AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI

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

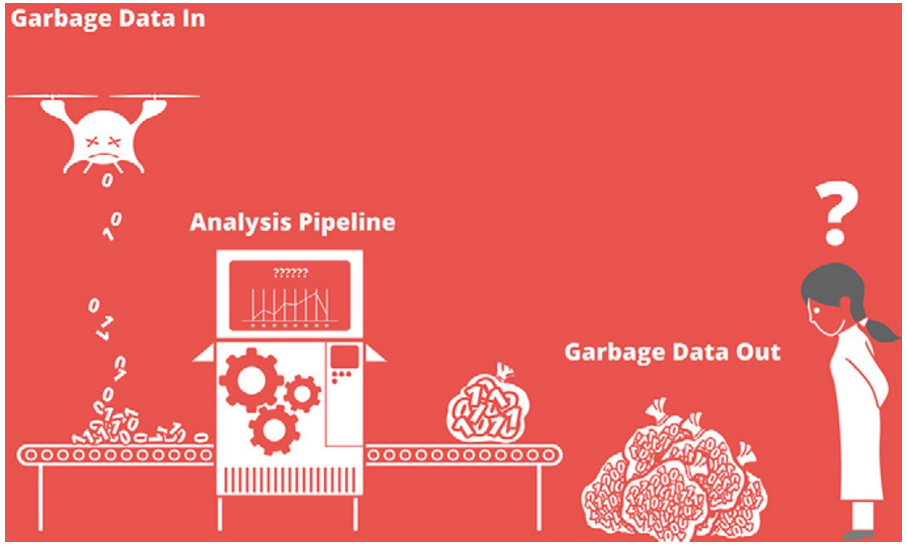
[1] The Ohio State University

[2] Lawrence Berkeley National Laboratory

[3] Argonne National Laboratory

Problem

- Garbage In, Garbage Out (GIGO), 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

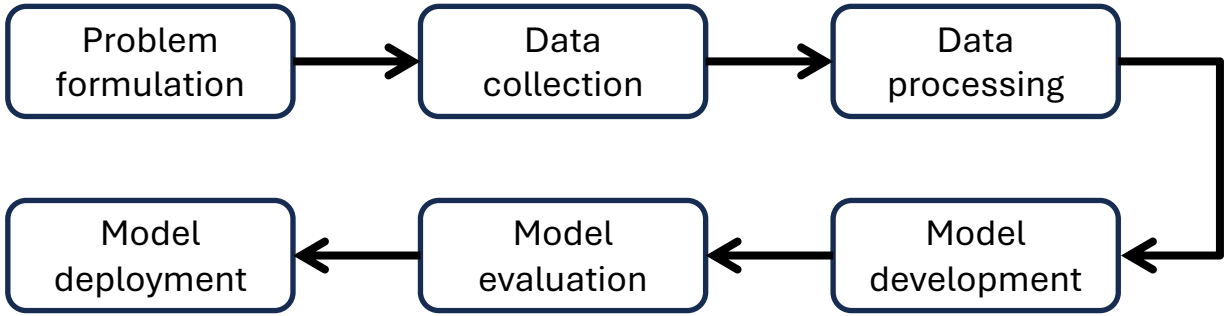
- Challenges facing data scientists
  - Poorly structured data from heterogeneous sources
  - Extensive time and effort are required for data preparation
  - Lack of standardized methods to assess data readiness for AI
- Objectives
  - What is data readiness for AI?
  - What are existing frameworks for 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*

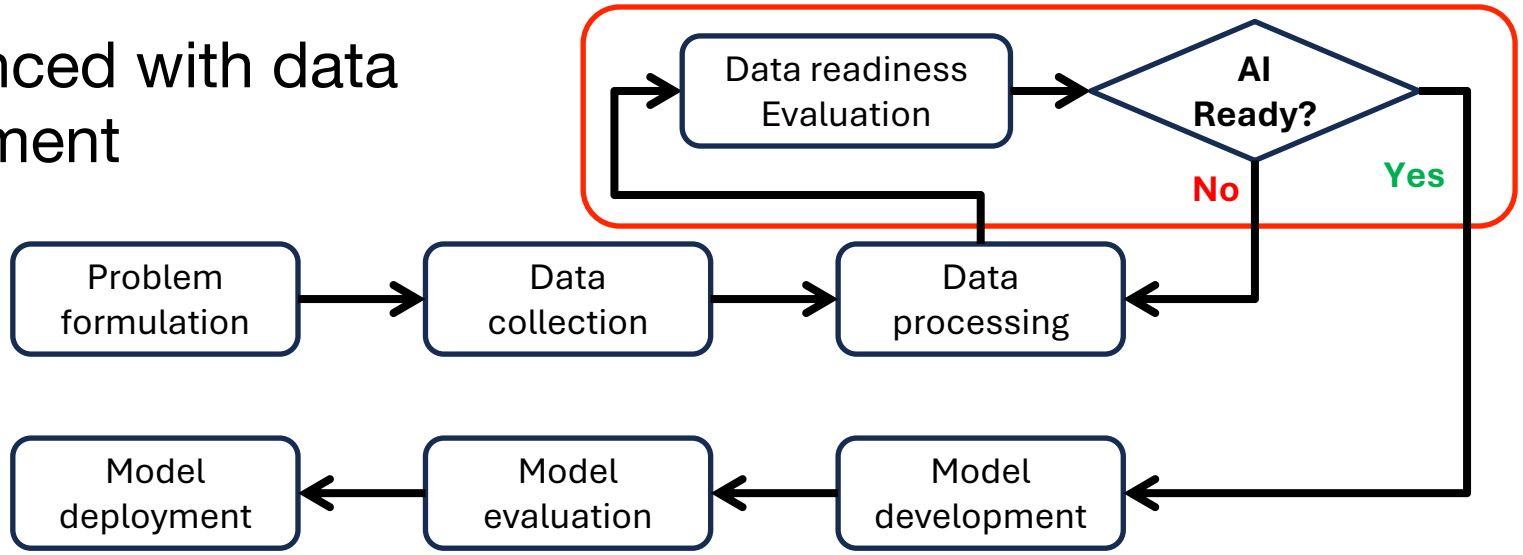
Image source: https://research.aimultiple.com/data-quality-ai/

Where does data readiness assessment play a role in ML/AI?

