Introduction to Big Data

Slides from Prof. Dr. Thoai Nam, Lecturer: Nguyen Quang Hung

High Performance Computing Lab (HPC Lab)

Faculty of Computer Science and Technology

HCMC University of Technology

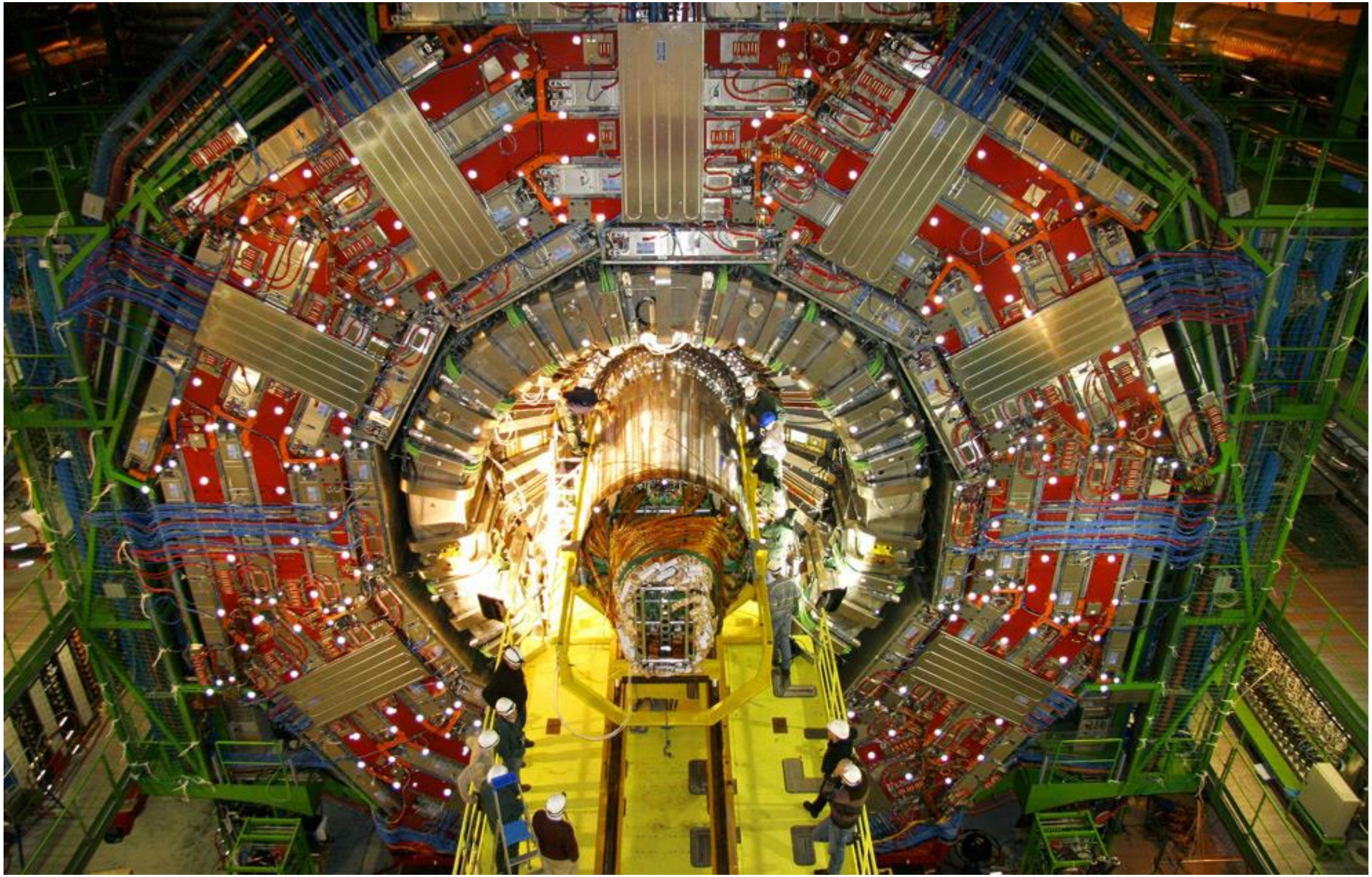
What’s Big Data?

“Data **too large & complex** to be effectively handled by standard database technologies currently founded in most organizations”

“Data whose **scale**, **diversity** and **complexity** require new **architectures**, **techniques**, **algorithms** and **analytics** to manage it and extract value and hidden knowledge from it”

Data sources?

CERN’s Large Hydron Collider (LHC) generates 15 PB a year



The Earthscope

- The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data.
- It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more.
- (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI)

Annual budget: $25,000,000
Construction cost: $197,000,000
Staff: 110
Physical size: 3.8 million square miles
Scientific utility: 10
WIIFY: 10
Wow factor: 10

