An accuracy-interpretability tradeoff?

**Why less can be more
and how to find it**

- Disclaimer and preliminaries
- The many problems of black boxes
- Simple models for the win
- When is it worth the effort?
- Coda: risks of interpretability

A fundamental distinction

- **Interpretable ML**: “not a black box”
- **Explainable ML**: use a proxy to explain a black box

A fundamental distinction

- **Interpretable ML**: “not a black box”
- **Explainable ML**: use a proxy to explain a black box

> A machine learning model is **interpretable if it is constrained** to make it “easier to understand” for user X

A fundamental distinction

- **Interpretable ML**: “not a black box”
- **Explainable ML**: use a proxy to explain a black box

> A machine learning model is **interpretable if it is constrained** to make it “easier to understand” for user X

- Many choices: sparsity, degree of non-linearity, low-order interactions…
- **Domain**- and **user-specific**

Our goal

To show why we should

prefer interpretability over explanations,

to see examples where

this does not necessarily incur a performance penalty,

and to look at some

theory supporting this preference.

A chamber of horrors

- Bad medical diagnosis / screening / treatments
- Unjust bail / parole decisions
- Wrong loan / credit decisions

...

A chamber of horrors

- Bad medical diagnosis / screening / treatments
- Unjust bail / parole decisions
- Wrong loan / credit decisions

  ...

Because of **typos** (!)

Stop explaining black box machine learning models for high stakes decisions... [14]

A chamber of horrors

- Bad medical diagnosis / screening / treatments
- Unjust bail / parole decisions
- Wrong loan / credit decisions

  ...

Because of **typos** (!)

Stop explaining black box machine learning models for high stakes decisions... [14]

Because of **bogus explanations**

Incorrect recommendations with easily interpretable explanations lead to reduction in treatment selection accuracy [11]

A chamber of horrors

- Bad medical diagnosis / screening / treatments
- Unjust bail / parole decisions
- Wrong loan / credit decisions

...

Because of **typos** (!)

Stop explaining black box machine learning models for high stakes decisions... [14]

Because of **bogus explanations**

Incorrect recommendations with easily interpretable explanations lead to reduction in treatment selection accuracy [11]

(...)

Black boxes and their “explanations”

- XAI: posthoc **proxies** for black boxes, e.g. LIME

Black boxes and their “explanations”

- XAI: posthoc **proxies** for black boxes, e.g. LIME
- *Explaining* a BB: now trust **two** models (and the data)

Black boxes and their “explanations”

- XAI: posthoc **proxies** for black boxes, e.g. LIME
- *Explaining* a BB: now trust **two** models (and the data)
- Proxies won’t be 100% accurate by definition

Black boxes and their "explanations"

- XAI: posthoc **proxies** for black boxes, e.g. LIME
- *Explaining* a BB: now trust **two** models (and the data)
- Proxies won't be 100% accurate by definition
- At best: ineffective/nonsensical, e.g. different domains for features [1]

Black boxes and their "explanations"

- XAI: posthoc **proxies** for black boxes, e.g. LIME
- *Explaining* a BB: now trust **two** models (and the data)
- Proxies won't be 100% accurate by definition
- At best: ineffective/nonsensical, e.g. different domains for features [1]
- At worst: detrimental (**under/overreliance**, see later) [13]

Black boxes and their "explanations"

- XAI: posthoc **proxies** for black boxes, e.g. LIME
- *Explaining* a BB: now trust **two** models (and the data)
- Proxies won't be 100% accurate by definition
- At best: ineffective/nonsensical, e.g. different domains for features [1]
- At worst: detrimental (**under/overreliance**, see later) [13]
- BBs hinder the **cyclic** nature of ML development

  Collect data, pre-process & model, evaluate, rinse, repeat

  $\Rightarrow$ Better understanding of the model leads to better models

Black boxes and their “explanations”

- XAI: posthoc **proxies** for black boxes, e.g. LIME
- *Explaining* a BB: now trust **two** models (and the data)
- Proxies won’t be 100% accurate by definition
- At best: ineffective/nonsensical, e.g. different domains for features [1]
- At worst: detrimental (**under/overreliance**, see later) [13]
- BBs hinder the **cyclic** nature of ML development

  Collect data, pre-process & model, evaluate, rinse, repeat

  $\Rightarrow$ Better understanding of the model leads to better models

But we probably don’t need BBs anyway…

A zoo of interpretable models

- Rule lists [2, 19, 15]
- Sparse scoring systems [17, 16]
- Sparse decision trees [10, 12, 20]
- Hierarchical models [5]
- Multilevel Bayesian modeling [8]
- Prototypes and concepts [4, 9, 7]

Rule lists

$$\min_{d \in \text{rule lists}} \underbrace{\hat{R}(d;S)}_{\text{misclassification error}} + \lambda \underbrace{|d|}_{\text{length of tuple } d}$$

Rules are tuples of **associations**, $r_k = p_k \rightarrow q_k$, followed by a default rule r_0

if $(age = 18-20)$ **and** $(sex = male)$ **then predict** *yes*
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** *yes*
else if $(priors > 3)$ **then predict** *yes*
else predict *no*

if p_1 **then predict** q_1
else if p_2 **then predict** q_2
else if p_3 **then predict** q_3
else predict q_0

Rule lists

$$\min_{d \in \text{rule lists}} \underbrace{\hat{R}(d; S)}_{\text{misclassification error}} + \lambda \underbrace{|d|}_{\text{length of tuple } d}$$

Rules are tuples of **associations**, $r_k = p_k \rightarrow q_k$, followed by a default rule r_0

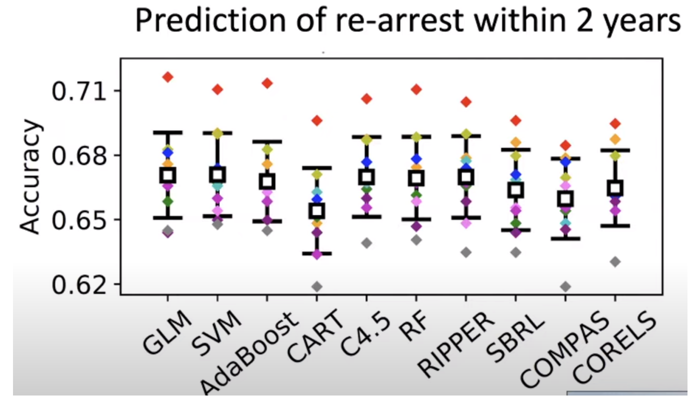
if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

if p_1 **then predict** q_1
else if p_2 **then predict** q_2
else if p_3 **then predict** q_3
else predict q_0

