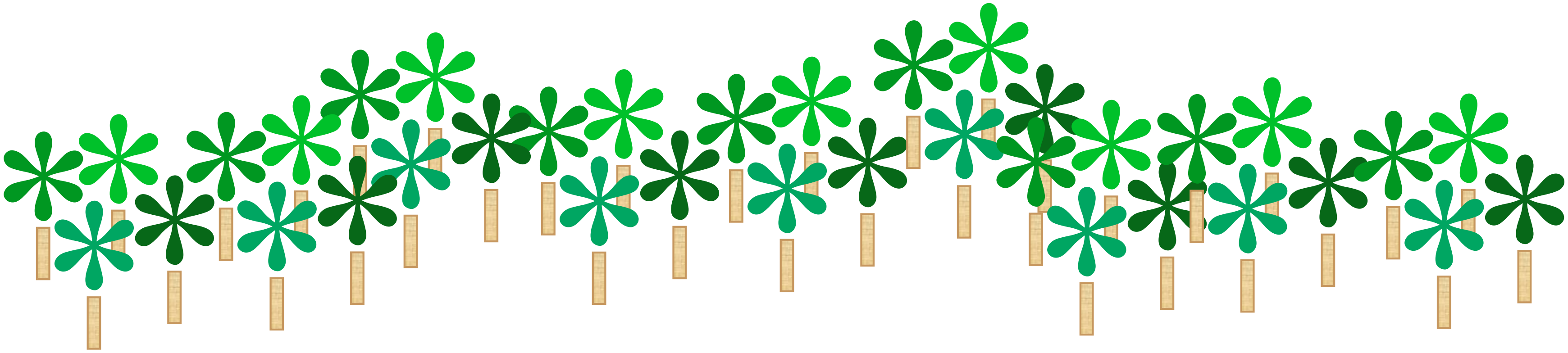
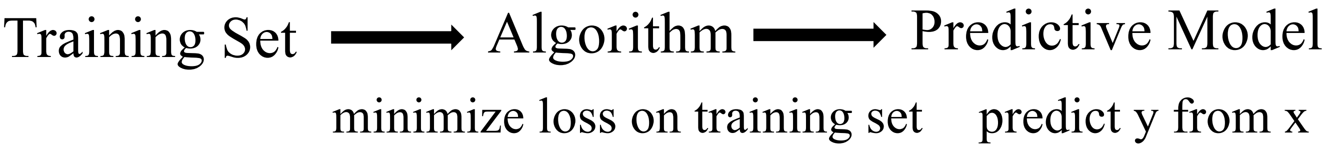
Exploring the Whole Rashomon Set of Sparse Decision Trees

Rui Xin*, Chudi Zhong*, Zhi Chen*, Takuya Takagi, Margo Seltzer, Cynthia Rudin



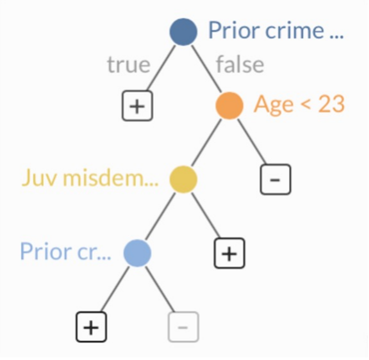
The Universal Paradigm of Machine Learning



