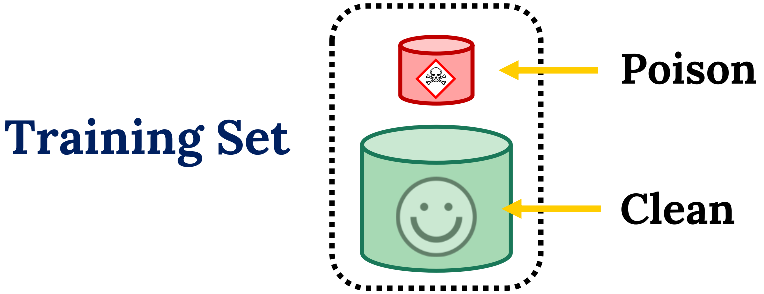
Reducing Certified Regression to Certified Classification for General Poisoning Attacks

Zayd Hammoudeh Daniel Lowd

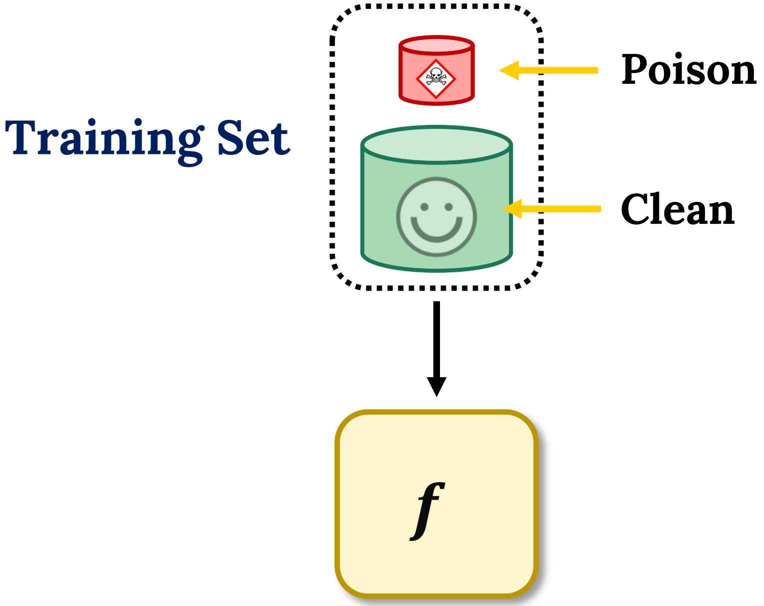
SaTML 2023 – Raleigh, NC

UNIVERSITY OF OREGON

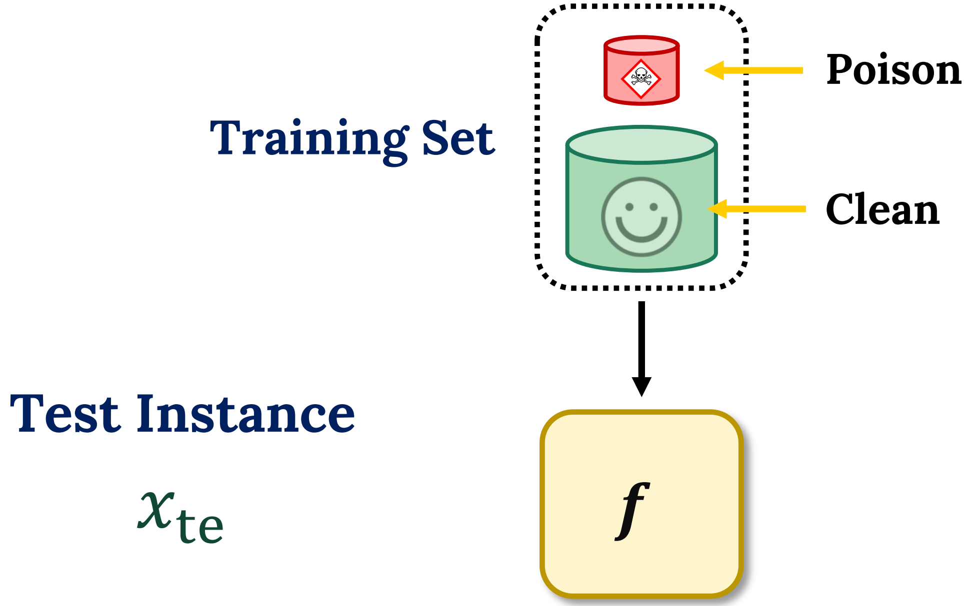
Data Poisoning Whirlwind Review



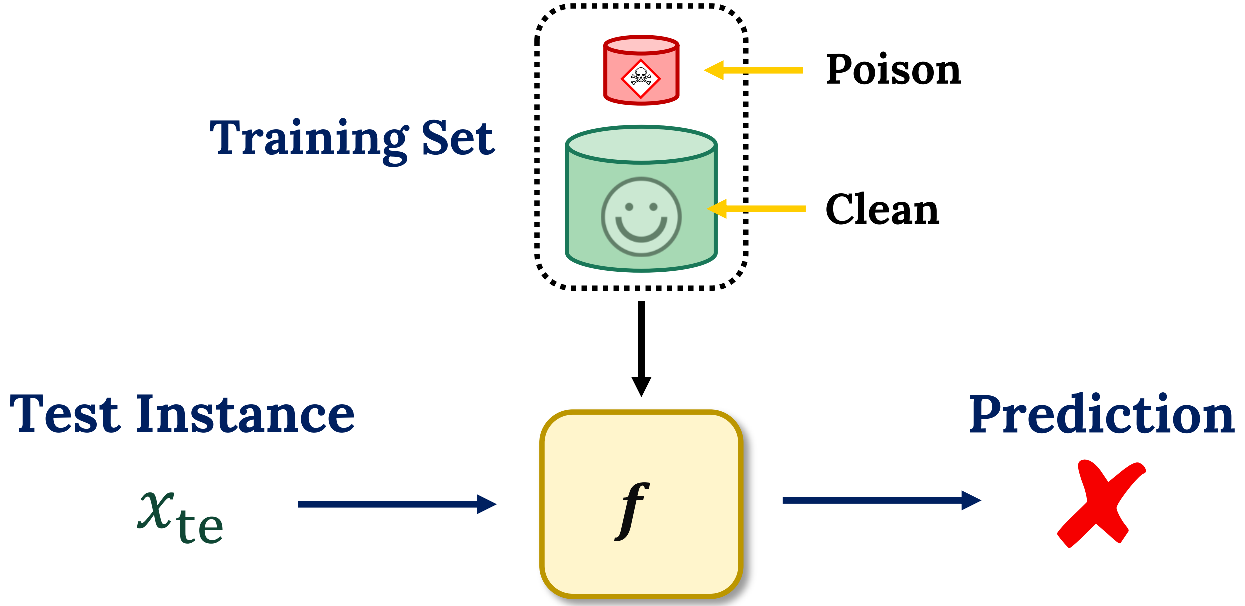
Data Poisoning Whirlwind Review



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:

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

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

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

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

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

Certified Poisoning Classifiers Show Promise

Certified Poisoning Classifiers Show Promise


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

arXiv:2006.14768v2 [cs.LG] 18 Mar 2021

Published as a conference paper at ICLR 2021

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

In AAAI Conference on Artificial Intelligence, 2022.

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

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

Wenxiao Wang [1] **Alexander Levine** [1] **Soheil Feizi** [1]

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%, 3.87% and 4.77%, respectively, while keeping the same clean accuracies as DPA's, 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 processing (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 (Shafahi et al., 2018), Convex Polytope (Zhu et al., 2019), Bullseye 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 backdooring (Turner et al., 2019) and hidden-trigger backdooring (Saha et al., 2020). While it is shown in (Schwarzschild 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)

[1]Department of Computer Science, University of Maryland, College Park, Maryland, USA. Correspondence to: Wenxiao Wang <wwx@umd.edu>.

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

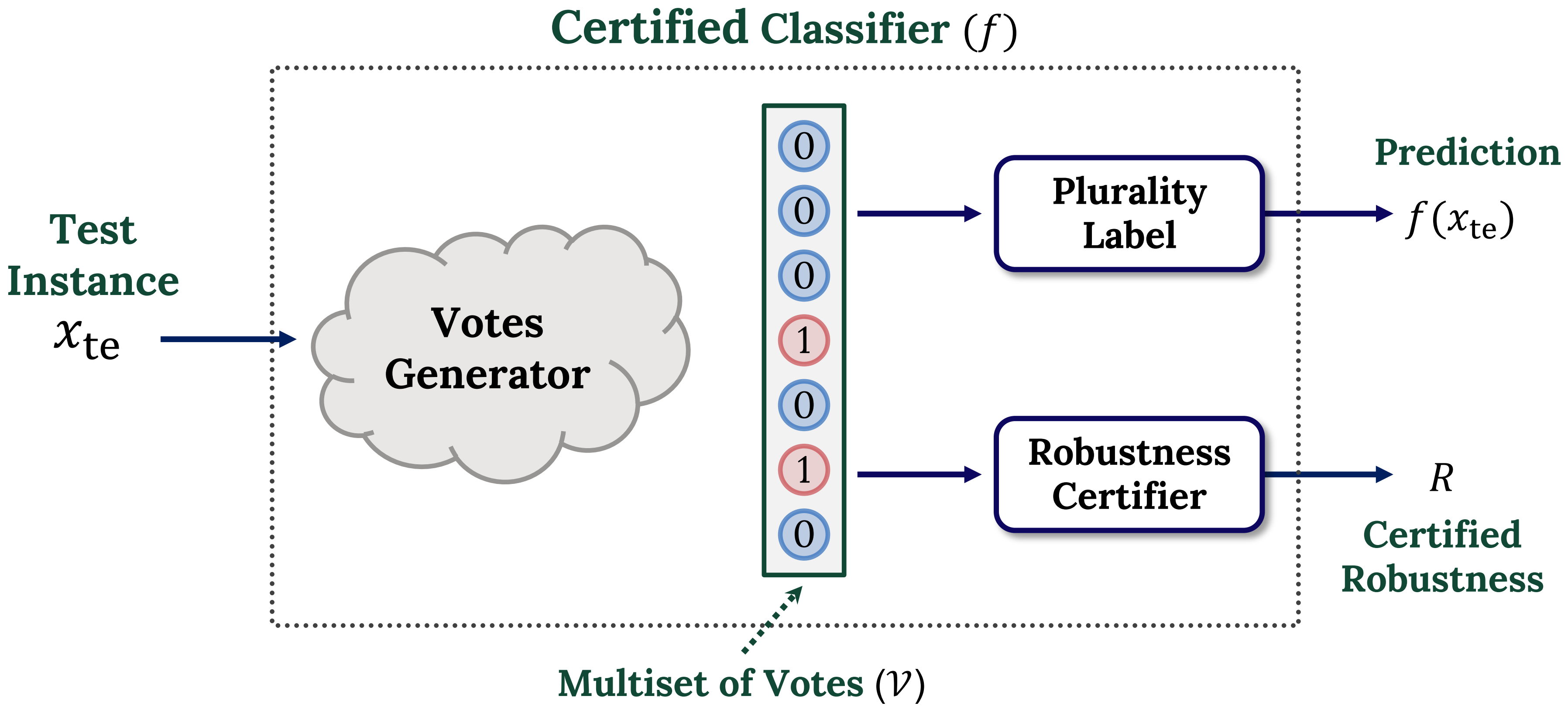
Strengths of Certified Poisoning Classifiers

