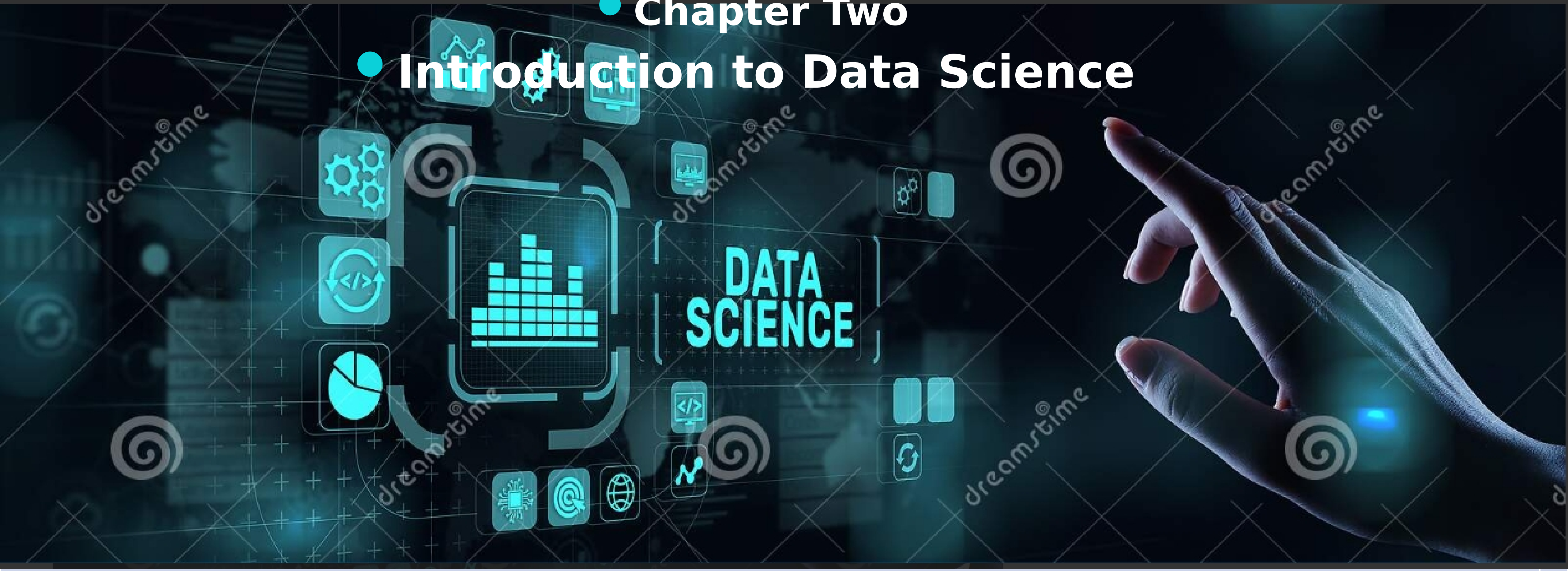
Introduction to Emerging Technology

Adama Science and Technology University

- **Chapter Two**
- **Introduction to Data Science**



Chapter : 2 Introduction to Data Science

Topics Covered

- Overview of Data Science
  - Definition of data and information
  - Data types and representation
- Data Value Chain
  - Data Acquisition
  - Data Analysis
  - Data Curating
  - Data Storage
  - Data Usage
- Basic concepts of Big data

Overview of Data Science

- **Data science** is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from **structured, semi structured and unstructured data.**
- **Data science enables** businesses to process huge amounts of structured and unstructured big data to detect patterns. This in turn allows companies to increase efficiencies, manage costs, identify new market opportunities, and boost their market advantage.
- **Data science** continues to evolve as one of the most promising and in-demand career paths for skilled professionals.
- Today, successful data professionals understand that they must advance past traditional skills of analyzing large amounts of data, data mining, and programming skills.

Overview of data science contd…

- **Data Science is** about data gathering, analysis and decision-making.
- **Data Science is** about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

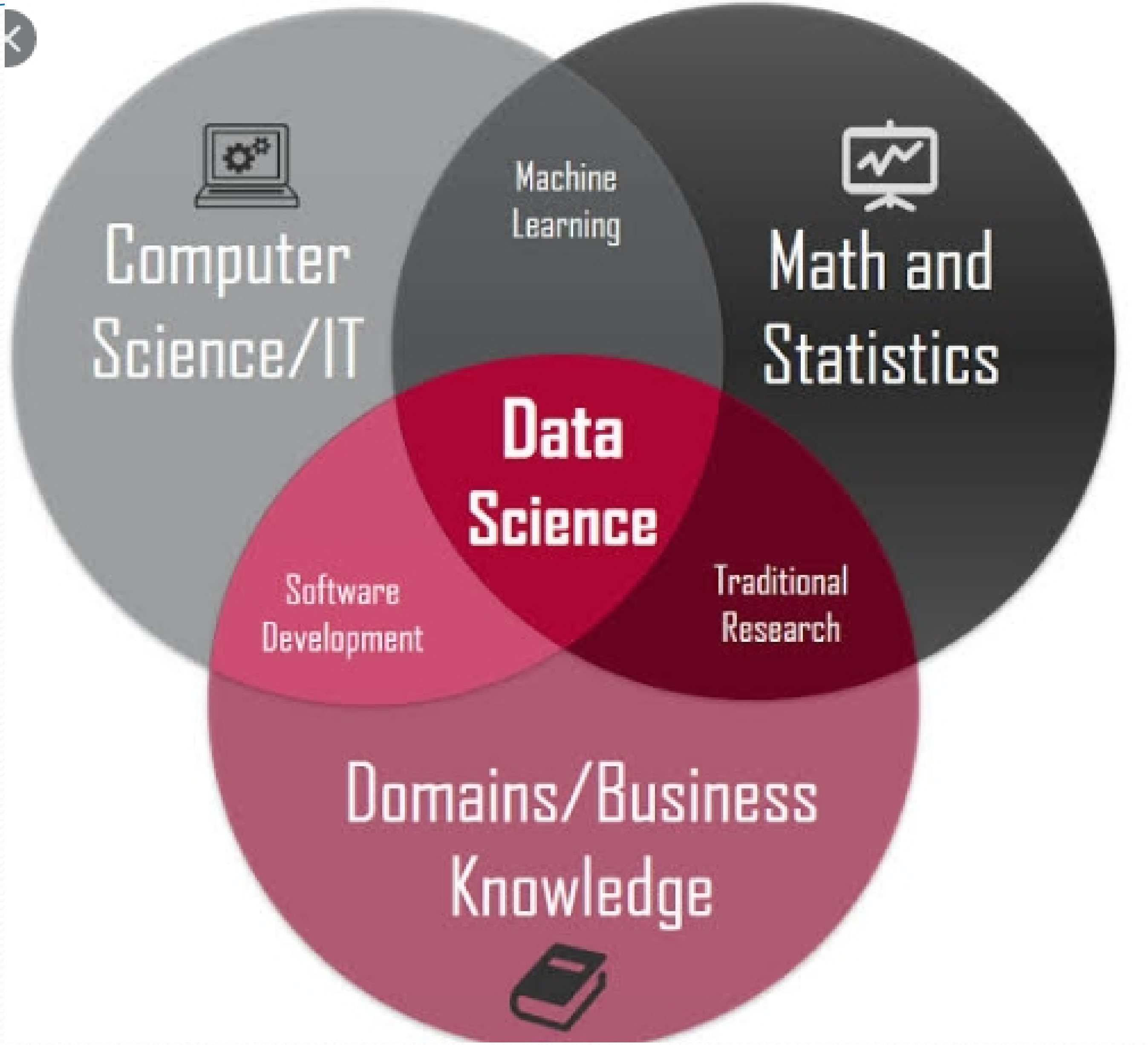
- **Better decisions (should we choose A or B)**
- **Predictive analysis (what will happen next?)**
- **Pattern discoveries (find pattern, or maybe hidden information in the data)**

Where is Data Science Needed?

Data Science is used in many industries in the world today, **e.g. banking, consultancy, healthcare, and manufacturing.**

Realtime Examples: For route planning: To discover the best routes to ship,

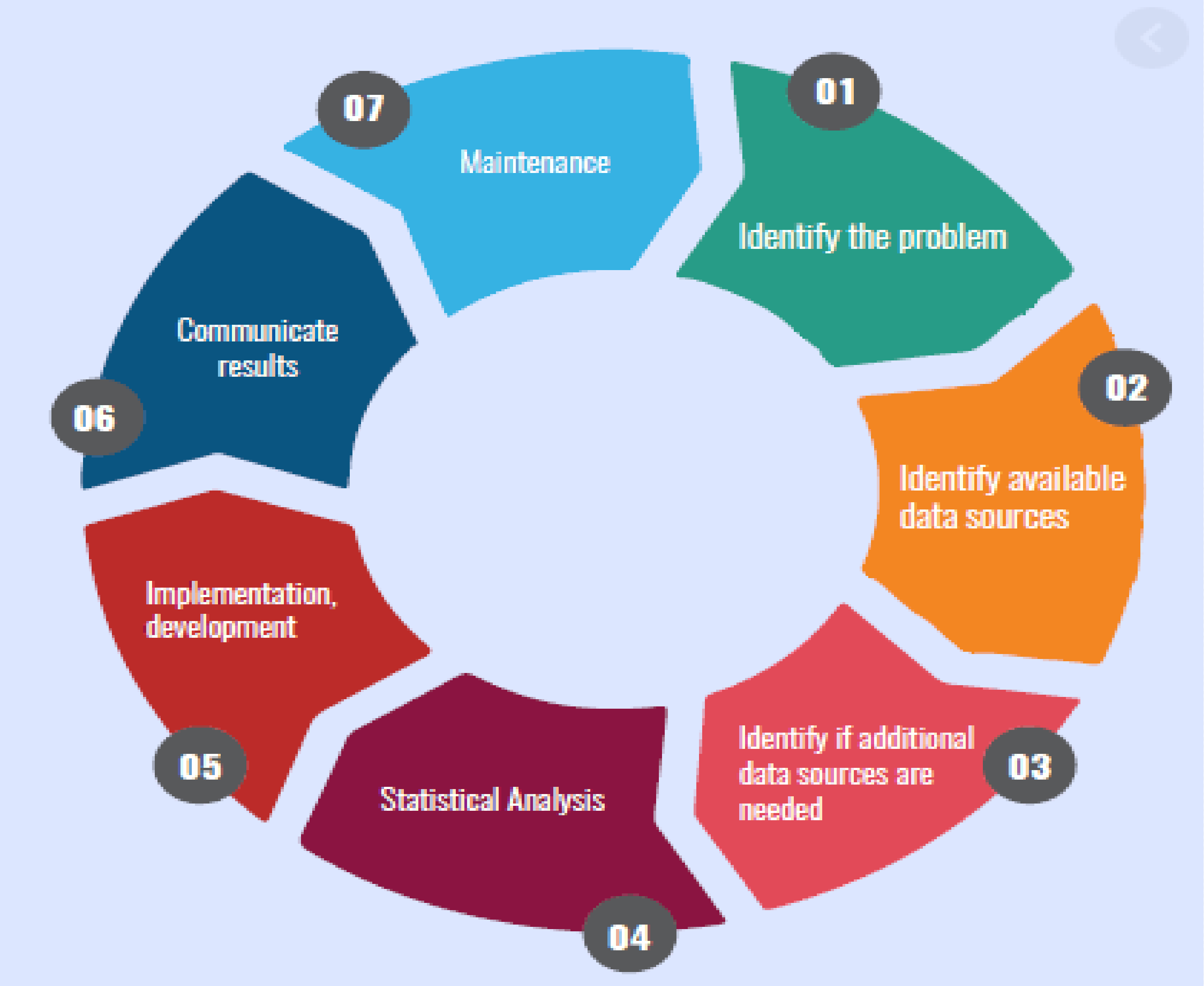
To foresee delays for flight/ship/train etc. (through predictive analysis),To create promotional offers,To find the best suited time to deliver goods,To forecast the next years revenue for a company,To analyze health benefit of training,To predict who will win elections


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

What is expected of a data scientist?

