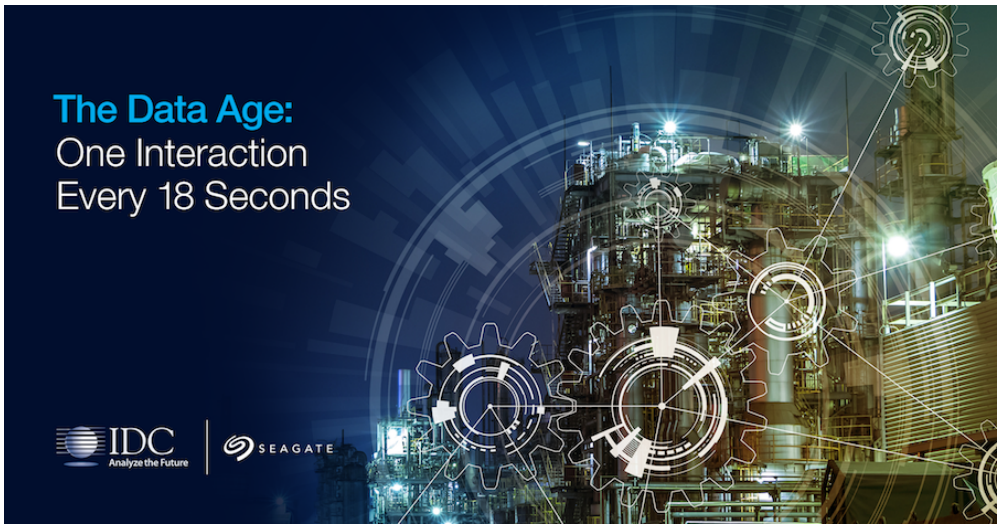
# Large Scale Data Processing CSE3025

Prof. Ramesh Ragala

July 18, 2019

- We are fast approaching a new era of **the Data age**
- From autonomous cars to humanoid robots and from intelligent personal assistants to smart home devices, the world around us is undergoing a fundamental change, transforming the way we live, work, and play.




- The average connected person anywhere in the world will interact with connected devices nearly **4,800 times per day** $\rightarrow$ basically one interaction every 18 seconds ...

43.9 Million
Wikipedia Articles

facebook

1.94 Billion Monthly users
1.28 Billion Active Users/day

flickr

3.5 Millions new images every day
1 million photos sharing every day

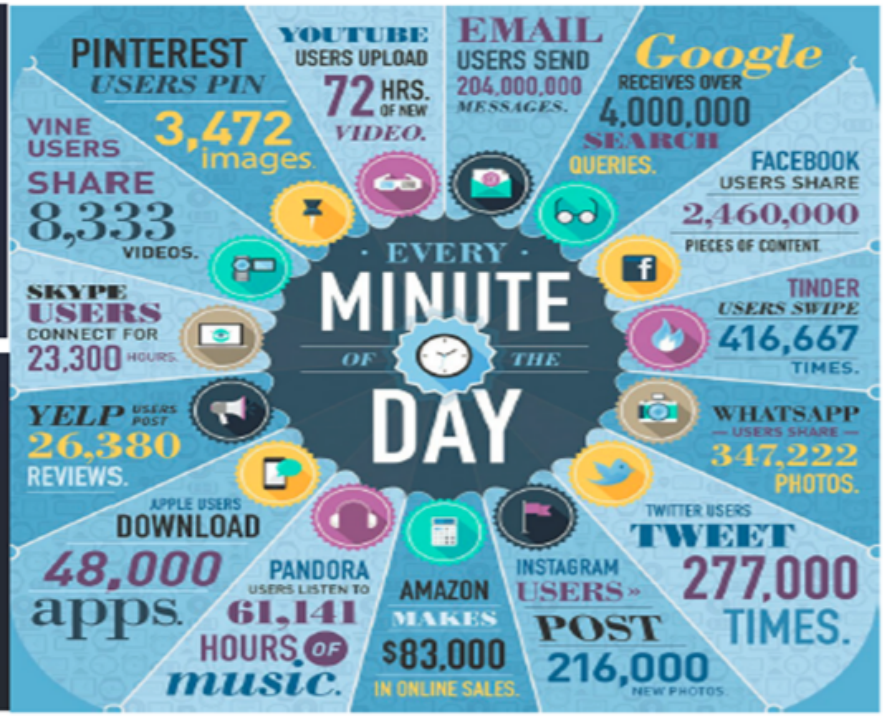
YouTube

1 billion user
300 hours of videos per minute

## Data grows fast!



## MORE IPHONES ARE SOLD THAN BABIES BORN



## The Model of Generating/Consuming Data has Changed

**Old Model:** Few companies are generating data, all others are consuming data




**New Model:** all of us are generating data, and all of us are consuming data



## What's Driving Data Deluge?

- Don't Focus on Big Data; Focus on the Data That's Big
- Data has become critical to all aspects of human life over the course of the past 30 years
- it's changed how we're educated and entertained, and it informs the way we experience people, business, and the wider world around us.
- It is the **lifeblood** of our rapidly growing digital existence
- This digital existence, as defined by the sum of all data **created**, **captured** and **replicated** on our planet in any given year is growing rapidly, and we call it the **global datasphere**
- we are as consumers of data and enjoying the benefits of a digital existence. $\rightarrow$ unique business opportunities are limitless.
- It's not easy to measure the total volume of data stored electronically, but an IDC estimate put the size of the datasphere at 16.1ZB in 2016 and is forecasting a tenfold growth by 2025 to 163ZB.

- Data Age 2025 describes five key trends that will intensify the role of data in changing our world:
  - The evolution of data from business background to life-critical.
    - Once siloed, remote, inaccessible, and mostly underutilized, data has become essential to our society and our individual lives.
    - In fact, IDC estimates that by 2025, nearly 20% of the data in the global datasphere will be critical to our daily lives and nearly 10% of that will be hypercritical.
  - Embedded systems and the Internet of Things (IoT)
    - Standalone Analog devices give way to connected digital devices $\rightarrow$ generate vast amount of data, chances to refine and improve our system
    - Big Data and metadata will eventually touch nearly every aspect of our lives
  - Mobile and real-time data.
    - Digital Transformation
    - By 2025, more than a quarter of data created in the global datasphere will be real time in nature, and real-time IoT data will make up more than 95% of this.

- Data Age 2025 describes five key trends that will intensify the role of data in changing our world:
  - Cognitive/artificial intelligence (AI) systems that change the landscape.
    - The flood of data enables a new set of technologies such as machine learning, natural language processing, and artificial intelligence → Cognitive Systems → to turn data analysis from an uncommon and retrospective practice into a proactive driver of strategic decision and action
    - the amount of analyzed data that is "touched" by cognitive systems will grow by a factor of 100 to 1.4ZB in 2025.
  - Security as a critical foundation
    - All this data from new sources open up new vulnerabilities to private and sensitive information.
    - By 2025, almost 90% of all data created in the global datasphere will require some level of security, but less than half will be secured.
- IDC estimates that in 2025, the world will create and replicate 163ZB of data, representing a tenfold increase from the amount of data created in 2016

- **Data From Business Background to Life-Critical**
- According to IDC, the data creation and use of compute data broadly classified into three eras
  - **I$^{st}$ Platform**: Before 1980
    - Data resided almost exclusively in Data Centers before 1980.
    - Access the data through remote terminals.
    - The data and processing ability remained centralized in mainframes.
    - The purpose of data generation and use was almost entirely business focused.
  - **II$^{nd}$ Platform** : 1980 - 2000
    - Rise of Personal Computers and Moore's law enabled democratic distribution of data and computing power.
    - Datacenters evolved from mere **data containers** to become **centralized hubs** that managed and distributed data across a network to end devices
    - These devices gained the ability to store and manage data for purely personal use by consumers.
  - **IIIrd Platform**: 2000 - today
    - The proliferation of wireless broadband and fast networks encouraged data's movement into the cloud, decoupling data from specific physical

