Data Modeling and Visualization (417522)

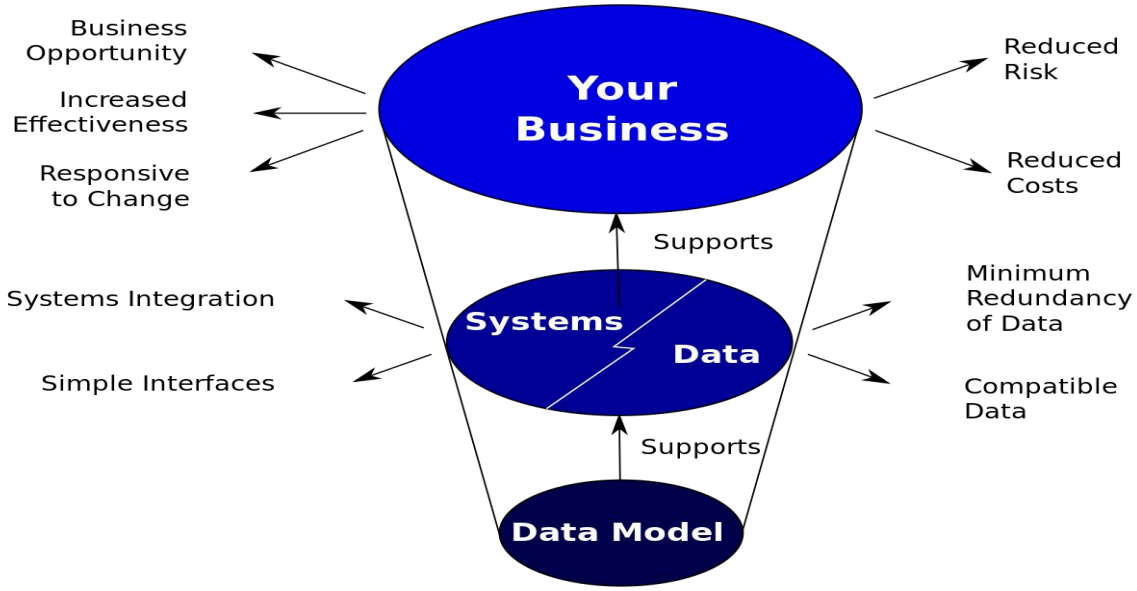
Introduction Data Modeling

Course: BE (AIDS)

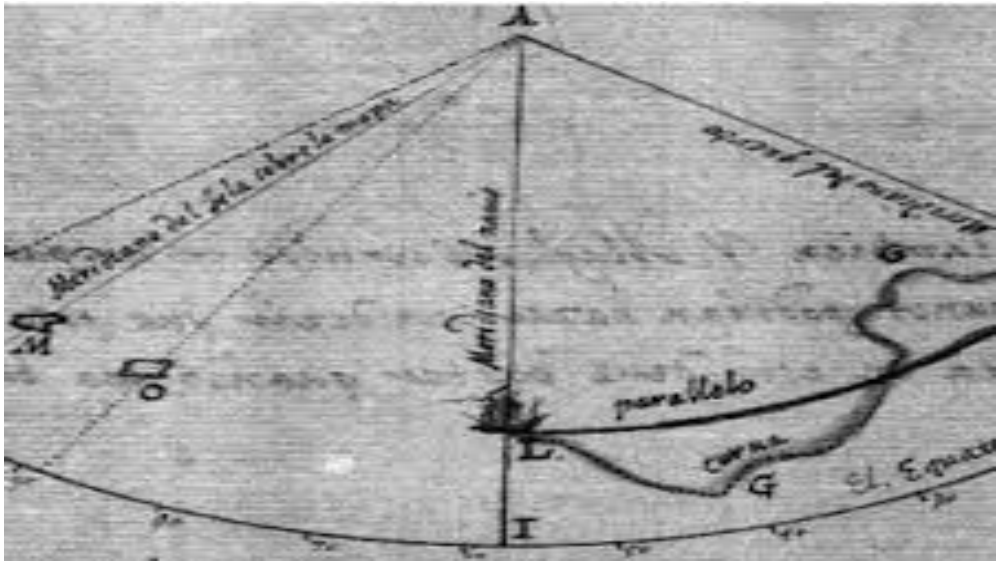
Subject Teacher
Mrs.R.L.Ghule

Data model concept :- Invention

- A **data model** is an abstractmodel that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities. For instance, a data model may specify that the data element representing a car be composed of a number of other elements which, in turn, represent the color and size of the car and define its owner.
- A data model explicitly determines the structure of data. Data models are typically specified by a data specialist, data librarian, or a digital humanities scholar in a data modeling notation. These notations are often represented in graphical form



Who is he????



- 17th Century Michael Florent Van Langren
- In 1644, the idea of statistical data presented in the form of graphical representation was attributed to Flemish astronomer Michael Florent Van Langren.

Savitribai Phule Pune University

Fourth Year of Artificial Intelligence and Data Science (2020 Course)

417522: Data Modeling and Visualization

Teaching Scheme: TH:
03 Hours/Week

Credit
03

Examination Scheme:
In-Sem (Paper): 30
Marks End-Sem (Paper):
70 Marks

Prerequisites Courses: Statistics (), Computer Graphics (), Database Management Systems ()

CO Course Objectives and Course Outcomes

Course Objectives:

- Creating an emerging data model for the data to be stored in a database
- Conceptualized representation of Data objects
- Create associations between different data objects, and the rules
- Organize data description, data semantics, and consistency constraints of data
- Identifying data trends
- Incorporate data visualization tools and reap transformative benefits in their critical areas of operations

Course Outcomes:

After completion of the course, learners should be able to-

CO1: Summarize data analysis and visualization in the field of exploratory data science

