Sustainability in Explainable AI

Public lecture series Sustainability in Computer Science

Moritz Grosse-Wentrup
Research Group Neuroinformatics
Faculty of Computer Science
University of Vienna

November 27, 2023

universität
wien

The General Data Protection Regulation (GDPR)

The data subject should have the right [...] to obtain an ***explanation of the decision reached*** *[...] and to* ***challenge the decision****.*

Correctional Offender Management Profiling for Alternative Sanctions

Correctional Offender Management Profiling for Alternative Sanctions

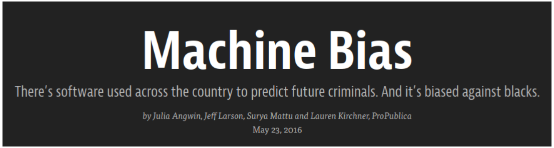
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions.

`https://en.wikipedia.org/wiki/COMPAS_(software)`

Correctional Offender Management Profiling for Alternative Sanctions

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions.

`https://en.wikipedia.org/wiki/COMPAS_(software)`



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

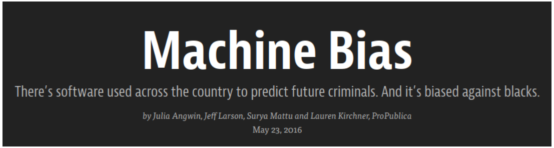
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Correctional Offender Management Profiling for Alternative Sanctions

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions.

`https://en.wikipedia.org/wiki/COMPAS_(software)`

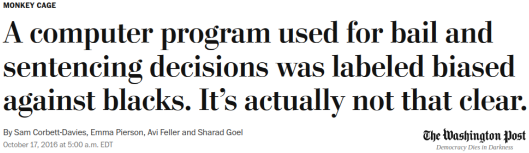


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



MONKEY CAGE

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

October 17, 2016 at 5:00 a.m. EDT

The Washington Post

Democracy Dies in Darkness

A Machine Learning Primer

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \ldots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \ldots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \dots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \dots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$
- Labels Y with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \ldots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$
- Labels Y with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \ldots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$
- Labels Y with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \boldsymbol{x}, y)$, e.g., zero-one loss.

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \ldots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$
- Labels Y with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \boldsymbol{x}, y)$, e.g., zero-one loss.
- Training set $\mathcal{S} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$ sampled iid from $P(Y, \boldsymbol{X})$

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \ldots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$
- Labels Y with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \boldsymbol{x}, y)$, e.g., zero-one loss.
- Training set $\mathcal{S} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$ sampled iid from $P(Y, \boldsymbol{X})$
- Learning algorithm $A(\mathcal{S}, \mathcal{H}, l)$

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \ldots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$
- Labels Y with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \boldsymbol{x}, y)$, e.g., zero-one loss.
- Training set $\mathcal{S} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$ sampled iid from $P(Y, \boldsymbol{X})$
- Learning algorithm $A(\mathcal{S}, \mathcal{H}, l)$
- Bayes-optimal hypothesis $h^* \in \mathcal{H}$ with (empirical) risk $L_{(\mathcal{S})}(h^*)$

A Machine Learning Primer

- Features $\boldsymbol{X} = [X_1, \ldots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$
- Labels Y with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \boldsymbol{x}, y)$, e.g., zero-one loss.
- Training set $\mathcal{S} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$ sampled iid from $P(Y, \boldsymbol{X})$
- Learning algorithm $A(\mathcal{S}, \mathcal{H}, l)$
- Bayes-optimal hypothesis $h^* \in \mathcal{H}$ with (empirical) risk $L_{(\mathcal{S})}(h^*)$
- Predictions $\hat{Y} = h^*(\hat{\boldsymbol{X}})$

A Machine Learning Primer

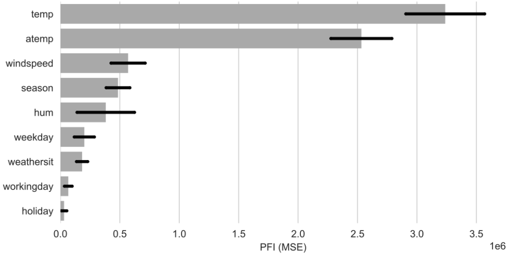
- Features $\boldsymbol{X} = [X_1, \dots, X_d]$; $\boldsymbol{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\boldsymbol{X}} = [\hat{X}_1, \dots, \hat{X}_d]$; $\boldsymbol{x} \in \mathcal{X}$
- Labels Y with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \dots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \boldsymbol{x}, y)$, e.g., zero-one loss.
- Training set $\mathcal{S} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$ sampled iid from $P(Y, \boldsymbol{X})$
- Learning algorithm $A(\mathcal{S}, \mathcal{H}, l)$
- Bayes-optimal hypothesis $h^* \in \mathcal{H}$ with (empirical) risk $L_{(\mathcal{S})}(h^*)$
- Predictions $\hat{Y} = h^*(\hat{\boldsymbol{X}})$
- Interventions do$\{\boldsymbol{X} = ()\}$ (on features) and do$\{\hat{\boldsymbol{X}} = ()\}$ (on measurements)

Example: The Permutation Feature Importance Score

Idea: Assess how removing each individual features affects model performance.

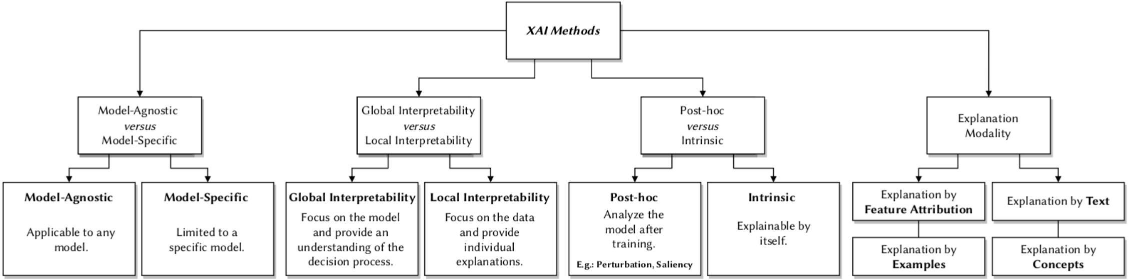
$$\mathrm{PFI}(X_i) = L(h^*, \{\boldsymbol{X} \backslash X_i, \tilde{X}_i\}, Y) - L(h^*, \boldsymbol{X}, Y)$$

with $\tilde{X}_i \sim P(X_i)$ and $\tilde{X} \perp\!\!\!\perp \{\boldsymbol{X} \backslash X_i, Y\}$.



Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

The xAI/IML Zoo



XAI Methods

Model-Agnostic
versus
Model-Specific

Model-Agnostic
Applicable to any model.

Model-Specific
Limited to a specific model.

Global Interpretability
versus
Local Interpretability

Global Interpretability
Focus on the model and provide an understanding of the decision process.

Local Interpretability
Focus on the data and provide individual explanations.

Post-hoc
versus
Intrinsic

Post-hoc
Analyze the model after training.
E.g.: **Perturbation, Saliency**

Intrinsic
Explainable by itself.

