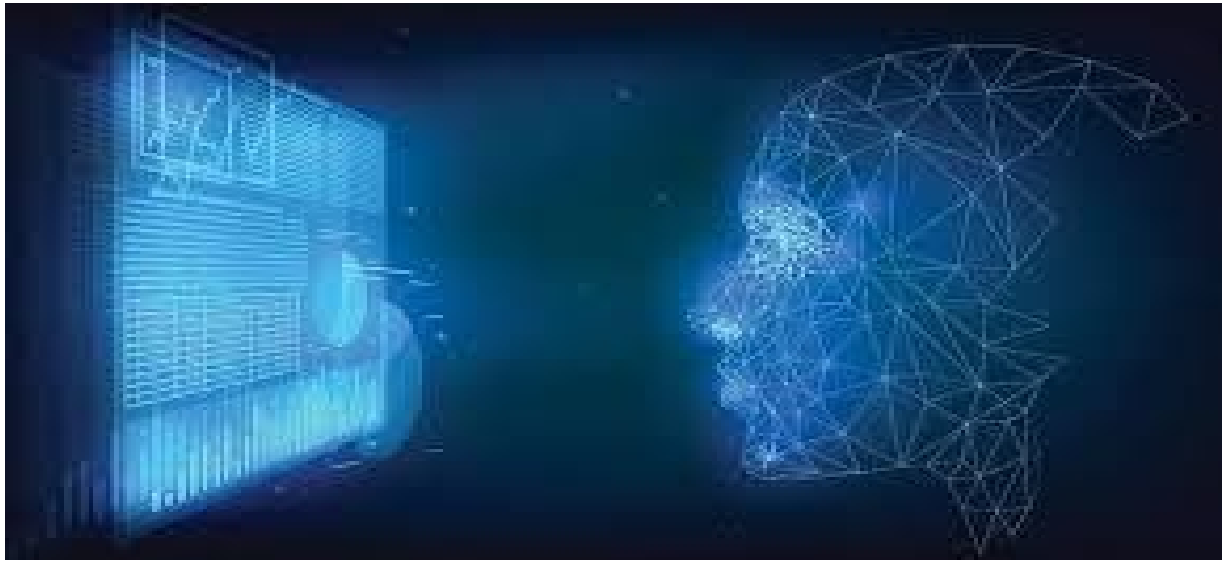
Introduction to Big Data

Welcome to the World of Big Data

Data- What Makes it Big??? How big is big?



No Single Definition

The term ‘big data’ is self-explanatory − a collection of huge data sets that normal computing techniques cannot process.

“***Big Data***” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

What is Big Data?

BIG DATA : is the term for a collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.

Real world examples of Big Data

- Facebook : has 40 PB of data and captures 100 TB / day
- Yahoo : 60 PB of data
- Twitter : 8 TB / day
- EBay : 40 PB of data, captures 50TB/ day

Three attributes stand out as defining Big Data characteristics

Huge volume of data: *Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.*

Complexity of data types and structures: *Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.*

Speed of new data creation and growth: *Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.*

Attributes Defining Big Data Characteristics

Volume

Big Data observes and tracks what happens from various sources which include business transactions, social media and information from machine-to-machine or sensor data. This creates large volumes of data.

Variety

Data comes in all formats that may be structured, numeric in the traditional database or the unstructured text documents, video, audio, email, stock ticker data.

Velocity

The data streams in high speed and must be dealt with timely.

The processing of data that is, analysis of streamed data to produce near

Look to the left, and you see Times Square at dusk. Look to the right, and you see the same location at midmorning. Internationally acclaimed photographer Stephen Wilkes's time-altering image of New York's Times Square is part of his body of work titled *Day to Night*. The image was created by blending more than 1,400 separate photos taken over the course of 15 hours—a meticulous process that took him nearly three months. PHOTO: STEPHEN WILKES

Big Facts About Big Data

As of 2013, experts believed that 90% of the world's data was generated from 2011 to 2012.

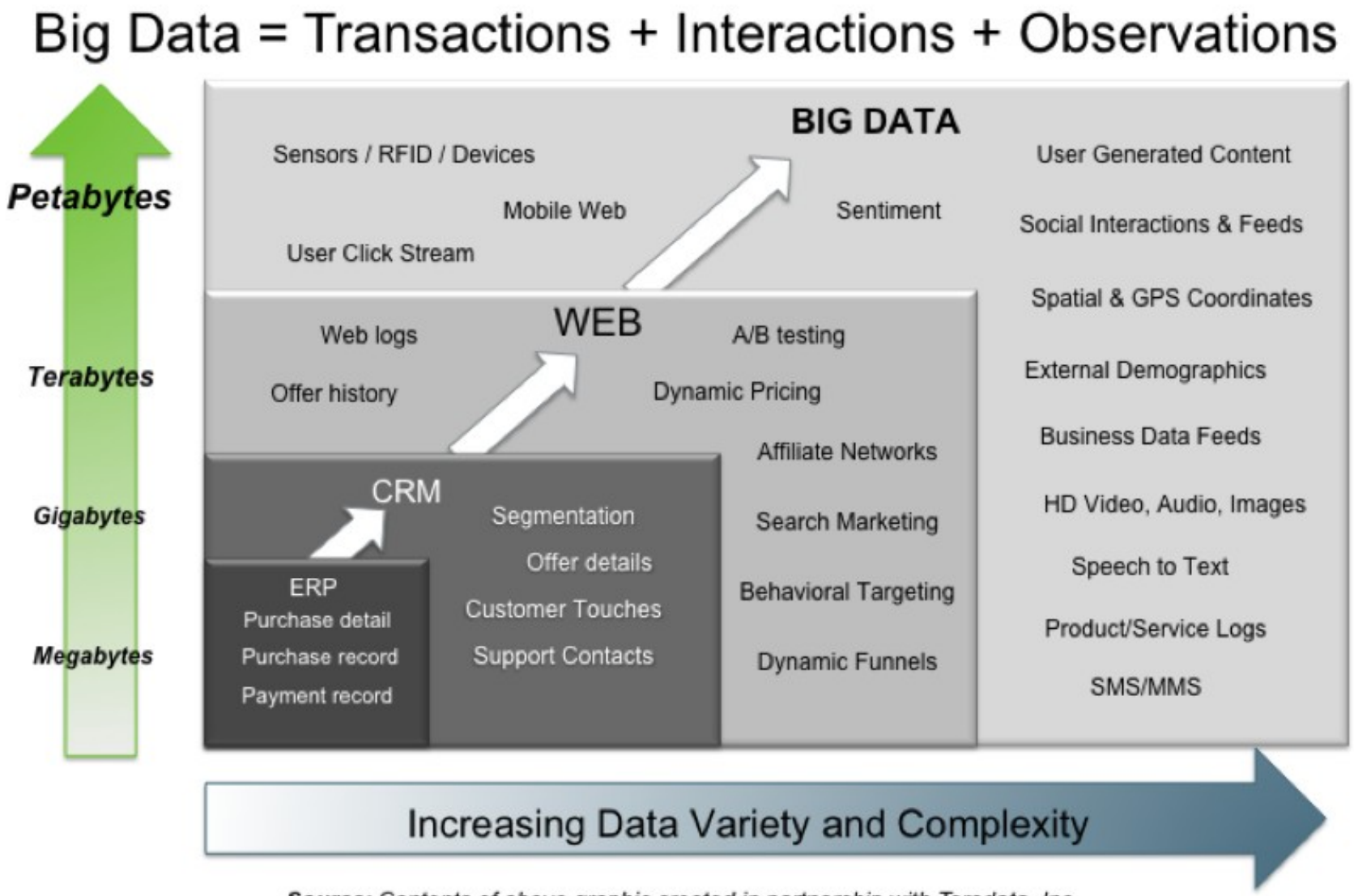
In 2018, more than 2.5 quintillion bytes of data were created every day.

The amount of data in the world was estimated to be 44 zettabytes at the dawn of 2020.

Google handles a staggering 1.2 trillion searches every year.

