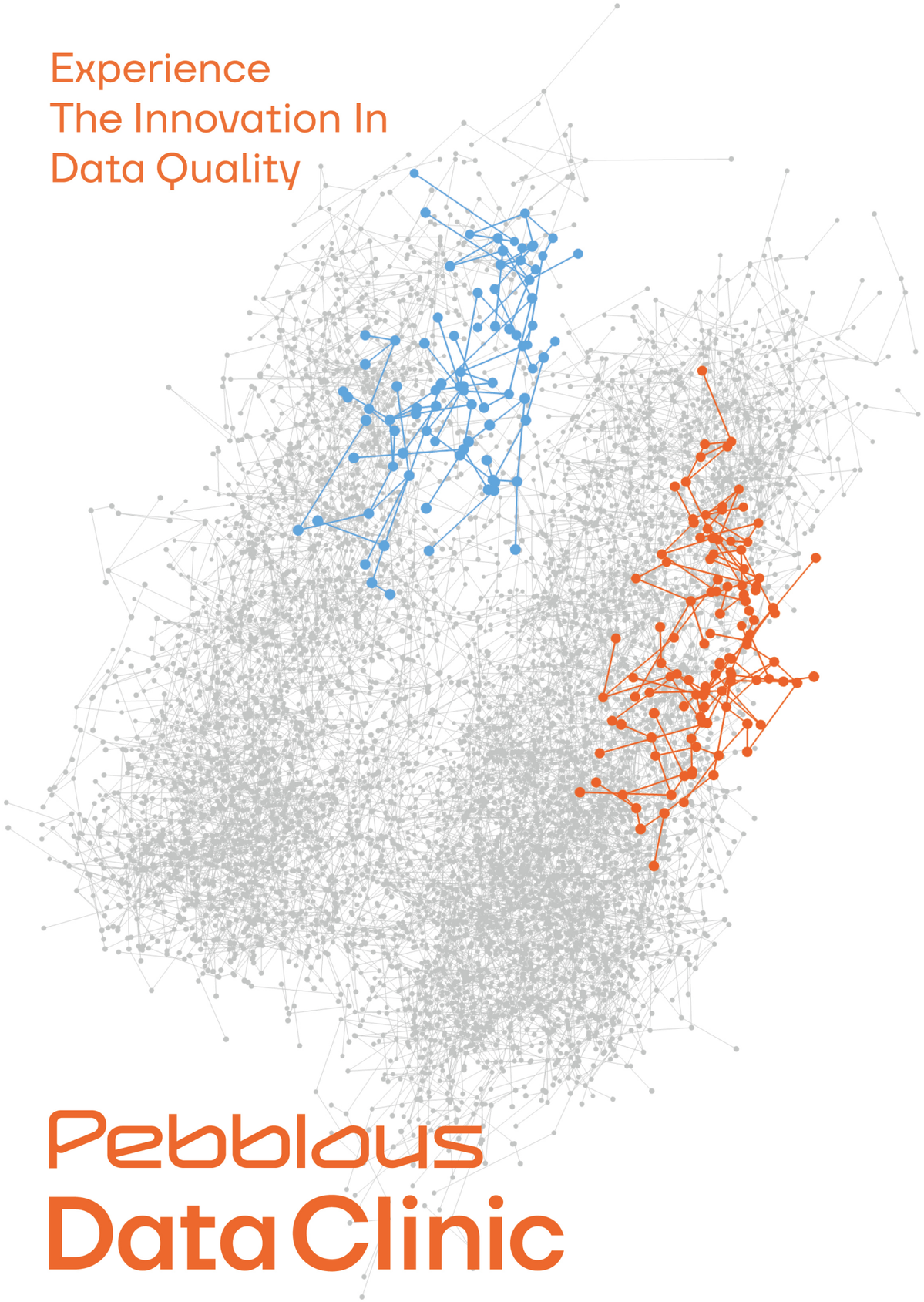
Pebblous Makes Data Tangible

Experience
The Innovation In
Data Quality



Pebblous Data Clinic

Pebblous

E-mail | contact@pebblous.ai
Tel | +82-44-589-3824
Web | pebblous.ai

Copyright © Pebblous Inc. All rights reserved.

Issue date | 2024.12.26

http://pebblous.ai

Pebblous

Power your innovation without wasting a watt.

We aim to contribute to sustainable artificial intelligence (AI) through **Green Data*** and **Data Greenhouse**.

*Green Data: High-efficiency, sustainable, and regulation-compliant data that meets AI readiness standards through advanced quality evaluation and optimization processes.

Current issues

01	02	03	04	05
Chronic data shortage	**Misguided** data collection	**Stagnant** AI performance	**Wasted** GPU resources	**Strengthened** AI regulations

Pebblous' solutions

Solution 01

Providing Green Data That Meets AI Readiness Standards

Green Data represents high-efficiency, eco-friendly, regulation-compliant data that is optimized for AI readiness. Through advanced data quality evaluation and optimization, we ensure your data aligns with cutting-edge standards while supporting sustainable AI practices.

- **Automated quality evaluation** for large-scale datasets.
- **Precise synthetic data generation** targeting critical gaps.
- **Data slimming** by eliminating duplicates and reducing biases.
- **Privacy-centric data flow** for secure management.
- **Semantics-based data generation** through exploration, inference, and intelligent combination.

Solution 02

Data Greenhouse SaaS Solution

A next-generation data management solution that continuously evaluates, optimizes, and enhances data for AI readiness.

Ensuring seamless compliance with evolving AI regulations through advanced, semantics-based data management.

Key Features:

- Flexible task handling powered by comprehensive data lifecycle management.
- Verified data safety and ethical integrity to support sustainable AI.

