합성데이터(Synthetic Data)를 통한 금융 AI 활성화 방안

: 생성AI를 이용한 금융빅데이터 활용 연구

한국신용정보원 기술데이터부

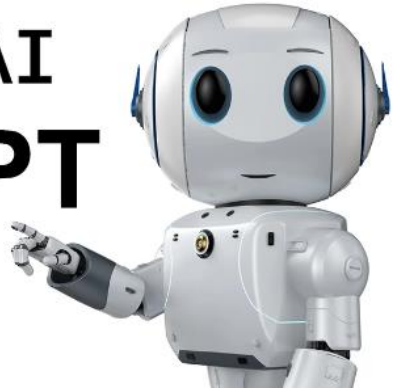
허용준 선임, Ph. D., MBA

Generative AI

(생성 AI 개요)

생성 AI(Generative AI)

OpenAI
ChatGPT

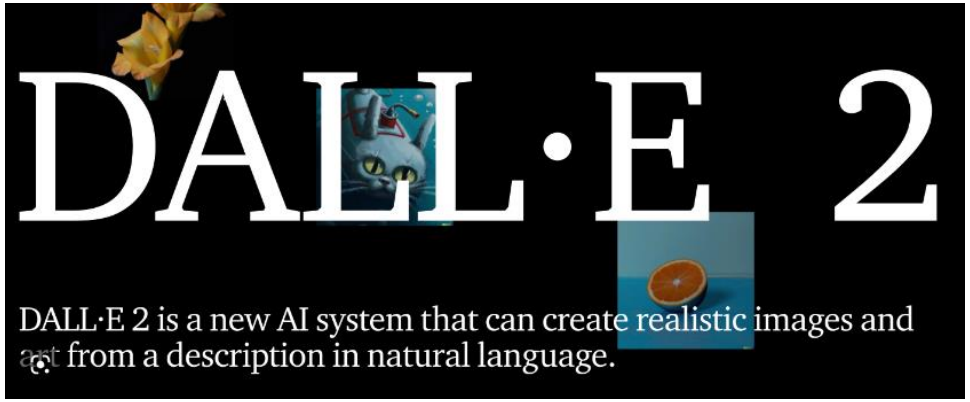


User What is unusual about this image?

Source: Barnorama

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

생성 AI(Genarative AI)

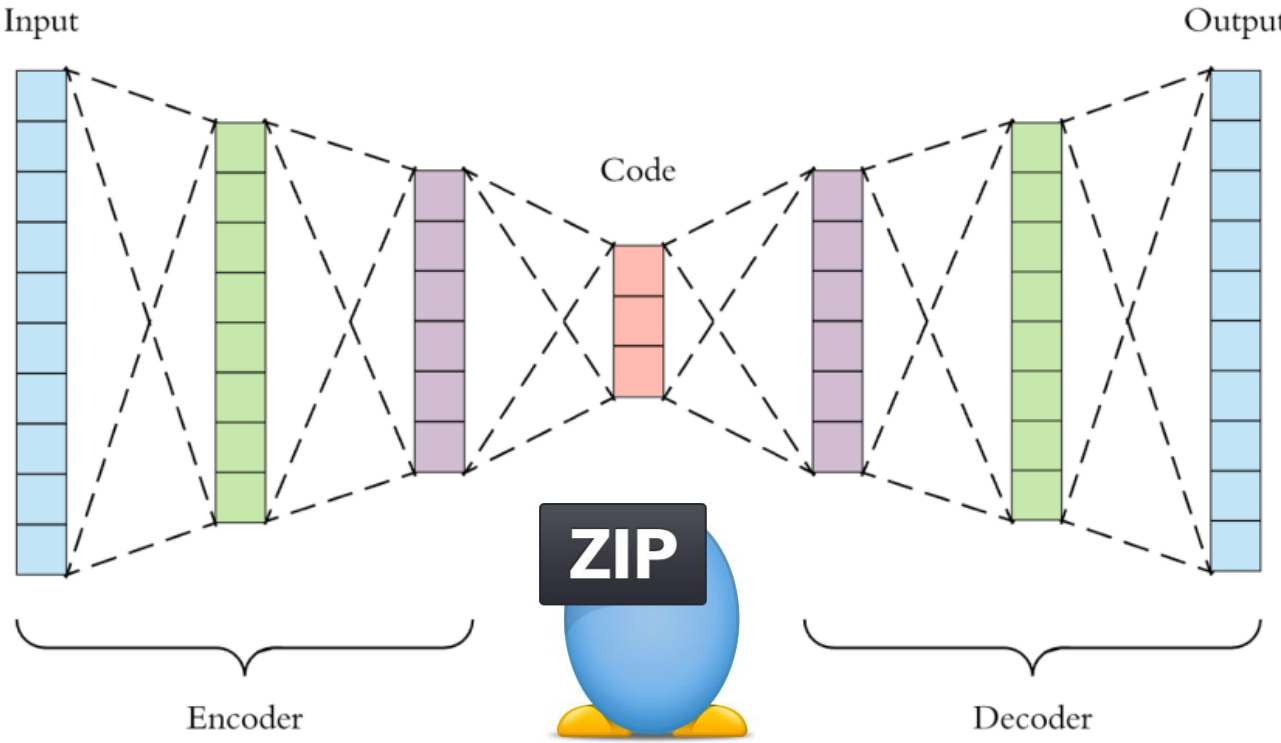


Text Prompt an armchair in the shape of an avocado. . . .

AI Generated images

생성 AI(Genarative AI)

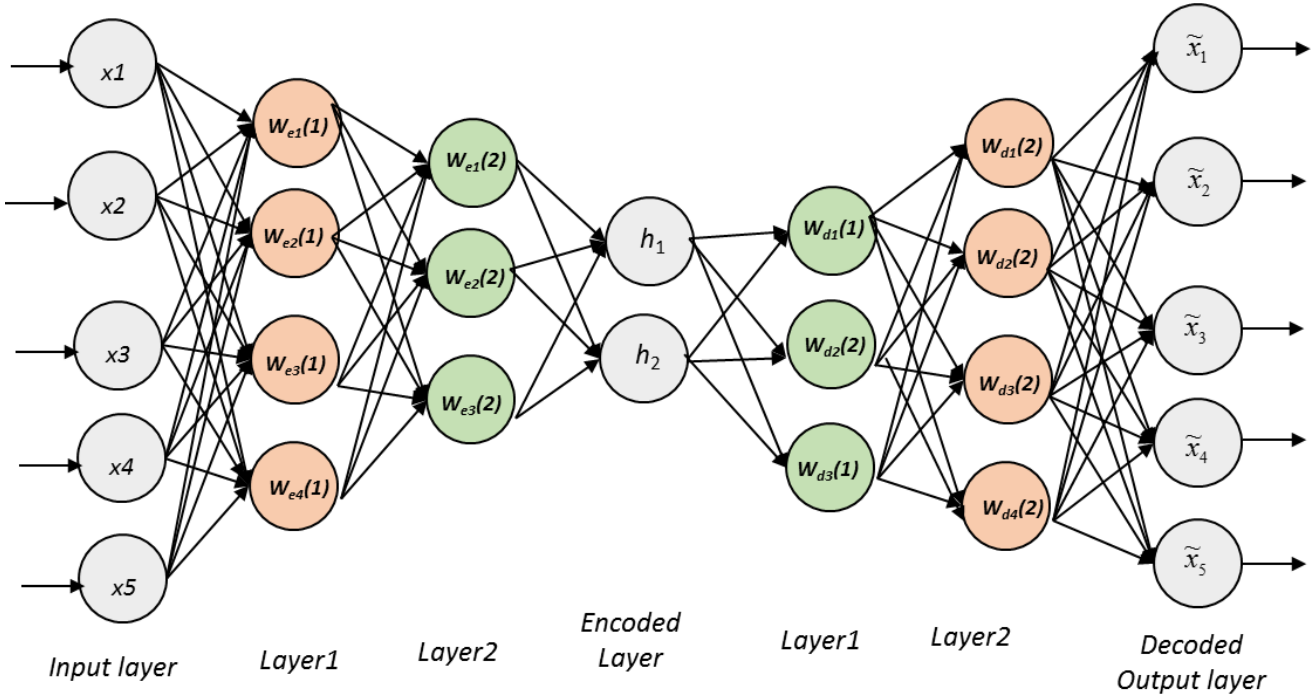
$$x \longrightarrow f(x) \rightarrow y \rightarrow f'(y) \longrightarrow x$$



생성 AI(Genarative AI)

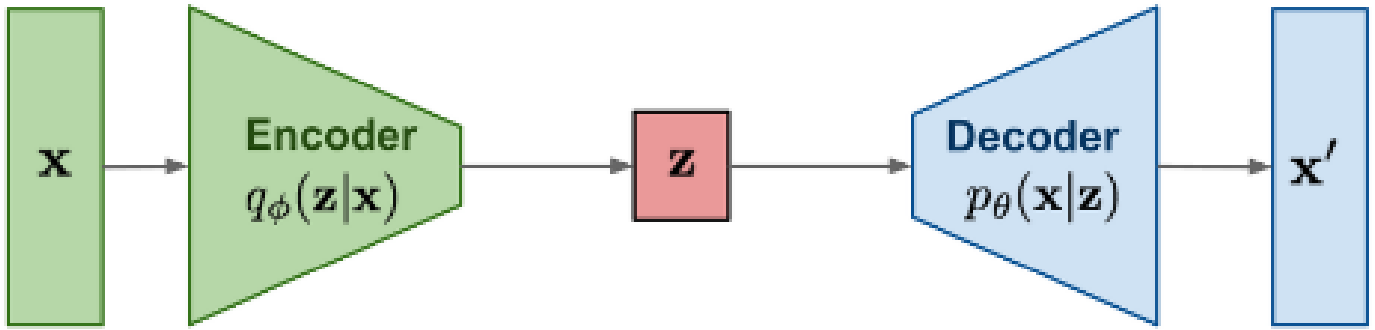
$$x \longrightarrow f(x) \rightarrow y \rightarrow g(y) \longrightarrow x'$$

Human face!

