Prologue

통계학이 아닌 통계

통계의 필요성

통계학보다 통계

- 영어 문법을 10년 배워도 영어로 말 한마디 꺼내기 힘듦

- 통계학을 배우는 것과 통계를 잘 쓰는 것은 차이가 있음

- 통계에 대한 일반적인 인식 :
  - “데이터의 요약”, “데이터 요약을 통해 만든 정보” 등 분석 결과를 가리킴

- 진짜 통계
  - 데이터로 표현된 세상을 바라보는 관점
  - 목표 : 데이터 속에 있는 차이를 확인하고 설명하는 것

통계는 시간으로 소통하는 언어

미래 = 불확실성

- 결정된 것은 없고 가능성만이 존재함
- 가능성을 수치화한 확률로 불확실성을 일부 통제

통계는 과거와 현재, 미래가 소통하는 언어

- 통계는 단순히 숫자를 계산하는 방법이 아님

- 과거의 정보를 담은 데이터를
- 현재의 관점으로 해석해서
- 미래의 불확실성을 다루는 소통의 언어

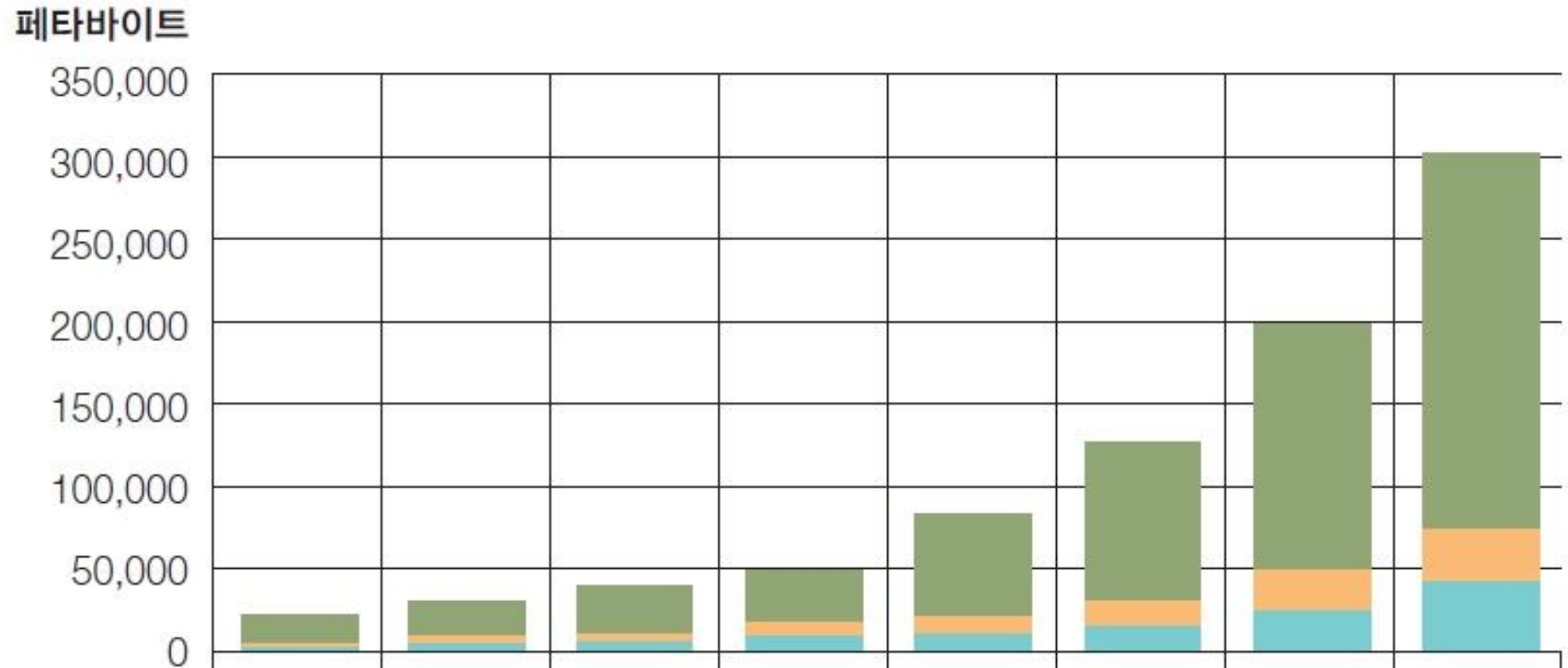
왜 갑자기 통계?

정보통신기술 발전에 따른 데이터의 변화 방향

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 (눈, 멀티미디어)	사물 정보, 인지 정보 (RFID, 센서, 사물 정보)
데이터 특성	구조화	다양성, 복합성, 소셜	현실성, 실시간성

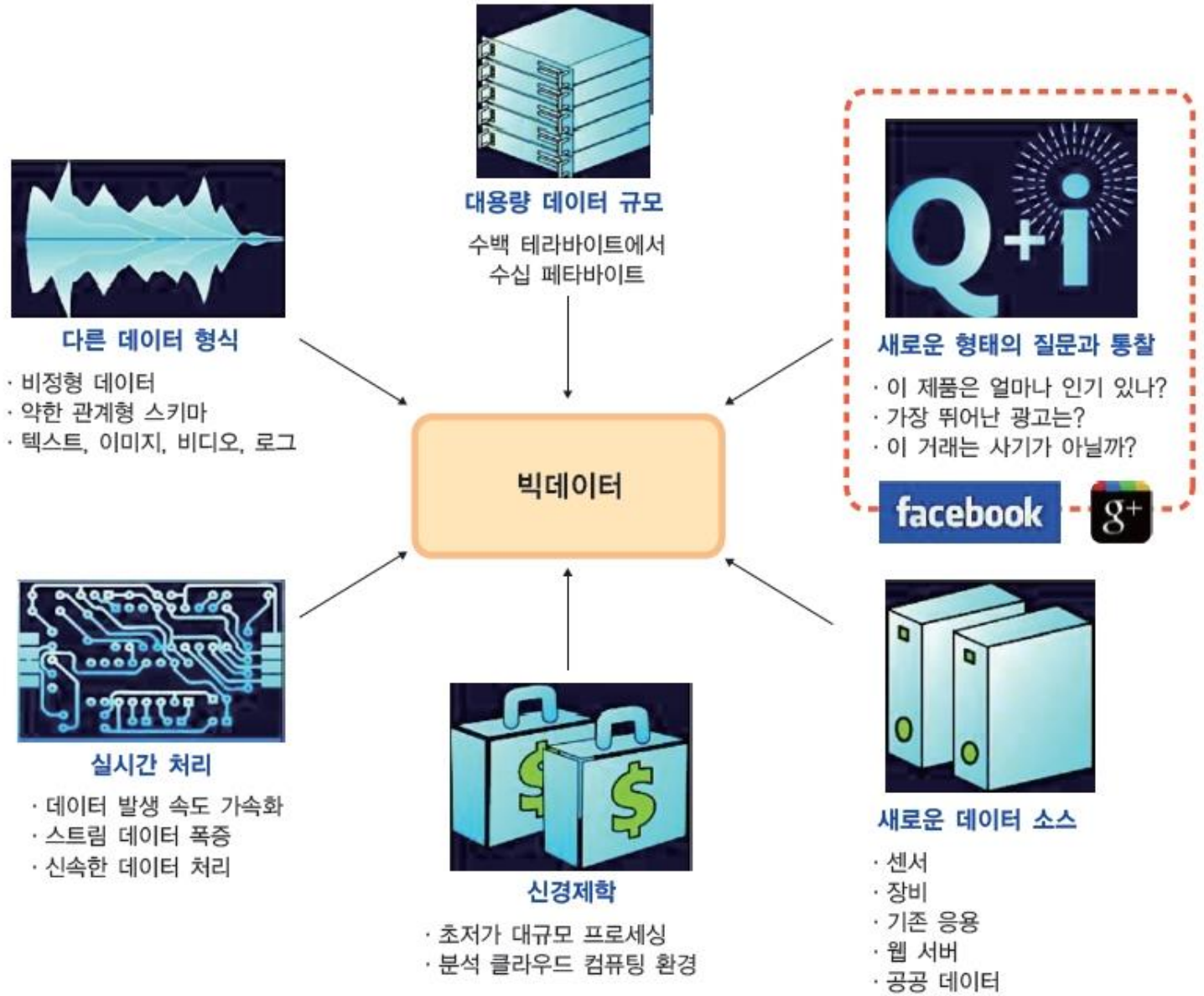
빅데이터?

정형과 비정형 데이터 유형의 변화



	2008	2009	2010	2011	2012	2013	2014	2015
비정형	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
정형(데이터베이스)	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
정형(이메일)	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

