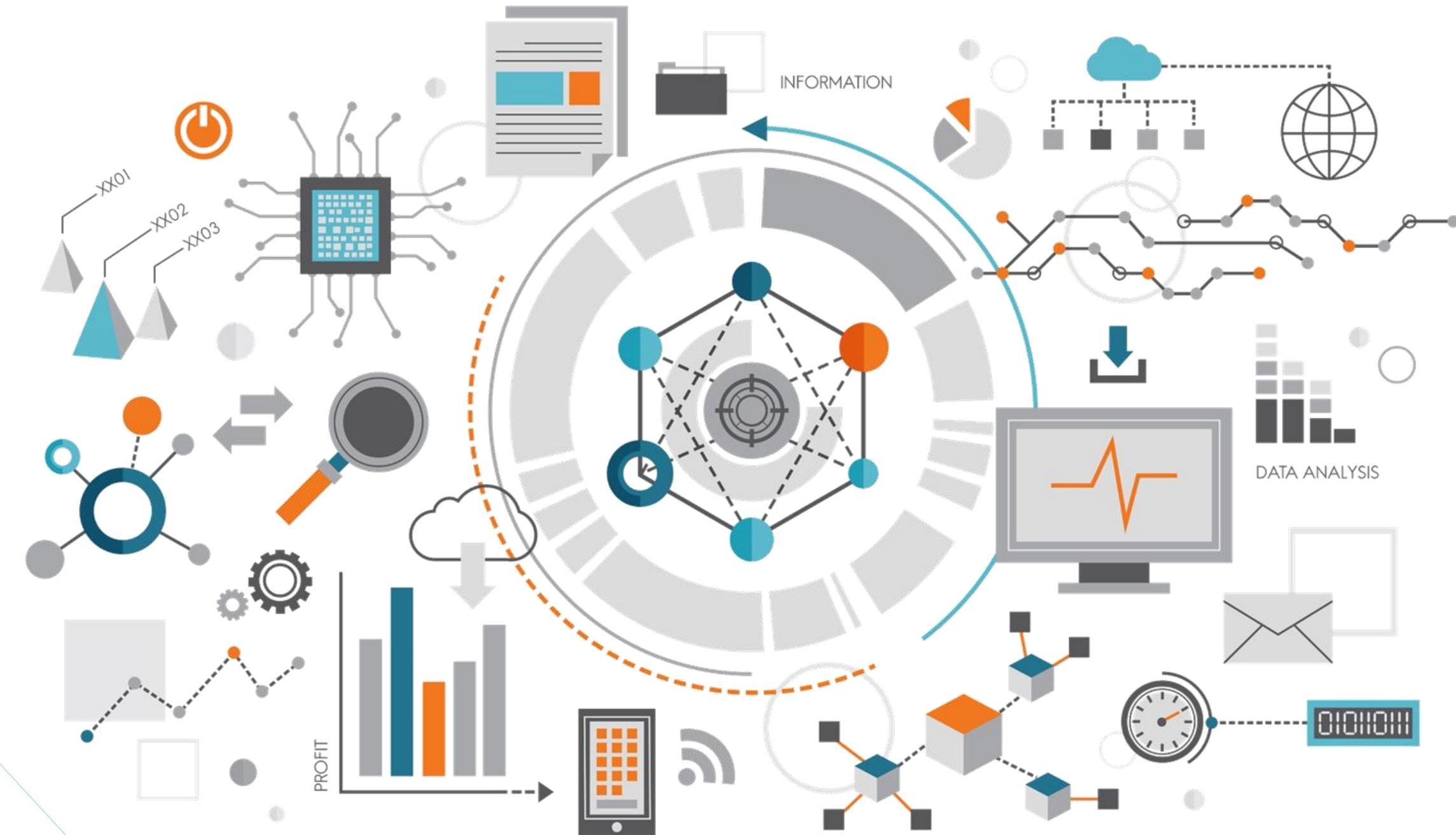


IBM What is Data Science?



Module II

Data science

Topics



Module II: Data science

Topics

How Big Data is Driving Digital Transformation?



Understanding Digital Transformation

- **Digital Transformation:** Overhauling business operations to leverage new technologies effectively.
- **Integration of Digital Technology:** Fundamental changes in operations and value delivery across all areas.
- **Driven by Data Science and Big Data:** Utilizing vast data resources for competitive advantage and innovation.
- **Industry Examples:** Netflix, Houston Rockets, and Lufthansa embracing digital transformation for success.
- **Core Organizational Change:** Digital transformation affects businesses fundamentally and culturally.
- **Example Case:** Houston Rockets' use of Big Data to revolutionize basketball strategy.

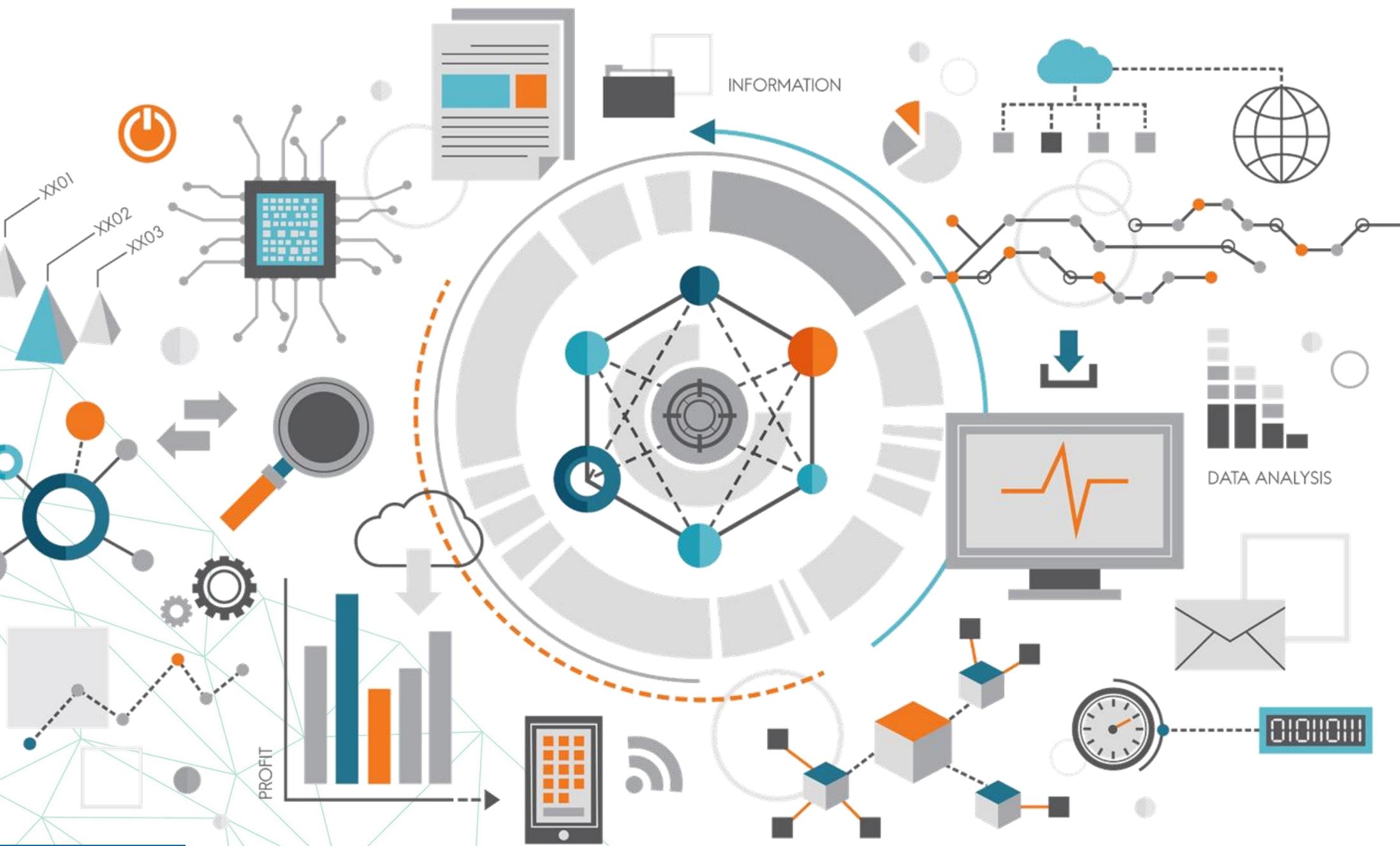
Key Aspects of Digital Transformation

- **Process Improvement:** In-depth analysis leads to enhancements in operations and workflows.
- **Organizational Culture:** Requires fundamental changes in approaches to data, employees, and customers.
- **Leadership Involvement:** Decision-makers at top levels crucial for successful implementation.
- **Executive Support:** CEO, CIO, and emerging role of Chief Data Officer pivotal in guiding transformation.
- **Whole Organization Approach:** Success relies on support from all levels and departments.
- **New Mindset:** Navigating challenges of digital transformation necessitates adopting a forward-thinking perspective.



Module II: Data science Topics

Introduction to Cloud



Understanding Cloud Computing

- **Cloud Computing:** Delivery of on-demand computing resources over the Internet.
- **Examples:** Online web apps, secure business applications, cloud-based storage platforms.
- **User Benefits:** Cost-effectiveness, access to latest application versions, collaborative work.
- **Essential Characteristics:** On-demand self-service, broad network access, resource pooling.
- **Characteristics Continued:** Rapid elasticity, measured service for transparent payment based on usage.
- **Transformation Impact:** Cloud computing changes how organizations consume compute services.

Cloud Deployment and Service Models

- **Deployment Models:** Public, private, and hybrid clouds based on infrastructure ownership.
- **Public Cloud:** Leverages services over the open internet, shared by multiple companies.
- **Private Cloud:** Infrastructure provisioned exclusively for a single organization.
- **Hybrid Cloud:** Seamless integration of public and private clouds.
- **Service Models:** Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS).
- **IaaS:** Access to physical computing resources without managing them.
- **PaaS:** Access to hardware and software tools for application development and deployment.
- **SaaS:** Centralized hosting and licensing of software on a subscription basis.

