Explainability vs. Interpretability

Ryan Rodriguez & Hayoung Jeong

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x
nature machine intelligence

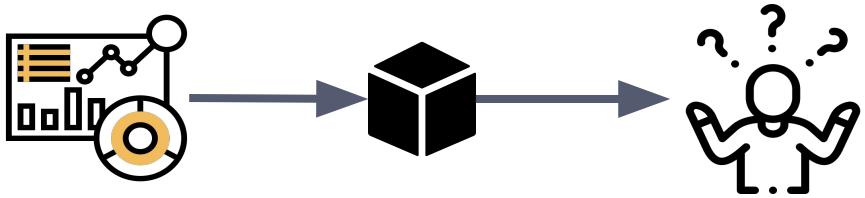
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

High-stake Prediction Applications

source: Dr. Sudeepa Roy's Lecture 6 & https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70

Black Box Models



TOO-COMPLICATED

- Highly recursive (e.g. deep learning models)
- Difficult to manually combine outside information (e.g. risk assessment)

TOO-COMPLICATED & PROPRIETARY

- Preserve secrecy

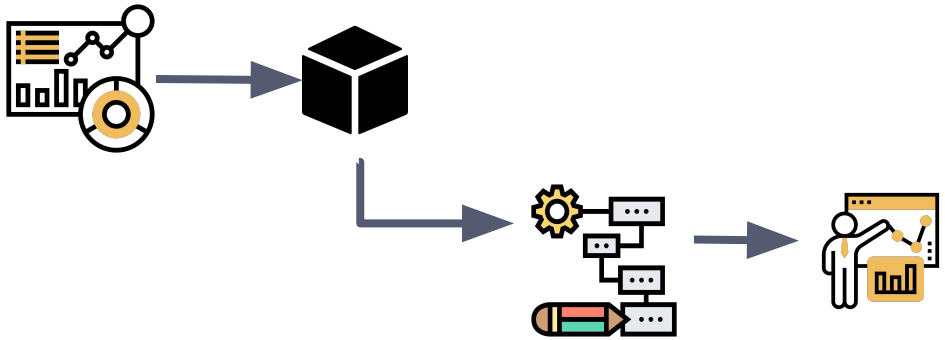
PROPRIETARY

- Limited transparency (e.g. training data, model selection)
- Example: COMPAS (recidivism prediction)

Explainable ML/AI

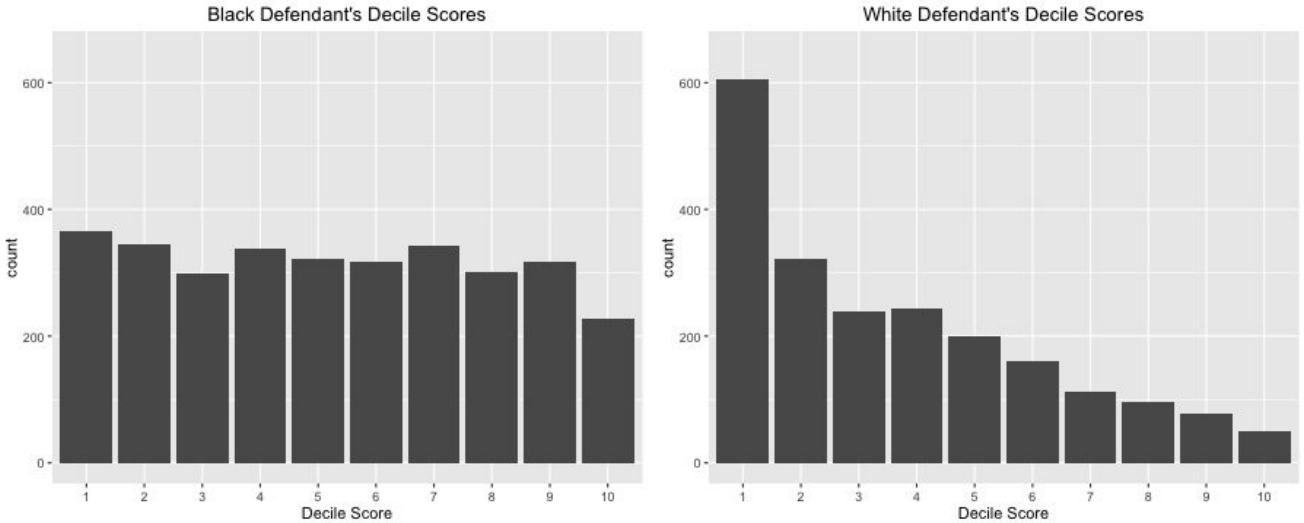
“tools and frameworks to help you understand and interpret predictions made by your machine learning models”

- Google Explainable AI



COMPAS and ProPublica

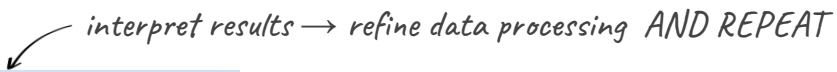
“many of the methods that claim to produce explanations instead compute useful summary statistics of predictions made by the original model.”



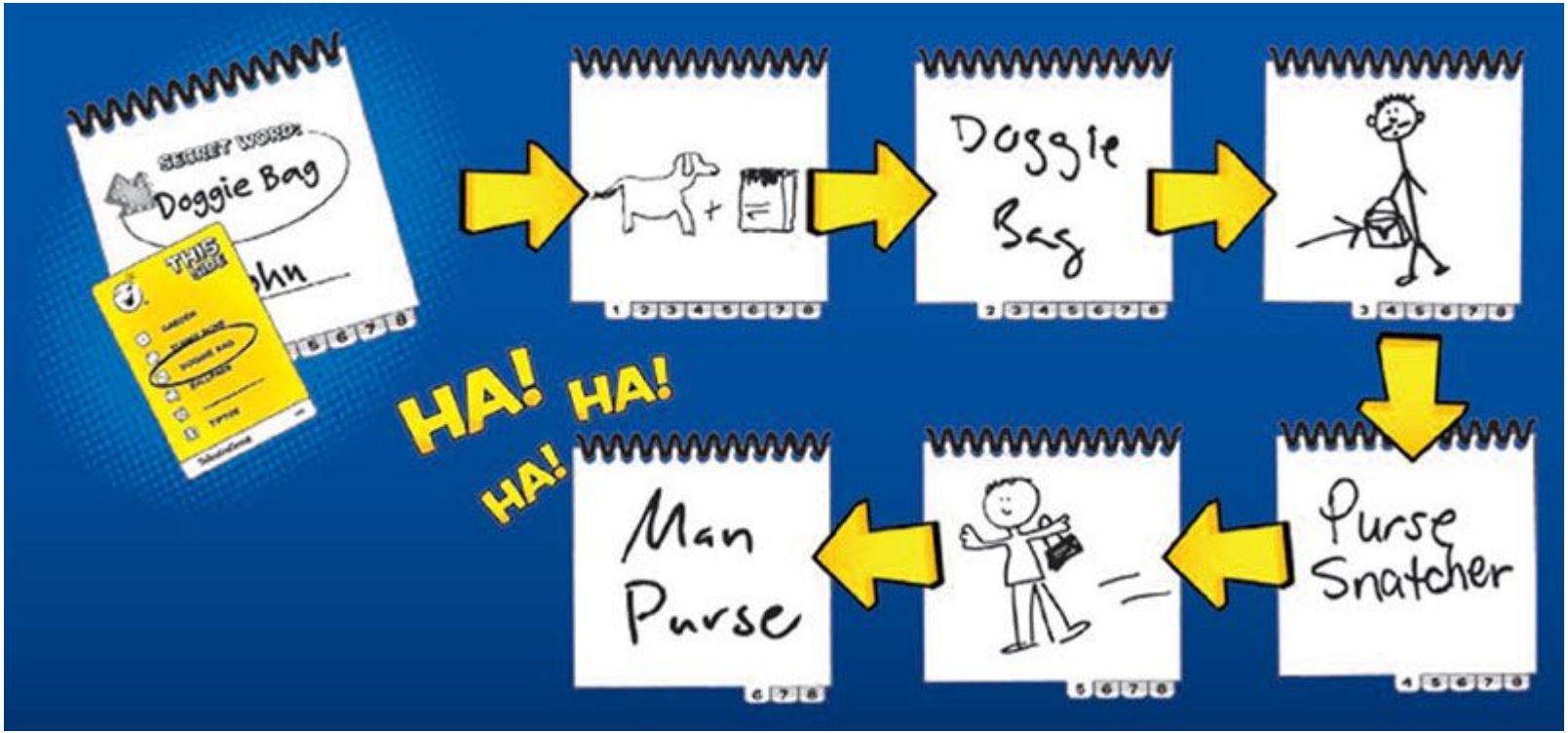
Key issues with explainable ML

“It is a myth that there is necessarily a trade-off between accuracy and interpretability.”

- Complicated model **does not** always give top performances.
- When data is nicely structured with meaningful features, simple models perform **similarly** to complicated models [e.g. Razavian et al., 2015]



- Model improvement is an iterative process. So the more interpretable the model, the easier to refine the data processing step.
  - Can reveal flaws in the dataset, false assumptions in data generation, find meaningful features



https://www.ultraboardgames.com/telestrations/game-rules.php

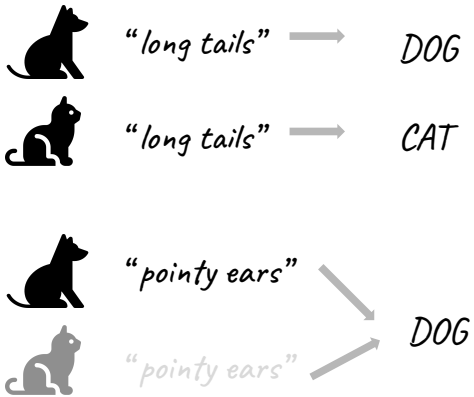
“Explainable ML methods provide explanations that are not faithful to what the original model computes.”

