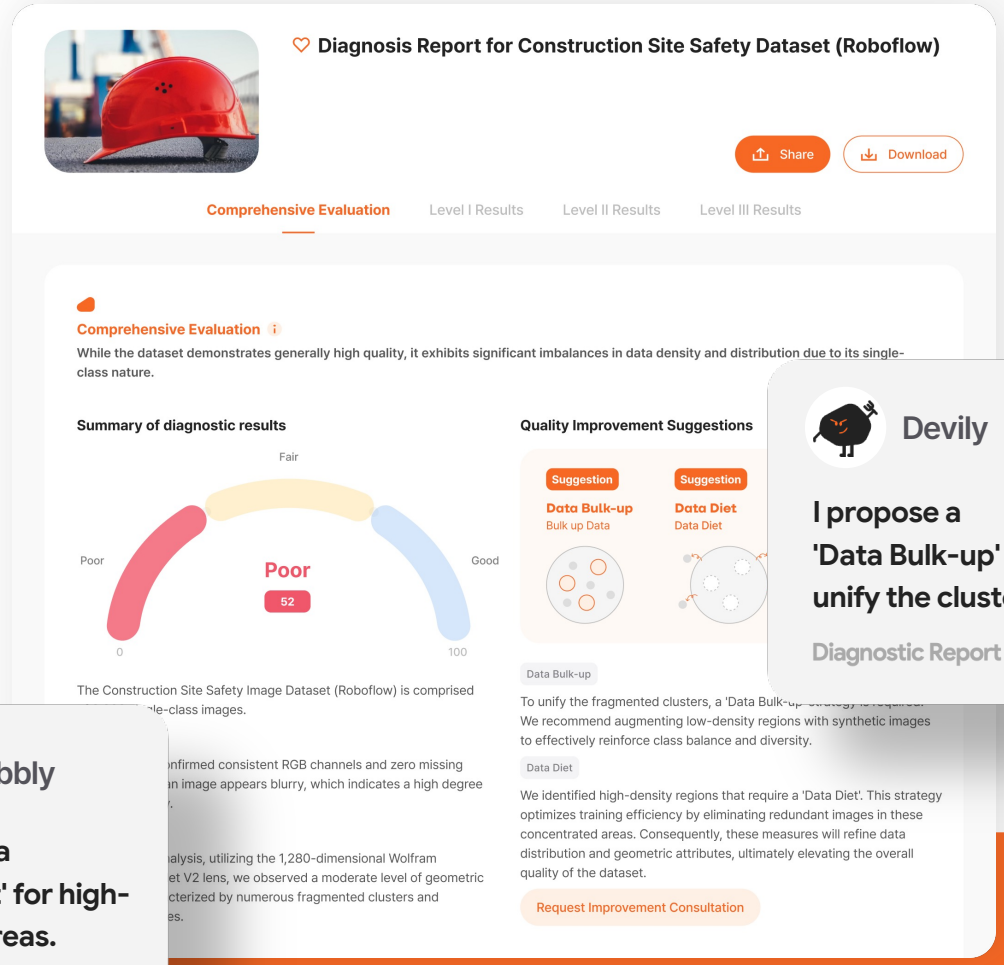
Turn Bad Data into AI-Ready Assets

The Ultimate Data Quality Management Guidebook

Boosting AI Performance by 200%



Pebblous

Why Your AI Hits a Ceiling: The Data Quality Gap

Can you mathematically prove your data is ready for production?

Move Beyond the Status Quo: A New Strategy is Required.

- **The Reality**: Model optimization has reached its limit. The real bottleneck isn't the code—it's the data. Without high-fidelity data, even the most advanced architectures deliver diminishing returns.

- **The Challenge**: Achieving 'AI-Ready' status requires more than just cleaning; it demands scientific validation. Yet, most enterprises are still tethered to legacy methods:

Relying solely on internal tribal knowledge for quality checks.

