Gaining Free or Low-Cost Transparency with Interpretable Partial Substitute

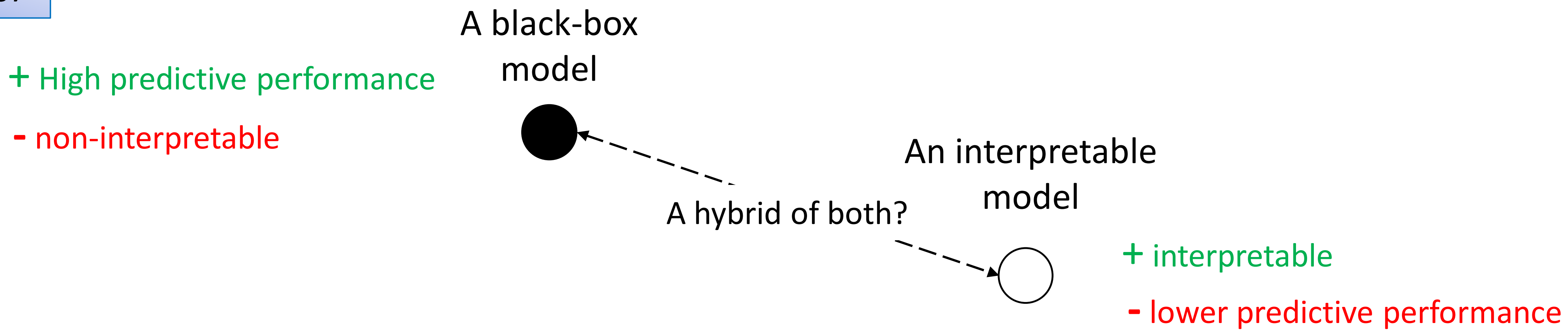
Tong Wang

University of Iowa

Tippie College of Business

tong-wang@uiowa.edu

Poster #67



A black-box model

+ High predictive performance

- non-interpretable

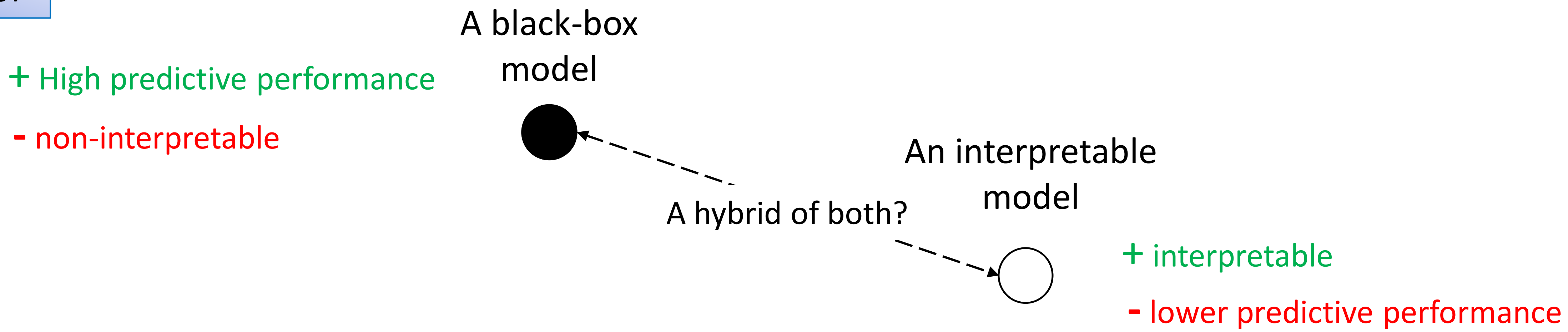
A hybrid of both?

An interpretable model

+ interpretable

- lower predictive performance

Poster #67



A black-box model

+ High predictive performance

- non-interpretable

A hybrid of both?

