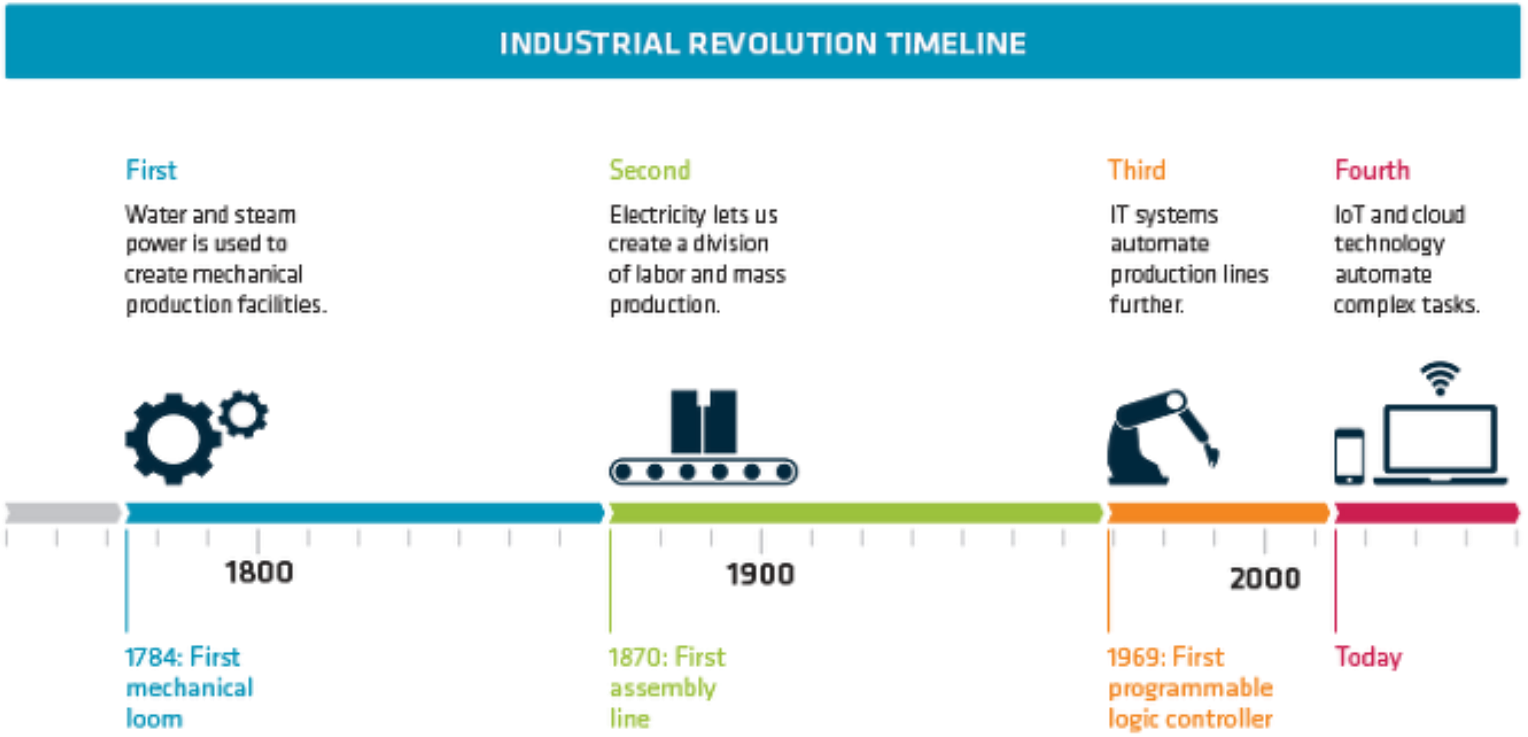
4차 산업혁명, 빅데이터 그리고 통계

불확실의 시대를 항해하는 히치하이커를 위한 가이드

4차 산업혁명?

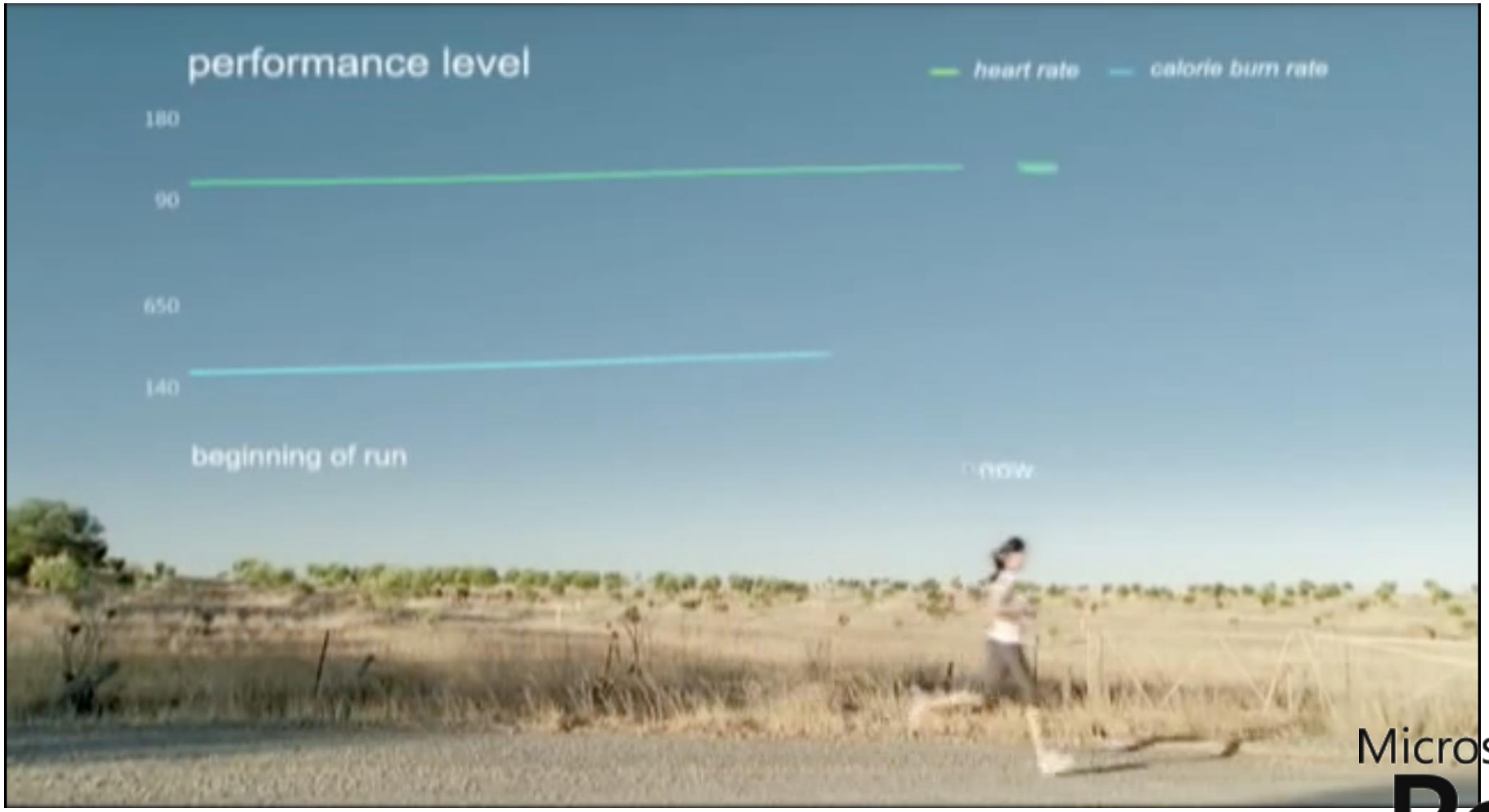
- 2016년 다보스 세계 경제포럼에서 알려진 개념
- 우리나라에서는 2016년 3월,
  이세돌 9단 vs 알파고의 바둑 대결을 계기로
  충격과 불안 속에서 맞이한 화두
- 정보통신 기술(ICT)의 융합으로 이루어낸 혁명 시대를 일컬음
- 6대 기술 : **인공지능**, 로봇공학, **사물인터넷**, 무인 운송 수단, **3차원 인쇄**, 나노 기술
- 제레미 러프킨(<The Third Industrial Revolution> 저자) : 현재 제 3차 산업혁명이 진행중이라고 주장

산업혁명의 구분



시기		관심의 중심
1차 산업혁명	기계화, 증기기관, 인쇄술, 석탄	기술효율화중심
2차 산업혁명	대량생산, 자동차, 전기, 석유	경영효율화중심
3차 산업혁명	전자화, 자동생산, 원자력, 재생에너지, 인터넷	지식효율화중심
4차 산업혁명	사물의 지능화(IoT)- Big Data - 인공지능, 신 재생에너지, 3D 프린터	개인&환경친화중심
5차 산업혁명(?)	인류를 지배하는 인공지능과 이에 맞서는 감성지능의 부각, 획기적 생명 연장	정신&형이상학중심

