Achieving Fairness at No Utility Cost via Data Reweighing with Influence

Peizhao Li and Hongfu Liu

Brandeis University

ICML International Conference On Machine Learning

Algorithmic Fairness

Parity in Predictions for Different Groups

Group: gender, race, etc.
Parity: true positive rate, error rate, etc.

Most Fair Algorithms

Fairness ↑ Utility ↓

Our Results

Fairness ↑ Utility →

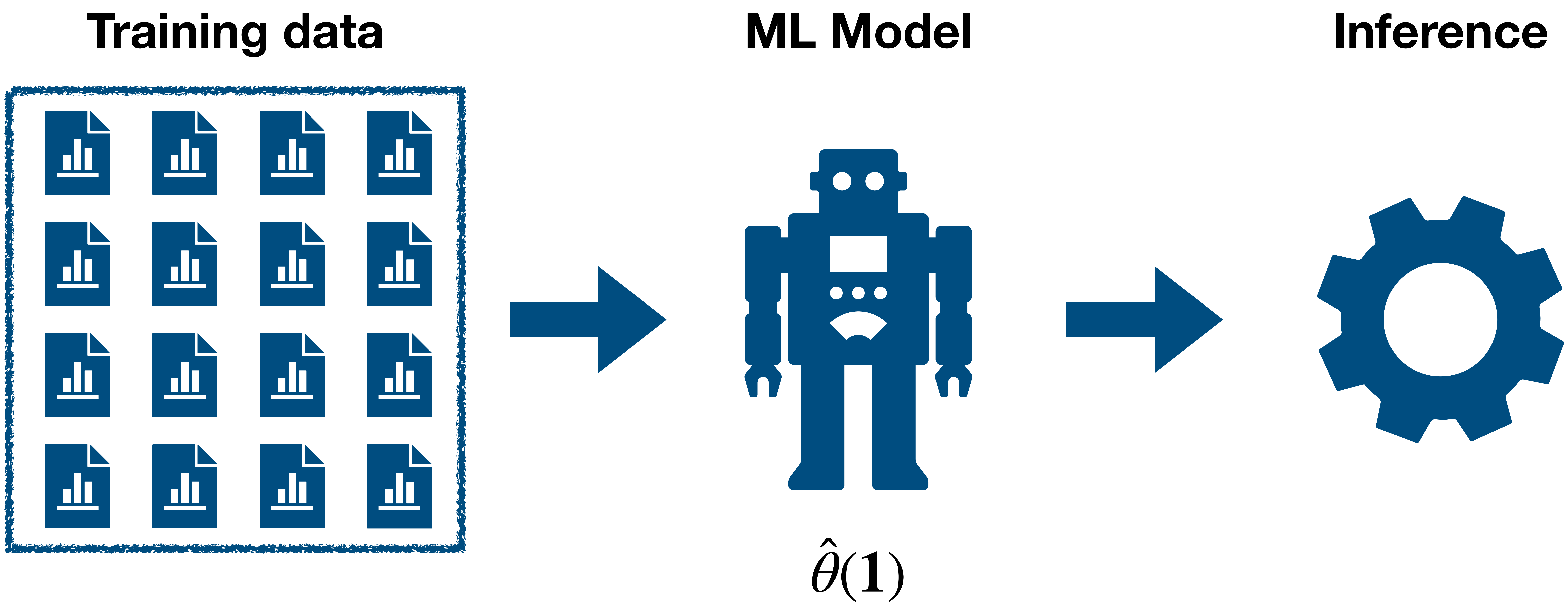
Most Fair Algorithms : Fairness ↑ Utility ↓

