ACMFACCT

MONTRÉAL 2026

The Illusion of the Best Model — Multiplicity, Interpretability, and Accountability in High-Stakes AI

Lesia Semenova

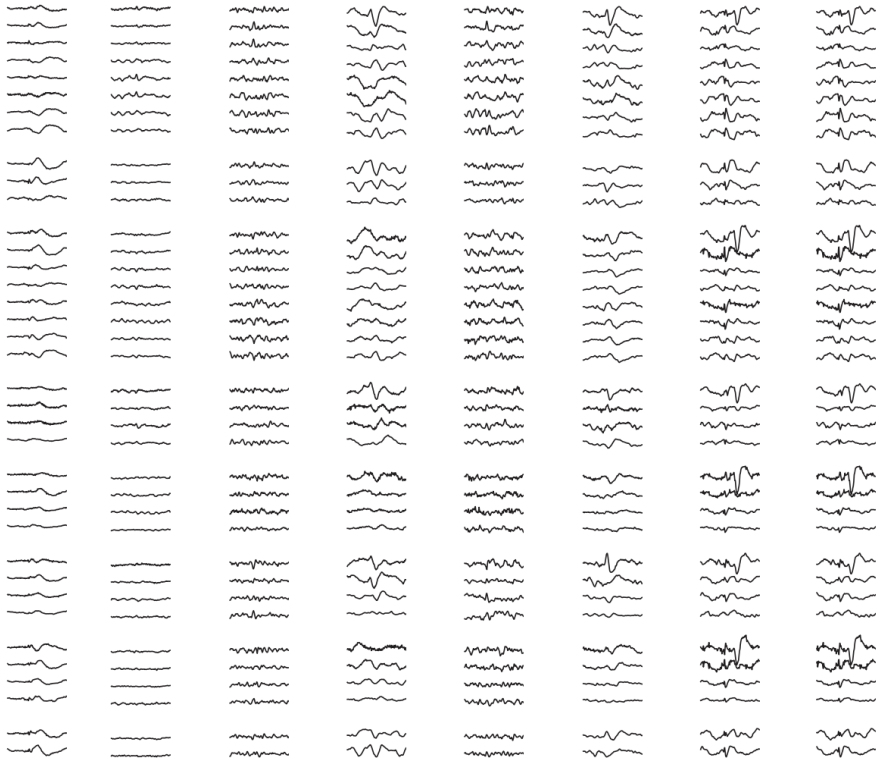
Chudi Zhong

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

The Illusion of the “Best” Model

Diagnosing epilepsy from EEG recordings:



The Illusion of the “Best” Model

Diagnosing epilepsy from EEG recordings:
10 different classifiers, every one beating the baseline at ~85% accuracy.

Model 1 ~85% accurate	**Model 2** ~85% accurate	**Model 3** ~85% accurate	**Model 4** ~85% accurate	**Model 5** ~85% accurate
Model 6 ~85% accurate	**Model 7** ~85% accurate	**Model 8** ~85% accurate	**Model 9** ~85% accurate	**Model 10** ~85% accurate

The Illusion of the “Best” Model

Diagnosing epilepsy from EEG recordings:
10 different classifiers, every one beating the baseline at ~85% accuracy.

Model 1 ~85% accurate	**Model 2** ~85% accurate	**Model 3** ~85% accurate	**Model 4** ~85% accurate	**Model 5** ~85% accurate
Model 6 ~85% accurate	**Model 7** ~85% accurate	**Model 8** ~85% accurate	**Model 9** ~85% accurate	**Model 10** ~85% accurate

Accuracy didn't choose the model. **We did!**

The **Problem**

Accuracy alone is not enough to choose a model

- Ensembling adds unnecessary complexity
- Ignoring it hides disagreement between models
- Keep exploring sequentially can be expensive

Every ”cheaper” fix doesn’t look at what each model does and why this phenomenon is happening

Multiplicity of well-performing models

All models are equally good in terms of accuracy,
but disagree on HOW they make predictions

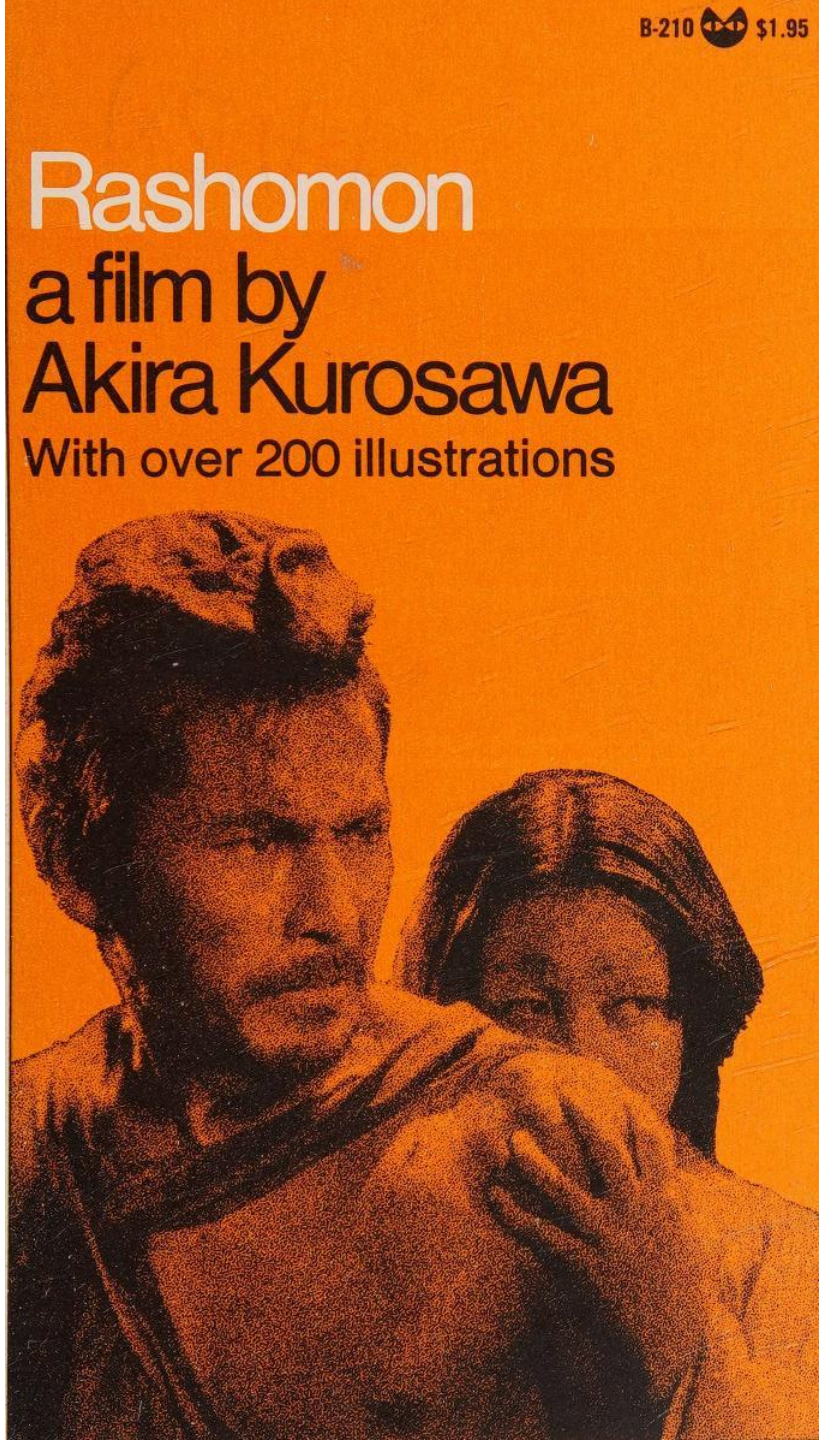
Rashomon Effect

Tutorial Outline

1. Intro and definitions
    - When does it occur?
    - Why does it occur?
2. Why should we care?
    - What does it mean?
    - What are the implications?
3. How can we explore it?
4. Future notes and outro