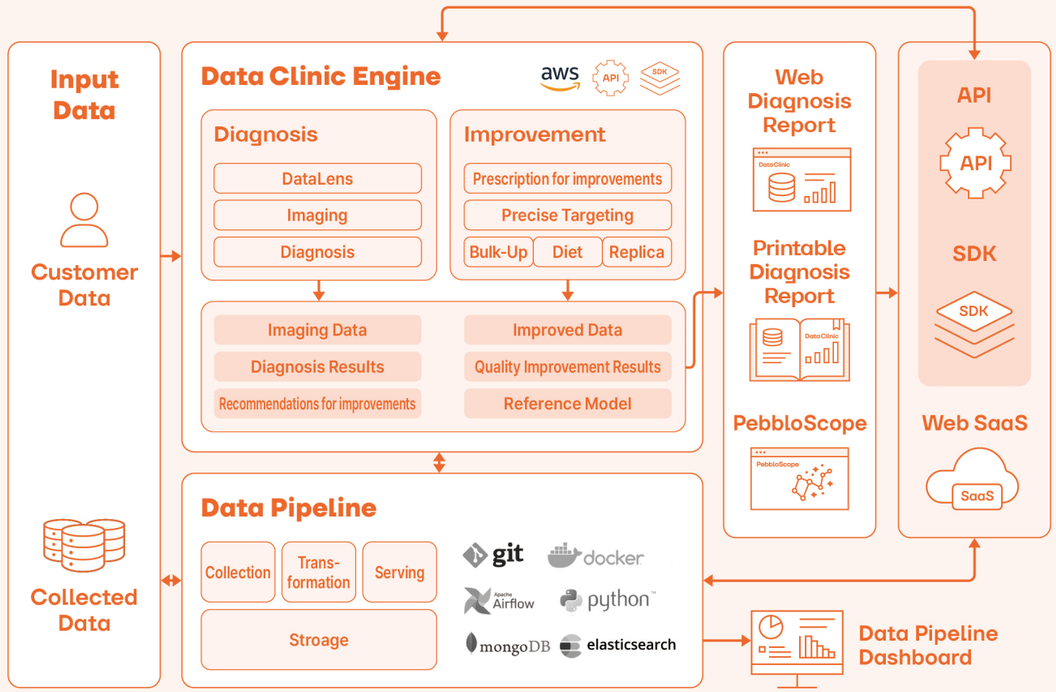
Pebblous Data Clinic

for AI-Ready Data Management SaaS

Pebblous Data Clinic is an all-in-one solution for evaluating and improving the quality of big data and AI training datasets.

01	Rapid Diagnosis	Quality evaluation of 100K image dataset **within 1 hour** Evaluation and visualization of the quality of AI training datasets
02	Improving AI Performances	5% synthetic data addition **2% improvement in model performance** Precise targeting synthetic data generation based on quality diagnosis
03	Improving Dev Efficiency	Reducing data volume by 80% **Achieving 5x GPU utilization** Optimizing training efficiency by reducing redundant data

Data Clinic SaaS Architecture



Pebblous DataClinic
AI Data Doctor

The Data Clinic, a comprehensive all-in-one solution for data, provides end-to-end services from data quality diagnosis to quality improvement.

Imaging — **Unlocking the Potential of Data**

- Our imaging technology enables deep dataset exploration.
- DataLens simplifies complex distributions, making datasets clear and actionable.

Diagnosis — **Precise Data Analysis**

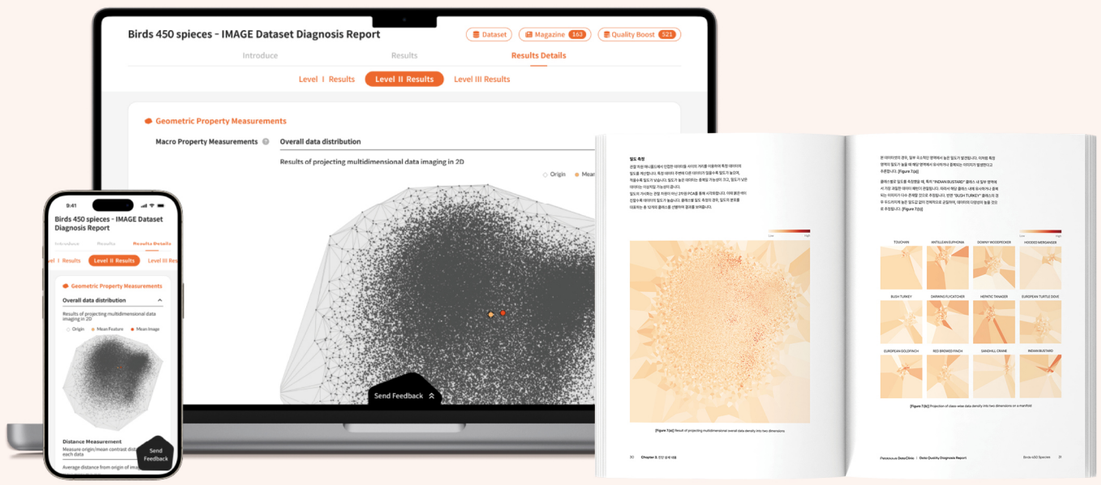
- Thorough evaluation of AI training datasets.
- Advanced neural networks reveal patterns and provide actionable insights.

Report — **Clear Evidence of Data Quality**

- Beyond charts, our reports explain AI dataset quality with clarity.
- Hidden patterns and improvement strategies are highlighted.

Improvement — **Optimized Quality Improvement**

- Tailored services include synthetic data generation (Data Bulk-Up) and compression (Data Diet).
- Solutions enhance AI performance with efficient dataset optimization.



Pebblous DataClinic
Key Distinctions

The Data Clinic leverages DataLens, powered by advanced deep learning neural networks, to accurately analyze the multidimensional characteristics of AI training datasets. *'DATALENS' is a U.S. trademark application.*

AI Training Data / Big Data



DataLens

Standard Type
Leverages pre-built deep learning models for rapid and efficient diagnosis.

Custom Type
Tailored deep learning models designed specifically for the customer's unique dataset.

Geometric and Statistical Properties

Analyzes multidimensional data spaces to reveal relationships and patterns with precision.

Data Visualization

Transforms complex datasets into intuitive visuals, enhancing understanding and decision-making.

Data Quality Diagnosis

Simplifies data complexity with tailored analytics and optimization strategies for superior results.

Pebblous Data Clinic

Key Features

The Data Clinic is a versatile, all-in-one data solution for artificial intelligence and big data analysis.

