Pebblous Makes Data Tangible

Pebblous.ai

Pebblous

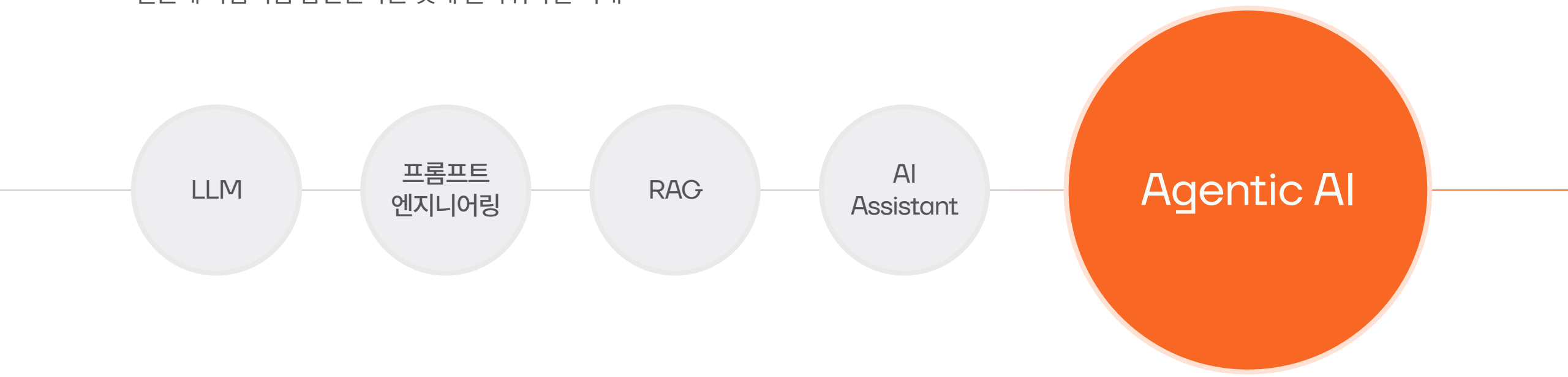
Better Data Makes Better AI

Copyright © Pebblous.ai | All rights reserved

Pebblous

AI 패러다임의 전환: Agentic 시대 도래

질문에 사람처럼 답변한다는 것에 놀라워하던 시대



의도를 이해하여 자율적인 액션과 계획이 가능한 시대

Agentic AI

AI 패러다임의 전환: Agentic 시대 도래



" 올해의 핵심 트렌드는

LLM 모델 자체의 발전보다

에이전트 워크플로우가 될 것 "

앤드류 응(Andrew Ng) 교수, 2024

" Agentic AI는 LLM 을 활용하여 동적 환경에서 기능하며,

생성적 출력을 특정 목표 달성에 적용하는 능력을

확장하는 시스템 "

IBM

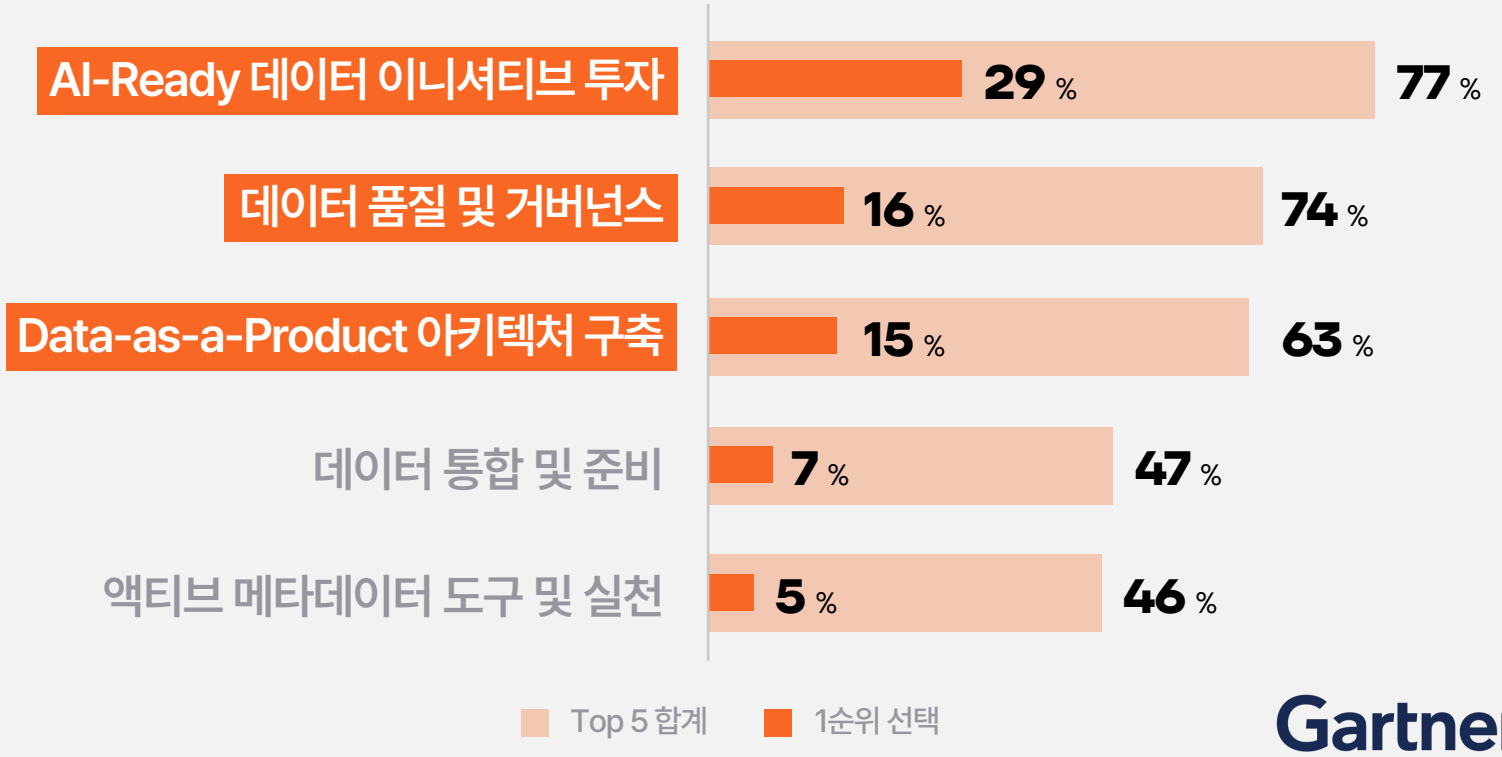
AI 전환(AX)의 병목은 인력 기반의 데이터 품질 관리

90%

“ AI 도입 시 데이터 처리에 소모되는 시간이 전체 절차의 90% “

- 현대자동차 OO부문 AI팀 -

데이터 관리 분야
투자 트렌드 Top 5는?



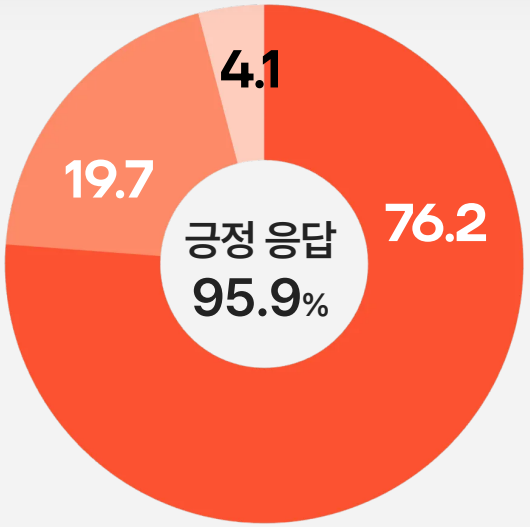
Gartner

※ 출처: 2024 데이터 관리의 진화 설문조사, Gartner, 247명 참여(응답 중 "잘 모르겠다" 선택 제외)

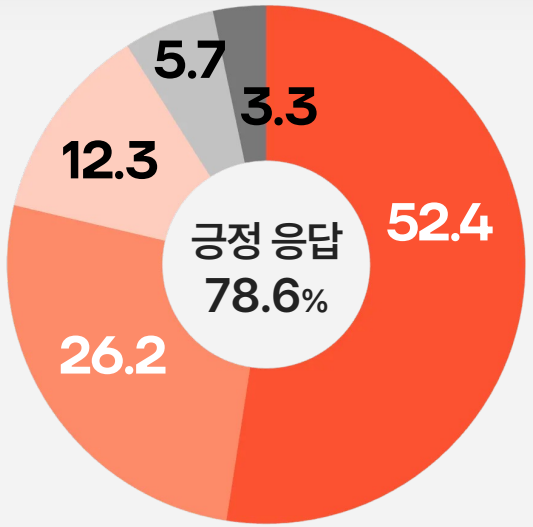
국내 기업/기관들의 데이터 품질 현황은?

(단위: %)

"데이터 품질 문제가 **비즈니스에 영향**을 준다



"데이터 품질 개선을 위한 **솔루션 도입이 시급**하다



매우 동의함 동의함 보통 동의하지 않음 전혀 동의하지 않음

※ 출처: Data Clinic 2.0— 에이전트 AI 데이터 과학자 (AADS) 기술을 적용한 차세대 데이터 클리닉 수요조사, 2025, 페블러스, 122명 참여

Pebblous

올인원 데이터 관리 솔루션

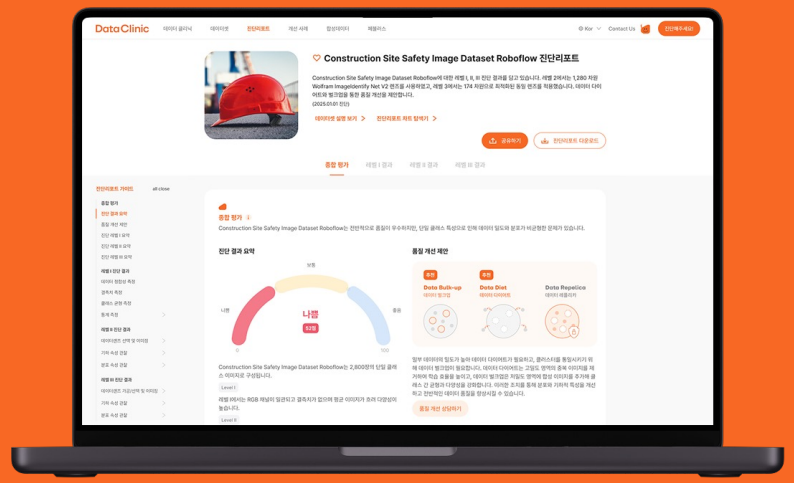
01
품질평가에서
개선까지

02
인터랙티브 가시화
커뮤니케이션

03
멀티모달
데이터셋

04
SaaS,
온프렘, API

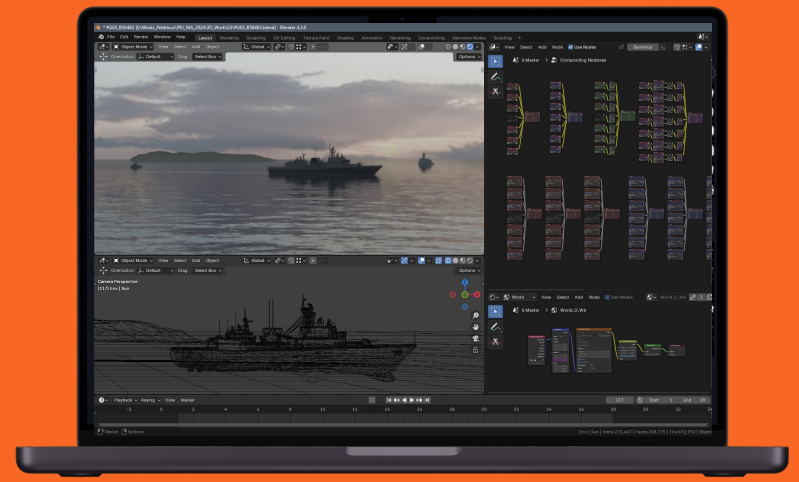
Data Clinic



PebbloScope



PebbloSim



Pebblous

AI 패러다임의 전환: Agentic 시대 도래

생성을 넘어 도구의 사용과 계획까지

01



생성

진단리포트 자동생성 및 챗봇

02



도구 수행

데이터클리닉 연동

03



계획

복잡한 품질 규제 대응

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.

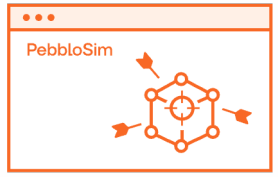
* EU AI Act는 2026년 8월 본격 시행 예정
AI 규제 완화안, 법 적용 1년 유예

