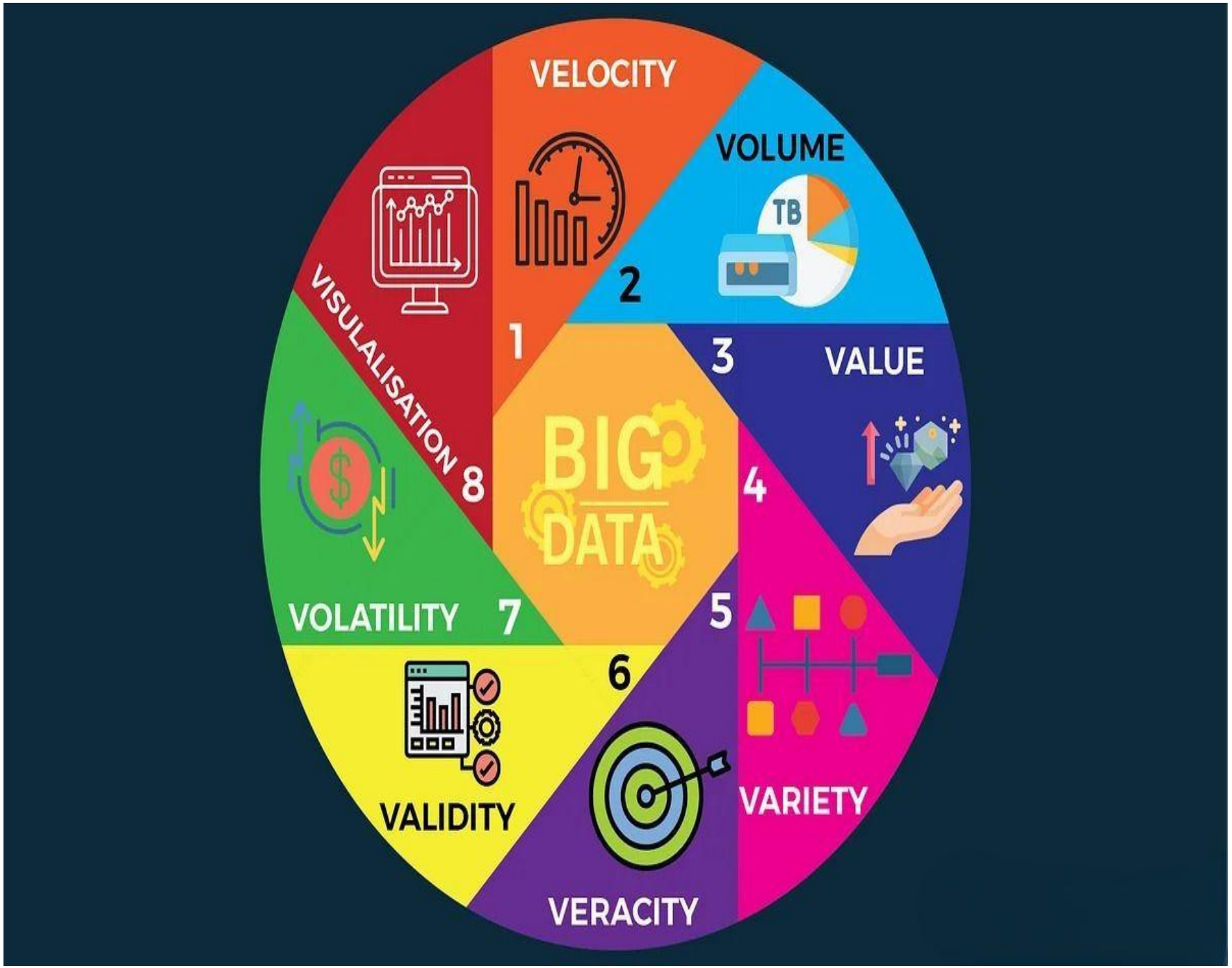


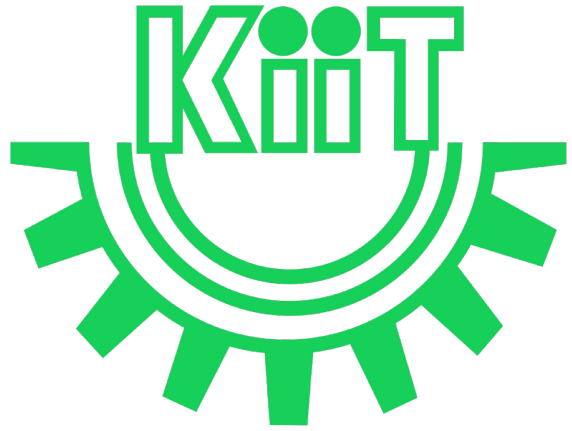
CS 3032: Big Data

Lec-1

Jan 06-17 , 2024

In this Discussion . . .

- Importance of Data
- Characteristics of Data
- Analysis of unstructured data
- Introduction to Big Data
- Challenges of conventional systems
- Elements of Big Data
- Any Questions



Data

- A representation of information, knowledge, facts, concepts or instructions which are being prepared or have been prepared in a formalized manner.
- Data is either intended to be processed, is being processed, or has been processed.
- It can be in any form stored internally in a computer system or computer network or in a person's mind.
- Since the mid-1900s, people have used the word data to mean computer information that is transmitted or stored.

Importance of Data

- In today's era, an integral part of piloting a successful organization involves gathering data that can be analyzed to gain greater insights into the business and its customers.

> Many of the largest companies in the world, such as Amazon, Google, and Netflix, have historically leveraged data for business purposes.

- However, advances in areas such as data processing and data visualization have shown the importance of data and have made reaping the benefits of big data more accessible to everyone.

Importance of Data

- “But what if we don’t have enough data?” **You do.**

> The total data generated is growing exponentially. Some figures suggest we collectively create roughly 2.5 quintillion bytes of data daily

- In other words, data is extremely readily available to business leaders to those willing to extract it. The more concerning issue is distinguishing what data is worth extracting and what is not since so much is being produced.

Importance of Data

- Nonetheless, data is a valuable asset for businesses in the 21st century.

In 2006, *Clive Humby* - a British mathematician - coined the phrase "data is the new oil" about the availability of both resources: neither oil nor data is valuable in its raw state; rather, value is derived when it is gathered rapidly, completely, accurately and is connected to other relevant data.

