

국가데이터연구 심포지엄

AI 학습용 데이터를 위한 데이터보호 신기술 연구개발 현황

국가데이터처의 역할과 데이터보호 기술 연구개발 추진 방향

발표 일자 2026. 6. 25.

주 최 국가데이터연구원

발표 목차

Table of Contents

01

발표 개요 및 목적

심포지엄 발표 배경과 목표

03

데이터보호 신기술 동향

국내외 적용 사례 및 연구개발 현황

05

향후 추진 방향

범정부 연계 활용을 위한 필요 사항

02

AI시대 데이터 활용 환경 변화

학습데이터 부족·데이터 안보·연계 활용 한계

04

국가데이터처 R&D 추진 현황

비식별화·재현자료·동형암호 3대 신기술

06

데이터센터의 미래상

신기술 적용 후 데이터센터 발전 방향

발표 개요 및 목적

Background and Objective



발표 배경

AI 모델의 성능은 학습데이터의 양과 질에 좌우되며, 양질의 데이터 확보가 국가 경쟁력의 핵심 요소로 부상

개인정보·기밀정보 등 민감 데이터의 보호 필요성과 활용 가치 사이의 균형이 시급한 과제로 대두

국가데이터처는 데이터의 안전한 연계·활용을 위한 정보보호 신기술의 연구개발을 본격 추진



발표 목적

01

현황 공유

AI 시대 데이터 활용 환경과 보호 기술의 국내외 동향 공유

02

역할 정립

국가데이터처의 데이터보호 신기술 R&D 추진 현황 및 역할 소개

03

협력 모색

범정부 데이터 연계 활용을 위한 정책·기술적 협력 방안 논의

AI시대 학습데이터 부족

데이터 수요 폭증과 공급 정체의 구조적 불일치

정책적 배경: 과학기술정보통신부 인공지능 대전환(AX) 전략
국가 가용 자원을 총동원하여 대한민국 AI G3 도약 지원 ('25.2월)

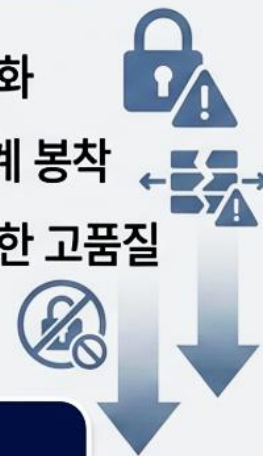
수요 (Demand) 폭발

- AI 모델 학습을 위한 고부가가치 데이터 절실
- 공공데이터 전면 개방 요구
- 민간 수요가 높은 비정형 원본데이터 확충
- 제조 및 산업 데이터 연계 급증



공급 (Supply) 한계

- 개인정보 침해 및 유출 우려 심화
- 현행 비식별 조치의 기술적 한계 봉착
- 실제 연계 및 분석에 활용 가능한 고품질 원본 데이터 턱없이 부족



결론: 단순한 정책 완화를 넘어선 '기술적 돌파구'를 통한 고품질 데이터의 안전한 공급망 구축 필수

AI시대 학습데이터 부족

데이터 수요 폭증과 공급 정체의 구조적 불일치

초거대 AI 시대 도래로 양질의 학습데이터 수요가 폭증하는 반면, 활용 가능한 공공·민간 데이터는 보호 규제로 제한적

수요 측면

$10^4 \times$

최근 5년간 글로벌 AI 학습데이터
요구량 증가 추정치

공급 측면

$< 30\%$

공공 데이터 중 AI 학습에 직접
활용 가능한 비율

보호 제약

법·기술

개인정보보호법·민감정보 처리
제약과 비식별화 한계

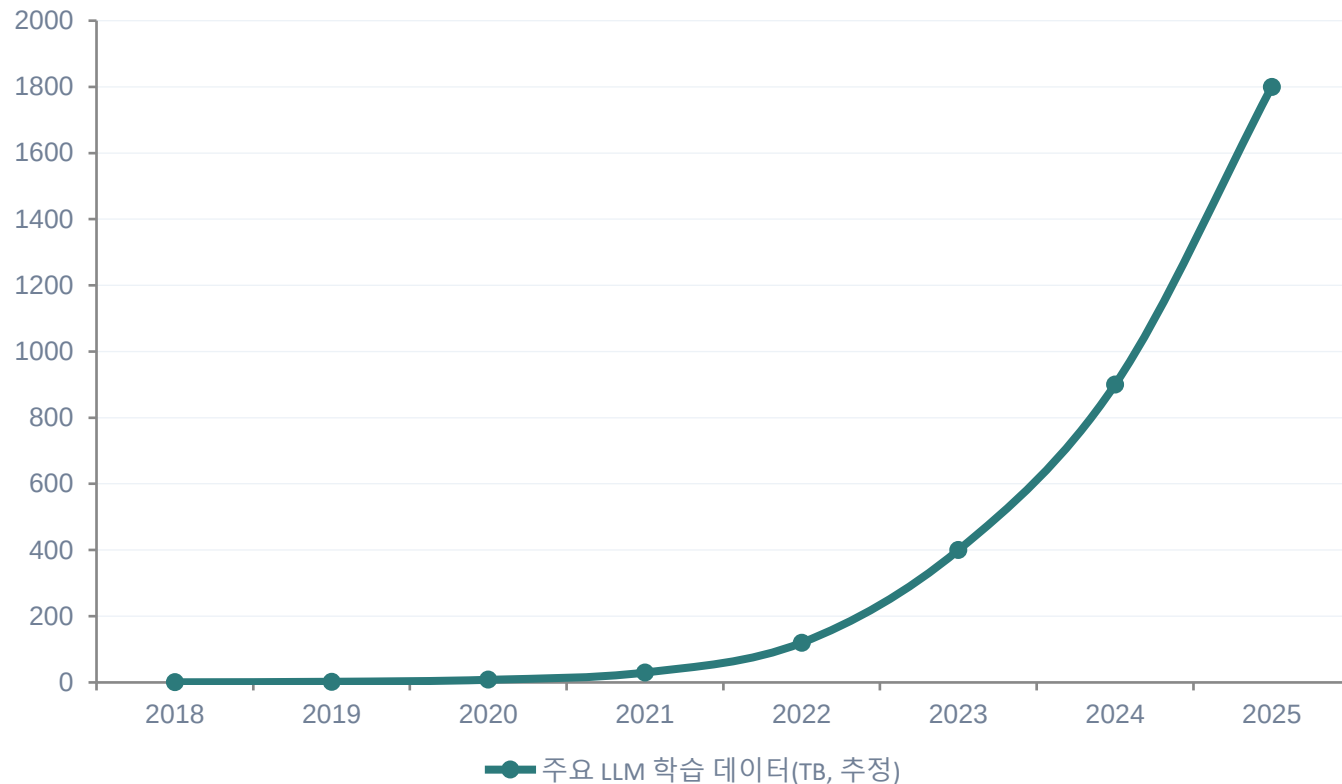
→ 시사점

데이터의 보호와 활용을 동시에 가능하게 하는 신기술 도입 없이는 AI 경쟁력 확보 곤란

학습데이터 수요·공급 격차 심화

주요 AI 모델의 학습 데이터 규모 증가 추세

주요 대규모 AI 모델 학습 데이터 규모 (TB, 추정)



