

Reducing Certified Regression to Certified Classification for General Poisoning Attacks

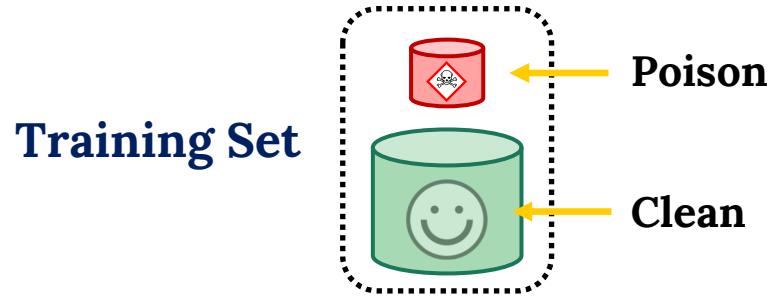
Zayd Hammoudeh Daniel Lowd

SaTML 2023 – Raleigh, NC

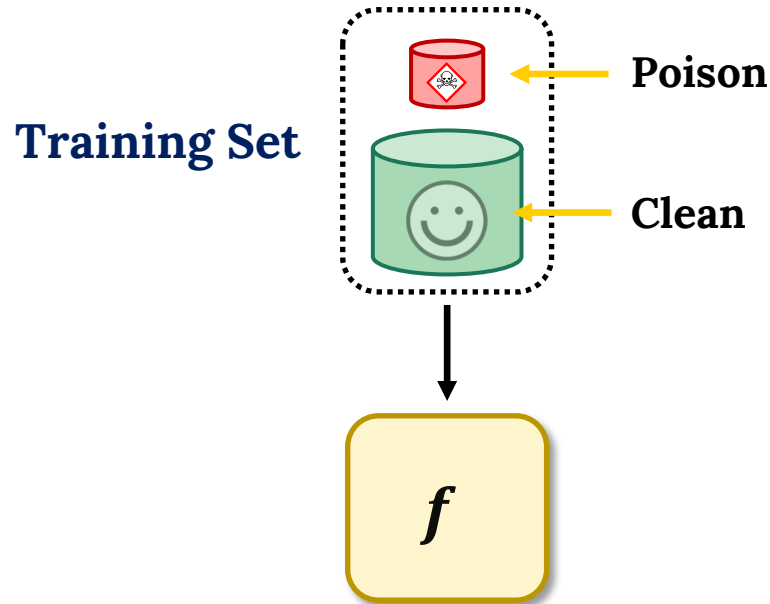


UNIVERSITY OF
OREGON

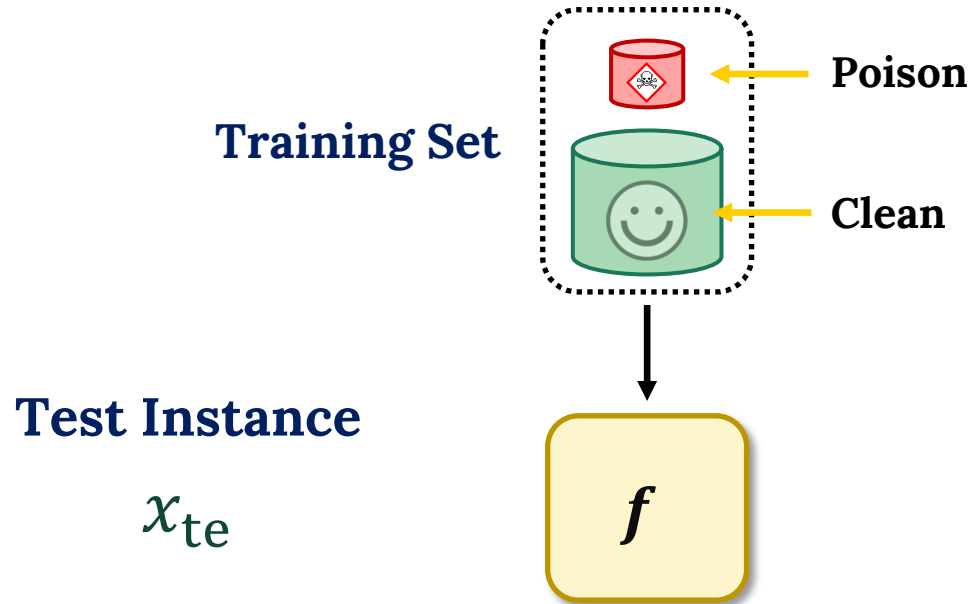
Data Poisoning Whirlwind Review



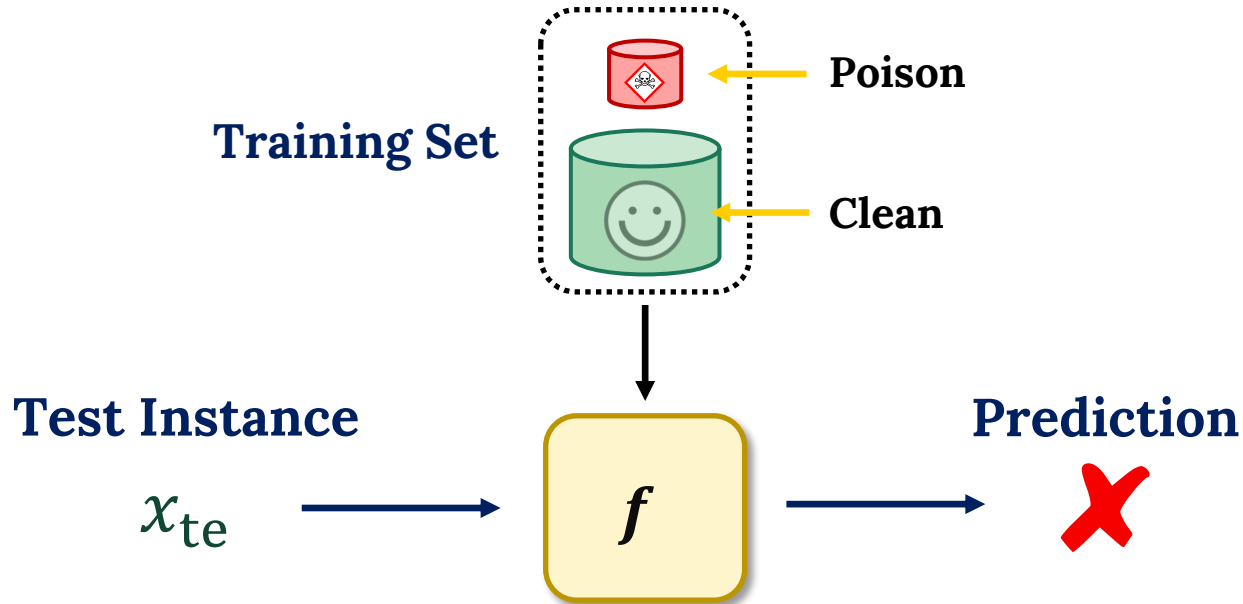
Data Poisoning Whirlwind Review



Data Poisoning Whirlwind Review



Data Poisoning Whirlwind Review



Certified Regression against Poisoning

Goal: Certify *pointwise* robustness R – the number of arbitrary instances that can be inserted or deleted from the training set with it guaranteed that:

$$\alpha \leq f(x_{te}) \leq \beta$$

- $\alpha, \beta \in \mathbb{R}$: User specified constants

Certified Regression against Poisoning

Goal: Certify *pointwise* robustness R – the number of arbitrary instances that can be inserted or deleted from the training set with it guaranteed that:

$$\alpha \leq f(x_{te}) \leq \beta$$

- $\alpha, \beta \in \mathbb{R}$: User specified constants

Structure of this Talk

Not the Focus: Our six certified regressors

Focus of this Talk: Our reduction

Specialized Robust Regressors Under Outliers & Poison

Session: Defense against Poisoning

Robust Linear

Chang
University of California
liuchang@eeecs.berkeley.edu

Yevgeniy Vorobeychik
Vanderbilt University
yevgeniy.vorobeychik@vanderbilt.edu

ABSTRACT

The effectiveness of supervised learning algorithms is ubiquitous in research and practice. In supervised learning commonly relies on minimizing the loss to improve performance and identify the best predicting outcomes. However, the enemy has made it a natural target for a training data, which we term poisoning. We deal with robust supervised learning about the nature of the feature matrix, and sub-Gaussian noise with low variance method for robust regression that assuming only that the feature matrix is by a low-rank matrix. Our technique uses low-rank matrix approximation and matrix regression, and yield strong performance. We experimentally show that our method outperform state-of-the-art robust regression prediction error.

1 INTRODUCTION

Machine learning has become widely used applications. An important class of applications is security defenses, such as spam analysis, and fraud detection [2, 8, 27]. In such of the machine learning system is crucial powerful adversaries, but strong increase in efficacy (e.g., to bypass spam filters).

An important factor in building a system is the availability of a collected samples. To achieve this, practitioners crowd-sourcing services, such as Amazon Mechanical Turk to collect training data sets.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that the copyright notice, this document title, the ACM copyright notice, and the date of publication are preserved. Copyright for components of this work that were reproduced from external sources must be honored. Abstracting with credit is permitted, but must be done with credit to the source. For more information, contact the ACM at www.acm.org/copyright.

To see the sensitivity of the samples x_i 's are samples provides us the inliers well. How closer to the positive multi-one and the label that LR is indeed frag shown via calculating

Robust

ABSTRACT

We consider propose a new method for robust regression that assuming only that the feature matrix is by a low-rank matrix. Our technique uses low-rank matrix approximation and matrix regression, and yield strong performance. We experimentally show that our method outperform state-of-the-art robust regression prediction error.

1 Introduction

Logistic regression is extensively used in many applications, such as spam analysis, and fraud detection [2, 8, 27]. In such of the machine learning system is crucial powerful adversaries, but strong increase in efficacy (e.g., to bypass spam filters).

An important factor in building a system is the availability of a collected samples. To achieve this, practitioners crowd-sourcing services, such as Amazon Mechanical Turk to collect training data sets.

Efficient A

Adam R.

We give the first polynomial-time algorithm for robust regression that assuming only that the feature matrix is by a low-rank matrix. Our technique uses low-rank matrix approximation and matrix regression, and yield strong performance. We experimentally show that our method outperform state-of-the-art robust regression prediction error.

arXiv:1805.11643v3 [cs.LG] 29 May 2019

arXiv:1803.03241v3 [cs.LG] 4 Jun 2020

High Dimensional

Liu Liu
liuliu@utexas.edu

We consider propose a new method for robust regression that assuming only that the feature matrix is by a low-rank matrix. Our technique uses low-rank matrix approximation and matrix regression, and yield strong performance. We experimentally show that our method outperform state-of-the-art robust regression prediction error.

1 Introduction

Learning in the presence of outliers has a long history in Robust Statistics. The high dimensional setting is essentially impossible via computationally efficient algorithms, and the first covariates and/or response outliers mean corruption. We assume that the data is from an uncorrupted distribution $x_i \sim \mathcal{N}(0, \Sigma)$ and $y_i \sim \mathcal{N}(\theta^T x_i)$. To model the corruption and replace them with adversarial samples (Definition 1.1), we propose to delete and corrupt samples independently. Outlier-robust regression is the low-dimensional setting, until recent breakthroughs in Mean Estimation [13, 33]. In the sparse setting, previous recovery guarant

Manipulating Inference and Countermeasures

Matthew Jagielski*, Alina Oprea
*Northeastern University, Boston, MA

Abstract—As machine learning becomes ubiquitous in research and practice, it is becoming a natural target for a training data, which we term poisoning. We deal with robust supervised learning about the nature of the feature matrix, and sub-Gaussian noise with low variance method for robust regression that assuming only that the feature matrix is by a low-rank matrix. Our technique uses low-rank matrix approximation and matrix regression, and yield strong performance. We experimentally show that our method outperform state-of-the-art robust regression prediction error.

arXiv:1804.00308v3 [cs.CR] 28 Sep 2021

Robust High Dimensional Sparse Regression and Matching Pursuit

Yudong Chen, Constantine Caramanis and Shie Mannor

Abstract

In this paper we consider high dimensional sparse regression, and develop strategies able to deal with arbitrary—possibly, severe or coordinated—errors in the covariance matrix X . These may come from corrupted data, persistent experimental errors, or malicious respondents in surveys/recommender systems, etc. Such non-stochastic error-in-variables problems are notoriously difficult to treat, and as we demonstrate, the problem is particularly pronounced in high-dimensional settings where the primary goal is support recovery of the sparse regressor. We develop algorithms for support recovery in sparse regression, when some number n_1 out of n total covariate/response pairs are arbitrarily (possibly maliciously) corrupted. We are interested in understanding how many outliers, n_1 , we can tolerate, while identifying the correct support. To the best of our knowledge, neither standard outlier rejection techniques, nor recently developed robust regression algorithms (that focus only on corrupted response variables), nor recent algorithms for dealing with stochastic noise or erasures, can provide guarantees on support recovery. Perhaps surprisingly, we also show that the natural brute force algorithm that searches over all subsets of n covariate/response pairs, and all subsets of possible support coordinates in order to minimize regression error, is remarkably poor, unable to correctly identify the support with even $n_1 = O(n/k)$ corrupted points, where k is the sparsity. This is true even in the basic setting we consider, where all authentic measurements and noise are independent and sub-Gaussian. In this setting, we provide a simple algorithm—no more computationally taxing than OMP—that gives stronger performance guarantees, recovering the support with up to $n_1 = O(n/(\sqrt{k} \log p))$ corrupted points, where p is the dimension of the signal to be recovered.

I. INTRODUCTION

Linear regression and sparse linear regression seek to express a response variable as the linear combination of (a small number of) covariates. They form one of the most basic procedures in statistics, engineering, and science. More recently, regression has found increasing applications in the high-dimensional regime, where the number of variables, p , is much larger than the number of measurements or observations, n . Applications in biology, genetics, as well as in social networks, human behavior prediction and recommendation, abound, to name just a few. The key structural property exploited in high-dimensional regression, is that the regressor is often sparse, or near sparse, and as much recent research has demonstrated, in many cases it can be efficiently recovered, despite the grossly underdetermined nature of the problem (e.g., [8], [6], [4], [12], [31]). Another common theme in large-scale learning problems—particularly problems in the high-dimensional regime—is that we not only have big data, but we have dirty data. Recently, attention has focused on the setting where the output (or response) variable and the matrix of covariates are plagued by erasures, and/or by stochastic additive noise [23], [26], [27], [9], [10]. Yet many applications, including those mentioned, may suffer from persistent errors, that are ill-modeled by stochastic distributions; indeed, many applications, particularly those modeling human behavior, may exhibit maliciously corrupted data.