- Explanations can be an inaccurate representation of the original model in parts of the feature space.
  - Perfect fidelity is infeasible
- Would you trust if explanations are said to be correct 90% of the time?
- Instead of the word “explanation,” call them ”‘summaries of predictions”, “summary statistics” or “trends”

“Explanations often do not make sense or do not provide enough detail to understand what the black box is doing.”

- Explanation may omit information about how relevant information is being used
- Explanation may be the same for multiple classes
- Recent works show explanations only for the observation’s correct label

Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University



“Black box models with explanations can lead to an overly complicated decision pathway that is ripe for human error.”

- Data entry errors (i.e. typographical errors) can happen
- *“If typographical errors by humans entering these data into a survey occur at a rate of 1%, then more than one out of every two surveys on average will have at least one typographical error”*
- Difficult to troubleshoot and would not be able to discover flaws in our model

“Counterfactual explanations’ of black boxes are insufficient.”

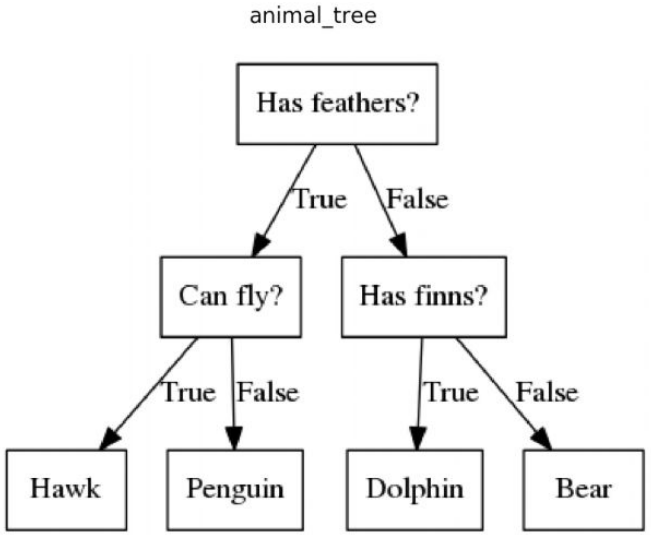
- Counterfactual explanations need to be realistic, an action that is reversible
- Example of counterfactual explanation: *You will qualify for the loan you were previously rejected…*
  - *If you have reduced your debt by $5000 and increased your savings by 50%*
  - *If you had gotten a job that pays $500 more per week*
- **Minimal change in input may lead to different conclusion per individual**
- **Which explanation is lowest cost for the user cannot be decided**

Interpretable Model

Table 1 | Machine learning model from the CORELS algorithm

IF	age between 18–20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21–23 and 2–3 prior offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

This model from ref. [39] is the minimizer of a special case of equation (1) discussed later in the challenges section. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data from Florida needed to produce this model.



Key issues with interpretable ML

“Corporations can make profits from the intellectual property afforded to a black box.”

Table 1 | Machine learning model from the CORELS algorithm

IF	age between 18–20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21–23 and 2–3 prior offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

This model from ref. [39] is the minimizer of a special case of equation (1) discussed later in the challenges section. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data from Florida needed to produce this model.

- **Certifiably Optimal Rule Lists** (CORELS) that looks for if–then patterns in data.
- CORELS is **EQUALLY ACCURATE** as COMPAS

“Corporations can make profits from the intellectual property afforded to a black box.”

Table 2 | Comparison of COMPAS and CORELS models

COMPAS	CORELS
Black box; 130+ factors; might include socio-economic info; expensive (software licence); within software used in US justice system	Full model is in Table 1; only age, priors, gender (optional); no other information; free, transparent

COMPAS claims that it needs to be proprietary in order to avoid revealing the trade secret.

- prevents them from being gamed or reverse-engineered.

Discussion: Is there ANY incentives for companies to strive for interpretability?

Rudin's Counterarguments on Proprietary Models🤫

TRUSTING BLACKBOX == TRUSTING THE DATASET

- Dataset may not represent all possible situations
- Dataset may be biased toward particular class
- Accuracy can drop significantly in real practice

TRANSPARENT == POSITIVE IMPACT

*"If the ratings are accurate measures of quality, then **making the ratings more transparent** could have a uniformly positive impact: it would help companies to **make better rated** products, it would help consumers to have these **higher quality products**, and it would **encourage** rating companies to receive **feedback** as to whether their rating systems **fairly represent quality**."*

“Interpretable models can entail significant effort to construct in terms of both computation and domain expertise”

- Interpretability often requires set of application-specific constraints on the model → harder to solve/computationally costly
    - Explanation methods are usually based on simple derivatives, which lead to easier gradient-based optimization.
- Definition of interpretability vary depending on the domain
    - Thus, domain knowledge is crucial.

Rudin's Counterargument on the Cost & Effort

- Analysis and computational cost and time are less expensive than "the cost of having a flawed or overly complicated model."
- Creating high-quality model will pay off!

“Scientists’ false belief: Black box models seem to uncover ‘hidden patterns.”

- Black box uncover subtle hidden patterns in the data
  - Pattern recognition

Rudin’s Counterargument

- If the pattern was THAT important, interpretable model can also locate & use for its prediction
  - Will require researcher’s ability to create a model that is capable of uncovering the **interpretable** patterns

Algorithmic challenges in interpretable ML

Challenge 1: constructing optimal logical models.

- Heuristic/greedy methods are not designed to choose a globally best choice (i.e., optimal solution)
- It is difficult to tell if poor performance is due to the choice of algorithm (not optimizing its objective) or combination of choice of model class and constraints (not enough flexibility to fit the data)

Challenge 1: constructing optimal logical models.

An optimization problem: "find a model that **minimizes** a combination of the fraction of **misclassified training points** and the **size of the model**."

Classification error: how much you are willing to sacrifice in order to have one fewer term (e.g. 0.01 = sacrifice 1% training accuracy to reduce size by one)

$$\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{[\text{training observation } i \text{ is misclassified by } f]} + \lambda \times \text{size}(f) \right) \quad (1)$$

Family of logical models (e.g., decision trees)

Size of model: # of logical conditions (e.g. leaves)

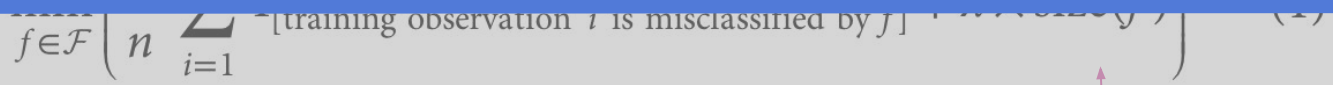
Challenge 1: constructing optimal logical models.

An optimization problem: "find a model that **minimizes** a combination of the fraction of **misclassified training points** and the **size of the model**."

Classification error: how much you are willing

COMPUTATIONALLY HARD– *will take forever to iterate all models & lists*

CAN WE SOLVE THIS IN PRACTICAL WAYS?



Family of logical models (e.g., decision trees)

Size of model: # of logical conditions (e.g. leaves)

Challenge 1: constructing optimal logical models.

Example: **CORELS algorithm–** able to solve the optimization in < 1 minute

- Reduced search space of rule lists using a **set of theorems**
- Built custom **fast bit-vector library** for fast exploration of the search space
- Set **specialized data structures** to keep track of **intermediate computations and symmetries.**

Challenge 1: constructing optimal logical models.

Example: **CORELS algorithm–** able to solve the optimization in < 1 minute

- Reduced search space of rule lists using a **set of theorems**
- Built custom **fast bit-vector library** for fast exploration of the search space
- Set **specialized data structures** to keep track of **intermediate computations and symmetries.**

Theoretical + system-level techniques are needed!

Challenge 2: construct optimal sparse scoring systems.

Table 3 | Scoring system for risk of recidivism

1.	Prior arrests ≥ 2	1 point	⋯
2.	Prior arrests ≥ 5	1 point	+⋯
3.	Prior arrests for local ordinance	1 point	+⋯
4.	Age at release between 18 to 24	1 point	+⋯
5.	Age at release ≥ 40	−1 point	+⋯
		Score	= ⋯

Score	−1	0	1	2	3	4
Risk (%)	11.9	26.9	50.0	73.1	88.1	95.3

This system is from ref. [21], which was developed from refs. [29,46]. The model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

Modified Early Warning Score					MEWS		
Score	3	2	1	0	1	2	3
Resp		< 9		9-14	15-20	21-29	≥ 30
Puls/min		≤ 40	41-50	51-100	101-110	111-129	≥ 130
Syst.bltr	≤ 70	71-80	81-100	101-199		≥ 200	
Temp ° C		≤ 35	35.1-36	36.1-38	38.1-38.5	> 38.5	
CNS			Nytillkommen förvirring	Alert	Reagerar på tilltal	Reagerar på smärta	Reagerar ej

Vid allvarlig oro över hur patientens tillstånd utvecklas, om saturationenen akut försämras till < 90% trots syrgas givet med avdelningens förutsättningar eller om diuresen är < 200 ml på 4 timmar: Kontakta dagtid; vårdlagsansvarig läkare, Kontakta jourtid; op.jour 97140

Challenge 2: construct optimal sparse scoring systems.

- **Same optimization problem:** "find a model that **minimizes** a combination of the fraction of **misclassified training points** and the **size of the model**."
  - Computationally very hard because **domain** over which we solve the optimization problem is the **integer lattice**

$$\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{[\text{training observation } i \text{ is misclassified by } f]} + \lambda \times \text{size}(f) \right) \qquad (1)$$

Size of model:
# of terms in the model

- **Logistic regression**: use the coefficients as the "scores" but…
  - Lack accuracy
  - Uninterpretable coefficients (not 1, -1)

Challenge 2: construct optimal sparse scoring systems.