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs **of log data every day**

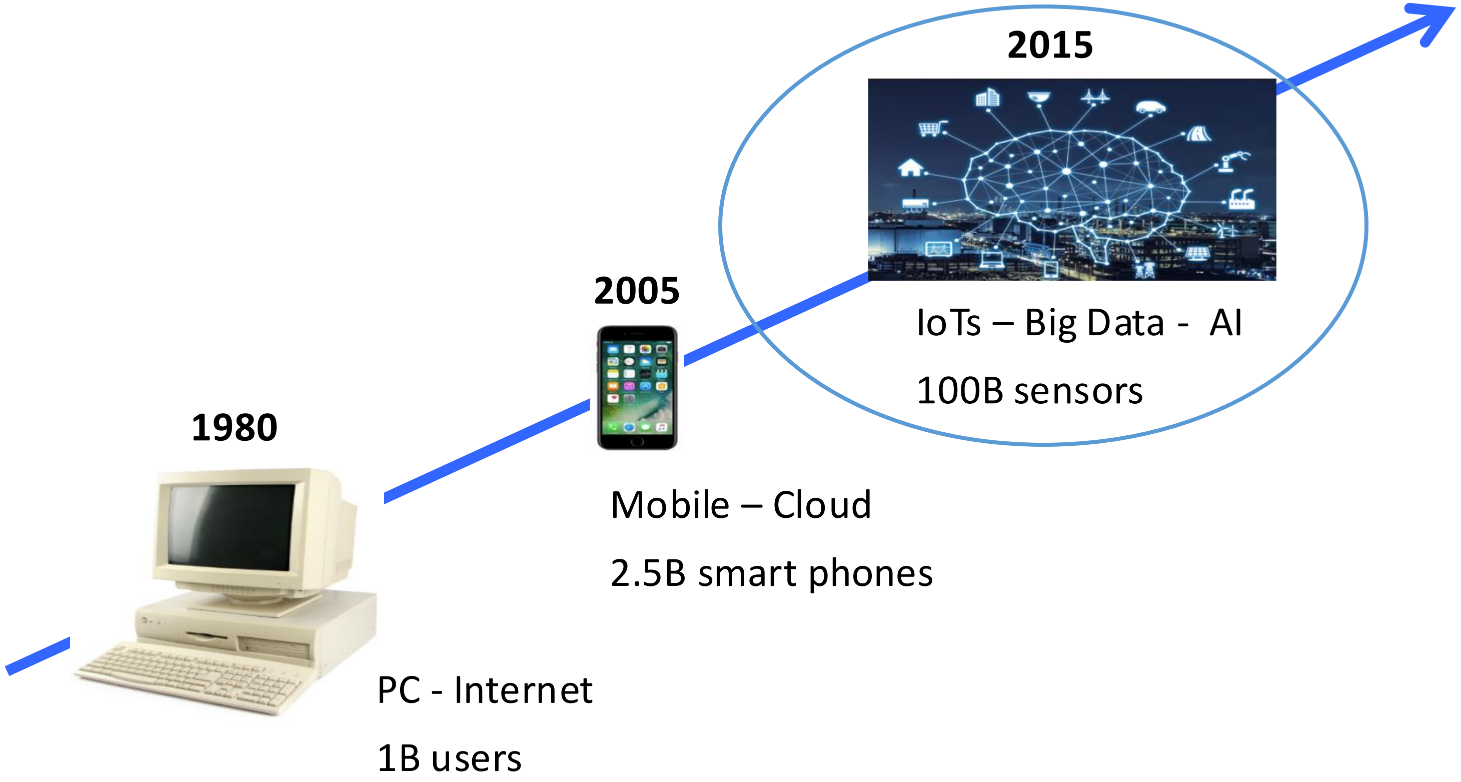
30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

