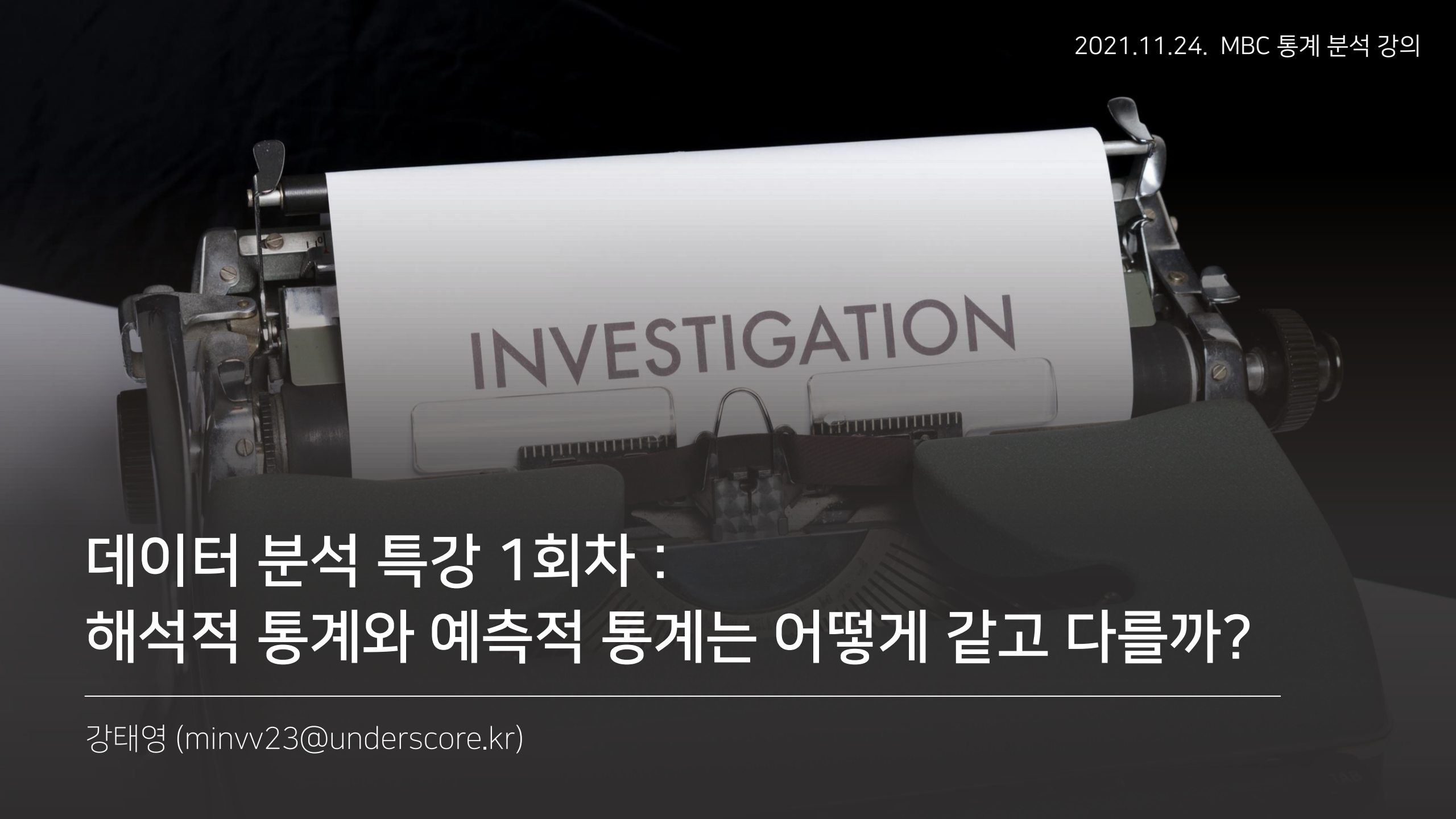
2021.11.24. MBC 통계 분석 강의



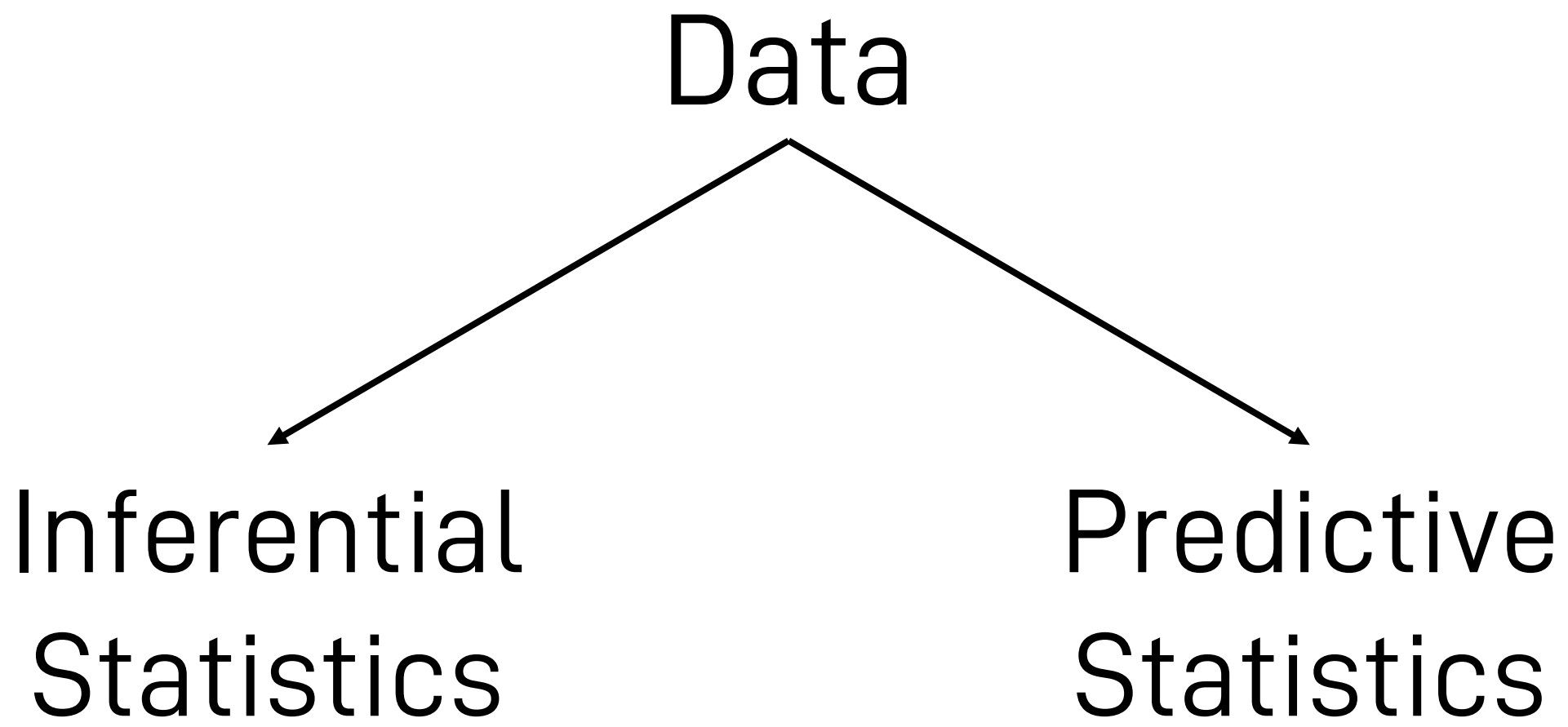
데이터 분석 특강 1회차 : 해석적 통계와 예측적 통계는 어떻게 같고 다를까?

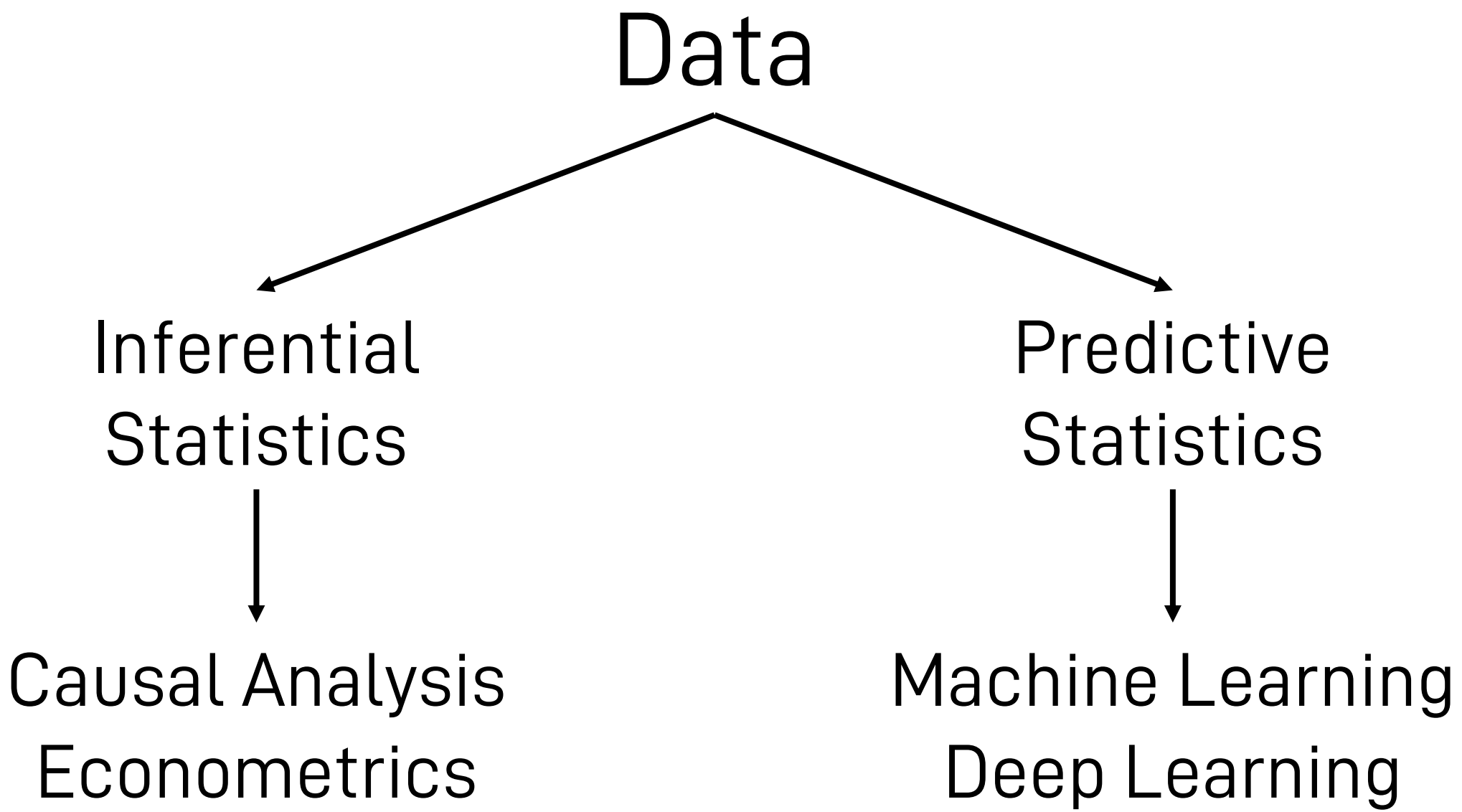
강태영 (minvv23@underscore.kr)