Rigid Rule-Based Silos: Inability to decode complex semantic relationships and cross-modal consistency in unstructured datasets.

Gartner

Source: Gartner

Pebblous

Breaking the 'PoC Trap': Engineering Data for Production-Grade AI

AI projects stall not because of the model, but because the data lacks the integrity required for real-world scaling.

01 The Multimodal Explosion

As AI moves into the physical world, the volume of video, sensor, and audio data is exploding. We provide the specialized infrastructure needed to manage this complexity, where traditional text-based tools fail.

02 Semantic-First Validation: Hidden Pattern Detection

Leveraging Vector Embeddings and Ontology, we move beyond rigid rules. Our semantic engine calculates deep-level similarities to pinpoint subtle errors, omissions, and 'data voids' that others miss.

03 Privacy-Preserving Utility: Synthetic & Replicated Data

We solve the privacy-utility trade-off. Our high-fidelity synthetic data and replicas serve as a robust de-identification layer, preserving the original data's 'DNA' while eliminating compliance risks (GDPR, EU AI Act).

Pebblous

Precision-Engineered Synthetic Data

Don't Just Generate—Validate.

Physical Fidelity & Domain Suitability

AI models must operate in the real world. Generating physically impossible scenarios is a waste of GPU resources. We ensure every data point adheres to rigid physical laws and domain-specific constraints.

Strategic Diversity: Conquering the Long-Tail

Synthetic data often inherits biases from the source. To mitigate this, we strategically augment rare edge cases—The Long-Tail—to ensure robust AI performance in unpredictable real-world environments.

The Gold Standard for AI Evaluation

Treat this as the "Final Exam" for your AI. Just as a student needs high-quality questions to be tested properly, AI requires rigorous, synthetic-based evaluation sets to verify and benchmark true performance improvements.

Success Story

Solving the 'Data Void' in Specialized Domains

Target Domain:

Autonomous Agricultural Robotics Company (Wildlife & Safety Object Detection)

How Pebblous eliminated model hallucinations through physically-accurate data synthesis.

Pebblous

| Challenge

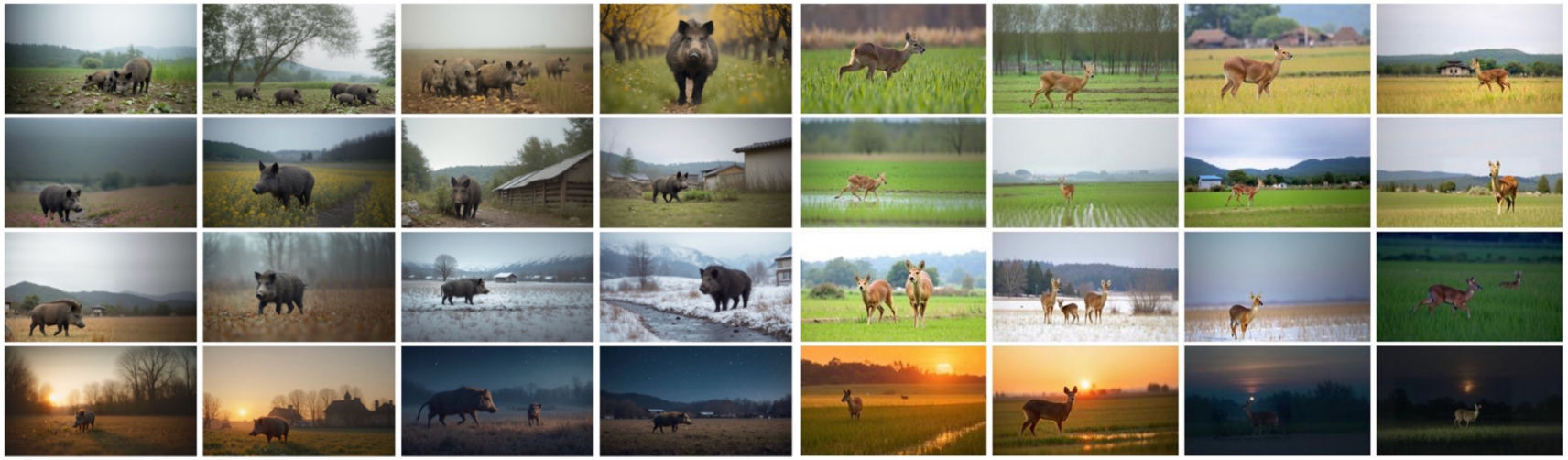
- **Extreme Data Scarcity**: Critical lack of high-fidelity imagery for indigenous wildlife (e.g., Korean Water Deer).
- **AI Hallucination**: Standard models produced unrealistic assets, failing to meet specialized field requirements.

