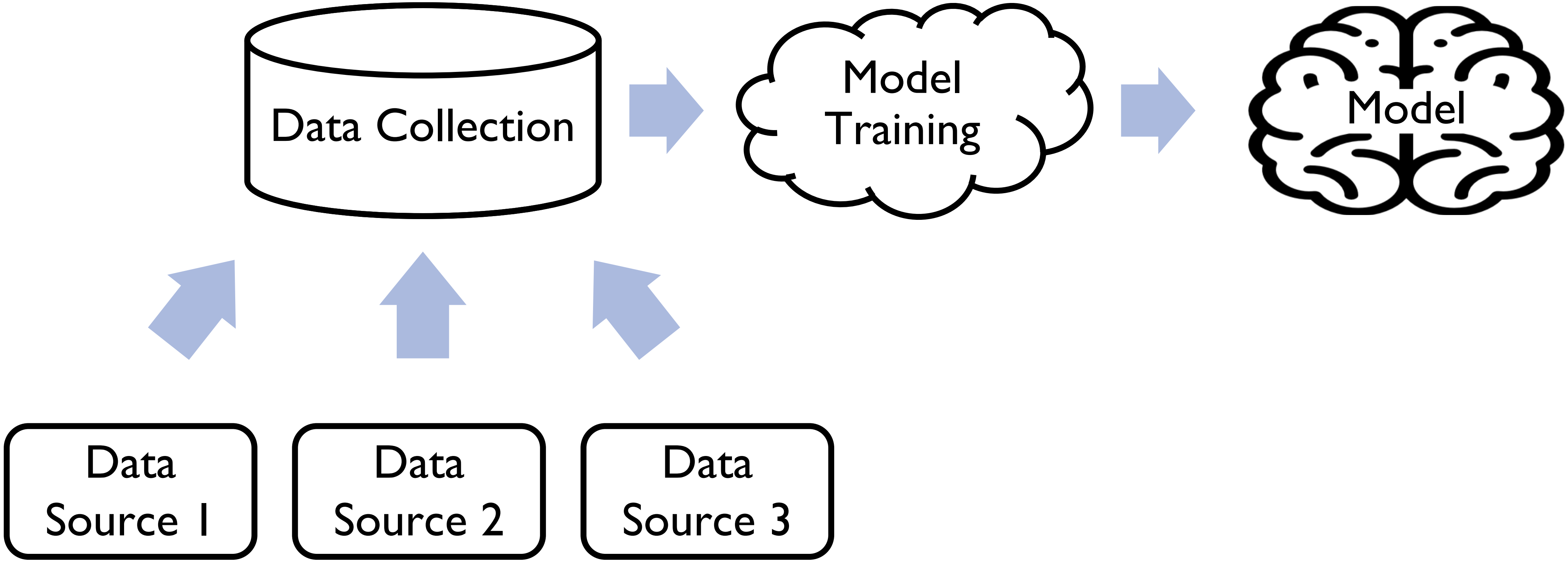
NEURAL INFORMATION PROCESSING SYSTEMS

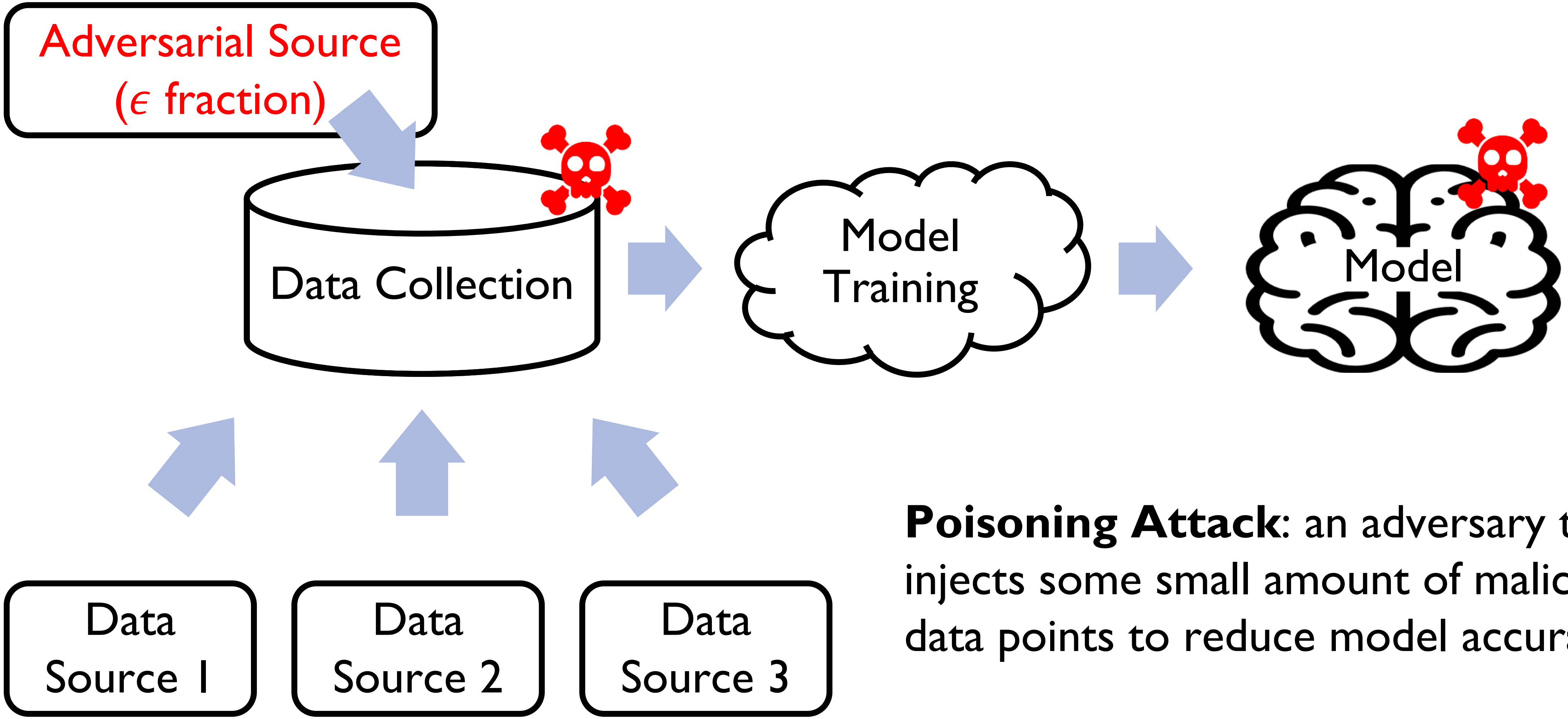
What Distributions are Robust to Indiscriminate Poisoning Attacks for Linear Learners?