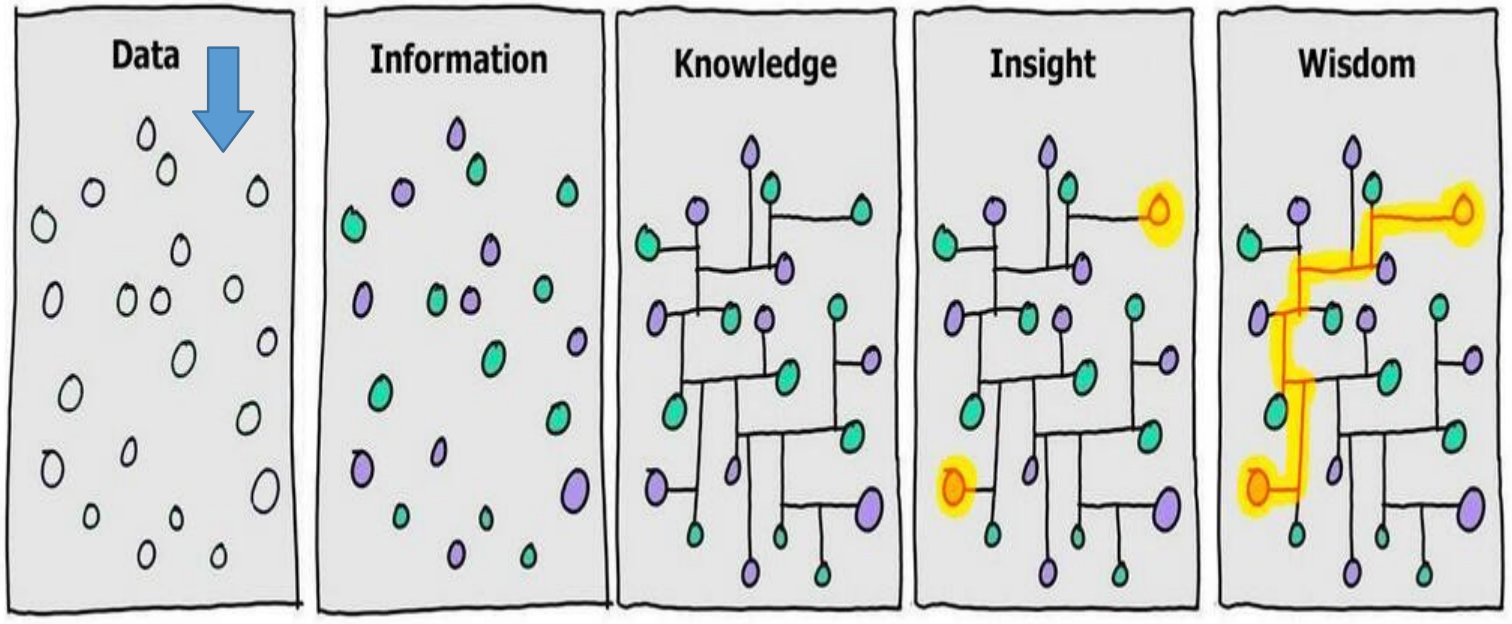
CO2: Analyze the characteristics and requirements of data and select an appropriate data model

CO3: Describe to load, clean, transform, merge and reshape data CO4: Design a probabilistic data modeling, interpretation, and analysis

CO5: Evaluate time series data

CO6: Integrate real world data analysis problems

The Information Continuum



Cartoon by David Somerville, based on a two pane version by Hugh McLeod

Types of Data

Quantitative Data

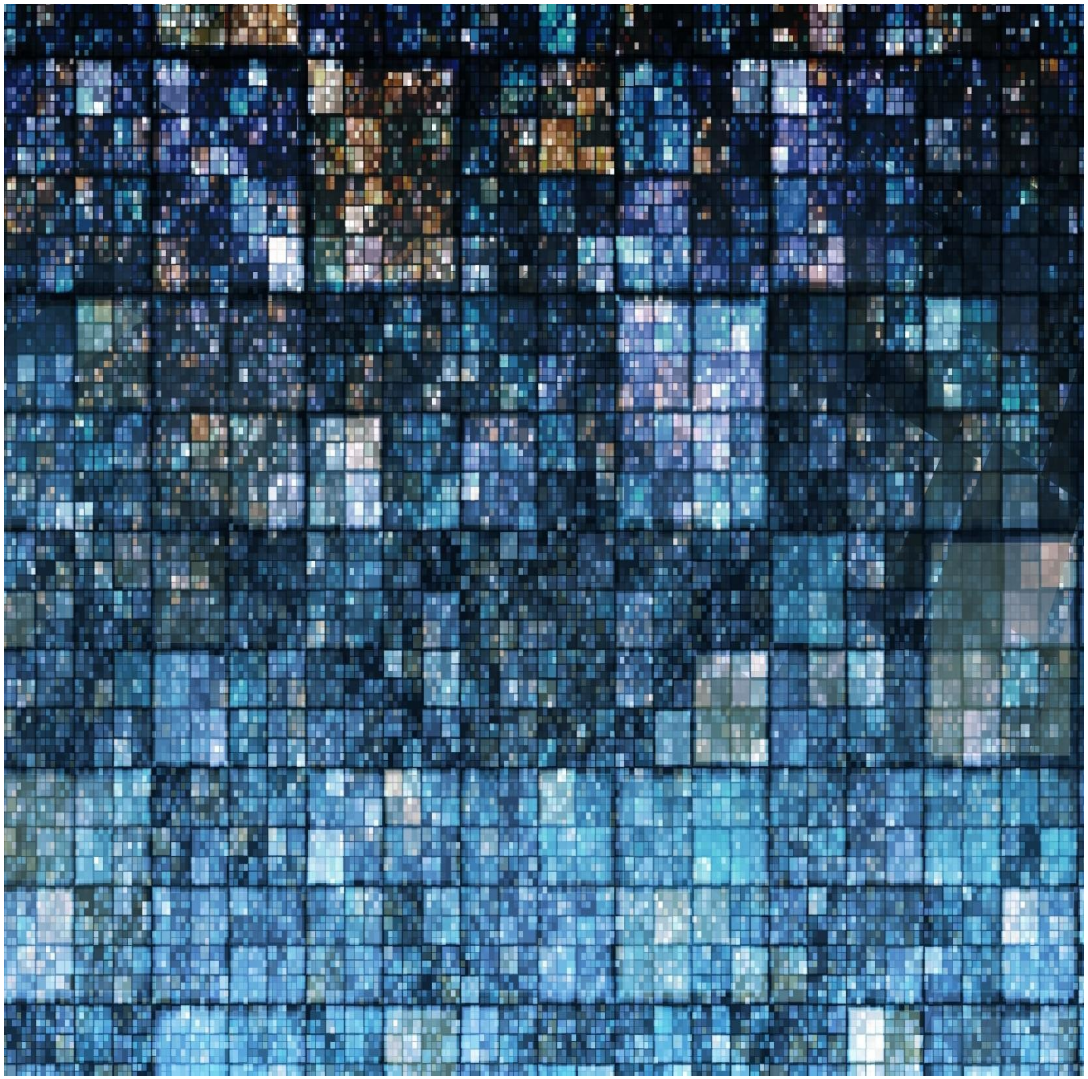
- Measurable
- Collected through measuring things that have a fixed reality
- Close ended

Qualitative Data

- Descriptive
- Collected through observation, field work, focus groups, interviews, recording or filming conversations
- Open ended

Big Data

Data that is too large or too complex to be managed using traditional data processing, analysis, and storage techniques.



