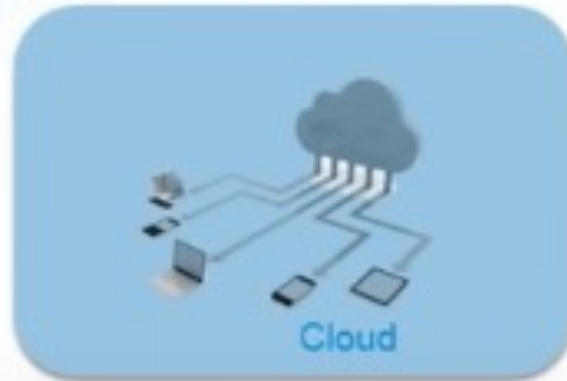


# Road Map

- Evolution of Technology
- Types of Data
- Big Data - Definition Aspect
- Big data Vs Not Big data
- Challenges of big data

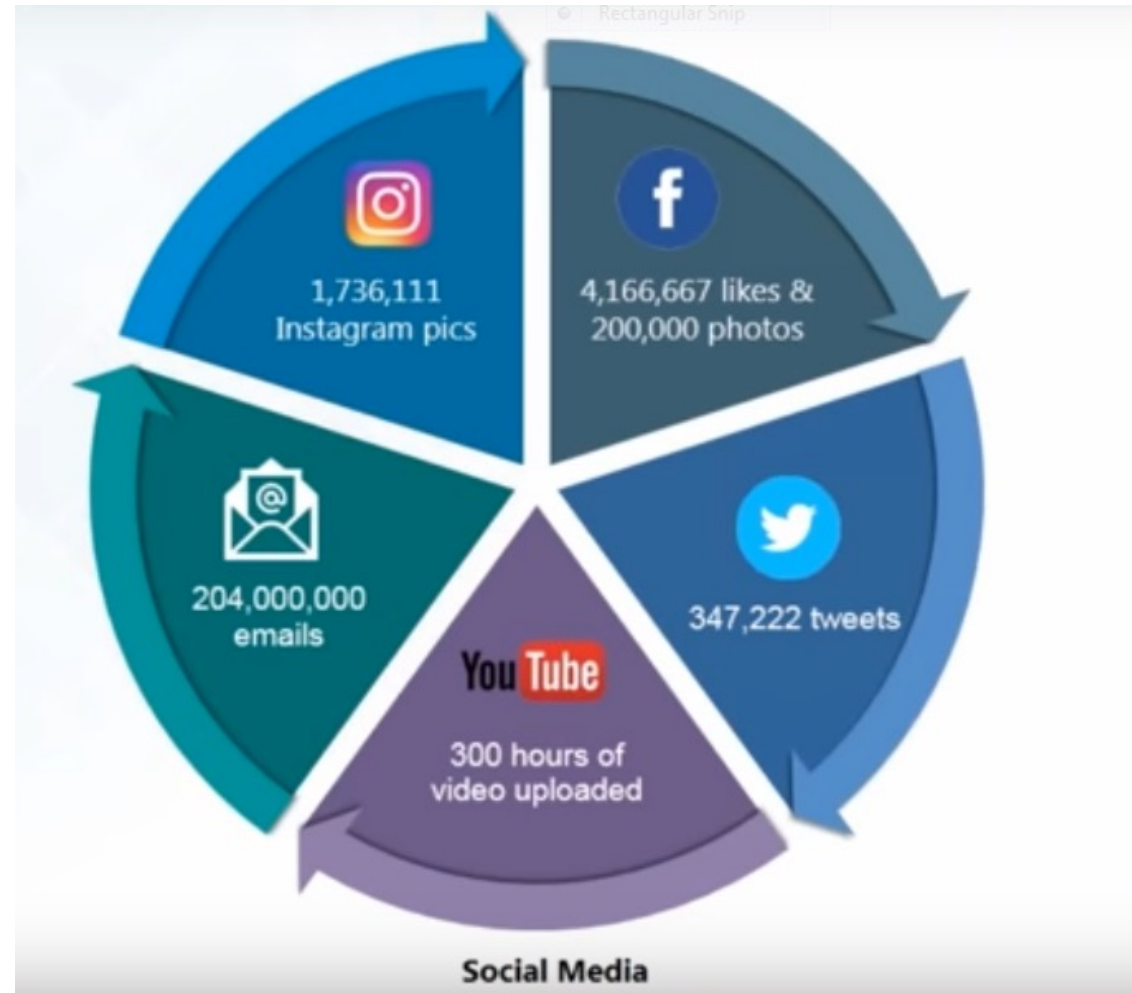
# Evolution of Technology



# Internet of Things

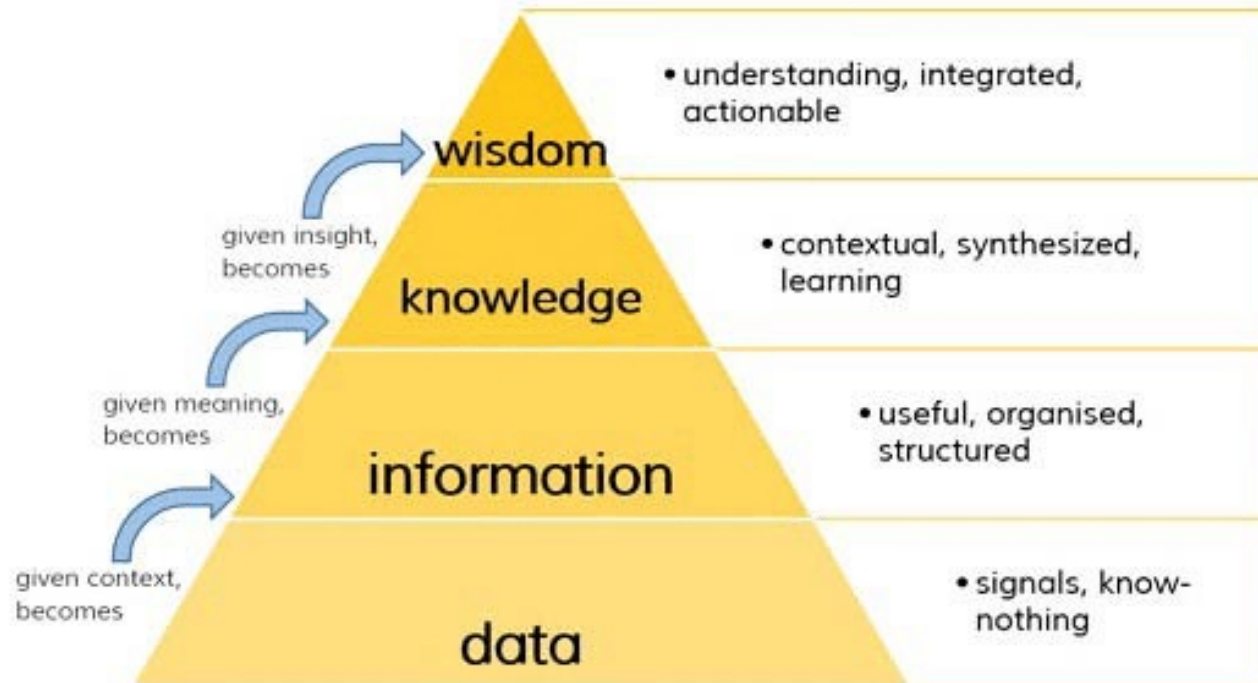


## Social Media Usage



# Classification of Digital Data

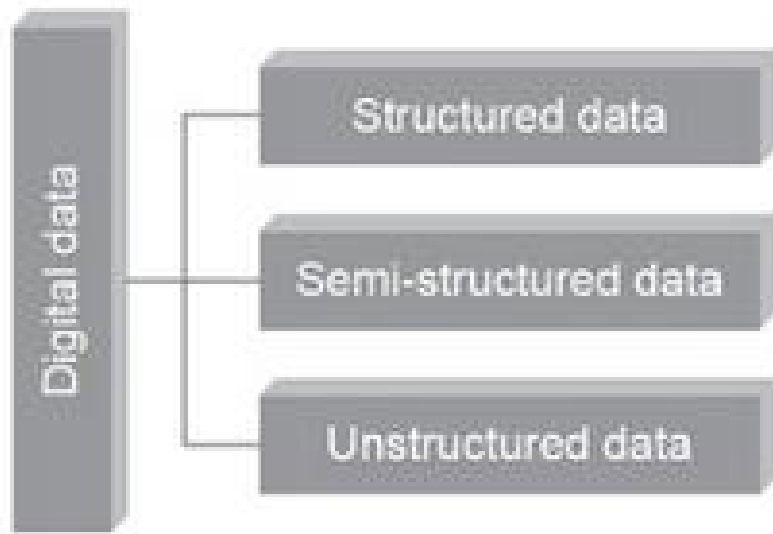
Data → Information  
Information → Insights