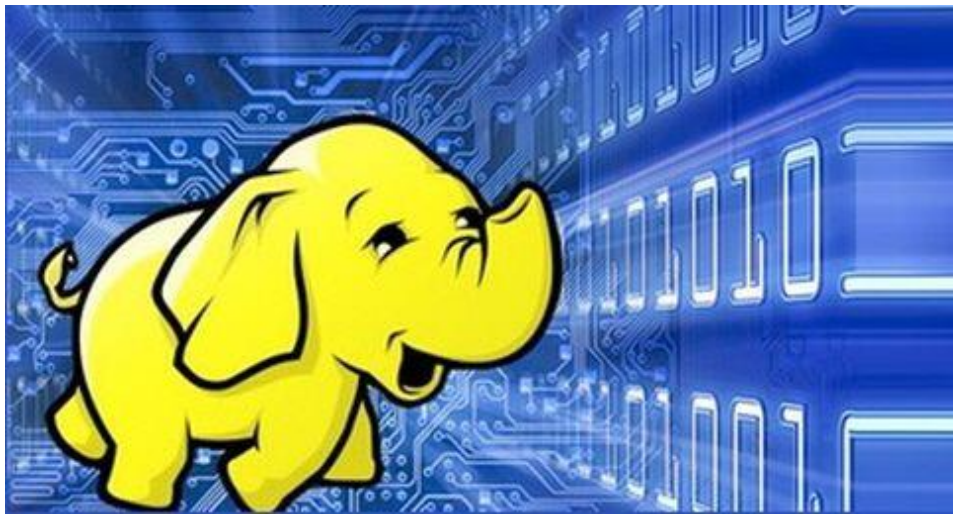
빅데이터 처리 특징

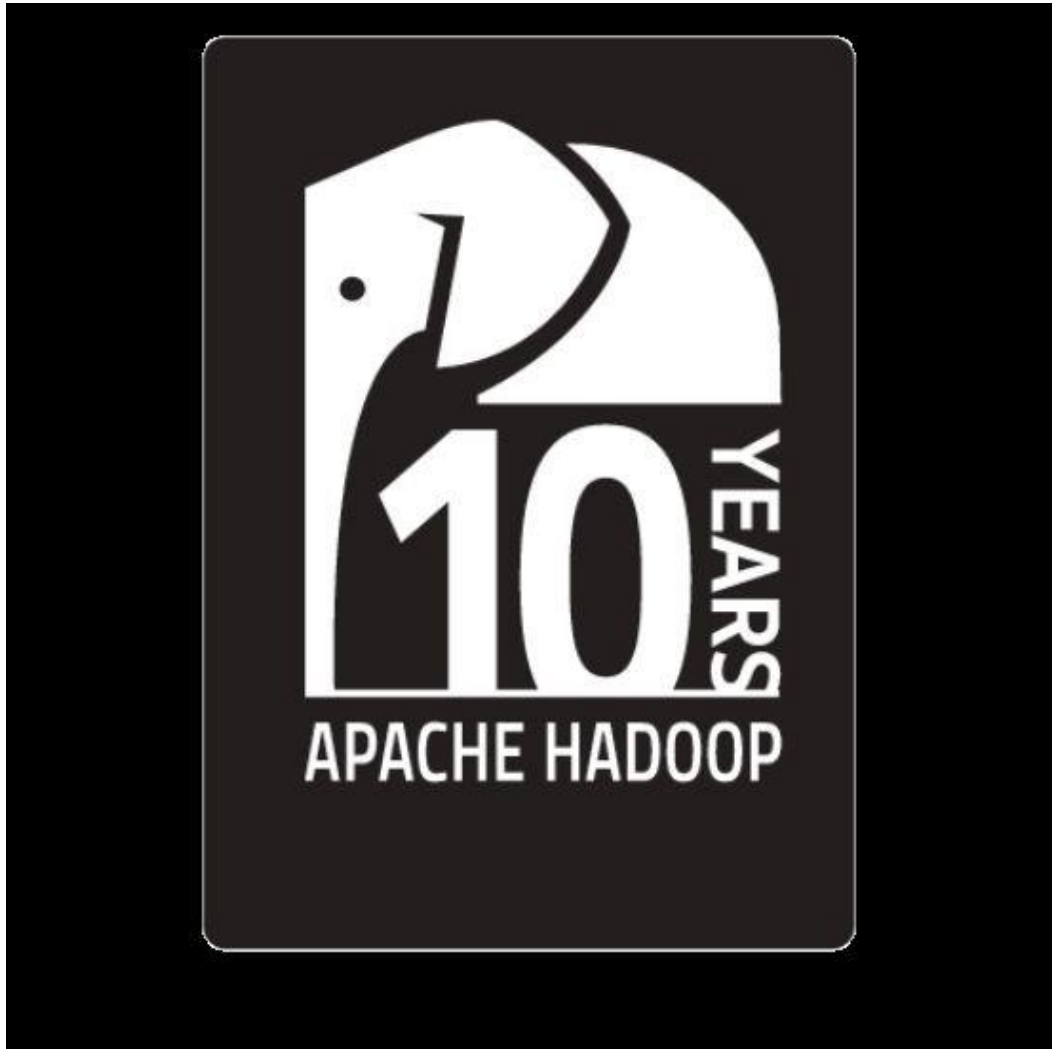

대용량 데이터 규모
수백 테라바이트에서
수십 페타바이트
다른 데이터 형식
· 비정형 데이터
· 약한 관계형 스키마
· 텍스트, 이미지, 비디오, 로그
새로운 형태의 질문과 통찰
· 이 제품은 얼마나 인기 있나?
· 가장 뛰어난 광고는?
· 이 거래는 사기가 아닐까?
facebook
빅데이터
실시간 처리
· 데이터 발생 속도 가속화
· 스트림 데이터 폭증
· 신속한 데이터 처리
신경제학
· 초저가 대규모 프로세싱
· 분석 클라우드 컴퓨팅 환경
새로운 데이터 소스
· 센서
· 장비
· 기존 응용
· 웹 서버
· 공공 데이터

빅데이터 처리 과정

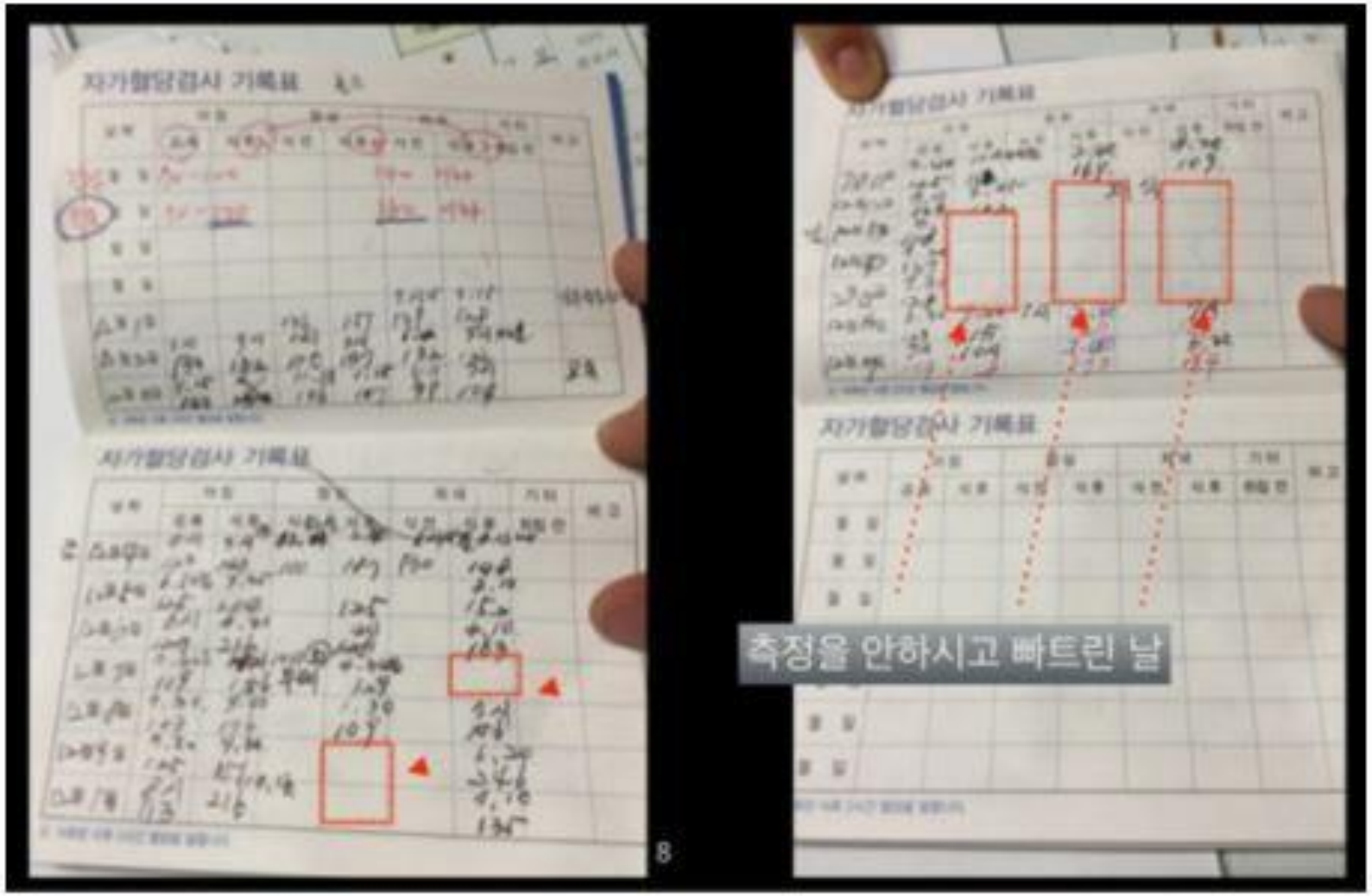

데이터 소스
수집
저장
처리
분석
표현
내부 데이터
외부 데이터
미디어
수동
자동
로그 수집기
크롤링
센싱
정형·비정형
데이터 등급
장치
일괄 처리
실시간 & 배치
NoSQL
관계형DB
분산 병렬 처리
전처리
분석 방법
Machine Learning

빅데이터와 하둡, 그리고 10년

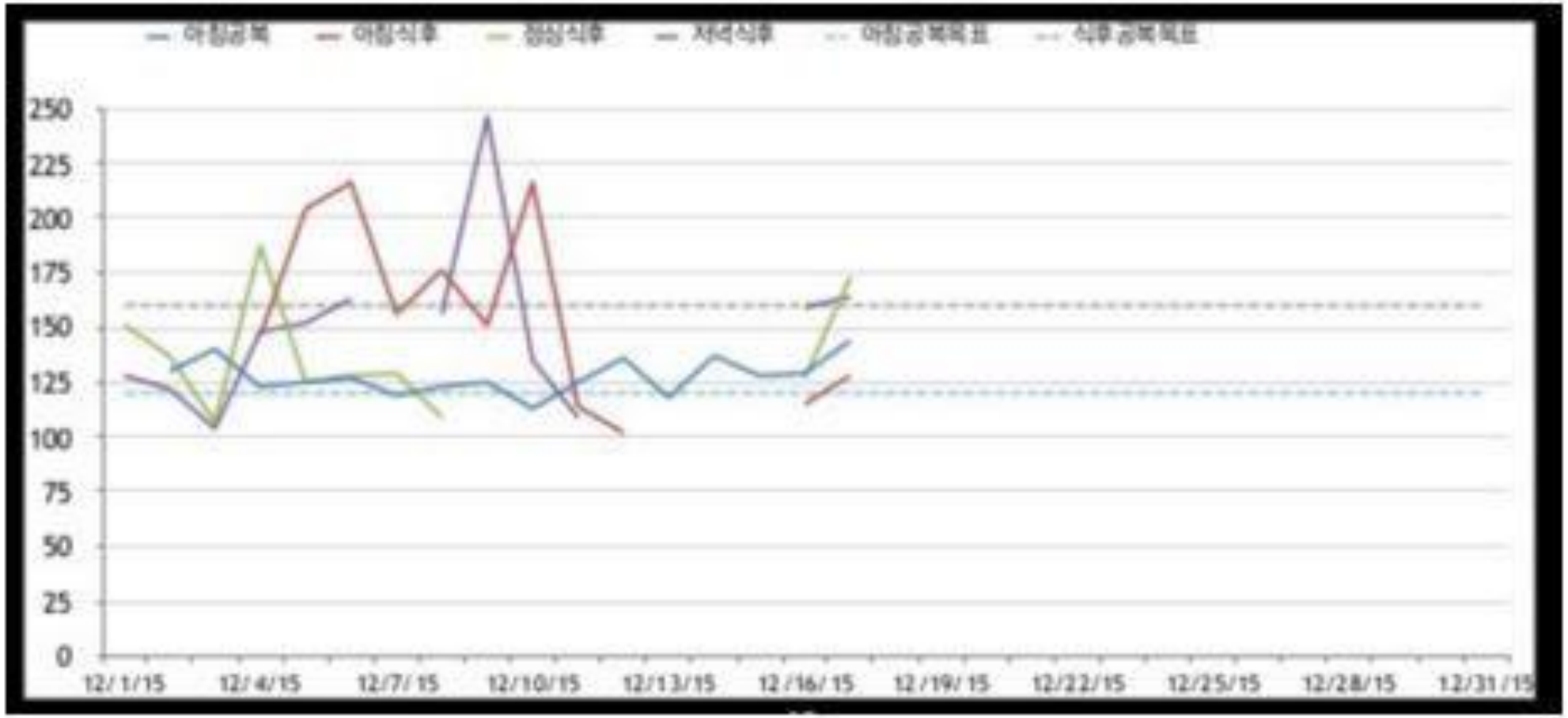




우리의 일상과 데이터 분석 (1)



