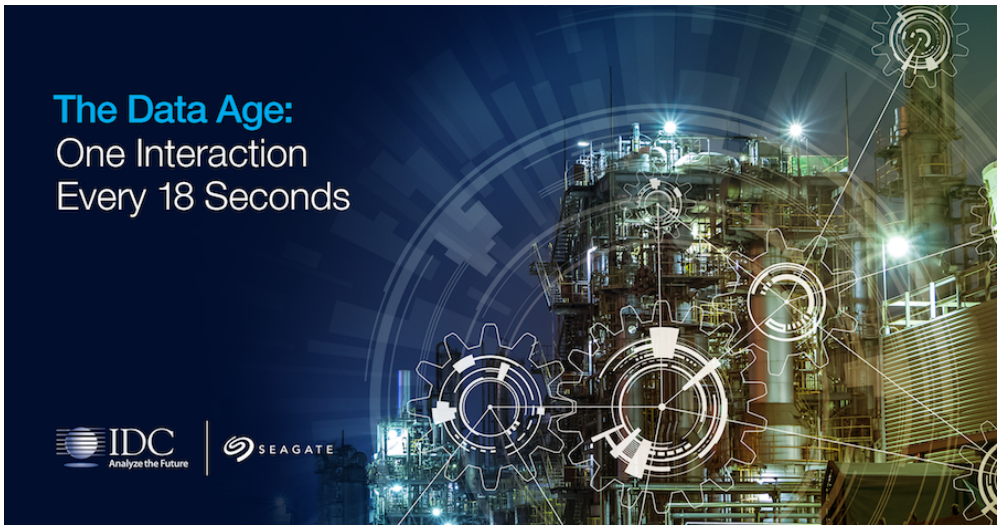
Large Scale Data Processing CSE3025

Prof. Ramesh Ragala

February 19, 2021

- Overview on Data, Information and Knowledge
  - **Data**: Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities
  - **Information:** it is the outcome of extraction and processing activities carried out on data, and it appears **meaningful** for those who receive it in a specific domain.
  - **Knowledge**: Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions.

- 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

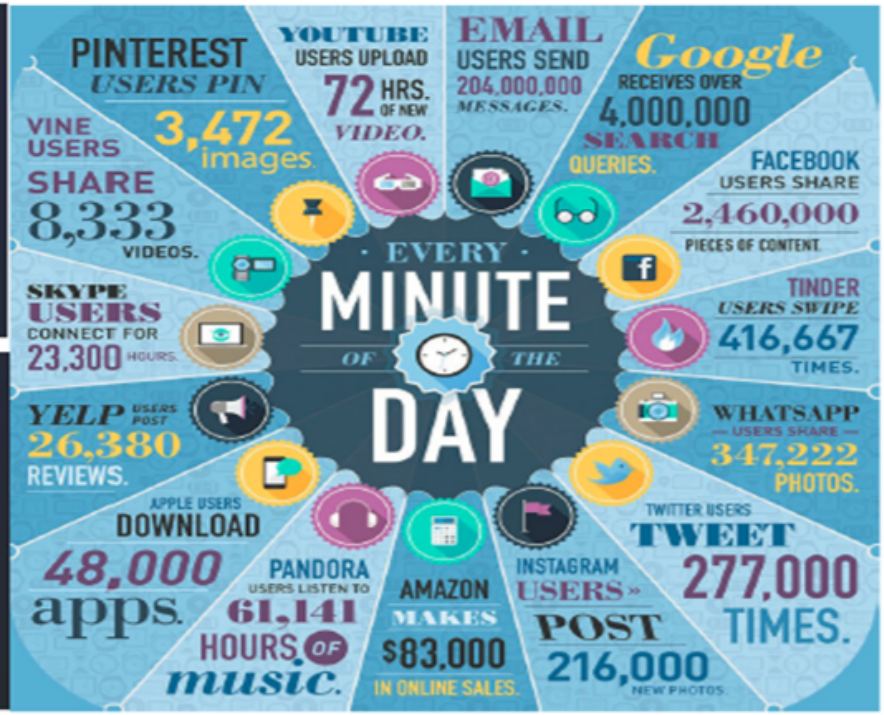
flickr

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!



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
  - **Ist 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.
  - **IInd 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 devices

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

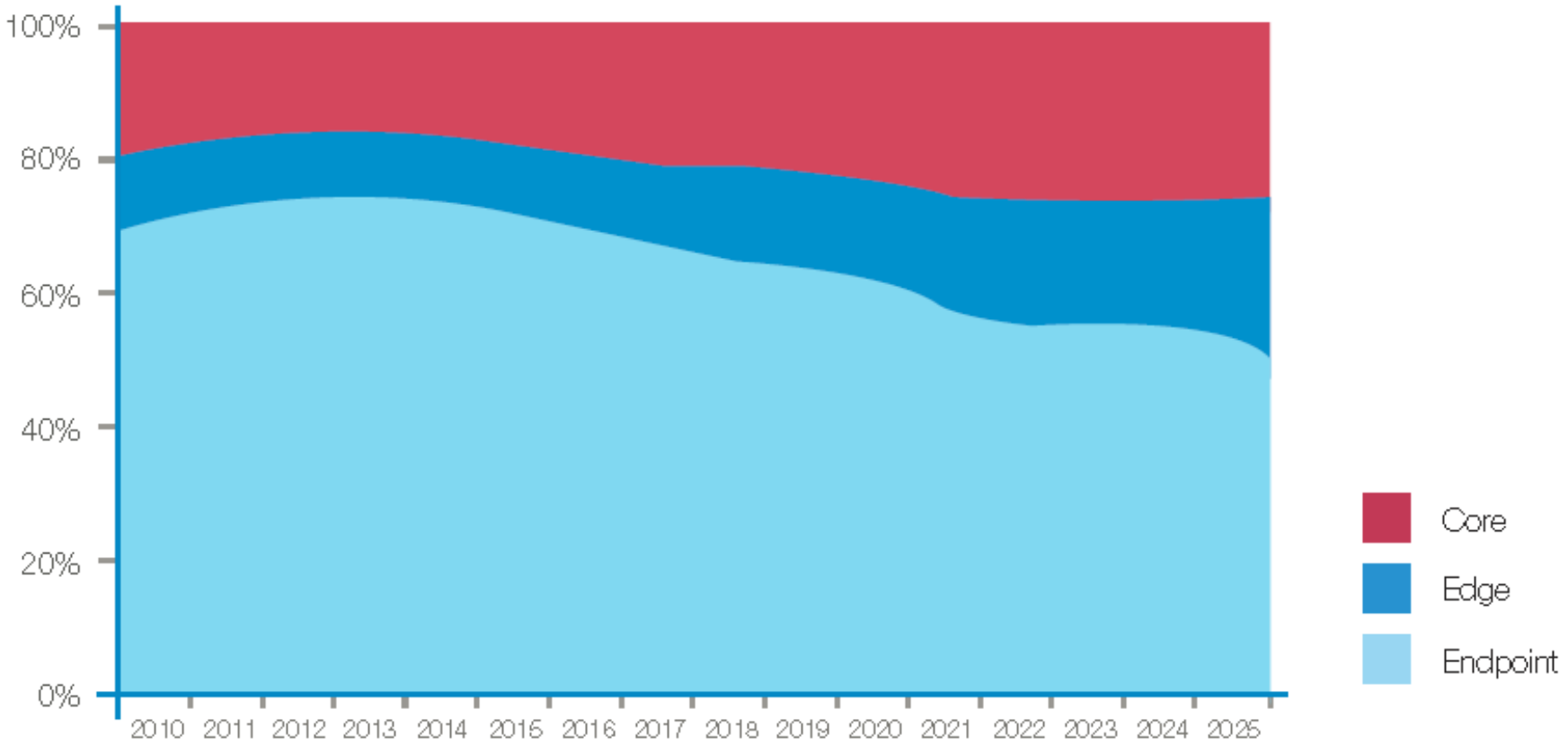
2000 to Today

- 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

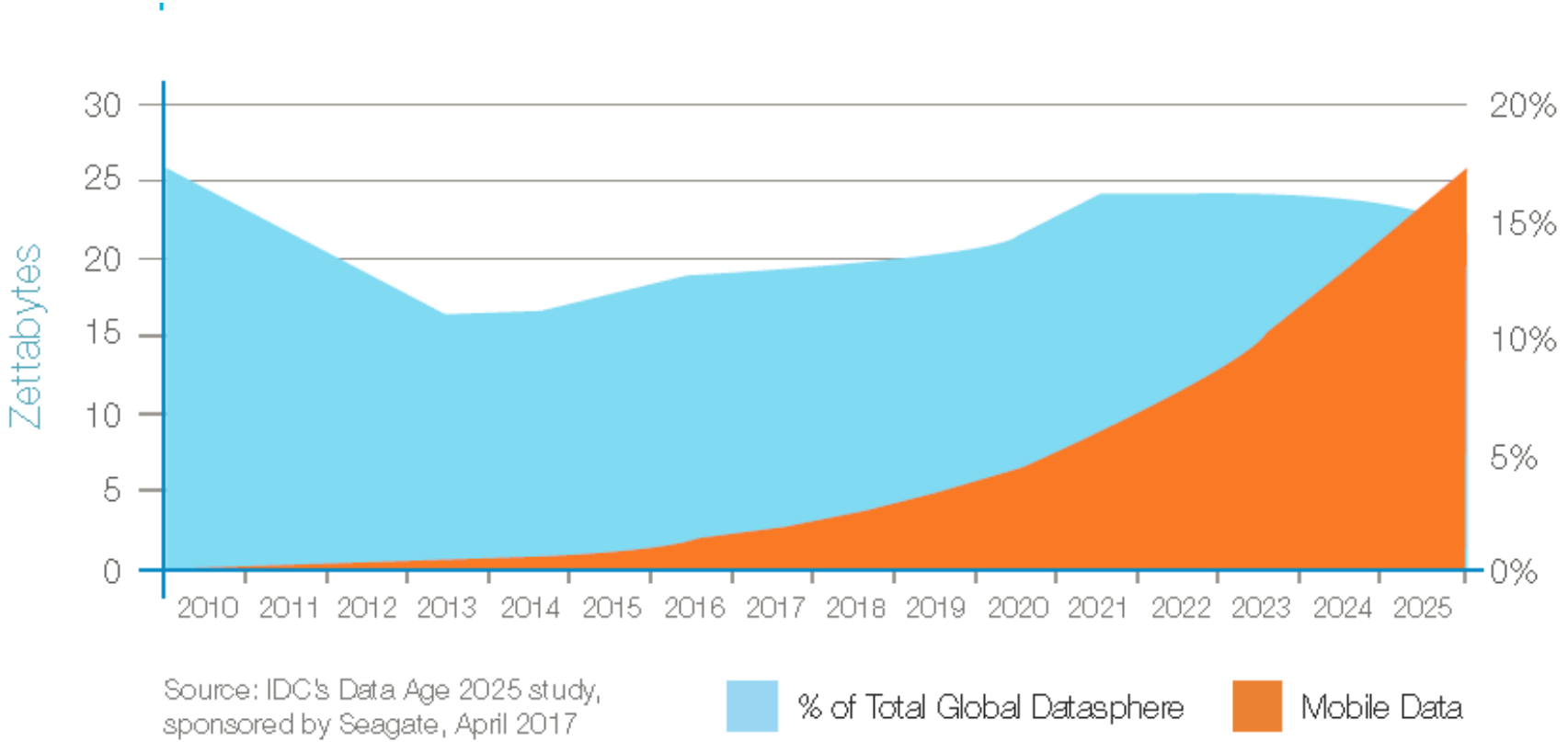
- Data evolutionary role in the world become rapidly apparent in the utilization of data by different computing platform over time
- These locations are classified into three categories:
  - **Core:** It refers a designated computing data center and cloud
  - Example: Public, Private and Hybrid Clouds, 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