| Solution

- **Semantic Imbalance Audit**: Pinpointed precise 'gaps' in species-specific datasets.
- **Hybrid Synthesis Pipeline**: Orchestrated a proprietary CG + GenAI pipeline to ensure biological accuracy and environmental consistency.

| Result

- **Production-Ready Assets**: Delivered 900+ high-fidelity synthetic images with 100% domain-alignment.
- **Zero-Hallucination Training**: Resolved critical detection errors by providing accurate training sets for rare species, enabling reliable field deployment.



Success Story

Conquering Data Scarcity in Wildfire Detection

Bridging the 'Sim-to-Real' gap with high-fidelity synthetic assets for mission-critical AI.

Pebblous

| Challenge

- **Quality Over Quantity**: Collected 4 million real-world images, but 90% were unusable due to poor quality.
- **Domain Gap**: Critical lack of night-time data. Standard day-to-night style transfer attempts resulted in low-fidelity images.

| Solution

- **Scenario-based Synthesis**: Generated sequence data simulating smoke behavior in night-time environments.
- **Model Optimization**: Separated the primary detection model from the secondary classification model.

| Result

- Doubled the effective dataset size.
- Achieved precision detection of fine smoke at a 9km distance.
- Model can now accurately distinguish between smoke and fog. Client requested expansion of quality improvement to their entire dataset.

(Original) Day Images

(Synthetic Data) Night Images

Success Story

Maximizing Efficiency in Industrial Safety AI

Eliminating redundancy and perception noise through Precision Pruning and Targeted Synthesis.

Pebblous

| Challenge

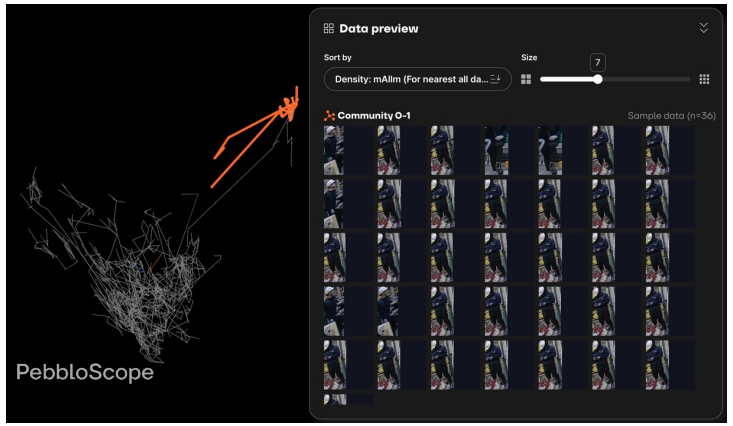
- **Redundancy Bottleneck**: Excessive duplicate frames from continuous CCTV feeds severely degraded training efficiency and inflated compute costs.
- **High False-Positive Rates**: Shadows, workwear, and cables were frequently misidentified as human threats, compromising system reliability.

