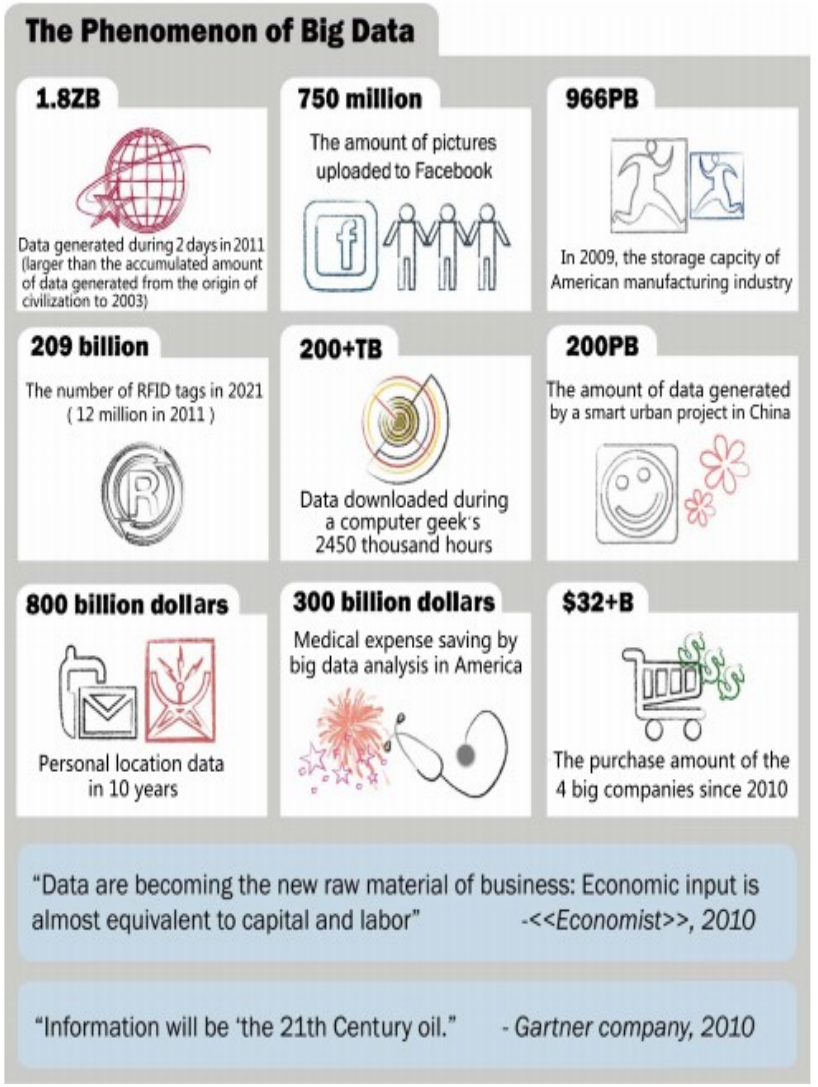
Big Data: An Introduction

K. M. AZHARUL HASAN

What is Big data?

- In general, big data shall mean the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time.
- According to a report from International Data Corporation (IDC), in 2011, the overall created and copied data volume in the world was 1.8ZB which increased by nearly nine times within five years . This figure will double at least every other two years in the near future.



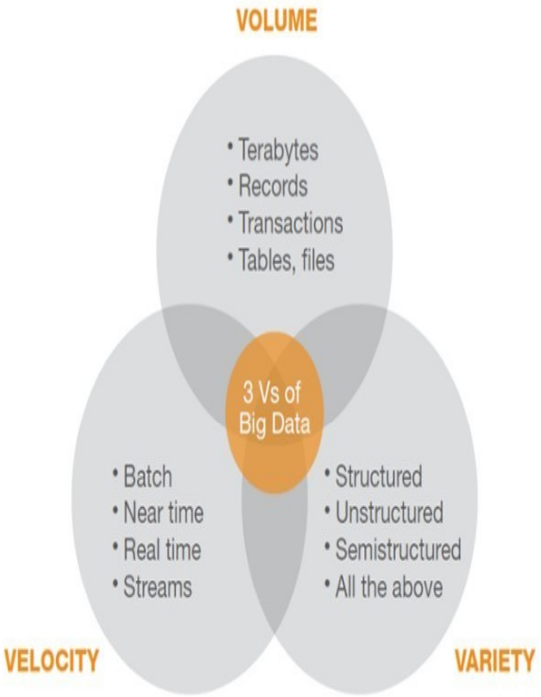
Storage Capacity revisited

1024	KB	kilobyte	2^{10}
1024^2	MB	megabyte	2^{20}
1024^3	GB	gigabyte	2^{30}
1024^4	TB	Terabyte	2^{40}
1024^5	PB	Petabyte	2^{50}
1024^6	EB	Exabyte	2^{60}
1024^7	ZB	Zettabyte	2^{70}
1024^8	YB	Yottabyte	2^{80}

As of 2015, there are no approved standard sizes for anything bigger than a Yottabyte. However, the two standards that have been proposed are the Hellabyte or Brontobyte.

What is Big data?

- In another way, Big Data simply means that huge amount of structured, semi-structured and unstructured data that has the potential to be processed for information.
- Big data also brings about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and also incurs new challenges, e.g., how to effectively organize and manage such datasets.
- Big data is defined by three Vs (some times four vs).
  - ***Volume*** : amount of data
  - ***Variety***: the number of types of data
  - ***Velocity*** : the speed of data processing.