Explanation Modality

Explanation by **Feature Attribution**

Explanation by **Examples**

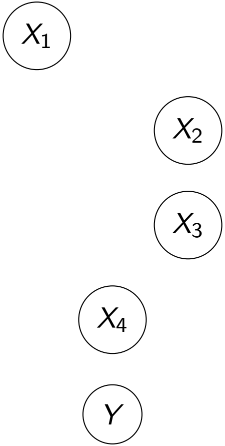
Explanation by **Text**

Explanation by **Concepts**

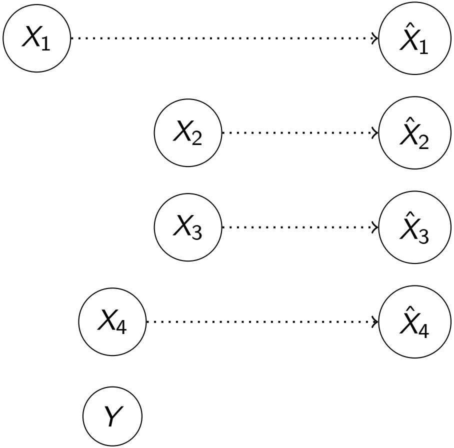
Patricio et al. "Explainable Deep Learning Methods in Medical Image Classification: A Survey." ACM Computing Surveys 56.4 (2023): 1-41.

xAI/IML is a Causal Problem

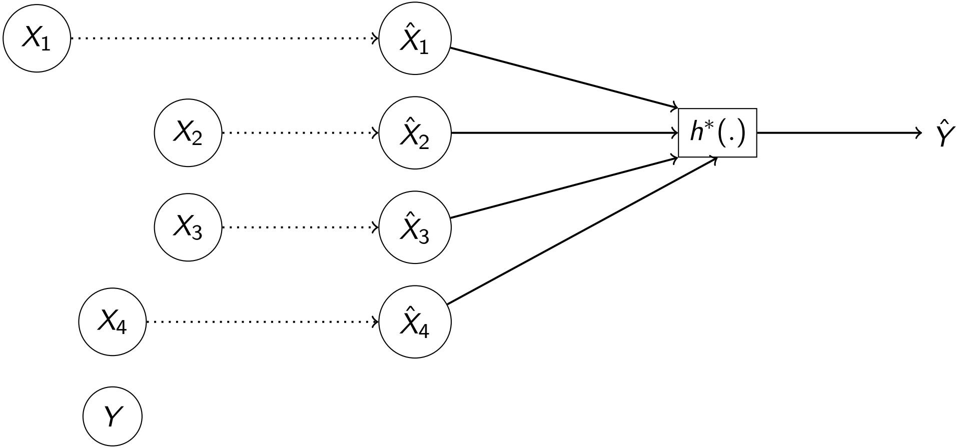
xAI/IML is a Causal Problem



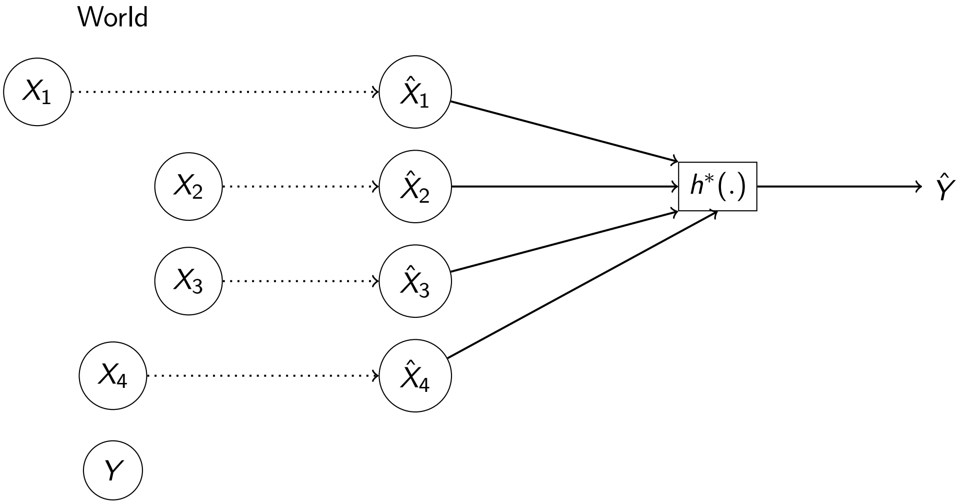
xAI/IML is a Causal Problem



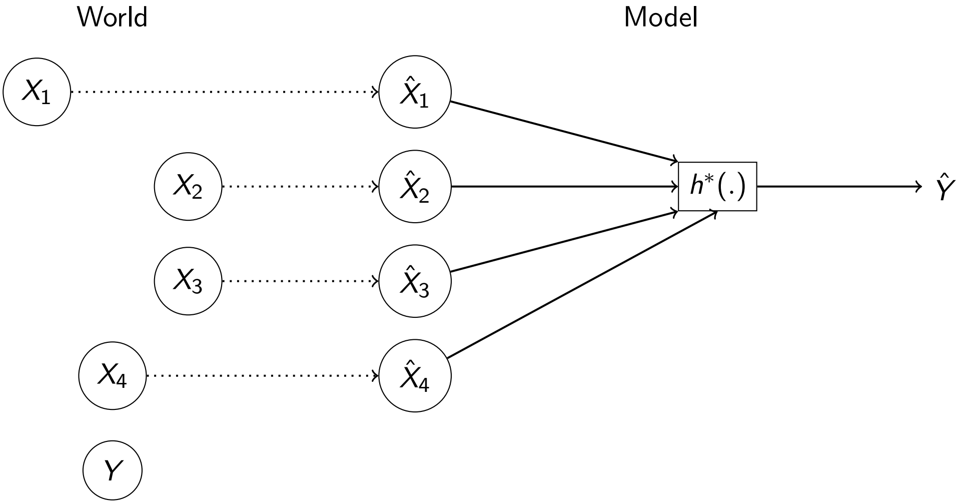
xAI/IML is a Causal Problem



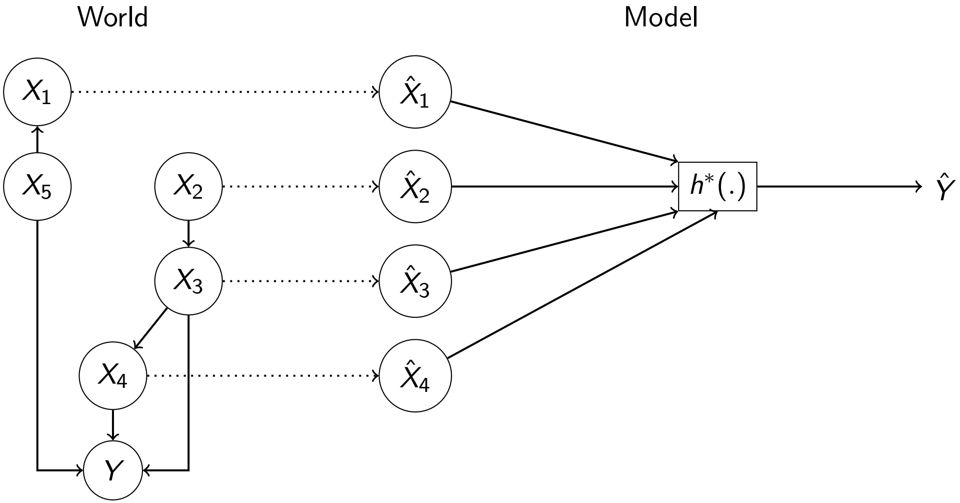
xAI/IML is a Causal Problem



xAI/IML is a Causal Problem



xAI/IML is a Causal Problem



Structural Causal Models (SCMs)

J. Pearl, *Causality*. 2000.

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^{N}$ a *structural causal model* (SCM) is defined by

$$X_i = f_i(\mathrm{pa}_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^{N}$ exogenous noise terms and the *parents* $\mathrm{pa}_i \subset \mathcal{X} \backslash X_i$ chosen such that the corresponding graph contains no cycles.

