Introduction to Big Data

교재 (참고) : 빅데이터 컴퓨팅 기술 (박두순 외, 한빛아카데미)

참고교재 (참고)

□ 빅데이터 컴퓨팅 기술
- ○ 박두순, 문양세 외
- ○ 한빛 아카데미
- ○ 2014년 06월 19일
- ○ 장점
  - ✓ IT 학생이 학습해야 할 빅데이터 관련 다양한 기술 전반
  - ✓ 다양한 SW, 시스템 소개
- ○ 단점
  - ✓ Introduction 수준 정도 (적은 분량)
    - – 실습 등을 위해서는 별도의 교재, 학습 필요
  - ✓ 출간 후 update 없음
- ○ 강의자료에 "교재 p.15" 등 언급은 이 참고교재

□ 수업 진행
- ○ **교재의 순서 반영 (내용은 일부 참고 정도)**
- ○ 인터넷 등에서 자료 보완
- ○ 실습이나 심화 학습은 개별적인 학습 필요

IT 기술의 변화

□ IT 기술의 패러다임 변화

표 1-1 정보 기술의 패러다임 변화 [01]

	PC 시대	인터넷 시대	모바일 시대	스마트 시대
패러다임 변화	디지털화, 전산화	온라인화, 정보화	소셜화, 모바일화	지능화, 개인화, 사물 정보화
정보 기술 이슈	PC, PC통신, 데이터베이스	초고속 인터넷, www, 웹 서버	모바일 인터넷, 스마트폰	빅데이터, 차세대 PC, 사물 네트워크Machine to Machine;M2M
핵심 분야(서비스)	PC, OS	포털, 검색 엔진, Web 2.0	스마트폰, 웹 서비스, SNS	미래 전망, 상황 인식, 개인화 서비스
대표 기업	MS, IBM	구글, 네이버, 유튜브	애플, 페이스북, 트위터	구글, 삼성, 애플, 페이스북, 트위터
정보 기술 비전	1인 1PC	클릭 e-Korea	손 안의 PC, 소통	IT everywhere, 신가치창출

IT 기술의 변화

□ IT 기술의 주도권이 Data로 이동 (빅데이터는 미래 경쟁력과 가치 창출의 원천)

데이터 규모	EB(엑사바이트) (1990년대 말=100EB)	ZB(제타바이트) 진입 (2011년대 말=1.8ZB)	ZB 본격화 시대 (2020년=2011년 대비 50배 증가)
데이터 유형	정형 데이터 (데이터베이스, 사무 정보)	비정형 데이터 (SNS, 멀티미디어)	사물 정보, 인지 정보 (RFID, 센서, 사물 정보)
데이터 특성	구조화	다양성, 복합성, 소셜	현실성, 실시간성

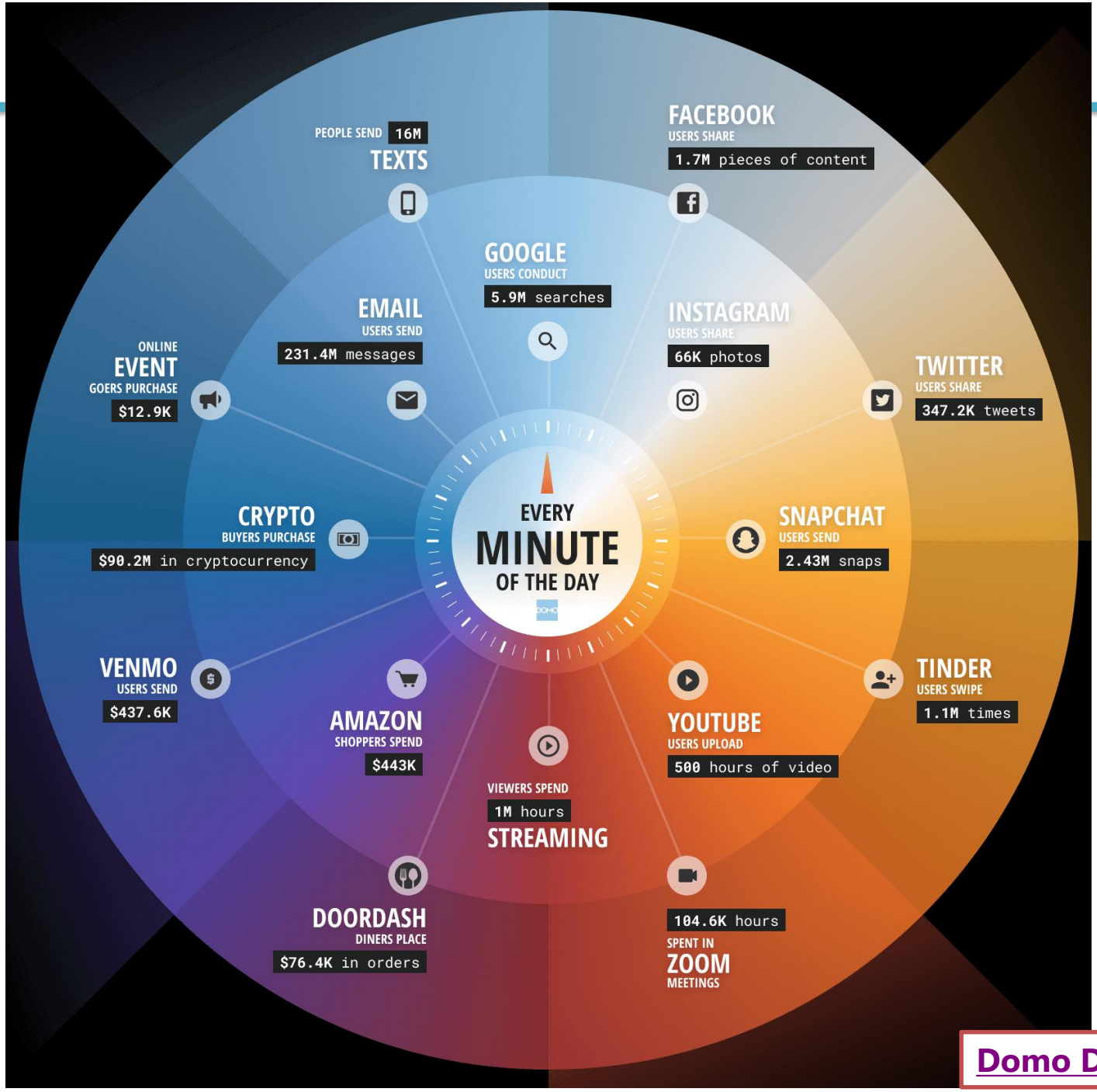
Value	Metric	
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

그림 1-1 정보 통신 기술 발전에 따른 데이터의 변화 방향 [01]

What's Big Data? No single definition!

□ Big Data

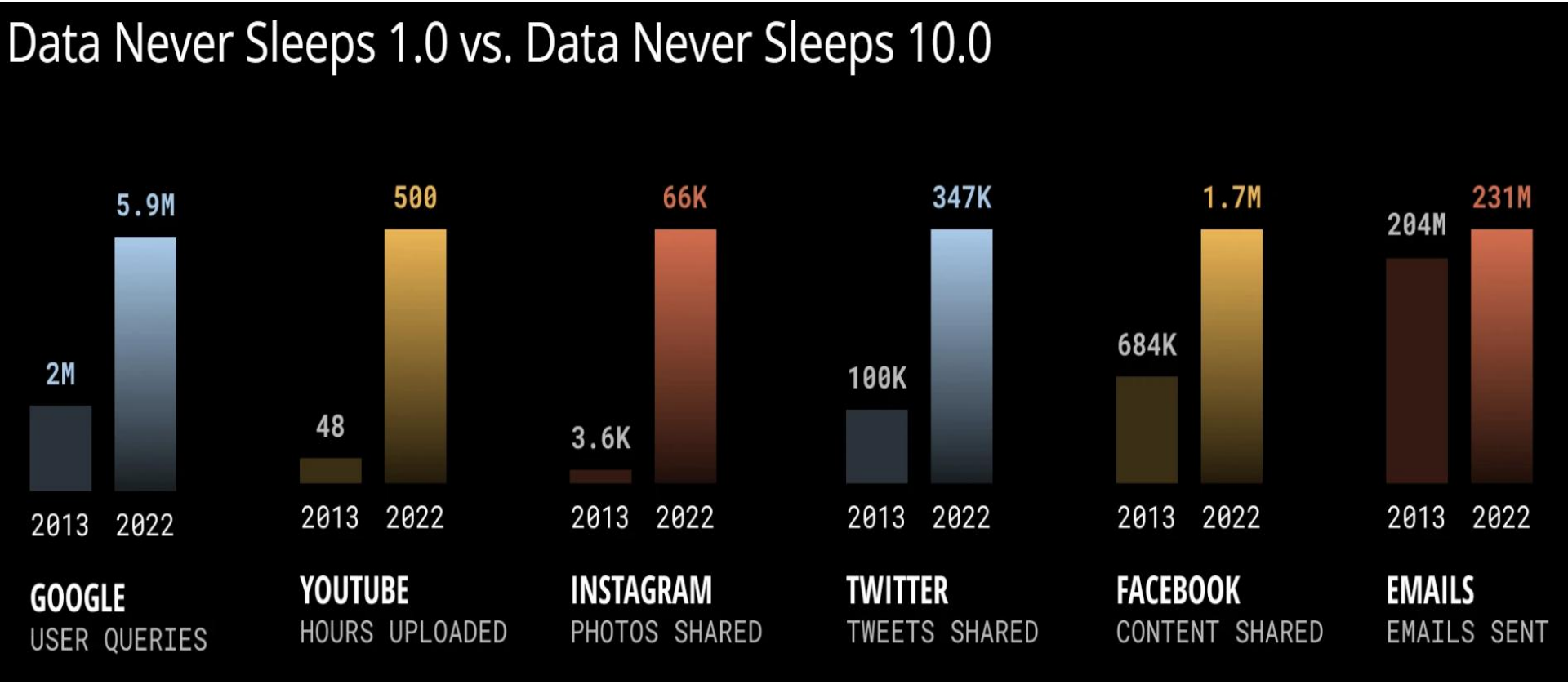
- ○ "collection of data sets so large and complex"
- ○ 기존의 DBMS나 데이터 처리 방법으로는 처리하기 곤란
  - ✓ 정형, 비정형의 데이터 집합 포함
  - ✓ capture, storage, search, sharing, transfer, analysis, visualization 기술 포함



□ 2022년 기준
(1분당 ...)