- Traditional ML pipeline



- ML pipeline enhanced with data readiness assessment



What is data readiness for AI?

- Numerous dimensions
- A 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 – *In review*

Kaveen Hiniduma, Suren Byna, Jean Luca Bez, "Data Readiness for AI: A 360-Degree Survey", in submission (An draft version is available on arXiv (https://arxiv.org/abs/2404.05779)

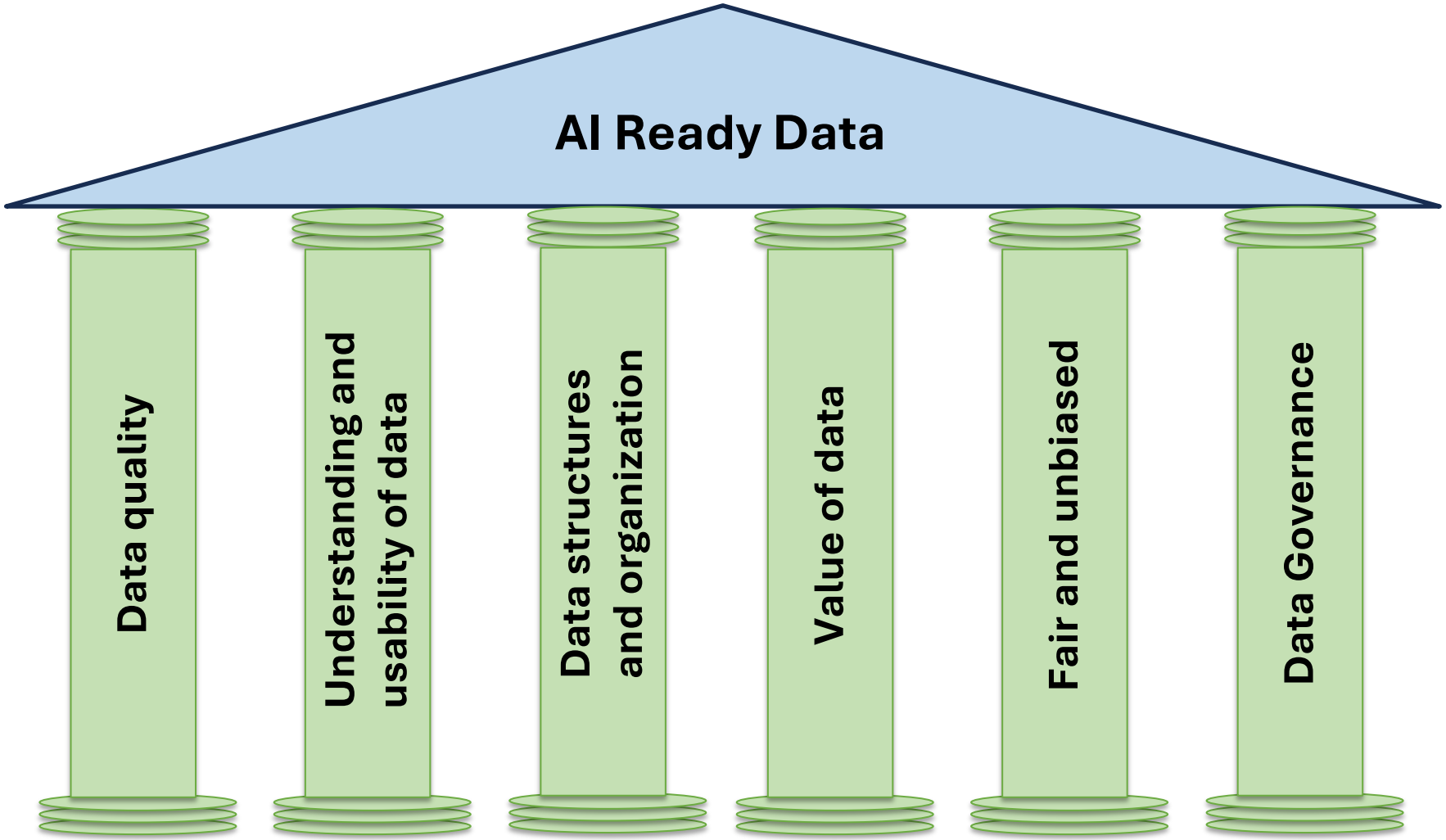
Requirements for defining AI readiness

- Standardized assessment methods
  - **Metrics**: Widely-accepted metrics for evaluating data readiness
  - **Quantitative and qualitative analysis**: Comprehensive approach combining both types of metrics (i.e., quantitative and qualitative) for a comprehensive assessment

- Efficient data preparation
  - **Time and effort reduction**: Streamline data preparation processes to save time and effort.
  - **Informed decision-making**: Enable data scientists to make well-informed decisions about data readiness for AI applications.

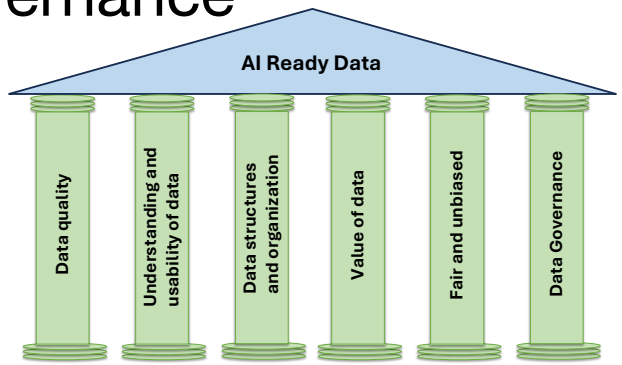
- Informative visualizations and reports
  - **Clear insights**: Provide visualizations to help understand data readiness status.
  - **Actionable reports**: Generate detailed reports to guide further data preparation steps.

