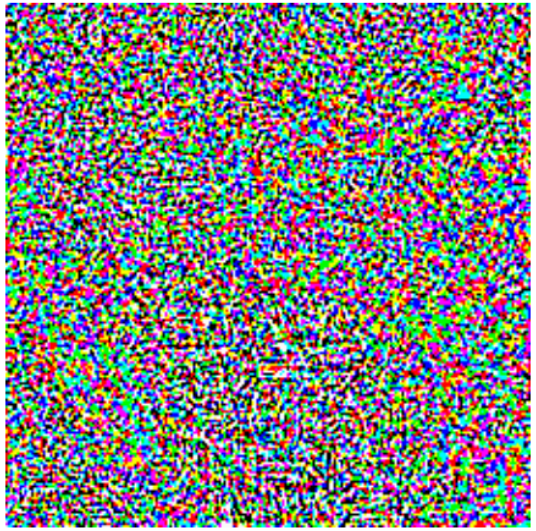
Data Poisoning Attacks

Shusen Wang

Data Evasion Attack

$+\ .007\ \times$



$=$

"panda"
57.7% confidence

"gibbon"
99.3% confidence

Reference

- Goodfellow, Shlens, and Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

Data Evasion v.s. Data Poisoning

- **Data Evasion attack** [1, 2] happens at test time.
- Perturb a test sample so that the model makes a classification error.

Reference

1. Biggio, et. al. Evasion attacks against machine learning at test time. In *ECML*, 2013.
2. Szegedy et al. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

Data Evasion v.s. Data Poisoning

- **Data Evasion attack** [1, 2] happens at test time.
- Perturb a test sample so that the model makes a classification error.

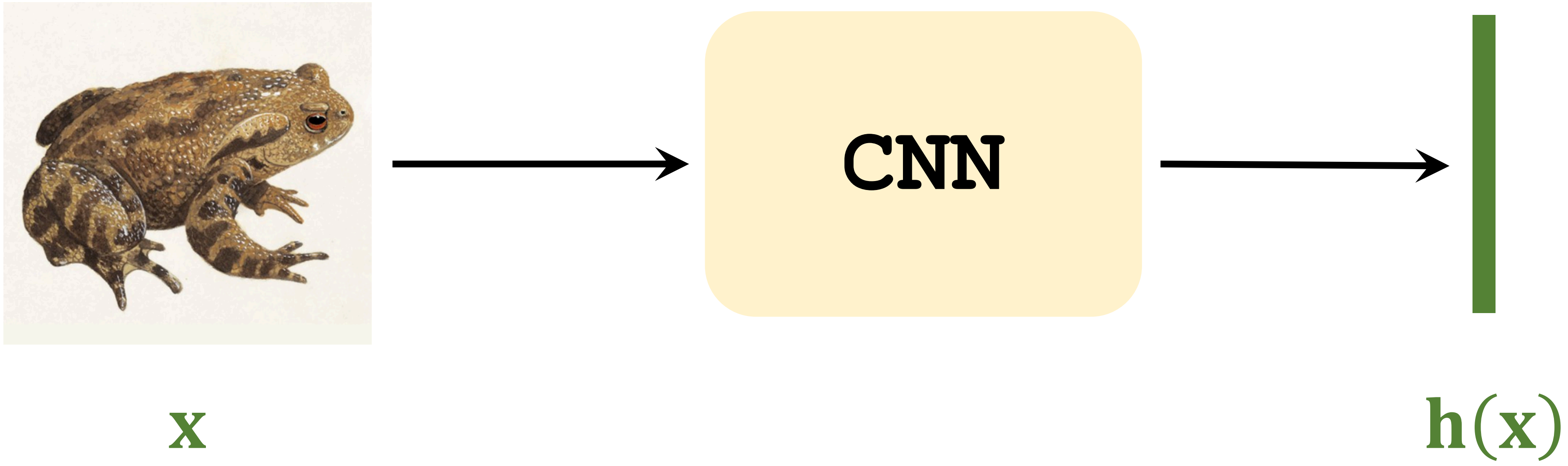
- **Data poisoning attack** [3] happens at training time.
- Add a poison sample to the training set.
- The trained model will make the mistake as the attacker planned.

