

*2016 MACHINE LEARNING CAMP*

# Introduction to Machine Learning

강필성

고려대학교 산업경영공학부

pilsung\_kang@korea.ac.kr

# 목차

I

데이터 사이언스

II

데이터 사이언스 활용 분야

III

데이터 사이언스 속의 기계학습

IV

데이터 사이언스 절차

# Simple Test

1

다음 세 개의 개체 중에서 유사한 두 개를 고르기



(from EBS 다큐프라임 東과 西)

# Simple Test

다음 세 개의 개체 중에서 유사한 두 개를 고르기

2

HD

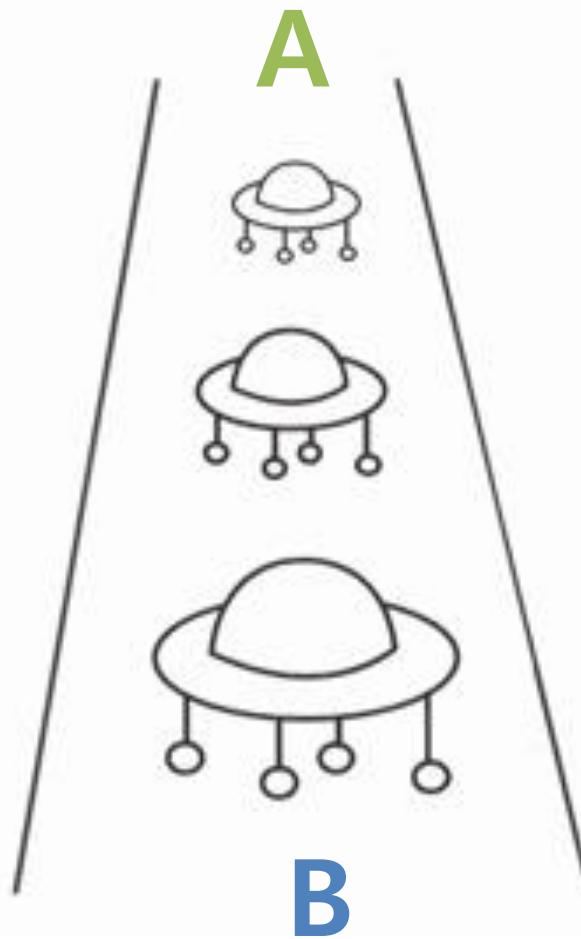


(from EBS 다큐프라임 東과 西)

# Simple Test

어느 쪽이 앞쪽?

3



(from EBS 다큐프라임 東과 西)

# Simple Test

나라면 친구의 사진을 어떻게 찍어줄 것인가?

4

A



B



(from EBS 다큐프라임 東과 西)

# Simple Test

가운데 아이는 행복해 보이나요?

5



(from EBS 다큐프라임 東과 西)

# Simple Test

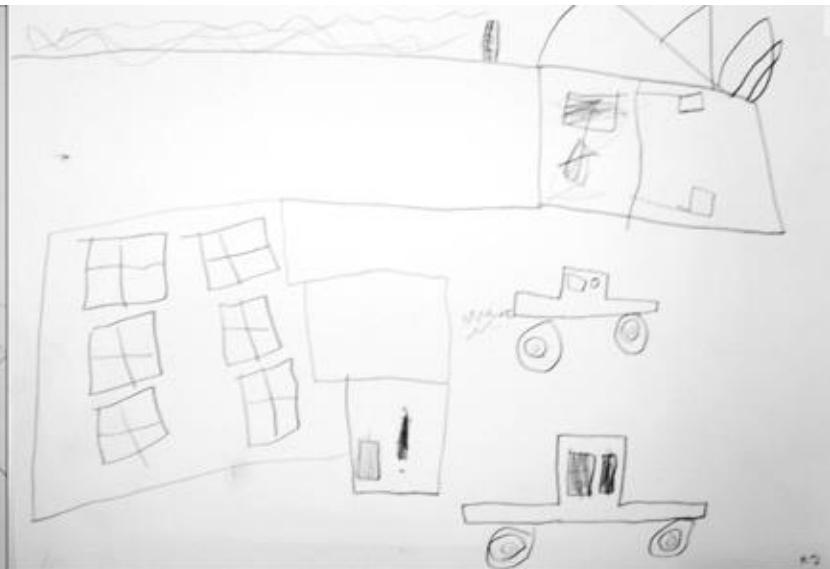
내가 일곱살로 돌아간다면 우리 집을 어떻게 그릴까?

A



6

B



(from EBS 다큐프라임 東과 西)

# Simple Test

손님 찻잔이 비었을 때 어떻게 말할 것인가?



# Simple Test

아래 꽃은 A그룹과 B그룹 중 어디에 속하는가?



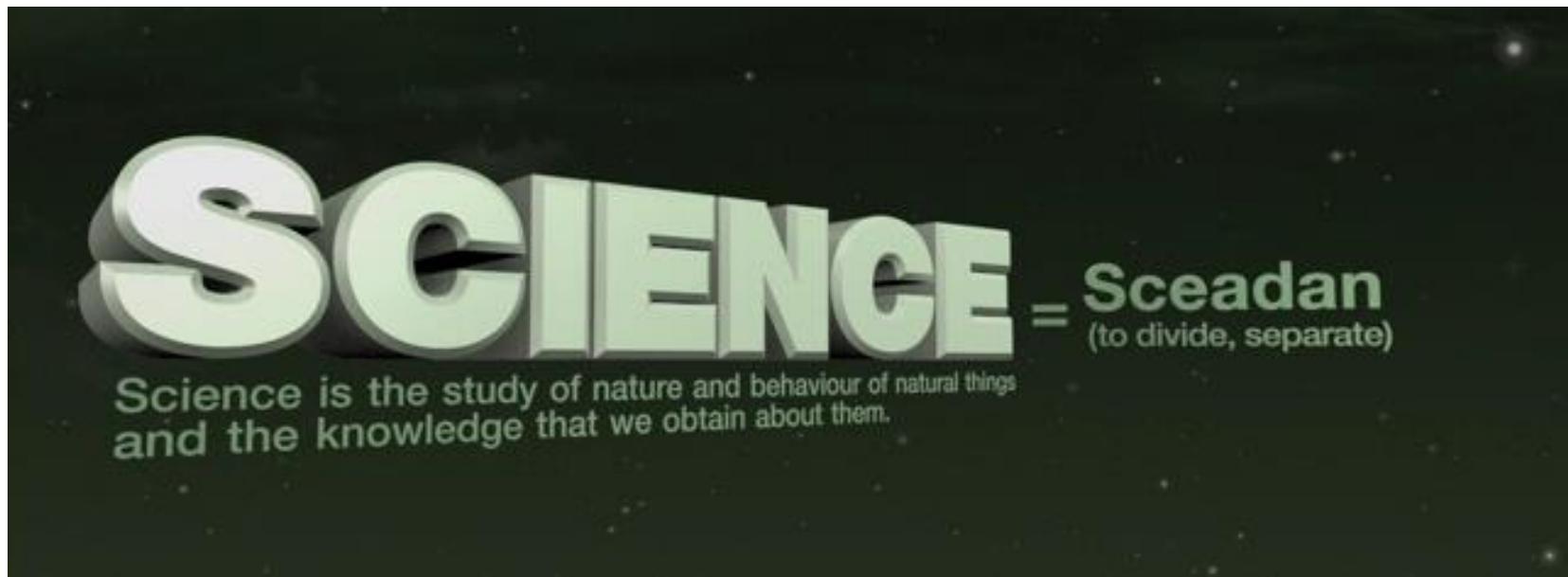
A



B



# 과학(SCIENCE)의 의미



**Science = Sceadan (to devide, separate)**

The study of nature and behavior of natural things and the knowledge that we obtain about them.

knowledge that we obtain about them.

# 데이터 마이닝?

규칙: A속성의 사람들은 인사를 하고 B속성의 사람들은 악수를 한다

A와 B는 무엇일까?



# 데이터 마이닝?



# 데이터 마이닝



# 데이터 마이닝



# 데이터 마이닝



# 데이터 마이닝

## ❖ Definitions

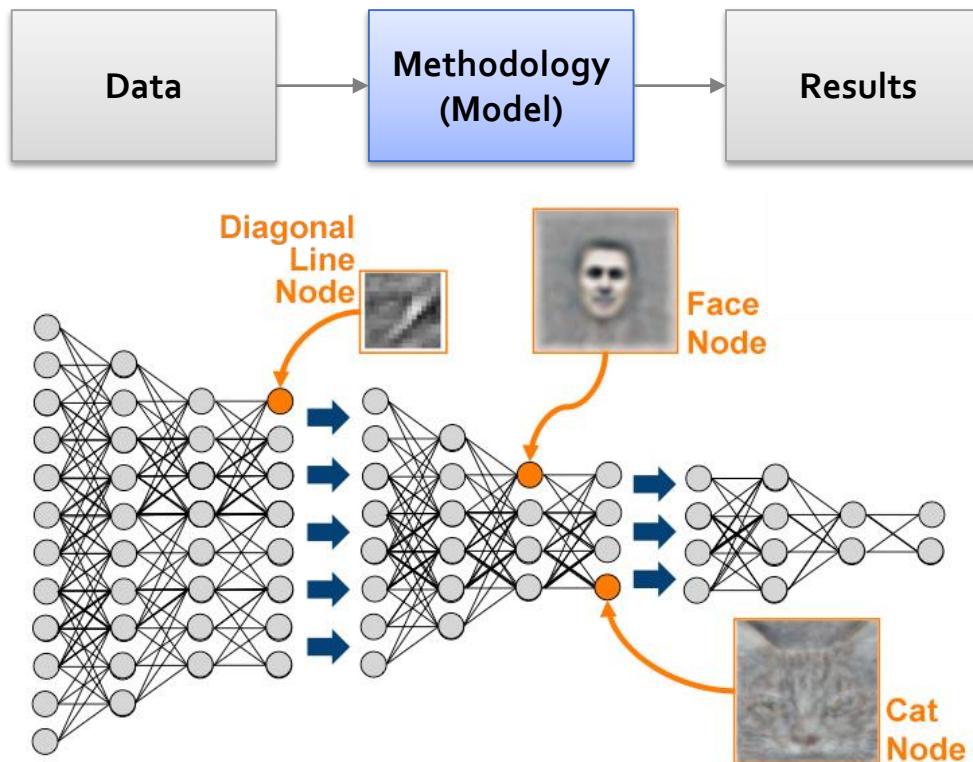


- Extracting useful information from **large datasets**. (Hand et al., 2001)
- The process of **exploration and analysis**, by automatic or semi-automatic means, of **large quantities of data** in order to **discover meaningful patterns and rules**. (Berry and Linoff, 1997, 2000)
- The **process of discovering meaningful new correlations, patterns and trends** by sifting through **large amount data** stored in repositories, using **pattern recognition technologies as well as statistical and mathematical techniques**. (Gartner Group, 2004)

# Machine Learning

## ❖ Definitions

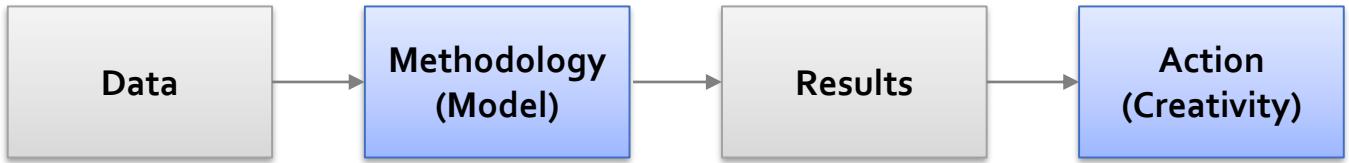
- A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E," – Mitchell (1997)



# 인공 지능(Artificial Intelligence)

## ❖ Definition

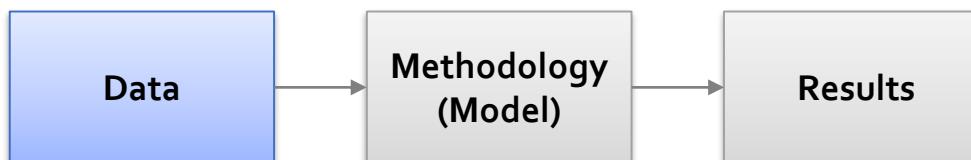
- Computers and computer software that are capable of intelligent behavior
- Intelligent agent perceives its environment and takes actions that maximize its chance of success



# 빅 데이터 (Big Data)

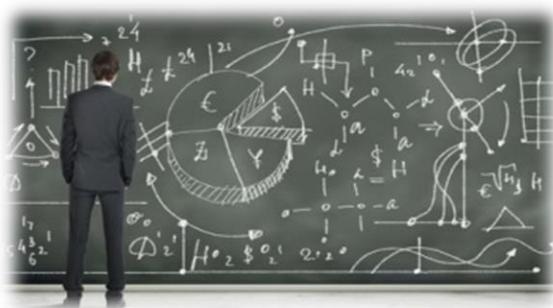
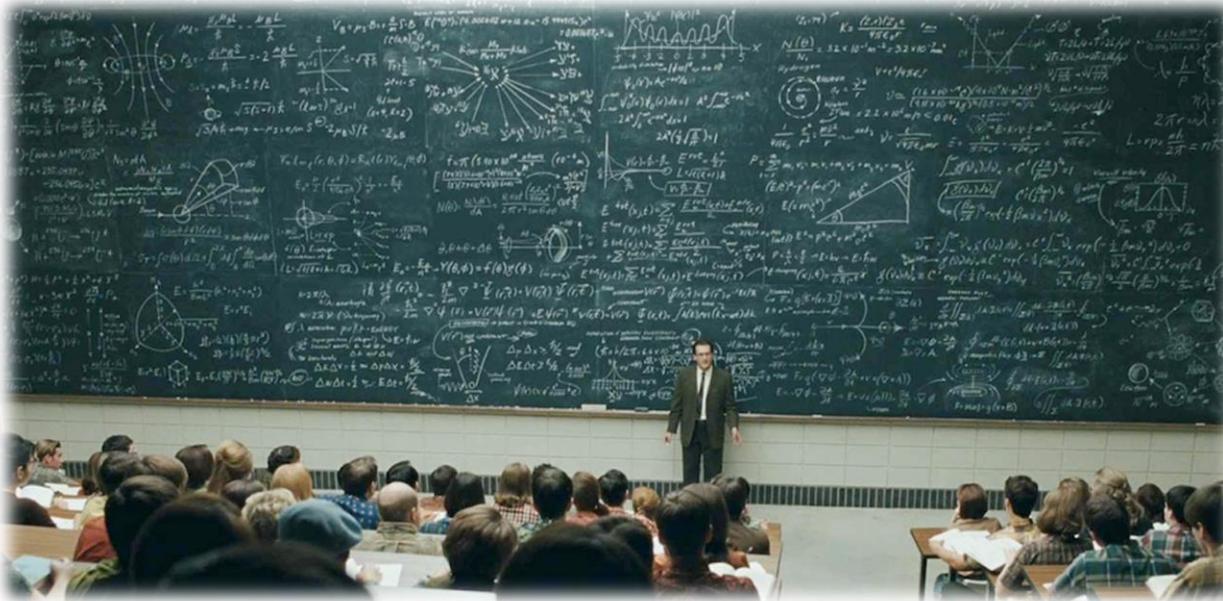
## ❖ Definition

- 데이터베이스 규모에 초점을 맞춘 정의 (McKinsey, 2011)
  - ✓ 일반적인 데이터베이스 SW가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터
- 업무 수행에 초점을 맞춘 정의 (IDC, 2011)
  - ✓ 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처



