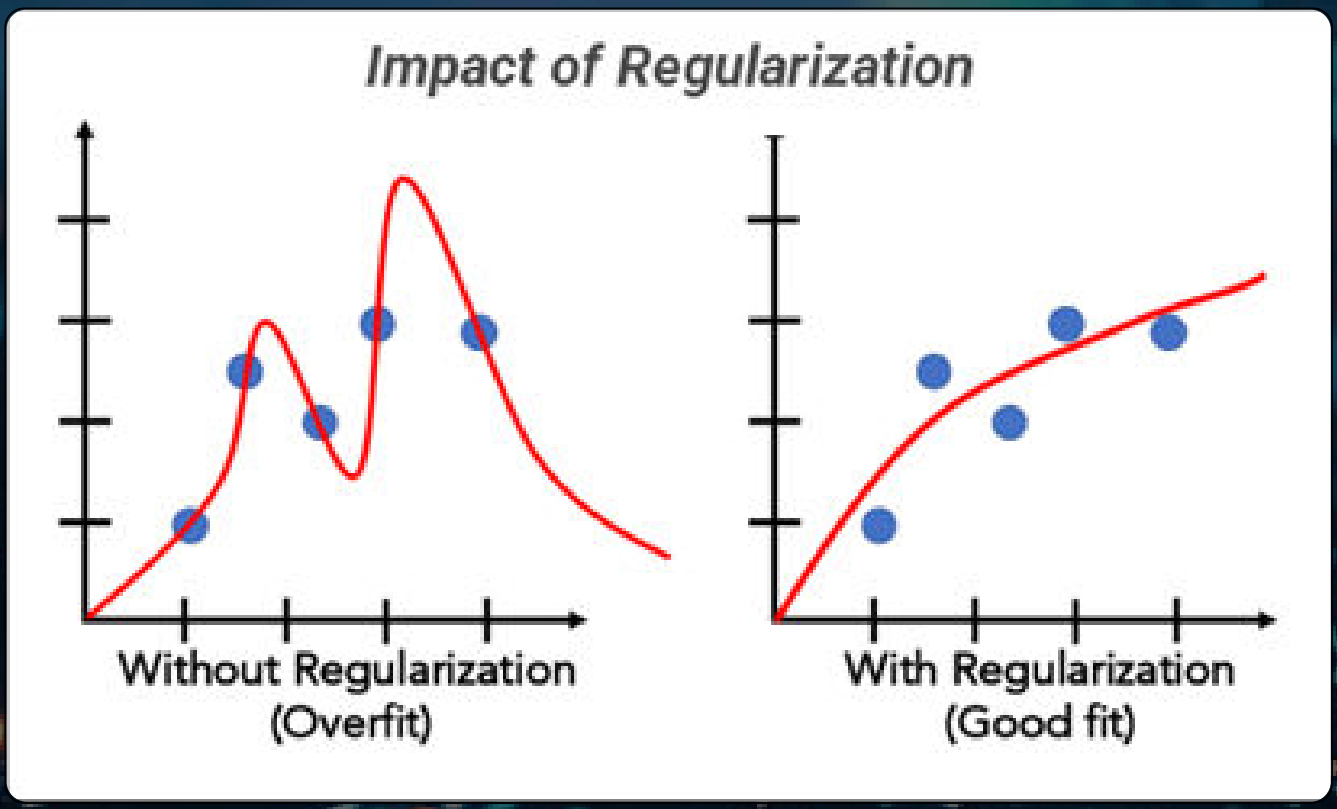
The effectiveness of a neural network model is measured by its ability to make accurate predictions and minimize prediction errors even with new input data that the model has never seen before. This ability is known as generalization.



Overfitting is an error that occurs when a neural network learns too much detail or noise from the training data. This results in poor performance on unseen data and poor generalization performance on test data.

Overfitting may happen because the network is too complex (too many parameters), the training data is too small or too noise, or the network trained too long.

Regularization is a technique used to reduce overfitting and improve a neural network's generalization. It works by introducing penalty terms (weight) on the network's parameters during training which encourage the model to avoid overly complex parameter values. As such, it promotes a balance between model complexity and performance, leading to better generalization on unseen data.



There are five main types of regularization, which are: data augmentation, addition of noise, L1 and L2 regularization, early stopping, and dropout.

Exercise

A. Multiple Choice Questions

1. What is the main of the input layer in a neural network?
A. To output the final result
B. To transform hidden signals
C. To receive and pass data into the network
D. To adjust model weights

2. What does overfitting mean in neural networks?
A. The model performs well on unseen data
B. The network fails to learn any pattern
C. The network learns too much detail from training data and performs poorly on new data
D. The training data is too large for the model to handle

3. How does regularization improve a neural network's performance?
A. By making the model memorize the training data
B. By simplifying the model and preventing overfitting
C. By adding more neurons and layers
D. By skipping the loss calculation step

4. What is the purpose of the hidden layers in a neural network?
A. To display final predictions to the user
B. To take raw input from sensors

C. To process and extract features from input data before passing it to the output layer
D. To store training data

B. Fill in the Blanks

1. A neural network is a machine learning model inspired by the structure and function of the _____.
2. The three main layers of a neural network are the _____ layer, _____ layer, and _____ layer.
3. The ability of a neural network to perform well on new, unseen data is known as _____.

C. True or False

1. Backward propagation is the process where the network adjusts its internal parameters to minimize error.
2. Overfitting happens when a neural network performs well on both training and unseen data.
3. Regularization helps improve generalization by preventing the model from becoming too complex.

D. Arrange the Order

a) Itteration
b) Loss calculation
c) Forward propagation
d) Input
e) Backward propagation

Answer

Multiple Choice Questions

1. C
2. C
3. B
4. C

Fill in the Blanks

1. Human brain
2. Input, hidden, output
3. Generalization

True or False

1. True
2. False
3. True

Arrange

d - c - b - e - a

CH4 Architecture & Models

4.1 Basic Architecture of Neural Networks

Connectionist AI refers to a branch of artificial intelligence that utilizes neural networks to process information similarly to the human brain. Connectionist systems are capable of recognizing patterns, making predictions, and adapting through experience.

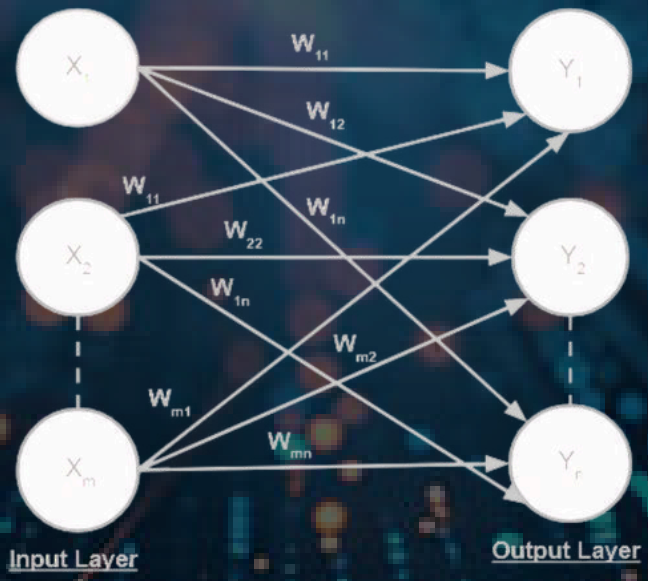
The basic architecture is formed by many interconnected nodes called artificial neurons. Stacking these neurons in layers enables them to learn patterns; this model is called neural network. There are several ways to connect the neurons, and the architecture we choose for a particular solution depends on these options.

4.2 Feed Forward Network (FNN)

4.2.1. Single-layer Feed forward network

The most simple and basic architecture, only consisting of two layers: input and output layers. There is no hidden layer, meaning all the input neurons are connected to the output directly. The computations, like weighted sum, are performed in the output layer. The signals always flow from input → output, which is why it's called Feed Forward.

However, since there are no hidden layers, the network cannot learn complex patterns; it can only solve problems that are linearly separable.

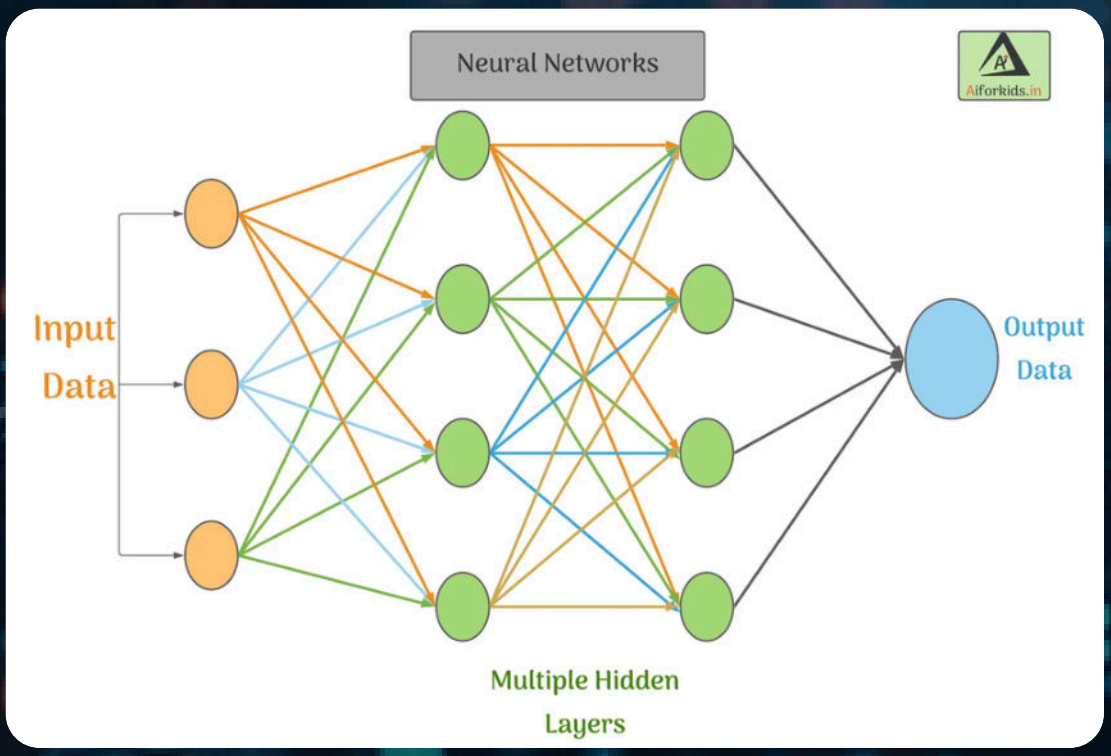


This means it can classify data between two categories only if they are separated by a straight line. Meanwhile, real-world data like images and text are highly non-linear, making single-layer network ineffective.

4.2.2. Multi-layer Feed forward network

The problem with single-layer network is solved with multi-layer, due to the existence of the hidden layers. In multi-layer there are 3 layers, where the input neurons are passed on to the hidden layer first.

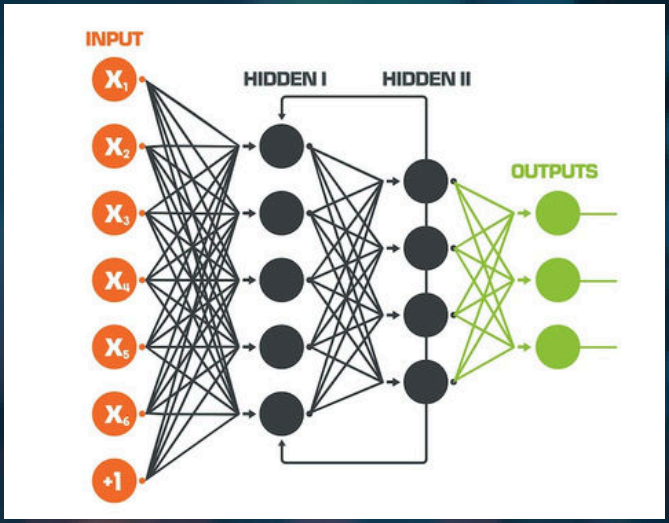
The computations are done in the hidden layers instead of output layer, so more complex features can be built by the network. When the neurons go to the hidden layers, they pass through a nonlinear activation function, and because these functions are stacked together, the network can bend the input space. Thus, the model can produce curved boundaries.