Reference

1. Biggio, et. al. Evasion attacks against machine learning at test time. In *ECML*, 2013.
2. Szegedy et al. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.
3. Shafahi et al. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*, 2018.

Feature extraction using CNN

- $\mathbf{x}$: input image.
- $\mathbf{h}(\mathbf{x})$: feature vector extracted by CNN.
- Function $\mathbf{h}$ includes the layers between input layer and flatten layer.



Create a poison sample

- $\mathbf{x}_{\text{victim}}$: victim sample (an image not in the training set).
- Add perturbation $\boldsymbol{\delta}^\star$ to $\mathbf{x}$ such that $\mathbf{h}(\mathbf{x}+\boldsymbol{\delta}^\star) \approx \mathbf{h}(\mathbf{x}_{\text{victim}})$.

Create a poison sample

- $\mathbf{x}_{\text{victim}}$: victim sample (an image not in the training set).
- Add perturbation $\boldsymbol{\delta}^{\star}$ to $\mathbf{x}$ such that $\mathbf{h}(\mathbf{x}+\boldsymbol{\delta}^{\star}) \approx \mathbf{h}(\mathbf{x}_{\text{victim}})$.

- Find the perturbation by optimization:

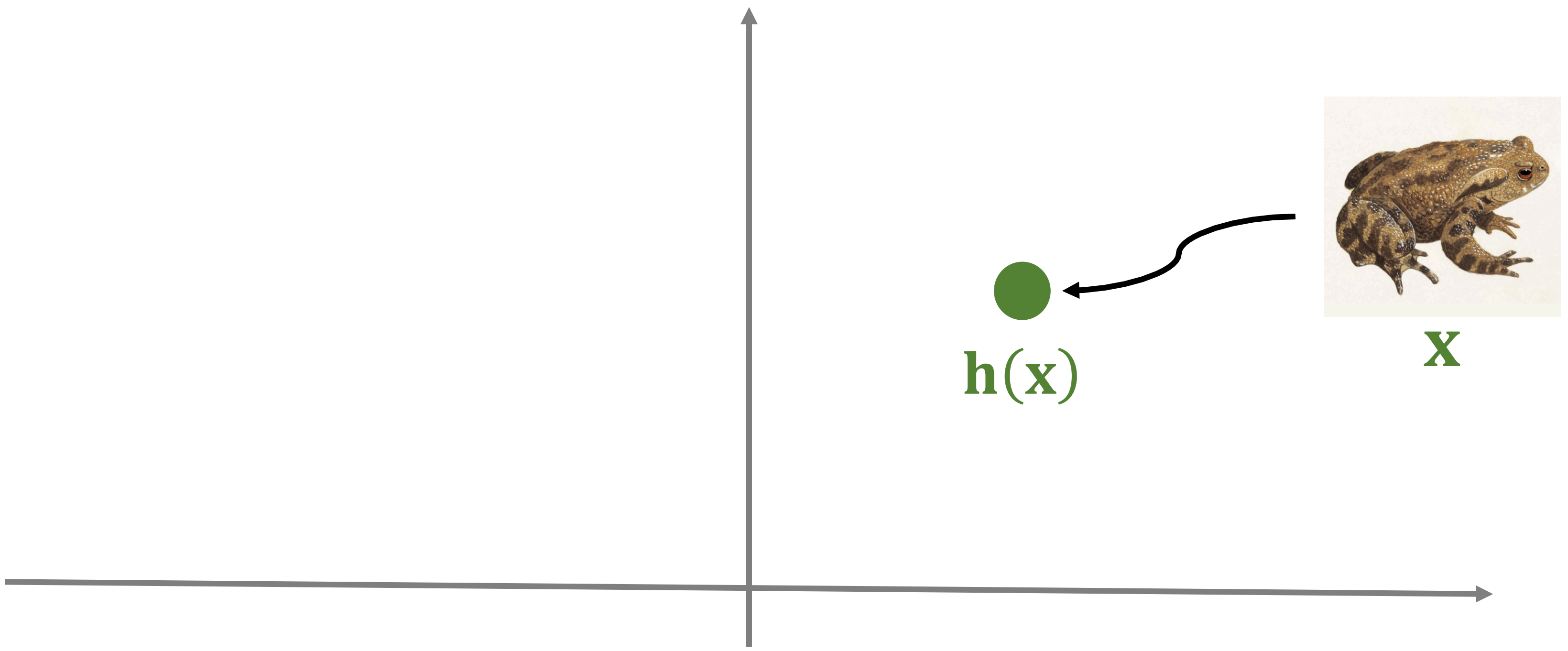
$$\boldsymbol{\delta}^{\star} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \ \left|\left|\mathbf{h}(\mathbf{x}+\boldsymbol{\delta}) - \mathbf{h}(\mathbf{x}_{\text{victim}})\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\delta}\right|\right|_2^2 .$$