Fnu Suya, Xiao Zhang, Yuan Tian, David Evans

Machine Learning Pipeline

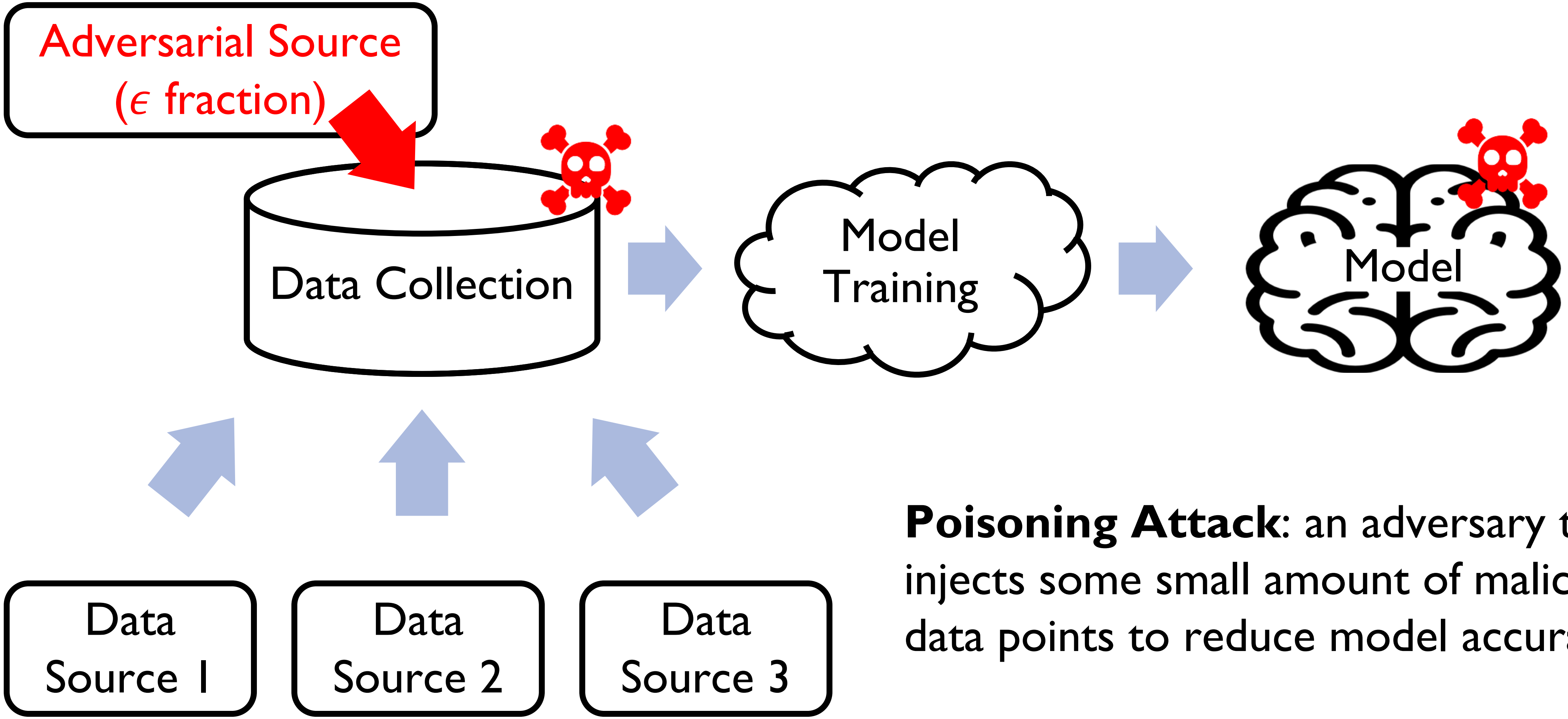


Indiscriminate Poisoning Attacks



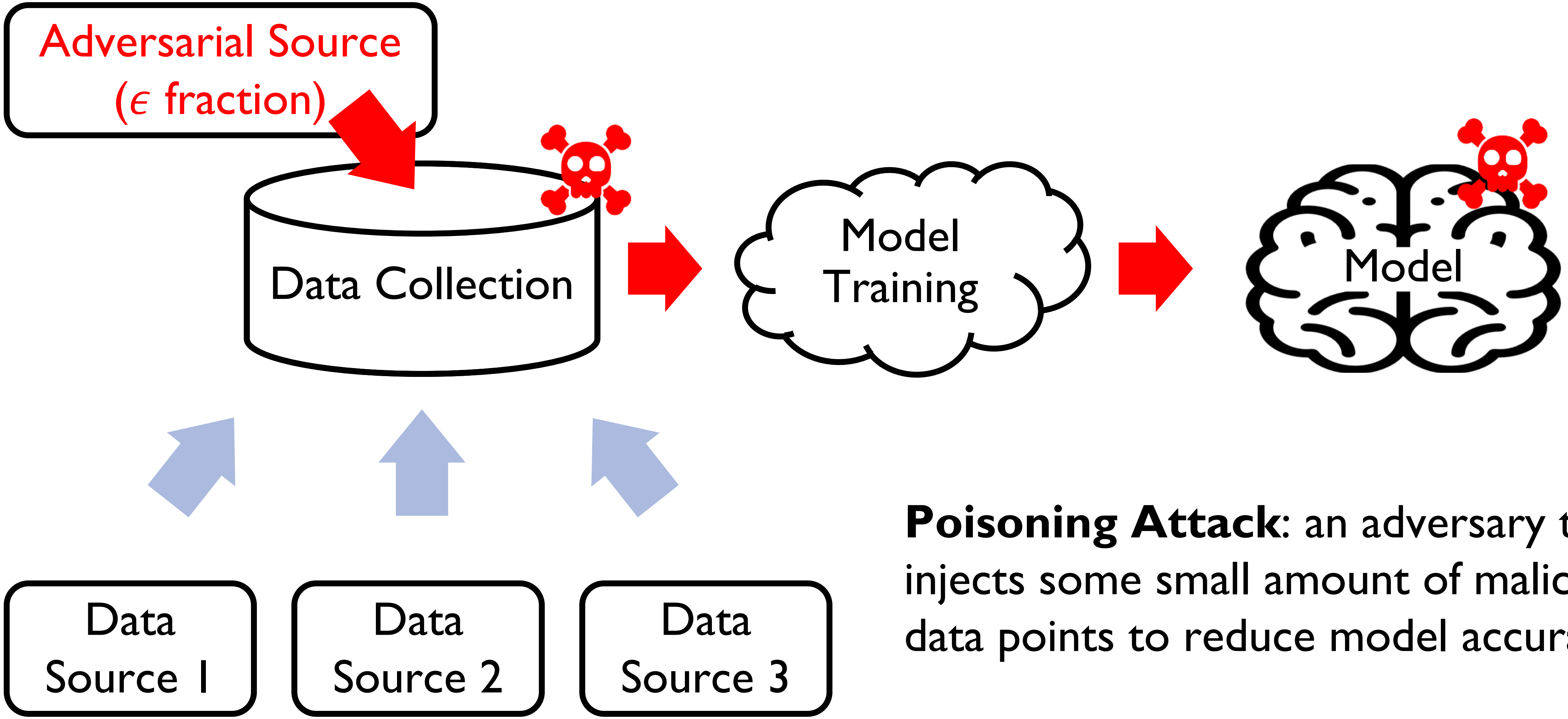
Poisoning Attack: an adversary that injects some small amount of malicious data points to reduce model accuracy

Indiscriminate Poisoning Attacks



Poisoning Attack: an adversary that injects some small amount of malicious data points to reduce model accuracy

Indiscriminate Poisoning Attacks



Poisoning Attack: an adversary that injects some small amount of malicious data points to reduce model accuracy

Diverse Indiscriminate Attacks

Attacks:
KKT [1], MTP [2], Min-Max [1,3], Influence [1]

Threat Model

known training algorithm, clean train/test data, model architecture, possible defenses

with defenses, effective on some datasets while ineffective on others

Are these attacks always effective without defenses?