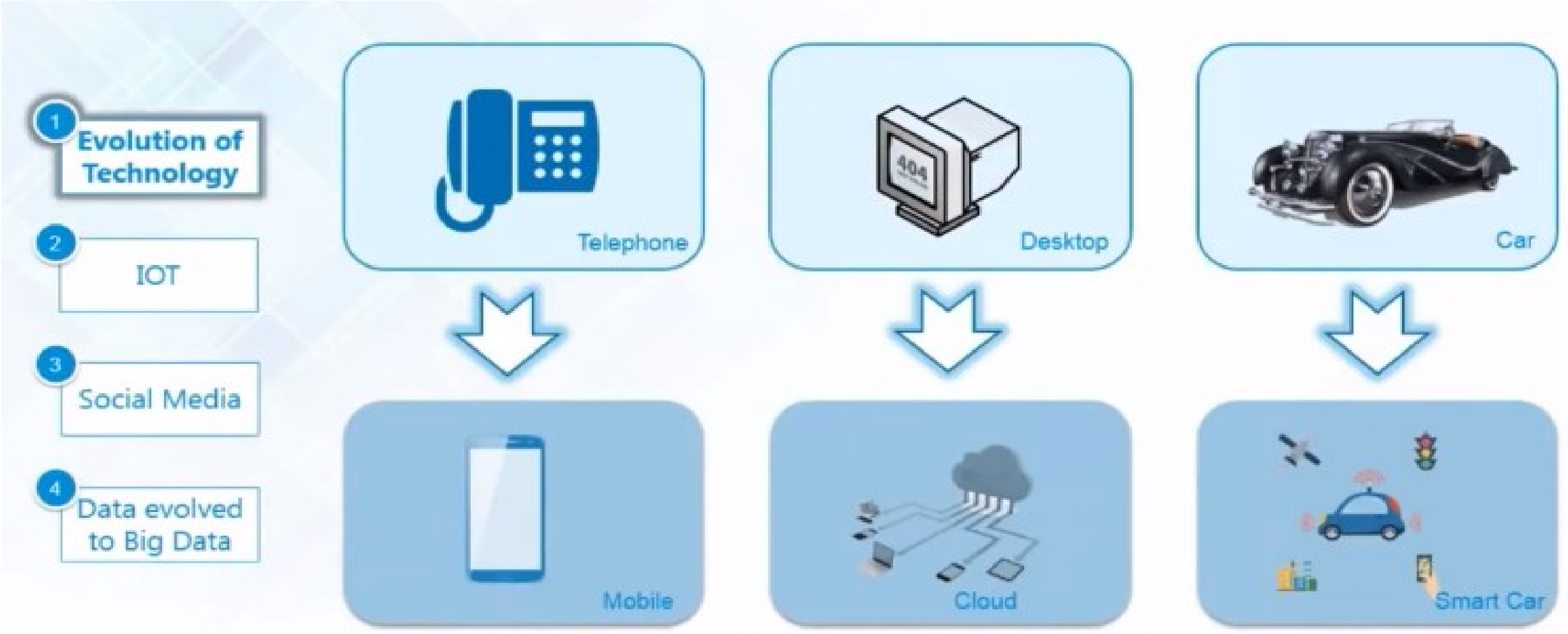
Unit of Data size	Exact size	Approximate Size	Examples
KB (kilobyte)	2^{10} or 1024 bytes	(10^3 or one thousand) bytes	A typical joke =1KB
MB(megabyte)	2^{20} bytes	(10^6 or one million) bytes	Complete work of Shakespeare =5MB
GB (gigabyte)	2^{30} bytes	(10^9 or one billion) bytes	Ten yards of books on a shelf = 1GB
TB (terabyte)	2^{40} bytes	(10^{12} or one trillion) bytes	All the X-rays for a large hospital =1TB Tweets; created daily =121TB;
PB (peta byte)	2^{50} bytes	(10^{15} or one quadrillion) bytes	All U.S. academic research libraries = 2PB Data processed in a day by Google =24PB
EB (exa byte)	2^{60} bytes	(10^{18} or one Quintillion) bytes	Total global data created in 2006 = 161EB
ZB (zetta byte)	2^{70} bytes	(10^{21} or one Sextillion) bytes	Total amount of global data created in 2012 = 2.7 ZB and expected 44 ZB by 2020
YB (yotta byte)	2^{80} bytes	(10^{24} or one Septillion) bytes	

