AIDRIN 2.0: A Framework to Assess Data Readiness for AI

Kaveen Hiniduma[1], Dylan Ryan[1], Suren Byna[1], Jean Luca Bez[2], Ravi Madduri[3]

[1]The Ohio State University, [2]Lawrence Berkeley National Laboratory, [3]Argonne National Laboratory

{hiniduma.1, ryan.1989, byna.1}@osu.edu, jlbez@lbl.gov, madduri@anl.gov

HIGHLIGHT

- **Solves the GIGO Problem (Garbage In, Garbage Out)**
  Unready data leads to poor AI outcomes. AIDRIN prevents this by assessing data across six critical pillars: quality, fairness, usability, structure, AI impact, and privacy.
- **Centralized & Decentralized, Privacy-Preserving by Design**
  Supports data readiness assessments in both centralized and decentralized settings, integrating seamlessly with the APPFL[3] framework for secure, privacy-aware evaluations across edge and sensitive environments.
- **Saves Time, Compute & Rework**
  Detects data issues early in the ML pipeline, optimizing efficiency and accelerating trustworthy AI development.

PILLARS OF DATA READINESS FOR AI

We comprehensively explored key metrics and dimensions from the literature to define data readiness for AI[1,2]
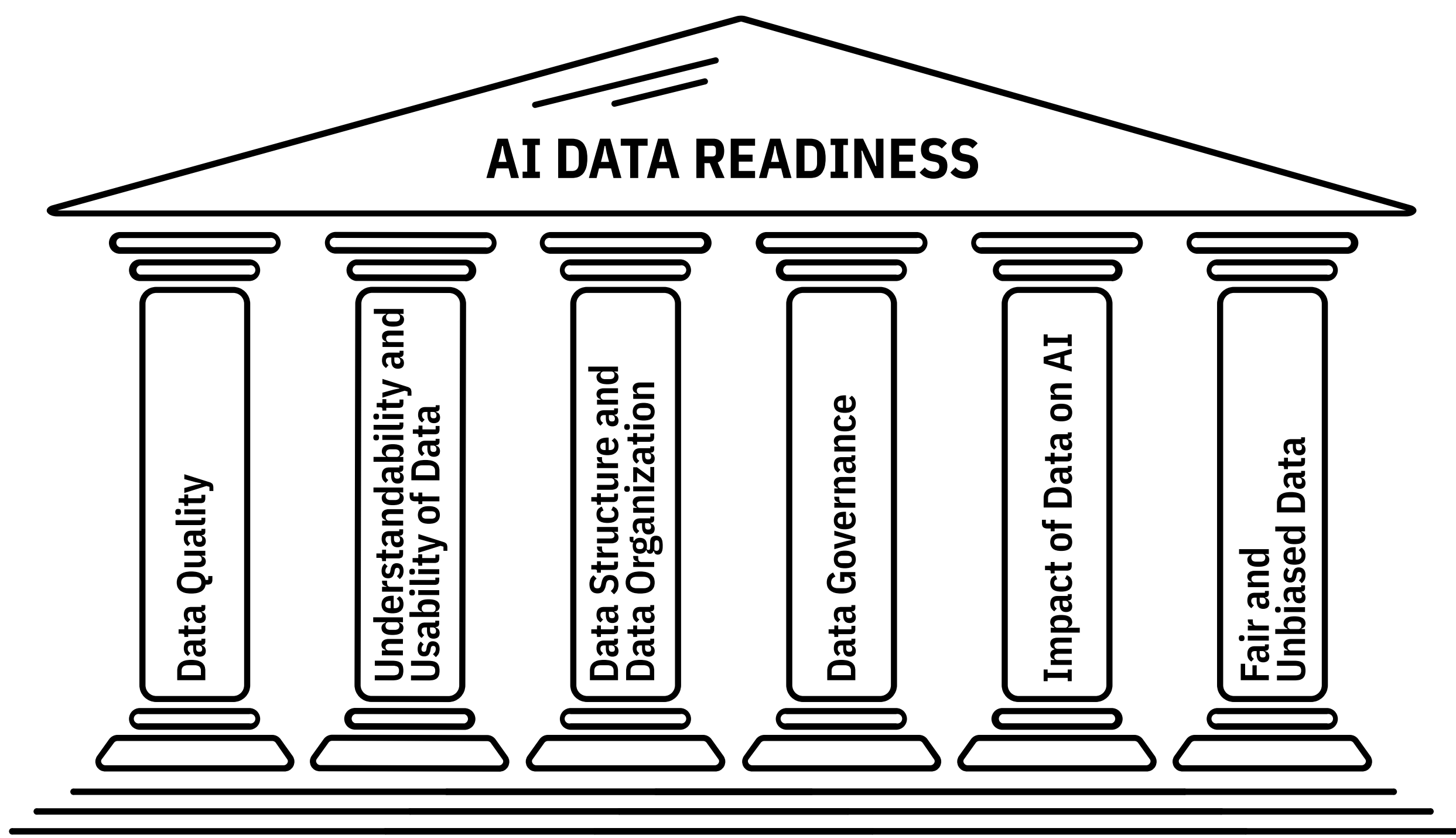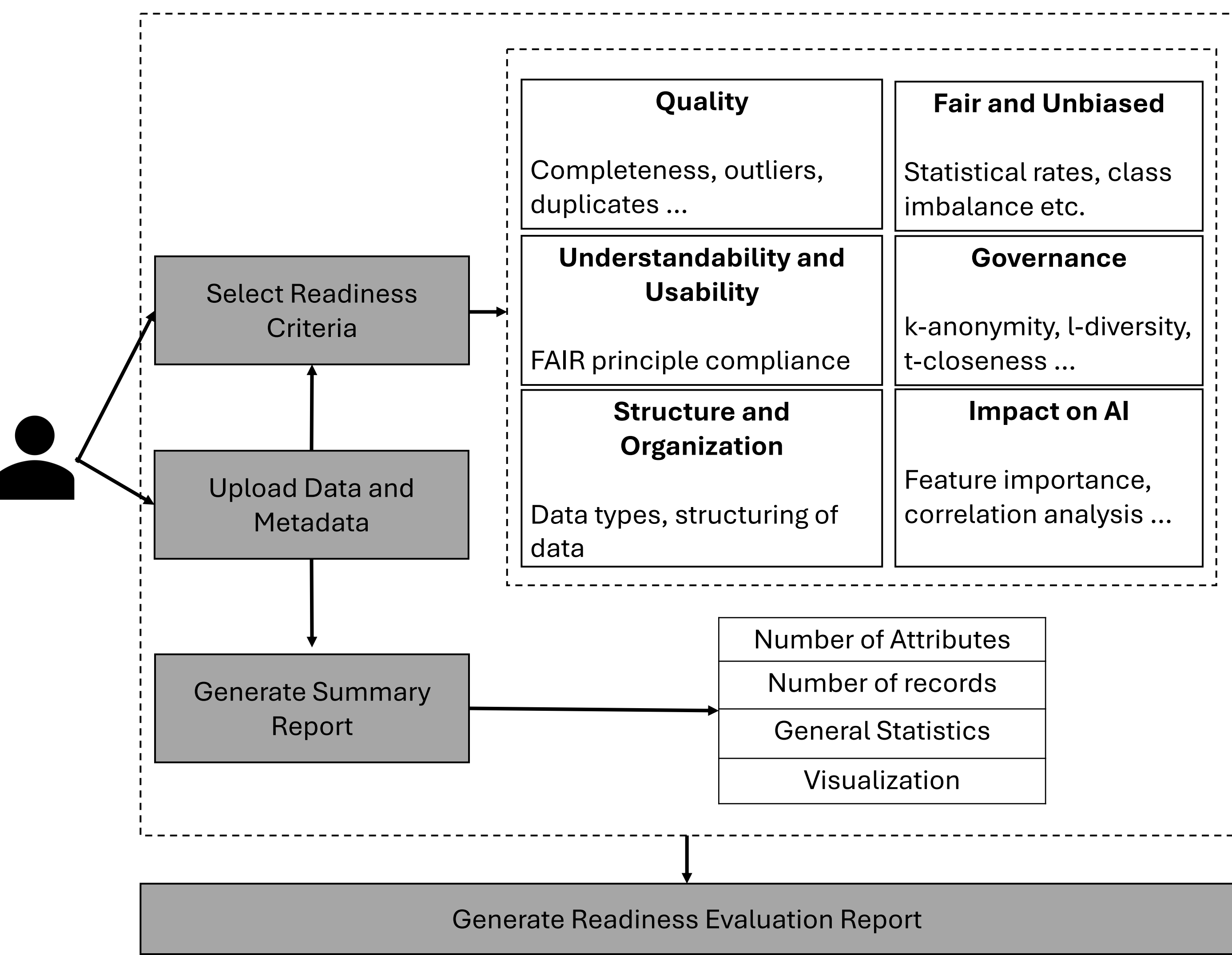

AIDRIN assesses data readiness across these pillars to provide a comprehensive evaluation of data suitability for AI

This work enhances AIDRIN by:
- Implementing an intuitive, interactive user interface with clear visual insights for centralized AI environments
- Enabling integration with APPFL[3] framework to support secure, federated and decentralized evaluations

AIDRIN OVERVIEW

UI Enhancements

Shown below are some user interfaces of AIDRIN 2.0

APPFL Integration

AIDRIN integrates to evaluate data characteristics at each edge client and assess their impact on overall model performance.

