

# 기상기후 빅데이터와 지능정보기술 활용과정

빅데이터 분석기법 (심화 – AI 기술 이해)

Training BigData Experts & Empowering AI Technology  
Kim Jin Soo

# Profile

KNU, 경북대 전자공학과 (시스템공학) 졸업  
KNU, 경북대 전자공학과 (정보통신) 석사 졸업  
CUK, 고려사이버대 디지털경영학과

## 경력 및 자격

前 대우정보시스템 기술연구소 선임연구원  
前 SBS 미디어넷 기술개발CP 개발팀장  
前 E-Biz 기업솔루션 전문벤처기업 기술연구소장  
前 기업통합보안솔루션 벤처기업 CTO  
前 산업보안전문가포럼 회장  
前 경기창조경제혁신센터 경기콘텐츠진흥원 CSP

**빅데이터기술전문가(Bigdata Technical Expert)**  
**산업보안전문가(Industrial Security Professional)**



**김 진 수**  
**CEO / DataActionist**

現 서울T직업전문학교 빅데이터 겸임교수  
現 멀티캠퍼스 IT기술교육 & 전문강사  
現 한국경제신문 빅데이터센터 위촉위원  
現 한국SW기술진흥협회 기술위원  
現 경기문화창조허브 스타트업플래너  
  
現 INNOTURN 경영컨설팅 대표컨설턴트  
現 중기부 중소기업지원단 현장클리닉위원  
現 고용노동부 일터혁신컨설팅 컨설턴트  
現 산인공 NCS/일학습병행제 컨설턴트  
現 서울시 정보통신분과 청년취업 멘토

# 빅데이터 분석기법

인공지능 기술이해

# 생각해 봅시다!!



- ◆ 정보화 시대, 데이터와 정보의 차이
- ◆ 컴퓨터가 바라본 데이터의 조건
- ◆ 빅데이터의 서막, 오픈소스진영의 반란 !!
- ◆ 기술을 알면 보이는 것도 달라진다
- ◆ 빅데이터가 없는 인공지능은 무용지물
- ◆ 빅데이터에 AI기술 활용

# 빅데이터 시대!!



# မြန်မာစာမျက်နှာ

사람들의 마음속을 들여다봤는지, 각자의 취향에 맞춰 필요한 물건을 알아서 추천해주는 시대. 누구나 건강 상태를 실시간으로 체크할 수 있는 시대. 수명까지 예측해주는 시대. 결코 먼 미래의 이야기가 아니다. 이미 우리 삶 속으로 빠르게 다가오고 있다. 그 중심에는 빅데이터가 있다. 하지만 정작 빅데이터가 뭐냐고 물었을 때 정확히 대답하는 사람은 드물다. 빅데이터가 도대체 무엇인지 제대로 파악해보자.

글 김경한 기자(dgude@donga.com) 일러스트 김경한 알앤씨 도움 한윤경과수수과학연구소 음용수학연구부 연구책임자, 카이스트(KAIST) 문화융합대학원 교수, 박진현(AIST 전산부록 세미나팀장) 박지현(김시아), 박재민(한국수학전집), 김기식(AIST 산업활성화사업단장) 교수, 최수영(미래수학 교수) 험고 서희 예술사대학원(박태이너의 다음 단계는 예술분야이다), 노루수, 손상미(미술) 박신은 미술이론가이다. 김경한(동아일보) (별도로 명시하지 않음은 고백)

## PART 1 크다고 다 빅데이터일까?

## PART 2 빅데이터 퀘哪家보는 수학

### PART 3 나도 데이터과학자!



# 빅데이터 시대!!



The collage includes the following titles:

- BIG DATA** (Large image, left side)
- 빅데이터 경영을 바꾼다!** (Large image, bottom left)
- 빅데이터의 충격** (Large image, center)
- 세상을 바꾸고 나를 변화시키는 보이지 않는 것의 힘** (Large image, top center)
- 빅 데이터 세상을 이해하는 새로운 방법** (Large image, middle center)
- 4차 산업혁명 인공지능 빅데이터** (Large image, right center)
- 데 이 터 는 알 고 있 다** (Large image, bottom right)
- BIG DATA** (Large image, far right)
- THE BIG DATA REVOLUTION** (Small image, top right)

Text elements from the books:

- "R을 이용한 중·고급"
- "세계 최고의 인재를 기우는 기업"
- "데이터의 파도가 사업 전략을 바꾼다!"
- "격차한 '빅 데이터'를 바꾸다, 인공지능은 정부다, 빅데이터는 세상을 바꾼다!"
- "데이터는 답을 알고 있다! 빅데이터 시대의 새로운 기회를 찾는다."
- "데이터를 활용한 IT 기술 혁명"
- "실무자를 위한 빅데이터 교과서"
- "전설을 말하고 세상을 만드는 빅 데이터의 모든 것!"

# 빅데이터의 의미

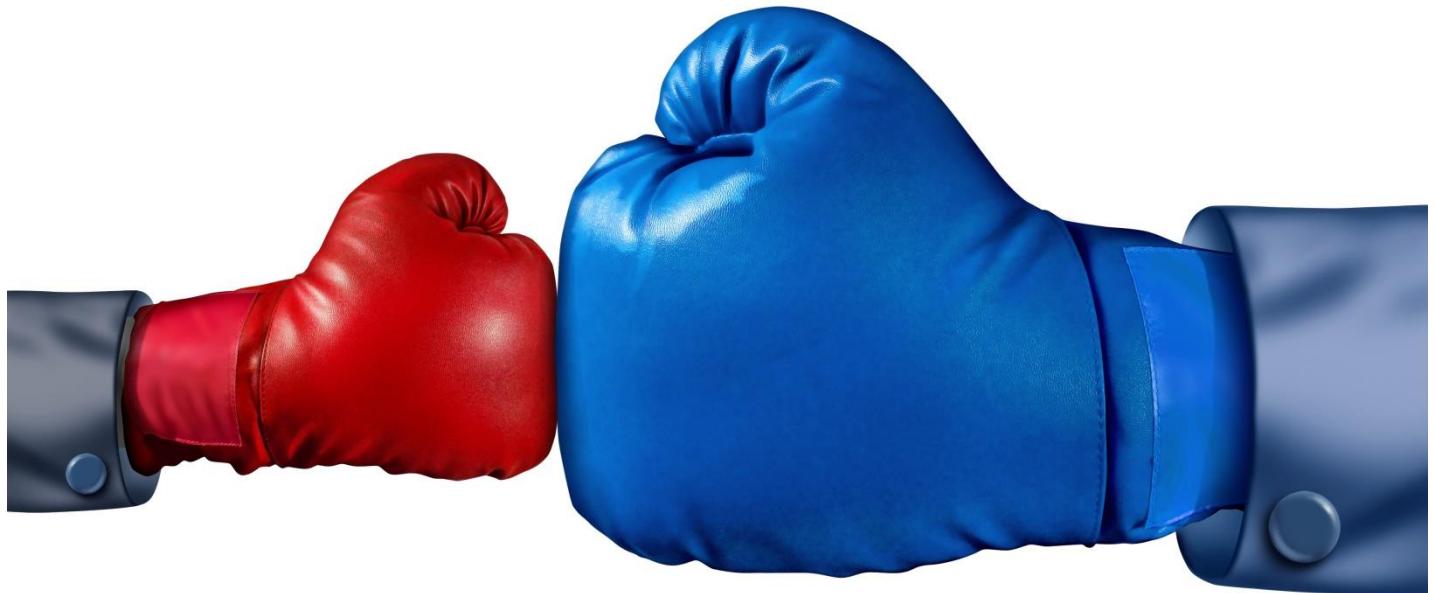


“빅데이터”라는 용어는 널리 알려져 있지만,  
그 구체적인 개념을 명확히 정의하기엔 모호하다!

# 빅데이터의 의미



- 많은 용량의 데이터가 빅데이터를 의미하는 건 아니다!



# 빅데이터의 의미



## 빅데이터 연상개념

- 대용량 데이터
- 대량의 데이터
- 소셜 미디어 데이터
- 차세대 데이터
- 실시간 데이터

- 빅(Big)이란 의미는 훨씬 포괄적인 의미를 내포하고 있다!

# 빅데이터의 의미



- 실제 오늘날 사람들은 매일 어마어마한 데이터를 쌓아가고 있다.

20세기 까지  
쌓은 데이터

요즘 하루 동안  
만든 데이터

## 빅데이터 - Word Cloud





## 정보화 시대 → 빅데이터 시대 !!

- 우리가 다루는 것은? 데이터는 무엇인가?
- 도대체 무엇을 수집하고, 어떻게 표현하고,  
관리는 왜 해야하고, 활용은 어떤 방식으로 해야할까?



