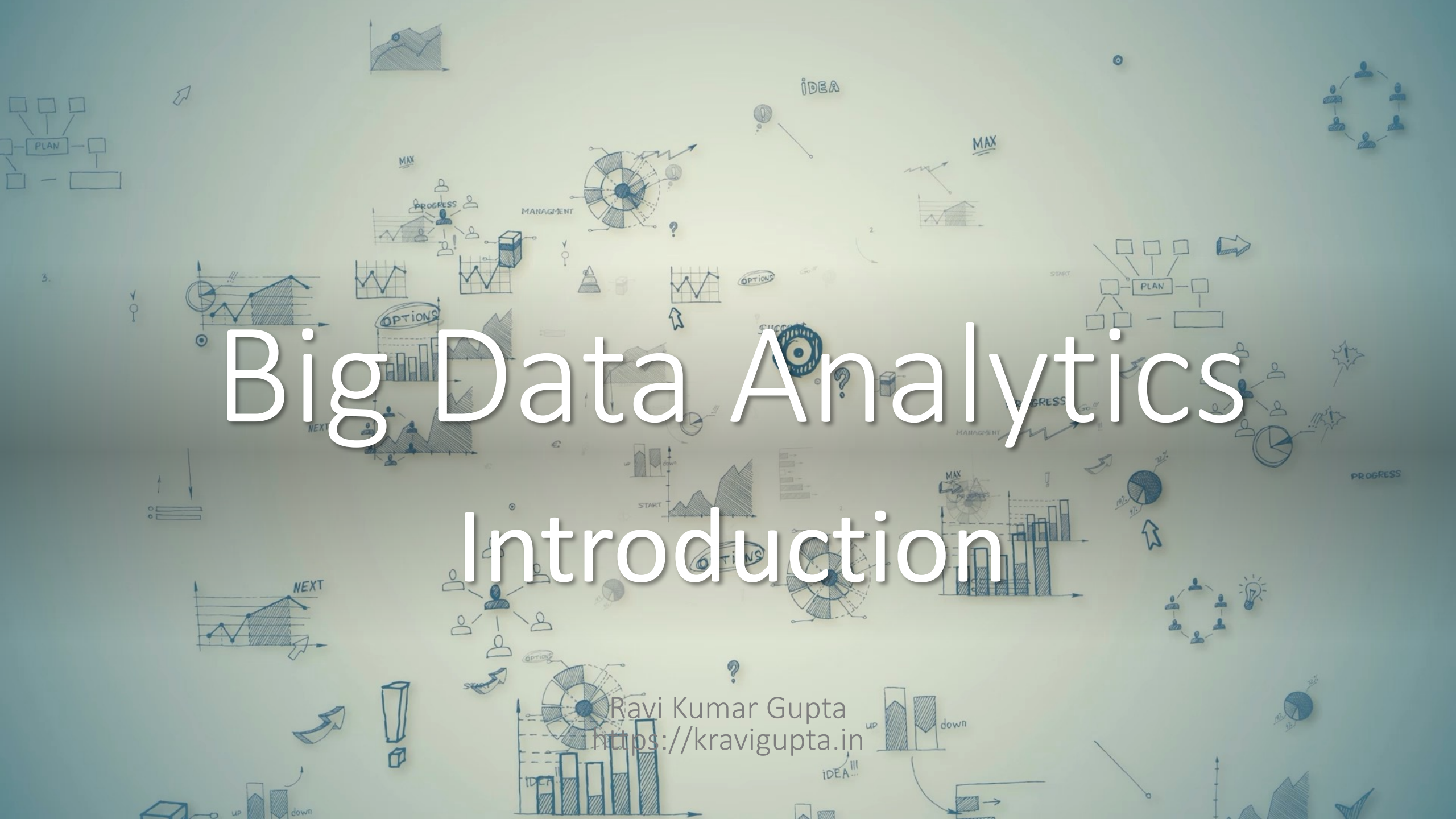
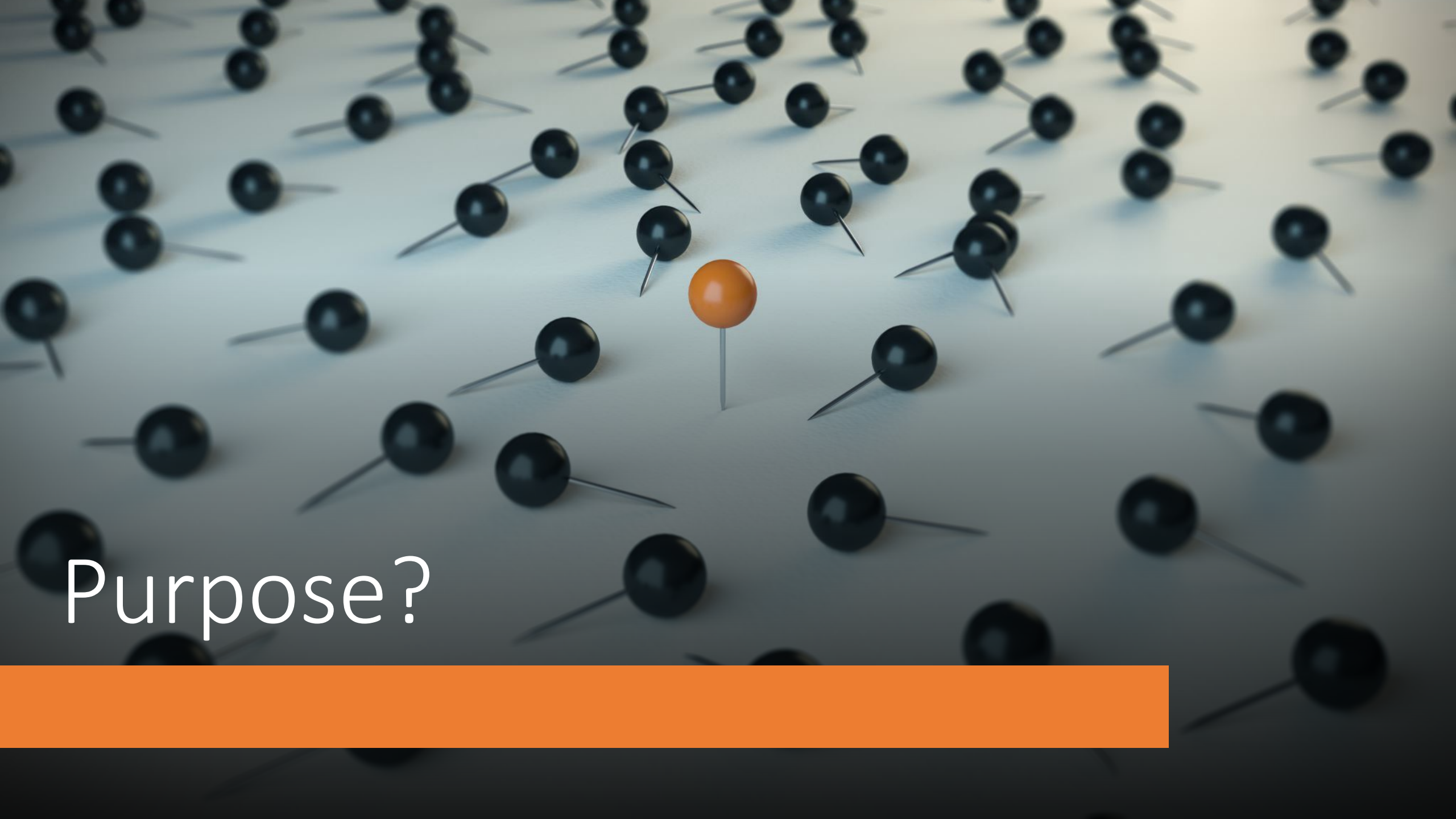


Big Data Analytics

Introduction

Ravi Kumar Gupta
https://kravigupta.in



Purpose?

Course Details

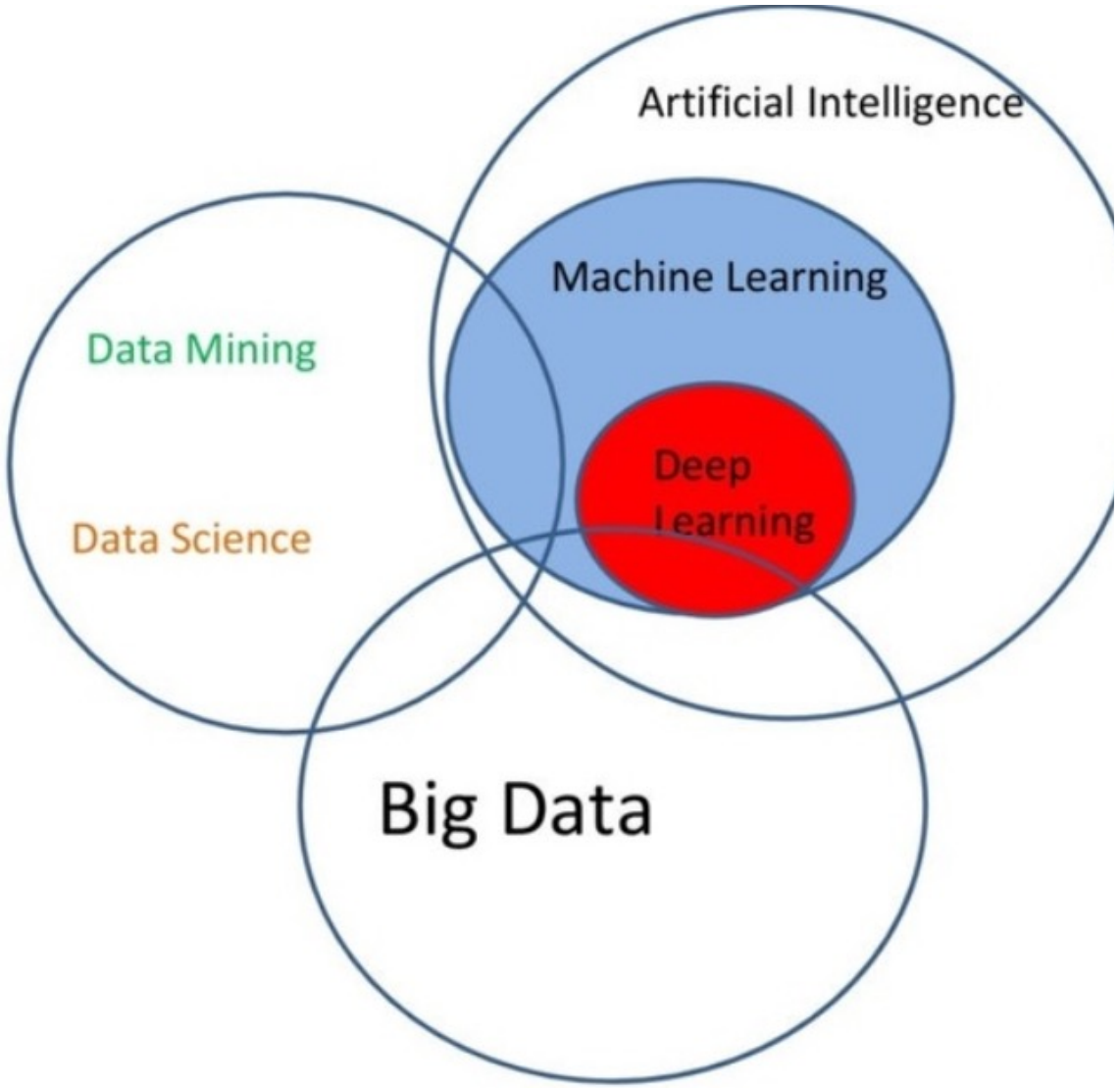
Course Code	Course Name	Teaching Scheme (Contact Hours)			Credits Assigned			
		Theory	Practical	Tutorial	Theory	Practical	Tutorial	Total
ECCDLO 7014	Big Data Analytics	03	--	--	03	--	--	03

Course Code	Course Name	Examination Scheme							
		Theory Marks				Exam Duration (Hrs.)	Term Work	Practical and Oral	Total
		Internal Assessment			End Sem. Exam.				
		Test1	Test2	Avg.					
ECCDLO 7014	Big Data Analytics	20	20	20	80	03	--	--	100

What will you learn?

