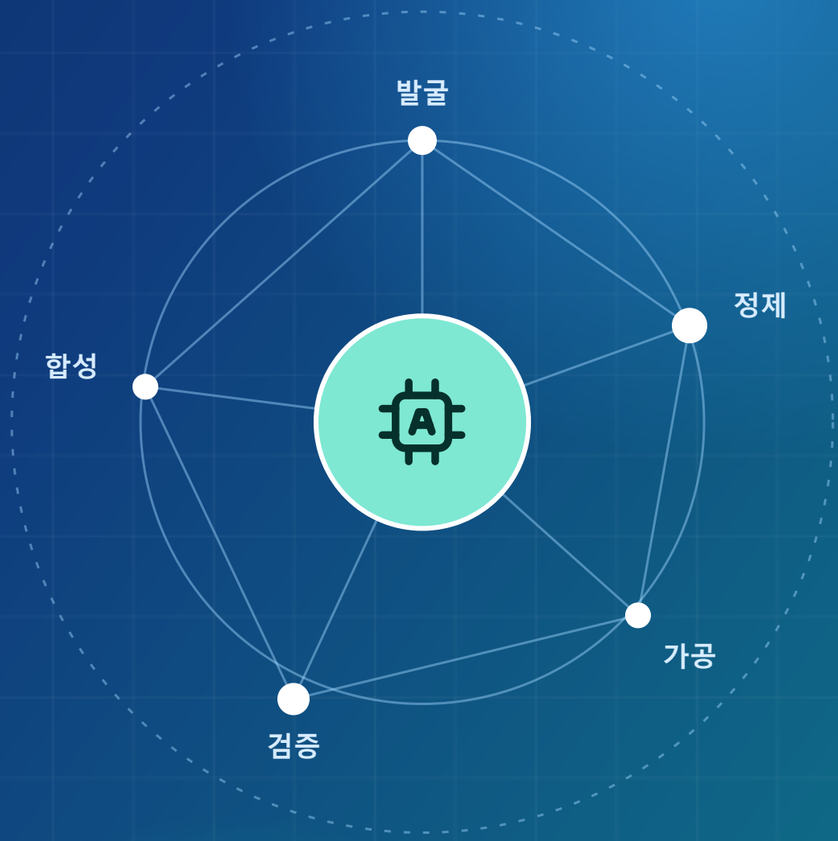
AI TRAINING DATA ENGINEERING

AI 학습용
데이터 구축

학습에 쓸 데이터를 찾아내고, 깨끗하게 정제하고,
목적에 맞게 가공 · 검증해 AI 가 바로 학습하는 데이터를 만듭니다.



데이터 발굴 | 정제 · 전처리 | 가공 · 라벨링 | 품질 검증 | DQSyn 합성데이터

MITSOFT IT · Data Consulting & Software

WHY

01 AI 의 성능은 데이터가 결정합니다

MITSOFT

모델은 공개되어 누구나 쓰지만, **우리 업무에 맞는 좋은 학습 데이터**는 직접 만들어야 합니다. 그 과정에서 현장은 같은 벽에 부딪힙니다.

무엇을 모아야 할지

기관 · 기업 안에 데이터는 많지만 **어떤 데이터가 학습에 쓸모 있는지** 판단하기 어렵습니다.

지저분한 원천 데이터

결측 · 중복 · 오류 값, 깨진 글자 · HTML 찌꺼기가 섞여 **그대로는 학습에 쓸 수 없습니다.**

개인정보 · 권리 문제

개인정보와 저작권 때문에 **좋은 데이터일수록 밖으로 꺼내기 어렵습니다.**

부족하고 치우친 데이터

드문 사례는 몇 건뿐이라 **모델이 꼭 맞혀야 할 경우를 배우지 못합니다.**

"좋은 학습 데이터는 많이 모은 데이터가 아니라,
목적에 맞게 고르고 · 다듬고 · 검증한 데이터입니다."

