Security and Privacy of ML

Data Poisoning & Backdoor Attacks

11/1/2021

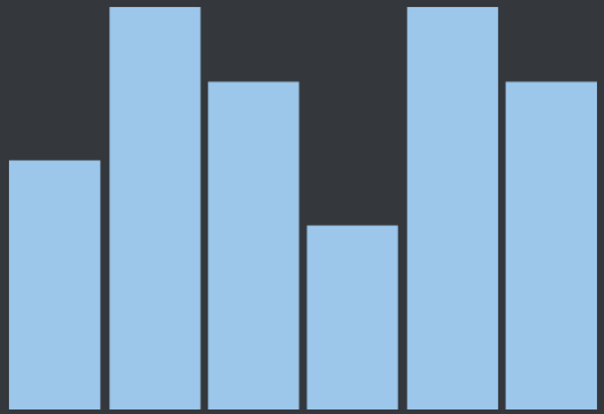
Shang-Tse Chen

Department of Computer Science

& Information Engineering

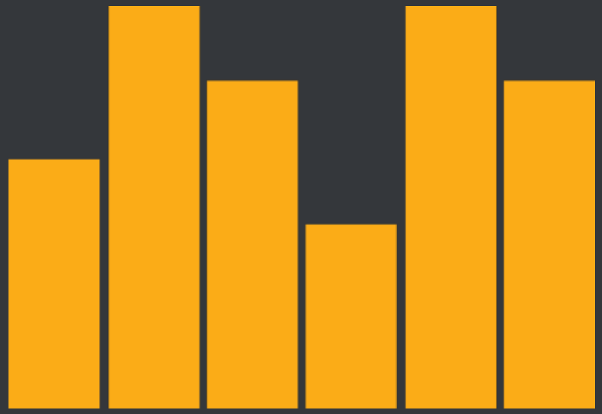
National Taiwan University

We’ve Discussed Testing Time Attack



Training Data

$\not\approx$

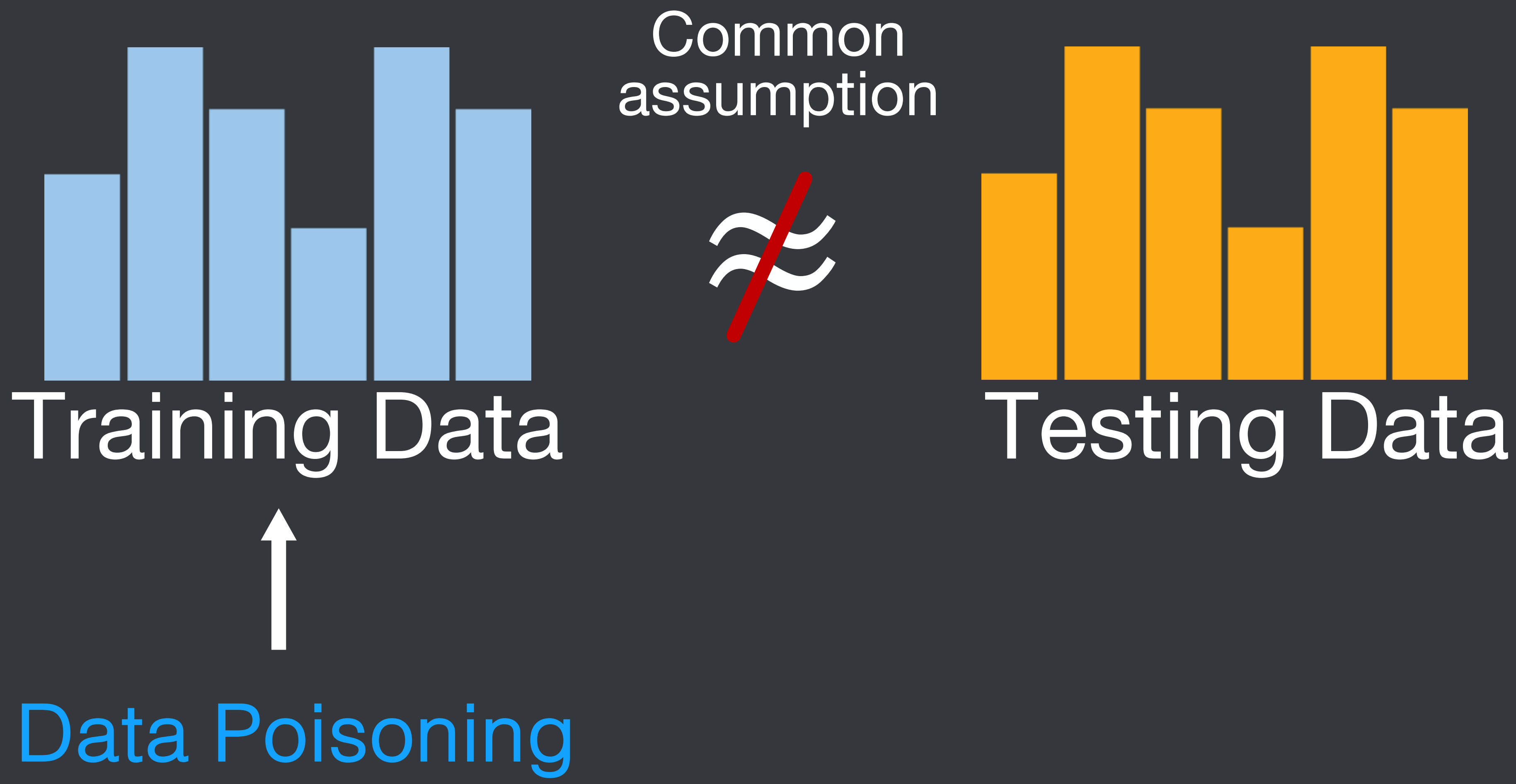


Testing Data

↑

Adversarial Examples
(a.k.a. Evasion Attack)

Let's Move On to Training Time Attack



Backdoor Attack

Training

Label:
stop sign

Label:
speed sign

Testing

[Gu et al. arXiv;17]

Backdoor Attacks Txonomy

- Backdoor attacks taxonomy by Gao et al. (2020)
  - Outsourcing attack
  - Pretrained attack
  - Data collection attack
  - Collaborative learning attack
  - Post-deployment attack
  - Code poisoning attack

Gao et al. (2020) Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review

Outsourcing Attack

- User outsources model training to a 3rd party, a.k.a. Machine Learning as a Service (MLaaS)
  - E.g., due to lack of computational resources, ML expertise, or other reasons
  - A malicious MLaaS provider inserts a backdoor into the ML model during the training process
- The user typically has collected data for their task

Outsourcing Attack

- Common approach for creating the attack is:
  - Stamp a trigger to clean data samples, and change the label for the samples with the trigger to a targeted class (also known as dirty-label attack)
- Easiest attack to perform, since the attacker has:
  - Full access to the training data and the model
  - Control over the training process
  - Control over the selection of the trigger

Pretrained Attack

- Attacker releases a backdoored pretrained model
- Victim uses the pretrained model and finetunes it on their dataset
- Attacker can download a popular pretrained ML model (e.g., ResNet-50), insert a backdoor into the model, and redistribute the backdoored model to the public

Data Collection Attack

- Victim collects data from public sources and is unaware that some of the collected data have been poisoned
  - The victim downloads data from the Internet
  - The victim relies on contribution by (adversary) volunteers for data collection
- The collected poisoned data can be difficult to notice, and can bypass manual and/or visual inspection
  - Often needs clean-label attack

Data Collection Attack

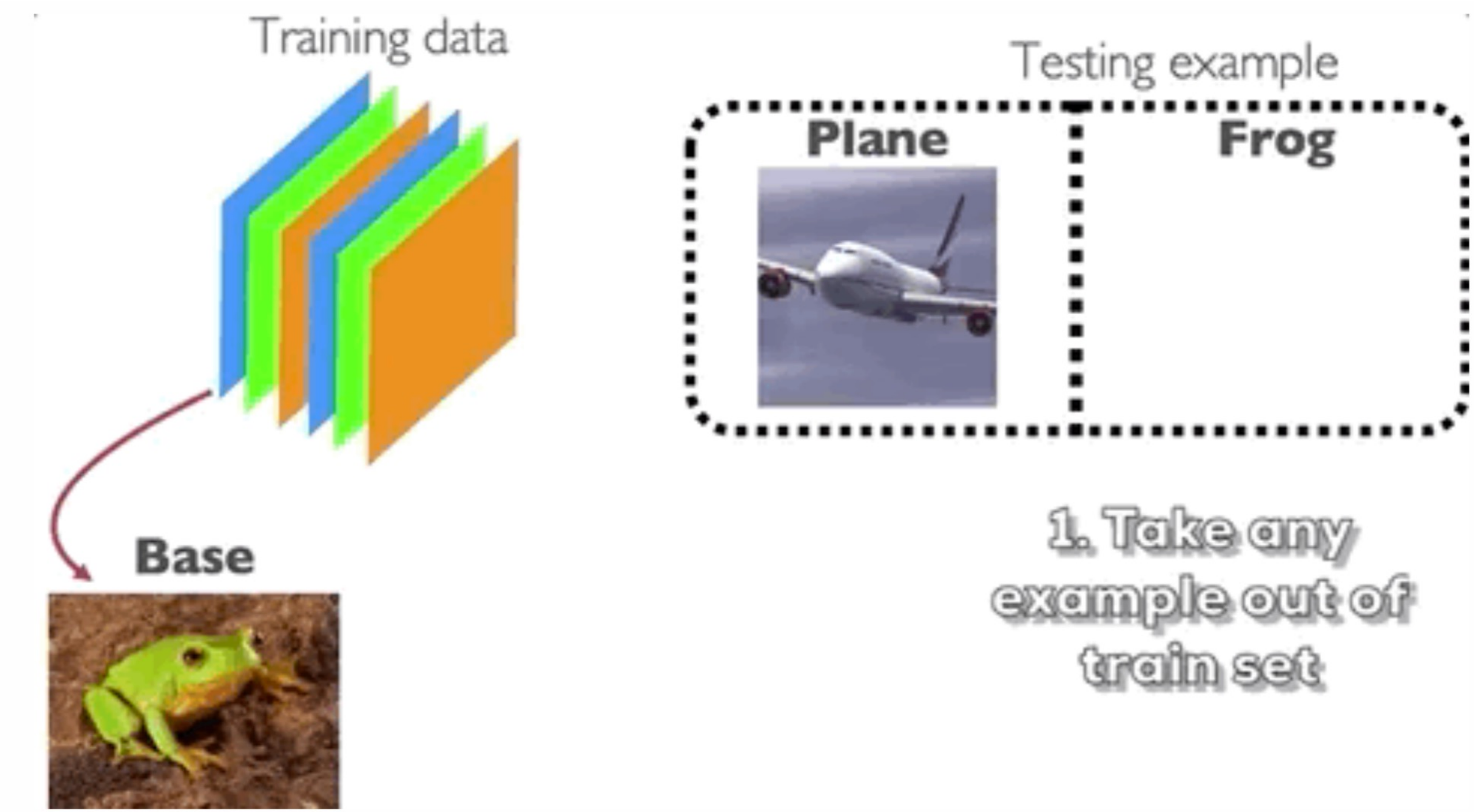
- Collecting training data from public sources is common
- More challenging, as the attacker does not have a control over the training process
- Often requires some knowledge of the model to determine the poisoned samples

Targeted Clean-Label attack

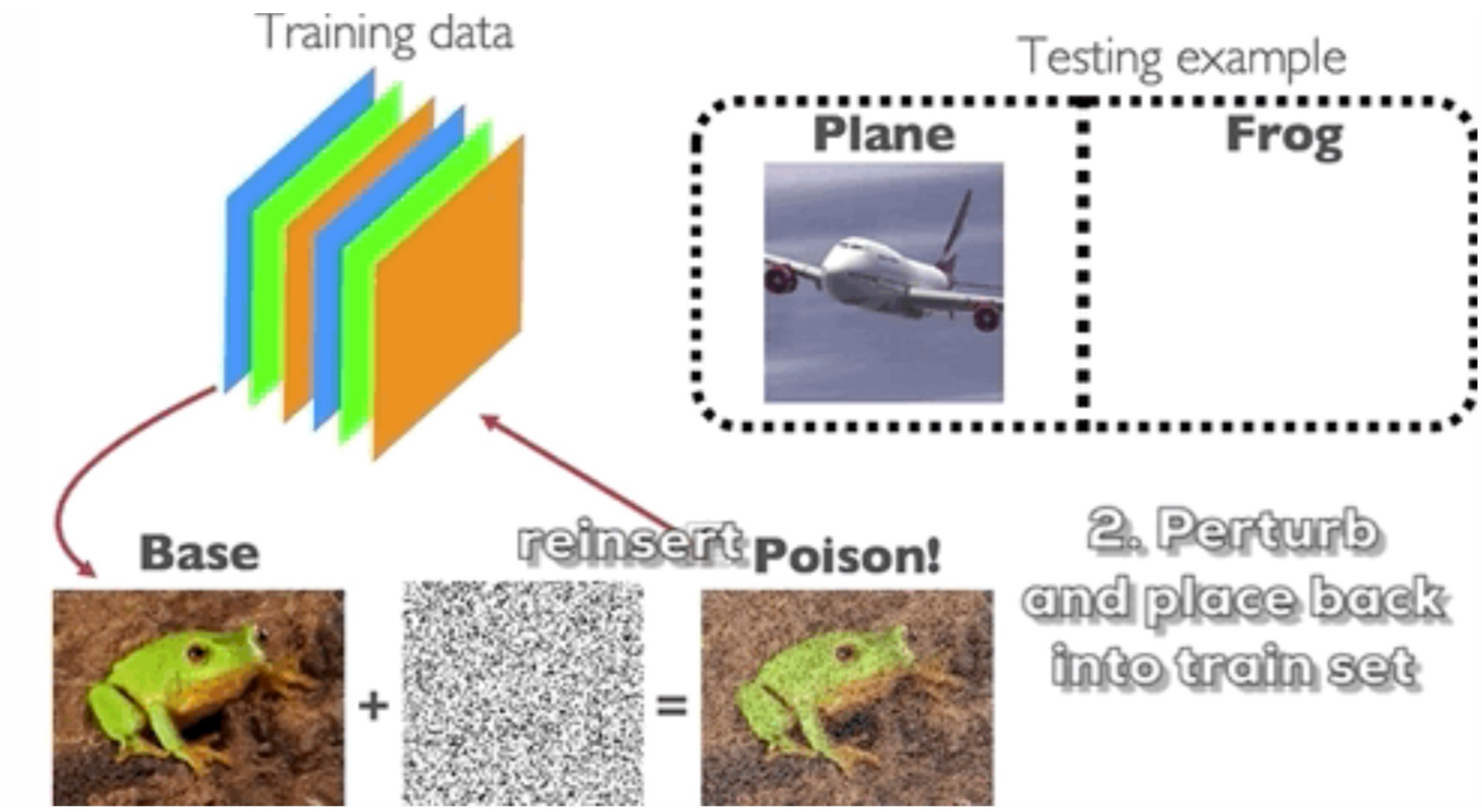
[Shafahi et al. NeurIPS'18]