Our proposal – Pillars of "AI Ready Data" assessment



Our proposal for a definition of "AI data readiness"

- High quality data
  - Completeness
  - Outliers
  - Duplication
- Understanding the data
  - Metadata availability and quality
  - Provenance
- Ensuring structural quality of data
  - Data types used
  - Data schema quality and consistency
  - File format and storage system
- Understanding the value of data
  - Feature importance
  - Label quality
  - Data point impact
  - Uncertainty quantification
- Ensure fairness and mitigate bias
  - Class imbalance
  - Bias mitigation
- Effective data governance
  - Security
  - Privacy
  - Collection



- **AI application-specific metrics**

A framework for assessment – Existing work

Tool	Completeness	Outliers	Duplicates	Privacy	Fairness	FAIR Compliance	Feature Correlations	Feature Relevancy	Class Imbalance
Informatica [30]	✓								
DQLearn [47]	✓	✓	✓						
Gupta et al [25]	✓	✓	✓		✓		✓	✓	✓
Data Readiness Report [5]	✓	✓				✓	✓		
AI360 [8]					✓				
FAIR Cookbook [43]						✓			
FAIRassist [19]						✓			
ESS-DIVE FAIR [18]						✓			

Proposed framework: AI Data Readiness INspector (AIDRIN)

- Features
  - **Comprehensive framework:** AIDRIN assesses data readiness across a variety of dimensions
  - **Key metrics:** Evaluates completeness, accuracy, duplicates, feature importance, fairness, privacy, and FAIR principle compliance
  - **Extensible**: Designed to be modular to easily add calculations of new metrics
- Standardized evaluation:
  - **Accepted metrics:** Uses consistent assessment with widely-accepted metrics.
  - **Quantitative and qualitative analysis:** Provides both types of analysis for thorough data evaluation.

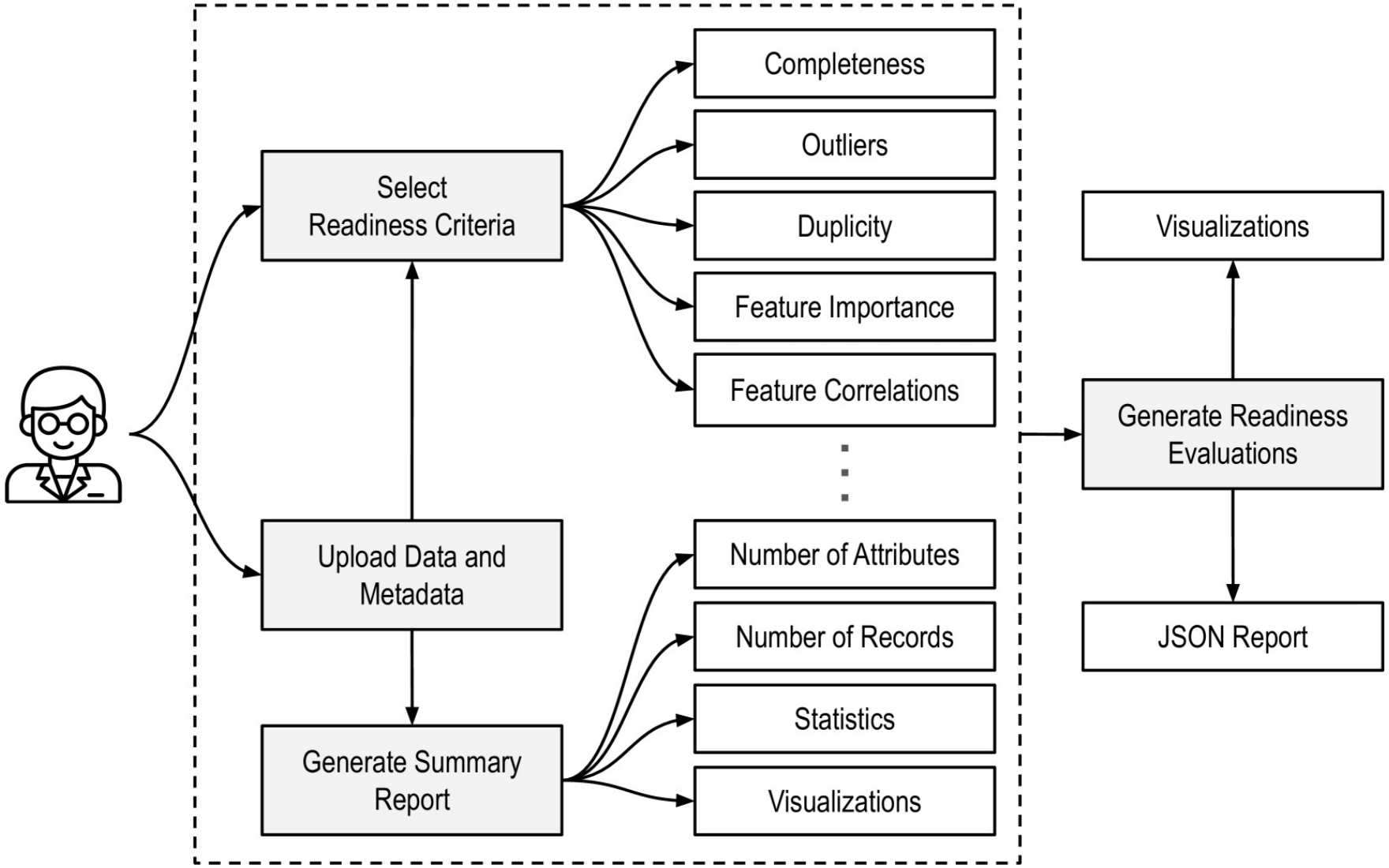
Proposed framework: AI Data Readiness INspector (AIDRIN)

- Efficiency and effectiveness:
  - **Informed decision-making:** Empowers data scientists with actionable insights for data readiness.
  - **Goal-oriented data preparation:** Aimed to reduce time and effort in data preparation.
- Visualizations and reporting:
  - **Extensive visualizations:** Presents data quality and readiness visually.
  - **Actionable reports:** Delivers detailed reports to guide further data preparation steps.
- Deployment
  - **Web application**: Web-based interface for easy accessibility (To be released)
  - **PyPI package**: Available on PyPI for simple installation and integration into Python environments. (https://test.pypi.org/project/aidrin)