2+ billion people on the Web by end 2011

76 million smart meters in 2009… 200M by 2014



1980

PC - Internet

1B users

2005

Mobile – Cloud

2.5B smart phones

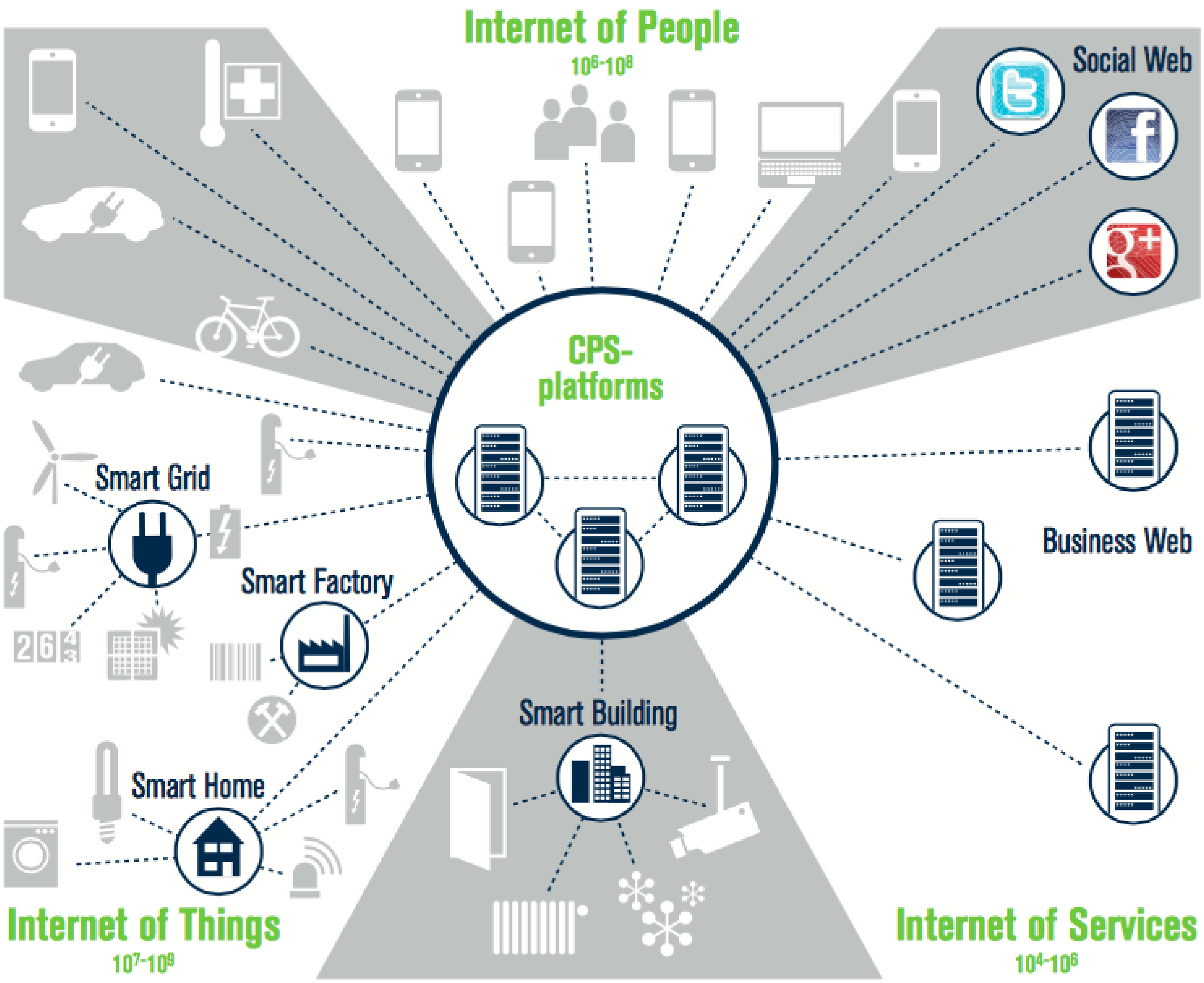
2015

IoTs – Big Data - AI

100B sensors

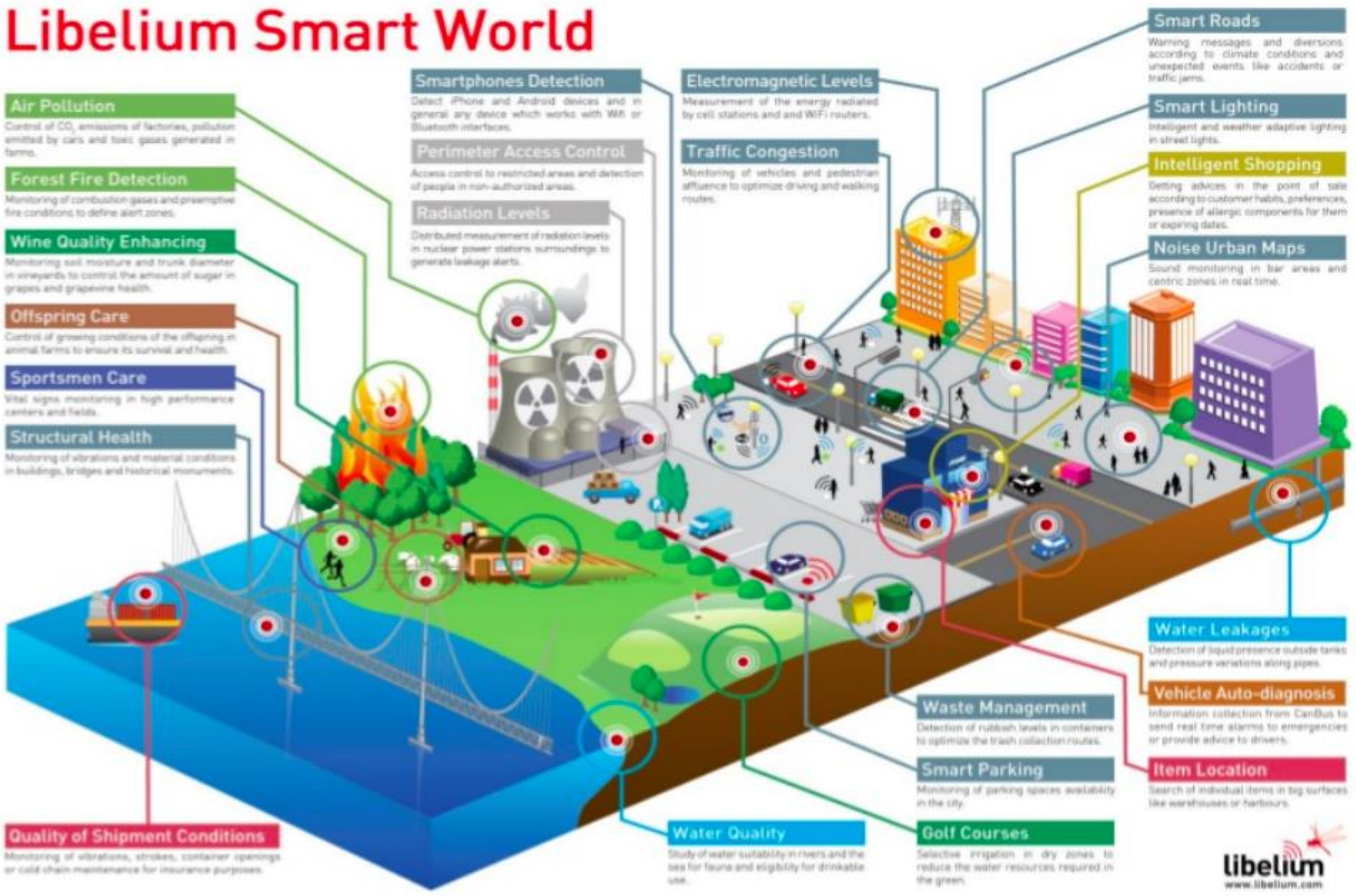
IoT and Services

Figure 4:
The Internet of Things and Services – Networking people, objects and systems



Source: Bosch Software Innovations 2012

Smart cities



Libelium Smart World

Air Pollution
Control of CO$_2$ emissions of factories, pollution emitted by cars and toxic gases generated in farms.

Forest Fire Detection
Monitoring of combustion gases and preemptive fire conditions to define alert zones.

Wine Quality Enhancing
Monitoring soil moisture and trunk diameter in vineyards to control the amount of sugar in grapes and grapevine health.

Offspring Care
Control of growing conditions of the offspring in animal farms to ensure its survival and health.

Sportsmen Care
Vital signs monitoring in high performance centers and fields.

Structural Health
Monitoring of vibrations and material conditions in buildings, bridges and historical monuments.

Smartphones Detection
Detect iPhone and Android devices and in general any device which works with Wifi or Bluetooth interfaces.

Perimeter Access Control
Access control to restricted areas and detection of people in non-authorized areas.

Radiation Levels
Distributed measurement of radiation levels in nuclear power stations surroundings to generate leakage alerts.

Electromagnetic Levels
Measurement of the energy radiated by cell stations and and WiFi routers.

Traffic Congestion
Monitoring of vehicles and pedestrian affluence to optimize driving and walking routes.

Smart Roads
Warning messages and diversions according to climate conditions and unexpected events like accidents or traffic jams.