Pebblous

데이터 그린하우스 데이터가 스스로 성장하는 지능형 운영 체제

Zero
Physical
Hallucination

ISO
Certification
Ready



PebbloSim

스스로 생성하고

- 핵심 실행 엔진. 디지털 트윈 기반
- 물리 법칙 기반의 합성 데이터 생성
- 부족한 Edge Case를 정밀 타격하여 보강

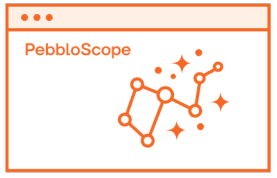


Data Clinic

데이터를 진단하고

- 데이터의 결핍과 과밀을 정밀 진단
- '무엇을 고쳐야 하는가' 자동 처방
- 개선을 위한 진단리포트 발행

Data Communication



PebbloScope

투명하게 통제함

- Human-in-the-Loop 게이트웨이
- 뉴로-심볼릭 양방향 시각화
- 시뮬레이션 과정 실시간 모니터링

Data Diet

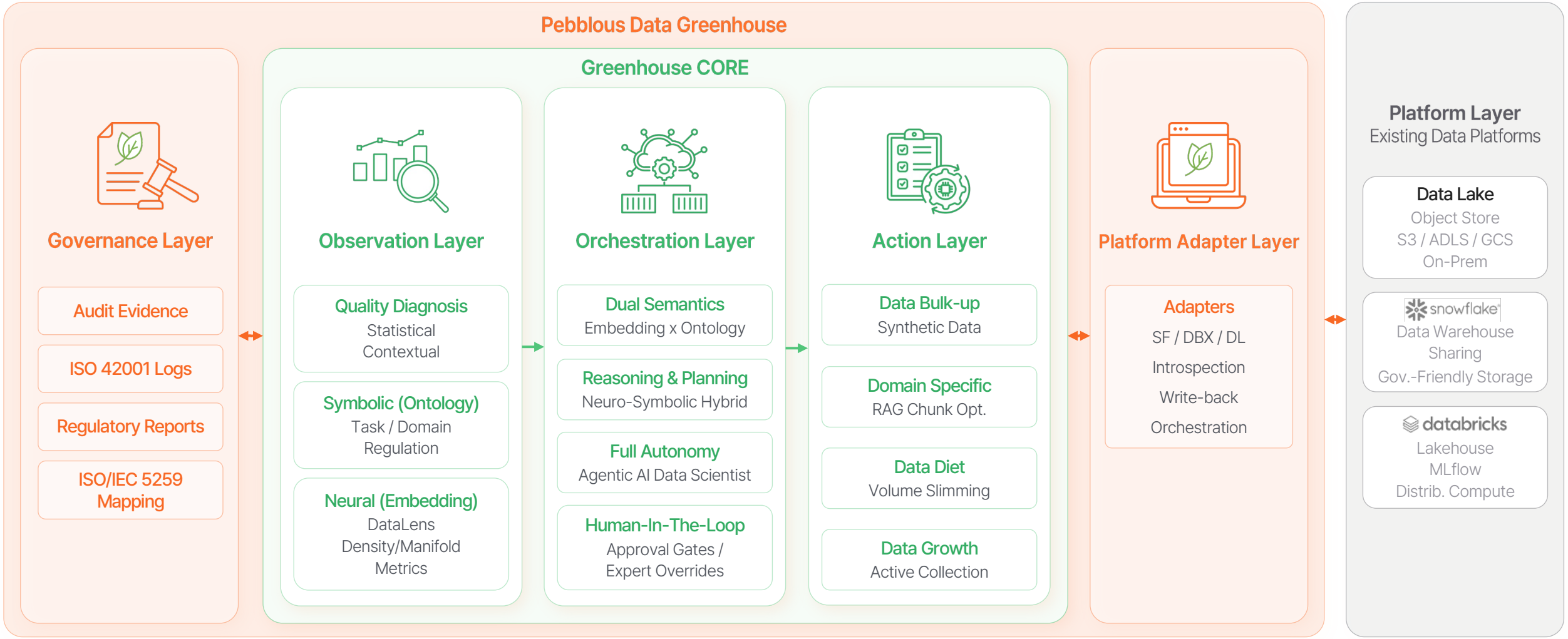
데이터 그린하우스
OS

Data Bulk-up

Data Greenhouse: System Architecture

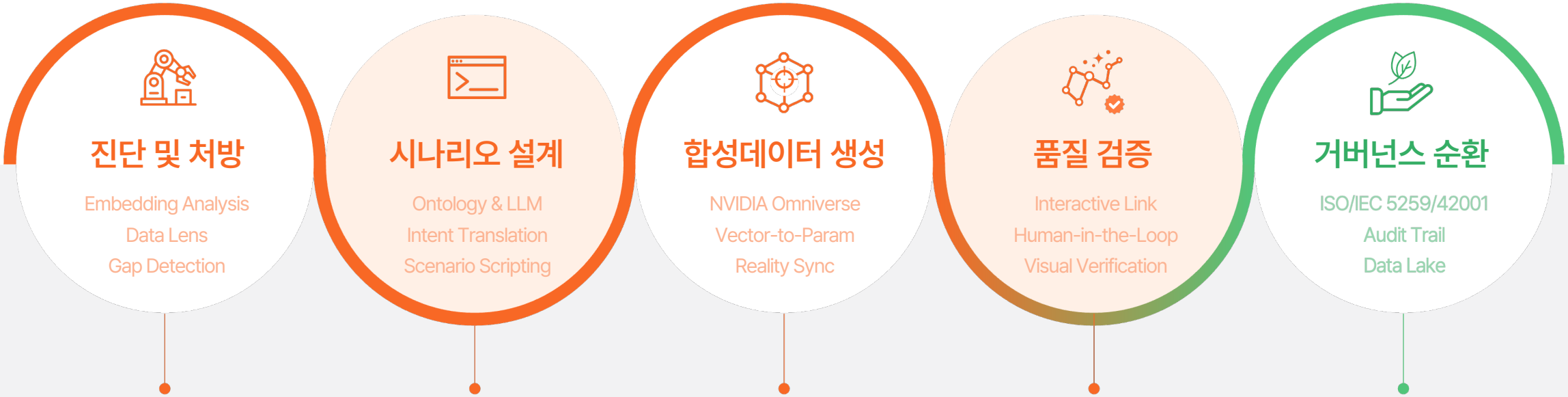
Pebblous

데이터 그린하우스 AI 데이터 플랫폼의 '책임 레이어'



Pebblous

데이터 그린하우스 AADS 기반의 'End-to-End' 자율 운영 파이프라인



데이터클리닉

데이터 품질을 관리하여
정밀 시나리오 설계도로 활용

GenSim 관리자

추상적 명령을 구체적
시뮬레이션 스크립트로 번역

PebbloSim 생성 엔진

실제 공장 환경 구축하여
부족한 데이터 정밀 타기팅

PebbloScope 시각화

데이터가 의도한 영역에
분포하는지 시각화 모니터링

데이터 그린하우스

AI 모델 재학습에 활용
감사 로그 기록 및 규제 대응

Pebblous

데이터 그린하우스 적용 시나리오

	모빌리티	조선	국방
	현대자동차 스마트팩토리 용접 불량 검출	**한국 조선소 건조 일정 예측 & 최적화**	**육군/해병대 전술 훈련 합성데이터 생성**
문제 상황	연간 희귀 불량 패턴 발생률 0.001% 미만 **→ 데이터 부족, 결함 검출 한계, 비용 손실**	수천 개의 공정이 복잡하게 얽힌 초대형 과제 **→ 데이터 복잡성, 소통 장벽, 비용 손실**	극도로 제한된 환경에서 이뤄지는 군사 훈련 **→ 데이터 희소성, 보안 제약, 비용 손실**
솔루션 적용	• 현장 데이터 수집 • 합성 불량 데이터 생성 • AI 모델 재학습 & 배포	• 멀티모달 데이터 수집 • 데이터 통합 & 디지털 트윈 구축 • PebbloScope 3D 시각화 & 현장 소통 강화	• 기존 데이터 수집 & 품질 진단 • Hyper-Synthetic 전투 시나리오 생성 • AI 모델 학습 & 시뮬레이션 통합
기대 효과	불량 검출 정확도 ▲94 % 향상 불량률 ▼78.75% 감소 데이터 수집 시간 ▼99% 단축 수동 검사 시간 ▼83% 단축	건조 일정 예측 정확도 ▲76 % 향상 현장 관리자 데이터 활용률 ▲ 300% 향상 병목 발견 시간 ▼95% 단축 일정 지연 건수 ▼67% 감소	AI 모델 성능 ▲50 % 향상 데이터 생성 기간 ▼99.9% 단축 훈련 비용 ▼95% 감소 데이터 라벨링 비용 ▼100% 감소







Pebblous

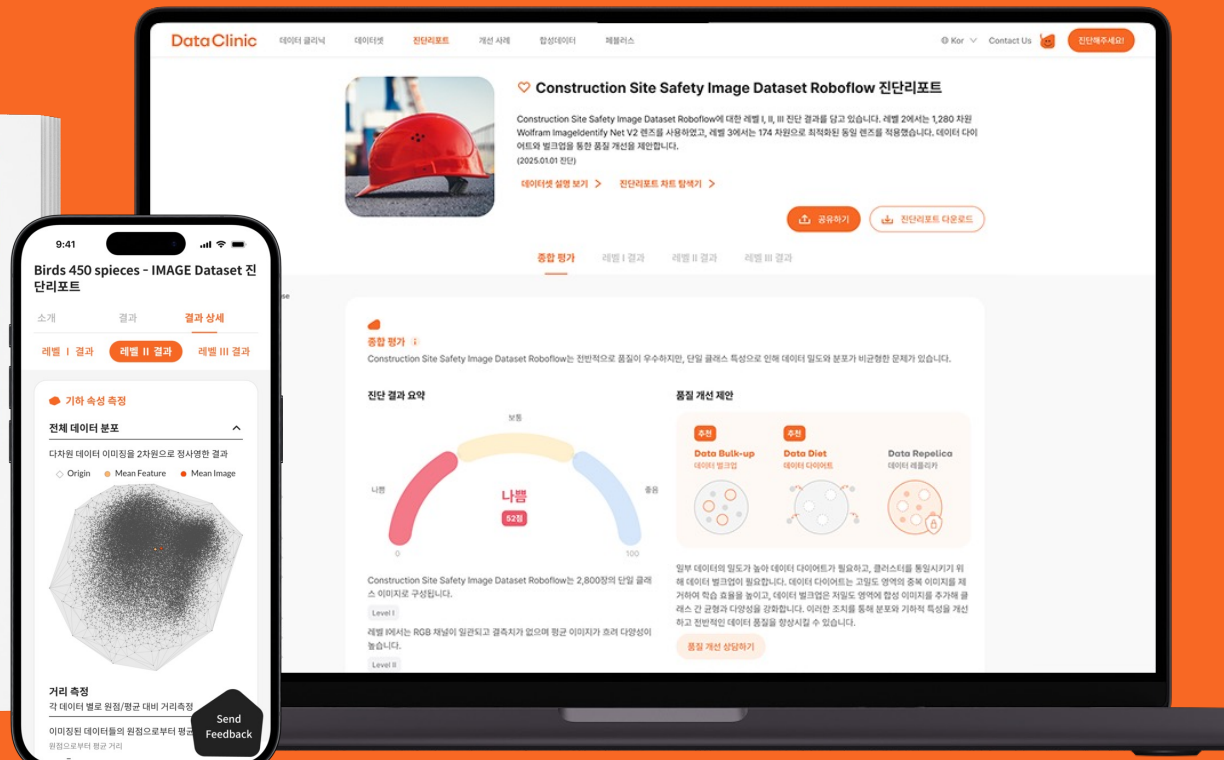
DataClinic

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

웹 버전

PDF 버전





모바일 버전

Pebblous

AI 데이터셋의 품질로 인한 문제들

모호한 데이터 가격 체계, 모델 성능 저하, GPU 자원 낭비, 높아가는 AI 규제 대응 미흡

01



학습 데이터

품질 불량 / 고갈

02



데이터센터

운용 비용 / 효율

03



AI 관련

법 / 규제 강화

EU AI Act: 전체 매출 7% 또는 3,500만 달러 벌금 부과
한국 2026년 AI기본법 시행

Pebblous

통합적인 데이터 품질 관리 솔루션

데이터의 인공지능 적합성(AI-Ready)을 고려한 지속적인 품질 진단 및 개선

01



AI 학습 데이터
품질 진단/관리

모델 성능 개선

02



데이터 경량화
합성데이터

학습 효율 개선

03

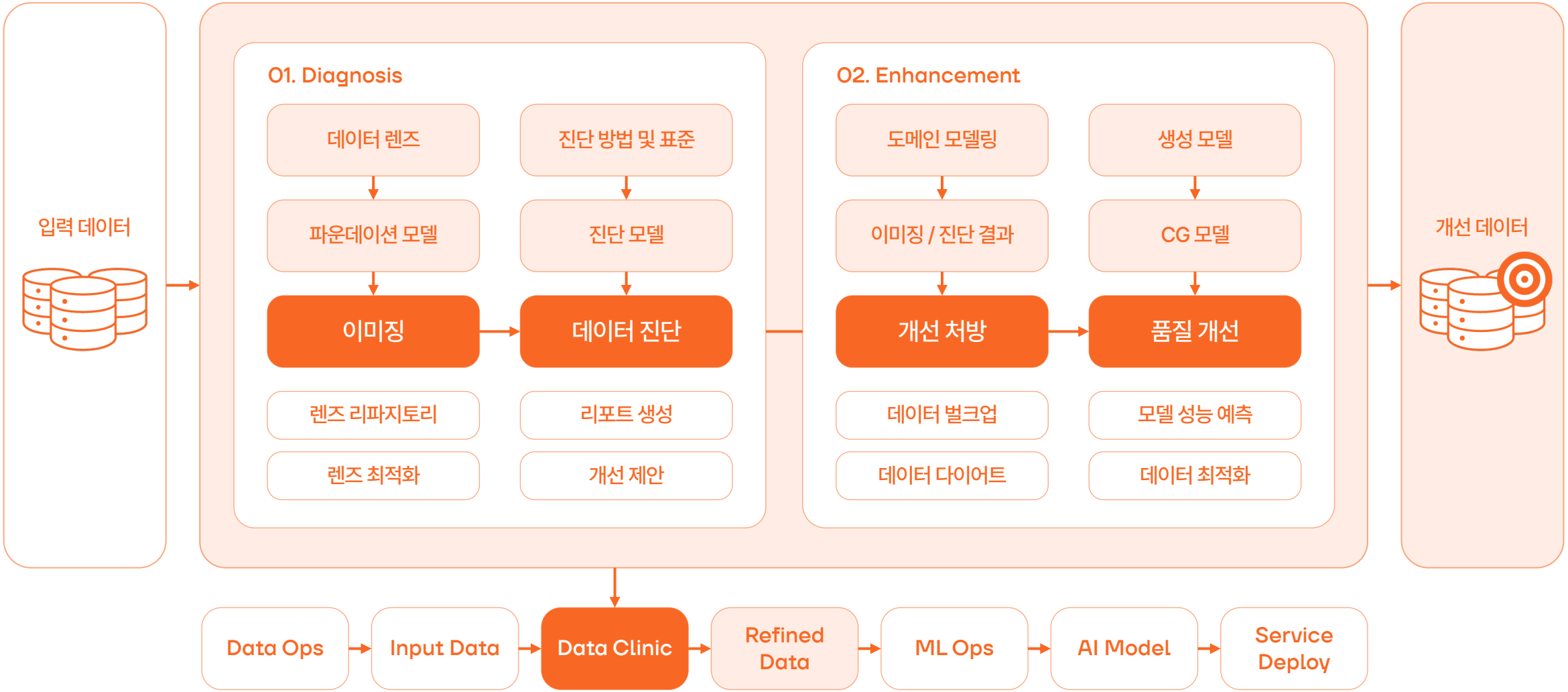


AI 적합 데이터
기준 모니터링

거버넌스 구축

Pebblous

데이터 품질 진단 및 개선 솔루션



데이터 품질 개선 솔루션

비어있는 데이터 영역 · 중복 데이터 영역 · 실제 데이터 · 합성데이터

데이터 레플리카

원본의 내용을 보호하면서도
통계적·분포적 특성이 유사한 가상데이터를 생성

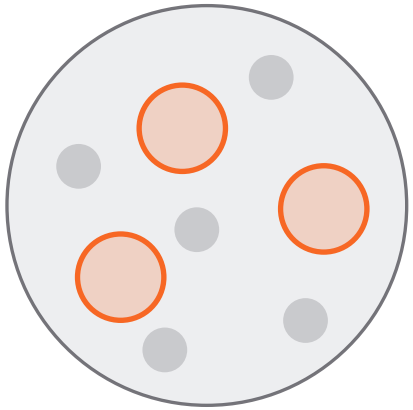


→ 개인정보 유출 리스크 예방

활용 예) 개인정보 보호를 위한 데이터 익명화

데이터 벌크업

데이터가 부족한 부분에 데이터셋에 적합한
정밀 타게팅 합성데이터를 생성하고 제공

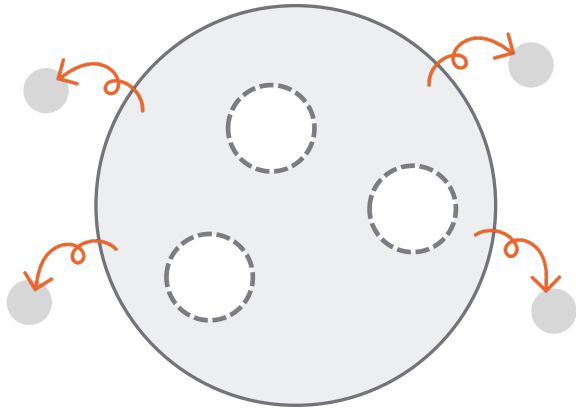


→ 데이터 수집 비용 절감, AI 성능 향상

활용 예) 예외 상황에 대한 합성데이터 생성

데이터 다이어트

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



→ AI 학습 비용 절감, 개발 효율 제고

활용 예) 학습 과정에서의 데이터 경량화

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.

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
© 2024 Gartner, Inc. and/or its affiliates. All rights reserved. CM_GTS_2952789

기존에서 완전히 벗어나, 새로운 전략이 필요합니다.