CORELS matches / beats COMPAS with **three** rules [2]

Branch & bound to search among pre-mined rules

Limiting factor: # of features (~ 30)



Rule lists

$$\min_{d \in \text{rule lists}} \underbrace{\hat{R}(d;S)}_{\text{misclassification error}} + \lambda \underbrace{|d|}_{\text{length of tuple } d}$$

Rules are tuples of **associations**, $r_k = p_k \rightarrow q_k$, followed by a default rule r_0

if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

if p_1 **then predict** q_1
else if p_2 **then predict** q_2
else if p_3 **then predict** q_3
else predict q_0

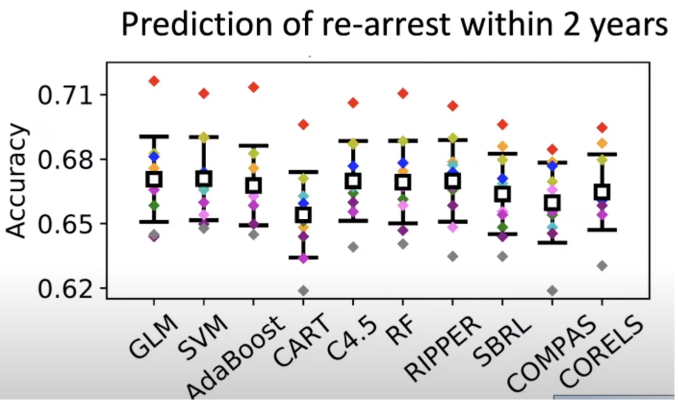
CORELS matches / beats COMPAS with **three** rules [2]

Branch & bound to search among pre-mined rules

Limiting factor: # of features (~ 30)

"Optimal decision lists using SAT" [19]

Learns rules, off-the-shelf solver, perfect or sparse



Sparse scoring systems

- 2HELPS2B for seizure risk prediction. Equal accuracy to SoTA, doctors can decide to ignore recommendations, can recalibrate with new variables

1.	Any cEEG Pattern with Frequency > 2 **Hz**	1 point		…
2.	**E**pileptiform Discharges	1 point	+	…
3.	Patterns include **L**PD or LRDA or BIPD	1 point	+	…
4.	**P**atterns Superimposed with Fast, or Sharp Activity	1 point	+	…
5.	Prior **S**eizures	1 point	+	…
6.	**B**rief Rhythmic Discharges	2 points	+	
		SCORE	=	

SCORE	0	1	2	3	4	5	6+
RISK	<5%	12%	27%	50%	73%	88 %	>95%

5-CV mean test CAL/AUC of 2.7%/0.819

Sparse scoring systems

- 2HELPS2B for seizure risk prediction. Equal accuracy to SoTA, doctors can decide to ignore recommendations, can recalibrate with new variables

1.	Any cEEG Pattern with Frequency > 2 **Hz**	1 point		…
2.	**E**pileptiform Discharges	1 point	+	…
3.	Patterns include **L**PD or LRDA or BIPD	1 point	+	…
4.	**P**atterns Superimposed with Fast, or Sharp Activity	1 point	+	…
5.	Prior **S**eizures	1 point	+	…
6.	**B**rief Rhythmic Discharges	2 points	+	
		SCORE	=	

SCORE	0	1	2	3	4	5	6+
RISK	<5%	12%	27%	50%	73%	88 %	>95%

5-CV mean test CAL/AUC of 2.7%/0.819

- Clinical decision, risk assessment, infrastructure reliability, repair crews… (HITL)

Sparse scoring systems

- 2HELPS2B for seizure risk prediction. Equal accuracy to SoTA, doctors can decide to ignore recommendations, can recalibrate with new variables

1.	Any cEEG Pattern with Frequency > 2 **Hz**	1 point		…
2.	**E**pileptiform Discharges	1 point	+	…
3.	Patterns include **L**PD or LRDA or BIPD	1 point	+	…
4.	**P**atterns Superimposed with Fast, or Sharp Activity	1 point	+	…
5.	Prior **S**eizures	1 point	+	…
6.	**B**rief Rhythmic Discharges	2 points	+	
		SCORE	=	

SCORE	0	1	2	3	4	5	6+
RISK	<5%	12%	27%	50%	73%	88 %	>95%

5-CV mean test CAL/AUC of 2.7%/0.819

- Clinical decision, risk assessment, infrastructure reliability, repair crews… (HITL)

- RISK-SLIM: sparse, linear, small integer coefficients, calibrated, high rank accuracy [16]

$$\min_{\theta} \underbrace{\hat{R}(\theta; S)}_{\text{logistic loss}} + \lambda \, \|\theta\|_0, \ \text{s.t.} \ \theta \ \textbf{admissible} \text{ and in } \mathbb{Z}^{d+1}$$

Sparse scoring systems (contd.)

About the admissible set:

Model Requirement	**Example**
Feature Selection	Choose between 5 to 10 total features
Group Sparsity	Include either *male* or *female* in the model but not both
Optimal Thresholding	Use at most 3 thresholds for a set of indicator variables: $\sum_{k=1}^{100} \mathbb{1}\left[age \leq k\right] \leq 3$
Logical Structure	If *male* is in model, then include *hypertension* or $bmi \geq 30$
Side Information	Predict $\Pr\left(y = +1 \vert \boldsymbol{x}\right) \geq 0.90$ when $male = \text{TRUE}$ and $hypertension = \text{TRUE}$

Table 1: Model requirements that can be addressed by adding operational constraints to RiskSlimMINLP.

