# Data-Blind ML

## Building privacy-aware machine learning models without direct data access

**Javier Pastorino** – Ashis Kumer Biswas

Machine Learning Laboratory

IEEE AIKE 2021 – *Virtual*

Department of Computer Science and Engineering
UNIVERSITY OF COLORADO
**DENVER | ANSCHUTZ MEDICAL CAMPUS**

# Agenda

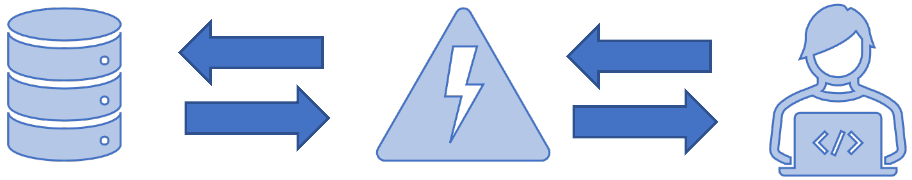
- Motivation
- Problem Description
- Related Work and Limitations
- Methodology
- Experiments – Datasets
- Results
- Limitations and Future Work

# Motivation

- ML Developers require data access to conduct analysis
- Lack of expertise/infrastructure triggers outsourcing
- Data may have privacy constraints
  - usually requires complex setups
- Data owners may have more control on how the data is accessed/used

# Problem Description

- **Building a framework allowing data owners outsource ML developments without sharing sensitive data.**
- Do not require ML Expertise for data owners



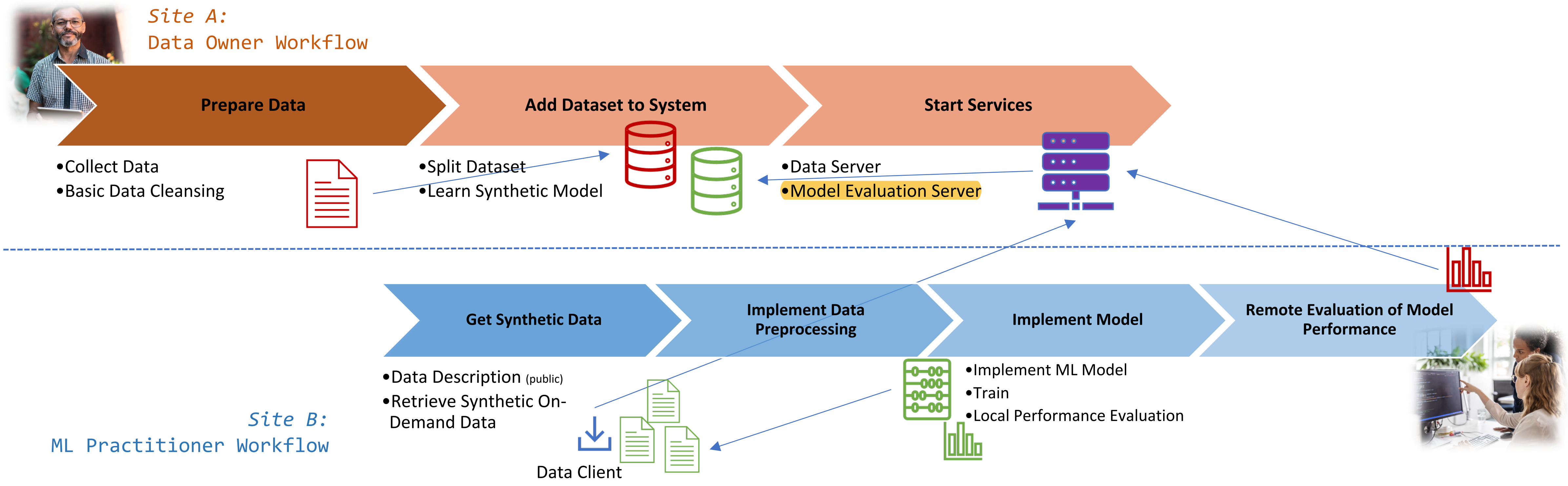
# Related Work and Limitations

- **Data Anonymization** (DA)
  - Data Sanitization
  - Large quantity of data anonymization algorithms [1-5]