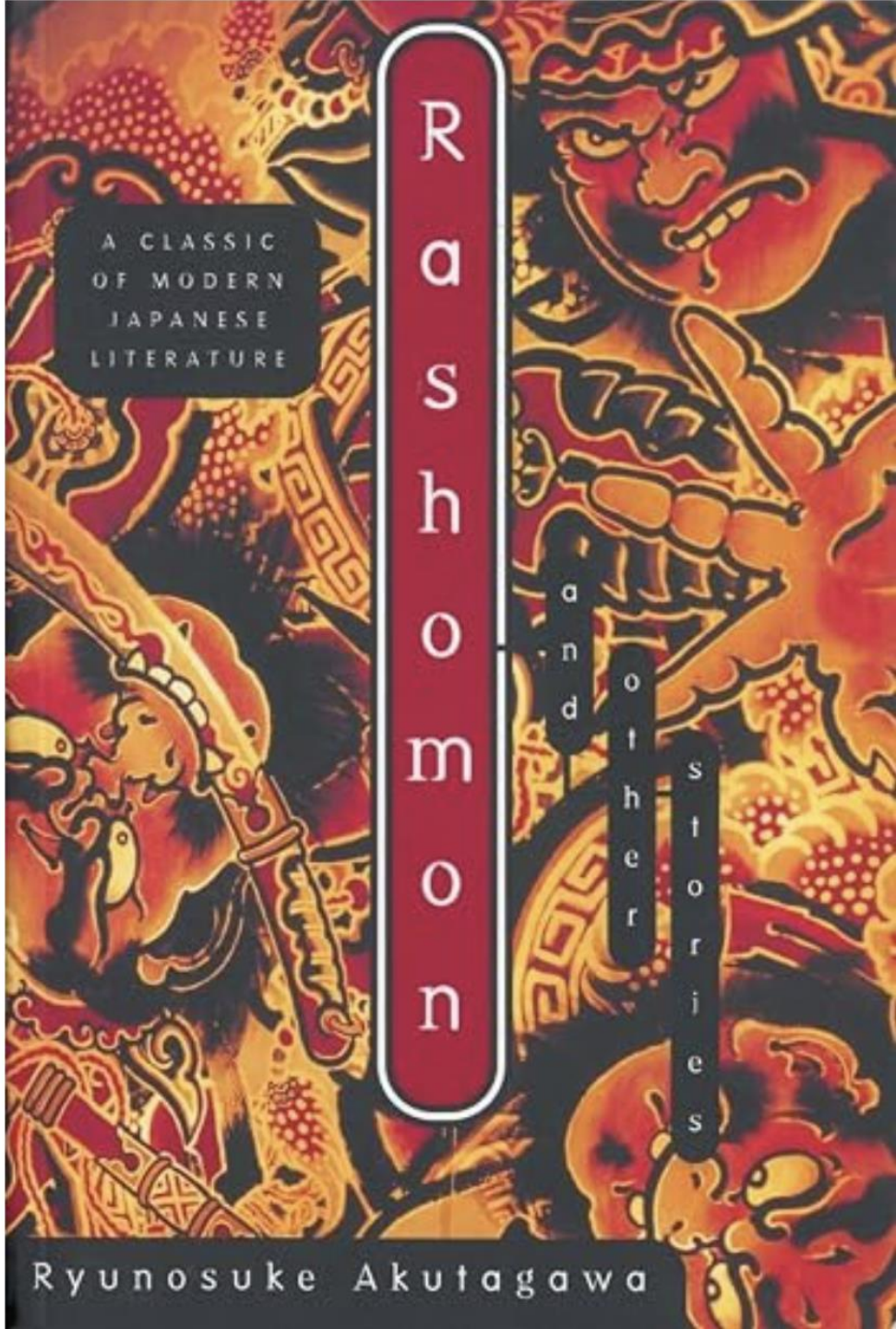
Motivation and Definitions

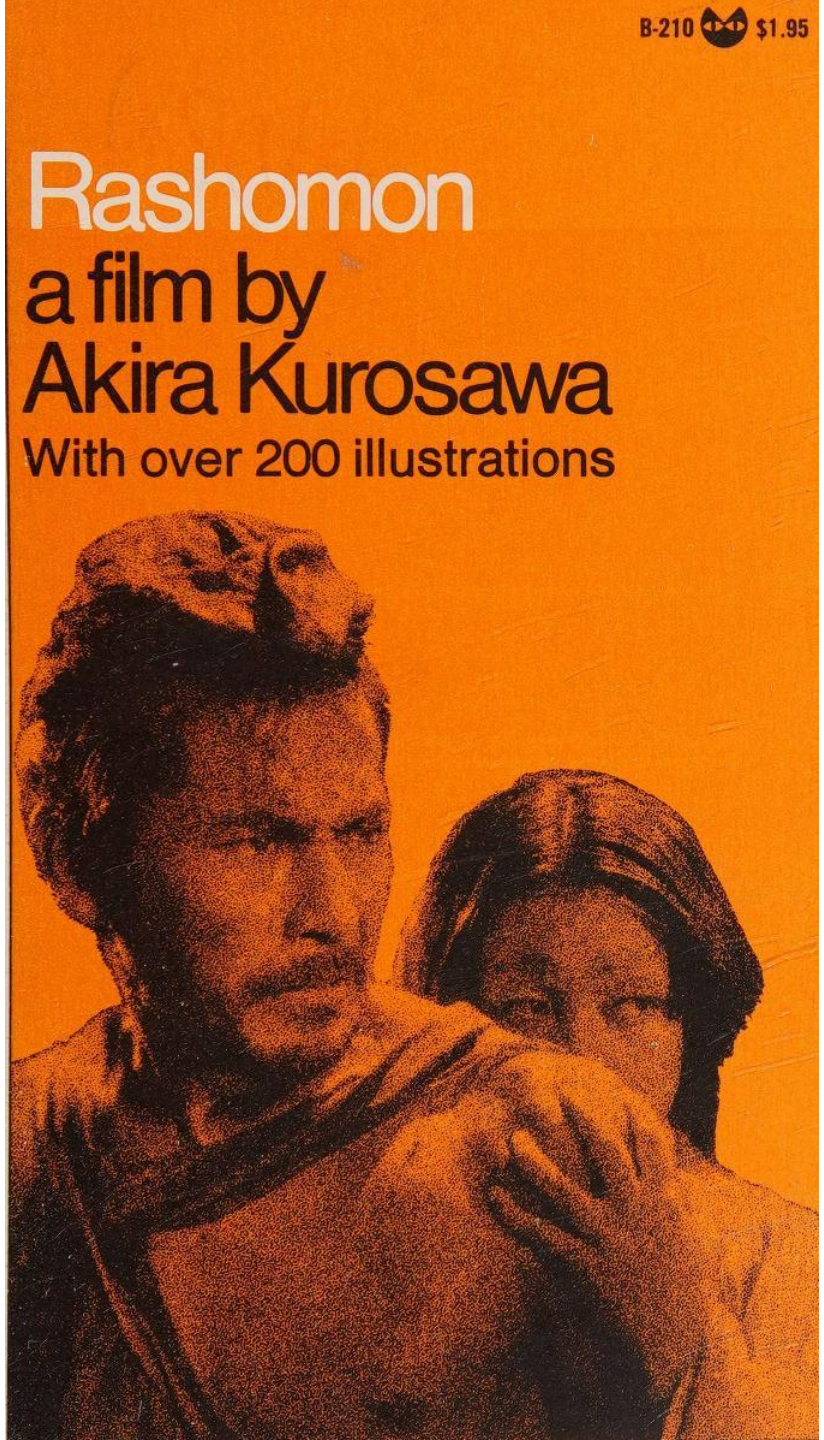
Why should we care?



"Rashōmon”
(The Setting & Title)

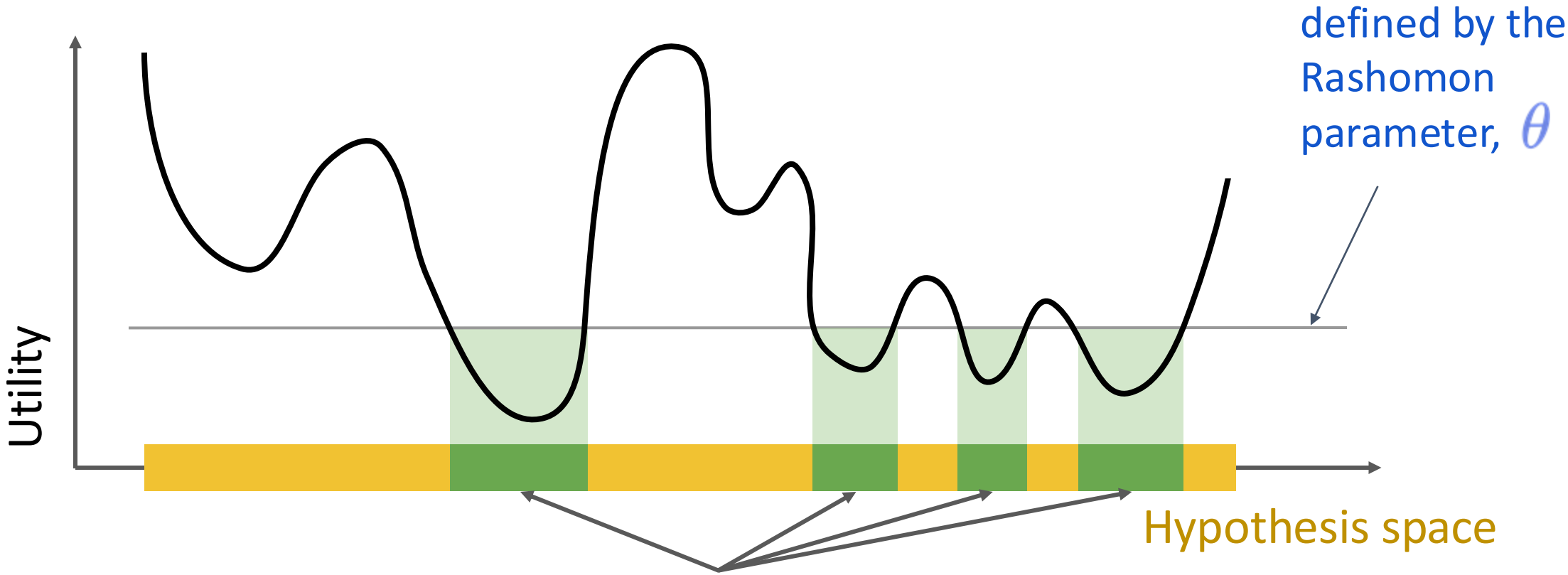
"In a Grove”
(The Actual Plot)





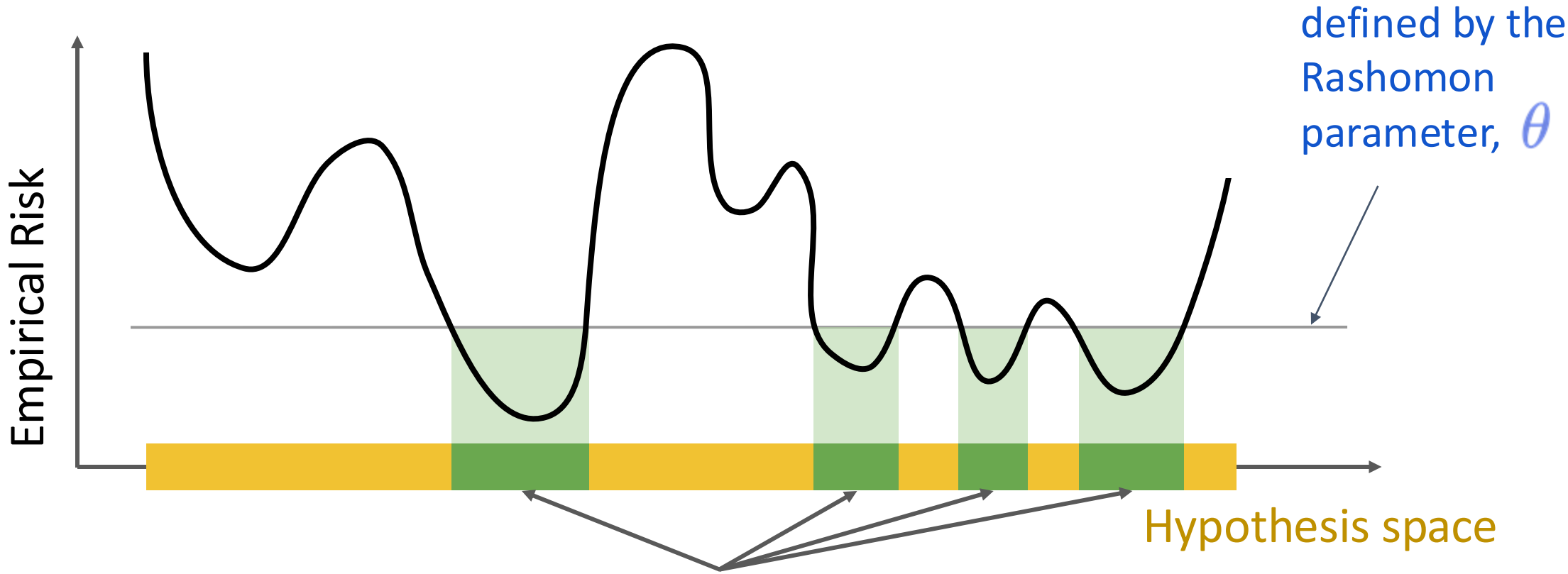
"What I call the **Rashomon Effect** is that there is often a multitude of different descriptions in a class of functions giving about the same minimum error rate."

Leo Breiman (2001)



models that perform approximately-equally well

Semenova et al., FAccT 2022



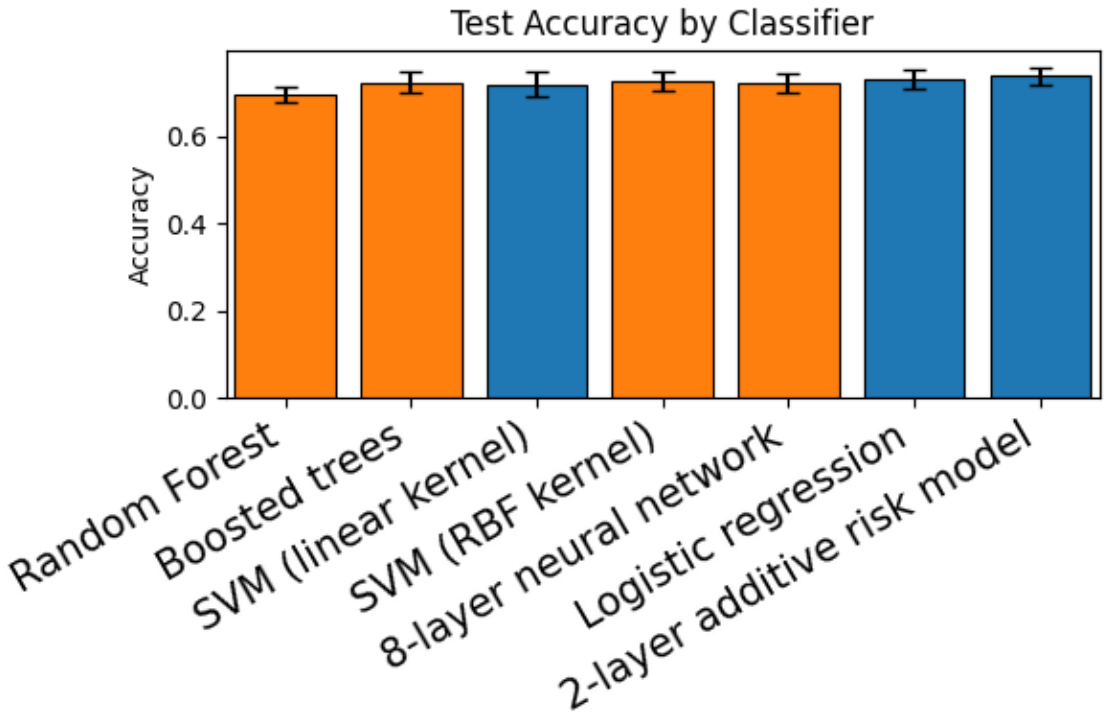
models that perform approximately-equally well

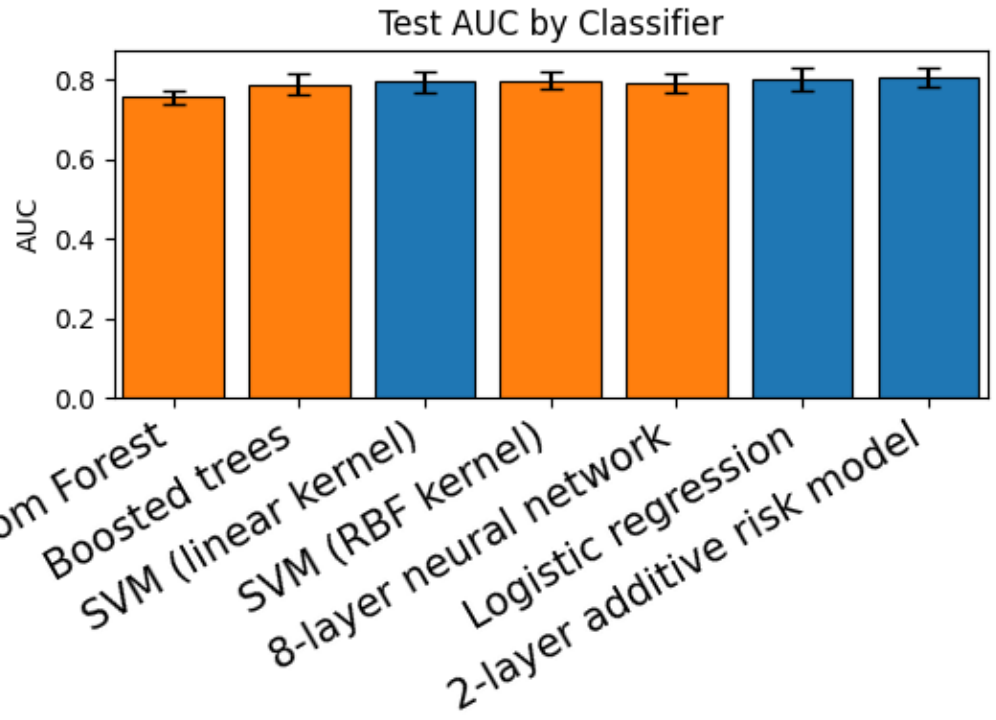
$$R_{set} = \{f : Risk_{Emp}(f) \leq Risk_{Emp}(f_{Emp}^{best}) + \theta\}$$

Semenova et al., FAccT 2022

The Rashomon Effect is everywhere

- Explainable ML Challenge – FICO dataset





Simpler/interpretable models

Black box models

Rudin et al., ICML 2024

The Rashomon Effect is everywhere, …including Gen AI

- **Epistemic Uncertainty.** If the prompt for language model is underspecified, there are multiple plausible answers (Hou et al., ICML 2024)
- **Representation Ambiguity.** Two BERT models with different seeds can achieve identical test scores, yet one learns robust linguistic rules while the other relies on fragile keyword-matching shortcuts (McCoy et al., ACL 2020, D'Amour et al., JMLR 2020)
- **Diversity in reasoning**. One model identifying a bird species by its "wing pattern", while another based on its "beak shape" (Feng et al., 2025)

Arbitrary choices at every step

and they compound!

