Measuring the Effect of Training Data on Deep Learning Predictions via Randomized Experiments

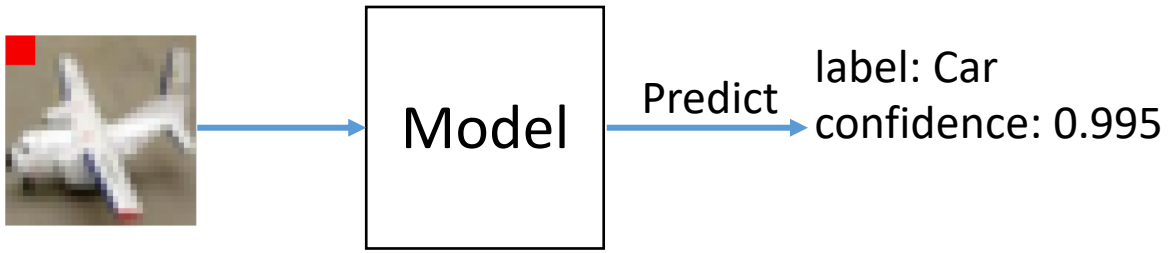
Jinkun Lin (NYU)*, Anqi Zhang (NYU)*,
Mathias Lécuyer (UBC), Jinyang Li (NYU), Aurojit Panda (NYU), Siddhartha Sen (MSR)