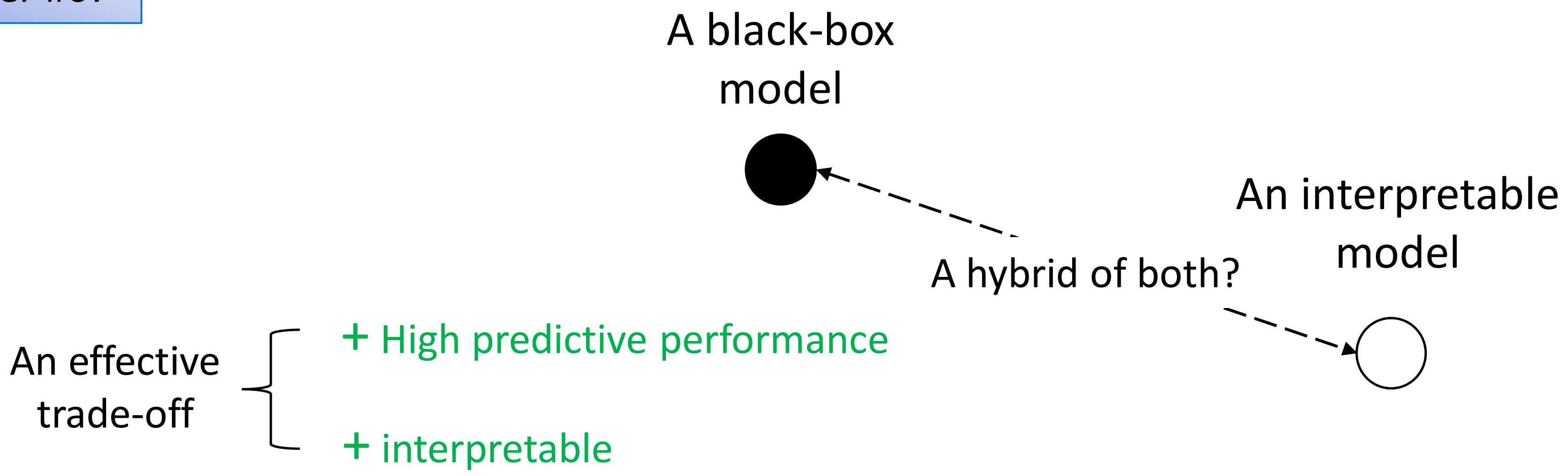
An interpretable model

+ interpretable

- lower predictive performance

A key observation: there might exist a subspace where a black-box is ***overkill*** and a simple interpretable model can perform just as well as the black-box

Poster #67



A black-box model

A hybrid of both?

An interpretable model

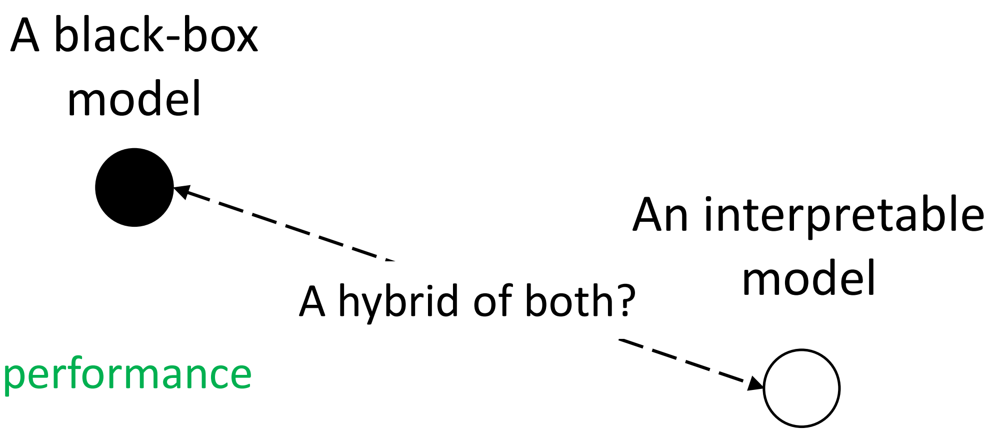
An effective trade-off

- + High predictive performance
- + interpretable

A key observation: there might exist a subspace where a black-box is ***overkill*** and a simple interpretable model can perform just as well as the black-box

The proposed solution: to substitute the black-box model with an interpretable model, where there is no or low-cost of predictive performance

