

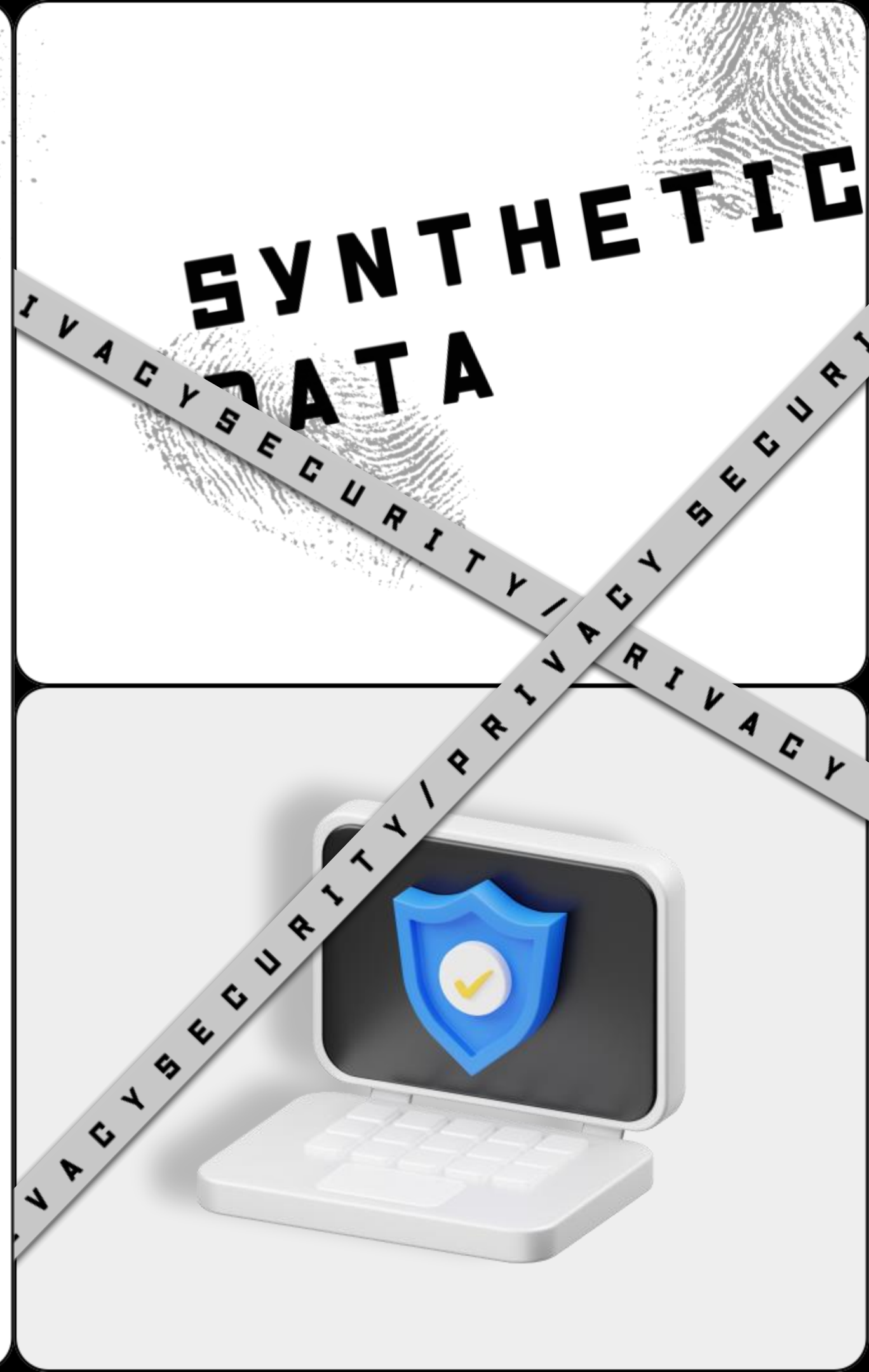
◆ 데이터사이언스 캡스톤디자인 최종발표

# 복지분야 재현데이터 생성기 개발

Synthesizing tabular data

Peakers

통계학과 2018110493 정정룡  
통계학과 2018110497 김평진  
통계학과 2018110498 황영우





# 목차

table of Contents

**01 프로젝트 주제 및 배경**

**02 재현데이터 생성기법**

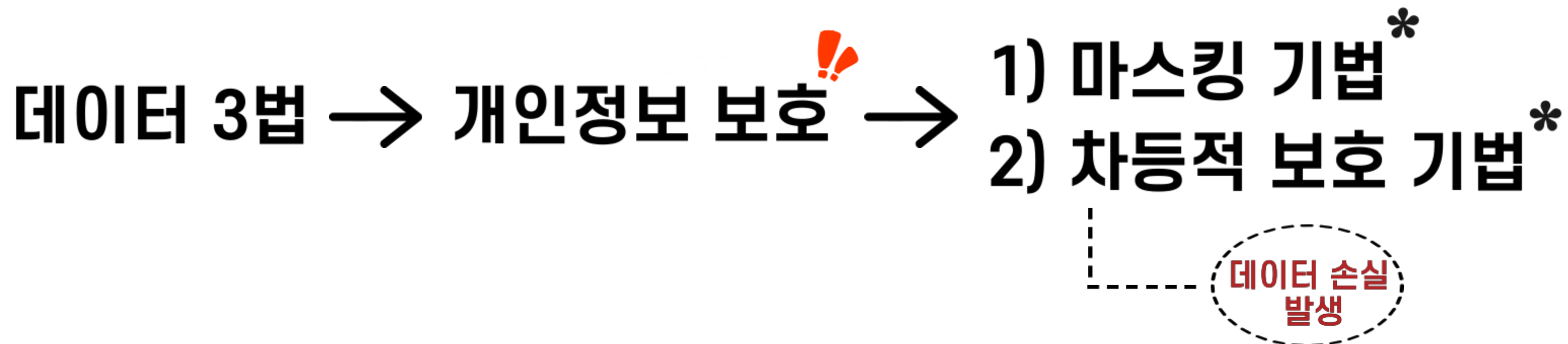
**03 검증 방법**

**04 프로젝트 결과 및 결론**

SECURITY / SYNTHESIZING TABULAR DATA

## 프로젝트 배경

과거



현재

∴ 재현데이터

- 개인정보 관련 규제로부터 자유로움
- 원 데이터와 흡사한 수준의 효용성

\* 마스킹기법: 마스킹은 원자료의 적절한 변환을 통해 민감한 정보를 가리는 기법을 말하며, 마스킹된 자료란 변환된 자료와 변환에 관한 모든 정보를 의미

\* 차등적 보호기법: 데이터에 포함된 개인정보를 보호하기 위해 해당 데이터세트에 임의의 노이즈를 삽입함으로써 개인정보가 제3자에게 노출되지 않도록 보호하는 기법

프로젝트 주제  
및 목표

# 복지분야 재현데이터 생성기 개발

AS-IS

민감한 개인정보가 포함된 복지분야의 특성으로 인해  