- **Optimization problem** for a mixed-integer-nonlinear program whose domain is the integer lattice.
  - "find coefficients b_j, j = 1...p for the **linear predictive model**"

Table 3 | Scoring system for risk of recidivism

1.	Prior arrests ≥ 2	1 point	⋯
2.	Prior arrests ≥ 5	1 point	+⋯
3.	Prior arrests for local ordinance	1 point	+⋯
4.	Age at release between 18 to 24	1 point	+⋯
5.	Age at release ≥ 40	−1 point	+⋯
		Score	= ⋯

Score	−1	0	1	2	3	4
Risk (%)	11.9	26.9	50.0	73.1	88.1	95.3

This system is from ref. [21], which was developed from refs. [29,46]. The model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

$$f(\mathbf{z}) = \sum_j b_j z_j$$ ← j^{th} covariate of observation z

Constrain to integers

$$\min_{b_1,b_2,\ldots,b_p \in \{-10,-9,\ldots,9,10\}} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-\sum_{j=1}^{p} b_j x_{i,j}\right)\right) + \lambda \sum_j 1_{[b_j \neq 0]}$$

Training observations

covariates

Size of model: number of non-zero coefficients

Challenge 2: construct optimal sparse scoring systems.

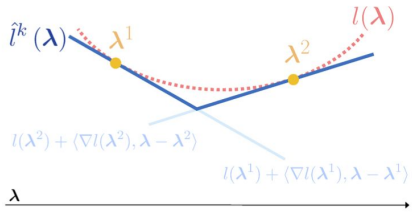
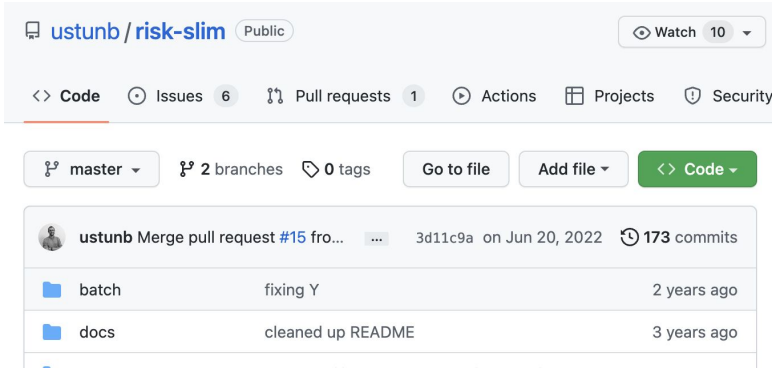


Figure 2: A convex loss function $l(\boldsymbol{\lambda})$ and its cutting plane approximation $\hat{l}^2(\boldsymbol{\lambda})$ built using cuts at the points $\boldsymbol{\lambda}^1$ and $\boldsymbol{\lambda}^2$.



Algorithm 2 Lattice Cutting Plane Algorithm (LCPA)

Input

$(\boldsymbol{x}_i, y_i)_{i=1}^N$	*training data*
$\mathcal{L}$	*constraint set for* RiskSlimMINLP
C_0	*ℓ_0 penalty parameter*
$\varepsilon^{\text{stop}} \in [0, 1]$	*optimality gap of acceptable solution*
RemoveNode	*rule to pick a node from a node set (provided by MIP solver)*
SplitPartition	*rule to split a partition into disjoint subsets (provided by MIP solver)*

Initialize

$k \leftarrow 0$	*number of cuts*
$\hat{l}^0(\boldsymbol{\lambda}) \leftarrow \{0\}$	*cutting-plane approximation of loss function*
$(V^{\min}, V^{\max}) \leftarrow (0, \infty)$	*bounds on the optimal value*
$\varepsilon \leftarrow \infty$	*optimality gap*
$\mathcal{P}_0 \leftarrow \text{conv}(\mathcal{L})$	*partition for initial node*
$v_0 \leftarrow V^{\min}$	*lower bound for initial node*
$\mathcal{N} \leftarrow \{(\mathcal{P}_0, v_0)\}$	*initial node set*

while $\varepsilon > \varepsilon^{\text{stop}}$ **do**
  $(\mathcal{P}_n, v_n) \leftarrow \text{RemoveNode}(\mathcal{N})$ ▷ *n is index of removed node*
  solve RiskSlimLP$(\hat{l}^k(\cdot), \mathcal{P}_n)$
  $\boldsymbol{\lambda}^{\text{LP}} \leftarrow$ coefficients from optimal solution to RiskSlimLP$(\hat{l}^k(\cdot), \mathcal{P}_n)$
  $v^{\text{LP}} \leftarrow$ optimal value of RiskSlimLP$(\hat{l}^k(\cdot), \mathcal{P}_n)$
  if optimal solution is integer feasible **then**
    compute cut parameters $l(\boldsymbol{\lambda}^{\text{LP}})$ and $\nabla l(\boldsymbol{\lambda}^{\text{LP}})$
    $\hat{l}^{k+1}(\boldsymbol{\lambda}) \leftarrow \max\{\hat{l}^k(\boldsymbol{\lambda}), l(\boldsymbol{\lambda}^{\text{LP}}) + \langle \nabla l(\boldsymbol{\lambda}^k), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{\text{LP}} \rangle\}$ ▷ *update approximation $\forall \boldsymbol{\lambda}$*
    if $v^{\text{LP}} < V^{\max}$ **then**
      $V^{\max} \leftarrow v^{\text{LP}}$ ▷ *update lower bound*
      $\boldsymbol{\lambda}^{\text{best}} \leftarrow \boldsymbol{\lambda}^{\text{LP}}$ ▷ *update best solution*
      $\mathcal{N} \leftarrow \mathcal{N} \setminus \{(\mathcal{P}_s, v_s) \mid v_s \geq V^{\max}\}$ ▷ *prune suboptimal nodes*
    end if
    $k \leftarrow k + 1$
  else if optimal solution is not integer feasible **then**
    $(\mathcal{P}', \mathcal{P}'') \leftarrow \text{SplitPartition}(\mathcal{P}_n, \boldsymbol{\lambda}^{\text{LP}})$ ▷ *$\mathcal{P}', \mathcal{P}''$ are disjoint subsets of $\mathcal{P}_n$*
    $(v', v'') \leftarrow (v^{\text{LP}}, v^{\text{LP}})$ ▷ *v^{LP} is lower bound for $\mathcal{P}', \mathcal{P}''$*
    $\mathcal{N} \leftarrow \mathcal{N} \cup \{(\mathcal{P}', v'), (\mathcal{P}'', v'')\}$ ▷ *add child nodes to $\mathcal{N}$*
  end if
  $V^{\min} \leftarrow \min_{\mathcal{N}} v_s$ ▷ *lower bound is smallest lower bound among nodes in $\mathcal{N}$*
  $\varepsilon \leftarrow 1 - V^{\min}/V^{\max}$ ▷ *update optimality gap*
end while
Output: $\boldsymbol{\lambda}^{\text{best}}$ *ε-optimal solution to* RiskSlimMINLP

RiskSlimLP$(\hat{l}(\cdot), \mathcal{P})$ is a LP relaxation of RiskSlimMIP$(\hat{l}(\cdot))$ over the partition $\mathcal{P} \subseteq \text{conv}(\mathcal{L})$:

$$\begin{aligned} \min_{\theta, \boldsymbol{\lambda}, \boldsymbol{\alpha}} \quad & \theta + C_0 \sum_{j=1}^{d} \alpha_j \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{P} \\ & \theta \geq \hat{l}(\boldsymbol{\lambda}) \\ & \alpha_j = \max(\lambda_j, 0)/\Lambda_j^{\max} + \min(\lambda_j, 0)/\Lambda_j^{\min} \text{ for } j = 1 \ldots d. \end{aligned} \tag{4}$$

https://blog.acolyer.org/2019/11/01/optimized-risk-scores/

Challenge 2: construct optimal sparse scoring systems.

The World Health Organization Adult Attention-Deficit/Hyperactivity Disorder Self-Report Screening Scale for *DSM-5*

Berk Ustun, MS, Lenard A. Adler, MD, [...], and Ronald C. Kessler, PhD

Additional article information

Key Points

Question

Can a brief screening scale based on patient responses to structured questions detect *DSM-5* adult attention-deficit/hyperactivity disorder in the general population?

Questions in the Optimal RiskSLIM *DSM-5* ASRS Screening Scale[a]

1. How often do you have difficulty concentrating on what people say to you, even when they are speaking to you directly? (*DSM-5* A1c)

2. How often do you leave your seat in meetings or other situations in which you are expected to remain seated? (*DSM-5* A2b)

3. How often do you have difficulty unwinding and relaxing when you have time to yourself? (*DSM-5* A2d)

4. When you're in a conversation, how often do you find yourself finishing the sentences of the people you are talking to before they can finish them themselves? (*DSM-5* A2g)

5. How often do you put things off until the last minute? (Non-*DSM*)

6. How often do you depend on others to keep your life in order and attend to details? (Non-*DSM*)

Challenge 3: define interpretability for specific domains and create methods accordingly, including computer vision.

- **What constitutes interpretability in computer vision** (visual classification tasks)?
  - Able to pay attention to different parts of the image and explain why these parts of the image were important in their reasoning process

---

***This* Looks Like *That*: Deep Learning for Interpretable Image Recognition**

Chaofan Chen*
Duke University
cfchen@cs.duke.edu

Oscar Li*
Duke University
oscarli@alumni.duke.edu

Chaofan Tao
Duke University
chaofan.tao@duke.edu

Alina Jade Barnett
Duke University
abarnett@cs.duke.edu

Jonathan Su
MIT Lincoln Laboratory†
su@ll.mit.edu

Cynthia Rudin
Duke University
cynthia@cs.duke.edu

Challenge 3: define interpretability for specific domains and create methods accordingly, including computer vision.