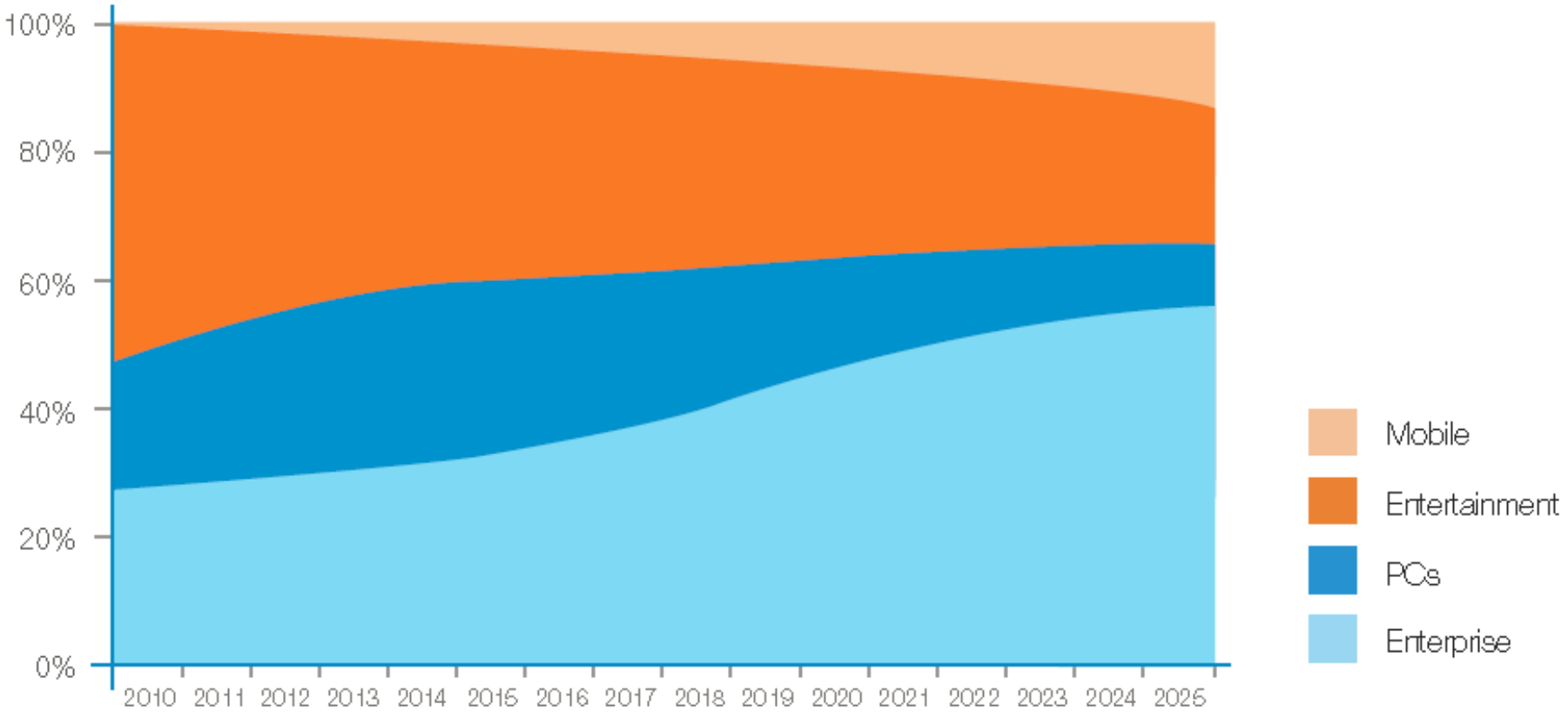
Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

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

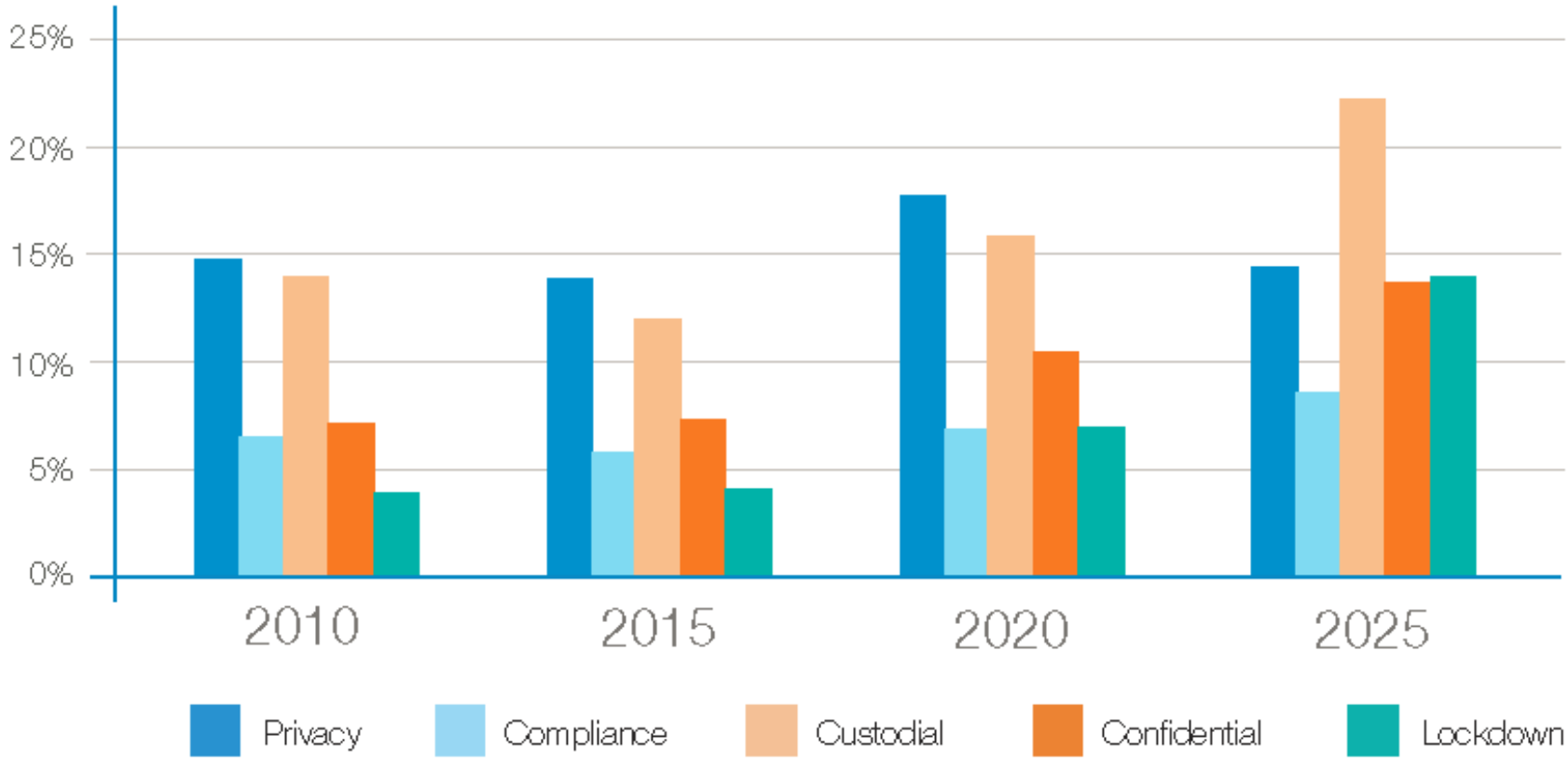


Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

- 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.