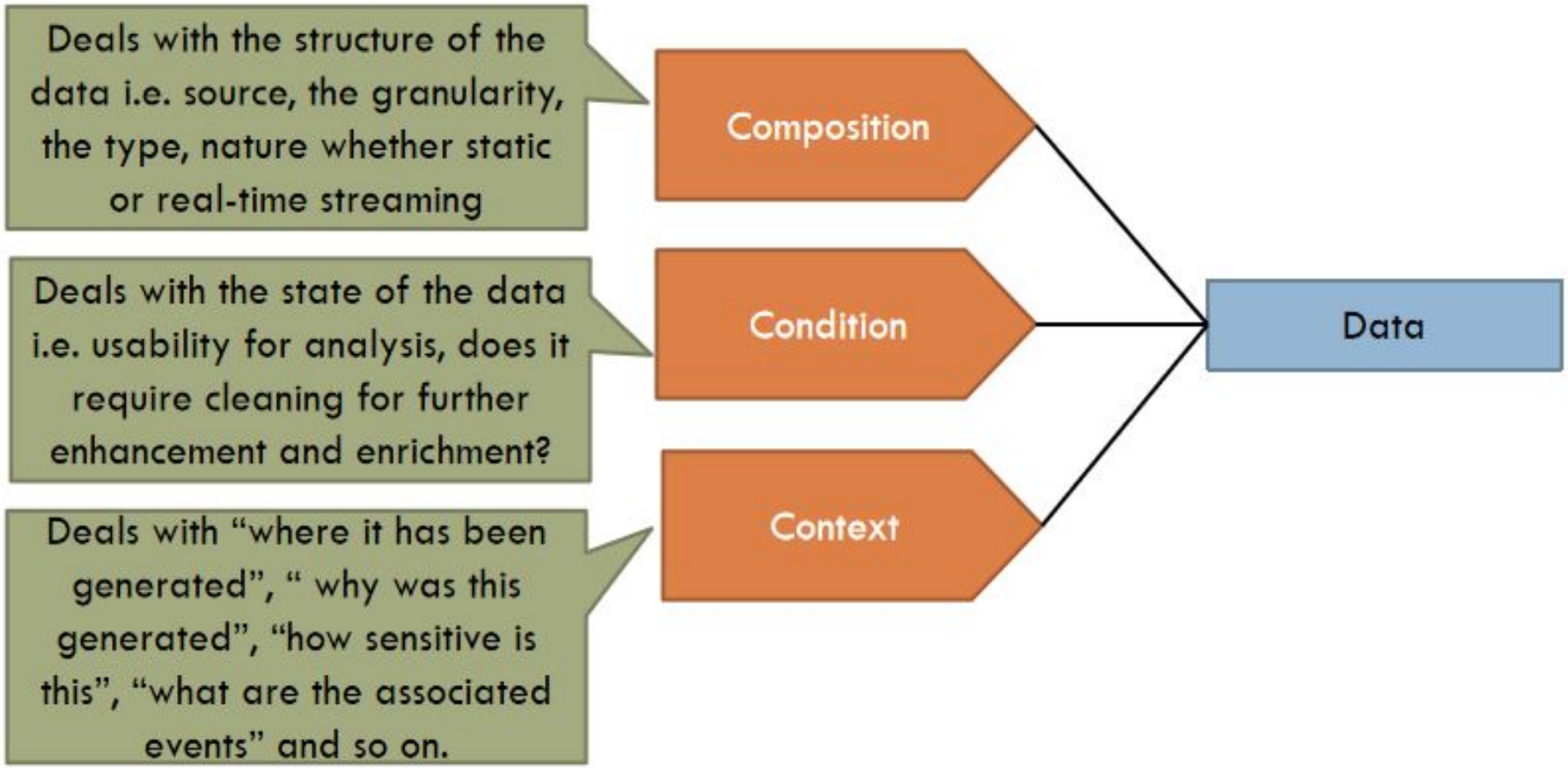
Broad areas significantly increasing Importance of Data

For Informed Decision-Making	• With real-time intelligence at their disposal, leaders can make informed decisions on the direction to take their company. • Using data enables leaders to make less risky decisions based on facts provided by the data.
For Problem-Solving	• Data enables organizations to track and review the health of various business processes and essential systems. • The aid this provides businesses is two-fold: 1) hindsight and 2) foresight. • In hindsight, businesses can review data to uncover stages where performance breakdowns occur. • In foresight, organizations can effectively enforce quality monitoring, enabling them to respond to challenges before they become a major issue.

Broad areas significantly increasing Importance of Data

For Greater Understanding	• Business leaders must understand how each aspect of the business is performing against key targets and goals. • Data is important to get an accurate insight into performance. Ex- sports industry.
For Improving Processes	• Data helps business leaders better understand and improve processes that reduce the number of wasted resources • This knowledge can help businesses get their products into customers' hands faster and cheaper.
For Understanding Behaviour	• At the center of every successful business is a deep passion for understanding and meeting the customer's needs. • If potential and existing customers are to believe a business has their best interest at heart, the business must know, understand, and meet their needs.

Characteristics of Data



Human Vs. Machine Readable Data

Machine Readable	Human Readable
• Data in a data format that can be automatically read and processed by a computer, such as CSV, JSON, XML, etc. • Machine-readable data must be structured data.	• Data in a format that can be conveniently read by a human. • Some human-readable formats, such as PDF, are not machine-readable as they are not structured data, i.e. the representation of the data on disk does not represent the actual relationships present in the data.

