Reducing Certified Regression to Certified Classification

Zayd Hammoudeh, Daniel Lowd

zayd@cs.uoregon.edu

UNIVERSITY OF OREGON

Key Result: Voting-based certified classifiers become certified regressors when using **median** as the decision function

Overview

Problem: Most model predictions are brittle
- One outlier can change a model prediction *arbitrarily*

Certified Model: Provide a provable guarantee on a prediction's robustness
- ***R*** **– Certified robustness**

Certified Poisoning Classifier: Certifies **prediction** $\boldsymbol{f(x)}$ will not change under R arbitrary insertions or deletions in the training set

Reduction: A method for transforming a problem Q into another problem Q' is solvable efficiently

What is Certified Regression?

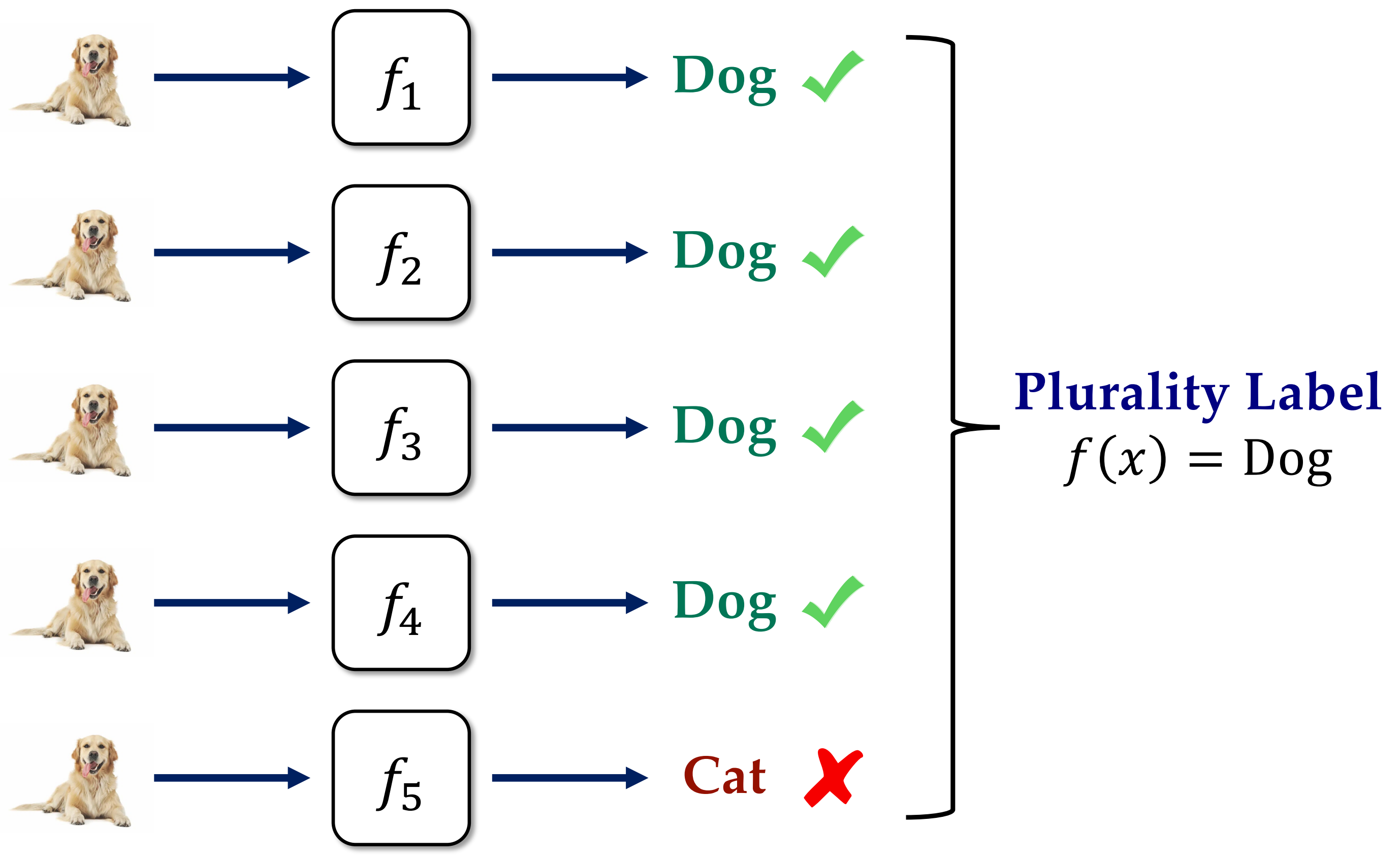
Goal: Certify the robustness that

$$\xi_l \leq f(x) \leq \xi_u$$

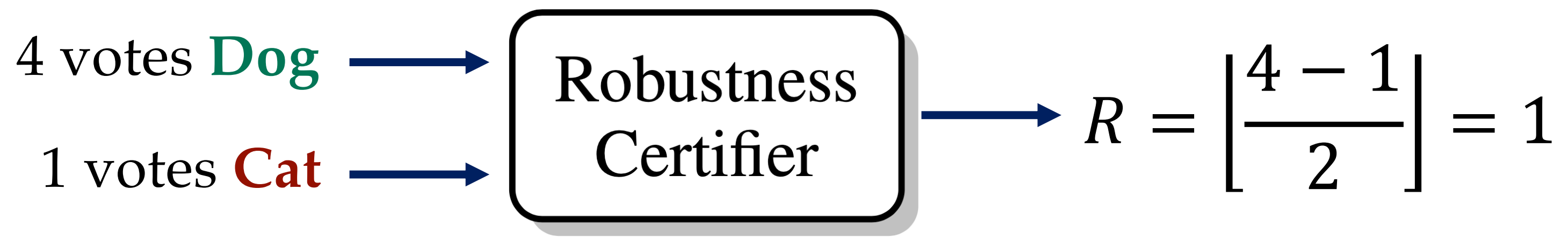
- $\xi_l, \xi_u \in \mathbb{R}$: User specified constants

Voting-Based Certified Classification

Example: Binary Classification Ensemble with Independent Submodels [LF21]



Robustness Certifier: Calculates certified robustness (R)
- Different for each certified classifier

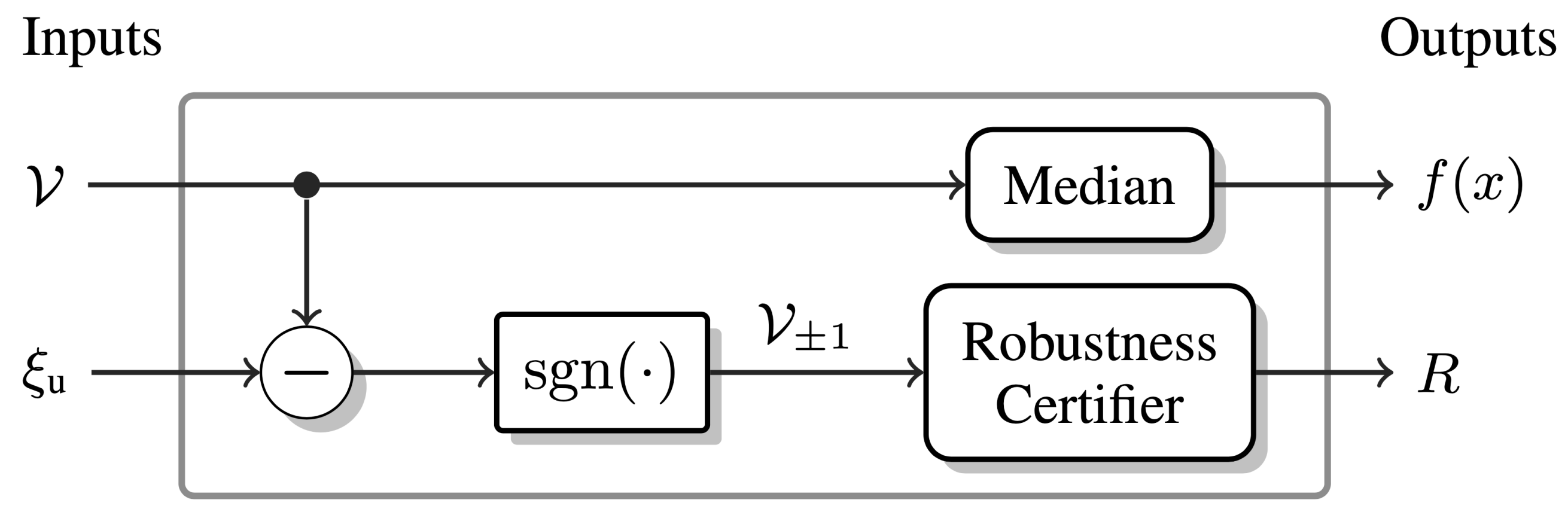


Don't Reinvent the Wheel

Our Solution: Reuse existing certified classifiers to certify regression

Key Insight: Certifying voting-based binary classifications is equivalent to certifying perturbation to a multiset's median