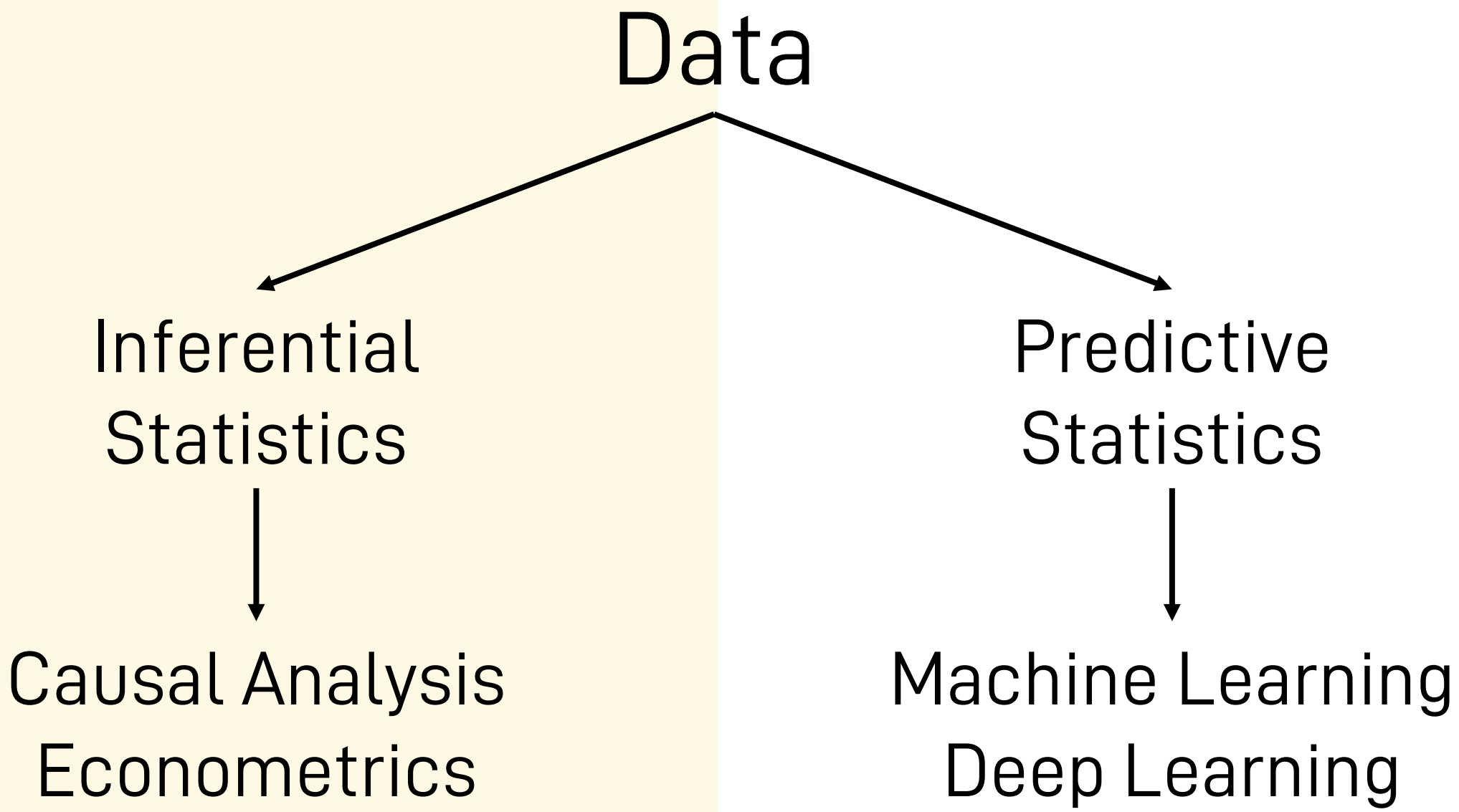
Underscore

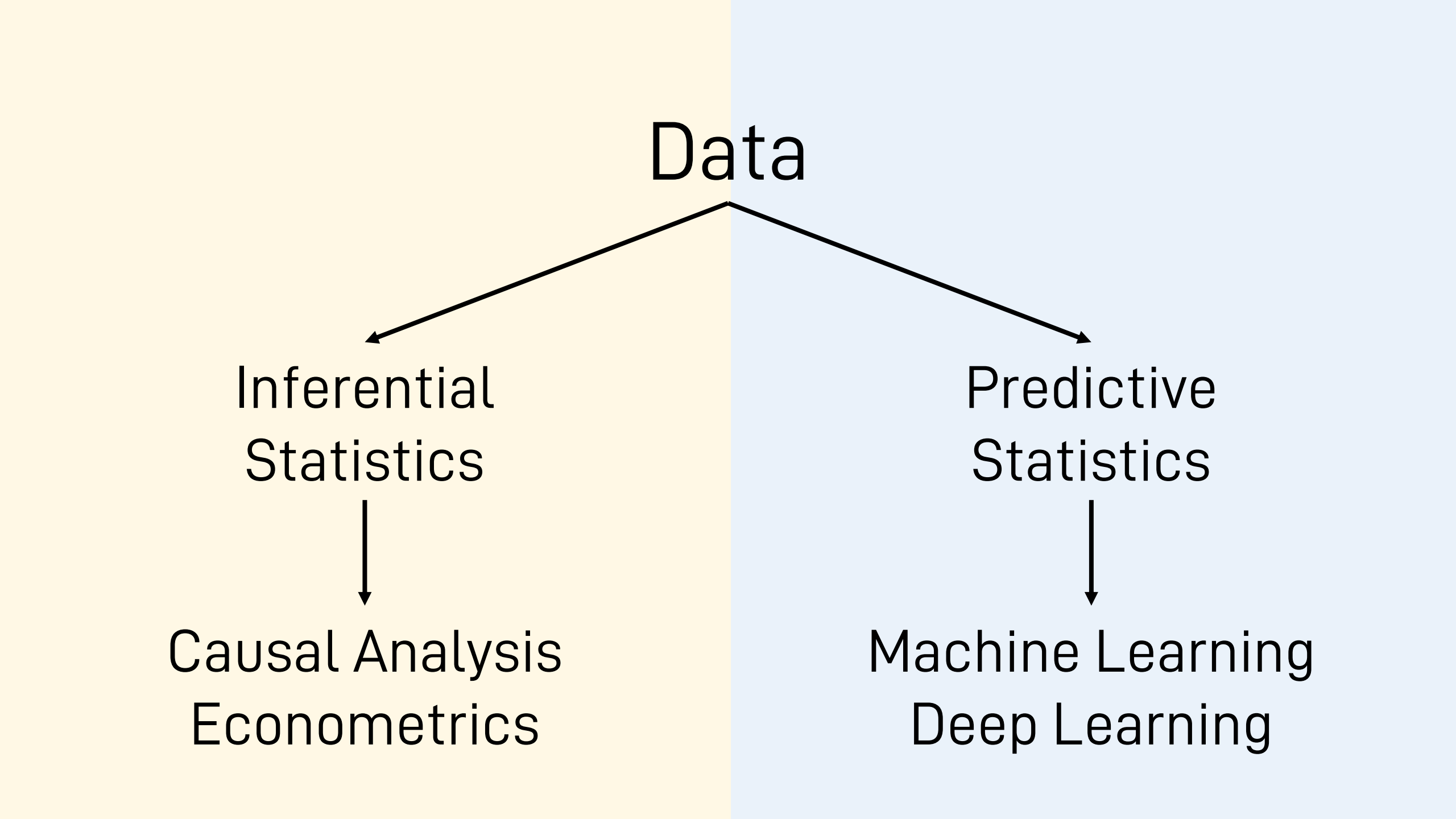
데이터 연구의 두 가지 방식

Data


Data
Inferential Statistics
Predictive Statistics

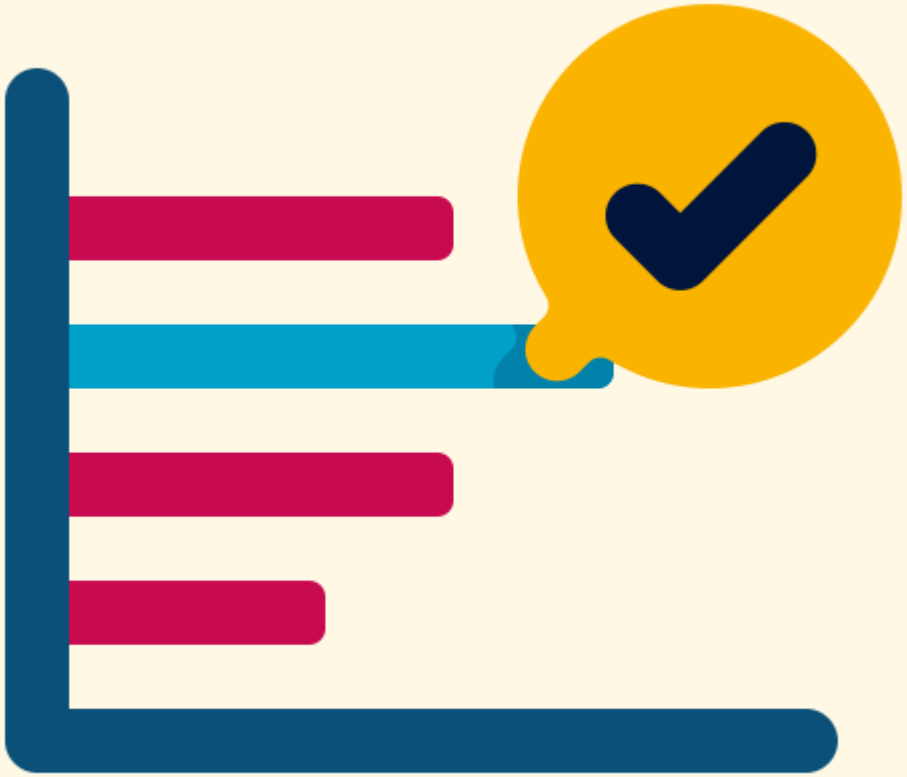

Data
Inferential Statistics
Predictive Statistics
Causal Analysis
Econometrics
Machine Learning
Deep Learning

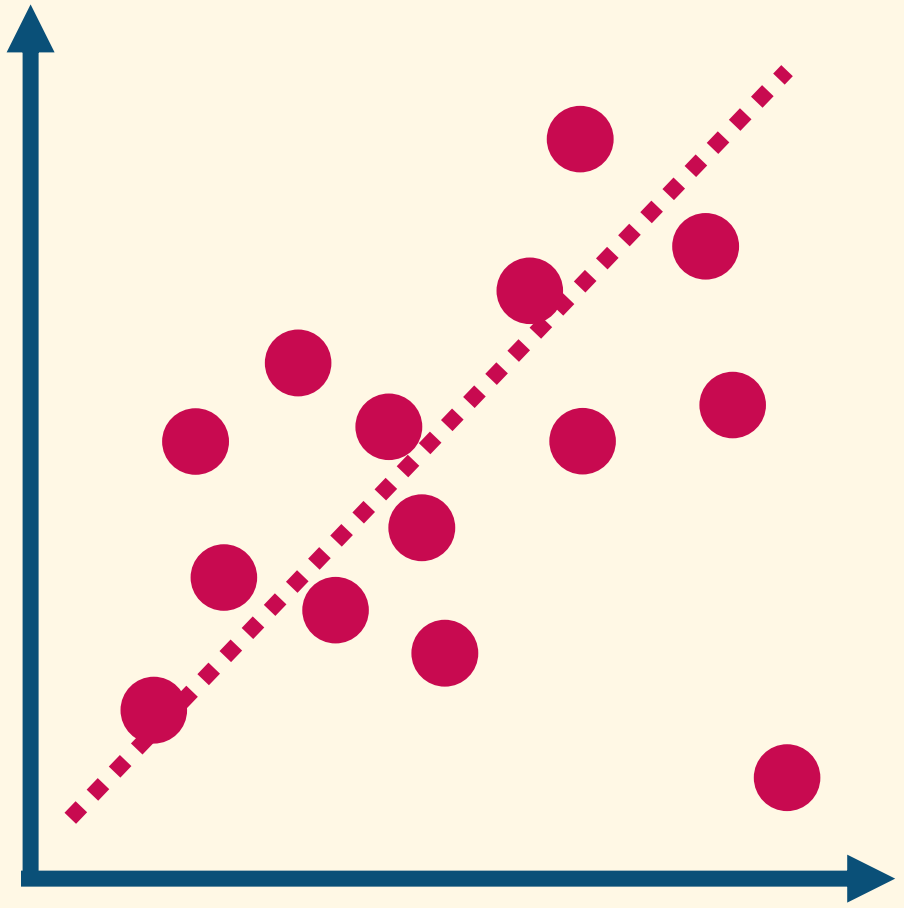

Data
Inferential Statistics
Causal Analysis
Econometrics
Predictive Statistics
Machine Learning
Deep Learning