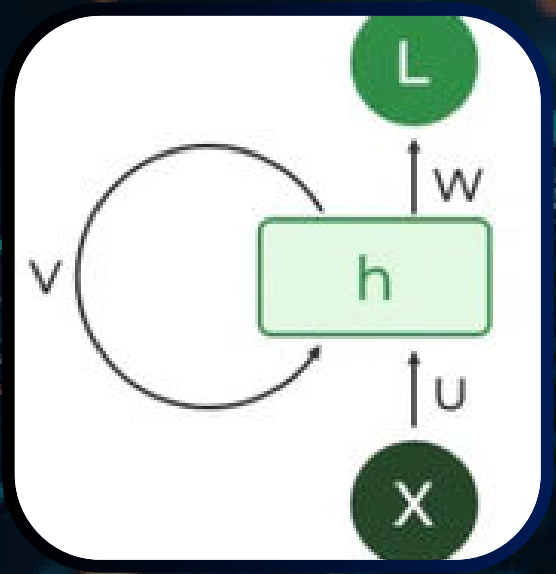


4.3 Recurrent Forward Network (RNN)

A connectionist AI model designed to process sequential data—where the order of data matters. It has a kind of memory called a hidden state that stores information from previous inputs.

Each input is not processed independently, but the output becomes an input in the next step thanks to the feedback loop.



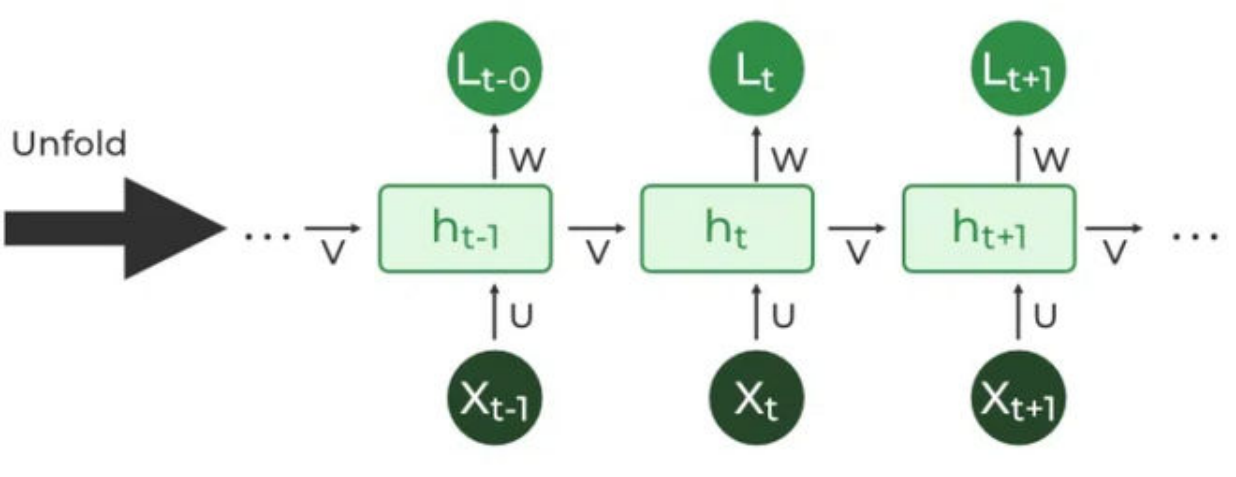


In RNN, there is a process called RNN unfolding—the process of expanding the recurrent structure over time steps. Every stage of the sequence is shown as a distinct layer in a series during unfolding—connected one after another and sharing the same weights.

It is known that hidden state stores past information, which is why the current hidden state (ht) depends on the previous hidden state (ht-1). The hidden state can be updated using the formula below:

$$h_t = f\left(U \times x_t + W \times h_{t-1} + B\right)$$

Where:
f = Activation function
U and W = weight matrices
B = bias



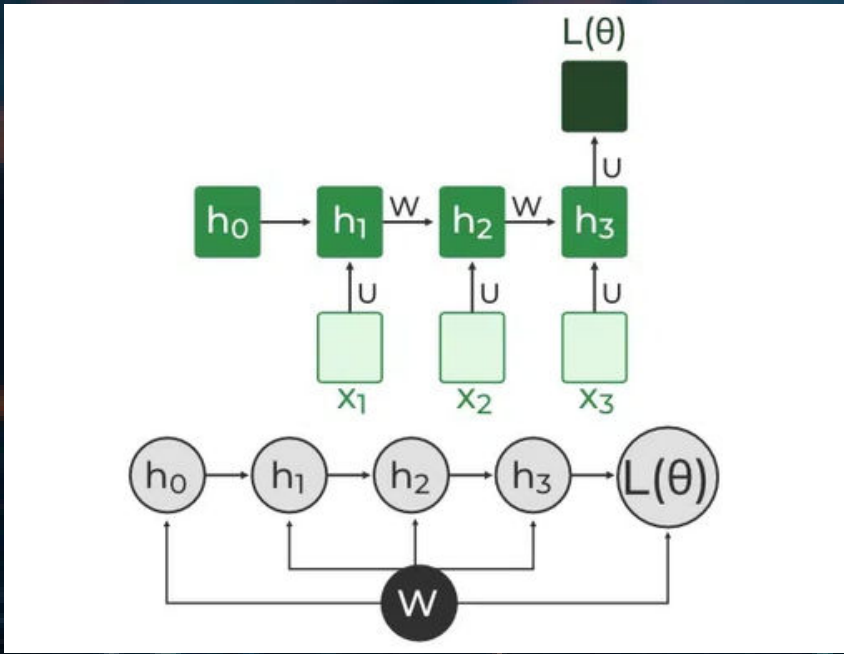
The output (yt) will be calculated from the hidden state at the same time step(ht), using the formula below:

$$y_t = f\left(W \times h_t + B\right)$$

4.3.1 How it works

How an RNN works can be summarized into three short steps:
Receiving input → updating hidden state → producing output.

These chains of steps will continue until the last step, where the final output is made. But after the final output is produced, a loss function (L) measures how accurate the predictions are from the correct values.



The network is "unrolled" over all the time steps, meaning the error (loss) is propagated backward.

It is done through each time step to update the weights, from ht to ht-1, ht-2, and so on. This is called the Backpropagation Through Time (BPTT). After all gradients are computed, the RNN's weights (shared across time steps) are updated to minimize the overall loss.

4.3.2. Limitation

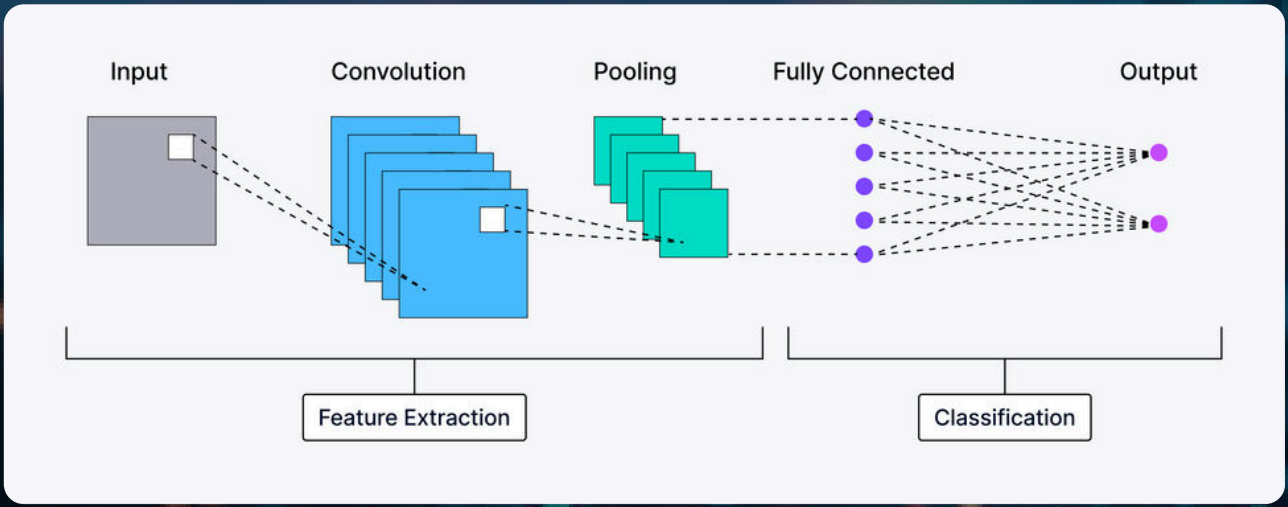
When BPTT is performed, gradients are multiplied by recurrent weights (W) repeatedly every time it passes through each time step.

If W < 1, the gradients shrink exponentially, potentially becoming close to zero. As a result, RNN can only learn short patterns. This problem is called the vanishing gradient problem.

On the other hand, if W > 1, the gradients increase exponentially, potentially becoming very large. As a result, there might be instability in the model—the result is not accurate. This is known as the exploding gradient problem.

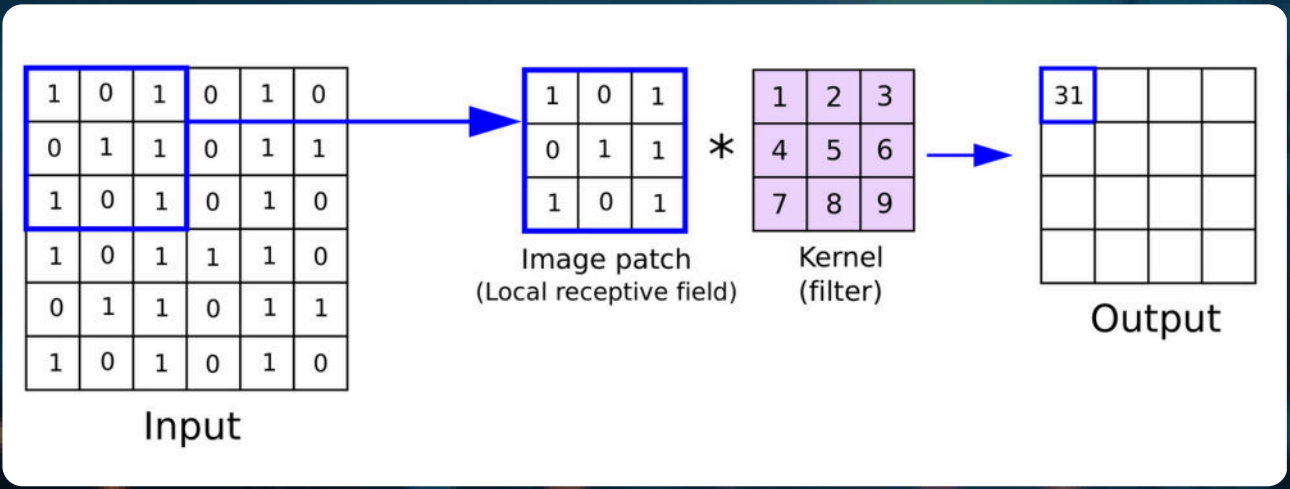
4.4 Convolutional Forward Network (CNN)

A deep learning framework, primarily made to handle data that has a grid-like matrix. It can extract features from spatial patterns (edge and texture), making it useful for visual datasets like images or videos. Due to the complex processes, there are more layers in CNN compared to FNN and RNN, 5 in total: input, convolutional, activation, pooling, fully connected, and lastly output layer.



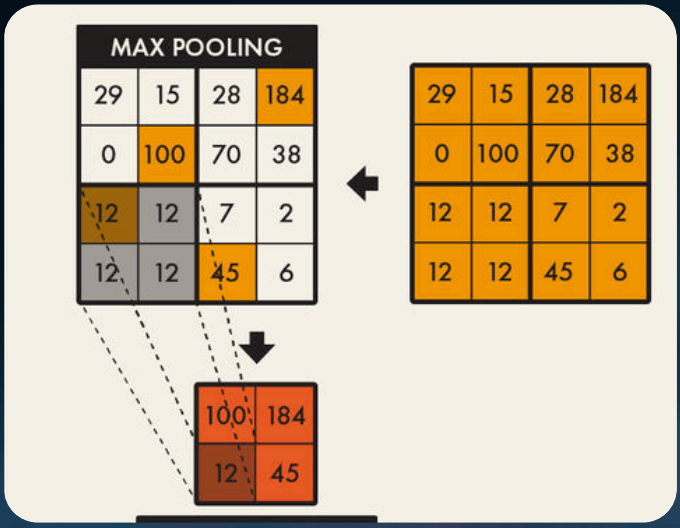
It first starts with the input layer, where raw pixel data are taken. Then, through the convolutional layer which consists of kernels - small matrix of weights - that have small width and height but same depth as the input volume (e.g., 3x3x3 or 5x5x3 in a 34x34x3 image dimension).

The kernel is moved across the whole input volume step by step, where each step is called a stride, scanning the image piece by piece. During each stride, the dot product will then be calculated between the kernel weights and input patch. The 2D output of the feature will be created in feature maps. But since a kernel/filter only extract a specific feature, multiple filters will be needed to detect different type of features.