A framework for assessment – AIDRIN comparison

Tool	Completeness	Outliers	Duplicates	Privacy	Fairness	FAIR Compliance	Feature Correlations	Feature Relevancy	Class Imbalance
Informatica [30]	✓								
DQLearn [47]	✓	✓	✓						
Gupta et al [25]	✓	✓	✓		✓		✓	✓	✓
Data Readiness Report [5]	✓	✓				✓	✓		
AI360 [8]					✓				
FAIR Cookbook [43]						✓			
FAIRassist [19]						✓			
ESS-DIVE FAIR [18]						✓			
AIDRIN	✓	✓	✓	✓	✓	✓	✓	✓	✓

AIDRIN workflow



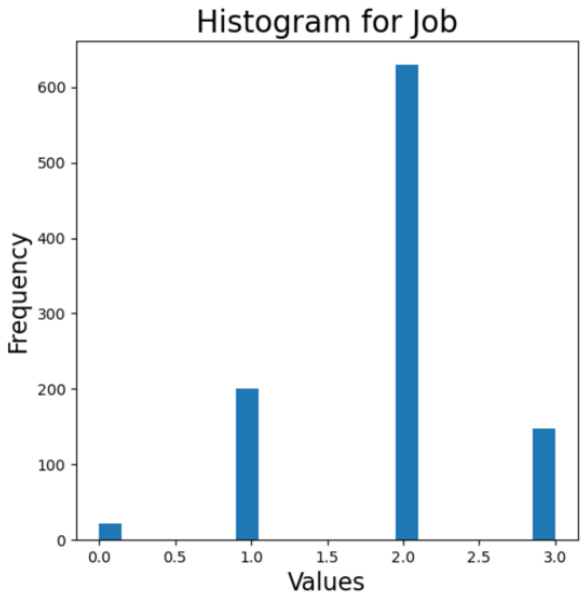
Target Standard Deviation (TSD) metric for bias calculation

- **Addressing limitations of traditional metrics**
  - Traditional metrics often limited to binary-sensitive attributes.
  - Non-binary attributes remain unaddressed in many evaluations.
- **Target Standard Deviation (TSD) metric**
  - **Beyond binary groups:** Considers average differences across all subgroups for a given target.
  - **Unified scalar value:** Facilitates straightforward and interpretable evaluations.
  - **Comparative insights:** Enables effective comparison and prioritization of decisions.
- **Benefits and Limitations**
  - **Benefit:** Captures variations within non-binary attributes.
  - **Limitation:** May not consider contextual factors or multiple sensitive attributes simultaneously.

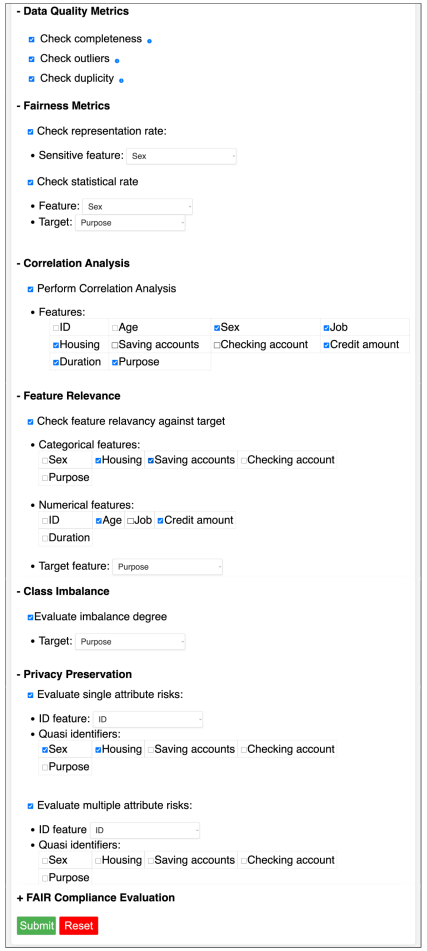
AIDRIN – Web Interface – Summary statistics

