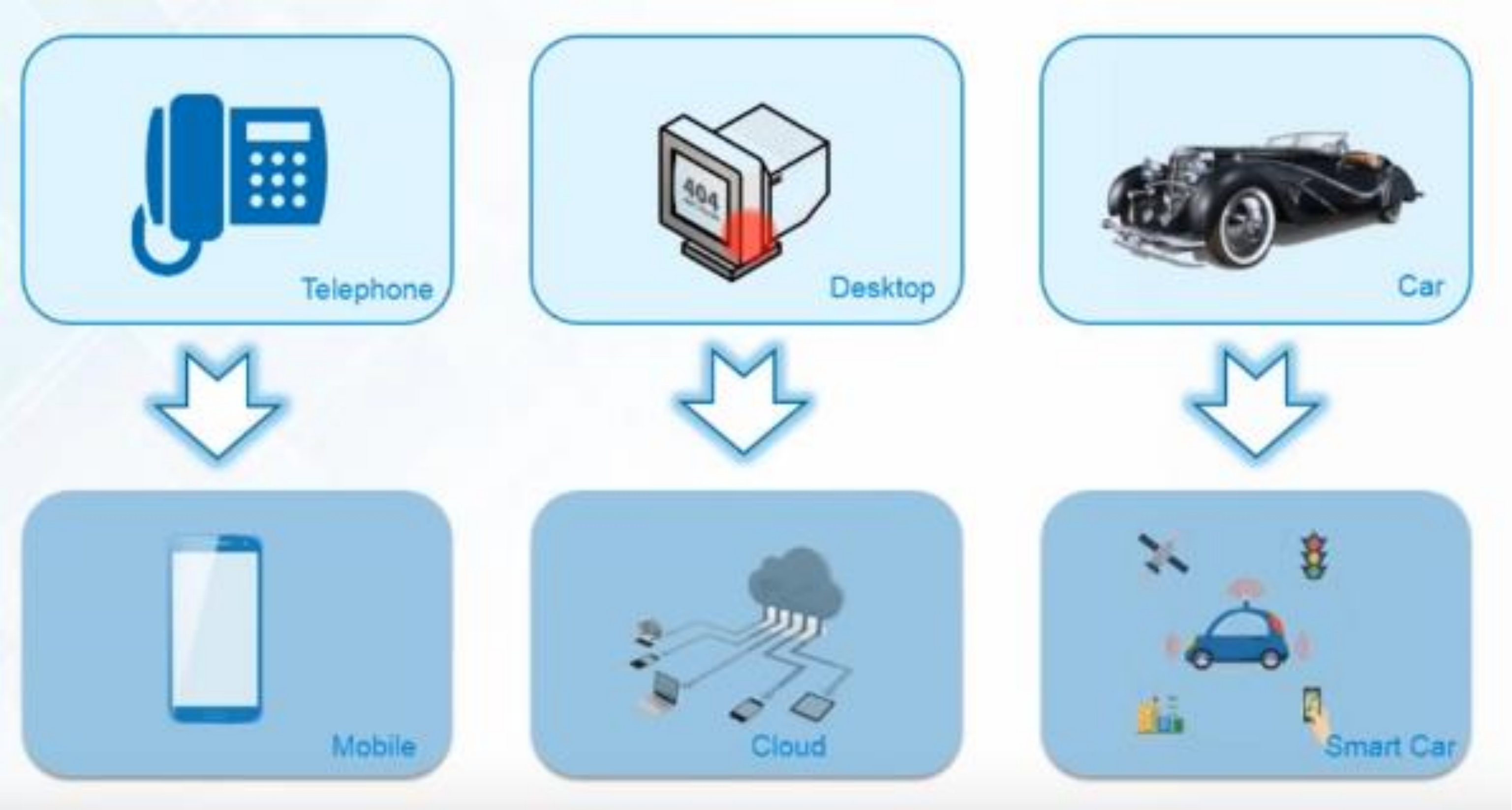
Road Map

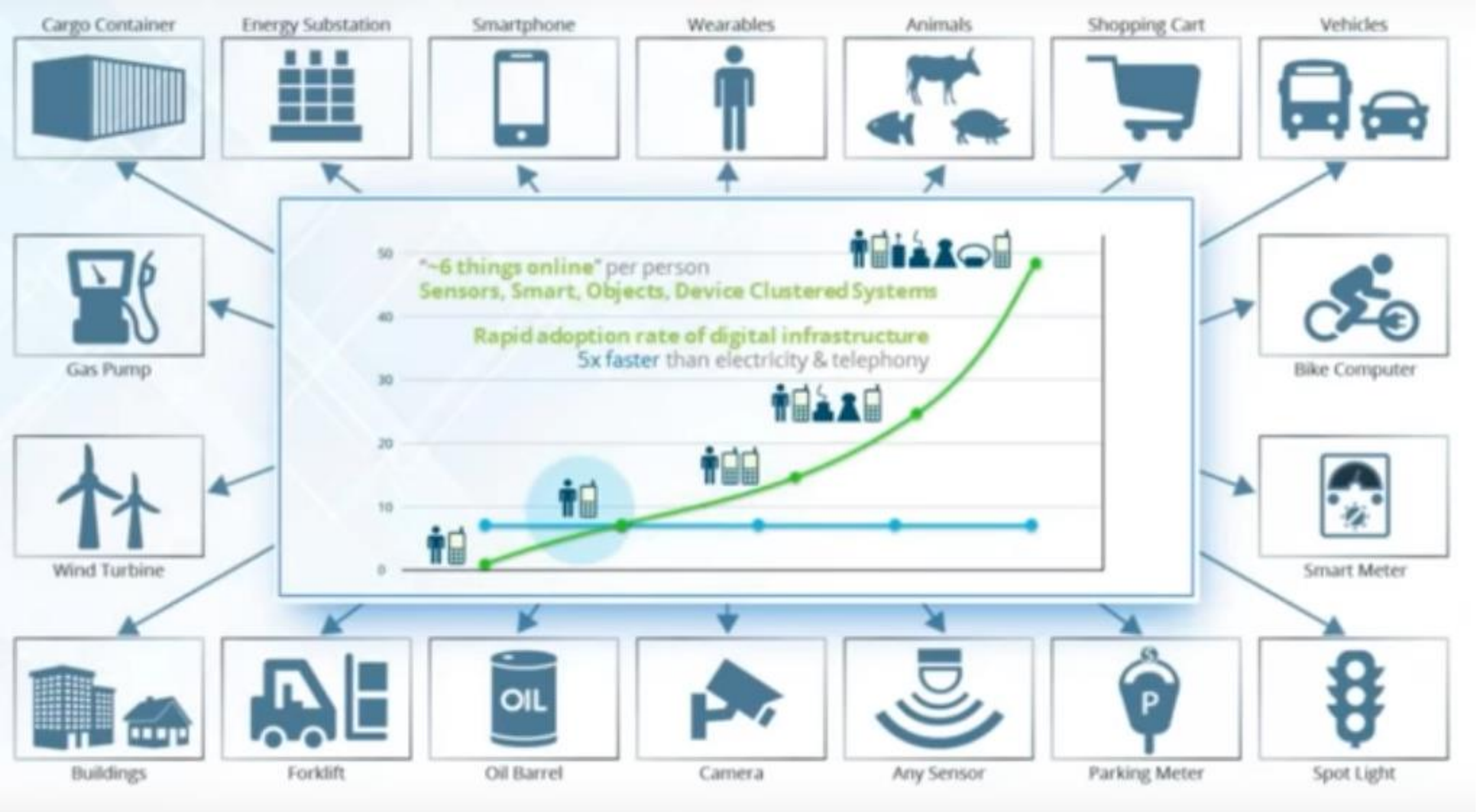
- Evolution of Technology
- Types of Data
- Big Data- Definition Aspect
- Big data Vs Not Big data
- Challenges of big data

