

Session 2

Challenges for XAI (and some XAI methods)

Gunnar König
October 21 2026

Why XAI?

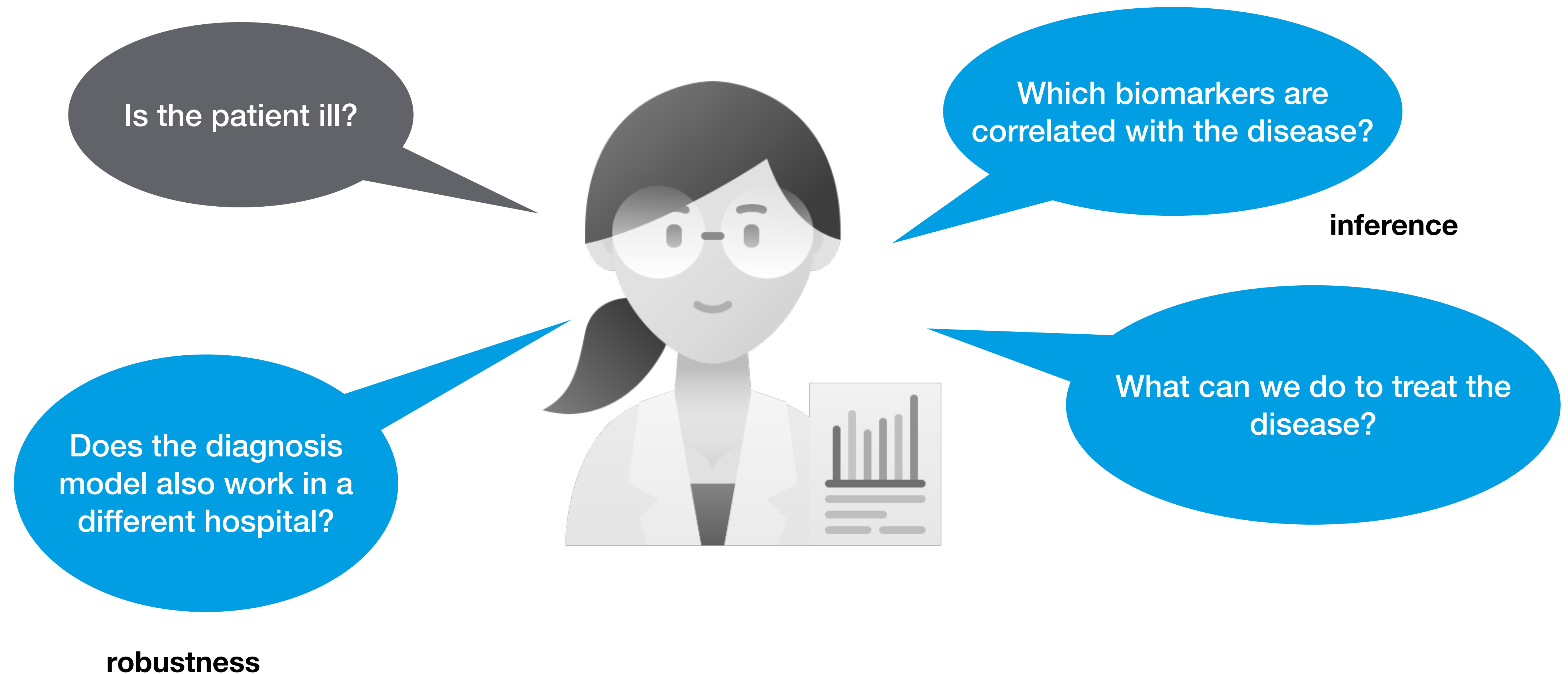
Q: What are concrete questions that one might like to answer with XAI?

Common Narrative

- ML excels at *accurate prediction* within a given distribution.
- Often we are not (only) interested in accurate prediction.

Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Example: Scientist



Example: Individual Affected by ML Decision



Example: LLM researcher

Does the model deceive us /
pursue misaligned goals?

safety



Can I steer the model to
behave differently?

control

Common Narrative

- ML excels at *accurate prediction* within a given distribution.
 - Often we also/actually want **robustness, contestability, inference, recourse, safety, control ...**
 - Hope: If we “understood” the model, we could assess/fulfill the aforementioned requirements.
- ⇒ We need “interpretability” / “explainability”.

*Zachary Lipton, “Interpretability: of what, for whom, why, and how?”
Machine Learning for Physics and the Physics of Learning (2019).*

Example1: Permutation Feature Importance (PFI)

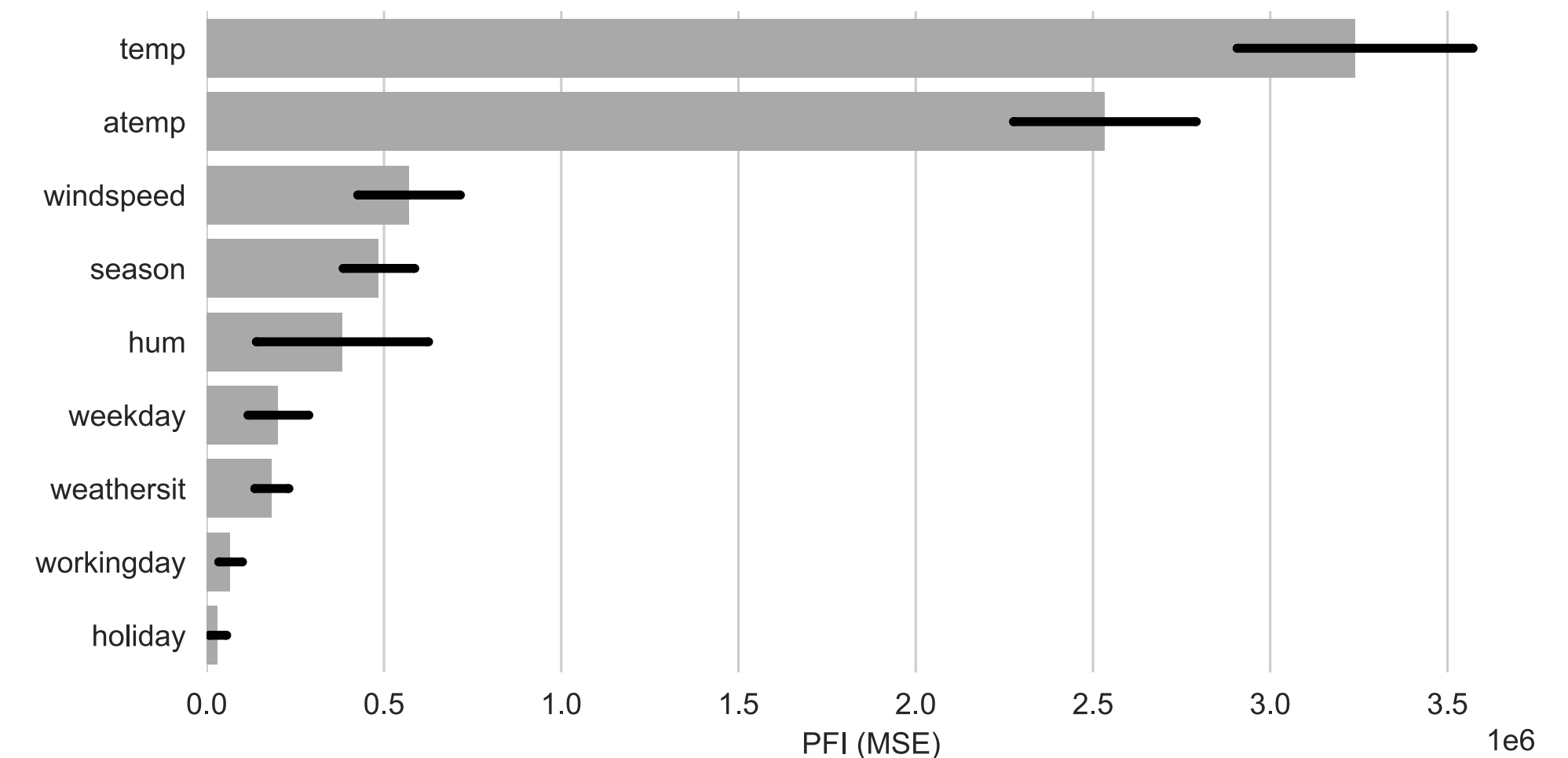
Model-agnostic, global, feature importance method

Idea: Gain understanding by assessing how destroying a feature affects the model's performance.

Formally: The difference in risk before and after perturbing X_j .

$$PFI_j := R(\hat{f}(\tilde{X}_j, X_{-j}), Y) - R(\hat{f}(X), Y)$$

with $\tilde{X}_j \sim P(X_j)$ and $\tilde{X}_j \perp X_{-j}, Y$.



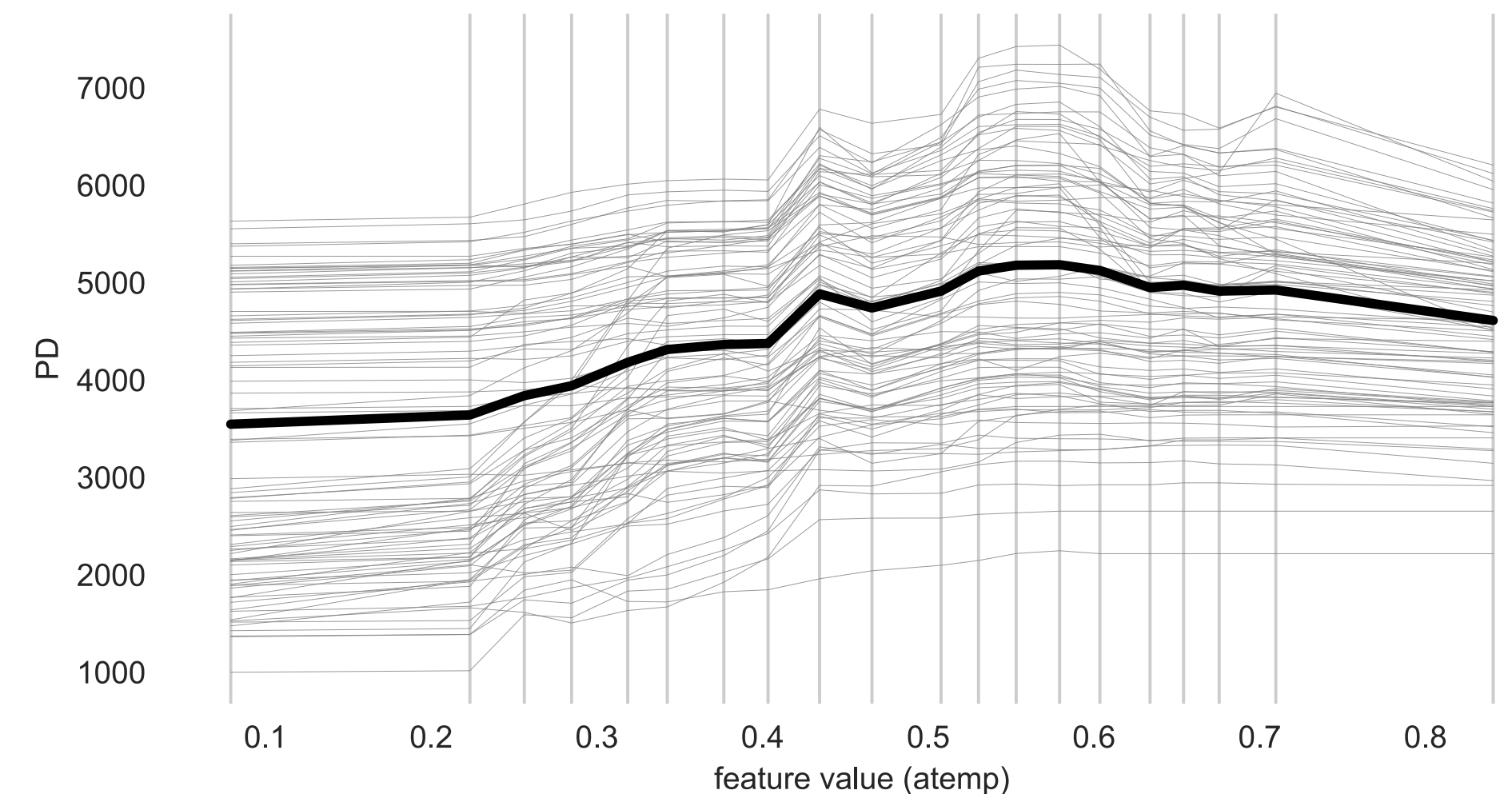
Example 2: Partial Dependence Plots (PDP)

Model-agnostic, global, feature effect method

Idea: Understand model by assessing how changing a feature affects the prediction.

Formally: Marginalize out all features but the feature of interest x_j .

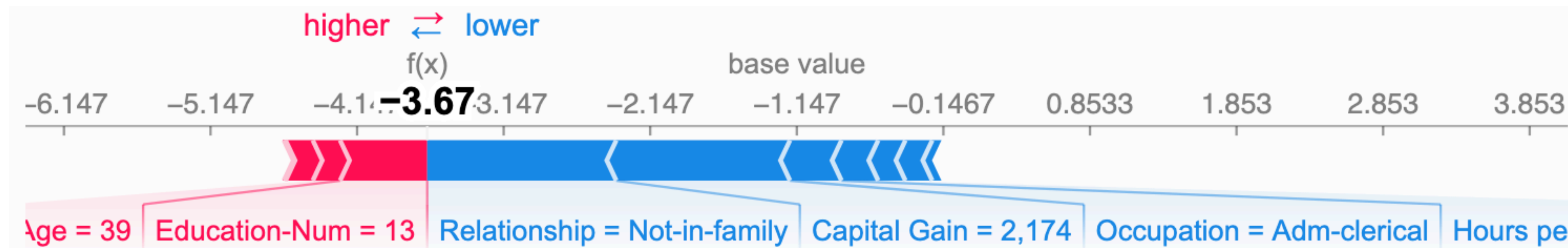
$$\hat{f}_j(x'_j) = E_{X_{-j}}[\hat{f}(x'_j, X_{-j})]$$



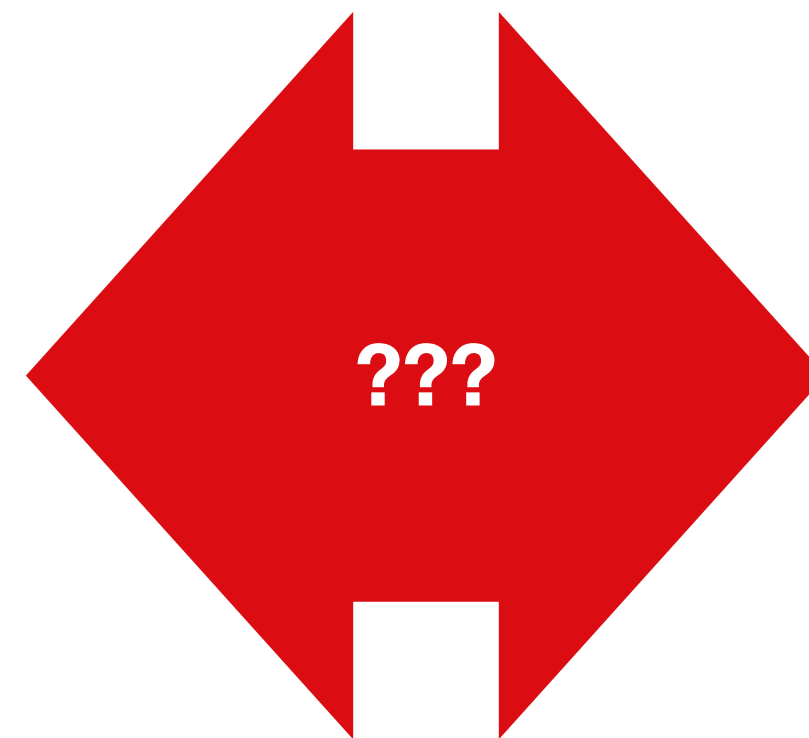
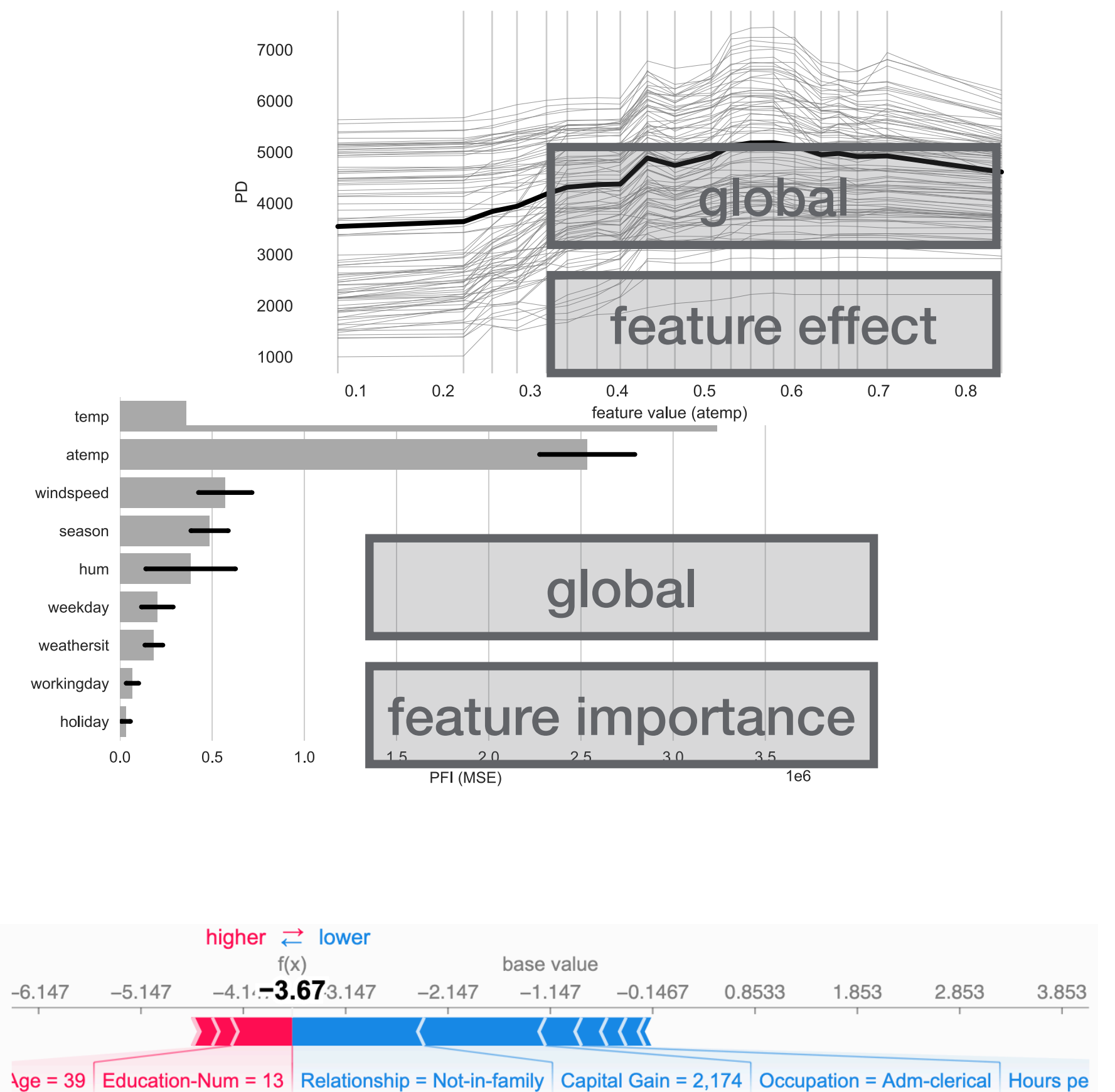
Example 3: SHAP

Model-agnostic, local, feature effect method

Idea: Fairly attribute the prediction to the individual features using game-theoretic Shapley values.