# 빅데이터의 서막 - 오픈소스진영의 반란

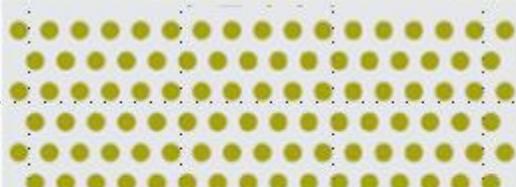


- ❖ 2005년 이전의 일반적인 데이터 관리, 대용량데이터
- ❖ 고가가 아닌 저가 시스템, 중앙집중방식이 아닌 분산처리방식
- ❖ Hadoop 기술의 등장
- ❖ 빅데이터 생태계(EcoSystem)를 이루는 다양한 기술들
- ❖ CAP이론  
→ Pick Two : Consistency, Availability, Partition Tolerance

# 빅데이터의 주요특징 - 3V, 4V, 5V

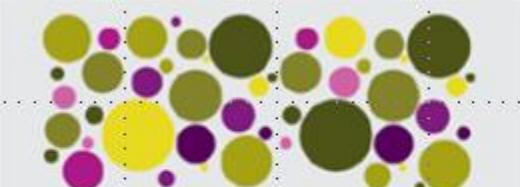


## Volume



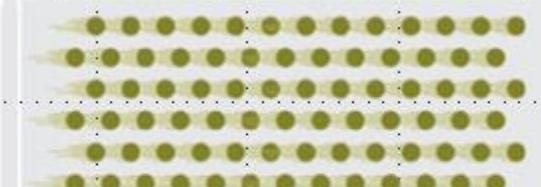
데이터의 양  
TB부터 PB  
정도의 데이터

## Variety



데이터의 다양한 형태  
정형/비정형  
텍스트, 멀티미디어

## Velocity



데이터의 이동  
몇 분의 1초 사이에 의사결정을  
가능하게 해주는 스트리밍  
데이터 분석

## Value

### 미래 가치 창출

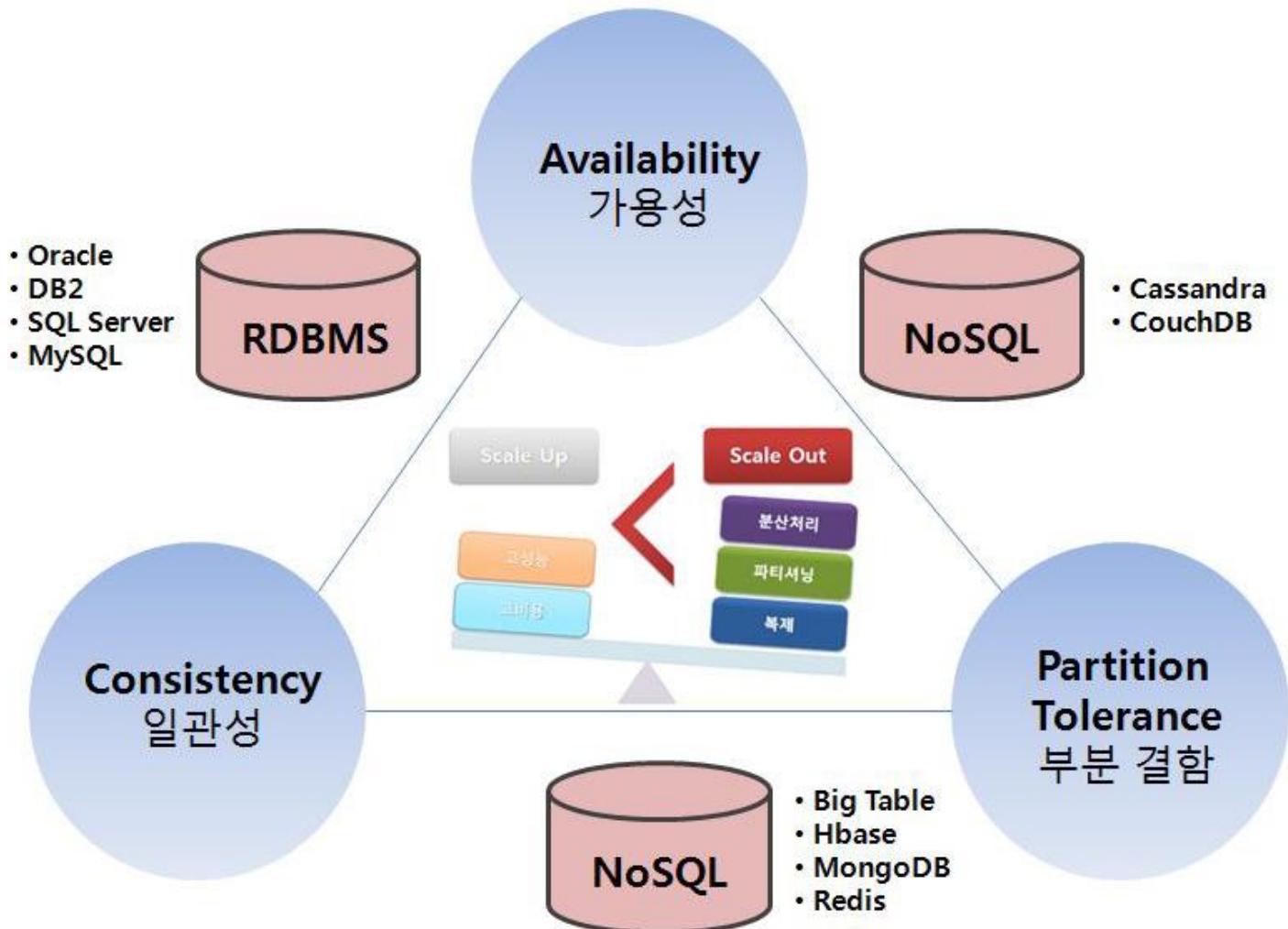
경제적 가치 뿐만 아니라,  
문제 해결을 통한 사회적 가치  
를 창출

### 데이터의 불확실성

본질적으로 불확실한 데이터  
유형의 신뢰성과 예측 가능성  
관리

## Veracity

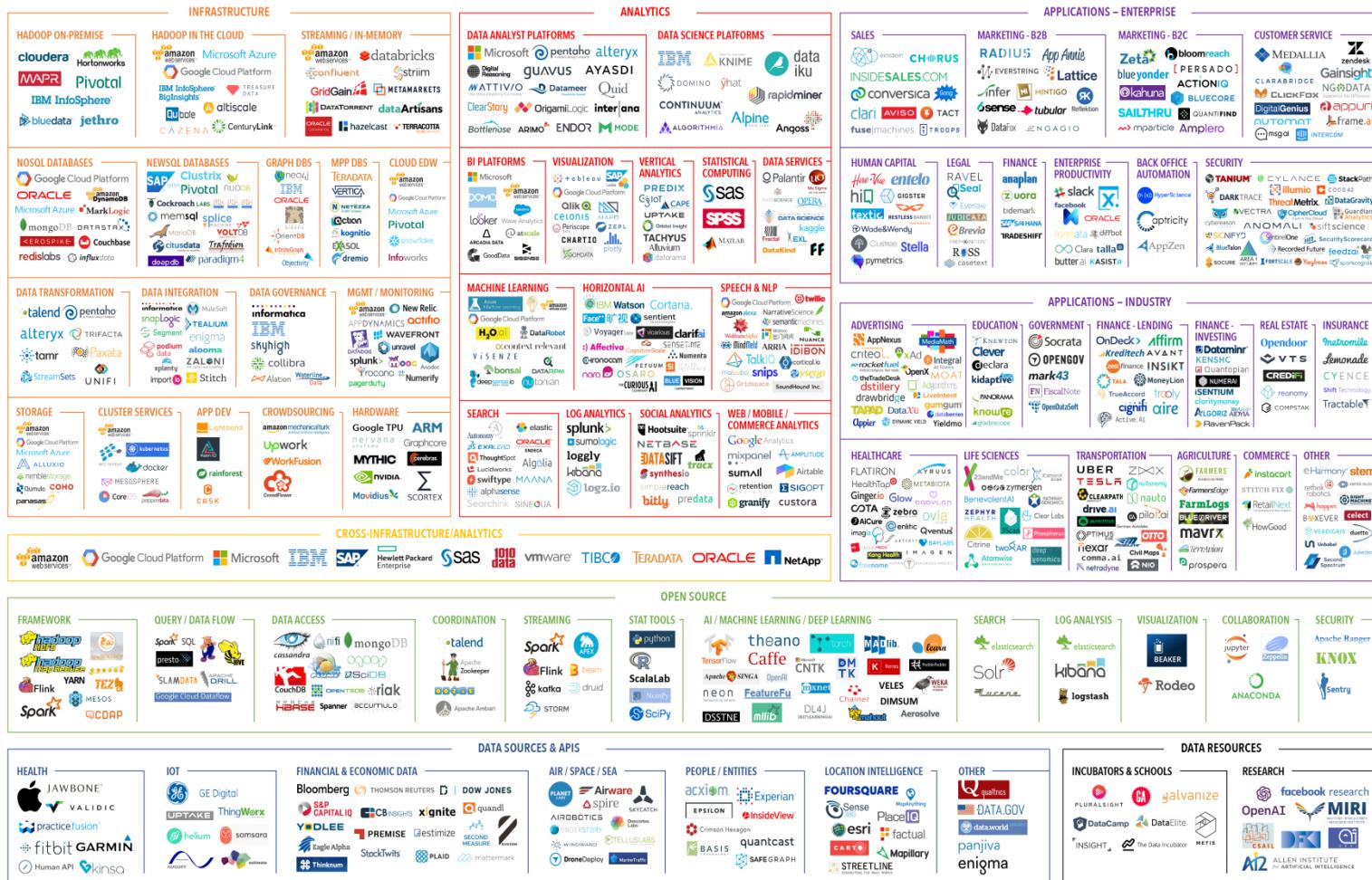
# CAP 이론에 기반한 빅데이터 제품



# 빅데이터 관련 기술 - BigData Landscape



BIG DATA LANDSCAPE 2017



V2 – Last updated 5/3/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)