The Evolution Big Data

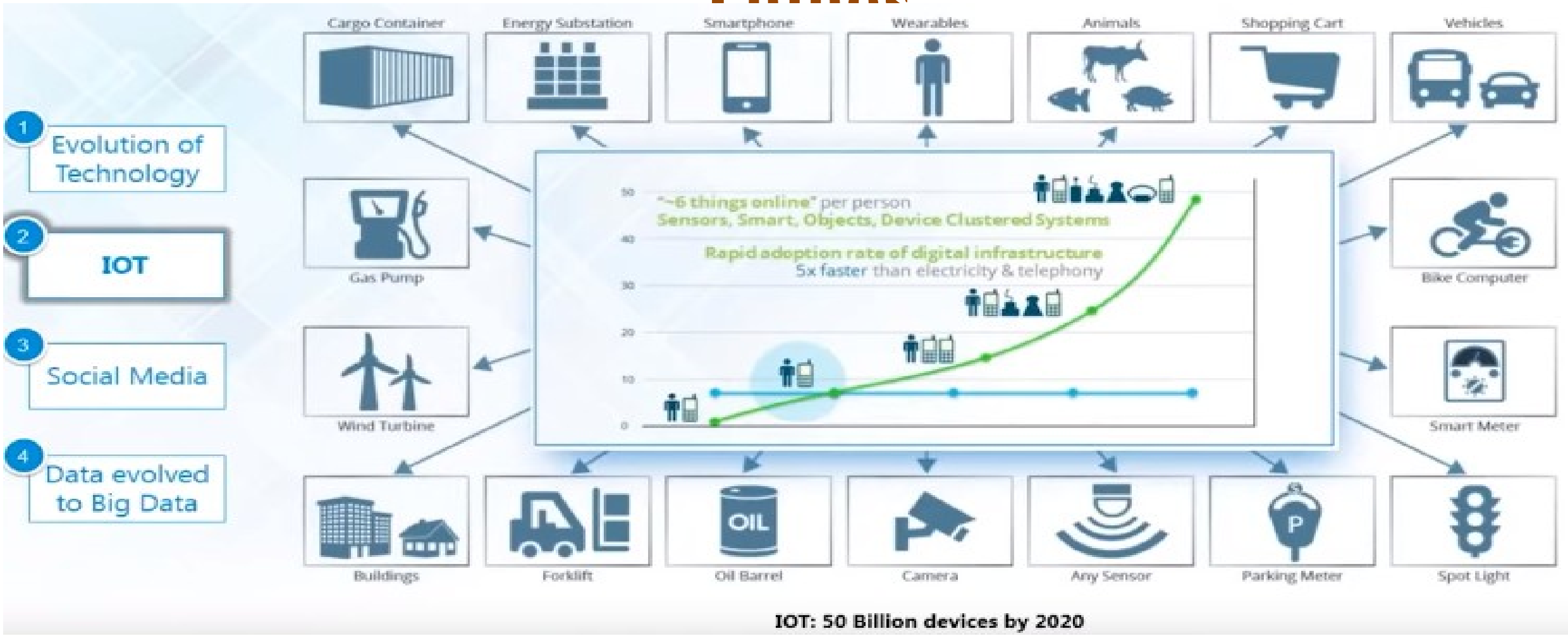


***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Evolution of Big Data by technology

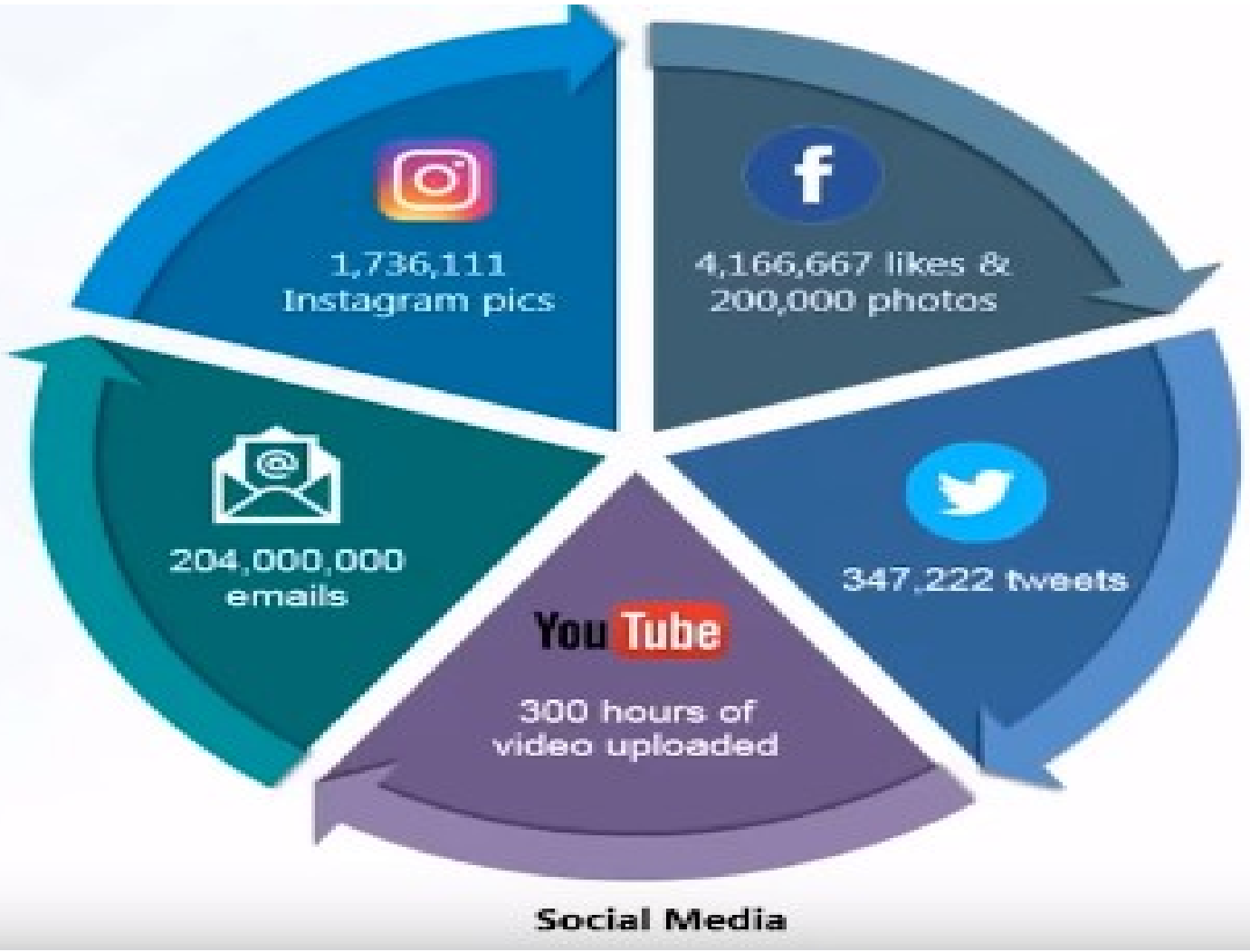


Evolution of Big Data by Internet Of Things



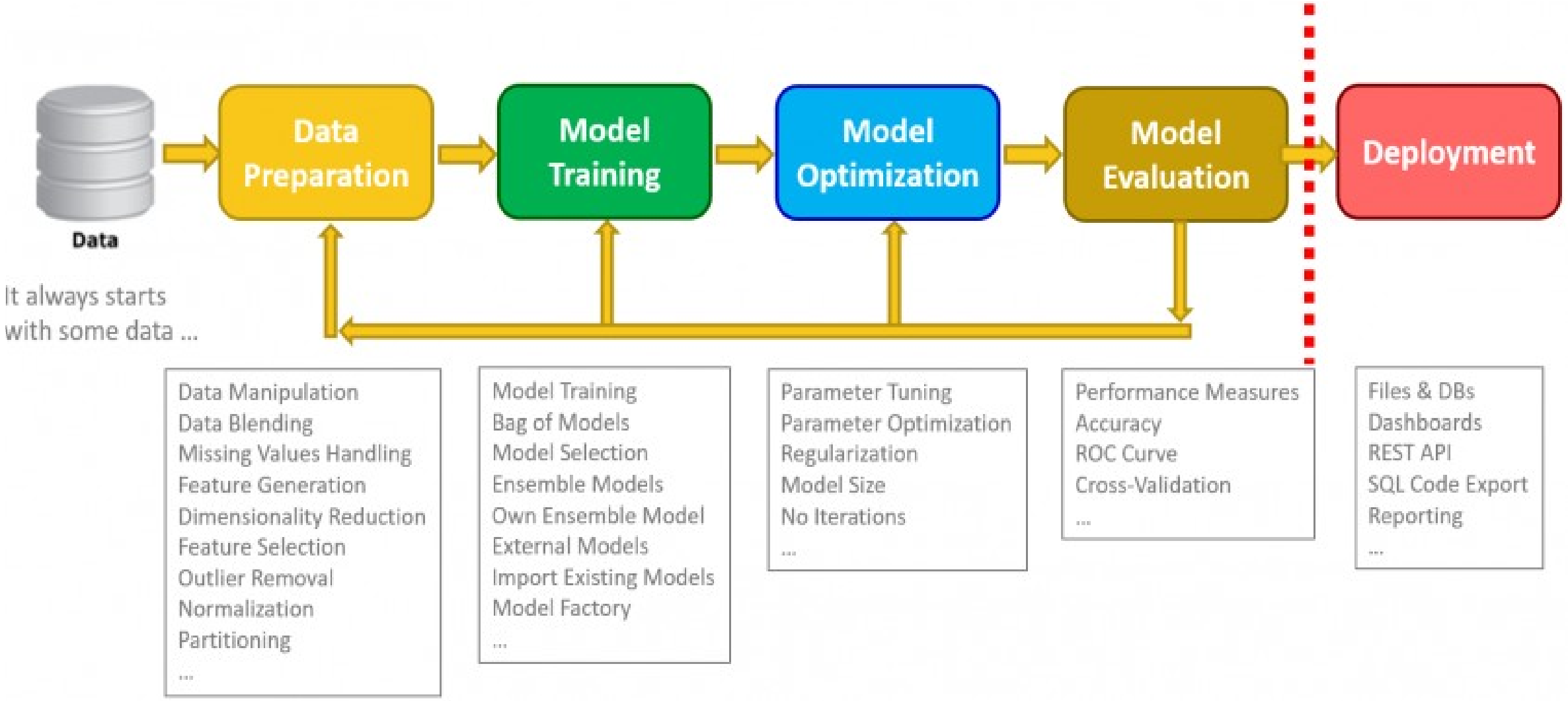
Evolution of Big Data by Social Media

1. Evolution of Technology
2. IOT
3. **Social Media**
4. Data evolved to Big Data



Evolution of Big Data by other factors

Best practices for Big Data Analytics



Three attributes stand out as defining Big Data characteristics

Huge volume of data: Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.

Complexity of data types and structures: Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.

Speed of new data creation and growth: Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

Attributes Defining Big Data Characteristics

Volume

Big Data observes and tracks what happens from various sources which include business transactions, social media and information from machine-to-machine or sensor data. This creates large volumes of data.