This paper is about extending the power of regression, and in particular, sparse high-dimensional regression, to be robust to this type of noise. We call this *deterministic or cardinality constrained robustness*, because rather than restricting the magnitude of the noise, or any other such property of the noise, we merely assume there is a bound on how many data points, or how many coordinates of every single covariate, are corrupted. Other than this number, we make absolutely no assumptions on what the adversary can do—the adversary is virtually unlimited in computational power and knowledge about the data, and about the robustness of the algorithms we are using. We show that, with

*Preprint of the work accepted for publication in Security and Privacy, San Francisco, CA
†Note: a prior version of this paper had the TRIM algorithm, leading to (non)convergent TRIM will significantly outperform prior algorithms.

*UT Austin klivans@cs.utexas.edu
†Princeton University and University of California, Berkeley

Adversarially Robust Regressors Make **Strong Assumptions**

Data Distribution Assumptions

- Sparsity/low rank
- Linear data distribution with AWGN

Model architecture assumptions

- Linear model

Distributional Guarantees Only

- No insight into individual predictions' robustness

Our Goal

Provably robust regressors that are **general**:

- ◎ No data distribution assumptions
- ◎ Model architecture agnostic
- ◎ Stop reinventing the wheel.
 - **Consistently** state-of-the-art with **minimal effort**

A Bit of a Detour



UNIVERSITY OF
OREGON

Certified Poisoning **Classifiers** Show Promise

Certified Poisoning **Classifiers** Show Promise

arXiv:2205.13176v2 [cs.CR] 6 Sep 2022

On Colle
DEEP P
PROVAE
ATTACK
Ab
Bootstrap aggregating ensemble protocol, which robustness by its major cent works further prove ness certificates for cert partition aggregation), forms, in this paper, w the tight robustness ag ing attack. Specifica num number of simul tions via solving a bin ming (BLLP) problem. robustness of vanilla ba bound of the tolerable this analysis, we prop prove the robustness of free. This is achieved subsampling in vanilla deterministic subsamp ling the influence scop ple universally. Our co the notable advantage and robustness. Our co //github.com/Em

1. Introduction
Bagging (Breiman, 1996), protocol that trains sub-cla trainees and makes predic is a commonly used metho works (Biggio et al., 2014, 2021) show its superior ce data poisoning attacks. Mo

Department of Computer S Key Lab of Artificial Intellig Shanghai, China. Tie Li and AI Laboratory, Shanghai, Chi lijies@sjtu.edu.cn.

Proceedings of the 39th Inter Learning, Baltimore, Maryland right 2022 by the author(s).

Published as a conference paper at ICLR 2021

DEEP P
PROVAE
ATTACK

Alexander
Department
University of
College Park
(alex@umd.edu)

Adv
beh
samp
the
move
DP
the
by
it
num
of
creat
a
vot
is
a
algori
thm
intrin
sic
mecha
nism
poison
ing
CIFAR
10
these
prot
for
futu

1. INTRO
Adversarial p attacks, an ad sifications of the literature, et al., 2017), of the training s a poisoning at on the magni classification. **General poi** number of sa cardinally of

arXiv:2012.03765v3 [cs.CR] 2 Dec 2021

Certif

1. Intro
Data poi soning (Chen et al., 2019; Steinhardt et al., 2019) is a system via car some training classifier make corrupted down certain trigger machine learni cybersecurity analytics (Moz et al., 2019). Re recently devel and backdoor accuracy on a a threshold (e bound of the examples with

arXiv:2202.02628v3 [cs.LG] 14 Jul 2022

In AAAI Conference on Artificial Intelligence, 2022.

Improved Certified Defenses against Data Poisoning with (Deterministic) Finite Aggregation

Wenzhao Wang¹ Alexander Levine¹ Soheil Feizi¹

Abstract

Data poisoning attacks aim at manipulating model behaviors through distorting training data. Previously, an aggregation-based certified defense, Deep Partition Aggregation (DPA), was proposed to mitigate this threat. DPA predicts through an aggregation of base classifiers trained on disjoint subsets of data, thus restricting its sensitivity to dataset distortions. In this work, we propose an improved certified defense against general poisoning attacks, namely **Finite Aggregation**. In contrast to DPA, which directly splits the training set into *disjoint* subsets, our method first splits the training set into smaller disjoint subsets and then combines duplicates of them to build larger (but not disjoint) subsets for training base classifiers. This reduces the worst-case impacts of poison samples and thus improves certified robustness bounds. In addition, we offer an alternative view of our method, bridging the designs of deterministic and stochastic aggregation-based certified defenses. Empirically, our proposed Finite Aggregation consistently improves certificates on MNIST, CIFAR-10, and GTSRB, boosting certified fractions by up to 3.05%, 3.97% and 4.77%, respectively, while keeping the same clean accuracies as DPAs, effectively establishing a new **state of the art** in (pointwise) certified robustness against data poisoning.

1. Introduction

Over the past years, we have witnessed the increasing popularity of deep learning in a variety of domains including computer vision (He et al., 2016), natural language process

¹Department of Computer Science, University of Maryland, College Park, Maryland, USA. Correspondence: Wenzhao.Wang<www@umd.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

ing (Devlin et al., 2019), and speech recognition (Xiong et al., 2016). In many cases, such rapid developments depend heavily on the increased availability of data collected from diverse sources, which can be different users or simply websites from all over the Internet. While the richness of data sources greatly facilitates the advancement of deep learning techniques and their applications, it also raises concerns about their *reliability*. This makes the data poisoning threat model, which concerns the reliability of models under adversarially corrupted training samples, more important than ever (Goldblum et al., 2020).

In this work, we use a general formulation of data poisoning attacks as follows: The adversary is given the ability to insert/remove a bounded number of training samples in order to manipulate the predictions (on some target samples) of the model trained from the corresponding training set. Here, the number of samples that the adversary is allowed to insert/remove is referred to as the attack size.

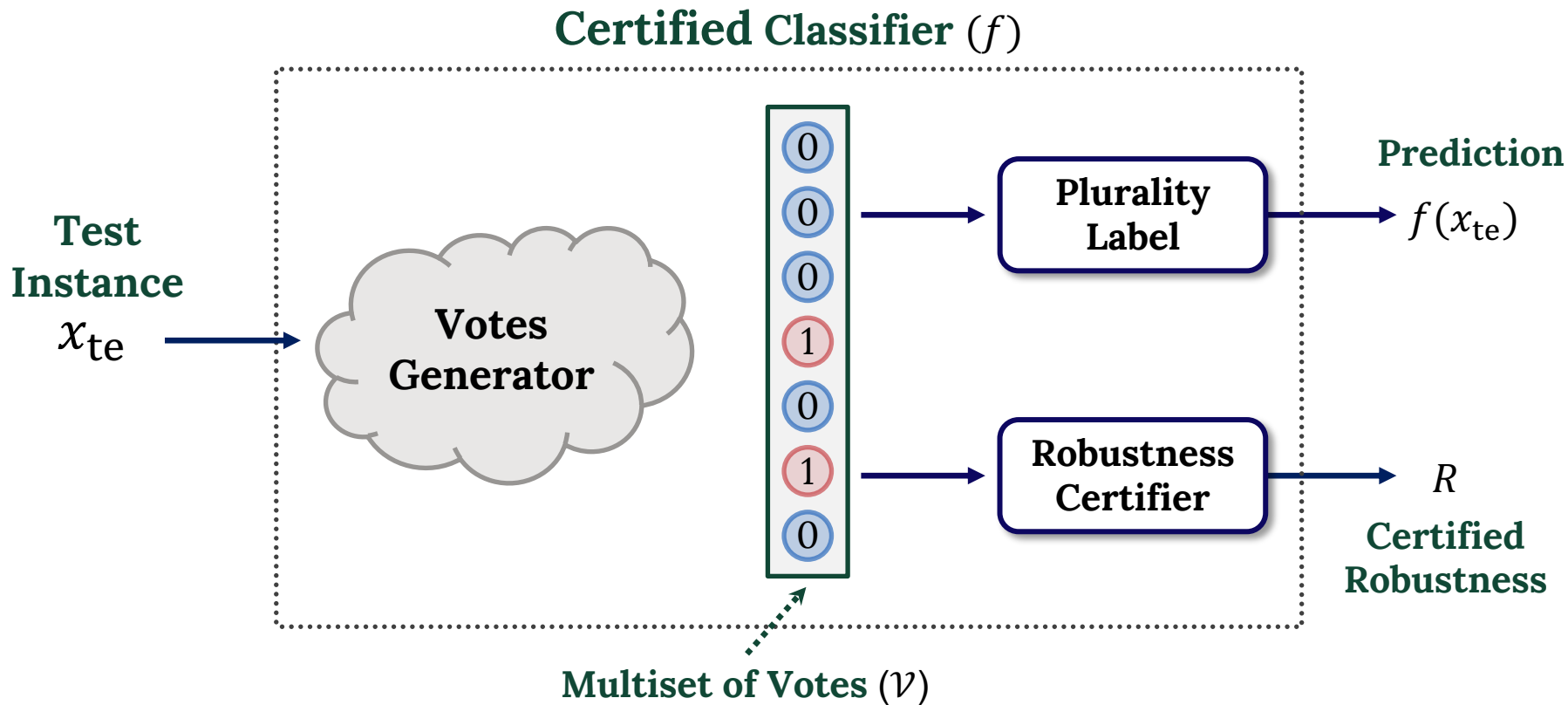
Many variants of empirical poisoning attacks targeting deep neural networks have been proposed, including Feature Collision (Shafiqi et al., 2018), Convex Polytope (Zhu et al., 2019), Bulyte Polytope (Aghakhani et al., 2021) and Witches' Brew (Geiping et al., 2021). These attacks are also referred to as triggerless attacks since no modification to the targets is required. Unlike triggerless attacks, backdoor attacks are poisoning attacks that allow modifications of the target samples, for which a variety of approaches have been developed including backdoor poisoning (Chen et al., 2017), label-consistent backdoor (Turner et al., 2019) and hidden-trigger backdoor (Saha et al., 2020). While it is shown in (Schwarzshild et al., 2021) that the evaluation settings can greatly affect the success rate of many data poisoning attacks to deep models, the vulnerability issues against poisoning attacks remain because (i) the current attacks can still succeed in many scenarios, and (ii) stronger adaptive poisoning attacks can potentially be developed in the future, posing practical threats.

In this work, we focus on developing *provably* robust defenses against general poisoning attacks. In particular, aggregation-based techniques, including a deterministic one (Levine & Feizi, 2021) and stochastic ones (Jia et al., 2021; Chen et al., 2020), have been adopted to offer (pointwise)

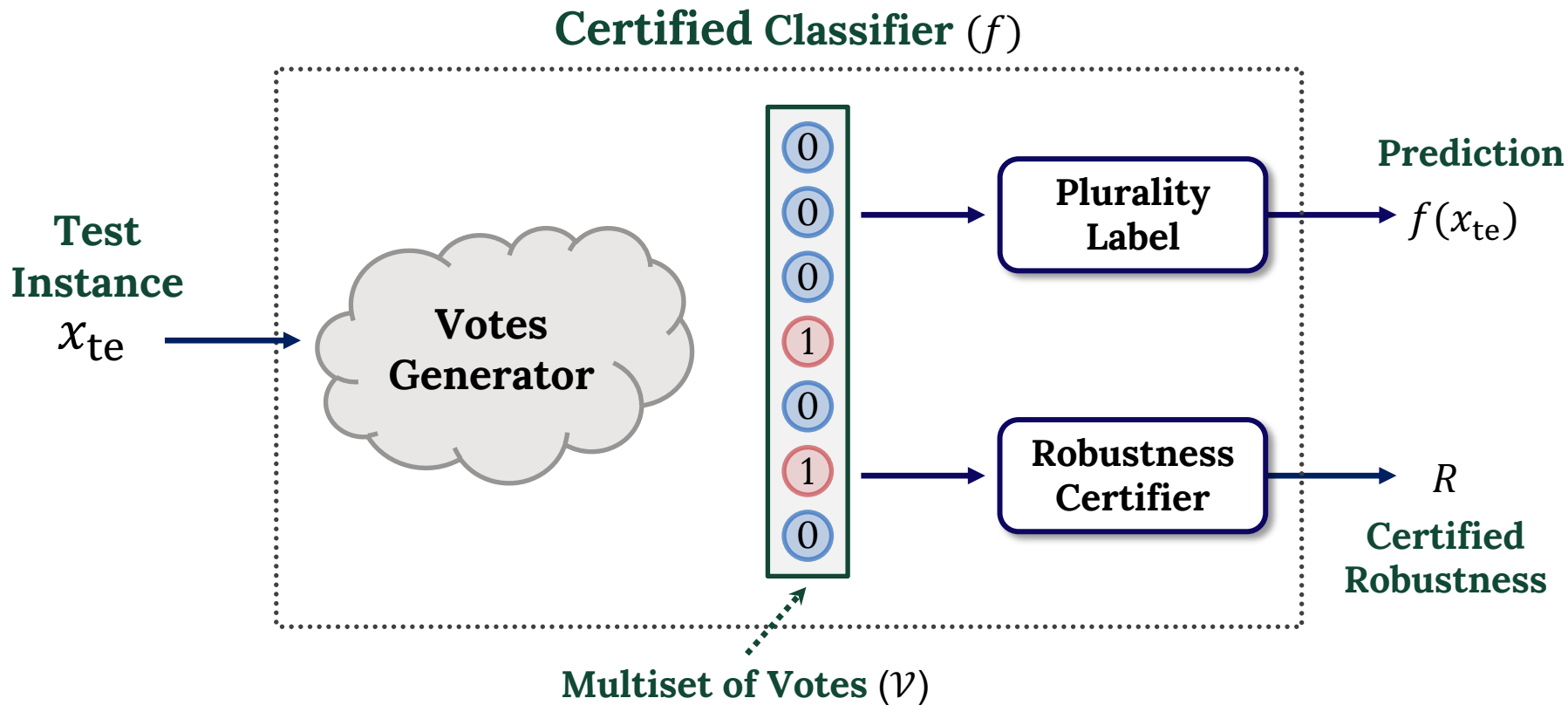
Strengths of Certified Poisoning **Classifiers**

1. No data distribution assumptions
2. Model architecture agnostic
3. Strong empirical performance
 - Certify 65% of MNIST predictions up to 0.8% arbitrary poison
 - Certify 16% of CIFAR10 predictions up to 0.1% arbitrary poison