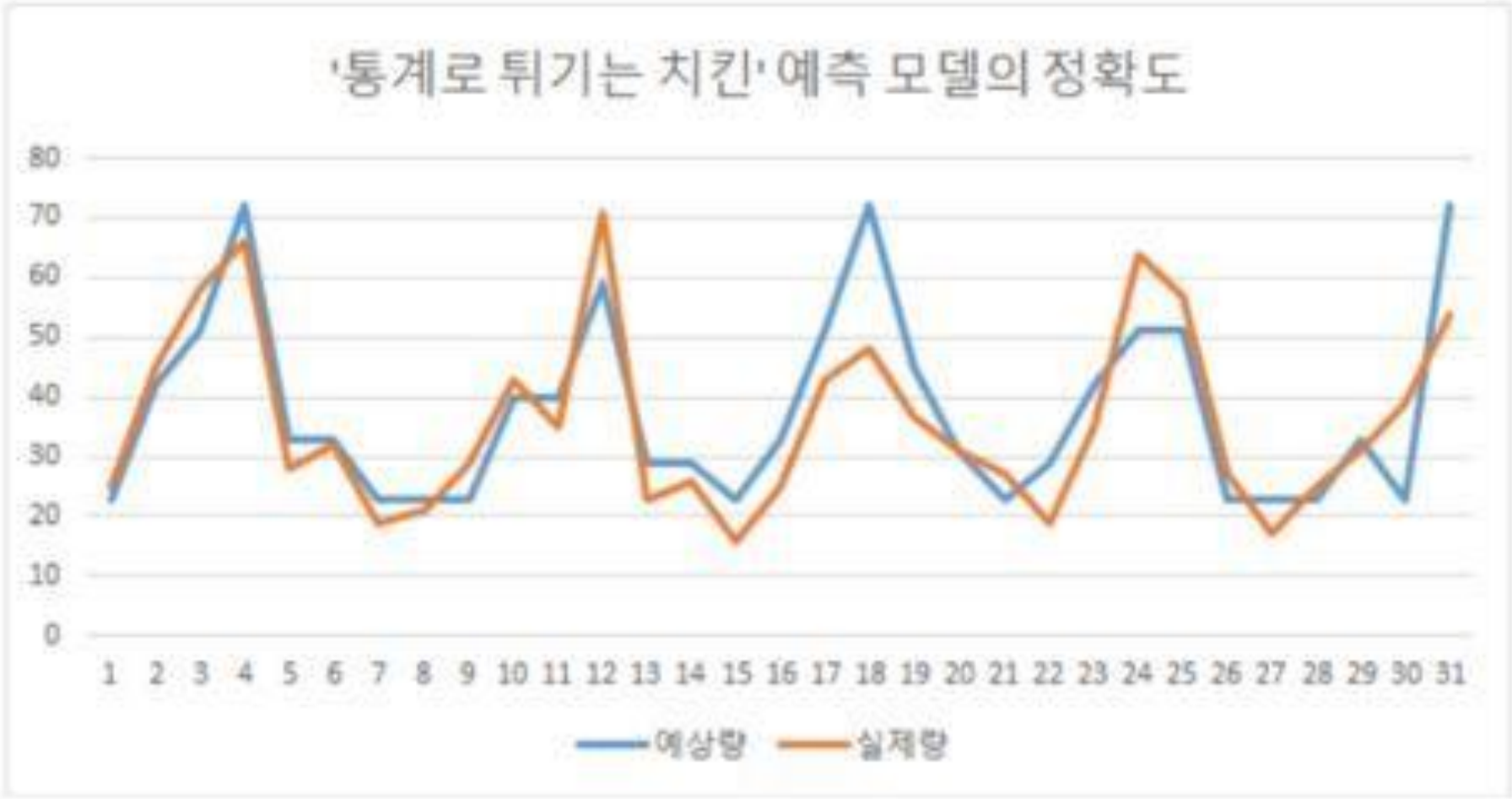
우리의 일상과 데이터 분석 (2)



우리의 일상과 데이터 분석 (3)

가중치	1.8	1.42	1.73	2.5	1.34	1.26	예상량	실제량
8명	계절	날씨	어벤트1	어벤트2	어벤트3	어벤트4		
1	여름						23	25
2	여름	비				야구	42	45
3	여름		주말			야구	51	58
4	여름	비	주말			야구	72	66
5	여름	비					33	28
6	여름	비					33	32
7	여름						23	19
8	여름						23	21
9	여름						23	29
10	여름		주말				40	43

우리의 일상과 데이터 분석 (4)



19대 대통령 선거와 빅데이터 분석 (1)

[후보자별 어휘 복잡성 분석]

후보자	어휘 복잡도	*어휘 복잡도 설명
문재인	7.09	약 중학교 1학년 수준의 어휘
안철수	7.88	약 중학교 2학년 수준의 어휘
홍준표	6.78	약 초등학교 6학년 수준의 어휘
유승민	9.29	약 중3 ~고1 수준의 어휘
심상정	9.45	약 중3~고1 수준의 어휘
평균	8.10	

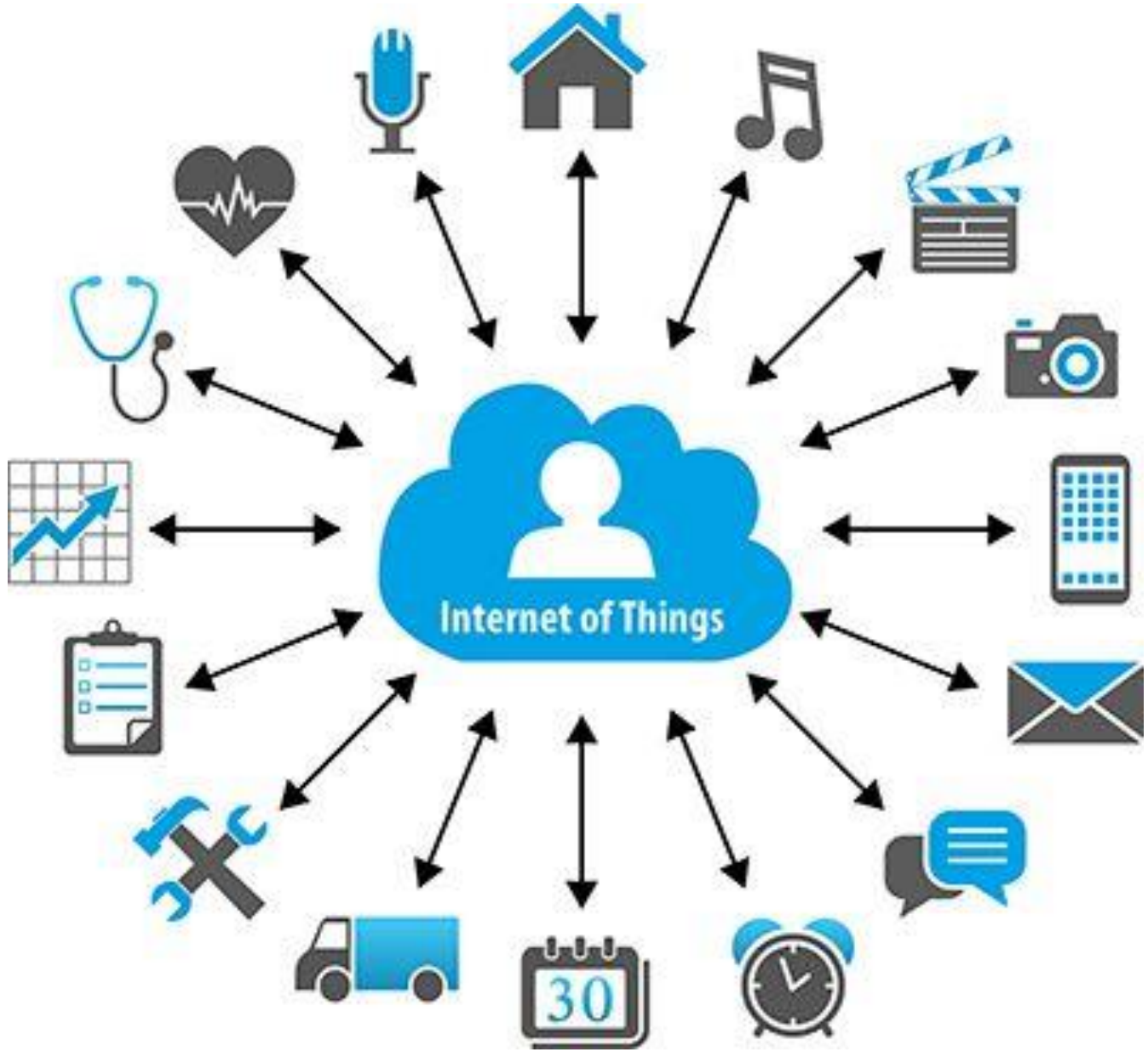
2017-04-13 SBS TV 토론회 텍스트 마이닝 결과

19대 대통령 선거와 빅데이터 분석 (2)

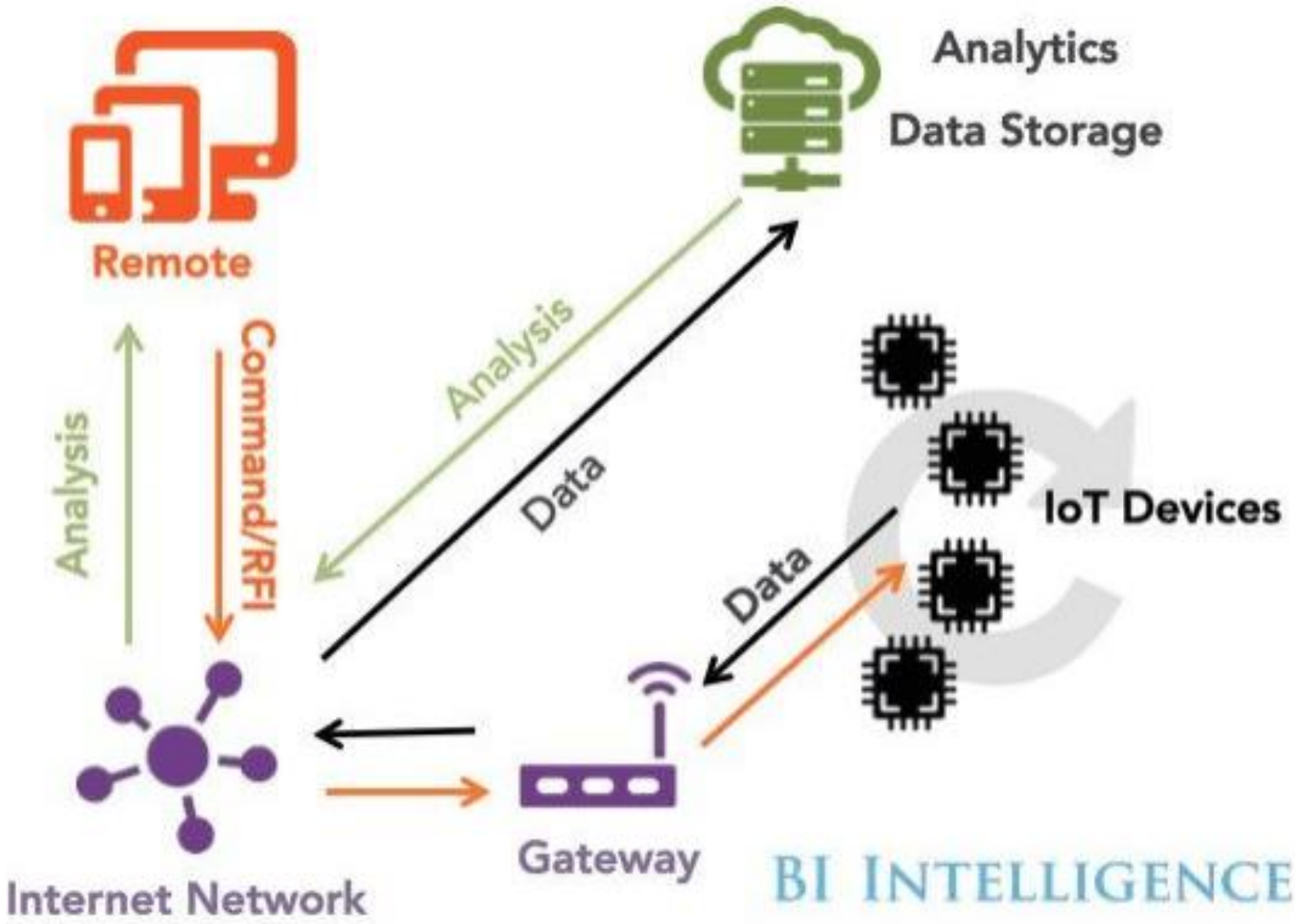
[후보자별 사용 단어 분석]

후보자	단어수	문장수	단어 상위 랭킹
문재인	566	130	탕감, 다음정부, 기본적, 차떼기, 동반성장, 국민성장
안철수	536	92	시작, 유능, 시도, 국가교육위원회, 적폐세력, 결국
홍준표	456	87	채용, 세월호, 해외, 고법, 문재인 후보, 강남좌파
유승민	540	79	보수, 위안부합의, 세금, 안보위기, 계속 반대, 보수표
심상정	556	71	정경유착, 실현, 이재용, 평등, 안보이용(정치적 이용)
평균	531	92	