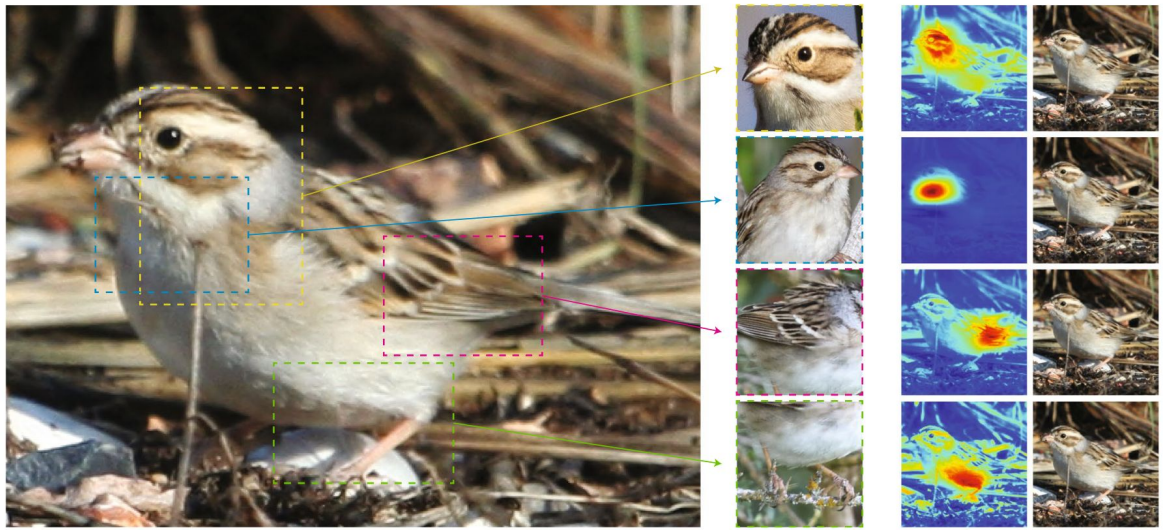


Fig. 3 | Image from the authors of ref. [48], indicating that parts of the test image on the left are similar to prototypical parts of training examples. The test image to be classified is on the left, the most similar prototypes are in the middle column, and the heatmaps that show which part of the test image is similar to the prototype are on the right. We included copies of the test image on the right so that it is easier to see to what part of the bird the heatmaps are referring. The similarities of the prototypes to the test image are what determine the predicted class label of the image. Here, the image is predicted to be a clay-coloured sparrow. The top prototype seems to be comparing the bird's head to a prototypical head of a clay-coloured sparrow, the second prototype considers the throat of the bird, the third looks at feathers, and the last seems to consider the abdomen and leg. Credit: Image constructed by Alina Barnett, Duke University

Encouraging responsible ML governance

Right to Explanation

Article 22
EU GDPR
"Automated individual decision-making, including profiling"

=> Recital: 71, 72
=> administrative fine: Art. 83 (5) lit b
=> Dossier: Automated Decision In Individual Cases, Profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
=> Article: 4

Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For

16 Duke Law & Technology Review 18 (2017)

67 Pages • Posted: 24 May 2017 • Last revised: 6 Dec 2017

stories" that have shaped recent attitudes in this domain. Firstly, the law is restrictive, unclear, or even paradoxical concerning when any explanation-related right can be triggered. Secondly, even navigating this, the legal conception of explanations as "meaningful information about the logic of processing" may not be provided by the kind of ML "explanations" computer scientists have developed, partially in response. ML explanations are restricted both by the type of explanation sought, the dimensionality of the domain and the type of user seeking an explanation. However, "subject-centric" explanations (SCEs) focussing on particular regions of a model around a query show promise for interactive exploration, as do explanation systems based on learning a model from outside rather than taking it apart (pedagogical versus decompositional explanations) in dodging developers' worries of intellectual property or trade secrets disclosure.

Discussion: What regulations can/will encourage interpretability?

Rudin's Proposal # 1

No black box should be deployed when there exists an interpretable (transparent) model with the same level of performance.

PRO

Companies receive compensation for developing an interpretable model

CONS

1. **No more proprietary models**! Not as much as profit but will be "useful for public good applications would make these problems appeal to academics and charitable foundations."
2. Might reduce industrial participation

Rudin's Proposal # 2

Organizations that introduce black box models are mandated to report the accuracy of interpretable modelling methods.

PRO

1. Identify the accuracy and/or interpretability trade-off
2. Encourages use/development of interpretable models

CONS

1. Longer development period (finding and/or interpretable model for comparison)
2. Might reduce industrial participation

THE HILL

OPINION > TECHNOLOGY

How dangerous is AI? Regulate it before it's too late

BY CYNTHIA RUDIN, OPINION CONTRIBUTOR - 02/08/23 5:00 PM ET

Conclusion

Rudin proposes why we should strive for interpretable models (especially for HIGH STAKE DECISIONS):

- **Rashomon set argument:** consider that the data permit a large set of reasonably accurate predictive models to exist. Because this set of accurate models is large, we should expect at least one model that is interpretable.

- If there are many diverse yet good models, it means that algorithms may not be stable; an algorithm might choose one model, and a small change to that algorithm or to the data set may yield a completely different (but still accurate) model.

CAUSAL INTERPRETATIONS OF BLACK-BOX MODELS

Qingyuan Zhao [1], Trevor Hastie [1]

Overview:

- Background & Motivation
- Partial Dependence Plots (PDP)
- Mediation Analysis
- Conclusions & Discussion

Objectives:

- When & how can we draw causal interpretations from black box algorithms?
- Mediation Analysis: Causal inference under uncertain causal graph
- Useful tools: PDP & Individual Conditional Expectation (ICE)

Motivation

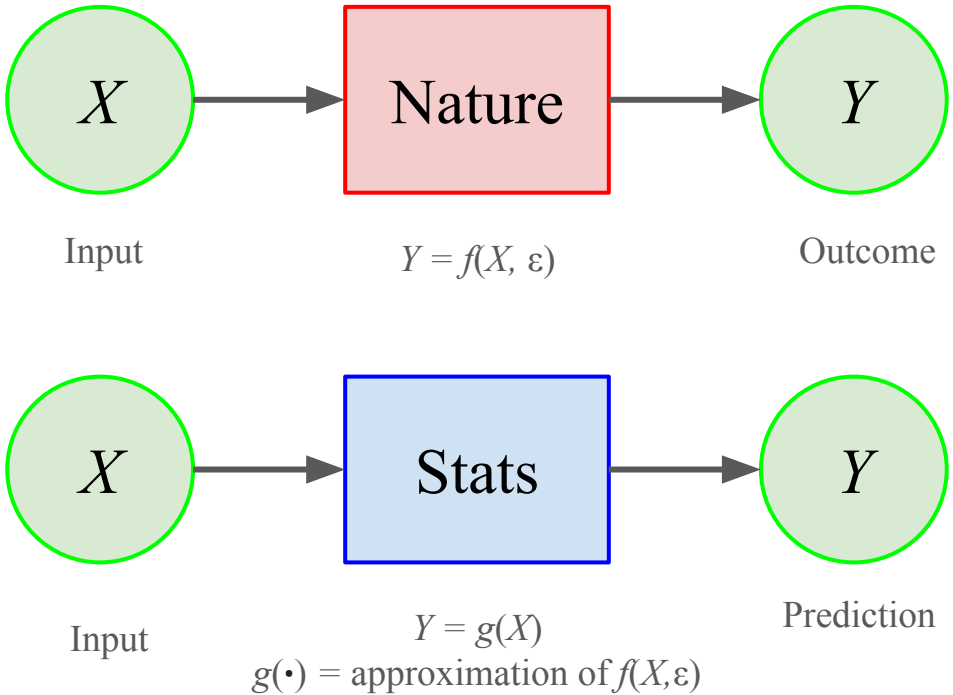
Nature: Some Assumptions

Input

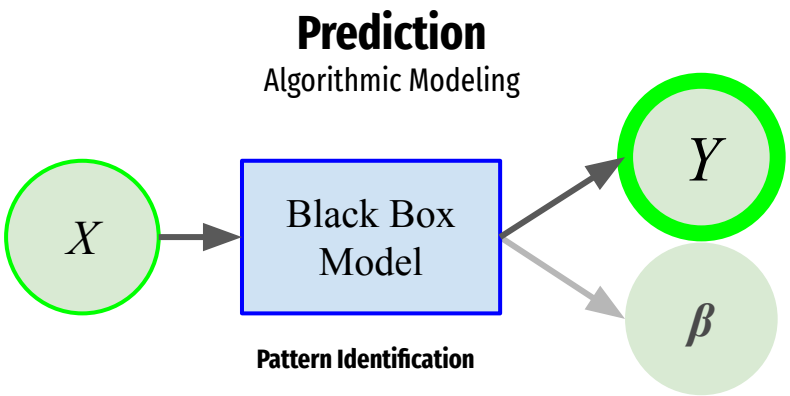
$Y = f(X,\varepsilon)$
ε = error
$f(\cdot)$ = law of nature

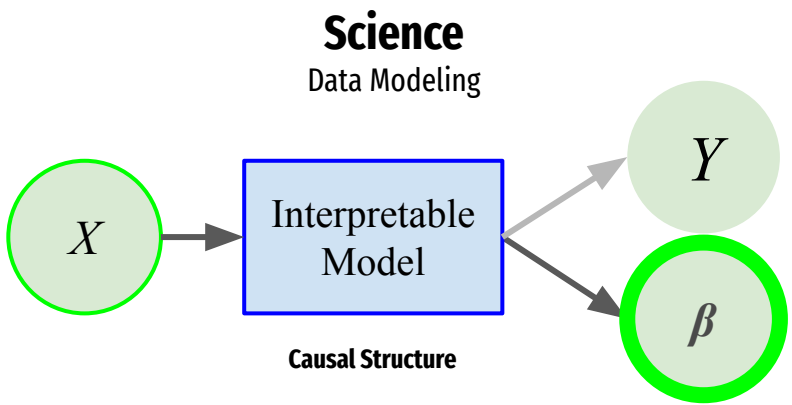
Outcome

Statistics: Modeling $f(X, \varepsilon)$



Competing Objectives & Cultures



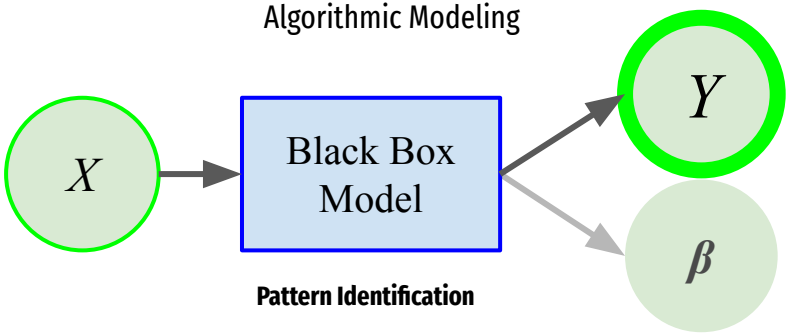


Why trust data modeling as the proper model of nature when it is often less accurate? (Breiman, 2001)

