2023 - AI - **E8E-55F** - Fairness Debugging of Tree-based Models using Machine Unlearning - tsurve@purdue.edu - Tanmay Surve

CERIAS
The Center for Education and Research in Information Assurance and Security

Fairness Debugging of Tree-based Models using Machine Unlearning

Student: *Tanmay Surve* **Mentor**: *Dr. Romila Pradhan*

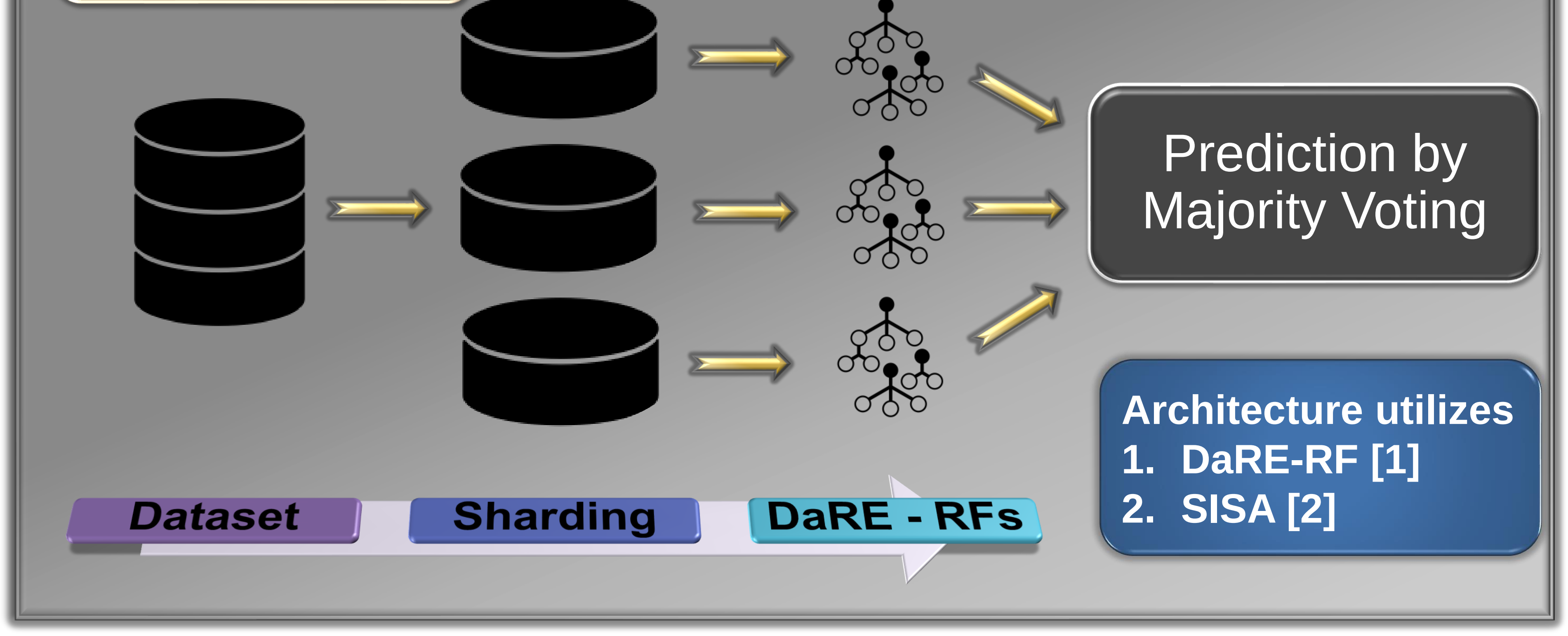
PURDUE POLYTECHNIC COMPUTER AND INFORMATION TECHNOLOGY

Motivation and Introduction

- Concern continues to mount that ML systems reinforce systemic biases and discrimination often reflected in their training data.
- There has not been much work on understanding and debugging tree-based classifiers in the context of fairness.
- Tree-based ML models, such as decision trees and random forests, are one of the most widely used ML models.
- Following machine unlearning advances are utilized to find fairness-based bias inducing subsets in the training dataset:

1. **DaRE – RF:** Data Removal-Enabled Random Forests [1]
2. **SISA:** Sharded, Isolated, Sliced, and Aggregated [2]

MODEL ARCHITECTURE



Prediction by Majority Voting

Architecture utilizes
1. DaRE-RF [1]
2. SISA [2]