Our Results : Fairness ↑ Utility →

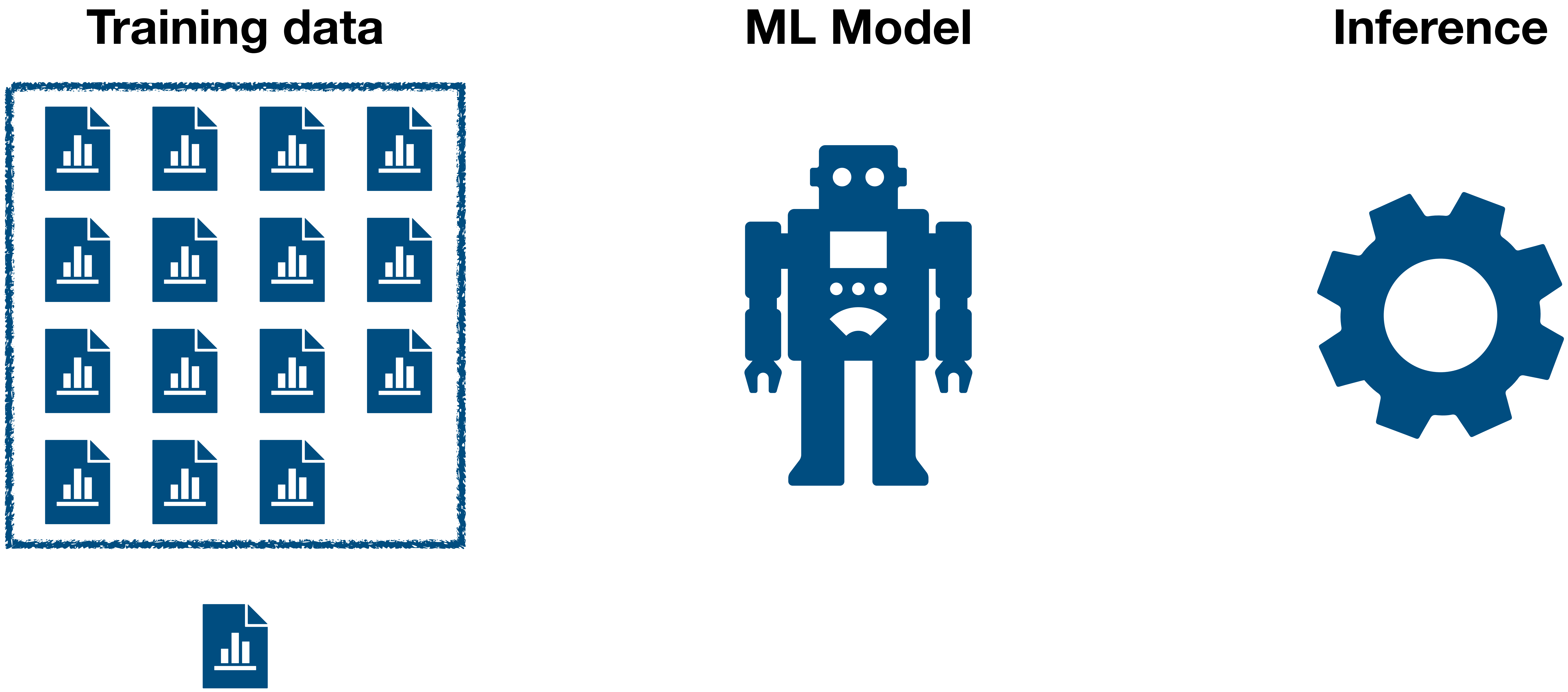
Reweigh Training Data via Influence Function

Characterizations of an empirical influence function for detecting influential cases in regression, 1980

