CADRE: Customizable Assurance of Data Readiness in Privacy-Preserving Federated Learning

Kaveen Hiniduma, Zilinghan Li, Aditya Sinha, Ravi Madduri, Suren Byna

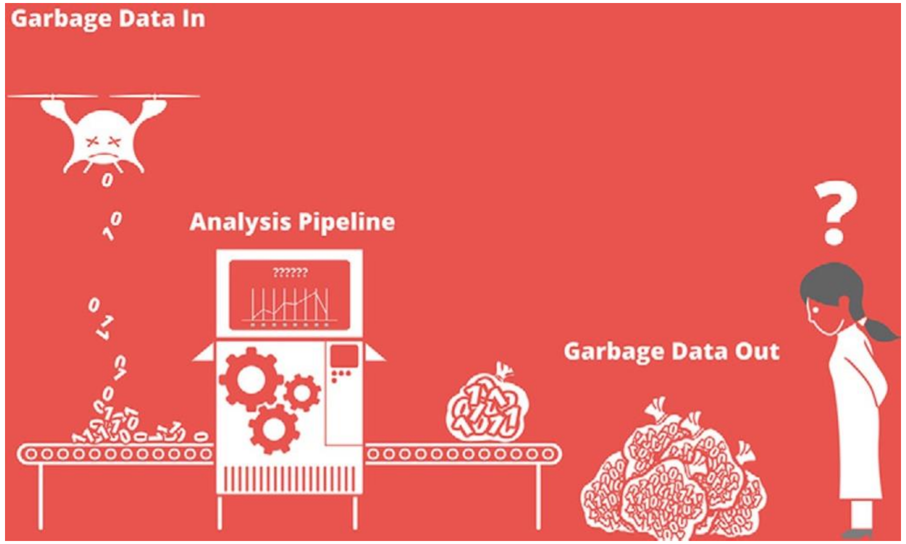
IEEE eScience, September 15-18, 2025, Chicago, USA

THE OHIO STATE UNIVERSITY

Argonne NATIONAL LABORATORY

Why should the data be ready for AI?

- Garbage In, Garbage Out, not only specific to AI
  - Data is the essential fuel for AI applications.
  - Low quality, biased data leads to ineffective and unreliable AI models
    - More critical when AI is used in decision making systems
  - High-quality data ensures accurate, fair, and robust AI outcomes



Nadia Shakoor et al., "Big Data Driven Agriculture: Big Data Analytics in Plant Breeding, Genomics, and the Use of Remote Sensing Technologies to Advance Crop Productivity", https://acsess.onlinelibrary.wiley.com/doi/full/10.2135/tppj2018.12.0009

Challenges and objectives

- Challenges facing data scientists
  - Poorly structured data from heterogeneous sources
  - Extensive time and effort are required for data preparation
  - Lack of standardized methods to assess data readiness for AI
- Objectives
  - What is data readiness for AI?
  - What are existing frameworks for assessing data and what are the gaps?
  - What are the requirements of a standardized, quantitative approach for assessing AI data readiness?

"If 80 percent of our work is data preparation, then ensuring data quality is the important work of a machine learning team."

Andrew Ng, Professor of AI at Standford University and founder of DeepLearning.AI

Image source: https://research.aimultiple.com/data-quality-ai/

What is data readiness for AI?

- Numerous dimensions to assess data quality and readiness
- A standard definition is still evolving

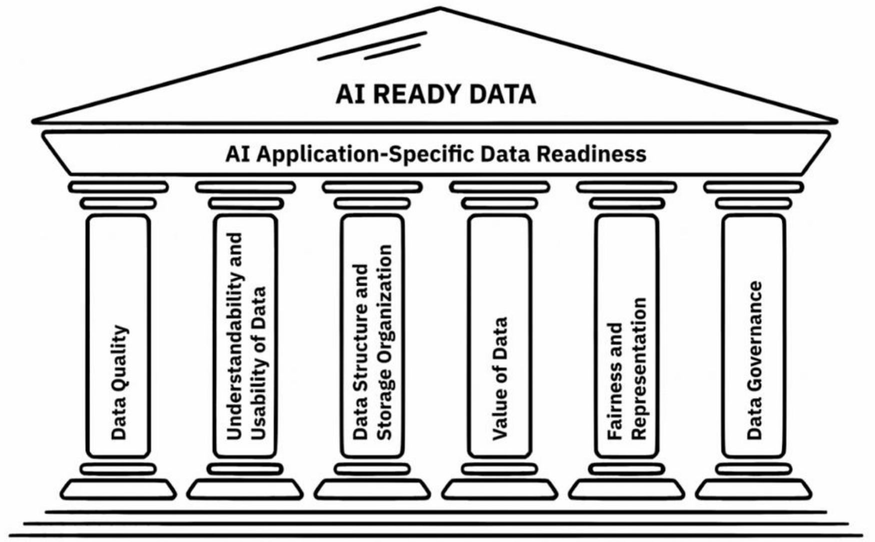
- Common factors considered in data processing now
  - Quality → Diverse definitions for structured and unstructured data
  - Findable, Accessible, Interoperable, and Reusable (FAIR) principles for data