Module II: Data science Topics

Cloud for Data Science



Benefits of Cloud Computing for Data Scientists

- **Centralized Storage:**
 - Eliminate physical storage limits by securely storing vast amounts of data in the Cloud.
- **Access to High-Performance Machines:**
 - Deploy complex data analytics on advanced computing machines without needing expensive hardware.
- **Scalable Algorithm Deployment:**
 - Run advanced algorithms at scale using cloud-based platforms designed for data-intensive tasks.
- **Collaborative Work:**
 - Enable global teams to work on the same data simultaneously, enhancing productivity and collaboration.
- **Instant Access to Technologies:**
 - Quickly leverage open-source technologies like Apache Spark, TensorFlow, and more, without setup delays.
- **Up-to-Date Tools and Libraries:**
 - Benefit from the latest tools, libraries, and frameworks automatically updated by cloud providers, reducing maintenance overhead.

Cloud Accessibility and Collaboration

- **Anytime, Anywhere Access:** Cloud technologies accessible from various devices globally.
- **Enhanced Collaboration:** Simultaneous data access enables easier collaboration among teams.
- **Pre-Built Environments:** Big tech companies offer Cloud platforms like IBM Cloud, AWS, Google Cloud.
- **IBM Skills Network Labs:** Access to tools like Jupyter Notebooks and Spark clusters for data science projects.
- **Productivity Enhancement:** Cloud dramatically enhances productivity for data scientists with practice.
- **Global Availability:** Cloud services available across different time zones, fostering collaboration and innovation.

Module II: Data science Topics

Foundations of Big Data



Understanding Big Data

- **Definition by Ernst and Young:** Dynamic, large, disparate data created by people, tools, machines.
- **Common Elements:** Velocity, volume, variety, veracity, value - the V's of Big Data.
- **Velocity:** Speed of data accumulation, processed in near or real-time.
- **Volume:** Scale of data, driven by increased sources and scalable infrastructure.
- **Variety:** Diversity of data, structured and unstructured, from various sources.
- **Veracity:** Quality, origin, accuracy of data, vital for trustworthiness and insights.

Harnessing Big Data

- **Value:** Ability to derive value from data beyond profit, including social and personal benefits.
- **Examples of V's in Action:** YouTube uploads, world population's digital interactions, data types.
- **Coping with Big Data Challenges:** Traditional tools insufficient, alternative tools like Apache Spark and Hadoop.
- **Distributed Computing Power:** Tools enable extraction, loading, analysis, and processing of massive data sets.
- **Enriching Services:** Insights from Big Data allow organizations to better connect with customers.
- **Personal Data Journey:** From devices to Big Data analysis, data impacts services and returns to users.

Module II: Data science Topics

Data Science and Big Data



Programming Background and Data Science Trends

- Diverse programming backgrounds: Some have basic skills, others are proficient.
- Importance of computational thinking: Essential for data science regardless of programming expertise.
- Rise of data science and analytics: Fueled by new tools, approaches, and abundant data.
- Growth in demand: Employers increasingly recognize the need for data science skills.
- Expansion in industries: From initial adoption in specific fields to widespread integration.
- Increasing enrollment: Significant rise in students pursuing data-related courses.

Understanding Big Data and its Evolution

- Definition varies: From handling large volumes to exceeding traditional database capabilities.
- Origins in Google's challenges: Originated from efforts to manage vast web page data.
- Evolution beyond storage: Includes new analytical and statistical techniques for massive datasets.
- Hadoop as a key player: Adapted Google's approach, leading to widespread use in big data clusters.
- Continued advancement: Constant evolution towards handling and analyzing larger datasets.
- Future prospects: Potential exploration of deep learning and further innovations in data science.

Module II: Data science Topics

What is Hadoop?



Introduction to Big Data Clusters

- Traditional data processing involved bringing data to the computer and running programs on it.
- Big data clusters, pioneered by Larry Page and Sergey Brin, distribute and replicate data across thousands of computers for parallel processing.
- Map and reduce processes enable handling large datasets and scaling linearly with server additions.
- Hadoop, a popular big data architecture, emerged as Yahoo adopted Google's approach in 2008.

Module II: Data science Topics

Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark



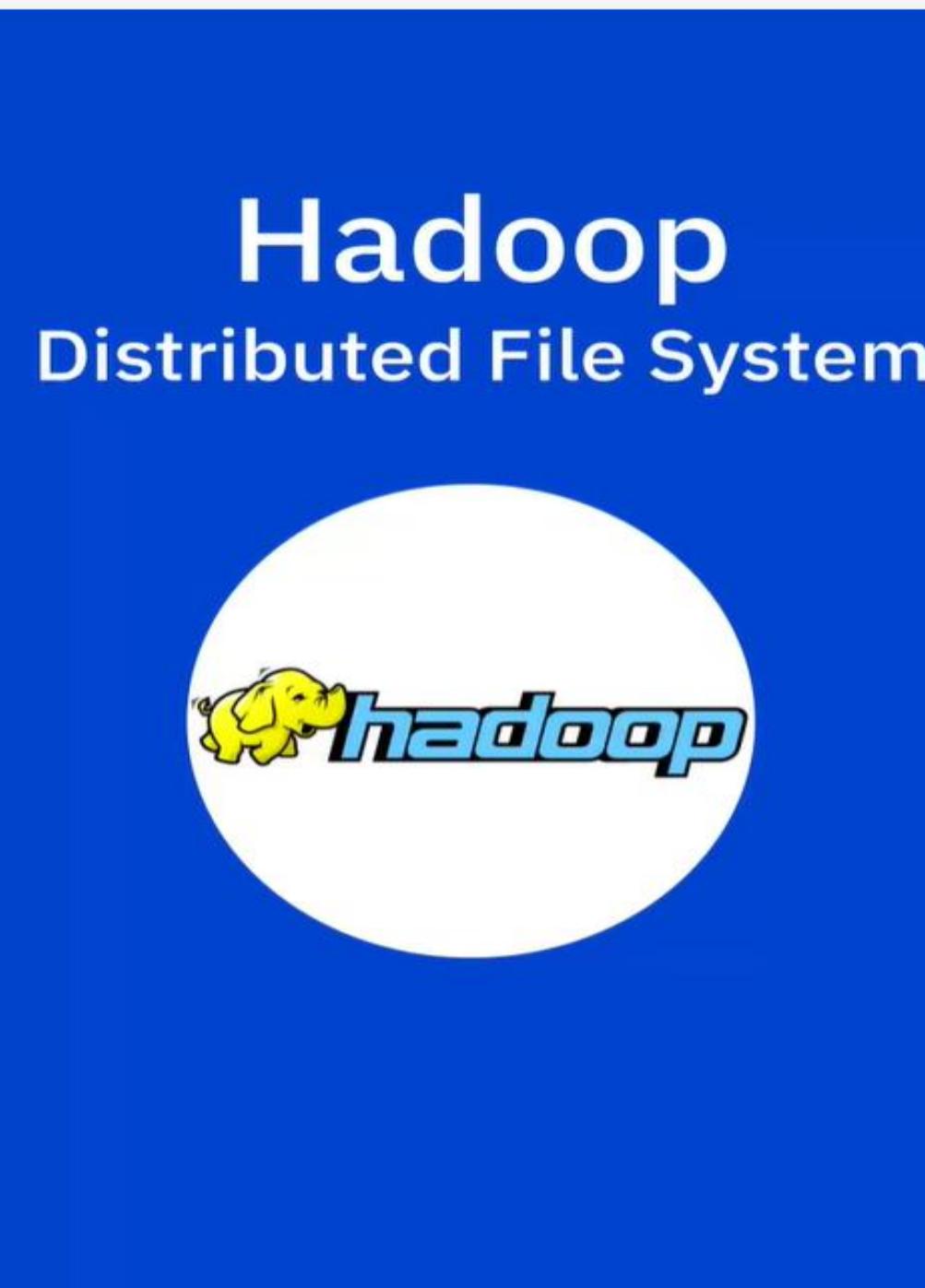
Introduction to Big Data Processing Technologies

- **Big Data Processing**
- Big Data processing enables handling large sets of structured, semi-structured, and unstructured data.
- **Overview of Apache Hadoop, Apache Hive, and Apache Spark in Big Data Analytics:**
- **Hadoop:**
 - Offers distributed storage and processing capabilities for large datasets.
- **Hive:**
 - Serves as a data warehouse on top of Hadoop for querying and analyzing data.
- **Spark:**
 - A distributed analytics framework designed for complex real-time data analytics.

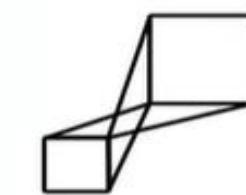
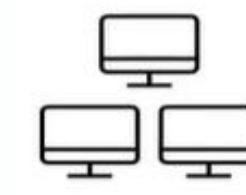
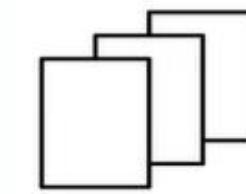
These technologies provide scalable, reliable, and cost-effective solutions for big data storage and processing.

Understanding Hadoop and HDFS

- Hadoop facilitates distributed storage and processing across clusters of computers.
- HDFS partitions files over multiple nodes, allowing parallel access and computations.
- Replication of file blocks ensures fault tolerance and availability in case of node failures.
- Data locality minimizes network congestion and increases throughput by moving computations closer to data nodes.
- Benefits of using HDFS include fast recovery, access to streaming data, scalability, and portability.
- Hadoop's ability to consolidate and optimize data storage across the organization enhances enterprise data warehouse efficiency.



Hadoop Distributed File System, or HDFS, is a storage system for big data that runs on multiple commodity hardware connected through a network.

-  Provides scalable and reliable big data storage by partitioning files over multiple nodes
-  Splits large files across multiple computers, allowing parallel access to them
-  Replicates file blocks on different nodes to prevent data loss



Exploring Hive and Spark

- Hive provides data warehousing capabilities for large datasets stored in HDFS or other systems.
- Designed for long sequential scans, Hive is suited for ETL(Extract, transform, and load), reporting, and data analysis tasks.
- Hive's read-based approach makes it less suitable for transaction processing requiring fast response times.
- Spark is a versatile data processing engine for various applications, including interactive analytics and machine learning.
- Utilizing in-memory processing, Spark significantly accelerates computations, spilling to disk only when memory is constrained.
- Spark interfaces with major programming languages and can access data from various sources, including HDFS and Hive.