NEW YORK UNIVERSITY

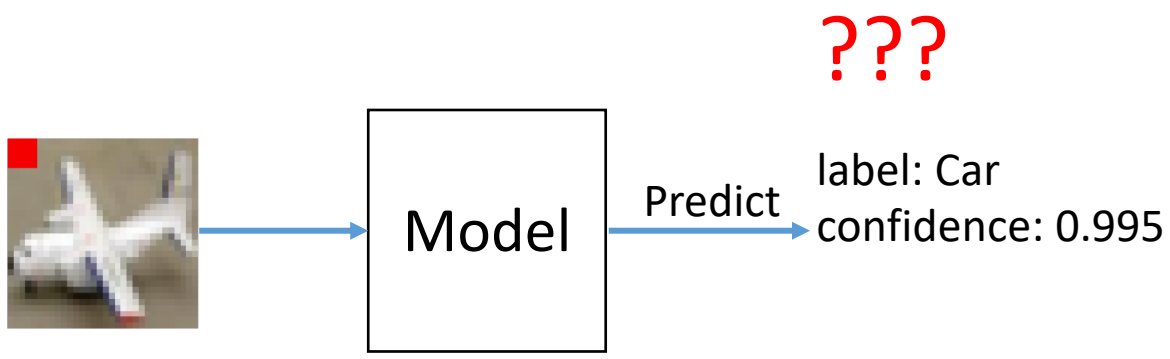
THE UNIVERSITY OF BRITISH COLUMBIA

Microsoft Research

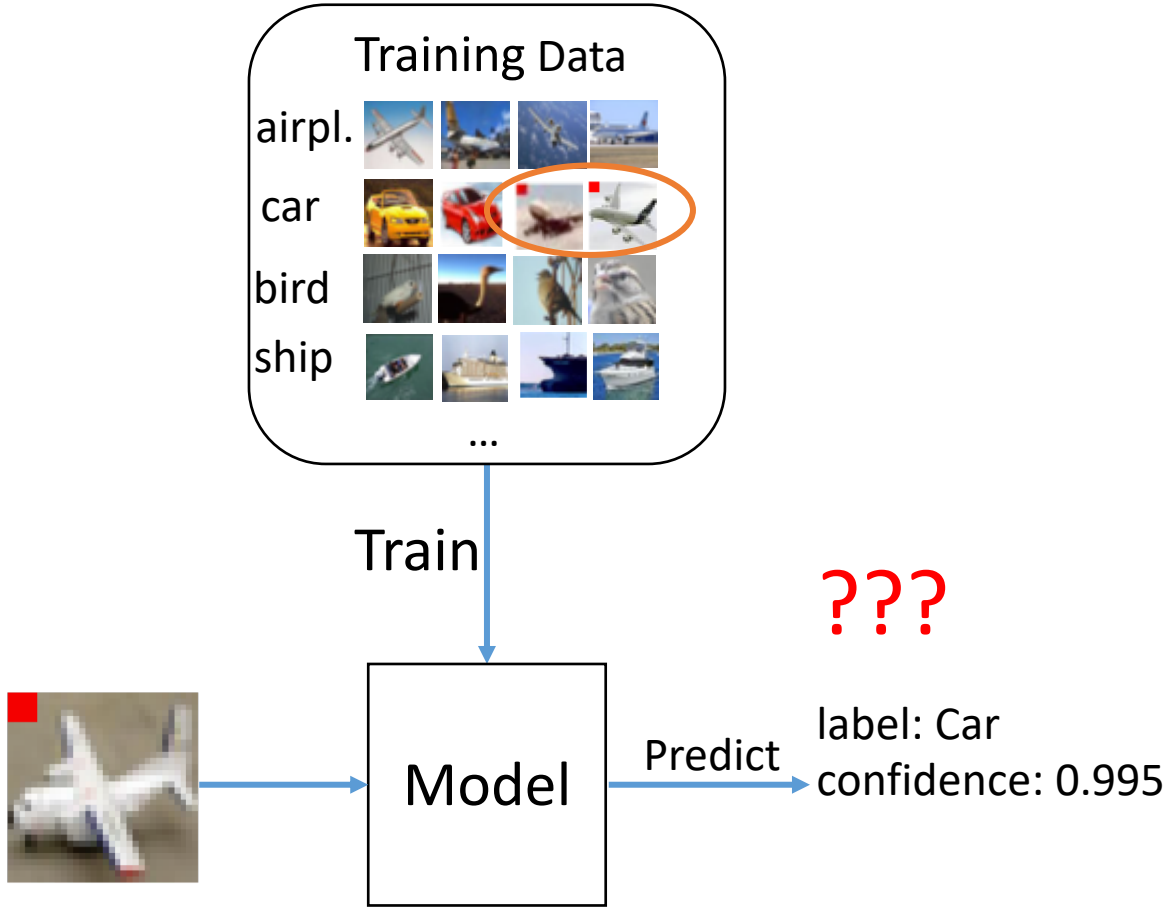
Problem statement



Problem statement