Challenges for XAI



Challenges for XAI

Challenge I: Unclear Goal



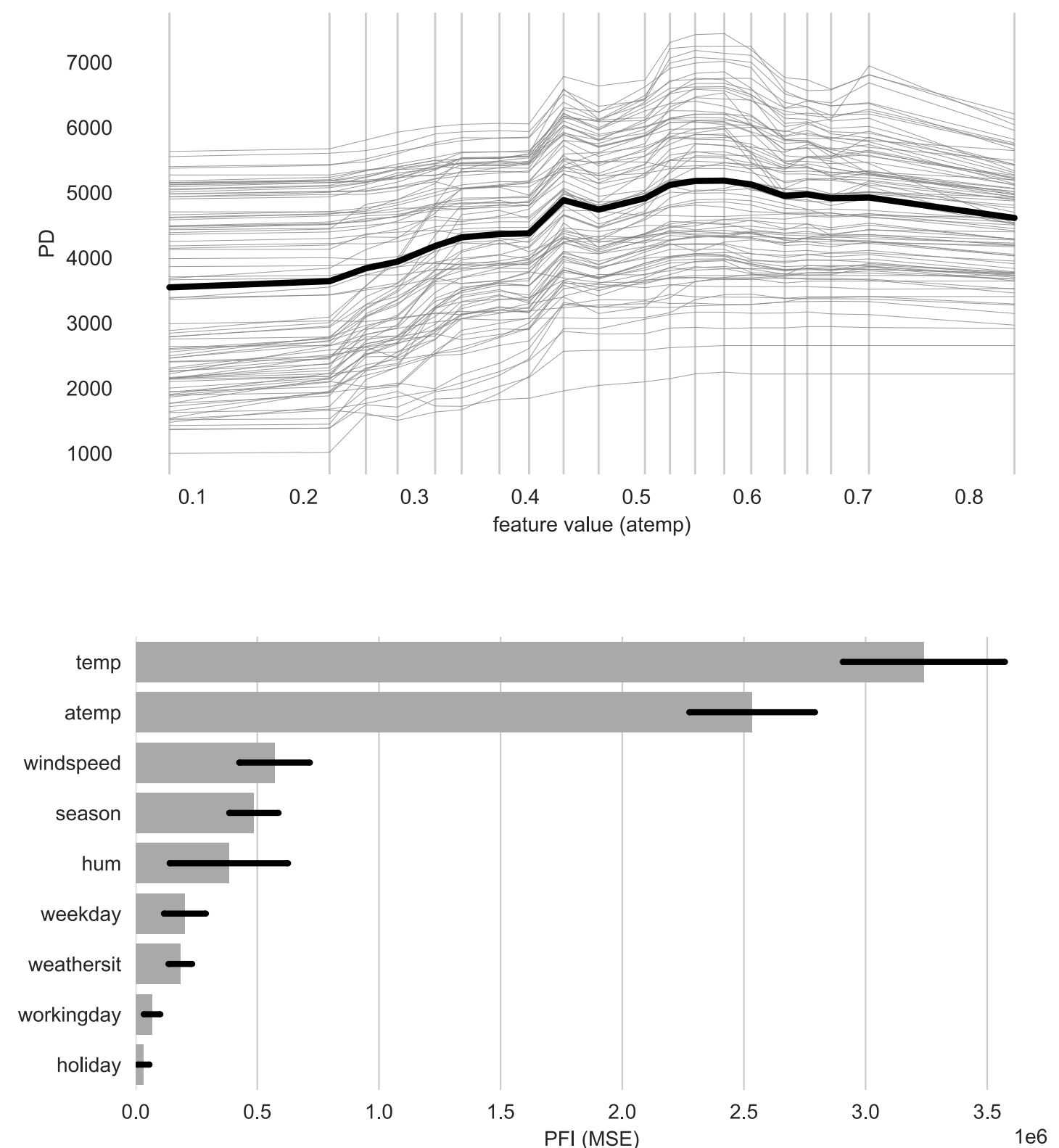
- “Explainable” is not well-defined and therefore does not serve as a goal
- reason: different incompatible subgoals are conflated
- Subgoals must be disentangled

Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16.3 (2018): 31-57.

Challenges for XAI

Challenge II: Misinterpretation

- Unclear about what insight methods provide (practitioners rely on superficial criteria instead, Krishna, 2022)
 - missing interpretation rules
 - Practitioners not aware of them
- Methods should be intelligible for their target audience!



Krishna, Satyapriya, et al. "The disagreement problem in explainable machine learning: A practitioner's perspective." *arXiv preprint arXiv:2202.01602* (2022).

Anchors

Illustrative Example

Why didn't I get the loan?



savings: 200€

age: 37

salary: 1000€

Anchor: *“You are between 30 and 40 years old and have between 10 and 500€ in savings, and any such applicant is rejected.”*

Local rule that anchors the decision:
“Everyone” in the scope get the same decision as explanation receiver x .

Anchors in Image / Text



original image



anchor for “beagle”

Where is the **dog**?
What **color** is the **wall**?
When was this picture taken?
Why is he lifting his paw?

input (anchor highlighted bold)

on the floor
white
during the day
to play

prediction

Scope

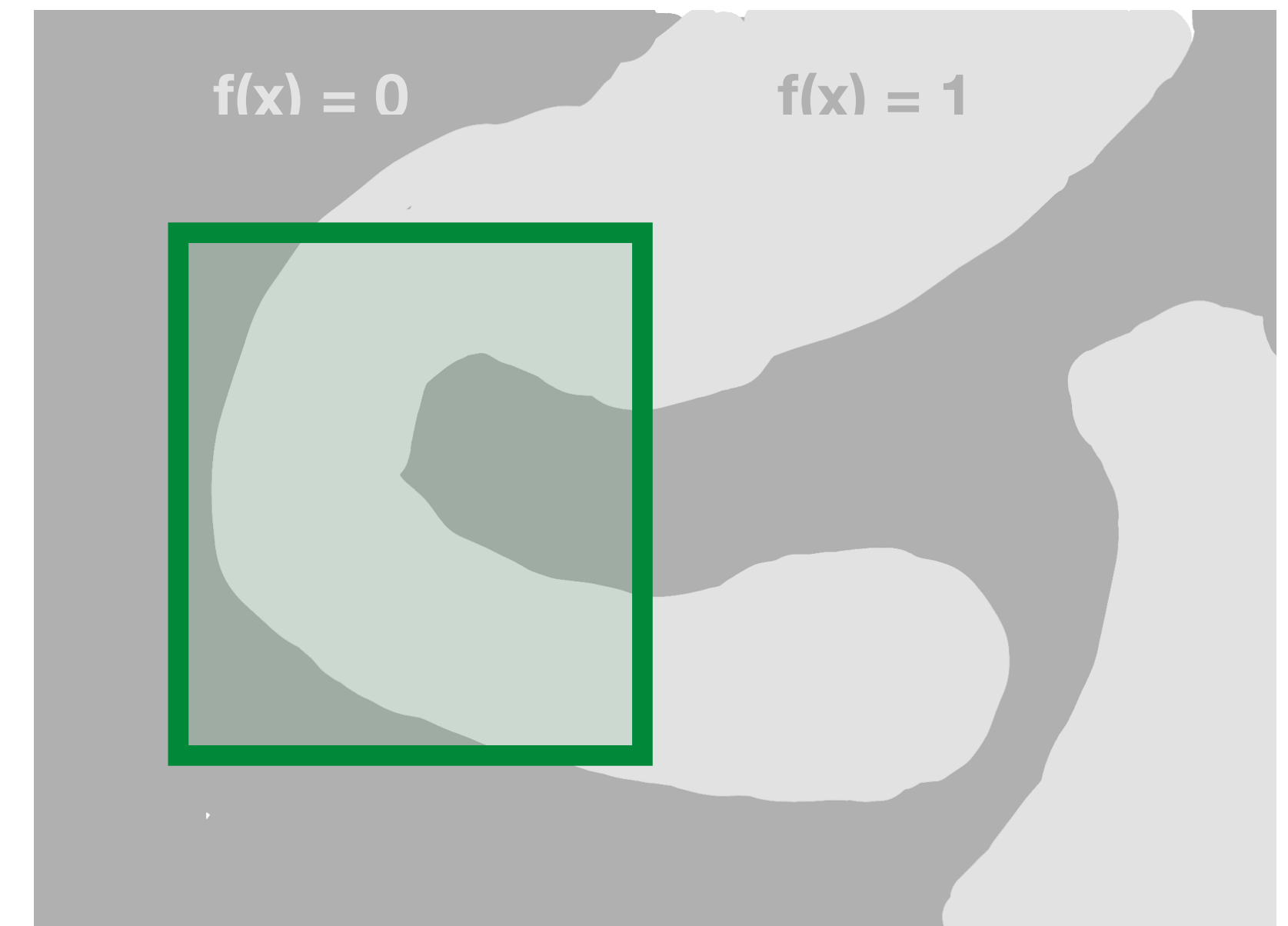
Binary mapping $A : \mathcal{X} \mapsto \{0,1\}$, data points z with $A(z) = 1$ are in the scope.

To ensure *intelligibility*, anchors construct the scope as *conjunctions of feature predicates* $a_j : \mathcal{X}_j \mapsto \{0,1\}$.

$$A(z) := a_k(x_k) \wedge a_l(x_l) \wedge \dots$$

Example:

$$A(\text{savings}, \text{age}) := \mathbf{1}_{10 \leq \text{savings} \leq 500} \wedge \mathbf{1}_{30 \leq \text{age} \leq 40}$$



Feature space $\mathcal{X}$

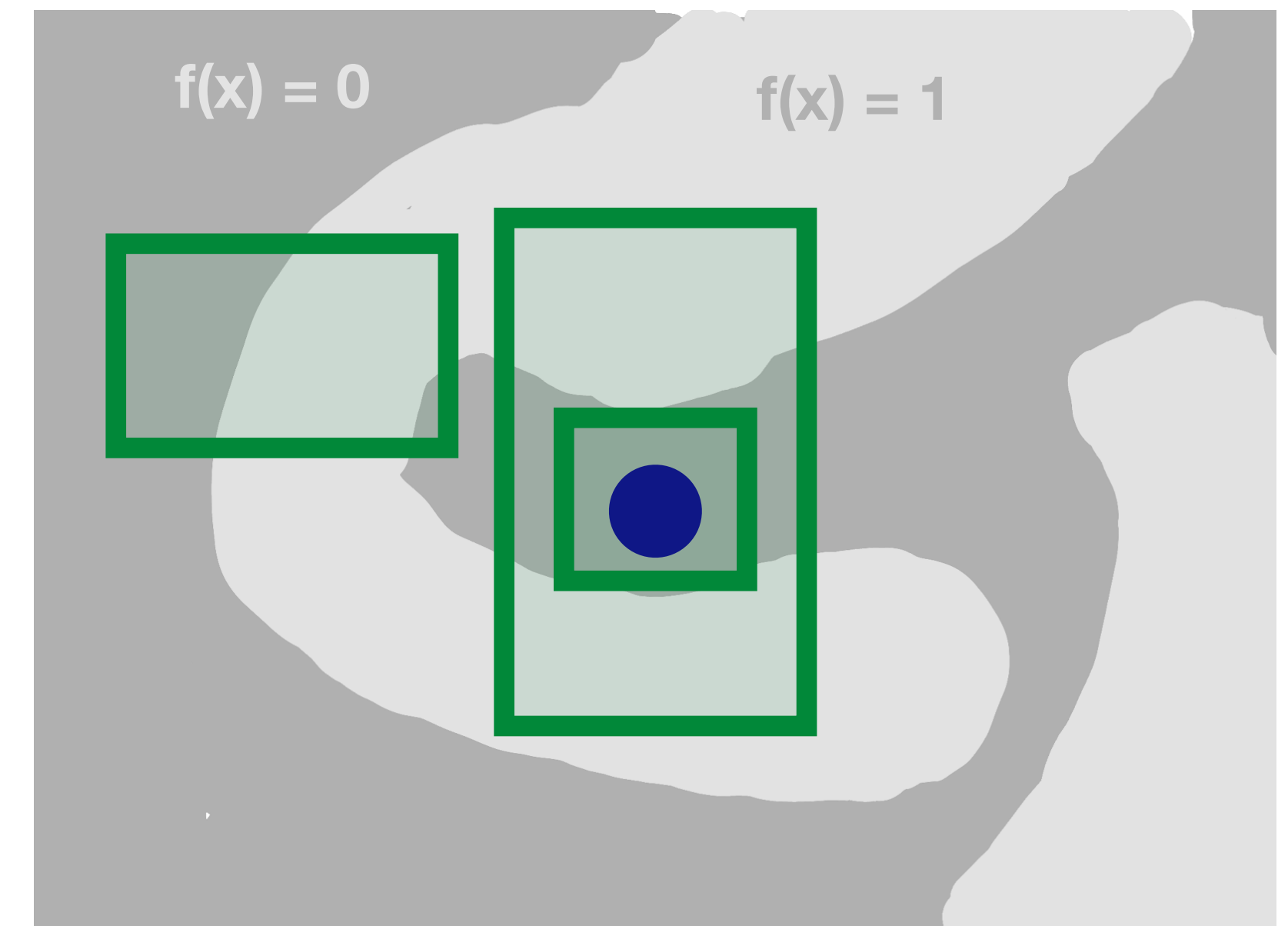
Anchor

Intuition: “Everyone” in the **scope** gets the same decision as explanation receiver x .

Let f be the predictor, x the explained observation. Let $\tau \in (0, 1]$. Then A is an **anchor** for x and f if:

1. $A(x) = 1$.
2. A is precise, formally $\text{prec}(A) \geq \tau$, where

$$\text{prec}(A) := \mathbb{P}_{Z=z|A=1} (f(z) = f(x) \mid A = 1).$$



Coverage

Let $A(z) := \mathbf{1}_{z=x}$. Then A is an anchor:

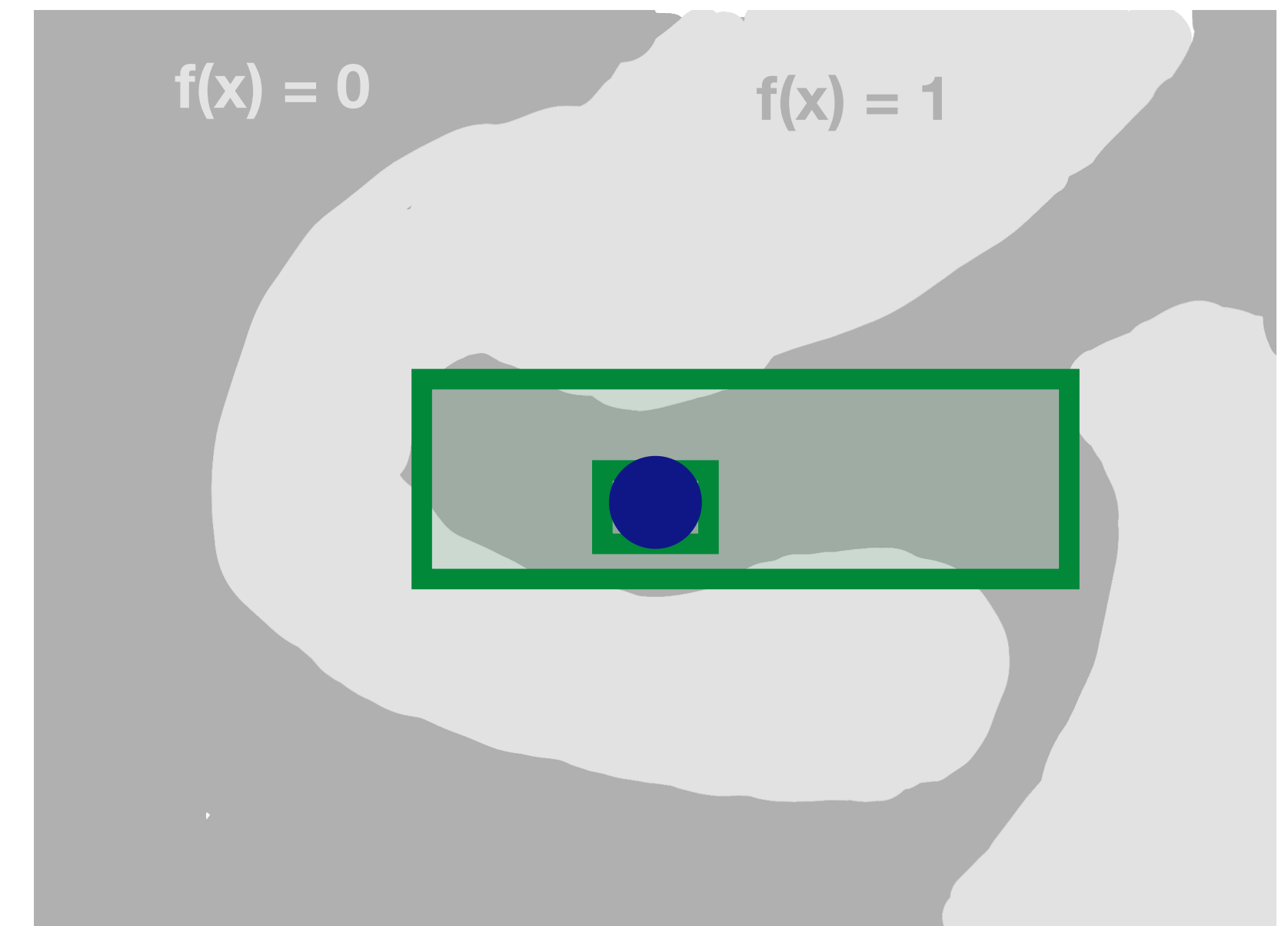
✓ $A(x) = 1$

✓ $\text{prec}(A) = 1$

But: The anchor tells us nothing new!