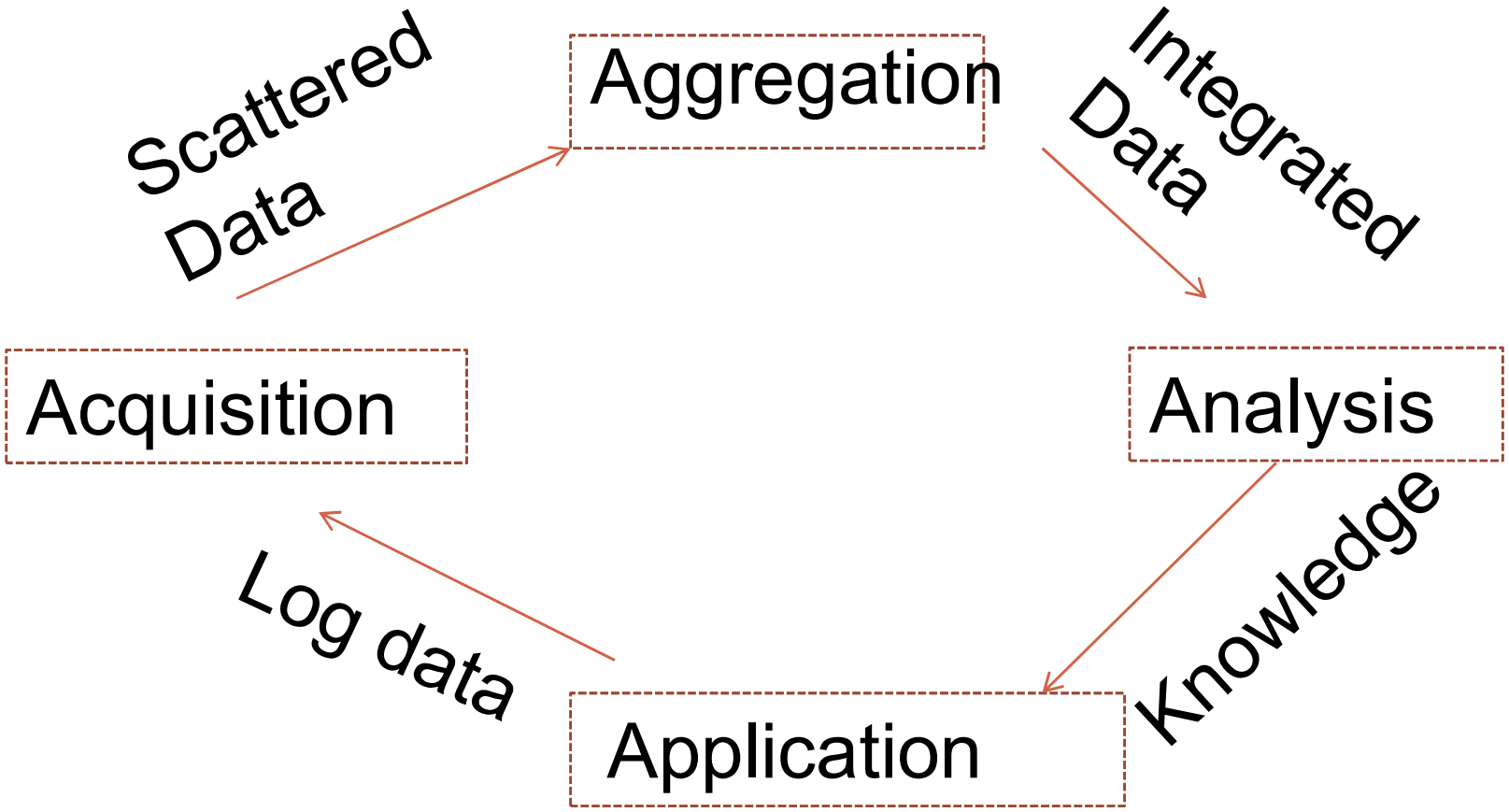
Challenges

- Traditional data management and analysis systems are based on the relational database management system (RDBMS).
- However, such RDBMSs only apply to structured data, other than semi-structured or unstructured data.
- The traditional RDBMSs could not handle the huge volume and heterogeneity of big data.
- For solutions of permanent storage and management of large-scale disordered datasets, distributed file systems and NoSQL databases are good choices.

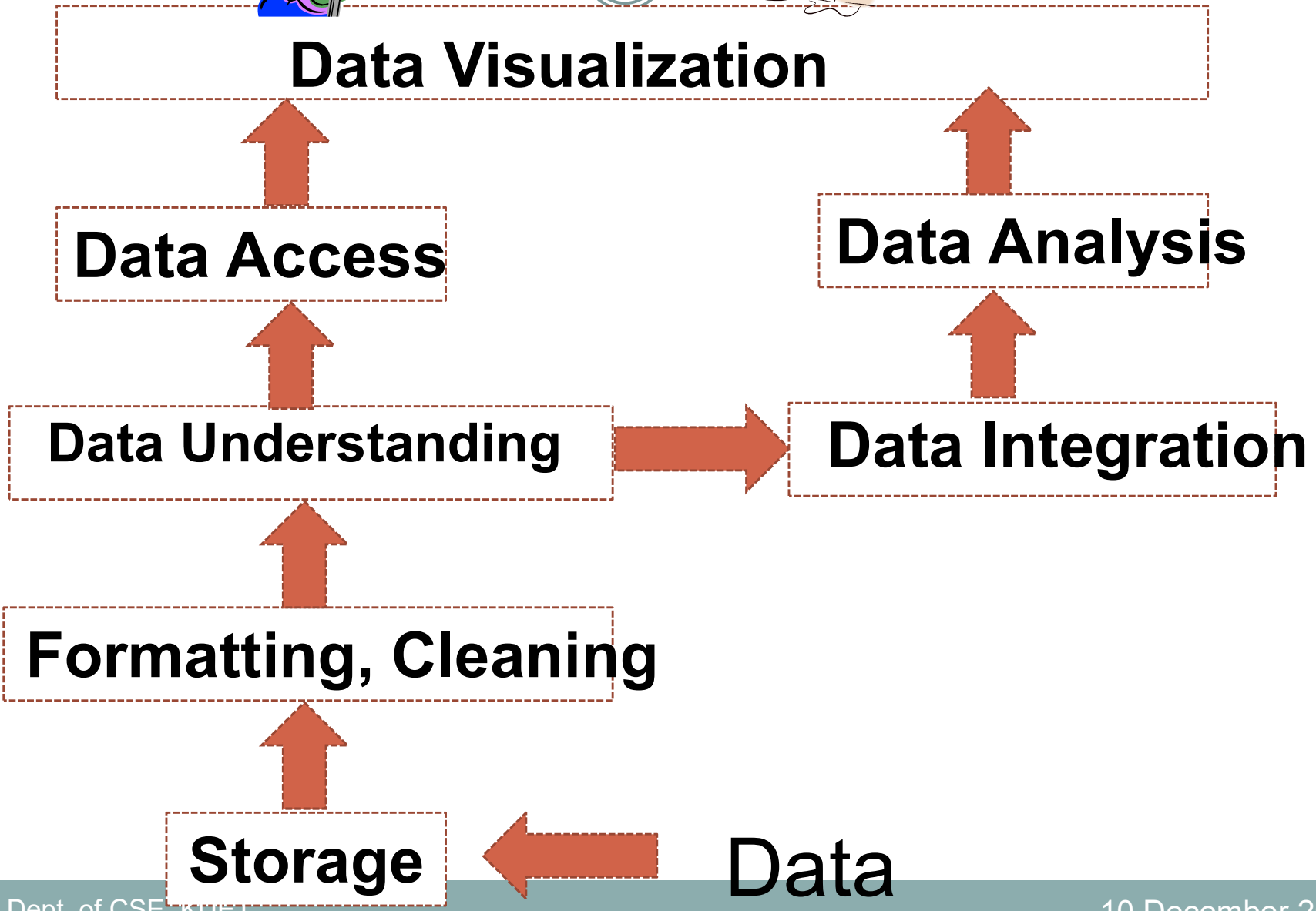
Challenges of big data applications

- The key challenges:
  - Data representation
  - Redundancy reduction and data compression
  - Data life cycle
  - Analytical mechanism
  - Data confidentiality
  - Energy management
  - Expendability and scalability
  - Cooperation

Lifecycle of Data: 4 “A”s



Computational View of Big Data



Big Data & Related Topics/Courses

Human-Computer Interaction

Data Visualization

Databases

Information Retrieval

Machine Learning

Data Access

Data Analysis

Computer Vision

Speech Recognition

Data Mining

Data Understanding

Data Integration

Natural Language Processing

Data Warehousing

Formatting, Cleaning

Signal Processing, Statistics

Many Applications!

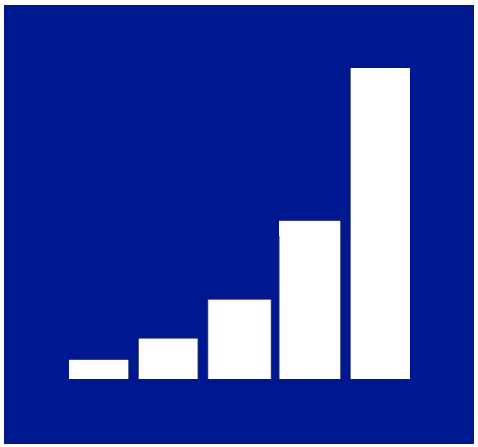
Data

Storage

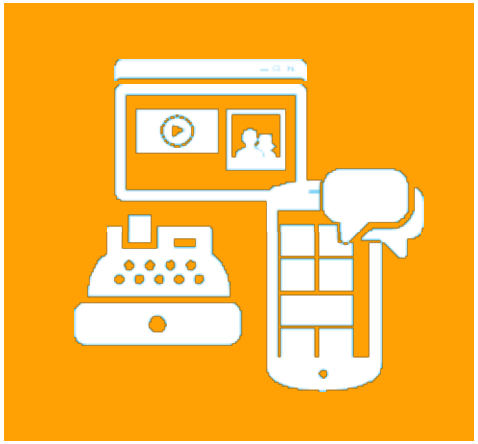
Information Theory

Introducing Big Data

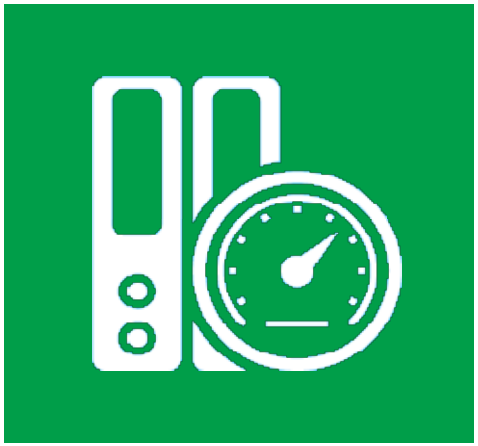
Big data solutions deal with the complexities of:



VOLUME
(Size)



VARIETY
(Structure)



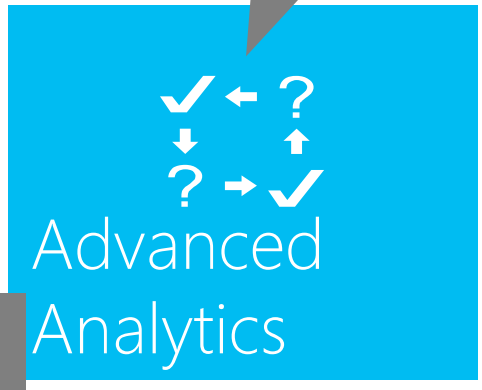
VELOCITY
(Speed)

Big Data: Responding to New Questions

What's the social sentiment of my product?

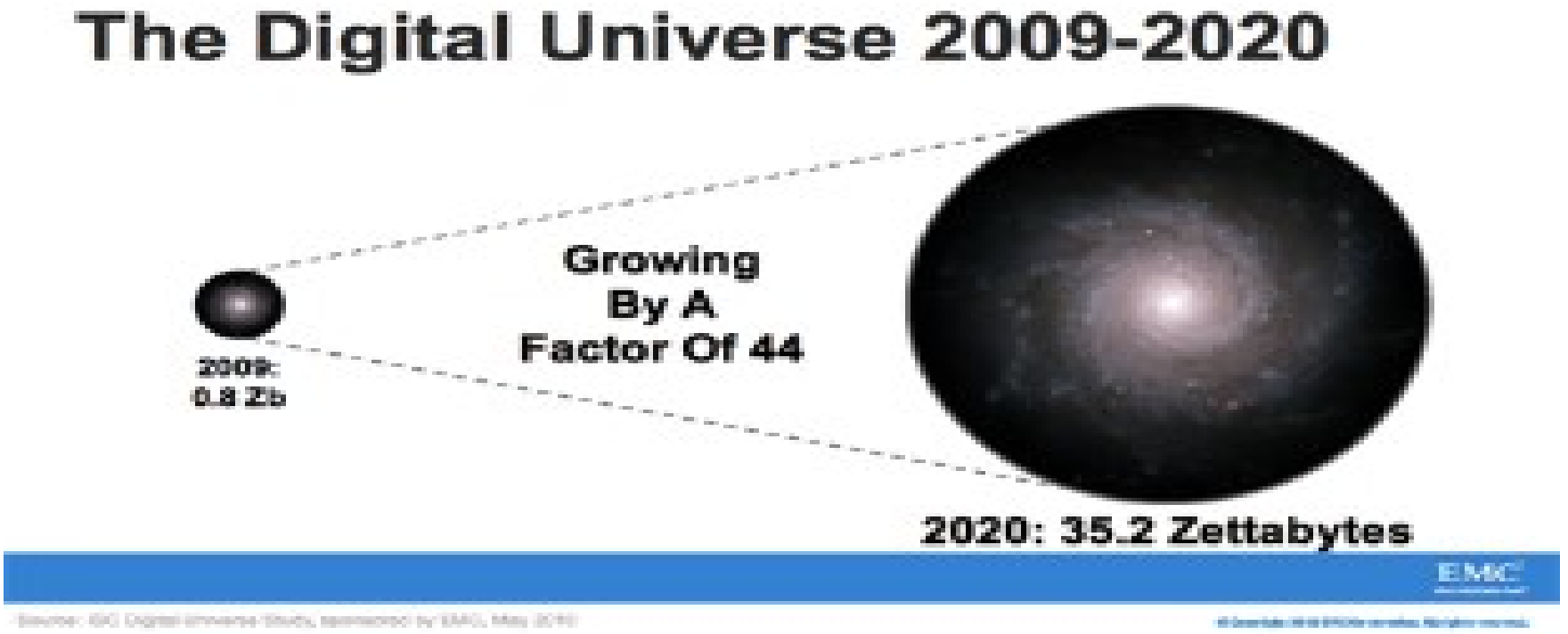
How do I optimize my services based on patterns of weather, traffic, etc.?

How do I better predict future outcomes?

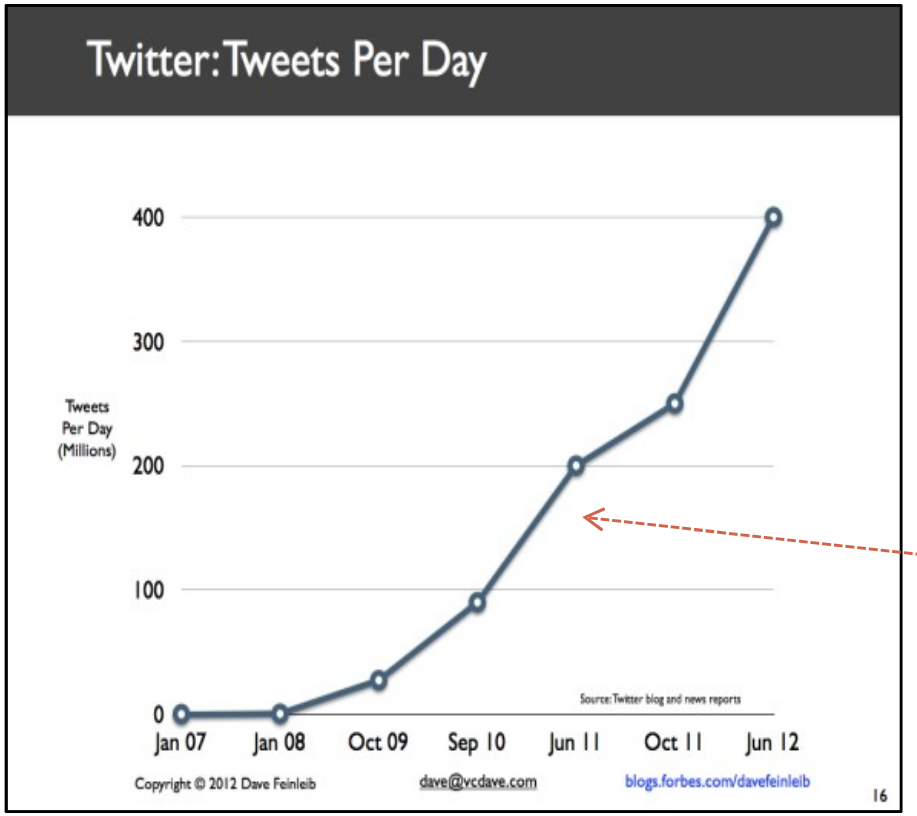


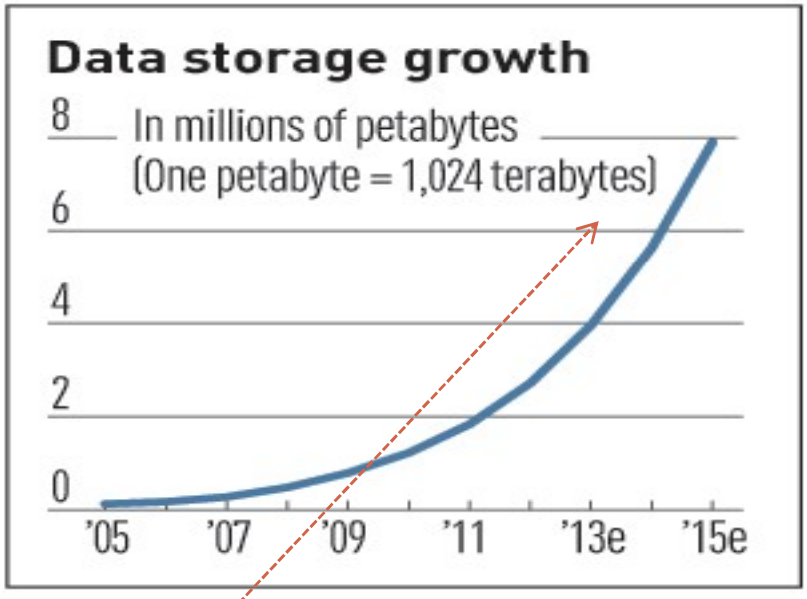
Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially



Volume (Cont...)