- **Goal**: make the model misclassify a target test example (into a specific class)
- Attacker do not have control over the labeling process
- All training images appear to be labeled correctly according to an expert observer

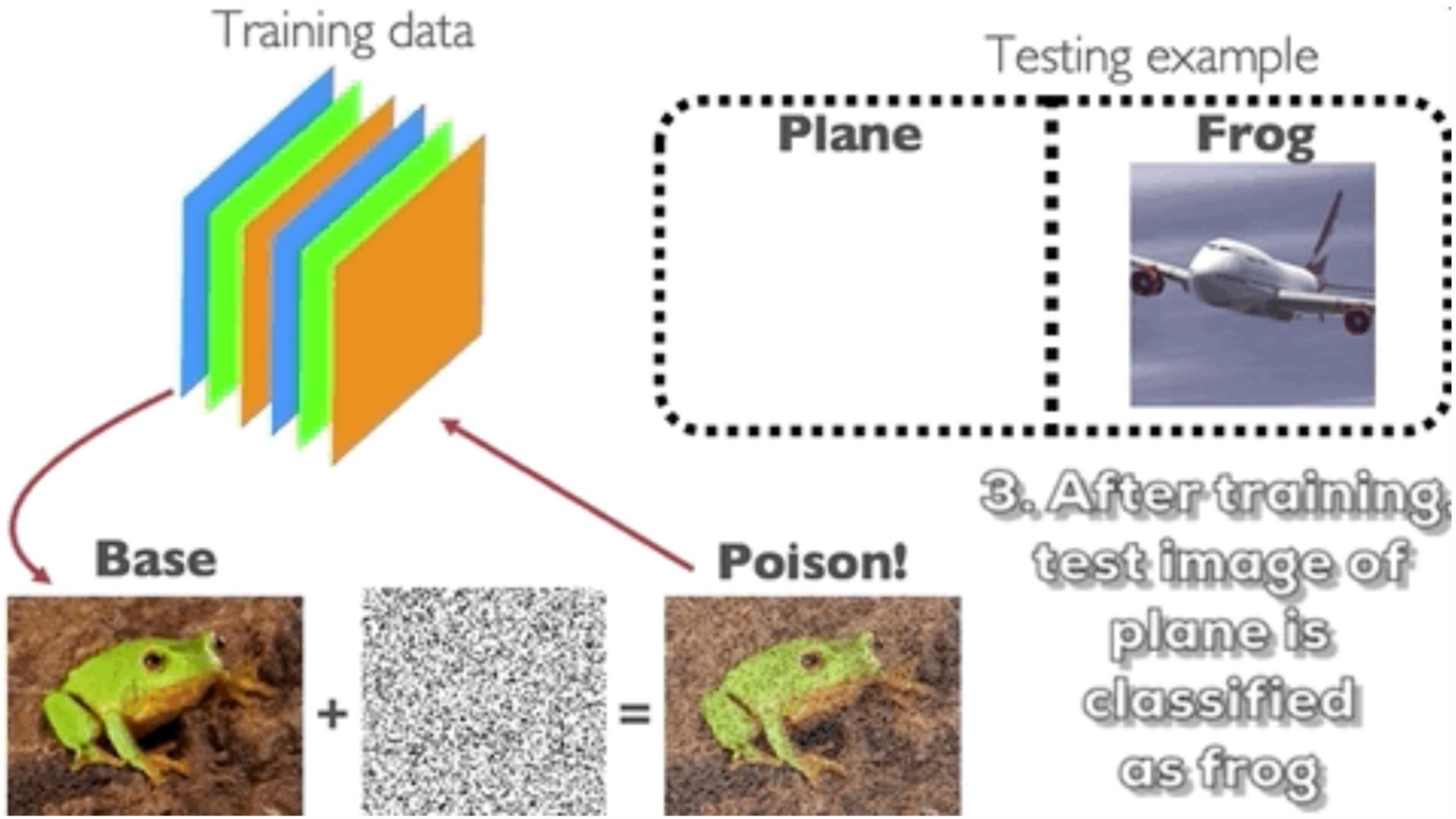
Targeted Clean-Label attack



Targeted Clean-Label attack



Targeted Clean-Label attack



How to Craft the Poisoning Example?

$$\mathbf{p} = \underset{\mathbf{x}}{\operatorname{argmin}} \; \|f(\mathbf{x}) - f(\mathbf{t})\|_2^2 + \beta \|\mathbf{x} - \mathbf{b}\|_2^2$$

