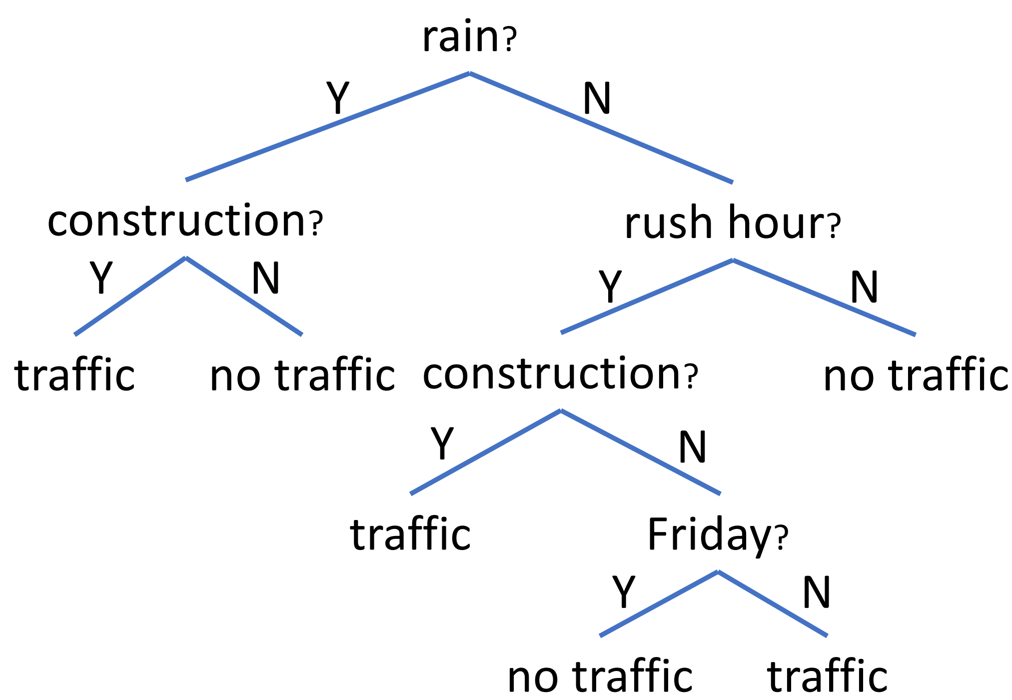
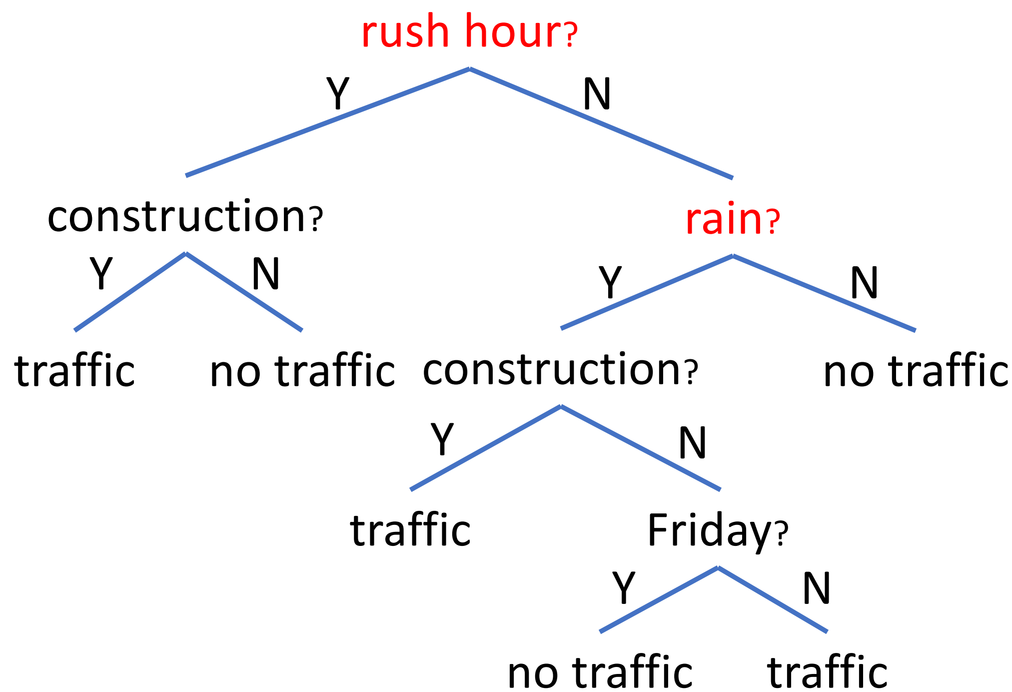
Optimal Sparse Decision Trees

