Fundamentals of Big Data

Derbew Felasman
Debre Brihan University

Data

- **What is Data?**
- Anything and everything is Data

Data: Where Does it come from???

- It comes from Everywhere:
- We speak
- We Move
- Sensors
- Computers
- Documents

Classification of Data

- Structured
- Semi-Structured
- Unstructured
- Human Generated Data – email, blogs, videos, pictures, etc.
- Machine Generated Data – Automatics alerts, logs, sensor data

What is Big Data

- **Big data** is the term and technology for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” -- Gartner

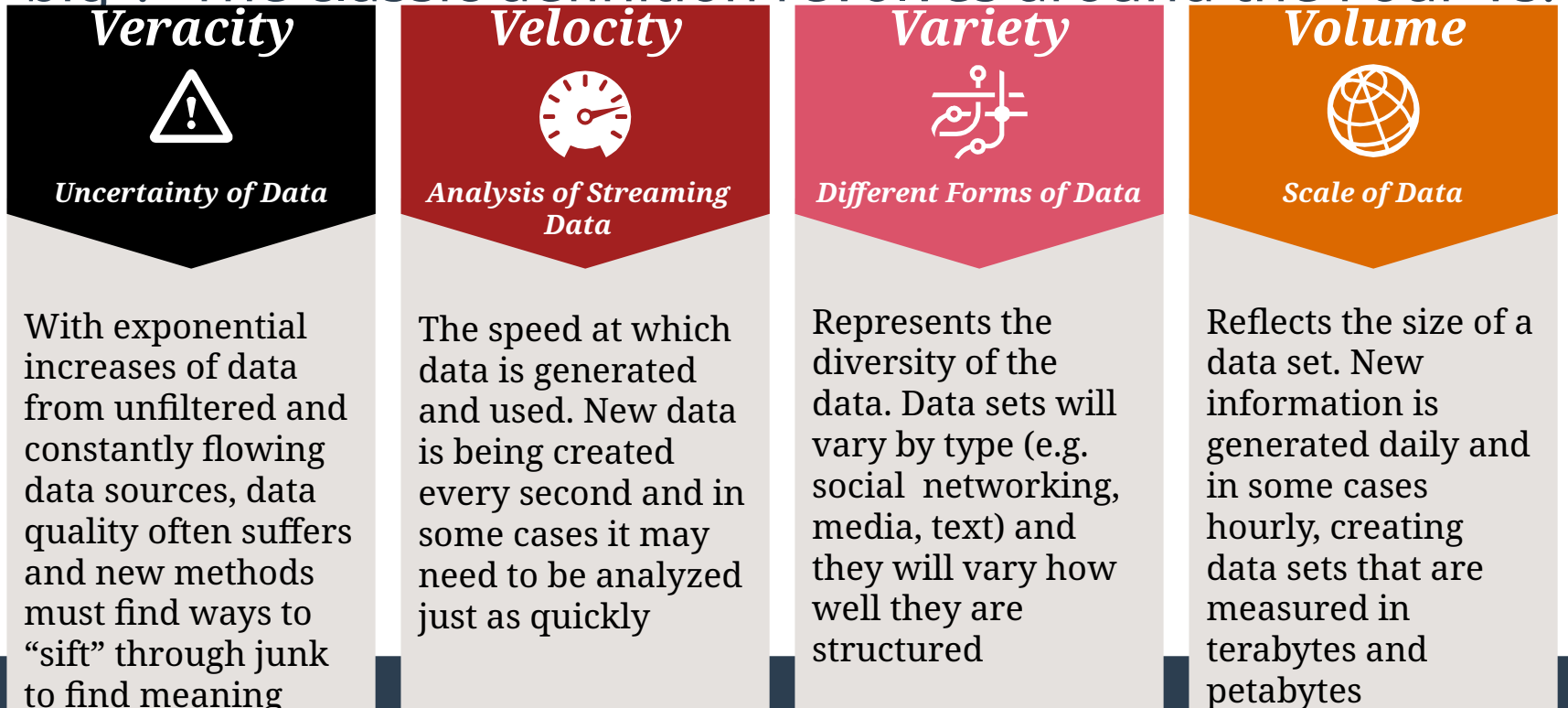
Cont..

- It can be structured, semi-structured, and unstructured.
- Unstructured data collectively account for 80 to 90% of big data.
- Challenges include analysis, capture, search, sharing, storage, transfer, visualization, querying, updating and information privacy.

what makes the data “big”?

Cont..