Domo Data Never Sleeps 10.0

□ 1분당 …



□ According to Statista,

○ the total amount of data (created, captured, copied and consumed globally)

✓ 2022 : 97 zettabytes

✓ 2025 : 181 zettabytes

단위

출처 : by Lauro Rizzatti

Multiplying Factor	SI Prefix	Scientific Notation	Name
1 000 000 000 000 000 000 000 000	Yotta (Y)	10^{24}	1 septillion
1 000 000 000 000 000 000 000	Zetta (Z)	10^{21}	1 sextillion
1 000 000 000 000 000 000	Exa (E)	10^{18}	1 quintillion
1 000 000 000 000 000	Peta (P)	10^{15}	1 quadrillion
1 000 000 000 000	Tera (T)	10^{12}	1 trillion
1 000 000 000	Giga (G)	10^{9}	1 billion
1 000 000	Mega (M)	10^{6}	1 million
1 000	kilo (k)	10^{3}	1 thousand
0 001	milli (m)	10^{-3}	1 thousandth
0 000 001	micro (u)	10^{-6}	1 millionth
0 000 000 001	nano (n)	10^{-9}	1 billionth
0 000 000 000 001	pico (p)	10^{-12}	1 trillionth
0 000 000 000 000 001	femto (f)	10^{-15}	1 quadrillionth
0 000 000 000 000 000 001	atto (a)	10^{-18}	1 quintillionth
0 000 000 000 000 000 000 001	zepto (z)	10^{-21}	1 sextillionth
0 000 000 000 000 000 000 000 001	yocto (y)	10^{-24}	1 septillionth

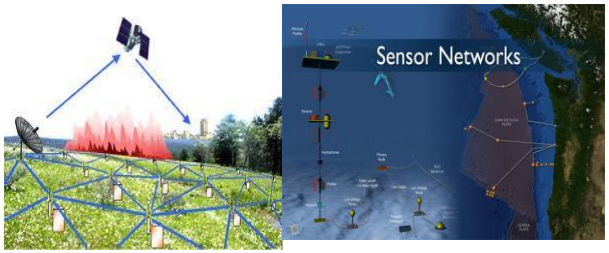
Big Data의 수집 사례

□ Big Data의 수집
- ○ Web data, e-commerce
- ○ purchases at department/grocery stores
- ○ Bank/Credit Card transactions
- ○ SNS, YouTube
- ○ Smartphone, Smartwatch
- ○ Science
- ○ IoT, Sensor Network

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

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



Sensor technology and networks
(measuring all kinds of data)

초연결 (hyper-connected) 시대

출처 : by Ruoming Jin

Big Data 활용 사례

1. 집계 (Aggregation) 및 통계 (Statistics)
   - ○ 데이터 웨어하우스 (Data Warehouse)
     - ✓ 과거부터 현재까지 수집된 중앙집중적, 전사적 데이터 저장소
   - ○ OLAP (OnLine Analytical Processing)
     - ✓ 다양한 각도 (다차원 분석 질의)에서 ①사용자가 직접, ②대화식으로, ③다양한 도구의 지원으로 정보를 분석하는 과정
       - – 저장, 관리 : 데이터 웨어하우스
       - – 전략적 정보 변환 : OLAP

2. 검색 : 색인 (Indexing), 질의 (Querying)
   - ○ 키워드 기반 검색
   - ○ 패턴 매칭 (XML/RDF)
   - ○ 자연언어 검색

3. 지식 추출 (Knowledge discovery)
   - ○ 데이터 마이닝 (Data Mining)
     - ✓ Knowledge Discovery, Data Discovery
     - ✓ 대규모의 데이터로부터 특정한 패턴이나 경향을 찾아내는 SW 기술
   - ○ 통계적 모델링, 기계 학습 (AI)

Big Data: 3V 모델 (초기, 기본 특성)

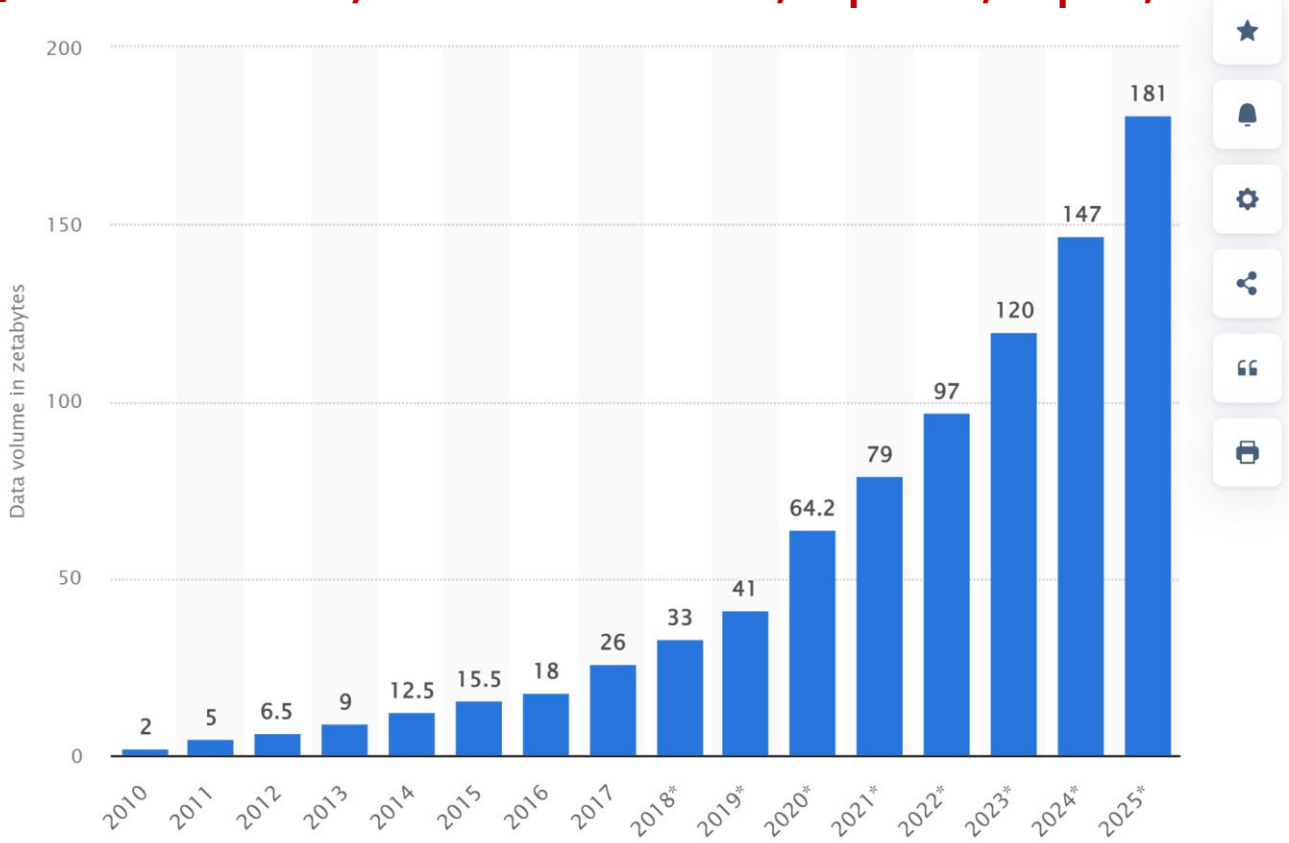
"Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Gartner 2012)

Big Data : 3V (출처)

3V (Gartner)
4V (IBM)
5V
7V
.
.
.

Volume (Scale)

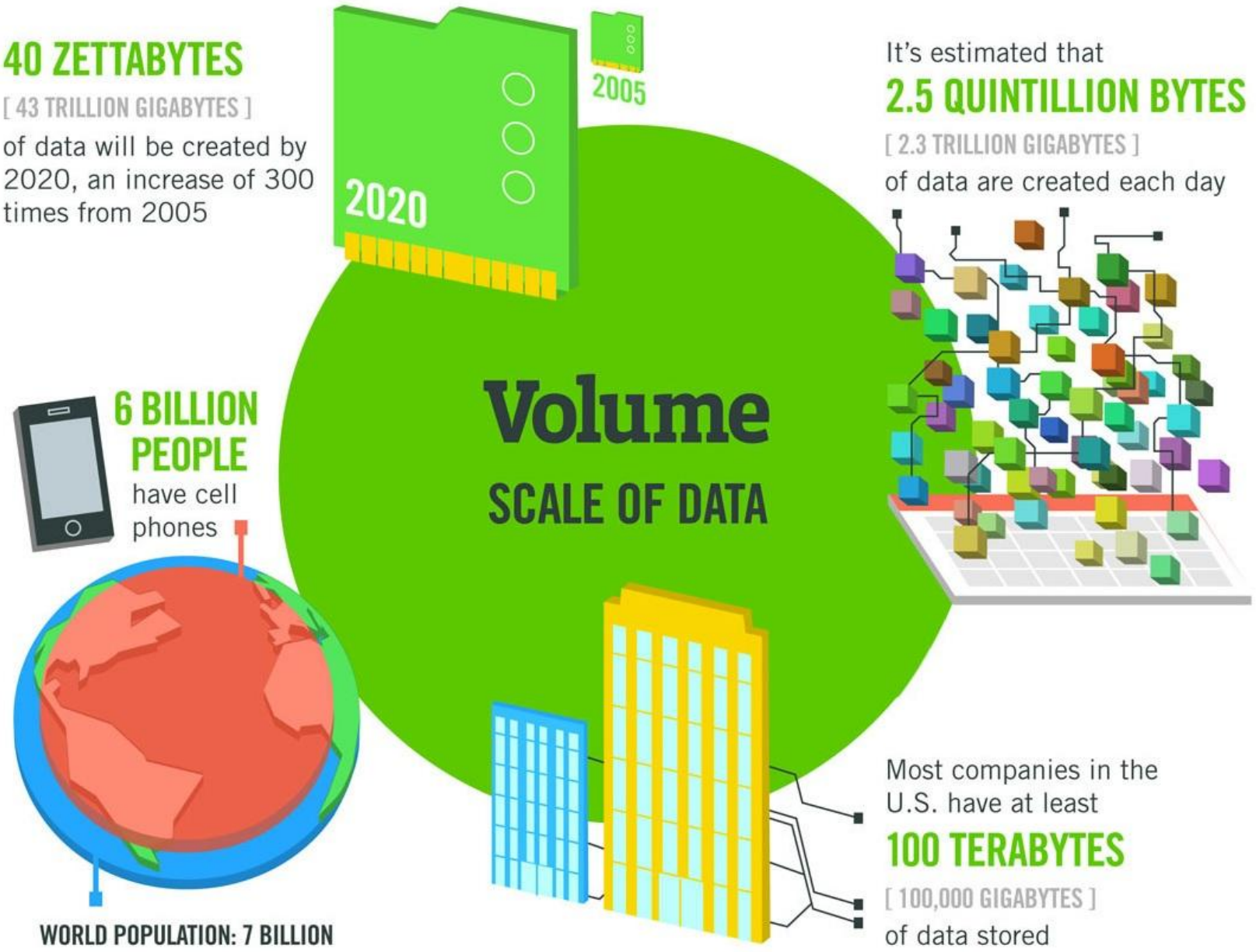
- **Data Volume**
  - 40x increase from 2010 – 2021, and so on
  - From 2 zettabytes to 79zb
- Data volume is increasing exponentially (smartphone, IoT, ...)
- **Volume of data/information created, captured, copied, and consumed worldwide**