1. No data distribution assumptions

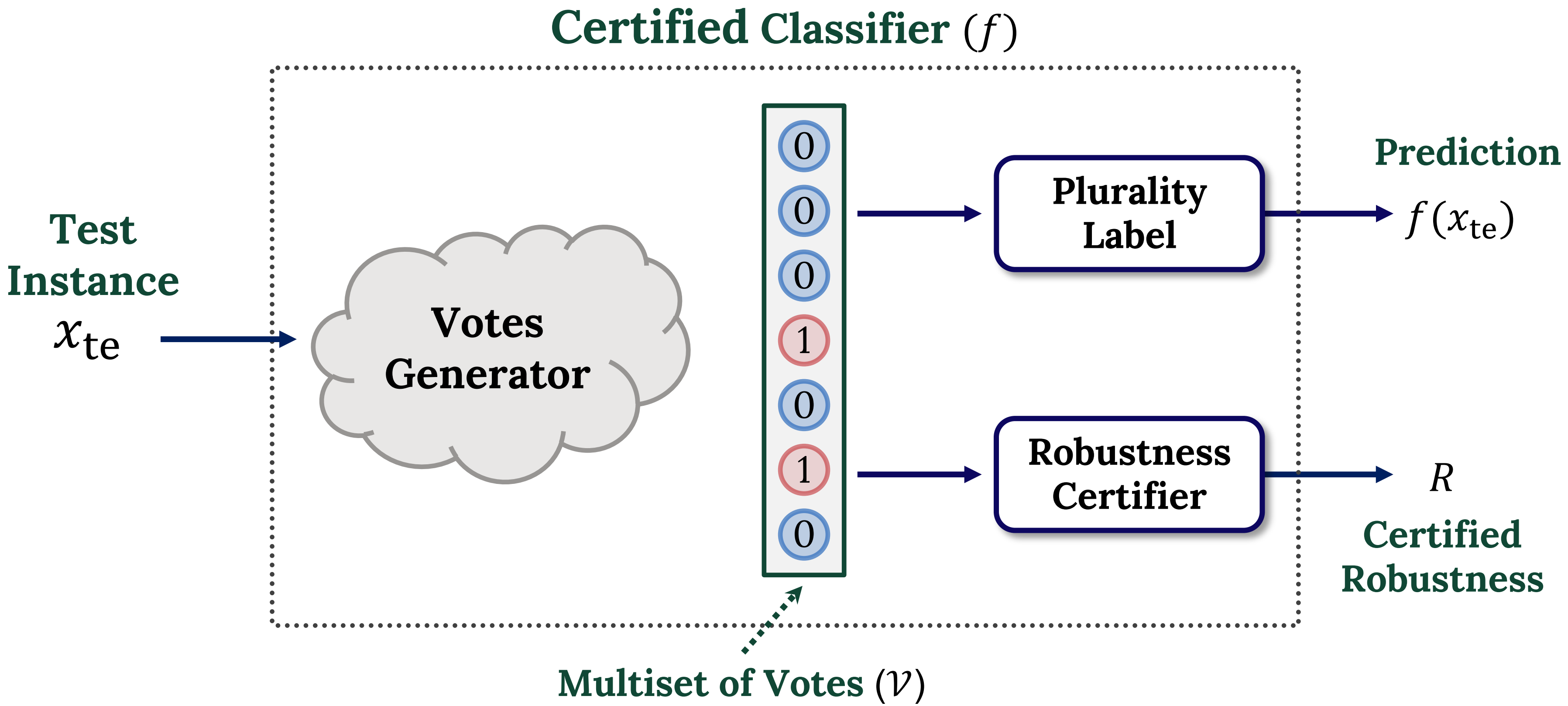
2. Model architecture agnostic

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

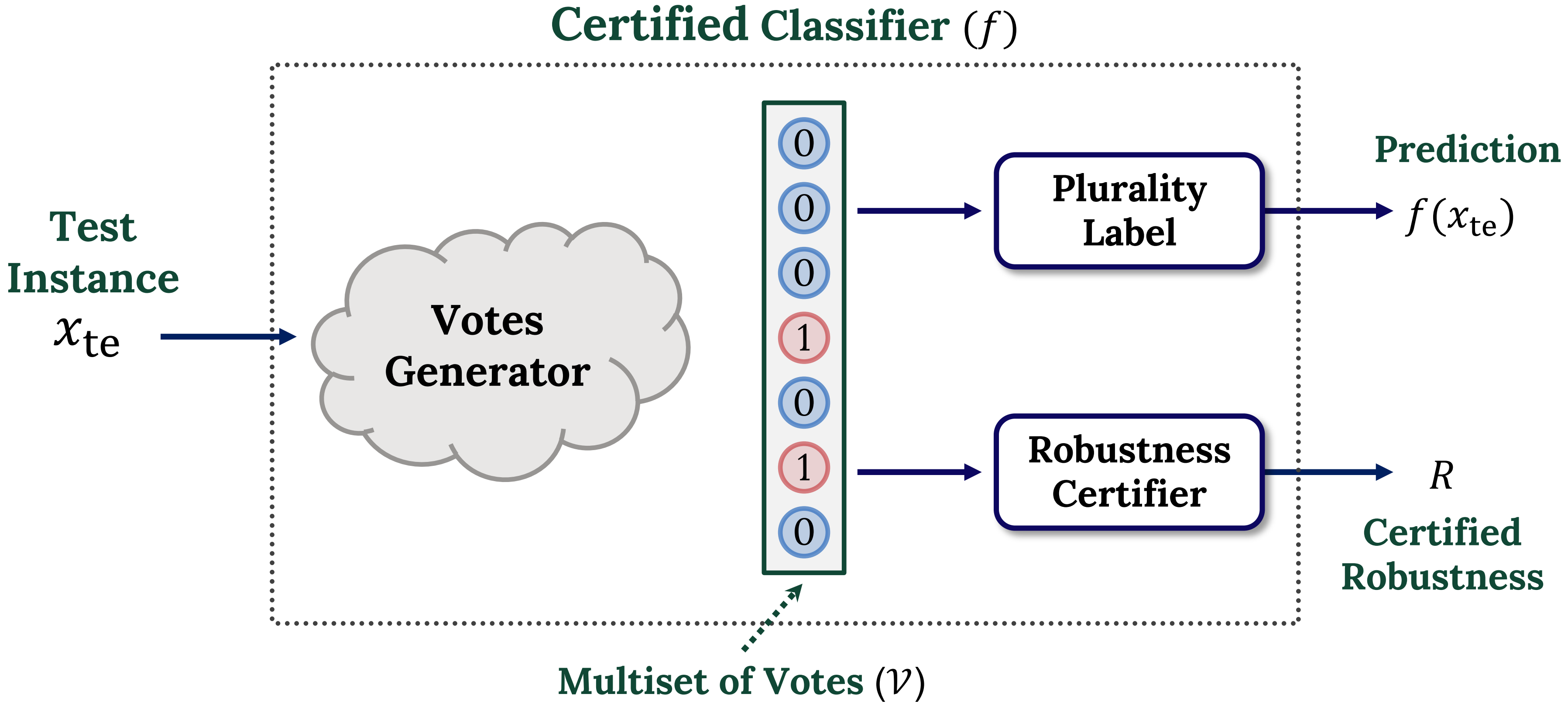
General Structure of a Certified Poisoning Classifier



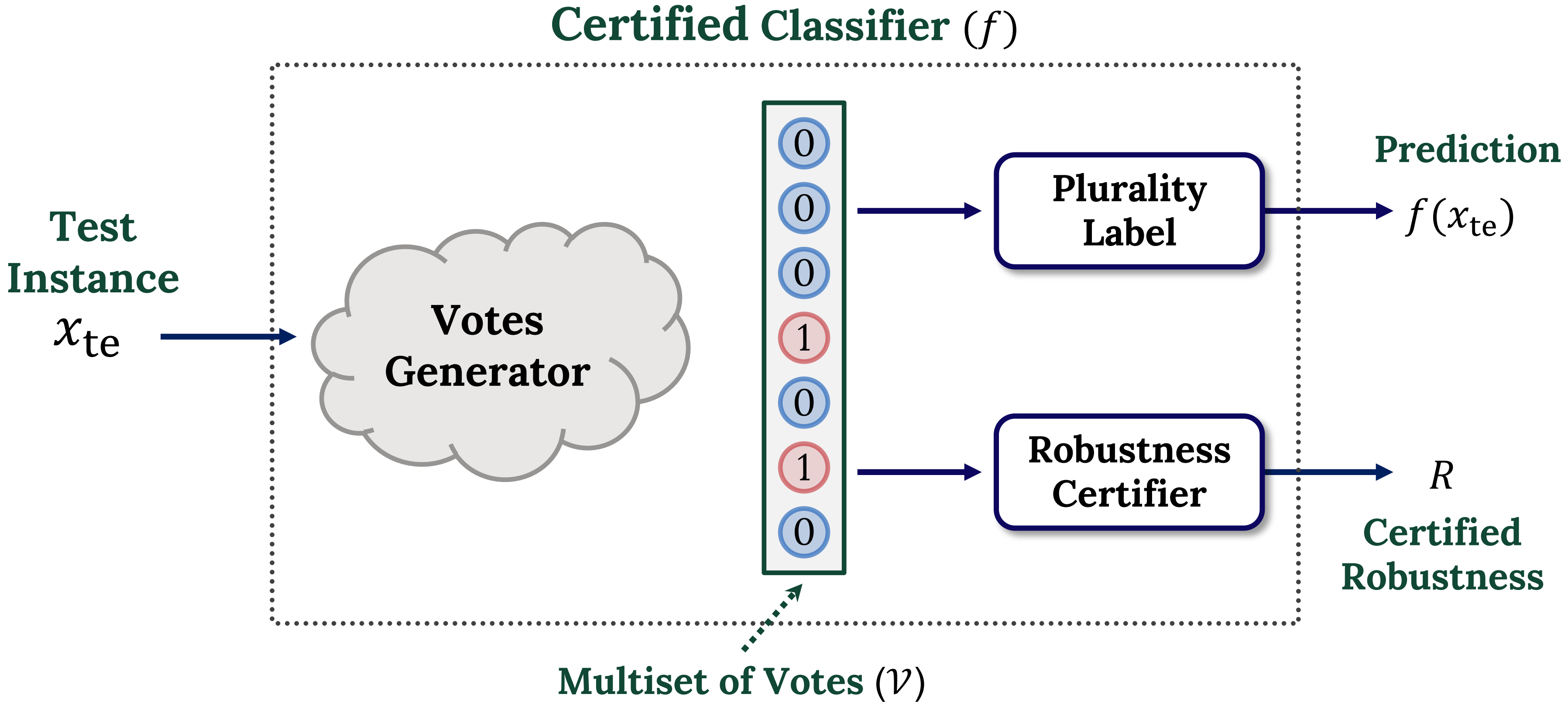
General Structure of a Certified Poisoning Classifier



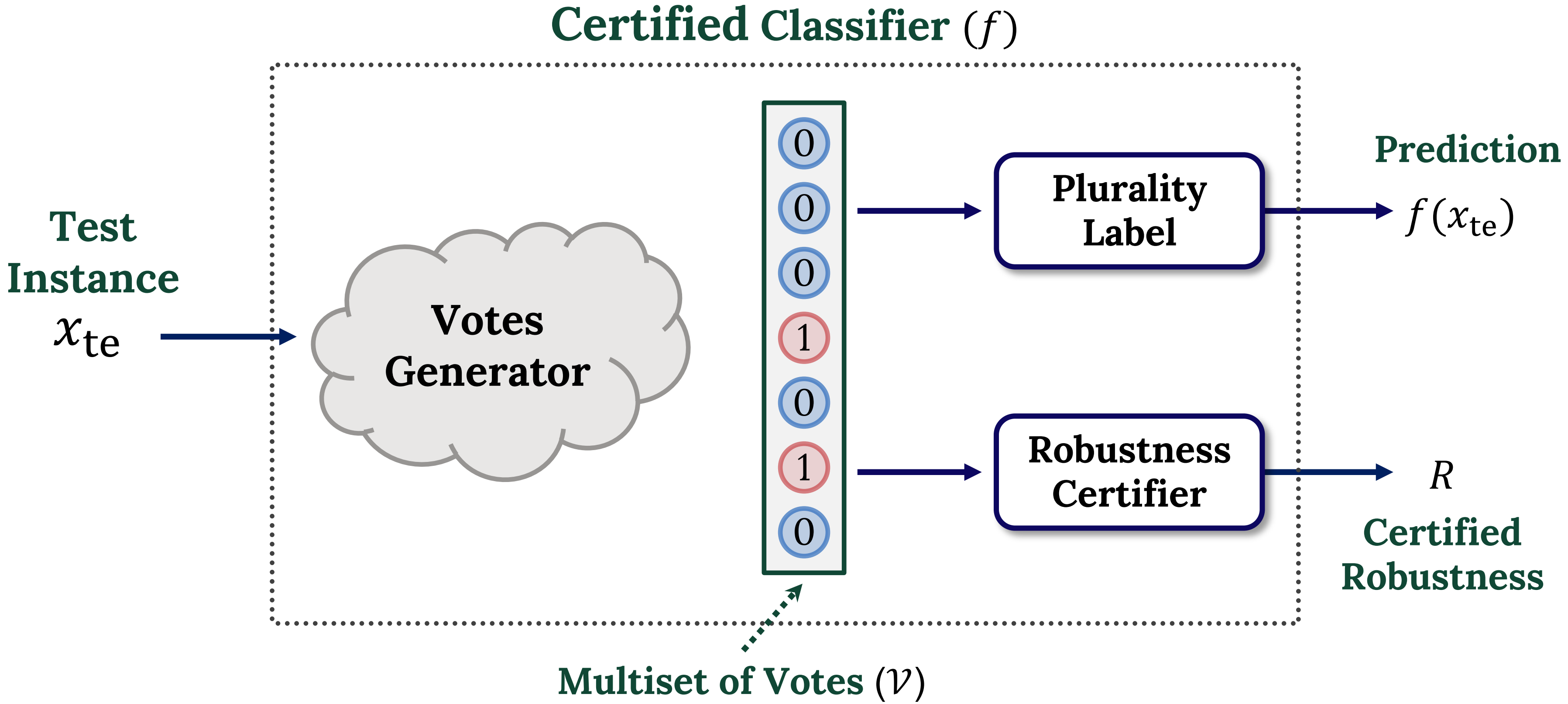
General Structure of a Certified Poisoning Classifier