Poster #67

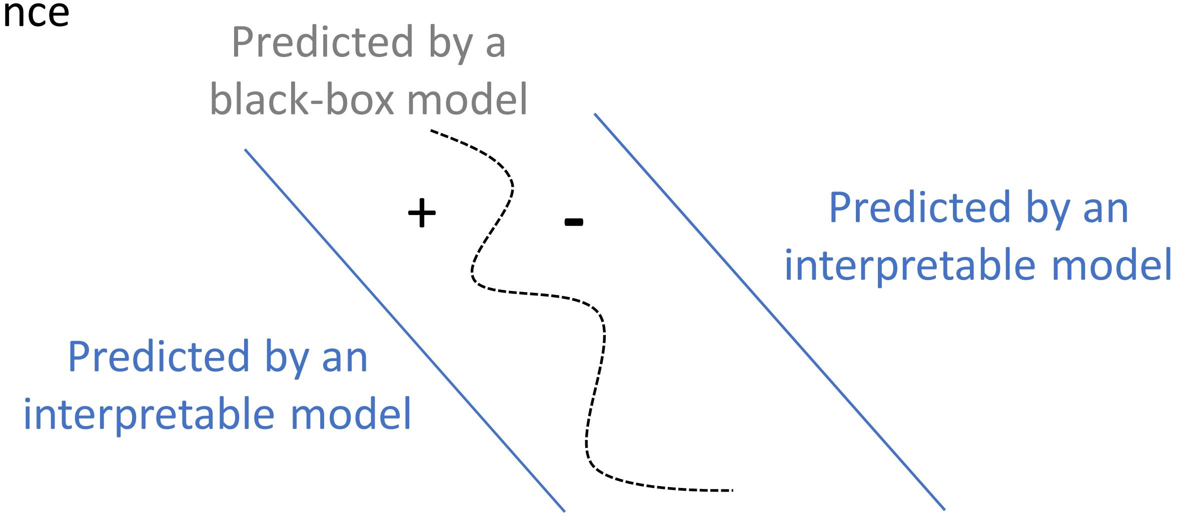


An effective trade-off

- + High predictive performance
- + interpretable

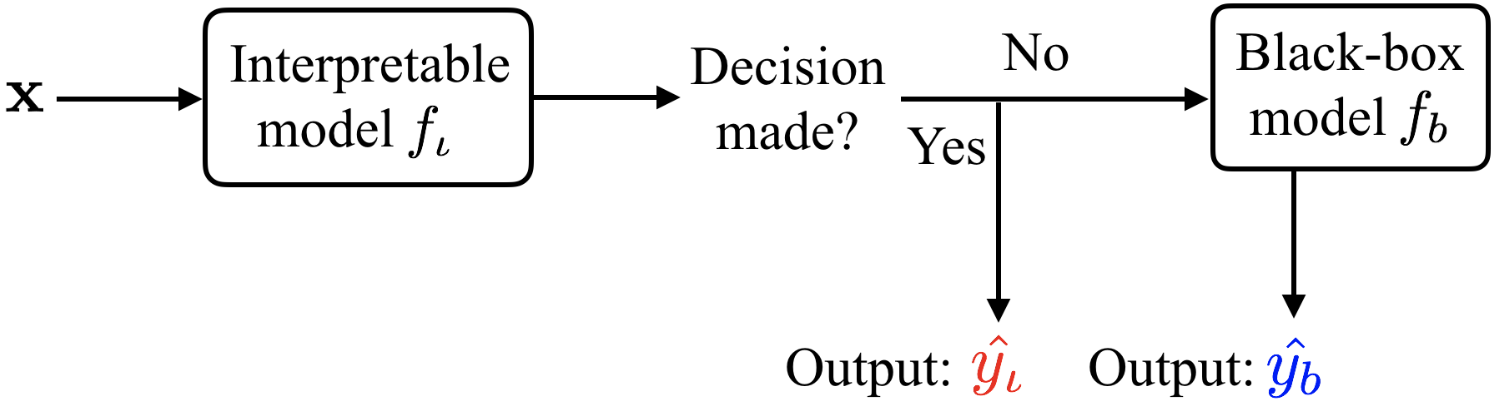
A key observation: there might exist a subspace where a black-box is ***overkill*** and a simple interpretable model can perform just as well as the black-box

The proposed solution: to substitute the black-box model with an interpretable model, where there is no or low-cost of predictive performance


Predicted by a black-box model
+
-
Predicted by an interpretable model
Predicted by an interpretable model

Poster #67

A hybrid predictive model



$\mathbf{x}$

Interpretable model f_ι

Decision made?