Apache Hive
a data warehouse for data query and analysis



Apache Spark
a distributed analytics framework for complex,
real-time data analytics

Introduction to Apache HBase

- **What is HBase?**
 - A distributed, scalable NoSQL database that runs on top of HDFS.
 - Designed to handle large amounts of sparse, unstructured data in real-time.
- **Key Features:**
 - **Column-Oriented Storage:** Ideal for querying specific columns efficiently.
 - **NoSQL:** Flexible data storage without relying on predefined schemas.
 - **Real-Time Data Access:** Supports fast read/write operations on large datasets.



HBase in the Hadoop Ecosystem

- **Scalability:**
 - Horizontally scales across thousands of servers, managing petabytes of data.
- **Integration with Hadoop:**
 - Uses HDFS for fault-tolerant storage.
 - Supports MapReduce for processing large-scale data.
- **Use Cases:**
 - Web analytics, log data analysis, and applications needing fast access to big data.
- **Why HBase?**
 - Real-time access and processing, complementing Hadoop's batch processing capabilities.



Introduction to Data Mining

- **What is Data Mining?**
 - The process of discovering patterns, trends, and useful insights from large datasets.
 - Involves techniques from statistics, machine learning, and database systems.
- **Purpose of Data Mining:**
 - To transform raw data into actionable information.
 - Helps in decision-making, predictive analytics, and identifying hidden patterns.



Key Techniques in Data Mining

Classification:

- Assigning data to predefined categories based on its attributes.
- Commonly used in spam detection and fraud detection.

Clustering:

- Grouping similar data points together based on shared characteristics.
- Used for market segmentation and customer profiling.

Association Rule Learning:

- Identifying relationships between variables in large datasets.
- Commonly used for market basket analysis (e.g., "customers who buy X also buy Y").



Applications of Data Mining

Business Intelligence:

- Helps companies understand customer behavior, market trends, and product performance.

Healthcare:

- Used for diagnosing diseases, predicting patient outcomes, and optimizing treatment plans.

E-commerce and Retail:

- Analyzes customer purchase patterns to optimize sales strategies and personalized recommendations.

Finance:

- Detects fraudulent transactions, assesses credit risk, and manages investment portfolios.





Module II: Data science Topics

Lesson Summary: Big Data and Data Mining



Fundamentals of Big Data and Cloud Computing

- Big data impacts various societal aspects, including business operations and sports.
- Understanding key attributes and challenges associated with big data is crucial.
- Big data drives digital transformation by necessitating fundamental changes in business approaches.
- The five characteristics of big data include value, volume, velocity, variety, and veracity.
- Cloud computing enables access to on-demand computational resources via the internet.
- Cloud computing features on-demand access, network accessibility, resource pooling, elasticity, and measured service.

Leveraging Cloud Technologies for Big Data Processing

- Cloud computing addresses scalability, collaboration, accessibility, and software maintenance challenges.
- Instant access to technologies and updated versions without installation is a benefit of cloud computing.
- Popular open-source tools for big data processing include Apache Hadoop, Hive, and Spark.
- Hadoop provides distributed storage and processing across computer clusters.
- Hive serves as a data warehouse for large datasets stored in Hadoop File System (HDFS) or Apache HBase.
- Spark is a versatile data processing engine suitable for various applications, leveraging cloud advantages for big data mining.

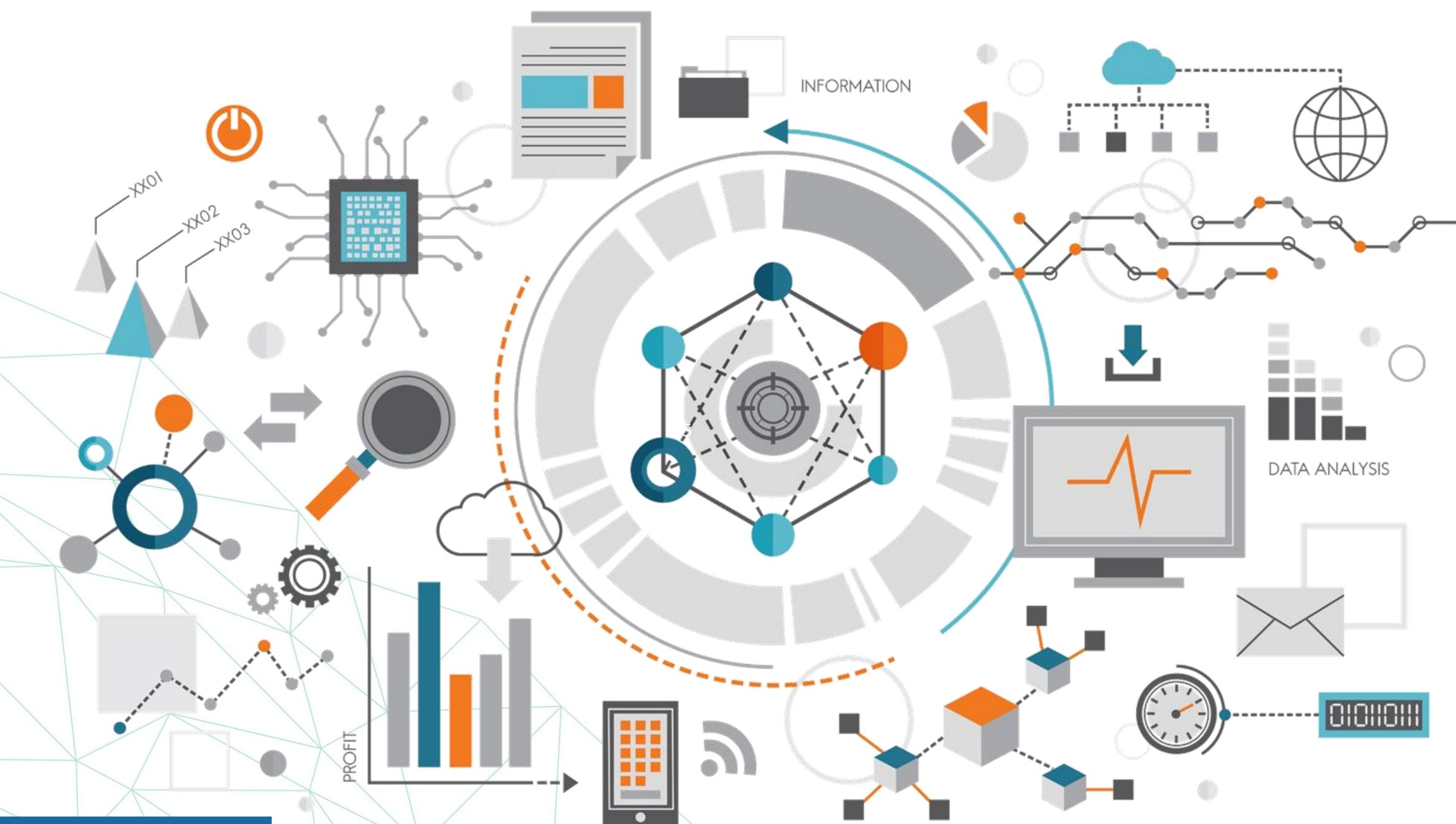
Data mining process

1. Goal set → Identify key questions
2. Select data → Identify data sources
3. Preprocess → Clean the data
4. Transform → Determine storage needs
5. Data mine → Determine methods and analyze
6. Evaluate → Assess outcomes, share results



Module II: Data science Topics

Artificial Intelligence and Data Science



Differentiating AI and Data Science

Definition:

- Artificial Intelligence (AI): The simulation of human intelligence processes by machines, especially computer systems. Involves learning, reasoning, and self-correction.
- Data Science: A multidisciplinary field that uses scientific methods, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

Key Focus:

- AI: Focuses on creating intelligent agents that can perform tasks without explicit instructions.
- Data Science: Focuses on analyzing and interpreting complex data to inform decision-making.

Differentiating AI and Data Science

Techniques and Tools:

- AI: Machine learning, neural networks, natural language processing, and computer vision.
- Data Science: Statistical analysis, data mining, data visualization, and predictive modeling.

Applications:

- AI: Chatbots, recommendation systems, autonomous vehicles, and facial recognition.
- Data Science: Market analysis, risk assessment, customer segmentation, and healthcare analytics..

Artificial Intelligence and Data Science

Data Science
is the process and method for extracting knowledge and insights from large volumes of disparate data.

It involves
mathematics
statistical analysis
data visualization
machine learning
and more...

It could use
machine learning algorithms

deep learning models

It's a broad term
encompasses the entire data processing methodology

AI includes everything that allows computers to learn how to solve problems and make intelligent decisions

Both AI and Data Science can involve the use of **Big Data**

Module II: Data science Topics

Generative AI and Data Science



Understanding Generative AI

- Generative AI creates new data rather than analyzing existing datasets.
- Models like GANs and VAEs are foundational to generative AI.
- Generative AI mimics human creations in images, music, language, and more.
- Applications span diverse industries, from content creation to healthcare and gaming.
- Examples include GPT-3 for text generation and medical image synthesis.
- Generative AI aids in fashion design, game development, and creating artworks.

Application in Data Science

- Generative AI augments data science by generating synthetic data for model training.
- Synthetic data closely mimics real data, aiding in analysis and model creation.
- It enables data scientists to overcome limitations in data availability and diversity.
- Coding automation with generative AI revolutionizes analytics, allowing focus on higher-level tasks.
- Generative AI enhances decision-making by autonomously exploring data and uncovering hidden patterns.
- Tools like IBM's Cognos Analytics leverage generative AI to provide comprehensive insights and reports.

Module II: Data science Topics

Neural Networks and Deep Learning



Evolution of Neural Networks

Early Beginnings (1940s - 1980s):

- 1943: Warren McCulloch and Walter Pitts introduce the concept of artificial neurons.
- 1958: Frank Rosenblatt develops the Perceptron, the first simple neural network.
- 1980s: Backpropagation algorithm re-emerges, allowing for training of multi-layer networks.

AI Winter (1990s):

- Neural networks lose popularity due to limited computational power and challenges in training deep networks.
- Research and funding decline, leading to reduced interest in neural networks.