J. Pearl, *Causality*. 2000.

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a *structural causal model* (SCM) is defined by

$$X_i = f_i(\mathrm{pa}_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the *parents* $\mathrm{pa}_i \subset \mathcal{X} \backslash X_i$ chosen such that the corresponding graph contains no cycles.

Example:

J. Pearl, *Causality*. 2000.

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a *structural causal model* (SCM) is defined by

$$X_i = f_i(\mathrm{pa}_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the *parents* $\mathrm{pa}_i \subset \mathcal{X} \backslash X_i$ chosen such that the corresponding graph contains no cycles.

Example:

SCM

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \epsilon_2 \\ X_3 &= X_1 \cdot X_2 + \epsilon_3 \end{aligned}$$

J. Pearl, *Causality*. 2000.

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a *structural causal model* (SCM) is defined by

$$X_i = f_i(\mathrm{pa}_i, \epsilon_i)$$

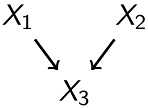
with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the *parents* $\mathrm{pa}_i \subset \mathcal{X} \backslash X_i$ chosen such that the corresponding graph contains no cycles.

Example:

SCM

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \epsilon_2 \\ X_3 &= X_1 \cdot X_2 + \epsilon_3 \end{aligned}$$

DAG



J. Pearl, *Causality*. 2000.

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^{N}$ a *structural causal model* (SCM) is defined by

$$X_i = f_i(\mathrm{pa}_i, \epsilon_i)$$

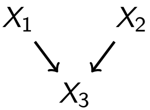
with $\{\epsilon_i\}_{i=1}^{N}$ exogenous noise terms and the *parents* $\mathrm{pa}_i \subset \mathcal{X} \backslash X_i$ chosen such that the corresponding graph contains no cycles.

Example:

SCM

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \epsilon_2 \\ X_3 &= X_1 \cdot X_2 + \epsilon_3 \end{aligned}$$

DAG



Data

$$\begin{aligned} \boldsymbol{\epsilon} &\sim p(\boldsymbol{\epsilon}) \\ x_i &= f_i(\mathrm{pa}_i, \epsilon_i) \\ \boldsymbol{x} &\sim p(\boldsymbol{x}) \end{aligned}$$

J. Pearl, *Causality*. 2000.

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a *structural causal model* (SCM) is defined by

$$X_i = f_i(\mathrm{pa}_i, \epsilon_i)$$

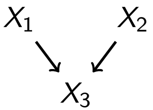
with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the *parents* $\mathrm{pa}_i \subset \mathcal{X} \backslash X_i$ chosen such that the corresponding graph contains no cycles.

Example:

SCM

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \epsilon_2 \\ X_3 &= X_1 \cdot X_2 + \epsilon_3 \end{aligned}$$

DAG



Data

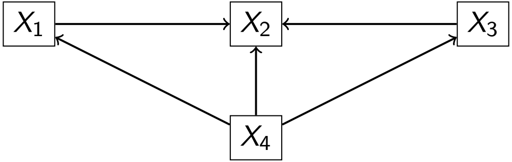
$$\begin{aligned} \boldsymbol{\epsilon} &\sim p(\boldsymbol{\epsilon}) \\ x_i &= f_i(\mathrm{pa}_i, \epsilon_i) \\ \boldsymbol{x} &\sim p(\boldsymbol{x}) \end{aligned}$$

Def.: X_i is a cause of X_j, iff there exist values of X_i and X_j such that $p(x_j|\mathrm{do}\{x_i\}) \neq p(x_j)$.

J. Pearl, *Causality*. 2000.

Causal reasoning with SCMs

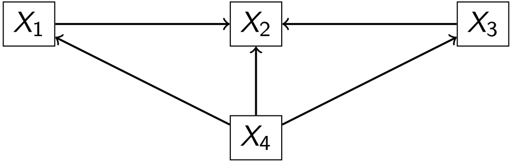
Causal factorization:



(J. Pearl, *Causality*. 2000.)

Causal reasoning with SCMs

Causal factorization:

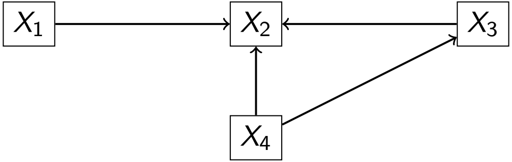


$$P(\boldsymbol{X}) = P(X_2|X_1, X_3, X_4)P(X_1|X_4)P(X_3|X_4)P(X_4)$$

(J. Pearl, *Causality*. 2000.)

Causal reasoning with SCMs

Causal factorization:



$$P(\boldsymbol{X}) = P(X_2|X_1, X_3, X_4)P(X_1|X_4)P(X_3|X_4)P(X_4)$$

Interventions are represented by the do-operator, e.g.,

$$P\left(\boldsymbol{X}|\mathrm{do}(X_1 = x_1)\right) = P(X_2|X_1 = x_1, X_3, X_4)P(X_3|X_4)P(X_4).$$

(J. Pearl, *Causality*. 2000.)

Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal Markov Condition (CMC): Every node in $\mathcal{X}$ is conditionally independent of its nondescendents given its parents.

Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal Markov Condition (CMC): Every node in $\mathcal{X}$ is conditionally independent of its nondescendents given its parents.

Faithfulness: There are no further independence relations among the nodes in $\mathcal{X}$ beyond those implied by d-separation.

Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal Markov Condition (CMC): Every node in $\mathcal{X}$ is conditionally independent of its nondescendents given its parents.

Faithfulness: There are no further independence relations among the nodes in $\mathcal{X}$ beyond those implied by d-separation.

d-separation: Let A, B, C non-intersecting subsets of $\mathcal{X}$. A and B are d-separated given C iff

- for all nodes on the path where the arrows meet head-to-tail ($\rightarrow . \rightarrow$) or tail-to-tail ($\leftarrow . \rightarrow$) the node is in C,
- for all nodes where the arrows meet head-to-head ($\rightarrow . \leftarrow$) neither the node or any of its descendants are in C.

Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal Markov Condition (CMC): Every node in $\mathcal{X}$ is conditionally independent of its nondescendents given its parents.

Faithfulness: There are no further independence relations among the nodes in $\mathcal{X}$ beyond those implied by d-separation.

d-separation: Let A, B, C non-intersecting subsets of $\mathcal{X}$. A and B are d-separated given C iff

- for all nodes on the path where the arrows meet head-to-tail ($\rightarrow . \rightarrow$) or tail-to-tail ($\leftarrow . \rightarrow$) the node is in C,
- for all nodes where the arrows meet head-to-head ($\rightarrow . \leftarrow$) neither the node or any of its descendants are in C.