데이터 (2011~2020)
예측치 (2021~2025)
단위 : zb

Statista 2022 (출처)

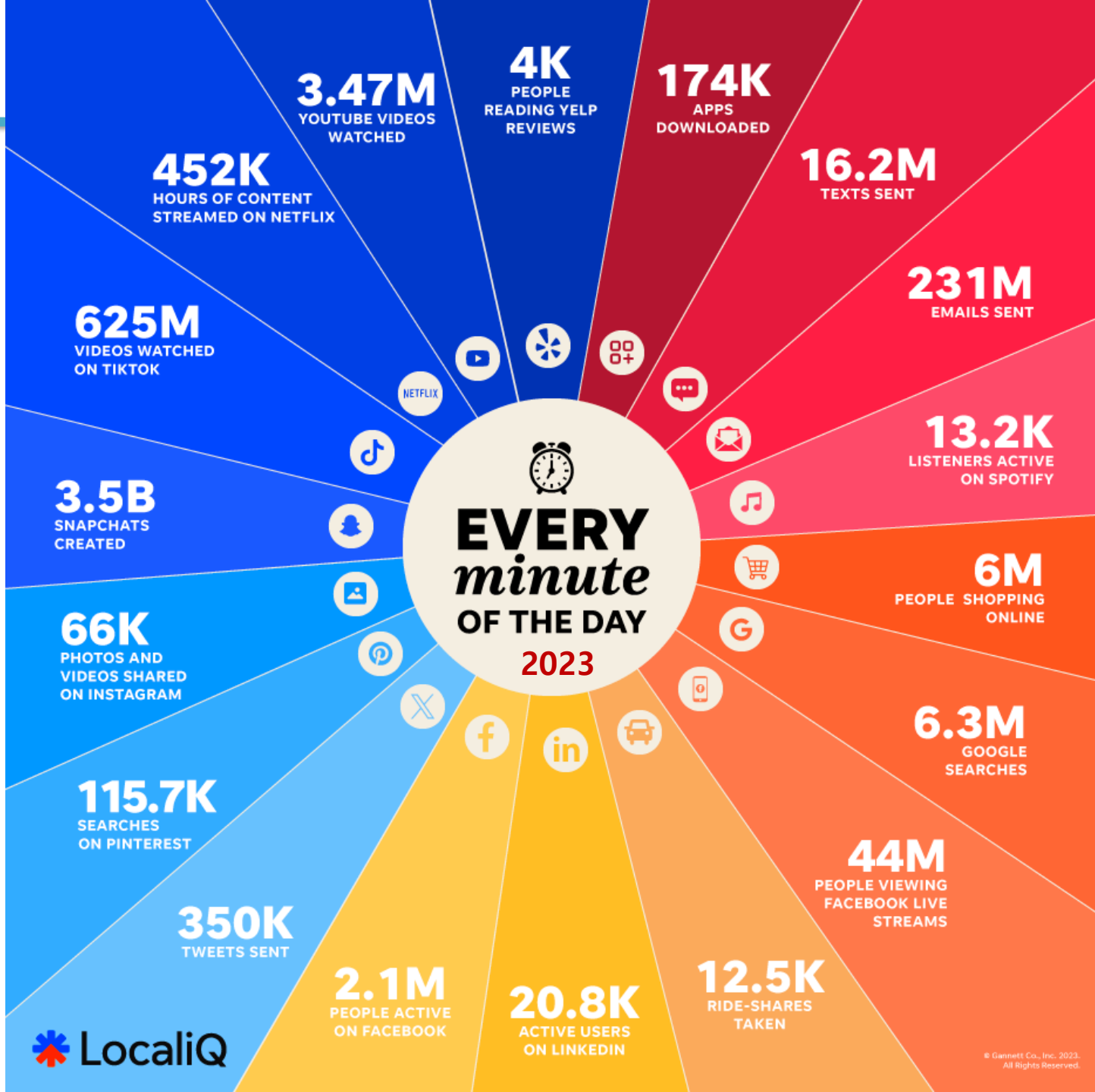
Volume (Scale)



Data Volume (from IBM) (출처)

Volume (Scale)

What happens
in an internet minute?



LOCALiQ(출처)

•95 million photos and videos are shared on **Instagram** each day. That translates to **65,972 each minute!**

•The average US user spends 53 minutes on Instagram per day. That's 297 hours per year.

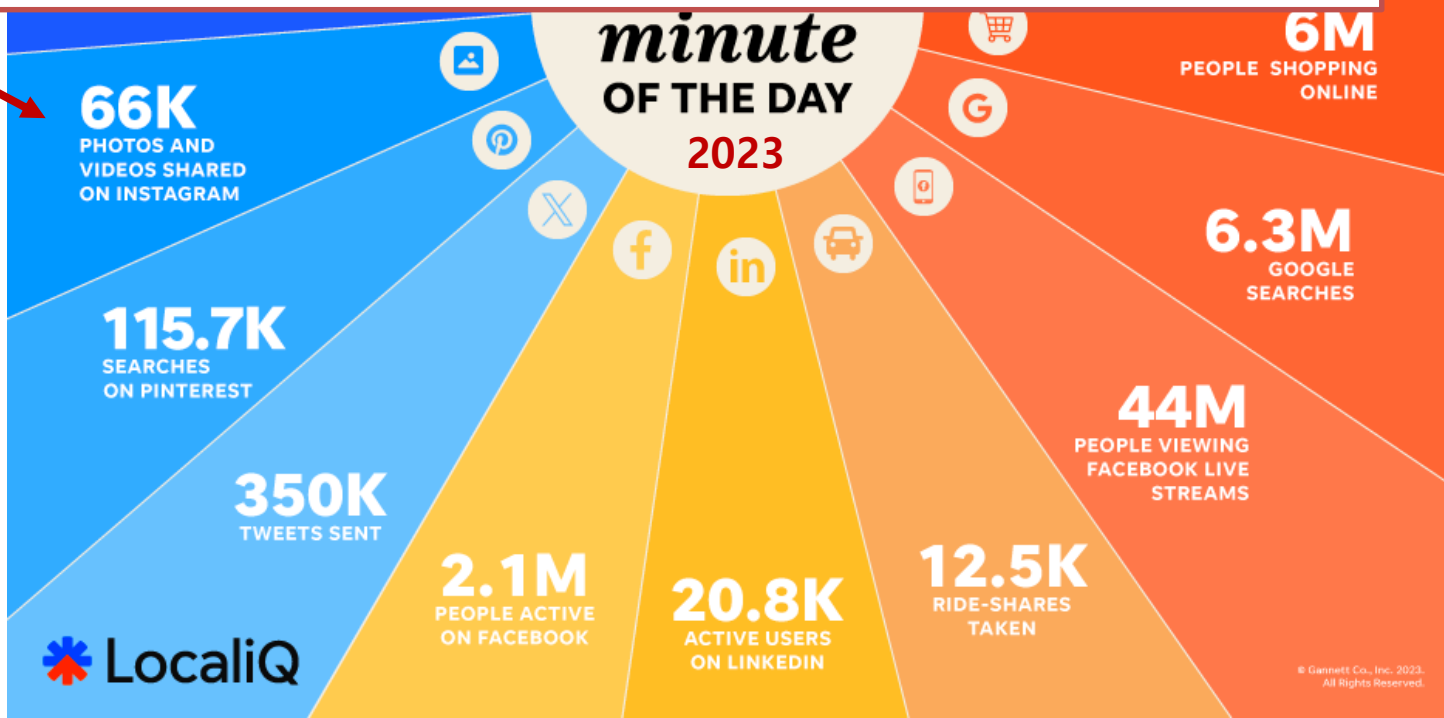
•Many Instagrammers are actively seeking out fresh content, as Instagram's explore page is viewed by 200 million accounts daily. That's 138,888 exploring new accounts each minute. (Be sure your business is showing across placements like this one by following Instagram's ranking algorithm.)

•81% of people use Instagram to research products and services.

•50% of people have visited a website to make a purchase after seeing it on Instagram.

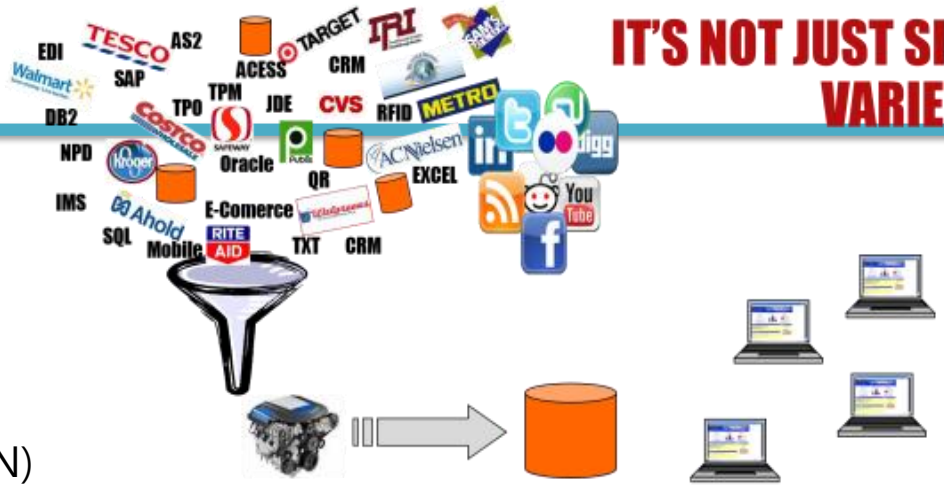
•Nearly 140,000 users are visiting a business's page on Instagram each minute.

•Businesses doing local marketing on Instagram are sure to see success, as Instagram posts with a location get 79% more engagement.