[mattturck.com/bigdata2017](http://mattturck.com/bigdata2017)

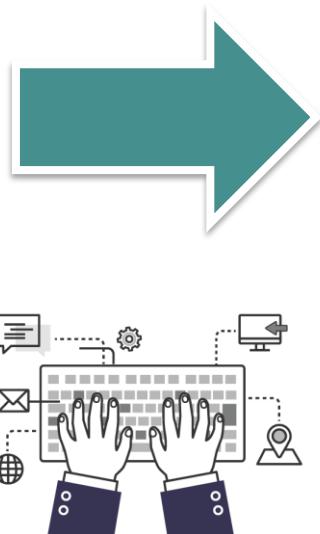
FIRSTMARK  
EARLY STAGE VENTURE CAPITAL



# 빅데이터의 목적



## 무엇을 얻어내려고 분석하는가?



Creating  
Shared  
Value

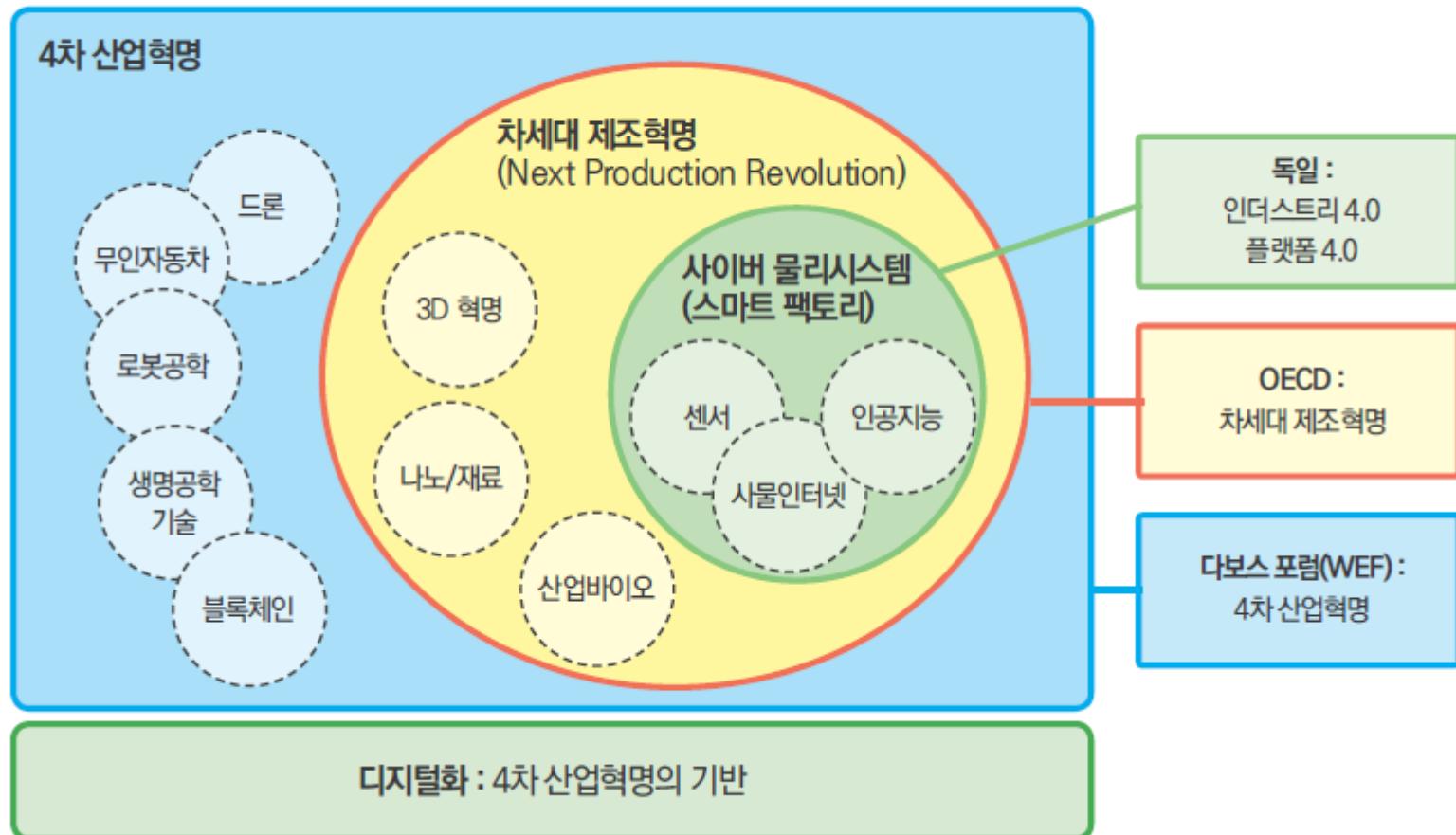


2018 다보스 포럼 의제  
분절된 세계에서 공유의 미래 창조



# 4차산업혁명에 대한 이해

## 4차 산업혁명 관련 개념 관계도



<발췌: 과학기술정책연구원 장필성님 칼럼>

# 데이터과학, Data Science



**WIKIPEDIA**  
The Free Encyclopedia

## Data science

From Wikipedia, the free encyclopedia

*Not to be confused with information science.*

**Data science**, also known as **data-driven science**, is an interdisciplinary field about scientific processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,<sup>[1][2]</sup> which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics,<sup>[3]</sup> similar to Knowledge Discovery in Databases (KDD).

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the **data deluge**.<sup>[4][5]</sup>

데이터 사이언스(data science)이란 데이터와 관련된 연구를 하는 학문이다.

데이터의 구체적인 내용이 아닌 서로 다른 성질의 내용이나 형식의 데이터에 공통으로 존재하는 성질, 또는 그것들을 다루기 위한 기술의 개발에 착안점을 둔다는 특징을 가진다.

사용되는 기술은 여러분야에 걸쳐있으며 수학, 통계학, 계산기과학, 정보공학, 패턴인식, 기계학습, 데이터마이닝, 데이터베이스 등과 관련이 있다.

데이터 사이언스는 생물학, 의학, 공학, 사회학, 인문과학 등의 여러 분야에 응용되고 있다.

# 빅데이터 전문가, Data Scientist



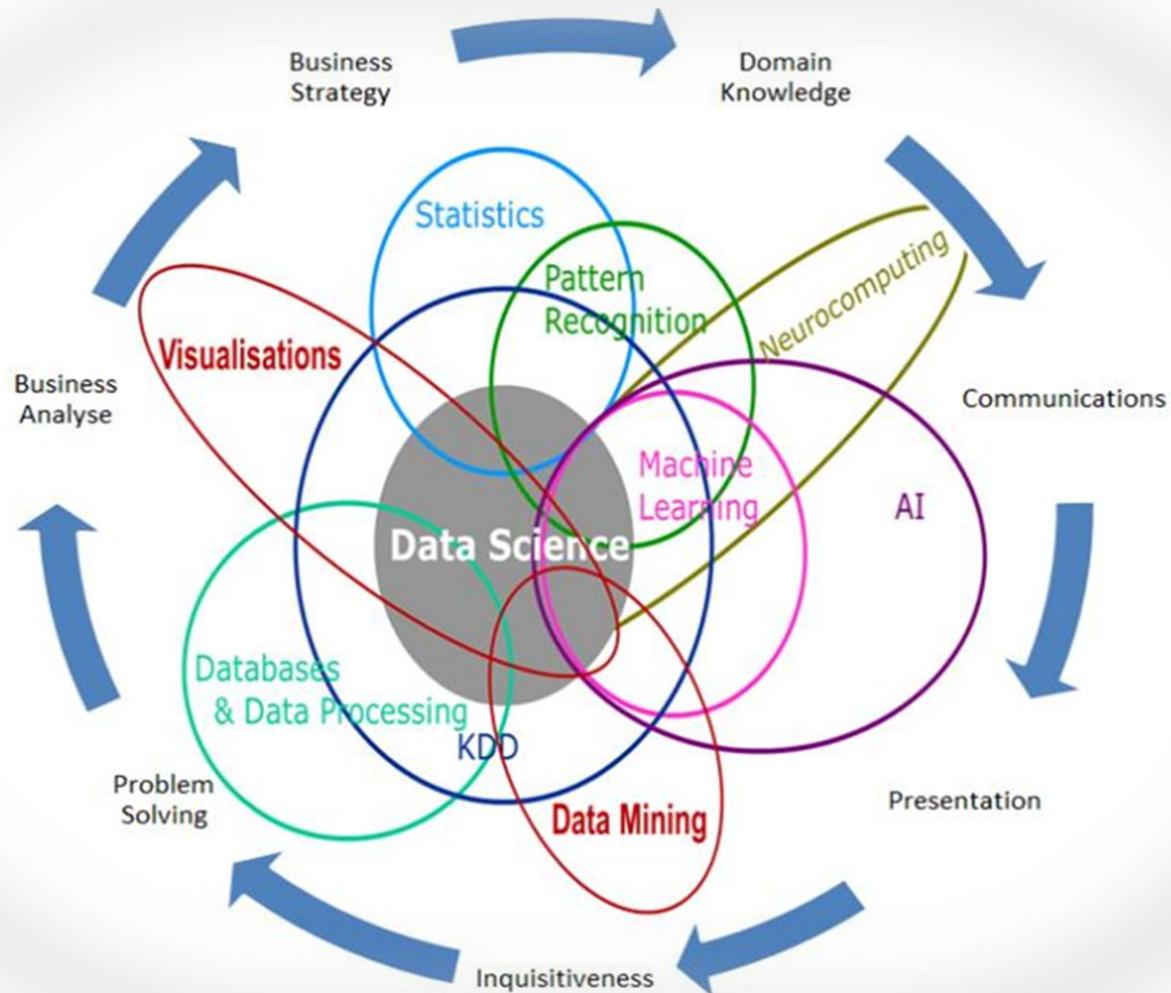
## Data scientist [edit]

The Intel logo is visible on the left side of the slide.