- Data Readiness for AI metrics survey[1]

[1] 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. https://doi.org/10.1145/3722214

Surveyed data readiness for AI

- A taxonomy of metrics to evaluate data readiness for AI training, covering structured/unstructured data
- Method
  - Surveyed 140+ papers from ACM, IEEE, Springer, and expert articles to identify gaps in standardized metrics
- Key Insight
  - Poor-quality and not ready data leads to inaccurate and unethical AI model outcomes ("garbage in, garbage out")



Gaps and Challenges Identified

- **Lack of a unified framework** to assess readiness across structured, unstructured, multimodal data, and in different ML settings such as FL
- **Limited scope of existing tools** (e.g., IBM DQT) mostly focused on structured data.
- **Scalability issues** when evaluating readiness in large, complex datasets.
- **Evolving and domain-specific metrics** complicate standardization across applications.
- **Interpretability challenges** as stakeholders may struggle to understand complex readiness metrics.
- **Privacy, fairness, and human bias** introduce subjectivity in assessments.
- **Lack of clear rules** to define acceptable levels of data readiness.

AI data readiness - Evaluation metrics

Data Quality	Understandability and Usability of Data	Data Structure and Organization	Data Governance	Impact of Data on AI	Fair and Unbiased Data
Completeness	FAIR principal compliance	Sample size	Privacy leakage	Feature relevance	Discrimination/ bias index
Correctness					
Timeliness		Appropriate data split ratios (train/validation/test)		Data point impact	Class imbalance
Mislabeling					
Multimedia data quality					Class separability

Kaveen Hiniduma, Suren Byna, Jean Luca Bez, and Ravi Madduri, "AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI", SSDBM 2024

Integration of AIDRIN into PPFL (Privacy-Preserving Federated Learning Framework)

