

Pebblous Makes Data Tangible

Experience
The Innovation In
Data Quality

Pebblous

E-mail | contact@pebbulous.ai

Tel | 044-589-3824

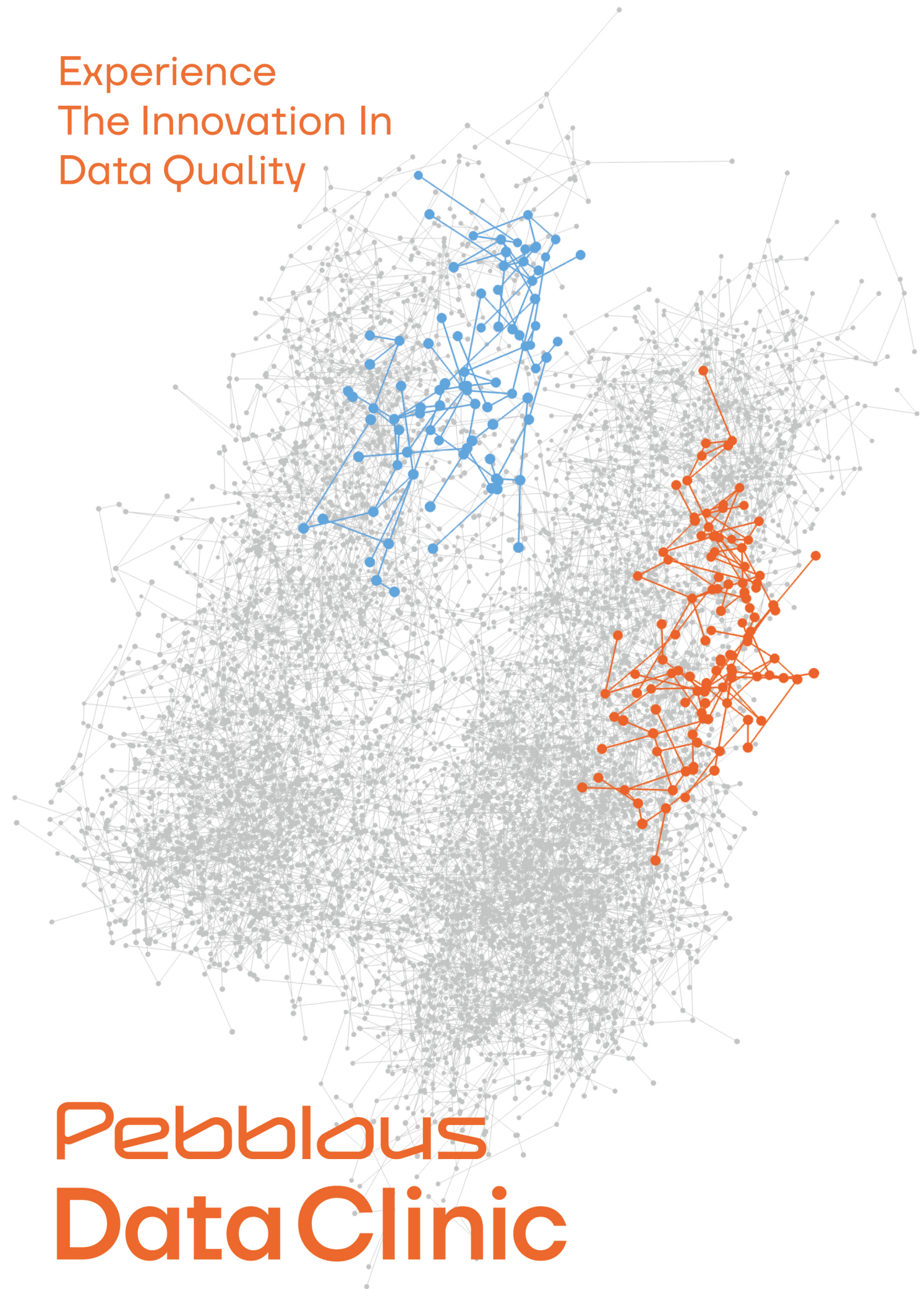
Web | pebbulous.ai

Copyright © Pebblous Inc. All rights reserved. 발행일 | 2024.12.26

(주) 페블러스 홈페이지



Pebblous Data Clinic



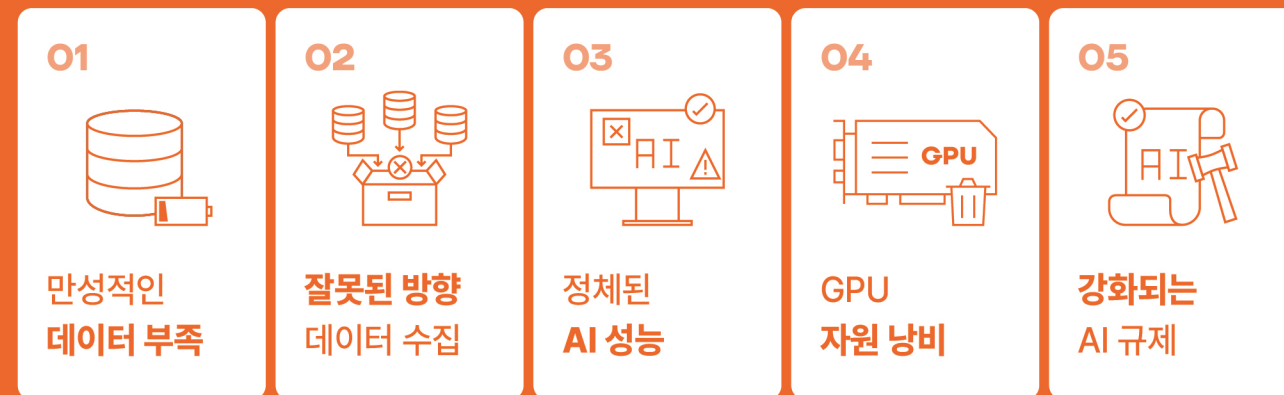
Pebblous

Power your innovation without wasting a watt.

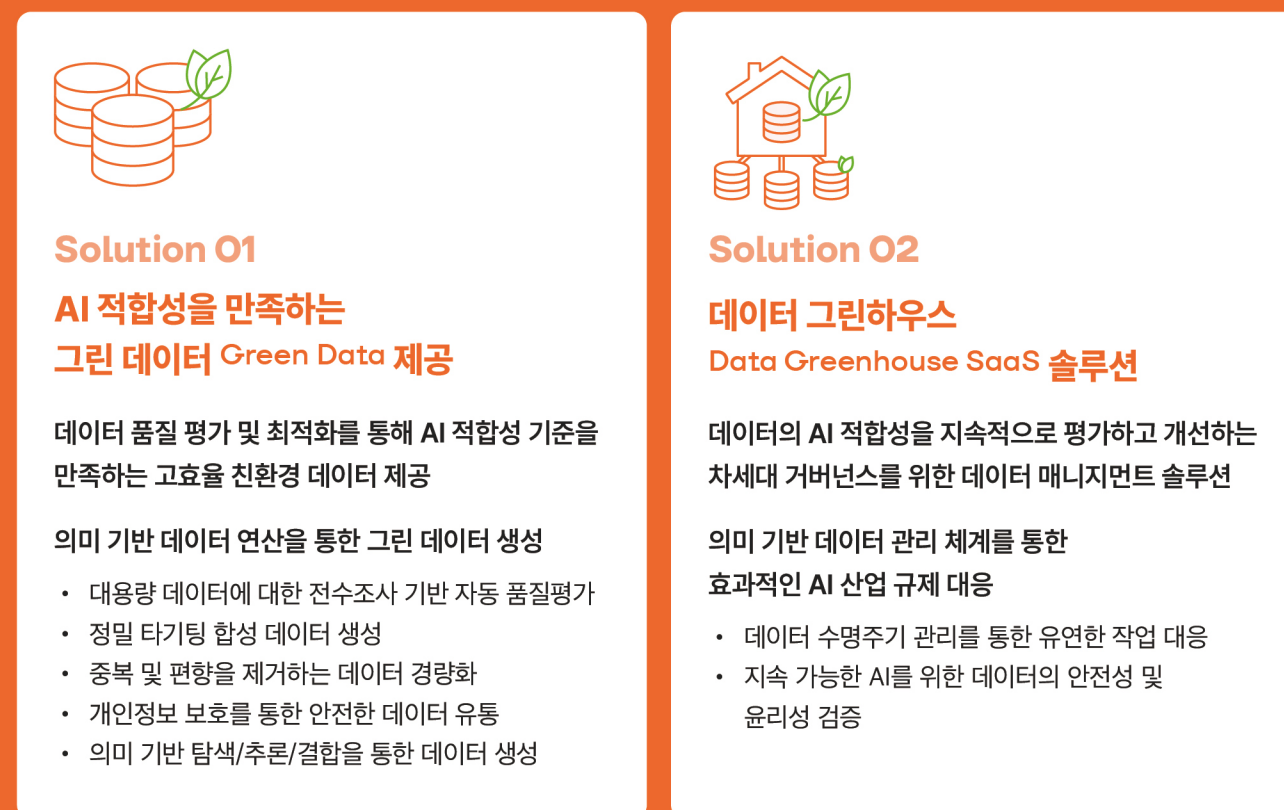
페블러스는 그린 데이터 Green Data *와 데이터 그린하우스 Data Greenhouse를 통해 지속가능한 인공지능에 기여하고자 합니다.