## Before 1980

- Data sits almost exclusively in datacenters
- Data and compute centralized
- Business-focused

## 1980—2000

- Data and compute are distributed
- Datacenters expand role in managing data
- Quick expansion in entertainment

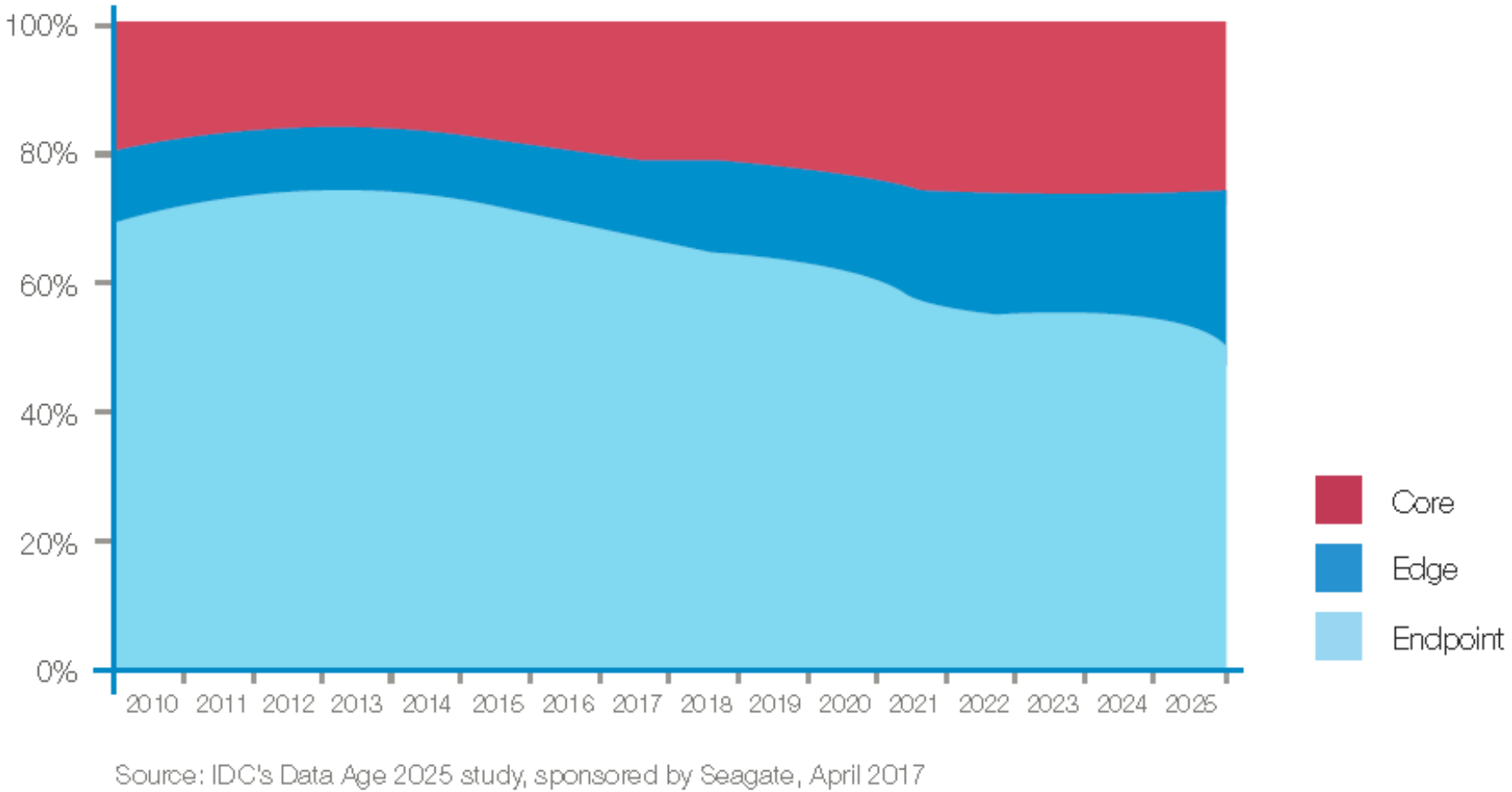
## 2000 to Today

- Datacenters expand to cloud infrastructures
- Compute continues to be distributed; data begins to contract
- Add social to the mix

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

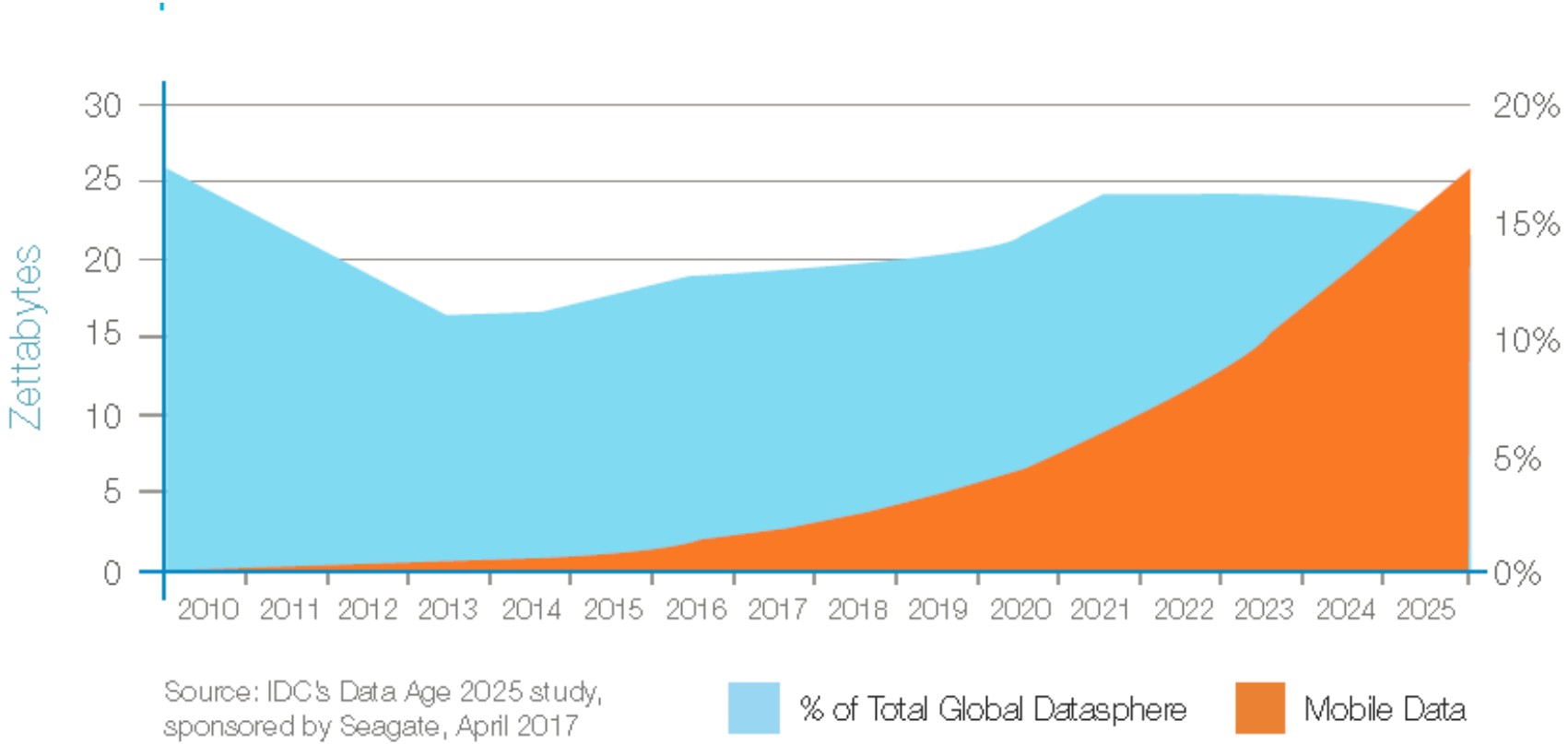
- Data evolutionary role in the world become rapidly apparent in the utilization of data by different computing platform over time
- These locations are classified into three categories:
  - **Core:** It refers a designated computing data center and cloud
  - Example: Public, Private and Hybrid Clouds, Operational control center of electric grid or telephone.
  - **Edge:** It refers the enterprise-hardened computer or appliances.
  - These are not in core data centers
  - Example: Server Rooms, Servers in Fields, Regional small data centers
  - **Endpoint:** It refers the all devices on the edge of network
  - Example: PCs, Phones, Cameras(Security), Autonomous Cars, Wearable Devices and Sensors
- Endpoints are given more contribution in the percentage of total data creation from 2013 onward.