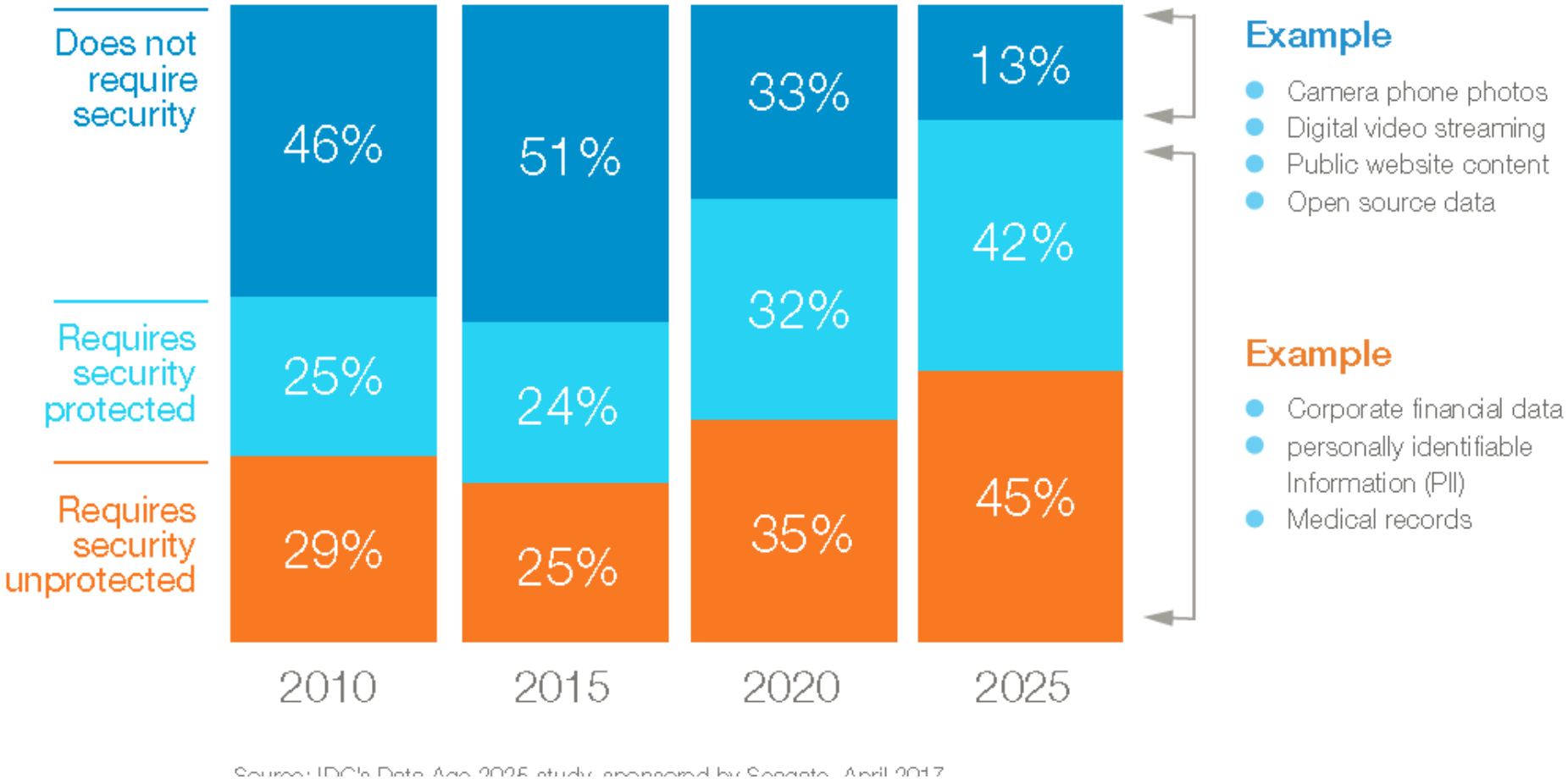
- Five different types of securities.
  - **Lockdown**: Highest Security. Ex: financial transactions, military intelligence, etc
  - **Confidential**: Information that the originator wants to protect. Ex: trade secrets, customer list, memos etc.
  - **Custodial**: Account information that, if breached, 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

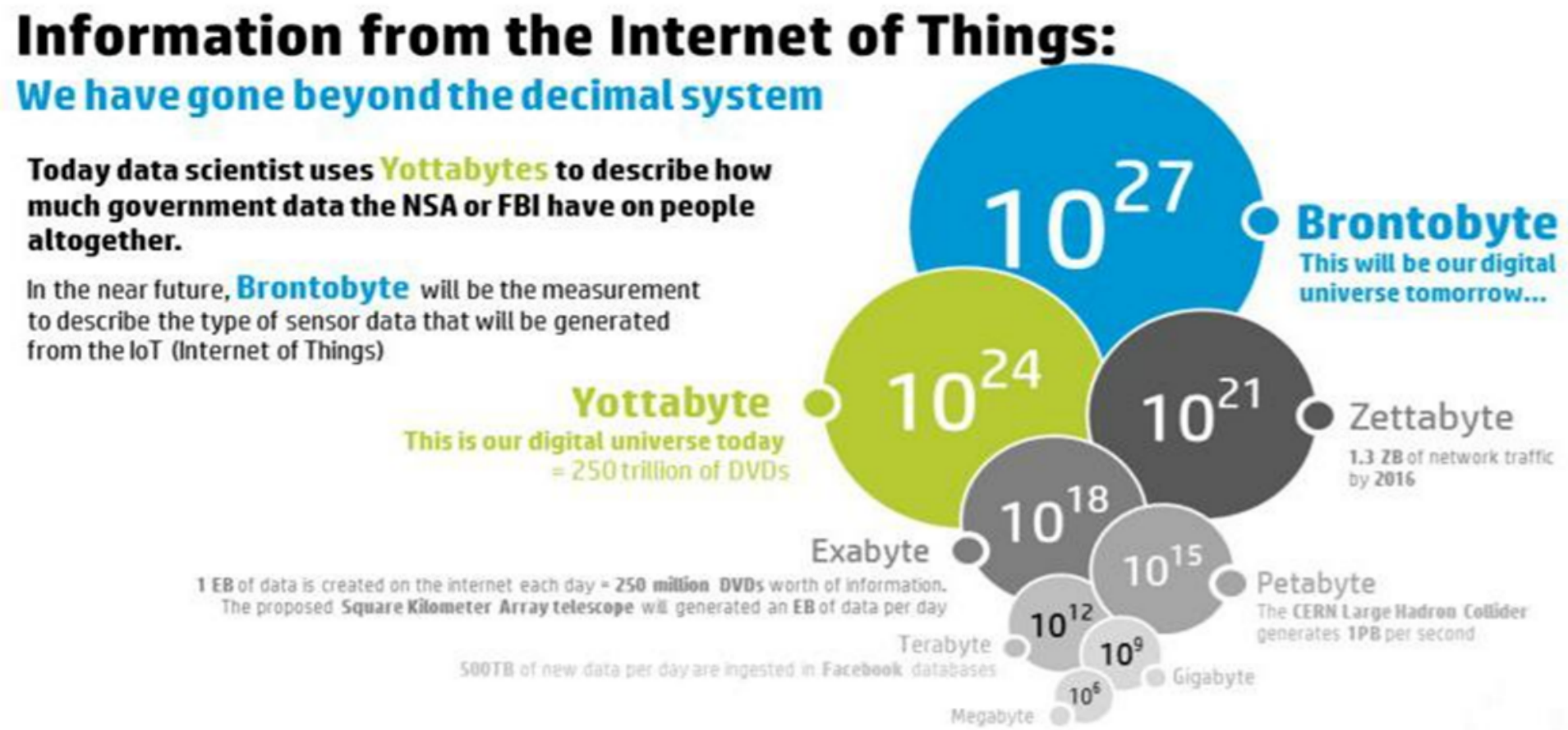


Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

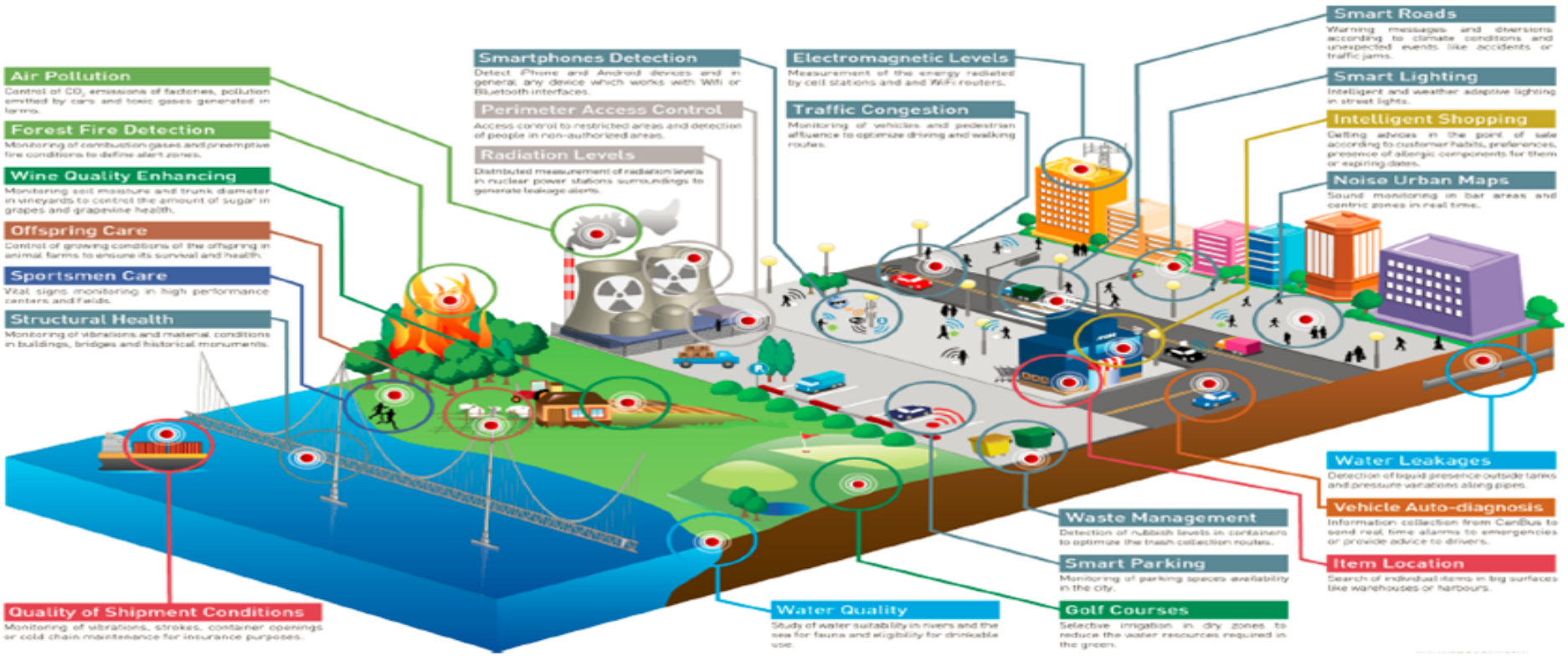
Actual Status of Data Security



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)
10^{27}
Brontobyte
This will be our digital universe tomorrow...
10^{24}
Yottabyte
This is our digital universe today
= 250 trillion of DVDs
10^{21}
Zettabyte
1.3 ZB of network traffic by 2016
10^{18}
Exabyte
1 EB of data is created on the internet each day = 250 million DVDs worth of information.
The proposed Square Kilometer Array telescope will generated an EB of data per day
10^{15}
Petabyte
The CERN Large Hadron Collider generates 1PB per second
10^{12}
Terabyte
500TB of new data per day are ingested in Facebook databases
10^{9}
Gigabyte
10^{6}
Megabyte

Proliferation of Data sources



- Due to the advent of low-cost data storage technologies and the wide availability of Internet connections have made it easier for individuals and organizations to access large amounts of data.
- This data may be heterogeneous.
- Is it possible to convert such data into information and knowledge $\rightarrow$ used for decision making

- It is defined as a set of **mathematical models** and **analysis methodologies** that exploit the available data to generate **information** and **knowledge** useful for complex **decision-making processes**.
- Previously, Knowledge workers are used to take decisions using easy and intuitive methodologies $\rightarrow$ experience, knowledge of the application domain and the available information.
- This approach leads to a **stagnant** decision-making style which is **inappropriate** for the **unstable conditions** $\rightarrow$ frequent and rapid changes in the environment.
- Decision making Process in today's organizations should dynamic, requires rigorous attitude based on analytical methodologies and mathematical models.