Sparse trees

- Classical tree algorithms: top-down, greedy back-tracking pruning (C4.5, CART)

Sparse trees

- Classical tree algorithms: top-down, greedy back-tracking pruning (C4.5, CART)
- (**G**eneralised) **O**ptimal **S**parse **D**ecision **T**rees [10, 12]

$$\min_{\tau \in \text{trees}} \underbrace{\hat{R}(\tau; S)}_{\text{misclassification}} + \lambda \underbrace{|\tau|}_{\#\text{leaves in } \tau}$$

Certificate of optimality: no better training performance possible at sparsity level

Sparse trees

- Classical tree algorithms: top-down, greedy back-tracking pruning (C4.5, CART)
- (**G**eneralised) **O**ptimal **S**parse **D**ecision **T**rees [10, 12]

$$\min_{\tau \in \text{trees}} \underbrace{\hat{R}(\tau; S)}_{\text{misclassification}} + \lambda \underbrace{|\tau|}_{\#\text{leaves in } \tau}$$

Certificate of optimality: no better training performance possible at sparsity level

Branch & bound with: strong analytical bounds, caching, leaf representation, fast impl.

Followup GOSDT: continuous variables and imbalanced data

Sparse trees

- Classical tree algorithms: top-down, greedy back-tracking pruning (C4.5, CART)
- (**G**eneralised) **O**ptimal **S**parse **D**ecision **T**rees [10, 12]

$$\min_{\tau \in \text{trees}} \underbrace{\hat{R}(\tau; S)}_{\text{misclassification}} + \lambda \underbrace{|\tau|}_{\#\text{leaves in } \tau}$$

Certificate of optimality: no better training performance possible at sparsity level

Branch & bound with: strong analytical bounds, caching, leaf representation, fast impl.

Followup GOSDT: continuous variables and imbalanced data

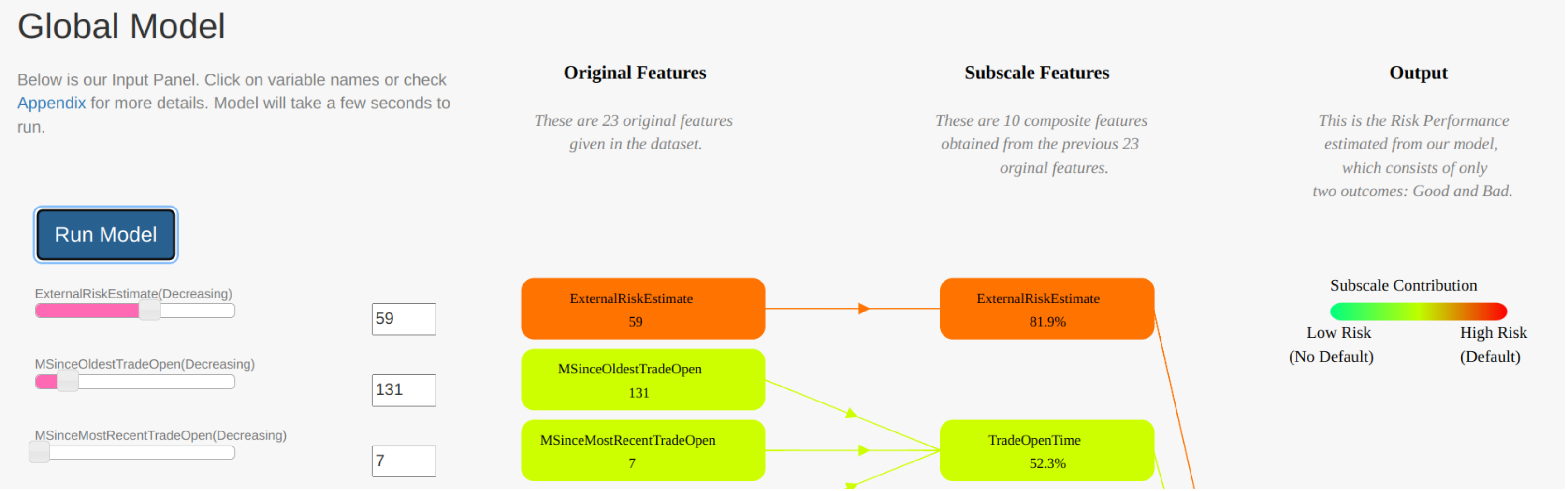
- Followup: Optimal sparse *regression* trees [20]

Hierarchical models

- Aggregation of simple models (e.g. stacked logistic regression)
- Example: two-layer additive risk model

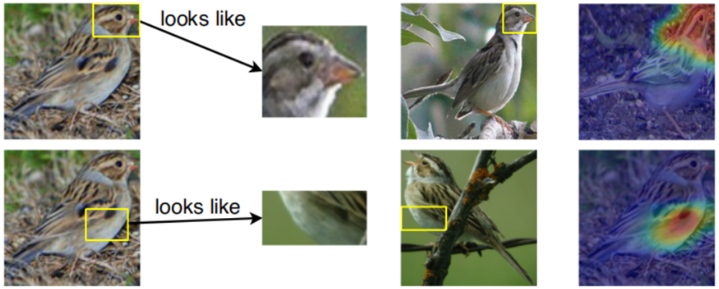
  Learned "subscale features"

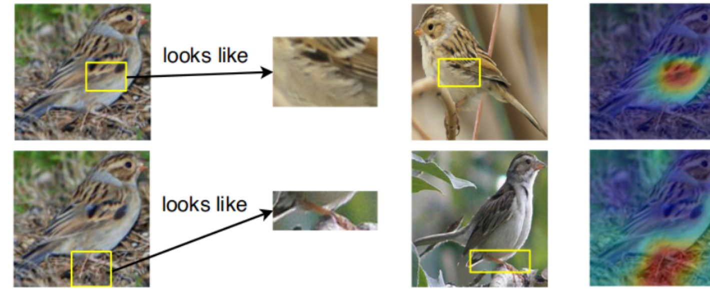
dukedatasciencefico.cs.duke.edu [5]