Data
Inferential Statistics
Predictive Statistics
Causal Analysis
Econometrics
Machine Learning
Deep Learning

Inferential Statistics

Causal Analysis
Econometrics





특정범죄 가중처벌 등에 관한 법률 일부개정법률안

민식이법

법 제정

사고 감소

법 제정

주의↑

사고 감소

법 제정

주의↑

사고 감소

COVID19

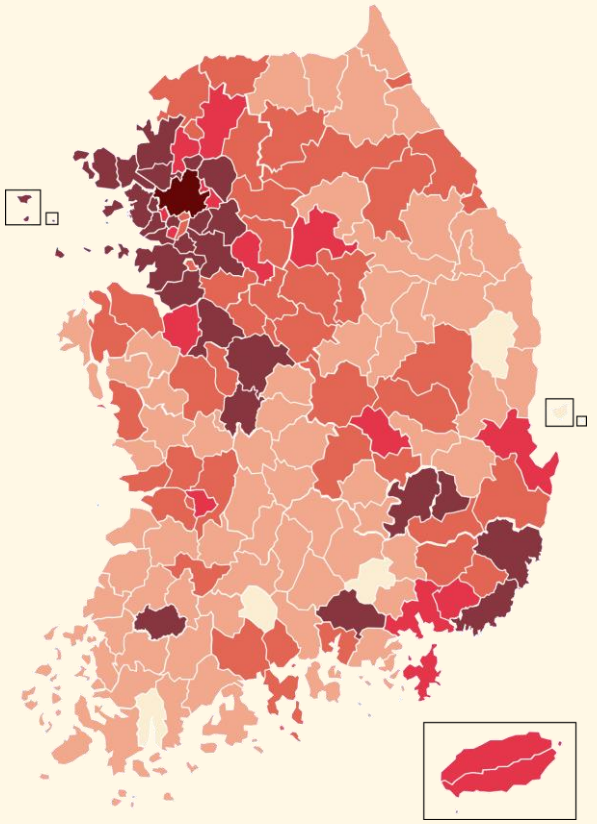
등교금지

주의↑

사고 감소

VID19

등교금지

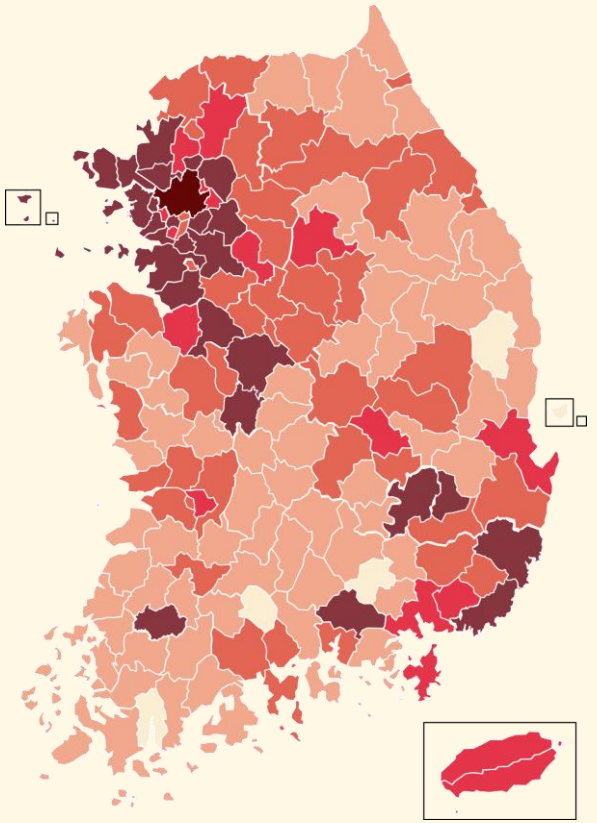


주의↑

사고 감소

VID19

등교금지

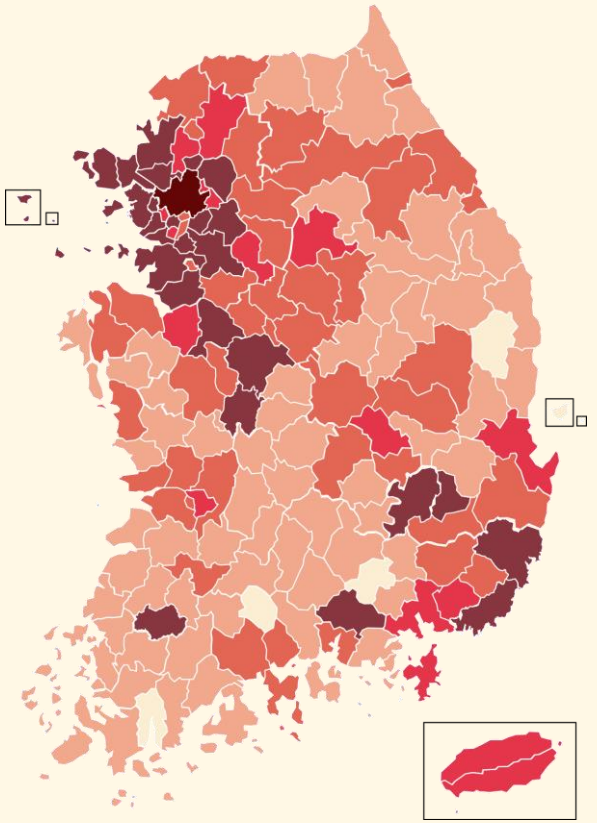


주의↑

사고 감소

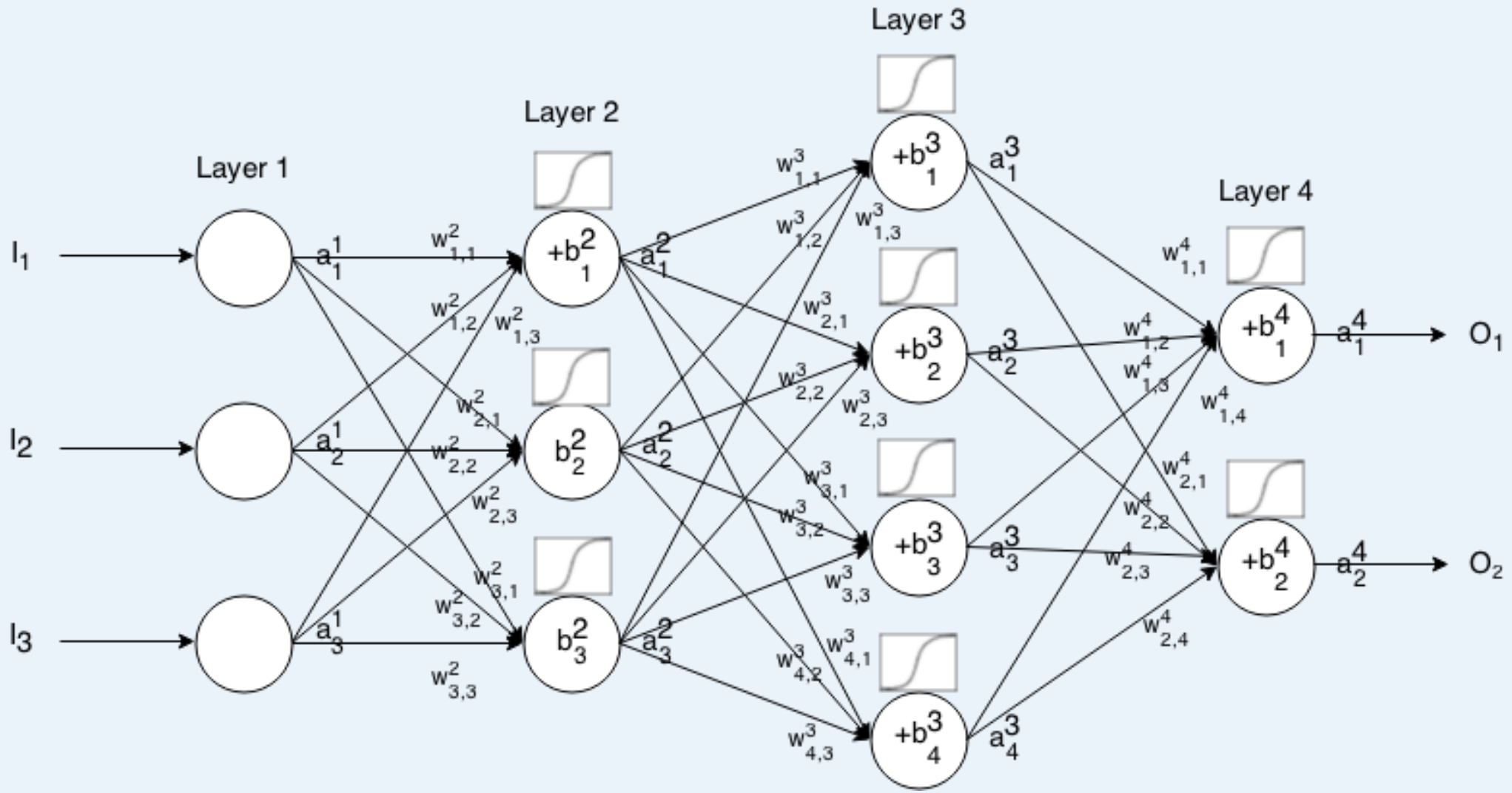
VID19

등교금지



Predictive
Statistics

Machine Learning
Deep Learning


Layer 1
Layer 2
Layer 3
Layer 4
I_1
I_2
I_3
O_1
O_2
$+b^2_1$
b^2_2
b^2_3
$+b^3_1$
$+b^3_2$
$+b^3_3$
$+b^3_4$
$+b^4_1$
$+b^4_2$
a^1_1
a^1_2
a^1_3
a^2_1
a^2_2
a^2_3
a^3_1
a^3_2
a^3_3
a^3_4
a^4_1
a^4_2
$w^2_{1,1}$
$w^2_{1,2}$
$w^2_{1,3}$
$w^2_{2,1}$
$w^2_{2,2}$
$w^2_{2,3}$
$w^2_{3,1}$
$w^2_{3,2}$
$w^2_{3,3}$
$w^3_{1,1}$
$w^3_{1,2}$
$w^3_{1,3}$
$w^3_{2,1}$
$w^3_{2,2}$
$w^3_{2,3}$
$w^3_{3,1}$
$w^3_{3,2}$
$w^3_{3,3}$
$w^3_{4,1}$
$w^3_{4,2}$
$w^3_{4,3}$
$w^4_{1,1}$
$w^4_{1,2}$
$w^4_{1,3}$
$w^4_{1,4}$
$w^4_{2,1}$
$w^4_{2,2}$
$w^4_{2,3}$
$w^4_{2,4}$