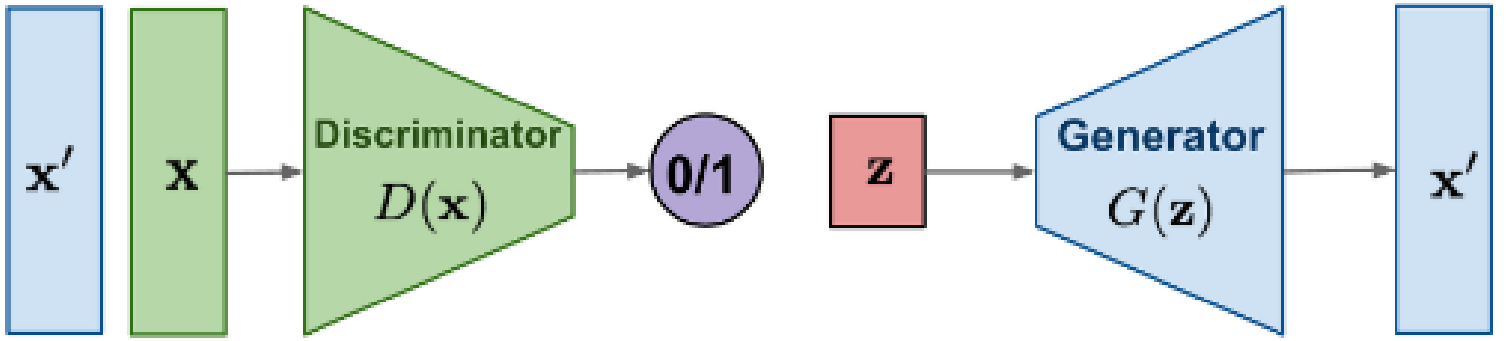


생성 AI(Genarative AI)

VAE: maximize variational lower bound



GAN: Adversarial training



GAN(Generative Adversarial Network)

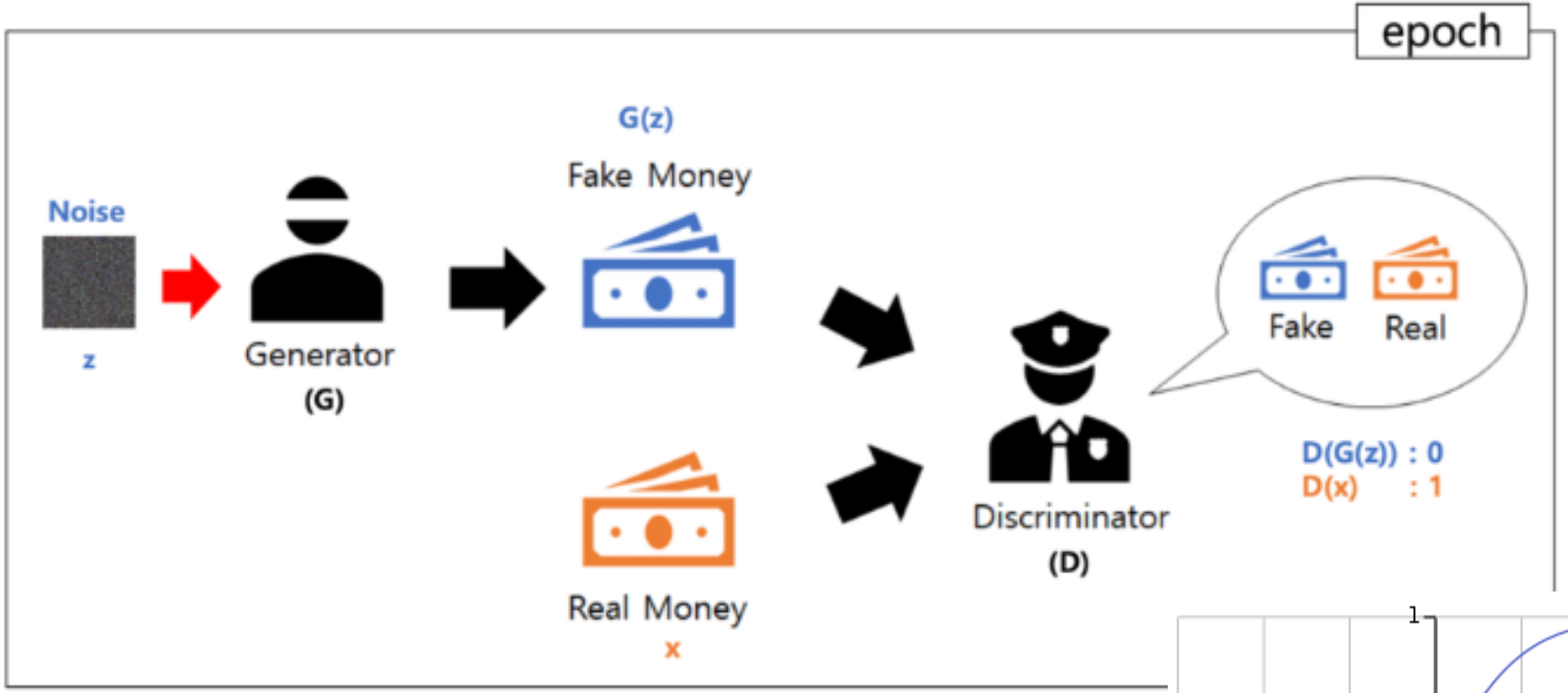
적대적 생성 신경망

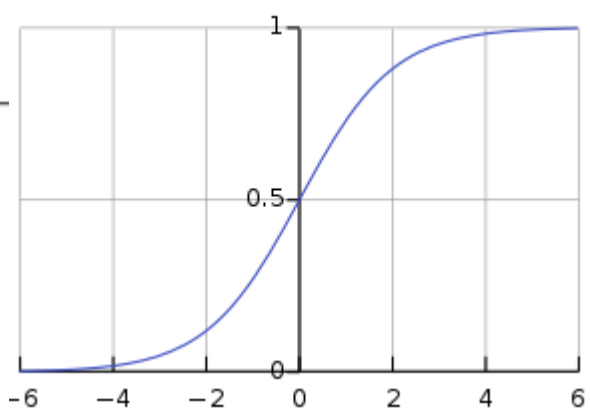
Generative Adversarial Nets

Ian J. Goodfellow,* Jean Pouget-Abadie,† Mehdi Mirza, Bing Xu, David Warde-Farley,
Sherjil Ozair,‡ Aaron Courville, Yoshua Bengio§
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

GAN(Generative Adversarial Network)

적대적 생성 신경망





금융 데이터 적용

이미지(비정형 데이터)

1000X(0~255)

1000X

RGB
3X

1000 X 1000 X 3 = 3000000 Columns

10000 X Rows

금융 데이터(정형 데이터)

100 Columns ...

Rows

...