[1]: Koh et al., "Stronger Data Poisoning Attacks Break Data Sanitization Defenses", Machine Learning 2021.
[2]: Suya et al., "Model-Targeted Poisoning Attacks with Provable Convergence", ICML 2021
[3]: Steinhard et al., "Certified Defenses Against Data Poisoning Attacks", N(eur)IPS 2017

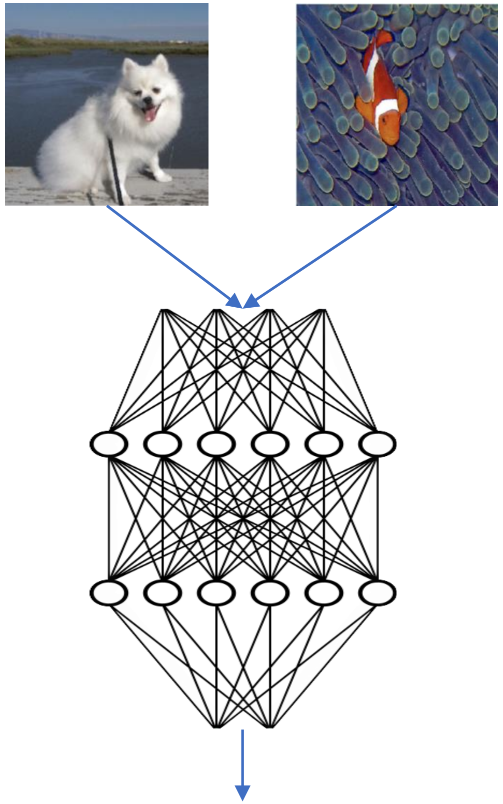
Evaluation without Defenses

Datasets to train Linear SVM

MNIST 1-7


Digits of "1" and "7" from MNIST

Dogfish

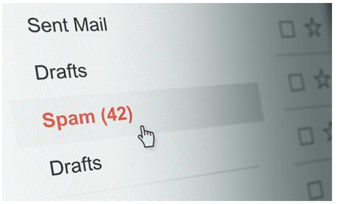


Features

Enron

Collection of Spam emails

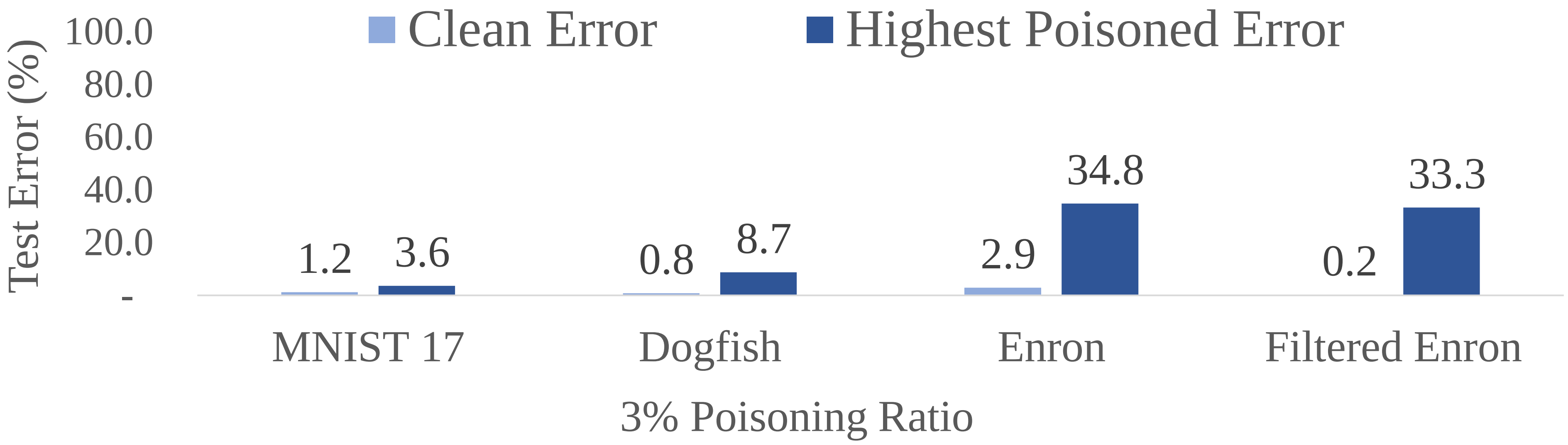
Filtered Enron



Filter our 3% near boundary points from **Enron**

Results of more models and datasets are in the paper.

Best Attack Effectiveness Varies



Are some datasets (e.g., MNIST 1-7) just robust to state-of-the-art poisoning attacks or inherently robust to any poisoning attacks?

