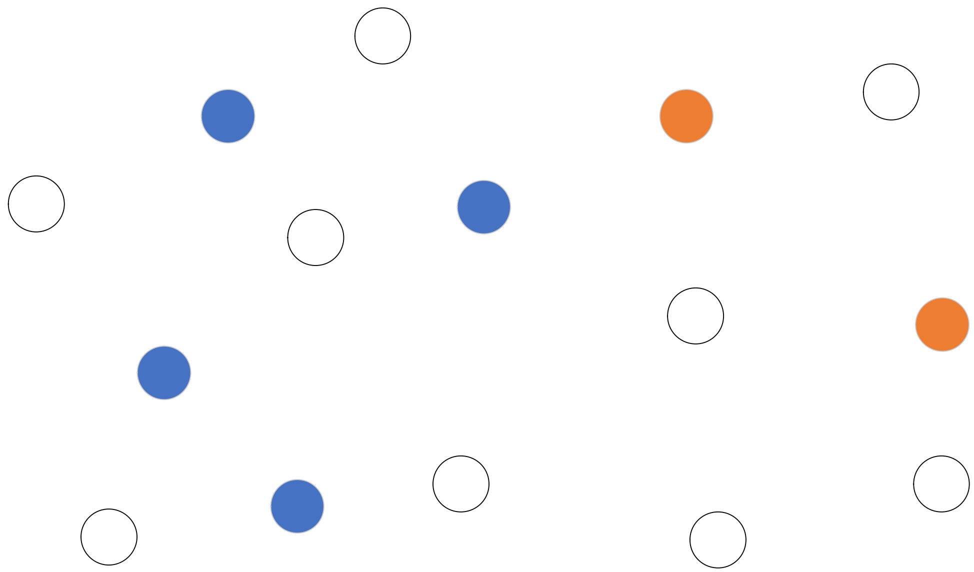
UNIVERSITY of LOUISIANA LAFAYETTE

Poisoning the Unlabeled Dataset of Semi-Supervised Learning

Nicholas Carlini, Google
USENIX Security Symposium, 2021

Semi-supervised learning

- **Collecting "Labeled" Data is Expensive.**



Unlabeled dataset $\mathcal{U}$

Labeled examples $\mathcal{X}$