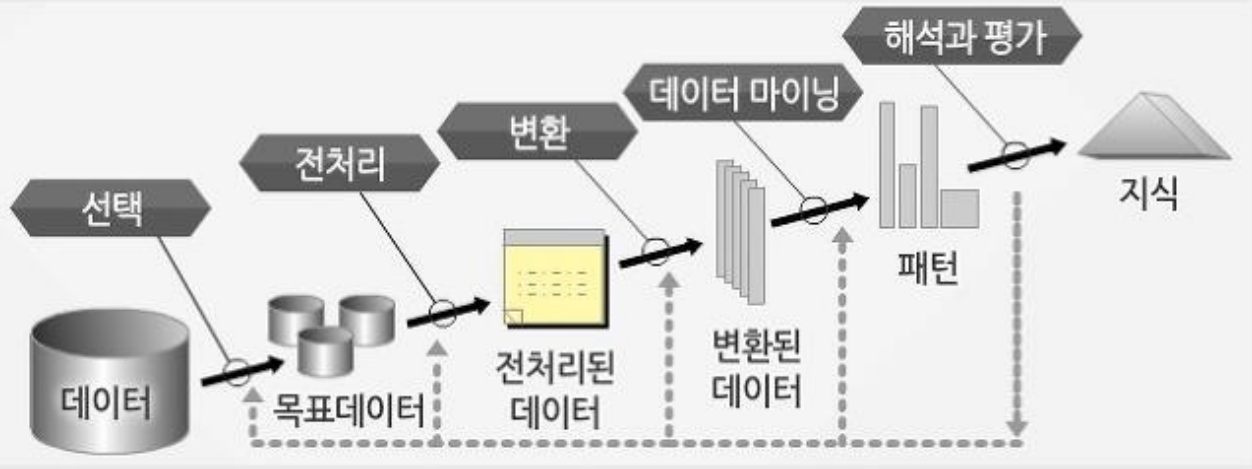
	A	B	C	D	E	F	G	H	I	J	K
1	ISSU1_NO	IT1M_531	IT1M_532	IT1M_533	IT1M_534	IT1M_535	IT1M_536	IT1M_537	IT1M_538	IT1M_539	IT1M_531(
2	7015-2022	7	10	7	7	7	9	10	9	10	5
3	7015-2022	3	2	9	4	6	1	3	10	3	5
4	7015-2022	5	7	6	6	1	6	5	3	4	3
5	7015-2022	9	10	10	6	8	4	10	10	9	10
6	7015-2022	9	10	9	9	6	7	8	3	10	9
7	7015-2022	9	10	8	6	6	7	6	3	6	7
8	7015-2022	8	10	8	5	5	1	2	1	2	4
9	7015-2022	10	10	8	2	5	7	2	1	4	3
10	7015-2022	5	5	8	6	8	1	10	1	10	6
11	7015-2022	7	10	8	1	1	1	1	1	1	7
12	7015-2022	8	10	8	8	4	6	10	8	6	8
13	7015-2022	3	1	9	2	1	1	5	1	1	4
14	7015-2022	10	10	9	6	8	4	10	4	10	7
15	7015-2022	9	10	9	6	8	6	3	1	6	6
16	7015-2022	1	1	4	2	1	1	2	1	4	1
17	7015-2022	9	10	8	4	6	8	2	1	6	3
18	7015-2022	8	10	6	4	6	8	5	4	3	4
19	7015-2022	7	9	9	4	1	1	2	9	2	3
20	7015-2022	6	6	8	4	6	10	1	1	2	3
21	7015-2022	8	9	8	4	1	1	3	1	2	4

Let's add Text data(NLP)...

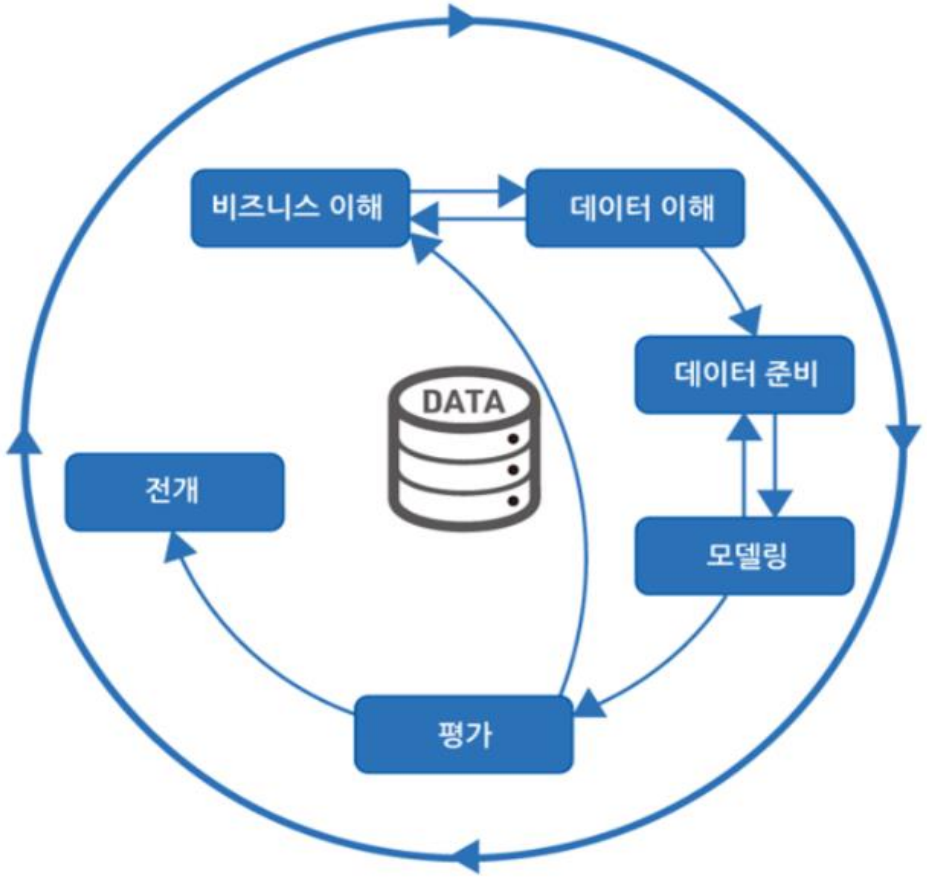
Synthetic Data

(합성, 생성, 재현 데이터)

Data Mining
(데이터 경험과 인공지능 모델 개발)



KDD(Knowledge Discovery in Databases, Farrad, 1996)

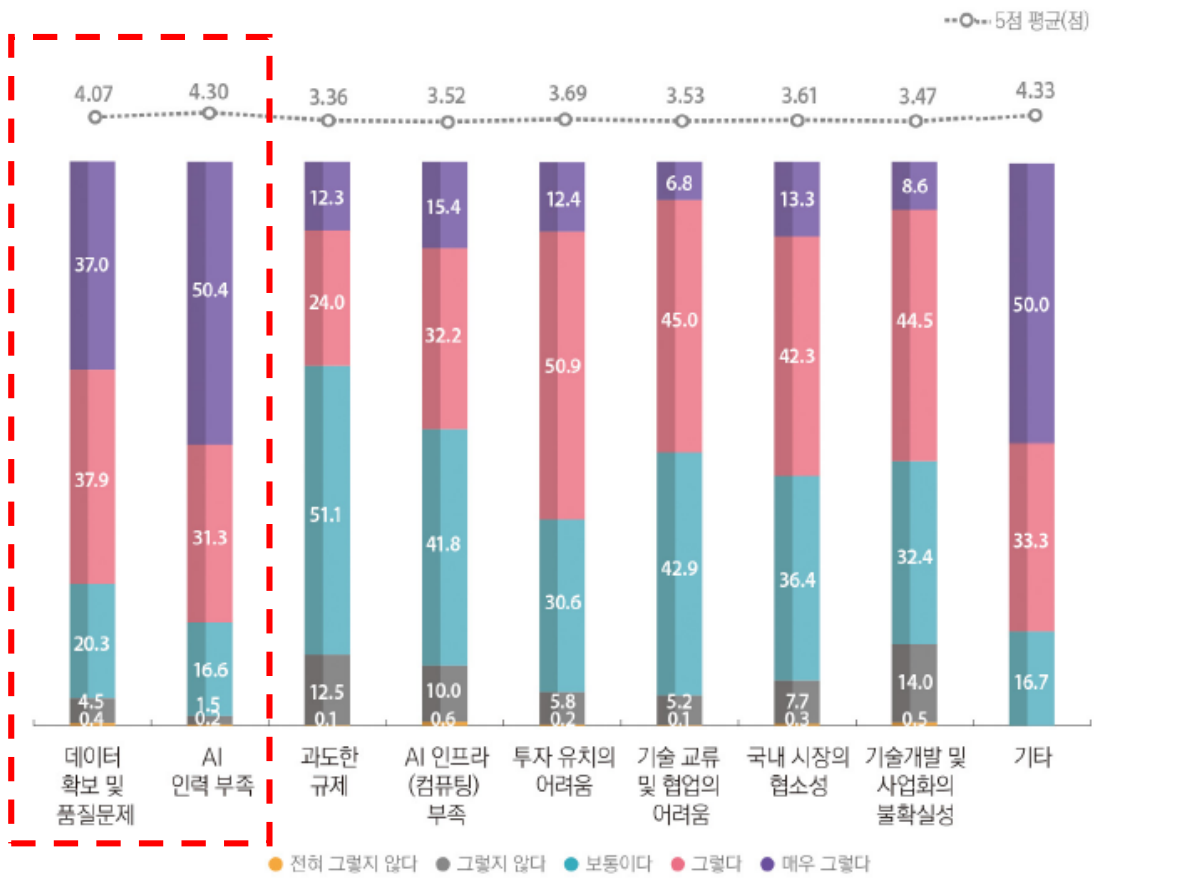


CRISP-DM
(Cross Industry Standard Process For Data Mining, ESPRIT, 1996)