*그린 데이터: 데이터 품질 평가 및 최적화를 통해 AI 적합성 기준을 만족하는 고효율 친환경 데이터

현장의 문제



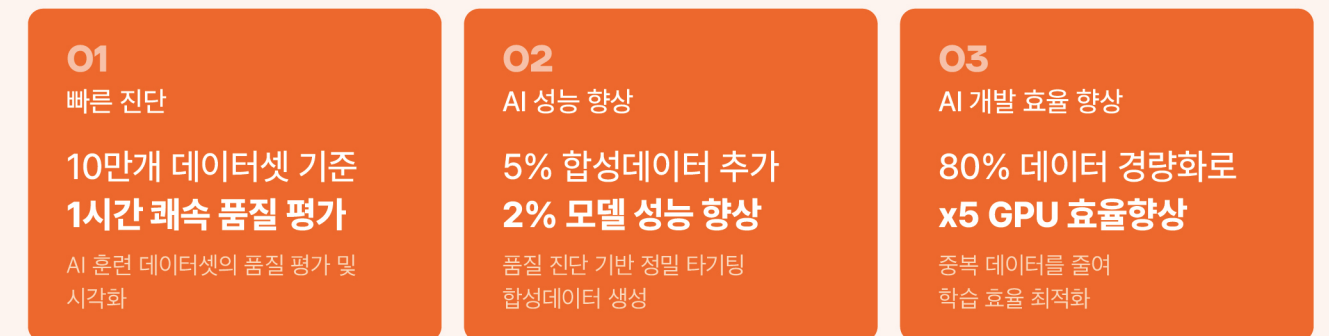
페블러스의 해결 방안



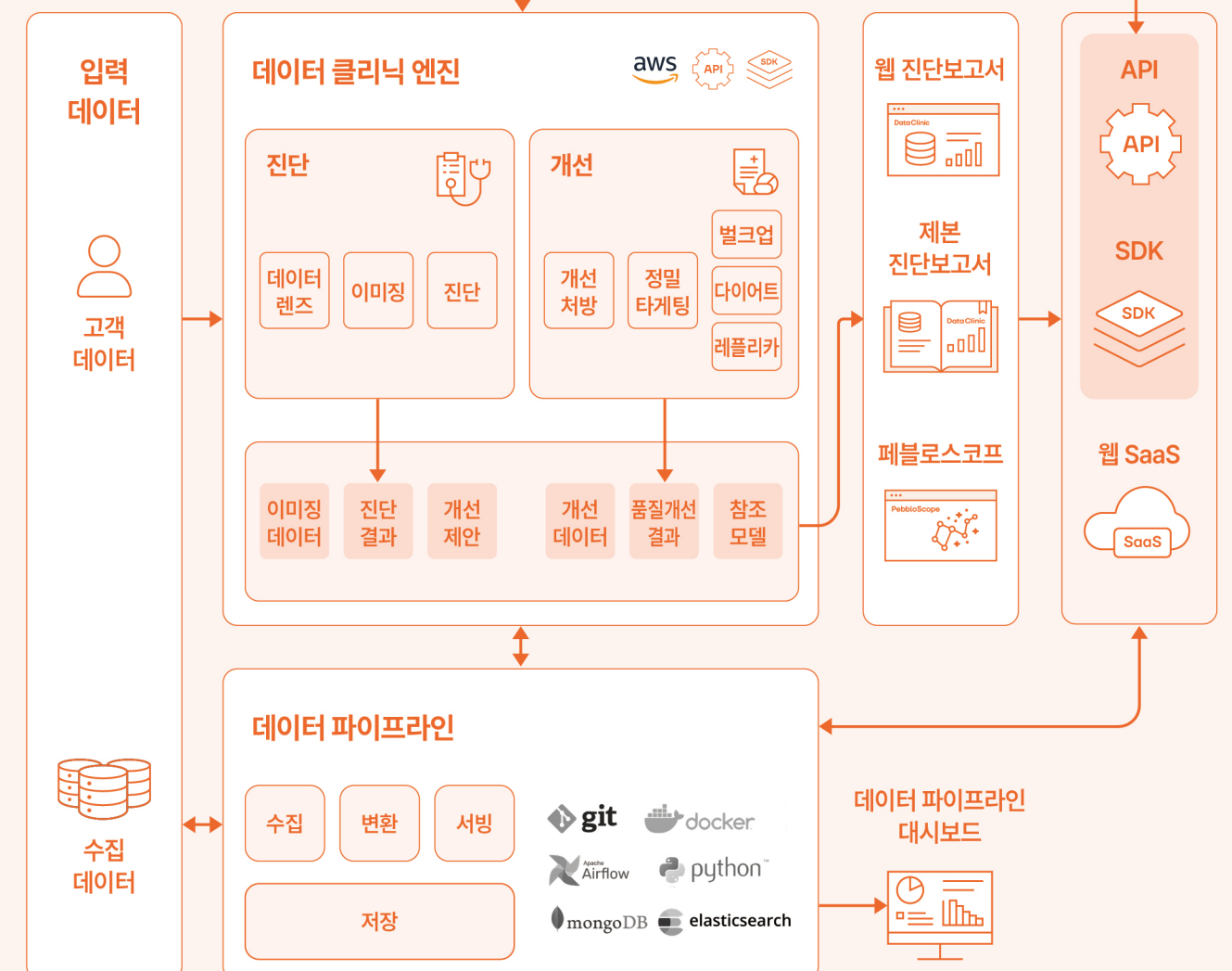
Pebblous Data Clinic

for AI-Ready Data Management SaaS

페블러스 Data Clinic은 빅데이터 및 AI 학습데이터의 품질 평가와 개선을 위한 올인원(All-in-one) 솔루션입니다.



데이터 클리닉 SaaS 아키텍처



Pebblous Data Clinic

AI Data Doctor

데이터를 위한 종합병원 Data Clinic은
데이터의 품질 진단에서 품질 개선까지의 전과정의 서비스를 제공합니다.

이미징 데이터의 잠재력 발굴

- ‘데이터 이미징(Data Imaging)’ 기술을 통해 데이터셋을 심도 있게 탐색 가능합니다.
- ‘데이터 렌즈(DataLens)’는 복잡한 데이터셋 공간 분포를 명확하고 관찰 가능한 형태로 변환하여 데이터셋에 대한 이해를 돕습니다.

진단 정밀한 데이터 분석

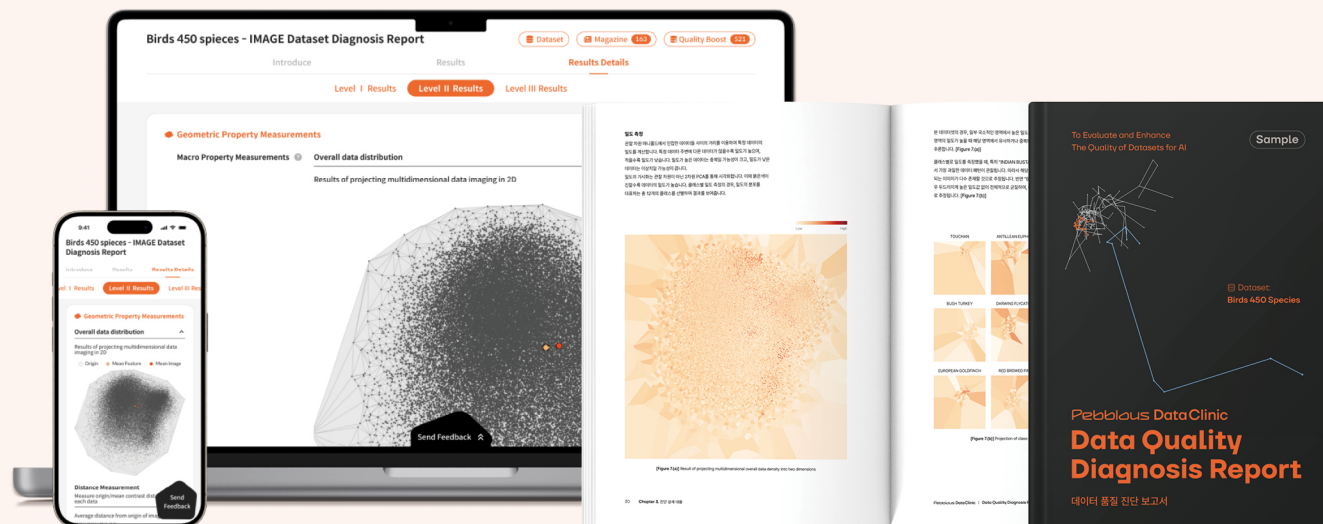
- 데이터 품질 진단을 통해 AI 학습데이터셋의 상태를 면밀히 검토합니다.
- ‘데이터 렌즈(DataLens)’를 구성하는 최첨단 인공지능망을 활용하여, 데이터셋의 특징과 숨겨진 패턴을 밝히고 필요한 개선방안을 제시합니다.

보고서 데이터 품질에 대한 명료한 설명과 근거

- 단순 검증 통계 값과 차트를 넘어서서 AI 학습데이터셋의 품질에 대하여 과학적으로 설명 가능한 진단 보고서를 제공합니다.
- 데이터에 숨겨진 패턴과 구조를 설명하고 다양한 지표(Metric)를 이용한 품질 평가 결과에 기반하여 데이터의 문제상황과 개선방안을 제시하는 분석 문서입니다.