The data deluge

Evolution of Technology

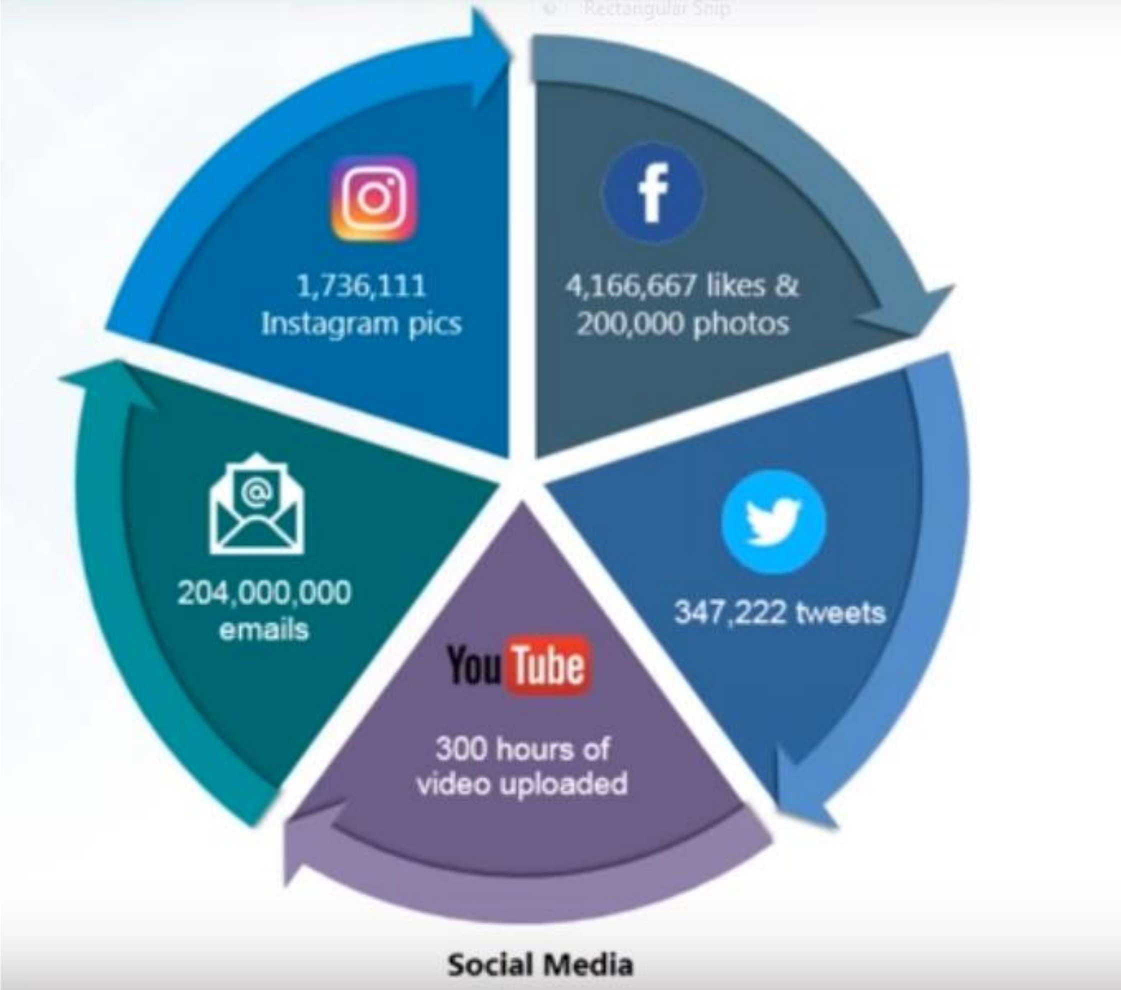


Internet of Things

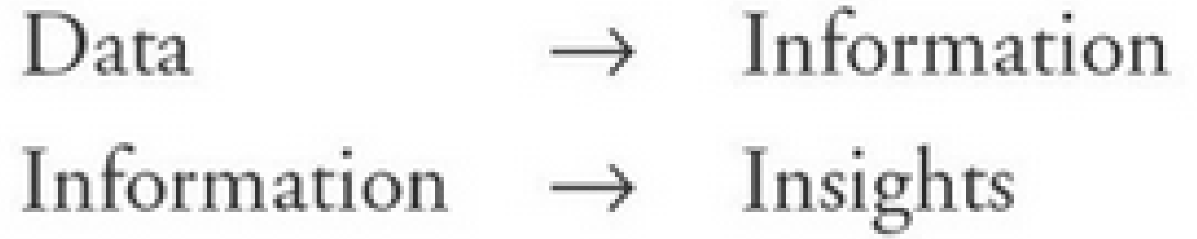


IOT: 50 Billion devices by 2020

Social Media Usage



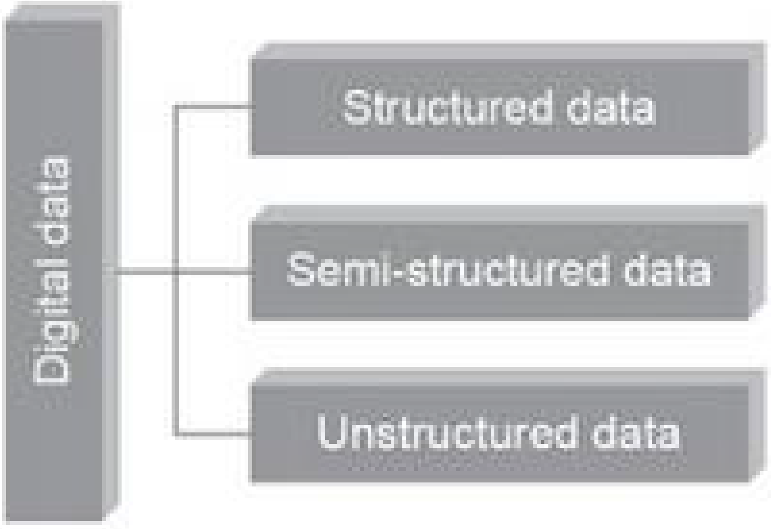
Classification of Digital Data



Data → Information

Information → Insights

Digital Data



Structured data

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

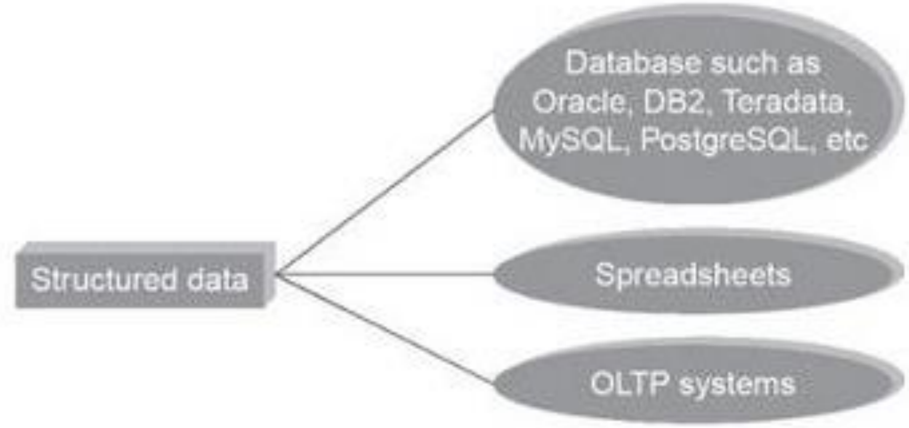
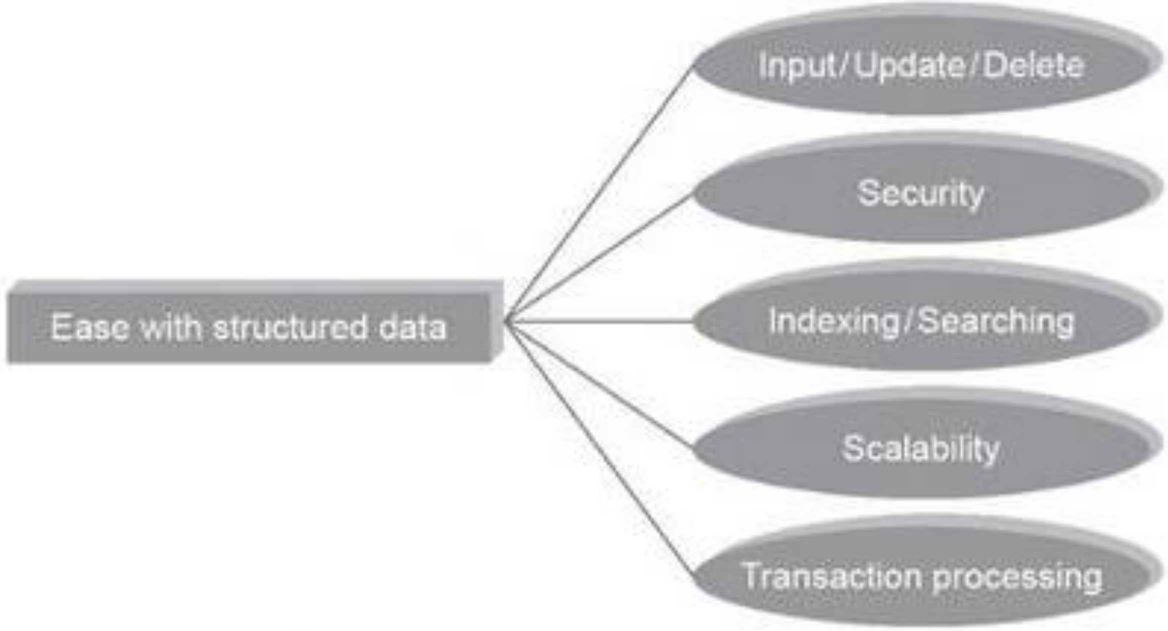

Structured data
Database such as Oracle, DB2, Teradata, MySQL, PostgreSQL, etc
Spreadsheets
OLTP systems

Figure 1.4 Sources of structured data.

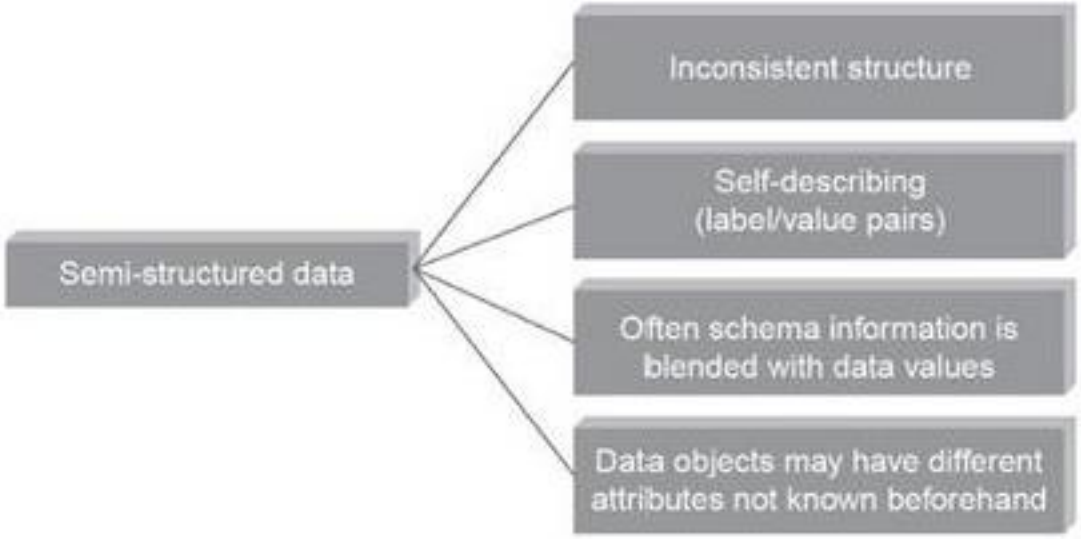
Working with structured data

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

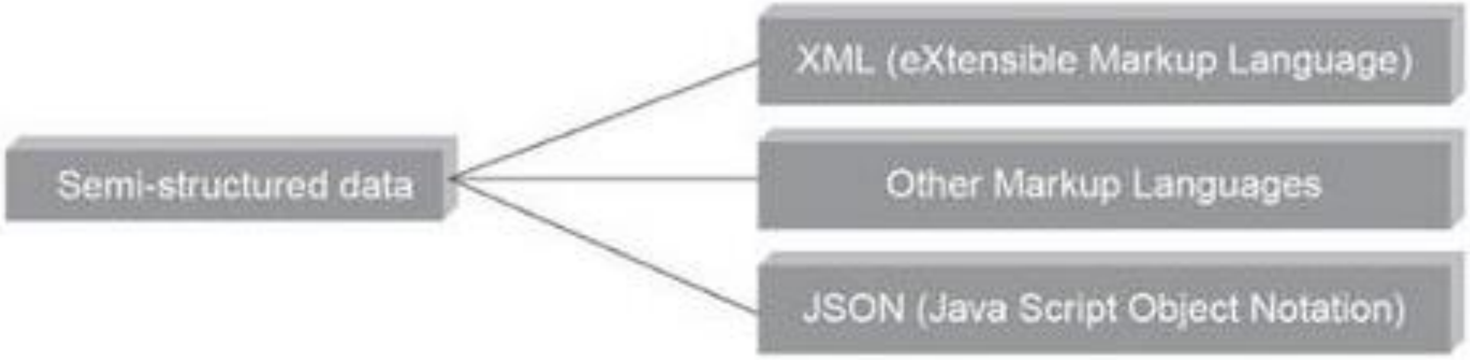


Semi-structured data

- 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


Semi-structured data
Inconsistent structure
Self-describing (label/value pairs)
Often schema information is blended with data values
Data objects may have different attributes not known beforehand

Sources of semi-structured data



Semi-structured data

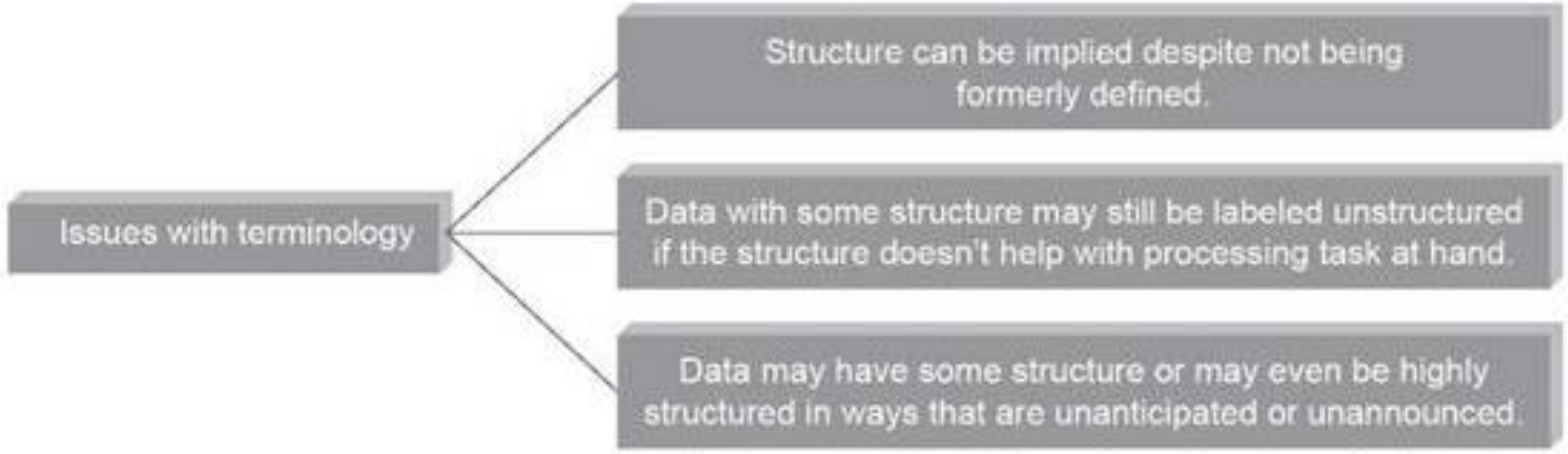
XML (eXtensible Markup Language)

Other Markup Languages

JSON (Java Script Object Notation)