데이터 품질 · 표준화 전문 기업의 방법으로 AI 학습용 데이터를 만듭니다.

정책 환경

AI 시대, 데이터 정책은 'AI-Ready' 로 바뀌고 있습니다

공공데이터 AI-Ready 품질관리

AI 가 읽고 학습할 수 있는 형태로 데이터를 정비 · 개방하는 방향으로 전환

AI 친화 · 고가치 데이터 개방

AI 기업 수요에 맞는 핵심 데이터의 발굴과 개방 확대

합성데이터 활용 제도화

개인정보보호위원회 합성데이터 안내서 - 유용성 · 안전성 검증 기준 제시

END-TO-END PROCESS

02 발굴부터 활용까지, 한 회사가 끝까지

MITSOFT

AI 학습용 데이터 구축의 **여섯 단계**를 나누지 않고 한 흐름으로 수행해, 단계 사이에서 품질이 새지 않게 합니다.

GOAL AI 모델이 바로 학습하고, 성능으로 증명되는 데이터

STEP 1 발굴

- 활용 목적 · 과제 정의
- 보유 데이터 조사
- 가치 · 권리 검토
- 구축 계획 수립

STEP 2 수집

- 원천 데이터 확보
- 수집 기준 · 분포 설계
- 메타데이터 기록
- 수집 진척 · 분포 점검

STEP 3 정제

- 결측 · 오류 · 중복 처리
- 표준화 · 형식 통일
- 개인정보 비식별
- 처리 전 · 후 기록

STEP 4 가공

- 가공 지침 정의
- 라벨링 · 어노테이션
- 단계별 검수
- 작업 일관성 관리

STEP 5 검증

- 품질 지표 측정
- 학습 유효성 시험
- 품질 보고서
- 외부 검증 대응

STEP 6 활용

- 납품 · 개방 등록
- 모델 학습 연계
- 갱신 · 유지관리
- 운영 교육 · 매뉴얼

DQSyn
합성데이터 구축 플랫폼

데이터가 **부족하거나, 치우쳐 있거나, 개인정보 때문에 쓸 수 없을 때**는 자체 플랫폼 DQSyn 으로 원본의 특성을 닮은 합성데이터를 만들어 정제 · 가공 단계에 더합니다. 유용성 · 안전성 검증과 결과보고서까지 함께 나옵니다.

DATA DISCOVERY

03 쓸 데이터를 찾아내는 일, 발굴

MITSOFT

데이터 표준화 · 메타데이터 정비 경험으로 **기관 · 기업이 가진 데이터를 빠짐없이 조사**하고, 학습 가치와 위험을 함께 따져 구축 대상을 정합니다.

발굴 절차

01
활용 목적 · 과제 정의
어떤 AI 를 무엇에 쓸지, 모델이 맞혀야 할 대상과 성능 목표를 먼저 정합니다.

02
보유 데이터 인벤토리
DB · 파일 · 문서 · 이미지를 메타데이터 기준으로 조사해 후보 목록을 만듭니다.