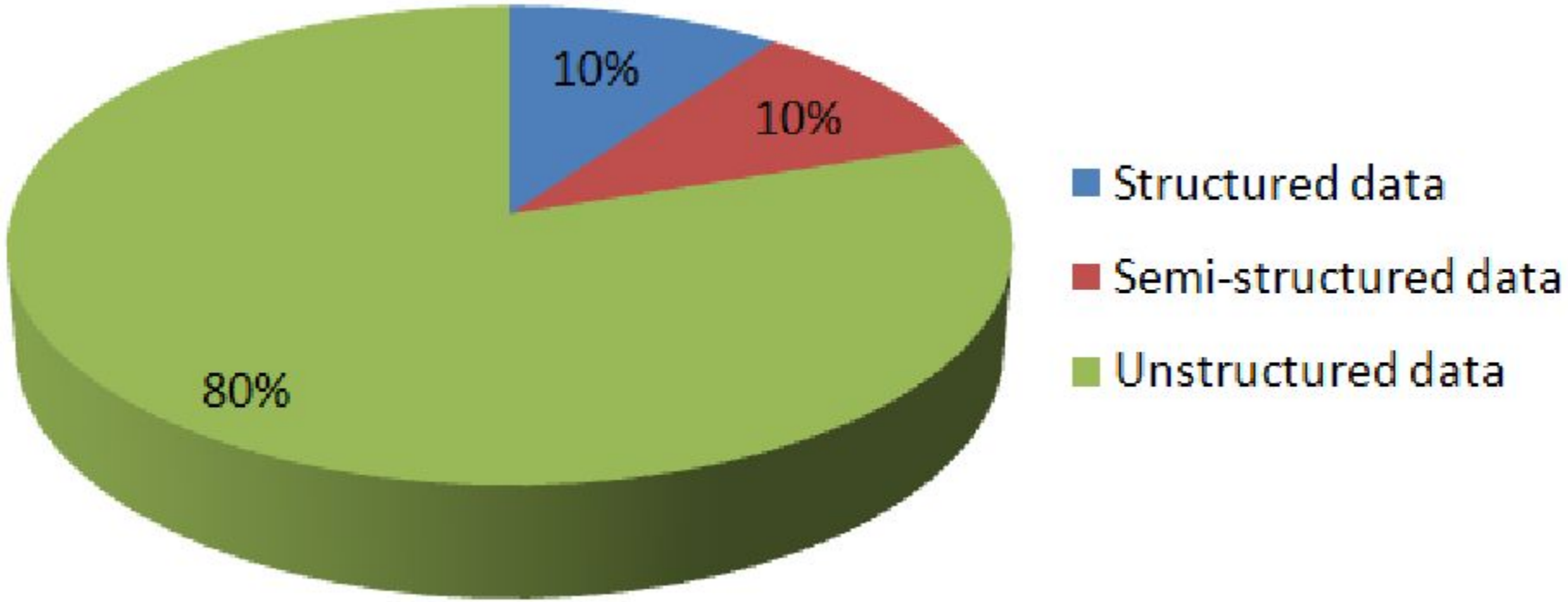
Classification of Digital Data

- Digital data is classified into the following categories:

  i) Structured data

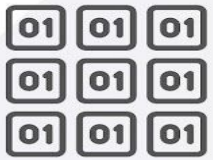
  ii) Semi-structured data,

  and iii) Unstructured data



Classification of Digital Data

Structured data



Characteristics

Predefined data models
Easy to search
Text-based
Shows what's happening

Resides in

Relational databases
Data warehouses

Stored in

Rows and columns

Examples

Dates, phone numbers, social security numbers, customer names, transaction info

Unstructured data

Characteristics

No predefined data models
Difficult to search
Text, pdf, images, video
Shows the why

Resides in

Applications
Data warehouses and lakes

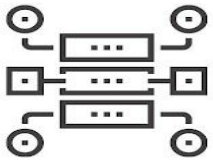
Stored in

Various forms

Examples

Documents, emails and messages, conversation transcripts, image files, open-ended survey answers

Semi-structured data



Characteristics

Loosely organized
Meta-level structure that can contain unstructured data
HTML, XML, JSON

Resides in

Relational databases
Tagged-text format

Stored in

Abstracts & figures

Examples

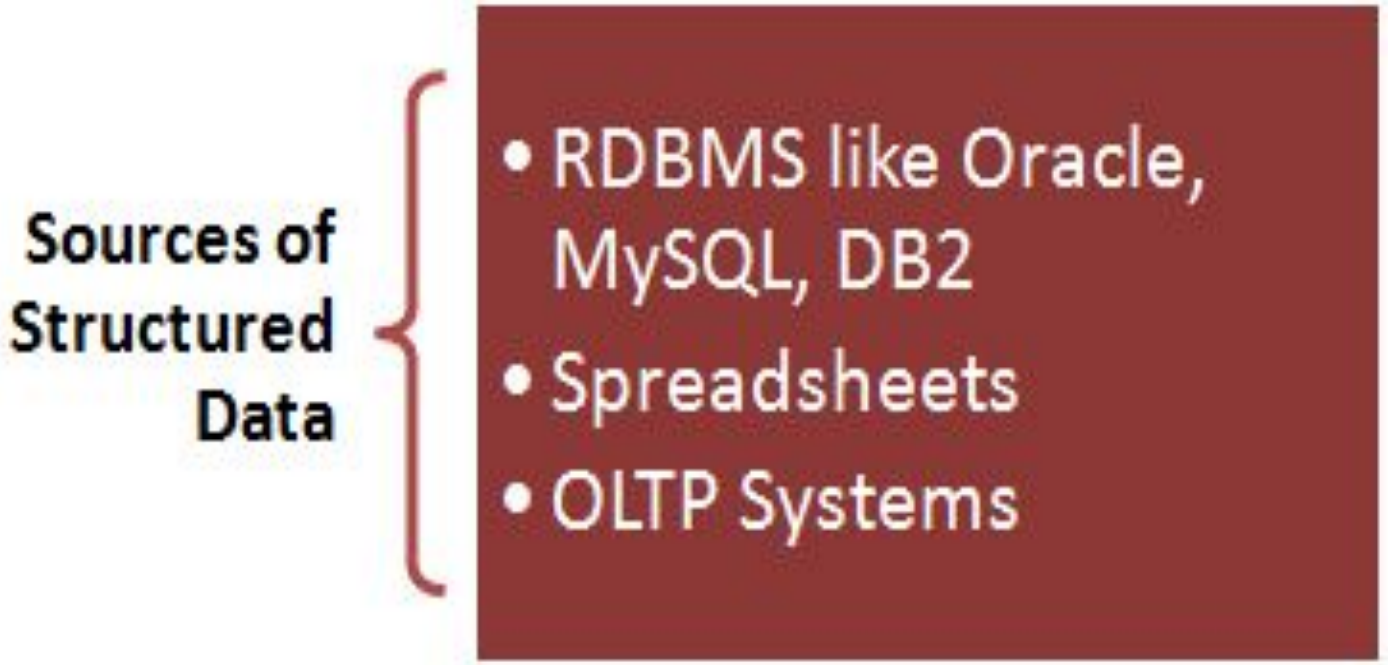
Server logs, tweets organized by hashtags, emails sorting by folders (inbox; sent; draft)