주요 시사점

지수적 증가

AI 모델 학습 데이터는 7년간 약 3,000배 이상 증가

공공 데이터 한계

공공·웹 크롤링 데이터의 양적·질적 포화 시점 도래

민감 데이터 의존

차세대 AI는 의료·금융 등 민감 영역 데이터 활용 불가피

기술적 돌파구

재현데이터·비식별화·동형암호 등 신기술 적용 필요성 증대

AI시대 데이터 연계 활용 한계

기관 간 데이터 결합·분석의 현행 제약 사항

부처·기관 간 데이터를 결합하여 활용하면 더 큰 정책적·산업적 가치를 창출할 수 있으나, 현재는 여러 제약으로 제한적 활용에 머무름



법·제도적 한계

개인정보보호법, 분야별 개별법(의료·금융·통신 등)의 칸막이 규제로 결합·재활용에 절차적 부담



기술적 한계

전통적 비식별화는 데이터 효용 저하가 크고, 재식별 위험 평가가 모호하여 안심하고 공개·결합 곤란



운영·인프라 한계

기관별로 분산된 데이터 저장·관리 체계, 상호운용 표준 부재로 안전한 연계 환경 구축 비용 증대



신뢰·수용성 한계

데이터 주체의 동의·통제 메커니즘 미비, 사회적 수용성 부족으로 활용 확대에 제약

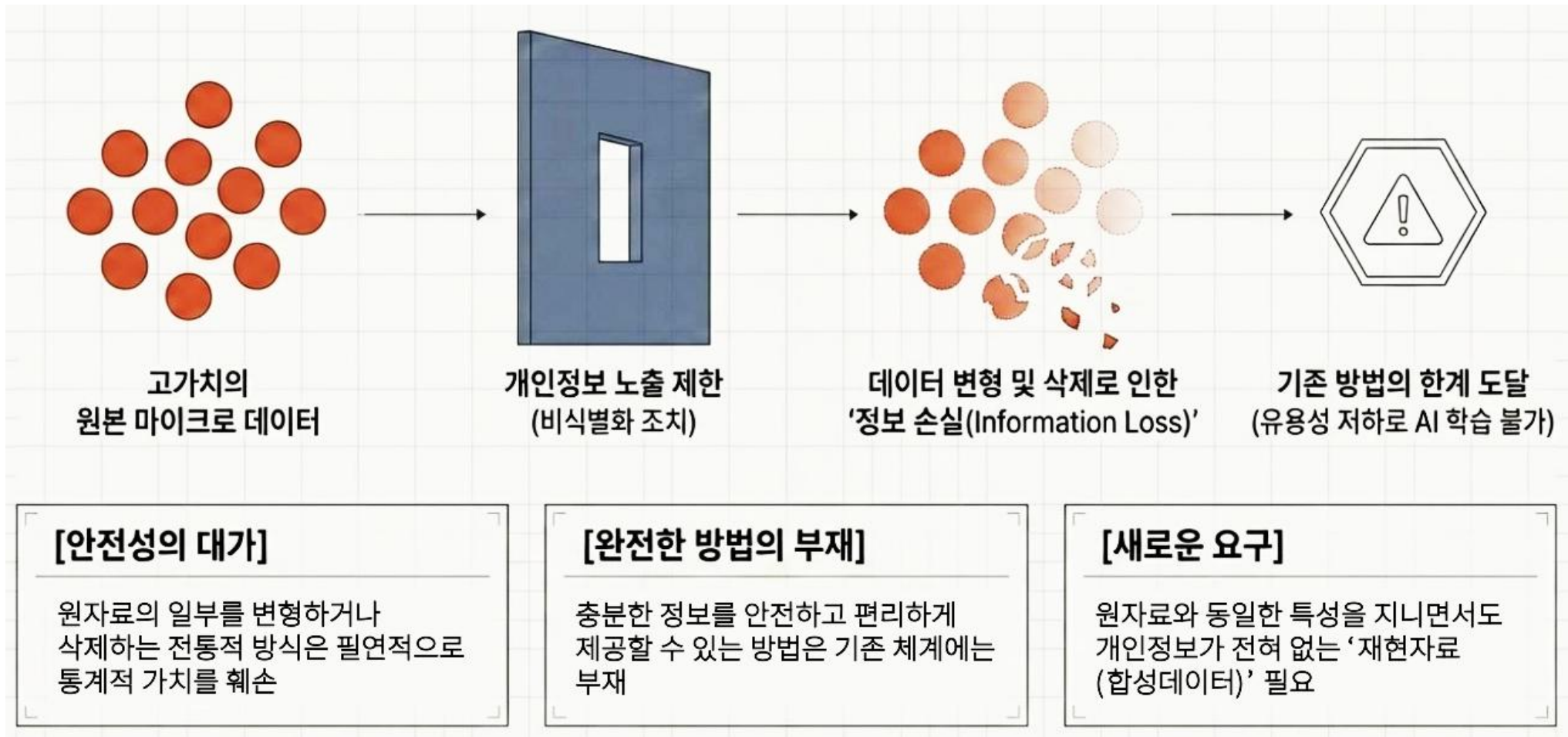
AI시대 데이터 연계 활용 한계

기관 간 데이터 결합·분석의 현행 제약 사항

	데이터 입수	데이터 연계	데이터 활용
현황	 행정자료 입수 327종 327종 행정자료 및 67종 민간자료 입수 198종 행정DB 구축 및 제공 ('25.12. 기준)	 융합데이터 구축 7종 아동가구, 청년, 취업활동, 연금, 소득이동, 생활인구, 중소기업	 통계데이터센터(SDC) 운영 16개소 행정자료·통계등록부로 통계 생산 지원 SDC·KOSIS 등 데이터 서비스 플랫폼 운영
문제점	 입수 체계 분산 및 데이터 공백 발생 - 입수 체계 분산 - 수요 대비 데이터 공백 발생 - 서식 및 데이터 형식 불일치	 데이터 제공 시의성 부족 - 융합데이터 제공 시의성 한계 - 신규 데이터 요청시 건별 신속한 대응에 한계	 물리적·시간적 제약, 데이터 환류 부족 - 물리적 제약: 직접 방문 필요 - 시간적 제약: 연계, 반출 시간 소요 - 자료제공기관의 데이터 활용 한계

AI시대 데이터 연계 활용 한계

기관 간 데이터 결합·분석의 현행 제약 사항



AI시대 데이터보호 신기술 적용 가능 분야

보호와 활용을 동시에 가능케 하는 3대 핵심 기술

AI 기반 비식별화

De-identification

원본 데이터에서 식별 가능 정보를
자동으로 탐지·처리

적용 분야

- 공공기관 행정통계 공개
- 의료영상 환자정보 마스킹
- 민간 마케팅 데이터 가공



재현(합성) 데이터

Synthetic Data

원본의 통계적 특성을 보존한 인공
데이터 생성

적용 분야

- 금융 신용·이상거래 모델 학습
- 의료 희귀질환 연구 데이터
- 교통·인구 수요 예측 시뮬레이션



동형암호

Homomorphic Encryption

암호화된 상태에서 직접 연산·분석을
수행

적용 분야

- 기관 간 통계 결합 분석
- 프라이버시 보존 머신러닝
- 다자간 데이터 연합 분석

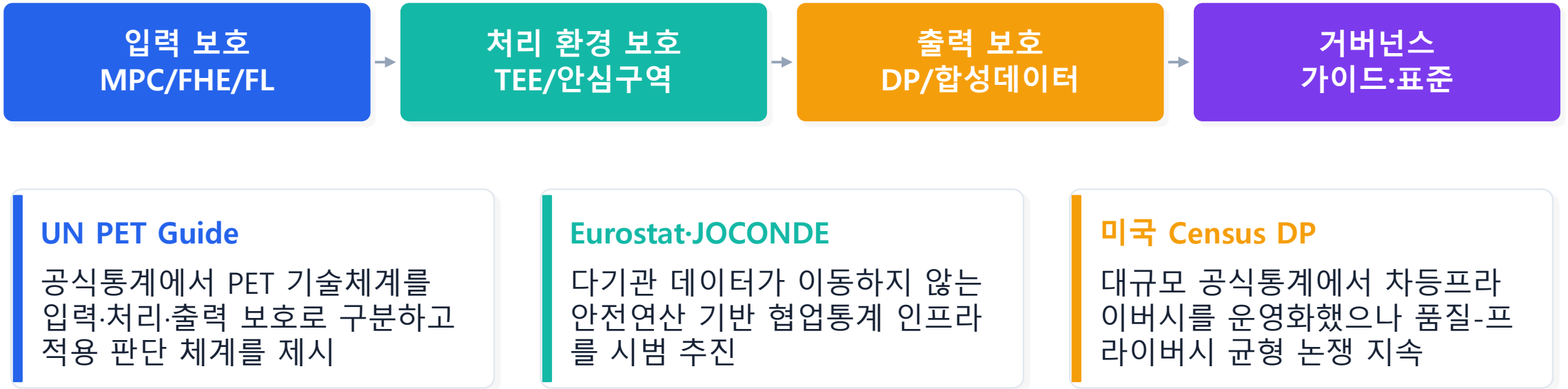
SECTION 03

데이터보호 신기술 국내외 연구개발 현황

R&D Status at Home and Abroad

국제 PET 동향: 입력 보호와 출력 보호의 병행

Strengths and Weaknesses of Domestic R&D



시사점: PET는 단일 기술이 아니라 데이터 유형·활용 목적·위험 수준에 맞춘 조합 전략

국제 PET 동향: 패러다임의 전환

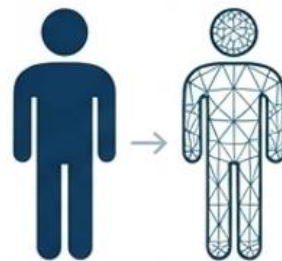
Strengths and Weaknesses of Domestic R&D



동형암호 (Homomorphic Encryption)

데이터 노출 없는 실시간 분석

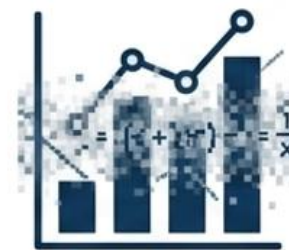
평문 데이터를 암호화한 상태 그대로
복호화 과정 없이 AI 모델 연산을
수행하는 차세대 암호 기술



재현자료 (Synthetic Data)