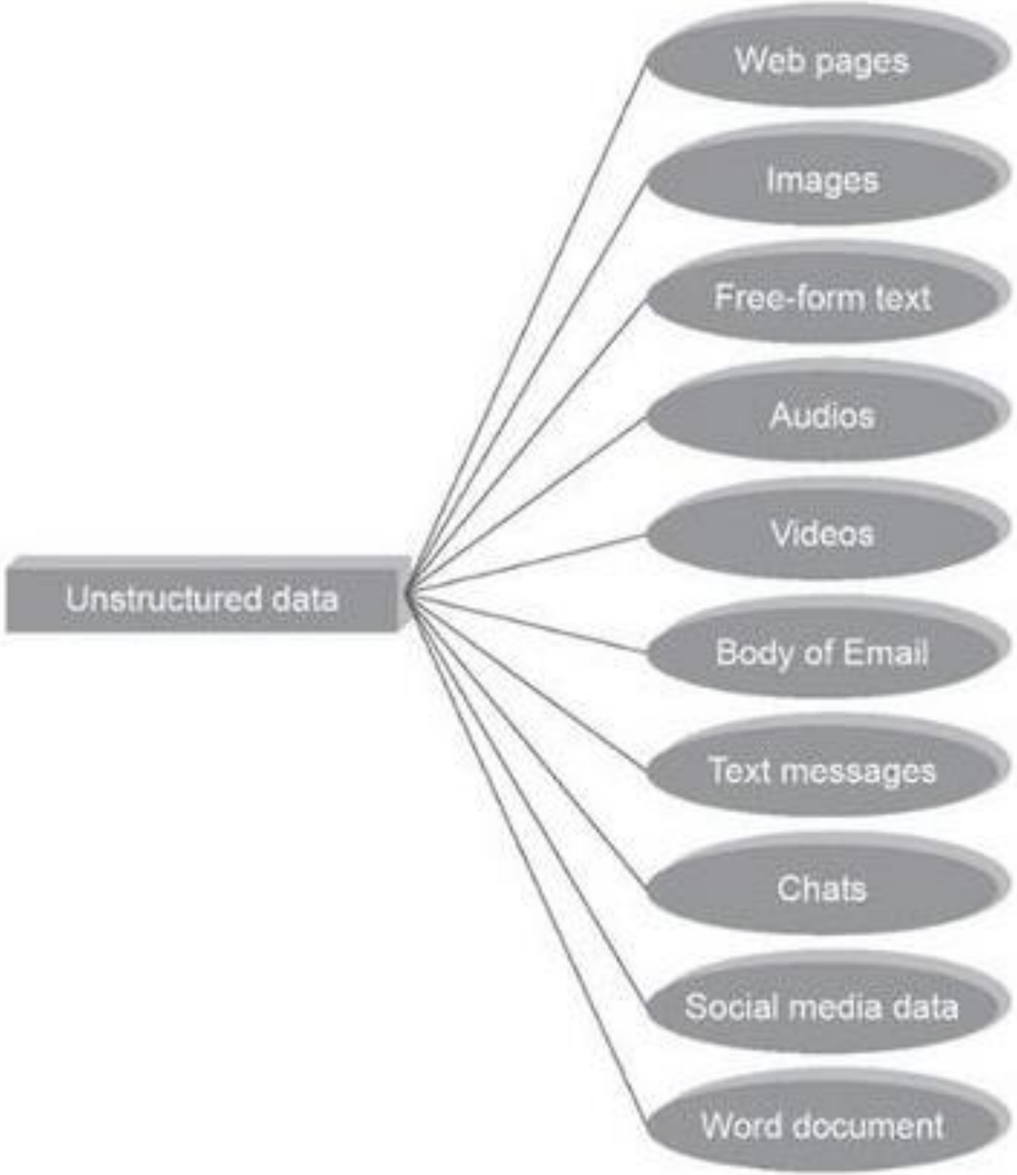
Unstructured data

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



Issues with terminology
Structure can be implied despite not being formerly defined.
Data with some structure may still be labeled unstructured if the structure doesn't help with processing task at hand.
Data may have some structure or may even be highly structured in ways that are unanticipated or unannounced.

Sources of unstructured data



Unstructured data

Web pages

Images

Free-form text

Audios

Videos

Body of Email

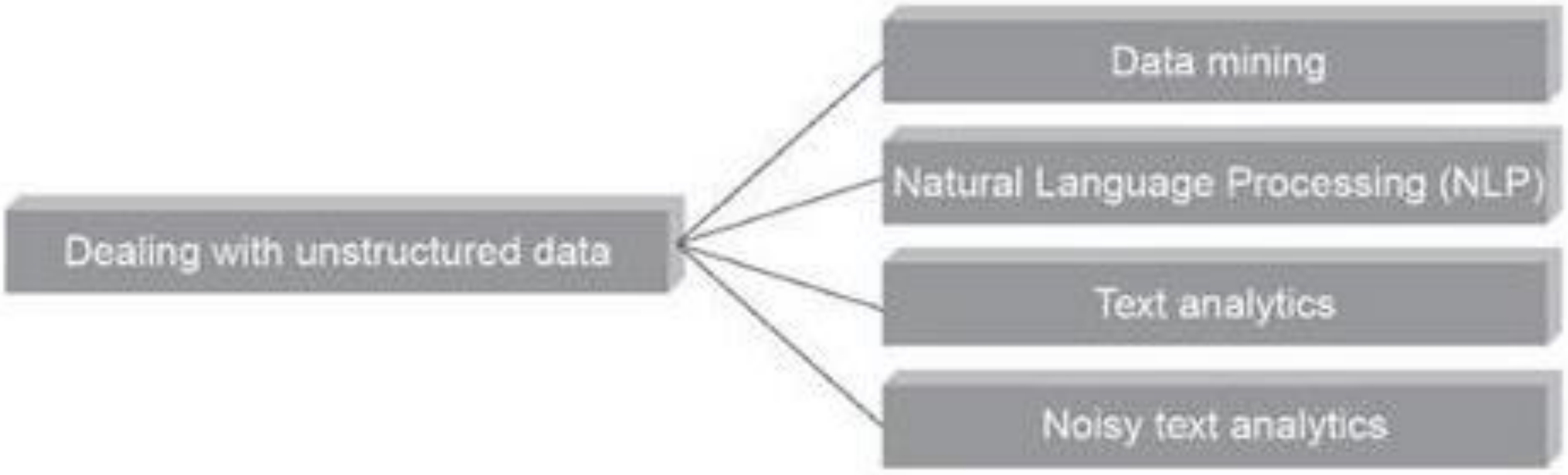
Text messages

Chats

Social media data

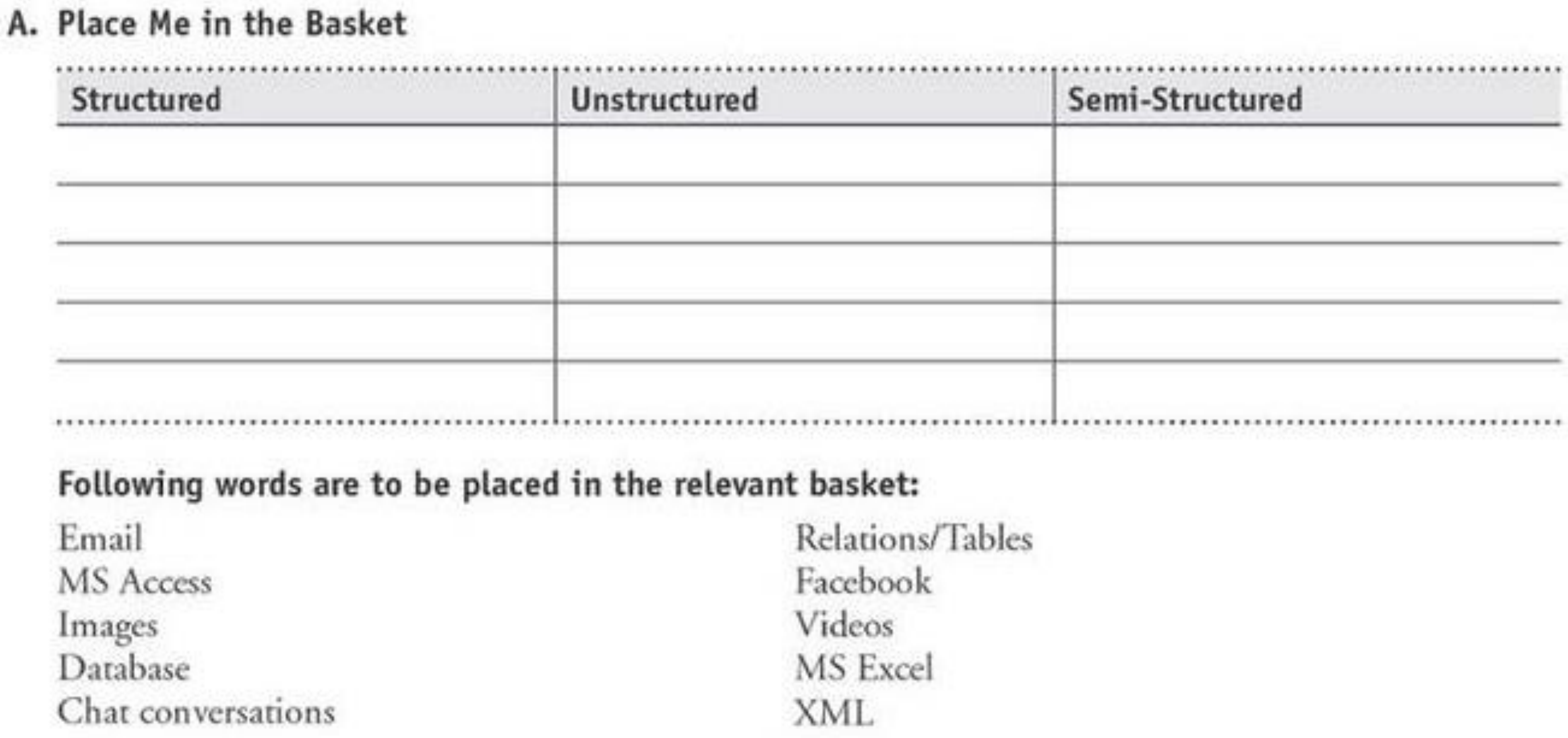
Word document

How to deal with unstructured data?



Dealing with unstructured data
Data mining
Natural Language Processing (NLP)
Text analytics
Noisy text analytics