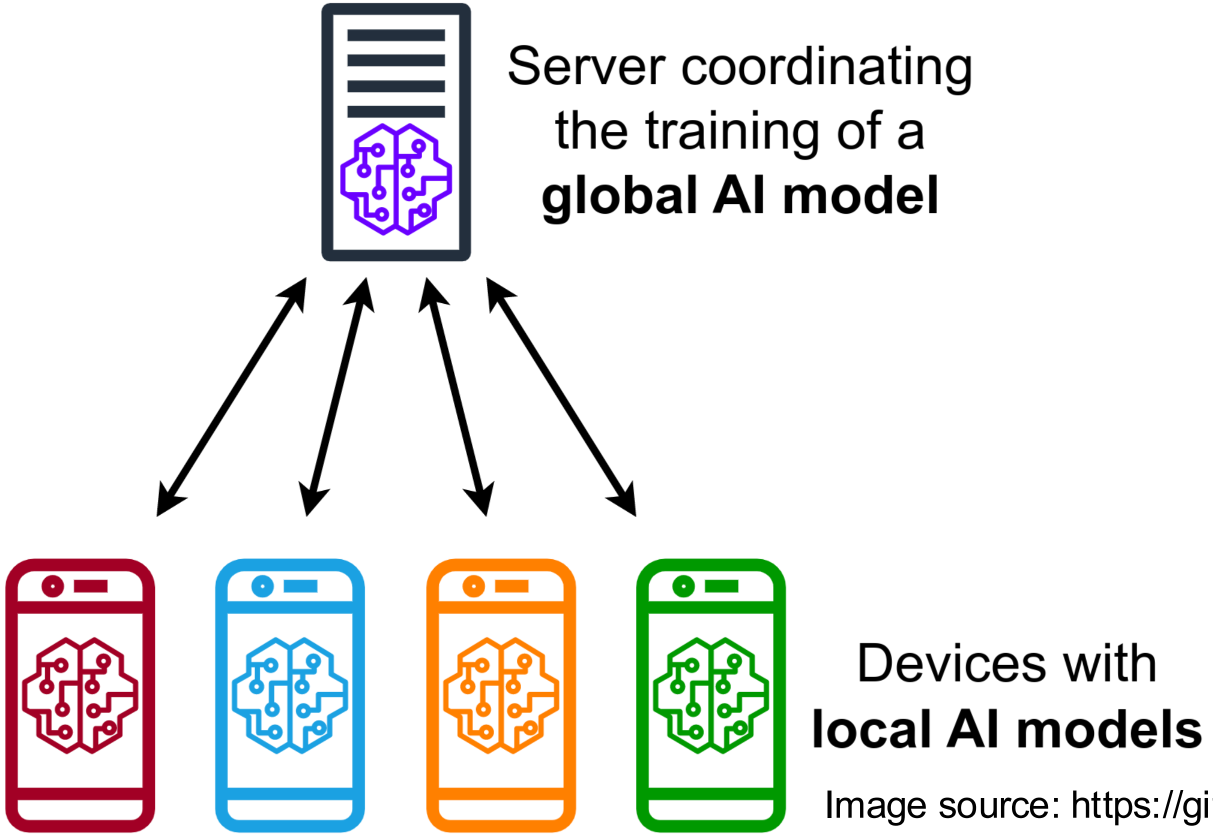


Image source: https://github.com/cs-joy/federated-learning

Data Readiness Challenges in PPFL

- Data heterogeneity and quality issues (e.g., noise, imbalance)
- Hard to detect/fix unprepared data due to privacy constraints
- Unprepared data leads to poor model performance, resource waste and deployment failures