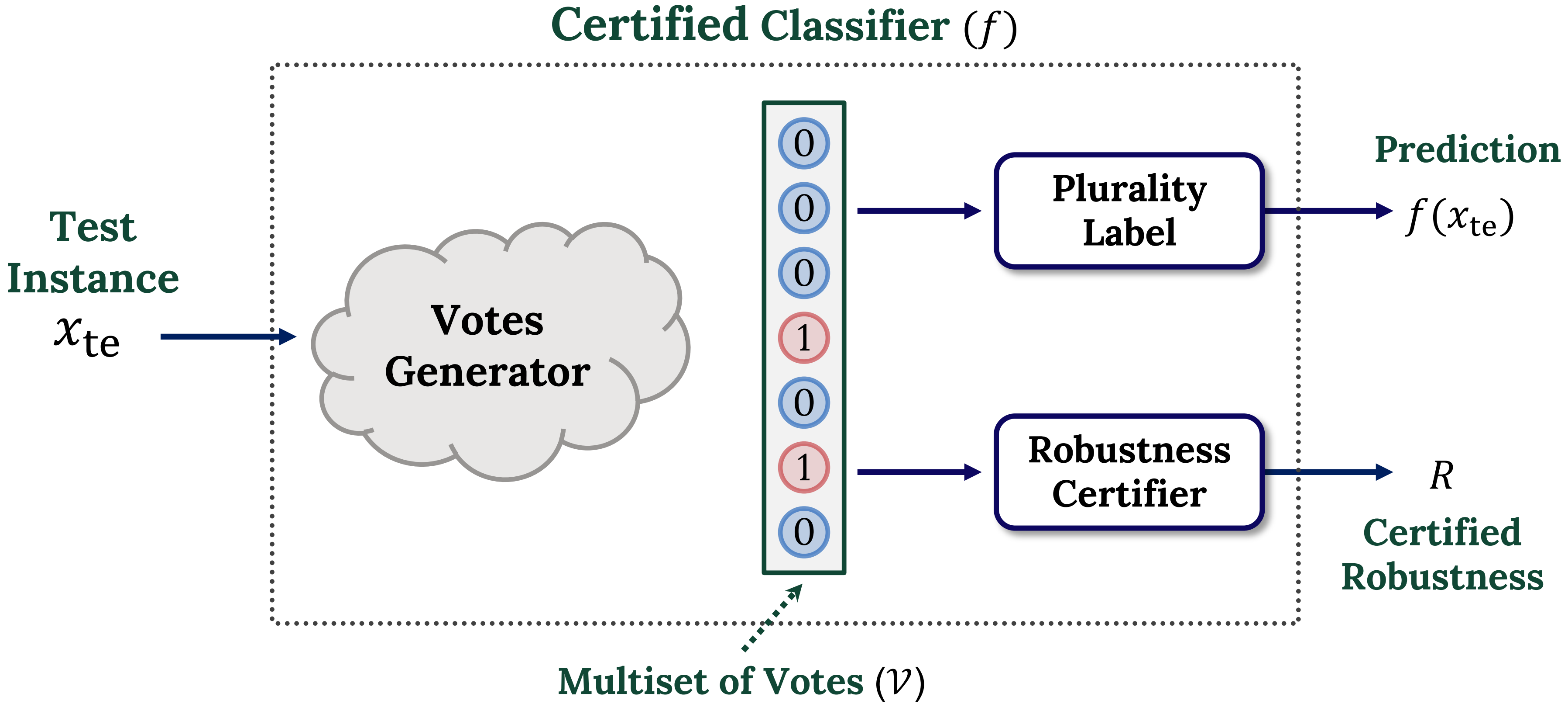
General Structure of a Certified Poisoning Classifier



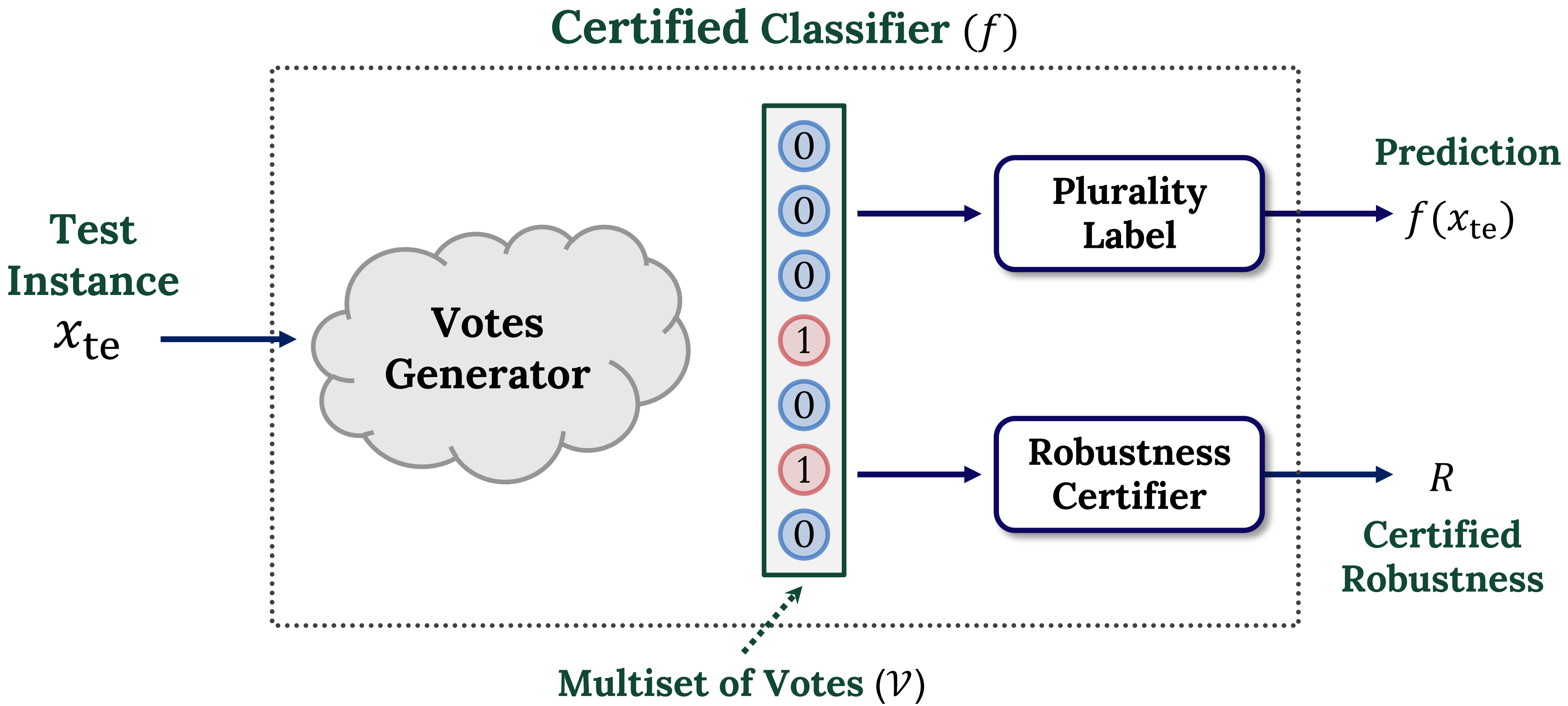
General Structure of a Certified Poisoning Classifier



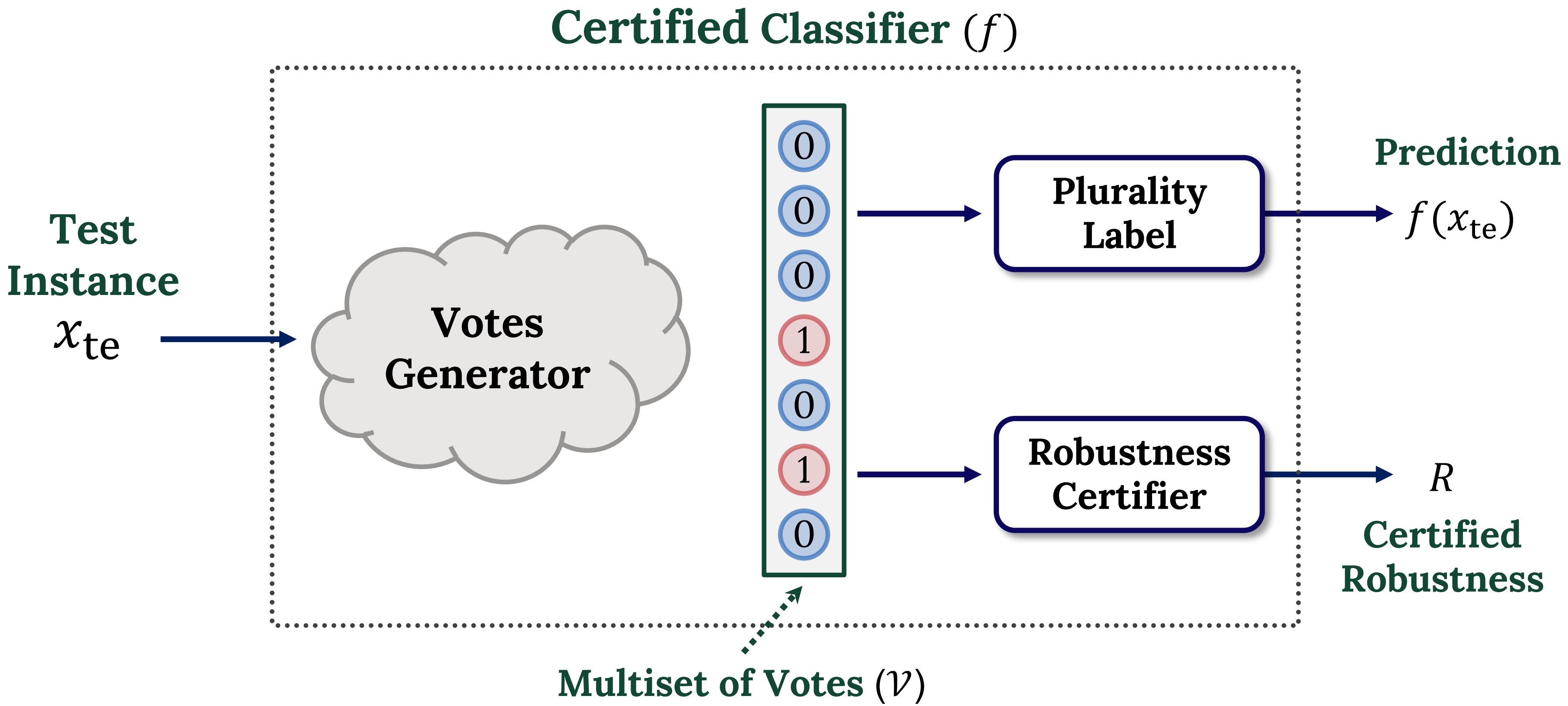
General Structure of a Certified Poisoning Classifier