Assuming the CMC and faithfulness, $\mathrm{dSep}(A, B|C) \Leftrightarrow A \perp\!\!\!\perp B|C$.

Example

Example

The chain

$X_1 \to X_2 \to X_3$

Example

The chain

$$X_1 \rightarrow X_2 \rightarrow X_3$$

The fork

$$X_1 \leftarrow X_2 \rightarrow X_3$$

Example

The chain
$X_1 \to X_2 \to X_3$

The fork
$X_1 \leftarrow X_2 \to X_3$

The collider
$X_1 \to X_2 \leftarrow X_3$

Example

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$		

Example

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$		
$X_1 \perp\!\!\!\perp X_3 \vert X_2$		

Example

The chain

$X_1 \rightarrow X_2 \rightarrow X_3$

$X_1 \not\perp\!\!\!\perp X_3$

$X_1 \perp\!\!\!\perp X_3 | X_2$

The fork

$X_1 \leftarrow X_2 \rightarrow X_3$

$X_1 \not\perp\!\!\!\perp X_3$

The collider

$X_1 \rightarrow X_2 \leftarrow X_3$

Example

The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$	
$X_1 \perp\!\!\!\perp X_3 \| X_2$	$X_1 \perp\!\!\!\perp X_3 \| X_2$	

Example

The chain

$$X_1 \rightarrow X_2 \rightarrow X_3$$

$$X_1 \not\perp\!\!\!\perp X_3$$

$$X_1 \perp\!\!\!\perp X_3 | X_2$$

The fork

$$X_1 \leftarrow X_2 \rightarrow X_3$$

$$X_1 \not\perp\!\!\!\perp X_3$$

$$X_1 \perp\!\!\!\perp X_3 | X_2$$

The collider

$$X_1 \rightarrow X_2 \leftarrow X_3$$

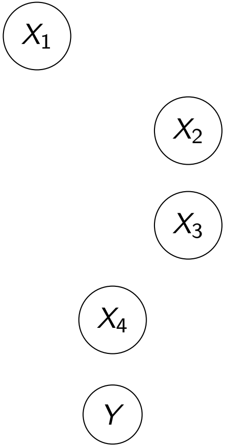
$$X_1 \perp\!\!\!\perp X_3$$

Example

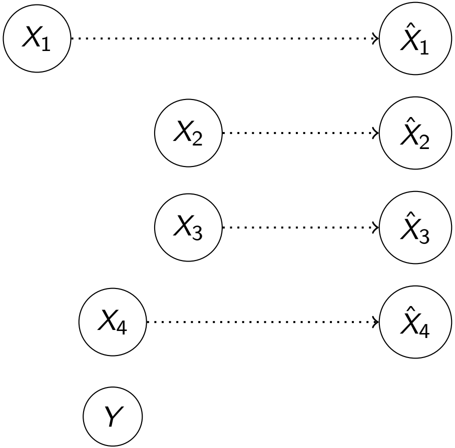
The chain	The fork	The collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3$
$X_1 \perp\!\!\!\perp X_3 \vert X_2$	$X_1 \perp\!\!\!\perp X_3 \vert X_2$	$X_1 \not\perp\!\!\!\perp X_3 \vert X_2$