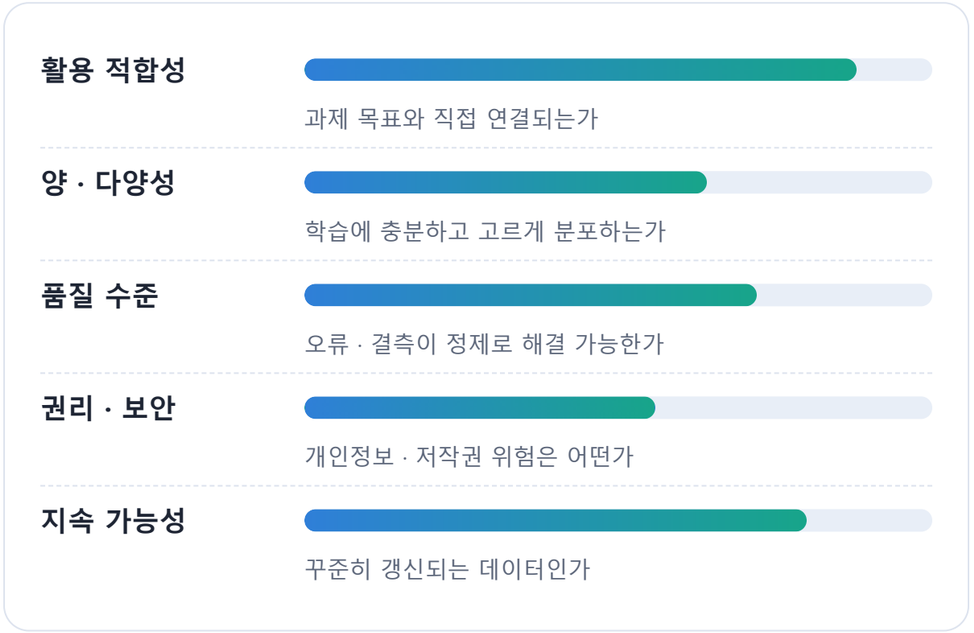
03
가치 · 품질 평가
양 · 다양성 · 품질 · 갱신 주기를 점수로 비교해 우선순위를 정합니다.

04
권리 · 개인정보 검토
저작권 · 이용 조건과 개인정보 포함 여부를 확인해 비식별 · 합성 대상을 가립니다.

05
구축 설계
수량 · 클래스 분포 · 스키마 · 가공 유형을 정한 구축 계획서를 만듭니다.

06
부족분 보완 계획
모자라거나 치우친 부분은 추가 수집 또는 DQSyn 합성으로 채울 계획을 세웁니다.

후보 데이터 평가 예시



발굴 산출물
데이터 인벤토리 · 후보 평가표 · 구축 계획서(수량 · 분포 · 스키마 · 가공 지침 초안)

CLEANSING & PREPROCESSING

04 품질 전문 기업의 정제 · 전처리

MITSOFT

공공데이터 품질진단과 표준화를 해 온 방법 그대로, **데이터 유형마다 정해진 기준과 자체 도구로** 학습 전에 문제를 걷어냅니다.

정형 데이터
DB · CSV · 엑셀

- 결측 · 이상값 · 특수값 처리
- 중복 행 · 키 중복 제거
- 코드값 · 단위 · 날짜 형식 표준화
- 업무 규칙 위반 값 검출

텍스트
문서 · 게시글 · 상담 기록

- 깨진 글자 · 보이지 않는 문자
- HTML 찌꺼기 · 상용구 제거
- 제각각인 빈 값 표기 통일
- 중복 · 유사 문장 정리

이미지 · 문서 · 영상
비정형 데이터

- 해상도 · 형식 기준 선별
- 중복 · 저품질 파일 제외
- 문서 텍스트 자동 추출
- 파일별 메타데이터 정리

개인정보 보호
모든 유형 공통

- 주민 · 연락처 · 계좌 등 탐지
- 가명 · 마스킹 · 범주화
- 재식별 위험 점검
- 필요하면 합성데이터로 대체

자체 솔루션으로
사람 손에만 맡기지 않습니다

DQube
값 품질 진단과 AI-Ready 점검 - 깨진 글자 · 개인정보 · 중복을 DB · 파일에서 찾고 고침

DQMeta
표준용어 · 코드 기준으로 컬럼 이름과 값을 맞춰 데이터셋 간 의미를 통일

DQSyn
자동 EDA · 전처리 확정 - 처리 전 · 후 통계를 비교하고 기록으로 남김

LABELING & BUILD

05 목적에 맞는 가공과 이중 검수

MITSOFT

가공 지침을 먼저 세우고 작업자 교육 · 단계별 검수로 **누가 작업해도 같은 기준의 라벨**이 나오게 합니다.

