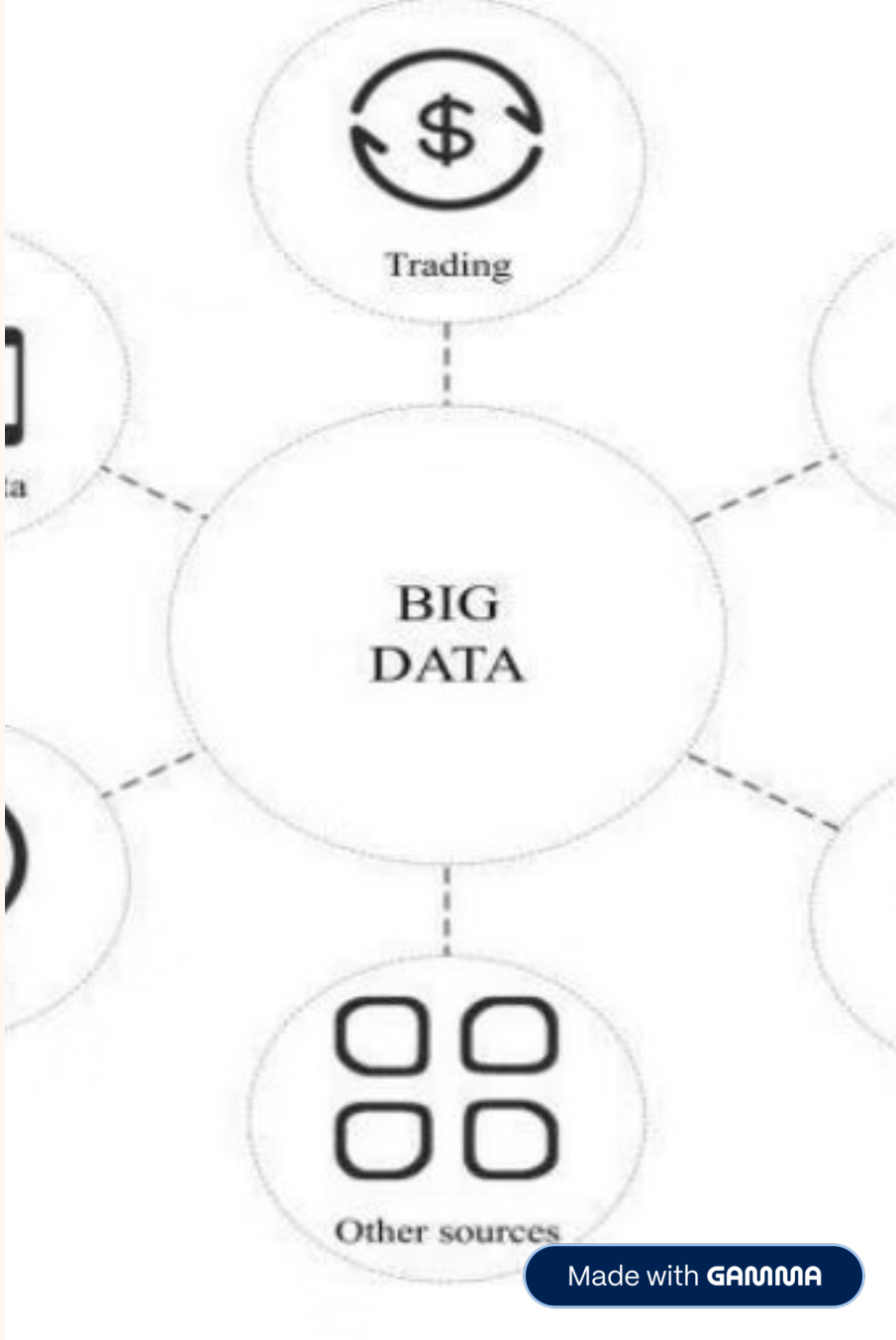
Introduction to Big Data and Data Analytics

Explore the world of Big Data, its characteristics, and how it's transforming data analysis to unlock new possibilities.



Learning Objectives

1 **Understand Big Data**

Grasp the concept and its evolution in the digital era.

2 **Role in AI & Data Science**

Appreciate Big Data's significance in these fields.

