Materials Data in Action

Boltzmann Trees:

A Physically Inspired Randomization for Robust Modeling of Physical Data

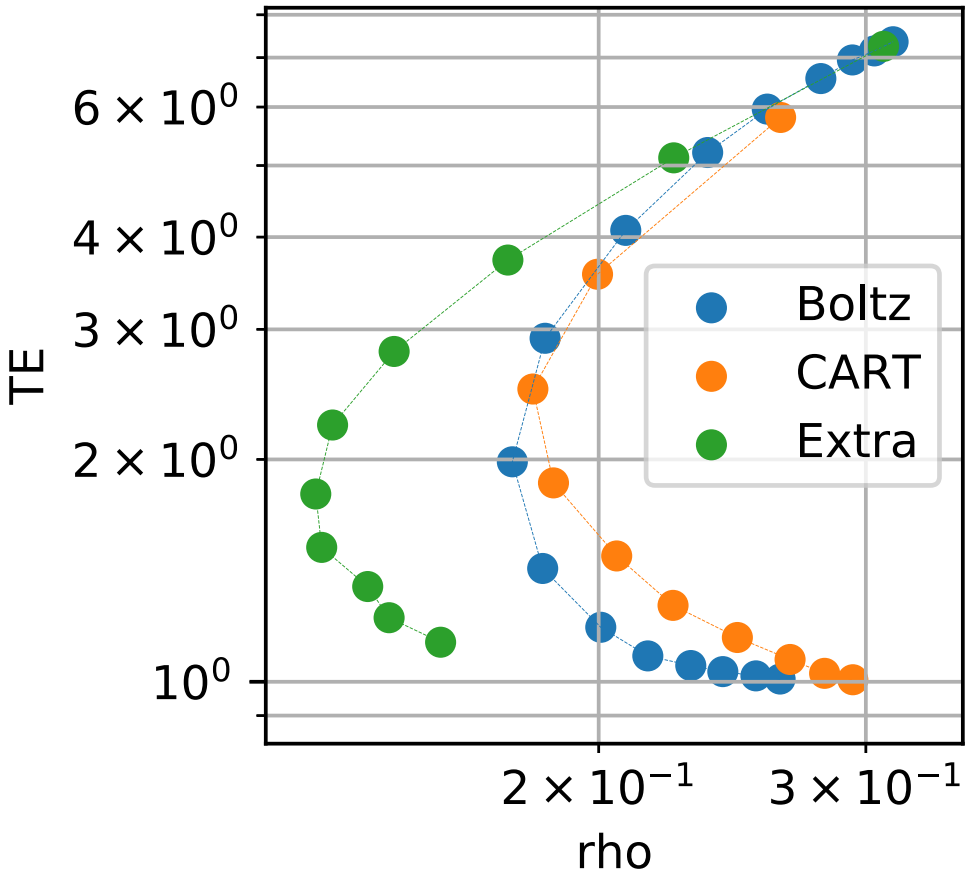
IPAM Workshop IV

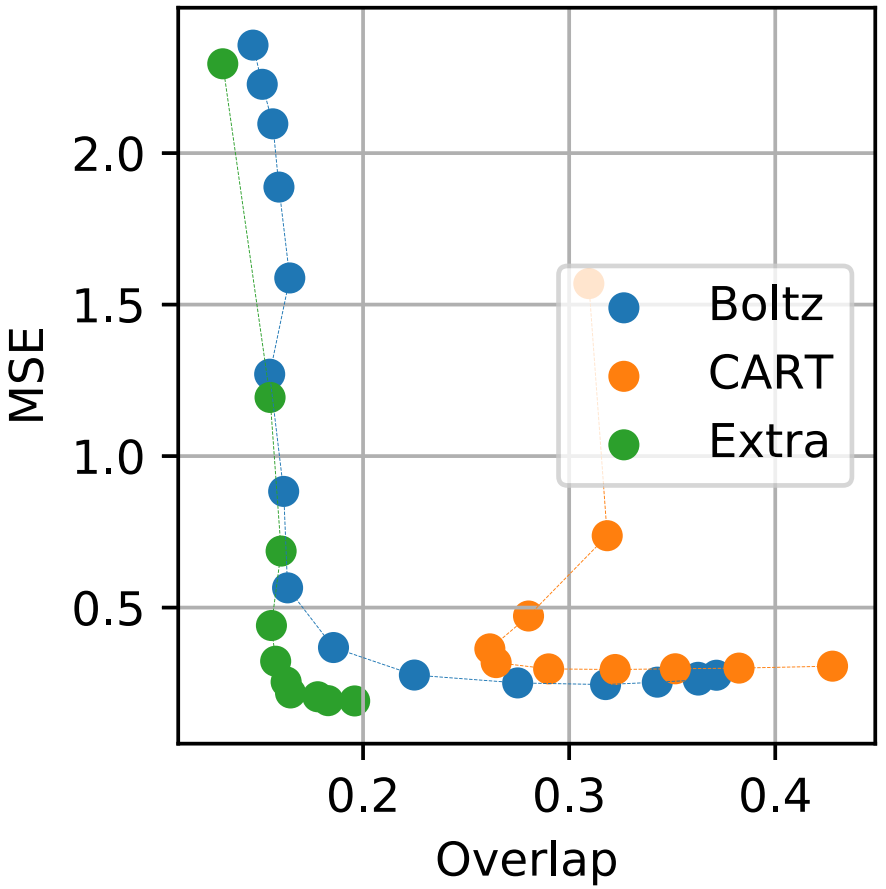
19 November 2019, UCLA

Max Hutchinson,

Scientific Software Eng.

SNEAK PEEK: TREE ERROR AND CORRELATION



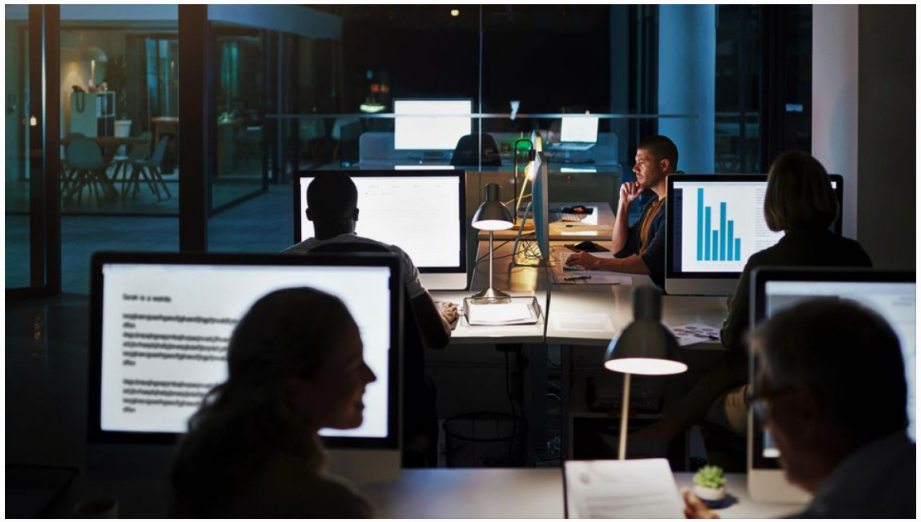


CITRINE INTRODUCTION

25 Machine Learning Startups To Watch In 2018

Louis Columbus Contributor ⓘ
Enterprise & Cloud

Forbes



ISTOCK

c&en CHEMICAL & ENGINEERING NEWS TOPICS MAGAZINE COLLECTIONS VIDEOS JOBS

POLYMERS

Lanxess teams with Citrine on AI in materials development

by Rick Mullin
MAY 18, 2019 | APPEARED IN VOLUME 97, ISSUE 20

c&en

INFORMATICS

BASF taps Citrine for artificial intelligence

by Rick Mullin
JUNE 30, 2018 | APPEARED IN VOLUME 96, ISSUE 27

3M Legal | Privacy

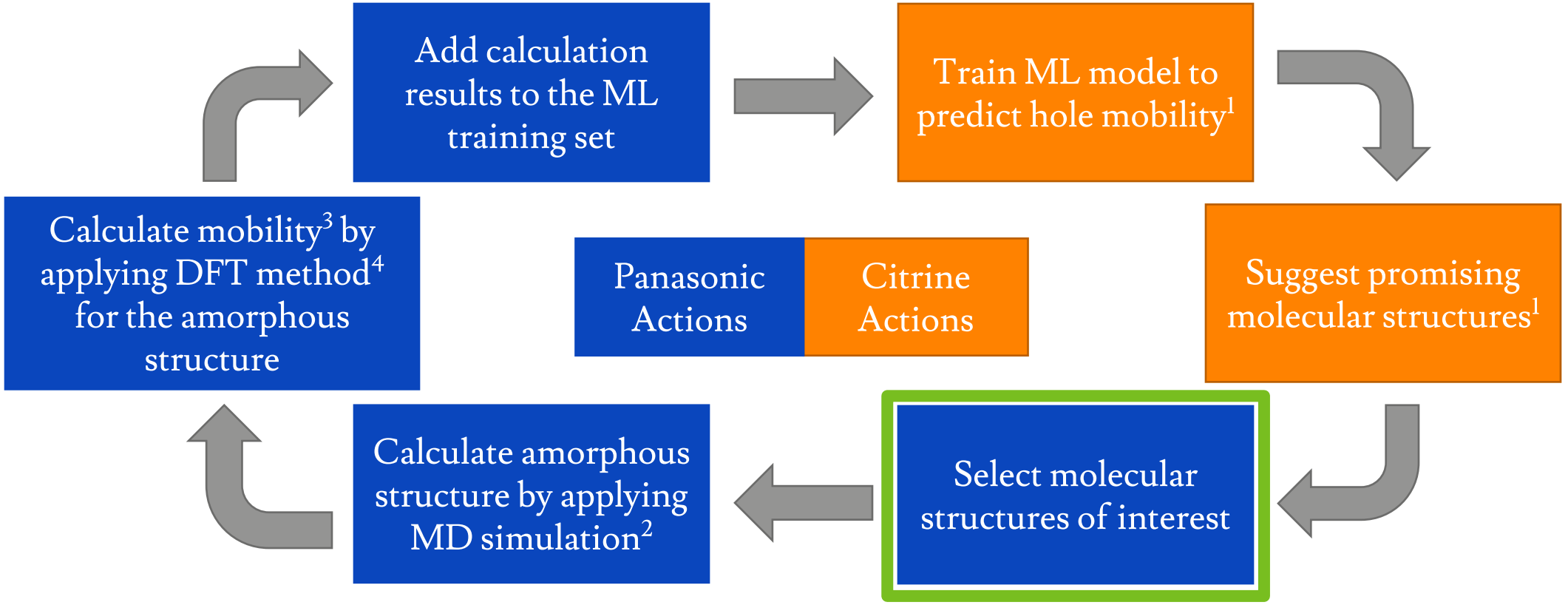
POWERED BY CITRINE INFORMATICS

CITRINE APPROACH

Let materials researchers leverage all their sources of information to make informed decisions.

— Julia Ling, CTO

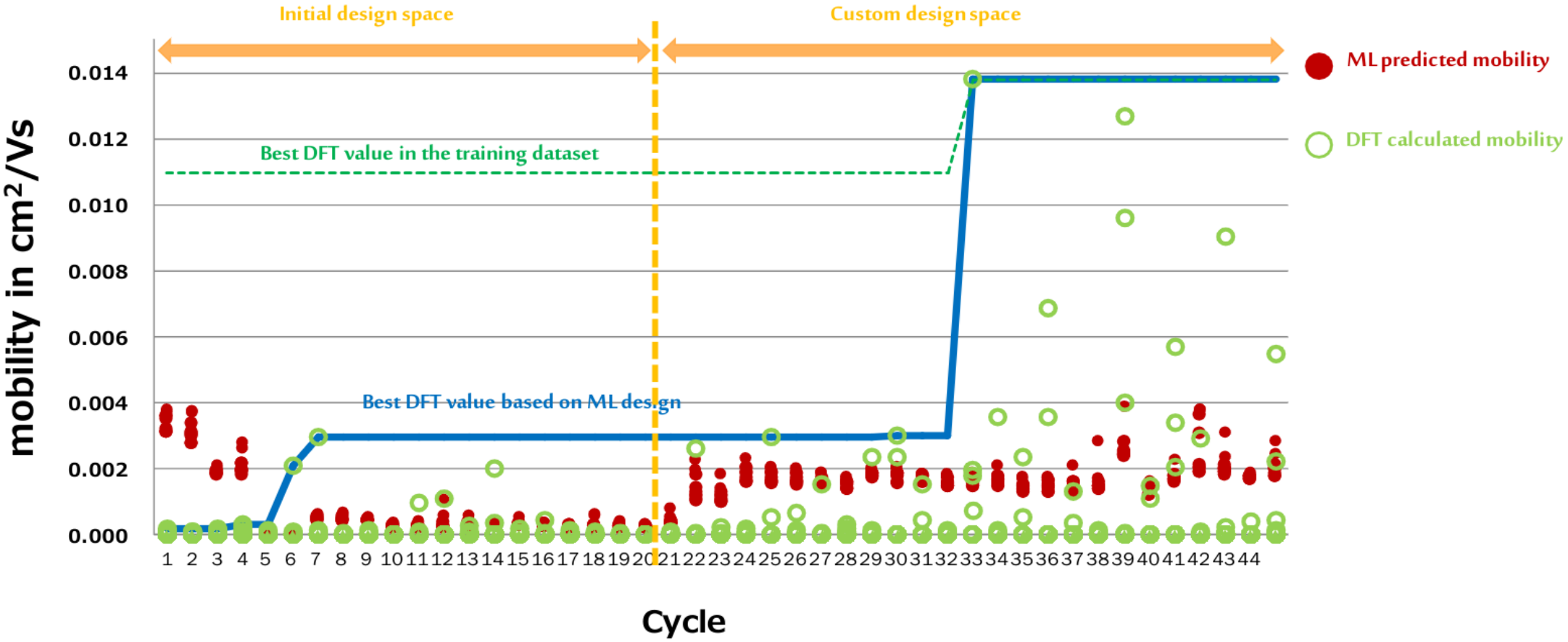
SEQUENTIAL LEARNING WORKFLOW[1]



1: Ling, J., Hutchinson, M., Antono, E., Paradiso, S., & Meredig, B. (2017). High-Dimensional Materials and Process Optimization using Data-driven Experimental Design with Well-Calibrated Uncertainty Estimates. Integr Mater Manuf Innov 6: 207.
2: Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY, 2016. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2016.
3: Evans D R, Kwak H S, Giesen D J, Goldberg A, Halls M D, Oh-e M, 2016 Estimation of charge carrier mobility in amorphous organic materials using percolation corrected random-walk model, Org. Electronics 29, 50.
4: Bochevarov A D, Harder E, Hughes T F, Greenwood J, Braden D, Rinaldo D, Halls M D, Friesner R A 2013 Jaguar: a high-performance quantum chemistry software program with strengths in life and materials sciences, Int. J. Quantum Chem. 113 2110. The software is available from Schrödinger, New York.