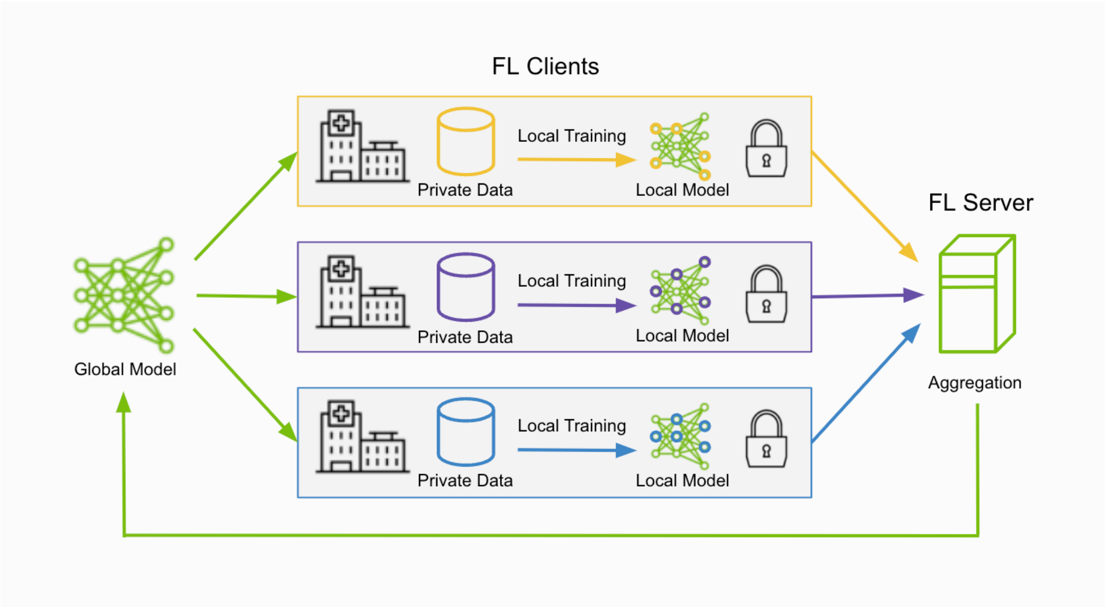
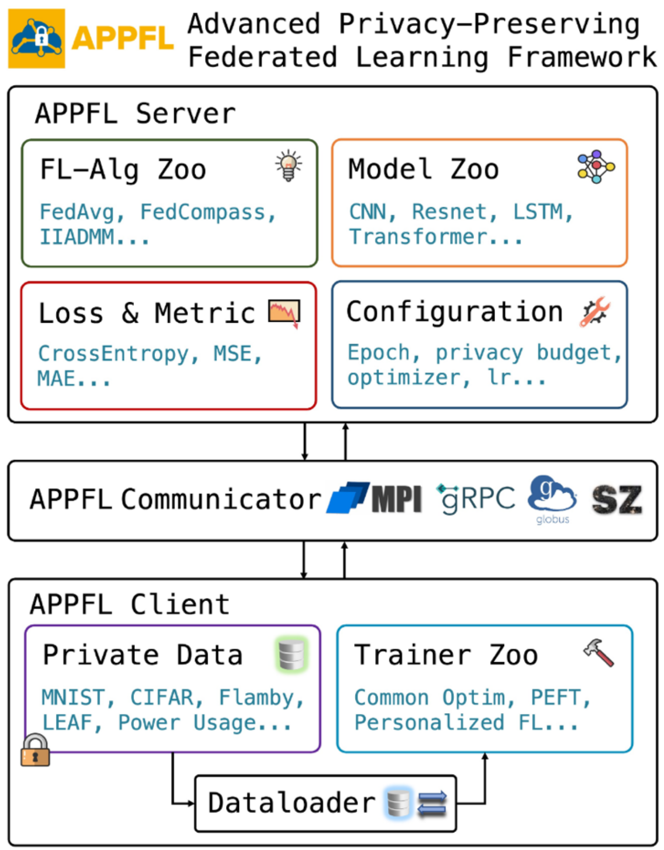
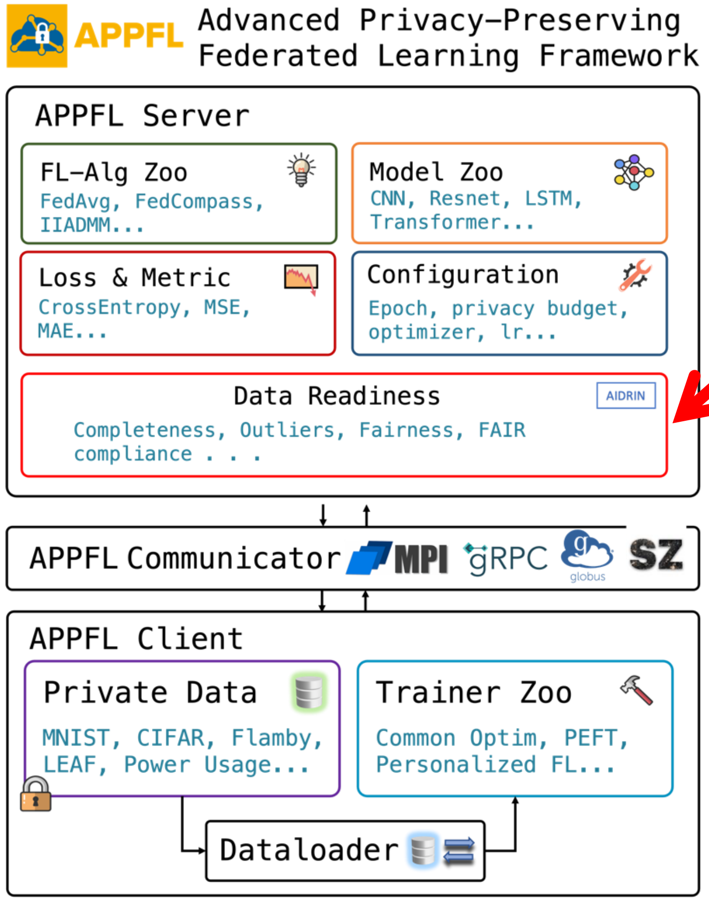


Image source: NVIDIA FLARE Documentation – Introduction to Federated Learning. https://nvflare.readthedocs.io/en/main/fl_introduction.html

Integration of AIDRIN into APPFL (Advanced PPFL)