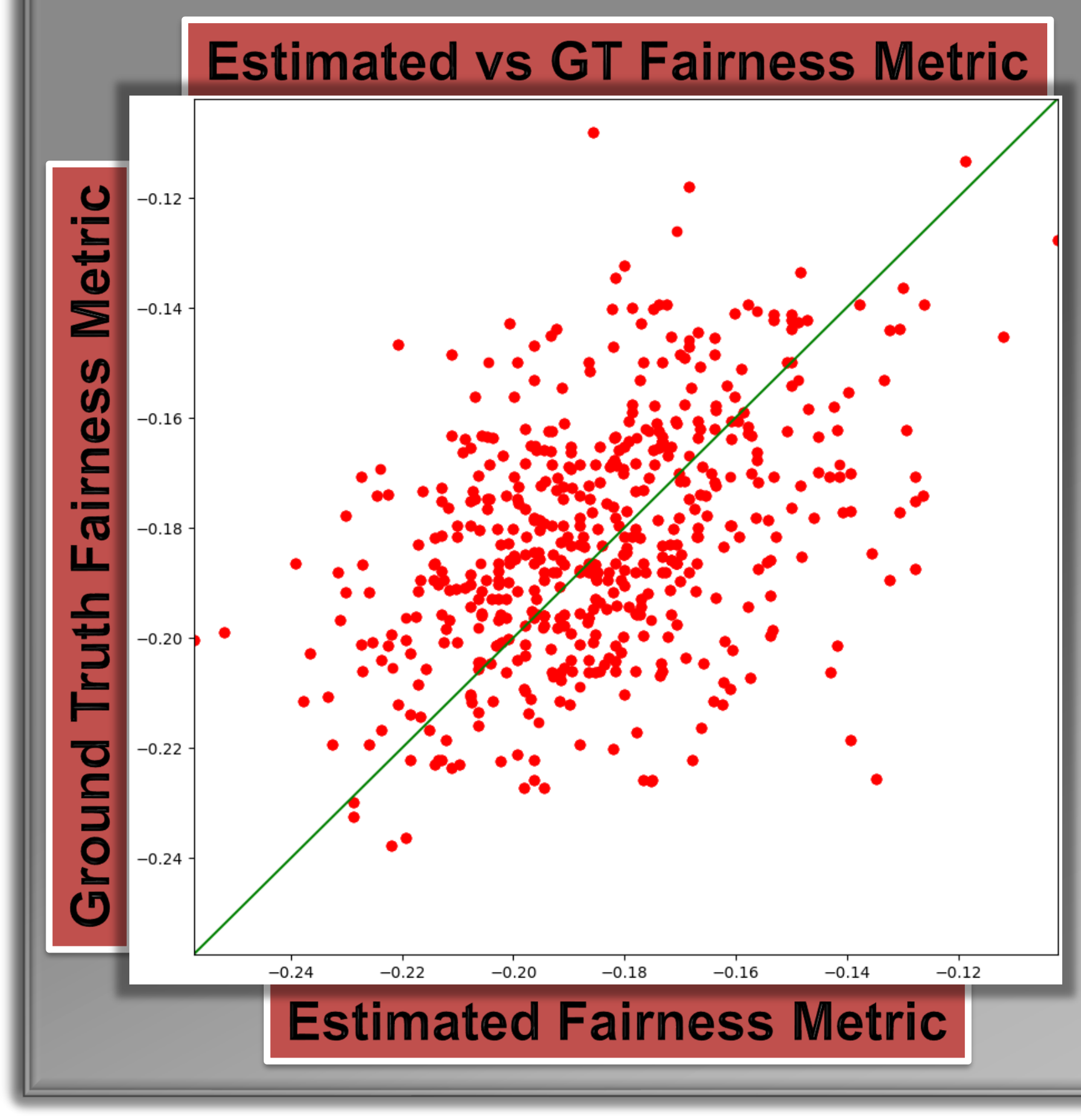
Experimental Setup

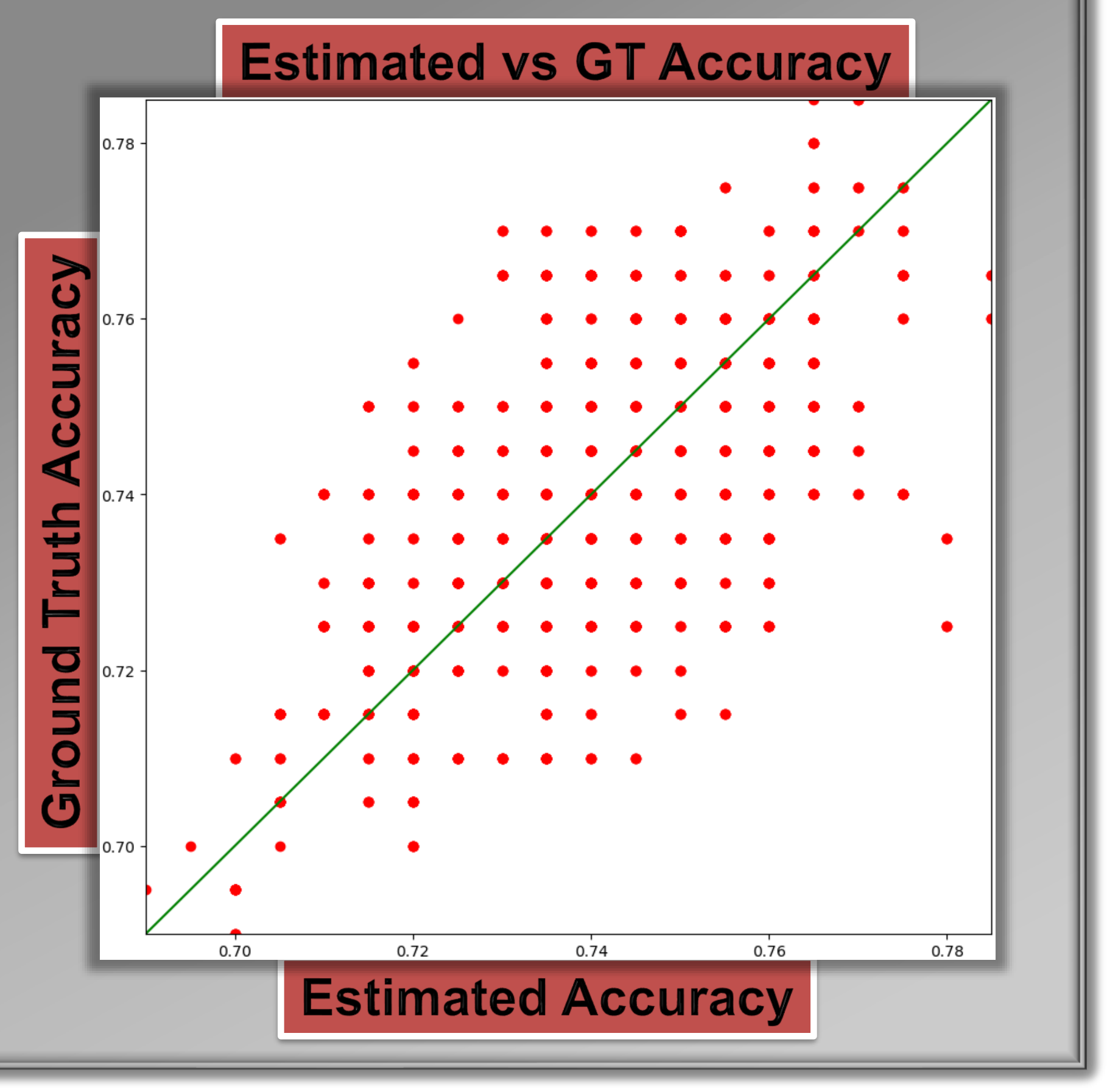
- German Credits dataset [3] (1000 data points, 20 features, sensitive attribute – *"age",* prediction task: binary classification , who should receive loan*).*
- Model – Random Forest Classifier (DaRE – RF [1] version).
- Fairness Metric – Predictive Parity [4].