Regression to Classification Reduction



$\mathcal{V}$: Set of real-valued votes

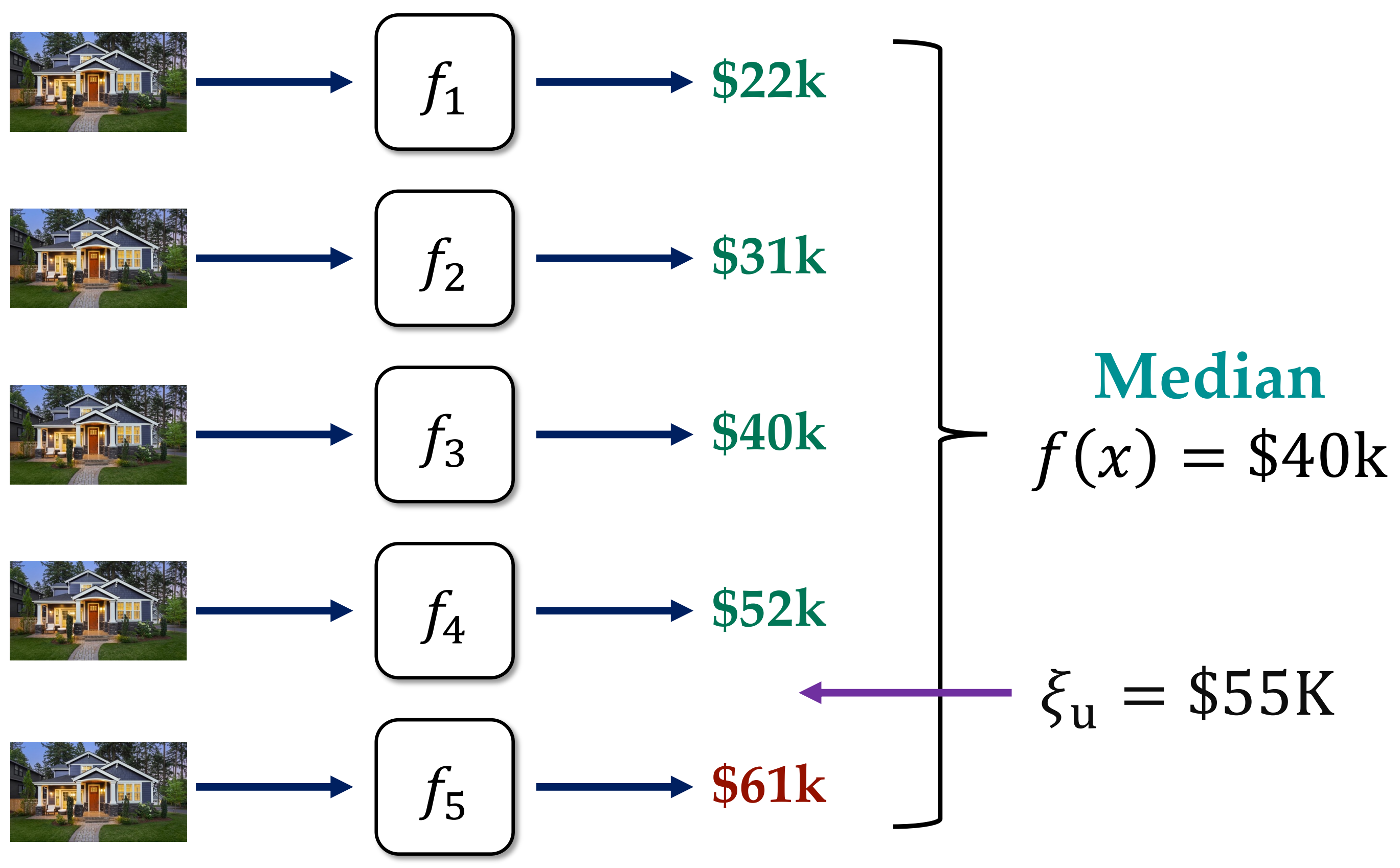
ξ_u: Upper threshold

$\mathcal{V}_{\pm 1}$: Binarized vote set where

$$\mathcal{V}_{\pm 1} \coloneqq \{\operatorname{sgn}(v - \xi_u) \colon v \in \mathcal{V}\}$$