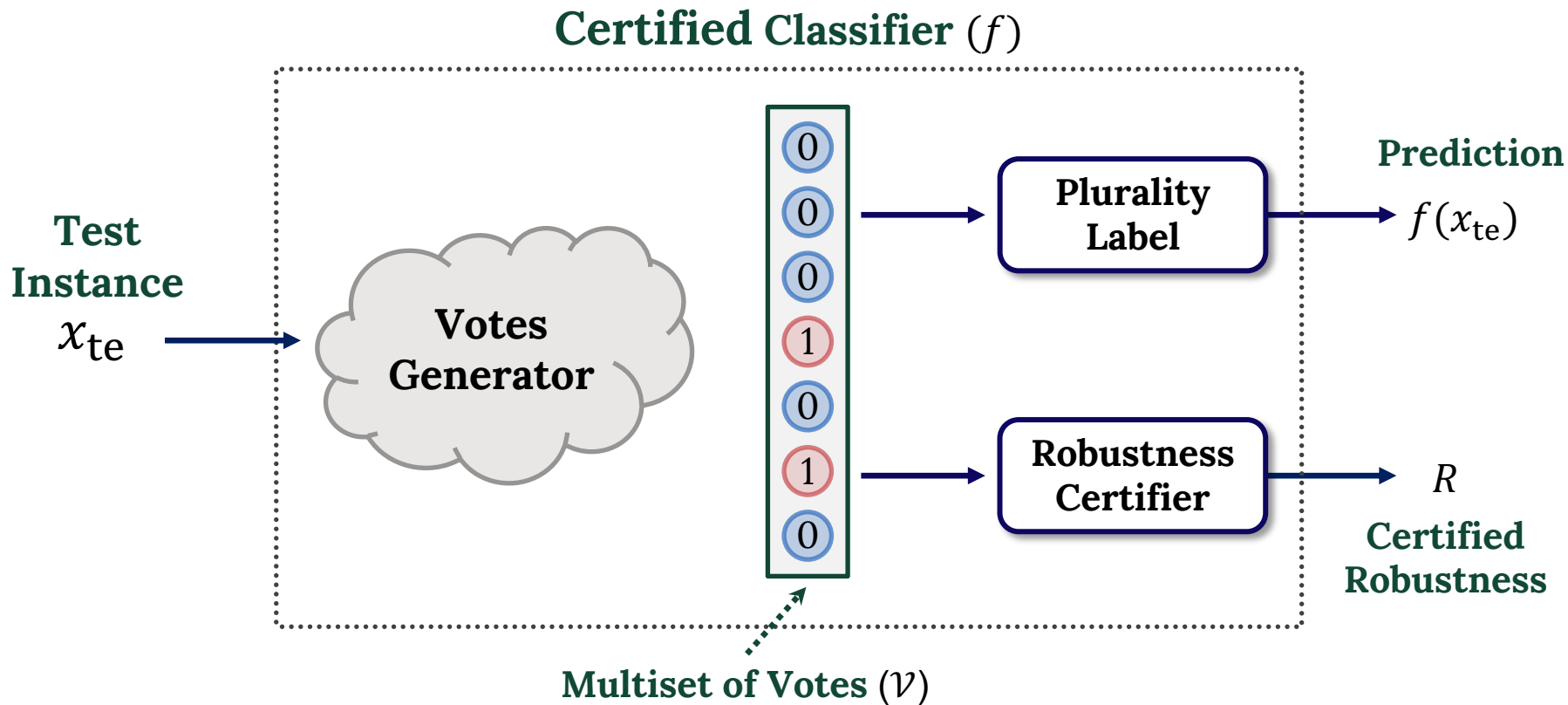
General Structure of a Certified Poisoning Classifier



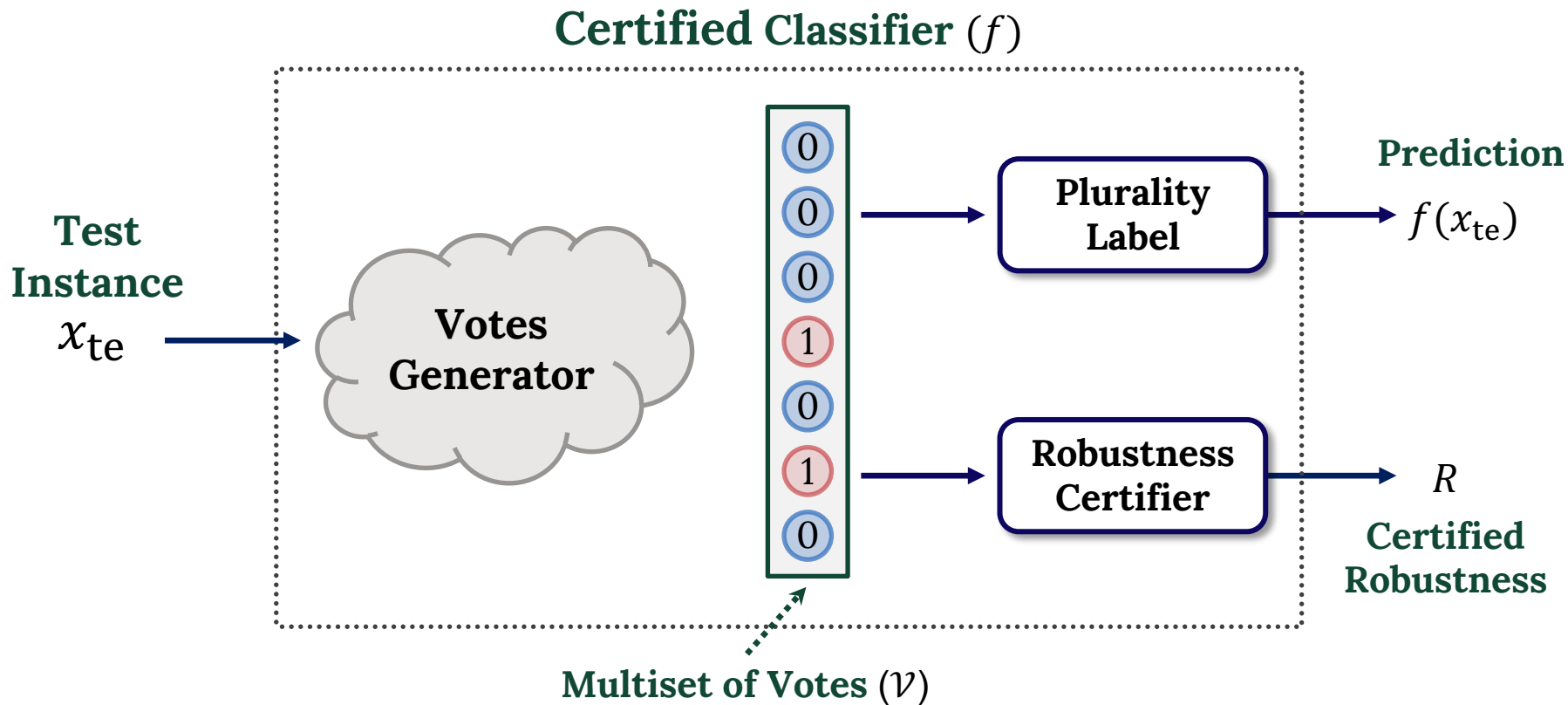
General Structure of a Certified Poisoning Classifier



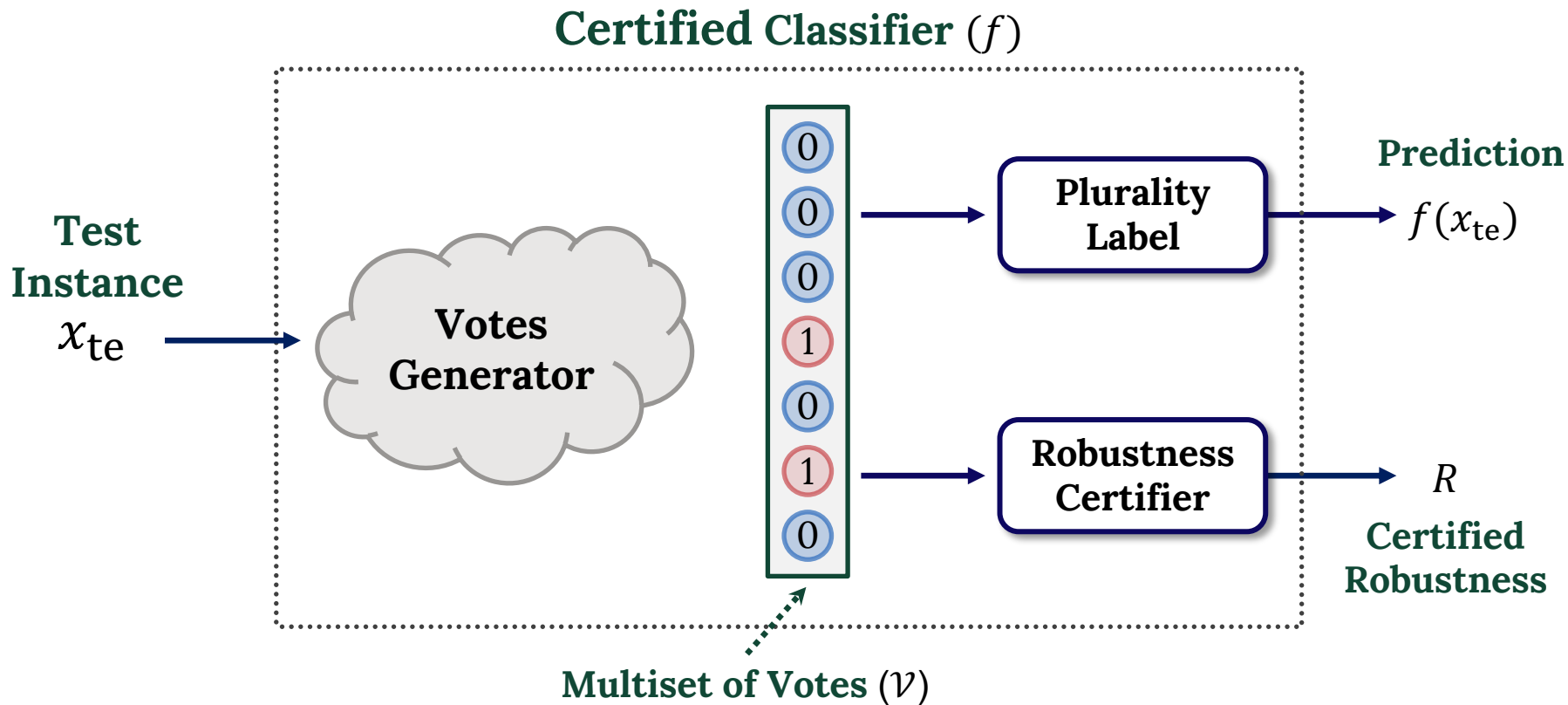
General Structure of a Certified Poisoning Classifier



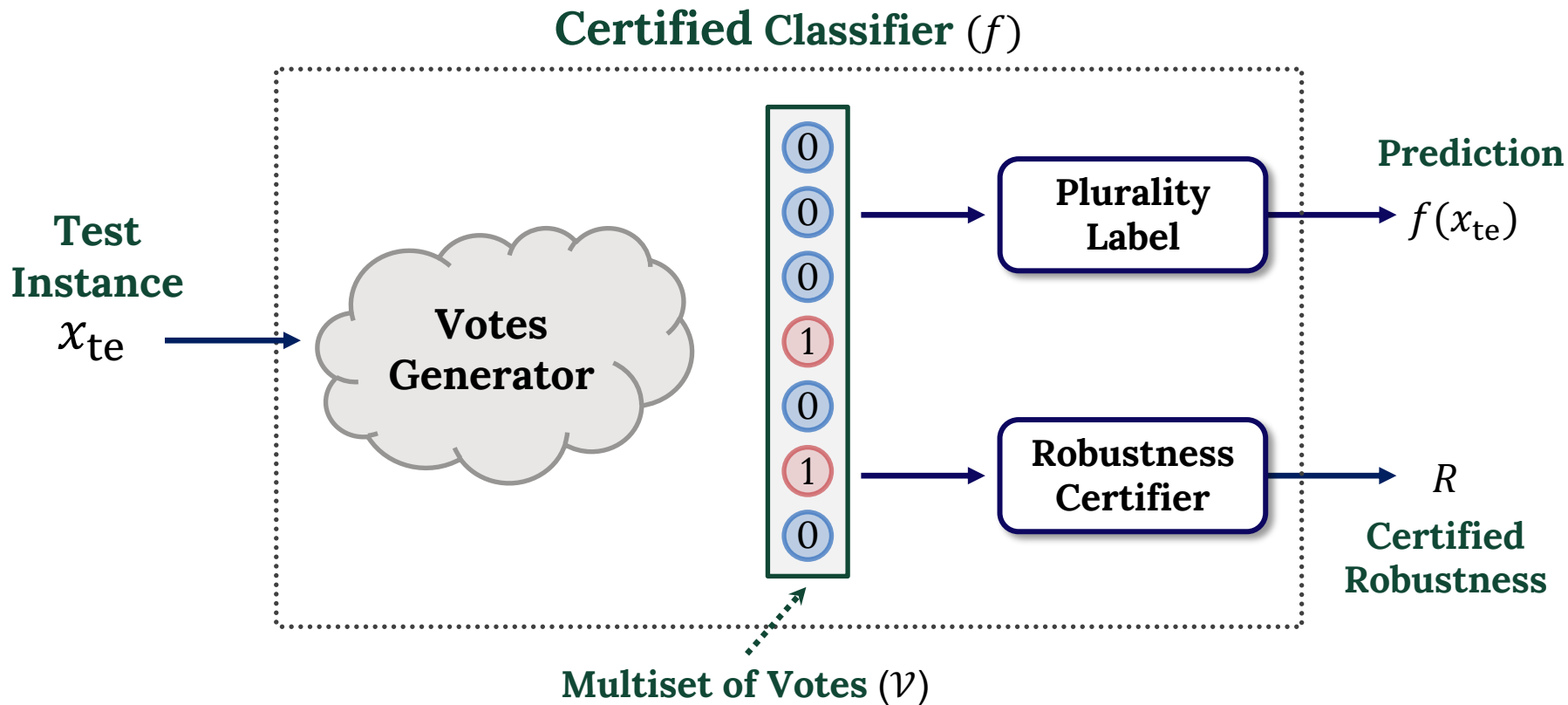
General Structure of a Certified Poisoning Classifier