The feature vectors are similar.

The perturbation is small.

Create a poison sample

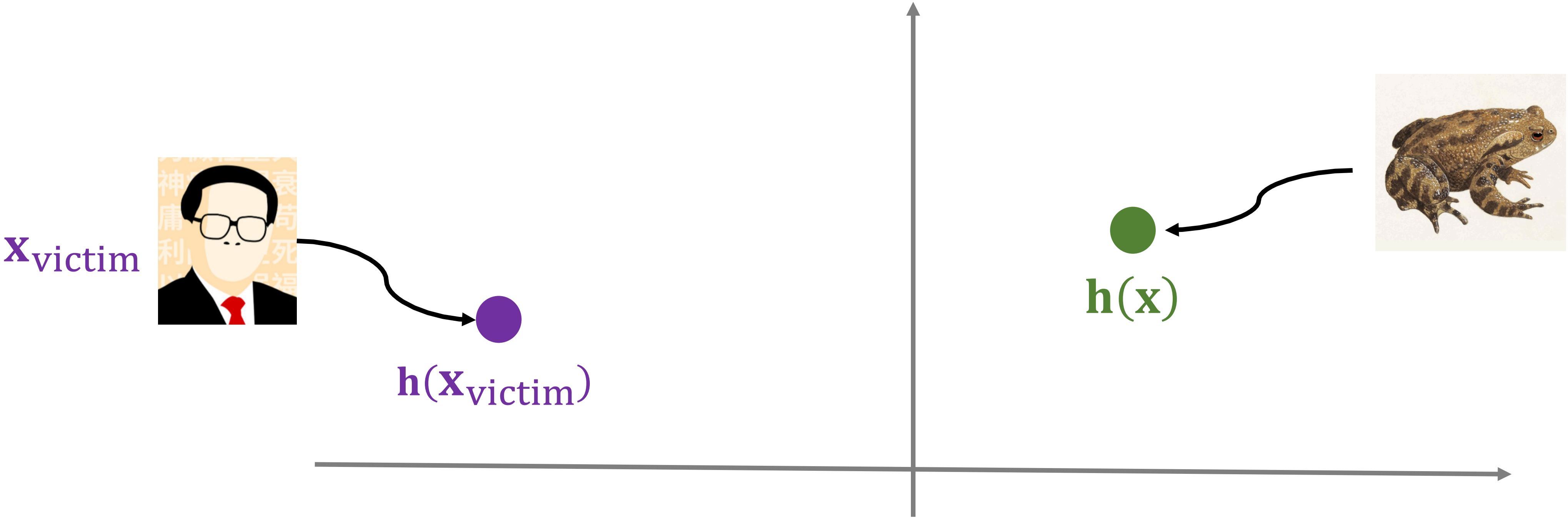
- $\mathbf{x}_{\text{victim}}$: victim sample (an image not in the training set).
- Add perturbation $\boldsymbol{\delta}^{\star}$ to $\mathbf{x}$ such that $\mathbf{h}(\mathbf{x}+\boldsymbol{\delta}^{\star}) \approx \mathbf{h}(\mathbf{x}_{\text{victim}})$.