Data 필요성
(데이터 경험과 인공지능 모델 개발)

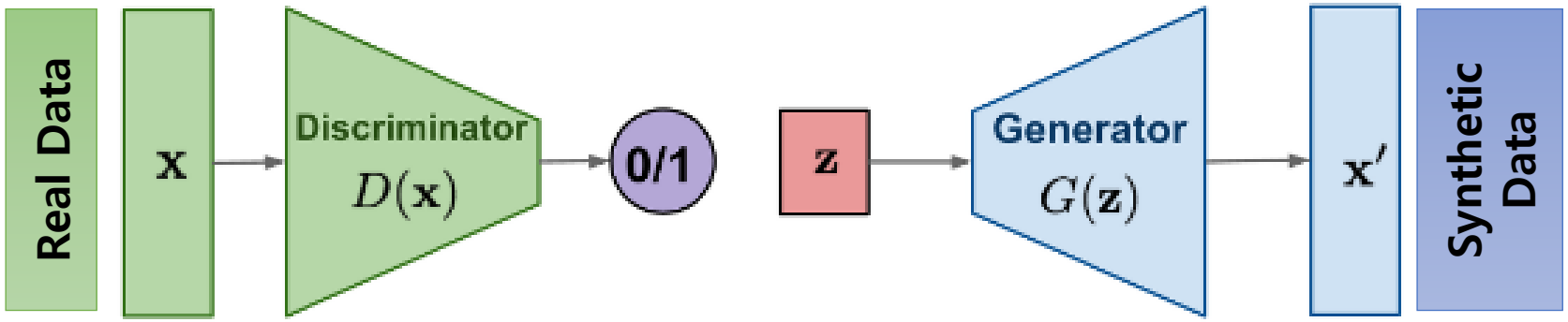
그림47. 인공지능 사업 운영상 느끼는 애로사항 – 전체 항목 비교

[Base= 모집단 전체, n=1,915, 단위: %]



(2022 인공지능산업 실태조사 보고서, 과학기술정보통신부, 2023)

Synthetic Data(합성, 생성, 재현 데이터)



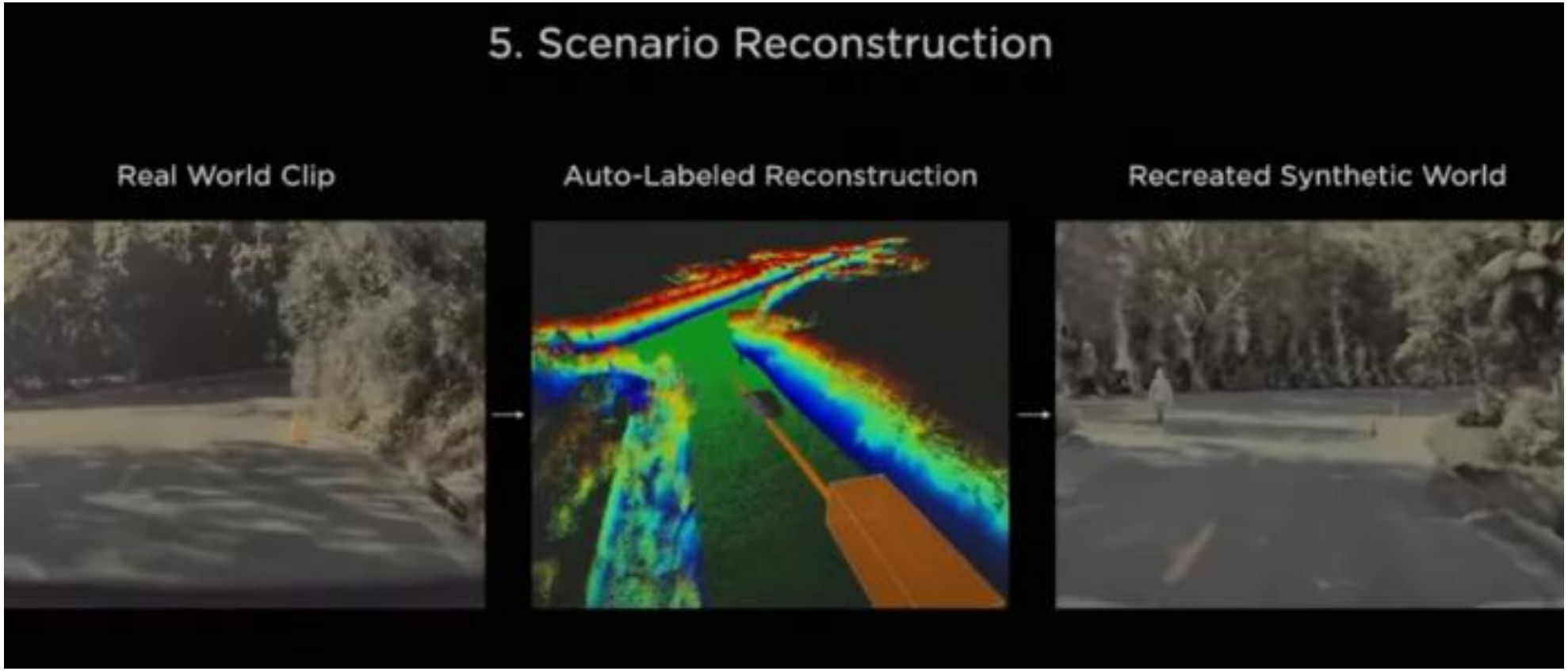
◆ 정의

- 실제로 측정된 데이터(Real Data)를 생성하는 모형이 존재한다고 가정하고, 통계적 방법이나 기계학습 방법 등을 이용하여 추정된 모형에서 새롭게 생성한 모의 데이터(Simulated Data)

◆ 특징

- 모집단의 통계적 특성들을 유지하면서도 민감한 정보를 외부에 직접 공개하지 않음
- 개인이 제공한 데이터가 아닌 임의로 생성한 데이터로 개인정보 관련 법규의 규제로부터 자유로움(익명화)

Synthetic Data Use case(Tesla)



Synthetic Data Use case(Financial services)

J.P.Morgan

Solutions Insights News About Us Contact Us

Synthetic Data > Payments data for Fraud Detection

Synthetic Data

Overview

Anti-Money Laundering (AML)

Customer Journey Event

Markets Execution Data

Payments data for Fraud Detection

Synthetic Documents for Layout Recognition

Synthetic Equity Market Data

Payments data for Fraud Detection

Data representing transactions from a subject-centric view with the goal of identifying fraudulent transaction. This data contains a large variety of transaction types representing normal activities as well as abnormal/fraudulent activities that are introduced with predefined probabilities. The data was generated by running an AI planning-execution simulator and translating the output planning traces into tabular format. Parameters of the data generation model include the number of clients, time duration and probabilities of fraud.