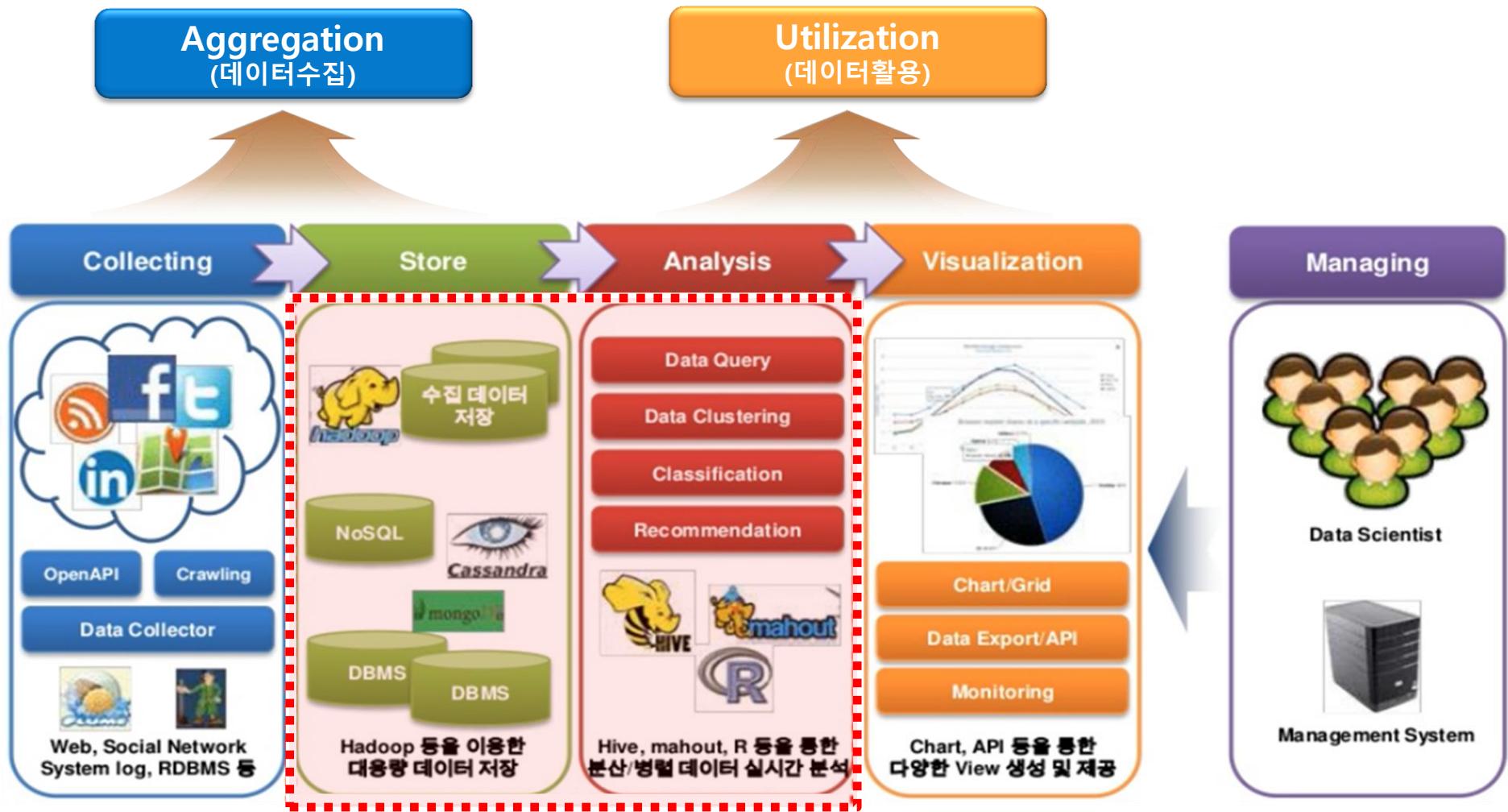
Data scientists use their data and analytical ability to find and interpret rich data sources; manage large amounts of data despite hardware, software, and bandwidth constraints; merge data sources; ensure consistency of datasets; create visualizations to aid in understanding data; build mathematical models using the data; and present and communicate the data insights/findings. They are often expected to produce answers in days rather than months, work by exploratory analysis and rapid iteration, and to produce and present results with dashboards (displays of current values) rather than papers/reports, as statisticians normally do.<sup>[8]</sup>

"Data Scientist" has become a popular occupation with Harvard Business Review dubbing it "The Sexiest Job of the 21st Century" <sup>[9]</sup> and McKinsey & Company projecting a global excess demand of 1.5 million new data scientists.<sup>[10]</sup> Universities are offering masters courses in data science.<sup>[11]</sup> Shorter private bootcamps are also offering data science certificates including student-paid programs like General Assembly to employer-paid programs like The Data Incubator.<sup>[12]</sup>

