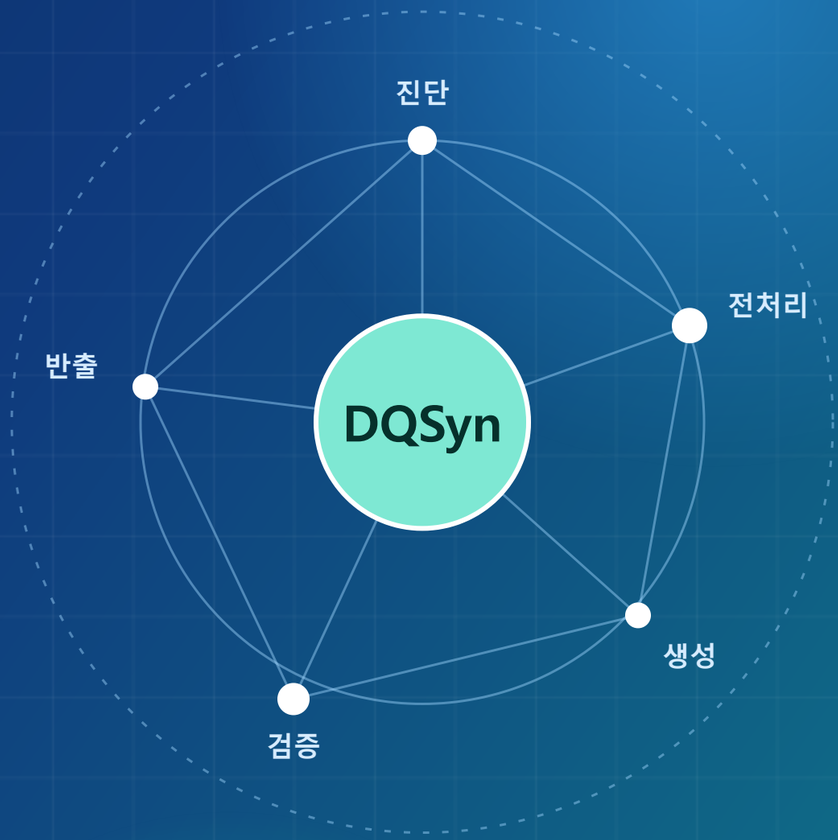
SYNTHETIC DATA CONSULTING

합성데이터 구축 컨설팅

자체 플랫폼 DQSyn 으로, 쓸 만하고 안전하다는 것을
숫자와 서류로 증명하는 합성데이터를 만듭니다.



활용 목적 진단 | 전처리 | 생성 · 검증 | 심의 · 반출 | DQSyn 플랫폼

MITSOFT IT · Data Consulting & Software

WHY

01 합성데이터, 만드는 것보다 증명이 어렵습니다

MITSOFT

원본의 통계적 특성은 살리고 실제 개인과는 연결되지 않는 데이터. **도구를 돌리는 것만으로는 끝나지 않는 일**이 현장에 남아 있습니다.

어디에 쓸 것인가

AI 학습 · 개방 · 개발 시험은 **요구 수준이 서로 다릅니다.** 목적이 흐리면 기준도 정할 수 없습니다.

원본이 지저분하면

결측 · 특수값 · 오류까지 **그대로 닮은 합성데이터**가 나옵니다.

어느 정도면 충분한가

원본과 얼마나 닮아야 하는지, **재식별 위험은 어디까지 허용**할지 판단이 막막합니다.

심의 · 반출 서류

안내서의 지표와 서식을 **손으로 맞추려면** 시간과 전문 인력이 듭니다.

"합성데이터의 가치는 생성이 아니라,
쓸 만하고 안전하다는 것을 증명하는 데서 나옵니다."

데이터 품질 전문 기업의 기준으로 생성부터 검증 · 서류까지 함께 수행합니다.