MATERIAL PERFORMANCE

At the 33rd cycle, the best DFT value of ML designed compounds exceeded that of compounds in the initial training dataset.

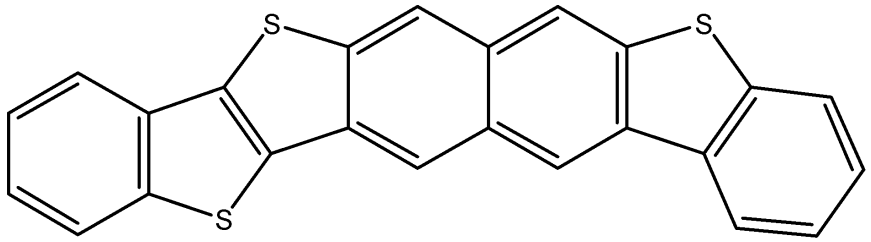


Panasonic

N Matsuzawa, E Antono, et al.

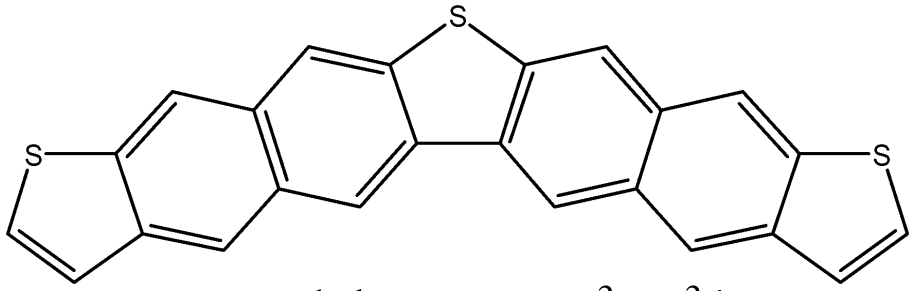
RESULTS

These four molecules were surfaced as promising candidates for an RFID development project:

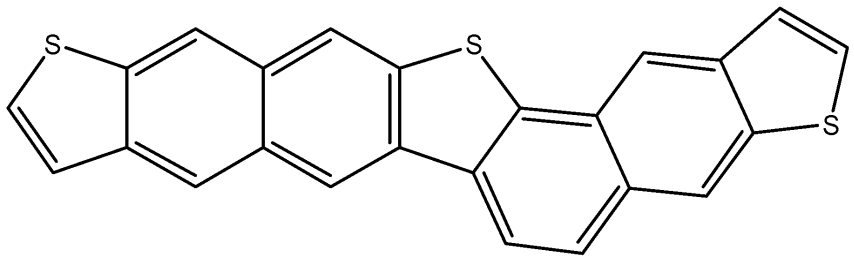


DFT mobility: 1.4×10^{-2} cm^2/Vs

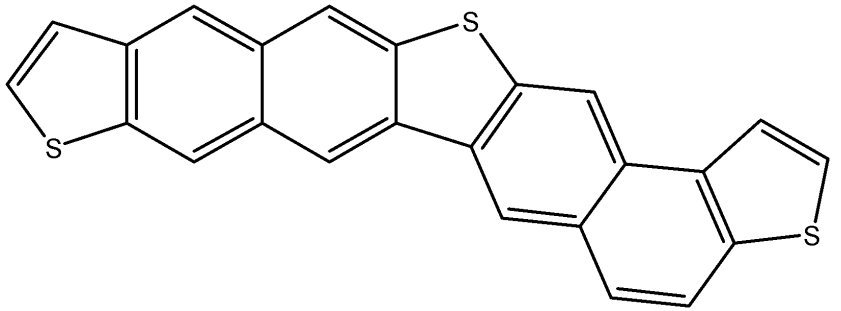
25% improvement over best training candidate



DFT mobility: 1.3×10^{-2} cm^2/Vs



DFT mobility: 9.1×10^{-3} cm^2/Vs



DFT mobility: 6.9×10^{-3} cm^2/Vs

Patents are pending for the structures above.

MATERIALS: A UNIQUE ML USE CASE

	TRADITIONAL ML APPLICATIONS	MATERIALS ML APPLICATIONS
DATA VOLUME	Big, dense (up to ~10^8 examples)	Small, sparse (~10^2 examples)
DATA REPRESENTATION	Can often be optimized by algorithms	Requires deep domain knowledge
PREDICTION TASK	Accurately pattern-match common cases	Predict unusual or "extreme" materials
ESTABLISHED DOMAIN KNOWLEDGE	Not applicable—rely on data to learn patterns	Must be physics-aware
SAMPLE BIAS	Often present	Experiments correlated, negatives stigmatized
UNCERTAINTY IN DATA AND MODELS	Usually unimportant	Always important
INTERPRETABILITY	Usually unimportant	Often required by scientists & engineers

MATERIALS: A UNIQUE ML USE CASE

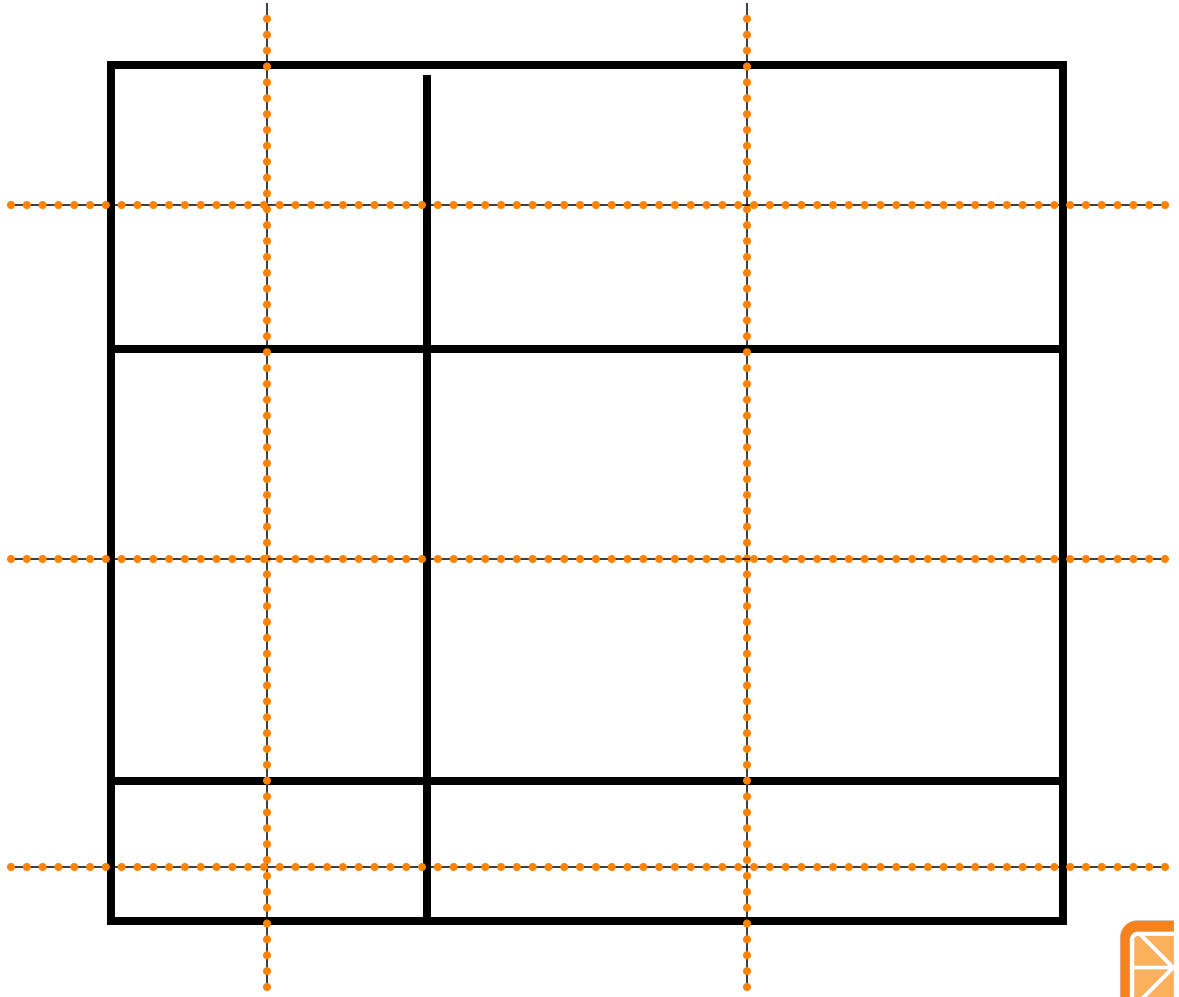
	TRADITIONAL ML APPLICATIONS	MATERIALS ML APPLICATIONS
DATA VOLUME	Big, dense (up to ~10^8 examples)	Small, sparse (~10^2 examples)
DATA REPRESENTATION	Can often be optimized by algorithms	Requires deep domain knowledge
PREDICTION TASK	Accurately pattern-match common cases	Predict unusual or "extreme" materials
ESTABLISHED DOMAIN KNOWLEDGE	Not applicable—rely on data to learn patterns	Must be physics-aware
SAMPLE BIAS	Often present	Experiments correlated, negatives stigmatized
UNCERTAINTY IN DATA AND MODELS	Usually unimportant	Always important
INTERPRETABILITY	Usually unimportant	Often required by scientists & engineers

EXAMPLE: GRIDDED DATA

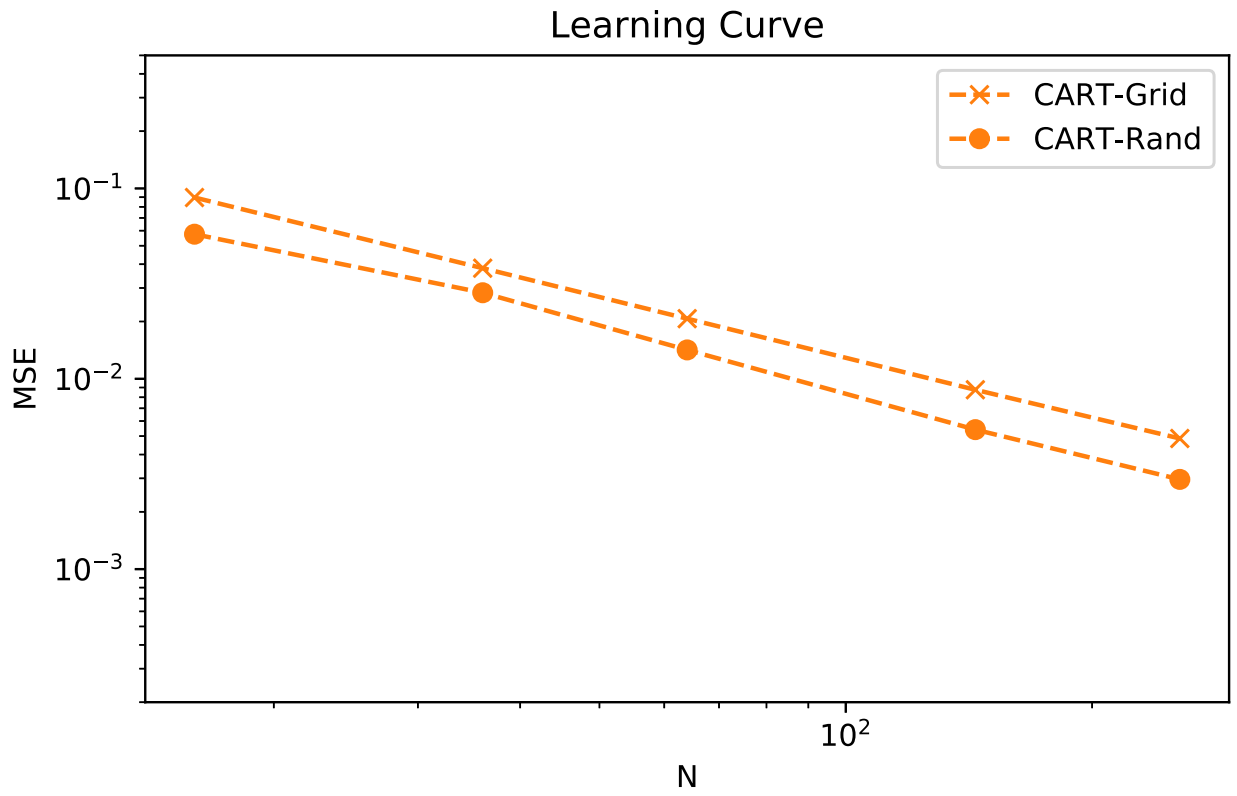
- We thought we had a 5-dimensional real continuous space
- The training data was actually on a 3x4x2x2x7x2 grid
  - Only 2 + 3 + 1 + 1 + 1 + 6 + 1 = 15 possible decisions!