No

Yes

Black-box model f_b

Output: $\hat{y}_\iota$

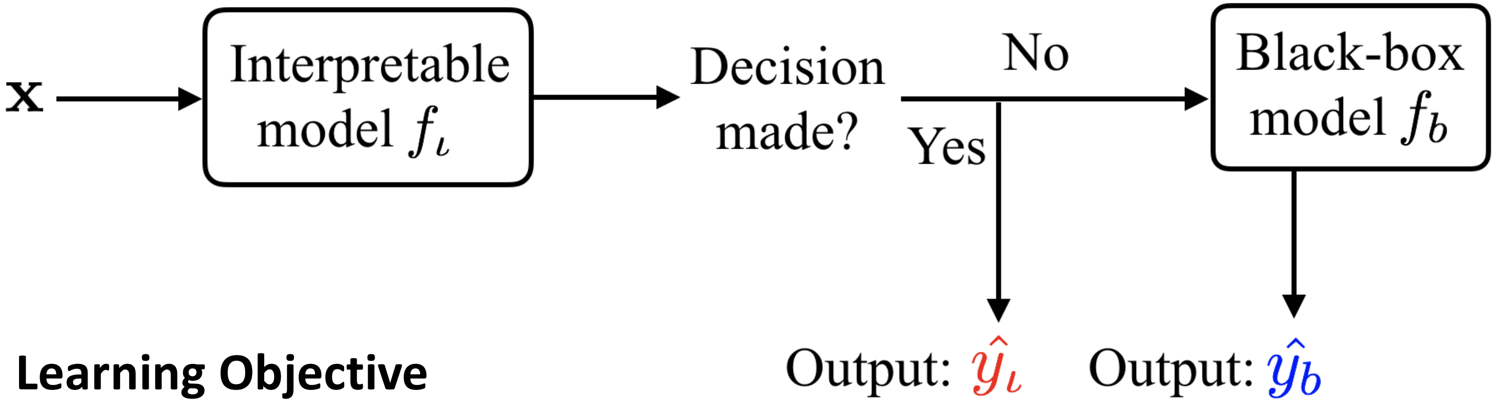
Output: $\hat{y}_b$

$$\mathcal{D} = \mathcal{D}_\iota \cup \mathcal{D}_b$$

Define transparency of model: $\frac{D_i}{D}$

Poster #67

A hybrid predictive model



$\mathbf{x}$

Interpretable model f_ι

Decision made?

No

Yes

Black-box model f_b

Output: $\hat{y}_\iota$

Output: $\hat{y}_b$

Learning Objective

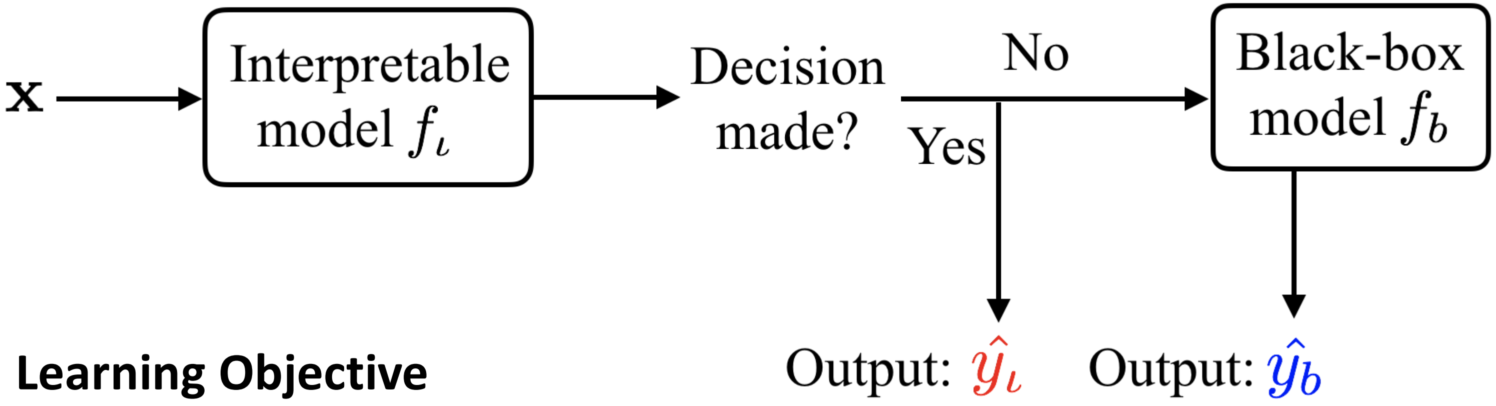
- Predictive performance
- Interpretability of f_i
- Transparency