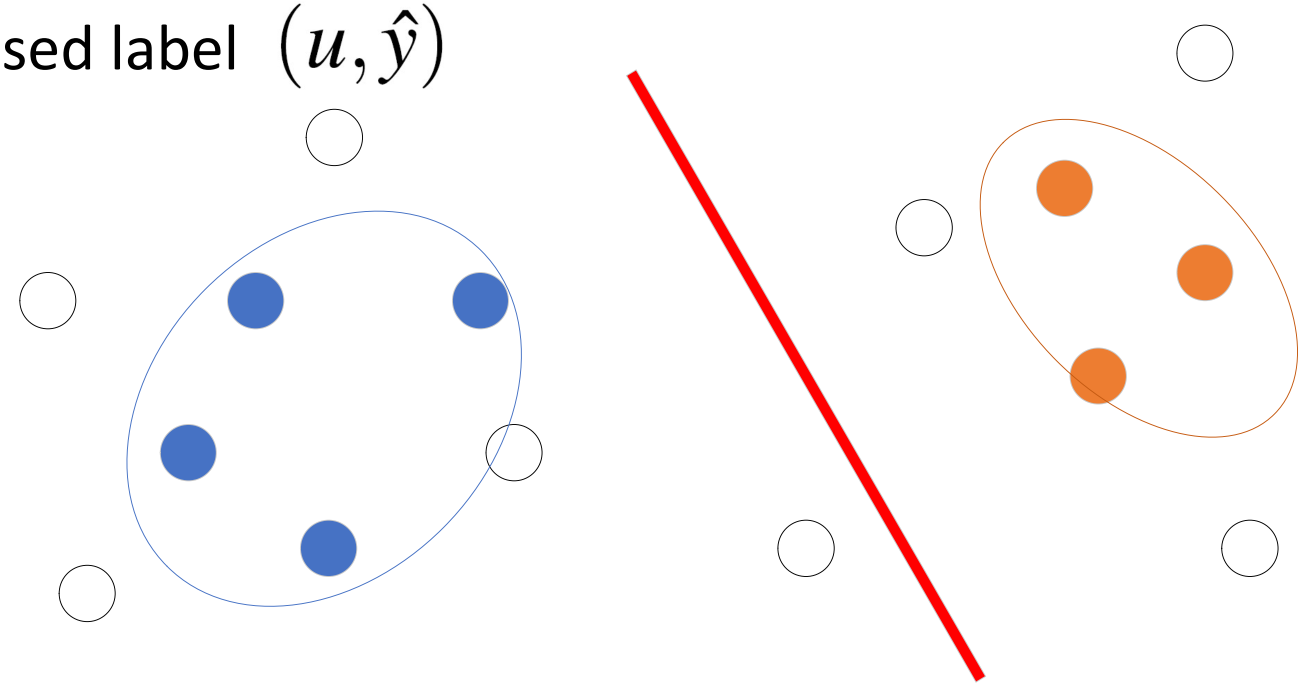
Supervised model f

$$f_\theta \leftarrow \mathcal{T}_s(f, \mathcal{X}, \mathcal{U})$$

Semi-supervised learning

- **Transform into Fully-supervised Problem**

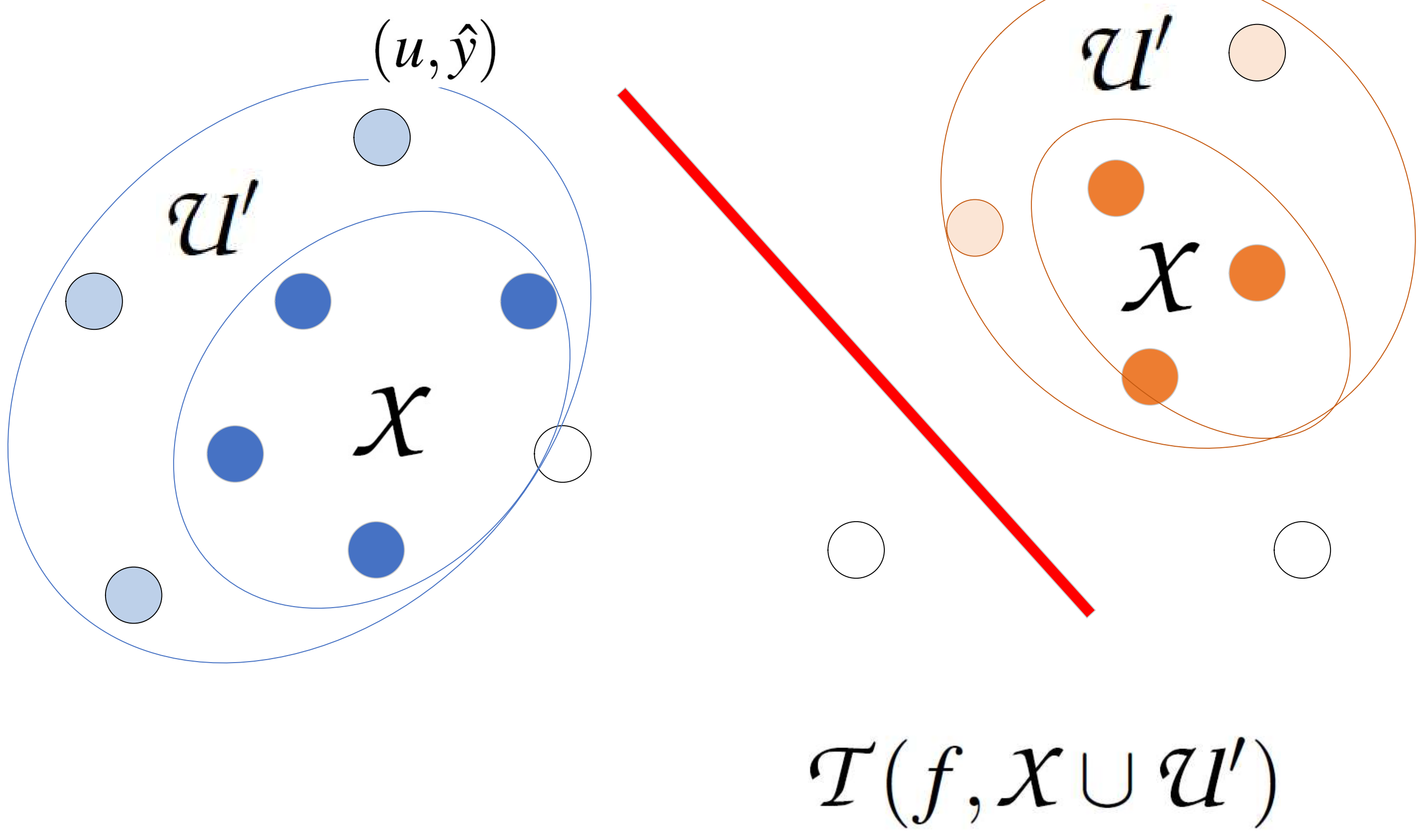
guessed label $(u, \hat{y})$



$$\hat{y} = f(u; \theta_i)$$