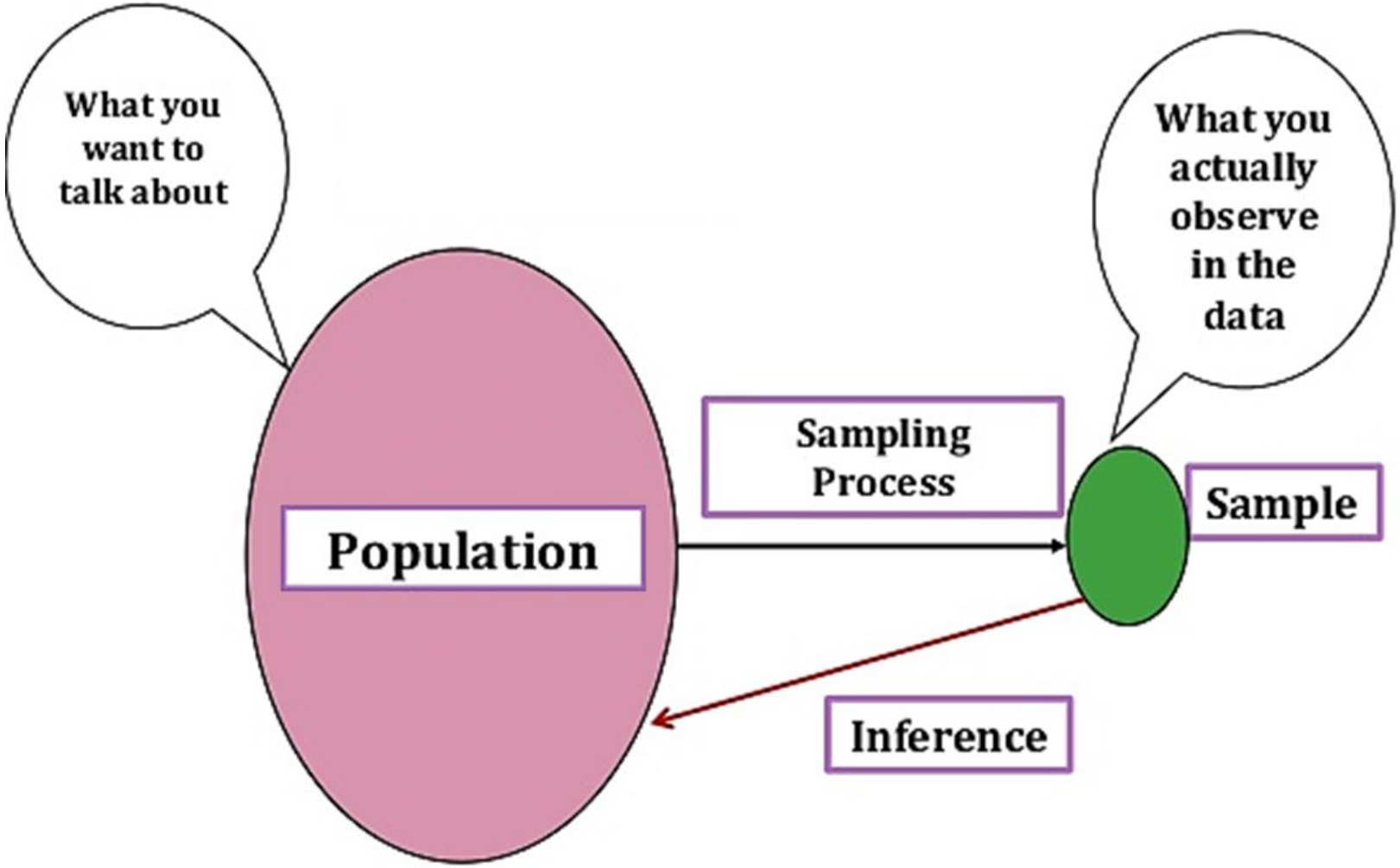
'19년도 하반기 SK이노베이션 역량기반지원

No.	시스템 ID	자소서 ID	4. Teamwork 발휘			시스템 ID	자소서 ID	1. 동기부여 / 높은 목표 설정						
			지원서	평가자 1	평가자 2			평가자 1				평가자 2		
								높은 목표 수준	[illegible]	원점수	표준점수	높은 목표 수준	[illegible]	원점수
1	SKI_경영지원_배터리경영관리_1	[illegible]	[Sharing and Reporting with	한국인 교수님의 피드백을 즉	글로벌 기업의 브랜드 전략을	SKI_경영지원_배터리경영관리_1	[illegible]	3	1	4	**4.32**	1	2	3
2	SKI_경영지원_배터리경영관리_2	[illegible]	[자동차 배터리 생산라인 재고	정확한 가이드라인 제공을 위	경영관리 담당자로 자동차 배	SKI_경영지원_배터리경영관리_2	[illegible]	2	2	4	**4.32**	2	3	5
3	SKI_경영지원_배터리경영관리_3	[illegible]	[인도에 울려퍼진 아리랑] 소통	공연 미션을 성공적으로 마치	소통을 통한 팀워크로 / 인도	SKI_경영지원_배터리경영관리_3	[illegible]	3	3	6	**6.42**	2	3	5
4	SKI_경영지원_배터리경영관리_4	[illegible]	협업은 작은 문제라도 함께 고	팀원들과 함께 복지관을 직접	크라우드펀딩을 진행해 1300	SKI_경영지원_배터리경영관리_4	[illegible]	2	2	4	**4.32**	2	3	5
5	SKI_경영지원_배터리경영관리_5	[illegible]	교양수업 팀프로젝트에서 팀원	제가 주도적으로 팀을 이끌었	교양수업 팀프로젝트에서 / 교	SKI_경영지원_배터리경영관리_5	[illegible]	1	2	3	**3.26**	2	2	4
6	SKI_경영지원_배터리경영관리_6	[illegible]	산호세 지역에 있는 쿠퍼티노	다른 강사님들과 교실 안과 밖	산호세 지역에 / 한인들을 위	SKI_경영지원_배터리경영관리_6	[illegible]	1	2	3	**3.26**	2	2	4
7	SKI_경영지원_배터리경영관리_7	[illegible]	구성원들과의 적극적인 피드백	발표 대본을 만들고 발표 연습	적극적인 피드백을 기반으로	SKI_경영지원_배터리경영관리_7	[illegible]	1	1	2	**2.21**	2	2	4
8	SKI_경영지원_배터리경영관리_8	[illegible]	#협력을 통해 이뤄낸 `KCC 와	이를 위해 소셜, 재정, 홍보, 아	대학 시절 한인경영학생회	SKI_경영지원_배터리경영관리_8	[illegible]	4	3	7	**7.47**	3	4	7
9	SKI_경영지원_배터리경영관리_9	[illegible]	[함께 이루어낸 쇼룸의 성공]	대표님과 상의 후 민속촌에서	학내 컨설팅학회에서 / 광화문	SKI_경영지원_배터리경영관리_9	[illegible]	3	1	4	**4.32**	2	3	5
10	SKI_경영지원_배터리경영관리_10	[illegible]	대학 입학 후에 활동했던 동아	선수들의 기본기가 어느 정도	축구동아리에서 저는 선수 겸	SKI_경영지원_배터리경영관리_10	[illegible]	1	1	2	**2.21**	1	2	3
11	SKI_경영지원_배터리경영관리_11	[illegible]	[목표 달성을 위한 필수 요소,	구매팀과 창고와 계속 소통하	물류팀 재직 시 / 중국 B2B를	SKI_경영지원_배터리경영관리_11	[illegible]	2	2	4	**4.32**	2	2	4
12	SKI_경영지원_배터리경영관리_12	[illegible]	학부 때 JTBC와 YTN에서 인턴	회사의 선배님들을 찾아가 조	인턴 / 서울시에서는 브랜드	SKI_경영지원_배터리경영관리_12	[illegible]	1	2	3	**3.26**	1	2	3
13	SKI_경영지원_배터리경영관리_13	[illegible]	[협의의 구심점이 되다] 올해까	5명이 모두 달라붙어 총 350개	경영전략학회에서 핀테크 스타	SKI_경영지원_배터리경영관리_13	[illegible]	2	2	4	**4.32**	1	2	3
14	SKI_경영지원_배터리경영관리_14	[illegible]	[연령대별 언어적 특성을 반영	프로젝트를 성공시키기 위해	데이터 분석의 이론을 학습한	SKI_경영지원_배터리경영관리_14	[illegible]	2	3	5	**5.37**	2	2	4
15	SKI_경영지원_배터리경영관리_15	[illegible]	중·고등학교부터 탁구를 즐겨	탁구부원과 짝이 되더라도 빠	탁구부'에서 / 전국대학탁구대	SKI_경영지원_배터리경영관리_15	[illegible]	1	1	2	**2.21**	1	1	2
16	SKI_경영지원_배터리경영관리_16	[illegible]	군대는 나라를 지키는 공동의	저는 분대장 직책을 인수인계	분대, 소대, 중대원들과 화합하	SKI_경영지원_배터리경영관리_16	[illegible]	2	2	4	**4.32**	2	2	4
17	SKI_경영지원_배터리경영관리_17	[illegible]	#1주일 만에 최고 고참이 되	먼저 아르바이트를 찾고 있던	서래마을에서 / 최고 경력자가	SKI_경영지원_배터리경영관리_17	[illegible]	1	1	2	**2.21**	1	1	2
18	SKI_경영지원_배터리경영관리_18	[illegible]	[앙숙 관계였던 라이벌 축구 팀	고민 결과 교내가 아닌 외부 대	중앙축구동아리와 경영학과 축	SKI_경영지원_배터리경영관리_18	[illegible]	1	2	3	**3.26**	2	2	4

Underscore

해석적 통계 분석 이해하기

population(모집단), sample(표본), sampling(표집)



평균

사회학자 A의 연봉이 7000만원
통계학자 B의 연봉이 800만원
물리학자 C의 연봉이 1200만원

이들 세 사람의 평균 연봉은?
→ (7000+800+1200)/3명=3000만원

근데 정확히 **'평균'**의 개념을 사용하는 이유가 뭘까?