원본 없는 안전한 통계 모방

원자료의 합계, 평균, 분포 등 통계적
특성은 동일하게 유지하되 가상으로
새롭게 합성하여 생성된 인공 데이터



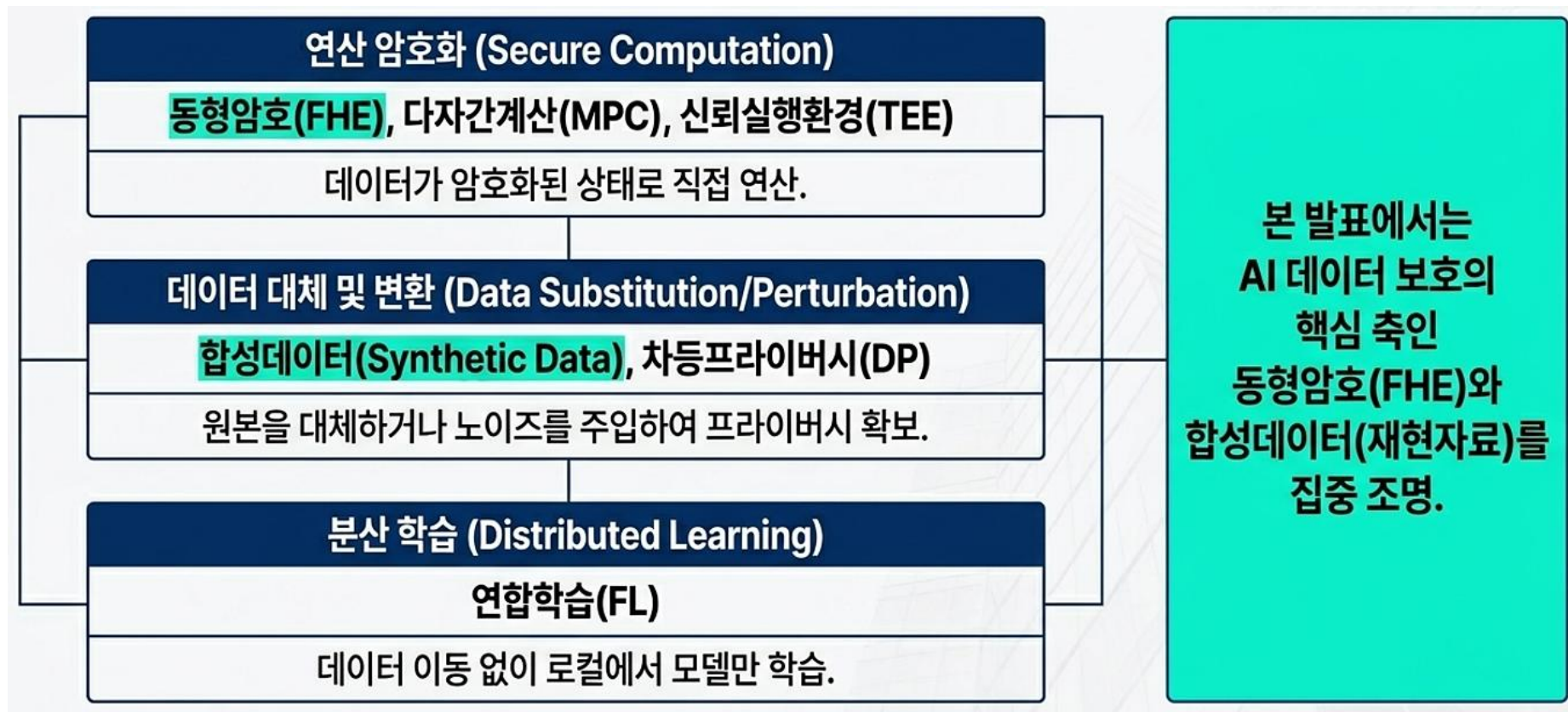
차등정보보호 (Differential Privacy)