→ Anchors should be general:

$$\text{coverage}(A) = \mathbb{P}_X (A(x))$$



Optimization Problem

For explanation receiver x and precision level τ we optimize

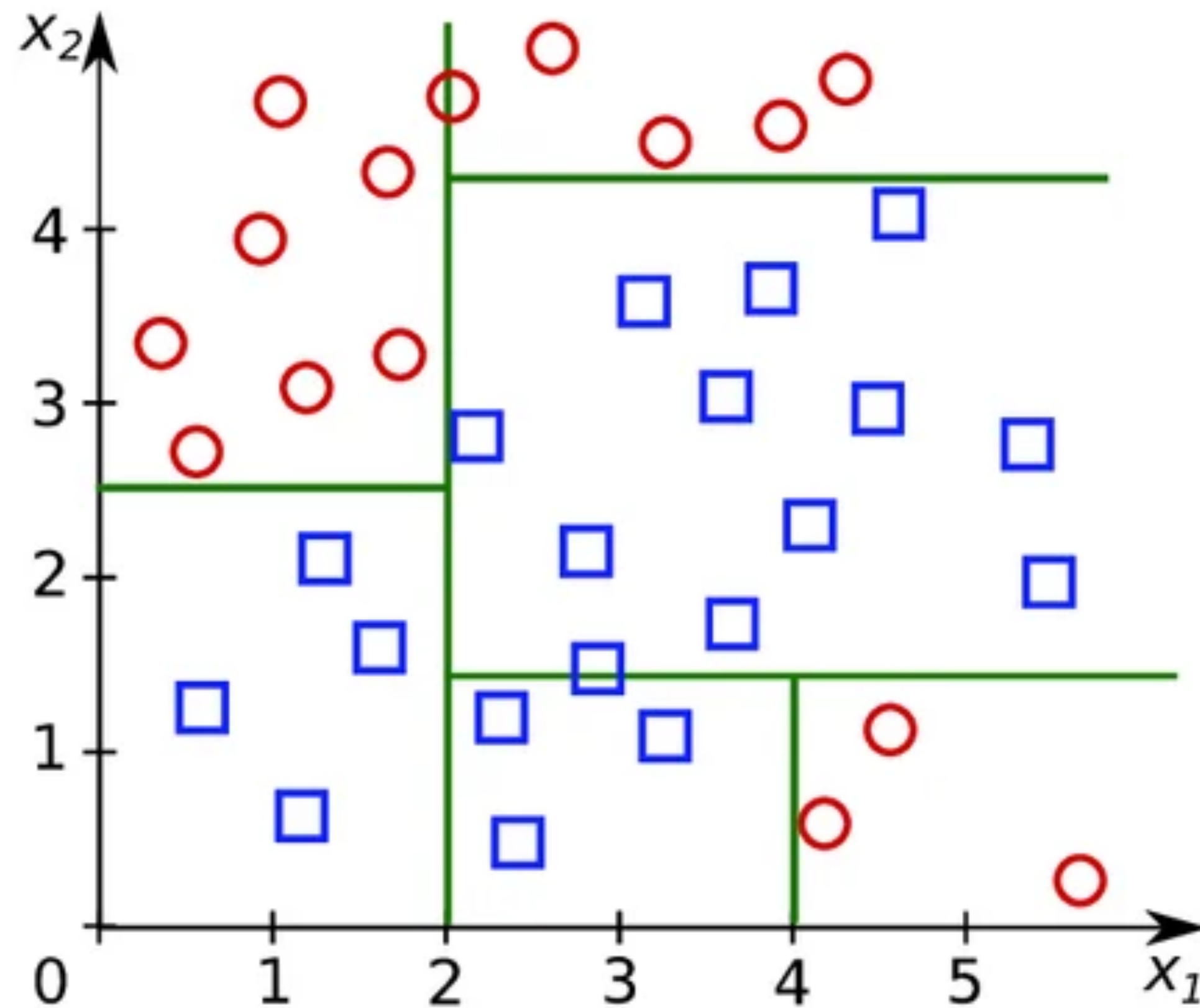
$$\max_{A \in \mathcal{A}} \text{coverage}(A) \quad \text{s.t.} \quad \text{prec}(A) \geq \tau \text{ and } A(x) = 1,$$

where $\mathcal{A}$ is the space of scopes.

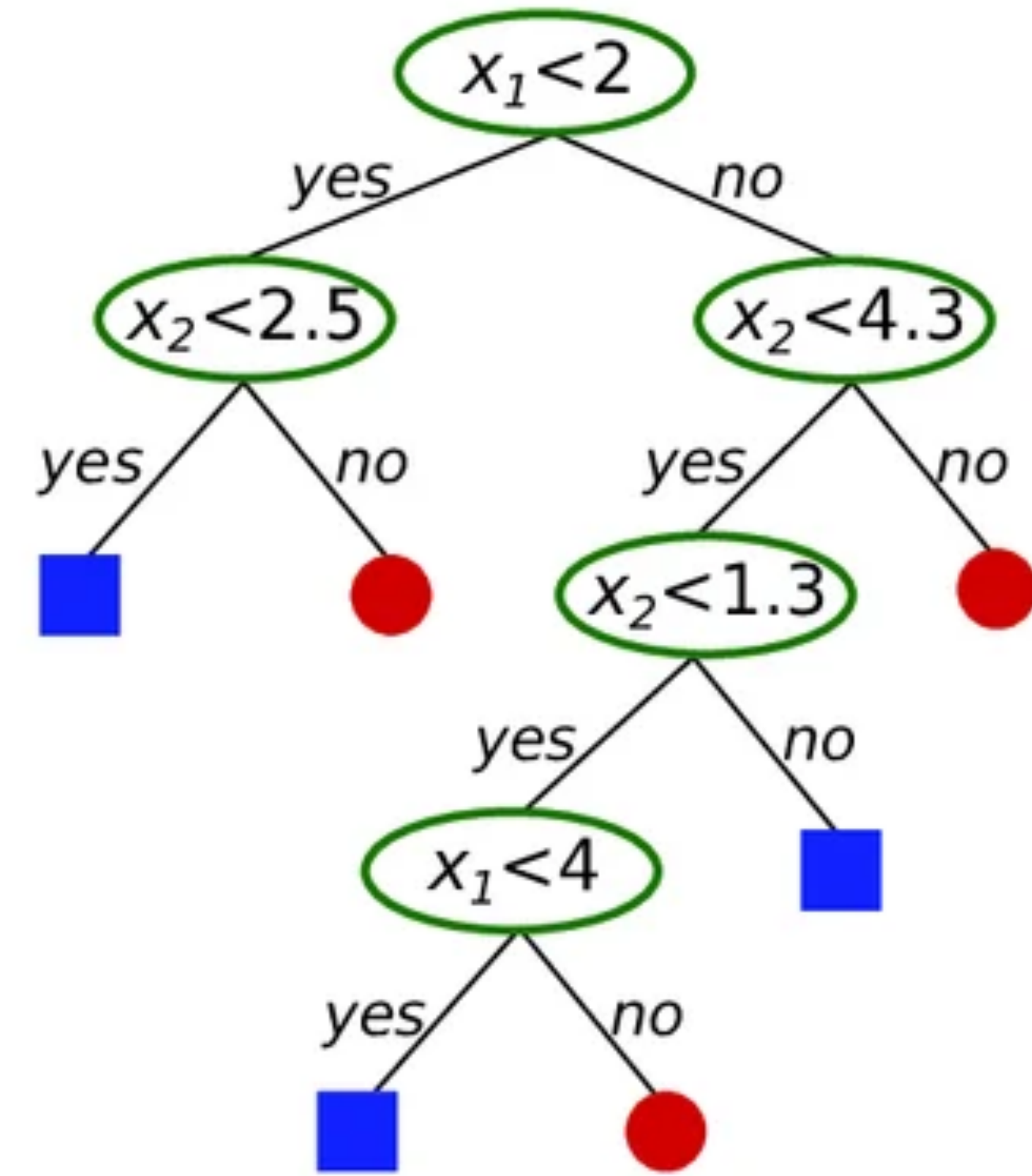
- ⚠ Large search space (combinatorial explosion), non-convex.
- ⚠ Exact evaluation of precision not feasible.
- ⚠ Approximate evaluation of precision is expensive.

Anchors and Decision Trees

A) feature space



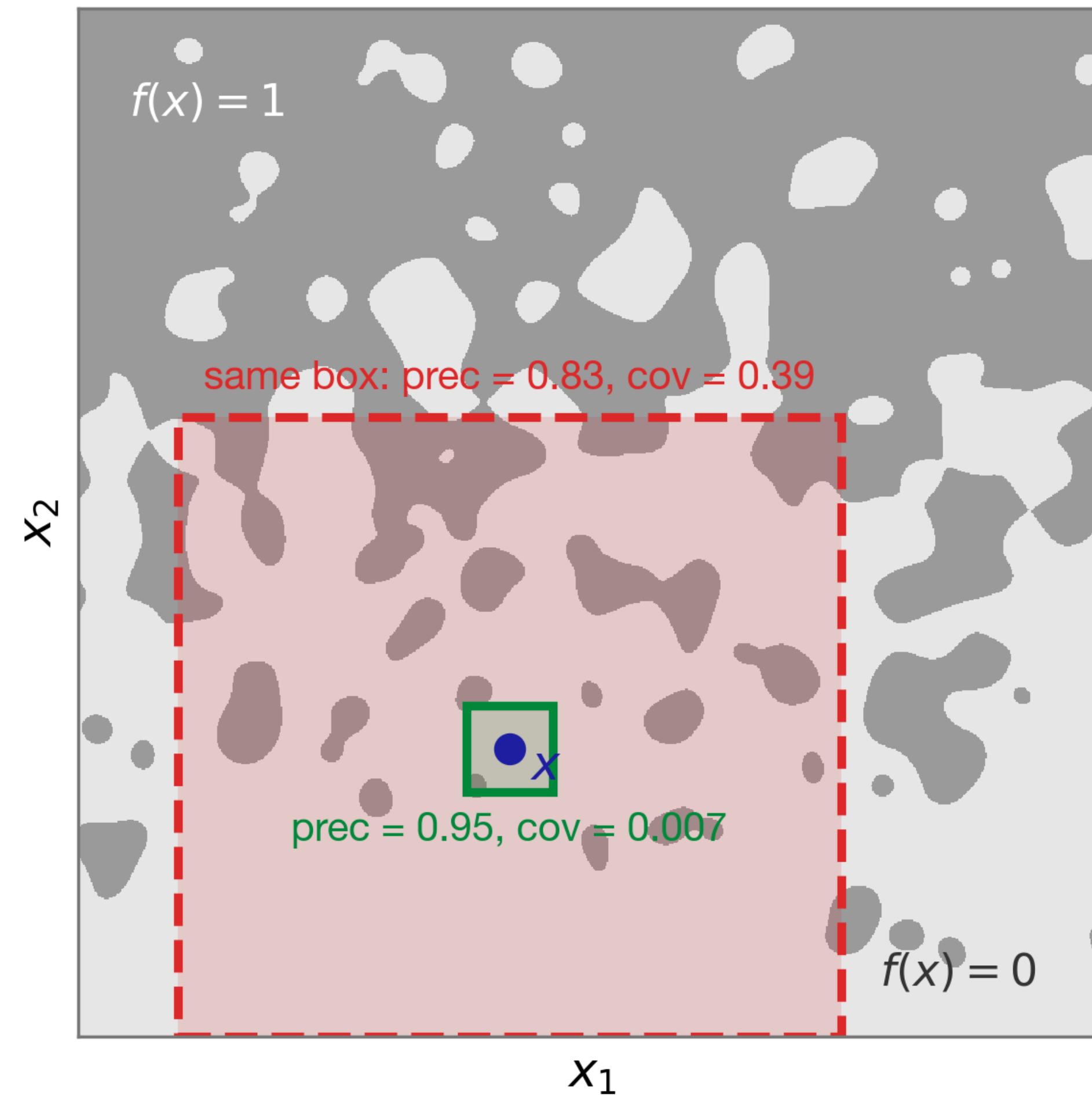
B) decision tree



Q: Are Anchors Useful? What for?

Anchors for Complex Models

(b) complex model



Challenges for XAI

Challenge III: Simple Explanations for Complex Models

- When the goal is to reconstruct the model, explanation must assume some form of simplicity
- The notion of simplicity of our explanation must match the model
- Some models might simply not be ‘explainable’

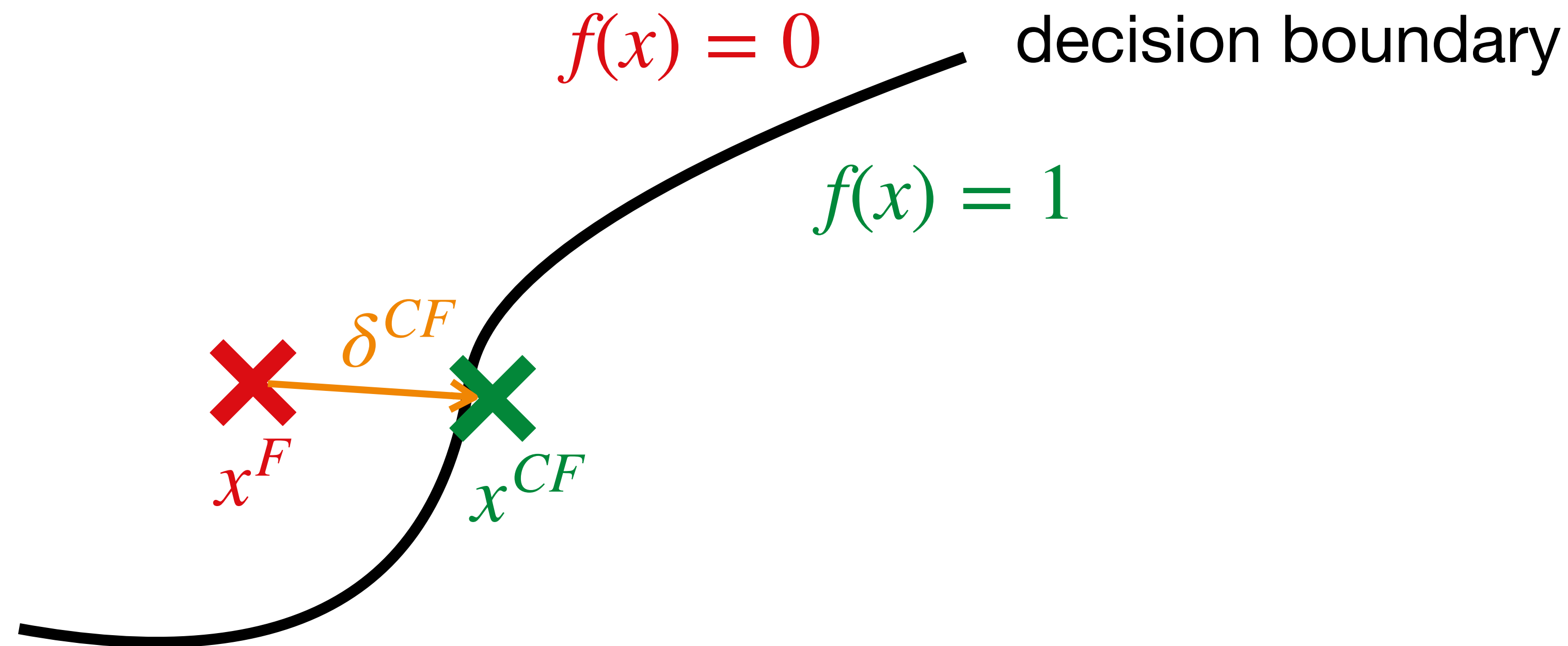
Counterfactual Explanations and Algorithmic Recourse

Counterfactual Explanations (CEs)

Idea: Explain prediction for observation x^F by assessing the minimal changes necessary to get a desirable prediction.

Example: If your debt would be 10000€ smaller, you would be have gotten the loan.

Counterfactual Explanations



Formally: For x^F with $f(x) = 0$, the counterfactual is defined as $x^{CF} := x^F + \delta^{CF}$ with

$$\delta^{CF} \in \operatorname{argmin}_{\delta} \|\delta\|_2 \quad s.t. \quad f(x^F + \delta) = y'.$$

y' is the desired outcome.