평균이란

(1) 제곱법에 기초하여

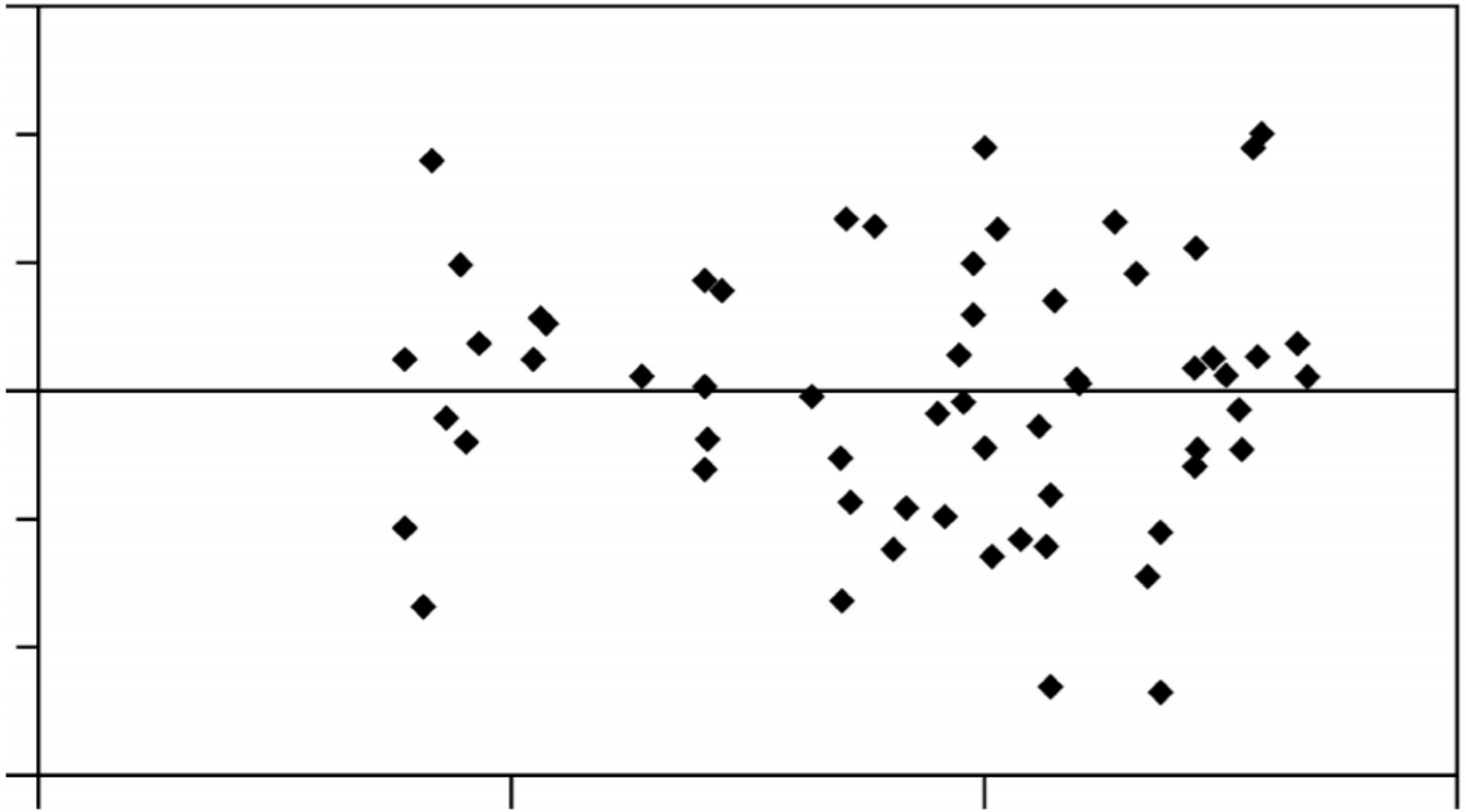
(2) 측정값에 포함되어 있는 차이를

(3) 가장 작게 만듬으로서

특정 데이터의 정보를 가장 잘 보여주는 값!

분산이란 데이터가 퍼져있는 정도!

$\hat{\mu}$



$$Y_i = \hat{\mu} + \hat{e}_i$$

$$\hat{e}_i = Y_i - \hat{\mu}$$

$\hat{\mu}$에 대해서... $\min \sum_{i=1}^{n} \hat{e_i^2} = \min \sum_{i=1}^{n} (Y_i - \hat{\mu})^2$

$$\frac{\partial \sum_{i=1}^{n} \hat{e_i^2}}{\partial \hat{\mu}} = \sum_{i=1}^{n} 2(Y_i - \hat{\mu})(-1) = 0$$

$$\sum_{i=1}^{n} (Y_i - \hat{\mu}) = \sum_{i=1}^{n} Y_i - n\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

GPA = -3×500cc 한 잔+ 5×열람실체류시간 + 2×10점만점설문

키 = 4×일평균칼로리+ 3×조깅시간 + 1.2×부모키 + 5×소득10분위

정치적 보수성 = -5×출생년도 + (-7)×소득10분위+ (-11)×교육년수

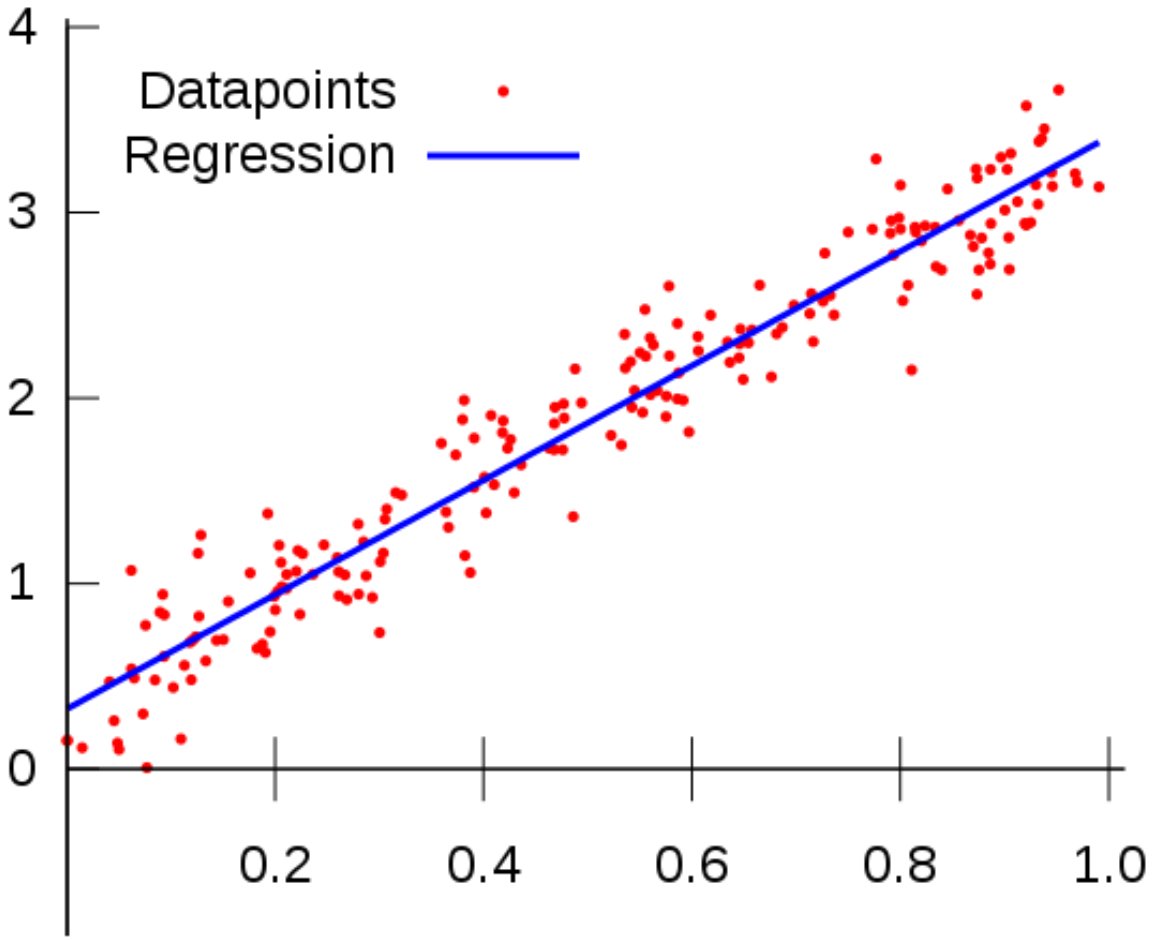
회귀모형

(regression model)