Structured Data Readiness

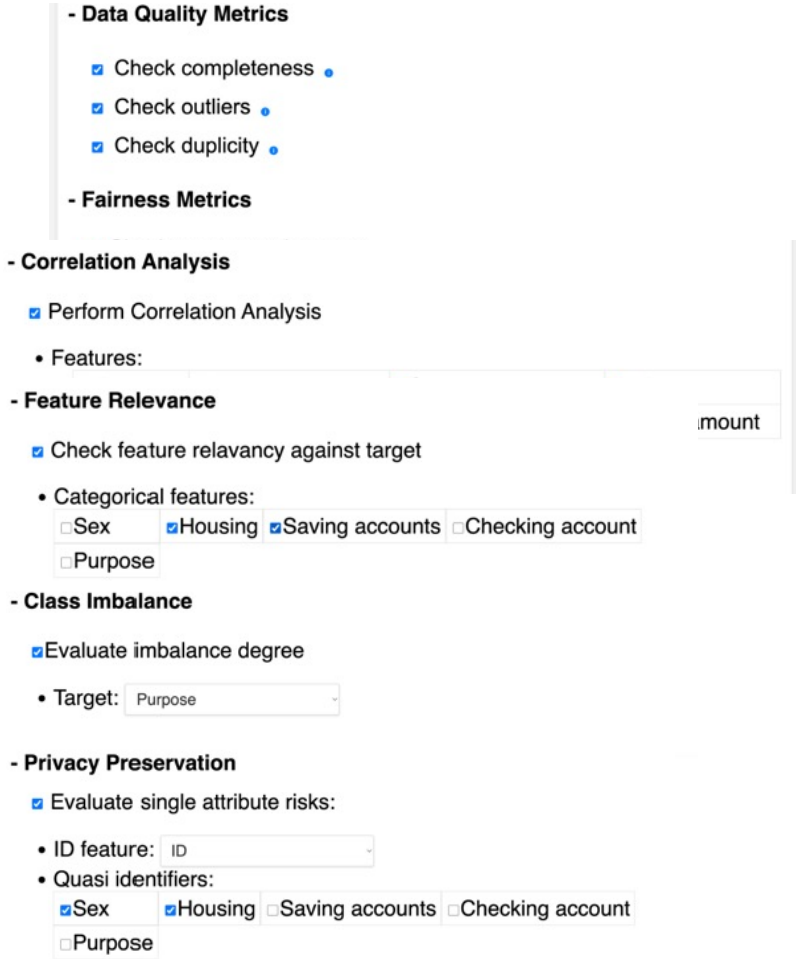
Summary Statistic Plots



AIDRIN – Web Interface – Quality, privacy, fairness metrics



Select readiness criteria



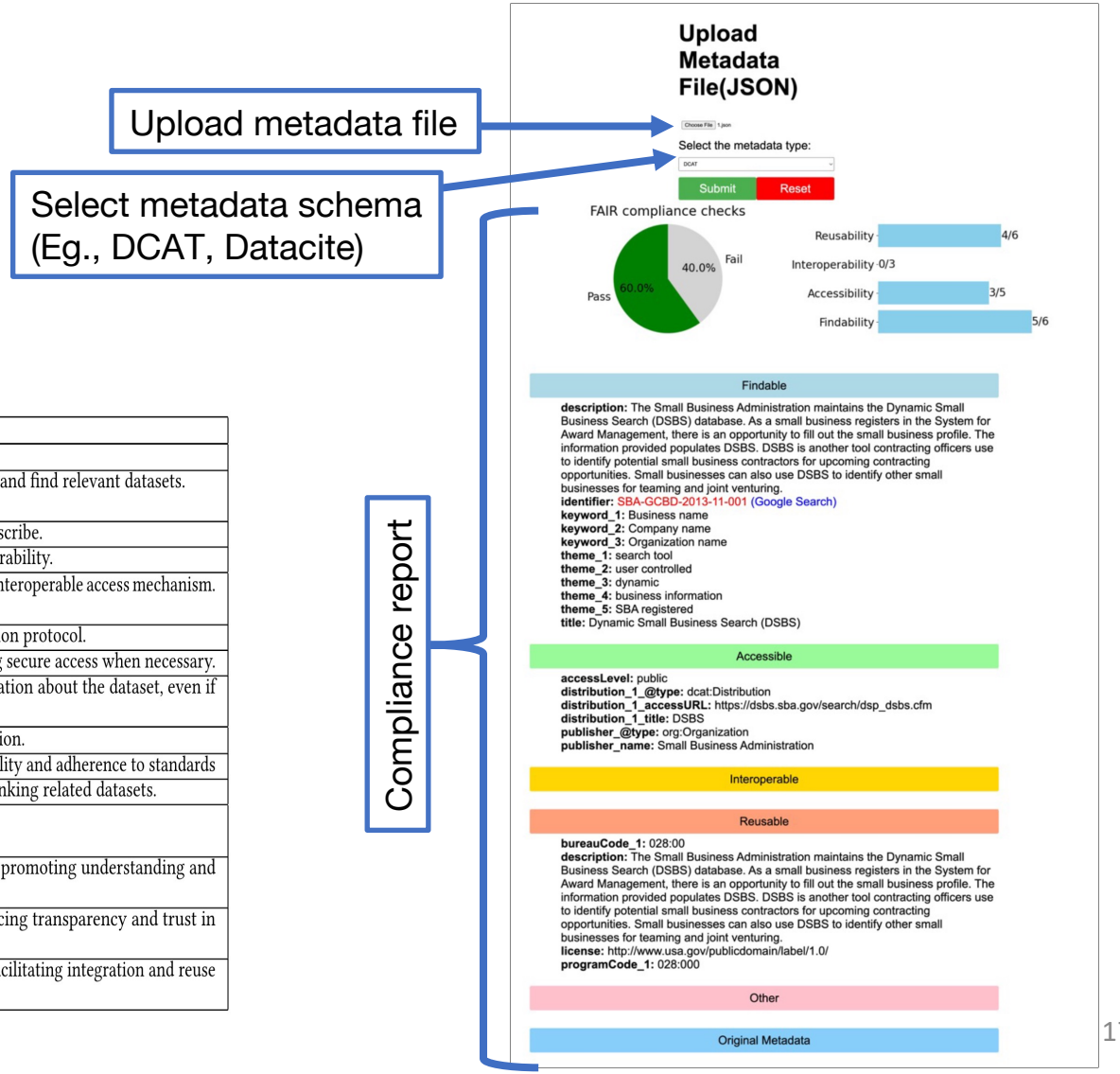
Output visualizations

Details of these plots are in the paper

FAIR principle compliance

- Findable, Available, Interoperable, and Reusable (FAIR) principles
- Evaluation by identifying keys available under each FAIR principle.

Principle	Subcategory	Elements	Reasoning
Findable	F1	identifier	Uniquely identifies the metadata, ensuring global uniqueness and persistence.
	F2	title, description, keyword, theme	Provide rich metadata, making it easier for users and computers to understand and find relevant datasets.
	F3	identifier	Reiterating the importance of explicitly linking metadata with the data they describe.
	F4	landingPage	Registering and indexing metadata in a searchable resource, enhancing discoverability.
Accessible	A1	distribution, downloadURL	Enable retrieval of data using a standardized protocol, providing a consistent and interoperable access mechanism.
	A1.1	format	Ensures openness, freedom, and universality in implementing the communication protocol.
	A1.2	accessLevel	Include information about authentication and authorization procedures, ensuring secure access when necessary.
	A2	publisher	The publisher information serves as a reference point for users seeking information about the dataset, even if the data is no longer accessible.
Interoperable	I1	format, conformsTo	Contribute to formal, accessible, and broadly applicable knowledge representation.
	I2	format, conformsTo	Support the use of vocabularies aligned with FAIR principles, ensuring compatibility and adherence to standards
	I3	references	Allows metadata to include qualified references, fostering interoperability by linking related datasets.
Reusable	R1	format, description, license	Contribute to rich metadata descriptions, supporting optimal reuse.
	R1.1	license	Ensures that data are released with clear and accessible licensing information, promoting understanding and adhering to usage terms.
	R1.2	programCode, bureauCode	Provides detailed information about the origin and history of the data, enhancing transparency and trust in data reuse.
	R1.3	conformsTo	Ensures that metadata and data follow domain-relevant community standards, facilitating integration and reuse within specific communities.