분류 · 속성
문서 · 기록 · 이미지의 유형과 속성 값 부여

영역 지정
이미지 · 문서 속 객체와 영역 표시

개체 · 관계
문장 속 이름 · 기관 · 날짜와 그 관계 표시

질의응답 · 요약
LLM 학습용 질문 · 답변, 요약문 작성

정형 정답값
예측 · 분류 모델용 목표 변수 정의와 검증

지침 정의	작업자 교육	1차 가공	검수	전문가 확인
· 라벨 체계 · 판단 기준	· 지침 교육 · 시험 작업	· 자동 추출 · 사전 라벨	· 검수자 전수 · 표본 검사	· 분야 전문가 최종 확인
· 헷갈리는 사례 모음	· 합격자만 투입	· 작업자 확인 · 보정	· 반려 · 재작업	· 지침 보완 환류
· 품질 기준 합의	· 작업 도구 안내	· 진척 · 분포 확인	· 오류 유형 집계	· 최종 승인

작업 일관성 관리
작업자 간 일치도를 재고, 차이가 큰 유형은 지침을 보완합니다.

진척 · 이력 관리
건별 작업자 · 검수자 · 반려 이력을 남겨 추적할 수 있게 합니다.

분포 관리
구축 중에도 클래스 분포를 확인해 치우침을 조기에 바로잡습니다.

SYNTHETIC DATA · DQSyn

06 DQSyn 으로 만드는 합성데이터

MITSOFT

원본의 통계적 특성은 살리고 실제 개인과는 연결되지 않는 합성데이터로, **부족 · 불균형 · 개인정보** 세 가지 문제를 한 번에 풉니다.

부족한 데이터를 늘림
원본과 같은 분포 · 관계를 가진 데이터를 필요한 만큼 생성

치우친 분포를 고름
드문 사례를 늘리는 재균형 · 증강으로 모델의 사각지대 축소

개인정보 없이 활용
재식별 위험을 검증한 데이터로 외부 학습 · 개방 · 공유

LLM 학습용 텍스트
문장 속 개인정보를 가리고 새 문장을 만들어 JSONL 로 제공

DQSyn 엠아이티소프트 합성데이터 구축 플랫폼

CSV · 엑셀을 올리면 분석부터 검증 · 결과보고서까지 자동으로

01 파일 올리기
여러 테이블 · 관계 자동 인식

02 데이터 확인
분포 · 민감정보 · 업무 규칙 분석

03 전처리 확정
처리 전 · 후 비교 후 확정

04 생성 · 검증
모델 4종 자동 선택, 미달 시 자동 개선

05 결과 · 보고서
합성데이터 · 보고서 · 검토 결과서

유용성 검증
분포 · 열 관계 재현, 원본과의 구별 정도, 합성데이터로 학습한 모델의 원본 예측 성능

안전성 검증
원본 복제 · 근접 기록, 준식별자 고유성, 소속 추론 · 연결 · 추론 공격 시험

합성데이터
CSV · 엑셀 · JSON · SQL · Parquet

결과보고서
HTML · Word · 한글, 반복 개선 이력

심의 서류
안전성 자체 검토 결과서 · 명세서

재현 자료
재현 패키지 · 보고서 해시 · 반출 대장

개인정보위 안내서 지표 자동 산출 | 안전성 자체 검토 결과서 | 폐쇄망 서버 설치 | 테스트 데이터 생성

QUALITY ASSURANCE

07 숫자로 증명하는 데이터 품질

MITSOFT

구축이 끝난 뒤가 아니라 **구축하는 동안 계속** 품질을 재고, 마지막에는 실제 모델 학습으로 데이터의 쓸모를 확인합니다.