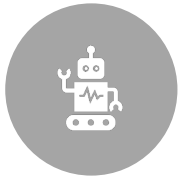
- Data Analytics

Science of analyzing the raw data in order to make meaningful conclusions



What’s not covered?

ARTIFICIAL INTELLIGENCE



ROBOTICS

MACHINE LEARNING

DEEP LEARNING

NEURAL NETWORKS

PROGRAMMING LANGUAGES

WEB/MOBILE APP DEVELOPMENT

Requirements

Technical Requirements

- Softwares
  - VSCode
  - XAMPP or similar
  - MongoDB
  - Python
- Github
- Cloud
  - AWS

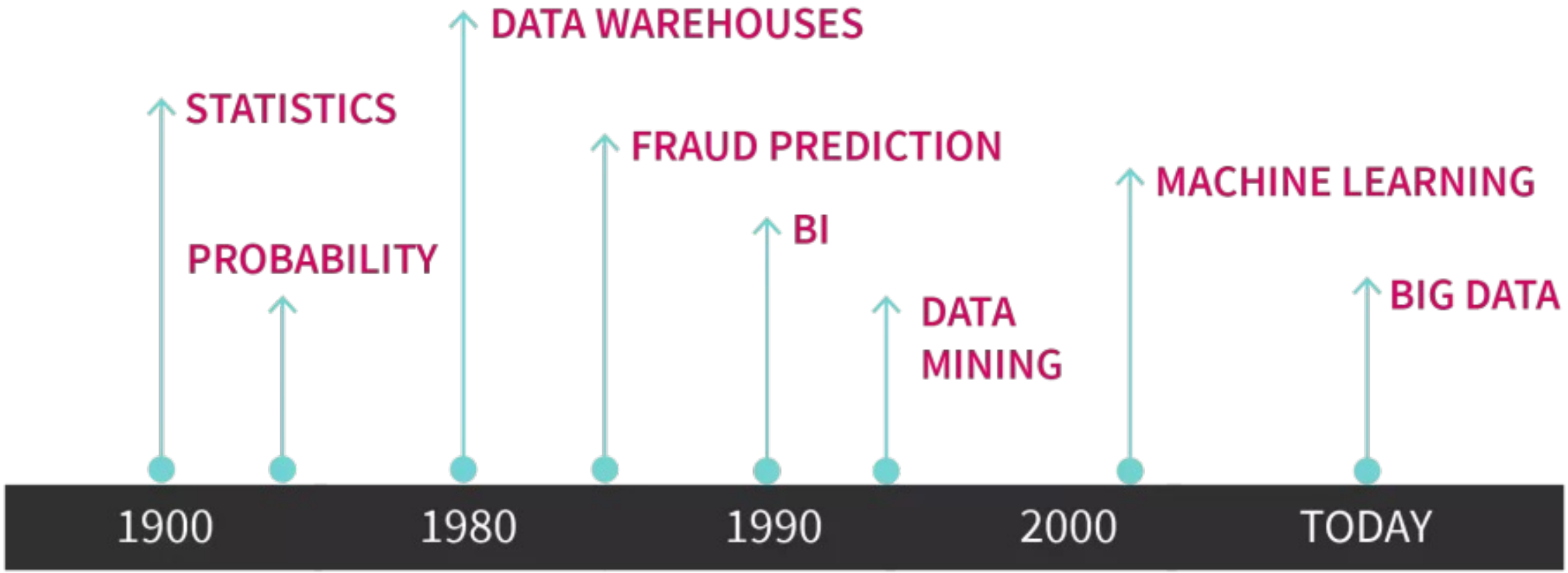
Skills

- Technical
  - Familiarity with Database and related concepts
- Teamwork
- Problem Solving Approach
- Attitude

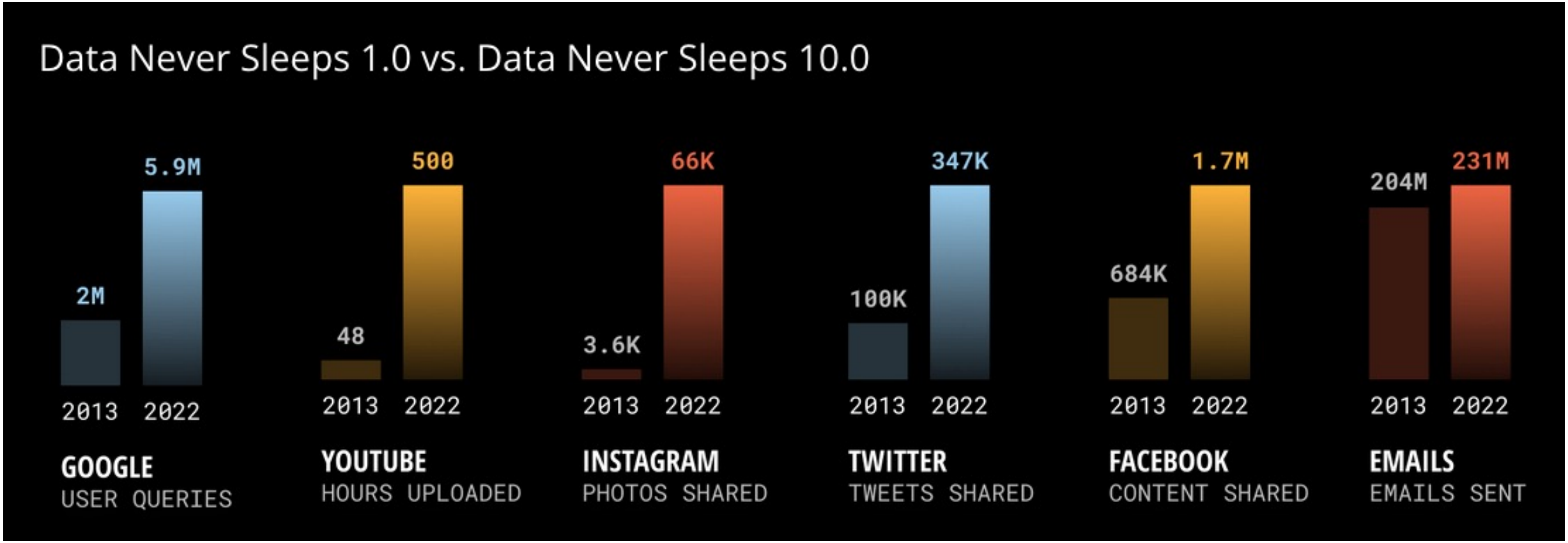

What exactly is Big Data?

What qualifies as Big Data?

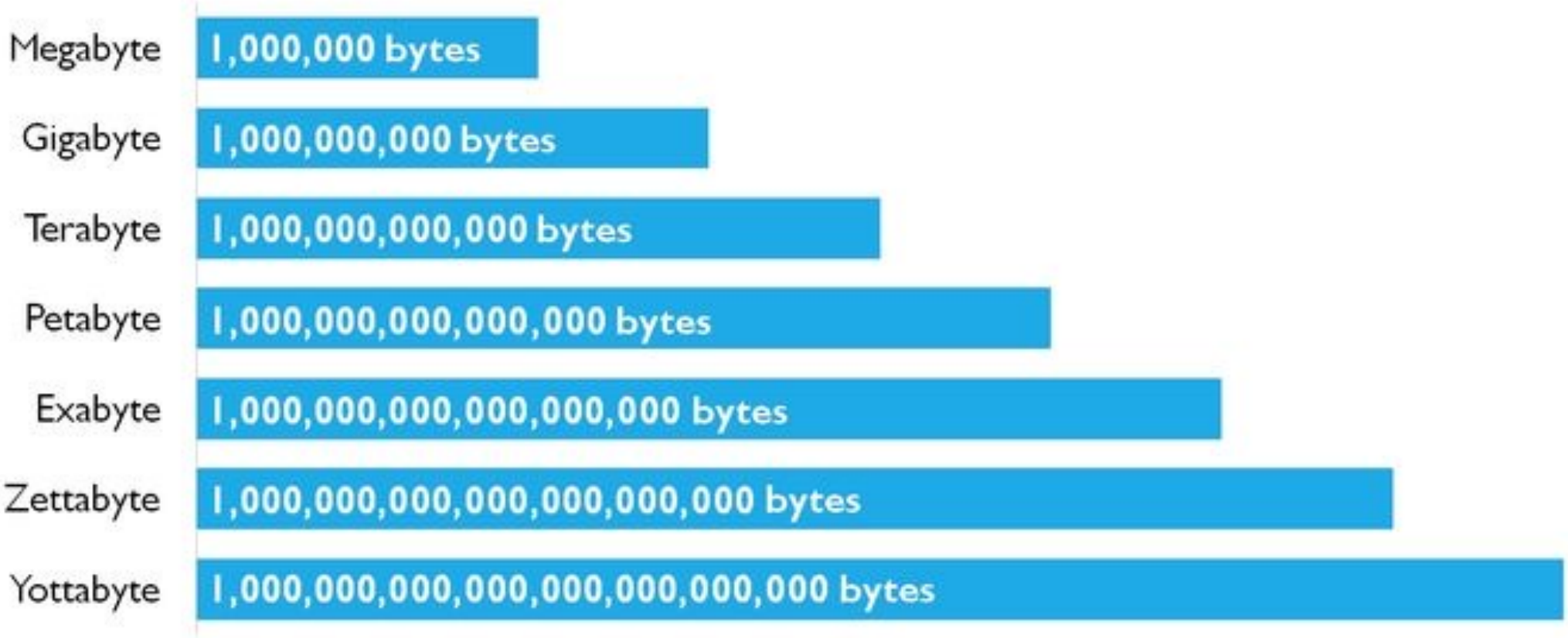
- Share market data?
- Google Search queries?
- Tweets?
- Photos/Videos shared online?


DATA WAREHOUSES
STATISTICS
FRAUD PREDICTION
MACHINE LEARNING
BI
PROBABILITY
DATA MINING
BIG DATA
1900
1980
1990
2000
TODAY

One minute of your life



Computer Memory



...

Table 1: Data Measurement Units

Unit	Abbreviation	Decimal Value	Binary Value	Decimal Size
bit	b	0 or 1	0 or 1	1/8 of a byte
byte	B	8 bits	8 bits	1 byte
kilobyte	KB	$1{,}000^1$ bytes	$1{,}024^1$ bytes	1,000 bytes
megabyte	MB	$1{,}000^2$ bytes	$1{,}024^2$ bytes	1,000,000 bytes
gigabyte	GB	$1{,}000^3$ bytes	$1{,}024^3$ bytes	1,000,000,000 bytes
terabyte	TB	$1{,}000^4$ bytes	$1{,}024^4$ bytes	1,000,000,000,000 bytes
petabyte	PB	$1{,}000^5$ bytes	$1{,}024^5$ bytes	1,000,000,000,000,000 bytes
exabyte	EB	$1{,}000^6$ bytes	$1{,}024^6$ bytes	1,000,000,000,000,000,000 bytes
zettabyte	ZB	$1{,}000^7$ bytes	$1{,}024^7$ bytes	1,000,000,000,000,000,000,000 bytes
yottabyte	YB	$1{,}000^8$ bytes	$1{,}024^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

