# **MDI+**: A Flexible Random Forest Feature Importance Framework

**Tiffany Tang**

June 18, 2023

Joint with:

Abhi Agarwal

Ana Kenney

Yan Shuo Tan

Bin Yu

# Why Random Forests?

+ A **powerful, nonparametric prediction algorithm**, which often outperforms deep learning on moderate-sized tabular datasets

> “ … the method that performs consistently well across all dimensions is **random forests**, ” followed by neural nets, boosted trees, and SVMs. [11 datasets]
>
> - Caruana, Karampatziakis, Yessenalina (2008)

> “ The classifiers most likely to be the bests are the **random forest** versions. ” [121 data sets, 179 models]
>
> - Fernandez-Delgado, Cernadas, Barro, Amorim (2014)

> “ *Why do tree-based models still outperform deep learning on tabular data?*
> … tree-based models [i.e., **random forests**, XGBoost] remain state-of-the-art on medium-sized data (~10K samples) even without accounting for their superior speed. [45 data sets] ”
>
> - Grinsztajn, Oyallon, Varoquaux (2022)

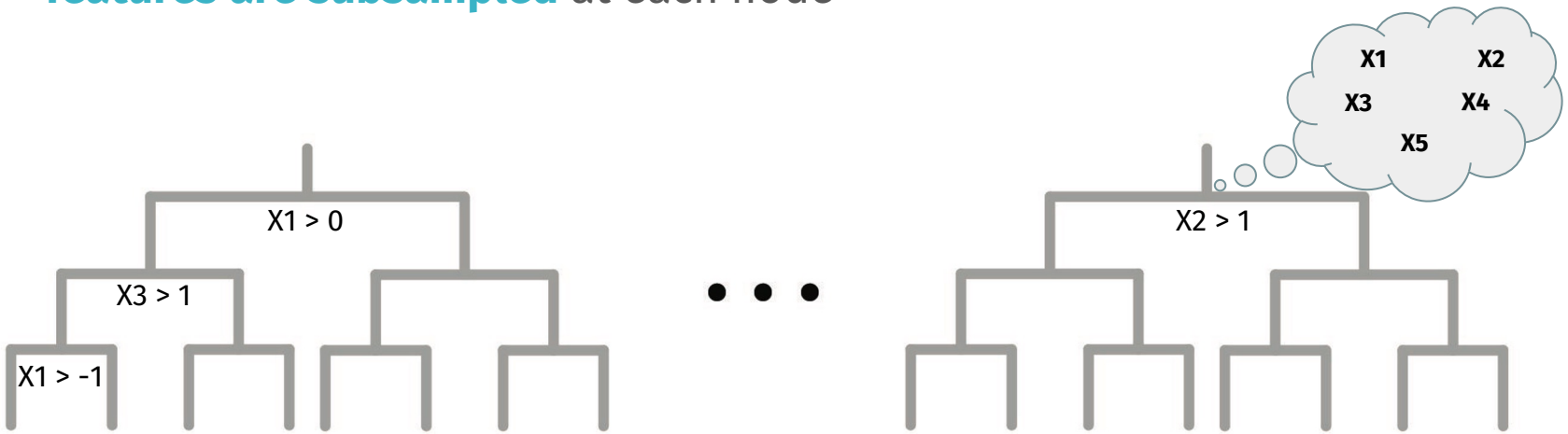
# Why Random Forests?

- A **powerful, nonparametric prediction algorithm**, which often outperforms deep learning on moderate-sized tabular datasets
- Numerous feature importance measures exist to enable **interpretability**
    - **Mean Decrease in Impurity (MDI):** most popular in practice (and default feature importance in sklearn)