# 데이터 마이닝: “빅데이터”의 시대

## ❖ Data Scientist: The sexist job of the 21<sup>st</sup> Century

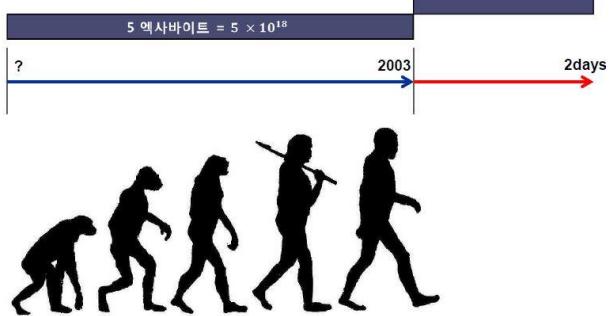


(Harvard Business Review, Oct. 2012, 내용 요약은 최윤섭 팀장 블로그 참조([Link](#))

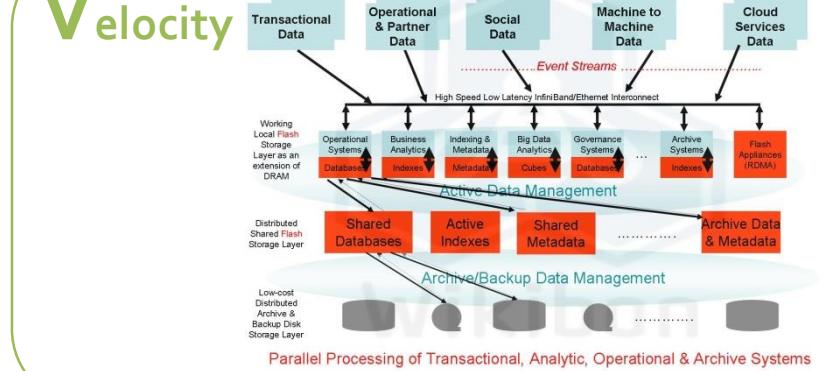
# 데이터 마이닝: “빅데이터”의 시대

## ❖ What is “Big Data”

### Volume



### Velocity



### Variety



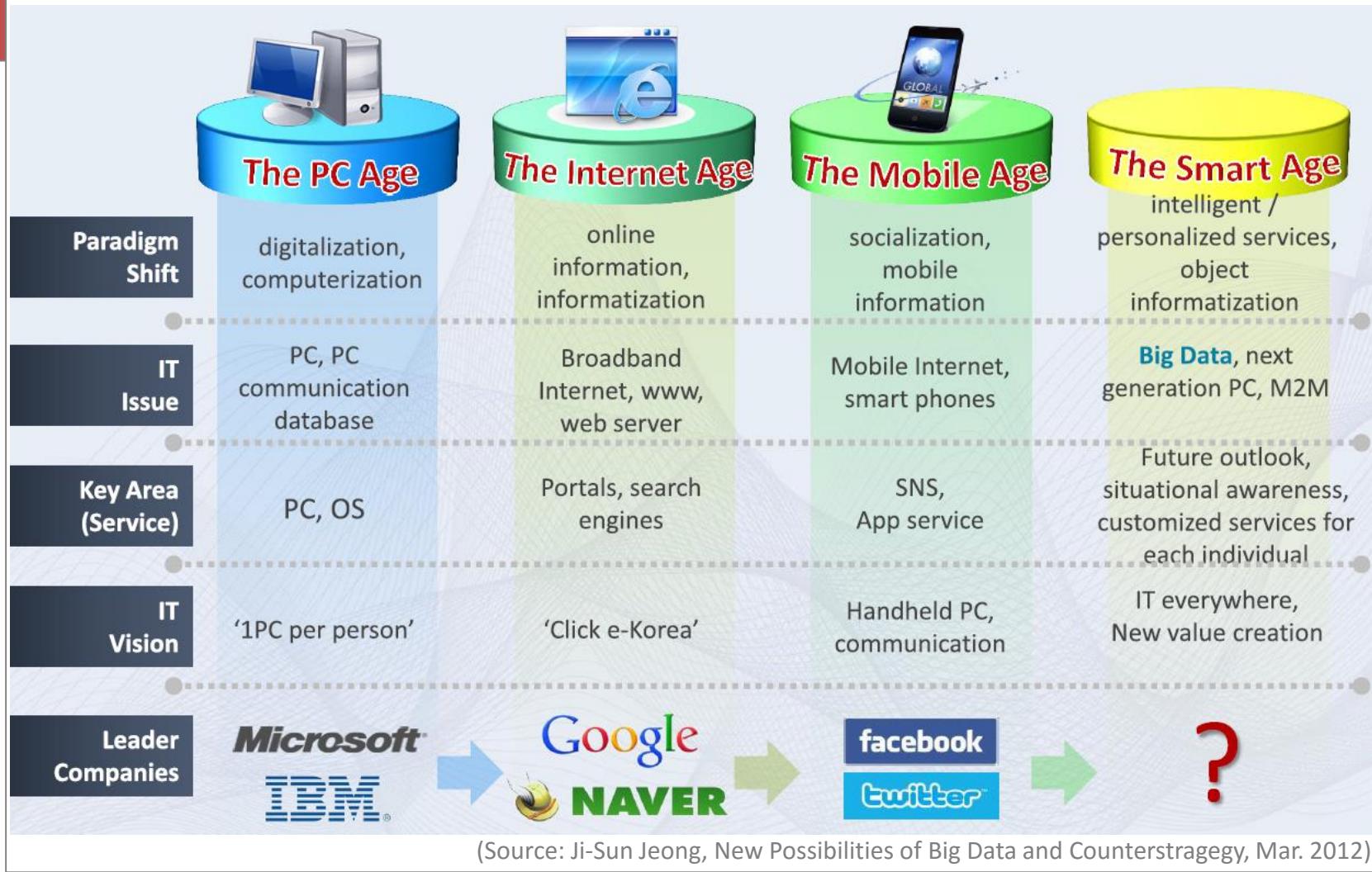
### Value



# 빅데이터 등장 배경

1

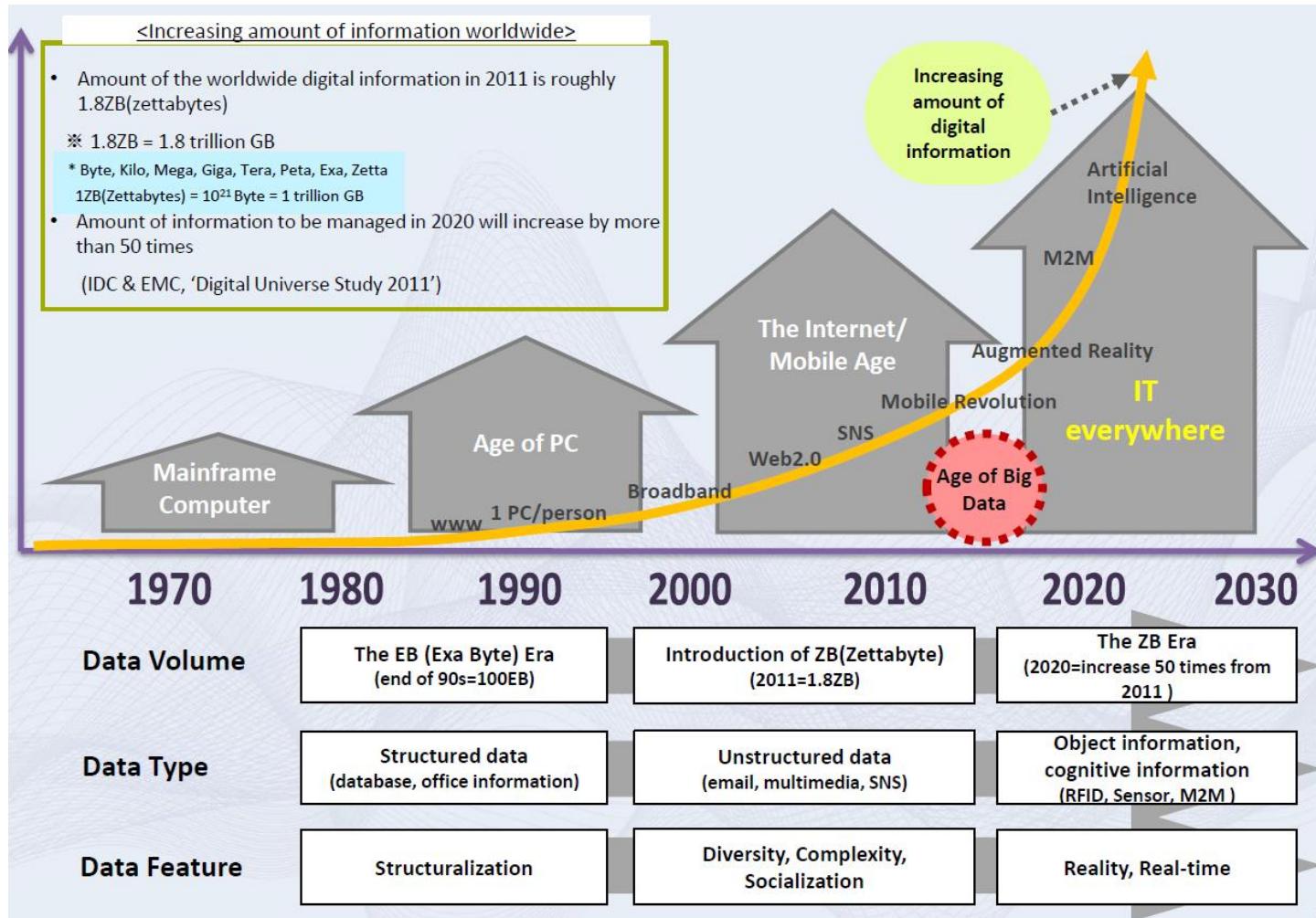
## ICT 패러다임의 변화



# 빅데이터 등장 배경

## 기하급수적으로 증가하는 데이터의 생산 및 저장

2



# 빅데이터 등장 배경

## 데이터의 본질적 가치 증가

3

Type	Institution	Forecast
Economic Feasibility in Industries	Economist (2010)	<input type="checkbox"/> “ Data are becoming the <b>new raw material of business</b> : an economic input almost on a par with capital and labour.”
	Gartner (2011)	<input type="checkbox"/> "Intelligence about Information is the <b>Oil of the 21<sup>st</sup> Century</b> ." Future competitive advantage depends on data. <input type="checkbox"/> Winning organizations understand the <b>stage of the data economy</b> and overcome information silos through effective information sharing.
	McKinsey (2011)	<input type="checkbox"/> Big Data is the <b>next frontier for innovation, competitiveness and productivity</b> <input type="checkbox"/> Big Data will create values worth more than \$600 billion in 5 areas including medicine and public administration
National Competitiveness	US PCAST	<input type="checkbox"/> Advisors emphasize that US government organizations should focus on the strategy for <b>transformation of data into knowledge, and of knowledge into action</b> .
	Singapore	<input type="checkbox"/> Singapore looks to evaluate threatening <b>risks</b> and detect environmental changes based on data

(Source: 김현곤, 빅데이터 국가 비전과 전략, Nov. 2012)

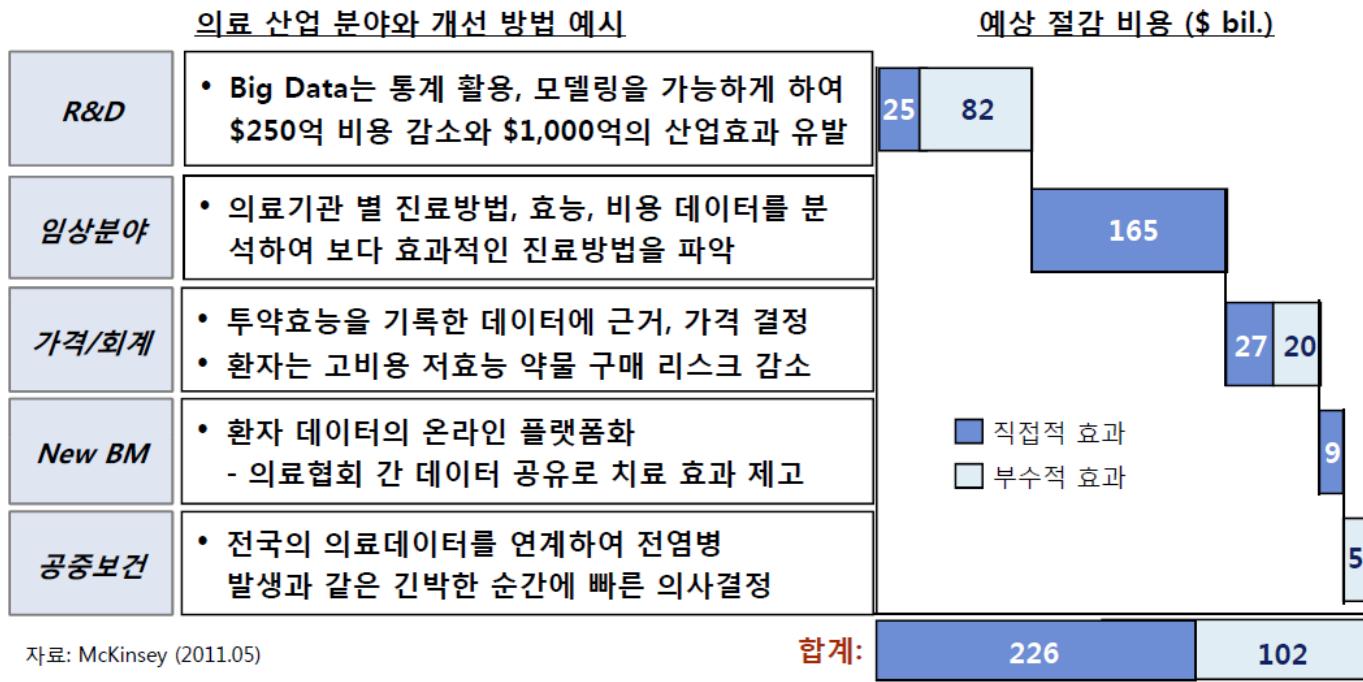
# 빅데이터 등장 배경

## 데이터의 본질적 가치 증가

美 의료부문은 Big Data 활용으로 연간 \$3,300억의 직간접적 비용 절감 효과 기대 가능  
(미 정부 의료 예산의 약 8%에 해당하는 규모)

- Big Data를 이용한 의료 분야의 직간접적 비용 개선 효과는 약 \$3,300억에 이를 전망
- 부수적으로 발생하는 가치는 약 \$1,000억에 이를 것으로 전망됨
- 직접적 효과는 '임상 분야'에서 제일 크게 발생 (약 \$1,650억)

3



# 빅데이터 등장 배경

3

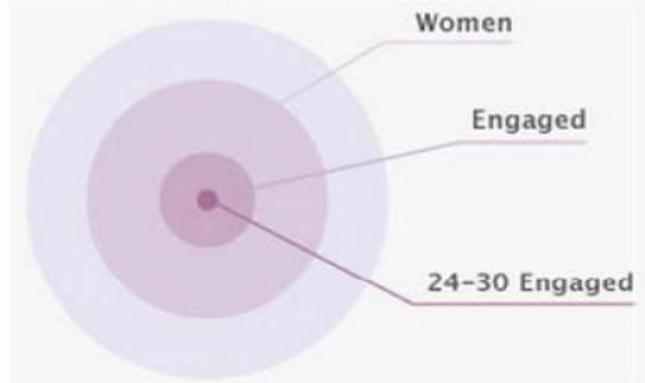
## 데이터의 본질적 가치 증가



- 영국의 O2는 Placecast와 협력하여 LBS 기반 실시간 Starbucks 프로모션을 모바일로 제공
  - O2 가입자 중 약 100만 명이 이 서비스에 가입
  - 가입자가 스타벅스 매장 근처에 도달하면 문자 메시지와 함께 프로모션 쿠폰이 전송됨
- 스마트폰 확산에 따라 Groupon처럼 SNS와 결합된 LBS 기반 모바일 서비스가 급증 추세
  - 스마트폰 사용자 50% 이상이 이미 LBS 기반 프로모션을 통해 쇼핑한 경험을 갖고 있음
- ➔ Big data를 신속, 저렴하게 처리할 수 있는 클라우드 컴퓨팅의 활용성 증가



- FB의 CM Photographic은 이용자가 입력한 검색조건을 즉각 처리하여 광고 타겟 대상 제시
  - 성별은 여성
  - 결혼/연애 상태에서 약혼으로 표시된 가입자
  - 23~30살의 연령



Recently engaged?