Understanding Optimal Attacks

Theorem 1 (Informal): optimal finite-sample poisoning attacks are consistent estimators of optimal distributional poisoning attacks if:

1) hypothesis class satisfies uniform convergence

2) surrogate loss for model training is strongly-convex

3) risk of the model is Lipchitz continuous.

Finite-sample optimal poisoning attacks (practice): relevant to practical applications	Sufficient Samples $\approx$	Distributional optimal poisoning attacks (theory): convenient for analysis
Generate poisoned dataset to maximize risk		Generate poisoned distribution to maximize risk

Useful to study distributional optimal attacks as they still connect to finite-sample attacks in practice!

Using Maximum Poisoning Ratio

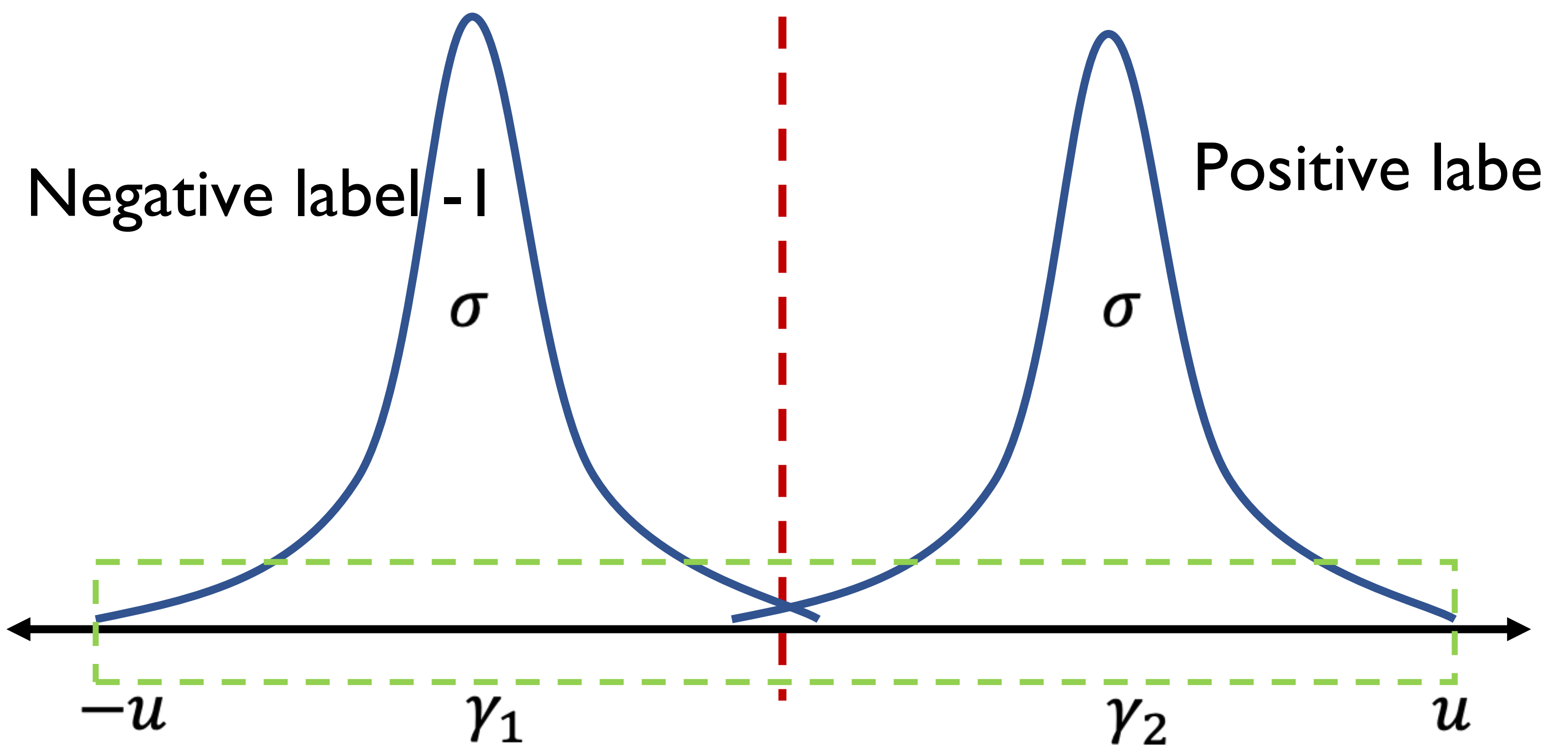
Theorem 2 (Informal): for convex hypothesis class, optimal distributional poisoning is achieved with maximum poisoning ratio ϵ if either condition is satisfied:

1) clean data points are not filtered during training

2) For any model $\boldsymbol{\theta}$, there is a distribution μ such that gradient w.r.t. μ is $\mathbf{0}$.

When studying distributional optimal poisoning attacks, we can use the maximum poisoning ratio!

Characterize Optimal Attacks in 1-D Gaussian



Linear SVM on1-D two Gaussian mixtures

Poisoning points are in constraint set [$-u, u$]
with **constraint size** $2u$.