$$\mathcal{D} = \mathcal{D}_\iota \cup \mathcal{D}_b$$

Define transparency of model: $\frac{D_i}{D}$

Poster #67

A hybrid predictive model



$$\mathcal{D} = \mathcal{D}_\iota \cup \mathcal{D}_b$$

Define transparency of model: $\frac{D_i}{D}$

Learning Objective

- Predictive performance
- Interpretability of f_i
- Transparency

A Hybrid Rule Set

$$\textbf{if } \mathbf{x}_i \text{ obeys } \mathcal{R}_+, Y = 1$$
$$\textbf{else if } \mathbf{x}_i \text{ obeys } \mathcal{R}_-, Y = 0$$
$$\textbf{else } Y = f_b(\mathbf{x}_i)$$

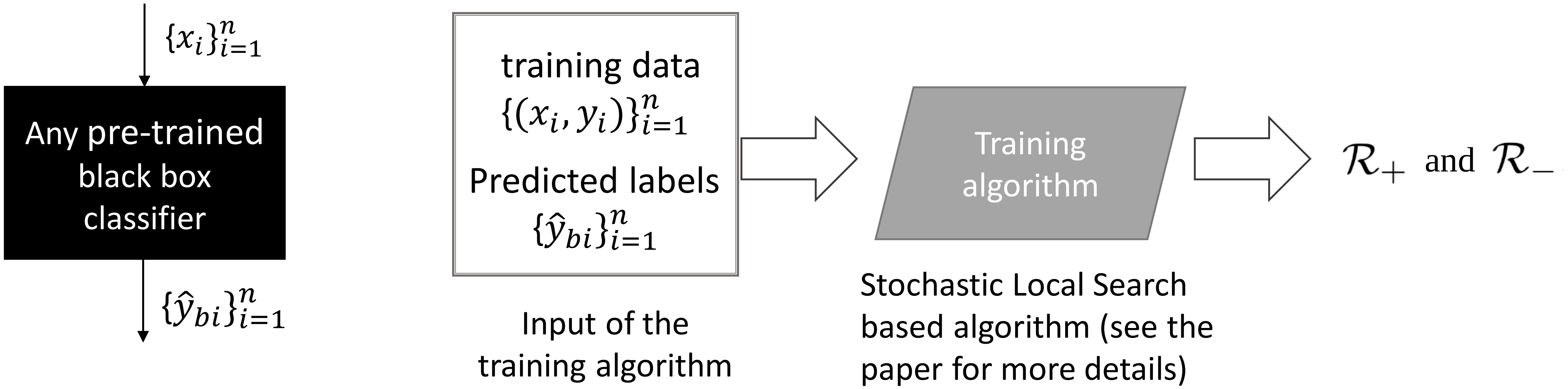
Table 1. An example of a HyRS model

	Rules	**Model**
if	age$<$ 35 *and* maximum heart rate $\geq$ 178 OR serum cholestorol $\geq$ 234 *and* thal $\neq$ 3 *and* the number of vessels $\geq$ 1 $\rightarrow Y = 1$ (heart disease)	$\mathcal{R}_+$
else if	chest pain type $\neq$ 4 *and* age $>$ 40 $\rightarrow Y = 0$ (no heart disease)	$\mathcal{R}_-$
else	$\rightarrow Y = f_b(\mathbf{x})$	f_b