- **Synthetic Data** (SD)
  - Generate synthetic, fake, data following same data distribution as real data[6, 8]
  - Use rigorous differential privacy solutions [7]

- **Limitations**
  - **DA**: no "gold-standard" to effectively anonymize data without risk of disclosure [6]
  - **SD**: techniques require extensive ML domain knowledge
  - There is no automatic toolset

# Methodology



Methodology

# Experimental Layout & Evaluation

- **Traditional pipeline development as baseline**
  - Using the entire real data, train and evaluate a model
  - **Real-trained model**
- **Develop pipeline using Data-blind ML**
  - Using synthetic data and following the framework methodology
  - developed and trained a model using synthetic data
  - Evaluate with real data using Data-Blind ML API
  - **Synthetic-trained model**
- **Evaluation**
  - Comparison based on accuracy difference between real-trained and synthetic-trained models

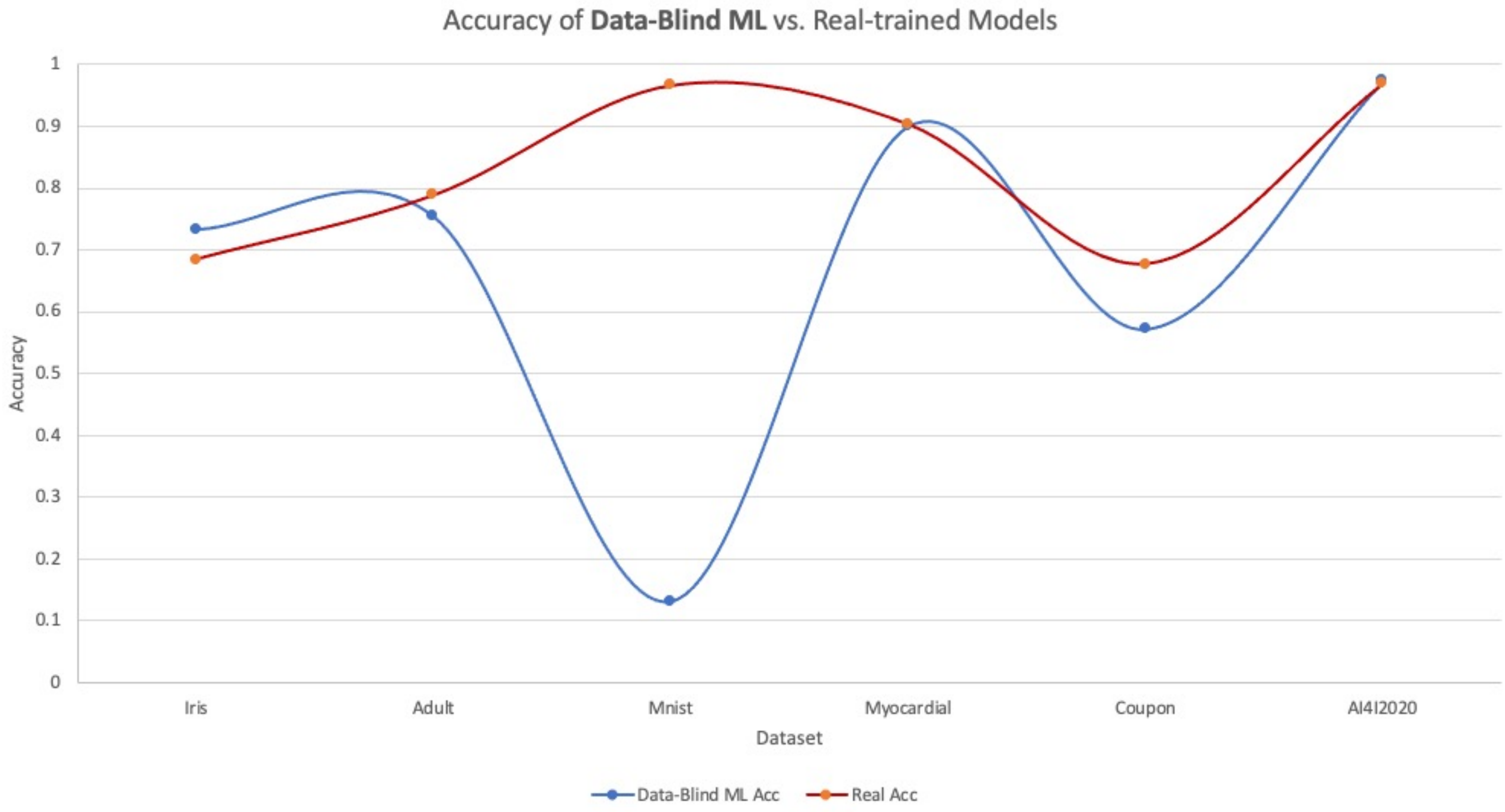
# Experiments – Datasets

| Dataset | | # Features | #Data Samples | Load Time | Synt. Model Samples 1,000 | 2,000 |
|---|---|---|---|---|---|---|
| Iris | | 5 | 150 | 0.05 s. | 3.84 s. | 3.84 s. |
| Adult | | 15 | 32,561 | 0.12 s. | 12.29 s. | 24.56 s. |
| MNIST | *tabular* | 785 | 42,000 | 2.11 s. | 154.34 s. | 276.77 s. |
| Myocardial | | 124 | 1,700 | 0.03 s. | 52.27 s. | 52.27 s. |
| Coupon | | 26 | 12,684 | 0.10 s. | 14.51 s. | 29.40 s. |
| AI4I2020 | | 13 | 10,000 | 0.07 s. | 16.06 s. | 33.98 s. |

- **Synthetic learning runtime:**
  - Linearity to the number of samples used
  - Proportional to the number of features in the training set.

# Results – Performance

| Dataset | Delta Accuracy |
|---|---|