- The are many examples of "data", but what makes some of it “big”? The classic definition revolves around the Four Vs.



Characteristics of Big Data

Volume

Data at scale

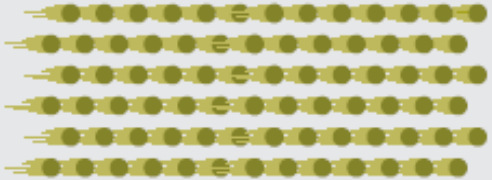
Terabytes to petabytes of data

Variety

Data in many forms

Structured, unstructured, text, multimedia

Velocity



Data in motion

Analysis of streaming data to enable decisions within fractions of a second

Veracity

Data uncertainty

Managing the reliability and predictability of inherently imprecise data types

Cont..

- Even more important than its definition is what Big Data promises to achieve: intelligence in the moment.

	Traditional Techniques & Issues	*Big Data Differentiators*
Veracity	• **Does not account for biases, noise and abnormality in data**	• Data is stored, and **mined meaningful to the problem being analyzed** • **Keeps data clean and processes to keep 'dirty data' from accumulating in your systems**
Velocity	• **No real time analysis**	**In real-time:** • **Dynamically analyze data** • Consistently integrate new information • Auto deletes unwanted to ensure **optimal storage**
Variety	• **Compatibility issues** • Advanced analytics **struggle with non-numerical data**	• Frameworks **accommodate varying data types and data models** • **Insightful analysis with very few parameters**
Volume	• Analysis is limited to **small data sets** • Analyzing large data sets = **High Costs & High Memory**	• Scalable for **huge amounts of multi-sourced data** • Facilitation of **massively parallel processing** • **Low-cost data storage**

Big Data: 6V

Big Data

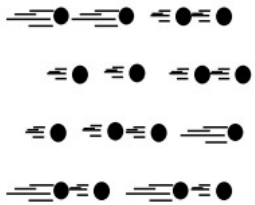
Open Data

Volume

Data at Rest

Terabytes to exabytes of existing data to process

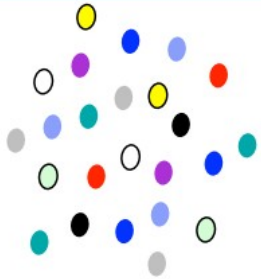
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

Variety



Data in Many Forms

Structured, unstructured, text, multimedia

Veracity

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Visibility

Data in the Open

Open data is generally open to anyone. Which raises issues of privacy. Security and provenance

Value

Data of Many Values

Large range of data values from free (data philanthropy to high value monetization)

Who is generating Big Data?

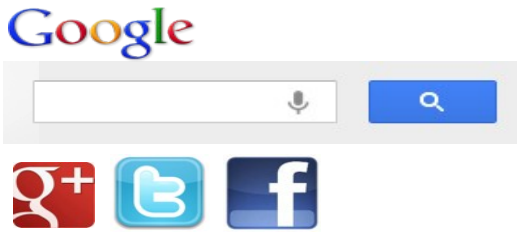
Social

User Tracking & Engagement

eCommerce

Financial Services

Real Time Search



Why is Big Data valuable?

Accessibility to Data	Enhanced visibility of relevant information and better transparency to massive amounts of data. Improved reporting to stakeholders.
Decision Making	Next generation analytics can enable automated decision making (inventory management, financial risk assessment, sensor data management, machinery tuning).
Marketing Trends	Segmentation of population to customize offerings and marketing campaigns (consumer goods, retail, social, clinical data, etc).
Performance Improvement	Exploration for, and discovery of, new needs, can drive organizations to fine tune for optimal performance and efficiency (employee data).
New Business Models/Services	Discovery of trends will lead organizations to form new business models to adapt by creating new service offerings for their customers. Intermediary companies with big data expertise will provide analytics to 3rd parties.

Cont..

$1 Trillion

One study estimated the potential value of big data in the U.S. health care, European public sector administration, global personal location data, U.S. retail, and global manufacturing to be over $1 trillion U.S. dollars per year[1].

Another study estimated the value of big data in the areas of customer intelligence, supply chain intelligence, performance improvements, fraud detection, and quality and risk management to be $41 billion per year in the UK alone[2].

$41 Billion

(1) J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey & Company, 2011.
(2) Centre for Economics and Business Research, “Data equity: unlocking the value of big data,” SAS, 2012.

Cont..