Training Set → Algorithm → Predictive Model

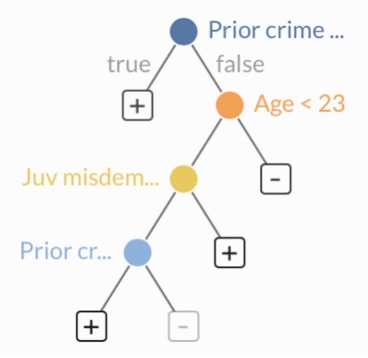
minimize loss on training set

predict y from x



Training Set ⟶ Algorithm ⟶ Predictive Model

minimize loss on training set   predict y from x



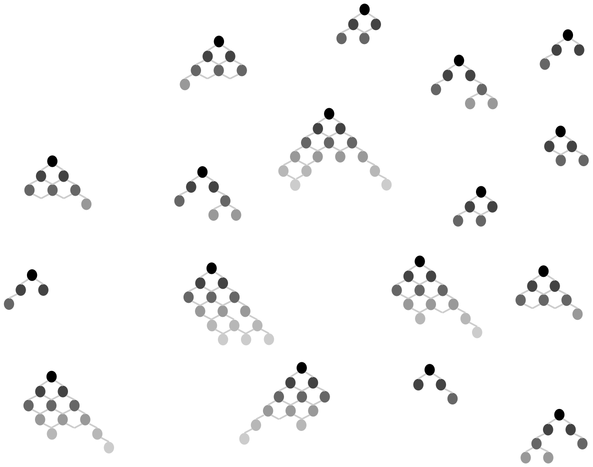
😕

“Uhh, there’s something wrong with this model”

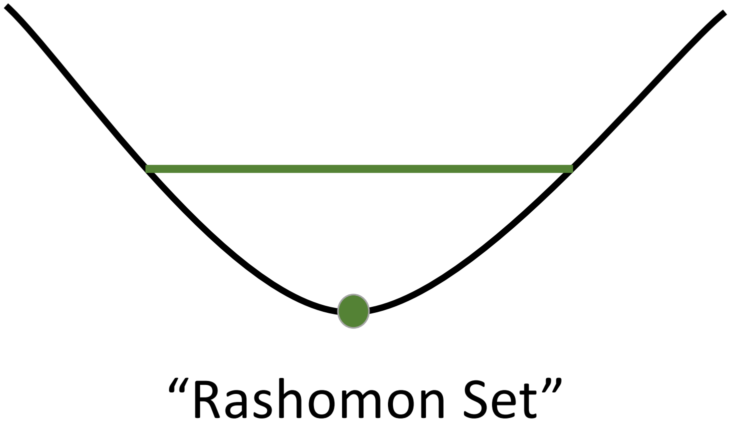
A New Paradigm of Machine Learning

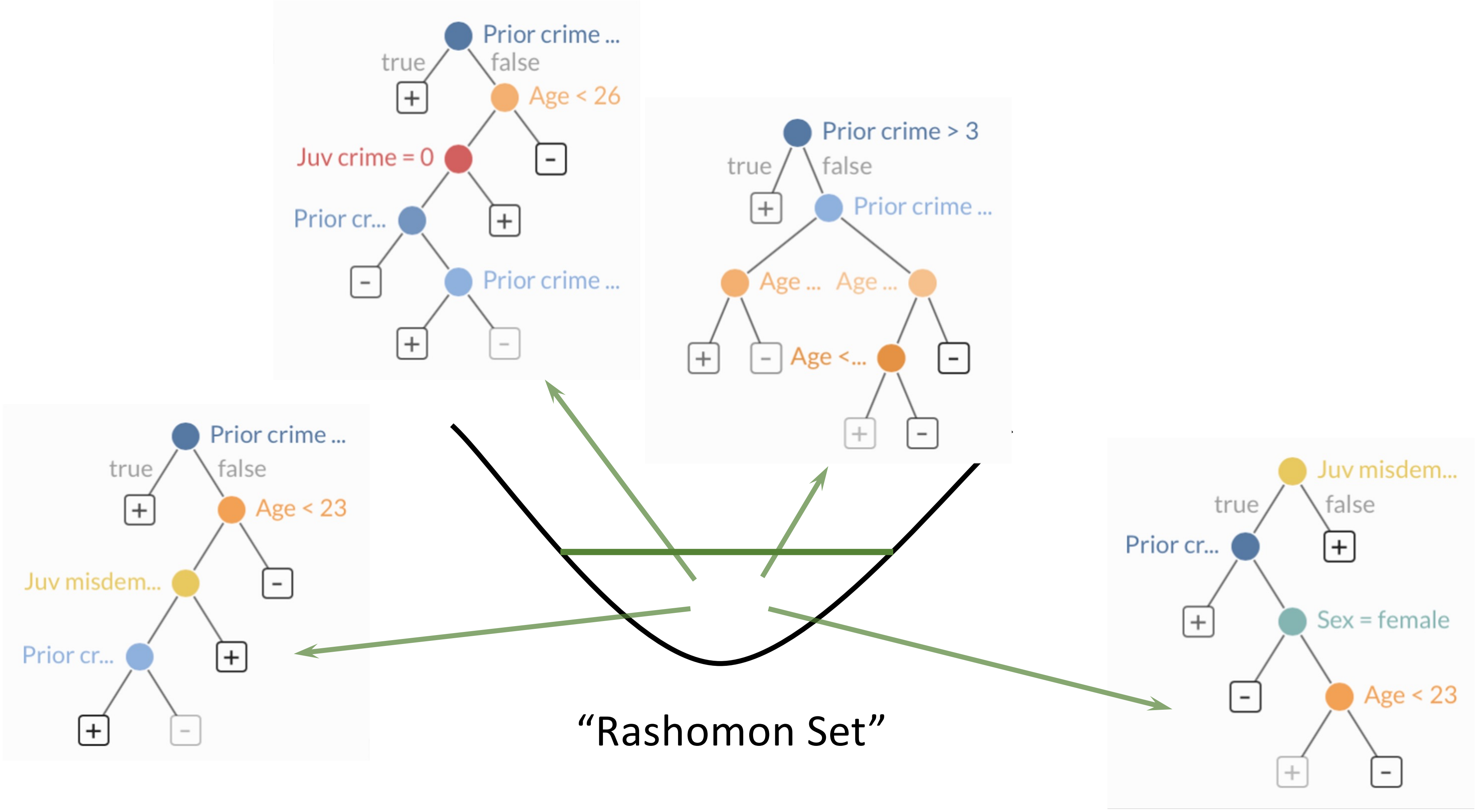
Training Set ⟶ Algorithm ⟶ Many Predictive Models