Inclass#exercise



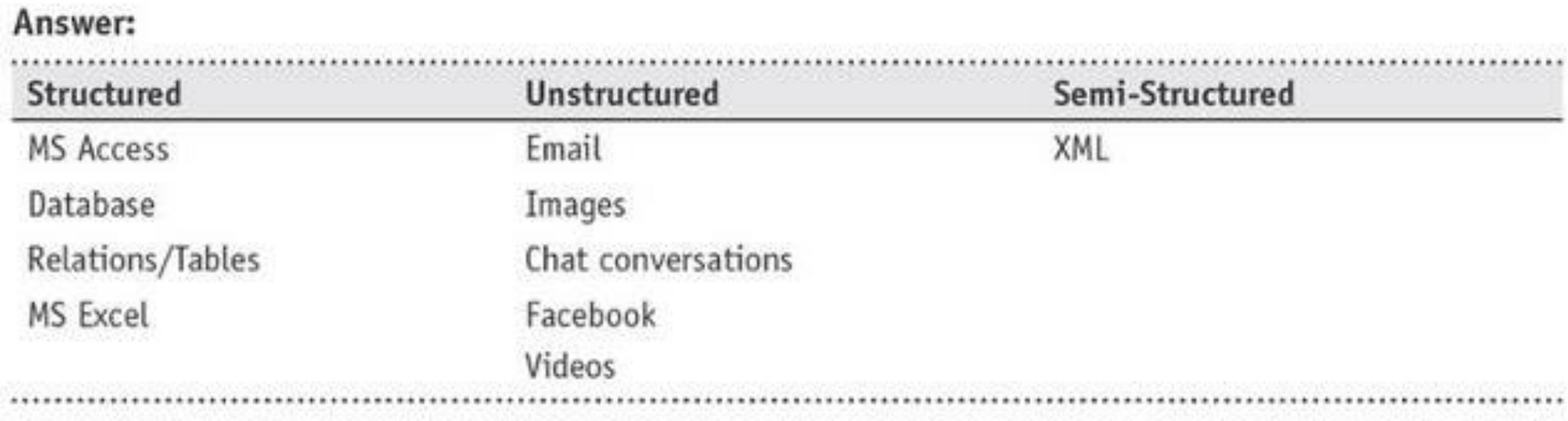
A. Place Me in the Basket

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

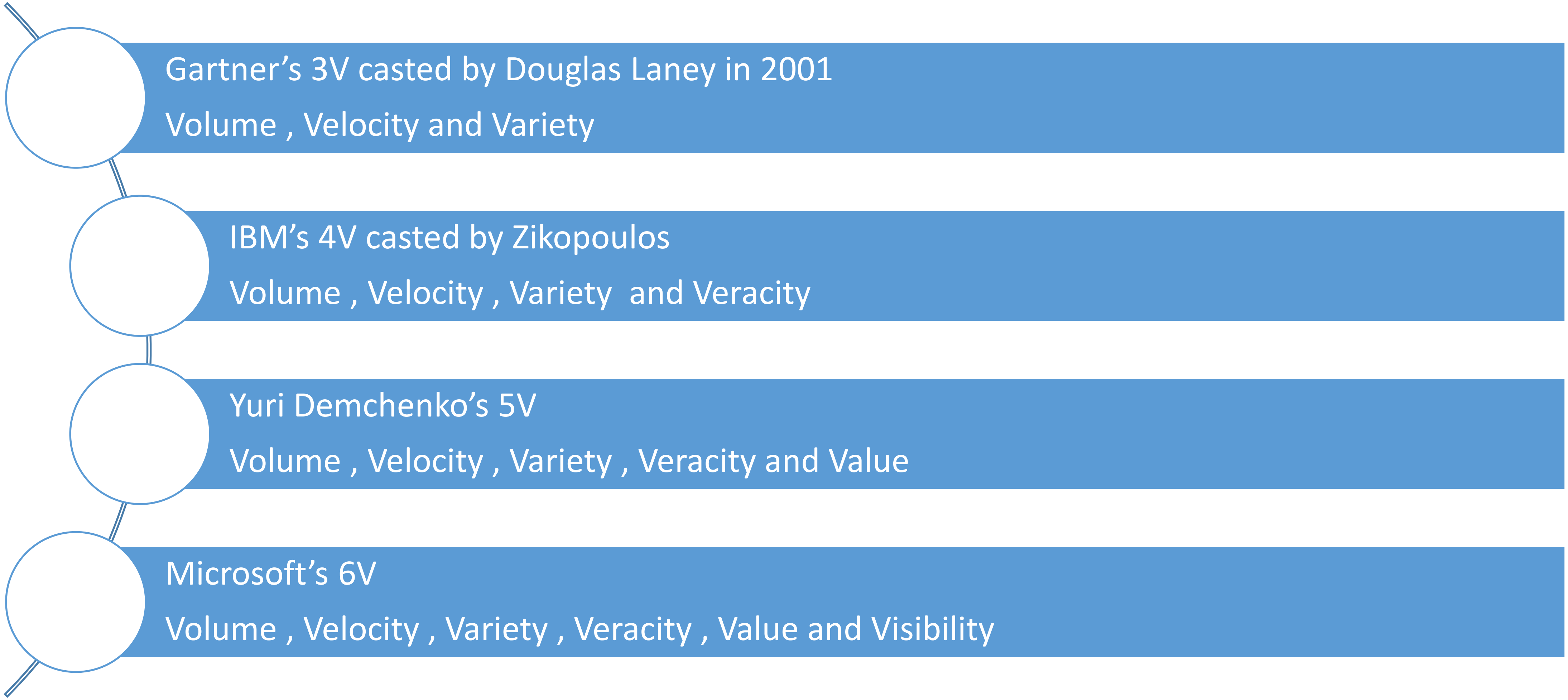
- Why email in unstructured category?
- Where should we put CCTV footage?