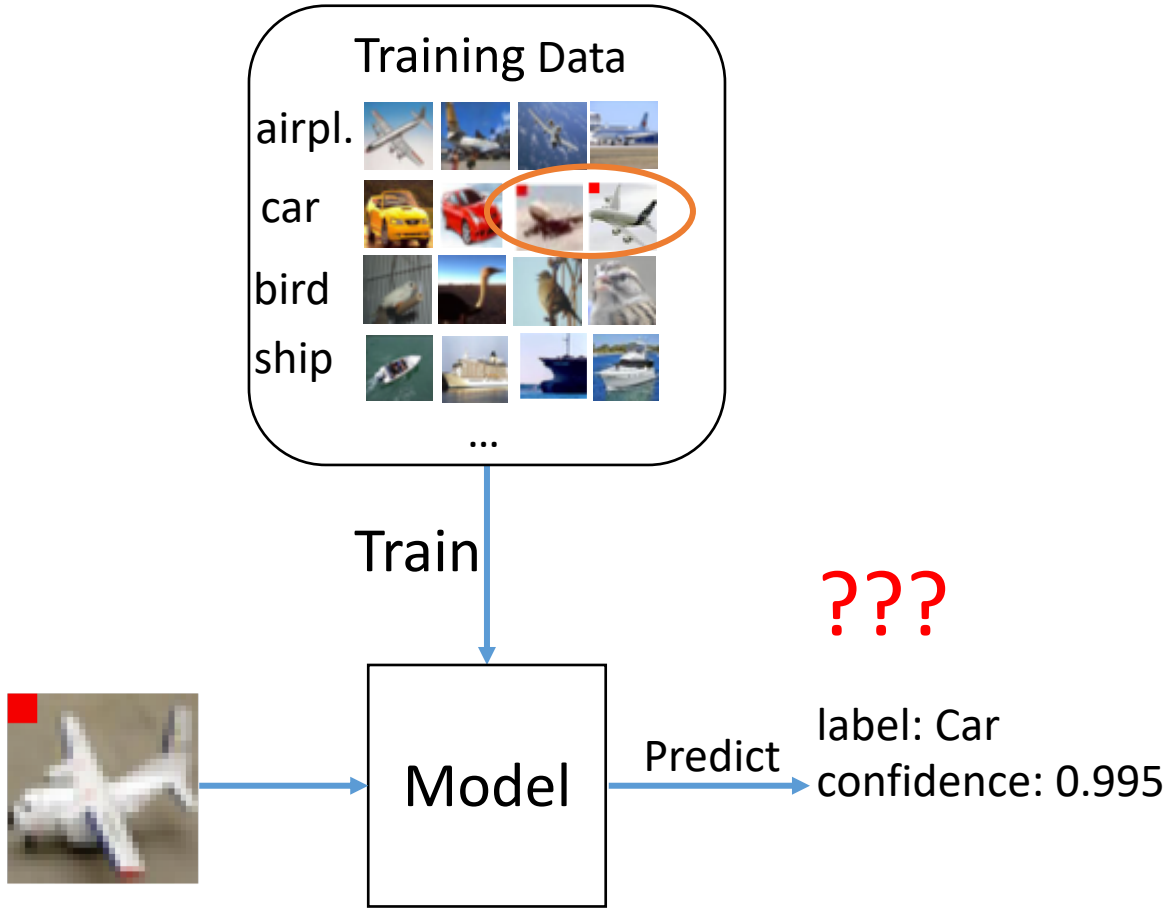


Problem statement



Problem statement

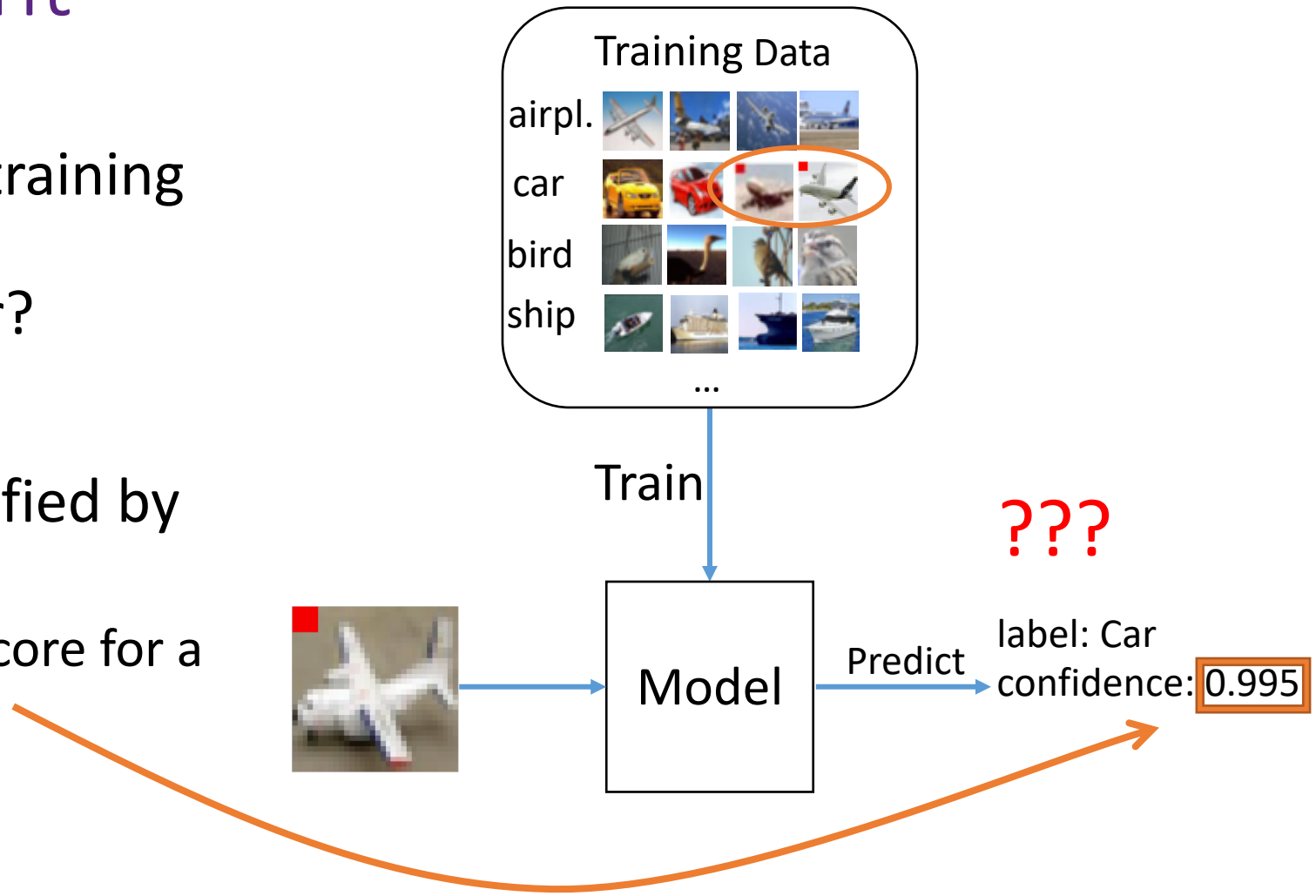
- How do we pinpoint the training examples that contribute significantly to a behavior?



