Scientific Inference with Interpretable Machine Learning

Analyzing Models to Learn About Real-World Phenomena

Timo Freiesleben

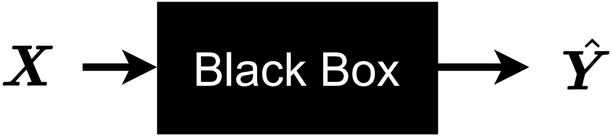
Joint work with G. König, C. Molnar, & A. Tejero-Cantero

University of Tübingen, Cluster of Excellence 'Machine Learning for Science'

Outline

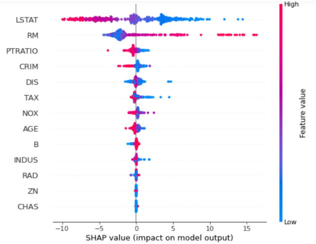
Motivation: Interpretable ML

- Also called XAI
- Ingredients:
  - Data $\boldsymbol{D}$
  - inputs $\boldsymbol{X}$ and prediction $\hat{\boldsymbol{Y}}$
  - Trained ML model $\hat{\boldsymbol{m}}$

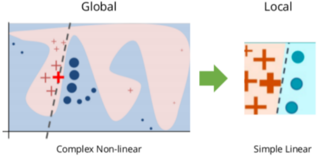


Motivation: Method zoo

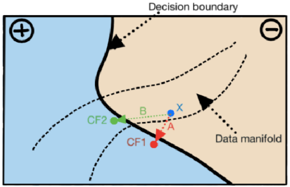
Shapley Values



LIME



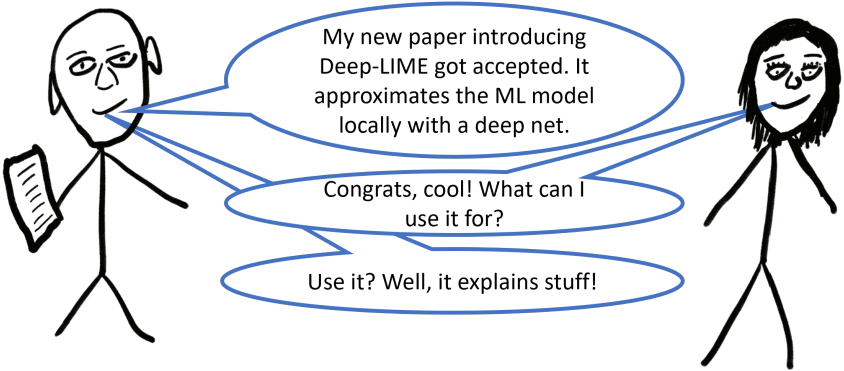
Counterfactual explanations



Saliency maps

Images by: Idit Cohen, Mokuwe et al. [2020], Ribeiro et al. [2016], Verma et al. [2020]

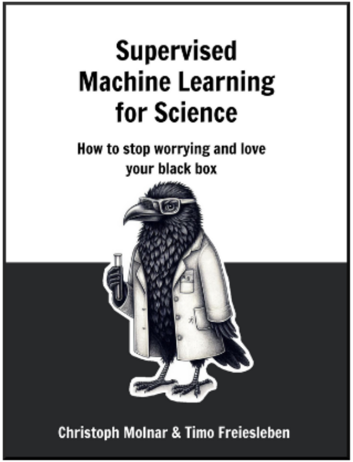
Motivation: What real problems are solved?



Dear XAI community, we need to talk! [Freiesleben and König, 2023]

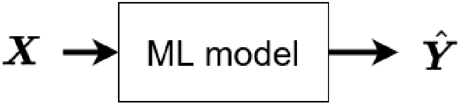
Motivation: Can we use (interpretable) ML for science?