## Data Creation



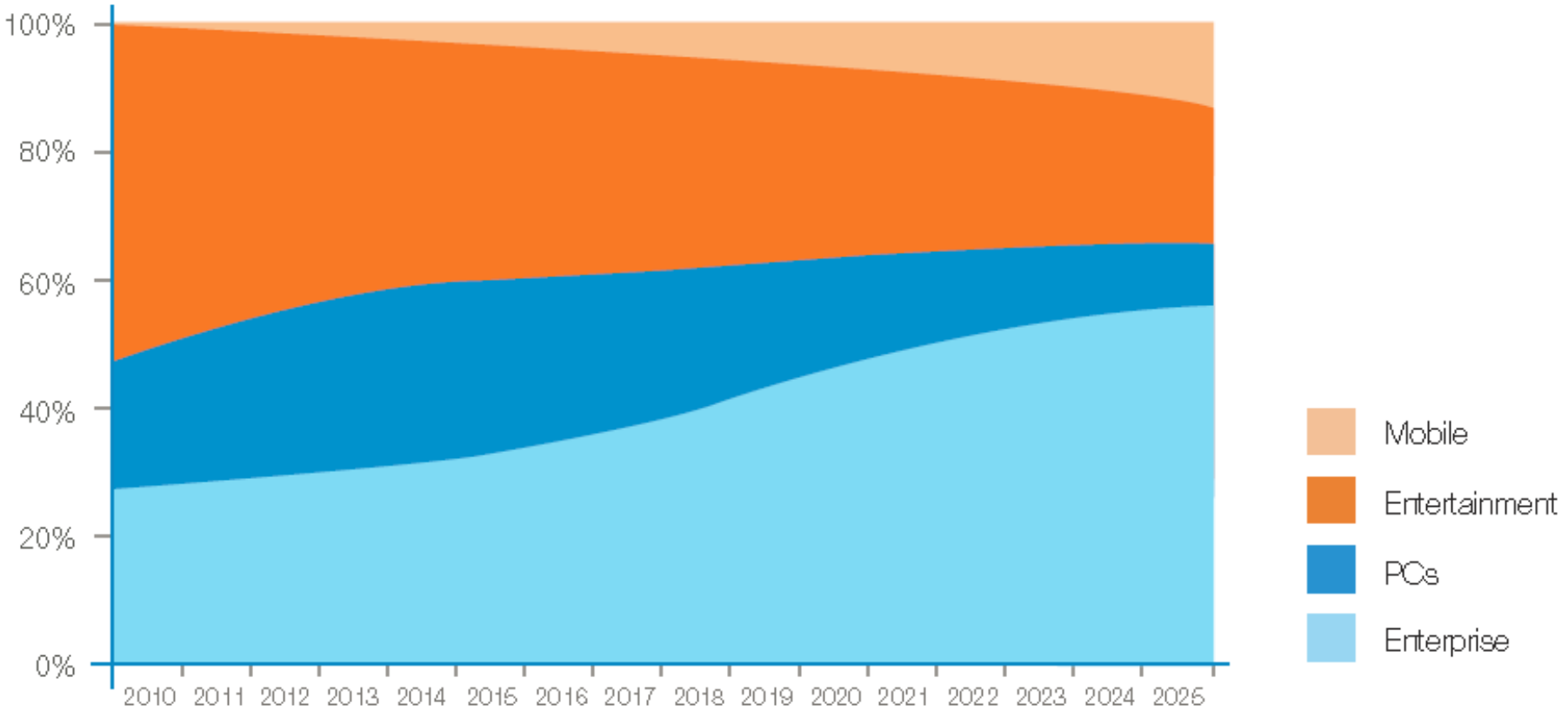

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

## Mobile Data


Zettabytes
30
25
20
15
10
5
0
20%
15%
10%
5%
0%
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017
% of Total Global Datasphere
Mobile Data

- Rapid change landscape in data storage platforms
- From 1980 to the early 2000, PCs and entertainment media dominated data creation and consumption.
- Rapid growth in network and IP connectivity $\rightarrow$ Streaming services $\rightarrow$ less need to store data locally to mobile devices, PCs, etc.
- Data Storage in IIIrd Platform $\rightarrow$ Cloud Storage