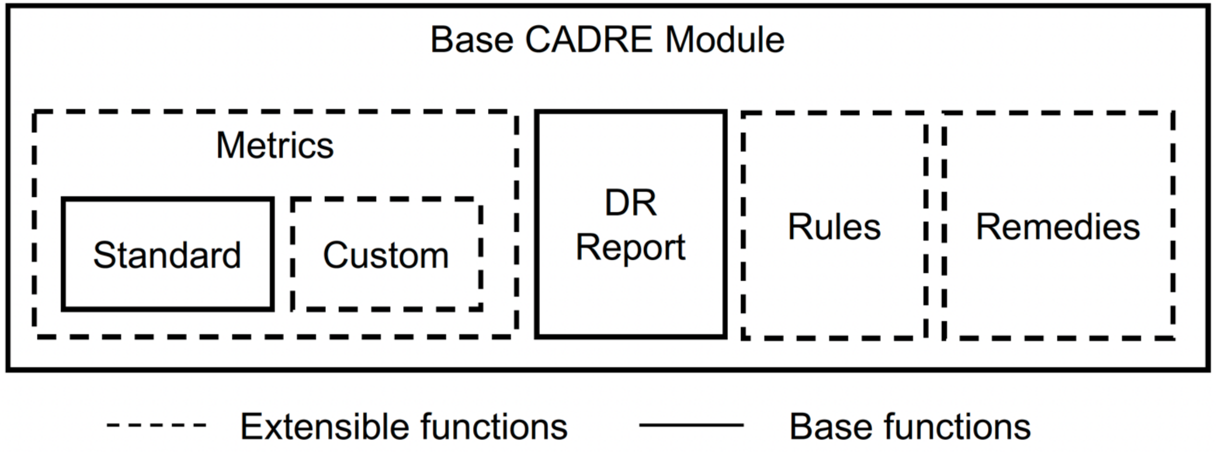


AIDRIN integration to study data characteristics at each site and impact on the model performance

https://github.com/APPFL/APPFL/tree/main

Customizable Assurance of Data Readiness (CADRE)

- Customizable framework to evaluate and assure data readiness standards – metrics, rules, remedies
- Users can define and verify data readiness standards while preserving privacy through local execution



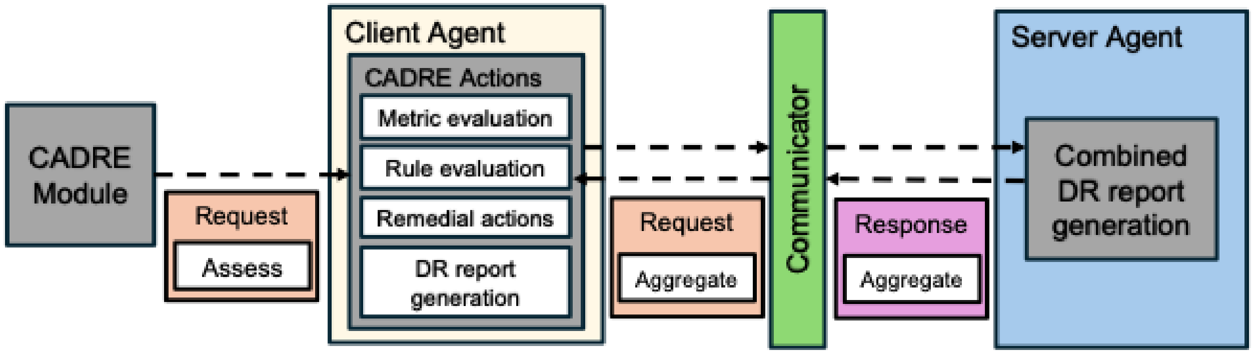
Defining custom metrics, rules, and remedies

```python
from appfl.misc.data_readiness import BaseCADREModule

class MyCustomCADREModule(BaseCADREModule):
    def __init__(self, train_dataset, **kwargs):
        super().__init__(train_dataset, **kwargs)

    def metric(self, **kwargs):
        # Compute and return your metric as a dictionary
        # Example: return {"my_metric1": 0.5, "my_metric2": 0.8, ...}
        pass

    def rule(self, metric_result, threshold=0.0):
        # Define the logic to check if a problem exists (optional)
        # Example: return metric_result["my_metric"] > threshold
        pass

    def remedy(self, metric_result, logger, **kwargs):
        # Apply remedy and return updated dataset in dictionary format (optional)
        # Example:
        # return {"ai_ready_dataset": self.train_dataset, "metadata": None}
        pass
```

```yaml
cadremodule_configs:
    cadremodule_path: ./resources/configs/mnist_dr/cadre_module/handle_ci.py
    cadremodule_name: CADREModuleCI       # Name of the class inside the .py file
    remedy_action: true                   # Apply remedy if supported
```

Customizable Assurance of Data Readiness (CADRE)