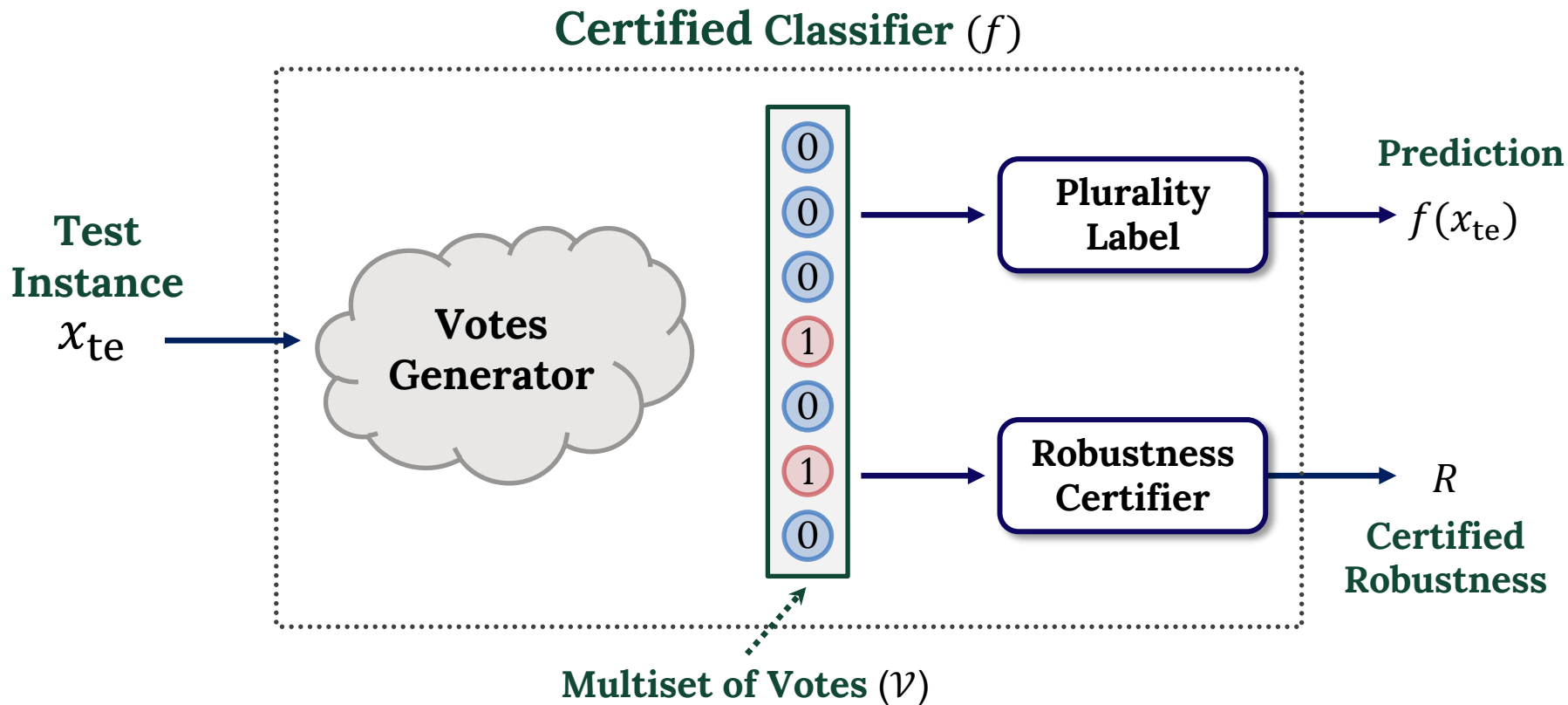
General Structure of a Certified Poisoning Classifier



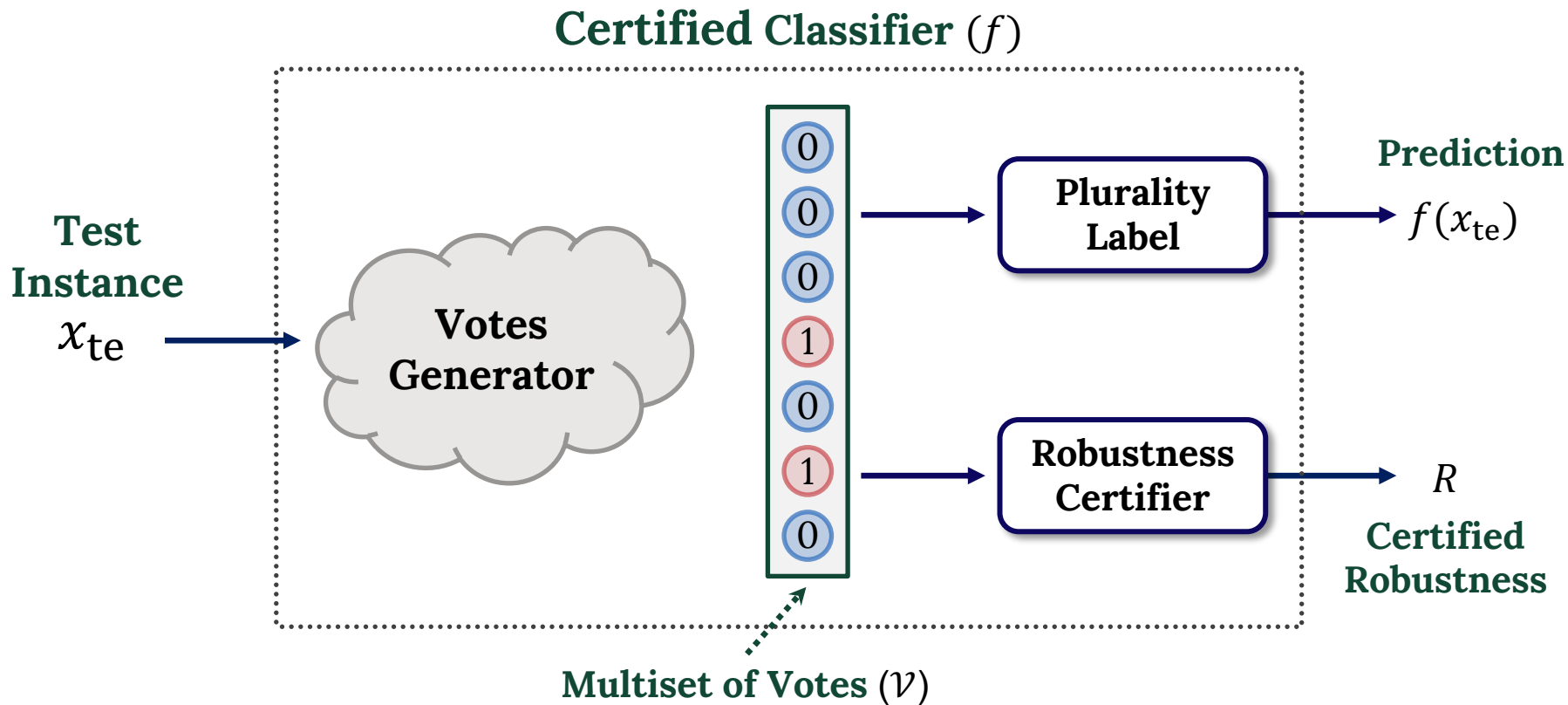
General Structure of a Certified Poisoning Classifier



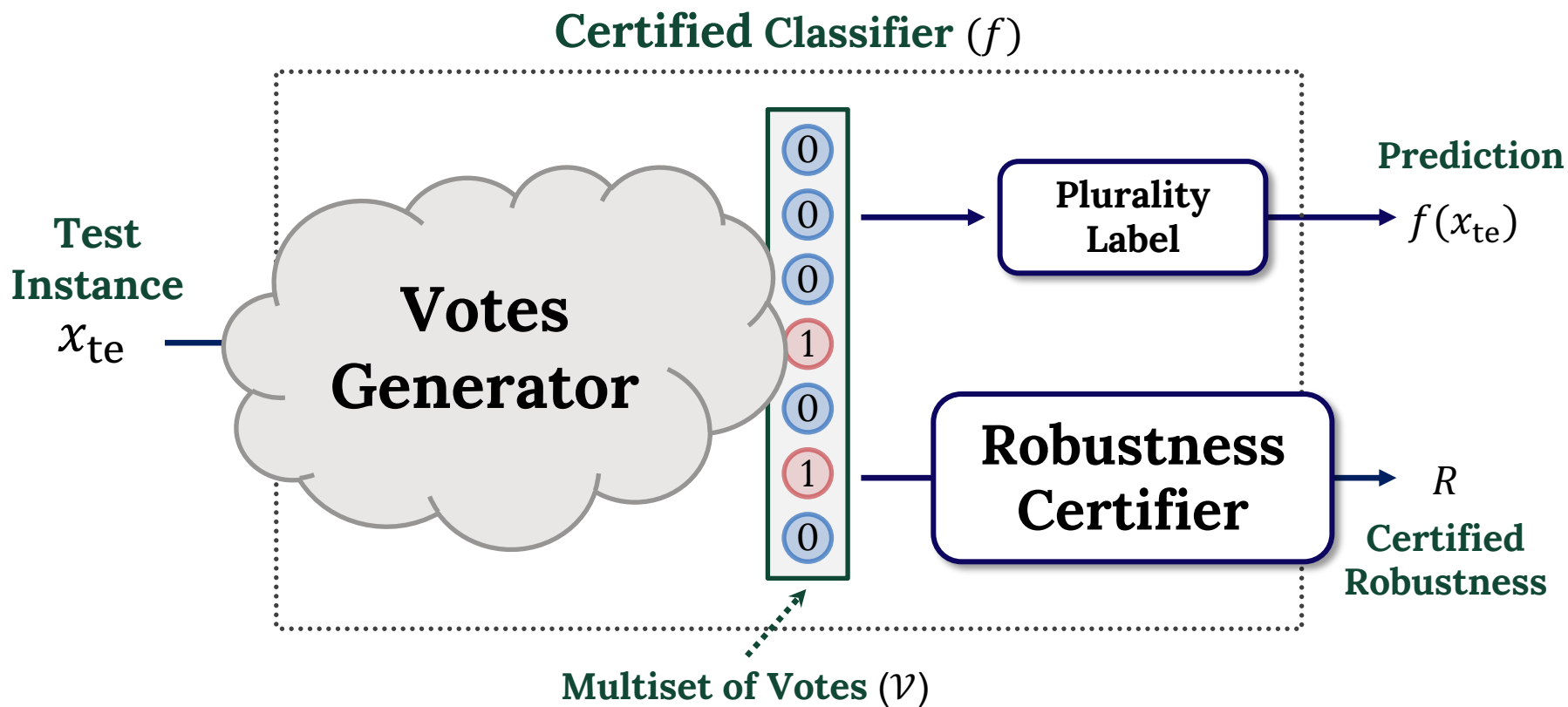
General Structure of a Certified Poisoning Classifier



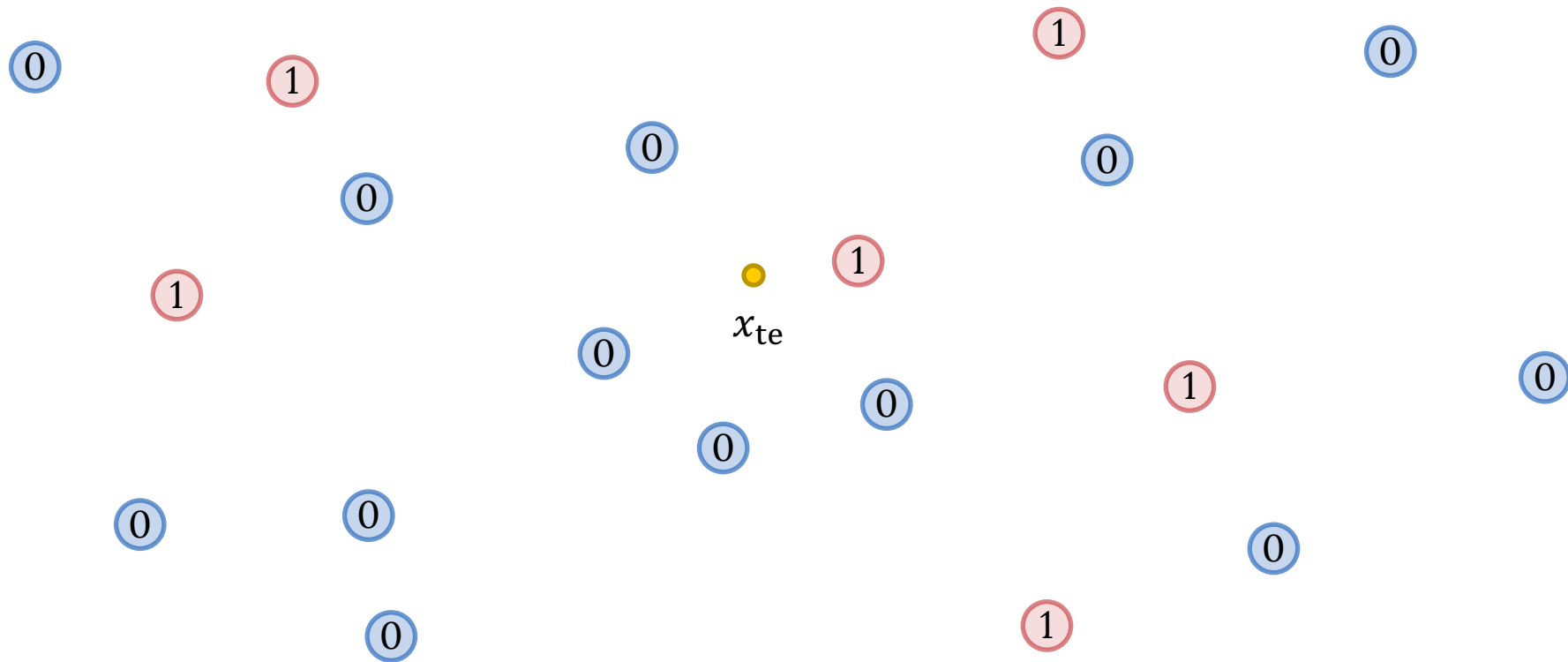
General Structure of a Certified Poisoning Classifier



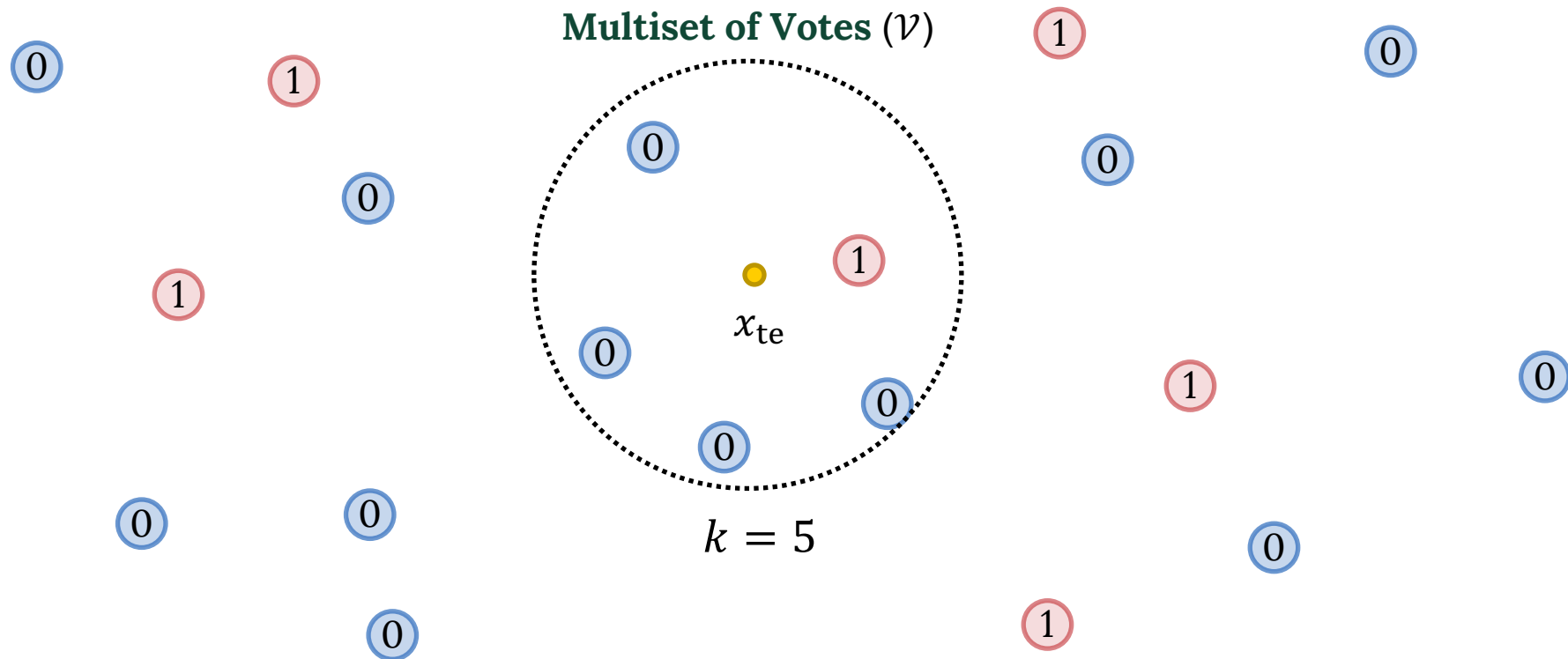
General Structure of a Certified Poisoning Classifier



