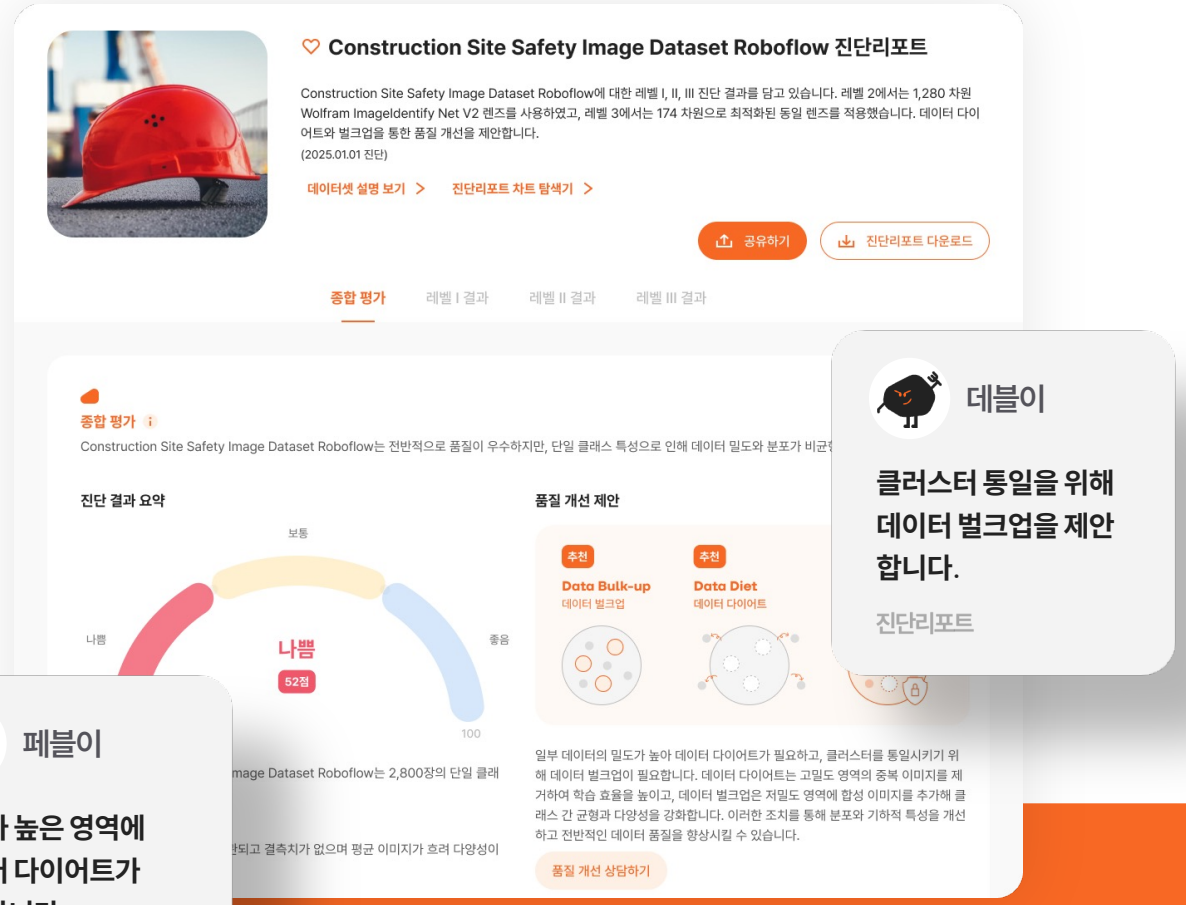
주식회사 페블러스

AI 성능을 200% 상승시키는 데이터 품질 관리 가이드북



데블이

클러스터 통일을 위해 데이터 벌크업을 제안합니다.

진단리포트

페블이

밀도가 높은 영역에 데이터 다이어트가 필요합니다.

진단리포트

기존에서 완전히 벗어나, 새로운 전략이 필요합니다.

파라미터 조정, 모델 변경 등 모델링 중심 품질 개선을 해도,
정작 본질적인 문제 '데이터의 품질'이 떨어진다면 어떨까요?
아무리 똑똑한 AI를 개발하더라도 알고리즘은 점점 망가집니다.

AI가 급속도로 발전하는 시대, 이제는 AI-Ready Data를 준비해야 합니다.
그러나 여전히 많은 기업들이 한계에 갇혀 있습니다.

내부 노하우에만 의존하여 품질을 점검하고 있습니다.