# Random Forest (RF)

A **collection of decision trees**, where

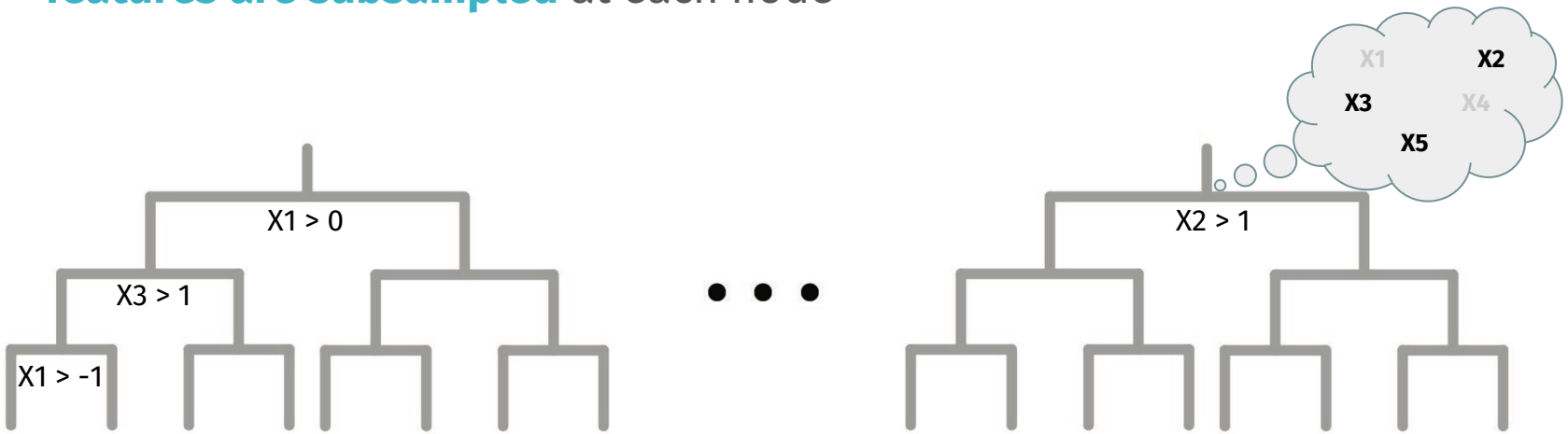
- each tree is fitted on a different **bootstrap** version of the data
- **features are subsampled** at each node



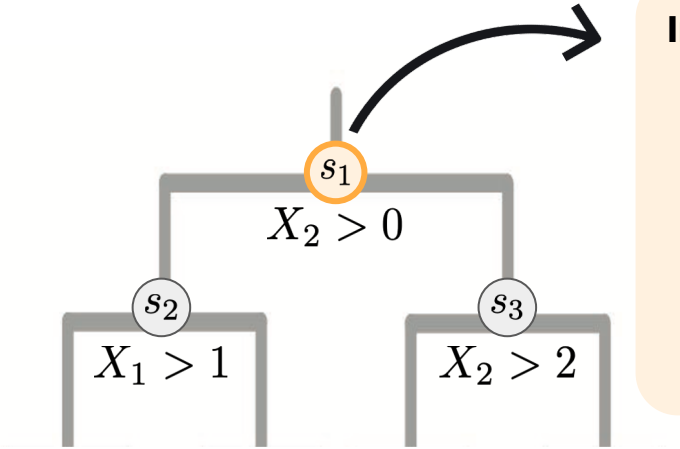
# Random Forest (RF)

A **collection of decision trees**, where

- each tree is fitted on a different **bootstrap** version of the data
- **features are subsampled** at each node



# Mean Decrease in Impurity (MDI)

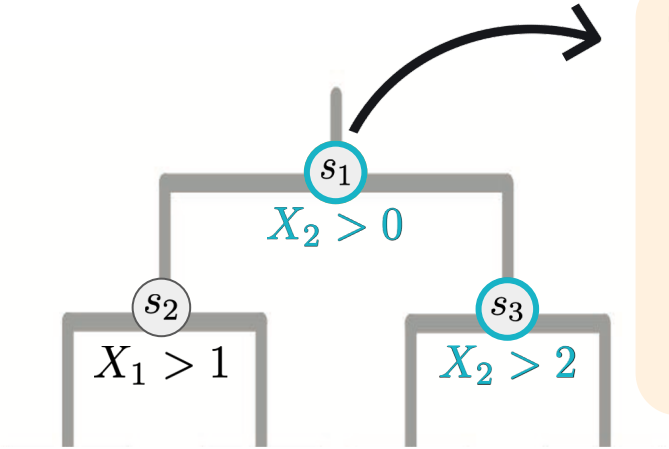



**Impurity decrease** at $s_1$

"Measures decrease in variance from making the split"

$$\hat{\Delta}(s_1) = \underbrace{\sum_{\mathbf{x} \in s_1} (y_i - \bar{y}_{s_1})^2}_{\text{Var(node of interest)}} - \underbrace{\sum_{\mathbf{x} \in s_2} (y_i - \bar{y}_{s_2})^2}_{\text{Var(left child node)}} - \underbrace{\sum_{\mathbf{x} \in s_3} (y_i - \bar{y}_{s_3})^2}_{\text{Var(right child node)}}$$