LOCALiQ(출처)

Variety (Complexity)



Data Variety (출처)

□ 다양한 유형의 데이터
- ○ Relational Data (Tables/Transaction/Legacy Data)
- ○ Text Data (Web), Image, Video
- ○ Semi-structured Data (XML, JSON)
- ○ Graph Data
  - ✓ Social Network, Semantic Web (RDF), …
- ○ Streaming Data
  - ✓ You can only scan the data once

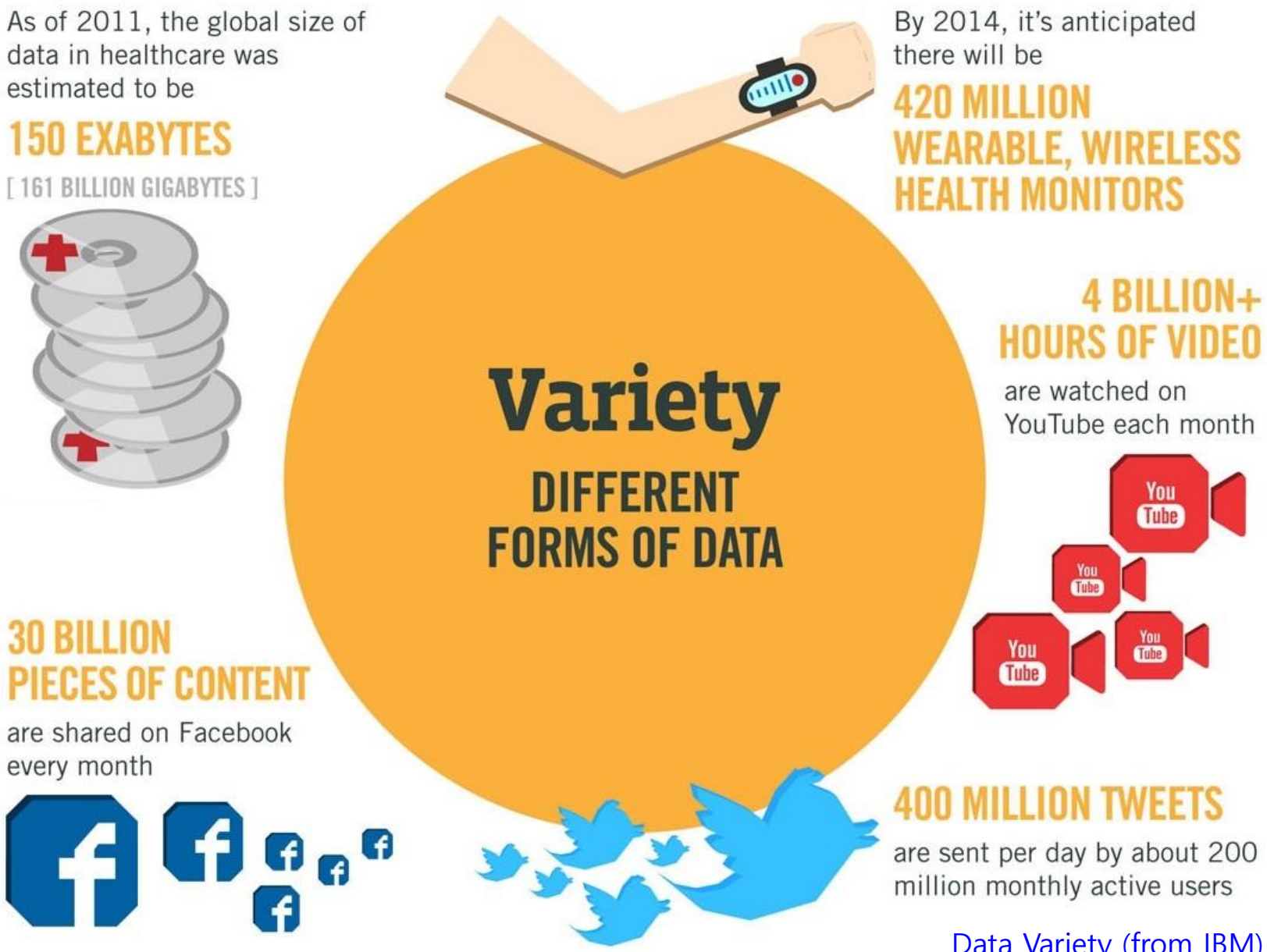
□ 하나의 어플리케이션이 다양한 유형의 데이터를 생산 또는 수집

□ 대규모 공공 데이터 (online, weather, finance, etc.)

□ 이러한 다양한 유형의 데이터들이 통합하여 분석 관리되어야

참고 : by Ruoming Jin

Variety (Complexity)



Data Variety (from IBM) (출처)

Velocity (Speed)

□ 데이터가 빠르게 생성되므로, 빠른 분석이 필요
- ○ Online Data Analytics
- ○ 데이터가 쌓이는 속도보다 분석 속도가 늦다면
  - → 데이터 누적, 지연, 처리 불가 …
- ○ 실시간 처리의 중요성
  - ✓ Streaming Data의 분석 등

□ 사례
- ○ 뉴스 검색, 주식 검색
  - ✓ 새로 생성된 뉴스 기사가 즉시 검색되어야
- ○ E-Promotions
  - ✓ 현재 위치, 과거 구매 이력, 취향 데이터 분석
  - → 적절한 장소 (ex. 상점 인근), 시간 (ex. 계절) 등에 따른 promotion
- ○ Healthcare monitoring
  - ✓ Wearable 장비, 휴대폰 등의 센서를 통한 정보 수집
  - → 비정상 상태로 판단, 즉각 조치 필요

참고 : by Ruoming Jin

Velocity (Speed)

The New York Stock Exchange captures

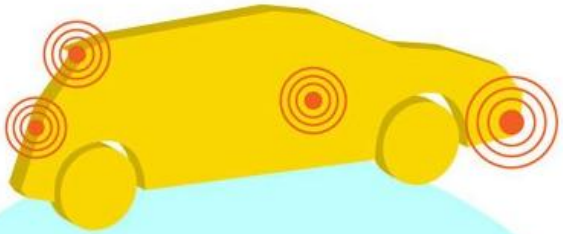
1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure



Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

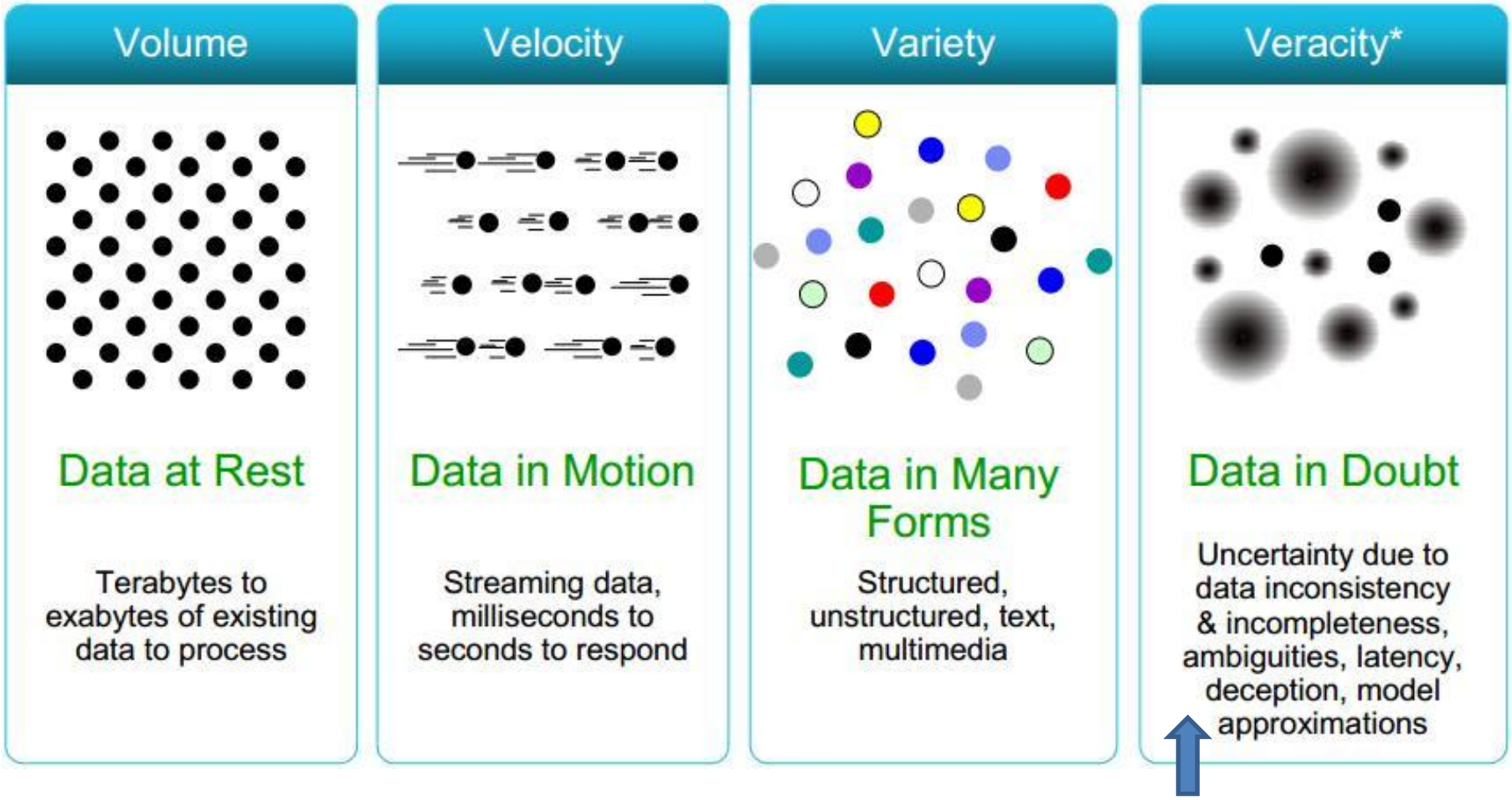
Data Velocity (from IBM) (출처)