Example-based reasoning

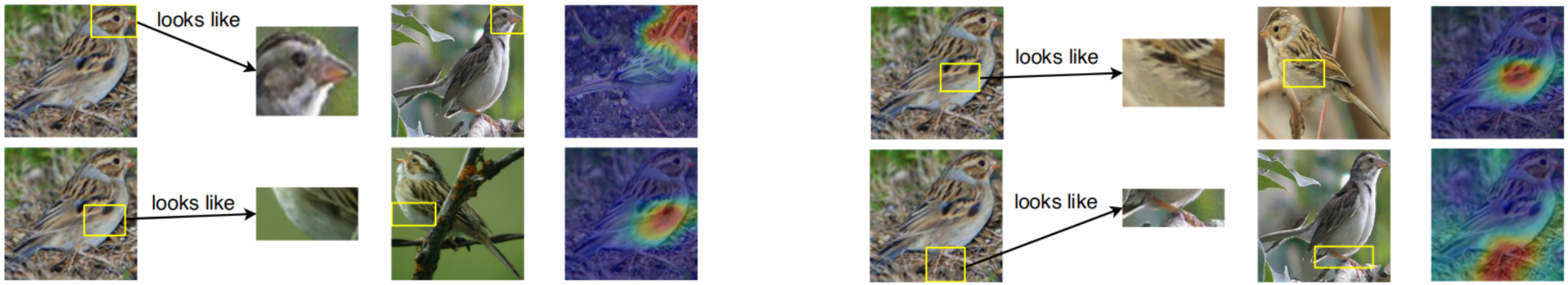
- Prototype images [4]



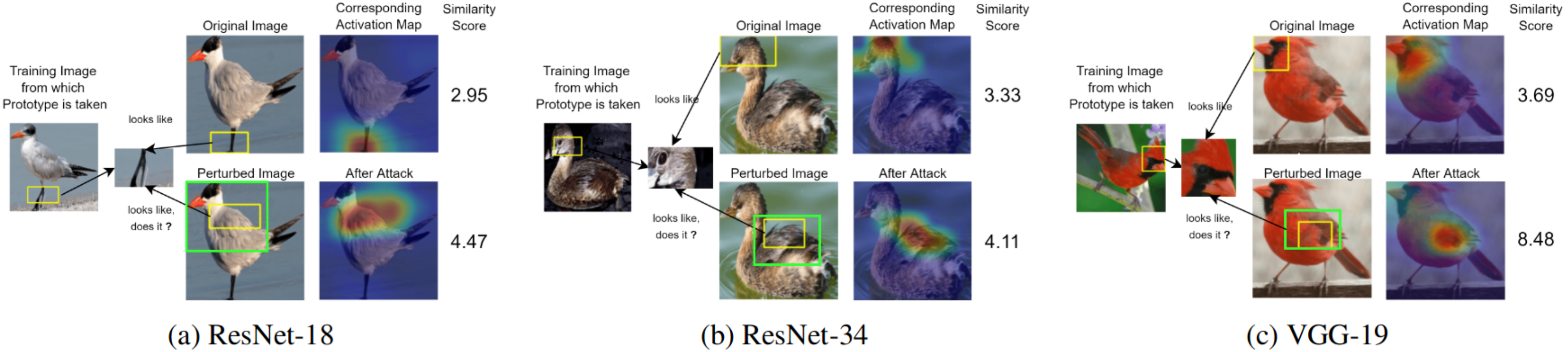


Example-based reasoning

- Prototype images [4]



- Issues: latent representations [9]

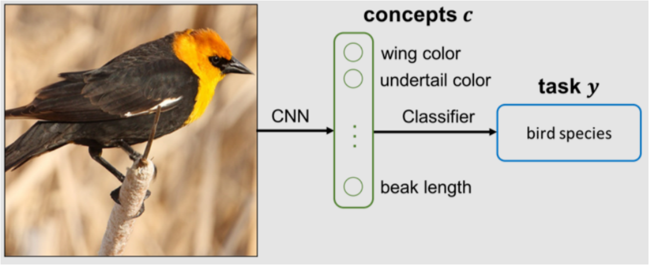


Example-based reasoning

- **Concept bottlenecks** (loss of accuracy)

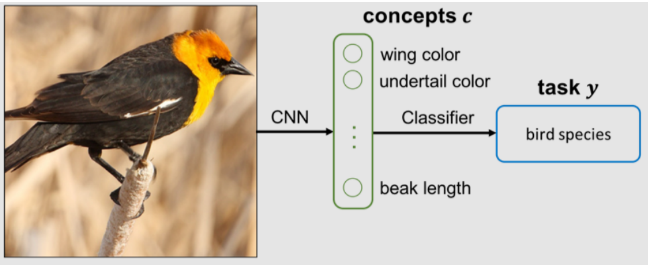
Example-based reasoning

- **Concept bottlenecks** (loss of accuracy)

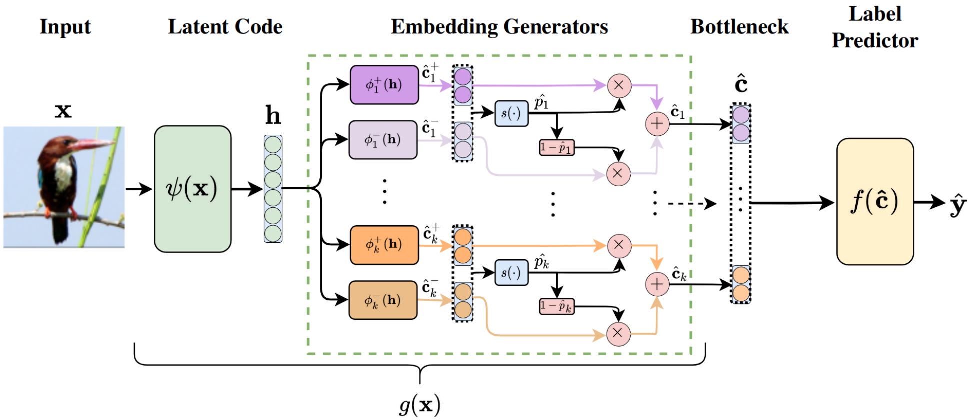


Example-based reasoning

- **Concept bottlenecks** (loss of accuracy)



- **Concept embeddings** (still prescribed concepts) [7]



So... is there a tradeoff?

- Simple, interpretable models *can* match black boxes [14]
  - With tabular data
  - With image data

So... is there a tradeoff?

- Simple, interpretable models *can* match black boxes [14]
  - With tabular data
  - With image data
- Typically intractable, might require domain knowledge. But $\exists$ **off-the-shelf solutions**

So... is there a tradeoff?

- Simple, interpretable models *can* match black boxes [14]
  - With tabular data
  - With image data
- Typically intractable, might require domain knowledge. But $\exists$ **off-the-shelf solutions**
- [move to end Black boxes are sometimes necessary, sometimes better, sometimes worse [3]]

So... is there a tradeoff?