개인 식별의 완벽한 통제

원본 데이터 세트에 정교하게 계산된
임의의 노이즈를 삽입하여 특정 개인의
포함 여부를 역추할 수 없게 차단하는 기술

국제 PET 동향: 패러다임의 전환

Strengths and Weaknesses of Domestic R&D



SECTION 04

국가데이터처 데이터보호 기술 R&D 현황

R&D Status of the Korea Data Agency

국가데이터처 데이터보호 기술 연구개발 현황

Mission, Strategy, and Three R&D Pillars

MISSION

국가데이터처의 역할

범정부 데이터의 안전한 **연계·활용** 환경 조성

데이터보호 신기술 R&D 추진을
통한 기술 자립과 표준화 선도

정책·제도·기술이 결합된 **통합 거버넌스**를 통해
데이터 주권과 산업 경쟁력 강화

3대 핵심 R&D 영역

01



AI 기반 집계자료 자동 비식별화

공공·통계 데이터의 자동 비식별화 처리 시스템 개발

02



AI 기반 재현자료 자동 생성·평가

통계적 특성을 보존하는 합성데이터 자동 생성·품질 평가

03



동형암호 기반 데이터 연계·활용

암호화 상태에서의 다자간 데이터 결합·분석 환경 구축

AI 기반 집계자료 자동 비식별화

기술 개요 및 기존 방식과의 차별점

연구 목적

집계자료 특성에 맞게 비식별화 방법을 자동 추천하고, 처리 결과의 품질을 자동 평가하는 시스템 개발

핵심 범위

데이터 탐색, RAG 기반 지식 활용, 처리 기준 자동 적용, 결과 보고서 생성까지 업무 전주기 지원

데이터 입력

위험 탐지

처리 추천

검증·보고

시간 단축

반복 처리와 보고서 생성 자동화

품질 일관화

가이드라인·사례 기반 추천

전문가 보조

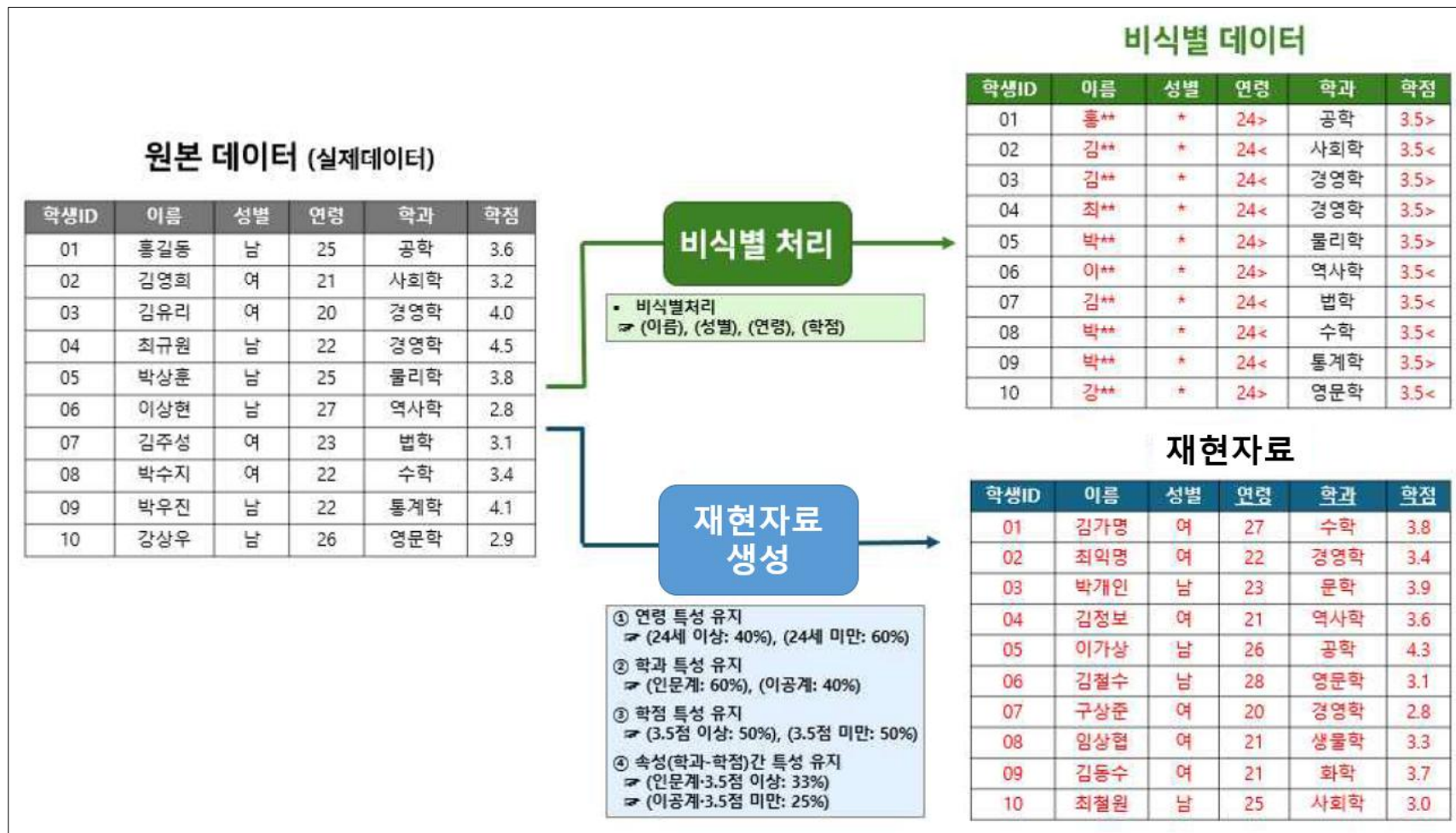
근거 제시형 의사결정 도구

환류 체계

처리 이력과 평가결과 축적

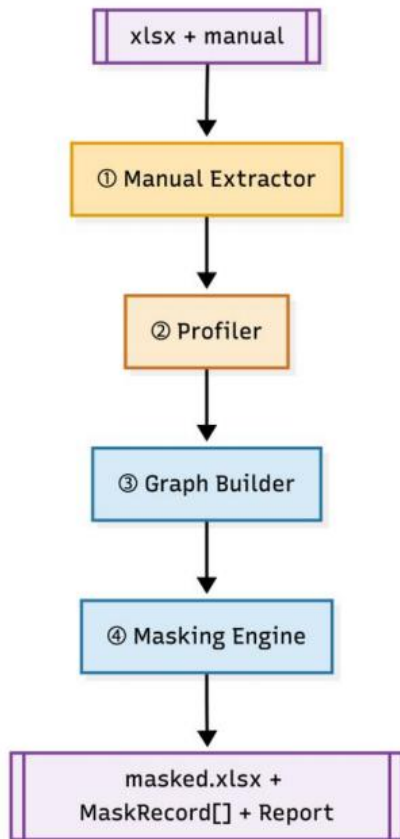
AI 기반 집계자료 자동 비식별화

기술 개요 및 기존 방식과의 차별점

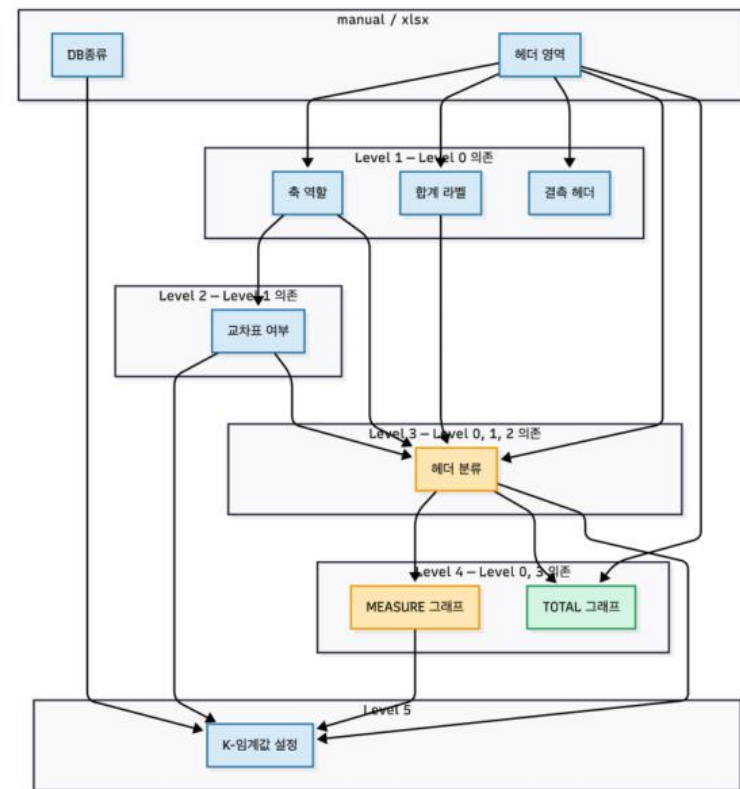


AI 기반 집계자료 자동 비식별화

기술 개요 및 기존 방식과의 차별점

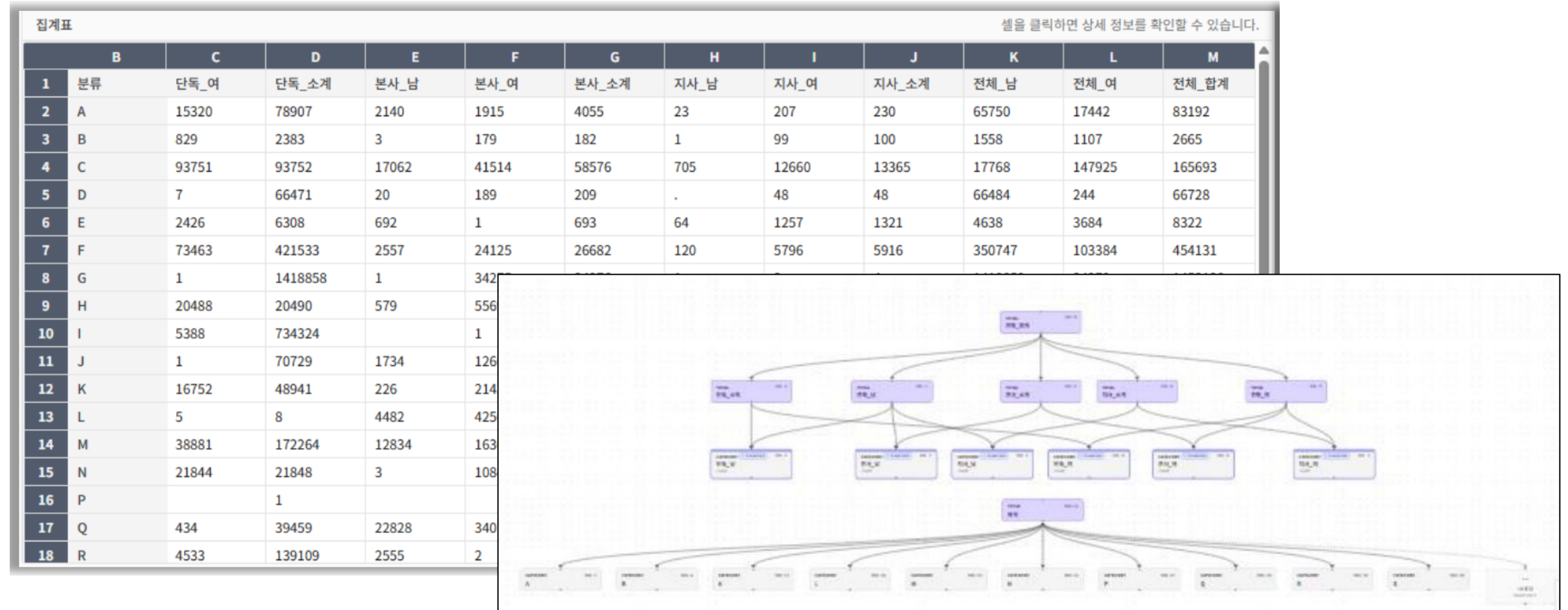


프로파일러 분석 항목	산출
DB종류	business / population (K=3 vs 5)
헤더 영역	헤더 영역 경계 (label_rows / label_cols)
축 역할	축 역할 (category / measure)
교차표 여부	crosstab 분기 (양 축 category)
합계 라벨	소계 / 총계 라벨 식별
헤더 분류	헤더 3분류 + kind (17 종)
결측 헤더	결측 행 / 열 (K-익명성 제외)
MEASURE 그래프	count ↔ measure 종속
TOTAL 그래프	소계 = Σ children + powerset 자동 발견
K-임계값 설정	base_k + per-count override



AI 기반 집계자료 자동 비식별화

기술 개요 및 기존 방식과의 차별점



AI 기반 집계자료 자동 비식별화

기술 개요 및 기존 방식과의 차별점

개념도

- 1 원본 데이터**
개인정보 포함 행정·통계 데이터
- 2 AI 자동 분석**
재식별 위험 항목 자동 탐지·평가
- 3 비식별 산출물**
안전하게 공개·활용 가능한 데이터

기존 방식 대비 차별점

수작업 의존 → **AI 자동화**
전문가 수작업 → AI 모델 기반 자동 처리

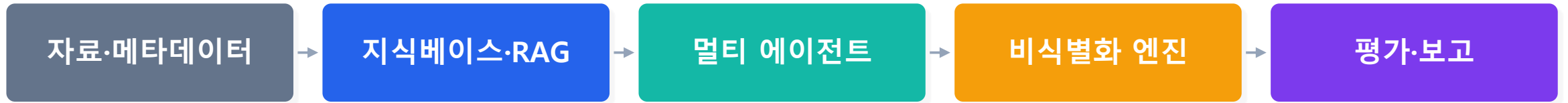
획일적 처리 → **맥락 인식**
고정 규칙 → 데이터 특성에 맞는 적응형 처리