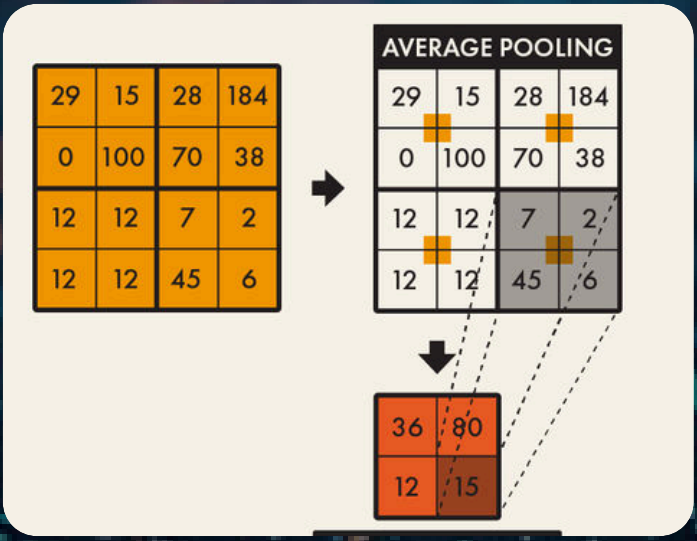
Activation function will then be implemented to introduce non linearity, usually ReLU or sigmoid function. After going through the activation function, it reaches the pooling layer. Its main purpose is to reduce the size of feature maps. Processing every pixel from every feature maps requires a lot of computing.

To make computing more efficient, there are two ways to reduce the data volume, namely max and average pooling. Suppose we use 2x2 pooling with stride = 2 like the pictures below:



In max pooling, the maximum value in each region is taken (e.g., 100 is chosen from 29, 15, 0, 100).

In average pooling, the average value is taken from each region (e.g., 36 is calculated from adding 29, 15, 0, and 100 and then divided by 4)



Since there are 4 regions, the pooling will be done 4 times. It is then reshaped into a 2 by 2 grid.

Lastly, the fully connected layer, which is responsible for flattening the output into a 1D vector. Then the final classification or prediction is produced by combining all of the features that were gathered.

Exercise

Multiple Choice Questions

1. Which of the following neural networks is most suitable for image recognition?
A. Single-layer Feedforward Neural Network
B. Convolutional Neural Network (CNN)
C. Recurrent Neural Network (RNN)
D. Multi-layer feedforward neural Network

2. The hidden state in an RNN is mainly used to:
A. Produce the final output of the network
B. Store information from previous time steps
C. Randomly initialize the weights
D. Reduce the size of the input data

3. In a Feedforward Neural Network, what happens if there are no hidden layers?
A. The model can only solve linear equations
B. The model gains the ability to recognize complex patterns
C. The model cannot be trained
D. The model can be used to analyze images

4. In a CNN, what does the **pooling layer** primarily do?
A. Adds non-linearity using activation functions
B. Reduces the spatial size of feature maps to make a faster model
C. Computes the loss function for training
D. Producing the output in feature maps

5. Which of the following best explains the vanishing gradient problem in RNN?
1. The gradient remains constant regardless of the weight value.
2. The gradient fluctuates randomly
3. The gradient increases exponentially, caused by a weight higher than 1.
4. The gradient shrinks exponentially, caused by a weight lower than 1.

Fill in the Blanks

1. A _____ neural network allows information to flow in one direction, from input to output, without forming cycles.
2. In RNNs, the _____ from the previous step is used as an input for the next step, keeping memory of past information.
3. The main limitation of a single-layer feedforward network is that it cannot learn _____ relationships.
4. In CNN, the layer responsible for flattening the features into a one-dimensional vector is the _____ layer.

Answers

Multiple Choice Questions

1. B
2. B
3. A
4. B
5. D

Fill in the blanks Questions

1. Feedforward
2. Hidden state
3. Nonlinear
4. Fully connected

CH5 Applications Of Connectionist AI

5.1 Healthcare

1) Disease Prediction

What it does

Uses patient data to predict the likelihood of developing chronic diseases such as heart disease, diabetes, or cancer. Helps healthcare providers act early and personalize preventive care.

How it works

Neural networks analyze patterns in EHR and lifestyle data to forecast disease onset.

Example

AI trained on Framingham data predicts cardiovascular risk.

Benefits

Enables early diagnosis and preventive treatment.

2) Medical Imaging Analysis

What it does

Automatically detects and classifies medical anomalies in scans such as X-rays, CTs, and MRIs. Assists radiologists in identifying tumors, fractures, or organ damage.

How it works

Convolutional Neural Networks (CNNs) identify patterns in image data.

Example

Google Health AI detects breast cancer from mammograms.

Benefits

Speeds up diagnosis, improves accuracy, reduces human error

3) Drug Discovery

What it does

Accelerates the process of finding and designing new drug compounds by simulating how molecules interact with biological targets.

How it works

Deep learning models simulate molecule-protein interactions.

Example

Atomwise uses neural networks to screen antiviral compounds.

Benefits

Reduces R&D time and cost.

5.2 Finance

1) Fraud Detection

What it does

Detects unusual or unauthorized financial activities by monitoring transactions in real time. Helps prevent identity theft and credit card fraud.

How it works

Neural networks detect abnormal patterns in transaction behavior.

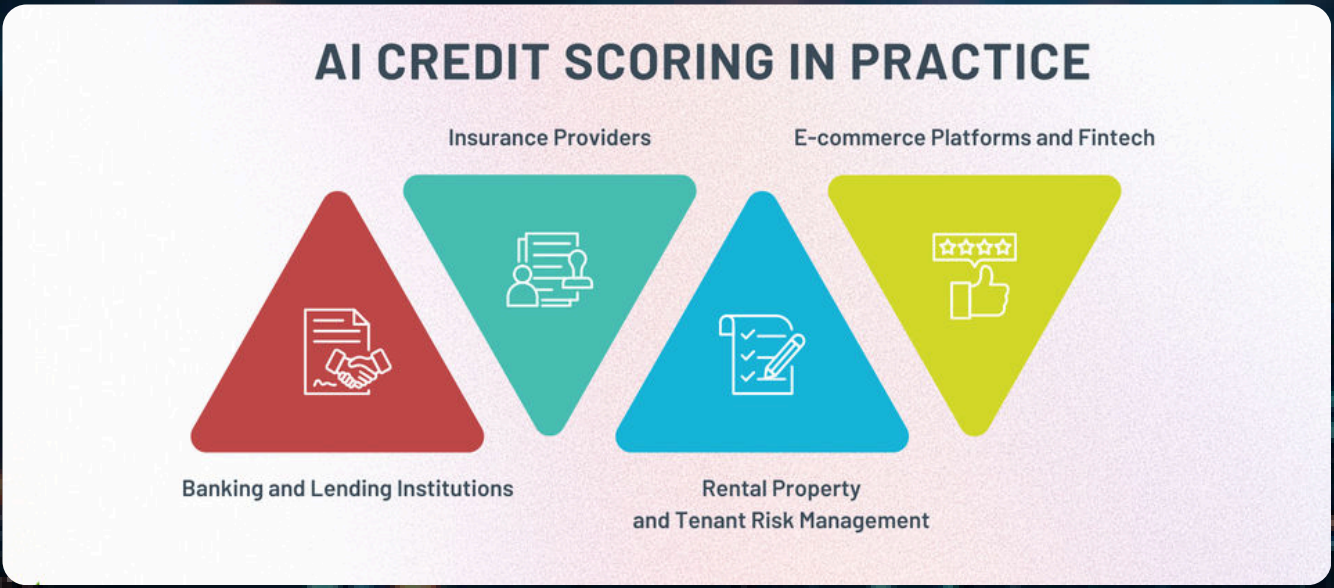
Example

JPMorgan Chase uses AI to catch fraud in real time.

Benefits

Prevents loss and protects customer trust.

2) Credit Scoring



What it does

Assesses a person's creditworthiness using a broader range of data including social, financial, and behavioral patterns, not just traditional credit scores.

How it works

AI analyzes both traditional and alternative data sources.

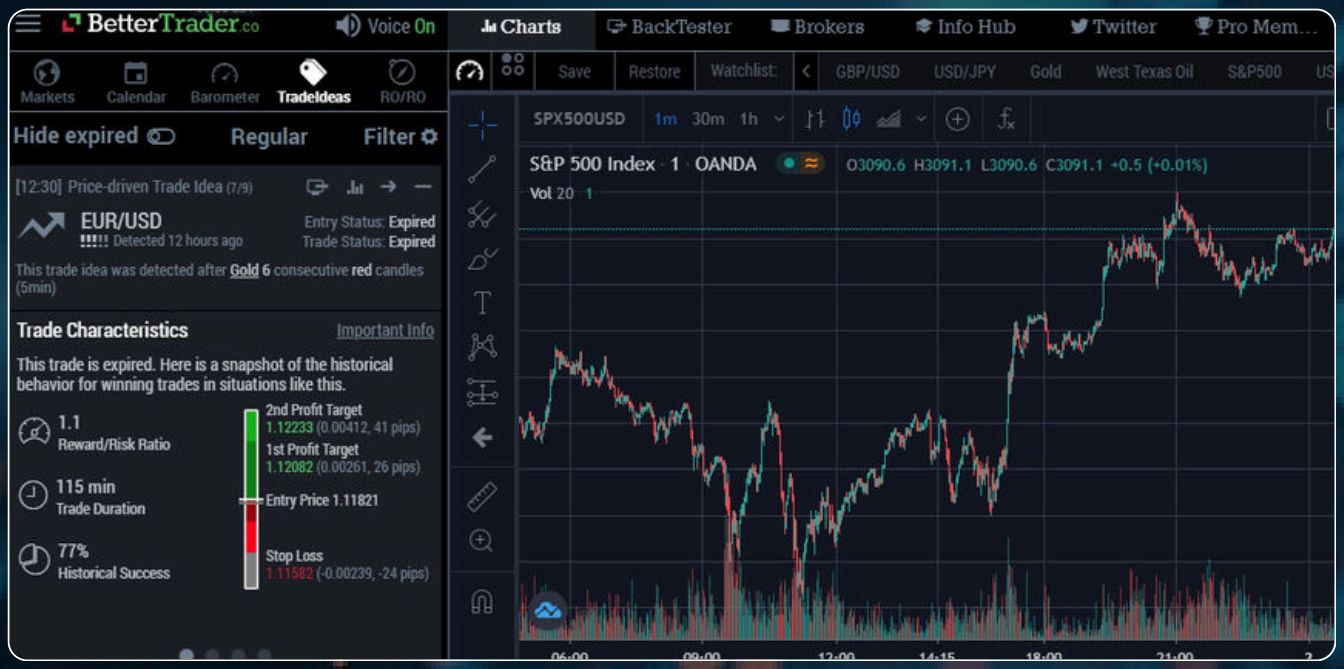
Example

Lenddo scores unbanked individuals using mobile data.

Benefits

Expands credit access and reduces loan default risk.

3) Algorithmic Trading



What it does

Automates financial trading by analyzing market trends, predicting price changes, and executing buy/sell orders within milliseconds

How it works

Deep learning models predict asset price movements.

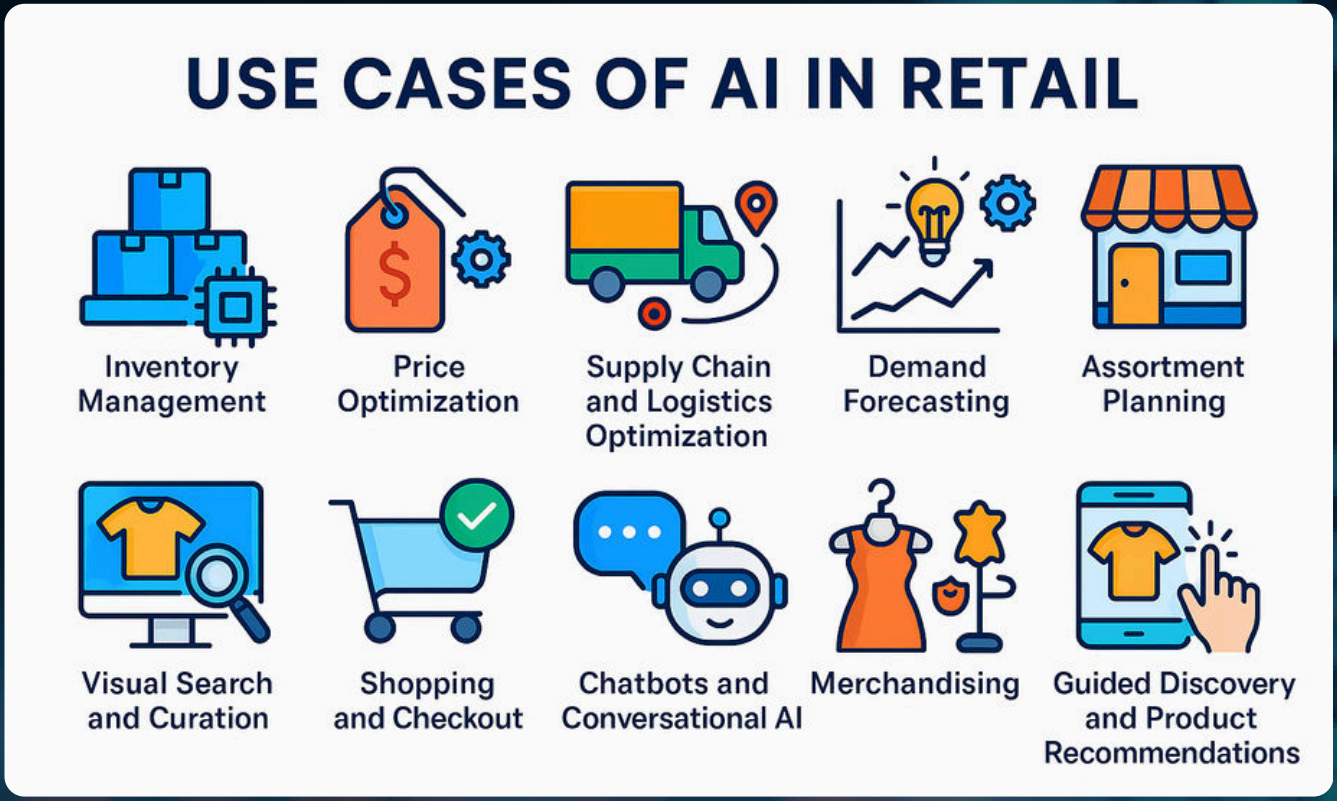
Example

Quant funds use neural networks to react to market signals.