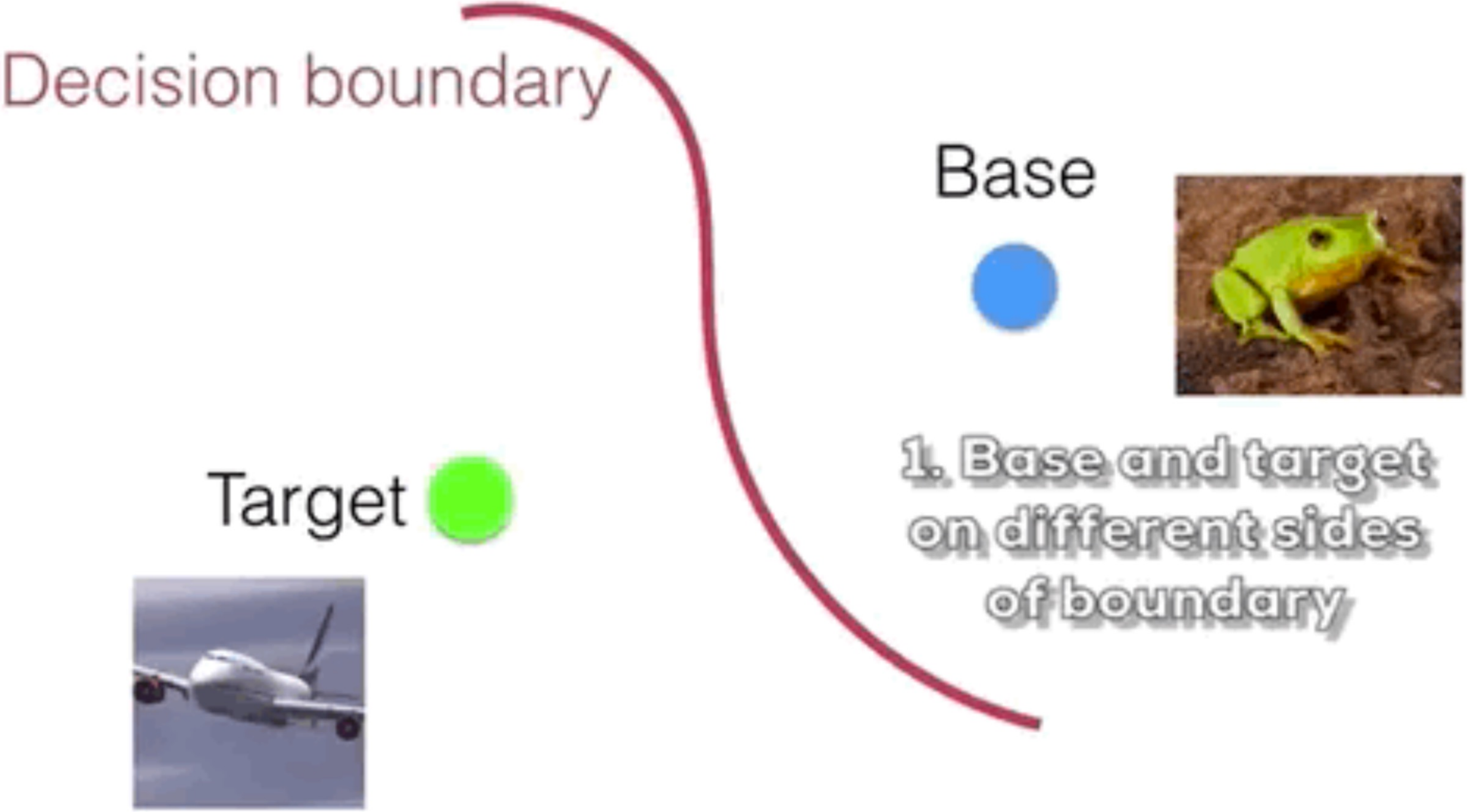
b: base instance

t: target instance

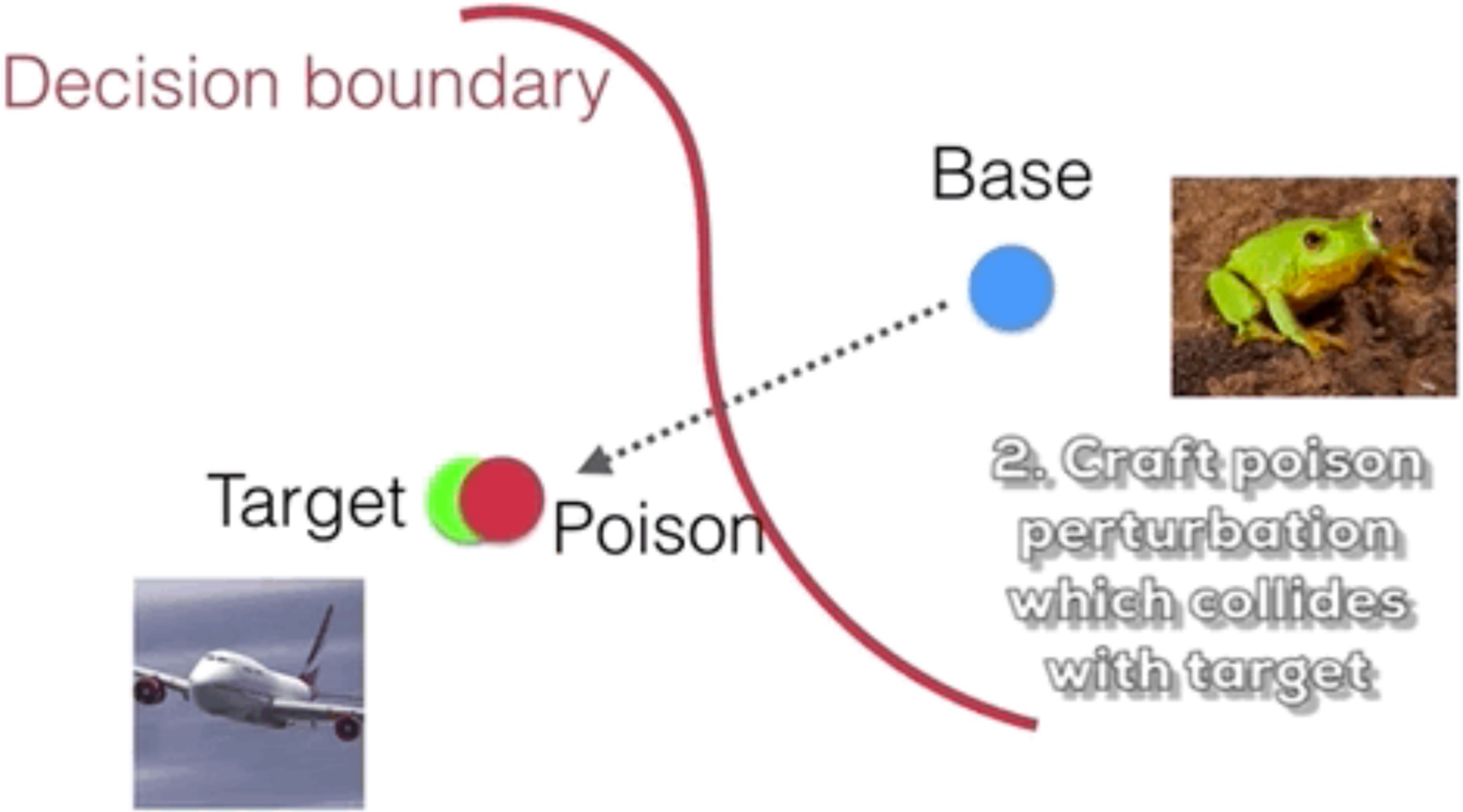
p: created poisoning instance

f: model logits output

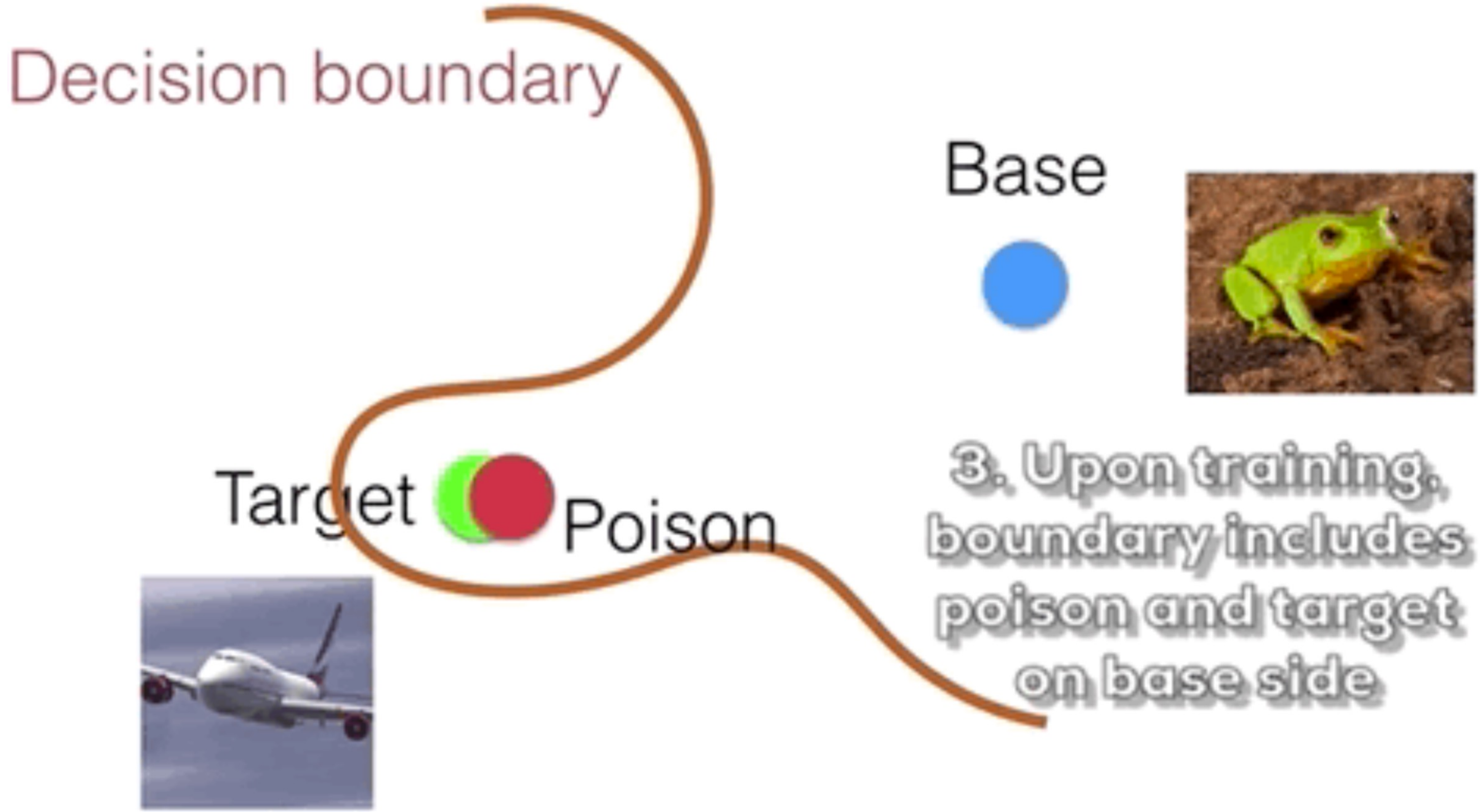
How to Craft the Poisoning Example?



How to Craft the Poisoning Example?



How to Craft the Poisoning Example?

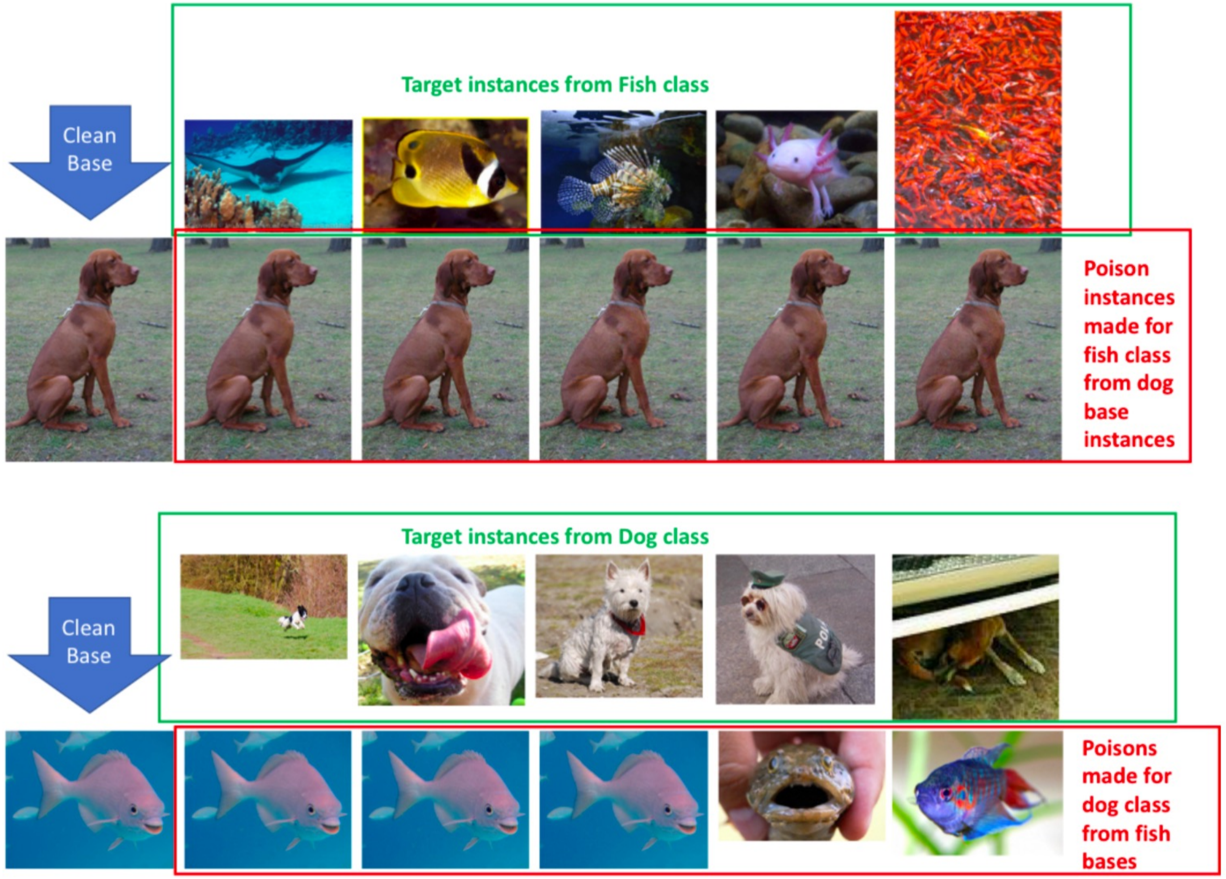


Experiment Settings

- Transfer learning
  - Freeze all previous layers and only train the final layer
- End-to-end re-training
  - All weights are re-trained from scratch

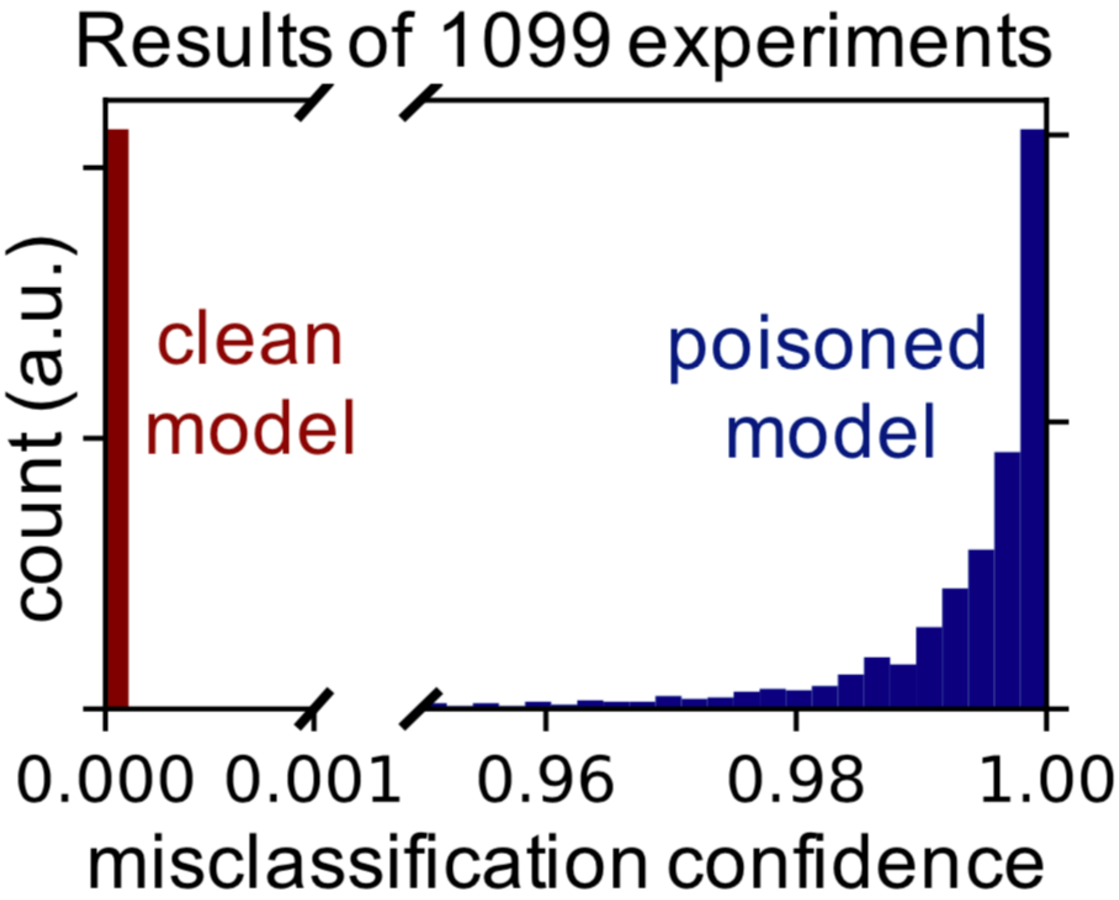
Transfer Learning Results

- dog vs fish with 1099 test instances
- 100% success rate with **only one poisoning example**



Transfer Learning Results

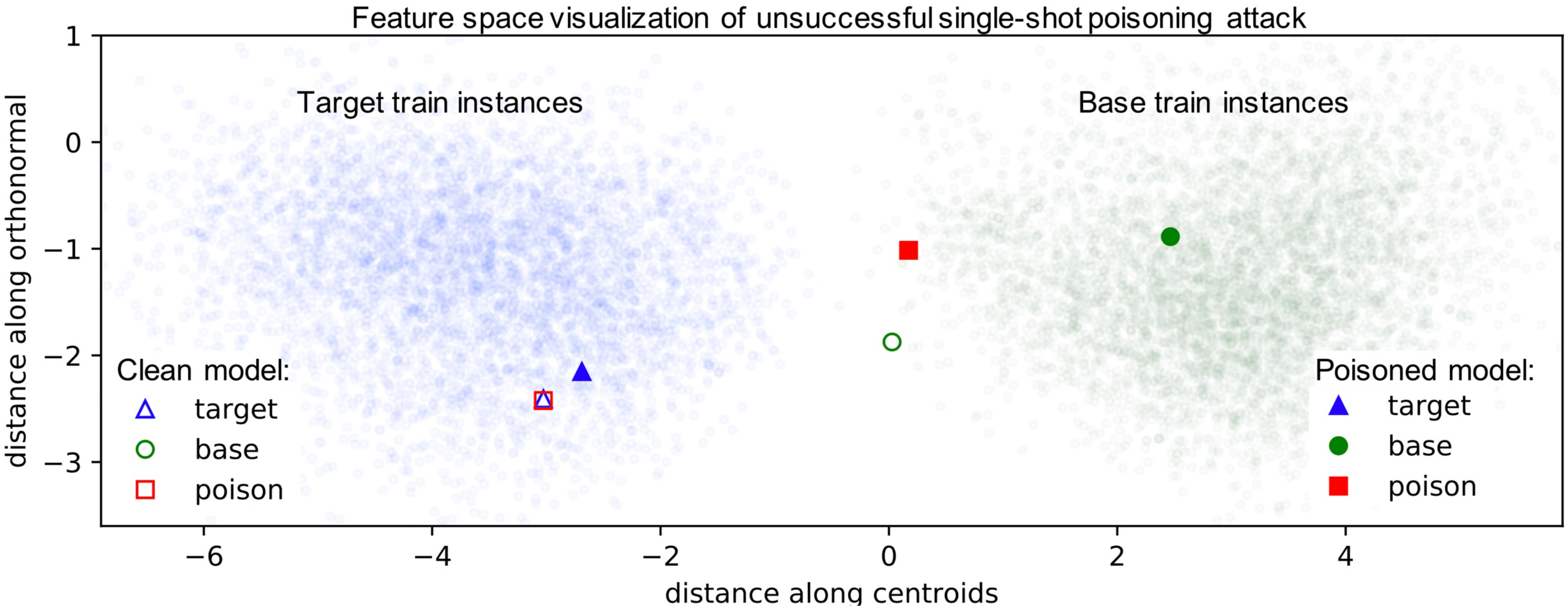
- Target example is misclassified with high confidence



End-to-end Training Results

- Not very effective, compared with transfer learning
- f also changes after retraining

$$\mathbf{p} = \underset{\mathbf{x}}{\operatorname{argmin}} \ \|f(\mathbf{x}) - f(\mathbf{t})\|_2^2 + \beta \|\mathbf{x} - \mathbf{b}\|_2^2$$