Health Future Vision



Health Future Vision

Microsoft Research

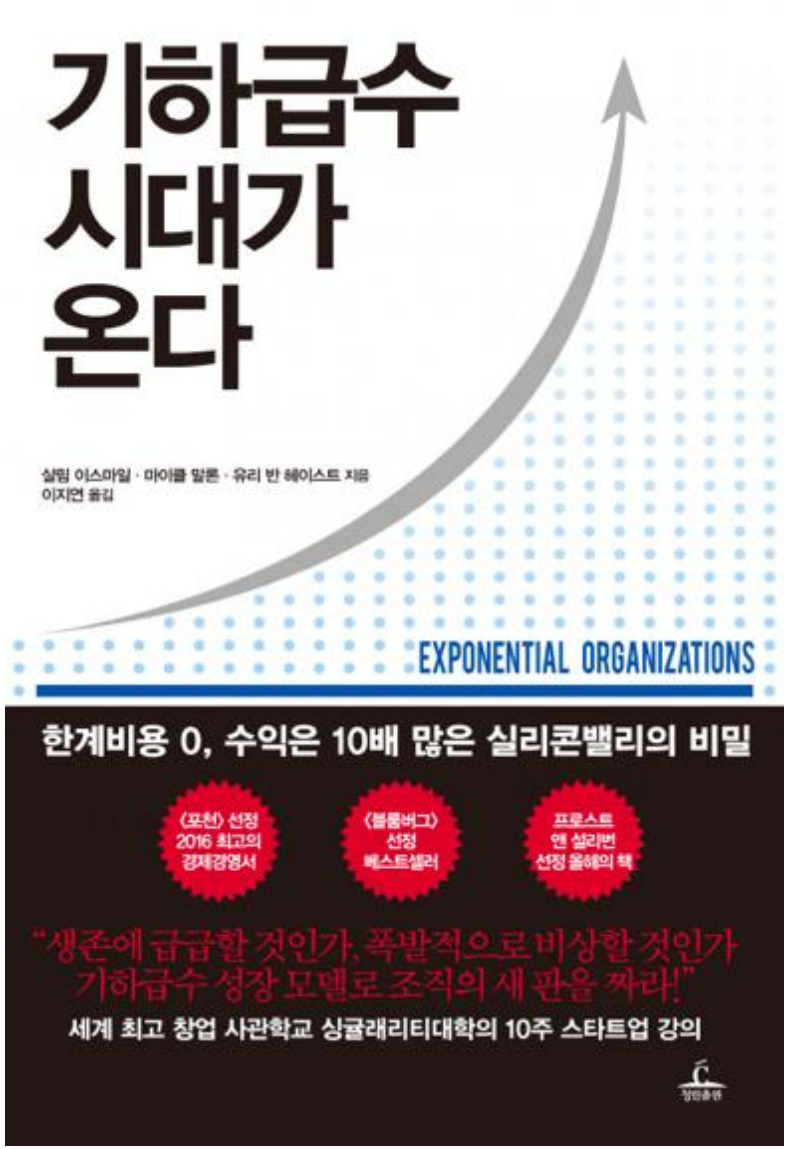
4차 산업혁명의 실체

AI(인공지능) + 초연결 + 3D 프린터

4차 산업혁명의 실체
= 기하급수적 시대

- 기하급수적 시대
  - 예측이 어렵다
  - 어느 순간 갑자기 성장속도가 빨라진다

- Unicorn (가치 10억 달러 기업)이 되는 시간
  - Google : 9년
  - Uber : 3년
  - Oculus : 2년

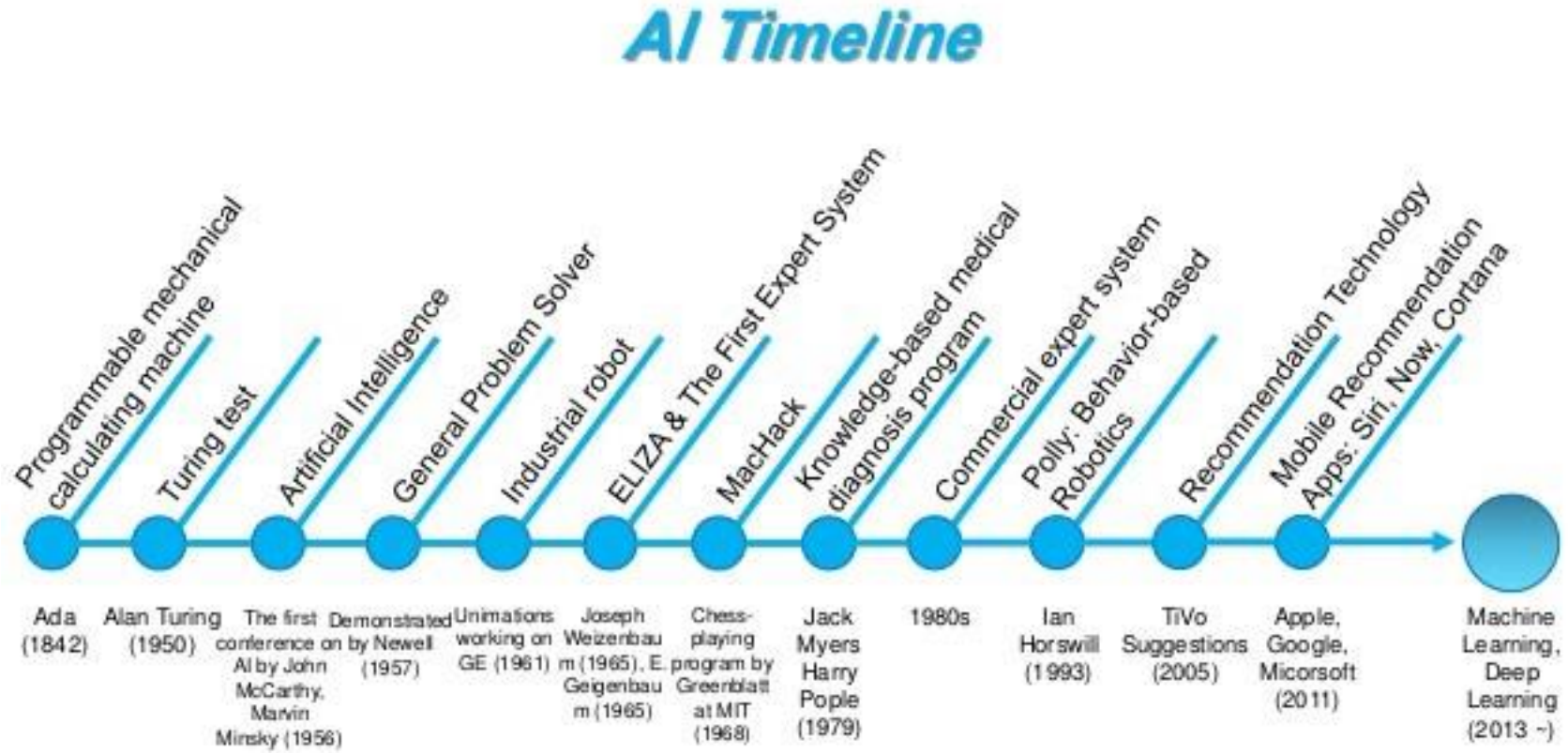


AI
: 인간을 닮은 기술

- Artificial Intelligence
  - 사전적 의미: 철학적 개념으로써 인간이나 지성을 갖춘 존재
    또는 시스템에 의해 만들어진 인공적인 지능을 의미
  - 컴퓨터 과학 관점의 의미
    수리적 모델을 이용하여 지적 능력을 연구하는 학문

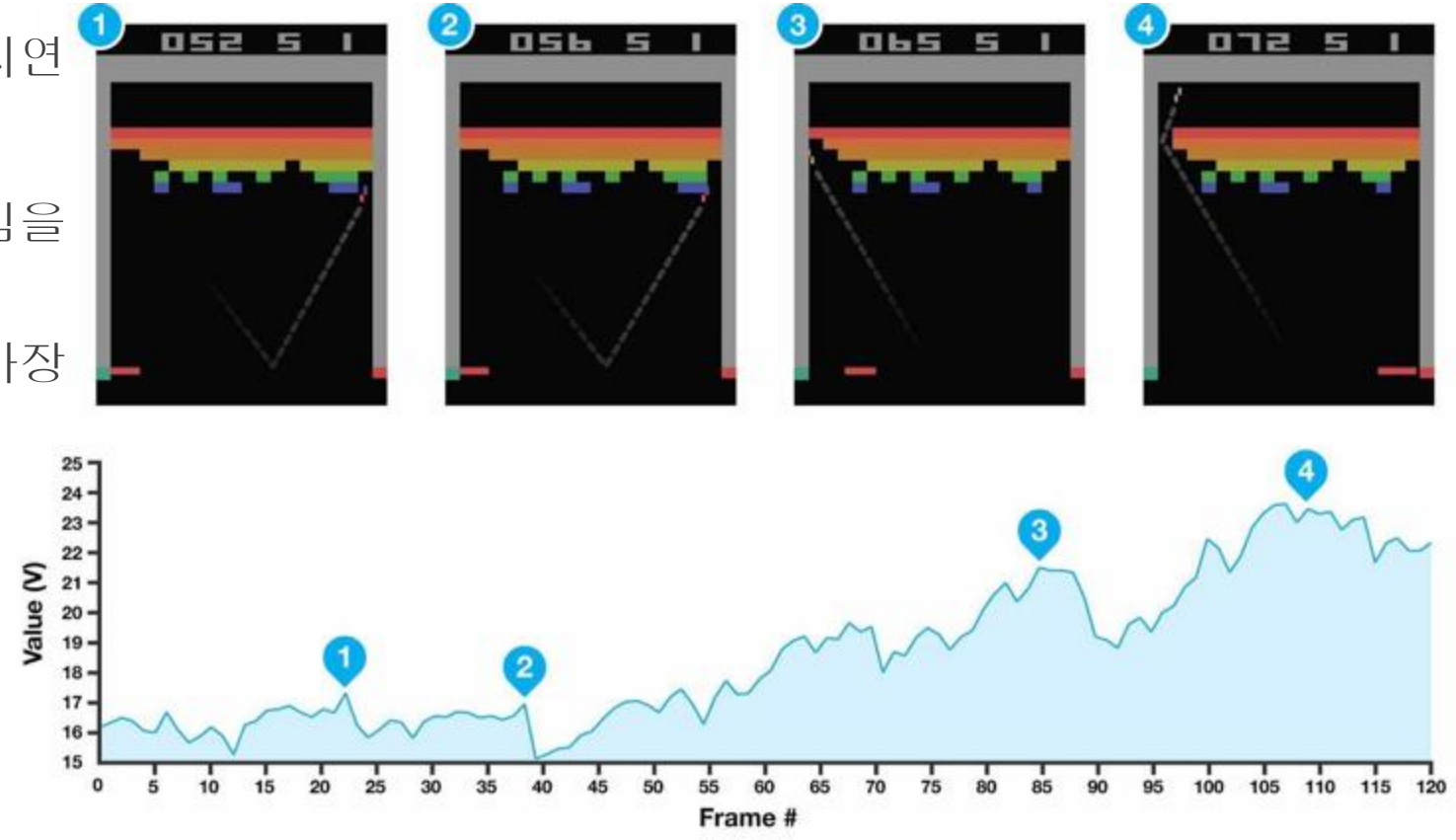
- 철학적 관점에서의 인공지능
  - 약인공지능(Weak AI)
  - 강인공지능(Stoing AI), 혹은 범용인공지능 AGI)

AI의 역사


AI Timeline
Programmable mechanical calculating machine
Turing test
Artificial Intelligence
General Problem Solver
Industrial robot
ELIZA & The First Expert System
MacHack
Knowledge-based medical diagnosis program
Commercial expert system
Polly: Behavior-based Robotics
Recommendation Technology
Mobile Recommendation
Apps: Siri, Now, Cortana
Ada (1842)
Alan Turing (1950)
The first conference on AI by John McCarthy, Marvin Minsky (1956)
Demonstrated by Newell (1957)
Unimations working on GE (1961)
Joseph Weizenbaum (1965), E. Geigenbaum (1965)
Chess-playing program by Greenblatt at MIT (1968)
Jack Myers Harry Pople (1979)
1980s
Ian Horswill (1993)
TiVo Suggestions (2005)
Apple, Google, Micorsoft (2011)
Machine Learning, Deep Learning (2013 ~)

기계가 학습을 시작했다
: Deep Q-Learning AI와 Breakout

- 딥마인드 인공지능의 벽돌깨기 학습 시연
- 기본적인 규칙과 목표만 알려주고 게임을 진행
- 게임을 진행하는 횟수가 거듭되면서 가장 효율적인 게임 방법을 스스로 찾아냄



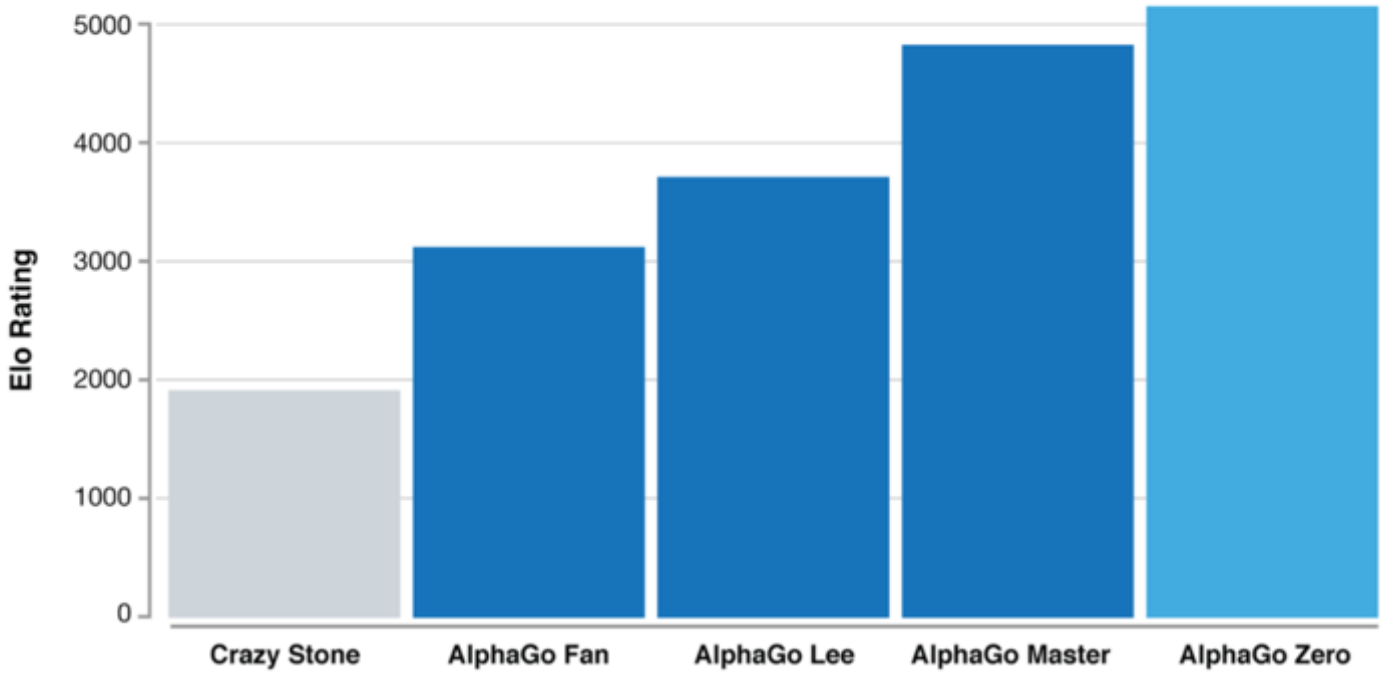
http://happinessbeyondthought.blogspot.kr/2015/06/wherehow-ai-is-beating-our-problem.html

https://www.youtube.com/watch?v=EfGD2qveGdQ&feature=youtu.be

AlphaGo Story
: 인공지능은 인간을 능가할 것인가?