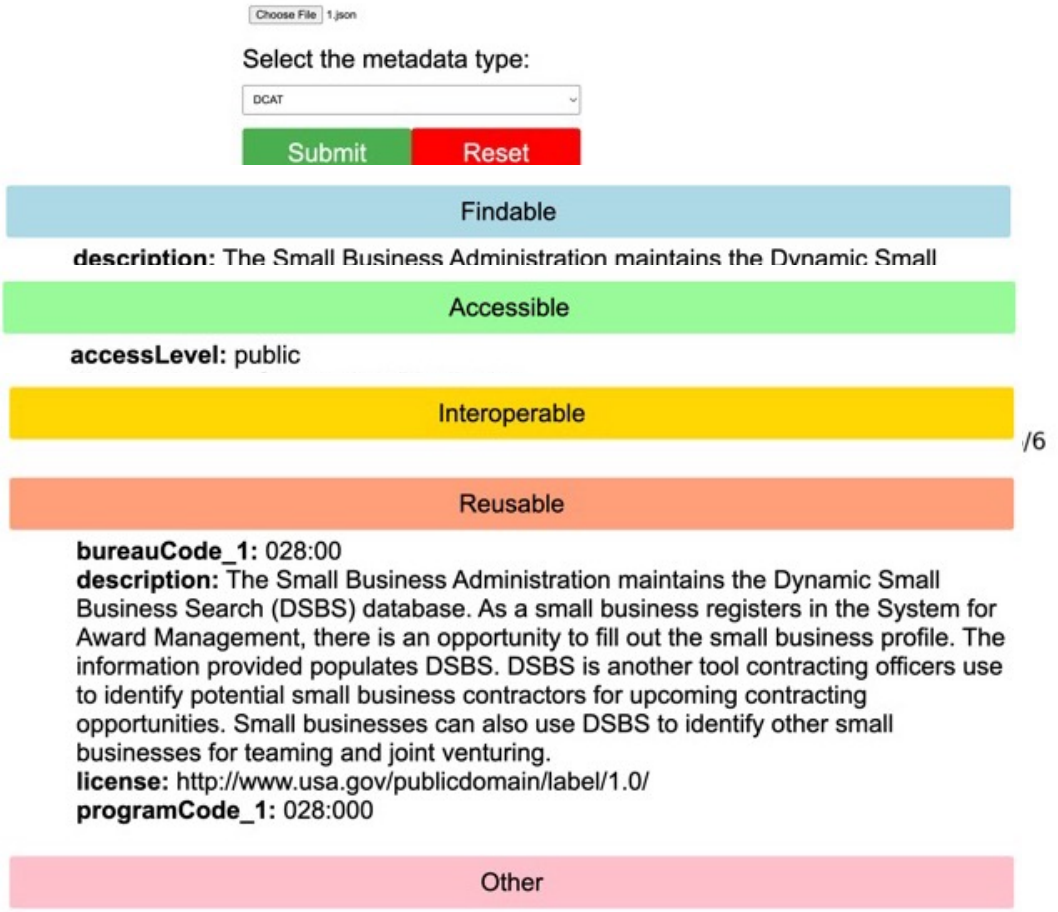


FAIR principle compliance

- Findable, Available, Interoperable, and Reusable (FAIR) principles
- Evaluation by identifying keys available under each FAIR principle.

DCAT FAIR Compliance Categorization

Principle	Subcategory	Elements	Reasoning
Findable	F1	identifier	Uniquely identifies the metadata, ensuring global uniqueness and persistence.
	F2	title, description, keyword, theme	Provide rich metadata, making it easier for users and computers to understand and find relevant datasets.
	F3	identifier	Reiterating the importance of explicitly linking metadata with the data they describe.
	F4	landingPage	Registering and indexing metadata in a searchable resource, enhancing discoverability.
Accessible	A1	distribution, downloadURL	Enable retrieval of data using a standardized protocol, providing a consistent and interoperable access mechanism.
	A1.1	format	Ensures openness, freedom, and universality in implementing the communication protocol.
	A1.2	accessLevel	Include information about authentication and authorization procedures, ensuring secure access when necessary.
	A2	publisher	The publisher information serves as a reference point for users seeking information about the dataset, even if the data is no longer accessible.
Interoperable	I1	format, conformsTo	Contribute to formal, accessible, and broadly applicable knowledge representation.
	I2	format, conformsTo	Support the use of vocabularies aligned with FAIR principles, ensuring compatibility and adherence to standards
	I3	references	Allows metadata to include qualified references, fostering interoperability by linking related datasets.
Reusable	R1	format, description, license	Contribute to rich metadata descriptions, supporting optimal reuse.
	R1.1	license	Ensures that data are released with clear and accessible licensing information, promoting understanding and adhering to usage terms.
	R1.2	programCode, bureauCode	Provides detailed information about the origin and history of the data, enhancing transparency and trust in data reuse.
	R1.3	conformsTo	Ensures that metadata and data follow domain-relevant community standards, facilitating integration and reuse within specific communities.



Evaluation

- Performance evaluation on a MacBook Pro laptop

Dataset	Features	# of Records	Completeness	Duplicates	Outliers	Fairness	Feature Relevance	Correlation Analysis	Privacy	Time(s)
Regensburg Pediatric Appendicitis	58	782	✓	✓	✓	✓	✓	✓	✓	0.8
Statlog German Credit	10	1K	✓	✓	✓	✓	✓	✓	✓	0.98
MIDRC Medical Cases	21	60K	✓	✓	✓	✓	✓	✓	✓	1.5
Single Elder Home Monitoring	10	416K	✓	✓	✓	N/A	✓	N/A	N/A	4.8
MetroPT-3	17	1.5M	✓	✓	✓	N/A	✓	N/A	N/A	5.3

R. Marcinkevičs et al. 2023. Regensburg Pediatric Appendicitis Dataset (1.01) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7669442
Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77
MIDRC. [n.d.]. The Medical Imaging and Data Resource Center (MIDRC). https: //www.midrc.org/
D. Marín López, D. Marín, J. Fonollosa, J. Llano, A. Perera, and Z. Haddi. 2023. Single Elder Home Monitoring: Gas and Position. UCI ML Repository.
N. Davari, B. Veloso, R. Ribeiro, and J. Gama. 2023. MetroPT-3 Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5VW3R

Evaluation - User study

- Conducted a user study involving a diverse group of participants:
  - Three AI researchers
  - Three PhD students
  - One postdoctoral scholar
  - Three computer / data management scientists
- Participants have experience in developing data management tools and AI algorithms.
- Insights from the study helped refine AIDRIN and expand its potential applications including federated learning.