종속변수 = 독립변수1 +독립변수2 + … +독립변수n

종속변수 : 삶의 만족도

독립변수 : 연령, 성별, 교육수준, **소득수준** 등등…



종속변수 기울기 잔차

$$Y_i = a + b_1 X_i + e_i$$

y 절편 독립변수

$$Y_i = \mu + \beta_1 X_i + \beta_2 X_i + \ldots + \beta_k X_i$$

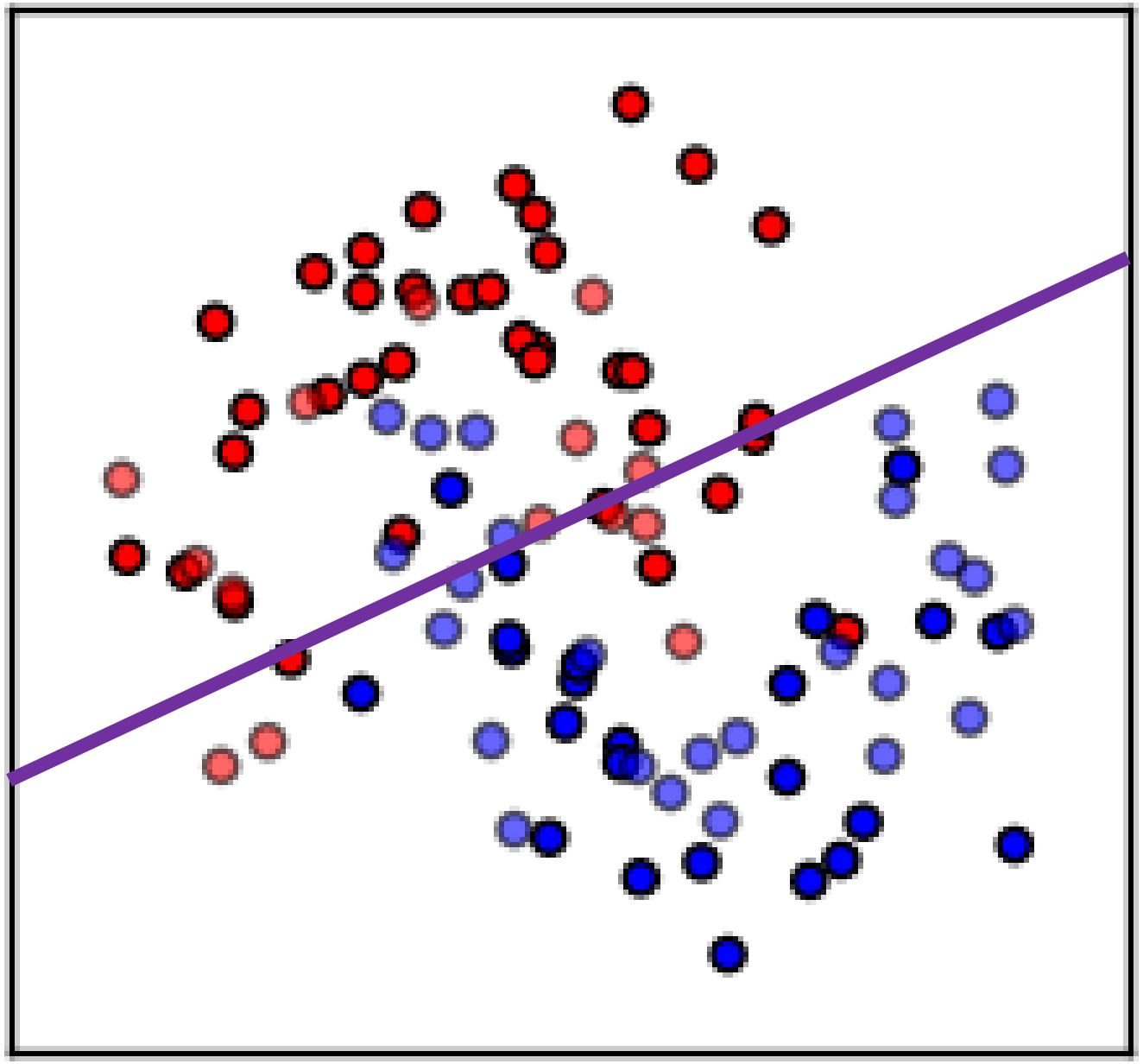
$$Y_i = \widehat{Y}_i + e_i$$

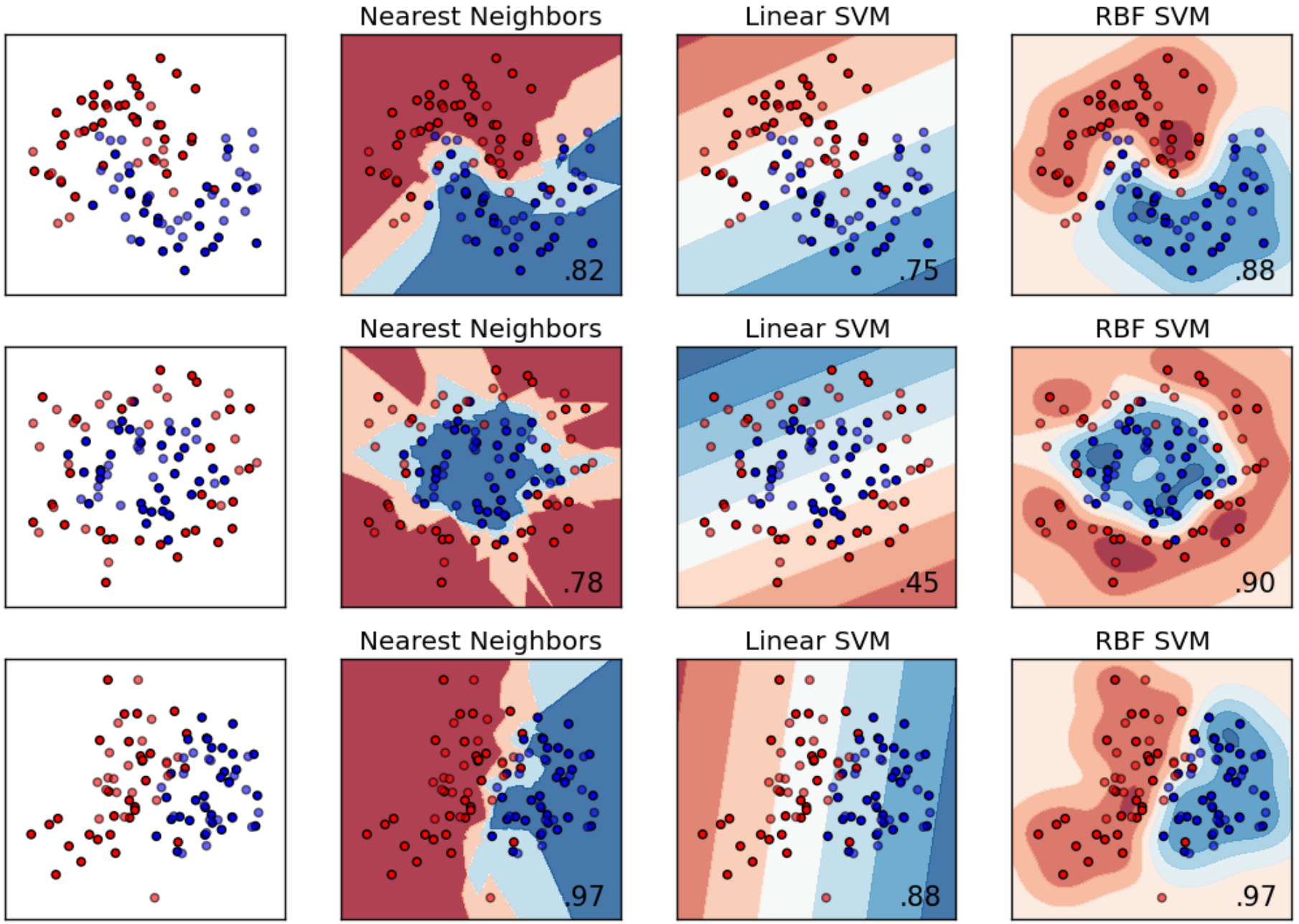
$$Y_i = \hat{\mu} + \widehat{\beta}_1 X_i + \widehat{\beta}_2 X_i + \ldots + \widehat{\beta}_k X_i + e_i$$

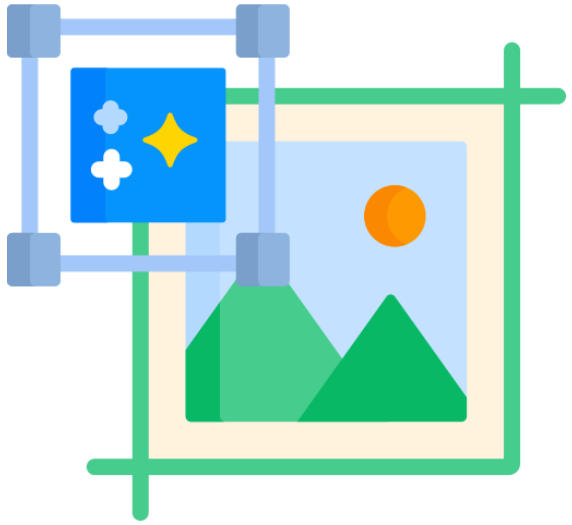
$$Y_i = a + b_1 X_i + b_2 X_i + \ldots + b_k X_i + e_i$$

Underscore

머신러닝과 딥러닝의 이해



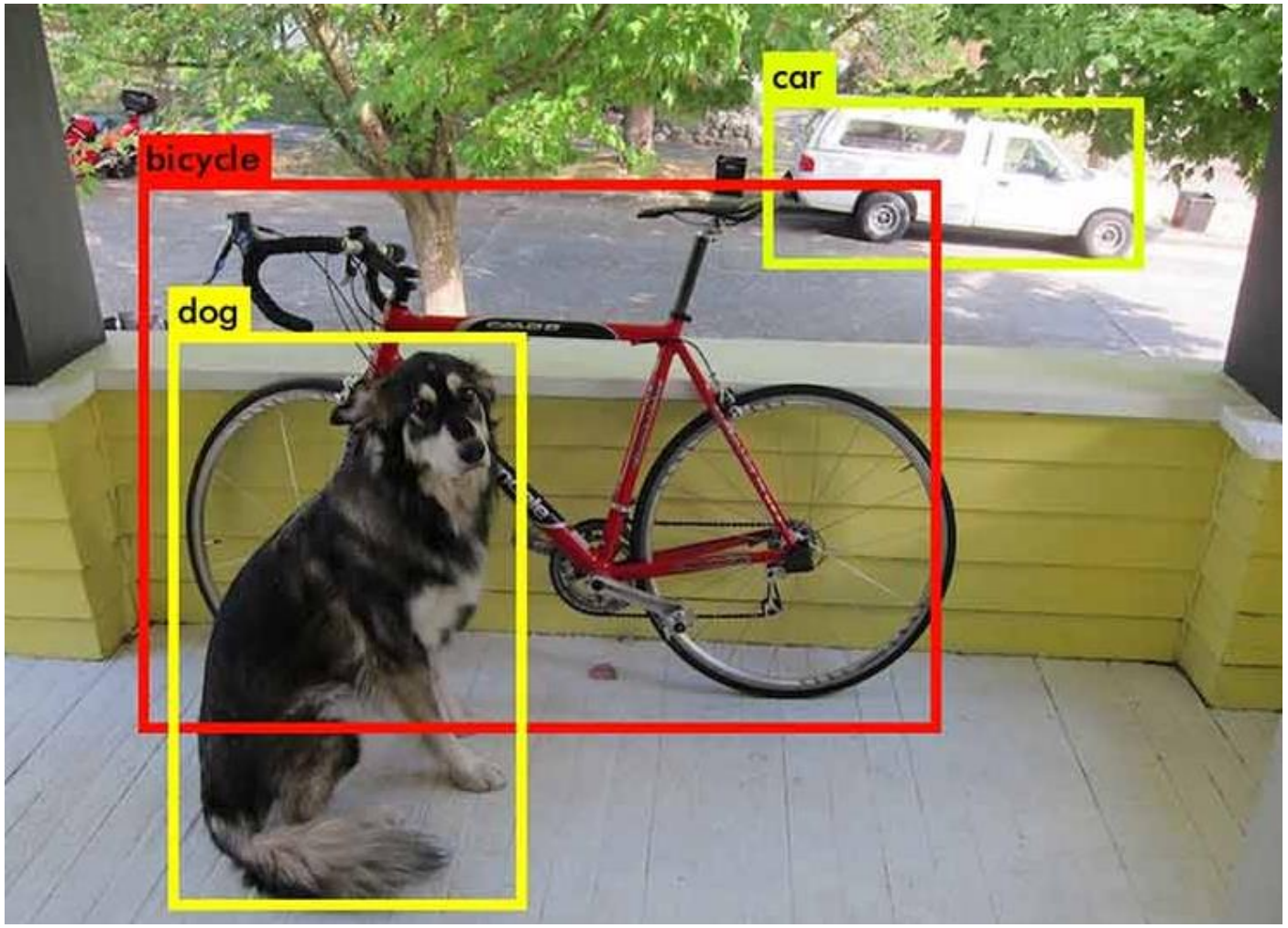

Nearest Neighbors
Linear SVM
RBF SVM
.82
.75
.88
Nearest Neighbors
Linear SVM
RBF SVM
.78
.45
.90
Nearest Neighbors
Linear SVM
RBF SVM
.97
.88
.97