On a grid, the number of likely decisions scales with a*d, but the amount of data scales with a**d

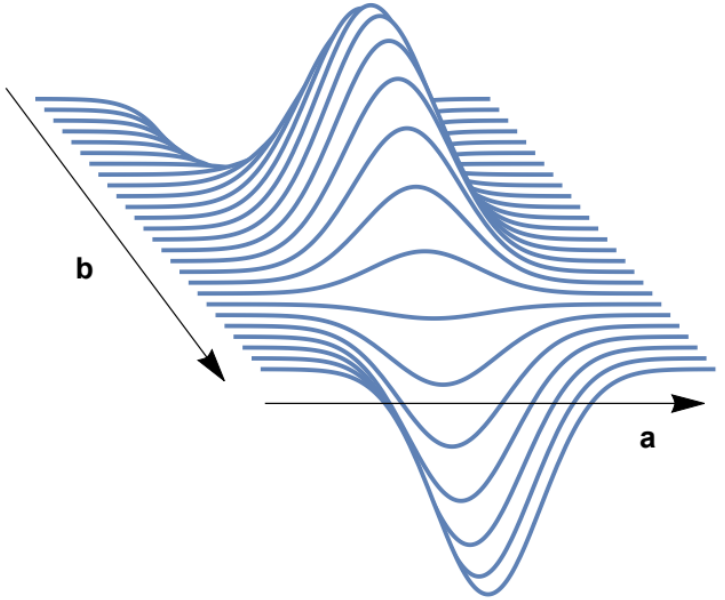
- Trees split in all the same places



EXAMPLE: GRIDDED DATA



$$f(x) = \sin(2\pi x_1)\exp[-4x_2^2]$$



RANDOM FORESTS

RANDOM FOREST ARE... ENSEMBLES

Take the average of "weak" learners, h:

$$f(x) = \frac{1}{N} \sum_i^N h(x, \Theta_i)$$

The generalization error (MSE) can be decomposed:

$$MSE(f) \leq \bar{\rho} \cdot MSE(h)$$

For a weighted correlation:

$$\bar{\rho} = \frac{E_i E_j [E_x (Y - h(x, \Theta_i))(Y - h(x, \Theta_j)))]}{E_i E_x (Y - h(x, \Theta_i))^2}$$

Breiman, L. "Random forests." *Machine learning* (2001).

RANDOM FOREST ARE... ENSEMBLES

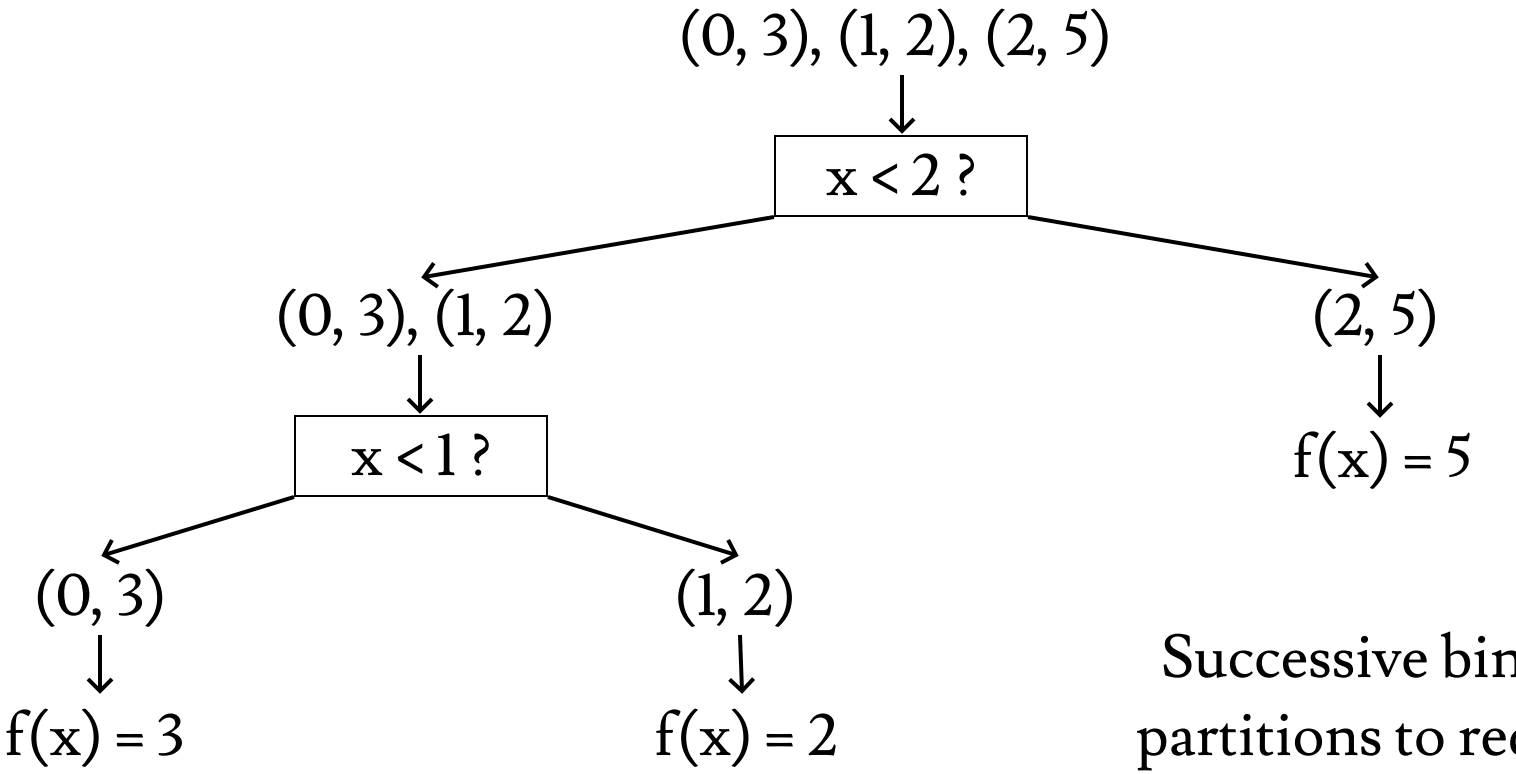
$$MSE(f) \leq \bar{\rho} \cdot MSE(h)$$

Either we can reduce $\bar{\rho}$ or we can reduce MSE(h)

"Bagging" is training the model on random data subsets, drawn with replacement

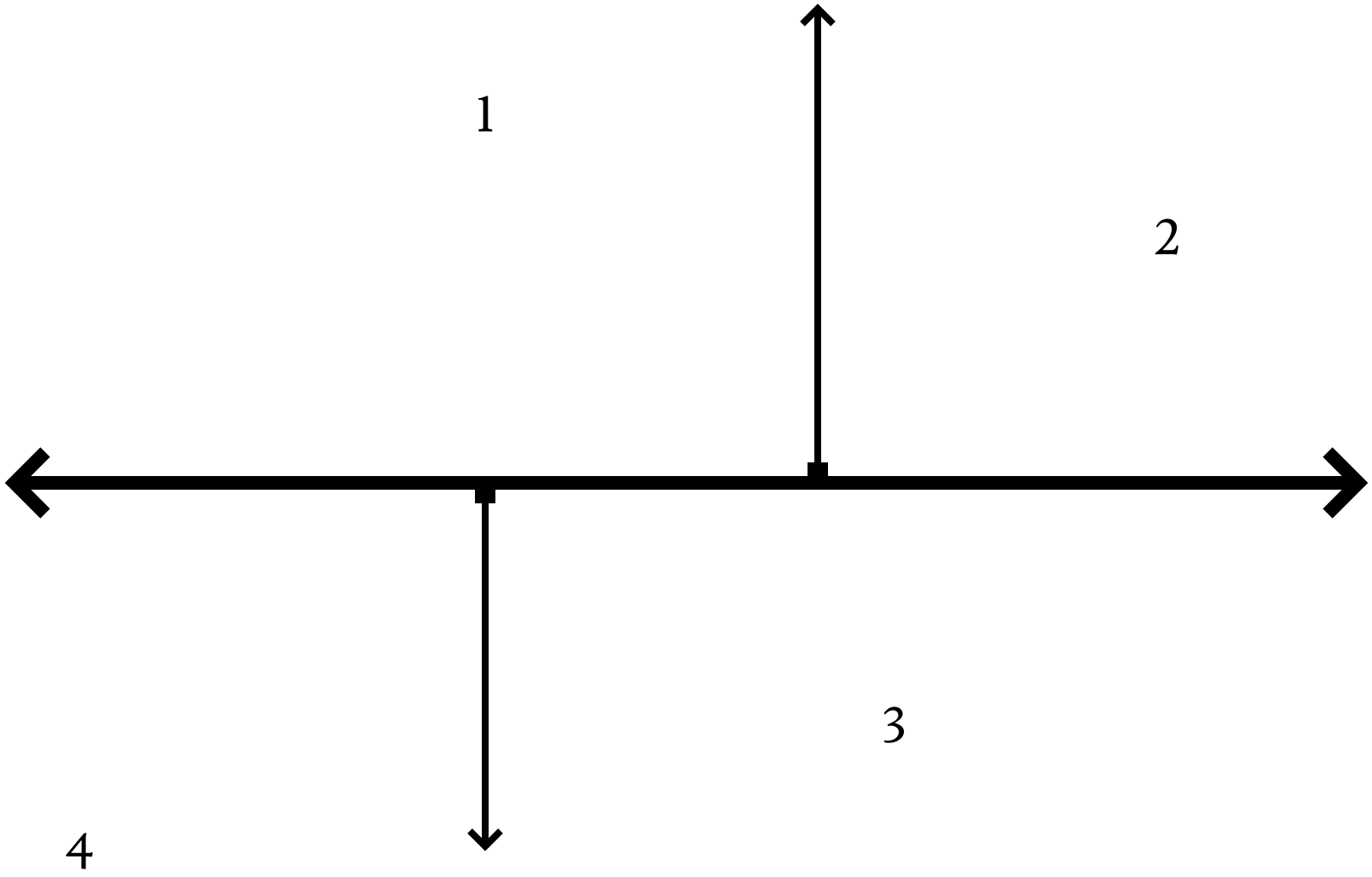
- Decreases the correlation
- Increases the single-tree error (less data)

RANDOM FOREST ARE... ENSEMBLES OF TREES



Successive binary partitions to reduce impurity (variance)

TREES ARE SIMPLE PARTITIONS



TREES ARE DEFINED BY THEIR "SPLITTER"

- "CART" splitter optimizes the reduction in "impurity"
  - in regression, "impurity" is the total variance
  - only K randomly selected features are considered for each split
  - PDF of the split is non-zero at d-points, with exp-decaying probability
- Feature randomization is typically combined with bagging

Breiman, L. "Random forests." *Machine learning* (2001).

CART

EXTREMELY RANDOMIZED TREES

Extremely randomized trees optimize the feature but not the cut-point

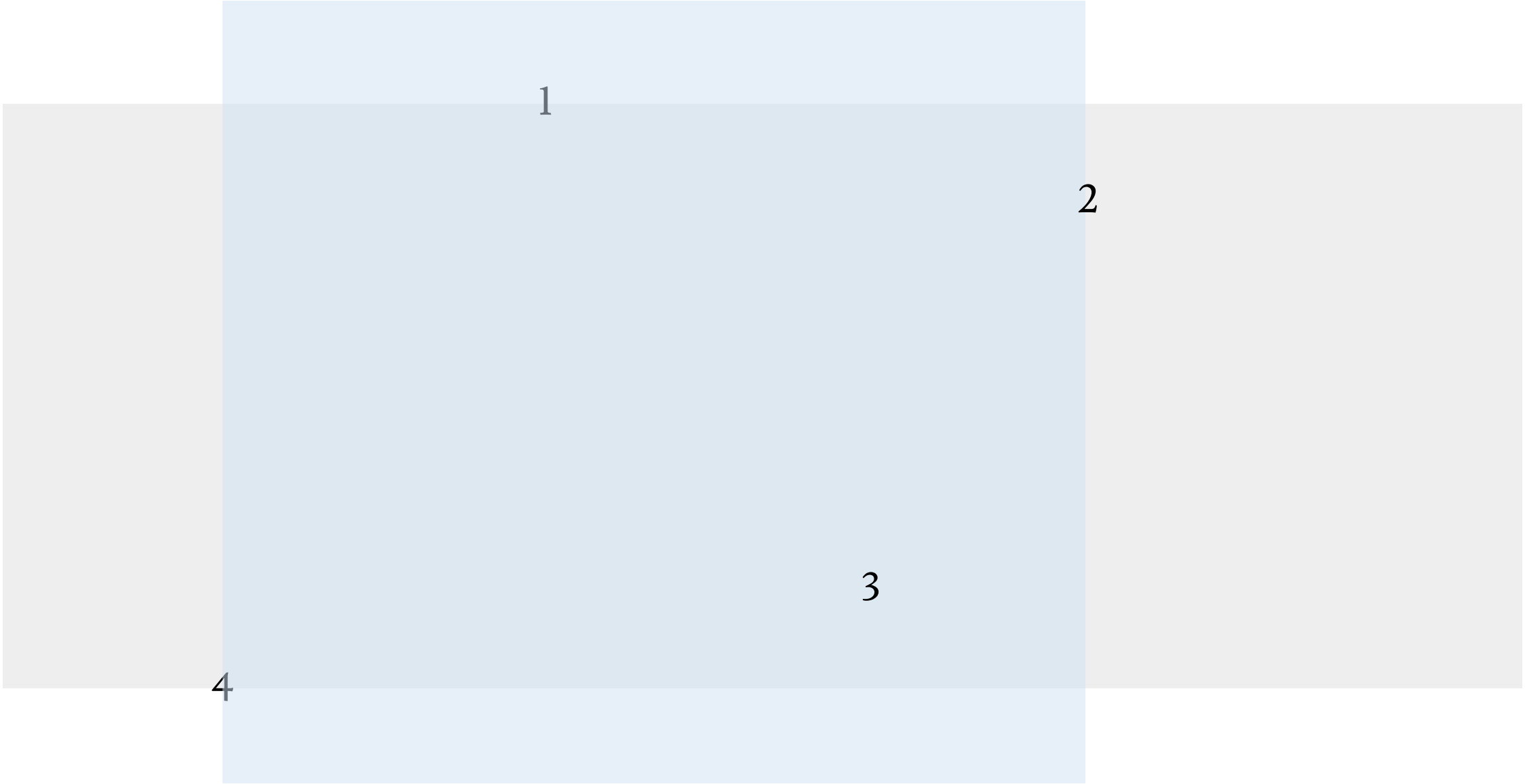
- The cut-point is chosen uniformally on the range of the feature
- only K randomly selected features are considered for each split
- PDF of the split is uniform** on each feature but exp-decaying across features

Bagging is typically **not** used.