Benefits

Enables high-frequency, data-driven investment strategies

5.3 Retail



1) Product Recommendation

What it does

Suggests products to customers based on their browsing and purchase history, improving engagement and sales.

How it works

Neural networks learn user preferences from browsing and purchase history.

Example

Amazon's recommendation engine.

Benefits

Increases sales and enhances user experience.

2) Customer Sentiment Analysis

What it does

Evaluates customer feedback and online reviews to understand public perception, identify product issues, and refine marketing.

How it works

NLP models extract emotion and feedback trends.

Example

Retailers use sentiment analysis to refine product design.

Benefits

Improves products and brand perception.

3) Inventory Forecasting

What it does

Anticipates future inventory requirements to maintain availability and minimize overstock based on historical data and market trends.

How it works

AI models analyze historical and real-time sales data.

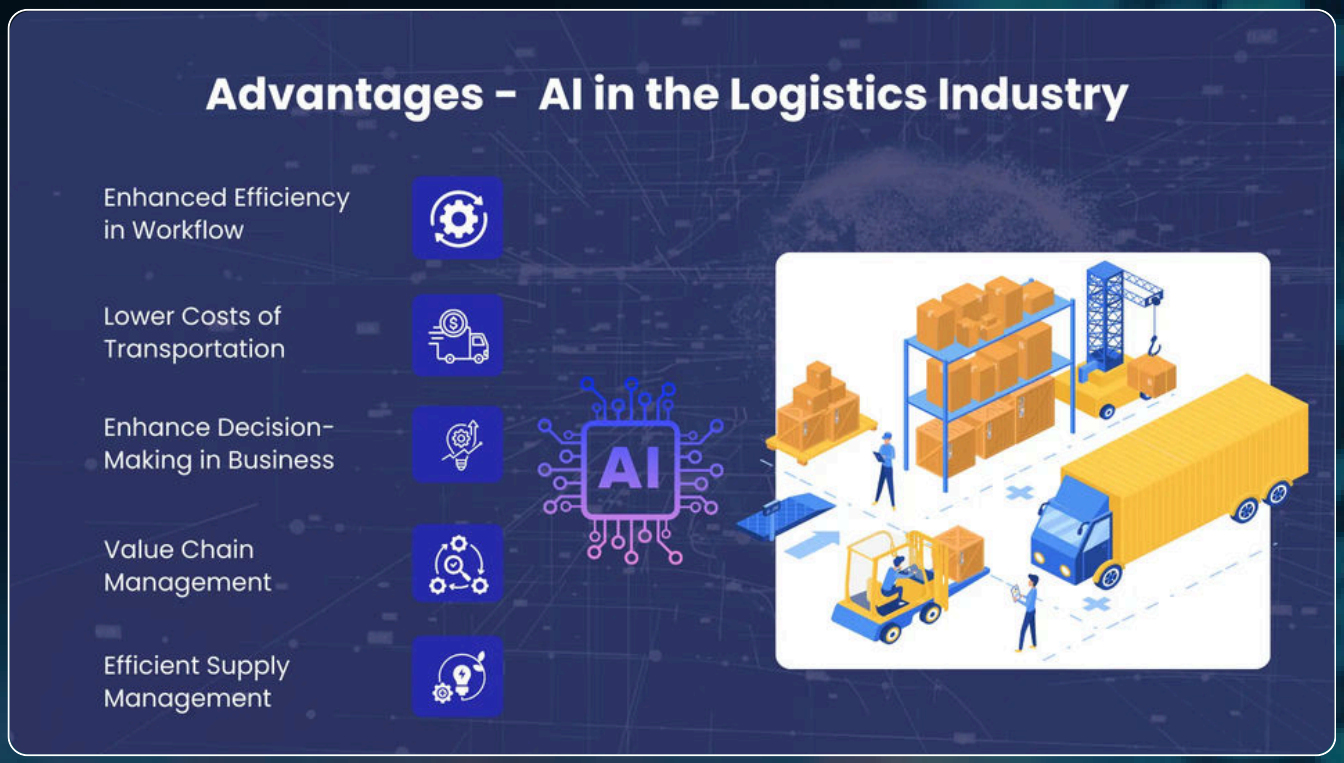
Example

Retailers use neural networks to restock efficiently.

Benefits

Reduces stockouts and overstock.

5.4 Logistics



1) Predictive Maintenance

What it does

Forecasts potential failures in vehicles or machines before they occur by monitoring sensor data.

How it works

Neural networks monitor sensor data from trucks/machines.

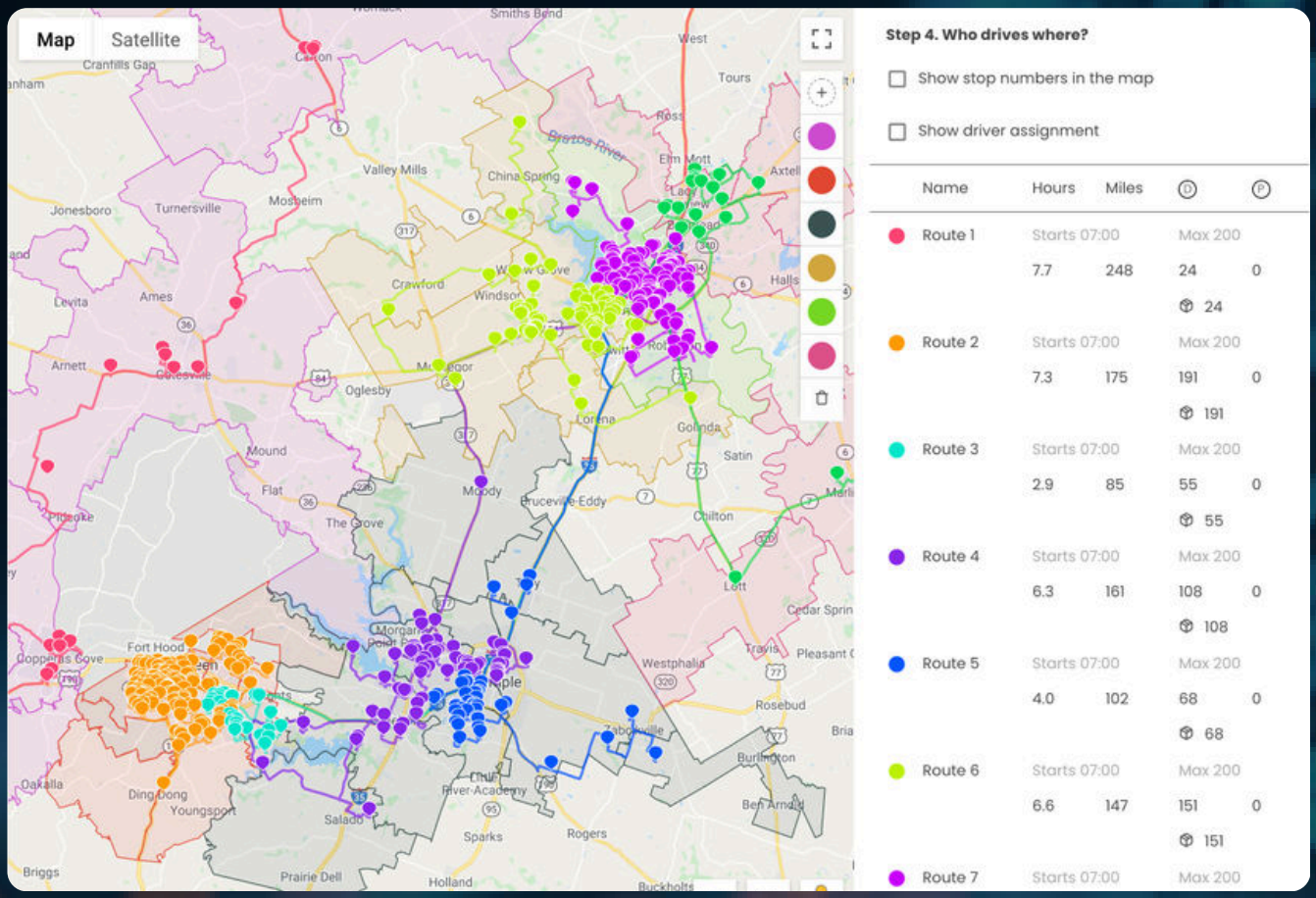
Example

Logistics companies use AI to avoid truck breakdowns.

Benefits

Reduces downtime and repair costs.

2) Route Optimization



What it does

Plans the most efficient delivery routes by analyzing real-time traffic, weather, and road conditions.

How it works

Neural networks evaluate traffic, weather, and delivery patterns.

Example

AI reroutes deliveries to avoid delays.

Benefits

Saves fuel and delivery time.

Exercise

A. Multiple Choice Questions

1. What is a primary use of Connectionist AI in medical imaging?
A. Generating synthetic scans
B. Detecting anomalies in diagnostic images
C. Recording patient feedback
D. Scheduling appointments

2. Which AI technology helps retailers suggest products based on browsing history?
A. Blockchain
B. Decision Trees
C. Neural Networks
D. Genetic Algorithms

3. In logistics, how does Connectionist AI support predictive maintenance?
A. It estimates customer delivery times
B. It scans QR codes for tracking
C. It monitors sensor data for machine failure
D. It sends marketing emails

4. Which of the following is NOT a benefit of using Connectionist AI in finance?
A. Real-time fraud detection
B. Better credit risk analysis

C. Weather prediction
D. Personalized investment advice

B. Fill in the Blanks

1. Connectionist AI enables _____ learning from data, which is useful in tasks such as medical diagnosis, _____ analysis, and _____ recommendations.
2. Connectionist AI uses _____ networks to learn from _____ and make predictions in areas like _____ and dynamic pricing.

C. True or False

1. Connectionist AI is only useful in structured datasets like spreadsheets or databases.
2. Neural networks can be used to predict future inventory needs in retail businesses.
3. Connectionist AI is incapable of handling image or audio data.