Goal

Analyze the impact of distributional properties on optimal poisoning attacks that have maximum risk on clean distribution

Distributional Factors on Optimal Attack

Theorem 3 (Informal): distributions with smaller $|\gamma_1 - \gamma_2|/\sigma$ (separability ratio) and larger $2u$ (larger constraint size) are inherently more vulnerable to poisoning attacks and vice-versa.

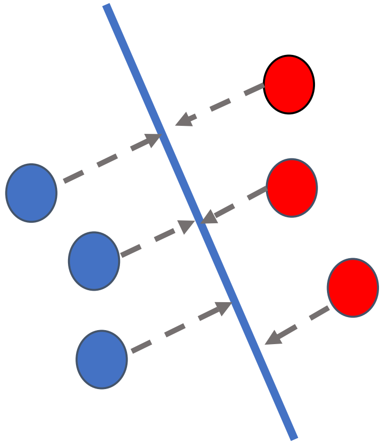
$|\gamma_1 - \gamma_2|/\sigma$: small ratio implies more near-boundary points and more prone to misclassifications

Larger constraint size $2u$: moves the decision boundary more with poisoning points

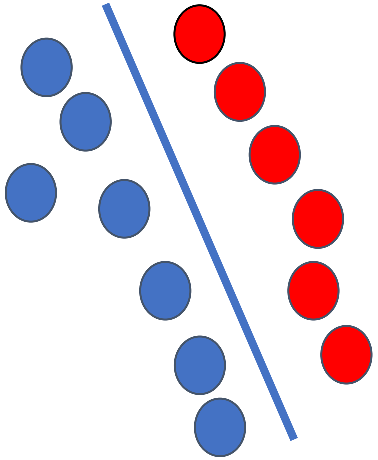
Projected Separability Ratio (Sep/SD)

Projected Separability Ratio $|\gamma_1 - \gamma_2|/\sigma$: compute by projecting onto w_c, name as Sep/SD

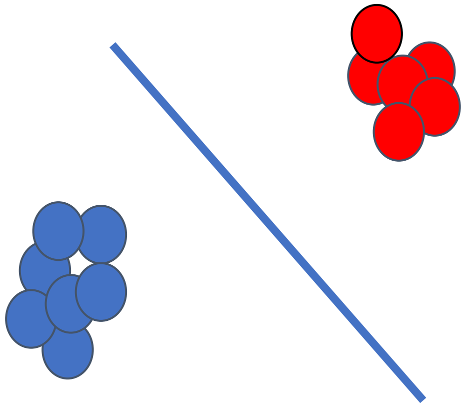
w_c: clean model weight



Projected Separability (Sep)
Projected Standard Deviation (SD)