Smart Lighting
Intelligent and weather adaptive lighting in street lights.

Intelligent Shopping
Getting advices in the point of sale according to customer habits, preferences, presence of allergic components for them or expiring dates.

Noise Urban Maps
Sound monitoring in bar areas and centric zones in real time.

Water Leakages
Detection of liquid presence outside tanks and pressure variations along pipes.

Vehicle Auto-diagnosis
Information collection from CanBus to send real time alarms to emergencies or provide advice to drivers.

Item Location
Search of individual items in big surfaces like warehouses or harbours.

Waste Management
Detection of rubbish levels in containers to optimize the trash collection routes.

Smart Parking
Monitoring of parking spaces availability in the city.

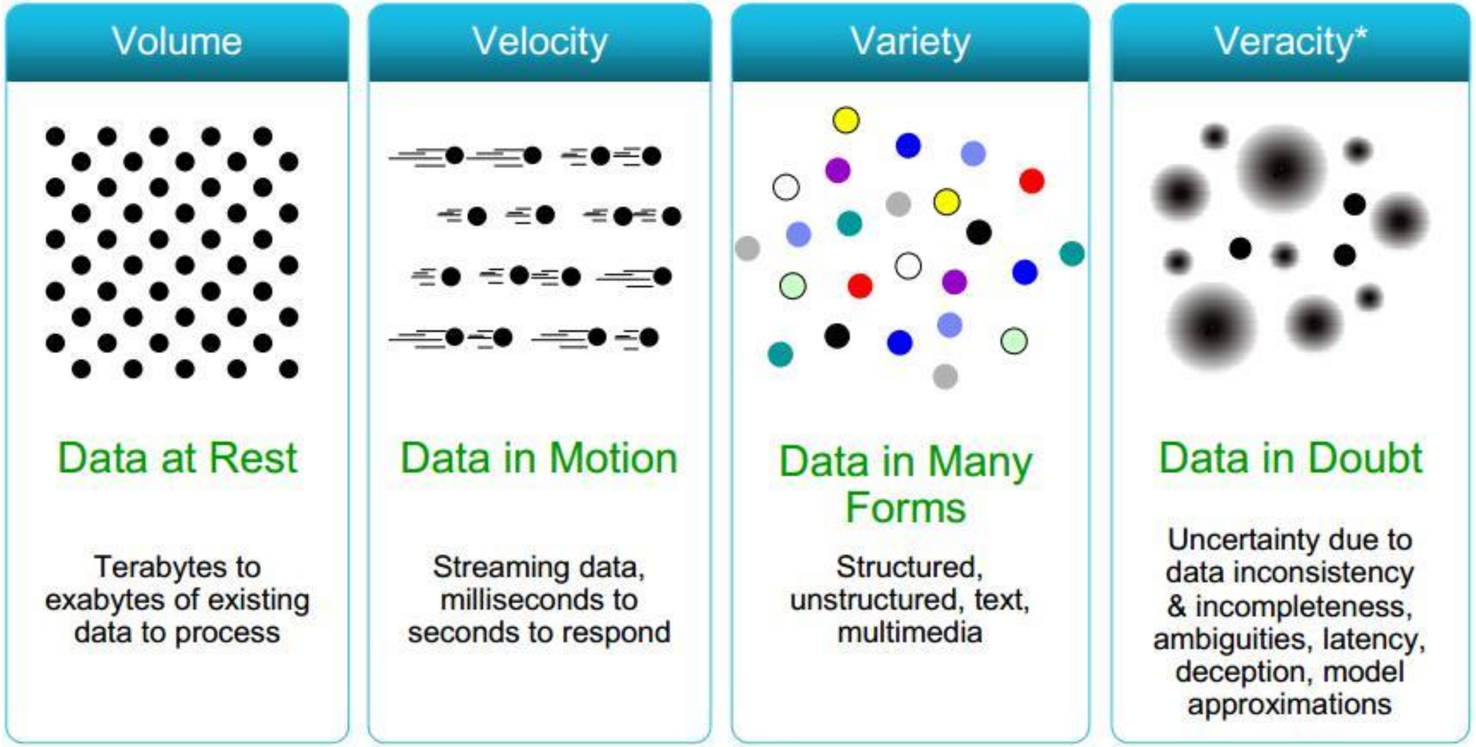
Golf Courses
Selective irrigation in dry zones to reduce the water resources required in the green.

Water Quality
Study of water suitability in rivers and the sea for fauna and eligibility for drinkable use.

Quality of Shipment Conditions
Monitoring of vibrations, strokes, container openings or cold chain maintenance for insurance purposes.

http://www.libelium.com/libelium-smart-world-infographic-smart-cities-internet-of-things/

Some make it 4V's



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Volume

- Data volume increases exponentially over time
- 33 ZB in 2018 to 175 ZB in 2025