D. Arrange the Order

Arrange the steps below in the correct order:
a) AI analyzes patient data
b) Optimal treatment is recommended
c) Medical history and genetics are collected
d) Outcome is monitored and fed back into the system

Answer

Multiple Choice Questions

1. B
2. C
3. C
4. C

Fill in the Blanks

1. Pattern, trend, personalized
2. Neural, large datasets, demand forecasting

True or False

1. False
2. True
3. False

Arrange the Order

c - a - b - d

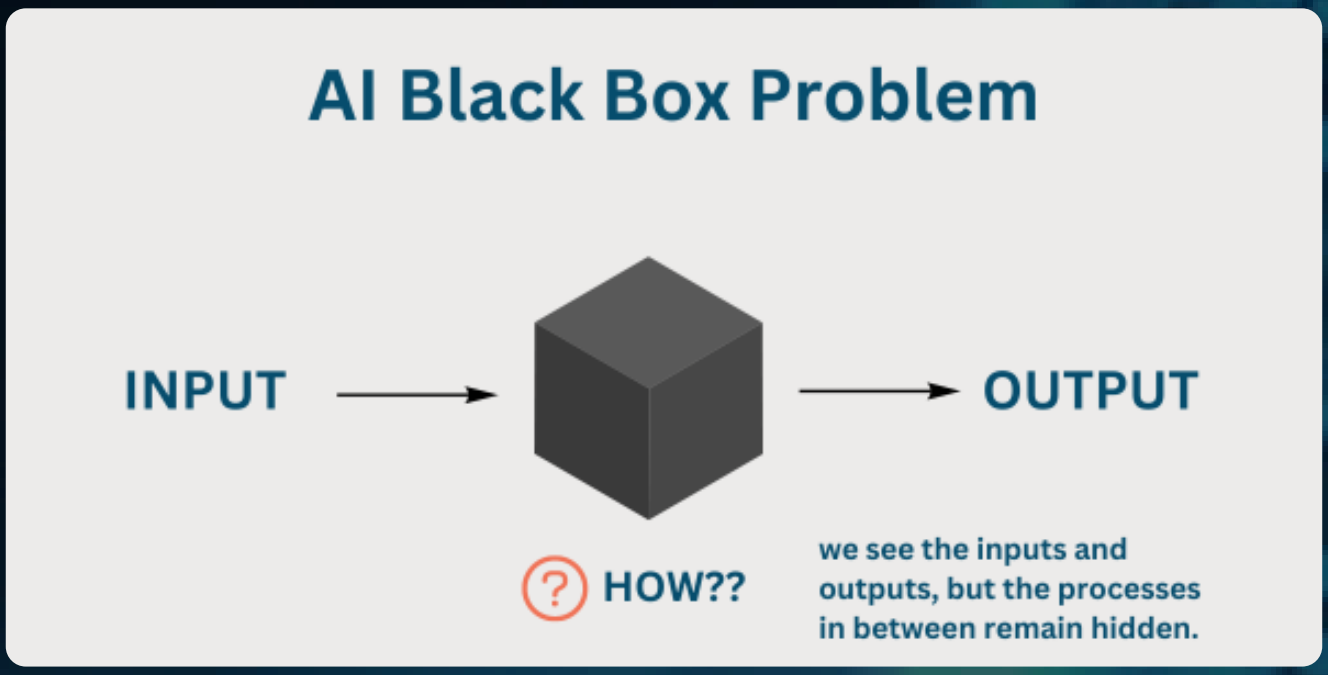
CH6 Connectionist AI Challenges and Future Directions

6.1 Connectionist AI Challenges

Connectionist AI, mainly neural networks, simulate interconnected artificial neurons to recognize patterns and learn from data. Despite their successes, they face several deep, ongoing challenges:

1) Lack of Interpretability (The Black Box Problem)

Connectionist models work through complex weighted connections distributed across many neurons. This internal representation is opaque to humans, making it extremely difficult to understand or explain exactly how decisions are made.



Unlike Symbolic AI, which uses explicit logical rules, connectionist AI provides no clear reasoning path — leading to trust and accountability issues, especially in critical domains.

What is the Black Box Problem?

Connectionist AI (like deep neural networks) makes decisions through complex layers of interconnected neurons. The internal process—how the model transforms inputs into outputs—is hidden and not understandable to humans.

This lack of transparency is called the "black box problem."

Why is it a Problem?

Trust: Users and developers can't easily trust or verify AI decisions without understanding their reasoning.

Accountability: In critical areas (e.g., autonomous cars, healthcare, credit scoring), unexplained AI decisions make responsibility and liability unclear.

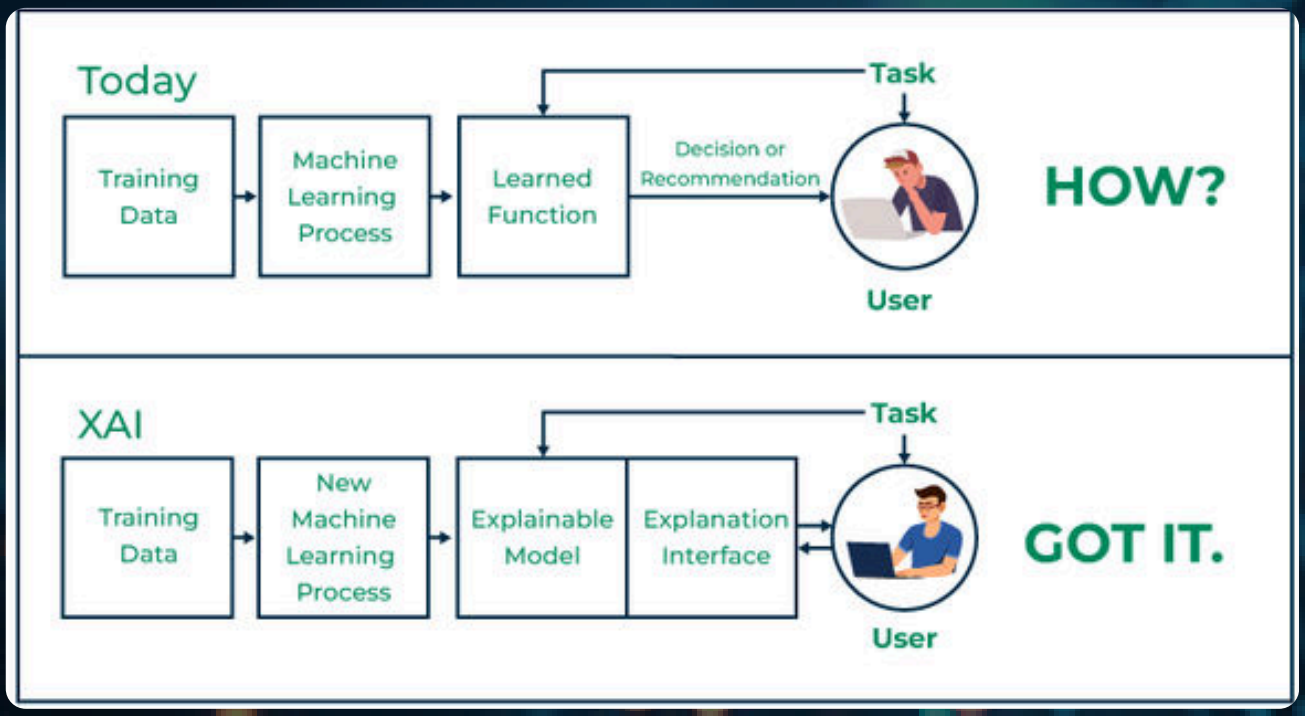
Error diagnosis: When AI fails (e.g., crashes or wrong predictions), it's hard to trace and fix the cause.

Example:

If a self-driving car hits a pedestrian, we cannot easily determine why the system failed because the AI's decision-making path is not visible. This limits how quickly the system can improve or be safely deployed in varied conditions.

Challenges in Solving the Black Box Problem:
Neural networks learn distributed representations, not explicit rules. The sheer number and complexity of parameters make it difficult to extract meaningful explanations. Training data can never cover every possible situation, so unexplained failures can still happen.

Current Approaches to Mitigate It:



Explainable AI (XAI): Techniques that try to approximate or visualize parts of the model's reasoning (e.g., feature importance, attention maps).

Hybrid Neuro-Symbolic Systems: Combining neural networks with symbolic logic to gain better transparency.

2) Difficulty with Abstract and Symbolic Reasoning

What Is It?

Abstract reasoning involves manipulating complex ideas and relationships beyond direct sensory inputs. Symbolic reasoning uses explicit symbols and strict rules to perform logical operations like mathematics, or language syntax.

Challenges in Connectionist AI:

- Uses distributed representations rather than explicit symbols, making direct manipulation difficult.
- Lacks systematicity (ability to combine known concepts into new configurations) and productivity (handling novel combinations).
- Poor at multi-step logical inference and formal reasoning.
- Represents knowledge as statistical correlations, lacking explicit hierarchical or causal structures.
- Relies heavily on training examples and struggles with truly novel problems.

3) Learning Generalization and Catastrophic Forgetting in Connectionist AI

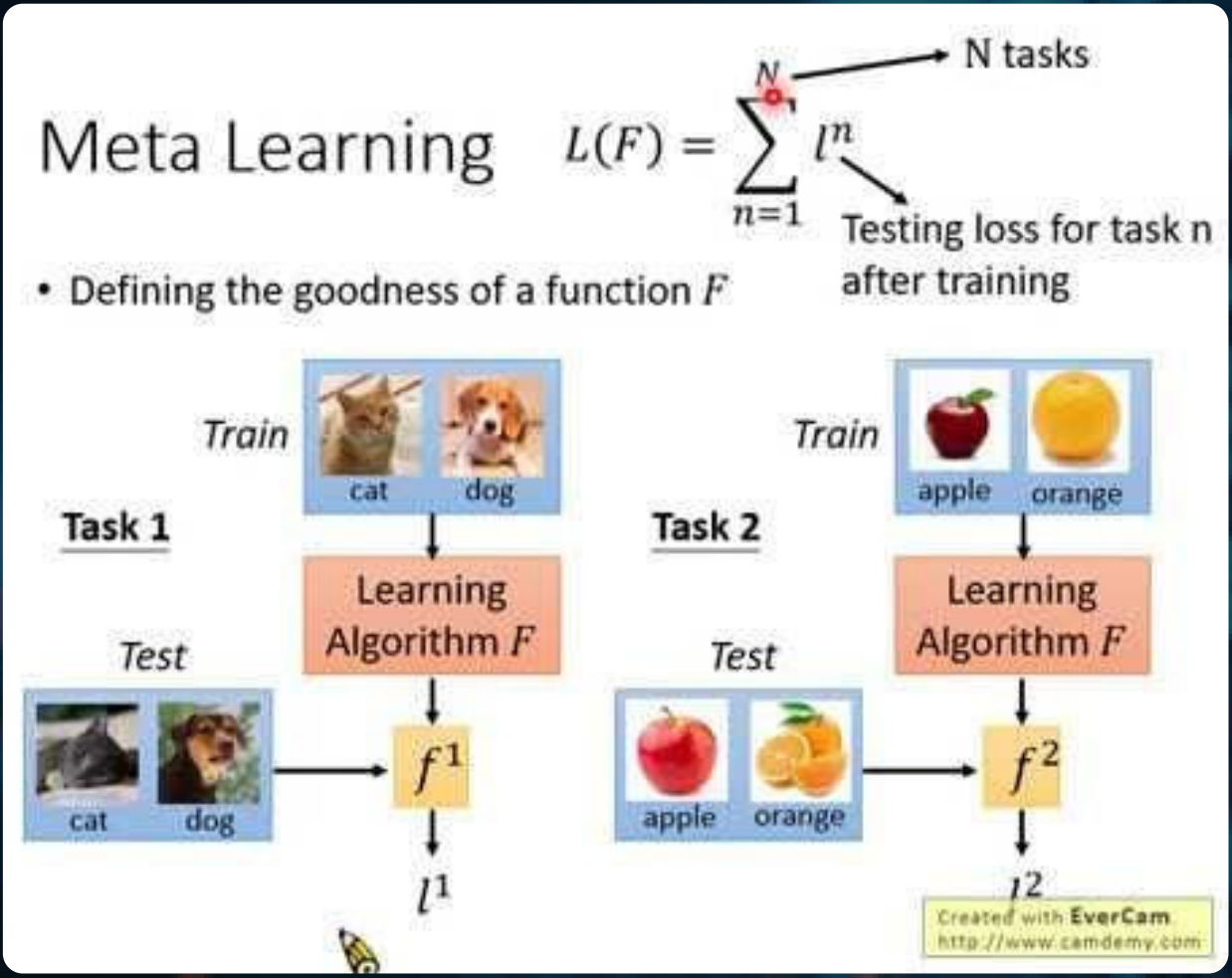
Generalization is an AI system's ability to apply what it has learned from training examples to new, unseen inputs that follow the same underlying rules or patterns.
It means solving new problems based on prior knowledge without needing explicit retraining for every variation.

Limited Systematic Generalization:
Neural networks often struggle when test examples are constructed by new combinations of known components (called compositional or systematic generalization).
For example, a model trained to recognize “red square” and “blue circle” may fail to correctly interpret “blue square” if it hasn't explicitly seen it before.

Data Hunger:
To cover all possible input variations with sufficient examples, an impractically large number of training data points is often required.
Without massive data, models may fail to generalize well to novel stimuli.

Research Directions and Solutions

- **Meta-learning:** Training models to learn how to learn, enabling them to generalize better to new tasks with fewer examples.



- **Complementary Learning Systems:** Inspired by the brain, combining fast learning (hippocampus-like) and slow, stable learning (neocortex-like) to reduce forgetting.
- **Architectural innovations:** Methods like regularization, rehearsal (replaying past examples), and memory-augmented neural networks help mitigate forgetting.