Universal Compatibility with All Data Types

Evaluates and improves text, images, time series, and tabular datasets.

Versatility Across Tasks and Industries

Delivers insights by analyzing data distribution in multidimensional spaces, unrestricted by task type or domain.

High-Quality Synthetic Data Generation

Creates targeted synthetic data using diagnostic results, AI models, and domain knowledge.

Specialized Services for Structured Domains

Offers quality diagnosis, replicable data generation, and pseudonymization for structured datasets.

Efficient AI Data Management

Optimizes AI training datasets with “Data Diet,” reducing volume while enhancing GPU and storage efficiency.

Advancing Explainable AI and AI Readiness

Supports ethics, fairness, and safety by prioritizing data evaluation to ensure explainable AI practices and AI readiness.

Pebblous Data Clinic

From Data Quality Diagnosis to Improvement

Customers can secure AI-ready data with safety, efficiency, and reliability through data treatment based on Data Clinic's precise data quality diagnosis.

01 Data Quality Diagnosis

Pebblous uses DataLens to analyze AI training data, assessing quality with scientific metrics and visual tools.

Task-Neutral Quality Diagnosis

- Evaluates datasets’ statistical and geometric characteristics across domains.
- Ensures consistency, integrity, and completeness.

Task-Specific Quality Diagnosis

- Tailored to specific AI tasks like classification or prediction.
- Focuses on separability, density, and outlier detection.

02 Data Quality Improvement

Enhances datasets for better AI performance using diagnosis-based solutions.

Empty Data Areas · Duplicate Data Areas · Original Data · Synthetic Data

Data Replica

- Replicates statistical characteristics while preserving data security.
- Prevents personal data leakage.

→ Prevents risks of personal data leakage.

Data Bulk-Up

- Creates targeted synthetic data to fill gaps in datasets.
- Reduces data collection costs and boosts performance.

→ Reduces costs for data collection and improves AI performance.leakage.

Data Diet

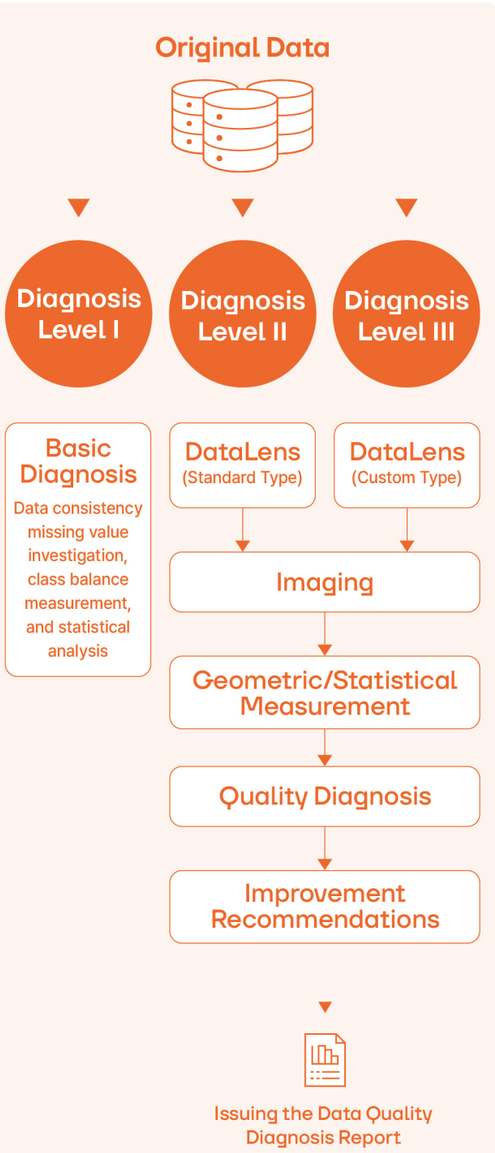
- Eliminates duplicates to minimize dataset size.
- Lowers costs while maintaining AI efficiency.

→ Reduces AI training costs and improves development efficiency

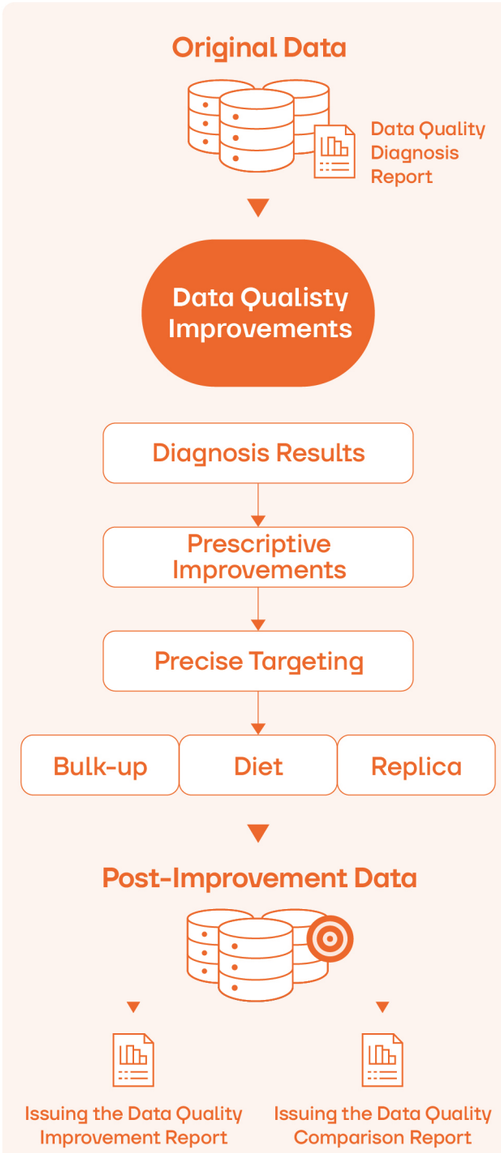
Level I, II, III Diagnosis Framework of **Data Clinic**

The Data Clinic's diagnostic framework systematically evaluates and improves data quality and structure, enabling effective data management tailored to customer needs.