- Simple, interpretable models *can* match black boxes [14]
  - With tabular data
  - With image data
- Typically intractable, might require domain knowledge. But $\exists$ **off-the-shelf solutions**
- [move to end Black boxes are sometimes necessary, sometimes better, sometimes worse [3]]
- How can we reason about this?

  Can we predict whether an interpretable model exists?

What can SLT tell us?

Dataset $S := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $(X, Y) \sim \mathcal{D}$

Hypothesis class $\mathcal{F} \subset Y^X$

Optimal $f^\star \in \mathcal{F}$ minimises **risk** $R(f) := \mathbb{E}_{\mathcal{D}}[l(f(X), Y)]$, for some **loss** $l\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

- Minimise **empirical risk** $\hat{R}_S(f) := \frac{1}{n} \sum l(f(x_i), y_i)$ to obtain estimator $f_S := f(S) \in \mathcal{F}$.

What can SLT tell us?

Dataset $S := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $(X, Y) \sim \mathcal{D}$

Hypothesis class $\mathcal{F} \subset Y^X$

Optimal $f^\star \in \mathcal{F}$ minimises **risk** $R(f) := \mathbb{E}_{\mathcal{D}}[l(f(X), Y)]$, for some **loss** $l \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

- Minimise **empirical risk** $\hat{R}_S(f) := \frac{1}{n} \sum l(f(x_i), y_i)$ to obtain estimator $f_S := f(S) \in \mathcal{F}$.
- **Interpretable model class** $\mathcal{F}_I \subsetneq \mathcal{F}$ [6]

What can SLT tell us?

Dataset $S := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $(X, Y) \sim \mathcal{D}$

Hypothesis class $\mathcal{F} \subset Y^X$

Optimal $f^\star \in \mathcal{F}$ minimises **risk** $R(f) := \mathbb{E}_\mathcal{D}[l(f(X), Y)]$, for some **loss** $l \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

- Minimise **empirical risk** $\hat{R}_S(f) := \frac{1}{n} \sum l(f(x_i), y_i)$ to obtain estimator $f_S := f(S) \in \mathcal{F}$.
- **Interpretable model class** $\mathcal{F}_I \subsetneq \mathcal{F}$ [6]
  - trees of depth $\leqslant k$
  - linear classifiers with $|\theta|_0 \leqslant k$
  - classifiers that can be well approximated by some class of surrogates (...)
  - Leaves out: dependency of $\mathcal{F}_I$ on the data (local interpretability), different user groups

What can SLT tell us?

Dataset $S := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $(X, Y) \sim \mathcal{D}$

Hypothesis class $\mathcal{F} \subset Y^X$

Optimal $f^\star \in \mathcal{F}$ minimises **risk** $R(f) := \mathbb{E}_{\mathcal{D}}[l(f(X), Y)]$, for some **loss** $l: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$.

- Minimise **empirical risk** $\hat{R}_S(f) := \frac{1}{n} \sum l(f(x_i), y_i)$ to obtain estimator $f_S := f(S) \in \mathcal{F}$.
- **Interpretable model class** $\mathcal{F}_I \subsetneq \mathcal{F}$ [6]
  - trees of depth $\leqslant k$
  - linear classifiers with $|\theta|_0 \leqslant k$
  - classifiers that can be well approximated by some class of surrogates (...)
  - Leaves out: dependency of $\mathcal{F}_I$ on the data (local interpretability), different user groups
- **Can we predict** whether we will lose accuracy with $\mathcal{F}_I$?

What can SLT tell us?

- $\mathcal{F}_I \subsetneq \mathcal{F}$

What can SLT tell us?

- $\mathcal{F}_I \subsetneq \mathcal{F}$

  $\Rightarrow$ Best risk in $\mathcal{F}_I =: R_I^\star \geqslant R^\star :=$ best risk in $\mathcal{F}$ (**modelling bias**) ... **Is that it?**

What can SLT tell us?

- $\mathcal{F}_I \subsetneq \mathcal{F}$

  $\Rightarrow$ Best risk in $\mathcal{F}_I =: R_I^\star \geqslant R^\star :=$ best risk in $\mathcal{F}$ (**modelling bias**) ... **Is that it?**

- $R(f) \leqslant \hat{R}(f) + \mathcal{O}\left(\sqrt{C/n}\right)$ with $C = \mathcal{O}(\log \text{“capacity”}(\mathcal{F}))$)

  $\Rightarrow$ Better generalization for $\mathcal{F}_I \subsetneq \mathcal{F}$ ... ??

What can SLT tell us?

- $\mathcal{F}_I \subsetneq \mathcal{F}$

  $\Rightarrow$ Best risk in $\mathcal{F}_I =: R_I^\star \geqslant R^\star :=$ best risk in $\mathcal{F}$ (**modelling bias**) ... **Is that it?**

- $R(f) \leqslant \hat{R}(f) + \mathcal{O}\left(\sqrt{C/n}\right)$ with $C = \mathcal{O}(\log \text{“capacity”}(\mathcal{F}))$)

  $\Rightarrow$ Better generalization for $\mathcal{F}_I \subsetneq \mathcal{F}$ ... ??

- Let $\hat{f}_I \in \operatorname*{argmin}_{f \in \mathcal{F}_I} \hat{R}_S(f)$ and $\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_S(f)$. The gap $R(\hat{f}_I) - R(\hat{f})$ depends on: [6]

What can SLT tell us?

- $\mathcal{F}_I \subsetneq \mathcal{F}$

  $\Rightarrow$ Best risk in $\mathcal{F}_I =: R_I^\star \geqslant R^\star :=$ best risk in $\mathcal{F}$ (**modelling bias**) ... **Is that it?**

- $R(f) \leqslant \hat{R}(f) + \mathcal{O}\left(\sqrt{C/n}\right)$ with $C = \mathcal{O}(\log \text{“capacity”}(\mathcal{F}))$)

  $\Rightarrow$ Better generalization for $\mathcal{F}_I \subsetneq \mathcal{F}$ ... ??