## Data Storage

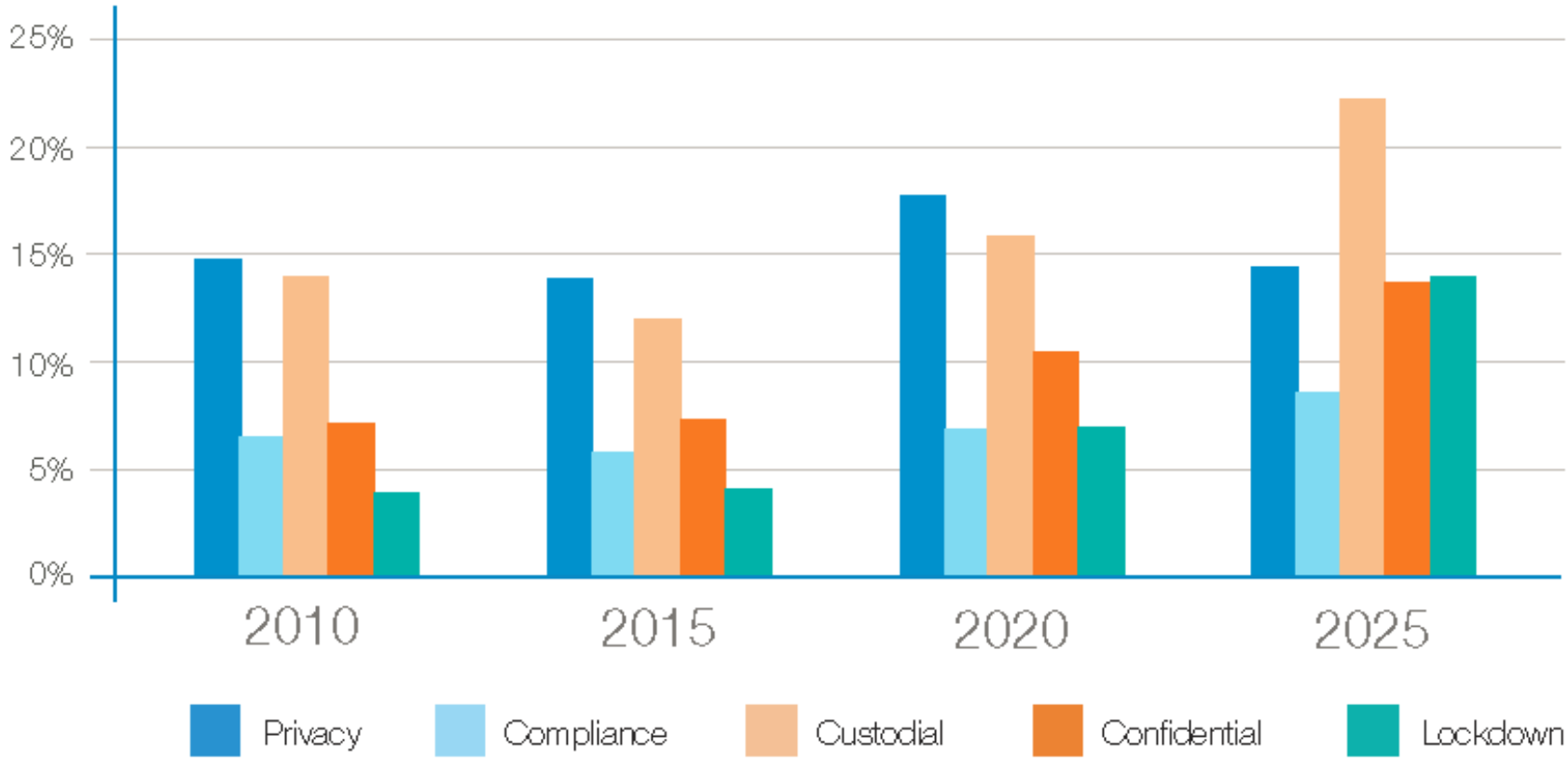



Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

- With the changes in data sources, usage and value, Security becomes crucial foundation in datasphere.
- Enterprises has more challenging and responsibility in managing privacy and security risk of personal data.
- Some data types do not carry hard security requirements today, including camera phone photos, digital video streaming, public website content, and open source data.
- However most data do, such as corporate financial data, personally identifiable information (PII), and medical records.
- 90% of data need high-end security by 2025.

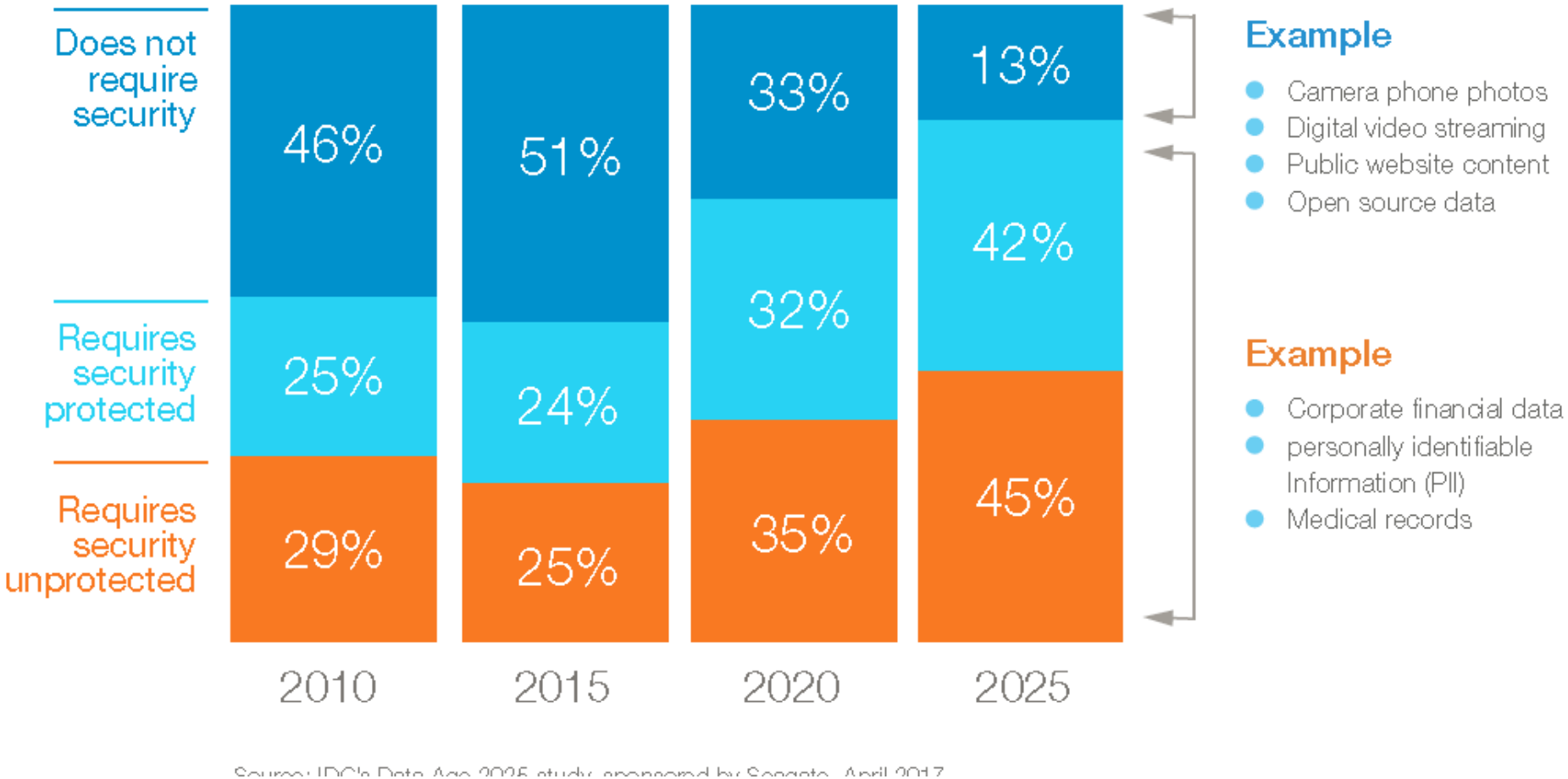
- Five different types of securities.
  - **Lockdown**: Highest Security. Ex: financial transactions, military intelligence, etc
  - **Confidential**: Information that the originator wants to protect. Ex: trade secrets, customer list, memos etc.
  - **Custodial**: Account information that, if breached, could lead to or aid in identity theft
  - **Compliance - Driven**: Information such as emails that might be discoverable in litigation or subject to a retention rule
  - **Private**: Information such as an email address on a YouTube upload