** The coupling between the independent cut-point selection (uniform) and the optimization across the features results in complicated cut-point distributions

P. Geurts, D. Ernst, L. Wehenkel. "Extremely randomized trees." *Machine learning* (2006)

EXTREMELY RANDOMIZED TREES



BOLTZMANN TREES

(or Gibbs trees if you prefer)

BOLTZMANN TREES

Boltzmann trees split with a probability that's related to the impurity:

$$P(x,i) \sim \frac{1}{\Delta x} \exp\left[-\frac{I(x,i)}{kT}\right]$$

Splits with differences of impurity that are on the order of kT are similarly likely.

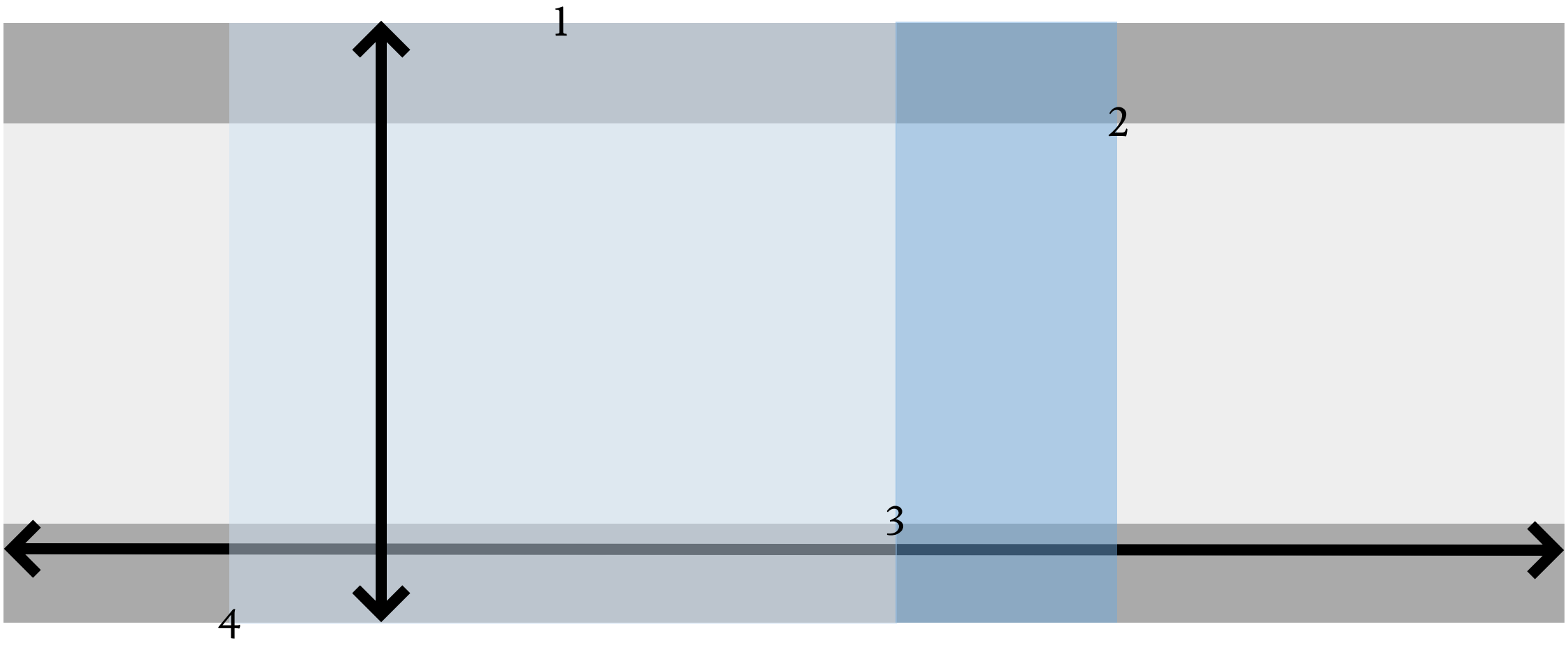
- kT controls the impurity scale, trading off tree-error for tree-correlation
- cut-points are uniformly drawn between consecutive data values

Works well when k is the impurity of the current (un-split) node

- Makes T dimension-less, encourages early randomization

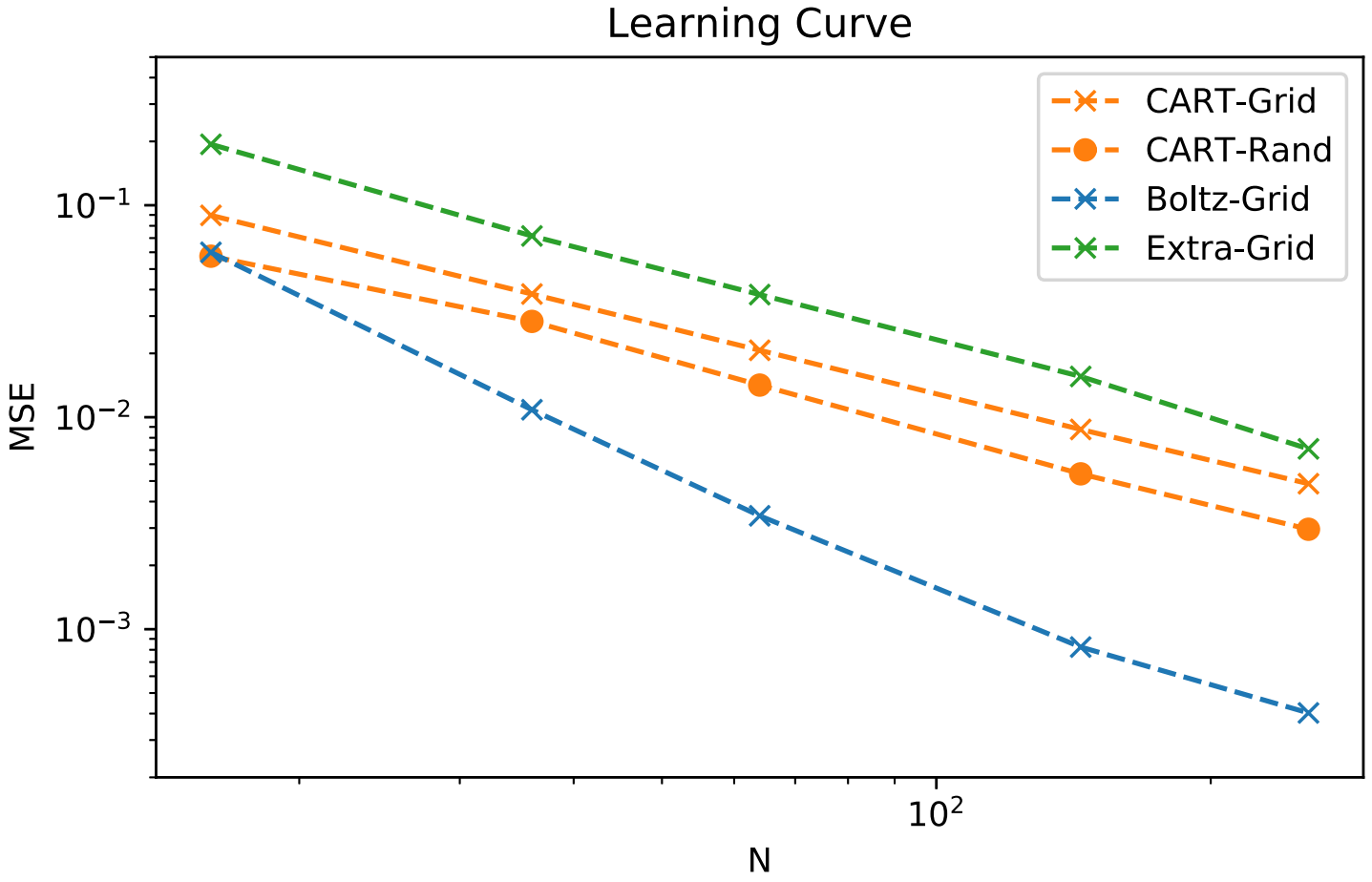
T = 0: LIKE CART, BUT UNIFORM INTER-DATA

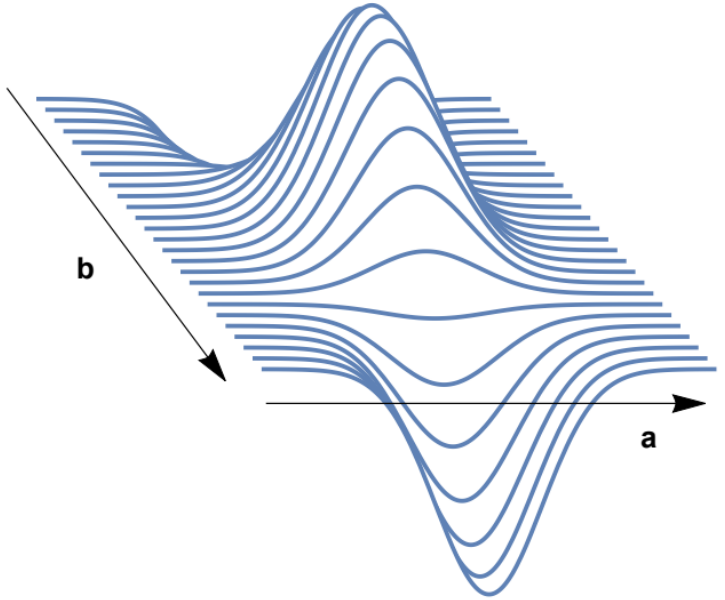
T >> 1: ~ DATA DENSITY



EXAMPLE: GRIDDED DATA

$$f(x) = \sin(2\pi x_1)\exp[-4x_2^2]$$





EMPIRICAL ANALYSIS

FRIEDMAN-SILVERMAN FUNCTION

For $x \in [0,1]^{10}$

$$f(x) = 0.1 \exp[4x_1] + \frac{4}{1+\exp[-20(x_2-0.5)]} + 3x_3 + 2x_4 + x_5$$

Gradient is anisotropic and non-uniform

$$\frac{\partial}{\partial x_2} f(x) = \frac{80 \exp[10-20x]}{(\exp[10-20x]+1)^2}$$

Friedman, Jerome H., and Bernard W. Silverman. "Flexible parsimonious smoothing and additive modeling." Technometrics 31, no. 1 (1989): 3-21.