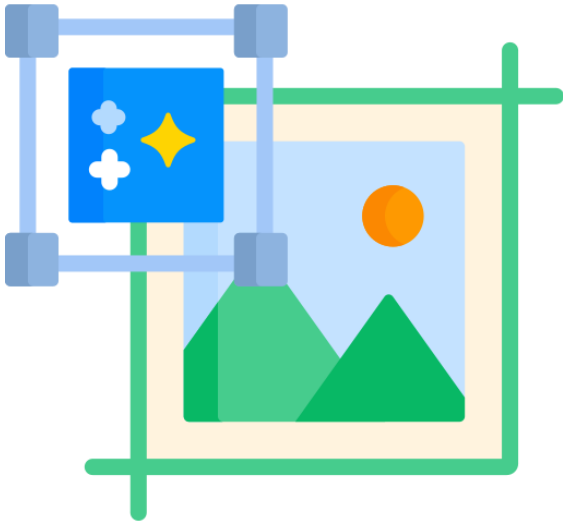


Computer Vision

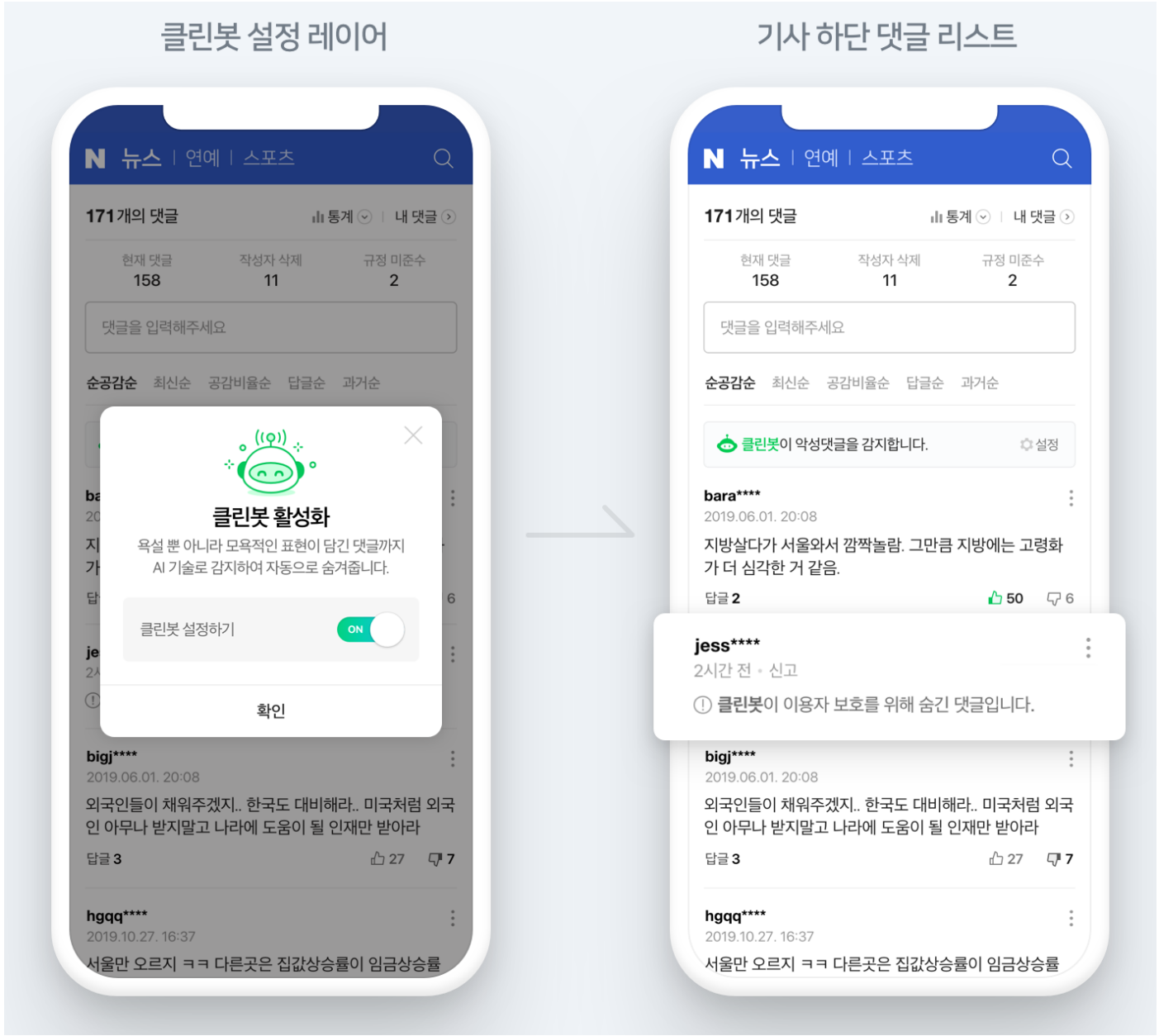
Natural Language Processing

Signal Processing

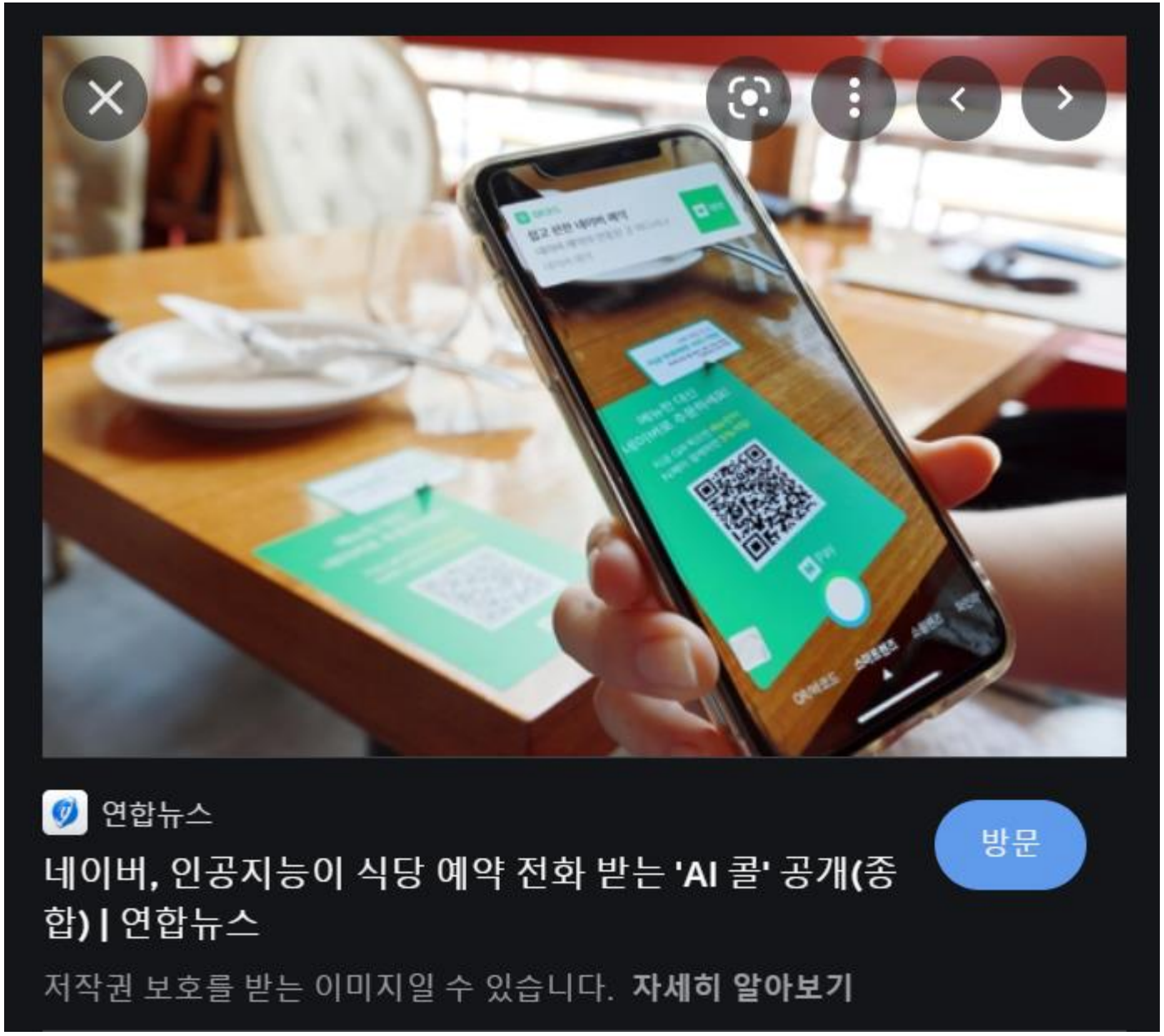




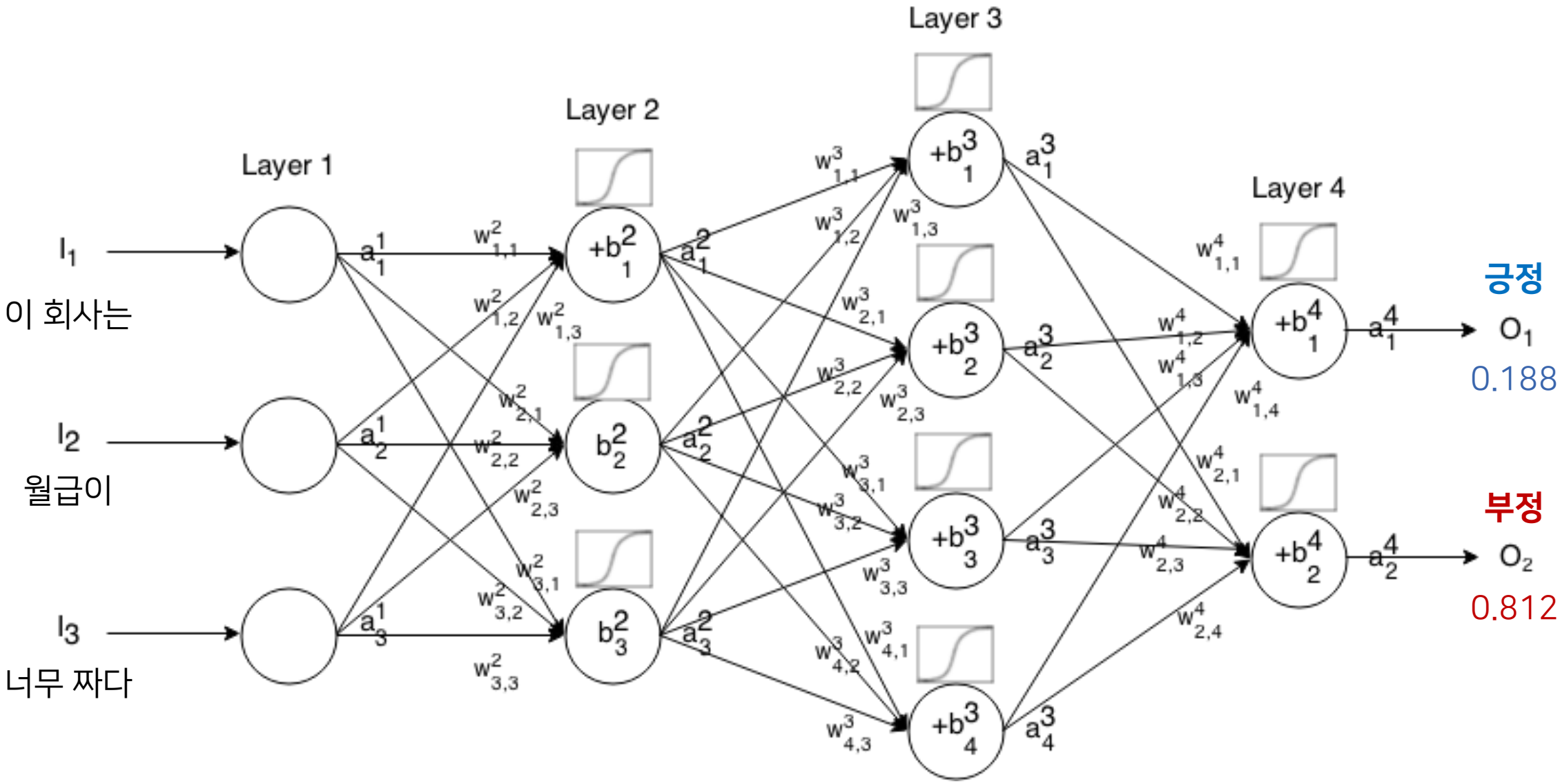
Computer Vision



Natural Language Processing



Signal Processing


Layer 1
Layer 2
Layer 3
Layer 4
I1
이 회사는
I2
월급이
I3
너무 짜다
긍정
O1
0.188
부정
O2
0.812

Underscore

연구·분석 예시

왜 학생들은 교실 문이 닫히자 학원과 온라인 커뮤니티로 떠났을까?

- 코로나 시대의 교육 불평등 -

강태영 (M.S. Management Engineering, KAIST)

강현제 (Ph.D. Candidate in Economics, Stony Brook University)

김선함 (Ph.D. Candidate in Economics, Purdue University)

초 록

코로나19의 대유행은 다양한 사회경제적 효과를 야기했으며, 전례 없는 광범위한 휴교령 역시 그 중 하나이다. 기존 감염병 대응과 달리, 이번 코로나19 대응은 상반기 전국 등교 연기와 온라인 개학, 원격 수업, 수업일수 축소를 포함했기 때문이다. 본 연구는 이러한 공교육 부재·축소에 따른 변화를 다음 세 가지 방식으로 분석했다.

첫 번째, 전국단위 자료인 사교육비조사와 한국노동패널, 교육부 제공 등교일수 자료를 결합하여 횡단면 및 패널 회귀분석(reduced form analysis)으로 등교일수 손실이 사교육 패턴에 미친 영향을 탐색했다. 두 번째, 이를 통해 얻은 추정치를 기반으로 구조모형인 토너먼트 모형(tournament model)을 추정하고 시뮬레이션을 진행했다. 그 과정에서 식별한 모수를 기초로 등교일수 손실이 현실에 비해 적었을 경우, 저소득층에게 교육 바우처 또는 기본소득이 주어졌을 경우의 학업성취도를 반사실 시뮬레이션(counterfactual simulation)하여 최종적으로 등교일수 손실이 야기한 교육불평등 및 정책대응에 관한 함의를 도출했다. 마지막으로 세 번째, 텍스트 데이터에 대한 딥러닝 모델인 자연어처리(natural language processing) 알고리즘을 활용해 국내 주요 온라인 수험생 커뮤니티 데이터를 수집 후 정성적 분석을 진행했다.

경제학 모형을 분석한 결과, 등교일수가 10일 감소할 때 초등학생 사교육지출은 5.3% 하락했고 (노동패널), 반대로 고등학생 사교육지출은 5.6% 증가했다 (경기종단). 이를 근거로 시행한 구조모형 시뮬레이션에 의하면 2020년 등교일수 감소가 학업성취도 불평등을 확대시켰다. 공교육 부재 시 사교육이 대체재로 기능할 수 있으나 이는 가계소득의 영향을 받으며, 따라서 소득불평등이 교육불평등으로 전이될 가능성이 크다.

