Lecture 2: Introduction to the Big Data

Big Data System Design

Table of Contents

Part 1

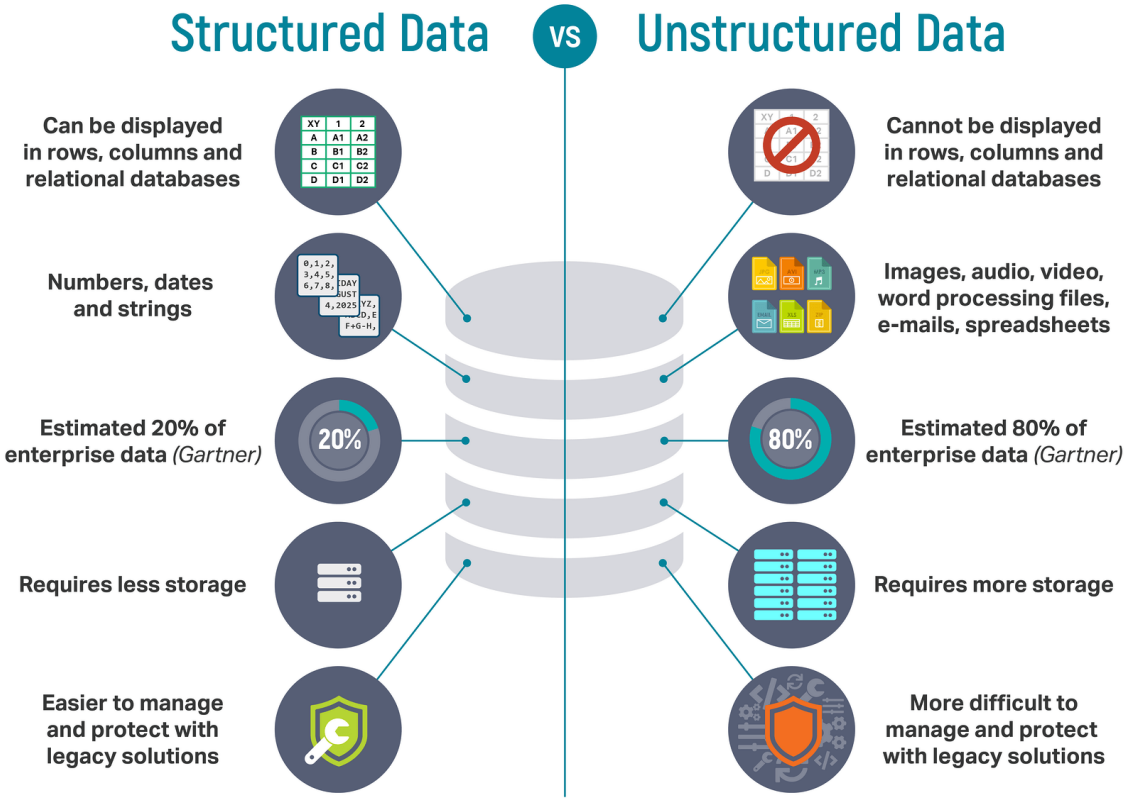
INTRODUCTION TO BIG DATA

Intro to Big Data

- What is data?
  - Since birth, we are surrounded with data
  - From the advent of written language, human observations have been recorded
  - The advent of computer technologies in 1950s, data most commonly refers to information that is transmitted or stored electronically
  - The electronic sensors has additionally contributed to the volume and richness of recorded data

Intro to Big Data

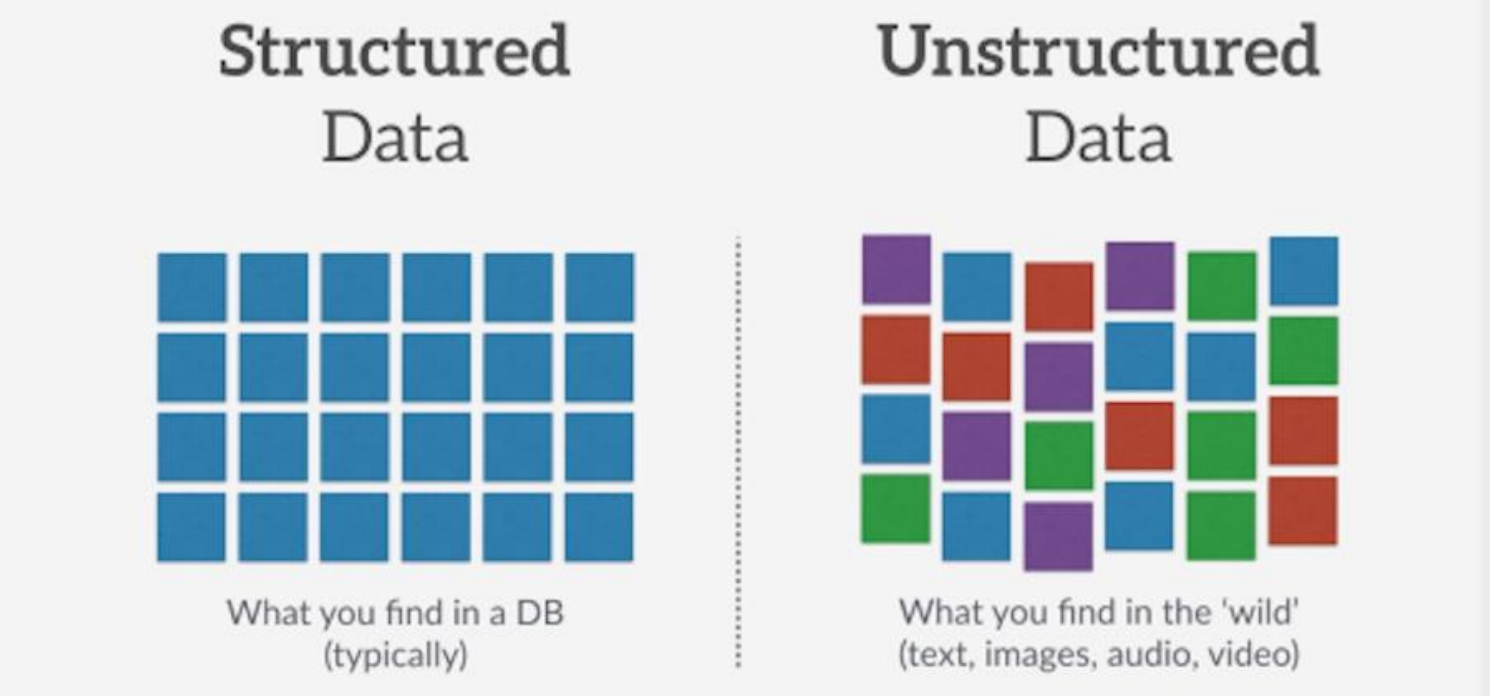
- Types of data
  - Structured data vs. unstructured data