REFRESHER

	CART	Boltzmann	Ex. Random
Bagging	X	X	
Feature Subset	X		X
Temperature		X	

REFRESHER

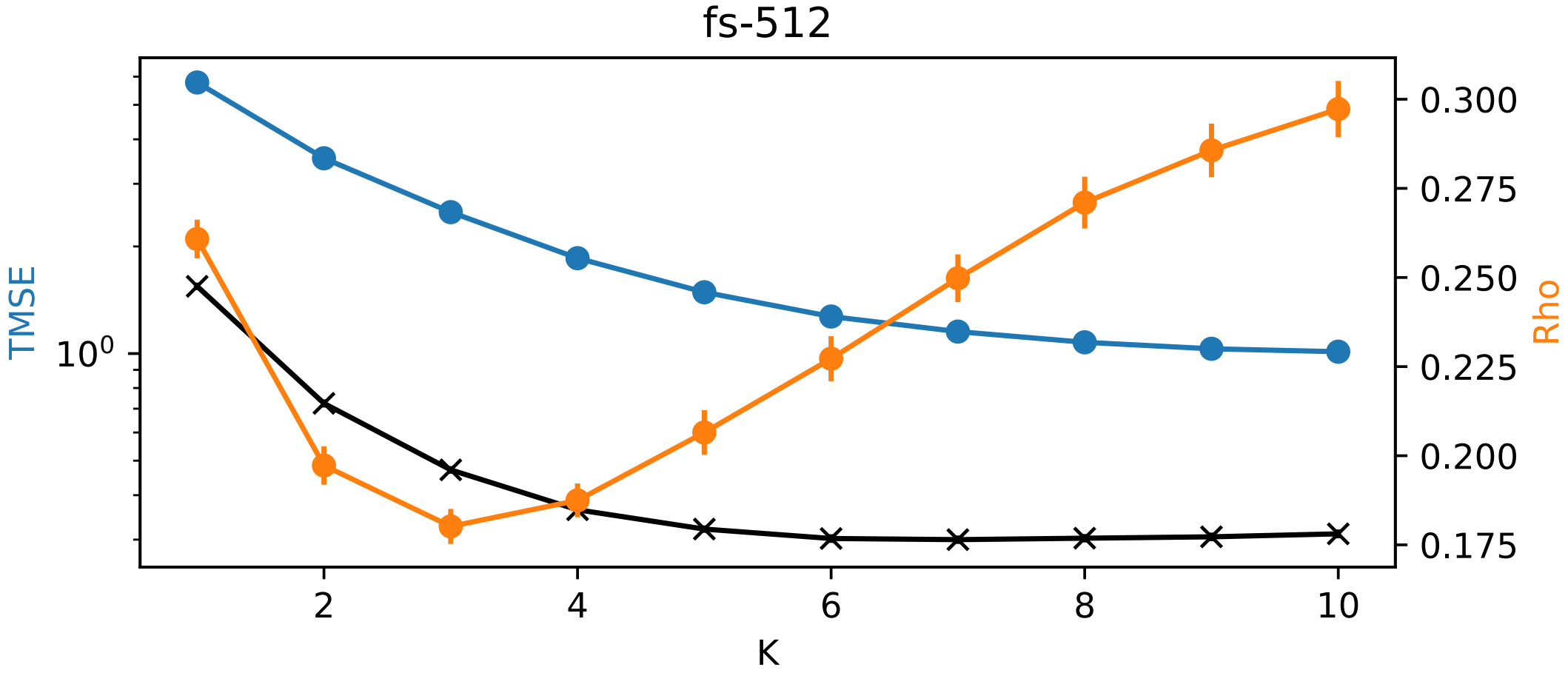
The generalization error (MSE) can be decomposed:

$$MSE(f) \leq \bar{\rho} \cdot MSE(h)$$

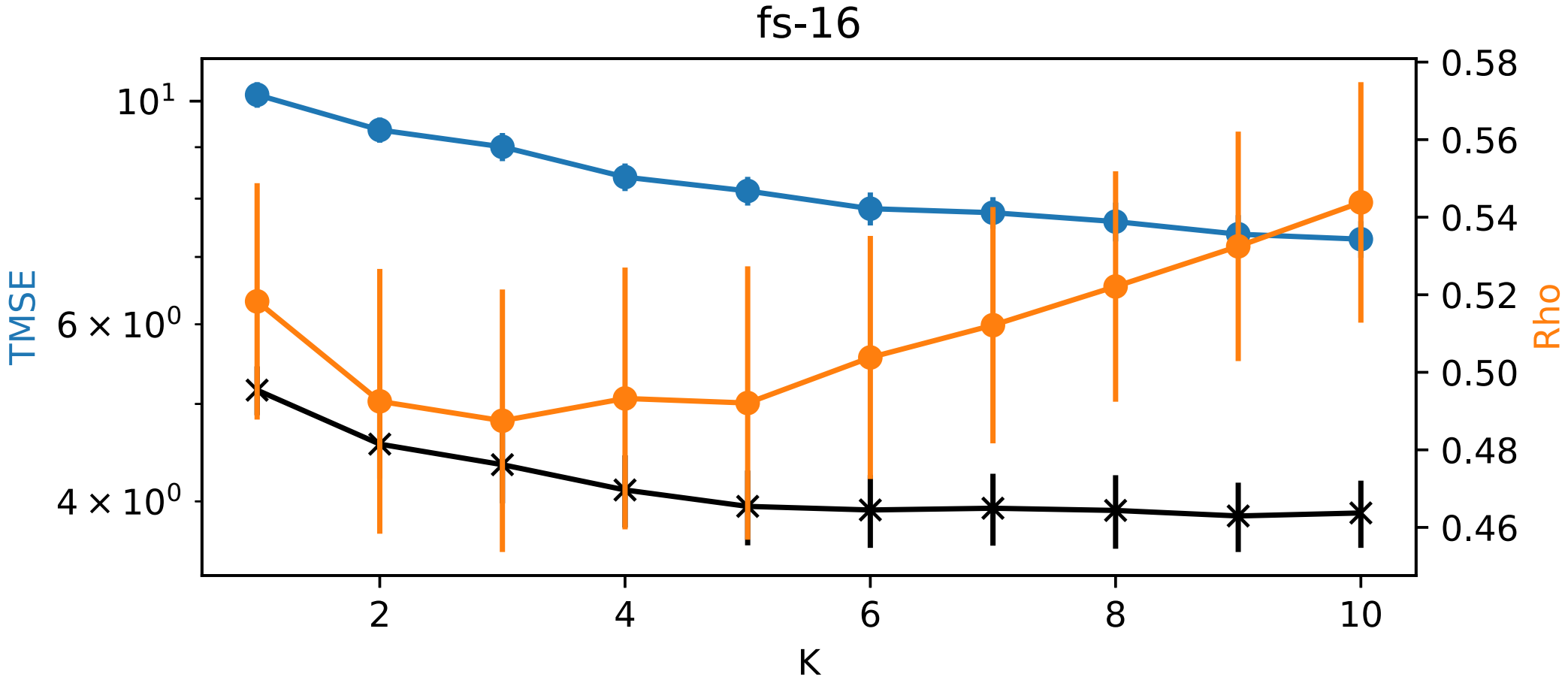
For a weighted correlation:

$$\bar{\rho} = \frac{E_i E_j [E_x (Y - h(x, \Theta_i))(Y - h(x, \Theta_j)))]}{E_i E_x (Y - h(x, \Theta_i))^2}$$