What is Big Data?

- Big Data is a collection of large datasets that cannot be adequately processed using traditional processing techniques. Big data is not only data it has become a complete subject, which involves various tools, techniques and frameworks.

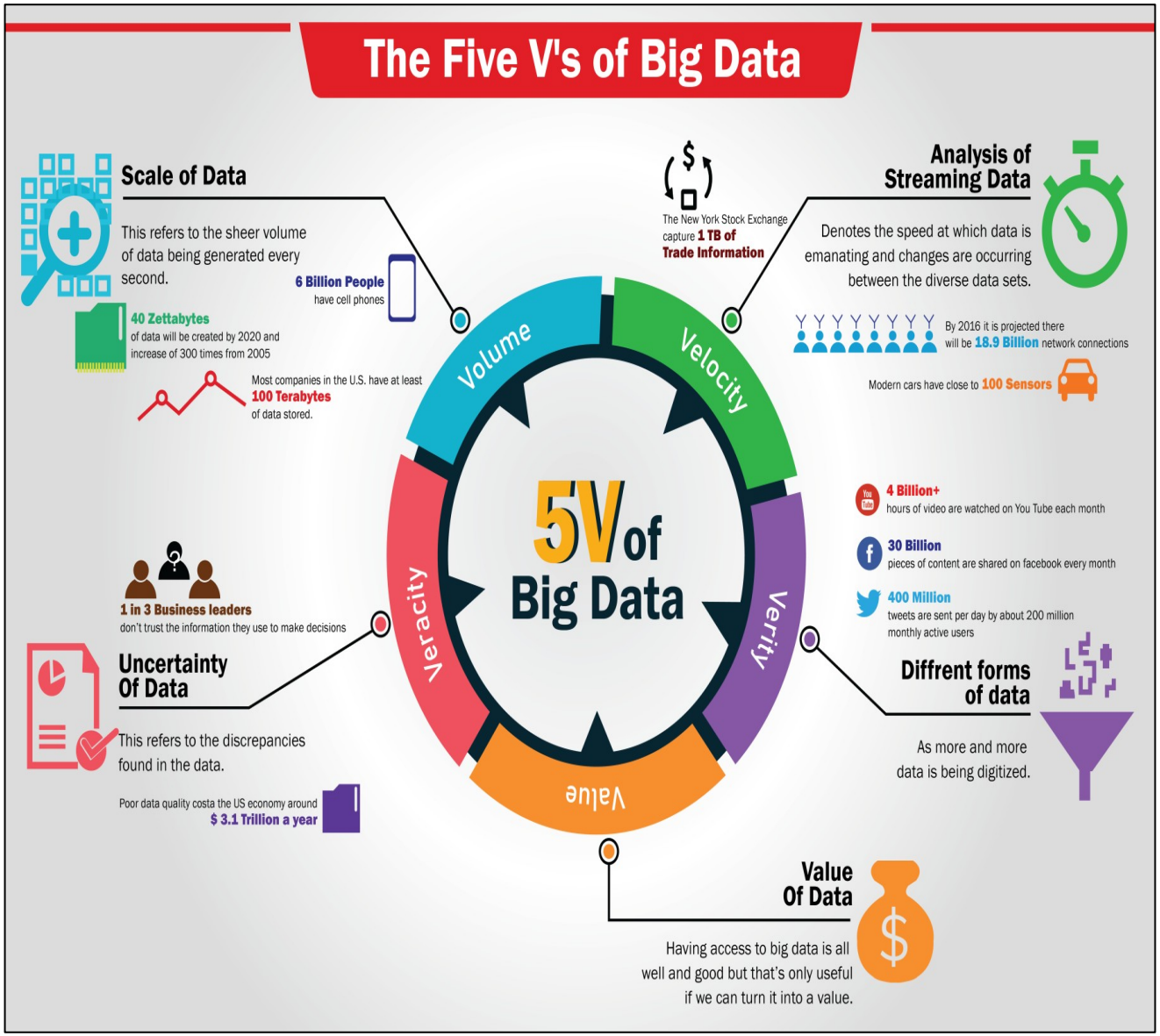
- Big data term describes the volume amount of data both structured and unstructured manner that adapted in day-to-day business environment. It's important that what organizations utilize with these with the data that matters.