Feature space

Create a poison sample

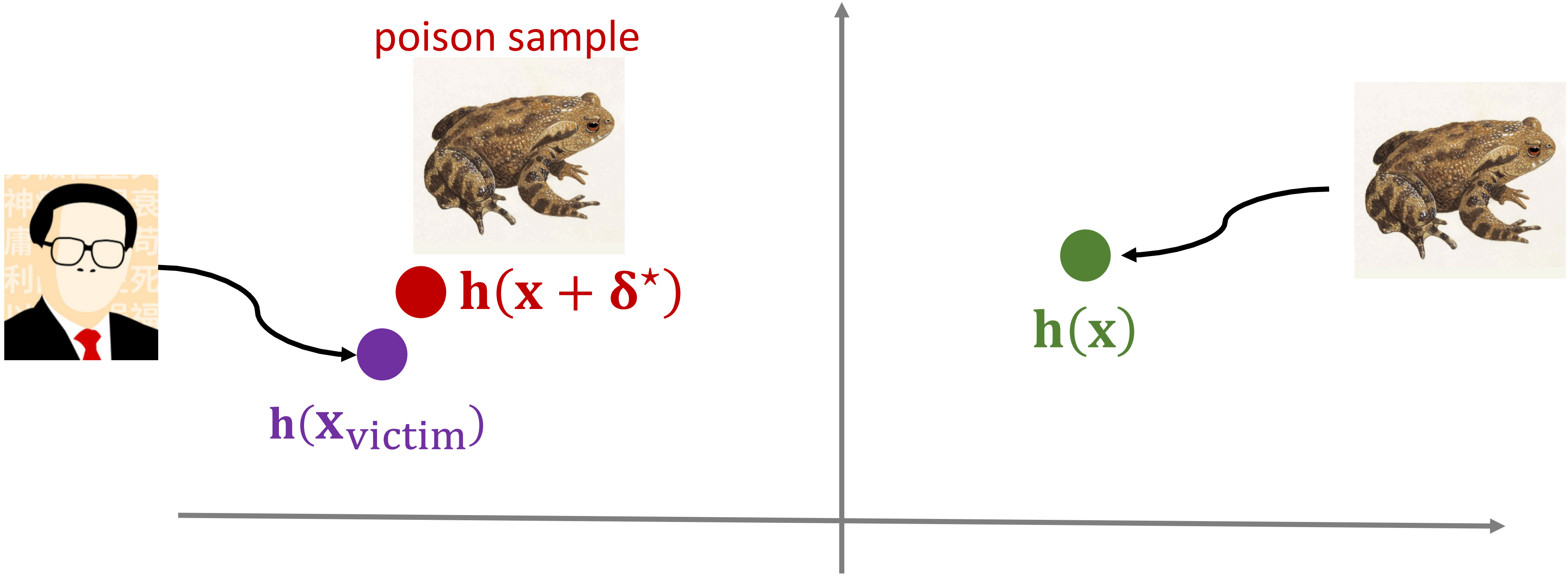
- $\mathbf{x}_{\text{victim}}$: victim sample (an image not in the training set).
- Add perturbation $\boldsymbol{\delta}^{\star}$ to $\mathbf{x}$ such that $\mathbf{h}(\mathbf{x} + \boldsymbol{\delta}^{\star}) \approx \mathbf{h}(\mathbf{x}_{\text{victim}})$.