- Let $\hat{f}_I \in \underset{f \in \mathcal{F}_I}{\operatorname{argmin}} \hat{R}_S(f)$ and $\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}_S(f)$. The gap $R(\hat{f}_I) - R(\hat{f})$ depends on: [6]

  a) Increase in modeling bias $R_I^\star - R^\star \geqslant 0$

  b) Change in estimation error $R(\hat{f}_I) - R_I^\star$ vs $R(\hat{f}) - R^\star$

  This depends on the **capacity** of $\mathcal{F}_I$ and **dataset size** $\Rightarrow$ can be small

What can SLT tell us?

- $\mathcal{F}_I \subsetneq \mathcal{F}$

  $\Rightarrow$ Best risk in $\mathcal{F}_I =: R_I^\star \geqslant R^\star :=$ best risk in $\mathcal{F}$ (**modelling bias**) ... **Is that it?**

- $R(f) \leqslant \hat{R}(f) + \mathcal{O}(\sqrt{C/n})$ with $C = \mathcal{O}(\log \text{“capacity”}(\mathcal{F}))$

  $\Rightarrow$ Better generalization for $\mathcal{F}_I \subsetneq \mathcal{F}$ ... ??

- Let $\hat{f}_I \in \operatorname*{argmin}_{f \in \mathcal{F}_I} \hat{R}_S(f)$ and $\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_S(f)$. The gap $R(\hat{f}_I) - R(\hat{f})$ depends on: [6]

  a) Increase in modeling bias $R_I^\star - R^\star \geqslant 0$

  b) Change in estimation error $R(\hat{f}_I) - R_I^\star$ vs $R(\hat{f}) - R^\star$

  This depends on the **capacity** of $\mathcal{F}_I$ and **dataset size** $\Rightarrow$ can be small

Derived from Excess risk: $R(f_S) - R^\star = \underbrace{R_I^\star - R^\star}_{\text{modelling bias}} + \underbrace{R(f_S) - R_I^\star}_{\text{estimation error}}$

The effect of ERM

- Recall $\hat{f}_I \in \underset{f \in \mathcal{F}_I}{\operatorname{argmin}} \hat{R}_S(f)$ and $\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}_S(f)$.

The effect of ERM

- Recall $\hat{f}_I \in \operatorname*{argmin}_{f \in \mathcal{F}_I} \hat{R}_S(f)$ and $\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_S(f)$.

- The gap $R(\hat{f}_I) - R(\hat{f})$ can also be seen to depend on: [6]

  a) Change in empirical risk $\hat{R}(\hat{f}_I) - \hat{R}(\hat{f})$

  b) Change in generalization error $R(\hat{f}_I) - \hat{R}(\hat{f}_I)$ vs $R(\hat{f}) - \hat{R}(\hat{f})$

  Will depend on **dataset size** and **capacity** of $\mathcal{F}$

The effect of ERM

- Recall $\hat{f}_I \in \operatorname*{argmin}_{f \in \mathcal{F}_I} \hat{R}_S(f)$ and $\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_S(f)$.

- The gap $R(\hat{f}_I) - R(\hat{f})$ can also be seen to depend on: [6]

  a) Change in empirical risk $\hat{R}(\hat{f}_I) - \hat{R}(\hat{f})$

  b) Change in generalization error $R(\hat{f}_I) - \hat{R}(\hat{f}_I)$ vs $R(\hat{f}) - \hat{R}(\hat{f})$

  Will depend on **dataset size** and **capacity** of $\mathcal{F}$

  Derived from standard generalization bounds $R(f) \leqslant \hat{R}(f) + \mathcal{O}(\sqrt{C/n})$

“Conclusions” from SLT

- Dependence on sample size and capacity

“Conclusions” from SLT

- Dependence on sample size and capacity
- Change hard to quantify

"Conclusions" from SLT

- Dependence on sample size and capacity
- Change hard to quantify
- SLT bounds look at $|R(\hat{f}) - \hat{R}(\hat{f})|$. We want: $|R^\star - \hat{R}(\hat{f}_I)|$

“Conclusions” from SLT

- Dependence on sample size and capacity
- Change hard to quantify
- SLT bounds look at $|R(\hat{f}) - \hat{R}(\hat{f})|$. We want: $|R^\star - \hat{R}(\hat{f}_I)|$
- So... Not very informative [6]

“Conclusions” from SLT

- Dependence on sample size and capacity
- Change hard to quantify
- SLT bounds look at $|R(\hat{f}) - \hat{R}(\hat{f})|$. We want: $|R^\star - \hat{R}(\hat{f}_I)|$
- So... Not very informative [6]

Can we do better?

The Rashomon set

Observation

Often, **if the data represent well the problem, most models perform similarly**

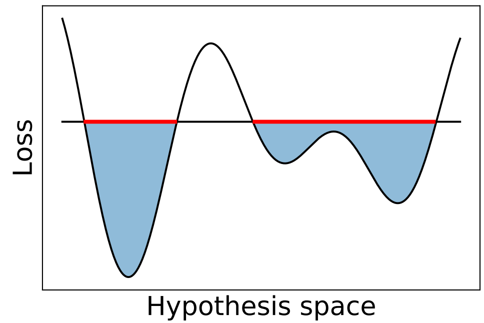
The Rashomon set

- **Observation**

  Often, **if the data represent well the problem, most models perform similarly**

- The (empirical) **Rashomon set** is the set of **almost-optimal models** [Breiman 2001]

$$\hat{\mathfrak{R}}(\mathcal{F},\gamma) := \{f \in \mathcal{F} : \hat{R}(f) - \hat{R}^{\star} \leqslant \gamma\}$$

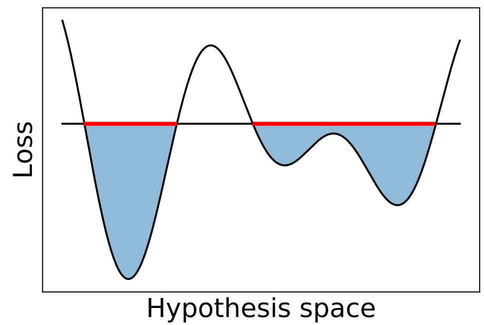


- ## Observation

  Often, **if the data represent well the problem, most models perform similarly**

- The (empirical) **Rashomon set** is the set of **almost-optimal models** [Breiman 2001]

$$\hat{\mathfrak{R}}(\mathcal{F},\gamma) := \{f \in \mathcal{F} : \hat{R}(f) - \hat{R}^\star \leqslant \gamma\}$$