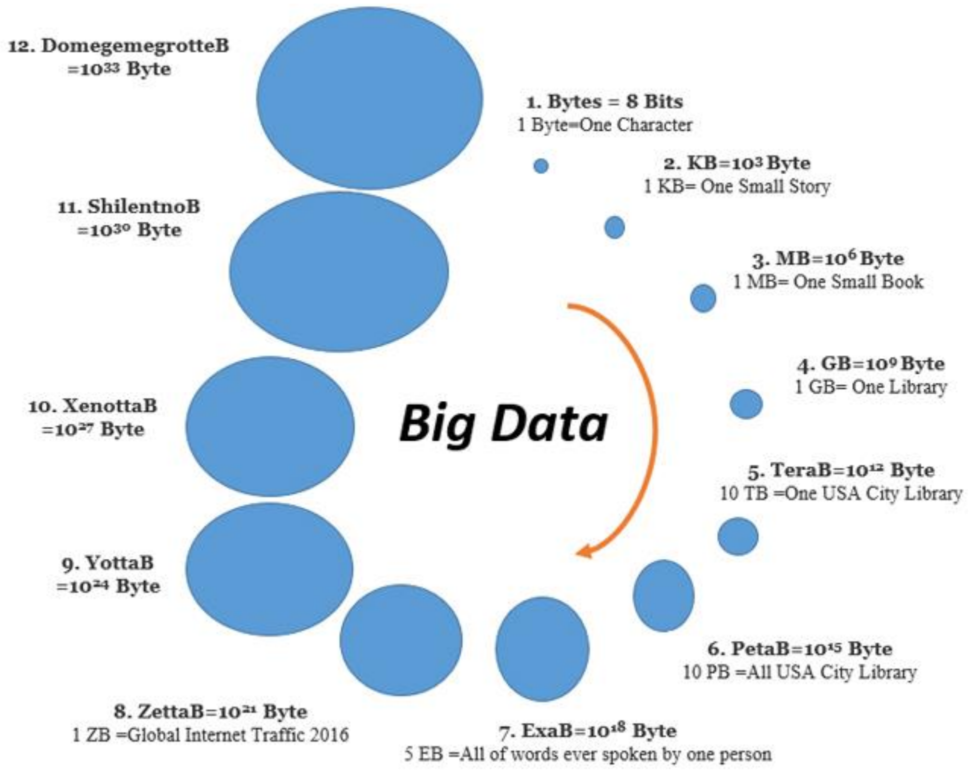
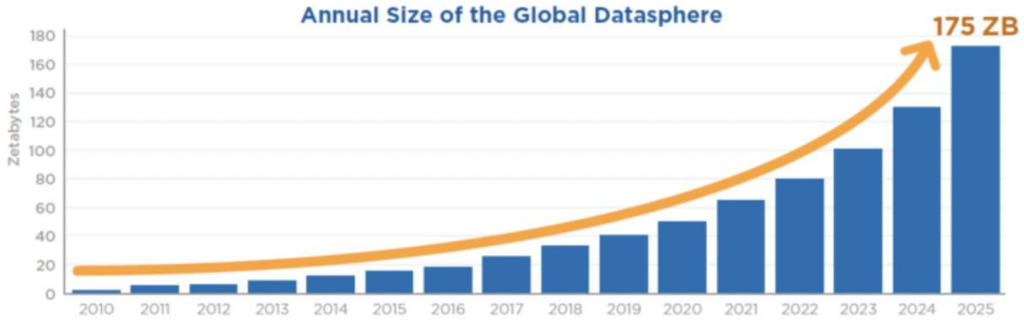


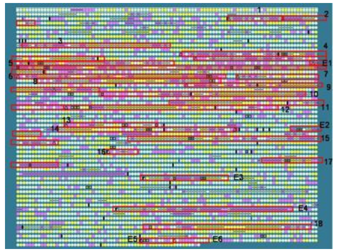
Figure 1 – Annual Size of the Global Datasphere

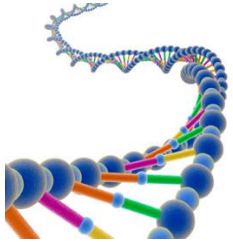


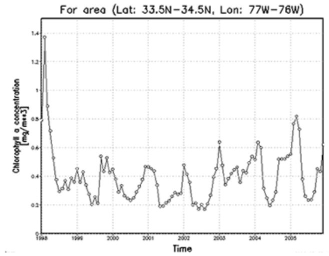
Variety (Complexity)

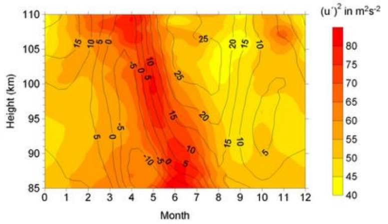
- Various formats, types and structures
  - Numerical data, image data, audio, video, text, time series
  - Relational Data (Tables/Transaction/Legacy Data)
  - Text Data (Web)
  - Semi-structured Data (XML)
  - Graph Data
    - Social Network, Semantic Web (RDF), …
- A single application can be generating/collecting many types of data
  - Heterogeneous data
  - Complex data integration problem.