Source: https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

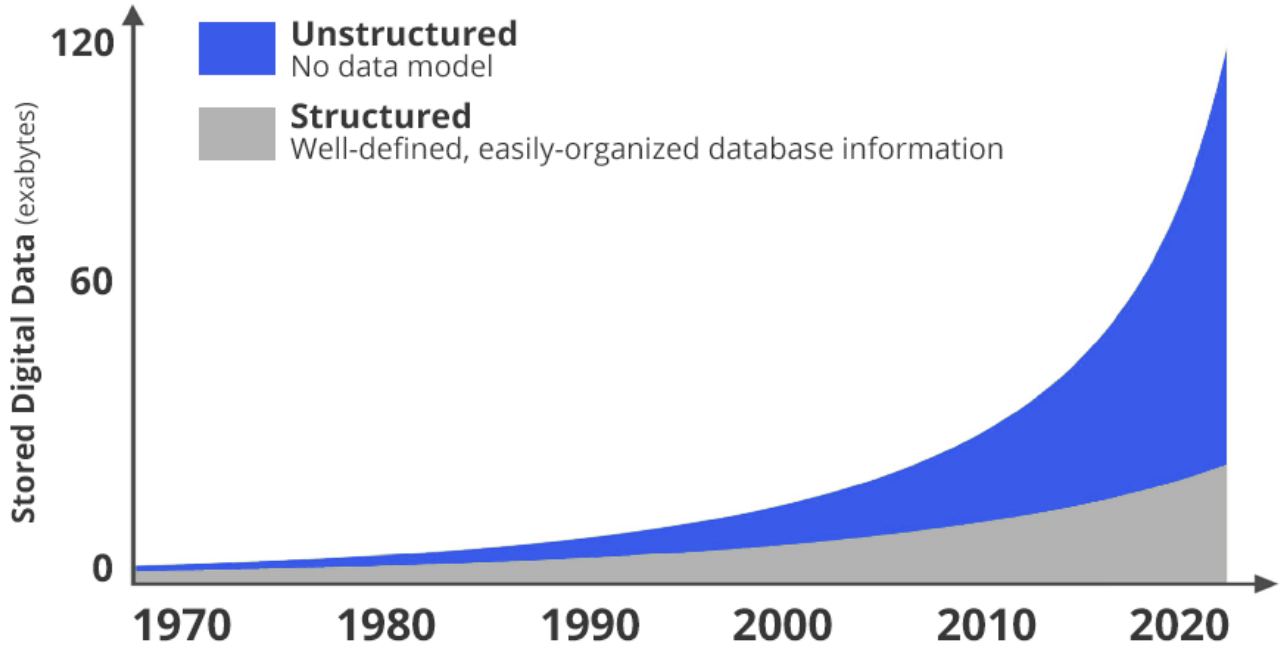
Intro to Big Data

- Types of data
  - Structured data vs. unstructured data



Intro to Big Data

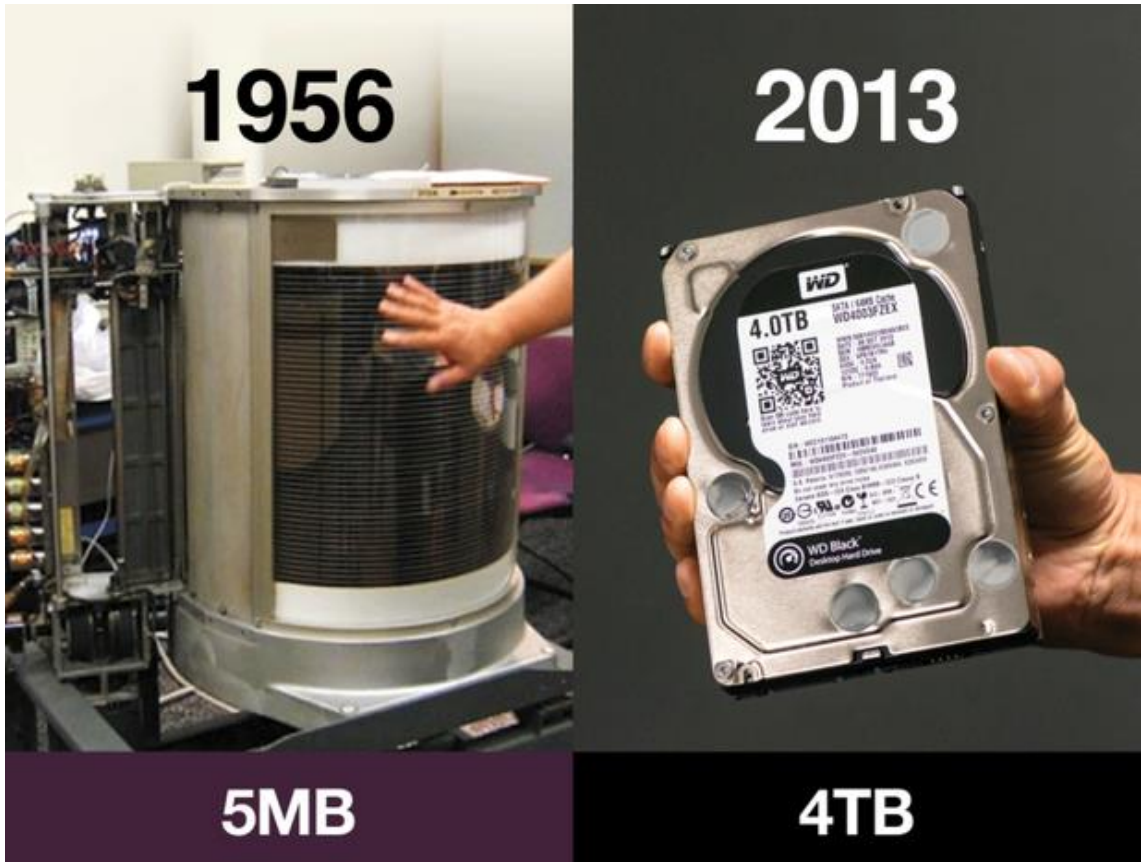
- Types of data
  - Structured data vs. unstructured data



Source: https://www.m-files.com/blog/what-is-structured-data-vs-unstructured-data/

Intro to Big Data

- Reason for growth
  - Rapid advance in computer hardware



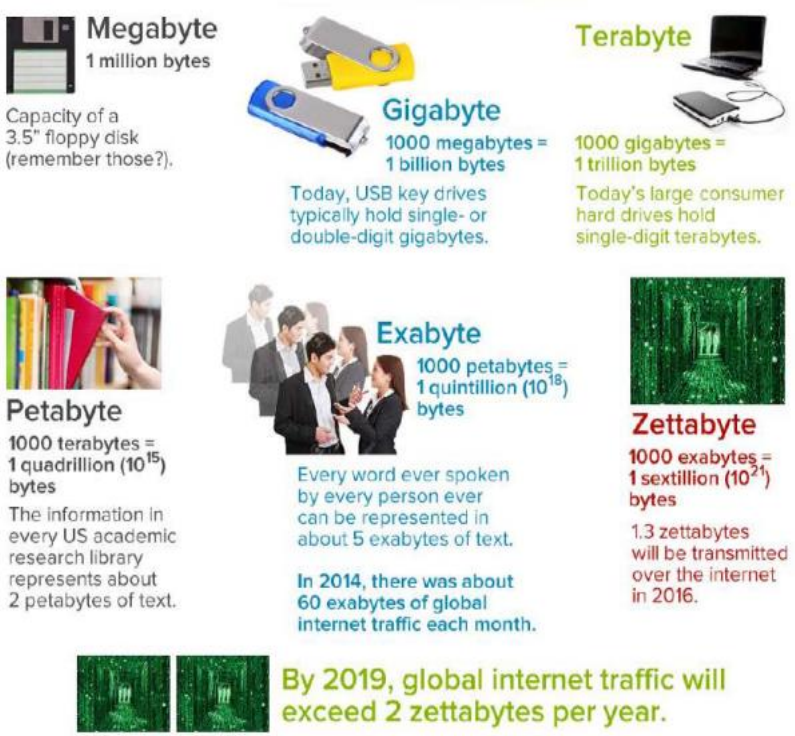
Source: https://www.computerworld.com/article/2473980/data-storage-solutions-143723-storage-now-and-then.html