Notions of Importance

Prediction

Algorithmic Modeling

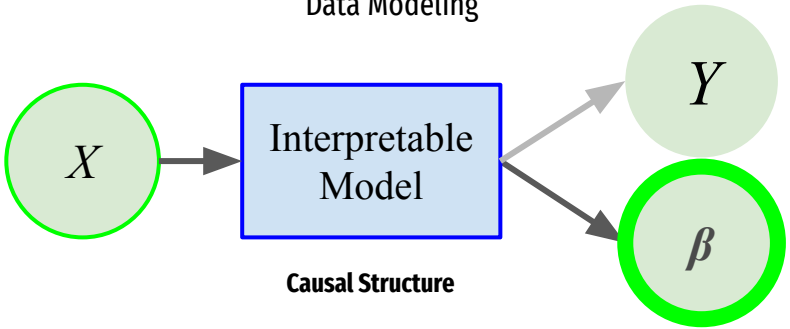


Important Variables:

- **Determined by association**
- Have high impact on model variance
- Are crucial to model performance

Science

Data Modeling

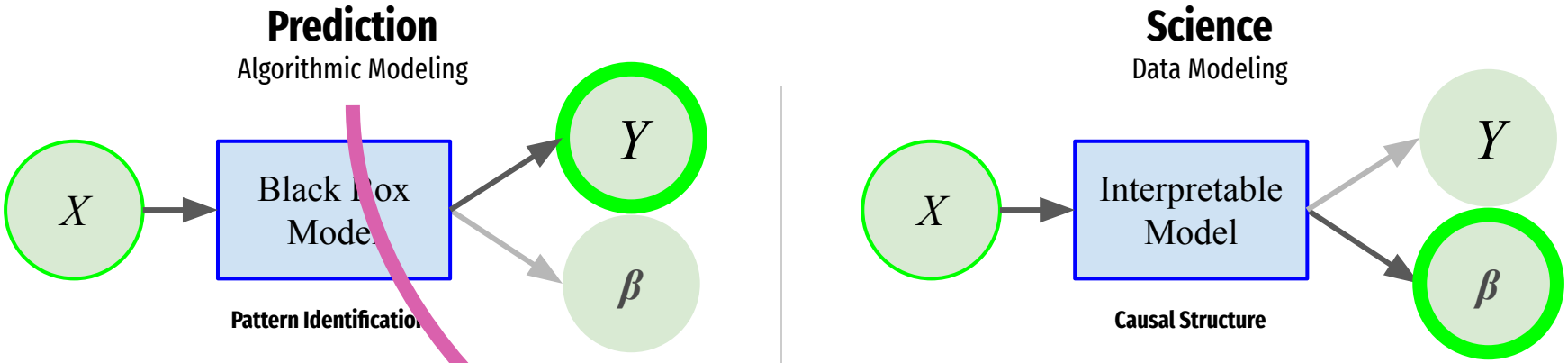


Important Variables:

- Focused on implications for **intervention or counterfactual reasoning**
- Are causally related to outcomes

How can we leverage the powerful algorithmic tools used in prediction to understand natural relationships? (Zhao, Hastie)

Notions of Importance



Important Variables:

- **Determined by association**
- Have high impact on model variance
- Are crucial to model performance

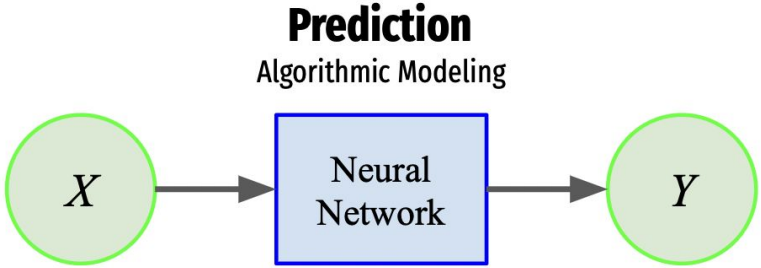
Important Variables:

- Focused on implications for **intervention or counterfactual reasoning**
- Are causally related to outcomes

How can we leverage the powerful algorithmic tools used in prediction to understand natural relationships? (Zhao, Hastie)

Feature Importance in Black Box Models

Partial Dependence Plots



Important Variables:

- Have high impact on model variance
- Are crucial to model performance

Partial Dependence Plot:

- Characterizes average relationship between feature and model prediction
- Model agnostic

Partial Dependence Plots

Expected value of g(X), marginalizing XS across the all other features

$$g_{\mathcal{S}}\left(x_{\mathcal{S}}\right) = \mathrm{E}_{X_{\mathcal{C}}}\left[g\left(x_{\mathcal{S}}, X_{\mathcal{C}}\right)\right]$$

$$= \int g\left(x_{\mathcal{S}}, x_{\mathcal{C}}\right) \mathrm{d}P\left(x_{\mathcal{C}}\right)$$

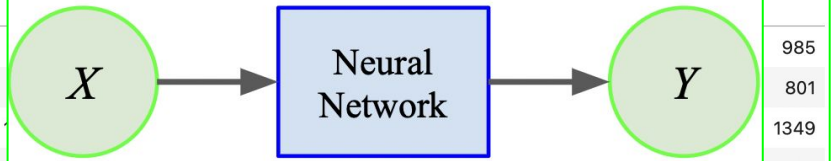
$\mathbf{X}_S$: Variable of interest
$\boldsymbol{X}_C$: Complement of $\mathbf{X}_S$

Partial Dependence Plots: Example

How does the temperature affect predicted bike rentals?

Dataset: 2 years of bike rental data

	X_S	X_C		
day	temp	hum	windspeed	cnt
0	0.34	0.805833	0.160446	985
1	0.36	0.696087	0.248539	801
2	0.20	0.437273	0.248309	1349
3	0.20	0.590435	0.160296	1562
4	0.23	0.436957	0.186900	1600



Outcome: Total daily bike rentals

X: Temperature, humidity, windspeed

Partial Dependence Plots: Example

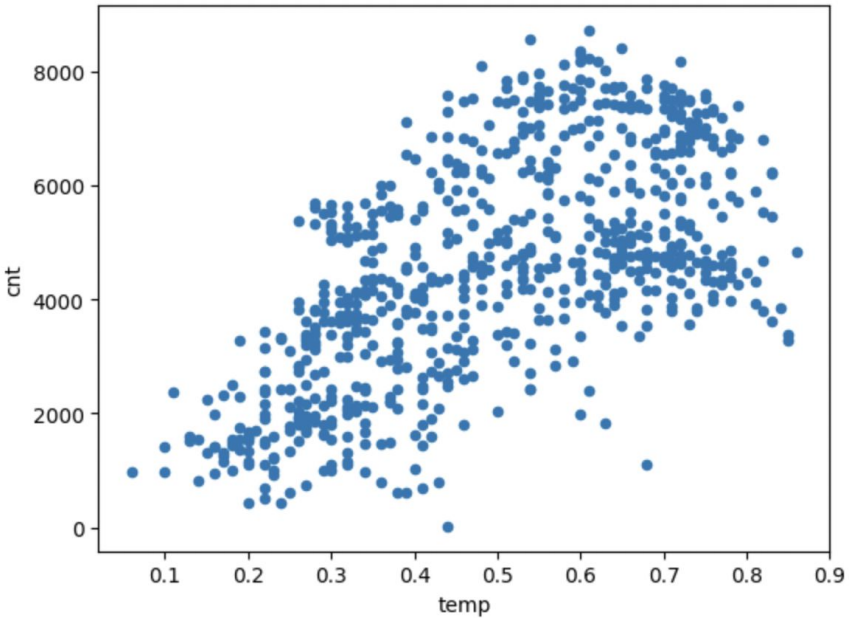
Calculating Partial Dependence Plot

1. Train model on original dataset
2. Upsample dataset to generate N observations for each unique value in $\mathbf{x}_S$ value
3. Get predictions for upsampled new dataset, and average each value of $\mathbf{x}_S$

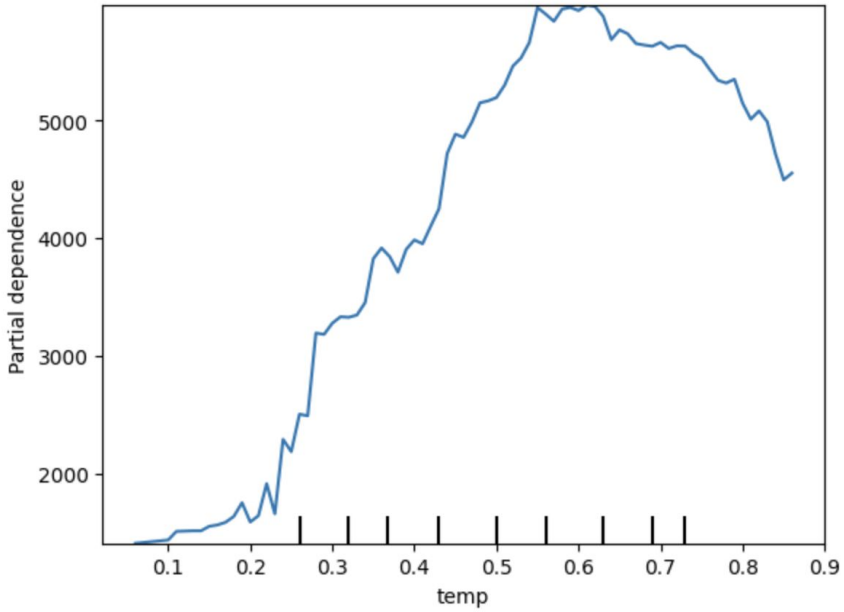
day	temp	hum	windspeed
0	0.34	0.805833	0.160446
1	0.36	0.696087	0.248539
2	0.20	0.437273	0.248309
3	0.20	0.590435	0.160296