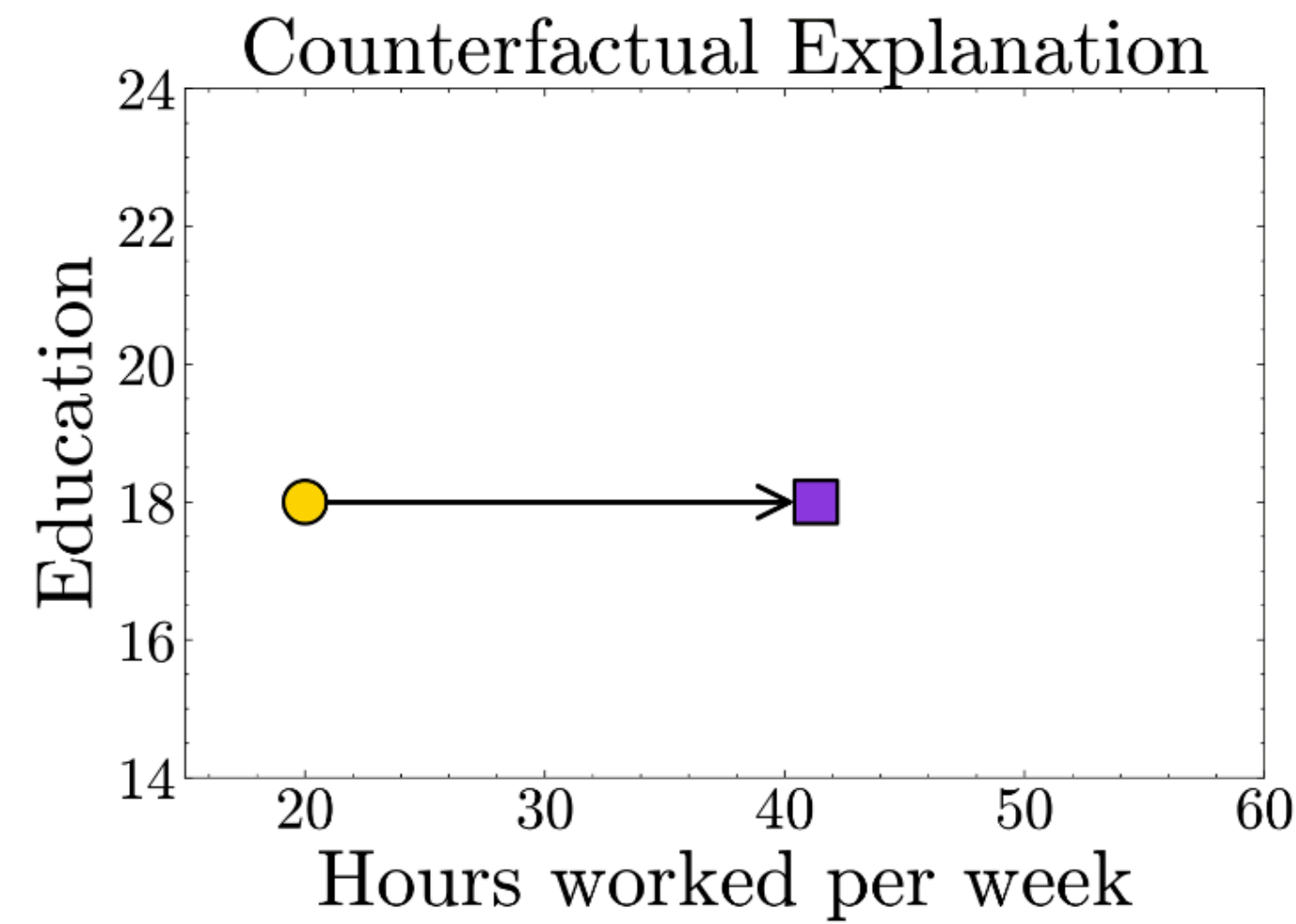
Recourse Setting



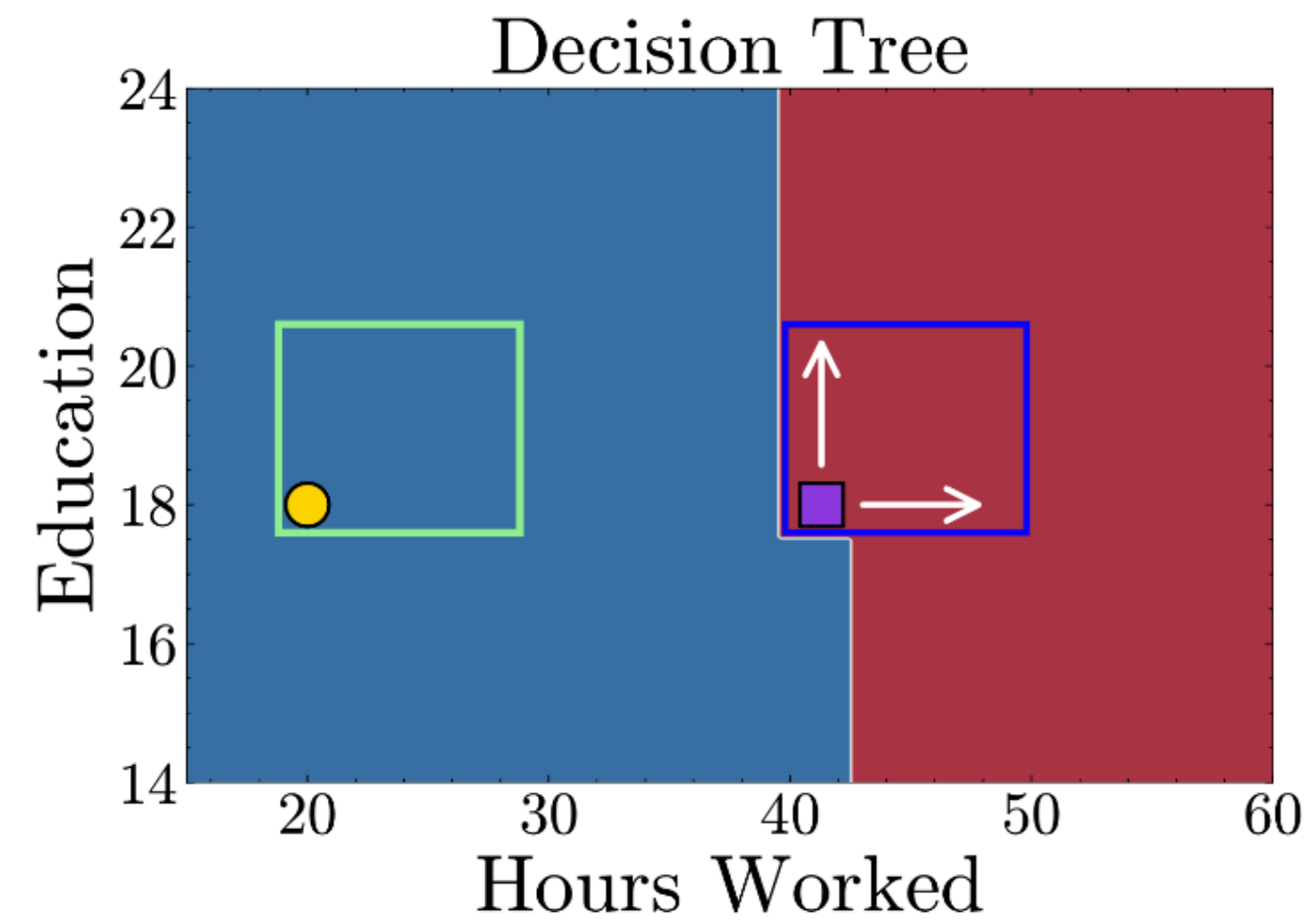
- suppose a ML model rejects your request (job application, loan application, hospital admission,)
- **recourse recommendations** tell you what to do to get accepted

**Are counterfactual explanations
suitable for algorithmic recourse?**

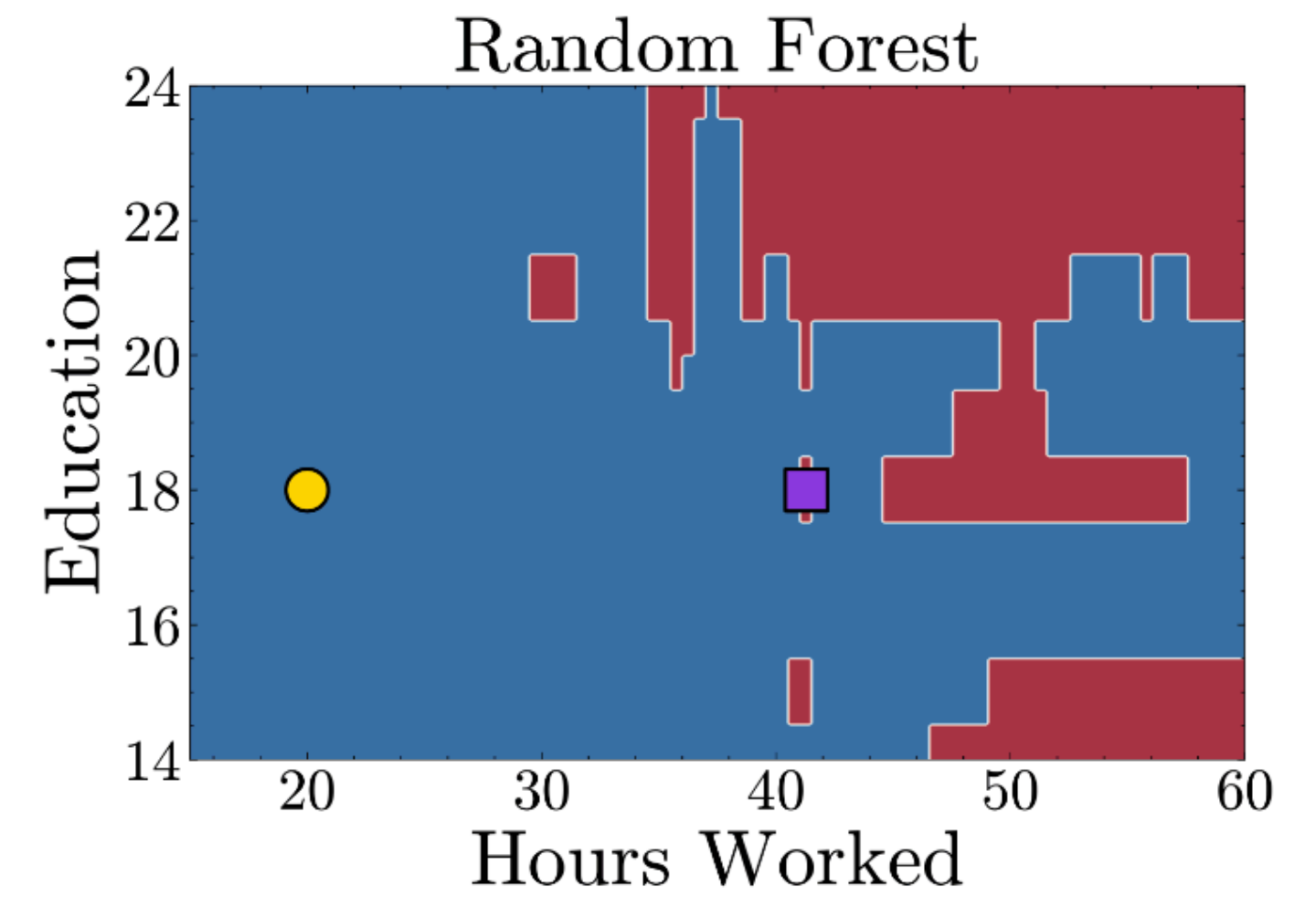
Robustness



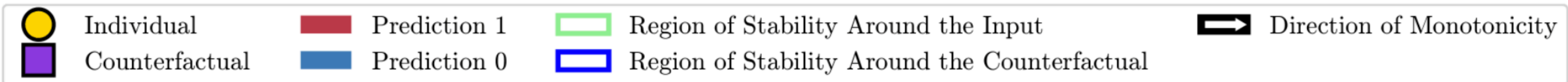
(a) A recipient obtains a counterfactual explanation.



(b) What the recipient believes the decision boundary looks like.



(c) How the decision boundary may actually look like.



Actionability

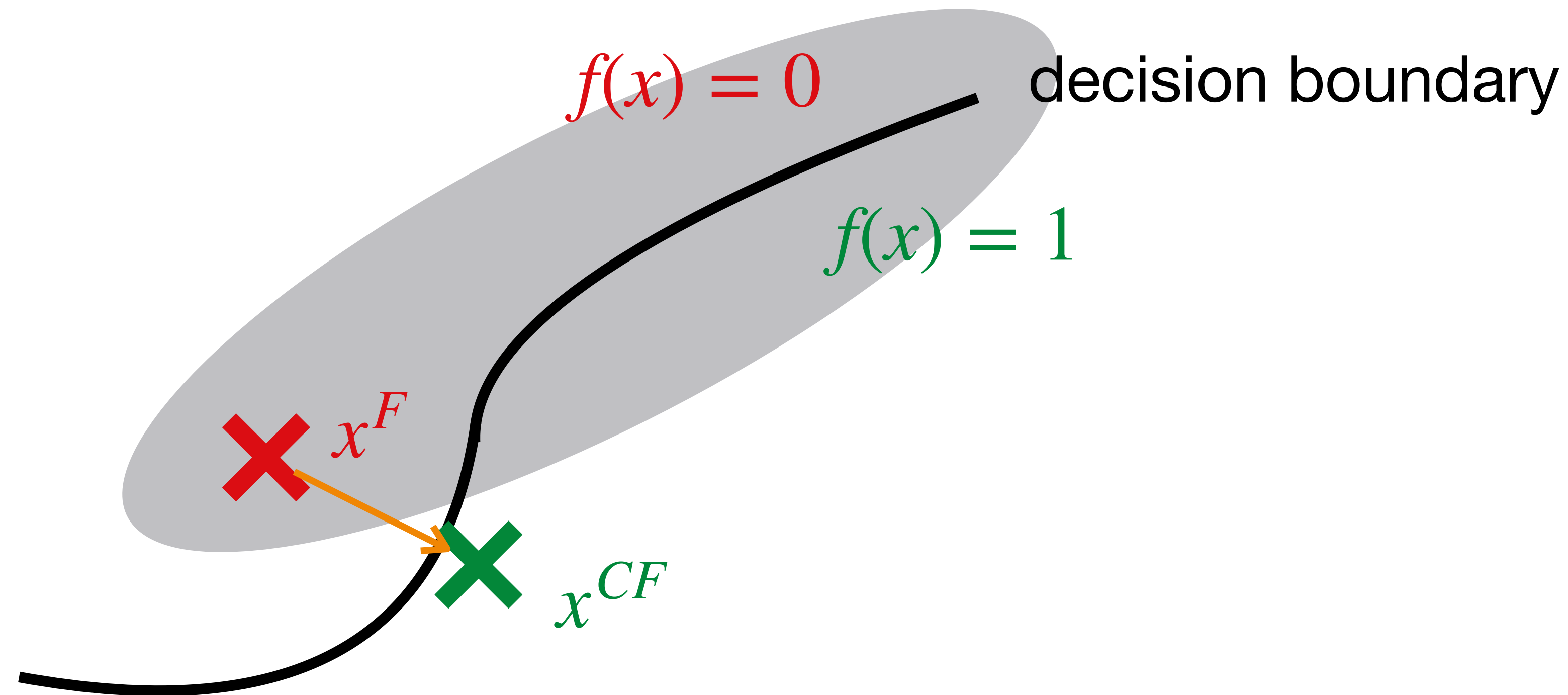
“If you were 10 years younger, you would be accepted”.

⇒ Replace $\|\delta\|_2$ with more sensible cost function $c(\delta)$

*Ustun, Berk, Alexander Spangher, and Yang Liu. "Actionable recourse in linear classification."
Proceedings of the conference on fairness, accountability, and transparency. 2019.*

Plausibility

“If you were a CEO (and still earned 1000\$/mo), you would have been accepted”



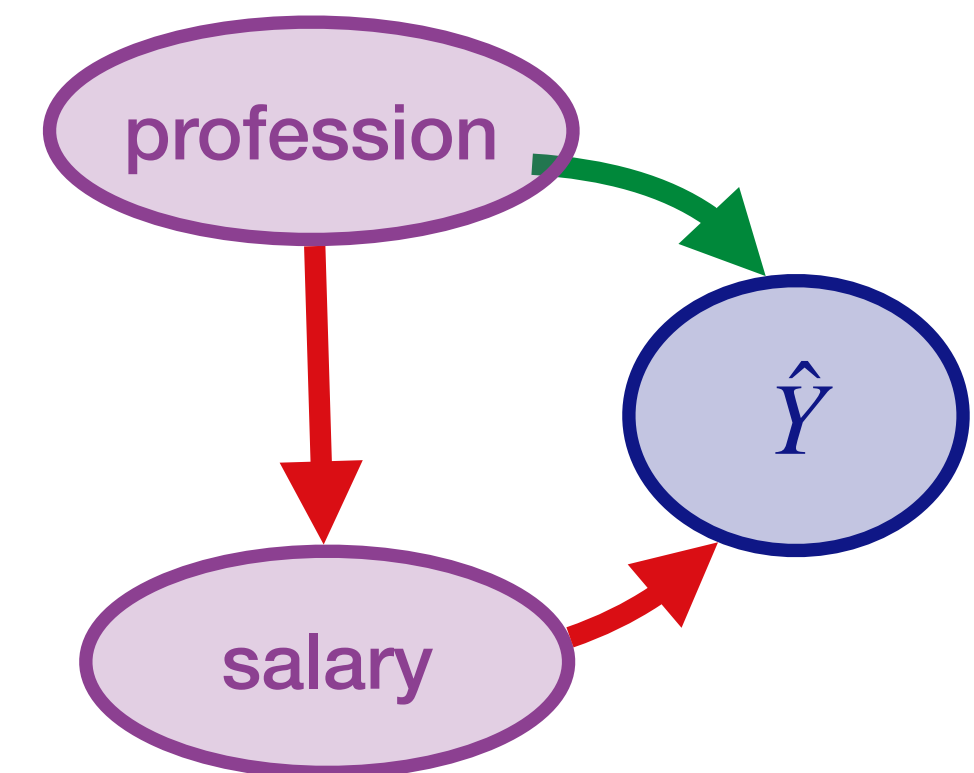
$\Rightarrow$ add constraint that $p(x^{CF}) > 0$

Causal Dependencies

“If you were a CEO (and still had the same salary), you would be accepted”.

Q: What’s wrong here?