이럴 때 필요합니다

원본을 그대로 쓸 수 없을 때

외부 제공 · 개방

개인정보가 담겨 내보내기 어려운 데이터를 안전하게 공유

AI 모델 학습

데이터가 부족하거나 드문 사례가 적어 학습이 어려울 때

개발 · 시험

운영 데이터 없이 현실적인 시험 데이터가 필요할 때

CONSULTING PROCESS

02 활용 목적부터 반출까지, 여섯 단계

MITSOFT

컨설턴트가 **판단과 기준**을 맡고 DQSyn 이 **분석 · 생성 · 계산**을 맡아, 각 단계의 결과를 기록으로 남깁니다.

GOAL 목적에 맞게 쓸 수 있고, 안전성이 검증된 합성데이터

STEP 1
진단
- 활용 목적 정의
- 합성 적합성 검토
- 목표 지표 합의
- 구축 계획 수립

STEP 2
분석
- 구조 · 분포 분석
- 민감정보 식별
- 업무 규칙 확인
- 테이블 관계 파악

STEP 3
전처리
- 결측 · 이상값 처리
- 특수값 정리
- 처리 전 · 후 비교
- 전처리 확정

STEP 4
생성
- 생성 모델 선택
- 업무 규칙 반영
- 가상 식별값 대체
- 재균형 · 증강

STEP 5
검증
- 유용성 평가
- 안전성 평가
- 공격 시험
- 미달 시 개선 반복

STEP 6
반출
- 결과보고서
- 자체 검토 결과서
- 승인 · 반출 기록
- 교육 · 내재화

DQSyn
합성데이터 구축 플랫폼

파일을 올리면 **분석 → 전처리 → 생성 → 검증 → 보고서**가 한 흐름으로 이어집니다. 기관 · 기업 서버에 설치해 **원본이 밖으로 나가지 않는 환경**에서 수행합니다.

DIAGNOSIS

03 목적이 정해져야 기준도 정해집니다

MITSOFT

합성데이터로 풀 수 있는 문제인지부터 확인하고, **활용 목적에 맞는 목표 지표**를 생성 전에 합의합니다.

진단 절차

01
활용 목적 정의
누가, 어디에, 어떤 범위로 쓸지 정하고 활용 계획으로 정리합니다.

02
대상 데이터 조사
테이블 · 파일 구성과 건수, 개인정보 포함 여부를 확인합니다.

03
합성 적합성 검토
건수 · 품질 · 관계 구조를 보고 합성이 알맞은지 판단합니다.

04
위험 시나리오 정의
공개 범위와 공격자의 사전 지식을 가정해 위험을 정의합니다.

05
목표 지표 합의
유용성 · 안전성의 항목과 통과 기준을 담당자와 합의합니다.

06
구축 계획 수립
대상 · 일정 · 역할 · 산출물과 심의 절차를 계획합니다.

목적별 중점 기준

AI 모델 학습
합성데이터로 학습한 모델의 원본 예측 성능, 드문 사례 재균형

외부 제공 · 개방
재식별 위험과 공격 시험을 가장 엄격하게, 필요하면 차분프라이버시

통계 · 분석
열별 분포와 열 사이 관계의 재현, 테이블 간 참조 무결성

개발 · 시험
스키마 · 업무 규칙 준수, 열 정의만으로 만드는 테스트 데이터

진단 산출물
활용 계획서 · 대상 데이터 명세 · 목표 지표 합의서 · 구축 계획서

ANALYSIS & PREPROCESSING

04 원본을 정확히 아는 데서 시작합니다

MITSOFT

품질 진단에서 쌓은 기준으로 원본을 살피고, **생성 전에 전처리를 확정**해 오류까지 닮은 합성데이터가 나오지 않게 합니다.