디시인사이드의 주요 수험생 갤러리로부터 대규모 온라인 데이터를 직접 수집해 분석한 결과는 다음과 같다. 첫 번째, 코로나 확진자 수의 증가는 수험생 커뮤니티의 활성도를 높였고, 두 번째, 수험생 커뮤니티 게시물들의 공격성 역시 높아졌으며 마지막으로 세 번째, 입시 관련 논의를 다룬 주제들이 감소하며 커뮤니티 내부의 동학을 강화하는 주제들이 등장하는 비율이 더 증가한 것으로 나타났다.

All Play and No School Makes Jack a Dull Boy

- School Closure during the Pandemic and Excessive Online Community Usage -

Anonymous ACL submission

Abstract

After the COVID-19 pandemic, the South Korean government decided the nationwide school closure. As it is an unprecedented policy decision and the Korean economy was not hugely affected by the pandemic, it enabled accurate estimation of the spillover effect of the school attendance. Based on the 14.3K online community posts data, we analyzed that the students wrote more postings and became more aggressive. Also, the proportion of the topics dealing with the internal dynamics of the online communities increased after the pandemic. By combining econometric research design approach, text classification model, and document clustering algorithm, we suggest that the pandemic might have negatively influenced the students who are less focused on the face-to-face education.

1 Introduction

The various policy level countermeasures during the COVID-19 pandemic and their spillover effects have been major concerns of social science research.

In the case of South Korea, the nationwide school closure was especially an unprecedented decision. Since the Korean government hasn't been adopting the national level cessation of school during previous pandemics including MERS in 2015, H1N1 in 2009, SARS in 2004, and cholera in 1969, we can concretely conceptualize this educational decision as school closure, instead of abstractly defining it as public health crisis. Also, as the pandemic's damage to the Korean economy was relatively low, it offers the opportunity of more accurate estimation on the effect of the school attendance, compared to the other countries.

While previous research focused on the macro-economic and educational effect of the school closure (Agostinelli et al, 2021; Aum et al. 2021a, b), this research focuses on analyzing the micro-level behavior of students via online community data. Due to the competitive nature of the Korean education system, the university entrance exam is always a trending topic for teenagers which even led to the growth of Korean CSAT (College Scholastic Ability Test) related online communities.

Since the teenage users, especially 12th graders are actively using these communities for knowledge sharing, chatting, and sometimes for seeking *power competition* with other competing communities, analyzing the posting pattern before and after the pandemic would help us to understand the behavioral change after the school closure.

By combining the econometric research design approach, supervised classification model, and the unsupervised document clustering algorithm, we will investigate the school closure's impact on the students' online behavior.

2 Research Objectives

Based on the text data from the largest online community in South Korea, our research focuses on answering following three questions.

First, did COVID-19 increase the posting activity?

Second, was the level of community posts' profanity exacerbated after the pandemic?

Third, how did the topic distribution of community posts change after the pandemic?

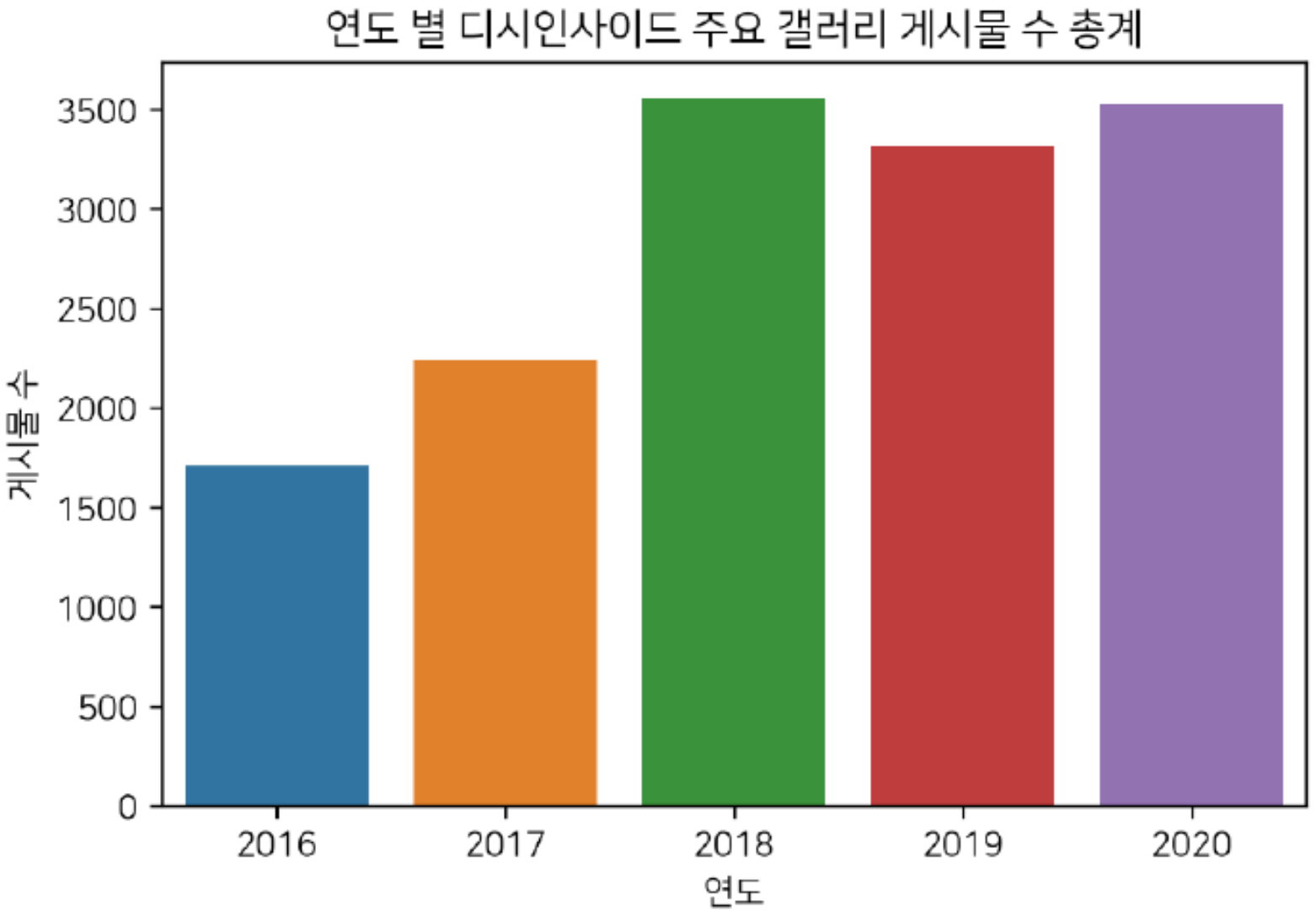
To answer the second question, we need a model that can distinguish the malice of given text. The third question with the qualitative understanding of the text would require both document clustering and statistical pre-post comparison.

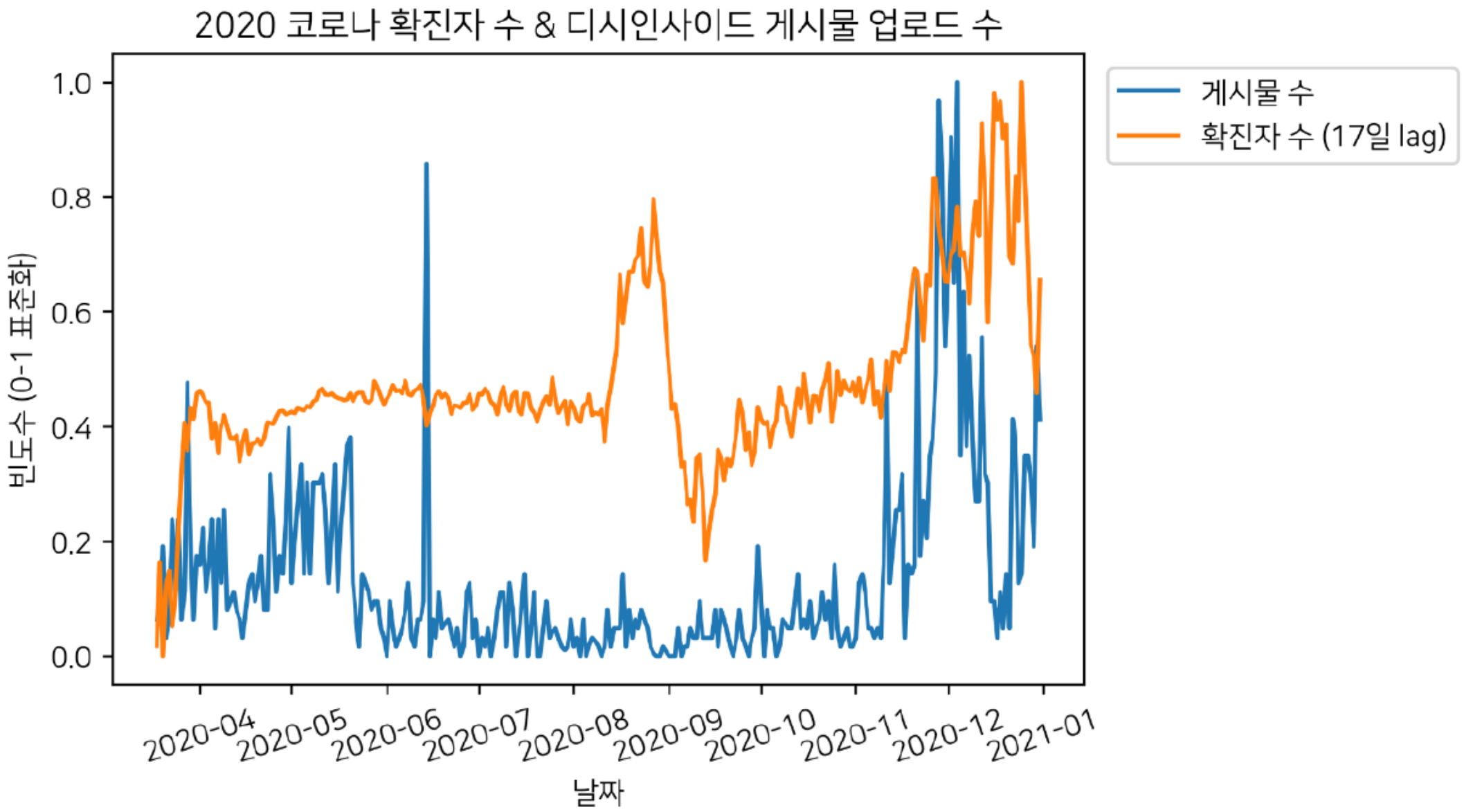
3 Data and Methods

3.1 Data

We collected 14.3K postings from *DCInside*, the largest online community in South Korea.

<그림 7> 디시인사이드 게시물 수 통계




2020 코로나 확진자 수 & 디시인사이드 게시물 업로드 수
게시물 수
확진자 수 (17일 lag)
빈도수 (0-1 표준화)
1.0
0.8
0.6
0.4
0.2
0.0
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
날짜

<표 14> 확진자 수 - 게시물 수 VAR 모형

확진자 수 시차 변수 (lagged variable)	**2019 lag12 model**	2019 lag17 model	2020 lag12 model	**2020 lag17 model**
8	0.002	-0.001	0.011	0.021
9	-0.117***	-0.022**	-0.012	-0.017
10	0.001	0.002	0.029**	0.023**
11	-0.012	-0.014	0.000	-0.001
12	0.010	0.005	0.006	0.005
13		-0.000		-0.018
14		-0.008		-0.010
15		0.008		-0.012
16		-0.002		-0.002
17		0.001		0.028**

(***p<0.01, **p<0.05, *p<0.1)