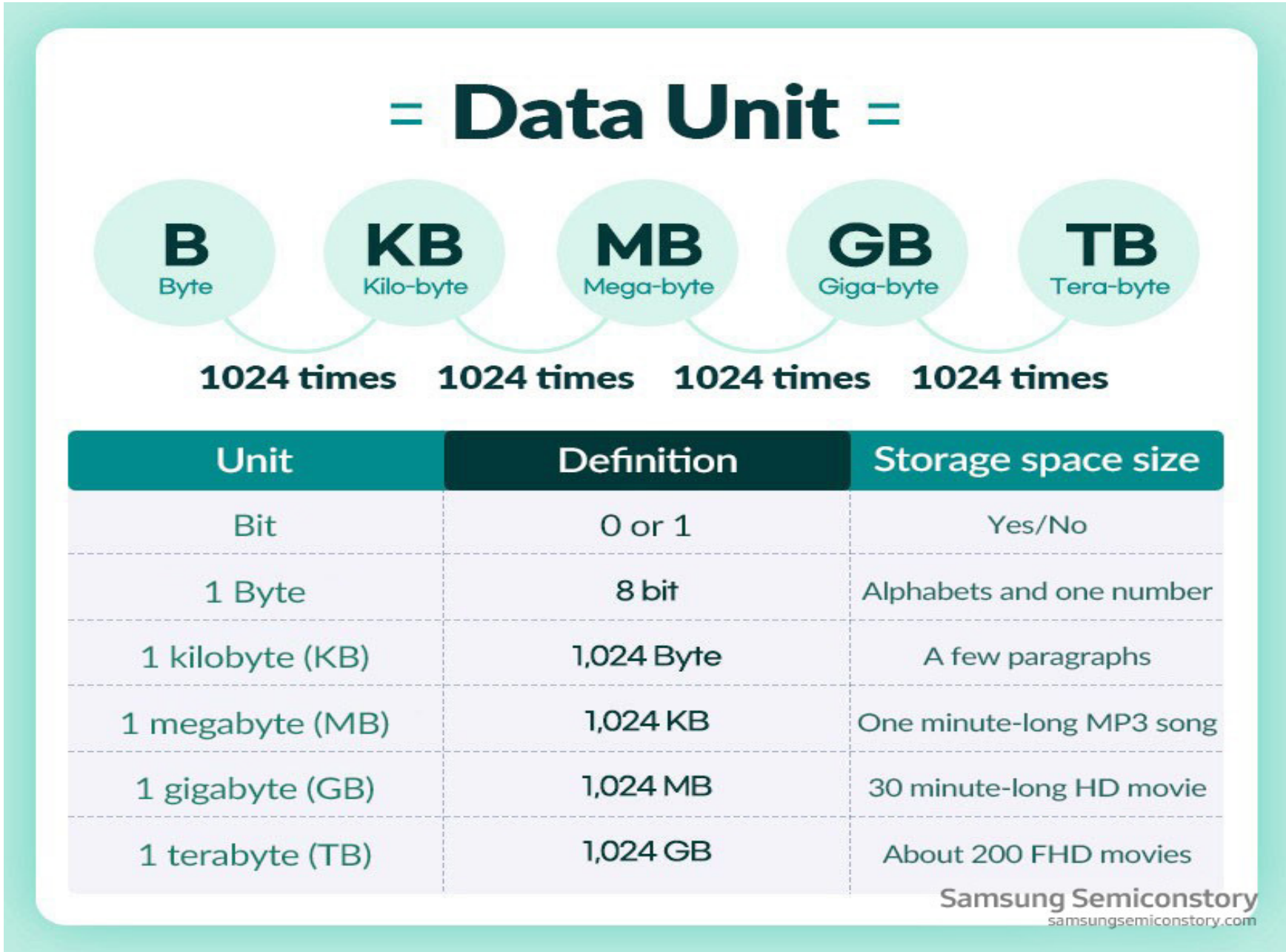
Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

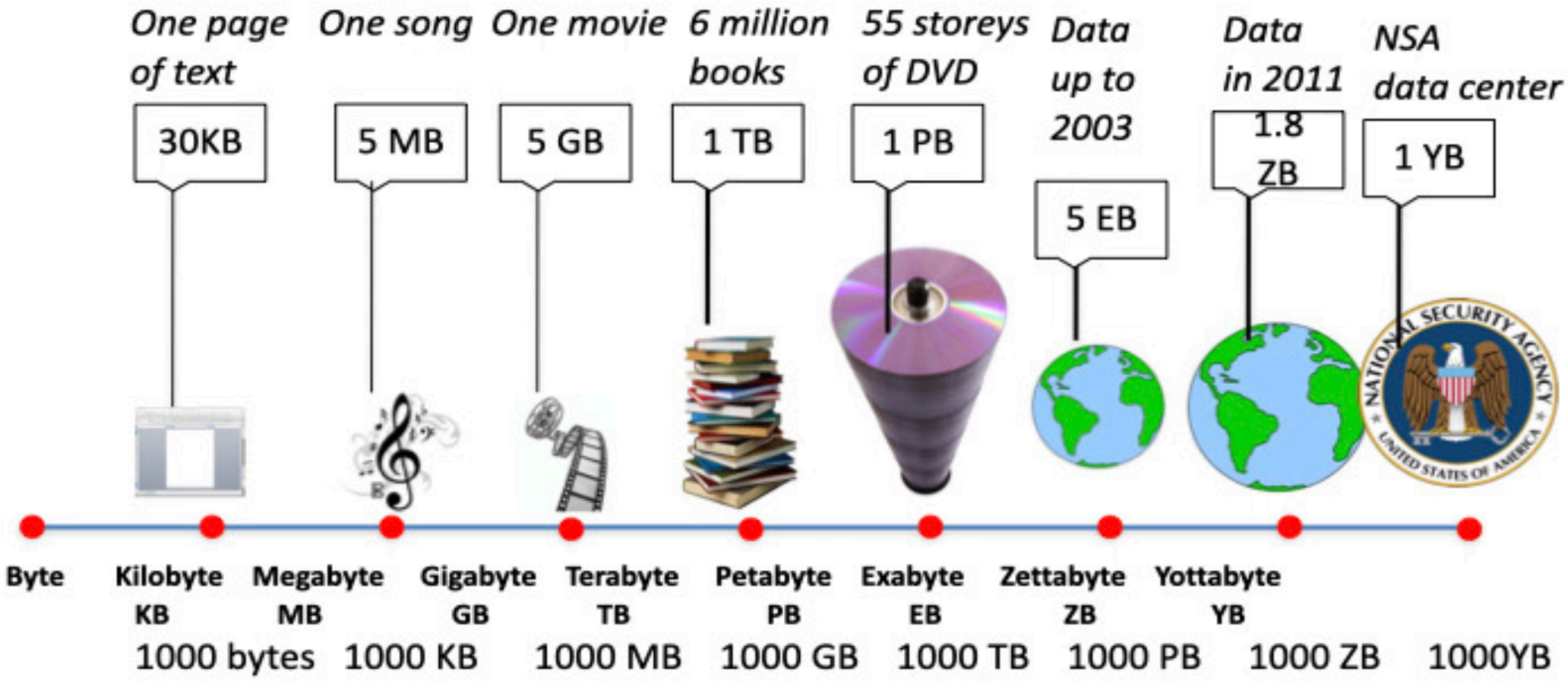
Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

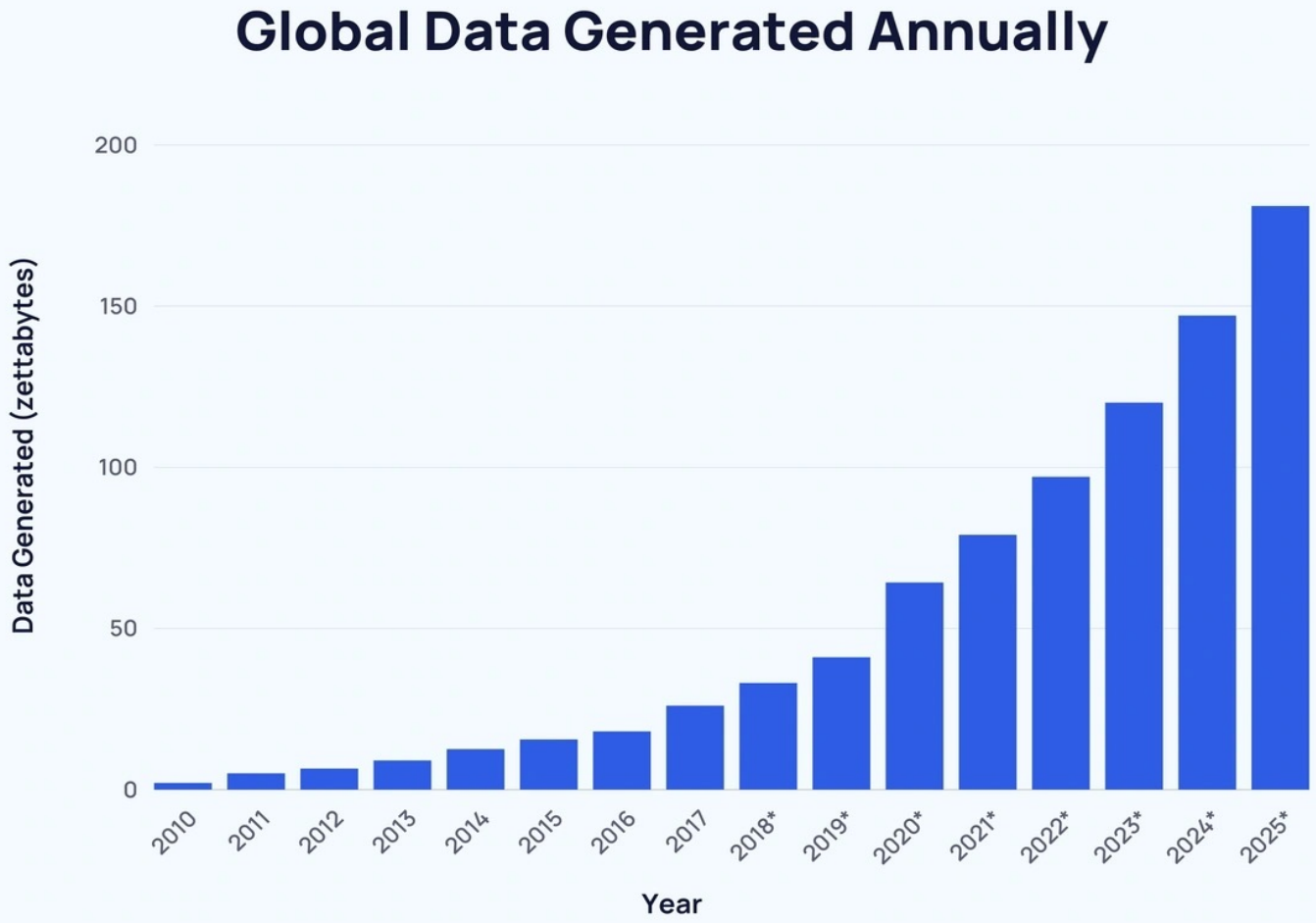


Source: https://semiconductor.samsung.com/support/tools-resources/dictionary/bits-and-bytes-units-of-data/

Big Data: Volume



Data Generated per year..



Cost of Data Storage

1 TB HDD COSTS AROUND $30

HOW MUCH FOR 1 ZB?

So, when we call it, “The Big Data”

Big Data Characteristics

The 3 V's of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3 V's: *volume*, *velocity* and *variety*.

Volume	Velocity	Variety
The amount of data from myriad sources.	The speed at which big data is generated.	The types of data: structured, semistructured, unstructured.