제한적인 활용

TO-BE

원자료와 매우 유사한 모의 데이터(Simulated Data)를  
생성하여 개인정보보호로 연구 및 교육 분야에 적극 활  
용 가능

데이터 소개

한국복지패널조사(2022) - 한국보건사회연구원

필요 변수 선정 (raw data)

조사표 유형	조사 영역
가구용 (원·신규)	I. 가구 일반 사항
	II. 건강 및 의료 A
	III. 경제활동 상태
	IV. 사회보험, 퇴직(연)금, 개인연금 가입
	V. 재산
	VI. 생활 여건
	.
	.
	.
가구원용 (원·신규)	I. 사회보험, 퇴직금, 개인연금 수금
	II. 근로
	III. 생활실태·만족 및 의식
	.
	.
	.
복지 인식 부가조사	I. 전반적인 사회적·정치적 인식과 태도
	II. 복지 자원 및 대상 범위
	III. 정치 참여와 성향



	저소득층 여부	나이	교육수준	혼인상태	가구형태	주택유형	주택점유 형태	총생활비	총소득	가구 서비스	노인 서비스	아동 서비스
0	저소득층 가구	78	중등	사별	단독	다가구용 단독주택	자가	3300	0	이용	이용	해당없음
1	저소득층 가구	75	초등	사별	단독	다가구용 단독주택	보증부월세	2868	2064	이용	이용	해당없음
							■					
							■					
							■					
7864	일반가구	59	대학	유배우	기타	일반 아파트	전세	6168	7880	미이용	해당없음	해당없음