https://ml-science-book.com/

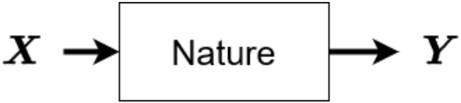


Motivation: Model audit vs scientific inference

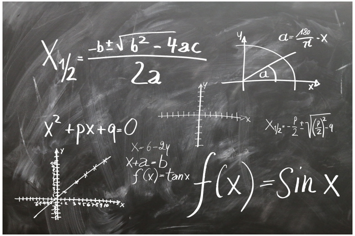
- Model audit



- Scientific inference



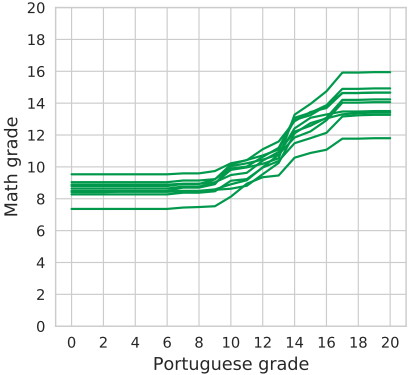
Motivation: Laura example



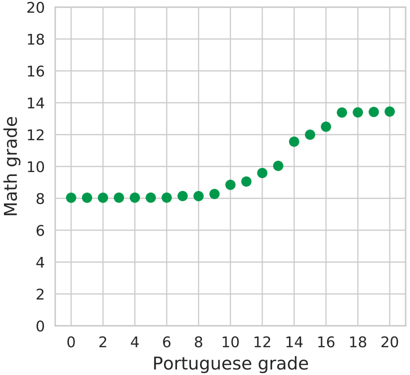
- How relate language and math skills?
- Data [Cortez and Silva, 2008]:
  - students grades,
  - parent's jobs/education,
  - age, tutoring, absences, etc.

Motivation: Partial dependence plot

- Timo= (Port grade: 0; tutoring: *no*; absences: 5; . . .)



ICE curves



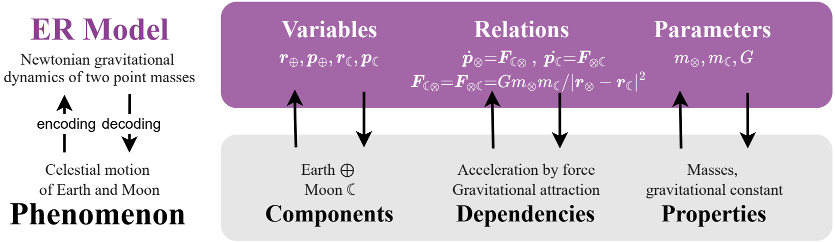
Partial Dependence Plot (PDP)

Where Are We?

Traditional scientific inference: Elementwise representation

Definition: Elementwise Representationality

A model is *elementwise representational* (ER) if all model elements represent an element in the phenomenon.

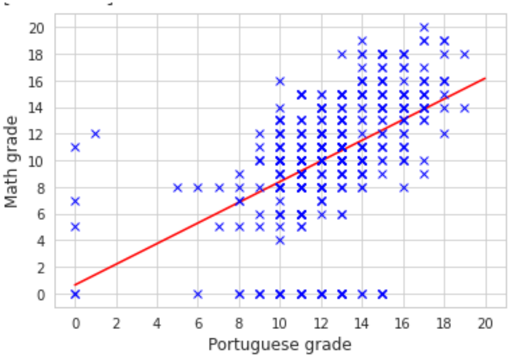


Traditional scientific inference: Why ER?

- ER is cognitively appealing
- ER eases model construction
- **ER allows inference from model to world**

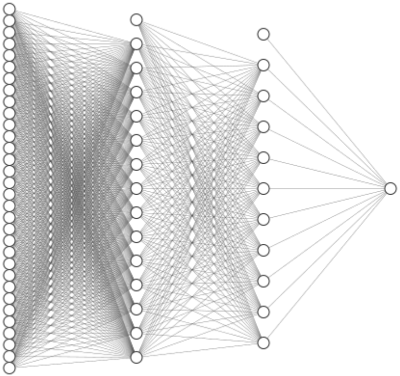
Traditional scientific inference: Example

- $\boldsymbol{Y}$=math and $\boldsymbol{X}_p$=Portuguese
- $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_p + \epsilon$
- Least squares: $\hat{m}_{LIN}(x_p) = 10.46 + 0.77x_p$
- Confidence Intervals $\mathrm{CI}_{\hat{\beta}_0} = [10.05; 10.88]$ and $\mathrm{CI}_{\hat{\beta}_1} = [0.63; 0.91]$