- Customizable framework to evaluate and assure data readiness standards – metrics, rules, remedies
- Users can define and verify data readiness standards while preserving privacy through local execution



https://appfl.ai/en/latest/tutorials/examples_dr_integration.html

Custom Metrics, Rules, and Remedies

- Validated on 6 datasets covering 7 DR challenges across diverse data modalities and tasks

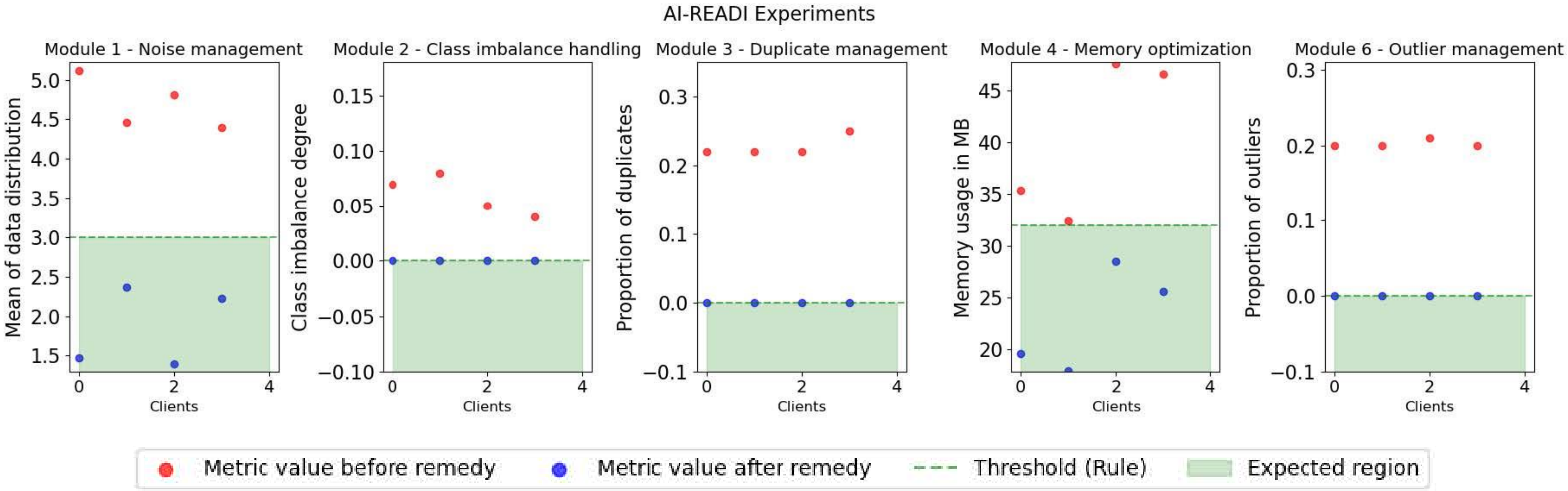
CADRE Module ID	Category	Metric	Rule	Remedy
1	Noise Management	Mean magnitude of the data (image intensities or feature values)	Applied remedy when the data distribution mean exceeded a threshold (e.g., > 0.37 for MNIST).	Data points with noisy indices were removed.
2	Class Imbalance Handling	Class imbalance degree [33]	Applied when imbalance degree > 0.	SMOTE [34] was used to oversample the minority class.
3	Duplicate Management	Proportion of duplicates	Applied when duplicates proportion > 0.	Duplicates were identified and removed.
4	Memory Optimization	Memory usage in megabytes (MB) to store the client's data	Applied when memory usage was excessively high.	Data types were optimized or duplicates removed depending on the dataset's pollution method.
5	Bias Handling	Statistical parity difference [35] for Adult Income dataset and representative rate difference for TCGA-BRCA dataset	Applied when metric value > 0.	Stratified resampling [36] to balance sensitive groups and labels in the Adult Income dataset, while SMOTE to oversample the minority group in the TCGA-BRCA dataset.
6	Outlier Management	Proportion of outliers using Interquartile range (IQR) method [37]	Applied when outliers proportion > 0.	Outliers were clipped at IQR bounds.
7	K-anonymity Handling	K-anonymity level [38]	Applied when anonymity level ≤ 1.	Data records with low anonymity levels were suppressed to ensure the desired level of anonymity.

Datasets

MNIST – Image classification, CIFAR-10 – Object recognition, Adult Income – Tabular data, income classification, Flamby TCGA-BRCA – Clinical data, survival analysis, Flamby IXI – 3D brain images, image segmentation, AI-READI – Retinal images, severity classification

Evaluation of AI-READI data with CADRE

- AI-READI (Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights) dataset
  - To evaluate data readiness, manually **polluted** the data (e.g., noise) and applied rules and remedies.
  - Iteratively apply remedies until the rules/standards are achieved.