- "리스크가 사람 책임 Skills-managed 에서 시스템 책임 System-based 으로 넘어갈 때 비로소 데이터는 진짜 AI-Ready가 된다."
- AI-Ready 데이터를 위한 기능적 요구사항
  - 모드별 필수/선택 항목 점검
- 깊이 우선(Depth-first)
  - 자동화보다 "메타데이터 충실도"가 리스크 결정
- 메타데이터 그래프(Metadata Graph)
  - 구조·비정형 자산을 체이닝 (chaining)하여 그래프/벡터 관리
  - 대표성·맥락 격차 상시 감시

범주	데이터 요구 사항	의미	평가의견
기본 정합성 *Traditional Alignment*			
Quantification	**정량화**	정형·비정형 데이터 자산의 레코드/오브젝트 수량이 파악됨	
Observability Metrics	**가시성 지표**	값 분포·변화 패턴·접근률·업무 도메인별 사용 사례 수를 모니터링	
Consistent Content	**내용 일관성**	동일한 정보가 다른 형식·구조로 존재할 때 교차 참조로 값 일치 확인	
Verified Source Capture	**소스 수집 검증**	설계 규칙에 따라 원천 데이터가 정확히 수집되었는지 품질 규칙으로 확인	
공정 사용 적합성 *Qualified Usage ("Fairness")*			
Contextual Quality	**맥락 품질**	동일 개념을 나타내는 여러 값(예: boy/girl vs. male/female)의 동등성 확인	
Lineage	**데이터 계보**	데이터 출처에서 사용 사례까지의 전체 경로 및 분기(branch)를 추적	
Validated Authority	**권한 검증**	승인된 시스템/제공자에서만 데이터가 공급되는지를 계보로 확인	
Diversity	**다양성 확보**	같은 출력에 기여할 다양한 데이터 유형/값(예: PSI 측정값 & 계산값) 확보	
SLA Performance	**SLA 성능**	읽기·전송·편집 지연이 결과 계산에 미치는 영향 인지	
Variable Access Rights	**가변 접근 권한**	권한 있는 경우만 보호 데이터 접근, 무권한 시스템은 차단	
Multi-content Metadata (RAG 등)	**다중콘텐츠 메타데이터 연결**	태깅·감정·톤 등 정성 메타데이터를 일관되게 부여해 RAG 활용	
Semantic Context ↔ KG	**의미 맥락-지식그래프 연계**	편향 제거, 그래프 상 대안 경로 확보·검증	
Use-Case Commonality	**사용 사례 공통성**	같은 문제를 다루는 여러 팀·시스템이 "권한·데이터·로직"을 조직 전체에서 일관되게 적용	
운영 신뢰성 인증 *Certified for Production ("Trust")*			
Versioning	**버전 관리**	데이터·스키마 버전을 문서화, 추세 분석 가능	
Train↔Prod Data Verification	**학습↔운영 데이터 검증**	보호·품질·무결성 위반 발생 시 변동 기록 및 패턴 분석	
Risk Compliance	**위험 준수**	기밀·개인·경쟁·법적 책임 등 데이터 분류 제공	
Regulatory Compliance	**규제 준수**	보고 최소 기준·참조 데이터·카테고리 변환 사항을 문서화	
AI Use-case ↔ Technique Alignment	**AI 기술-과제 적합성**	모델·기법이 사용 목적에 대해 안정·효율·빈도 면에서 적합	
Ethical Exposure Risk	**윤리 노출 위험**	외부·기업·지역별 거버넌스 요구를 구분, 매체·타이포 규범 파악	
Data Share Logs	**데이터 공유 로그**	의도적·비의도적 데이터 누출 모두 기록·분석, 제한 위반 여부 체크	
Observability of Content Use Cases	**콘텐츠 사용 가시성**	유사·미사용 콘텐츠 교차 참조, 흡수·폐기율 추적	
확장·합성 대비 *Augmentation-Enabled Data Readiness*			
Regression Testing (Validation, Poisoning)	**회귀 테스트**	데이터 변경이 기존 분석 결과를 바꾸는지 지속 검증	
Continuous Profiling & Analytics	**연속 프로파일링·분석**	값·무결성·크로스참조 패턴 변동을 그래프 분석	
Change Recognition & Alerting	**변화 인지·알림**	변화 속도 가속/감속/불안정 시 예외 지표로 모델 경고	
Explainable Inferred Calculations	**추론 계산 설명성**	모든 산출값의 원천(Seed) 데이터를 계산 경로로 추적 가능	
Presumed Reuse of Derivations	**파생 재사용 가정**	Seed 식별 불가 파생값은 모델/합성 데이터로 명시해 가정 과다 노출	

Gartner사의 AI-Ready Data 평가 항목 (재가공)

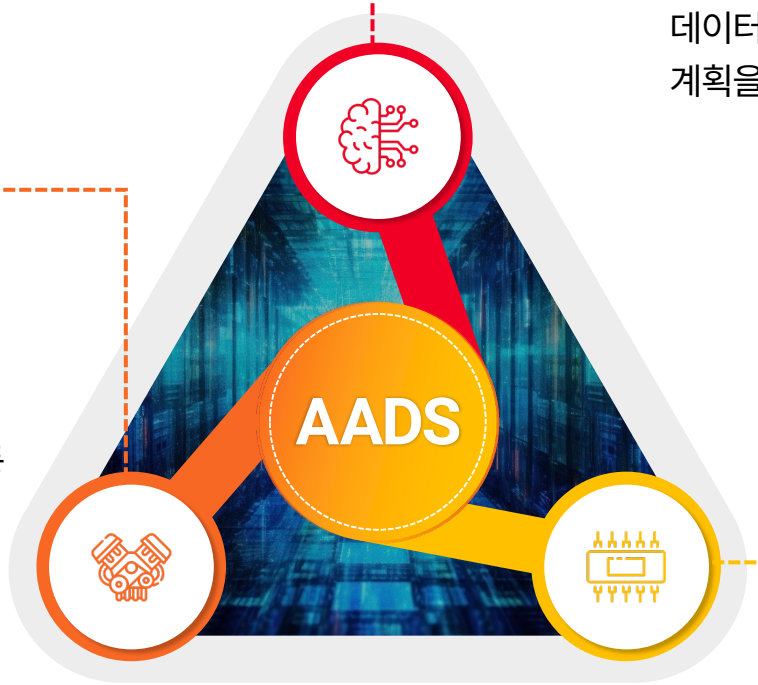
AADS 시스템 구조

The Brain 파운데이션 모델

- **KONI LLM (KISTI) :** 과학기술 분야에 특화된 LLM. 데이터의 의미와 맥락을 깊이 이해하고, 복잡한 문제 해결 계획을 수립하는 **두뇌 역할 수행**.

The Engine 핵심 실행 엔진

- **Data Clinic® (페블러스) :** 데이터 이상치 탐지, 결측치 보정, 편향성 제거 등 **실제 데이터 품질을 개선**하는 30건 이상의 특허 기반 엔진.
- **Governance Engine :** ISO 42001, EU AI Act 등 복잡한 규제 요건을 실시간으로 검증하고 감사 리포트를 자동으로 생성하는 규제 대응 전문 모듈.



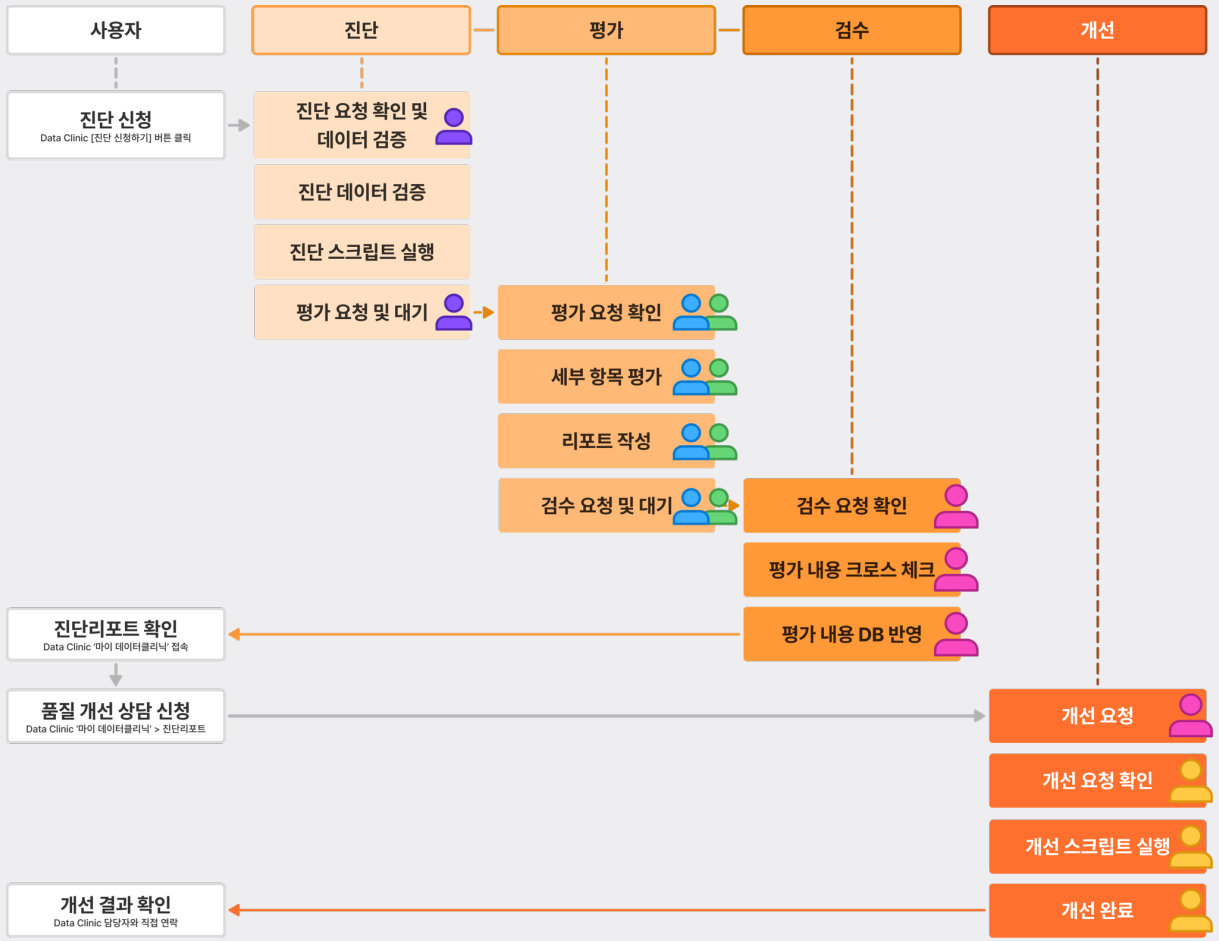
The Conductor 에이전트 오케스트레이션

- **Planner :** 사용자의 목표(데이터 품질 개선)를 받아 최적의 작업 계획(Workflow)을 자율적으로 설계.
- **Tool Router :** 계획에 따라 필요한 도구(API)를 순서에 맞게 호출하고 실행을 총괄 지휘.