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

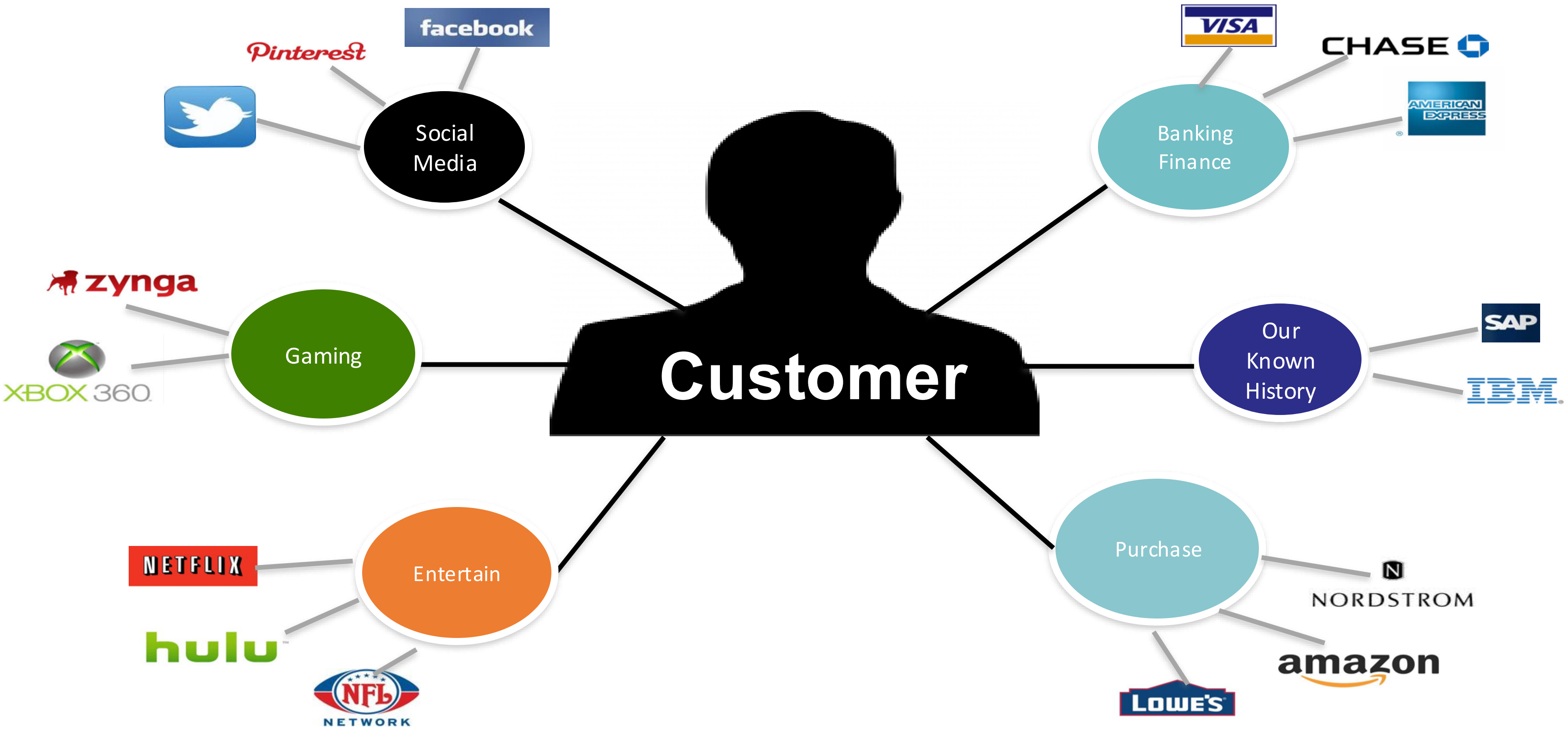






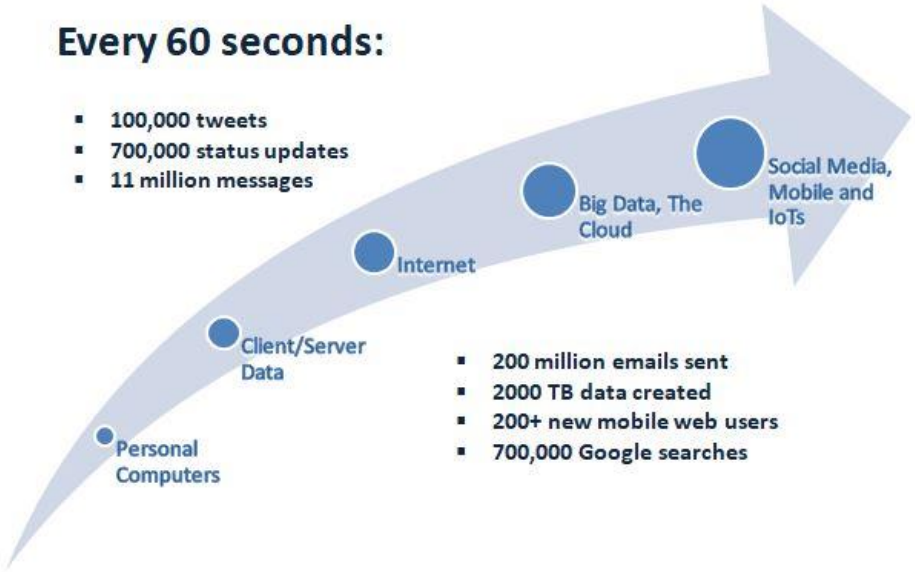


A single view to the customer



a Single Customer View *'is an aggregated, consistent and holistic representation of the data known by an organization about its customers'*

Velocity (Speed)



Every 60 seconds:

- 100,000 tweets
- 700,000 status updates
- 11 million messages

- 200 million emails sent
- 2000 TB data created
- 200+ new mobile web users
- 700,000 Google searches

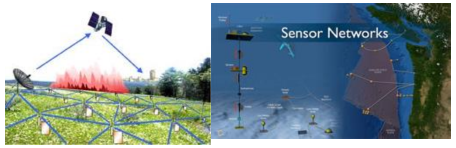
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:**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

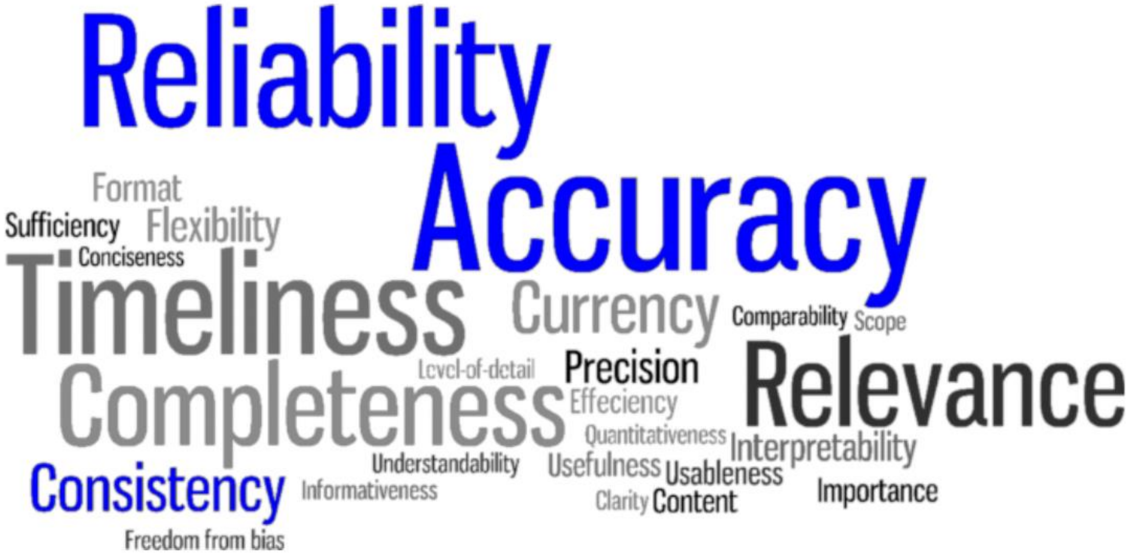


Sensor technology and networks
(measuring all kinds of data)