(High stakes) ML

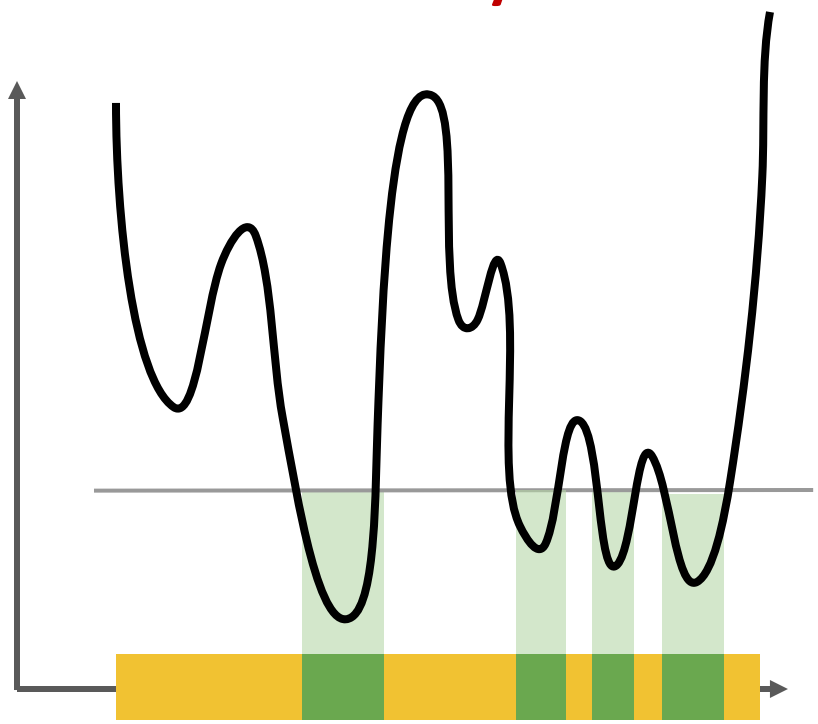
- outlier handling
- imputation method for missing data
- feature encoding and engineering
- model class selection
- random seed and train/test split
- regularization strength
- decision threshold

Generative / foundation-model AI

- base model and checkpoint
- fine-tuning data selection
- adaptation method (full / LoRA / RAG)
- prompt and system message
- preference data and reward model
- decoding temperature and sampling
- refusal / safety cutoff

One dataset can support many stories

Technically



Accountability

- **Different people affected**
- **Different explanations**
- **Different disparities**
- **Different institutional reasons**

A decision you can't explain, can't reproduce, and can't appeal isn't a decision — it's a coin-flip nobody's accountable for.

The Rashomon set can enable the turn of a hidden modeling choice into inspectable, discussable, and documentable policy object

Why does the Rashomon Effect Occur

It happens, but how common?

The Rashomon Effect

Many distinct models that perform approximately equally well

Breiman, 2001; Rudin et al., ICML 2024; Semenova et al., FAccT 2022

Underspecification

When a system or learning problem isn't tightly defined by objectives, data, or evaluation, so several models can satisfy it

D'Amour et al., 2020, JMLR

Model Uncertainty

Captures our inability to know which of the "good" models is the right one (epistemic uncertainty)

Predictive Multiplicity

When models give different predictions for the same input, even though they all perform well

Marx et al., ICML 2020; Black et al., FAccT 2022

For systematization: Ganesh et al., AIES 2025; Parikh, 2026

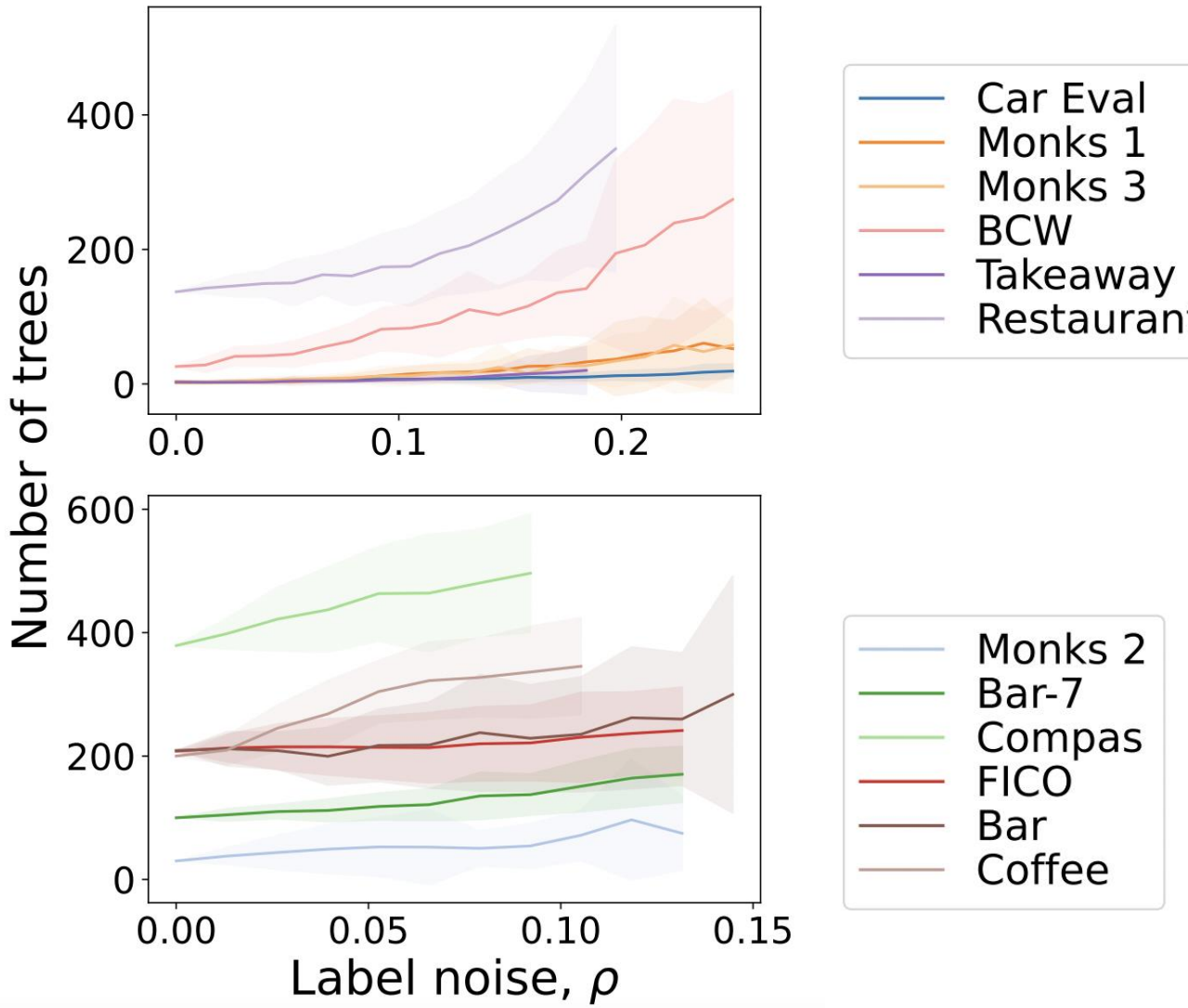
Why does the Rashomon Effect occur?

Clean data

Noisy data

Noise as one of the causes

The size of the Rashomon set **increases with the random label noise** for the hypothesis space of sparse decision trees



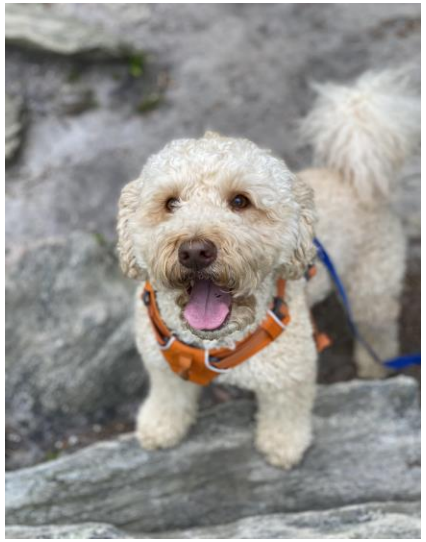
Semenova et al., NeurIPS 2024

Informally, noise distorts the feature signal, which allows many different models to achieve similar performance on the same dataset.

Thus, we can **expect larger Rashomon sets.**

Theorem 8, Boner et al., NeurIPS 2025

vision,
purely predictive tasks



Outcomes are **certain**

High-stakes decision domains
tabular data

healthcare

lending

criminal justice

Outcomes are **uncertain**

Semenova et al., NeurIPS 2024

We do not need to know

exactly how much noise is present.

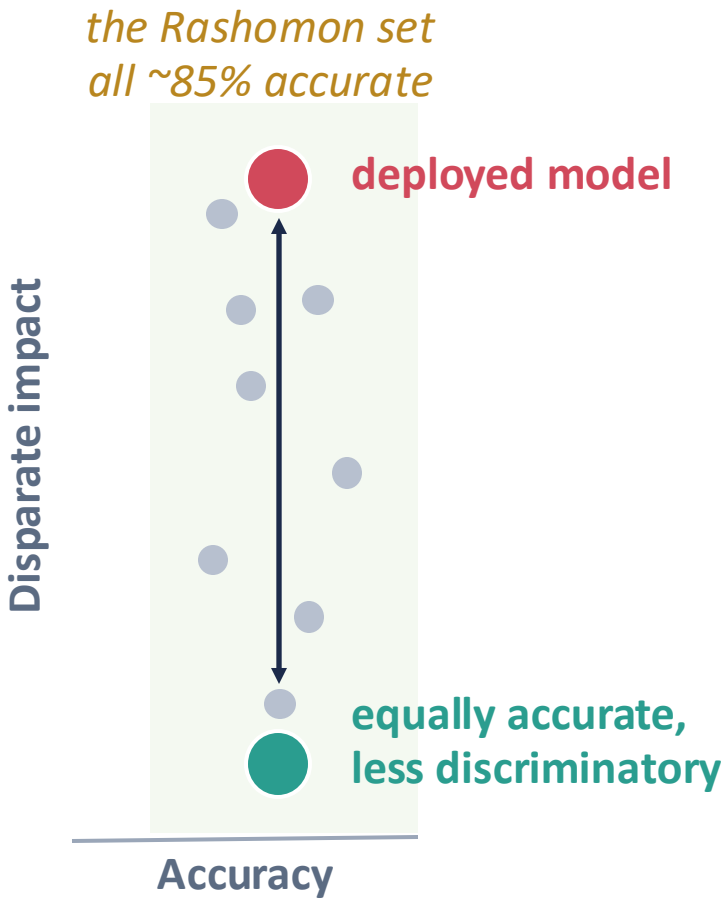
As long as it is present, think **multiplicity**!

What are the Implications?

Interpretability, Fairness, Robustness, Privacy

The Rashomon set can enable the turn of a hidden modeling choice into inspectable, discussable, and documentable policy object

The fairer model was already there

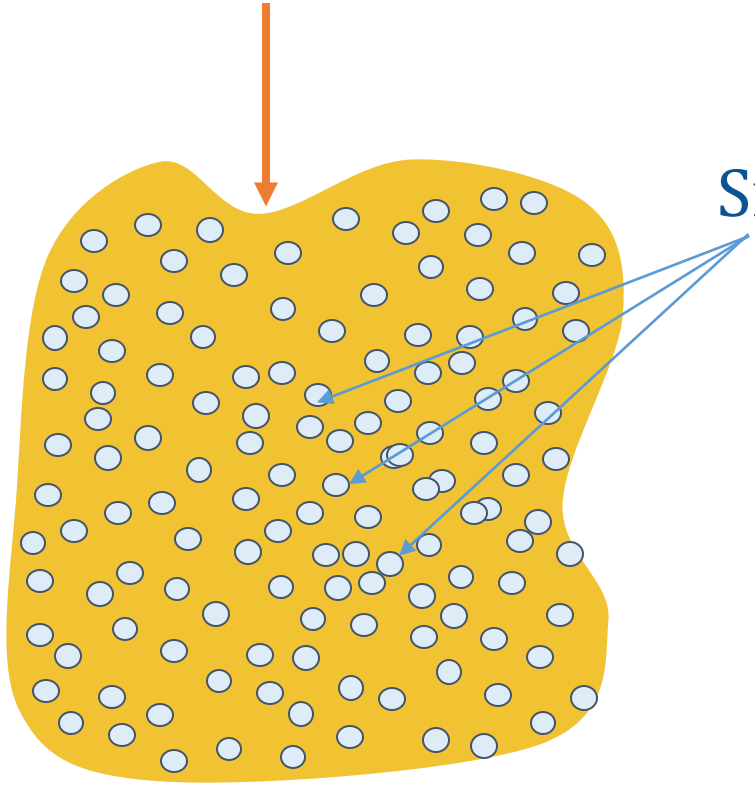


Disparate-impact law has long asked whether a **less discriminatory alternative** was available. Model multiplicity shows one **almost always exists** — so failing to look for it becomes legally and ethically relevant, not a neutral default.