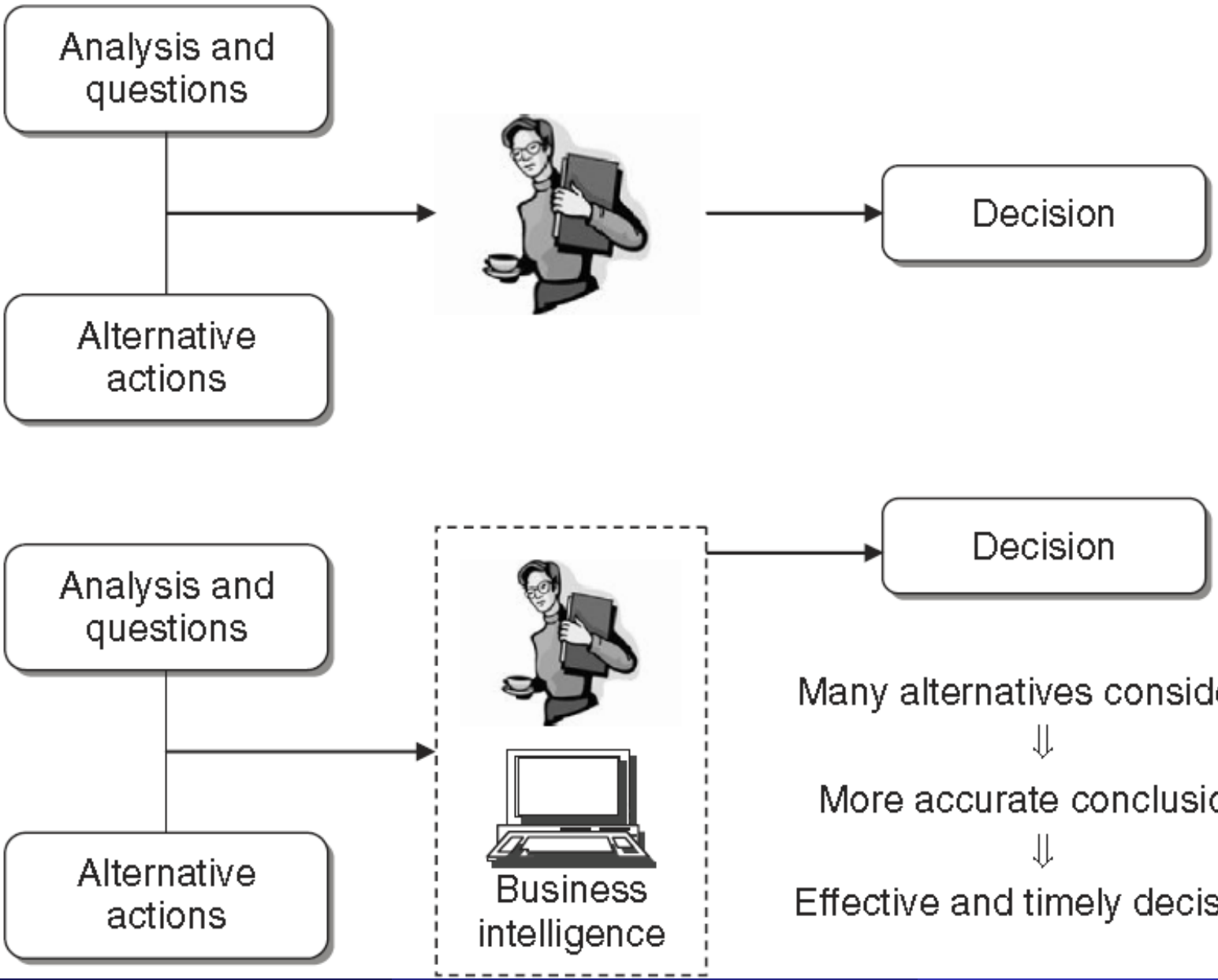
CASE STUDY-1

Retention in the mobile phone industry:The marketing manager of a mobile phone company realizes that a large number of customers are discontinuing their service, leaving her company in favor of some competing provider. As can be imagined, low customer loyalty, also known as customer attrition or churn, is a critical factor for many companies operating in service industries. Suppose that the marketing manager can rely on a budget adequate to pursue a customer retention campaign aimed at 2000 individuals out of a total customer base of 2 million people. Hence, the question naturally arises of how she should go about choosing those customers to be contacted so as to optimize the effectiveness of the campaign. In other words, how can the probability that each single customer will discontinue the service be estimated so as to target the best group of customers and thus reduce churning and maximize customer retention? By knowing these probabilities, the target group can be chosen as the 2000 people having the highest churn likelihood among the customers of high business value. Hence it requies an advanced mathematical models and data mining techniques to derive a reliable estimate of the churn probability and to determine the best recipients of a specific marketing campaign.

CASE STUDY-2

Logistics planning:The logistics manager of a manufacturing company wishes to develop a medium-term logistic-production plan. This is a decision-making process of high complexity which includes, among other choices, the allocation of the demand originating from different market areas to the production sites, the procurement of raw materials and purchased parts from suppliers, the production planning of the plants and the distribution of end products to market areas. In a typical manufacturing company this could well entail tens of facilities, hundreds of suppliers, and thousands of finished goods and components, over a time span of one year divided into weeks. The magnitude and complexity of the problem suggest that advanced optimization models are required to devise the best logistic plan.

- The **main purpose** of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make **effective** and **timely decisions**.
- **Effective Decision**:
  - The application of rigorous analytical methods allows decision makers to rely on **information** and **knowledge** which are more dependable.
  - It ensures in-depth examination and thought lead to a deeper awareness and comprehension of the underlying logic of the decision-making process.
- **Timely decisions**:
  - The ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company.

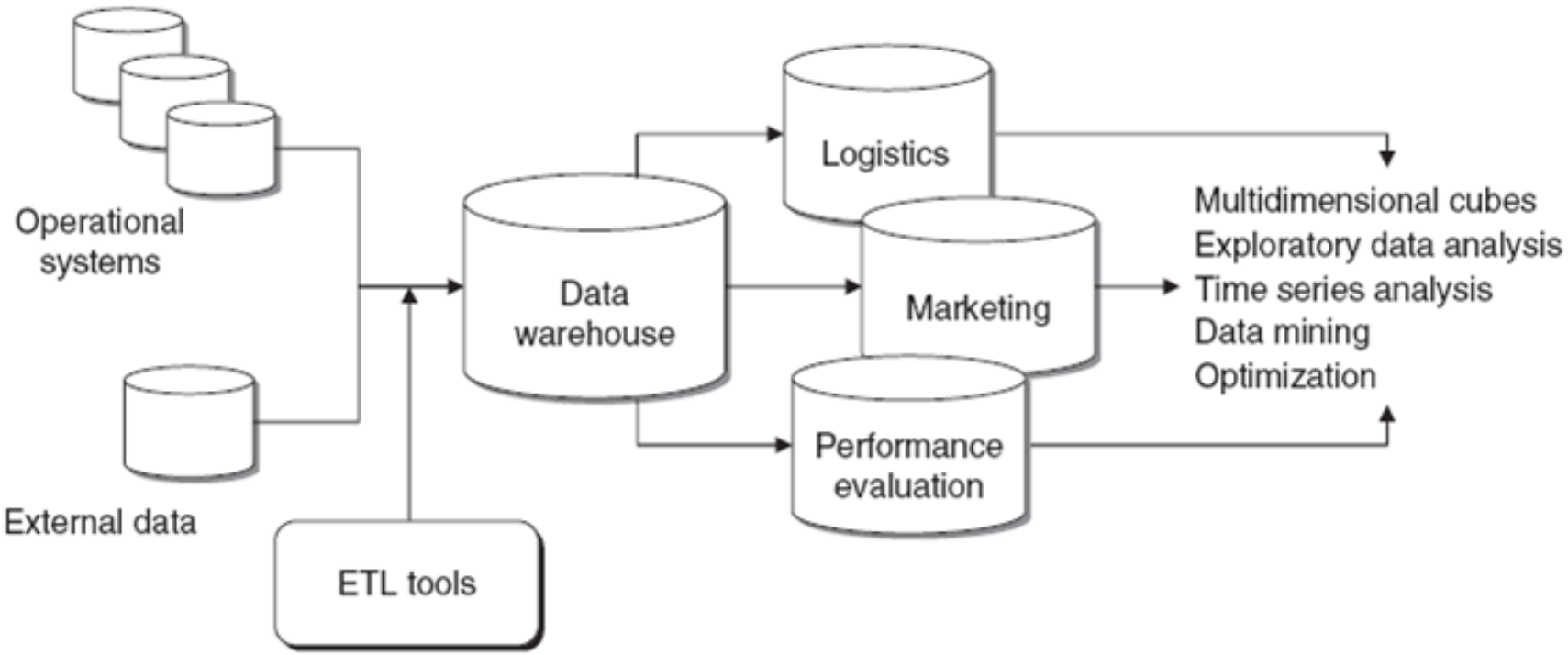

Analysis and questions
Alternative actions
Decision
Analysis and questions
Alternative actions
Business intelligence
Decision
Many alternatives considered
⇓
More accurate conclusions
⇓
Effective and timely decisions

- If the decision makers uses BI system, then overall quality of the system will be greatly improved.
- With the help of mathematical models and algorithms, BI has a possibility of analyzing a larger number of alternative actions, achieve more accurate conclusions and reach effective and timely decisions
- The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as **knowledge management**