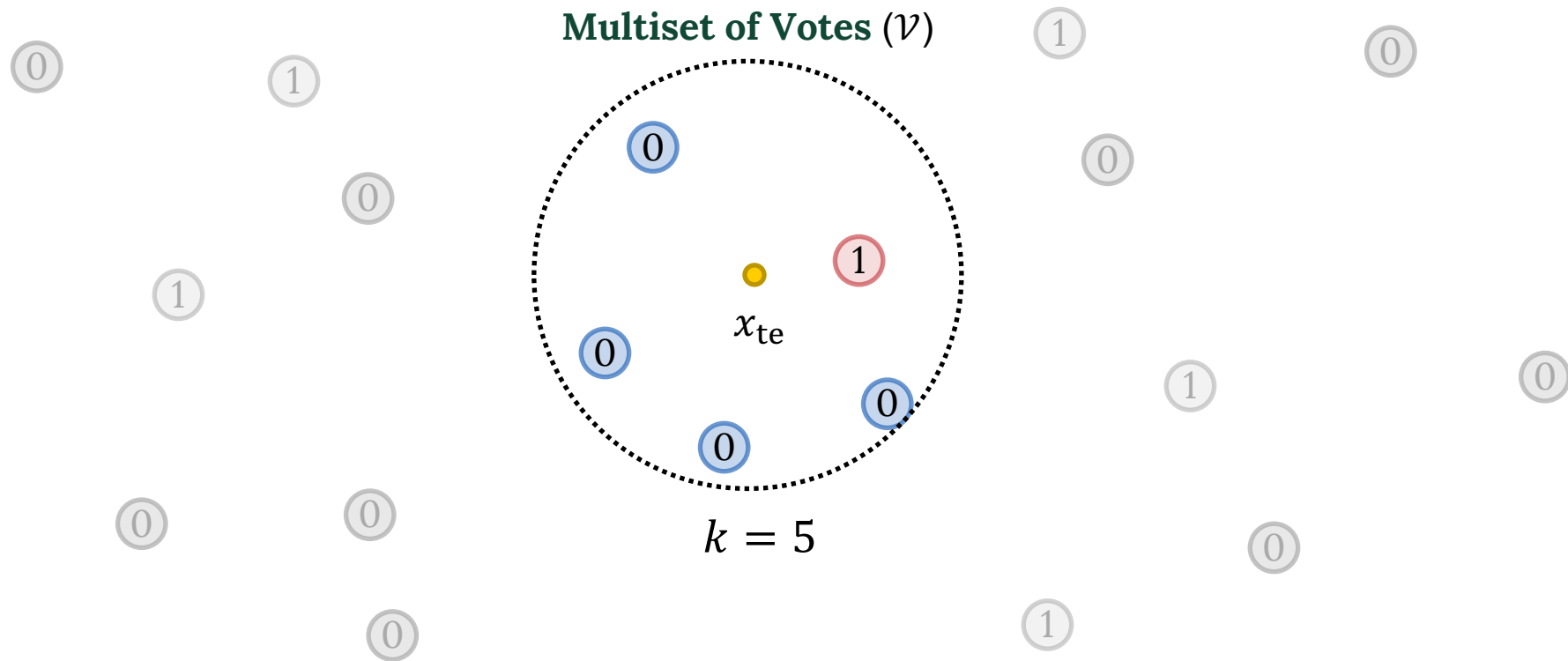
Example: kNN Certified Classifier [Jia+22]



Example: kNN Certified Classifier [Jia+22]



Example: kNN Certified Classifier [Jia+22]



Example: kNN Certified Classifier [Jia+22]



Vote Distribution:

- 4 votes label 0
- 1 vote label 1

Robustness Certifier: At least two votes must change to perturb the plurality label

$$R = \left\lceil \frac{4 - 1}{2} \right\rceil = 1$$

Example: Certified Ensemble Classifier [LF21]

Example: Certified Ensemble Classifier [LF21]

Ensemble

f_1

f_2

f_3

f_4

f_5

Example: Certified Ensemble Classifier [LF21]

Test Instance

Ensemble

x_{te}

f_1

x_{te}

f_2

x_{te}

f_3

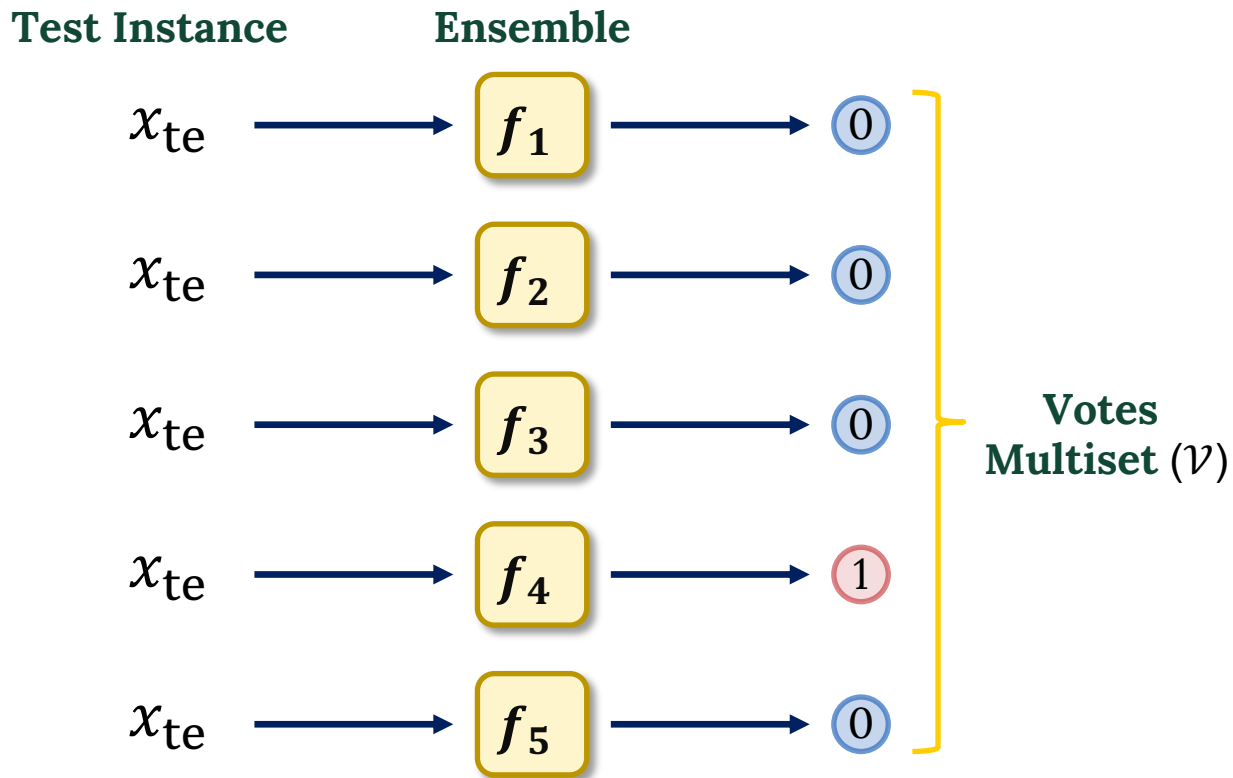
x_{te}

f_4

x_{te}

f_5

Example: Certified Ensemble Classifier [LF21]

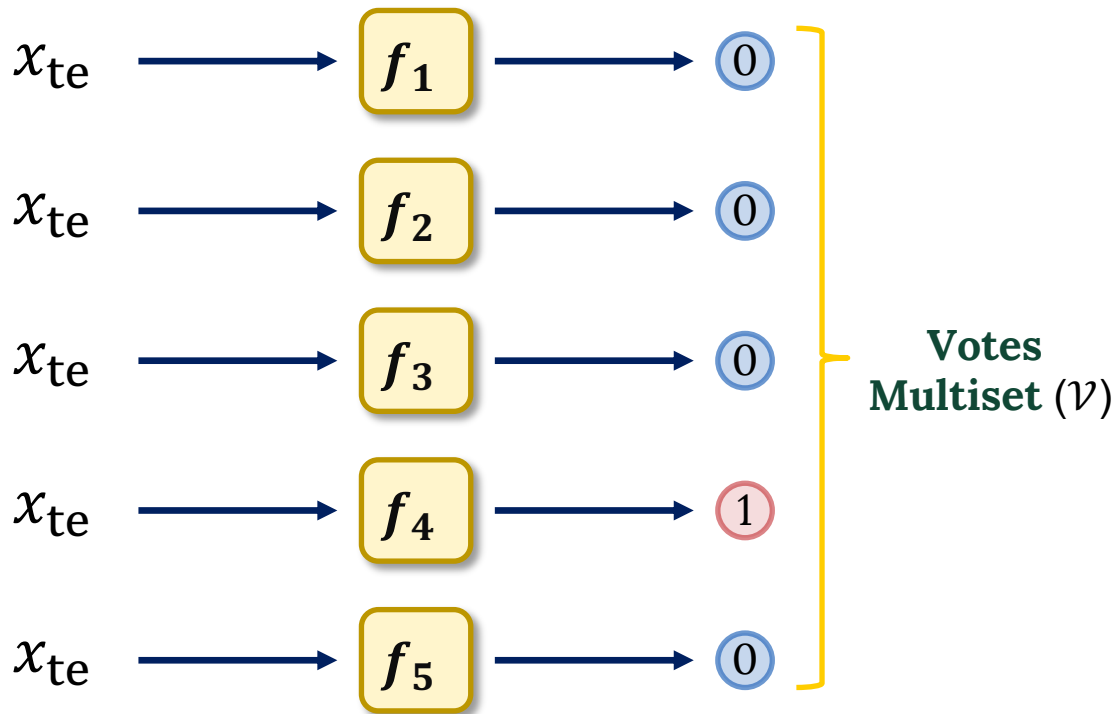


Example: Certified Ensemble Classifier [LF21]

Test Instance

Ensemble

Robustness Certification

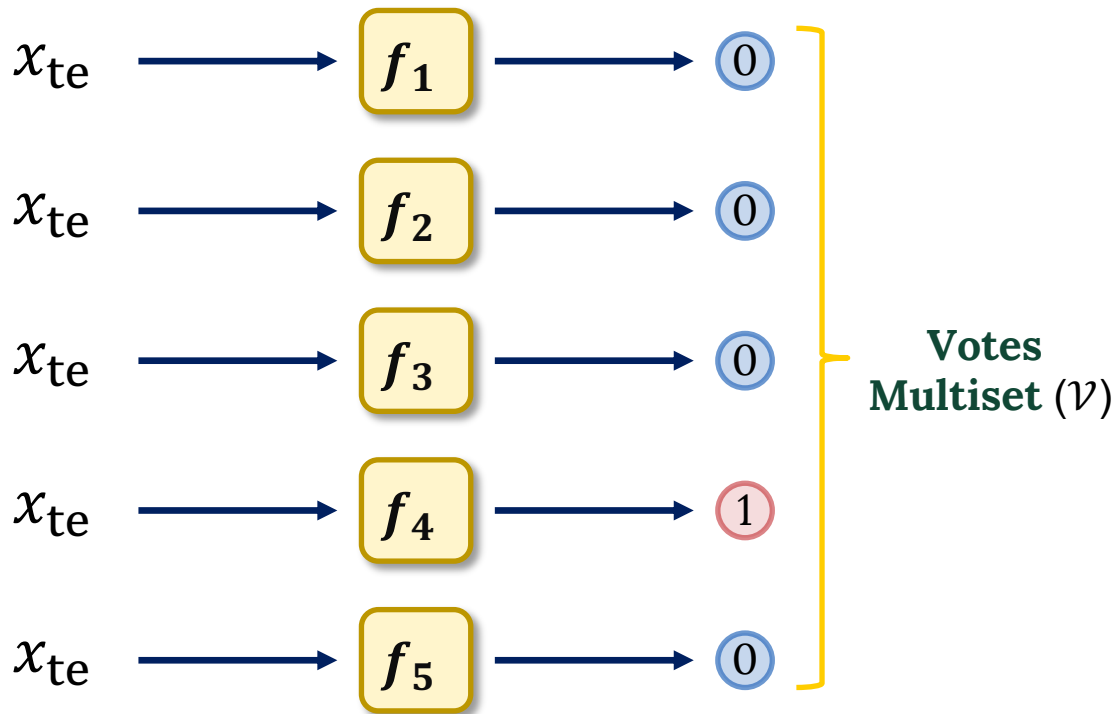


Example: Certified Ensemble Classifier [LF21]