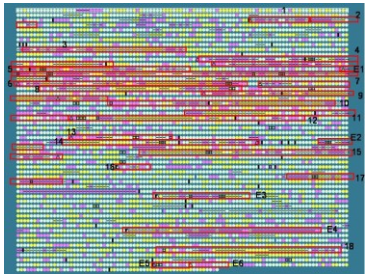


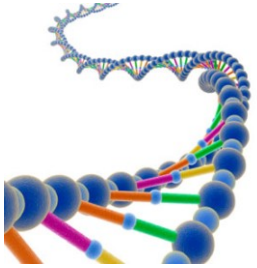


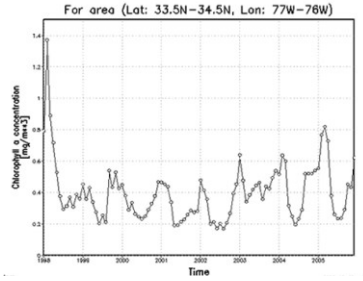
Exponential increase in collected/generated data

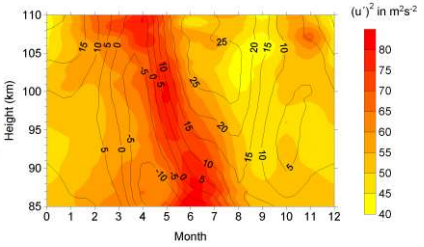
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)



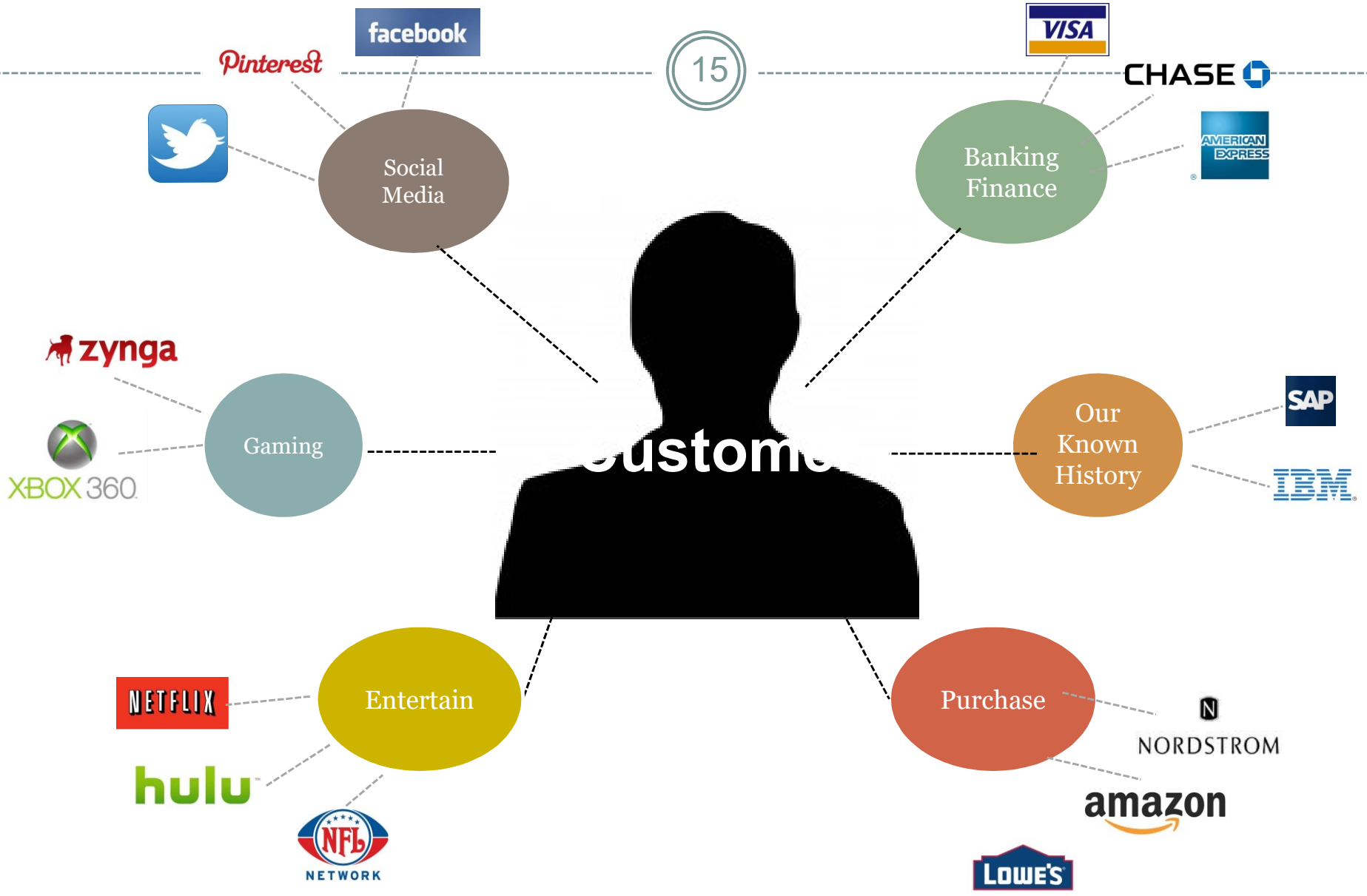






To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



Velocity (Speed)

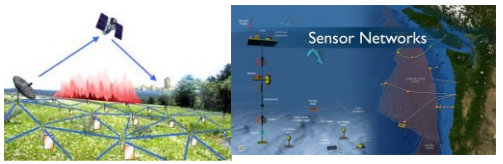
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