01 Diagnosis



02 Improvement



DataLens and Imaging

Data Imaging

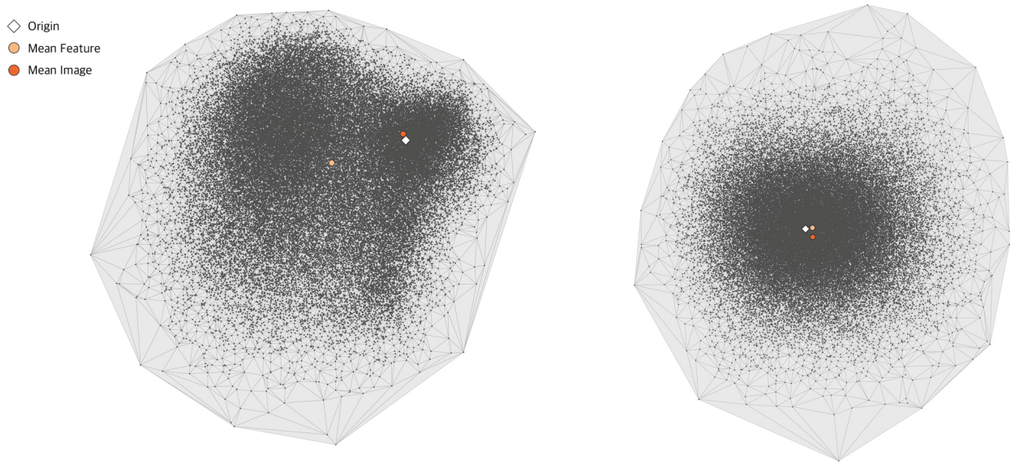
Pebblous' patented technology visualizes and interprets complex datasets by transforming them into embedding spaces, revealing patterns and supporting predictive analysis.

Intrinsic Dimension of Data

Determines the minimal feature vector dimensionality for optimal dataset representation, differentiating between standard and custom lenses.

Level II vs Level III Diagnosis

	Level II Diagnosis	Level III Diagnosis
DataLens	Uses pre-built neural networks for quick deployment.	Uses custom neural networks tailored for tasks.
Diagnosis Scope	Evaluates geometric and distributional properties.	Offers deeper insights and advanced metrics.
Problem Detection	Identifies biases in general embedding spaces.	Detects issues precisely for targeted datasets and tasks.



Measurement Metrics and Various Visualizations
for Effectively Explaining Diagnosis Results

The Pebblous Quality Diagnosis Report presents data quality findings through easy-to-understand diagnostic metrics and visualizations.

Measuring Distances Between Data

DataLens calculates distances between data points, systematically explaining their relationships. It identifies outliers and duplicates effectively, revealing the overall data structure through clear charts and examples.

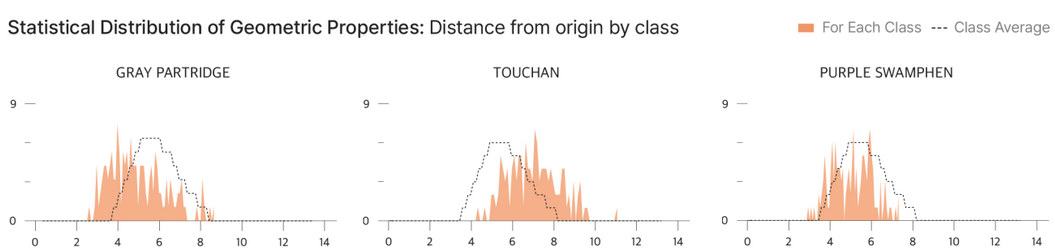
10 Nearest Data Points from a specific data point



10 Farthest Data Points from a specific data point



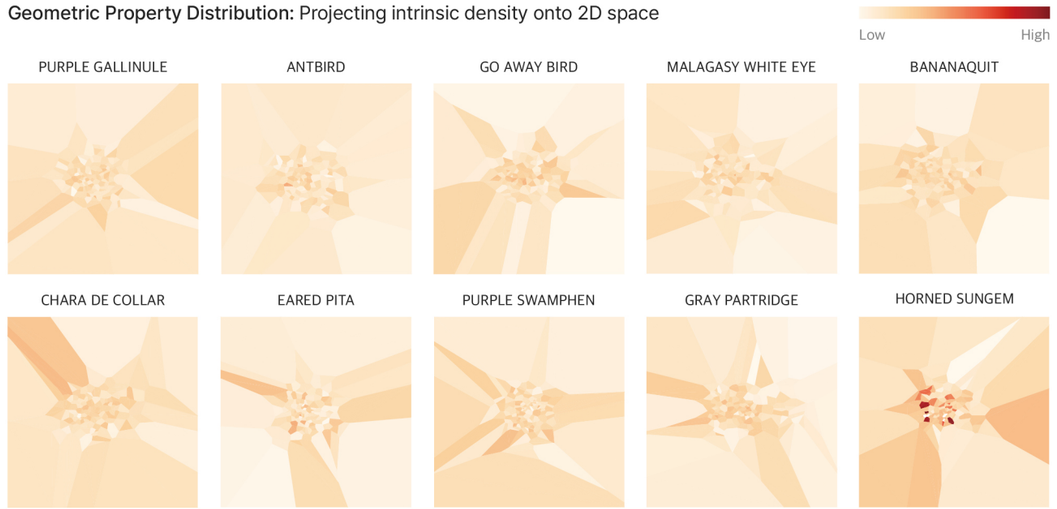
Statistical Distribution of Geometric Properties: Distance from origin by class