정형 데이터는 규칙 기반 검사에 머무릅니다.
데이터 간 관계나 정합성을 깊게 확인하기란 어렵죠.

1단계 PoC(개념 검증)에만 머무르고 있습니다.

대부분의 기업은 학습 목적에 맞지 않는 데이터를 수집하거나, 레이블 형태가
부적절하거나, 다양성이 부족합니다. 개념 검증 단계에만 머무르기 때문이죠.

Gartner

Source: Gartner

Pebblous

데이터 품질관리 기술도 그에 따라 변화합니다.

01 데이터 비중 변화

이제는 텍스트 데이터 관리만으론 부족합니다. **영상, 센서, 음성 데이터**의 비중이 크게 확대되고 있습니다. 그에 따라 그데이터의 품질 관리 방식도 복잡해지는 것이죠.

02 의미 기반 품질 관리

벡터 임베딩과 온톨로지를 동시에 활용한 의미 기반 품질 관리는 데이터 간 '의미적 유사도'를 계산해 규칙 기반 품질 관리만으로는 잡히지 않는 오류, 누락, 패턴까지 발견할 수 있습니다.

03 안전한 데이터 확보

민감정보를 안전하게 다루면서도 데이터 활용도를 높이려면? **합성 데이터(재현 데이터)를 통한 가명 처리**가 필요합니다. 원본과 유사한 패턴을 유지하고, 개인정보 노출 위험을 크게 낮추죠.

Pebblous

버려지는 합성데이터가 아닌, 정확한 합성데이터 생성 비결

AI-Ready Data에 적합한 데이터를 만드는 전략 중 합성데이터 생성이 대표적이죠.
AI에게 적합한 데이터, 정확한 합성데이터를 생성하려면 이 세 가지 조건을 지키셔야 합니다.

물리적 타당성 & 도메인 적합성

데이터는 '현실에서' 활용되어야 의미가 있습니다. 만약 현실에선 거의 일어나지 않는 장면을 생성한다면? GPU만 낭비하게 됩니다.
물리적 법칙, 도메인 제약을 꼭 체크하세요!

다양성 & 대표성 확보

합성데이터를 생성할 경우 실제 데이터의 특성을 반영합니다.
이때 데이터셋이 가진 '편향'을 그대로 반영한다면 오히려 편향된 AI가 만들어지고, 성능이 저하됩니다.
희귀, 소수 케이스를 포함하여 다양한 데이터를 의도적으로 보강해야 편향을 줄일 수 있어요.

평가용 합성데이터 활용

평가 데이터란 AI의 성능이 적절한지 평가하는 용도의 데이터를 말합니다. '시험 문제'와 같은 역할이죠.
AI는 좋은 문제를 잘 풀어낼수록 한층 똑똑해지기 때문에, 평가 데이터의 품질도 반드시 신경쓰셔야 합니다.

Knowhow

AI 데이터의 한계를 극복한 노하우

A기업

농업용 로봇의 야생동물 및
안전사고 객체 탐지 솔루션

Pebblous

| 문제 상황

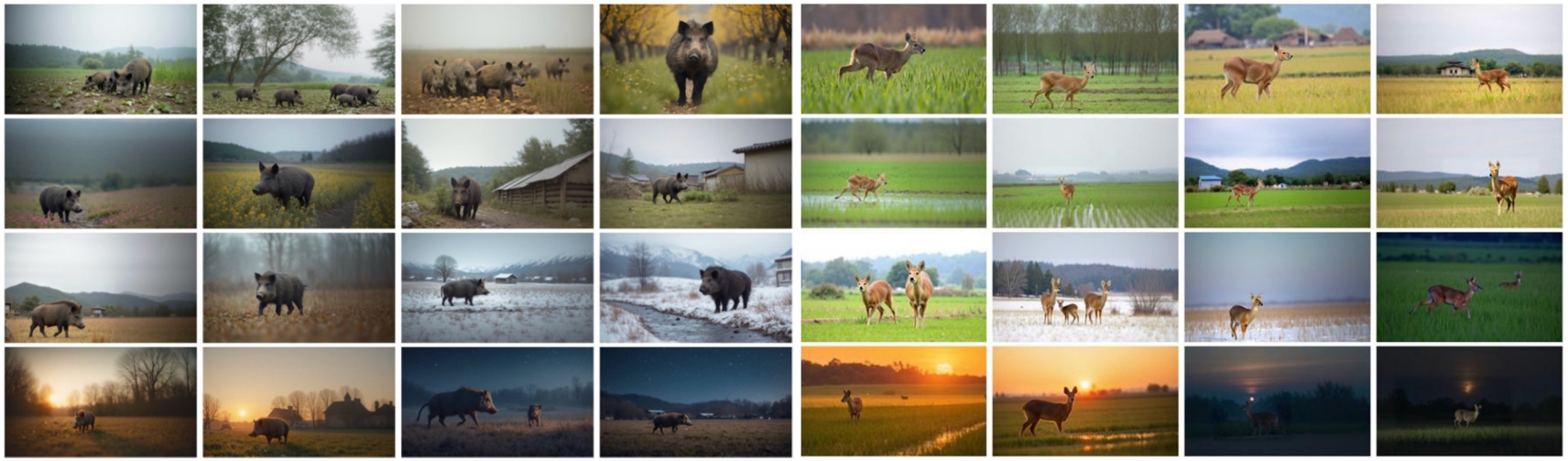
야생동물 이미지가 현저히 부족했습니다.
특히 그 중 '고라니' 이미지가 부족한 상황에서 해외 기반 생성형 AI로 고라니를 생성하면 사슴으로 잘못 생성하는 경우가 다반사였습니다. 또한 국내 농촌 환경에 맞지 않는 비현실적인 이미지도 다수 생성되었죠.