자료: Facebook 웹사이트

지난 12개월 동안 CM Photographic는 600달러를 투자해 Facebook에 광고를 냄으로써 4만달러에 이르는 수익을 창출했습니다. 광고를 통해 CM Photographic 웹사이트를 방문하게 된 Facebook 사용자 중 60%는 선도 이용자가 되었고 더 많은 정보에 대한 관심을 적극적으로 표현했습니다.

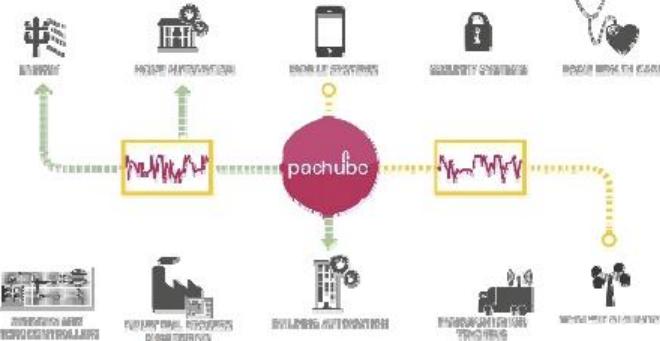
(Source: Big Data 미래를 여는 비밀열쇠, 2012)

# 빅데이터 등장 배경

## 데이터의 본질적 가치 증가

3

### Pachube의 Data App Store 사례



- 2010년 영국에서 창업한 Pachube는 다양한 영역에 (전기, 가전, 휴대폰, 가로등 등) 부착된 센서를 통해 수집된 정보를 저장/분석/제공하는 전문 기업
- 일반적인 수준의 데이터 이용은 대중에 무료로 공개
- App 개발 또는 분석된 Data의 독점적 이용을 원하는 고객들에게 유료로 제공하는 BM

### Sparked의 특화된 Solution 사업



- 2008년 네덜란드에서 창업 후 유럽전역으로 확대 중
- 소에 센서를 부착해 소에 대한 정보를 수집(연간 200MB)
- 기후변화 등 외부 Big Data와의 결합을 통해 축산업자가 소에 대한 움직임, 건강을 수시로 확인 가능한 시스템 제공
- 소 한 두당 세계 최고 수준의 우유생산량 달성을 및 사육 밀도를 높여 더 많은 소를 건강하게 키울 수 있도록 지원

(Source: Big Data 미래를 여는 비밀열쇠, 2012)

# 빅데이터 등장 배경

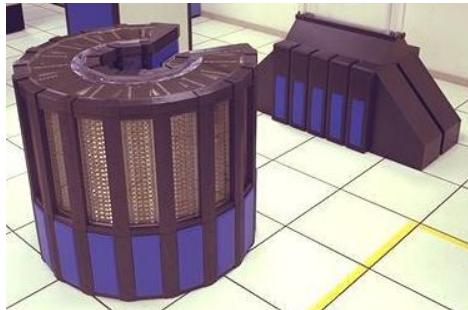
## 연산 능력의 비약적인 발전



# 빅데이터 등장 배경

## 연산 능력의 비약적인 발전

4



품명	무게	CPU	RAM	HD	가격
Cray-2s	2000kg	4개	1Gb	40Gb	270억
Parallel Computing System with 3 PCs	30kg	16개	64Gb	4.5Tb	500만

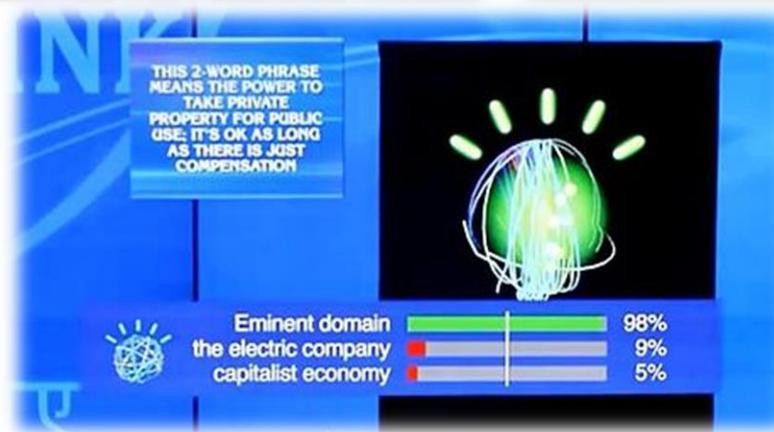
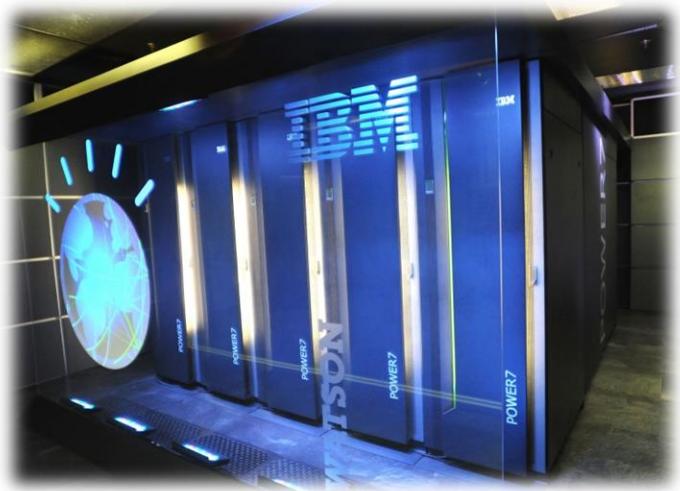
↑ 70배     
 ↓ 4배     
 ↓ 64배     
 ↓ 100배     
 ↑ 5400배

# 빅데이터 등장 배경

## 연산 능력의 비약적인 발전

IBM's super computer Watson won an overwhelming victory in final Jeopardy! In Feb. 2011

4



# 빅데이터 등장 배경

## 점점 심화되는 경쟁체제



The screenshot shows the official Netflix Prize website. At the top, the Netflix logo is on the left, and a large yellow banner across the middle reads "Netflix Prize" in white text. To the right of the banner is a red "COMPLETED" stamp. Below the banner, there's a navigation bar with links for "Home", "Rules", "Leaderboard", and "Update". The main content area features a dark background with a blurred image of a movie screen showing a movie cover for "The Big One". On the right side of this area, a large blue "Congratulations!" message is displayed. Below this message, there is explanatory text about the prize and links to the Grand Prize winner's algorithm, the Leaderboard, and the Forum.

5

**Netflix Prize**

Home Rules Leaderboard Update

**Congratulations!**

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about [their algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

We applaud all the contributors to this

# 빅데이터 등장 배경

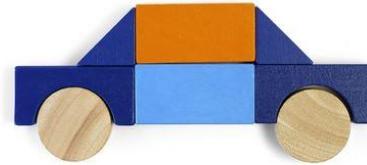
## 점점 심화되는 경쟁체제

5

Active Competitions		Active Competitions	
All Competitions			
		<b>Allstate Purchase Prediction Challenge</b> Predict a purchased policy based on transaction history	2 months 185 teams \$50,000
		<b>March Machine Learning Mania</b> Tip-off college basketball by predicting the 2014 NCAA Tournament	15 days 150 teams \$15,000
		<b>Flu Forecasting</b>  Predict when, where and how strong the flu will be	3.9 days 50 teams
		<b>Walmart Recruiting - Store Sales Forecasting</b> Data Scientist at Walmart Various Locations	2 months 90 teams Jobs
		<b>Galaxy Zoo - The Galaxy Challenge</b> Classify the morphologies of distant galaxies in our Universe	35 days 203 teams \$16,000
		<b>Loan Default Prediction - Imperial College Lon...</b> Constructing an optimal portfolio of loans	14 days 501 teams \$10,000
		<b>PAKDD 2014 - ASUS Malfunctional Componen...</b> Predict malfunctioning components of ASUS notebooks	32 days 348 teams \$8,500
		<b>CONNECTOMICS</b> Reconstruct the wiring between neurons from fluorescence imaging of neural activity	2 months 48 teams \$3,000

[Competition Details](#) » [Get the Data](#) » [Make a submission](#)

Predict a purchased policy based on transaction history



As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. This is represented in this challenge as a series of rows that include a customer ID, information about the customer, information about the quoted policy, and the cost. Your task is to predict the purchased coverage options using a limited subset of the total interaction history. If the eventual purchase can be predicted sooner in the shopping window, the quoting process is shortened and the issuer is less likely to lose the customer's business.

Using a customer's shopping history, can you predict what policy they will end up choosing?

Started: 3:04 pm, Tuesday 18 February 2014 UTC  
Ends: 11:59 pm, Monday 19 May 2014 UTC (90 total days)

(<https://www.kaggle.com>)

# 빅데이터 등장 배경

데이터 분석을 가능하게 하는 인프라와 (오픈소스) 소프트웨어

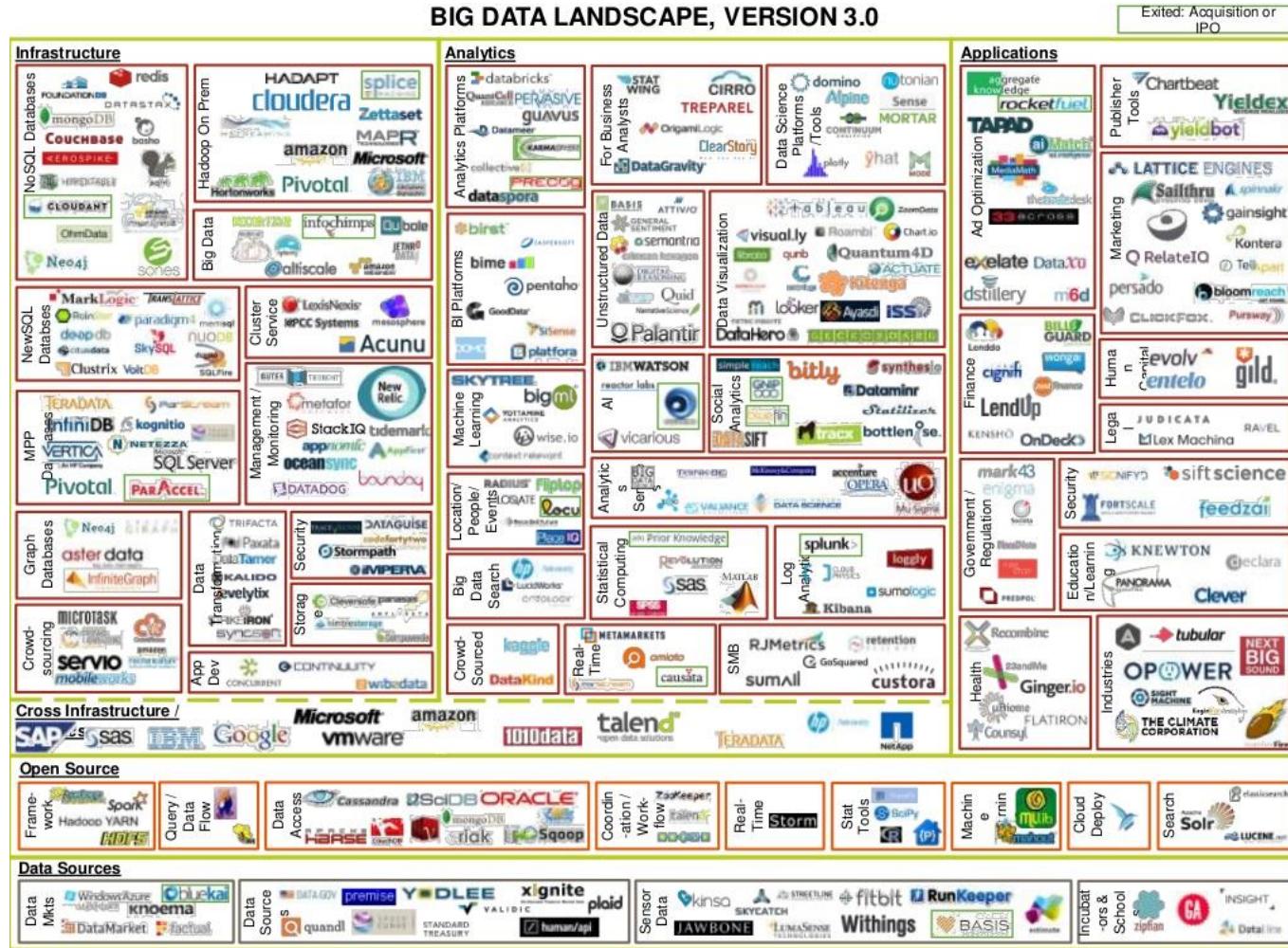


6

# 빅데이터 등장 배경

## 데이터 분석을 가능하게 하는 인프라와 (오픈소스) 소프트웨어

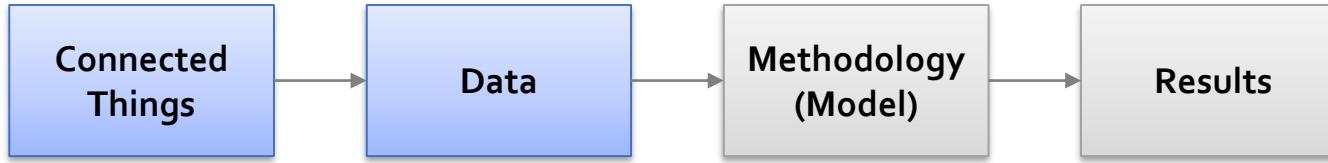
BIG DATA LANDSCAPE, VERSION 3.0



# 사물 인터넷 (Internet of Things: IoT)

## ❖ Definition

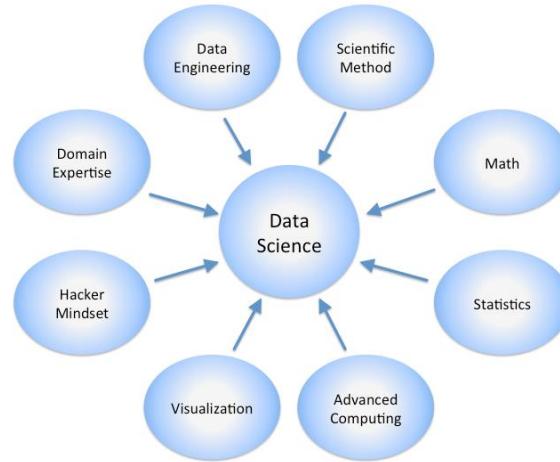
- The network of physical objects embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data