- "… the sexy job in the next 10 years will be data expert/Data miniers/statisticians," Hal Varian, Google Chief Economist
- the U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011
- New Big Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,…
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley
  - One proposal (elsewhere) for an MS in "Big Data Science"

It’s not just about the data…

- It is important to understand the distinction between Big Data sets (large, unstructured, fast, and uncertain data) and ‘Big Data Analytics’.

Big Data
Refers to the DATA only

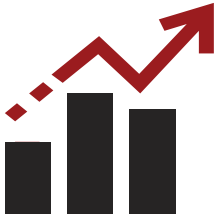
\+

Big Data Analytics
Methods of using Big Data to generate insight

1	**Machine Learning/Deep Learning**	• Leveraging a computer’s ability to learn without being explicitly programmed to solve business problems
2	**IoT (Internet of Things) & Sensor Analytics**	• Understanding value drivers from the ever-growing network of connected physical objects and the communication between them
3	**Modeling Willingness-to-Pay**	• Mining product reviews to estimate willingness-to-pay for product features
4	**Natural Language Processing**	• Understanding human speech as it is spoken through application of computer science, AI, and computational linguistics
5	**Analyzing Data @ Scale**	• Using distributed computing and machine learning tools to analyze hundreds of gigabytes of data
6	**Creating a Streaming Consumer Behavior Data Lake**	• Mining social data in real time to understand when and where consumers are making choices

It’s also about what, how, and why you use it

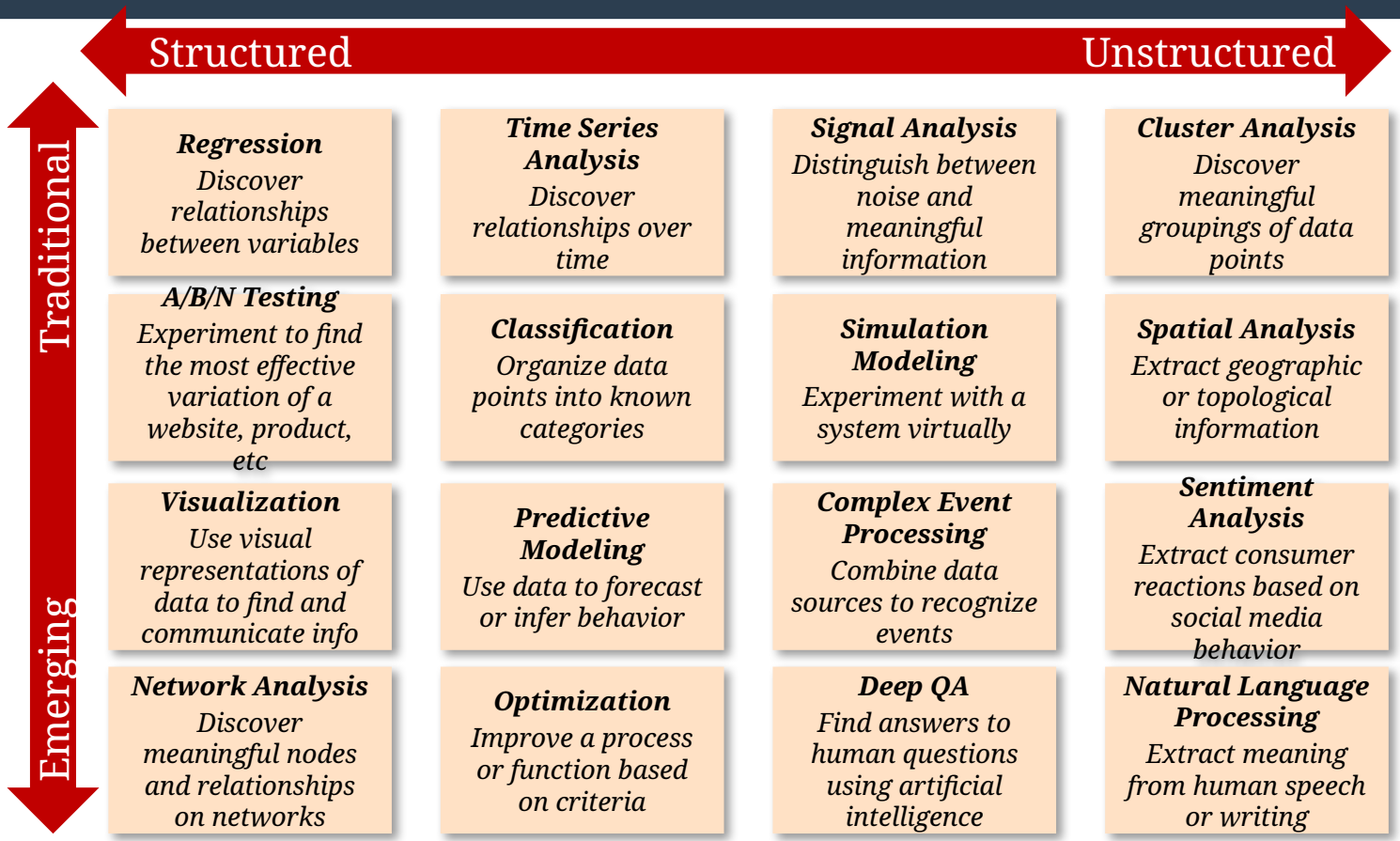
Big Data Analytics – the process of harnessing Big Data to yield actionable insights – is a combination of five key elements:

Decisions	***Analytics***	***Data***	***Technology***	***Mindset & Skills***
The value of Big Data Analytics is driven by the unique decisions facing leaders, companies, and countries today. In turn, the type, frequency, speed, and complexity of decisions drive how Big Data Analytics is deployed.	To leverage the variety and volume of Big Data while managing its volatility, advanced analytical approaches are necessary, such as natural language processing, network analysis, simulative modeling, artificial intelligence, etc.	Big Data Analytics is about operationalizing new and more data, but it is also about data quality, data interoperability, data disaggregation, and the ability to modularize data structures to quickly absorb new data and new types of data.	To store, manage, and use Big Data often requires investments in new technologies and data processing methods, such as distributed processing (e.g., Hadoop), NoSQL storage, and Cloud computing.	Big Data Analytics requires firm commitment to using analytics in decision-making; a decisive mentality capable of employing in-the-moment intelligence; and investment in analytical technology, resources, and skills.
				

What is big data analytics?

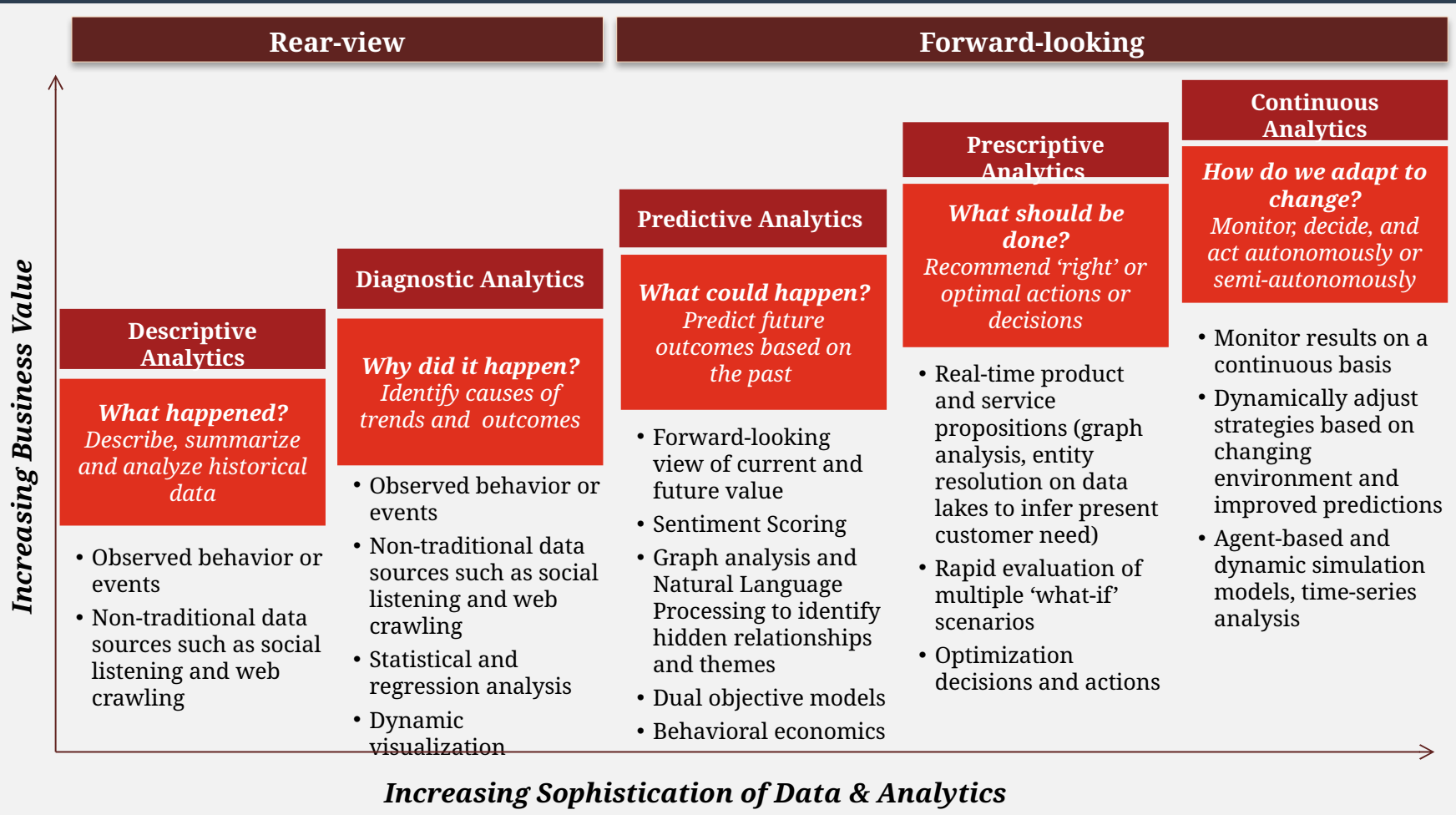
- Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions.