4V 모델

□ + Veracity (정확성) : IBM

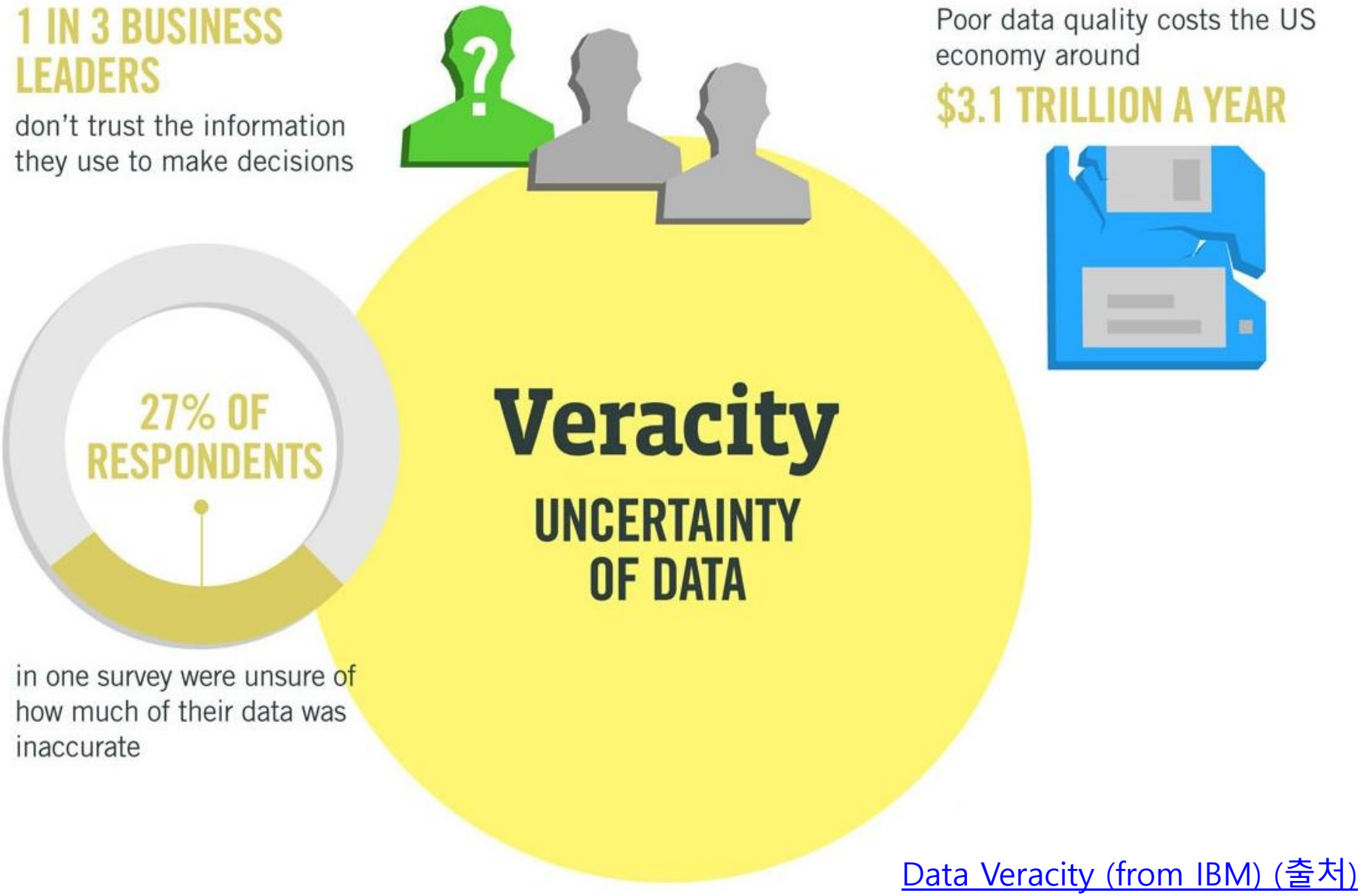
○ 정확하지 않은 데이터 포함 (-> Uncertainty)

○ 가공을 통해 높은 정확성을 유지해야



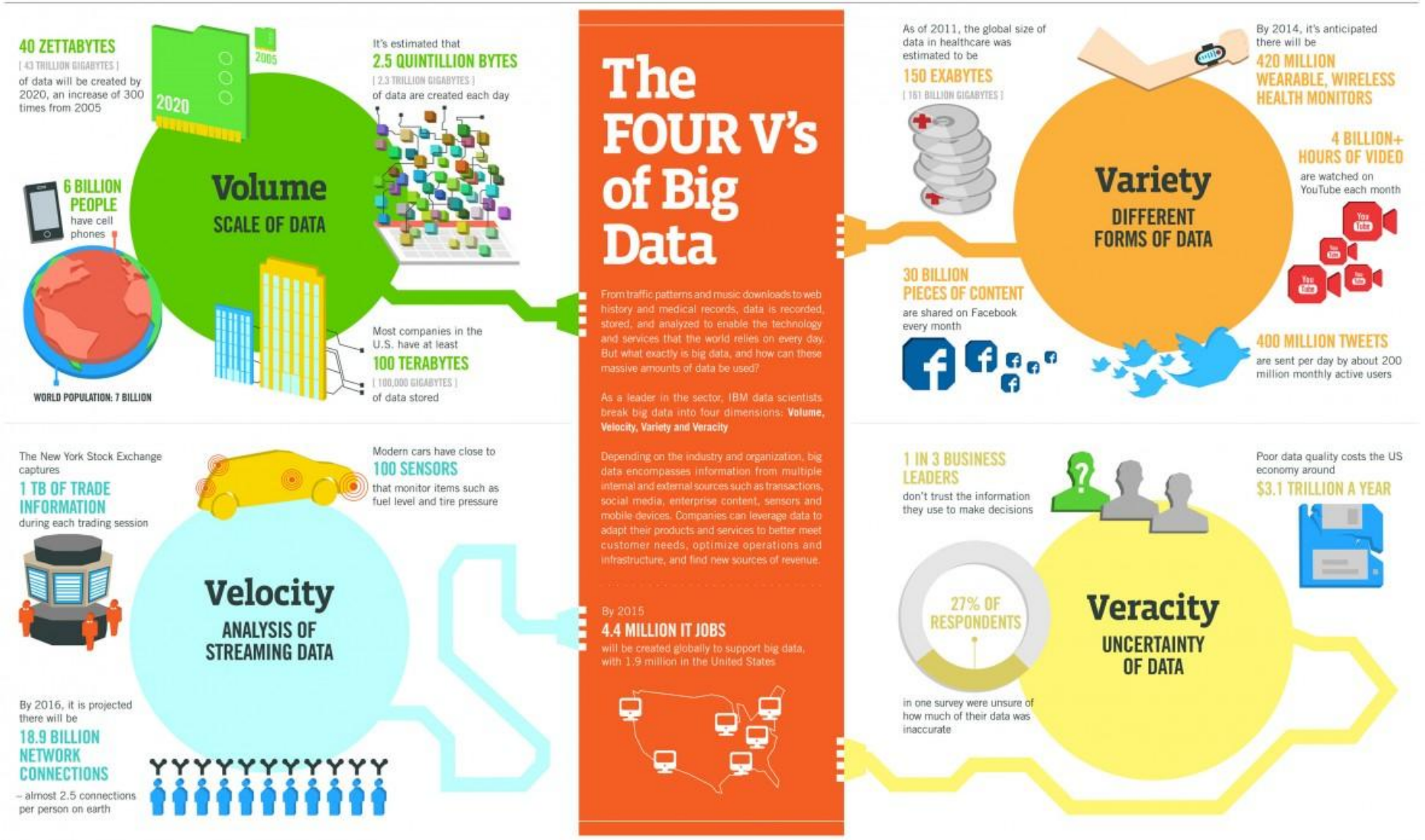
Big Data 정의 : 4V (출처)

Veracity (Uncertainty)



[Data Veracity (from IBM) (출처)]

Big Data: 4V 모델



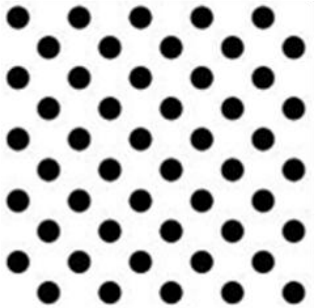
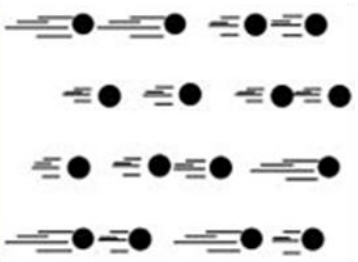
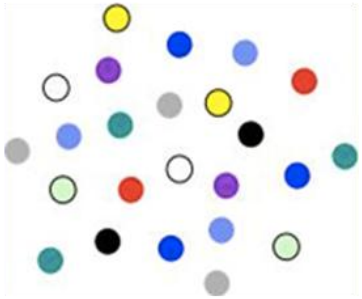
[The Four V's of Big Data (출처)]

5V 모델

□ + Value (가치)

○ 가치 있는 결과를 만들어 내는 데이터이어야 (품질이 낮은 데이터가 많아봐야 의미 없음)

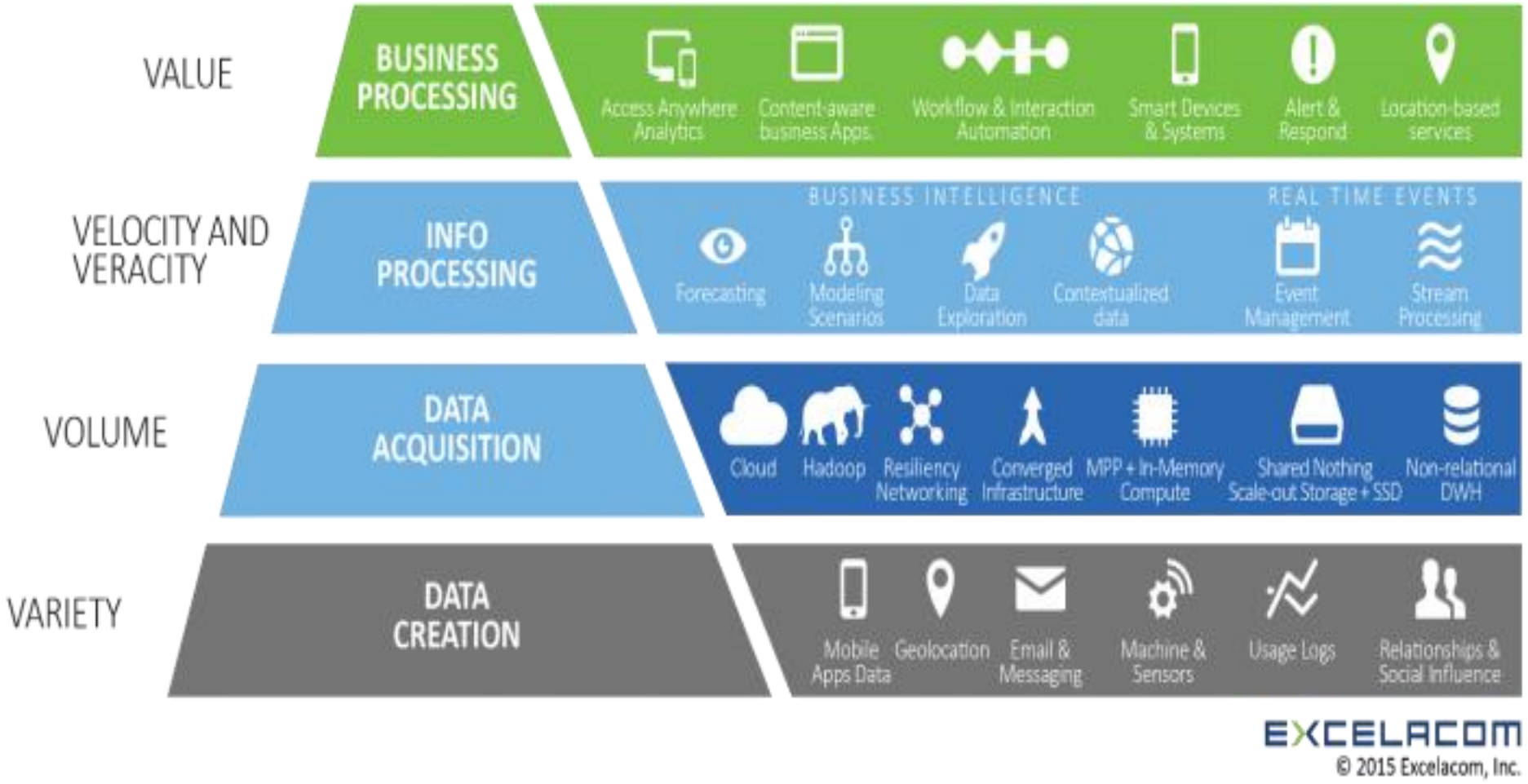
○ 정확성, 시간 (적시) 등과 연관 -> 가치를 창출할 수 있는 Business model이 있어야