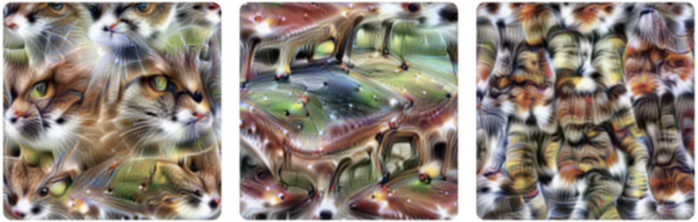
Traditional scientific inference: ML models not ER

- ML models are less assumption laden
- Most model elements (weights, activation functions, etc) have no meaning



Traditional scientific inference: Inference with ML

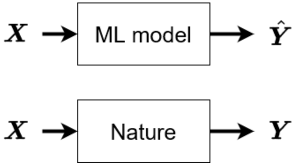
- Option 1: [Bokulich, 2011]
  - Scientific inference without ER is impossible
- Option 2: [Olah et al., 2020]
  - ML models are ER too



- Option 3: [Cichy and Kaiser, 2019]
  - IML for scientific inference

Traditional scientific inference: IML for inference

- **Problems**
  - Current IML focuses on model audit
  - Not every audit allows for inference
  - Audit and inference are complementary goals



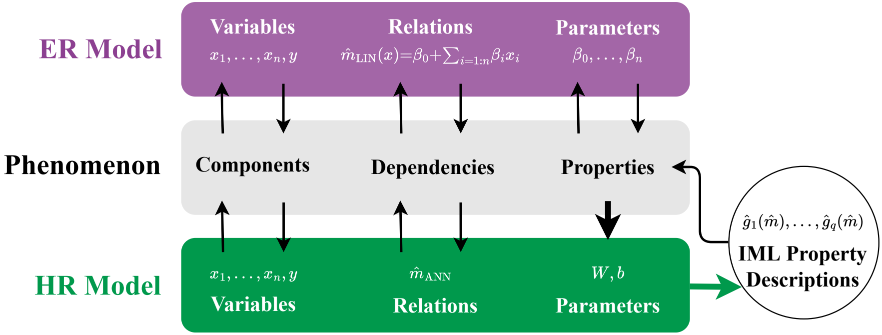
Where Are We?

1. Motivation
2. Traditional scientific inference
3. Theory of property descriptors
4. Discussion

Theory of property descriptors: Hollistic representation

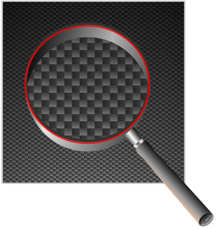
Definition: Holistic representationality

A model is *holistically representational* (HR) if the whole model represents aspects of the phenomenon.

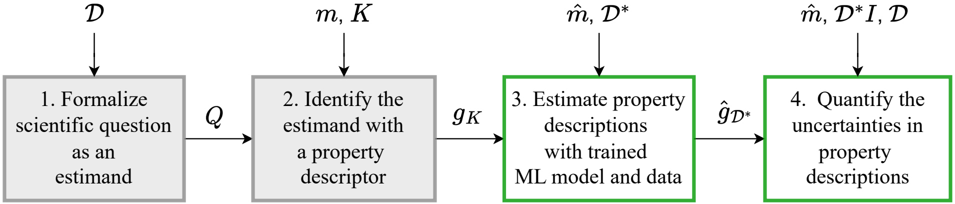


Theory of property descriptors: What ML models represent?