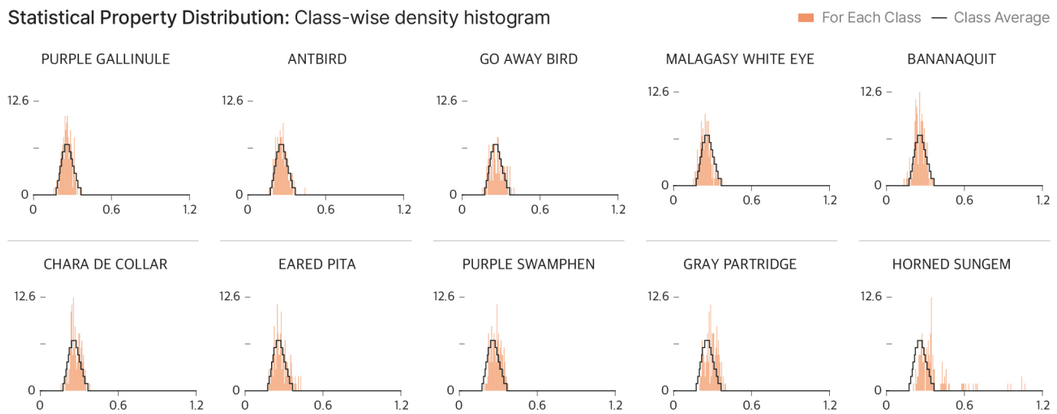
Measurement of Internal Data Density

Density measurement evaluates data concentrations in embedding spaces to reveal structural patterns and relationships.
The Data Quality Diagnosis Report uses density charts to identify high and low concentrations, helping detect critical regions, clusters, and outliers.

Geometric Property Distribution: Projecting intrinsic density onto 2D space



Statistical Property Distribution: Class-wise density histogram



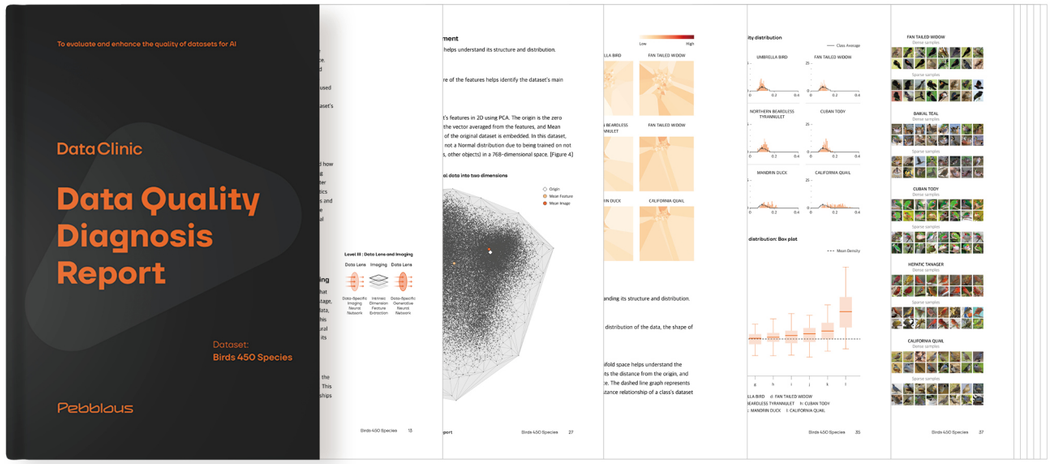
Comprehensive Data Quality Analysis Reports
for Intuitive Insights into AI Training Data

Pebblous reports provide clear analyses of data quality, helping customers understand their datasets through objective explanations and visuals.

Download Sample Diagnosis Report
https://bit.ly/DiagnosisReport

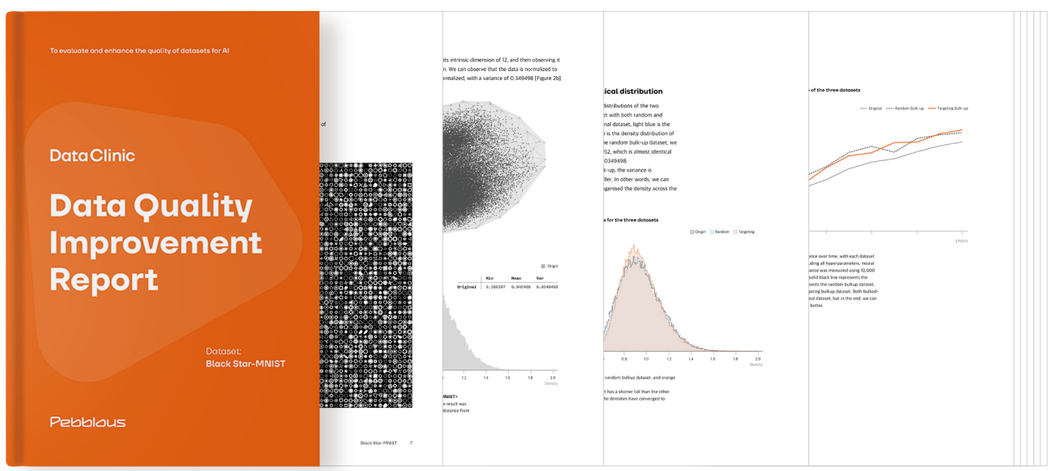
01 Data Quality Diagnosis Report

Comprehensive diagnosis results with improvement recommendations for AI training datasets.



02 Data Quality Improvement Report

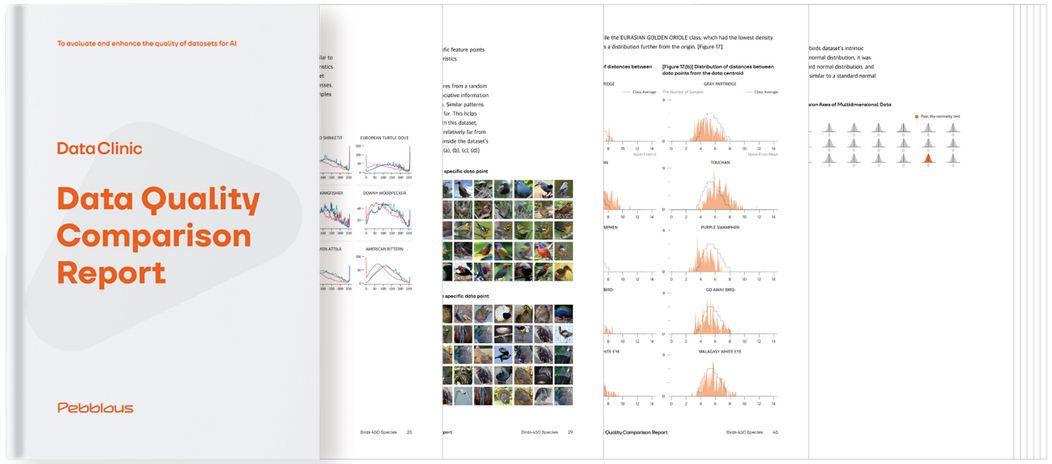
Explains the improvement process and outcomes based on diagnosis results.