xAI/IML is a Causal Problem

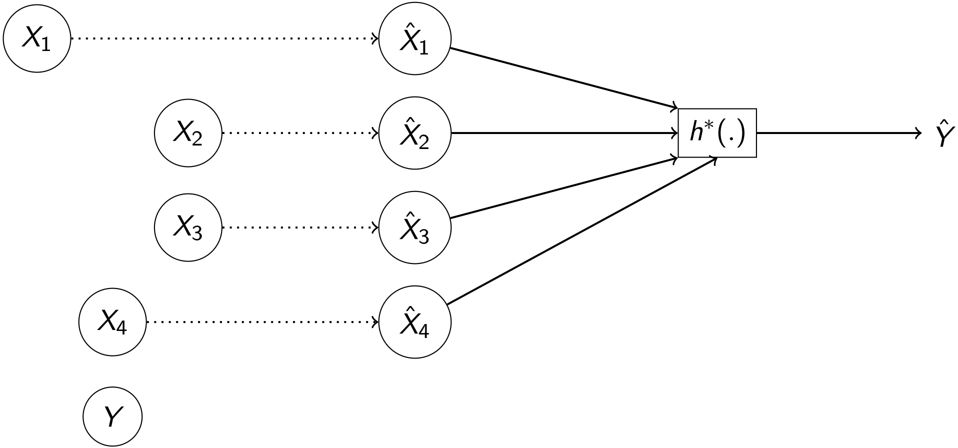
xAI/IML is a Causal Problem



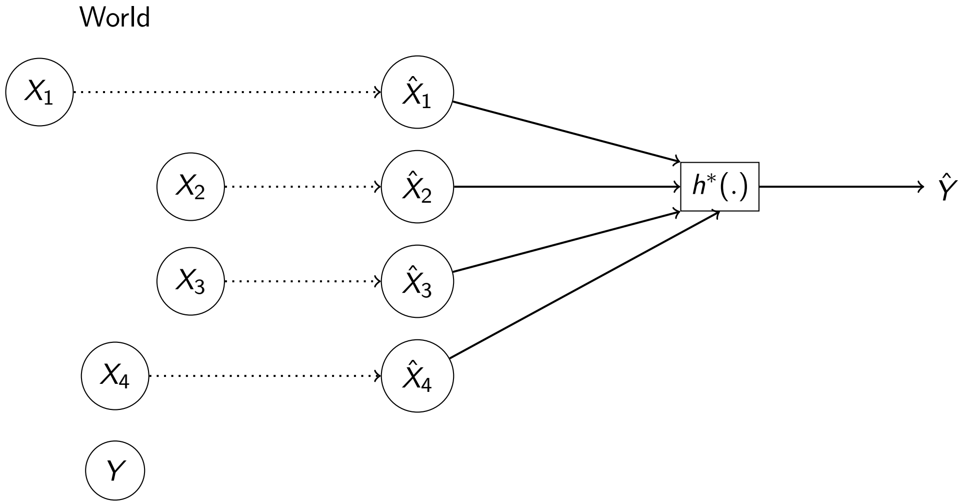
xAI/IML is a Causal Problem



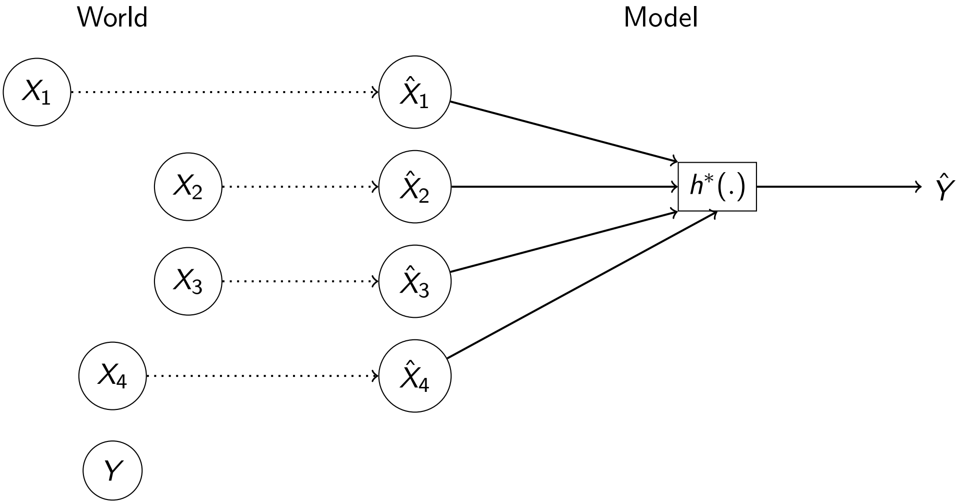
xAI/IML is a Causal Problem



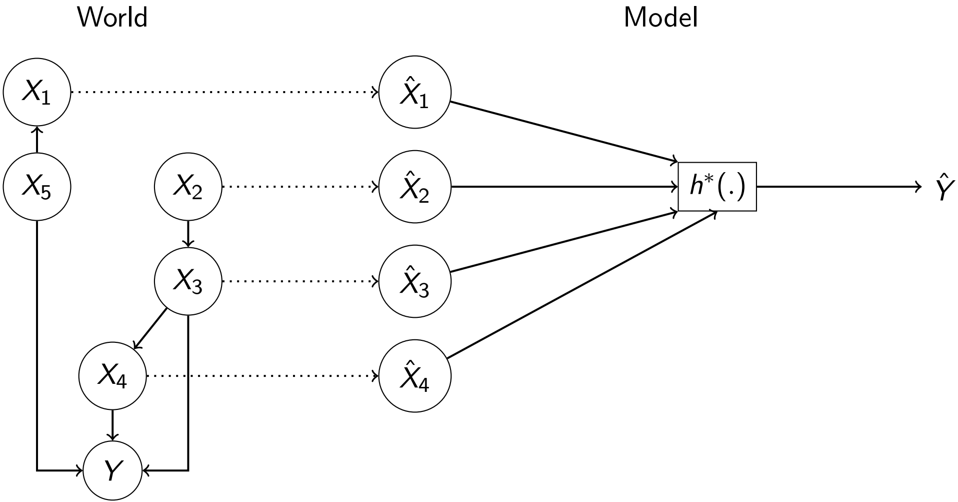
xAI/IML is a Causal Problem



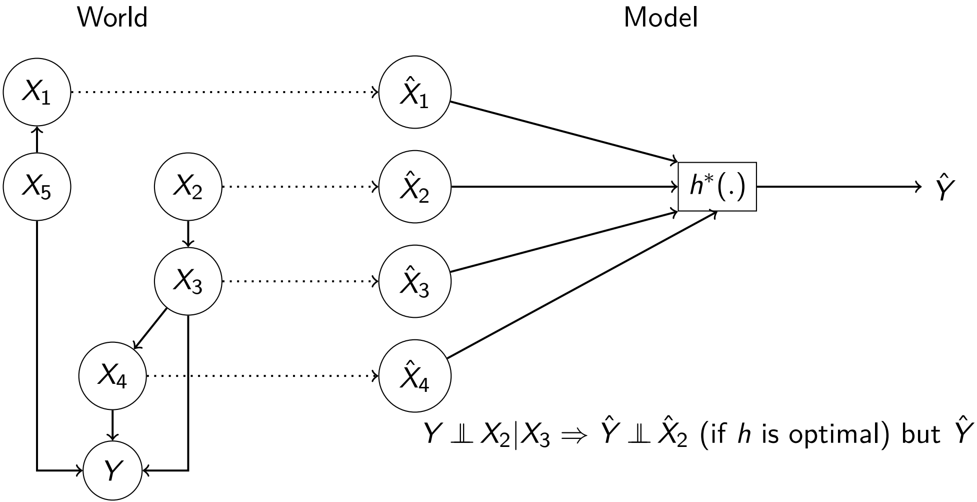
xAI/IML is a Causal Problem



xAI/IML is a Causal Problem

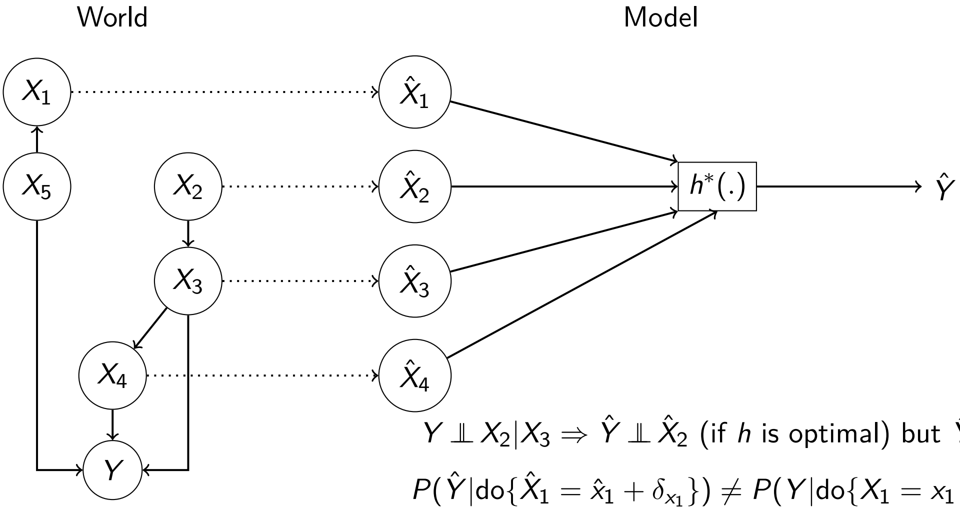


xAI/IML is a Causal Problem



$Y \perp\!\!\!\perp X_2 | X_3 \Rightarrow \hat{Y} \perp\!\!\!\perp \hat{X}_2$ (if h is optimal) but $\hat{Y} \not\!\perp\!\!\!\perp X_2$

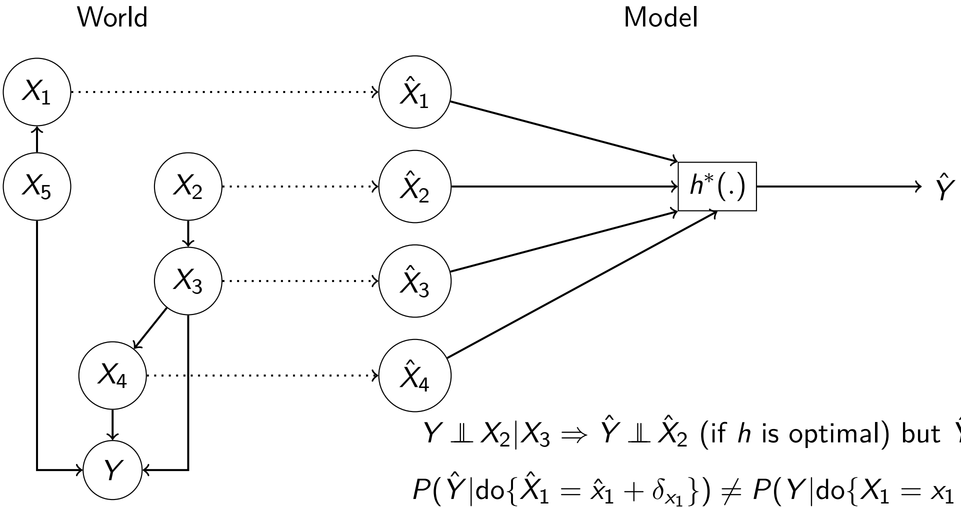
xAI/IML is a Causal Problem



$$Y \perp\!\!\!\perp X_2 | X_3 \Rightarrow \hat{Y} \perp\!\!\!\perp \hat{X}_2 \text{ (if } h \text{ is optimal) but } \hat{Y} \not\!\perp\!\!\!\perp X_2$$