개선 최적화된 품질 개선 서비스

- 데이터 진단 결과에 기반하여 All-in-One으로 데이터셋에 최적화된 개선 서비스를 제공합니다. (*개선 옵션이 포함된 서비스 이용 시)
- 합성데이터(Data Bulk-Up)와 효율적인 데이터 경량화(Data Diet)가 결합된 개선방안으로 AI 성능을 위해 학습데이터셋 품질을 최적화합니다.

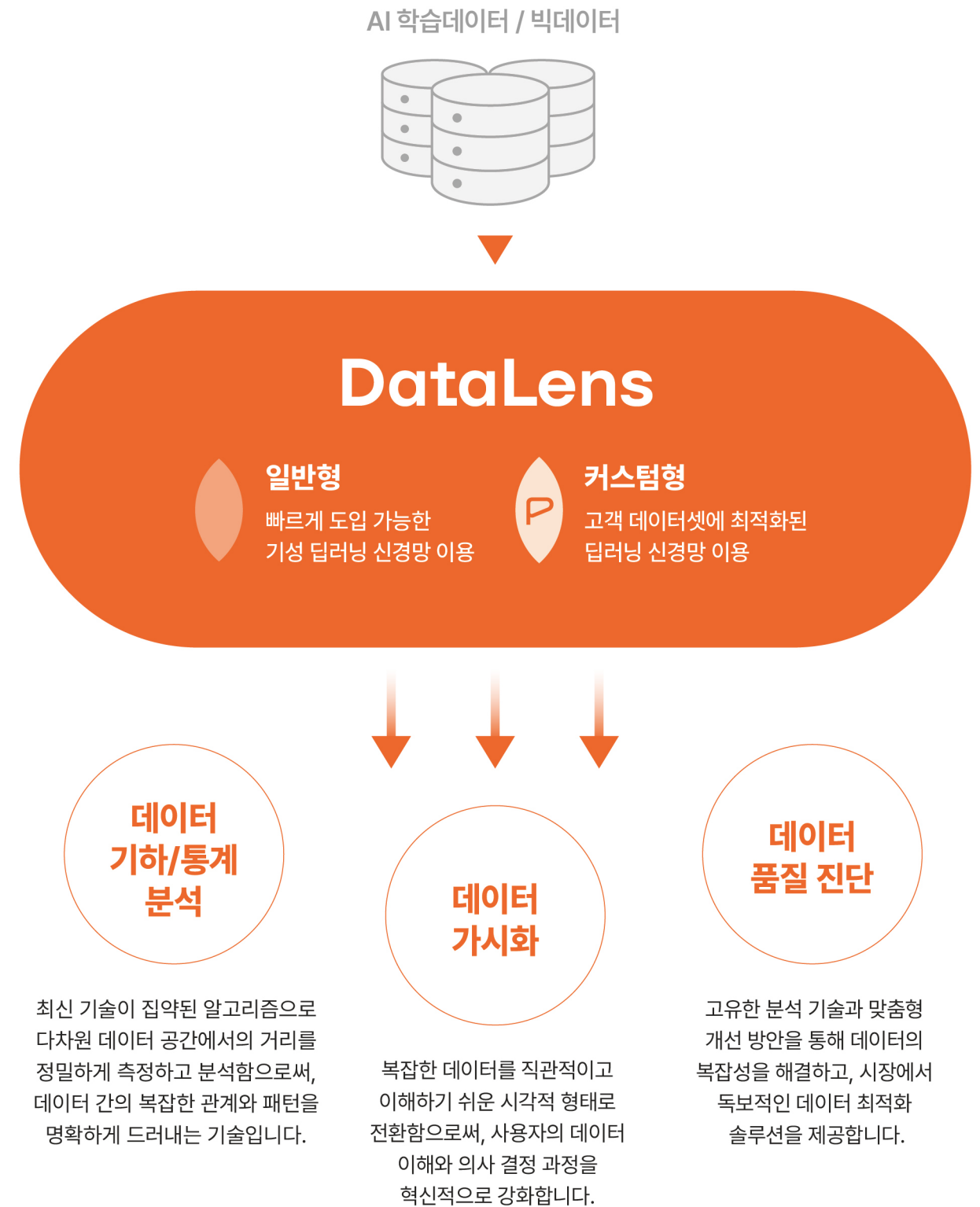


Pebblous Data Clinic

차별점

데이터 클리닉은 최신 딥러닝 신경망으로 구성된 ‘DataLens’를 통해
인공지능 학습데이터의 다차원적 특성을 정밀하게 관찰할 수 있습니다.

*"DATALENS" 미국상표출원



Pebblous Data Clinic

주요 특징

다재다능한 데이터 클리닉은 인공지능과 빅데이터 분석을 위한 올인원(All-in-One) 데이터 솔루션입니다.



데이터 유형에 상관없이 작동합니다.

텍스트, 이미지, 시계열, DB 데이터까지 'Data Clinic'은 모든 유형의 데이터에 대해서 품질평가와 데이터 품질 개선이 가능합니다.



작업 유형이나 산업 도메인에 상관없이 작동합니다.

'Data Clinic'은 다차원 기하 공간에서 데이터의 분포적 특성을 통해 데이터의 품질을 해석하므로 특정 산업이나 작업 유형에 상관없이 데이터 분석과 인사이트를 제공합니다.



정밀 타게팅으로 고품질 합성데이터를 생성합니다.

데이터셋 품질 진단 결과와 데이터셋을 활용할 AI 모델, 분석 분야, 도메인 지식을 활용하여 최적의 위치에 최적 볼륨의 합성데이터를 생성합니다.



정형 도메인에 특화된 진단/개선 서비스를 제공합니다.

결합데이터의 결합 품질 평가, 가명성과 익명성이 보장된 완전 재현데이터 생성 등 정형 데이터 분야에서 기존 기술과는 차별화되는 다양한 품질 진단 및 합성 기술들을 제공합니다.



AI 학습데이터를 효율적으로 운영하고 관리합니다.

'데이터 다이어트'를 통해 불필요한 데이터를 제거하여 AI 학습 데이터셋의 볼륨을 최적화할 수 있고, 이를 통해 인공지능의 성능은 물론 GPU와 스토리지 자원의 효율을 극대화할 수 있습니다.



설명 가능한 인공지능에 기여합니다.

인공지능에게는 단순한 성능 외에도 윤리, 공정, 안전에 대한 기준들이 높아지고 있습니다. 설명 가능한 인공지능은 분포적 특성을 고려한 데이터 품질 평가에서 출발합니다.

Pebblous Data Clinic

데이터 품질 진단에서 개선까지

고객은 데이터 클리닉의 정확한 데이터 품질 진단에 근거한 데이터 치료를 통해 안전성, 효율성, 신뢰성을 확보한 AI-Ready 데이터를 확보할 수 있습니다.

O1. 데이터 품질 진단

페블러스는 '데이터렌즈(DataLens)' 인공지능망을 통해 고객의 AI 학습데이터를 임베딩 공간에 특징 벡터(Features)로 표현합니다. 이렇게 관찰 가능하고 측정 가능하도록 변환된 '데이터 이미징' 결과에 대해서 페블러스만의 다양한 과학적 지표로 데이터의 품질을 측정하고 이를 가시화 등을 통해 설명하는 진단 서비스를 제공하고 있습니다.



작업 중립적 Task-neutral 품질 진단

- 데이터셋의 도메인 혹은 이용 목적과 관계없이 데이터셋 자체에 대한 통계적, 물리적 특성을 진단하는 과정
- 정합성, 무결성, 완전성 등의 기초 품질 진단 지표들과 더불어 임베딩 공간의 데이터 분포에 대한 정량적 및 정성적 고급 진단 수행

작업 특이적 Task-specific 품질 진단

- 적용 도메인과 분류, 예측, 생성 등의 고객이 구성하고자 하는 AI 모델 및 작업에 적합하게 설계되는 맞춤형 진단 과정
- 임베딩 공간의 데이터 분포 및 타겟 작업의 특성을 동시에 고려하여 클래스 구별력, 데이터의 밀도, 이상치 탐지 등 진단 목표를 설정하여 고급 진단 수행

O2. 데이터 품질 개선

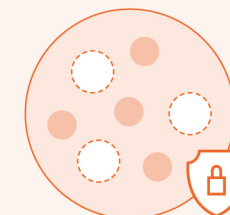
AI 모델 성능 개선을 위해 데이터셋의 품질과 기능을 향상시키는 기술입니다. 데이터 진단한 결과에 기반하여, 데이터의 사용목적에 최적화된 정밀 타게팅 합성데이터를 처방하거나 불필요한 데이터를 제거하는 서비스입니다.