⇒ We want to estimate the causal effect of intervening in the real-world



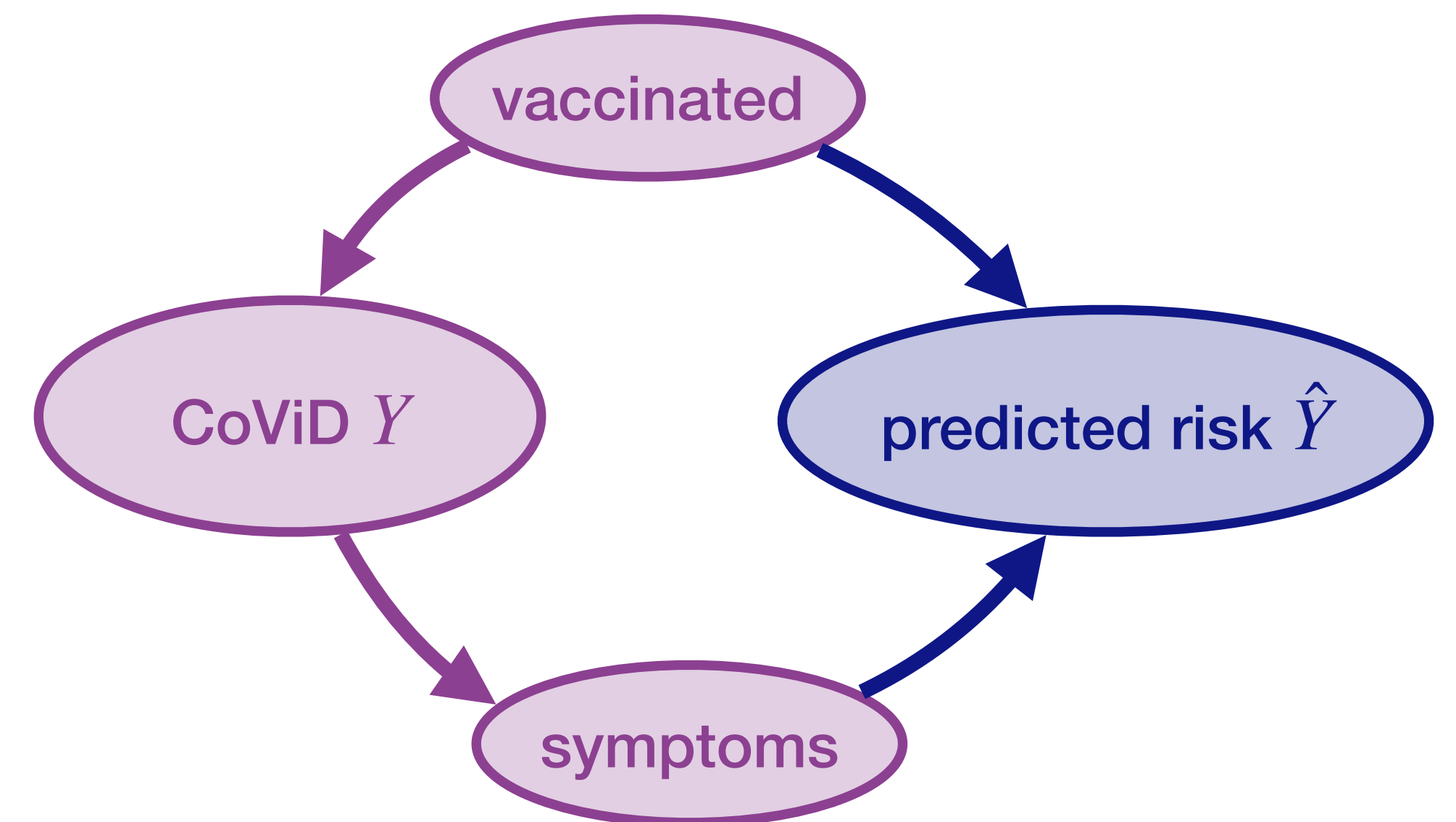
Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera. "Algorithmic recourse: from counterfactual explanations to interventions." Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021.

Acceptance vs. Improvement

“If you treat your symptoms by taking antipyretics, you would be allowed to enter.”

Q: What’s wrong here?

⇒ We want applicants to *improve* ($Y = 1$)



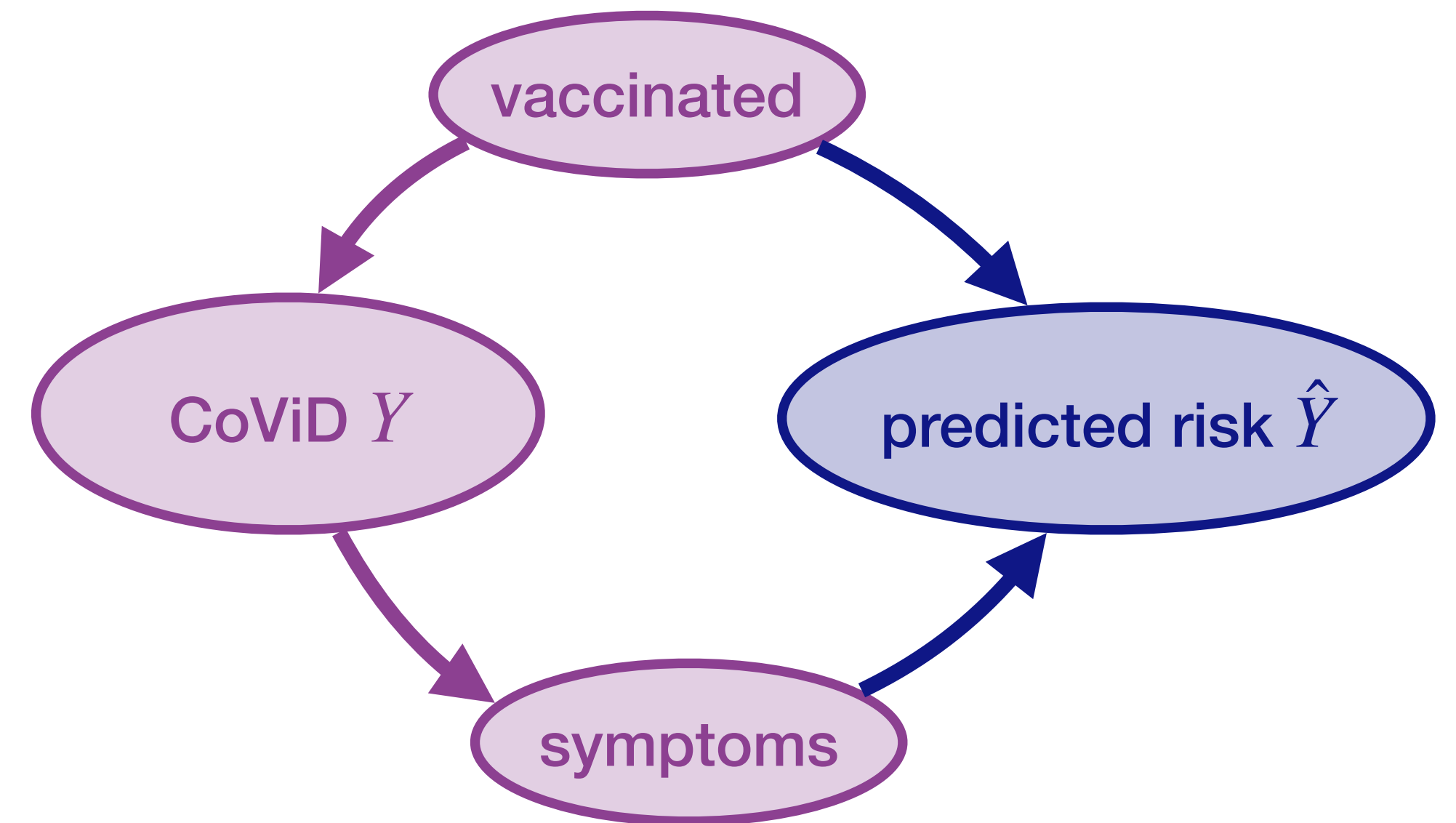
Challenges for XAI

Challenge IV: Estimation

- Many interesting techniques are difficult and costly to estimate
- Many interesting techniques require additional knowledge (user cost function, statistical relationships, causal relationships, ...)

!! Be careful to distinguish... !!

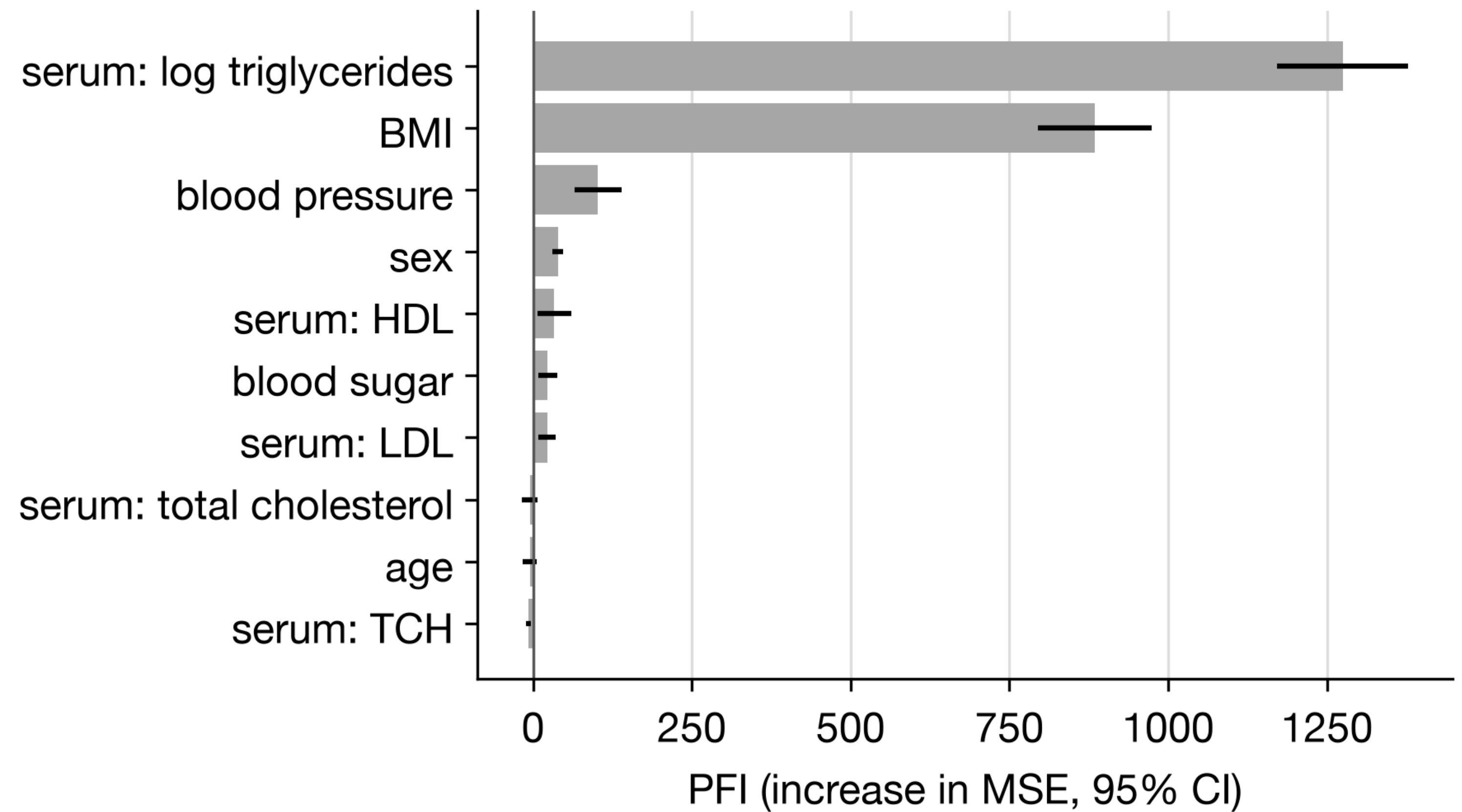
- **model-level causation**
- **association**
- **data-level causation**



(Global) Feature Attribution

S

explain outcomes by
assigning **one score**
per feature



Simple Example: Permutation Feature Importance (PFI)

Permutation Feature Importance (PFI)

$X:$	x_1	...	x_j	...	x_4
	1.4		0		-21
	2.8		1		2.4
	-2		2		3.7
	...		...		...
$\tilde{X}:$	x_1	...	$\sim x_j$	...	x_4
	1.4		1		-21
	2.8		0		2.4
	-2		2		3.7
	...		...		...

$$PFI_j = R(f(\tilde{X}), Y) - R(f(X), Y)$$

Permutation Feature Importance (PFI)

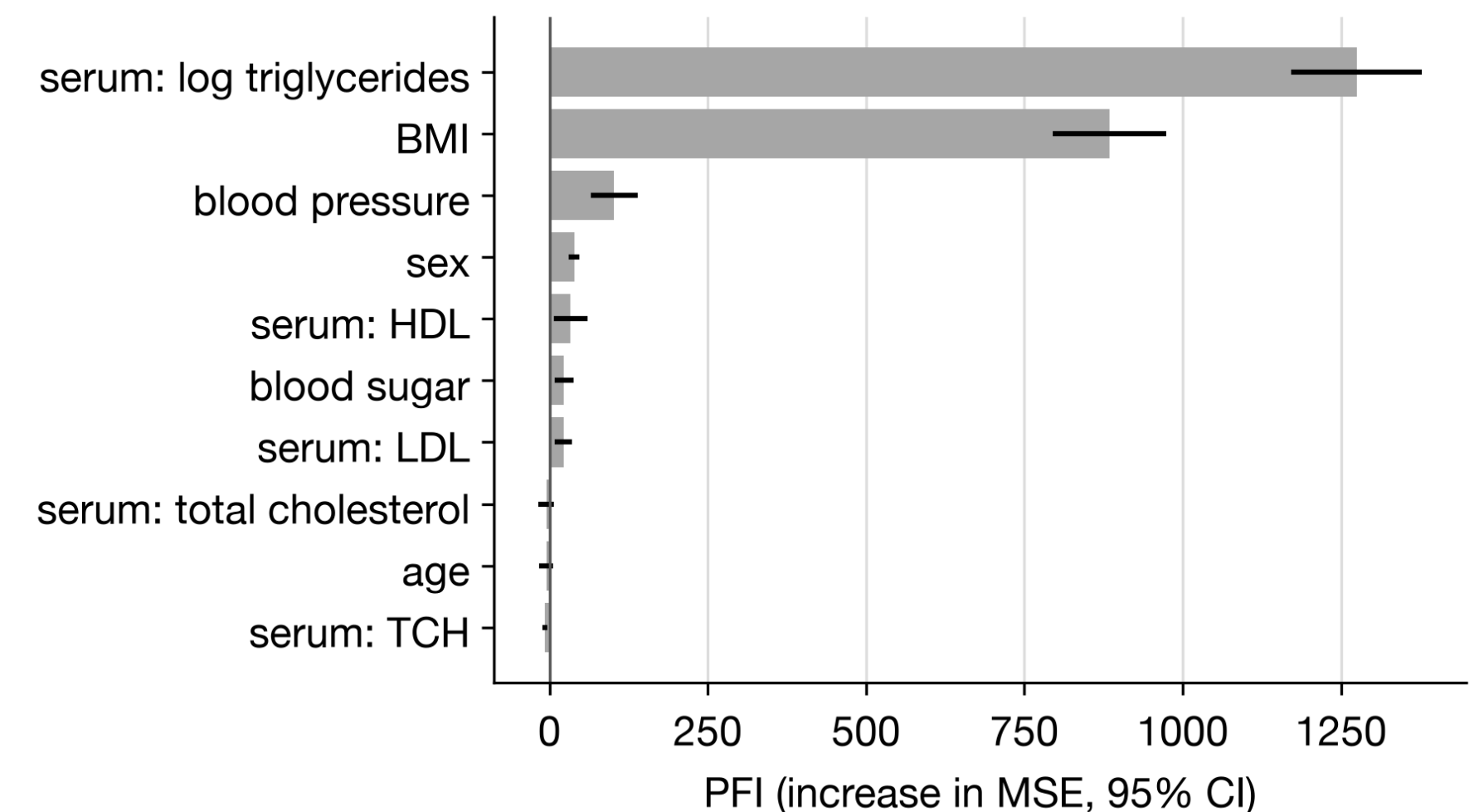
Model-agnostic, global, feature importance method

Idea: Gain understanding by assessing how destroying a feature affects the model's performance.

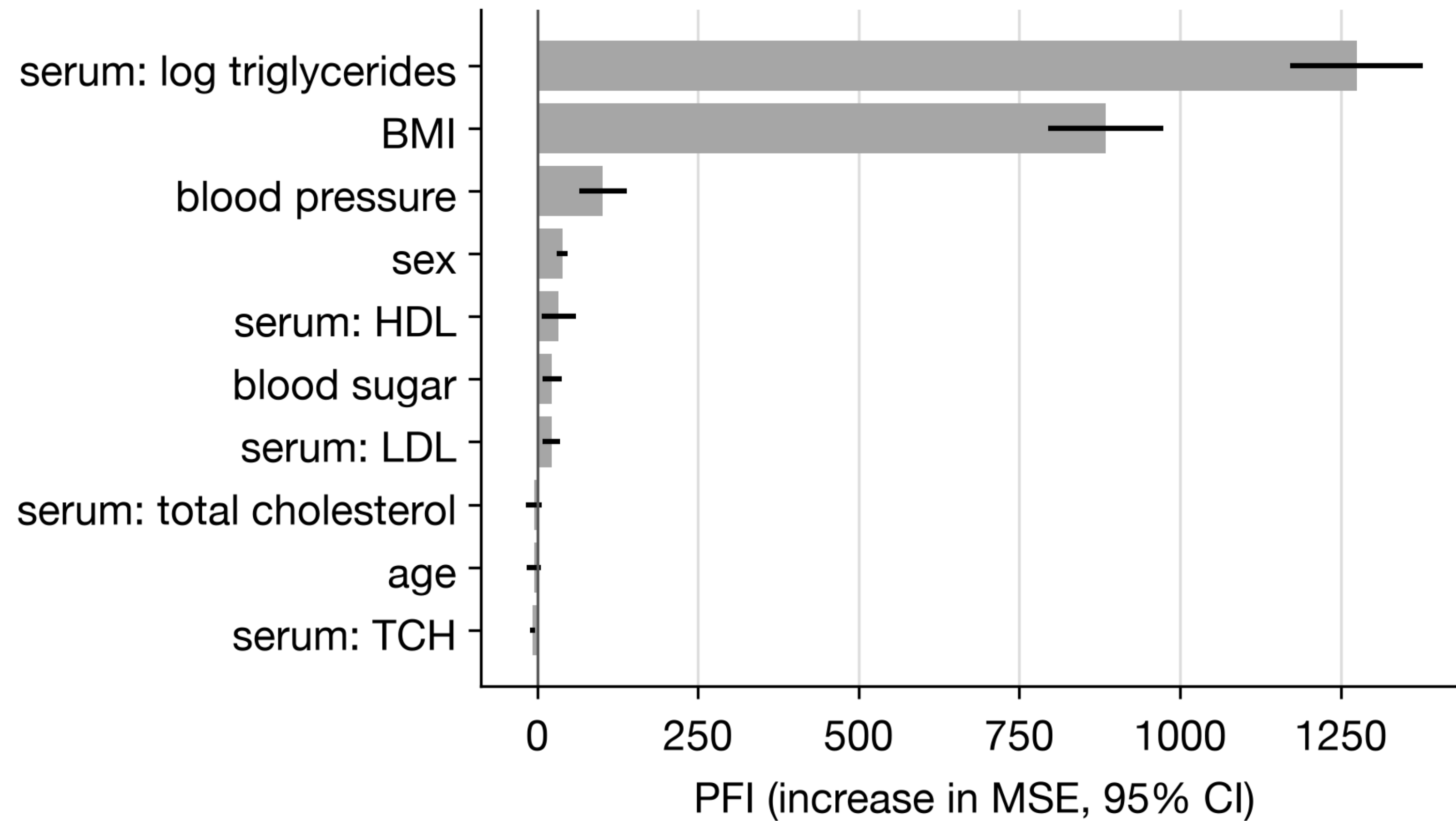
Formally: The difference in risk before and after perturbing X_j .

$$PFI_j := R(\hat{f}(\tilde{X}_j, X_{-j}), Y) - R(\hat{f}(X), Y)$$

with $\tilde{X}_j \sim P(X_j)$ and $\tilde{X}_j \perp X_{-j}, Y$.