# Mean Decrease in Impurity (MDI)




**Impurity decrease** at $s_1$

"Measures decrease in variance from making the split"

$$\hat{\Delta}(s_1) = \underbrace{\sum_{\mathbf{x} \in s_1} (y_i - \bar{y}_{s_1})^2}_{\text{Var(node of interest)}} - \underbrace{\sum_{\mathbf{x} \in s_2} (y_i - \bar{y}_{s_2})^2}_{\text{Var(left child node)}} - \underbrace{\sum_{\mathbf{x} \in s_3} (y_i - \bar{y}_{s_3})^2}_{\text{Var(right child node)}}$$

For each feature k, MDI(k) is the weighted sum of impurity decreases across nodes that split on $X_k$, e.g.,

$$MDI(X_2) = \frac{n_1}{n}\hat{\Delta}(s_1) + \frac{n_3}{n}\hat{\Delta}(s_3)$$

# Mean Decrease in Impurity (MDI)

## Advantages of MDI:

Conceptually simple

Fast to compute

## Well-known drawbacks of MDI:

Unstable in **low-signal** problems

Biased against features are highly **correlated** or have low **entropy**

Inefficient measure if **additive structure** is present

Nicodemus, K. K. and Malley, J. D. “Predictor correlation impacts machine learning algorithms: implications for genomic studies.” *Bioinformatics* (2009)
Nicodemus, K. K. “On the stability and ranking of predictors from random forest variable importance measures.” *Briefings in Bioinformatics* (2011)
Tan, Y. S., Agarwal, A., and Yu, B. “A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds.” AISTATS (2022)

# MDI+:
# A generalized mean decrease in impurity

# Overview of MDI+

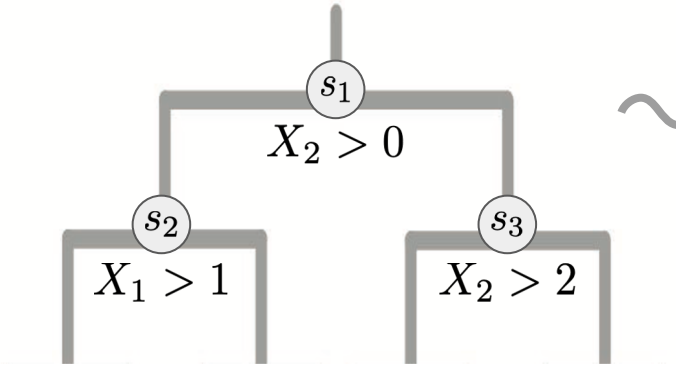
**MDI+ provides a flexible framework for computing feature importances using RFs**

- Avoids aforementioned drawbacks of MDI
- Allows the analyst to tailor the feature importance computation to the data/problem structure (e.g., handle outliers, classification vs. regression)

**Key idea:** Leverage a connection between decision trees and linear regression

# Connecting decision trees to linear regression

Step 1: Obtain engineered "stump" features $\psi(\cdot\,; s_k)$ from decision tree




node

$$\psi(\mathbf{x}; s_k) = \begin{cases} 0 & \text{if } \mathbf{x} \notin s_k \\ \frac{-N_R}{\sqrt{N_L N_R}} & \text{if } \mathbf{x} \in \text{left child of } s_k \\ \frac{N_L}{\sqrt{N_L N_R}} & \text{if } \mathbf{x} \in \text{right child of } s_k \end{cases}$$

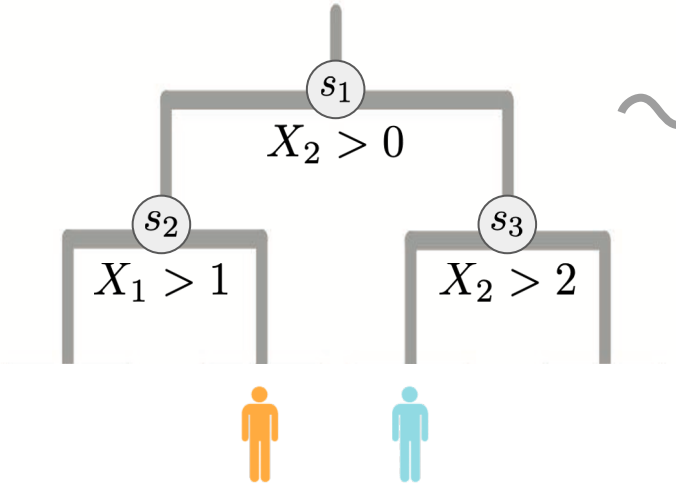
Input data $\mathbf{x} \in \mathbb{R}^p$