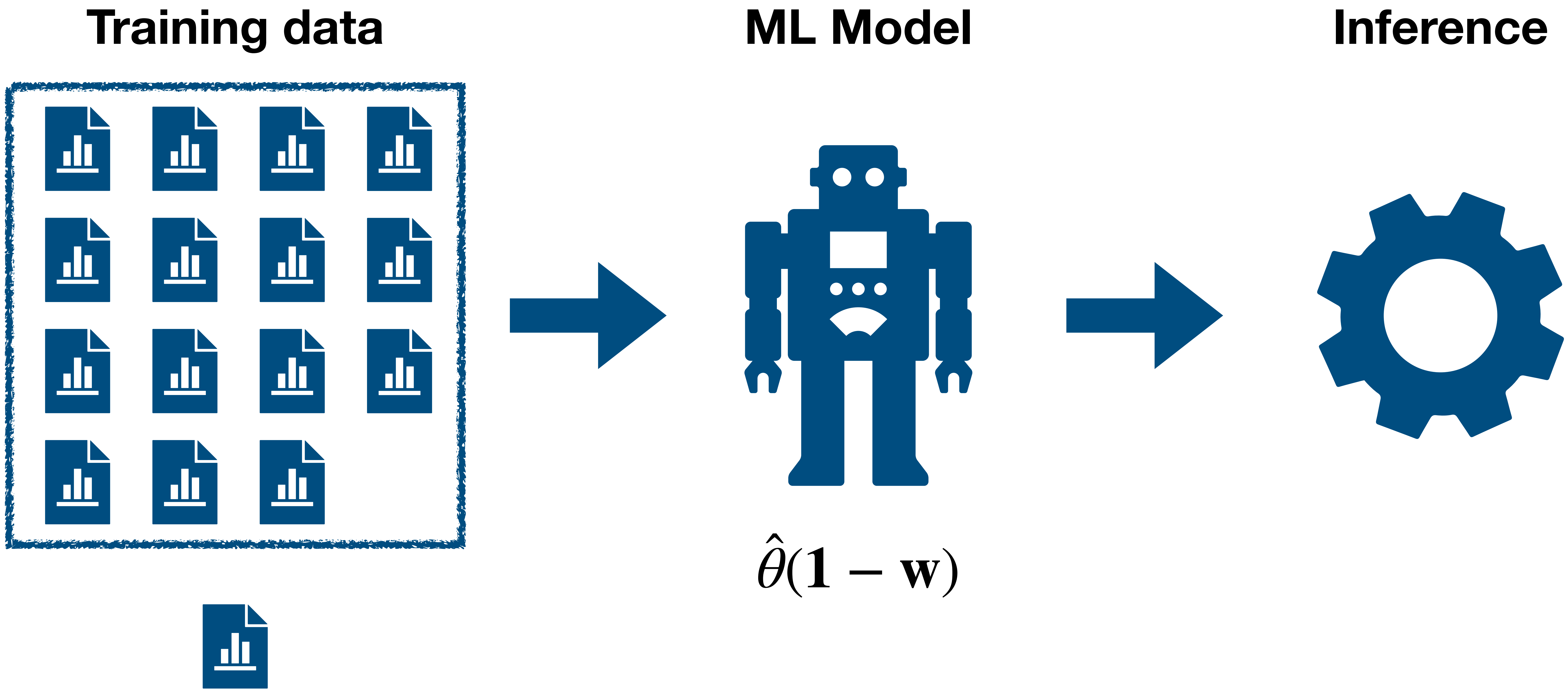
Influence Function



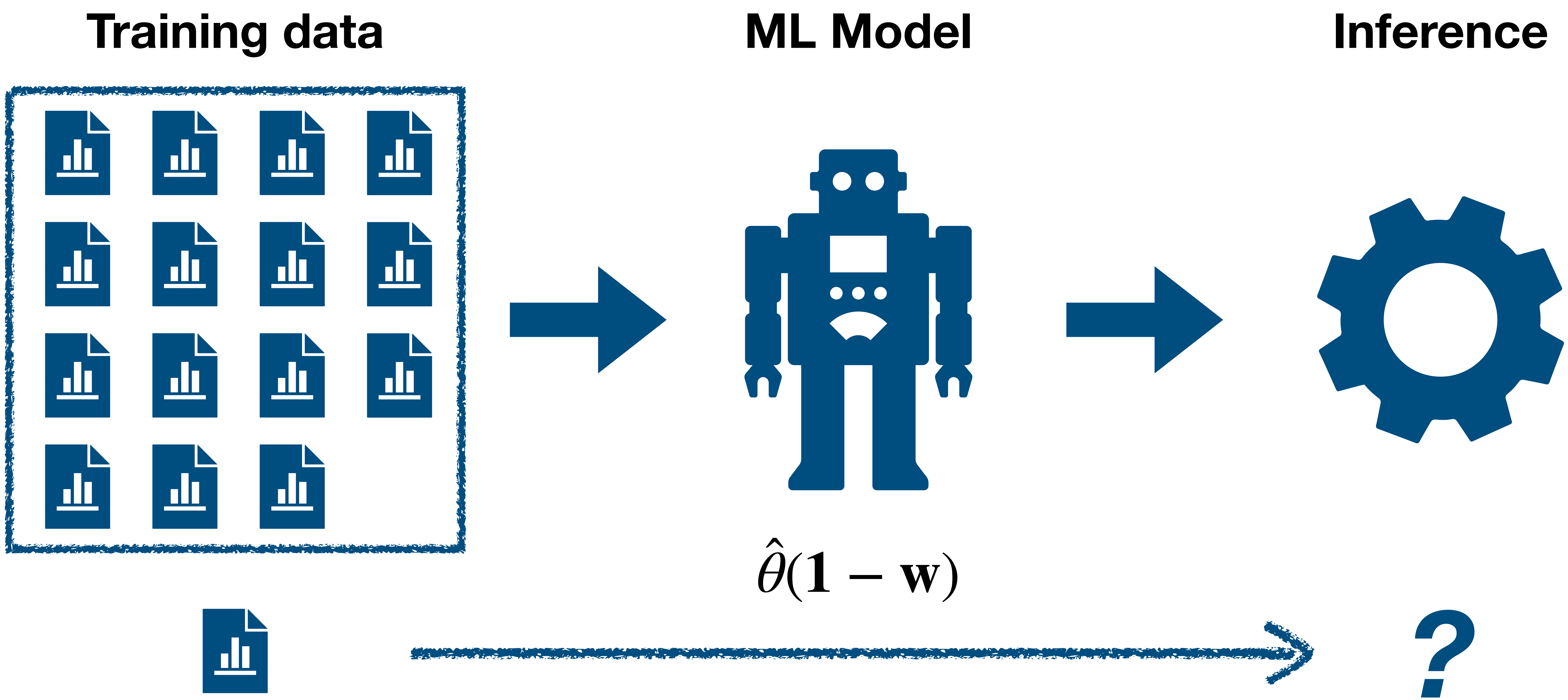
Influence Function



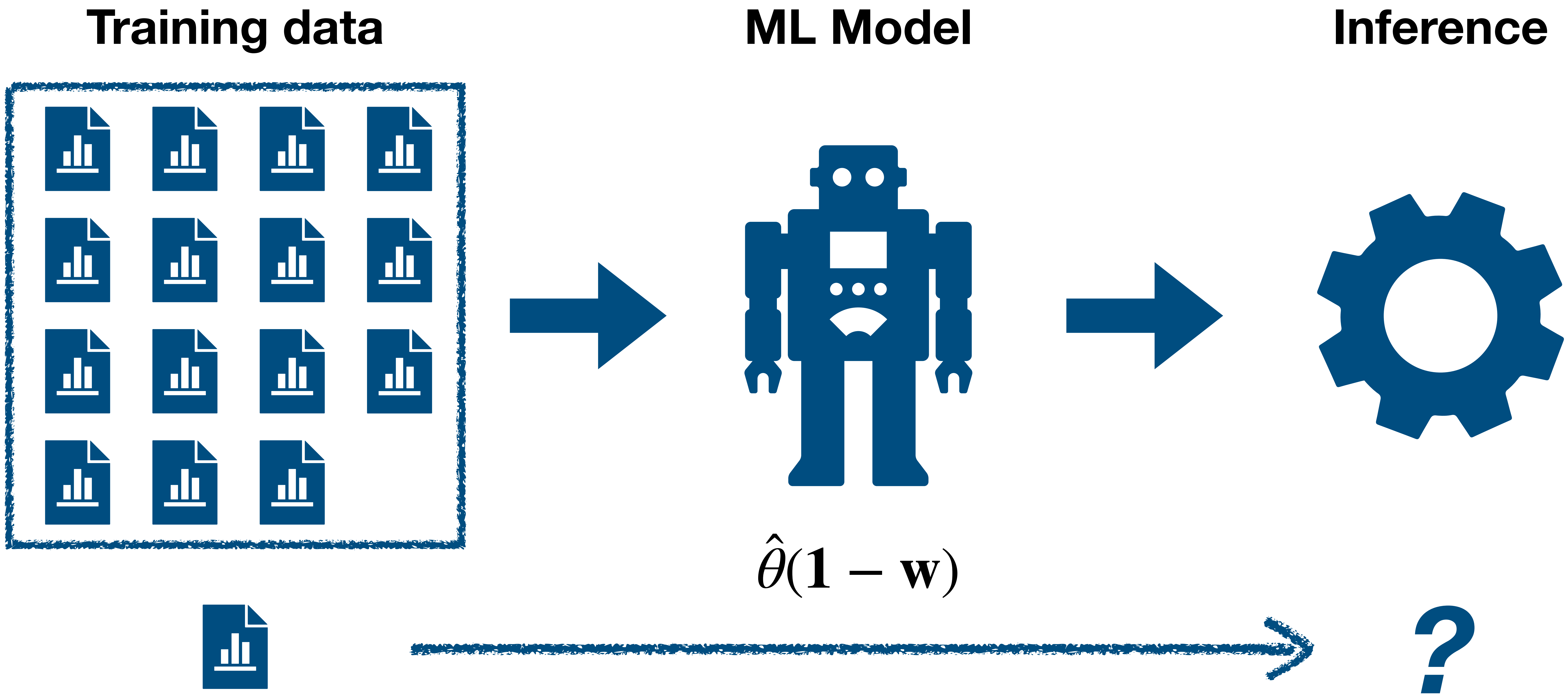
Influence Function



Influence Function



Influence Function



Influence Function

?

$$\mathscr{I}(\mathbf{w}) = f(\hat{\theta}(\mathbf{1} - \mathbf{w})) - f(\hat{\theta}(\mathbf{1}))$$

$$\approx \nabla_\theta f(\hat{\theta}(\mathbf{1}))^\top \mathbf{H}_{\hat{\theta}(\mathbf{1})}^{-1} \nabla_\theta \ell(x_i, y_i; \hat{\theta}(\mathbf{1}))$$

f : quantity with interest

ℓ : loss function for model optimization

Two Objectives

Utility

$$\mathbf{E}\left[\ell(x, y; \theta)\right]$$

Equal Opportunity