Test Instance

Ensemble

Robustness Certification



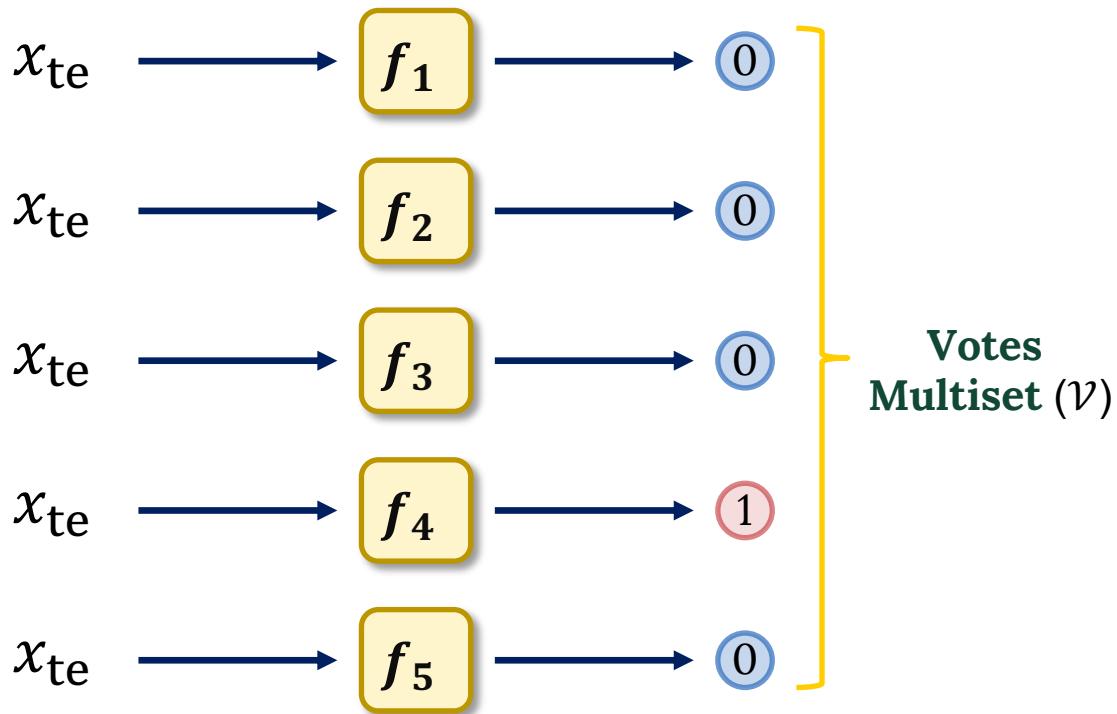
Vote Distribution:

- 4 votes label 0
- 1 vote label 1

Example: Certified Ensemble Classifier [LF21]

Test Instance

Ensemble



Robustness Certification

Vote Distribution:

● 4 votes label 0

● 1 vote label 1

Certified Robustness:

$$R = \left\lfloor \frac{4 - 1}{2} \right\rfloor = 1$$

Reducing Certified Regression to Voting- Based Certified Classification



UNIVERSITY OF
OREGON

“Don’t Reinvent the Wheel”

Reduction: An algorithm for converting a problem Q into a **different problem** Q' that can be **readily solved**.

) to certified classification (Q')

“Transform certified poisoning classifiers into certified regressors”

Benefits:

- Inherit the strengths of the certified classifiers
- Each improved certified classifier improves certified regression

“Don’t Reinvent the Wheel”

)

Reduction: An algorithm for converting a problem Q into a **different problem** Q' that can be **readily solved**.

Our Idea: Reduce certified regression (Q) to certified classification (Q')

“Transform certified poisoning classifiers into certified regressors”

“Transform certified poisoning classifiers into certified regressors”

Benefits:

- Inherit the strengths of the certified classifiers
- Each improved certified classifier improves certified regression

“Don’t Reinvent the Wheel”

)

Reduction: An algorithm for converting a problem Q into a **different problem** Q' that can be **readily solved**.

Our Idea: Reduce certified regression (Q) to certified classification (Q')

“Transform certified poisoning classifiers into certified regressors”

“Transform certified poisoning classifiers into certified regressors”

Benefits:

- Inherit the strengths of the certified classifiers
- Each improved certified classifier improves certified regression
Each improved certified classifier improves certified

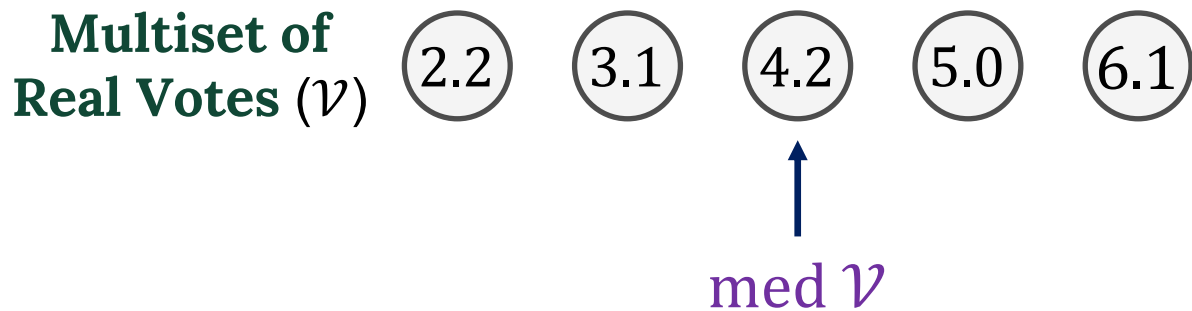
Key Insight of the Reduction

For any binary multiset,
the **plurality label** and **median** have
equivalent robustness

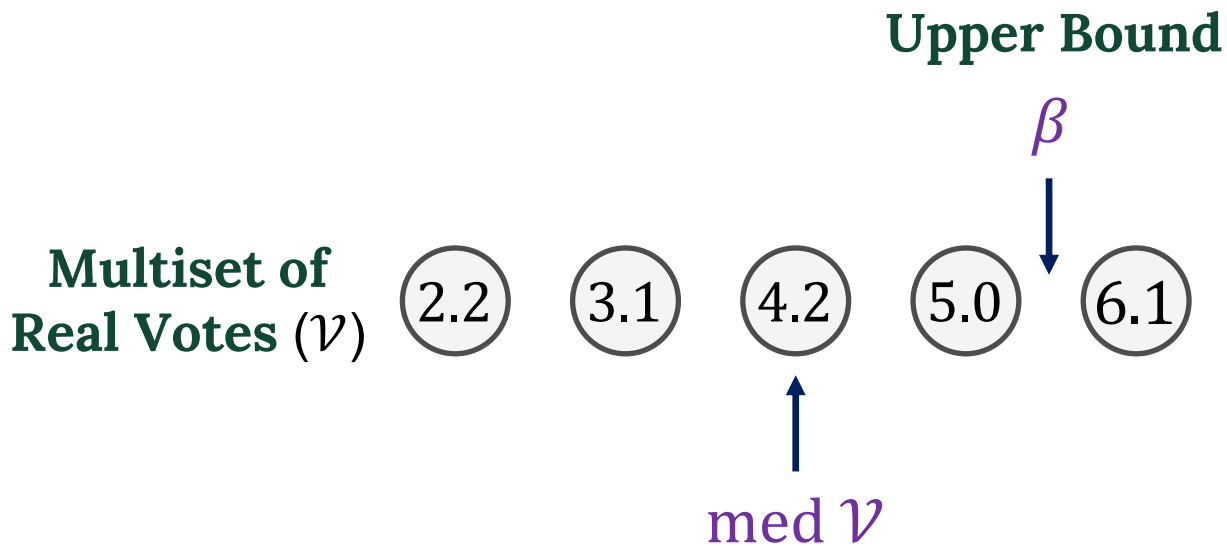
Relating Real-Valued and Binary Robustness

**Multiset of
Real Votes (\mathcal{V})** (2.2) (3.1) (4.2) (5.0) (6.1)

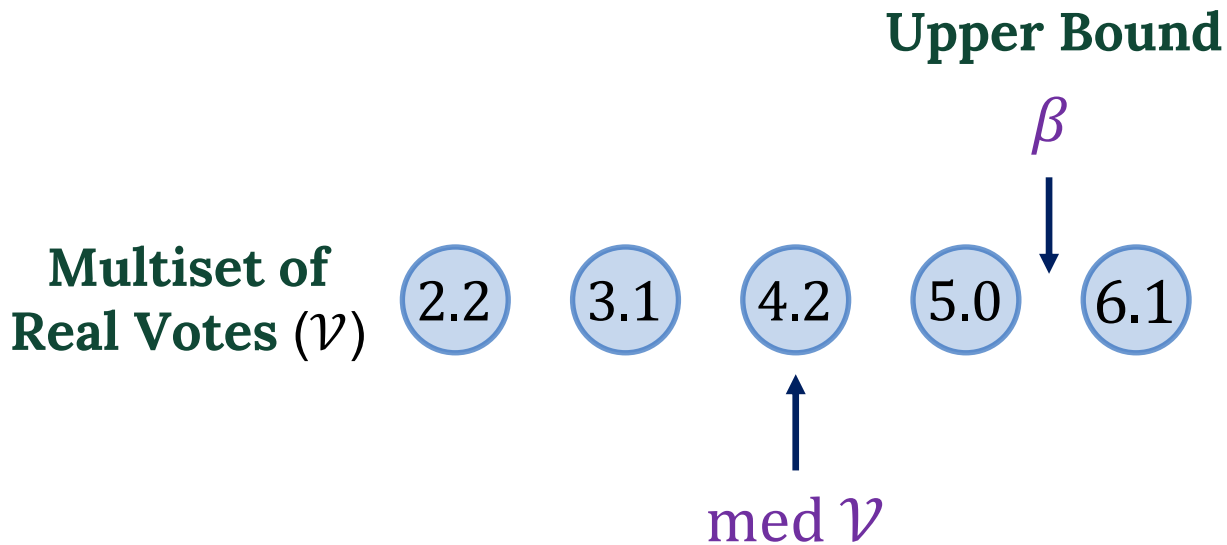
Relating Real-Valued and Binary Robustness



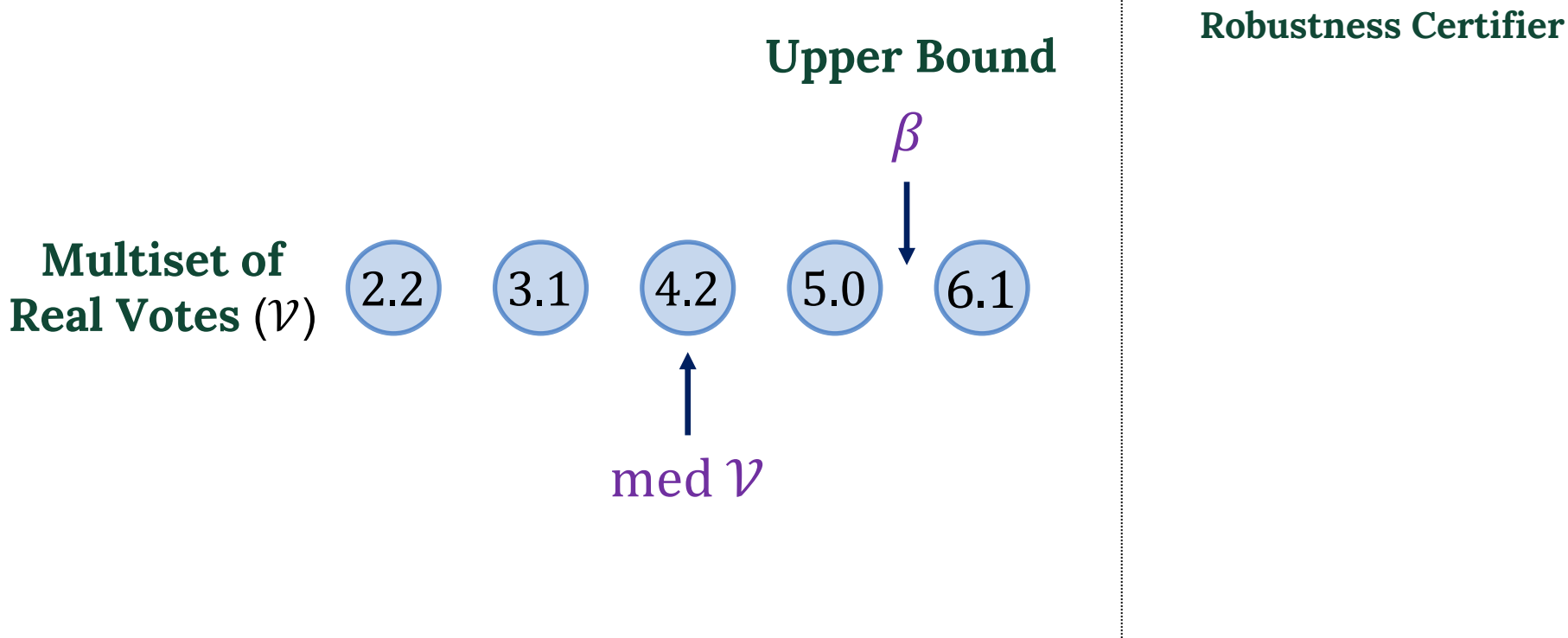
Relating Real-Valued and Binary Robustness



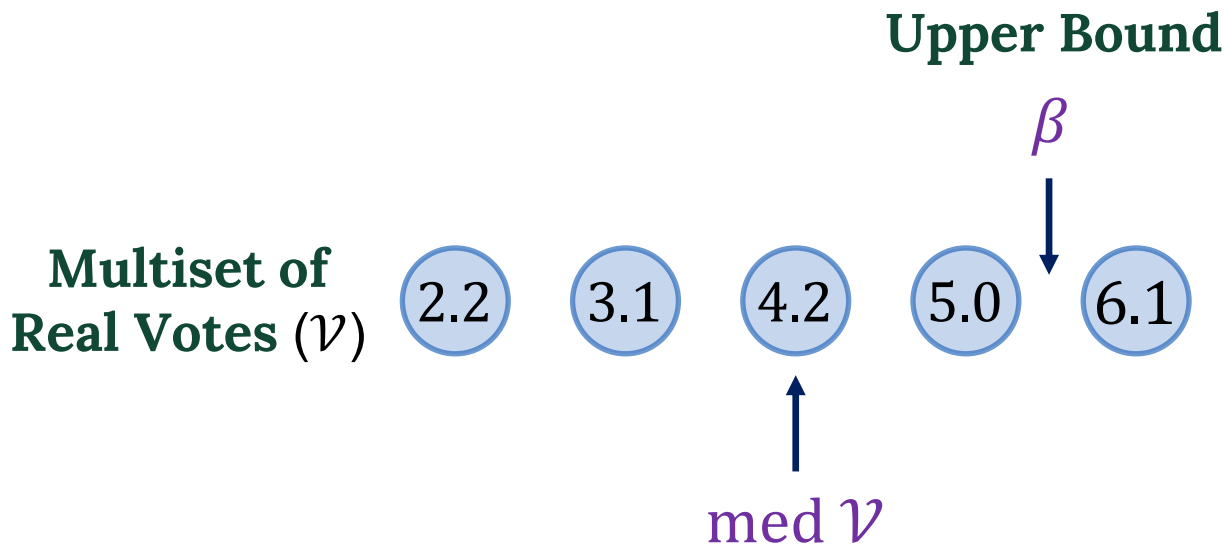
Relating Real-Valued and Binary Robustness