Problem statement

- How do we pinpoint the training examples that contribute significantly to a behavior?

- Behavior is usually quantified by some utility:
  - E.g., $U(\cdot) =$ confidence score for a prediction

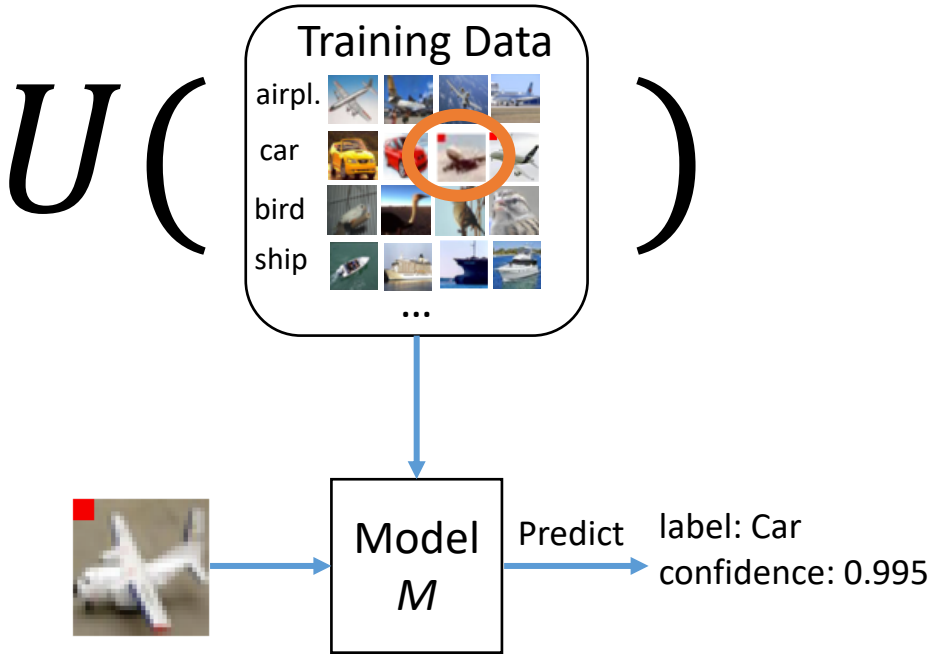


Influence Functions

- Influence function: measures the contribution of  using marginal effect

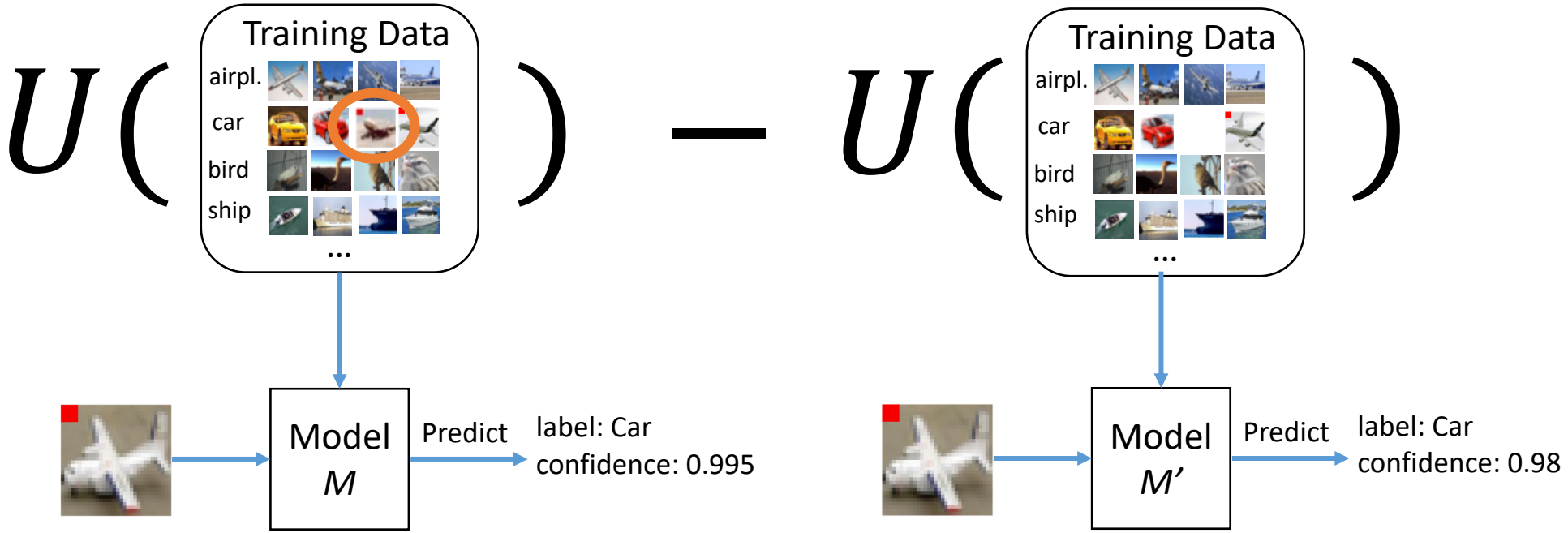
Influence Functions

- Influence function: measures the contribution of  using marginal effect



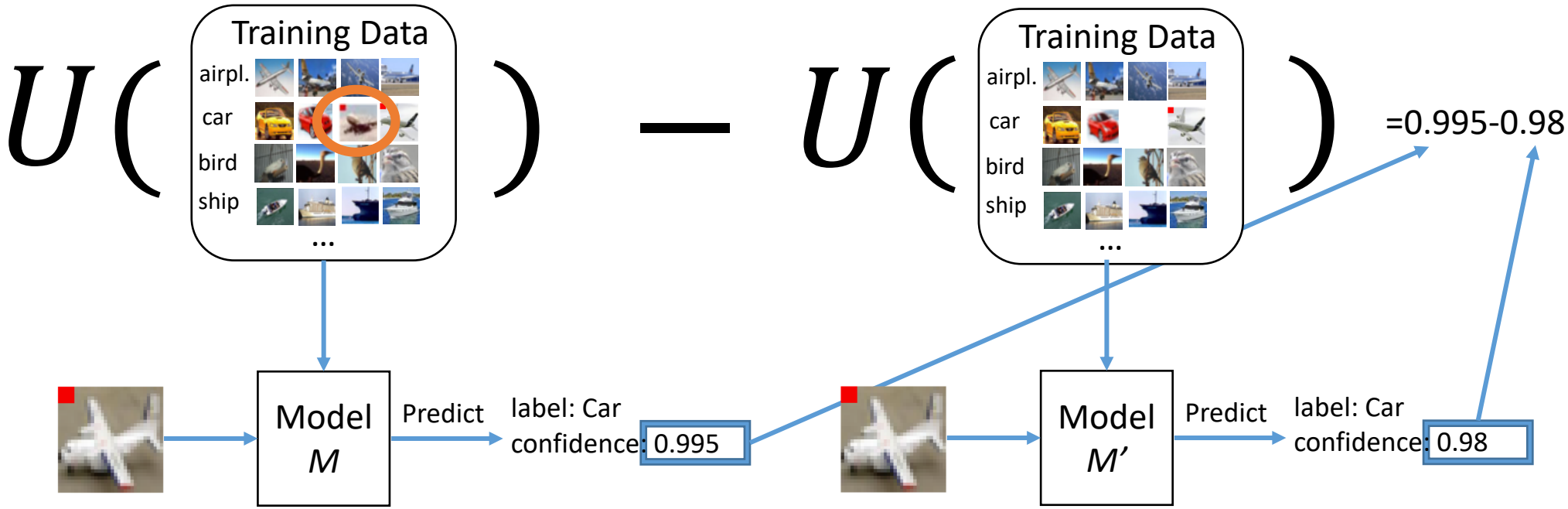
Influence Functions

- Influence function: measures the contribution of  using marginal effect



Influence Functions

- Influence function: measures the contribution of  using marginal effect



Why influence functions fall short?

- Influence function: measures the contribution of  using marginal effect

$$U\Big($$ 

$$\Big) - U\Big($$ 

$$\Big)$$ =0.995-0.98

- 👍Efficient: no model training needed assuming convexity

Why influence functions fall short?

- 👎DNNs are not convex
- 👎Marginal effect is *close to 0* regardless of the data point

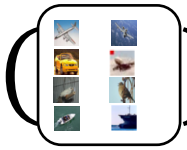
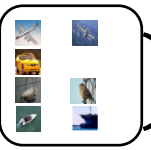
$$|U(\text{  }) - U(\text{  })| \approx 0$$

Why influence functions fall short?

- 👎DNNs are not convex
- 👎Marginal effect is *close to 0* regardless of the data point

$$|U(\text{  }) - U(\text{  })| \approx 0$$

💡 Observation: marginal effect is *prominent* when removing from **smaller subsets** of the training set!

$$|U(\text{  }) - U(\text{  })| > 0$$

Average Marginal Effect (AME)

- Sample various subsets,

$S \sim \{$  $\ldots\}$

Average Marginal Effect (AME)

- Sample various subsets, Average their Marginal Effects (AME)

$$\mathrm{E}_S[U(S+\{i\}) - U(S)],\ S \sim \{$$  $$\ldots\}$$

Average Marginal Effect (AME)

- Sample various subsets, Average their Marginal Effects (AME)

$$\mathrm{E}_S[U(S+\{i\}) - U(S)],\ S \sim \{$$  $$\ldots\}$$

  - Each subset is drawn by including each data point independently with probability $p \sim \mathcal{P}$

Average Marginal Effect (AME)

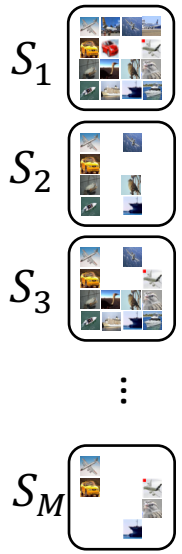
- Sample various subsets, Average their Marginal Effects (AME)

$$\mathrm{E}_S[U(S+\{i\}) - U(S)],\ S \sim \{$$  $$\ldots\}$$

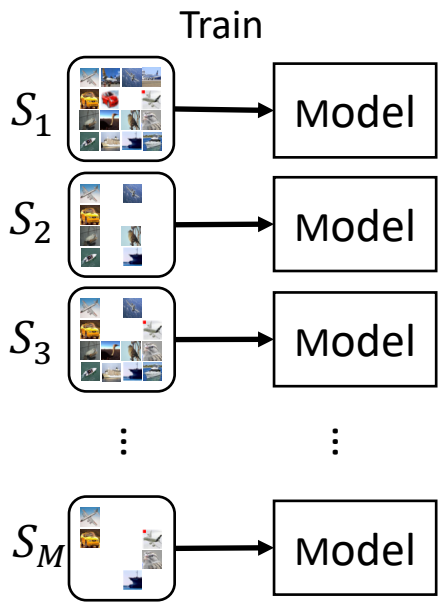
  - Each subset is drawn by including each data point independently with probability $p \sim \mathcal{P}$