achieve low loss on training set



Obj = Misclassification Error + λ(#leaves)




Prior crime ...
true
false
+
Age < 26
Juv crime = 0
-
Prior cr...
+
-
Prior crime ...
+
-
Prior crime > 3
true
false
+
Prior crime ...
Age ...
Age ...
+
-
Age <...
-
+
-
Prior crime ...
true
false
+
Age < 23
Juv misdem...
-
Prior cr...
+
+
-
Juv misdem...
true
false
Prior cr...
+
+
Sex = female
-
Age < 23
+
-
“Rashomon Set”

TreeFARMS

Trees FAst RashoMon Sets

- Finds all optimal and almost-optimal sparse decision trees.
- Let users choose between trees.

Ingredients:

- Dynamic programming formulation
- Theorems that reduce the search space
- The model set representation: data structure for efficiently storing and evaluating lots of trees.

Do other methods produce all almost-optimal models?

They do not.

In 46 seconds on Monk2…

BART	3 / 488 unique trees
Random Forest	0 / 20,731 unique trees
CART+sampling	7 / 20,398 unique trees
TreeFARMS	105,782,431 / 105,782,431 unique trees

Applications:

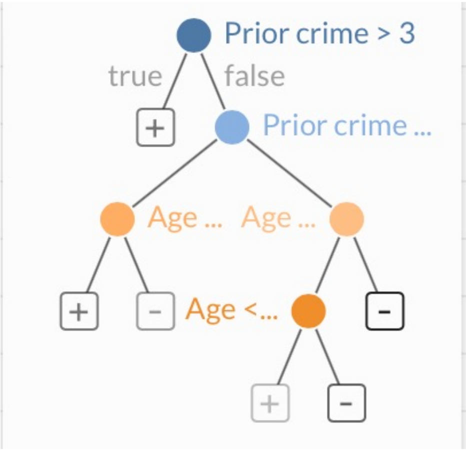
- Model-free Variable importance analysis
- Rashomon sets for other metrics
- Robustness to removal of data

Applications:

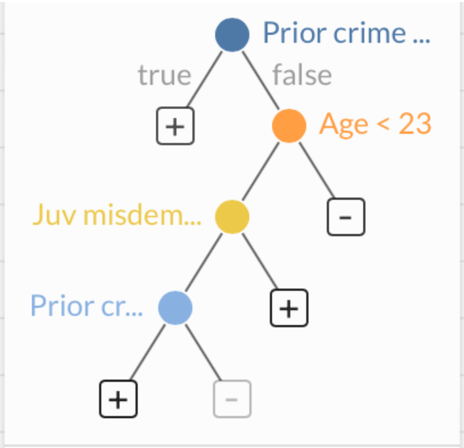
- Model-free Variable importance analysis
- Rashomon sets for other metrics
- Robustness to removal of data

Is variable *v* important to *all* good models?