2017-04-13 SBS TV 토론회 텍스트 마이닝 결과

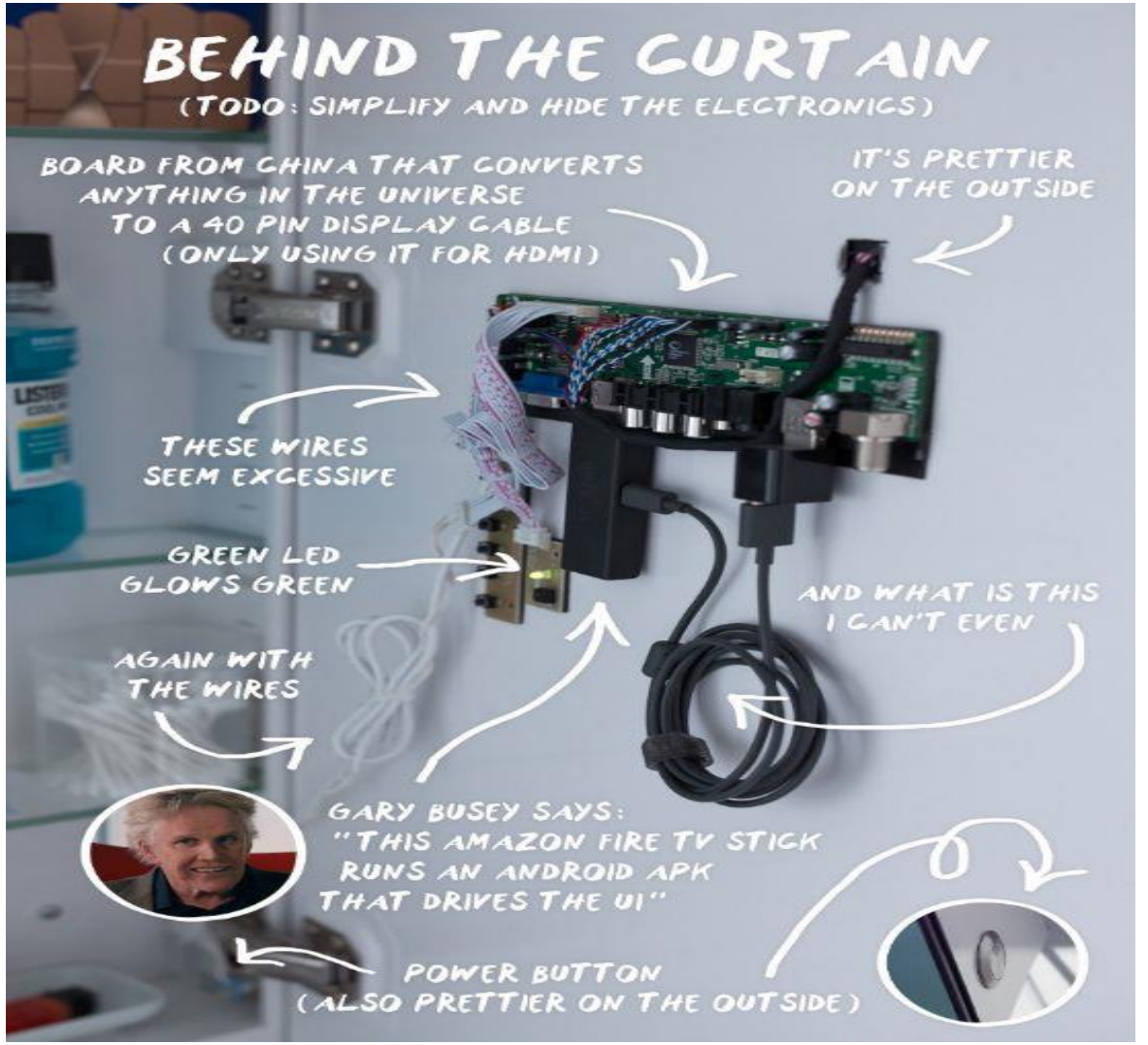

Internet of Things
30

IoT 생태계


Analytics
Data Storage
Remote
Analysis
Command/RFI
Analysis
Data
Data
IoT Devices
Internet Network
Gateway
BI INTELLIGENCE

스마트 거울

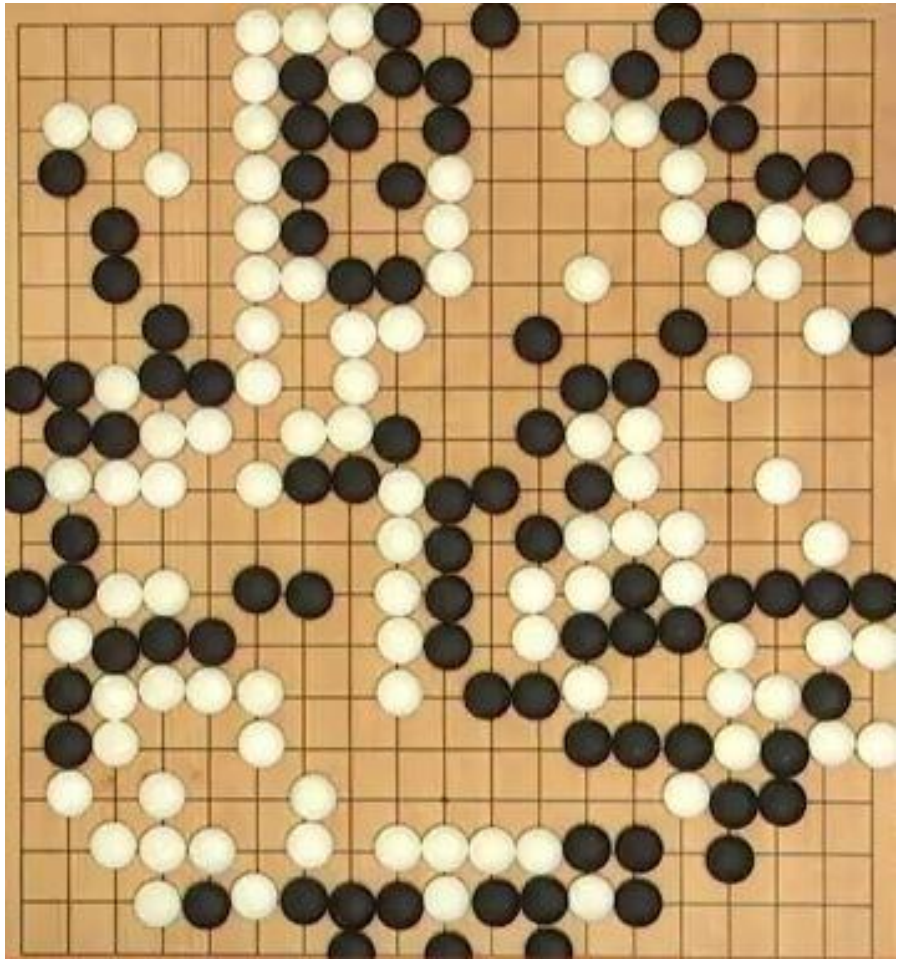

59°
3:20


BEHIND THE CURTAIN
(TODO: SIMPLIFY AND HIDE THE ELECTRONICS)
BOARD FROM CHINA THAT CONVERTS ANYTHING IN THE UNIVERSE TO A 40 PIN DISPLAY CABLE (ONLY USING IT FOR HDMI)
IT'S PRETTIER ON THE OUTSIDE
THESE WIRES SEEM EXCESSIVE
GREEN LED GLOWS GREEN
AND WHAT IS THIS I CAN'T EVEN
AGAIN WITH THE WIRES
GARY BUSEY SAYS: "THIS AMAZON FIRE TV STICK RUNS AN ANDROID APK THAT DRIVES THE UI"
POWER BUTTON (ALSO PRETTIER ON THE OUTSIDE)

인공지능은 우리 곁에 얼마나 가까이 다가왔는가? (1)

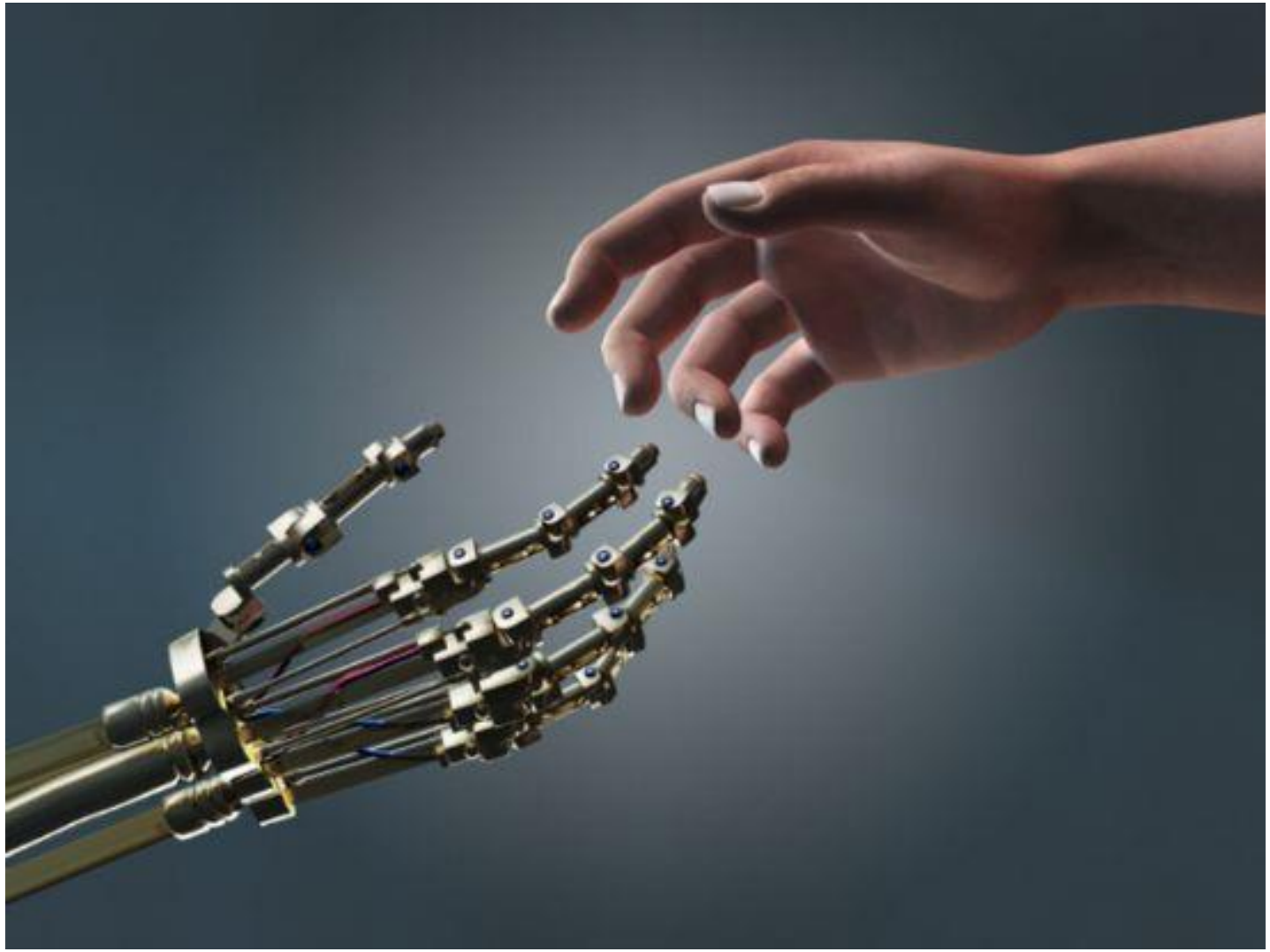
IBM의 Watson프로그램의 Jeopardy 경기, 왼쪽 Ken Jennings, 오른쪽 Brad Rutter

인공지능은 우리 곁에 얼마나 가까이 다가왔는가? (2)