Feature space

Create a poison sample

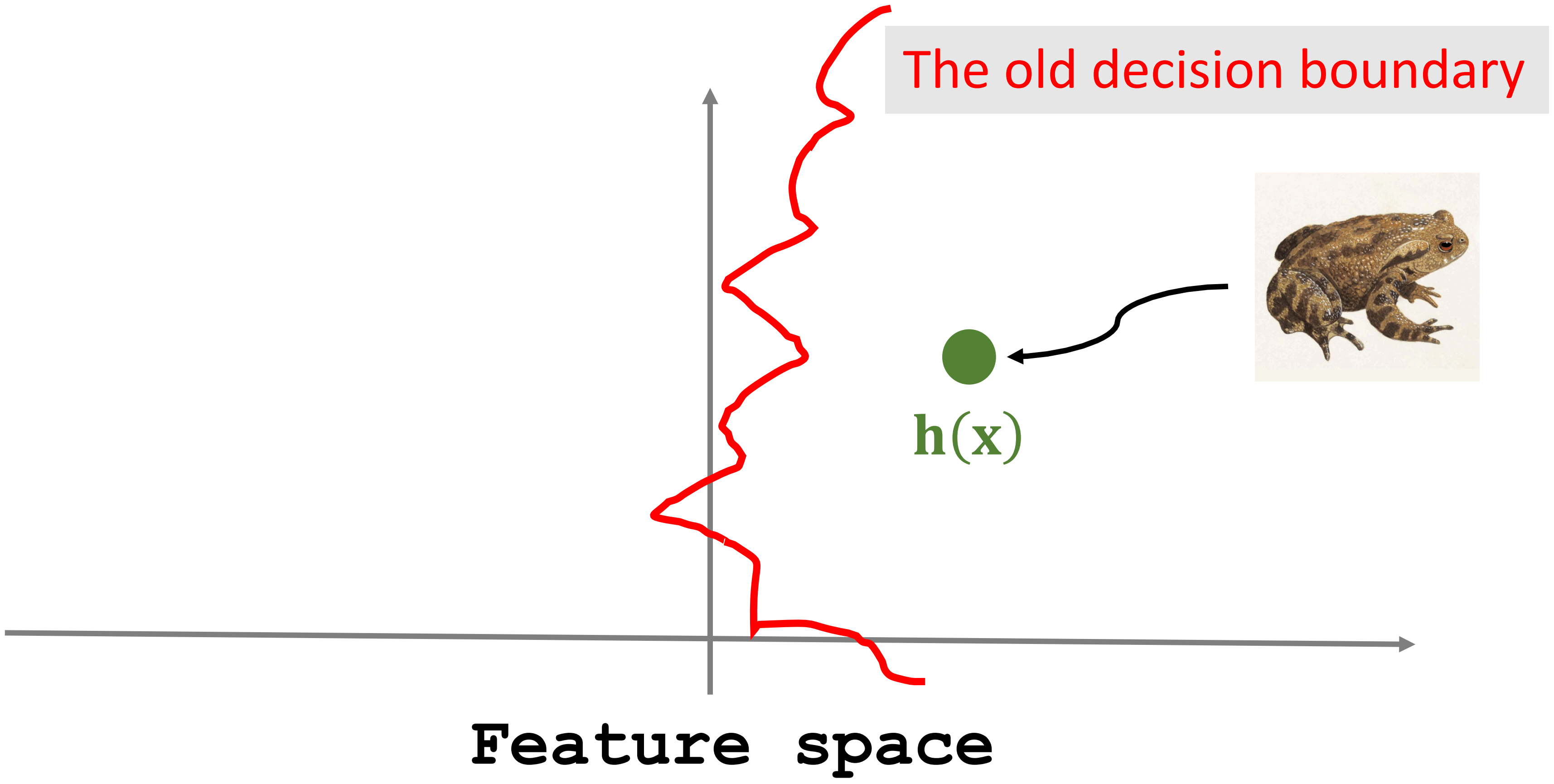
- $\mathbf{x}_{\text{victim}}$: victim sample (an image not in the training set).
- Add perturbation $\boldsymbol{\delta}^{\star}$ to $\mathbf{x}$ such that $\mathbf{h}(\mathbf{x}+\boldsymbol{\delta}^{\star}) \approx \mathbf{h}(\mathbf{x}_{\text{victim}})$.