## Data Requiring Security




Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

## Actual Status of Data Security



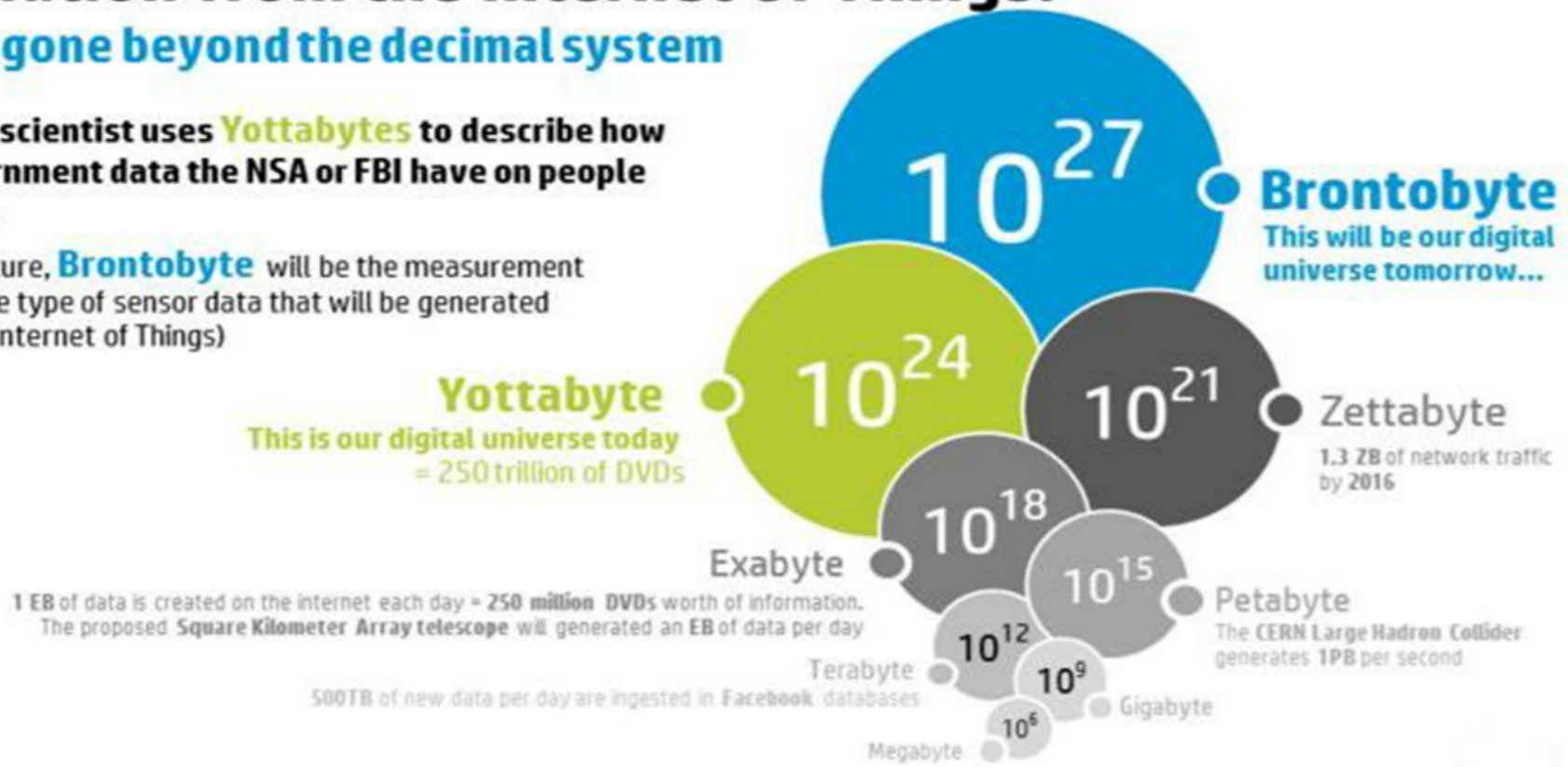

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

# Information from the Internet of Things:

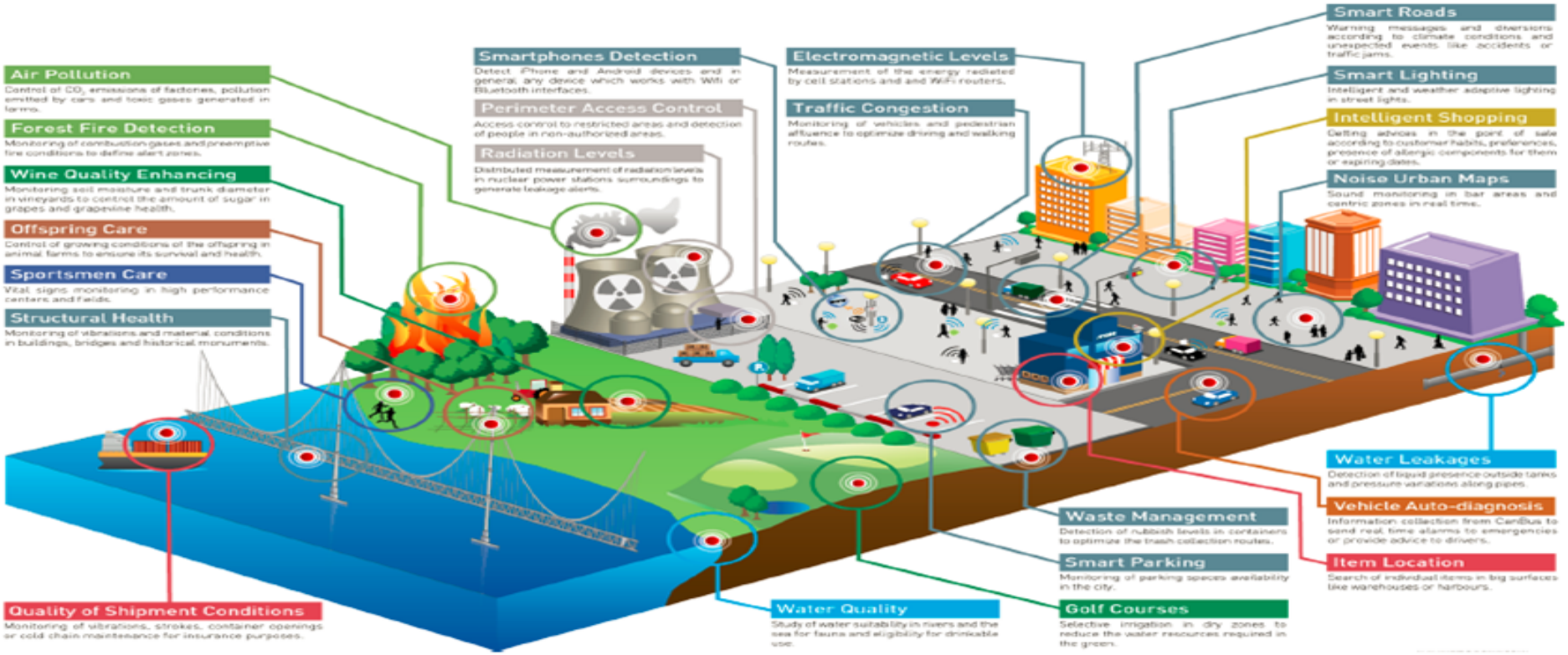
## We have gone beyond the decimal system

**Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.**

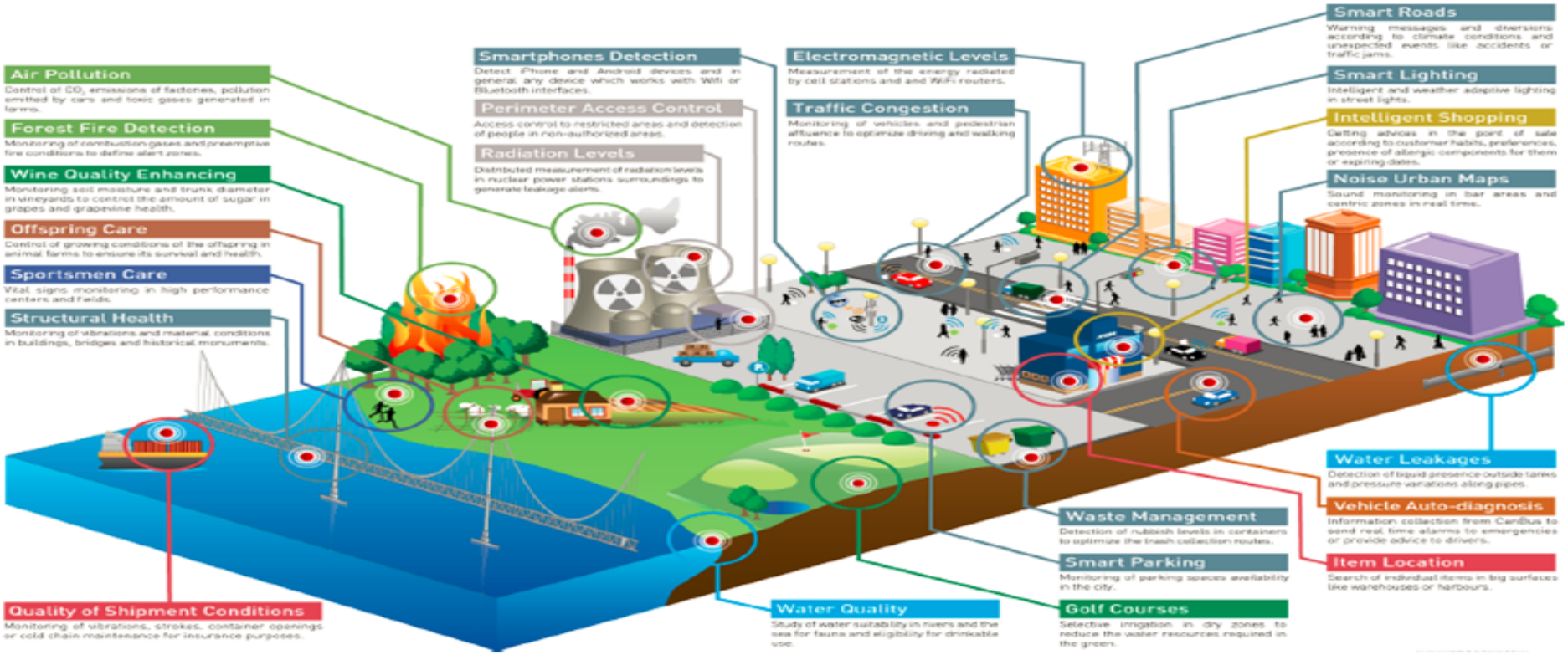
In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)



# Proliferation of Data sources

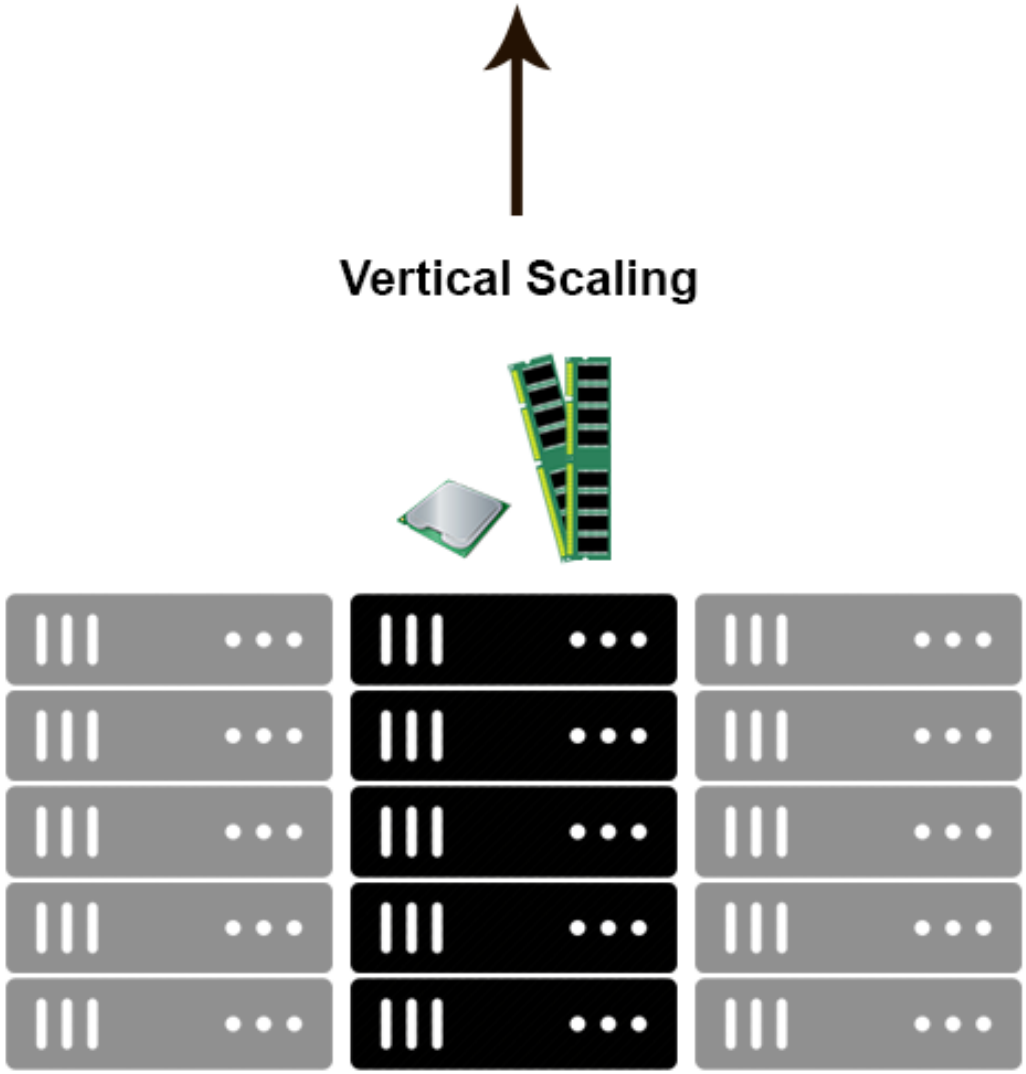


# Proliferation of Data sources



- **Vertical Scaling**
  - It means that scaling by adding more power (CPU, RAM) to an existing machine.
  - Often limited to the capacity of single machine $\rightarrow$ beyond that capacity often involves downtime and comes with an upper limit.
  - Example : MySQL
- **Horizontal Scaling**
  - It means that scaling by adding more machines into pool of resources
  - easy to scale dynamically by adding more machines into existing pool of resources
  - Example: MongoDB, etc.

- Vertical Vs Horizontal Scaling



# Parallelism

- Wide variety of different parallel architecture
  - GPUs
  - Multi-core
  - Clusters
- Design and implement parallel learning Systems???
- Low Level Parallel Primitives
  - Threads, Locks and Messages
  - It tunes for a specific platform.
  - Difficult to extends