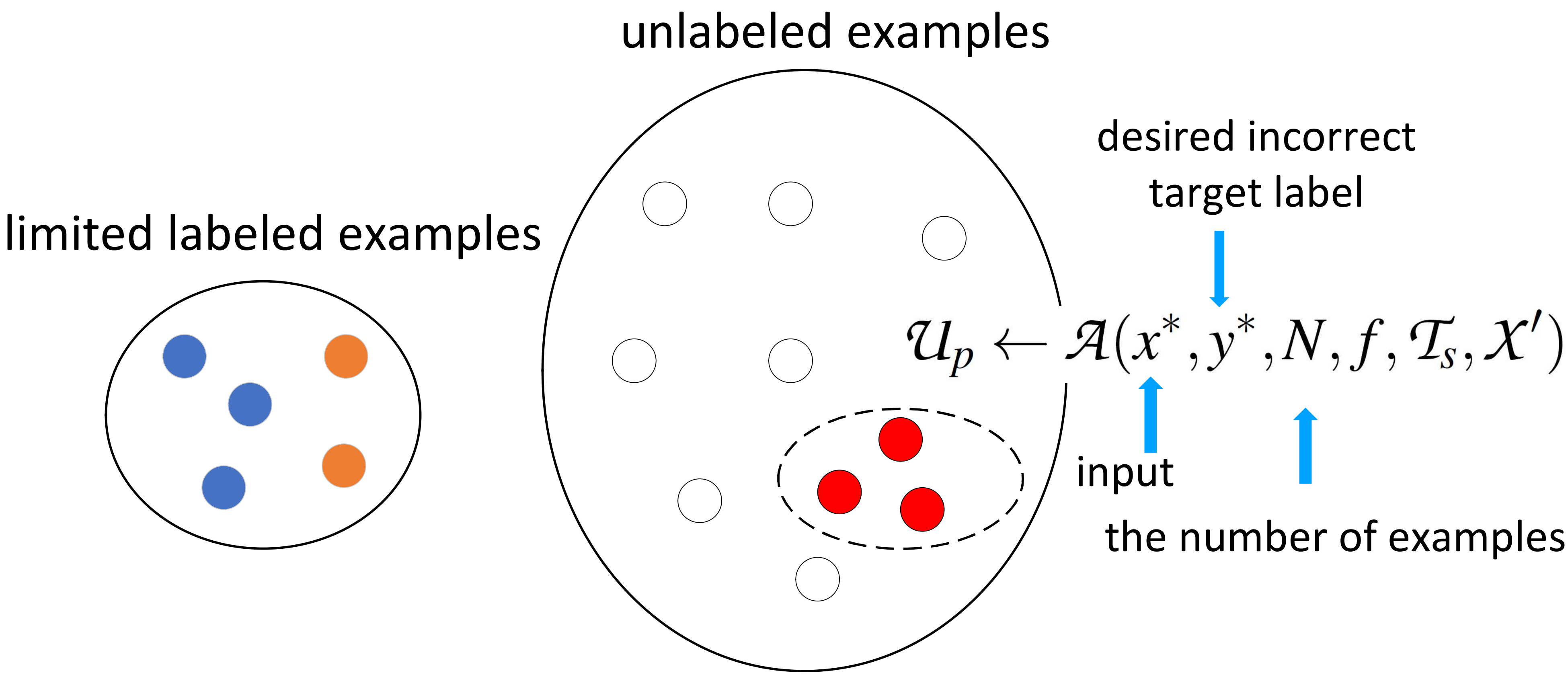
Semi-supervised learning

- **Transform into Fully-supervised Problem**



Threat Model

- **Transform into Fully-supervised Problem**

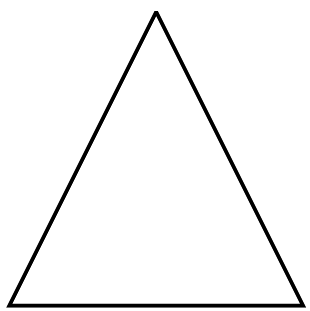


$$f_\theta \leftarrow \mathcal{T}_s(X, \mathcal{U} \cup \mathcal{U}_p) \Longrightarrow f_\theta(x^*) = y^*$$