## 통계 기반 모델



## 딥러닝 기반 모델

Synthpop package



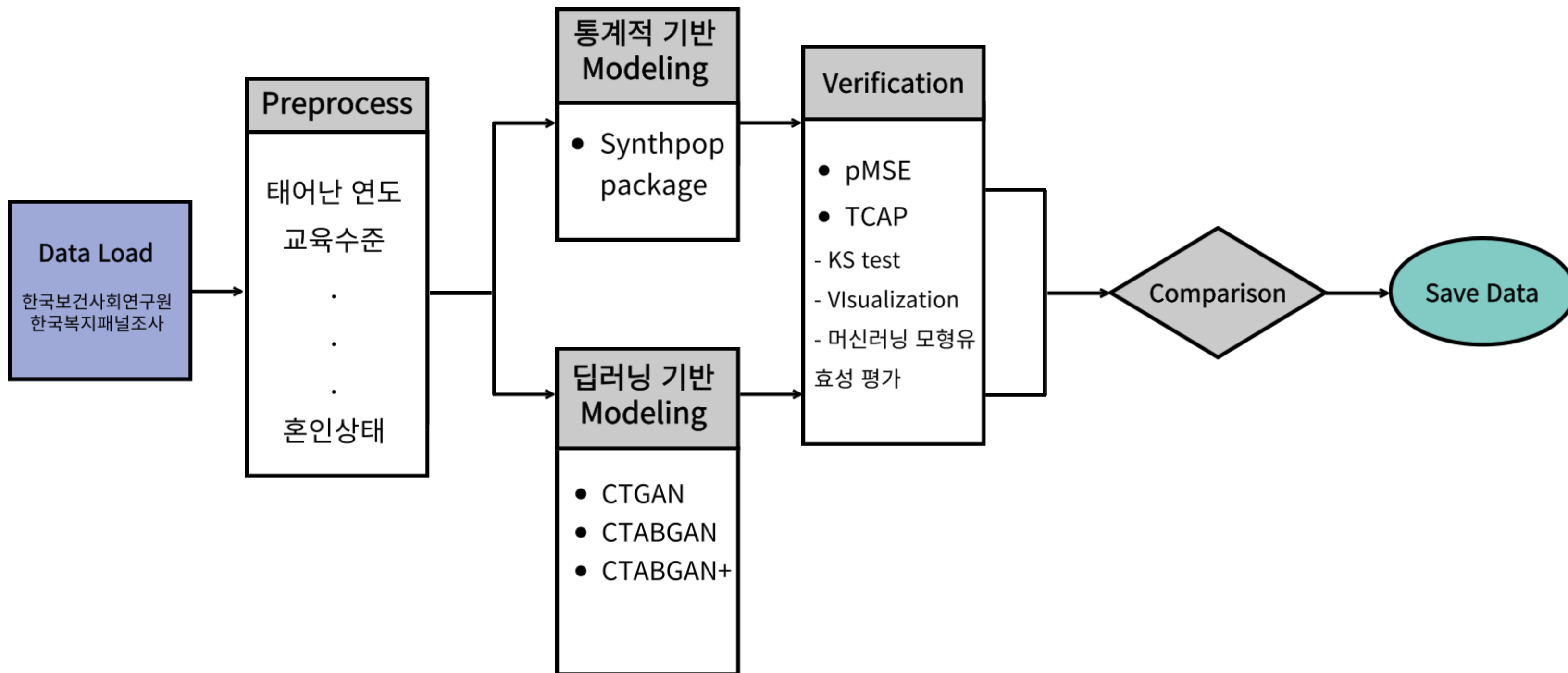
- 
- 비모수적 다중대체 모형

GAN model

- 
- CTGAN
  - CTABGAN
  - CTABGAN+



## Flow chart



# Synthpop

순차적 조건부  
확률분포

비모수적  
다중대체 모형

**CART** \*  
algorithm

~~개별 말단 노드의 반응변수 평균~~

Bayesian bootstrap 사용

\* CART(Classification And Regression Tree) : 주어진 여러 설명 변수에 기반하여 분할 규칙에 따라 의사결정나무를 만들고 반응 변수의 값을 예측하는 비모수적 방법



## GAN\*

## Model



Generator



Discriminator

synthetic data :  $G(z)$ 

Real data (Target)

## Training

NOISE( $z$ )Generator ( $G$ )Fake Money  
 $G(z)$ Discriminator  
( $D$ )

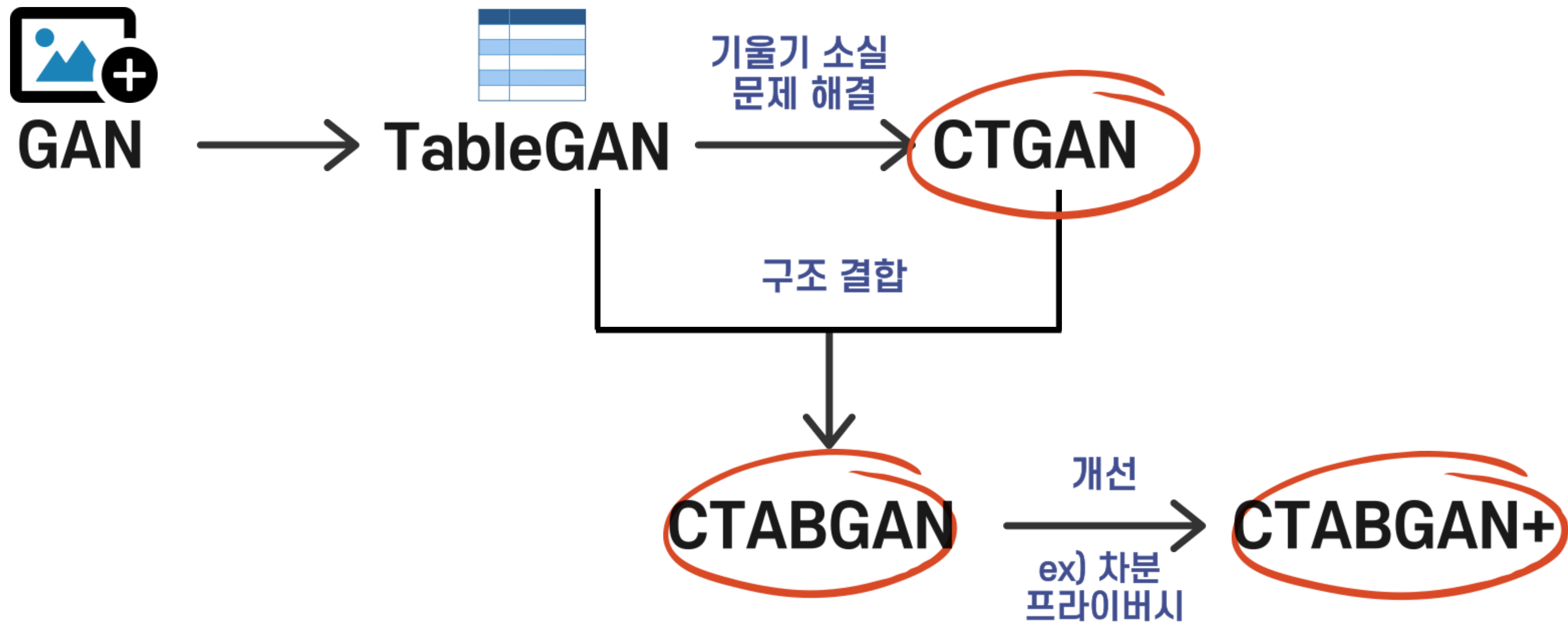
Fake

Real

Real Money  
 $x$  $D(G(x)) : 0$  $D(x) : 0$ 

\* GAN(Generative Adversarial Network) : 생성적 적대 신경망 모형은 임의의 난수를 추출하여 원 데이터와 유사하도록 생성하는 모형으로 생성자와 판별자로 구성되어 있는 인공지능 모형