Intro to Big Data

- Reason for growth
  - Rapid advance in social networking

Intro to Big Data

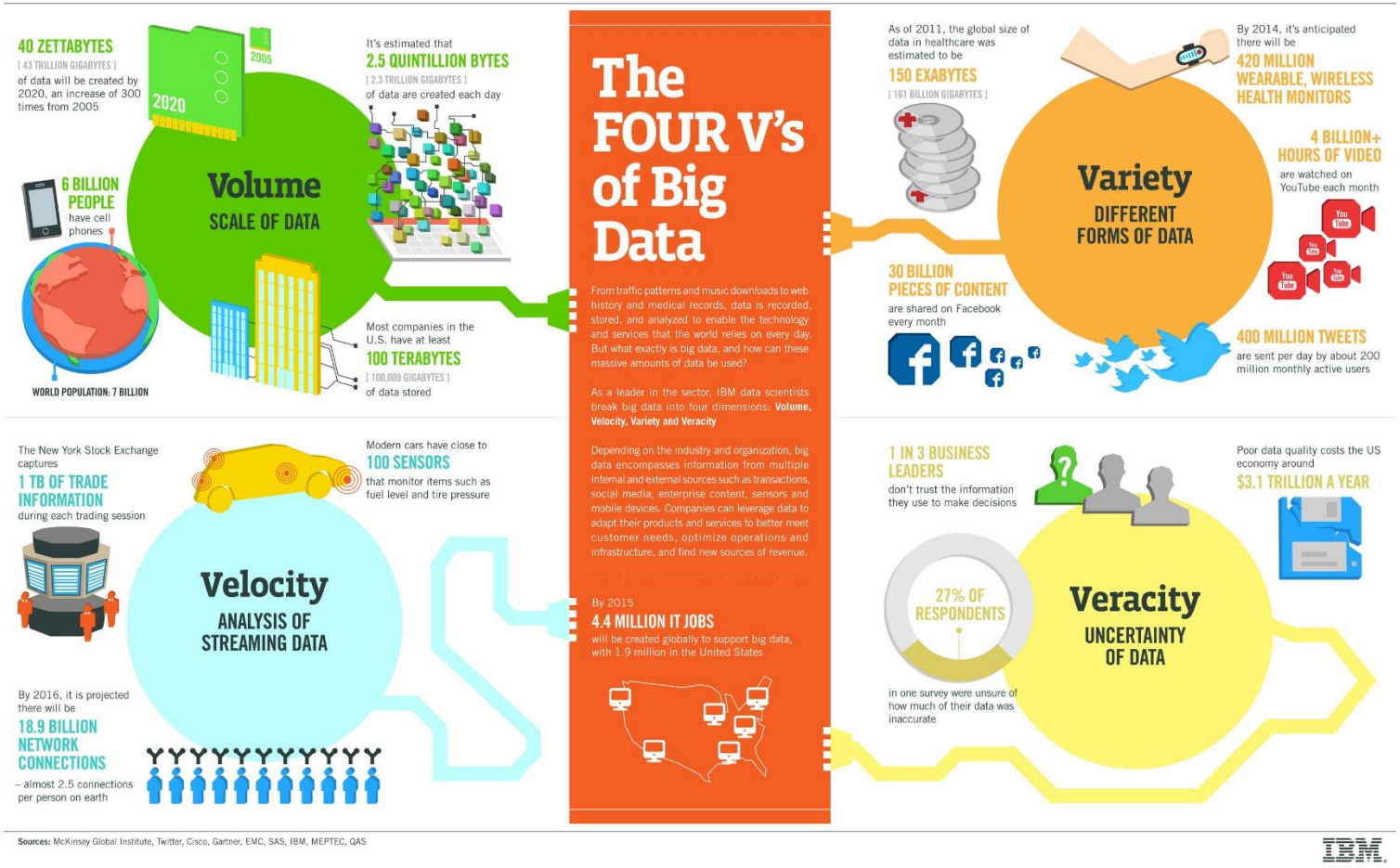
- What is Big Data?
  - According to Seagate, the volume of data generated worldwide will increase from 33 in 2018 to about 175 zettabytes in 2025



1 kilobyte	1,000
1 megabyte	1,000,000
1 gigabyte	1,000,000,000
1 terabyte	1,000,000,000,000
1 petabyte	1,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

Intro to Big Data

- What is Big Data?



Source: https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Intro to Big Data

❖ Hospital Big Data Example

1. Volume

- Hospitals around the world generate a massive amount of data in the form of patient records and test results
- According to IBM, 2.314 Exabytes of medical data collected annually around the world

3. Velocity

- According to IBM, medical data is experiencing a 48 percent annual growth rate

2. Variety

- Hospital can collect medical records in variety form, such as structured and unstructured data
- It can be textual information, excel or images (e.g., X-Ray images)

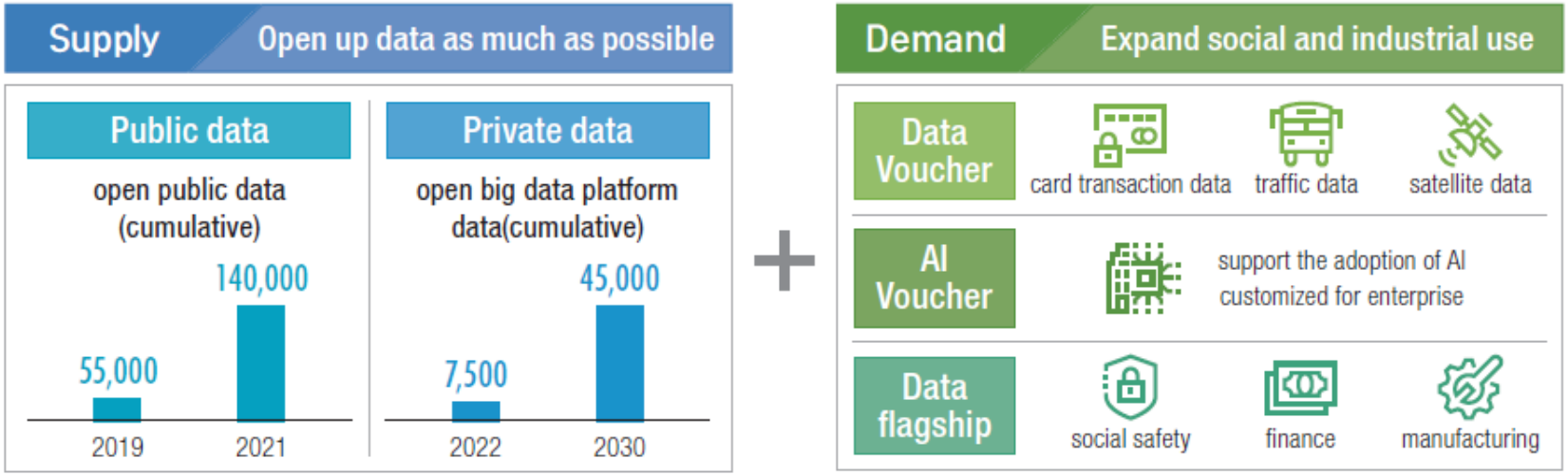
4. Veracity

- Since its healthcare field, the accuracy and trustworthiness of the data must be very high
- High accuracy in medical examination, prediction of disease