Poisoning the Unlabeled Dataset

- **Problem**

How a task should be completed

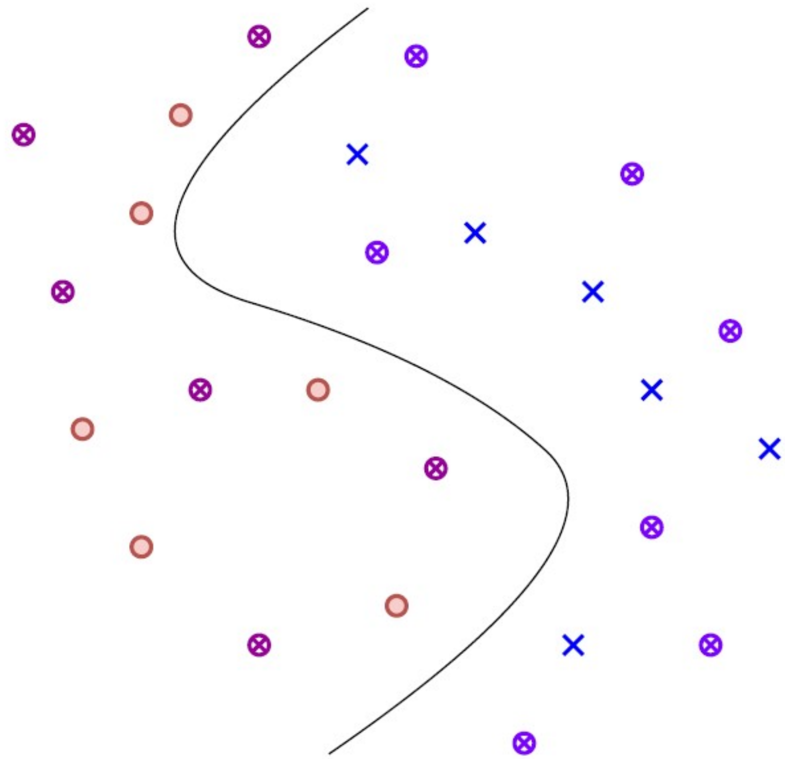


What should be done

Teach itself from this unlabeled data

Poisoning the Unlabeled Dataset

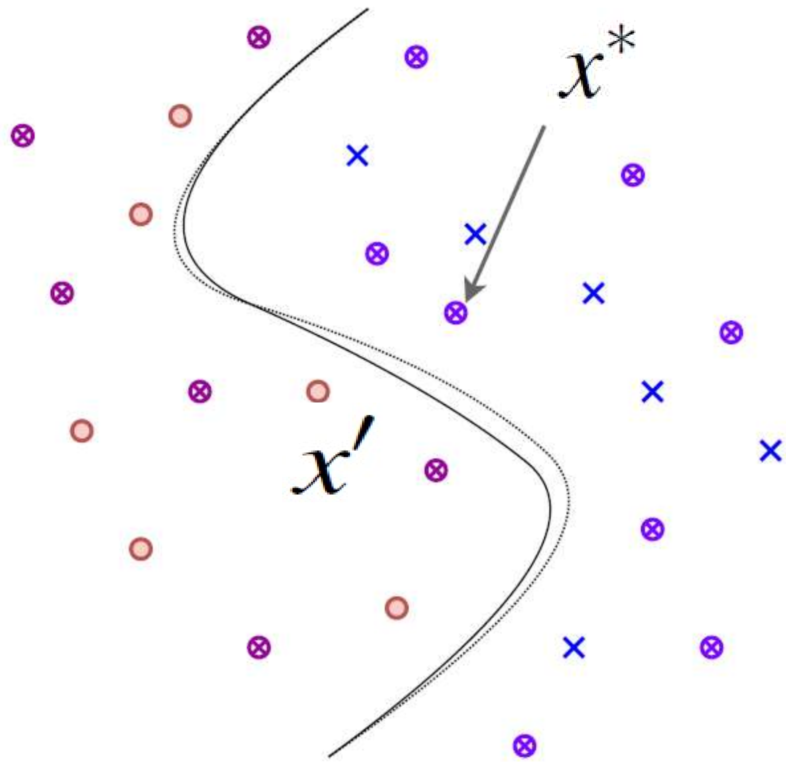
- **Interpolation Consistency Poisoning**



(a) A classifier trained on a semi-supervised dataset of red $\odot$s, blue $\times$s, and *unlabeled* $\otimes$s. During training the unlabeled $\otimes$s are given pseudo-labels such that the correct original decision boundary is learned.