Xiyang Hu, Cynthia Rudin, Margo Seltzer

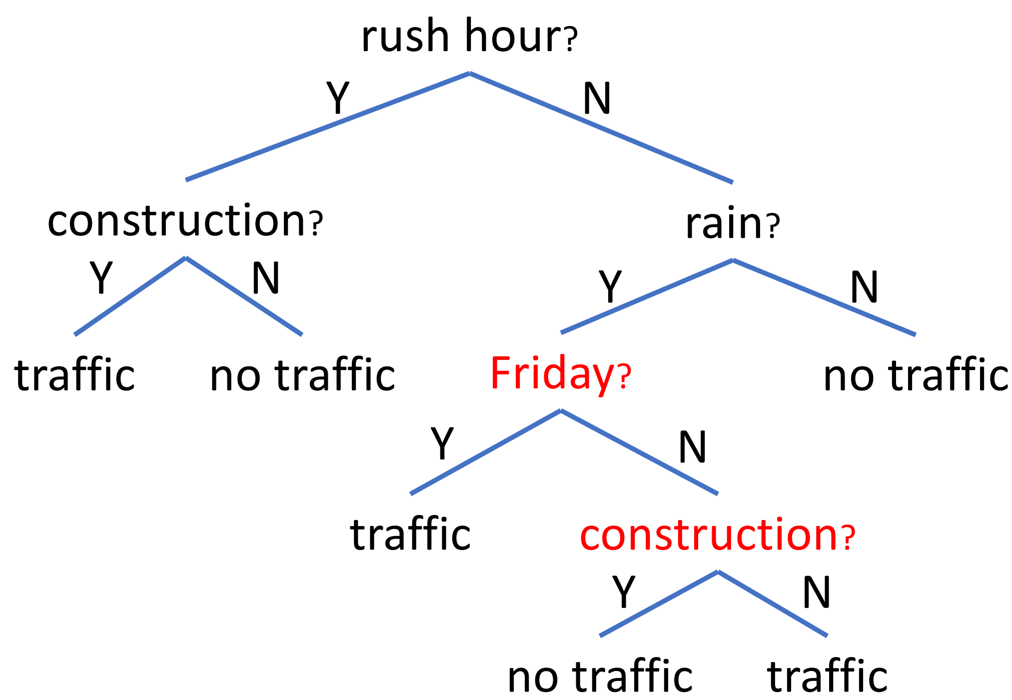
Optimal Sparse Decision Trees



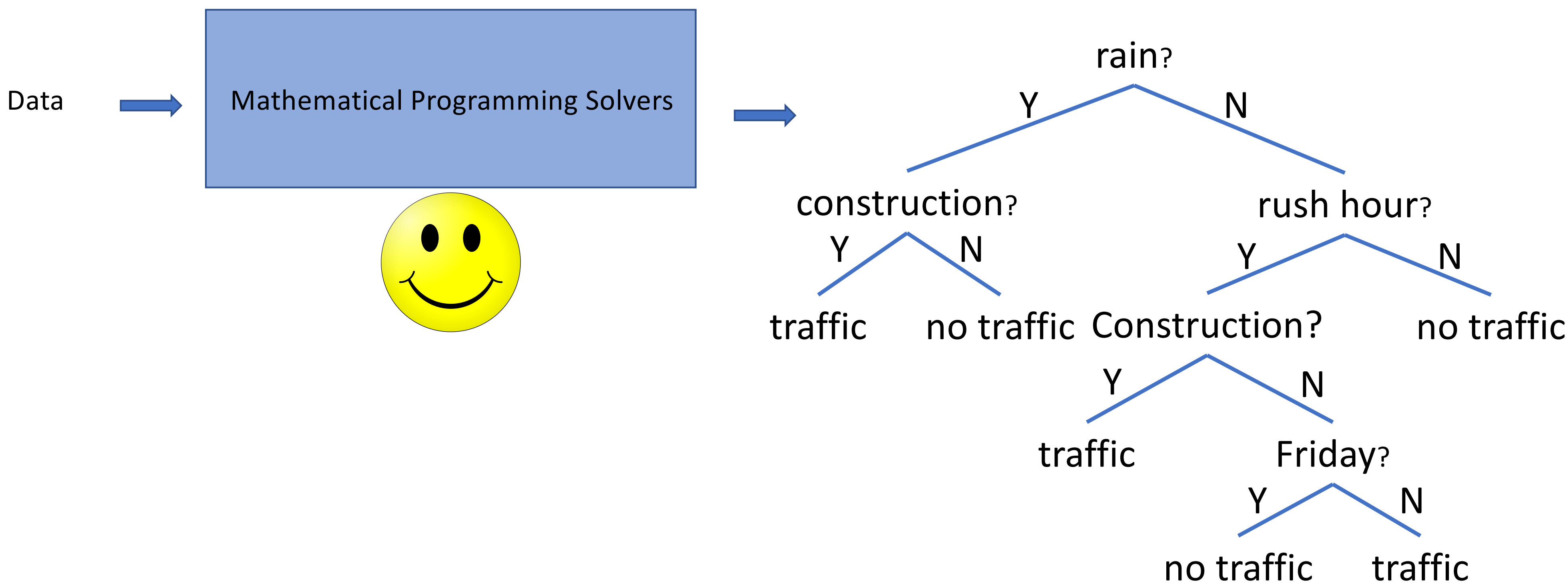
Optimal Sparse Decision Trees



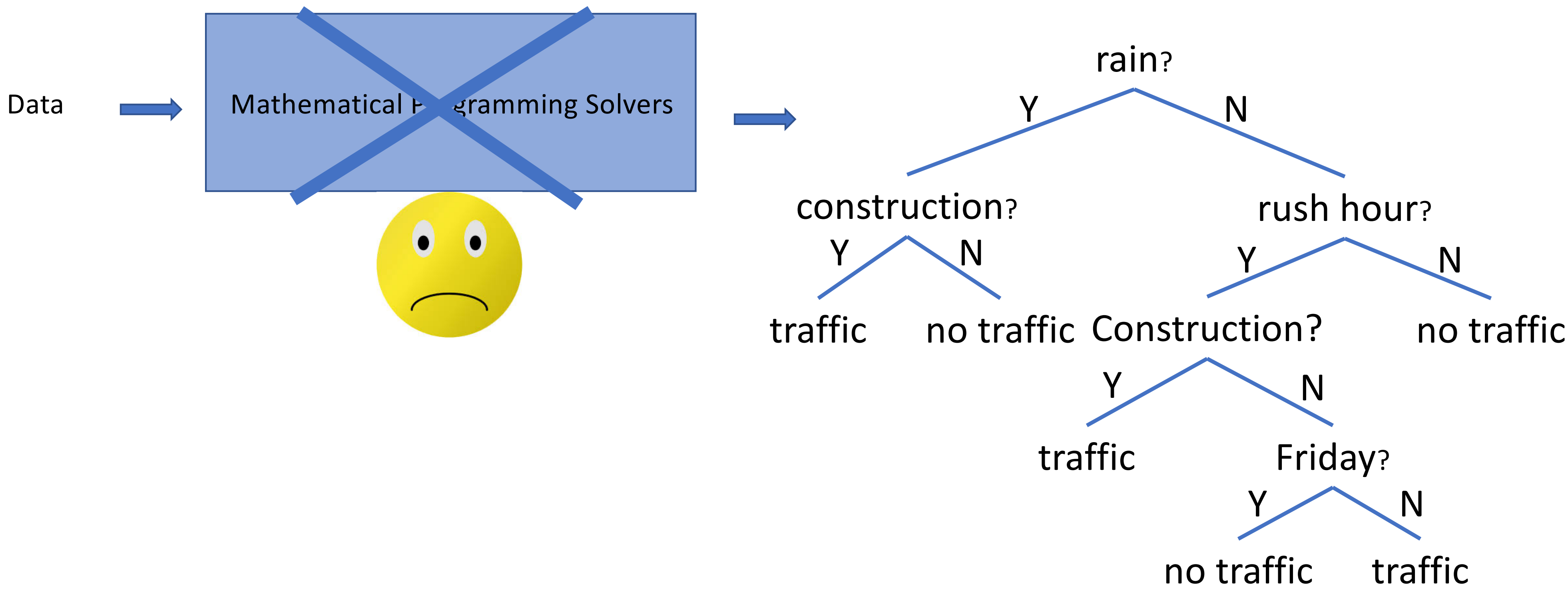
Optimal Sparse Decision Trees



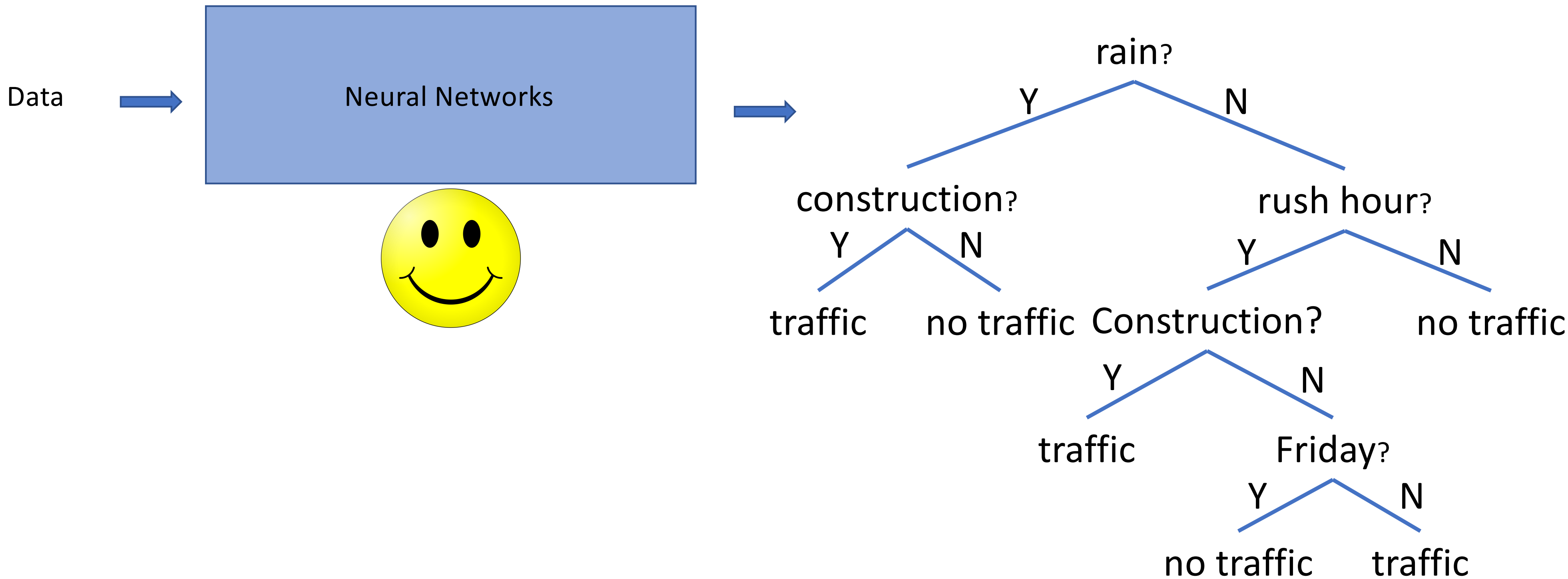
Optimal Sparse Decision Trees



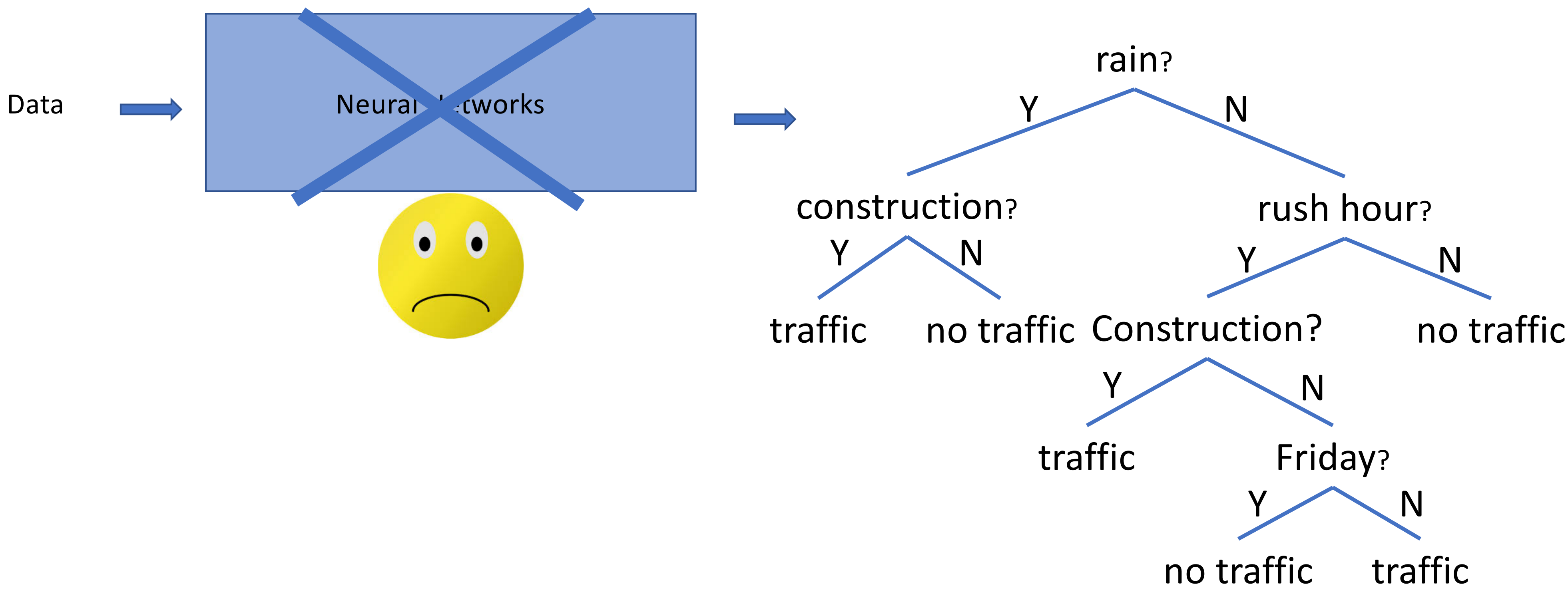
Optimal Sparse Decision Trees



Optimal Sparse Decision Trees



Optimal Sparse Decision Trees



Optimal Sparse Decision Trees

$$\min_{\text{tree}} \hat{L}(\text{tree}, \{(x_i, y_i)\}_i) \text{ where}$$

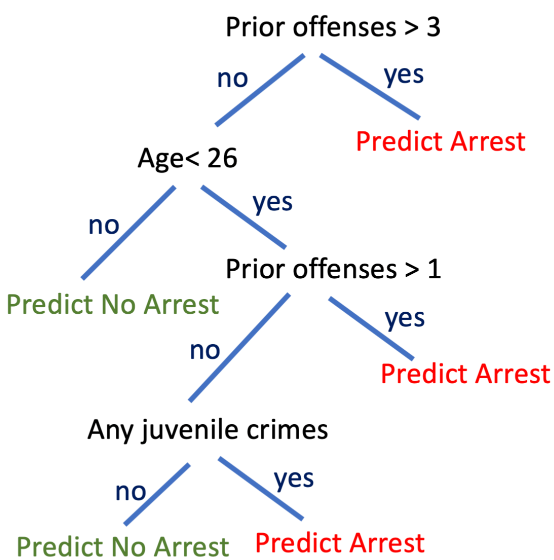
$$\hat{L}(\text{tree}, \{(x_i, y_i)\}_i) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} 1_{[\text{tree}(x_i) \neq y_i]}}_{\text{Misclassification error}} + \underbrace{C(\#\text{ leaves in tree})}_{\text{Sparsity}}$$