ICONS: PRESSURELUA/GETTY IMAGES

Volume

Massive amount of data generated every minute

- Social Media
- Stocks

Beyond the capability of traditional databases

- Limitations on Memory
- Processing capabilities

Influenced by many factors

- IoT, Business Transactions, Social Media

Variety

Different types of data

- Structured
- Semi-structured
- Unstructured

Integration challenges due to varying formats

- Ingesting data from different sources
- Data of different format

Influenced by diverse data sources

- Databases, text, images, videos, other formats

Velocity

Speed of data generation

- Real time data streams
- User generated data
- Machine generated data

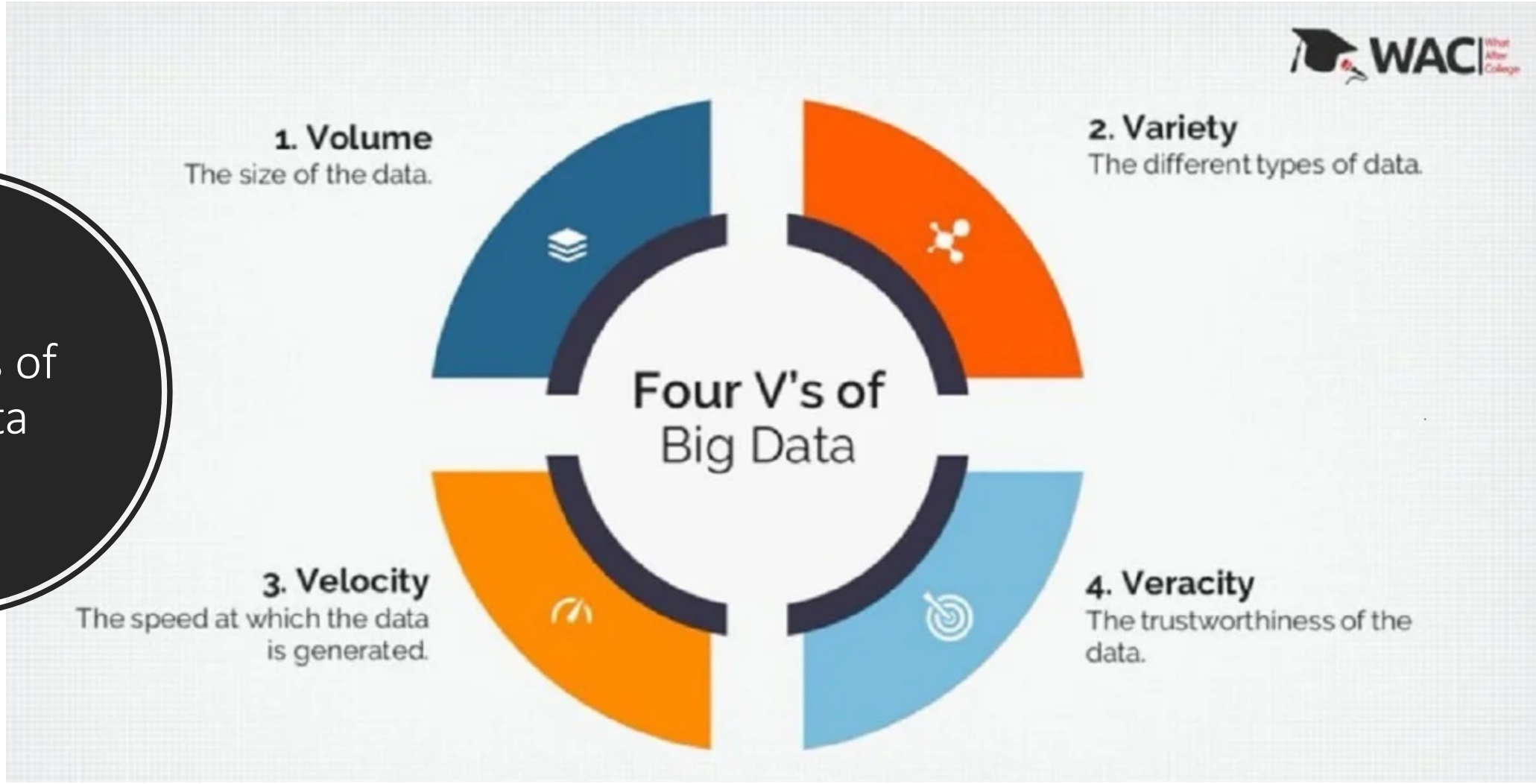
Speed of data processing

- Real time analysis
- Stream processing
- Data in motion – Vs – Data at rest

Influenced by diverse data sources

- IoT, Business Transactions, Social Media, live streaming, sensor data

The 4 Vs of Big Data



Veracity

Data Quality

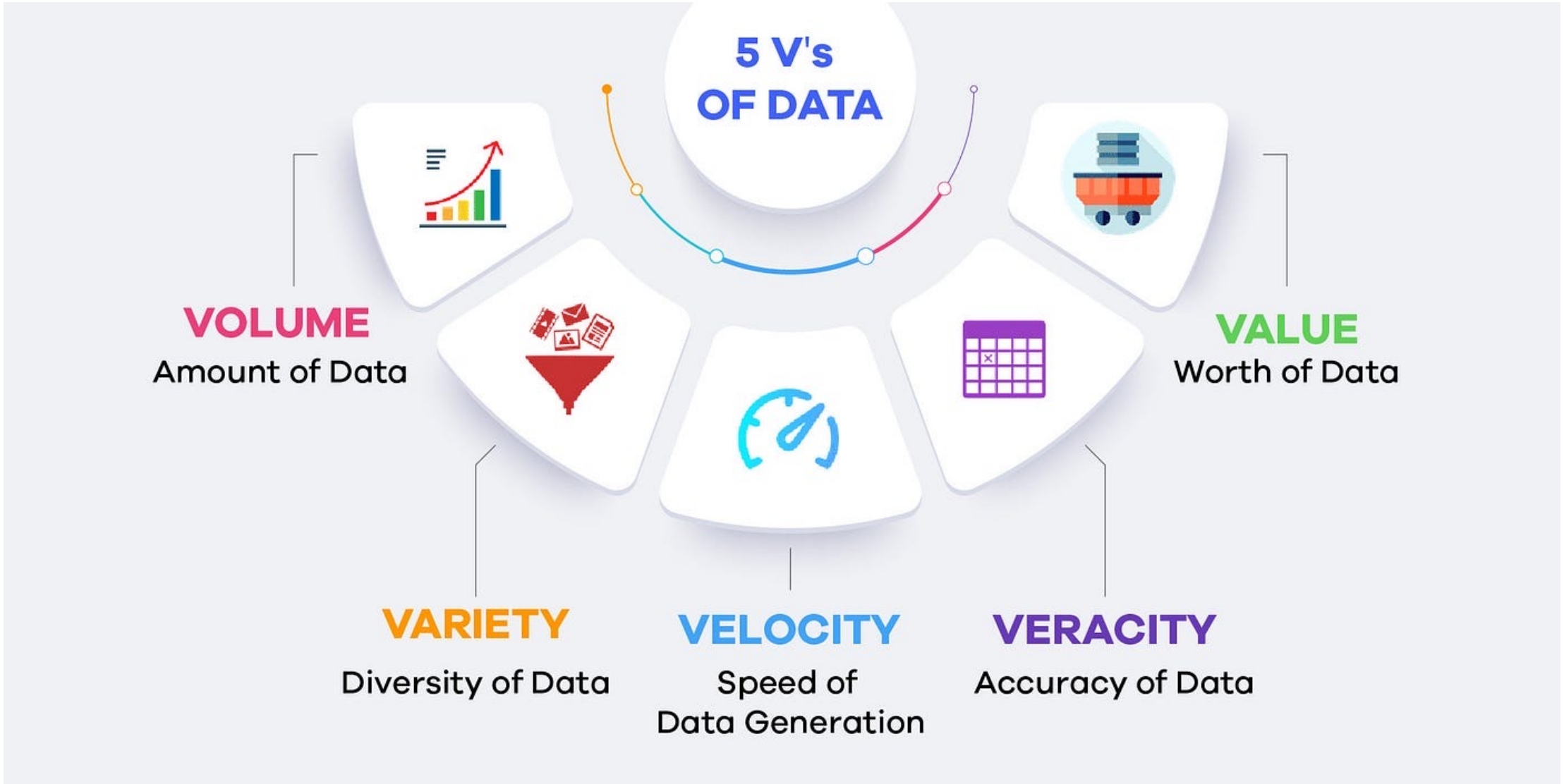
- Noisy, incomplete, or inconsistent
- Poor quality can lead to inaccurate analysis and decisions

Data Trustworthiness

- Depends on the source of data
- Critical to verify the authenticity and reliability

Data Uncertainty

- Data collection process
- Sampling errors
- Missing Data, Sampling errors and more..


5 V's
OF DATA
VOLUME
Amount of Data
VALUE
Worth of Data
VARIETY
Diversity of Data
VELOCITY
Speed of
Data Generation
VERACITY
Accuracy of Data

Value

Usefulness of data

- Decision Making
- Benefits to the organization, people

Data Insights

- Drives strategy
- Creates business intelligence

Finances

- Cost Saving, Data Monetization
- Improved customer experience
- Risk Mitigation,

5 Vs of Big Data

Volume:	Refers to the massive amount of data being generated every second.
Velocity:	Refers to the speed at which data is being generated and processed.
Variety:	Refers to the different types of data: structured, semi-structured, and unstructured.
Veracity:	Refers to the reliability and accuracy of data, emphasizing the importance of data quality.
Value:	Refers to the ability to turn raw data into meaningful insights that can lead to beneficial outcomes.

Need More Vs?

There are more Vs indeed!!!

Variability

Peaks and lows in data sets

Variable data flow rates

Visualization

Ability to present complex data visually

May be a chart, table, image etc.

Easy to Understand and Interpret the trends and patterns

Validity | Vulnerability

Different sources – Different opinion

Vs of Big Data

Velocity

Volume

Variety

Veracity

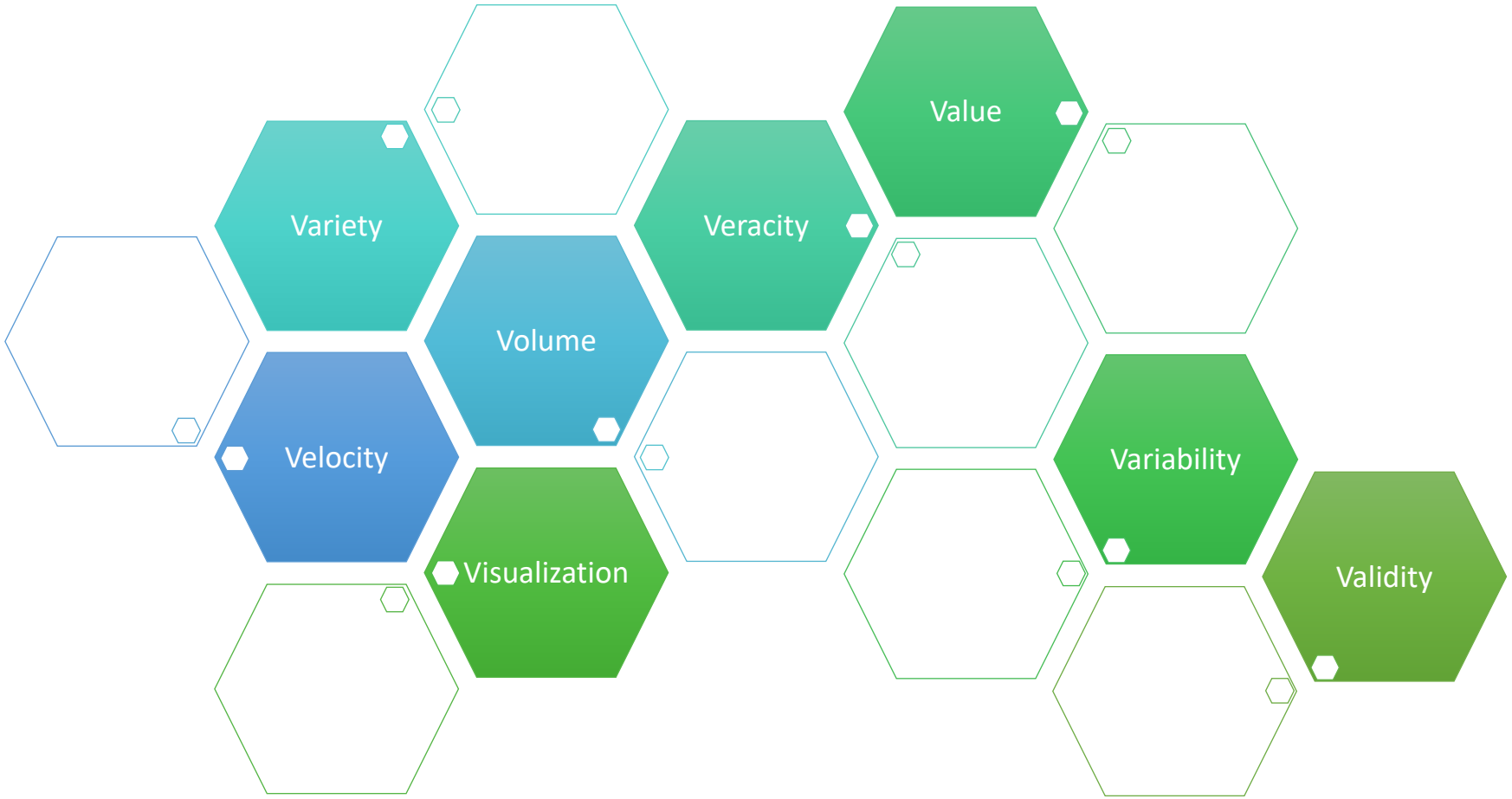
Value

Variability

Visualization

Validity

Vs of Big Data



So, Big Data

- Refers to
  - Massive Volume of Data
  - Collected from Variety of data sources
- Big Data is -
  - Both Structured and Unstructured data
  - Large and Complex
  - Variable in nature
  - Difficult to process using traditional databasesx