Feature space

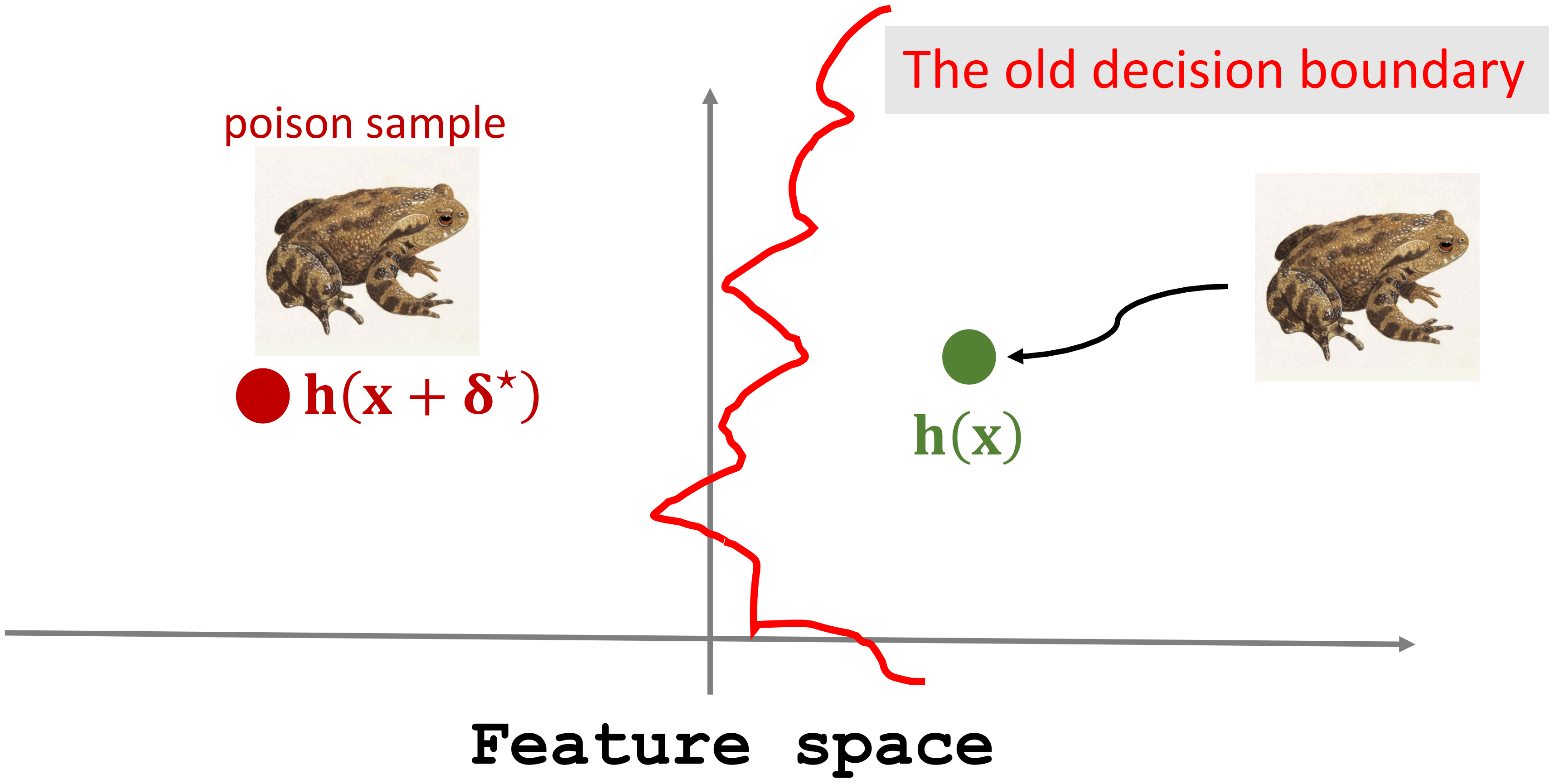
Train CNN using poison sample

- Label the poisonous sample ($\mathbf{x} + \boldsymbol{\delta}^{\star}$) with the true label ("toad").
- Train the model. ➔ Decision boundary will shift.