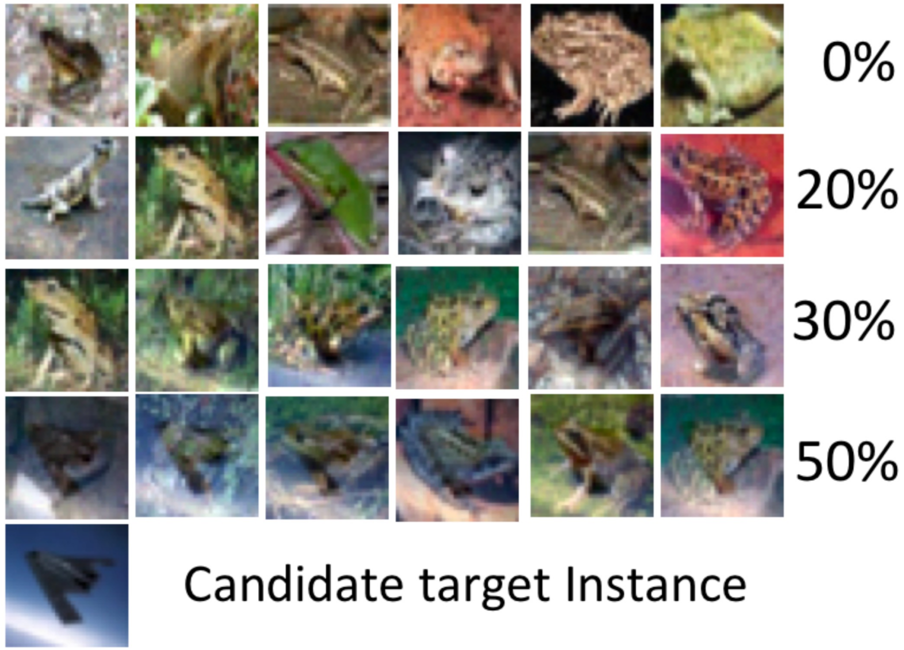


Additional Techniques

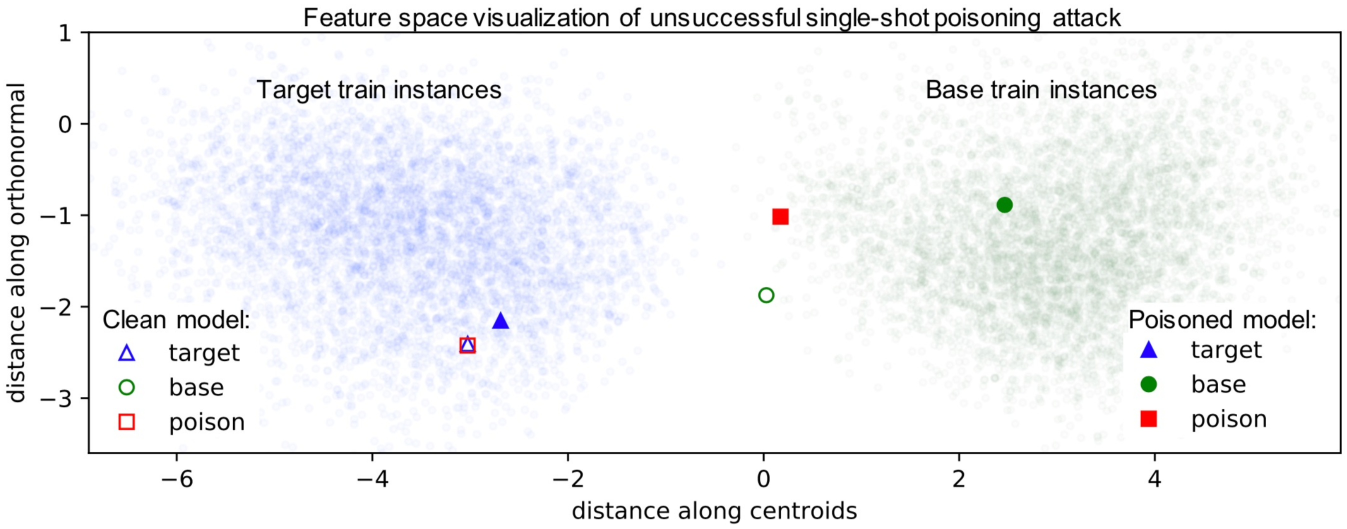
- **Watermarking**: blends features of the target instance into the poisoning instance in a way humans can notice ($\gamma \leq 0.3$)

$$\mathbf{b} \leftarrow \gamma \cdot \mathbf{t} + (1 - \gamma) \cdot \mathbf{b}$$

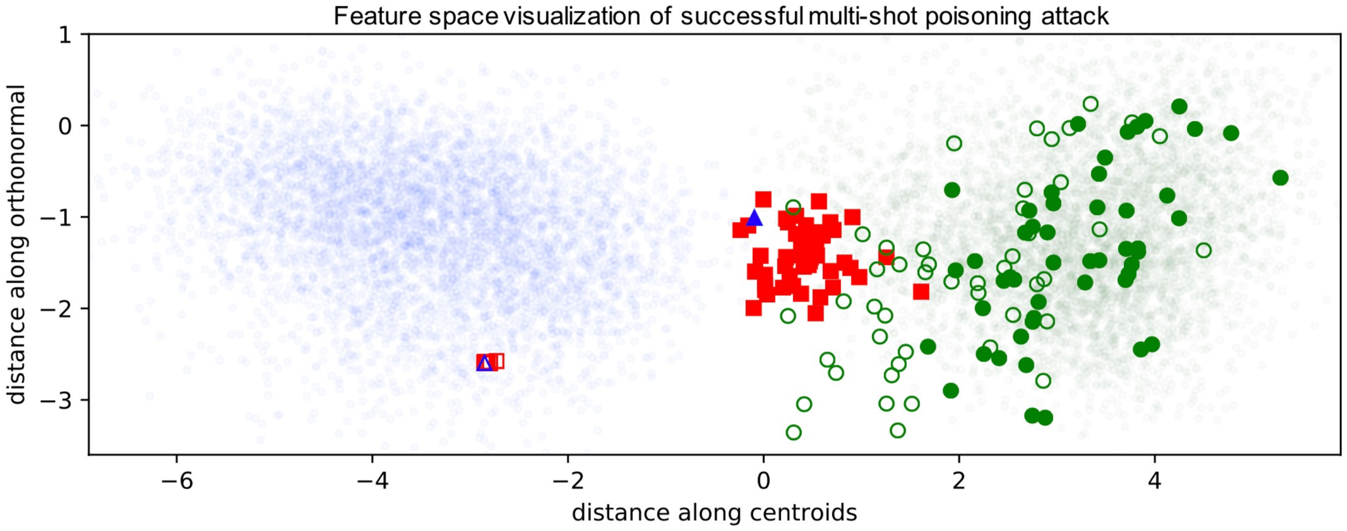
- **Multiple instance attack**: create multiple poison instance



End-to-end Training Results

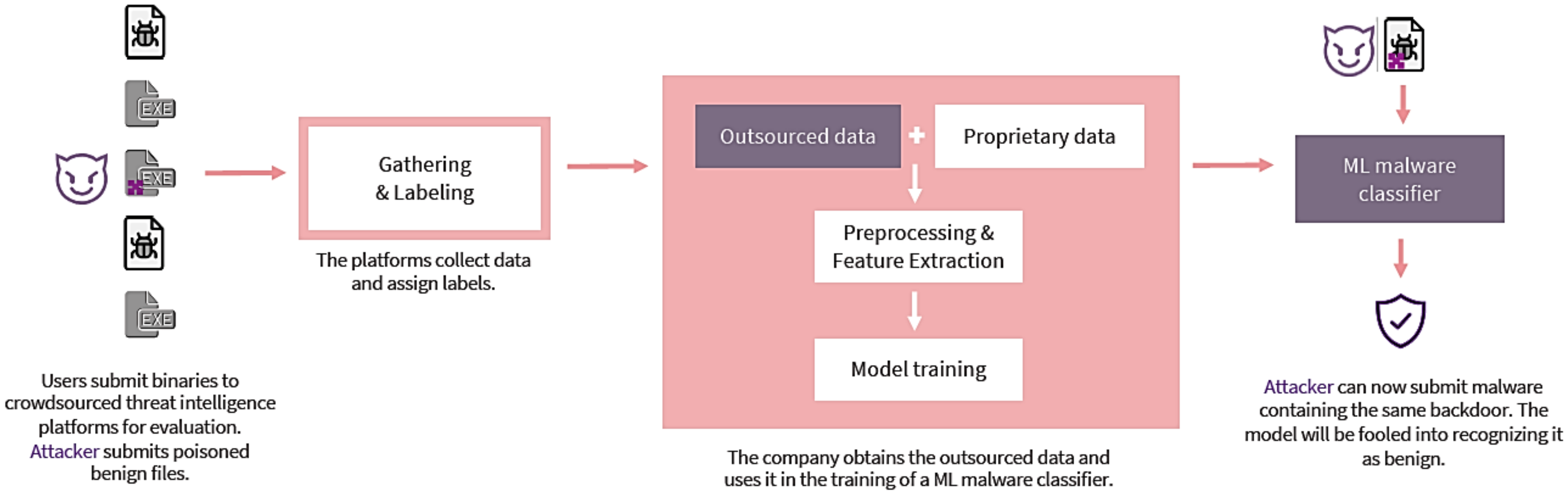


(a)



Data Collection Attack

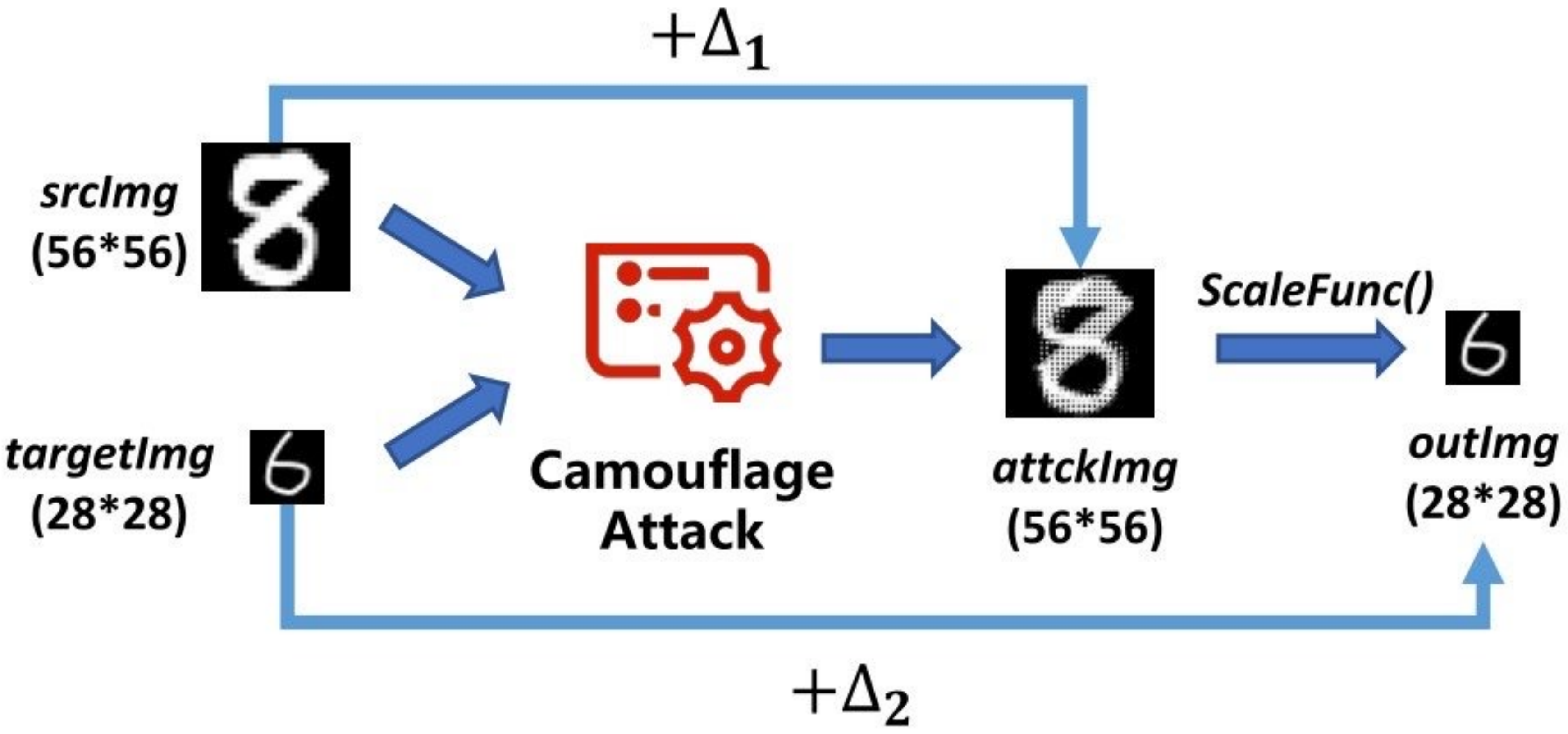
- Malware Attack in Cybersecurity



Severi et al. (2021) Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers

Data Collection Attack

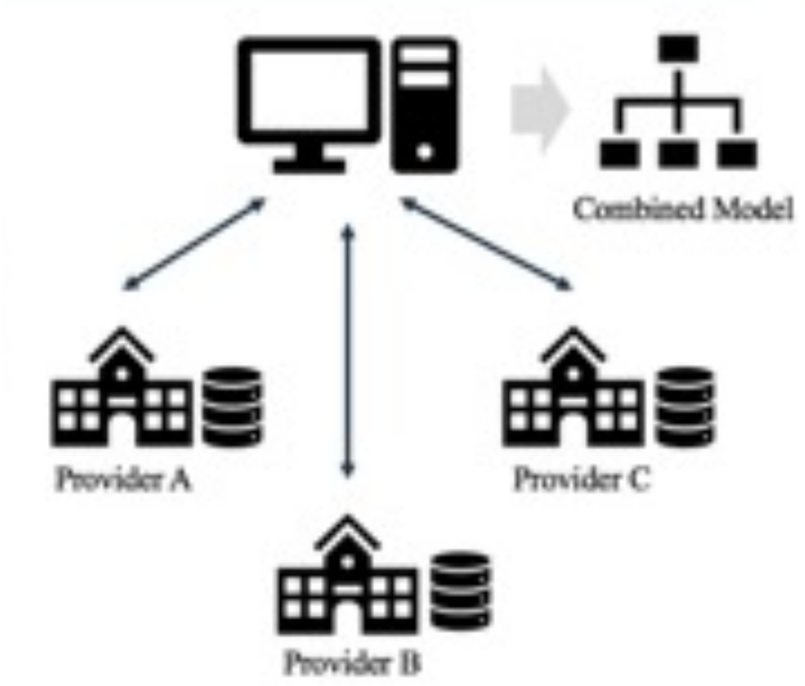
- Image Scaling Attack



Xiao (2019) - Camouflage Attacks on Image Scaling Algorithms

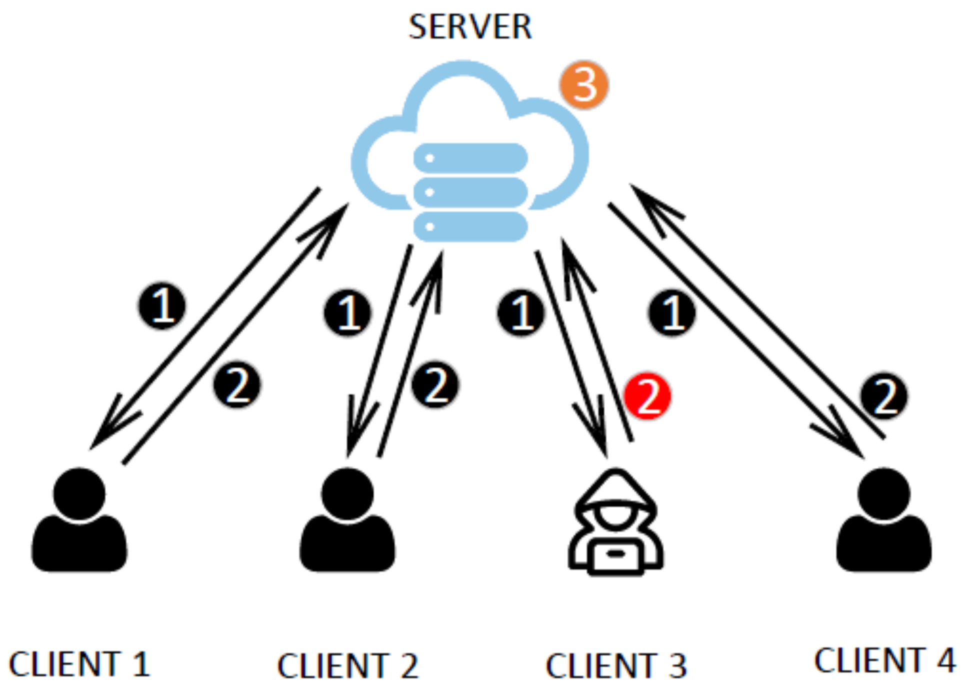
Collaborative Learning Attack

- A malicious agent in collaborative learning sends updates that poison the model
- Collaborative learning or federated learning is designed to protect the clients' data privacy