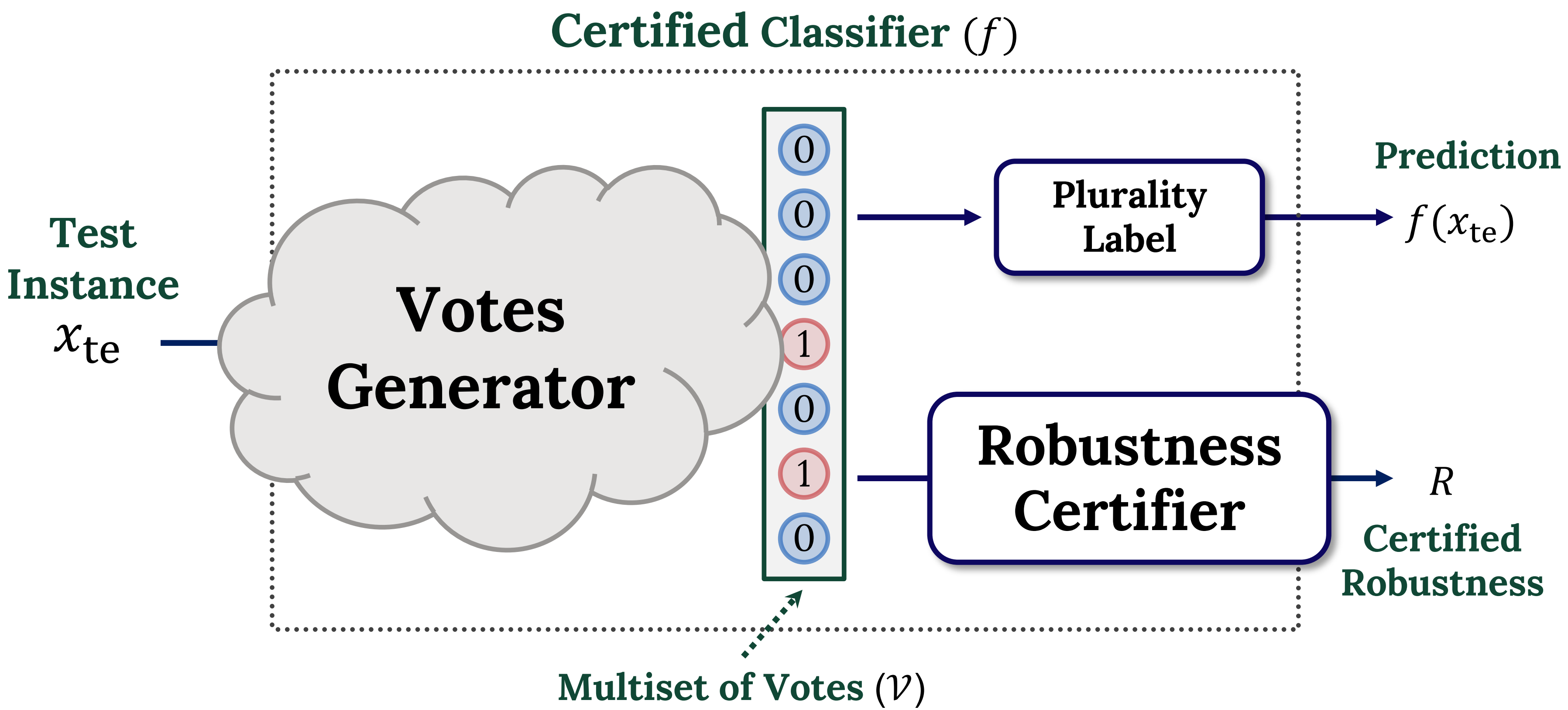
General Structure of a Certified Poisoning Classifier



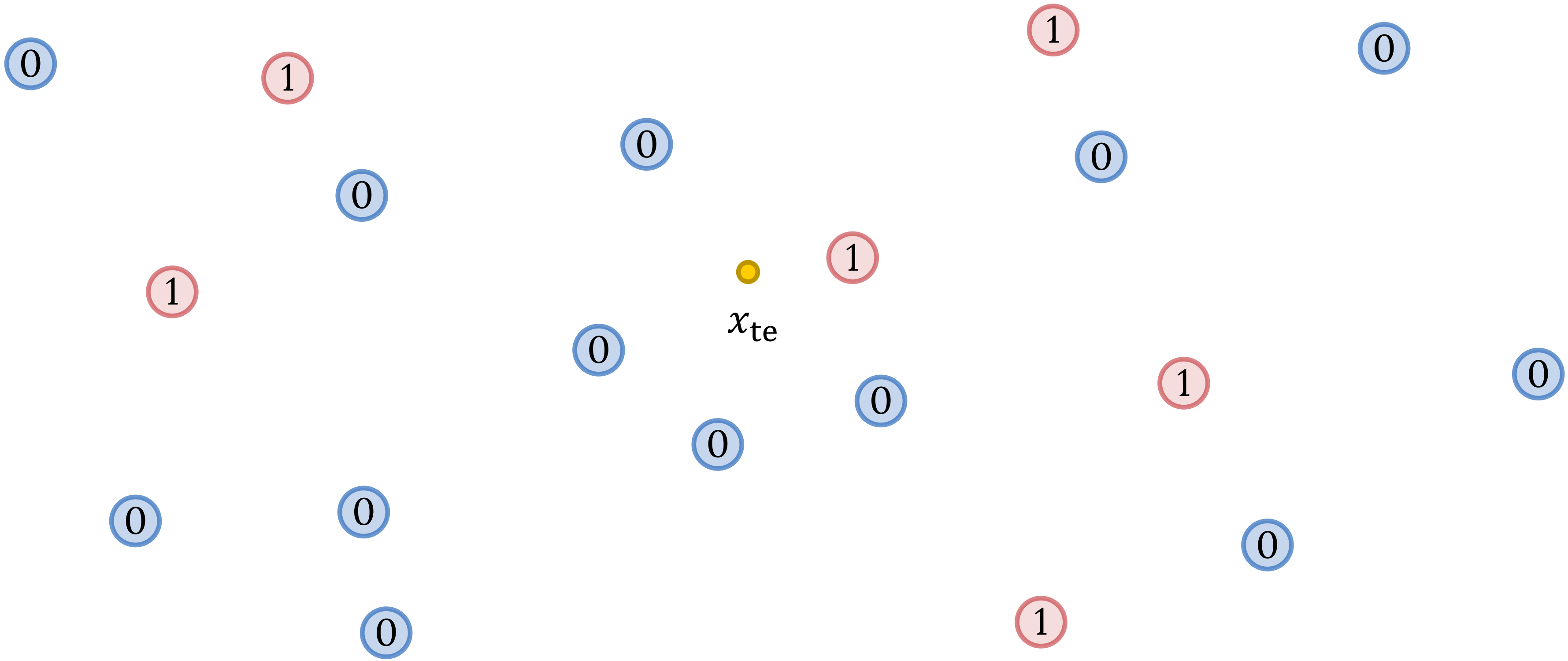
General Structure of a Certified Poisoning Classifier



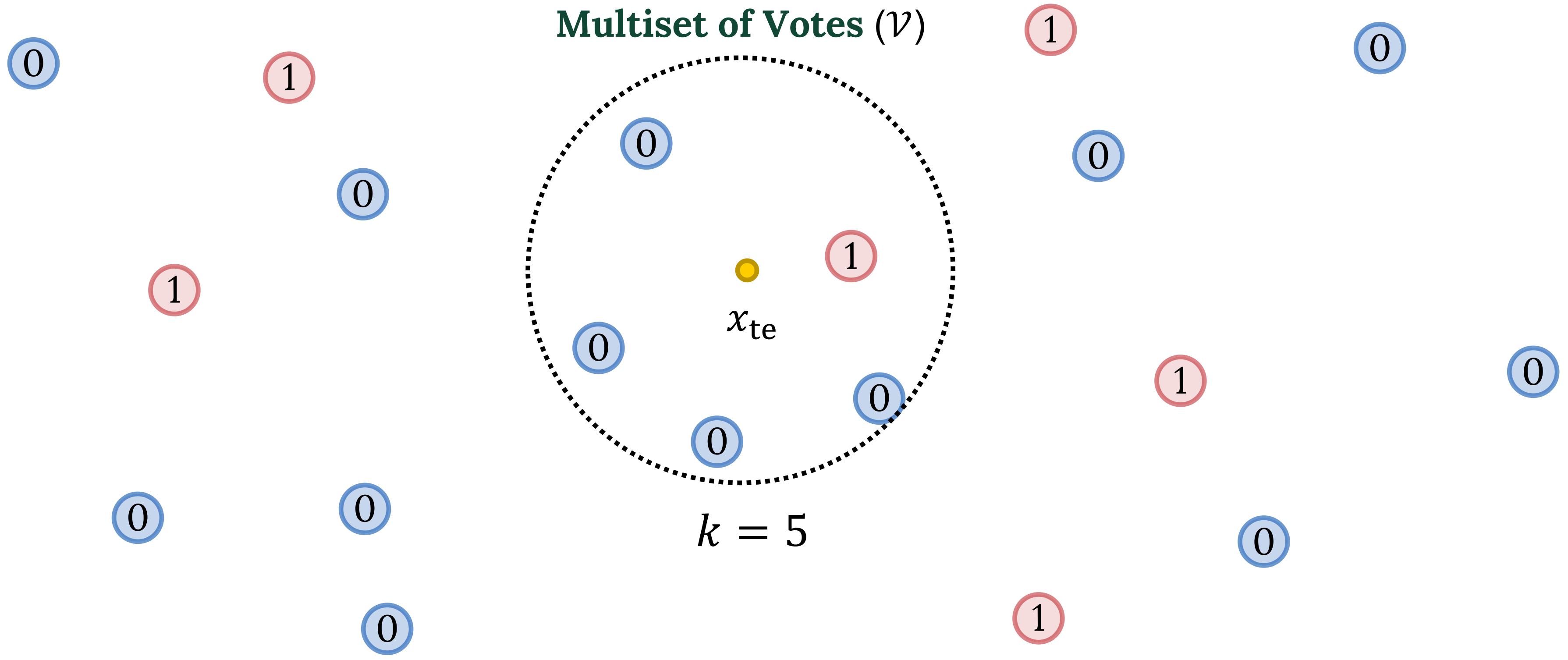
General Structure of a Certified Poisoning Classifier



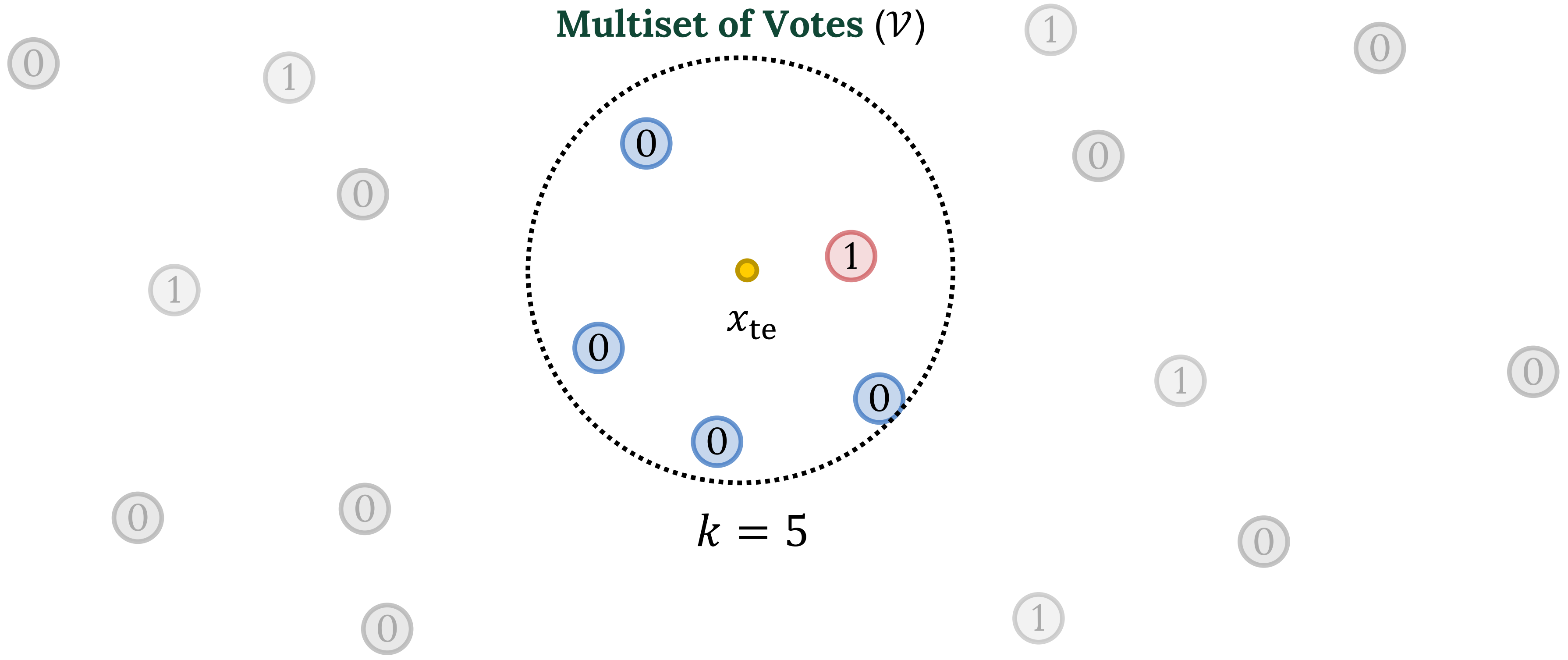
Example: kNN Certified Classifier [Jia+22]



Example: kNN Certified Classifier [Jia+22]



Example: kNN Certified Classifier [Jia+22]



Example: kNN Certified Classifier [Jia+22]