We solve this to optimality.
No greedy splitting and pruning like C4.5 and CART
The key: very efficient branch & bound combined with computer systems.

Optimal Sparse Decision Trees

$$\min_{\text{tree}} \hat{L}(\text{tree}, \{(x_i, y_i)\}_i) \text{ where}$$

$$\hat{L}(\text{tree}, \{(x_i, y_i)\}_i) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} 1_{[\text{tree}(x_i) \neq y_i]}}_{\text{Misclassification error}} + \underbrace{C(\#\,\text{leaves in tree})}_{\text{Sparsity}}$$

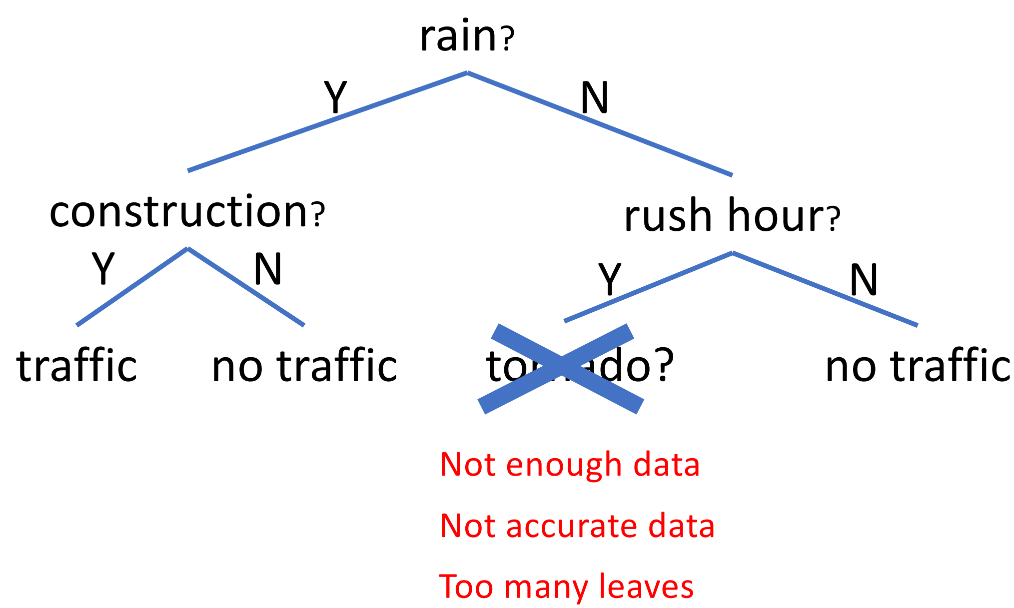


An example of an optimal tree on the re-arrest data

Optimal Sparse Decision Trees

Analytical Bounds Reduce the Search Space

This collection of theorems show that some partial trees can never be extended to form optimal trees.