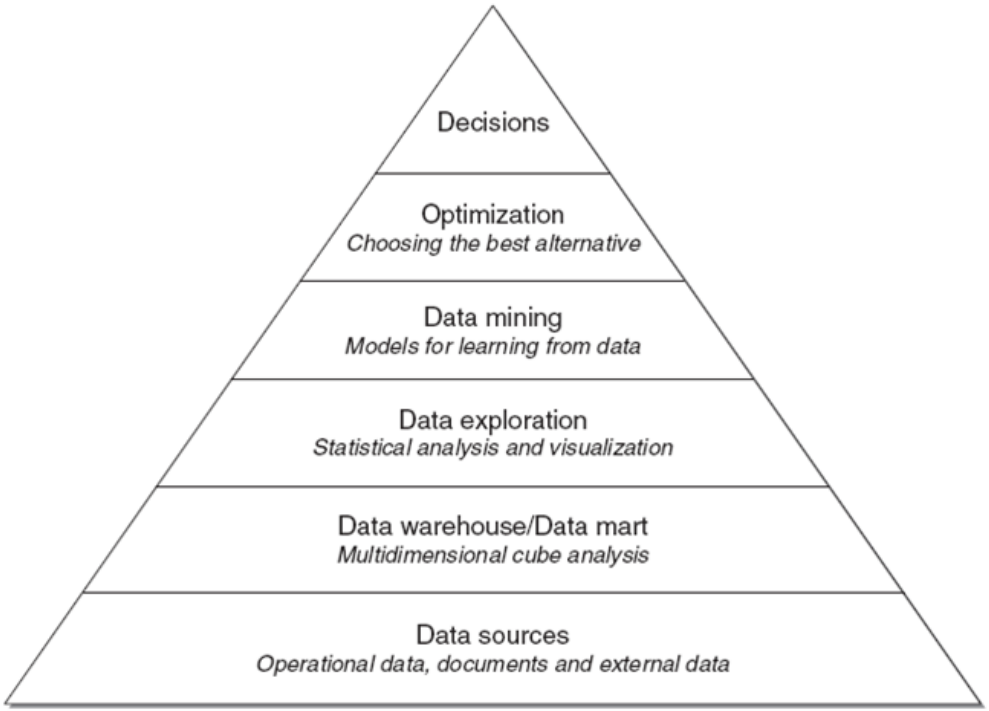
- BI and knowledge management share some degree of similarity in their objectives → both are helping knowledge worker for decision making process.
- Boundary between BI and Knowledge-Management Systems
  - Knowledge Management methodologies are focusing on the treatment of information that is usually **unstructured**, at times implicit, contained mostly in documents, conversations and past experience.
  - BI systems are based on **structured information**, most often of a quantitative nature and usually organized in a database.
  - Recently BI, with the help of other technologies, it is also started to process unstructured data up to some level.

Business Intelligence Architecture

- Business Intelligence Architecture consists of 3 levels
- 1. Data sources 2. Data Warehouses and Data marts and 3. BI methodologies



- Main Components of BI Architecture



Business Intelligence Architecture

- Data sources:
  - operational systems,
  - may also include unstructured documents, such as emails and data received from external providers.
- Data warehouses and data marts:
  - The data originating from different sources are uses **ETL** tools to store in database, which intends to support BI analysis.
- Business intelligence methodologies:
  - Several decision support applications may be implemented:
  - 1. multidimensional cube analysis
  - 2. exploratory data analysis
  - 3. time series analysis
  - 4. inductive learning models for data mining
  - 5. optimization models

- **Data Exploration** (Passive Business Intelligence Analysis)
  - It consists of
    - 1. Query and Reporting Systems
    - 2. Statistical Methods.
  - Decision makers are expects prior hypotheses or defined data extraction and then use the analysis tools to find answers and confirm their original insight.
- **Data Mining** (Active Business Intelligence Methodologies)
  - Extraction of information and knowledge from data
  - These include mathematical models for pattern recognition and Machine Learning
  - The models of an active kind do not require decision makers to formulate any prior hypothesis to be later verified.
  - Their purpose is instead to expand the decision maker's knowledge.
- **Optimization**
  - To determine the best solution out of a set of alternative actions.
  - This technique is usually fairly extensive and sometimes even infinite.
- **Decision**
  - The choice and the Actual Adoption of a Specific Decision

Data Repositories from an Analyst Perspective

- **Spreadsheets and Data marts**
  - Used for record keeping
  - Analyst depends upon data extracts
- **Data Warehouses**
  - Centralized data containers in purpose-built space
  - Support BI and Reporting, but restrict robust analysis
  - Analyst dependent on IT and DBAs for data access and schema changes
  - Analyst must spend time to get aggregated and dis-aggregated data extracts from multiple sources
- **Analytic Sandbox or Workspace**
  - Data assets are gathered from multiple sources and technologies for analysis
  - Enables flexible and high performance analysis, leverages in database processing
  - Analyst owned rather than DBA owned

Business Driver	**Examples**
Optimized Business Operations	Sales, Pricing, etc
Identify Business Risk	Customer Churn, Fraud, etc
Predict New Business Opportunities	Upsell, Cross-sell, best new customer prospects

- Rather than only performing standard reportings, organizations can apply advanced analytical techniques to optimize processes and derive more value from these common tasks.
- Requires variety of analytical techniques to address these problems.

Data Analytics Life Cycle

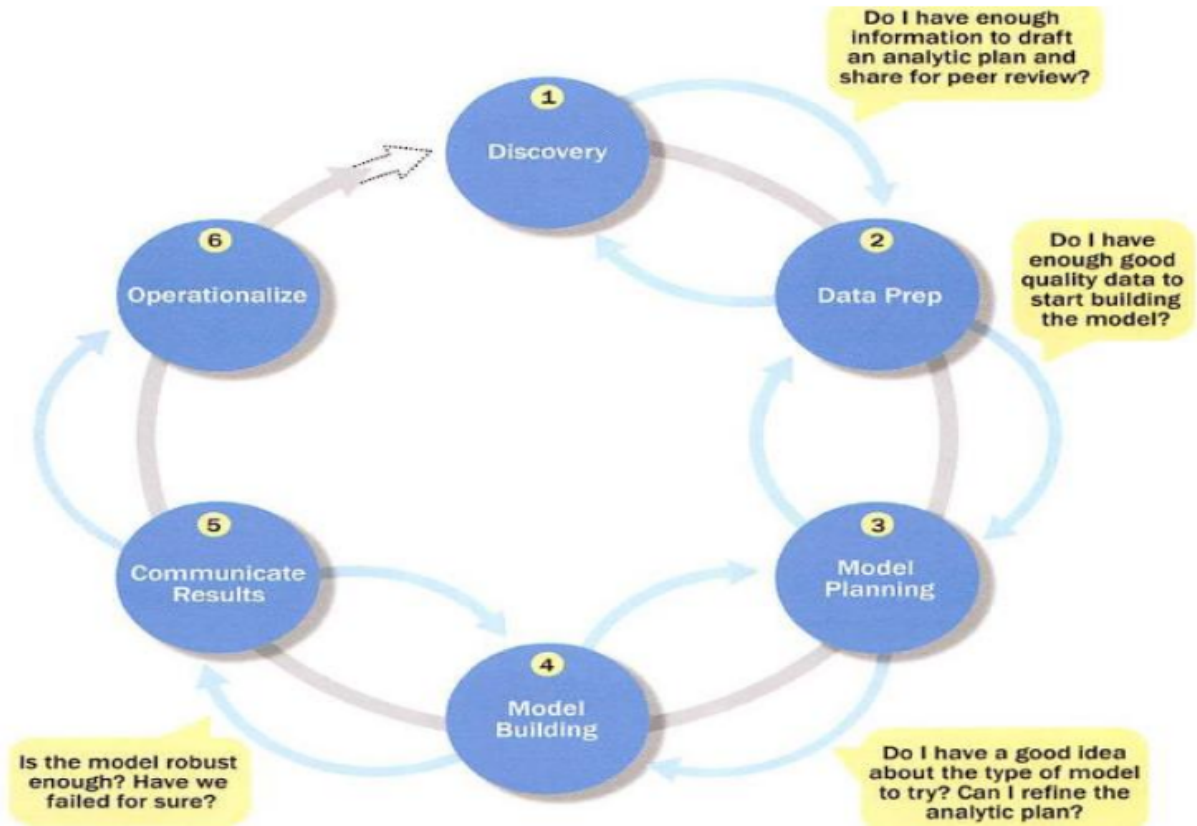
- Data science projects differ from BI projects
  - More exploratory in nature
  - Critical to have a project process
  - Participants should be thorough and rigorous
- Break large projects into smaller pieces
- Data Analytics Lifecycle defines **the analytics process and best practices from discovery to project completion**

Data Analytics lifecycle Phases:

- Discovery Phase
- Data Preparation Phase
- Model Planning Phase
- Model Building Phase
- Communicate Result
- Operationalize

Data Analytics Life Cycle

- With six phases the project work can occur in several phases simultaneously
- The cycle is iterative to portray a real project
- Work can return to earlier phases as new information is uncovered



- **Phase - 1: Discovery**
  - Learning the Business Domain
  - Resources
  - Framing the Problem
  - Developing Initial Hypotheses
  - Identifying Potential Data Sources

- **Phase - 2: Data Preparation**
  - It requires analytical sandbox in which you can perform analytics for the entire duration of the project
  - Includes steps to
    - Explore
    - Preprocess
    - Conditional Data
  - Data preparation tends to be the most labor-intensive step in the analytics lifecycle
    - Often at least 50% of the data science project's time
  - The data preparation phase is a iterative process

- **Phase - 2: Data Preparation**
  - In ETL users perform extract, transform, load
  - Data Analytics lifecycle → ELT or ETLT → Extract, Transform, Load and Transform
    - early load preserves the raw data which can be useful to examine.
  - **Example:**
    - In credit card fraud detection, outliers can represent high-risk transactions that might be inadvertently filtered out or transformed before being loaded into the database