Game



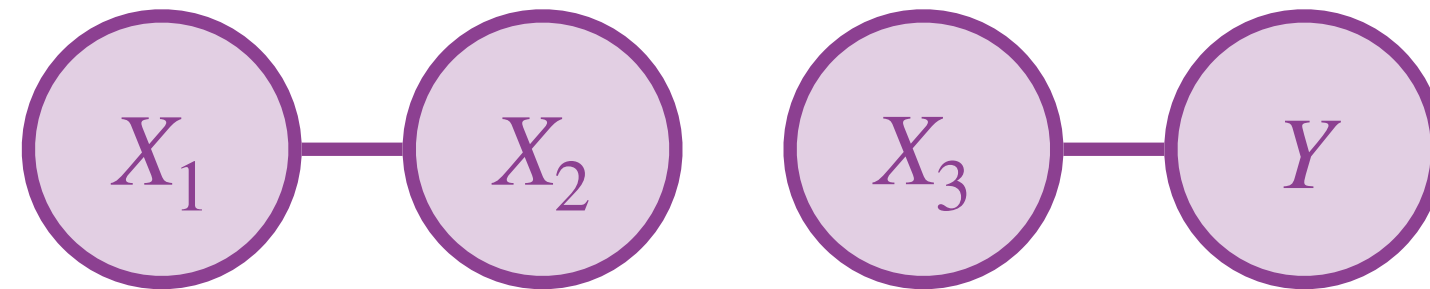
Which variables are most “informative” about diabetes progression?

What could we do to reduce diabetes progression?



Example

Data: X_1 and X_2 nearly perfectly correlated, but irrelevant for Y .
Only X_3 is relevant for Y .



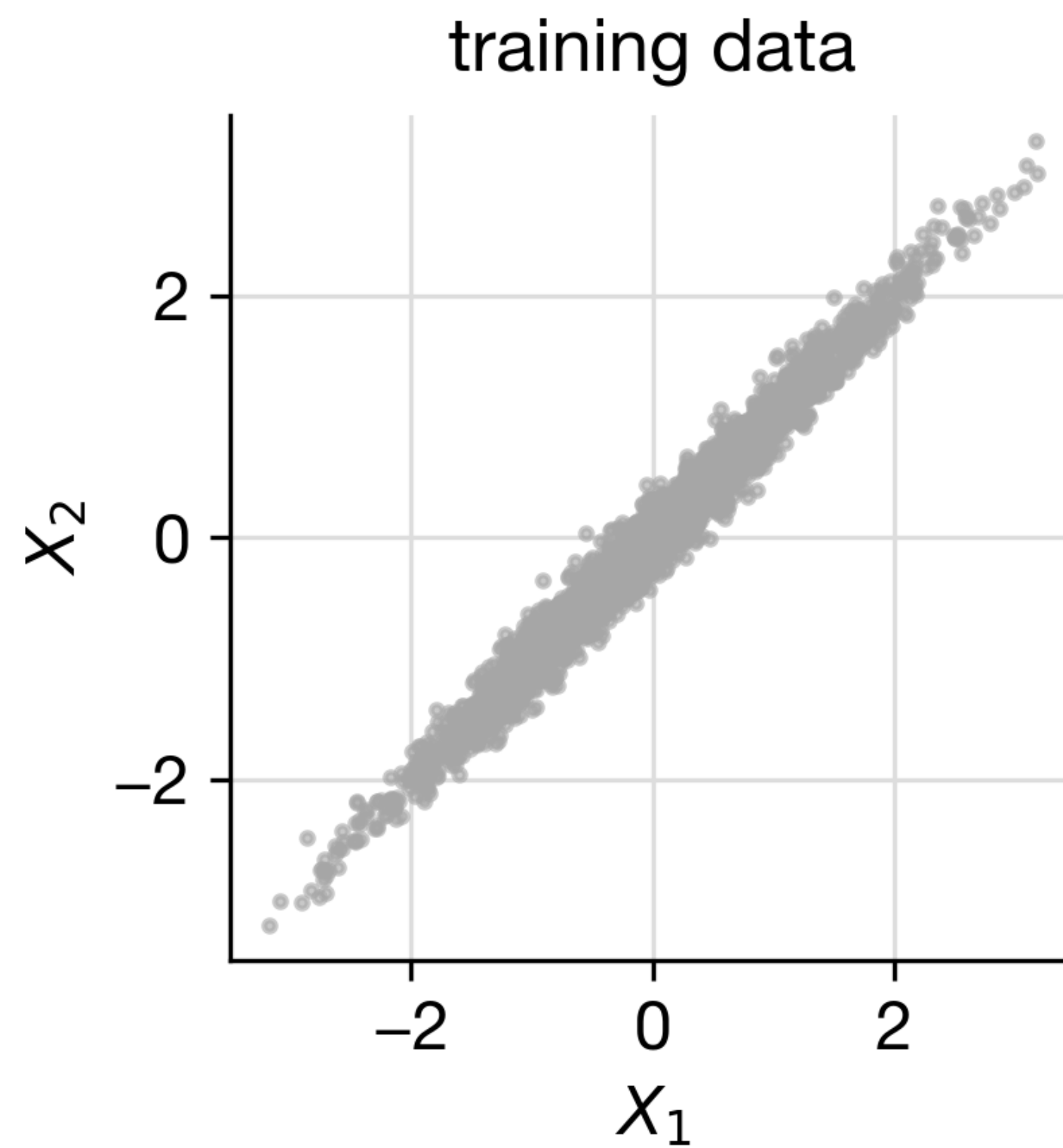
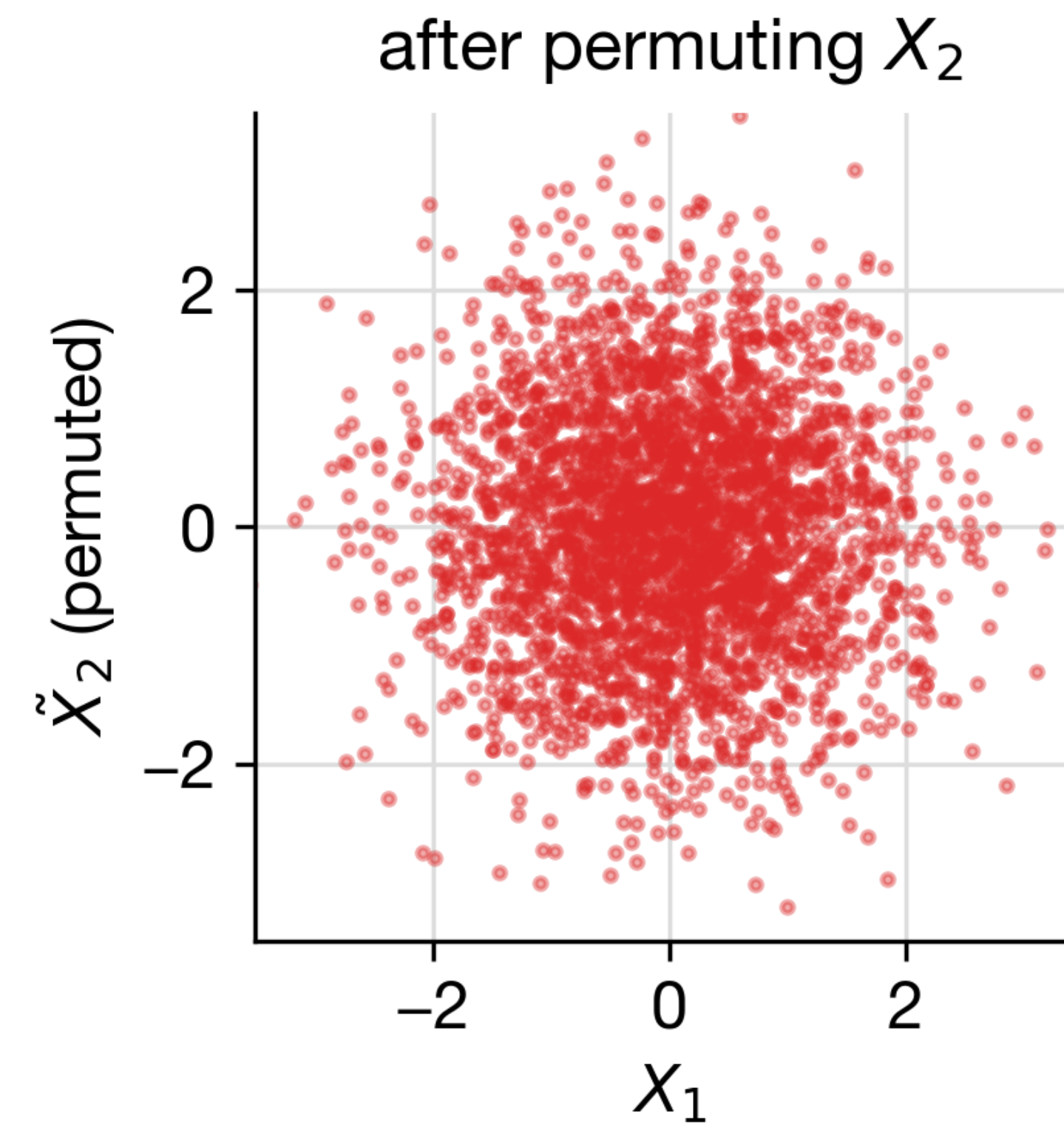
Loss-optimal model:

$$\hat{f}(x) = x_3 + (x_1 - x_2) \approx x_3$$

Q: Which features does PFI consider to be important?
What do you conclude about the suitability of PFI for inference?

PFI and Extrapolation

$X_2 \approx X_1$ (irrelevant for Y), $Y = X_3 + \varepsilon_y$, model $f(x) = x_3 + (x_1 - x_2)$



Observational Value Functions

Observational Value Function

DGP $(X, Y) \sim P$, features $D := \{1, \dots, d\}$, ML model $\hat{f}$, risk R

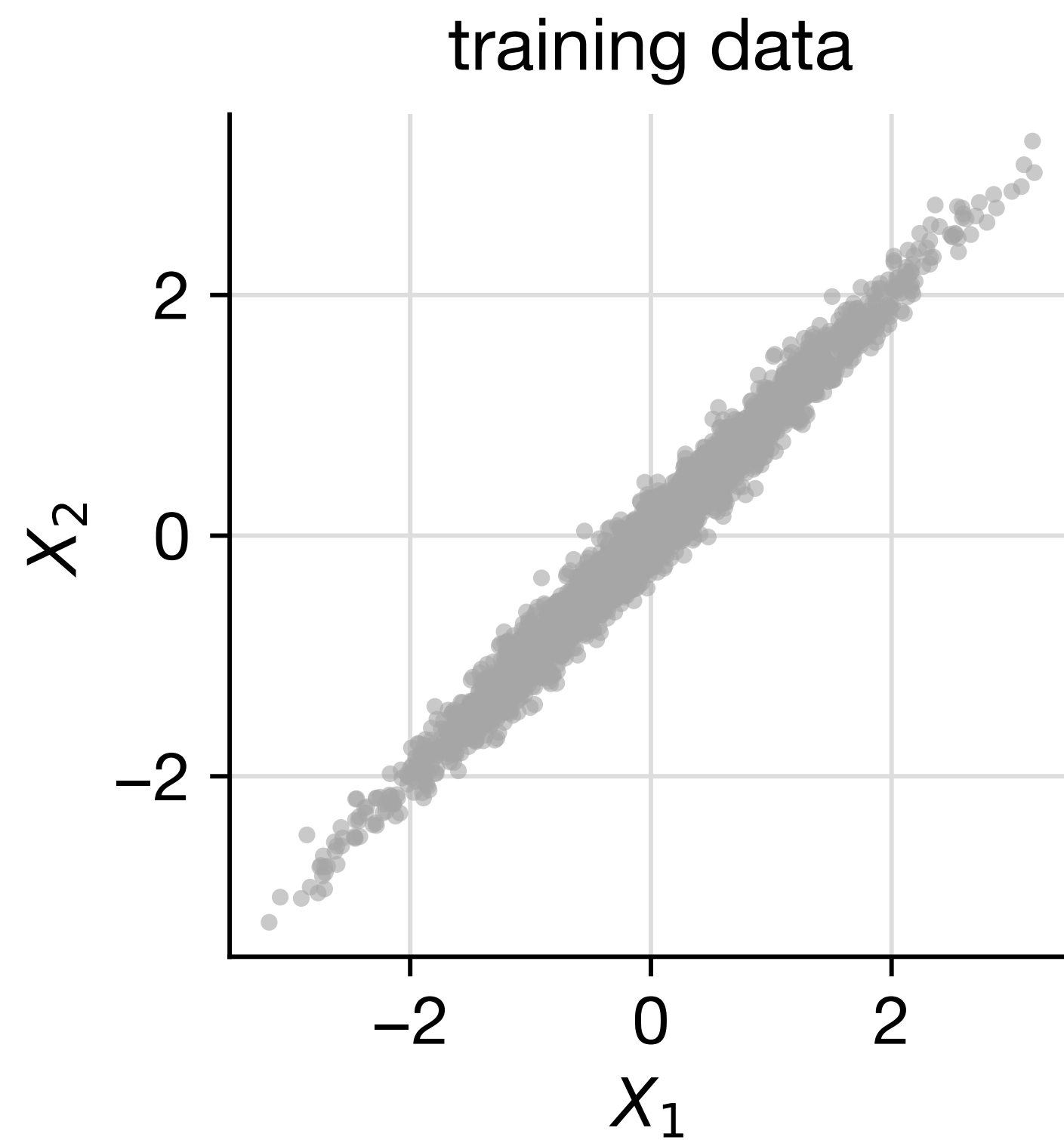
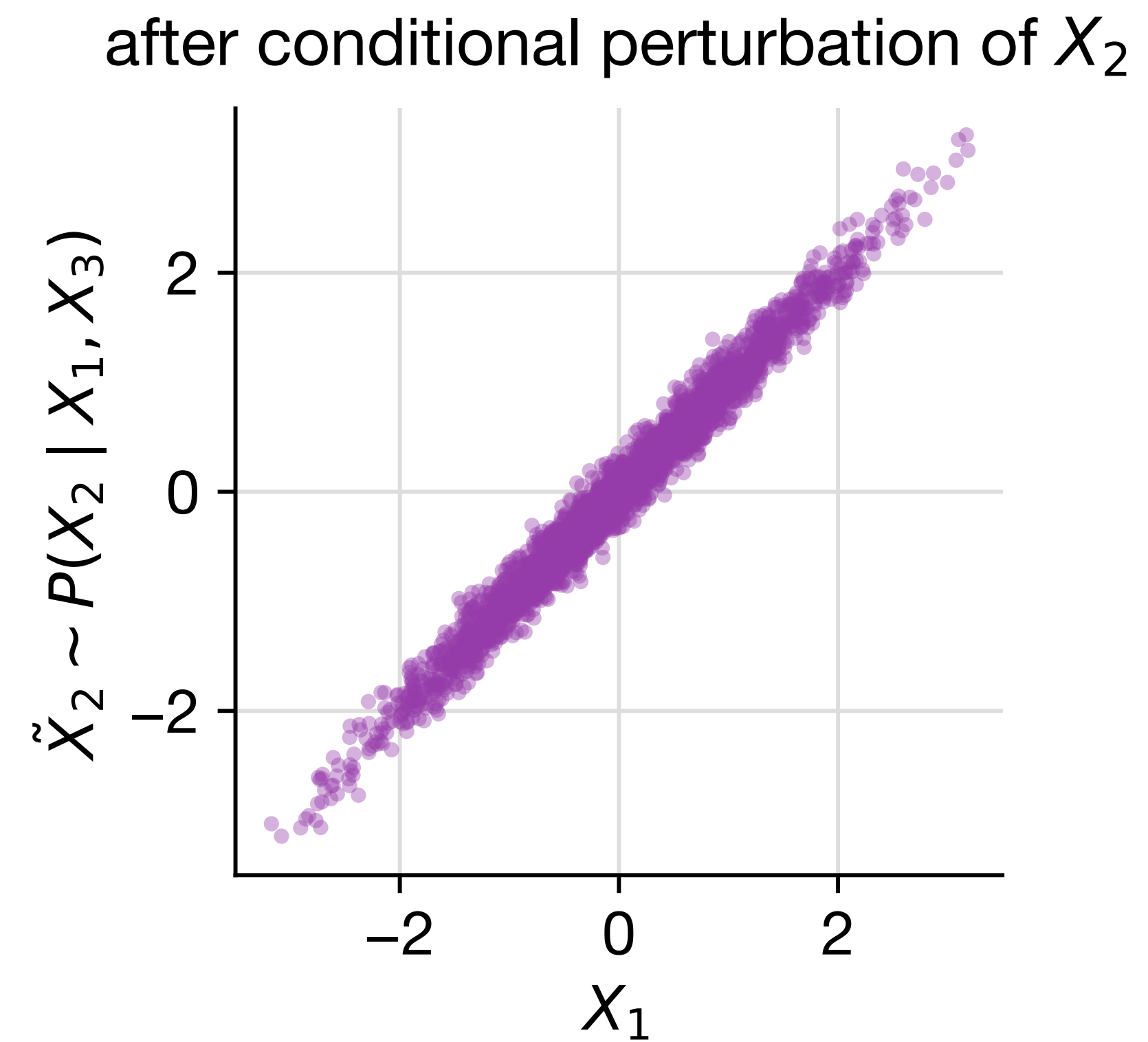
Restricted Model: $\hat{f}_S$ with $S \subseteq D$,

$$g_S(x_S) := E[g(x_S, X_{-S}) \mid X_S = x_S]$$

Value Function: $v_f(S) := R(Y, \hat{f}_\emptyset) - R(Y, \hat{f}_S(X_S))$

Observational Perturbations

$X_2 \approx X_1$ (irrelevant for Y), $Y = X_3 + \varepsilon_y$, model $f(x) = x_3 + (x_1 - x_2)$