Intro to Big Data

❖ Big Data in South Korea

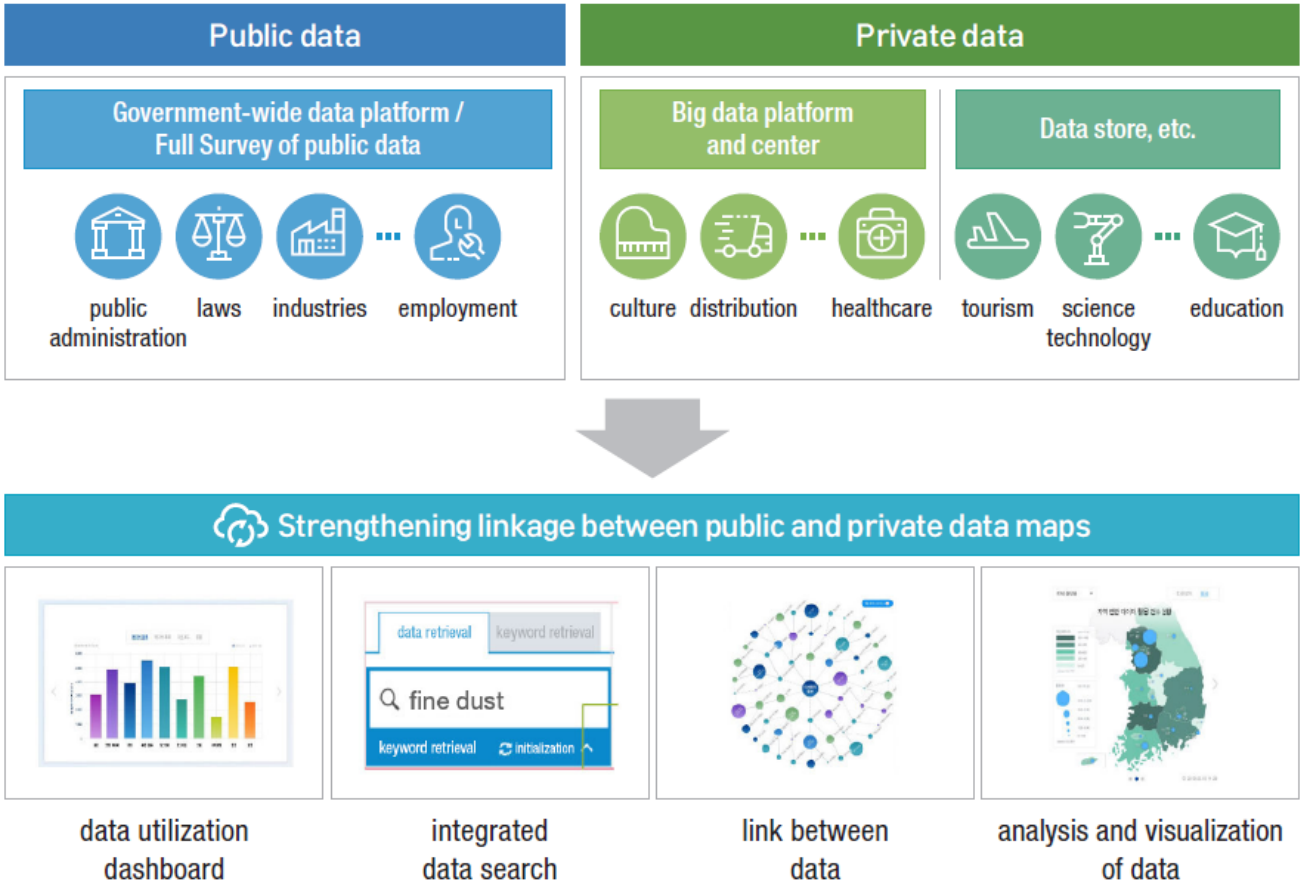
- Promotion of Opening Up Data and Reuse
  - Expanding the construction of AI learning data and securing of AI development infrastructure through the 'AI Hub' platform supply



Intro to Big Data

❖ Big Data in South Korea

- Strengthening Linkage between Public/Private Data Map



Part 2

WHAT IS BIG DATA LIFE CYCLE ?

Concept of big data life cycle

Importance and Usage of Big data

- Speed for analyzing large volumes of data
- Flexibility for various types of data such as unstructured data

  - Used in various fields such as trend analysis, marketing, and decision making by deriving meaningful information in real time.
  - Used to detect various changes such as consumer taste and behavior
  - Used to support quick decision-making without going through people

Concept of big data life cycle

Role of Big data in the future

Characteristics of the future society	Role of Big Data in the future	
Uncertainty	Insight	-Pattern analysis and future outlook based on social phenomena and physical world data -Response strategy in consideration of various situations with big data
Risk	Responsiveness	-Discover danger signs through big data analysis -Recognize and analyze issues in advance, and support quick decision-making and real-time response
Smart	Competitiveness	-Create context awareness and artificial intelligence service through big data analysis -Reinforce product competitiveness through trend change analysis based on big data
Convergence	Creativity	-Create a new convergence market such as smart city such as smart manufacturing through the use of big data

Concept of big data life cycle

Big data life cycle

- Consider 5V characteristics(volume, velocity and variety, veracity, value) of the data being processes.
- Organize the activities and tasks involved with acquiring, saving processing, analyzing and repurposing data

A specific data analytics lifecycle that organizes and manages the tasks and activities associated with the analysis of Big Data

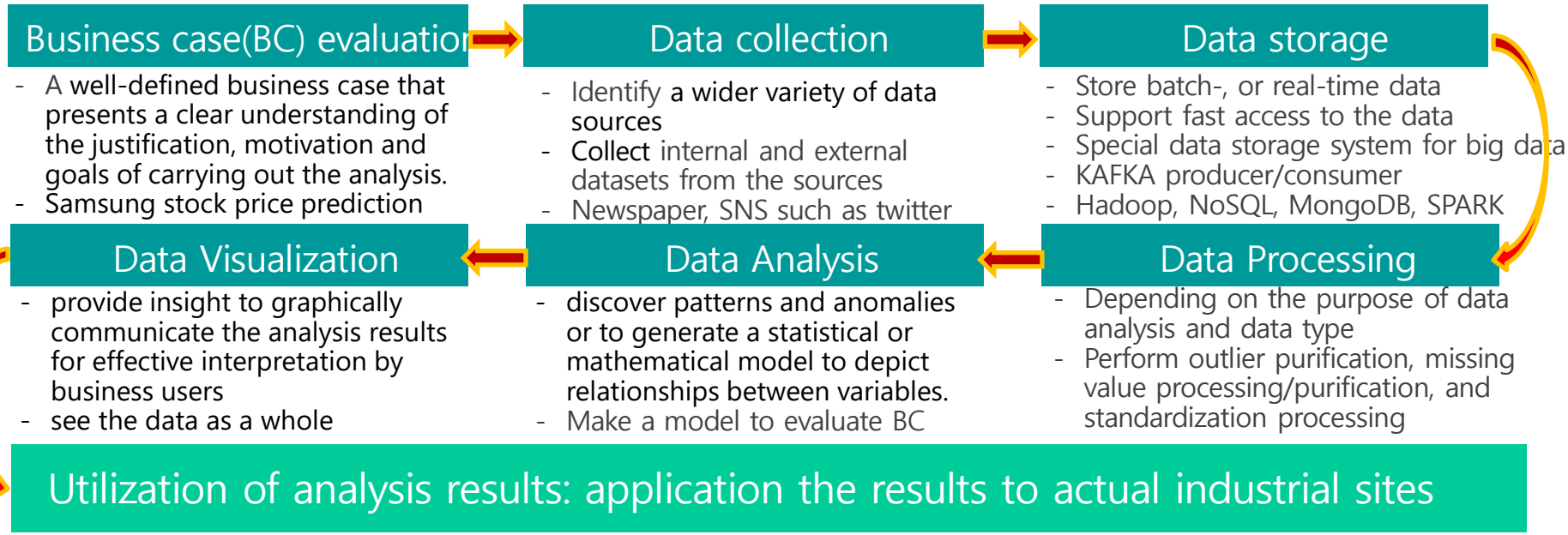
Tasks:

-Adoption and planning perspective

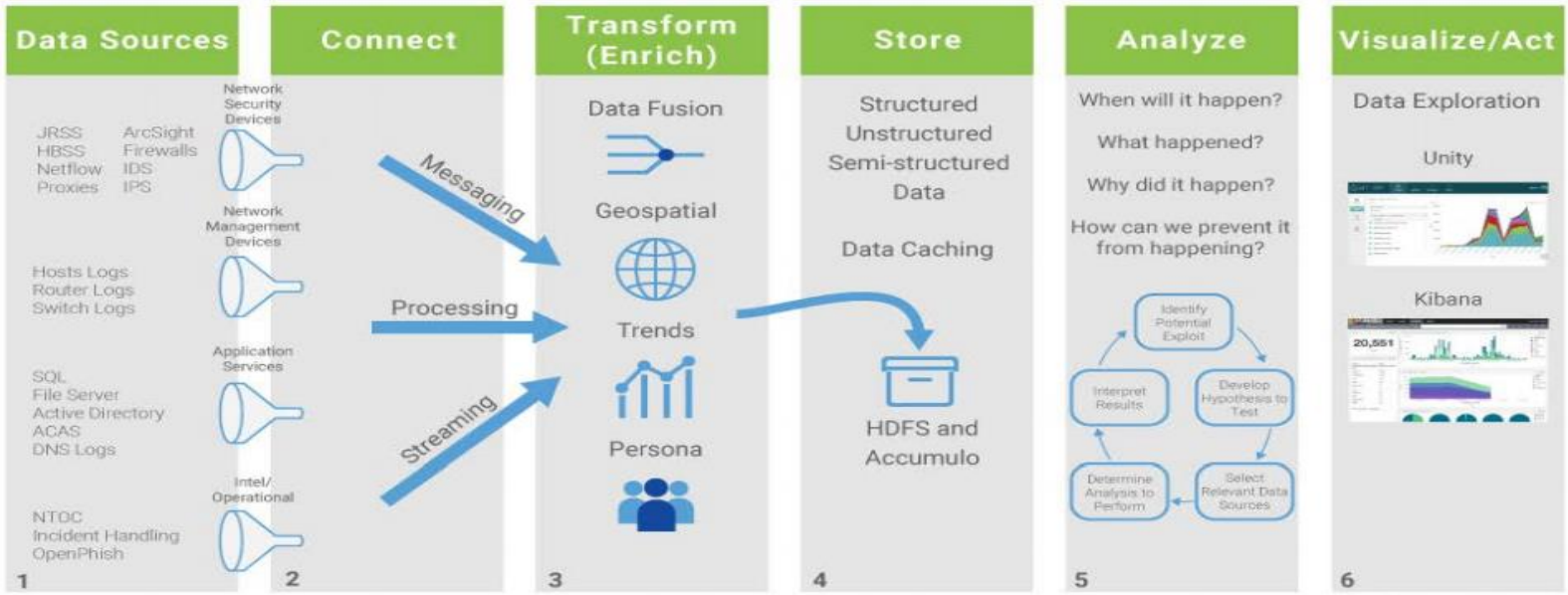
-Training, education, tooling and staffing