- In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.
- Data scientists need to be curious and result-oriented, with exceptional industry-specific knowledge and communication skills that allow them to explain highly technical results to their non-technical counterparts.
- Data science need a strong quantitative background in statistics and linear algebra as well as programming knowledge with focuses in data warehousing, mining, and modeling to build and analyze algorithms.
- This chapter cover basic definitions of data and information, data types and representation, data value change and basic concepts of big data.

Data Science Life cycle

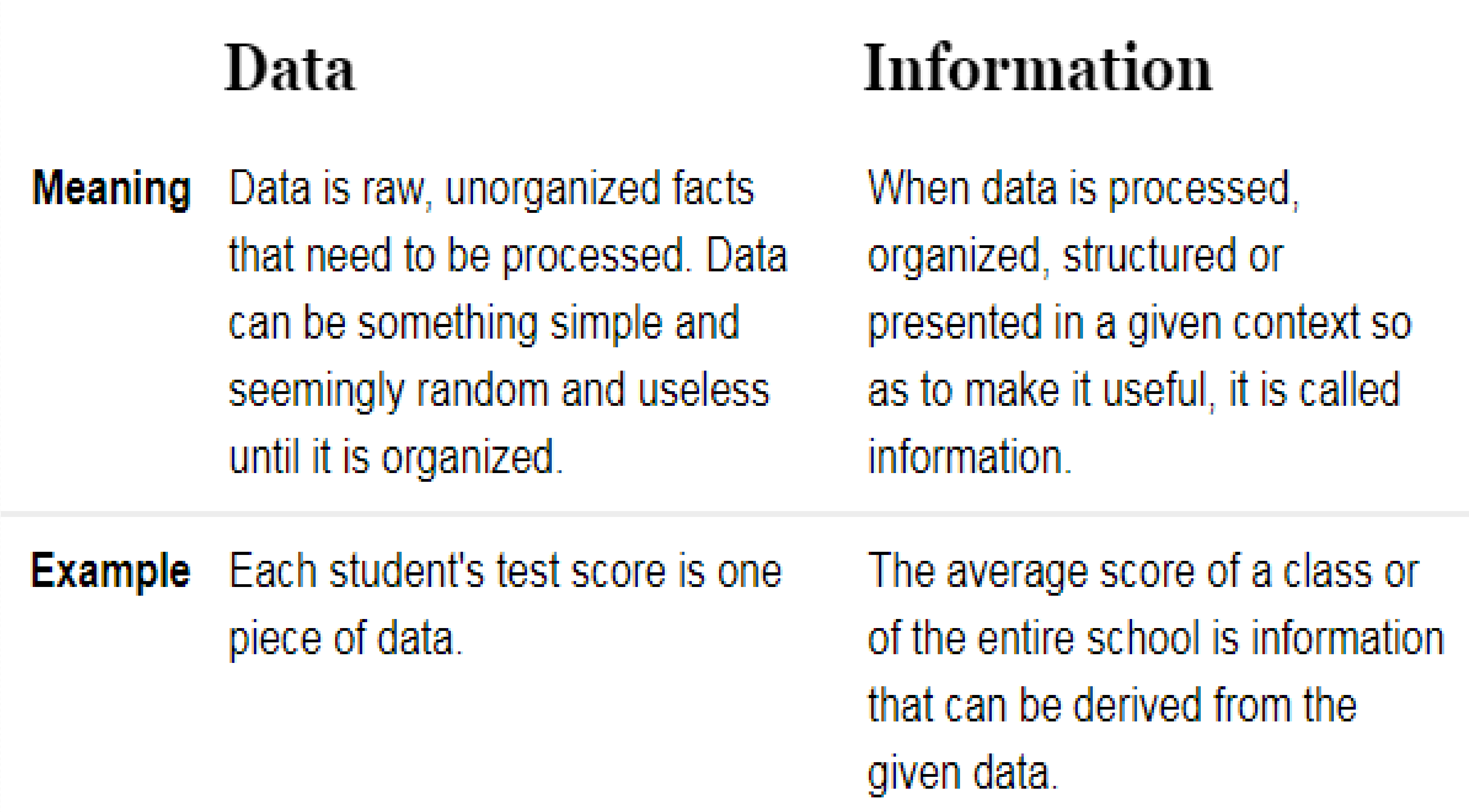

01
Identify the problem
02
Identify available data sources
03
Identify if additional data sources are needed
04
Statistical Analysis
05
Implementation, development
06
Communicate results
07
Maintenance

What is data?

- Data can be defined as a representation of facts, concepts, or instructions in a formalized manner, which should be suitable for communication, interpretation, or processing by human or electronic machine.
- Data is represented with the help of characters such as alphabets (A-Z, a-z), digits (0-9) or special characters (+,-,/,*,<,>,= etc.)

What is Information?

- Information is organized or classified data, which has some meaningful values for the receiver. Information is the processed data on which decisions and actions are based.
- Information is a data that has been processed into a form that is meaningful to recipient and is of real or perceived value in the current or the prospective action or decision of recipient.
- For the decision to be meaningful, the processed data must qualify for the following characteristics −
  - Timely − Information should be available when required.
  - Accuracy − Information should be accurate.
  - Completeness − Information should be complete.



	Data	Information
Meaning	Data is raw, unorganized facts that need to be processed. Data can be something simple and seemingly random and useless until it is organized.	When data is processed, organized, structured or presented in a given context so as to make it useful, it is called information.
Example	Each student's test score is one piece of data.	The average score of a class or of the entire school is information that can be derived from the given data.

Summery: Data Vs. Information

Data	Information
Described as unprocessed or raw facts and figures	Described as processed data
Cannot help in decision making	Can help in decision making
Raw material that can be organized, structured, and interpreted to create useful information systems.	Interpreted data; created from organized, structured, and processed data in a particular context.
‘groups of non-random’ symbols in the form of text, images, and voice representing quantities, action and objects'.	Processed data in the form of text, images, and voice representing quantities, action and objects'.

Data Processing Cycle

- Data processing is the re-structuring or re-ordering of data by people or machine to increase their usefulness and add values for a particular purpose.
- Data processing consists of the following basic steps - input, processing, and output. These three steps constitute the data processing cycle.
- **Input step** − the input data is prepared in some convenient form for processing.
- The form depends on the processing machine.
- For example - when electronic computers are used – input medium options include magnetic disks, tapes, and so on.
- **Processing step** − the input data is changed to produce data in a more useful form.
- For example - pay-checks can be calculated from the time cards, or a summary of sales for the month can be calculated from the sales orders.
- **Output step** − the result of the proceeding processing step is collected.
- The particular form of the output data depends on the use of the data.
- For example - output data may be pay-checks for employees.

2.1.2 Data types and its representation – based on programming language

- Data type or simply type is an attribute of data which tells the compiler or interpreter how the programmer intends to use the data.
- Almost all programming languages explicitly include the notion of data type. Common data types include:
  - Integers
  - Booleans
  - Characters
  - floating-point numbers
  - alphanumeric strings
- A data type constrains the values that an expression, such as a variable or a function, might take.
- This data type defines the operations that can be done on the data, the meaning of the data, and the way values of that type can be stored.
- On other hand, for the analysis of data, there are three common types of data types or structures: Structured data, unstructured data, and semi-structured data.

