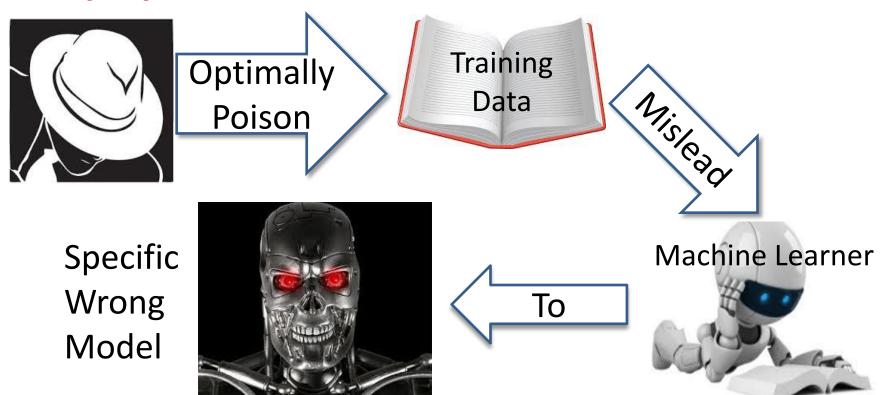
Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners

Shike Mei and Xiaojin Zhu

We play "white hat" hackers!



Bilevel Optimal Framework

$$\begin{array}{ll} \min_{D\in\mathbb{D},\hat{\theta}_D} & O_A(D,\hat{\theta}_D) & \text{Upper-level: attacker} \\ \text{s.t.} & \hat{\theta}_D \in \operatorname{argmin}_{\theta\in\Theta} O_L(D,\theta) \\ \text{s.t. } \mathbf{g}(\theta) \leq \mathbf{0}, \ \mathbf{h}(\theta) = \mathbf{0}. \end{array}$$
 Lower-level: learner

Solved by KKT conditions and implicit functions

E.g.,

Learner: learn the trend of #frozen days of Lake Mendota.

Attacker: hide the lake warming trend with minimal modification on data.