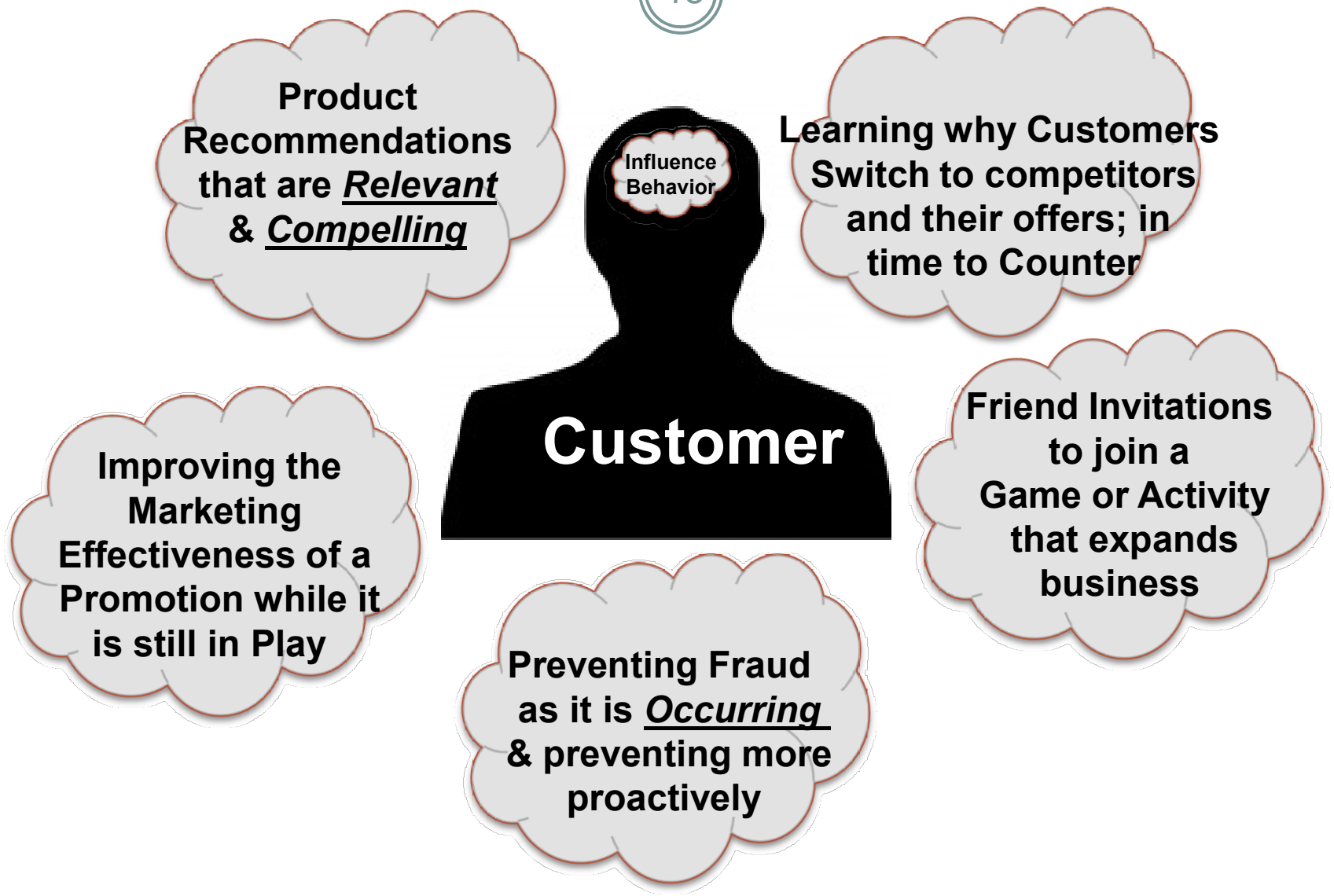
Mobile devices
(tracking all objects all the time)



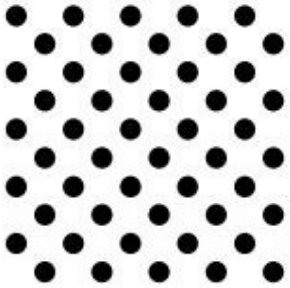
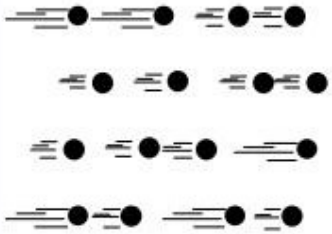
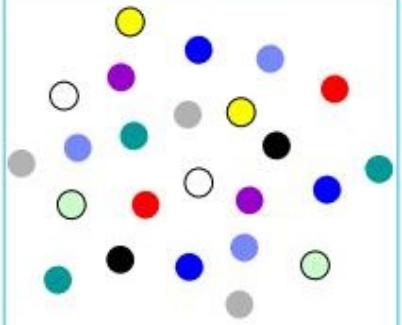
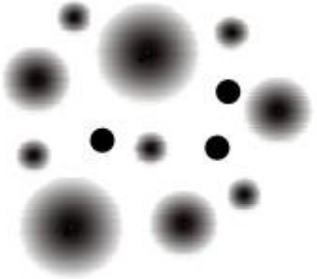
Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

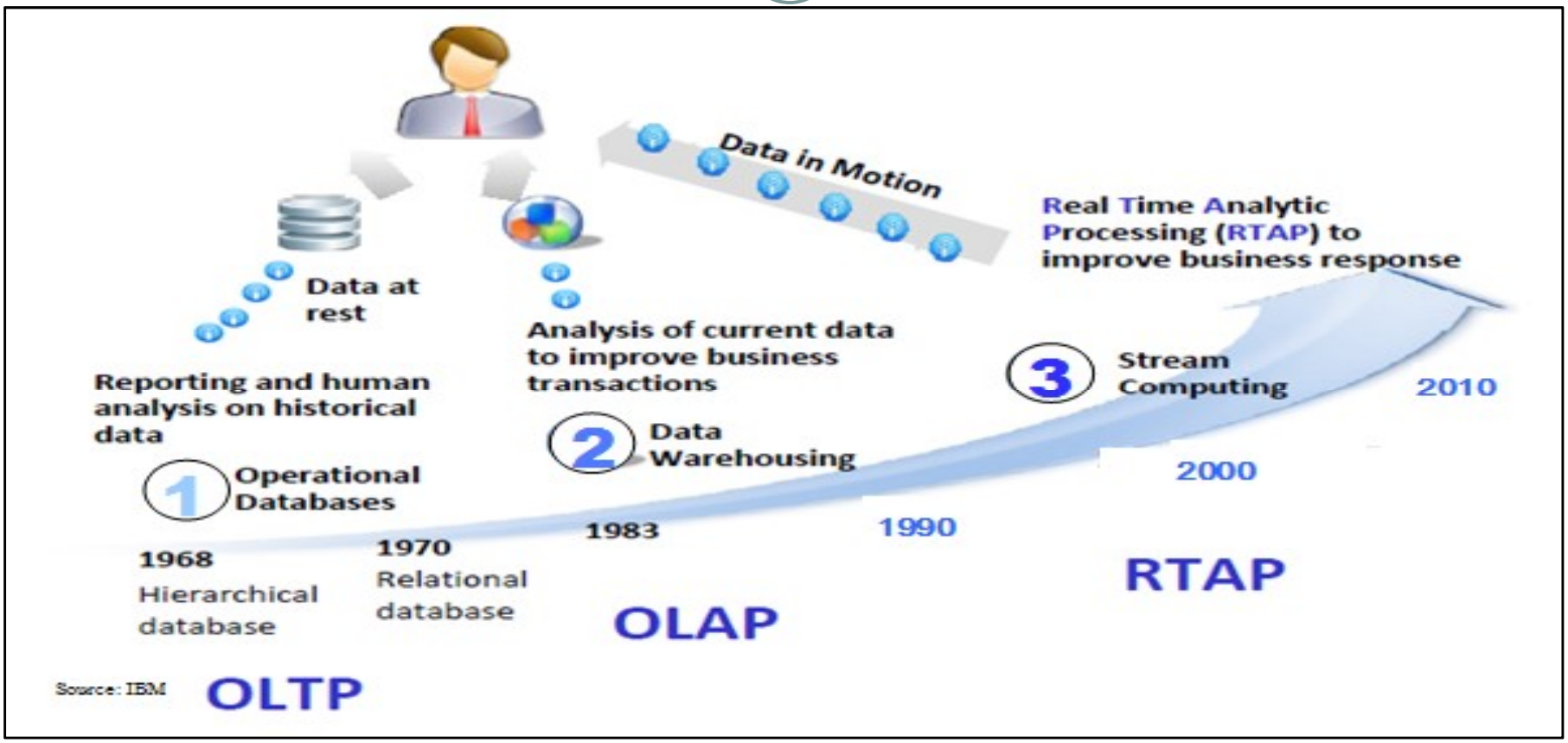
Real-Time Analytics/Decision Requirement



Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

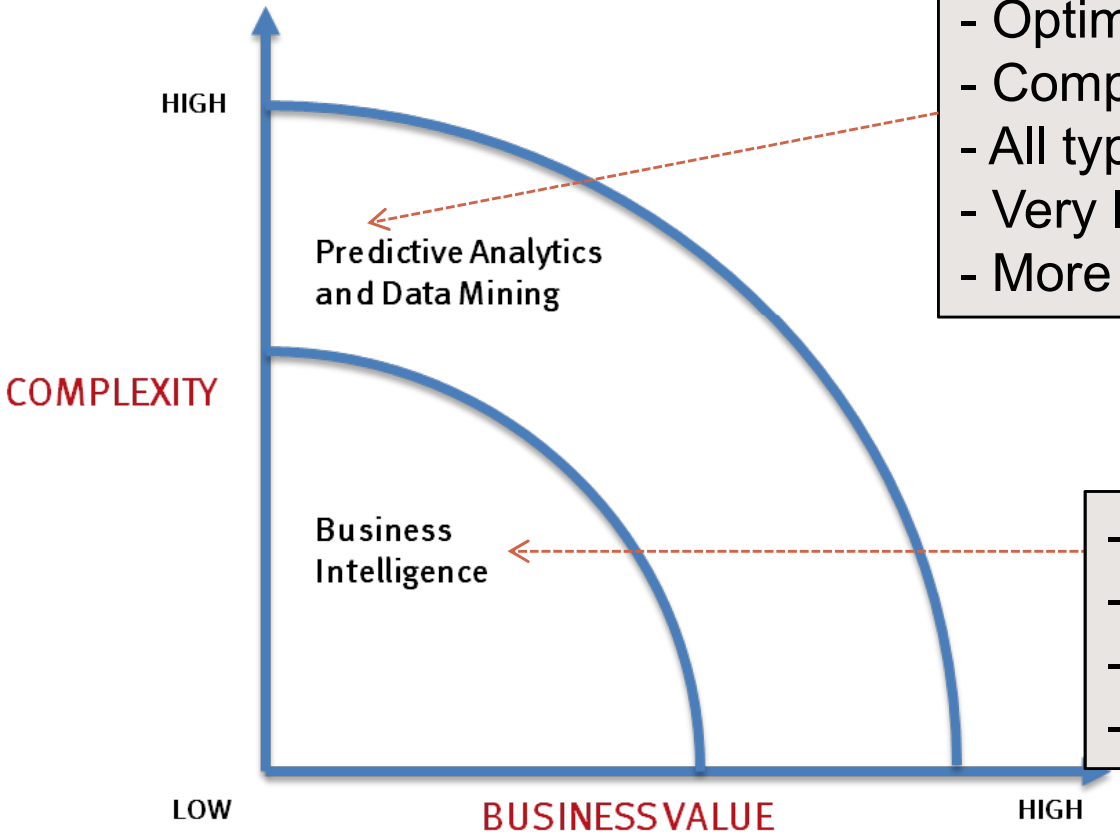
The Model Has Changed…

- **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 Big Data

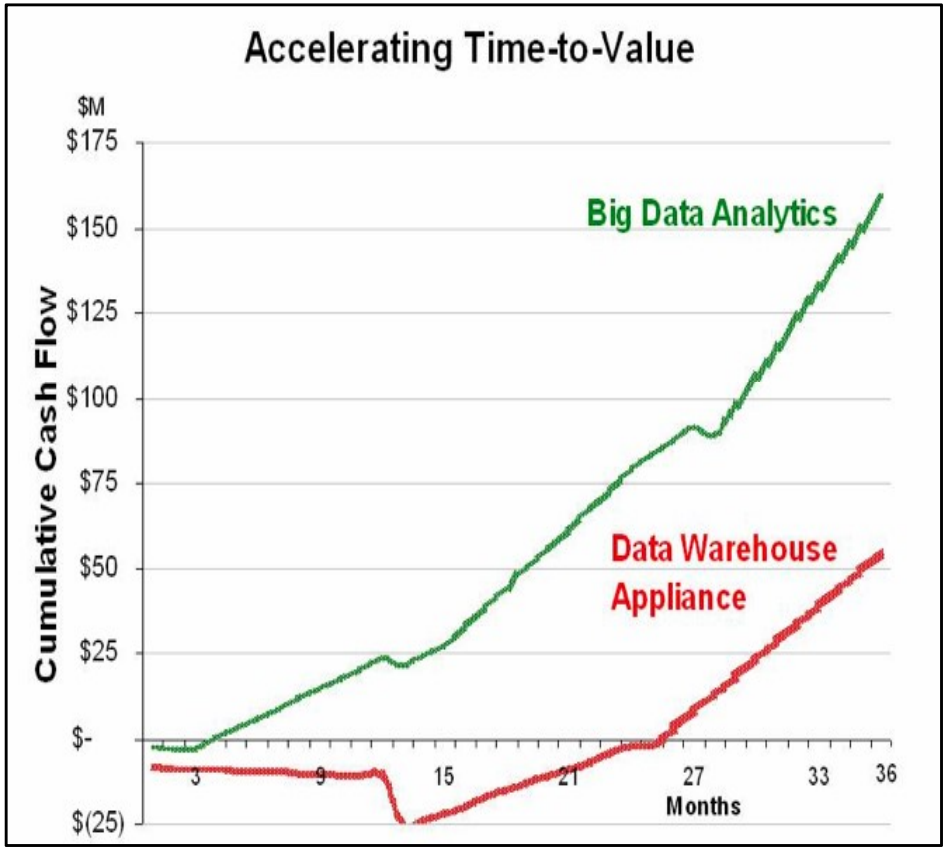


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

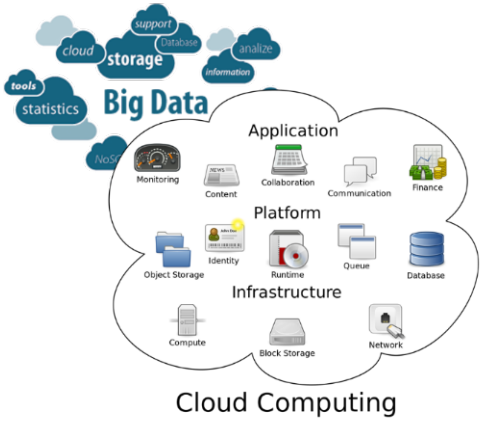
Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Cloud Computing and Big Data

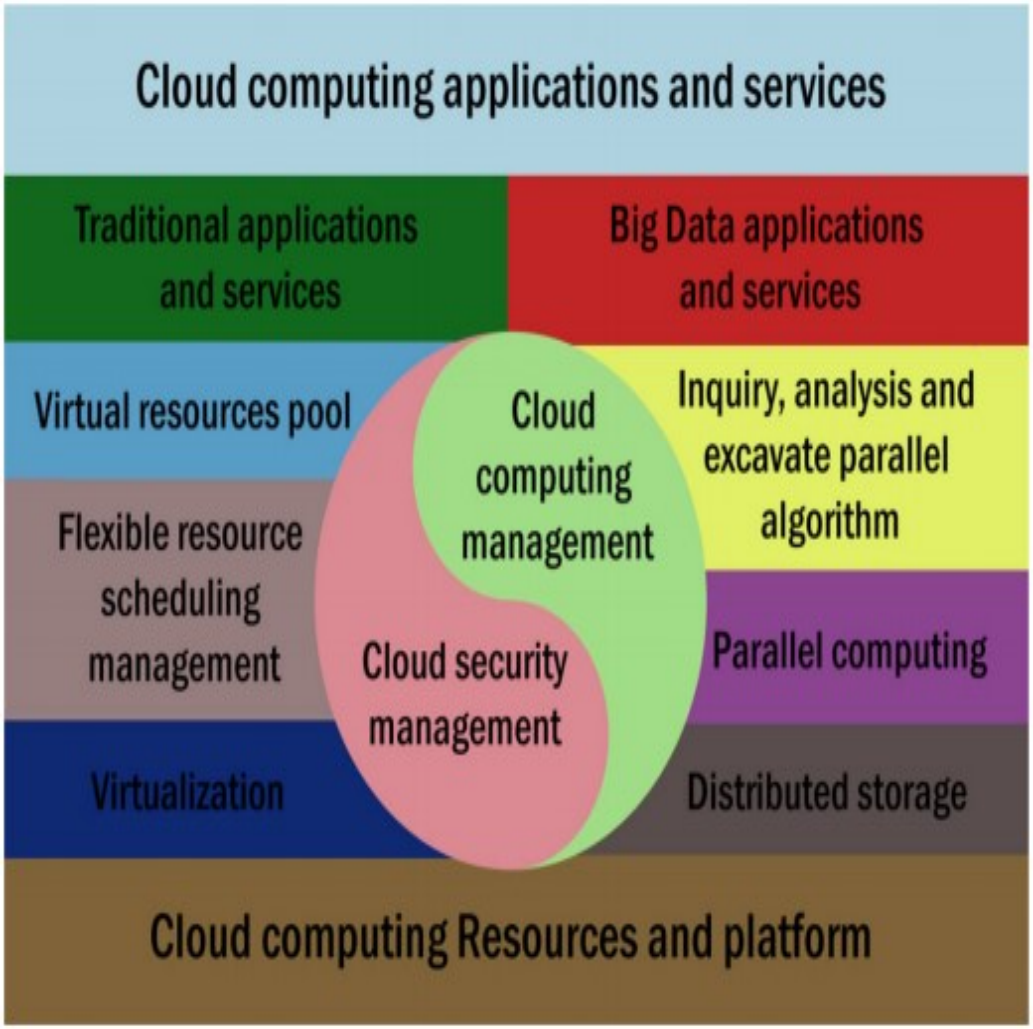
- Big data is data whose volume makes it challenging enough to use from a computational standpoint that tools like cloud is necessary.
- Cloud computing is one way to get enough computing power and storage to handle big data in a manner that doesn't involve investing in infrastructure assets.
- Cloud Computing provides a simple way to access servers, storage, databases and a broad set of application services over the Internet. Cloud Computing providers such as Amazon Web Services etc.