Partial Dependence Plots: Example

Scatter Plot: Temperature and Bike Sales



Predicted Bike Sales By Temperature



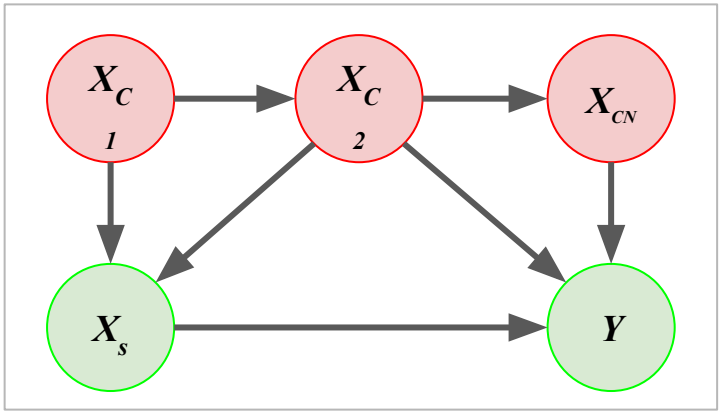
Partial Dependence Plots: The Math

Expected value of g(X), marginalizing X_S across the all other features

$$g_{\mathcal{S}}\left(x_{\mathcal{S}}\right) = \mathrm{E}_{X_{\mathcal{C}}}\left[g\left(x_{\mathcal{S}}, X_{\mathcal{C}}\right)\right]$$

$$= \int g\left(x_{\mathcal{S}}, x_{\mathcal{C}}\right) \mathrm{d}P\left(x_{\mathcal{C}}\right)$$

Revisiting the Backdoor Criterion



If:

1. No node in X_C a descendent of $\mathbf{X_S}$
2. X_C d-separates $\mathbf{X_S}$ and $\mathbf{Y}$?

Then:

$$\mathrm{E}[Y|do(X_S = x_S)] = \int \mathrm{E}[Y|X_S = x_S, X_C = x_C]\,\mathrm{d}P(x_C)$$

An Interesting Coincidence? 🤔 💡

If Backdoor Criterion are Satisfied:

1. Model is causally structured
2. No node in X_C a descendent of $\mathbf{X}_S$
3. X_C d-separates $\mathbf{X}_S$ and $\mathbf{Y}$

Then:

$$\mathrm{E}[Y|do(X_S = x_S)] = \int \mathrm{E}[Y|X_S = x_S, X_C = x_C]\,\mathrm{d}P(x_C)$$

$$\implies \mathrm{E}\left[Y|do(X_S = x_s)\right] = g_S(X_S)$$

An Interesting Coincidence? 🤔 💡

If Backdoor Criterion are Satisfied:

1. Model is causally structured
2. No node in X_C a descendent of $\mathbf{X}_S$
3. X_C d-separates $\mathbf{X}_S$ and $\mathbf{Y}$

Then:

$$\mathrm{E}[Y|do(X_S = x_S)] = \int \mathrm{E}[Y|X_S = x_S, X_C = x_C]\,\mathrm{d}P(x_C)$$

$$\implies \mathrm{E}\left[Y|do(X_S = x_s)\right] = g_S(X_S)$$

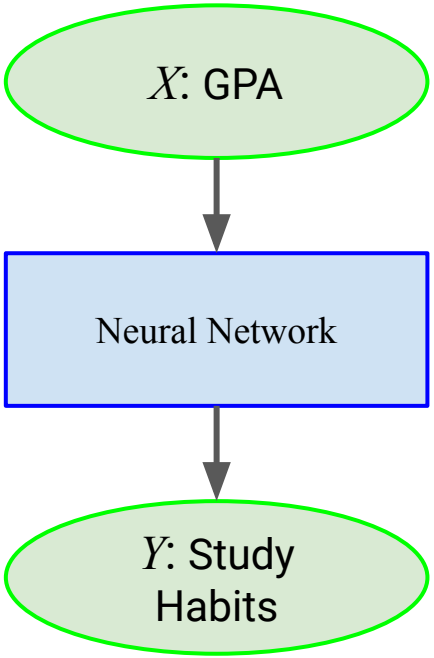
Under the right conditions,
PDP allows for causal inference in black box models!

Example 1: GPA->Study Habits

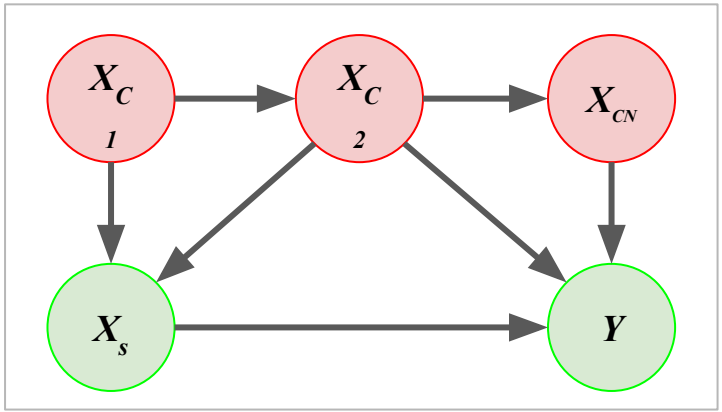
$\mathbf{x}$ $\hat{y}$

	GPA	Predicted Hours Studying Per Week
Student 1	2.82	17
Student 2	4.00	50
...		
Student N	3.58	25

Can PDP extract a causal interpretation between GPA and predicted study habits?

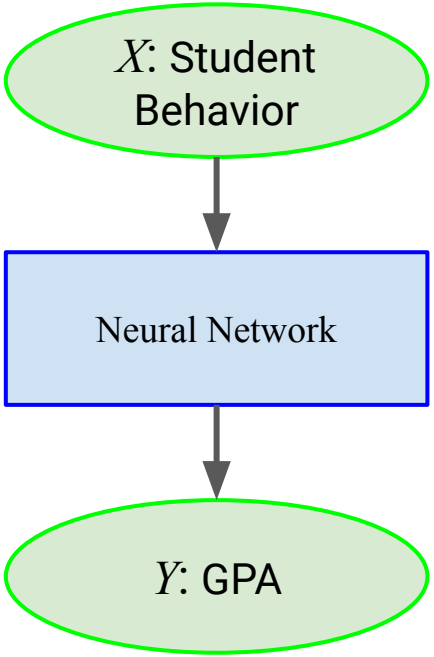


Example 2: Study Habits->GPA

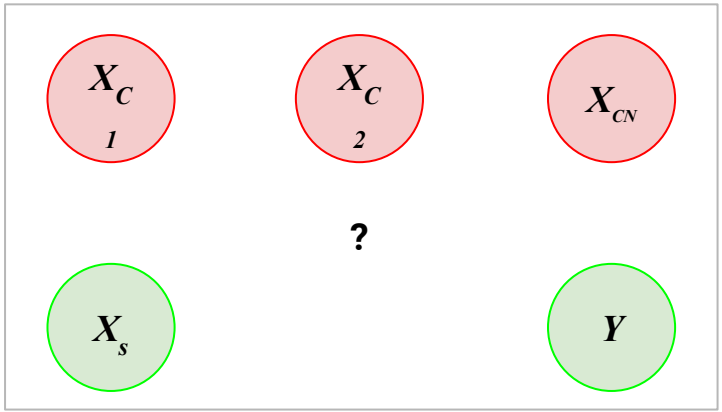


X: Student behavior data (weekly hours studying, sleep habits, department, etc)
X_S: Hours studying
X_C: Student Behavior Data (hours studying, sleep habits, department, etc)
Y: GPA

Can PDP extract a causal interpretation between study and predicted GPA?
Assume the DAG to be accurate

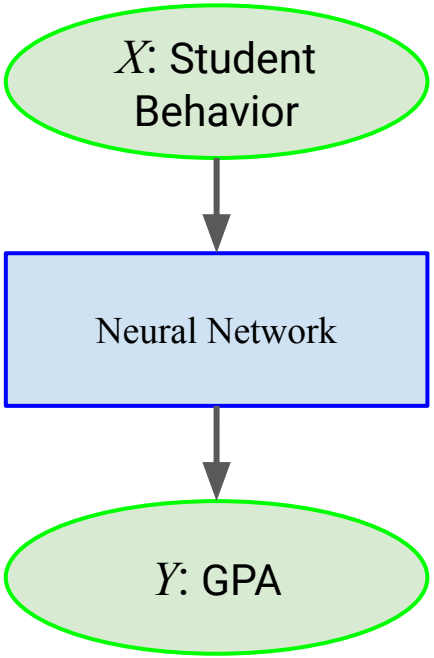


Example 3: Uncertainty

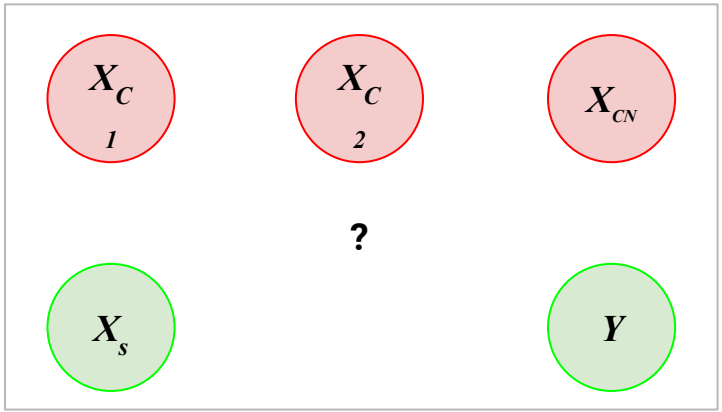


X: Student behavior data (weekly hours studying, sleep habits, department, etc)
X_S: Hours studying
X_C: Student Behavior Data (hours studying, sleep habits, department, etc)
Y: GPA

Can PDP extract a causal interpretation between study and predicted GPA?