Structured Data

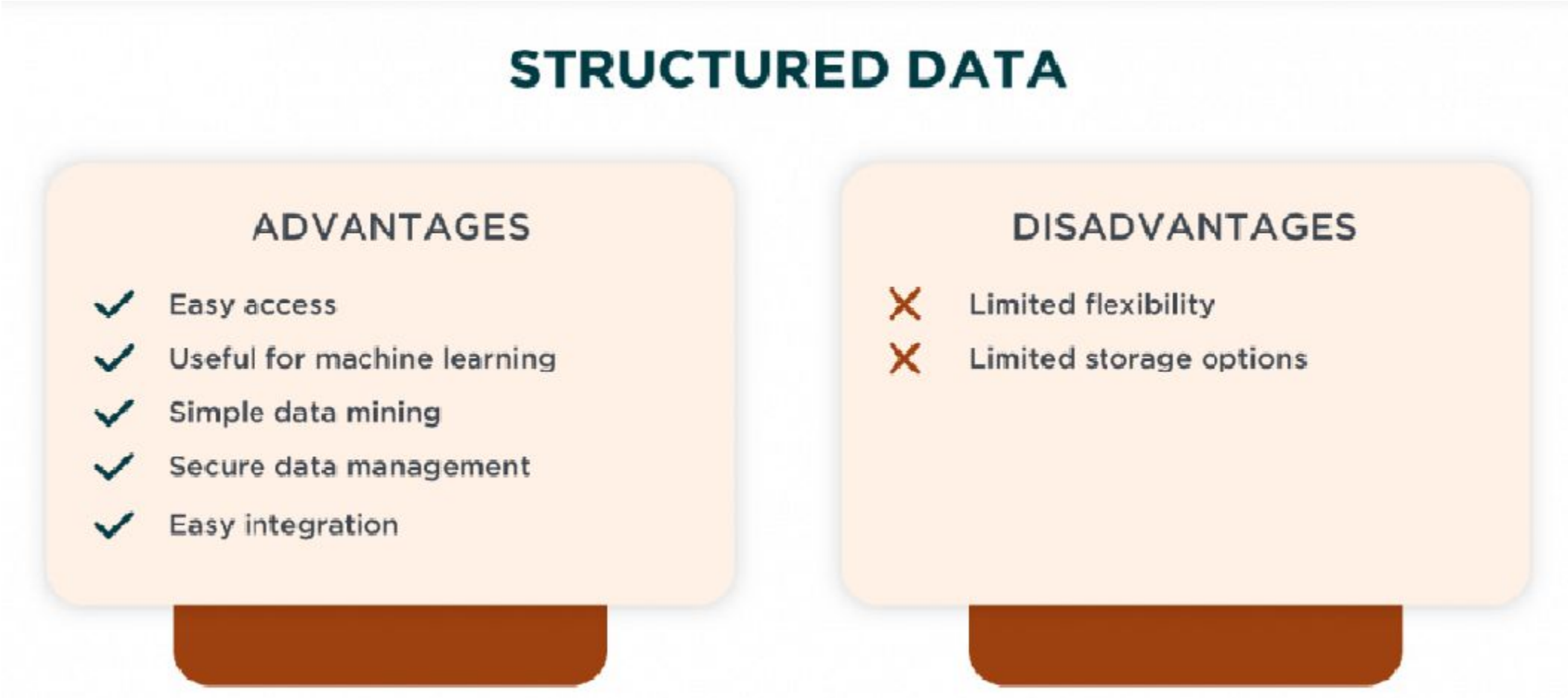
- Structured data types, also known as relational data types, can be managed and searched in a relational database or through a relational database management system (RDBMS). This includes integers, decimals, dates, time, strings, and Booleans.
- These data types are easily arranged in rows and columns and can be queried using programming languages like SQL to return relevant search information.

The **benefit** of structured data is its **labelling** to **describe** its **attributes and relationships with other data**. This data structure is easily searchable using a human or algorithmically generated query.

Major Sources of Structured Data



Structured Data: Advantages and Disadvantages



Semi-Structured Data

- When data deviates a little from a formal structure but is still relevant in its context, it is called **semi-structured data**.

  - ❏ Unlike relational databases or other types of data tables, semi-structured data does not adhere to the tabular structure of structured data.
  - ❏ Nevertheless, it includes tags or markers to segregate semantic pieces and impose hierarchies of records and fields within the data.
  - ❏ As a result, it is also known as a self-descriptive structure.

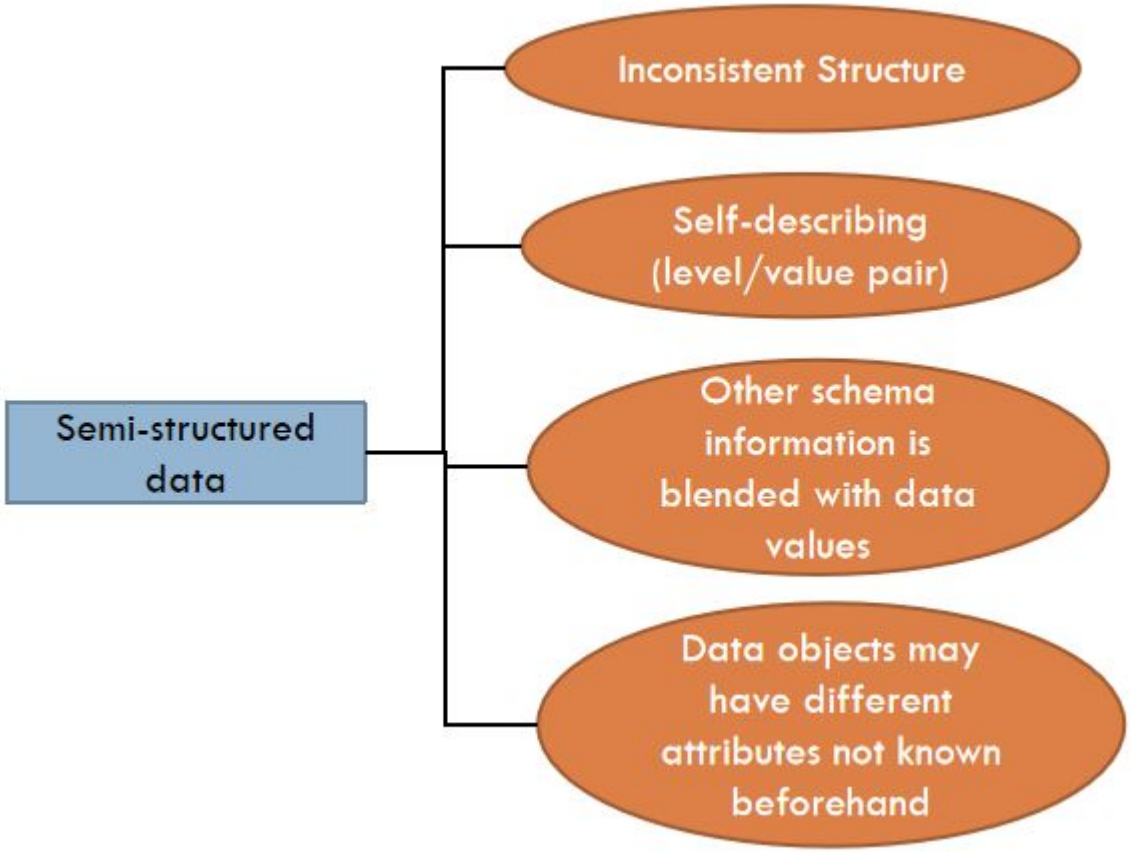
- Semi-structured data is recognized by relational databases with the help of identifiers (tags) to place the data under hierarchical categories. Only then can it be analyzed with the help of data management tools.