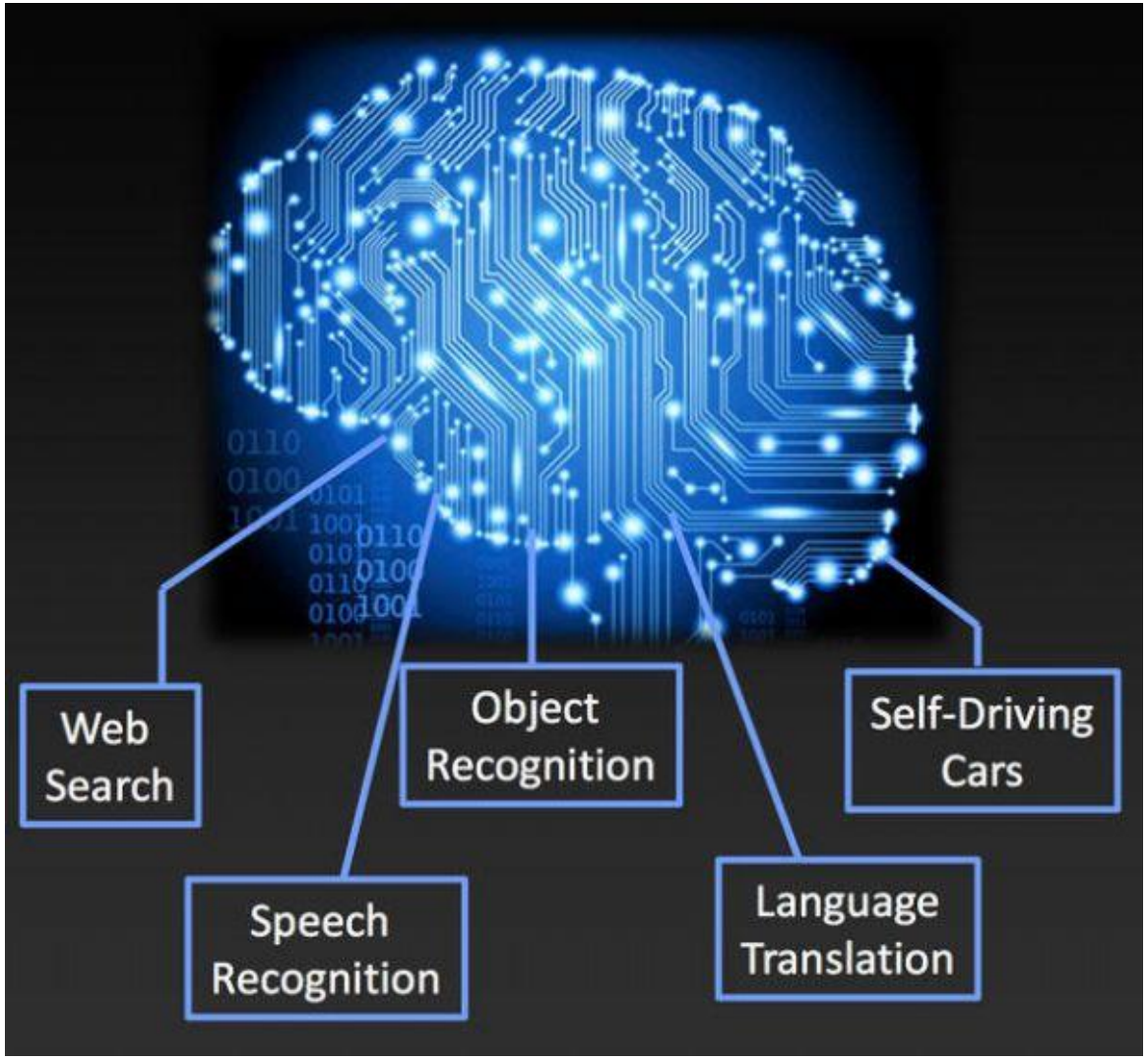


인공지능은 우리 곁에 얼마나 가까이 다가왔는가? (3)

머신러닝인가? 미신러닝인가?

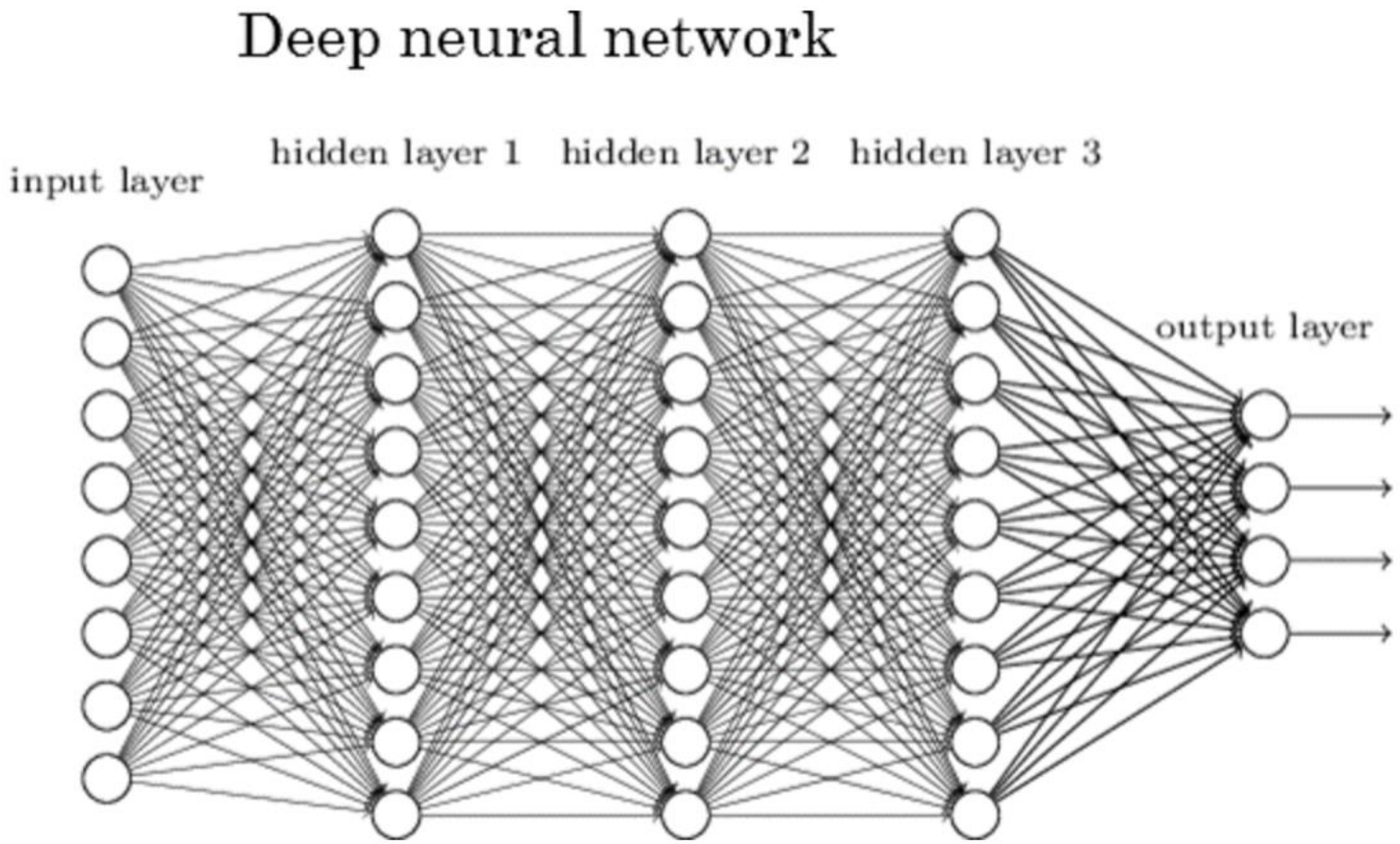


머신러닝 신기루...

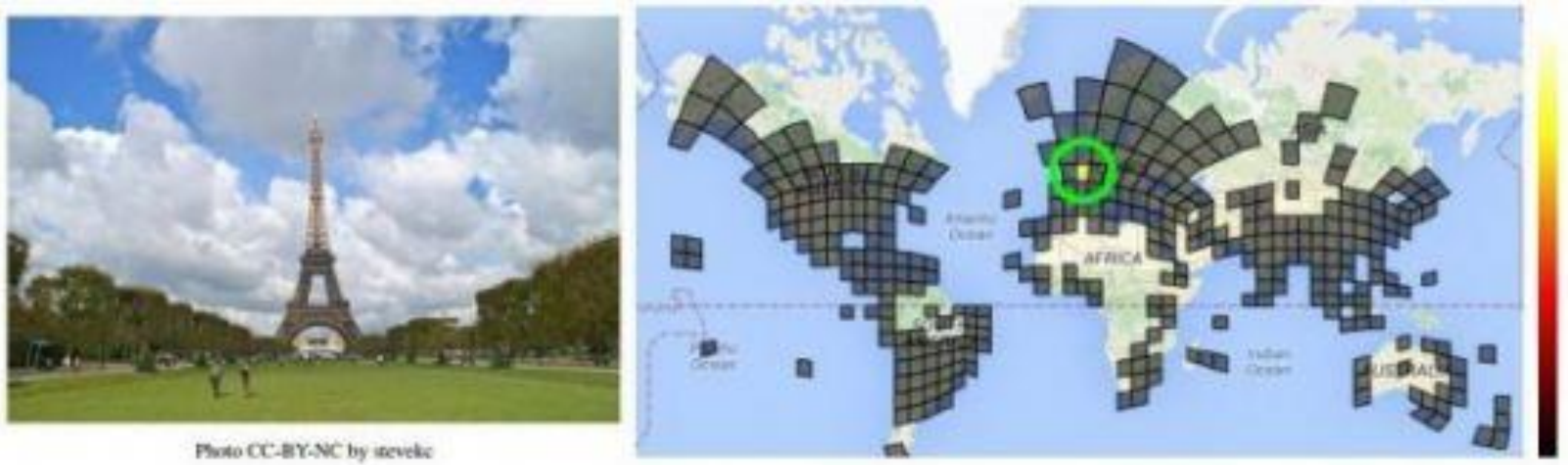

Web Search
Object Recognition
Self-Driving Cars
Speech Recognition
Language Translation

인간을 흉내 낸 인공신경망과 딥러닝 (1)

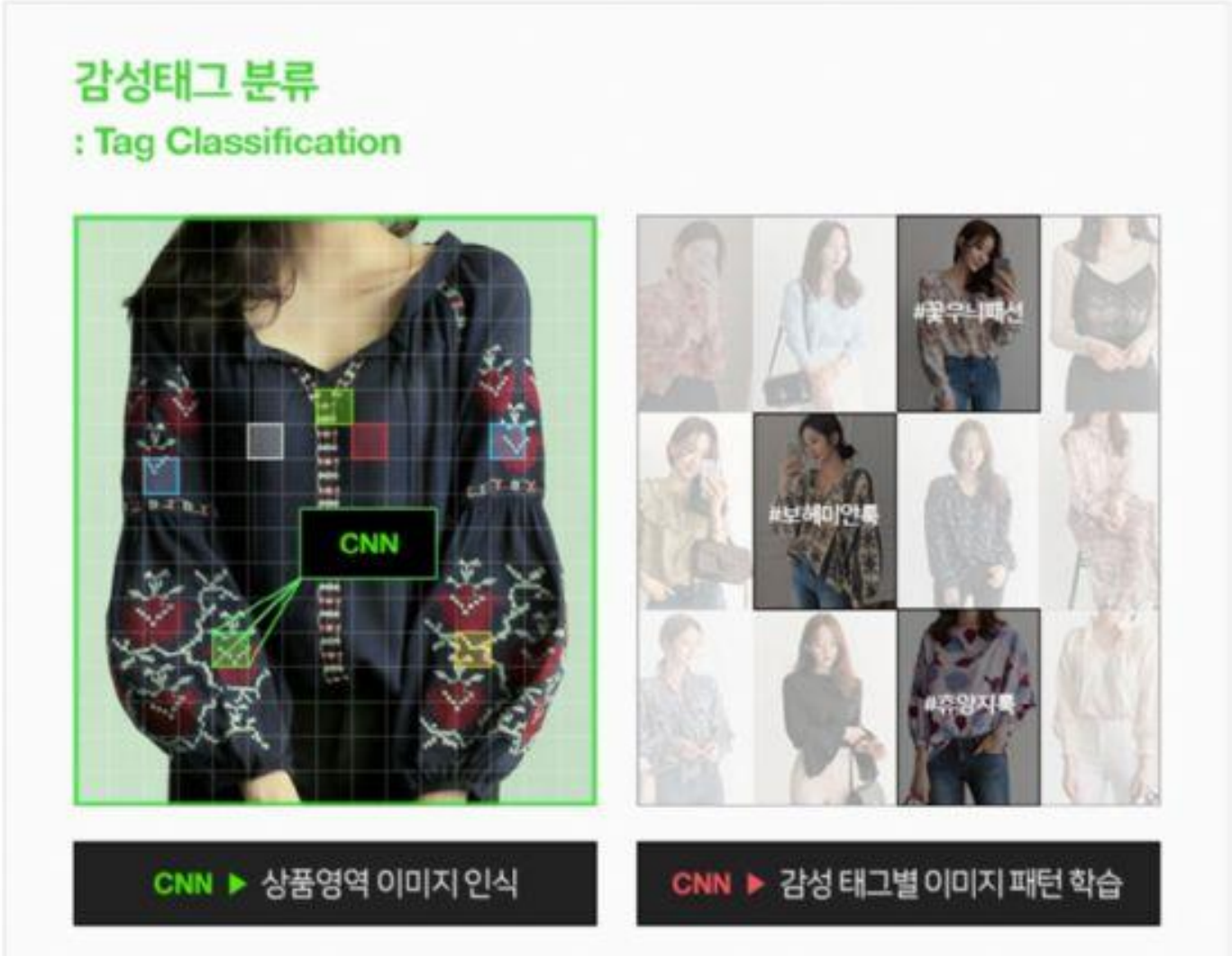
인간을 흉내 낸 인공신경망과 딥러닝 (2)



인간을 흉내 낸 인공신경망과 딥러닝 (3)



네이버 쇼핑은 CNN을? (1)



네이버 쇼핑은 CNN을? (2)

예적금보다 로보어드바이저?