# 데이터 사이언스란?

## ❖ 데이터 사이언스

- 다양한 학제간 학문이 융합되어 데이터 기반 의사결정 및 문제 해결을 목적으로 하는 학문



# 데이터 사이언스 관련 학문

## Statistics

- Deterministic & parametric approaches

## Computer science

- Non-parametric approaches
  - Machine learning pattern recognition algorithms

## Data Science

## Database

- Efficient data structure & query
  - Database management

## Industrial Engineering

- Operations research and mathematical optimization
  - Business intelligence

# 목차

I

데이터 사이언스

II

데이터 사이언스 활용 분야

III

데이터 사이언스 속의 기계학습

IV

데이터 사이언스 절차

# 데이터 사이언스 활용 분야

1

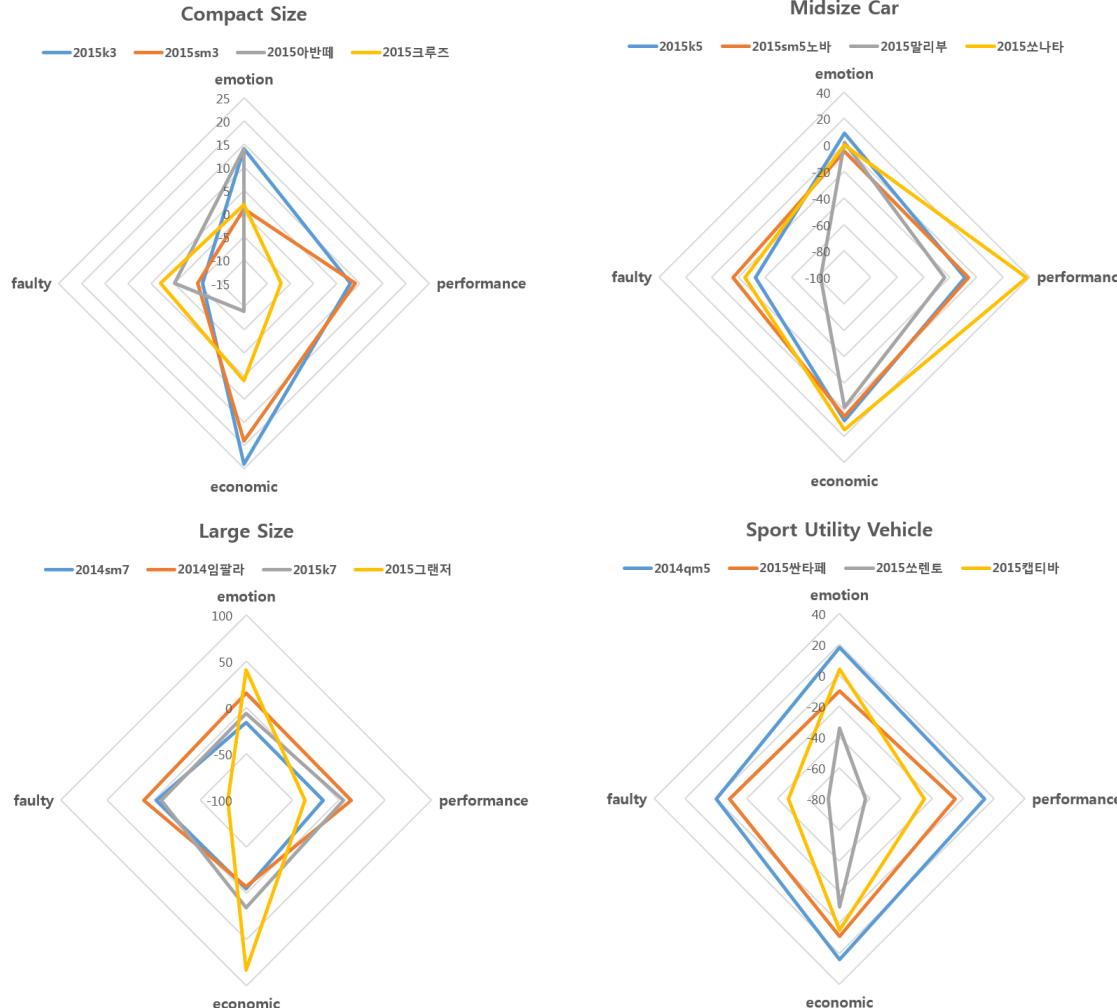
## 보다 직관적인 이해를 돋기 위한 시각화



# 데이터 사이언스 활용 분야

1

## 보다 직관적인 이해를 돋기 위한 시각화



```
> perf_cars$'2015쏘나타'
ngram morpho_pattern
166 2 트 보
266 2 지동 작
270 2 가솔린 좋
290 2 경쟁사 별벗
1064 2 것 같
267 2 작 미만
292 2 오 게임
275 3 조용 진동 작
279 3 운행자 가솔린 좋
294 3 옵션 경쟁사 별벗
```

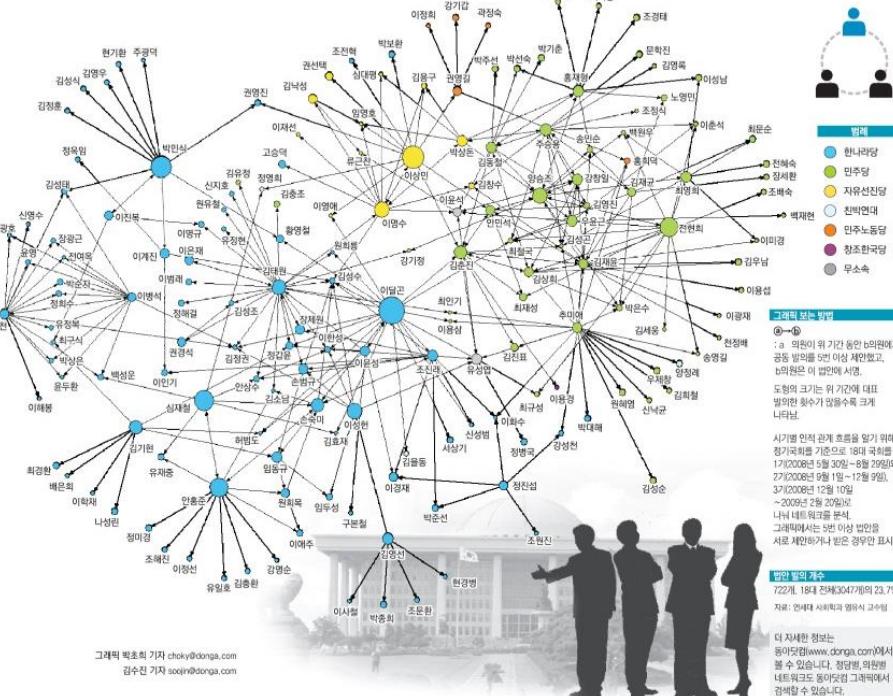
```
> econ_cars$'2015그랜저'
ngram morpho_pattern
1899 2 수 있
16504 2 차 틀리
1017 2 좋 차
3053 2 시승 보
16505 2 틀리 복합연비
16508 3 디젤 자 틀리
6889 2 사 것
316 2 타 보
3780 2 연비 좋
12577 2 연비 잘나
```

# 데이터 사이언스 활용 분야

1

## 보다 직관적인 이해를 돋기 위한 시각화

의원 입법발의를 통해 본  
18대 국회의원 인적 네트워크 3기 : 2008년 12월 10일~2009년 2월 20일



18대 국회의원 시기별 대표발의 '허브 의원' 순위 1기: 2008년 5월 30일~8월 29일 / 2기: 2008년 9월 1일~12월 9일 / 3기: 2008년 12월 10일~2009년 2월 20일



'허브 의원' 정수는 영유식 교수팀이 자체 개선한 것으로 1) 수치가 높을수록 많은 법의 때 의원들 사이에서 중개 역할을 많이 한 것으로 볼 수 있다.

**politiz**

"Politiz.org"

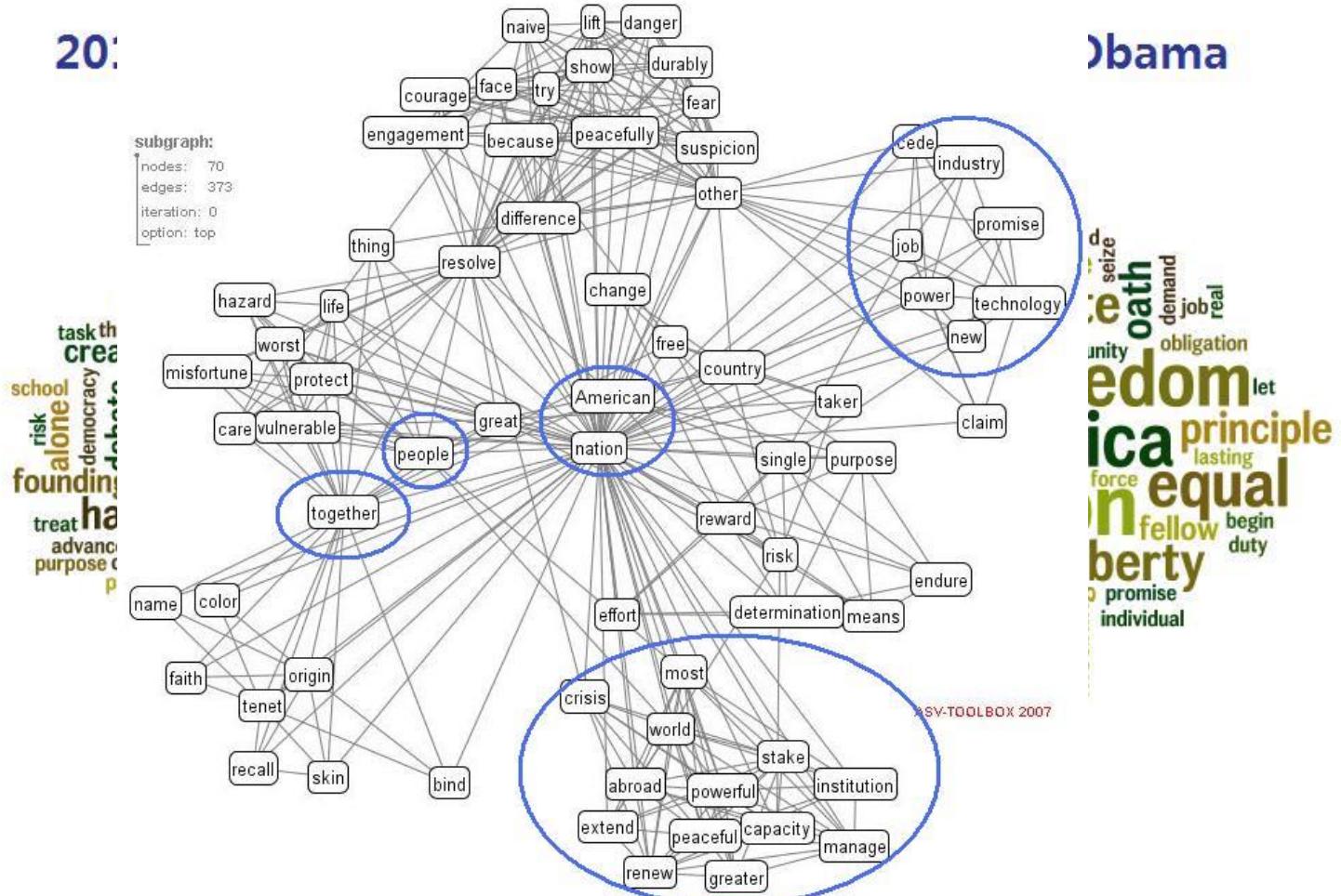
# 데이터 사이언스 활용 분야

1

## 보다 직관적인 이해를 돋기 위한 시각화

2013년 오바마 취임 연설 주제어 연결망 - nation

20:

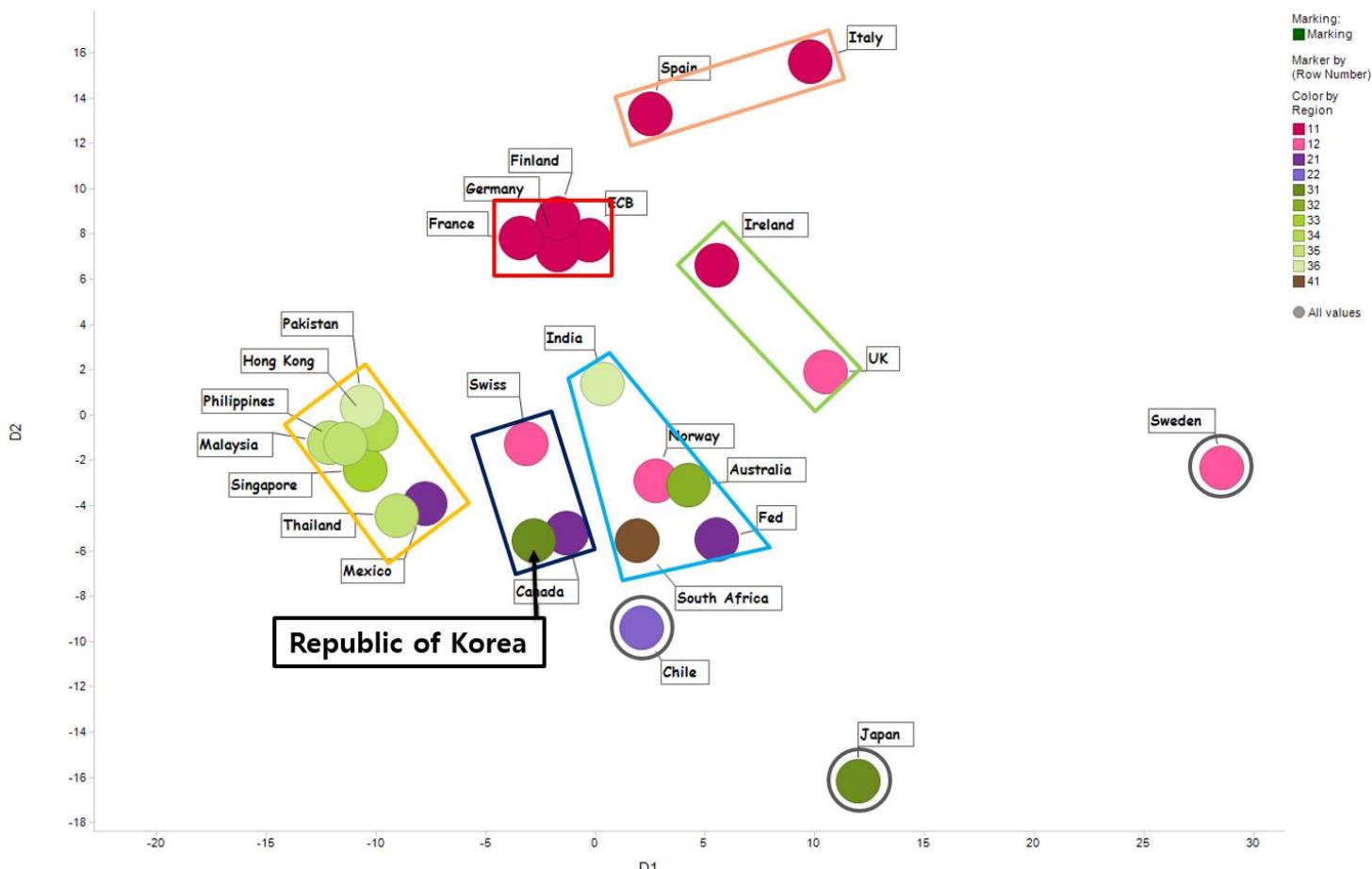


# 데이터 사이언스 활용 분야

1

## 보다 직관적인 이해를 돋기 위한 시각화

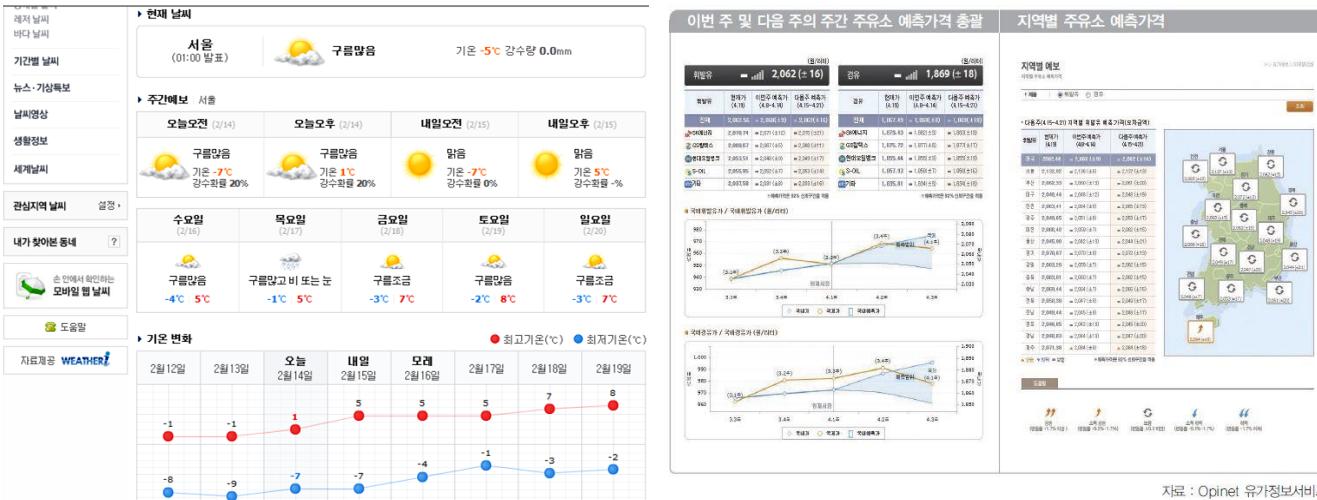
세계 각국 중앙은행 총재 연설문 분석



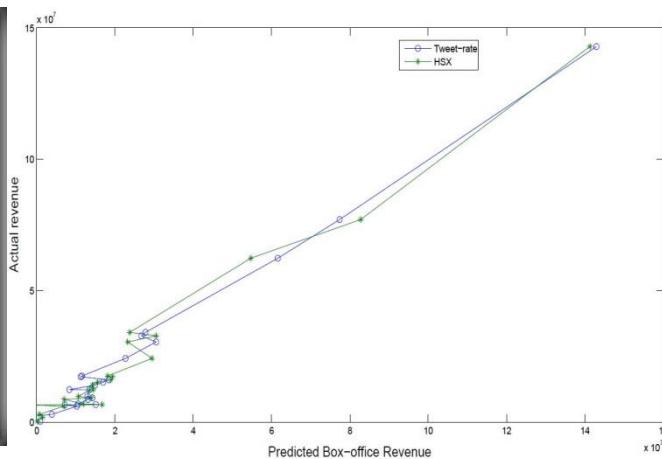
# 데이터 사이언스 활용 분야

2

## 미래의 예측, 진단 및 탐지



자료 : Opinet 유가정보서비스

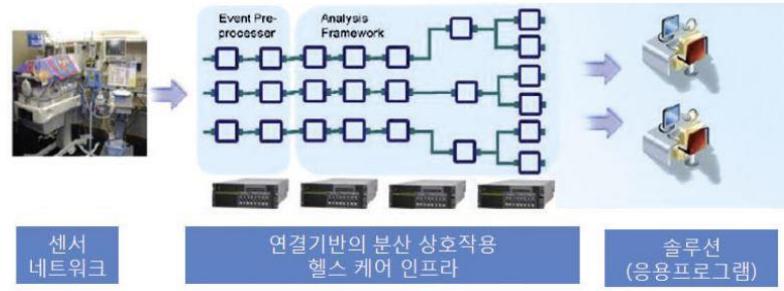


Asur and Huberman (2010) Predicting the Future with Social Media, WI-IAT10: 492-499

# 데이터 사이언스 활용 분야

## 미래의 예측, 진단 및 탐지

2



자료 : Anjul Bhambhani, Smarter Analytics for Big Data, IBM, 2011.6.7

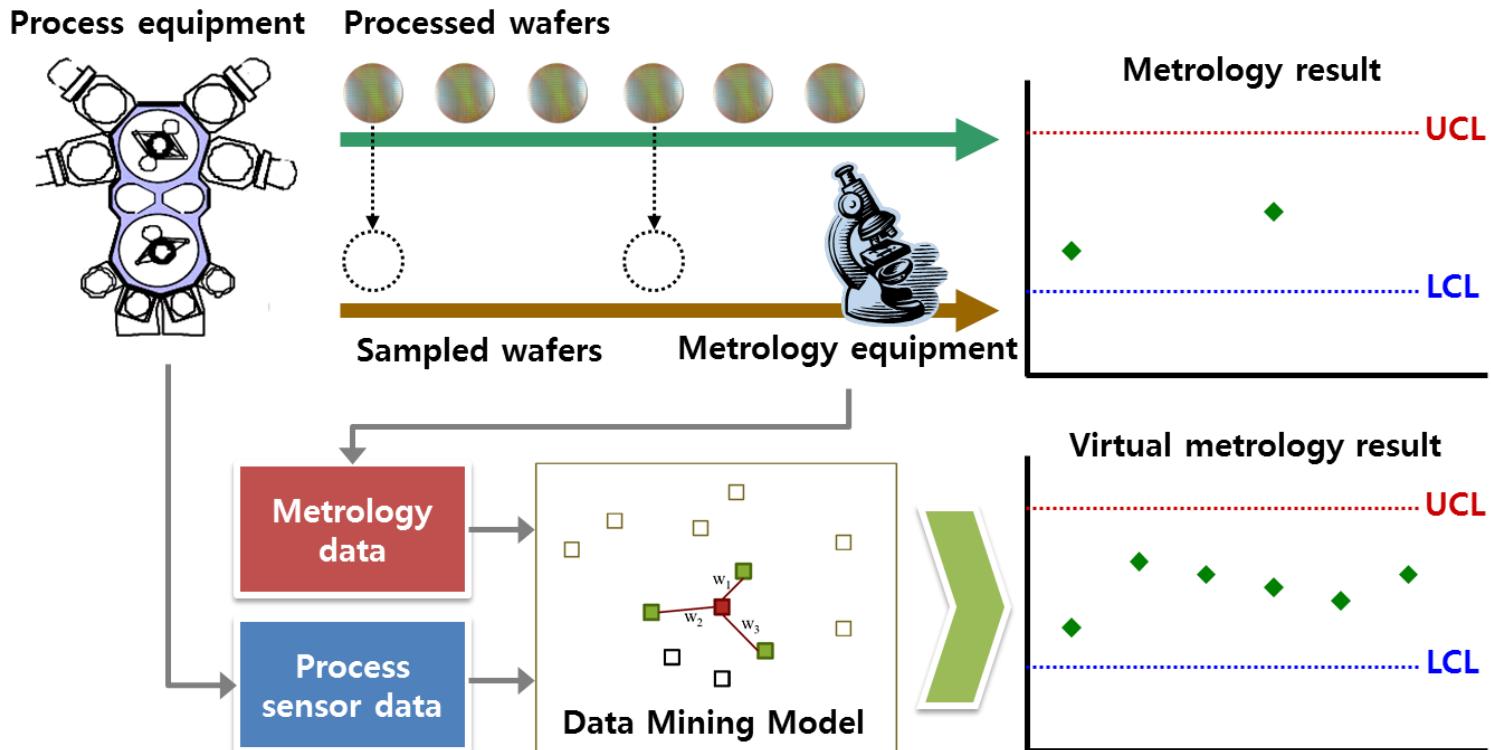
- 신생아의 혈압, 체온, 심전도, 혈중산소포화도 등 미숙아 모니터링 장비에서 생성되는 환자당 일 9,000만 건 이상의 생리학 데이터스 트림 실시간 분석
- 의료진보다 24시간 전에 감염 사실을 밝혀냄으로써 조기 치료 가능

# 데이터 사이언스 활용 분야

## 미래의 예측, 진단 및 탐지

2

### Virtual Metrology in Semiconductor Manufacturing



# 데이터 사이언스 활용 분야

## 일상생활에서의 의사결정 지원



Pilsung's Amazon.com | Today's Deals | Gift Cards | Help

Shop by Department ▾

Search

Books ▾

data mining and business intelligence

Go

kindle fire HD  
from \$199

Hello, Pilsung  
Your Account ▾

Join Prime ▾

0 Cart ▾

### Frequently Bought Together

3



Price for both: \$121.17

Add both to Cart

Add both to Wish List

Show availability and shipping details

This item: Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Galit Shmueli Hardcover \$89.67

Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Gordon S. Linoff Paperback \$31.50

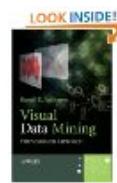
### Customers Who Bought This Item Also Bought

Page 1 of 25



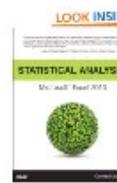
Data Mining Techniques:  
For Marketing, Sales, ...  
Gordon S. Linoff

(34)  
Paperback  
\$31.50



Visual Data Mining: The  
VisMiner Approach  
Russell K. Anderson

Hardcover  
\$62.36



Statistical Analysis:  
Microsoft Excel 2010  
Conrad Carlberg

(10)  
Paperback  
\$22.53



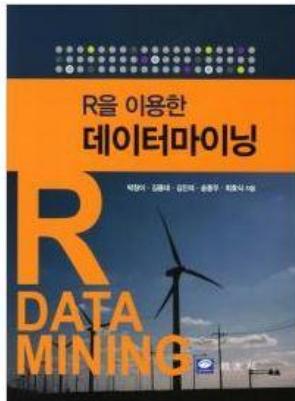
Customer Relationship  
Management: A ...  
V. Kumar

(2)  
Paperback  
\$93.75

# 데이터 사이언스 활용 분야

## 일상생활에서의 의사결정 지원

3

[크게보기](#)

이 책의 영업점 진열 위치

광화문 | 강남 | 잠실 | 분당 | 목동

[다른영업점 보기](#) | [영업점 전체재고](#)

정 가 : 22,000원

판 매 가 : **22,000원** [0%↓ 0원 할인] [장가계 Free](#)청구할인가 : **19,800원** 최대 10%할인 현대카드 M포인트 결제할인 [안내](#)통합포인트 : 220원 [1%적립] [안내](#) 0.5% 추가적립 [안내](#)13년 2학기 대학교재전 최저가 막강혜택! **캠퍼스 라이프 업그레이드!**· 배송비 : 무료 [배송비 안내](#)· 도착예정일: 지금 주문하면 **내일(29일,목) 도착 예정** > [도착예정일 안내](#)서울 종로구 종로1가 교보생명빌딩 [지역변경](#)

서울,수도권

부산

대구,창원

24시까지 주문하면  
**내일(29일,목) 도착**24시까지 주문하면  
**내일(29일,목) 도착**24시까지 주문하면  
**내일(29일,목) 도착**

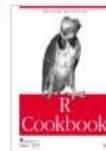
### ■ 이 책과 함께 구매한 도서

 전체선택 장바구니에 담기 보관함에 담기

R을 이용한 누구나 하는 데이터마이닝 기법과 응용  
**15,000원**  
[0%+3%P]



**28,000원**  
[0%+1%P]



**28,000원**  
[10%+10%P]



데이터마이닝(비즈니스 인텔리전스를 위한)  
**33,000원**  
[0%+3%P]

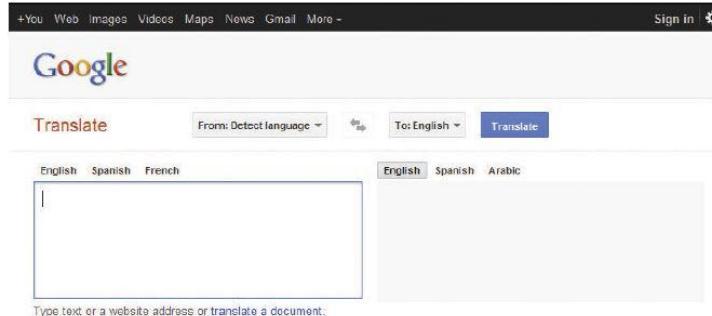


**27,000원**  
[0%+1%P]

# 데이터 사이언스 활용 분야

## 삶의 질 향상

### 구글 번역 홈페이지와 애플리케이션



4

- 언어의 문법적 구조를 컴퓨터가 이해할 수 있는 로직으로 변환하는 기존의 방법 탈피
- 6개 국어로 번영되는 UN 회의록과 23개 국어로 번역된 유럽의회 회의록을 번역 엔진에 입력한 뒤 통계적 추론 기법을 학습

# 데이터 사이언스 활용 분야

## 삶의 질 향상

4



- 다양한 경로로부터 축적된 민원 데이터를 종합적/체계적으로 분석하여 정책에 활용
- 키워드 추출, 연관규칙 분석, 군집화 등 의미기반 민원분석을 통해 선제적 대응

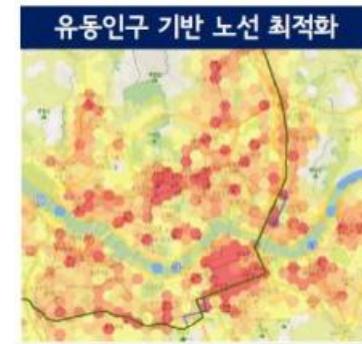
# 데이터 사이언스 활용 분야

## 삶의 질 향상

4



- 데이터에 의한 정량적 유동인구 분포도 작성: 서울시를 1km 반경의 1,250개 헥사셀 단위로 구분 → KT 휴대전화 이력 데이터로 심야시간 (0시~5시) 통화량 분석 → 유동 인구 기반 노선 최적화 및 배차간격 조정



# 목차

I

데이터 사이언스

II

데이터 사이언스 활용 분야

III

데이터 사이언스 속의 기계학습

IV

데이터 사이언스 절차

# 기계 학습: Machine Learning

## ❖ 정의

- A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E," – Mitchell (1997)

### 지도학습: Supervised Learning

- 목적: 특정한 타겟을 예측하는 것
- 독립변수X와 종속변수Y사이의 관계(인과관계 혹은 상관관계)를 찾음
- 과거에 타겟 정보가 있는 데이터를 사용하여 모델 학습
- 새로운 데이터를 사용하여 모델 평가

### 비지도학습: Unsupervised Learning

- 데이터에 내재된 특징을 분석
- 데이터의 분포 추정, 고객집단의 구분, 연관규칙 분석 등
- 예측하고자 하는 지정된 타겟 변수가 존재하지 않음

# 기계학습의 종류: 1st Dimension

## ❖ Target(정답)의 유무에 따른 구분

- Supervised, semi-supervised, and unsupervised

### Supervised Learning

A given dataset  $\mathbf{X}$  &  $\mathbf{Y}$

	Var. 1	Var. 2	...	Var. d	$\longrightarrow$	$\mathbf{Y}$
Ins. 1	..	..	...	..		..
Ins. 2	..	..	...	..	$y = f(x)$	..
...	...	...	...	...		...
Ins. N	..	..	..	..		..

### Semi-supervised Learning

A given dataset  $\mathbf{X}$  &  $\mathbf{Y}$

	Var. 1	Var. 2	...	Var. d	$\longrightarrow$	$\mathbf{Y}$
Ins. 1	..	..	...	..		..
Ins. 2	..	..	...	..	$y = f(x)$	..
...	...	...	...	...		...
Ins. N	..	..	..	..		..
...	...	...	...	...		...
...	...	...	...	...		...
...	...	...	...	...		...
...	...	...	...	...		...
Ins. M	..	..	..	..		..

### Unsupervised Learning

A given dataset  $\mathbf{X}$

	Var. 1	Var. 2	...	Var. d
Ins. 1	..	..	...	..
Ins. 2	..	..	...	..
...	...	...	...	...
Ins. N	..	..	..	..

## 지도학습 vs. 비지도학습

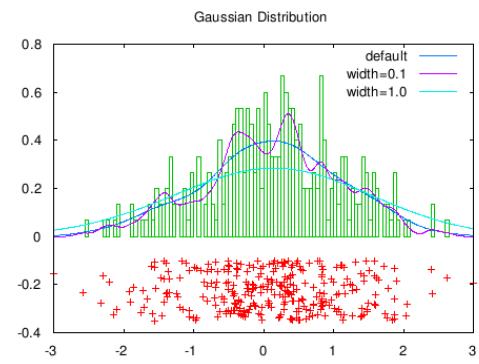
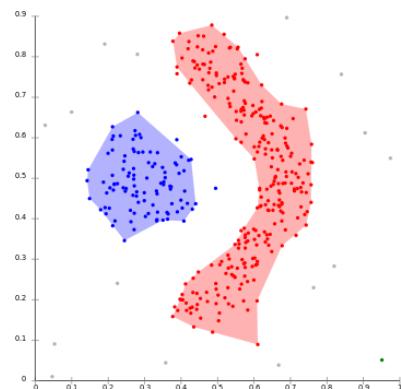
# ❖ 비지도학습

## A given dataset **X**

	Var. 1	Var. 2	...	Var. d
Ins. 1	..	..	...	..
Ins. 2	..	..	...	..
...	...	...	...	...
Ins. N	..	..	..	..