Google DeepMind

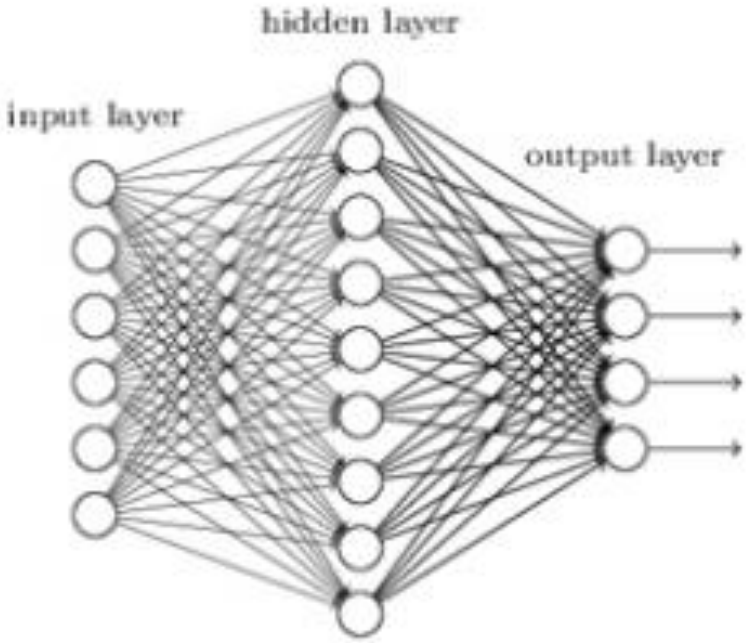
AlphaGo

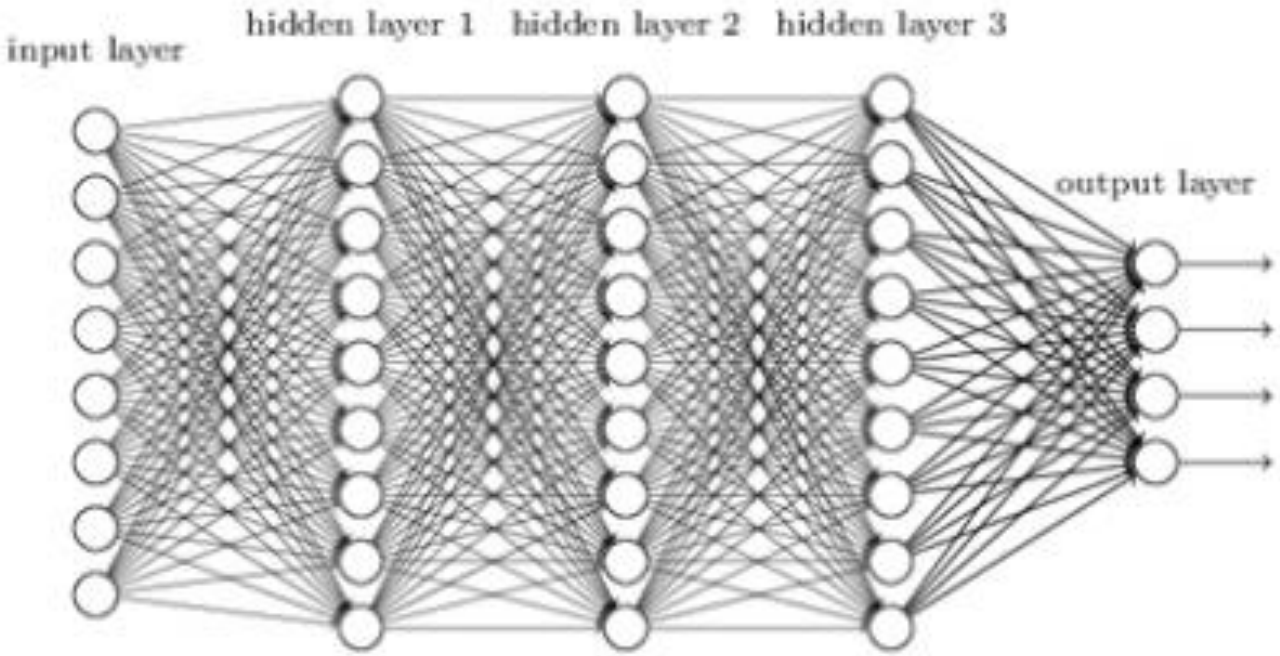


Elo ratings - a measure of the relative skill levels of players in competitive games such as Go - show how AlphaGo has become progressively stronger during its development