Cloud Computing

Cloud Computing and Big Data (Cont…)

- On the other hand, the emergence of big data also accelerates the development of cloud computing.
- The distributed storage technology based on cloud computing can effectively manage big data;
- The parallel computing capacity by virtue of cloud computing can improve the efficiency of acquisition and analyzing big data.



Key components of cloud computing

Cloud Computing and Big Data (Cont…)

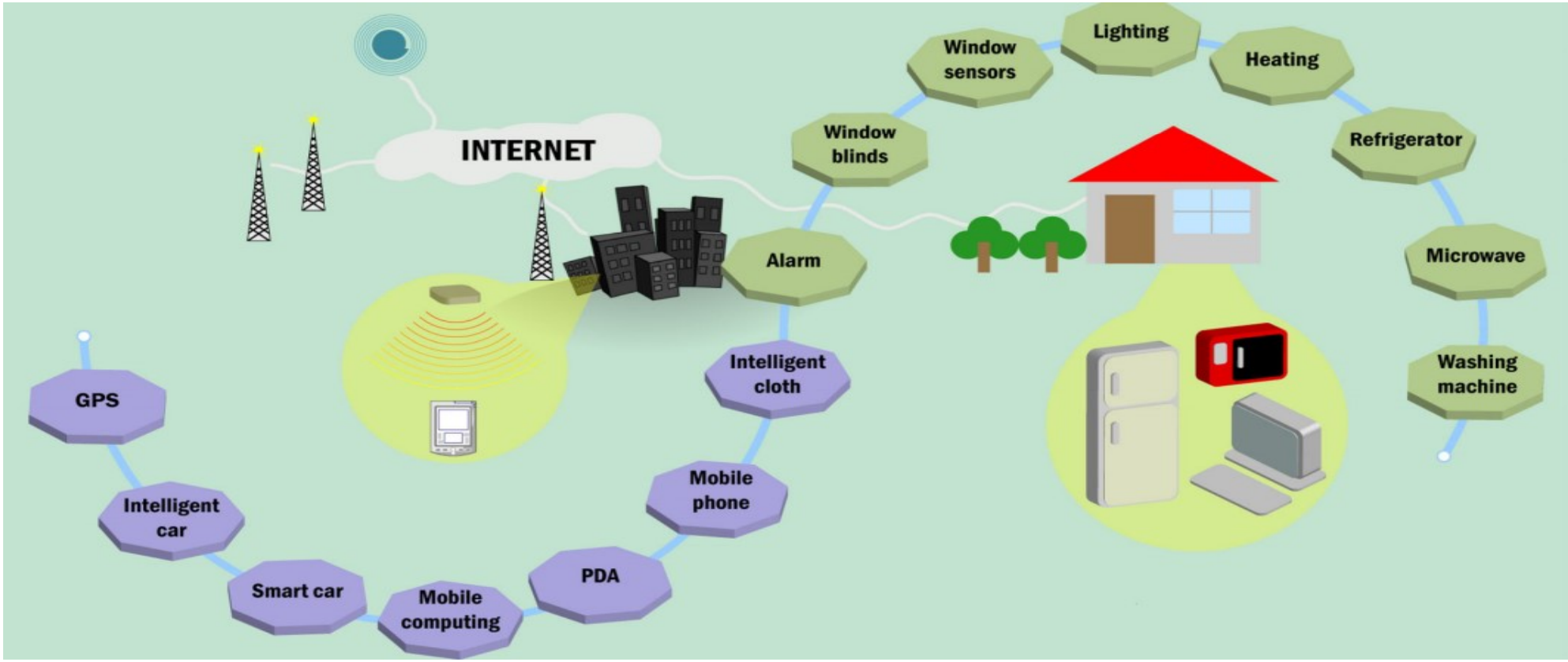
Differences	
Cloud Computing	Big Data
Cloud computing transforms the IT architecture.	Big data influences business decision-making. However, big data depends on cloud computing as the fundamental infrastructure for smooth operation.
Cloud computing is a technology and product targeting an advanced IT solution	Big data is a product focusing on business operations. Since the decision makers may directly feel the pressure from market competition, they must defeat business opponents in more competitive ways.

Cloud Computing and Big Data (Cont…)

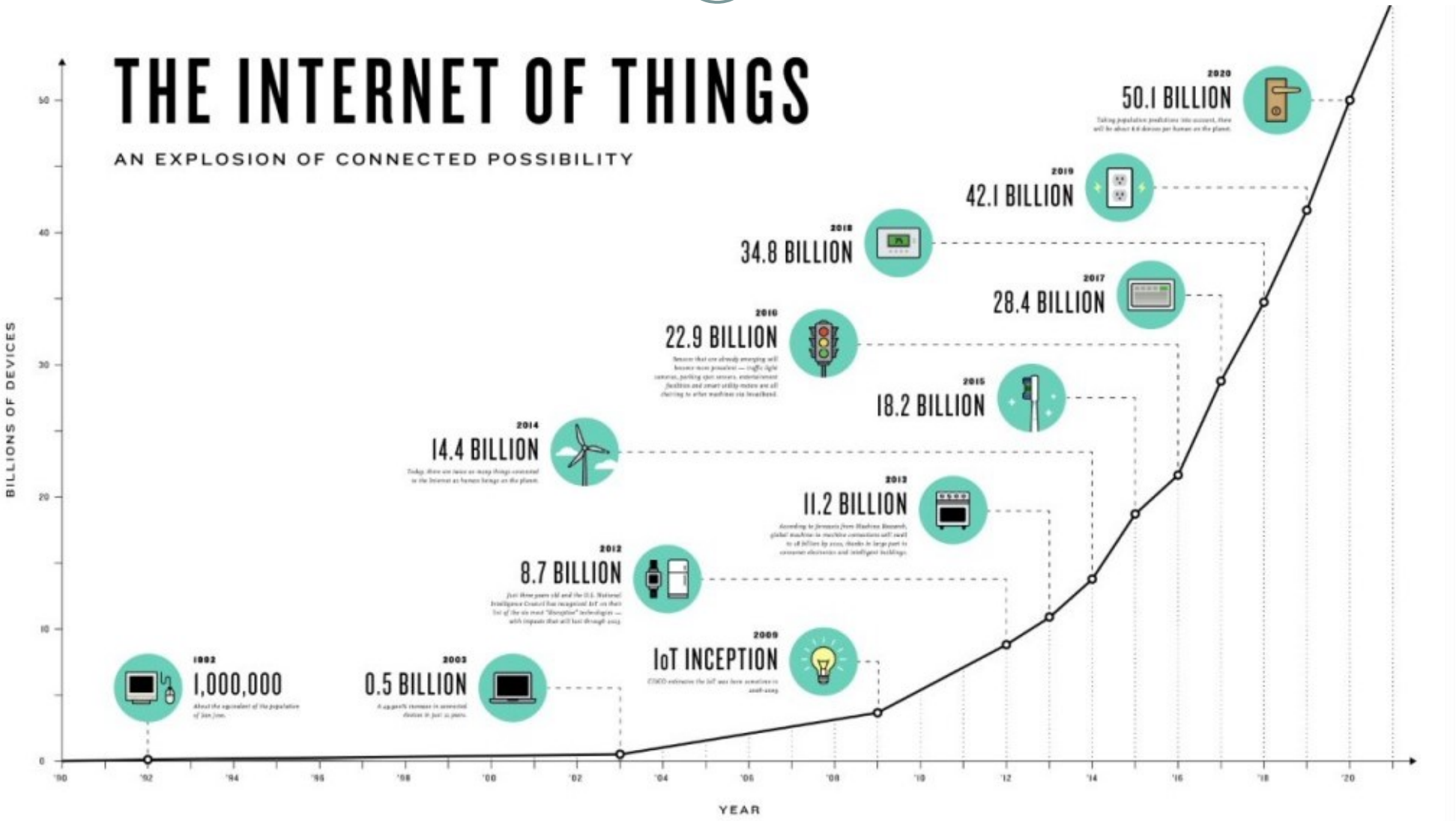
- Benefits of using cloud computing in Big Data for storage and computing the data:
  - Easy to use
  - Low cost
  - Reduce the use of equipment
- As a conclusion, Big Data represents ***content*** and Cloud Computing is ***infrastructure***.