[illegible]

# Classification of Digital Data

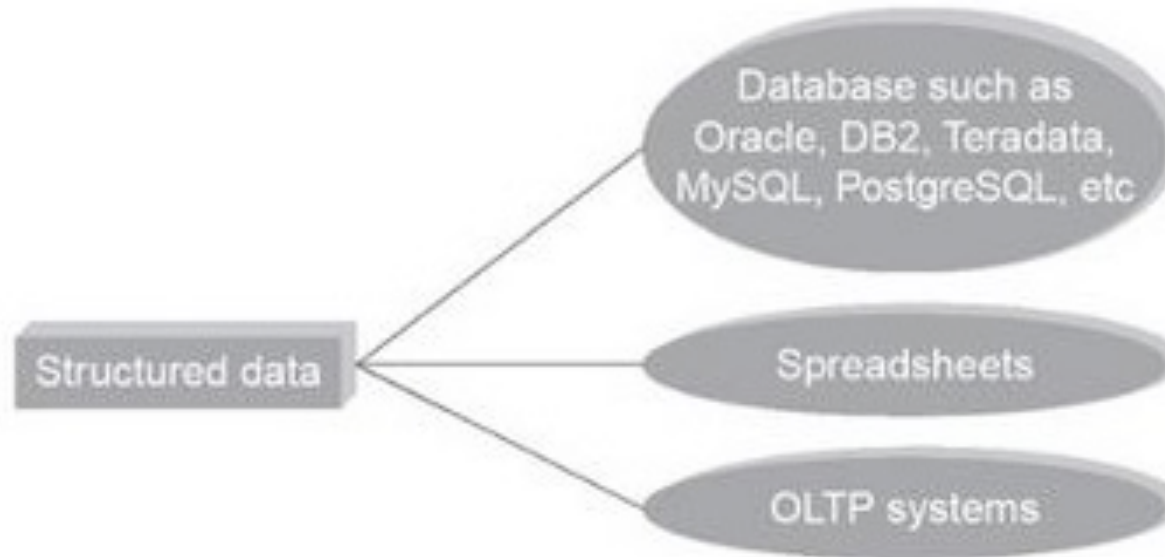
Data → Information

Information → Insights



# Structured data

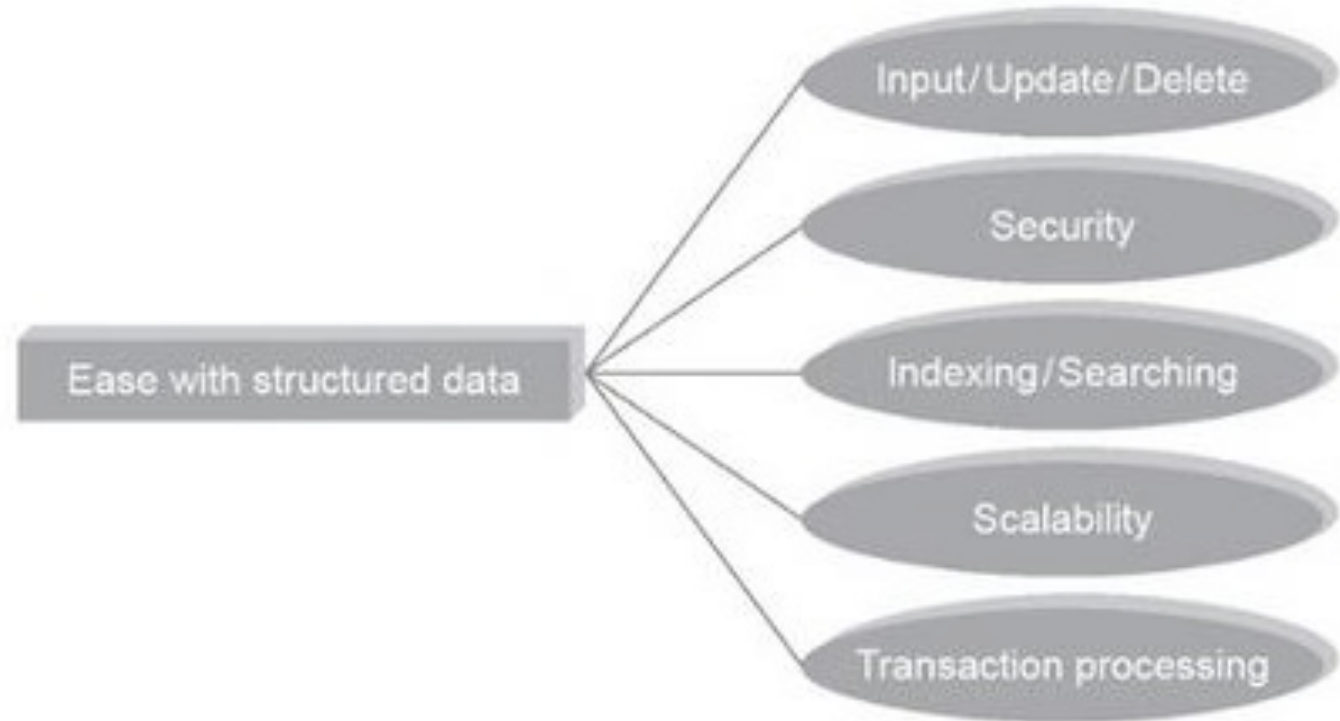
- When do we say that the data is structured??
- Sources of structured data