where $N_R =$ number of samples in right child of $s_k$

$N_L =$ number of samples in left child of $s_k$

# Connecting decision trees to linear regression

Step 1: Obtain engineered "stump" features $\psi(\cdot\,; s_k)$ from decision tree




node

$$\psi(\mathbf{x}; s_k) = \begin{cases} 0 & \text{if } \mathbf{x} \notin s_k \\ \frac{-N_R}{\sqrt{N_L N_R}} & \text{if } \mathbf{x} \in \text{left child of } s_k \\ \frac{N_L}{\sqrt{N_L N_R}} & \text{if } \mathbf{x} \in \text{right child of } s_k \end{cases}$$

Input data $\mathbf{x} \in \mathbb{R}^p$

where $N_R$ = number of samples in right child of $s_k$

$N_L$ = number of samples in left child of $s_k$

$$\Psi(\mathbf{X}; \mathcal{S}) := \begin{array}{ccc} s_1 & s_2 & s_3 \\ \hline - & + & 0 \\ + & 0 & - \\ \vdots & \vdots & \vdots \end{array}$$

**A new basis using supervised tree features**

# Connecting decision trees to linear regression

Step 2: Fit OLS on stump features

$$\mathbf{y} \sim \Psi(\mathbf{X}, \mathcal{S})$$

**Key Connection: OLS predictions = original tree predictions** [Klusowski 2021]

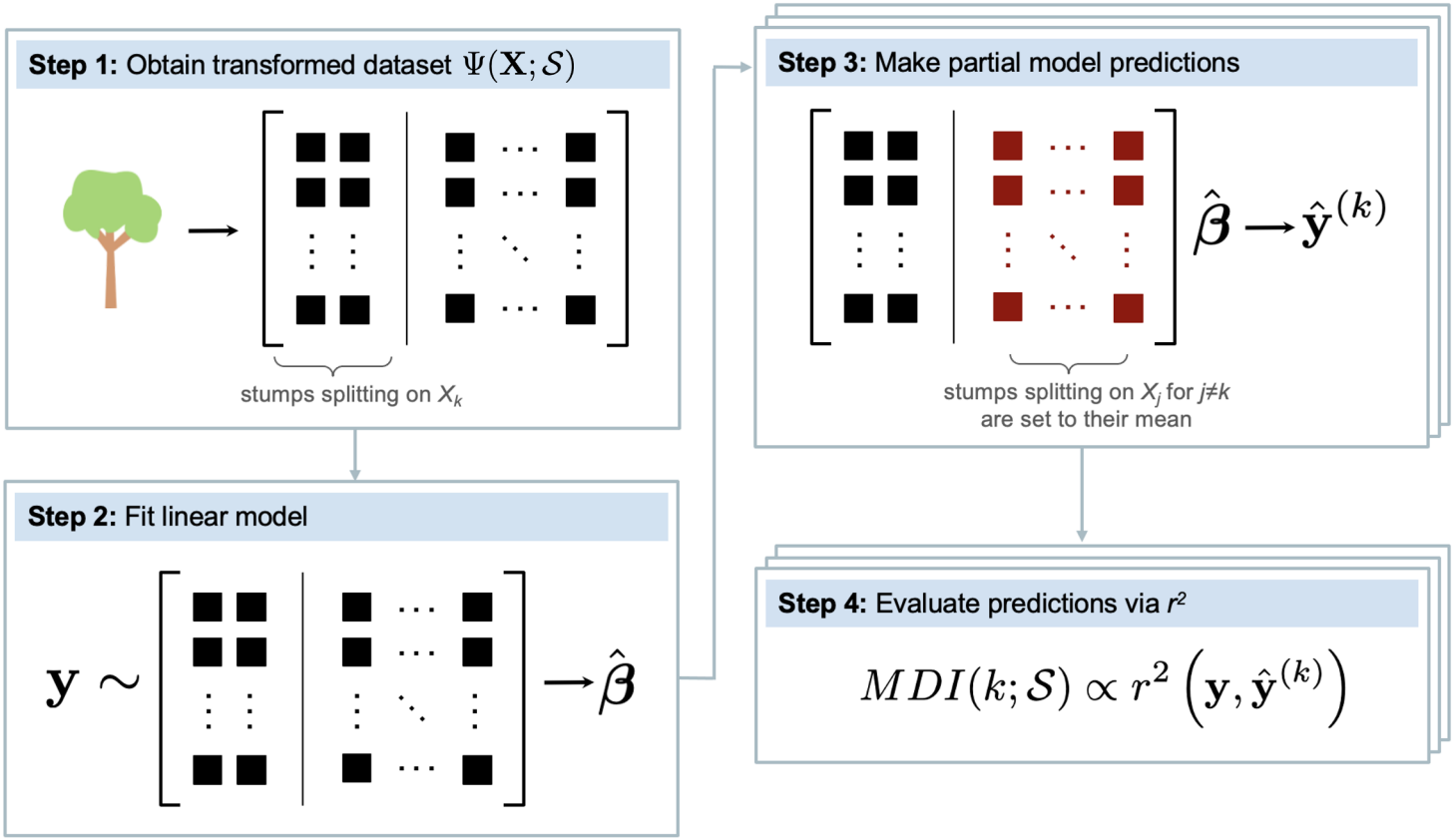
↘ assuming tree prediction = mean response per leaf node (e.g., in CART)