정성 평가 → **정량 평가**
재식별 위험을 수치화하여 객관적 측정

원본 손실 큼 → **효용 보존**
분석에 필요한 통계적 특성 최대한 유지

AI 기반 집계자료 자동 비식별화

기술 개요 및 기존 방식과의 차별점



지식관리 에이전트

가이드라인, 과거 처리사례, 평가기준을 검색·정리해
추천 근거 제공

정제·탐지 에이전트

결측, 이상치, 희소 셀, 비밀보호 규칙 위반 가능성을
사전 탐지

처리·튜닝 에이전트

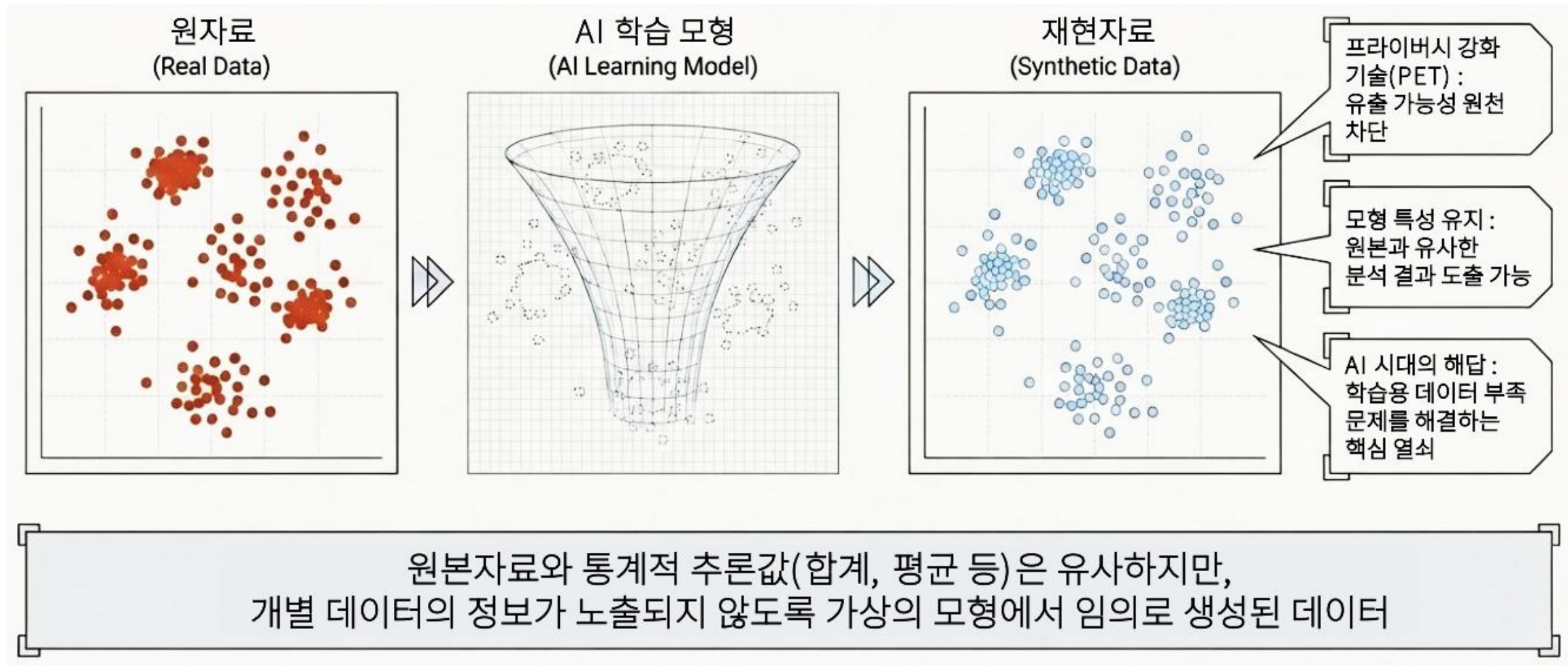
목표 안전성 기준을 만족하도록 마스킹·범주화·노이즈
수준을 조정

보고 에이전트

위험 감소, 정보손실, 처리 이력, 재현 가능성을 보고
서로 자동 정리

AI 기반 재현자료 자동 생성·평가

기술 개요 및 활용 영역



AI 기반 재현자료 자동 생성·평가

기술 개요 및 활용 영역



개념

원자료의 통계적 분포와 분석 결과를 유사하게 유지 하되, 개별 원자료는 포함하지 않는 합성 데이터

가치

개인정보 노출 위험을 낮추면서 연구자의 사전 분석, AI 학습, 데이터 품질 점검을 지원

핵심은 “완벽한 복제”가 아니라 목적별 안전성-유용성의 최적점을 찾는 것

재현자료 생성 기법 비교

기술 개요 및 활용 영역

샘플 기반

관측 샘플을 변형·혼합. 운영은 단순하나 속성노출 위험 관리가 어려울 수 있음

통계모형 기반

분포를 추정해 샘플링. 해석 가능성이 높지만 복잡한 구조 반영에 한계

딥러닝 기반

CTGAN, TVAE, Diffusion 등 복잡한 비선형 관계 포착에 강하나 튜닝 비용 증가

LLM·멀티모달 확장

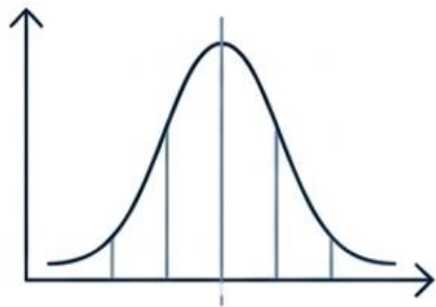
텍스트·표·이미지까지 확장 가능하나 사실성·구조 보존 검증이 중요

재현자료 생성 기법 비교

기술 개요 및 활용 영역

통계 모형 기반

(Statistical Model-based, 예: Synthpop)



- **핵심 원리:** 원본의 통계적 분포 특성을 추정하여 확률적으로 생성 (조건부 확률분포 이용).
- **장점:** 이해 및 적용이 직관적, 연산 리소스가 적게 소모 됨.
- **단점:** 변수 생성 순서(도메인 지식)에 따라 성능 편차 발생.

딥러닝 모형 기반

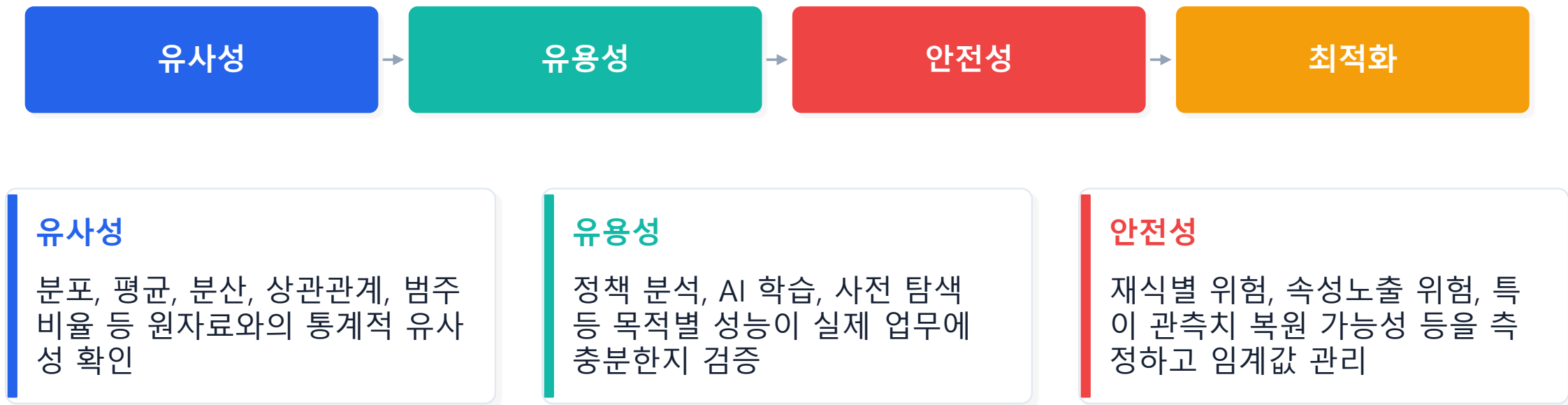
(Deep Learning Model-based, 예: CTGAN)



- **핵심 원리:** 인공신경망을 이용해 복잡한 분포 특성을 심층 학습하여 근사치 생성.
- **장점:** 정형 데이터뿐만 아니라 이미지, 텍스트 등 비정형 데이터 확장성 극대화.
- **단점:** 모델 학습 및 생성에 막대한 컴퓨팅 비용과 시간 소요 (리소스 집약적).

재현자료 평가 프레임워크: 유사성·유용성·안전성

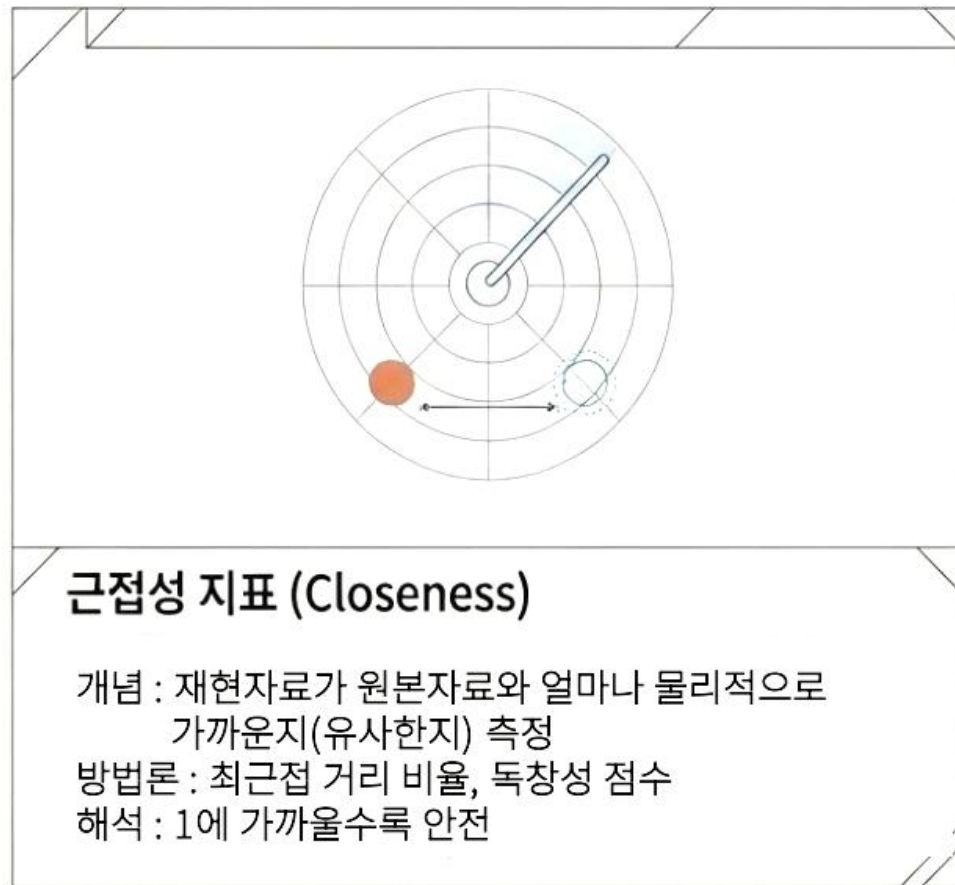
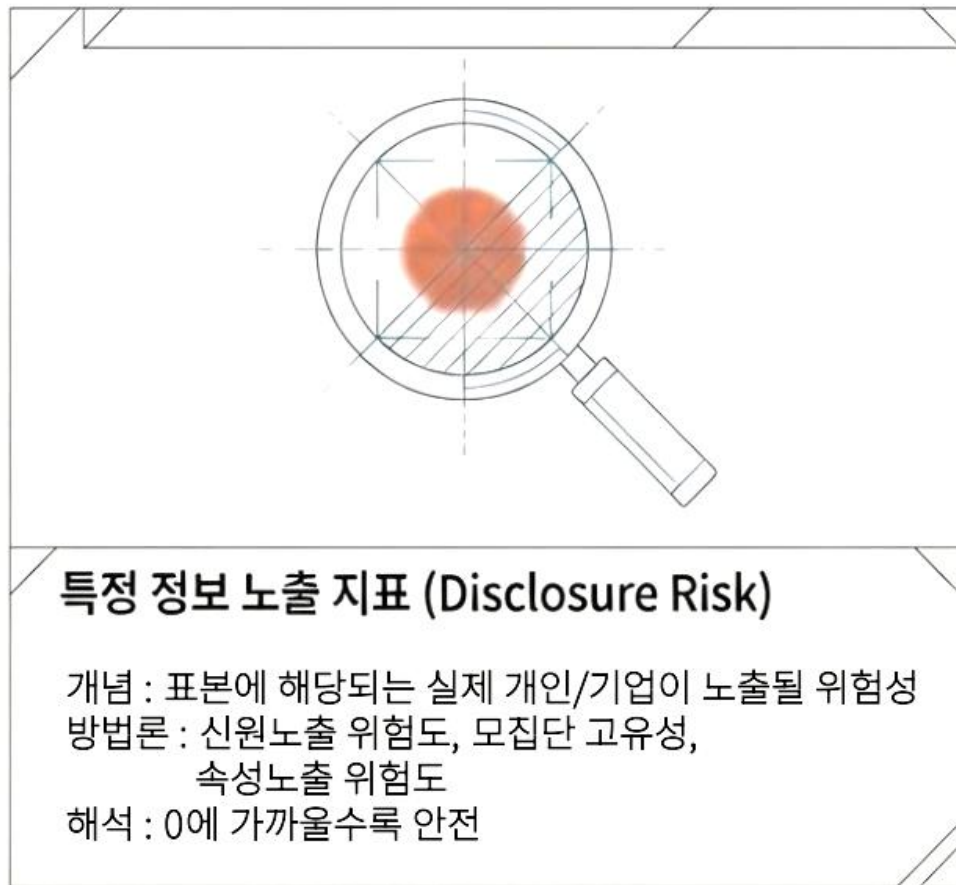
기술 개요 및 활용 영역