Semi-structured data Examples

- No one can say that semi-structured data will never fit a preset data model or schema. Some examples of semi-structured data are:

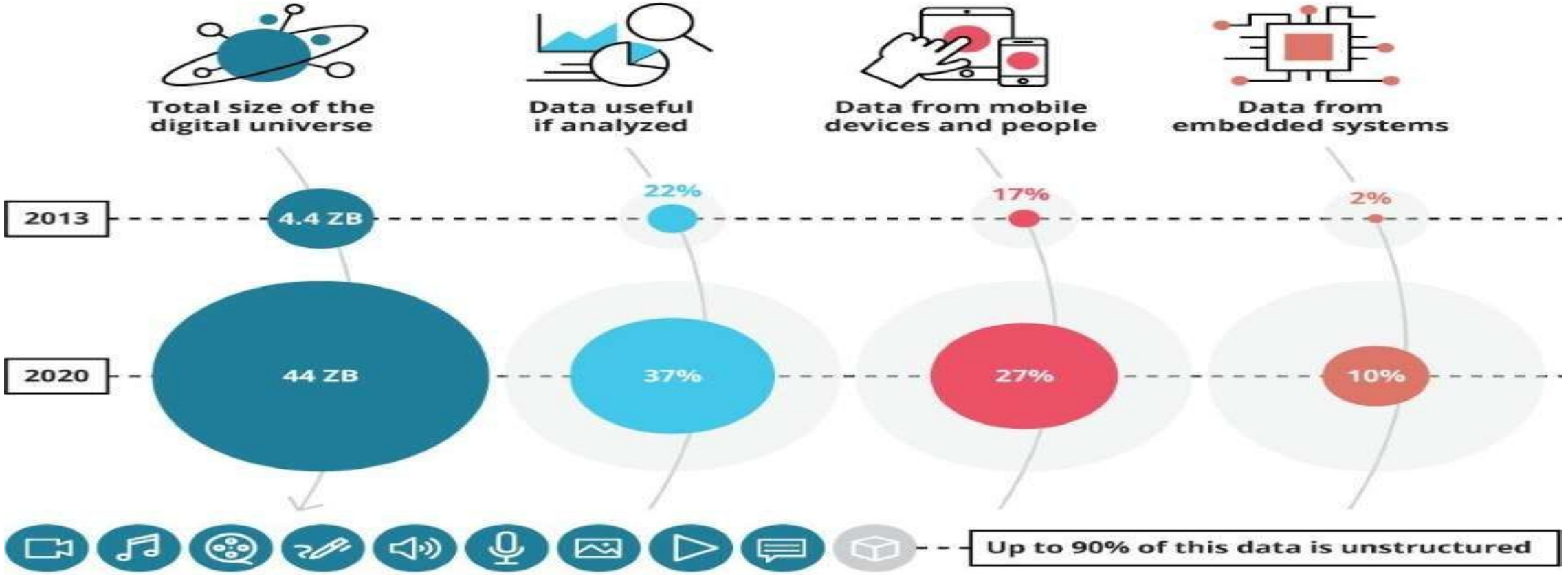
  - Email
  - NoSQL databases
  - CSV, XML, and JSON documents:
  - Electronic data interchange (EDI)
  - HTML
  - RDF

Characteristics of Semi-Structured Data



Unstructured Data

- **Unstructured data is any data type that cannot be stored in a relational database or RDBMS.** Data in the form of images, videos, audio, webpages, free text, and even social media content fall into this category.
- This type of data is too complex to be parsed and interpreted. Of course, this data can still be stored, retrieved, and edited by other means.



Unstructured Data

- Unstructured data:
  - has an **internal structure** (i.e. bits and bytes)
  - but is **not structured via pre-defined data models or schema**, i.e. not organised and labelled to identify meaningful relationships between data
- It may be textual / non-textual. It may be human / machine-generated. It might also be stored within a non-relational database like NoSQL.