Boston Housing: Will People Pay for Better Air?

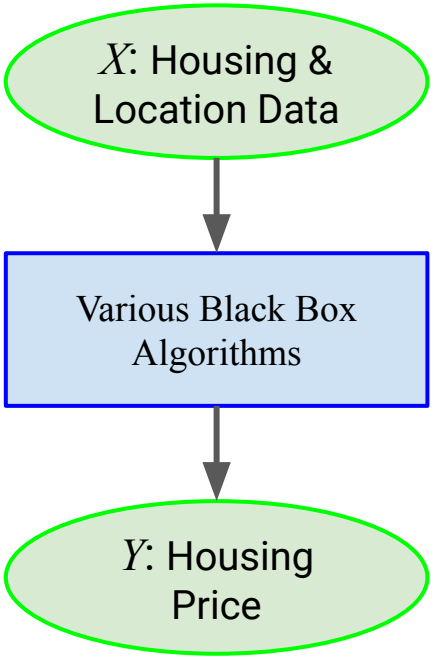


X: Housing & Location Data

X_S: Air Quality

X_C: Remaining Housing Info

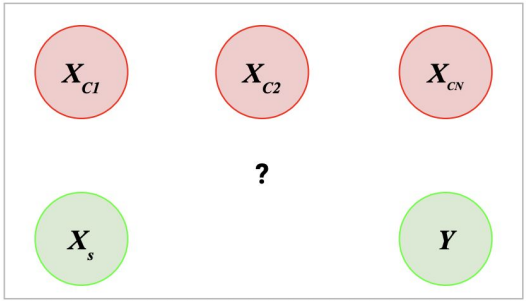
Y: Median House Price



Author's Assumptions:

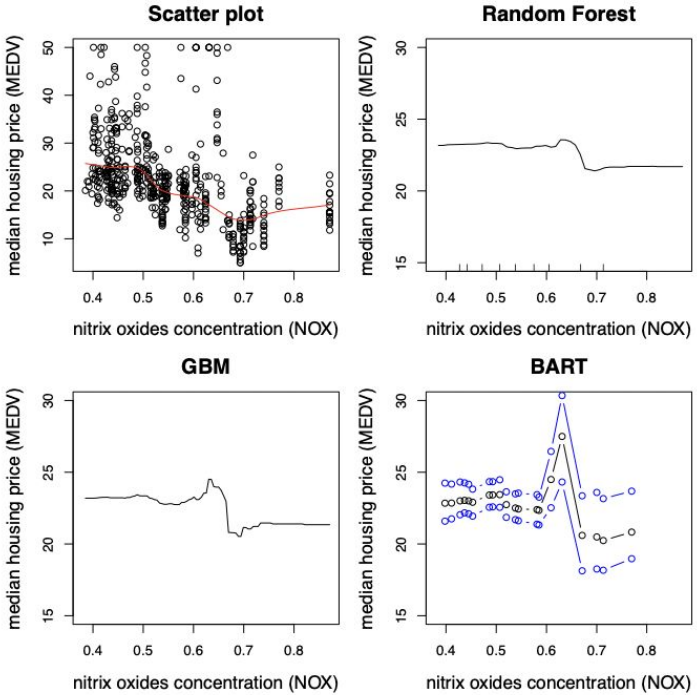
1. Data causally structured? **Yes**
2. No node in X_C a descendent of $\mathbf{X}_S$? **Likely**
3. X_C d-separates $\mathbf{X}_S$ and $\mathbf{Y}$? **Maybe**

Findings: Will People Pay for Better Air?



X: Student behavior data (weekly hours studying, sleep habits, department, etc)
X_S: Hours studying
X_C: Student Behavior Data (hours studying, sleep habits, department, etc)
Y: GPA

Author's Assumptions
Data causally structured? **Yes**
No node in X_C a descendent of $\mathbf{X_S}$? **Likely**
X_C d-separates $\mathbf{X_S}$ and $\mathbf{Y}$? **Maybe**



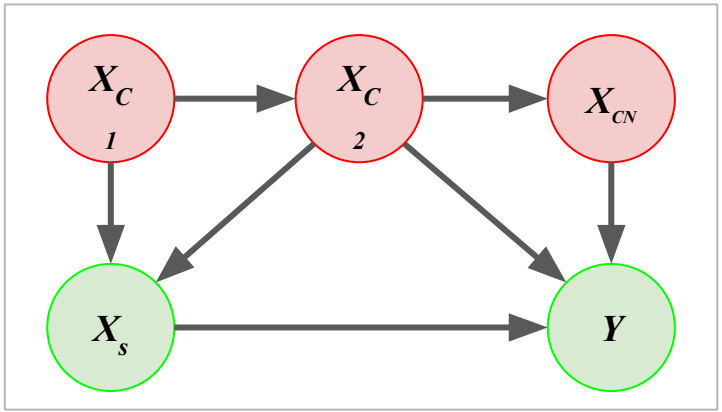
Author's Conclusion:

Plausible evidence of causal nonsmooth relationship

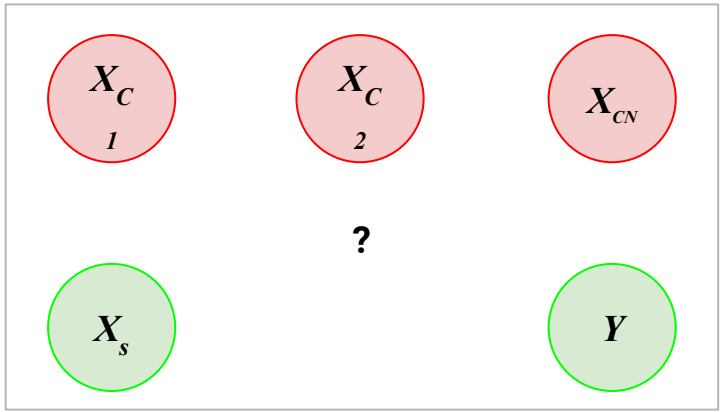
Additional analysis required to make any causal claim!

Finer Analysis

Required Assumptions

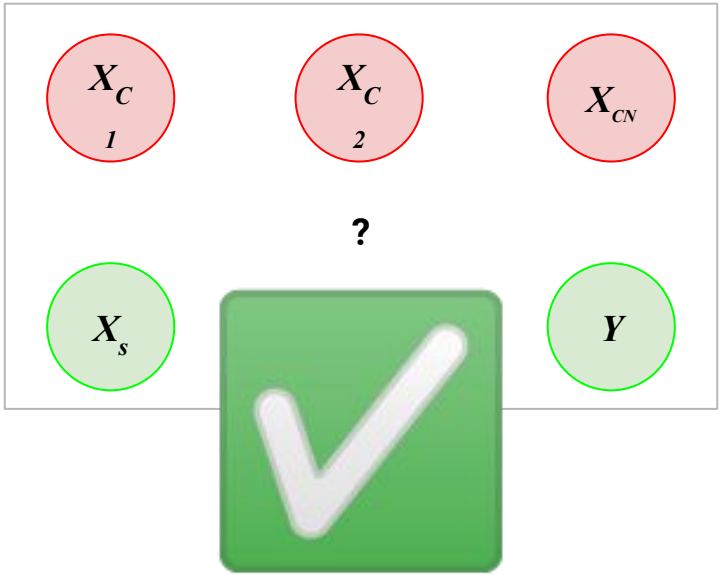


Real Life

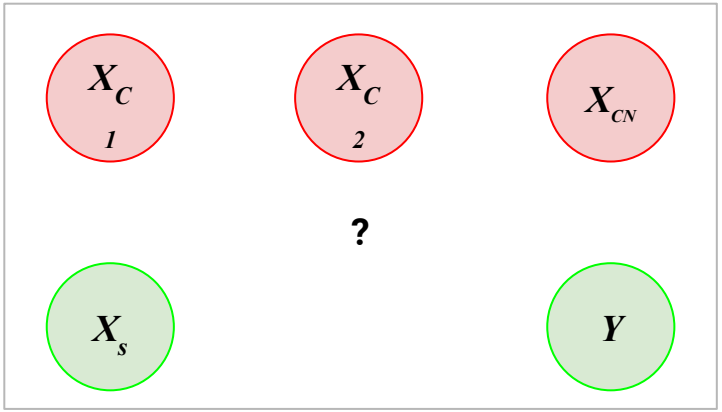


What if the Backdoor Criteria are Not Verifiable

Option A: Accepting the Uncertainty

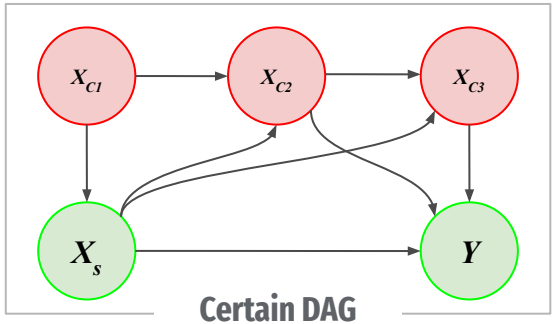


Option B: Probing for further evidence

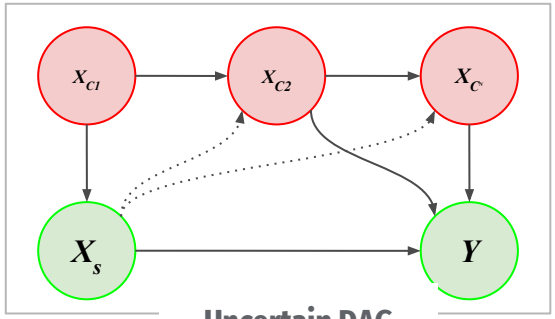


Mediation Analysis ✅

What if X_c contains (or may contain) descendants of X_s?