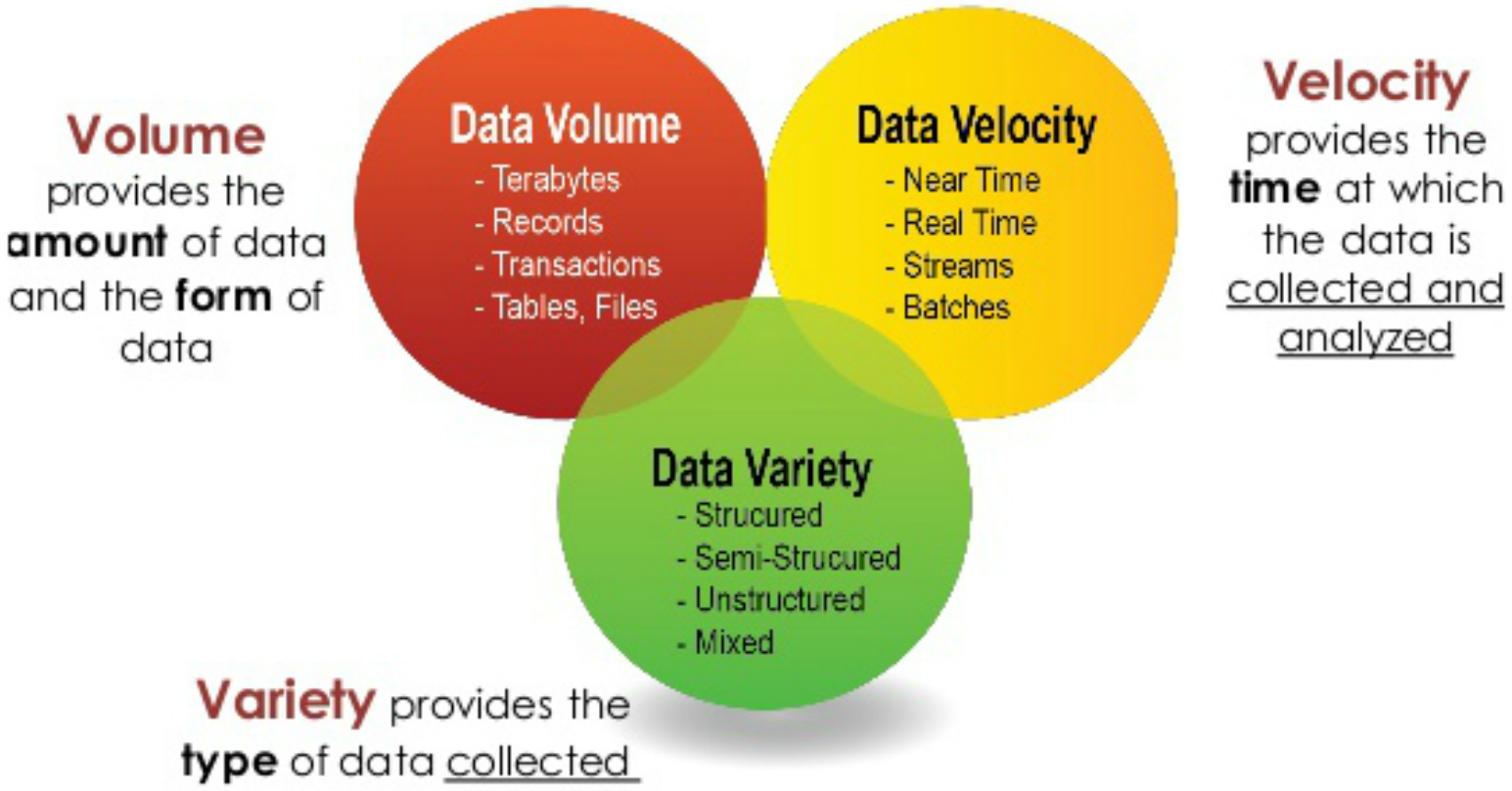
# What is Big Data

- **It's the data that is too large, complex, and dynamic such that it is impractical for any conventional hardware and/or software tools and systems to manage and process.**
- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges that we face with DBMS tools and other technologies are capture, curation, storage, search, sharing, transfer, analysis, and visualization

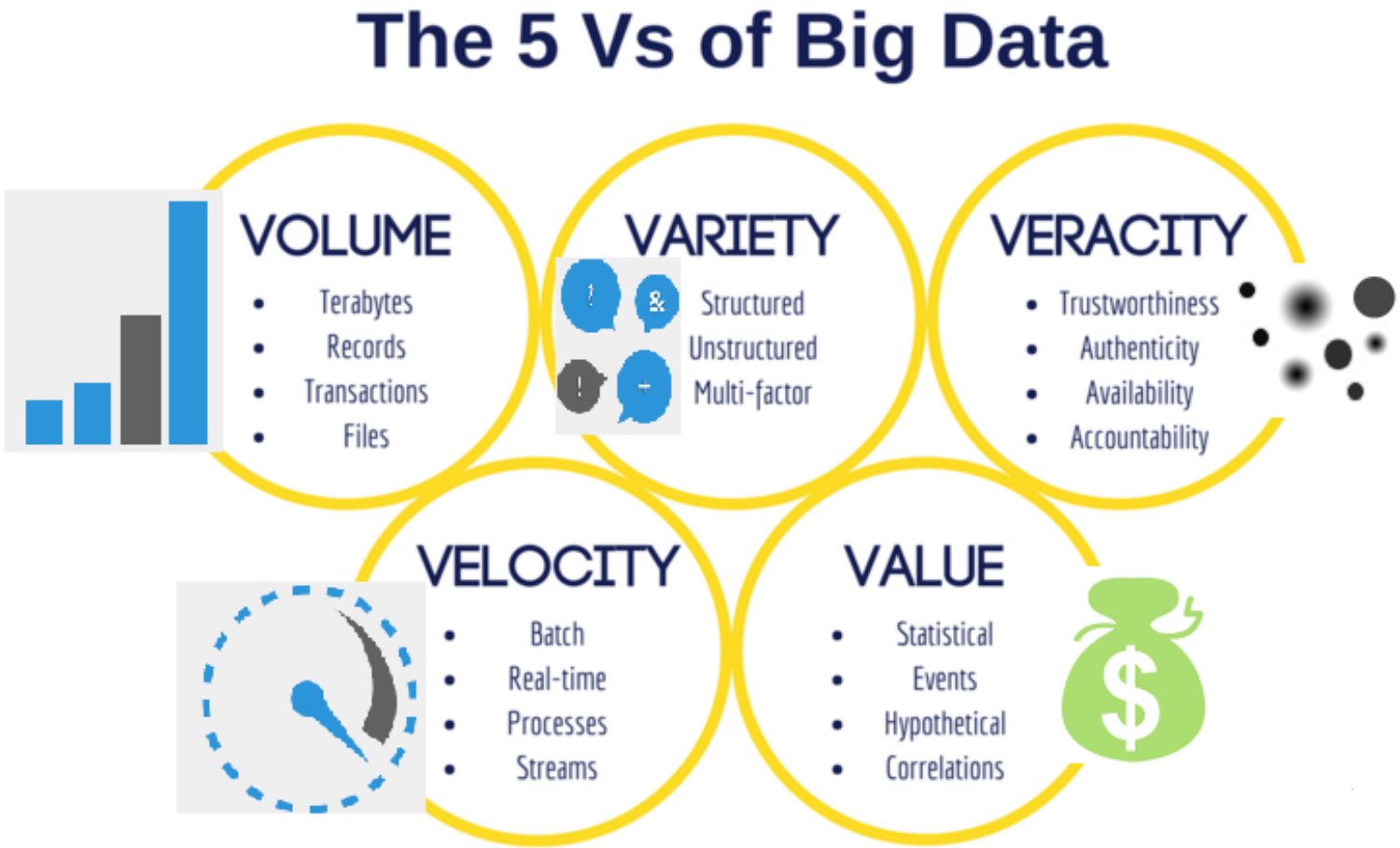
# Why is Big Data

- Key enablers for the appearance and growth of 'Big-Data' are:
  - Increase in storage capabilities
  - Increase in processing power
  - Availability of data