- ## Hypothesis

  [15]

  if many models perform similarly well, then there usually is an interpretable one

The Rashomon ratio

The **Rashomon ratio** is the **fraction of models that have low loss**

$$\mathfrak{Ra}(\mathcal{F}, \gamma) := \frac{|\hat{\mathfrak{R}}(\mathcal{F}, \gamma)|}{|\mathcal{F}|} \qquad = \frac{|\{f \in \mathcal{F} \colon \hat{R}(f) - \hat{R}^\star \leqslant \gamma\}|}{|\mathcal{F}|}$$

The Rashomon ratio

The **Rashomon ratio** is the **fraction of models that have low loss**

$$\mathfrak{Ra}(\mathcal{F},\gamma) := \frac{|\hat{\mathfrak{R}}(\mathcal{F},\gamma)|}{|\mathcal{F}|} = \frac{|\{f\in\mathcal{F}\colon \hat{R}(f)-\hat{R}^{\star}\leqslant\gamma\}|}{|\mathcal{F}|}$$

Theorem. *If* $\exists\, \hat{f}_I \in \hat{\mathfrak{R}}(\mathcal{F},\gamma)$ *then with high probability*

$$|\hat{R}(\hat{f}_I) - R^{\star}| \leqslant \gamma + \mathcal{O}\left(\sqrt{\log(|\mathcal{F}_I|)/n}\right)$$

The Rashomon ratio

The **Rashomon ratio** is the **fraction of models that have low loss**

$$\mathfrak{Ra}(\mathcal{F},\gamma) := \frac{|\hat{\mathfrak{R}}(\mathcal{F},\gamma)|}{|\mathcal{F}|} = \frac{|\{f \in \mathcal{F} : \hat{R}(f) - \hat{R}^\star \leqslant \gamma\}|}{|\mathcal{F}|}$$

Theorem. *If* $\exists \hat{f}_I \in \hat{\mathfrak{R}}(\mathcal{F},\gamma)$ *then with high probability*

$$|\hat{R}(\hat{f}_I) - R^\star| \leqslant \gamma + \mathcal{O}\left(\sqrt{\log(|\mathcal{F}_I|)/n}\right)$$

Theorem. *Assume that* $\mathcal{F}_I$ *is "dense enough" in* $\hat{\mathfrak{R}}(\mathcal{F},\gamma)$, *and* $\hat{\mathfrak{R}}(\mathcal{F},\gamma)$ *is "wide enough". With prob* $1-\varepsilon$ *there exist* $f_1,\ldots,f_m \in \mathcal{F}_I$ *s.t.*

$$|R(f_i) - \hat{R}(f_i)| \leqslant C\,\mathrm{Rad}(\mathcal{F}_I) + \mathcal{O}\left(\sqrt{\log(1/\varepsilon)/n}\right)$$

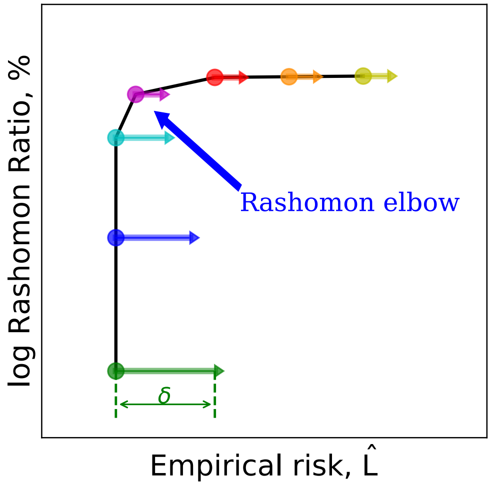
Rashomon curves

Some empirical observations across *many* datasets

Tower of model classes $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_k \subset \mathcal{F}$

As $\mathfrak{Ra}(\mathcal{F}_i, \gamma) = \dfrac{|\hat{\mathfrak{R}}(\mathcal{F}_i, \gamma)|}{|\mathcal{F}_i|}$ [decreases (higher $|\mathcal{F}_i|$), so does $\hat{R}$ up until the "elbow", see video]

After some point, all $\mathcal{F}$ perform equally, and higher $|\mathcal{F}_i|$ worsens generalization



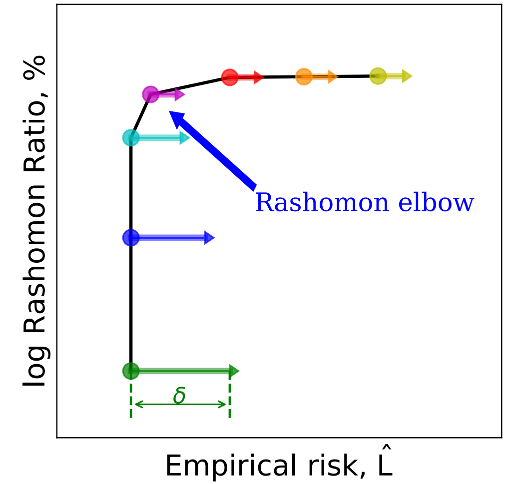
Rashomon curves

Some empirical observations across *many* datasets

Tower of model classes $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_k \subset \mathcal{F}$

As $\mathfrak{Ra}(\mathcal{F}_i, \gamma) = \frac{|\hat{\mathfrak{R}}(\mathcal{F}_i, \gamma)|}{|\mathcal{F}_i|}$ [decreases (higher $|\mathcal{F}_i|$), so does $\hat{R}$ up until the "elbow", see video]

After some point, all $\mathcal{F}$ perform equally, and higher $|\mathcal{F}_i|$ worsens generalization



Rule of thumb

If off-the-shelf methods do similarly, try constraining for interpretability

Bonus: the Rashomon set of sparse trees

- TreeFARMS: complete enumeration of $\mathfrak{R}$ for sparse trees [18]

  (Can be sampled when it is too large)

- Applications

  a) pick among **all** almost-optimal models

  b) study variable importance for the set of almost-optimal trees

  c) $\mathfrak{R}$ for accuracy $\Rightarrow$ can enumerate $\mathfrak{R}$ for balanced accuracy and F_1-score

  d) $\mathfrak{R}$ for a dataset $\Rightarrow$ $\mathfrak{R}$ for subsets

A panacea?

- For many critical applications, there is no tradeoff

A panacea?