Big Data Analytical Capabilities



Continuing increases in processing capacity have opened the door to a range of advanced algorithms and modeling techniques that can produce valuable insights from Big Data.

Forward-Looking vs. Rear-View Analytics



Big Data Analytic improves the speed and efficiency with which we understand the past, and opens up entirely new avenues for preparing for an adapting to the future.

What is Data Mining?

- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.
- Definition (Fayyad et. al): The non-trivial discovery of novel, valid , comprehensible and potentially useful patterns from data.
- What is a pattern? A relationship in the data. E.g.,
- On Thursday nights people who buy diapers also tend to buy beer
- People with good credit ratings are less likely to have accidents
- Male consumers, 37+, income bracket 50K-75K spend between $25-$50 per catalog order

Web Mining

- Discovering interesting and useful information from Web content and usage
- Three types: Web usage, Web structure, Web content
- Examples:
  - Web search, e.g. Google, Yahoo, MSN, Ask, …
  - Specialized search: e.g. Froogle (comparison shopping), job ads (Flipdog)
  - eCommerce :
    - Recommendations: e.g. Netflix, Amazon
    - improving conversion rate: next best product to offer
  - Advertising, e.g. Google Adsense
  - Improving Web site design and performance

Business intelligence (BI)

- BI is a broad term encompassing technologies, methodologies, and applications that enable organizations to collect, analyze, and transform data into actionable insights.
- It refers to a process that helps organizations transform their data into actionable insights.
- It comprises a set of software tools and methodologies (data mining, data management, data visualization, etc.) used to collect, store, access, and analyze relevant data to assist in making sound business decisions.

Cont..

- BI empowers organizations to:
  - Understand past performance and identify trends
  - Make informed decisions based on data-driven insights
  - Optimize processes, enhance efficiency, and improve productivity
  - Gain a competitive edge through data-driven innovation

Components of a BI Solution

- BI solutions typically consist of:
  - **Data warehousing:** A central repository for storing and organizing vast amounts of data from various sources
  - **Data mining:** Techniques for extracting hidden patterns and insights from data
  - **Data visualization:** Tools for transforming data into interactive charts, graphs, and dashboards
  - **Reporting:** Mechanisms for generating comprehensive reports and summaries of key performance indicators (KPIs)

BI vs. Big Data Analytics: Understanding the Key Differences

- The data-driven world has witnessed the emergence of two crucial concepts: business intelligence (BI) and big data analytics.
- While both aim to extract valuable insights from data, they differ in their scope, methodology, and application.
- Understanding these distinctions is essential for organizations to make informed decisions about their data management strategies.

Cont..

- **Data Type and Volume**
  - BI primarily deals with structured, organized data, typically stored in data warehouses and databases.
  - Big data analytics, on the other hand, encompasses a broader spectrum of data, including unstructured and semi-structured data, such as social media posts, sensor readings, and machine-generated content.
  - The sheer volume of big data poses a unique challenge, demanding specialized tools and techniques for processing and analysis.

Cont..

- **Purpose and Outcomes**
  - BI focuses on historical data analysis, providing insights into past trends, customer behavior, and market performance.
  - Big data analytics, in contrast, extends beyond historical data, enabling real-time analysis and predictive forecasting.
  - The goal of big data analytics is to uncover hidden patterns, identify potential risks and opportunities, and make data-driven decisions that shape future strategies.