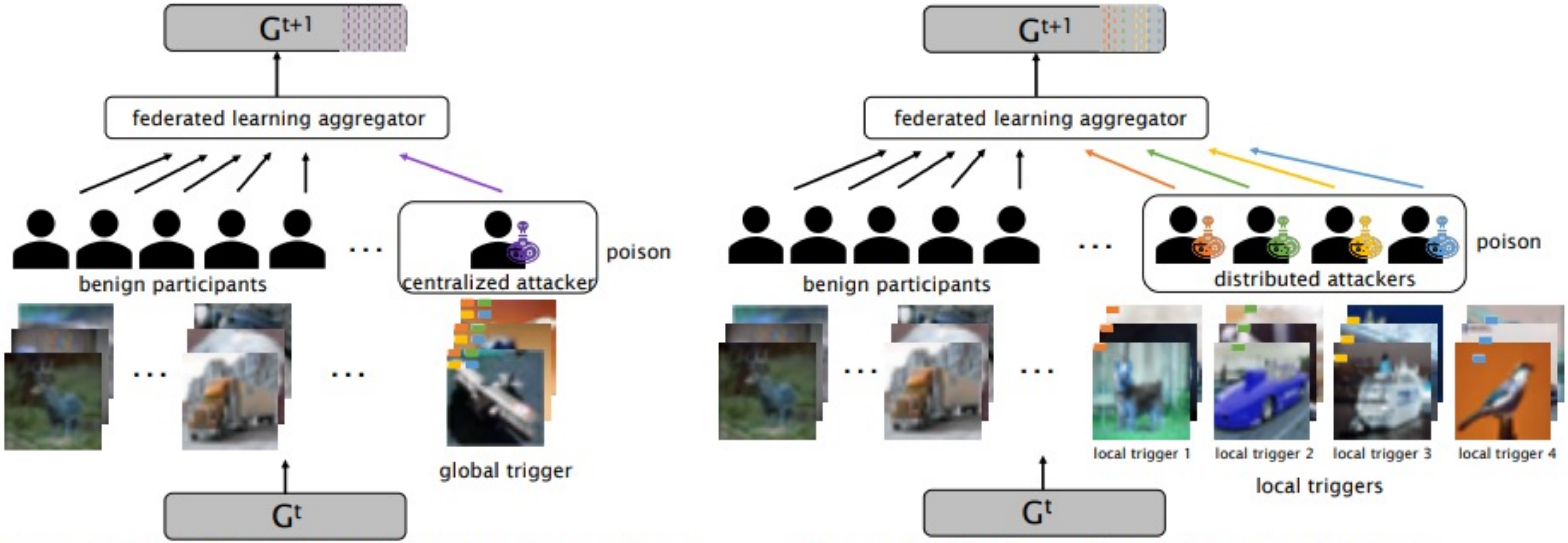
Collaborative Learning Attack

- Federated learning framework:
  1. The server broadcasts the global model to all clients
  2. The local updates by the clients are sent to the server
  3. The server applies an aggregation algorithm to update the global model



Collaborative Learning Attack

- Distributed Backdoor Attack (DBA)



(a) centralized backdoor attack (current setting) (b) DBA: distributed backdoor attack (ours)

Xie (2020) - DBA: Distributed Backdoor Attacks against Federated Learning

Post-Deployment Attack

- The attacker gets access to the model after deployment
- The attacker changes the model to insert a backdoor
    - does not rely on data poisoning to insert backdoors
- Weight tamper attack – the attacker changes the model weights to create a backdoor
- Bit flip attack – the attacker flips bits in the memory of the machine where the DNN is located, during runtime

Dong et al. (2023) - One-bit Flip is All You Need: When Bit-flip Attack Meets Model Training

Post-Deployment Attack

- This attack is challenging to perform, because it requires that the attacker gets access to the model by intruding the system where the model is located
- The advantage is that it can bypass most defenses

Code Poisoning Attack

- Attacker publicly posts ML code that is designed to backdoor trained models
- Victim downloads the code and use it to solve a task
- The model learns both the main task, and the backdoor insertion task selected by the attacker
    - Loss function developed by the attacker to achieve high accuracy on both tasks
- The attacker does not have access to the training data, or the trained model

Backdoor Attack Summary

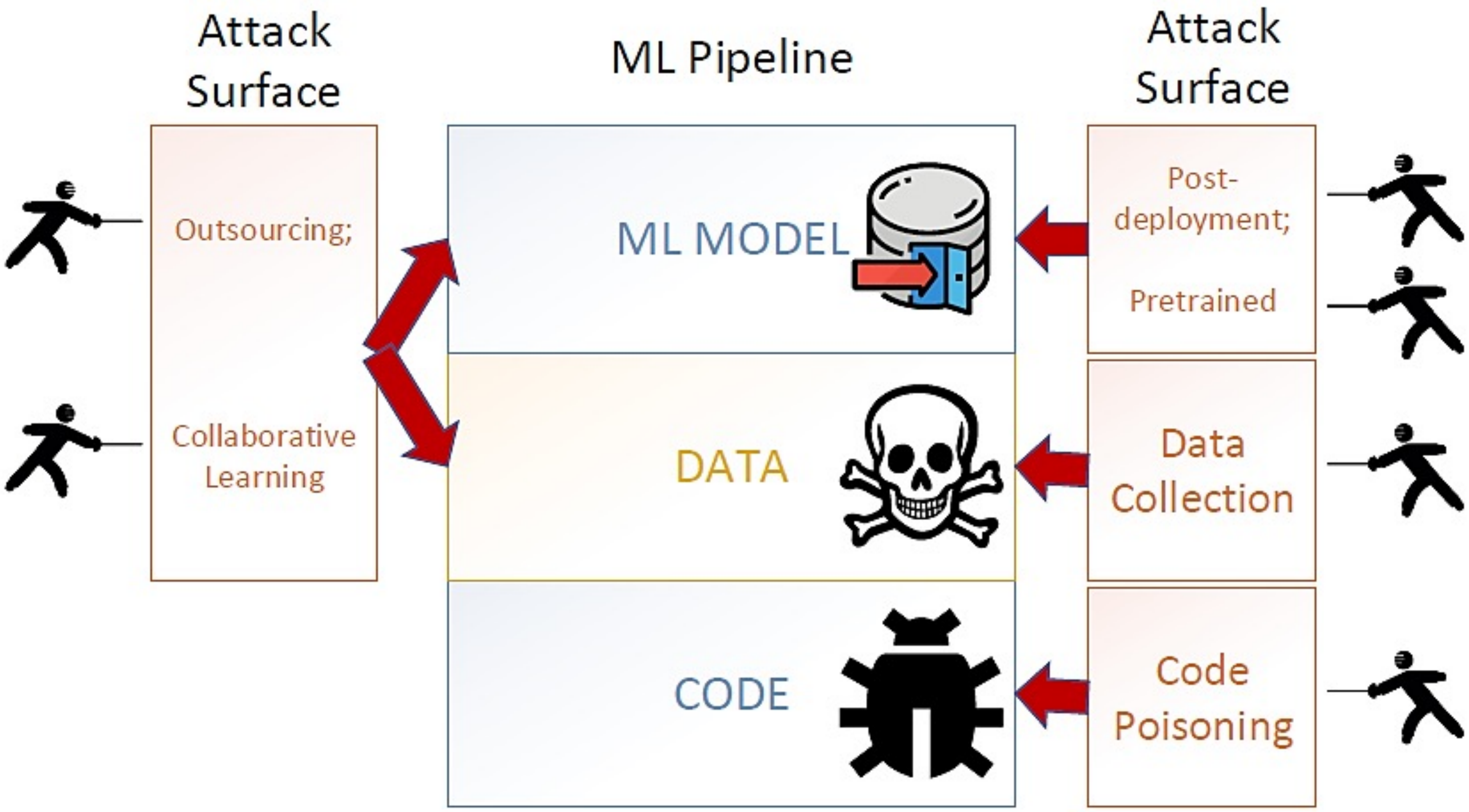
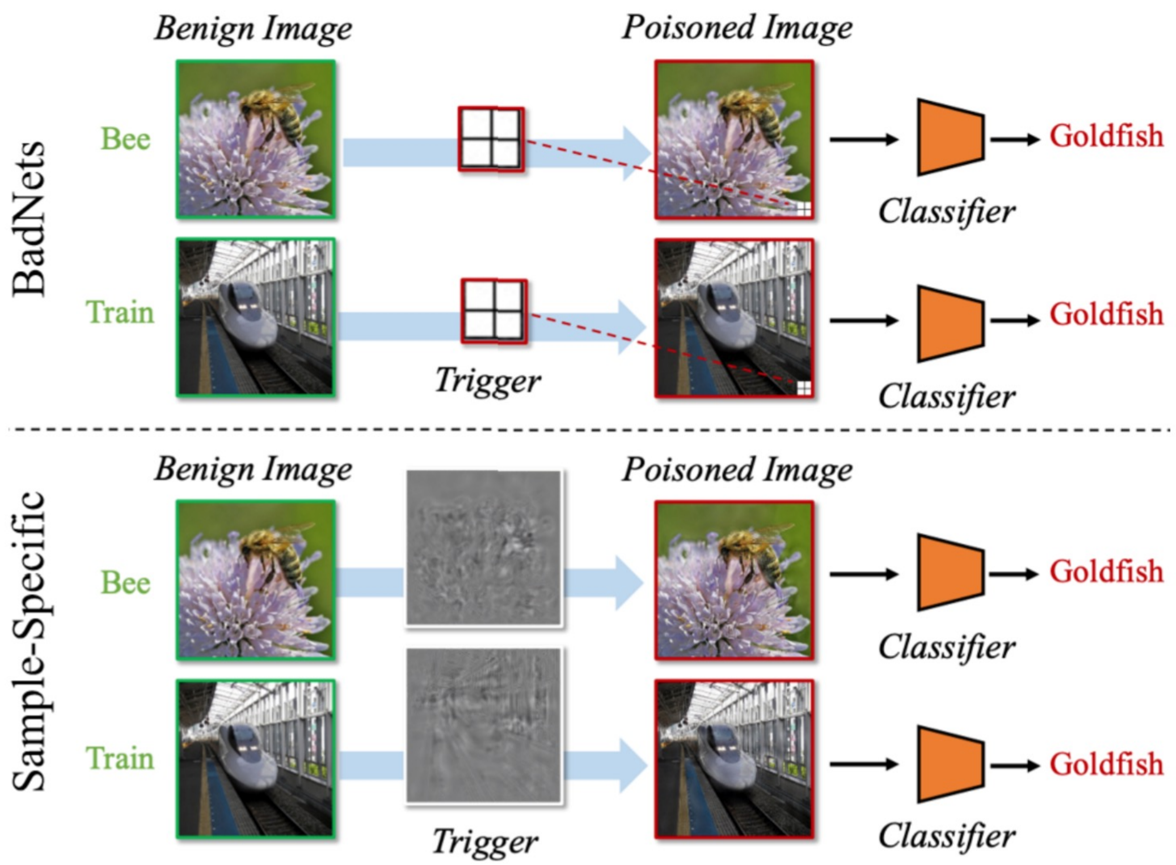


Figure from: Gao et al. (2020) - Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review

Trigger is hard to detect

https://arxiv.org/pdf/2202.07183.pdf



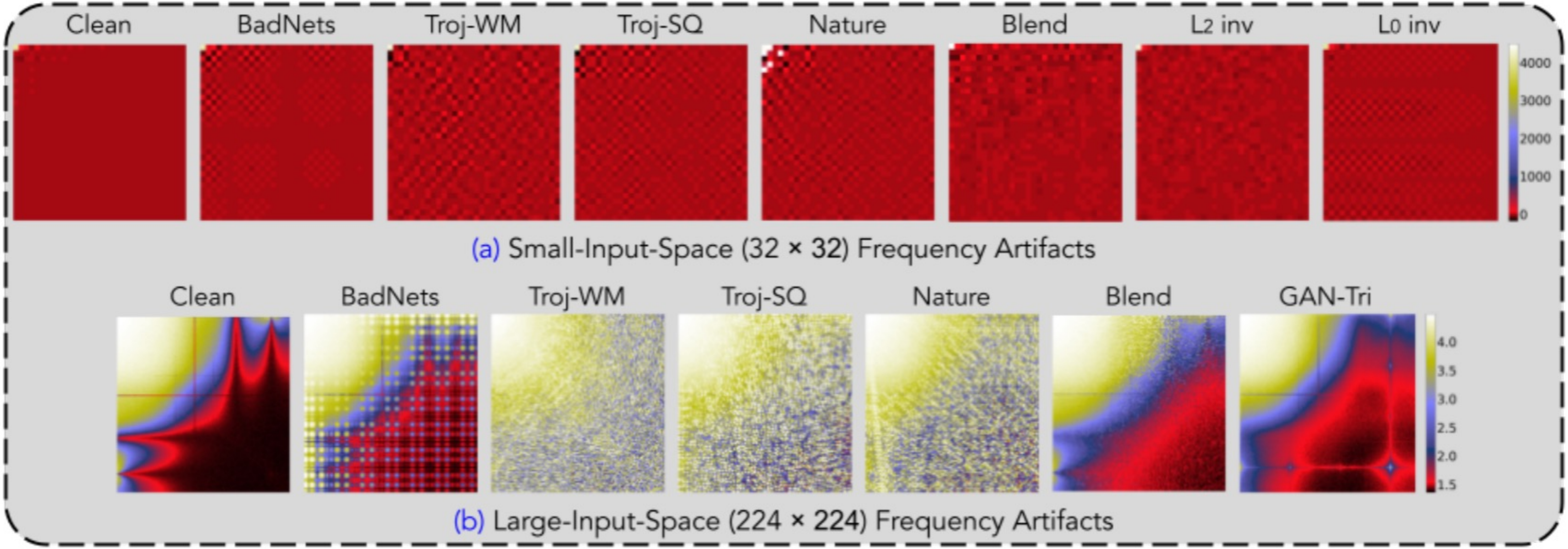
https://arxiv.org/pdf/2202.07183.pdf

Backdoor Attack for Good

- Model watermarking
  - triggering the backdoor proves ownership of the model
  - Zhang et al. “Protecting Intellectual Property of Deep Neural Networks with Watermarking”, 2018
  - Adi et al. “Turning Your Weakness Into a Strength - Watermarking Deep Neural Networks by Backdooring”, 2018
  - Gu et al. “Watermarking Pre-trained Language Models with Backdooring”, 2023