$$P(\hat{Y}|\mathrm{do}\{\hat{X}_1 = \hat{x}_1 + \delta_{x_1}\}) \neq P(Y|\mathrm{do}\{X_1 = x_1 + \delta_{x_1}\})$$

xAI/IML is a Causal Problem



$Y \perp\!\!\!\perp X_2 | X_3 \Rightarrow \hat{Y} \perp\!\!\!\perp \hat{X}_2$ (if h is optimal) but $\hat{Y} \not\perp\!\!\!\perp X_2$

$$P(\hat{Y}|\text{do}\{\hat{X}_1 = \hat{x}_1 + \delta_{x_1}\}) \neq P(Y|\text{do}\{X_1 = x_1 + \delta_{x_1}\})$$

$$P(\hat{Y}|\text{do}\{\hat{X}_3 = \hat{x}_3\}) \neq P(Y|\text{do}\{X_3 = x_3\})$$

A Taxonomy of xAI/IML

A Taxonomy of xAI/IML

What object is explained?

A Taxonomy of xAI/IML

What object is explained?

- The prediction $\hat{Y}$?

A Taxonomy of xAI/IML

What object is explained?

- The prediction $\hat{Y}$?
- The target variable Y?

A Taxonomy of xAI/IML

What object is explained?

- The prediction $\hat{Y}$?
- The target variable Y?
- Their relationship R (as measured by the loss)?

A Taxonomy of xAI/IML

What object is explained?

- The prediction $\hat{Y}$?
- The target variable Y?
- Their relationship R (as measured by the loss)?

On what level is the object explained?

A Taxonomy of xAI/IML

What object is explained?

- The prediction $\hat{Y}$?
- The target variable Y?
- Their relationship R (as measured by the loss)?

On what level is the object explained?

- Associations $X = x$?

A Taxonomy of xAI/IML

What object is explained?

- The prediction $\hat{Y}$?
- The target variable Y?
- Their relationship R (as measured by the loss)?

On what level is the object explained?

- Associations $X = x$?
- Model-level interventions $\mathrm{do}(\hat{X} = x)$?

A Taxonomy of xAI/IML

What object is explained?

- The prediction $\hat{Y}$?
- The target variable Y?
- Their relationship R (as measured by the loss)?

On what level is the object explained?

- Associations $X = x$?
- Model-level interventions $\text{do}(\hat{X} = x)$?
- World-level interventions $\text{do}(X = x)$?

Nine Perspectives on Model and Data

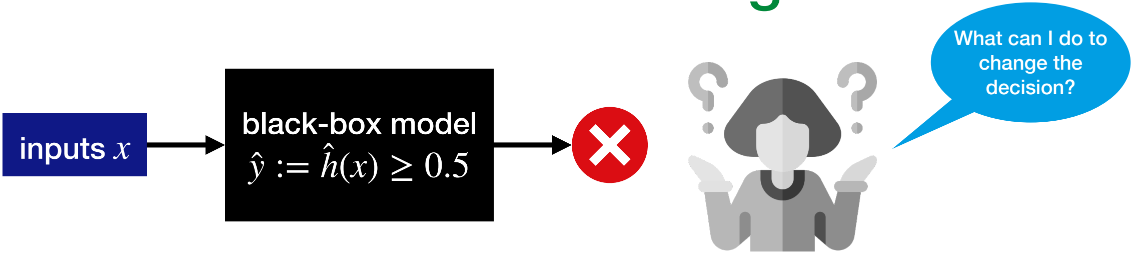
	$\hat{Y}$	R	Y
$do(\underline{X} = x')$	Does the model's mechanism rely on gender?		
$X = x'$			Which biomarkers are correlated with the disease?
$do(X = x')$	What can I do to get accepted?	Does the diagnosis model also work in a different hospital?	What can we do to treat the disease?

Contribution I: Improvement-Focused Causal Recourse (ICR)

König, G., *Freiesleben, T., & Grosse-Wentrup, M. (2023).*
Improvement-focused causal recourse (ICR).
In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 11847-11855).

Recourse Setting

Recourse Setting



inputs x

black-box model

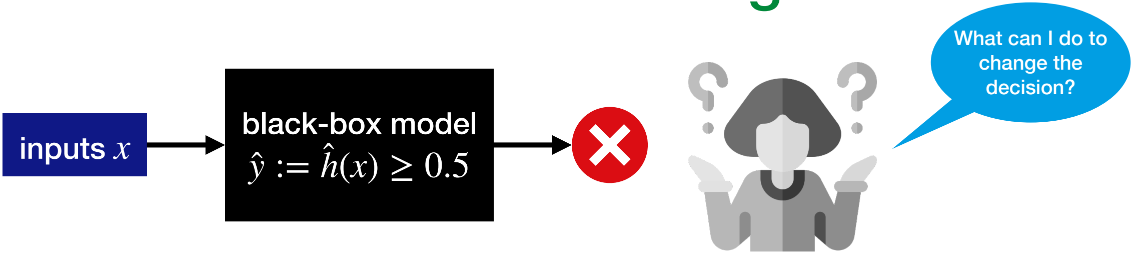
$\hat{y} := \hat{h}(x) \geq 0.5$

What can I do to change the decision?

- suppose a ML model rejects your request (job application, loan application, hospital admission, ….)

- **recourse recommendations** tell you what to do to get accepted

Recourse Setting



inputs x

black-box model

$\hat{y} := \hat{h}(x) \geq 0.5$

What can I do to change the decision?

- suppose a ML model rejects your request (job application, loan application, hospital admission, ….)

- **recourse recommendations** tell you what to do to get accepted

- existing methods: counterfactual explanations (CE) [Wachter et al.], causal recourse (CR) [Karimi et al.]

Problem Terminology

Problem

- existing methods (CE, CR) may suggest to **game** the predictor

Terminology

gaming*:* tricking the predictor into falsely believing that one is qualified

$\hat{y} = 1$ but $y = 0$

Problem

- existing methods (CE, CR) may suggest to **game** the predictor
- recourse should guide towards both **acceptance** *and* **improvement**

Terminology

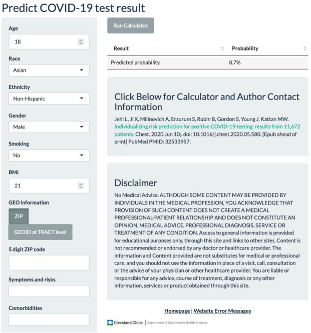
gaming*:* tricking the predictor into falsely believing that one is qualified $\hat{y} = 1$ but $y = 0$

acceptance: reverting the model's decision $\hat{y} = 1$

improvement*:* reverting the underlying real-world state $y = 1$

Illustrative Example

Goal: predict CoViD risk to decide whether you are allowed to enter a hospital (without testing for CoViD)



Predict COVID-19 test result

Age: 18

Race: Asian

Ethnicity: Non-Hispanic

Gender: Male

Smoking: No

BMI: 21

GEO information

5 digit ZIP code

Symptoms and risks

Comorbidities

Result	Probability
Predicted probability	8.7%

Click Below for Calculator and Author Contact Information

Jehi L, Ji X, Milinovich A, Erzurum S, Rubin B, Gordon S, Young J, Kattan MW. Individualizing risk prediction for positive COVID-19 testing: results from 11,672 patients. Chest. 2020 Jun 10;. doi: 10.1016/j.chest.2020.05.580. [Epub ahead of print] PubMed PMID: 32533957.