# 데이터 과학의 활용 분야



# 빅데이터 처리 4단계



빅데이터 요소기술이 투입됨

BigData Solution의 기능 및 처리 흐름과 관리구조

# 데이터 관련 산업 종사자들

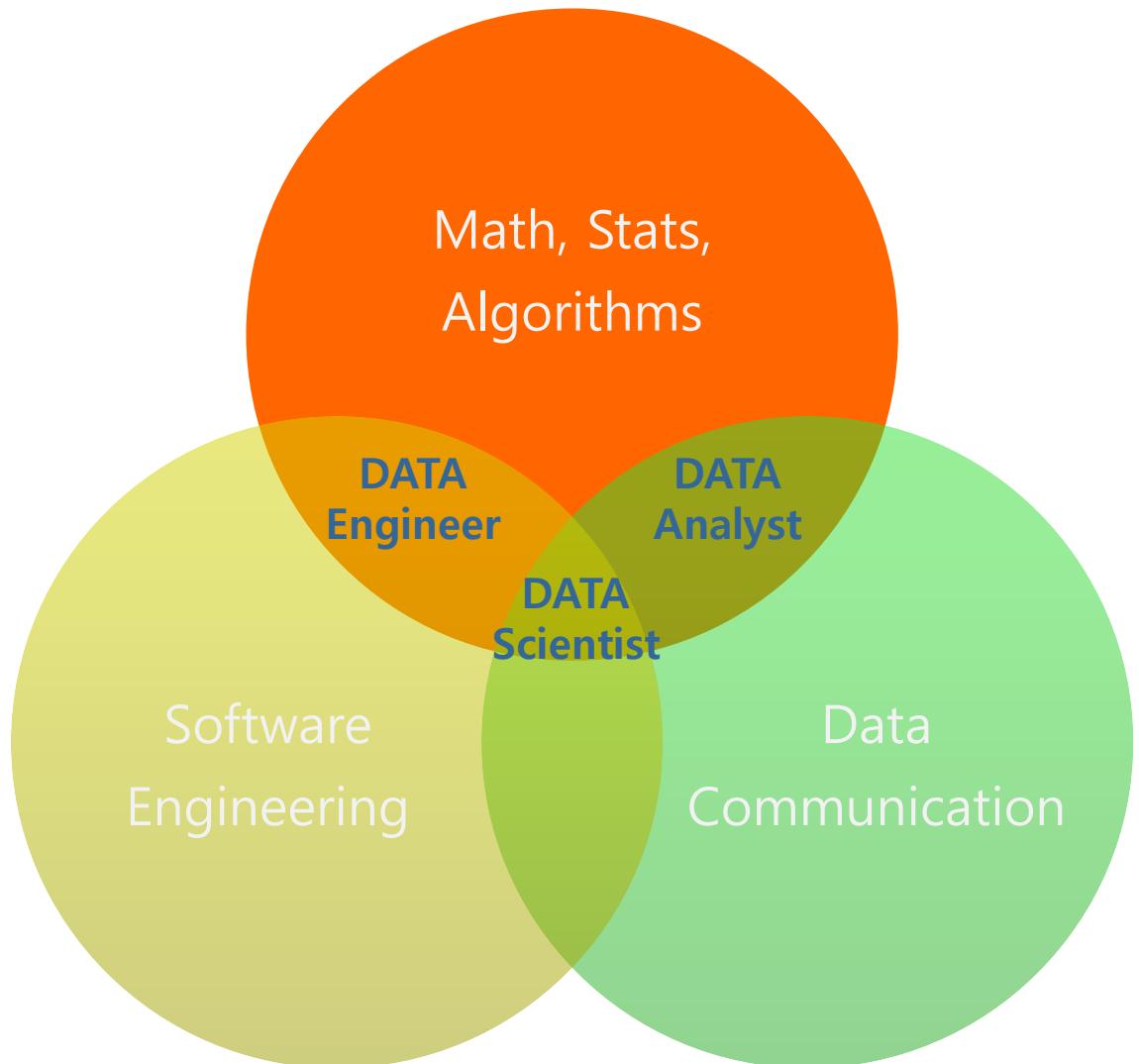


## 고전적 관점

- 데이터 분석전문가
- 데이터 기술전문가

## 최신 트랜드

- 데이터 과학자

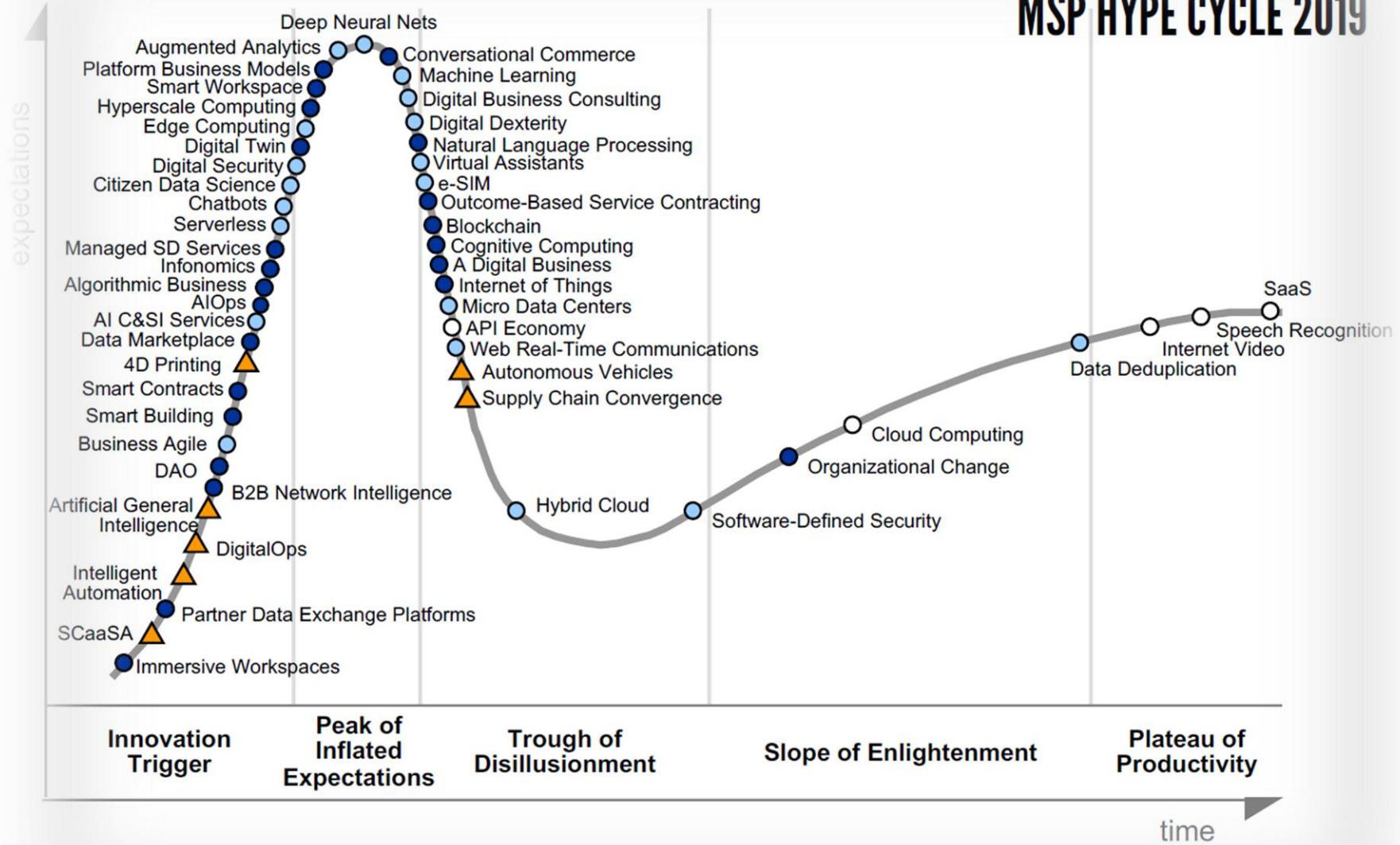


# 데이터 과학의 동향



portland<sup>®</sup>

## MSP HYPE CYCLE 2019





**BigData** ❤ **Software**

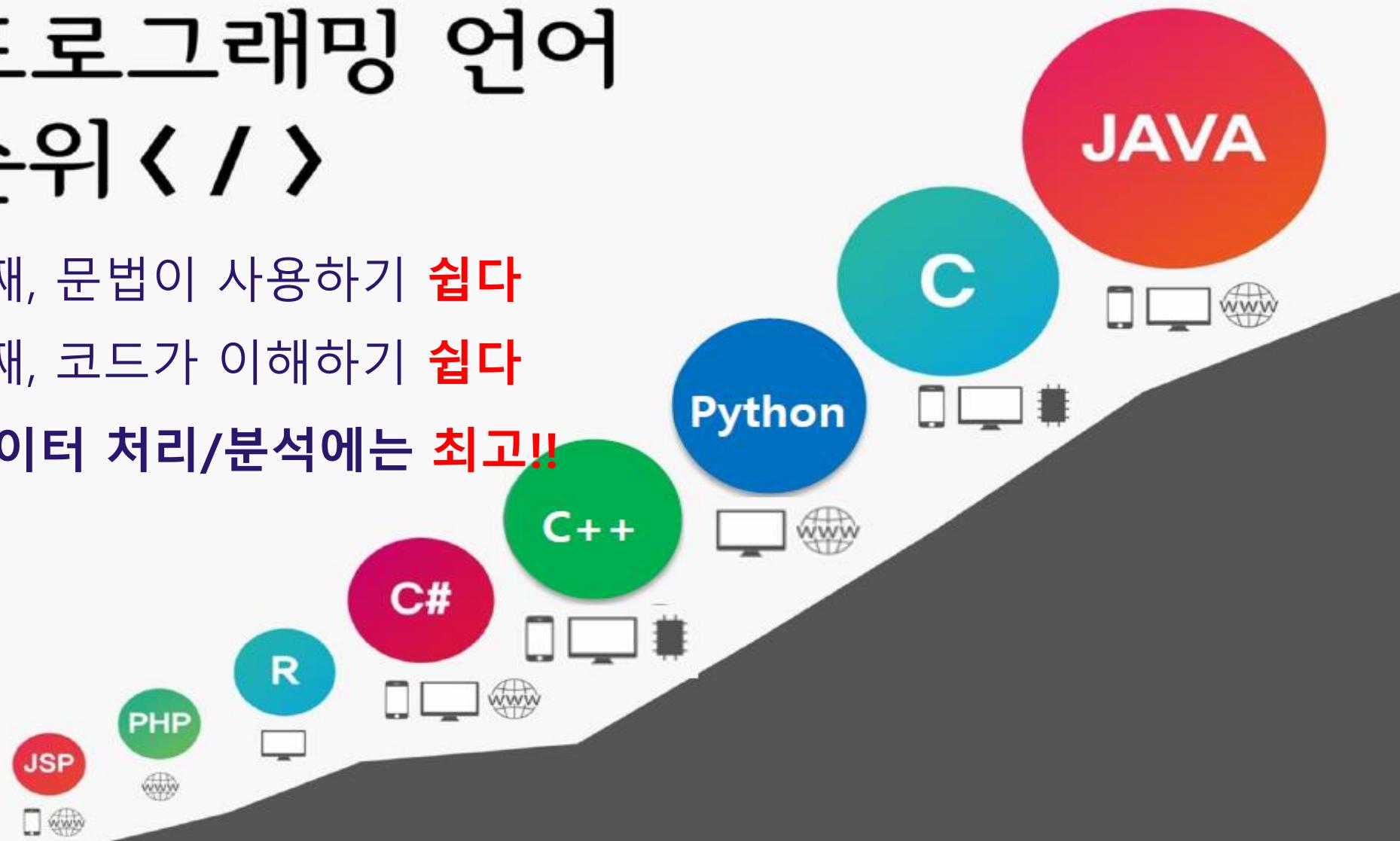


## 프로그래밍 언어 순위 < / >

첫째, 문법이 사용하기 **쉽다**

둘째, 코드가 이해하기 **쉽다**

∴ 데이터 처리/분석에는 **최고!!**



# 데이터 분야 종사자들



## 데이터 분석전문가

- MS Excel, SPSS(통계툴) 등을 잘 다룸
- 데이터를 통해 문제를 해결하는데 집중된 능력
- 대량의 데이터 (빅데이터) 를 다루거나, 수학적 알고리즘이나 모델을 만드는 능력은 기대되지 않음

## 데이터 기술전문가

- Hadoop, MapReduce, MySQL (RDBMS), 프로그래밍 등을 할 줄 안다
- 데이터 분석가를 위해 **대량의 데이터로부터 핵심 데이터를 추출 (ETL작업)**
- 개발 부분에 포커스 하기 때문에 데이터 분석이나 머신러닝 같은 능력은 기대되지 않음

# 인공지능의 핵심기술

머신러닝/딥러닝의 메커니즘 이해

# 머신러닝(Machine Learning)의 개념



**Using known data, develop a model  
to predict unknown data**

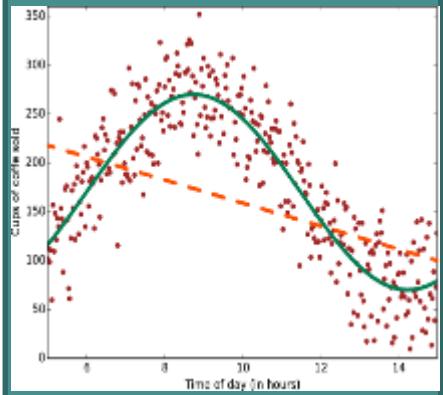
**알려진 데이터를 사용하여, 알려지지 않은 데이터를  
예측하는 모델을 개발하는 기법**