Variety

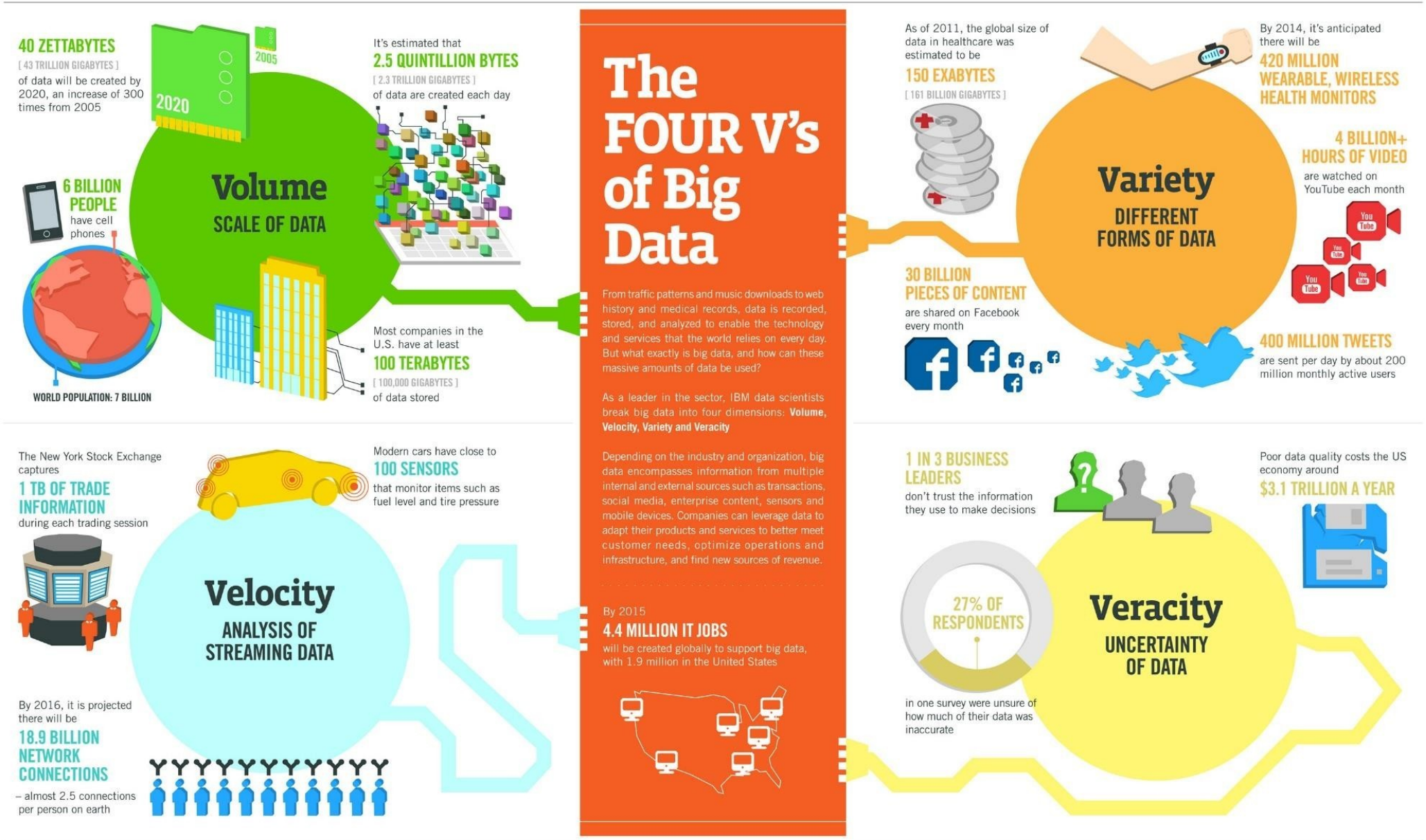
Data comes in all formats that may be structured, numeric in the traditional database or the unstructured text documents, video, audio, email, stock ticker data.

Velocity

The data streams in high speed and must be dealt with timely.

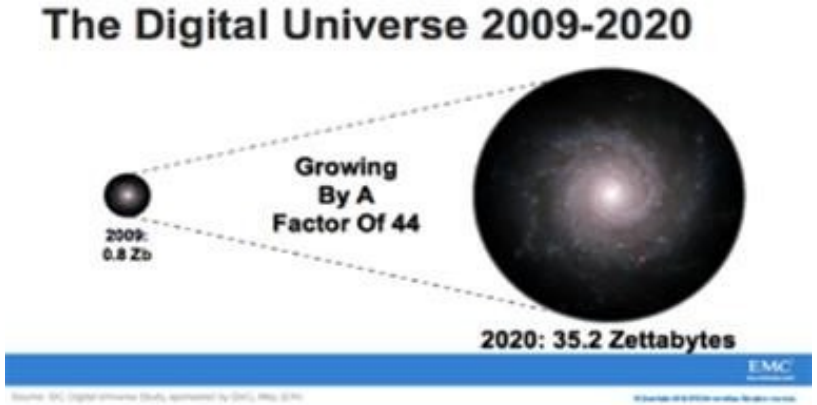
The processing of data that is, analysis of streamed data to produce near

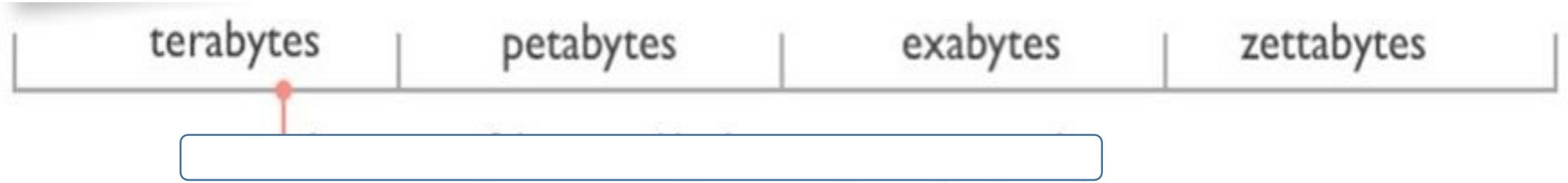
Big Data- the


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Characteristics of Big Data: 1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
- Data volume is increasing exponentially





Characteristics of Big Data:2-Complexity (Variety)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data

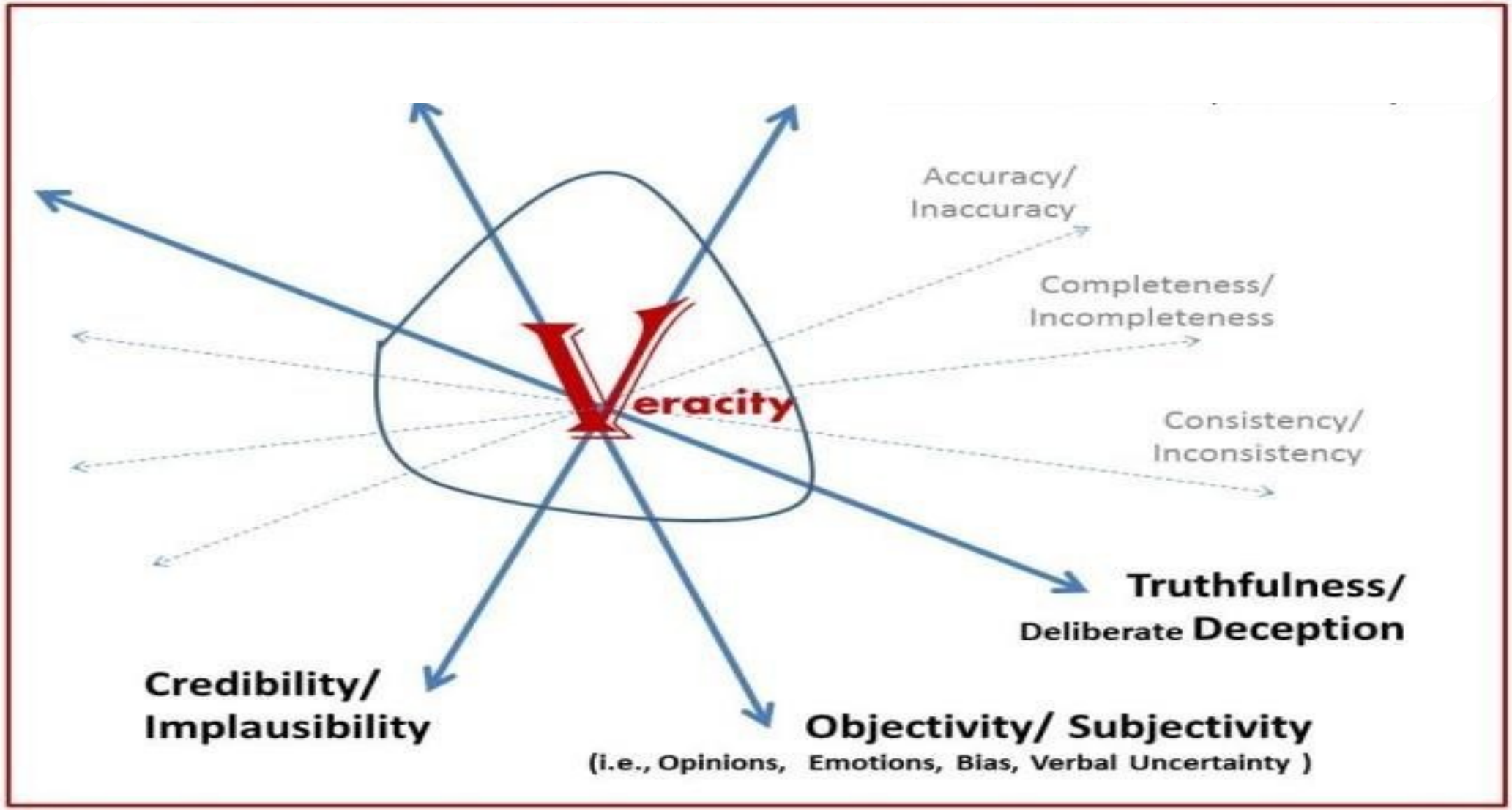
To extract knowledge➔ all these types of data need to linked together