Machine Learning

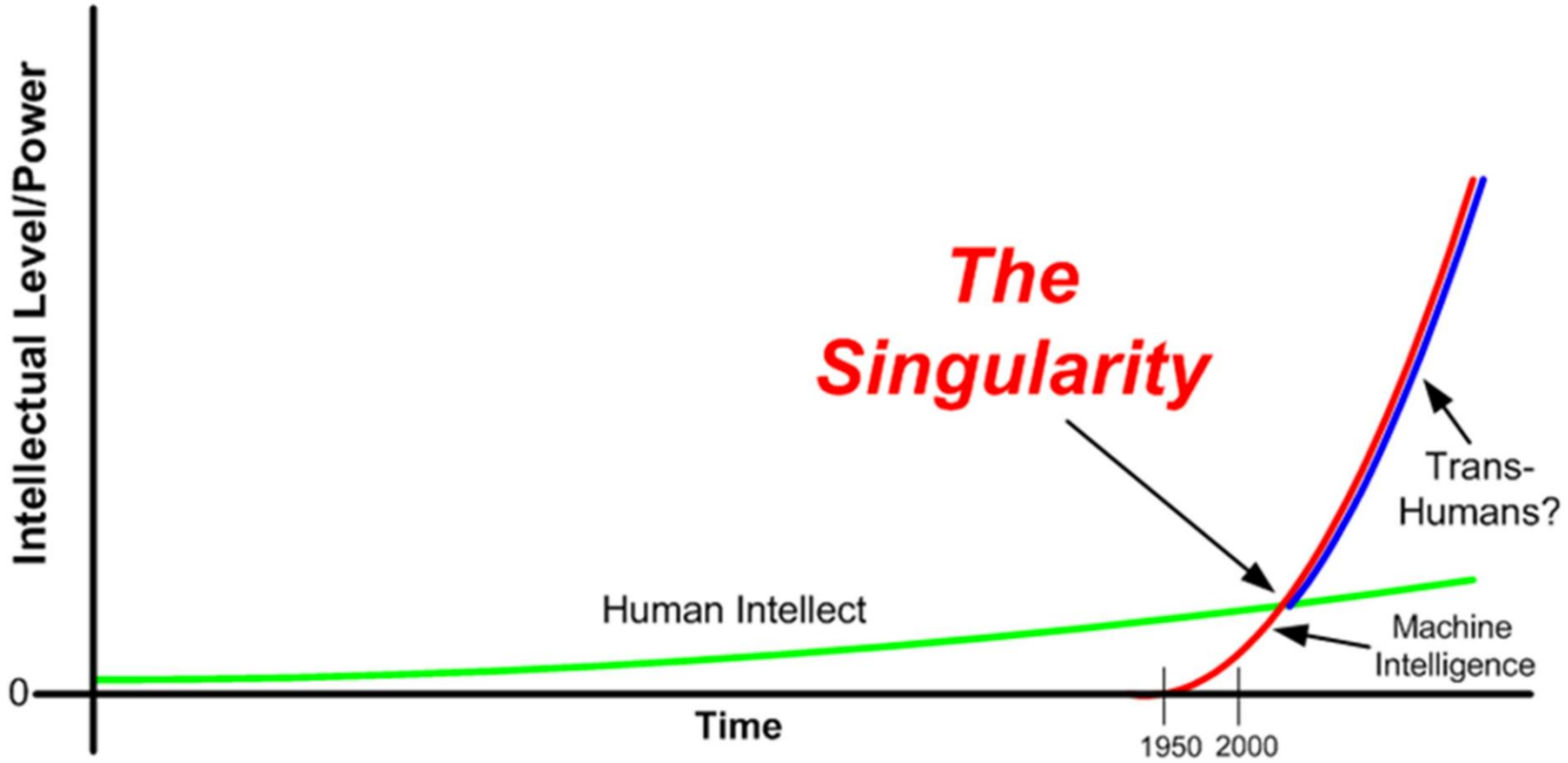
: Artificial Neural Network





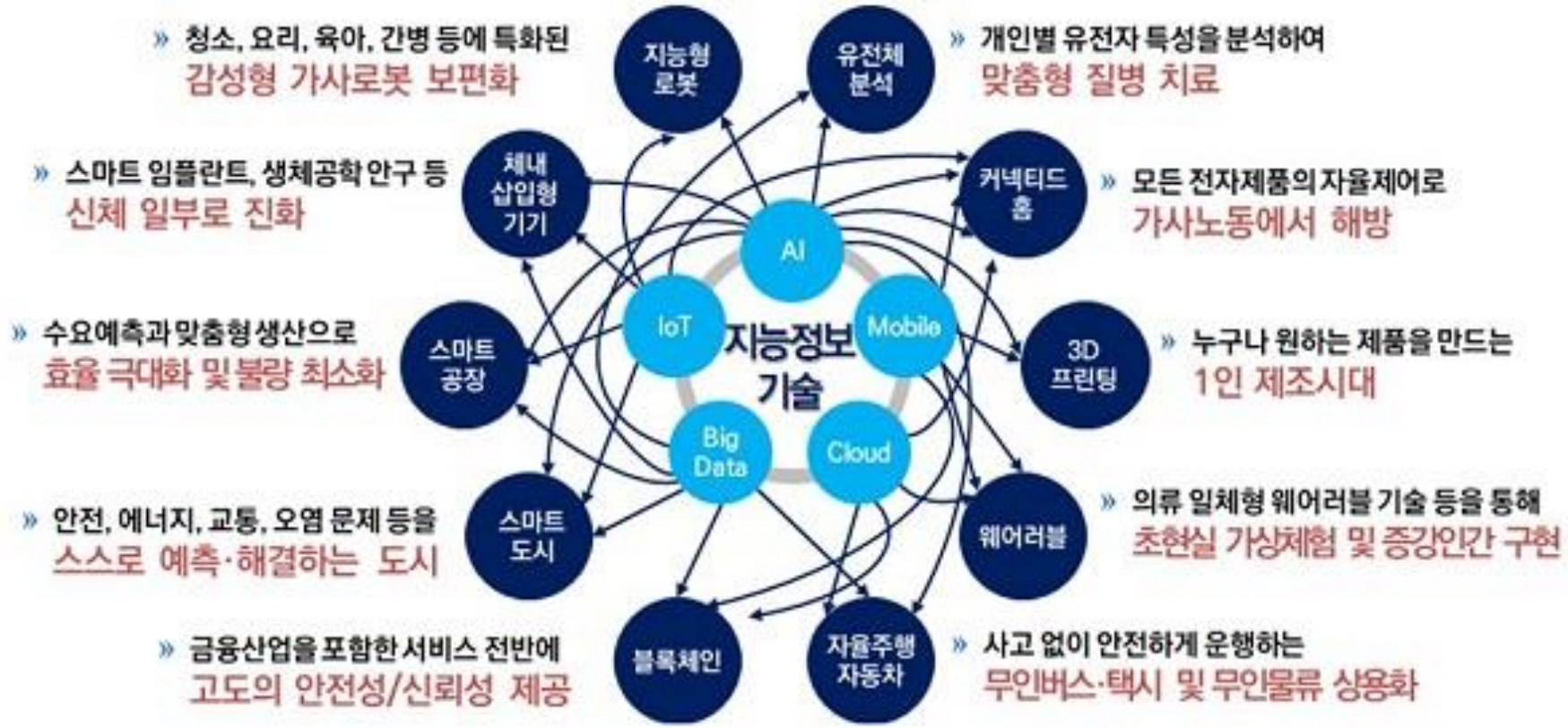
AI와 Singularity

특이점: 레이 커즈와일



초연결의 시대

- 초연결 사회 플랫폼
  - 신인류를 모음 (Phono Sapiens)
  - 새로운 시장의 개막
  - 벤처기업과 대기업의 콜라보레이션



Wearables

- 안경, 시계, 의복 등 착용할 수 있는 형태의 컴퓨터
- 궁극적 목표는 사용자가 거부감 없이 항상 착용하고 사용할 수 있으며 인간의 능력을 보완하거나 배가시키는 것
- 데이터의 입장에서 보면, 인간의 활동 영역에 대한 상세 정보나 신체 변화를 지속적으로 수집할 수 있다는 장점

IoT + Wearable
: Hoggies Tweet Pee의 사례

- 전통 제조업체 Hoggies의 IoT + Wearable 응용 사례

IoT + Wearable
: Nike Fuel Band + Nike+

- 전통 제조업체 Nike의 IoT + Wearable 응용 사례

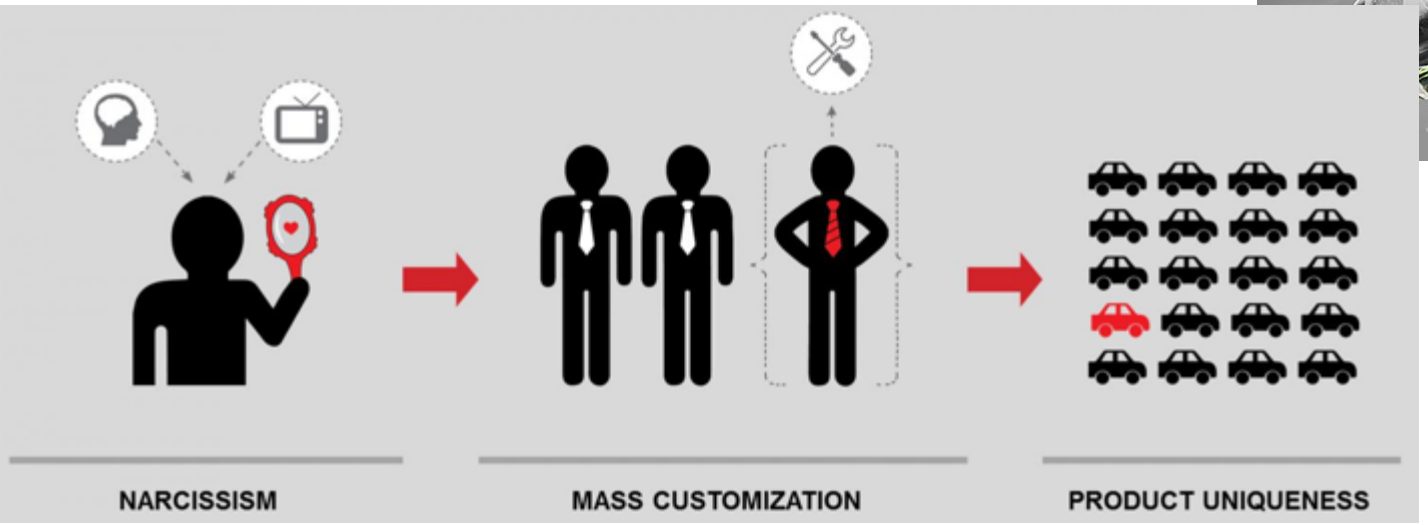
Manufacturing Future Vision

Manufacturing Future Vision

Microsoft® Research

Manufacture 2.0
: Mass Production에서 Mass Customization으로

- 3D 프린터가 바꾸는 제조업의 미래
- 기성품 대량 생산 대량 소비에서
  대량 맞춤 시대로의 이행
- 나이티(아디다스)의 3D Printed Shoes 사례
- 제조업의 민주화

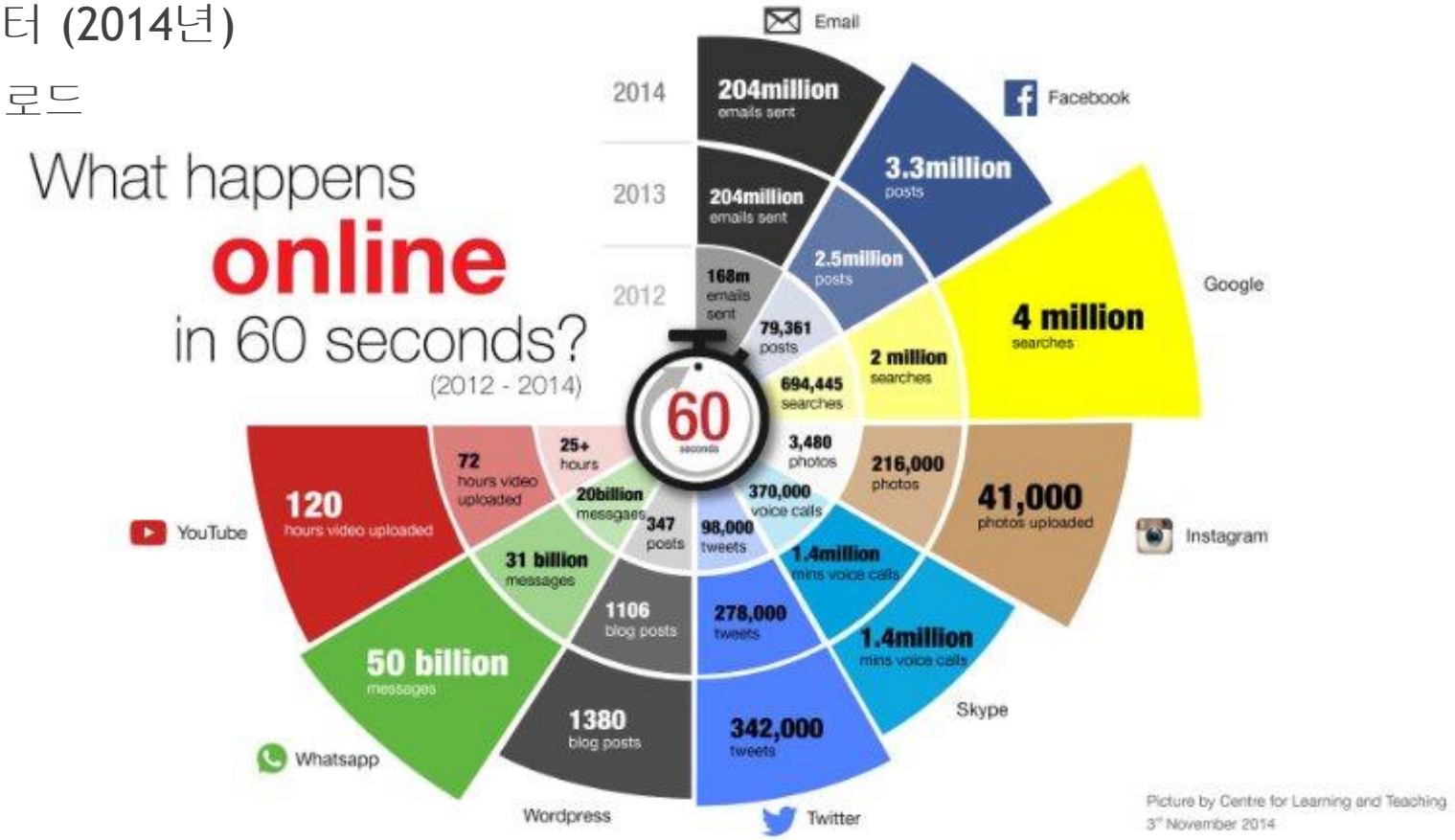


What is Nike Flyprint?

우리가 사는 세상
: 빅데이터의 출현과 개념

- 2010년 전후, 스마트폰의 대중화와 소셜 미디어(SNS)의 성공과 발전을 기점으로 데이터가 폭발적으로 증가

- 1분동안 인터넷에서 발생하는 데이터 (2014년)
  - 유튜브 : 120시간 분량의 동영상 업로드
  - 트위터 : 34만여 트윗
  - 페이스북 : 3천 3백여만 포스트
  - 인스타그램 : 4만여 건 사진 업로드
  - 이메일 : 2억여 이메일 전송



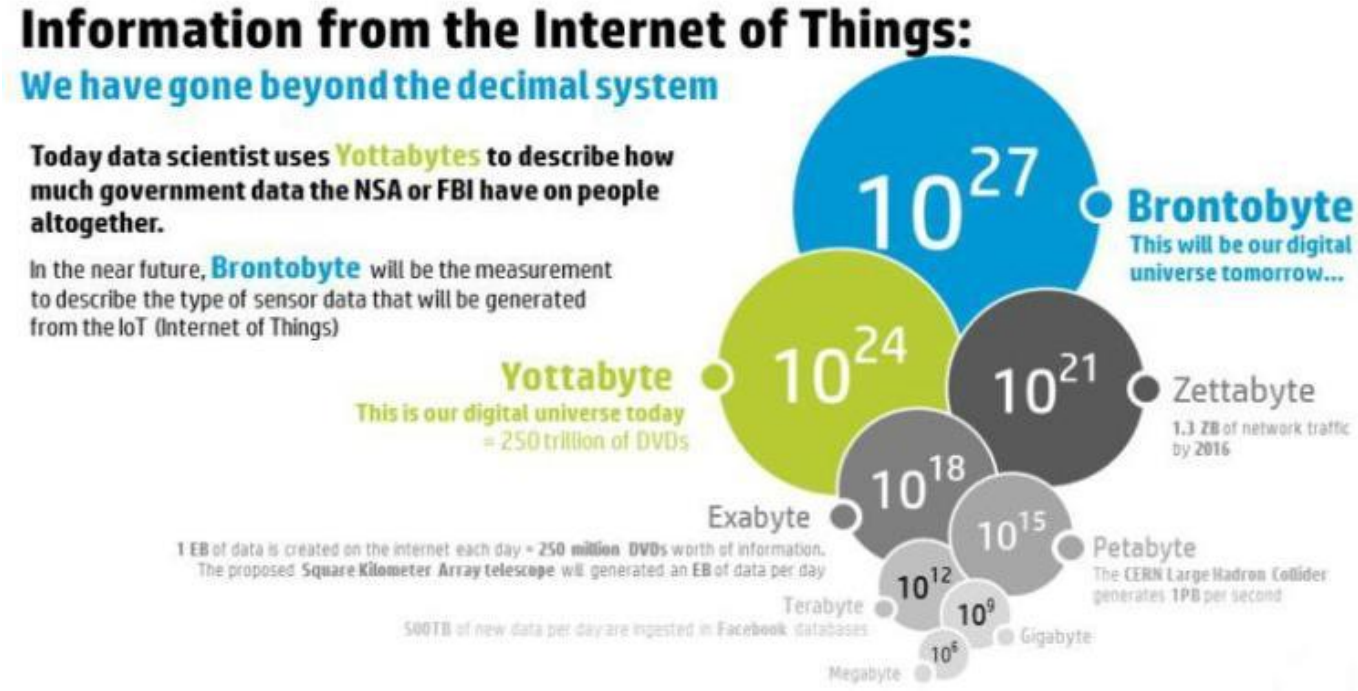
IoT와 데이터의 폭증

: 빅데이터의 출현과 개념

- 2016년 전후로 사람, 사물, 정보가 하나로 연결(융합)되는 4차 산업혁명이 시작됐으며 인공지능 + 사물 인터넷 + 무인 자동차 + 로봇산업 등 4차 산업 혁명 주요 기술들의 핵심 기반으로 빅데이터를 주목