| Solution

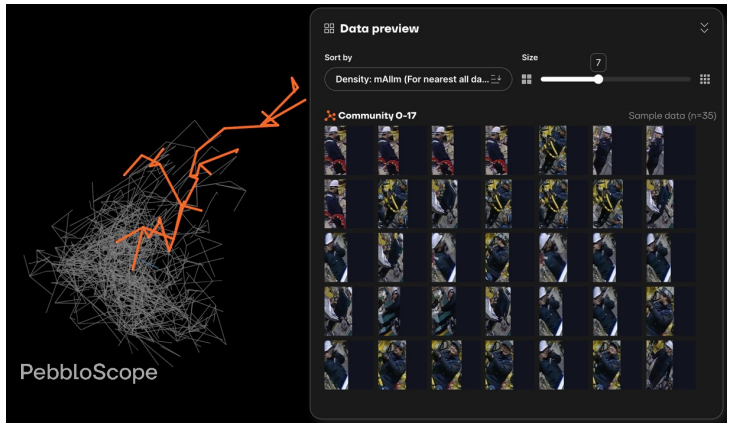
- **Semantic Data Diet:** Applied intelligent de-duplication to remove redundant frames, optimizing the dataset for high-impact learning.
- **Context-Aware Enrichment**: Generated synthetic data reflecting diverse lighting, gear, and clothing to build environmental resilience.
- **Precision Class Separation**: Implemented targeted labeling for high-risk distractors (robots, forklifts) to eliminate cross-class confusion.

| Result

- **Operational Trust**: Drastically cut false alarms, enabling reliable 24/7 autonomous monitoring.
- **Cost Efficiency**: Shorter training cycles and lower storage overhead with zero loss in accuracy.



(Original) Many Duplicates



(Data Diet) De-duplication

See PebbloScope in Action

Visualizing the Invisible: Interactive 3D Data Diagnostics

Pebblous

The Age of AI Accountability

Is Your Data Compliance-Ready?

Navigating the Shift from Technical Quality to Regulatory Integrity.

Beyond the Hype: The Hidden Costs of AI Risks

Cognitive bias, model hallucinations, and privacy infringements are no longer just technical glitches—they are systemic liabilities that threaten the core of AI adoption.

The Cost of Non-Compliance

AI is now strictly regulated. Under the EU AI Act, violations can cost up to €35M or 7% of Global Revenue.

ISO/IEC
AI Data Quality Standards

25024	5259	42119
Classic Data Measures	Standard for AI-Ready Data	AI Risk & Safety Control

Trust: Your Best Asset, Your Biggest Risk

Fines are temporary. Lost trust is permanent. Don't let data bias erode your brand.

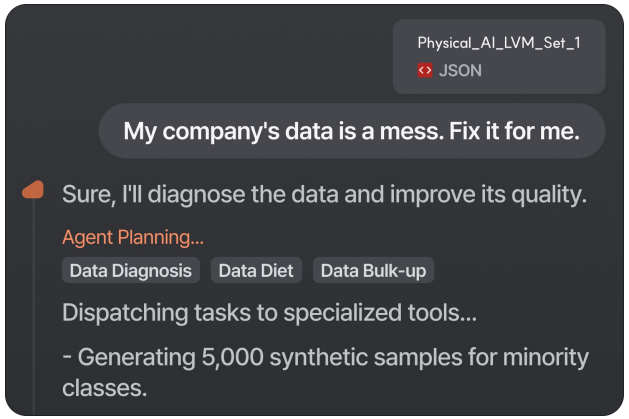
* The EU AI Act will be fully applicable by August 2026, marking a new era of strict AI regulation.

Meet
“Agentic Data Clinic”

The Neural Engine for Physical AI Data

The Bottleneck: Physical AI is paralyzed by the extreme cost and scarcity of real-world data.

The Cure: Autonomous AI Data Scientists that diagnose, synthesize, and optimize Physical AI assets 24/7.