# Characteristics of Big Data



# Extended (3+n)V Model



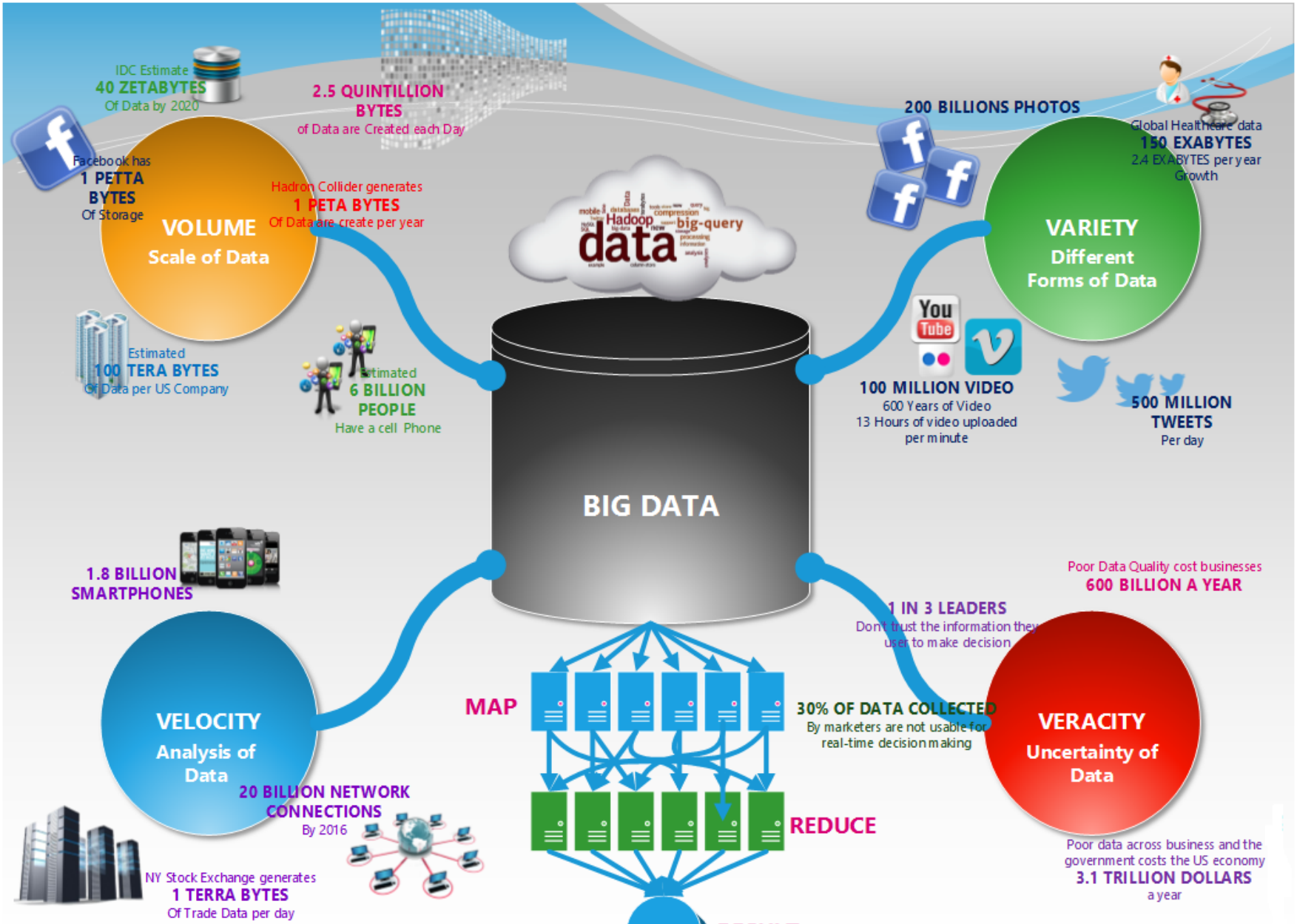
# Extended (3+n)V Model

- **Volume**
  - The amount of data generated every second.
  - Challenges to store and process ( how to index and retrieve)
  - Terabytes, Zettabytes, Brontobytes
- **Velocity**
  - Speed-issues to consider
  - How fast is the data available for analysis?
  - How fast can we do something with it
- **Variety**
  - Different kinds of data $\rightarrow$ curse of dimensionality
  - **Structured Data**
  - **Semi-structured Data**
  - **Unstructured Data**
- **Veracity**
  - Trustworthiness of data
  - With many forms of data $\rightarrow$ Quality and Accuracy are less controllable
  - Example: Twitter post with hash tags, abbreviations, typos and colloquial speech

# Extended (3+n)V Model

- **Value**
  - Big data can generate huge competitive advantages
- **Variability**
  - data flows can be highly inconsistent with periodic peaks
- **Veracity**
  - untrusted, uncleaned data, etc

# Extended (3+n)V Model



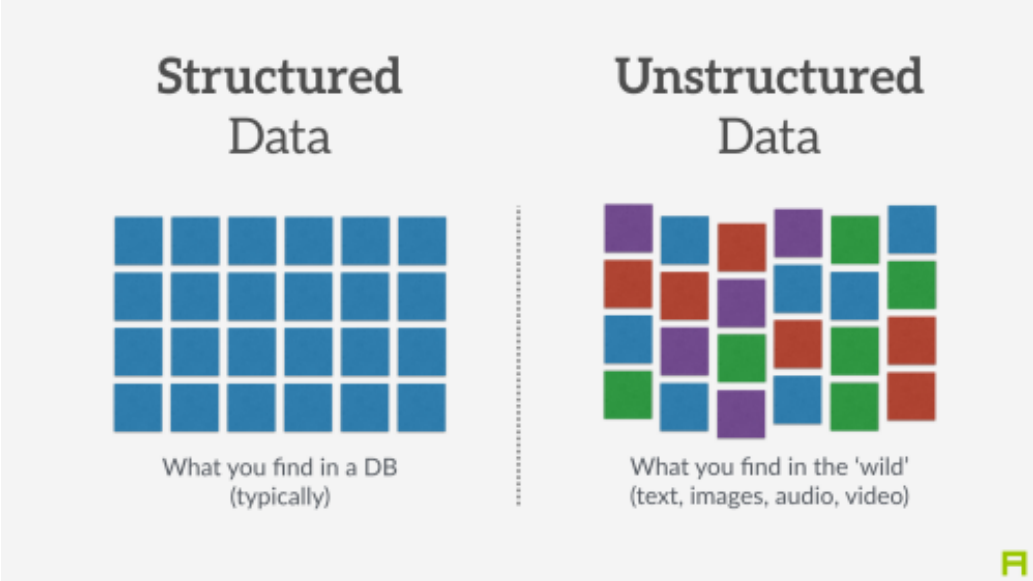
- **Structured Data**
  - Pre-defined schema imposed on the data
  - Highly structured
  - Usually stored in a relational database system
  - Structured data is relatively simple to enter, store, query, and analyze, but it must be strictly defined in terms of field name and type
- Example:
  - numbers: 20, 3.1415,. . .
  - dates: 21/03/1978
  - strings: "Hello VIT"
- **Roughly 10% of all data out there is structured**

- **Semi-Structured Data**
  - Data may have certain structure but not all information collected has identical structure
  - Inconsistent structure. it mixed with schema
  - Cannot be stored in rows and tables in a typical database.
  - Information is often self-describing (label/value pairs).
- Example:
  - XML, SGML,. . .
  - BibTeX files
  - logs, tweets, sensor feeds

# Extended (3+n)V Model

- **Unstructured Data**
  - Does not reside in traditional databases and data warehouses
  - May have an internal structure, but does not fit a relational data model
  - Lacks structure or parts of it lack structure.
- Example:
  - multimedia: videos, photos,
  - audio files,. . .
  - email messages
  - word processing documents
- **Experts estimate that 80 to 90 % of the data in any organization is unstructured**

# Extended (3+n)V Model




Structured Data


Unstructured Data