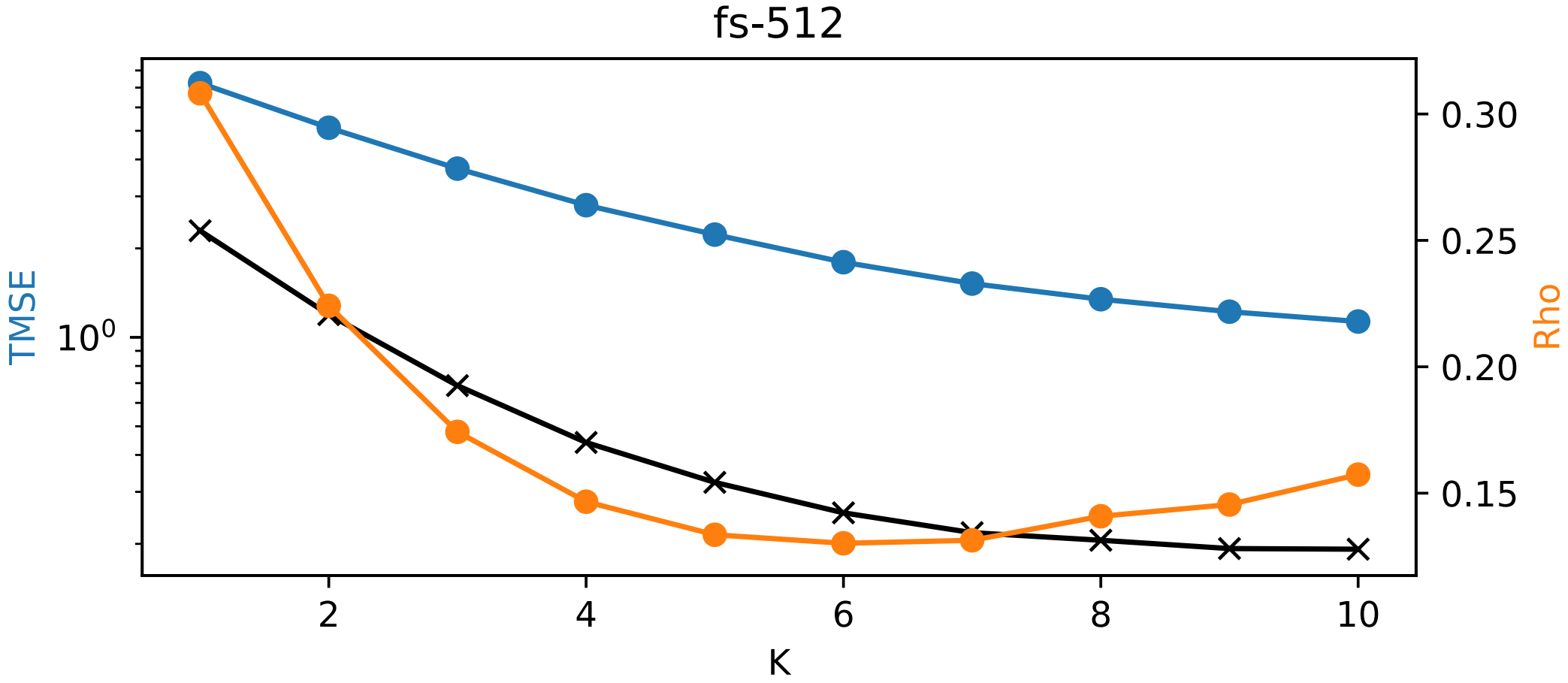
RANDOMIZATION IN CART



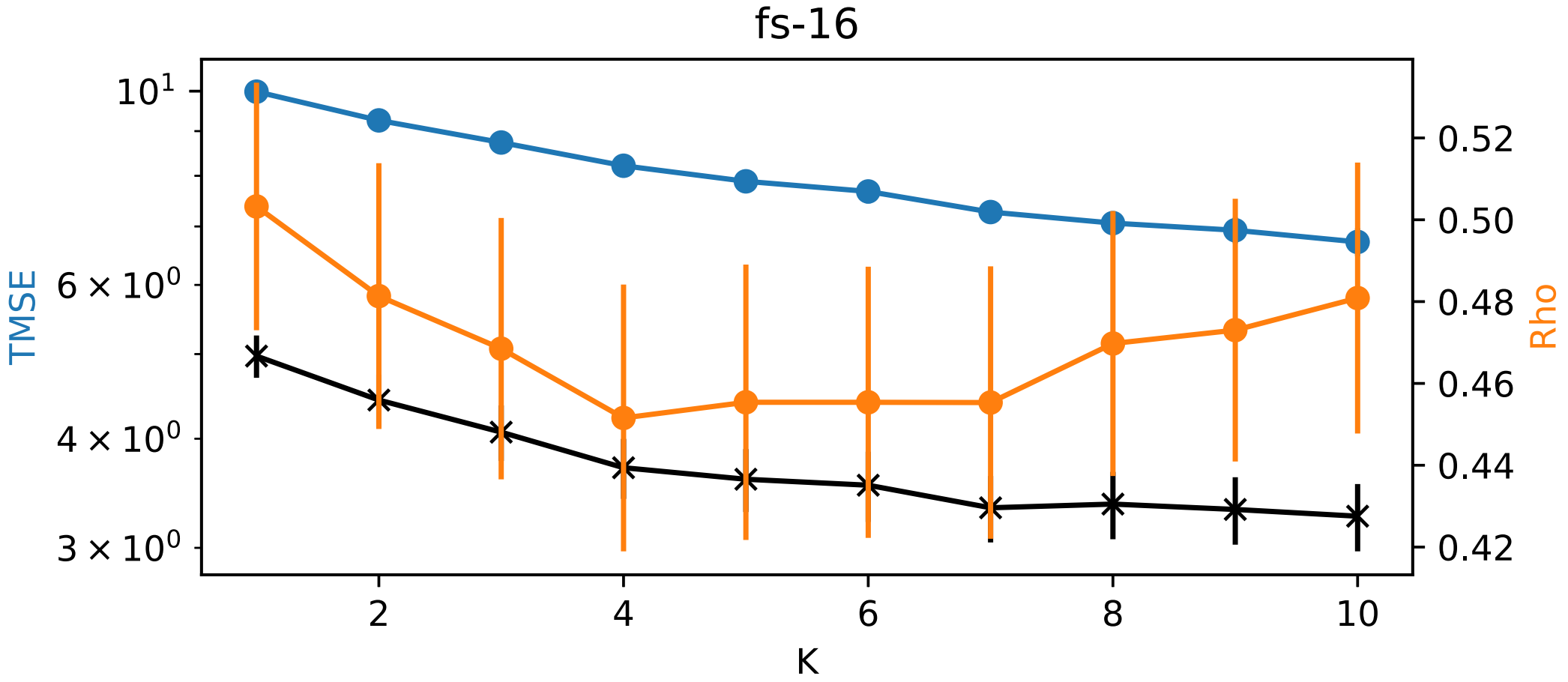
RANDOMIZATION IN CART



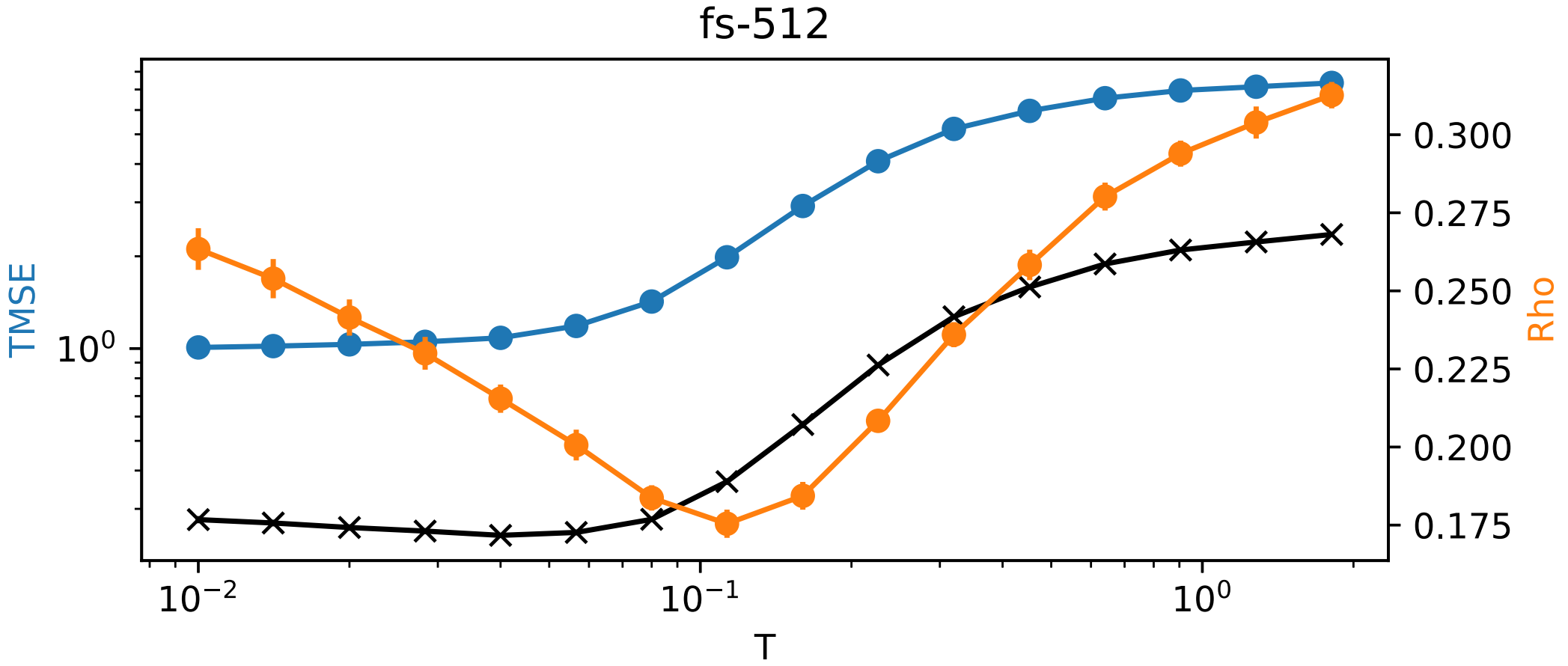
RANDOMIZATION IN EXTREMELY RANDOM TREES



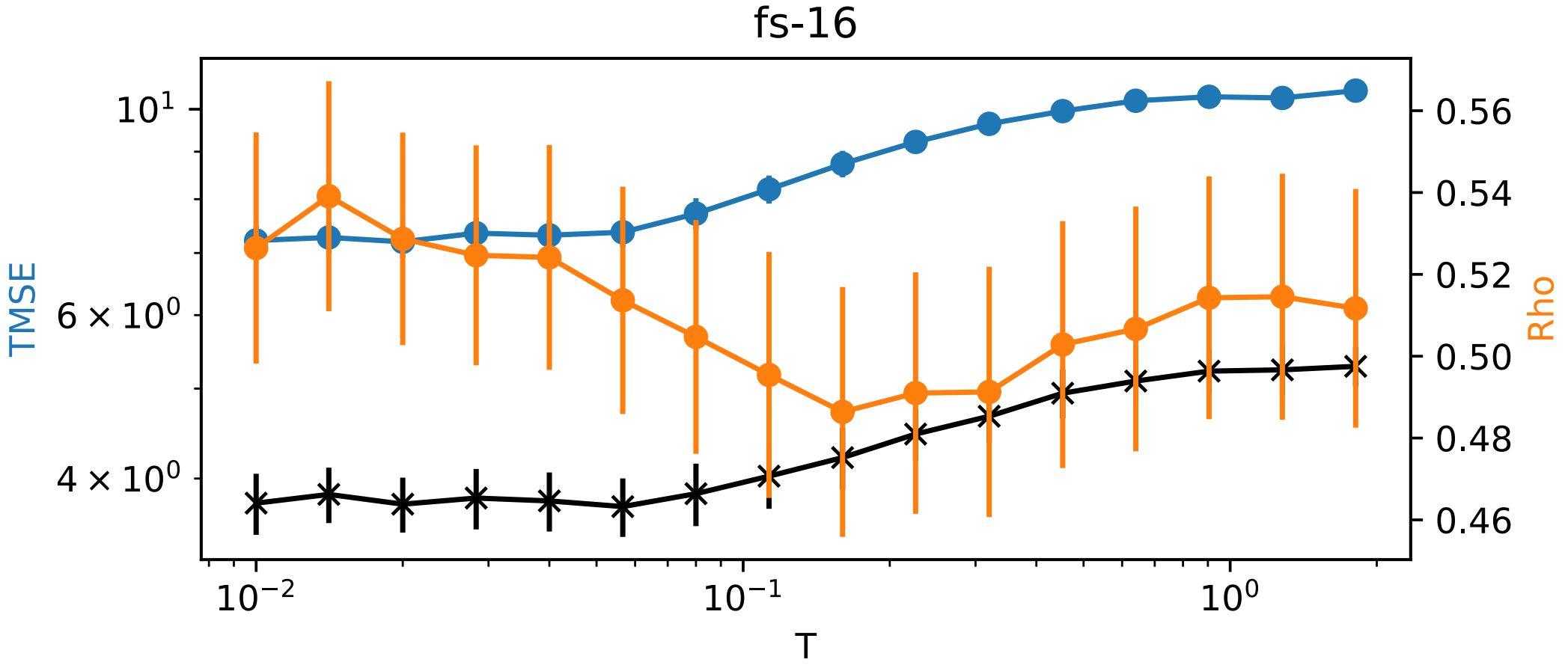
RANDOMIZATION IN EXTREMELY RANDOM TREES



RANDOMIZATION IN BOLTZMANN TREES



RANDOMIZATION IN BOLTZMANN TREES



REFRESHER

The generalization error (MSE) can be decomposed:

$$MSE(f) \leq \bar{\rho} \cdot MSE(h)$$

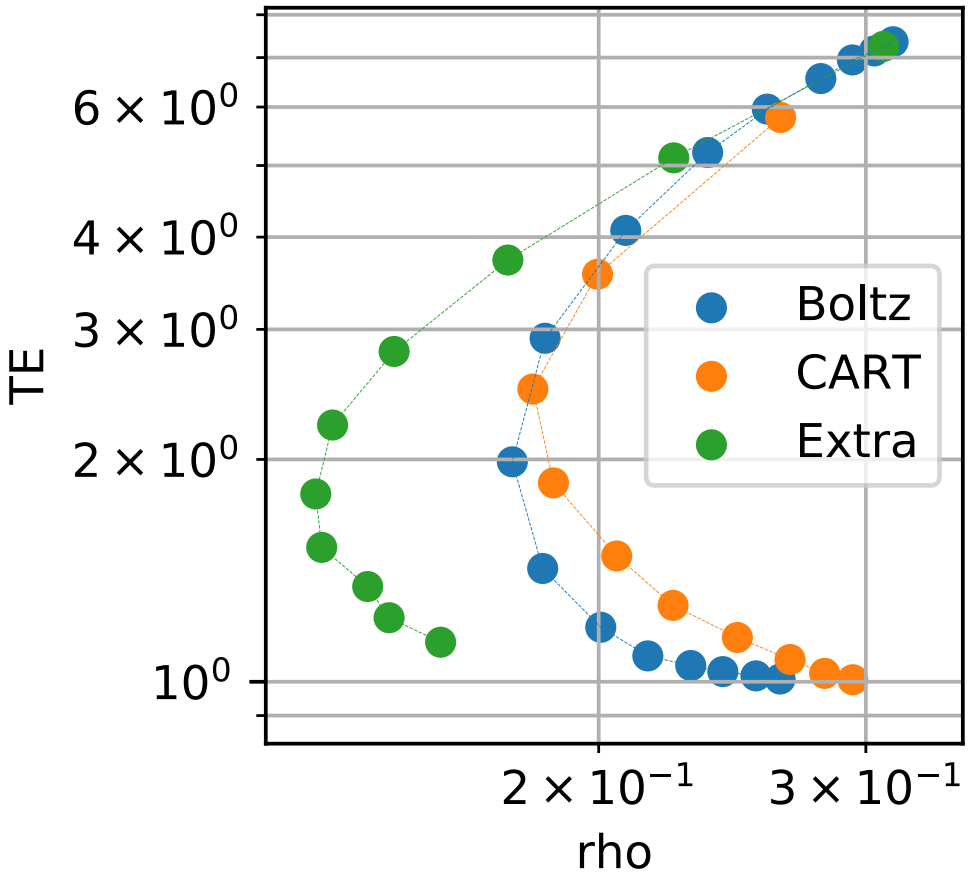
For a weighted correlation:

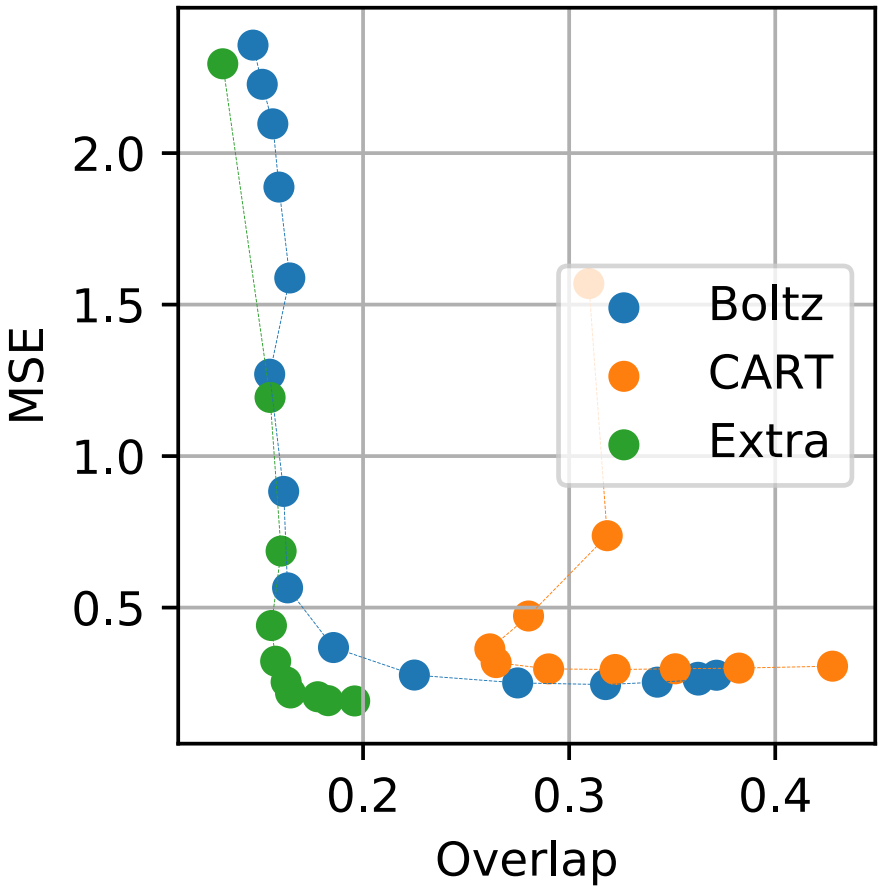
$$\bar{\rho} = \frac{E_i E_j [E_x (Y - h(x, \Theta_i))(Y - h(x, \Theta_j)))]}{E_i E_x (Y - h(x, \Theta_i))^2}$$

Define an "overlap" that grows with the correlation between the trees:

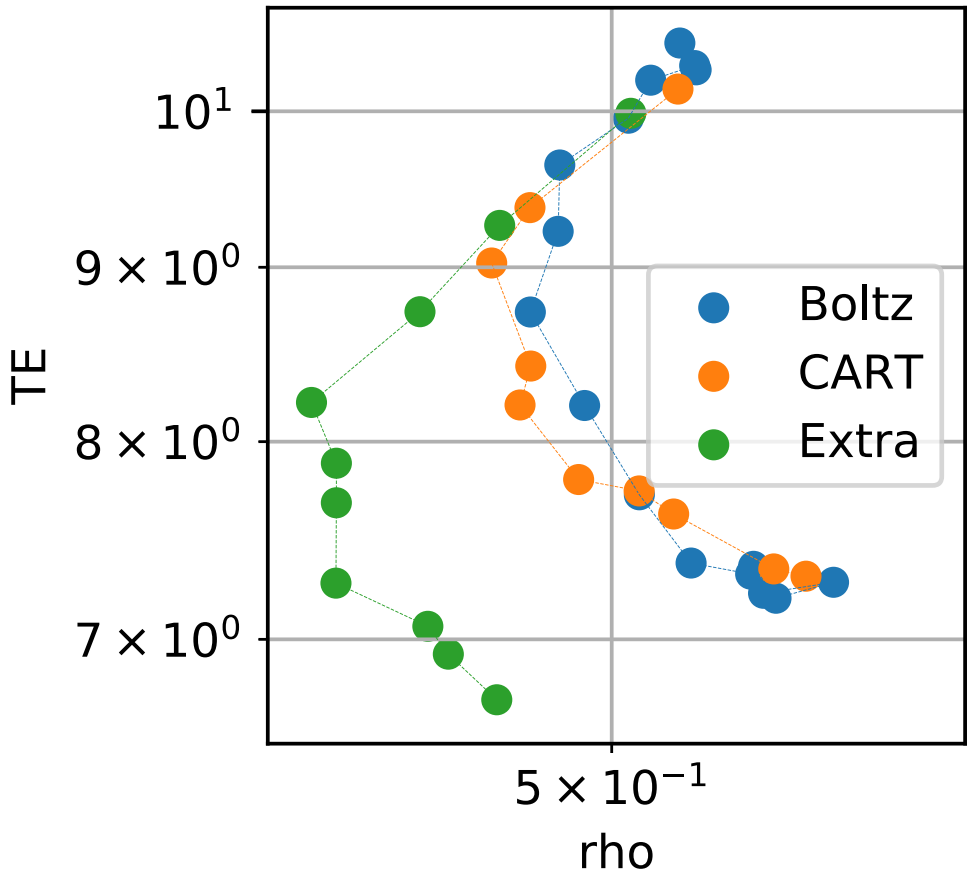
$$\mathcal{S} = \frac{\sum_i \lambda_i^2(\tilde{\rho})}{\sum_i \tilde{\rho}_{i,i}^2} - 1$$