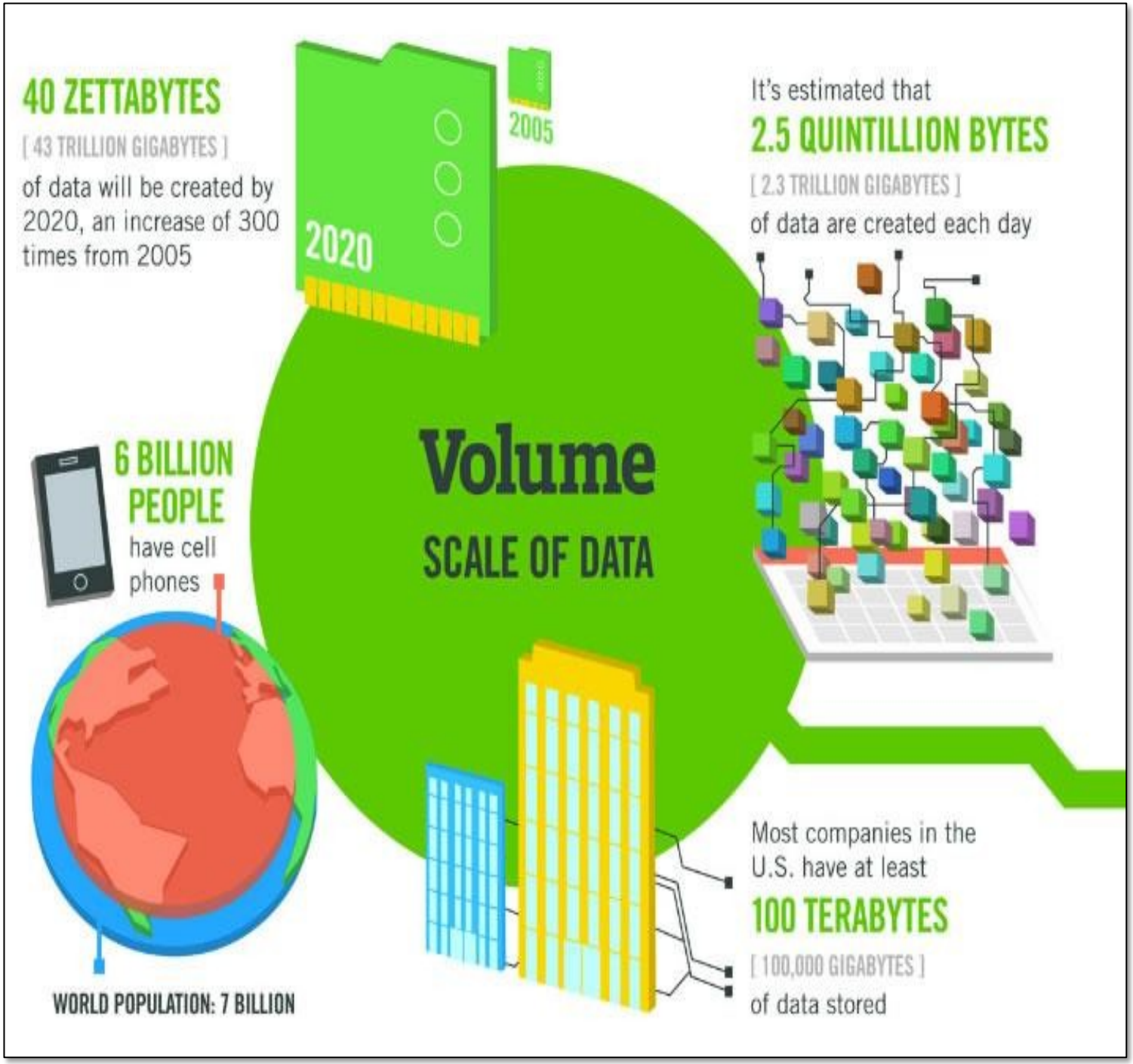
- Big data helps to analyze the in-depth concepts for the better decisions and
strategic taken for the development of the

The Evolution of Big Data

The concept of Big Data came into existence in the early 2000s when Industry analyst Doug Laney the definition of big data as the three categories as follows:

- **Volume:** Organizations collects the data from relative sources, which includes business transactions, social media and information from sensor or machine-to-machine data. Before, storage was a big issue but now the advancement of new technologies (such as Hadoop) has reduced the burden.
- **Velocity:** Data streams unparalleled speed of velocity and have improved in timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in real time operations.
- **Variety:** Data comes in all varieties in form of structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.


The Five V's of Big Data
Scale of Data
This refers to the sheer volume of data being generated every second.
6 Billion People have cell phones
40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005
Most companies in the U.S. have at least 100 Terabytes of data stored.
The New York Stock Exchange capture 1 TB of Trade Information
Analysis of Streaming Data
Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.
By 2016 it is projected there will be 18.9 Billion network connections
Modern cars have close to 100 Sensors
Volume
Velocity
5V of Big Data
Veracity
Verity
Value
1 in 3 Business leaders don't trust the information they use to make decisions
Uncertainty Of Data
This refers to the discrepancies found in the data.
Poor data quality costa the US economy around $ 3.1 Trillion a year
4 Billion+ hours of video are watched on You Tube 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
Diffrent forms of data
As more and more data is being digitized.
Value Of Data
Having access to big data is all well and good but that's only useful if we can turn it into a value.


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Volume: scale of data

Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000^1 bytes	1,000 bytes
megabyte (MB)	1000^2 bytes	1,000,000 bytes
gigabyte (GB)	1000^3 bytes	1,000,000.000 bytes
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes

Volume: scale of data

- 90% of today's data has been created in just the last 2 years
- Every day we create 2.5 quintillion bytes of data or enough to fill 10 million Blu-ray discs
- 40 zettabytes (4o trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005, and the equivalent of 5,200 gigabytes of data for every man, woman and child on Earth
- Most companies in the US have over 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