Optimal Sparse Decision Trees

Represent a tree by its leaves

rain & construction & traffic

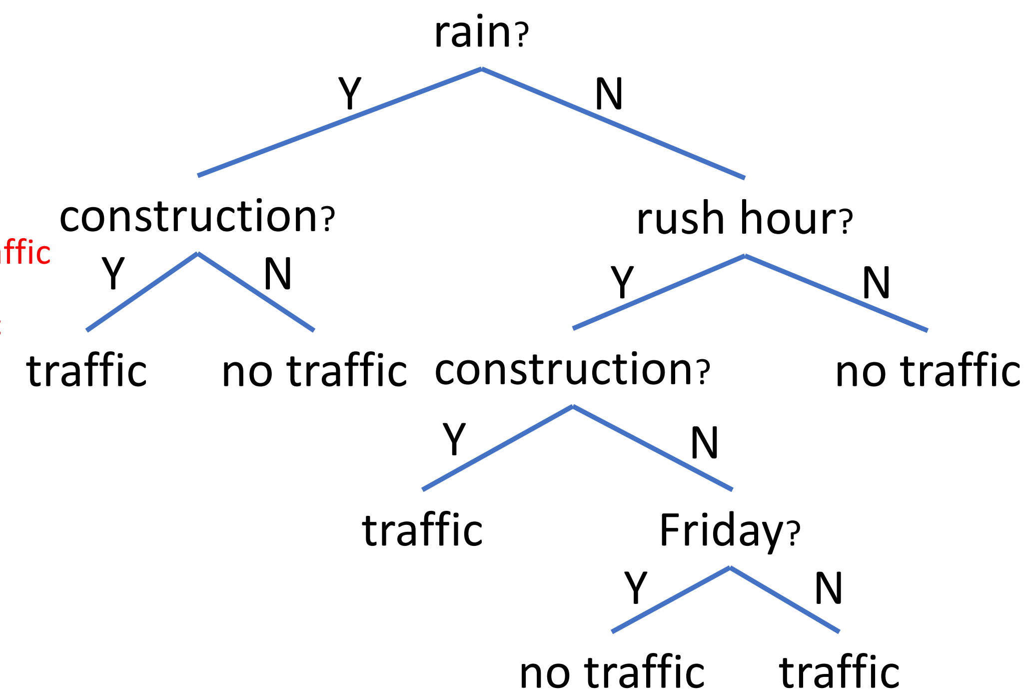
rain & no construction & no traffic

no rain & rush hour & construction & traffic

no rain & rush hour & no construction & Friday and no traffic

no rain & rush hour & no construction & Friday and traffic

no rain & no rush hour & no traffic



Optimal Sparse Decision Trees

Permutation map: Discover identical trees already evaluated

rain & construction & traffic

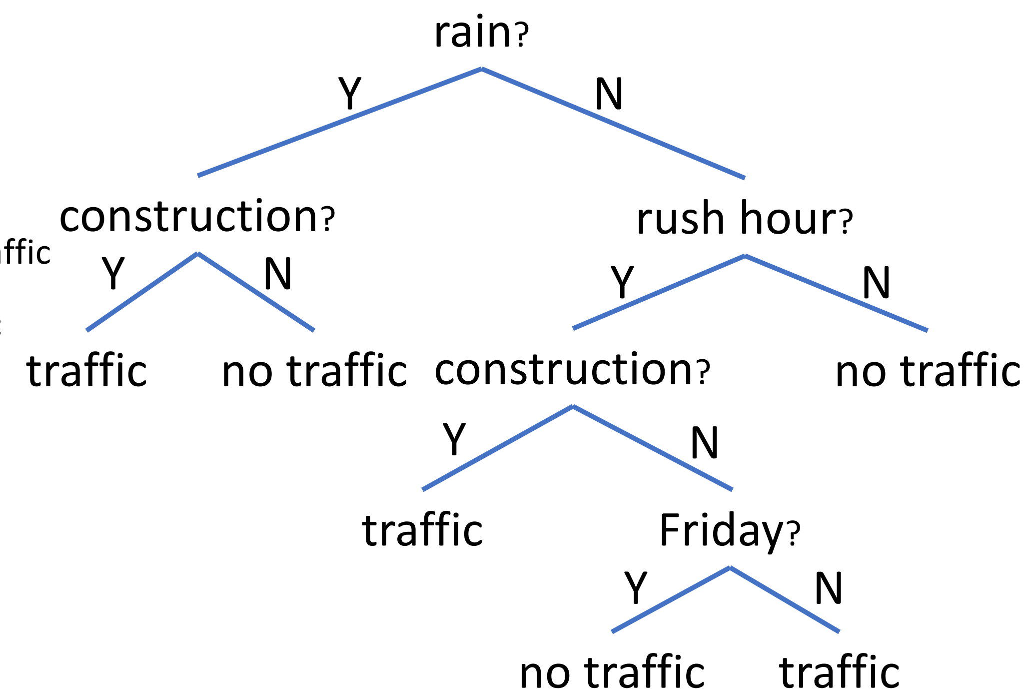
rain & no construction & no traffic

no rain & rush hour & construction & traffic

no rain & rush hour & no construction & Friday and no traffic

no rain & rush hour & no construction & Friday and traffic

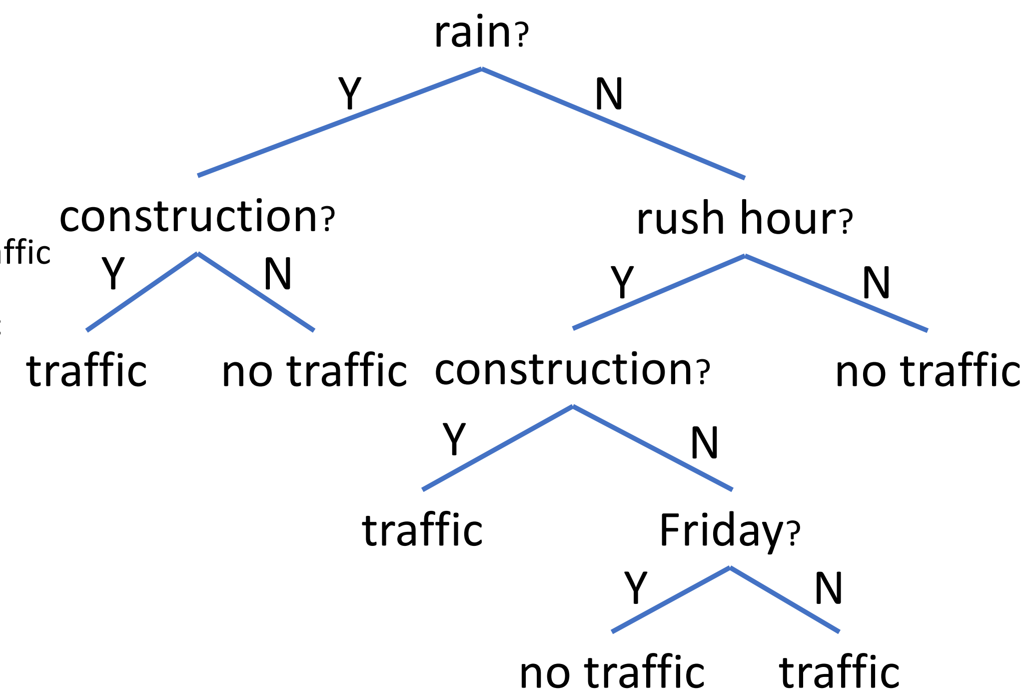
no rain & no rush hour & no traffic



Optimal Sparse Decision Trees

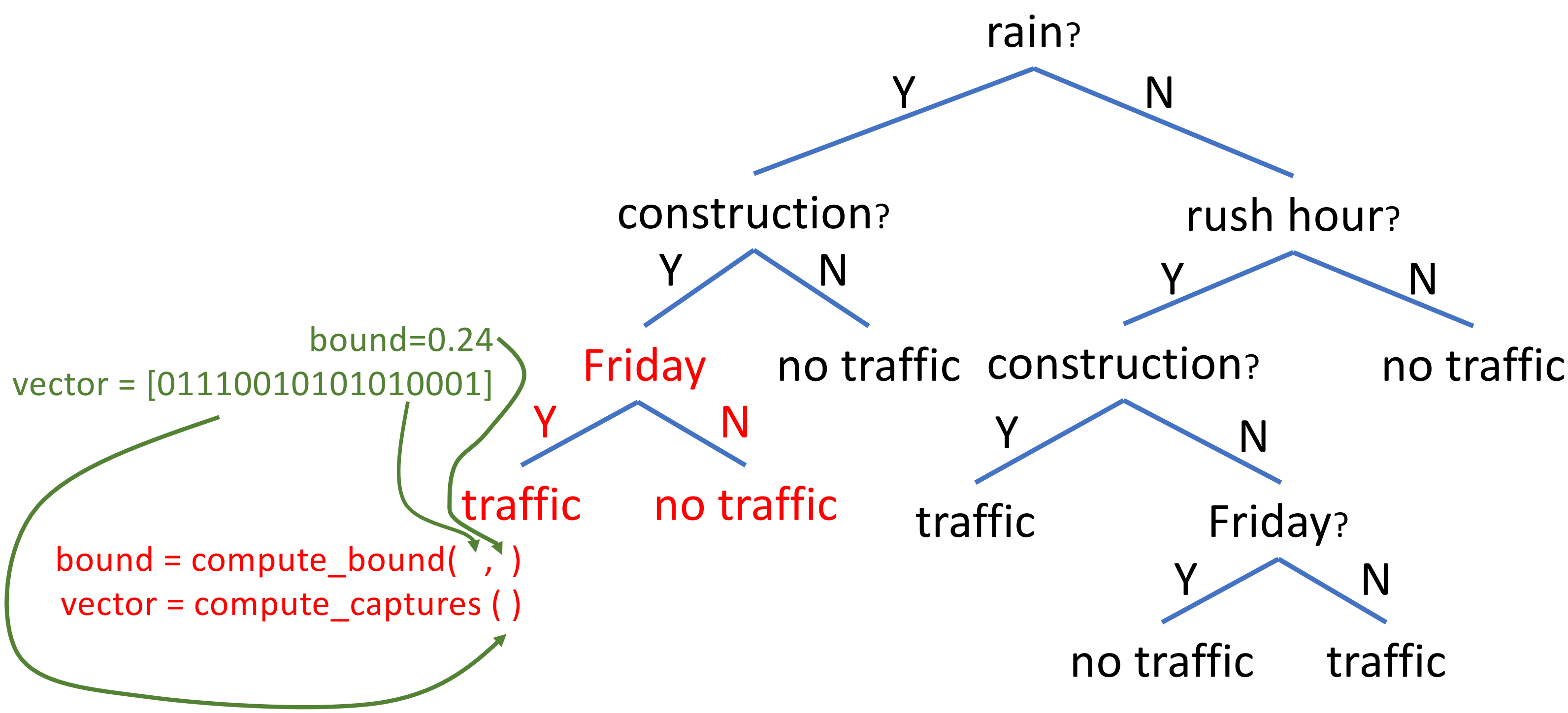
Bit-vectors describe data represented by each leaf

rain & construction & traffic
[10000100010011100000.........................0]
rain & no construction & no traffic
[0110001000000000110.........................1]
no rain & rush hour & construction & traffic
[0001000100000001000.........................0]
no rain & rush hour & no construction & Friday and no traffic
[0000100000000000001.........................0]
no rain & rush hour & no construction & Friday and traffic
[0000000010000000000.........................0]
no rain & no rush hour & no traffic
[0000000000011000000.........................0]