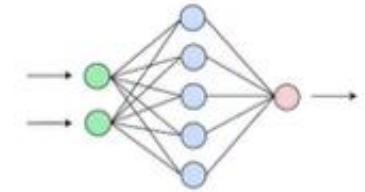
Evolution of Neural Networks

Resurgence (2000s - Present):

- 2006: Geoffrey Hinton revives interest in deep learning with the introduction of deep belief networks.
- 2012: Breakthrough in image recognition with AlexNet, demonstrating the power of convolutional neural networks (**CNNs**).
- 2010s: Rapid advancements in deep learning applications, including NLP, computer vision, and reinforcement learning.

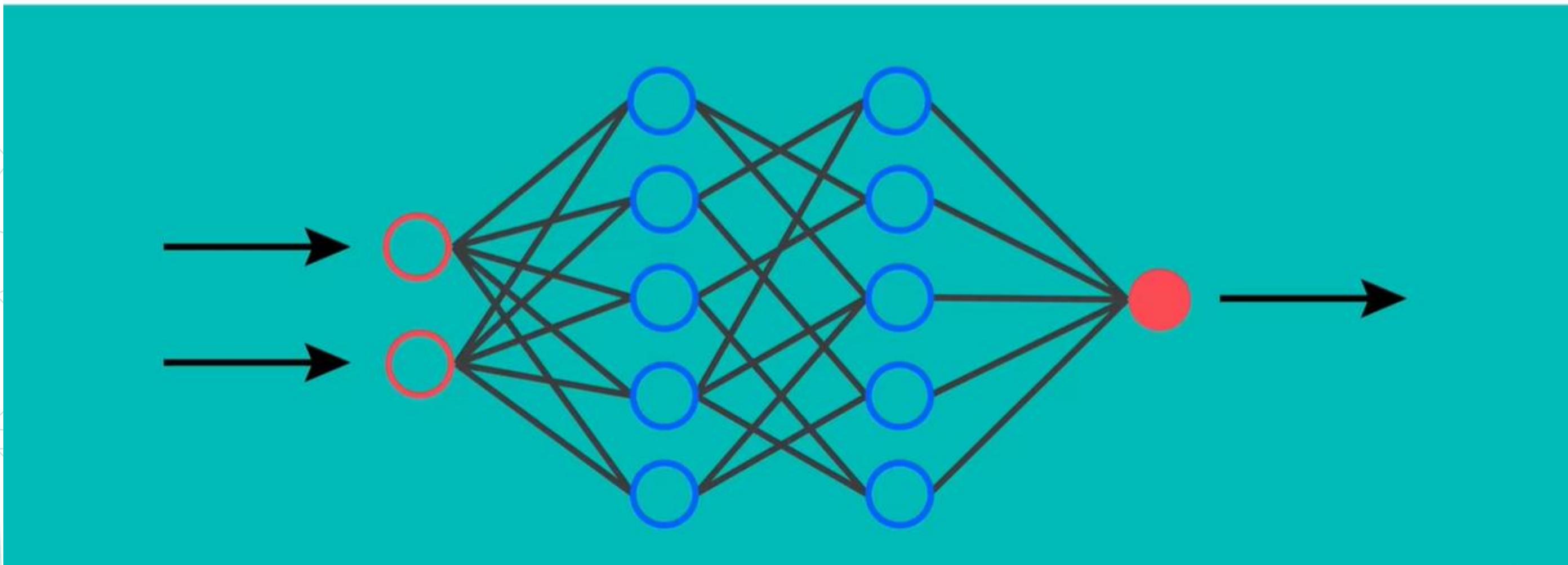
Current Trends:

- Continued innovation in neural network architectures (e.g., **Transformers**, **GANs**).
- Integration of neural networks in various fields such as healthcare, autonomous vehicles, and robotics.
- Ongoing research into explainable AI and ethical considerations.



Neural Networks

Take inspiration from biological neural networks, although they work quite a bit differently



Deep Learning and Computational Requirements

- Deep learning extends neural networks to tackle larger and more complex tasks.
- It involves training networks to learn patterns and make decisions autonomously.
- Deep learning requires understanding linear algebra for matrix operations.
- Packages exist to simplify deep learning, but understanding underlying concepts is valuable.
- High-powered computational resources, like GPUs, are essential for deep learning tasks.
- Deep learning applications include speech recognition, image classification, and natural language processing.



Module II: Data science Topics

Applications of NLP in Data Science



1. Chatbots

Chatbots are a form of artificial intelligence that are programmed to interact with humans in such a way that they sound like humans themselves.



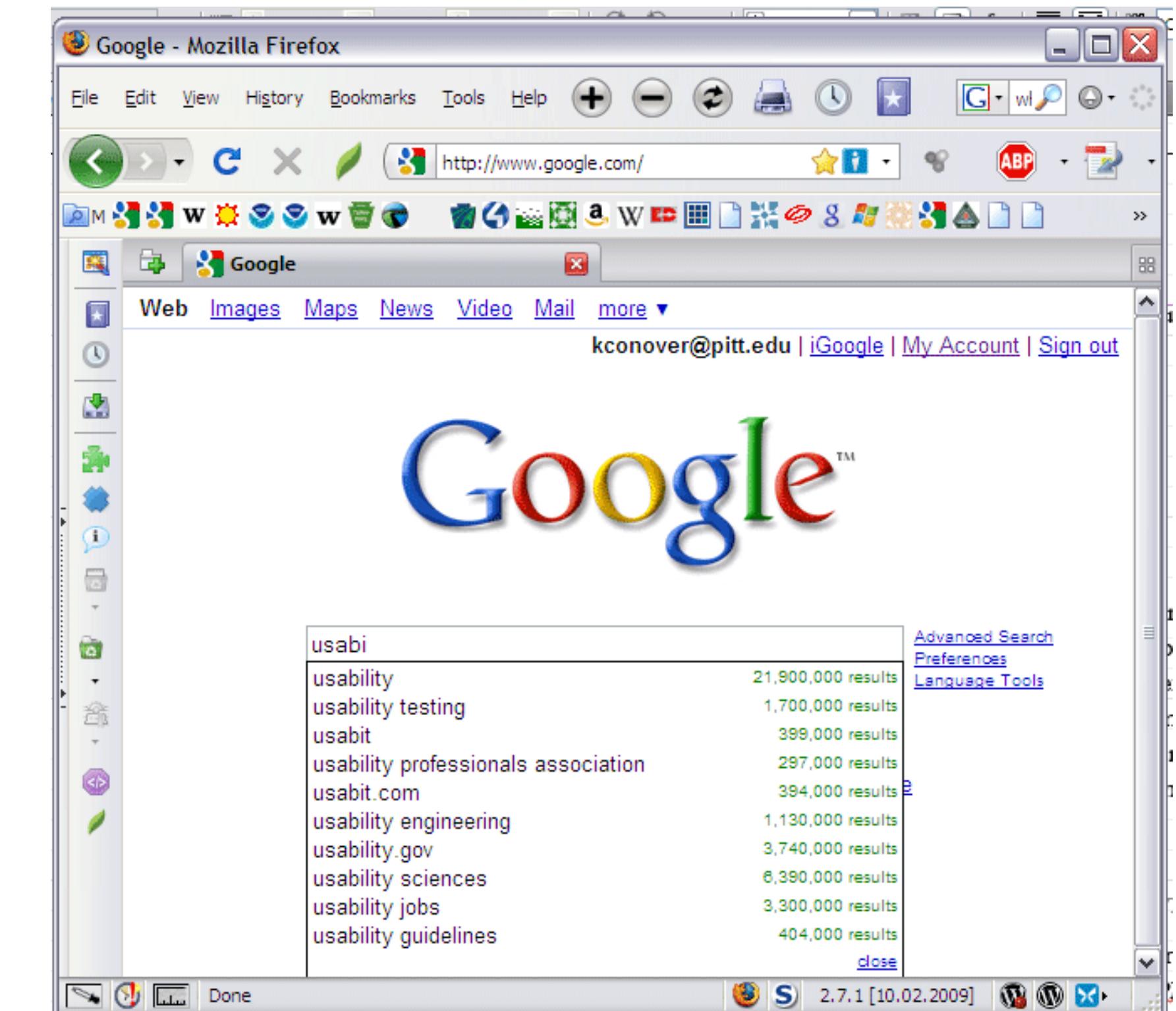


2. Autocomplete in Search Engines

Have you noticed that search engines tend to guess what you are typing and automatically complete your sentences?

For example, On typing “game” in Google, you may get further suggestions for “game of thrones”, “game of life” or if you are interested in math then “game theory”.