- 인터넷 시대의 데이터
  - 센서 : 수치, 정량적인 데이터
  - 스마트폰 : 사람들의 행동 패턴을 측정

- 빅데이터 시대의 거대 빅소스
  = 소셜미디어 + 센서들의 인터넷



Global Data Growth
: 빅데이터의 출현과 개념

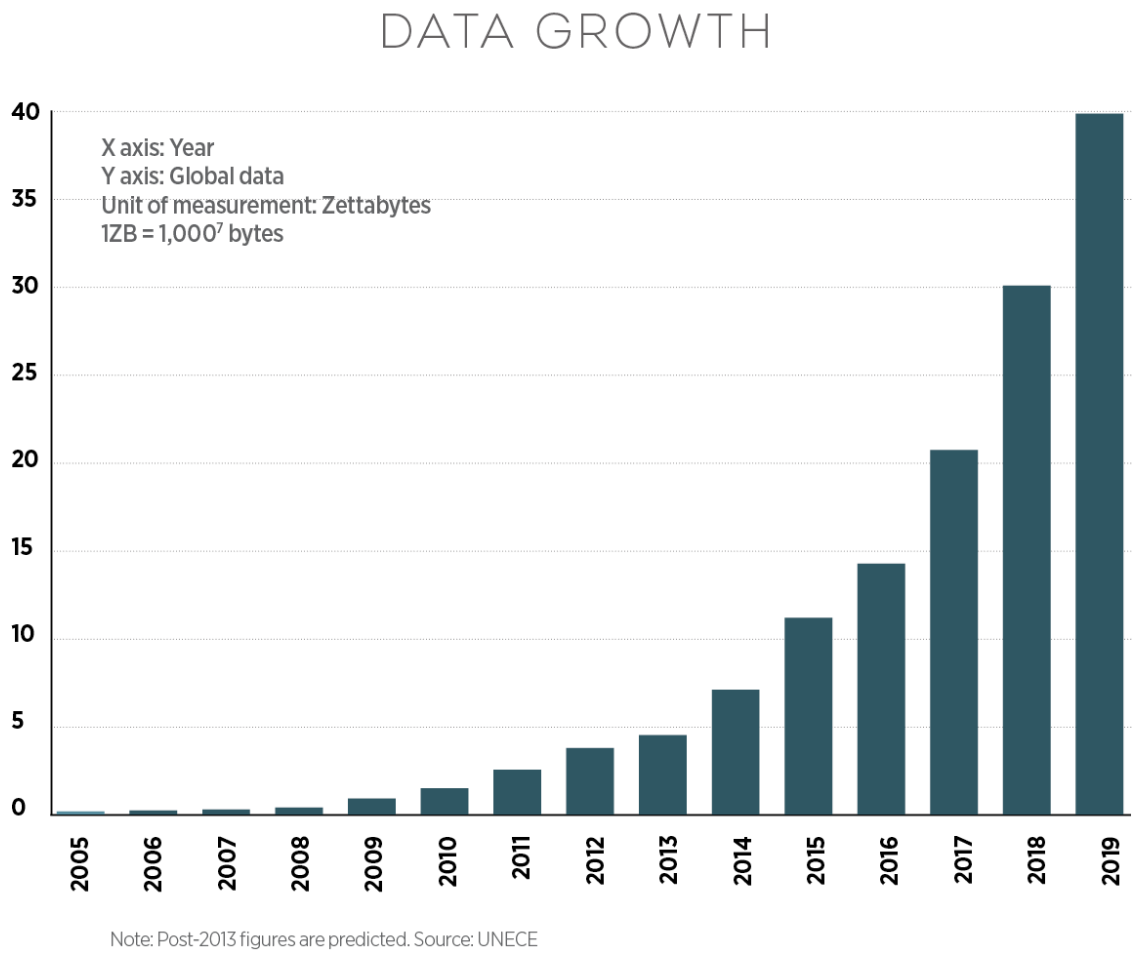
- 최근 발생한 데이터가 전 세계 데이터의 80%를 차지하며 2020년 전 세계 데이터는 44ZB까지 늘어날 전망(IDC, 2014)

- 데이터의 양의 방대한 증가와 함께 저장/처리를 위한 하드웨어 가격의 하락(인프라 발전)

- 데이터 전문가들은 비정형 중심의 데이터 증가를 '빅데이터' 라 부르며 인사이트(가치와 의미)에 대한 연구가 활발

- 하드웨어 인프라의 발전뿐만 아니라 빅데이터 시스템 기술 (소프트웨어 엔지니어링) 발전하고 있는 중

- 빅데이터 시스템 운영(분석 운영, 데이터 사이언스)의 성공 사례가 많아지면서 중요성이 산업, 사회, 미디어 등 여러 분야에 커지고 있음



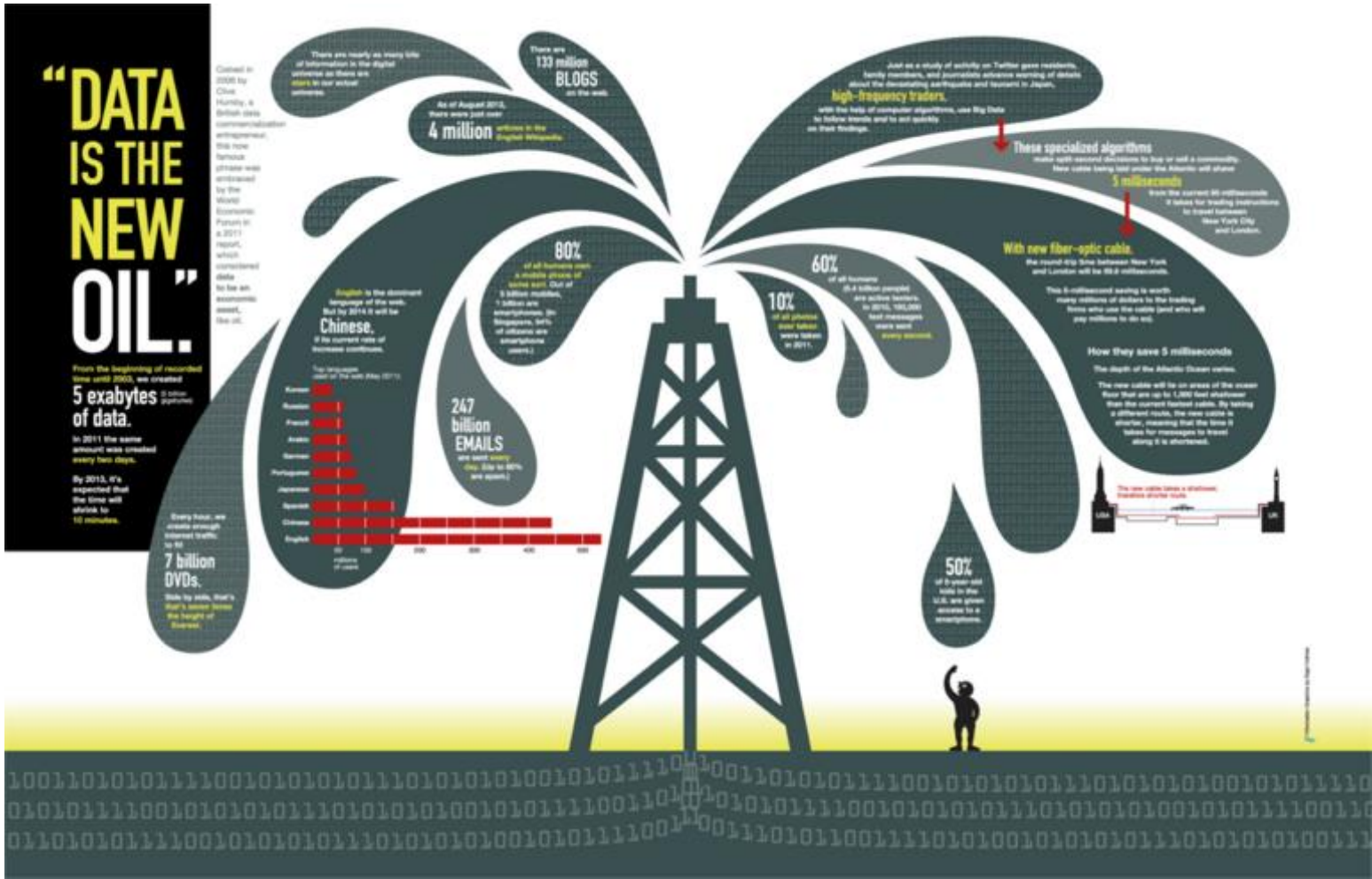
Note: Post-2013 figures are predicted. Source: UNECE

빅데이터의 시대
: Data is the New Oil

- 빅데이터
  - IT 업계 최대의 화두
  - 기업과 공공부문 빅데이터에 대한 투자 활발
  - 데이터 산업혁명 주장 대두
- 빅데이터 원전 3대 기업
  - 구글 : 검색엔진 + App 데이터
  - 아마존 : 전자상거래 데이터
  - 페이스북 : 소셜 데이터

”데이터는 21세기의 오일이며 이를 분석하는 것은 연소 엔진에 해당한다" - Peter Dondergaard, Gartner

”We need to find it, extract it, refine it, distribute it and monetize it”
David Buckingham



빅데이터의 시대
: 이론의 종말

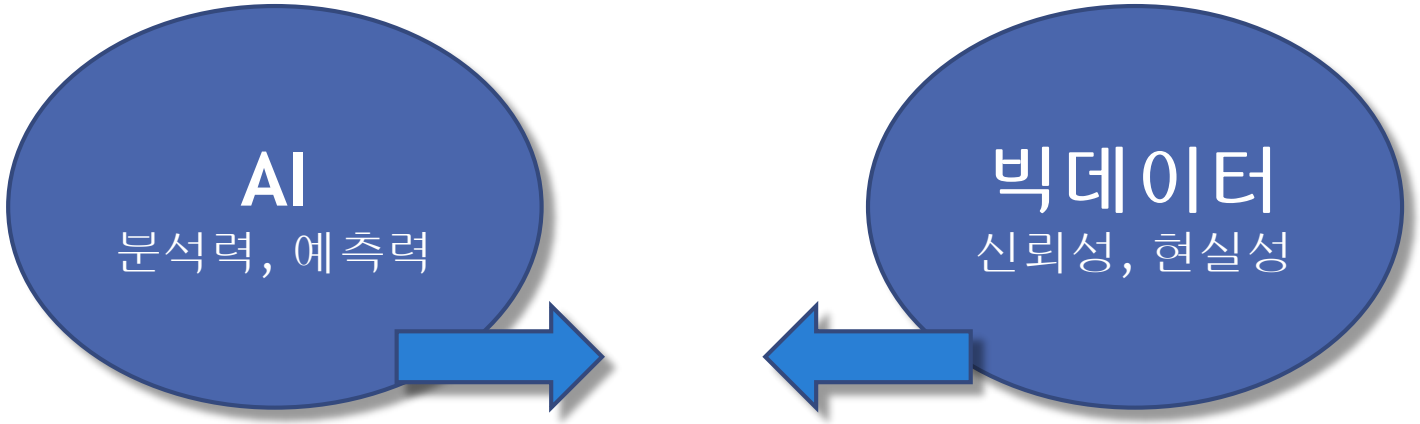
- 크리스 앤더슨
  - 빅데이터의 등장으로 전통적인 과학 연구방법론이 퇴색하고 있다고 주장
  - 인식의 한계치를 넘어선 데이터(팩트가 아닌 패턴)