□ 비어있는 데이터 영역 ● 중복 데이터 영역 ○ 실제 데이터 ○ 합성데이터

데이터 레플리카 Data Replica

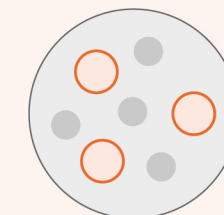
원본의 내용을 보호하면서도 통계적·분포적 특성이 유사한 가상데이터를 생성



→ 개인정보 유출 리스크 예방

데이터 벌크업 Data Bulk-Up

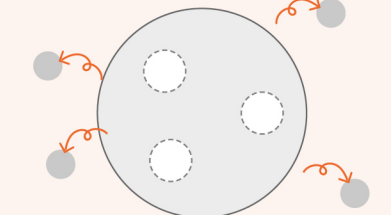
데이터가 부족한 부분에 데이터셋에 적합한 정밀 타게팅 합성데이터를 생성하고 제공



→ 데이터 수집 비용 절감, AI 성능 향상

데이터 다이어트 Data Diet

AI 모델의 성능을 보장하는 선에서 데이터셋의 중복/유사 데이터를 제거하고 경량화



→ AI 학습 비용 절감, 개발 효율 제고

레벨 I, II, III 진단 체계에 따른 DataClinic 프레임워크

데이터 클리닉의 진단 체계는 데이터의 품질과 구조를 진단 레벨 별로 평가하고 개선하는 프레임워크를 통하여, 고객의 데이터를 효과적으로 진단하고 관리할 수 있도록 설계되었습니다.

01. 진단



02. 개선



데이터 렌즈와 이미징

데이터 이미징 Data Imaging

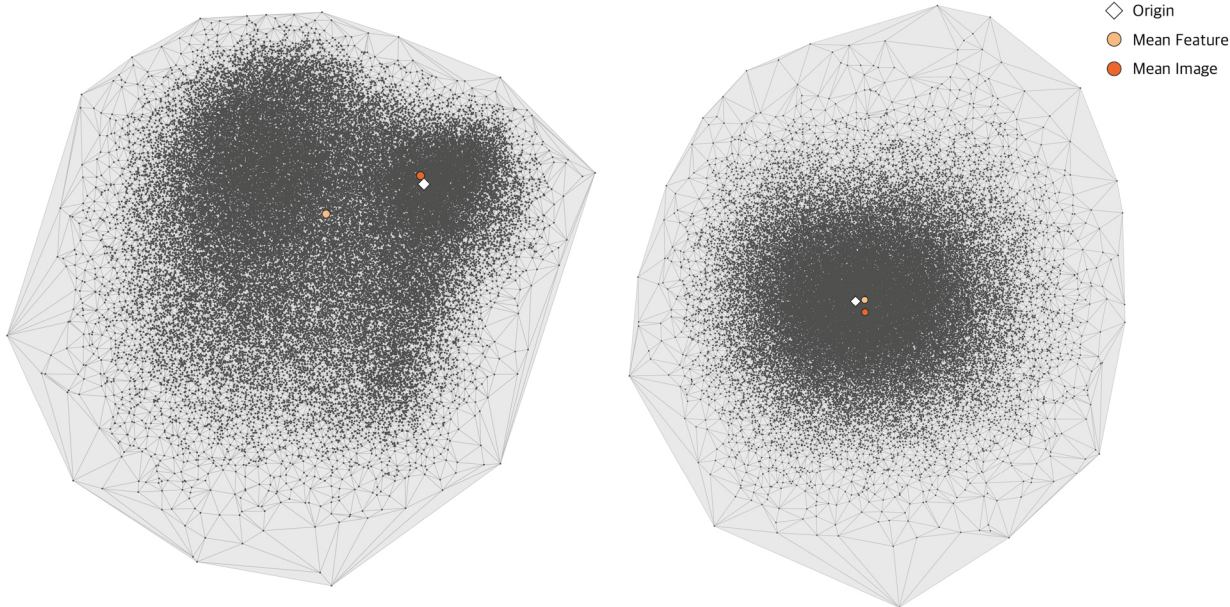
‘Data Imaging’은 복잡한 다차원의 데이터셋을 시각화하고 해석하기 위한 기술로서, 특허로 등록되어 있는 페블러스의 핵심 기술입니다. 대규모 데이터셋을 목적에 맞게 구성된 임베딩 공간으로 변환하여, 신경망이 데이터의 구조적 특성을 이해하는 단계입니다. 데이터의 패턴과 관계에 대한 심층적인 인사이트를 확보하여 정보에 입각한 의사 결정과 예측 분석을 돕습니다.

데이터의 내재적 차원 Intrinsic Dimension of Data

데이터셋 전체를 ‘Data Lens’에 통과시켜 특징(Features) 벡터들을 추출하는 과정에서 입력 데이터셋에 최적화된 특징 벡터들의 최소 차원이 산출되는데, 이 최소 차원을 관찰 차원 혹은 내재적 차원(Intrinsic Dimension)이라 정의합니다. 관찰 차원은 일반형 렌즈에서 출력된 특징 벡터들의 최소 차원이고, 내재적 차원은 커스텀 렌즈에서 출력된 특징 벡터들의 최소 차원입니다.

레벨 II vs. 레벨 III 진단

	레벨II 진단	레벨III 진단
데이터 렌즈	빠르게 도입이 가능한 기성 딥러닝 신경망을 활용하여 ‘데이터 렌즈’를 구성합니다.	작업과 도메인에 최적화된 커스텀 인공신경망 을 이용하여 ‘데이터 렌즈’를 구성합니다.
진단 범위	데이터의 기하 속성 및 분포 속성 측정을 통해 데이터의 품질과 구조에 대한 평가에 초점을 맞추어 개선이 필요한 영역 정보를 식별합니다.	레벨2 진단에 더해서, 심층 진단이 가능한 고급 지표를 통해 데이터의 복잡한 문제에 대한 인사이트와 개선 솔루션을 제공 합니다.
문제 식별력	범용 임베딩 공간에서의 분포 분석을 통해 학습데이터셋의 편향과 같은 문제를 식별할 수 있습니다.	데이터 특이적 임베딩 공간 에서 학습데이터셋의 문제를 더욱 정밀하게 발견하고 정밀 타게팅 합성데이터와 같은 개선 방안을 찾을 수 있습니다.



진단 결과를 효과적으로 설명할 수 있는 측정 항목과 다양한 시각화

페블러스 품질 진단 보고서의 진단 항목과 다양한 시각화들은
데이터의 품질 진단 결과를 직관적으로 전달합니다.

데이터 간 거리 측정

데이터 렌즈를 통하면 두 데이터 사이의 거리 측정이 가능해지고, 이를 통해 두 데이터의 유사성과 차이를 체계적으로
설명할 수 있습니다.
두 데이터 사이의 거리 측정은 다중 데이터들로 확장될 수 있고 이를 통해 전반적인 데이터의 구조를 계산할 수 있습니다.
예를 들어, 이상치나 중복 데이터의 존재 여부를 확인하는 데 효과적입니다.
이러한 데이터의 기하학적 속성들은 다양한 차트와 예제를 통해 설명됩니다.