6.2 Connectionist AI Future Directions

1) Adaptive and Meta-Learning Architectures in Connectionist AI

Adaptive AI architectures refer to systems that can dynamically modify their structure, learning rules, or parameters based on new information or changing environments.

Meta-learning (or "learning to learn") is a subfield of machine learning where models learn from a variety of tasks to develop the ability to quickly adapt to new problems with minimal data and retraining.

How Meta-Learning Works

The AI is trained over multiple tasks (meta-training) to extract generalizable knowledge about learning itself.
When faced with a new task (meta-testing), it adapts rapidly using its learned experience.

This approach contrasts with traditional models trained on a specific task and dataset, enabling greater flexibility and efficiency.

Benefits

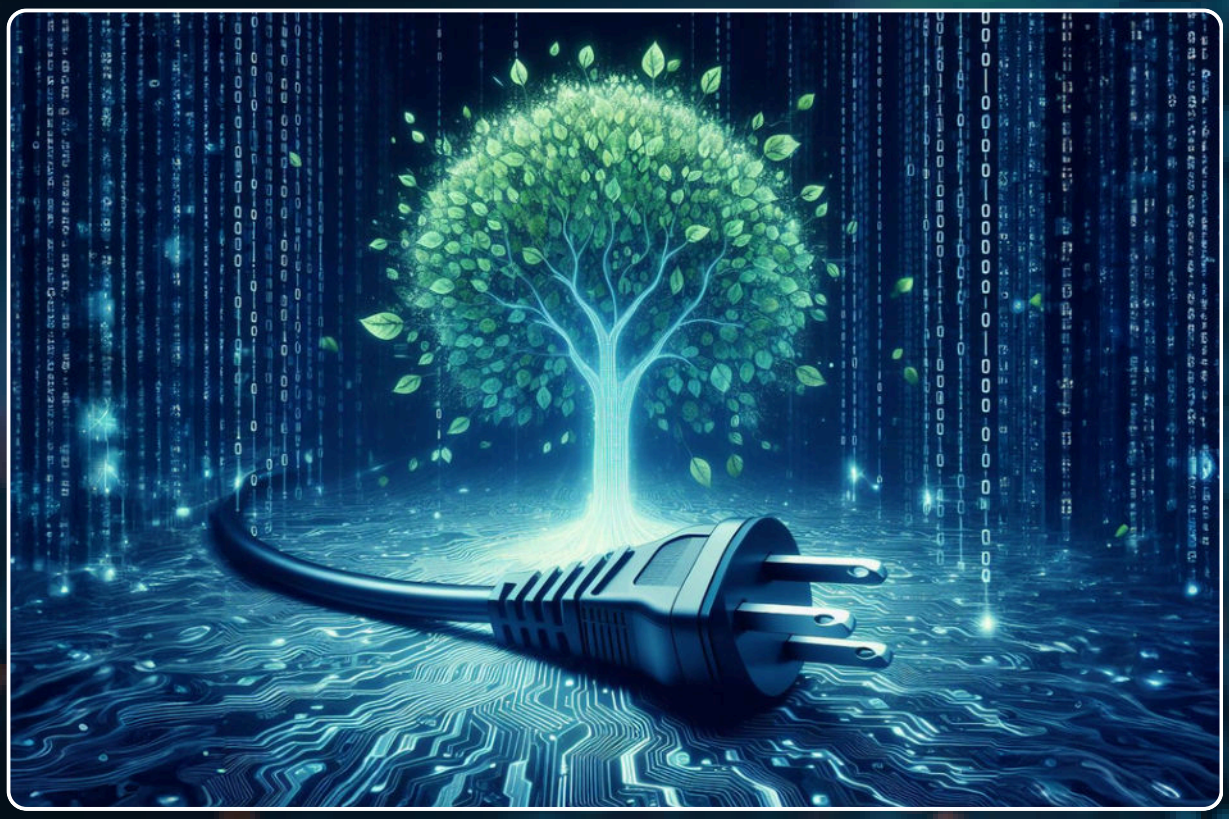
- Fast adaptation to new domains or tasks with few examples.
- Improved generalization beyond training data.
- Reduces the need for large labeled datasets in each new application.
- Enhances AI's ability to function in dynamic or uncertain environments.

Challenges

- Ensuring robustness when new tasks differ significantly from training tasks.
- Improving interpretability of the meta-learning process.
- Extending meta-learning for continual and lifelong learning scenarios.
- Combining with modular and neuro-symbolic approaches for enhanced flexibility.

2) Energy-Efficient Neural Computing

Training and running large neural networks require enormous amounts of computational power and energy. In 2025, energy consumption by AI data centers is approaching levels comparable to the electrical usage of entire countries, raising environmental and cost concerns.



Why Energy Efficiency Matters

Environmental impact: High energy use contributes to carbon emissions and climate change.

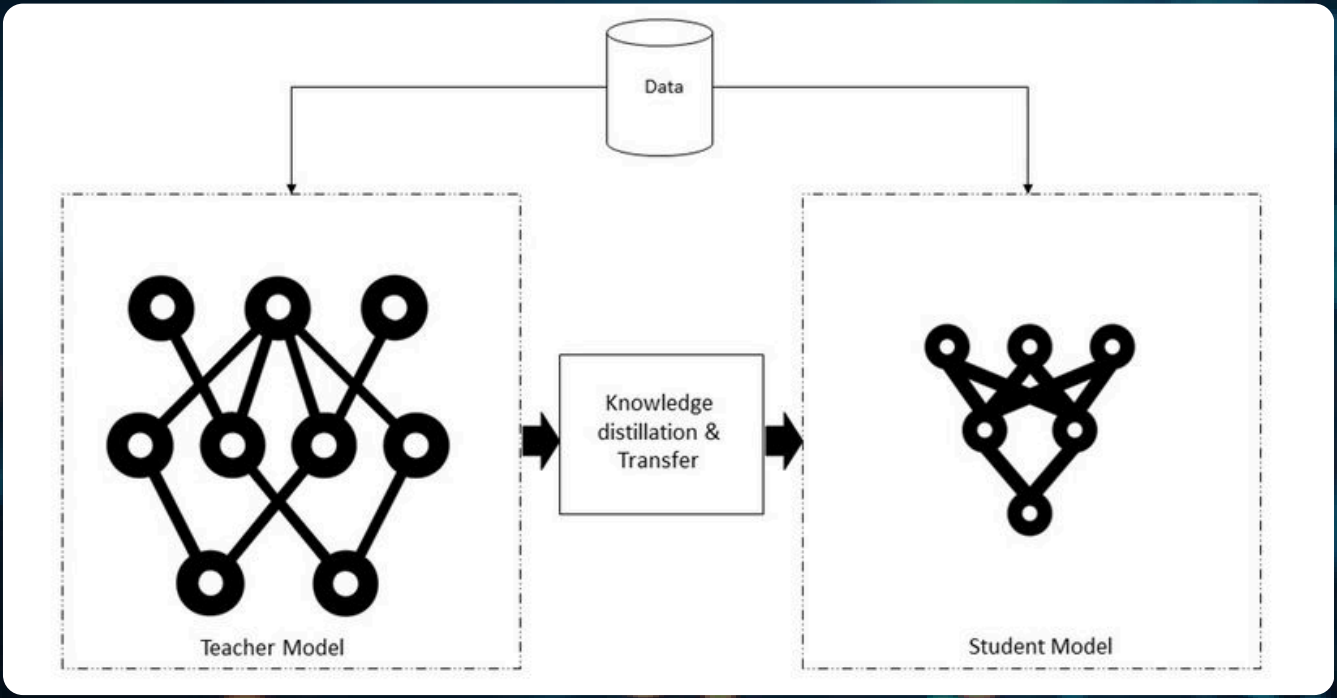
Cost: Data center and hardware running costs limit accessibility and scalability.

Feasibility: Lower energy consumption enables AI deployment on edge devices like smartphones, robots, and autonomous vehicles.

Techniques for Energy Efficiency

a) Model Compression

- Pruning: Removing redundant neurons/connections after training.
- Quantization: Reducing precision of model weights and activations from 32-bit floats to smaller bit-widths.
- Knowledge Distillation: Training smaller "student" models to imitate larger "teacher" models.



b) Hardware Optimizations

Development of specialized AI chips (neuromorphic, FPGA-based systems) optimized for low energy consumption. Integration of memory and processing units to reduce data movement energy costs.

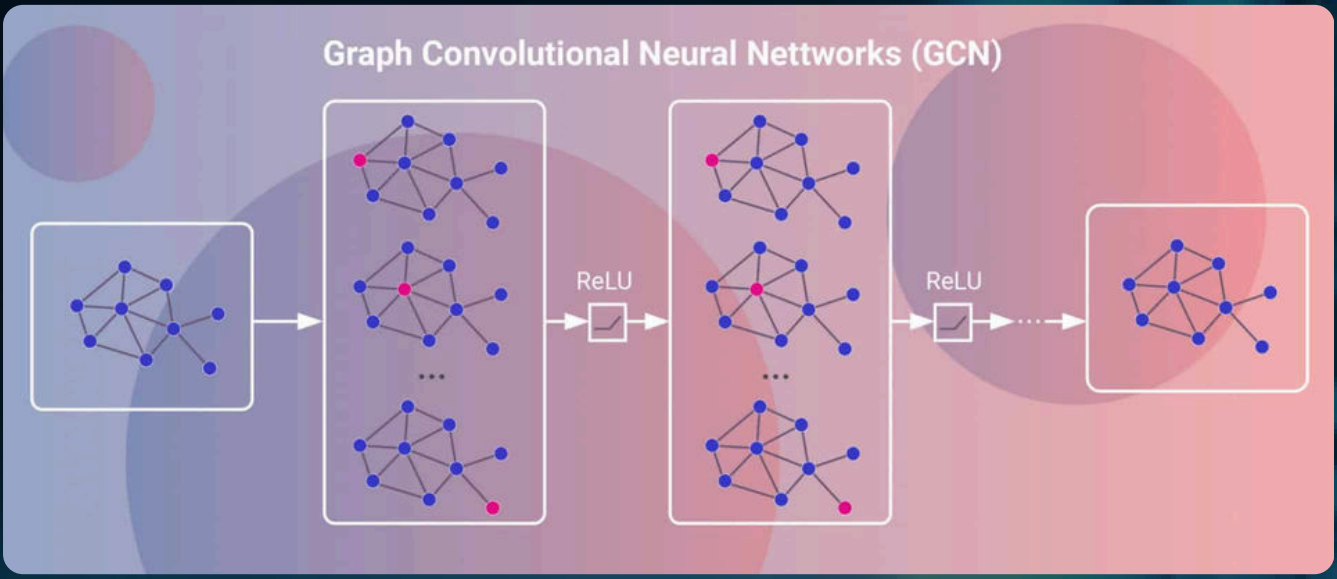
c) Efficient Training Algorithms
Faster convergence methods that reduce training epochs. Layer freezing to avoid retraining stable parts of networks. Early stopping to limit unnecessary training cycles.

d) Neuromorphic Computing Mimics the brain's architecture integrating memory and computation. Significantly reduces energy use while performing complex tasks.

Future Directions

- Scaling energy-efficient AI to large foundation models.
- Combining hardware and software innovations for maximal efficiency.
- Developing open standards for benchmarking AI energy consumption.
- Increasing research investments into sustainable AI infrastructure.

3) Graph Neural Networks and Complex Data Structures



GNNs are neural networks designed specifically to process graph-structured data, which consists of nodes (vertices) and edges (connections).

Unlike traditional neural networks optimized for fixed-grid data like images, GNNs handle graphs with variable sizes, irregular connectivity, and diverse node features.

They achieve this by passing information along edges to learn representations that combine structural and feature data for each node.