GAN 모델  
in table



→ : 발전 방향

## 사용 GAN모델

## CTGAN

- 연속형 데이터에 정규 혼합 분포 추정에 기반한 최빈값 기준 정규화
- 범주형 데이터에 조건부 벡터 도입하여 생성자 설계
- 기울기 소실 문제 및 범주형 데이터에 대한 취약성 해결

## CTABGAN

- CTGAN 모형의 구조 모두 적용
- TableGAN의 정보 손실 함수 및 분류 손실 함수 추가 반영

TableGAN + CTGAN = CTABGAN

## CTABGAN+

- 학습의 안정화
- 차분 프라이버시 도입
- DP-SGD기법 적용하여 노출 위험성에 효과적으로 대처
- CTABGAN을 개선한 생성 기법

Synthpop,  
CTGAN  
재현데이터  
생성

Synthpop

파라미터

- 원 저자의 기본 설정
- 주어진 변수 순서에 따라 순차적으로 학습 후 재현데이터 생성

	저소득층 여부	나이				가구 서비스	노인 서비스	아동 서비스
0	저소득층 가구	82				이용	이용	해당없음
1	저소득층 가구	72				이용	이용	해당없음
7864	일반가구	55				이용	해당없음	이용

CTGAN

파라미터

- 기본 설정
- epoch = 150
- batch size : 100

	저소득층 여부	나이				가구 서비스	노인 서비스	아동 서비스
0	저소득층 가구	70				미이용	이용	해당없음
1	저소득층 가구	81				이용	이용	해당없음
7864	일반가구	65				이용	이용	이용

CTABGAN,  
CTABGAN+  
재현데이터  
생성

CTABGAN

파라미터

- 기본 설정
- epoch = 150
- batch size : 100

	저소득층 여부	나이				가구 서비스	노인 서비스	아동 서비스
0	저소득층 가구	39				이용	이용	해당없음
1	저소득층 가구	56				이용	이용	해당없음
7864	일반가구	51				이용	해당없음	해당없음

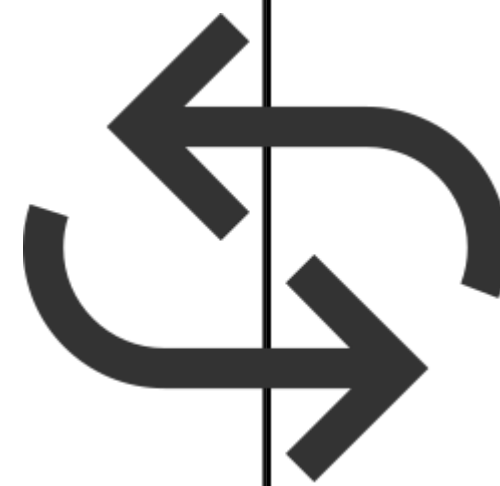
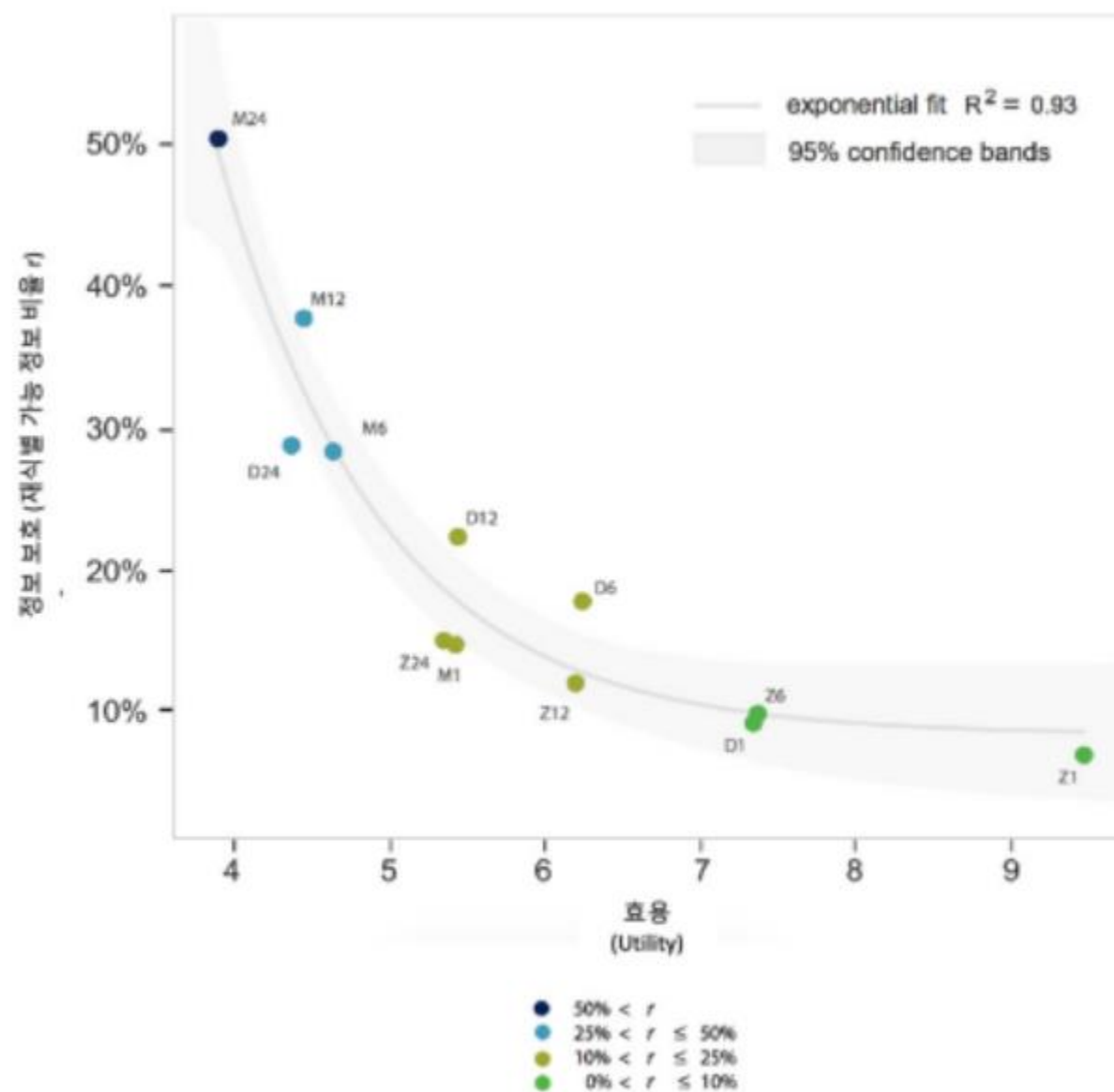
CTABGAN+