Types of Big Data

Types of Big Data

- Structured Data
- Semi Structured Data
- Unstructured Data

Structured Data

Structured Data

- Tables from Relational Databases
- Examples
  - Customer information in CRM (Customer Relationship Management)
  - E-Commerce Transaction data
  - Employees Data in HRMS

Semi Structured Data

Unstructured Data

Semi-Structured Data

Structured Data

Semi Structured Data

- Mix of Structured and unstructured Data
- Examples
  - XML, JSON, YAML files
  - Email Data

Unstructured Data

Unstructured Data

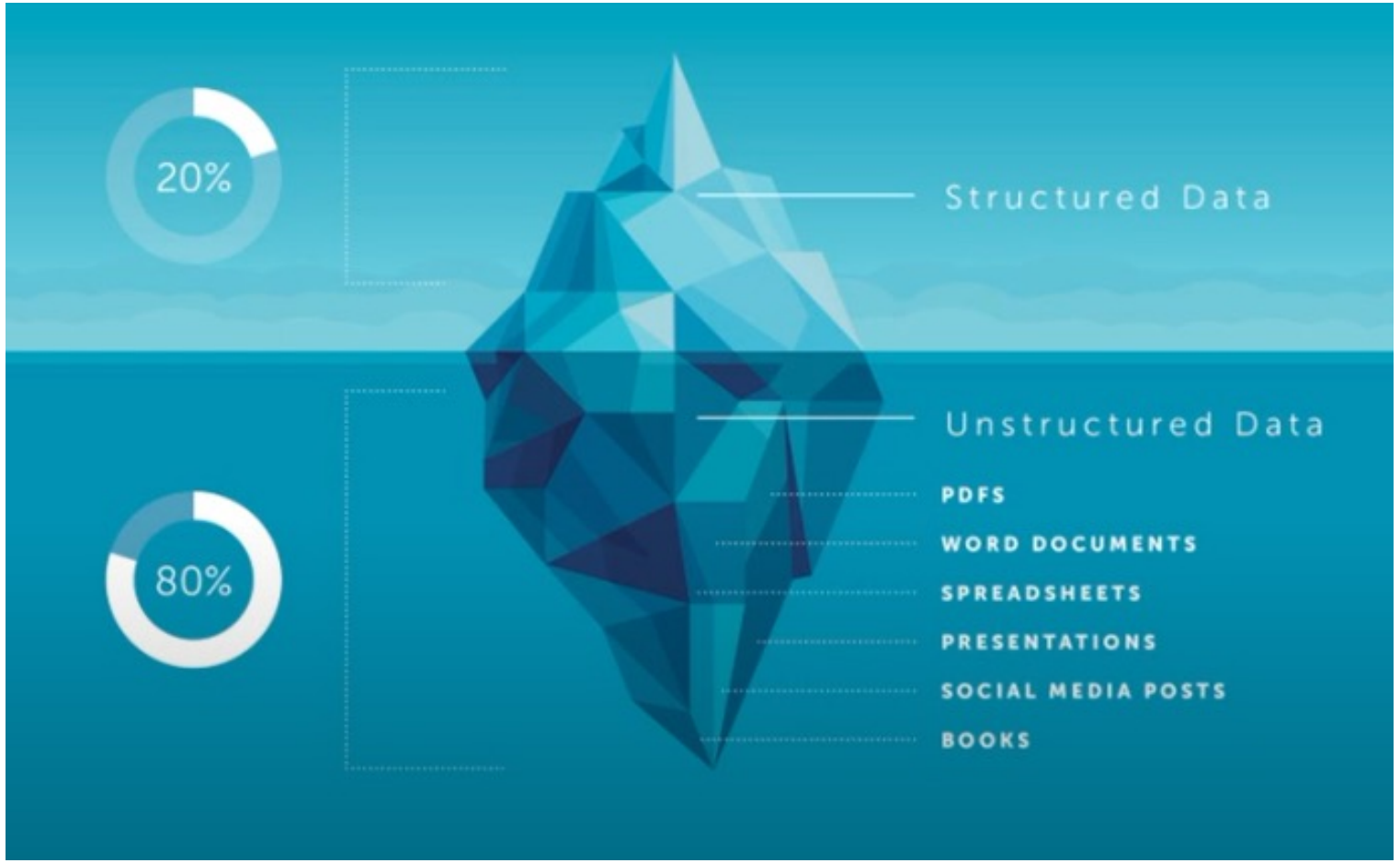
Structured Data

Semi Structured Data

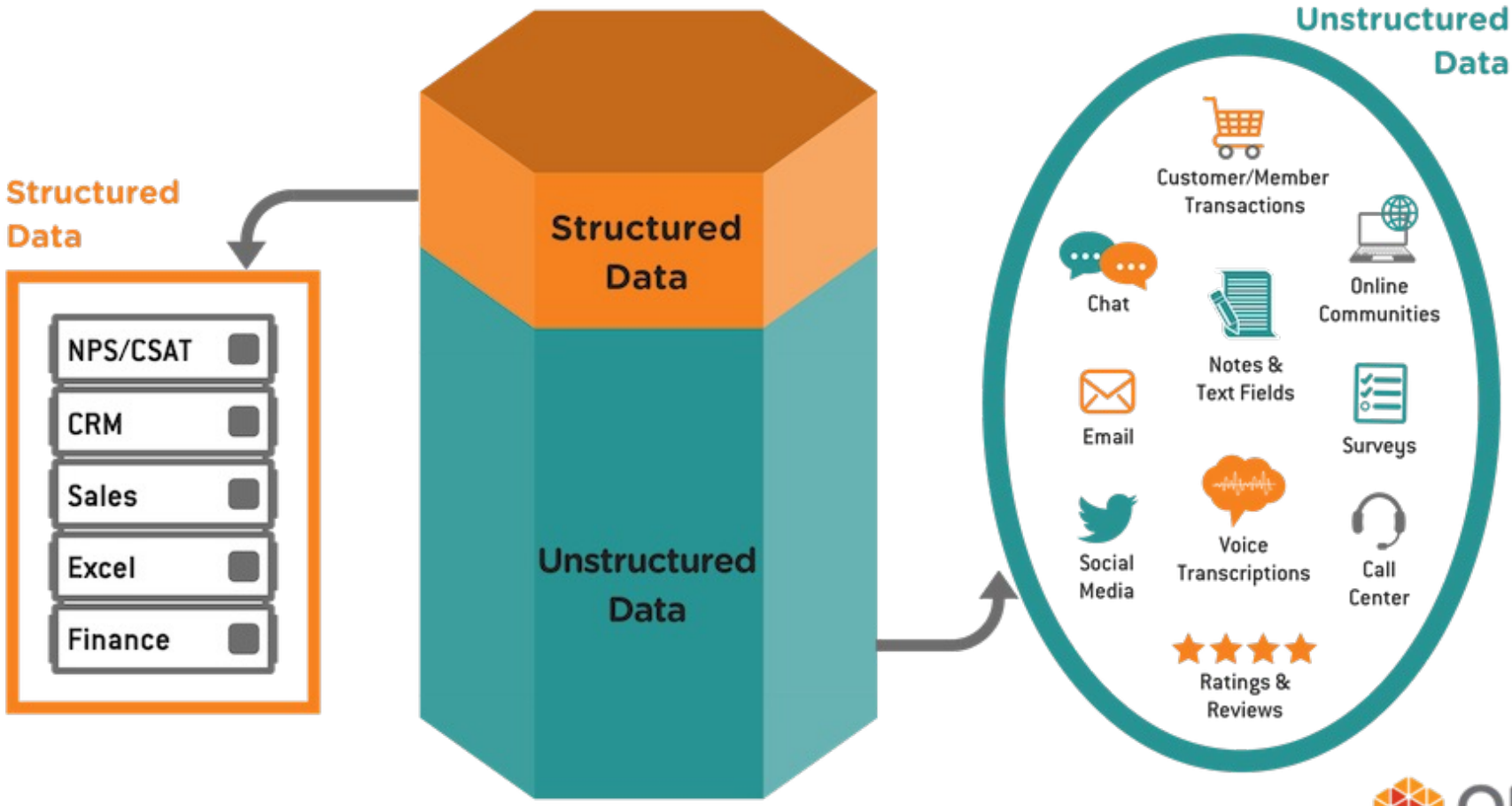
Unstructured Data

- Lacks predefined format
- Difficult to Collect, Process and Analyze.
- Examples
  - Text Documents
  - Social Media Posts, Videos, Audio, Images
  - Mobile activity
  - Websites content

Data in Real World



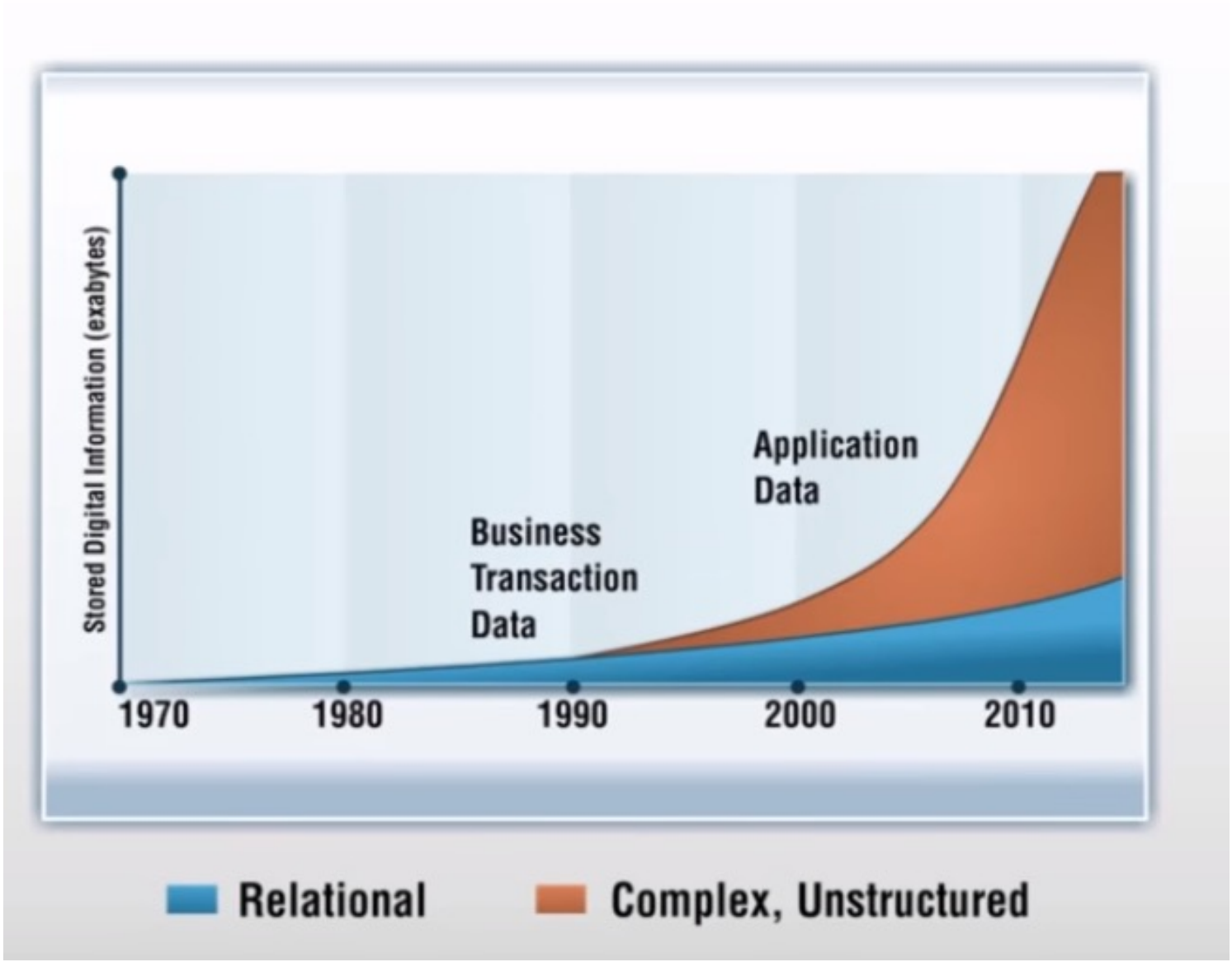
What's Hiding in Your Unstructured Data?