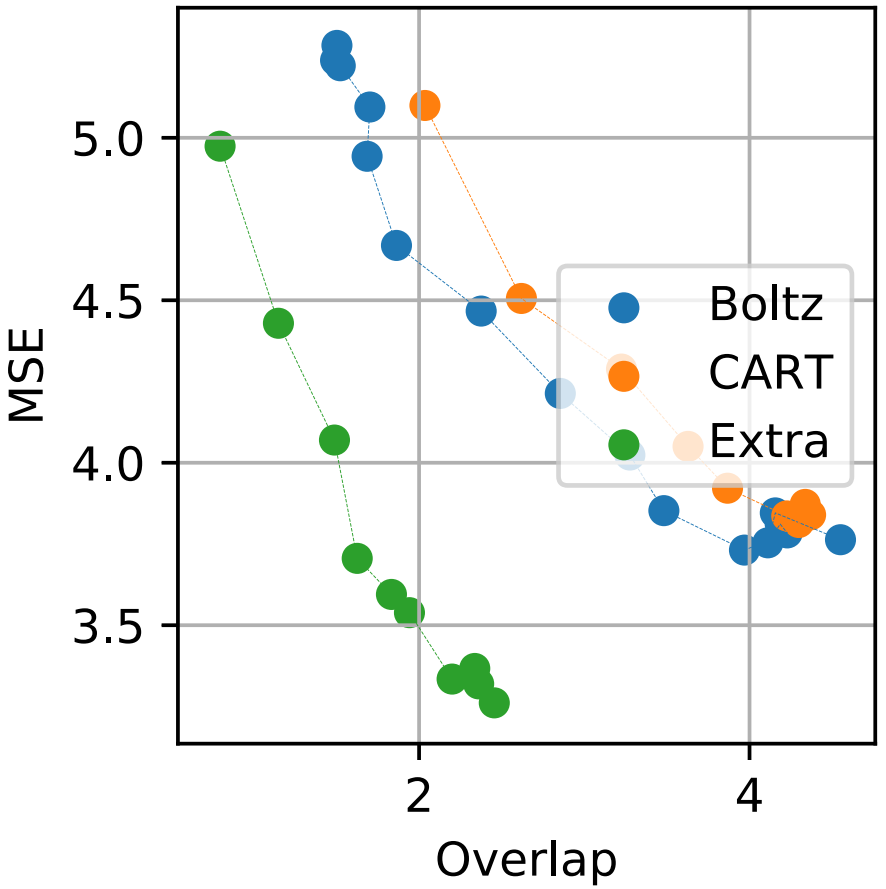
CORRELATION AND TREE ERROR, N=512



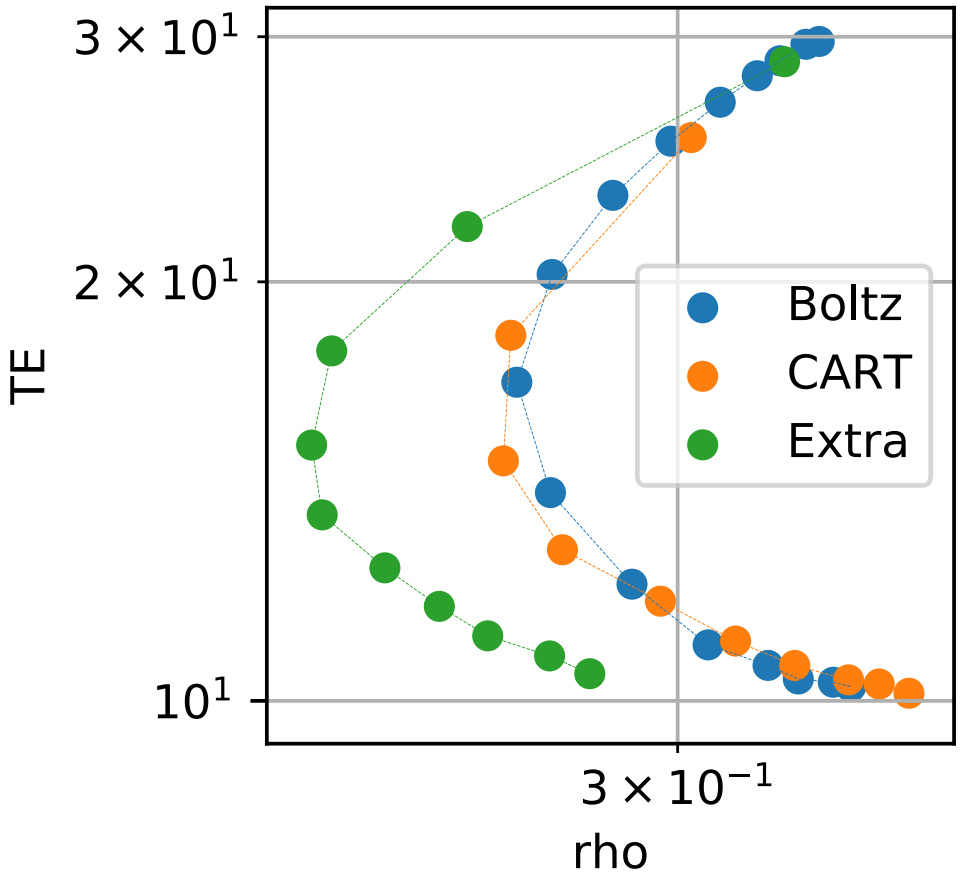


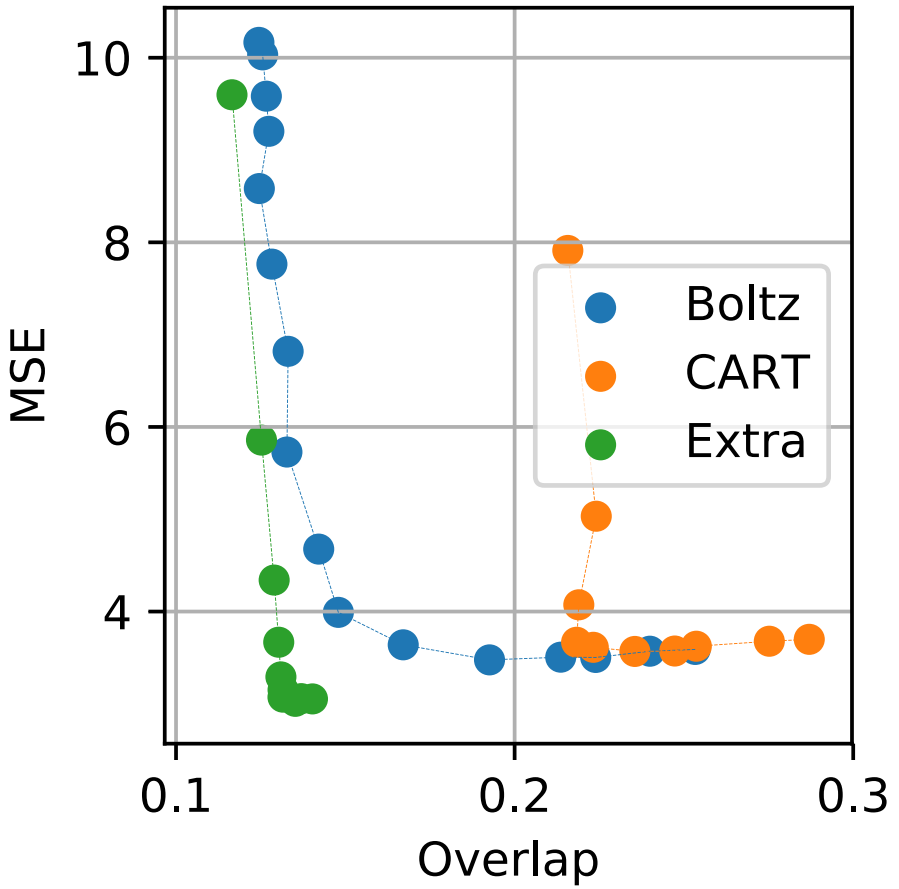
CORRELATION AND TREE ERROR, N=16



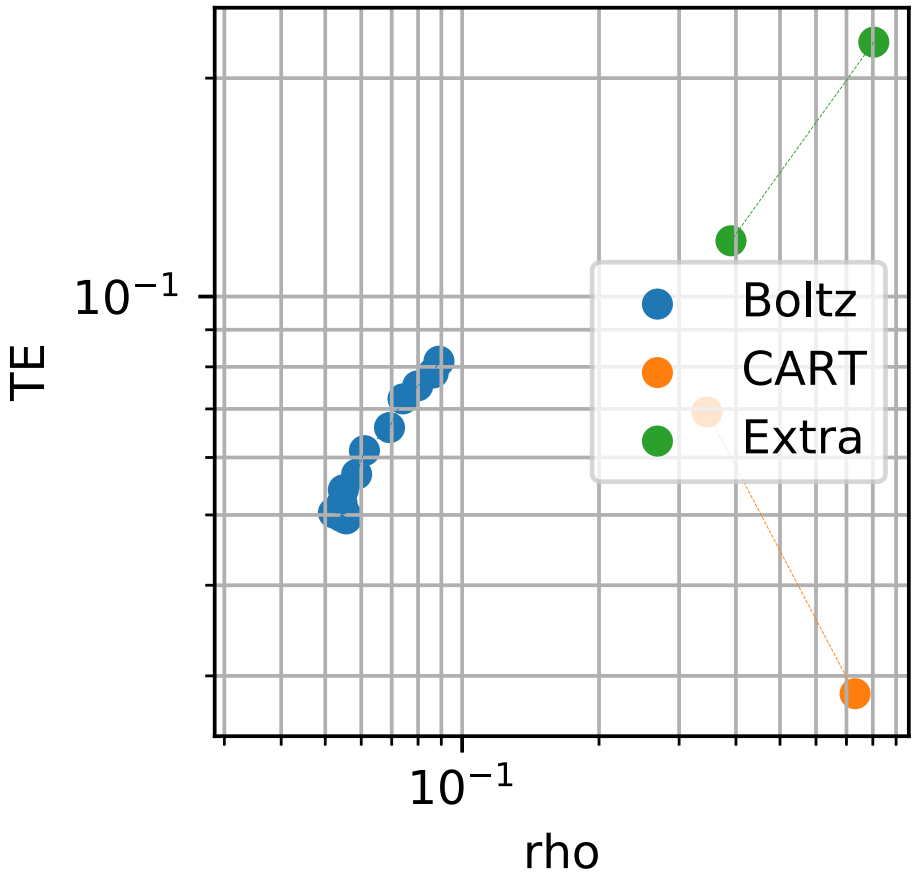


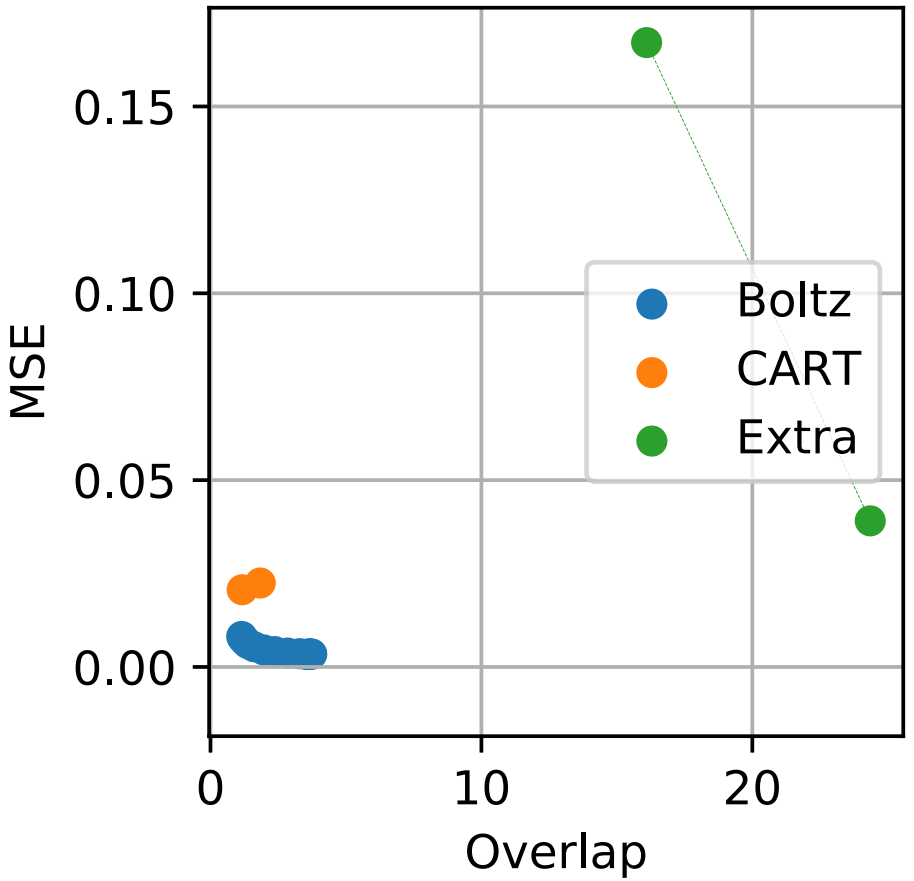
CORRELATION AND TREE ERROR, N=512, F.G.S.