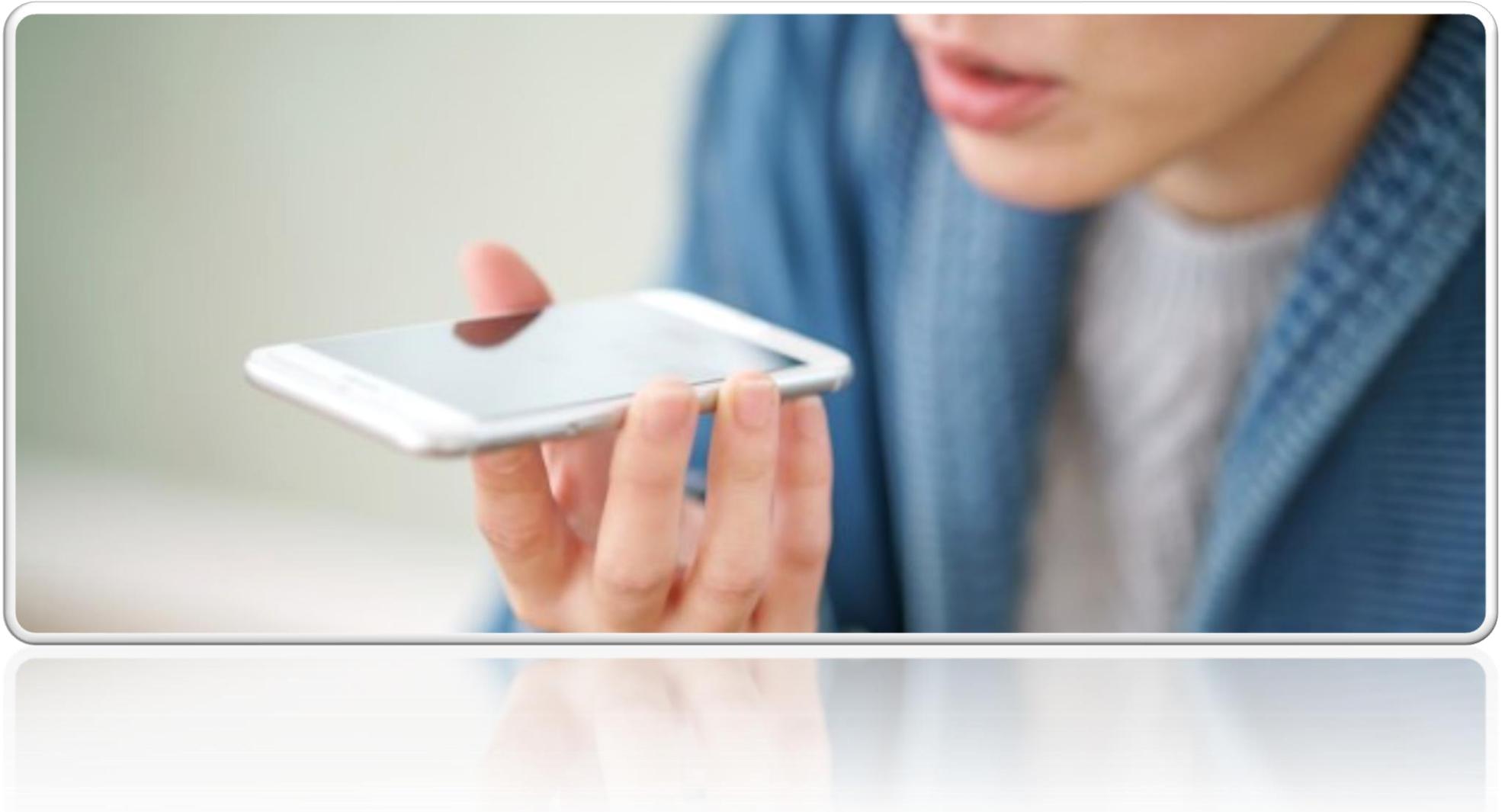
All these suggestions are provided using autocomplete that uses Natural Language Processing to guess what you want to ask.



3. Voice Assistants

These days voice assistants are all the rage! Whether its Siri, Alexa, or Google Assistant,

almost everyone uses one of these to make calls, place reminders, schedule meetings, set alarms, surf the internet, etc.

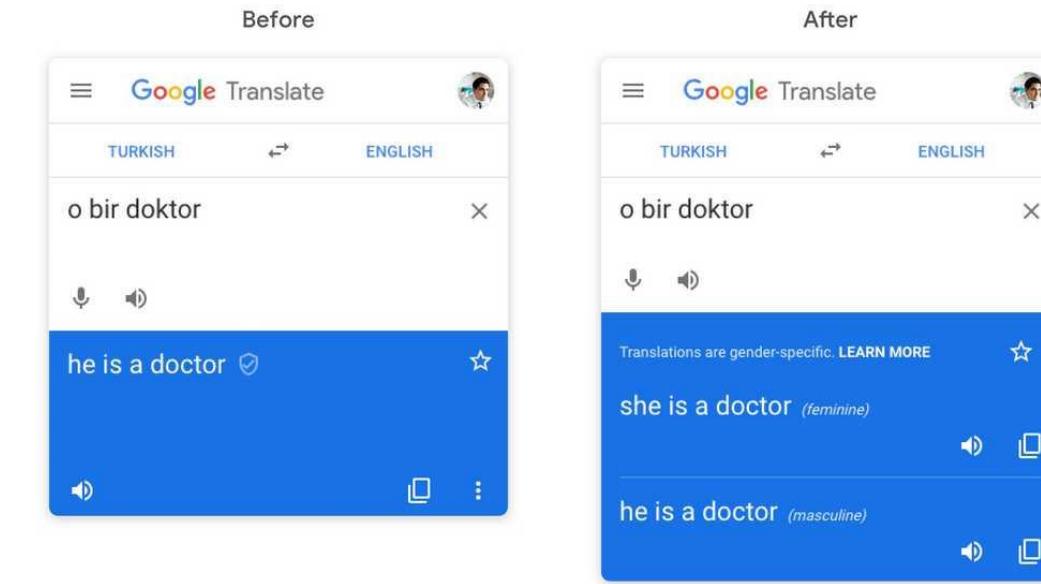


4. Language Translator

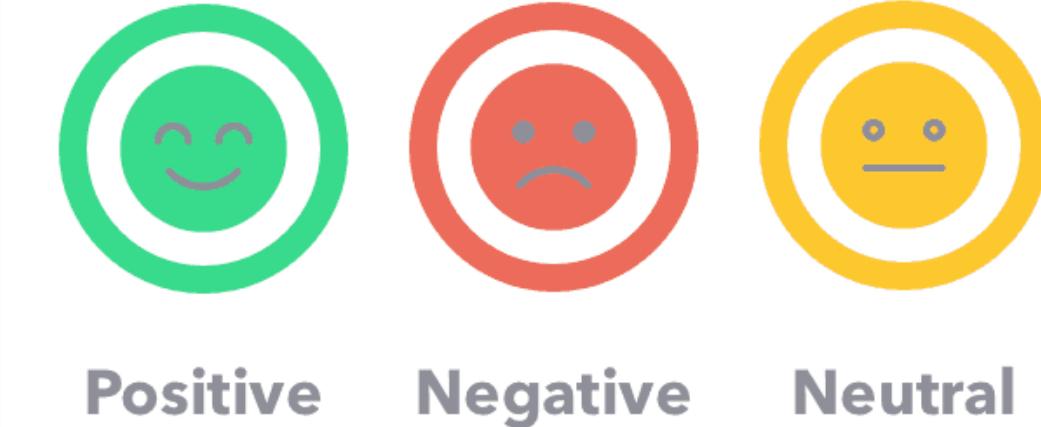
5. Sentiment Analysis

6. Grammar Checkers

7. Email Classification and Filtering



Sentiment Analysis



Module II: Data science Topics

Applications of Machine Learning



Applications of Machine Learning

- Recommender systems are significant applications of machine learning.
- Predictive analytics utilizes techniques like decision trees and Bayesian analysis.
- Understanding precision versus recall and overfitting is crucial in applying machine learning.
- Machine learning finds applications in various sectors, including fintech.
- Recommendations in fintech mirror those in platforms like Netflix or Facebook.
- Fraud detection, particularly in retail banking, is a critical area for machine learning.

Machine Learning in Fintech

- Machine learning models analyze previous transactions to identify fraudulent activities.
- Real-time decision-making in fraud detection is essential for timely intervention.
- Machine learning enhances risk management and security measures in fintech.

Enhanced Fraud Detection

Generative AI models can simulate various fraudulent scenarios to improve detection algorithms, making fraud prevention systems more robust and responsive.

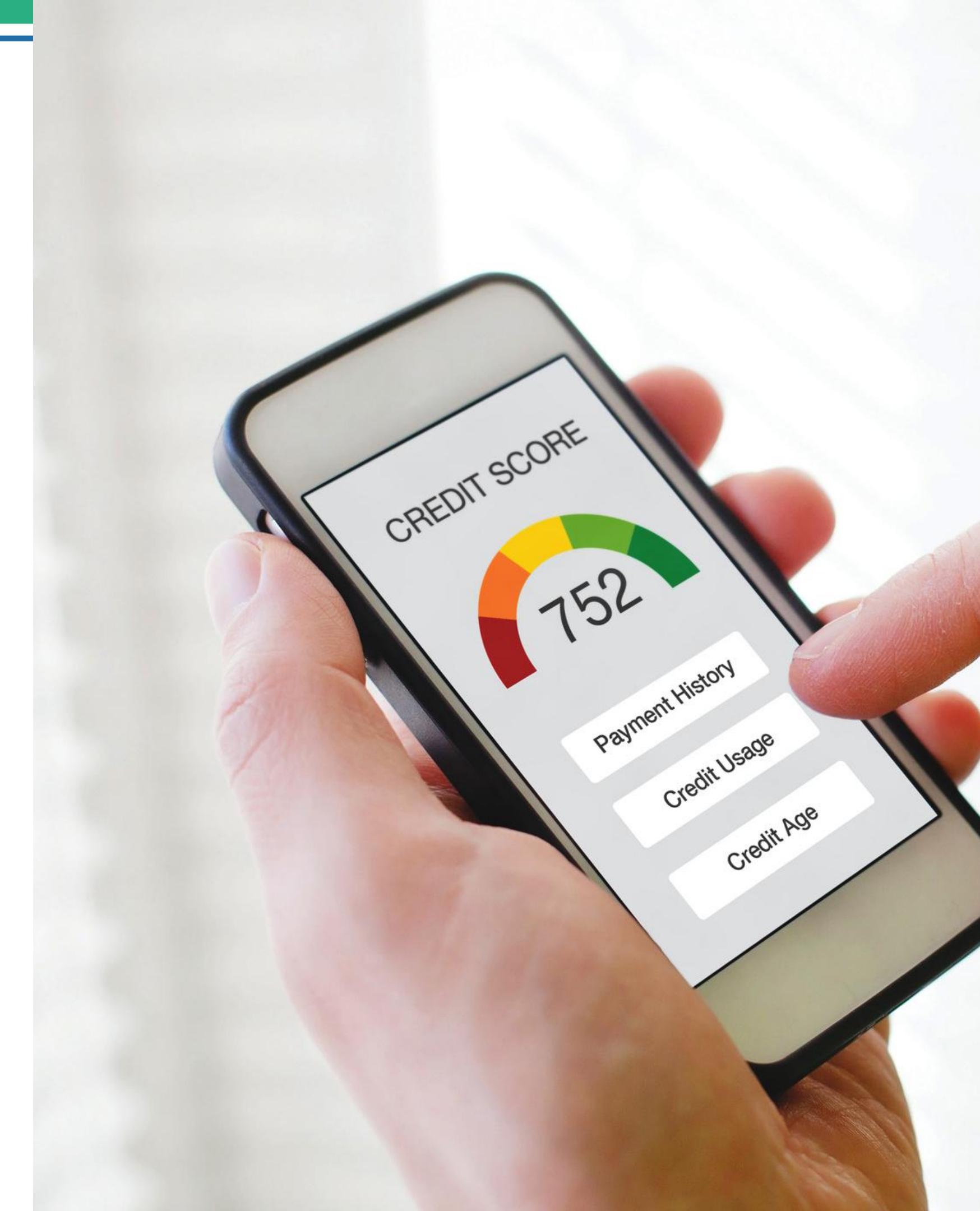


Risk Assessment And Credit Scoring:

Generative AI is reshaping risk assessment and credit scoring in the banking sector.