Data types/structure – based on analysis of data

- **Structured Data, unstructured data, semi-structured data, and metadata**

Structured Data

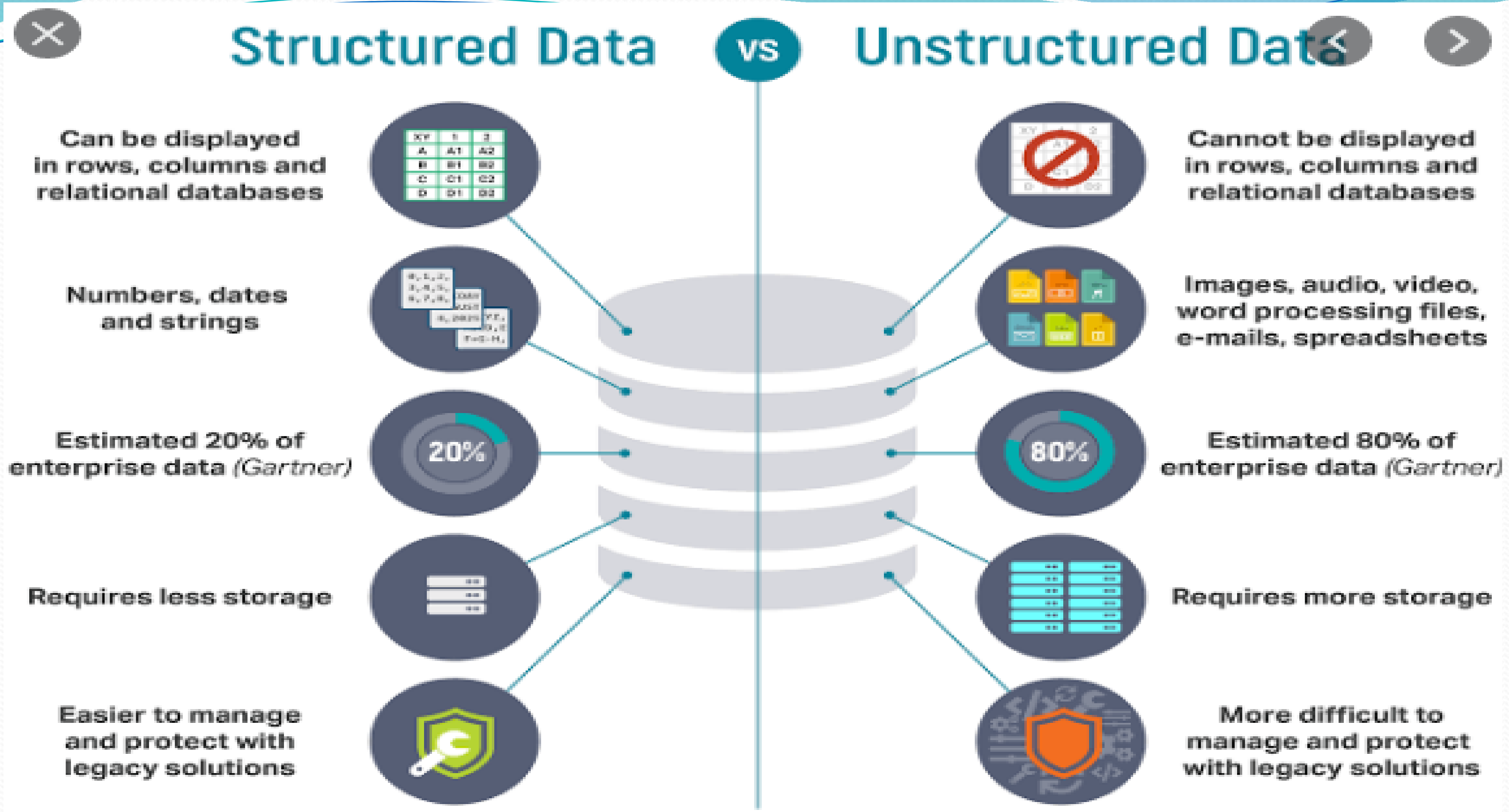
- Structured data is data that adheres to a pre-defined data model and is therefore straightforward to analyze.
- Structured data conforms to a tabular format with relationship between the different rows and columns. Common examples are Excel files or SQL databases.
- Each of these have structured rows and columns that can be sorted.
- Structured data depends on the existence of a data model – a model of how data can be stored, processed and accessed.
- Because of a data model, each field is discrete and can be accesses separately or jointly along with data from other fields.
- This makes structured data extremely powerful: it is possible to quickly aggregate data from various locations in the database.
- Structured data is considered the most ‘traditional’ form of data storage, since the earliest versions of database management systems (DBMS) were able to store, process and access structured data.

Unstructured Data

- Unstructured data is information that either does not have a predefined data model or is not organized in a pre-defined manner.
- It is without proper formatting and alignment
- Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well.
- This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in structured databases.
  - Common examples include: audio, video files or No-SQL databases.
- The ability to store and process unstructured data has greatly grown in recent years, with many new technologies and tools coming to the market that are able to store specialized types of unstructured data. For example:
  - MongoDB is optimized to store documents.
  - Apache Graph - is optimized for storing relationships between nodes.
- The ability to analyze unstructured data is especially relevant in the context of Big Data, since a large part of data in organizations is unstructured. Think about pictures, videos or PDF documents.
- The ability to extract value from unstructured data is one of main drivers behind the quick growth of Big Data.

Semi-structured Data

- Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables,
- but nonetheless contain tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure.
- Fore example: JSON and XML are forms of semi-structured data.
- The reason that this third category exists (between structured and unstructured data) is because semi-structured data is considerably easier to analyze than unstructured data.
- Many Big Data solutions and tools have the ability to ‘read’ and process either JSON or XML. This reduces the complexity to analyze structured data, compared to unstructured data.



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

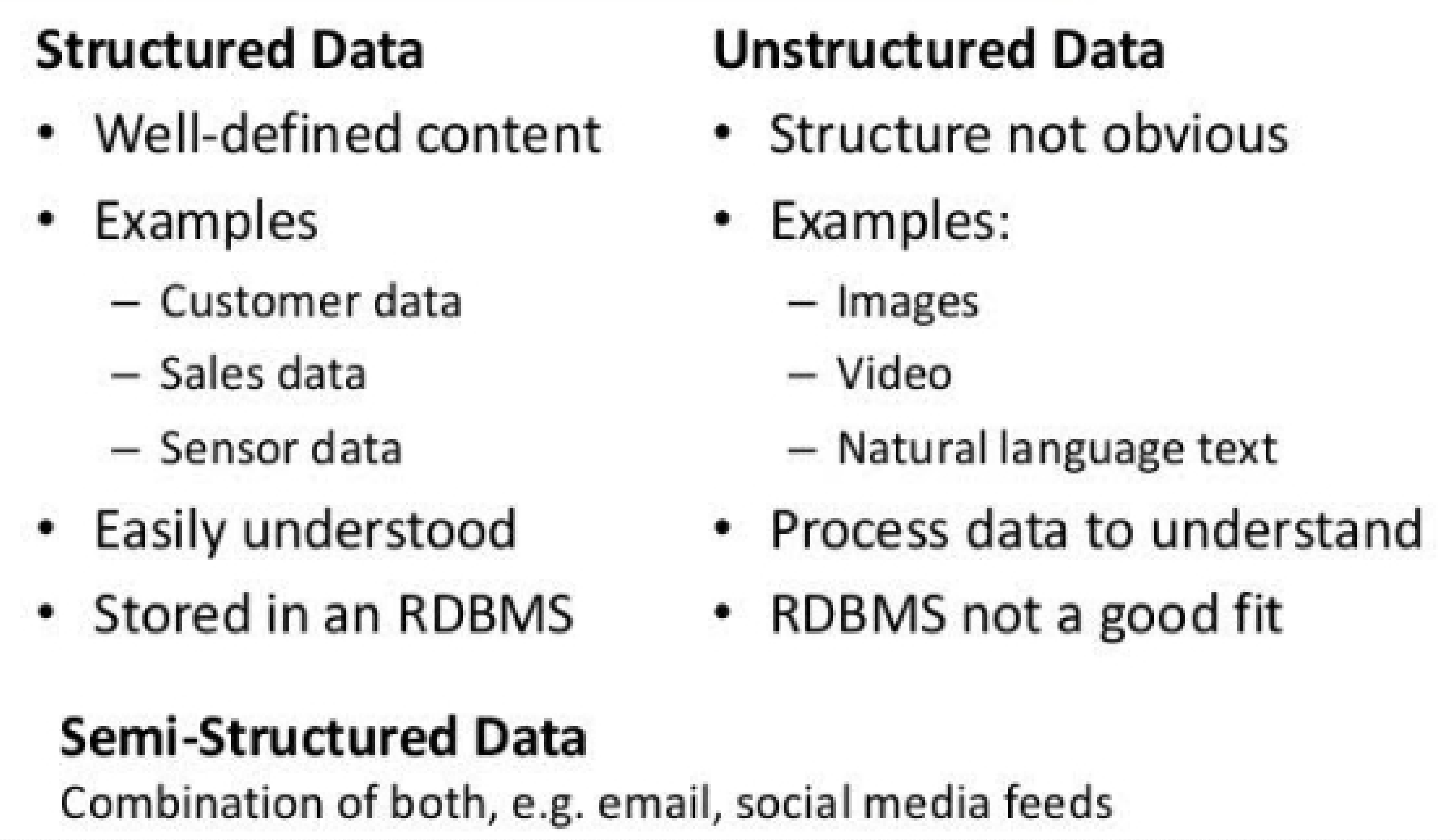
Cannot be displayed in rows, columns and relational databases

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



Structured Data

- Well-defined content
- Examples
  - Customer data
  - Sales data
  - Sensor data
- Easily understood
- Stored in an RDBMS

Unstructured Data

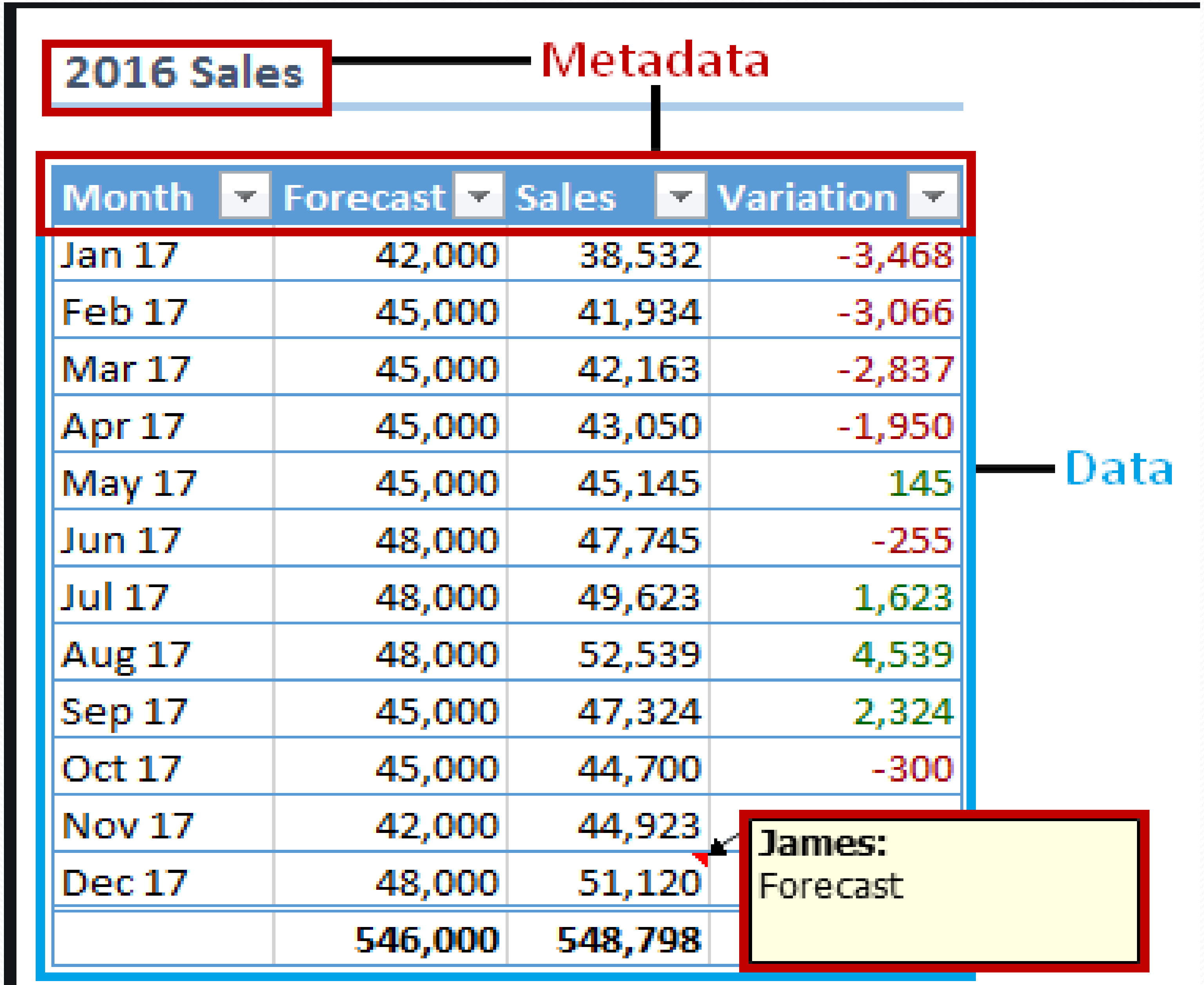
- Structure not obvious
- Examples:
  - Images
  - Video
  - Natural language text
- Process data to understand
- RDBMS not a good fit