Poster #67

Poster #67

Model Training



$\{x_i\}_{i=1}^{n}$

Any pre-trained black box classifier

$\{\hat{y}_{bi}\}_{i=1}^{n}$

training data $\{(x_i, y_i)\}_{i=1}^{n}$

Predicted labels $\{\hat{y}_{bi}\}_{i=1}^{n}$

Input of the training algorithm

Training algorithm

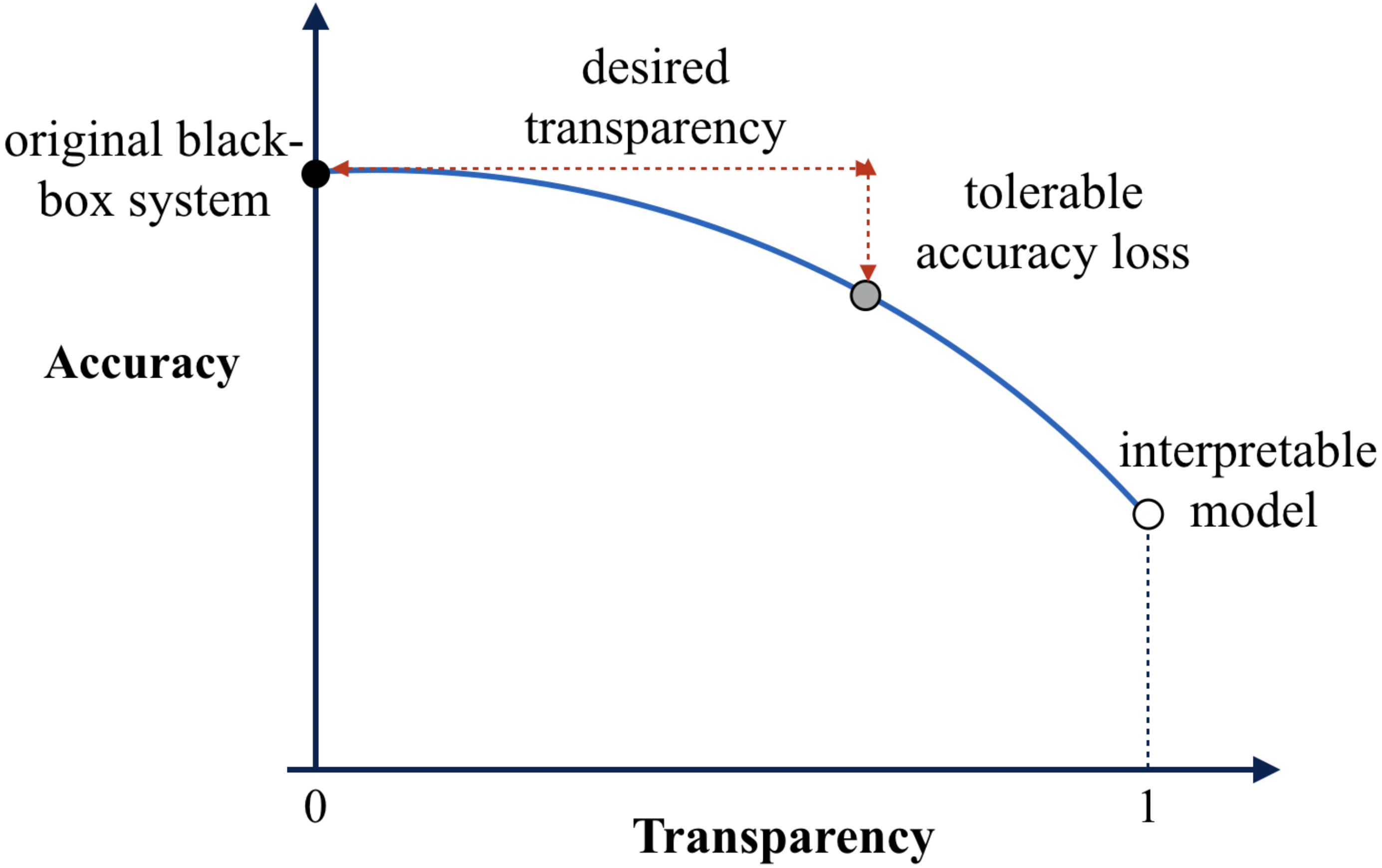
Stochastic Local Search based algorithm (see the paper for more details)