**Figure 1.4** Sources of structured data.

# Working with structured data

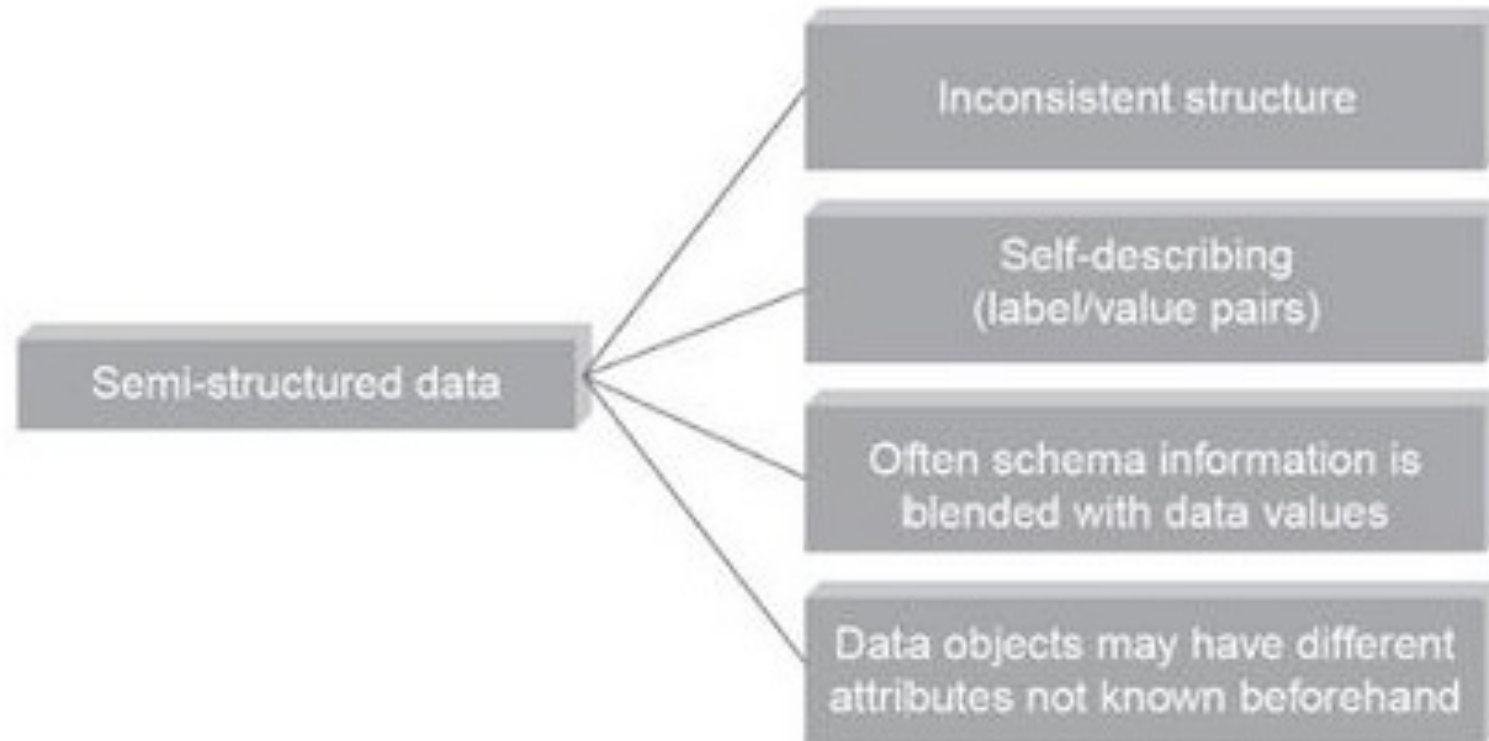
- Insert/update/delete
- Indexing
- Transaction processing
- Security
- Scalability



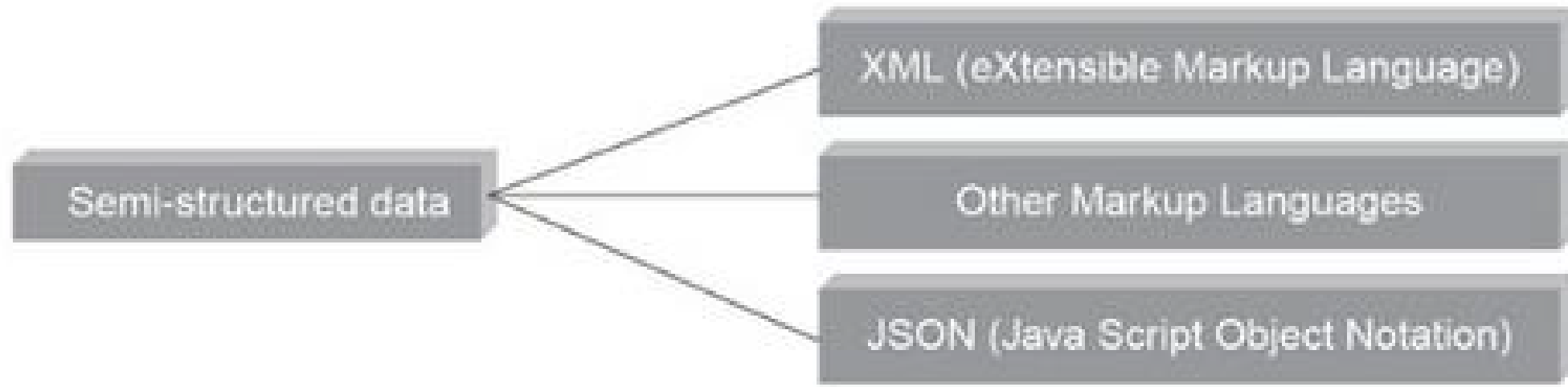


# Semi-structured data

- It does not conform to the data models that one typically associates with relational databases or any other form of data tables
- It uses tags to segregate semantic elements