- 날것 그대로의 데이터 -> 지식을 넘어 통찰로
  - 예전에는 처리 불가능했던 날것 그대로의 데이터를 저장-분석하여 패턴을 추출하면 지식의 영역을 넘어 지혜의 영역에 도달할 수 있을 것이라는 믿음

빅데이터가 AI를 만났을 때

- AI와 빅데이터가 만나면 상화 약점 보완 가능
- AI와 빅데이터는 상호보완적 역할을 수행, 시너지 기대 가능

- 빅데이터가 진정한 힘을 발휘하려면
  -> 데이터로부터 **패턴**을 발견, 그 속에서 **통찰**을 찾아내야 한다
- 일반적으로 데이터에서 가치를 찾아 지혜와 통찰로 발전시키는 것은 의사결정자(Decision Maker)의 몫이지만 잘 훈련된 AI에 빅데이터를 전달하면 **"통찰 비슷한 것"**을 얻을 수 있다

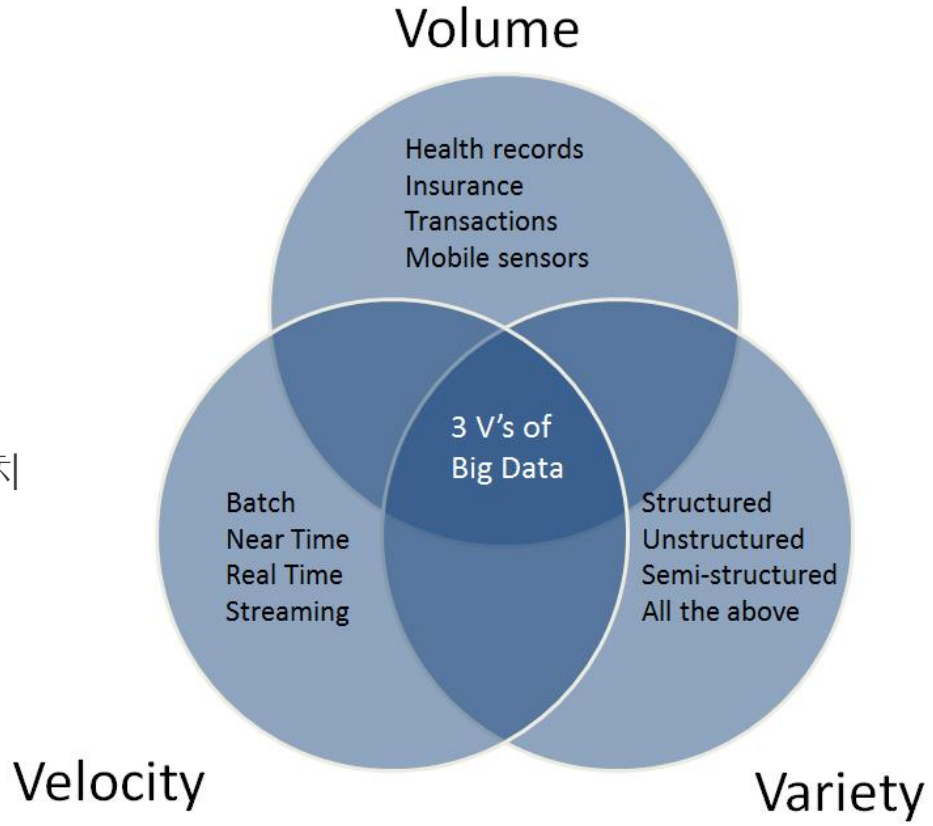


빅데이터의 정의

- 다양한 전문가(전문기관)의 정의
  - 서버 한 대로는 처리할 수 없는 규모의 데이터
    - 아마존 데이터 과학자 존 라우저(John Rauser), 2012년 아마존 클라우드 컨퍼런스
    - 가지고 있는 데이터 처리를 위해 분산 환경이 필요한가?
  - 기존의 SW(일반적인 데이터베이스)가 처리(저장/관리/분석)할 수 없는 규모의 데이터
    - 맥킨지(McKinsey), 2011년 5월 'Big data: The next frontier for innovation, completion, and productivity)
    - 기존 데이터베이스는 많은 경우 분산 환경을 염두에 두지 않고 만들어진 소프트웨어
    - 빅데이터 시스템은 스케일 업(Scale-up)보다는 스케일 아웃(Scale-out) 방식을 선호
  - 향상된 인사이트(Insight)와 더 나은 의사결정을 위해 사용되는 비용효율이 높고 혁신적이며 대용량, 고속, 다양성의 특성을 가진 정보
    - 가트너(Gartner) 2011년 1월, 'Big Data Analytics'
  - 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, (데이터의) **초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처 "**
    - IDC 2011년 보고서 'Extracting Value from Chaos'

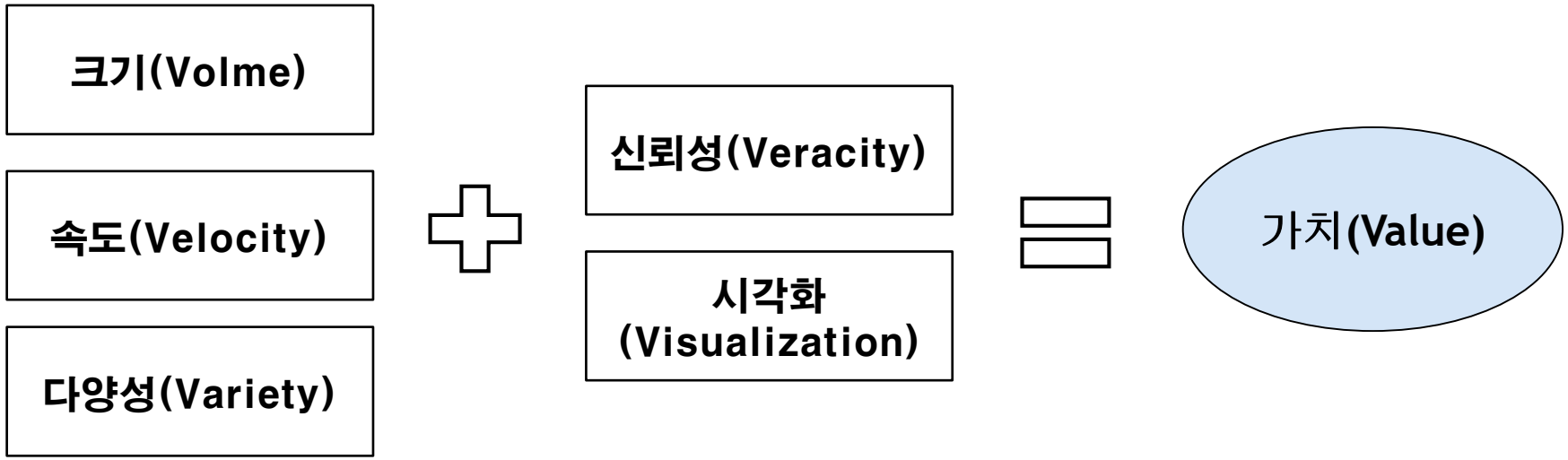
빅데이터 3대 요소
: 3V

- 2011년 가트너 애널리스트 더그 레이니(Doug Laney)
- Volume (규모)
  - 데이터의 급증은 빅데이터의 큰 특징
  - 적게는 PB에서 많게는 ZB 이상을 기준으로 보지만 정량적으로 정해져 있지는 않음
  - 데이터의 양은 산업별, 규모별 차이가 있지만 대량이라는 점에서 일치
- Velocity (속도)
  - 변화의 속도 또는 유통의 속도가 빠른 데이터
  - 주식, 환율, 항공 경로 등 매우 짧은 시간 내에 계속 변경되는 데이터
- Variety (다양성)
  - 소셜 데이터, 위치 데이터, 텍스트, 센서 데이터, 비디오, 오디오 등 다양한 형태의 데이터들이 발생
  - 어떻게 다양한 데이터를 수집, 저장, 처리, 분석하느냐가 이슈로 등장



빅데이터 개념의 확대
: 6V

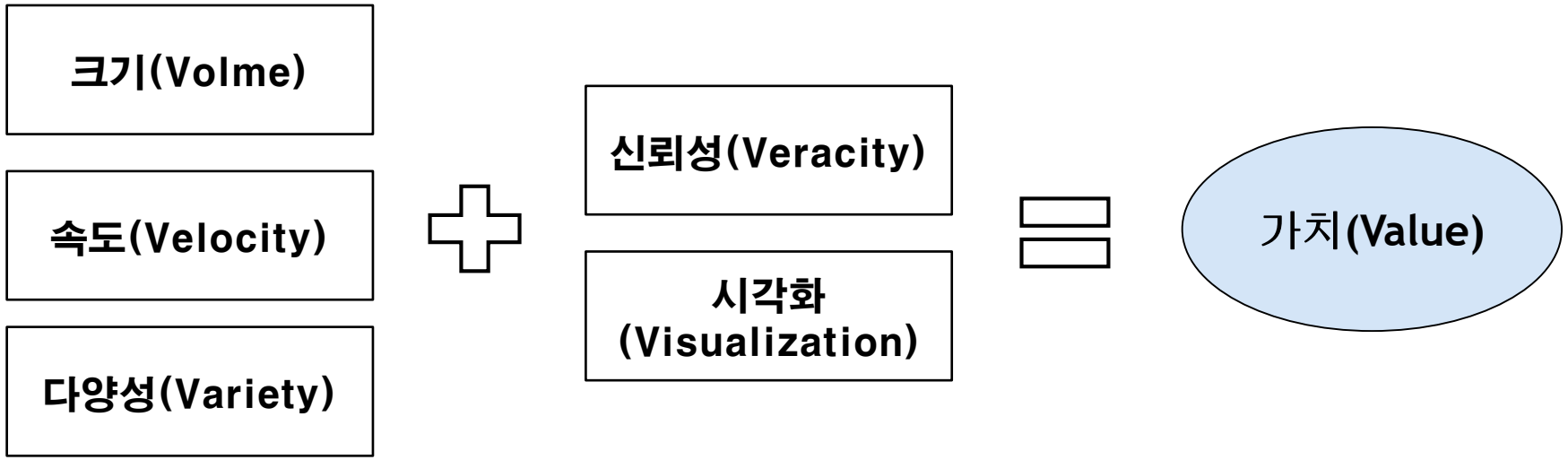
- 3V 정의에 따른 빅데이터
  = 대규모, 고속의 다양한 데이터를 분석하여 인사이트(Insight)와 가치(Value)를 주는 기술
- 이후 IBM은 Veracity(신뢰성)을 추가, 4V 정의를 내림
- 현재는 6V까지 확장된 개념이 널리 통용
  - 3V(Volume, Variety, Velocity) : 빅데이터의 본질(정의)
  - + 2V(Veracity, Visualization) : 빅데이터 분석 개념
  - + Value : 궁극적 목적



빅데이터 개념의 확대
: 6V

- 6V 정의에 따른 빅데이터

  = **대규모(Volume)**로 **빠르게(Velocity)** 발생하고 있는 **다양한(Variety)** 데이터를 수용하고
  정확한 분석을 통하여 **신뢰성(Veracity)**을 확보하고 **시각화(Visualization)**하여
  새로운 **가치(Variety)**를 창출하는 기술