**Upshot: Can build upon this connection to reinterpret MDI**

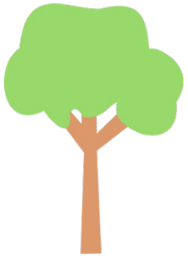


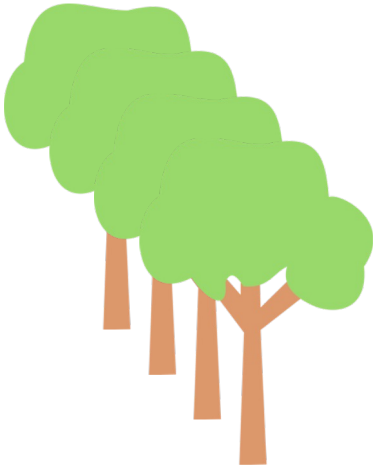
Jason M. Klusowski. "Universal consistency of decision trees in high dimensions." (2021)

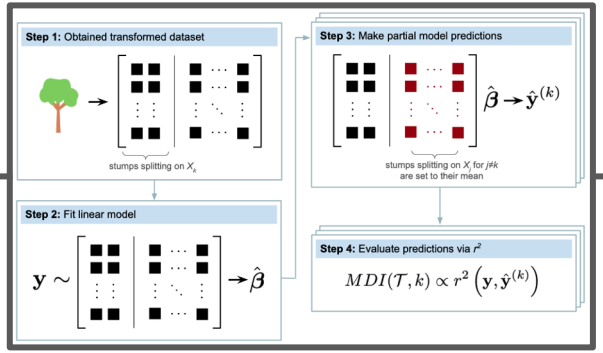
# Reinterpreting MDI



# Reinterpreting MDI



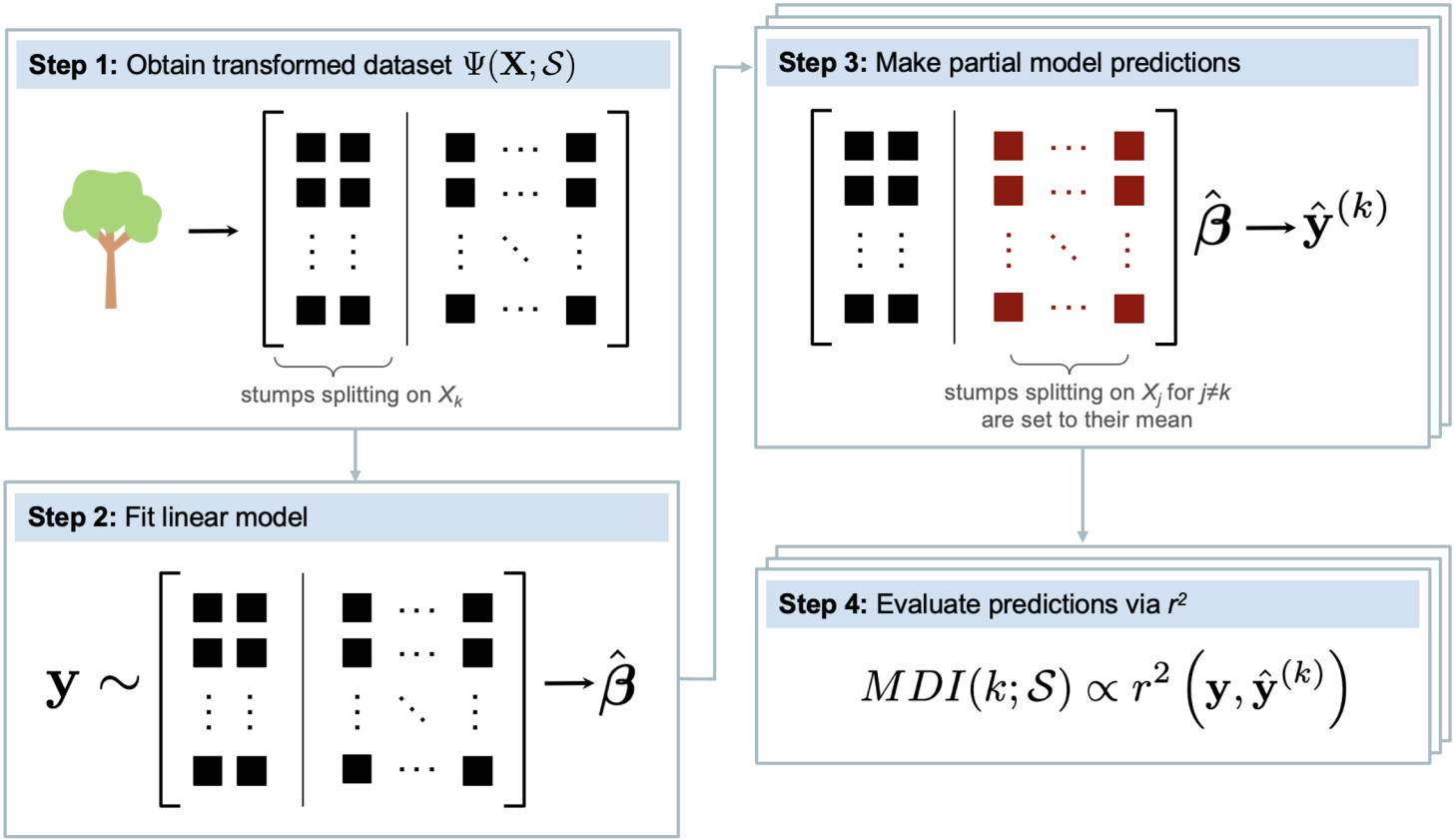





$$MDI(k; \mathcal{S})$$

$$MDI(k) = \frac{1}{n_{\text{trees}}} \sum_{i=1}^{n_{\text{trees}}} MDI(k; \mathcal{S}_i)$$

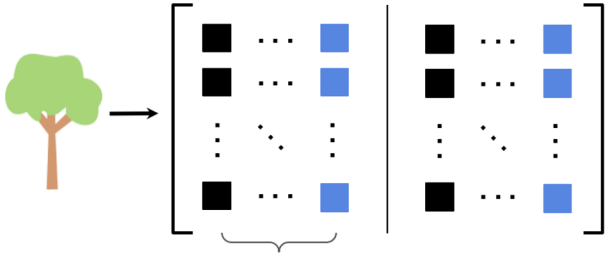
# Reinterpreting MDI



# MDI+: A Generalized Mean Decrease in Impurity

Approximate leave-one-out predictions can be computed without refitting the RF

**Step 1:** Obtain **augmented** transformed dataset



stumps splitting on $X_k$ **+ additional features derived from $X_k$ (e.g., the original feature $X_k$)**

**Step 2:** Fit ~~linear~~ **regularized GLM** model

$$GLM_\lambda\left(\mathbf{y} \sim \left[\begin{array}{ccc|ccc} \blacksquare & \cdots & \blacksquare & \blacksquare & \cdots & \blacksquare \\ \blacksquare & \cdots & \blacksquare & \blacksquare & \cdots & \blacksquare \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \blacksquare & \cdots & \blacksquare & \blacksquare & \cdots & \blacksquare \end{array}\right]\right) \rightarrow \hat{\boldsymbol{\beta}}$$