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

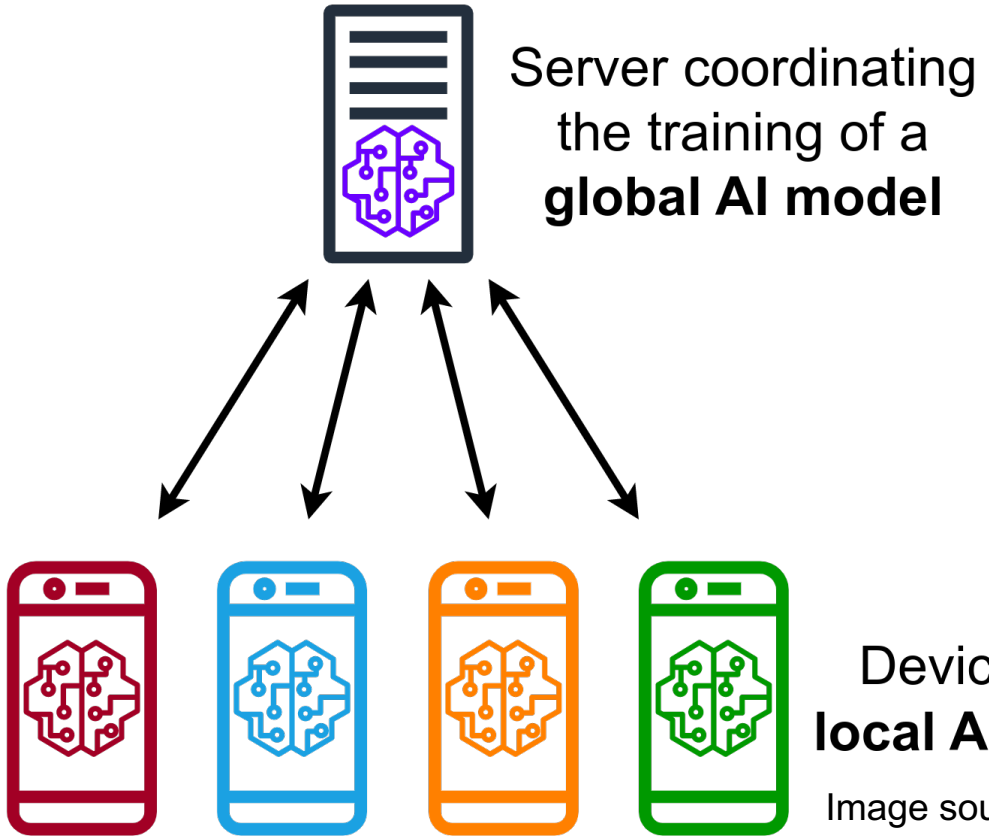
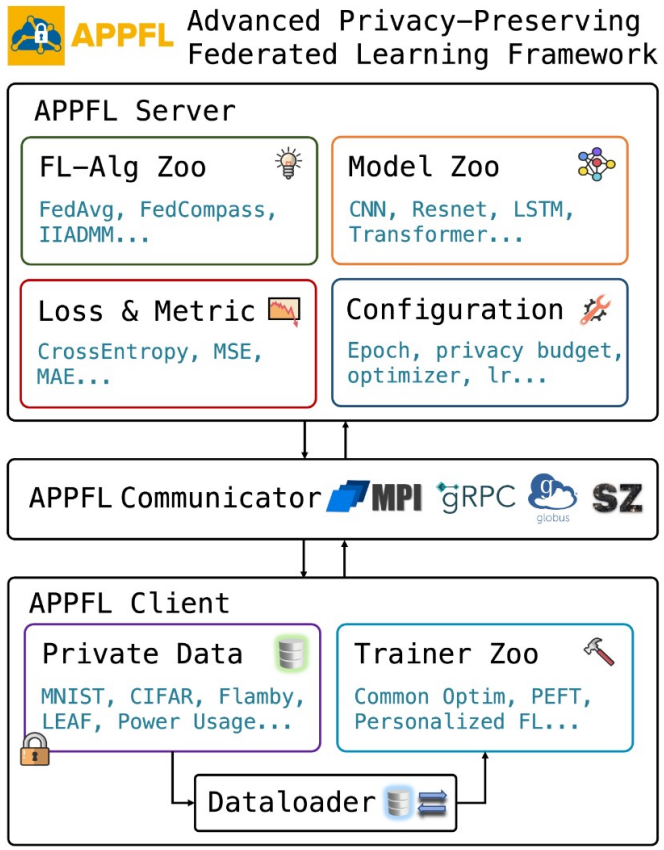
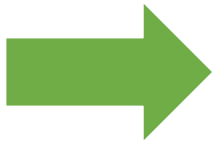