Black, Koepke, Kim, Barocas & Hsu — "Less Discriminatory Algorithms," Georgetown Law Journal 113(1), 2024.

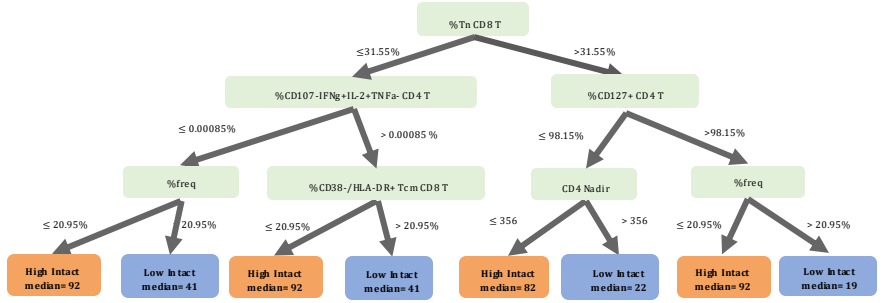
The interpretable model is likely to be there

Hypothesis space: all models

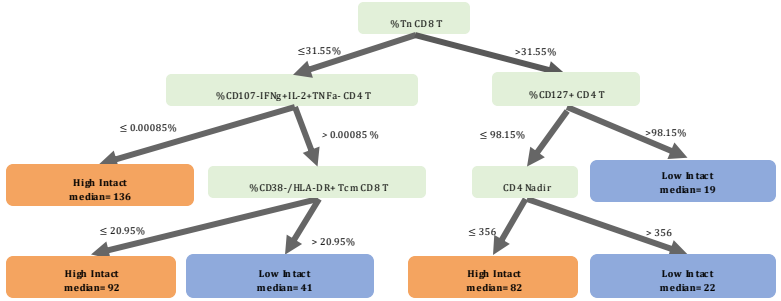


Simpler models: cover over the set of all models

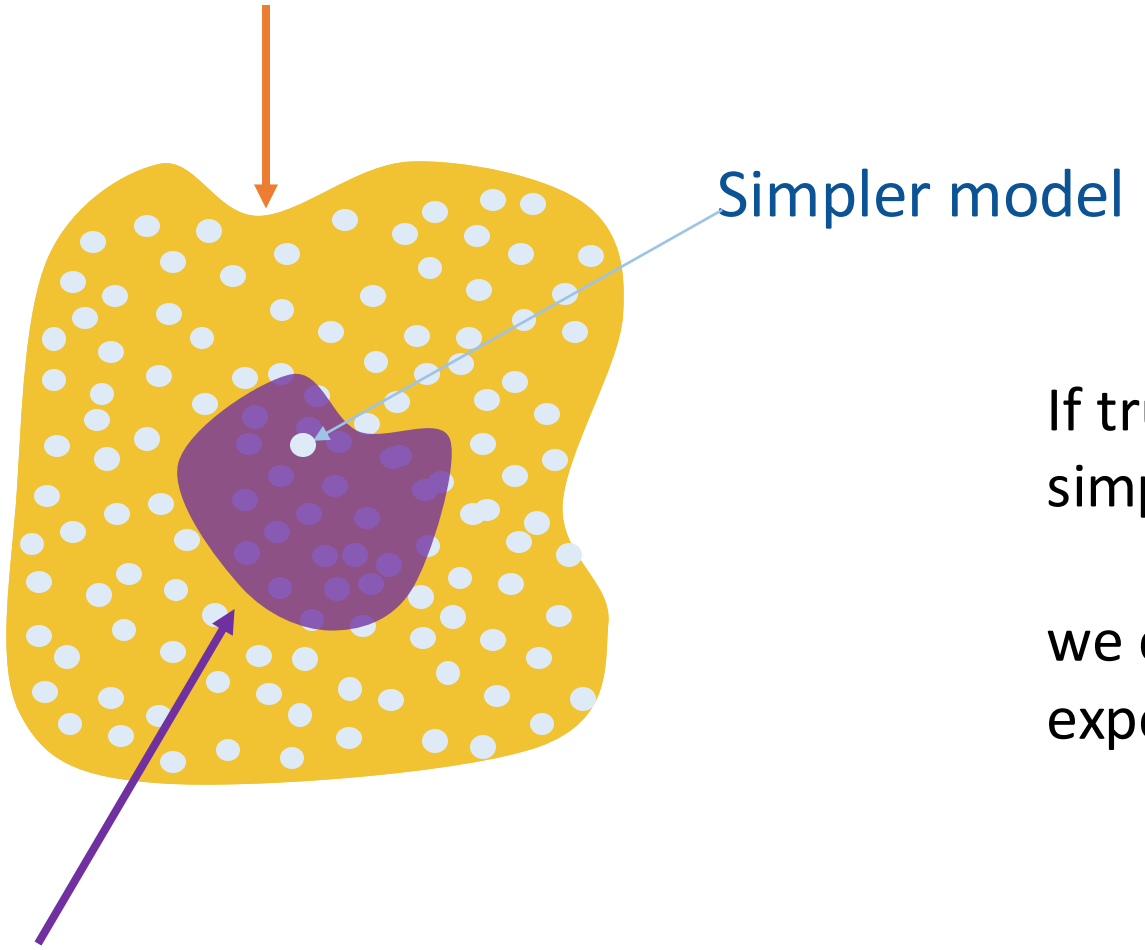
Fully-grown tree



Sparse tree



Hypothesis space: all models



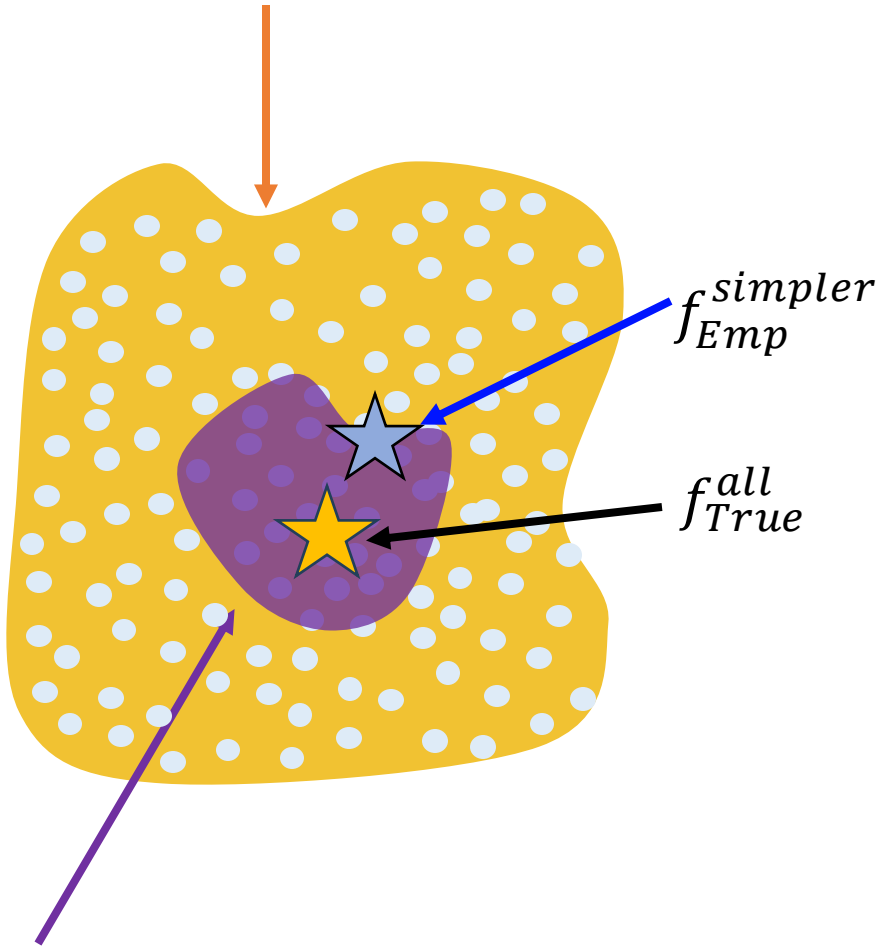
If true Rashomon set contains at least one simple model,

we can optimize for simpler class and expect good generalization!

True Rashomon set: models with true loss less than the threshold

Semenova et al., FAccT 2022

Hypothesis space: all models



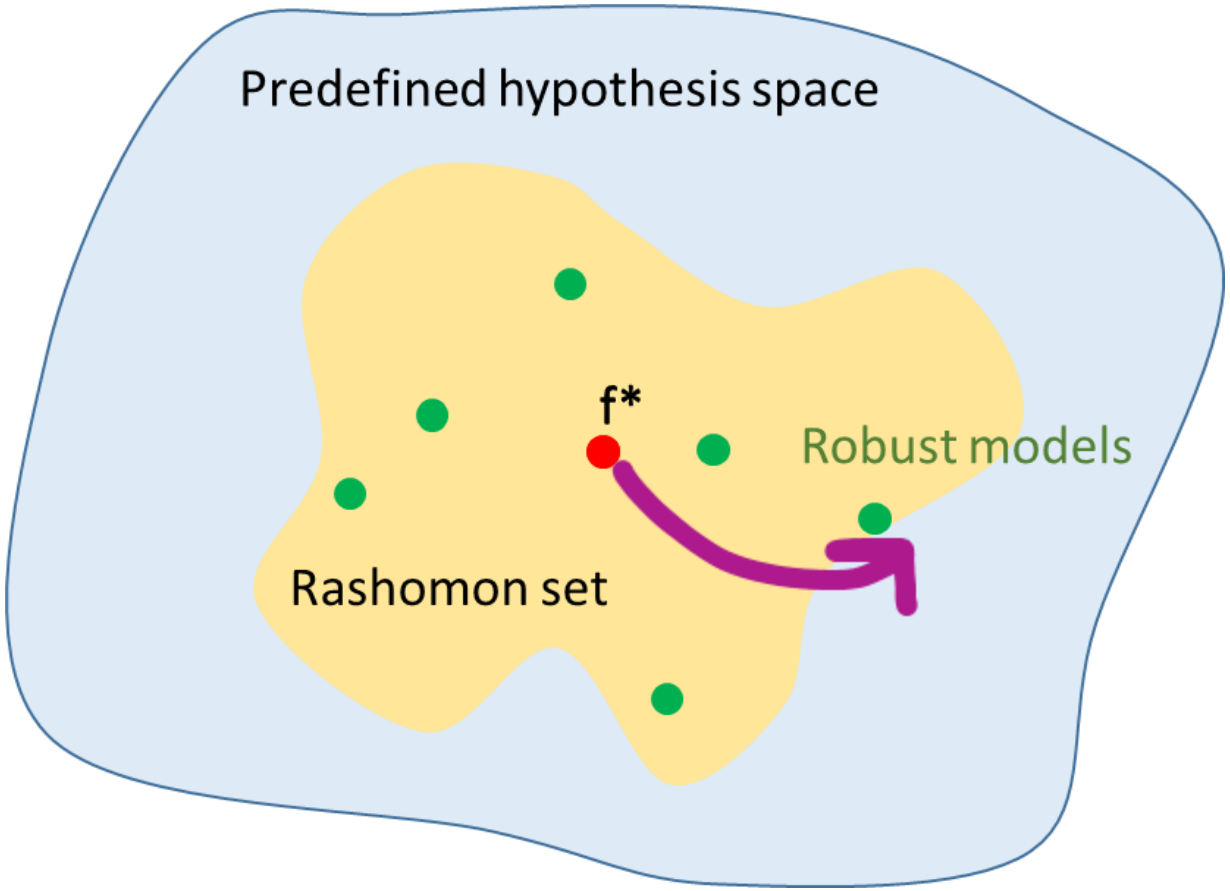
True Rashomon set: models with true loss less than the threshold

If true Rashomon set contains at least one simple model,

we can optimize for simpler class and expect good generalization!

Semenova et al., FAccT 2022

If we do not like the properties of one model, we can switch to another



Diversity between models in the Rashomon set enable ***reactive robustness***, where we can switch to another model if one was compromised.

Hsu et al., ICML 2026

However, there are risks and caveats!

The same freedom enables ***"d-hacking"*** — selecting a model that looks fair while still discriminating.

The Rashomon set reveals the choices available.
It does not tell you which one is just.
Legal, institutional, and participatory mechanisms still have to.

Privacy. Releasing multiple models from the Rashomon set enables information leakage, where adversary can learn more about the data with more models.