| 해결책

클래스별 데이터 불균형을 분석했고, 비현실적인 이미지를 우선 제거했습니다. 이후 한국 야생동물을 그대로 반영할 수 있도록 CG와 생성형 AI를 결합한 합성 파이프라인을 구축했습니다.

| 결과

다양한 농촌 배경을 반영한 합성데이터 900장을 생성했습니다.
고라니, 멧돼지도 문제 없이, 품질 높은 데이터로 생성할 수 있었습니다.



Knowhow

AI 데이터의 한계를 극복한 노하우

B기업

산불 화재 감지 AI 솔루션

Pebblous

| 문제 상황

4백만 장의 실제 데이터를 수집했으나, 90%가 사용이 불가할 정도로 품질이 낮았습니다. 야간 데이터가 거의 없었고, 낮 이미지에서 밤 이미지로 합성을 시도했으나 품질이 저조한 데이터만 생성되었습니다.

| 해결책

야간 환경이 반영된 연기 합성 시퀀스 데이터를 제작했습니다.
또한 1차 감지 모델, 2차 분류 모델을 분류하였습니다.

| 결과

기존보다 2배 많은 데이터 수량을 확보한 결과, 9km 거리의 미세 연기를 감지하고, 연기와 안개를 정확히 구분할 정도로 성능이 높아졌습니다. 또한 B기업의 팀장님은 야간 합성데이터 뿐만 아니라 추후 전체 데이터에 대해서도 품질 개선을 요청하고 싶다고 말씀하셨습니다.

(원본) 주간 이미지

(합성데이터) 야간 이미지

Knowhow

AI 데이터의 한계를 극복한 노하우

C기업

공장 안전 감지 AI 솔루션

Pebblous

| 문제 상황

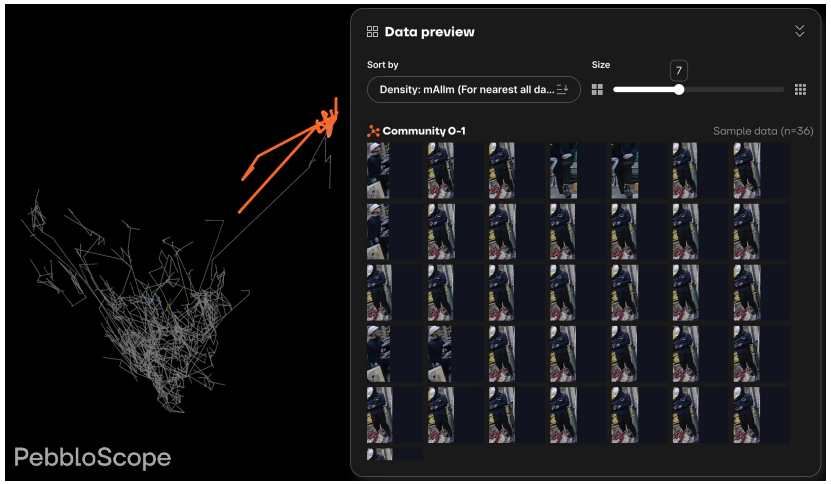
안전 감지 솔루션의 학습 효율이 떨어지는 원인은 크게 2가지였습니다. 첫 번째, CCTV 영상에서 연속된 프레임을 추출하다보니 중복 데이터가 과다했습니다. 두 번째, 그림자, 작업복, 케이블을 사람으로 오인식하고 있었습니다.