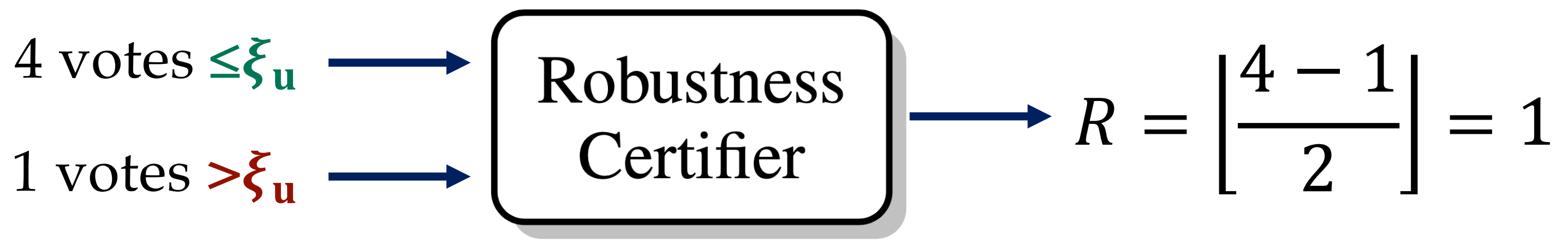
Partitioned Certified Regression

Example: Housing Price Prediction [De11]



Set of real-valued votes:

$$\mathcal{V} \coloneqq \{\$22K, \$31K, \$40K, \$52K, \$61K\}$$



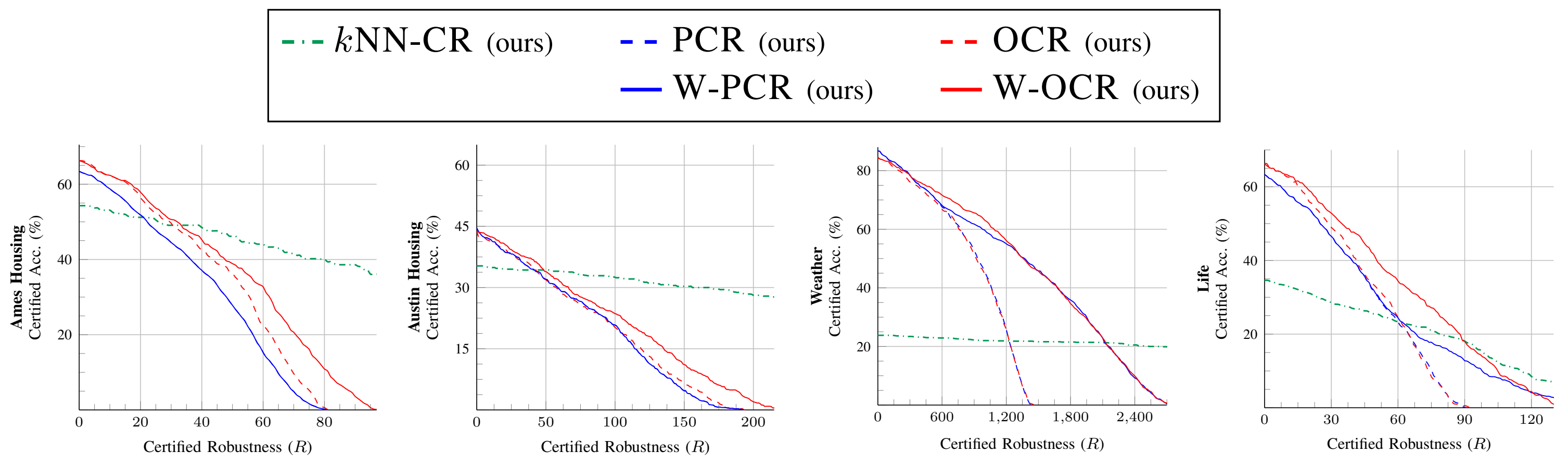
Six Proposed Certified Regressors

Two Based on Nearest-Neighbors Certified Classifiers [Jia+22]
- kNN Certified Regression
- rNN Certified Regression