Semi-Structured Data

Combination of both, e.g. email, social media feeds

Metadata – Data about Data

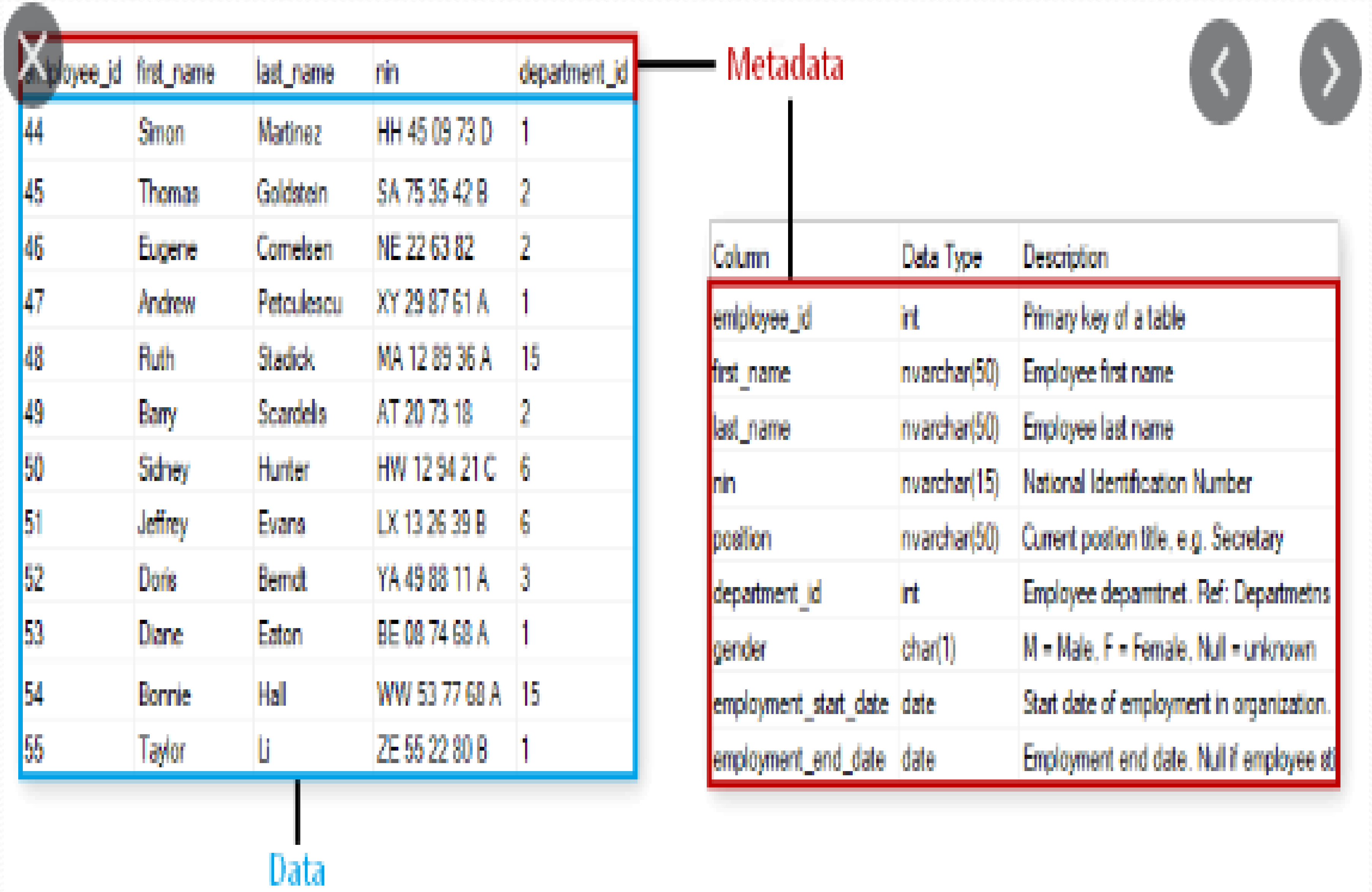
- A last category of data type is metadata. From a technical point of view, this is not a separate data structure, but it is one of the most important elements for Big Data analysis and big data solutions.
- Metadata is data about data.
- It provides additional information about a specific set of data.
- In a set of photographs, for example, metadata could describe when and where the photos were taken. The metadata then provides fields for dates and locations which, by themselves, can be considered structured data.
- Because of this reason, metadata is frequently used by Big Data solutions for initial analysis.



2016 Sales

Month	Forecast	Sales	Variation
Jan 17	42,000	38,532	-3,468
Feb 17	45,000	41,934	-3,066
Mar 17	45,000	42,163	-2,837
Apr 17	45,000	43,050	-1,950
May 17	45,000	45,145	145
Jun 17	48,000	47,745	-255
Jul 17	48,000	49,623	1,623
Aug 17	48,000	52,539	4,539
Sep 17	45,000	47,324	2,324
Oct 17	45,000	44,700	-300
Nov 17	42,000	44,923	
Dec 17	48,000	51,120	
	546,000	**548,798**	

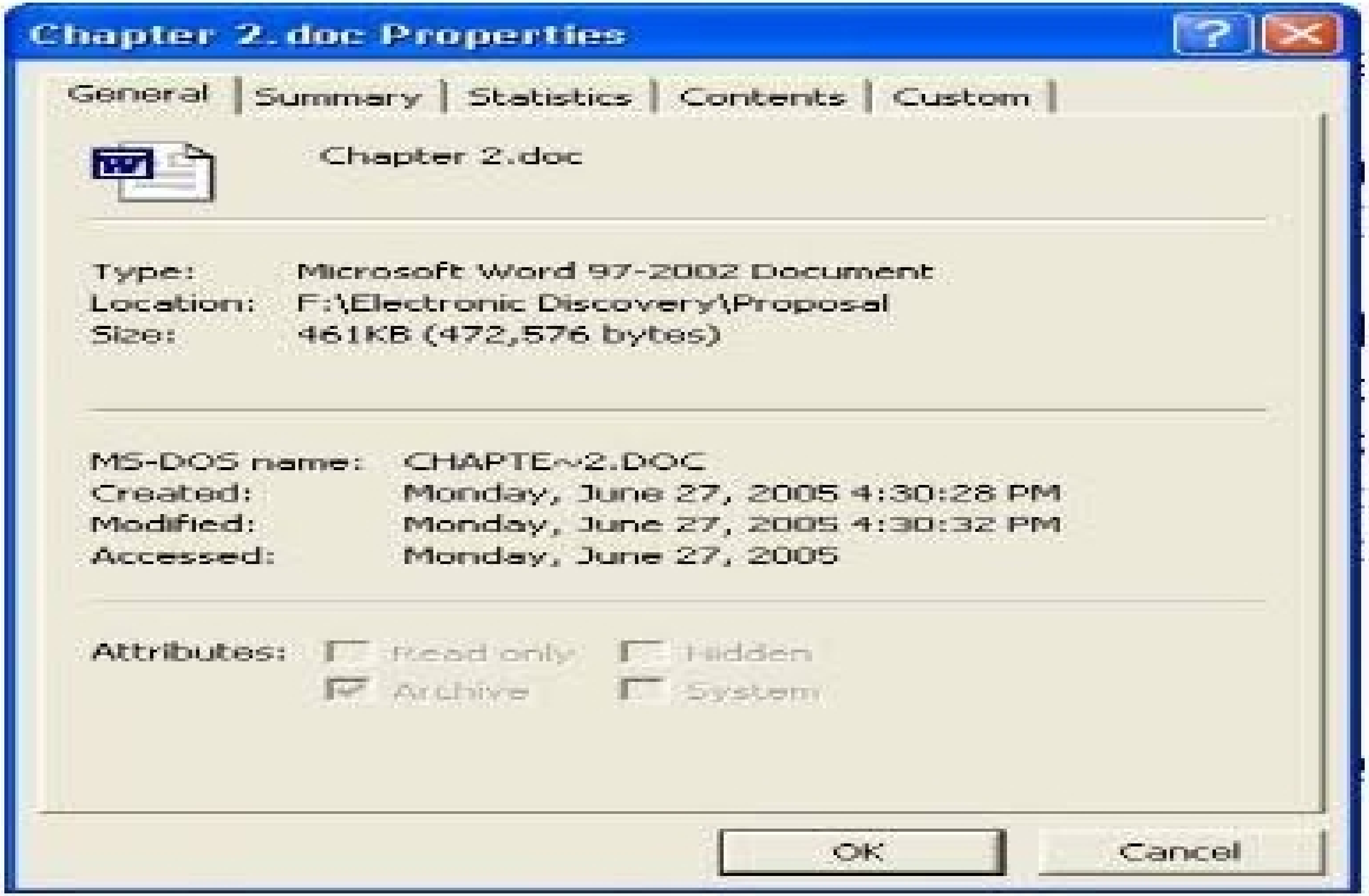
James:
Forecast



employee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 56 22 80 B	1

Column	Data Type	Description
employee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current position title, e.g. Secretary
department_id	int	Employee depamtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee st

Word Document Metadata Example



Chapter 2.doc

Type: Microsoft Word 97-2002 Document
Location: F:\Electronic Discovery\Proposal
Size: 461KB (472,576 bytes)

MS-DOS name: CHAPTE~2.DOC
Created: Monday, June 27, 2005 4:30:28 PM
Modified: Monday, June 27, 2005 4:30:32 PM
Accessed: Monday, June 27, 2005

Attributes: □ Read only □ Hidden
☑ Archive □ System

2.2 Data value Chain

- The Data Value Chain is introduced to describe the information flow within a big data system as a series of steps needed to generate value and useful insights from data.
- The Big Data Value Chain identifies the following key high-level activities:

Data Acquisition

- It is the process of gathering, filtering, and cleaning data before it is put in a data warehouse or any other storage solution on which data analysis can be carried out.
- Data acquisition is one of the major big data challenges in terms of infrastructure requirements.
- The infrastructure required to support the acquisition of big data must deliver low, predictable latency in both capturing data and in executing queries; be able to handle very high transaction volumes, often in a distributed environment; and support flexible and dynamic data structures.