| 해결책

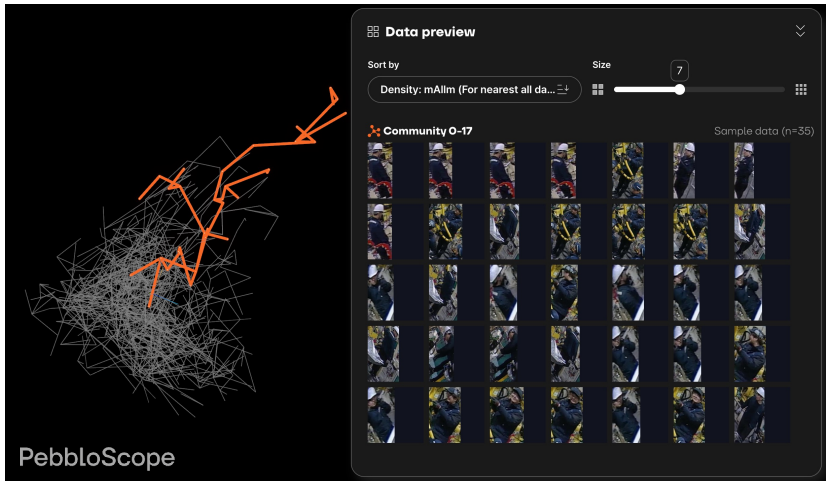
중복 및 유사 프레임을 제거하여 학습 효율을 증가시켰습니다. 옷의 색상, 조명, 복장 등 다양성을 반영한 합성데이터를 생성했습니다. 그 외에 로봇, 지게차 등 오인식 유발 객체에 별도로 라벨링을 진행했습니다.

| 결과

데이터 다이어트를 통해 중복 및 유사 프레임을 제거하여 오인식률이 감소하고, 다양한 환경에서도 안정적으로 작동하는 모델이 만들어졌습니다.



(원본) 중복 많음



(데이터 경량화) 중복 제거

PebbloScope 샘플 보기

- 데이터 3D 인터랙티브 시각화 도구

Pebblous

AI 데이터 규제 시대

여러분은 준비되어 있나요?

데이터 품질뿐 아니라 규제까지 꼼꼼히 살펴봐야 하는 이유

부작용 증가

AI가 급속히 확산되며 정보 편향, 허위 정보, 프라이버시 침해와 같은 부작용이 증가하고 있습니다.

최대 529억 벌금 약 € 31,000,000

이러한 상황을 법적으로 규제하고 있습니다. 특히 EU AI Act의 경우 위반 시 **최대 529억 원의 벌금**을 납부해야 합니다.

ISO/IEC
AI 데이터 품질 국제 표준

25012	5259	42119
데이터 품질 평가/개선	데이터 품질 평가/관리	AI 시스템 테스트/검증

기업 신뢰도 하락

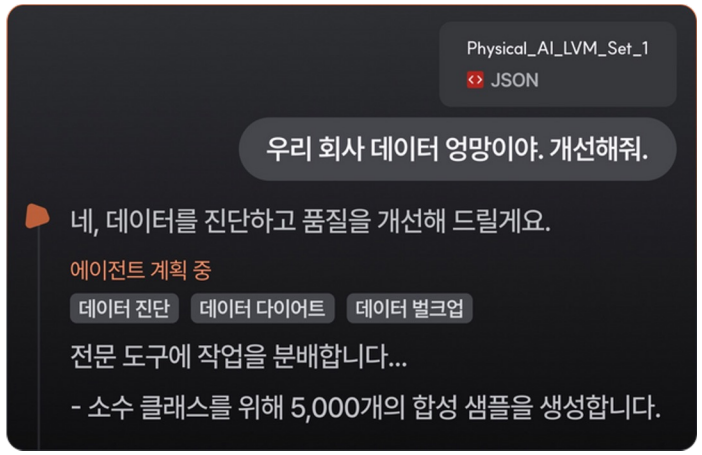
국내 AI 기본법의 경우, 해외보다 벌금의 액수는 낮은 편입니다.
그러나 벌금의 액수를 떠나서, 고객 입장에서 기업의 신뢰도가 하락할 위험이 높습니다.

* EU AI Act는 2026년 8월 본격 시행 예정
AI 규제 완화안, 법 적용 1년 유예

이 모든 것을 관리하는 솔루션, 데이터클리닉 2.0입니다.