빅데이터 활용 사례
: Case Study – [제조] GE Aviation

- GE Aviation : 항공기 엔진 제작 업체
- 제작 엔진에 다수의 센서 부착,
  정기적으로 데이터를 서버로 전송
  엔진의 상태를 분석,
  - 점검 및 교체 시기 판단
- 항공사를 대상으로 "엔진 유지보수 대행"

데이터 분석을 활용한 제조업에서 서비스업으로의 사업영역 확장

GE Aviation

빅데이터 활용 사례
: Case Study – [유통] Amazon

- 데이터를 가장 잘 활용하는 기업 중 하나

- 상품 추천 서비스
  - 고객의 구매 패턴을 분석, 향후 구매도가 높을 것 같은 제품을 추천
  - 관심 제품 추천 서비스를 통한 매출이 아마존 전체 매출의 절반에 육박

- 신선제품 배송 서비스
  - 고객의 구매 패턴을 분석, 미래의 구매를 예측, 유통기한이 짧은 신선식품을 미리 고객의 근거리 창고에 배치

- 자사에서 연구, 개발한 추천 로직, 모델 및 구현 관련 서비스들을 자사의 클라우드 서비스를 통해 제공 (Amazon Web Service)

amazon.com

빅데이터 활용 사례
: Case Study – [유통] Target

- Target : 1800개 매장을 보유한 미국 내 2위의 대형 할인점

- 고객의 구매 패턴을 분석, 구매 패턴의 변화가 발생하면 인생의 중대사가 발생했음을 예측
  - 해당 상황에 걸맞는 상품의 카탈로그를 고객에게 전송
  - 사례: 임신 지수 모델(Pregnancy Model)

  - 고객의 불안감과 불쾌감을 고려, 예측에 적합한 상품 구매 쿠폰과 예측과 전혀 상관 없는 상품 구매 쿠폰을 섞어서 발송

TARGET

빅데이터 활용 사례
: Case Study – [금융]

- 결제 사고 예방 시스템
  - 사용자의 결제 패턴을 분석 사기 결제 여부 판단
  - 금융 사기 탐지, 소비자 보호

- 향후 경제 환경 예측
  - 10년치 BIS 중앙은행 총재 연설문의 단어 빈도 분석
  - 세계 경제 흐름 예측
  - 시장 변화에 선제 대응

- Robo Adviser를 활용한 투자 상품 추천 및 운용
  - 과거 주가 데이터를 분석, 패턴을 추출 투자자의 투자 성향에 맞는 상품 추천
  - 데이터를 기반으로 한 향후 주가 예측 및 투자 상품 운용

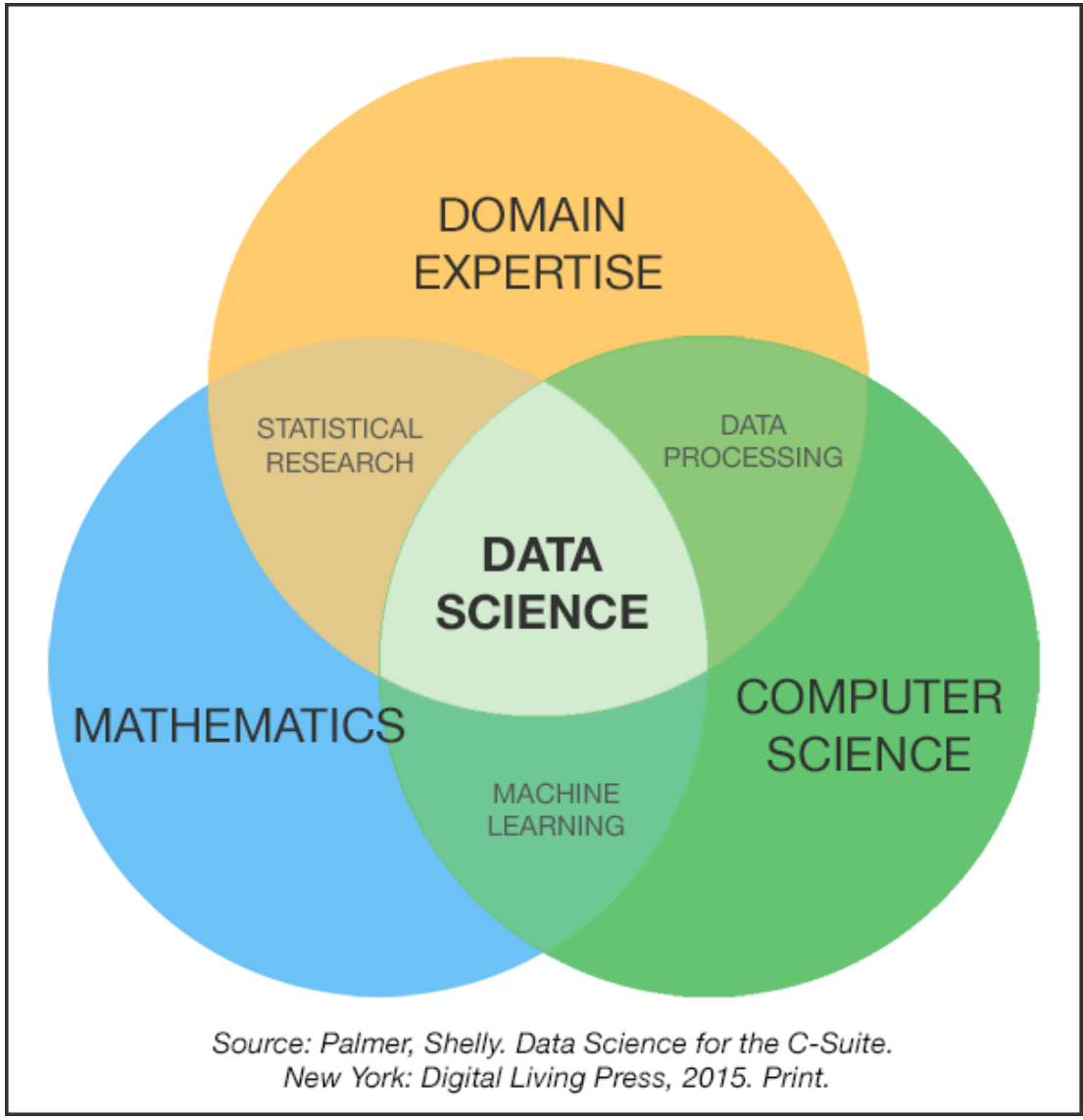
데이터 사이언스

- 데이터 사이언스는 단일 학문이 아님
  - 도메인 지식
  - 수학 및 통계
  - 컴퓨터 과학의 융합형 학문

- 데이터 사이언스 : 데이터를 과학적으로 다루는 일
  - 데이터 사이언스의 일

가치를 더할 수 있는 일을 찾고
데이터를 이용해서 문제를 해결하는 것

데이터로부터 머신러닝, 컴퓨터, 통계 등의
기술을 활용, 인사이트를 도출하는 일



데이터 사이언스 밴 다이어그램

데이터 사이언스
: 데이터 사이언스와 데이터

- 데이터
  - 기록 또는 자료
  - 관찰이나 측정을 통해 수집한 사실 또는 값
  - 예전 : 대부분의 데이터가 버려졌다
  - 현재 : 컴퓨팅 파워의 증가와 저장 기술의 발달
    -> 대부분의 데이터를 저장, 활용 가능

- 데이터 사이언스
  - 데이터 마이닝과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합 분야

    - 위키피디아

  - 데이터와 연관된 모든 것을 의미

    - Journal of Data Science

  - 데이터 사이언스에 필요한 역량은 프로그래밍, 수학과 통계 그리고 특정 분야에 대한 전문성이다.

    - 컨웨이(경제학자)

빅데이터에 대한 오해와 한계

- 빅데이터를 바라보는 두 가지 시선

"빅데이터는 아무 것도 해주는 것이 없다"

vs

"빅데이터는 우리가 가진 모든 문제를 해결해 줄 것이다

- 빅데이터의 한계 = 그 데이터가 가지고 있는 데이터 자체의 한계
  - 분석하고자 하는 빅데이터가 어떤 한계를 가지고 있는지를 항상 염두에 두고
  - 데이터의 속성을 제대로 이해해야

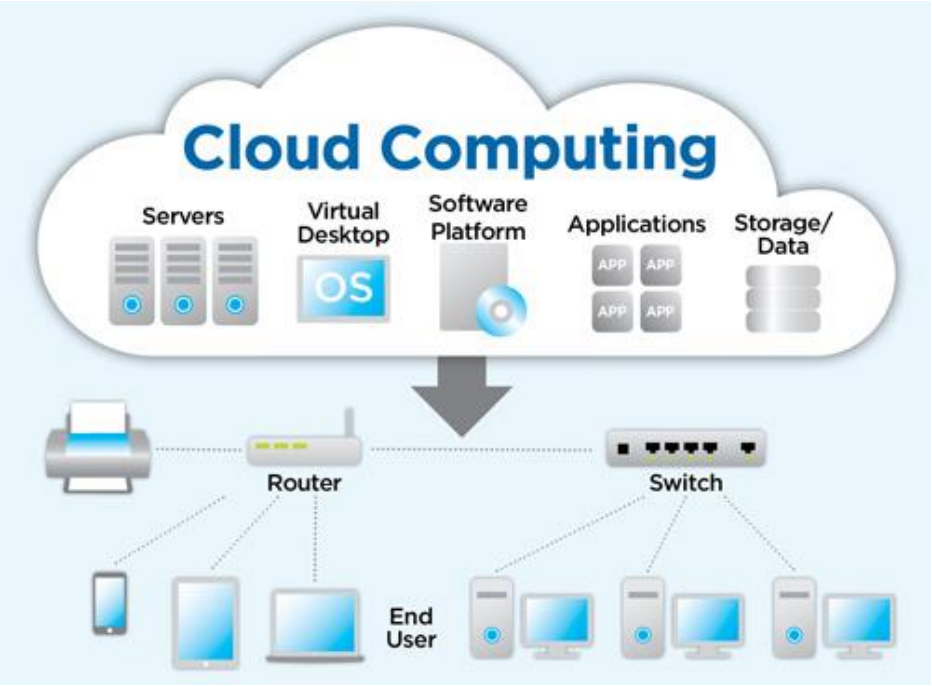
유의미한 가치 창출로 이어질 수 있다

데이터 그 자체는 아무 것도 아니다
데이터에 대한 정확한 이해, 분석, 기획이 있을 때
진정한 가치가 있는 무기가 될 수 있을 것

빅데이터를 지탱하는 기술
: 클라우드 컴퓨팅

- 인터넷 기반 컴퓨팅의 일종
  - 자신의 컴퓨터가 아닌 인터넷에 연결된 다른 컴퓨터로 처리하는 기술
  - 필요할 때 필요한 위치에서 필요한 만큼의 자원을 활용하고 사용한 만큼 대가를 지불하는 방식

- 클라우드 도입의 장점
  - 인프라스트럭처 개발에 시간을 들이는 대신 핵심사업에 집중
  - 유동적이고 예측 불가능한 사업 수요를 충족시키기 위해 자원을 빠르게 조절할 수 있도록 함