Concept of big data life cycle

7 Steps of big data life cycle

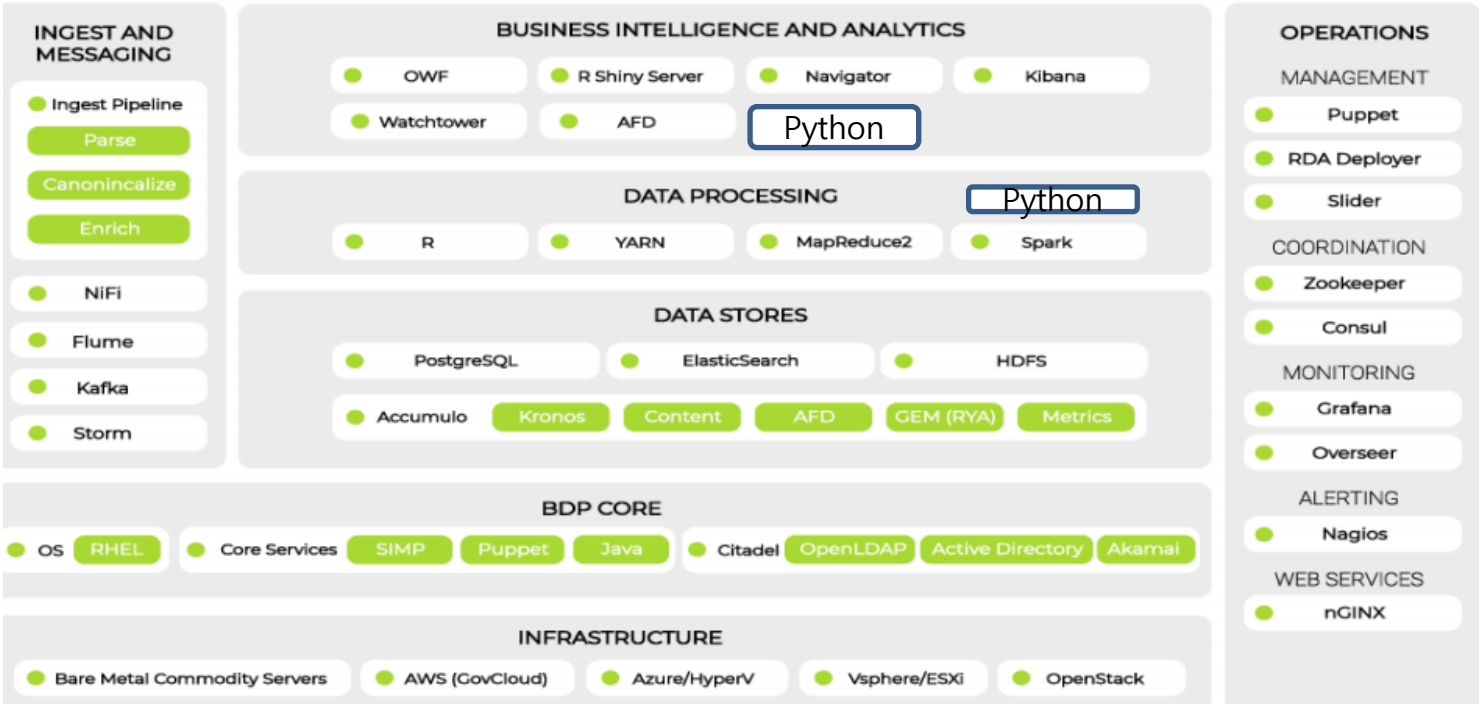


Flow of Big Data



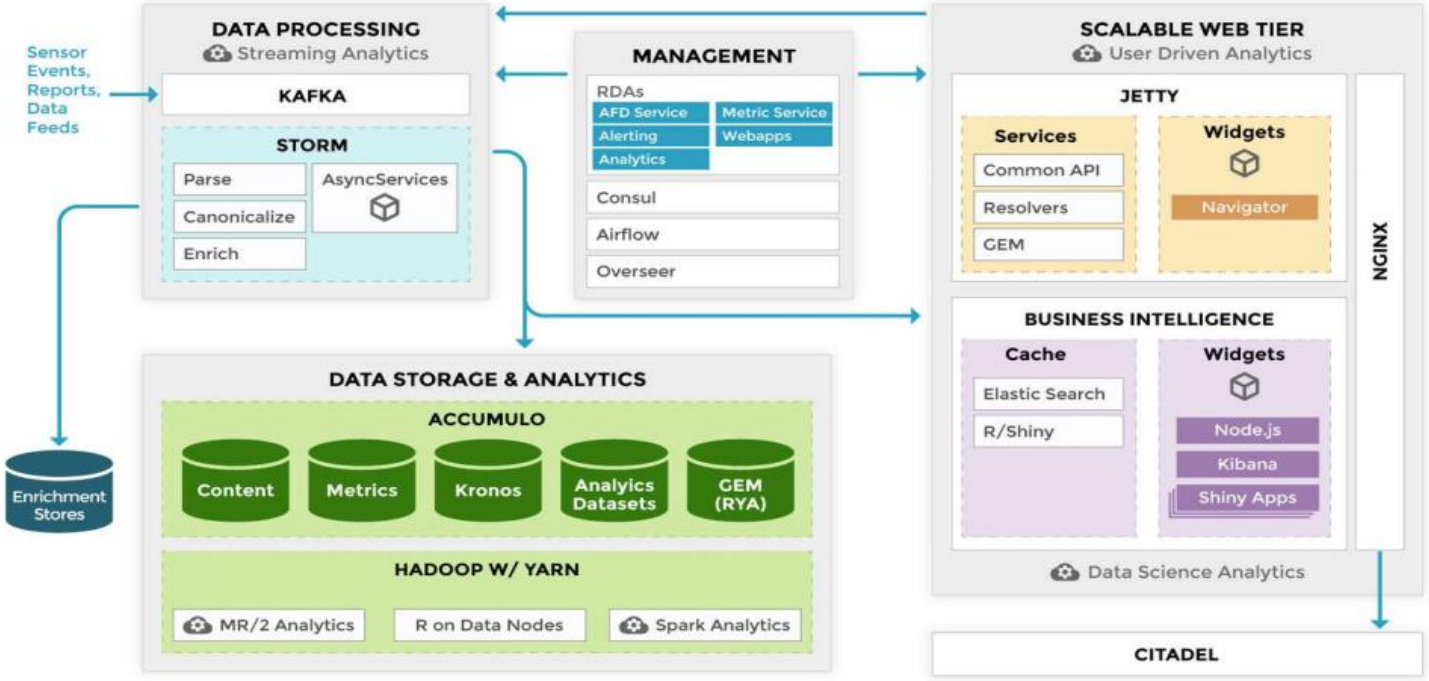
Source: Enlighten IT Consulting, a MacAulay-Brown company

Big data platform technology



Source: Enlighten IT Consulting, a MacAulay-Brown company

Big data platform architecture



Source: Enlighten IT Consulting, a MacAulay-Brown company

Big Data Sources

- Find out the sources to collect data which is used to achieve the purpose of business plan
- Kinds of big data

Kinds	Description
Structured data	Data stored in fixed fields -Relational database, spreadsheet
Semi-structured data	Data that is not stored in a fixed field, but contains metadata or schema -XML, HTML Text
Unstructured data	Data that is not stored in a fixed field -Text document, image, video, audio

Big Data Collection

- Internal data collection: internal file system, database owned by itself collecting structured data from management systems, sensors
- External data collection: unstructured data from outside connected to the Internet

Kinds	Description
Log collector	Web logging, transaction logging, click logging
Crawling	Visiting web data, Internet data
Sensing	Data collection from any kinds of sensors
RSS Reader/Open API	Data collection from shared data center
ETL(Extraction, Transformation, Loading)	Extraction after collecting data from any kinds of sensors

Big Data Storage

- Efficiently store and manage data to extract meaningful information
- A storage method that can accommodate large capacity, unstructured, and real-time

Kinds	Products
Distributed file system	Google File System(GFS) HDFS(Hadoop Distributed File System) Amazon S3 file system
NoSQL	Clouddata, Hbase, Cassandra
Parallel DBMS	VoltDB, SAP HANA, Vertica, Greenplum, Netezza
Network Storage System	SAN(Storage Area Network), NAS(Network Attached Storage)

Processing of Big Data

- Process of processing data suitable for analysis
  - Data set verification, missing value processing
  - Outlier handling, feature engineering

Characteristics of the data
- Check null data
- Check data quality

Data analysis time
- Ratio of data processing time: 80%~90%
- Ratio of time to perform the data analysis itself: 10% to 20%

Data processing: Any task that fixes data to make it easier to analyze

Processing of Big Data

Any task that fixes data to make it easier to analyze

Kinds	Description
Handling null data	Remove null data, replace null data
Detecting outlier	Detect outlier in the data Remove and replace the detected outlier
Converting data	Convert category data to numerical data
Normalizing data	Normalization of data(Min-Max, Standards, etc)

Big Data Analysis

-Discover patterns and anomalies of data

-Generate a statistical or mathematical model to depict relationships between variables of data

- Exploratory data analysis (EDA)
- Statistical data analysis
- Machine learning
  - Supervised Running: Regression, Classification
  - Unsupervised running: clustering
- Deep learning
  - Artificial neural network (ANN): Perceptron, Multilayer perceptron
  - Convolution neural network (CNN)
  - Recurrent neural network (RNN)
  - Long short time memory (LSTM)

Examples of Big Data Analysis

Titanic Data

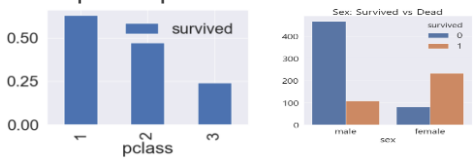
- X values: survived, pclass, sex, sibsp, parch, fare, embarked, deck class, adult, alone
- Y value: alive (yes or no)

Data Processing

- Process null data such as deck
- Convert "who" and "adult_male" data to numerical data

Exploratory Data Analysis

- Explore pclass vs alive



	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False

Statistical Data Analysis

- Apply mathematical equation
- Mean, variance, standard deviation
- F-test

Machine learning

- Supervised learning: regression, classification
- Unsupervised learning: clustering

Deep learning

- Convolutional neural network

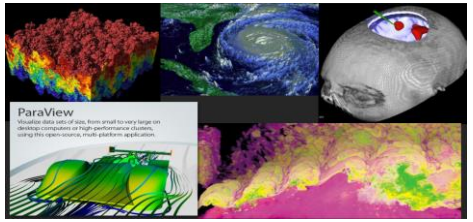
Big Data Visualization

- Understand what makes a visualization effective through the study of core principles
- Tools to graphically communicate the analysis results for effective interpretation by business users

Source:Alark Joshi, Yale Univ.

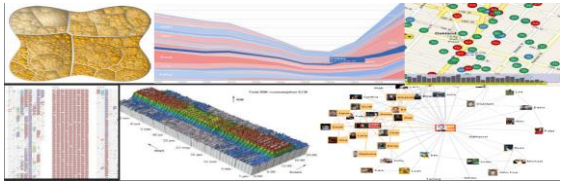
Scientific visualization

Structural Data
- Seismic, Medical,



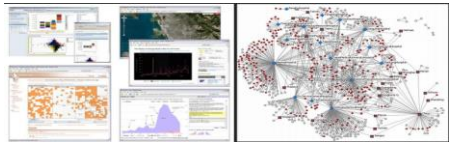
Information visualization

No inherent structure
- News, stock market, top grossing movies, facebook connections



Visual analytics

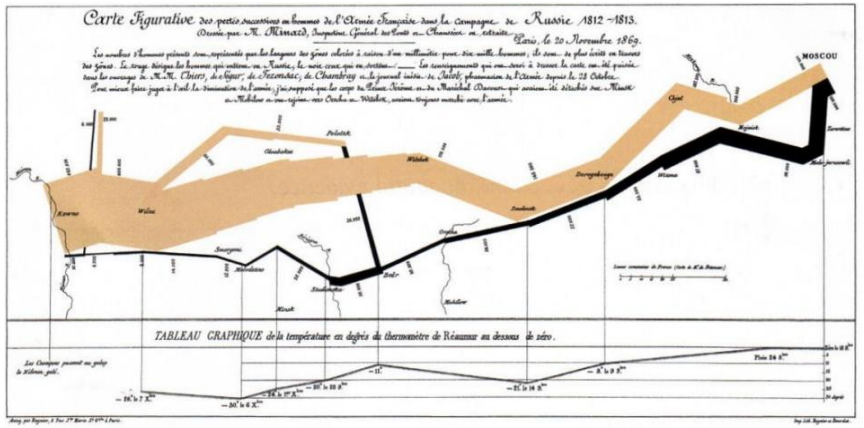
Use visualization to understand and
synthesize large amounts of multimodal data
- audio, video, text, images, networks of people



Integration of interactive visualization with analysis techniques
to answer a growing range of questions in science, business, and analysis.