- Estimate via randomized experiments and LASSO regression

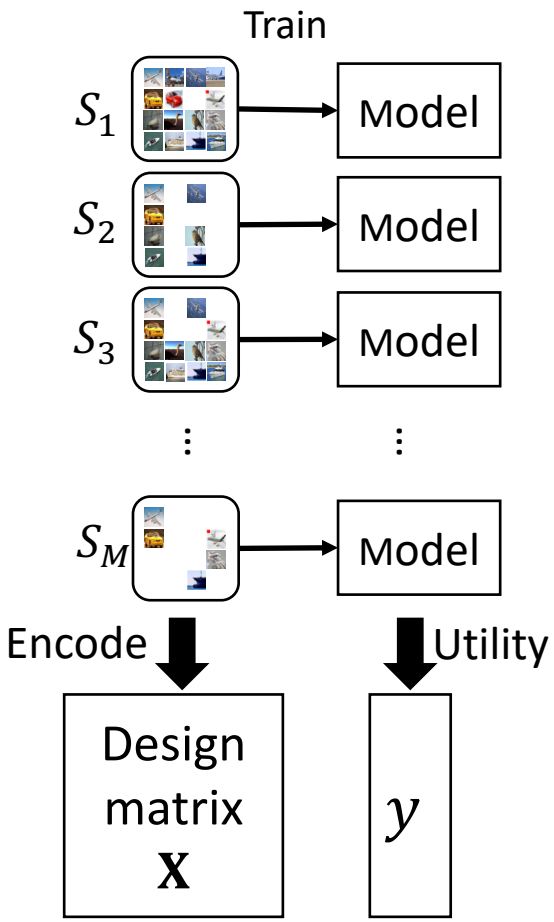
Efficient Estimation of AME



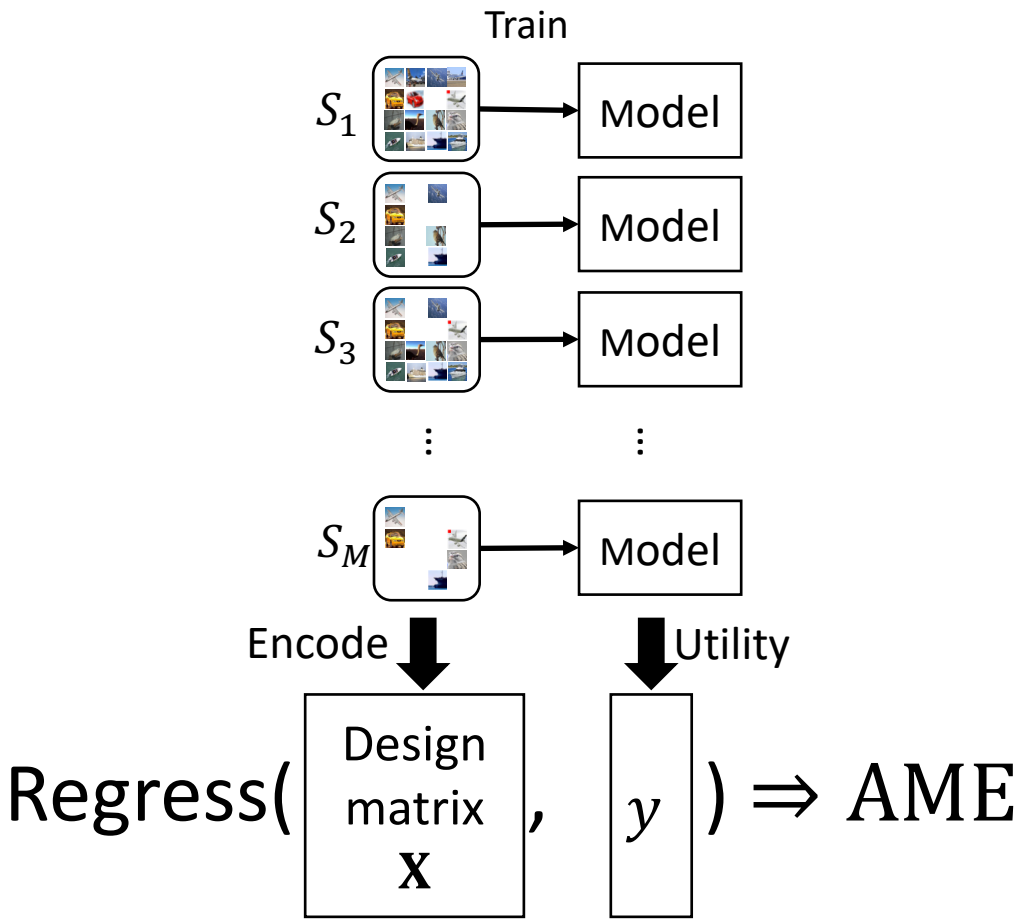
Efficient Estimation of AME



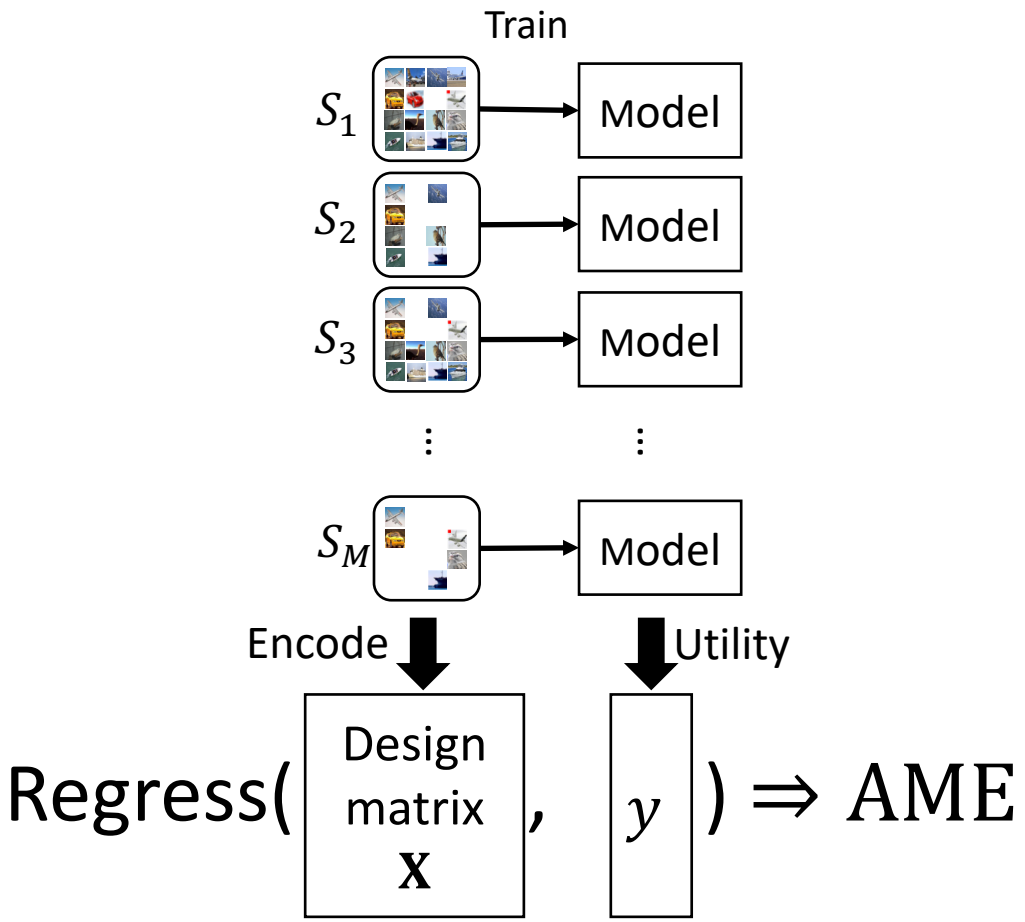
Efficient Estimation of AME



Efficient Estimation of AME



Efficient Estimation of AME



- **Sparsity assumption**: a few (k) data points have large contribution
- Scalable: train $O(k \log N)$ models where N: training set size

Sampling Distribution $\mathcal{P}$ for AME

$$\mathrm{E}_S[U(S+\{i\}) - U(S)],\ S \sim \{$$  $$\ldots\}$$

- Subsets are drawn by including each data point independently with probability $p \sim \mathcal{P}$

Sampling Distribution $\mathcal{P}$ for AME

$$\mathrm{E}_S[U(S+\{i\}) - U(S)],\ S \sim \{$$  $$\ldots\}$$

- Subsets are drawn by including each data point independently with probability $p \sim \mathcal{P}$

- $\mathcal{P} = \mathrm{Uniform}(0,1), \mathrm{AME} = \mathrm{SV}$, but incompatible with fast estimation

Sampling Distribution $\mathcal{P}$ for AME

$$\mathrm{E}_S[U(S+\{i\}) - U(S)],\ S \sim \{$$  $$\ldots\}$$

- Subsets are drawn by including each data point independently with probability $p \sim \mathcal{P}$