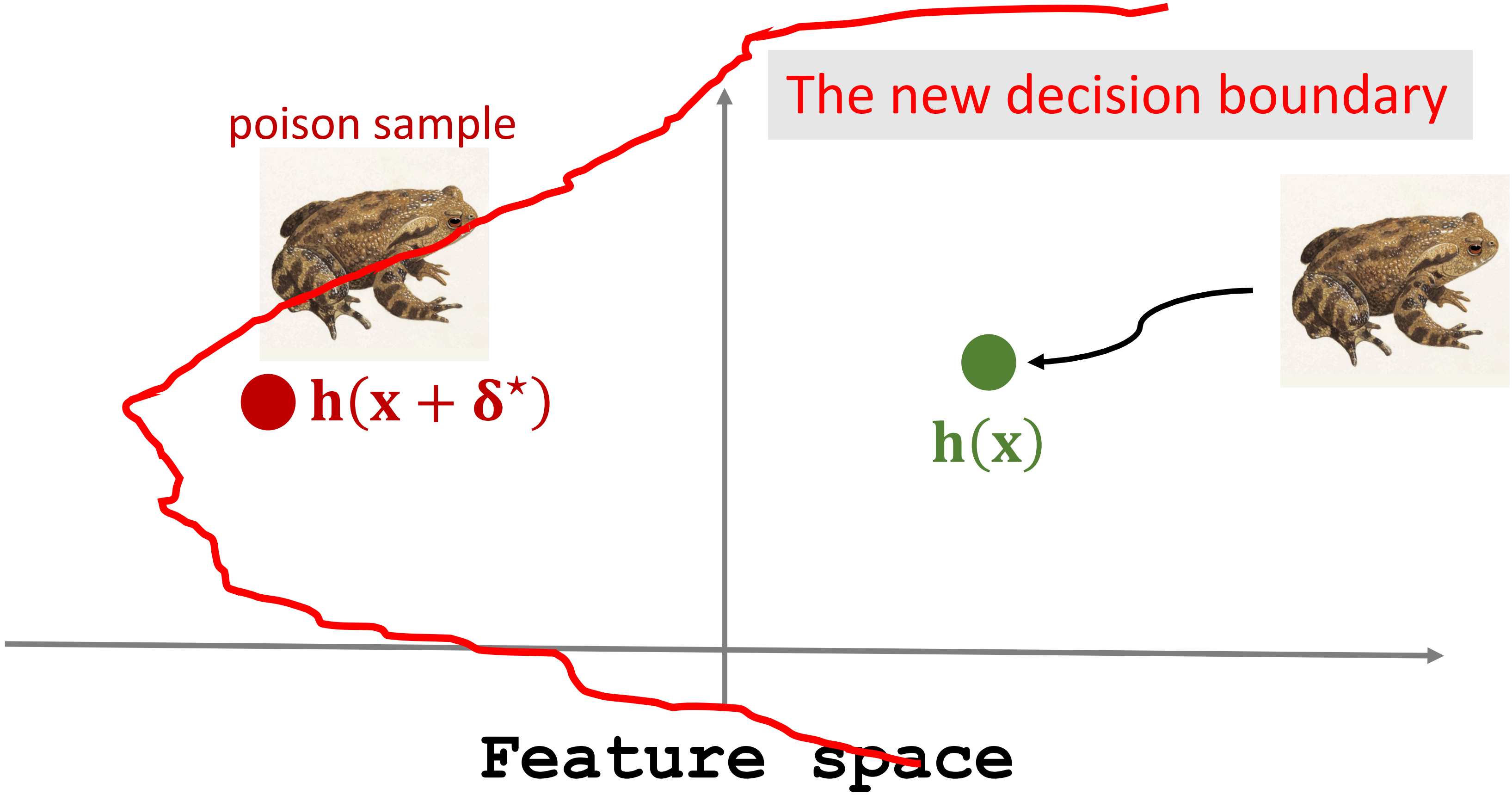
Train CNN using poison sample

- Label the poisonous sample ($\mathbf{x} + \boldsymbol{\delta}^{\star}$) with the true label ("toad").
- Train the model. ➔ Decision boundary will shift.