- Known Data :  
과거의 모든 빅데이터, 이전에 관측된 데이터,
- Unknown Data :  
누락된 데이터, 보이지 않는 데이터, 존재하지 않는 미래데이터
- Model : Known Data + Algorithms(ML algorithm)

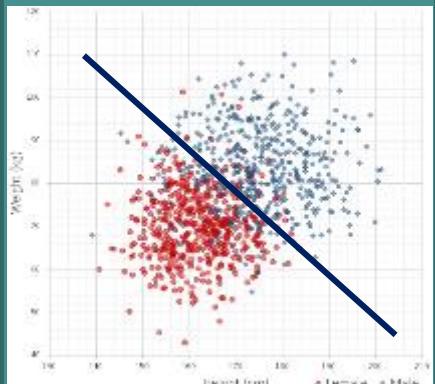
# 머신러닝 알고리즘



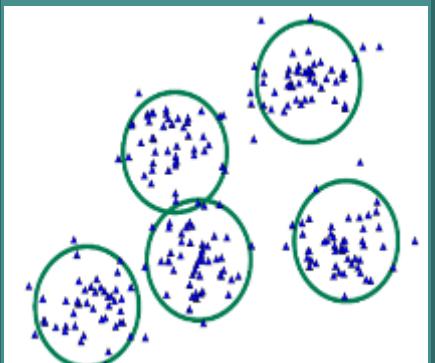
## Regression



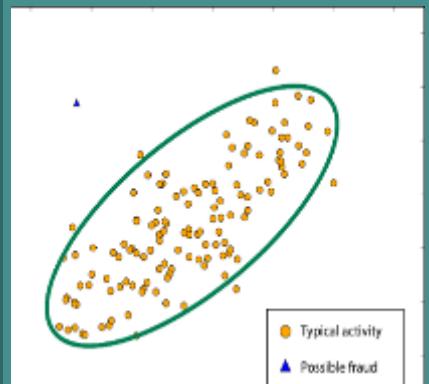
## Classification



## Clustering



## Anomaly Detection



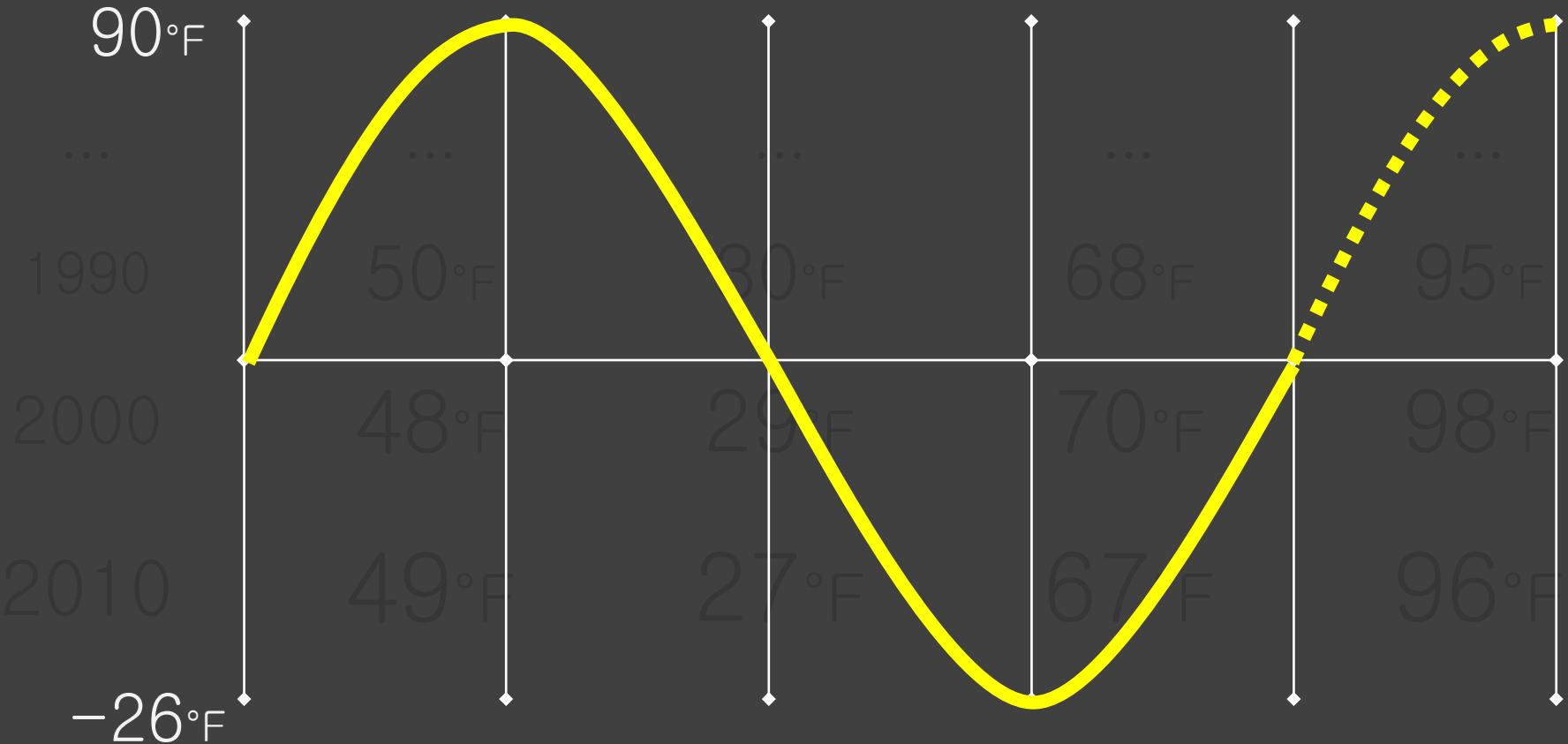
# 머신러닝 모델 : Regression



...	...	...	...	...
1990	50°F	30°F	68°F	95°F
2000	48°F	29°F	70°F	98°F
2010	49°F	27°F	67°F	96°F
2020	?	?	?	?

Using known data, develop a model to predict unknown data.

# 머신러닝 모델 : Regression

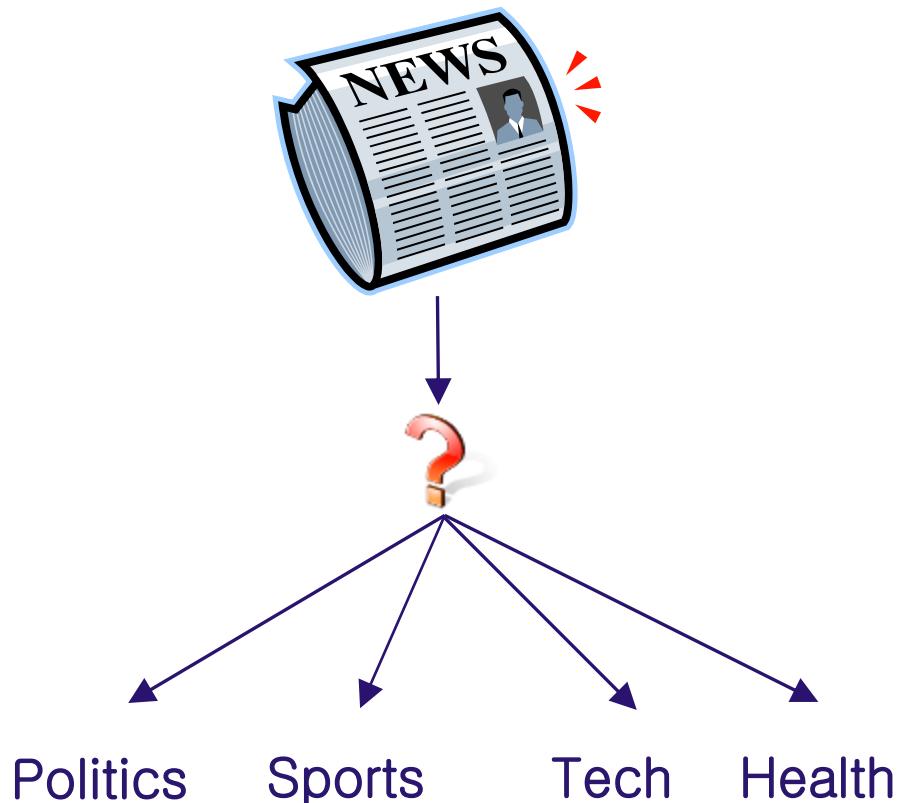


Using known data, develop a model to predict unknown data.

# 머신러닝 모델 : Classification



Classify a news article as (politics, sports, technology, health, ...)



