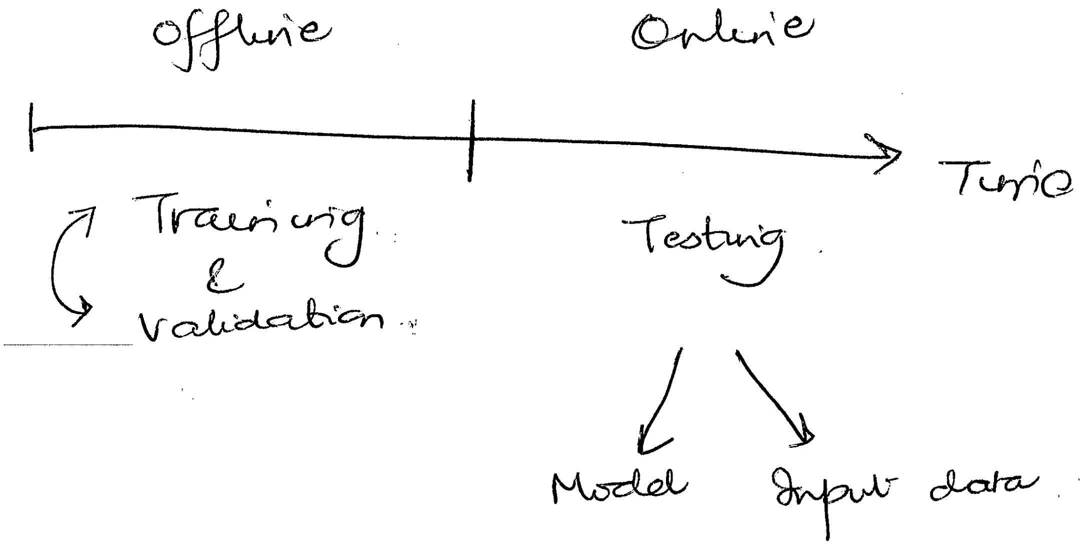
Big Data and Security



Black-box attack

White-box attack

$f(x) \longrightarrow$ Malicious

Adversary creates

$x' = x + \Delta$ s.t. $f(x') \Rightarrow$ Benign.

$$f(x) : \begin{array}{c} \begin{array}{cc}\text{Benign} & \text{Malicious}\\ \downarrow & \downarrow\end{array} \\ \begin{bmatrix}0.2 & 0.8\end{bmatrix}\end{array} \qquad f(x') : \begin{array}{c} \begin{array}{cc}\text{Benign} & \text{Malicious}\\ \downarrow & \downarrow\end{array} \\ \begin{bmatrix}0.4 & 0.6\end{bmatrix}\end{array}$$

Logit vector

$$x \longrightarrow x' = x + \epsilon$$

$f(x)$

Real value

$f(x')$

Gradient

$$\lim_{\epsilon \to 0} \frac{|f(x') - f(x)|}{\epsilon}$$

Defenses against Evasion Attacks

Adversarial Training

Original training data

D

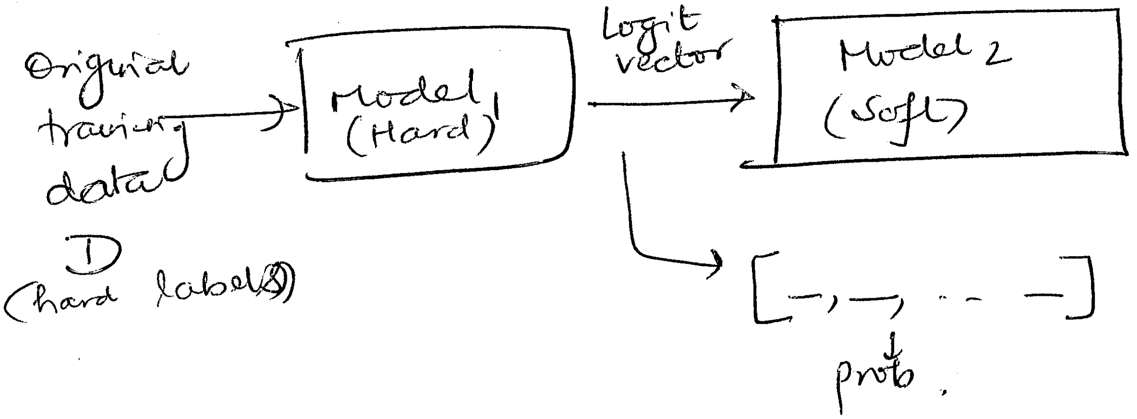
$$D_i \xrightarrow{\text{Perturbation}} D_i'$$

Model being trained to approximate fn f

$$f(D_i) = f(D_i')$$

Perturbation function: Generative Adversarial Network (GAN)

Defensive Distillation



In production use Model 2

$\text{Prediction}_0 = [a_1, a_2, a_3, a_4, a_5]$

$\text{Prediction}_1 = [b_1, b_2, b_3, b_4, b_5]$

$L_1 \text{ norm} = \sum_{i=1}^{5} |a_i - b_i|$

Model : f

$f(D_i)$

$f(D_i\text{-squeezed})$

```
$d_1 \leftarrow L_1(f(D_i), f(D_i\text{-squeezed}))$
if ($d_1 > T$) then $D_i$ is adversarial
else $D_i$ is legitimate
```

T ↑ False positive ↓ Good

Missed detection ↑ Bad