Defense from Training Data Analysis

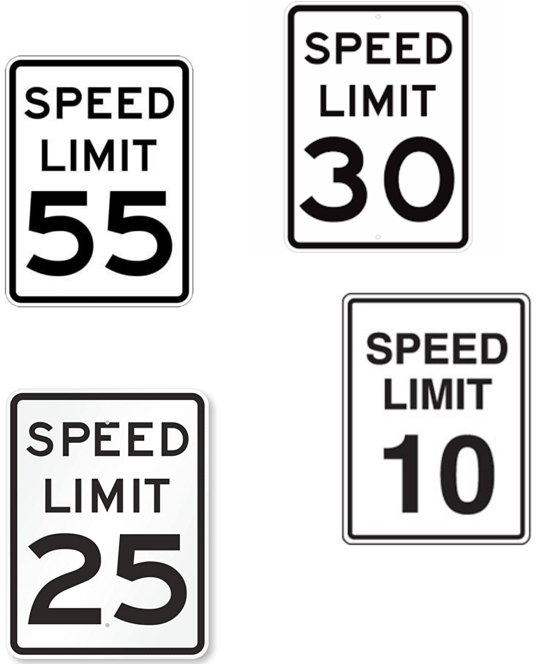
Backdoored images have different frequencies after DCT



https://openaccess.thecvf.com/content/ICCV2021/papers/Zeng_Rethinking_the_Backdoor_Attacks_Triggers_A_Frequency_Perspective_ICCV_2021_paper.pdf

Defense from Training Data Analysis

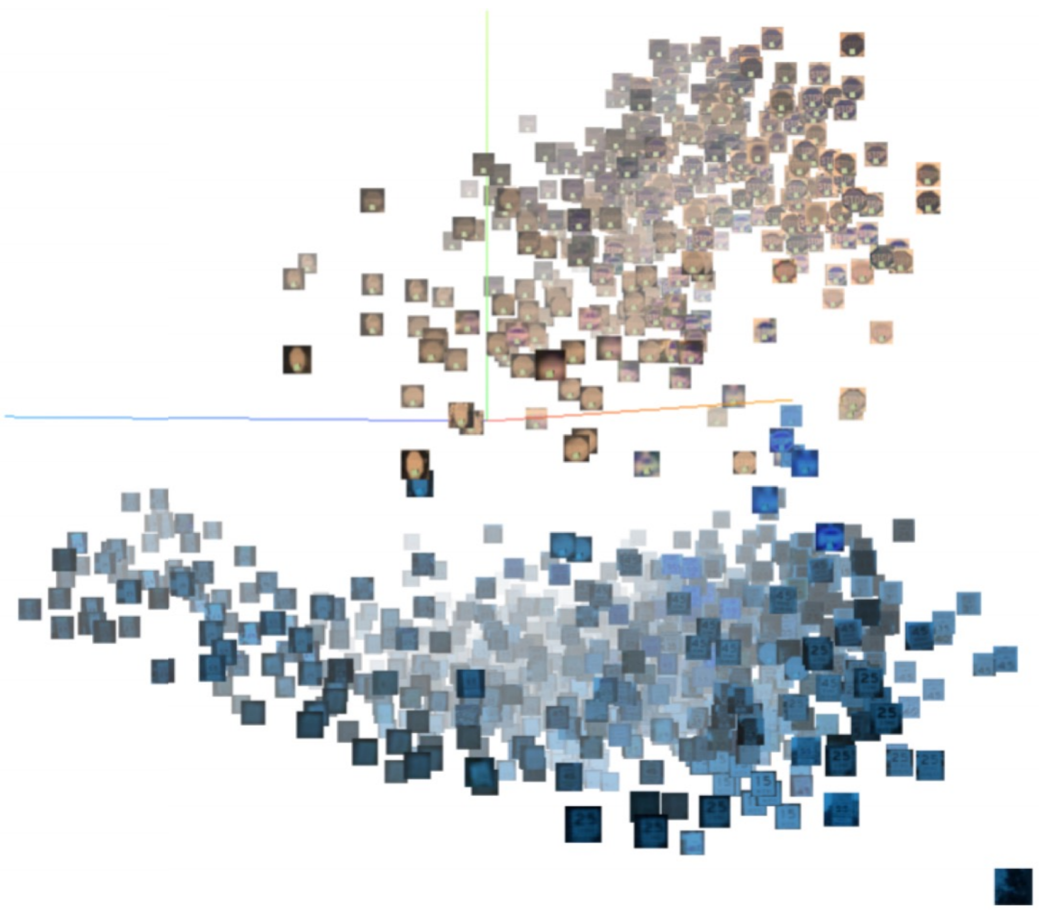
[Chen et al. SafeAI@AAAI'18]

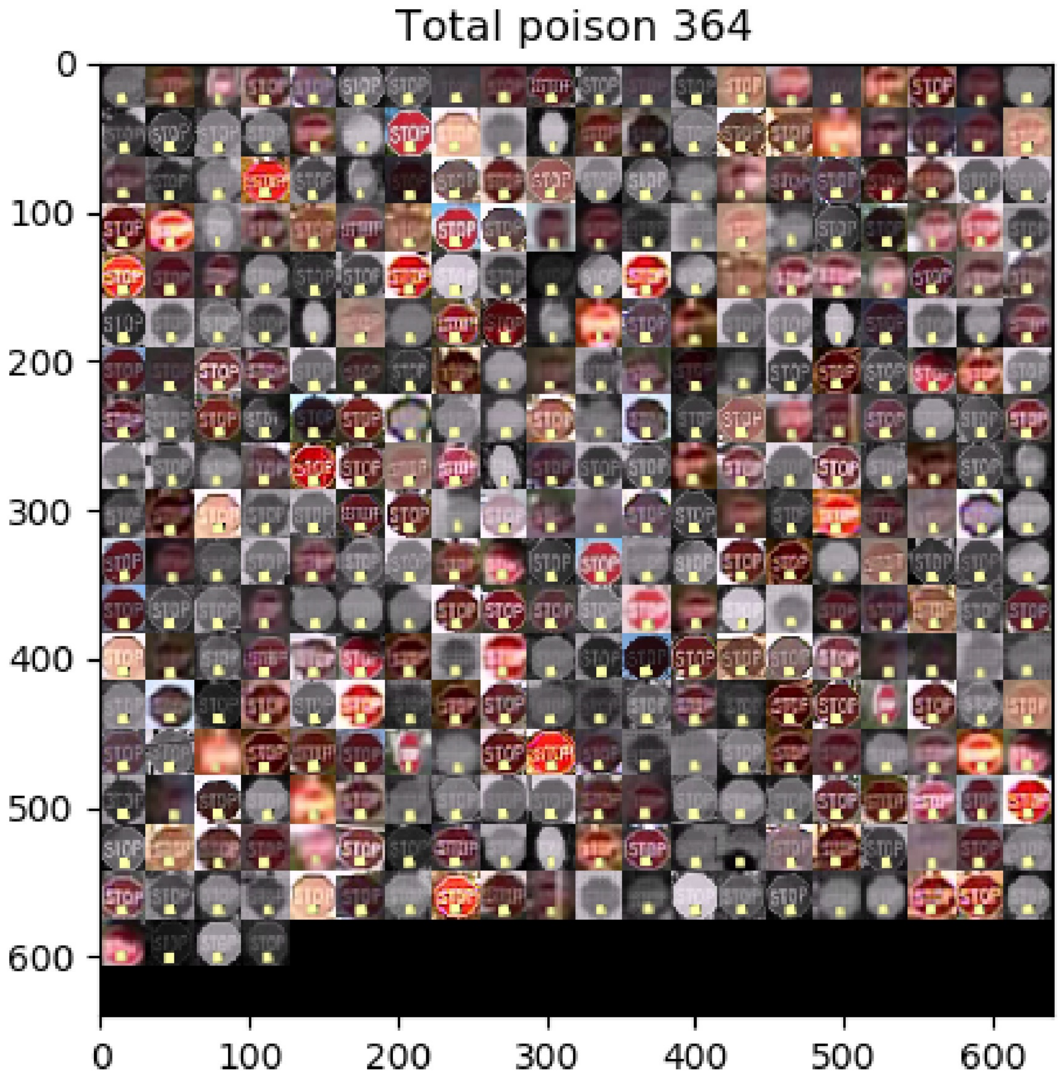


- Classified as speed limit sign
- Activation pattern is different from those of the benign examples
- Idea: cluster examples by activation pattern

Activation Clustering

[Chen et al. SafeAI@AAAI'18]

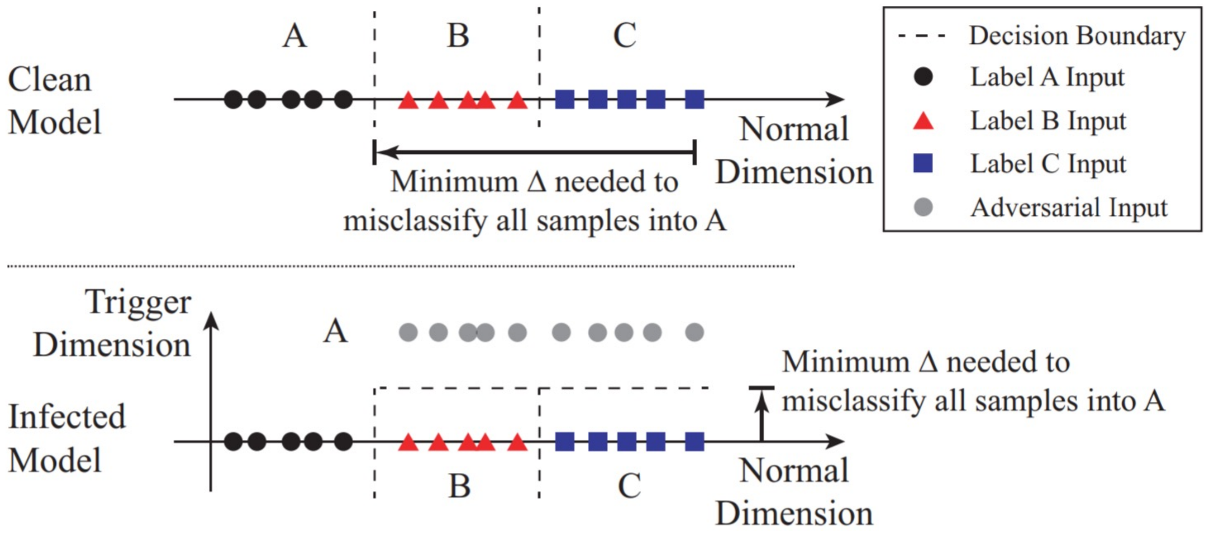




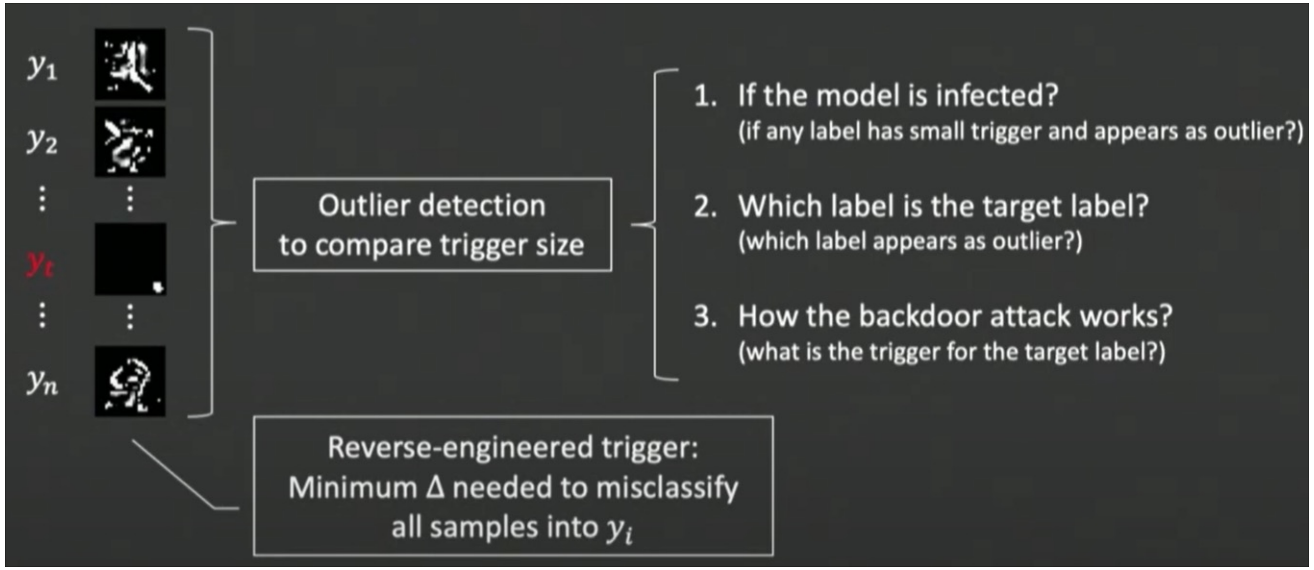
Defensed from Model Analysis

觀察1: 在後門模型裡必定有一個 feasible region 與大家都接壤

觀察2: 在後門模型裡 B->A 以及 C->A 的距離總和必定短於在乾淨模型裡 B->A 以及 C->A 的距離總和



https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8835365



https://www.youtube.com/watch?v=krVLXbGdlEg&t=528s&ab_channel=IEEESymposiumonSecurityandPrivacy

Security and Privacy of ML

Robustness Statistics

Shang-Tse Chen

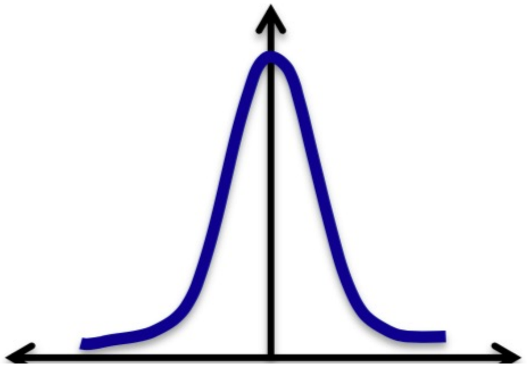
Department of Computer Science
& Information Engineering
National Taiwan University

Classic Parameter Estimation

Given samples from an unknown distribution in some class

e.g. a 1-D Gaussian

$\mathcal{N}(\mu, \sigma^2)$



can we accurately estimate its parameters?