Hsu et al., ICML 2026; Black et al., FaccT 2024

And there is more work

- Differential Privacy
- Active learning
- Explainability and algorithmic recourse
- Variable Importance

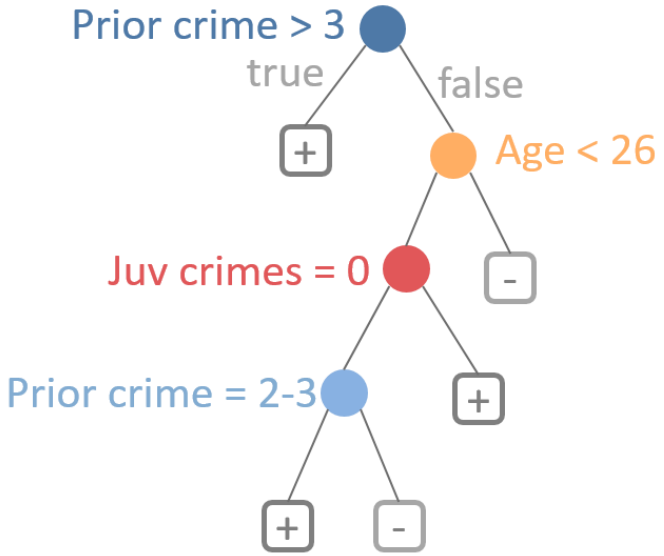
… still, we are not there to fully understand all the impact on DS/ML/AI pipelines

For systematization: Ganesh et al., AIES 2025; Parikh, 2026

Constructing and Exploring the Rashomon Set

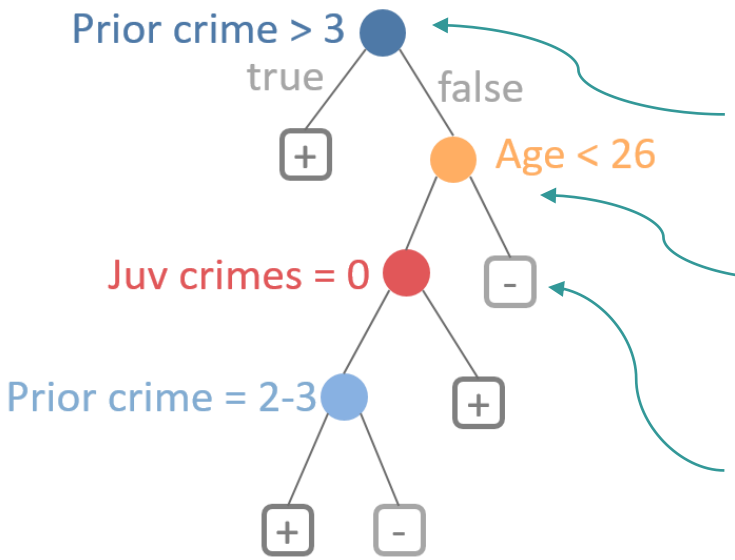
Decision trees, generalized additive models, risk scores, neural networks

Decision Trees



- First algorithm ~1960's
- are nonlinear
- are robust to outliers
- easily handle multiclass
- great for categorical features
- hard to optimize
  - Mathematical optimization solvers, SAT solvers
    - Bennett mid-1990's,.., Blanquero et al., 2018, 2020, Menickelly et al., 2018; Vilas Boas et al., 2019, Verwer & Zhang, 2019, Aghaei et al., 2021, Gunluk et al., 2021,..
  - Branch-and-bound / dynamic programming
    - Garofalakis et al., DTC, 2003, Nijssen & Fromont, DL8, 2007, 2010, Aglin et al., DL8.5, 2020, Lin et al., GOSDT, 2020, McTavish et al. 2022, Demirovic et al., 2022, …

Generalized optimal sparse decision trees (GOSDT)



maximize accuracy + simplicity

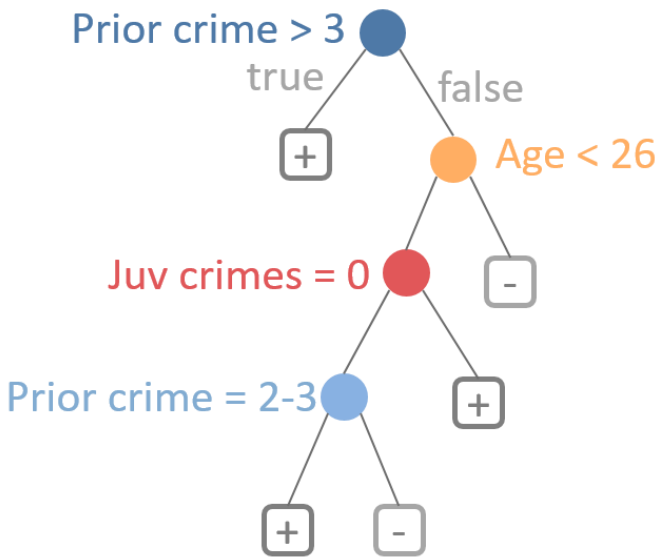
To figure out the optimal split at the top

Figure out the optimal split beneath it.

And the one below that which eventually is a leaf.

Dynamic Programming + Branch and Bound

Decision Trees



Dynamic Programming + Branch and Bound

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]

TreeFARMS, NeurIPS 2022 oral

TreeFARMS (Xin et al., 2022)

Dynamic programming reuses subproblem solutions

Branch & bound efficiently prunes the search space

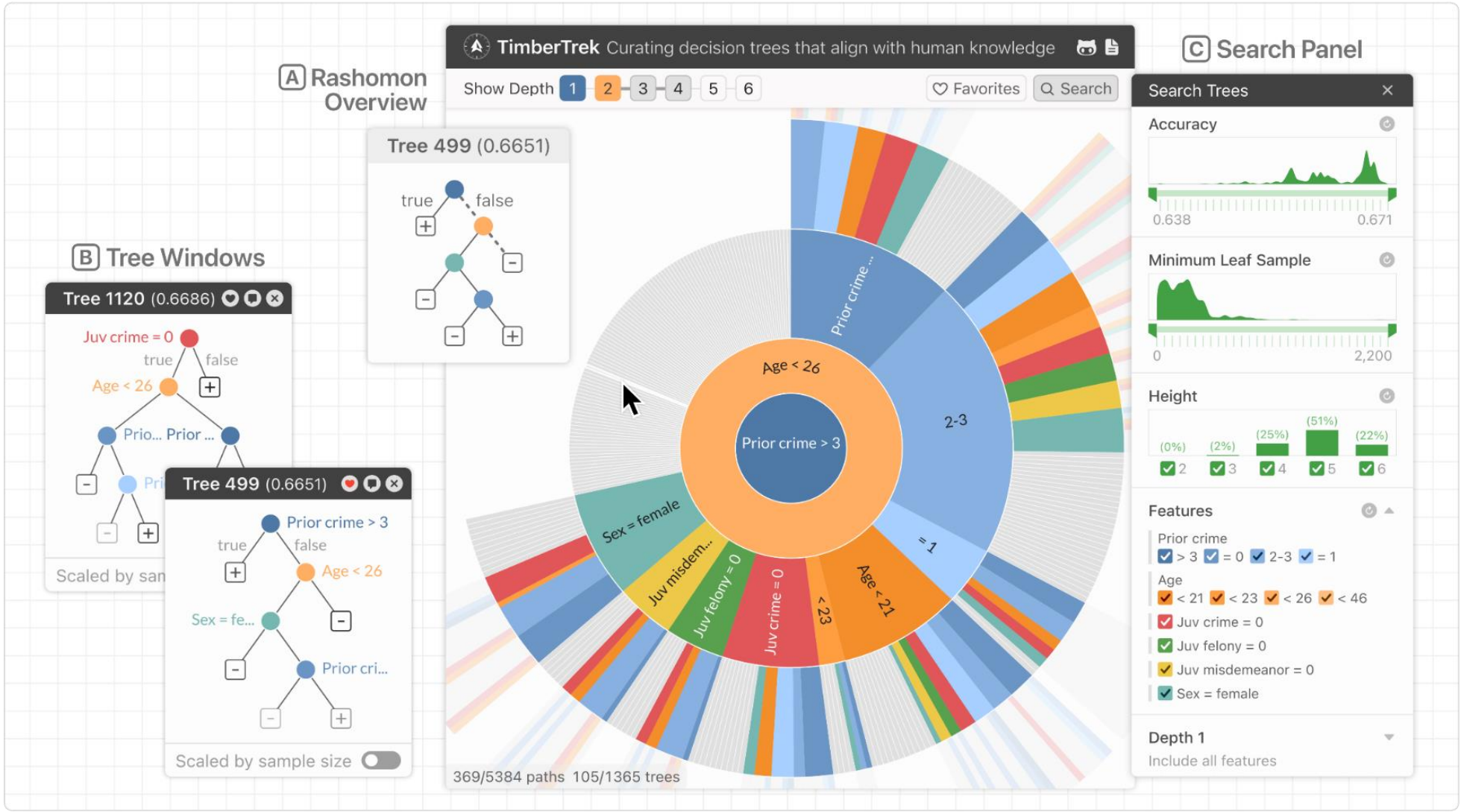
Model set representation efficiently extracts and stores trees

Find millions of good trees within minutes

TreeFARMS vs Baselines

In 46 seconds

Method \ Dataset	Monk2	COMPAS	Bar
	# trees in Rset	# trees in Rset	# trees in Rset
TreeFARMS	**105,782,431**	**265,176**	**5,096,307**
BART	3	0	187
Random Forest	0	196	1,239
CART + Sampling	7	139	1,006
GOSDT (optimal tree) + Sampling	15	6	12

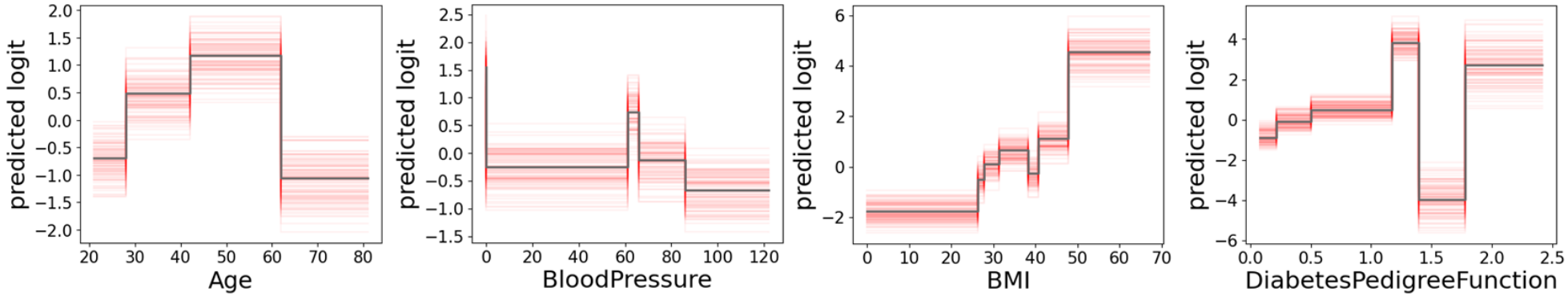


TimberTrek, 2022

Can we expand the Rashomon set to other interpretable model classes?

Other interpretable model classes

- Generalized additive models (Zhong et al., 2023)



- Scoring systems (Liu et al., 2022)

1.	Oval Shape	-2 points		...
2.	Irregular Shape	4 points	+	...
3.	Circumscribed Margin	-5 points	+	...
4.	Spiculated Margin	2 points	+	...
5.	Age $\geq$ 60	3 points	+	...
		SCORE	=	

SCORE	-7	-5	-4	-3	-2	-1
RISK	6.0%	10.6%	13.8%	17.9%	22.8%	28.6%

SCORE	0	1	2	3	4	$\geq$ 5
RISK	35.2%	42.4%	50.0%	57.6%	64.8%	71.4%

1.	Irregular Shape	2 points		...
2.	Circumscribed Margin	-2 points	+	...
3.	Spiculated Margin	1 point	+	...
4.	Age $\geq$ 30	2 points	+	...
5.	Age $\geq$ 60	1 point	+	...
		SCORE	=	