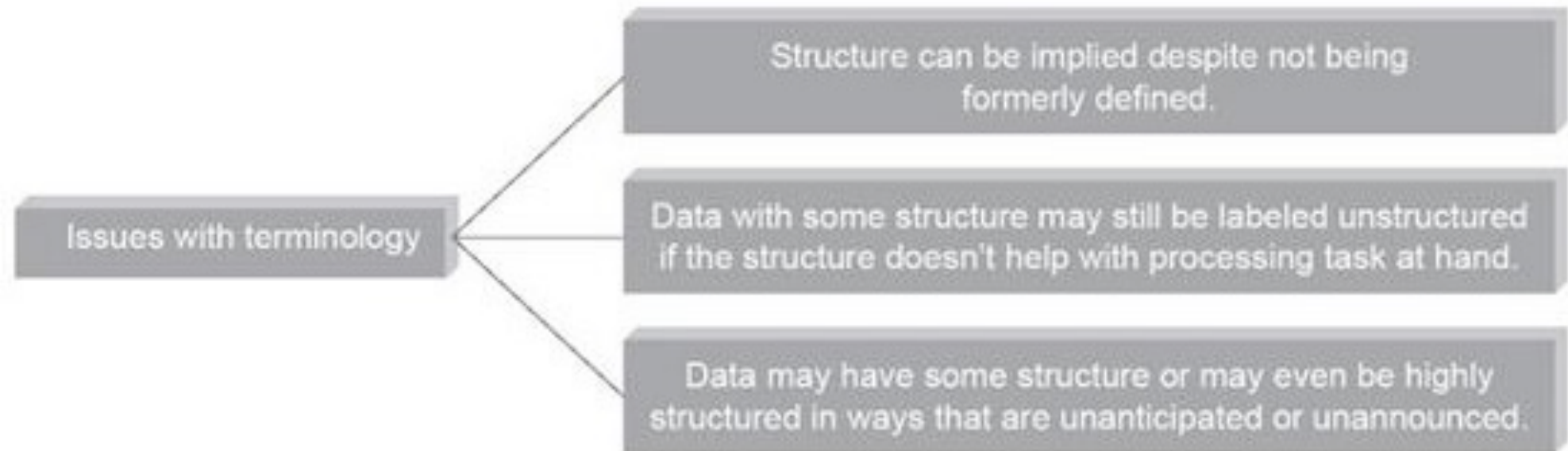


# Sources of semi-structured data

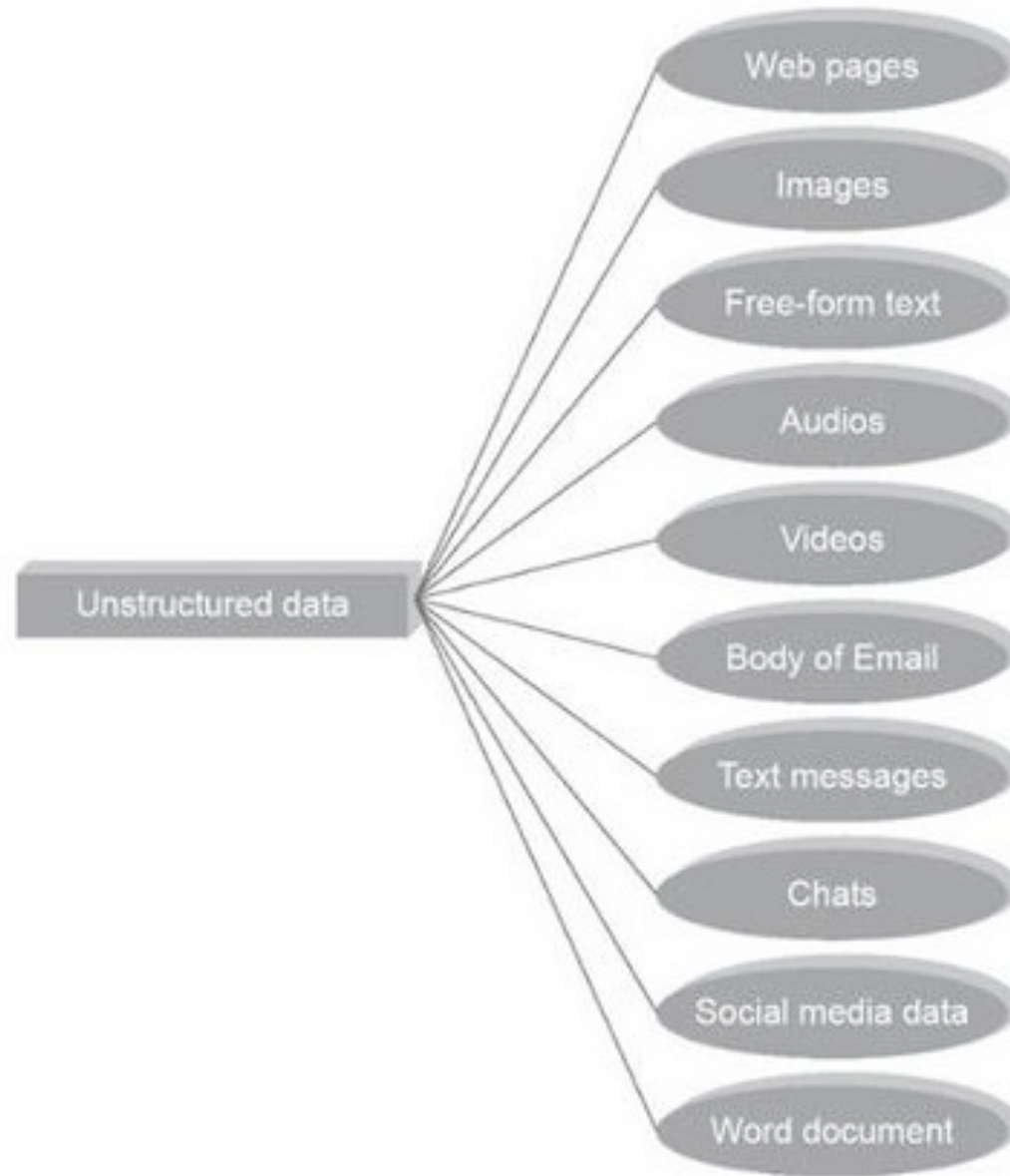


# Unstructured data

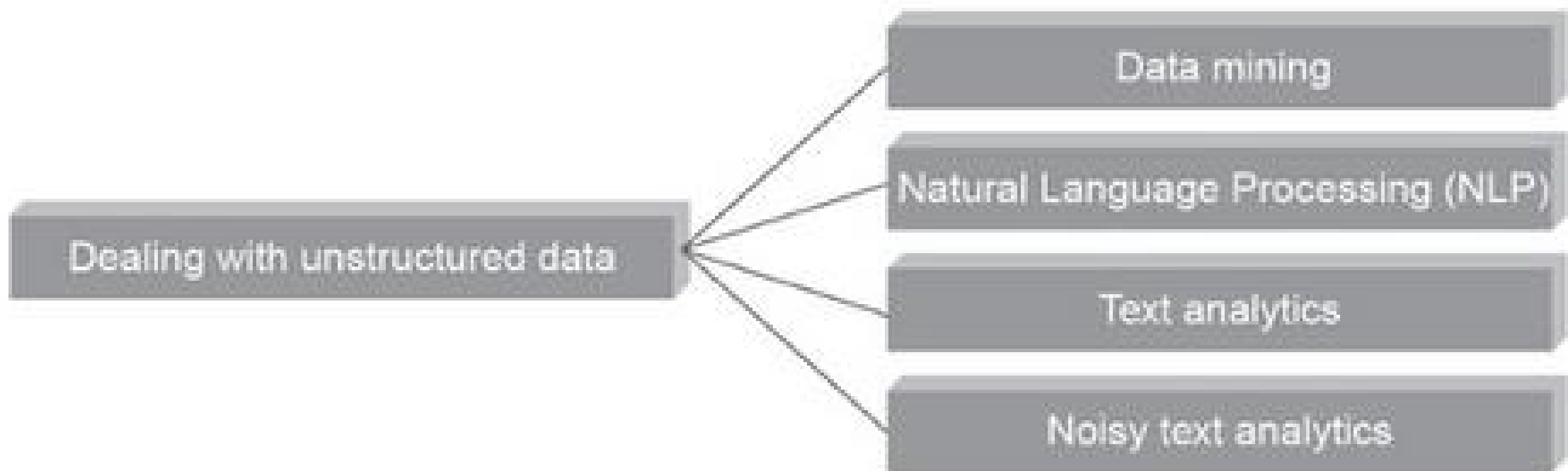
- Does not conform to any predefined data model
- The structure can be unpredictable.



# Sources of unstructured data



# How to deal with unstructured data?



# Inclass#exercise

## A. Place Me in the Basket

Structured	Unstructured	Semi-Structured

Following words are to be placed in the relevant basket:

Email  
MS Access  
Images  
Database  
Chat conversations

Relations/Tables  
Facebook  
Videos  
MS Excel  
XML

# Solution

Answer:

Structured	Unstructured	Semi-Structured
MS Access	Email	XML
Database	Images	
Relations/Tables	Chat conversations	
MS Excel	Facebook	
	Videos	

# Let's Discuss

- Why email in unstructured category?
- Where should we put CCTV footage?