파라미터

- 기본 설정
- epoch = 150
- batch size : 100

	저소득층 여부	나이				가구 서비스	노인 서비스	아동 서비스
0	일반가구	64				이용	이용	이용
1	일반가구	61				이용	이용	해당없음
7864	일반가구	44				이용	해당없음	이용

## 검증 방법

노출 위험도 **TCAP**데이터 효용성 **pMSE**

노출 위험도와  
데이터 효용성  
은 상충관계

## + 보조지표

- 머신러닝 활용 유효성 평가
- 통계적 유사성
- 시각적 비교분석



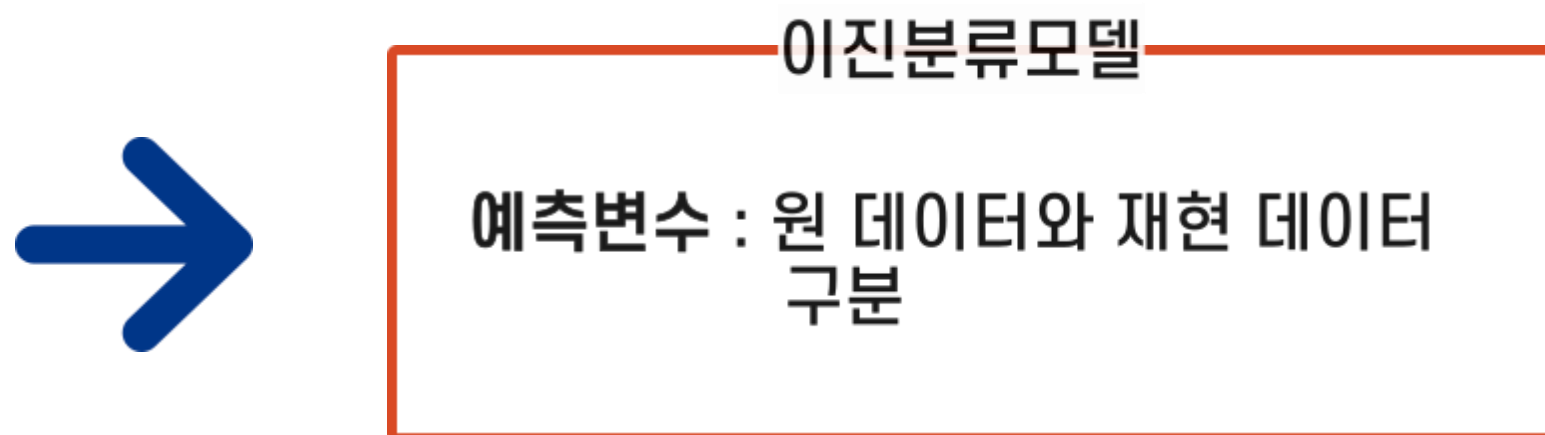
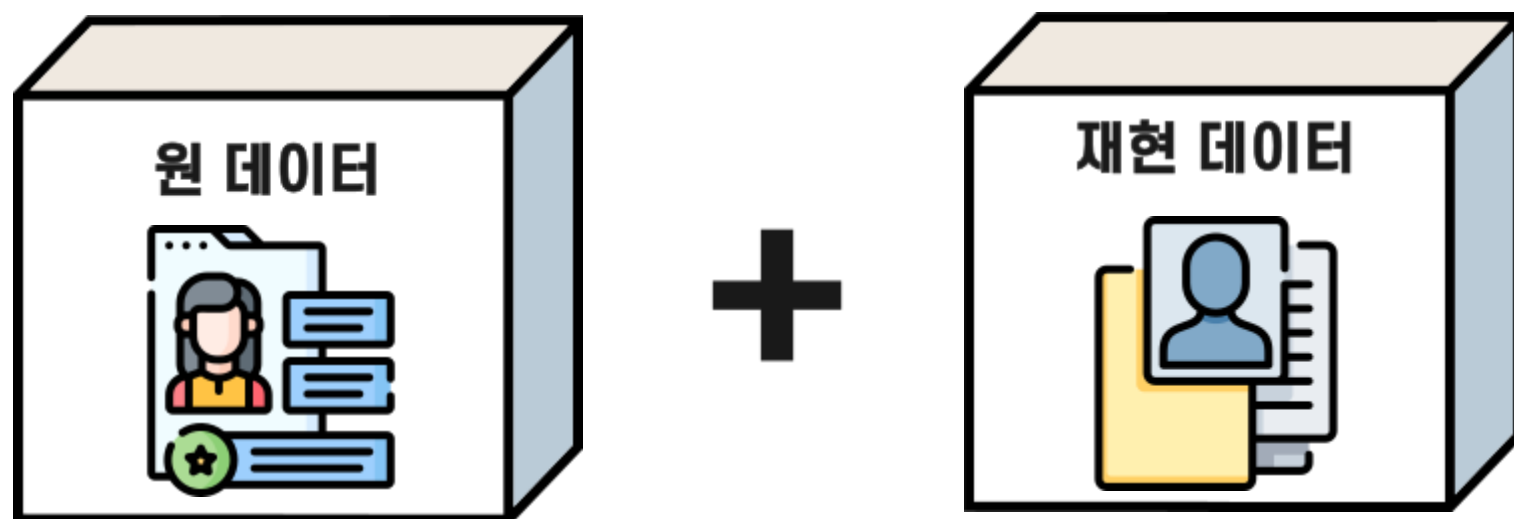
## TCAP