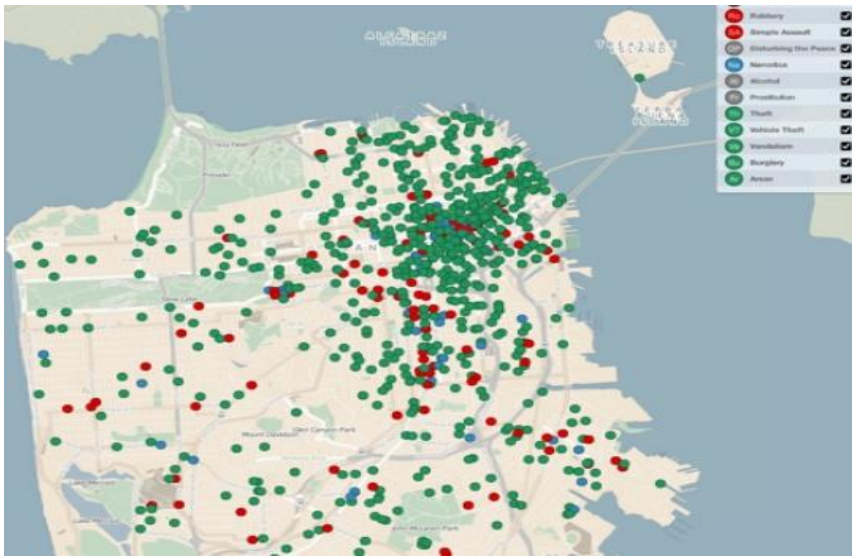
Examples of data visualization

Visualization of Napoleon's Army



This map drawn by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen, the thick band shows the size of the army (422,000 men) as it invaded Russia. The width of the band indicates the size of the army at each position. In September, the army reached Moscow with 100,000 men. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. The remains of the Grande Armée struggled out of Russia with 10,000 men. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn. Napoleon's March poster $14 postpaid; English/French version $18 postpaid.

Sanfrancisco crime map



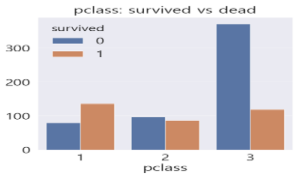
Examples of data visualization

- Titanic Data Visualization

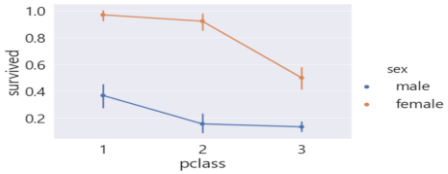
Alive vs plcass

survived	0	1	All
pclass			
1	80	136	216
2	97	87	184
3	372	119	491
All	549	342	891

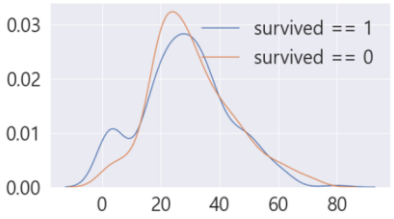
Alive-dead vs plcass



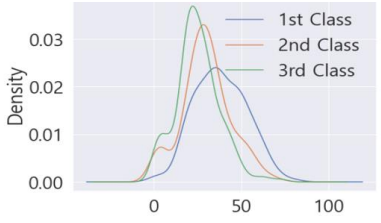
Alive-sex-plcass



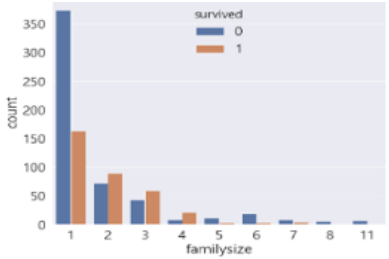
Alive vs age



Pclass vs age



Alive vs family size



Part 3

WHAT IS BIG DATA ANALYTICS

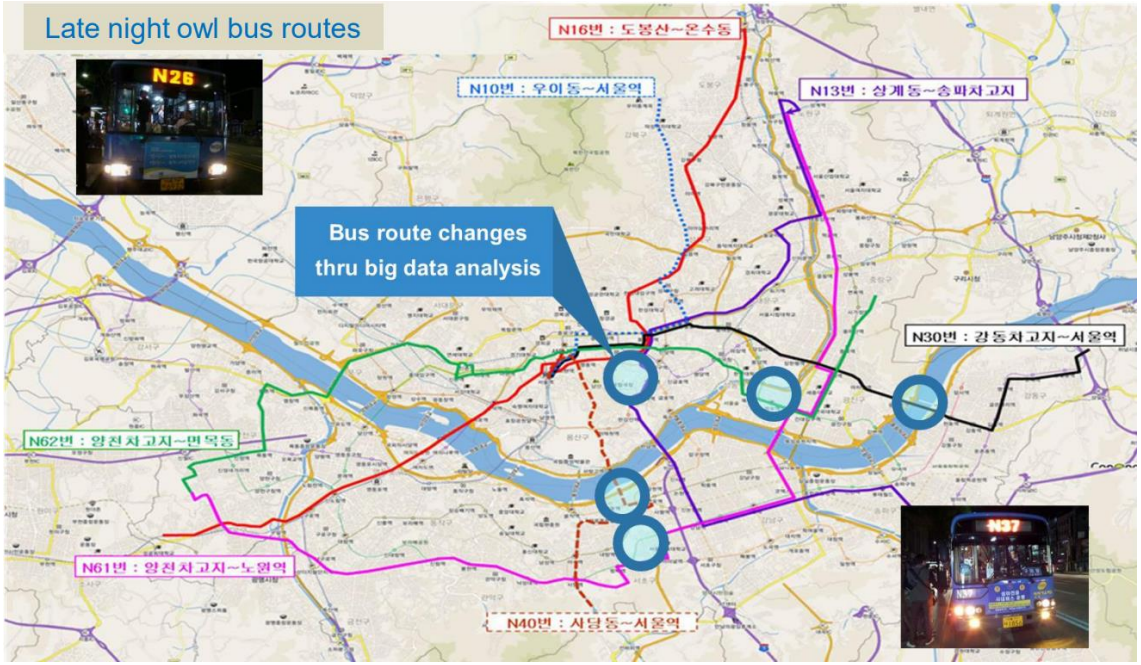
What is Big Data Analytics?

- Big Data analytics is a process used to extract meaningful insights
  - hidden patterns
  - unknown correlations
  - market trends
  - customer preferences

- Big Data analytics provides various advantages
  - It can be used for better decision making, preventing fraudulent activities, reduce cost among other things.

What is Big Data Analytics?

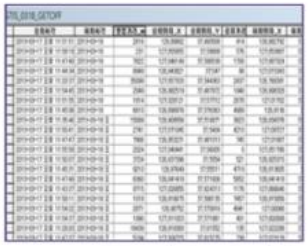
- Example of Big Data Analytics (The OWL Service)
  - Taxi at night is expensive and difficult to catch
  - Through a partnership with Korea Telecom, Seoul Government gained access to anonymized mobile communication data
    - 3 billion mobile call logs, 5 million taxy ride data



What is Big Data Analytics?