Characteristics of Big Data: 3-Speed (Velocity)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

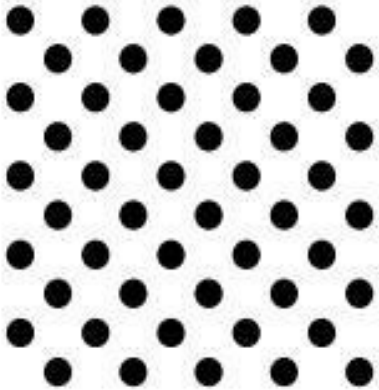
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction.

Characteristics of Big Data: 4-Accuracy/ Trustworthiness



The 4Vs in a Nutshell

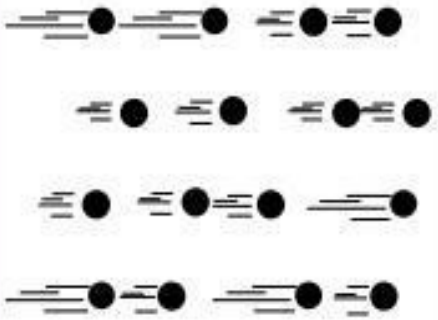
Volume



Data at Rest

Terabytes to exabytes of existing data to process

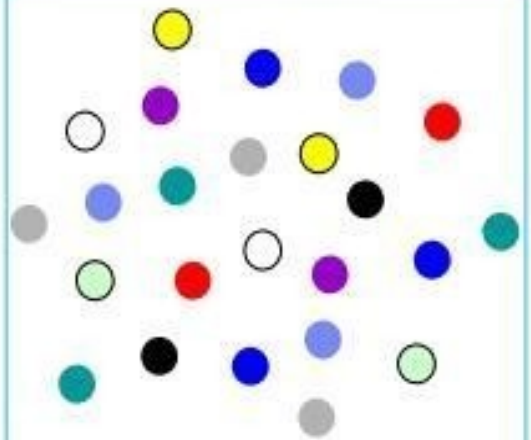
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

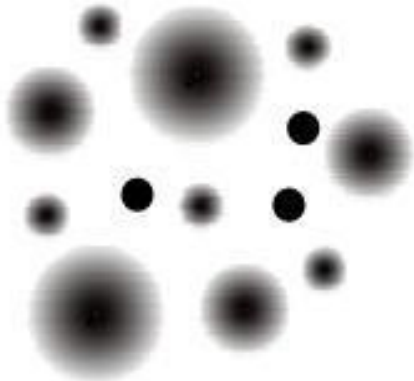
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*



Data in Doubt

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

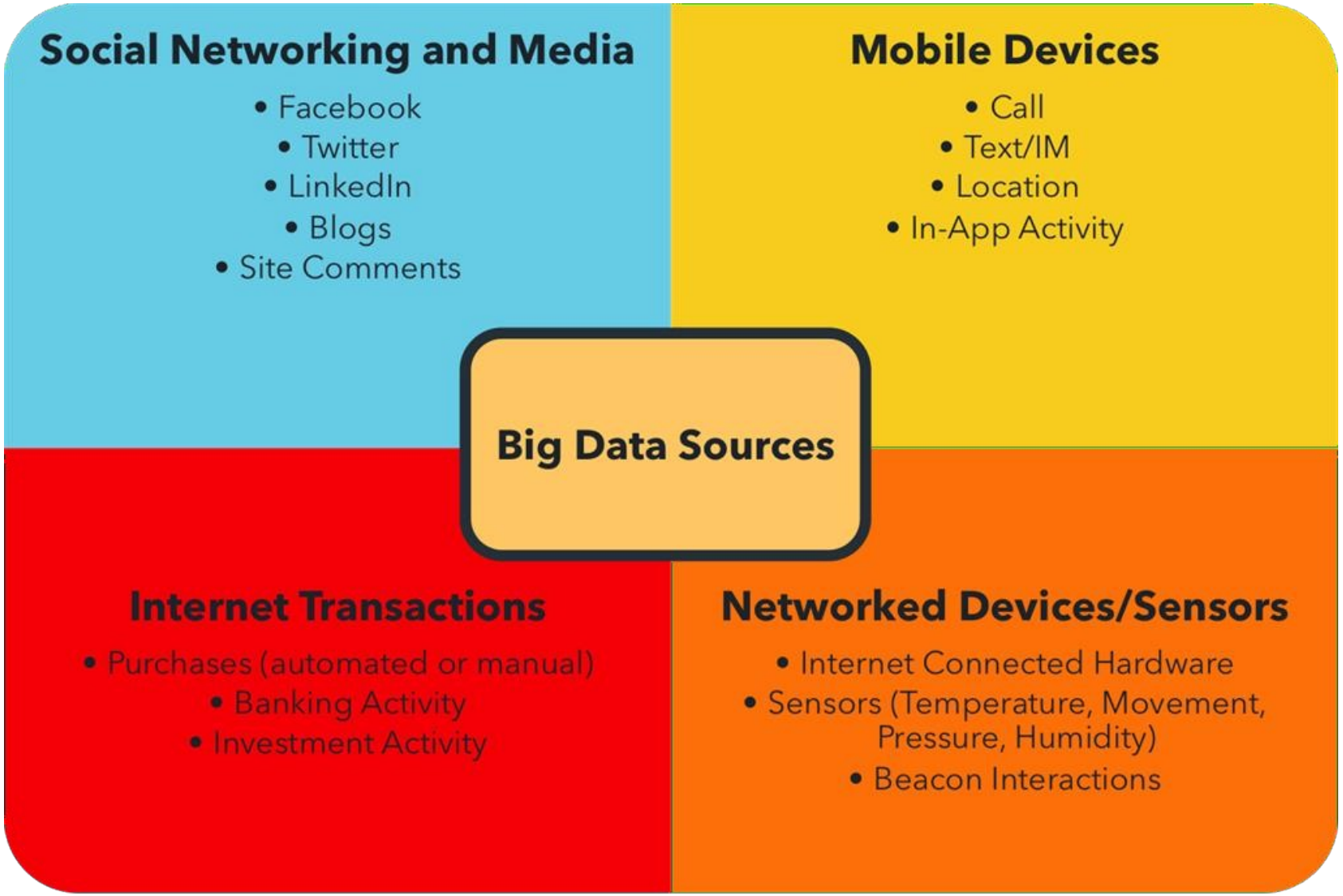
Why Big Data Analytics

- **Cost Savings :** help in identifying more efficient ways of doing business.
- **Time Reductions :**helps businesses analyzing data immediately and make quick decisions based on the learnings.
- **New Product Development :** By knowing the trends of customer needs and satisfaction through analytics you can create products according to the wants of customers.
- **Understand the market conditions :** By analyzing big data you can get a better understanding of current market conditions.
- **Control online reputation:** Big data tools can do