Optimal Sparse Decision Trees

Incremental computation of objective and bounds



Optimal Sparse Decision Trees

Strong analytical bounds

+

Leaf-based representation

+

Permutation map

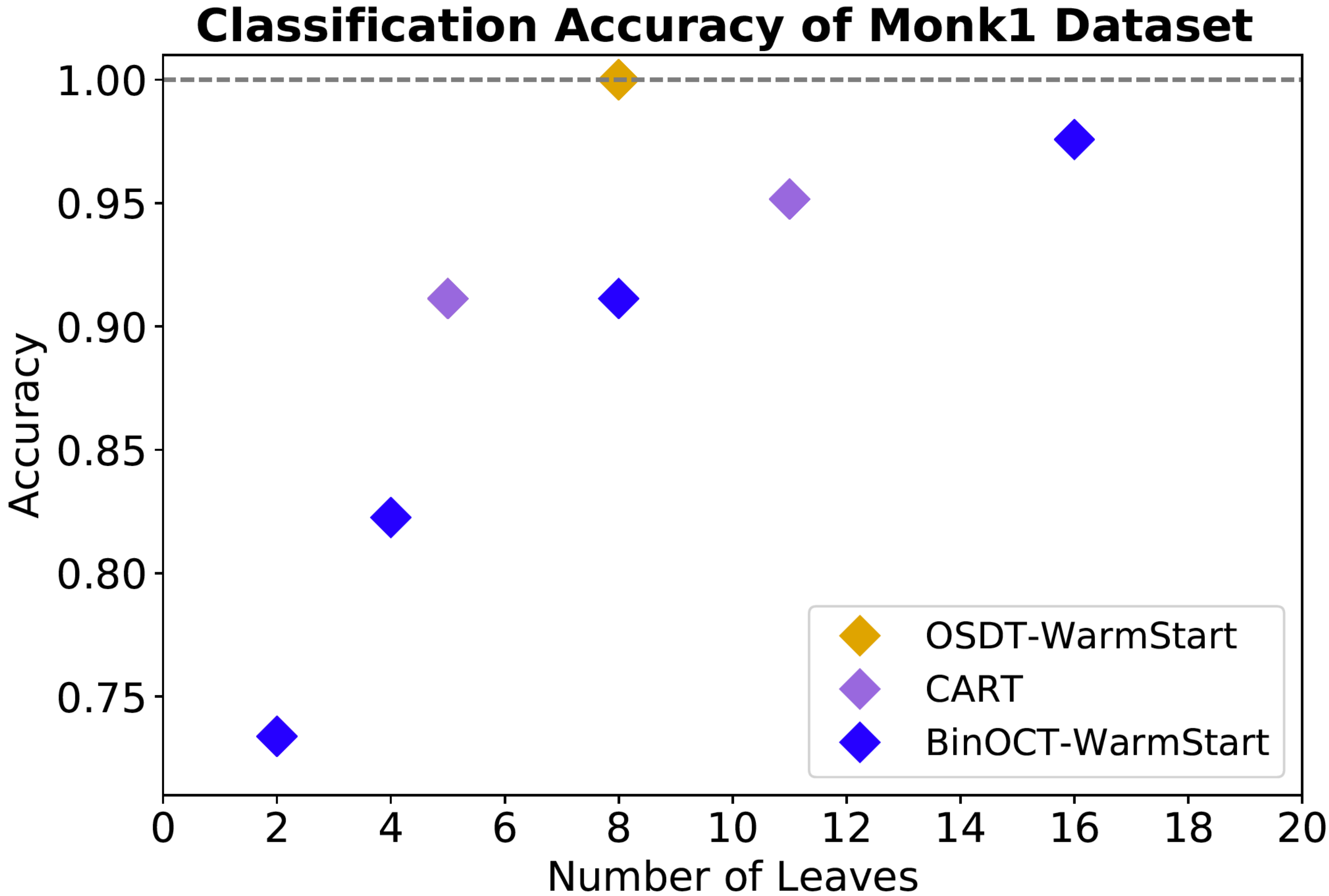
+

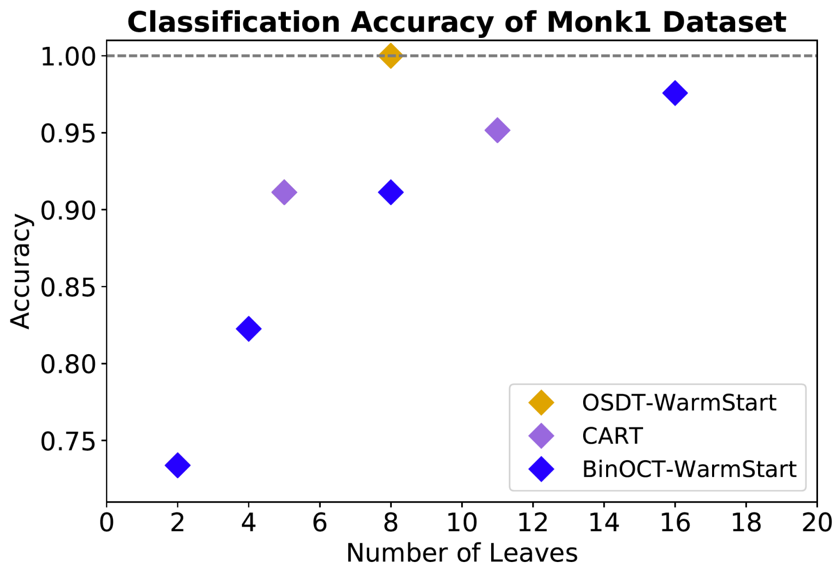
Caching of intermediate results

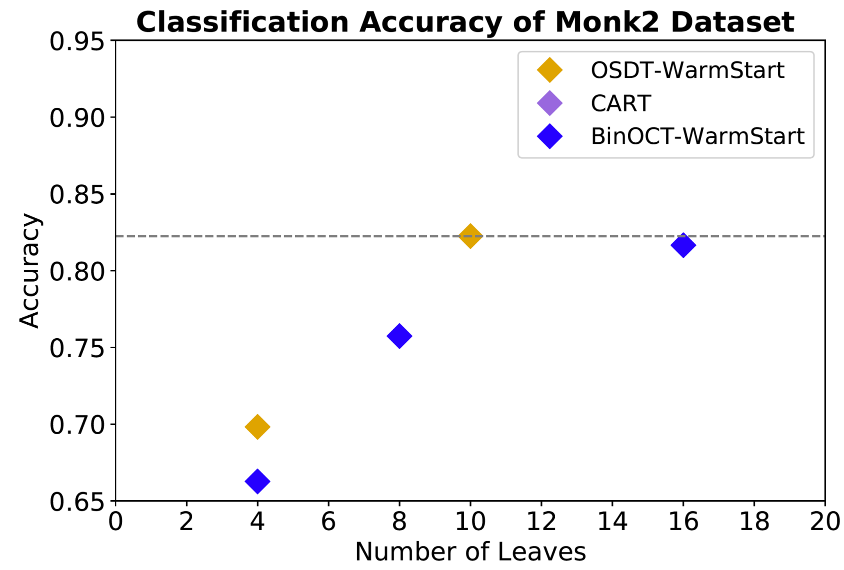
+

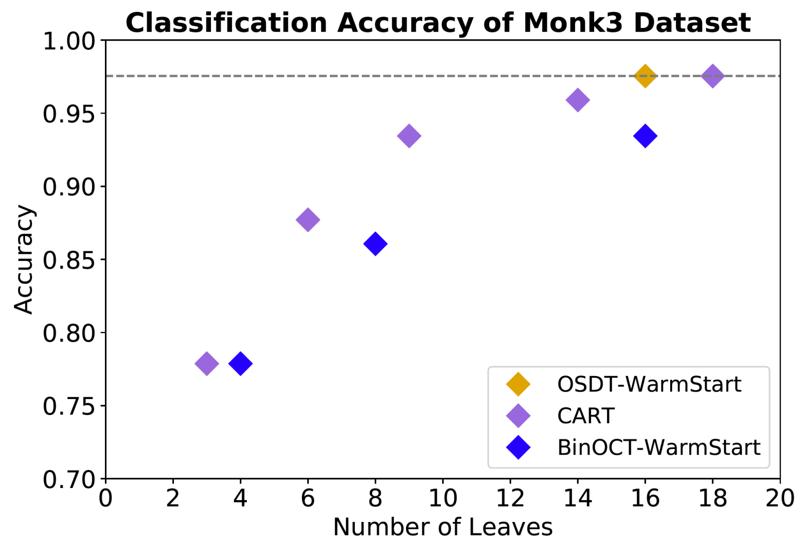
Incremental computation

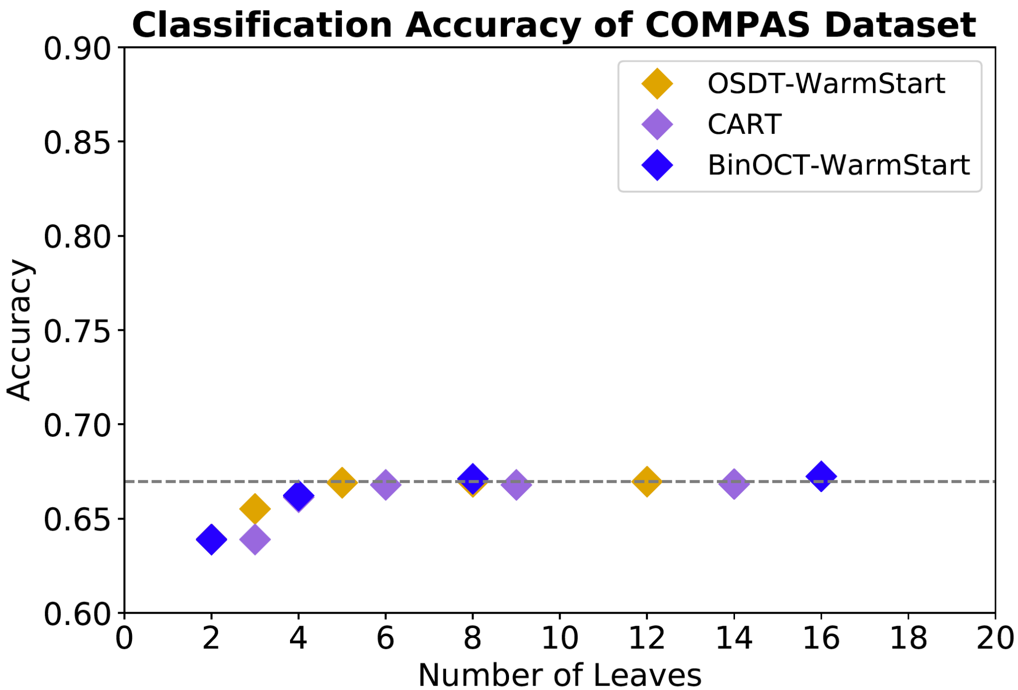
= Fast Implementation