# 머신러닝 모델 : Classification



Documents consist of unstructured text.  
Machine learning typically assumes a more structured format of examples

Process the raw data

Documents      Labels



Tech



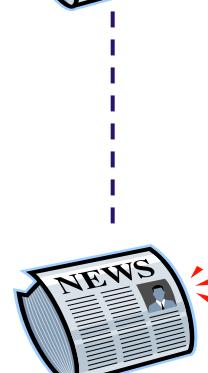
Health



Politics



Politics



Sports

# 머신러닝 모델 : Classification

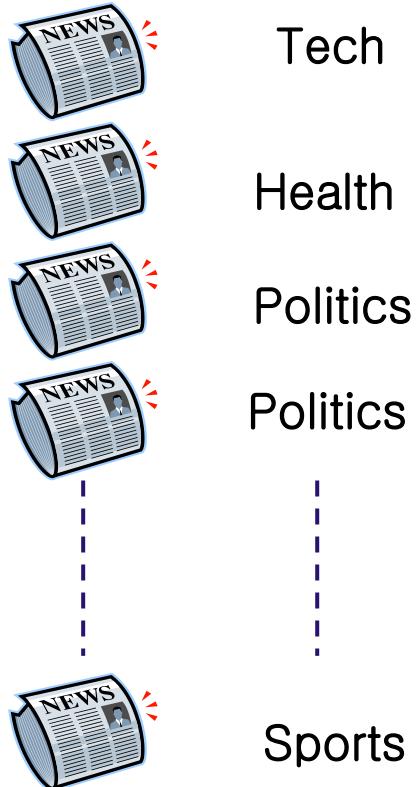


# Known data (Training data)

## Documents

## Labels

Process each data instance to represent it as a feature vector



# Feature

## Documents

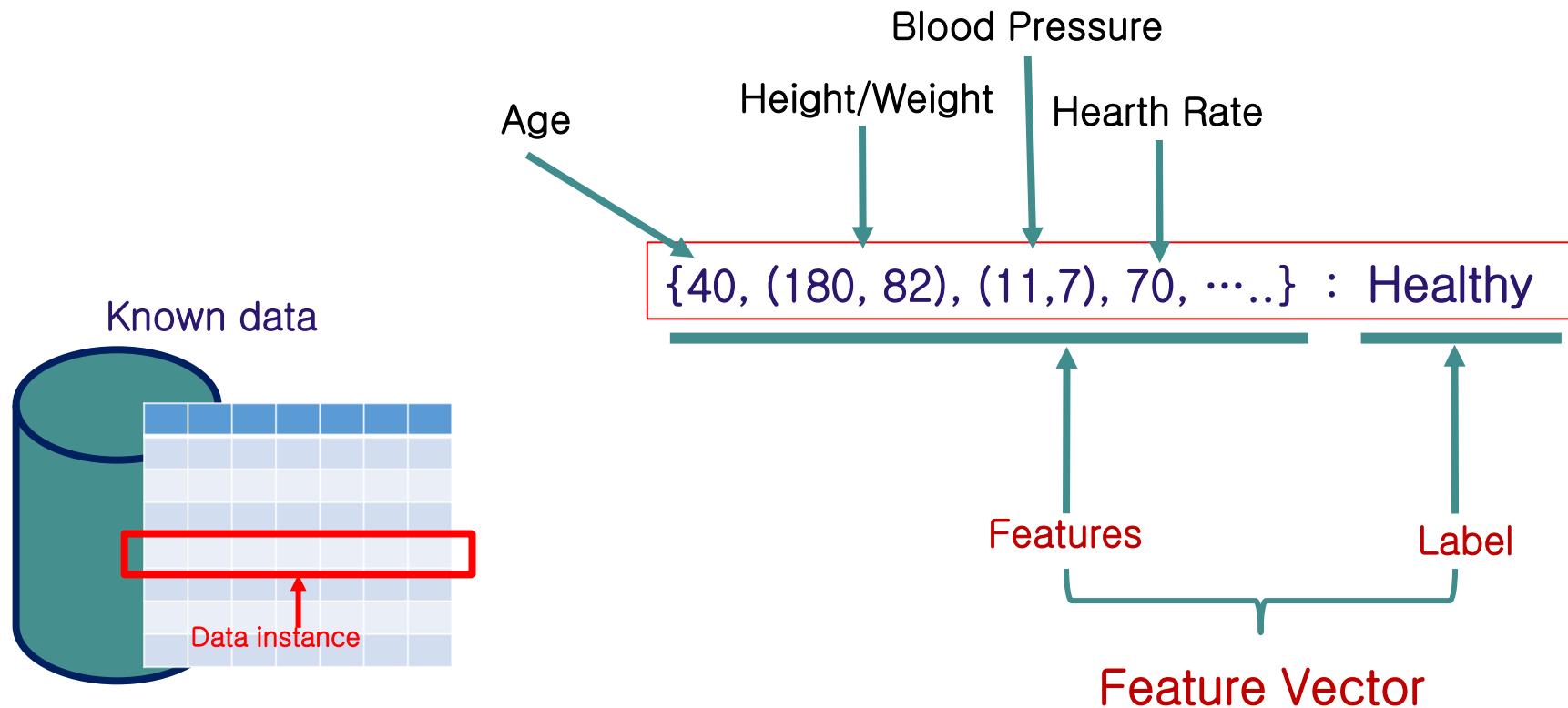
## Labels

# 머신러닝 모델 : Classification



## Feature vector

i.e.



# 머신러닝 모델 : Classification



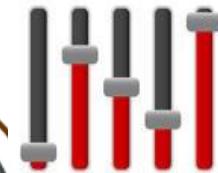
## Developing a Model

Training data

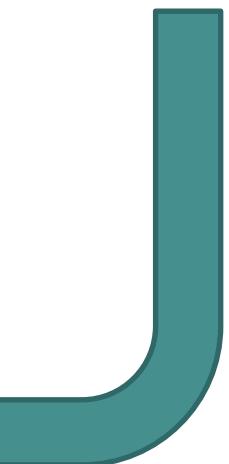
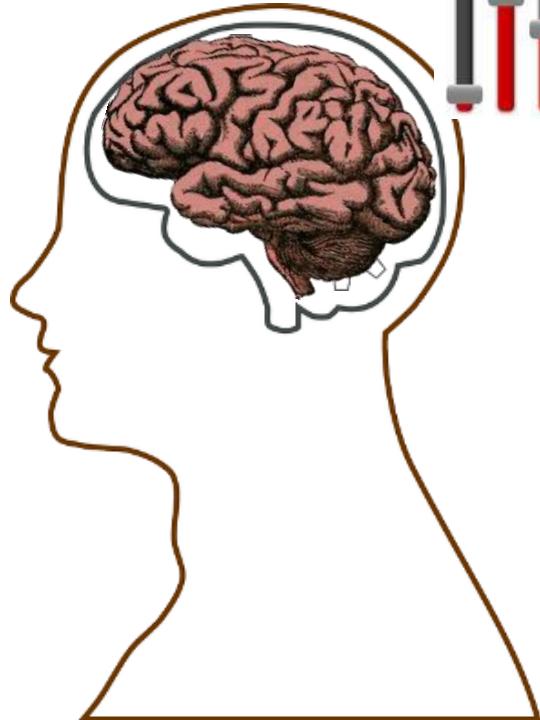
Documents	Labels	Feature Vectors
NEWS	Tech	
NEWS	Health	
NEWS	Politics	
NEWS	Politics	
NEWS	Sports	