SCORE	-2	-1	0	1	2
RISK	2.3%	4.7%	9.5%	18.2%	32.0%

SCORE	3	4	5	6
RISK	50.0%	68.0%	81.8%	90.5%

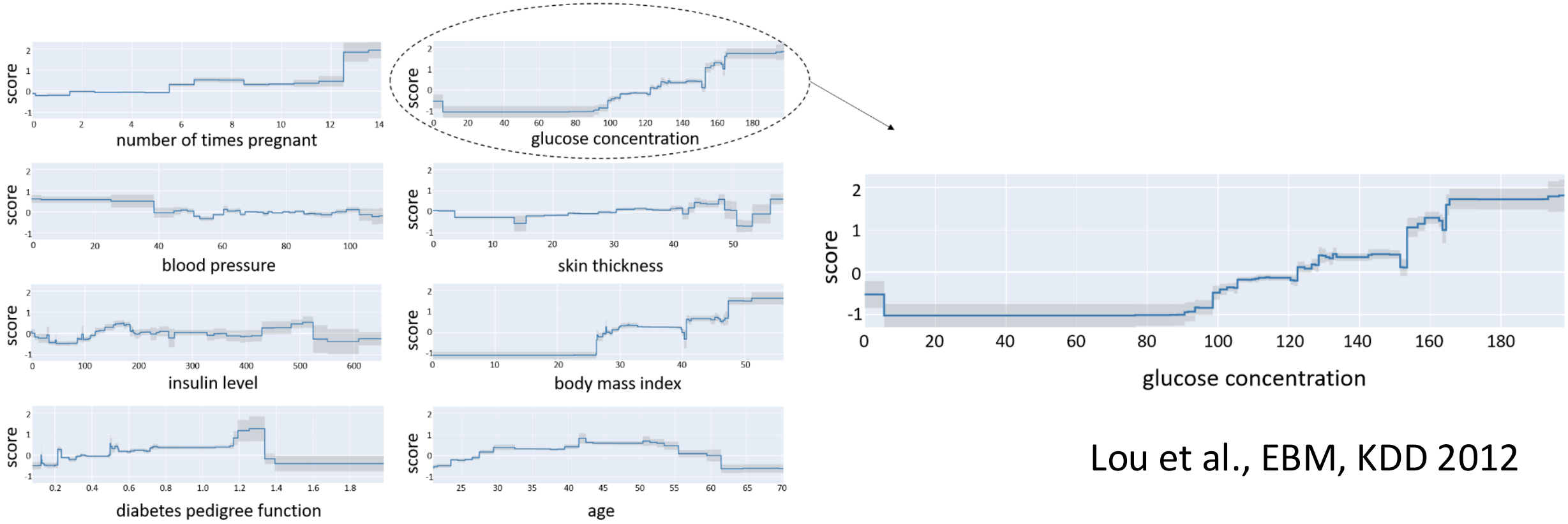
1.	Irregular Shape	5 points		...
2.	Circumscribed Margin	-5 points	+	...
3.	Microlobulated Margin	2 points	+	...
4.	Spiculated Margin	2 points	+	...
5.	Age $\geq$ 60	3 points	+	...
		SCORE	=	

SCORE	-5	-3	-2	-1	0	2	3
RISK	8.6%	14.6%	18.6%	23.5%	29.2%	42.7%	50.0%

SCORE	4	5	7	8	9	10	12
RISK	57.3%	64.3%	76.5%	81.4%	85.4%	88.7%	93.4%

Generalized additive models (GAMs)

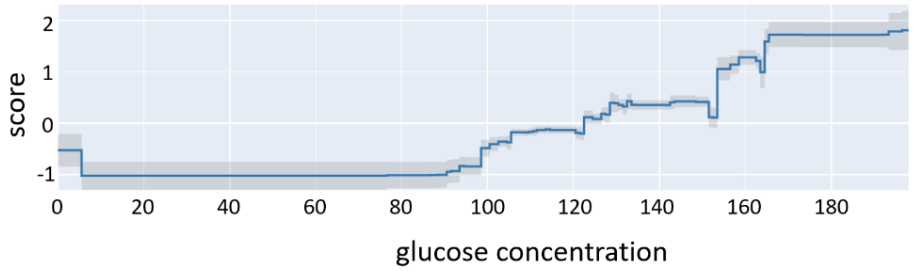
$$g(E[\boldsymbol{y}]) = w_0 + f_1(\boldsymbol{X}_{\cdot,1}) + f_2(\boldsymbol{X}_{\cdot,2}) + \cdots + f_p(\boldsymbol{X}_{\cdot,p})$$



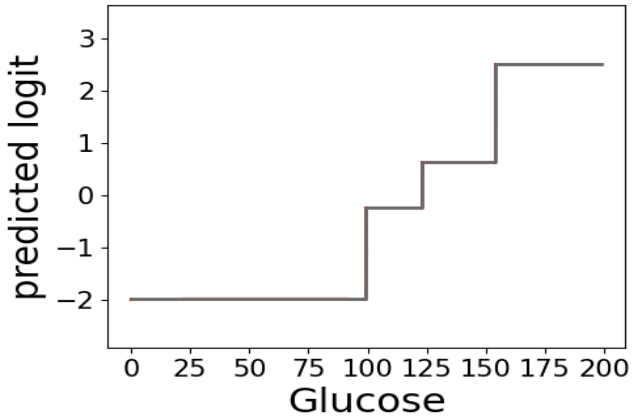
Lou et al., EBM, KDD 2012

Generalized additive models (GAMs)

$$g(E[\boldsymbol{y}]) = w_0 + f_1(\boldsymbol{X}_{\cdot,1}) + f_2(\boldsymbol{X}_{\cdot,2}) + \cdots + f_p(\boldsymbol{X}_{\cdot,p})$$



Lou et al., EBM, KDD 2012



Liu et al., Fastsparse, AISTATS 2022

- very interpretable
  - sum of univariate component functions
  - 2d-plot to show contributions from each feature
- very powerful
  - each component function can be non-linear
- can be trained using boosting or other ML techniques
- generally, few pairwise interaction terms
- great for continuous features, not good for categorical features
- doesn't easily handle missing data or multiclass

Rashomon set of generalized additive models

Exploring and Interacting with the Set of Good Sparse Generalized Additive Models

Chudi Zhong[1*] Zhi Chen[1*] Jiachang Liu[1] Margo Seltzer[2] Cynthia Rudin[1]

NeurIPS 2023

Approximate the Rashomon set by an ellipsoid

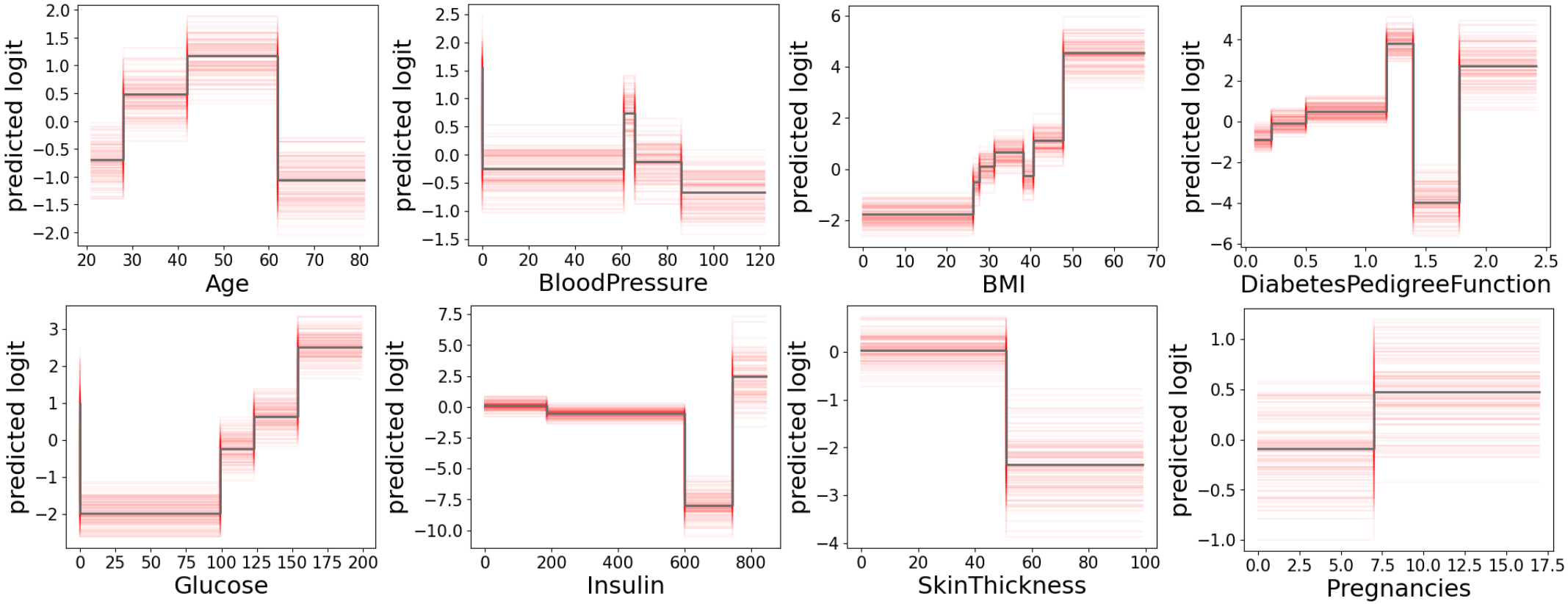
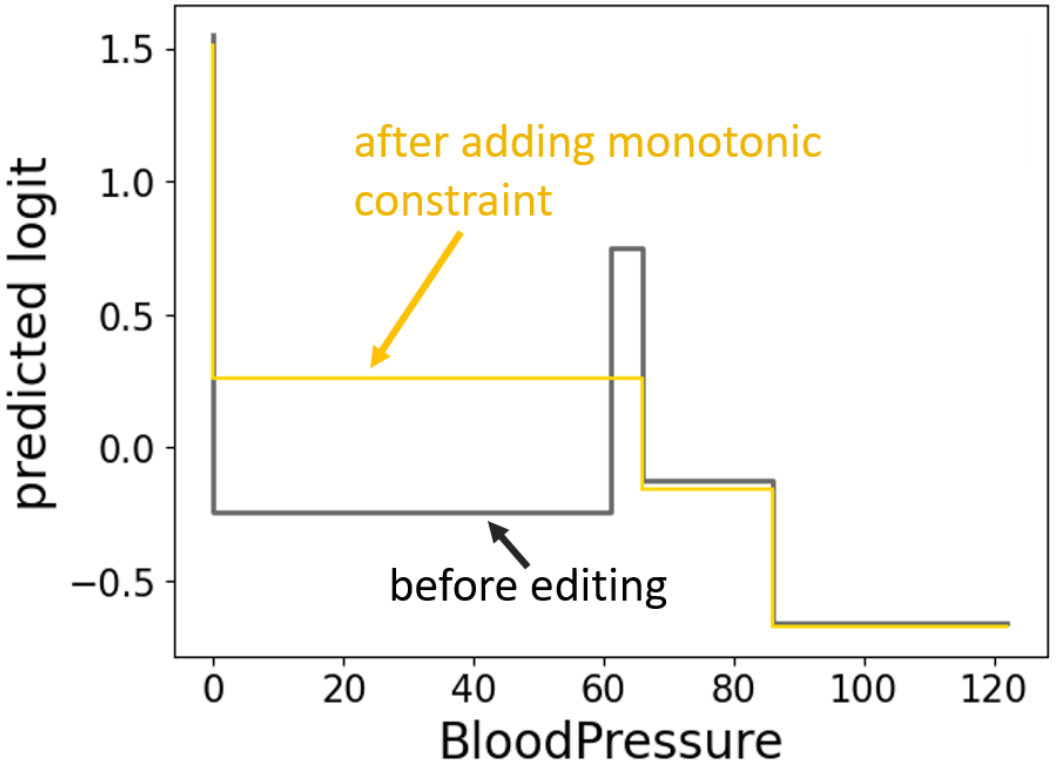


Fig: 100 different shape functions sampled from the approximated Rashomon set on the Diabetes dataset. All of them perform almost equally well in prediction.

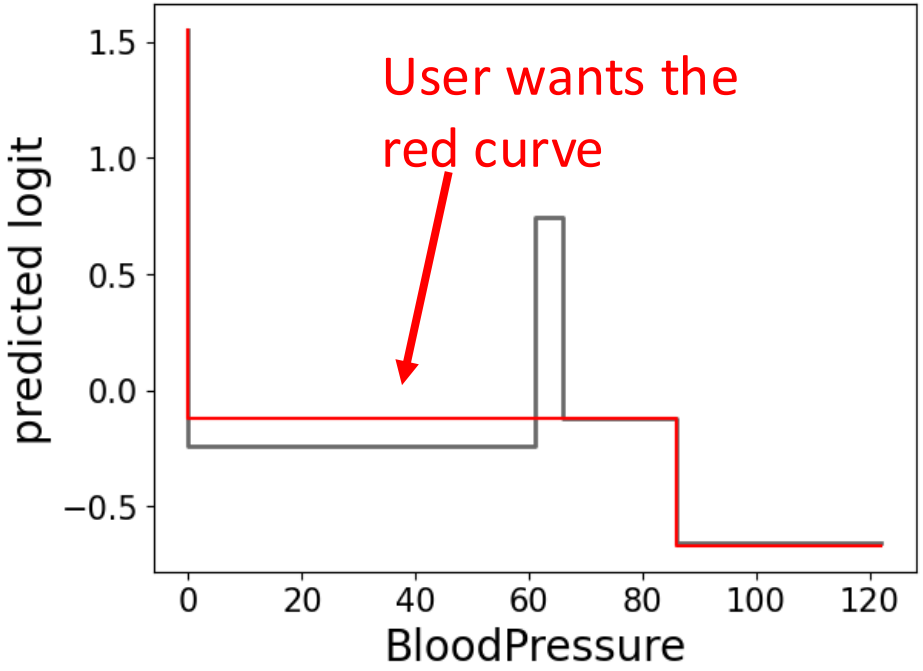
Find a model with monotonic constraints

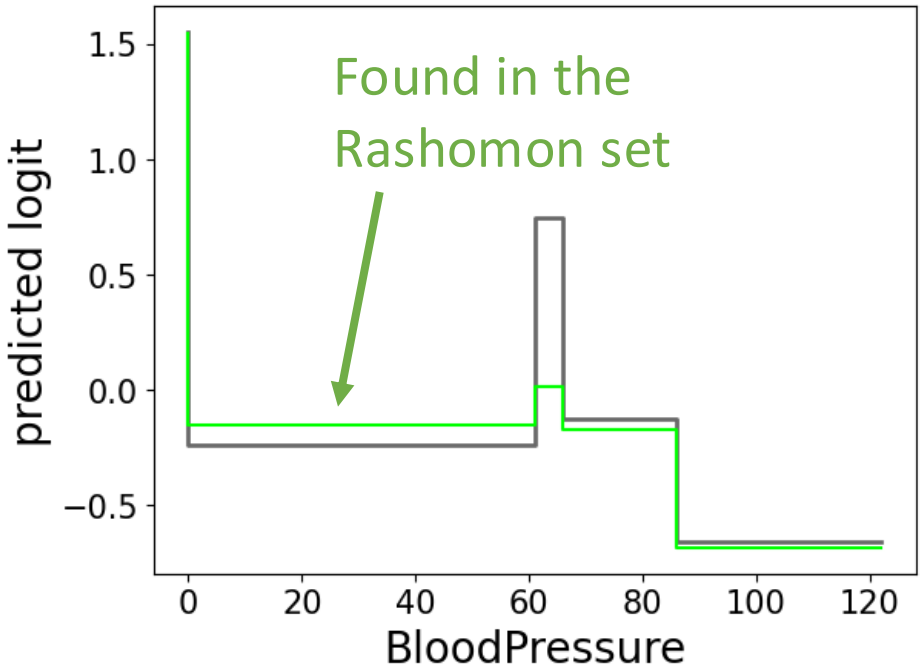
Solve a quadratic programming problem with linear constraints