Data Analysis

- It is concerned with making the raw data acquired amenable to use in decision-making as well as domain-specific usage.
- Data analysis involves exploring, transforming, and modelling data with the goal of highlighting relevant data, synthesizing and extracting useful hidden information with high potential from a business point of view.
- Related areas include data mining, business intelligence, and machine learning (covered in Chapter 4).

Data Curation

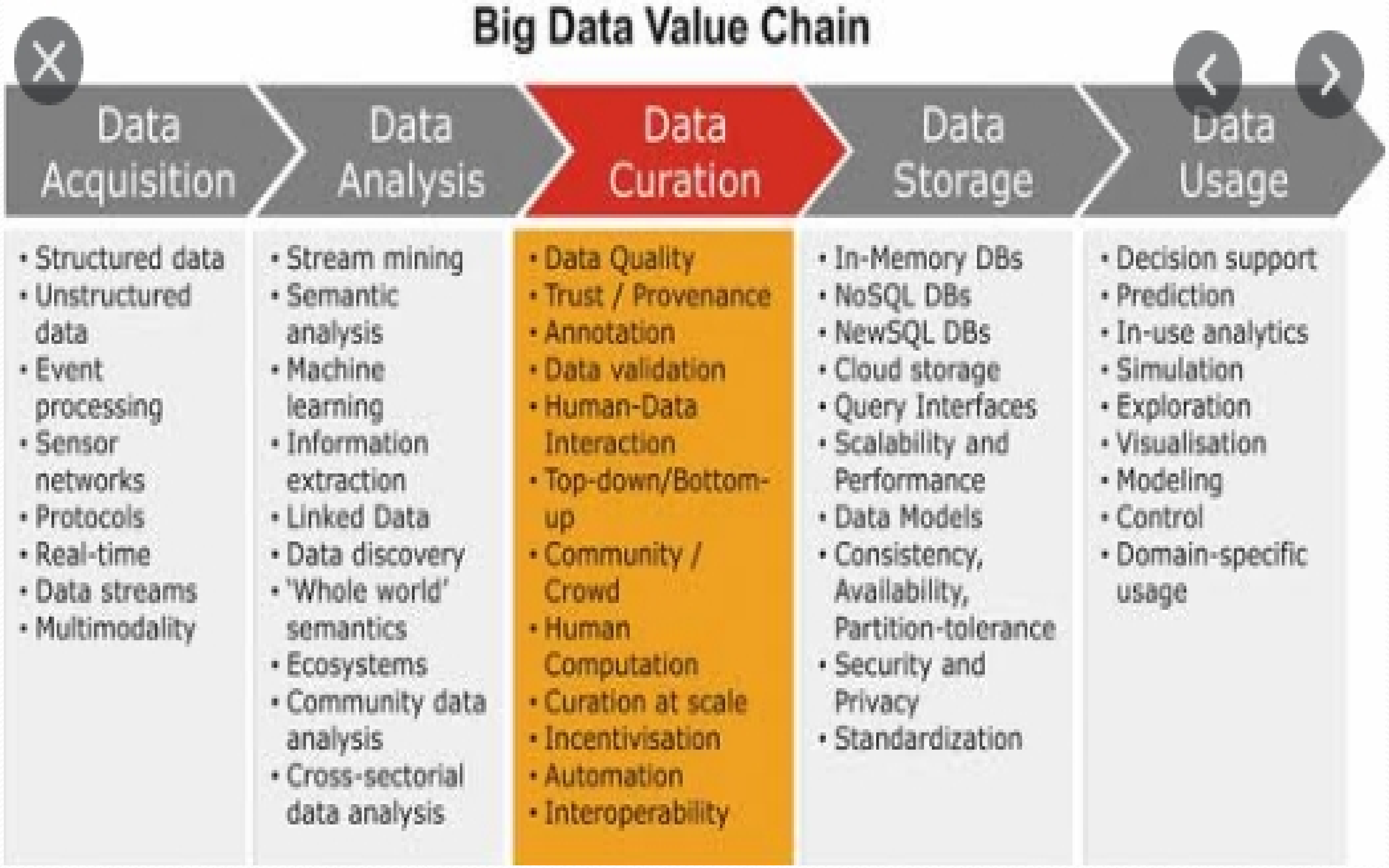
- It is the active management of data over its life cycle to ensure it meets the necessary data quality requirements for its effective usage.
- Data curation processes can be categorized into different activities such as content creation, selection, classification, transformation, validation, and preservation.
- Data curation is performed by expert curators that are responsible for improving the accessibility and quality of data.
- Data curators (also known as scientific curators, or data annotators) hold the responsibility of ensuring that data are trustworthy, discoverable, accessible, reusable, and fit their purpose.
- A key trend for the curation of big data utilizes community and crowd sourcing approaches.

Data Storage

- It is the persistence and management of data in a scalable way that satisfies the needs of applications that require fast access to the data.
- Relational Database Management Systems (RDBMS) have been the main, and almost unique, solution to the storage paradigm for nearly 40 years.
- However, the ACID (Atomicity, Consistency, Isolation, and Durability) properties that guarantee database transactions lack flexibility with regard to schema changes and the performance and fault tolerance when data volumes and complexity grow, making them unsuitable for big data scenarios.
- NoSQL technologies have been designed with the scalability goal in mind and present a wide range of solutions based on alternative data models.

Data Usage

- It covers the data-driven business activities that need access to data, its analysis, and the tools needed to integrate the data analysis within the business activity.
- Data usage in business decision-making can enhance competitiveness through reduction of costs, increased added value, or any other parameter that can be measured against existing performance criteria



Big Data Value Chain

Data Acquisition

- Structured data
- Unstructured data
- Event processing
- Sensor networks
- Protocols
- Real-time
- Data streams
- Multimodality

Data Analysis

- Stream mining
- Semantic analysis
- Machine learning
- Information extraction
- Linked Data
- Data discovery
- 'Whole world' semantics
- Ecosystems
- Community data analysis
- Cross-sectorial data analysis

Data Curation

- Data Quality
- Trust / Provenance
- Annotation
- Data validation
- Human-Data Interaction
- Top-down/Bottom-up
- Community / Crowd
- Human Computation
- Curation at scale
- Incentivisation
- Automation
- Interoperability

Data Storage

- In-Memory DBs
- NoSQL DBs
- NewSQL DBs
- Cloud storage
- Query Interfaces
- Scalability and Performance
- Data Models
- Consistency, Availability, Partition-tolerance
- Security and Privacy
- Standardization

Data Usage

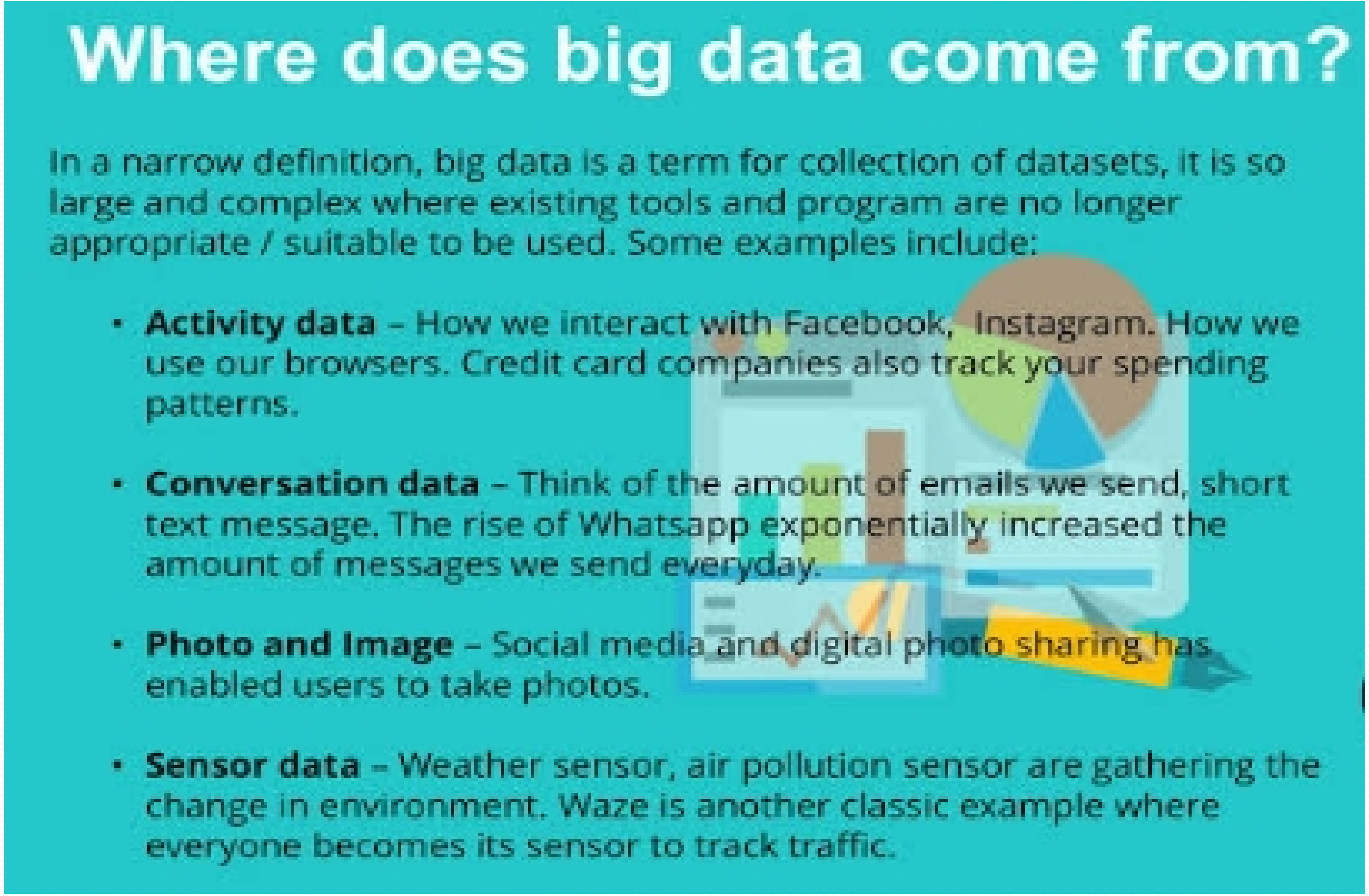
- Decision support
- Prediction
- In-use analytics
- Simulation
- Exploration
- Visualisation
- Modeling
- Control
- Domain-specific usage

2.3 Basic concepts of big data

- Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets.
- While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.
- In this section, we will talk about big data on a fundamental level and define common concepts you might come across.
- We will also take a high-level look at some of the processes and technologies currently being used in this space.