| **Iris** | 0.0491 |
| **Adult** | −0.0336 |
| **MNIST** *tabular* | −0.8345 |
| **Myocardial** | −0.0036 |
| **Coupon** | −0.1049 |
| **AI4I2020** | 0.0034 |



# Limitations and Future Work

- **Limitations**
  - Data-Blind ML underperform on image data
    - Uses a CTGAN Core
  - Quality of the synthetic data is linked with the samples used generator trainings

- **Future Work**
  - Currently incorporating generative models for images data
  - Analyze the trade-off between generator sampling and data quality


Thank you!
Questions

# References


- [1] G. Ghinita, et al. “Fast data anonymization with low information loss,” in Proceedings of the 33rd international conference on Very large data bases, 2007
- [2] T. Li, et al.“Slicing:  A new approach for privacy preserving data publishing,” IEEE transactions on knowledge and data engineering, 2010
- [3] J. Xu, et al., “Utility-based anonymization using local recoding,” in Proceedings of the 12th ACM SIGKDD, 2006
- [4] H. Lee, et al., “Utility-preserving anonymization for health data publishing,” BMC medical informatics and decision making,2017
- [5] C. C. Aggarwal, “On k-anonymity and the curse of dimensionality,” in VLDB, vol. 5, 2005, pp.901–909.
- [6] N. C. Abay, et al., “Privacy preserving synthetic data release using deep learning,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2018
- [7] C. Dwork, et al., “The algorithmic foundations of differential privacy.” Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3-4, pp. 211–407, 2014
- [8] L. Xu, et al., “Modeling tabular data using conditional gan,” in Advances in Neural Information Processing Systems


https://github.com/jpastorino/Data-Blind-ML