▲ Click to Watch Agentic Data Clinic in Action

Pebblous

Key Advantages of Agentic Data Clinic

Self-Governing Pipelines

Our AI Agents autonomously orchestrate the entire data lifecycle—**Diagnosis, Synthesis, and Optimization**—eliminating human bottlenecks.

Audit-Ready Governance

Fully aligned with global mandates like the EU AI Act & GDPR. We transform complex compliance requirements into automated, high-trust reports.

One Prompt. Total Control

Simply issue a command, and AADS handles the rest. From deep-dive diagnosis to strategic improvement and professional reporting—instantly.

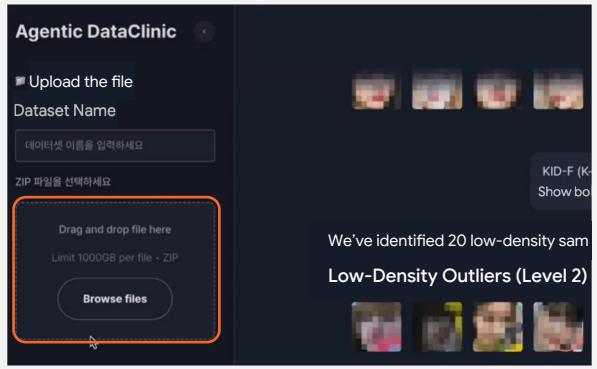
Pebblous

Agentic Data Clinic

Get Your Free AI Data Audit!

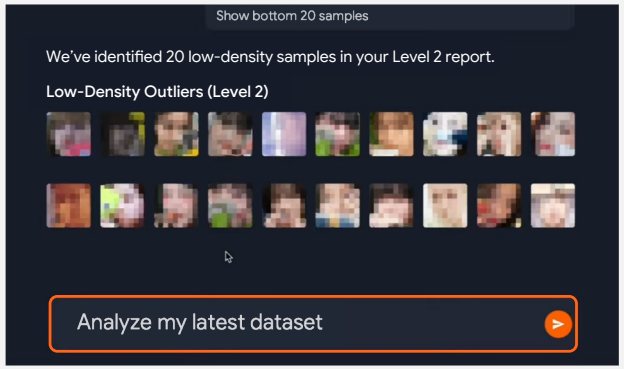
01 Upload Your Data:

Drag-and-drop your datasets for instant transformation.



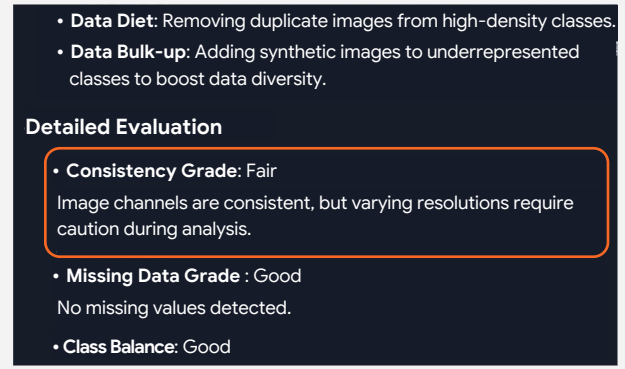
02 Issue a Command:

Command AADS to analyze and optimize Physical AI data.



03 Consult Your AI Expert:

Actionable insights and expert reports, delivered on-demand.



Contact Sales for Agentic Data Clinic

Pebblous

Data Clinic

The Secret to
Reducing Workload by 80%,
Boosting AI Performance by 200%.

Too good to be true?
We prove it's possible.