What Is Big Data?

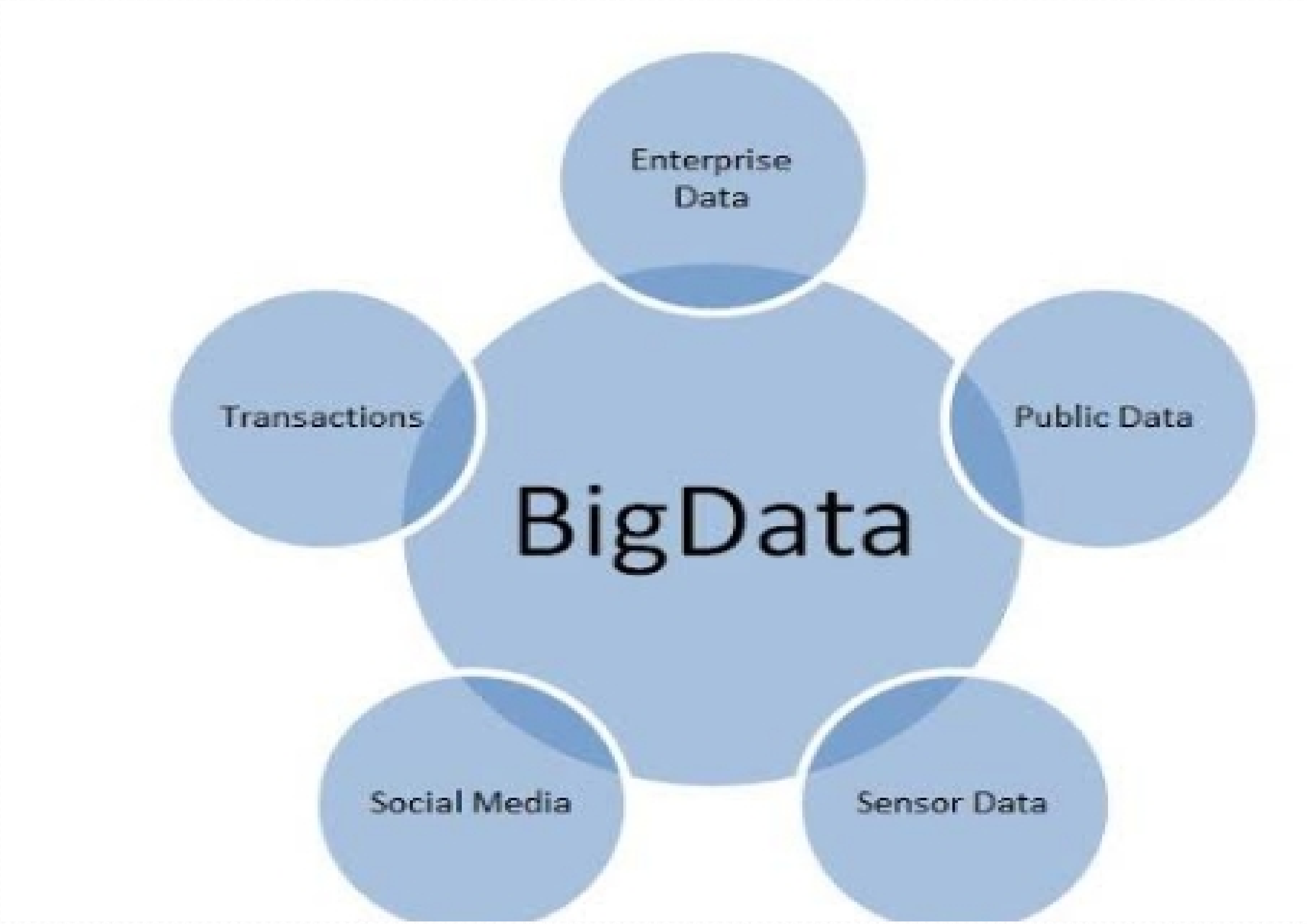
- An exact definition of "big data" is difficult to nail down because projects, vendors, practitioners, and business professionals use it quite differently. With that in mind, generally speaking, big data is:
  1. large datasets
  2. the category of computing strategies and technologies that are used to handle large datasets
- In this context, "large dataset" means a dataset too large to reasonably process or store with traditional tooling or on a single computer.
- This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization.



Where does big data come from?

In a narrow definition, big data is a term for collection of datasets, it is so large and complex where existing tools and program are no longer appropriate / suitable to be used. Some examples include:

- **Activity data** – How we interact with Facebook, Instagram. How we use our browsers. Credit card companies also track your spending patterns.
- **Conversation data** – Think of the amount of emails we send, short text message. The rise of Whatsapp exponentially increased the amount of messages we send everyday.
- **Photo and Image** – Social media and digital photo sharing has enabled users to take photos.
- **Sensor data** – Weather sensor, air pollution sensor are gathering the change in environment. Waze is another classic example where everyone becomes its sensor to track traffic.


Enterprise Data
Transactions
Public Data
BigData
Social Media
Sensor Data

Why Are Big Data Systems Different?

- The basic requirements for working with big data are the same as the requirements for working with datasets of any size.
- However, the massive scale, the speed of ingesting and processing, and the characteristics of the data that must be dealt with at each stage of the process present significant new challenges when designing solutions.
- The goal of most big data systems is to surface insights and connections from large volumes of heterogeneous data that would not be possible using conventional methods.
- In 2001, Gartner’s Doug Laney first presented what became known as the “three Vs of big data” to describe some of the characteristics that make big data different from other data processing:

Characteristics of Big Data – 3V's

Volume

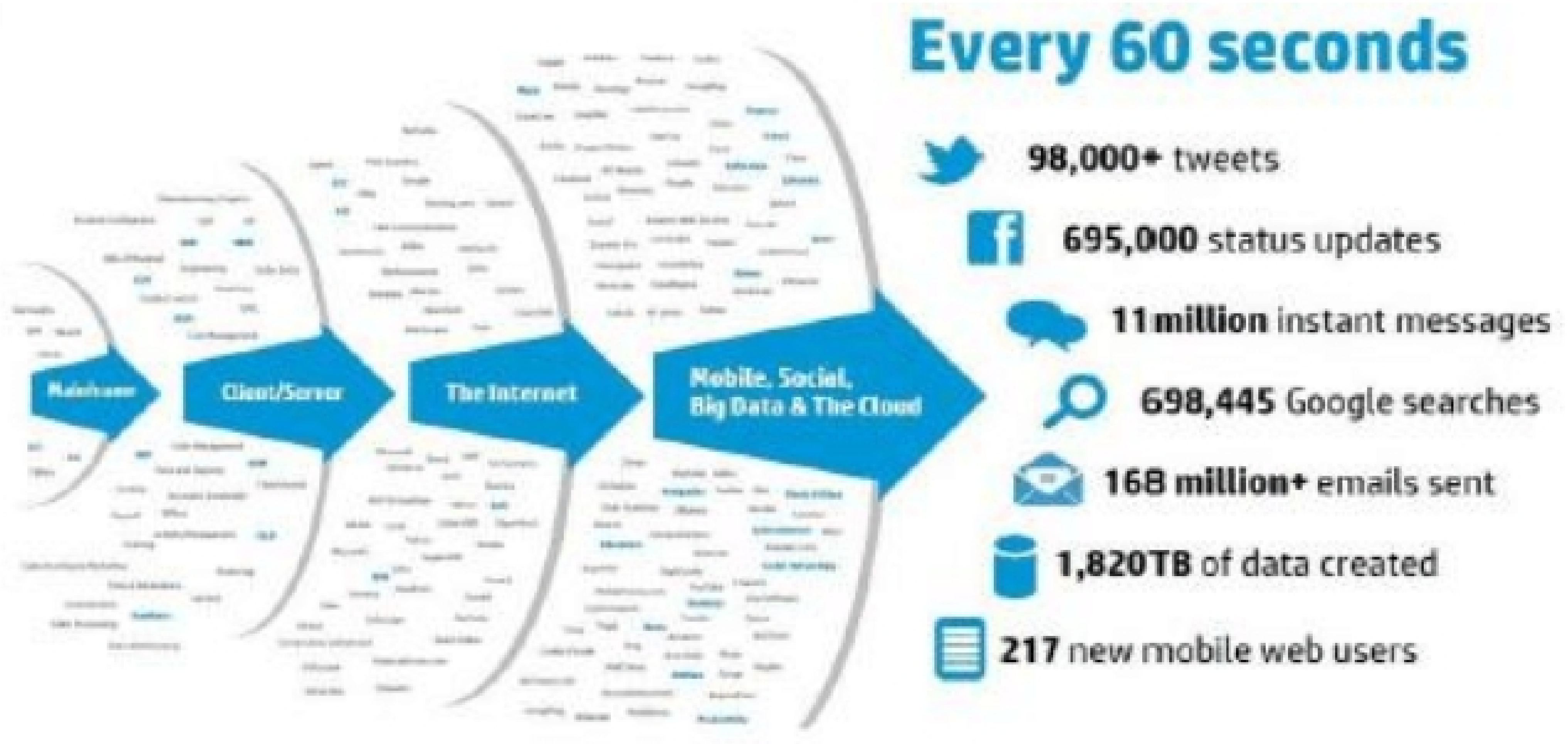
- The sheer scale of the information processed helps define big data systems.
- These datasets can be orders of magnitude larger than traditional datasets, which demands more thought at each stage of the processing and storage life cycle.
- Often, because the work requirements exceed the capabilities of a single computer, this becomes a challenge of pooling, allocating, and coordinating resources from groups of computers.
- Cluster management and algorithms capable of breaking tasks into smaller pieces become increasingly important.