Documentation and code

Demo video

- Shown below are two actual AIDRIN reports generated on the server during a federated learning task using the AI-READI dataset.

Data Readiness Report

Client IDs	Class Imbalance	Data Range	Data Shape	Data Mean
Client1	0.27	-2.12 to 2.62	(220, 3, 32, 32)	1.4
Client2	0.28	-2.10 to 2.62	(162, 3, 32, 32)	1.47
Client3	0.25	-2.12 to 2.29	(113, 3, 32, 32)	1.4
Client4	0.25	-2.12 to 2.64	(202, 3, 32, 32)	2.22

Distribution of Classes - Client 1, Distribution of Classes - Client 2, Distribution of Classes - Client 3, Distribution of Classes - Client 4

Combined PCA Plot

- Based on the client's selected metrics before training, the data is evaluated locally to preserve privacy. The resulting evaluations and visualizations are then sent to the server to generate the final aggregated data readiness report
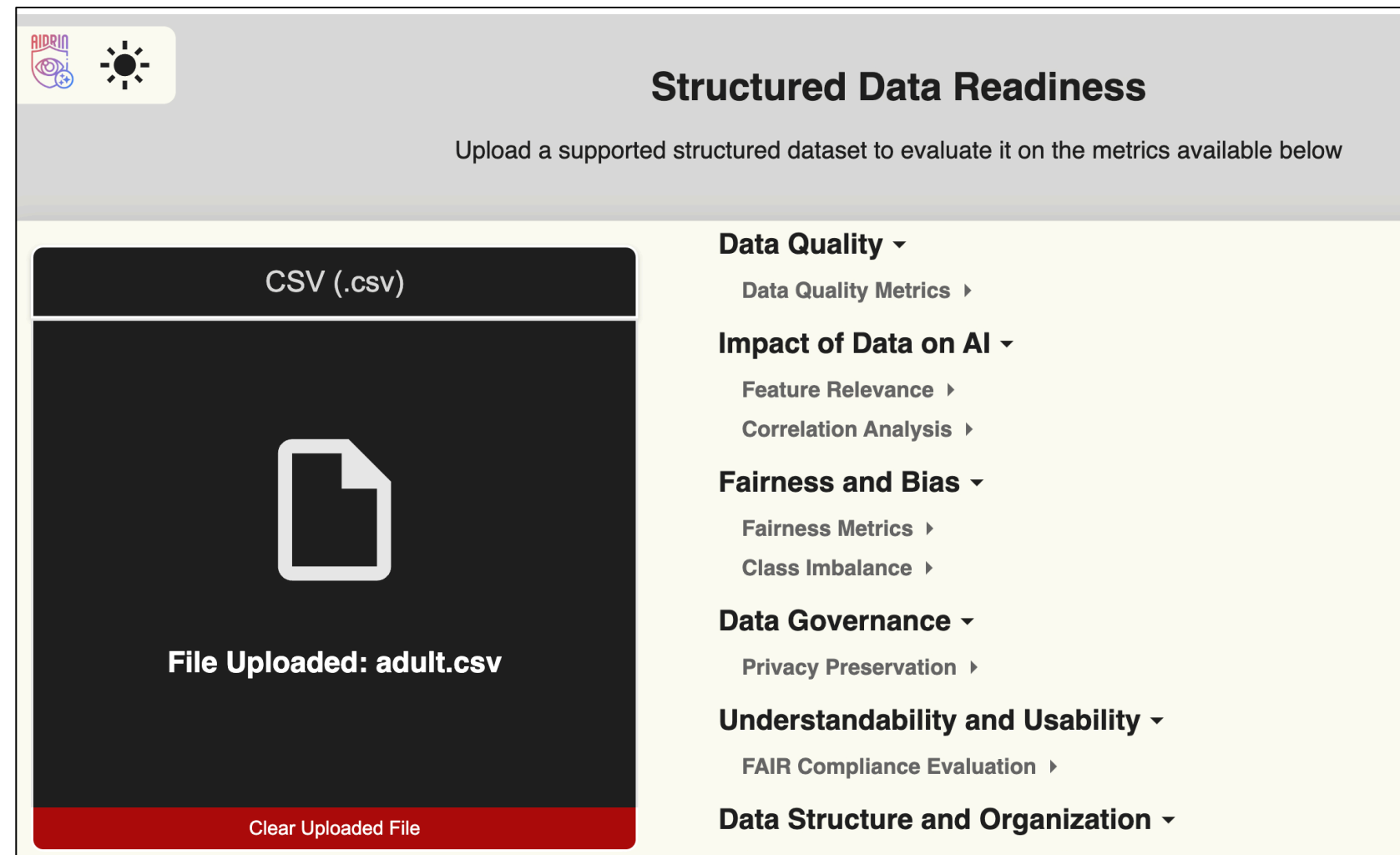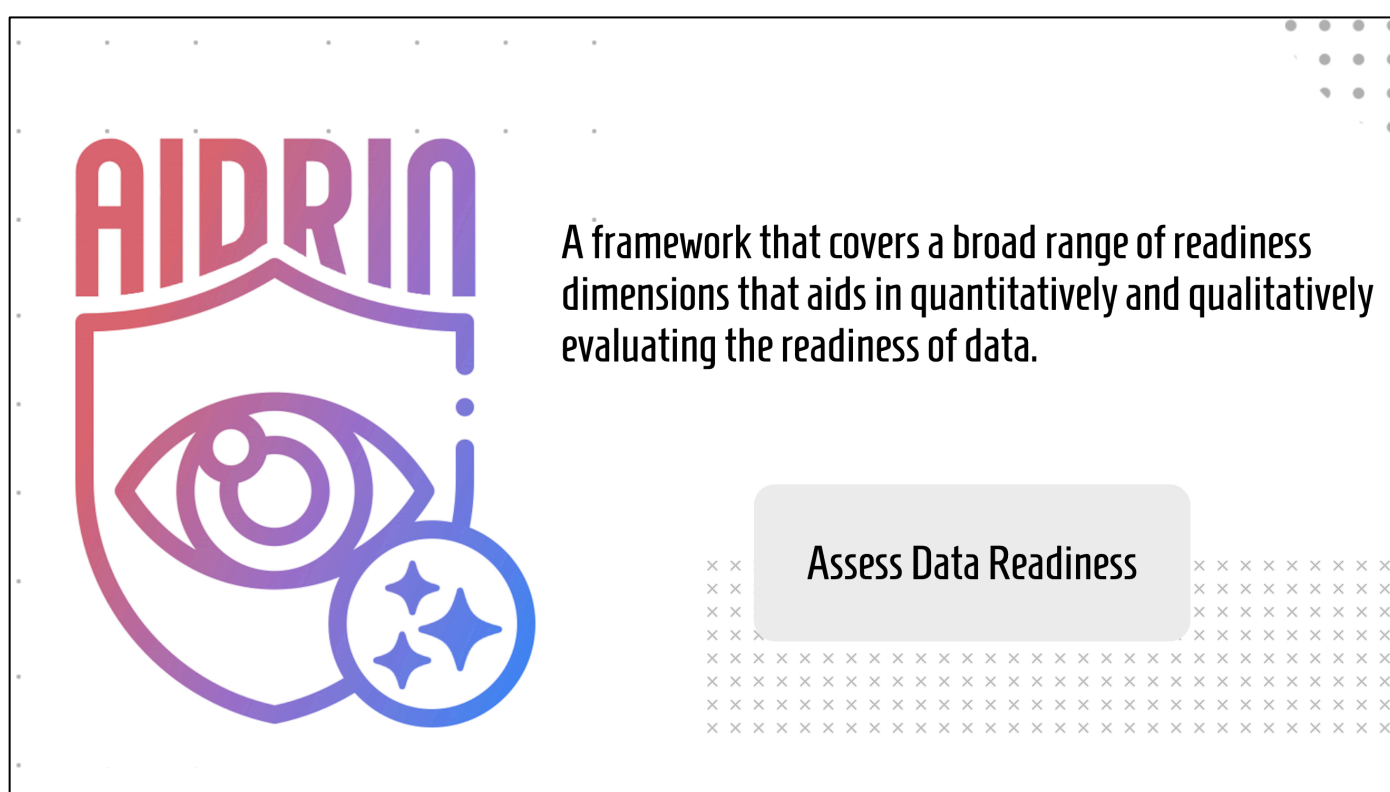
References