Find a model closest to users' requests

Solve a quadratic programming problem with quadratic constraint





Risk scores

1. Oval Shape	-2 points		…
2. Irregular Shape	4 points	+	…
3. Circumscribed Margin	-5 points	+	…
4. Spiculated Margin	2 points	+	…
5. Age $\geqslant$ 60	3 points	+	…
	SCORE	=	

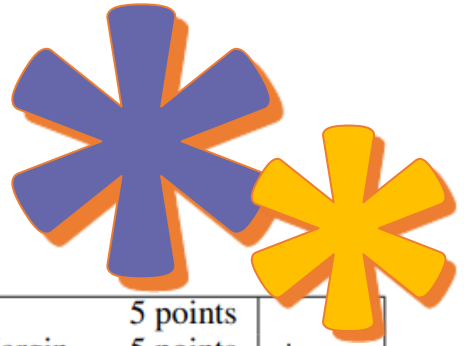
SCORE	-7	-5	-4	-3	-2	-1
RISK	6.0%	10.6%	13.8%	17.9%	22.8%	28.6%

SCORE	0	1	2	3	4	$\geqslant$5
RISK	35.2%	42.4%	50.0%	57.6%	64.8%	71.4%

FasterRisk, Liu et al., 2022

- are simple additive models
- assign integer points to features
- easy to compute by hand
- widely used in medicine, finance, and criminal justice
- highly interpretable and transparent
- Used to be developed by doctors but nowadays with ML algorithms

A pool of good risk scores from FasterRisk



1.	Oval Shape	-2 points		...
2.	Irregular Shape	4 points	+	...
3.	Circumscribed Margin	-5 points	+	...
4.	Spiculated Margin	2 points	+	...
5.	Age $\geq$ 60	3 points	+	...
		SCORE	=	

SCORE	-7	-5	-4	-3	-2	-1
RISK	6.0%	10.6%	13.8%	17.9%	22.8%	28.6%

SCORE	0	1	2	3	4	$\geq$ 5
RISK	35.2%	42.4%	50.0%	57.6%	64.8%	71.4%

1.	Irregular Shape	2 points		...
2.	Circumscribed Margin	-2 points	+	...
3.	Spiculated Margin	1 point	+	...
4.	Age $\geq$ 30	2 points	+	...
5.	Age $\geq$ 60	1 point	+	...
		SCORE	=	

SCORE	-2	-1	0	1	2
RISK	2.3%	4.7%	9.5%	18.2%	32.0%

SCORE	3	4	5	6
RISK	50.0%	68.0%	81.8%	90.5%

1.	Irregular Shape	5 points		...
2.	Circumscribed Margin	-5 points	+	...
3.	Microlobulated Margin	2 points	+	...
4.	Spiculated Margin	2 points	+	...
5.	Age $\geq$ 60	3 points	+	...
		SCORE	=	

SCORE	-5	-3	-2	-1	0	2	3
RISK	8.6%	14.6%	18.6%	23.5%	29.2%	42.7%	50.0%
SCORE	4	5	7	8	9	10	12
RISK	57.3%	64.3%	76.5%	81.4%	85.4%	88.7%	93.4%

1.	Irregular Shape	4 points		...
2.	Circumscribed Margin	-5 points	+	...
3.	SpiculatedMargin	2 points	+	...
4.	Age $\geq$ 45	1 point	+	...
5.	Age $\geq$ 60	3 points	+	...
		SCORE	=	

SCORE	-5	-4	-3	-2	-1	0	1	2
RISK	7.3%	9.7%	12.9%	16.9%	21.9%	27.8%	34.6%	42.1%
SCORE	3	4	5	6	7	8	9	10
RISK	50.0%	57.9%	65.4%	72.2%	78.1%	83.1%	87.1%	90.3%

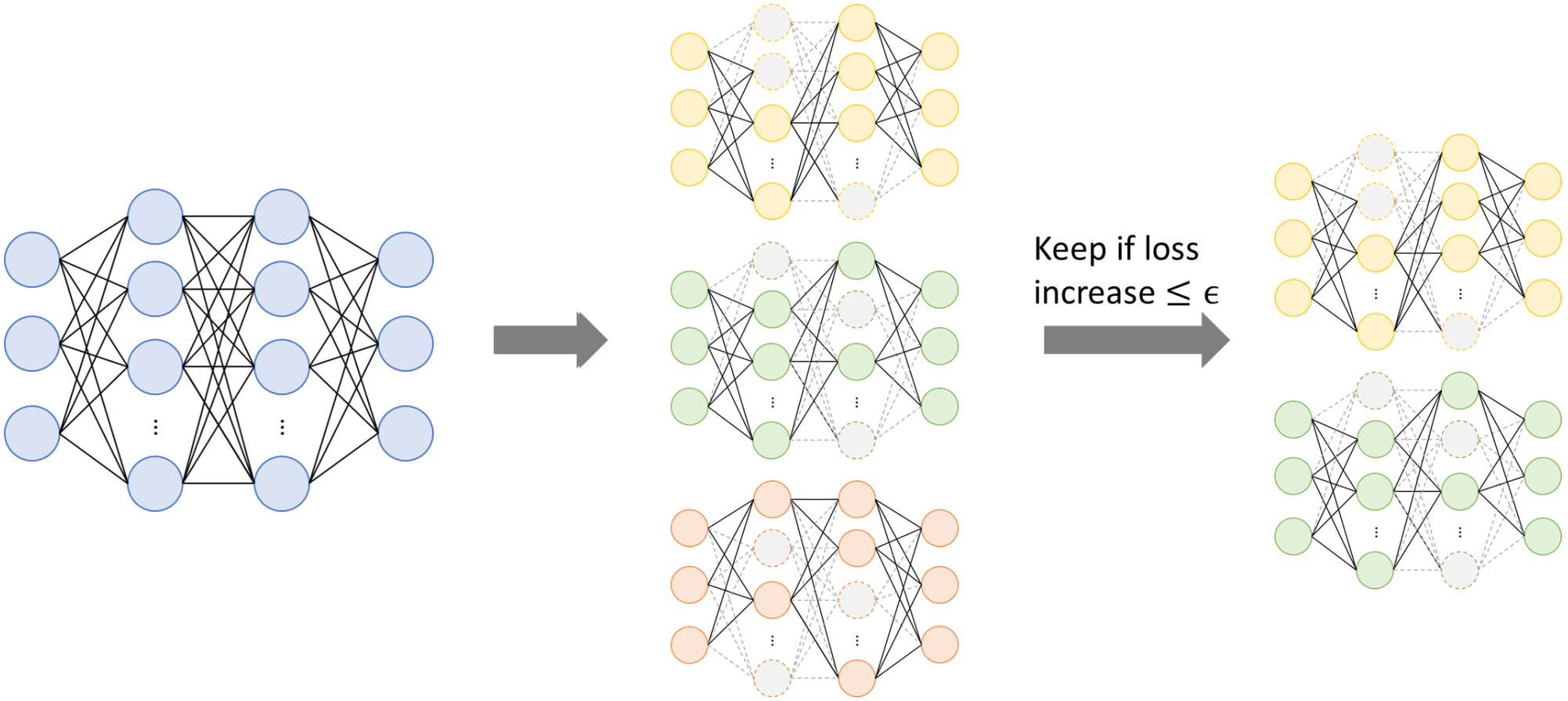
1.	Irregular Shape	4 points		...
2.	Circumscribed Margin	-5 points	+	...
3.	Obscure Margin	-1 points	+	...
4.	Spiculated Margin	2 points	+	...
5.	Age $\geq$ 60	3 points	+	...
		SCORE	=	

SCORE	-6	-5	-4	-3	-2	-1	0	1
RISK	6.8%	9.2%	12.3%	16.3%	21.3%	27.3%	34.2%	41.9%
SCORE	2	3	4	5	6	7	8	9
RISK	50.0%	58.1%	65.8%	72.7%	78.7%	83.7%	87.7%	90.8%

What about the Rashomon set of neural networks?