❖ Example of Big Data Analytics (The OWL Service)

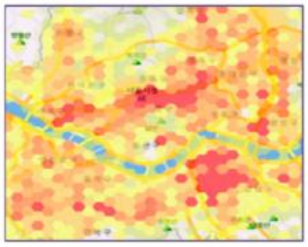
Big Data Analytics for Bus Route Optimization



1. Data collection and analysis

2. Layer modelling based on ride locations

3. Hexagon mapping



4. Floating population density analysis

5. Bus route optimization with floating population

6. Dispatch timetable adjusted accordingly

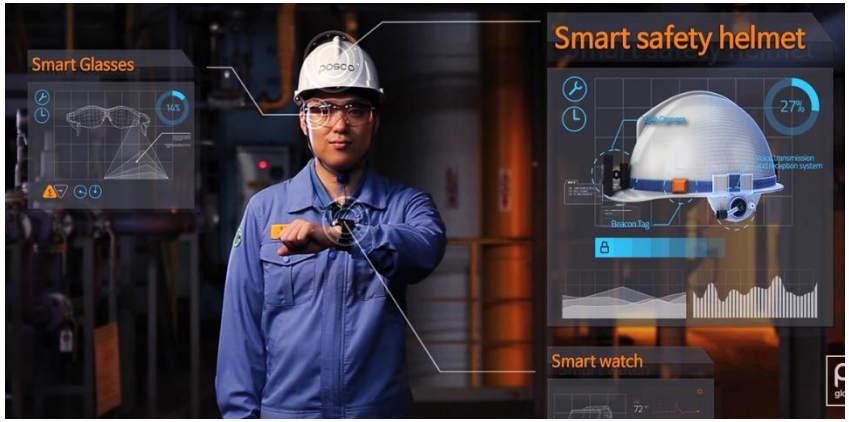
Impact

After three months of operating two routes

- Covers 42% of Seoul residents
- 7,900 passengers per day
- 2.3 million less car trips per year
- $13 million fare savings
- 500 metric tons reduction in greenhouse gas emission per year
- A service satisfaction score of 82 points (74.3 points for standard buses)

What is Big Data Analytics?

- Example of Big Data Analytics (POSCO)
  - POSCO is one of the largest hot rolling plant in the world
  - POSCO reduced energy input by 2% and save 1 billion won annually
    - Collecting and analyzing manufacturing environment data through sensors in factory
    - Maintaining the optimal working conditions through AI

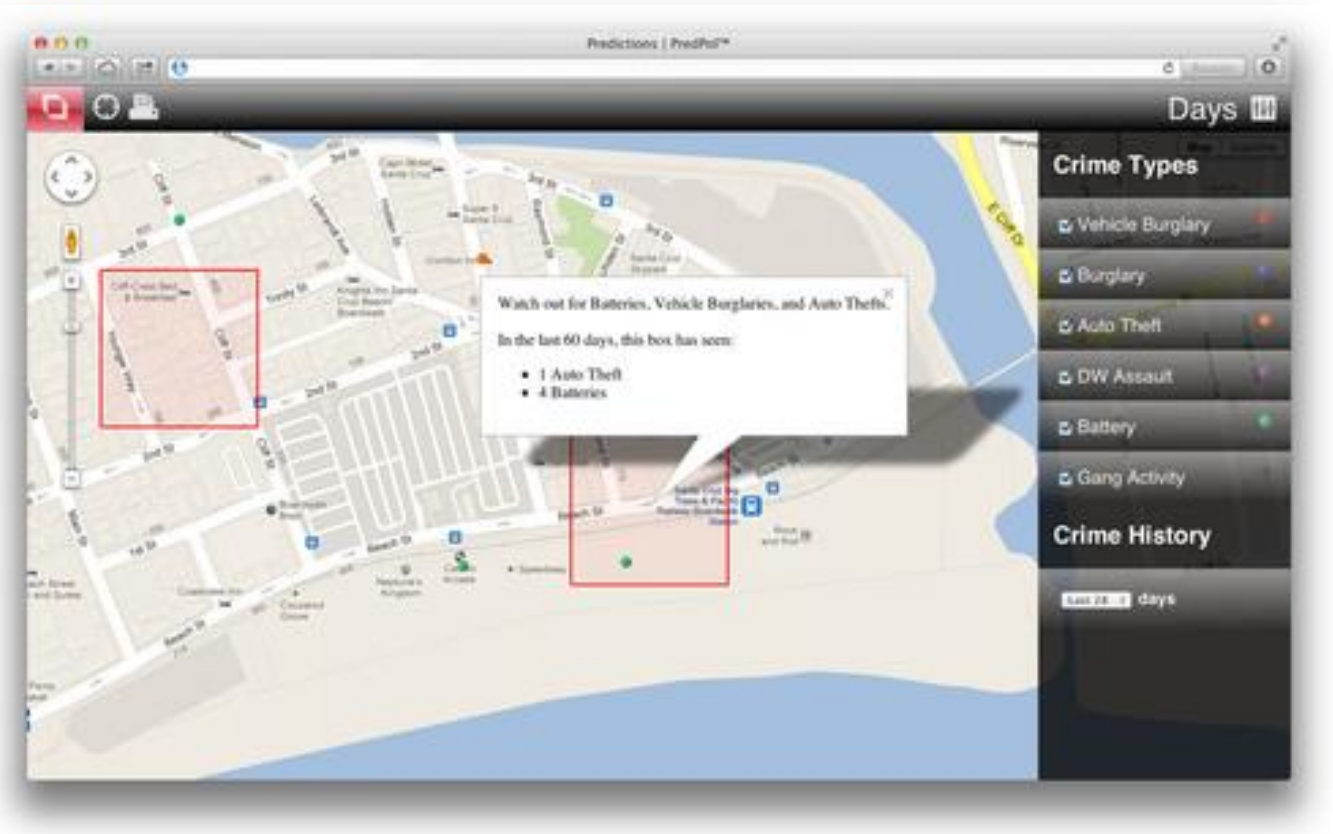


What is Big Data Analytics?

- Example of Big Data Analytics (Netflix)
  - With 115 million subscribers, Netflix collect a huge amount of data
    - Ratings, watch history, searchers and others
  - Recommend the next movie you should watch or smart advertising

What is Big Data Analytics?

- Example of Big Data Analytics (Predpol)
  - Projection of areas where criminal activity is most likely
  - Reduced crime rates in Los Angeles, US



What is Big Data Analytics?

- Example of Big Data Analytics
  - Drug data reveal sneaky side effects

nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

News & Comment > News > 2019 > May > Article

E-alert RSS Facebook Twitter

NATURE | NEWS

Drug data reveal sneaky side effects

Mining of surveillance data highlights thousands of previously unknown consequences when drugs are taken together.

Heidi Ledford

14 March 2012

Rights & Permissions

An algorithm designed by US scientists to trawl through a plethora of drug interactions has yielded thousands of previously unknown side effects caused by taking drugs in combination.

The work, published today in *Science Translational Medicine*[1], provides a way to sort through the hundreds of thousands of 'adverse events' reported to the US Food and Drug Administration (FDA) each year. "It's a step in the direction of a complete catalogue of drug–drug interactions," says the study's lead author, Russ Altman, a bioengineer at Stanford University in California.

What is Big Data Analytics?

- Example of Big Data Analytics
  - Shoplifting detection using artificial intelligence (AI)

Questions?

SEE YOU NEXT TIME!