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



Gartner’s 3V casted by Douglas Laney in 2001

Volume , Velocity and Variety

IBM’s 4V casted by Zikopoulos

Volume , Velocity , Variety and Veracity

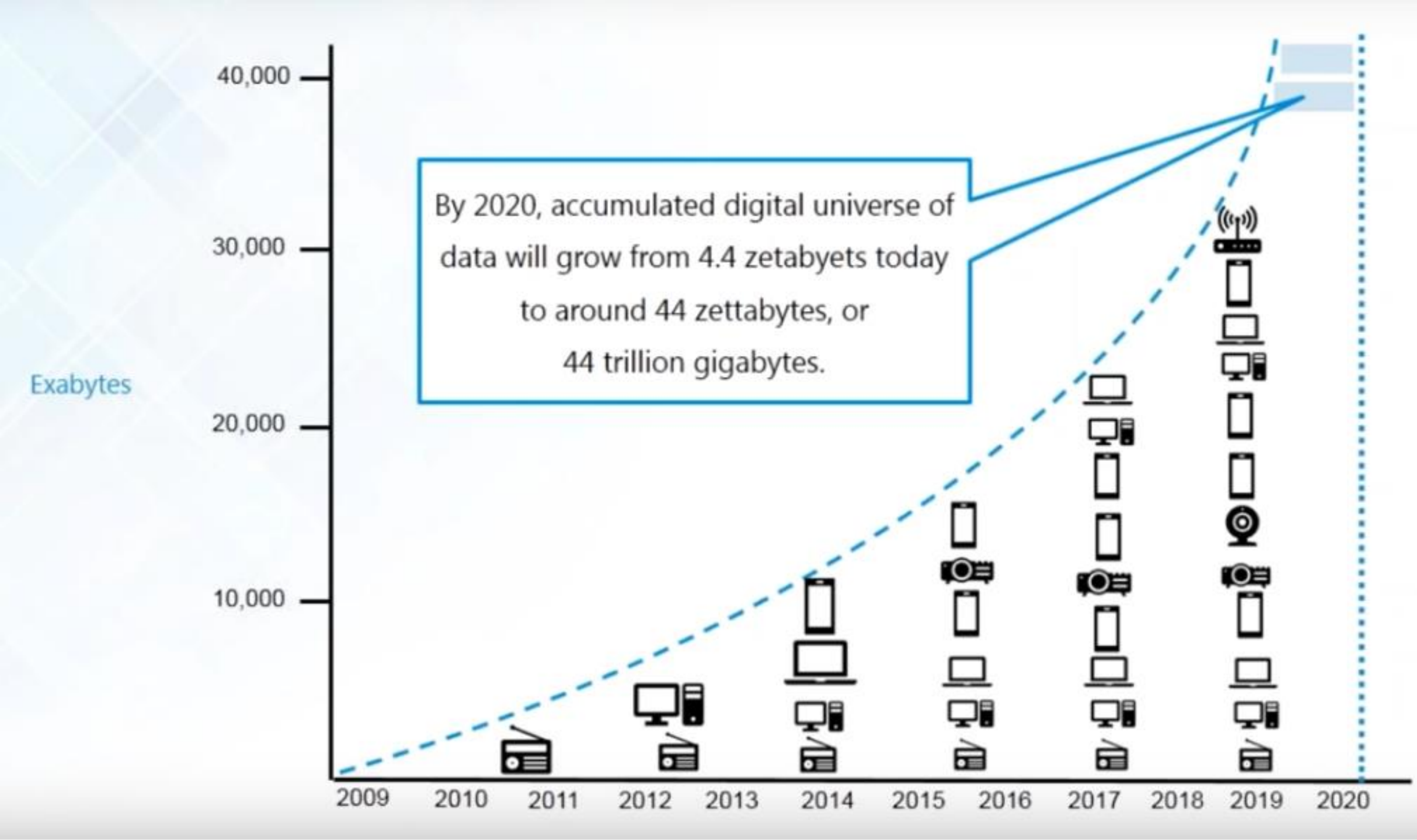
Yuri Demchenko’s 5V

Volume , Velocity , Variety , Veracity and Value

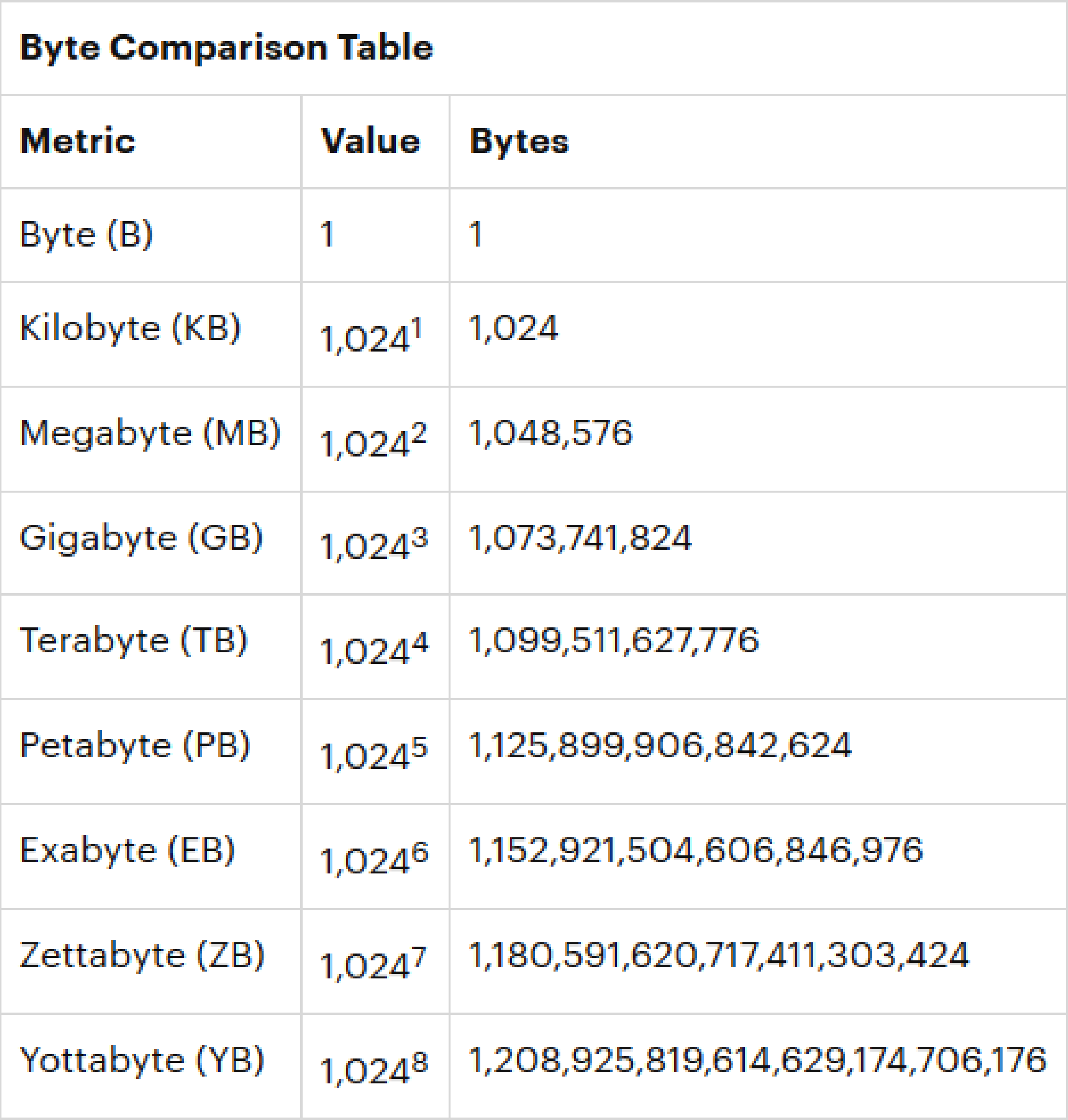
Microsoft’s 6V

Volume , Velocity , Variety , Veracity , Value and Visibility

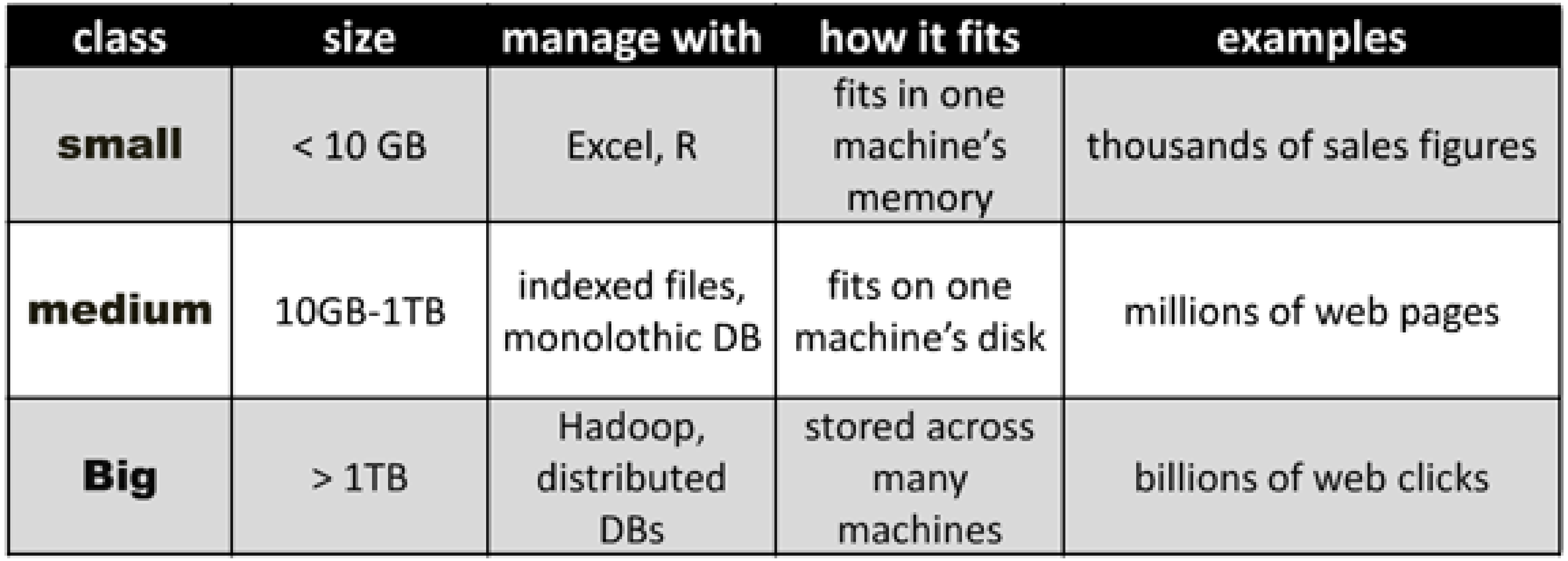
Volume



By 2020, accumulated digital universe of data will grow from 4.4 zetabyets today to around 44 zettabytes, or 44 trillion gigabytes.



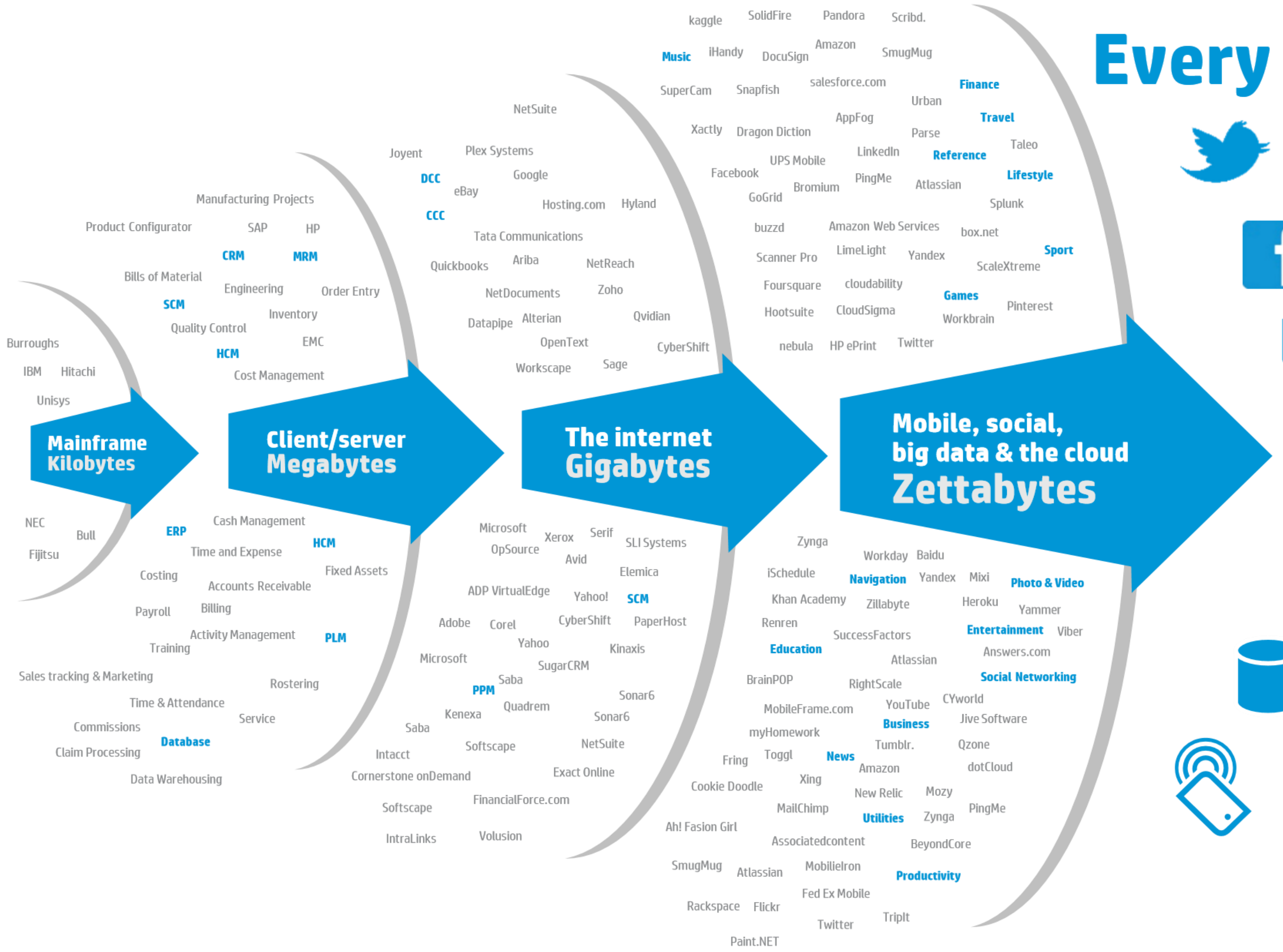
Byte Comparison Table		
Metric	**Value**	**Bytes**
Byte (B)	1	1
Kilobyte (KB)	1,024^{1}	1,024
Megabyte (MB)	1,024^{2}	1,048,576
Gigabyte (GB)	1,024^{3}	1,073,741,824
Terabyte (TB)	1,024^{4}	1,099,511,627,776
Petabyte (PB)	1,024^{5}	1,125,899,906,842,624
Exabyte (EB)	1,024^{6}	1,152,921,504,606,846,976
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
small	< 10 GB	Excel, R	fits in one machine's memory	thousands of sales figures
medium	10GB-1TB	indexed files, monolothic DB	fits on one machine's disk	millions of web pages
Big	> 1TB	Hadoop, distributed DBs	stored across many machines	billions of web clicks

Velocity

Accelerating innovation and time to value



Burroughs
IBM Hitachi
Unisys

Mainframe
Kilobytes

NEC Bull
Fijitsu

Manufacturing Projects
Product Configurator SAP HP
CRM MRM
Bills of Material Engineering Order Entry
SCM Inventory
Quality Control EMC
HCM
Cost Management

Client/server
Megabytes

ERP Cash Management
Time and Expense HCM
Fixed Assets
Costing Accounts Receivable
Payroll Billing
Activity Management PLM
Training
Sales tracking & Marketing Rostering
Time & Attendance
Service
Commissions
Database
Claim Processing
Data Warehousing

NetSuite
Joyent Plex Systems
DCC Google
eBay
Hosting.com Hyland
CCC
Tata Communications
Quickbooks Ariba NetReach
NetDocuments Zoho
Datapipe Alterian Qvidian
OpenText CyberShift
Workscape Sage

The internet
Gigabytes

Microsoft Xerox Serif SLI Systems
OpSource Avid
Elemica
ADP VirtualEdge Yahoo! SCM
Adobe Corel CyberShift PaperHost
Yahoo Kinaxis
Microsoft SugarCRM
Saba
PPM Sonar6
Kenexa Quadrem
Saba Sonar6
Softscape NetSuite
Intacct
Cornerstone onDemand Exact Online
Softscape FinancialForce.com
IntraLinks Volusion

kaggle SolidFire Pandora Scribd.
Music iHandy DocuSign Amazon SmugMug
SuperCam Snapfish salesforce.com Finance
Urban
AppFog Travel
Xactly Dragon Diction Parse
Taleo
UPS Mobile LinkedIn Reference
Facebook PingMe Atlassian Lifestyle
Bromium
GoGrid Splunk
buzzd Amazon Web Services box.net
Scanner Pro LimeLight Yandex Sport
ScaleXtreme
Foursquare cloudability
Games
Hootsuite CloudSigma Pinterest
Workbrain
nebula HP ePrint Twitter

Mobile, social,
big data & the cloud
Zettabytes

Zynga
Workday Baidu
iSchedule Navigation Yandex Mixi Photo & Video
Khan Academy Zillabyte Heroku Yammer
Renren Entertainment Viber
SuccessFactors
Education Answers.com
Atlassian
BrainPOP RightScale Social Networking
CYworld
MobileFrame.com YouTube
Business Jive Software
myHomework
Tumblr. Qzone
Fring Toggl News
Amazon dotCloud
Xing
Cookie Doodle New Relic Mozy
MailChimp PingMe
Utilities Zynga
Ah! Fasion Girl
Associatedcontent BeyondCore
SmugMug Atlassian MobilieIron Productivity
Fed Ex Mobile
Rackspace Flickr
TripIt
Twitter
Paint.NET