Source: Graphic adapted from January 2018 CXPA Presentation "The Why Behind the What," Jim Kitterman

ORI
Innovative Insights. Driving Results.

Type of Data..


Stored Digital Information (exabytes)
Application Data
Business Transaction Data
1970
1980
1990
2000
2010
Relational
Complex, Unstructured

Big Data Sources

- Social Media
- Machine Generated Data
- Website data
- Activity logs
- Sensors Data
- Transactional Data
- Data made available by Govt. – Research data
- Multimedia
- Telecom Data
- And more...

Problems in Big Data

Why RDBMS Fails?

- Unstructured data not supported
- Must have a schema
- Usage of Joins makes it slow
- Can not handle data coming at high speed – Velocity
- Massive Data - Volume
- Scalability

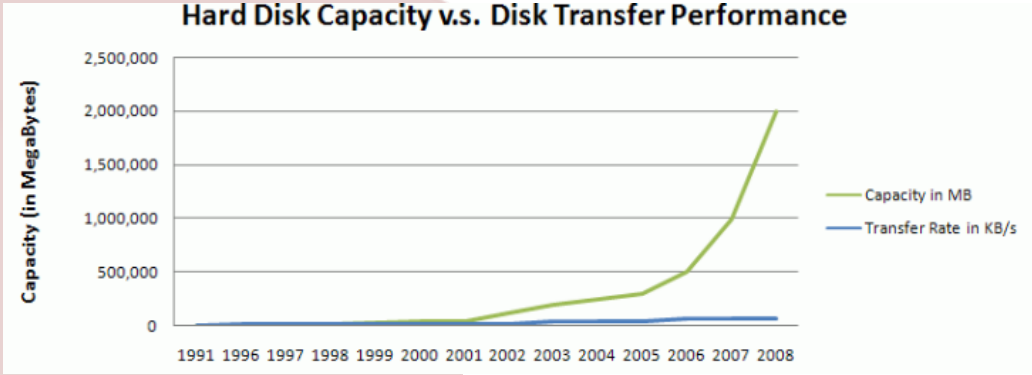
Problems in Big Data

STORAGE

COMPLEXITY OF DATA

PROCESSING SPEED

Problems in Big Data



Storage of huge datasets

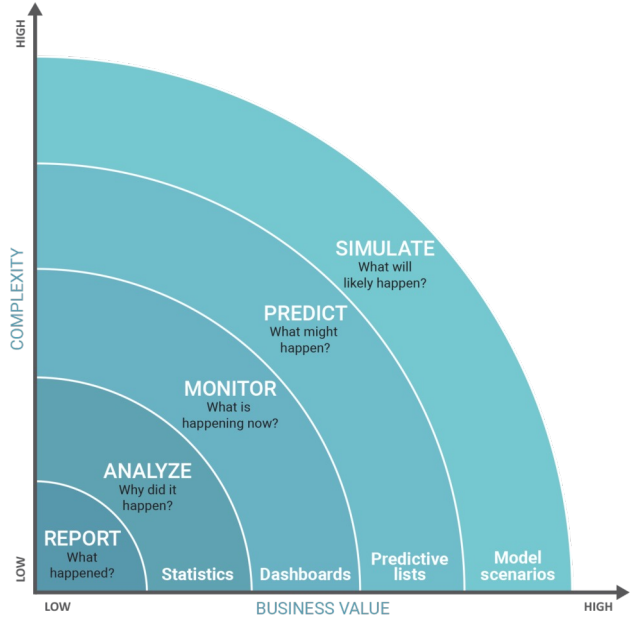
Complex Data structure

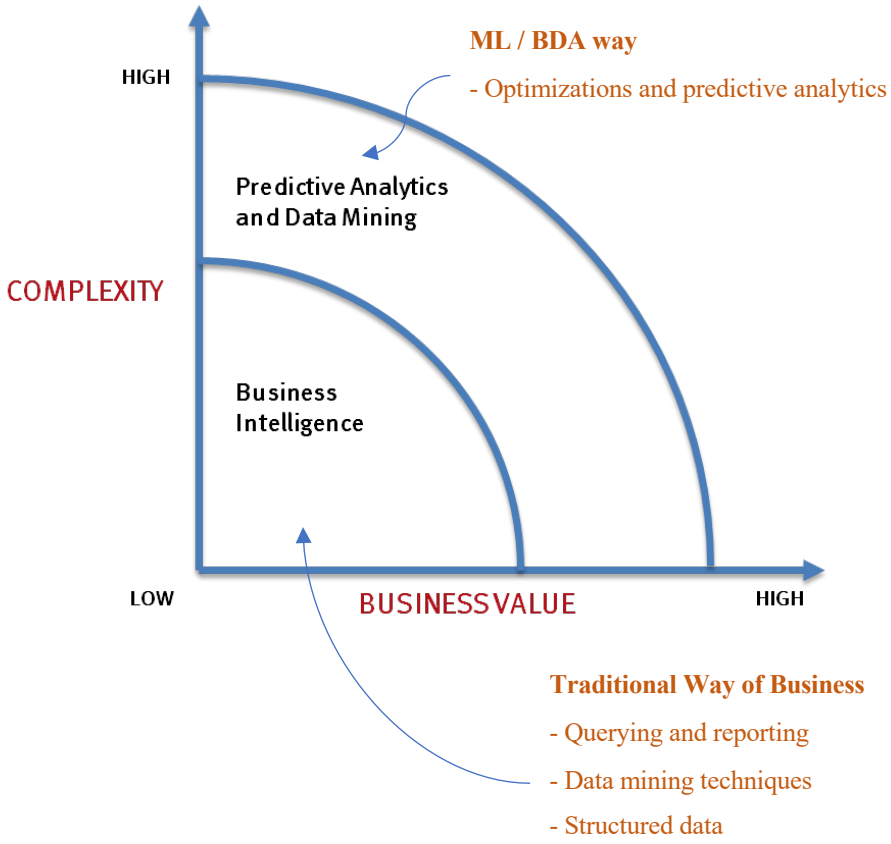
Faster processing of Data

Solution?

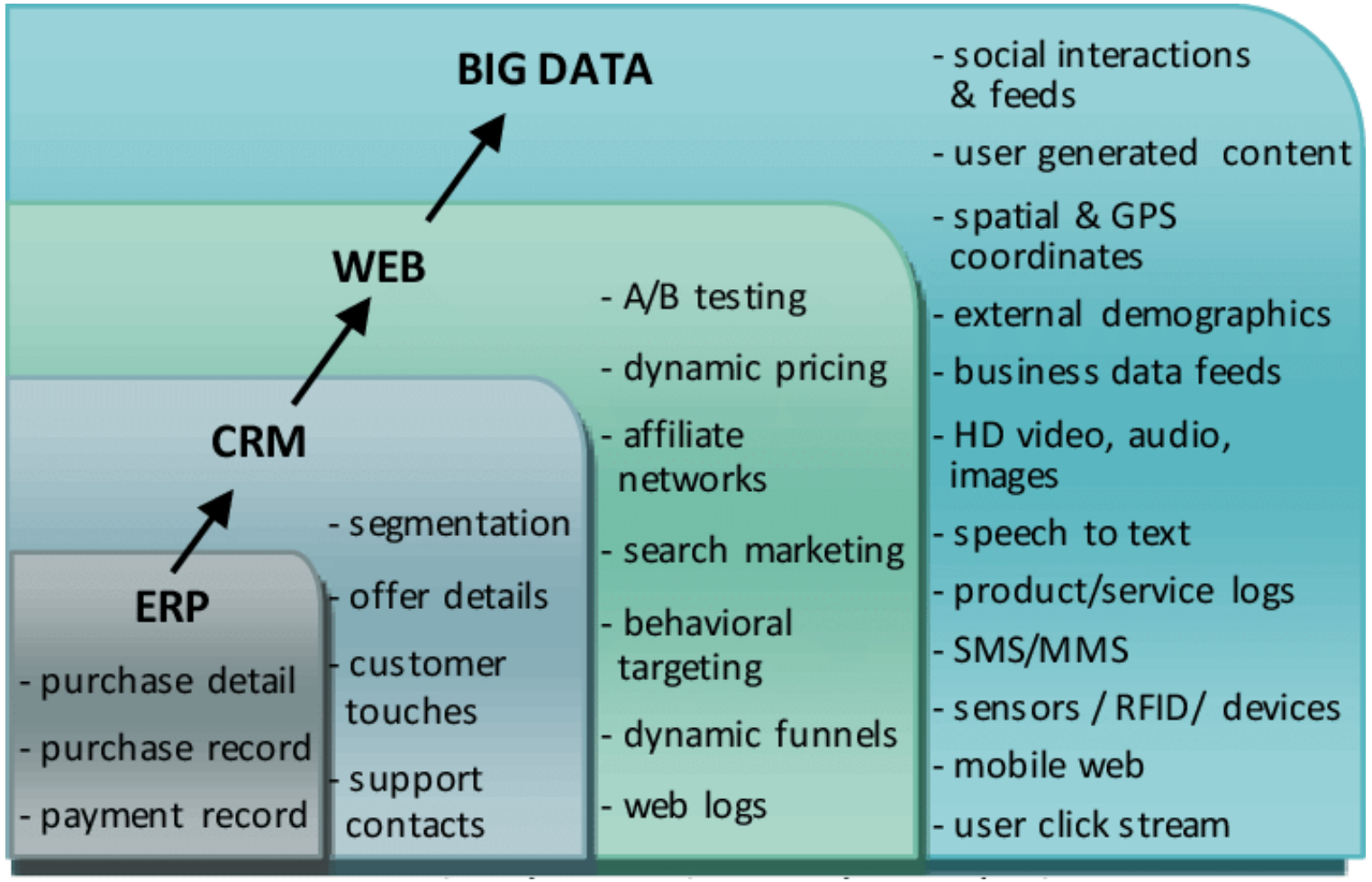
What’s Driving BDA?