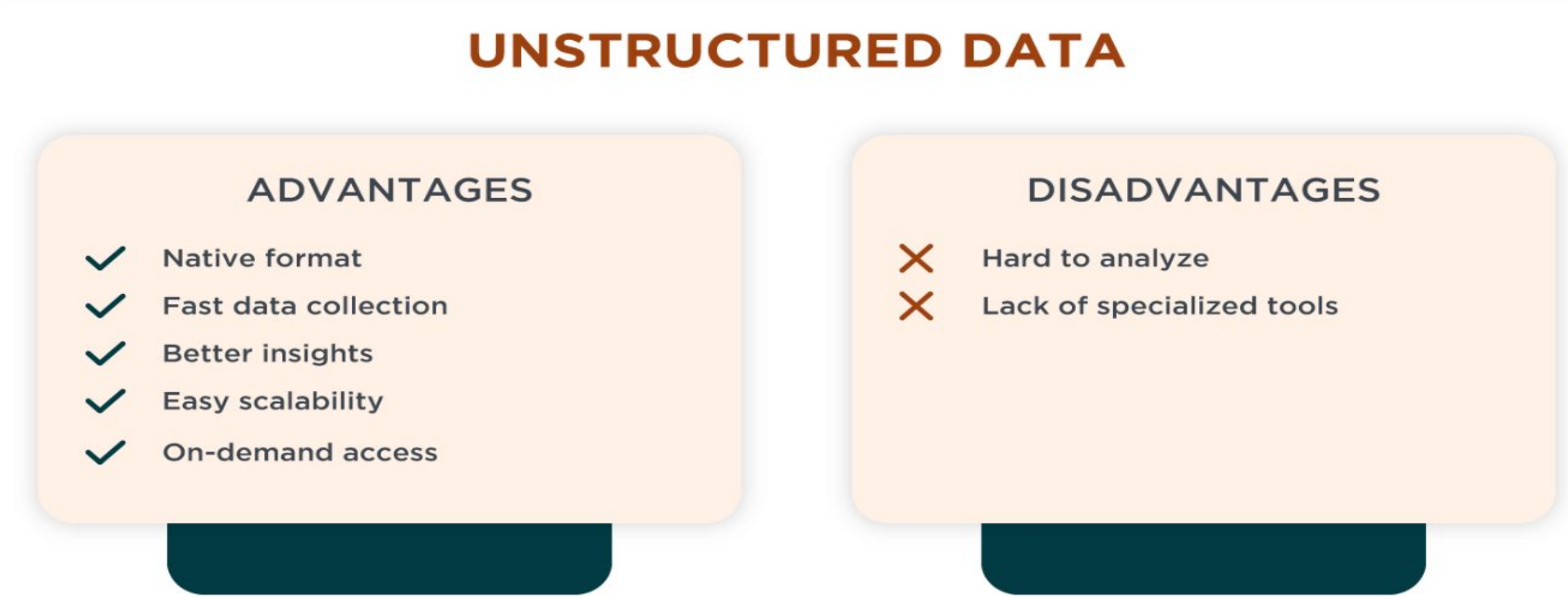
Unstructured Data

Human Generated	■ **Text files:** word processing files, spreadsheets, presentations, emails. ■ **Email:** largely text, but has some internal structure thanks to its metadata (e.g. including the visible “to”, “from”, “date / time”, “subject” entered to send an email) but also mixes in unstructured data via the message body. For this reason, email is also referred to as **semi-structured data**. ■ **Social Media:** like email, this is often **semi-structured data**, containing unstructured data (e.g. a Tweet) but also structured data (e.g. the number of “Likes”, “retweets”, “date”, “author” etc). ■ **Websites:** YouTube, Instagram etc contain lots of unstructured data, but also much structured data, e.g. like described above for Twitter ■ **Mobile data:** text messages, locations. ■ **Communications:** IMs, dictaphone recordings. ■ **Media:** MP3, digital photos, audio recordings and video files. ■ **Business applications:** MS Office documents, PDFs and similar.

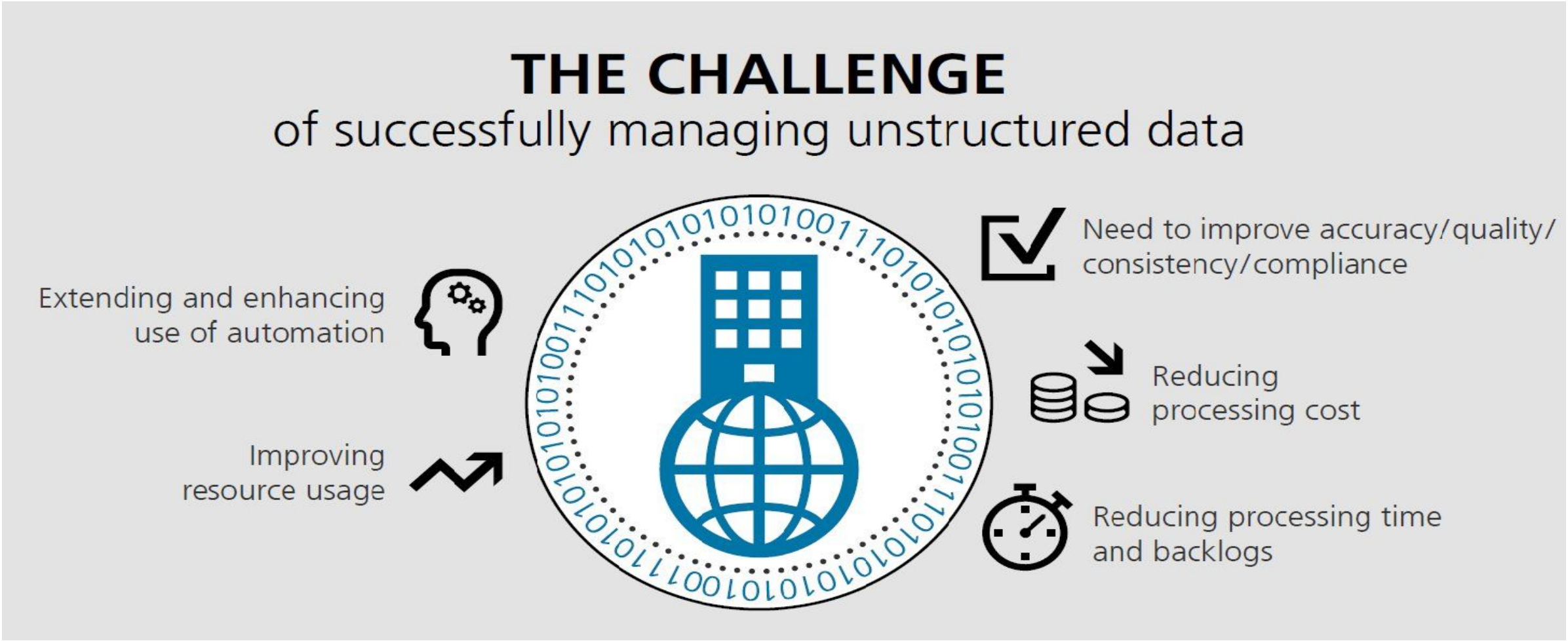
Unstructured Data

Machine Generated	■ **Satellite imagery:** weather data, geographic forms, military movements. ■ **Scientific data:** oil and gas exploration, space exploration, seismic imagery and atmospheric data. ■ **Digital surveillance:** CCTV.