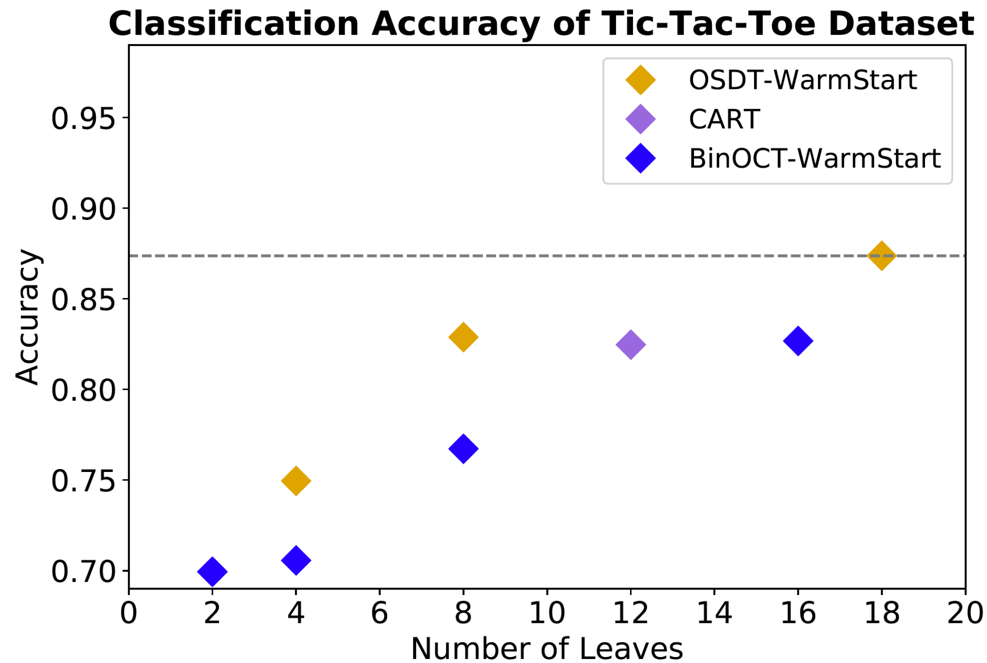

Classification Accuracy of Monk1 Dataset
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0
2
4
6
8
10
12
14
16
18
20
Number of Leaves
OSDT-WarmStart
CART
BinOCT-WarmStart

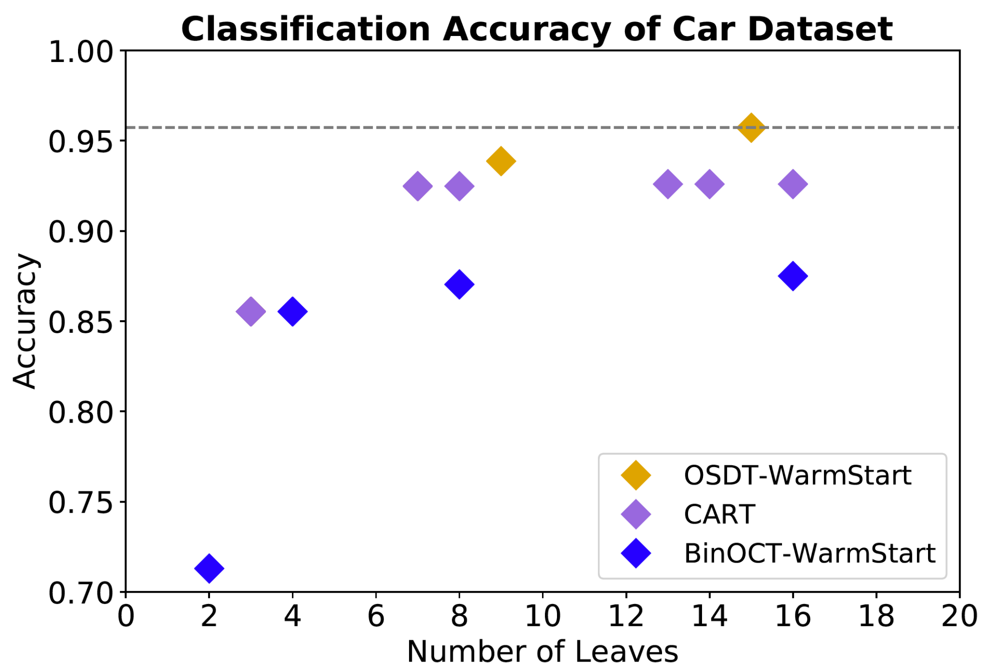

Classification Accuracy of Monk1 Dataset
Accuracy
Number of Leaves
OSDT-WarmStart
CART
BinOCT-WarmStart


Classification Accuracy of Monk2 Dataset
Accuracy
Number of Leaves
OSDT-WarmStart
CART
BinOCT-WarmStart

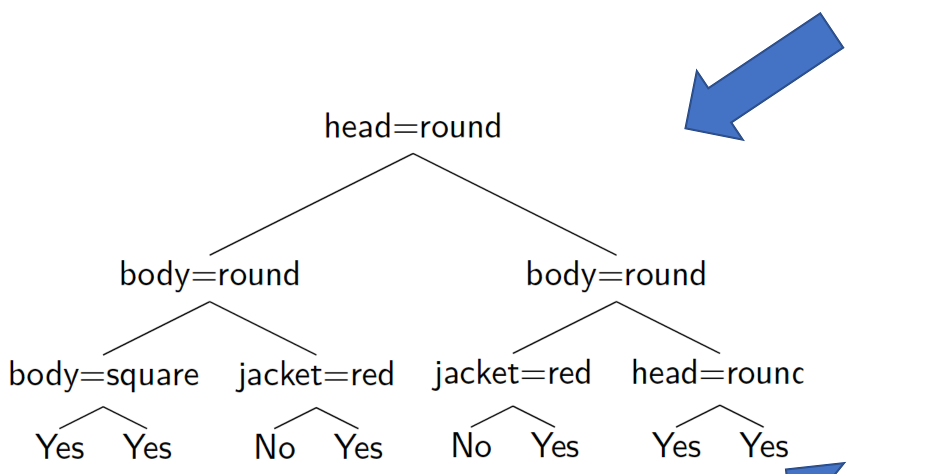