By creating detailed simulations of financial scenarios, generative AI tools provide deeper insights into credit risks.

This helps financial institutions improve the accuracy of their credit-scoring models, leading to smarter lending decisions.



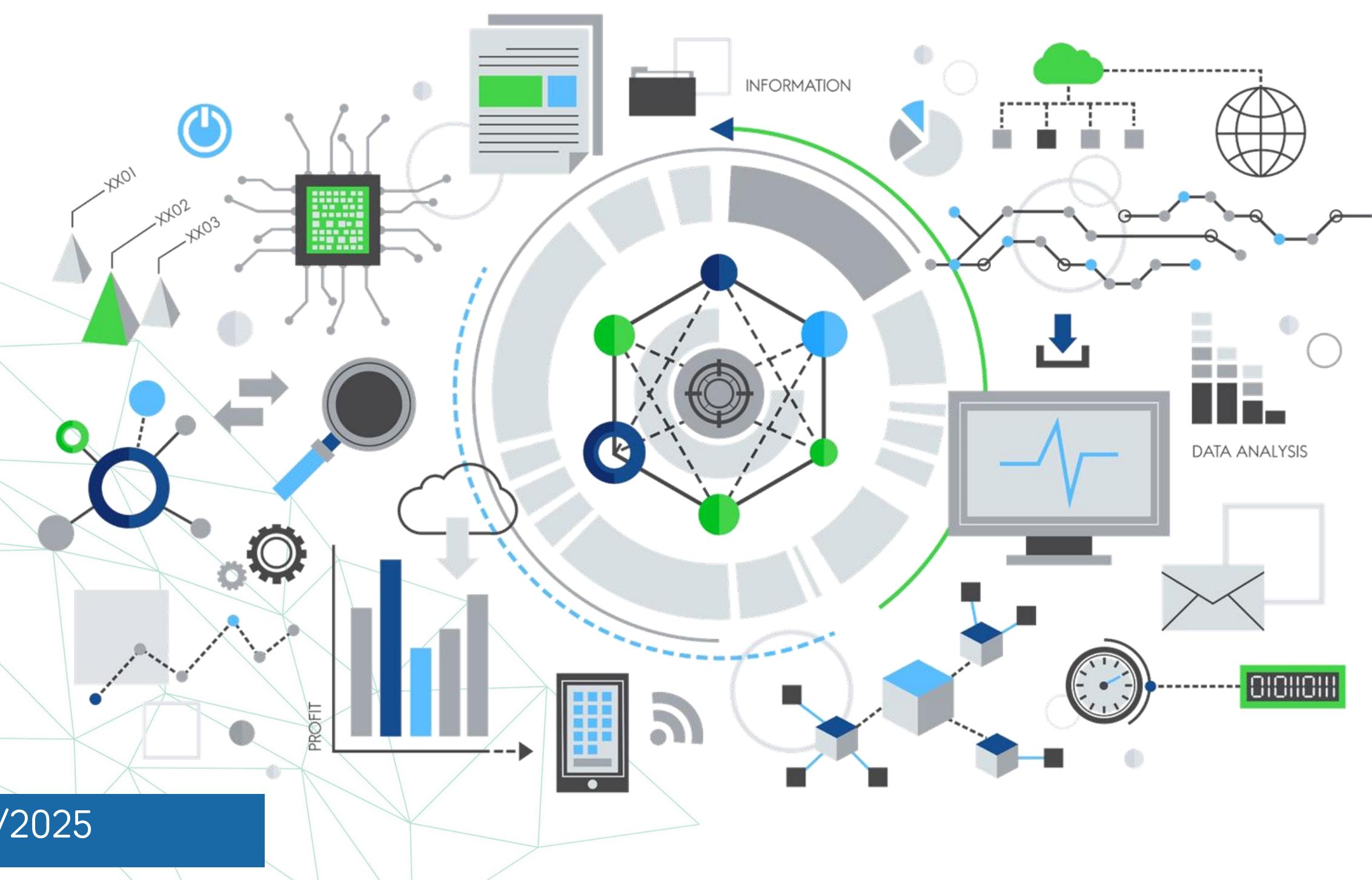
Document Processing Automation:

Generative AI excels in automating the generation and processing of complex banking documents, reducing errors, and increasing efficiency.



Module III: Data Literacy for Data Science

Understanding Data





Understanding Data Structures

Structured Data VS Unstructured Data

Can be displayed in rows, columns and relational databases



Numbers, dates and strings



Estimated 20% of enterprise data (Gartner)



Requires less storage



Easier to manage and protect with legacy solutions



Cannot be displayed in rows, columns and relational databases



Images, audio, video, word processing files, e-mails, spreadsheets



Estimated 80% of enterprise data (Gartner)



Requires more storage



More difficult to manage and protect with legacy solutions

Understanding Data Structures

- **Structured Data:**
 - Well-defined structure or data model.
 - Stored in databases with rows and columns.
 - Examples: SQL databases, spreadsheets, online forms.

Understanding Data Structures

- **Semi-structured Data:**

- Has some organizational properties but no fixed schema.
- Uses tags and metadata for grouping and hierarchy.
- Examples: XML, JSON, emails, TCP/IP packets, Zipped files.



Call center log



JSON data



Web pages



Tweets organized
by hashtag



Emails sorted
by folder



Server logs

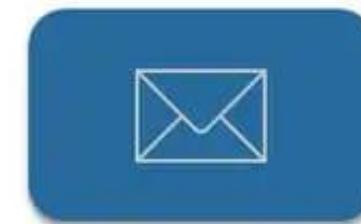
Understanding Data Structures

- **Unstructured Data:**

- Lacks identifiable structure, not stored in rows/columns.
- Includes web pages, social media feeds, multimedia files, and documents.
- Stored in files for a manual analysis or NoSQL databases for analysis tools.



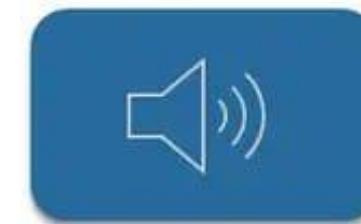
Text documents



Emails



Images



Audio files

Module III: Data Literacy for Data Science

Data Sources



Data Sources Overview

- Relational Databases:
 - Organize data in structured tables (SQL Server, Oracle, MySQL).
 - Used for internal applications and business activities.
- Flatfiles and XML Datasets:
 - Flat files store data in plain text format (CSV, spreadsheet files).
 - XML files use tags to mark up data with hierarchical structures.
- APIs and Web Services:
 - Provide data access for multiple users or applications.
 - Examples: Twitter, Facebook, Stock Market APIs, Data Lookup APIs.

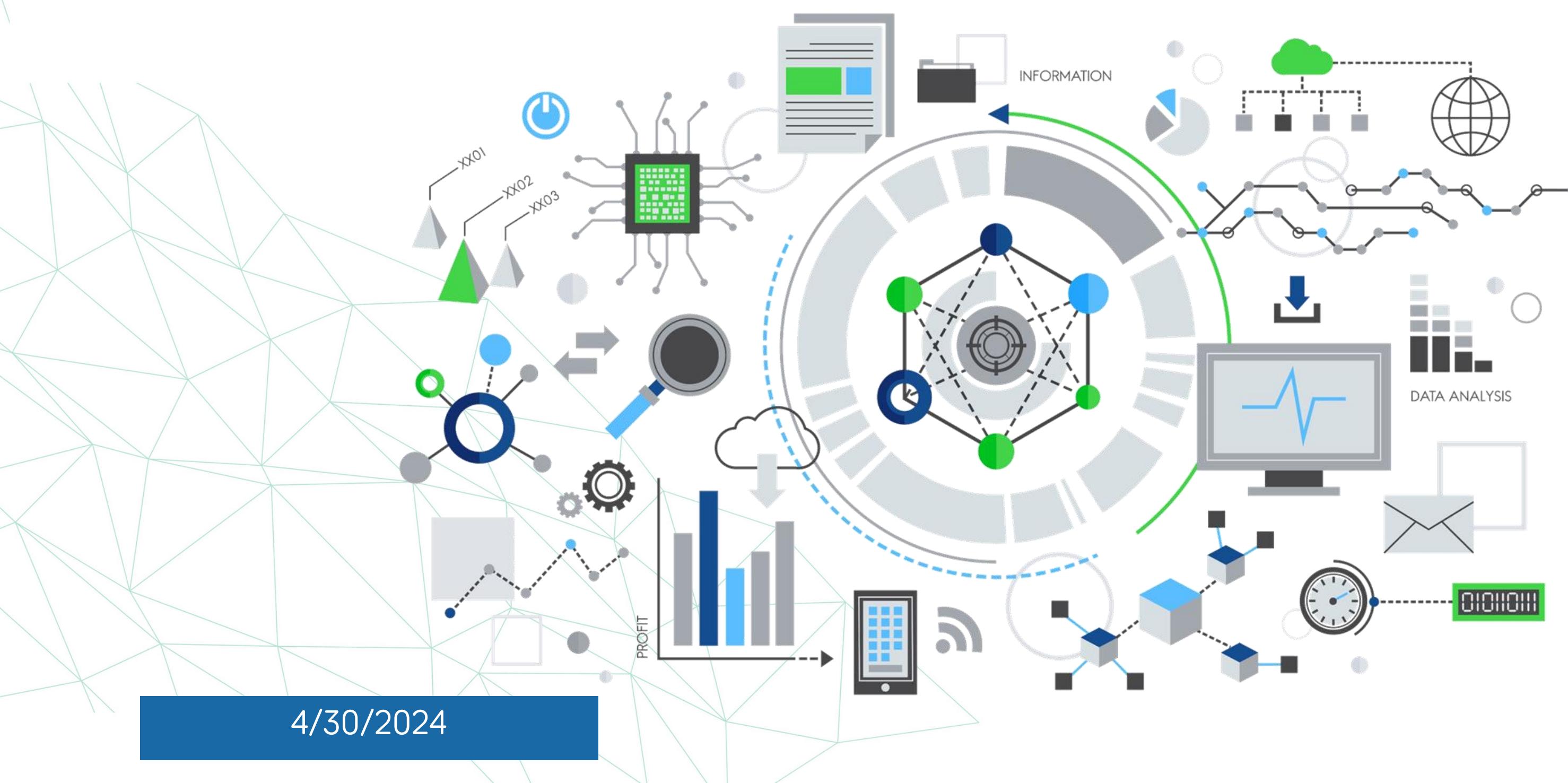
Advanced Data Sources

- Web Scraping:
 - Extracts data from unstructured web sources.
 - Used for product details, sales leads, forum posts, and more.
 - Popular tools: BeautifulSoup, Scrapy, and Selenium.
- Data Streams and Feeds:
 - Aggregates constant streams of data from various sources.
 - Applications in real time flights, surveillance, social media for sentiment analysis.
 - Popular technologies: Kafka, Spark, and Apache Storm.
- RSS (Really Simple Syndication) Feeds:
 - Captures updated data from online forums and news sites.
 - Streamed to user devices using feed readers for real-time updates.