전략: 최대 노출위험 수준을 고정하고, 그 안에서 유용성을 극대화

재현자료 평가 프레임워크: 유용성·안전성

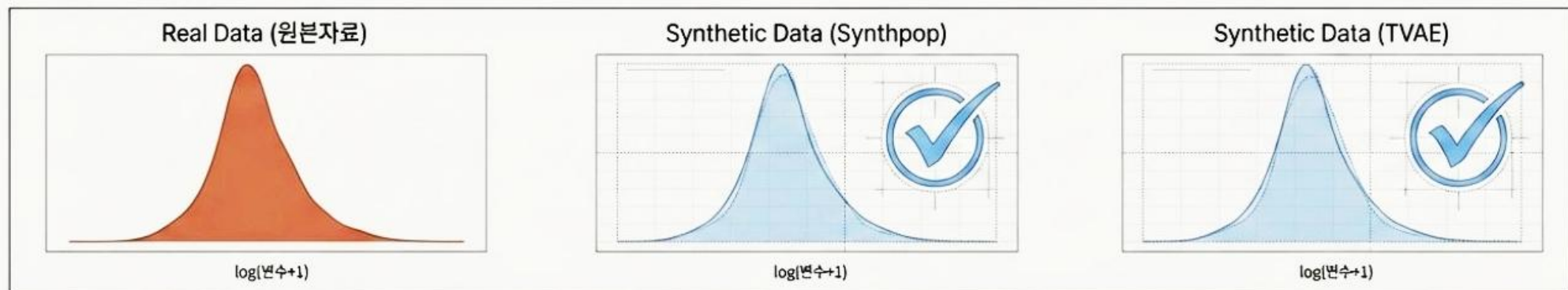
기술 개요 및 활용 영역



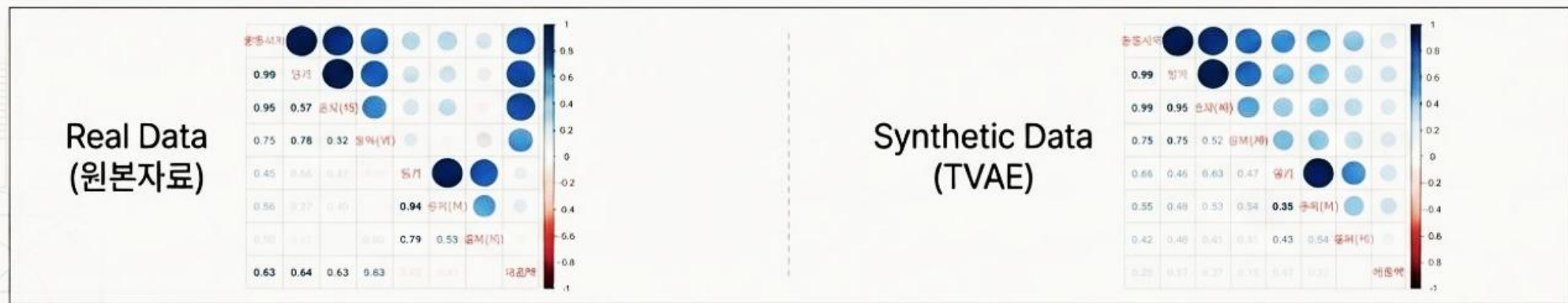
재현자료 평가 프레임워크: 유용성·안전성

기술 개요 및 활용 영역

[분포 비교]



[상관관계 비교]



재현자료 평가 프레임워크: 유용성·안전성

기술 개요 및 활용 영역



동형암호 기반 데이터 연계·활용

기술 개요 및 차별적 가치



동형암호란?

데이터를 **암호화된 상태로 두고** 직접 **덧셈·곱셈 등 연산을 수행**할 수 있는 차세대 암호 기술

기존 vs 동형암호

기존 방식 : 암호화 → 복호화 → 연산 → 재암호화
(복호화 시점 노출 위험)

동형암호 : 암호화 상태에서 바로 연산 (전 과정 보호)