Pebblous

AI 데이터 품질 관리, 자율형 AI 기술로 확장

기존 방식 Data Clinic 1.0



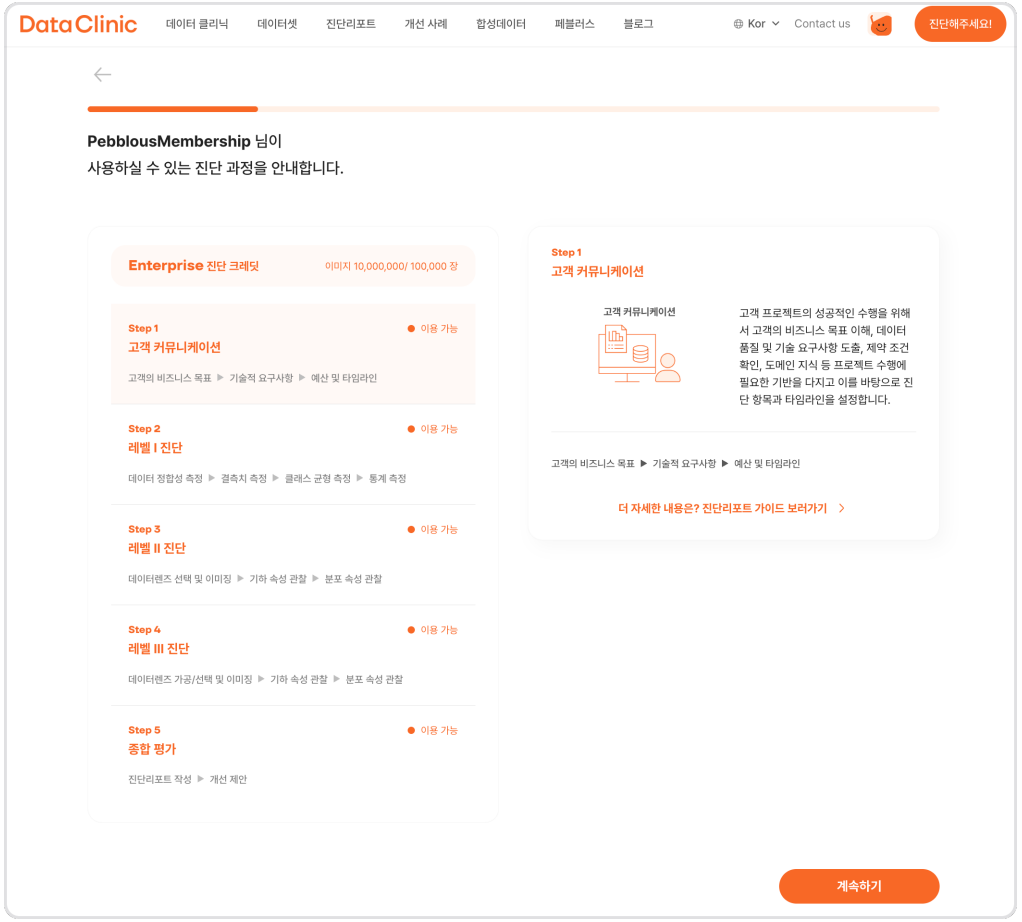
AADS Data Clinic 2.0



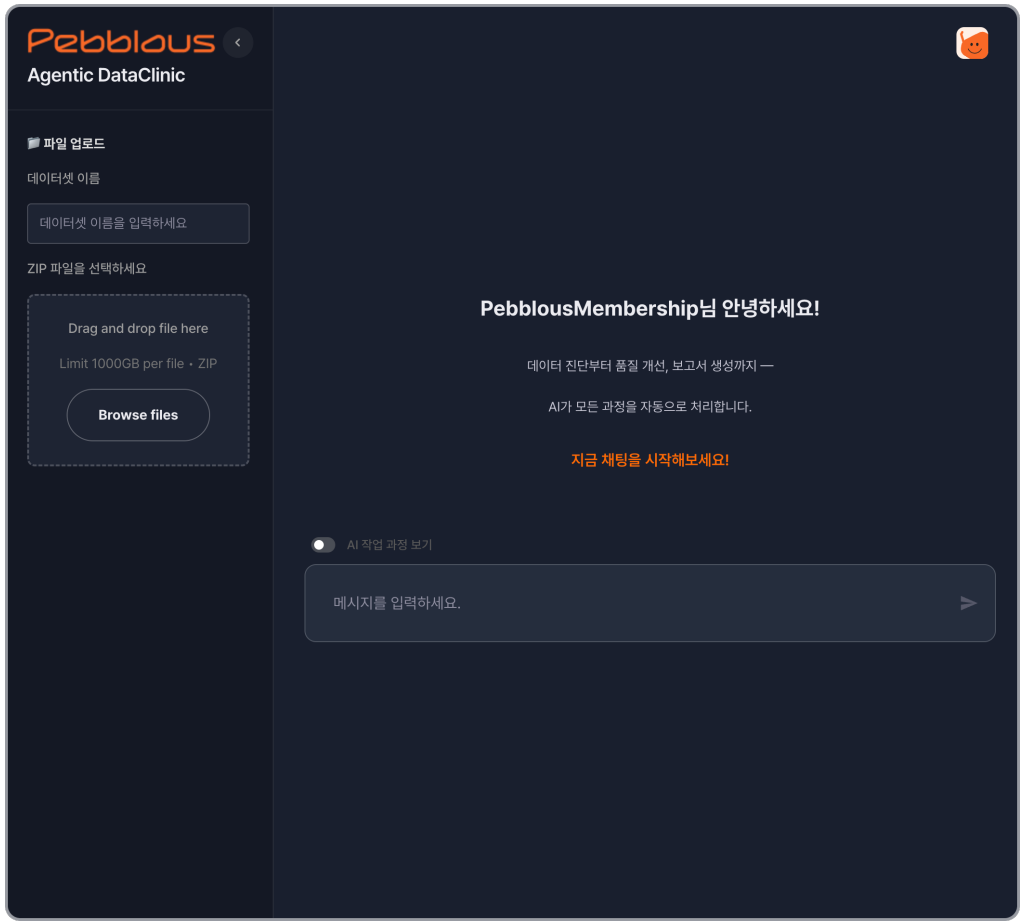
Pebblous

절차적 방식에서 Agentic AI로의 전환

기존 방식



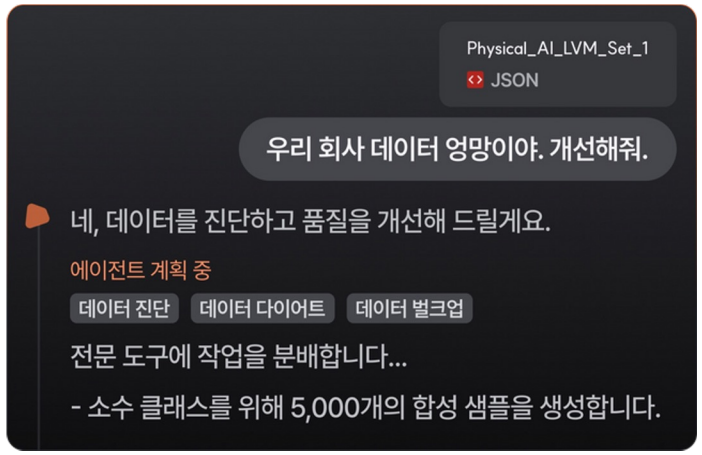
Agentic AI 방식



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.

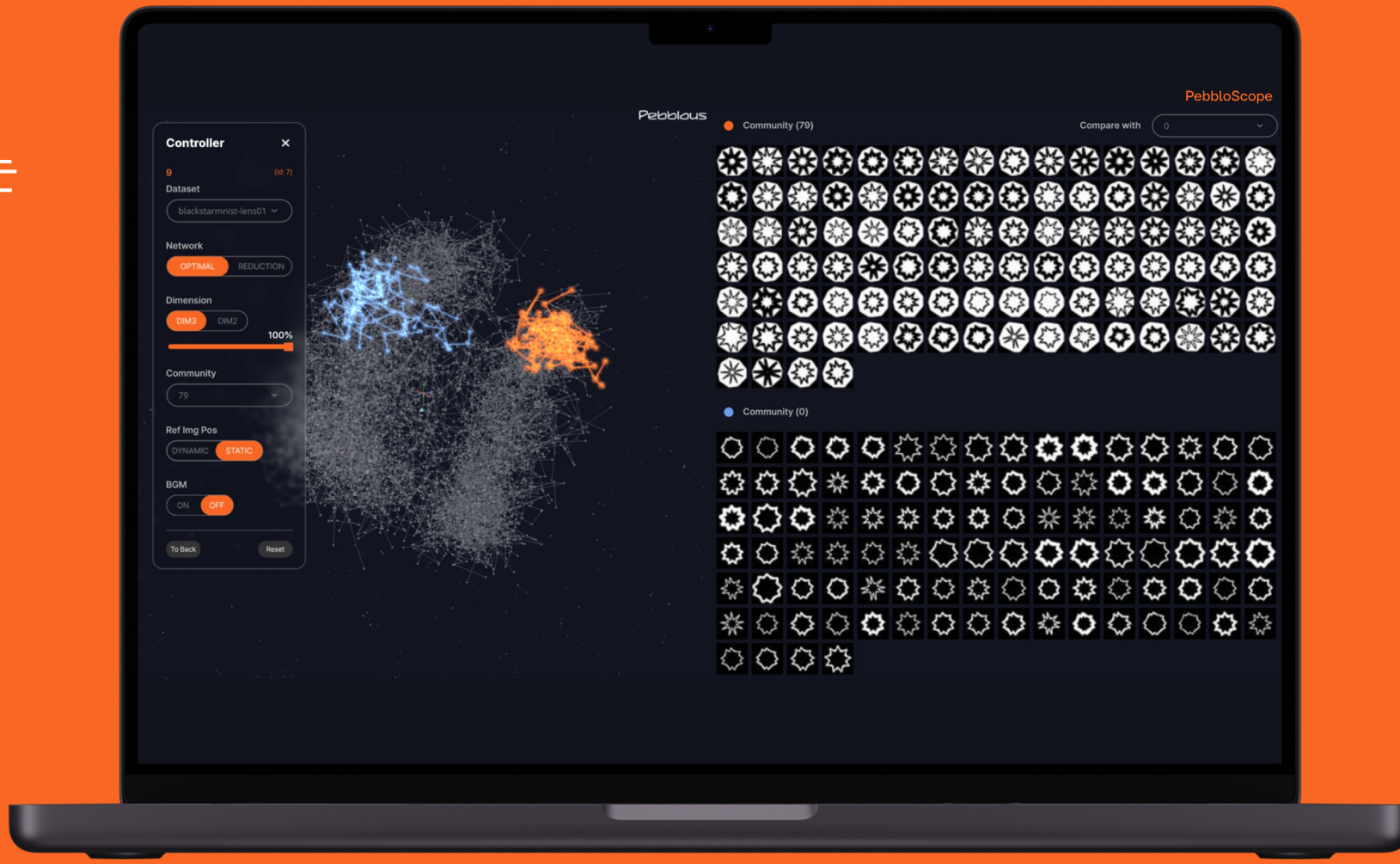
프롬프트 단 한 줄만으로 업무의
효율이 상승합니다. 데이터를
하나씩 들여다보지 않아도,
프롬프트 한 번이면 데이터 품질
진단, 개선, 보고서 추출까지
AADS가 대신해드립니다.

Pebblous

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 02

페블로스코프

Pebblous



시각화 패널

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.

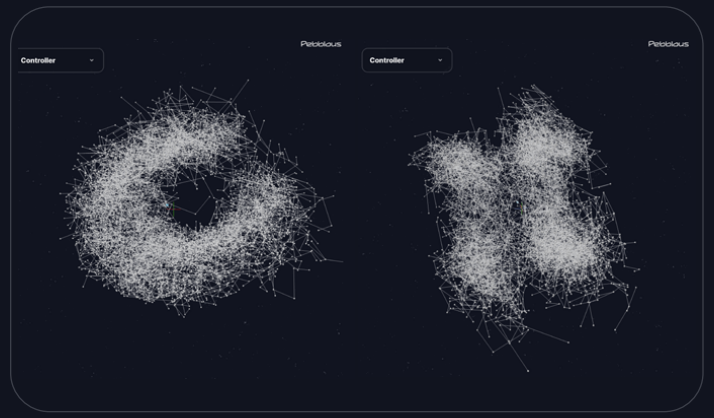
이미지 패널

시각화 패널에서 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.

페블로스코프 데모 영상

Product 02

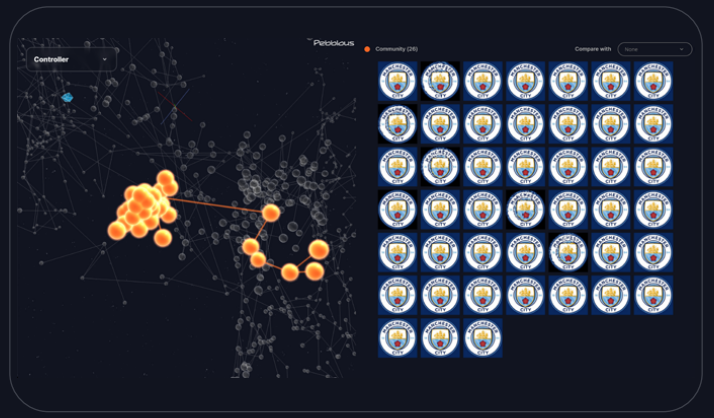
페블로스코프 특징



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.

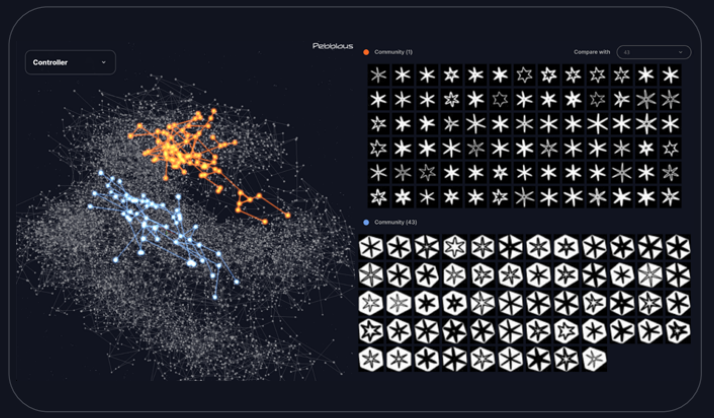
예) 관찰 시점 변경으로 클러스터의 형태 파악



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.