Poisoning the Unlabeled Dataset

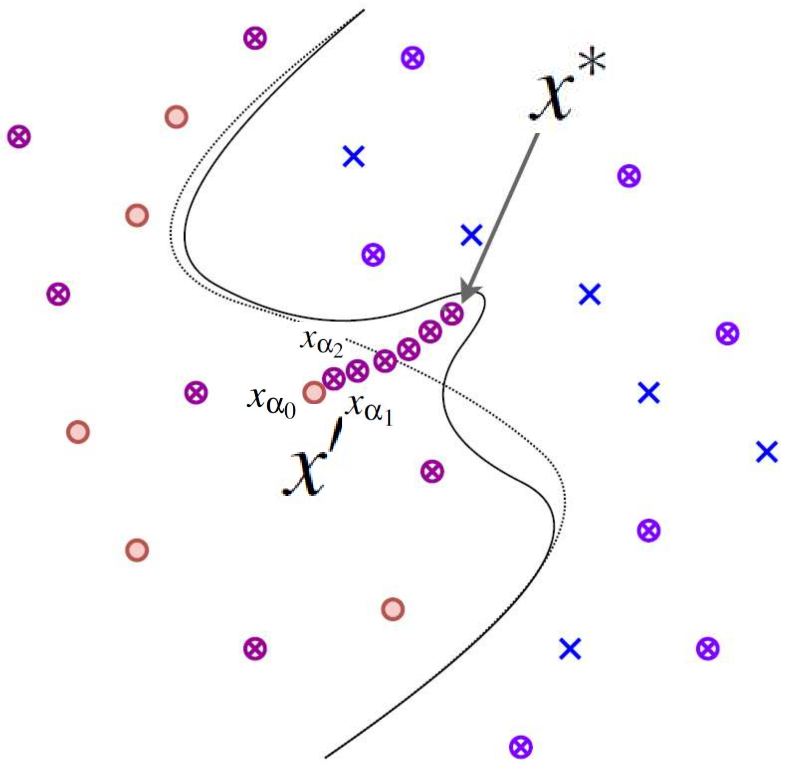
- **Interpolation Consistency Poisoning**



(b) When inserting just one new *unlabeled* poisoned example near the boundary, the model gives it the correct pseudo label of the blue $\times$s. The poisoning attempt fails, and the decision boundary remains largely unchanged.

Poisoning the Unlabeled Dataset

- **Interpolation Consistency Poisoning**



$$\{x_{\alpha_i}\}_{i=0}^{N-1} = \text{interp}(x', x^*, \alpha_i)$$

$$\text{interp}(x', x^*, 0) = x'$$

$$\text{interp}(x', x^*, 1) = x^*$$

$$f(x_{\alpha_0}) = f(x_{N-1}) = f(x^*) = y^*$$

Poisoning the Unlabeled Dataset

- **Interpolation Strategy**

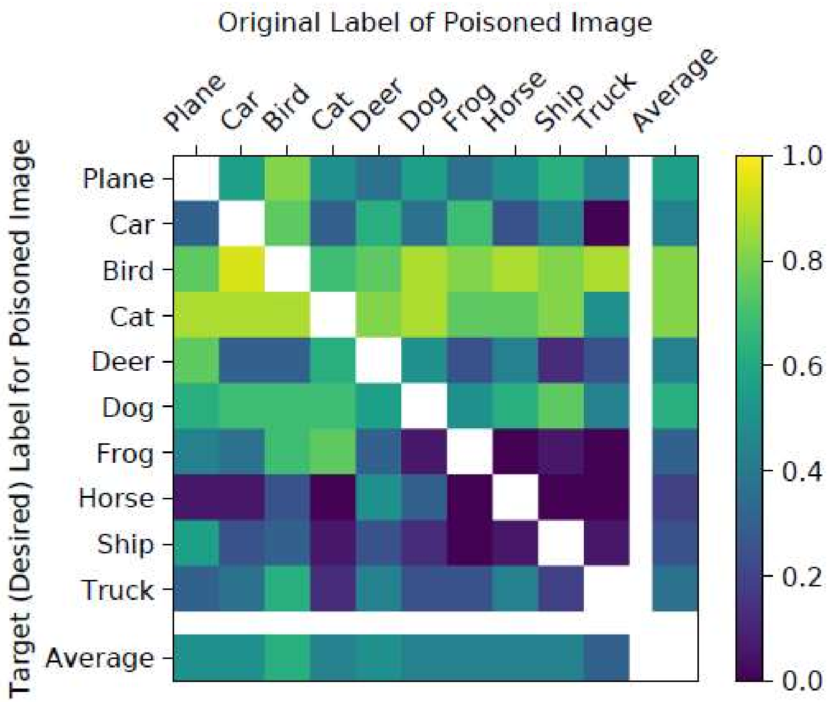
$$\text{interp}(x', x^*, \alpha) = x' \cdot (1 - \alpha) + x^* \cdot \alpha$$