Problem	Loss	$L(Y, \hat{m}(\boldsymbol{X}))$	Optimal predictor[a] m
Regression (Y continuous)	mean squared error	$(Y - \hat{m}(\boldsymbol{X}))^2$	$\mathbb{E}_{Y\mid\boldsymbol{X}}[Y \mid \boldsymbol{X}]$
	mean absolute error	$\lvert Y - \hat{m}(\boldsymbol{X})\rvert$	$\text{median}(Y \mid \boldsymbol{X})$
Classification (Y discrete)	0-1 loss	0 if $\hat{m}(\boldsymbol{X}) = Y$, else 1	$\arg\max_{y\in\mathcal{Y}} \mathbb{P}(Y{=}y \mid \boldsymbol{X})$
	cross entropy	$\sum_{r\in\mathcal{Y}} \mathbb{P}_Y(r) \log \mathbb{P}_{\hat{m}(\boldsymbol{X})}(r)$	$\mathbb{P}(Y \mid \boldsymbol{X})$



Theory of property descriptors: Four steps



Theory of property descriptors: 1. Formalize scientific question

- Scientists start by asking and formalizing questions.
- Question: How are language skills associated with math skills?
- Formalized Question: $\boldsymbol{Q} = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}_p}[\boldsymbol{Y} \mid \boldsymbol{X}_p]$

Theory of property descriptors: 2. Identify estimand

Definition: Question Identifiability

We say that a question is *identifiable* relative to probabilistic knowledge K if we can compute Q from m and K.

- Laura's question can be identified with $K = \mathbb{P}(\boldsymbol{X}_{-p} \mid \boldsymbol{X}_p)$

$$
\begin{aligned}
Q &:= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}_p}[\boldsymbol{Y} \mid \boldsymbol{X}_p] \\
&= \mathbb{E}_{\boldsymbol{X}_{-p}|\boldsymbol{X}_p}[\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[\boldsymbol{Y} \mid \boldsymbol{X}] \mid \boldsymbol{X}_p] \qquad \text{(by the \textit{tower rule})} \\
&= \mathbb{E}_{\boldsymbol{X}_{-p}|\boldsymbol{X}_p}[m(\boldsymbol{X}) \mid \boldsymbol{X}_p].
\end{aligned}
$$

Theory of property descriptors: 2. Identify estimand

Definition: Property Descriptor

A *property descriptor* is a continuous function g_K that identifies Q given K

$$g_K : \mathcal{M} \to \mathcal{Q} \quad \text{with} \quad g_K(m) = Q.$$

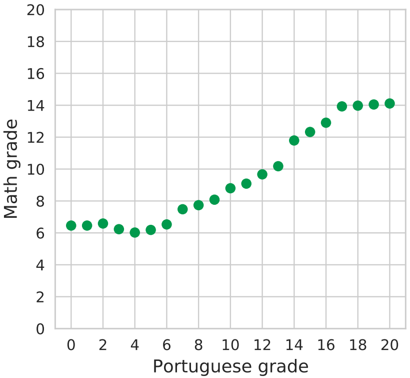
- In our example, this is the conditonal Partial Dependence Plot (cPDP): $g_K(\hat{m}) := \mathbb{E}_{\boldsymbol{X}_{-p}|\boldsymbol{X}_p}[\hat{m}(\boldsymbol{X}) \mid \boldsymbol{X}_p]$

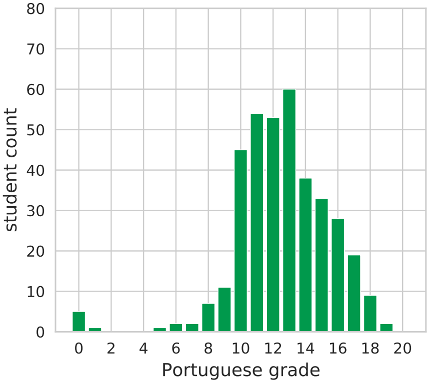
Theory of property descriptors: 3. Estimate property

In real life, we have limited access to $\boldsymbol{X}, \boldsymbol{Y}$. We have finite data.

Definition: Property Description Estimator

The *property description estimator* $\hat{g}_{\mathcal{D}^*}$ is an unbiased estimator of g_K.