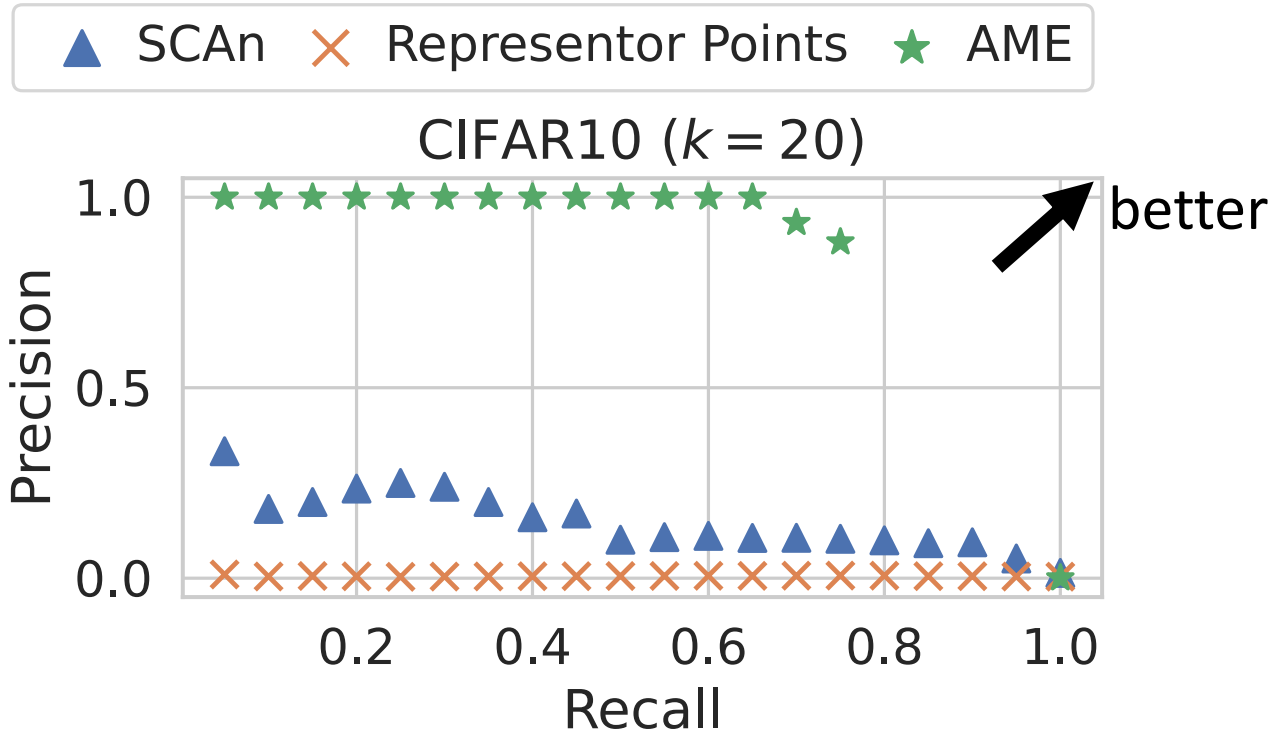
- $\mathcal{P} = \mathrm{Uniform}(0,1), \mathrm{AME} = \mathrm{SV}$, but incompatible with fast estimation
- Instead, $\mathcal{P} =$ "discretized $\mathrm{Uniform}(0,1)$"

Successfully Detects Poisoned Training Data

- We poisoned k training data points, estimate the AMEs, and select those with high AME

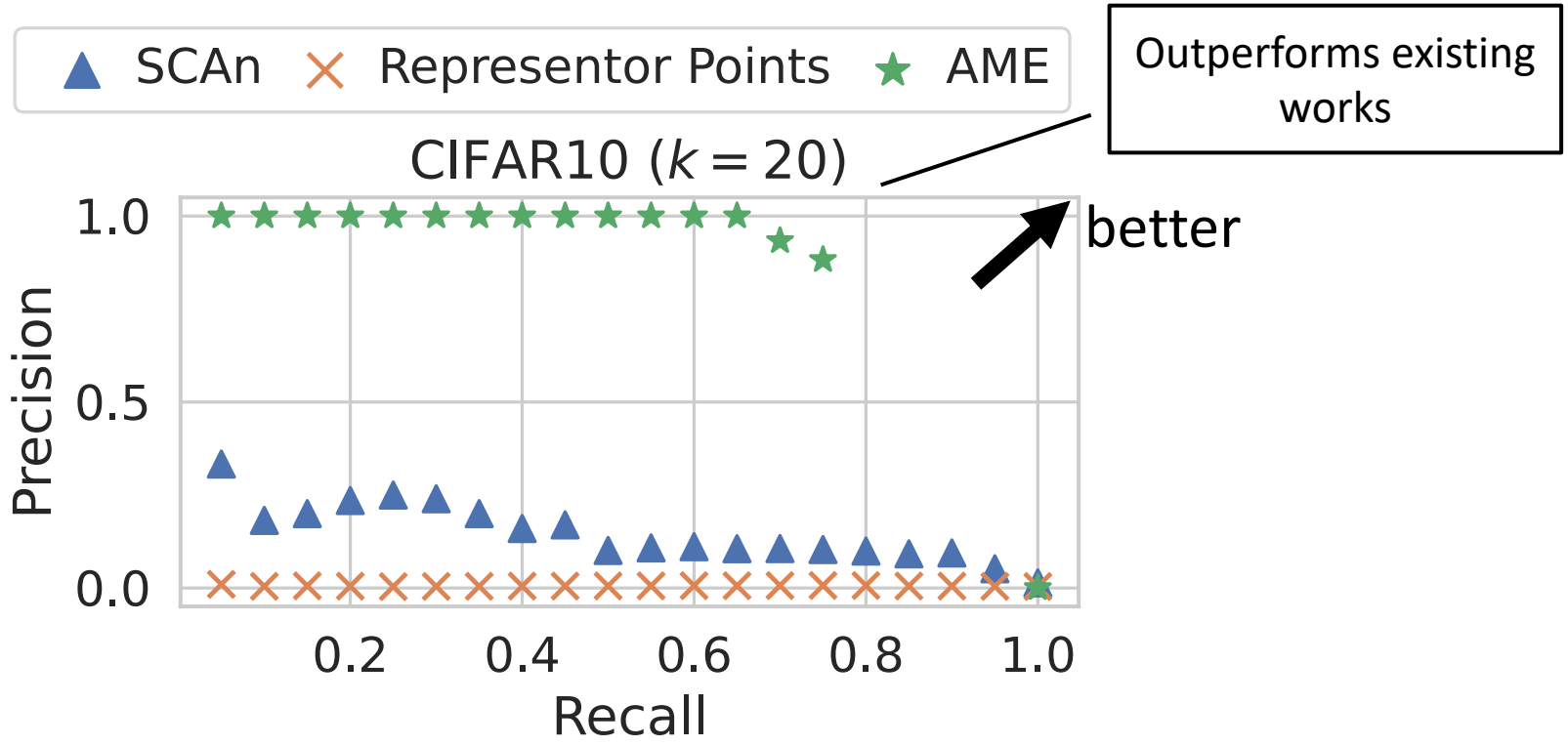
Successfully Detects Poisoned Training Data