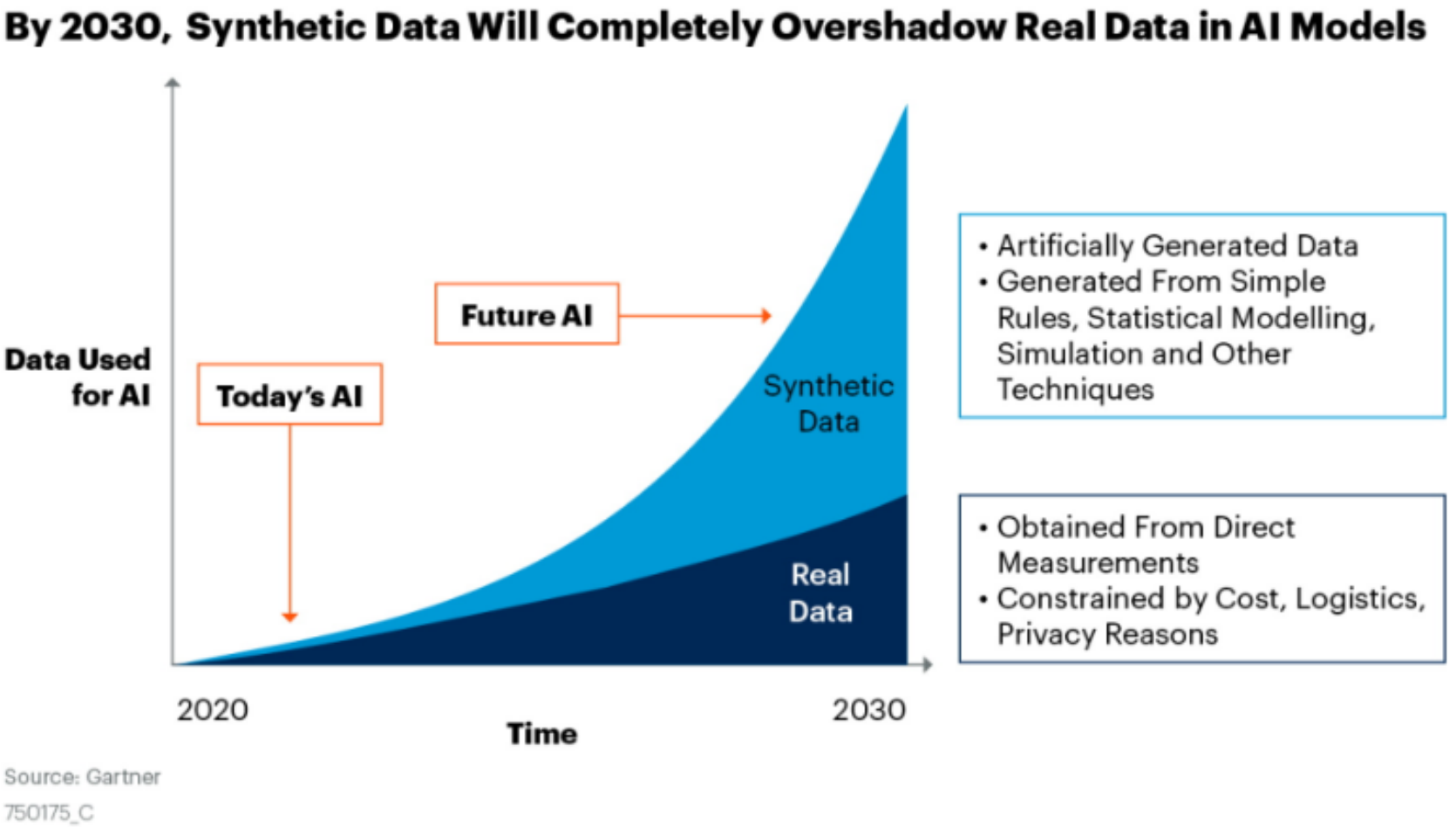
Sample data

Transaction_Id	Sender_Id	Sender_Account	Sender_Country	Sender_Sector	Sender_Iob	Bene_Id	Bene_Account	Bene_Country	USD_Amount	label	Transaction_Type
PAY-BILL-3589	CLIENT-3566	ACCOUNT-3578	USA	21264	CCB	COMPANY-3574	ACCOUNT-3587	GERMANY	492.67	0	MAKE-PAYMENT
WITHDRAWAL-3591	CLIENT-3566	ACCOUNT-3579	USA	18885	CCB				388.92	0	WITHDRAWAL
MOVE-FUNDS-3528	CLIENT-3508	ACCOUNT-3520	USA	4809	CCB	COMPANY-3516	ACCOUNT-3527	GERMANY	280.7	0	MOVE-FUNDS
WITHDRAWAL-3529	CLIENT-3508	ACCOUNT-3519	USA	7455	CCB				118.14	0	WITHDRAWAL
QUICK-DEPOSIT-3471						CLIENT-3442	ACCOUNT-3461	USA	105.16	0	DEPOSIT-CASH
QUICK-DEPOSIT-3473						CLIENT-3442	ACCOUNT-3460	USA	164.97	0	DEPOSIT-CASH
PAY-BILL-3404	CLIENT-3384	ACCOUNT-3395	USA	36316	CCB	COMPANY-3392	ACCOUNT-3401	GERMANY	456.89	0	MAKE-PAYMENT
QUICK-DEPOSIT-3406						CLIENT-3384	ACCOUNT-3396	USA	413.17	0	DEPOSIT-CASH
PAY-CHECK-3347	CLIENT-3330	ACCOUNT-3341	USA	36194	CCB	CLIENT-3333	ACCOUNT-3338	CANADA	377.65	0	PAY-CHECK
PAY-CHECK-3348	CLIENT-3330	ACCOUNT-3340	USA	20626	CCB	CLIENT-3333	ACCOUNT-3338	CANADA	338.03	0	PAY-CHECK
MOVE-FUNDS-3292	CLIENT-3272	ACCOUNT-3284	USA	21568	CCB	CLIENT-3275	ACCOUNT-3291	CANADA	100.85	0	MOVE-FUNDS
MOVE-FUNDS-3294	CLIENT-3272	ACCOUNT-3284	USA	29040	CCB	CLIENT-3273	ACCOUNT-3289	USA	276.66	0	MOVE-FUNDS
PAY-BILL-3232	CLIENT-3203	ACCOUNT-3222	USA	27393	CCB	COMPANY-3210	ACCOUNT-3218	GERMANY	234.88	0	MAKE-PAYMENT
QUICK-DEPOSIT-3234						CLIENT-3203	ACCOUNT-3222	USA	945.22	0	DEPOSIT-CASH
DEPOSIT-CASH-3163						CLIENT-3139	ACCOUNT-3154	USA	655.09	0	DEPOSIT-CASH
PAY-BILL-3162	CLIENT-3139	ACCOUNT-3153	USA	25066	CCB	COMPANY-3147	ACCOUNT-3160	GERMANY	675.37	0	MAKE-PAYMENT
WITHDRAWAL-3100	CLIENT-3075	ACCOUNT-3090	USA	22778	CCB				319.95	0	EXCHANGE
QUICK-PAYMENT-3099	CLIENT-3075	ACCOUNT-3091	USA	39013	CCB	CLIENT-3078	ACCOUNT-3087	TAIWAN	771.54	0	QUICK-PAYMENT
PAY-BILL-3036	CLIENT-3016	ACCOUNT-3028	USA	43951	CCB	COMPANY-3022	ACCOUNT-3033	GERMANY	730.69	0	MAKE-PAYMENT

References

1. Generating Synthetic Data in Finance: Opportunities, challenges and pitfalls. S Assefa, D Dervovic, M Mahfouz, R Tillman, P Reddy, T

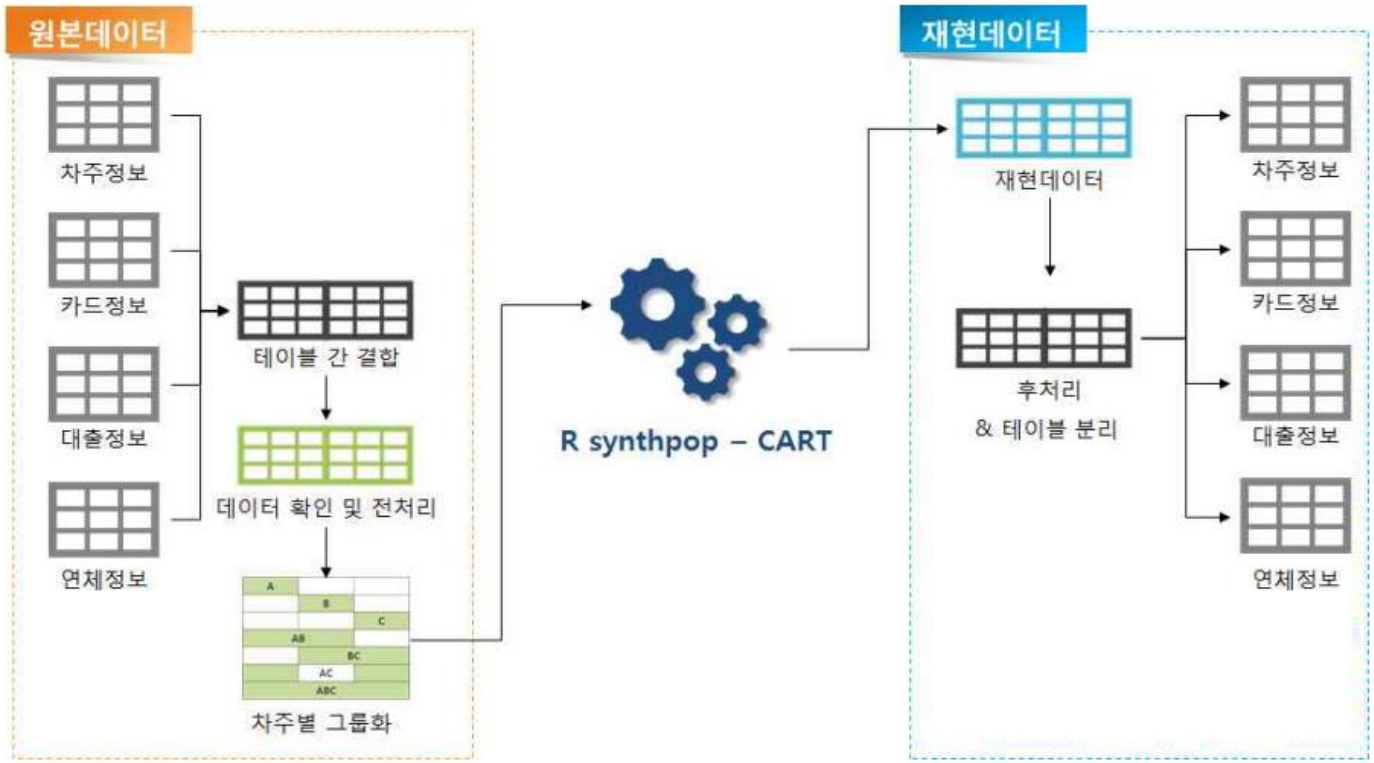
Synthetic Data



Gartner

CreDB Synthetic Data

[개인신용정보 재현데이터 생성 과정]