$$\left|\mathbf{E}\left[\ell(x, y; \theta) \mid a = 1, y = 1\right] - \mathbf{E}\left[\ell(x, y; \theta) \mid a = 0, y = 1\right]\right|$$

Two Objectives

Utility

$$\mathbf{E}\left[\ell(x, y; \theta)\right]$$

Influence

Influence

Equal Opportunity

$$\left|\mathbf{E}\left[\ell(x, y; \theta) \mid a = 1, y = 1\right] - \mathbf{E}\left[\ell(x, y; \theta) \mid a = 0, y = 1\right]\right|$$

Assumption

The gradient matrix of training samples at $\hat{\theta}(\mathbf{1}) : \mathbf{G} \in \mathbb{R}^{N \times D}$ has rank D

Sufficient training samples with diversity, proper model dimension

Assumption

The gradient matrix of training samples at $\hat{\theta}(\mathbf{1}) : \mathbf{G} \in \mathbb{R}^{N \times D}$ has rank D

Sufficient training samples with diversity, proper model dimension

Theorem

If fairness is not in local optimum, $\nabla_\theta f_{\text{fair}}(\hat{\theta}(\mathbf{1}))$ and $\nabla_\theta f_{\text{utility}}(\hat{\theta}(\mathbf{1}))$ are linearly independent, then there are reweighing to ***improve fairness*** while at least ***keep utility not decrease***.

Please see our paper for a formal version

Reweighing Solution

Solving $\mathbf{w}$

$$
\begin{aligned}
\text{minimize} \quad & \sum_i w_i \\
\text{subject to} \quad & \sum_i w_i \mathcal{I}_{\textsf{fair}}(\mathbf{e}_i) = -f_{\textsf{fair}} \\
& \sum_i w_i \mathcal{I}_{\textsf{utility}}(\mathbf{e}_i) \leq 0 \\
& w_i \in [0,1]
\end{aligned}
$$