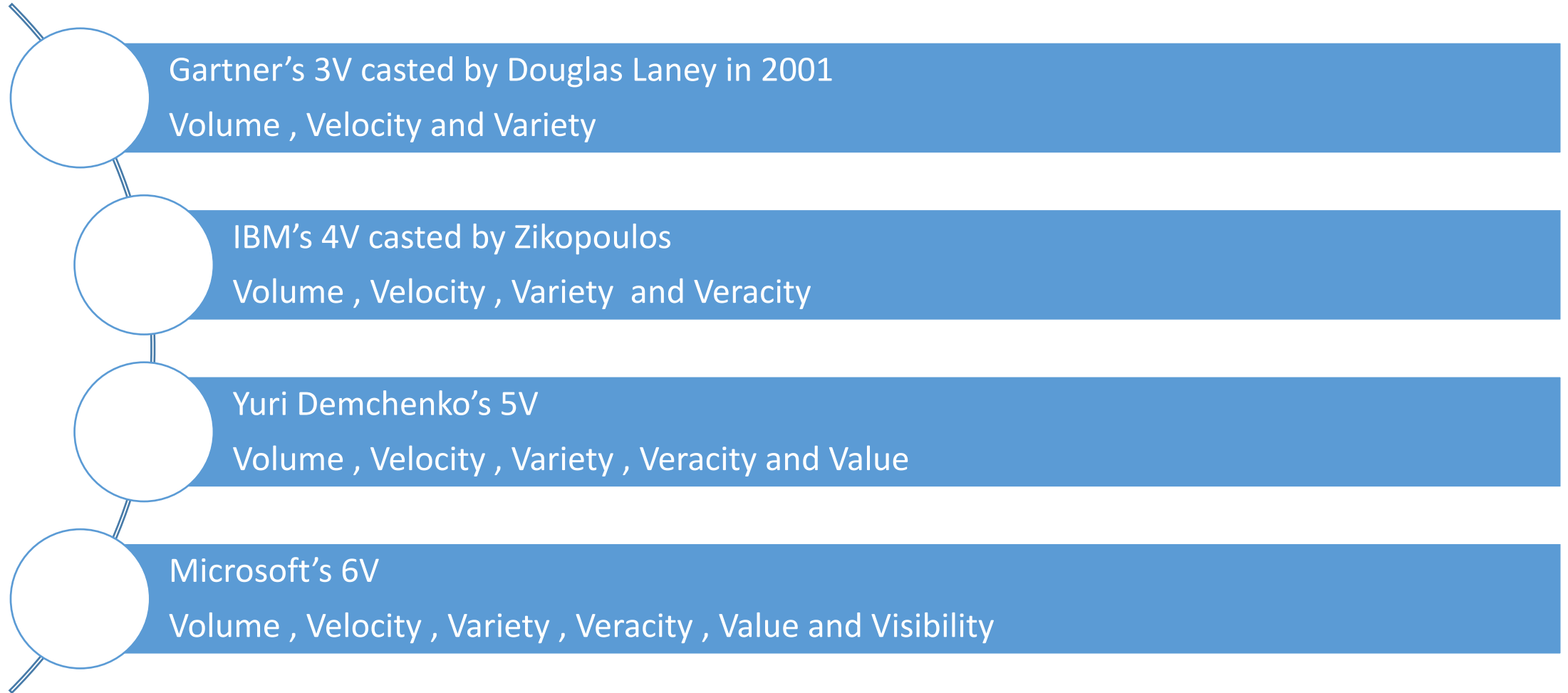


***You are at city shopping mall. You see few people are browsing the items. Some of them are looking for discounts. Some of them are filling feedback form. Few people are at billing counter. You may consider other things and events happening in this scenario. Think for while on the different types of data generated. Mention each of them with proper logic***

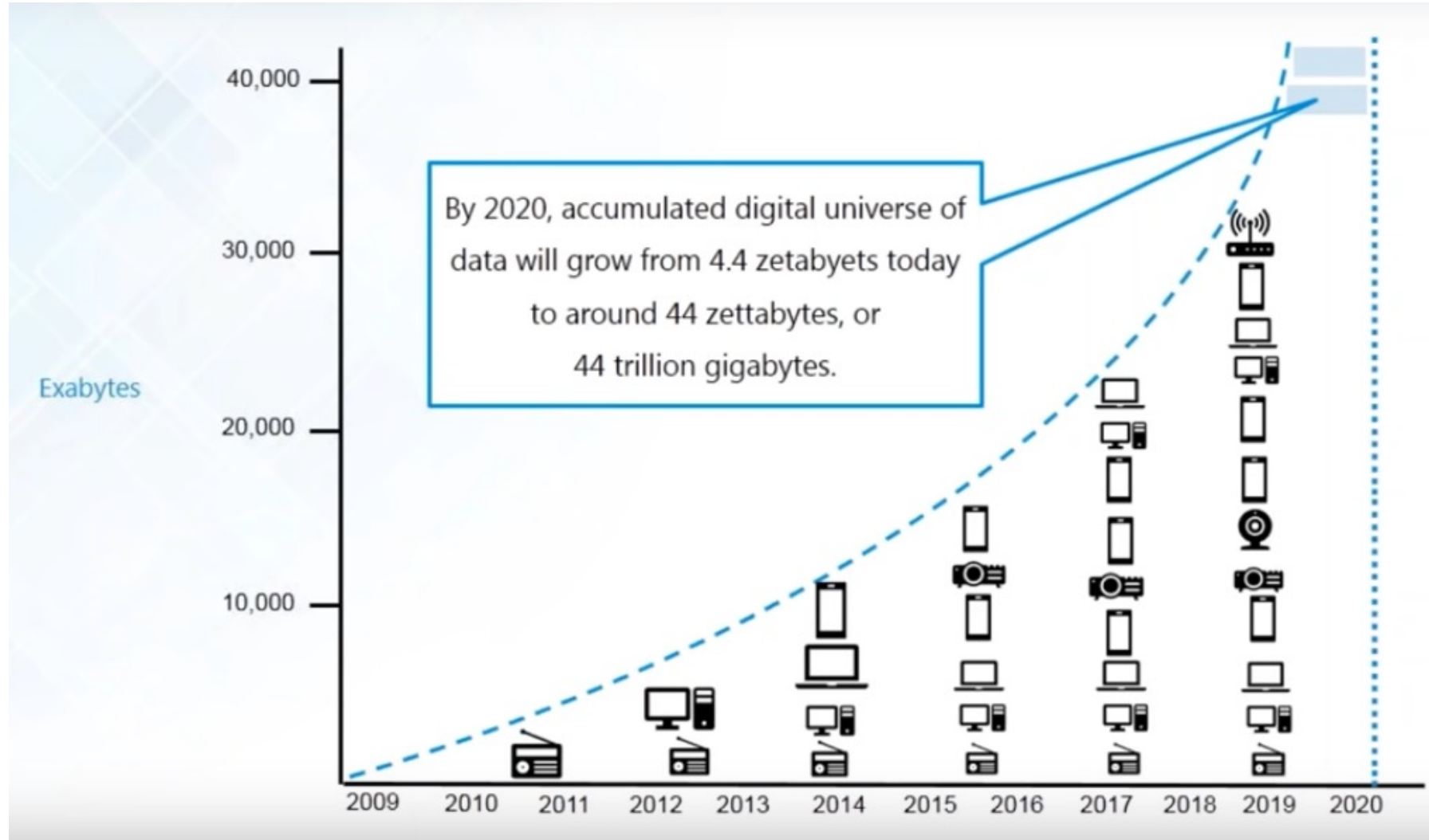
***You are at university library. You see few students browsing through the library catalog on kiosk. You see the working of librarians and other staff to issue/return books, magazines, and journals. Few students are using the e-library service, too. Which type of data is generated in this scenario? Support your answer by considering big data***