TCAP : 목표 변수를 식별할 확률 (0~1)


- 목표 변수 : 저소득층여부, 나이, 혼인상태, 가구형태, 총생활비
- 식별자 변수 : 교육수준, 주택유형, 주택점유형태, 가구서비스, 노인서비스, 아동서비스, 총소득

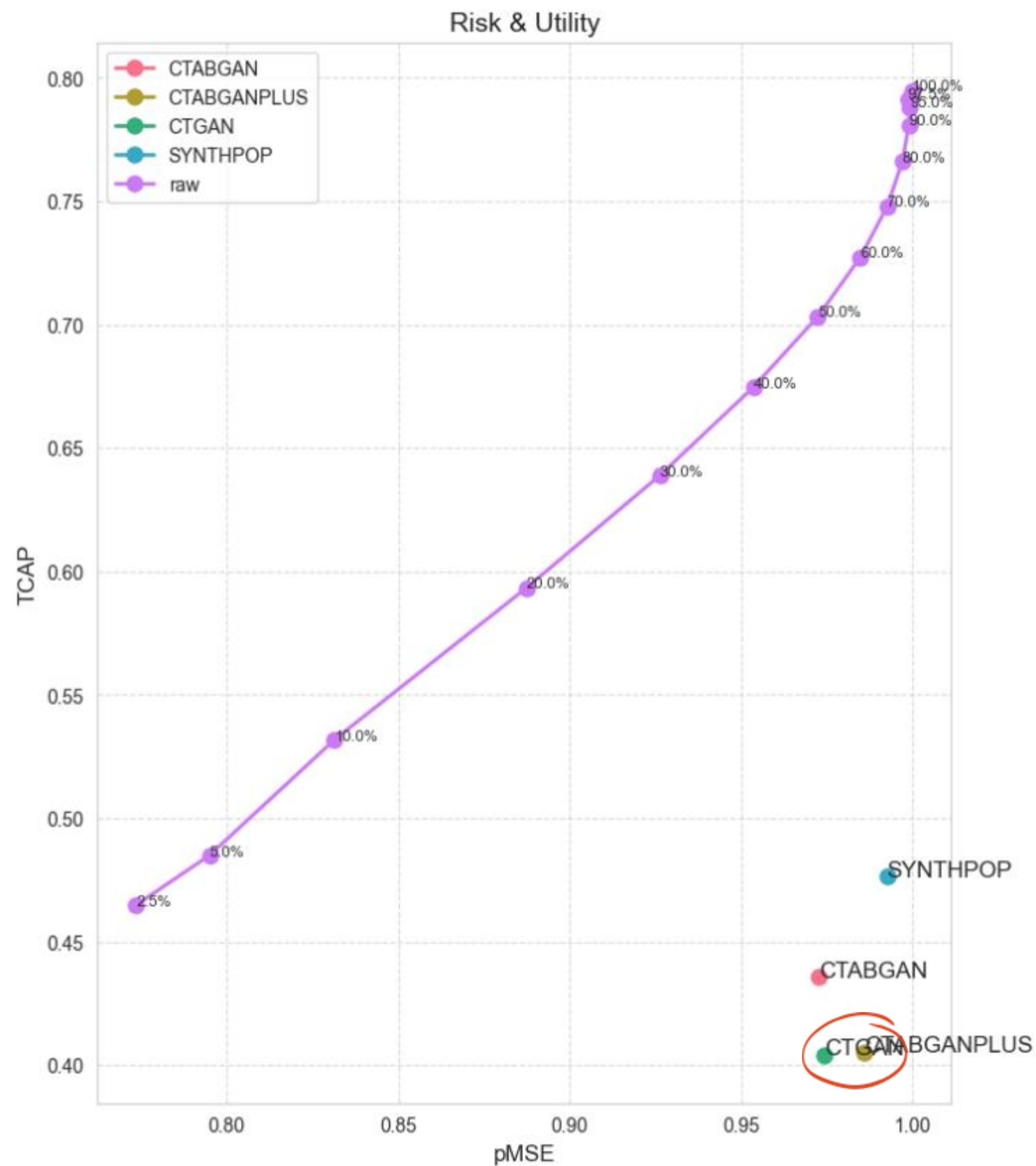
## pMSE



**pMSE** : 원 데이터와 재현 데이터의  
성향 점수를 기준으로 **유용성** 측정

$\hat{p}_i$  = 각각의 레코드마다 재현데이터라고  