- For many critical applications, there is no tradeoff
- Many off-the-shelf algorithms (with caveats)

A panacea?

- For many critical applications, there is no tradeoff
- Many off-the-shelf algorithms (with caveats)
- Even for image classification (with more caveats)

A panacea?

- For many critical applications, there is no tradeoff
- Many off-the-shelf algorithms (with caveats)
- Even for image classification (with more caveats)
- Interpretable can mean very different things

A panacea?

- For many critical applications, there is no tradeoff
- Many off-the-shelf algorithms (with caveats)
- Even for image classification (with more caveats)
- Interpretable can mean very different things
- Interpretable $\not\Rightarrow$ users make better decisions [11]

A panacea?

- For many critical applications, there is no tradeoff
- Many off-the-shelf algorithms (with caveats)
- Even for image classification (with more caveats)
- Interpretable can mean very different things
- Interpretable $\not\Rightarrow$ users make better decisions [11]
- Interpretable $\Rightarrow$ still lots to do [13]

Different target groups

- Developers need insights into data and model
- ML-literate users can benefit from “simple” models
- Scientifically-illiterate users can be overwhelmed even by simple systems
- High-stakes applications require 1:1 faithfulness of the explanations
- ...

Antecedents, mechanisms, and consequences of overreliance on AI

	[13]	Description	Mitigation
Antecedents	Individual differences	Differences in users' demographic, professional, social, and cultural traits affect their reliance on AI.	Provide personalized adjustments for users; Effectively onboard users; Give users choice
Mechanisms	Automation bias	Tendency to favor recommendations from automated systems, while disregarding information from nonautomated sources.	Effectively onboard users; Employ cognitive forcing functions; Provide personalized adjustments to users; Provide real-time feedback
	Confirmation bias	Tendency to favor information that aligns with prior assumptions, beliefs, and values.	Employ cognitive forcing functions; Effectively onboard users; Provide personalized adjustments to users; Provide real-time feedback
	Ordering effects	The order of presented information affects user perceptions and decisions. The timing of AI errors significantly affects user reliance.	Effectively onboard users; Provide personalized adjustments to users; Alter speed of interaction;
	Overestimating explanations	High-fidelity explanations can lead users to develop overreliance on AI.	Be transparent with users; Provide real-time feedback; Provide effective explanations
Consequences	Poor human+AI performance	Overreliance causes poor human+AI team performance compared to the human or AI working alone.	All

- Explaining models with black boxes can be dangerous

- Explaining models with black boxes can be dangerous
- When do we trust the proxy? If it were always right, we could just use it

- Explaining models with black boxes can be dangerous
- When do we trust the proxy? If it were always right, we could just use it
- Simple and interpretable models can often perform as well as complex, black boxes

- Explaining models with black boxes can be dangerous
- When do we trust the proxy? If it were always right, we could just use it
- Simple and interpretable models can often perform as well as complex, black boxes
- We should prefer simpler models for development (model and data debugging)

- Explaining models with black boxes can be dangerous
- When do we trust the proxy? If it were always right, we could just use it
- Simple and interpretable models can often perform as well as complex, black boxes
- We should prefer simpler models for development (model and data debugging)
- We should prefer simpler models for deployment with experts, when properly designed

- Explaining models with black boxes can be dangerous
- When do we trust the proxy? If it were always right, we could just use it
- Simple and interpretable models can often perform as well as complex, black boxes
- We should prefer simpler models for development (model and data debugging)
- We should prefer simpler models for deployment with experts, when properly designed
- Natural interpretability constraints don't always translate to better results down the line

- Explaining models with black boxes can be dangerous
- When do we trust the proxy? If it were always right, we could just use it
- Simple and interpretable models can often perform as well as complex, black boxes
- We should prefer simpler models for development (model and data debugging)
- We should prefer simpler models for deployment with experts, when properly designed
- Natural interpretability constraints don't always translate to better results down the line
- Rule of thumb: if many models perform similarly, there is probably a simple one

Learning more

- Many excellent talks by Cynthia Rudin (YouTube)
- Prototype networks and concept embeddings (this seminar, September)
- Sparse models, anyone?
- (but . . .)
- TransferLab

[1] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu. Does explainable artificial intelligence improve human decision-making? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6618–6626.

[2] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning Certifiably Optimal Rule Lists for Categorical Data. 18(234):1–78.

[3] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 248–266. ACM.

[4] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[5] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang. An Interpretable Model with Globally Consistent Explanations for Credit Risk. ArXiv.

[6] G. Dziugaite, S. Ben-David, and D. M. Roy. Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability.

[7] M. Espinosa Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, Z. Shams, F. Precioso, S. Melacci, A. Weller, P. Lió, and M. Jamnik. Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. 35:21400–21413.

[8] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. CRC Press.

[9] A. Hoffmann, C. Fanconi, R. Rade, and J. Kohler. This Looks Like That... Does it? Shortcomings of Latent Space Prototype Interpretability in Deep Networks.

[10] X. Hu, C. Rudin, and M. Seltzer. Optimal Sparse Decision Trees. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[11] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. 11(1):1–9.

[12] J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer. Generalized and Scalable Optimal Sparse Decision Trees. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6150–6160. PMLR.

[13] S. Passi and M. Vorvoreanu. Overreliance on AI: Literature review.

[14] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. 1(5):206–215.

[15] L. Semenova, C. Rudin, and R. Parr. On the Existence of Simpler Machine Learning Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1827–1858. Association for Computing Machinery.

[16] B. Ustun and C. Rudin. Learning Optimized Risk Scores. 20(150):1–75.

[17] B. Ustun, S. Tracà, and C. Rudin. Supersparse linear integer models for predictive scoring systems. In *Proceedings of the 17th AAAI Conference on Late-Breaking Developments in the Field of Artificial Intelligence*, AAAIWS'13-17, pages 128–130. AAAI Press.

[18] R. Xin, C. Zhong, Z. Chen, T. Takagi, M. Seltzer, and C. Rudin. Exploring the Whole Rashomon Set of Sparse Decision Trees.

[19] J. Yu, A. Ignatiev, P. L. Bodic, and P. J. Stuckey. Optimal Decision Lists using SAT.

[20] R. Zhang, R. Xin, M. Seltzer, and C. Rudin. Optimal Sparse Regression Trees. Association for the Advancement of Artificial Intelligence.