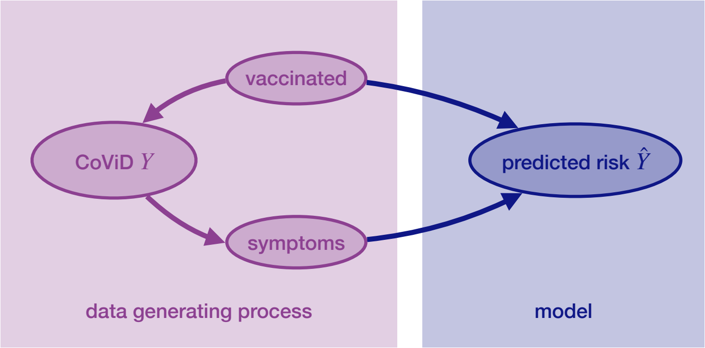
Disclaimer

No Medical Advice. ALTHOUGH SOME CONTENT MAY BE PROVIDED BY INDIVIDUALS IN THE MEDICAL PROFESSION, YOU ACKNOWLEDGE THAT PROVISION OF SUCH CONTENT DOES NOT CREATE A MEDICAL PROFESSIONAL-PATIENT RELATIONSHIP AND DOES NOT CONSTITUTE AN OPINION, MEDICAL ADVICE, PROFESSIONAL DIAGNOSIS, SERVICE OR TREATMENT OF ANY CONDITION. Access to general information is provided for educational purposes only, through this site and links to other sites. Content is not recommended or endorsed by any doctor or healthcare provider. The information and Content provided are not substitutes for medical or professional care, and you should not use the information in place of a visit, call, consultation or the advice of your physician or other healthcare provider. You are liable or responsible for any advice, course of treatment, diagnosis or any other information, services or product obtained through this site.

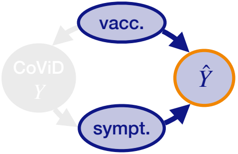
Homepage | Website Error Messages

Cleveland Clinic | Department of Quantitative Health Sciences

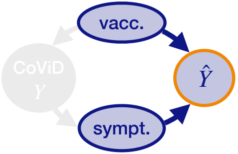
https://riskcalc.org/COVID19/



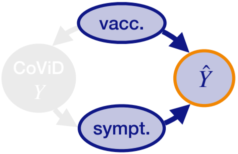
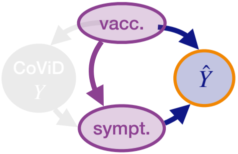
Illustrative Example

	Counterfactual Explanations (CE) [Wachter et al.]
viewpoint	
idea	What is the minimal $do(\underline{X}_I = x')$ such that $\hat{y} = 1$?
exemplary explanation	*“If you plug lower symptom values into the model, the prediction is favorable.”*

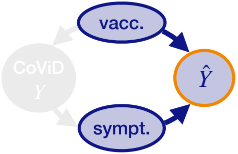
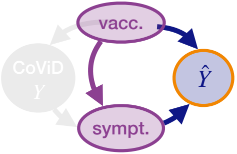
Illustrative Example

	Counterfactual Explanations (CE) [Wachter et al.]
viewpoint	
idea	What is the minimal $do(\underline{X}_I = x')$ such that $\hat{y} = 1$?
exemplary explanation	*"If you plug lower symptom values into the model, the prediction is favorable."*
acceptance $(\hat{y} = 1)$	not guaranteed

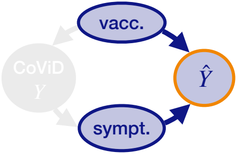
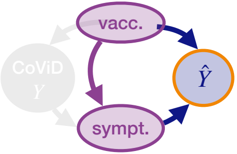
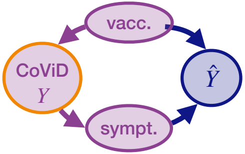
Illustrative Example

	Counterfactual Explanations (CE) [Wachter et al.]	**Causal Recourse (CR) [Karimi et al.]**
viewpoint		
idea	What is the minimal $do(\underline{X}_I = x')$ such that $\hat{y} = 1$?	What is the most cost-efficient $do(X_I = x')$ such that $\hat{y} = 1$?
exemplary explanation	*“If you plug lower symptom values into the model, the prediction is favorable.”*	*“If you treat your symptoms (take cough syrup), the model will accept you.”*
acceptance ($\hat{y} = 1$)	not guaranteed	yes

Illustrative Example

	Counterfactual Explanations (CE) [Wachter et al.]	**Causal Recourse (CR) [Karimi et al.]**
viewpoint		
idea	What is the minimal $do(\underline{X}_I = x')$ such that $\hat{y} = 1$?	What is the most cost-efficient $do(X_I = x')$ such that $\hat{y} = 1$?
exemplary explanation	*“If you plug lower symptom values into the model, the prediction is favorable.”*	*“If you treat your symptoms (take cough syrup), the model will accept you.”*
acceptance $(\hat{y} = 1)$	not guaranteed	yes
improvement $(y = 1)$	not guaranteed	not guaranteed

Illustrative Example

	Counterfactual Explanations (CE) [Wachter et al.]	Causal Recourse (CR) [Karimi et al.]	Improvement-Focused CR (ICR) [Koenig et al.]
viewpoint			
idea	What is the minimal $do(\underline{X}_I = x')$ such that $\hat{y} = 1$?	What is the most cost-efficient $do(X_I = x')$ such that $\hat{y} = 1$?	What is the most cost-efficient $do(X_I = x')$ such that $y = 1$?
exemplary explanation	*"If you plug lower symptom values into the model, the prediction is favorable."*	*"If you treat your symptoms (take cough syrup), the model will accept you."*	*"If you get vaccinated, you will decrease your CoViD risk and thus be accepted."*
acceptance $(\hat{y} = 1)$	not guaranteed	yes	guaranteed
improvement $(y = 1)$	not guaranteed	not guaranteed	yes

The Methods & The 9 Perspectives

	$\hat{Y}$	R	Y
$do(\underline{X} = x')$	CEs		
$X = x'$			
$do(X = x')$	CR		ICR

ICR: Optimization Problem

For a target improvement probability $\bar{\gamma}$, some cost function c and pre-recourse observation x^{pre}, the ICR action $a := do(X_{I_a} := \theta_{I_a})$ is found by optimizing

$$\text{argmin}_a \quad c(a; x^{pre}) \quad s.t. \quad \gamma(a; x^{pre}) \geq \bar{\gamma};$$

with $\gamma(a, x^{pre})$ being the improvement probability for action a and an individual with characteristics x^{pre}.

How to Define Improvement Confidence γ?

Goal: For an action $do(a)$, estimate probability of improvement while taking as many features as possible into account.

How to Define Improvement Confidence γ?