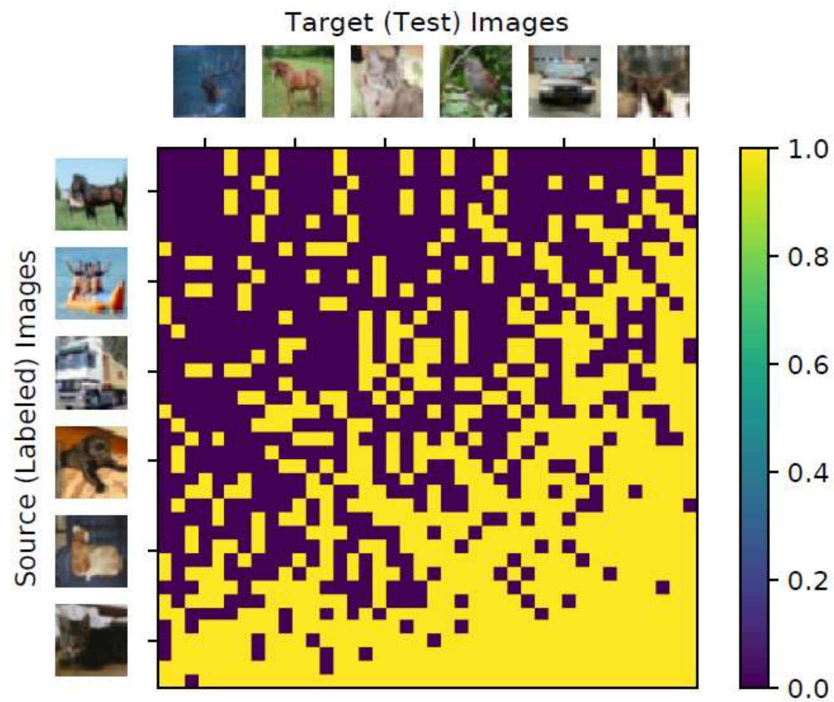
- **Density of poisoned samples.**

$$\hat{\rho}(x) = \rho(x) \cdot \left(\int_0^1 \rho(x)\, dx \right)^{-1} \quad \Pr[p < \alpha < q] = \int_p^q \hat{\rho}(x)\, dx$$

Evaluation

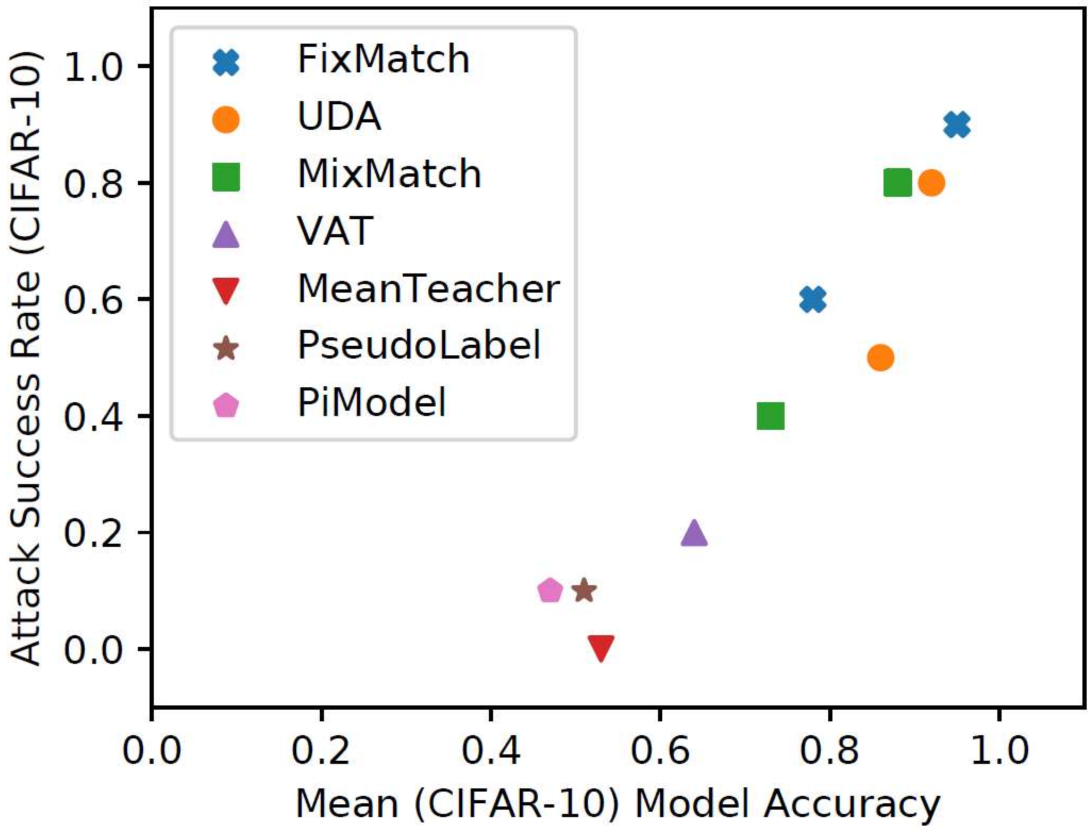
- **Evaluation across source- and target-image**