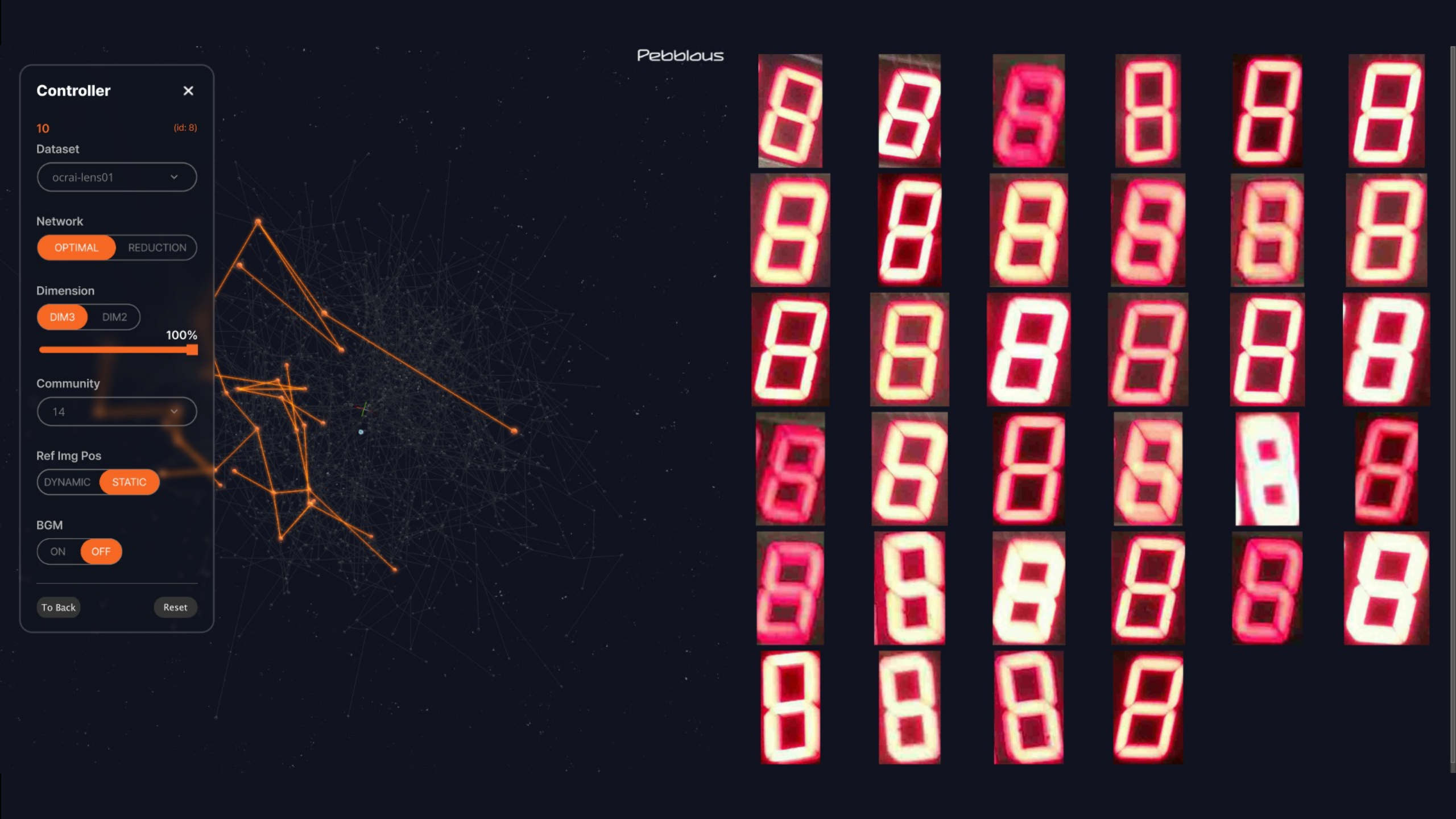
예) 캐글 로고 데이터셋의 과밀 클러스터의 확인

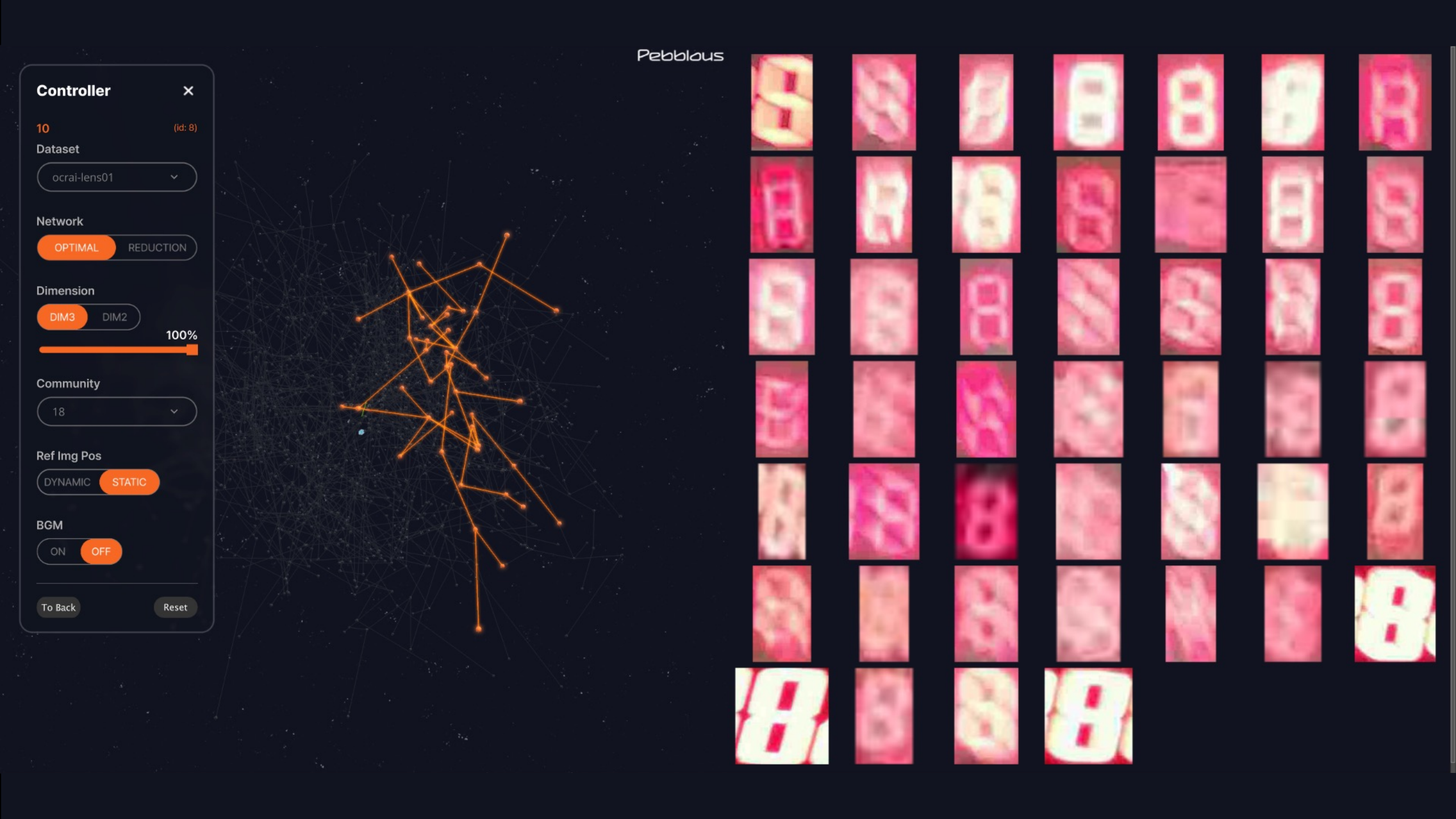


비교를 통한 데이터 특성 확인

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

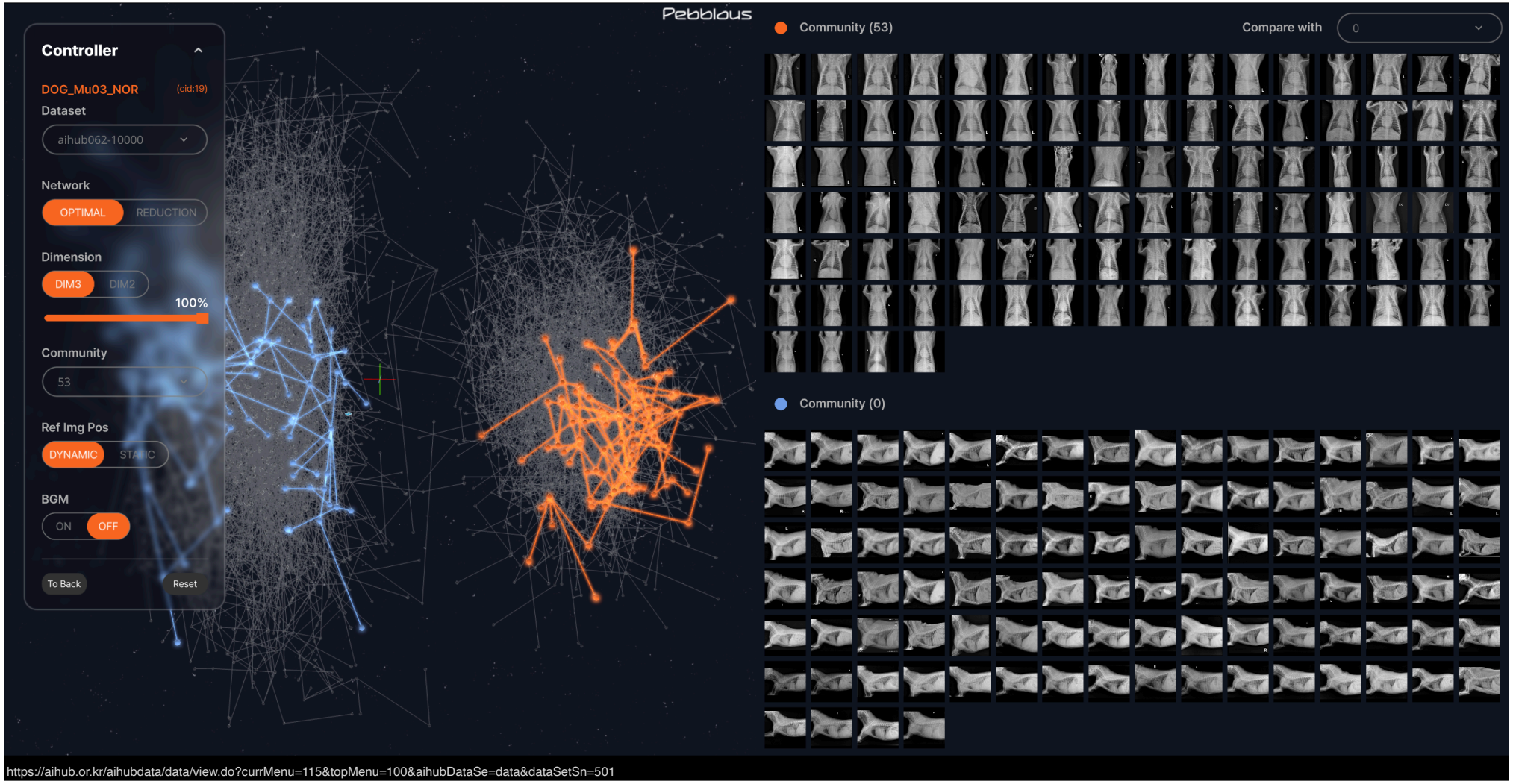
예) 동일 클래스 내의 유사 데이터 클러스터 비교


Pebblous
Controller
10
(id: 8)
Dataset
ocrai-lens01
Network
OPTIMAL
REDUCTION
Dimension
DIM3
DIM2
100%
Community
14
Ref Img Pos
DYNAMIC
STATIC
BGM
ON
OFF
To Back
Reset


Pebblous
Controller
10
(id: 8)
Dataset
ocrai-lens01
Network
OPTIMAL
REDUCTION
Dimension
DIM3
DIM2
100%
Community
18
Ref Img Pos
DYNAMIC
STATIC
BGM
ON
OFF
To Back
Reset

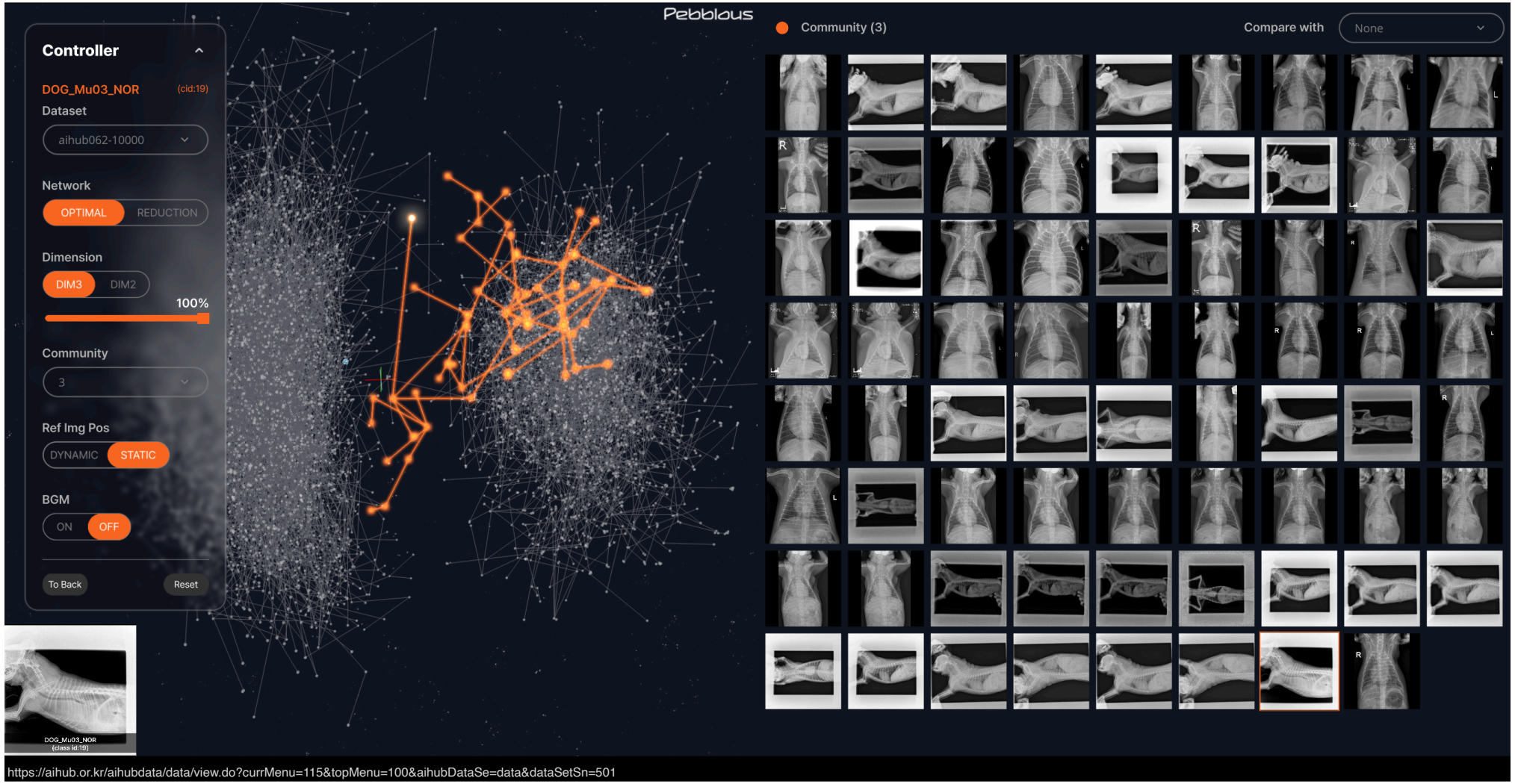
Pebblous

Pet-bone Dataset



Pebblous

Pet-bone Dataset



Menu

Dataset

WikiArt (Lv3.2_Lens1.10/dH...

Network

Optimization | Reduction

Dimensions

2D | 3D

Node color mapping

Normal

Selected community

3-1

Compare community

3-17

Node Tooltips

On | Off

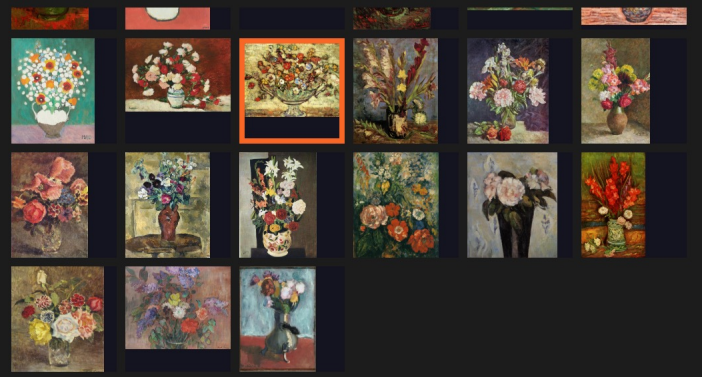
Reset



Data preview

Sort by: Density: mAllm (For nearest all da...

Size: 6



Community 3-17 Sample data (n=30)



E-mail | contact@pebblous.ai Tel | 044-589-3824 Office | South Korea Sejong Seoul Daejeon Pohang Web | pebblous.ai

Copyright © Pebblous.ai All rights reserved

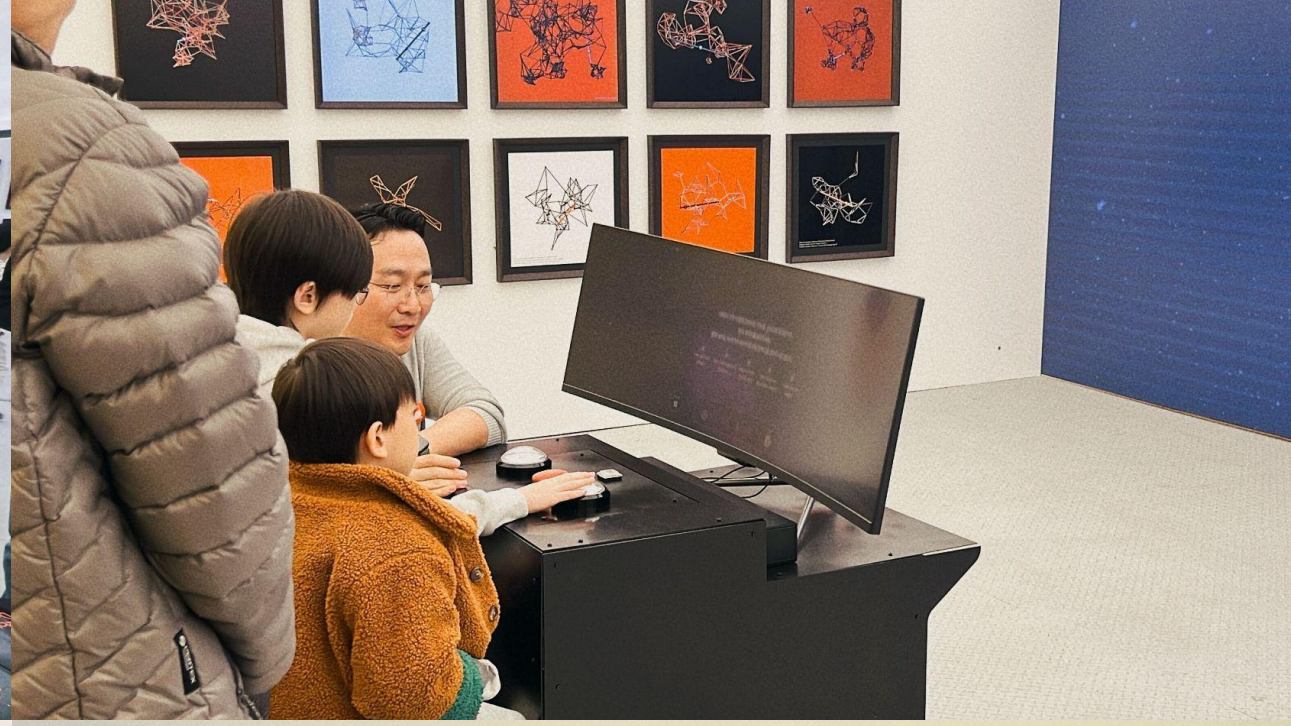
Exhibition #16 Tangible Data @ 현대차 제로원 크리에이터 데모데이. 2024.10






Journey Start
Select Destination
Explore Highlights
Create Your Own
Archive Creation
End Journey
Archived Data 0 / 5
Step 0 1 2 3 4 5 6 7 8 9 10
Restart
B
New Creation
L
Archive Creation
End Journey
R
B
L
R