Traditional vs Big Data Approach

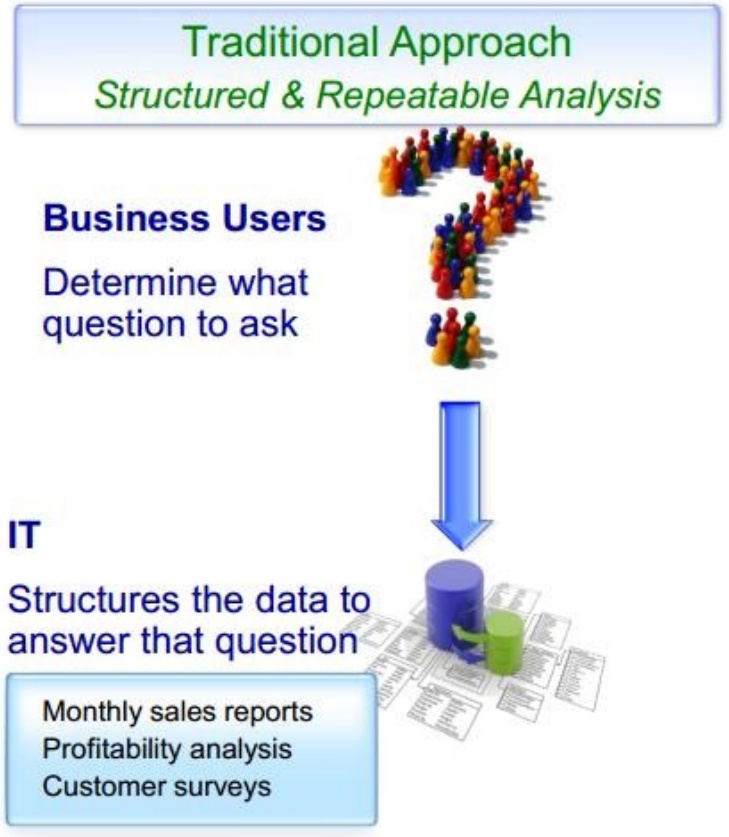


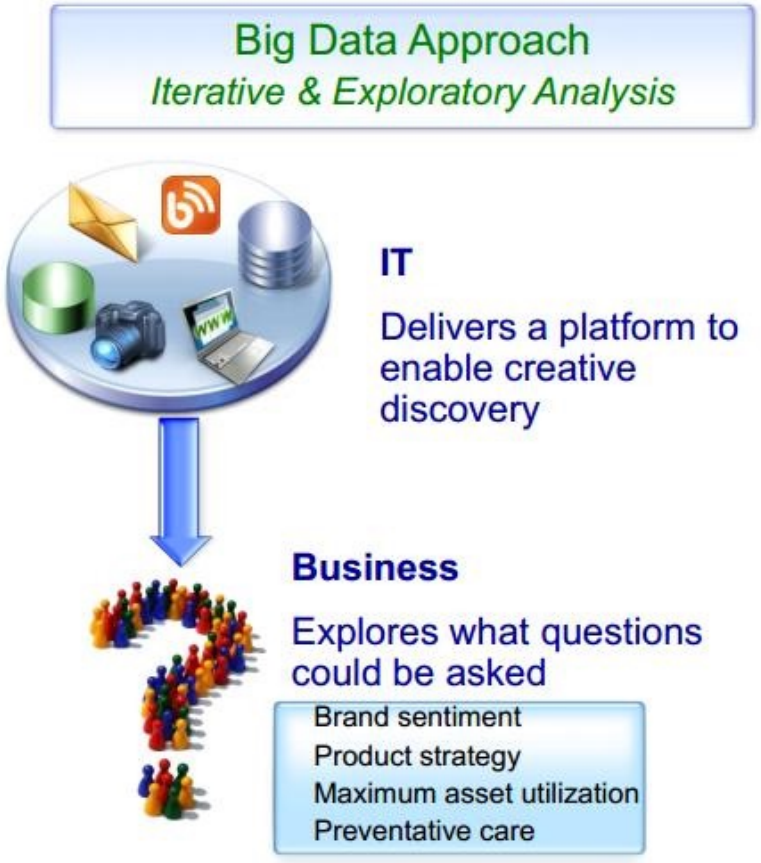


Traditional vs Big Data Approach



Traditional v/s Big Data Approach





Business Users Asking what to do.
IT delivering it.

IT delivering a platform
Business Users exploring the possibilities

Advantages of BDA

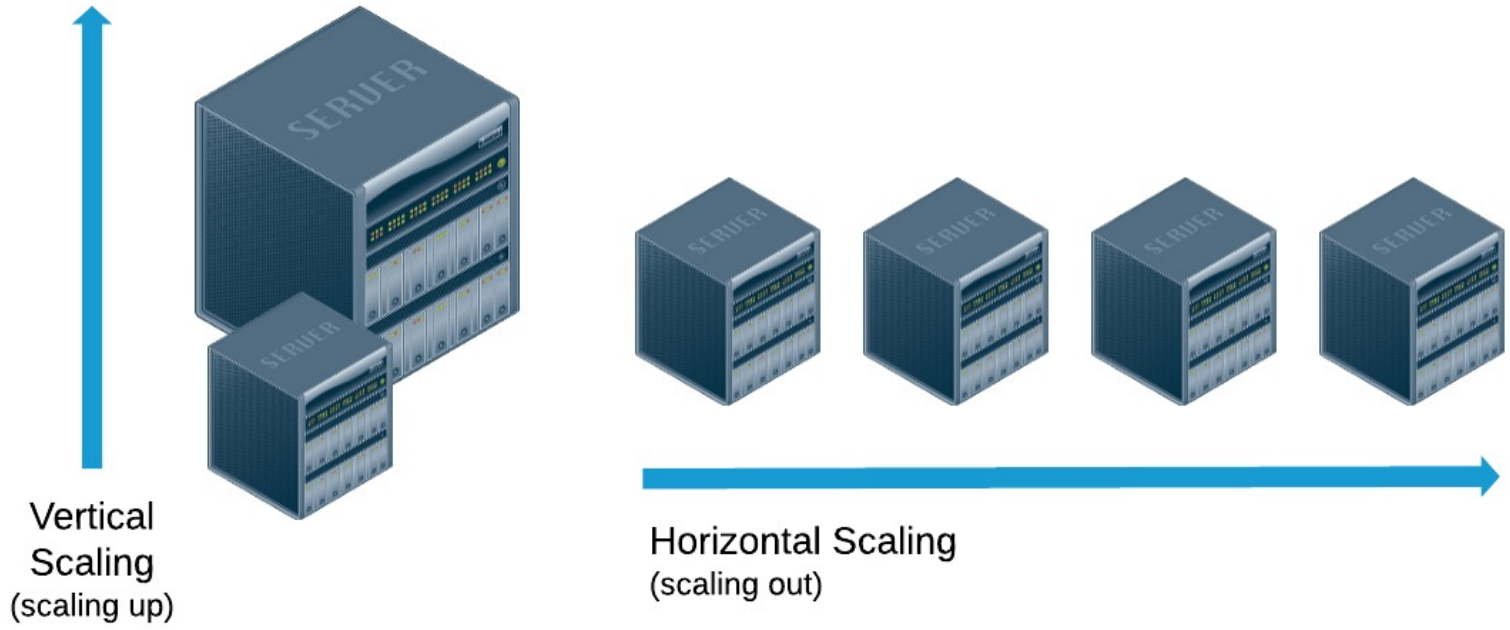
- Big Data Analysis Approach
  - Simple Model
    - Simple algorithm
    - Can be applied over large data
    - Produces more accurate analysis rather than Sophisticated models/algos.
  - Sophisticated model
    - Already built in models
    - But might not be suitable for the need
    - People improve or develop new algos

Big Data Challenges

- Data Quality
- Cost
- Data integration
- Storage and processing
- Data Security and Privacy
- Data Analysis and Interpretation
- Real time analysis
- Data Governance

Desired Properties of a Big Data EcoSystem

- Scalability
  - Storage
  - Processing
- Horizontal Scalability

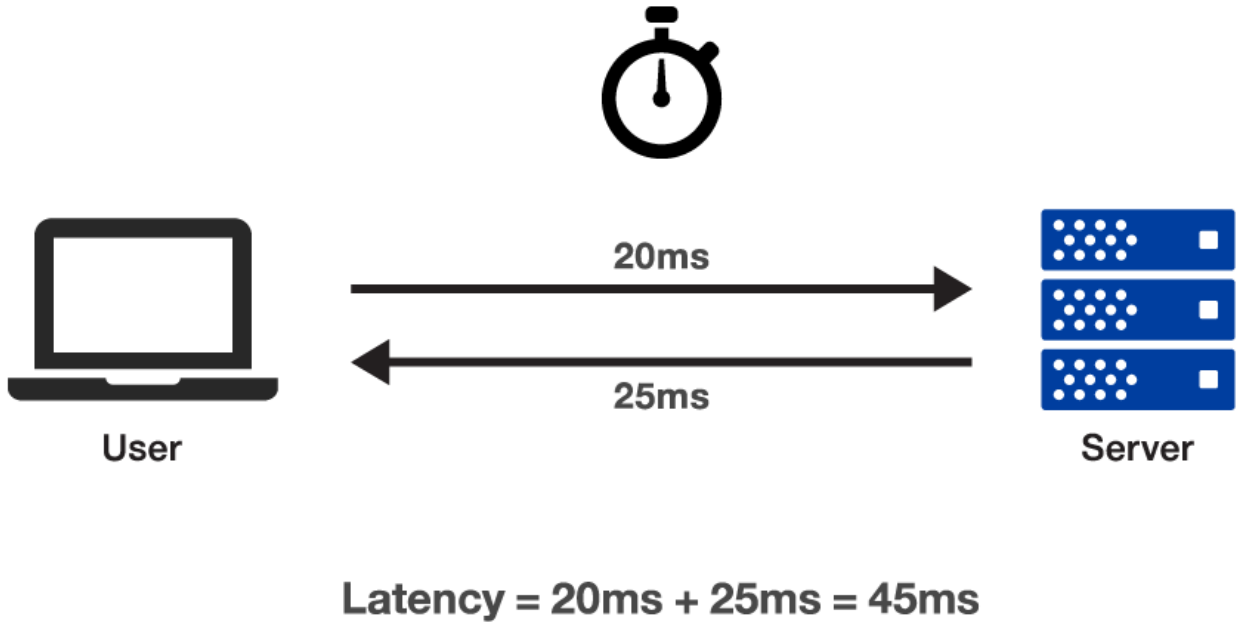


Desired Properties of a Big Data EcoSystem

- Robustness and Fault tolerance
  - Does not go down if one node goes down
  - Data remains consistent

Desired Properties of a Big Data EcoSystem

- Low Latency reads and updates



Desired Properties of a Big Data EcoSystem

- Integrity

Desired Properties of a Big Data EcoSystem

- Extensible
  - Allows new features at minimal cost/change

Desired Properties of a Big Data EcoSystem

- Cost Effective

Desired Properties of a Big Data EcoSystem

- Security – Secure system, Protects the data
- Ease of Use – User friendliness
- General – One system multiple usage or with minimal changes