구조 · 분포
자동 판독 · 탐색 분석

- 인코딩 · 구분자 · 머리행 판독
- 열 형식 추론과 분포 분석
- 결측 · 0 비율 · 이상값 탐지
- 9999999 같은 특수값 탐지

민감정보
직접 식별자 · 준식별자

- 주민번호 · 휴대폰 등 24종 탐지
- 준식별자 · 키 판별
- 담당자 검토와 분류 확정
- 개인정보 유형 분류표 작성

규칙 · 관계
업무 지식 반영

- 합계 · 곱 · 비율 규칙 발견
- 날짜 · 금액 순서 규칙 발견
- 테이블 부모 · 자식 관계
- 업종 팩으로 판정 보강

전처리 확정
처리 전 · 후 비교

- 보존 · 대체 · 자르기 · 삭제
- 자동 또는 담당자 직접 선택
- 처리 전 · 후 통계 비교
- 결정을 레시피로 재사용

자체 솔루션 연계
품질과 표준을 함께

DQube
원본 값 품질 진단으로 오류 · 중복 · 개인정보를 미리 확인

DQMeta
표준용어 · 코드 기준으로 열의 의미와 형식을 정리

DQSyn
분석 · 전처리 결과와 결정 내용을 작업 기록으로 저장

GENERATION

05 데이터에 맞는 모델로 생성합니다

MITSOFT

네 가지 모델을 표본으로 미리 학습해 본 점수로 **가장 알맞은 모델**을 고르고, 왜 골랐는지 결과 화면에서 설명합니다.

가우시안 코풀라
열별 분포와 상관을 재현하는 빠른 기본 모델

GMM 코풀라
봉우리가 여럿인 복잡한 숫자 분포에 적합

CART 코풀라
범주가 많은 데이터의 조합 관계를 재현

코풀라-VAE
딥러닝으로 비선형 관계와 범주 조합을 학습

차분프라이버시
더 강한 보호가 필요할 때 선택, 예산을 보고서에 기록

후보 학습	모델 선택	생성	규칙 적용	식별값 대체
• 표본으로 사전 학습 • 후보별 점수 비교	• 자동 또는 직접 지정 • 선택 근거 확인	• 원본 평가용 분리 • 필요 수량만큼 생성	• 계산 규칙 재계산 • 순서 규칙 유지	• 이름 · 연락처 가상값 • 키는 겹치지 않게

여러 테이블
부모 · 자식 관계와 '부모 합계 = 자식 합계' 같은 교차 규칙 유지

재균형 · 증강
적은 집단을 늘려 AI 모델의 사각지대를 줄이고 결과를 나란히 비교

텍스트 합성
문장 속 개인정보를 가린 새 문장을 LLM 학습용 JSONL 로 제공

VALIDATION

06 따로 떼어 둔 평가 데이터로 판정합니다

MITSOFT

원본을 얼마나 닮았는지(유용성)와 원본 속 개인을 알아낼 수 없는지(안전성)를 측정하고, **목표에 못 미치면 고쳐서 다시 검증**합니다.

분포 · 관계

KS · JSD, 상관 · 크래머 V, 참조 무결성

구별 · 예측

pMSE · 판별 AUC, 합성으로 학습한 모델의 원본 성능 (TSTR)

재식별 위험

원본 복제, DCR · NNDR, 준식별자 고유성, I-다양성 · t-근접성

공격 시험

소속 추론(MIA), 연결 위험 (CAP), 단독 식별 · 연결 · 추론 공격

안내서 대응

개인정보위 안내서 지표 계산, 체크리스트 19항목 판정

목표에 못 미치면 자동 개선

문제 기록

위험한 레코드를 찾아 새 레코드로 교체

분포 보정

분위 매핑과 순위 재배열로 분포를 맞춤

모델 재학습

다시 학습하거나 다른 모델로 전환, 준식별자 정밀도 조정

컨설턴트의 역할

- ✓ 지표 결과의 의미 해석
- ✓ 유용성 · 안전성 균형 조정
- ✓ 잔여 위험과 활용 범위 정리
- ✓ 다른 도구 합성데이터도 같은 기준 검증

DELIVERABLES

07 심의와 반출에 필요한 자료까지

MITSOFT

개인정보보호위원회 합성데이터 안내서를 기준으로, **기관의 심의 · 승인 절차에 바로 쓸 수 있는 문서**를 합성데이터와 함께 준비합니다.