Sources of Big Data Deluge

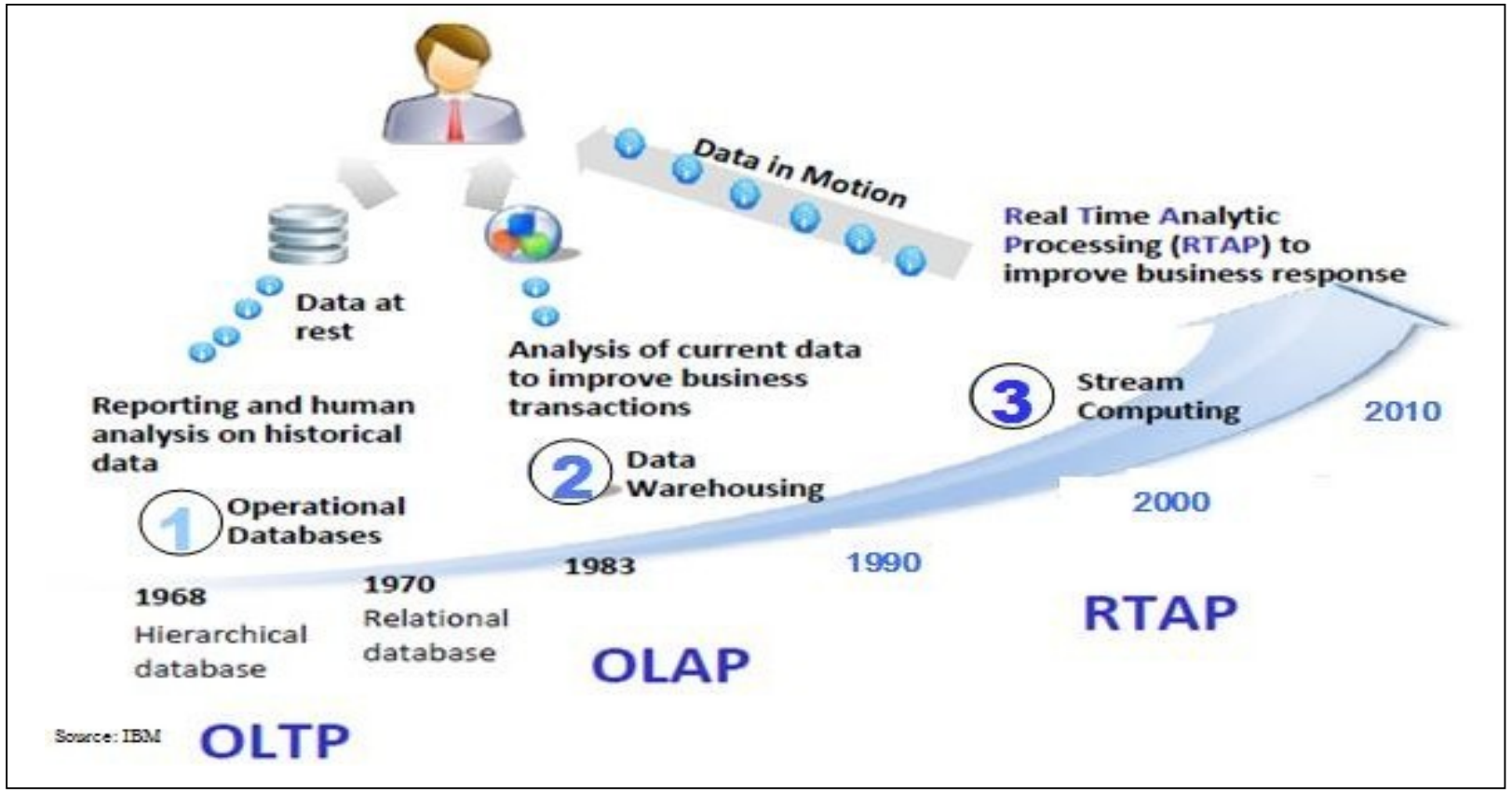
- Mobile sensors – GPS, accelerometer, etc.
- Social media – 700 Facebook updates/sec in2012
- Video surveillance – street cameras, stores, etc.
- Video rendering – processing video for display
- Smart grids – gather and act on information
- Geophysical exploration – oil, gas, etc.
- Medical imaging – reveals internal body structures
- Gene sequencing – more prevalent, less expensive, healthcare would like to predict personal illnesses

Sources of Big Data Deluge



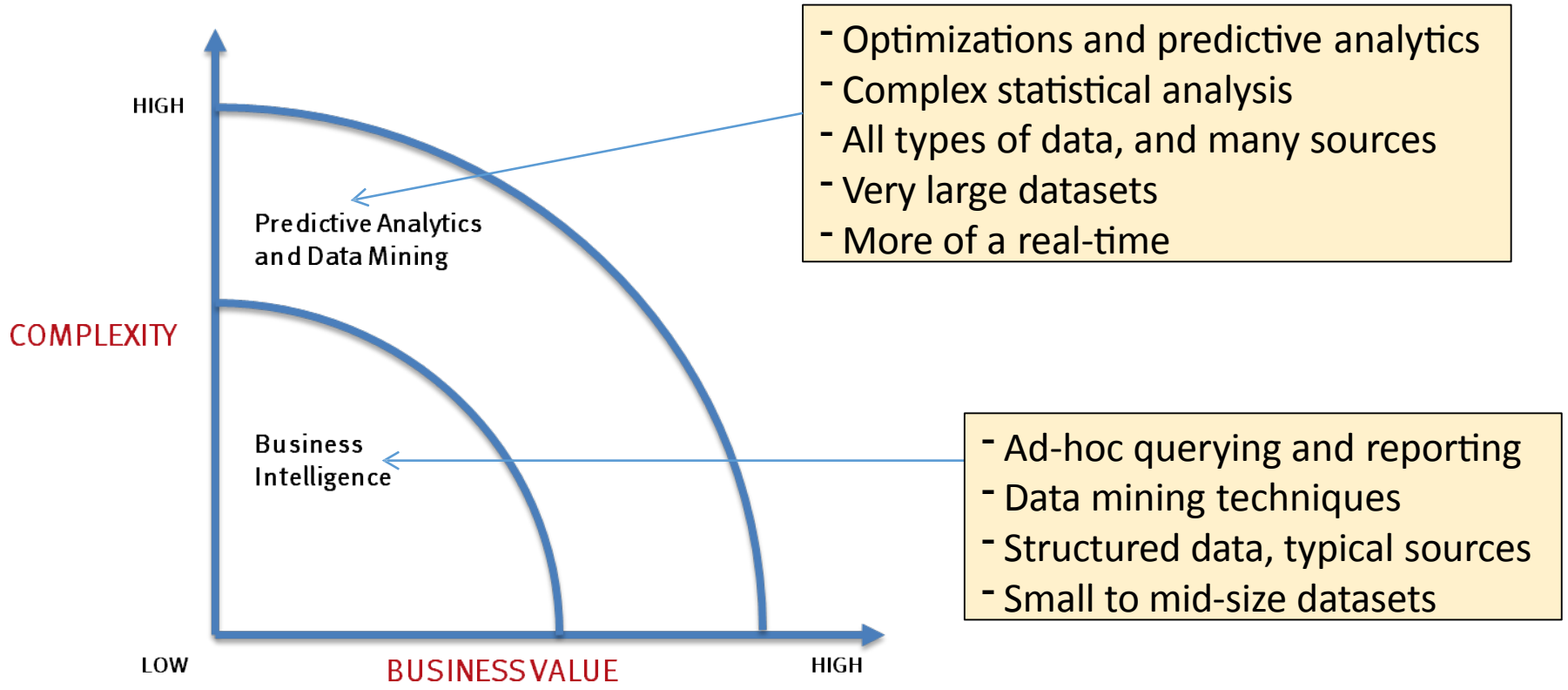
Sources of Big Data Deluge

Harnessing Big Data



- OLTP: Online Transaction Processing (DBMS)
- OLAP: Online Analytical Processing (Data Warehousing)
- RTAP: Real-Time Analytics Processing (Big Data Architecture &Technology)