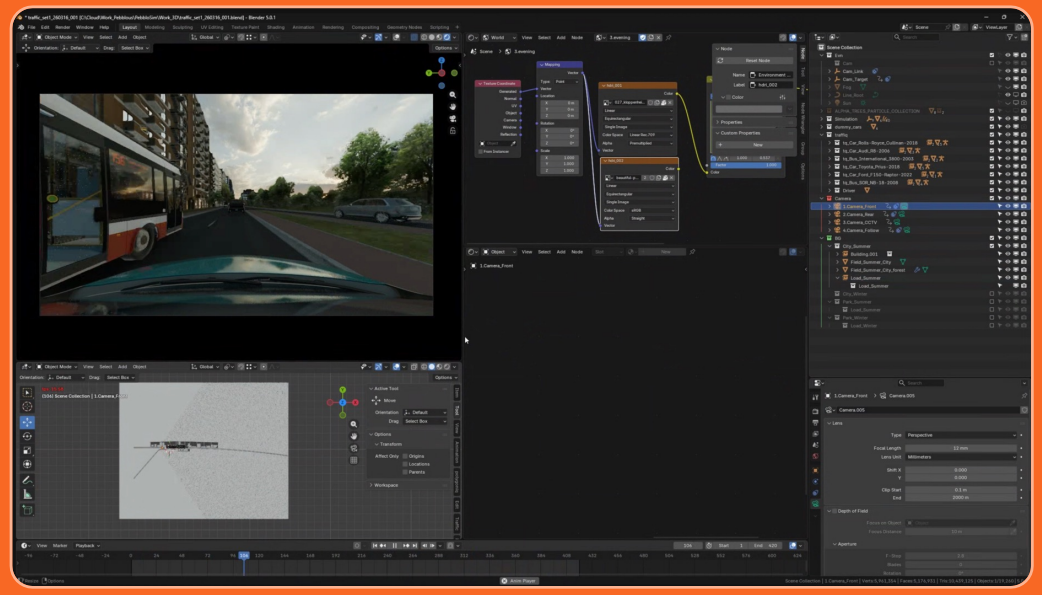

WikiArt
BIRDS 450 SPECIES

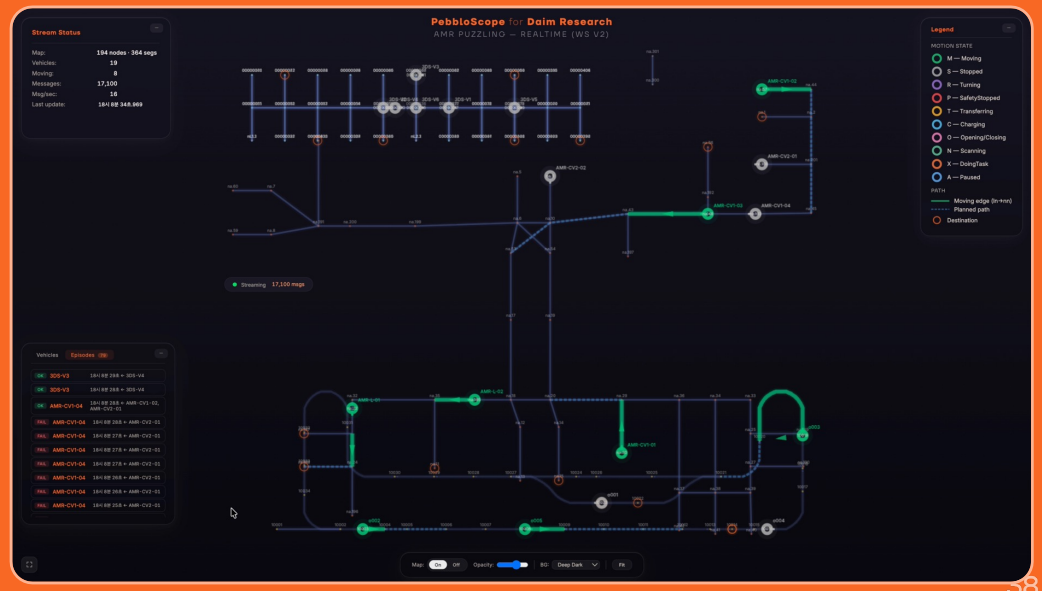

Pebblaus
ata Art Lab

PebbloSim

PebbloSim 기반 합성데이터 생성

① 자율 데이터 운영 체계 Data Greenhouse와 연동하여
② 디지털트윈 기반의 물리적 환각 없는 합성데이터를 생성하여
③ 피지컬 AI를 위한 AI-Ready Data를 지속적으로 생산하고 품질 증명





New Bottleneck

Pebblous

국내외 관련 분야 환경변화

국내 01

제조·국방·조선 등 주력 산업 중심

Physical AI 시장 급성장

- **디지털 챗봇 넘어, 물리 세계로 확장**
  삼성·LG(제조 지능화), HD현대(자율운항), 국방(무인체계) 등
- **고난도 멀티모달 데이터 수요 폭증**
  복잡한 변수(조명/마찰/날씨) 반영한 시각/센서 데이터 처리
- **산업 현장의 데이터 갭(Data Gap) 발생**
  희소한 결함/재난 데이터를 위한 고품질 합성데이터 기술 요구

국외 02

AI 성능 한계 돌파를 위한

데이터 품질, 합성데이터 수요 폭발

- **데이터 품질 시장의 폭발적 성장**
  Garbage In, Garbage Out 해결 (연평균 13%↑ 성장)
- **규제가 촉발한 합성데이터 전성기**
  개인정보 규제 회피 및 데이터 고갈 해결책 (연평균 46%↑)
- **설명 가능한 AI(XAI) 요구 강화**
  ISO/IEC 5259 국제 표준 기반, 투명한 데이터품질검증 의무화

연구개발 환경 03

외산 의존 탈피

소버린 AI 기반 모델로의 전략적 선회

- **베이스 모델의 전략적 변경**
  외산 오픈소스(Gemma) 대신, 국산 모델(KISTI KONI 등) 활용
- **한국형 AADS 구축을 통한 안보 강화**
  국방/공공 분야의 진입 장벽 해소 및 데이터 주권 확보
- **기술적 자립 및 내재화**
  외부 API 종속성 없는 독자적 생태계 구축 → 지속 가능한 구조

Data Bottleneck in Physical AI

Physical AI의 본질적 데이터 병목



Pebblous

01 Data Famine

희소 데이터(Edge Case)의 절대적 부족

사고, 재난 등 AI 안전에 필수적인 '위험 상황' 데이터는 현실 세계에서 수집 자체가 불가능하거나 비용이 막대함: **데이터 기근**(Data Famine)

02 Physical Hallucination

생성형 AI의 물리적 환각

기존 GenAI는 시각적으로만 그럴듯할 뿐, 중력·마찰 등 물리 법칙을 위배하여 정밀 제어 모델 학습에 치명적임

03 Data Complexity

멀티모달·3D·시계열 데이터 복잡성

이미지, 센서, 로그 등 이기종 데이터 간의 시공간 동기화(Sync) 및 인과관계 파악이 난해함

Pebblous

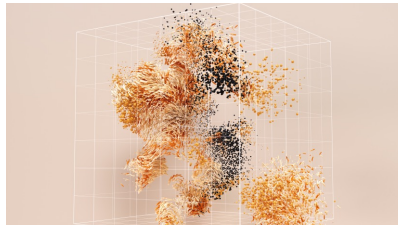
피지컬 AI 데이터 병목의 근본적 해결책: Data Greenhouse + PebbloSim

우리는 데이터를 수집하지 않고
'경작(Farming)'합니다.

① 물리 법칙을 복제한
PebbloSim과

② 자율 데이터 운영 체계
Data Greenhouse를 통해,

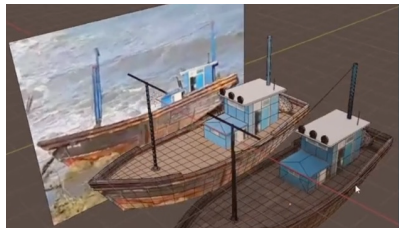
피지컬 AI를 위한 AI-Ready
데이터를 지속적으로 생산하고
품질을 증명합니다.



01 Agentic Data Synthesis

데이터 기근을 해결하는 '자율 데이터 벌크업'

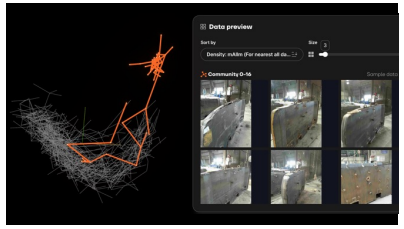
AADS 에이전트가 데이터 공백을 스스로 진단하고, 필요한 **엣지 케이스를 정밀 타격**(Targeting)하여 합성데이터 생성: On-Premise/Sovereign 환경 지원



02 Physically-Grounded

물리적 환각 없는 '뉴로-심볼릭' 합성 데이터

시뮬레이션의 **물리적 정합성** 위에 생성 AI의 **시각적 사실감**을 입혀,
물리적 환각을 구조적으로 원천 차단



03 Explainable Verification

인과관계를 규명하는 '설명 가능한 검증'

데이터의 **분포**와 비즈니스 **의미(Ontology)**를 실시간으로
상호 연결하여, **데이터 생성의 근거**를 투명하게 시각화하고 설명함

Pebblous

피지컬AI를 위한 초고품질 합성데이터 — Hyper-Synthetic Data for Physical AI

Gartner

Emerging Tech: Tech Innovators in Hyper-Synthetic Data for High-Fidelity Imaging

23 June 2025 - ID G00825904 - 26 min read

By: Nick Ingelbrecht, Vibha Chitkara, Tuong Nguyen, Kiumarse Zamanian, Ben Lee

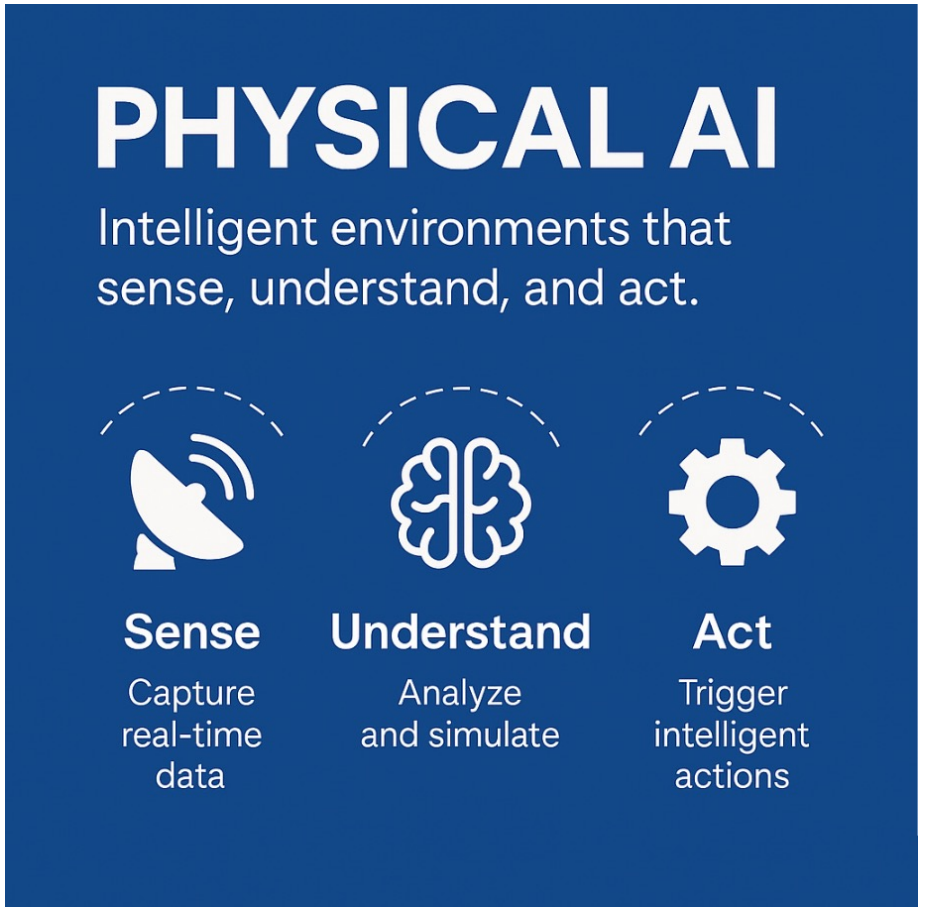
Initiatives:Emerging Technologies and Trends Impact on Products and Services; Increase Product Traction

> The surge in demand for computer vision training data is driving the adoption of hyper-synthetic data (HSD) solutions. Product leaders should capitalize on this emerging AI race opportunity by identifying and integrating high-fidelity spectral imaging innovations into their product roadmaps.

More on This Topic

This is part of an in-depth collection of research. See the collection:

- Emerging Tech Roundup: Navigating Emerging Disruptions and Challenges Ahead



Pebblous

Physical AI의 최전선: 산업별 난제를 해결하는 자율 데이터 운영 체계 Data OS

산업 도메인	주요 요구사항	구체적 페인 포인트	페블러스 솔루션
 조선	인력 절벽	AI 도입이 생존 문제로 직결 로봇 용접 · AI OCR(도면 분석) · 자율 운항	도면과 센서를 통합 이해하는 멀티모달 융합
	공기 단축 압박	수십만 개의 부품·도면이 디지털화 되지 않았음	
 국방	데이터 주권	데이터의 외부 반출이 불가능한 폐쇄망 환경 소버린 AI · 온프레미스 설치	데이터 주권(Sovereignty)을 보장하는 온프레미스(On-Prem)
	폐쇄망 환경	한 번의 오류가 치명적인 고위험 환경	
 모빌리티·제조	안전 필수	현실에서 1만 번 사고를 낼 수는 없는 '데이터 기근' 희소상황 데이터 · 안전규제 준수	물리적 환각(Hallucination) 없는 고정밀 합성 데이터
	물리적 정합성	기존 생성 AI는 물리 법칙을 무시하는 '환각' 발생	

Pebblous

Physical AI에서 페블러스가 갖는 결정적 우위



01

특허 기반 기술 해자

- 미국 특허 Data Imaging (US 12,481,720 B2)
- 단순 관측을 넘어 구조적 개선(Diet/Bulk-up)까지 보호받는 기술적 해자



02

단절 없는 End-to-End 데이터 OS

- 진단(Clinic) → 치료(Bulk-up) → 검증(Sim)의 유기적 통합
- 파편화된 도구가 아닌, AADS 기반의 완전 자율 운영 체계



03

공신력과 진입장벽

- KOLAS 기반의 국제 공인 시험성적서 발급 권한 (품질 보증)
- SaaS가 진입 불가한 '국방/제조 폐쇄망(On-Prem)' 구축 역량 (진입 장벽)

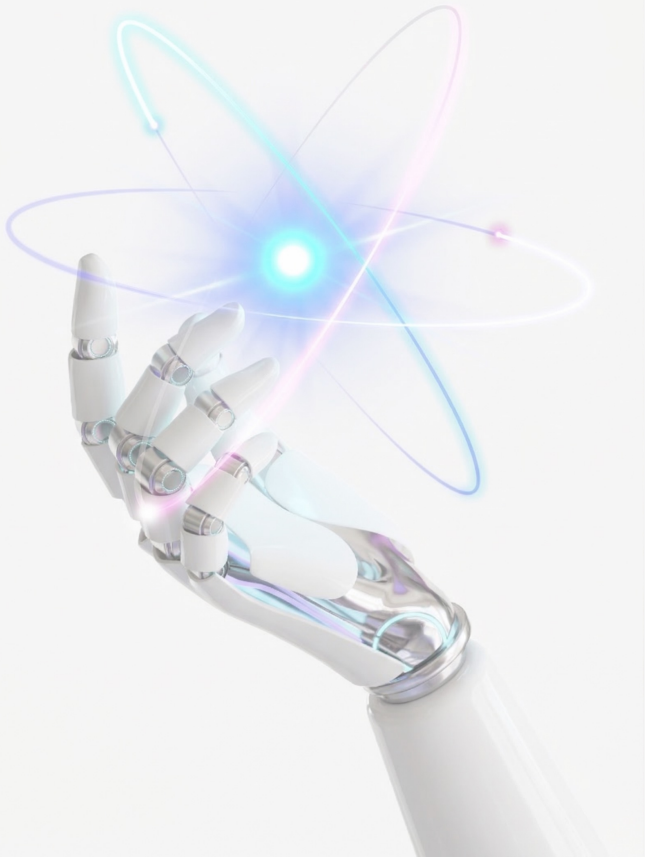


04

글로벌 표준 내재화

- ISO/IEC 5259 (데이터 품질)
- ISO/IEC 42001 (AI 거버넌스)
- 운영 로그 자체가 ISO 42001/5259 감사 증적이 되는 Embedded Governance