Internet of Things and Big data

- In the IoT paradigm, an enormous amount of networking sensors are embedded into various devices and machines in the real world.
- IoT generates big data which often needs to leverage cloud computing to scale cost effectively.



Internet of Things and Big data (Cont…)



Internet of Things and Big data (Cont…)

- The big data generated by IoT has some classical characteristics:
  - Heterogeneity
  - Variety
  - Unstructured feature
  - Noise
  - High redundancy.
- High growth of generated data by IoT provide the opportunity for the application and development of big data
- The application of big data technology to IoT also accelerates the research advances and business models of IoT.
- Big Data and IoT are inter-dependent.

Value Chain of Big Data

- Four phases of the value chain of big data, i.e.,
  - Data generation
  - Data acquisition
  - Data storage
  - Data analysis

Big Data Generation

- ***Enterprise Data***
  - Main sources of big data: The internal data of enterprises.
  - The internal data of enterprises mainly consists of online trading data and online analysis data, most of which are historically static data
- ***IoT Data***
  - Internet data as an example, huge amount of data in terms of searching entries, Internet forum posts, chatting records, and microblog messages, are generated
- ***Bio-medical data***
  - High-throughput bio-measurement technologies generate a huge number of gene related data.

Big Data Acquisition

- Big data acquisition includes:
  - ***Data collection*** : Acquire raw data from a specific data generation environment using different data collection methods like Log files, Sensing etc.
  - ***Data transmission***: Upon the completion of raw data collection, data will be transferred to a data storage infrastructure for processing and analysis.
  - ***Data pre-processing***: The collected datasets may sometimes include much redundant, noise or useless data, which unnecessarily increases storage space and affects the subsequent data analysis.

Big Data Acquisition: Data pre-processing

- Different techniques are used for the pre-processing steps:
  - ***Integration*** : Combine the data from different sources and it provides users with a uniform view of data. Historically, two methods have been widely recognized: data warehouse and data federation.
  - ***Cleaning*** : Data cleaning is a process to identify inaccurate, incomplete, or unreasonable data, and then modify or delete such data to improve data quality.
  - ***Redundancy Elimination*** : Data redundancy can increase the unnecessary data transmission expense and cause defects on storage systems, e.g., waste of storage space, leading to data inconsistency, reduction of data reliability, and data damage. Therefore, various redundancy reduction methods have been proposed, such as redundancy detection, data filtering, and data compression.

Big data Storage

- Big data storage refers to the storage and management of large-scale datasets while achieving reliability and availability of data accessing
- Different Storage systems are:
  - Massive storage systems
  - Distributed storage systems

Big data Storage

- ***Direct Attached Storage(DAS)***
  - Various hard disks are directly connected with servers
  - Storage devices are peripheral equipments
  - DAS is only suitable to interconnect servers with a small scale.
- ***Network Storage***:
  - It utilizes network to provide users with a union interface for data access and sharing. It classified as:
  - ***Network Attached Storage(NAS)***: It is directly connected to a network through a hub or switch through TCP/IP protocols. In NAS, data is transmitted in the form of files.
  - ***Storage Area Network(SAN) :*** Data storage management is relatively independent within a storage local area network and it achieves a maximum degree of data sharing and data management.

Big data Storage: Distributed Storage system

- To use a distributed system to store massive data, the following factors should be taken into consideration:
  - ***Consistency***: A distributed storage system requires multiple servers to cooperatively store data.
  - ***Availability***: A distributed storage system operates in multiple sets of servers.
  - ***Partition Tolerance:*** The distributed storage will still works well when the network is partitioned because of the link/node failures or temporary congestion.

Big data Storage

- Existing storage mechanisms of big data may be classified into three levels:
  - File systems.
  - Databases.
  - Programming models.

Storage mechanism for big data: *Database technology*

- Traditional relational databases cannot meet the challenges on categories and scales brought about by big data.
- NoSQL databases (i.e., non traditional relational databases) are becoming more popular for big data storage.
- Three main NoSQL databases:
  - ***Key-value databases :*** Data is stored corresponding to key-values. Every key is unique and customers may input queried values according to the keys. Different key-value databases are *Dynamo*, *Voldemort etc.*
  - ***Column-oriented databases:*** It's idea was came from Google's BigTable. The column-oriented databases store and process data according to columns other than rows. Both columns and rows are segmented in multiple nodes to realize expandability.

Storage mechanism for big data: *Database technology*

- ***Document-oriented databases:*** Document storage can support more complex data forms. Documents do not follow strict modes. In addition, key-value pairs can still be saved. Three important representatives of document storage systems, i.e.,
  - ***MongoDB***: MongoDB stores documents as Binary JSON (BSON) objects, which is similar to object. Every document has an ID field as the primary key.
  - ***SimpleDB*** : Data in SimpleDB is organized into various domains in which data may be stored, acquired, and queried.
  - ***CouchDB***: Data in CouchDB is organized into documents consisting of fields named by keys/names and values, which are stored and accessed as JSON objects. Every document is provided with a unique identifier.

Parallel Programming models for NoSQL

- ***MapReduce***:
  - Computing model only has two functions, i.e., Map and Reduce, both of which are programmed by users.
  - The Map function processes input key-value pairs and generates intermediate key-value pairs.
  - MapReduce will combine all the intermediate values related to the same key and transmit them to the Reduce function,
  - Reduce functions compress the value set into a smaller set.
- ***Dryad***:
  - A general-purpose distributed execution engine for processing parallel applications.
  - The operational structure of Dryad is a directed acyclic graph.
  - Dryad allows vertexes to use any amount of input and output data, while MapReduce supports only one input and output set.

Traditional Data Analysis

- Cluster Analysis
- Factor Analysis
- Correlation Analysis
- Regression Analysis
- Statistical Analysis
- Data Mining Algorithms

Big data Analysis

- According to timeliness requirements, big data analysis can be classified into two:
  - Real-time analysis
    - Used in E-commerce and finance
  - Off-line analysis
    - Used for applications without high requirements on response time, e.g., machine learning, statistical analysis, and recommendation algorithms.

Tools for big data mining and analysis

- ***R*:**
  - An open source programming language and software environment, is designed for data mining/analysis and visualization.
- ***Excel*:**
  - It provides powerful data processing and statistical analysis capabilities.
- ***Weka*:**
  - *It* is a free and open-source machine learning and data mining software written in Java.
  - Weka provides such functions as data processing, feature selection, classification, regression, clustering, association rule, and visualization, etc.

Thank you