Goal: For an action $do(a)$, estimate probability of improvement while taking as many features as possible into account.

	individualized **(structural causal model known):**
estimation based on	structural counterfactuals
precision	takes all features into account
definition	$\gamma^{ind} := P(Y^{post} = 1 \;\; x^{pre}, do(a))$

How to Define Improvement Confidence γ?

Goal: For an action $do(a)$, estimate probability of improvement while taking as many features as possible into account.

	individualized **(structural causal model known):**	***subpopulation-based*** **(only causal graph known):**
estimation based on	structural counterfactuals	conditional average treatment effect
precision	takes all features into account	only takes features that are not affected by the action into account
definition	$\gamma^{ind} := P(Y^{post} = 1 \;\; x^{pre}, do(a))$	$\gamma^{sub} := P(Y^{post} = 1 \;\; x^{pre}_{G_a}, do(a))$

How to Achieve Acceptance ($\hat{y} = 1$)?

How to Achieve Acceptance ($\hat{y} = 1$)?

Intuition: Classifiers remain accurate under ICR actions → acceptance ensures improvement.

How to Achieve Acceptance ($\hat{y} = 1$)?

Intuition: Classifiers remain accurate under ICR actions → acceptance ensures improvement.

Why?

How to Achieve Acceptance ($\hat{y} = 1$)?

Intuition: Classifiers remain accurate under ICR actions → acceptance ensures improvement.

Why?

- ICR only recommends interventions on causes

How to Achieve Acceptance ($\hat{y} = 1$)?

Intuition: Classifiers remain accurate under ICR actions → acceptance ensures improvement.

Why?

- ICR only recommends interventions on causes
- intervening on causes does not affect $P(Y\ \ X)$

How to Achieve Acceptance $(\hat{y} = 1)$?

Intuition: Classifiers remain accurate under ICR actions → acceptance ensures improvement.

Why?

- ICR only recommends interventions on causes
- intervening on causes does not affect $P(Y \quad X)$

→ the classifier is *stable* w.r.t. ICR actions

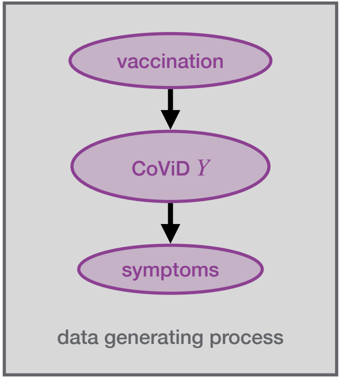
How to Achieve Acceptance ($\hat{y} = 1$)?

Intuition: Classifiers remain accurate under ICR actions → acceptance ensures improvement.

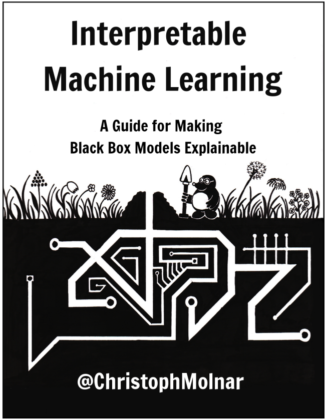
Why?

- ICR only recommends interventions on causes
- intervening on causes does not affect $P(Y \quad X)$

→ the classifier is *stable* w.r.t. ICR actions



Summary: A Causal Perspective on IML



Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

@ChristophMolnar

Pitfalls to Avoid when Interpreting Machine Learning Models

Christoph Molnar [1] Gunnar König [1 2] Julia Herbinger [1] Timo Freiesleben [3 4] Susanne Dandl [1]
Christian A. Scholbeck [1] Giuseppe Casalicchio [1] Moritz Grosse-Wentrup [2 5 6] Bernd Bischl [1]

Abstract

Modern requirements for machine learning (ML) models include both high predictive performance and model interpretability. A growing number of techniques provide model interpretations, but can lead to wrong conclusions if applied incorrectly. We illustrate pitfalls of ML model interpretation such as bad model generalization, dependent features, feature interactions or unjustified causal interpretations. Our paper addresses ML practitioners by raising awareness of pitfalls and pointing out solutions for correct model interpretation, as well as ML researchers by discussing open issues for further research.

1. Introduction

Traditionally, researchers have used parametric models, e.g., linear models, to conduct inference. However, a noticeable shift has happened over the last years towards more non-parametric and non-linear ML models. Models such as random forests, boosting or neural networks often outperform interpretable models on many prediction tasks, as most ML models handle feature interactions and non-linear effects automatically[1] (Fernández-Delgado et al., 2014).
Many disciplines benefit from the predictive performance of ML models and answer scientific questions using ML interpretation techniques. Examples of such efforts include modeling pre-evacuation decision making (Zhao et al., 2020), mapping canopy covers in savannas (Anchang et al., 2020), understanding wildlife diseases (Fountain-Jones et al., 2019), forecasting crop yield (Shahhosseini et al., 2020; Zhang et al., 2019), inferring behavior from smartphone usage (Stachl et al., 2019), and analyzing risk for teacher burnout (Posada-Quintero et al., 2020).
Practitioners are usually interested in the global effect that features have on the outcome and their importance for correct predictions. For certain model classes, e.g., linear models or decision trees, feature effects or importance scores can be inferred from the learned parameters and model structure. In contrast, complex non-linear models that, e.g., do not have intelligible parameters, make it more difficult to extract such knowledge. Therefore, interpretation methods necessarily simplify the relationships between features and the target, e.g., by marginalizing over other features. Prominent techniques for global feature effects include the partial dependence plot (PDP) (Friedman et al., 1991), accumulated local effects (ALE) (Apley & Zhu, 2016) and individual conditional expectation (ICE) (Goldstein et al., 2015). A common feature importance technique is the permutation feature importance (PFI) (Breiman, 2001; Fisher et al., 2019; Casalicchio et al., 2019). This paper will mainly focus on pitfalls of global interpretation techniques when the full functional relationship underlying the data is to be analyzed. Out of scope is the discussion of "local" interpretation methods such as LIME (Ribeiro et al., 2016) or counterfactual explanations (Wachter et al., 2017; Dandl et al., 2020), where individual predictions are to be explained – usually to explain decisions to individuals.
The shift towards ML modeling entails numerous pitfalls for model interpretations. ML models usually contain non-linear effects and higher-order interactions. Therefore, lower-dimensional or linear approximations can be inappropriate and misleading masking effects can occur. As interpretations are based on simplifying assumptions, the associated conclusions are only valid if we have checked that the assumptions underlying our simplifications are not substantially violated. In classical statistics this process is called "model diagnostics" (Fahrmeir et al., 2013) and we believe that a similar process is necessary for interpretable machine learning (IML) based techniques.
Contributions: We review pitfalls of global model-

[1]Department of Statistics, LMU Munich, Munich, Germany [2]Research Group Neuroinformatics, Faculty for Computer Science, University of Vienna [3]Munich Center for Mathematical Philosophy, LMU Munich [4]Graduate School of Systemic Neurosciences, LMU Munich [5]Research Platform Data Science @ Uni Vienna [6]Vienna Cognitive Science Hub. Correspondence to: Christoph Molnar <christoph.molnar@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] While the inclusion of non-linear and interactions effects in classical statistical models is possible, it comes with the increased cost of going more or less manually over many possible modelling options.

Gunnar König

Alex Markham