Reweighing Solution

Compensate for the group effect of Influence Function

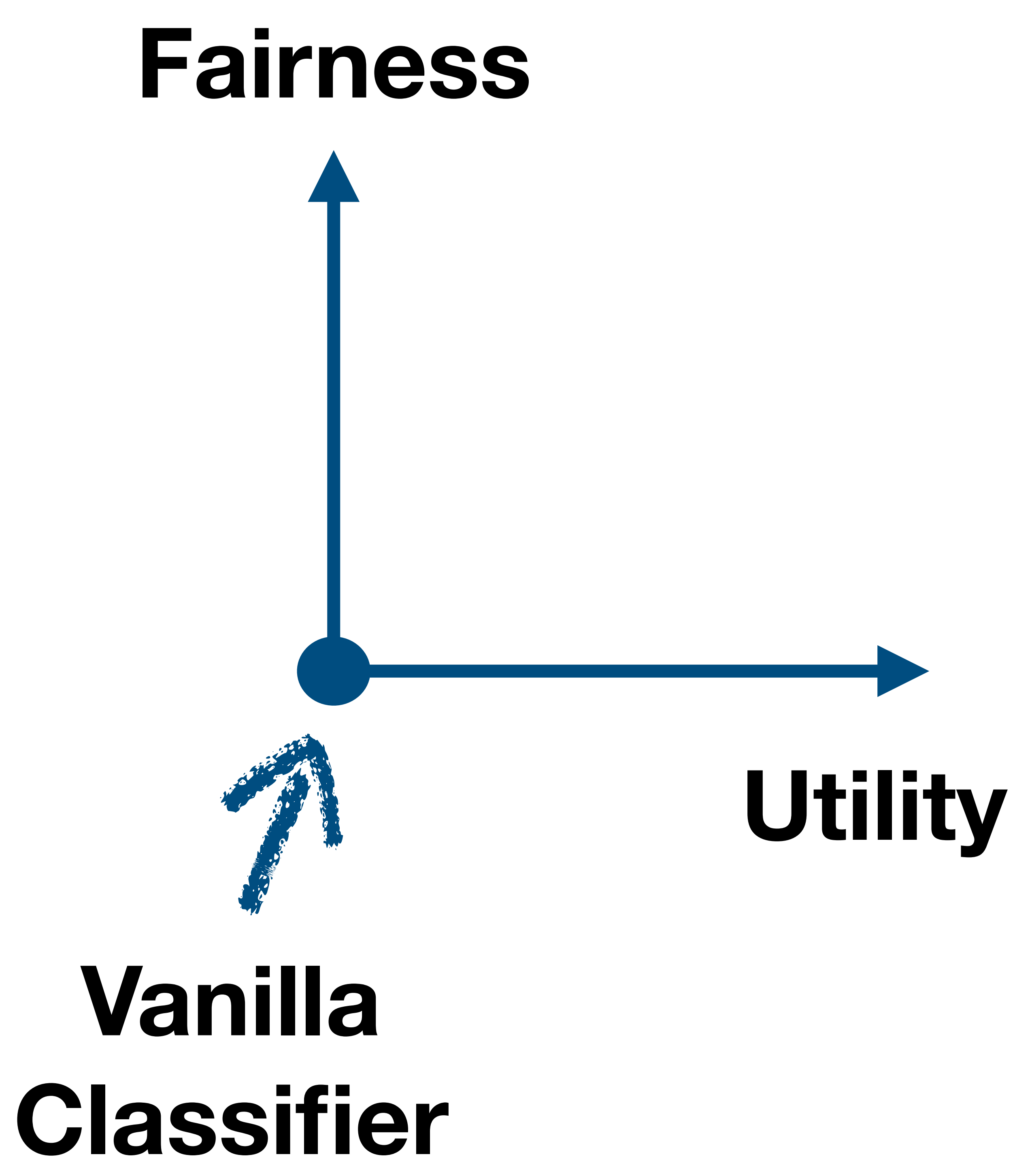
Solving $\mathbf{w}$

$$\text{minimize} \sum_i w_i$$

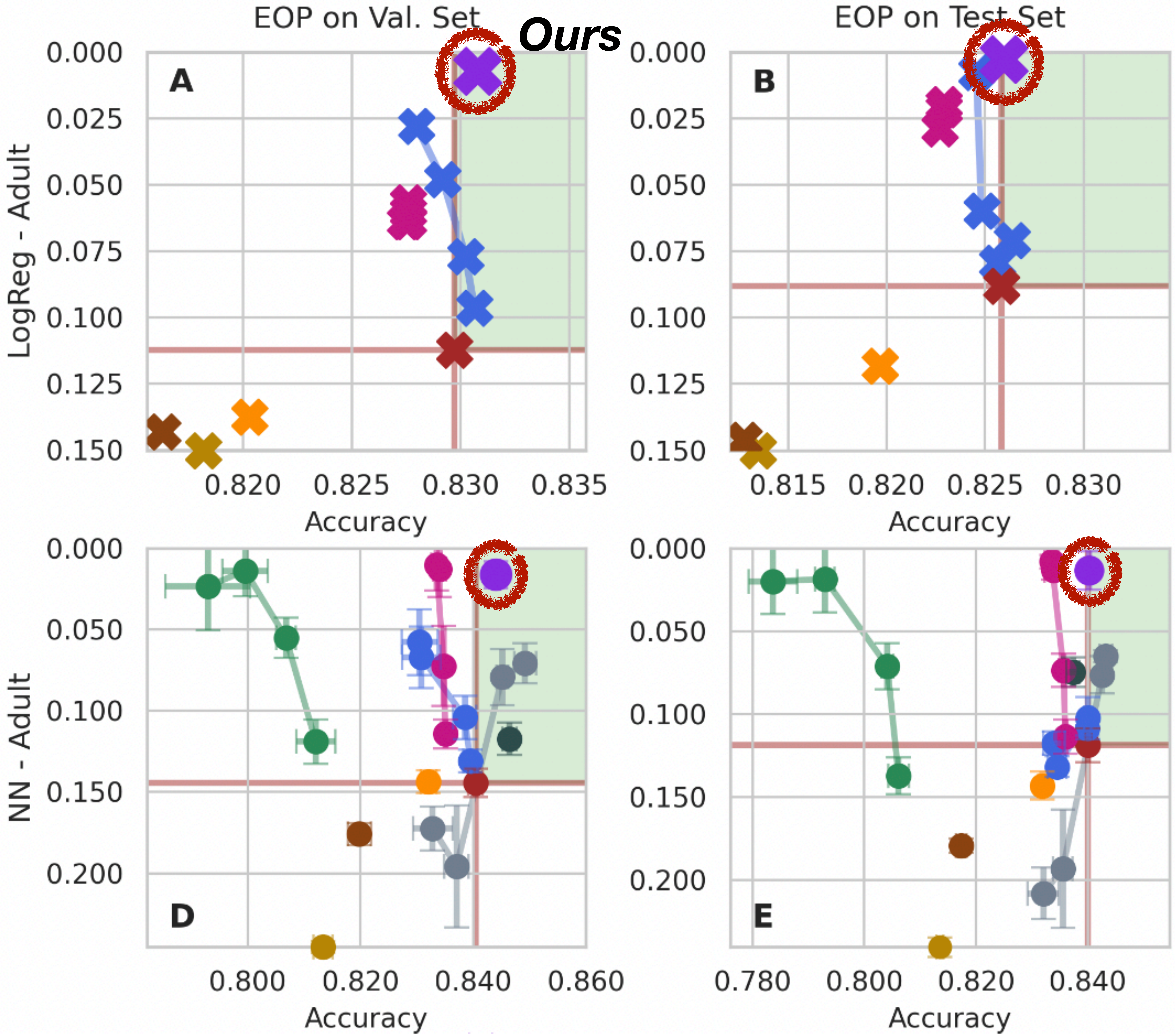
$$\text{subject to} \sum_i w_i \mathcal{I}_{\text{fair}}(\mathbf{e}_i) \leq -(1-\beta) f_{\text{fair}}$$

$$\sum_i w_i \mathcal{I}_{\text{utility}}(\mathbf{e}_i) \leq \gamma (\min_{\mathbf{v}} \sum_i v_i \mathcal{I}_{\text{utility}}(\mathbf{e}_i))$$

$$w_i \in [0,1]$$



Fairness
Utility
Vanilla Classifier



EOP on Val. Set
EOP on Test Set
Ours
LogReg - Adult
NN - Adult
A
B
D
E
Accuracy

Summary

1. A pre-processing approach for Algorithmic Fairness;

2. Better fairness and non-decreasing utility.

For our paper