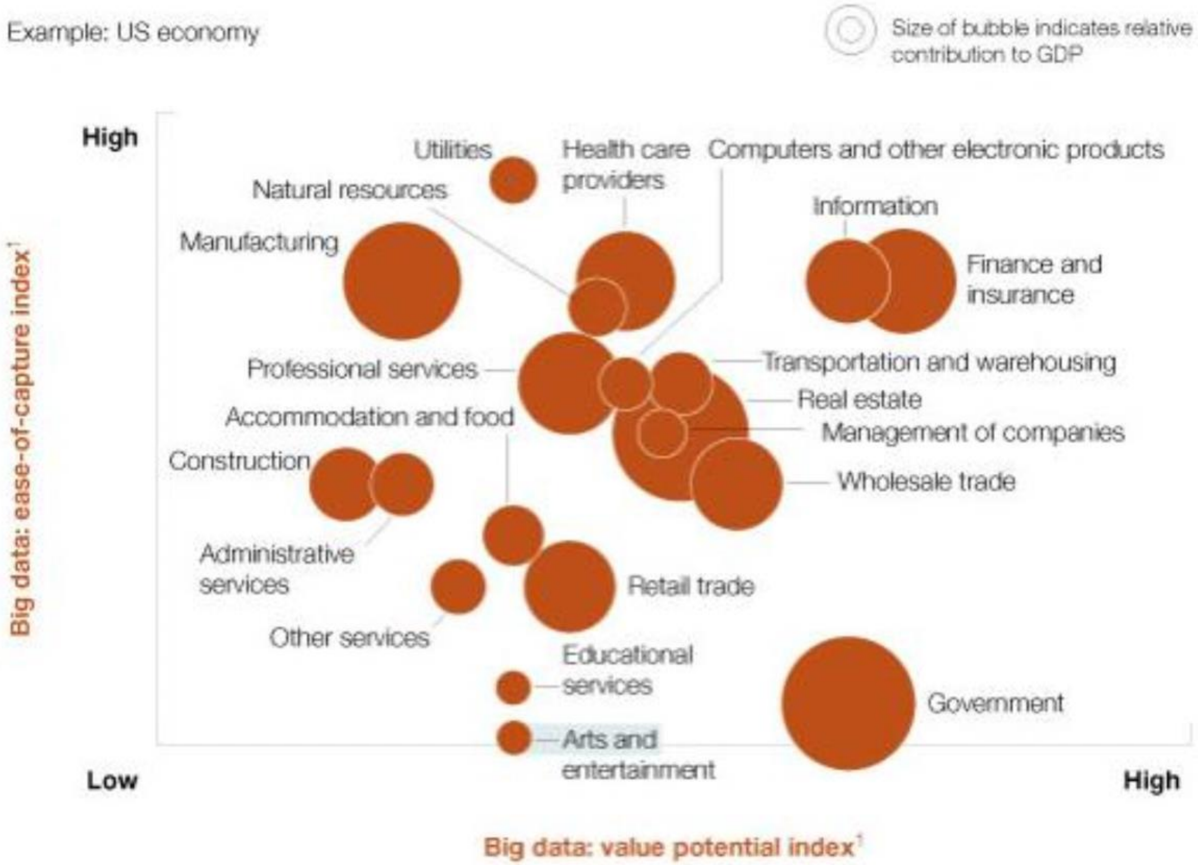
- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion.

Veracity

Data quality



Value



- ➢ Translate data into business advantage.

Answering tough questions

- **Problem**
  - sales for lollipops are going down
- **Data**
  - all sales data by customer, region, time, ...
- **Information**
  - lollipops bought by people older than 25
  - (but eaten by people younger than 10)
- **Knowledge**
  - moms believe: lollipops = bad teeth
- **Value**
  - dentists advertise your lollipops

Why is this difficult?

- You need more data than your data warehouse
  - you need more data that you have
  - logs, Twitter feeds, blogs, customer surveys, ...
- You need to ask the right questions
  - data alone is silent
- You need technology and organization that help you concentrate on asking the right questions.

What is Big Data?

- Three alternative perspectives
  - Philosophical
  - Business
  - Technical
- (Ultimately, it is a buzz word for everybody.)

Philosophical

- What is more valuable, if you had to pick one?
  - experience or intelligence?
- Traditional (computer) science: **logic!** [intelligence]
  - understand the problem, build model / algorithm
  - answer question from implementation of model
- New science: **statistics!** [experience]
  - collect data
  - answer question from data (what did others do?)

Data Science, 4th Paradigm

- New approach to do science
  - Step 1: Collect data
  - Step 2: Generate Hypotheses
  - Step 3: Validate Hypotheses
  - Step4: (Goto Step 1 or 2)
- Why is this a good approach?
  - it can be automated: no thinking, less error
- Why is this a bad approach?
  - how do you debug without a ground truth?

Is bigger = smarter?

- Yes!
  - tolerate errors
  - discover the long tail and corner cases
  - machine learning works much better
- But!
  - more data, more error (e.g., semantic heterogeneity)
  - with enough data you can prove anything
  - still need humans to ask right questions

What is Big Data?