Evaluation

- **Evaluation across training techniques**



Evaluation

- **Evaluation across datasets**

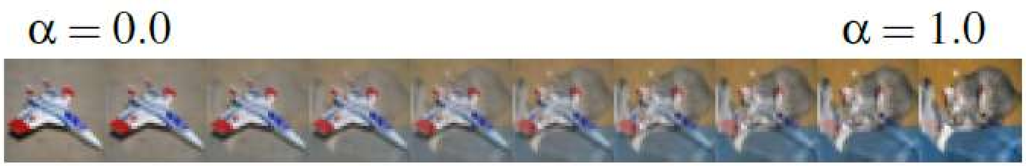
Dataset (% poisoned)	CIFAR-10 0.1%	CIFAR-10 0.2%	CIFAR-10 0.5%	SVHN 0.1%	SVHN 0.2%	SVHN 0.5%	STL-10 0.1%	STL-10 0.2%	STL-10 0.5%
MixMatch	5/8	6/8	8/8	4/8	5/8	5/8	4/8	6/8	7/8
UDA	5/8	7/8	8/8	5/8	5/8	6/8	-	-	-
FixMatch	7/8	8/8	8/8	7/8	7/8	8/8	6/8	8/8	8/8

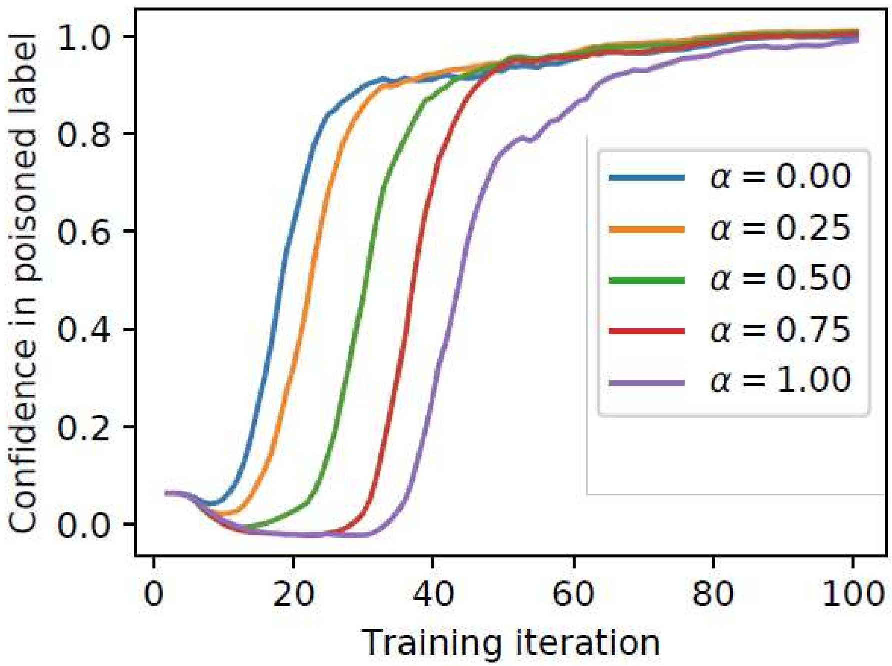
- **Evaluation across number of labeled examples**

Dataset (# labels)	CIFAR-10 40	CIFAR-10 250	CIFAR-10 4000	SVHN 40	SVHN 250	SVHN 4000
MixMatch	5/8	4/8	1/8	6/8	4/8	5/8
UDA	5/8	5/8	2/8	5/8	4/8	4/8
FixMatch	7/8	7/8	7/8	7/8	6/8	7/8