디지털 트윈 기반 피지컬 AI 학습 데이터 자율 생성 운영 체계



데이터 과제

01



희귀 사건

발생 빈도가 극히 낮아 현실에서
수집하기 매우 어려운
예외 상황(엣지 케이스) 데이터

02



위험 상황

현실에서 일부러 연출하기에는
너무 위험하고 막대한 비용이 드는
치명적인 사고 상황

03



환경적 공백

비 오는 날이나 야간처럼
특정 날씨/조명 조건에서 나타나는
구조적인 데이터 부족 현상

데이터 생성 프로세스

작업 → 장면-온톨로지 모델 → 시뮬레이션 → 영상, 텍스트, 물리 법칙

자동화된 데이터 루프



PebbloSim

Pebblous

KAIST 피지컬 AI 실증랩

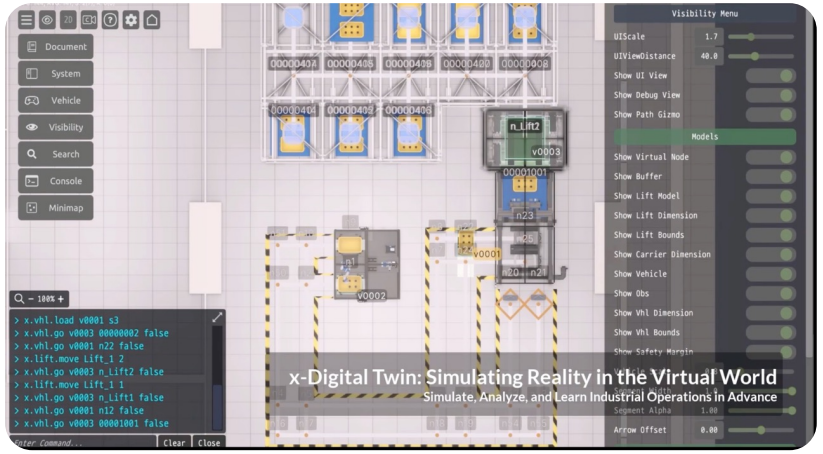
협력:

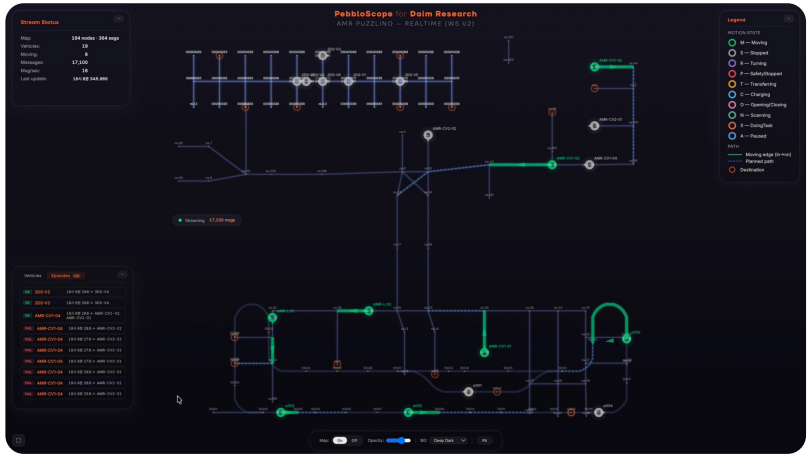
+ PebbloScope

+ PebbloSim

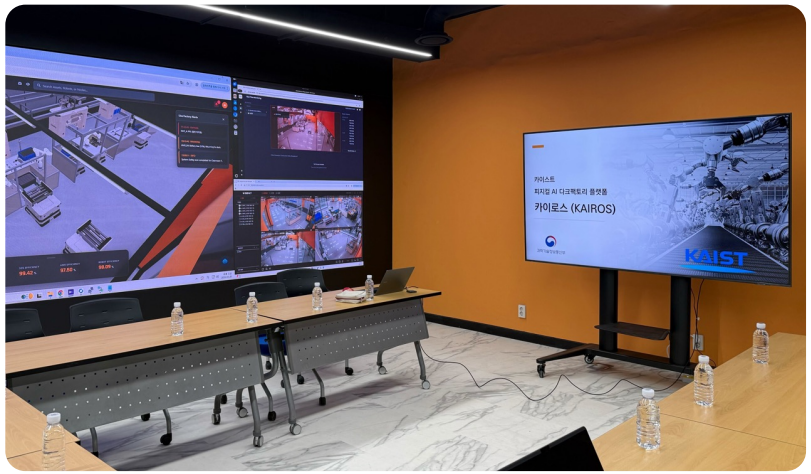
다크 팩토리 설계를 위한 제조 산업 디지털 트윈

디지털 트윈 시뮬레이션 기반 합성데이터 생성 및 품질 분석









Pebblous

데이터 그린하우스 SaaS — 데이터 품질관리에서 공급까지



AI 적합성을 만족하는 그린 데이터 (Green Data) 제공

01 데이터 품질 평가 및 최적화를 통해
AI 적합성 기준을 만족하는 **고효율 친환경 데이터 제공**

02 **의미 기반 데이터 연산**을 통한 그린 데이터 생성

- 대용량 데이터에 대한 전수조사 기반 자동 품질평가
- 정밀 타기팅 합성 데이터 생성
- 중복 및 편향을 제거하는 데이터 경량화
- 개인정보 보호를 통한 안전한 데이터 유통
- 의미 기반 탐색/추론/결합을 통한 데이터 생성



데이터 그린하우스 (Data Greenhouse) Agentic AI 솔루션

01 데이터의 AI 적합성을 지속적으로 평가하고 개선하는
차세대 거버넌스를 위한 **데이터 매니지먼트 솔루션**

02 **의미 기반 데이터 관리 체계**를 통한 효과적인 AI 산업 규제 대응

- 대용량 데이터에 대한 전수조사 기반 자동 품질평가
- 데이터 수명주기 관리를 통한 유연한 작업 대응
- 지속 가능한 AI를 위한 데이터의 안전성 및 윤리성 검증

피지컬 AI의 산업별 난제를 해결하는 자율 데이터 운영체계

Pebblous

현대차, 제조, 국방 등 주요 파트너사와의 성공적인
실증 경험을 바탕으로, 귀사와 함께 피지컬 AI 시대의
새로운 성공 사례를 만들어갈 준비가 되어 있습니다.



로봇 용접 · AI OCR(도면분석) · 자율 운항

조선·해양

인력 절벽 + 공기 단축 압박

도면과 센서를 통합 이해하는
멀티모달 융합으로
AI 도입의 생존 문제를 해결합니다.



소버린AI · 온프레미스 설치

국방·방산

데이터 주권 + 폐쇄망 환경

데이터 주권을 보장하는
온프레미스 솔루션으로
고위험 환경에 대응합니다.



회소상황 데이터 · 안전규제 준수

모빌리티·제조

안전 필수 + 물리적 정합성

물리적 환각 없는 고정밀
합성데이터로 1만 번의
사고 시나리오를 구축합니다.

Fabulous Data With

Pebblous

Better Data Makes Better AI

Pebblous.ai

Copyright ⓒ Pebblous.ai | All rights reserved | STRICTLY CONFIDENTIAL | ir@pebblous.ai

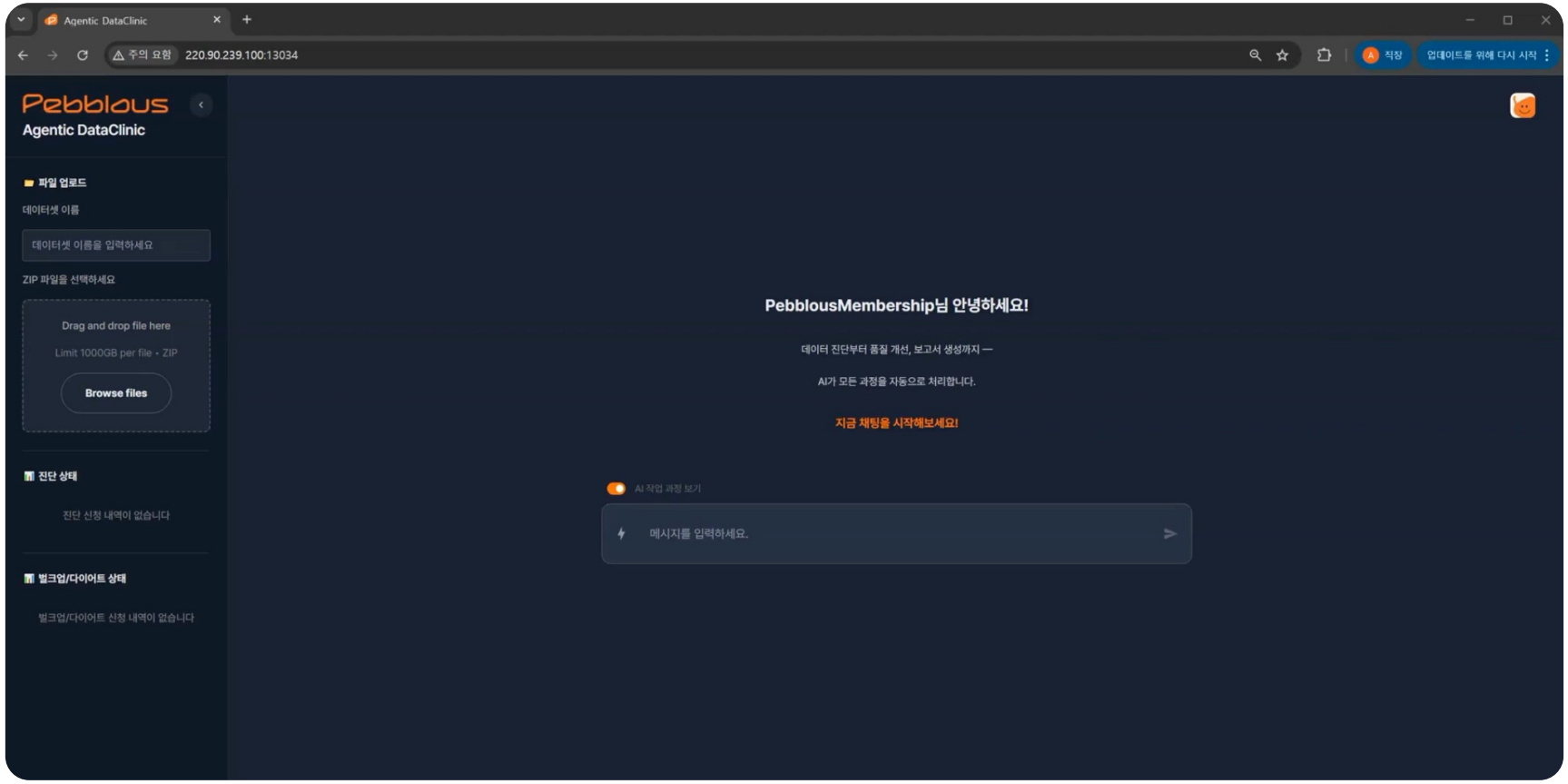
Pebblous

시연 - 진단

진단 API

" 방금 업로드한 데이터셋 진단해줘. "