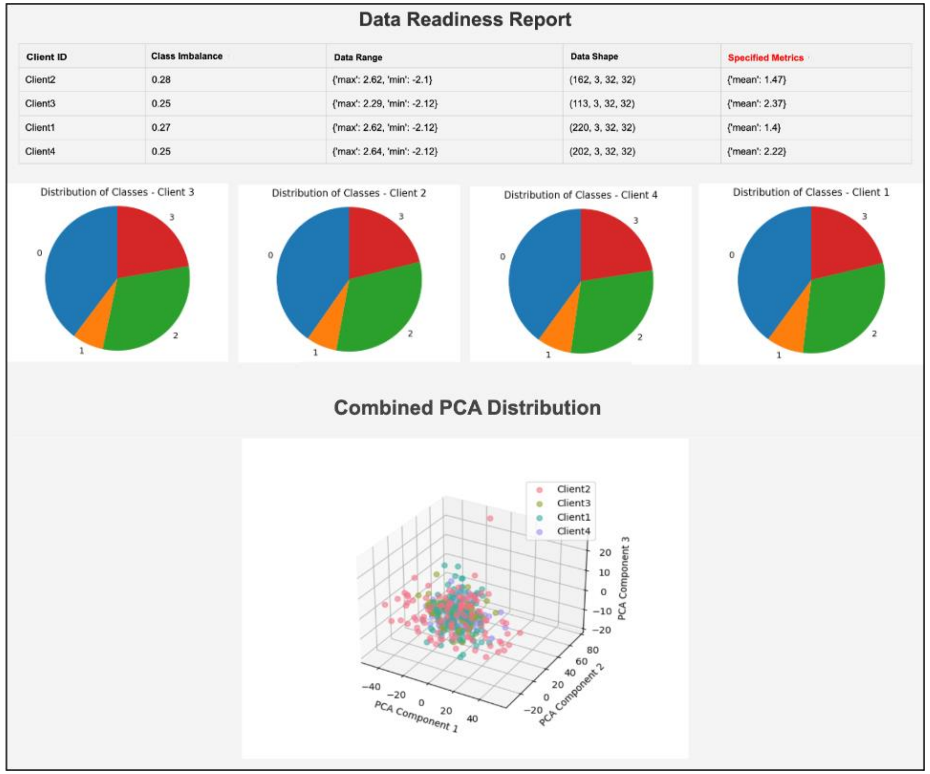
Data readiness reports

- Generates DR reports for easier understandability of data characteristics across the clients involved

Data Readiness Report

Client ID	Class Imbalance	Data Range	Data Shape	Specified Metrics
Client3	0.24	{'max': 5.0, 'min': -2.12}	(1535, 3, 32, 32)	{'mean': 5.12}
Client2	0.25	{'max': 5.0, 'min': -2.1}	(1677, 3, 32, 32)	{'mean': 4.46}
Client4	0.24	{'max': 5.0, 'min': -2.12}	(2257, 3, 32, 32)	{'mean': 4.81}
Client1	0.24	{'max': 5.0, 'min': -2.12}	(2207, 3, 32, 32)	{'mean': 4.4}