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,...	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

Big Data: Five V's (출처)

5V 모델



[Big Data: Five V's (by Kelly LeBoeuf)]

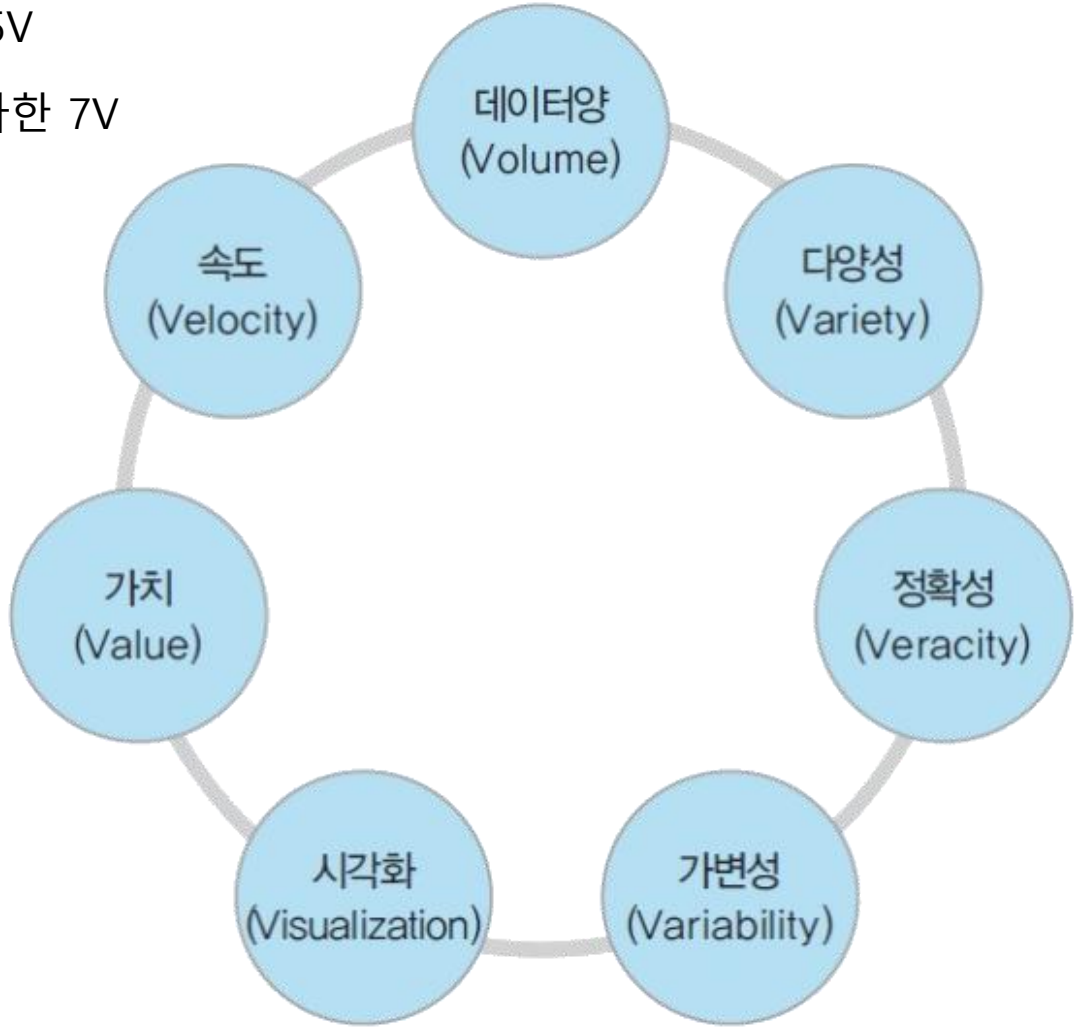
7V 모델

□ 빅데이터의 확장된 특징 : 7V

○ 활용 영역이 넓어지고 관련 기술이 발전하면서 빅데이터 특징도 확장됨

✓ 가치와 정확성을 추가한 5V

✓ 시각화와 가변성까지 추가한 7V



7V 모델

□ 빅데이터의 확장된 특징 : 정리

○ 가치(Value)

✓ 빅데이터 분석 결과는 의사 결정에 활용될 만한 유용한 가치를 가져야 함

○ 정확성(Veracity)

✓ 가치 있는 결과를 만들기 위해 빅데이터는 정확하고 신뢰할 수 있어야 함

✓ 가공 작업을 통해 높은 정확성을 유지하는 것이 중요

○ 시각화(Visualization)

✓ 빅데이터 분석 결과는 이해하기 쉽고 보기 좋게 그림이나 도표로 시각화 필요

○ 가변성(Variability)

✓ 빅데이터가 맥락에 따른 가변성이 있음을 인식하고 수집 및 분석 작업 시 데이터의 원래 의미가 그대로 반영될 수 있도록 노력해야 함

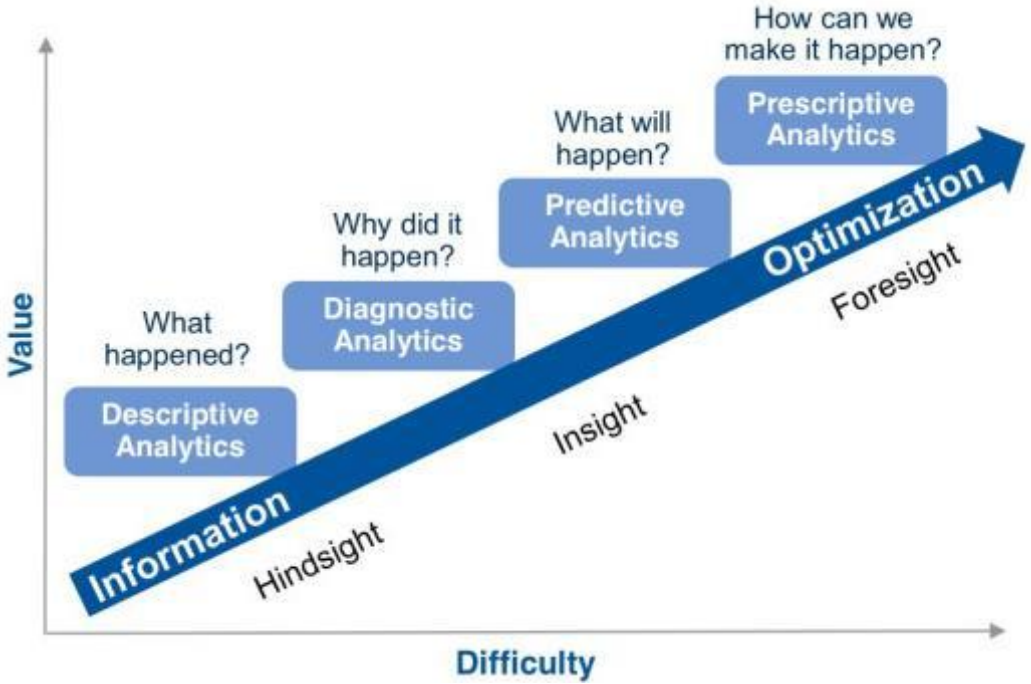
✓ 텍스트 형태의 비정형 데이터가 많아지면서 중요성 부각

7V 모델

빅데이터는 빠른 속도로 생성되는 대량의 다양한 데이터를 정확성을 유지하도록 가공하고, 맥락에 따른 가변성을 고려한 분석을 통해 가치 있는 결과를 도출해서 이를 이해하기 쉽게 시각화하여 제공하는 것이 중요!

데이터 분석 개념

- Data Set
  - 관련된 데이터들의 집합
- Data Analysis & Data Mining
  - 데이터를 분석하여 내재되어 있는 의미나 패턴을 찾아내는 과정
  - Data Analysis의 범위
    - ✓ Descriptive(서술), Diagnostic(진단), Predictive(예측), Prescriptive(처방) Analysis
- Data Analytics (데이터 애널리틱스)
  - 데이터의 수집과 정돈, 구성, 저장, 관리, 분석, 표현 등 전반적인 포괄 개념
  - Algorithms, SW Tools, Libraries, Techniques, Systems 들을 모두 포함



출처 : by http://www.rosebt.com/blog/

데이터 생성/활용 모델의 변화

□ 과거 모델

데이터 생산 : 일부 기업 등

데이터 소비 : 모든 사람

□ 새 모델

데이터 생산 : 모든 사람

데이터 소비 : 모든 사람

참고 : by Ruoming Jin

데이터의 유형

□ 정형 데이터 (Structured Data)
- ○ 관계형 데이터처럼 Schema에 따라 저장된 구조화된 데이터
- ○ 일반적으로 Table 형식으로 RDBMS에 저장
- ○ 예: 은행 계좌 정보 등의 데이터