AADS, Agentic AI Data Scientist

오류 감지를 넘어, 데이터 관리의 전 과정을 자율적으로
판단하고 실행하는 인공지능 기반 데이터 품질 관리 기술



▲ 클릭하여 데이터클리닉 2.0 작동 영상 확인하기

Pebblous

데이터클리닉 2.0 사용 후 이렇게 달라집니다.

데이터 수명주기를 고려한 자율적 품질 평가

데이터의 그 어떤 순간이든, 품질 상태를 항상 최상으로 유지합니다.

복잡한 규제, 완벽 대응 가능합니다.

데이터 품질 개선은 물론,
규제까지 완벽 대응 가능합니다.
기업 내부 규정, 공공기관의 경우
공공데이터 품질 관리 매뉴얼까지
완벽 학습하여 데이터를
개선합니다.

프롬프트 딱 한 줄로 충분합니다.

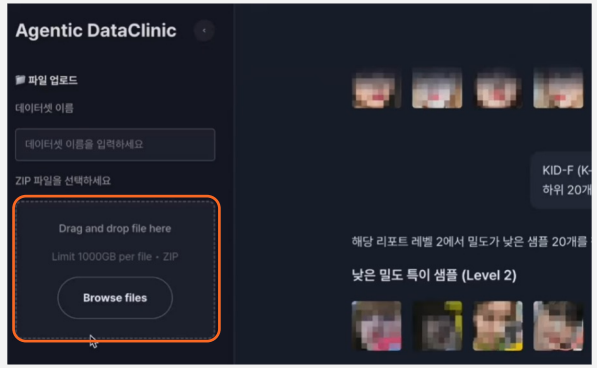
프롬프트 단 한 줄만으로 업무의
효율이 상승합니다. 데이터를
하나씩 들여다보지 않아도,
프롬프트 한 번이면 데이터 품질
진단, 개선, 보고서 추출까지
AADS가 대신해드립니다.

Pebblous

데이터클리닉 2.0

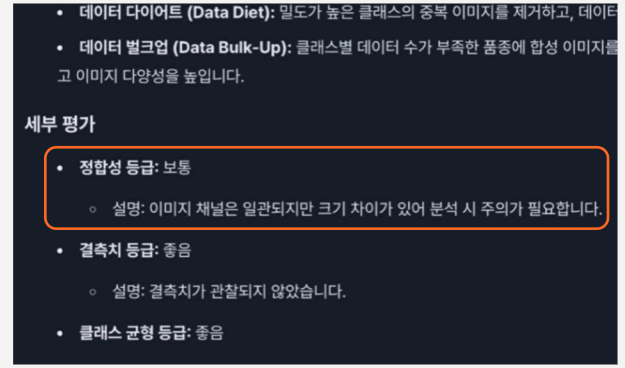
지금 무료로 사용해보세요!

01 진단 및 개선하고 싶은
데이터셋을 업로드해주세요.



02 '방금 업로드한 데이터셋 진단해줘'와
같이 프롬프트를 입력하여 데이터셋을
진단해보세요!

03 데이터 과학자에게 받는
컨설팅 그대로! 데이터클리닉 2.0에서
무료 진단을 받으세요.



데이터클리닉 2.0 무료 체험하기

업무량은 80% 줄이고,
AI 성능은 200% 올리는 비결

믿기지 않으시죠?
데이터클리닉 2.0에서 가능합니다.