Part 1

차이를 확인하는 데이터 요약

데이터의 구성

데이터는 일반적으로 변수와 관측치로 구성

변수(Variable)

- 관심 대상을 바라보는 관점

관측치(Observation)

- 각 관측 대상에 대해 변수별로 측정된 값

데이터와 데이터 공간

데이터 공간

- 변수와 관측치로 구성된 데이터가 만들어 내는 공간
- 변수 개수만큼의 차원과 관측치 개수만큼의 점으로 구성

데이터 공간의 예시

- 2개 변수(성별과 몸무게), 5개 관측치로 구성된 데이터는 2차원 공간으로 표현가능

이름	성별	몸무게
A	여자	55kg
B	여자	60kg
C	남자	65kg
D	여자	50kg
E	남자	80kg

알파벳을 활용한 데이터의 표현

일반적인 데이터의 크기 등을 표현하기 위해 영어 알파벳을 활용

- p : 변수의 개수
- n : 관측치의 개수
- $n \times p$: p개의 변수와 n개의 관측치로 구성된 데이터의 크기
- x : 1개의 변수(y나 z도 1개의 변수를 의미)
- x_1 : x라는 변수의 첫번째 관측치
- x_i : x라는 변수의 i번째 관측치
- Σ : 합계 ($\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$)

기술 통계량과 변수 요약

통계량(Statistics)

- 데이터로부터 계산된 모든 숫자

기술통계량(Descriptive Statistics)

- 하나의 변수나 변수 사이의 관계를 표현하기 위해 데이터로부터 계산된 숫자
- 변수의 구분에 따라 주로 활용하는 통계량이 다름
- 예시
  - 관측치들이 아주 다양한 숫자 값을 가지는 연속형 변수 : 평균과 표준편차
  - 관측치들이 정해진 몇가지 값을 가지는 범주형 변수 : 표

[1개 연속형 변수의 요약]
“Continuous”

정렬과 순서 통계량

정렬(Sort)

- 대표적인 연속형 변수의 요약 방법
- 일반적으로 작은 것부터 큰 순서로 나열하는 오름차순을 활용

순서 통계량(Order Statistics)

- 연속형 변수의 정렬된 값
- 대표적인 순서통계량
  - 최솟값(Minimum) : 정렬해서 가장 먼저 나오는 가장 작은 값
  - 최댓값(Maximum): 정렬해서 가장 나중에 나오는 가장 큰 값

정렬과 순서 통계량

50.8	50.9	54.5	55	56	56.7	57.4	58.2	59.1	60.4
60.9	61.4	61.4	61.6	61.7	61.8	62.2	62.4	63.2	63.3
64	64.1	64.1	64.2	64.3	64.6	64.7	66.4	66.4	66.7
67.6	67.8	67.9	68.1	68.5	68.6	68.7	68.7	68.8	69.1
70.4	70.5	71.8	73.2	73.2	73.6	73.6	73.8	74	74.7
75.1	75.2	75.2	75.4	75.5	75.7	75.9	76.3	77.2	77.3
77.8	78.1	78.3	78.3	78.5	78.5	79.1	80.3	80.7	81.1
81.7	81.9	82	82.2	82.7	82.8	82.8	83.1	83.1	83.2
83.4	83.4	83.5	83.6	83.8	84.4	84.4	84.6	84.8	84.8
84.9	85.8	86.4	89.8	90.7	92.5	93.7	94.3	96.2	98.8
99.7									

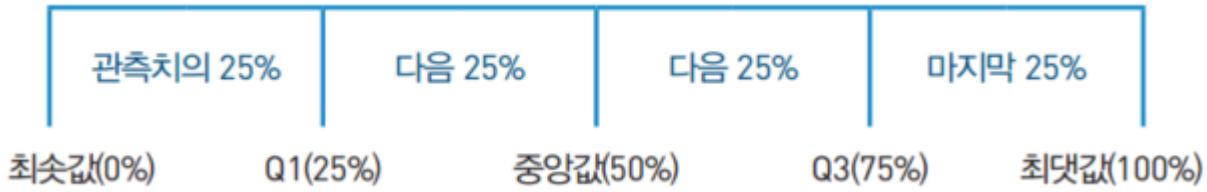
사분위수(Quartile)

분위수(Quantile)

- 관측치들을 정렬하고 특정한 비율로 나눌 때 활용하는 기준 값
- 100등분에 활용하는 백분위수(Percentile)이 대표적

사분위수(Quartile)

- 정렬된 관측치를 25%씩 4등분하는 5개 값
- 최솟값 : 0% 값
- Q1 (1st Quartile) : 25% 값
- 중앙값(Median) : 50% 값
  - 관측치가 100개일 때 : 50번째와 51번째 값의 평균
  - 관측치가 101개일 때 : 51번째 값
- Q3(3rd Quartile) : 75% 값
- 최댓값 : 100% 값



다섯숫자요약과 상자그림

다섯숫자요약과 상자그림

다섯숫자요약

- 사분위수 5개를 계산해서 변수의 분포를 확인하는 과정
- 일반적으로 그룹 간의 연속형 변수 비교를 위해 활용

상자그림(Boxplot)

- 다섯숫자를 수직선에 표현하는 시각화 방법
- Q1과 Q3를 연결해서 상자로 표현

히스토그램

도수분포표(Frequency distribution table)

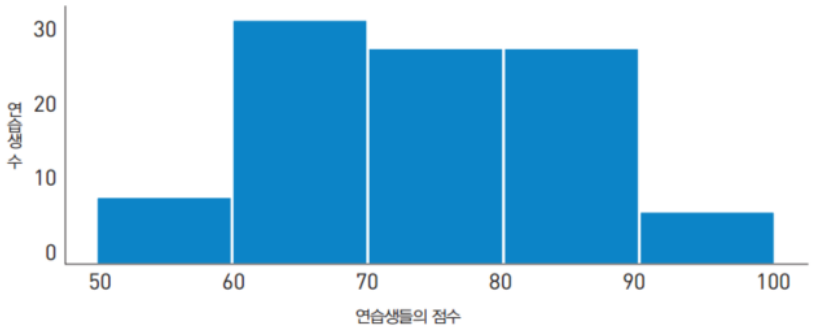
- 사전에 정한 각 구간에 포함된 관측치의 수를 정리한 표
- 구간은 목적에 따라 다양한 단위로 지정가능
- 일반적으로 각 구간의 길이는 동일
  - 예) 5살 간격, 10점 간격 등

히스토그램(Histogram)

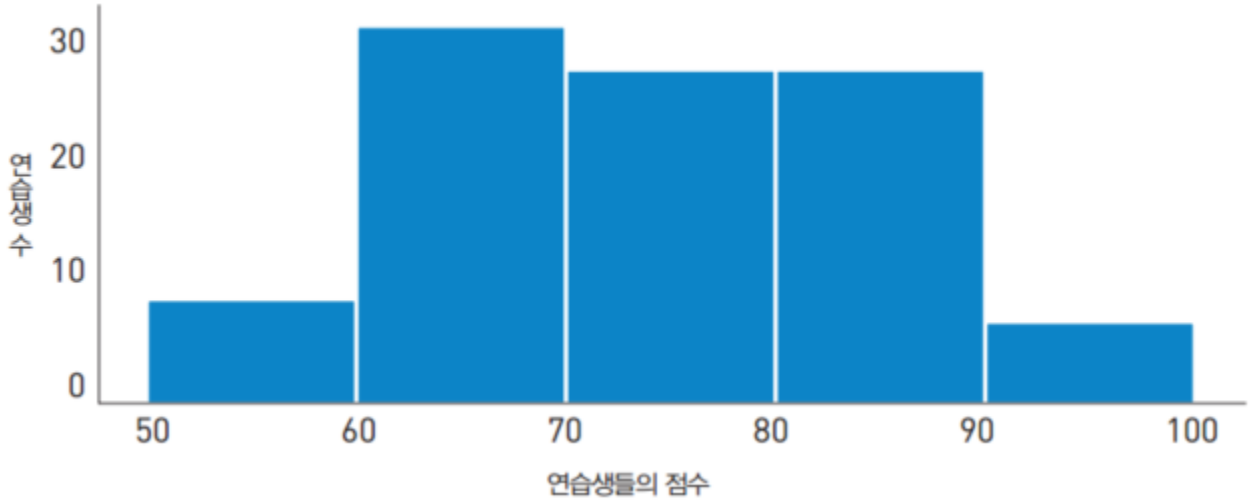
- 도수분포표의 숫자를 높이로 표현한 그림

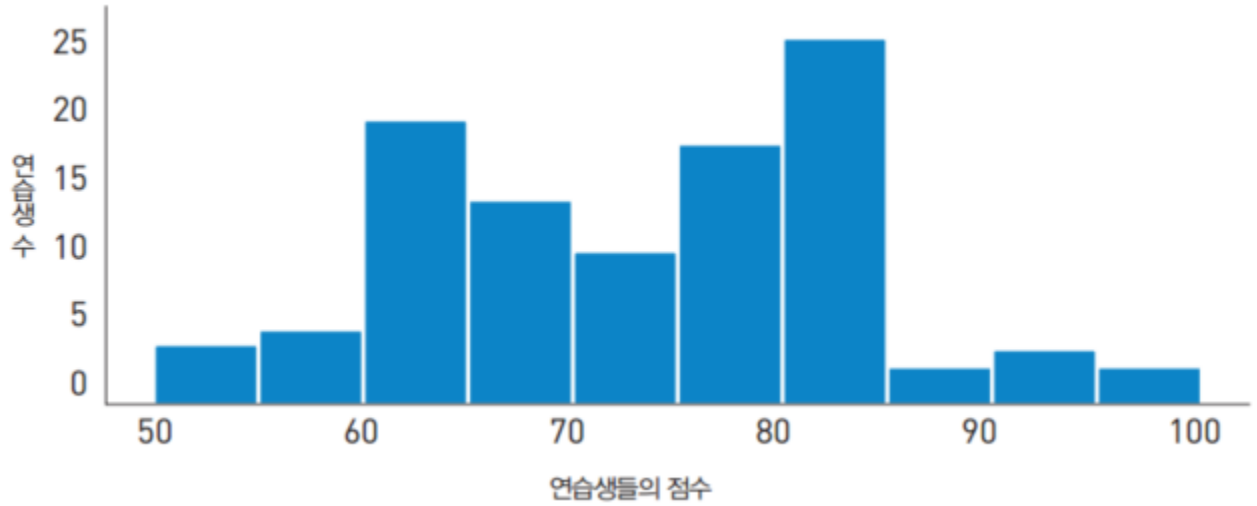
50.8	50.9	54.5	55	56	56.7	57.4	58.2	59.1	60.4
60.9	61.4	61.4	61.6	61.7	61.8	62.2	62.4	63.2	63.3
64	64.1	64.1	64.2	64.3	64.6	64.7	66.4	66.4	66.7
67.6	67.8	67.9	68.1	68.5	68.6	68.7	68.7	68.8	69.1
70.4	70.5	71.8	73.2	73.2	73.6	73.6	73.8	74	74.7
75.1	75.2	75.2	75.4	75.5	75.7	75.9	76.3	77.2	77.3
77.8	78.1	78.3	78.3	78.5	78.5	79.1	80.3	80.7	81.1
81.7	81.9	82	82.2	82.7	82.8	82.8	83.1	83.1	83.2
83.4	83.4	83.5	83.6	83.8	84.4	84.4	84.6	84.8	84.8
84.9	85.8	86.4	89.8	90.7	92.5	93.7	94.3	96.2	98.8
99.7									