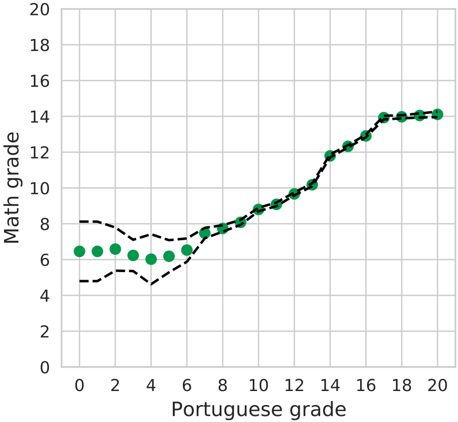


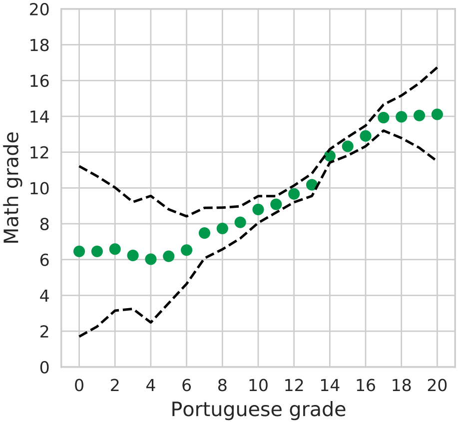


Theory of property descriptors: 4. Uncertainty quantification

We make two errors on the way:

1. we do not have the optimal model (model error), and
2. we only have finite data (estimation error).





Theory of property descriptors: Practical descriptors

	Global / local question	Estimand	IML method
Conditional contribution	How much worse can Y be predicted from $\boldsymbol{X}$ if we did not know X_p?	$\mathrm{EPE}_{\boldsymbol{X},Y}\, m_{\boldsymbol{X}}(\boldsymbol{X}) - \mathrm{EPE}_{\boldsymbol{X}_{-p},Y}\, m_{\boldsymbol{X}_{-p}}(\boldsymbol{X}_{-p})$	cFI Strobl et al. (2008)
	How much worse can Y be predicted from $\boldsymbol{X} = \boldsymbol{x}$ if we did not know X_p?	$L(y, m_{\boldsymbol{X}}(\boldsymbol{x})) - L(y, m_{\boldsymbol{X}_{-p}}(\boldsymbol{x}_{-p}))$	ICI Casalicchio et al. (2019)
Fair contribution	What is the fair share of feature X_p in the prediction of Y?	$\frac{1}{n}\sum_{S\subseteq N\setminus\{p\}} \binom{n-1}{\lvert S\rvert}^{-1}\Big(\mathrm{EPE}_{\boldsymbol{X}_{S\cup\{p\}},Y}\, m_{\boldsymbol{X}_{S\cup\{p\}}}(\boldsymbol{X}_{S\cup\{p\}}) - \mathrm{EPE}_{\boldsymbol{X}_S,Y}\, m_{\boldsymbol{X}_S}(\boldsymbol{X}_S)\Big)$	SAGE Covert et al. (2020)
	What is the fair share of feature X_p in the prediction of Y if $X = \boldsymbol{x}$?	$\frac{1}{n}\sum_{S\subseteq N\setminus\{p\}} \binom{n-1}{\lvert S\rvert}^{-1}\left(m_{\boldsymbol{X}_{S\cup\{p\}}}(\boldsymbol{x}_S, x_p) - m_{\boldsymbol{X}_S}(\boldsymbol{x}_S)\right)$	Conditional Shapley values Aas et al. (2021)

- Which student information should educators track? $\Rightarrow$ ICI, cFI, conditional SHAP & SAGE

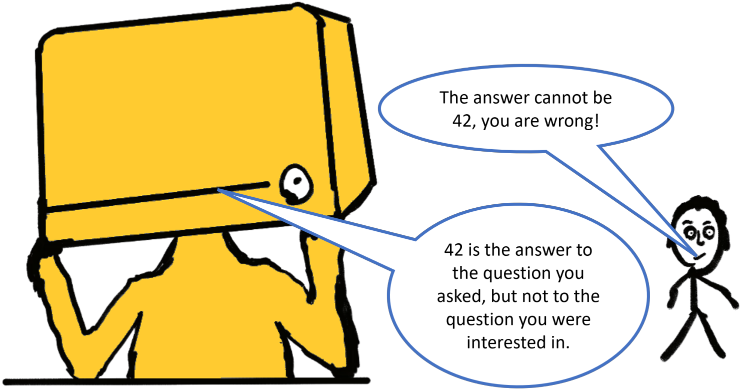
Theory of property descriptors: Practical descriptors