Contact Sales for Agentic Data Clinic

Pebblous

The Full Stack of Agentic Data Mastery

01 AI Ready Data

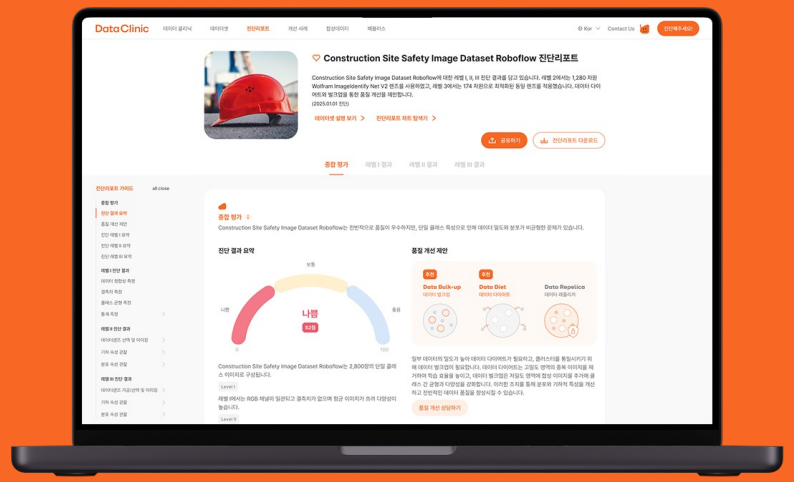
02 Observability Semantics Layer

03 Multi -modality

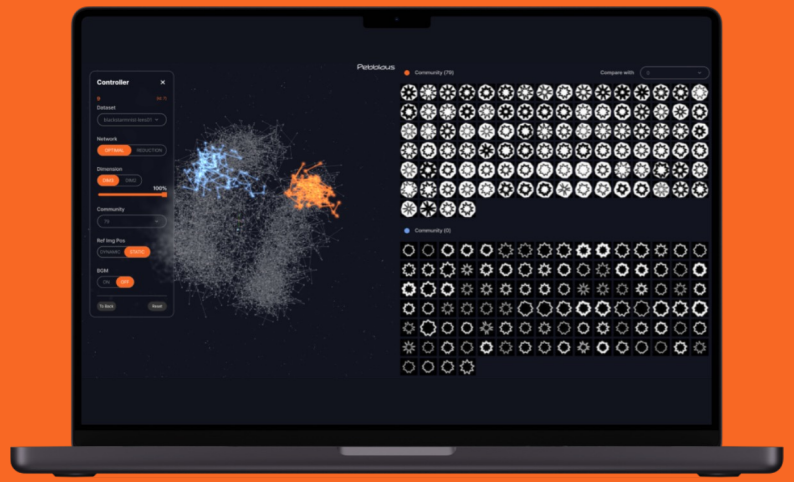
04 SaaS, On-Prem, API

Data Clinic, PebbloScope, and Synthetic Data:
The core pillars of your AI success.

