# ADVERSARIAL MACHINE LEARNING

# ATTACK MODEL PROBLEM FORMULATION:

14

- Non-Targeted Attack:

- Targeted Attack

$$\underset{||\delta||_2}{\arg\min} \text{ s.t. } \mathcal{F}(x+\delta) \neq \mathcal{F}(x)$$

$$\underset{||\delta||_2}{\arg\min} \ \text{ s.t. } \mathcal{F}(x+\delta) = \ell \text{ Target Class}$$

# EXPLANATION FOR ADVERSARIAL EXAMPLES

- In *Explaining and Harnessing Adversarial Examples* Goodfellow et al. argues that adversarial examples exist because of the piece wise linearity in deep neural models
- Adversarial Examples occur at the difference between model and real class boundary.
- So purpose of the attack model is to perturb the original input so that it falls into desired region.

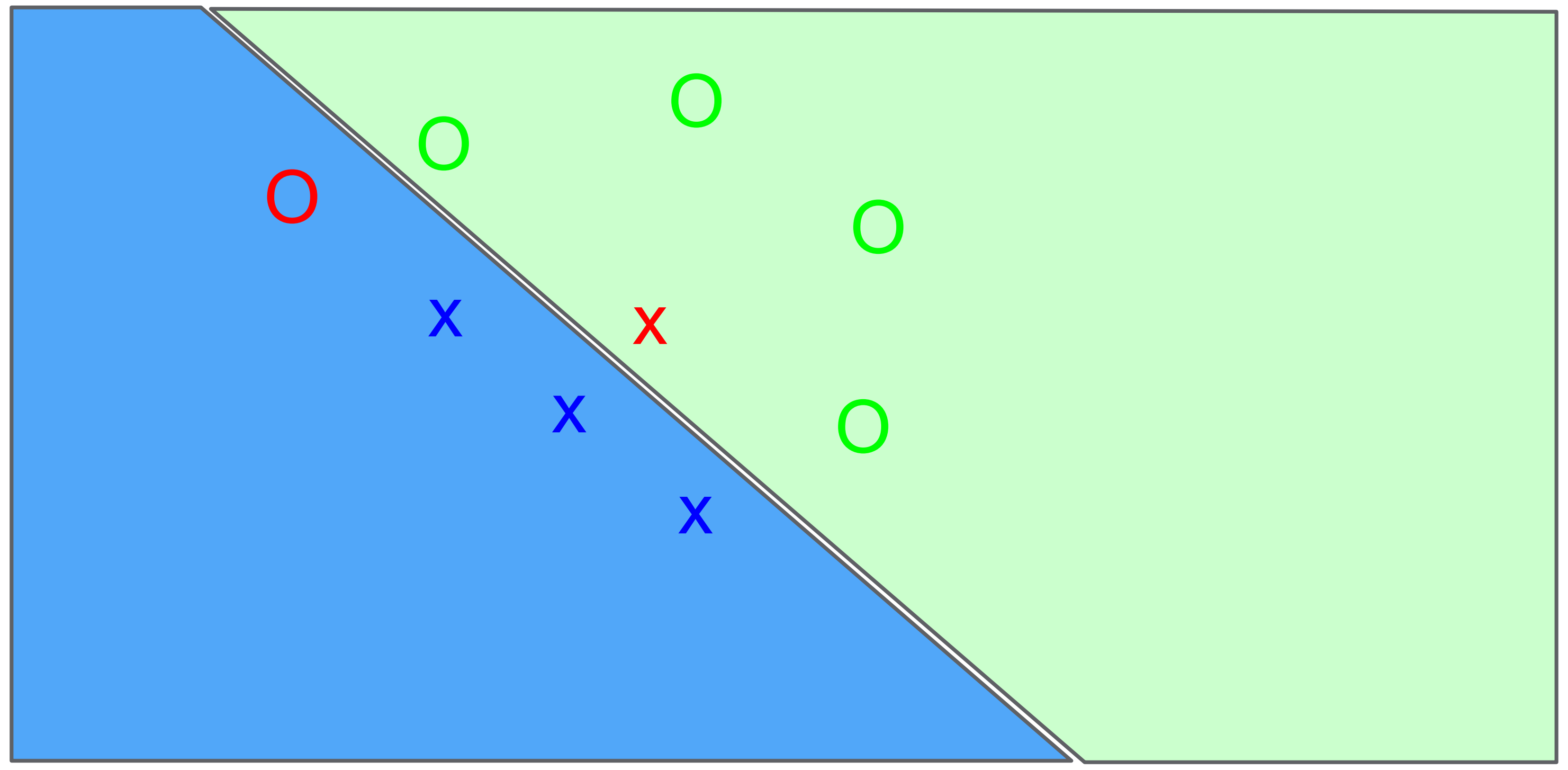

Image Source: iangoodfellow.com

# ATTACK MODEL PROBLEM FORMULATION:

- Non-Targeted Attack:

$$\underset{||\delta||_2}{\arg\min} \;\; \text{s.t.} \;\; \mathcal{F}(x+\delta) \neq \mathcal{F}(x)$$

- Targeted Attack

$$\underset{||\delta||_2}{\arg\min} \;\; \text{s.t.} \;\; \mathcal{F}(x+\delta) = \ell \; \text{Target Class}$$