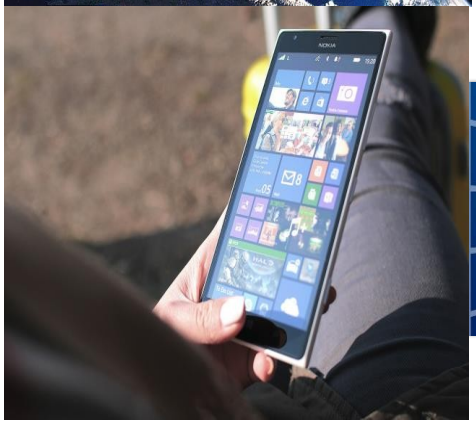
ANALYSIS OF STREAMING DATA

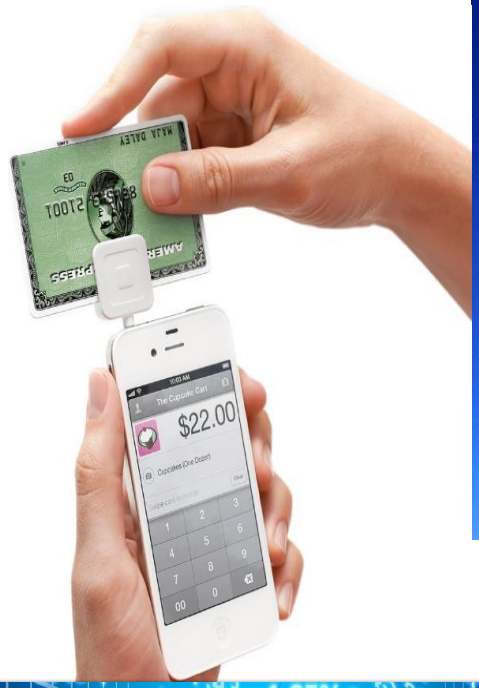
By 2016, it is projected there will be

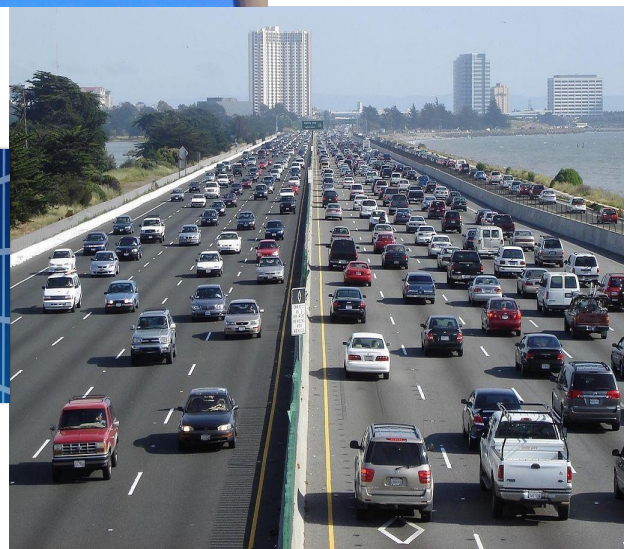
18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Velocity: analysis of streaming data







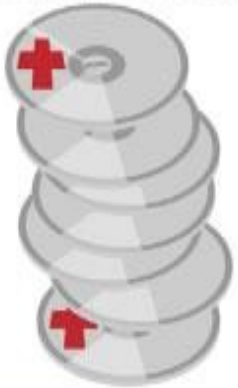
Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

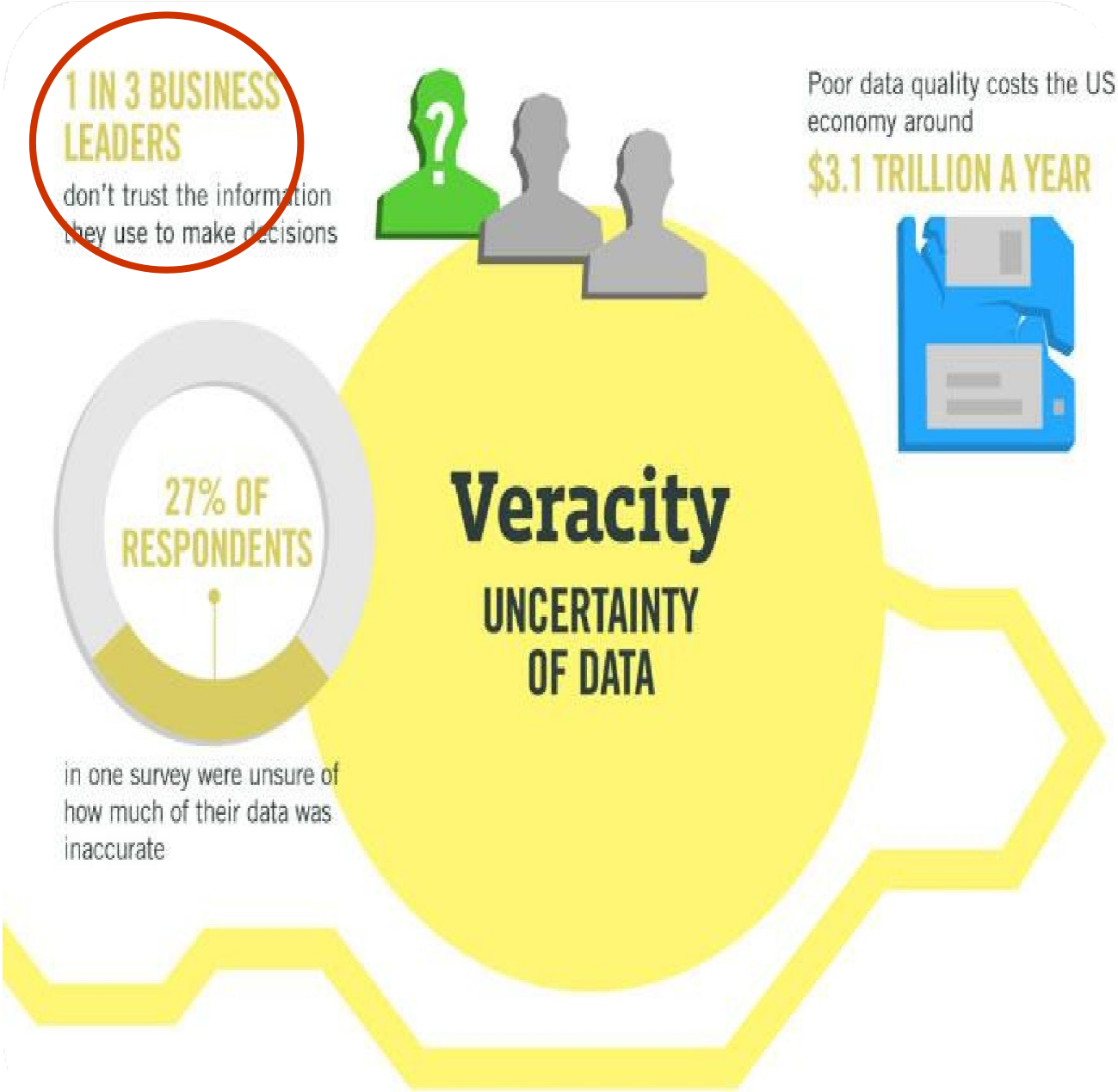
4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