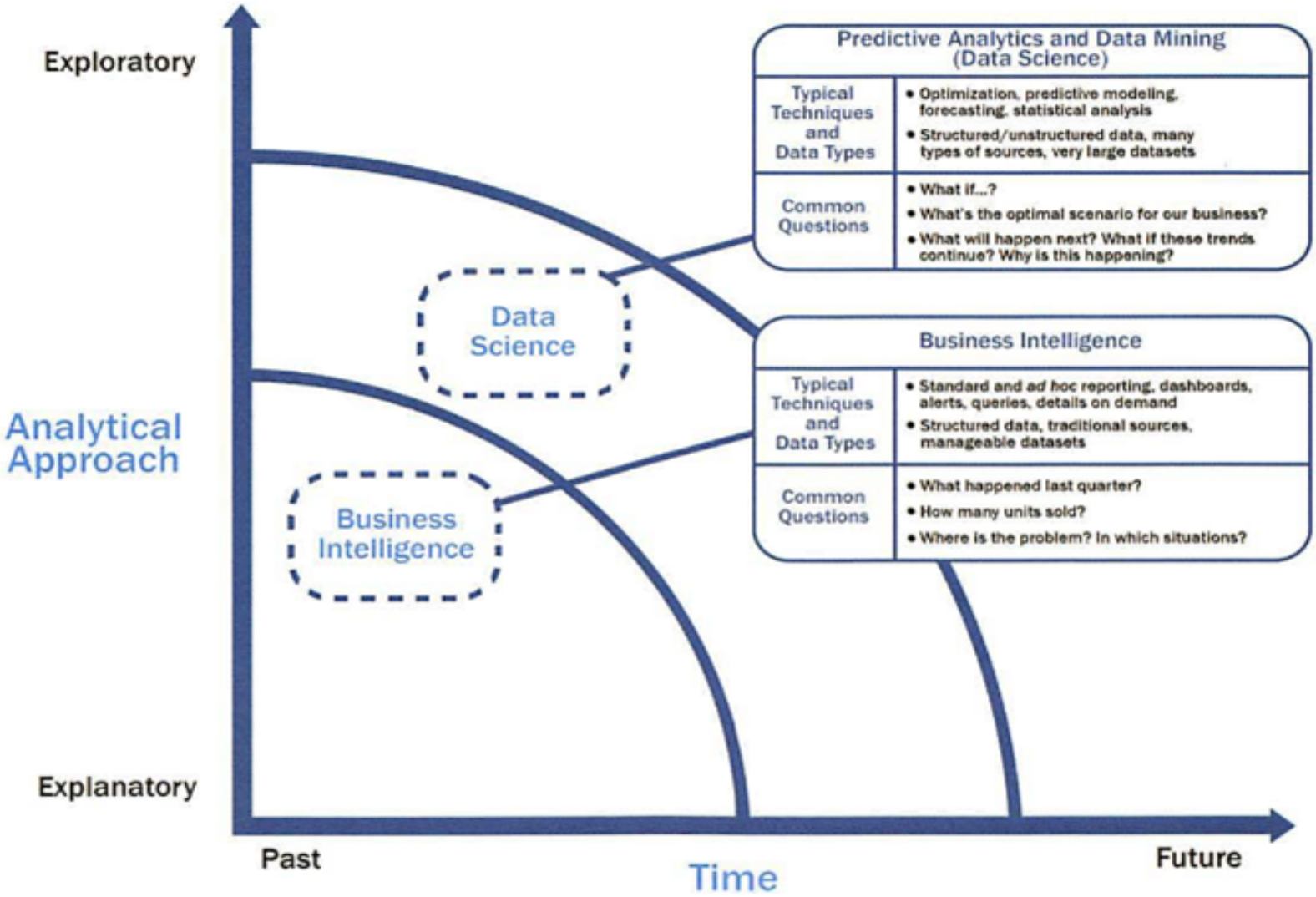
- **Phase - 3: Model Planning**
  - This determines the methods and techniques to extract relationships among variables.
  - These relationship patterns will set the base for algorithms which will be used in next phase
  - It uses Exploratory Data Analysis (EDA) using various statistical formulae and visualization tools.
  - Simply, it identifies candidate models to apply to the data for clustering, classifying, or finding relationships in data.
  - Activities to consider:
    - Assess the structure of the data
    - Ensure the analytic techniques enable the team to meet the business objectives and accept or reject the working hypotheses
    - Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow
    - Research and understand how other analysts have approached this kind or similar kind of problem

- **Phase - 4: Model Building**
  - Execute the models defined in Phase 3
  - Develop datasets for training, testing, and production
  - Develop analytic model on training data, test on test data.
  - It will consider whether your existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing).
  - You will analyze various learning techniques like classification, association and clustering to build the model.

- **Phase - 5: Communicate Results**
  - Determine if the team succeeded or failed in its objectives
  - Assess if the results are statistically significant and valid
    - If so, identify aspects of the results that present salient findings
    - Identify surprising results and those in line with the hypotheses
  - Communicate and document the key findings and major insights derived from the analysis
  - This is the most visible portion of the process to the outside stakeholders and sponsors

- **Phase - 6: Operationalize**
  - In this last phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way
  - Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout
  - During the pilot project, the team may need to execute the algorithm more efficiently in the database

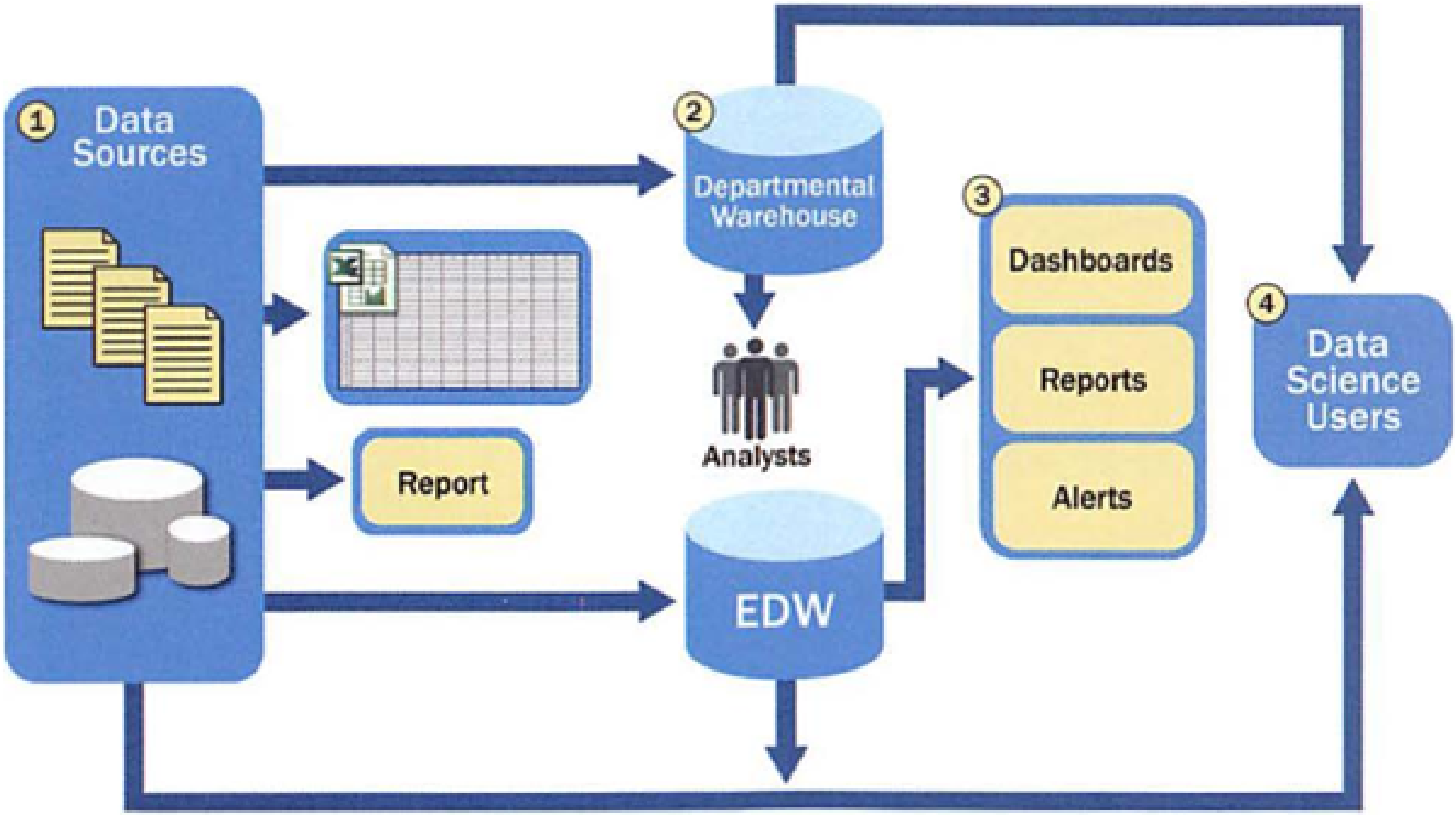
BI Vs. Data Science



BI vs. Data Science

- Evaluation can be based on time horizon and analytical approaches used in BI and Data Analytics
- BI tends to provide reports, dashboards and queries on business for the current period of in the past
- The questions tend to be closed-ended.
- BI answers the questions related to "When" and "Where" events occurred.
- BI requires highly structured data organized in rows and columns. $\rightarrow$ accurate reporting etc.
- Data Science tends to use disaggregated data in a more forward-looking, exploratory way, focusing on analyzing the present and enabling informed decisions about the future.
- Time series techniques are mostly used.
- Open-ended question can be answered by Data Science
- Data Science answers the questions related to "How" and "Why" events occurred.
- Data Science uses larges and unconvential datasets $\rightarrow$ Reports,