데이터클리닉 2.0 도입 문의하기

Pebblous

올인원 데이터 관리 솔루션

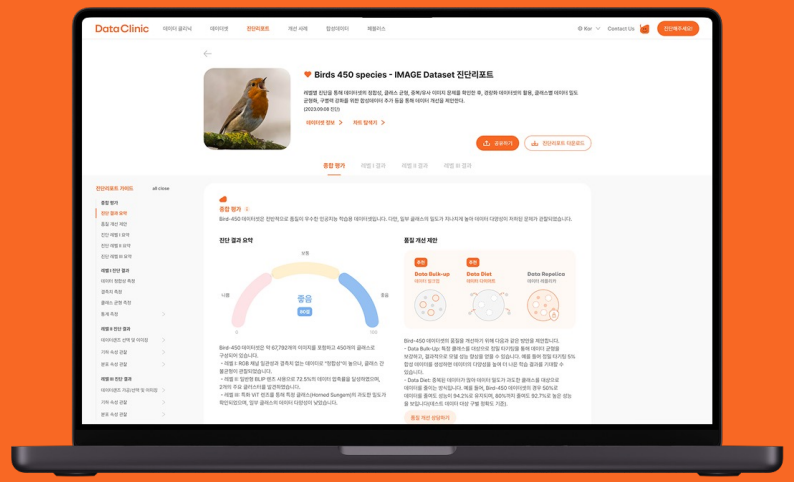
01
품질평가에서 개선까지

02
인터랙티브 가시화 커뮤니케이션

03
멀티모달 데이터셋

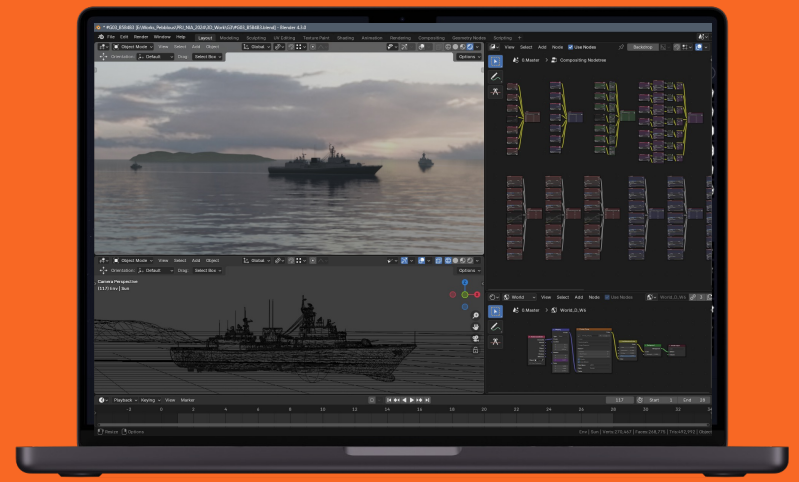
04
SaaS, 온프렘, API

DataClinic



PebbloScope

Synthetic Data



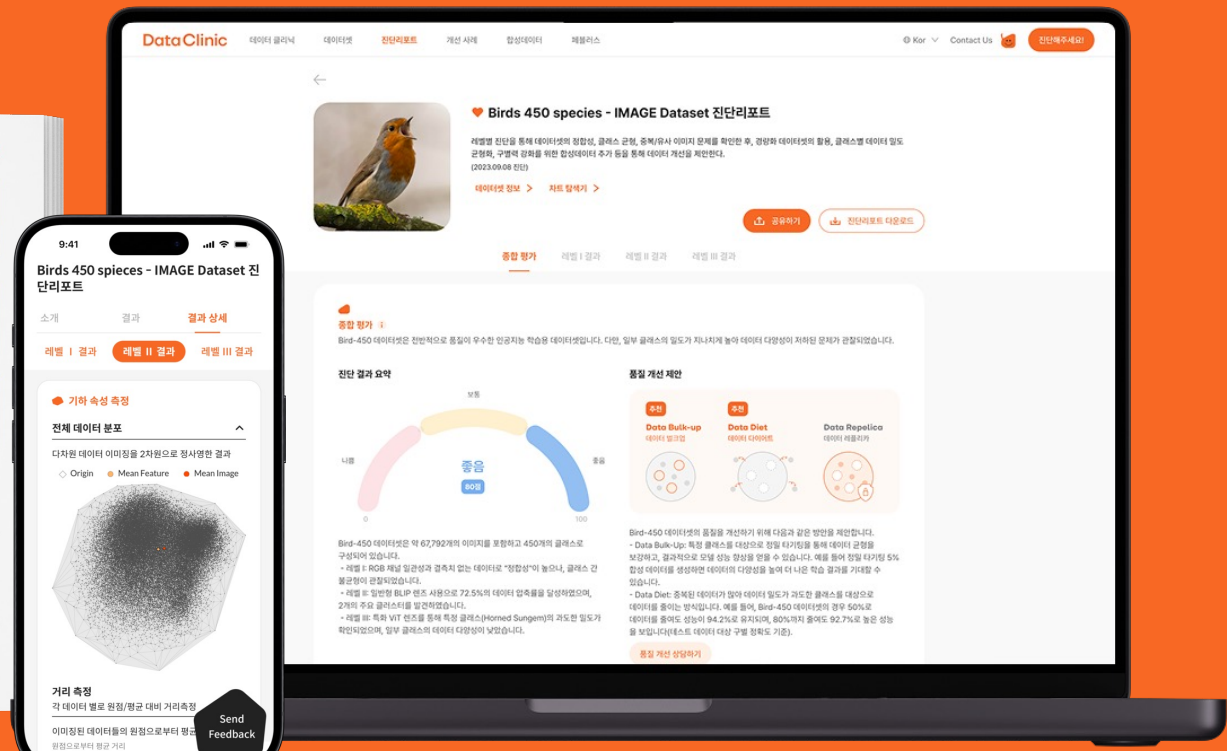
Pebblous

DataClinic

페블러스 데이터클리닉은 **데이터 종합병원**입니다.
AI 학습 데이터를 위한 품질 평가에서 합성데이터 생성까지의 모든 솔루션을 제공합니다.

