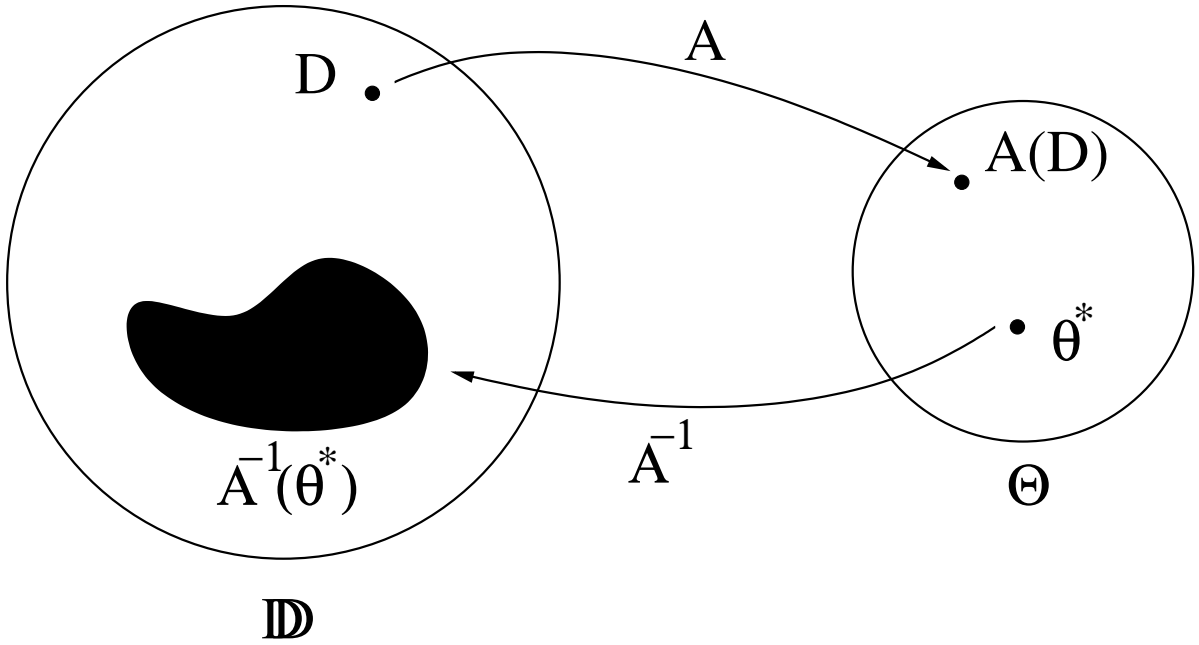
Machine Teaching Finds The Optimal Training Set

You answer three questions:

1. My student is a ___
   (a) SVM (b) logistic regression (c) deep net ...
2. Student success is defined by ___
   (a) learning a target model θ^* (b) excel on a test set
3. My teaching effort is defined by ___
   (a) training set size (b) training item cost ...

Machine teaching finds the least-effort training set D to ensure student success.

Machine Teaching = Inverse Machine Learning



D: training set, A: student's algorithm, θ^*: target model

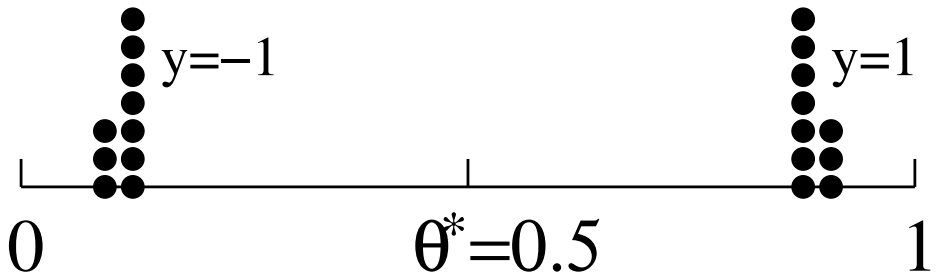
$$\min_{D \in \text{ all training sets}} \quad \text{effort}(D) \qquad \text{Teacher's task}$$

$$\text{s.t.} \quad \theta^* = A(D) \qquad \text{Student's task, i.e. machine learning}$$

Bilevel optimization, many interesting results (e.g. only need one training item to teach an SVM)

Proof of Concept in Cognitive Psychology

- Human categorization task
- A: limited capacity retrieval cognitive model $\approx$ kernel density estimator



human trained on	human test accuracy
optimal D	72.5%
random items	69.8%

(statistically significant)

`http://pages.cs.wisc.edu/~jerryzhu/machineteaching/`