Typical Data Analytical Architectur



Typical Data Analytical Architectur

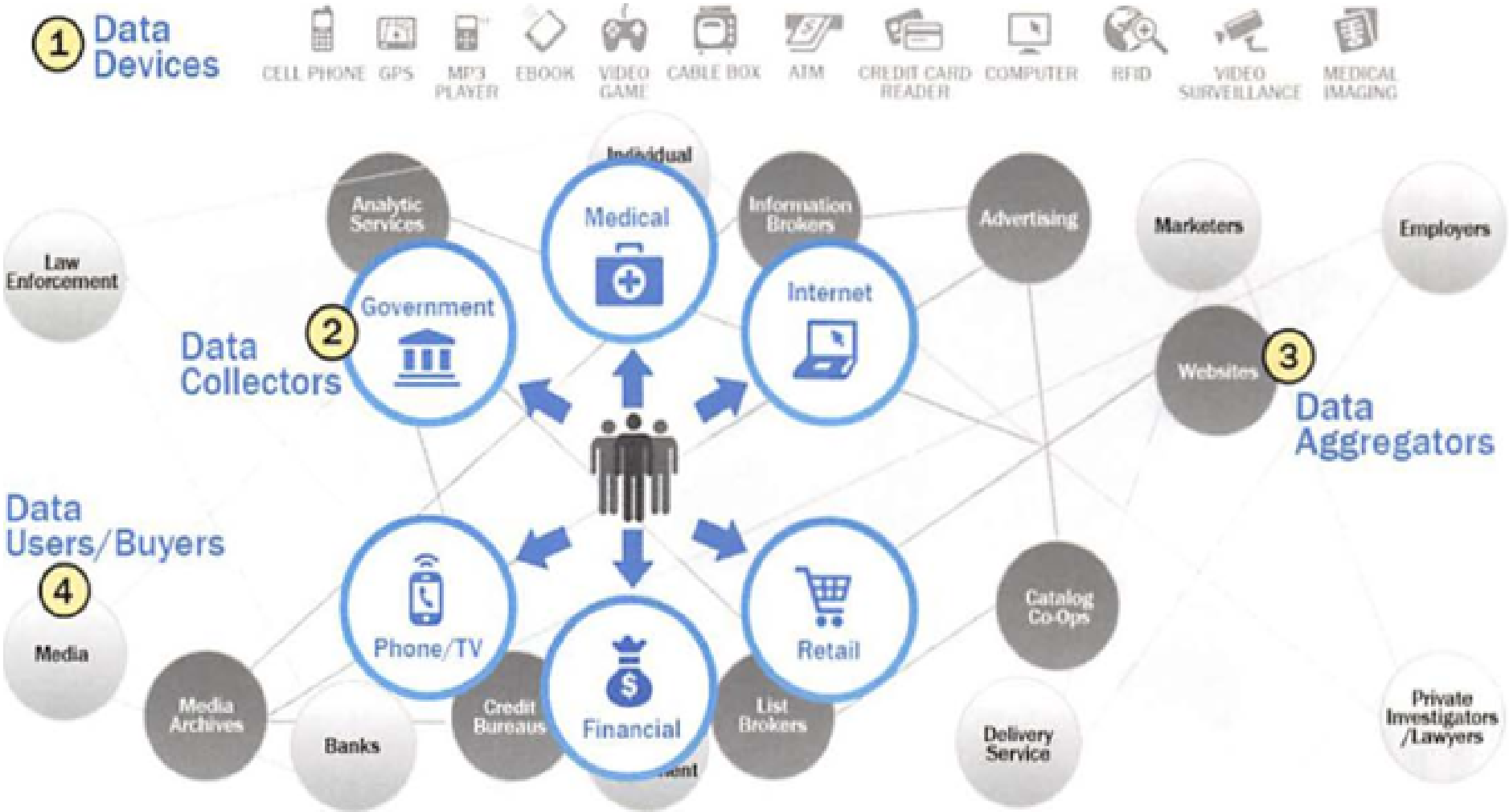
- Data Sources:
  - For data sources to be loaded into the data warehouse, data needs to be well understood, structured, and normalized with the appropriate data type definitions.
  - Must go through the preprocessing and check points
- Warehouse
  - Control on EDW, additional local file systems can be emerged inform of departmental warehouse and local datamarts.
  - Local datamarts allows to do some level of in-depth analysis.
  - one-off systems reside in isolation
- Reporting and Dashboard
  - data is read by additional applications across the enterprise for BI and reporting purposes
- Data science user
  - At the end of this workflow, analysts get data provisioned for their downstream analytics.
  - analysts create data extracts from the EDW to analyze data offline or other local analytical tools

New data sources slowly accumulate in EDW $\rightarrow$ rigorous validation and data structuring process

problems in typical Data Analytical Architecture

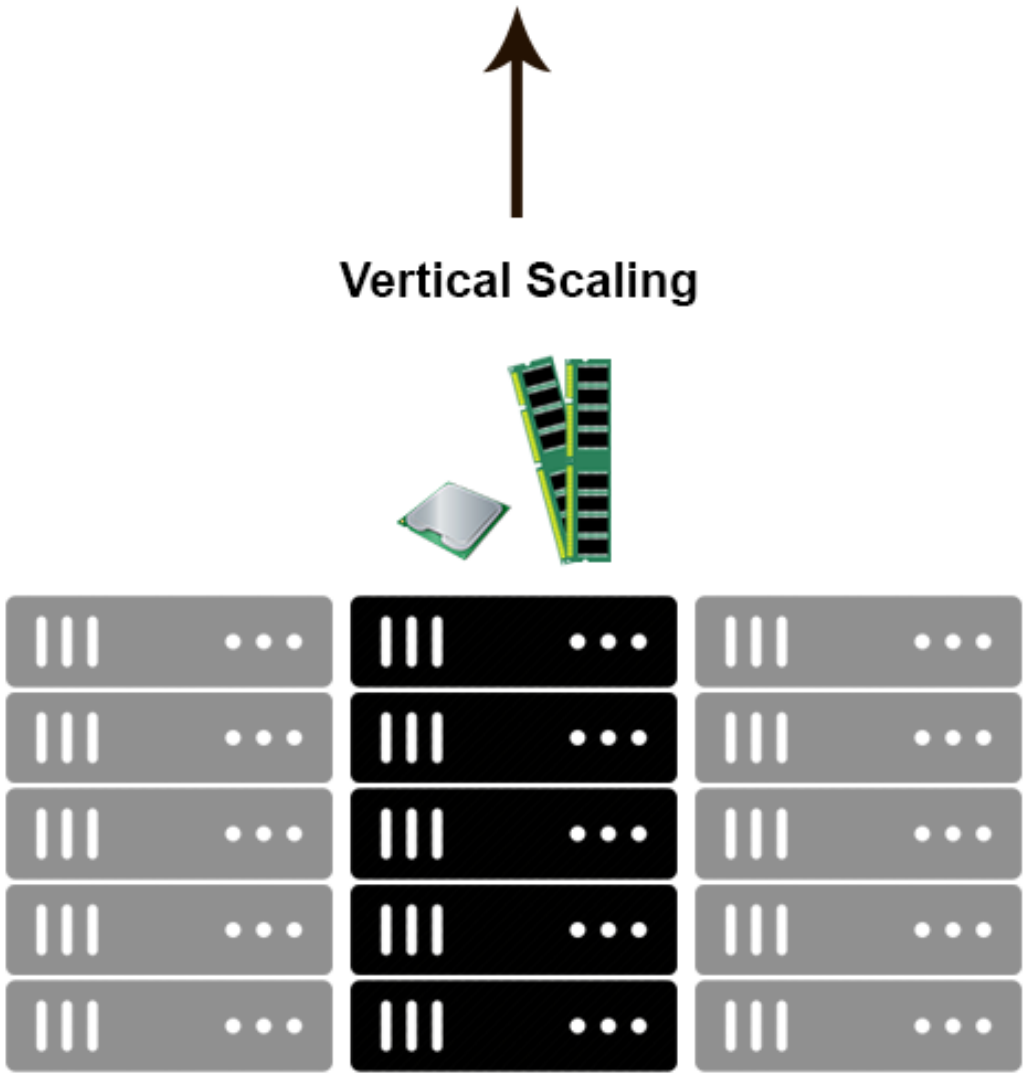
- Traditional data architectures have several additional implications for data scientists
  - High-value data is hard to reach and leverage
  - predictive analytics and data mining activities are last in line for data
  - More priority is given to operational process (at EDW).
  - Data moves in batches from EDW to local analytical tools. $\rightarrow$ limited to perform in-memory analytics $\rightarrow$ skew the model accuracy
  - Data Science projects will remain isolated and adhoc, rather than centrally managed
  - traditional data architecture is a slow "time-to-insight" $\rightarrow$ alternative $\rightarrow$ analytic sandboxes $\rightarrow$ it allows to perform advanced analytics in controlled way