Certain DAG



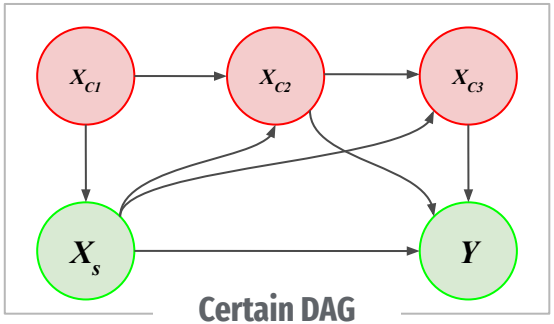
Uncertain DAG

Causal Interpretations:

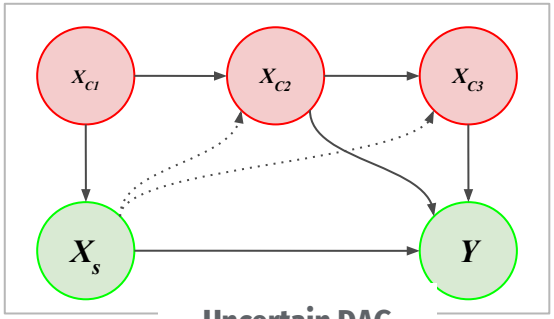
Total Effect: The causal impact of X_s on Y in total
Controlled Direct Impact: The causal impact of $\mathbf{X_s}$ on **Y** in for a fixed value of descendent nodes

Mediation Analysis ✅

What if X_c contains (or may contain) descendants of X_s?



Certain DAG



Uncertain DAG

Causal Interpretations:

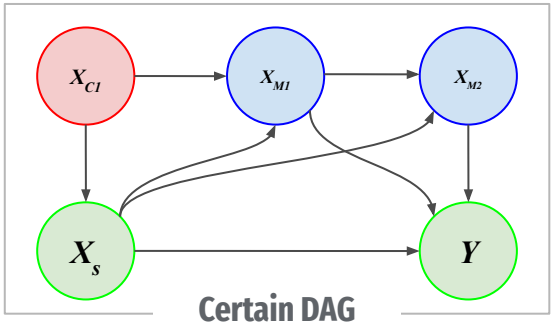
Total Effect: The causal impact of X_s on Y in total
Controlled Direct Impact: The causal impact of $\mathbf{X_s}$ on **Y** in for a fixed value $\mathbf{X_M = x_M}$

Notation:

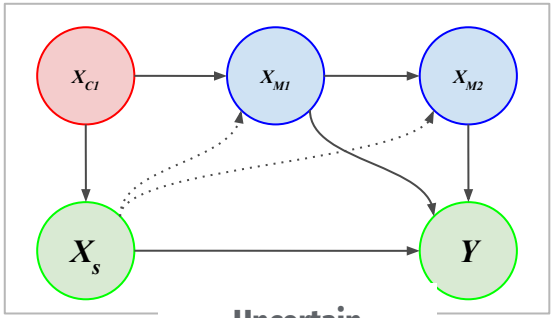
X_S: Variable of interest
X_M: Causal descendent variables of X_S
X_C: Set of variables assumed to satisfy backdoor criterion for set X_{S+M}
$X_M = h(X_S, X_C, \varepsilon_M)$
$Y = g(X_S, X_C, X_M, \varepsilon)$

Mediation Analysis

What if X_c contains (or may contain) descendants of X_s?



Certain DAG



Uncertain
Mediated DAG

Causal Interpretations:

Total Effect: The causal impact of X_s on Y in total

$$\mathrm{TE} = \mathrm{E}[f(x_{\mathcal{S}}, h(x_{\mathcal{S}}, X_{\mathcal{C}}, \epsilon_{\mathcal{M}}), X_{\mathcal{C}}, \epsilon)] - \mathrm{E}[f(x'_{\mathcal{S}}, h(x'_{\mathcal{S}}, X_{\mathcal{C}}, \epsilon_{\mathcal{M}}), X_{\mathcal{C}}, \epsilon)]$$

Controlled Direct Effect: The causal impact of $\mathbf{X_s}$ on **Y** in for a fixed value $\mathbf{X_M = x_M}$

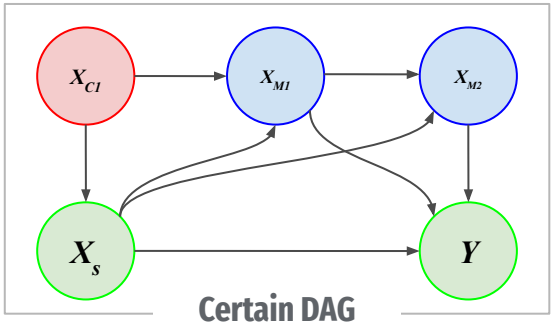
$$\mathrm{CDE}(x_{\mathcal{M}}) = \mathrm{E}[f(x_{\mathcal{S}}, x_{\mathcal{M}}, X_{\mathcal{C}}, \epsilon)] - \mathrm{E}[f(x'_{\mathcal{S}}, x_{\mathcal{M}}, X_{\mathcal{C}}, \epsilon)]$$

Notation:

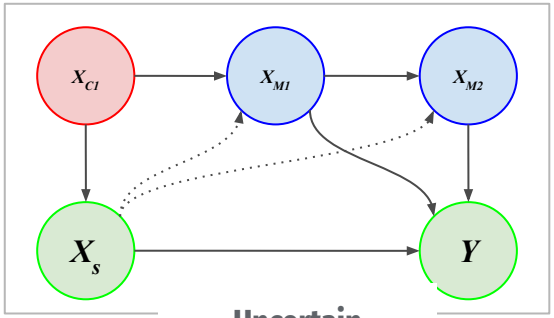
X_S: Variable of interest
X_M: Causal descendent variables of X_S
X_C: Set of variables assumed to satisfy backdoor criterion for set X_{S+M}
$X_M = h(X_S, X_C, \varepsilon_M)$
$Y = g(X_S, X_C, X_M, \varepsilon)$

Mediation Analysis

What if X_c contains (or may contain) descendants of X_s?



Certain DAG



Uncertain Mediated DAG

Causal Interpretations:

Total Effect: The causal impact of X_s on Y in total
Controlled Direct Impact: The causal impact of $\mathbf{X_s}$ on **Y** in for a fixed value of descendent nodes

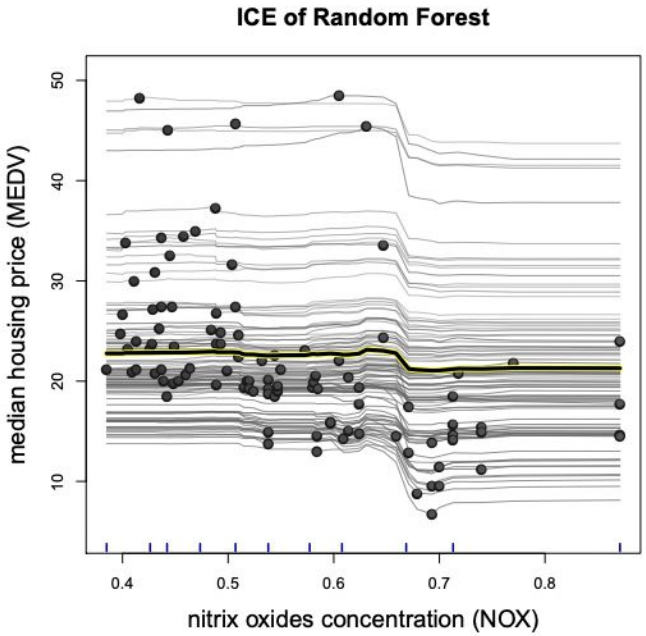
Important :

If X_C does satisfies backdoor criterion for X_S, TE=CDE
If X_C does satisfies backdoor criterion for X_S, **PDP** can visualize TE
If TE≠CDE, PDP cannot visualize CDE

Individual Conditional Expectation (ICE) 🔎

Searching for Mediating Variables

ICE



Key Idea

- **ICE** Marginalizes $\mathrm{E}(g(\mathbf{X}))$ across individual $\mathbf{X}_{\mathbf{Ci}}$
- **PDP is the average of each ICE,** as it marginalizes $\mathrm{E}(g(\mathbf{X}))$ across whole $\mathbf{X}_{\mathbf{C}}$
- **Consistent curves** provides evidence that **no** $\mathbf{X}_{\mathbf{Ci}}$ **is mediating t**he relationship between $\mathbf{X}_{\mathbf{C}}$ and **Y**
- **Inconsistent curves** suggest evidence there are **mediating** $\mathbf{X}_{\mathbf{Ci}}$ **variables**

Boston

- Consistent ICE lines show an additive relationship between NOX and MEDV
- NOX ***may*** have a non-smooth causal effect on MEDV

Conclusion & Discussion

Conclusions

ML algorithms are not deterministically uninterpretable. Under following conditions we can derive causal interpretations from black box models:

1. A good predictive model, so the estimated black-box function g is (hopefully) close to the law of nature f.
2. Some domain knowledge about the causal structure to assure the back-door condition is satisfied. **Not Trivial!**
3. Access to visualization tools such as the PDP and its extension ICE

Given the author's analysis of housing prices in Boston, do you consider their models of housing price to be interpretable?