기하 속성 관찰 > 국소 속성 관찰 > 밀도 측정 (II) > 데이터 등밀도선: 전체 데이터

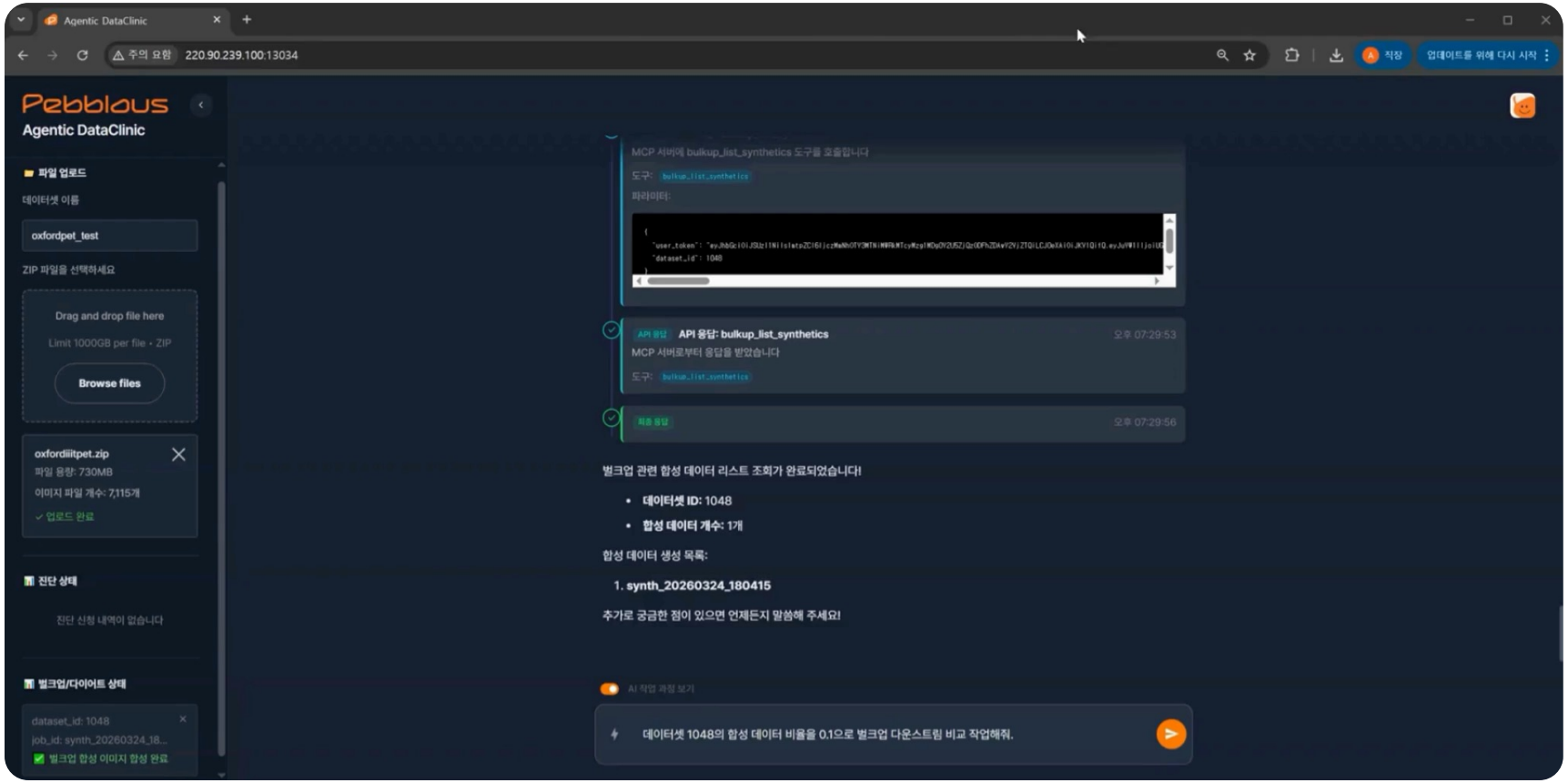


Pebblous

시연 – 데이터 벌크업

벌크업 API

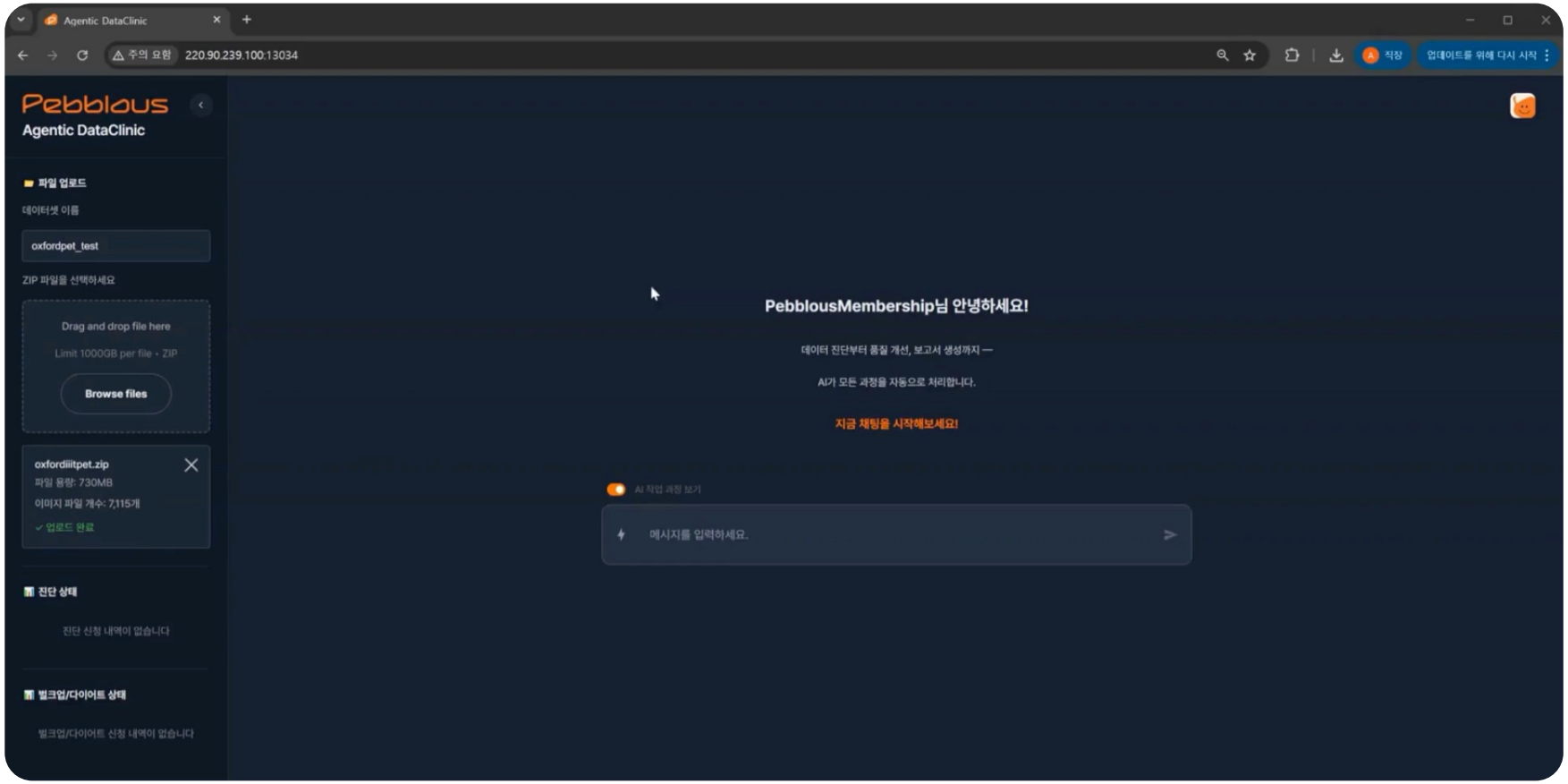
" 벌크업 다운스트림 성능 비교 시각화 차트 보여줘. "



시연

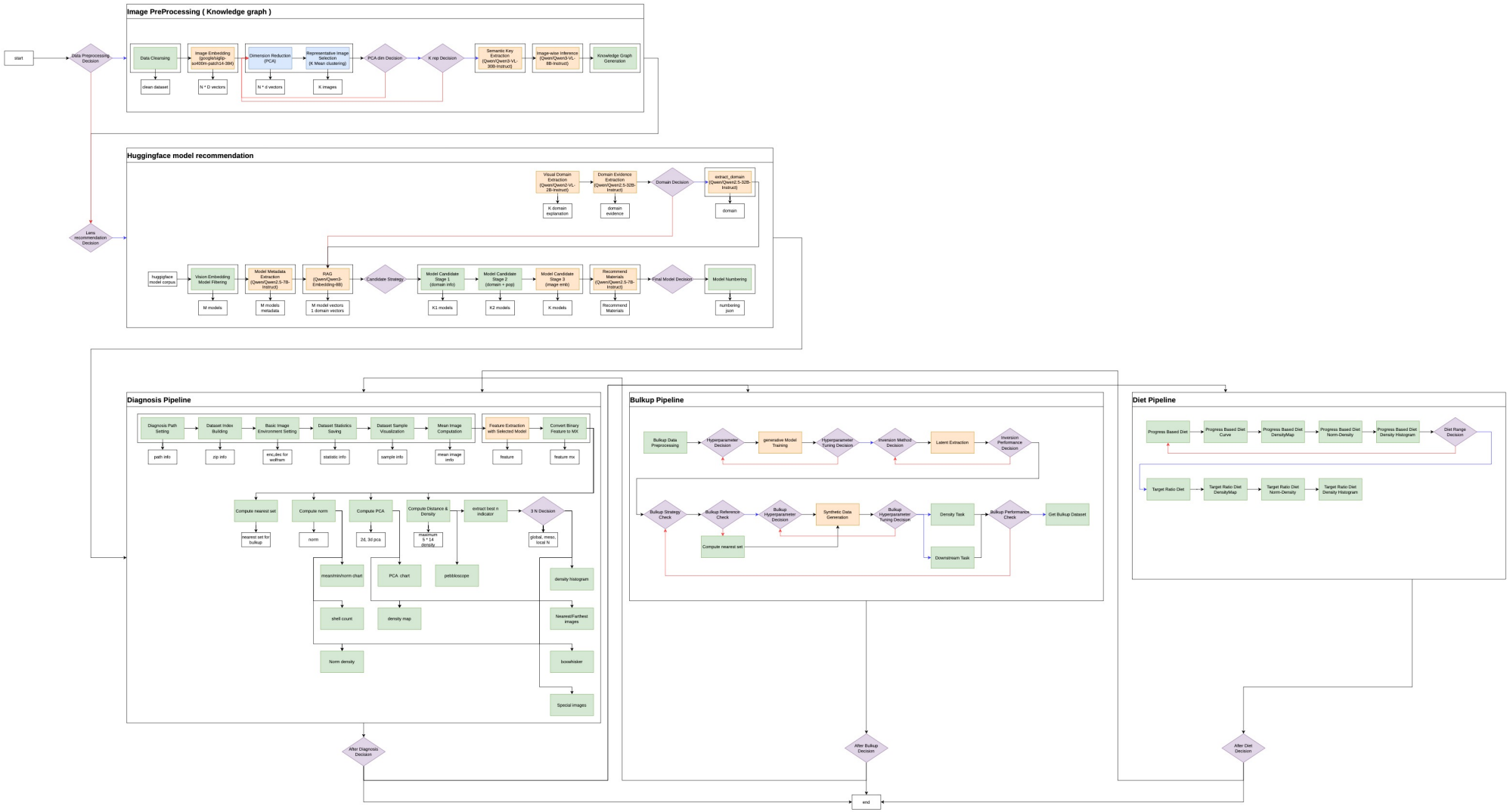
다이어트 API

" 데이터셋에 대해
다이어트 실행해줘. "



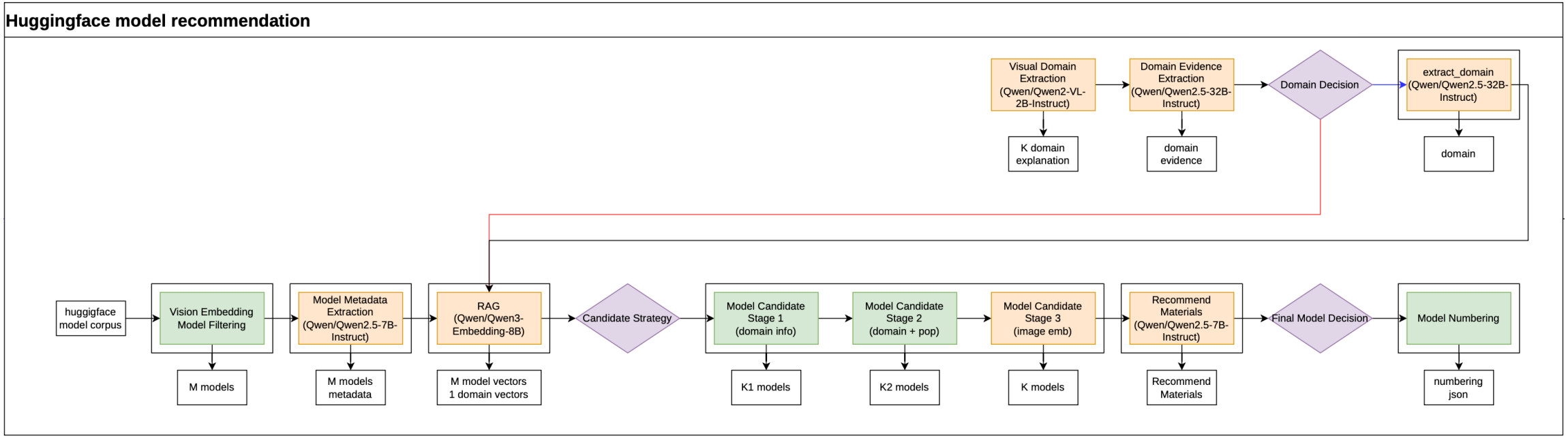
Pebblous

에이전트 워크플로우: 오버뷰



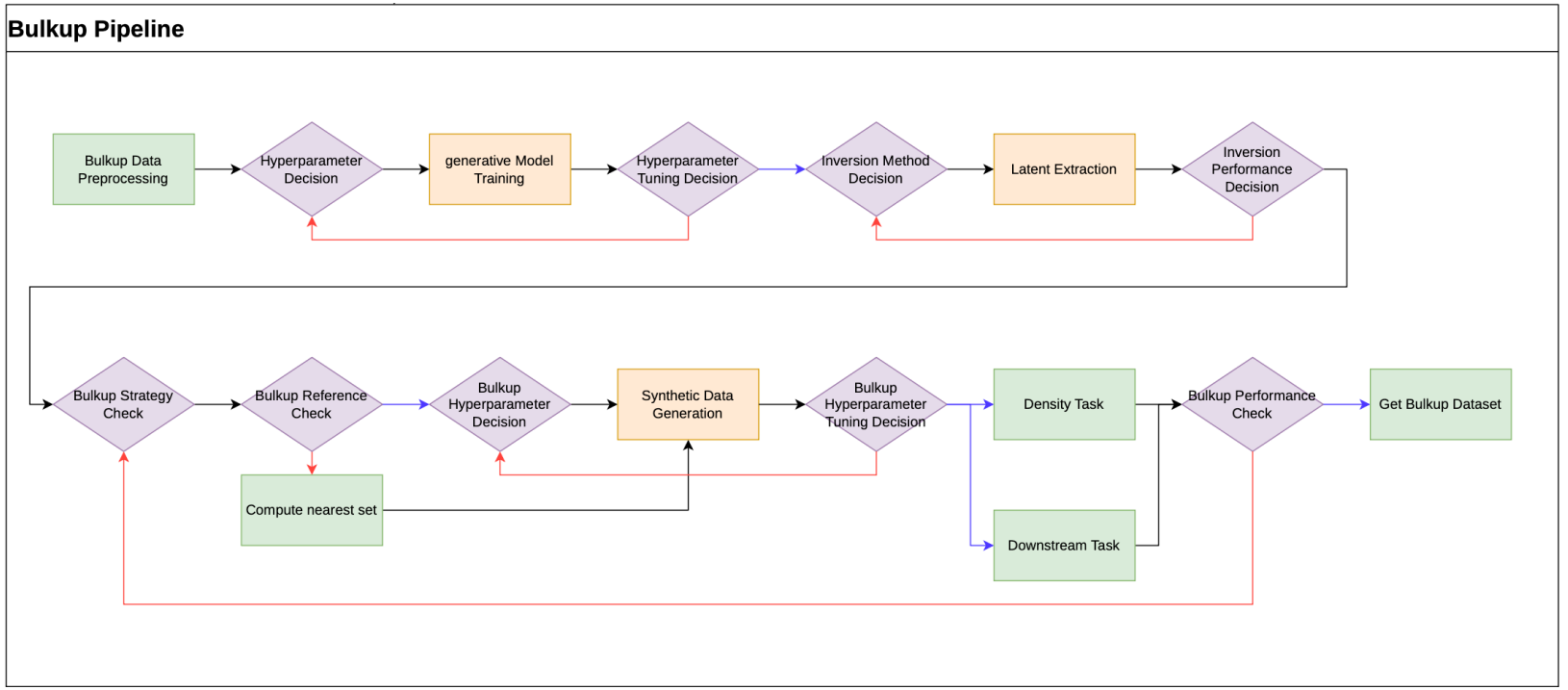
Agentic 렌즈 추천 파이프라인

- 도메인 이해
  - visual 정보에 근거하여 도메인 evidence 생성 후 최종 도메인 결정
  - 이 문제가 어떤 분야인지 정확히 규정
- 후보 모델 생성
  - 도메인 기준, 성능 및 대중성 고려하여 단계별 모델 필터링
  - HuggingFace 모델 풀에서 해당 도메인 적용에 유력한 모델들을 점점 좁혀나감
- 최종 추천 및 선택
  - 후보 모델 기반 에이전트 최종 판단
  - 후보 중에서 가장 적합한 모델 최종 선택



Agentic 데이터 벌크업 파이프라인

- 학습·표현 최적화 에이전트
  - 모델과 latent 표현을 가장 잘 만들도록 반복적 판단 및 조정
  - 좋은 생성 모델과 latent를 만들 때까지 계속 스스로 튜닝하는 에이전트
- 데이터 생성 전략 에이전트
  - 어떤 방식으로 데이터를 늘릴지 전략적으로 결정
  - 가장 효과적인 데이터 증식 전략을 선택하고 실행하는 에이전트
- 성능 검증·피드백 에이전트
  - 생성된 데이터가 실제로 유효한지 판단하고 다시 루프로 돌림
  - 데이터 품질을 검증하고 필요하면 전체 과정을 다시 돌리는 에이전트



Agentic 데이터 다이어트 파이프라인

- 단계별 데이터 다이어트
  - 다이어트 정도를 분위 기준으로 단계적으로 적용
  - 데이터 다이어트의 전반적인 흐름을 파악하기 위한 과정
- 다이어트 범위 결정 에이전트
  - 소수점 단위까지 최적화하기 위해 유망한 다이어트 범위를 결정
  - 단계별 데이터 다이어트를 점차 좁은 범위로 반복하며 최적 비율을 추정
- 목표 비율 데이터 다이어트
  - 결정된 기준에 따라 실제 데이터 비율을 조정
  - 추정된 최적 비율을 적용해 최종 데이터셋을 구성