Yes!

empirical mean:

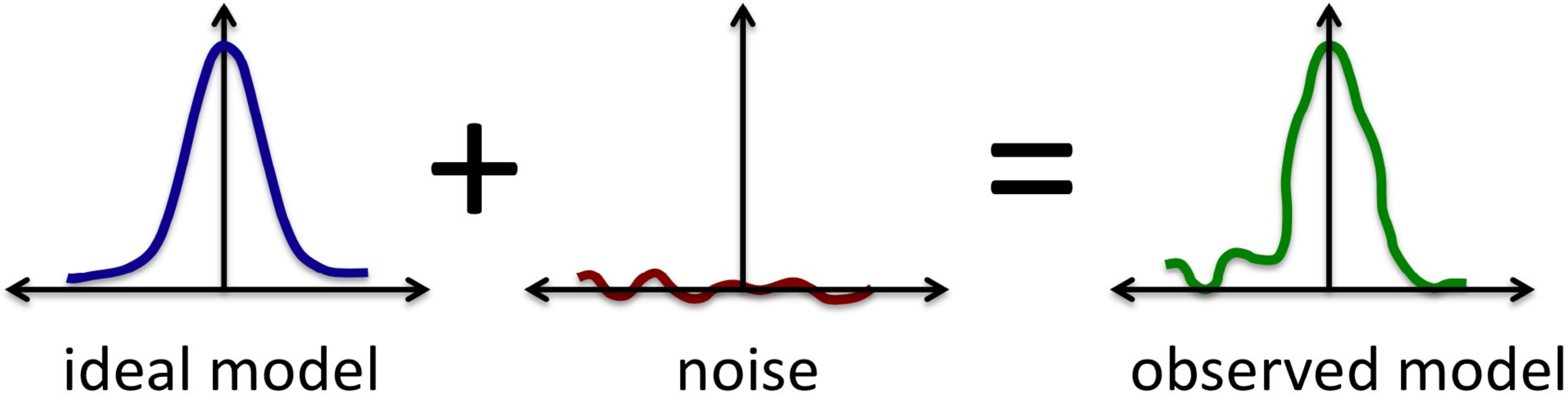
$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mu$$

empirical variance:

$$\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2 \to \sigma^2$$

Robust Parameter Estimation

Given **corrupted** samples from a 1-D Gaussian

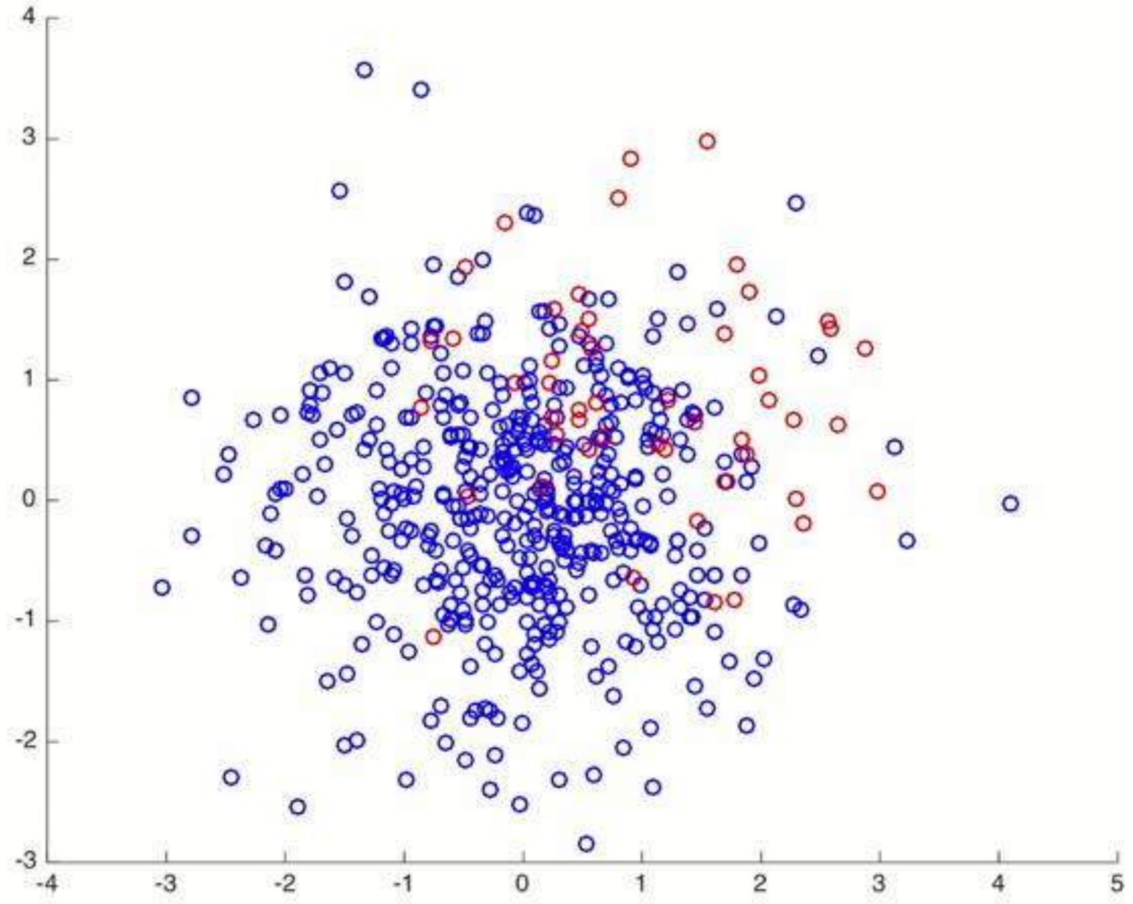


$\mathcal{N}(\mu, \sigma^2)$

can we accurately estimate its parameters?

Assumption on Noise

Adversary can arbitrarily corrupt ϵ-fraction of samples

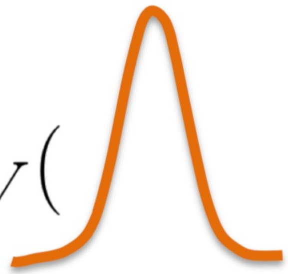


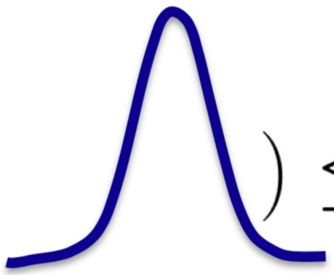
Total Variation Distance

Definition:

$$d_{TV}(f(x), g(x)) \triangleq \frac{1}{2}\int_{-\infty}^{\infty} \left| f(x) - g(x) \right| dx$$

Goal: find a 1-D Gaussian such that:

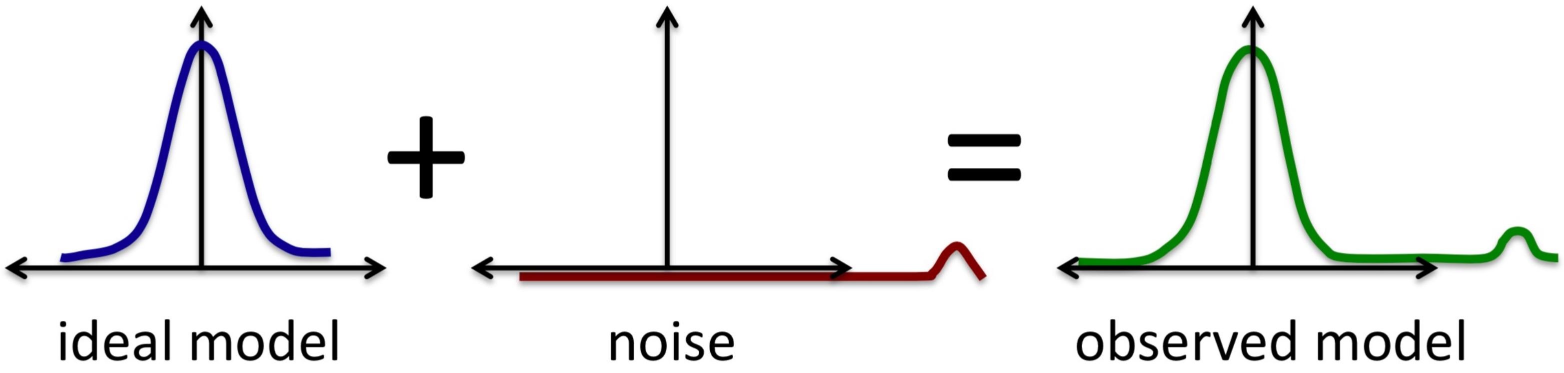
$d_{TV}($ 

, 

$) \leq O(\epsilon)$

Do empirical mean and variance work?

No!



But the median and median absolute deviation do work

$$\mathrm{MAD} = \mathrm{median}(|X_i - \mathrm{median}(X_1, X_2, ..., X_n)|)$$

Theorem (folklore)

Given ϵ-corrupted samples from a 1-D Gaussian $\mathcal{N}(\mu, \sigma^2)$

the median and MAD recover estimates that satisfy

$$d_{TV}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2)) \leq O(\epsilon)$$

where $\widehat{\mu} = \mathrm{median}(X),\ \widehat{\sigma} = \dfrac{\mathrm{MAD}}{\Phi^{-1}(3/4)}$

Median Without Noise

To prove that the median of $X \sim N\left(\mu, \sigma^2\right)$ is μ, we verify:

$$\Pr(X < \mu) = \int_{-\infty}^{\mu} f_X(x)\, \mathrm{d}x = \frac{1}{2}$$

$$\begin{aligned}
\int_{-\infty}^{\mu} f_X(x)\, \mathrm{d}x &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\frac{\mu-\mu}{\sqrt{2}\sigma}} \exp\left(-t^2\right) \mathrm{d}t \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{0} \exp\left(-t^2\right) \mathrm{d}t \\
&= \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-t^2\right) \mathrm{d}t \\
&= \frac{\sqrt{\pi}}{2\sqrt{\pi}}
\end{aligned}$$

substituting $t = \dfrac{x-\mu}{\sqrt{2}\sigma}$

Definite Integral of Even Function

Gaussian Integral

Median With Noise

$$\Phi(t) = \Pr_{X \sim \mathcal{N}(0,1)}[X \leq t]$$ is the cdf of the standard Gaussian

Theorem: Let S: ϵ-corrupted samples size n, $t \geq \Phi^{-1}(1/2 + \varepsilon)$

$$\Pr\left[|\mathrm{med}(S) - \mu| > t\sigma\right] \leq 2\exp(-2n(\Phi(t) - 1/2 - \varepsilon)^2)$$

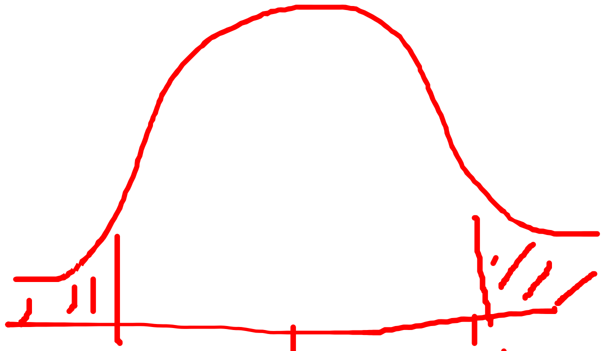
Proof

- We show $\Pr\left[\mathrm{med}(S) - \mu > t\sigma\right] \leq \exp(-2n(\Phi(t) - 1/2 - \varepsilon)^2)$
- By scaling, we can assume w.l.o.g. that $\sigma = 1$
- med(S) is at most $\left(\frac{1}{2} + \epsilon\right)$-quantile of S_{good} , since S_{bad} contains only ϵ fraction of points
- It suffices to show that the $\left(\frac{1}{2} + \epsilon\right)$-quantile of S_{good} is not too large

Proof (cont.)

- For each $i \in S_{\text{good}}$, let Y_i be a $\{0,1\}$-valued r.v.