03 Data Quality Comparison Report *(Upcoming)*

Provides detailed evaluations of data quality before and after improvements or between training and test datasets.

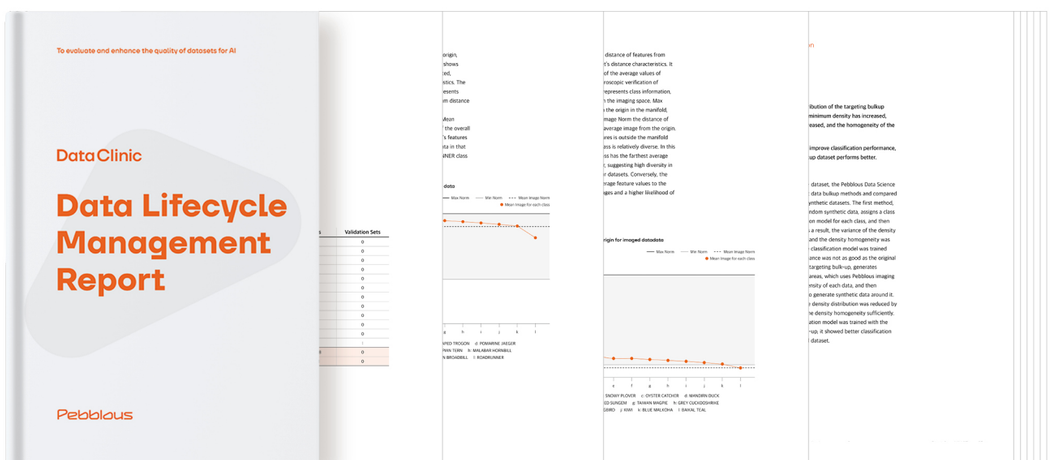
Sample



04 Data Lifecycle Management Report *(Upcoming)*

Offers analyses and recommendations for managing data throughout its lifecycle.

Sample



Precision-Targeted Synthetic Data

generation based on quality diagnosis

Delivers various types of synthetic data tailored to customer needs, based on quality diagnosis results.

Key Generation Techniques

Computer Graphics Technology

Advanced techniques like modeling, animation, and rendering to generate realistic visuals.

Generative AI Technology

Uses deep learning models to replicate patterns in real data.

Modeling & Simulation

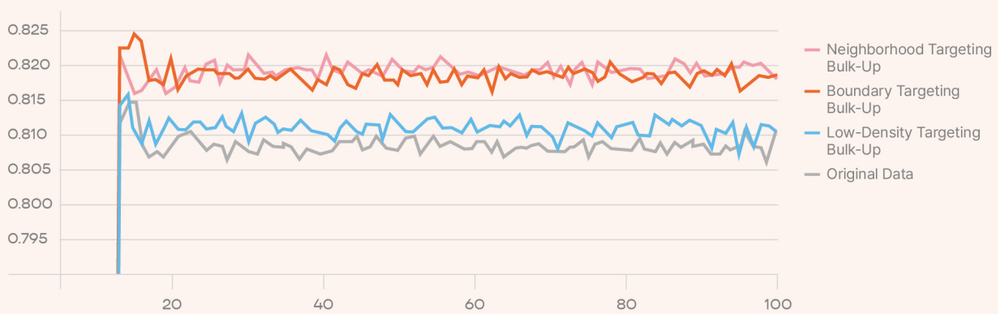
Simulates real-world environments with scenarios, mathematical equations, and physical models.

AI Synthetic Data Performance Improvement

Original Birds-450 Dataset vs. Precision-Targeted Synthetic Data



Improved AI Performance in Bird Classification with Synthetic Data



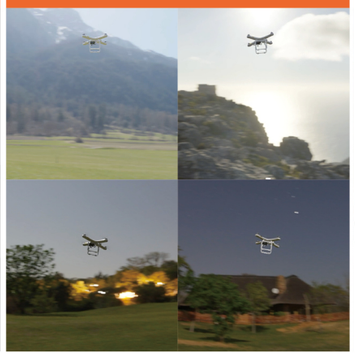
Single Image 3D Generation



Generates 3D models and synthetic data from a single photograph.

Joint research with POSTECH, achieving SOTA.

Drone Detection



3D synthetic data for drone recognition and tracking.

People / Poses in Specific Env



AI synthetic data for identifying individuals and poses in unique settings, such as vehicles.

Autonomous Driving



Synthetic data for indoor navigation robots using panoramic cameras and depth sensors.

Animal Behavior



AI data for mouse behavior classification and body part recognition.

Livestock Sectors



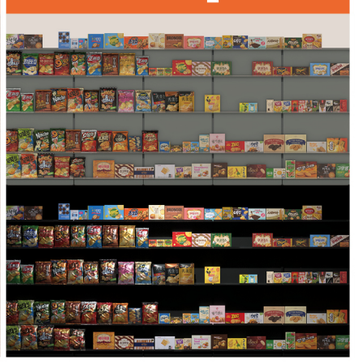
Synthetic images for analyzing cattle behavior with diffusion models.

Diet Monitoring



Monitors meal consumption using 3D models and generative techniques.

Retail & Logistics



3D data for inventory AI in unmanned stores.