Every 60 seconds

98,000+ tweets

695,000 status updates

11million instant messages

698,445 Google searches

168 million+ emails sent

1,820TB of data created

217 new mobile web users

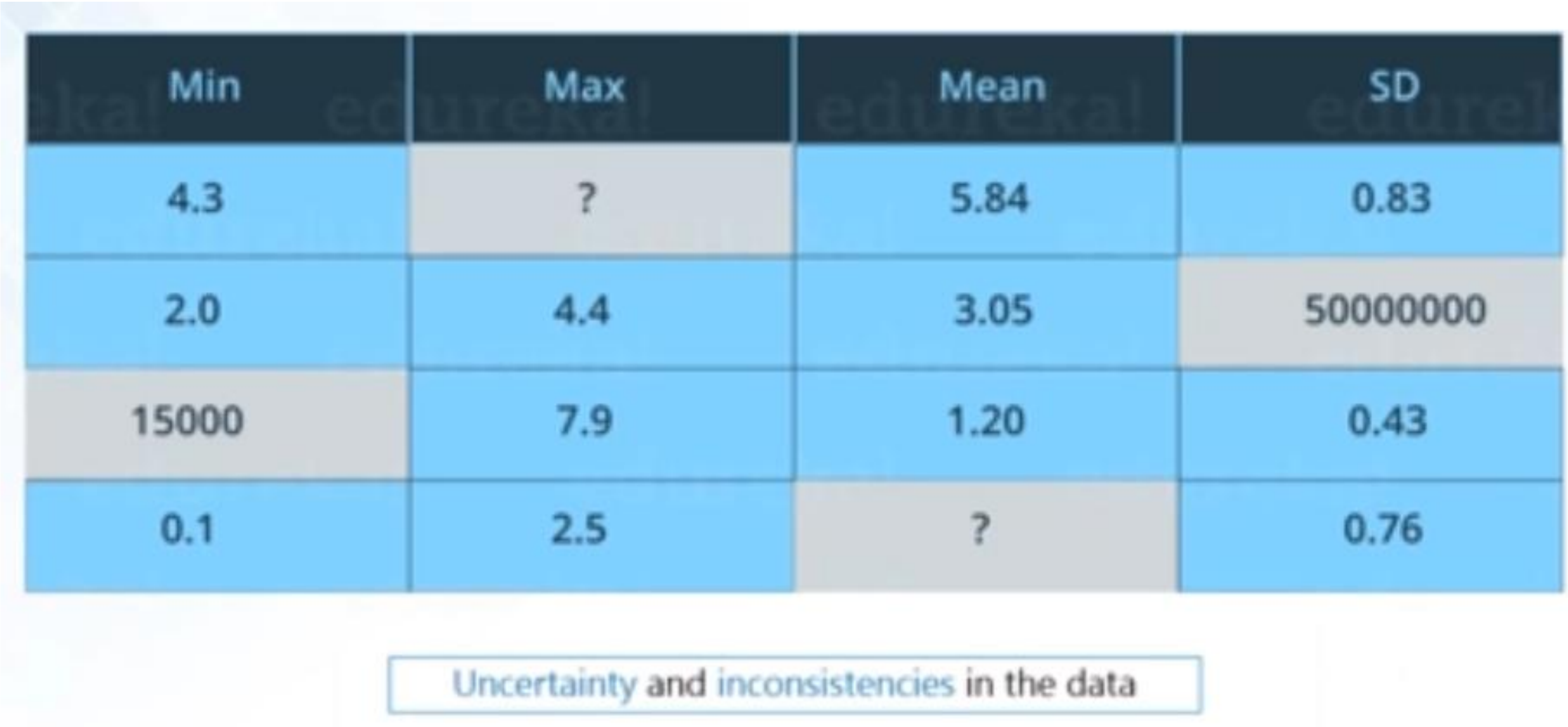
Yottabytes

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

Variety

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

Veracity

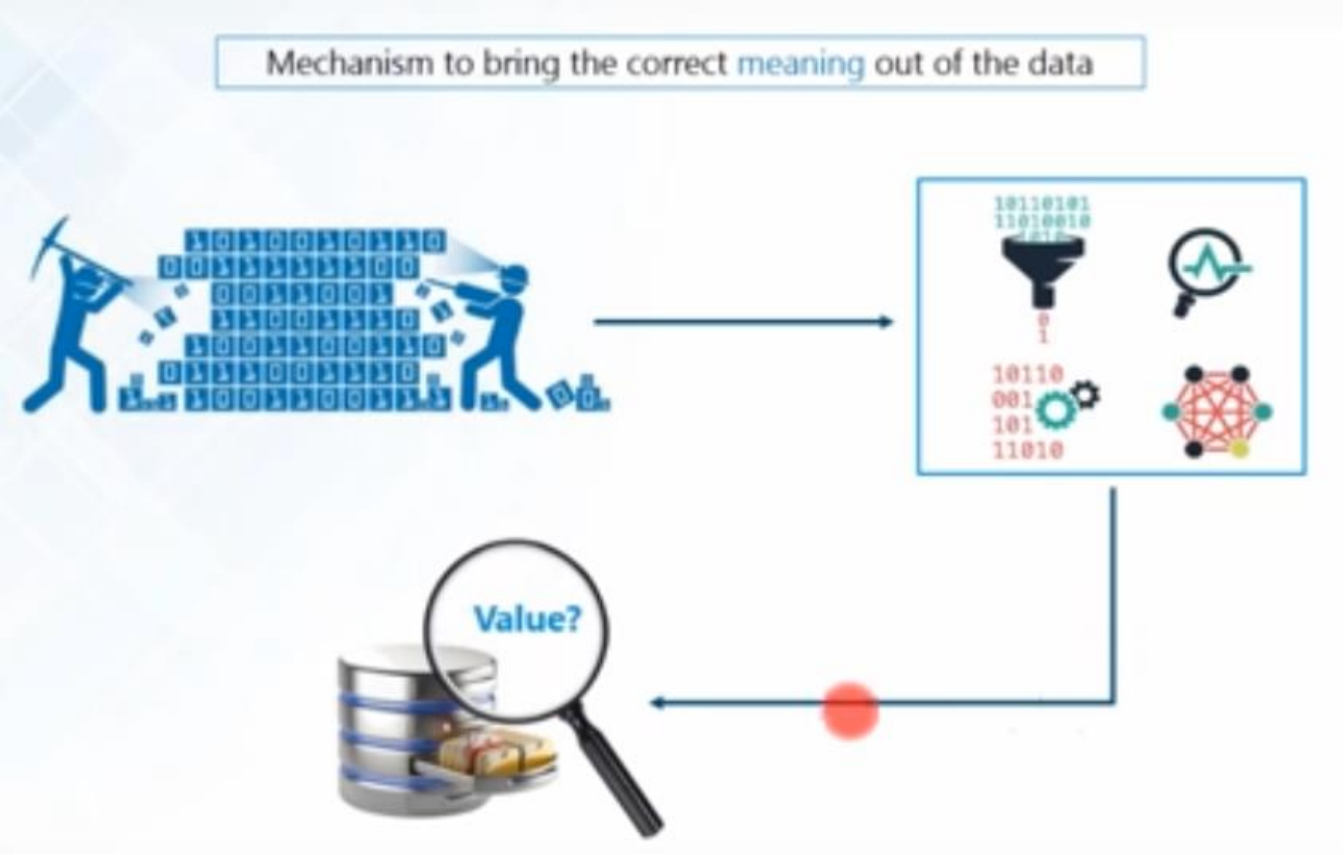


Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

Uncertainty and inconsistencies in the data

Value

Data by itself, regardless of its volume, usually isn’t very useful — to be valuable, it needs to be converted into insights or information, and that is where data processing steps in. By using custom processing software, you can derive useful insights from gathered data, and that can add value to your decision-making process.



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…)

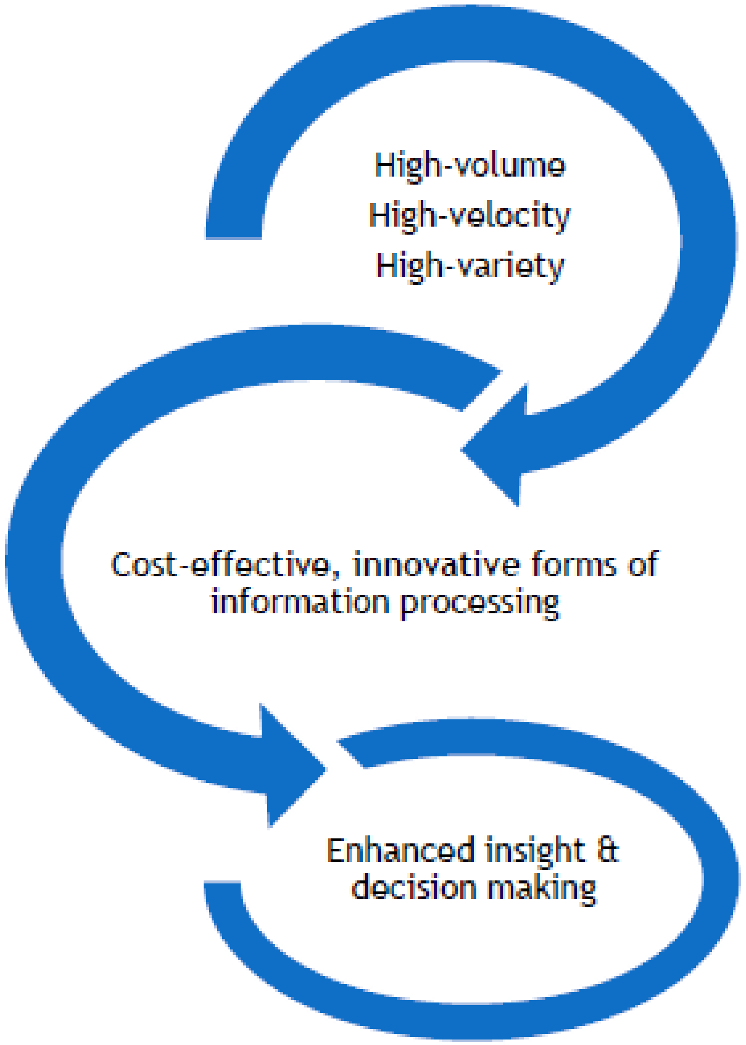
What is big data about - and not about?

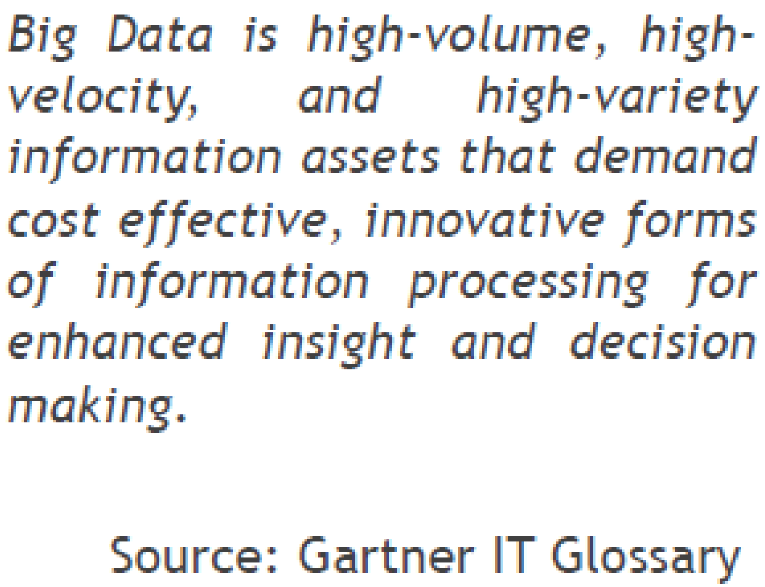
“*Big Data is not about the data*” (Gary King)

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

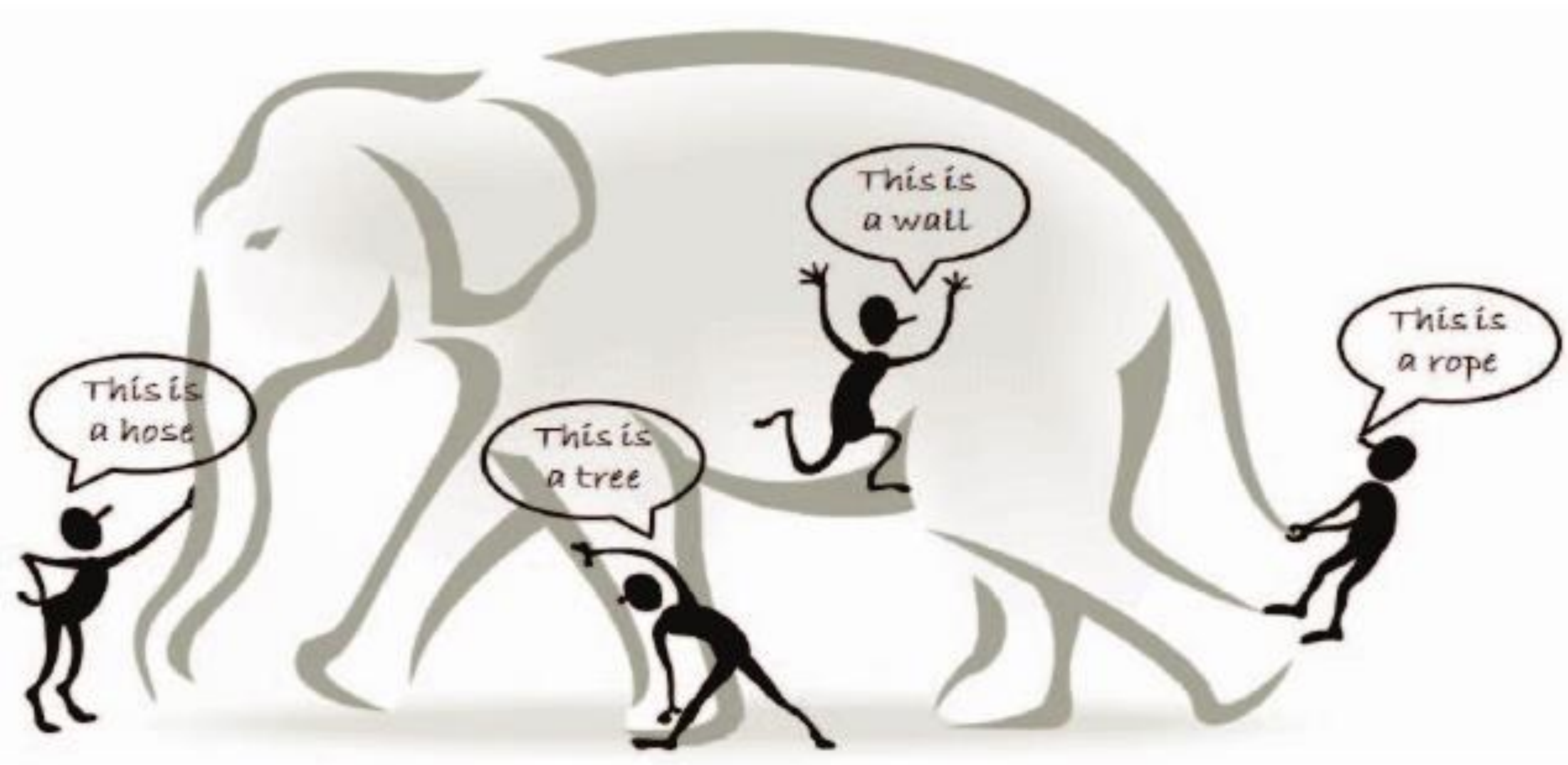
Big data Definition


High-volume
High-velocity
High-variety
Cost-effective, innovative forms of information processing
Enhanced insight & decision making