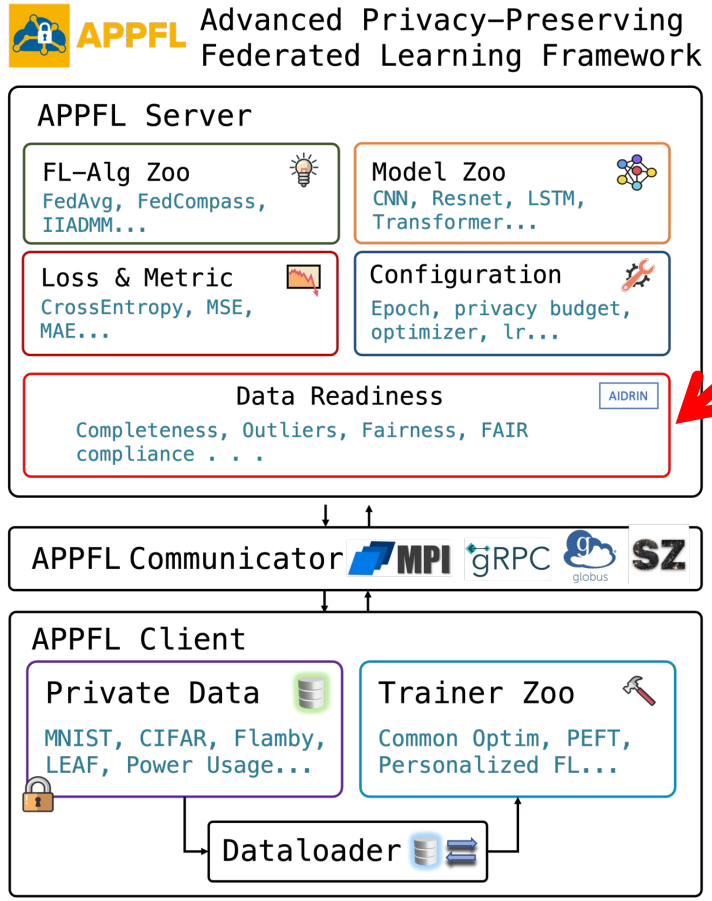
Image source: https://github.com/cs-joy/federated-learning

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

Integration of AIDRIN into APPFL







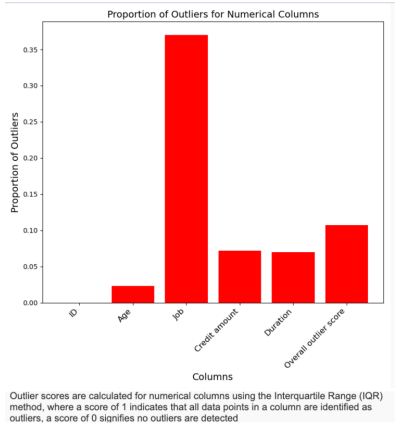
AIDRIN integration to study data characteristics at each site and impact on the model performance

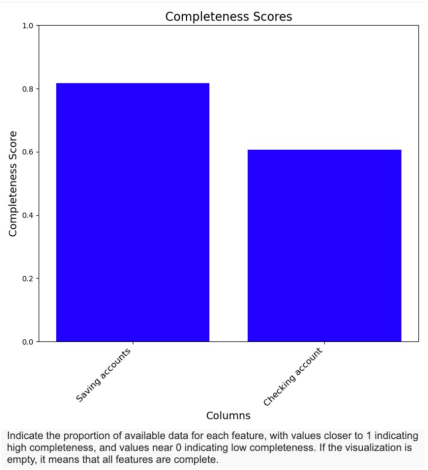
- Future work of AIDRIN privacy metrics on the client side

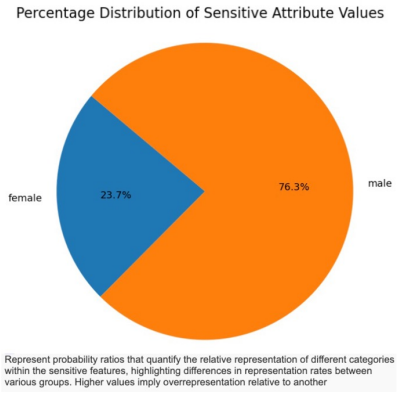
https://github.com/APPFL/APPFL/tree/main

Evaluation - Case study 1: SGC

- Statlog (German Credit Data – SGC) from UC Irvine
  - https://doi.org/10.24432/C5NC77
- Dataset analysis using the web interface
  - **Summary statistics:** Dimensions and detailed description of existing features.
  - **Completeness:** Two features at 80% and 60%, remaining features fully complete.
  - **Duplicates:** No duplicate records found.
  - **Outliers:** One feature with 35% outliers, others below 15%.
  - **Fairness:** Biased representation with over 75% males.
  - **Class imbalance:** "Purpose" attribute showing significant imbalance, most popular class at 33.7%, minority class at 1.2%, imbalance degree score of 4.49.
  - **Privacy risk:** Mean re-identification risk score of 0.45 for the 'Housing' feature.



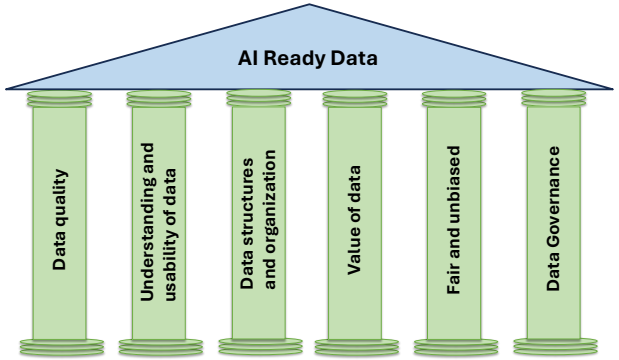




Evaluation - Case study 2: TCGA-LUAD

- The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) from the National Cancer Institute (NCI) Data Portal
  - https://www.cancerimagingarchive.net/collection/tcga-luad/
- Using Jupyter notebook
  - **Completeness:** Overall completeness at 56%,
    - Least complete column with 22% of completeness.
  - **Duplicates:** No duplicates ensuring unique patient information.
  - **Fairness:** Female representation at 65%, racial distribution with 87% Caucasian indicating potential bias.
  - **Class imbalance:** ID score of 0.27 for 'dead' and 'alive' class attributes.
  - **Discrimination risk:** TSD score of 0.01 for gender, 0.15 for racial groups (potential discrimination with racial groups).
  - **Insights and decisions:** Evaluation highlighted potential discrimination and class imbalance, providing crucial insights for informed decision-making.

Conclusion and future work



- AI-ready data is vital for confident AI-assisted decision-making
- Quantitative metrics and frameworks for measuring AI readiness of data are still evolving

- AIDRIN is a step towards developing a comprehensive framework for assessing AI data readiness

- AIDRIN is effective in federated learning; evaluates edge / client data
- AIDRIN is currently focused on structured / tabular data and reduced performance with large datasets

- Future work
  - A single AIDRIN Score that is meaningful to AI applications and to improve data → In progress
  - Integration of unstructured data assessment metrics → In testing
  - Studying the impact of data readiness on AI model performance → In development

Contact: byna.1@osu.edu