0 0 0 0 1

Vote Distribution:

- 4 votes label 0
- 1 vote label 1

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

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

Example: Certified Ensemble Classifier [LF21]

Example: Certified Ensemble Classifier [LF21]

Ensemble

f_1

f_2

f_3

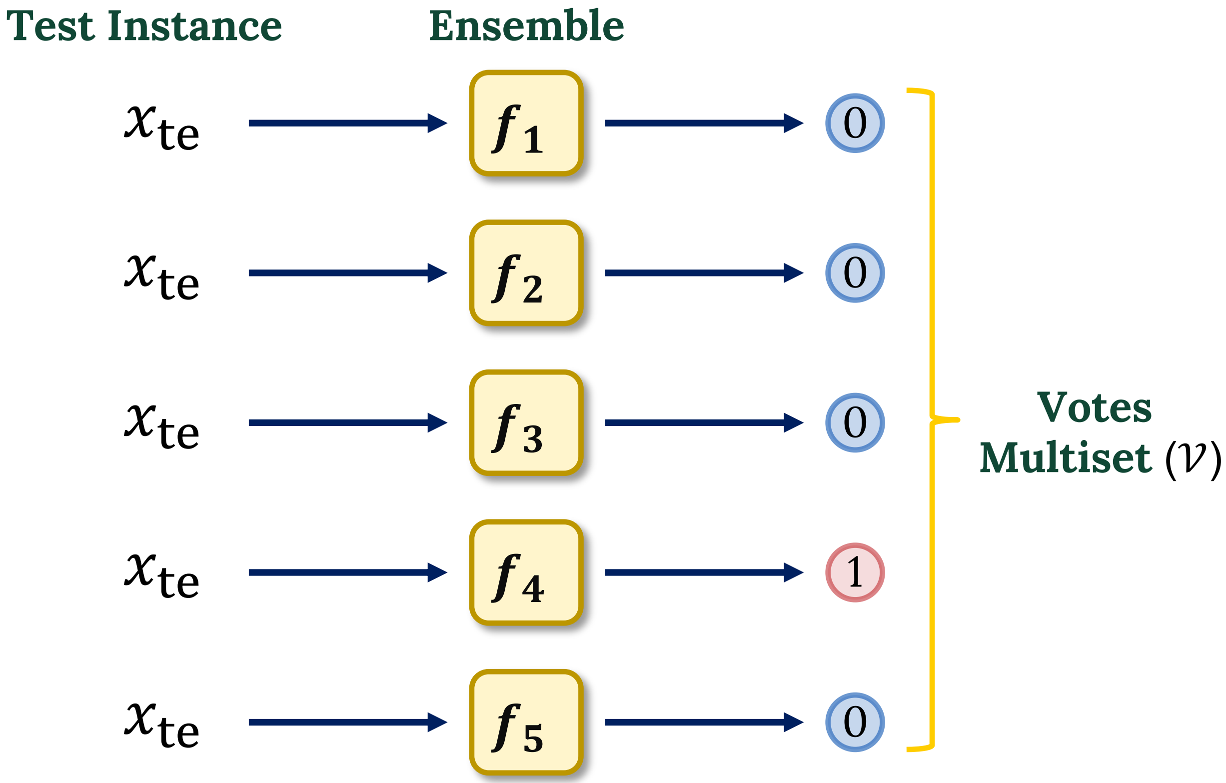
f_4

f_5

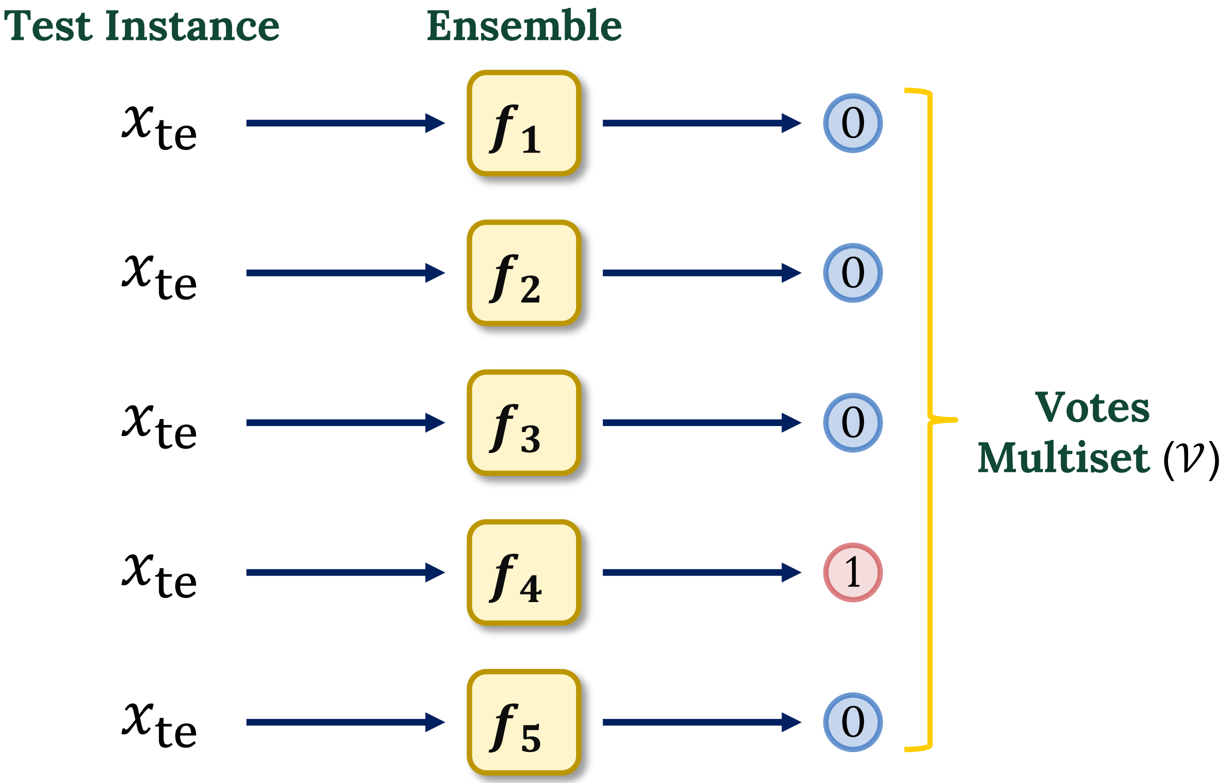
Example: Certified Ensemble Classifier [LF21]

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]

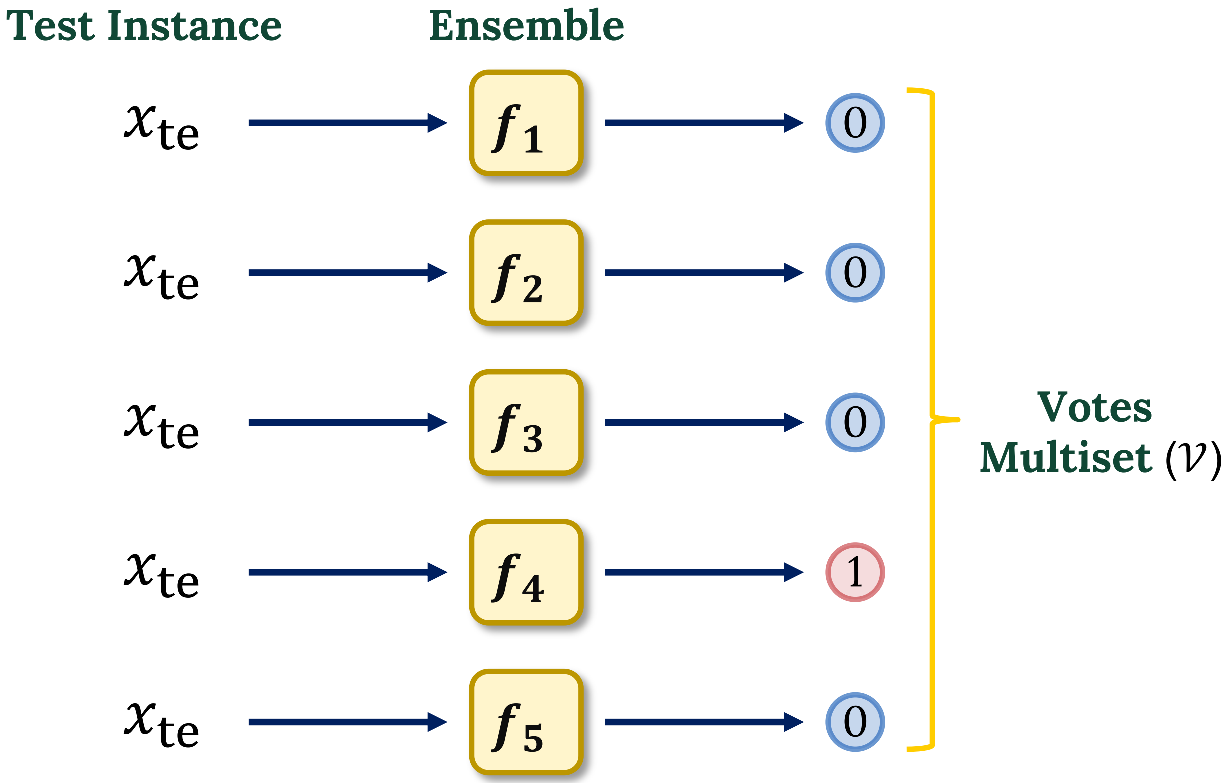


Example: Certified Ensemble Classifier [LF21]



Robustness Certification

Example: Certified Ensemble Classifier [LF21]

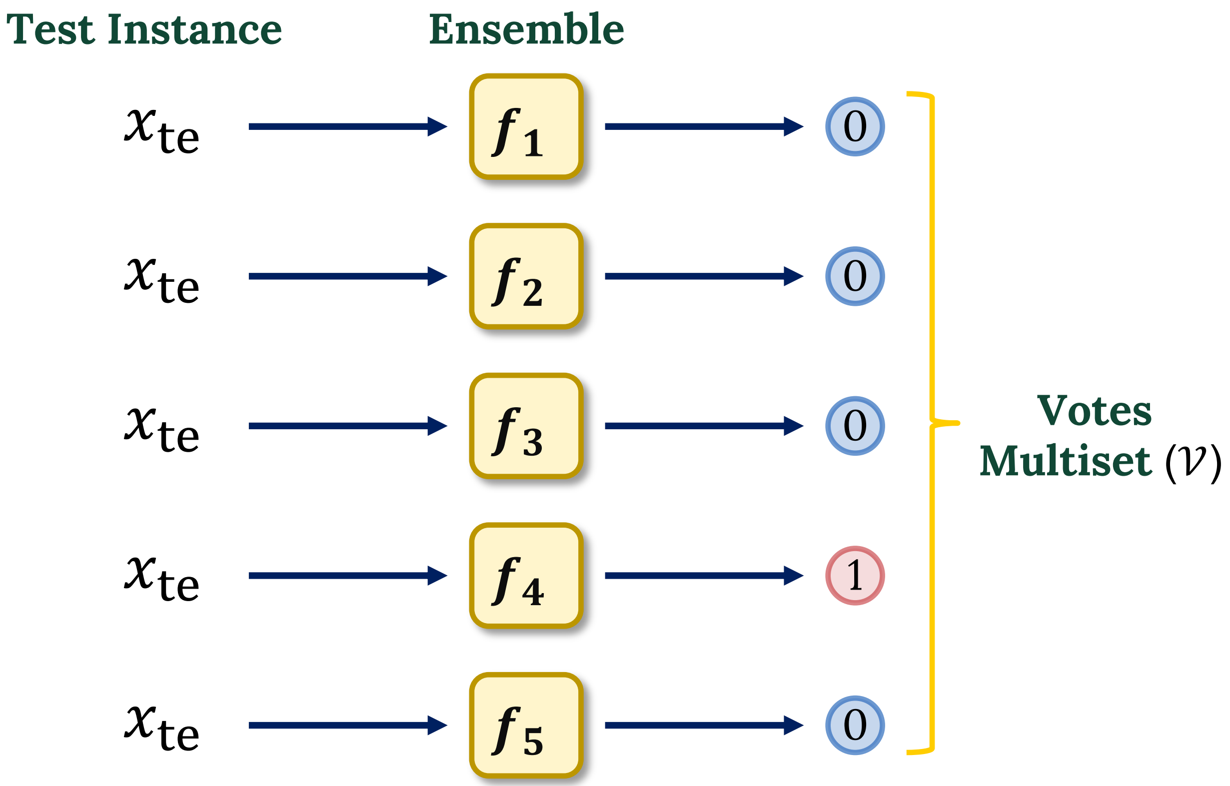


Robustness Certification

Vote Distribution:

- 4 votes label 0
- 1 vote label 1

Example: Certified Ensemble Classifier [LF21]



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 regressionEach improved certified classifier improves certified

Key Insight of the Reduction

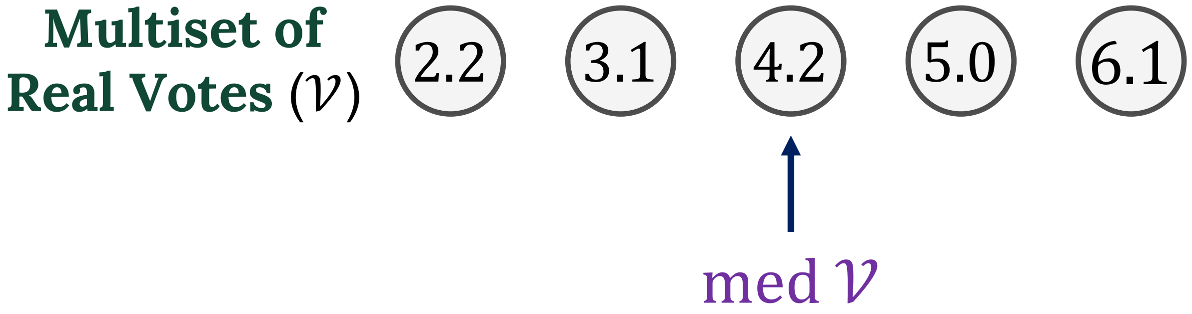
For any binary multiset,
the **plurality label** and **median** have
equivalent robustness

Relating Real-Valued and Binary Robustness

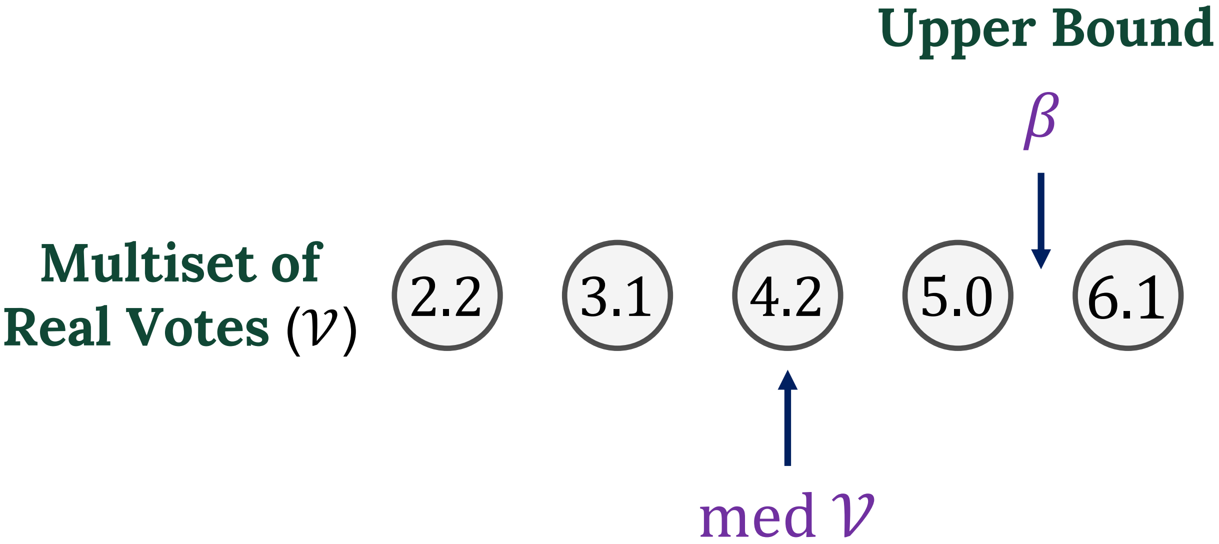
Multiset of Real Votes ($\mathcal{V}$)

2.2 3.1 4.2 5.0 6.1

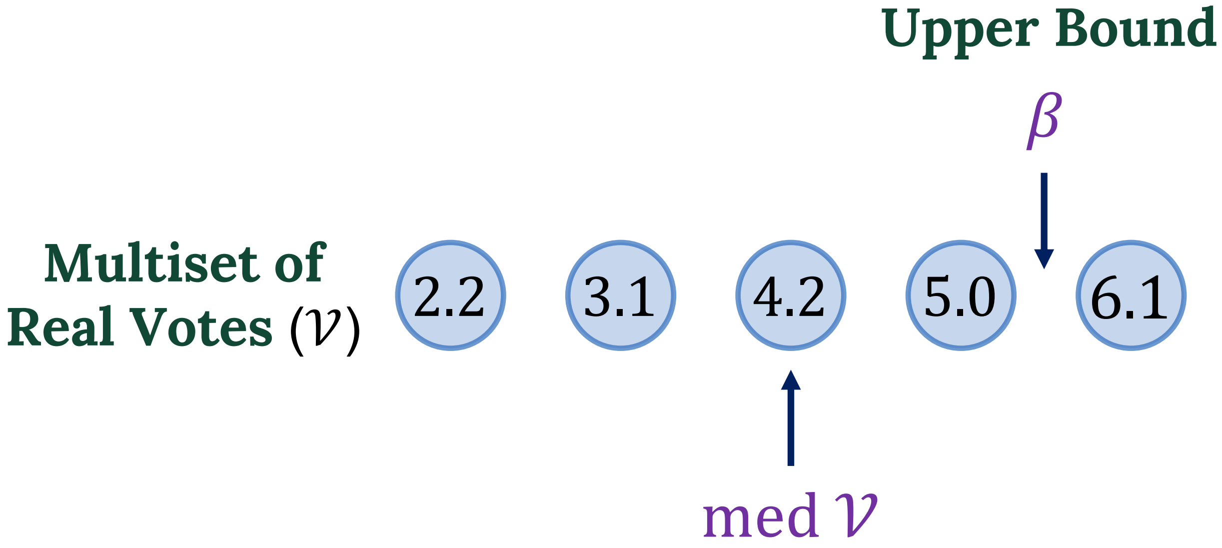
Relating Real-Valued and Binary Robustness



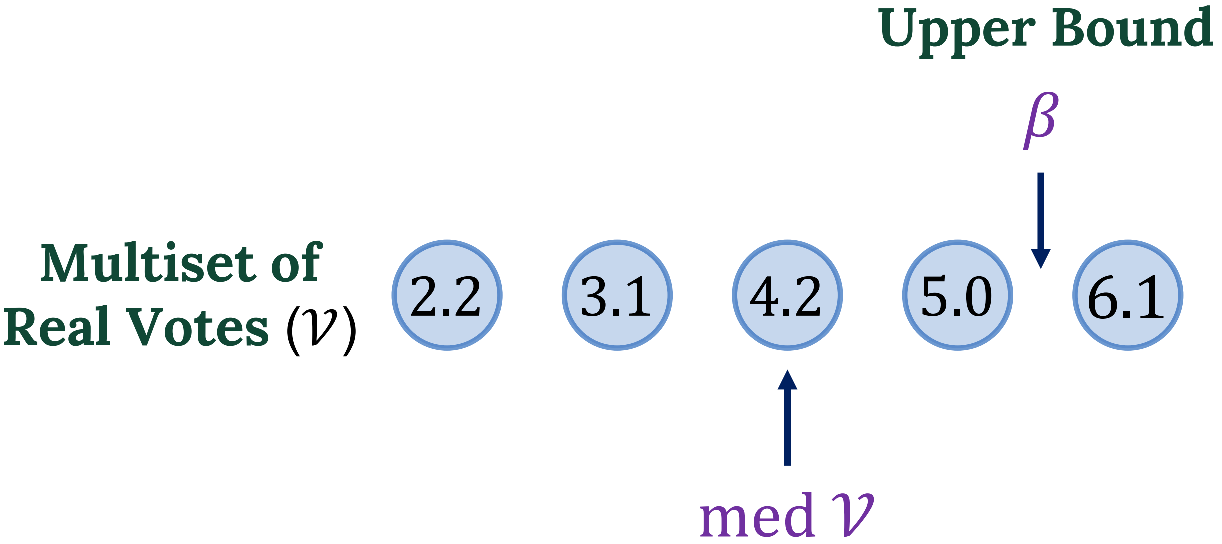
Relating Real-Valued and Binary Robustness



Relating Real-Valued and Binary Robustness

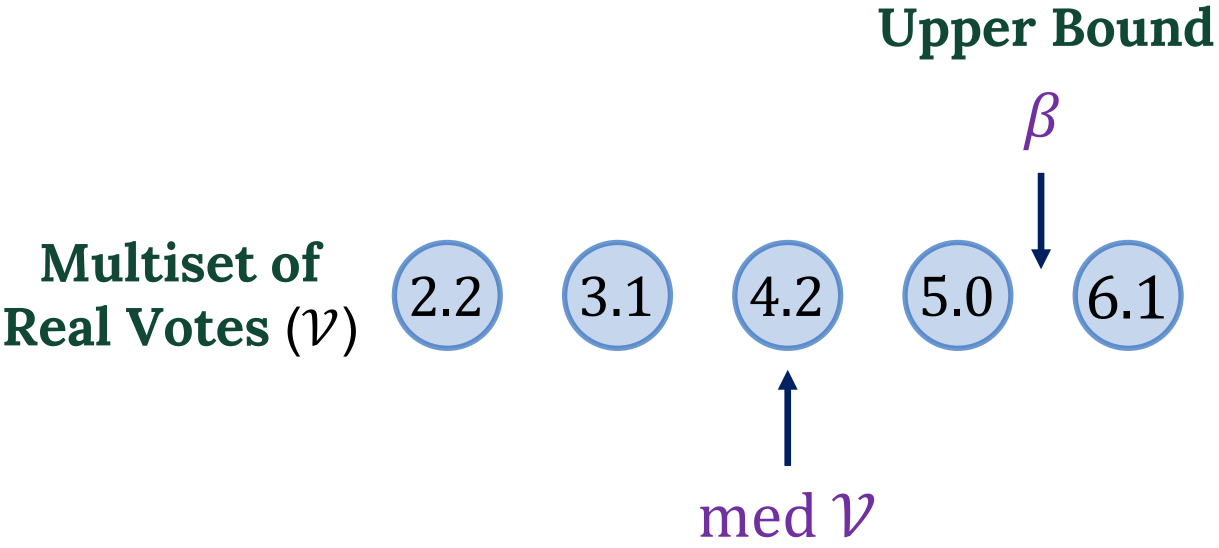


Relating Real-Valued and Binary Robustness



Robustness Certifier

Relating Real-Valued and Binary Robustness



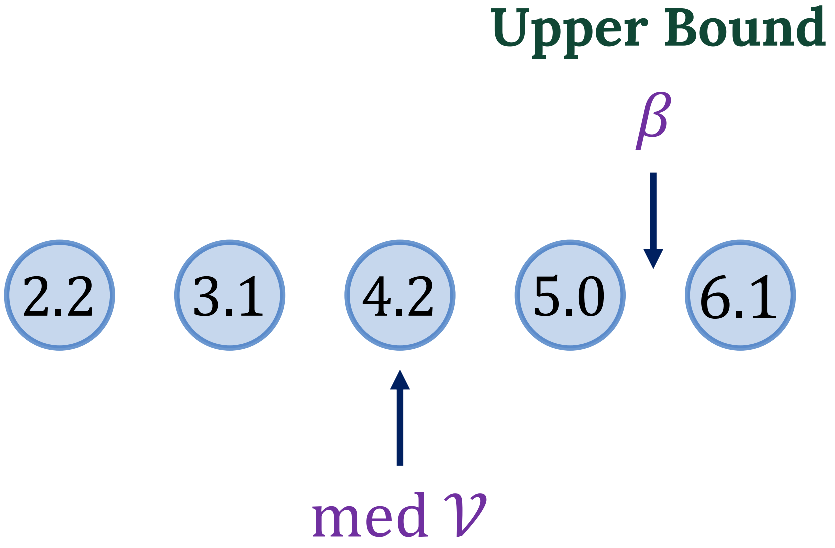
Robustness Certifier

Vote Distribution:

- 4 votes label 0
- 1 vote label 1

Relating Real-Valued and Binary Robustness

Multiset of Real Votes ($\mathcal{V}$)



Robustness Certifier

Vote Distribution:

- 4 votes label 0
- 1 vote label 1

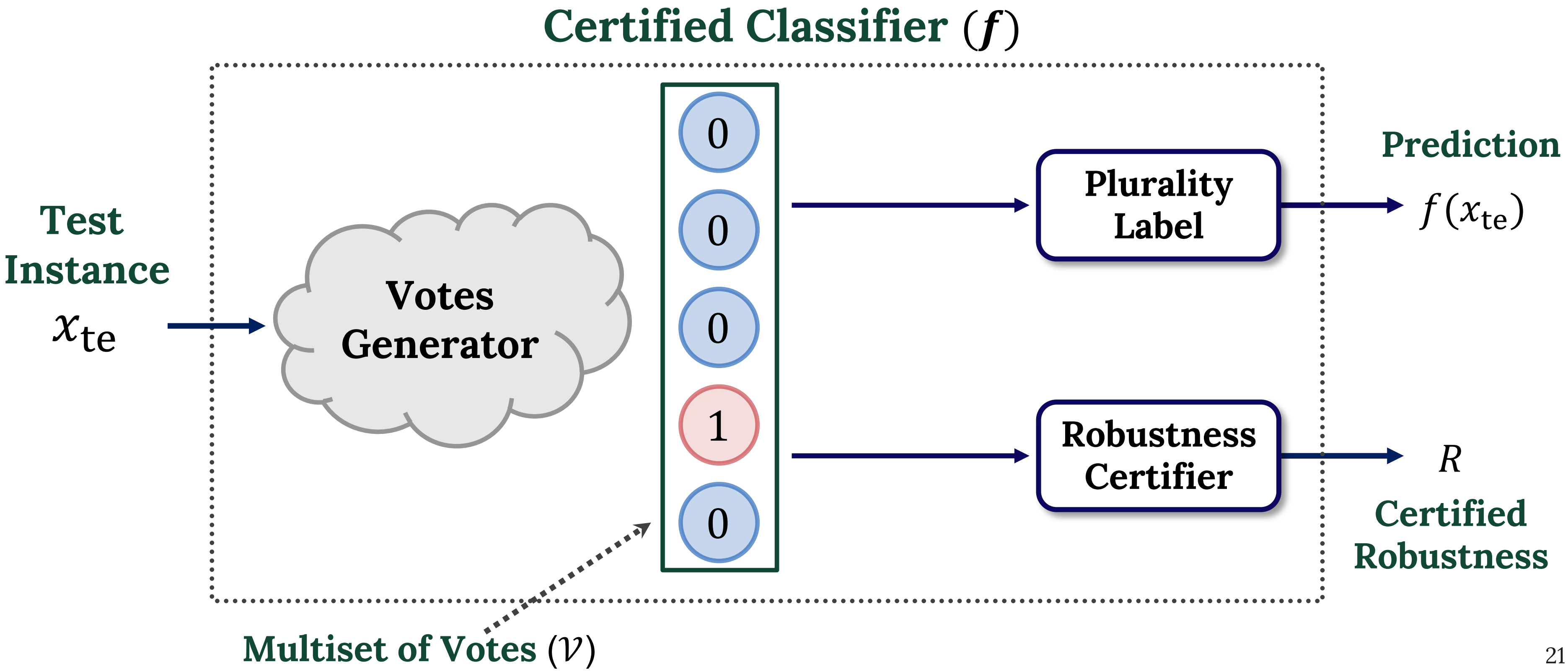
Certified Robustness:

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

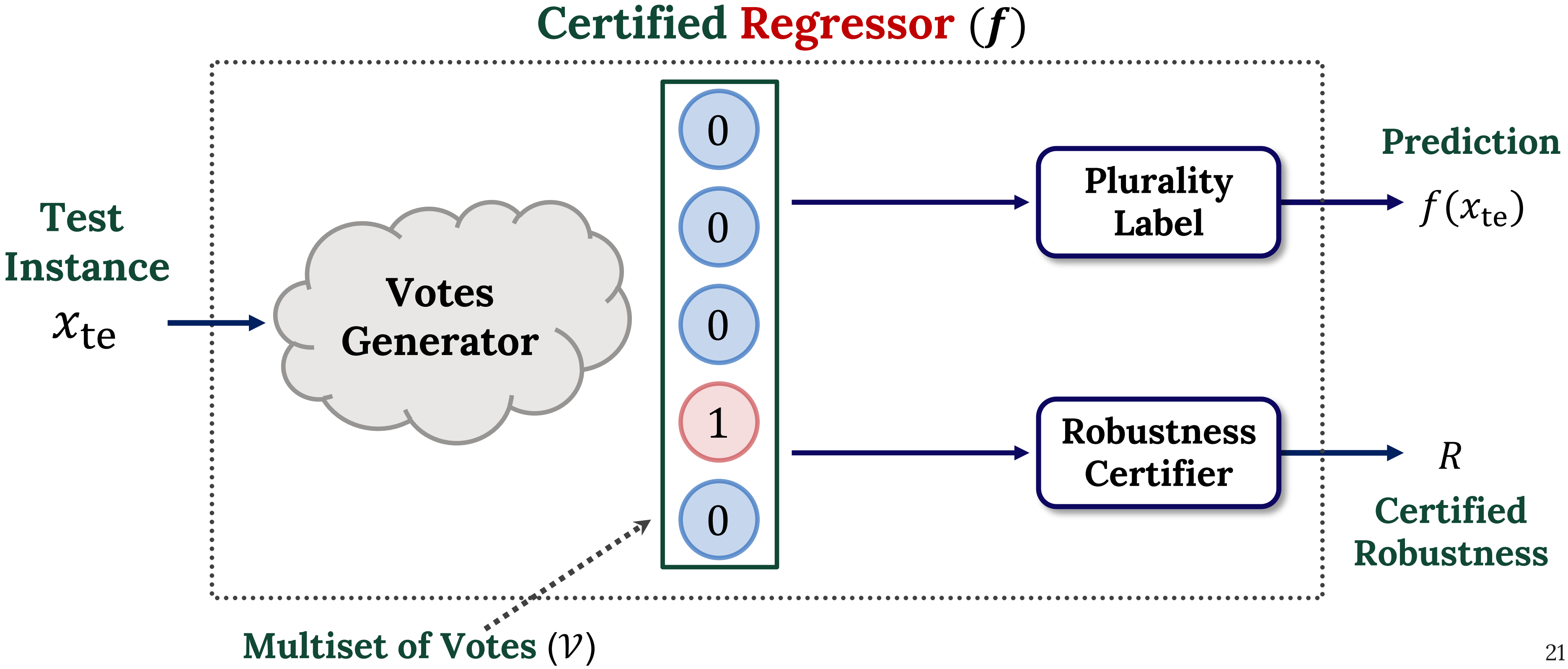
“Transform certified poisoning classifiers into certified regressors”

UNIVERSITY OF OREGON

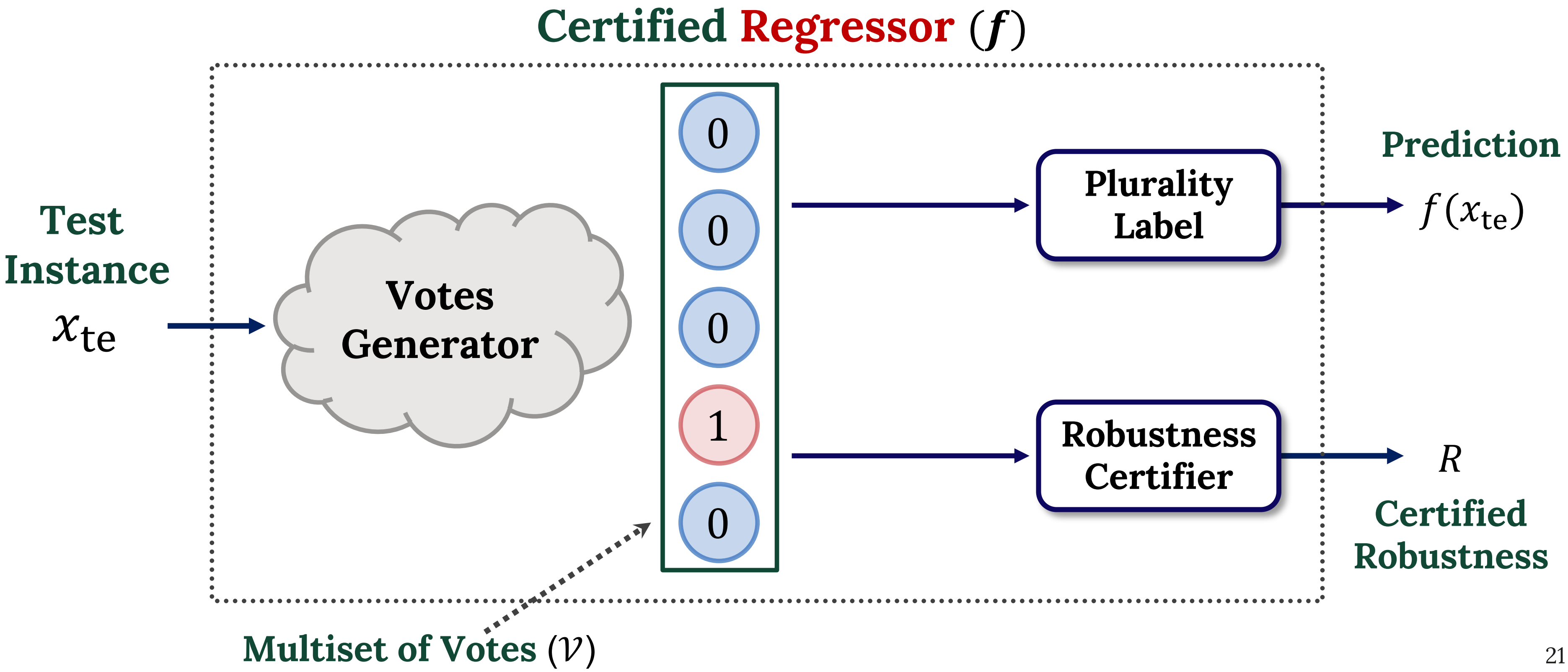
General Structure of a Certified Poisoning Classifier



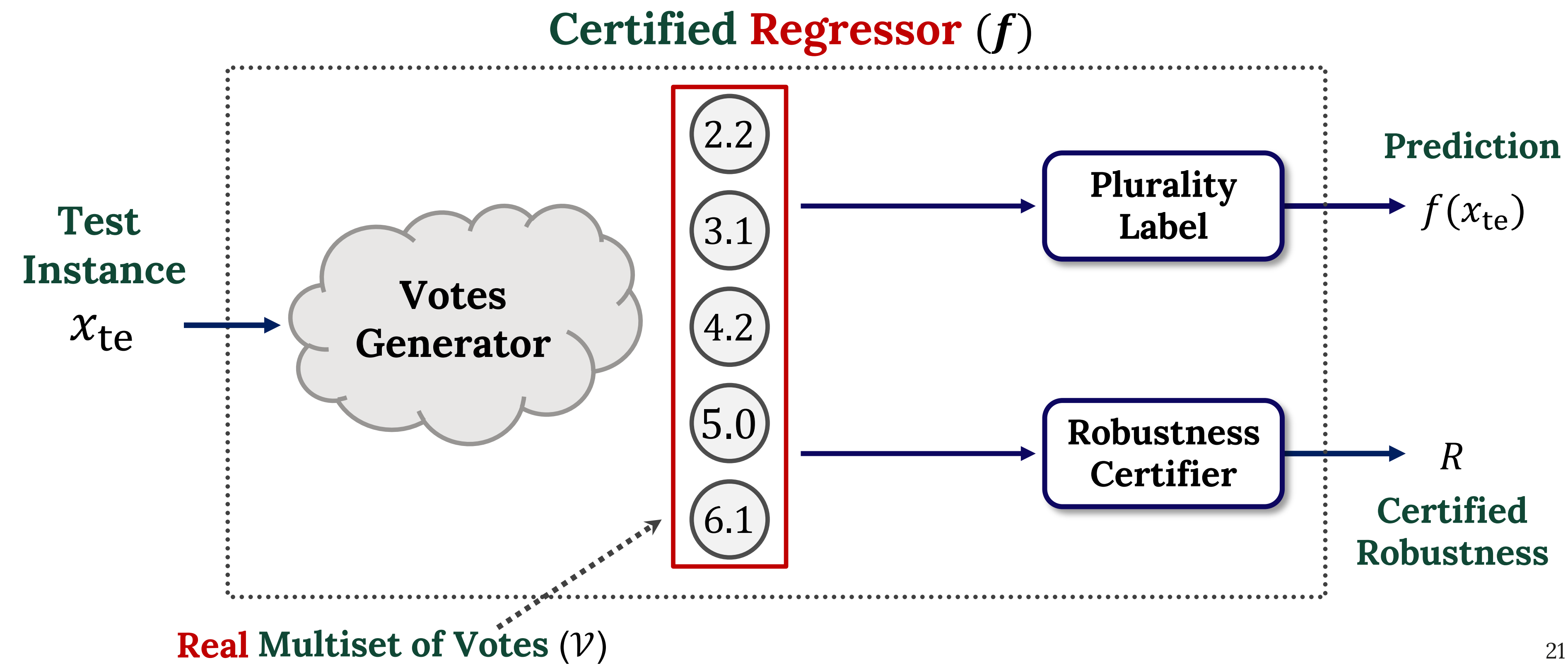
General Structure of a Certified Poisoning Regressor