빅데이터를 지탱하는 기술
: 클라우드 컴퓨팅

- Amazon vs Google Cloud Platform vs Microsoft Azure

amazon web services

On Demand

Google Cloud Platform

Microsoft Azure

빅데이터를 지탱하는 기술
: 분석 기법과 처리 기술의 발전

- 비정형 데이터의 폭발적 증가로 기존 분석 기술과 처리 기술의 한계에 마주함

- 빅데이터의 양대 기술
  - 분석 기법
  - 처리 기술

기술 분류	설명	세부기술	
분석기법	데이터 집합을 분석하는데 활용될 수 있는 통계 및 컴퓨터 공학 분야의 다양한 기법	· A/B testing · Association rule learning · Classification · Cluster analysis · Crowdsourcing · Data fusion and integration · Data mining · Ensemble learning · Genetic algorithms · Machine learing · Natural language processing · Network analysis · Neural networks	· Optimization · Pattern recognition · Predictive modeling · Regression · Sentiment analysis · Signal processing · Simulation · Spatial analysis · Statistics · Supervised learning · Time series analysis · Unsupervised learning · Visualization
처리기술	분석에 필요한 데이터를 수집, 조작, 관리하거나 분석기법을 지원하기 위해 개발된 기술	· Big Table · Business Intelligence · Cassandra · Cloud computing · Data mart · Data warehouse · Distributed system · Dynamo · Extract, transform and load · Google File System · Hadoop · Hbase	· MapReduce · Mashup · Metadata · Non-relational · R · Relational database · Semi-structured · SQL · Stream processing · Structured data · Unstructured data · Visualization

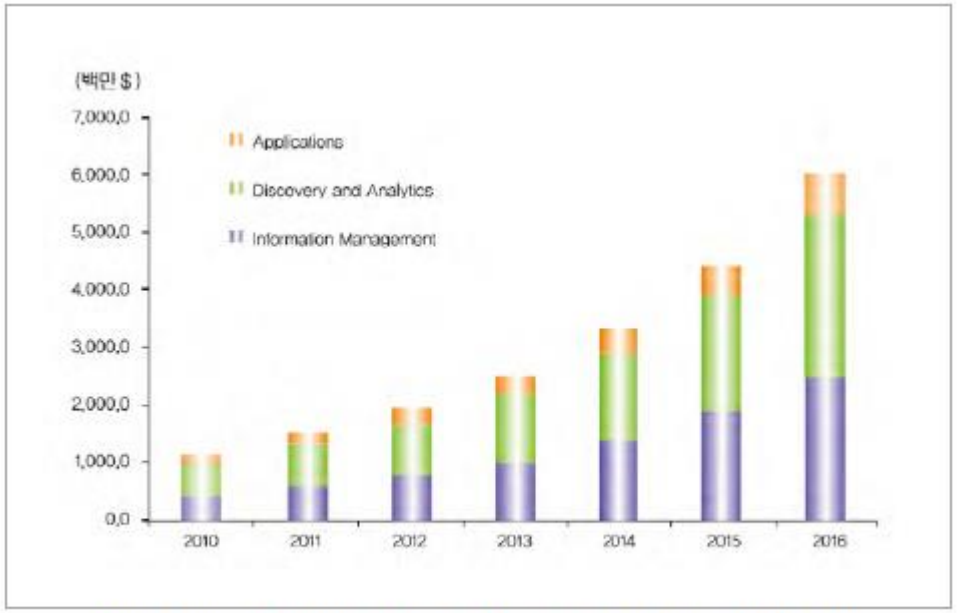
빅데이터 기술과 기법 분류 〈출처 : 정보통신산업진흥원〉

빅데이터에 대한 관심들
: 시장 전망

- 시장 조사기관 IDC, 전체 빅데이터 시장 규모가 2016년 238억 달러에 이를 것으로 예측

- IDC의 빅데이터 시장 구분
  - 기반시설 (Infrastructure)
  - 소프트웨어 (Software)
  - 서비스 (Service)

  - 이중 SW 시장만 60억 달러에 이를 것으로 추산



IDC가 전망한 세계 빅데이터 소프트웨어의 시장 성장 〈출처 : 정보통신산업진흥원〉

빅데이터에 대한 관심들
: 주요 기업들의 M&A

인수기업	시점	대상기업 및 솔루션	영역	세부분야
IBM	2010.09	Netezza	Information Management	DW Appliance & Analytics
	2011.08	i2 Limited	Analytics & Discovery	Analytics for crime & fraud prevention
	2012.05	Vivisimo	Analytics & Discovery	Enterprise Search SW
	2012.05	Varicent	Application	Sales Performance Analytics
	2012.06	Tealeaf Tecchnology	Application	Smarter Commerce Analytics SW
	2013.10	The Now Factory	Application	Mobile Networks Analytics
Oracle	2011.10	Endeca	Analytics & Discovery	Information Discovery
HP	2011.03	Vertica	Analytics & Discovery	BI, real-time analytics
	2011.10	Autonomy	Analytics & Discovery	Pattern matching technology
Teradata	2011.03	Asterdata	Information Management	MapReduce based analytics
	2012.05	eCircle	Application	Marketing Analytics
EMC	2010.07	Greenplum	Information Management	DW Appliance
	2012.03	Pivotal Labs	Information Management	Analytic Application Development Framework

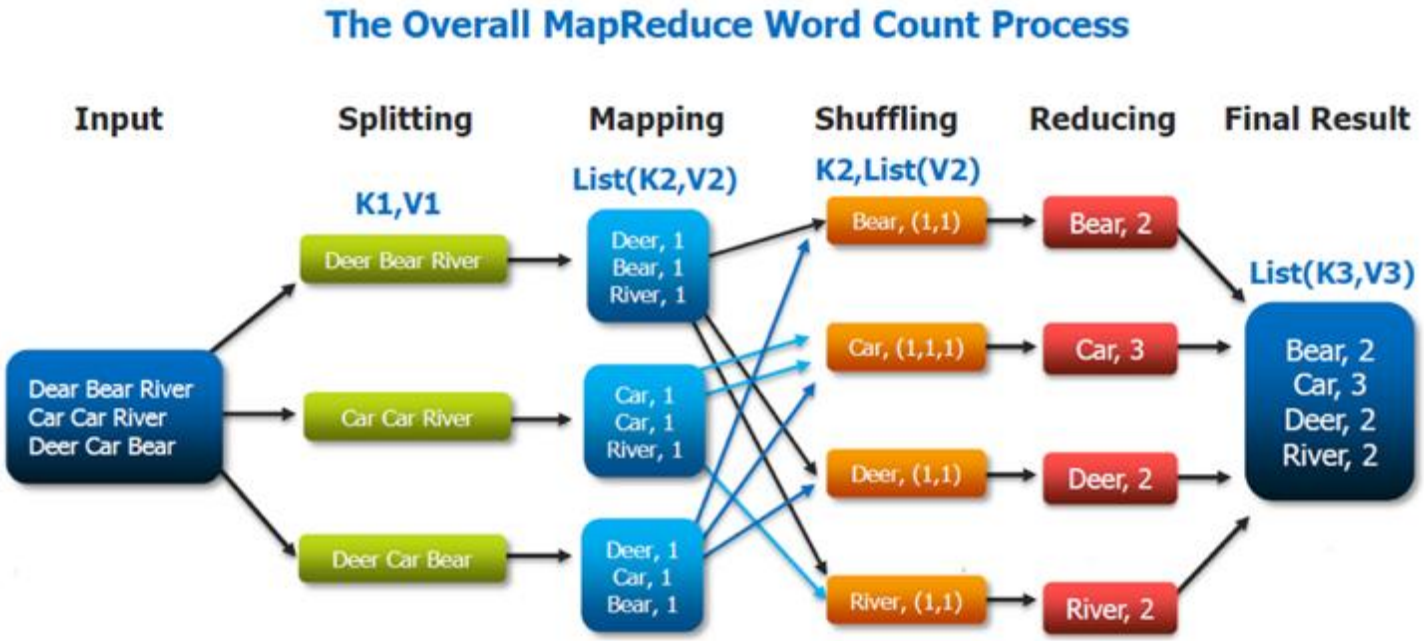
빅데이터 관련 주요 기업 합병 동향 〈출처 : 정보통신산업진흥원〉

빅데이터를 처리하는 기술
: Apache Hadoop

- 저가 서버와 하드디스크를 이용해 빅데이터를 상대적으로 쉽게 활용해 처리할 수 있는 분산 파일 시스템
  - 야후의 지원으로 개발
  - 현재는 아파치 소프트웨어의 프로젝트로 관리
- 빅데이터 플랫폼의 핵심기술이자 사실상의 표준
  - HDFS(분산파일시스템)과 MapReduce(분산 병렬 처리 기술)로 구성

빅데이터를 처리하는 기술
: Apache Hadoop

- 하둡의 용도
  - 검색엔진 색인 저장소(Indexing)
  - 데이터 분석 또는 통계 분석
  - 데이터 전처리(Table Precomputation and Rollup)
  - 정형 데이터의 저장소
    (Structured Data Storage)



빅데이터를 처리하는 기술
: R

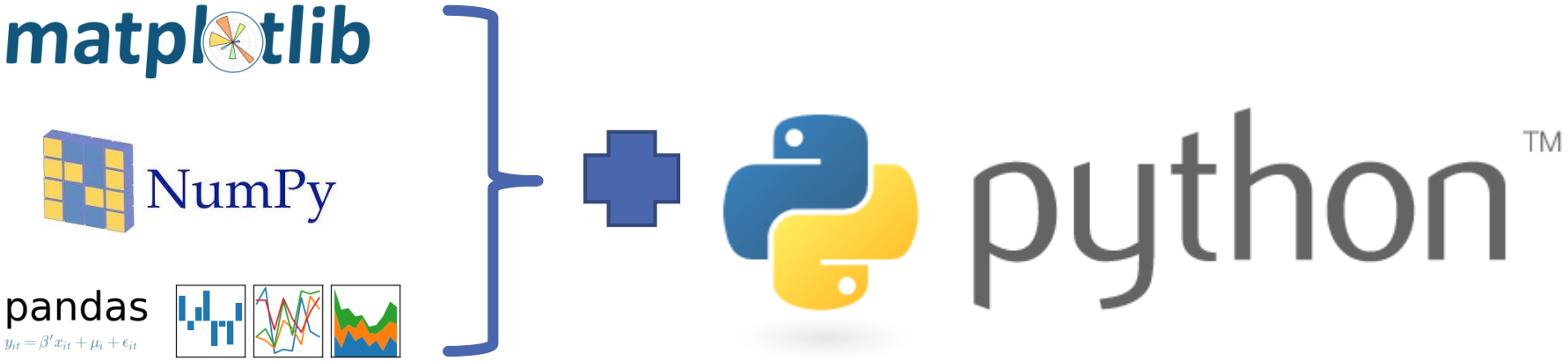
- 통계 계산 및 시각화를 위한 언어 및 개발환경
- 기본적인 통계 기법, 모델링, 최신 데이터 마이닝 기법까지 구현, 개선이 가능
- Java, C, Python 등 다른 프로그래밍 언어와 연결도 용이

- 통계 분석 분야에서 높은 인지도
- 하둡 환경 상에서 분산 처리를 지원하는 라이브러리
  - 페이스북, 아마존 등 빅데이터 분석이 필요한 기업에서 대용량 데이터 통계분석 및 데이터 마이닝 용도로 활용

빅데이터를 처리하는 기술
: Python

- 간결한 문법과 높은 확장성을 지닌 범용 프로그래밍 언어
- Glue Language로 폭넓은 커뮤니티를 보유, 방대한 오픈소스 라이브러리 활용 가능
- 데이터 처리 이외에도 다른 분야의 응용프로그램과의 연동이 용이

- 기본적으로는 통계 기법, 모델링, 분석 등 데이터 관련 도구들은 R에 비해 취약
- NumPy, Pandas, Matplotlob 등 데이터 처리 모듈들을 활용, 높은 수준의 데이터 처리 가능



빅데이터를 처리하는 기술
: NoSQL

- Not-Only SQL
- 전통적 관계형 데이터베이스와 다르게 구현된 비관계형 데이터베이스

- Cassandra, Hbase, MongoDB 등

- 스키마가 고정되지 않고, 수평적 확장이 용이하다는 장점
  : 빅데이터의 특징인 비정형, 대량의 데이터 처리에 유리