판단할 확률


$$pMSE = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - 0.5)^2$$

TCAP, PMSE  
그래프

기존 데이터의 사용 비율이 높을수록  
노출 위험도와 데이터 유용성이 높음

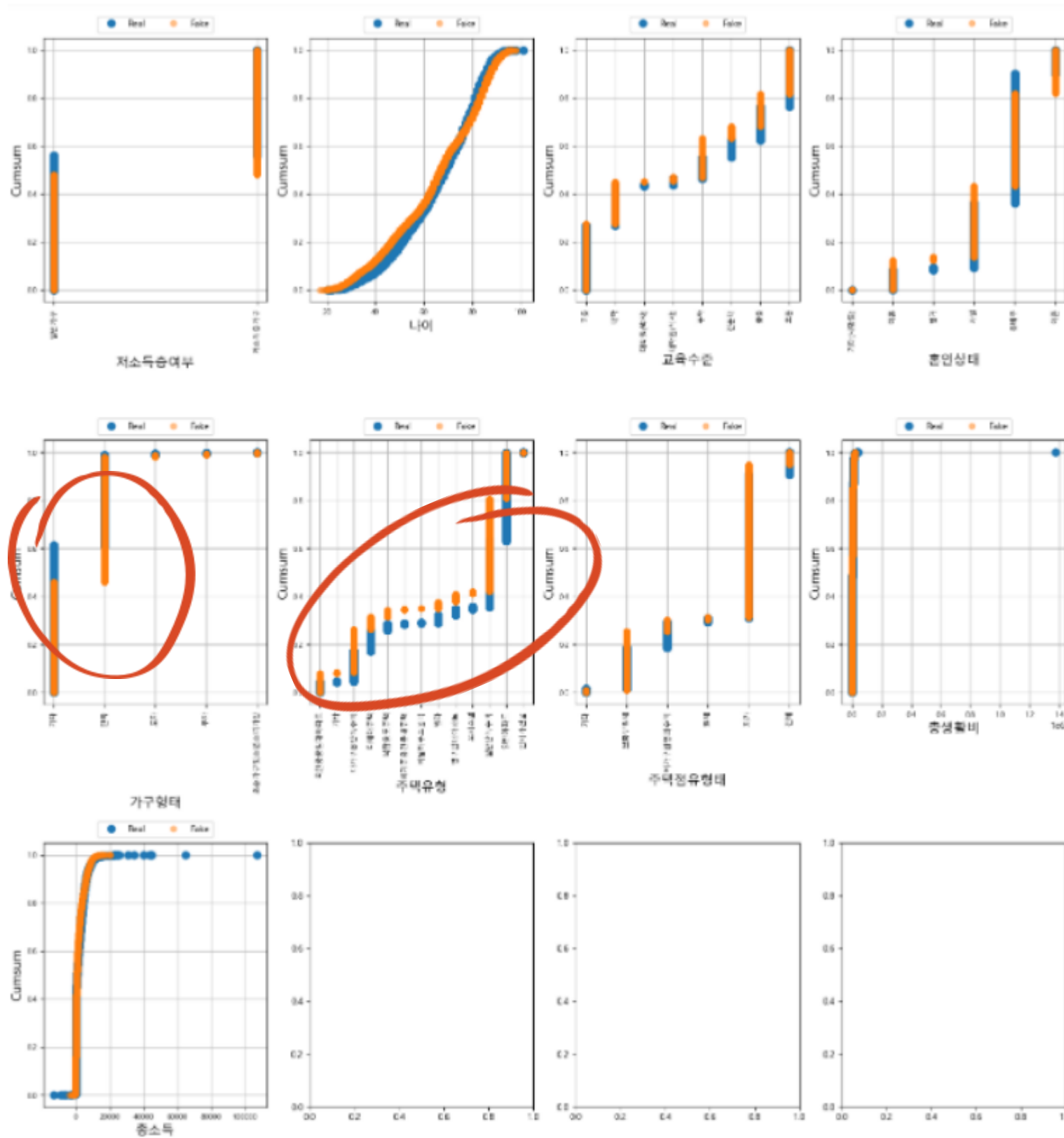
Synthpop과 CTABGAN은 유용성에 비해 노출  
위험도가 비교적 높게 나타남