웹 버전

PDF 버전

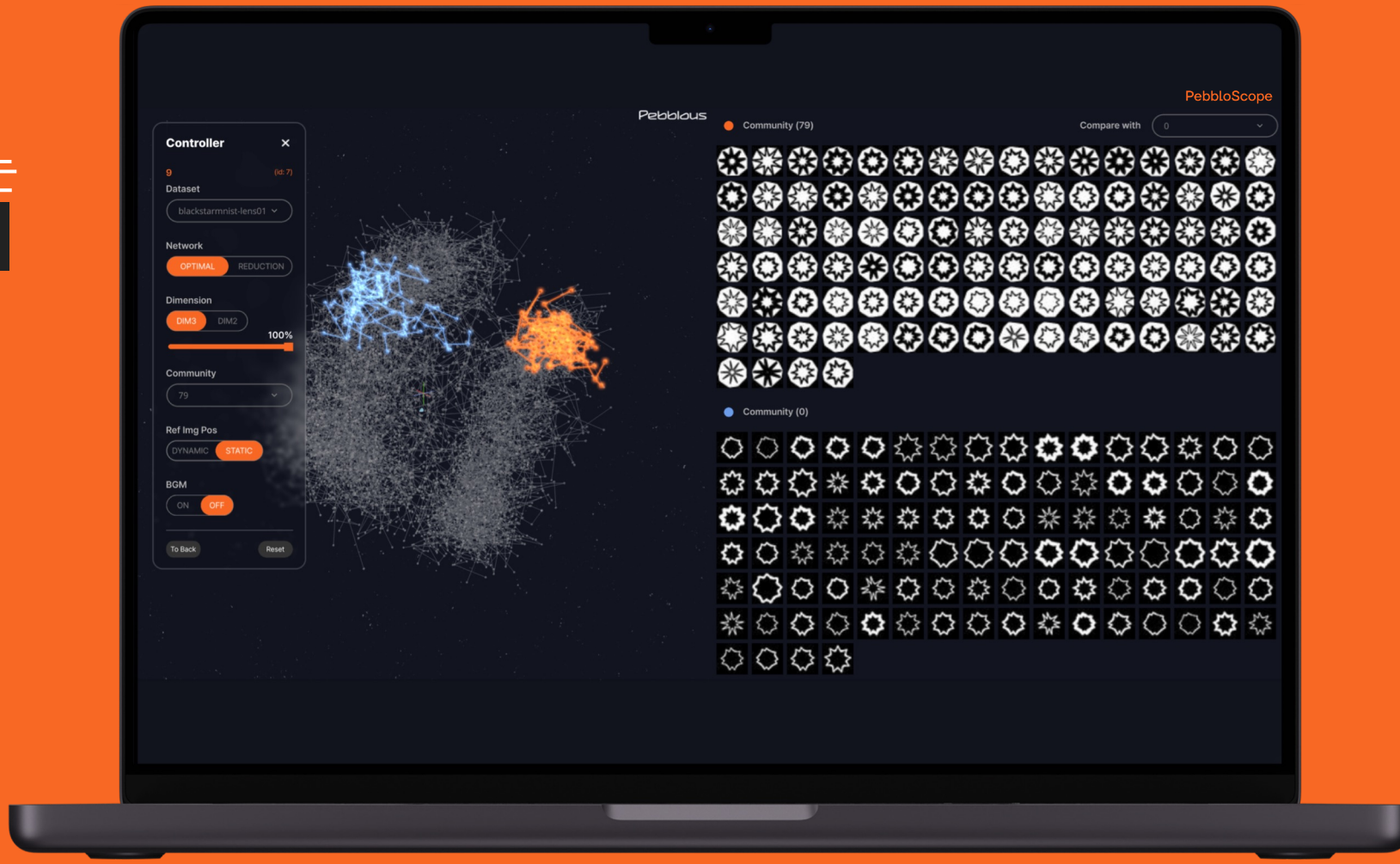


모바일 버전

PebbloScope

대용량 다차원 데이터에 대한
3D 인터랙티브 시각화를 지원하는
데이터 커뮤니케이션 도구입니다.