기준 적합성

구축 계획서의 수량 · 형식 · 스키마를 지켰는가

완전성

빠진 항목 · 파일 없이 모두 갖추었는가

정확성

값과 라벨이 사실 · 지침과 맞는가 (구문 · 의미)

다양성 · 분포

클래스 · 속성이 고르게 분포하는가

유효성

이 데이터로 학습한 모델이 목표 성능을 내는가

검증 방법

자동 전수 검사

형식 · 필수값 · 코드값 · 중복 · 개인정보를 DQube로 전체 점검

표본 정밀 검사

통계적 표본을 뽑아 값 · 라벨 정확도를 검수자가 직접 확인

학습 유효성 시험

기준 모델을 학습시켜 성능 지표로 판정, 합성데이터는 원본 예측 성능(TSTR)까지 확인

품질 산출물

- ✓ 데이터 품질 검증 보고서
- ✓ 데이터셋 명세서 · 메타데이터
- ✓ 가공 지침서 · 검수 이력
- ✓ 외부 품질 검증 대응 자료
- ✓ 합성데이터 검증 결과보고서

DELIVERY & USE

08 바로 쓰는 형태로 넘기고, 계속 쓰게 합니다

MITSOFT

데이터만 넘기지 않습니다. **학습 · 개방 · 갱신**까지 이어지도록 형식과 체계를 함께 만듭니다.

DATASET
학습용 데이터셋

- 학습 · 검증 · 시험 분할 세트
- CSV · JSON · JSONL · Parquet 등 요구 형식
- 원천 · 라벨 파일과 매핑 정보
- 데이터셋 명세서 · 메타데이터
- 구축 결과 · 품질 보고서

OPEN & SHARE
개방 · 공유

- 공공데이터 개방 형식으로 정비
- 개방 데이터셋 등록 · 설명 작성 지원
- DataPrism 으로 오픈API 제공
- 합성데이터로 안전한 외부 공유
- 활용 사례 · 이용 안내 자료

OPERATION
갱신 · 유지관리

- 정제 · 가공 규칙을 재사용 가능한 절차로
- DQube 로 신규 데이터 품질 상시 점검
- DQSyn 정제 레시피로 같은 구조 자동 처리
- 담당자 교육 · 운영 매뉴얼
- 데이터 추가 · 재학습 주기 설계

DQube
품질 진단
값 품질 · AI-Ready 점검

DQMeta
표준 · 메타
용어 · 코드 · 메타데이터 관리

DQSyn
합성데이터
생성 · 검증 · 결과보고서

DataPrism
개방 · 오픈API
코딩 없이 표준 API 제공

EXPECTED EFFECTS

09 기대효과와 엠아이티소프트의 강점

AI 학습용 데이터의 **발굴 · 정제 · 구축**을 데이터 품질 전문 기업의 기준과 자체 솔루션으로 수행합니다.

01

AI 성능 향상

목적에 맞게 고르고 다듬은 데이터로 모델의 정확도와 안정성 향상

02

안전한 데이터 활용

비식별 · 합성데이터로 개인정보 위험 없이 학습 · 개방 · 공유

03

구축 기간 · 비용 절감

자동 진단 · 전처리 · 합성으로 사람 손이 가는 반복 작업 축소

04

검증된 품질 · 자산화

품질 보고서와 재사용 가능한 절차로 데이터가 기관의 자산으로 남음

왜 엠아이티소프트인가

데이터 품질 · 표준화 전문
공공데이터 품질진단 · 품질인증 · 표준화 컨설팅을 해 온 품질 기준

데이터 발굴 · 개방 경험
공공데이터 · 국가중점데이터 발굴 · 개방, 문서 · 영상 등 비정형 데이터 정비

정제부터 합성까지 자체 솔루션
DQube 3.0 GS인증 1등급, DQMeta · DQSyn · DataPrism 직접 개발

AI · 데이터 전문 인력
데이터 분석 · AI 모델링, 품질관리 전문가가 직접 수행

MITSOFT IT · Data Consulting & Software

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