- Business Perspective
  - it is a new business model
- People pay with data
  - e.g. Facebook, Google, Twitter:
    - use service, give data
    - Google sells your data to advertisers • (you pay advertisers indirectly)
  - e.g, Amazone
    - pay service + give data
    - sells data and uses data to improve service

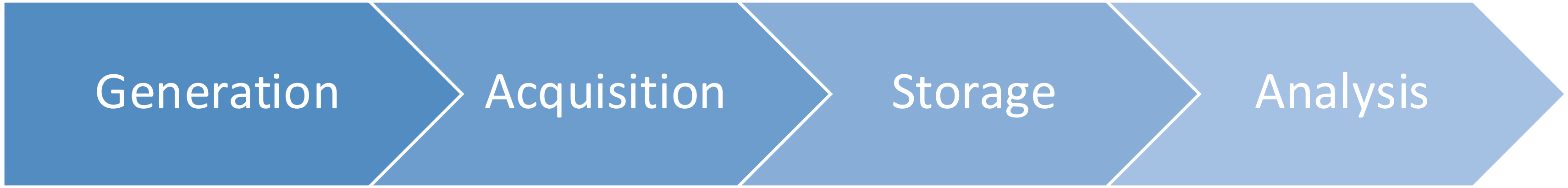
Business Perspective

- Bank
  - keeps your money securely (kind of...)
  - puts your money at work (lends it to others), interest
  - you keep ownership of money and take it when needed
- Databank
  - keeps your data securely (kind of...)
  - puts your data at work: interest or better service
  - (you keep ownership of data: hopefully to come)

Technical Perspective (?)

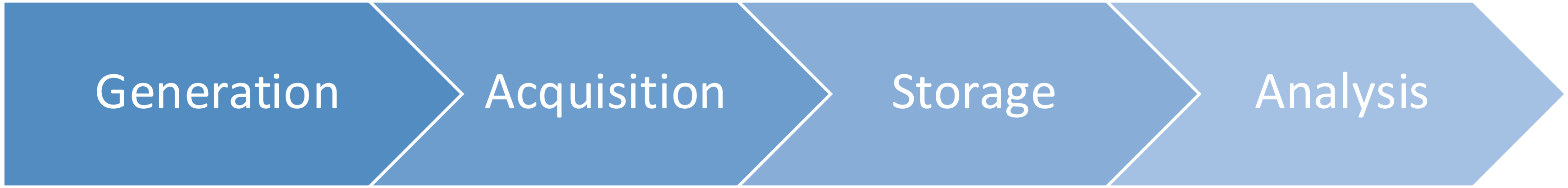
- You collect all data
  - the more the better -> statistical relevance, long tail
  - keeping all is cheaper than deciding what to keep
- You decide independently what to do with data
  - run experiments on data when question arises
- Huge difference to traditional information systems
  - design upfront what data to keep and why!!!
  - (e.g., waterfall model of software engineering!)

Big data value chain (1)



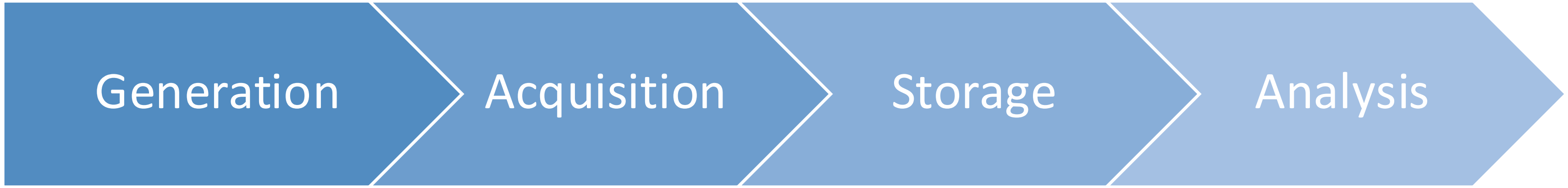
- Generation
  - Passive recording
    - Typical structured data
    - Bank trading transactions, shopping records, government sector archives
  - Active generation
    - Semi-structured or unstructured data
    - User-generated content, e.g., social networks
  - Automatic production
    - Location-aware, context-dependent, highly mobile data
    - Sensor-based Internet-enabled devices.

Big data value chain (2)



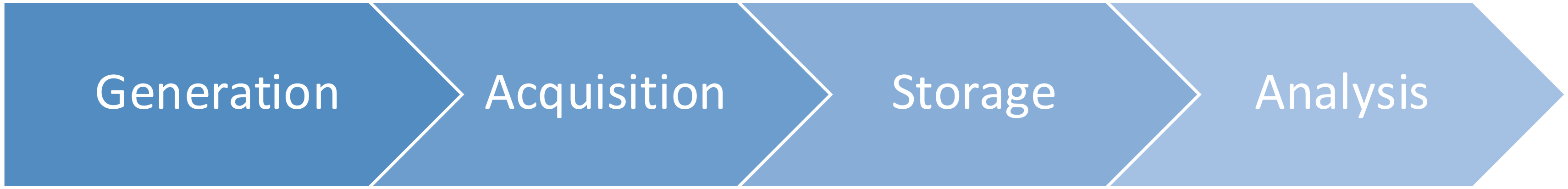
- Acquisition
  - Collection
    - Pull-based, e.g., web crawler
    - Push-based, e.g., video surveillance, click stream
  - Transmission
    - Transfer to data center over high capacity links
  - Preprocessing
    - Integration, cleaning, redundancy elimination.

Big data value chain (3)



- Storage
  - Storage infrastructure
    - Storage technology, e.g., HDD, SSD
    - Networking architecture, e.g., DAS, NAS, SAN
  - Data management
    - File systems (HDFS), key-value stores (Memcached), column-oriented databases (Cassandra), document databases (MongoDB)
  - Programming models
    - Map-Reduce, stream processing, graph processing.

Big data value chain (4)



- Analysis
  - Objectives
    - Descriptive analytics, predictive analytics, prescriptive analytics
  - Methods
    - Statistical analysis, data mining, text mining, network and graph data mining
    - Clustering, classification and regression, association analysis
  - Diverse domains call for customized techniques.

Big data challenges

- Technology and infrastructure
  - New architectures, programming paradigms and techniques are needed
- Data management and analysis
  - New emphasis on “data”
  - Data science.

The bottleneck

- Processors process data
- Hard drives store data
- We need to transfer data from the disk to the processor.

The solution

- Transfer the processing power to the data
- Multiple distributed disks
  - Each one holding a portion of a large dataset
- Process in parallel different file portions from different disks.