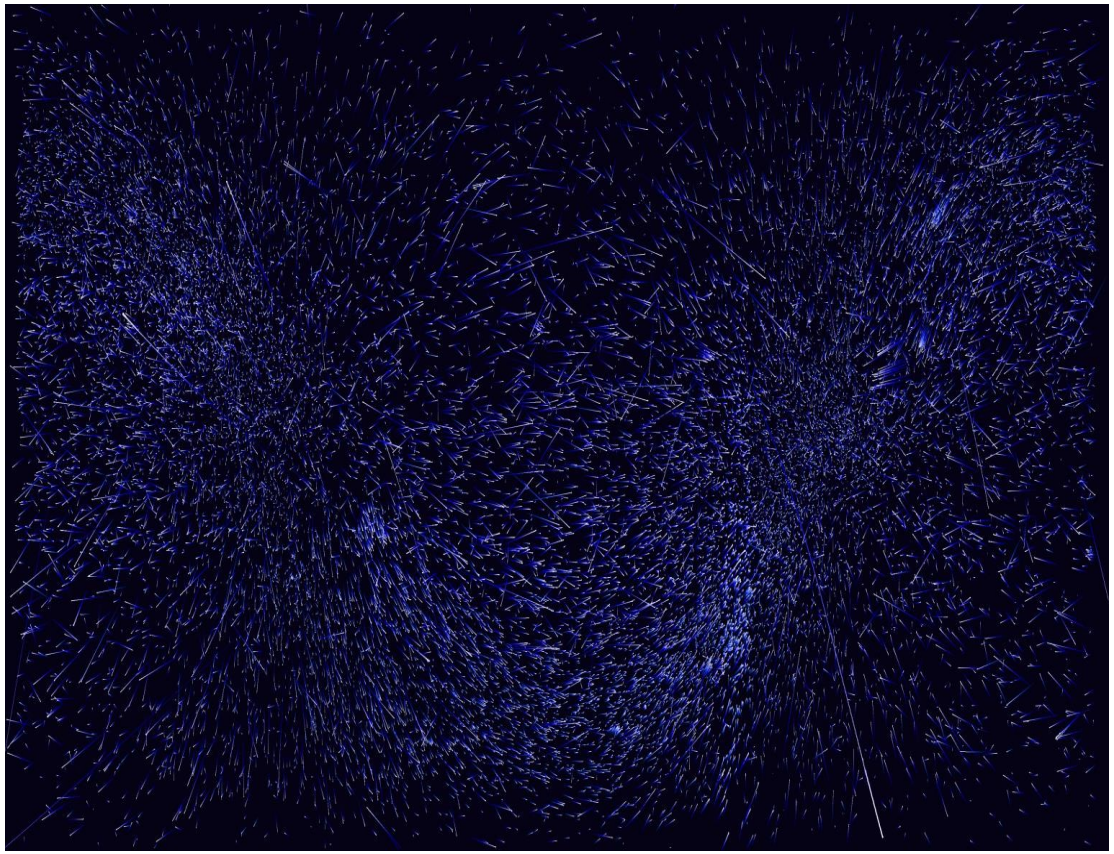
are sent per day by about 200 million monthly active users

Variety: different forms of data

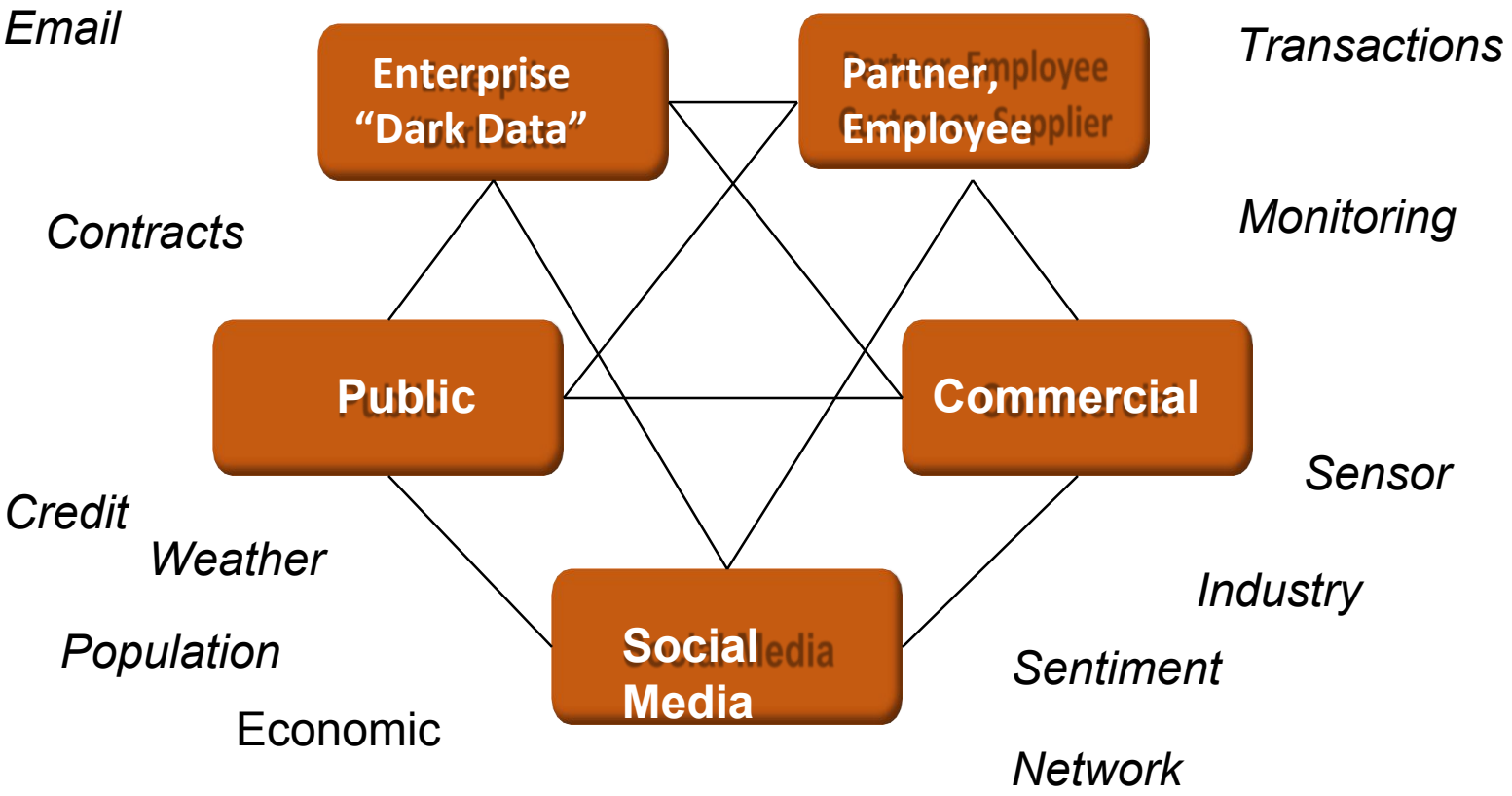

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Veracity: trustworthiness of data

- Origin
- Authenticity
- Trustworthiness
- Completeness
- Integrity



Where does Big Data come from?