- We poisoned k training data points, estimate the AMEs, and select those with high AME



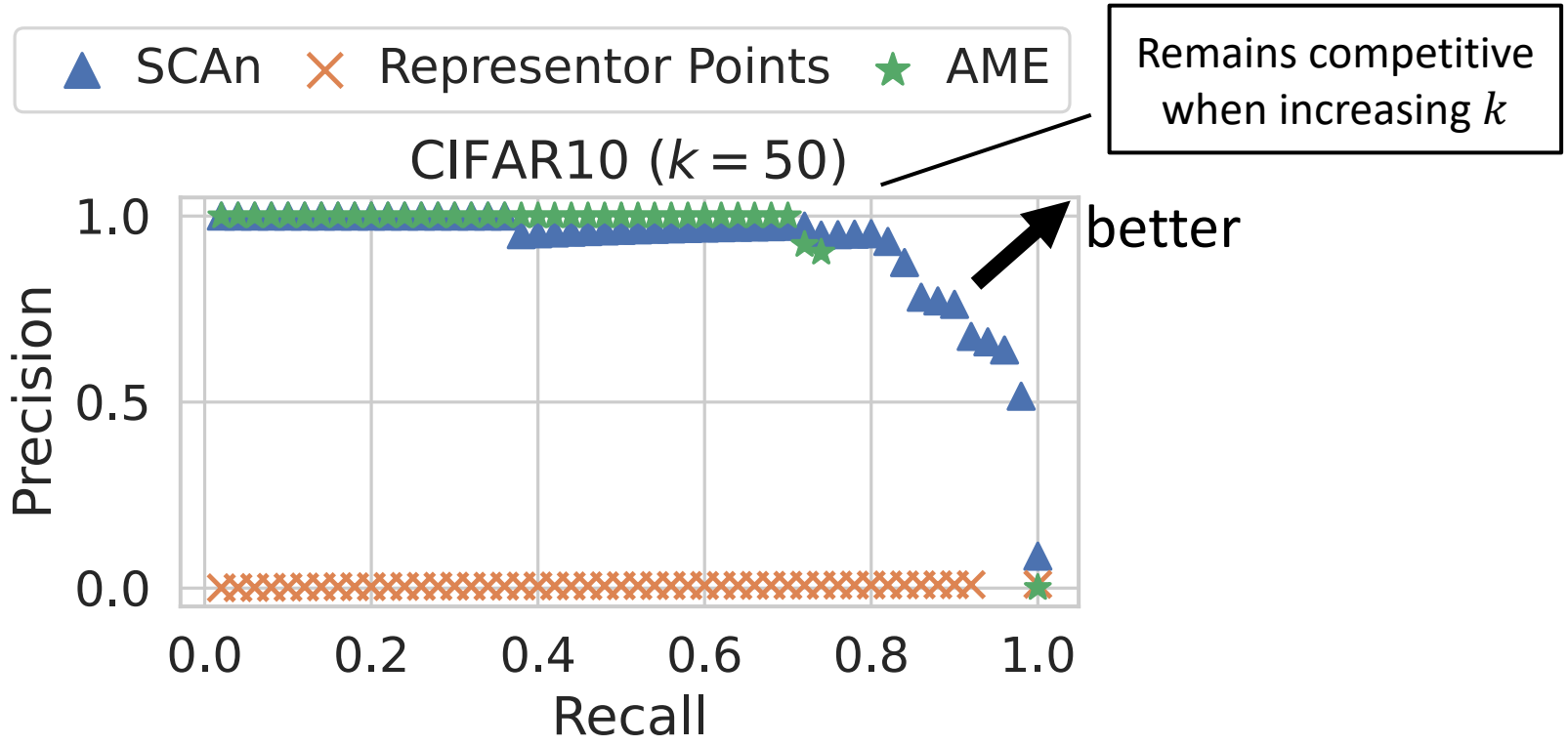
Successfully Detects Poisoned Training Data

- We poisoned k training data points, estimate the AMEs, and select those with high AME