차별적 가치



원본 미공개 분석

데이터 제공자가 원본을 공개하지 않고도 결과 도출 가능



다자간 안전 협업

여러 기관·기업이 데이터 결합 없이 통합 분석



법적 부담 경감

개인정보의 직접 처리가 아니므로 활용 제약 완화

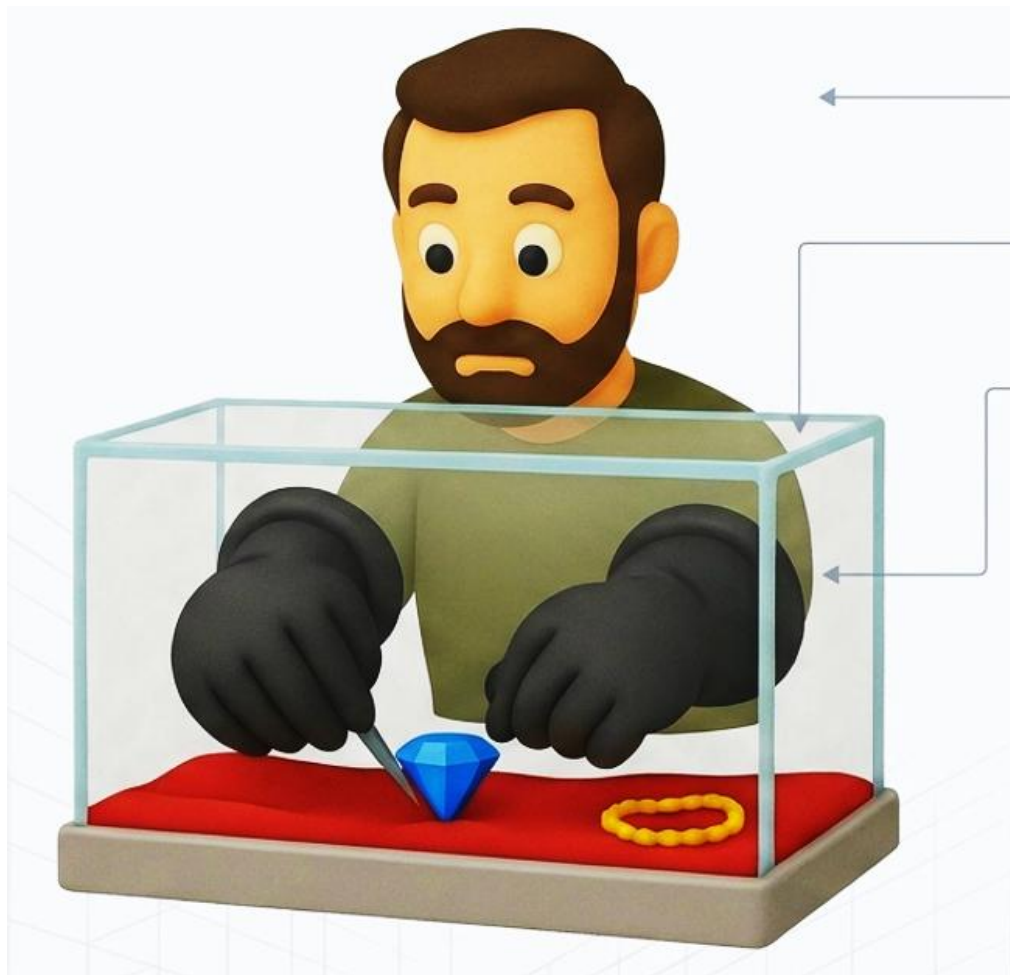


프라이버시 보존 AI

암호화 데이터로 AI 모델 학습·추론 가능

동형암호 기반 데이터 연계·활용

기술 개요 및 차별적 가치



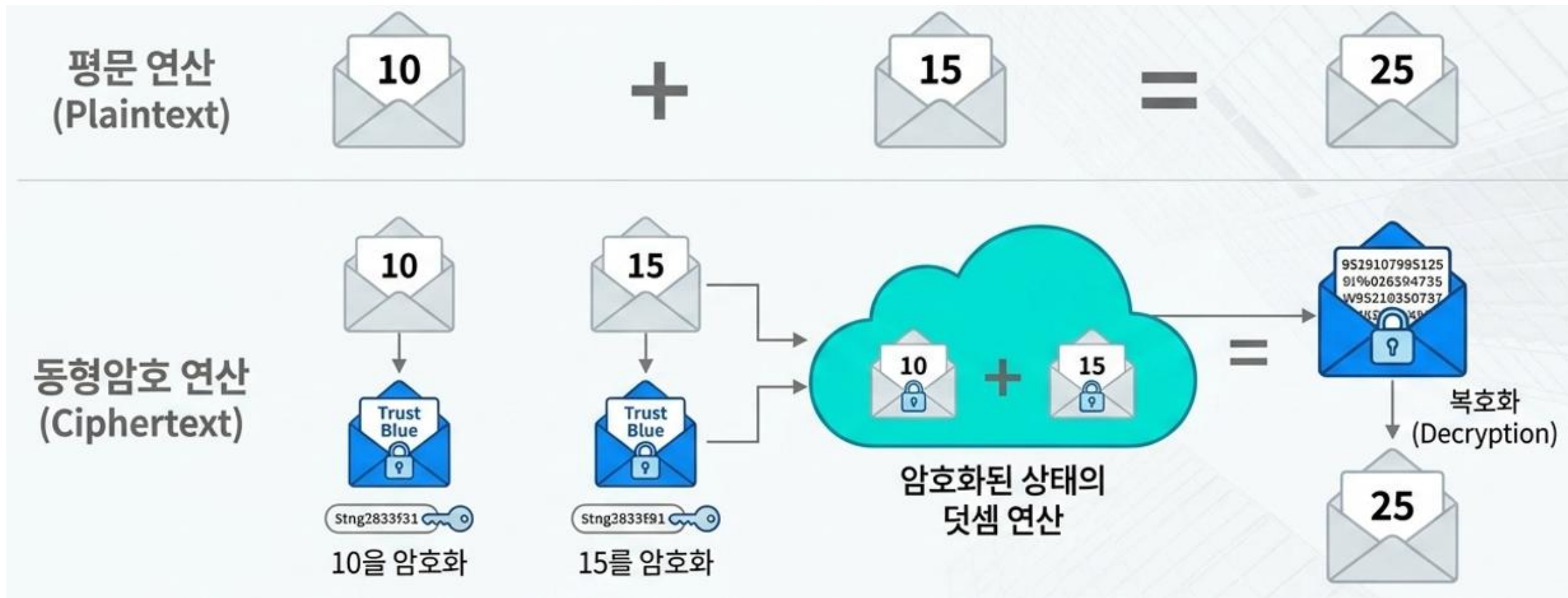
- 유리 상자 밖: 위험이 통제되지 않은 외부 환경 (클라우드, 분석가).
- 유리 상자 안: 원본 데이터(보석).
- 특수 장갑: 복호화 없이 데이터에 안전하게 조작(연산)을 가하는 동형암호 기술.

Key Characteristics

- 암호화된 상태(Ciphertext)에서 복호화(Decryption) 과정 없이 직접 수학적 연산 및 AI 추론 수행.
- 1978년 최초 개념 제안 후, 양자컴퓨터 시대에도 안전한(Quantum Resistance) 궁극의 데이터 보호 기술로 평가.