Categories of Big Data - I

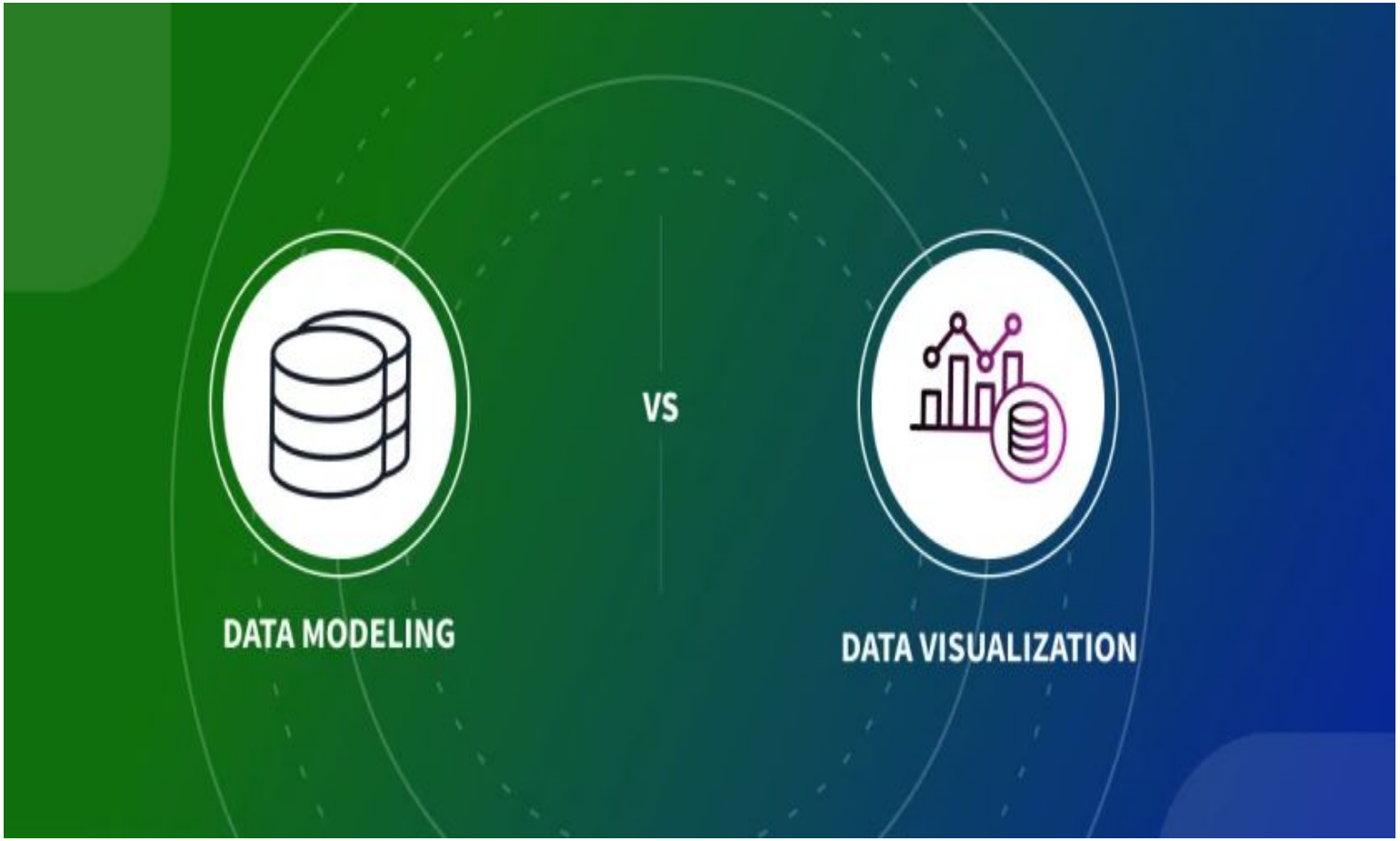
Big data works on the data produced by various devices and their applications.

- **Black Box Data:** It includes the conversation between crew members and any other communications (alert messages or any order passed) by the technical grounds duty staff.
- **Social Media Data:** Social networking sites such as Facebook and Twitter contains the information and the views posted by millions of people across the globe.
- **Stock Exchange Data:** It holds information (complete details of in and out of business transactions) about the 'buyer' and 'seller' decisions in terms of share between different companies made by the customers.

Who are the ones who use the Big Data Technology?

1. Banking
2. Government
3. Education
4. Health Care
5. Manufacturing
6. Retail

Data Modelling Vs Data Visualization



Data Modelling

- Data Modeling refers to the process of creating a visual representation of an entire information system or some of its parts to communicate the relationships between data points and structures.

- The purpose is to show the types of data stored in the system, the relationships among the data types, the formats and attributes of the data, and how the data can be grouped and organized.

- The Data Modeling process begins with the collection of information about business requirements from both stakeholders and end-users. The business requirements are then translated into data structures for the formulation of a concrete Database design.

- Today, Data Modeling finds its application across every sector you could possibly think of, from Financial Institutions to the Healthcare Industry. A study by LinkedIn rates Data Modeling as the fastest-growing profession in the present job market.

Why Data Model?