DGP Interpretation

For $L2$ -loss and optimal models (e.g., Covert2020):

$$f^*(x) = E[Y | X = x]$$

$$f_S^*(x_s) = E[Y | X_s = x_s]$$

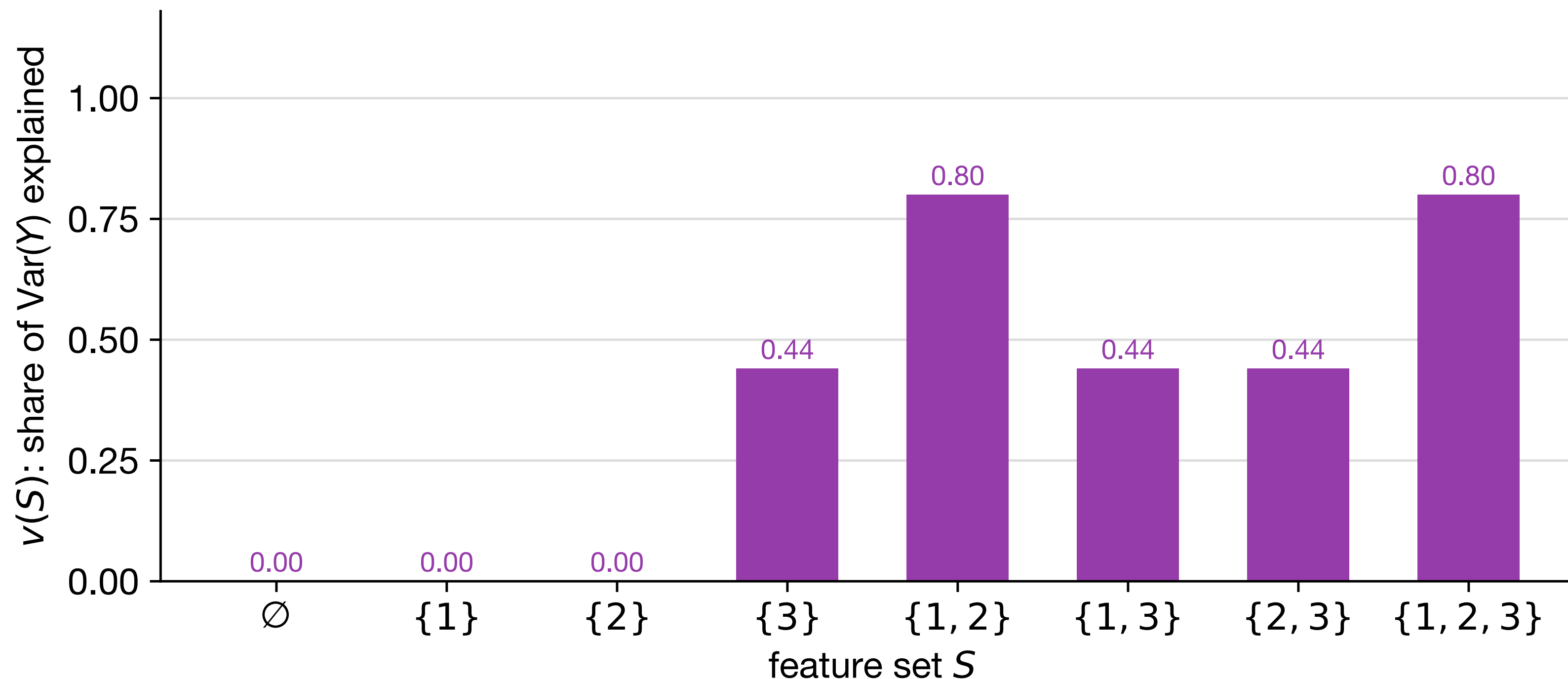
$$v(S) := \text{Var}(E[Y | X_S])$$

The Feature Attribution Problem

Game

Task: Come up with a way to construct a feature attribution score (one score/feature) based on v .

$X_1, X_2 \in \{-1, +1\}$ independent fair coins, $Z = X_1X_2$, $X_3 = Z + \varepsilon$, $Y = Z + \varepsilon_j$



Why is attribution difficult?

Task: What notion of simplicity does feature attribution assume?
What makes attribution difficult in practice?

Answer: Feature attribution “assumes” that features contribute individually.

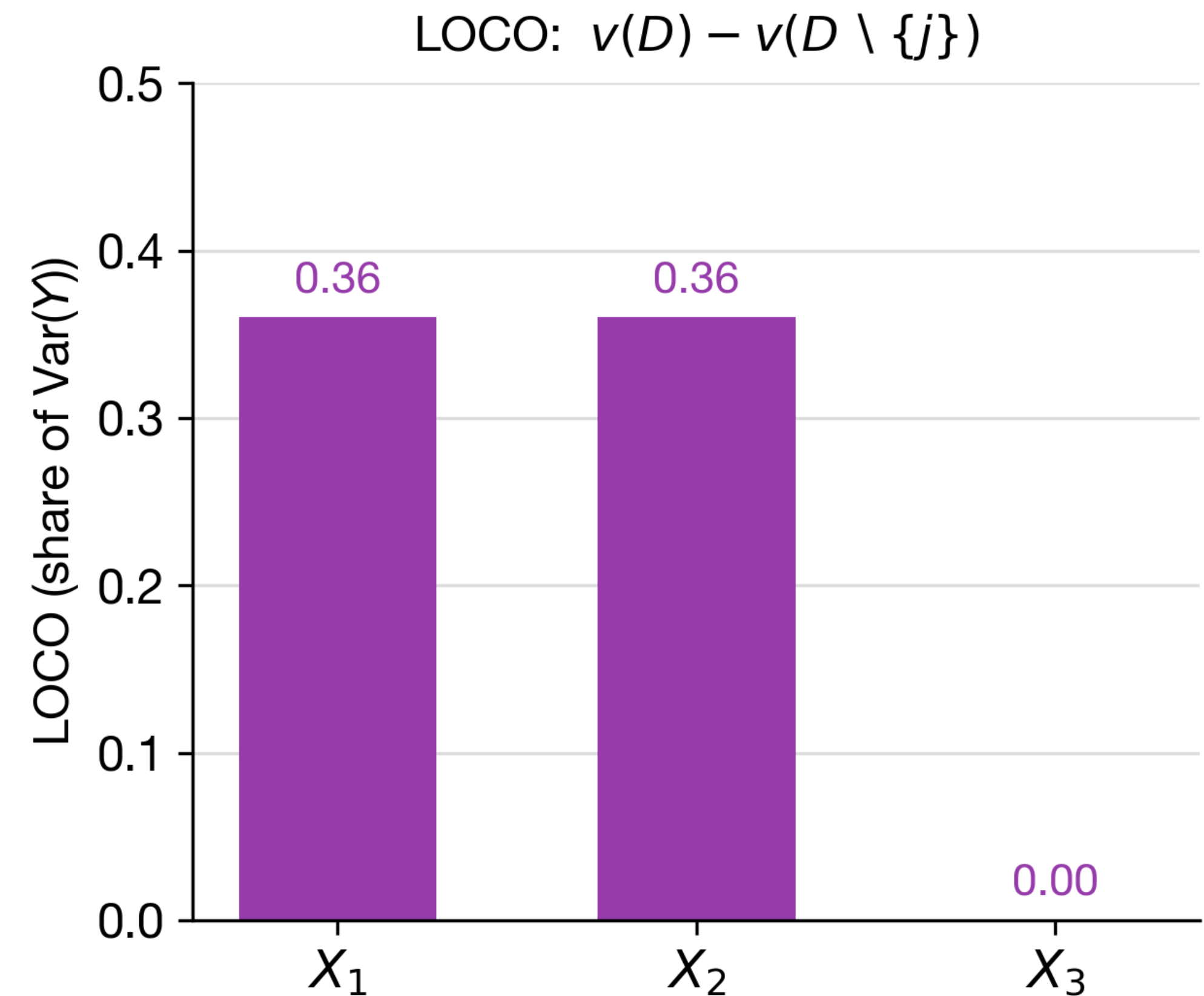
If there are no interactions and no dependencies

$$v(S) = \sum_{i \in S} v(i).$$

Leave-One-Covariate-Out (LOCO)

How much surplus does feature j contribute over all remaining features $-j$?

$$\phi_j = v(j \cup -j) - v(-j)$$

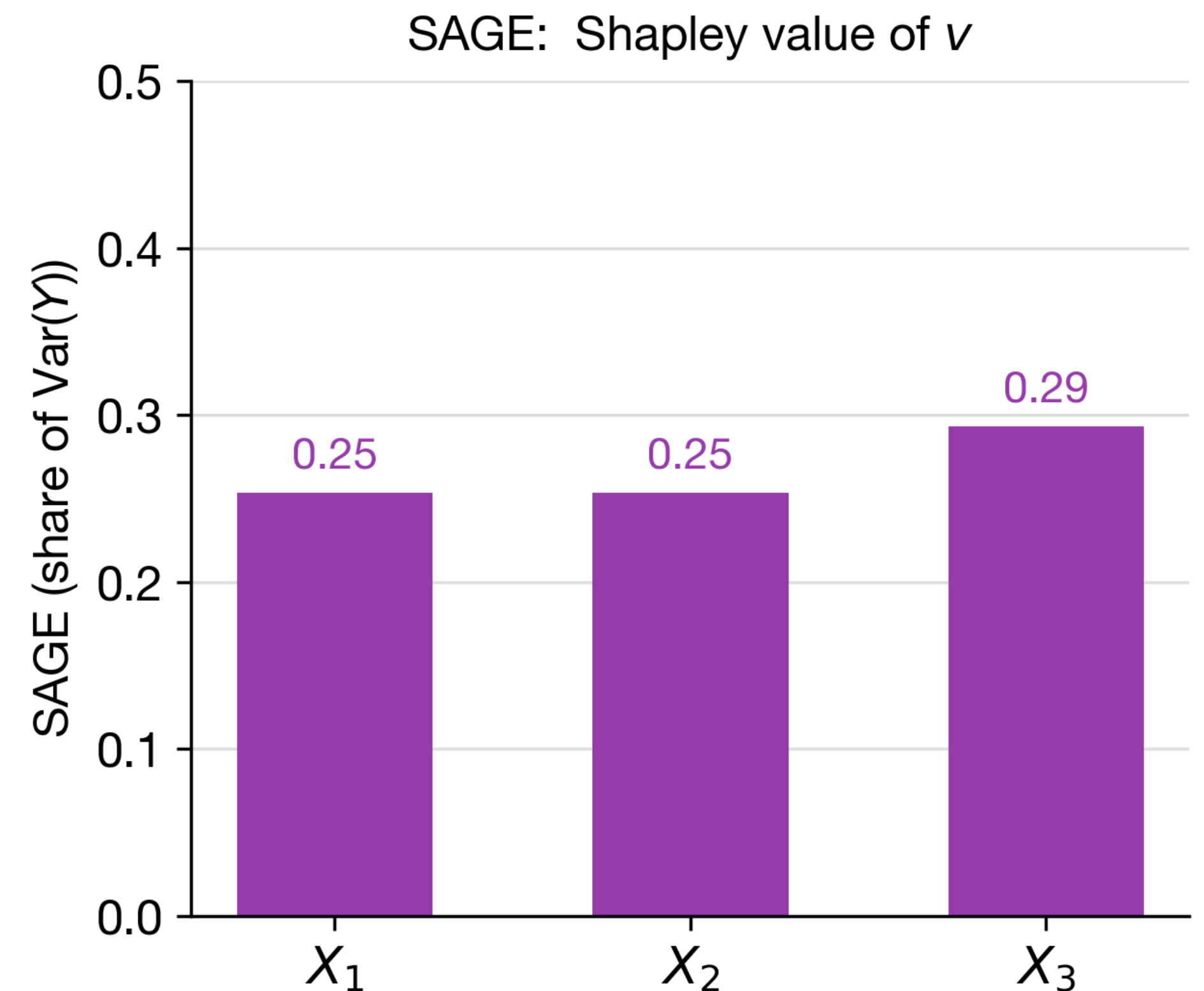


Shapley Value approach

How much does feature j contribute over features $S_{\pi,j}$ that appear before j in some ordering π , averaged over all orderings π .

$$\phi_j = \sum_{\pi \in \Pi} v(j \cup S_{\pi,j}) - v(S_{\pi,j})$$

⇒ More next week

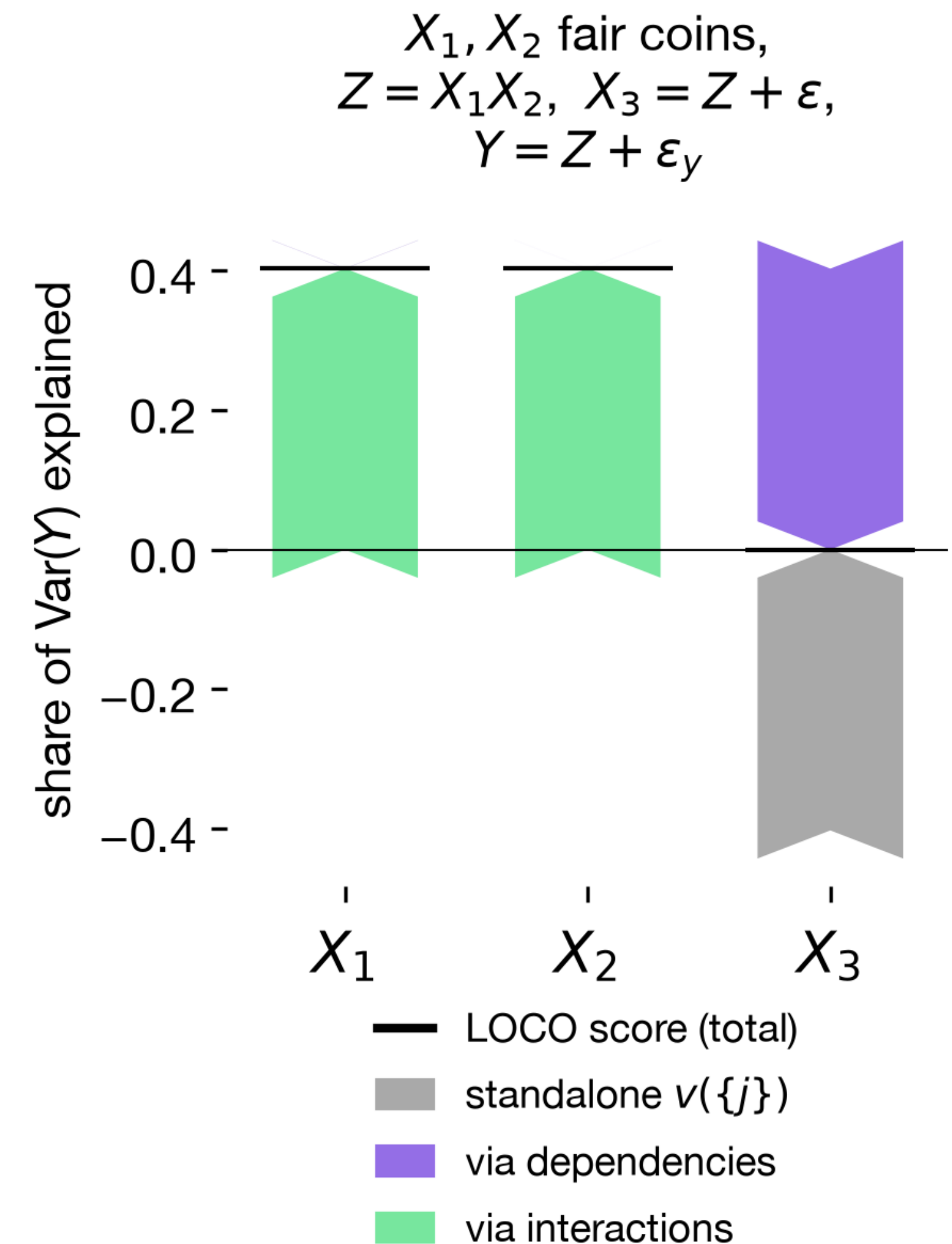


Outlook: DIP Decomposition

Move beyond one score per feature?

$$\phi_j = v(j) + \text{Int}(j, -j) + \text{Dep}(j, -j)$$

⇒ User can see whether simplicity notion is justified.