Relating Real-Valued and Binary Robustness



Relating Real-Valued and Binary Robustness

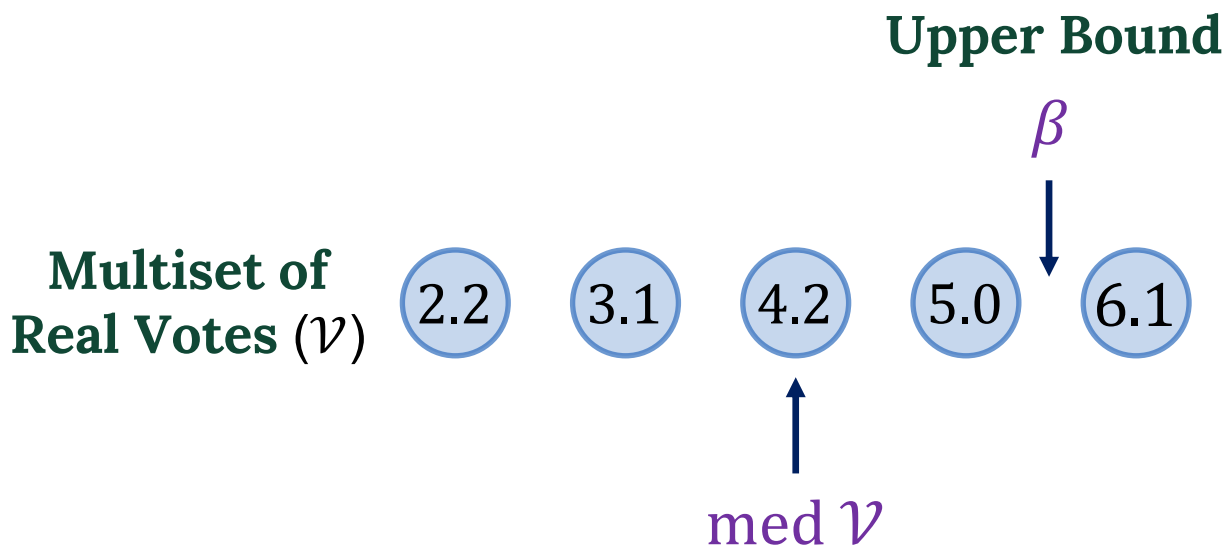


Robustness Certifier

Vote Distribution:

- 4 votes label 0
- 1 vote label 1

Relating Real-Valued and Binary Robustness



Robustness Certifier

Vote Distribution:

- 4 votes label 0
- 1 vote label 1

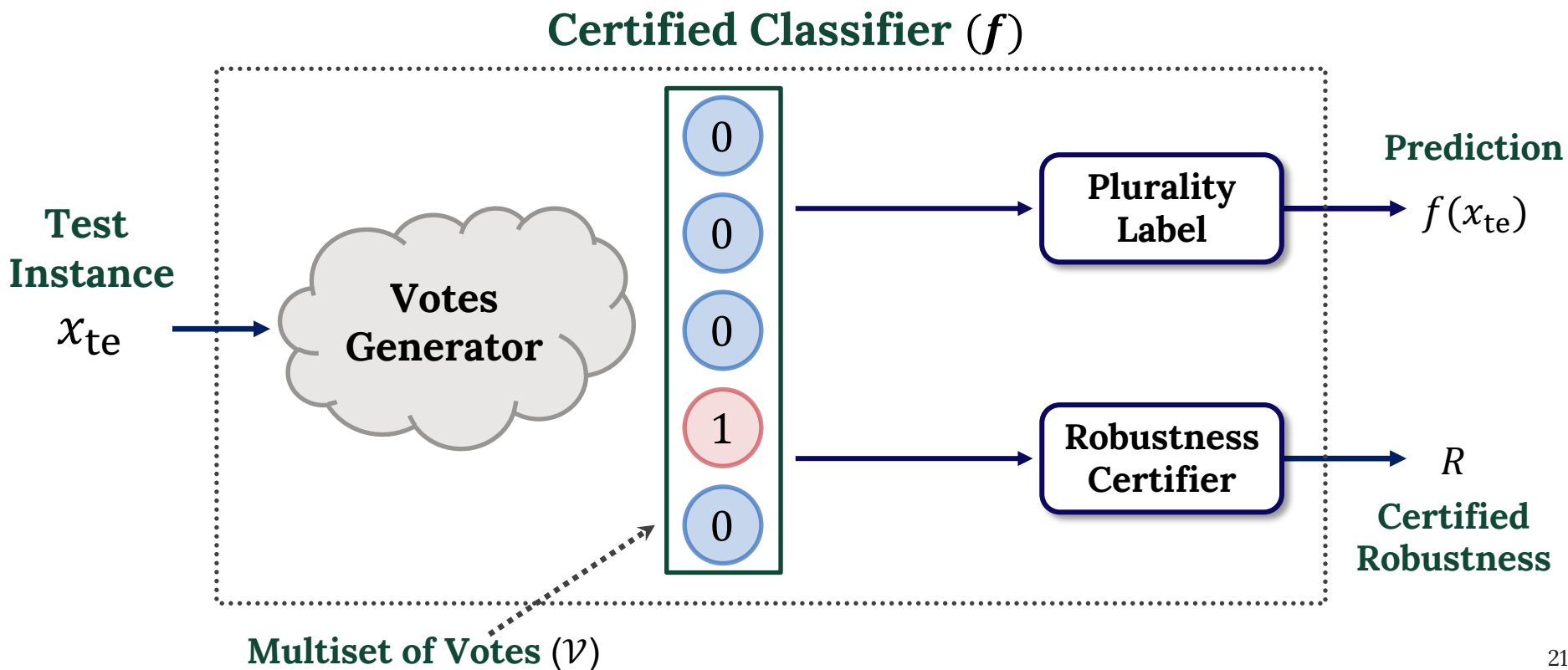
Certified Robustness:

$$R = \left\lfloor \frac{4 - 1}{2} \right\rfloor = 1$$

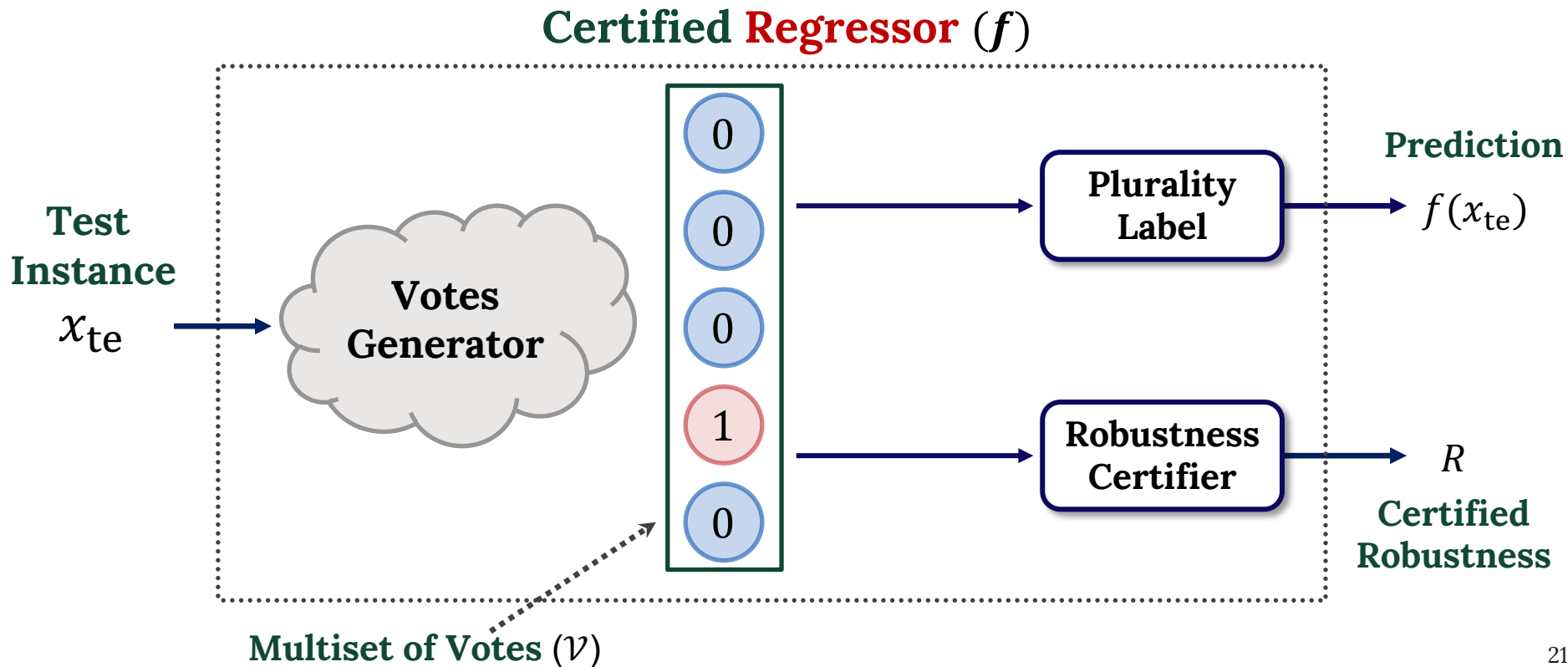
“Transform certified poisoning
classifiers into certified regressors”



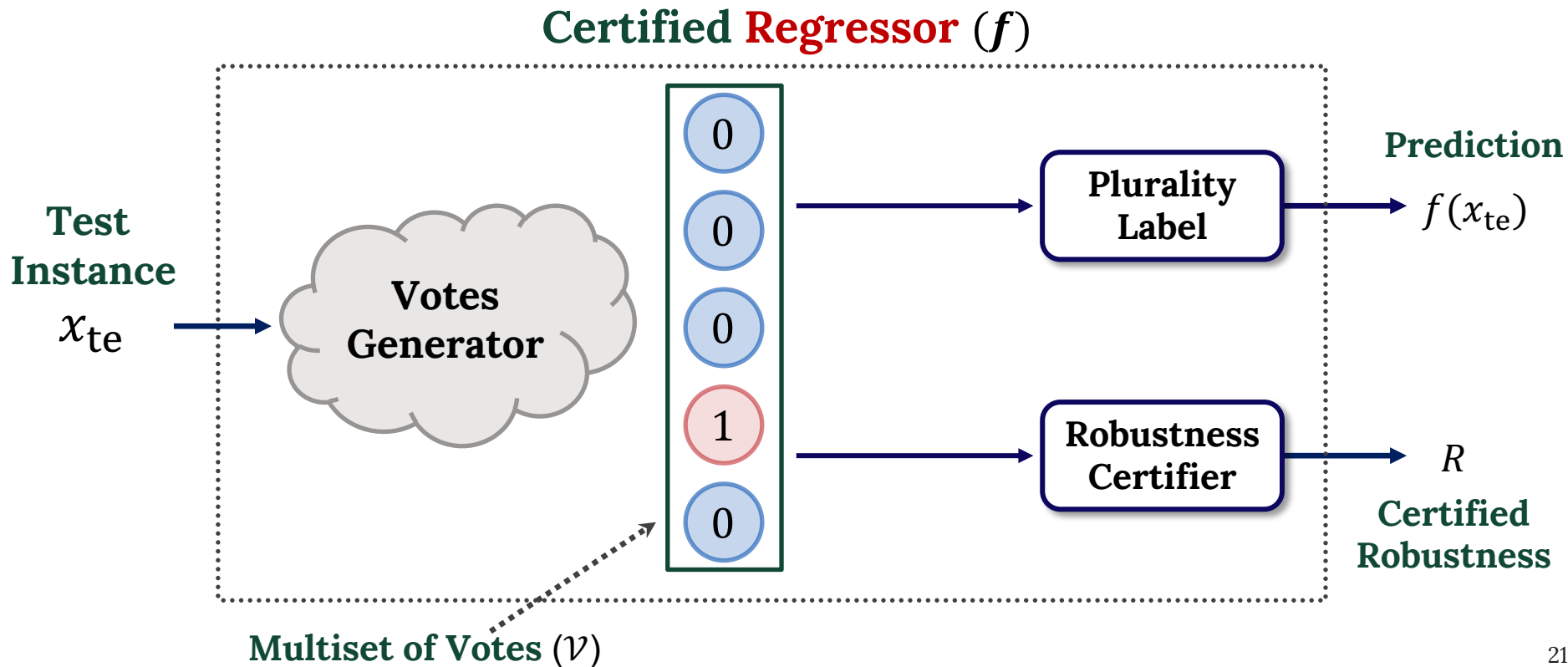
General Structure of a Certified Poisoning Classifier



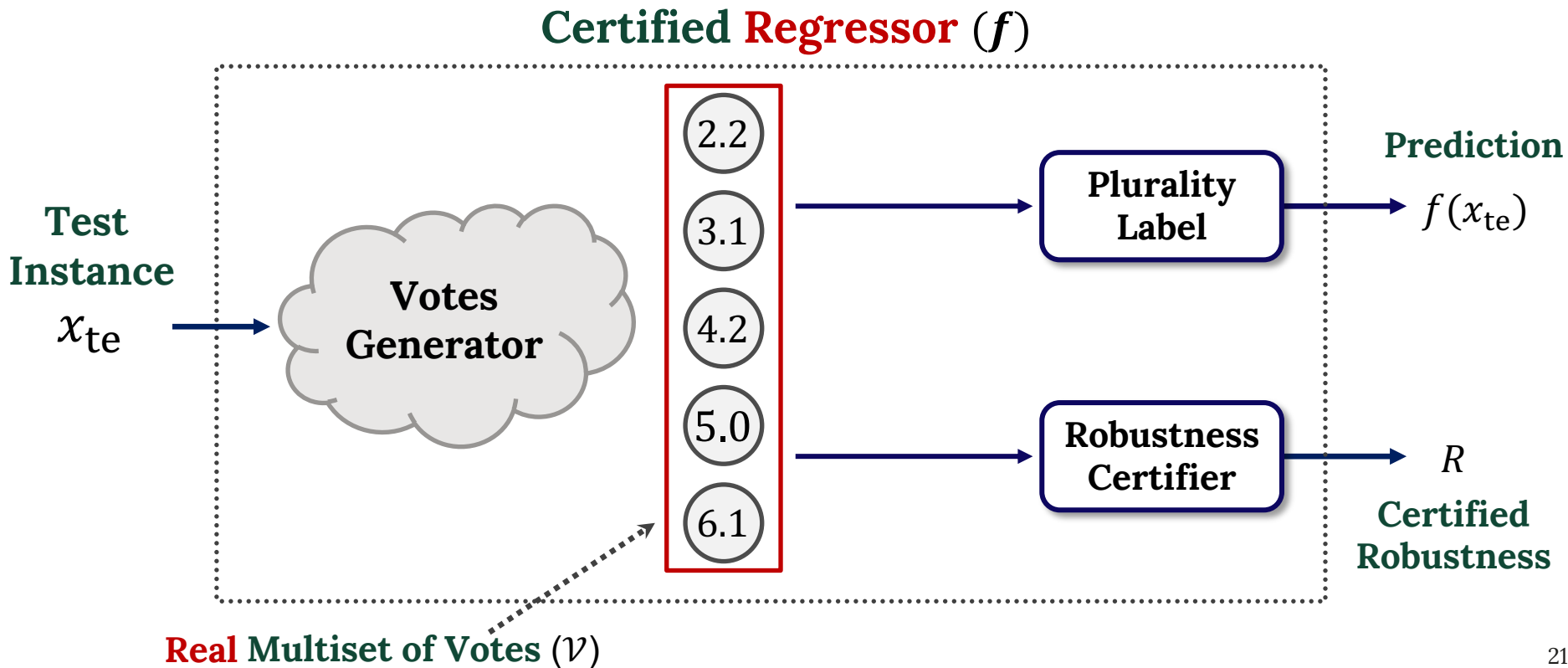
General Structure of a Certified Poisoning **Regressor**