$\mathcal{R}_+$ and $\mathcal{R}_-$

Poster #67

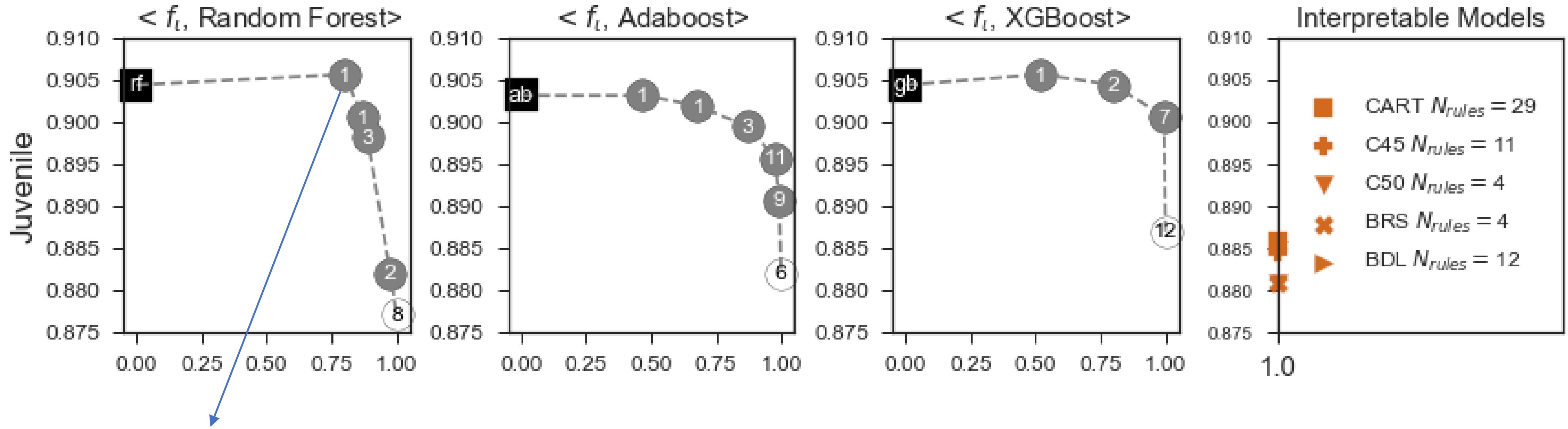
Evaluation:

An efficient frontier that characterizes the trade-off between transparency and accuracy

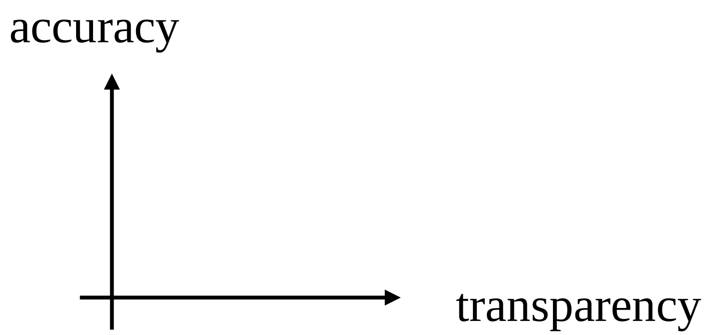


Poster #67

Performance on Juvenile dataset



if Has any of your family members or friends ever attacked you with a weapon $\neq$ Yes *and* Have your friends ever hit or threatened to hit someone without any reason? $\neq$ Yes *and* Have your friends ever broken into a vehicle or building to steal something $\neq$ Yes
then $Y = 0$
else $Y = f_b(\mathbf{x})$



Thank you!

Poster #67 in Pacific Ballroom