# Unsupervised learning

- Explores intrinsic characteristics
  - Estimates underlying distribution
  - Density estimation, clustering,  
association rule mining, network (graph)  
analysis, etc.



| 이 책과 함께 구매한 도서



# 지도학습 vs. 비지도학습

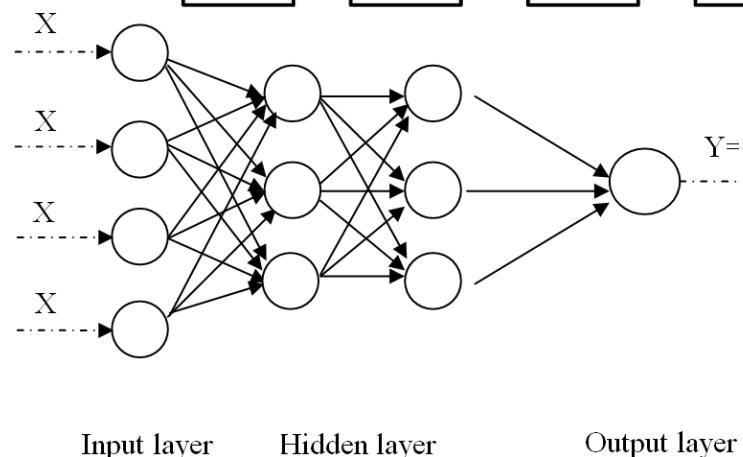
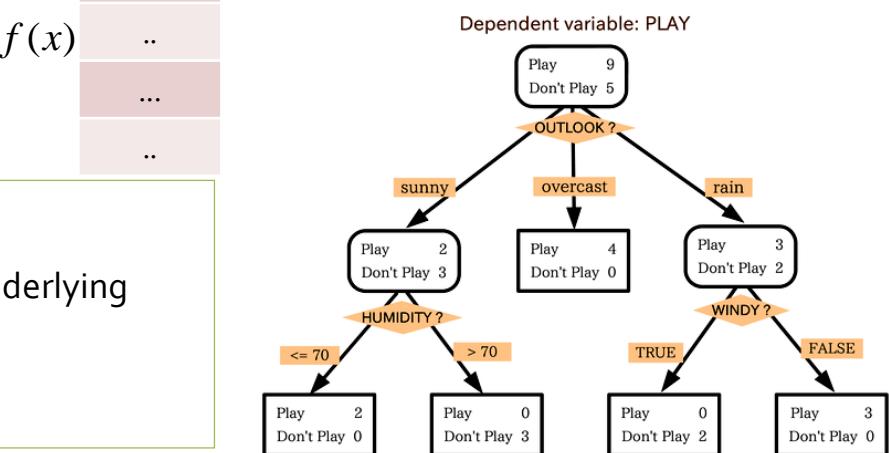
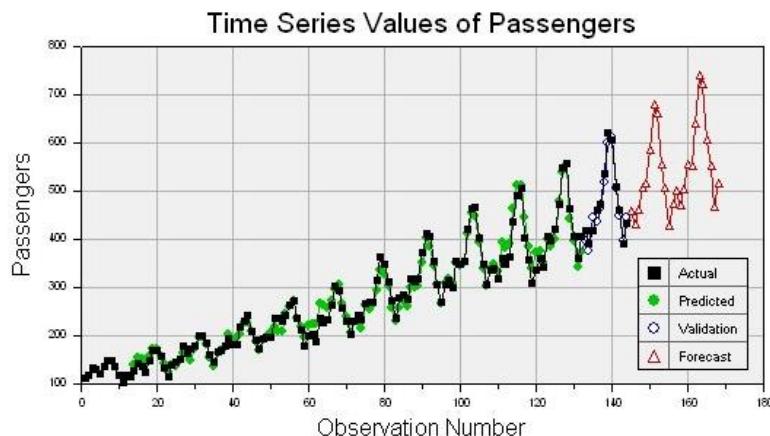
## ❖ 지도학습

A given dataset X & Y

	Var. 1	Var. 2	...	Var. d	→	Y
Ins. 1	..	..	...	..		..
Ins. 2	..	..	...	..	$y = f(x)$	..
...	...	...	...	...		...
Ins. N	..	..	..	..		..

### Supervised learning

- Finds relations between X and Y: estimate the underlying function  $y = f(x)$
- Classification, regression, novelty detection



# 기계학습의 종류: 2<sup>st</sup> Dimension

## ❖ 학습 목적에 따른 구분

### ▪ 분류(Classification)

- 명목형(categorical) 변수를 예측하는 방법론 (예: 익일 주식 상승/하락 예측)
- 로지스틱 회귀분석, 의사결정나무, 인공신경망, Support Vector Machine 등

### ▪ 회귀(Regression)

- 연속형(continuous) 변수를 예측하는 방법론 (예: 익일 주식 가격 예측)
- 다중선형회귀분석, 인공신경망, Support Vector Regression, 가우시안 프로세스 회귀 등

### ▪ 군집화(Clustering)

- 유사한 개체들의 집단을 판별하는 방법론
- K-평균 군집화, 계층적 군집화 등

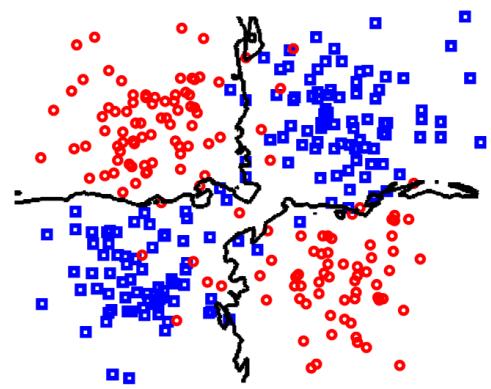
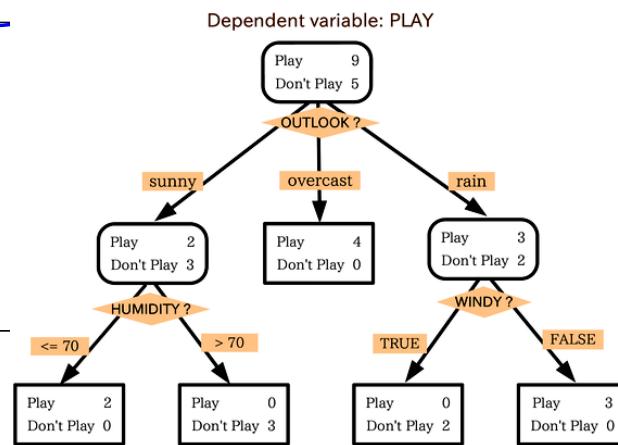
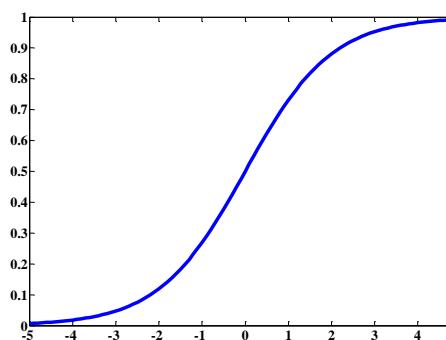
# 목적에 따른 기계학습 방법론

1

## 분류: Classification

- 목적: 범주형 타겟 변수를 예측

- ✓ 구매/비구매, 사기 거래/정상거래, 양성 반응/음성 반응 등
- ✓ 데이터의 각 행은 레코드에 해당하며 각 열은 설명변수에 해당
- ✓ 이진 분류(binary classification) 문제가 대다수



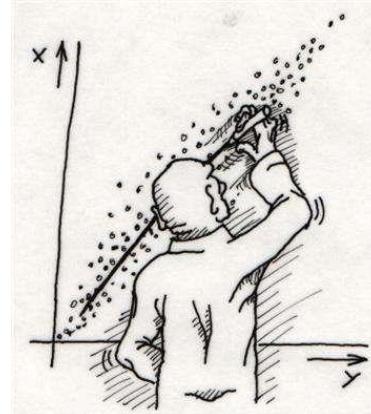
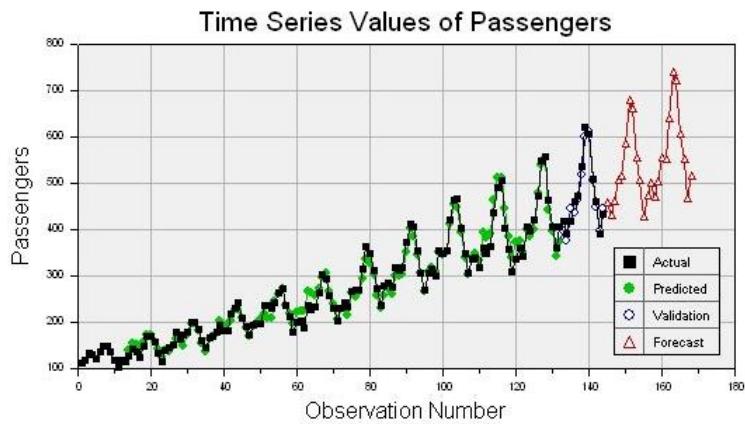
# 목적에 따른 기계학습 방법론

## 회귀: Regression

2

- 목적: 연속형 타겟 변수를 예측

- ✓ 구매 금액, 매출액, 이용 고객 수 등
- ✓ 데이터의 각 행은 레코드에 해당하며 각 열은 설명변수에 해당
- ✓ 분류 문제에 비해 모델의 복잡도가 높음



# 목적에 따른 기계학습 방법론

3

## 연관규칙분석: Association Rule Mining

- 목적: 연관성이 높은 아이템들로 구성된 규칙 집합을 생성

- ✓ 예시: 아이폰을 구매하는 고객들을 범퍼를 함께 구매한다.
- ✓ 데이터의 각 행은 트랜잭션에 해당
- ✓ 추천 시스템에 주로 사용되며 “장바구니 분석”이라고도 불림

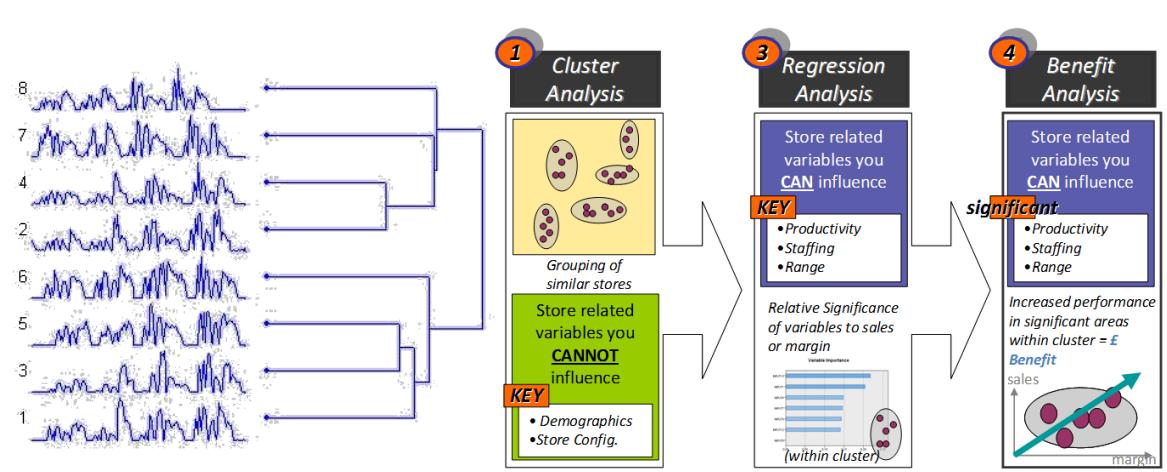
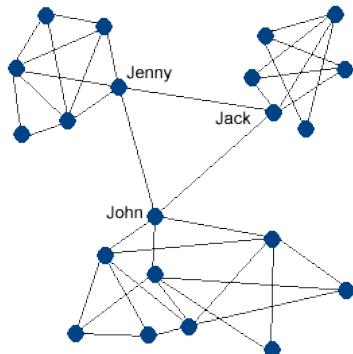


# 목적에 따른 기계학습 방법론

## 군집화: Clustering/Segmentation

- 목적: 전체 데이터를 보다 일관성과 응집성이 높은 세부 그룹으로 나눔
  - ✓ 동일그룹 내의 개체들은 유사할수록, 다른 그룹 내의 개체들은 서로 상이할수록 좋은 군집화 결과임
  - ✓ Market/Customer segmentation 등에 사용

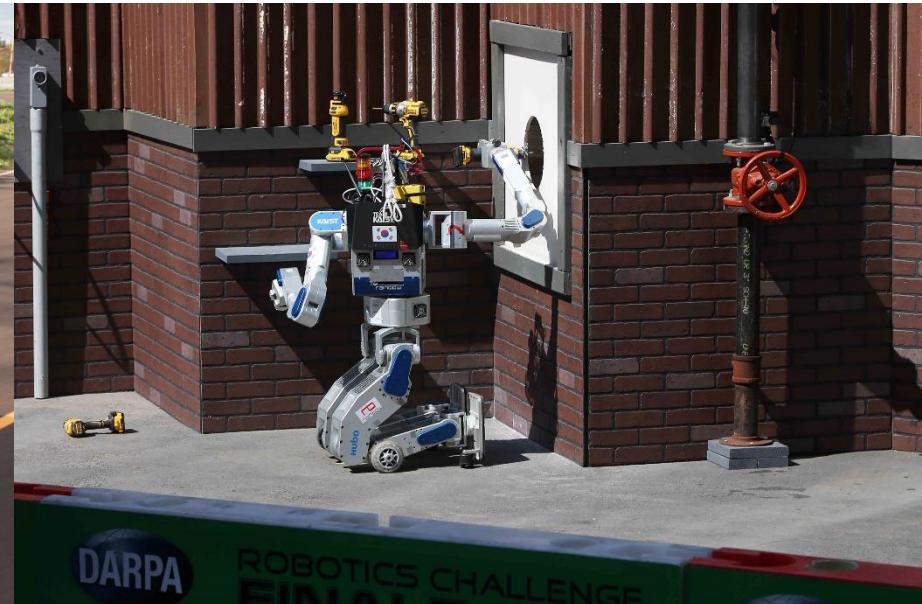
4



# 기계학습의 종류: 3<sup>rd</sup> Dimension

## ❖ 모델 업데이트와 학습의 최종 목적에 따른 구분

	Static Learning	Incremental (online) Learning	Reinforcement Learning
Objective Function	Short-term (snapshot)	Short-term (snapshot)	Long-term
Model update	Fully Updated	Partially Updated	Adaptively updated