고차원의 데이터를 3차원 공간으로 변환하여 다양한 속성들을 인터랙티브하게 탐색하며 데이터 분석을 위한 인사이트를 얻을 수 있는 데이터 커뮤니케이션 도구입니다.



Product 03

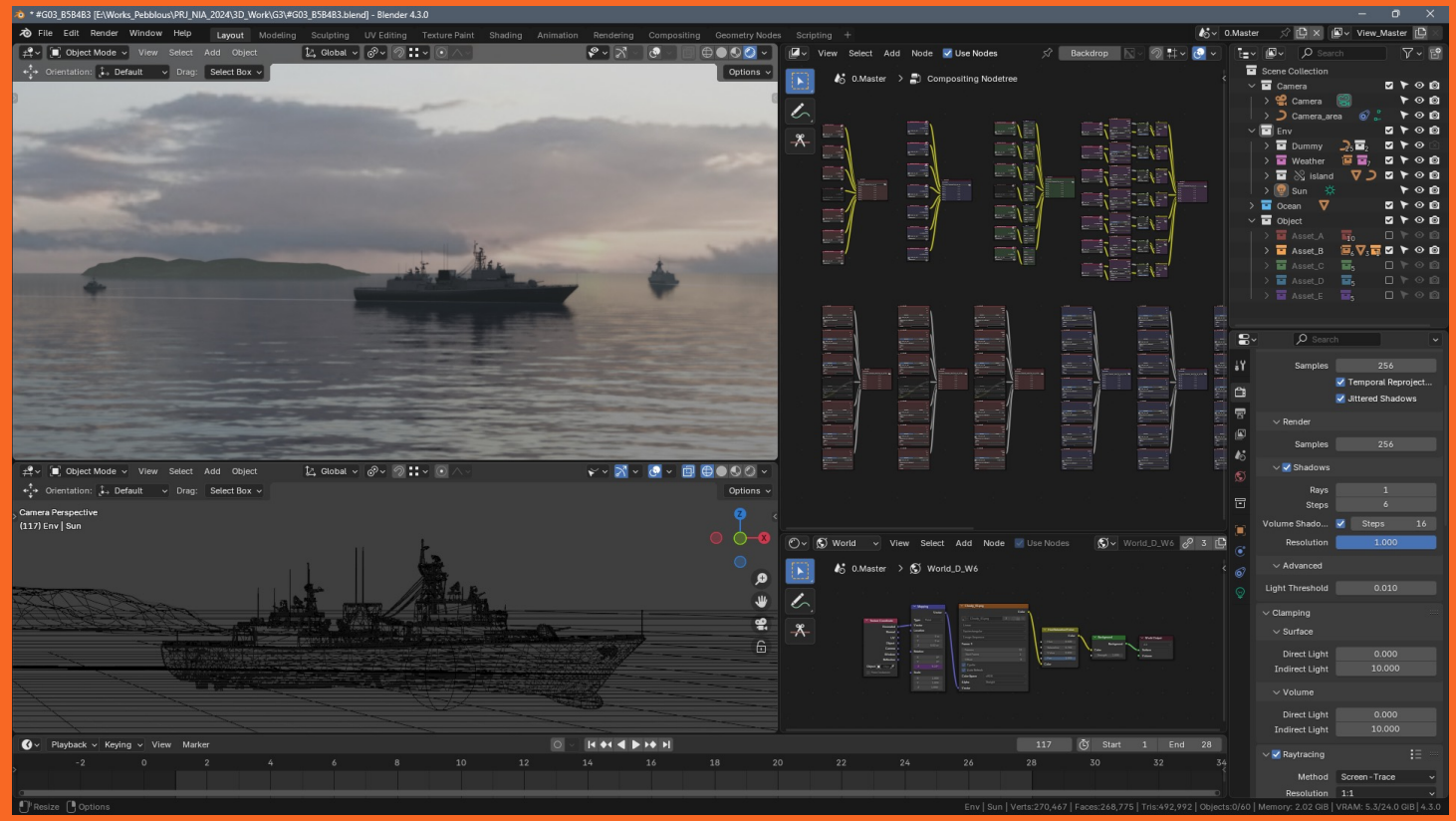
Pebblous

Synthetic Data

인공지능 학습을 위해

① 데이터 수량이 부족한 경우
② 실제 데이터를 구할 수 없는 경우
③ 다양한 환경에서의 데이터가 필요한 경우

합성데이터를 제작합니다.



Fabulous Data With

Pebblous

Better Data Makes Better AI

Pebblous.ai