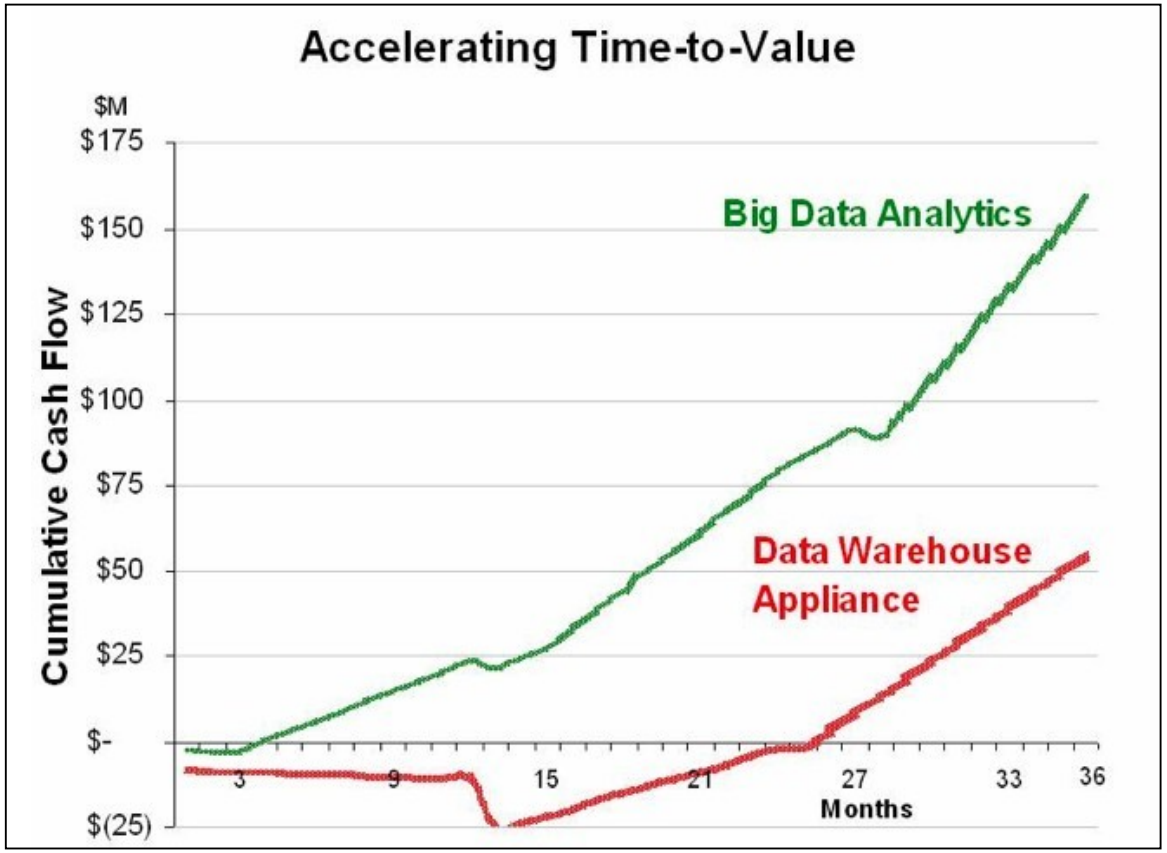
What's driving Big Data



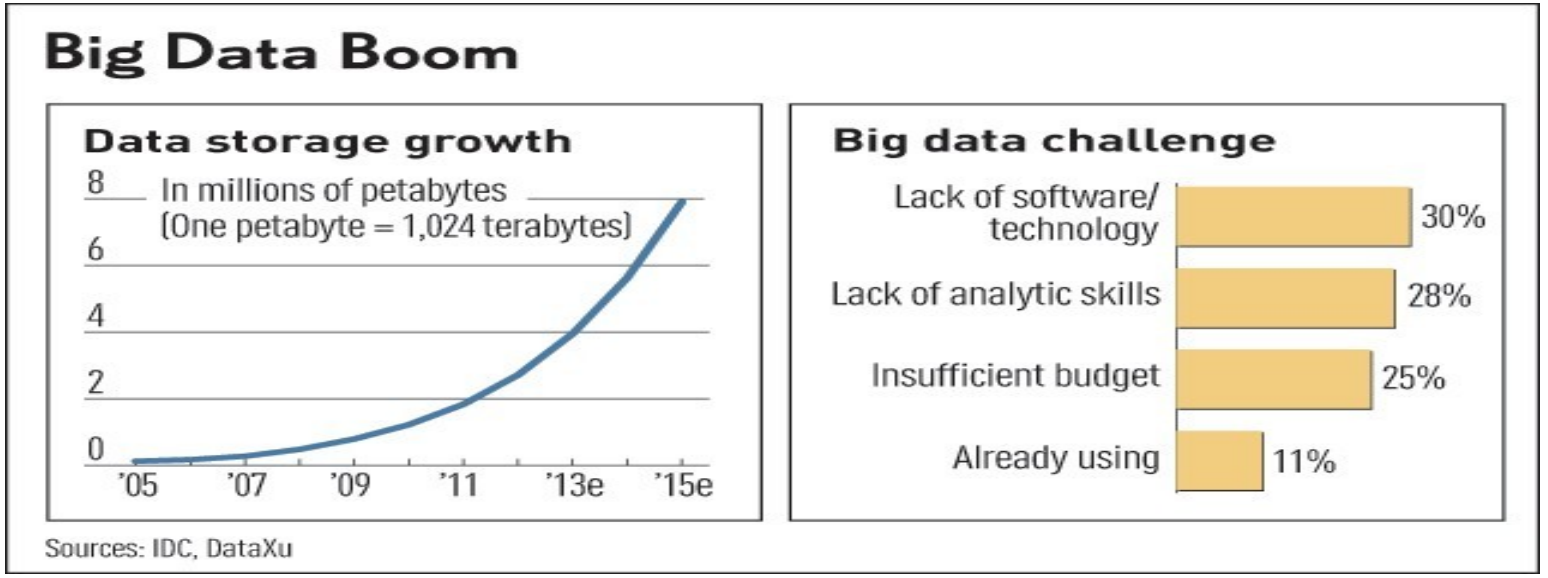
Value of Big Data Analytics

- Big data is more real-time in nature than traditional DW applications.

- Traditional DW architectures are not well-suited for big data apps



Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

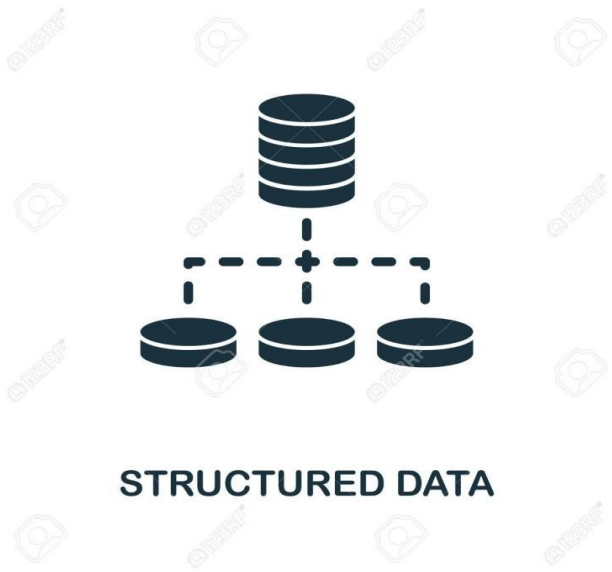
Types of big data

Big Data' could be found in three forms:

- **Structured**
- **Unstructured**
- **Semi-structured**

Structured

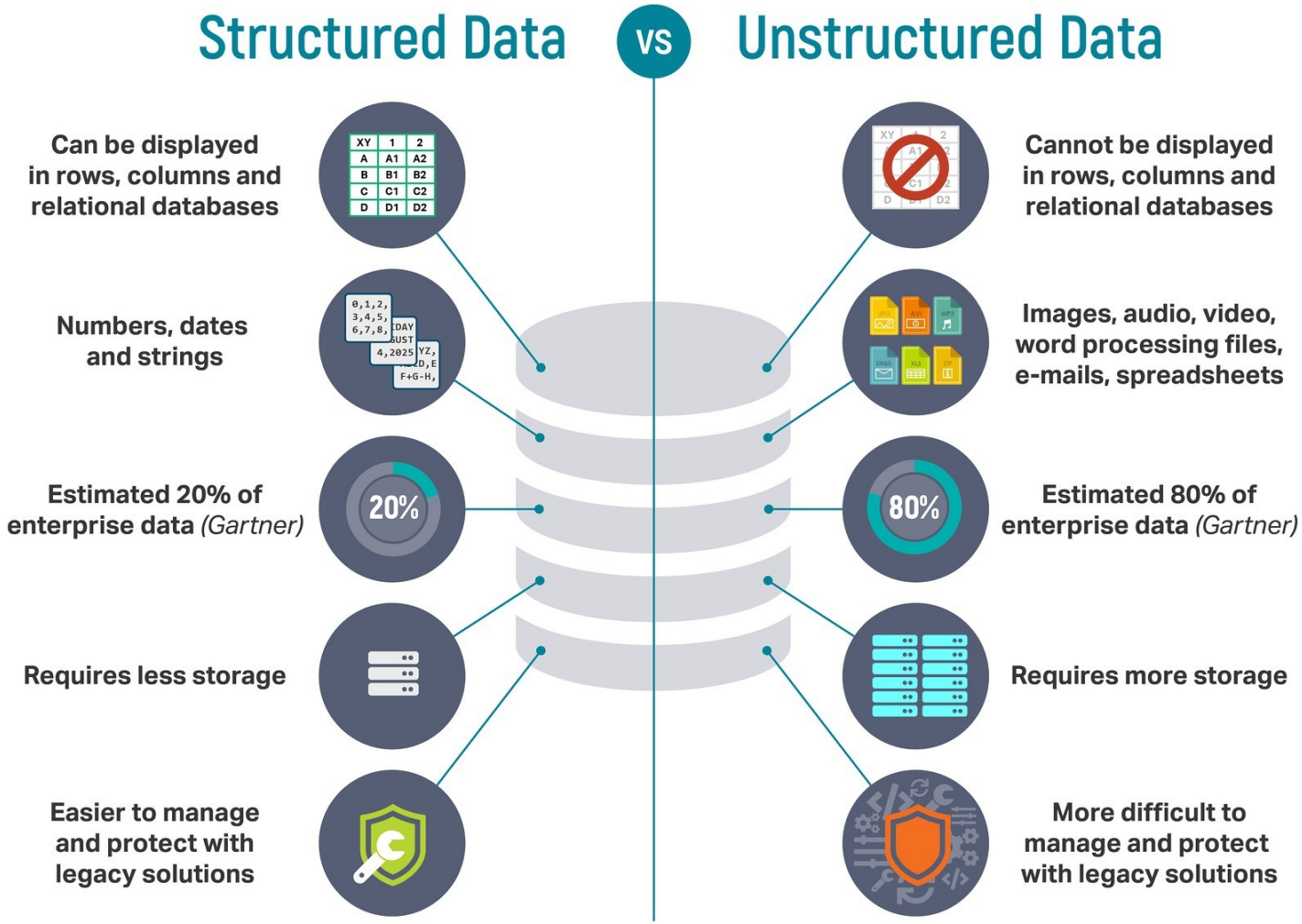
- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
- The format is well known in advance. Can derive value out of it.
- Currently typical sizes are being in the rage of multiple zettabytes.



Structured Data

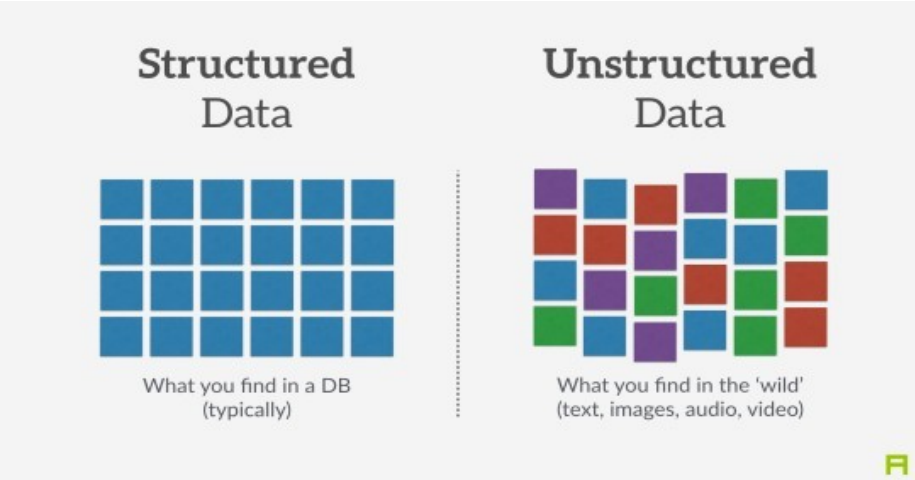
Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured Data



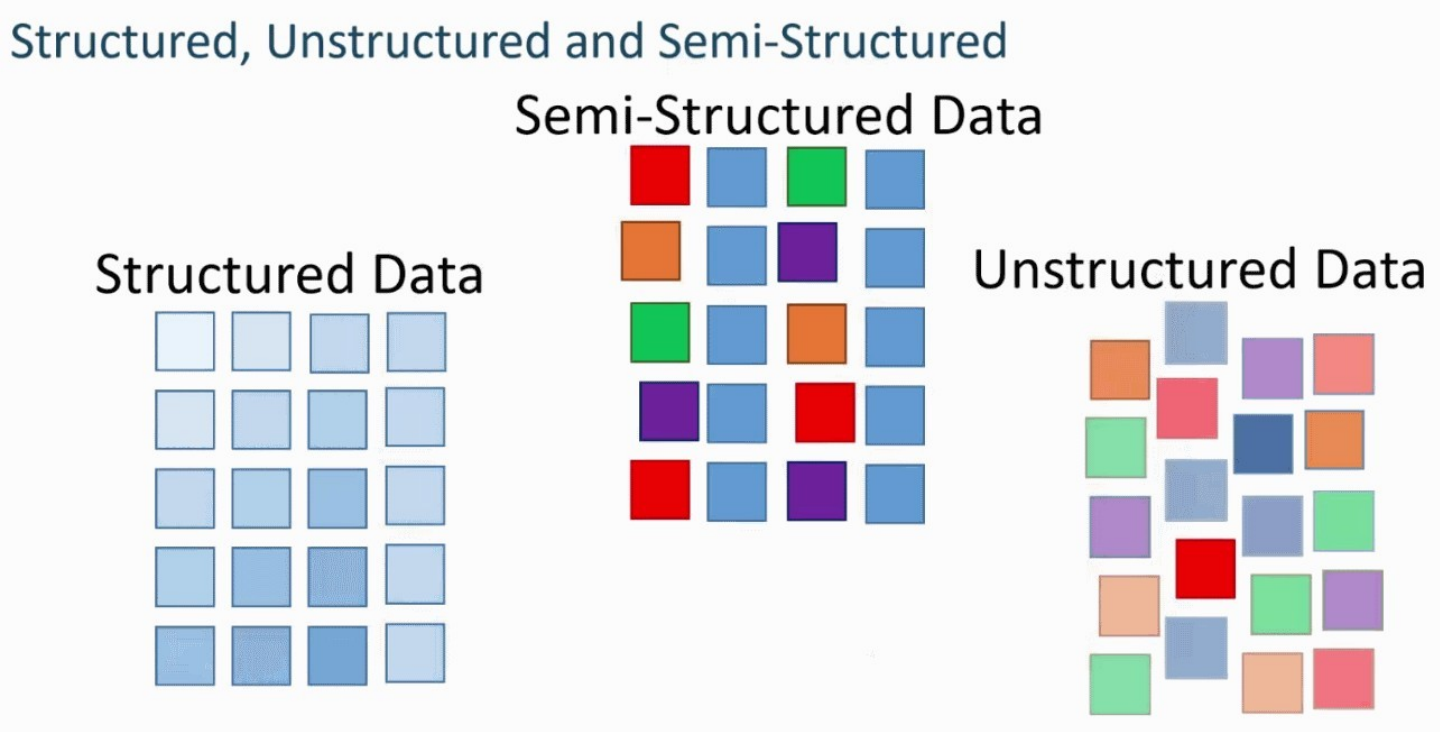
Unstructured Data

- Data with unknown form or the structure.
- Apart from being huge it poses multiple challenges in terms of its processing for deriving value out of it.
- A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.
- Raw form or unstructured format created complexity.



Semi-structured

- Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined like a table definition in relational DBMS.
- Example of semi-structured data is a data represented in an XML file.



Semi Structured

<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>

<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>

<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>

<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>

<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

A Contrast of the Three Types

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.


Banking And Securities
Media and Entertainment
Insurance
Big data Applications
HealthCare
Transportation
Energy and Utilities
Education
Manufacturing