**Step 3:** Make partial model predictions **via leave-one-out**

$$GLM_\lambda\left(\left[\begin{array}{ccc|ccc} \blacksquare & \cdots & \blacksquare & \blacksquare & \cdots & \blacksquare \\ \blacksquare & \cdots & \blacksquare & \blacksquare & \cdots & \blacksquare \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \blacksquare & \cdots & \blacksquare & \blacksquare & \cdots & \blacksquare \end{array}\right] \hat{\boldsymbol{\beta}}^{LOO}\right) \rightarrow \hat{\mathbf{y}}^{(k)}$$

stumps splitting on $X_j$ for $j \neq k$ are set to their mean

**Step 4:** Evaluate predictions via ~~$r^2$~~ **metric of choice**

$$MDI^+(k;\mathcal{S}) = metric\left(\mathbf{y}, \hat{\mathbf{y}}^{(k)}\right)$$

# Results

# Roadmap of Empirical Results

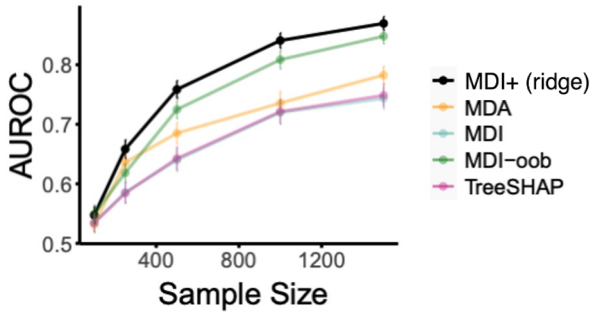
- **Correlation/entropy bias:** MDI+ overcomes correlation and entropy bias using out-of-sample prediction
- **Real data-inspired simulations:** MDI+ improves feature rankings in various regression, classification, and robust regression scenarios
  - Regression: MDI+ with ridge regression as GLM + $r^2$ metric
  - Classification: MDI+ with $l_2$-regularized logistic regression as GLM + log-loss metric
  - Robust regression: MDI+ with regularized Huber regression as GLM + Huber loss metric
- **Two real data case studies:** MDI+ identifies well-known gene predictors with greater stability than competing methods (for drug response prediction and breast cancer subtyping)

# Roadmap of Empirical Results

- **Correlation/entropy bias:** MDI+ overcomes correlation and entropy bias using out-of-sample prediction
- **Real data-inspired simulations:** MDI+ improves feature rankings in various regression, classification, and robust regression scenarios
  - Regression: MDI+ with ridge regression as GLM + $r^2$ metric
  - Classification: MDI+ with $l_2$-regularized logistic regression as GLM + log-loss metric
  - Robust regression: MDI+ with regularized Huber regression as GLM + Huber loss metric
- **Two real data case studies:** MDI+ identifies well-known gene predictors with greater stability than competing methods (for drug response prediction and breast cancer subtyping)