Applications and Tasks

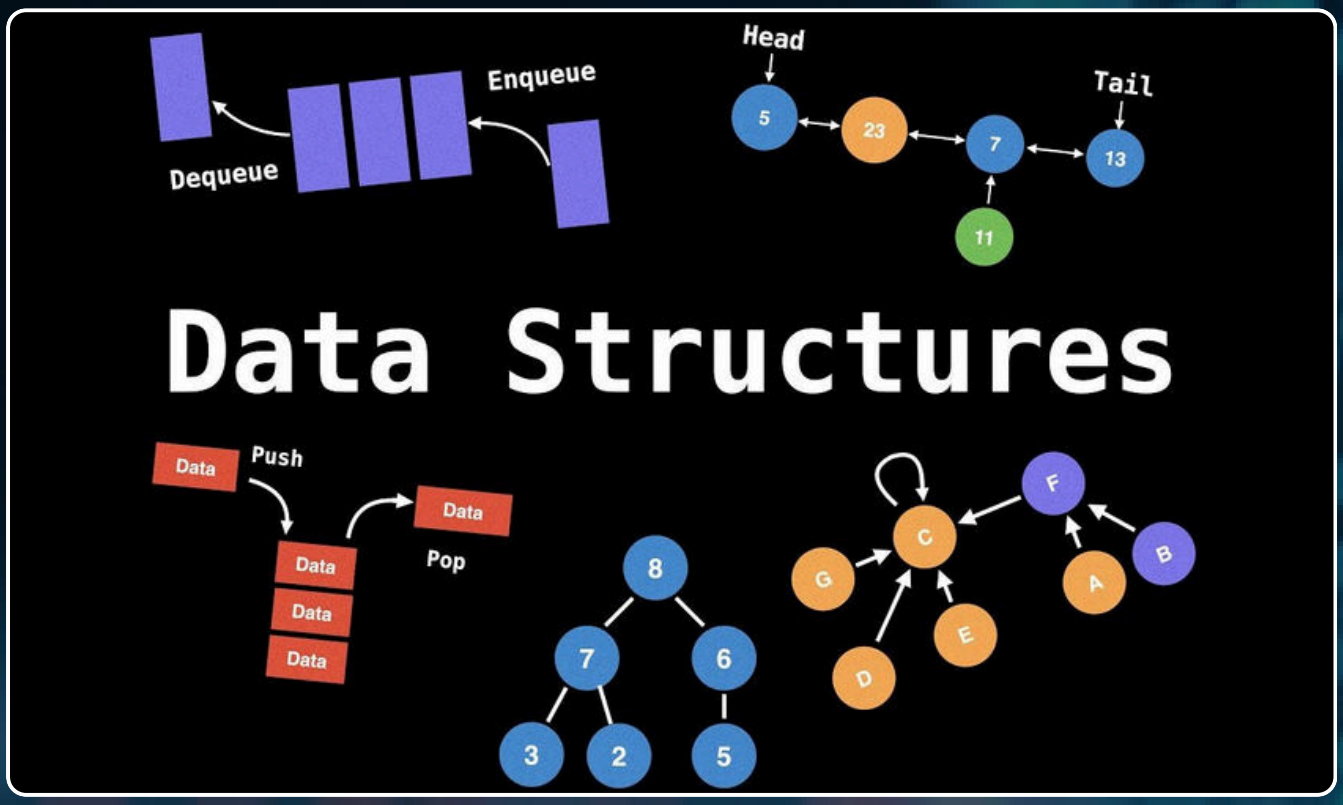
Node Classification: Assign labels to individual nodes based on their features and neighborhood (e.g., identifying influential users in social media).

Edge Prediction: Predict likelihood or properties of links between nodes (e.g., friend suggestions or molecular bonding).

Graph Classification: Classify the entire graph structure, such as predicting properties of molecules or social community types.

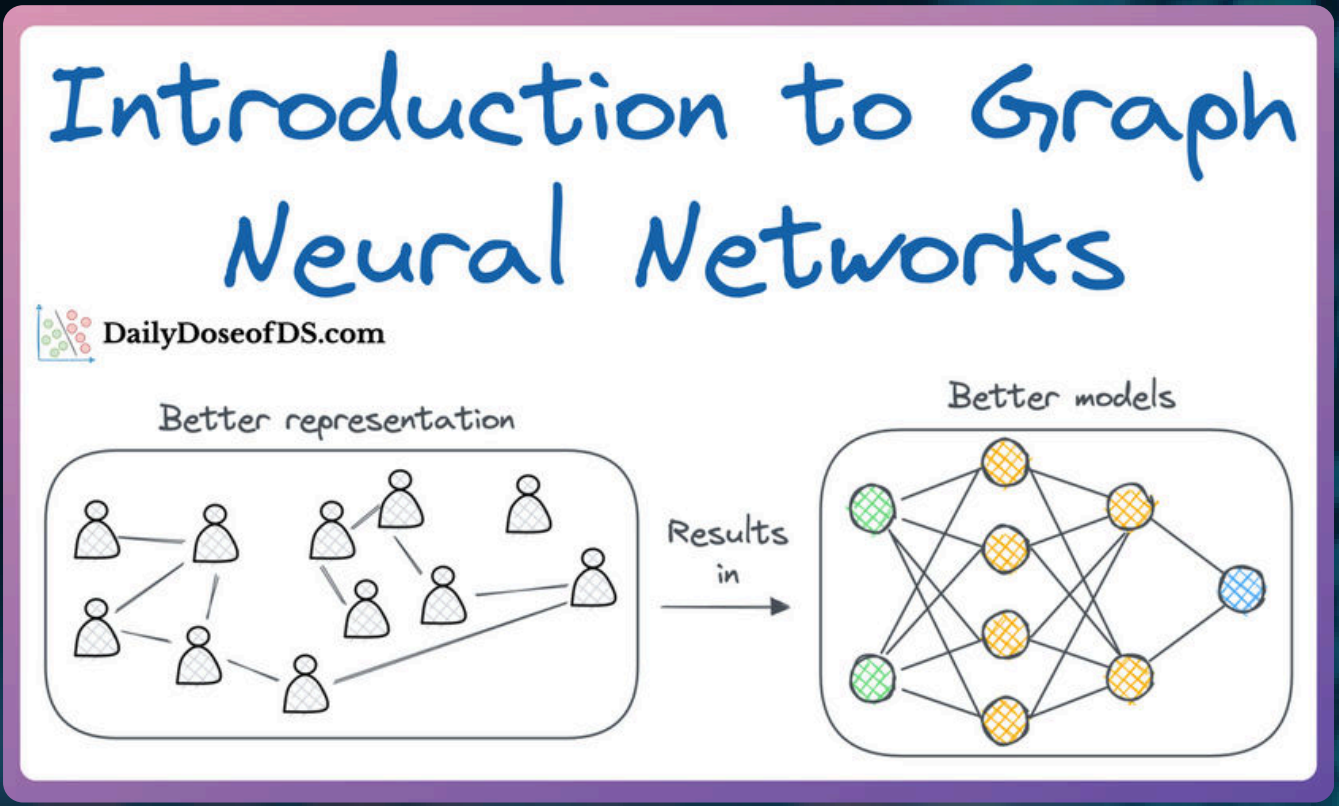
Recommendation Systems and Bioinformatics: Analyze complex, relational data like social networks and biological pathways.

Challenges with Complex Data Structures



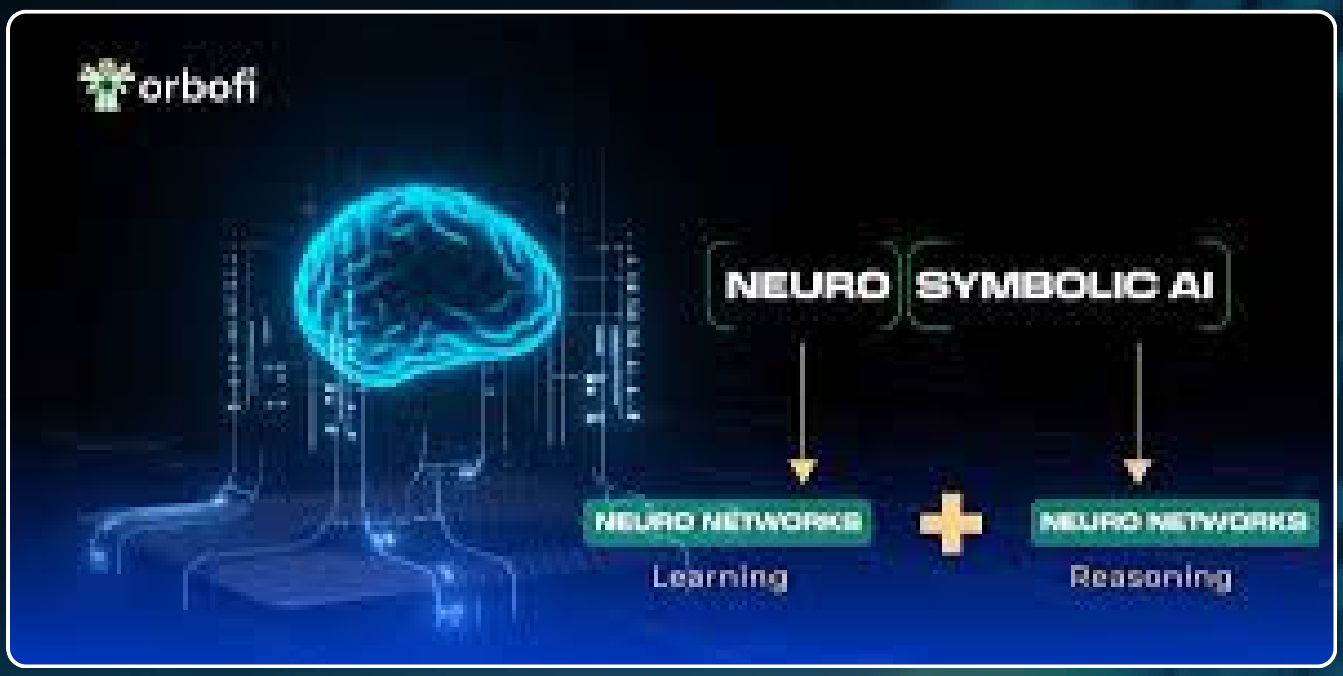
- Graphs are inherently irregular, nodes have varying degrees of connectivity, no fixed size or shape.
- Traditional machine learning models assume independent and identically distributed data, which is false for graphs where nodes are interdependent.
- Existing algorithms struggle with scalability and effectiveness on large, complex, and dynamic graphs

Future Directions



- Developing more scalable and expressive models for diverse graph types.
- Creating models to learn from multi-relational, heterogeneous, or dynamic graphs.
- Applying GNNs to real-world problems like drug discovery, transportation, and social network analysis with increasingly complex structures.

4. Hybrid Neuro-Symbolic AI



A fusion of connectionist AI (neural networks) and symbolic AI (logic-based reasoning). Neural networks excel at learning from raw, unstructured data such as images and natural language. Symbolic AI provides strong reasoning, explainability, and rule-based decision-making using explicit symbols, logic, and rules.

Combining these approaches leverages the strengths of both to build more robust, interpretable, and cognitively capable AI systems.

Neural components extract patterns and features from data. Symbolic components use rules, ontologies, or knowledge graphs to perform logical reasoning on learned representations.

Advantages

- **Explainability:** Provides clear, traceable reasoning paths alongside pattern recognition.
- **Robustness:** Better handles noisy or ambiguous input by combining statistical learning with rule-based checks.
- **Data Efficiency:** Symbolic knowledge reduces reliance on large data by encoding human expertise.
- **Generalization:** Combines neural adaptability with symbolic abstraction, improving transfer to new tasks.
- **Modularity:** Facilitates extending or updating knowledge bases and neural models independently.

Applications

- Visual reasoning and scene understanding.
- Complex language comprehension and dialogue systems.
- Robotics and autonomous systems requiring planning and perception.
- Scientific discovery and knowledge extraction from vast datasets.

Exercise

A. Essay

1. Define the Black Box Problem in connectionist AI and explain why it matters in high-stakes applications. (2-3 sentences)

2. What is catastrophic forgetting, and what general strategy is commonly proposed to mitigate it in neural networks? (2 sentences)

3. Briefly describe meta-learning and how it differs from standard supervised learning. (2-3 sentences)

B. Multiple Choice Questions

1. Which approach directly aims to improve interpretability of neural networks by outlining which features or regions influenced a decision?
a) Pruning
b) Attention visualization
c) Data augmentation
d) Hardware acceleration

2. Graph Neural Networks (GNNs) are particularly well-suited for:
a) Fixed-grid image data
b) Sequential tabular data

Exercise

c) Graph-structured data with variable connectivity
d) Pure symbolic reasoning

3. Hybrid Neuro-Symbolic AI primarily combines:
a) Large-scale data and label scarcity
b) Neural pattern recognition with symbolic reasoning
c) Hardware-centric optimization with software
d) Unsupervised learning with reinforcement learning

C. True or False

1. Meta-learning aims to learn a single task extremely well and then generalize to related tasks without any adaptation.

2. Energy efficiency in large neural networks is only a concern for researchers and has no impact on real-world deployment.

3. Explainable AI (XAI) techniques can fully reveal all internal decision pathways of deep networks without any trade-offs.

Answers

Essay

1. The Black Box Problem: neural networks' internal reasoning is not human-interpretable; this undermines trust and accountability in critical domains.
2. Catastrophic forgetting: neural networks forget previously learned tasks when trained on new ones; mitigations include replay/rehearsal, meta-learning, regularization, or architectural methods.
3. Meta-learning: learning to learn; models gain ability to adapt quickly to new tasks with limited data, unlike standard supervised learning which optimizes for a single task.