Medical & Pharma



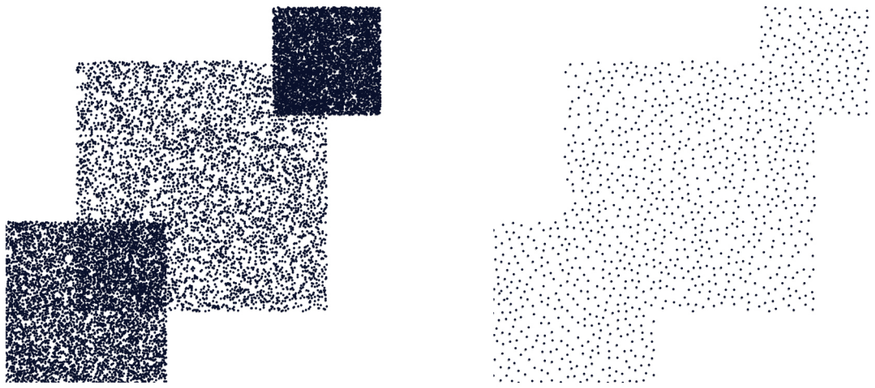
Synthetic data for vision inspection in pill dispensing and packaging systems.

Data Optimization Through Quality Diagnosis-Based Minimization **Data Diet**

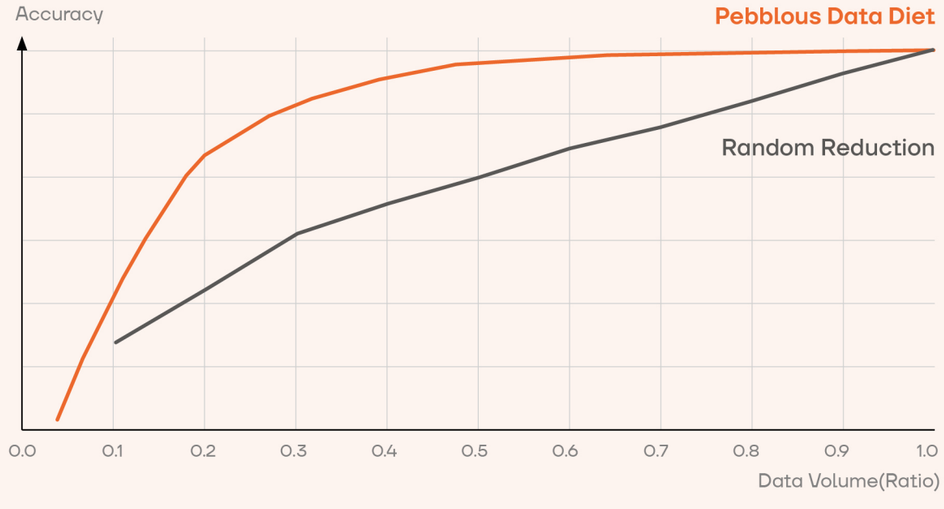
Reduces over-dense data points to maintain performance and optimize training time and GPU costs.

Concepts of Data Diet

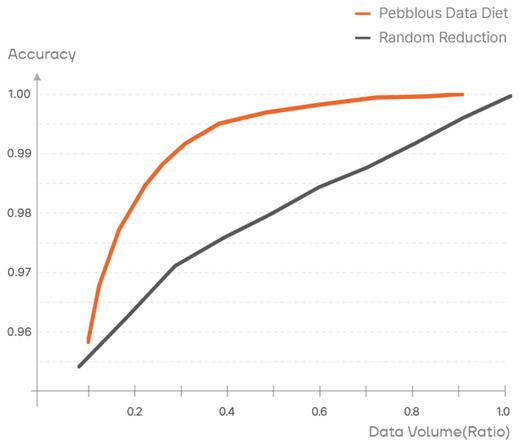
Before Data Diet vs. After Data Diet

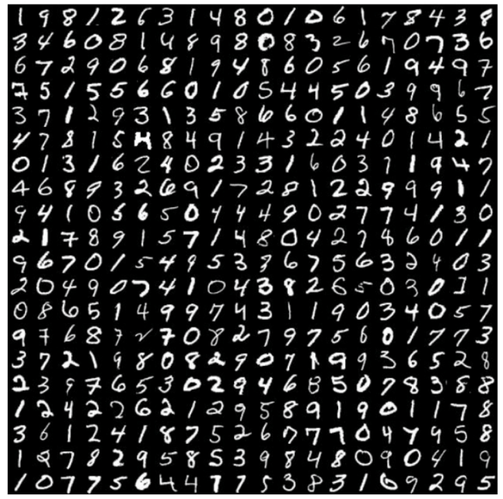


Data Diet Performance Curve

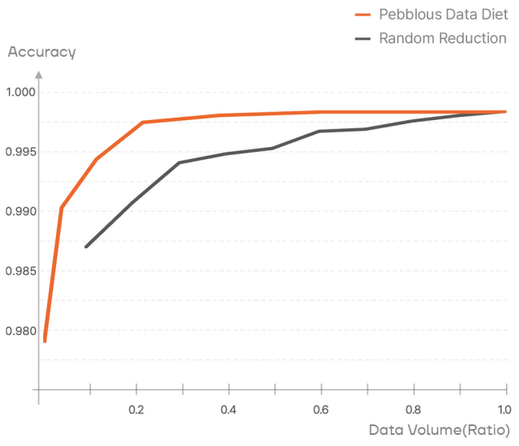


Case 01 MNIST Dataset



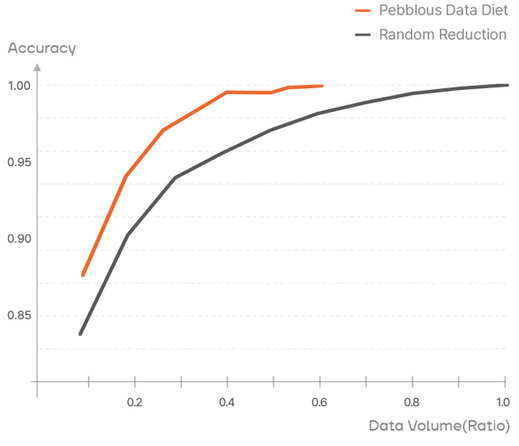


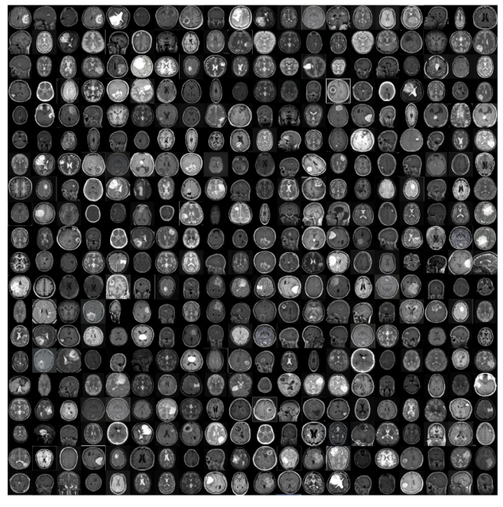
Case 02 Fire Detection Dataset





Case 03 Brain Tumor Dataset





PebbloScope

3D Interactive Data Communication Tool

PebbloScope visualizes data imaging and quality metrics from the Data Clinic, enabling interactive exploration.