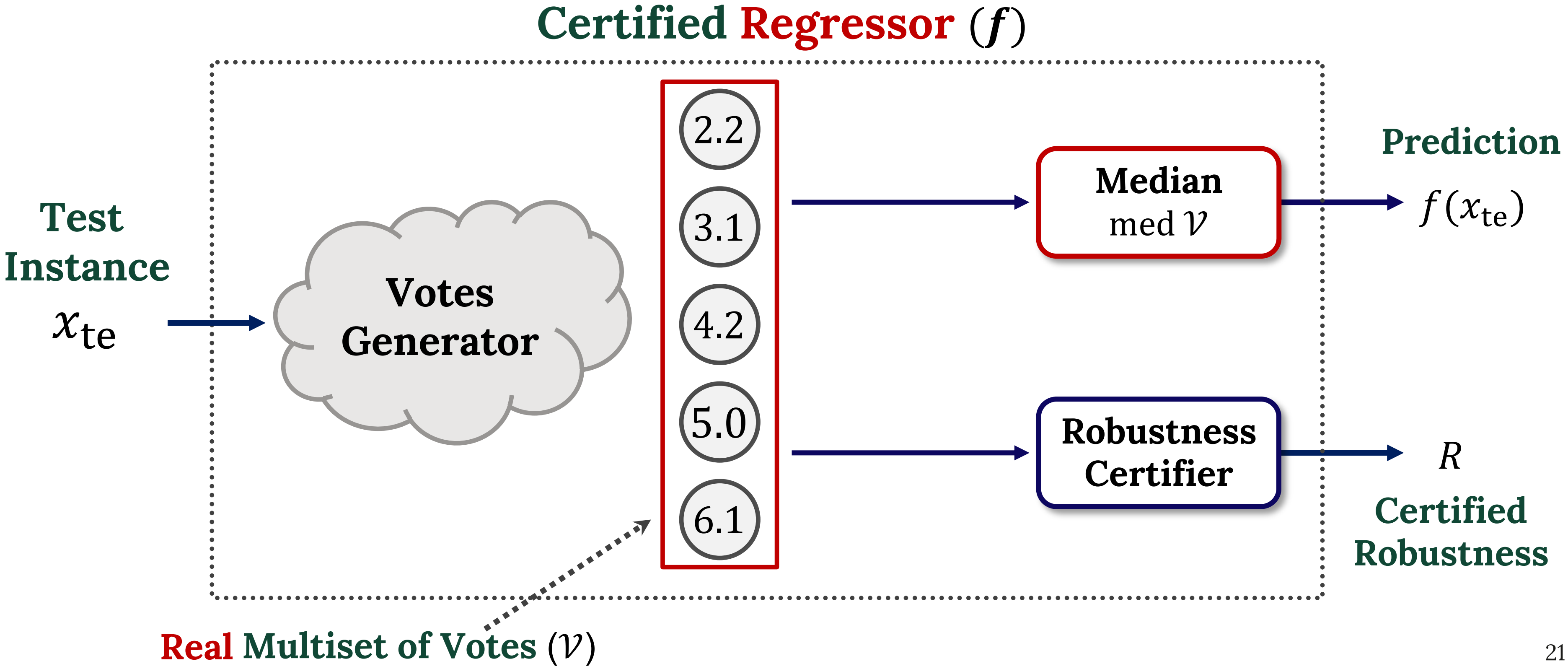
General Structure of a Certified Poisoning Regressor



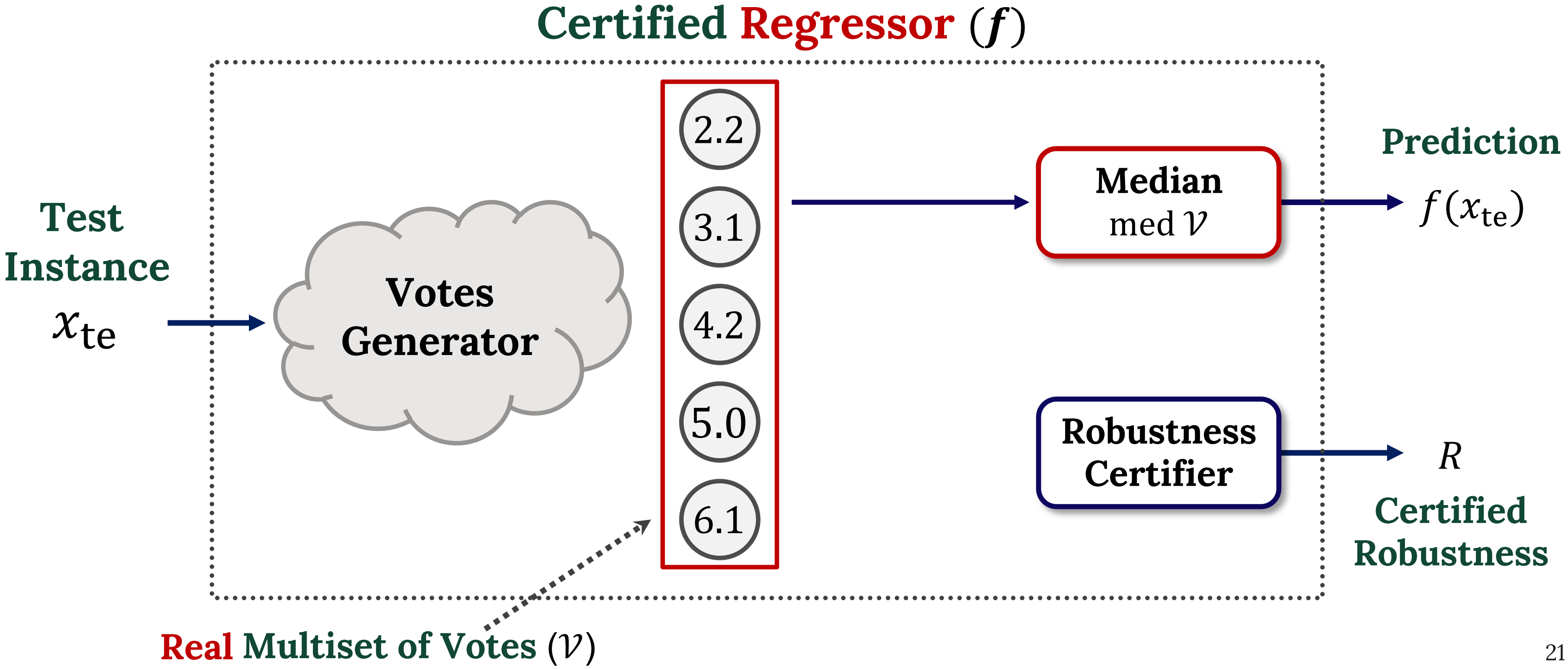
General Structure of a Certified Poisoning Regressor



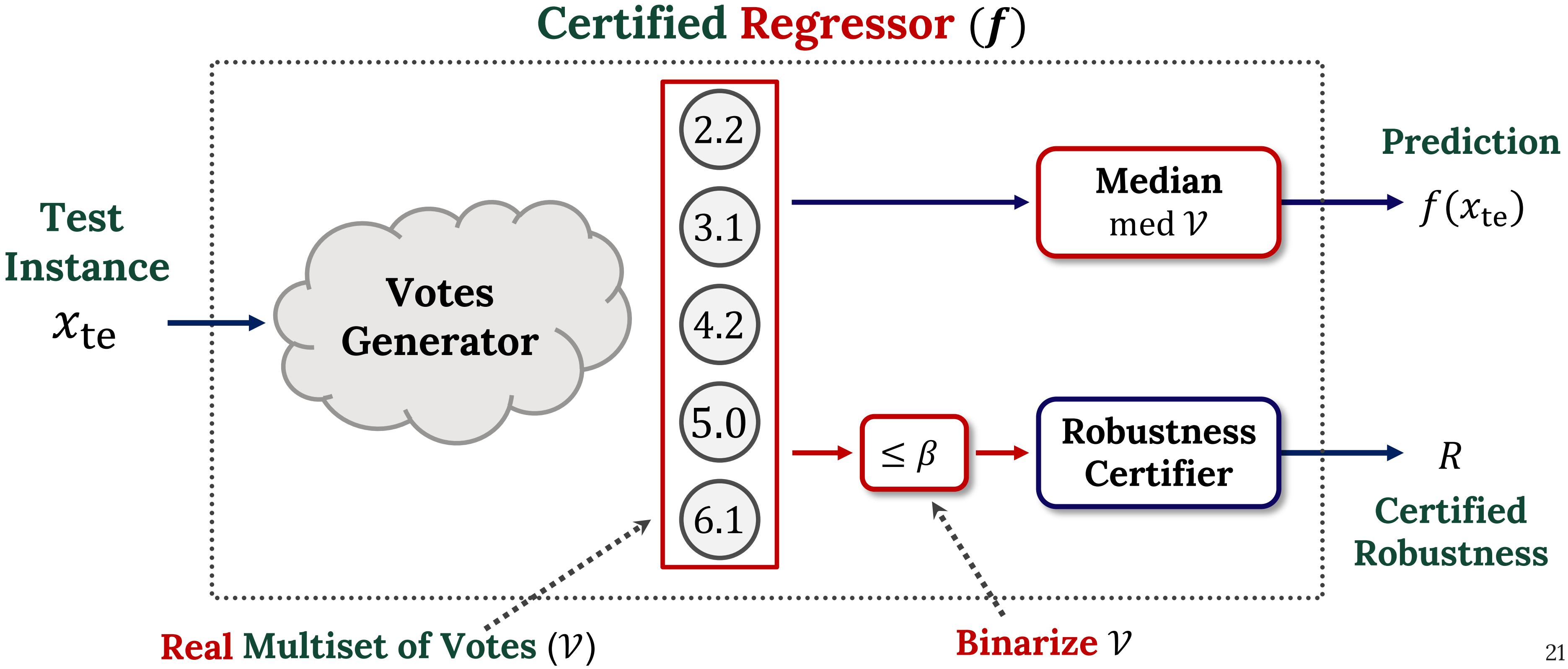
General Structure of a Certified Poisoning Regressor



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

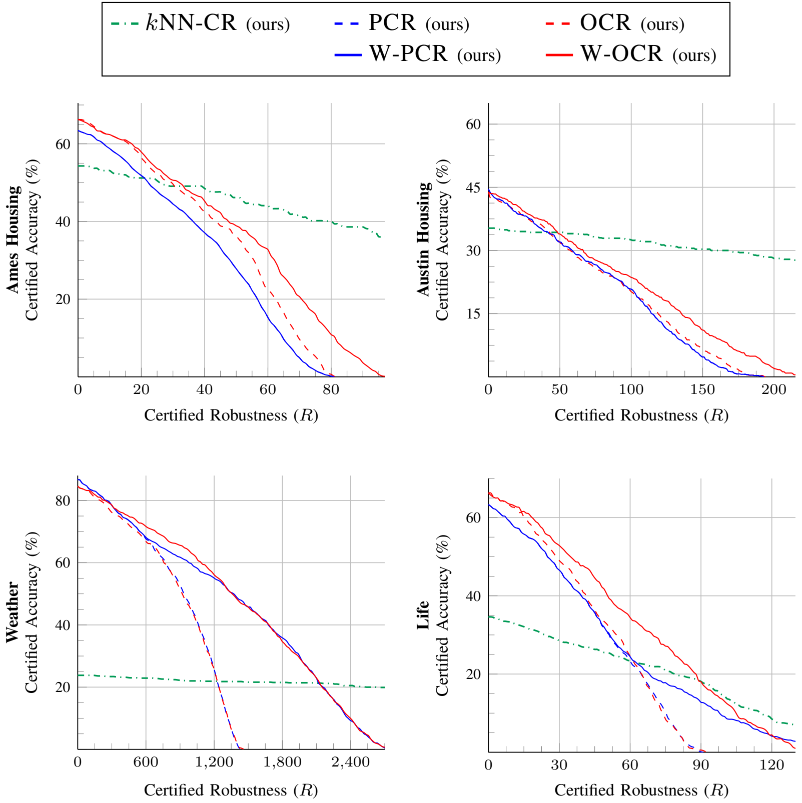
Empirical Evaluation

- **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.