Before CADRE



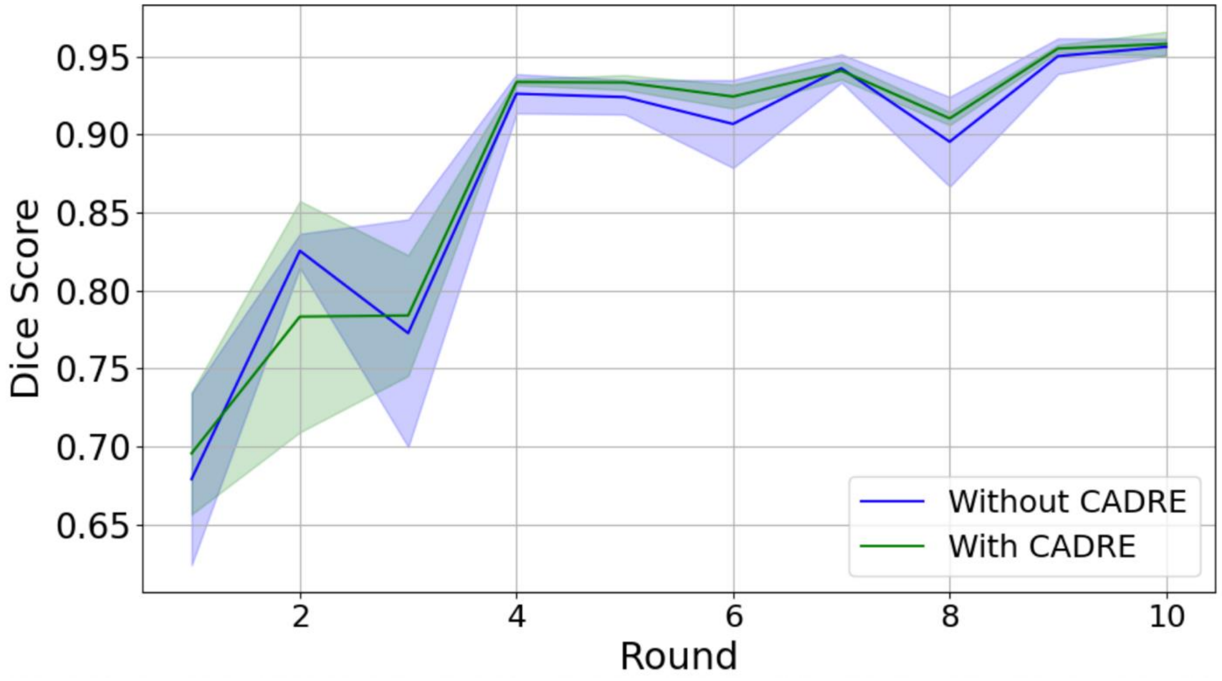
Data Readiness Report

Client ID	Class Imbalance	Data Range	Data Shape	Specified Metrics
Client2	0.28	{'max': 2.62, 'min': -2.1}	(162, 3, 32, 32)	{'mean': 1.47}
Client3	0.25	{'max': 2.29, 'min': -2.12}	(113, 3, 32, 32)	{'mean': 2.37}
Client1	0.27	{'max': 2.62, 'min': -2.12}	(220, 3, 32, 32)	{'mean': 1.4}
Client4	0.25	{'max': 2.64, 'min': -2.12}	(202, 3, 32, 32)	{'mean': 2.22}

After CADRE

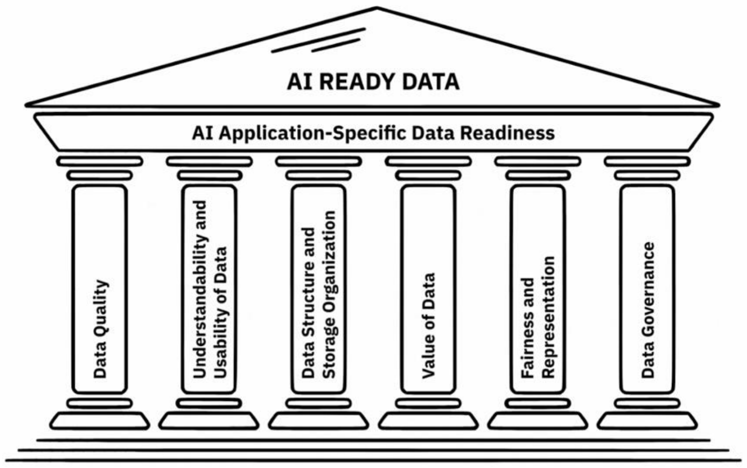
Impact on Model Performance

- **Model improvements:** Higher dice scores in segmentation (IXI Tiny) after cleaning noisy samples.
- Reduced variability, better generalization.
- CADRE boosts performance



Conclusion

- AI-ready data is vital for trustworthy AI-assisted decision-making
- AIDRIN with CADRE is a step towards developing a comprehensive framework
- APPFL integration
  - https://appfl.ai/en/latest/tutorials/examples_dr_integration.html
- Demo
  - https://www.youtube.com/watch?v=aBVgtV65t5M&t=1s
- Ongoing work
  - Composite AIDRIN Score that is meaningful to AI applications and to improve data
  - Parallel version of AIDRIN to evaluate massive datasets



github.com/idtlab/AIDRIN
aidrin.readthedocs.io
hiniduma.1@osu.edu

The Ohio State University
Argonne National Laboratory