Multiple Choice Questions

1. B
2. C
3. B

True or False

1. False
2. False
3. False

Case Study

Automated Detection of Diabetic Retinopathy

Diabetic Retinopathy (DR) is a leading cause of blindness. Screening millions of diabetic patients with human specialists is slow and expensive. We need an AI to analyze retinal eye scans and classify disease severity automatically.

Solution: Convolutional Neural Network (CNN)

A CNN is used because it processes images efficiently by learning hierarchical features:

- Early Layers: Detect simple patterns (edges, colors).
- Middle Layers: Detect complex shapes (blood vessels, textures).
- Final Layers: Identify signs of disease (hemorrhages, lesions).

Case Study

How it works

1, Input: Thousands of labeled retinal images.
2. Processing: The CNN analyzes the image through multiple layers.
3. Output: A probability score for each disease class (e.g., No DR, Mild, Moderate, Severe).
4. Learning: The model's incorrect predictions are used to adjust its internal connections (backpropagation), improving its accuracy over time.

Result

Google AI developed a CNN that achieved ~90% accuracy, matching the performance of certified ophthalmologists. This proves a connectionist system can perform a complex visual diagnostic task.

Critical Thinking Questions

1. Why is a CNN better for this task than a simple Multi-Layer Perceptron (MLP)?
2. What is the "black box" problem, and why is it a concern in medical diagnosis?

Overall Test

Multiple Choice Questions

1. In a Convolutional Neural Network (CNN), what is the main purpose of the convolution layer?

A. To flatten the data into a one-dimensional vector

B. To extract spatial features and patterns from the input

C. To store past information

D. To reduce the spatial size of feature maps

2. What problem does Backpropagation Through Time (BPTT) aim to solve?

A. Training RNNs by computing gradients over multiple time steps

B. Reducing data volumes in CNNs

C. Increasing the number of convolution kernels

D. Differentiate whether the image is a dog or a cat

3. What is the primary purpose of an activation function in an artificial neuron?

A. To calculate the weighted sum of the inputs.

B. To adjust the learning rate of the network during training.

C. To introduce non-linearity and map the output to a specific range.

D. To initialize the weights and biases before training begins.

5. How does Connectionist AI differ from Symbolic AI in terms of learning?
A. Connectionist AI requires human experts to manually encode rules, while Symbolic AI learns automatically from data.
B. Connectionist AI learns automatically from examples, while Symbolic AI requires manual rule encoding.
C. Symbolic AI excels in pattern recognition, while Connectionist AI handles logical reasoning.
D. Both Connectionist and Symbolic AI store information in human-readable symbols.

6. What is one of the main challenges faced by Connectionist AI models?
A. They can only process structured data.
B. They require a lot of predefined rules to function.
C. They are prone to overfitting and generalization issues.
D. They cannot handle high-dimensional data.

7. Which of the following best describes the black box problem in connectionist AI?
A. AI models are entirely interpretable and transparent.
B. The internal workings of neural networks are complex and not easily understood by humans.
C. AI models cannot make predictions from data.
D. Simulation models are less interpretable than AI.

8. What is one future direction of connectionist AI research?
A. Reducing the interpretability of models
B. Limiting AI to single-modal tasks
C. Developing neuro-symbolic systems for reasoning and transparency
D. Eliminating on-device computation

9. Which of the following is a real-world example of Connectionist AI in medical imaging analysis?
A. Siri using speech recognition
B. Google Health detecting breast cancer from mammograms
C. Amazon recommending health books
D. Tesla's Autopilot system

10. What is overfitting?
A. When the model performs well on unseen data
B. When the network forgets patterns
C. When the model memorizes training data and fails on new data
D. When data is not trained enough

True or False

1. A convolutional neural network (CNN) is primarily used for analyzing sequential data like text.
2. Without an activation function, a neural network—no matter how many layers—can only model linear relationships.

3. The bias in an artificial neuron acts as a weight for a special, constant input of 1.
4. Connectionist AI learns by adjusting the weights of connections based on the error between predicted and actual outcomes.
5. Deep neural networks in connectionist AI are easily interpretable and transparent.
6. Connectionist models excel at pattern recognition but often struggle with symbolic reasoning and long-term planning.
7. Neural networks can predict the likelihood of developing chronic diseases by analyzing lifestyle and medical data.
8. AI is currently unable to assist in personalizing cancer treatments based on genetic profiles.
9. Generalization refers to how well a model performs on new data.
10. Overfitting improves model accuracy on unseen data.

Short Answers

1. Explain how the architectural differences between Recurrent Neural Networks (RNNs) and Feedforward Neural Networks (FNNs) enable RNNs to perform better on sequential data such as speech or text.

2.What is one advantage of connectionist AI over symbolic AI

3.What is one vulnerability of connectionist AI models?

4. Name one advantage of using AI in drug discovery.

5. What does "loss function" measure?

Word Banks

1, A _____ neural network is designed to process data with a grid-like matrix, such as images, using filters/kernels. In this type of neural network, a layer called _____ is used to reduce the data volume and lower the computation workload.

Word Bank: {convolutional, recurrent, feedforward, pooling, activation}

2. 1._____ AI stores information in human-readable symbols and logical statements, making it easy to audit and modify. Connectionist systems, however, distribute knowledge across networks of artificial neurons, making them more 2._____ but harder to interpret. While Symbolic AI excels in tasks that require explicit knowledge and logical reasoning, Connectionist AI shines in 3._____, adaptive learning, and handling large amounts of data.

Word bank: {pattern recognition, symbolic, adaptable}

3. Connectionist AI models require large amounts of _____ to train effectively, which can make them resource-intensive. One of their strengths is the ability to adapt and improve through _____, allowing them to handle complex tasks like _____ and _____. However, these models can be vulnerable to _____ attacks, where small changes in input data can lead to incorrect outputs.

Word Bank: recognition adversarial, data, image recognition speech, learning

4. An artificial neuron, the fundamental unit of a neural network, processes information in a structured way. It begins by receiving multiple input signals. Each input is multiplied by its corresponding strength value, known as (1)_____. The neuron then sums these weighted inputs along with a tunable offset called the (2)_____ in a step known as (3)_____. This resulting sum is then passed to a non-linear (4)_____, such as ReLU or Sigmoid, which decides the final (5)_____ of the neuron, determining what signal is passed on to the next layer in the network.

Word Bank: Weights, Activation Function, Output, Bias, Aggregation

5. In the healthcare sector, neural networks help predict _____ risk and detect _____ from scans. In finance, they are used to assess _____ scoring, prevent _____ activity, and support _____ management.

Word bank: {disease, credit, risk, fraudulent, tumors}

6. _____ are the connections between nodes that determine signal strength. The process of adjusting weights to reduce errors is called _____. The ability of a network to perform well on unseen data is called _____. When a model learns too much detail from training data, it causes _____. _____ adds a penalty to prevent the model from becoming too complex. (Kevin)

Word Bank: (Input Layer, Hidden Layer, Output Layer, Backpropagation, Generalization, Overfitting, Regularization, Loss Function, Neurons, Weights)

Multiple Answers

1. Which of the following are examples of connectionist AI architectures?

A. Convolutional Neural Network (CNN)
B. Circulating neural network (CNN)
C. Recurrent Neural Network (RNN)
D. Feedforward Neural Network (FNN)

2. Which of the following are key features of Connectionist AI?

A. It learns from data.

B. It requires manual rule coding.

C. It uses activation functions.

D. It represents knowledge through symbols.

3. Which of the following statements are key purposes of an activation function in an artificial neural network?

A. It determines the final prediction at the output layer of the network.

B. It introduces non-linearity, allowing the network to learn complex patterns.

C. It initializes the optimal values for all weights before training.

D. It maps the neuron's output to a specific, desirable range (e.g., 0 to 1).

4. Which of the following are valid use cases of Connectionist AI in either Healthcare or Finance? (Select all that apply)

A. Tumor detection from MRI scans

B. Credit risk prediction

C. Ad targeting in social media

D. Personalized cancer therapy

E. Forecasting shopping trends

5. Which are examples of neural network applications?
A. Speech recognition
B. Fraud detection
C. Face recognition
D. Manual data entry

Answers

Multiple Choice Questions

1. B
2. A
3. C
4. C
5. B
6. C
7. B
8. C
9. B
10. C

True or False

1. False
2. True
3. True
4. True
5. False
6. True
7. True
8. False
9. True
10. False

Short Answer

1. RNNs have feedback loops that allow information from previous time steps to be used in the current time step in a kind of memory called hidden state. By keeping these memories, it enables them to predict a more accurate outcome based on past information. On the other hand, the FNN process inputs independently without managing the order or context of previous input, making it less effective on sequential data.

2. Connectionist AI is more adaptable and better at pattern recognition.
3. They are susceptible to adversarial and data poisoning attacks, where small changes in input or training data can lead to incorrect outputs
4. It speeds up the identification of effective drug compounds and reduces R&D costs.
5. The difference between predicted and actual outputs.

Word Bank

1.Convolutional, pooling
2. Symbolic, adaptable, pattern recognition
3. Data, learning, recognition, speech recognition, adversarial
4. Weights, bias, aggregation, activation function, output
5. Disease, tumors, credit, fraudulent, risk
6. Backpropagation, generalization, overfitting, regularization

Multiple Answers

1. A, C, D
2. A, C
3. B, D, A
4. A, B, D
5. A, B, C

Glossary

Big Data
Extremely large and complex datasets that cannot be handled by traditional data systems. It requires special tools and methods for storage, processing, and analysis.

5 Vs of Big Data
Key characteristics that define Big Data: Volume, Velocity, Variety, Veracity, and Value.

Hadoop Distributed File System (HDFS)
A storage system that splits big files into blocks and stores them across multiple computers, ensuring reliability through replication.

NoSQL
A flexible database designed to manage unstructured or semi-structured data efficiently, unlike traditional relational databases.

Cloud Storage
An online storage service that allows users to save and access data through the internet instead of local devices.

Data Warehouse
A structured data repository used for analytical processing (OLAP). Data here is extracted, transformed, and loaded (ETL) for business insights.

Glossary

Data Lake
A repository that stores raw, unprocessed data of all types (structured or unstructured) for future analysis or machine learning.

ETL (Extract, Transform, Load)
A process for moving data into a warehouse by extracting it from sources, transforming it for consistency, and loading it for analysis.

ELT (Extract, Load, Transform)
A process where raw data is loaded first (into a data lake), then transformed later for analysis.

Batch Processing
Processing data in large, pre-collected groups at scheduled times — used for large historical datasets.

Stream Processing
Continuous, real-time data processing as new information arrives — used for instant decisions like fraud detection.

Apache Kafka
A framework for handling real-time data streams between systems; often used for event-driven applications.

Glossary

Apache Spark
A fast data-processing engine that supports both batch and real-time analytics using in-memory computation.

Apache Flink
A real-time processing framework that handles continuous data flows with low latency.

Apache Superset
An open-source platform for visualizing big data and creating interactive dashboards.

Big Data Analytics
The process of examining massive datasets to uncover patterns, correlations, and insights for better decisions.

Four Types of Analytics

1. Descriptive - What happened?
2. Diagnostic - Why did it happen?
3. Predictive - What might happen next?
4. Prescriptive - What should we do about it?

Quantitative Analysis
Uses numerical data and statistics to measure and compare outcomes (e.g., mean, variance).

Glossary

Qualitative Analysis
Uses non-numerical data (like opinions or interviews) to understand reasons, emotions, or motivations.

Correlation
A statistical relationship between two variables; can be positive, negative, or none.

Regression
A method for predicting outcomes based on relationships between dependent and independent variables.

Machine Learning (ML)
A branch of AI that allows computers to learn from data and improve their performance without being explicitly programmed.

Supervised Learning
The model learns from labeled data — where inputs and correct outputs are known.

Unsupervised Learning
The model finds patterns or groupings in unlabeled data without predefined answers.

Glossary

Support Vector Machine (SVM)
A machine learning algorithm that separates data into classes using the best dividing line (hyperplane).

Artificial Neural Network (ANN)
A computing system inspired by the human brain, composed of layers of connected "neurons" that process data and learn patterns.

Activation Function
A mathematical rule that determines a neuron's output, helping the network model complex patterns.

Overfitting
When a model learns too closely from training data, performing well on that data but poorly on new inputs.

Regularization
A technique used to reduce overfitting by simplifying the model and improving generalization.

Real-Time Analytics
Analyzing data instantly as it is created, allowing for immediate decisions and responses.

Glossary

Business Intelligence (BI)
The practice of analyzing data to support decision-making and improve business performance.

OLTP (Online Transaction Processing)
A system that manages real-time transaction data, such as ATM withdrawals or online purchases.

OLAP (Online Analytical Processing)
A system used for analyzing historical data in large volumes to support complex decision-making.

RTAP (Real-Time Analytics Platform)
Combines OLAP analysis with real-time data updates for instant insights.