점수 구간	50.0~59.9	60.0~69.9	70.0~79.9	80.0~89.9	90~100
관측치 수	9	31	27	27	7

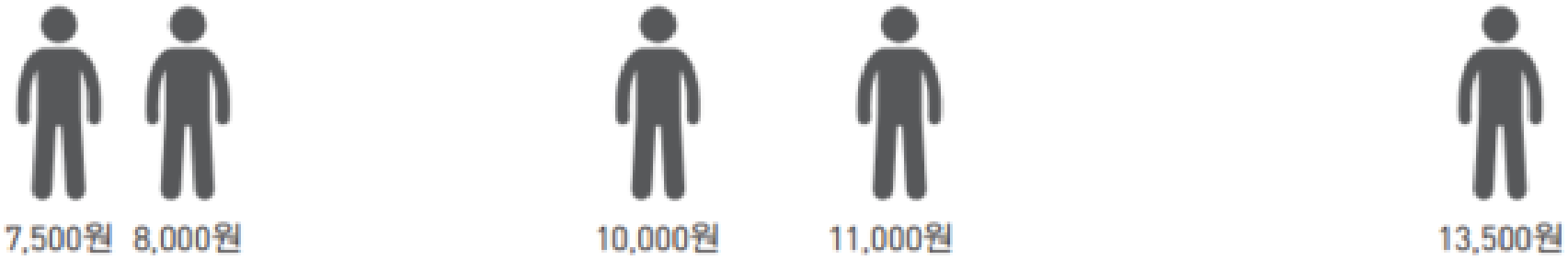


히스토그램


30
20
10
0
연습생 수
50
60
70
80
90
100
연습생들의 점수


25
20
15
10
5
0
연습생 수
50
60
70
80
90
100
연습생들의 점수

평균

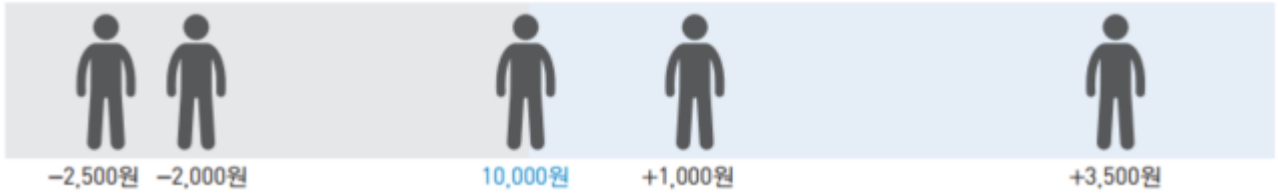


평균

평균(Mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- n개 관측치가 가지고 있는 값들의 전체 합계를 n개 관측치가 공평히 나눴을 때의 값
- 친구들과 함께 밥을 먹은 후 더치페이가 아닌 “N빵”을 했을 때 내가 내야할 금액



- 평균보다 작은 관측치들의 평균보다 작은 정도의 합과
- 평균보다 큰 관측치들의 평균보다 큰 정도의 합은 항상 동일

분산

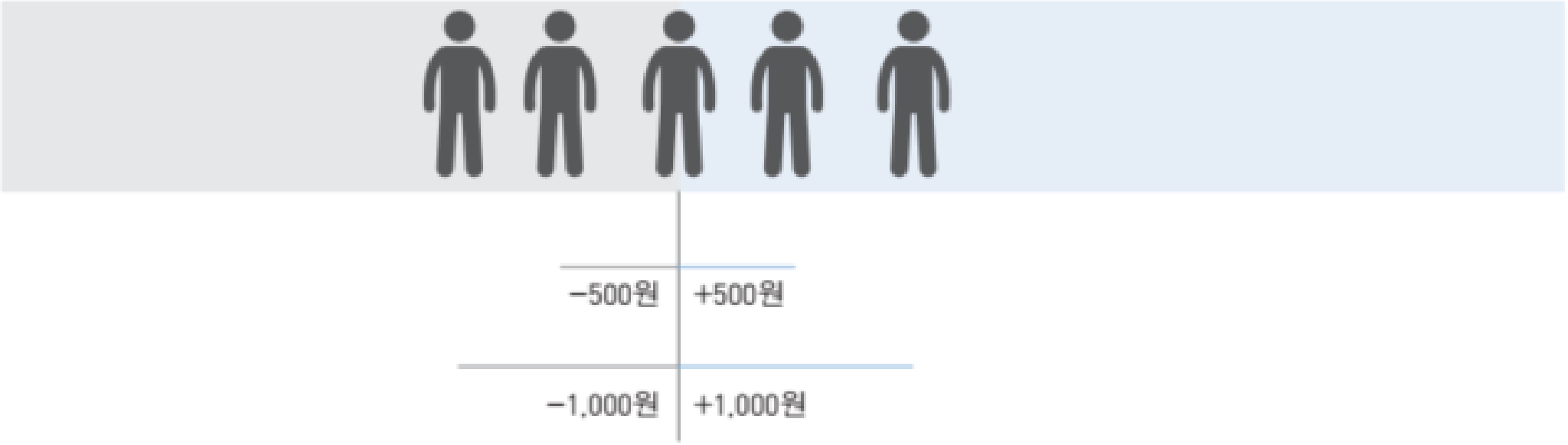
분산

분산(Variance)

$$s_x{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- n개 관측치들이 전반적으로 평균에서부터 얼마나 떨어져 있는지 계산한 값
- 계산 과정
  - 각 관측치들의 평균($\bar{x}$)으로부터의 거리 $(x_i - \bar{x})$를 계산
  - 계산된 거리를 제곱한 $(x_i - \bar{x})^2$를 모두 더해 $\sum_{i=1}^{n}(x_i - \bar{x})^2$를 계산
  - 계산된 제곱합을 $(n-1)$로 나눠 분산 $s_x{}^2$를 계산

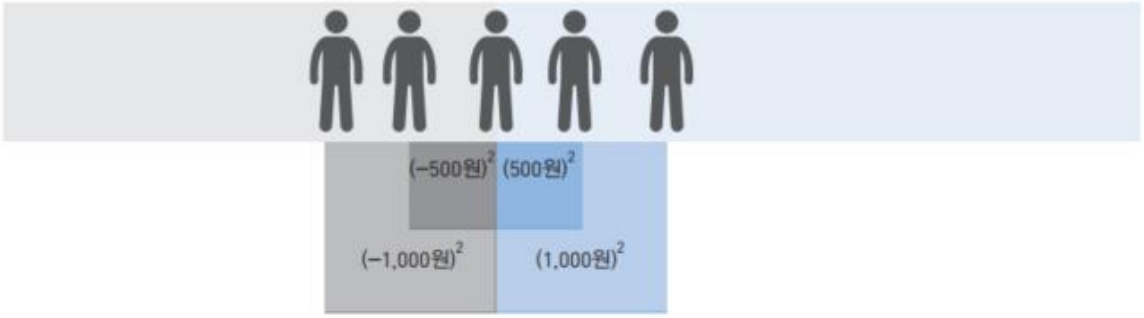
분산의 의미

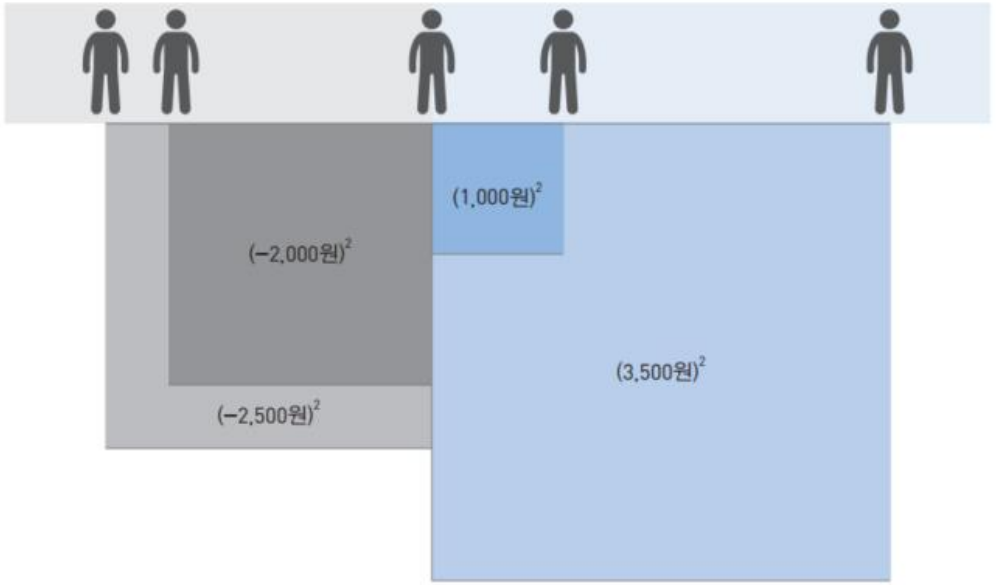


분산의 의미

분산이 클수록 관측치들 간의 차이가 큼

- 예제) 평균이 모두 10,000원인 두 그룹의 분산
  - 첫번째 그룹의 분산 : $625{,}000원^2$
  - 두번째 그룹의 분산 : $5{,}875{,}000원^2$





분산과 표준편차

분산의 단점

- 단위(Scale)가 큼
  - 고작 $10{,}000$원 정도의 밥을 먹었는데 분산은 $5{,}875{,}000$원^2
- 이해할 수 없는 단위(Unit)
  - 수식에서 계산된 "원^2"이라는 단위를 해석할 수가 없음

표준편차(Standard Deviation)의 활용

$$s_x = \sqrt{{s_x}^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 분산의 제곱근($\sqrt{\ }$). 단위 문제를 해결
  - ${s_x}^2 = 5{,}875{,}000$원^2 $\Rightarrow s_x = \sqrt{{s_x}^2} = 2{,}424$원

표준화(Standardization)

	2011	2013	2015	2017
1등급	79	92	100	92
2등급	72	83	96	88
3등급	64	75	90	83

연도별 수능 수리가형 원점수 등급 기준점수

표준화(Standardization)

표준화의 목적

- 특성이 서로 다른 두 연속형 변수의 값을 비교

표준화 과정

- 중심화(Centering)
  - 각 관측치에서 자신이 속한 변수의 평균을 빼는 과정
  - 평균으로부터 얼만큼 떨어졌는지 상대적인 값으로 변환