3 **Features & Handling**

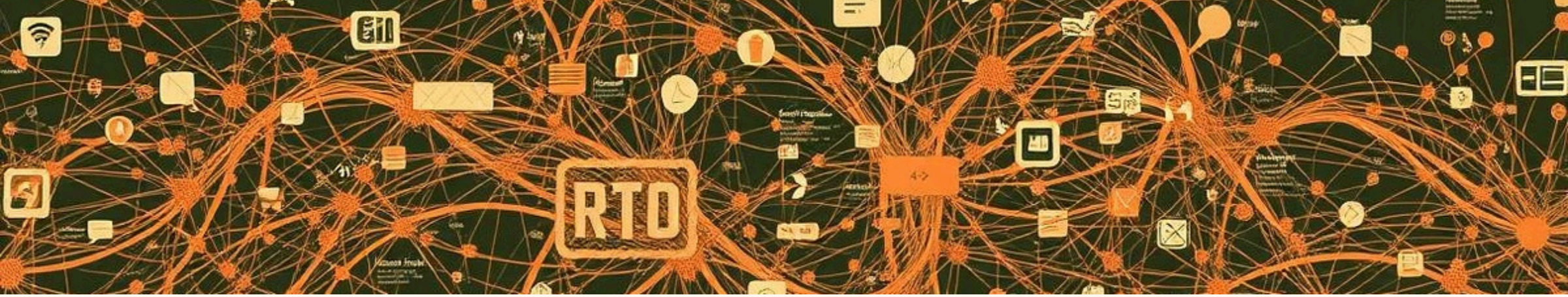
Learn Big Data characteristics and their management in analytics.

4 **Real-World Applications**

Discover how Big Data transforms data analysis across various fields.

5 **Mining Data Streams**

Comprehend the process of extracting insights from continuous data flows.



What is Big Data?

Small Data

Easily comprehensible datasets, ideal for everyday tasks and simple decision-making. Example: A small store tracking daily sales.

Big Data

Extremely large, complex datasets beyond traditional program handling. Sourced from transactional, machine, and social data. Requires special tools for analysis to gain insights. Example: Amazon/Netflix recommendations.

Types of Big Data

Definition	Quantitative data with a defined structure	Mix of quantitative and qualitative properties	No inherent structures or formal rules
Organization	Clearly defined columns	Less organized than structured data	No organization, exhibits variability
Examples	Customer info, transaction records	XML, JSON, HTML files	Audio, images, videos, emails, social media posts

Advantages of Big Data

Enhanced Decision Making

Data-driven decisions from large, diverse datasets.

Improved Efficiency & Productivity

Identify inefficiencies, streamline processes, optimize resources.

Better Customer Insights

Deeper understanding of behavior, preferences, and needs.

Competitive Advantage

Uncover market trends, identify opportunities, stay ahead.

Innovation & Growth

Develop new products, services, and business models.

Disadvantages of Big Data

Privacy & Security Concerns

Risks of unauthorized access, breaches, and misuse of personal info.

Data Quality Issues

Challenges in ensuring accuracy, reliability, and completeness.

Technical Complexity

Requires specialized skills and expertise for infrastructure and tools.

Regulatory Compliance

Challenges in meeting data protection laws like GDPR.

Cost & Resource Intensiveness

High costs for acquisition, storage, processing, and skilled staff.

Characteristics of Big Data: The 6Vs

Velocity

Speed of data generation, delivery, and analysis (e.g., 40,000 Google searches/sec).

Volume

Huge amount of data generated daily, typically exceeding gigabytes (e.g., 328.77 million TB/day).

Variety

Data in various formats: structured, unstructured, semi-structured (text, images, audio, video).

Veracity

Consistency, accuracy, quality, and trustworthiness of data; requires cleaning.

Value

Business value derived from data analysis; critical for insights.

Variability

Regularity and dependability of data stream structure despite unpredictability.

Working on Big Data Analytics

1. Gather Data

Collect structured and unstructured data from diverse sources like cloud, mobile apps, IoT sensors.

2. Process Data

Prepare data for accurate results using batch or stream processing.

3. Clean Data

Improve data quality by correcting formatting and eliminating duplicates/irrelevant entries.

4. Analyze Data

Turn cleaned data into insights using advanced analytics processes.

Tools: Tableau, Apache Hadoop, Cassandra, MongoDB, SAS.

Mining Data Streams

What is a Data Stream?

A continuous, real-time flow of data from sources like sensors, satellite images, and web traffic.

What is Mining Data Streams?

Extracting patterns and knowledge from real-time data flows without complete storage. Example: Analyzing website traffic spikes for election results.

Future of Big Data Analytics

Real-Time Analytics

Instantaneous data processing for immediate insights and actions.

Advanced Predictive Models

Integration of sophisticated ML/AI for precise trend forecasting.

Quantum Computing

Revolutionary processing power to solve complex problems faster.