Data Visualization

- Data Visualization refers to the process of representing data and information graphically. By the use of visual elements like graphs, charts, and maps, Data Visualization tools offer an accessible way to view and understand trends and patterns in data.
- Data Visualization helps organizations to analyze huge volumes of data and make data-driven decisions. It also makes it easy for individuals and companies to understand data. Data Visualization is very useful today as companies are generating and collecting huge data volumes. It can help them to unmask hidden gems from data, which are good for growth.
-

Categories of Data Visualization

Geospatial

1. Cartograms 2. Density Maps
3. Flow Maps 4. Heat Maps

Hierarchical

1. Ring Charts 2. Sunburst Diagrams
3. Tree Diagrams

Multi-dimensional

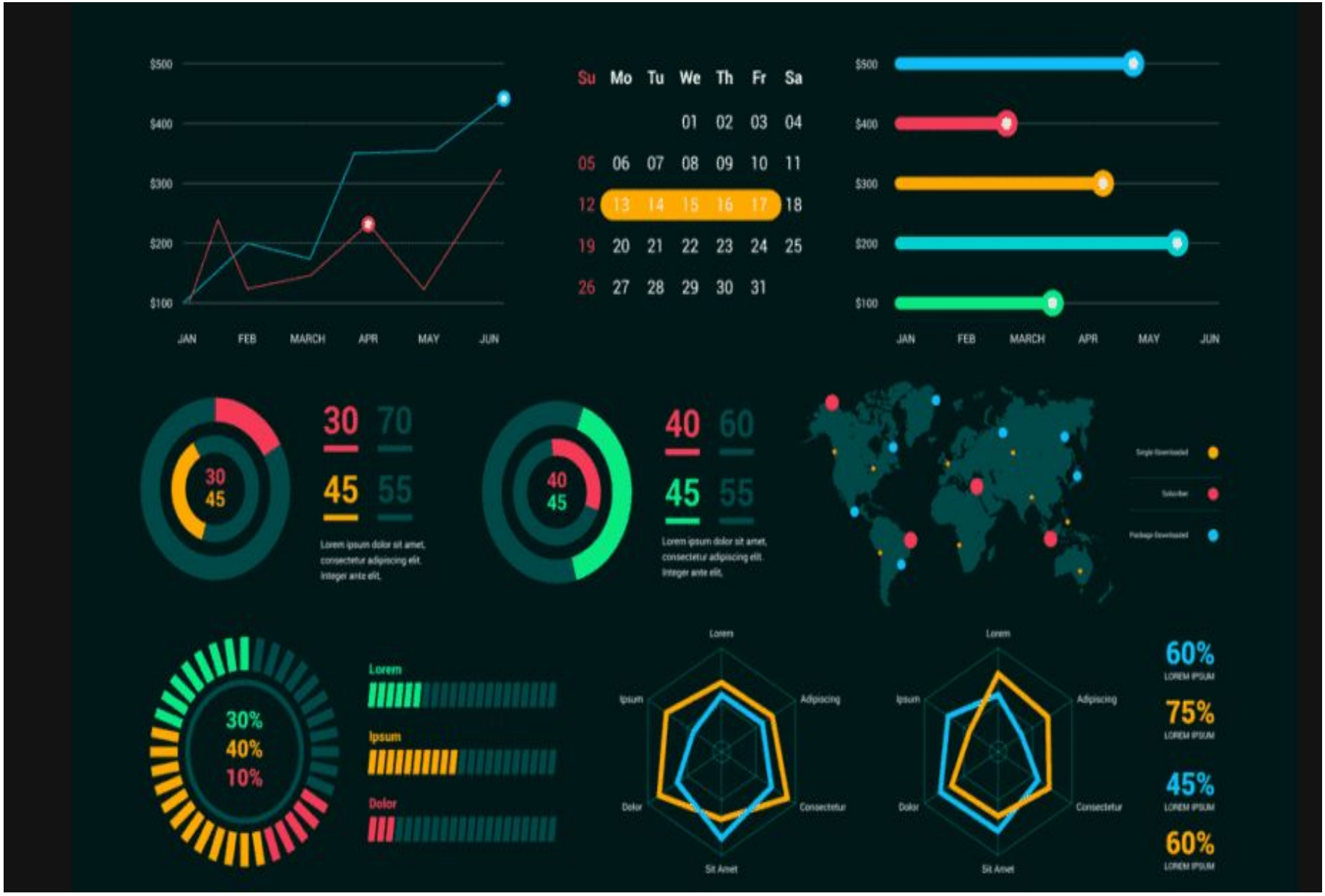
1. Histograms 2. Pie Charts
3. Stacked Bar Graphs 4. Venn Diagrams

Temporal

1. Line Graphs 2. Scatter Plots
3. Time Series Sequence 4. Timelines

Network

1. Matrix Charts 2. Alluvial Diagrams
3. Word Clouds 4. Node-link Diagrams


$500
$400
$300
$200
$100
JAN
FEB
MARCH
APR
MAY
JUN
Su Mo Tu We Th Fr Sa
01 02 03 04
05 06 07 08 09 10 11
12 13 14 15 16 17 18
19 20 21 22 23 24 25
26 27 28 29 30 31
30
45
30 70
45 55
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer ante elit.
40
45
40 60
45 55
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer ante elit.
30%
40%
10%
Lorem
Ipsum
Dolor
Lorem
Adipiscing
Consectetur
Sit Amet
Dolor
Ipsum
60%
LOREM IPSUM
75%
LOREM IPSUM
45%
LOREM IPSUM
60%
LOREM IPSUM

Difference between DM and DV

Feature	Data Modeling	Data Visualization
Definition	Data Modeling refers to **designing** the Entity-Relationship modeling for Database tables to establish the connections between tables. It also involves designing the schema for Data Warehouses. Thus, it shows how tables are connected in schema terms.	Data Visualization involves **presenting data** in a visual context to show hidden trends and patterns in data. Such trends and patterns may not be explicit in text data. Visualization makes data easy for anyone to understand.

Difference between DM and DV

Feature	Data Modeling	Data Visualization
Used For	Data Modeling is used to ensure that data is stored in a database and represented accurately. It shows the inherent structure of data by identifying data identities, attributes, and the relationship between the entities.	Data Visualization is used to communicate information clearly and efficiently to the users by presenting it using visual elements.
Benefits	Facilitate faster access to data across the entire organization. Data Modeling also makes it easy to establish the correct structure of data and enforce compliance standards.	Helps businesses understand their customers, products, and processes better. This is good for sound decision-making and making predictions. TT
Tools	Common data modelling tools include Erwin Data Modeler, ER/Studio, DbSchema, ERBuilder, HeidiSQL, Navicat Data Modeler, Toad Data Modeler, Archi, and others.	Data Visualization is done using tools such as Knowi, Tableau, Dygraphs, QlikView, DataHero, ZingCHhart, Domo, and others. It can also be done in programming languages such as Python and R.
Performed By	Data Architects and Modelers.	Data Engineers.