CART

(Classification And Regression Tree)

- 희귀분석기반 머신러닝 기법 활용
  (생성 가능 데이터 한계)
- 원격 분석 시스템 운영
  (데이터 접근성 및 활용성)

- **(목적)** 통계분석 및 교육 지원, 신용정보 교육 등
- **(규모)** 약 200만 명에 해당하는 가상 차주에 대한 시계열 데이터

Synthetic Data by GAN

```python
from ctgan import CTGAN

real_data = df_GAN

# Names of the columns that are discrete
discrete_columns = [
    'COMP_SCL_CD',
    'COMP_ADDR_prep',
    'COMP_INDU_CLSF_CD1',
    'COMP_INDU_CLSF_CD2',
    'TECH_CLSF_CD1',
    'TECH_CLSF_CD2',
    'OS_YN',
    'VNTR_CERTI_YN',
    'INNOBIZ_CERTI_YN',
    'CNTRY_RD_PFMC_RETN_YN',
    'ITEM_CD19'
]

ctgan = CTGAN(epochs=10)
ctgan.fit(real_data, discrete_columns)

# Create synthetic data
synthetic_data = ctgan.sample(100000)

# Save it to disk
synthetic_data.to_csv('TCB_synthetic_data_20230430.csv', index=False)

# Save the fitted CTGAN model to disk
ctgan.save('TCB_ctgan_model_20230430.pkl')
```

이산형, 범주형 변수

생성 데이터 수

생성 데이터 저장

생성 모델 저장

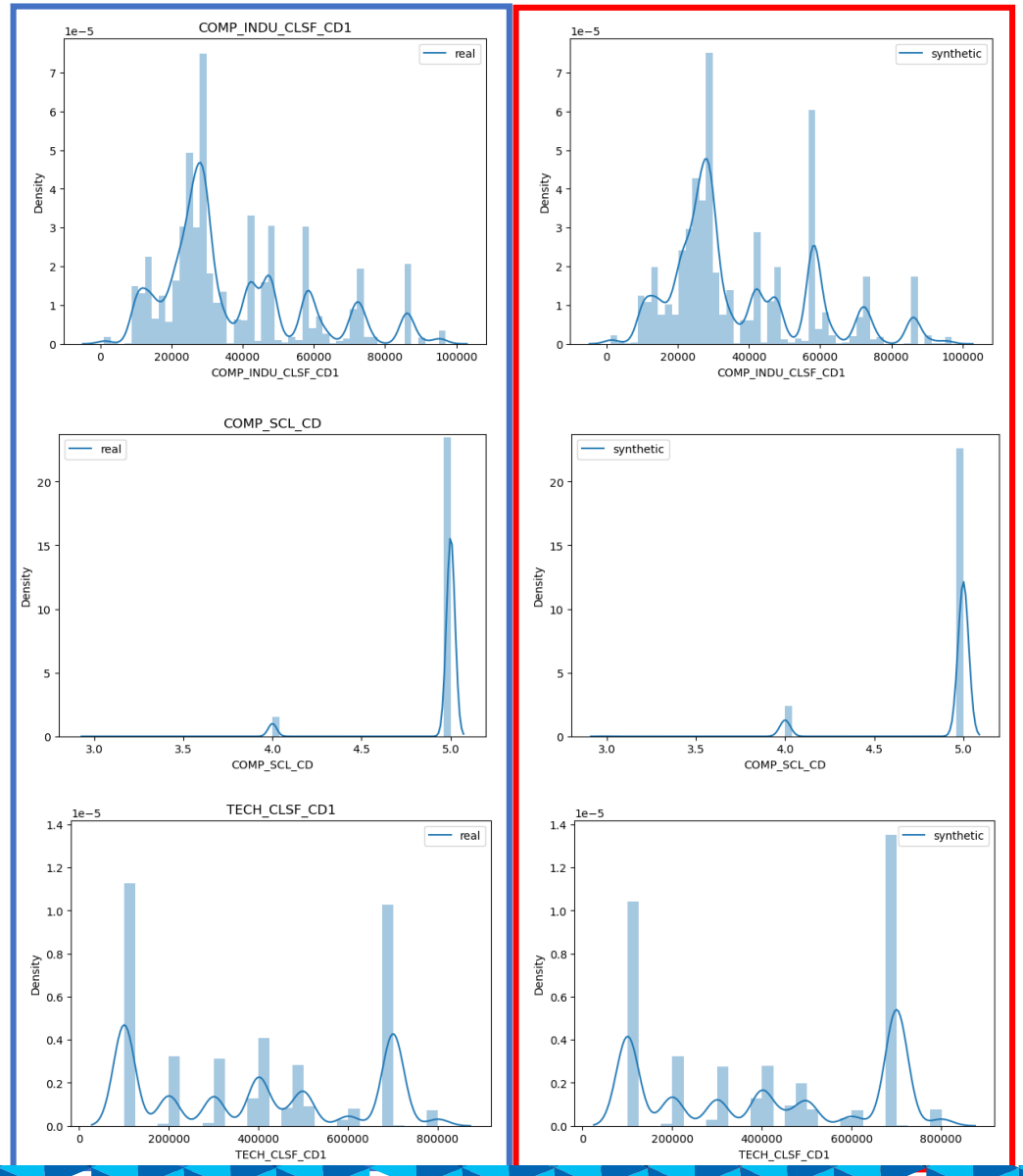
CTGAN

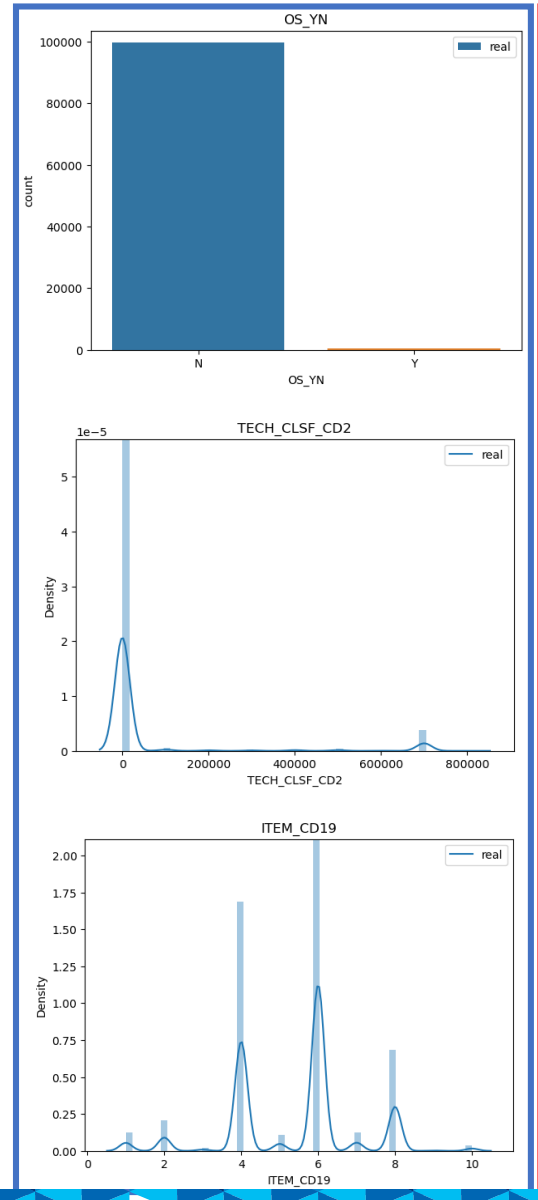
(Conditional Tabular Generative Adversarial Network)