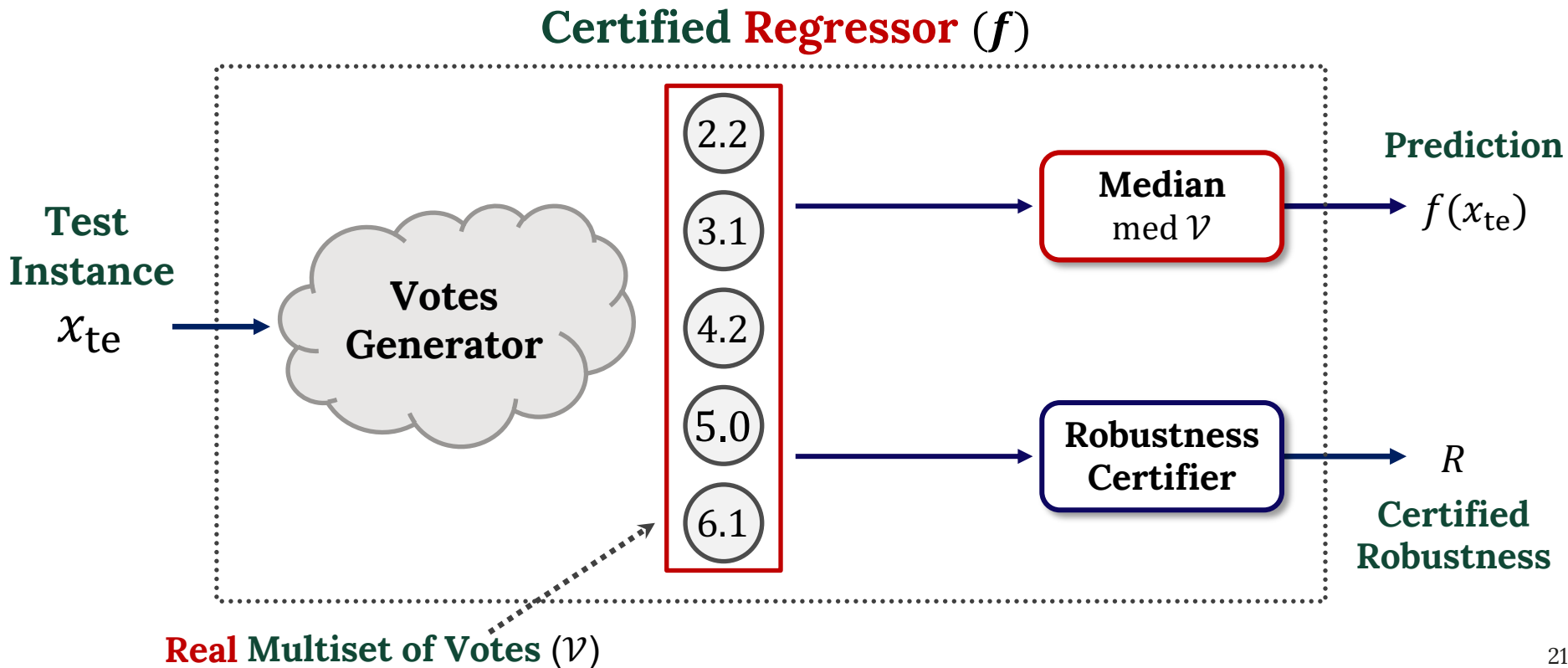
General Structure of a Certified Poisoning **Regressor**



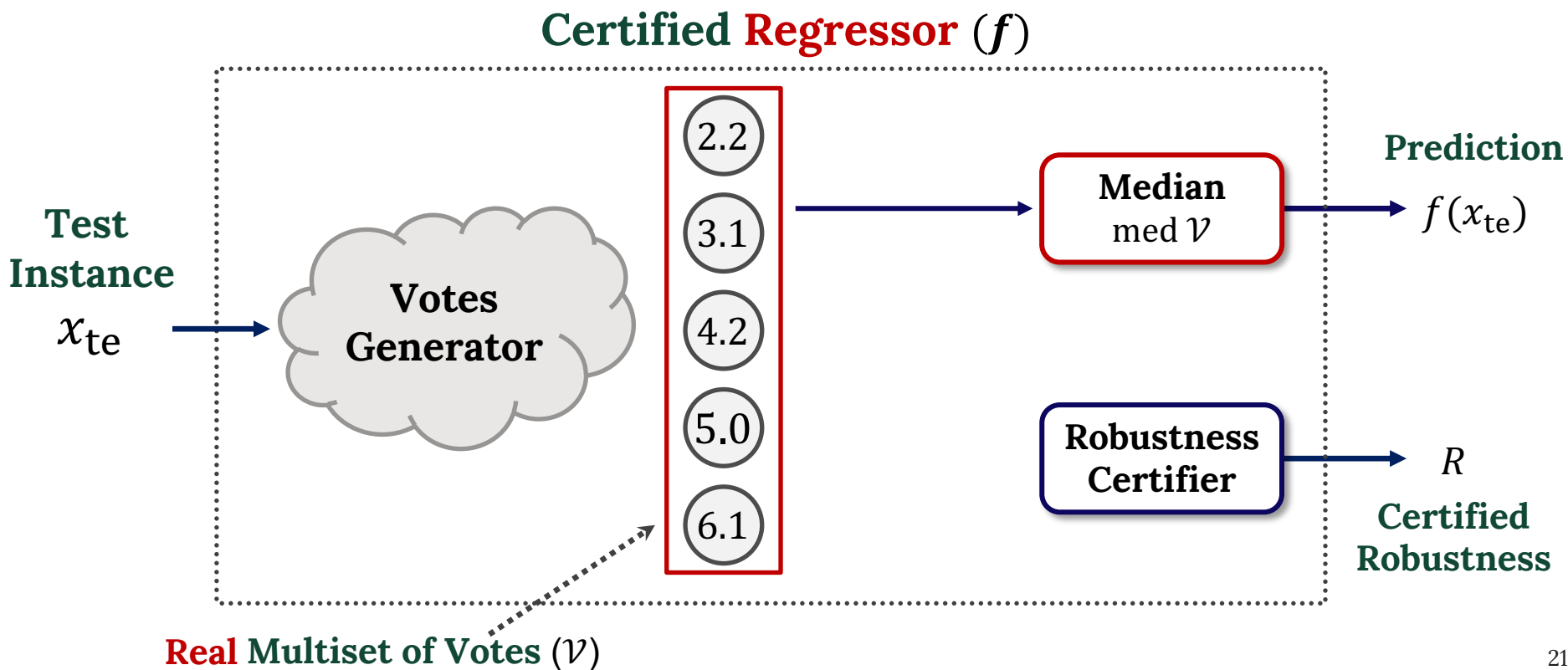
General Structure of a Certified Poisoning **Regressor**



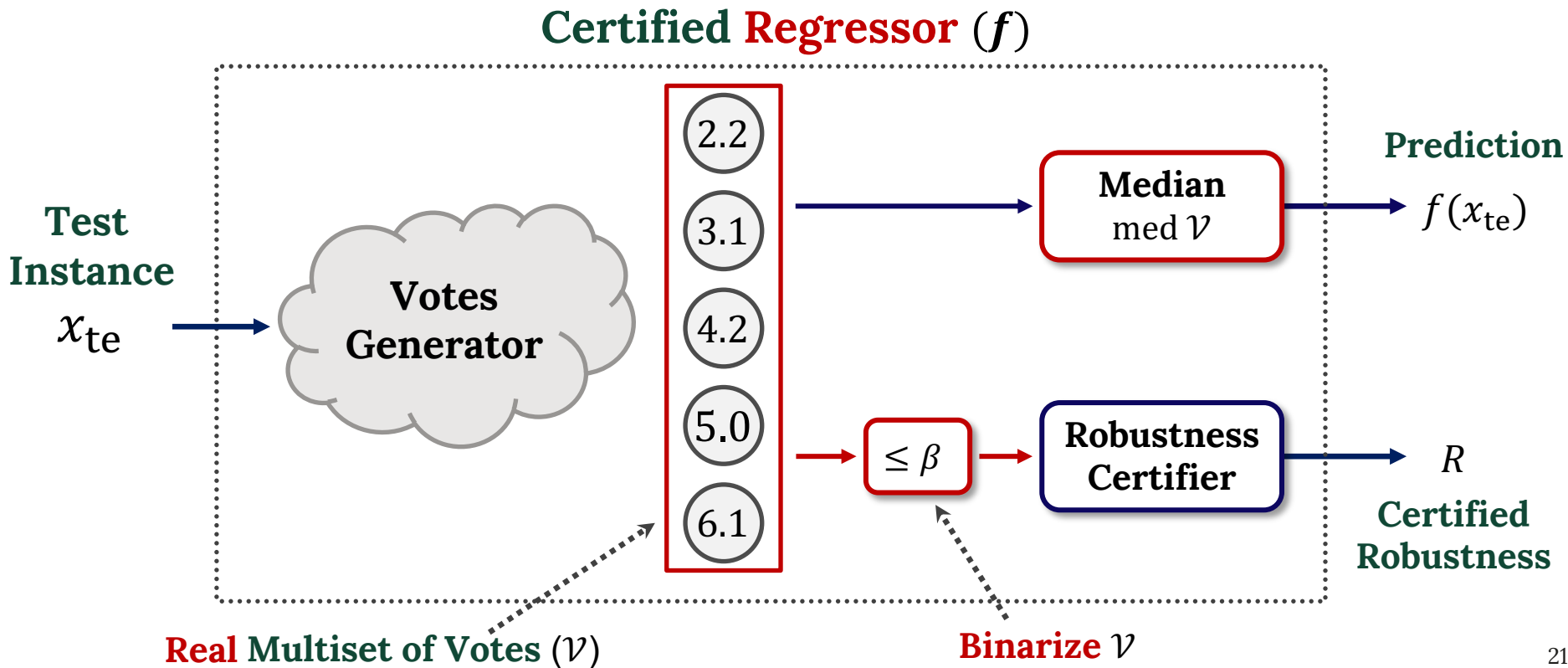
General Structure of a Certified Poisoning **Regressor**



General Structure of a Certified Poisoning **Regressor**



General Structure of a Certified Poisoning **Regressor**



Summary: Reducing Certified Regression to Voting-Based Certified Classification

Three simple steps:

- Generate real-valued votes instead of labels
- Use median as the decision function
- Binarize the real-valued votes \mathcal{V} using threshold β

Our Reduction Yields a Suite of Certified Regressors

We propose six certified regressors:

- Two based on certified **nearest neighbor** classifiers [Jia+22]
- Two based on certified **ensemble** classifiers [LF21, WLF22]
- Two based on **our improved** certified ensemble classifiers

Empirical Evaluation

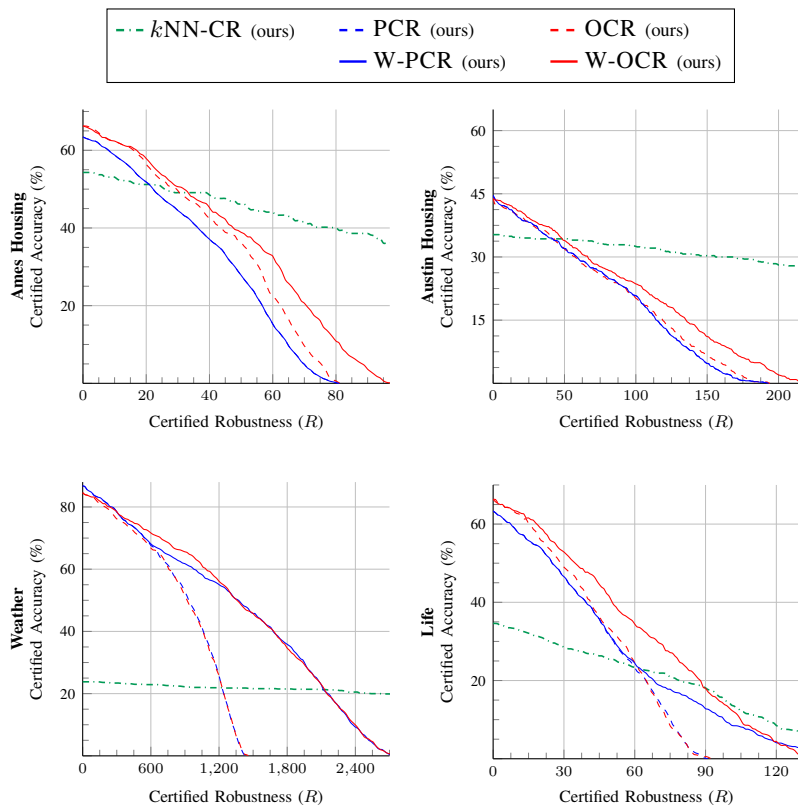


UNIVERSITY OF
OREGON

Empirical Evaluation

- **Datasets:** 5 Regression + 1 Binary Classification
- **Performance Metric: Certified accuracy**
 - Percentage of correctly predicted test instances given α and β with certified robustness $R \geq \psi$
- **Model Architecture Agnostic:** Decision trees and linear models

Certified Regression – Takeaways



Certified Accuracy:

- Half of predictions up to 1% poisoning
- Third of predictions up to 4% poisoning

Method Comparison:

- **Nearest Neighbors:** Better maximum robustness (R)
- **Ensemble:** Better accuracy

One more thing...



UNIVERSITY OF
OREGON

Reducing Certified Regression to Certified Classification for General Poisoning Attacks

Zayd Hammoudeh Daniel Lowd

SaTML 2023 – Raleigh, NC



UNIVERSITY OF
OREGON

Reducing Certified Regression to Certified Classification ~~for General~~ ~~Poisoning Attacks~~

Zayd Hammoudeh Daniel Lowd

SaTML 2023 – Raleigh, NC



UNIVERSITY OF
OREGON

References

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

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

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