CTGAN과 CTABGAN+ 모델 선정

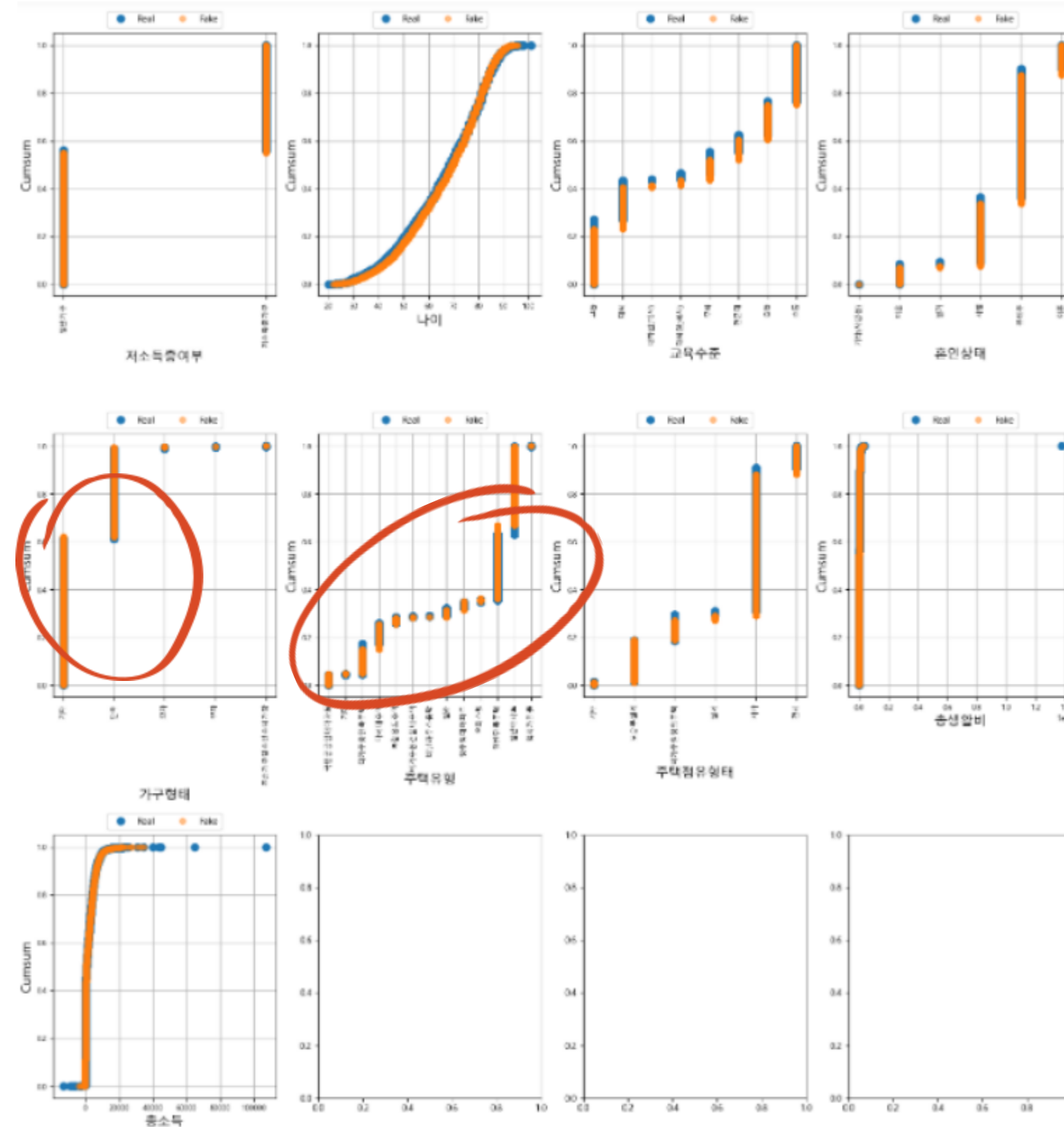
보조지표를 활용하여 최종 모델 선정

## 보조지표 - 시각적 비교분석

# CTGAN



# CTABGAN+



보조 지표

통계적 유사성

구분	CTGAN	CTABGAN+
Column Shapes(%)	87.0	94.81
Column Pair Trends(%)	67.39	75.73

**Column Shapes** : 원본 데이터와 재현 데이터의 누적 그래프가 서로 겹치는 비율

**Column Pair Trends** : 두 개의 데이터 추세선이 유사한 것을 나타내는 지표

ks test

: 누적분포함수간의 차이를 이용하여 분포의 유사함 비교

구분	CTGAN	CTABGAN+
나이	0.946	0.951
총 생활비	0.904	0.948
총 소득	0.759	0.804

보조 지표

머신러닝 유효성 평가

- 예측 변수 : 저소득층 여부
- 설명 변수 : 교육수준, 혼인상태, 가구형태, 주택유형, 주택점유형태, 가구서비스, 노인 서비스, 아동서비스
- 머신러닝 모델 : Logistic Regression, Random Forest, SVM
- 파라미터 : 기본 설정
- Train, Test 비율 : 8:2

	구분	원본데이터	CTGAN	CTABGAN+
Logistic Regression	f1 score	0.85	0.86	0.85
	accuracy	0.87	0.86	0.86
RandomForest	f1 score	0.89	0.85	0.89
	accuracy	0.90	0.85	0.90
SVM	f1 score	0.87	0.86	0.86
	accuracy	0.88	0.85	0.86



최종 모델 선정

보조 지표 결과

지표	세부구분	원본데이터	CTGAN	CTABGAN+
통계적 유사성	Column Shapes(%)		88.85	95.29
	Column Pair Trends(%)		71.78	75.3
ks test	나이		0.946	0.951
	총 생활비		0.904	0.948
	총 소득		0.759	0.804
Logistic Regression	f1 score	0.85	0.86	0.85
	accuracy	0.87	0.86	0.86
RandomForest	f1 score	0.89	0.85	0.89
	accuracy	0.90	0.85	0.90
SVM	f1 score	0.87	0.86	0.86
	accuracy	0.88	0.85	0.86
+ 총소득 음수 개수		0	131	0

최종 모델 : CTABGAN+

## 기대효과

01



재현데이터를 사용해도 원 데이터와  
거의 차이가 없는 성능의 학습 모델  
구축 가능

03



기존에 수행하기 어려웠던 다양한  
분석, 교육 등에 활용하는 방안 모  
색 가능

02



내부 이용자 외의 연구자들의 분석  
사용자 입장에서는 관련 분야와 기술  
에 대한 시각 확장 및 실력 향상

04



복지 분야 뿐만 아닌 금융, 의료 등 다  
양한 분야에서 활용 가능

# 복지분야 재현데이터 생성기 개발

team peakers

감사합니다. *Thanks for listening*



SECURITY / PRIVACY