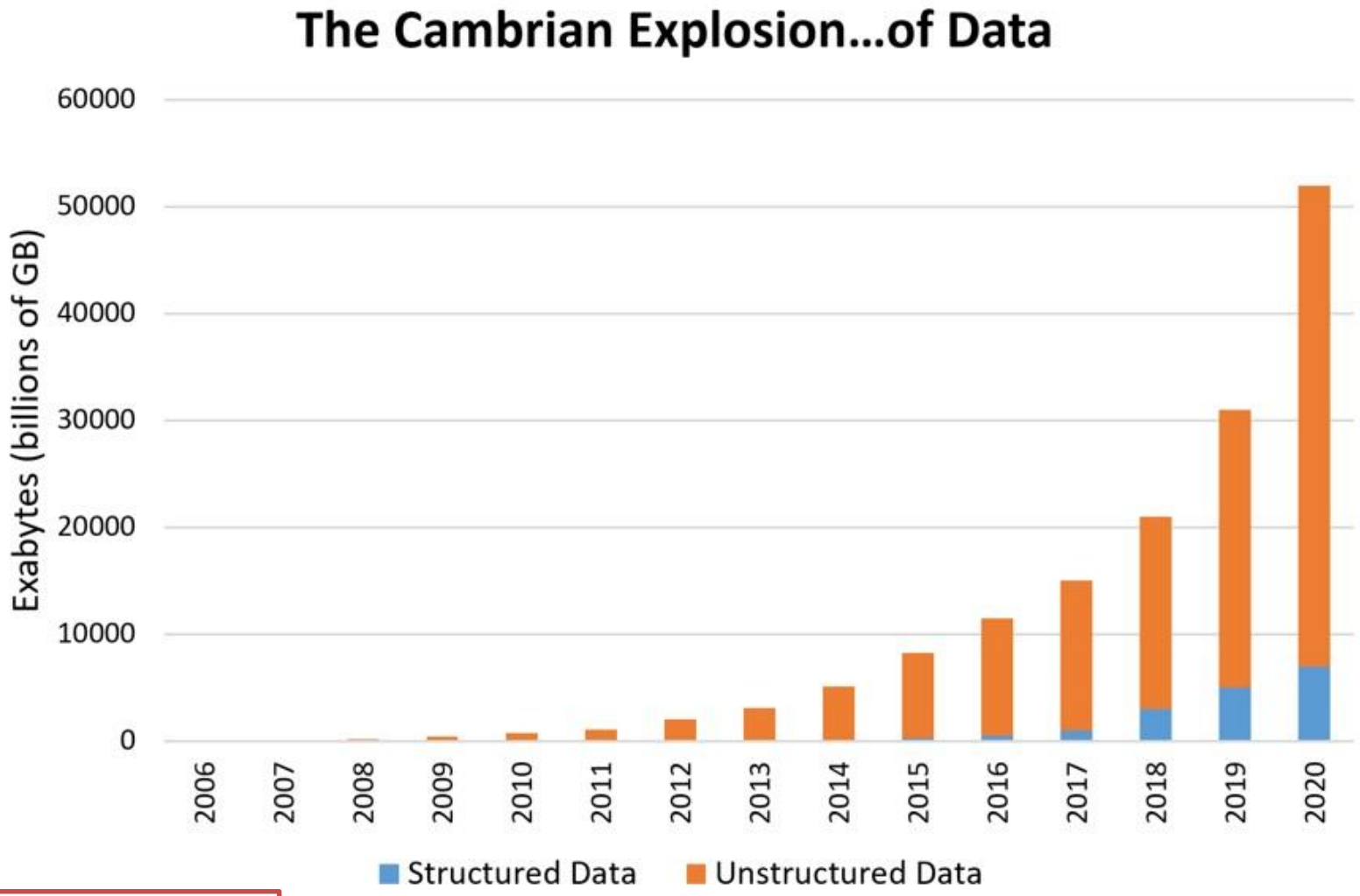
□ 비정형 데이터 (Unstructured Data)
- ○ Schema나 데이터 모델 없이 저장되는 데이터
- ○ 예: 일반 텍스트나 사진, 동영상 등의 데이터

□ 반정형 데이터 (Semi-Structured Data)
- ○ 고정된 필드를 가지고 있지는 않고 관계형 데이터 아님
- ○ 데이터에 Schema나 metadata와 같은 구조 정보를 포함하고 있어 일관성 유지
- ○ 계층적 또는 그래프 기반의 데이터
- ○ 예: XML, HTML 문서나 JSON, BSON 형태의 데이터

□ 참고 : **메타데이터 (metadata)**
- ○ data set의 특성 및 구조에 대한 정보 (data에 대한 data)
- ○ 자동으로 생성되어 data에 포함되는 경우가 일반적
- ○ 비정형이나 반정형의 빅데이터 분석에 주요 정보 제공
- ○ 예: 사진의 해상도, 작성일, 크기 정보, 문서의 작성일, 작성자, 버전 정보 등

데이터의 유형

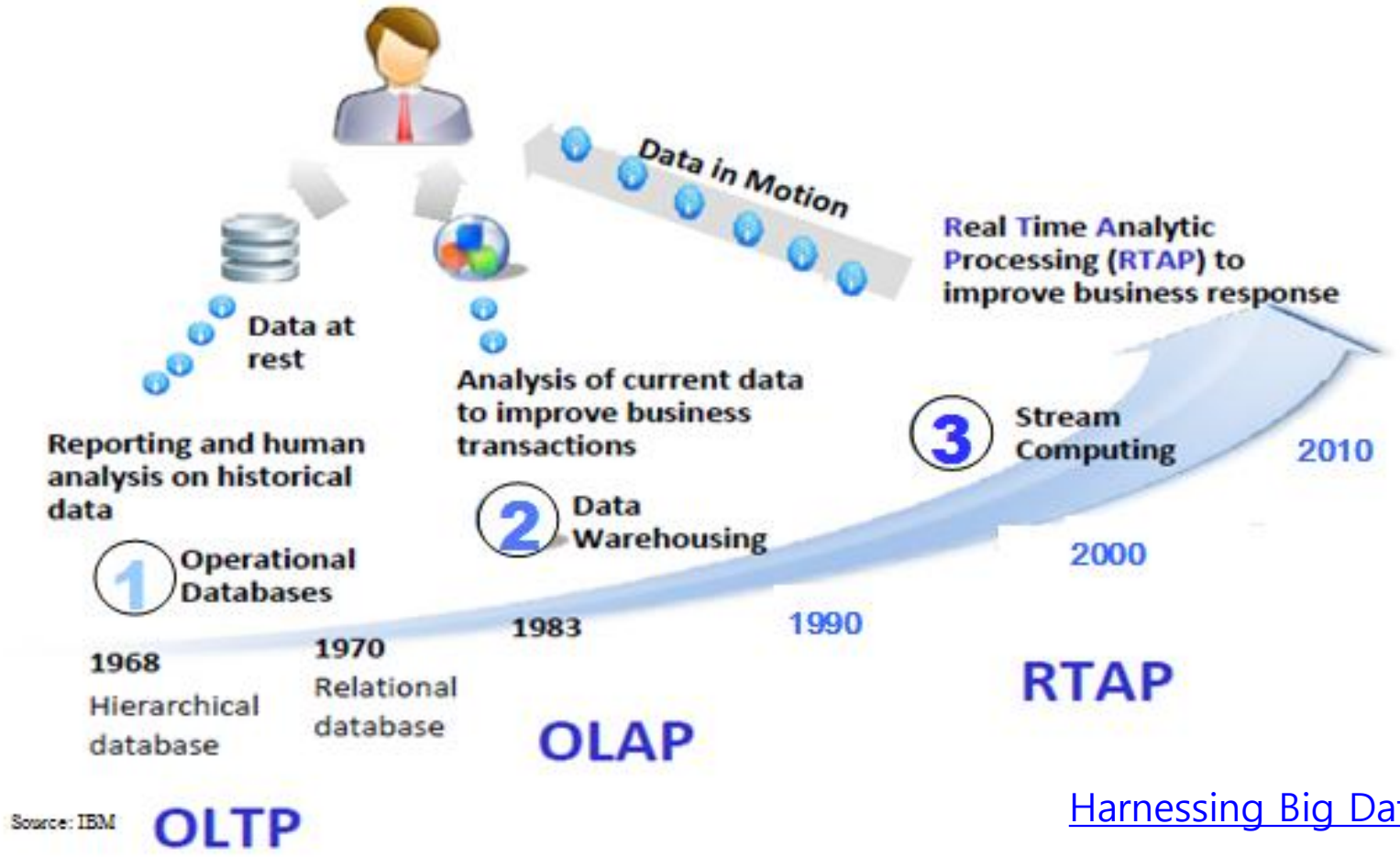
□ 비정형 데이터의 폭발적 증가



추세 참고용
정확도 있는 그림은 아님...

참고 : by Patrick Cheesman

Big Data의 이용



Harnessing Big Data

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

참고 : by Ruoming Jin

OLTP and OLAP 개념

□ OnLine Transaction Processing (OLTP)

○ Transaction 중심으로 데이터를 처리하는 시스템

○ Transaction

✓ 한번에 처리되어야 하는 단위 작업

✓ 비교적 짧은 처리 시간 (간단한 query)

✓ 예: 계좌이체 (2번의 데이터 변경)

○ RDBMS와 같은 저장 장치 활용

✓ 연산 : READ, WRITE, UPDATE, DELETE

John의 잔고는 얼마인가?

□ OnLine Analytical Processing (OLAP)

○ 다양한 각도 (다차원 분석 질의)에서

①사용자가 직접,
②대화식으로,
③다양한 도구의 지원으로 정보를 분석하는 과정

✓ 저장, 관리 : 데이터 웨어하우스 이용

✓ 연산 : READ 중심

✓ 분석 : Data Mining, Analytics, Decision Making 등 (복잡한 query)

– Drill down, Roll up, Pivot, …

✓ Reporting

A 상품을 구매한 사람 수를
1) 월별로
2) 지역별로
3) 나이별로 보여라

Data Lake, ETL 개념

Customer Relationship Management (CRM, 고객 관계 관리)
Supply Chain Management (SCM, 공급망 관리)
Enterprise Resource Planning (ERP, 전사적 자원 관리)

□ 데이터 레이크 (Data Lake)

- ○ 가공되지 않은 모든 종류의 데이터 저장하기 위한 중앙 집중식 저장소
- ○ 크기 제한 없이, 원시 데이터 형식으로 저장
- ○ 빅데이터에 활용 (...HDFS 파일)

□ ETL (Extract Transform Load; 추출-변환-적재) Process

- ○ Data Source로부터 원하는 Object로 데이터를 옮기는 처리 과정 (for DW)
  - ✓ ex) OLTP 스토리지로부터 OLAP 스토리지 (Data Warehouse, Data Mart 등)로 이동
- ○ 일반적으로 Big Data SW들은 ETL process에 해당하는 기능 포함
- ○ 과정
  - ✓ Extract
    - – OLTP, CRM, SCM, ERP database 등으로부터 필요한 데이터를 추출
  - ✓ Transform
    - – 규칙 등에 따라 데이터를 수정, 변형하여 원하는 형태로 변환
  - ✓ Load
    - – 대상 시스템에 데이터 적재
- ○ cf. ELT (적재 후, 대상 시스템에서 변환 수행)

Data Warehouse, Data Mart 개념

Customer Relationship Management (CRM)
Supply Chain Management (SCM)
Enterprise Resource Planning (ERP)

□ 데이터 웨어하우스 (Data Warehouse; DW)

- ○ 과거부터 현재까지 수집된 중앙집중적, 전사적 데이터 저장소
- ○ 조직 전체의 여러 소스들(ERP, CRM, DB, IoT, 파트너 시스템 등)로부터 데이터를 추출, 저장 (ETL)
- ○ 한번 저장되면 변경이 거의 없음 (cf. DBMS)
- ○ 저장 목적은 분석 (OLAP와 상호 작용)
- ○ Business Intelligence (BI)에서 주로 사용
- ○ DW의 대표적 특성
  - ✓ 주제 지향적 (Subject-oriented; 업무가 아니라 주제 중심)
  - ✓ 통합적 (Integrated; 전사적 데이터 표준화, 통일성)
  - ✓ 비휘발성 (Non-volatile; 갱신없는 조회 전용 데이터)
  - ✓ 시계열적 (Time Variant; 시간에 따른 변경을 항상 반영해야)

□ 데이터 마트 (Data Mart)

- ○ 데이터 웨어하우스에 저장된 데이터의 subset
- ○ 특정 부서, 사업팀 등 (고객 데이터 마트, 자재 데이터 마트, 재무 데이터 마트 등)

Business Intelligence 개념

주로 보고서나 시각화 중심의 개념
-> 무슨 일이 일어나는가?
(cf. 분석 포함 : Business Analytics 용어도 사용)