New Approach to Analytics



- **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

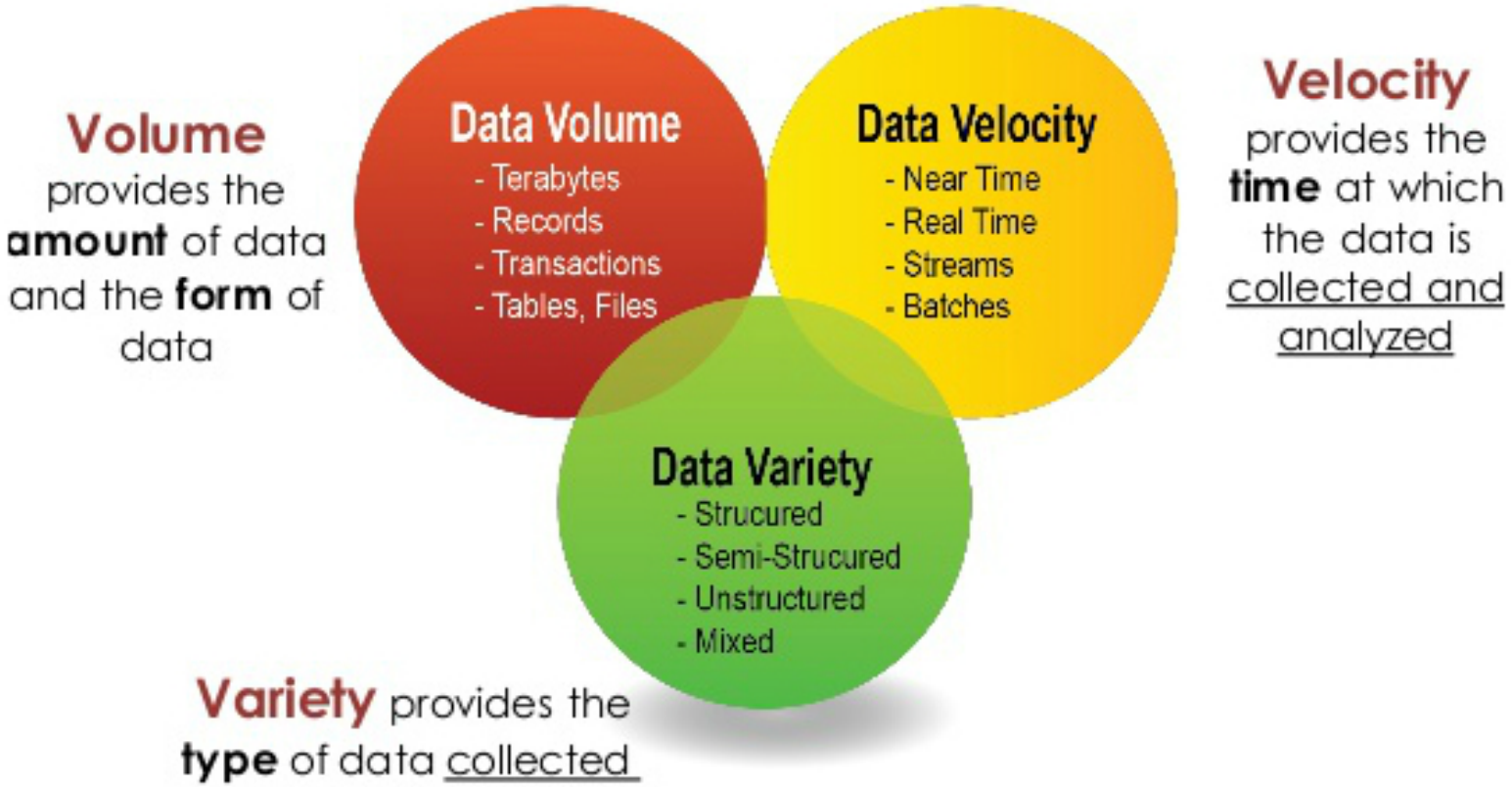
What is Big Data

- **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

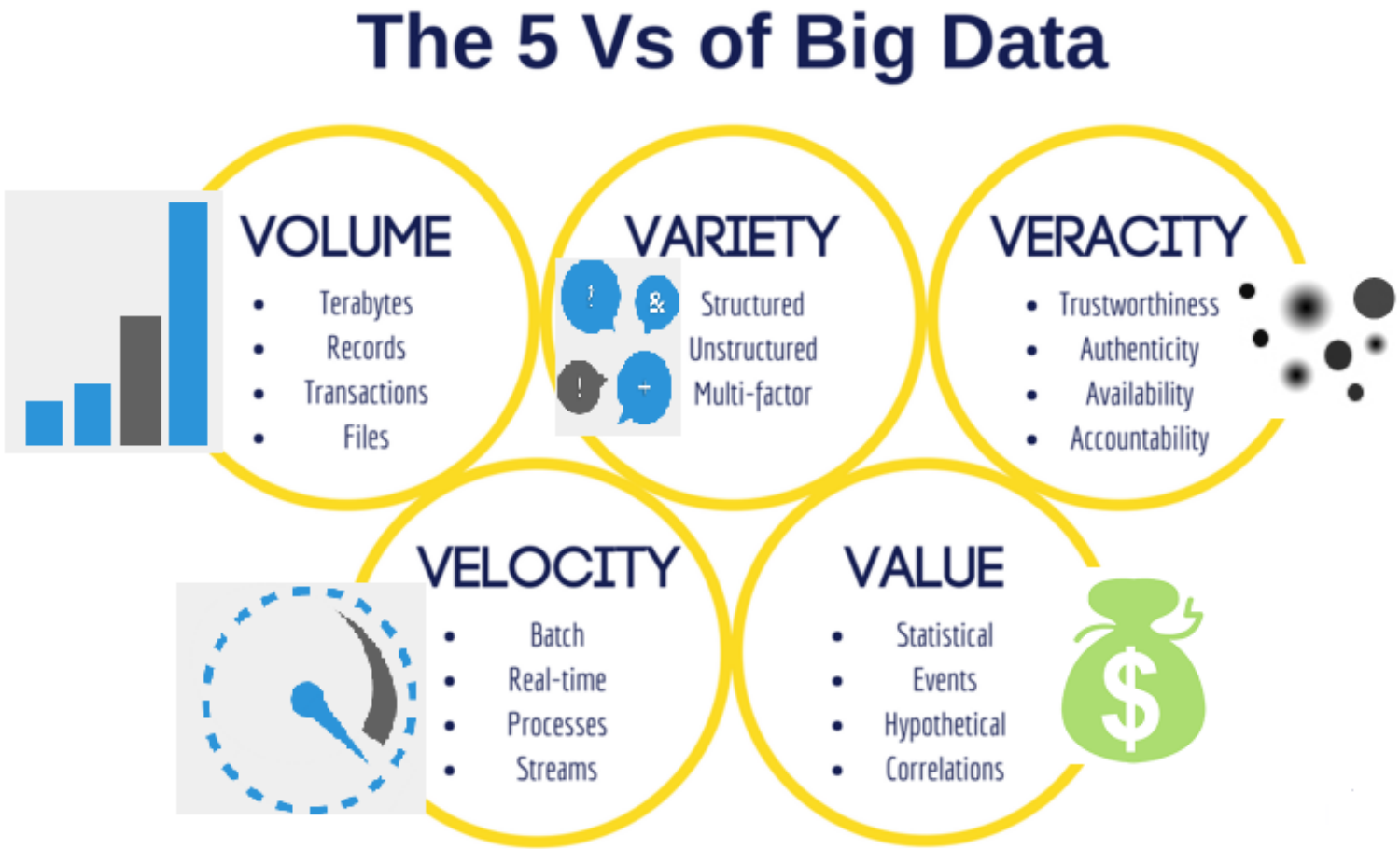
Why is Big Data

- 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



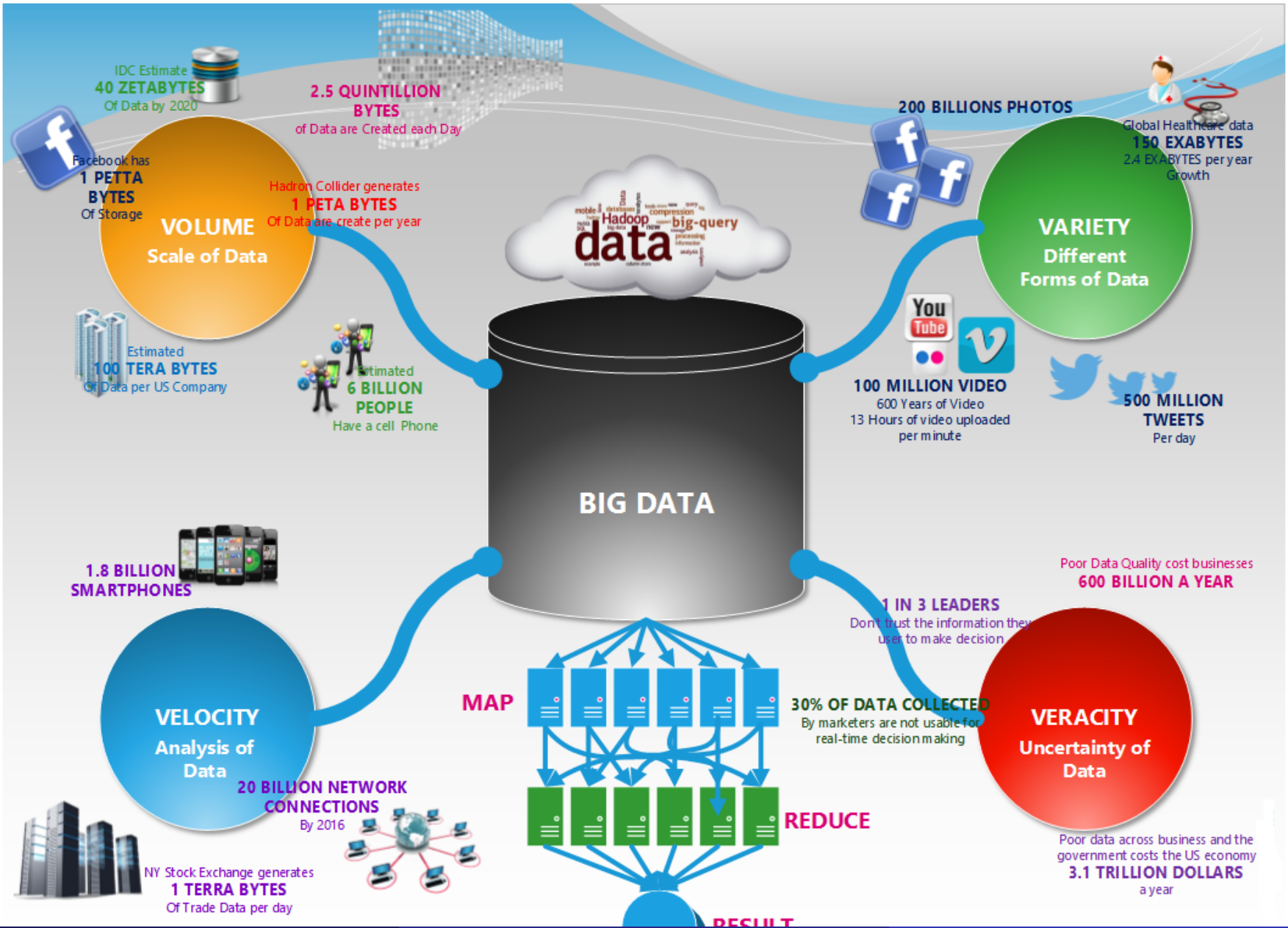
Extended (3+n)V Model

- **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



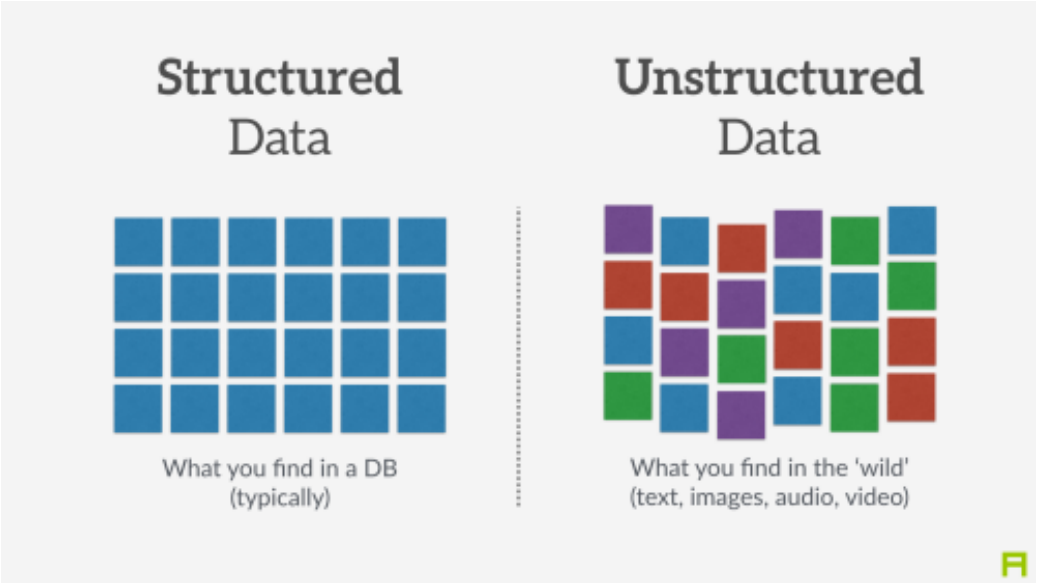
- **Structured Data**
  - Pre-defined schema imposed on the data
  - Highly structured
  - Usually stored in a relational database system
  - Structured data is relatively simple to enter, store, query, and analyze, but it must be strictly defined in terms of field name and type
- Example:
  - numbers: 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