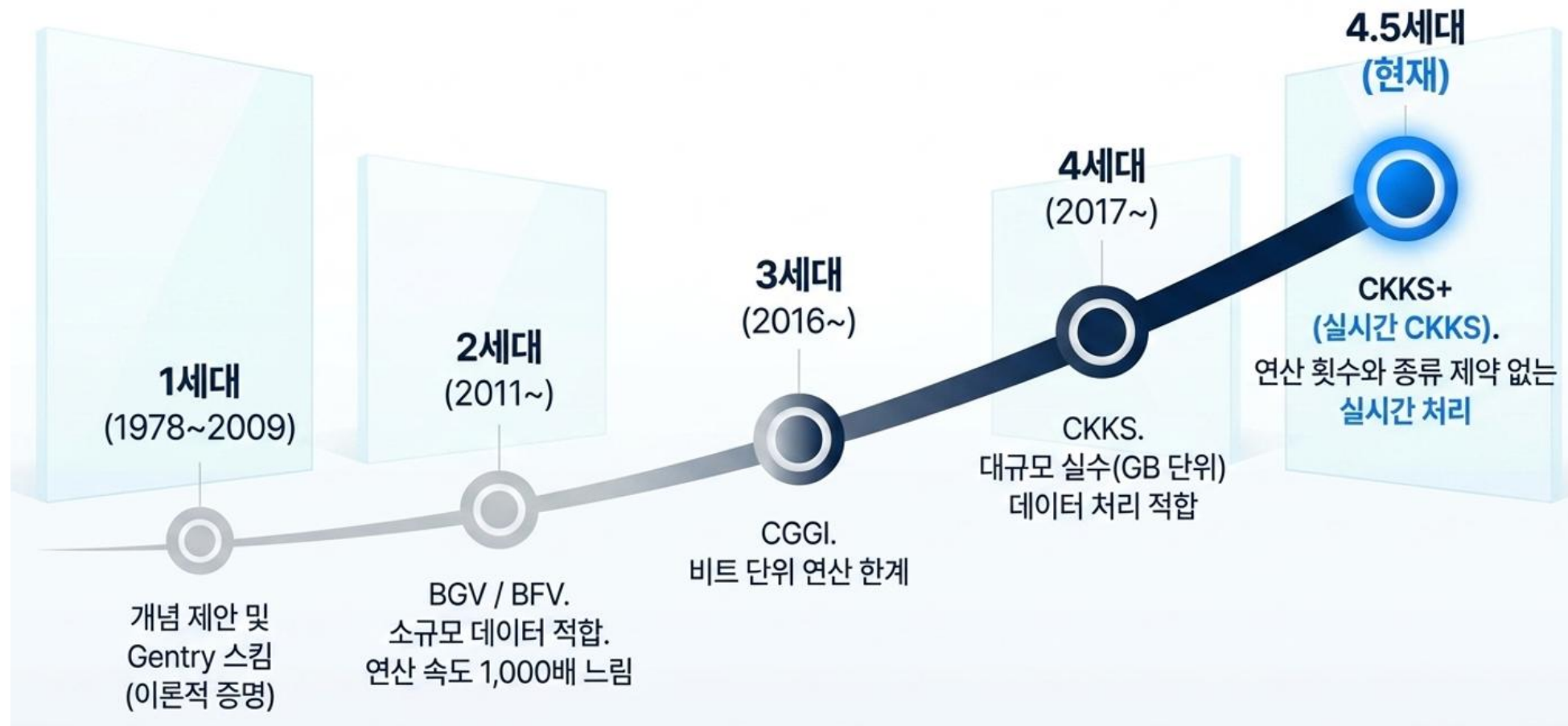
동형암호 기반 데이터 연계·활용

기술 개요 및 차별적 가치



동형암호 기반 데이터 연계·활용

기술 개요 및 차별적 가치



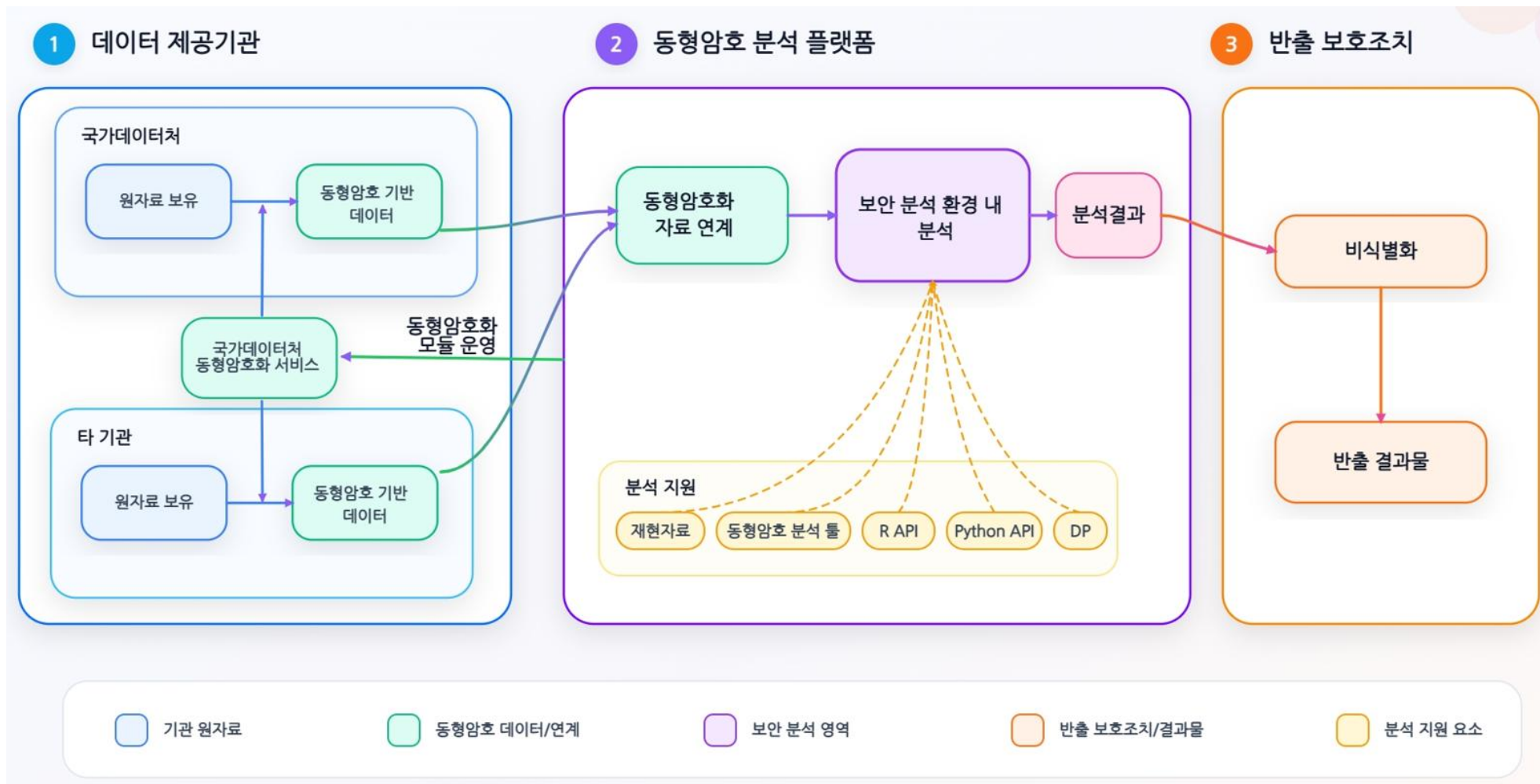
3대 R&D 기술 종합 비교

Comparison of Three Core Technologies

구분	AI 비식별화	재현(합성)데이터	동형암호
핵심 원리	식별정보 탐지·처리	원본 분포를 학습한 합성 생성	암호화 상태에서 직접 연산
주요 활용	공공·통계 데이터 공개	AI 학습·연구·교육 데이터	기관 간 결합·다자간 분석
강점	구현 용이·기존 데이터 활용	유연한 활용·무제한 생성	최고 수준 보호·원본 비공개
주의 사항	효용·재식별 위험 균형 필요	품질·안전성 평가 핵심	연산 비용·성능 최적화 필요

세 기술은 대체 관계가 아닌 보완적 관계로, 데이터 특성과 활용 목적에 따라 적절히 조합하여 적용 필요

동형암호 기반 안전 분석 · 반출



SECTION 05

향후 추진 방향 범정부 연계 활용을 위한 과제

Future Directions and Key Tasks

법정부 데이터 연계 활용 추진 필요 사항

추진전략

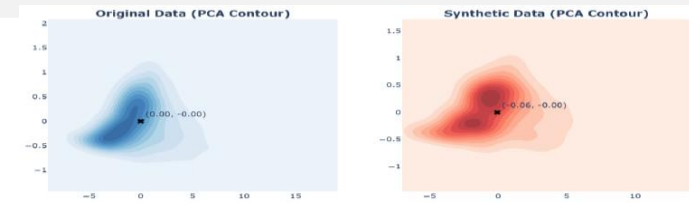
- **(필요성)** 활용 수요에 비해 폐쇄환경 의존으로 연계·분석 제약
- **(추진방안)** 동형암호 등 신기술을 통한 온라인 연계·분석환경 제공
 - ✓ (재현자료 제공) 데이터센터 이용전, 데이터를 사전 분석할 수 있는 합성(재현)자료를 제공, 효율적인 데이터 활용 환경 구축
 - ✓ (신기술 도입) "원자료는 노출되지 않는(Invisible)" 연계 체계 확립
- **(기대효과)** 민감데이터도 안전하고 신속하게 활용하는 체계 마련

추진일정

구분	내용	추진일정
단기	재현자료 활용 시스템 구축, 신기술 방법론 연구 및 실증, 동형암호·다자간연산 시범시스템 구축	~'27년
중장기	동형암호 외부활용 플랫폼 구축을 통한 온라인 분석체계 전환, 법정부 확산	~'30년

재현자료, 동형암호 등 정보보호 신기술

- ▶ **재현자료:** 원자료와 합계나 평균, 분포 등 통계적 추론값은 유사하지만 가상으로 합성된 자료



- ▶ **동형암호:** 암호화된 데이터를 복호화하지 않은 상태에서 연산을 수행할 수 있는 암호 기술



- ▶ **다자간 연산:** 데이터를 조각으로 나누어 여러 서버에 분산, 분석하는 기술

데이터보호 신기술 적용 후 데이터센터의 향후 모습

From Data Storage to Trusted Data Hub

AS-IS | 현재의 데이터센터

- ⚠ 데이터 저장·관리에 집중된 기능
- ⚠ 기관별로 분산된 사일로 형태 운영
- ⚠ 데이터 결합 시 별도 안전공간 필요
- ⚠ 비식별화는 수작업 중심, 효율 저조
- ⚠ 외부 결합·연계가 제도·기술적으로 제한

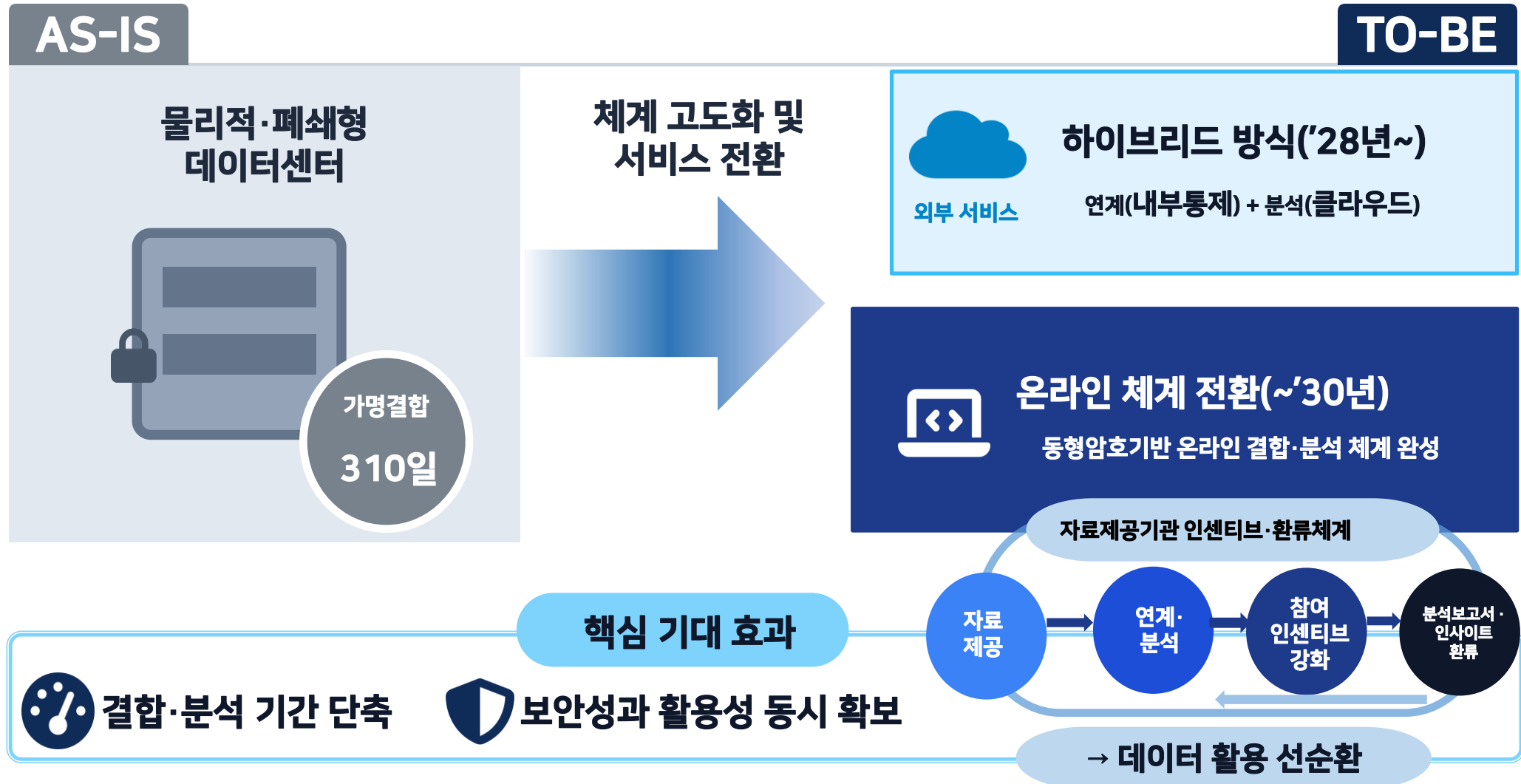


TO-BE | 미래의 데이터 허브

- ✓ 데이터 활용·가치 창출 중심으로 전환
- ✓ 범정부 통합 연계 플랫폼으로 진화
- ✓ 동형암호 기반 안전한 결합 분석 일상화
- ✓ AI 자동 비식별화·재현으로 즉시 활용
- ✓ 신뢰 기반 글로벌 데이터 협력 허브

데이터센터 = 단순 저장소를 넘어 신뢰 기반의 가치 창출 플랫폼으로 진화

법정부 데이터 연계 활용 추진 필요 사항



T H A N K Y O U

감사합니다