Classification Accuracy of Monk3 Dataset
Accuracy
Number of Leaves
OSDT-WarmStart
CART
BinOCT-WarmStart


Classification Accuracy of COMPAS Dataset
OSDT-WarmStart
CART
BinOCT-WarmStart
Accuracy
Number of Leaves
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0 2 4 6 8 10 12 14 16 18 20

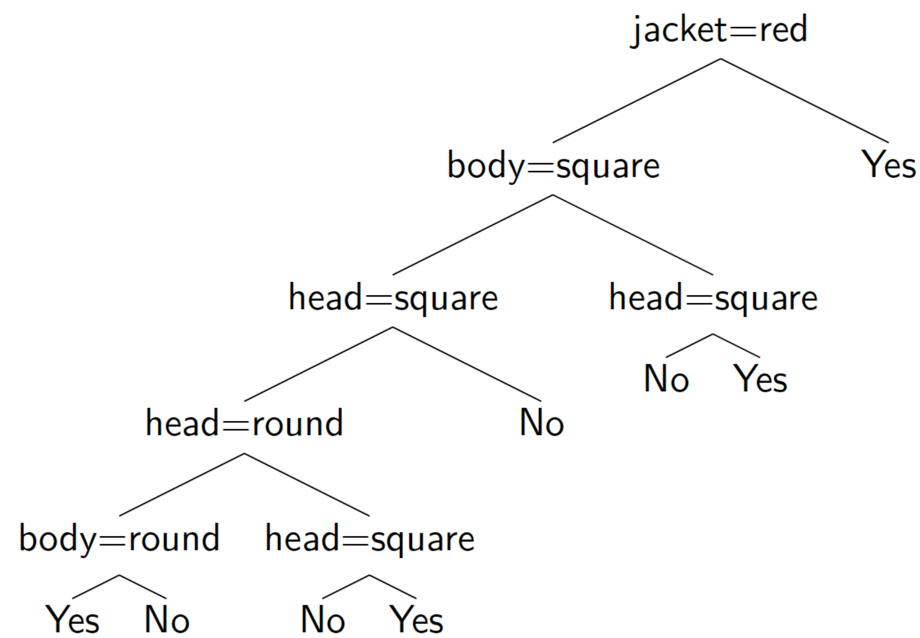

Classification Accuracy of Tic-Tac-Toe Dataset
OSDT-WarmStart
CART
BinOCT-WarmStart
Accuracy
Number of Leaves
0.95
0.90
0.85
0.80
0.75
0.70
0 2 4 6 8 10 12 14 16 18 20