Unstructured Data: Advantages and Disadvantages

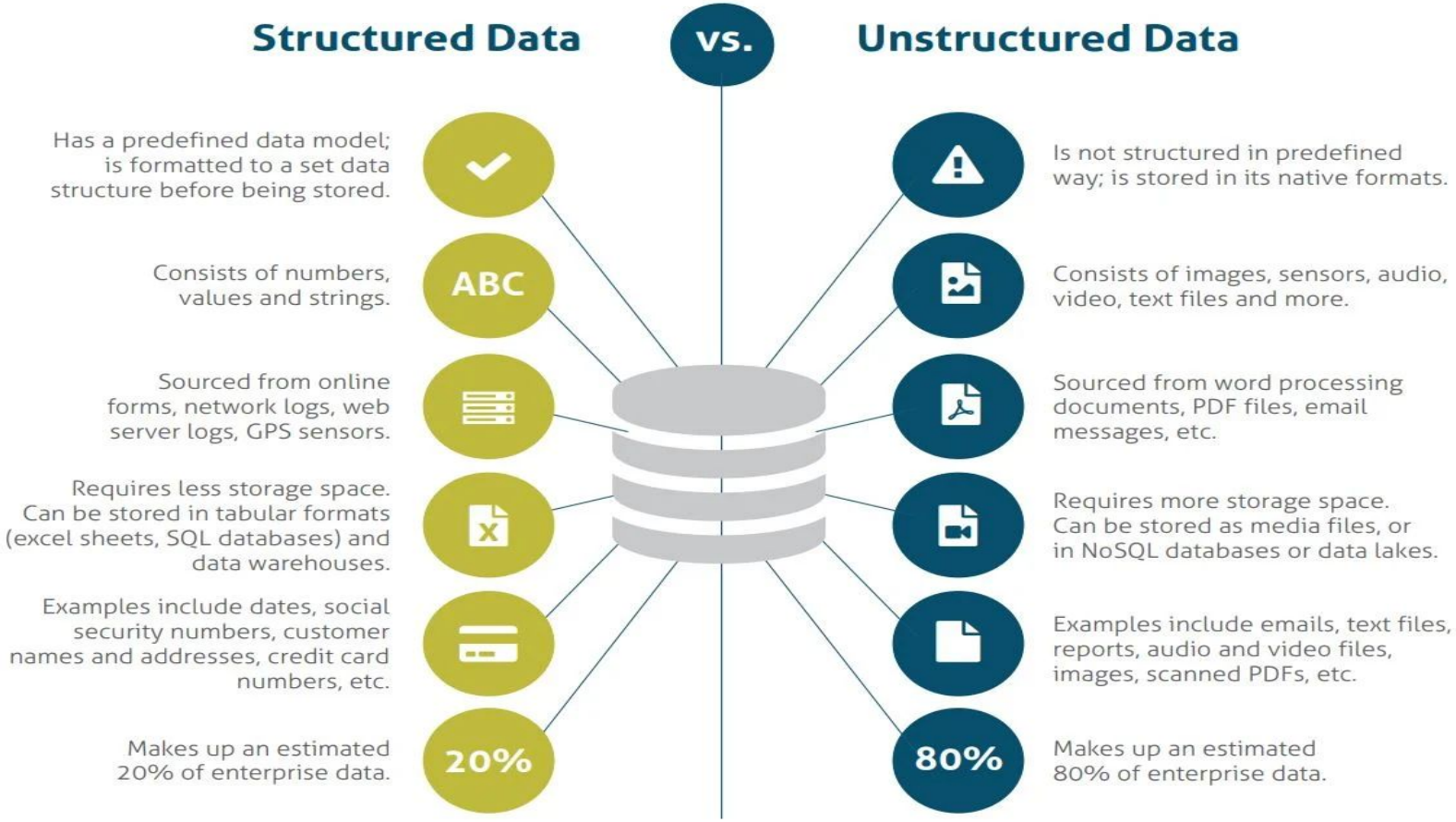


Challenges associated with Unstructured Data

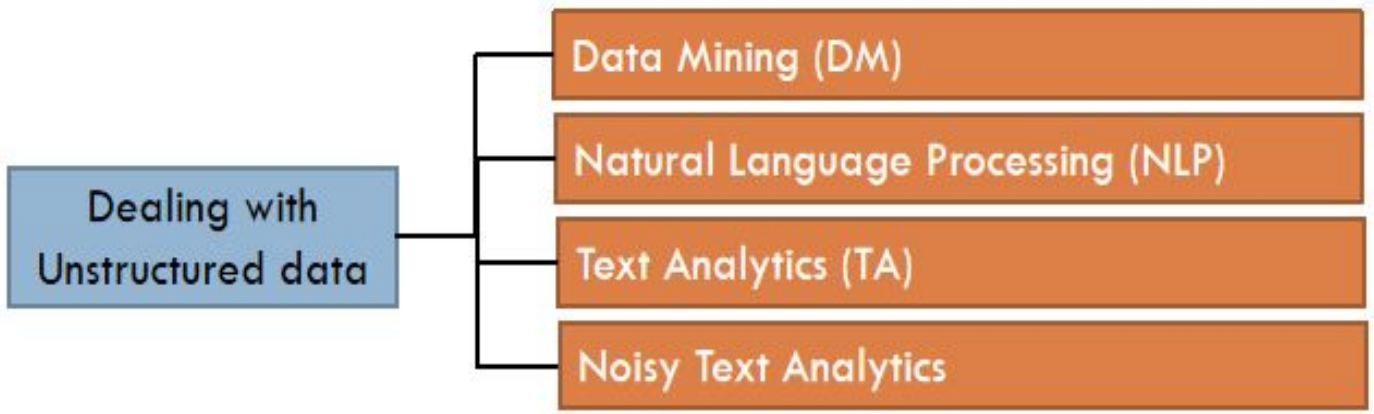


Challenges associated with Unstructured Data

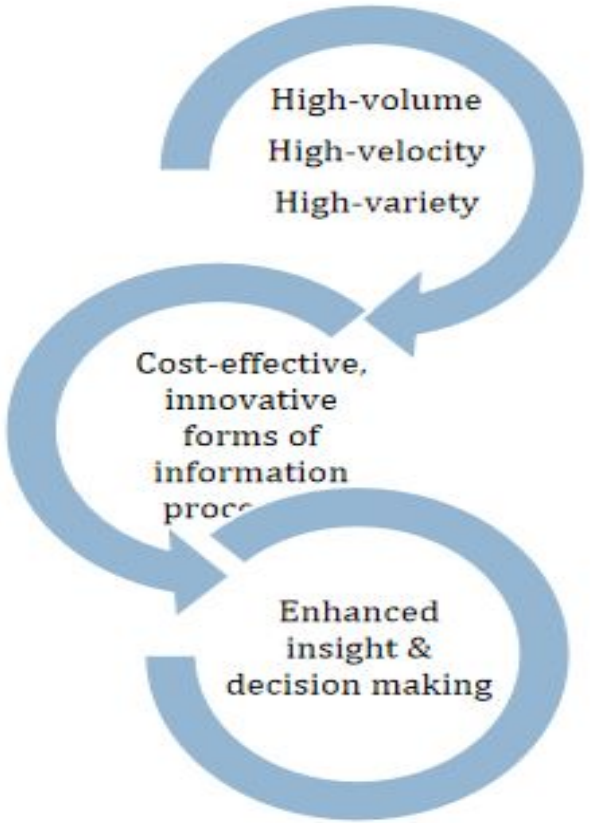
Structured Vs. Unstructured Data



Dealing with Unstructured Data

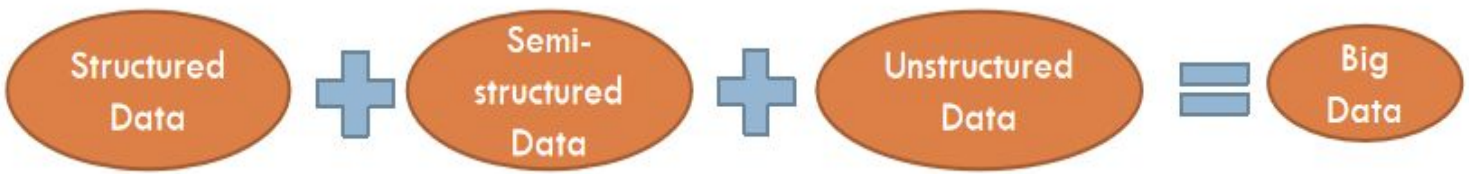


Big Data



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.