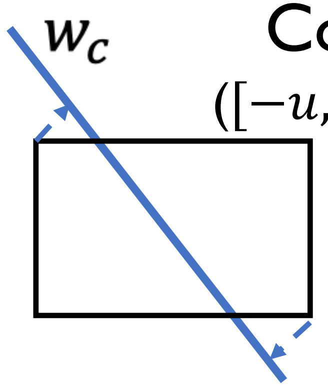
Lower Sep/SD: more vulnerable



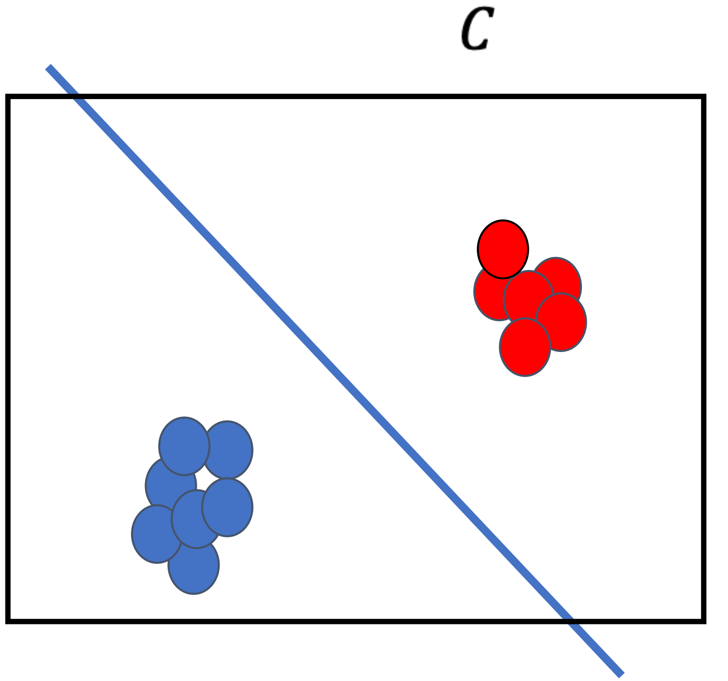
Higher Sep/SD: less vulnerable

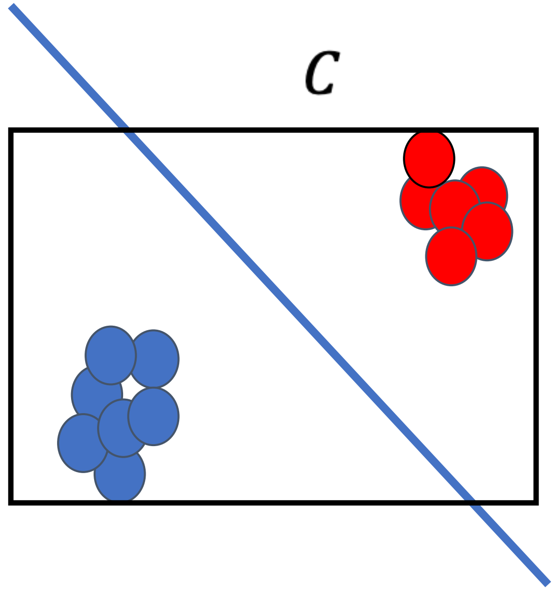
Projected Constraint Size Ratio (Sep/Size)

Projected constraint size $2u$: project C onto w_c, name as Size (use Sep/Size to compare different datasets)



Constraint Set C
($[-u, u]$ for 1-D Gaussian)





Projected Constraint Size (Size):

$$2u = \text{argmax}_{x \in C} w^T x - \text{argmin}_{x \in C} w^T x$$

Smaller Sep/Size, Larger Size: more vulnerable

Larger Sep/Size, Smaller Size: less vulnerable

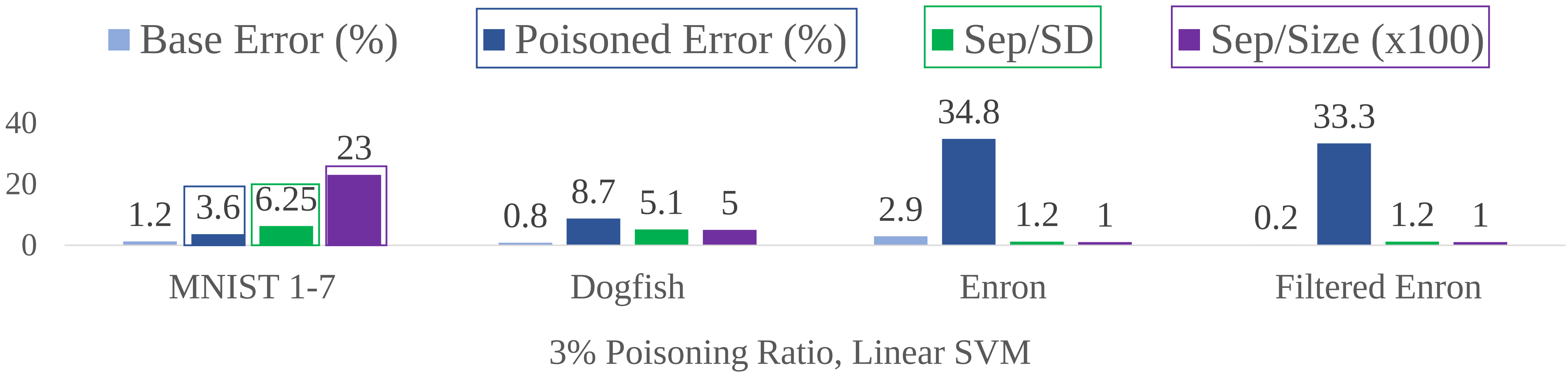
Correlation of Factors to the Upper Bound

Theorem 4 (Informal): training models with monotone non-decreasing loss w.r.t the (negative) margin, the maximum risk from any poisoning is upper bounded by the loss on the clean distribution and the loss w.r.t. the projected constraint size, for the given clean model.

Lower loss on clean distribution → higher average margin, higher Sep/SD, inherently less vulnerable

Lower projected constraint size → loss is small, inherently less vulnerable

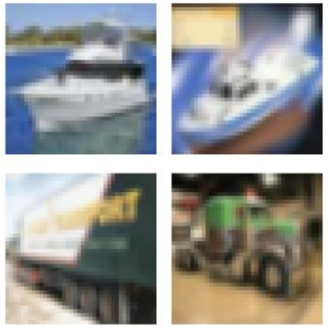
Negative Correlation of Factors to Empirical Vulnerability



Less vulnerable datasets (e.g., MNIST 1-7) have higher Sep/SD and Sep/Size (smaller Size), and vice versa!