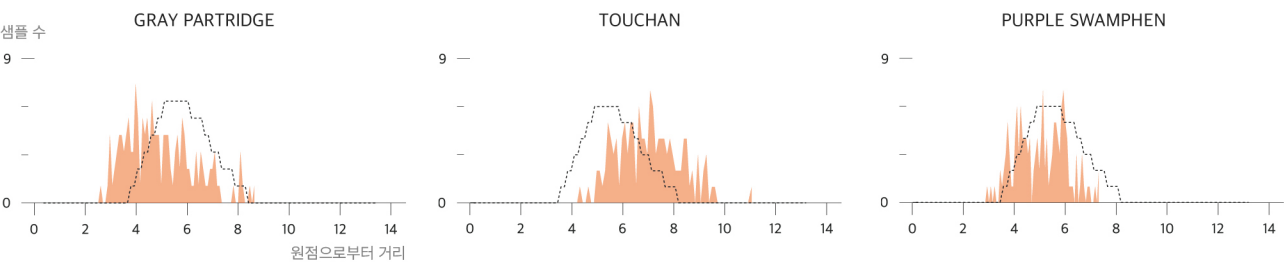
특정 데이터로부터 거리가 가장 가까운 10개 데이터



특정 데이터로부터 거리가 가장 먼 10개 데이터



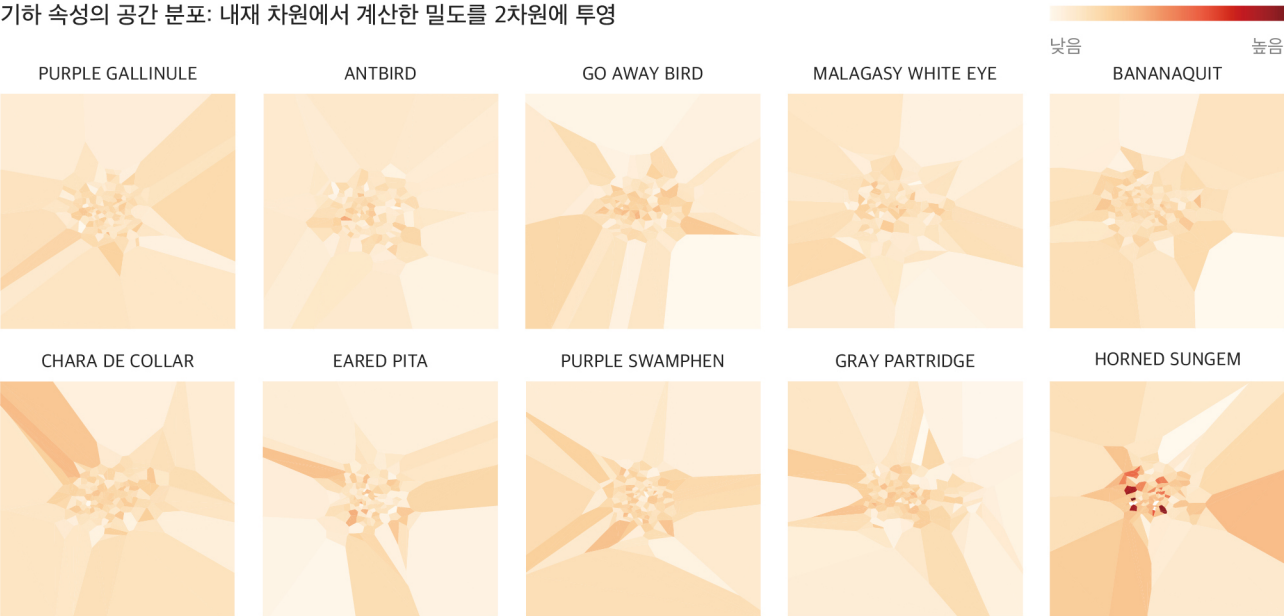
기하 속성의 통계적 분포: 클래스별 원점 거리



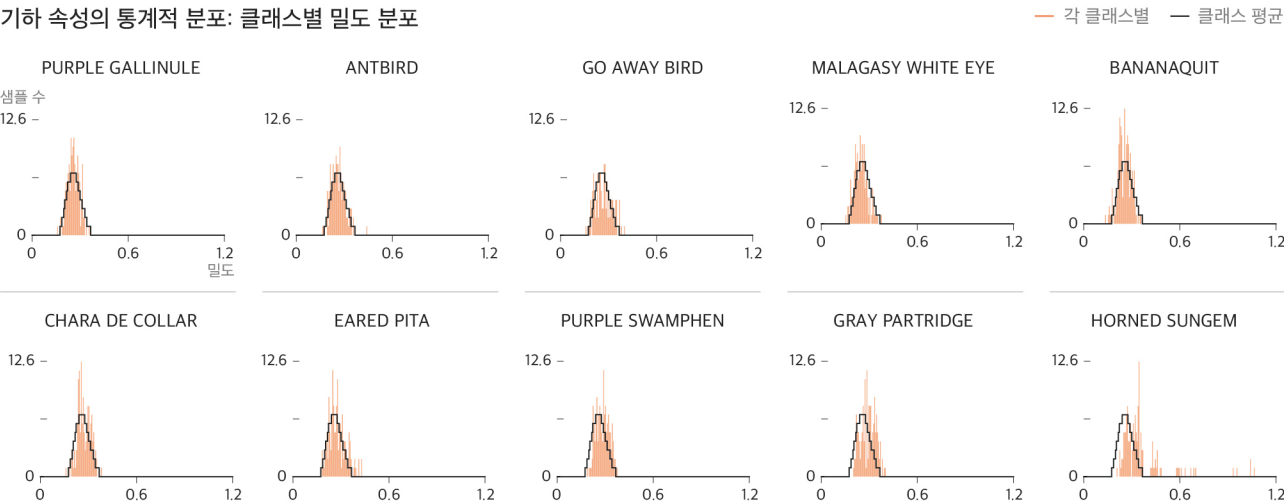
데이터 내부 밀도 측정

데이터 밀도 측정은 임베딩 공간에서 데이터 포인트들의 밀집도를 정량화하여 데이터의 구조와 복잡한 패턴을 해석하며,
유의미한 관계를 추론하기 위한 핵심적인 기술입니다.
데이터 품질 진단보고서에는 밀도 측정 결과 차트들은 과밀 영역 혹은 저밀도 영역을 식별하여 중요 영역의 데이터들과
잠재적 이상치를 구분하는 데 사용됩니다. 데이터 유사도, 클러스터링, 그리고 특이 패턴을 파악하는 데 효과적입니다.

기하 속성의 공간 분포: 내재 차원에서 계산한 밀도를 2차원에 투영



기하 속성의 통계적 분포: 클래스별 밀도 분포



AI 학습데이터의 문제를 직관적으로 전달하는 다양한 데이터 품질 분석 보고서

페블러스 데이터 품질 분석 보고서는 데이터의 품질에 대한 과학적 분석을 제공하며, 객관적인 설명과 차트를 통해 데이터에 대한 이해를 돕습니다.

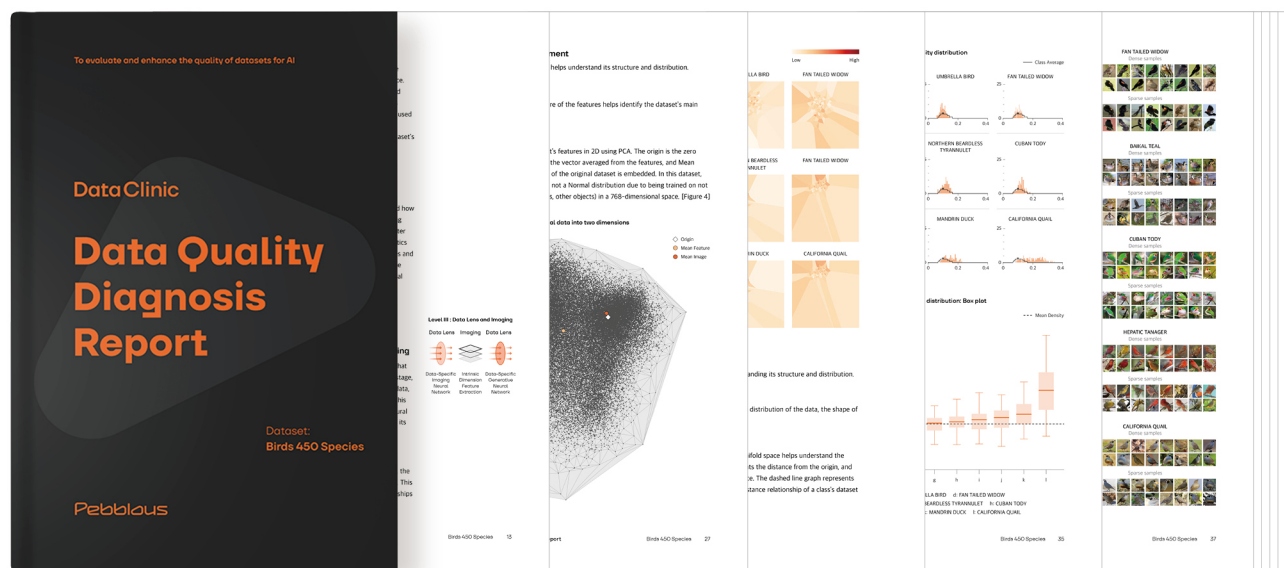
진단보고서 샘플 다운로드
<https://bit.ly/진단보고서>



O1. 데이터 품질 진단 보고서

AI 학습데이터셋에 대한 종합적인 품질 진단 결과와 개선방안이 담긴 보고서입니다.

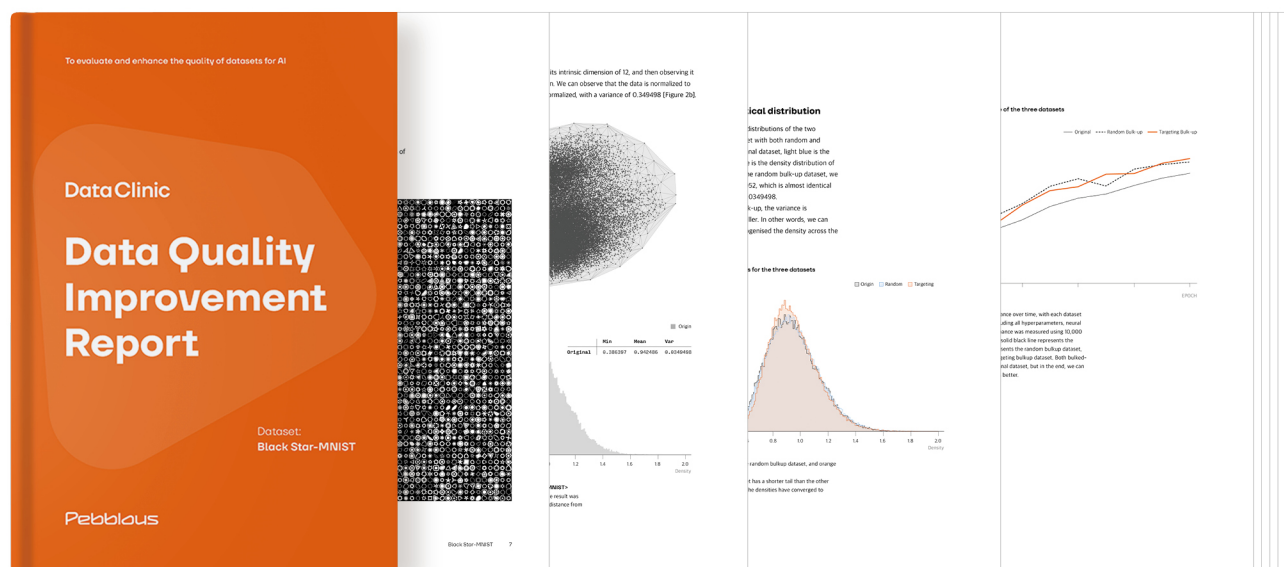
*영문 보고서 샘플: 미국 라스베이거스 CES 2024 공개



O2. 데이터 품질 개선 보고서

데이터 품질 진단 결과에 기반한 품질 개선 과정과 효과를 설명하는 보고서입니다.