Source: Gartner IT Glossary



Salient Features of Big Data

Challenges of Conventional Systems

- The main challenge in the traditional approach for computing systems to manage 'Big Data' because of immense speed and volume at which it is generated. Some of the challenges are:
  - Traditional approach cannot work on unstructured data efficiently
  - Traditional approach is built on top of the relational data model, relationships between the subjects of interests have been created inside the system and the analysis is done based on them. This approach will not adequate for big data.

Challenges of Conventional Systems

- The main challenge in the traditional approach for computing systems to manage "Big Data" because of immense speed and volume at which it is generated. Some of the challenges are:
  - **Traditional approach is batch oriented** and need to wait for nightly ETL (extract, transform and load) and transformation jobs to complete before the required insight is obtained
  - Traditional data management, warehousing, and analysis systems fizzle to analyze this type of data. Due to its complexity, big data is processed with parallelism. **Parallelism in a traditional system is achieved through costly hardware like MPP (Massively Parallel Processing) systems**

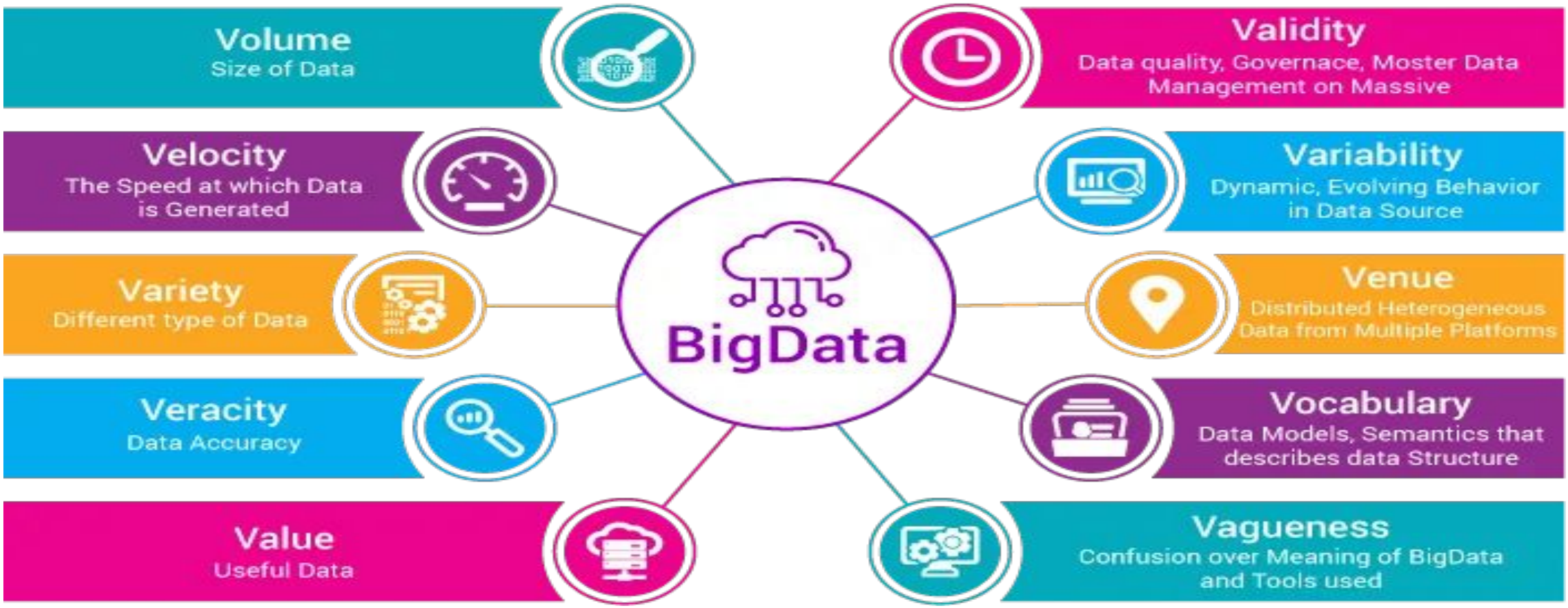
Challenges of Conventional Systems

- The main challenge in the traditional approach for computing systems to manage “Big Data” because of immense speed and volume at which it is generated. Some of the challenges are:
  - **Inadequate support of aggregated summaries of data**
  - **Data Challenges:**
    - **Volume, velocity, veracity, variety**
    - **Data discovery and comprehensiveness**
    - **Scalability**

Challenges of Conventional Systems

- The main challenge in the traditional approach for computing systems to manage “Big Data” because of immense speed and volume at which it is generated. Some of the challenges are:
  - **Process challenges**
    - Capturing Data
    - Aligning data from different sources
    - Transforming data into suitable form for data analysis
    - Modeling data(Mathematically, simulation)
  - **Management Challenges:**
    - Security
    - Privacy
    - Governance
    - Ethical issues

Elements of Big Data: V's of Big Data



Any Questions ??

References

1. https://www.datacamp.com/blog/importance-of-data-5-top-reasons
2. https://opendatahandbook.org/glossary/en/terms/machine-readable/
3. https://opendatahandbook.org/glossary/en/terms/human-readable/
4. https://levity.ai/blog/structured-vs-unstructured-data
5. https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/
6. https://jelvix.com/blog/structured-vs-unstructured-data
7. https://www.researchgate.net/figure/Sources-of-Structured-Data_fig3_322406625
8. https://www.simplilearn.com/semi-structured-data-article
9. https://www.loginworks.com/blogs/10-effective-ways-to-deal-with-structured-and-semi-structured-data/
10. https://twitter.com/SPSGlobal/status/937577253432684544/photo/1
11. https://slideplayer.com/slide/13495893/
12. https://blog.collabware.com/what-is-unstructured-data-intelligence-why-is-it-becoming-essential
13. https://www.analyticssteps.com/blogs/top-10-big-data-technologies-2020

References

14. https://medium.com/digital-transformation-and-platform-engineering/data-ingestion-processing-and-big-data-architecture-layers-3cb4988c07de
15. https://twitter.com/deadpoolmovie/status/718576599243751424/photo/1
16.