Velocity

- Another way in which big data differs significantly from other data systems is the speed that information moves through the system.
- Data is frequently flowing into the system from multiple sources and is often expected to be processed in real time to gain insights and update the current understanding of the system.
- This focus on near instant feedback has driven many big data practitioners away from a batch-oriented approach and closer to a real-time streaming system.
- Data is constantly being added, massaged, processed, and analyzed in order to keep up with the influx of new information and to surface valuable information early when it is most relevant.
- These ideas require robust systems with highly available components to guard against failures along the data pipeline.

Variety

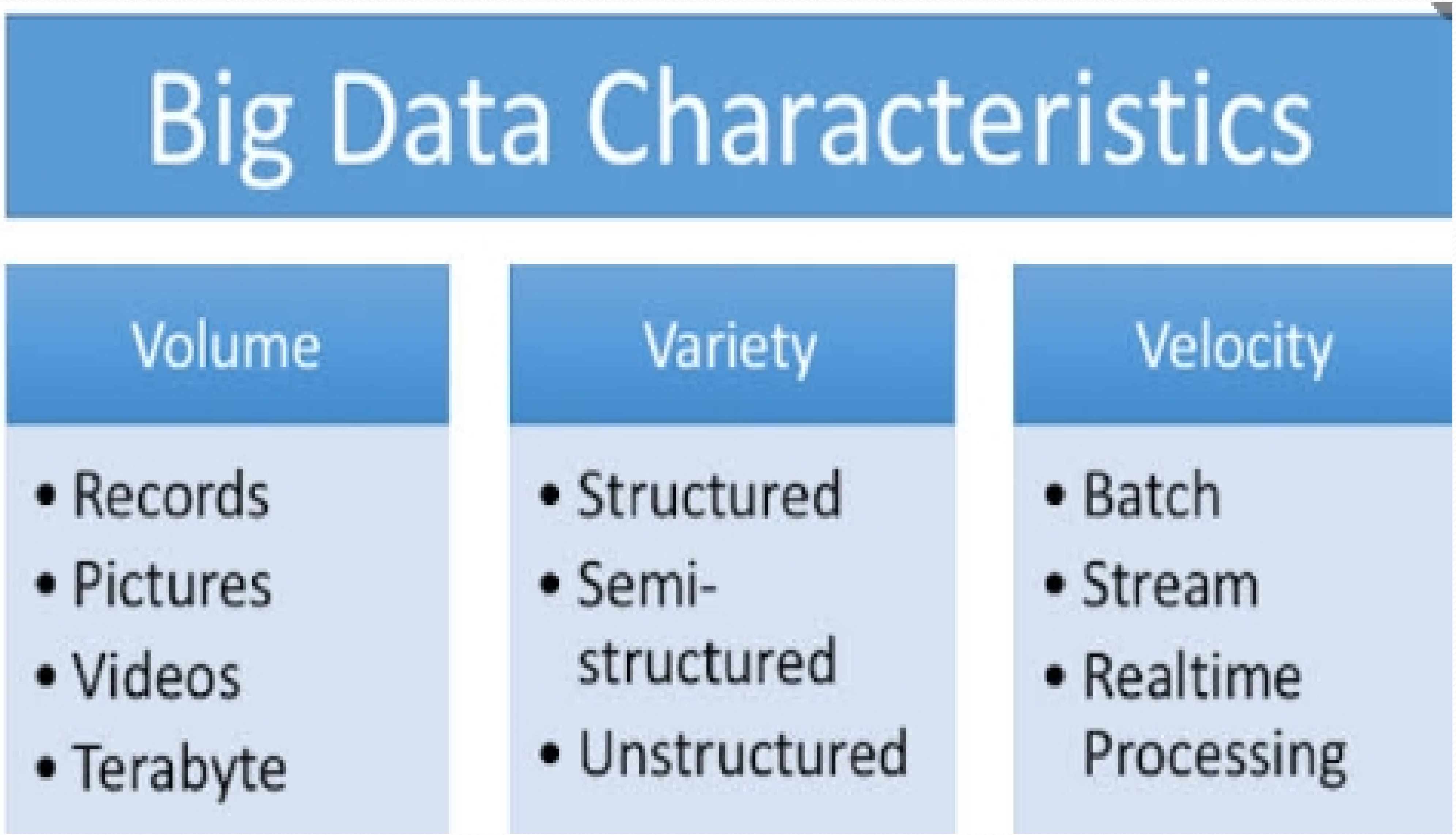
- Big data problems are often unique because of the wide range of both the sources being processed and their relative quality.
- Data can be ingested from internal systems like application and server logs, from social media feeds and other external APIs, from physical device sensors, and from other providers.
- Big data seeks to handle potentially useful data regardless of where it’s coming from by consolidating all information into a single system.
- The formats and types of media can vary significantly as well. Rich media like images, video files, and audio recordings are ingested alongside text files, structured logs, etc.
- While more traditional data processing systems might expect data to enter the pipeline already labeled, formatted, and organized, big data systems usually accept and store data closer to its raw state.
- Ideally, any transformations or changes to the raw data will happen in memory at the time of processing.



Every 60 seconds

- 98,000+ tweets
- 695,000 status updates
- 11 million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820TB of data created
- 217 new mobile web users