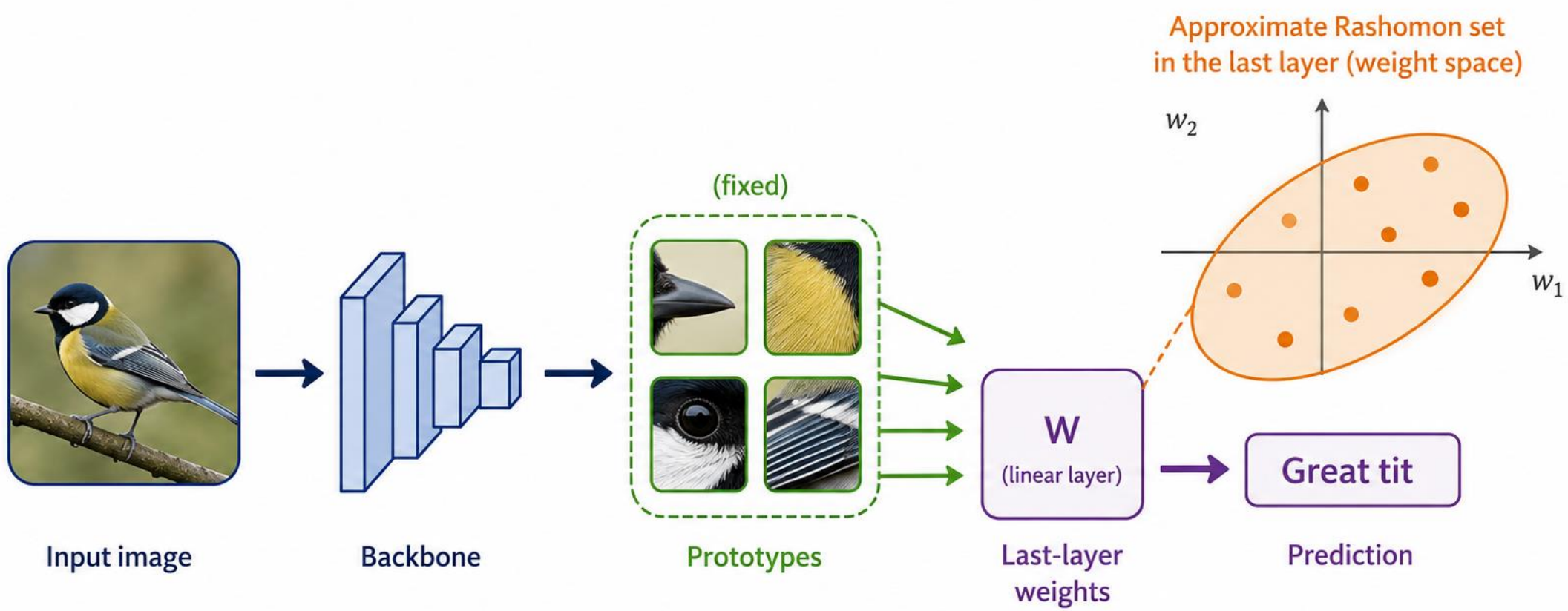
Rashomon set of neural networks

- Hsu et al., 2024: find an empirical set through dropout



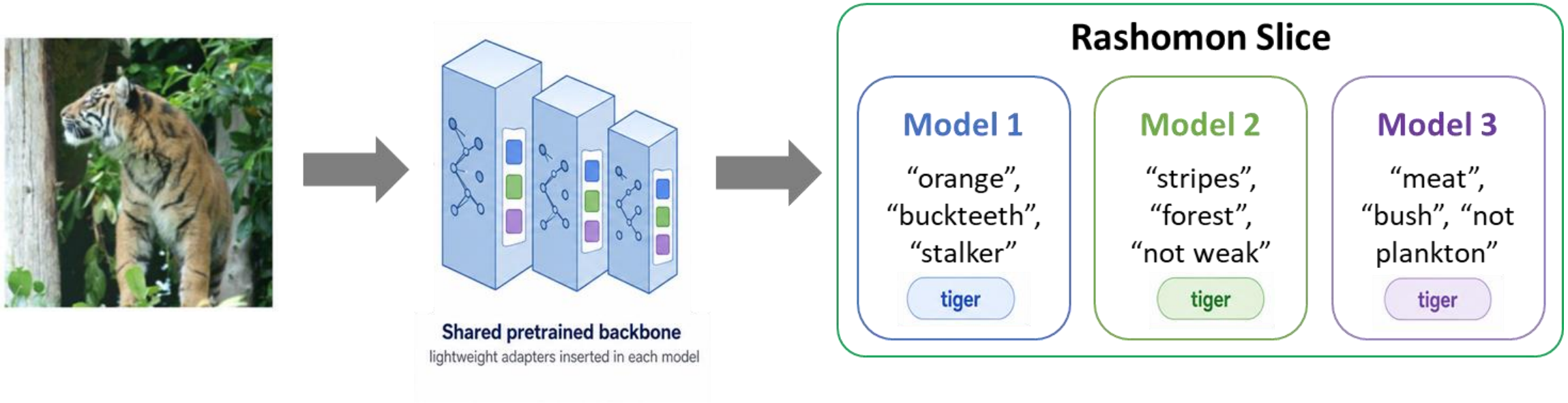
Rashomon set of neural networks

- Donnelly et al., 2025: approximate the Rashomon set for ProtoPNet

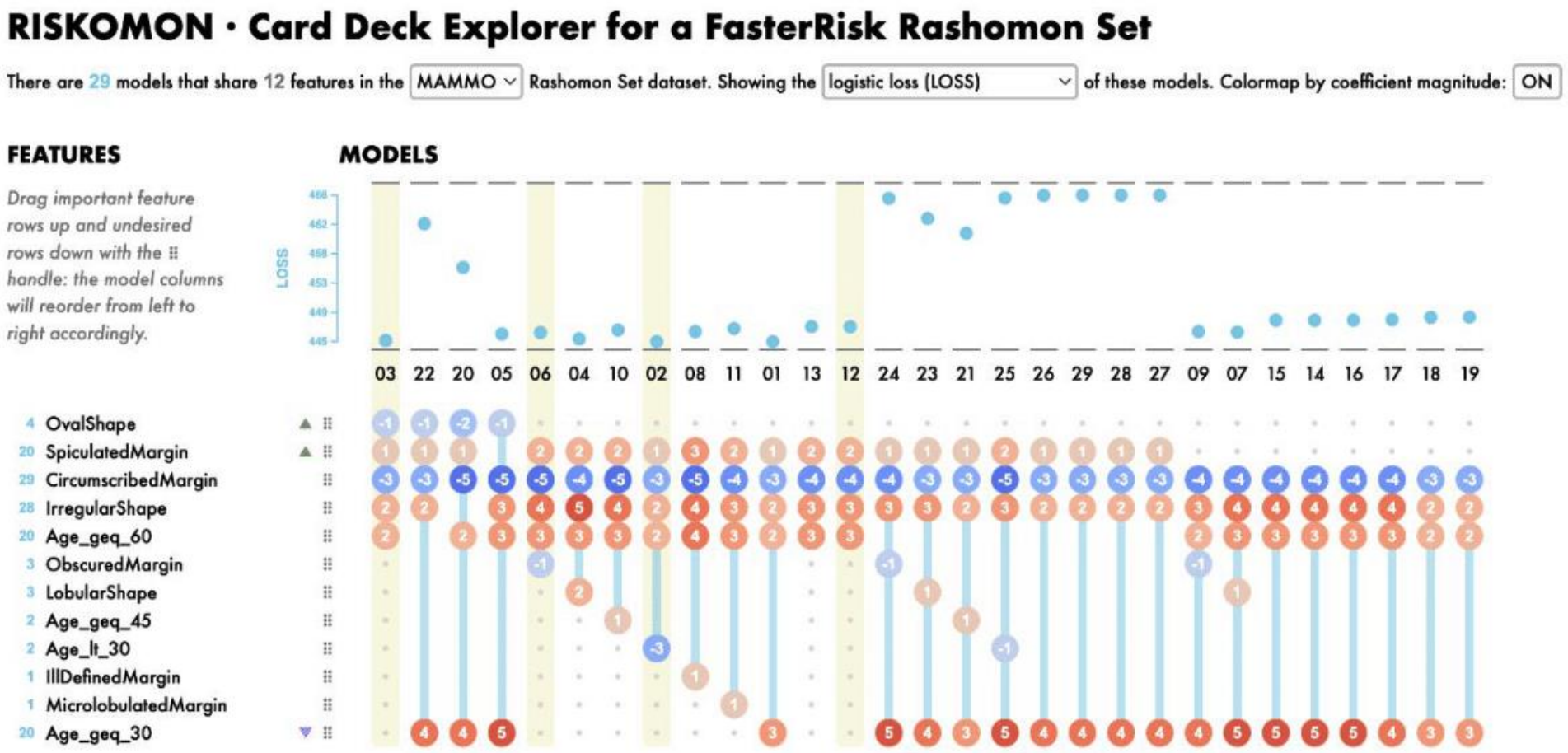


Rashomon set of neural networks

- Feng et al., 2025: find the Rashomon slice for concept-based models



Interactive session - Riskomon



https://riskomon.netlify.app/

Oddo et al., 2024

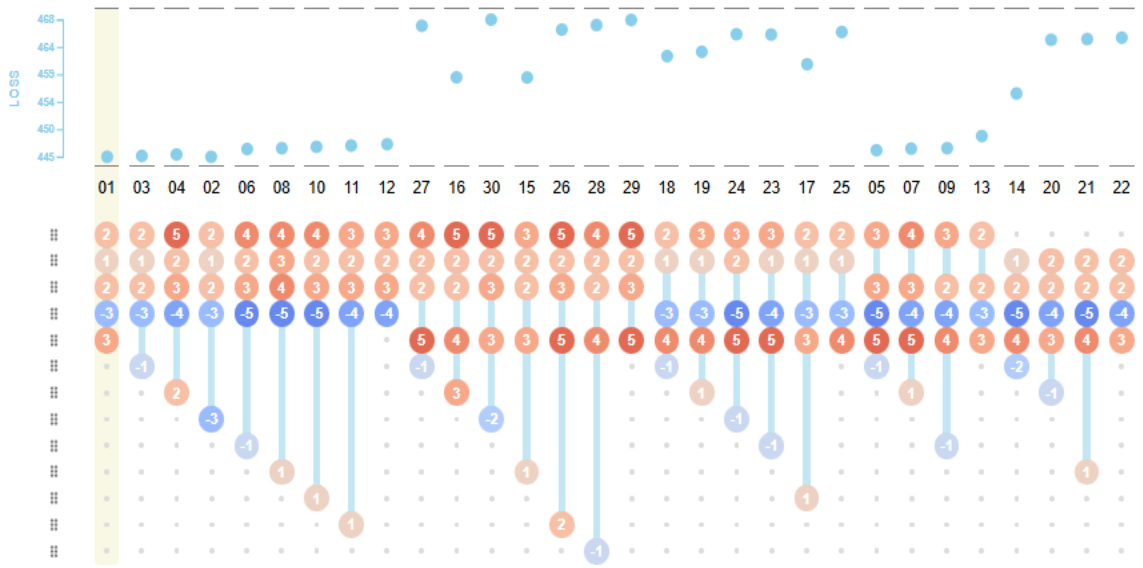
RISKOMON · Card Deck Explorer for a FasterRisk Rashomon Set (paper, source code)

There are 30 models that share 13 features in the MAMMO Rashomon Set dataset, here showing the logistic loss (LOSS) of these models, with colormap by coefficient magnitude turned ON

FEATURES

Drag important feature rows up and undesired rows down with the ⠿ handle: the model columns will reorder from left to right accordingly.

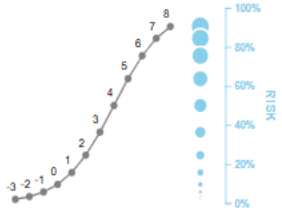
MODELS



CARDS

Model 01 Card

IrregularShape	+2
CircumscribedMargin	-3
SpiculatedMargin	+1
Age_geq_30	+3
Age_geq_60	+2



LOSS: 445.0 ACC: 0.80% AUC: 0.855 MAX RISK: 1.92%

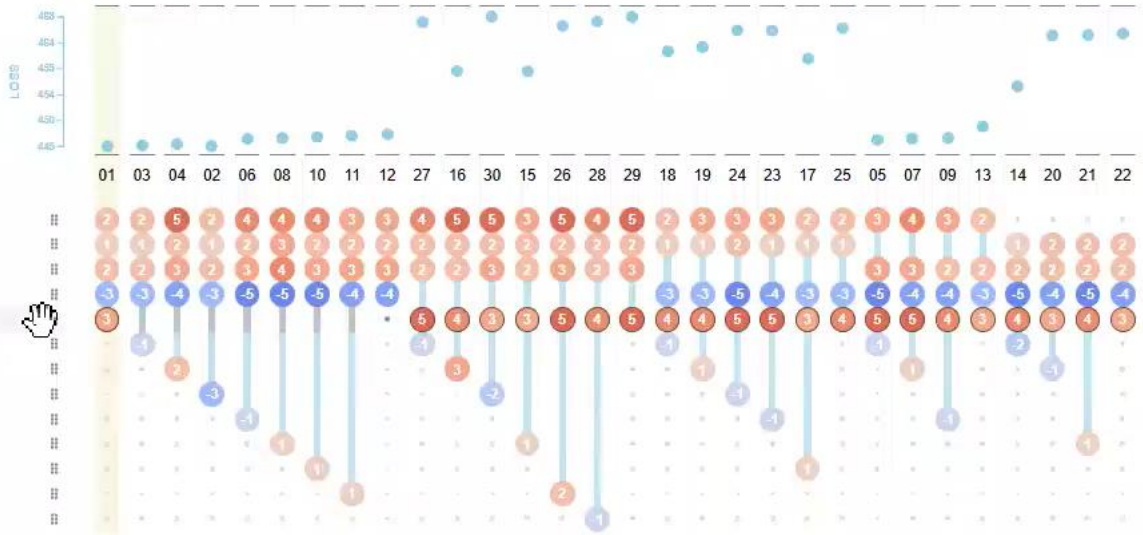
RISKOMON · Card Deck Explorer for a FasterRisk Rashomon Set (paper, source code)

There are 30 models that share 13 features in the MAMMO Rashomon Set dataset, here showing the logistic loss (LOSS) of these models, with colormap by coefficient magnitude turned ON

FEATURES

Drag important feature rows up and undesired rows down with the ⠿ handle: the model columns will reorder from left to right accordingly.

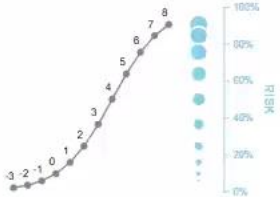
MODELS



CARDS

Model 01 Card

IrregularShape	+2
CircumscribedMargin	-3
SpiculatedMargin	+1
Age_geq_30	+3
Age_geq_60	+2



LOSS: 445.0 ACC: 0.80% AUC: 0.855 MAX RISK: 1.92%

Future steps and outro

What can we do next?

Ongoing computing/measuring problems

- What causes the Rashomon Effect? What determines its size?
  - how much multiplicity comes from (i) feature redundancy / collinearity, (ii) label noise, …
- How do we find/represent Rashomon sets at scale?
  - Especially for more complex model classes such as LLMs or VLMs
- How to display the Rashomon set for human-model interaction?

How to ease interaction and construction of model multiplicity

Scope and scale

- For LLMs the Rashomon set isn't even well-defined
  - (what object? what utility?), yet that's where consequential systems are heading.
- Beyond classification
  - The theory is about mostly about prediction real systems do ranking, allocation, generation, and agentic action sequences
- Measuring ecosystem-level multiplicity (monoculture)
  - Individual arbitrariness compounds when everyone ships similar models or shares a base model

How to scale it and measure where it matters

(Some) Accountability questions

- What does a good-faith search look?
  - The same freedom that lets you find a less-discriminatory model lets you cherry-pick one that looks compliant.
- Standards for multiplicity disclosure
  - If we require deployers to report that alternatives existed (procurement, impact assessments), what must the report contain, and how is it verified?
- Operationalizing contestability
  - If your outcome flipped because of which model was chosen, what can you demand? No legal or institutional mechanism exists.

How to operationalize the Rashomon set

What can we do right now

1. **Require multiplicity reporting**
   not just one accuracy number
2. **Ask for alternatives**
   simpler, less discriminatory, more stable
3. **Document the choice**
   criteria, rejected alternatives, stakeholders
4. **Audit individual arbitrariness**
   who changes outcomes across good models?
5. **Design contestation around alternatives**
   appeals can ask whether a viable alternative changes the outcome
6. **Ask for disclosure**
   vendors should disclose model-selection degrees of freedom

Tutorial slides:

STAY IN TOUCH

lesia.semenova@rutgers.edu
chudi@unc.edu