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

Challenges with Big data


This is a wall
This is a rope
This is a hose
This is a tree

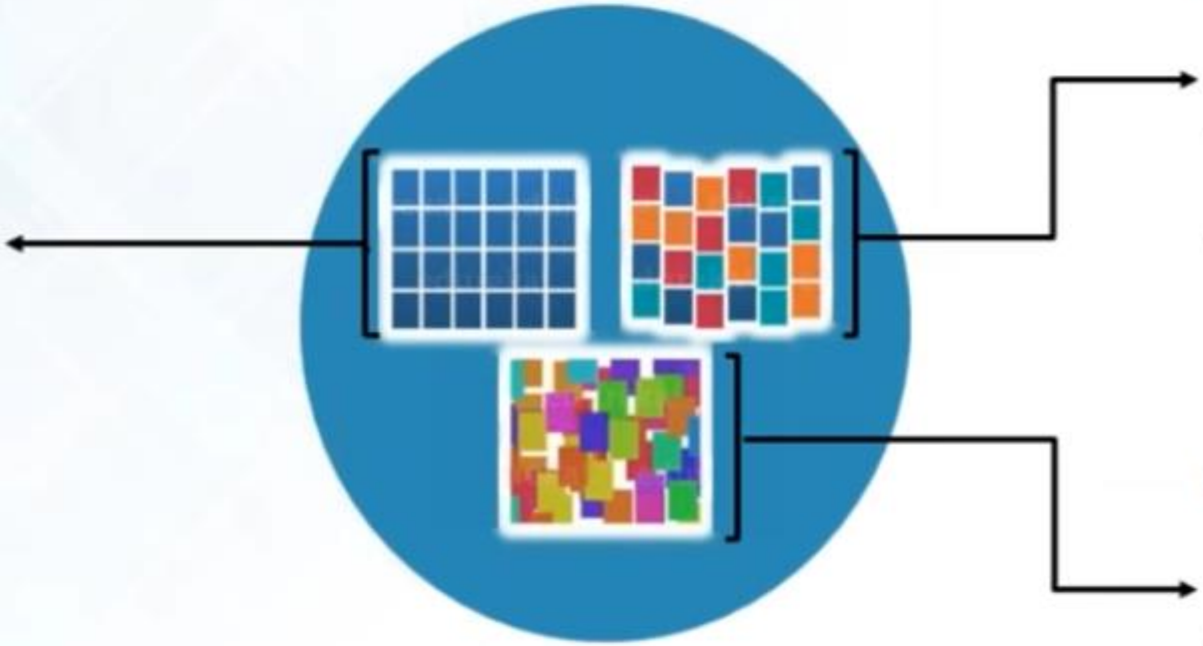
Problem 1: Storing exponentially growing huge datasets

- 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.

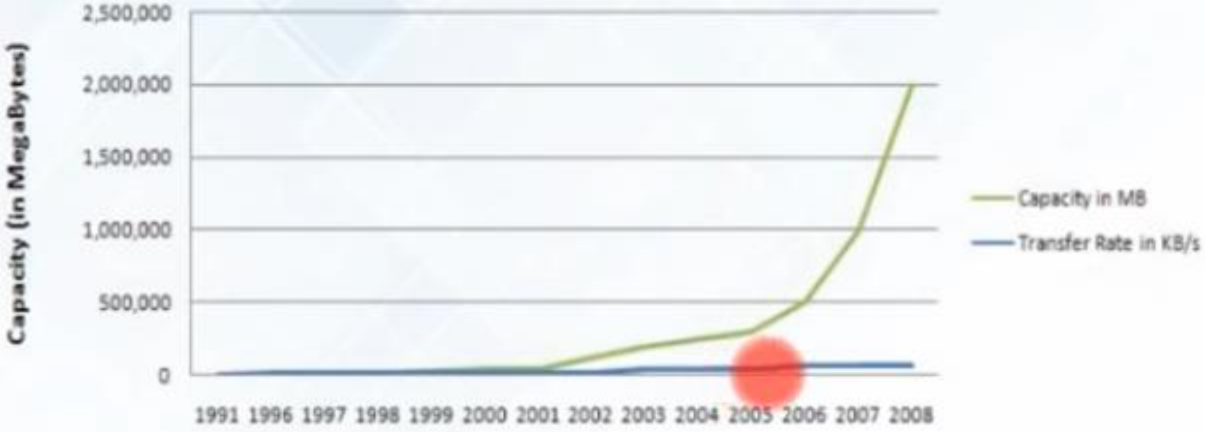
Unstructured

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

Problem 3: Processing data faster

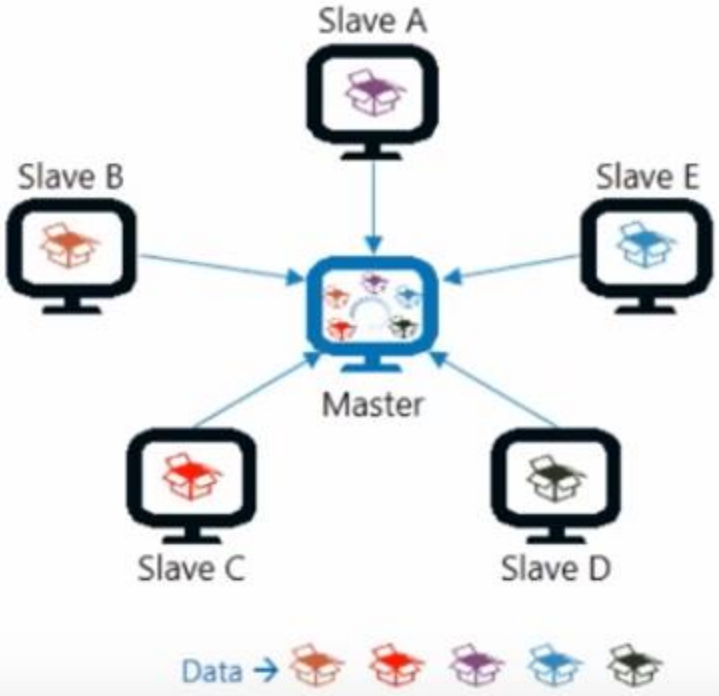
The data is growing at much faster rate than that of disk read/write speed

Relative Improvment
Hard Disk Capacity v.s. Disk Transfer Performance



Source: *Tom's Hardware*

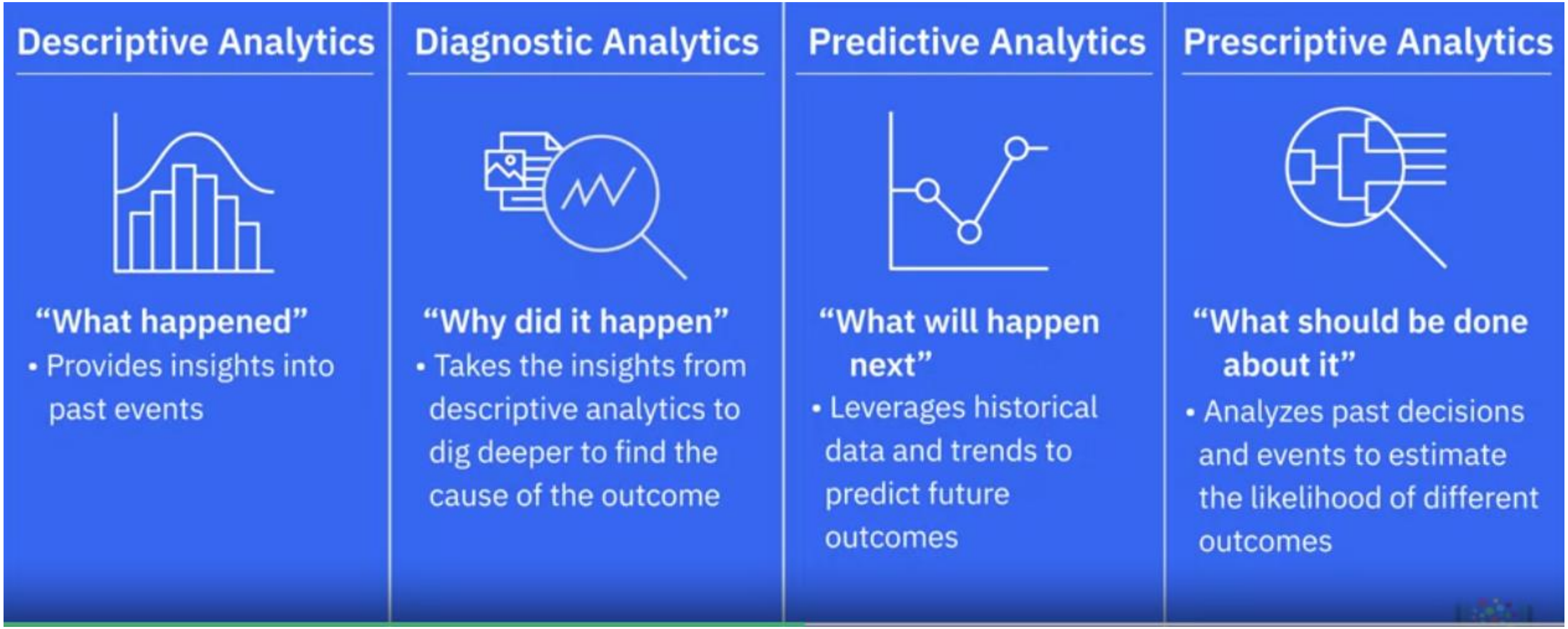
Bringing huge amount of data to computation unit becomes a bottleneck



A big data analytics cycle can be described by the following stage −

1. **Business Problem Definition**
2. **Data Identification**
3. **Data Acquisition & Filtering**
4. **Data Extraction**
5. **Exploratory Data Analysis**
6. **Data Preparation for Modeling and Assessment**
7. **Data Visualization**
8. **Analysis of Results**

Classification of Data Analytics



Descriptive Analytics

“What happened”

- Provides insights into past events

Diagnostic Analytics

“Why did it happen”

- Takes the insights from descriptive analytics to dig deeper to find the cause of the outcome

Predictive Analytics

“What will happen next”