Examples of big data velocity



Big Data Characteristics

Volume

- Records
- Pictures
- Videos
- Terabyte

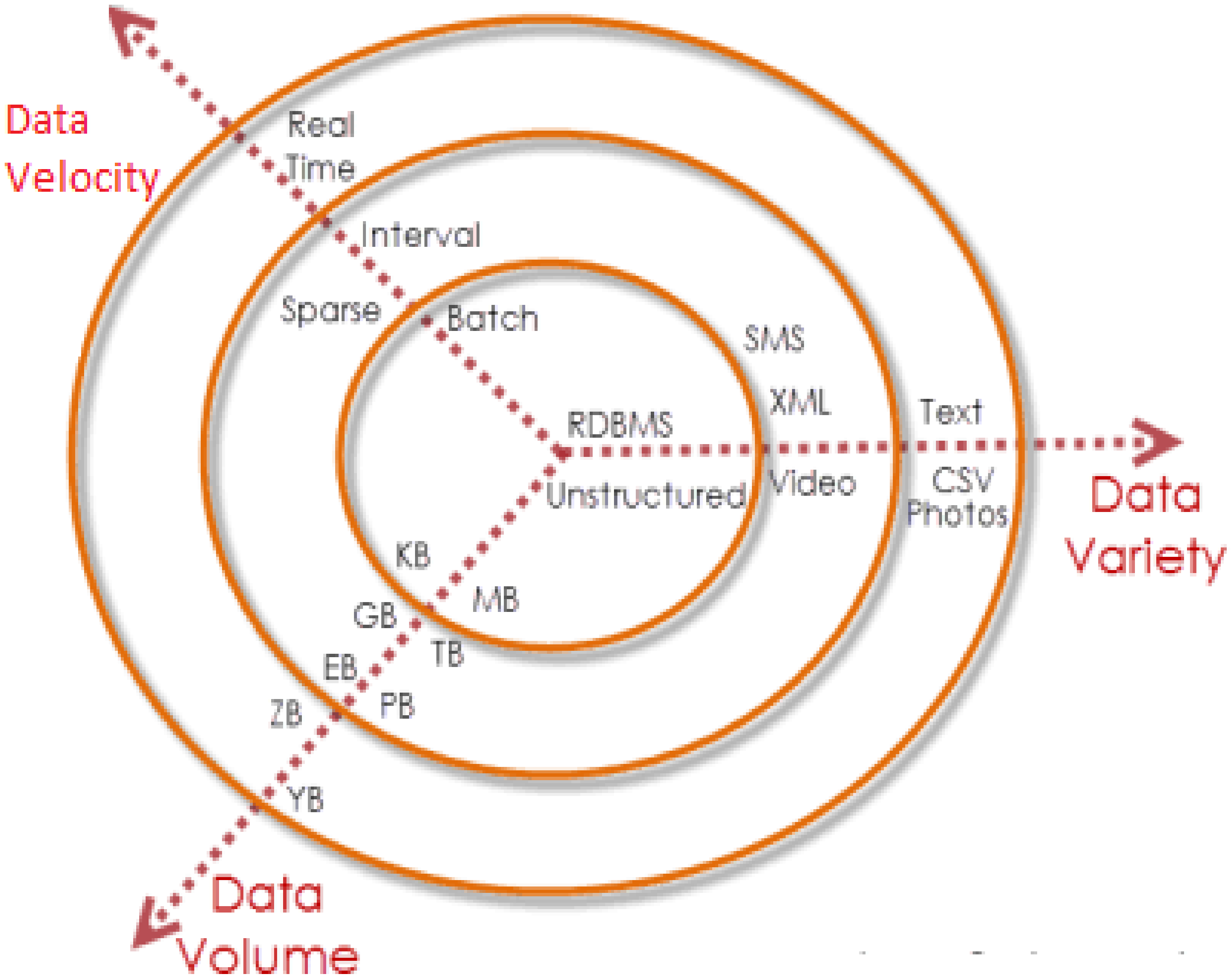
Variety

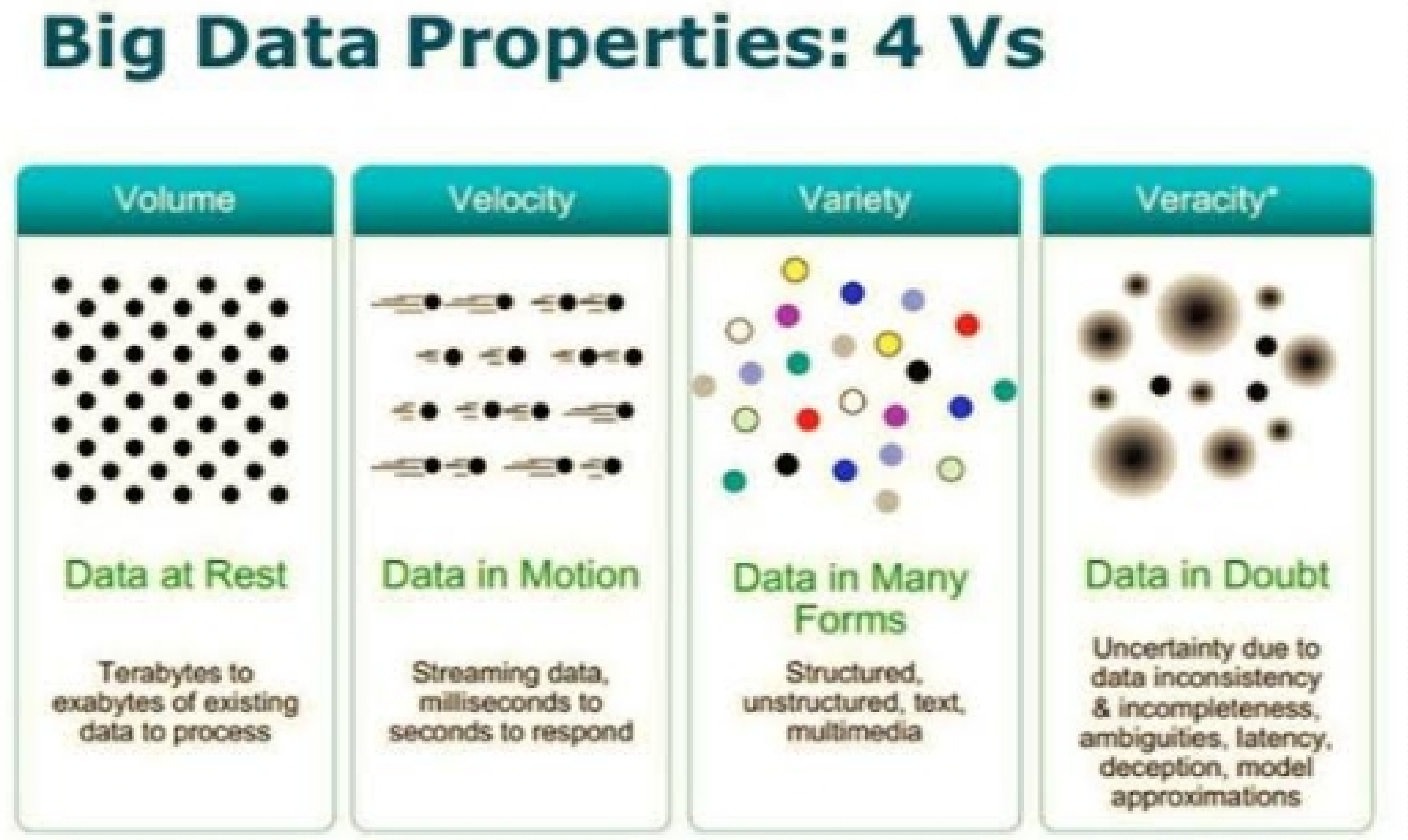
- Structured
- Semi-structured
- Unstructured

Velocity

- Batch
- Stream
- Realtime Processing

3 Vs of Big Data





Big Data Properties: 4 Vs

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

Other Characteristics of Big data – 6V’s

- Various individuals and organizations have suggested expanding the original 3Vs, which tended to describe challenges rather than qualities of big data. The additions include:
- **Veracity**: The variety of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data (and consequently, the quality of the resulting analysis)
- **Variability**: Variation in the data leads to wide variation in quality. Additional resources may be needed to identify, process, or filter low quality data to make it more useful.
- **Value**: The ultimate challenge of big data is delivering value. Sometimes, the systems and processes in place are complex enough that using the data and extracting actual value can become difficult.

Big Data Life Cycle – *ingesting, persisting, commuting & analyzing, and visualizing*

- So how is data actually processed with a big data system?
- While approaches to implementation differ, there are some commonalities in the strategies and software that we can talk about generally.
- Therefore, the widely adopted steps are presented below (note it might not be true in all cases).
- The general categories of activities involved with big data processing are:
  - *Ingesting data into the system*
  - *Persisting the data in storage*
  - *Computing and Analyzing data*
  - *Visualizing the results*
- Before discussing these steps, understanding of clustered computing - an important strategy employed by most big data solutions is important.

Clustered Computing

- Setting up a computing cluster is often the foundation for technology used in each of the life cycle stages.
- Because of the quantities of big data, individual computers are often inadequate for handling the data at most stages.
- Therefore, to address the high storage and computational needs of big data, computer clusters are a better fit.
- Big data clustering software combines the resources of many smaller machines, to provide a number of benefits:
- **Resource Pooling**: Combining the available storage space to hold data is a clear benefit, but CPU and memory pooling is also extremely important. Processing large datasets requires large amounts these three resources.
- **High Availability**: Clusters can provide varying levels of fault tolerance and availability guarantees to prevent hardware or software failures from affecting access to data and processing. This becomes increasingly important as we continue to emphasize the importance of real-time analytics.
- **Easy Scalability**: Clusters make it easy to scale horizontally by adding additional machines to the group. This means the system can react to changes in resource requirements without expanding the physical resources on a machine.
- Using clusters requires a solution for managing cluster membership, coordinating resource sharing, and scheduling actual work on individual nodes. Solution for cluster membership and resource allocation include:
  - software like Hadoop’s YARN (which stands for Yet Another Resource Negotiator) or Apache Mesos.
- The assembled computing cluster often acts as a foundation which other software interfaces with to process the data. The machines involved in the computing cluster are also typically involved with the management of a distributed storage system (*discuss in data persistence).*

Step 1: Ingesting Data into the System

- Data ingestion is the process of taking raw data and adding it to the system.
- Complexity of this operation depends - heavily on the format and quality of the data sources and how far the data is from the desired state prior to processing.
- Dedicated ingestion tools that can add data to a big data system are.
  - **Apache Sqoop** – technologies that can take existing data from relational databases and add it to a big data system.
  - Similarly, **Apache Flume and Apache Chukwa** are projects designed to aggregate and import application and server logs.
  - Queuing systems like **Apache Kafka** can also be used as an interface between various data generators and a big data system.
  - Ingestion frameworks like **Gobblin** can help to aggregate and normalize the output of these tools at the end of the ingestion pipeline.
- In the ingestion process - some level of analysis, sorting, and labelling usually takes place.
  - This process is sometimes called ETL (stands for extract, transform, and load).
- While this term conventionally refers to legacy data warehousing processes, some of the same concepts apply to data entering the big data system.
- Typical operations might include modifying the incoming data to format it, categorizing and labelling data, filtering out unneeded or bad data, or potentially validating that it adheres to certain requirements.
- With those capabilities in mind, ideally, the captured data should be kept as raw as possible for greater flexibility further on down the pipeline.

Step 2: Persisting the Data in Storage

- The ingestion processes typically hand the data off to the components that manage storage, so that it can be reliably persisted to disk.
- Although looks simple operation, the volume of incoming data, the requirements for availability, and the distributed computing layer make more complex storage systems necessary.
- This usually means leveraging a distributed file system for raw data storage.
- Solutions like Apache Hadoop’s HDFS filesystem allow large quantities of data to be written across multiple nodes in the cluster.
- This ensures that the data can be accessed by compute resources, can be loaded into the cluster’s RAM for in-memory operations, and can gracefully handle component failures.
- Other distributed filesystems can be used in place of HDFS including Ceph and GlusterFS.
- Data can also be imported into other distributed systems for more structured access.
- Distributed databases, especially NoSQL databases, are well-suited for this role because they are often designed with the same fault tolerant considerations and can handle heterogeneous data.
- Many different types of distributed databases available to choose from depending on how you want to organize and present the data.

Step 3: Computing and Analyzing Data

- Once the data is available, the system can begin processing the data to surface actual information.
- The computation layer is perhaps the most diverse part of the system.
  - the requirements and best approach can vary significantly depending on what type of insights desired.
- Data is often processed repeatedly - either iteratively by a single tool or by using a number of tools to surface different types of insights.
- **Two main method of processing: Batch and Real-time**
- **Batch processing** is one method of computing over a large dataset.
- The process involves: breaking work up into smaller pieces, scheduling each piece on an individual machine, reshuffling the data based on the intermediate results, and then calculating and assembling the final result.
- These steps are often referred: splitting, mapping, shuffling, reducing, and assembling, or collectively as a distributed map reduce algorithm. This is the strategy used by Apache Hadoop’s MapReduce.
- Batch processing is most useful when dealing with very large datasets that require quite a bit of computation.

- **Real-time processing** - While batch processing is a good fit for certain types of data and computation, other workloads require more real-time processing.
- Real-time processing demands that information be processed and made ready immediately and requires the system to react as new information becomes available.
  - One way of achieving this is stream processing, which operates on a continuous stream of data composed of individual items.
- Another common characteristic of real-time processors is in-memory computing, which works with representations of the data in the cluster’s memory to avoid having to write back to disk.
- Apache Storm, Apache Flink, and Apache Spark provide different ways of achieving real-time or near real-time processing.
- There are trade-offs with each of these technologies, which can affect which approach is best for any individual problem.
- In general, real-time processing is best suited for analyzing smaller chunks of data that are changing or being added to the system rapidly.
- The above examples represent computational frameworks. However, there are many other ways of computing over or analyzing data within a big data system. These tools frequently plug into the above frameworks and provide additional interfaces for interacting with the underlying layers.(see more on the module).

Step 4: Visualizing the Results

- Due to the type of information being processed in big data systems, recognizing trends or changes in data over time is often more important than the values themselves.
- Visualizing data is one of the most useful ways to spot trends and make sense of a large number of data points.
- Real-time processing is frequently used to visualize application and server metrics. The data changes frequently and large deltas in the metrics typically indicate significant impacts on the health of the systems or organization.
- Projects like Prometheus can be useful for processing the data streams as a time-series database and visualizing that information.
- Elastic Stack – is one popular way of visualizing data, formerly known as the ELK stack.
- Composed of Logstash for data collection, Elasticsearch for indexing data, and Kibana for visualization, the Elastic stack can be used with big data systems to visually interface with the results of calculations or raw metrics.
- A similar stack can be achieved using Apache Solr for indexing and a Kibana fork called Banana for visualization. The stack created by these is called Silk.
- Another visualization technology typically used for interactive data science work is a data “notebook”.
- These projects allow for interactive exploration and visualization of the data in a format conducive to sharing, presenting, or collaborating. Popular examples of this type of visualization interface are Jupyter Notebook and Apache Zeppelin.

Review questions

1. What does one understand by the term Data Science?
2. Differentiate Between Data Analytics and Data Science
3. Types of data like semi structured/structured/unstructured
4. What is Big data? And its life cycle
5. Quickly differentiate between Machine Learning, Data Science, and AI.
6. Why data cleaning plays a vital role in the analysis?
7. What is Big Data, and where does it come from? How does it work?
8. What are the 5 V’s in Big Data?
9. How can Big Data add value to businesses?
10. What are your experiences in big data?
11. What are some of the challenges that come with a big data project?
12. Why is Hadoop so popular in big data analytics?
13. How Will Big Data Help My Company?

Thank You! End of Chapter two