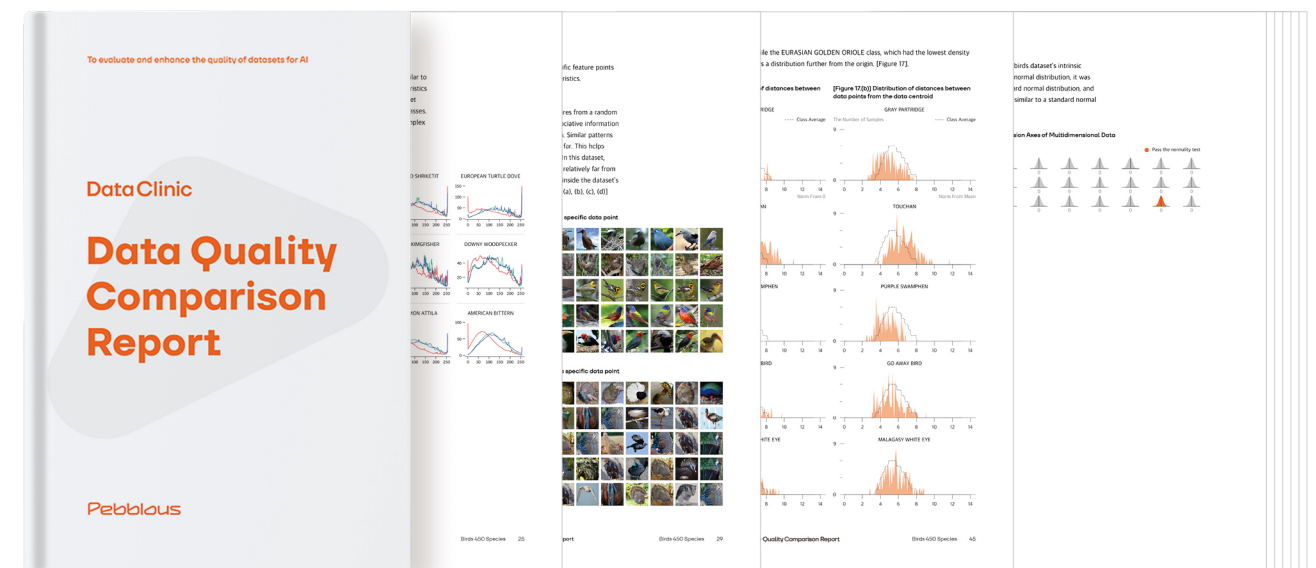
*영문 보고서 샘플: 미국 라스베이거스 CES 2024 공개



O3. 데이터 품질 비교 보고서 | 출시 예정

품질 개선 전후, 또는 학습데이터와 테스트 데이터의 품질을 상세히 비교평가하는 보고서입니다.

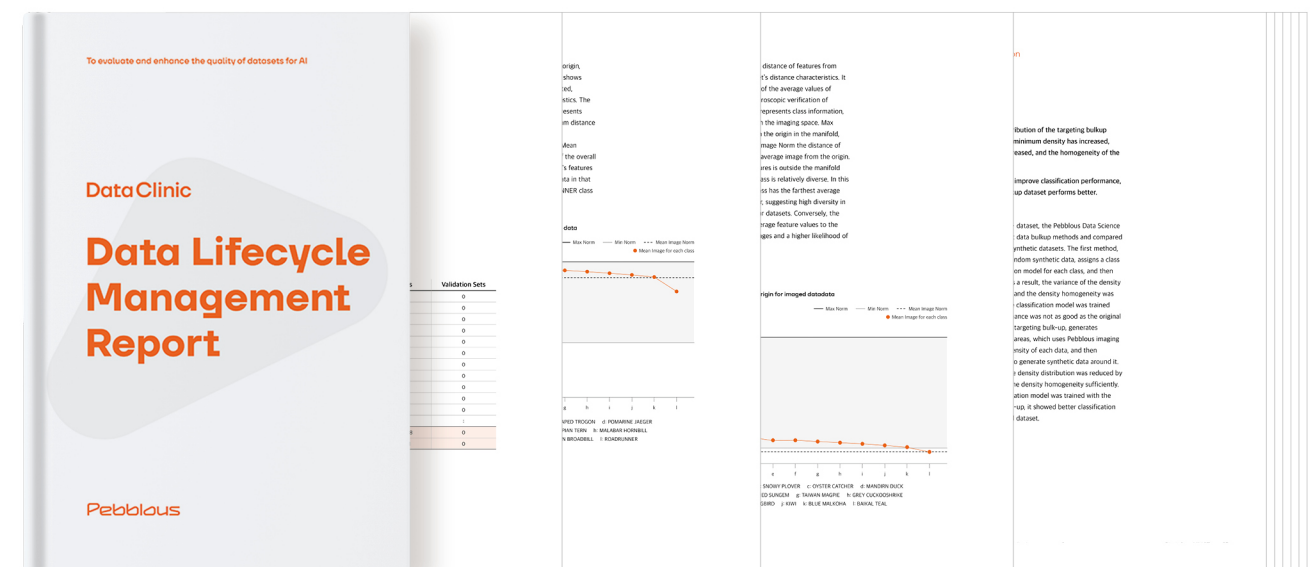
*예시



O4. 데이터 수명주기 관리 보고서 | 출시 예정

데이터 수명주기 관리에 관한 분석과 제안 내용이 담긴 보고서입니다.

*예시



품질 진단에 기반한 정밀 타게팅 합성데이터 생성

고객의 요청 및 잠재 수요를 발굴하여 생성한 다양한 합성데이터를 제공합니다.

주요 생성 기법



컴퓨터 그래픽스 기술

모델링, 애니메이션, 렌더링 등
고급 컴퓨터 그래픽스 기술을 사용하여
사실적인 비주얼 데이터를 생성합니다.



생성 AI 기술

딥러닝 모델을 사용하여 실제 데이터
패턴을 모방하는 새로운 데이터를
자동으로 생성합니다.



모델링 & 시뮬레이션

시나리오, 수학적, 물리 방정식으로
현실 세계의 환경과 프로세스를 모델링
하고 이를 시뮬레이션한 데이터를
생성합니다.

AI 합성데이터 성능 개선 사례

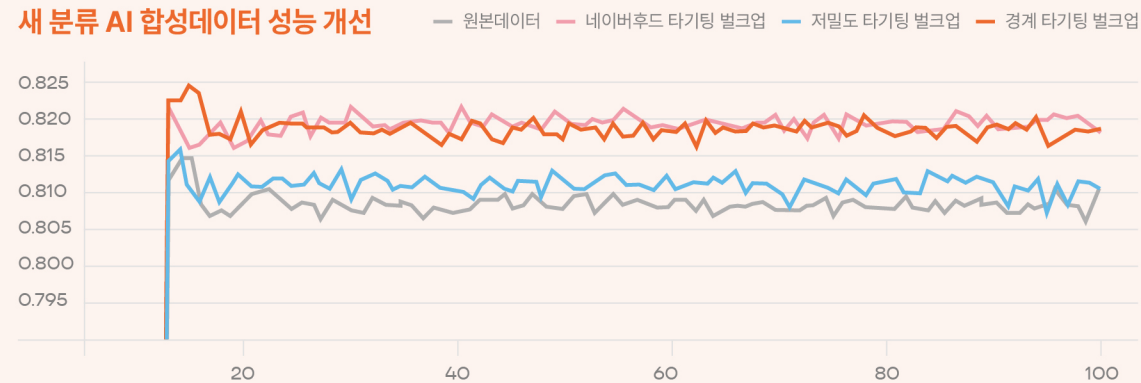
Birds-450 데이터셋 원본



정밀타기팅 합성데이터



새 분류 AI 합성데이터 성능 개선



단일 이미지 3D 생성



단일 사진 기반 3D 모델 생성 합성데이터
*POSTECH 공동연구, SOTA 수준 달성

드론 탐지



드론의 인식 및 추적을 위한
3D 합성데이터

특수환경의 인물 및 자세



차량 내부에서의 어린이 인식 등
특수환경에서의 인물 및 자세 인식을 위한
AI 학습용 합성데이터

자율 주행



실내 자율주행 로봇의 네비게이션을 위한
3D 합성데이터. 파노라마 카메라,
깊이 센서 등 시뮬레이션

동물 행동



쥐 행동 분류와 신체 부위 인식을 위한
AI 학습용 합성데이터

축산 분야



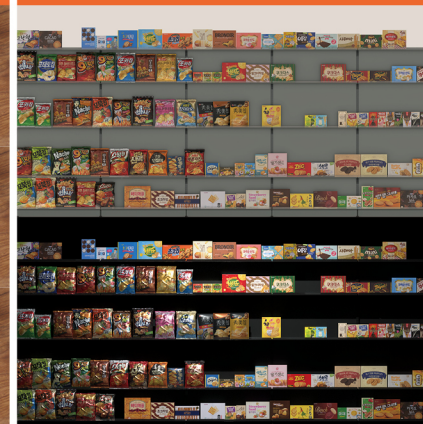
소 행동 양식 분석을 위한 이미지
합성데이터. 디퓨전 생성 모델 사용

식단 모니터링



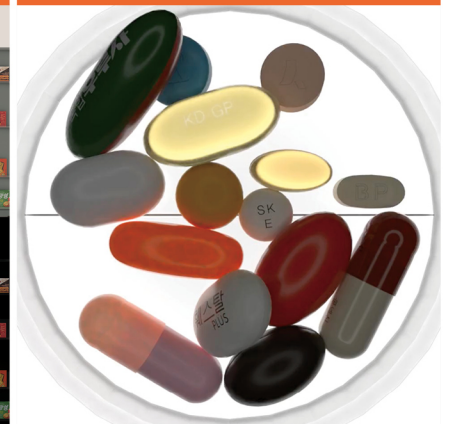
급식 섭취 모니터링을 위한 합성데이터.
3D 모델과 생성 모델을 하이브리드 사용