- Leverages historical data and trends to predict future outcomes

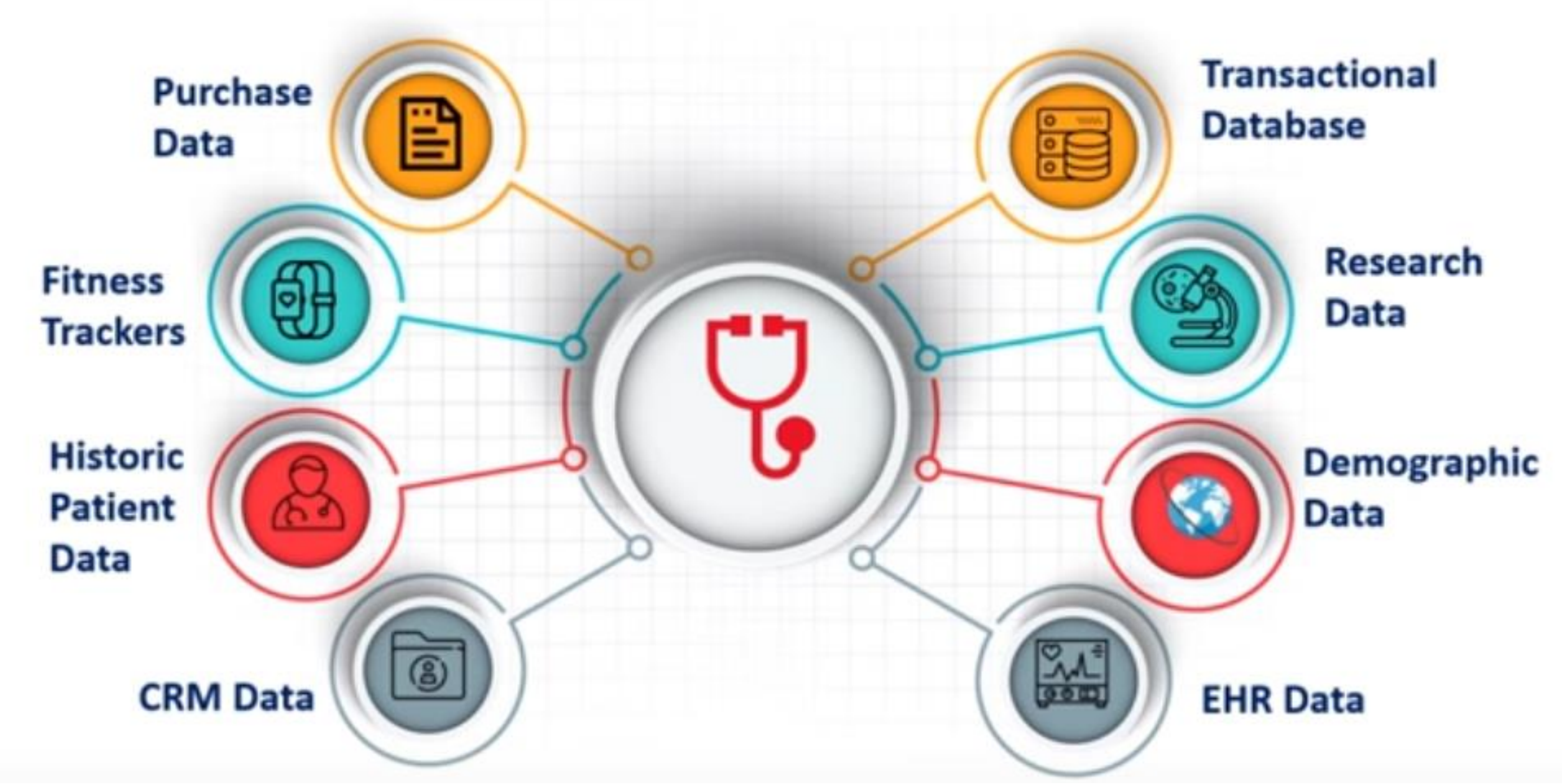
Prescriptive Analytics

“What should be done about it”

- Analyzes past decisions and events to estimate the likelihood of different outcomes

Big data Analytics-Case studies

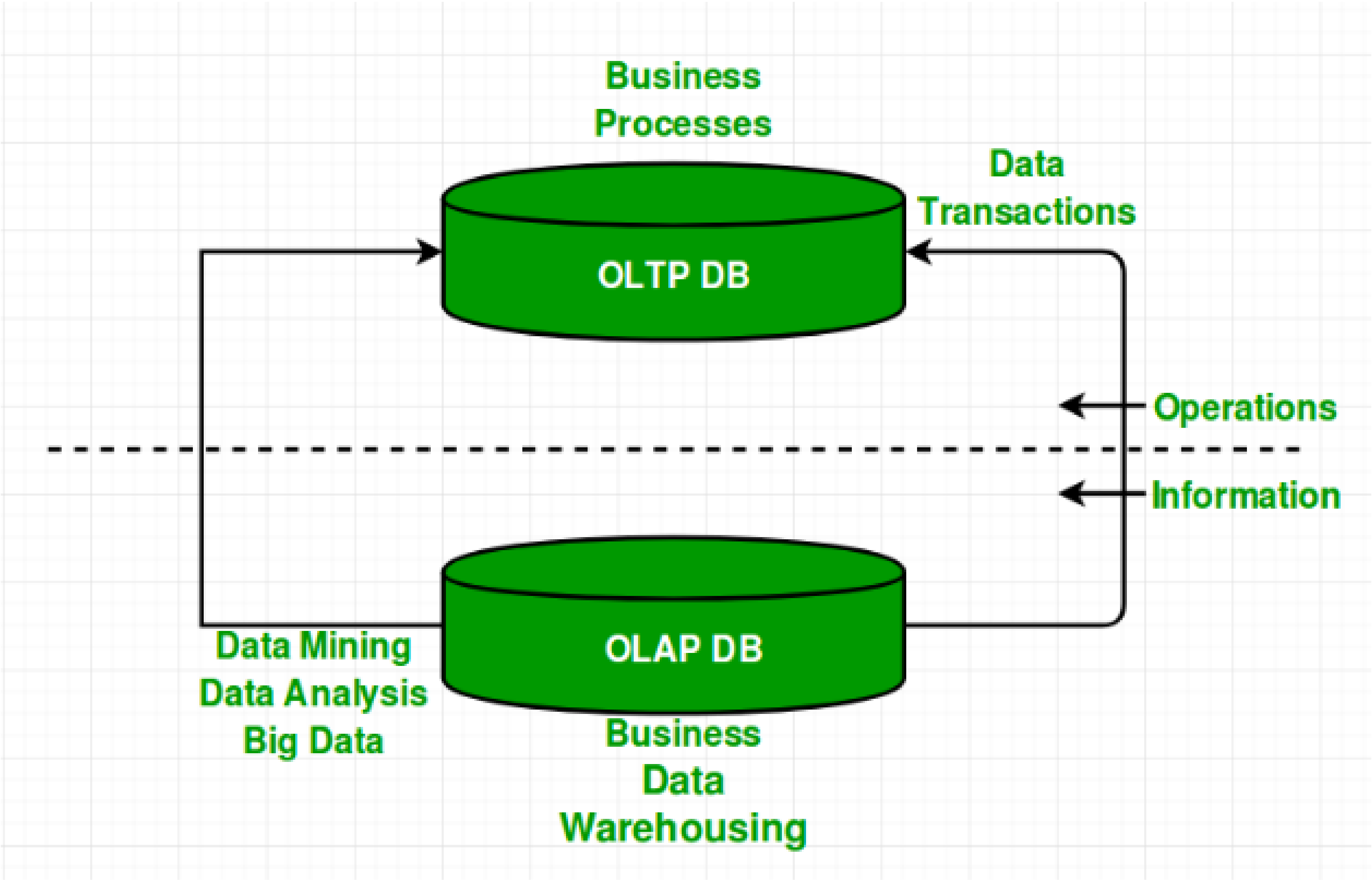
- Healthcare



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

  - https://www.geeksforgeeks.org/difference-between-olap-and-oltp-in-dbms/


Business Processes
Data Transactions
OLTP DB
Operations
Information
OLAP DB
Data Mining
Data Analysis
Big Data
Business Data Warehousing

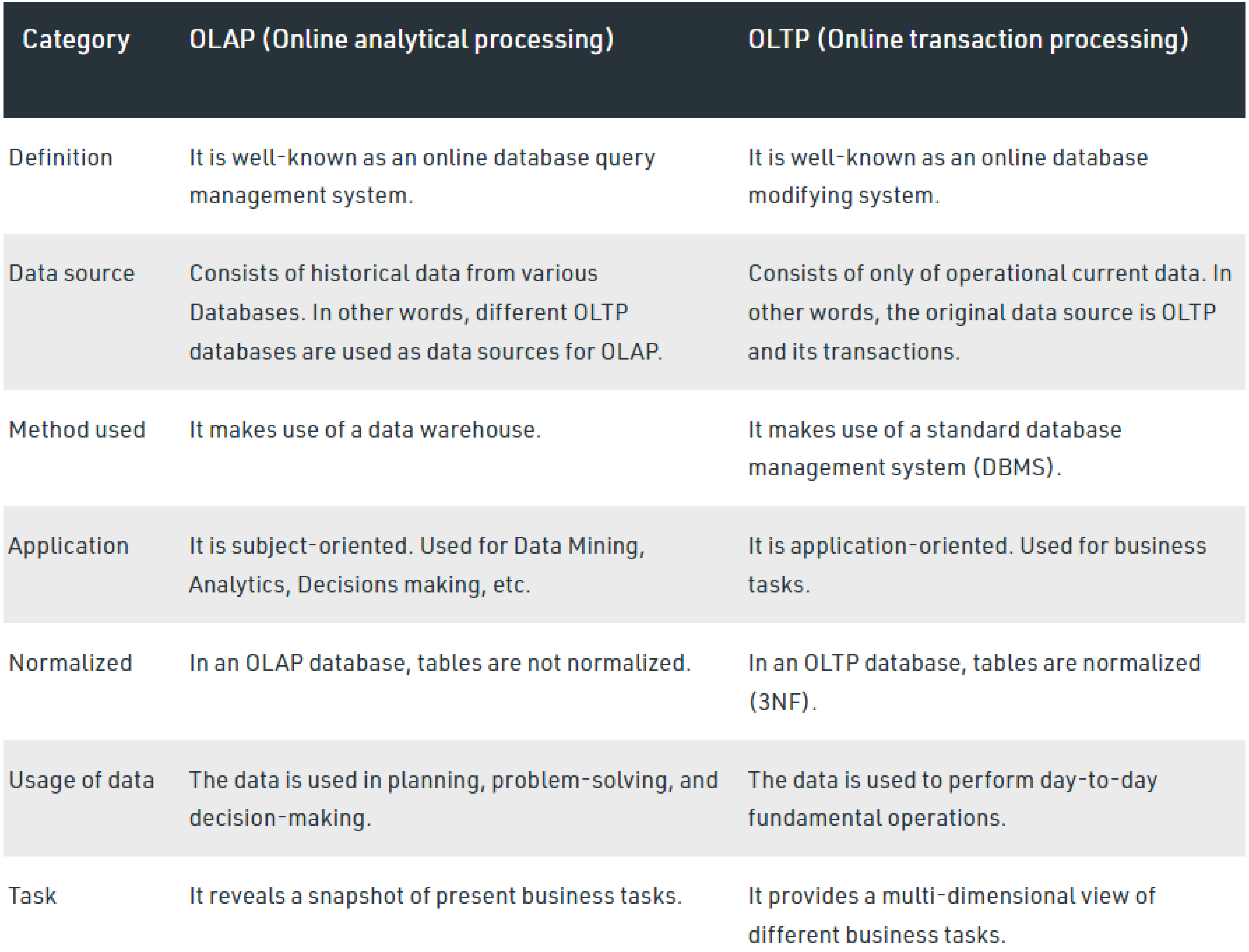
Online Analytical Processing (OLAP): Online Analytical Processing consists of a type of software tools that are used for data analysis for business decisions. OLAP provides an environment to get insights from the database retrieved from multiple database systems at one time. **Examples –** Any type of Data warehouse system is an OLAP system. The uses of OLAP are as follows:

- Spotify analyzed songs by users to come up with a personalized homepage of their songs and playlist.
- Netflix movie recommendation system.

Online transaction processing (OLTP): Online transaction processing provides transaction-oriented applications in a 3-tier architecture. OLTP administers the day-to-day transactions of an organization.

Examples: Uses of OLTP are as follows:

- ATM center is an OLTP application.
- OLTP handles the ACID properties during data transactions via the application.
- It’s also used for Online banking, Online airline ticket booking, sending a text message, add a book to the shopping cart.



Category	OLAP (Online analytical processing)	OLTP (Online transaction processing)
Definition	It is well-known as an online database query management system.	It is well-known as an online database modifying system.
Data source	Consists of historical data from various Databases. In other words, different OLTP databases are used as data sources for OLAP.	Consists of only of operational current data. In other words, the original data source is OLTP and its transactions.
Method used	It makes use of a data warehouse.	It makes use of a standard database management system (DBMS).
Application	It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc.	It is application-oriented. Used for business tasks.
Normalized	In an OLAP database, tables are not normalized.	In an OLTP database, tables are normalized (3NF).
Usage of data	The data is used in planning, problem-solving, and decision-making.	The data is used to perform day-to-day fundamental operations.
Task	It reveals a snapshot of present business tasks.	It provides a multi-dimensional view of different business tasks.