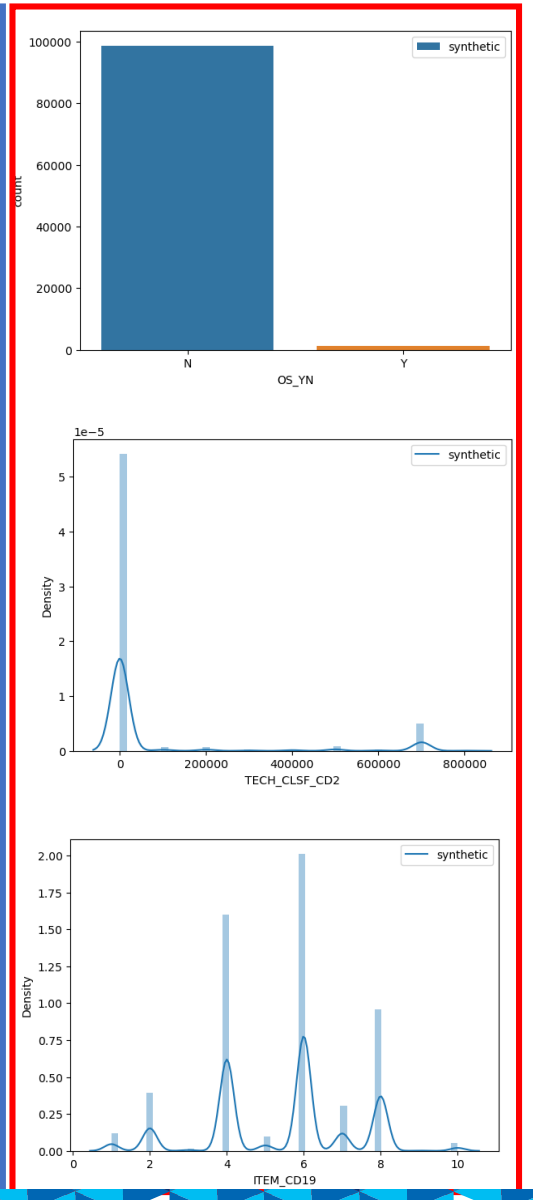
테이블 형태의 데이터는 정형 데이터 내 연속 데이터(수치)와
이산 데이터(범주, 분류)가 포함되어 있어, 기존 GAN이 두 형태가
혼합된 데이터의 생성 효율이 떨어지는 단점을 보완한 생성 AI 모델

→ 금융 데이터 생성에 적합한 생성 AI모델

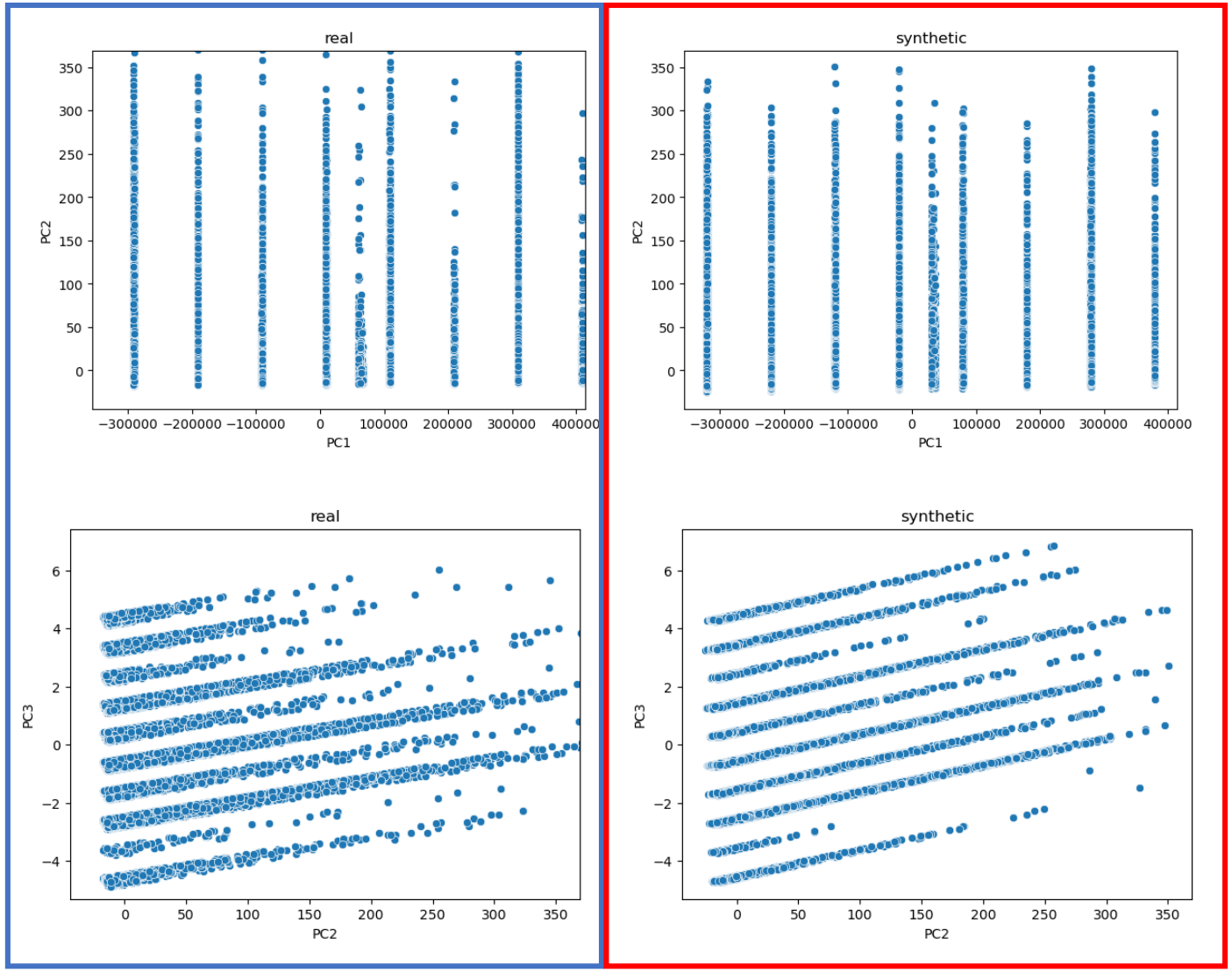
Synthetic Data







Synthetic Data

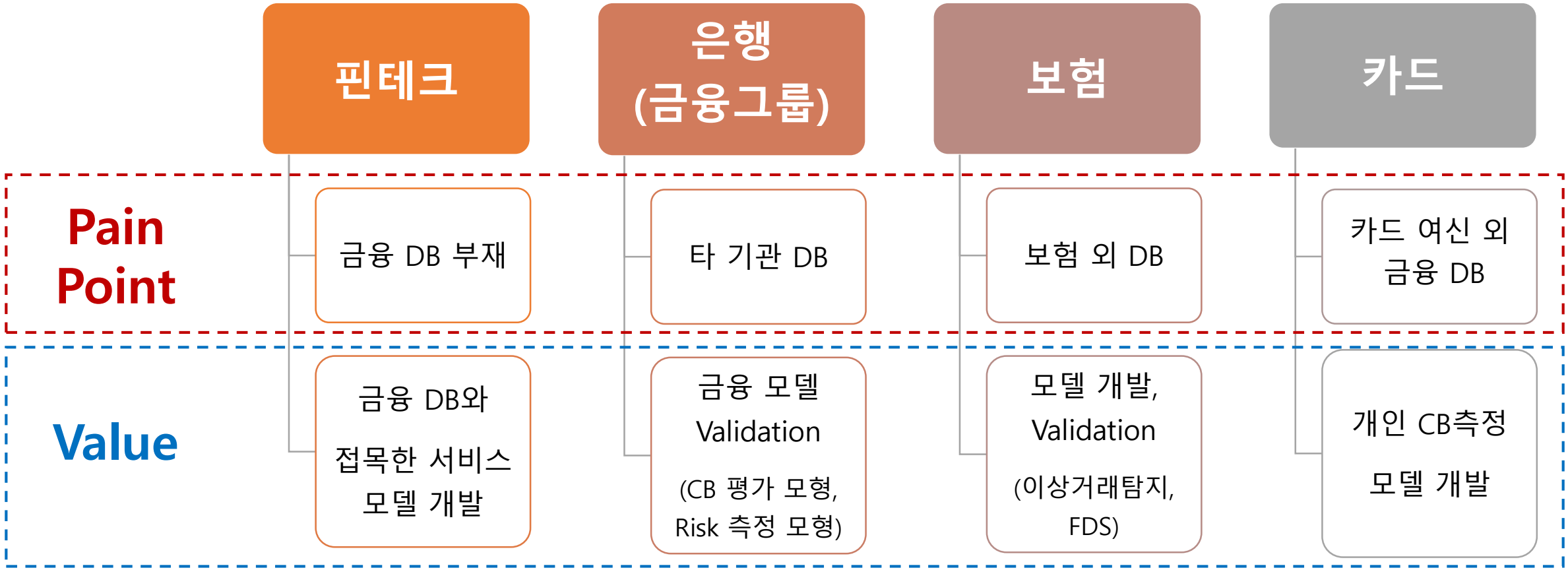


PCA(주성분 분석)