# Big Data – Definitional Aspects

# Characteristics of Big data



# Volume

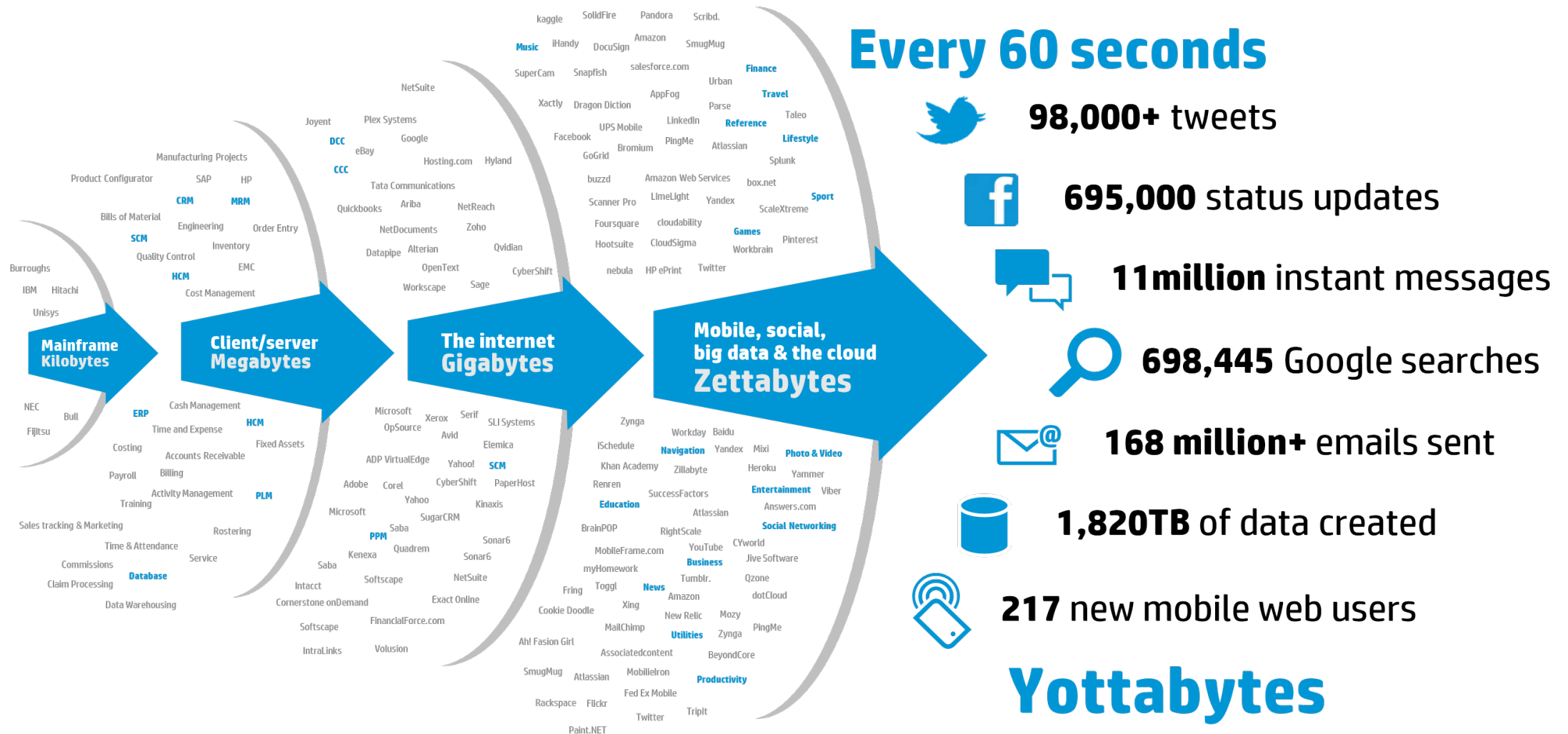


Byte Comparison Table		
Metric	Value	Bytes
Byte (B)	1	1
Kilobyte (KB)	$1,024^1$	1,024
Megabyte (MB)	$1,024^2$	1,048,576
Gigabyte (GB)	$1,024^3$	1,073,741,824
Terabyte (TB)	$1,024^4$	1,099,511,627,776
Petabyte (PB)	$1,024^5$	1,125,899,906,842,624
Exabyte (EB)	$1,024^6$	1,152,921,504,606,846,976
Zettabyte (ZB)	$1,024^7$	1,180,591,620,717,411,303,424
Yottabyte (YB)	$1,024^8$	1,208,925,819,614,629,174,706,176

class	size	manage with	how it fits	examples
<b>small</b>	< 10 GB	Excel, R	fits in one machine's memory	thousands of sales figures
<b>medium</b>	10GB-1TB	indexed files, monolithic DB	fits on one machine's disk	millions of web pages
<b>Big</b>	> 1TB	Hadoop, distributed DBs	stored across many machines	billions of web clicks

# Velocity

## Accelerating innovation and time to value



Taken from : Hewlett-Packard Development Company "truths and myths about big data",2013

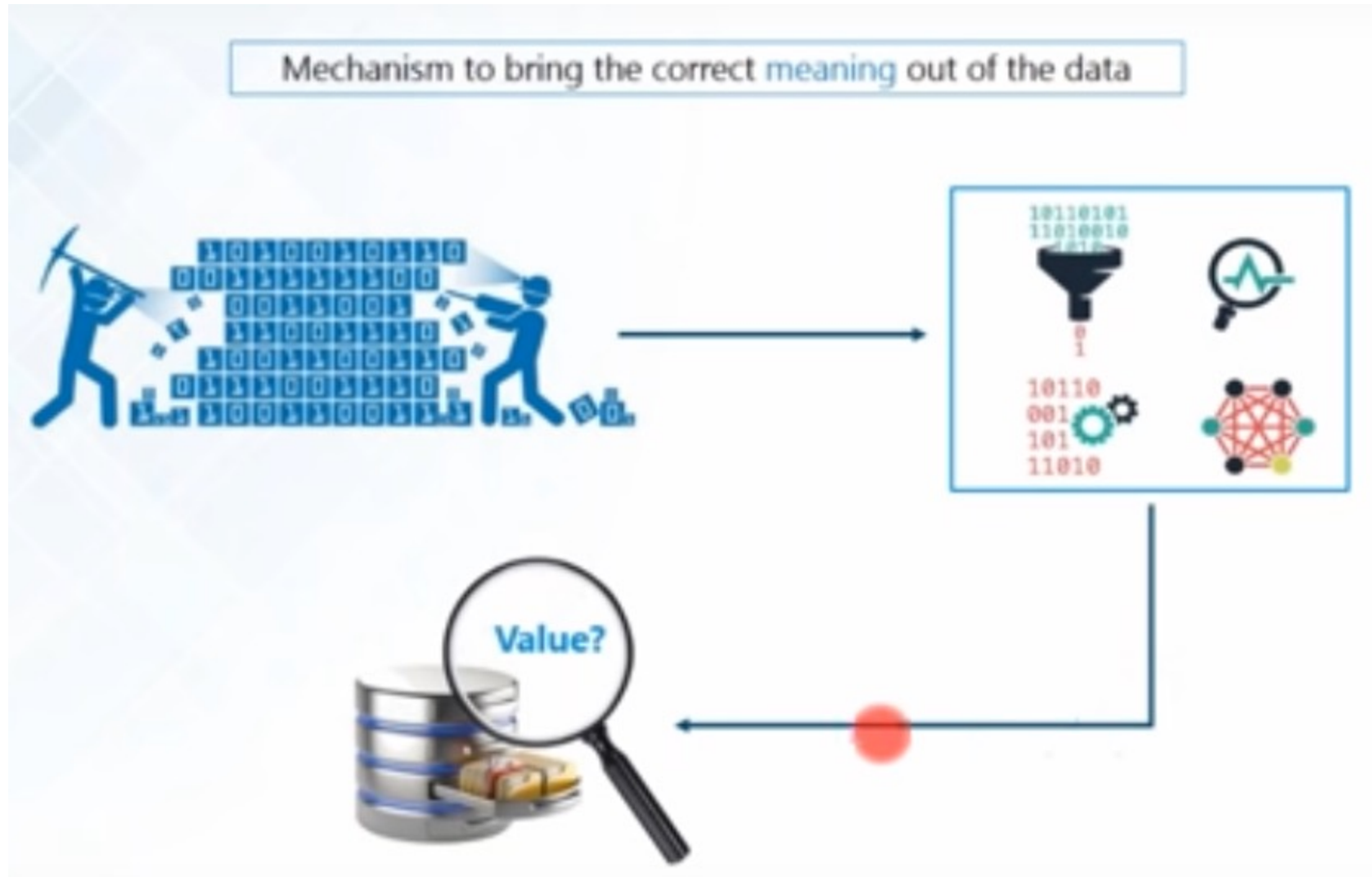


# Veracity

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

Uncertainty and inconsistencies in the data

# Value



# What is big data about?

Answers are often “too big to ....”

- Load into memory.....Store on a hard drive.....Fit in a standard database
- “Fast changing”.....Not just relational
- “Digital breadcrumbs” left behind (communication transactions..) —Hard little data particles left behind as people go about their daily lives
- Open web data/social media data (facebook, twitter, blogs, online news, videos....)
- Remote sensing (satellite, meters...)