리테일 / 물류



무인샵 재고관리 인공지능을 위한
합성데이터. 3D 모델링 사용

제약 / 의료



알약 자동 조제 시스템의 비전 검수를 위한
합성데이터. 3D 모델링

품질 진단에 기반한 데이터 최적 경량화 데이터 다이어트

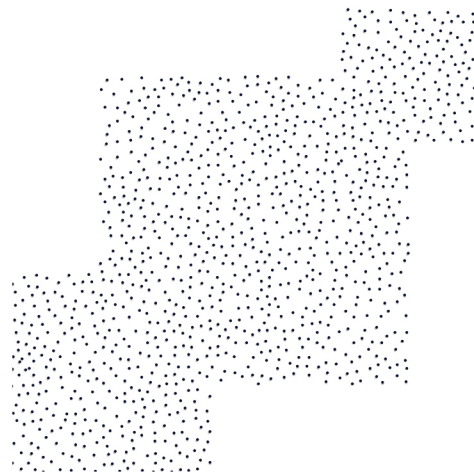
과밀 부분의 데이터 개수를 줄임으로써 성능을 유지하고,
학습 시간 및 GPU 비용 절감함으로써 인공지능 개발 효율을 향상시킬 수 있습니다.

데이터 다이어트 개념

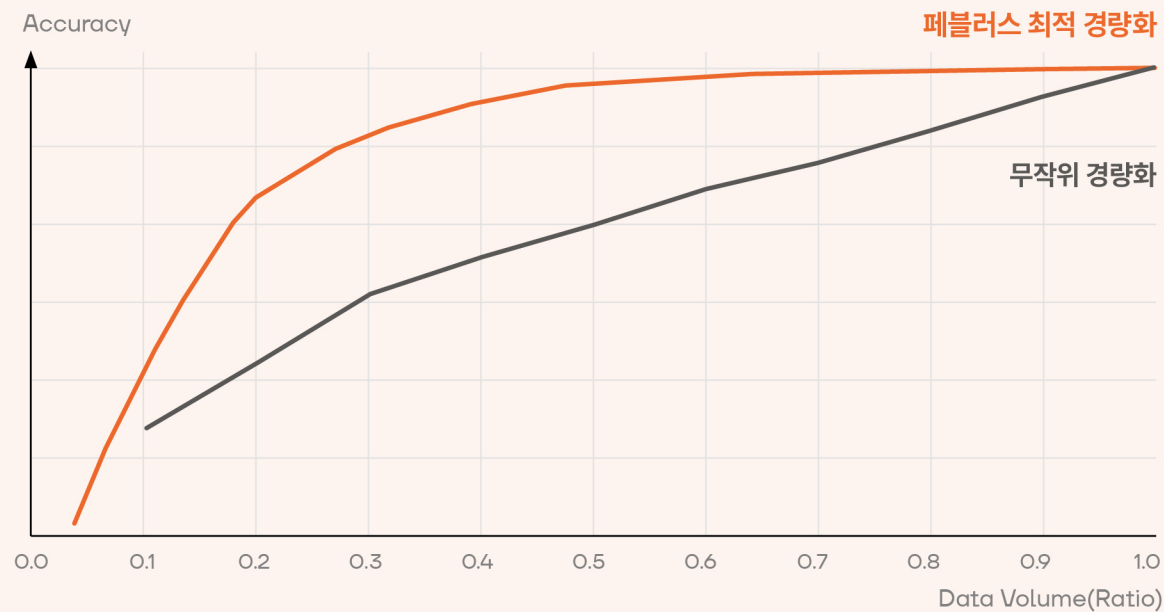
기존 데이터셋 품질 상태



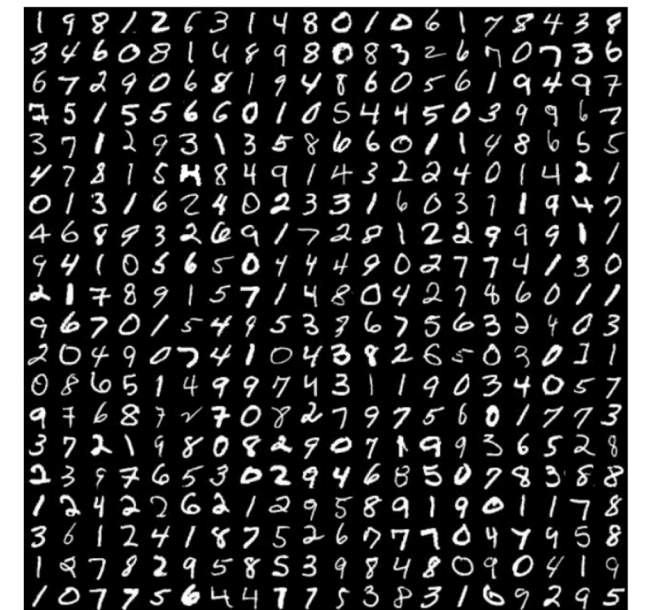
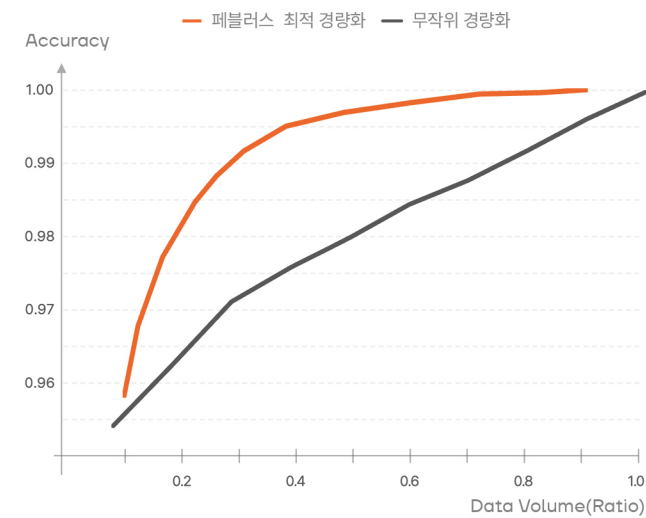
데이터 다이어트



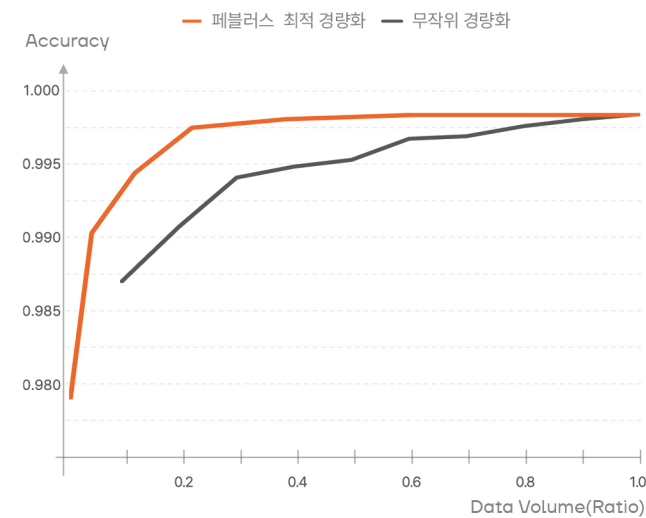
데이터 다이어트 성능 곡선



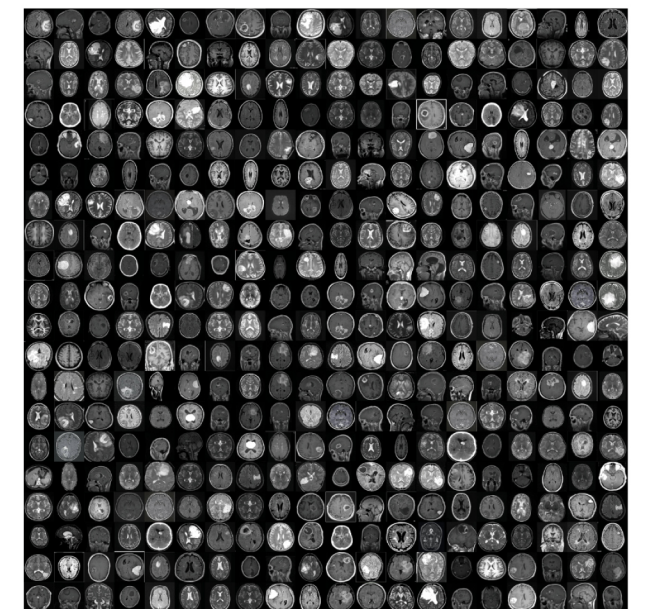
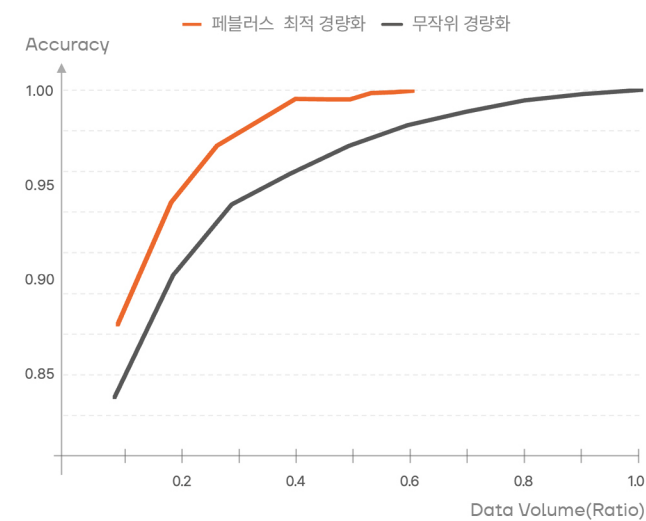
사례 O1. MNIST Dataset



사례 O2. Fire Detection Dataset



사례 O3. Brain Tumor Dataset



PebbloScope

3D 인터랙티브 데이터 커뮤니케이션 도구

‘페블로스코프’는 데이터 클리닉의 결과로 만든 데이터 이미징과 각종 품질 지표를 인터랙티브하게 가시화할 수 있는 데이터 커뮤니케이션 도구입니다.