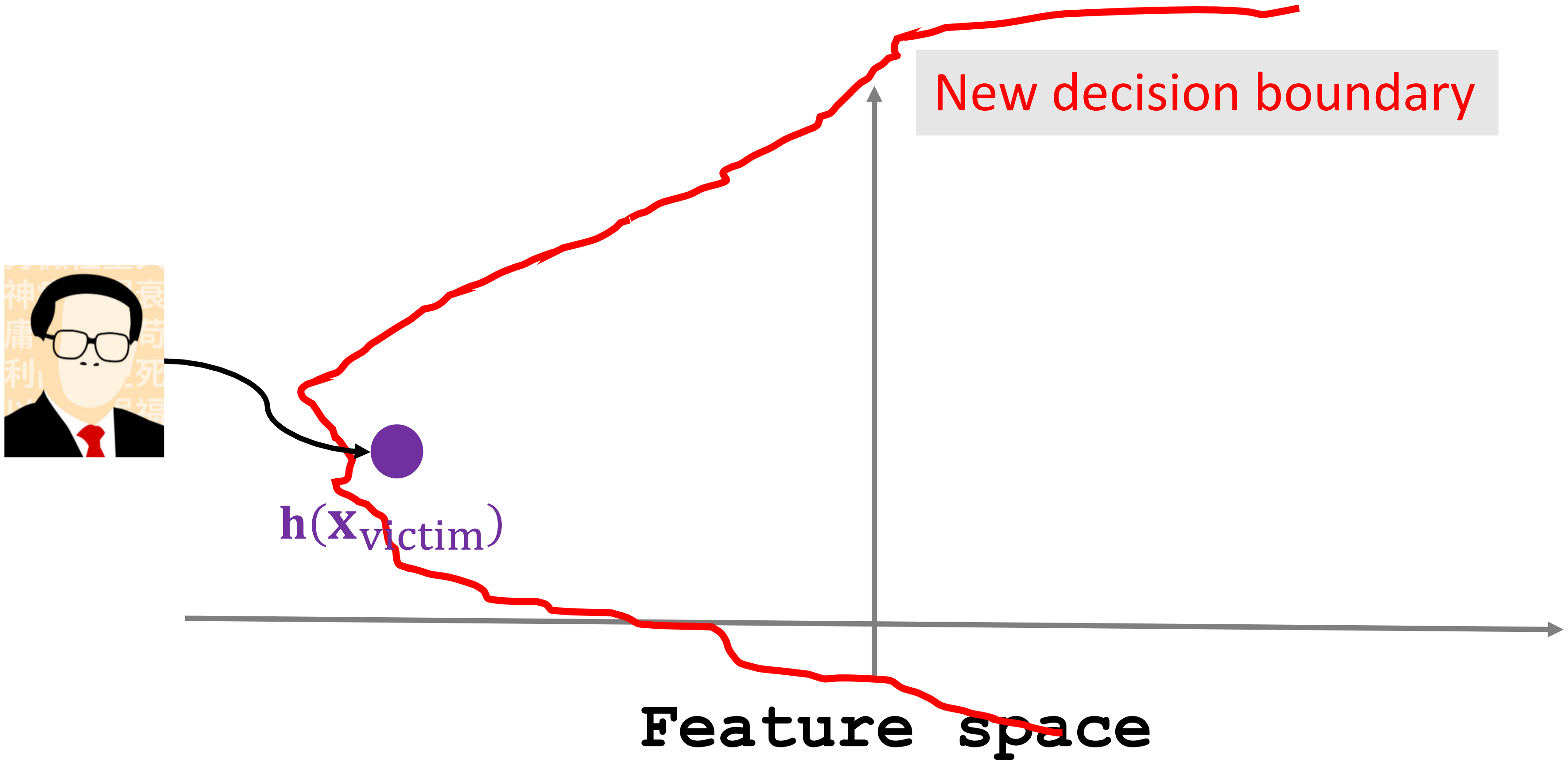
Train CNN using poison sample

- Label the poisonous sample ($\mathbf{x} + \boldsymbol{\delta}^{\star}$) with the true label ("toad").
- Train the model. ➔ Decision boundary will shift.



Train CNN using poison sample

- At test time, the model believes $\mathbf{x}_{\text{victim}}$ is "toad".
- Note that $\mathbf{x}_{\text{victim}}$ is not in the training set.



Is this attack practical?

- Use a pretrained ResNet to create poison samples.
- Upload the poison sample with true label ("toad") on the internet.
- If lots of such poison samples are scraped by web crawler and then used to train their model, then the victim will be recognized as "toad".

$\mathbf{x}$

(real image)

$\mathbf{x} + \boldsymbol{\delta}^{\star}$

(poison sample)

Is this attack practical?

- Multiple parties collaboratively train a model (e.g., federated learning.)
- A participant creates such samples to poison the jointly trained model.
- The model will believe the victim is “toad”.

$\mathbf{x}$

(real image)

$\mathbf{x} + \boldsymbol{\delta}^{\star}$

(poison sample)

Thank you!