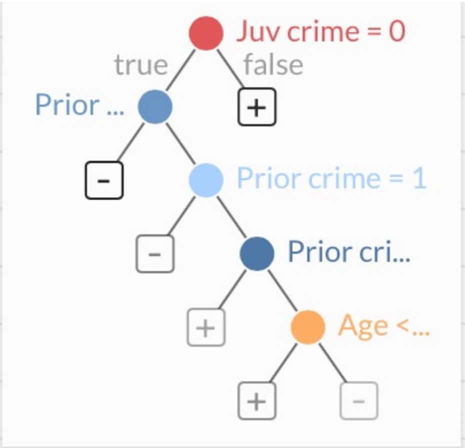
Is variable *v* important to *none* of the good models?



doesn’t depend on
variable at all



depends on
variable a lot

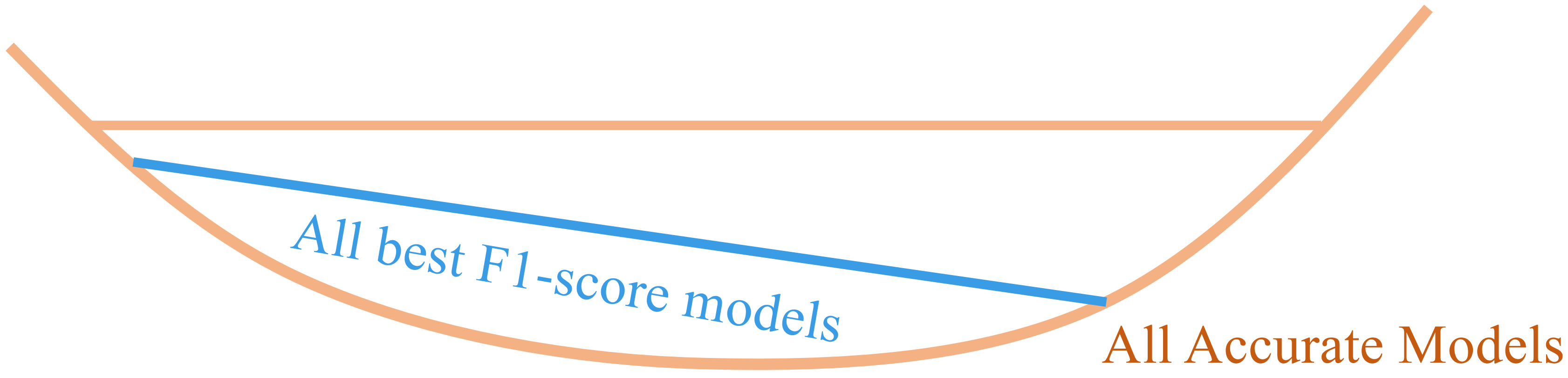


depends on
variable a little bit

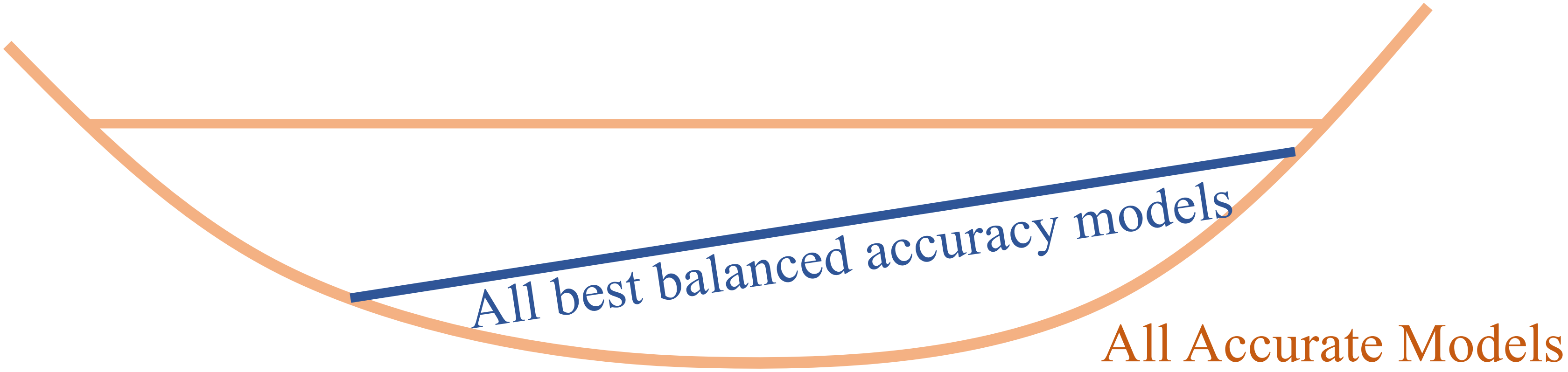
Applications:

- Model-free Variable importance analysis
- Rashomon sets for other metrics
- Robustness to removal of data

- The set of almost-optimal accurate models includes:
  - all almost-optimal F1-score models



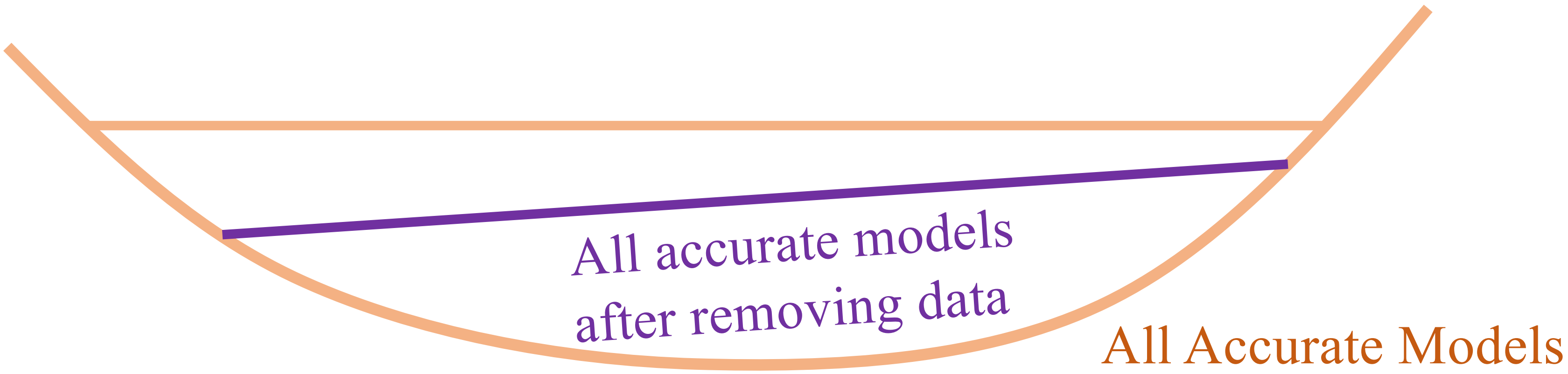
- The set of almost-optimal accurate models includes:
  - all almost-optimal F1-score models
  - all almost-optimal balanced accuracy models