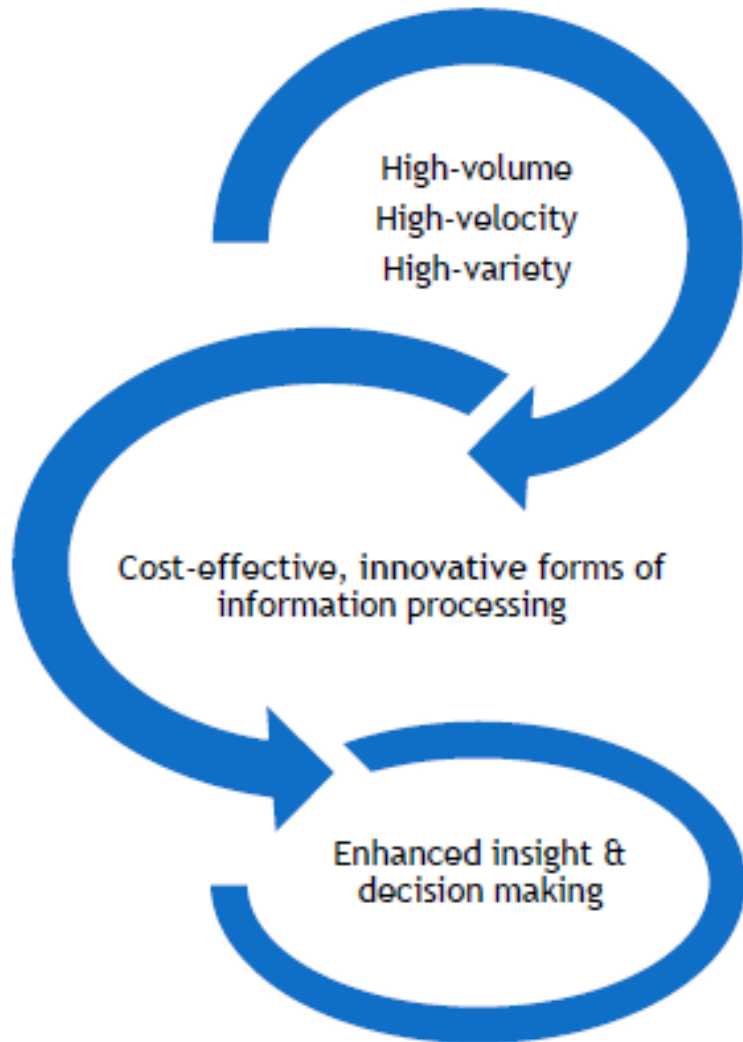
# What is big data about - and not about?

*“Big Data is not about the data”* (Gary King)

Institute for social science ,Harvard university

- It's about the analytics—the insights gleaned from the data; and the necessary capacities to do so—human, technological
- One step further: it's about knowledge: getting near to the 'true' meaning of a facebook status update;
- It's about sharing and diffusion – visualizations

# Big data Definition




*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


# Types of Big Data

## **Types of Big Data**


 Relational Data (Tables/Transaction/Legacy Data)

 Text Data (Web)

 Semi-structured Data (XML)

 Graph Data

- Social Network
- Semantic Web (RDF-Resource Description Framework)

 Streaming Data

- You can only scan the data once

# Challenges with Big data

**Problem 1:** Storing exponentially growing huge datasets

- Data generated in past **2 years** is more than the previous history in total
- By 2020, total digital data will grow to **44 Zettabytes** approximately
- By 2020, about **1.7 MB** of new info will be created every second for every person

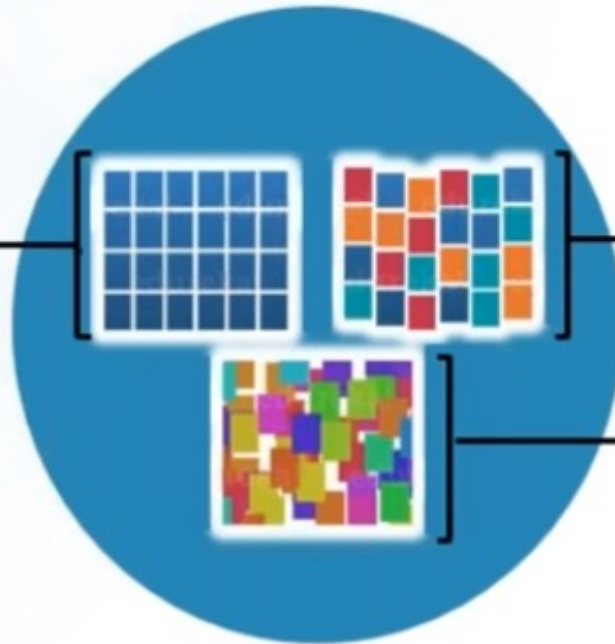




**Problem 2:** Processing data having complex structure

**Structured**

- Organized data format
- Data schema is fixed
- Ex: RDBMS data, etc.



**Semi – Structured**

- Partial organized data
- Lacks formal structure of a data model
- Ex: XML & JSON files, etc.

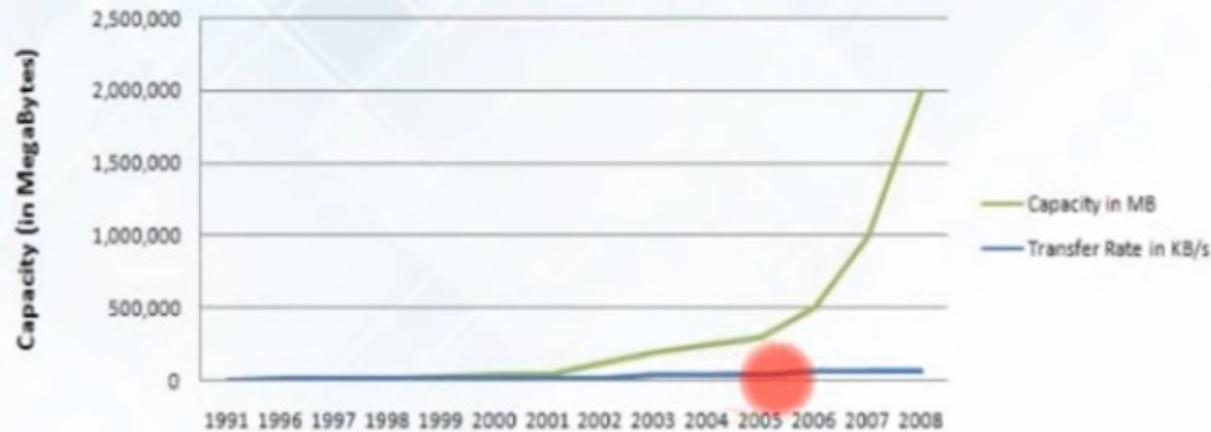
**Unstructured**

- Un-organized data
- Unknown schema
- Ex: multi-media files, etc.