Cont..

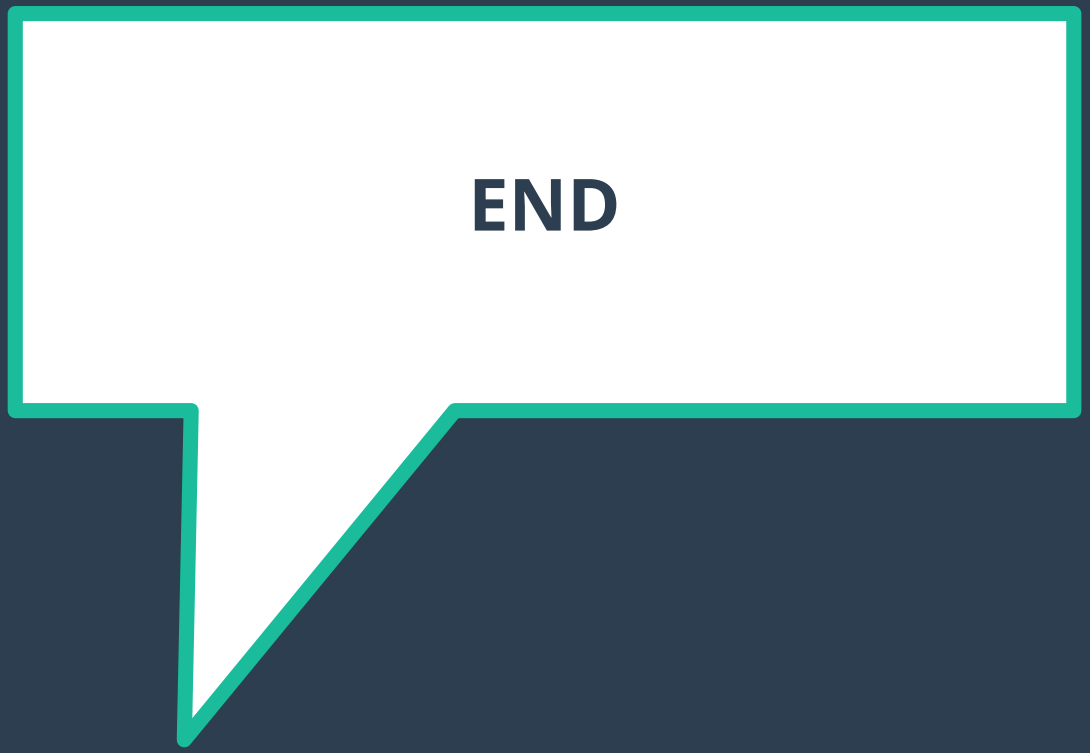
- **Tools and Technologies**
  - BI utilizes traditional data warehousing tools and reporting dashboards to visualize and analyze data.
  - Big data analytics employs a diverse set of technologies, including Hadoop, Spark, and cloud-based platforms, to handle the complexity and volume of big data.
  - Advanced analytics tools, such as machine learning and artificial intelligence, are increasingly integrated into big data analytics to extract deeper insights and predictive patterns.

Cont..

Features	Business Intelligence (BI)	Big Data Analytics
Data type	Primarily structured data from internal sources	Structured, semi-structured, and unstructured data from internal and external sources
Data volumes	Smaller volumes of data	Large volumes of data
Purpose	Historical data analysis and reporting	Real-time data analysis and predictive modeling
Tools and technologies	Data warehouses, data marts, reporting tools	Hadoop, Spark, cloud-based platforms, machine learning, artificial intelligence
Outcomes	Improved decision-making, operational efficiency	Innovation, competitive advantage

Which one is right for you?

- The best tool for your organization will depend on your specific needs and goals.
  - If you need to answer historical questions and make tactical decisions, then BI is a good choice.
  - If you need to answer predictive questions and make strategic decisions, then big data analytics is a better choice.
- In many cases, organizations will use both BI and big data analytics. BI can be used to get a high-level view of the data, while big data analytics can be used to drill down into the data and find more detailed insights.


END

Quiz

- What are the difference between data analytics, BI, data science, data mining, data engineering?
- Discuss the role of data visualization in business intelligence (BI) applications?
- Explain the significance of data governance in big data analytics?
- Describe the challenges associated with implementing big data analytics projects?
- Discuss the potential impact of artificial intelligence (AI) on the future of BI and big data analytics?
- Explain how organizations can effectively leverage BI and big data analytics to gain a competitive advantage?