□ Business Intelligence

1) 기업이나 조직에서 생성된 데이터를 분석하고 -AND-
   - ✓ 데이터 웨어하우스 활용
2) 기업에서 필요로 하는 데이터의 의미 (Insights)를 획득하는 기술 및 분야
   - ✓ 관리자나 경영진이 활용할 수 있도록, 분석 결과의 표현 (시각화)
   - ✓ Dashboard나 Reporting tool 등 활용

○ 큰 의미로는 DB, DW, ERP등을 포함하는 개념으로 사용 (애플리케이션과 기술의 집합)

참고 : Dashboard

빅데이터의 처리 과정

□ **5단계 처리 과정 (이번 학기동안 설명할 내용)**

- ○ 수집, Acquisition (Ingest)
  - ✓ collecting, aggregating, moving the big data
- ○ 저장, Storage & Database
- ○ 처리, Processing
  - ✓ batch, streaming, distributed processing
- ○ 분석, Analysis
  - ✓ algorithm, library & system
- ○ 표현, Visualization

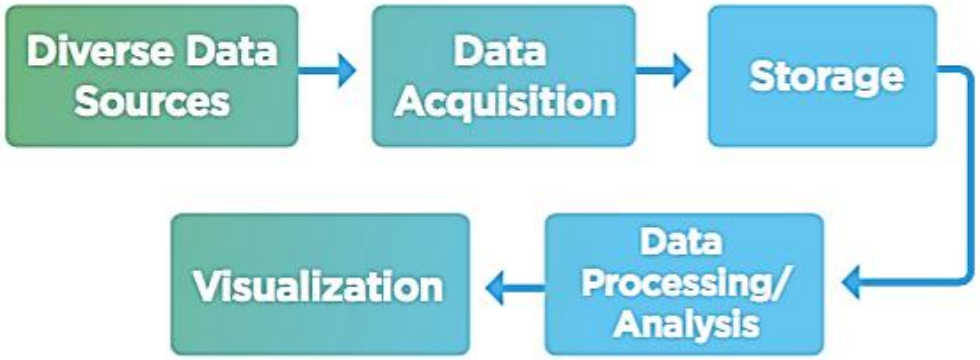


Fig. 1: Building blocks of a Big Data System

참고 : by Rohit Dhall

빅데이터의 처리 과정

□ Google Cloud Platform (GCP)의 4 Stage (처리 순서에 따른)

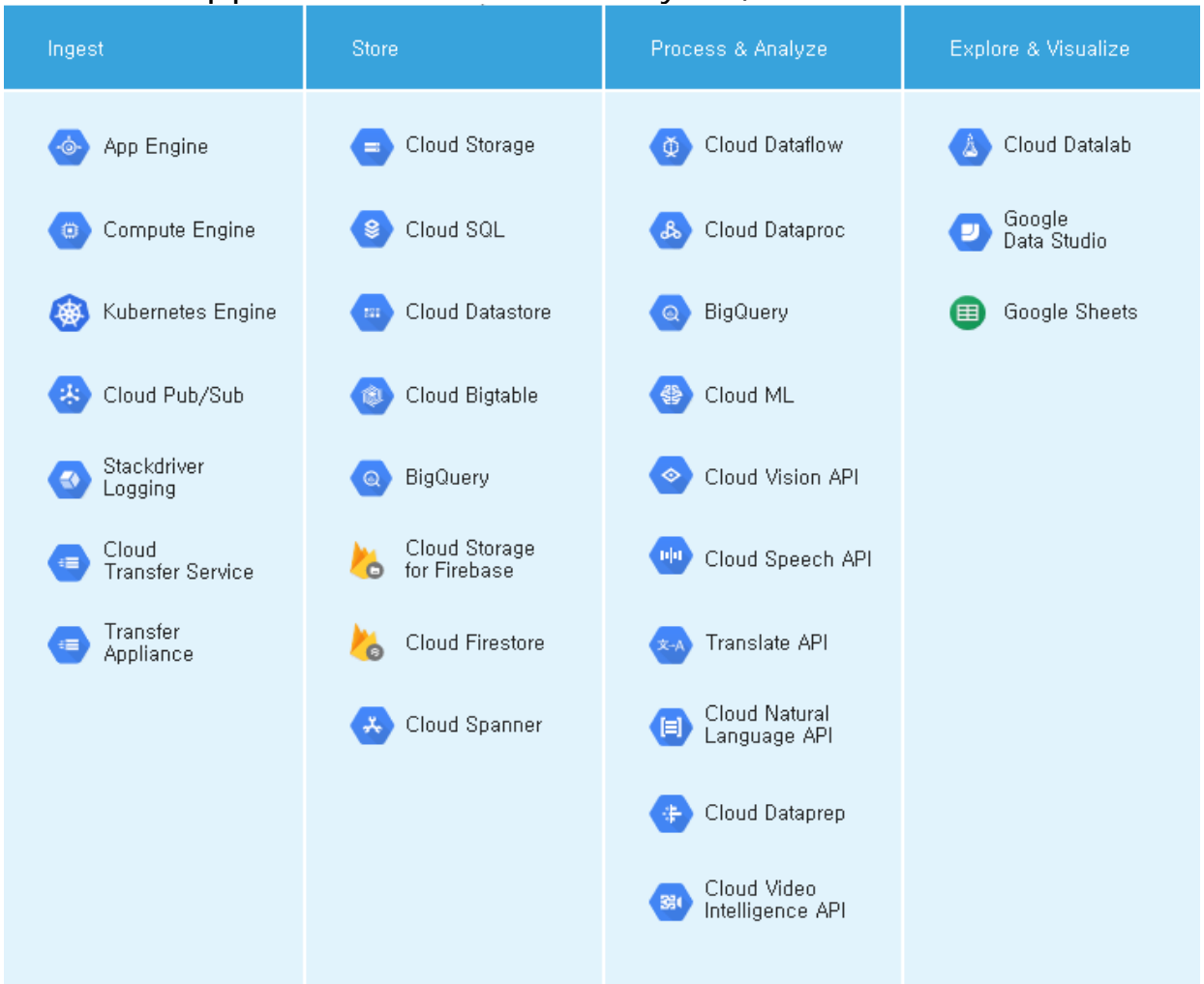
○ **Ingest (수집)**

✓ 원시 데이터의 수집 (streaming data from devices, on-premises batch data, application logs, or mobile-app user events and analytics)

○ **Store (저장)**

○ **Process and Analyze (처리와 분석)**

○ **Explore and Visualize (표현)**

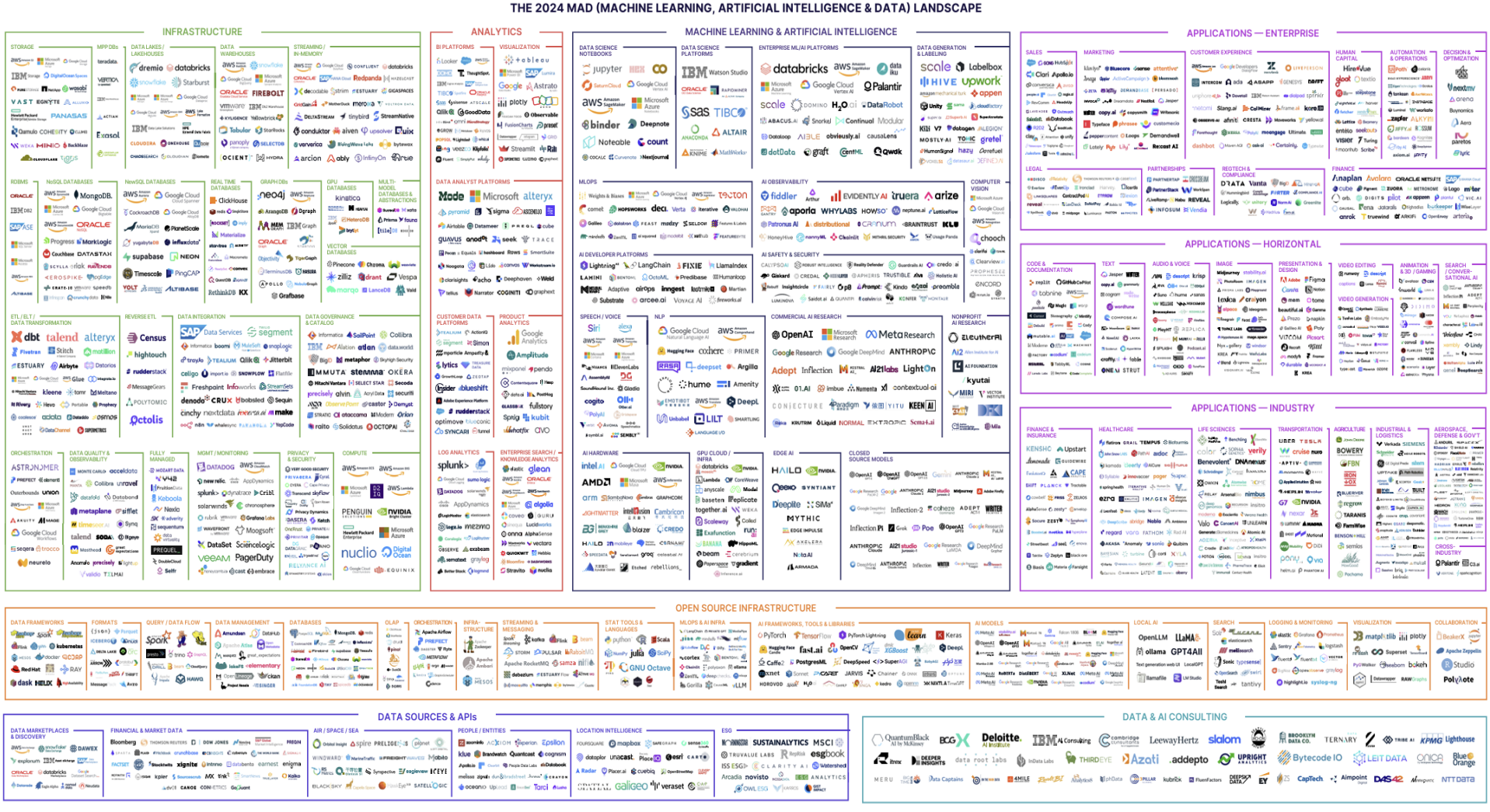


□ 물론,
다른 기준으로 분류도 가능

참고 : https://cloud.google.com

□ The 2024 Machine Learning, AI and Data (MAD) Landscape

- ○ big data 분야와 관련 분야 (Machine Learning, AI) 통합 반영 …
- ○ Website에서 part별 interactive version 참고



Version 1.0 - March 2024 © Matt Turck (@mattturck) , Aman Kabeer (@AmanKabeer11) & FirstMark (@firstmarkcap) Blog post: mattturck.com/MAD2024 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2024@firstmarkcap.com

FIRSTMARK

The 2024 MAD (Machine Learning, Artificial Intelligence & Data) Landscape

감사합니다.