Module III: Data Literacy for Data Science

Viewpoints: Working with Varied Data Sources and Types



Challenges in Working with Data Sources

- Diverse data formats require adapting data handling methods.
- SQL is crucial for data movement, structuring, and security.
- Migrating data between relational databases faces vendor changes and versioning challenges.
- Flexibility is key when working with various data sources.
- Evaluating multiple solutions is necessary for consistent and performant data movement.

Relational Databases and Alternatives

- Relational databases struggle with unstructured data like logs, XML, and JSON.
- Heavy write-intensive applications such as IoT pose challenges for relational databases.
- Alternatives like Google BigTable, Cassandra, and HBase gain popularity for specific data handling needs.
- Data engineers deal with standard formats (CSV, JSON, XML) and proprietary formats.
- Data integration spans relational databases, NoSQL databases, and Big Data repositories.

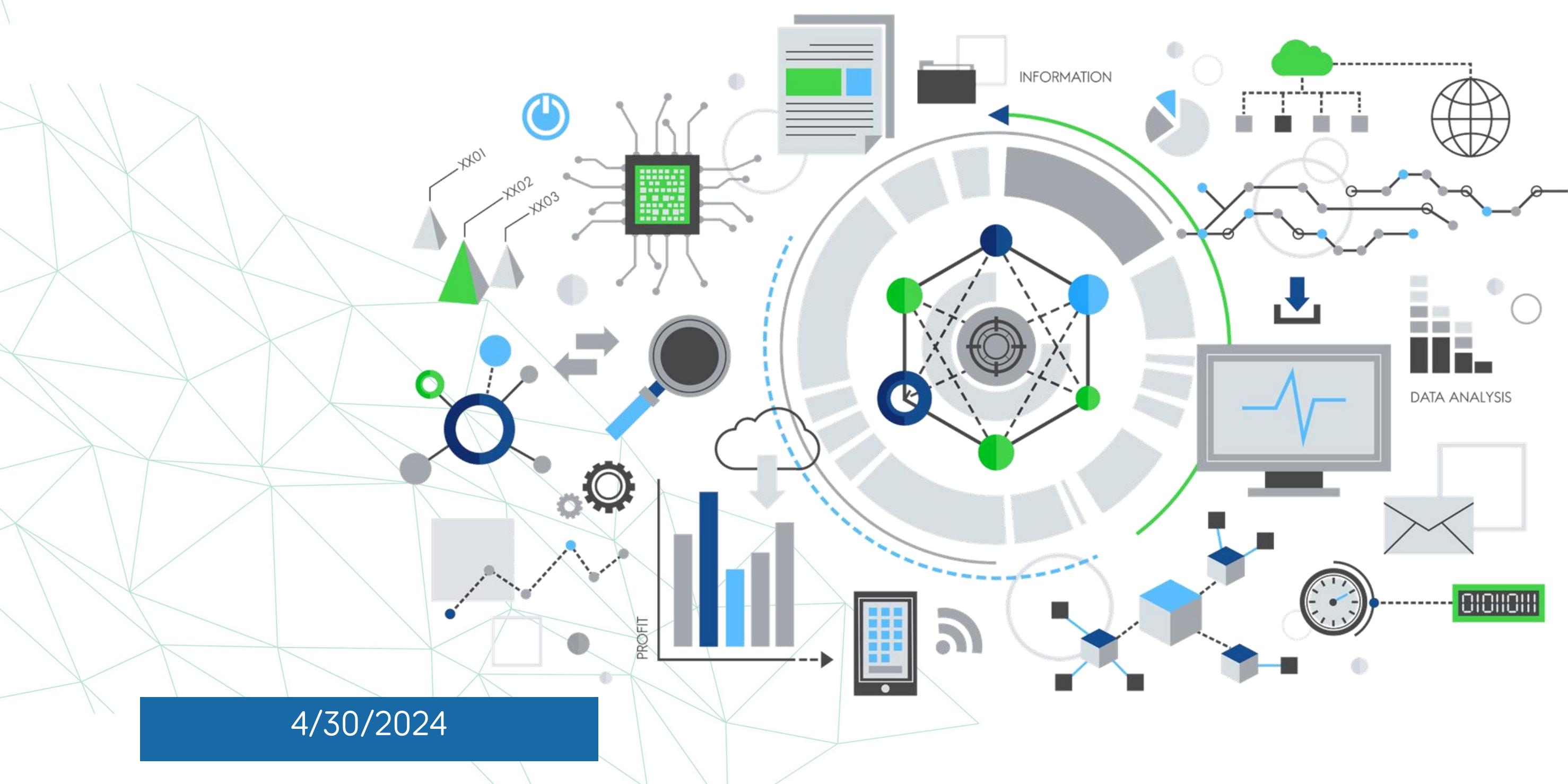
Handling Complex Data Formats

- Log data's unstructured nature demands custom parsing tools.
- XML data's resource intensity challenges efficient data handling.
- JSON's popularity stems from its simplicity and usage in RESTful APIs.
- Apache Avro gains traction for its efficient data storage capabilities.
- Import/export differences between Db2 and SQL Server
- present integration challenges.



Module III: Data Literacy for Data Science

Lesson Summary: Understanding Data



Understanding Data

- Data is foundational to data science, available in structured, semi-structured, or unstructured forms.
- Structured data adheres to a data model, stored in databases with well-defined schemas.
- Semi-structured data lacks a fixed schema but has organizational properties and metadata.
- Metadata categories include technical, process, and business, crucial for data organization.
- Unstructured data is heterogeneous, coming from various sources and requiring AI for analysis.
- Data can be sourced electronically, from internal applications, publicly available sets, or purchased proprietary data.

Data Storage and Access

- Flat files like CSV and spreadsheets were common, XML structured older data, and JSON is now prevalent.
- JSON allows flexible data transfer between evolving structures, accessible through RESTful APIs.
- APIs from platforms like Twitter and Facebook provide data for sentiment analysis and opinion mining.



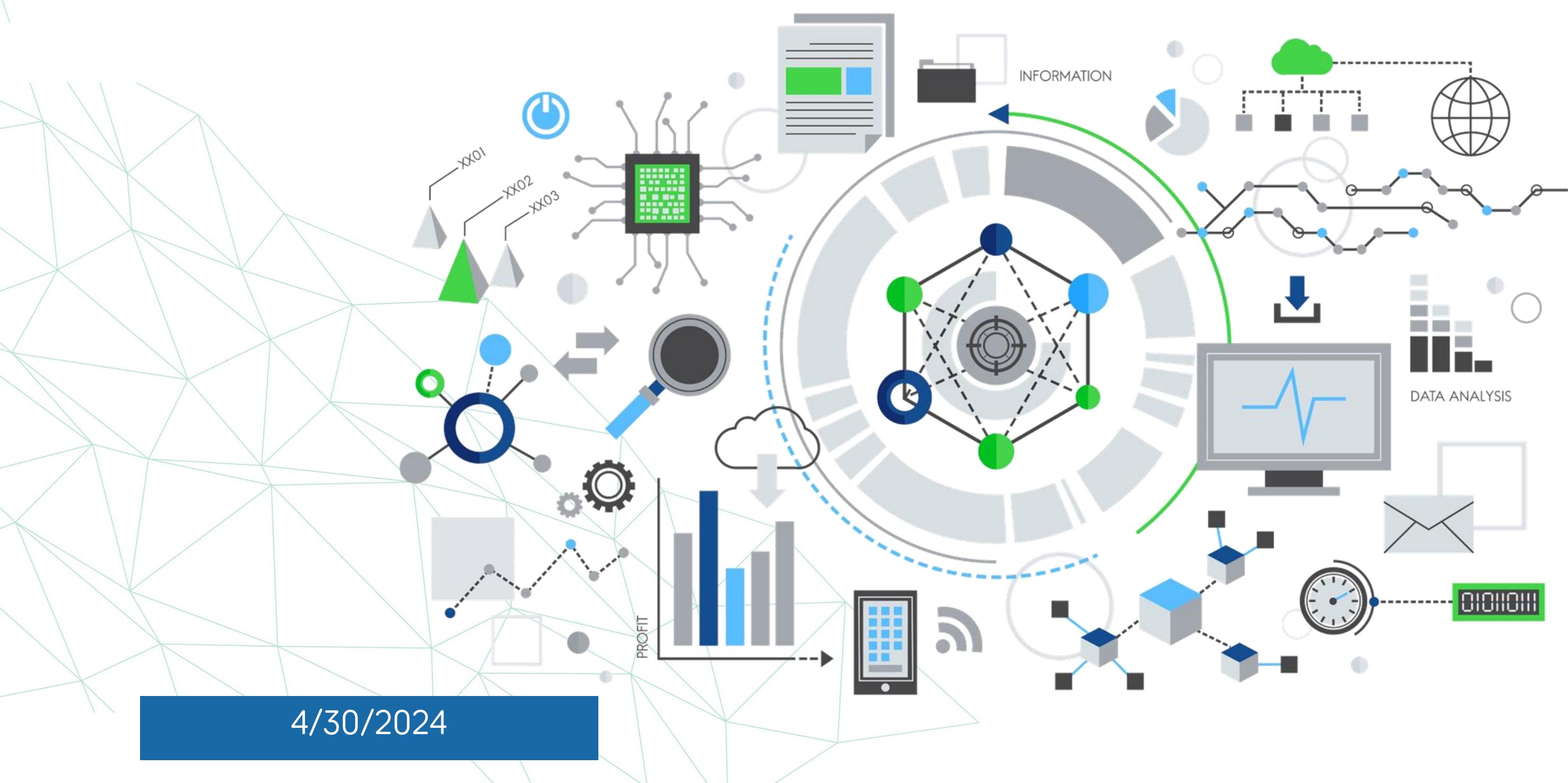
Data Storage and Access

- Data gathering and management are typically handled by data engineers.
- Data scientists deal with extensive, terabyte-sized data sets from sources like IoT and social media.
- Understanding modern data ecosystems is crucial for data scientists to analyze data effectively.