1. Kaveen Hiniduma, Suren Byna, Jean Luca Bez, and Ravi Madduri. 2024. AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI. In Proceedings of the 36th International Conference on Scientific and Statistical Database Management (SSDBM '24).
2. Kaveen Hiniduma, Suren Byna, and Jean Luca Bez. 2025. Data Readiness for AI: A 360-Degree Survey. ACM Comput. Surv. 57, 9, Article 219 (September 2025), 39 pages.
3. Z. Li, S. He, P. Chaturvedi, T.-H. Hoang, M. Ryu, E. Huerta, V. Kindratenko, J. Fuhrman, M. Giger, R. Chard et al., "APPFLx: Providing privacy-preserving cross-silo federated learning as a service," in 2023 IEEE 19th International Conference on e-Scie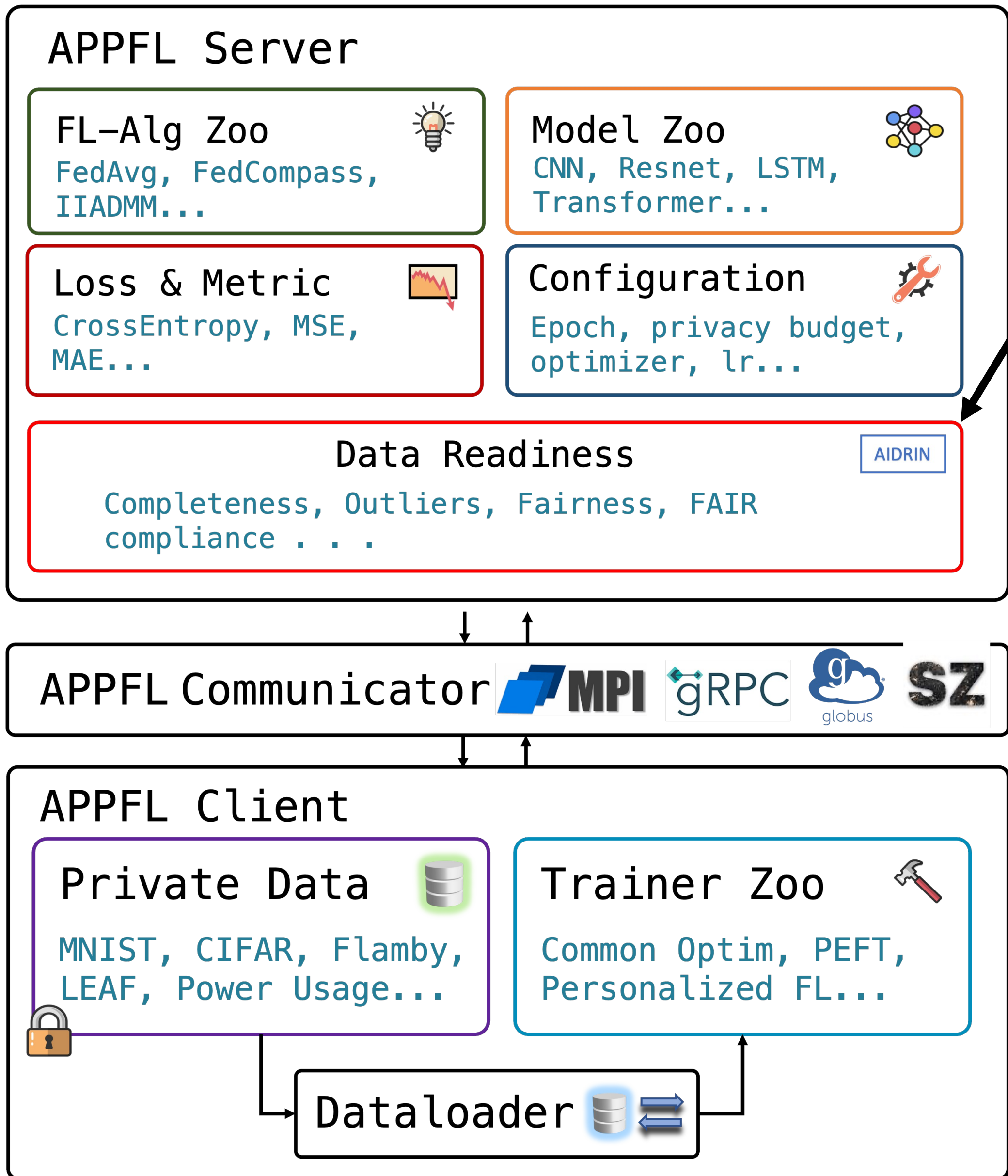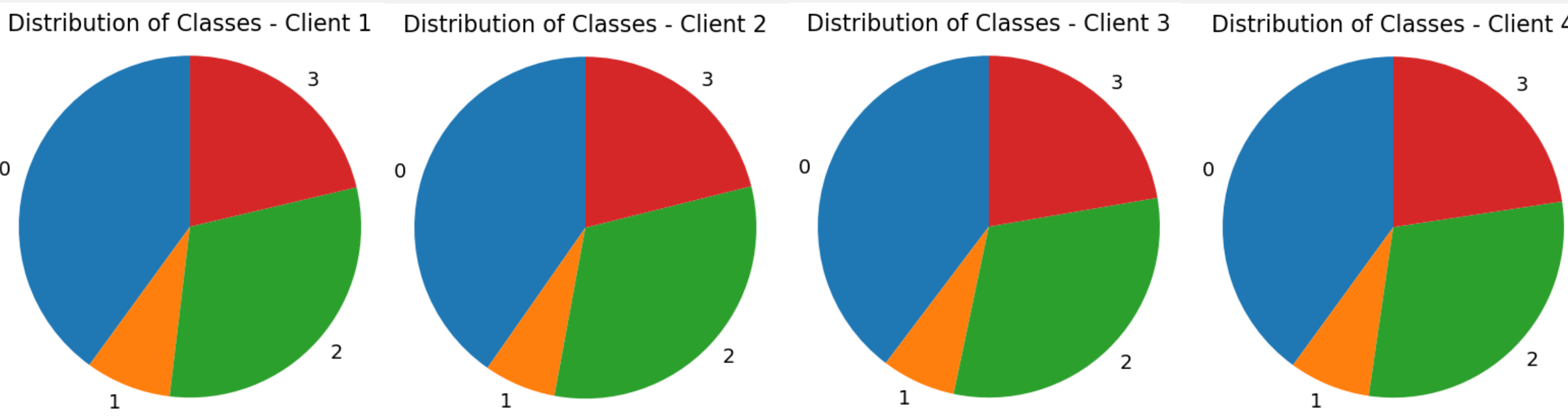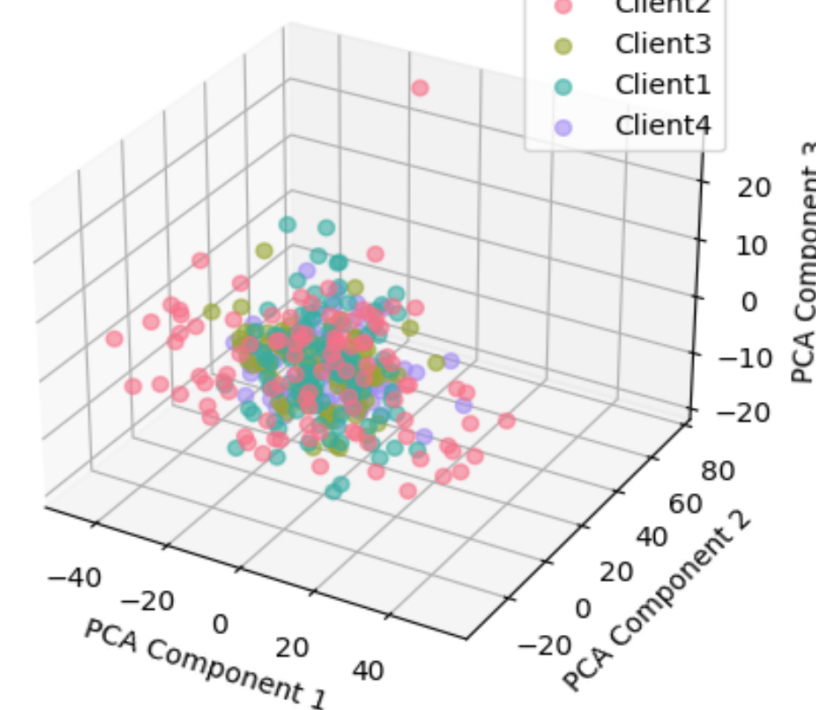nce (e-Science). IEEE, 2023, pp. 1-4.