# 목차

I

데이터 사이언스

II

데이터 사이언스 활용 분야

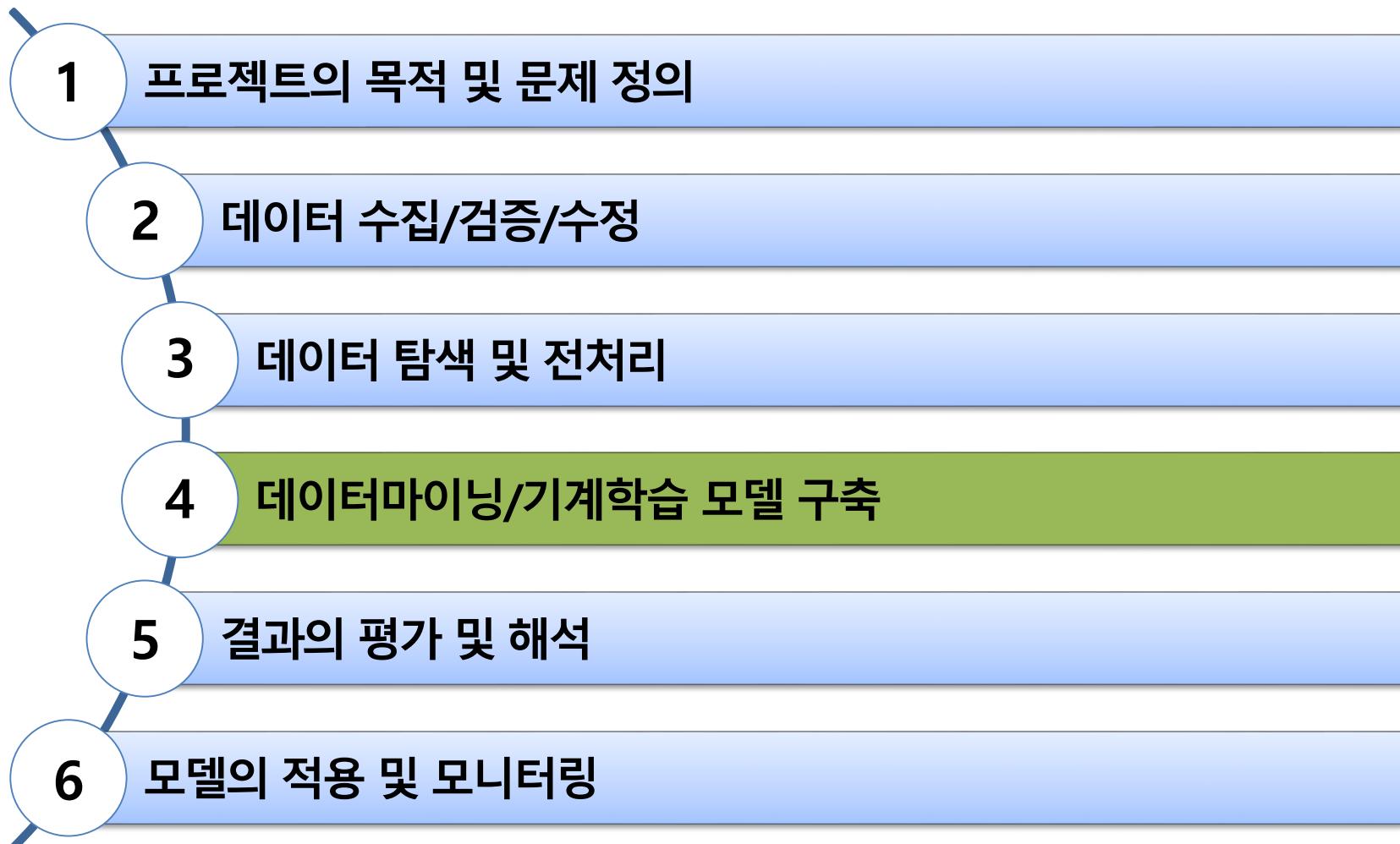
III

데이터 사이언스 속의 기계학습

IV

데이터 사이언스 절차

# 데이터 사이언스 절차



# 1단계: 프로젝트의 목적 및 문제 정의

1

## 프로젝트의 목적 및 문제 정의

- 왜 이 프로젝트를 진행하려고 하는가?

- ✓ 정성적 목표
- ✓ 고객 충성도의 증가
- ✓ 마케팅 효율성 증대

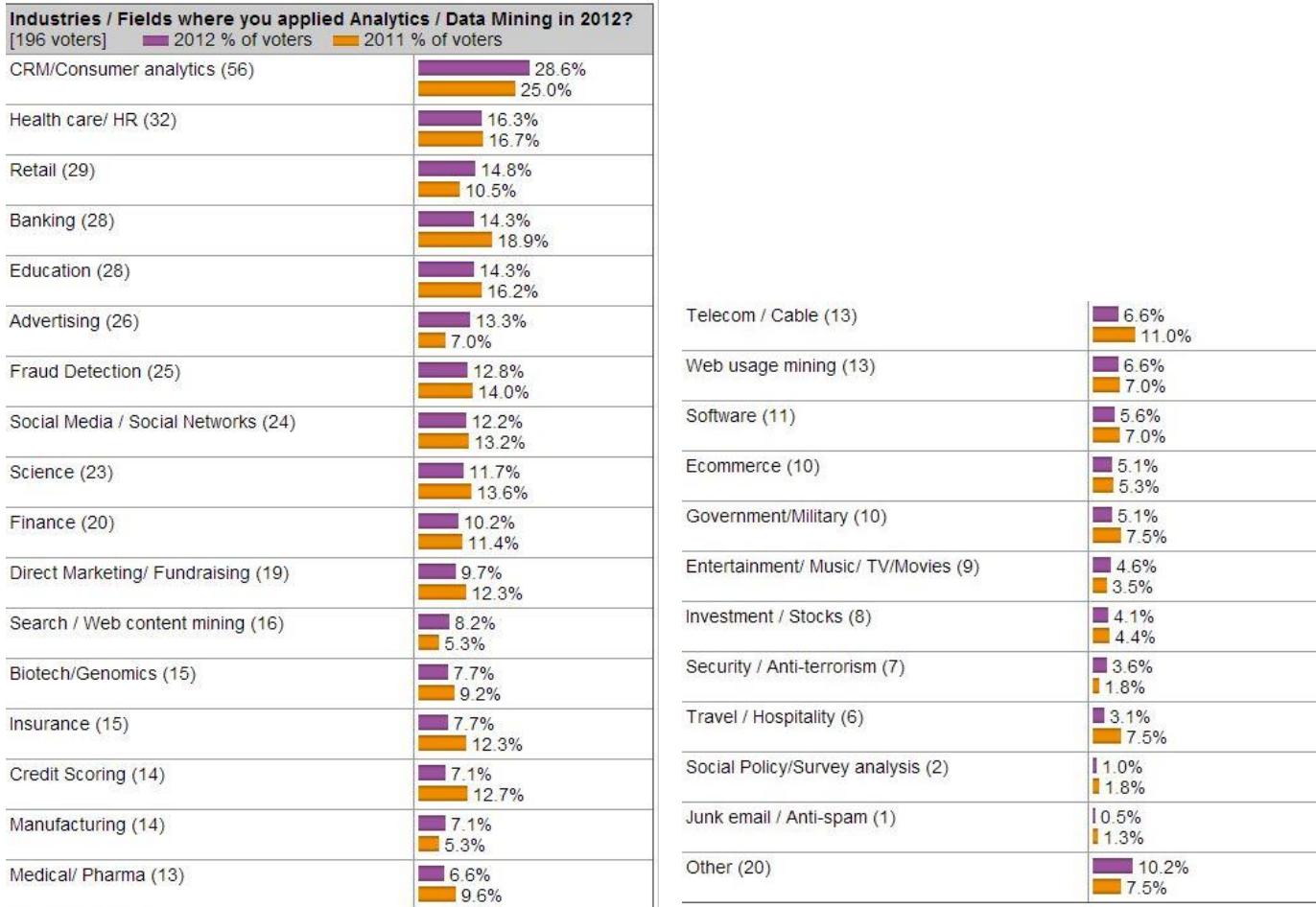
- 이 프로젝트가 성공하면 무엇을 기대할 수 있는가?

- ✓ 고객기대수명의 증가
- ✓ 매출 증대/비용 감소

# 1단계: 프로젝트의 목적 및 문제 정의

1

## 프로젝트의 목적 및 문제 정의



# 1단계: 프로젝트의 목적 및 문제 정의

1

## 프로젝트의 목적 및 문제 정의

- 데이터마이닝/기계학습으로 풀고자 하는 문제 정의

- ✓ 이 고객은 우리 제품을 구매하는데 한달에 얼마를 소비할 것인가?
- ✓ 현재 환자의 상태로 볼 때 암으로 판정될 가능성은 얼마나 되는가?

- 문제에 걸맞는 기계학습 알고리즘 선택

- ✓ 분류 방법론
- ✓ 예측 방법론
- ✓ 연관규칙분석 등

# 2단계: 데이터 수집/검증/수정

2

## 데이터 수집/검증/수정: 데이터 수집

### ▪ 데이터 원천

- ✓ 내부 데이터: 데이터 웨어하우스, 데이터 마트 등
- ✓ 외부 데이터: 구입 데이터, 공공 데이터 등

### ▪ 필요할 경우 독립변수와 종속변수 정의

- ✓ 예시: 신용카드 이탈고객 예측 모형
  - 독립변수: 나이, 성별, 사용 기간, 월평균 이용금액, 신용등급
  - 종속변수: 이탈 여부



# 2단계: 데이터 수집/검증/수정

2

## 데이터 수집/검증/수정: 데이터 검증

### ▪ 이상치 (Outlier)

- ✓ 값은 존재하되 실질적으로 가능하지 않은 값 (나이 990살, 키 30m)
- ✓ 다양한 이유로 인해 현실 DB에는 이상치가 존재함

### ▪ 결측치 (Missing value)

- ✓ 값이 있어야 할 곳에 값이 존재하지 않는 상황
- ✓ 운영 에러, 휴먼 에러 등

### ▪ 해결 방안

- ✓ 레코드의 수가 충분히 많다 → 제거
- ✓ 레코드의 수가 충분치 않다 → 합리적인 값으로 대치(imputation)

# 2단계: 데이터 수집/검증/수정

2

## 데이터 수집/검증/수정: 데이터 수정

### ▪ 정성적 변수

- ✓ 특정한 속성을 가진 자료 (성별: 남/여, 혈액형: A/B/O/AB 등)
- ✓ 일반적으로 사칙 연산 적용 불가능

#### 명목형 (Nominal)

- 자료값의 크기나 순서에 의미가 없음
- 각 속성에 대하여 편의상 숫자를 대응시켜 사용하기도 함
- 혈액형, 종교, 운동선수 등번호, 인종 등

#### 순서형 (Ordinal)

- 기준에 따라 자료값들의 순서에 의미를 부여
- 각 숫자는 순서의 의미만을 가지며, 차이/비율의 의미는 없음
- 에너지 효율 등급, 학점, 올림픽 메달 등

# 2단계: 데이터 수집/검증/수정

2

## 데이터 수집/검증/수정: 데이터 수정

### ▪ 정량적 변수

- ✓ 많고 적음을 나타내는 수치로 된 자료
- ✓ 사칙 연산 가능

#### 계수형/이산형 (Count/Discrete)

- 셀 수 있는 정수의 형태
- 형제 수, 보험 가입 건 수 등

#### 연속형 (Continuous)

- 셀 수 없는 소수점을 포함
- 키, 무게, 길이 등

#### 구간형 (Interval)

- 차이만 의미가 있음
- 온도: 20도는 10도보다 2배 뜨겁다(X)

#### 비율형 (Ratio)

- 차이와 비율이 모두 의미가 있음
- 20kg은 10kg보다 2배 무겁다 (O)

# 2단계: 데이터 수집/검증/수정

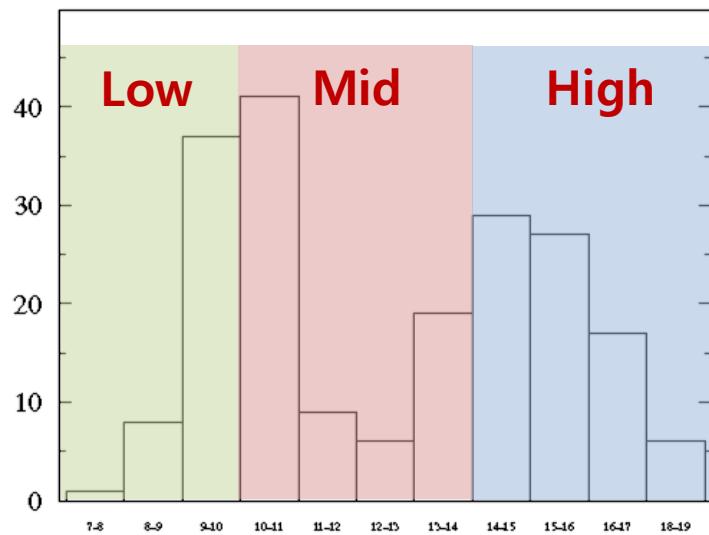
2

## 데이터 수집/검증/수정: 데이터 수정

### ▪ 변수 변환

#### Binning:

- 연속형 변수를 순서형 변수로



#### 1-of-C coding:

- 1개의 명목형 변수를 C개의 이진형 변수로

"Color: yellow, red, blue, green"

	d1	d2	d3	d4
yellow	1	0	0	0
red	0	1	0	0
blue	0	0	1	0
green	0	0	0	1

# 3단계: 데이터 탐색 및 전처리

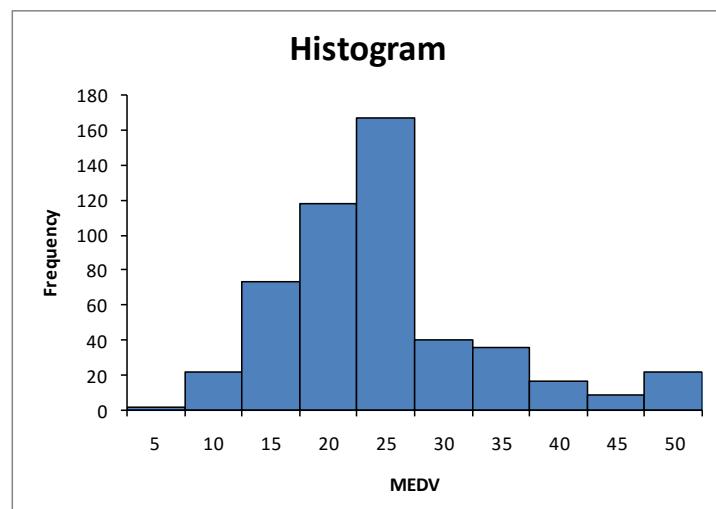
## 데이터 탐색 및 전처리: 데이터 탐색

### ▪ 단변량 분석

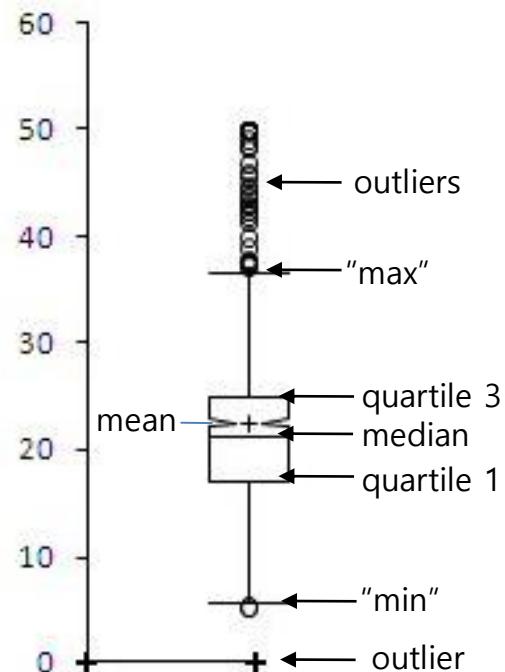
3

#### 히스토그램: Histogram

- 한 변수의 값이 갖는 분포를 파악
- 정규분포 가정 등에 이용



#### 상자그림: Box plot



# 3단계: 데이터 탐색 및 전처리

## 데이터 탐색 및 전처리: 데이터 탐색

- **다변량 분석**

- ✓ 상관계수:

3

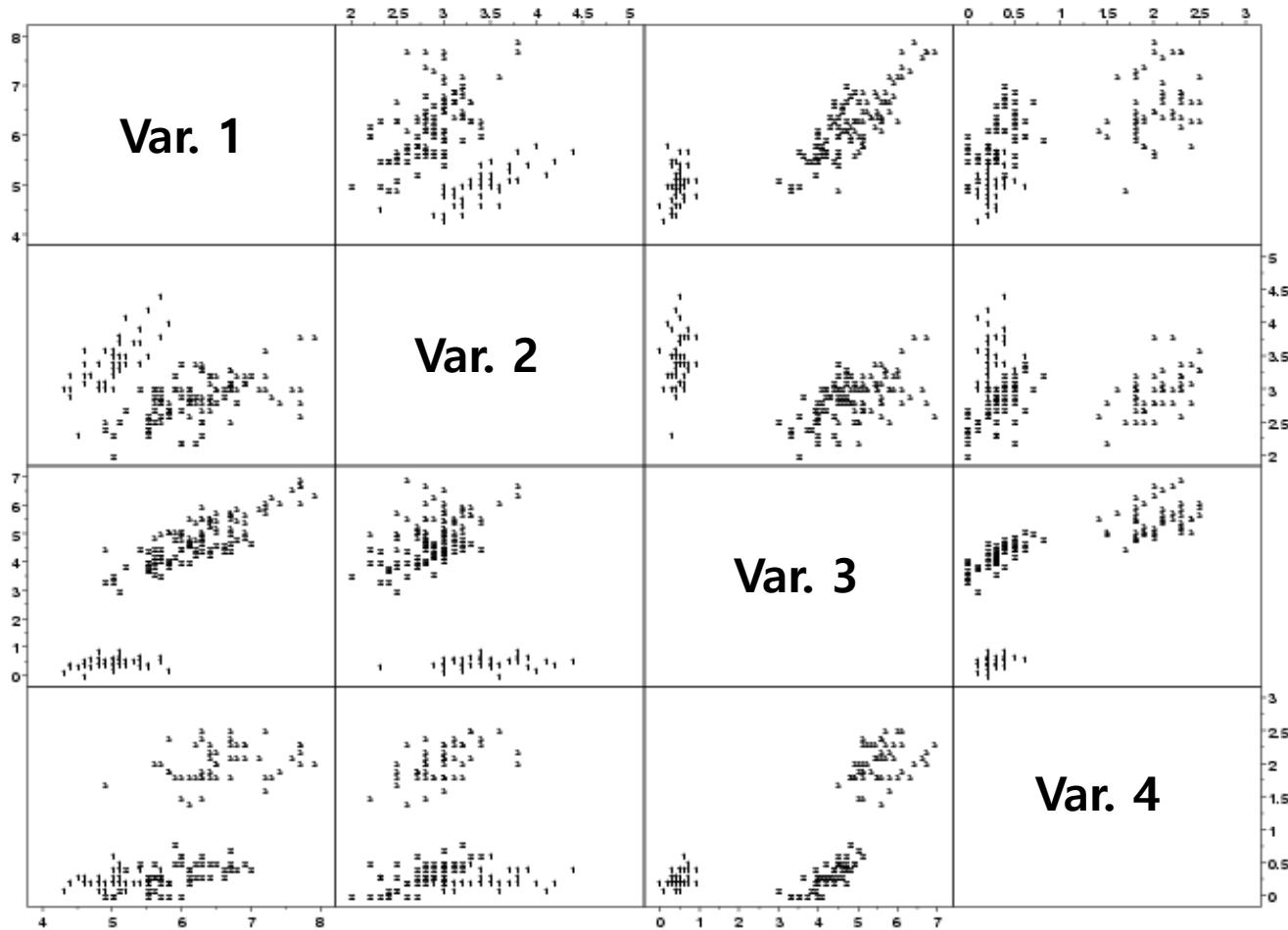
- 어떤 변수들이 서로 상관관계가 높은가?
  - 대표성을 갖는 적은 수의 변수를 선택할 때 사용 가능

	<i>CRIM</i>	<i>ZN</i>	<i>INDUS</i>	<i>CHAS</i>	<i>NOX</i>	<i>RM</i>
<i>CRIM</i>	1					
<i>ZN</i>	-0.20047	1				
<i>INDUS</i>	0.406583	-0.53383	1			
<i>CHAS</i>	-0.05589	-0.0427	0.062938	1		
<i>NOX</i>	0.420972	-0.5166	0.763651	0.091203	1	
<i>RM</i>	-0.21925	0.311991	-0.39168	0.091251	-0.30219	1

# 3단계: 데이터 탐색 및 전처리

## 데이터 탐색 및 전처리: 데이터 탐색

3



# 3단계: 데이터 탐색 및 전처리

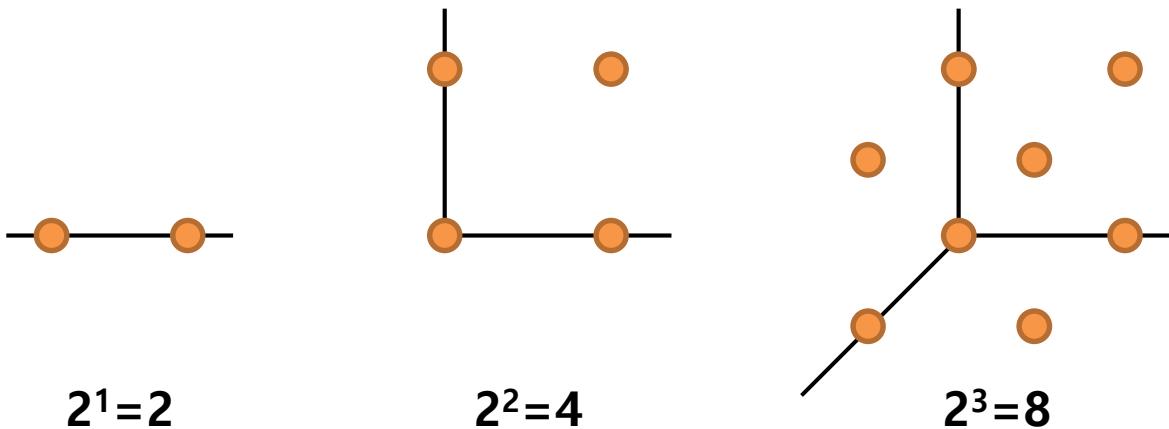
## 데이터 탐색 및 전처리: 데이터 전처리

*“If there are various logical ways to explain a certain phenomenon, the simplest is the best” - Occam’s Razor*

3

### ▪ 차원의 저주: Curse of dimensionality

- ✓ 변수가 증가할수록 동일한 설명력을 유지하기 위해 필요한 레코드의 수는 기하급수적으로 증가함



# 3단계: 데이터 탐색 및 전처리

3

## 데이터 탐색 및 전처리: 데이터 전처리

### ▪ 변수 선택: Variable Selection

- ✓ 전체 변수 중에서 유의미한 변수만을 선택
- ✓ 원래 변수의 형태가 그대로 보존됨

### ▪ 변수 추출: Variable Extraction

- ✓ 전체 데이터 집합을 잘 설명할 수 있는 적은 수의 변수를 생성
- ✓ 원래 변수의 형태가 그대로 보존되지 않음
- ✓ Original variables: Age, sex, height, weight
- ✓ Constructed variables:  $\text{Age} + 3 * \text{I}(\text{sex} = \text{female}) + 0.2 * \text{height} - 0.3 * \text{weight}$

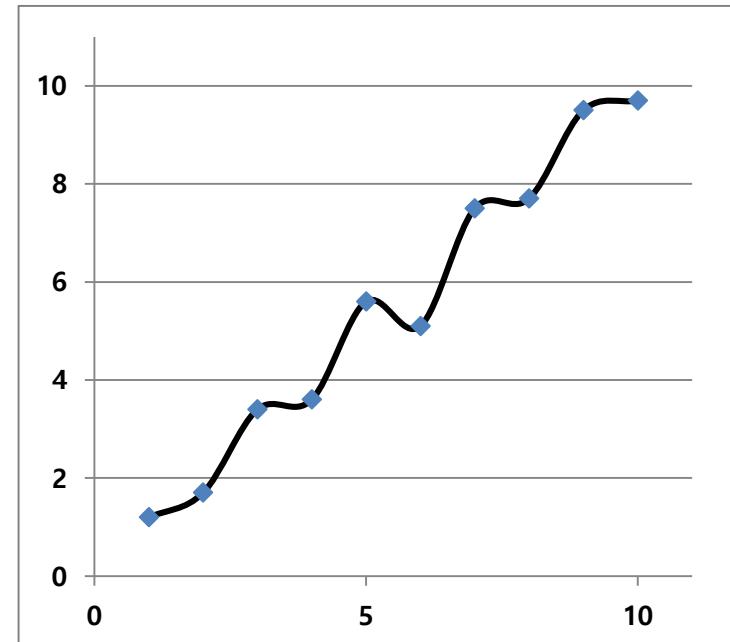
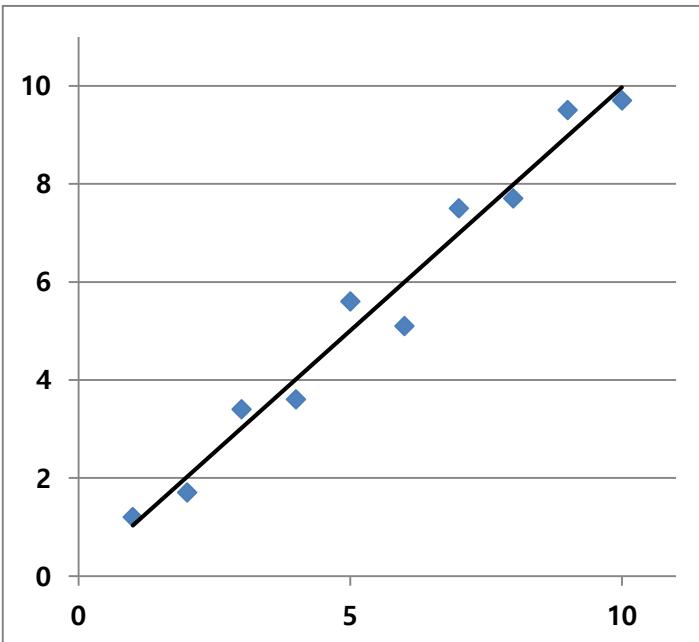
# 3단계: 데이터 탐색 및 전처리

## 데이터 탐색 및 전처리: 데이터 전처리 (데이터 분할)

- 과적합 (Over-fitting)

- ✓ 데이터마이닝 알고리즘이 데이터의 필요 없는 특징(noise)까지 외워버리는 상태

3



# 3단계: 데이터 탐색 및 전처리

3

## 데이터 탐색 및 전처리: 데이터 전처리 (데이터 분할)

- 학습 데이터: **Training Data**

- ✓ 데이터마이닝 모델을 구축하는데 사용

- 검증 데이터: **Validation Data**

- ✓ 모델의 최적 파라미터를 선택하는데 사용

- 테스트 데이터: **Test Data**

- ✓ 새로운 데이터에 적용하여 모델의 실제 예측력을 평가하는데 사용

**Training Data**

Algorithm A-1  
Algorithm A-2  
Algorithm A-3  
Algorithm B-1  
Algorithm B-2  
Algorithm B-3

**Validation Data**

Algorithm A-1  
Algorithm A-2  
Algorithm A-3  
Algorithm B-1  
Algorithm B-2  
Algorithm B-3

**Test Data**

Algorithm A-1  
Algorithm A-2  
Algorithm A-3  
Algorithm B-1  
Algorithm B-2  
Algorithm B-3

# 3단계: 데이터 탐색 및 전처리

3

## 데이터 탐색 및 전처리: 데이터 전처리 (데이터 정규화)

- 정규화: **Normalization**

- ✓ 각 변수들의 측정 단위가 다름으로 인해 나타날 수 있는 효과를 제거
- ✓ z-score:  $(\text{value}-\text{mean})/(\text{standard deviation})$ .

**Original data**

Id	Age	Income
1	25	1,000,000
2	35	2,000,000
3	45	3,000,000
...	...	...
Mean	35	2,000,000
Stdev	5	1,000,000

**Normalized data**

Id	Age	Income
1	-2	-1
2	0	0
3	2	1
...	...	...
Mean	0	0
Stdev	1	1

# 4단계: 데이터 마이닝/기계학습 모델 구축

4

## 데이터마이닝/기계학습 모델 구축

- 데이터마이닝/기계학습 모델

- ✓ 분류

- Logistic regression, k-nearest neighbor, naïve bayes, classification trees, neural networks, linear discriminant analysis.

- ✓ 예측

- Linear regression, k-nearest neighbor, regression trees, neural networks.

- ✓ 연관규칙 분석: A priori algorithm.

- ✓ 군집화: Hierarchical clustering, K-Means clustering.

# 5단계: 결과의 평가 및 해석

## 결과의 평가 및 해석

- 분류 성능 평가 지표

- ✓ 혼동 행렬(confusion matrix)

		Predicted	
		1(+)	0(-)
Actual	1(+)	True positive, Sensitivity (A)	False negative, Type I error (B)
	0(-)	False positive, Type II error (C)	True negative, Specificity (D)

5

- ✓ Simple accuracy:  $(A+C)/(A+B+C+D)$
- ✓ Balanced correction rate:  $\sqrt{\frac{A}{A+B} \cdot \frac{D}{C+D}}$
- ✓ Lift charts, receiver operating characteristic (ROC) curve, etc.

# 5단계: 결과의 평가 및 해석

## 결과의 평가 및 해석

- 회귀 성능 평가 지표

  - ✓  $y$ : actual target value,  $y'$ : predicted target value

  - Mean squared error, Root mean squared error

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2$$

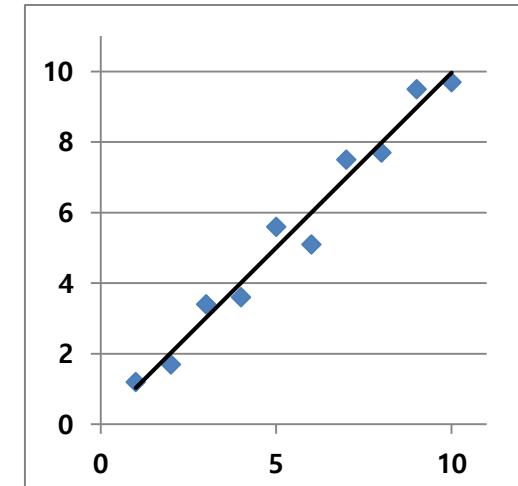
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2}$$

  - Mean absolute error

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i|$$

  - Mean absolute percentage error

$$MAPE = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i| / |y_i|$$



# 5단계: 결과의 평가 및 해석

## 결과의 평가 및 해석

### ▪ 군집화 성능 평가 지표

- ✓ 군집 내 분산: 동일 군집 내 레코드들은 얼마나 가까이 위치하는가
- ✓ 군집 간 분산: 다른 군집에 속하는 레코드들은 얼마나 멀리 떨어져 있는가
- ✓ 좋은 군집화 결과: 군집 내 분산은 작고 군집 간 분산은 큼

### ▪ 연관규칙분석 성능 평가 지표

5

- ✓ Support:  $P(A, B)$
- ✓ Confidence:  $P(A | B) = \frac{P(A, B)}{P(B)}$
- ✓ Lift  $\frac{P(A | B)}{P(B)} = \frac{P(A, B)}{P(A) \cdot P(B)}$

# 6단계: 적용 및 모니터링

## 적용 및 모니터링

### ▪ 모델 적용

- ✓ 구축된 기계학습 모델을 운영시스템에 이식 및 실행
- ✓ 기계학습 모델의 예측 결과를 바탕으로 실제 action 수행
  - "강필성 고객이 다음달에 우리회사 서비스를 이용하지 않을 확률이 80%이니 이탈하지 않도록 사용 쿠폰을 발송하시오."

### ▪ 모니터링

- ✓ 기계학습 예측 결과를 바탕으로 수행한 캠페인의 성과 분석
- ✓ 필요할 경우 모델의 업데이트 수행