Module III: Data Literacy for Data Science

Data Collection and Organization



Understanding Data Repositories

- A data repository encompasses organized data used for business operations and analysis.
- It includes small to large database infrastructures with one or more databases.
- Types of repositories include databases, data warehouses, and big data stores.
- Databases are designed for input, storage, retrieval, and modification of data.
- Relational databases (RDBMS) organize data into tables with SQL for querying.
- Non-relational databases (NoSQL) offer flexibility, speed, and scalability for big data.

Advanced Data Repository Concepts

- Data warehouses consolidate data from various sources for analytics and BI.
- The ETL process (Extract, Transform, Load) cleans and integrates data into warehouses.
- Data Marts and Data Lakes are subsets of warehouses for specific purposes.
- Both relational and non-relational repositories are used in data warehousing.
- Big Data Stores handle distributed storage and processing of large datasets.
- Repositories enhance data isolation and reporting efficiency and serve as archives.



Module III: Data Literacy for Data Science

Relational Database Management System



Introduction to Relational Databases

- **What is a Relational Database?**

Organized collection of data in tables.

Tables are linked based on common data.

Each table has rows (records) and columns (attributes).

- **Key Concepts**

Table Example: Customer table with Company ID, Name, Address, Phone.

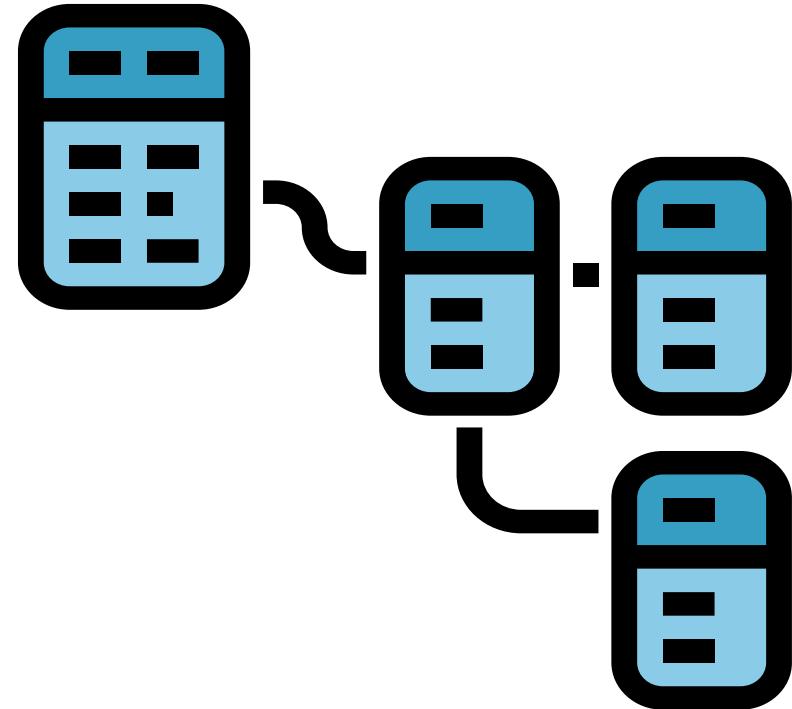
Linking Tables: Relating customer and transaction tables via Customer ID.

- **Advantages of Relational Databases**

Data Organization: Structured storage and retrieval of large volumes.

Data Integrity: Minimized redundancy, consistent data types.

Querying Power: Uses SQL for efficient data processing and retrieval.



Applications and Limitations

- **Use Cases of Relational Databases**
 - **OLTP:** Online Transaction Processing for fast, frequent data transactions.
 - **Data Warehouses:** Analyzing historical data for business intelligence.
 - **IoT Solutions:** Lightweight database for collecting and processing IoT data.
- **Limitations of Relational Databases**
 - **Data Type Limitations:** Not suitable for semi-structured or unstructured data.
 - **Schema Requirements:** Need identical schemas for data migration.
 - **Field Length Limitations:** Data fields have length restrictions.
- **Conclusion**
 - Despite limitations, relational databases remain essential for structured data management and common business applications.

Module III: Data Literacy for Data Science

NoSQL



Introduction to NoSQL Databases

- **What is NoSQL?**
 - Non-relational database design for flexible data storage.
 - Flexible schemas for scalability, performance, and ease of use.
- **Key Concepts**
 - Flexible Schemas: Not limited by fixed row/column structures.
 - Data Models: Four common types - Key-value store, Document-based, Column-based, and Graph-based.
- **Types of NoSQL Databases**
 - Key-value Store: Ideal for user session data and real-time recommendations.
 - Document-based: Flexible indexing for eCommerce and CRM platforms.
 - Column-based: Efficient for time-series and IoT data storage.
 - Graph-based: Visualizing and analyzing interconnected data.

Advantages and Differences

- **Advantages of NoSQL**
 - Scalability: Distributed systems for large data volumes.
 - Cost-Effective: Scale-out architecture with low-cost hardware.
 - Agility: Simplified design for better control and scalability.
- **Key Differences from Relational Databases**
 - Schema Flexibility: NoSQL allows schema-agnostic data storage.
 - Cost Considerations: Lower maintenance costs compared to RDBMS.
 - ACID Compliance: Relational databases offer transaction reliability.
- **Conclusion**
 - NoSQL databases offer scalability, cost-effectiveness, and flexibility, making them valuable for modern applications despite differences from traditional RDBMS.

Module III: Data Literacy for Data Science

Data Marts, Data Lakes, ETL, and Data Pipelines



Understanding Data Warehouses, Data Marts, and Data Lakes

- **Data Warehouse Overview**
 - Multi-purpose storage for analysis-ready data.
 - Single source of truth for historical and current data.
- **Data Mart Definition**
 - Sub-section of a data warehouse for specific business functions.
 - Provides isolated security and performance for targeted analytics.
- **Data Lake Concept**
 - Storage for structured, semi-structured, and unstructured data.
 - Retains raw data without predefined use cases.

Data Lake

Data Warehouse

Data Mart

Most Important Use

Group & Use-Cases

Predictive & Advanced Analytics

Multi-Purpose Enabler of Operational & Performance Analytics

Line of Business Specific Reporting & Analytics

Time-to-Market

Questions & Solutions



Weeks - Months



Hours - Days

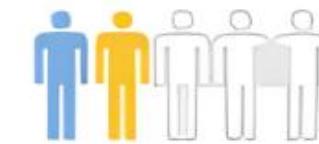


Minutes - Hours

Cost

Implementation & Ownership

\$\$\$\$\$

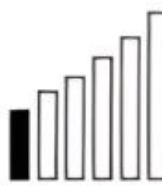
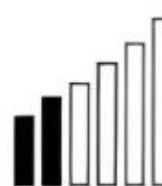


Users

(# & Types)

Data Growth

Volume & Variety



Exploring ETL Process and Data Pipelines

- **ETL Process Explanation**
 - Extract: Collecting raw data from various sources.
 - Transform: Cleaning, standardizing, and converting data for analysis.
 - Load: Transporting processed data to a data repository.
- **Types of ETL**
 - Batch Processing: Scheduled transfers in large chunks.
 - Stream Processing: Real-time data processing before loading.
- **Data Pipelines Overview**
 - Broader term including ETL for data movement.
 - Supports batch and streaming data processing for various applications.

Module III: Data Literacy for Data Science

Viewpoints: Considerations for Choice of Data Repository



Factors in Choosing a Data Repository

- **Data Type Consideration**
 - Structured, semi-structured, or unstructured data.
 - Impact on schema design and storage methods.
- **Performance and Storage Needs**
 - Performance requirements for data access and processing.
 - Volume of data and storage capacity needed.
- **Data Access and Encryption**
 - Frequency of data access and update requirements.
 - Data encryption needs for security and compliance.

Additional Considerations and Data Repository Types

- **Repository Compatibility and Purpose**
 - Compatibility with existing tools and programming languages.
 - Purpose of the repository: transactional, analytical, or archival.
- **Scalability and Security Features**
 - Scalability for long-term data growth.
 - Security features such as access control and data encryption.
- **Diverse Repository Types**
 - Preferred relational databases, open-source options, and unstructured data sources.
 - Considerations for specific use cases like product recommendations or analytics.

Module III: Data Literacy for Data Science

Data Integration Platforms



Data Integration Overview

- **Definition:** Gartner defines data integration as the practice, techniques, and tools for ingesting, transforming, and provisioning data across various types.
- **Usage Scenarios:** Includes data consistency, master data management, data sharing, migration, and consolidation.
- **Analytics and Data Science:** Involves accessing, transforming, merging, ensuring data quality, governance, and delivering integrated data for analytics.
- **Example:** Extracting customer data from sales, marketing, and finance systems for unified analytics.

Data Integration Capabilities

- **Modern Solutions:** Offer extensive connectors, open-source architecture, batch, and continuous processing, integration with Big Data sources, and additional functionalities.
- **Portability:** Supports cloud models, including single cloud, multi-cloud, or hybrid environments.
- **Market Overview:** Various platforms and tools available, including IBM's offerings, Talend, SAP, Oracle, Denodo, and others.
- **Evolution:** Data integration evolves with technology advancements and increasing data complexity in business decision-making.

Questions & Answers



Thank you!