REPORT
최종 결과보고서

- 원본 분석 · 전처리 · 생성 방법
- 유용성 · 안전성 결과와 개선 이력
- 활용 범위 · 안내서 대응
- HTML · Word · 한글 형식

REVIEW
심의 · 검토 서류

- 안전성 자체 검토 결과서
- 원본데이터 명세서
- 개인정보 유형 분류표
- 활용 계획서 · 생성 기초자료 명세서

DELIVERY
합성데이터 · 재현 자료

- CSV · 엑셀 · JSON · SQL · Parquet
- 재현 패키지(ZIP)
- 보고서 ID · 해시
- 반출 대장 기록

STEP 1
검토용 작성
승인 전 문서에는 '검토용' 표시

STEP 2
심의 · 승인
승인자가 결과를 확인하고 승인

STEP 3
반출 · 기록
승인 후 내려받고 반출 대장에 기록

STEP 4
원본 파기
보관 기간 관리와 즉시 파기

HOW WE WORK

08 상황에 맞게 시작하고, 기관에 남깁니다

MITSOFT

작은 데이터로 효과를 먼저 확인하고, 본 구축을 거쳐 **담당자가 직접 운영할 수 있을 때까지** 지원합니다.

공공데이터 개방
개인정보 때문에 닫혀 있던 데이터의 공유 · 개방 검토

AI 학습 데이터
부족하거나 치우친 학습 데이터 보완

개발 · 시험 환경
운영 데이터를 대신하는 시험 데이터 구축

분석 · 협업
외부 전문가와 안전하게 공유하는 분석용 데이터

DQSyn 컨설팅과 플랫폼을 함께 세 가지 방식으로 함께합니다

PILOT
시범 적용
보유 데이터 일부로 DQSyn을 직접 돌려 보고, 결과보고서로 효과를 확인합니다.

BUILD
구축 컨설팅
활용 목적 진단부터 생성 · 검증 · 심의 서류까지 컨설턴트가 함께 수행합니다.

TRANSFER
운영 내재화
플랫폼 설치, 담당자 교육, 레시피 정리로 기관이 스스로 반복하게 합니다.

폐쇄망 설치
외부 연결 없이 기관 서버에서 운영

권한 분리
관리자 · 승인자 · 작업자 3역할

감사 기록
생성 · 승인 · 내려받기 전부 기록

원본 관리
보관 기간 설정, 민감 열 가림

Windows 서버 설치 · 웹 브라우저 접속 · CPU 만으로 동작 · 딥러닝 모델 선택 설치

EXPECTED EFFECTS

09 기대효과와 엠아이티소프트의 강점

MITSOFT

합성데이터 생성 도구를 직접 만든 회사가, **데이터 품질 컨설팅의 기준**으로 구축 전 과정을 함께합니다.

01

못 쓰던 데이터 활용

개인정보 때문에 묶여 있던 데이터를 분석 · 학습 · 공유에 활용

02

검증된 안전성

재식별 위험과 공격 시험 결과를 수치와 문서로 확인

03

구축 기간 단축

분석 · 검증 · 서류 작성을 자동화해 반복 작업 축소

04

재사용되는 자산

절차 · 레시피 · 재현 자료가 기관에 남아 반복 구축 가능

왜 엠아이티소프트인가

자체 개발 플랫폼 DQSyn
기관 요구를 플랫폼과 절차에 직접 반영할 수 있는 개발사

데이터 품질 · 표준화 전문
공공데이터 품질진단 · 품질인증 · 표준화 컨설팅의 품질 기준

안내서 기준 산출물
개인정보위 안내서의 지표와 검토 서식에 맞춘 결과 문서

시범 적용 기관 모집 중
DQSyn 은 출시 준비 중인 제품으로, 시범 적용을 함께 진행합니다

MITSOFT IT · Data Consulting & Software

TEL 02-719-0380
E-MAIL sales@mitsoft.kr
WEB www.mitsoft.kr