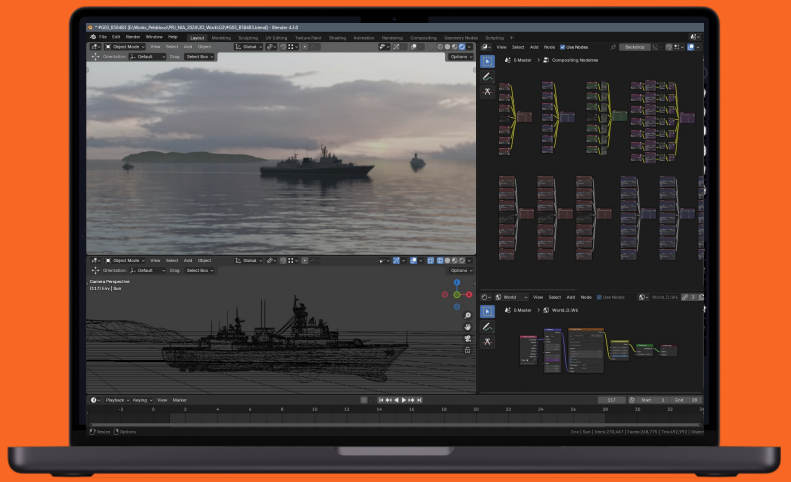
Data Clinic



PebbloScope



Synthetic Data



Pebblous

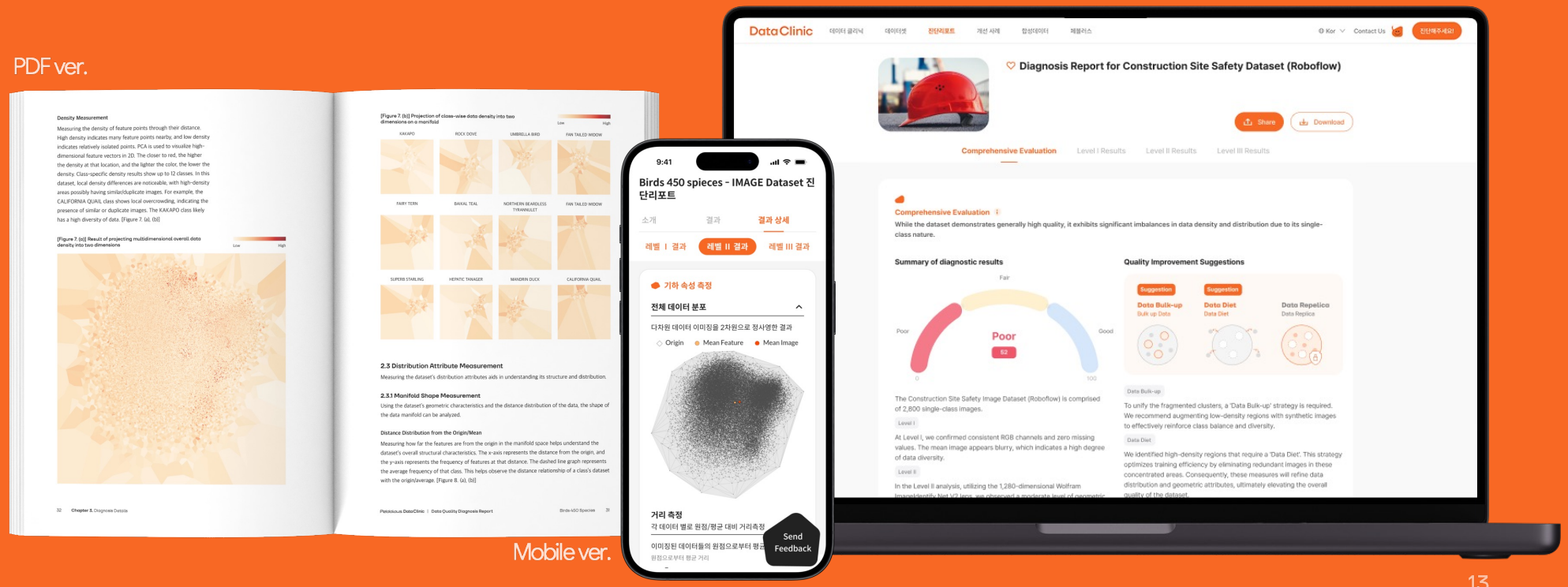
DataClinic

Pebblous Data Clinic is your All-in-One Data Care Center.

We offer comprehensive solutions for AI training data,
ranging from rigorous quality diagnostics to precise synthetic data generation.

Web ver.

PDF ver.



Mobile ver.

Pebblous

PebbloScope

Interactive 3D Data Communication Tool for visualization and sharing actionable insights

A data communication tool that transforms high-dimensional data into a three-dimensional space, allowing you to interactively explore different attributes and gain insights for data analysis.

Pebblous

Synthetic Data

Synthetic data is the strategic choice when:

① **Data Scarcity:**
Real-world samples are insufficient for robust training.

② **Accessibility Barriers:**
Data collection is restricted, dangerous, or impossible.

③ **Environmental Diversity:**
Testing requires diverse, edge-case scenarios beyond reality.

Fabulous Data With

Pebblous

Better Data Makes Better AI

Pebblous.ai