Evaluation

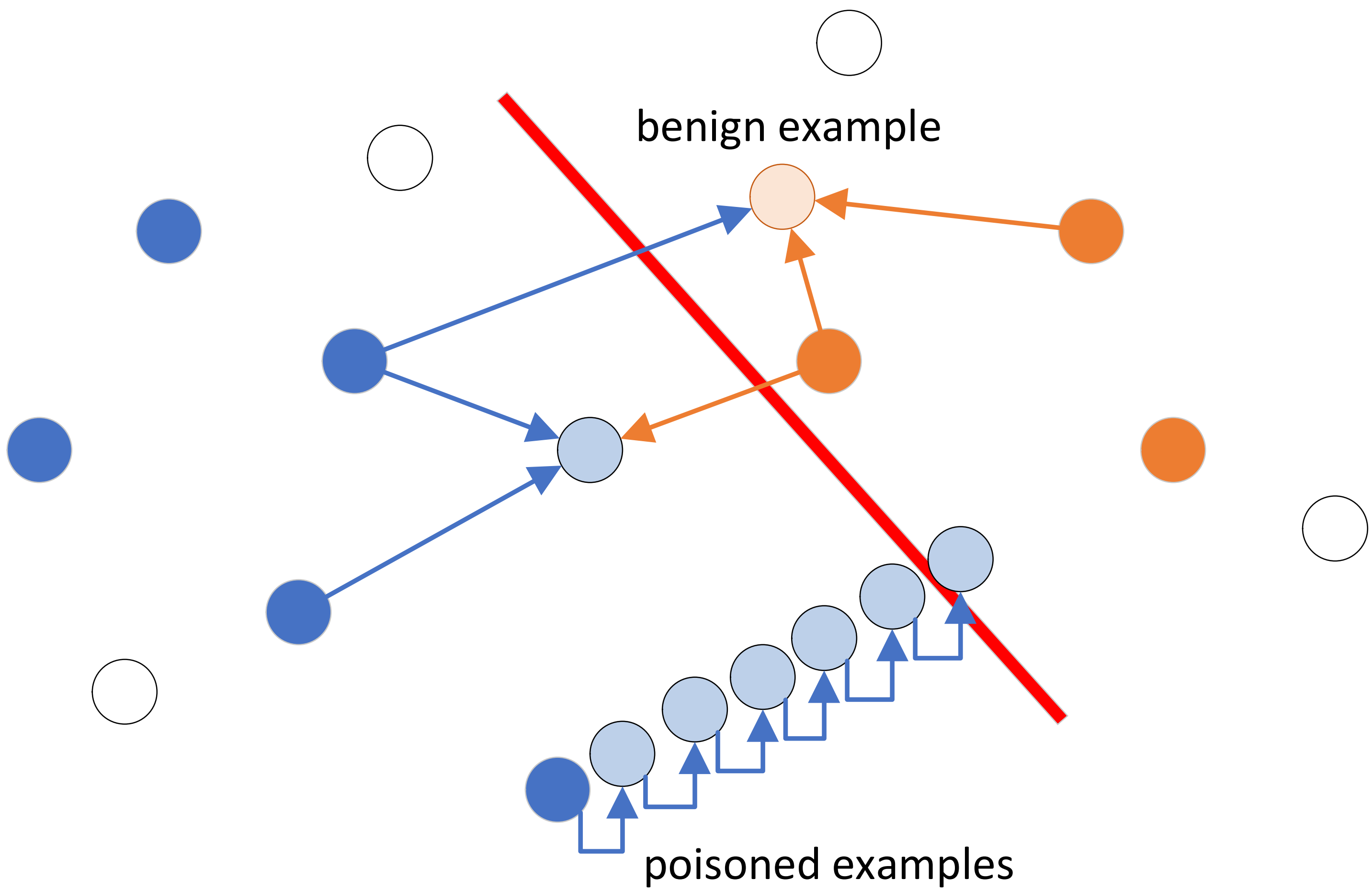
- **Why does this attack work?**





Defense

- **Monitoring Training Dynamics**



Defense

- **Computing Pairwise Influence**

Difference in the model f predictions on example j from i epoch to i+1.

$$\partial f_{\Theta_i}(u_j) = f_{\Theta_{i+1}}(u_j) - f_{\Theta_i}(u_j)$$

Model's collection of prediction difference from epoch a to epoch b.

$$\mu_j^{(a,b)} = \left[\partial f_{\Theta_a}(u_j) \quad \partial f_{\Theta_{a+1}}(u_j) \quad \dots \quad \partial f_{\Theta_{b-1}}(u_j) \quad f_{\Theta_b}(u_j)\right]$$

The influence of example u_i on u_j

$$\text{Influence}(u_i, u_j) = \|\mu_i^{(0,K-2)} - \mu_j^{(1,K-1)}\|_2^2$$

Defenses

- **Identifying Poisoned Example**