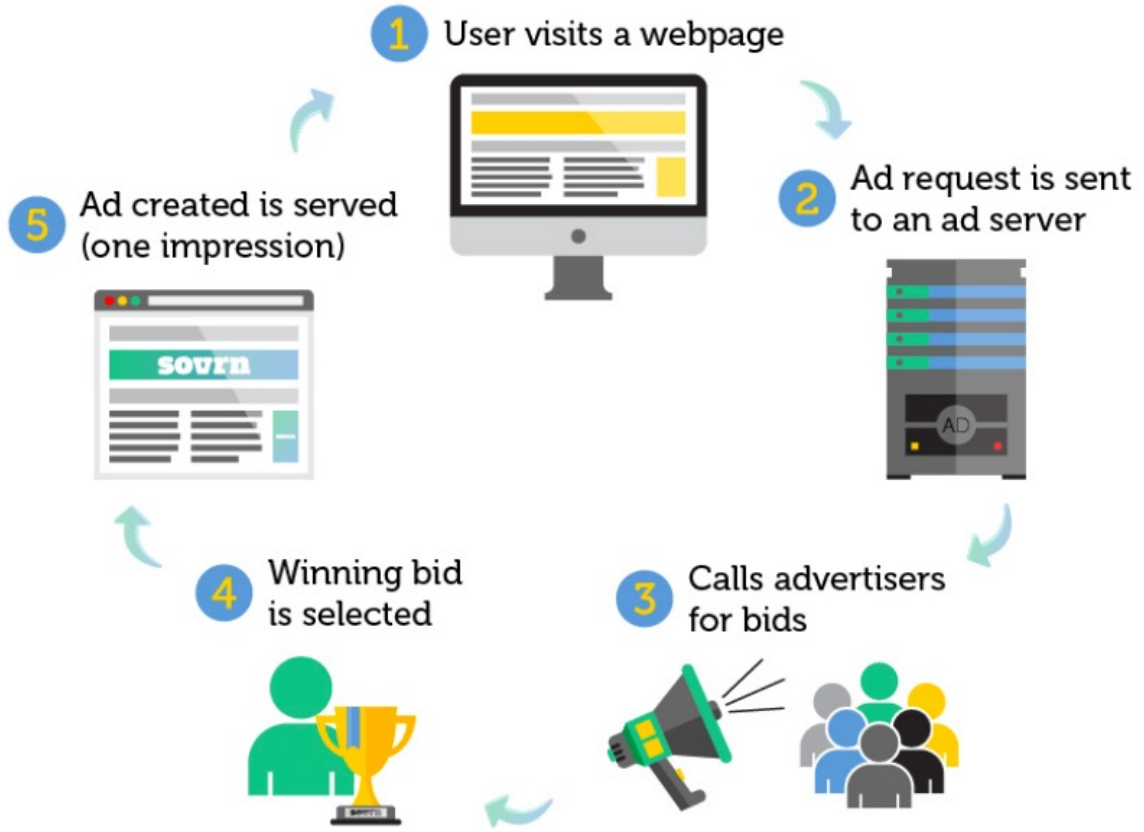
Case Studies

Big Data Applications

- Some of those…

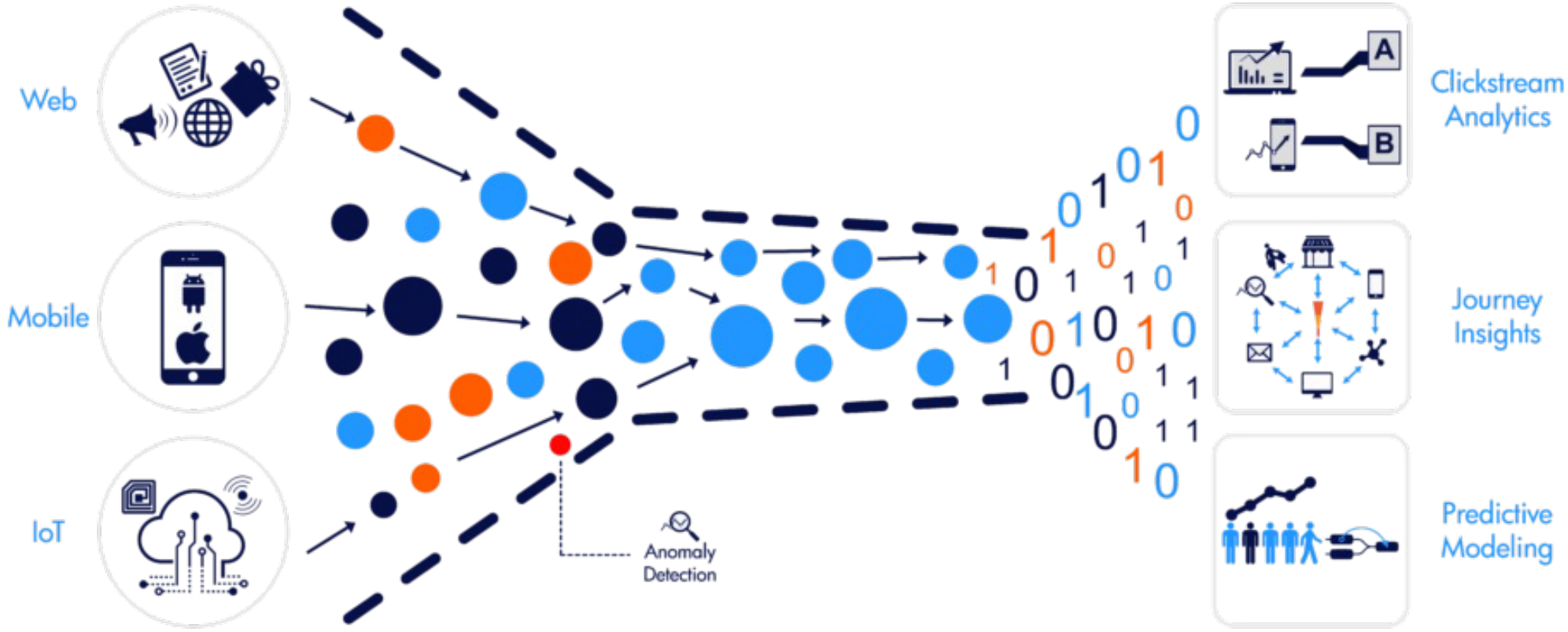
Case Study - Clickstream

- Clickstream Data
  - Information collected about a user
  - While they visit a website
  - In what order the pages are visited
  - How much time is spent?
  - Frequency – returning user?
  - Exit page? Or a Feature?
- Clickstream analysis
  - Process of collecting and analyzing the data
  - User behaviour analysis
- Used by
  - Search engines for ads
  - Other websites for ads
  - Personalised ads



Source: sovrn.com

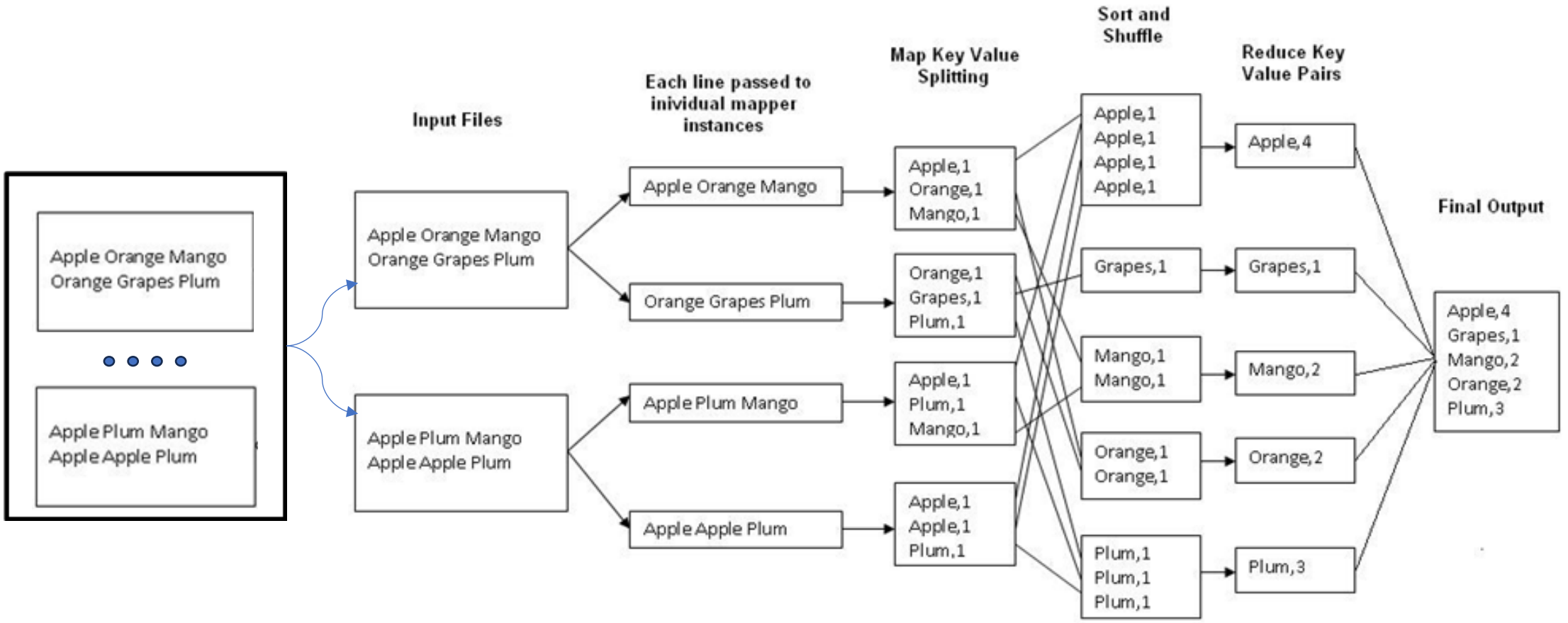
Case Study - Clickstream



Why would RDBMS not work here?

- Huge data
  - Billions of clicks each month for an e-commerce website
  - Historical data needed for predictions
- Pages are generated real-time, dynamically
  - Based on user clicks, search queries etc
  - Requires too much data to be processed at run-time, not possible with RDBMS.
  - Data may be coming from multiple data sources
    - Clickstream data from xyz service
    - Customer data from local e-commerce database
    - Other types of data..

Case Study – Word Count



Case Study – Sentiment Analysis

- For a given text
  - Guess the mood
- Applications?

POSITIVE

"Great service for an affordable price.
We will definitely be booking again."

NEUTRAL

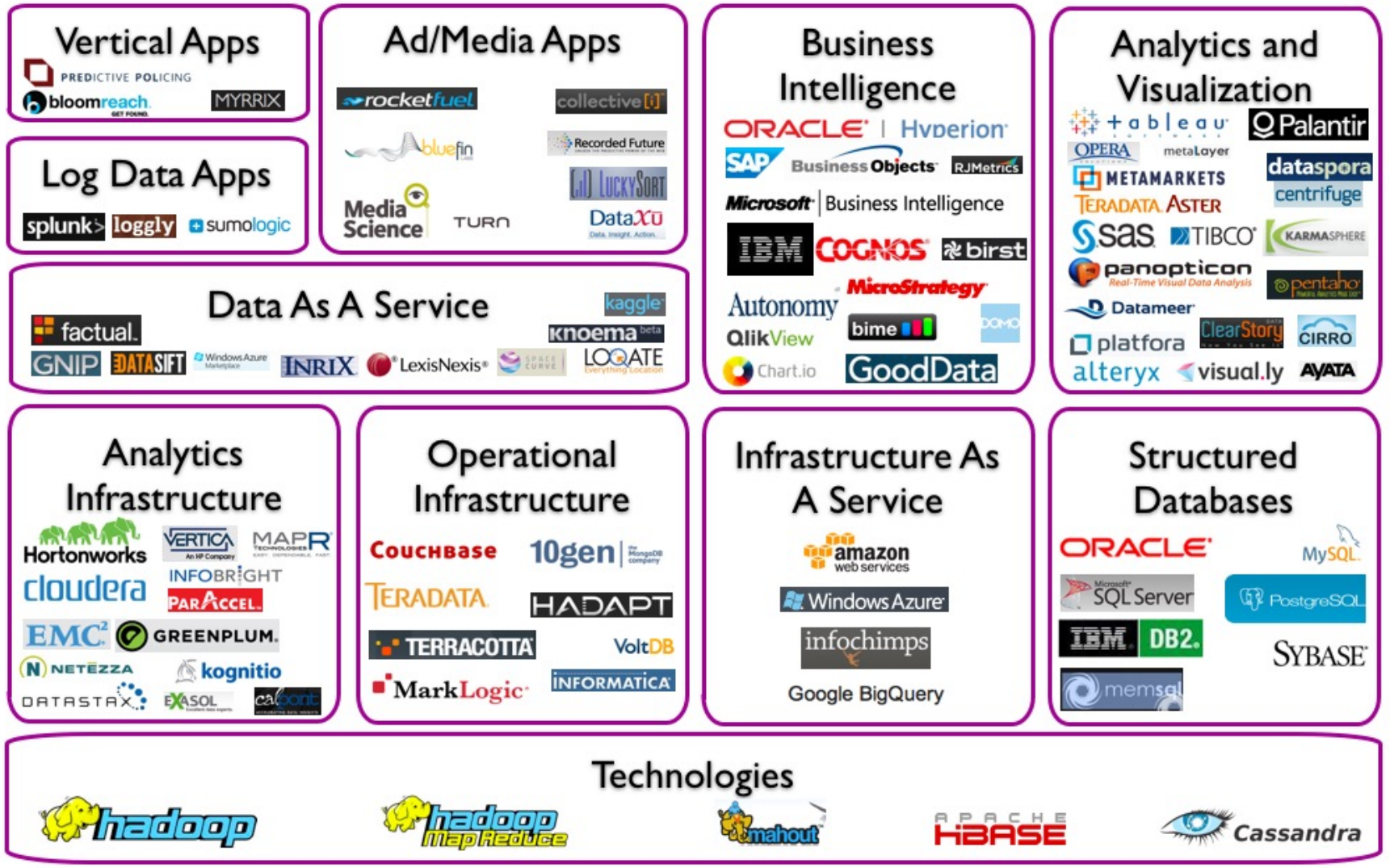
"Just booked two nights at this hotel."

NEGATIVE

"Horrible services. The room was dirty and unpleasant.
Not worth the money."

Case Studies – Sentiment Analysis

Big Data Technolgies



dave@vcdave.com

blogs.forbes.com/davefeinleib

Big Data Technolgies