	Global / local question	Estimand	IML method
Effect	What is the best estimate of Y if we only know X_p?	$m_{X_p}(X_p)$	cPDP Apley and Zhu (2020)
	How does the best estimate of Y change relative to X_p, knowing that $\boldsymbol{X}_{-p} = \boldsymbol{x}_{-p}$?	$m_{\boldsymbol{X}}(X_p, \boldsymbol{x}_{-p})$	ICE curve Goldstein et al. (2015)
Relevant value	Under which realistic conditions $\boldsymbol{X}$ can we observe relevant value y_{rel}?	$\arg\min_{\boldsymbol{x} \in \text{supp}\, \boldsymbol{X}} d_{\mathcal{Y}}(m_{\boldsymbol{X}}(\boldsymbol{x}), y_{\text{rel}})$	PRIM[a] Friedman and Fisher (1999)
	Under which realistic conditions similar to $\boldsymbol{x}$ can we observe relevant value y_{rel}?	$\arg\min_{\boldsymbol{x}' \in \text{supp}\, \boldsymbol{X}} d_{\mathcal{Y}}(m_{\boldsymbol{X}}(\boldsymbol{x}'), y_{\text{rel}}) + \lambda\, d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')$	Counterfactu-als [b] Dandl et al. (2020)

- How influences parents' education students math skills? $\Rightarrow$ ICE & cPDP
- What characterizes (more/less) successful students? $\Rightarrow$ counterfactuals & PRIM

Theory of property descriptors: Disagreement

- Methods can only meaningfully disagree if they have different estimands.
- The disagreement problem stems from a lack of clarity about the question asked.



Where Are We?

Discussion: Causality

- Most scientific questions are causal:
  - Would tutoring in Portuguese improve students math skills? (**Interventional**)
  - Did the student fail in math because of her Portuguese skills? (**Counterfactual**)
- Property descriptors describe associational quantities.
- Causal questions add another layer to the pipeline and require causal knowledge.

Discussion: Limitations

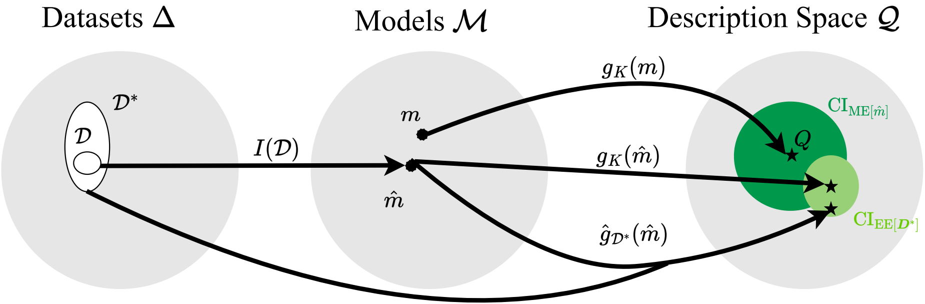
- Direct estimation often better! (e.g. targeted learning [Van der Laan and Rose, 2011])
- Conditional sampling is needed but hard!
- Formalizing questions on images and sound?

Conclusion

- **Problem:** Scientific inference via model elements is not available. Current IML mixes different desiderata.
- **Our Solution:** Smart interrogation with property descriptors allows to learn about the process.

Questions

Graph for Formal Depiction



A. Bokulich. How scientific models can explain. *Synthese*, 180(1):33–45, 2011.

R. M. Cichy and D. Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23 (4):305–317, 2019.

P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. 2008.

T. Freiesleben and G. König. Dear xai community, we need to talk! fundamental misconceptions in current xai research. In *World Conference on Explainable Artificial Intelligence*, pages 48–65. Springer, 2023.

M. Mokuwe, M. Burke, and A. S. Bosman. Black-box saliency map generation using bayesian optimisation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

M. J. Van der Laan and S. Rose. *Targeted learning*. Springer, 2011.

S. Verma, J. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2, 2020.