*설치형 출시 완료. 웹 버전 2024년 하반기 출시

페블로스코프 소개 영상
<https://bit.ly/pebbloscope>

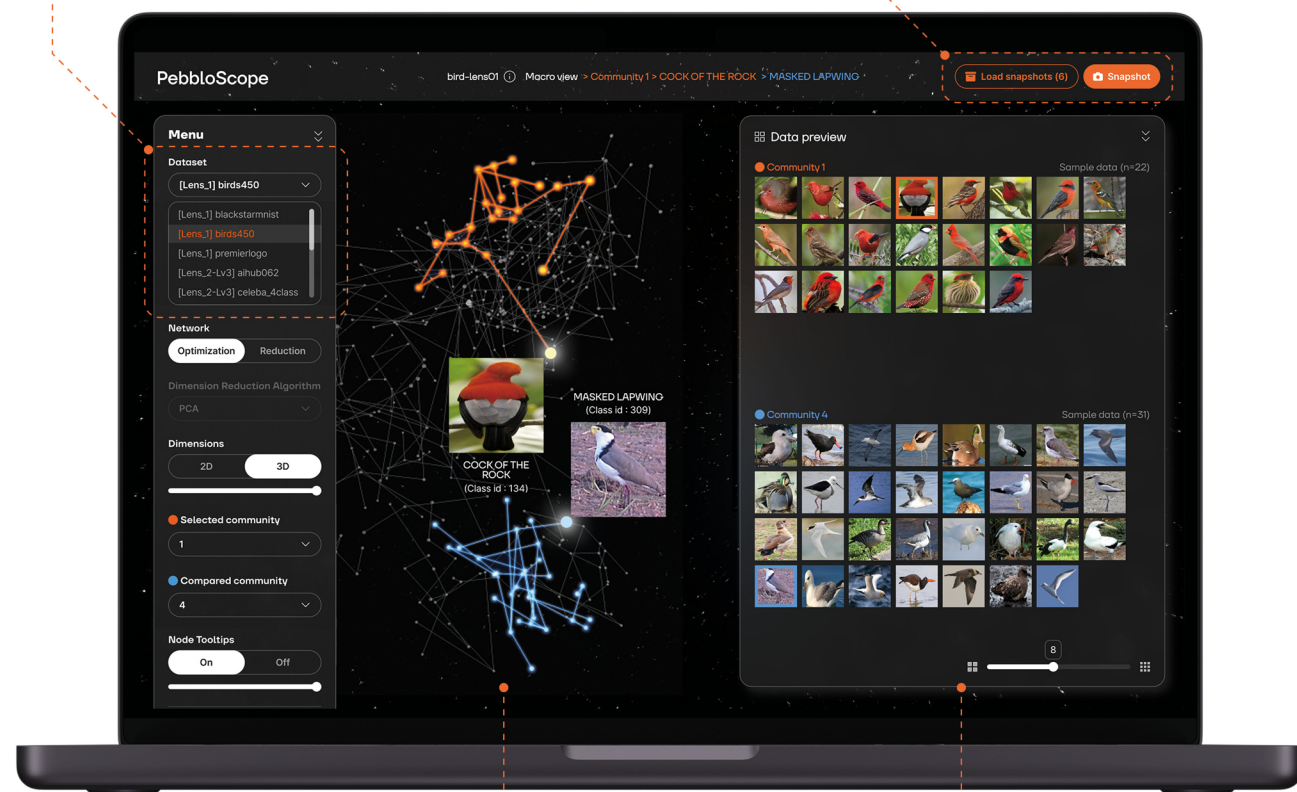


데이터셋 목록

탐색하고자 하는 다양한 도메인의 데이터셋을 선택할 수 있습니다.

스냅샷

선택된 데이터에 대한 3D 특정 시점을 주석 정보와 함께 기록할 수 있으며, 부여된 고유한 웹 주소를 통해 다른 사람들과 인사이트를 공유할 수 있습니다.



3D 시각화

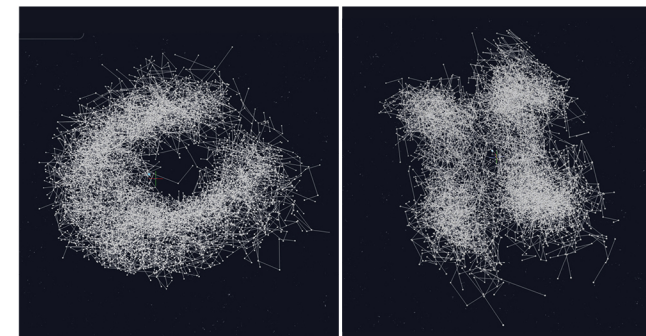
페블러스 데이터 렌즈로 처리된 최적 차원의 데이터를 3D 공간에서 마우스 인터랙션을 통해 탐색할 수 있습니다. 최적 차원 상에서 계산된 데이터 커뮤니티는 유사한 데이터 또는 과밀한 데이터의 분포를 확인하는 데 도움이 됩니다.

데이터 프리뷰

시각화의 노드/엣지에서 선택한 데이터 및 데이터 커뮤니티의 원본 이미지를 확인할 수 있습니다. 데이터 커뮤니티 간 비교를 통해 각각의 커뮤니티가 갖는 이미지의 특징을 더욱 직관적으로 확인할 수 있습니다.

*출시 예정 기능: 인터랙티브 합성데이터 생성 및 데이터 경량화

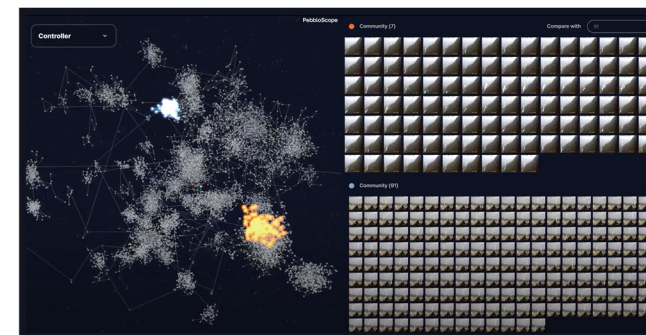
페블로스코프의 특징



최적 차원으로 처리된 데이터의 3차원 시각화

페블러스 데이터 렌즈를 통과한 고차원 인공지능 학습데이터셋이 3차원 공간에 가시화 됩니다. 3차원 공간 탐색을 통해 데이터의 고유하며 독특한 분포적 특성을 관찰할 수 있습니다.

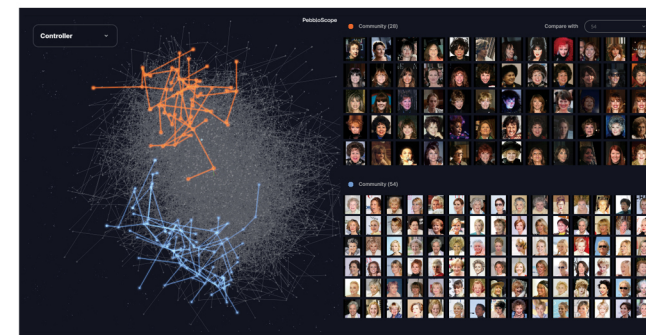
→ 예) 관찰 시점 변경으로 클러스터의 형태 파악



데이터 커뮤니티 탐색을 통한 직관적인 문제점 파악

데이터 커뮤니티는 최적 차원 상의 거리를 기반으로 계산되기 때문에 직관적으로 유사한 데이터를 확인하거나 과밀한 데이터를 파악하는 데 유용합니다.

→ 예) 산불 데이터셋의 과밀 클러스터의 확인



비교를 통한 데이터 특성 확인

데이터 커뮤니티 간 비교를 통해 각각의 클러스터가 갖는 특징을 확인할 수 있습니다. 독특한 분포를 갖는 데이터의 경우 더욱 심도 있는 인사이트를 얻을 수 있습니다.

→ 예) celeba_4class 데이터셋의 동일 클래스 내의 유사 데이터 클러스터 비교