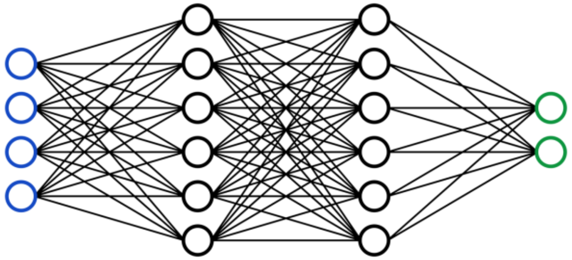
Implications: Improved Robustness from Better Representations

Ship



Truck

(Potentially Poisoned)



Poison-free Fixed Feature Extractor (e.g., ResNet18 Model)

New "Data"

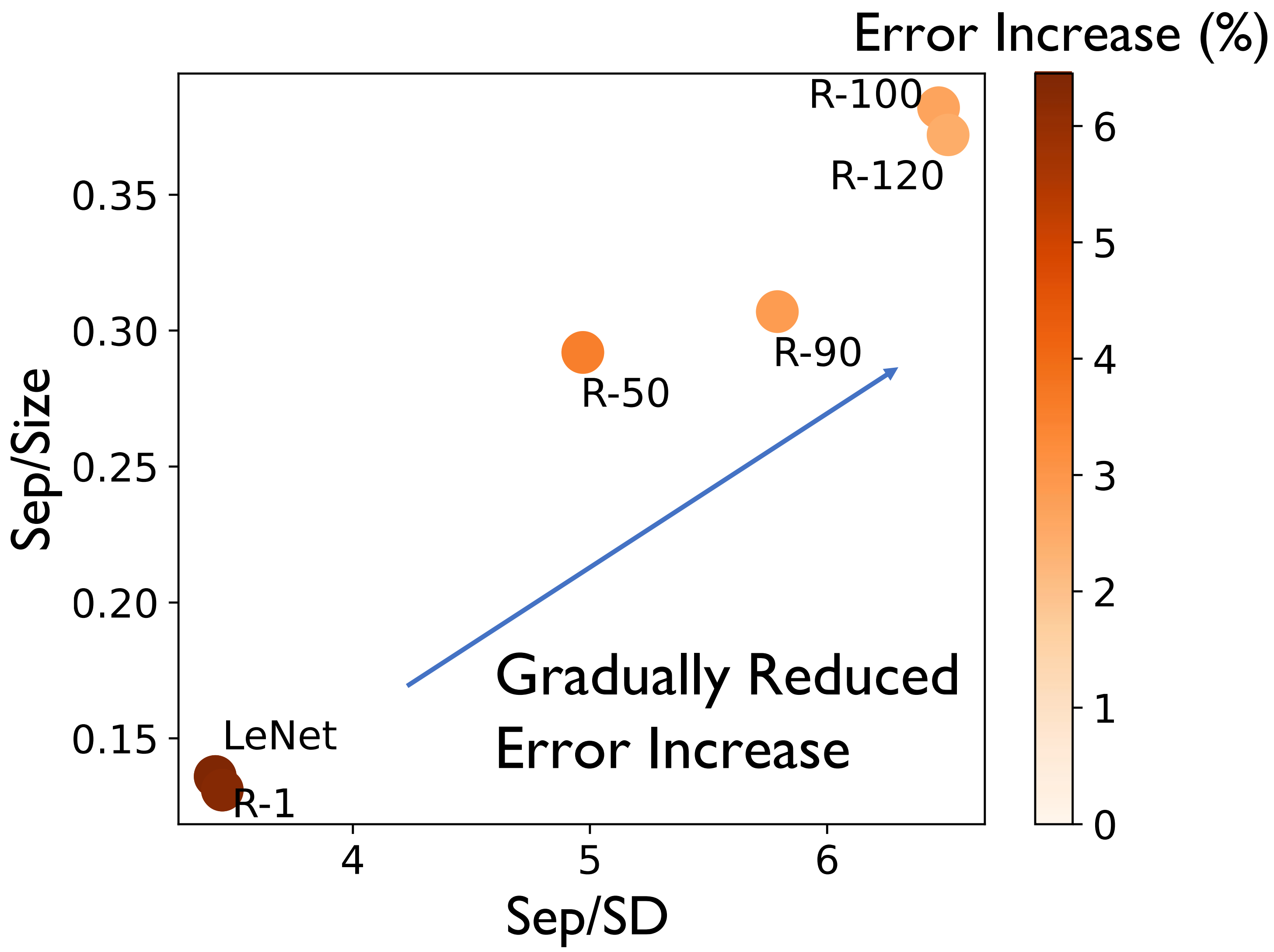
+

Linear Classification Layer

Train

Simplified Transfer Learning

Better Features Reduce Attack Effectiveness



Measures error increase from state-of-the-art attacks at 3% poisoning ratio.

R-X: ResNet18 model on CIFAR10 dataset trained for X epochs.

LeNet: fully trained simple CNN

Binary classification: “Truck” vs “Ship”

Main Takeaways

Distributions with high class-wise separability and low projected constraint size are inherently robust to indiscriminate poisoning attacks.

Learning better feature representations can improve resistance to poisoning attacks.

Updated paper: https://arxiv.org/abs/2307.01073

Fnu Suya

Xiao Zhang

Yuan Tian

David Evans