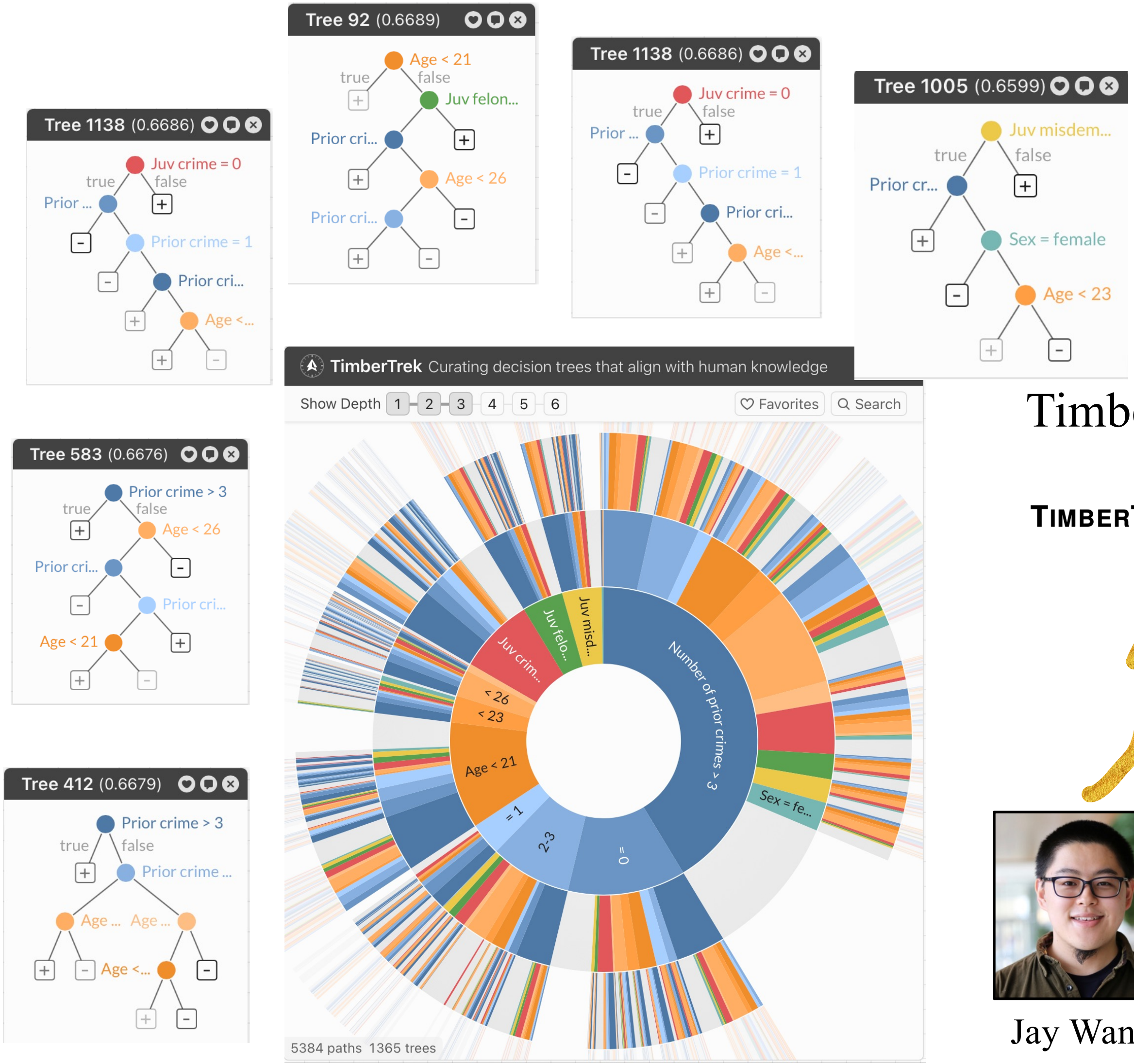


Applications:

- Model-free Variable importance analysis
- Rashomon sets for other metrics
- Robustness to removal of data

- The set of almost-optimal accurate models is robust to removal of some data points.





TimberTrek is an interface for TreeFARMS

TIMBERTREK: Exploring and Curating Sparse Decision Trees with Interactive Visualization

Zijie J. Wang[1] Chudi Zhong[2] Rui Xin[2] Takuya Takagi[3] Zhi Chen[2]
Duen Horng Chau[1] Cynthia Rudin[2] Margo Seltzer[4]

Jay Wang

bit.ly/timbertrek

Exploring the Whole Rashomon Set of Sparse Decision Trees

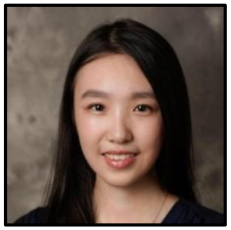
Rui Xin[1]* **Chudi Zhong**[1]* **Zhi Chen**[1]*

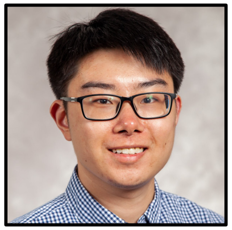
Takuya Takagi[2] **Margo Seltzer**[3] **Cynthia Rudin**[1]

[1] Duke University [2] Fujitsu Laboratories Ltd. [3] The University of British Columbia
{rui.xin926, chudi.zhong, zhi.chen1}@duke.edu
takagi.takuya@fujitsu.com, mseltzer@cs.ubc.ca, cynthia@cs.duke.edu

Paper: https://arxiv.org/abs/2209.08040
Code: https://github.com/ubc-systopia/treeFarms

Rui Xin


Chudi Zhong


Zhi Chen

Takuya Takagi

Margo Seltzer

Cynthia Rudin