EXPERIMENT RESULTS

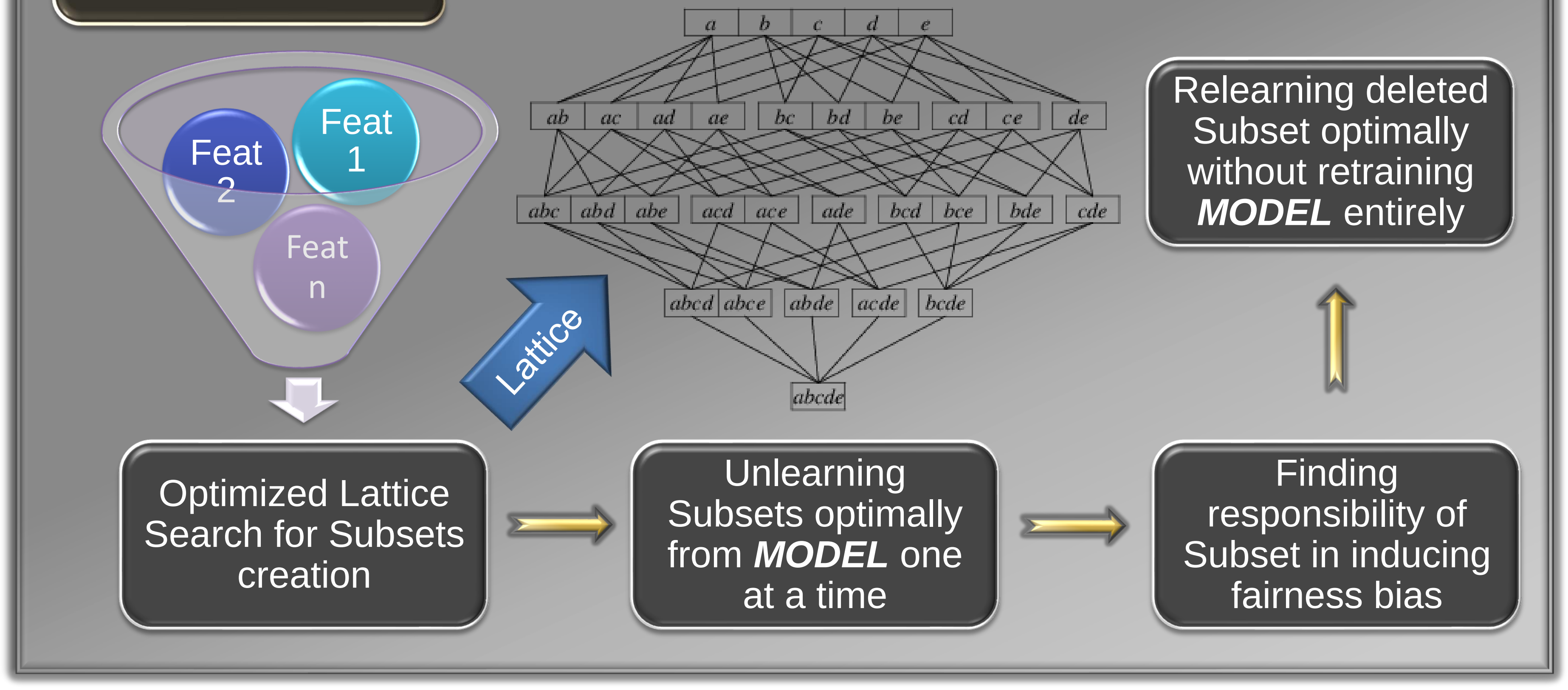
Subsets	Support	Bias Reduction	Accuracy Reduction
housing = rent	18.12 %	88.87 %	5.40 %
property = real estate	28.25 %	84.25 %	0.60%
cred_amt = high, *debtors* = none	21.87 %	79.78 %	3.37 %
age = young, *employment* = < 1 year	15.12 %	83.77 %	4.70 %
duration = high, *existing_creds* = high	17.25 %	84.60 %	4.05 %
foreign_worker = yes, *install_plans* = bank	13.75 %	76.19 %	4.72 %

Effectiveness of DaRE-RF


Estimated vs GT Fairness Metric


Estimated vs GT Accuracy

ALGORITHM FLOW



Optimized Lattice Search for Subsets creation

Unlearning Subsets optimally from ***MODEL*** one at a time

Finding responsibility of Subset in inducing fairness bias

Relearning deleted Subset optimally without retraining ***MODEL*** entirely

Conclusion and Future Works

- Fairness-based bias inducing subsets of training data (removing which has negligible accuracy loss) can be optimally found using techniques from machine unlearning.
- We plan to expand this study to other standard datasets and fairness metrics used in the fairness literature [4].

References

[1] *Brophy, J., & Lowd, D. (2021, July). Machine unlearning for random forests. In International Conference on Machine Learning (pp. 1092-1104). PMLR.*

[2] *Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021, May). Machine unlearning. In 2021 IEEE Symposium on Security and Privacy (SP) (pp. 141-159). IEEE.*

[3] *Dheeru Dua and Casey Graff. Uci machine learning repository, 2017.*

[4] *Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In Proceedings of the international workshop on software fairness (pp. 1-7).*

Acknowledgements

- This research was partially funded by a **Google Research Scholar Award**.