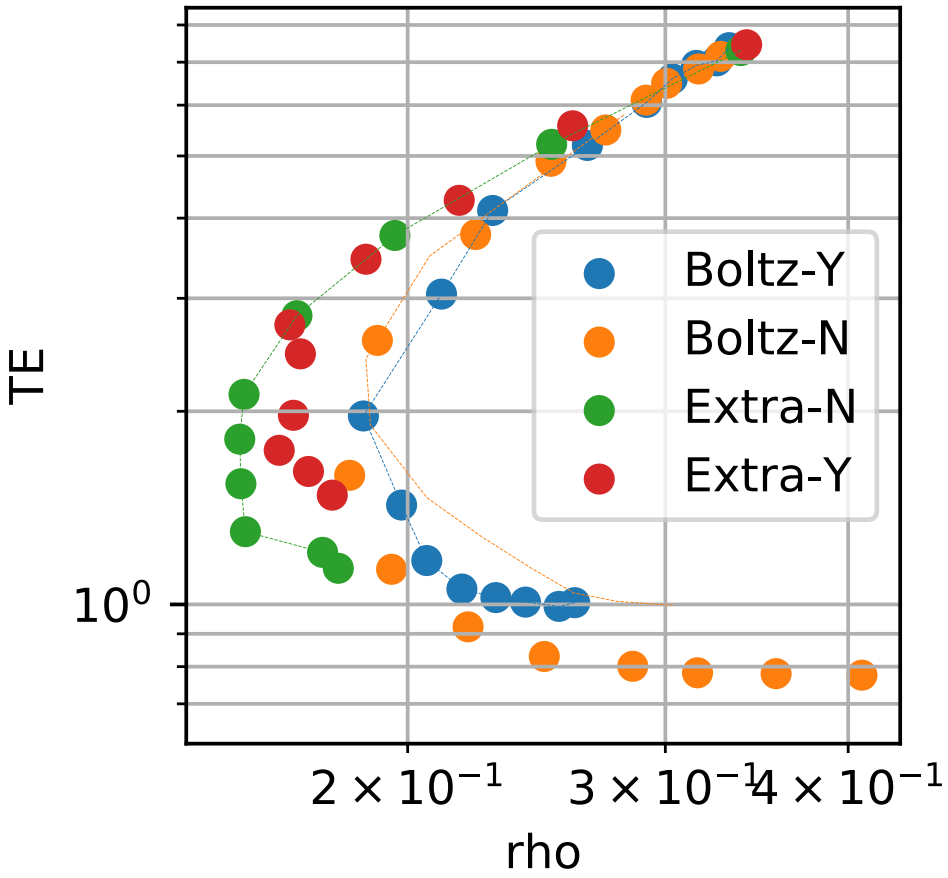
CORRELATION AND TREE ERROR, GRID

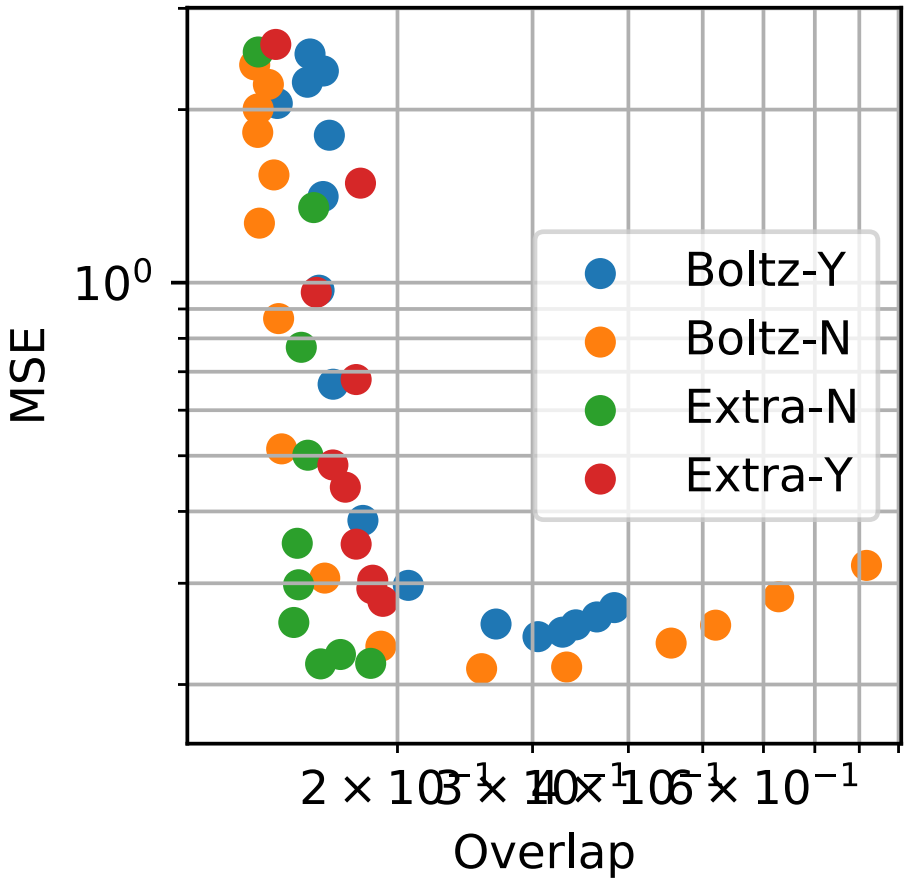




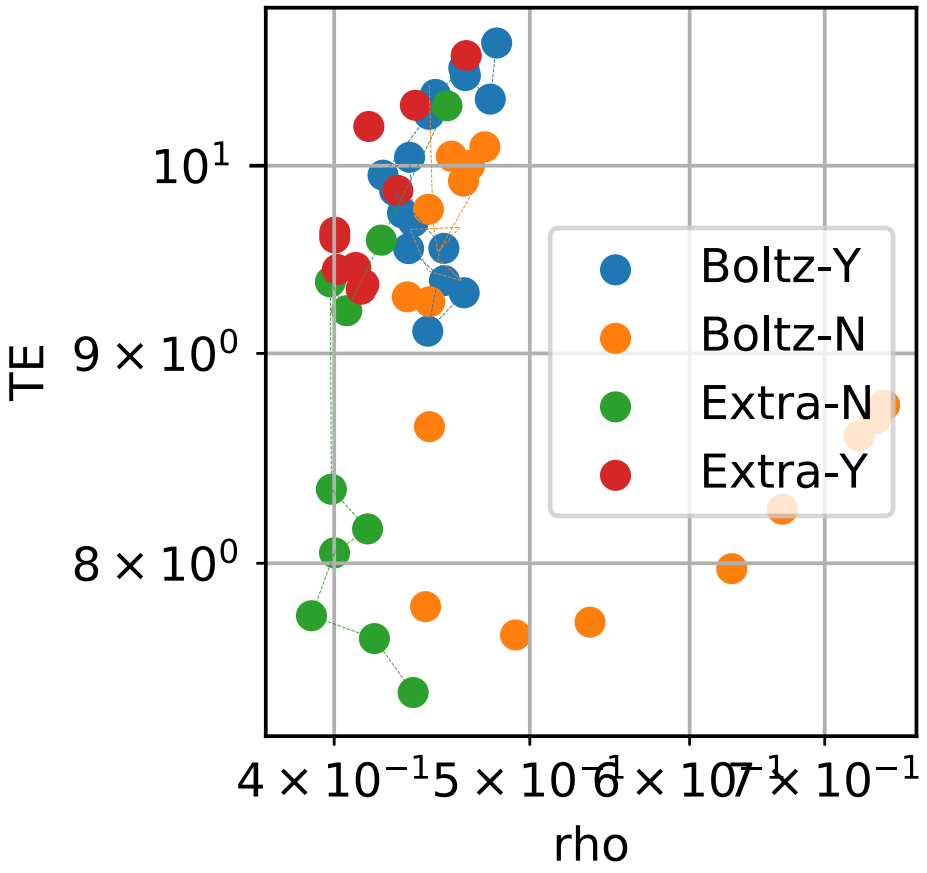
TO BAG OR NOT TO BAG?

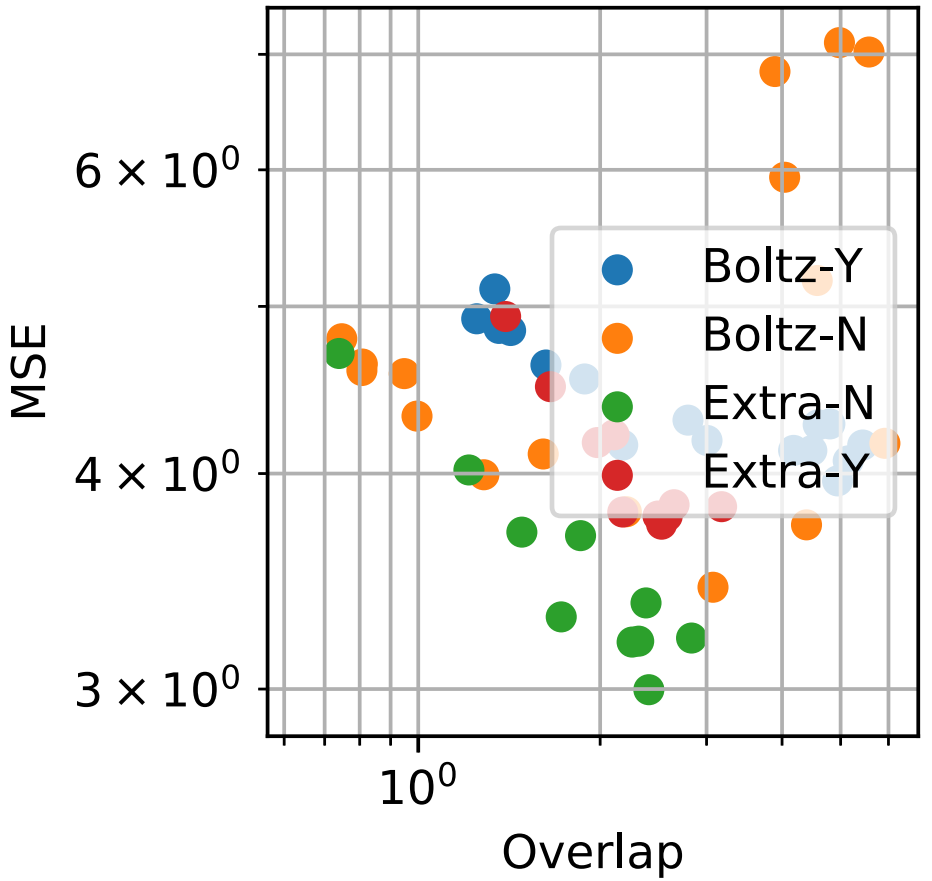
BAGGING VS NOT BAGGING, N=512





BAGGING VS NOT BAGGING, N=16





IN SUMMARY

BOLTZMANN TREES

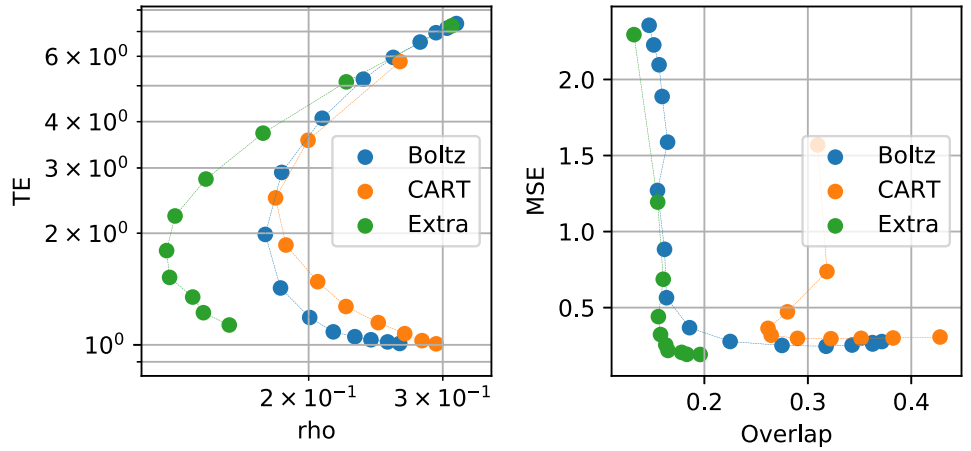
Boltzmann trees trade-off tree error and ensemble correlation via kT

$$P(x,i) \sim \frac{1}{\Delta x} \exp\left[-\frac{I(x,i)}{kT}\right]$$

Boltzmann trees often outperform CART, but extremely random trees can be better

Consider extremely random if you can, and Boltzmann if you cannot.

- clustered data
- uncertainty quantification



ENSEMBLE METHODS: SOME QUESTIONS

What is the optimal way to trade-off correlation and tree error?

- Can you randomize split locations better than a Boltzmann tree?
- How should the energy and the impurity relate?
- Is bagging even a good idea?

How well do these results generalize to more complex problems:

- How strong are extremely randomized trees for highly clustered data?
  - What does the split probability distribution look like?
- What happens when there is label noise?

ENSEMBLE METHODS: SOME MORE QUESTIONS

How do correlation and tree error affect other model results?

- Uncertainty quantification, e.g. Infinitesimal Jackknife
- Model interpretability

How does the correlation affect the "converged" ensemble size?

- Can we get away with smaller ensembles if there is less correlation?
- Can more expensive splitting methods be a net-win?
  - Especially when storing and applying

How do we measure the (positive) externality of randomization?

THANK YOU!

Job listings: citrine.io/careers

Newsletter: citrine.io/newsletter

Podcast: citrine.io/podcast

CITRINE INFORMATICS