A desktop version is available, with a web version launching in late 2024.

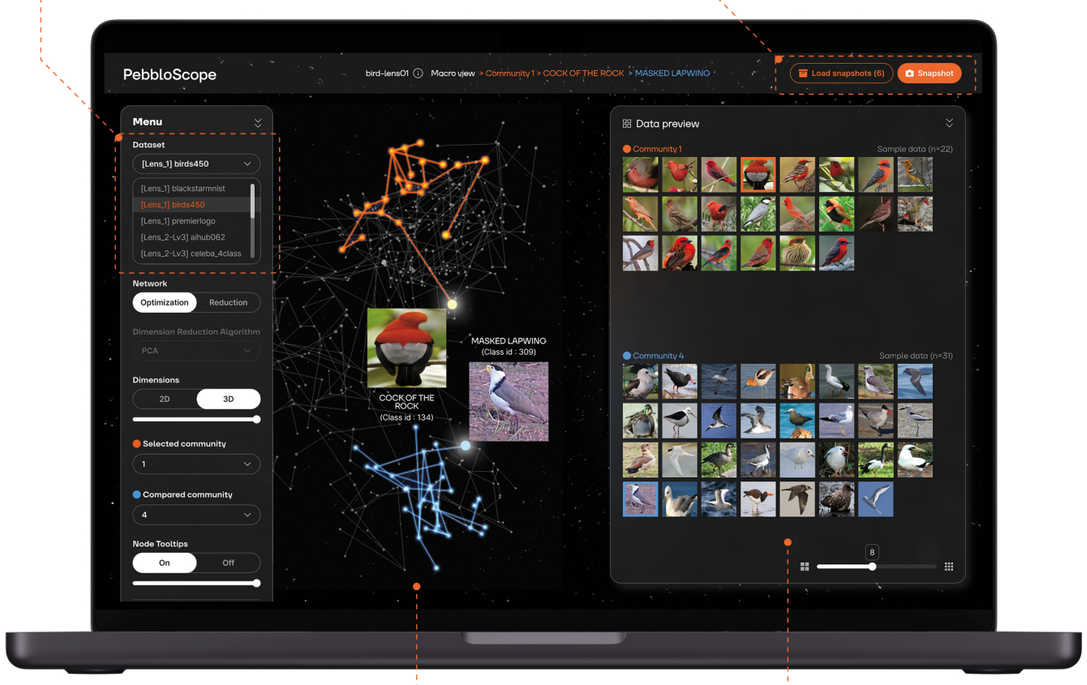
PebbloScope Introduction Video
https://bit.ly/pebbloscope

Dataset List

Select datasets from various domains for detailed exploration.

Snapshots

Capture 3D data points with annotations, shareable via unique URLs.



3D Visualization

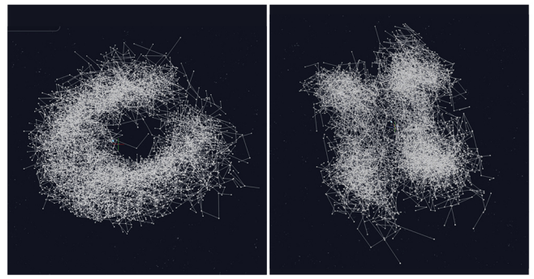
Explore optimized data dimensions in 3D, identifying patterns and distributions.

Data Preview

View original data and communities, compare nodes, and gain intuitive insights.

**Upcoming Features: Interactive synthetic data generation and data reduction.*

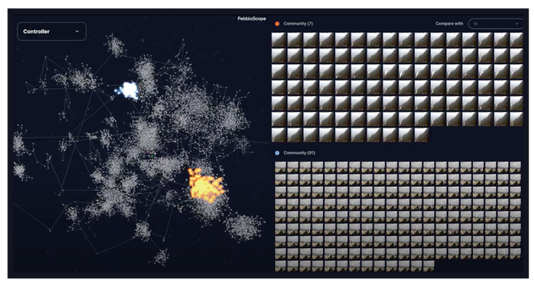
Features of PebbloScope



3D Visualization of Data in Optimal Dimensions

High-dimensional AI training datasets processed through Pebblous DataLens are visualized in a 3D space. Through 3D exploration, users can observe the unique and distinctive distributional characteristics of the data.

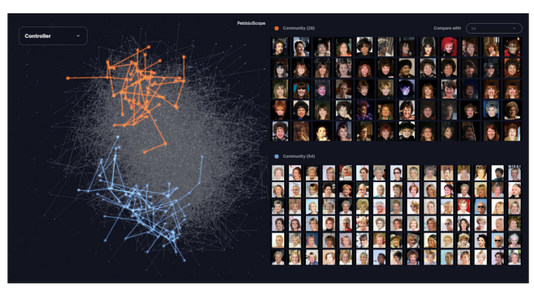
→ Example: Identify cluster shapes by adjusting the observation perspective.



Problem Identification via Data Community Exploration

Analyze data communities based on distances in optimal dimensions to detect similar data points or over-dense clusters.

→ Example: Spot over-dense clusters in a wildfire dataset.



Data Characteristics Through Comparison

Compare data communities to reveal cluster characteristics and gain deeper insights into unique distributions.

→ Example: Compare similar clusters in the CelebA_4class dataset.