Conclusion

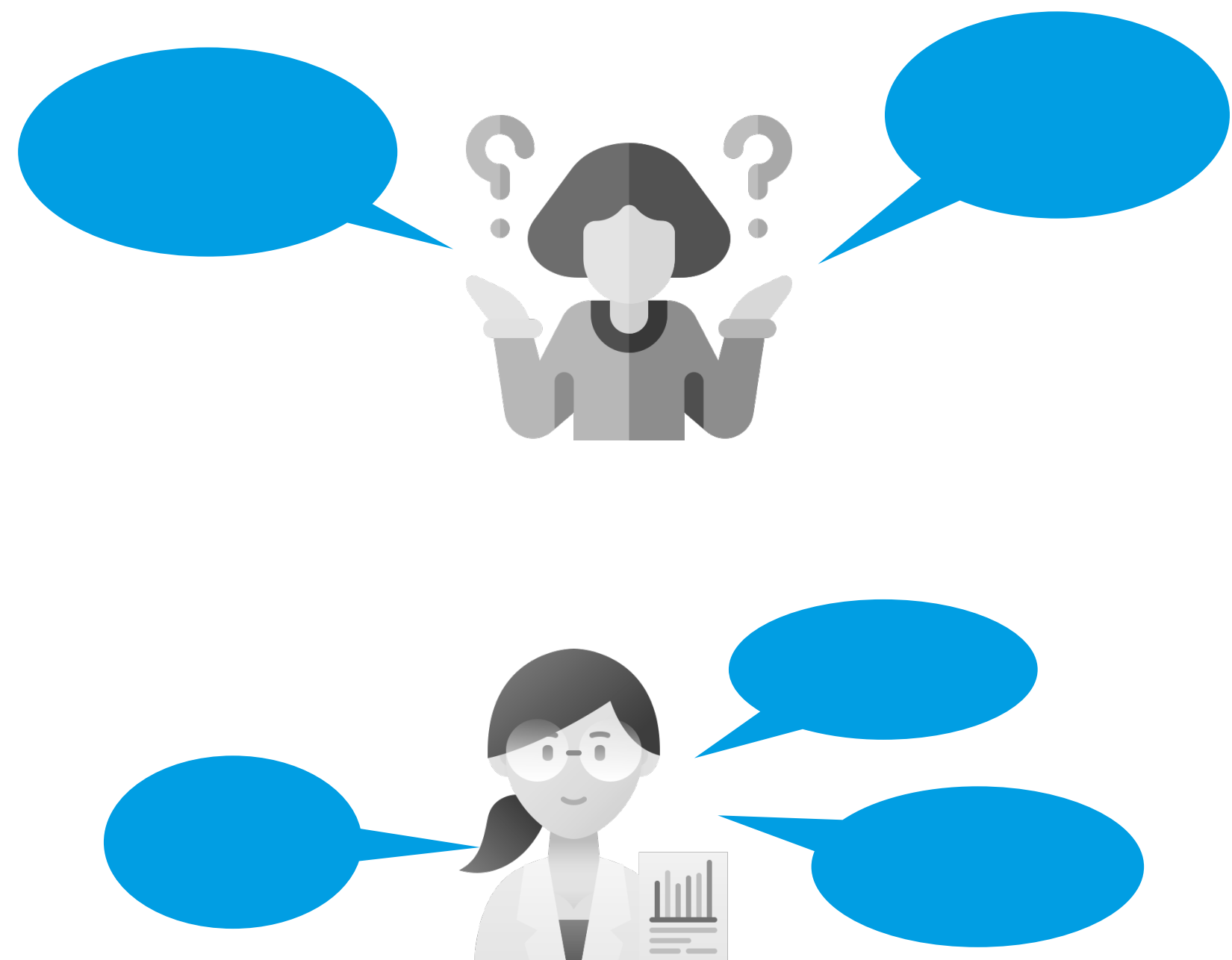
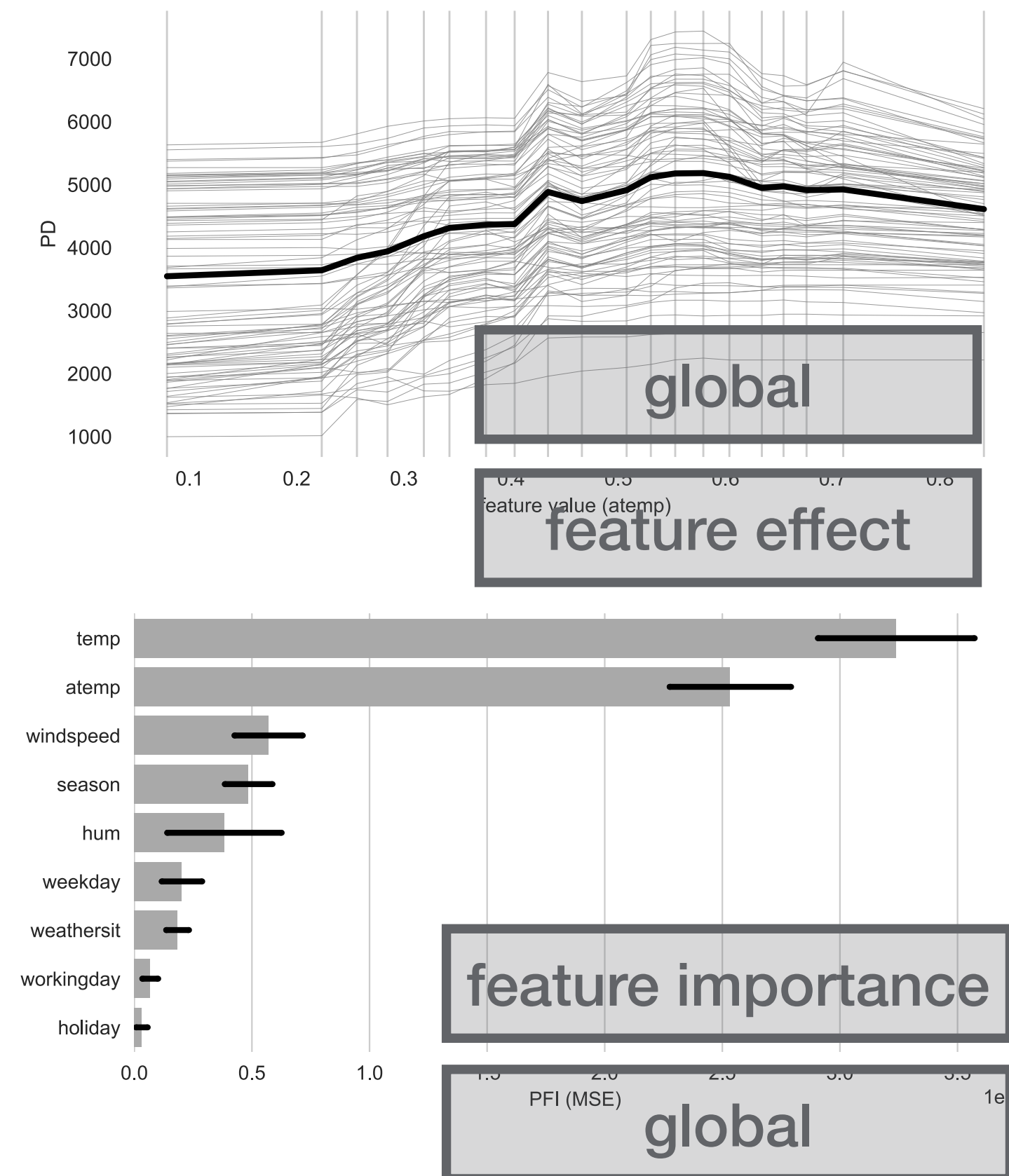
Conclusion: Be rigorous about ...

- What concrete question are we trying to answer?
- Does the XAI method allow answering it?
- Distinguish different levels of causation (model, data) and association.
- What notion of simplicity does the XAI method assume? Can the explanation user know whether the assumption holds?
- Can the XAI method be estimated in practice?

Thanks!

Supplement

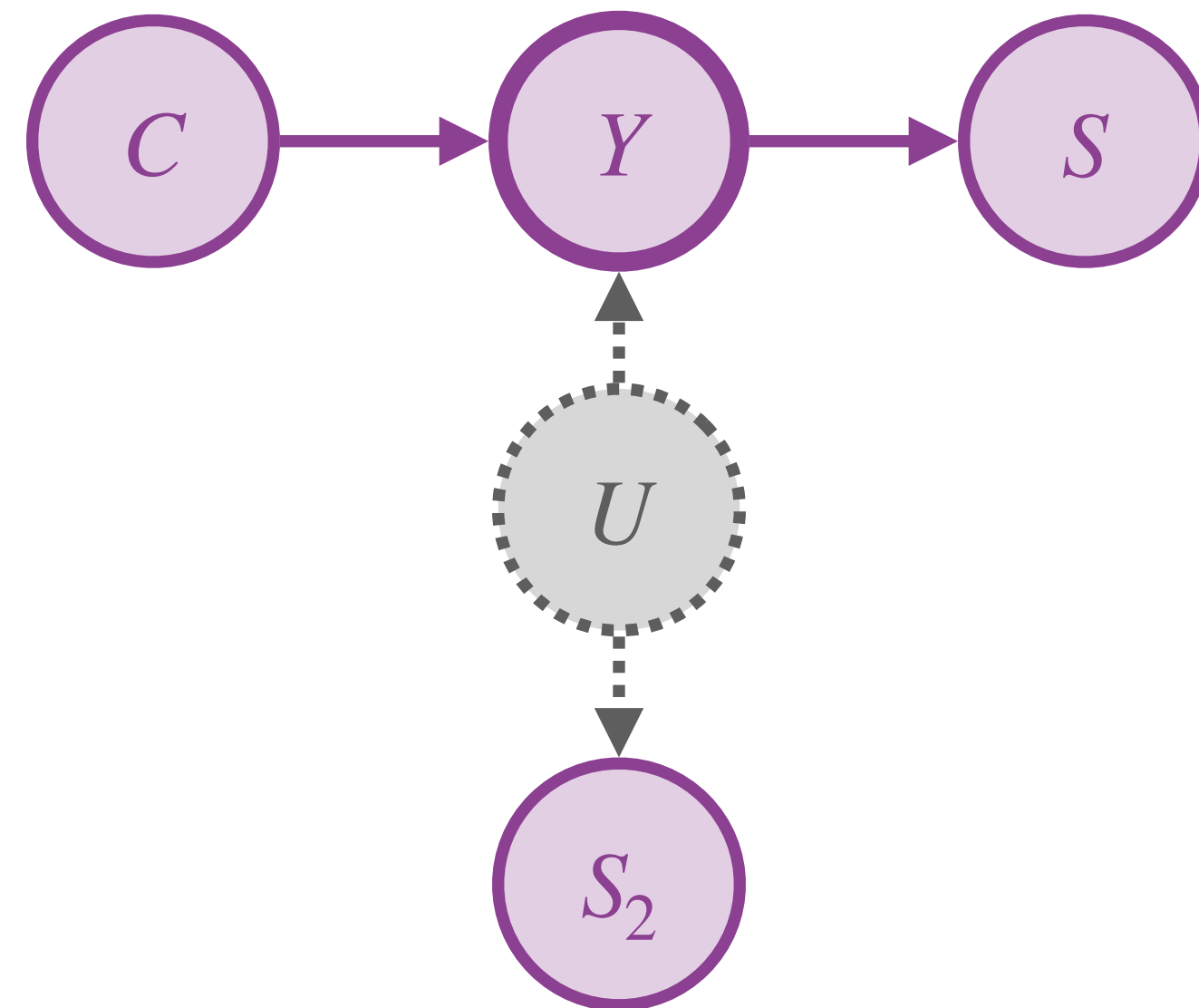
Importance of A Causal Perspective



Causality helpful both to formalize question (Challenge I)
& interpretation rules (Challenge II)

Causality: A Short Primer

Association $\neq$ Causality



disease Y , risk factors C , symptoms S , unobserved confounder U

The *do*-operator

Notation

$$P(Y = 1 \mid S = 1) \\ -P(Y = 1 \mid S = 0)$$

$$P(Y = 1 \mid do(S = 1)) \\ -P(Y = 1 \mid do(S = 0))$$

Translation

Are people with symptoms more likely to be sick than people without symptoms?

Does treating the symptoms affect the probability of being sick?

-> *association*

-> *causation*

Model-level vs Data-level Interventions

$$do(\underline{X} = x')$$

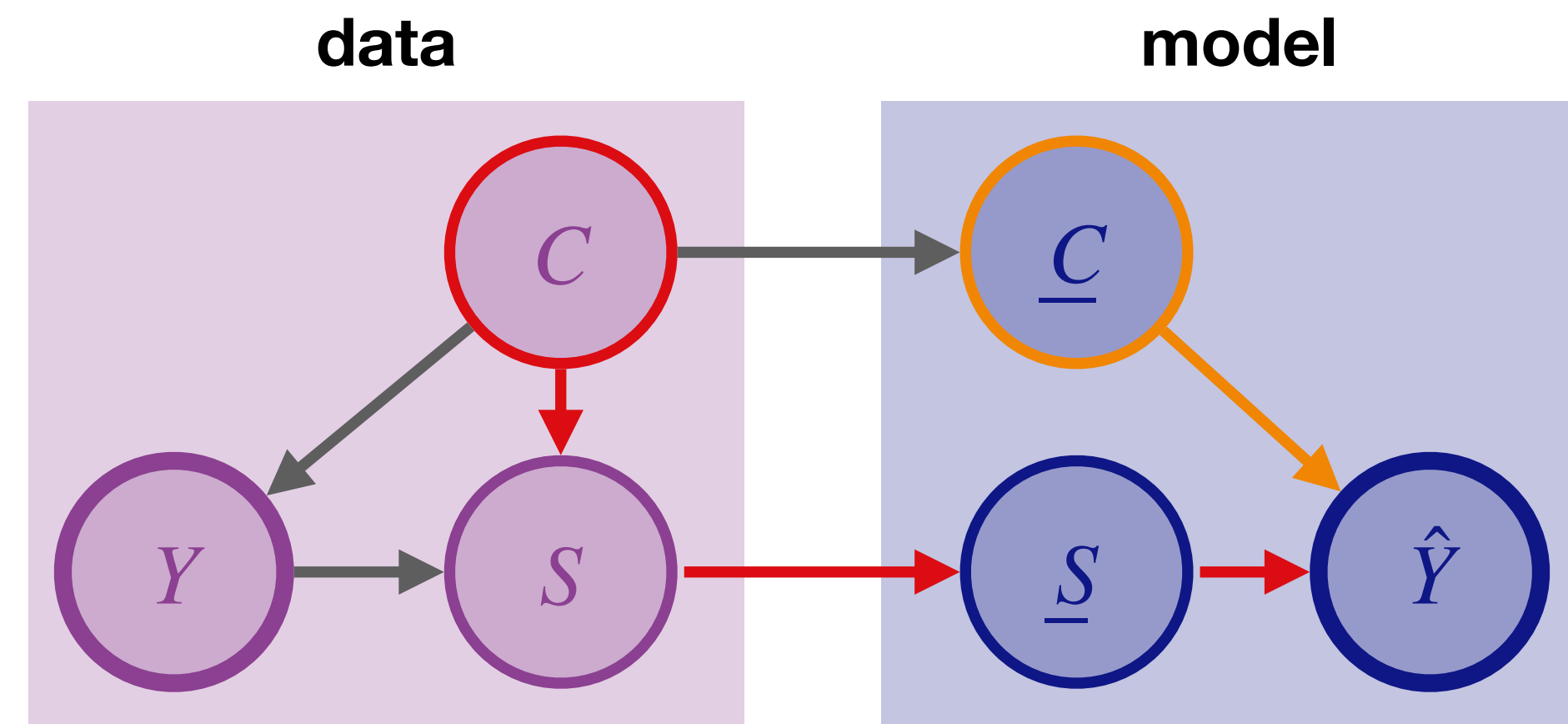
model-level interventions

- plugging different values into the model inputs
- Interventions on one model input does not affect other model inputs

$$do(X = x')$$

data-level interventions

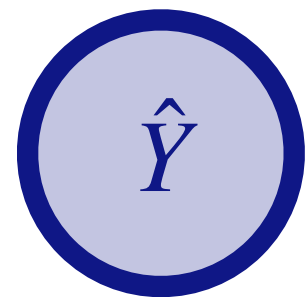
- Interventions in the real world
- May affect covariates as well



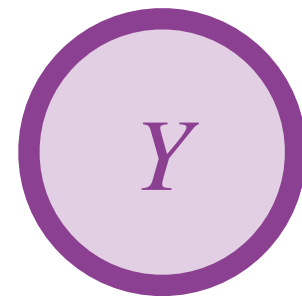
Nine Perspectives on Model and Data

Nine Perspectives on Model and Data

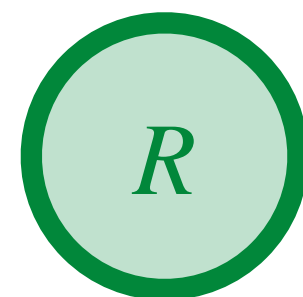
What object is explained?



prediction

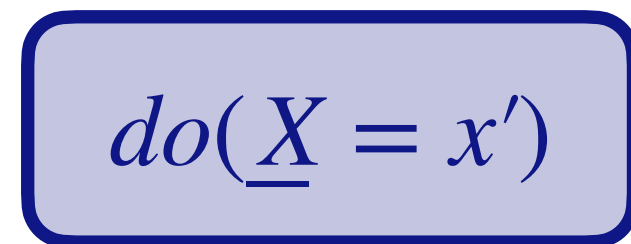


target

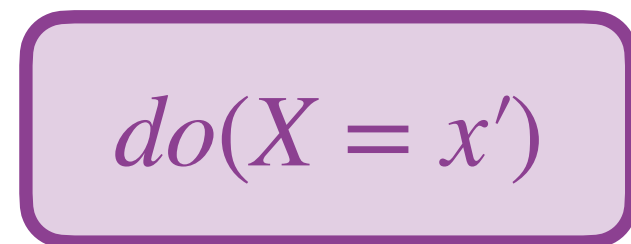


relationship

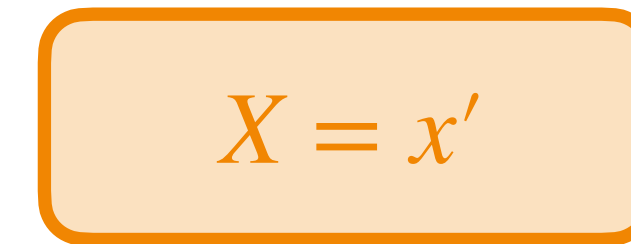
On what level is the object explained?



model-level interventions



data-level interventions



association

Nine Perspectives on Model and Data

	$\hat{Y}$	R	Y
$do(\underline{X} = x')$	Does the model rely on the feature “gender”?		
$X = x'$			
$do(X = x')$	What can I do to get accepted?		What can we do to treat the disease?

Nine Perspectives on Model and Data

	$\hat{Y}$	R	Y
$do(\underline{X} = x')$	Does the model rely on the feature “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?

Intermediate Conclusion

- In general, the nine perspectives on model and data are distinct
 - The questions we try to answer in XAI concern distinct perspectives
- ⇒ There is no universal XAI method
- ⇒ Requirements towards methods are not universal