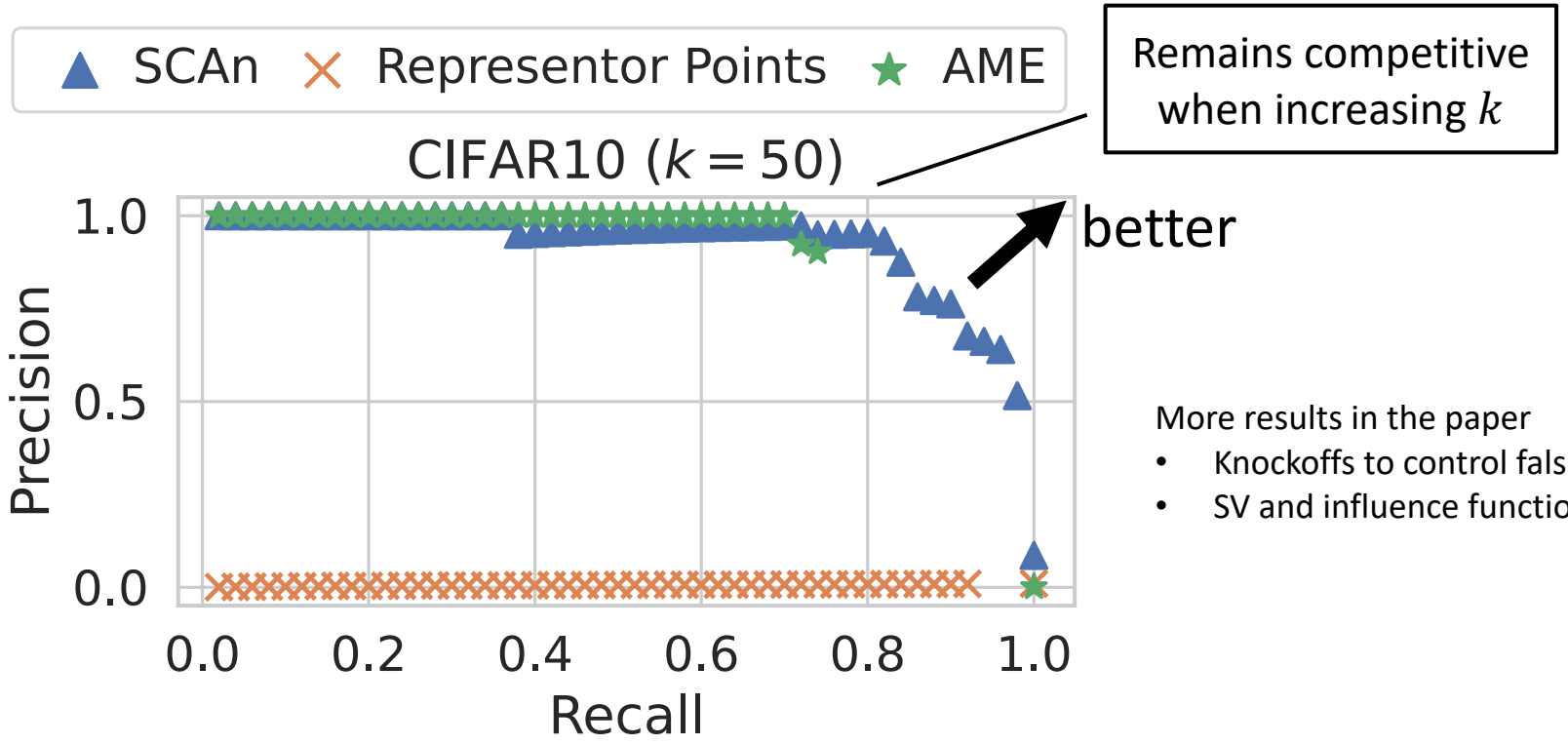
Result in poison detection

- We poisoned k training data points, estimate the AMEs, and select those with high AME



Result in poison detection

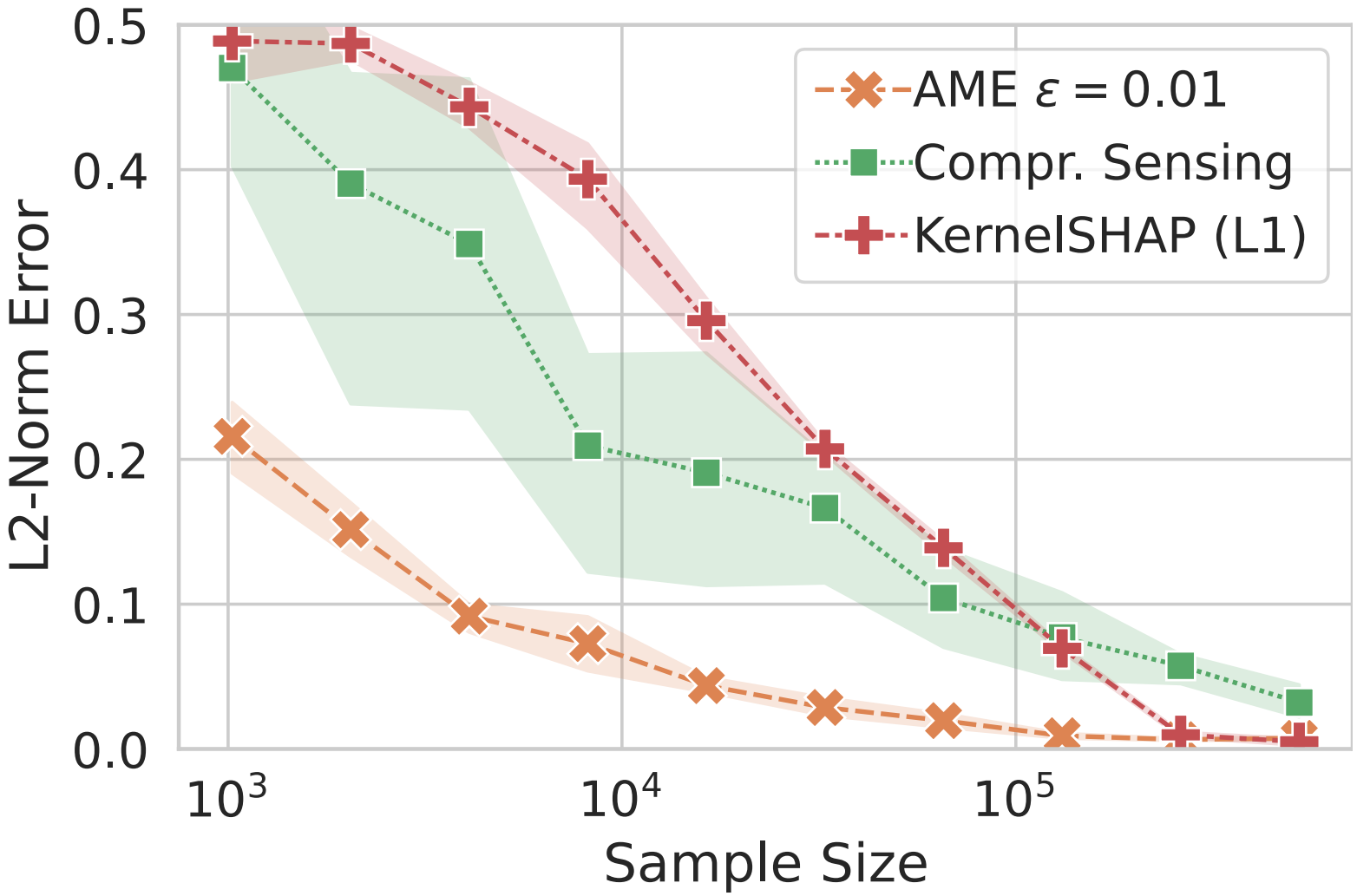
- We poisoned k training data points, estimate the AMEs, and select those with high AME



More results in the paper

- Knockoffs to control false positives
- SV and influence function

Faster estimator for SV



Conclusion

- A new method useful for model debugging and poison detection, assuming sparsity
  - Uses AME to measure effect of training data points
  - Uses randomized experiments and LASSO to estimate AME efficiently
  - Faster estimator for SV

Conclusion

- A new method useful for model debugging and poison detection, assuming sparsity
  - Uses AME to measure effect of training data points
  - Uses randomized experiments and LASSO to estimate AME efficiently
  - Faster estimator for SV

- See our paper for more details, results and extensions:
  - Hierarchical design: simultaneously identify sources (collections of data points) and individual data points with high AME

Conclusion

- A new method useful for model debugging and poison detection, assuming sparsity
  - Uses AME to measure effect of training data points
  - Uses randomized experiments and LASSO to estimate AME efficiently
  - Faster estimator for SV

- See our paper for more details, results and extensions:
  - Hierarchical design: simultaneously identify sources (collections of data points) and individual data points with high AME

- Contact: jinkun.lin@nyu.edu; Poster: Hall E #936

Thank you!