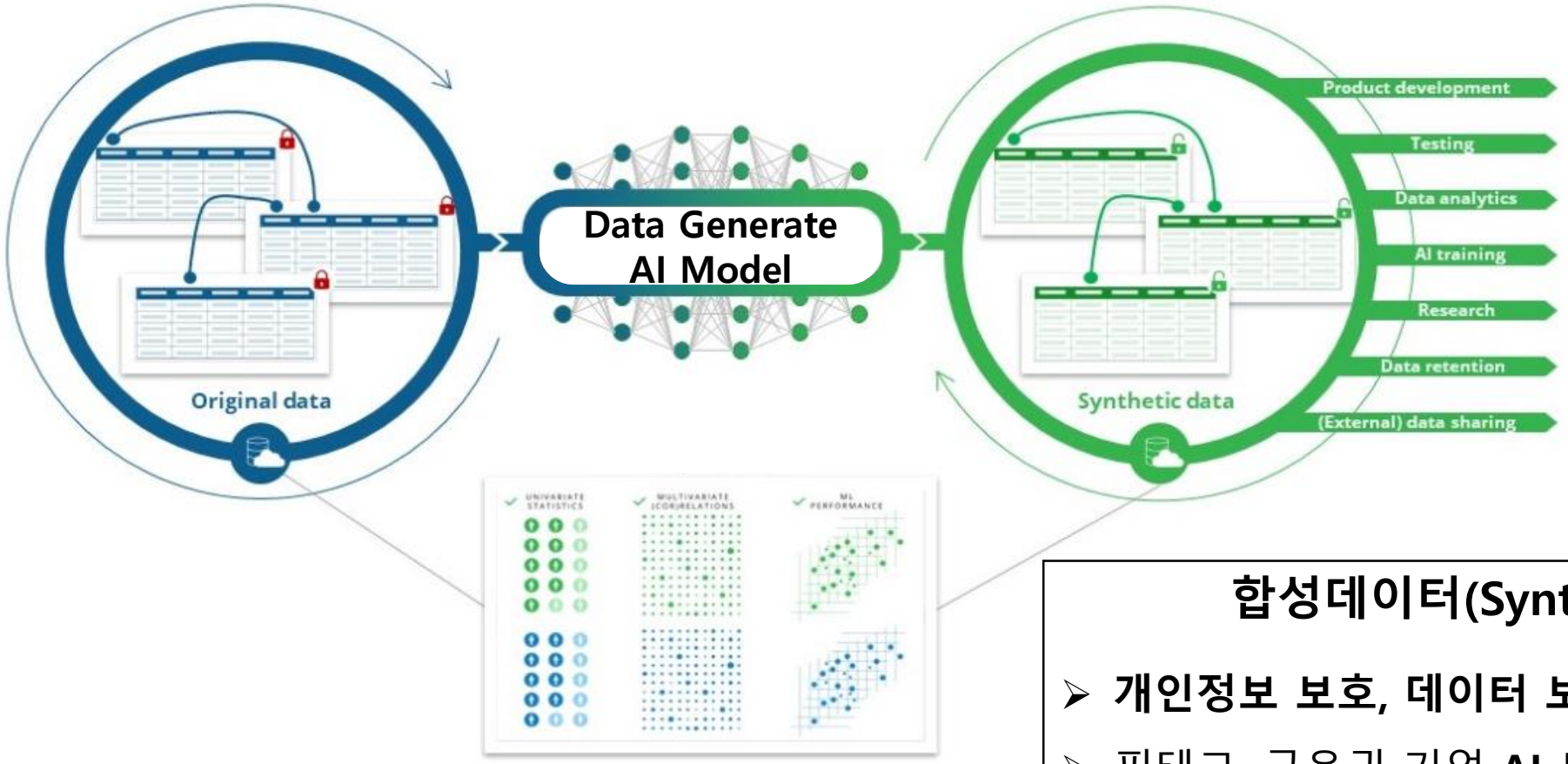
(Principal Component Analysis)

→ 데이터 변수 사이의 의존성 및 상관관계 유지

Synthetic Data Value



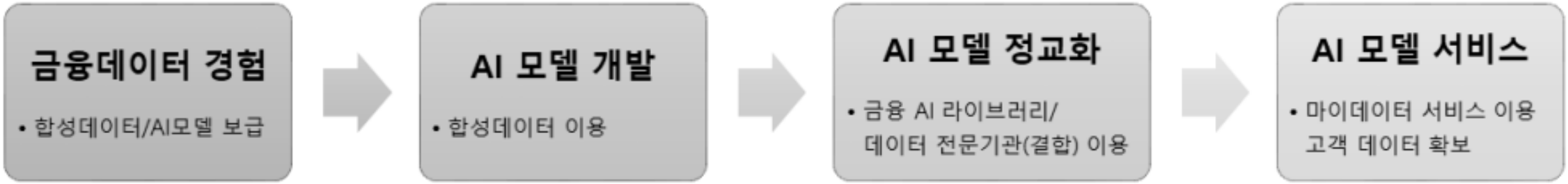
Synthetic Data 활용



합성데이터(Synthetic Data) 활용

- **개인정보 보호, 데이터 보안** 이슈 해결
- 핀테크, 금융권 기업 **AI 모델링에 활용**
- **소규모 샘플 DB –데이터 경험 확대**(외부 활용)
- **대규모 DB – AI 모델 분석 테스트베드 활용**

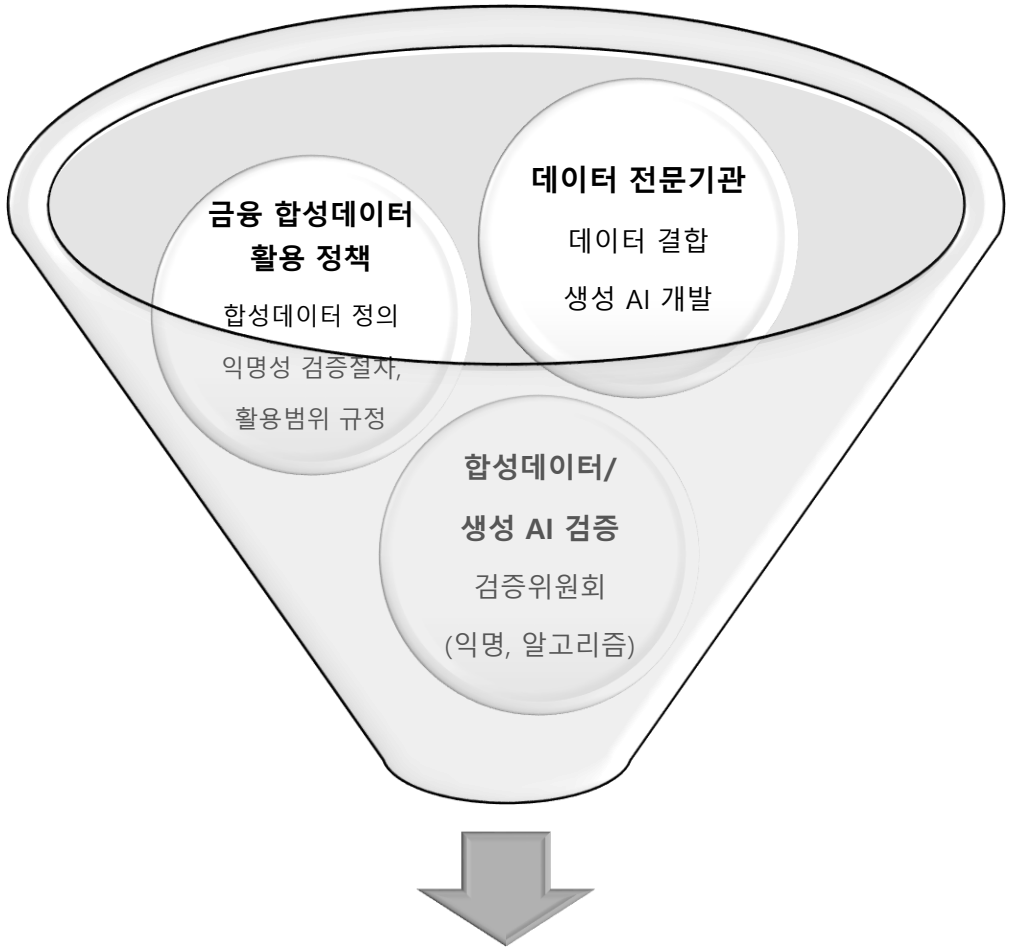
핀테크 산업 합성데이터 활용 프로세스



합성데이터(Synthetic Data) 활용 기대효과

- 합성데이터 확보를 위한 **데이터 결합 서비스 이용**
- 핀테크 서비스 제공을 위한 **마이데이터 서비스 활성화** 기대

합성데이터 운영 시스템



합성데이터 / AI 이용 활성화