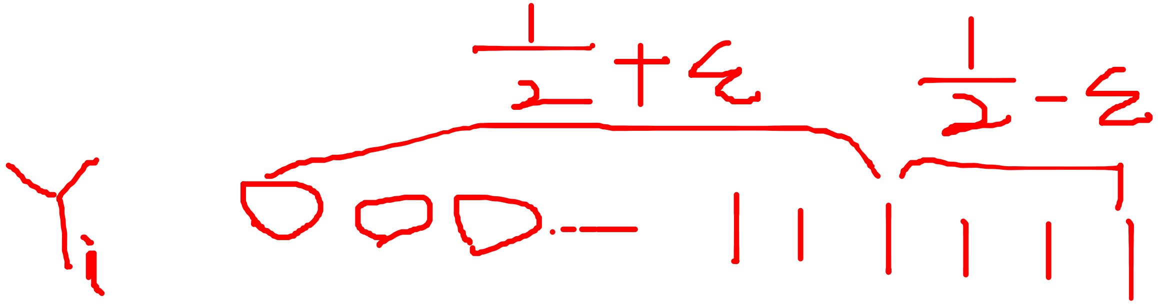
$$Y_i = \begin{cases} 1, & X_i - \mu > t \\ 0, & X_i - \mu \le t \end{cases}$$



- Y_i are i.i.d. Bernoulli r.v. and $\mathbb{E}[Y_i] = \Phi(-t) = 1 - \Phi(t)$

$-t$ μ $\mu+t$

- $\left(\frac{1}{2} + \epsilon\right)$-quantile of $S_{\text{good}} > \mu + t$ iff $\frac{1}{n}\sum_{i \in S_{\text{good}}} Y_i \ge 1/2 - \varepsilon.$



Proof (cont.)

- By Chernoff bound, for all s>0:

$$\Pr\left[\frac{1}{n}\sum_{i\in S_{\mathrm{good}}} Y_i > 1-\Phi(t)+s\right] \leq \exp(-2ns^2)$$

Plug in $s = \Phi(t) - 1/2 - \varepsilon$ proves the result

MAD Without Noise

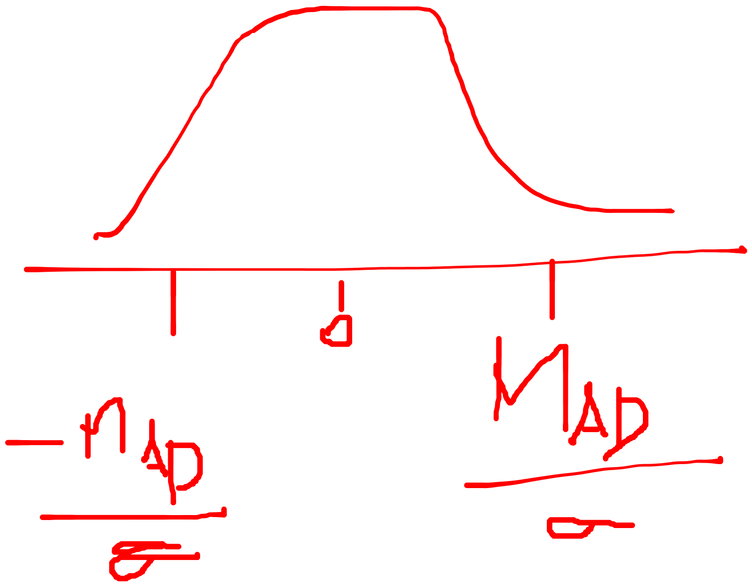
$$\text{MAD} = \text{median}(|X_i - \text{median}(X_1, X_2, \ldots, X_n)|)$$

$$\frac{1}{2} = P(|X-\mu| \le \text{MAD}) = P\left(\left|\frac{X-\mu}{\sigma}\right| \le \frac{\text{MAD}}{\sigma}\right) = P\left(|Z| \le \frac{\text{MAD}}{\sigma}\right)$$

$$\Phi\left(\text{MAD}/\sigma\right) - \Phi\left(-\text{MAD}/\sigma\right) = 1/2.$$

$$\Phi\left(-\text{MAD}/\sigma\right) = 1 - \Phi\left(\text{MAD}/\sigma\right)$$

$$\text{MAD}/\sigma = \Phi^{-1}\left(3/4\right) = 0.67449$$



Robustness in High Dimensions

Problem:

Given ϵ-corrupted samples from a d-dimensional Gaussian

$$\mathcal{N}(\mu, \sigma^2)$$

give an efficient algorithm to find parameters that satisfy

$$d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})) \leq \widetilde{O}(\epsilon)$$

Special Cases

(1) Unknown mean $\mathcal{N}(\mu, I)$

(2) Unknown covariance $\mathcal{N}(0, \Sigma)$

Can't We Learn Coordinate-wise?

- Each coordinate yields error $O(\epsilon)$, aggregating over all d dimensions yield an error of $O(\epsilon\sqrt{d})$
- Large error in high dimensions

Tukey Median

- Define Tukey depth of a point η in S:

$$\text{depth}(S, \eta) = \inf_{\|v\|_2 = 1} \frac{|\{X \in S : \langle X - \eta, v \rangle \geq 0\}|}{n}$$

- Then Tukey median is defined as

$$\text{Tukey}(S) = \arg\max_{\eta} \text{depth}(S, \eta)$$

Tukey Median

- Tukey median achieve true mean with error $O(\epsilon)$
- But it is **NP-hard** to find the Tukey median

Efficient Algorithm in High Dimensions

[Diakonikolas et al. FOCS'16]

The algorithm uses $N = \widetilde{O}(d^3/\epsilon^2)$ samples from a d-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$ with ϵ-corruption , and finds parameters that satisfy

$$d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})) \leq O(\epsilon \log^{3/2} 1/\epsilon)$$

Moreover, the algorithm runs in poly(N, d)

Unknown Mean Case

Lemma: $$d_{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\widehat{\mu}, I)) \leq \frac{\|\mu - \widehat{\mu}\|_2}{2}$$

This can be proven using Pinsker's Inequality

$$d_{TV}(f,g)^2 \leq \frac{1}{2}\, d_{KL}(f,g)$$

And properties of KL-divergence between Gaussians

Unknown Mean Case

Lemma:

$$d_{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\widehat{\mu}, I)) \leq \frac{\|\mu - \widehat{\mu}\|_2}{2}$$

Corollary: If our estimate (in the unknown mean case) satisfies

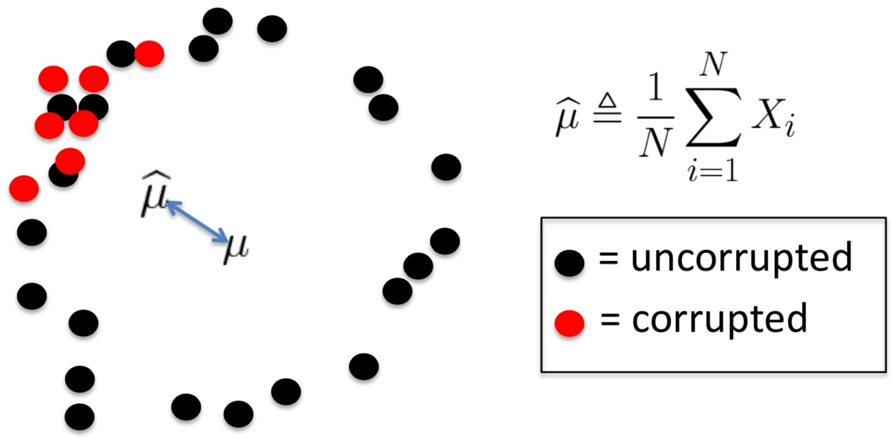
$$\|\mu - \widehat{\mu}\|_2 \leq \widetilde{O}(\epsilon)$$

then $d_{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\widehat{\mu}, I)) \leq \widetilde{O}(\epsilon)$

Our new goal is to be close in **Euclidean distance**

Detecting Corruptions

- If the corruption move the mean, they also change the covariance matrix



$$\widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i$$

- We know the naïve estimator has been compromised if there is a direction of large (>1) variance

Key Lemma

If $X_1, X_2, \ldots, X_N$ come from ϵ-corrupted $\mathcal{N}(\mu, I)$, and

$N \geq 10(d + \log\frac{1}{\delta})/\epsilon^2$, then for

(1) $\widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i$ (2) $\widehat{\Sigma} \triangleq \frac{1}{N}\sum_{i=1}^{N}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T$

with probability at least 1-δ

$$\|\mu - \widehat{\mu}\|_2 \geq C\epsilon\sqrt{\log 1/\epsilon} \longrightarrow \|\widehat{\Sigma} - I\|_2 \geq C'\epsilon \log 1/\epsilon$$

Filtering-based Algorithm

- Suppose that $\|\widehat{\Sigma} - I\|_2 \geq C'\epsilon \log 1/\epsilon$
- Find direction v of largest variance (top eigen vector)
- Project data in the direction of v and remove the largest data points in this direction
- Repeat until there are not corrupted points left

Running Time: $\widetilde{O}(Nd^2)$ **Sample Complexity:** $\widetilde{O}(d^2/\epsilon^2)$

Unknown Covariance Case

Again, by using Pinsker's Inequality:

$$d_{TV}(\mathcal{N}(0,\Sigma),\mathcal{N}(0,\widehat{\Sigma})) \leq O(\|I-\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}\|_F)$$

Our new goal is to find $\hat{\Sigma}$ that satisfies:

$$\|I-\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}\|_F \leq \widetilde{O}(\epsilon)$$

Unknown Covariance Case

Key Idea: Transform the data, look for restricted large eigenvalues

$$Y_i \triangleq (\widehat{\Sigma})^{-1/2} X_i$$

If $\widehat{\Sigma}$ were the true covariance, we would have $Y_i \sim N(0, I)$ for inliers, in which case:

$$\frac{1}{N}\sum_{i=1}^{N}\Big(Y_i \otimes Y_i\Big)\Big(Y_i \otimes Y_i\Big)^T - 2I$$

would have small restricted eigenvalues

Take-away: An adversary needs to mess up the (restricted) **fourth** moment in order to corrupt the **second** moment

Putting It All Together

1. Doubling trick: $X_i - X_i' \sim_\epsilon \mathcal{N}(0, 2\Sigma)$
   - Now use algorithm for unknown covariance
2. Transform into isotropic position

$$\widehat{\Sigma}^{-1/2} X_i \sim_\epsilon \mathcal{N}(\widehat{\Sigma}^{-1/2}\mu, I)$$

   - Now use algorithm for unknown mean

Beyond Robust Statistics

- Can we "robustify" more complicated objectives, like supervised learning? E.g., regression, SVM
- These problems can be solved in the framework of stochastic optimization:

Given a loss function $\ell(X, w)$ and a distribution $\mathcal{D}$ over X, minimize

$$f(w) = \mathbb{E}_{X \sim \mathcal{D}}\left[\ell(X, w)\right]$$

- Challenge: Given ϵ-corrupted samples from $\mathcal{D}$, minimize f

SEVER: Robust Stochastic Optimization

[Diakonikolas et al. ICML'2019]

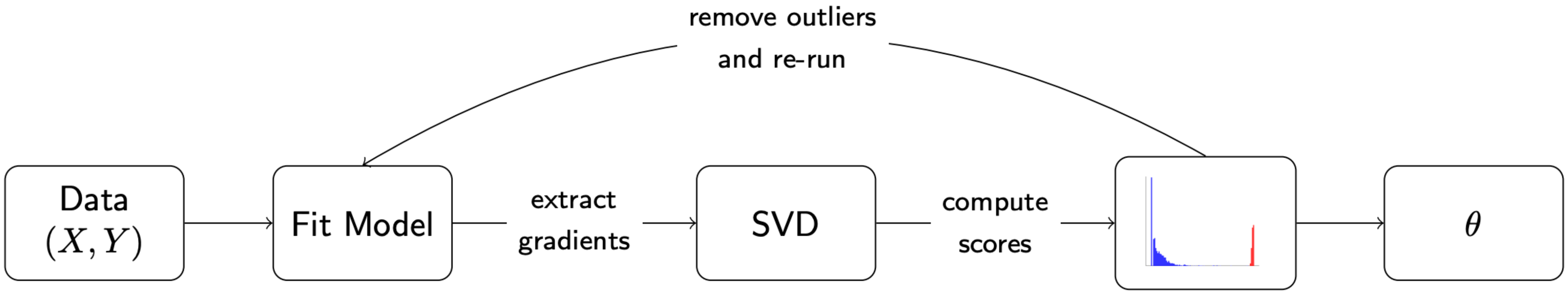
SGD with robust estimates

$$w_{t+1} \leftarrow w_t - \eta_t \cdot g_t$$

where g_t is a robust estimate of $\nabla f(w_t)$

This straightforward approach is slow

Idea: only filter at minimizer of the empirical risk



SEVER: Robust Stochastic Optimization

[Diakonikolas et al. ICML'2019]

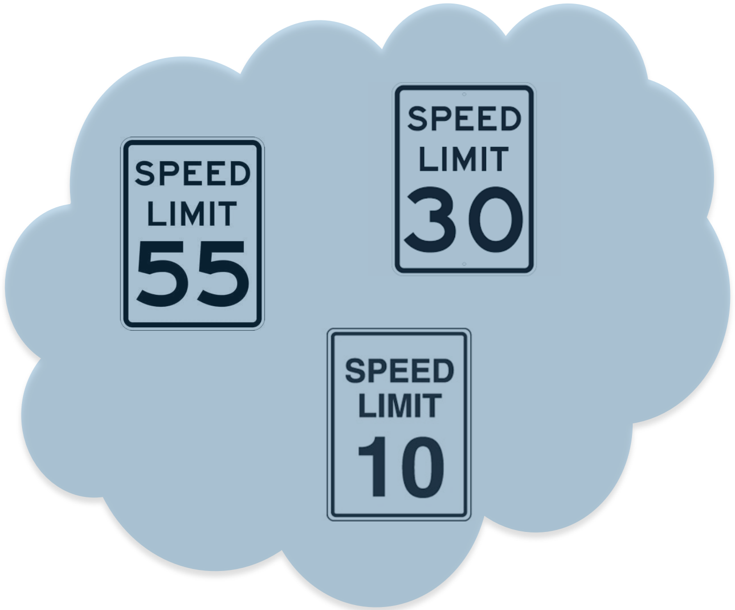
Theorem: Suppose ℓ is convex, and $\mathrm{Cov}\,[\nabla\ell(X, w)] \preccurlyeq \sigma^2 I$. Under mild assumptions on $\mathcal{D}$, then SEVER outputs a $\widehat{w}$ so that w.h.p.

$$f(\widehat{w}) - \min_{w} f(w) < O\left(\sqrt{\sigma^2 \varepsilon}\right).$$

Defense to Backdoor Attacks

[Tran et al. NeurIPS'18]

- Representation space of training data:



Empirically, attack causes noticeable perturbation in the covariance → Detect the corruption with previous algorithm

Summary

- There exists an efficient algorithm for learning a high-dimensional ϵ-corrupted Gaussian
- Can be used in stochastic optimization problems
- May be used in general outlier detection