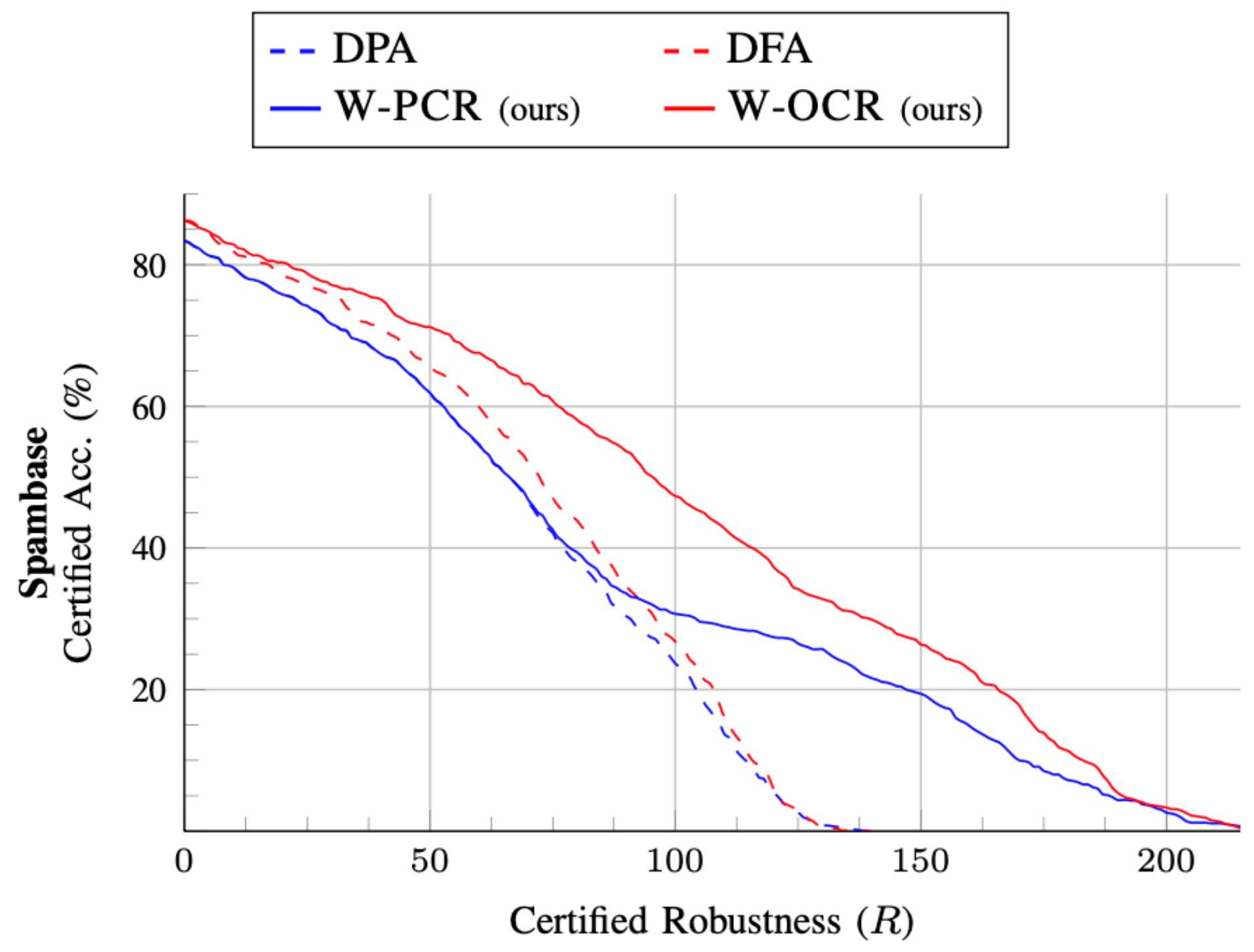
Two Based on Ensemble Certified Classifiers DPA [LF21] and DFA [WLF22]
- Partitioned Certified Regression (PCR)
- Overlapping Certified Regression (OCR)

Two Based on Novel Certified Classifiers
- Weighted Partitioned Certified Regression (W-PCR)
- Weighted Overlapping Certified Regression (W-OCR)

Empirical Evaluation



- Certify 50% of predictions up to 1% corruption
- Certify 30% of predictions up to 4% correction



State of the Art Certified Binary Classification

References

[De11] D. De Cock. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project". *Statistics Education* 2011.

[Jia+22] J. Jia, Y. Liu, X. Cao, and N. Gong. "Certified Robustness of Nearest Neighbors against Data Poisoning and Backdoor Attacks" AAAI, 2022.

[LF21] A. Levine and S. Feizi, "Deep partition aggregation: Provable defenses against general poisoning attacks," ICLR, 2021.

[WLF22] W. Wang, A. Levine, and S. Feizi "Improved Certified Defenses against Data Poisoning with (Deterministic) Finite Aggregation" ICML, 2022.