**Problem 3: Processing data faster**

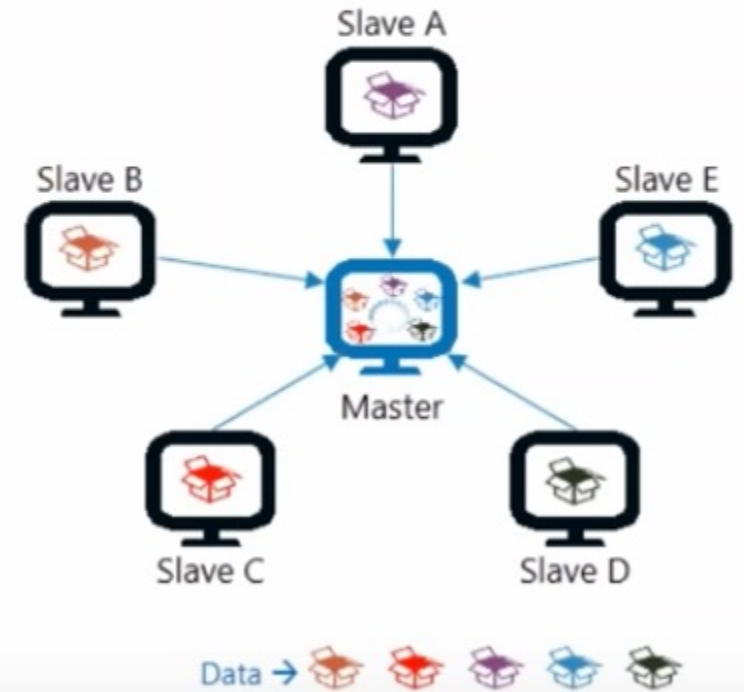
The data is growing at much faster rate than that of disk read/write speed

**Relative Improvement  
Hard Disk Capacity v.s. Disk Transfer Performance**



**Source:** Tom's Hardware

Bringing huge amount of data to computation unit becomes a bottleneck



# Classification of Data Analytics

Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
 <p><b>“What happened”</b></p> <ul style="list-style-type: none"><li>• Provides insights into past events</li></ul>	 <p><b>“Why did it happen”</b></p> <ul style="list-style-type: none"><li>• Takes the insights from descriptive analytics to dig deeper to find the cause of the outcome</li></ul>	 <p><b>“What will happen next”</b></p> <ul style="list-style-type: none"><li>• Leverages historical data and trends to predict future outcomes</li></ul>	 <p><b>“What should be done about it”</b></p> <ul style="list-style-type: none"><li>• Analyzes past decisions and events to estimate the likelihood of different outcomes</li></ul>

# Big data Analytics-Case studies

- Healthcare



# Traditional Vs Big data Approach

❖ **OLTP:** Online Transaction Processing

- DBMSs

❖ **OLAP:** Online Analytical Processing

- Data Warehousing

❖ **RTAP:** Real-Time Analytics Processing

- Big Data Architecture & Technology

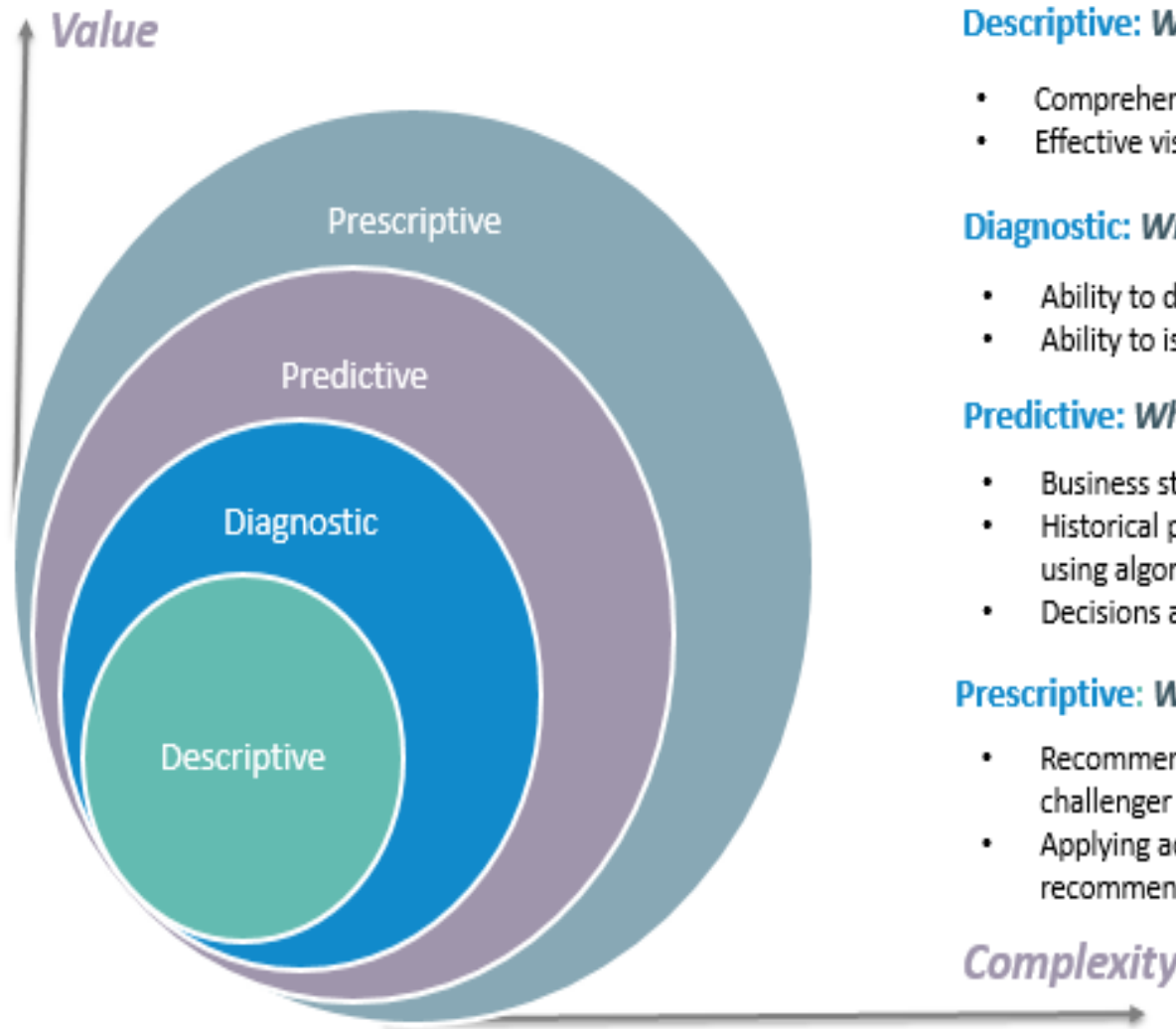
# Use cases

- Online money transfer
- Iterative computations
- Joins and aggregations computations
- Ad-hoc computations
- Customer churn analytics
- Customer segmentation for optimized offers
- Processing of large amount of small files
- Security applications



# Data analytics

## 4 types of Data Analytics



### What is the data telling you?

**Descriptive:** *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

**Diagnostic:** *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

**Predictive:** *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

**Prescriptive:** *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

# Use cases

- **Use Case:** A Company wants to analyse the effective utilization of office and lab resources.
- **Use Case:** A E-commerce company wants to identify the reason behind the sudden fall in sales during last month for a particular product.
- **Use Case:** The PAY-PAL company wants to protect their clients against fraudulent transactions.
- **Use case:** An airline company wants to maximize the airline profit.



# RDBMS vs. Hadoop

RDBMS	Hadoop
Traditional row-column based DB, used for data storage, manipulation and retrieval.	An open-source software used for storing data and running applications or processes concurrently.
Structured data is mostly processed.	Both structured and unstructured data is processed.
Best suited for OLTP environment.	Best suited for BIG data.
Less scalable than Hadoop.	Highly scalable.
Data normalization is required.	Data normalization is not required.
Stores transformed and aggregated data.	Stores huge volume of data.
No latency in response.	Some latency in response.
Data schema of RDBMS is static.	Data schema of Hadoop is dynamic.
High data integrity available.	Low data integrity available than RDBMS.
Cost is applicable for licensed software.	Free of cost, as it is an open source software.

# RDBMS vs. MapReduce

	Traditional RDBMS	MapReduce
Data size	Gigabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Transactions	ACID	None
Structure	Schema-on-write	Schema-on-read
Integrity	High	Low
Scaling	Nonlinear	Linear