Train  
the Model

Base  
Model



Adjust  
Parameters

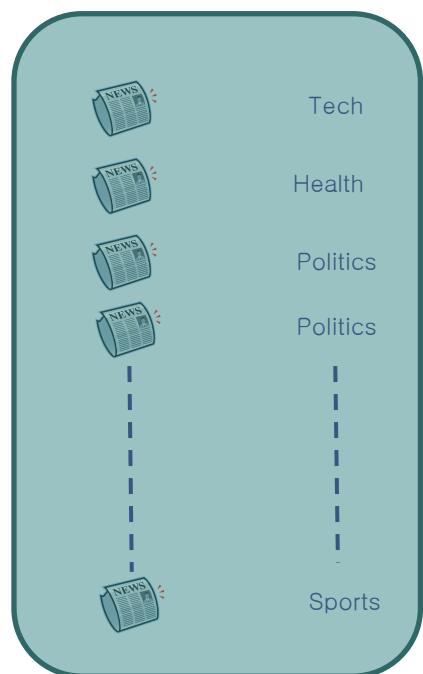


# 머신러닝 모델 : Classification



## Model's Performance

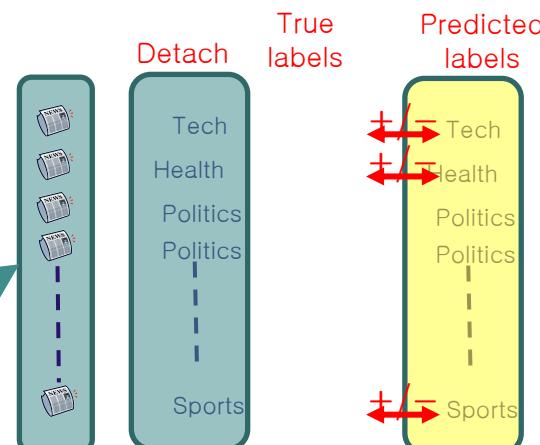
Known data with true labels



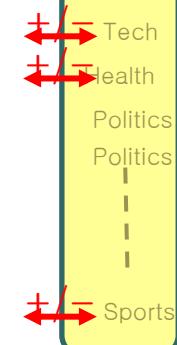
Split

Test data 20%

Training data 80%



True labels  
Predicted labels



Compare prediction with true labels

Test trained model with features

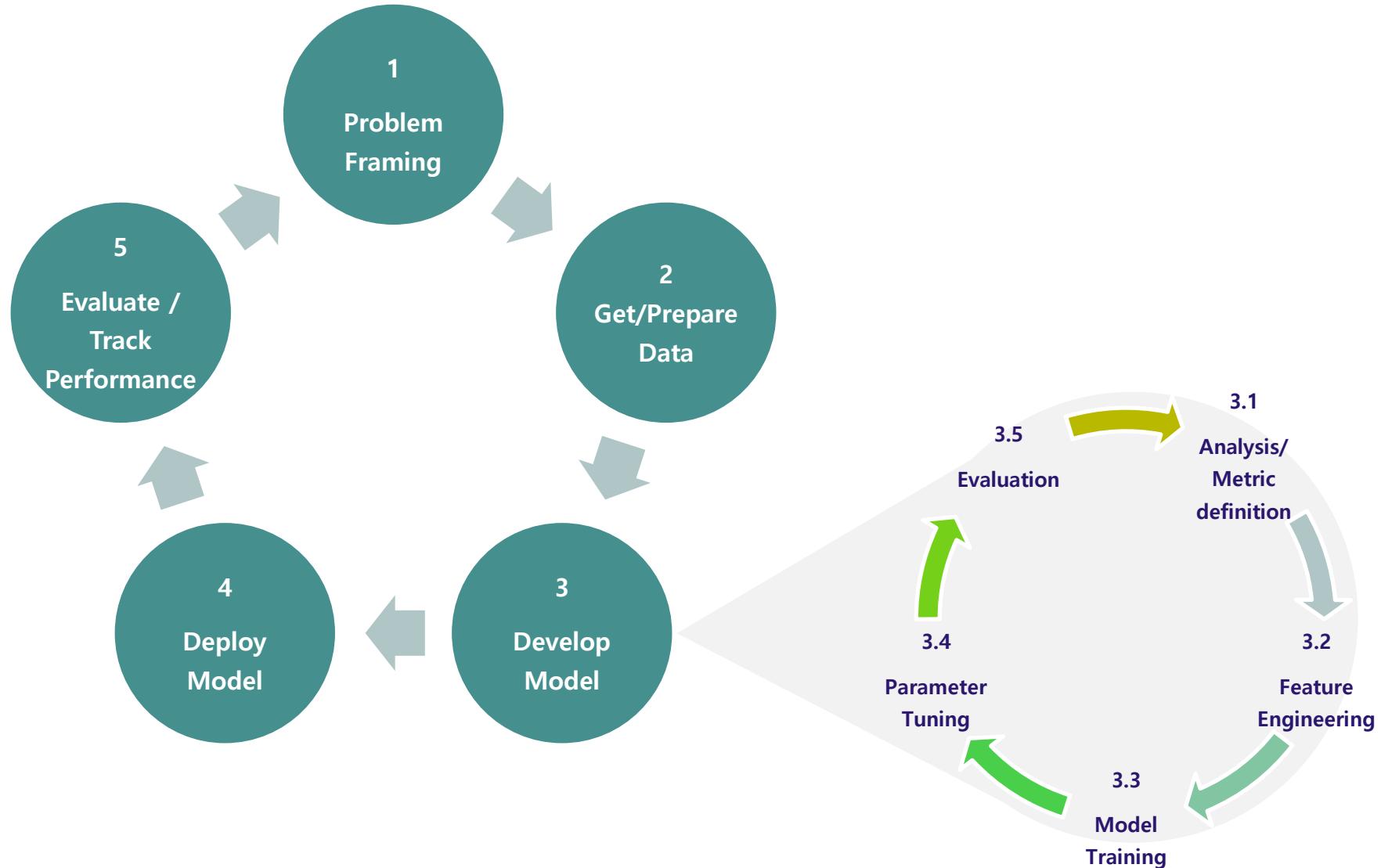
Train the Model

Model's Performance

Difference between "True Labels" and "Predicted Labels"



# 머신러닝 솔루션 구현 프로세스





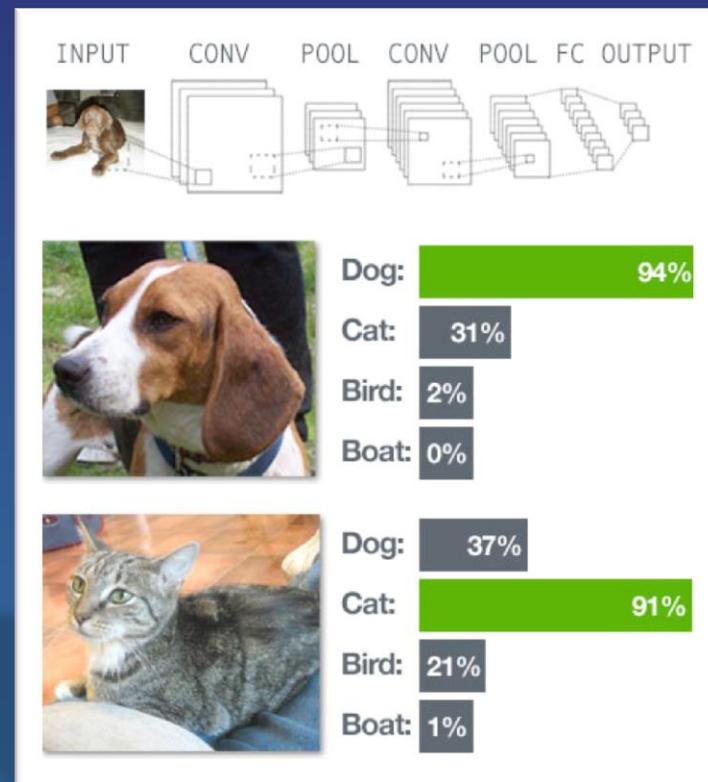
# 마지막 제언 !!

- ❖ 세상에 더 좋은 머신러닝은 없다.
  - 더 적합한 머신러닝만 있을 뿐...
- ❖ 잘 아는 것부터 점진적으로 접근해 나가라.
  - 블랙박스에 맡기는 것이 아니라, 하나씩 처방해 나가는 것이다.
- ❖ 머신러닝은 성능점수를 최적화 하는 것이다.
  - 즉, 성능측정기준을 무엇으로 하느냐가 중요하다.
  - 공부(운동) 잘했어? → 공부(운동)에 최적화 된 아이로 자란다.
  - 단순히 돈 벌고 싶다가 아니라,  
어떤 고객을 대상으로 어떤 상품을 얼마만큼 팔 수 있는가를 검증
- ❖ 딥러닝이 좋은 경우
  - 내가 세상의 모든 데이터를 다 가지고 있을 때
  - 내가 가지고 있는 지식이나 능력으로 해결되지 않을 때
- ❖ 데이터사이언스는 "프로그래밍"이 아니라 "디버깅"이다.
  - 데이터를 넣어보고, 왜 안 되는지를 끊임없이 고민
  - 머신러닝 = 러닝머신 ^^

# Appendix

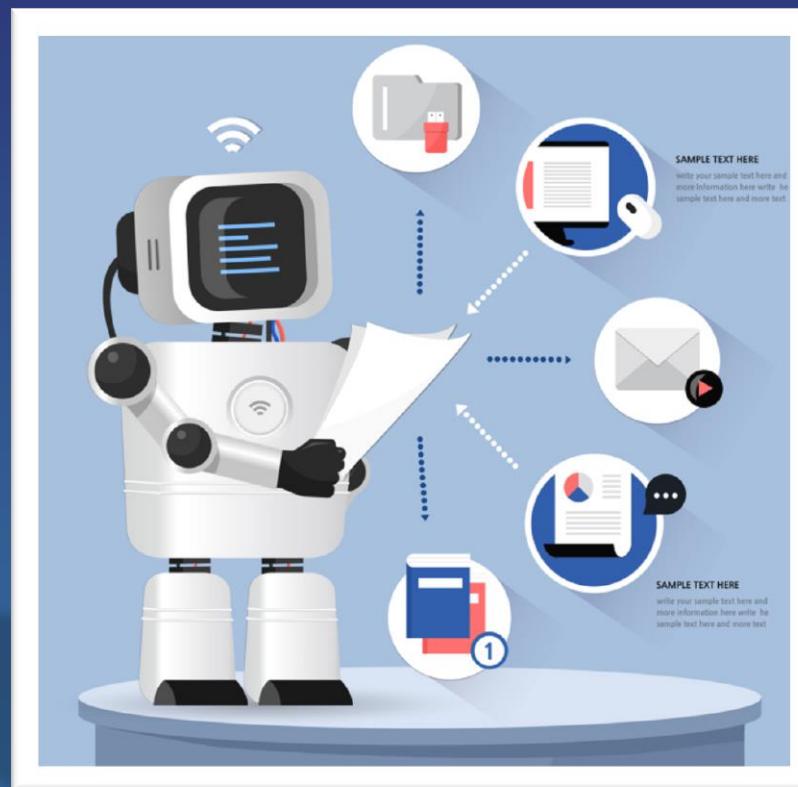
# AI 핵심기술1

## 이미지 인식의 은총알, CNN



# AI 핵심기술2

## 자연어 인식의 최적화, RNN





김 진 수  
CEO, Data Actionist

100-791 서울특별시 중구 청파로 463번지 3F BigData R&D Center

CP. 010-5670-3847      Tel. 02-360-4047      Fax. 02-360-4899

E-mail. [bigpycraft@gmail.com](mailto:bigpycraft@gmail.com)

<http://www.bigpycraft.com>

감사합니다!