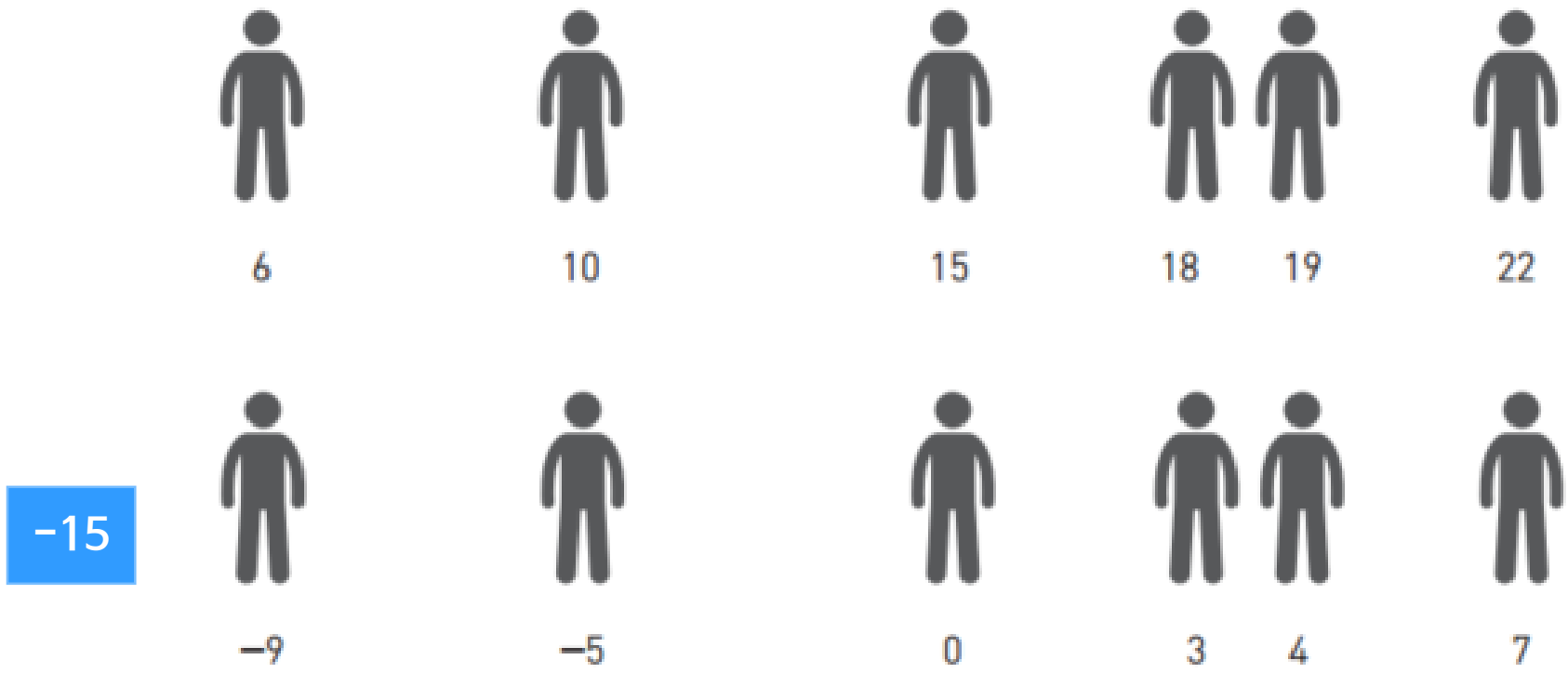
- 척도화(Scaling)
  - 각 관측치를 자신이 속한 변수의 표준편차로 나누는 과정
  - 절대적인 크기가 아닌 표준편차에 비해 몇배나 크고 작은 지 상대적인 값으로 변환
  - 단위(Unit)와 또다른 단위(Scale)의 차이에서 오는 비교 불가능 문제를 해결

표준화(Standardization)

표준화(Standardization)

표준화(Standardization)

표준화(Standardization)



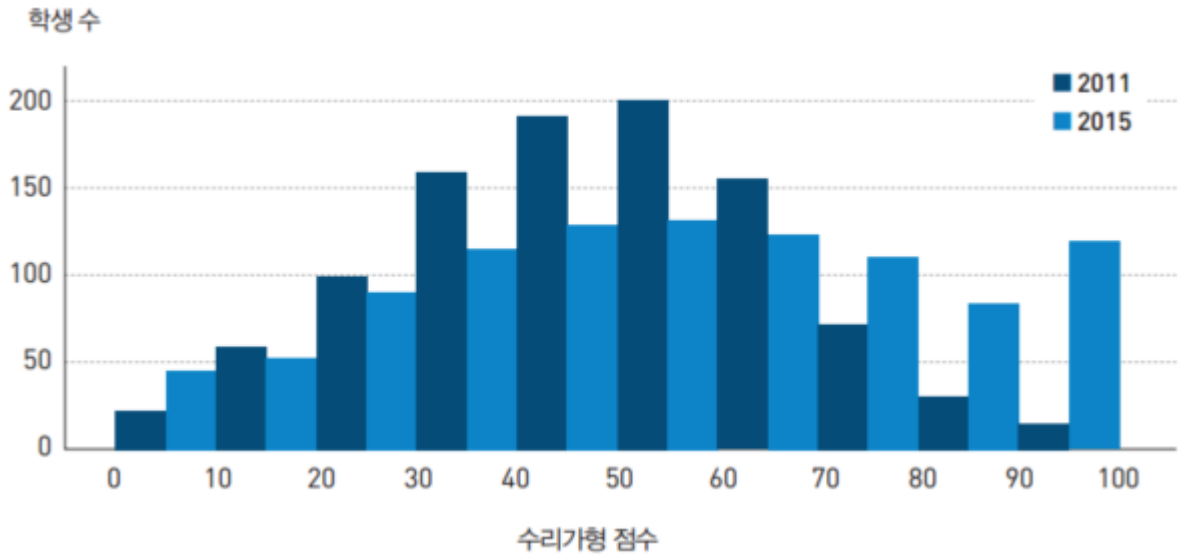
표준화(Standardization)

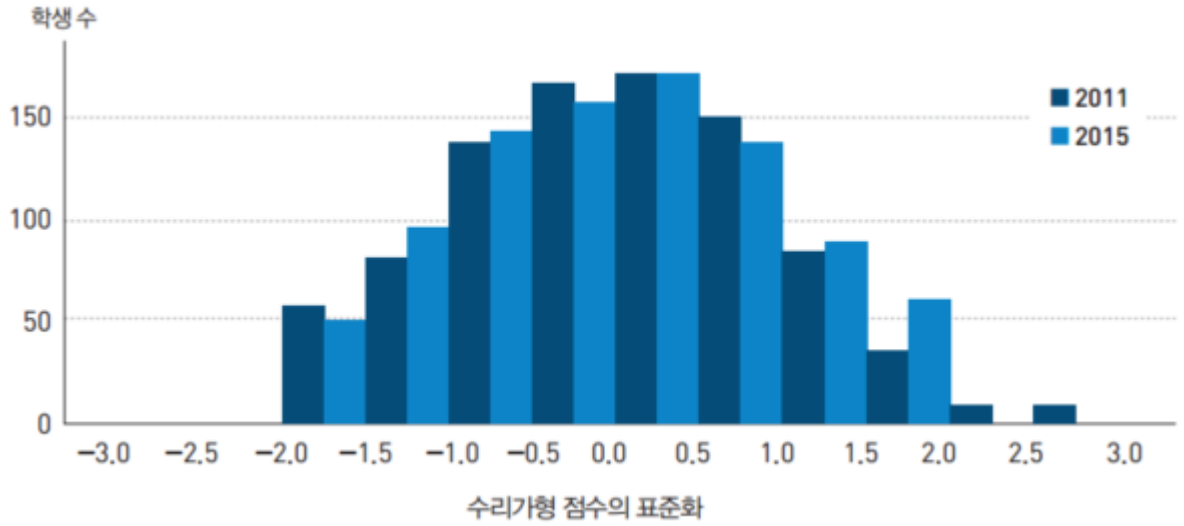
간단한 표준화 예제

대입수능의 표준점수가 대표적인 표준화 예제

- 2개 년도 수리 가형 성적 비교
  - 연도별이나 과목별로 서로 분포가 다른 점수를 비교 가능
  - 2011년(불수능) : 평균 47.8점, 표준편차 19.7점
  - 2015년(물수능) : 평균 55.4점, 표준편차 28.5점
- 2011년의 80점과 2015년의 100점 중 누가 더 성적이 좋다고 할 수 있을까?
  - 절대적인 차이로 비교 불가/표준화를 통해 상대적인 값으로 변환
- 중심화
  - 2011년 : $(80점 - 47.8점) = 32.2점 < (100점 - 55.4점) = 44.6점$: 2015년
- 표준화(중심화+척도화)
  - 2011년 : $(80점 - 47.8점)/19.7점 = 1.63$
  - 2015년 : $(100점 - 55.4점)/28.5점 = 1.56$
  - 난이도 차이를 고려했을 때 2011년이 상대적으로 더 높은 점수인 것을 확인

간단한 표준화 예제


학생 수
2011
2015
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
수리가형 점수


학생 수
2011
2015
150
100
50
0
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
수리가형 점수의 표준화

[1개 범주형 변수의 요약]
“Categorical”

범주형 변수와 수준(Levels)

범주형 변수

- 관측치의 개수와 상관없이 한정적인 값을 갖는 변수
- 간단한 예제
  - 5명의 중국집 음식 주문 : 짜장, 짜장, 짬뽕, 짬뽕, 볶음밥
  - 3번 동전 던지기 결과 : 앞면, 뒷면, 앞면

수준

- 범주형 변수가 갖는 한정적인 값 목록
- 간단한 예제
  - n명의 중국집 음식 주문 : 짜장, 짬뽕, 볶음밥
  - n번 동전 던지기 결과 : 앞면, 뒷면

표와 파이차트, 막대그래프

계수(Counting)

- 범주형 변수는 계수를 통해 요약. 수준별 관측치의 수를 계산
- 예제) 동전 10번 던지기 결과

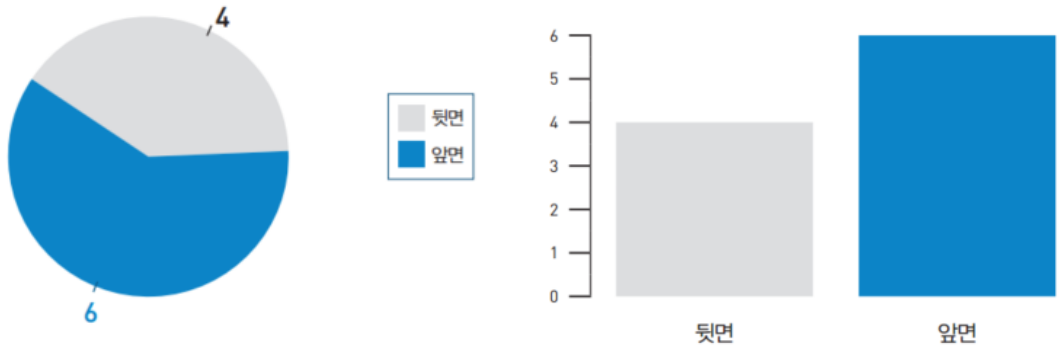
횟수	1	2	3	4	5	6	7	8	9	10
수준(앞/뒤)	앞	앞	뒤	앞	뒤	뒤	뒤	앞	앞	앞

→

수준	앞면	뒷면
횟수	6	4

파이차이와 막대그래프

- 요약된 표는 그래프를 통해 표현 가능



확률(Probability)

확률

- 아직 결과를 모르는 사건(Event)의 발생 가능성을 0부터 1사이의 숫자로 표현한 것
- 이론적(Theoretical) 확률
  - 논리적, 수리적으로 계산한 확률
  - 예제) 나눔로또 6/45의 1등 당첨 확률

$$\frac{1}{\binom{45}{6}} = \frac{1}{8{,}145{,}060} = 0.0000001227738 = 0.0000123\%$$

- 경험적(Empirical) 확률
  - 직접 관찰 또는 모의실험(Simulation), 데이터로 부터 계산된 확률

확률(Probability)

당첨 번호	1번	2번	3번	4번	5번	6번	7번	8번	9번
추첨 횟수	10	10	7	9	6	9	10	7	1
당첨 번호	10번	11번	12번	13번	14번	15번	16번	17번	18번
추첨 횟수	10	12	5	8	4	9	7	6	5
당첨 번호	19번	20번	21번	22번	23번	24번	25번	26번	27번
추첨 횟수	8	6	7	4	6	7	7	3	5
당첨 번호	28번	29번	30번	31번	32번	33번	34번	35번	36번
추첨 횟수	10	5	9	3	5	15	9	7	7
당첨 번호	37번	38번	39번	40번	41번	42번	43번	44번	45번
추첨 횟수	8	7	7	4	7	4	7	8	8

2016년 나눔로또 53회 각 번호 출현 횟수

데이터 분석과 확률

데이터 분석 ≃ 확률 계산

- 데이터 분석의 목적
  - 과거 데이터를 통해 불확실한 미래를 예측하는데 활용
  - 예제) 기대값(Expectation) : 확률을 고려했을 때 평균적으로 나올 것 같은 값
- 데이터 분석의 과정은 데이터로부터 확률을 계산하는 것으로 볼 수 있음

확률모형(Probability model)

- 데이터를 활용해서 관심있는 사건을 확률로 정의한 것
- 데이터 속에 있는 차이를 확률로 설명