Classification Accuracy of Car Dataset
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
2
4
6
8
10
12
14
16
18
20
Number of Leaves
OSDT-WarmStart
CART
BinOCT-WarmStart

Monk 1 dataset

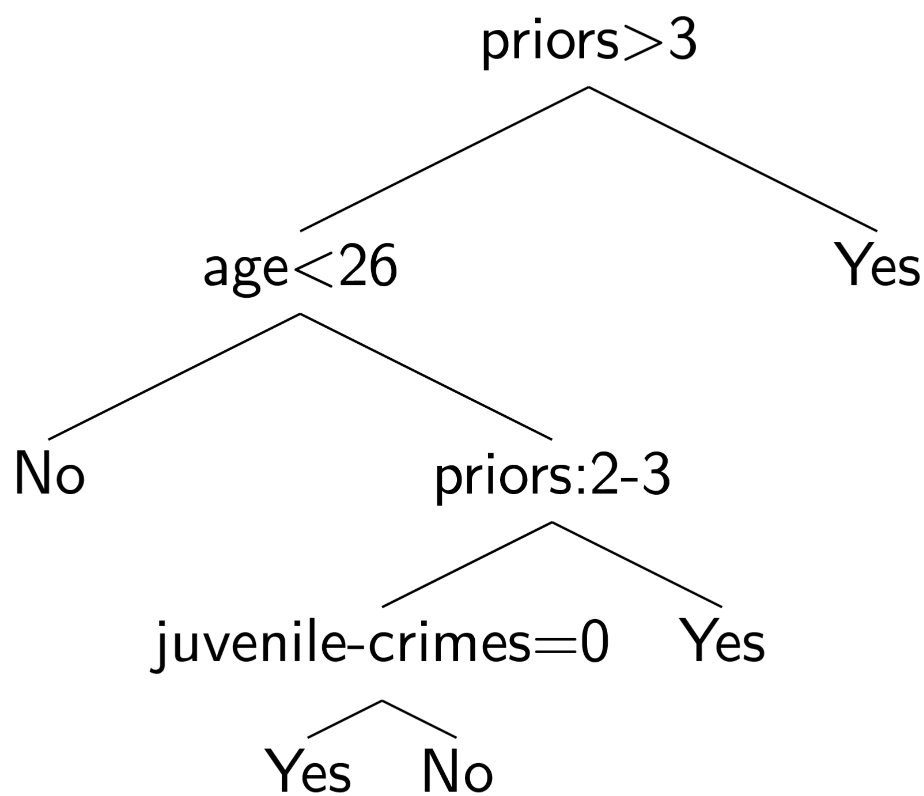


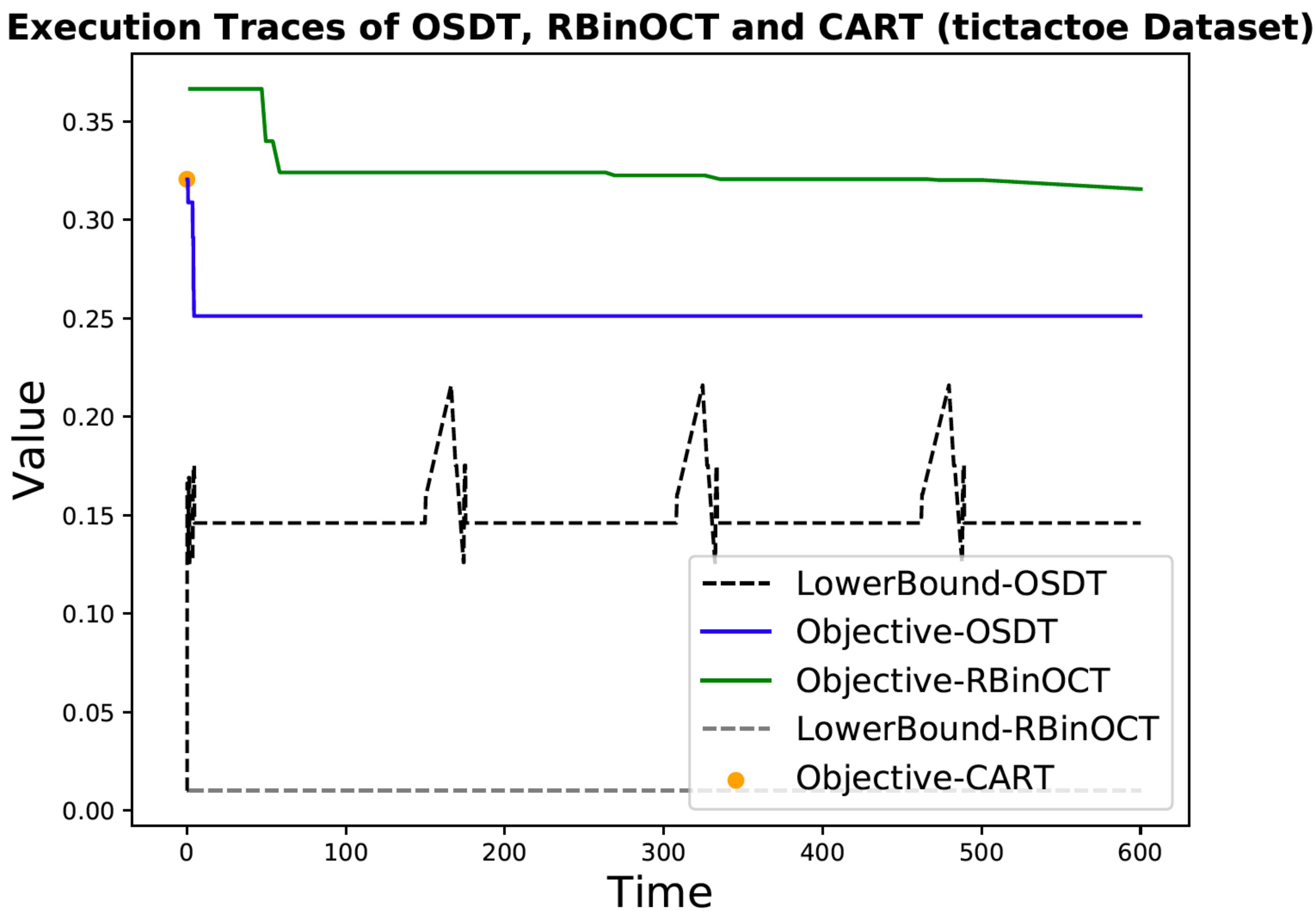
(a) Bin[illegible]CT (accuracy: 91.129%)



(b) OSDT (accuracy: 100%)

COMPAS dataset




Execution Traces of OSDT, RBinOCT and CART (tictactoe Dataset)
Value
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Time
0
100
200
300
400
500
600
LowerBound-OSDT
Objective-OSDT
Objective-RBinOCT
LowerBound-RBinOCT
Objective-CART

Summarize

- First practical method for optimal sparse binary split decision trees
- Current work:
  - Non-straightforward speedup for continuous variables
  - Generalization to other objectives

Xiyang Hu, Cynthia Rudin, Margo Seltzer. Optimal Sparse Decision Trees.
https://arxiv.org/abs/1904.12847