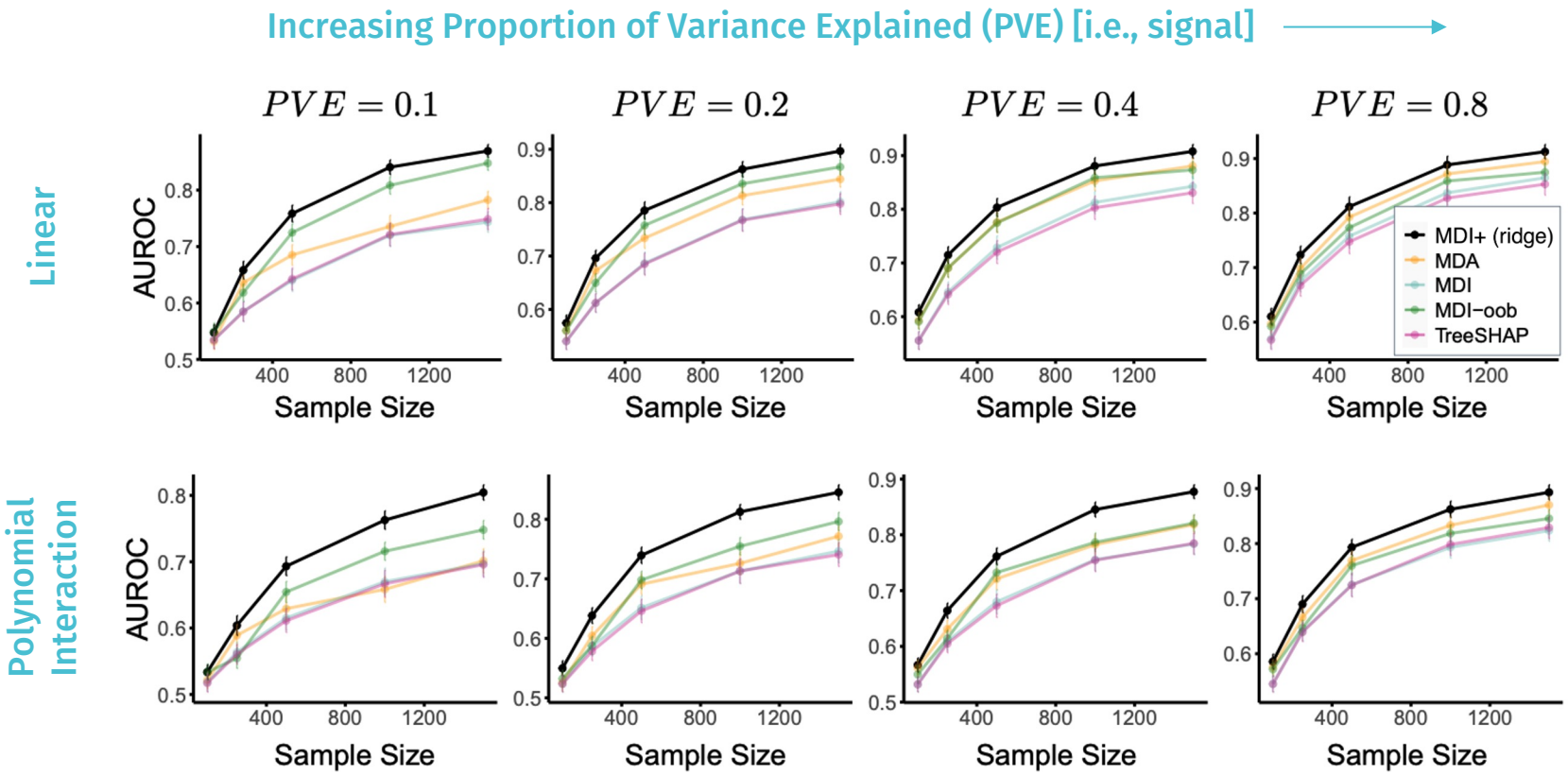
# Regression simulation results

Linear




* X = Splicing dataset

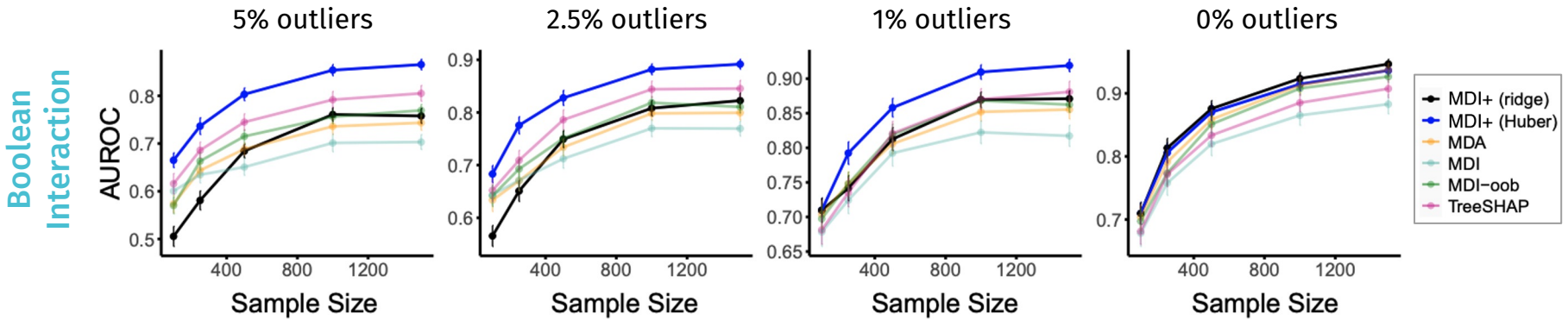
# Regression simulation results




* X = Splicing dataset

# In the presence of outliers

## Tailoring GMDI to the problem setting improves feature ranking accuracy






* X = Enhancer activity dataset

# Summary and Discussion

- MDI+ builds upon the **$r^2$ interpretation of MDI**
- MDI+ provides a **flexible framework for feature importances using RFs** that
  - Overcomes many of the inductive biases of decision trees and limitations of MDI
  - Allows the analyst to tailor the feature importance computation to the data/problem structure
- **Connection between decision trees and linear regression** opens the door to
  - A new class of prediction algorithms that leverage the tree basis/stump features
  - Possibility to build upon familiar linear regression tools (e.g., for inference)
- Code in `imodels` python package: https://github.com/csinva/imodels/tree/master
  - Notebook for example usage: https://github.com/csinva/imodels/blob/master/notebooks/mdi_plus_demo.ipynb


Thank you!