Model-agnostic explanations

and their limitations
(featuring: Causal Dependence Plots)

Joshua Loftus, Lucius Bynum, Sakina Hansen, Kateryna Koval

Explanations and interpretability

Some historical context/dialectics

- Behaviorism vs cognitivism
- Two cultures (Breiman, 2001) data modeling and algorithm modeling
- Algorithm vs inference (Efron and Hastie, 2016)
- Interpretability is a constraint, hence SOTA methods for prediction tasks tend to be opaque (“black box”)

Causality is hard

Udny Yule, inventing multiple regression in 1897:

> “The investigation of causal relations between economic phenomena presents many problems of peculiar difficulty, and offers many opportunities for fallacious conclusions.
>
> Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and discuss as best he can the relations of a whole group of changes; **he cannot, like the physicist, narrow down the issue to the effect of one variation at a time. The problems of statistics are in this sense far more complex than the problems of physics**.”

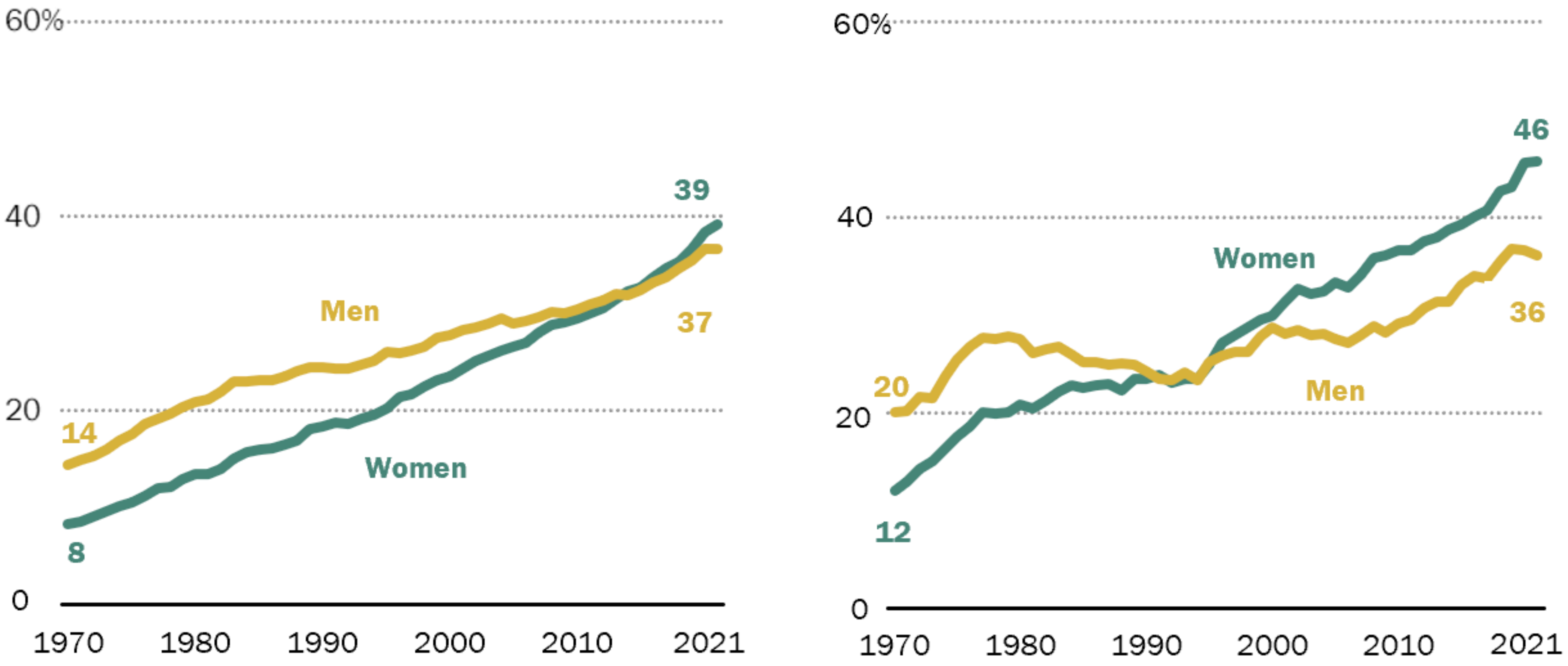
Motivating problem

Algorithmic discrimination

- EU Equality Directive, CHAPTER I, Article 2 (b):
  - indirect discrimination [...] would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons, *unless that provision, criterion or practice is objectively **justified*** by a legitimate aim and the means of achieving that aim are appropriate and necessary.
- US civil rights law:
  - If the evidence establishes a prima facie case of adverse disparate impact [...] courts then determine whether the recipient has articulated a "*substantial legitimate **justification***" [...]

Discrimination: “Direct” vs “indirect”

Women in the U.S. are outpacing men in college graduation



*% of adults **ages 25 and older** with a bachelor’s degree*

*% of adults **ages 25 to 34** with a bachelor’s degree*

Source: Pew Research Center analysis of Current Population Survey Annual Social and Economic Supplement (IPUMS).

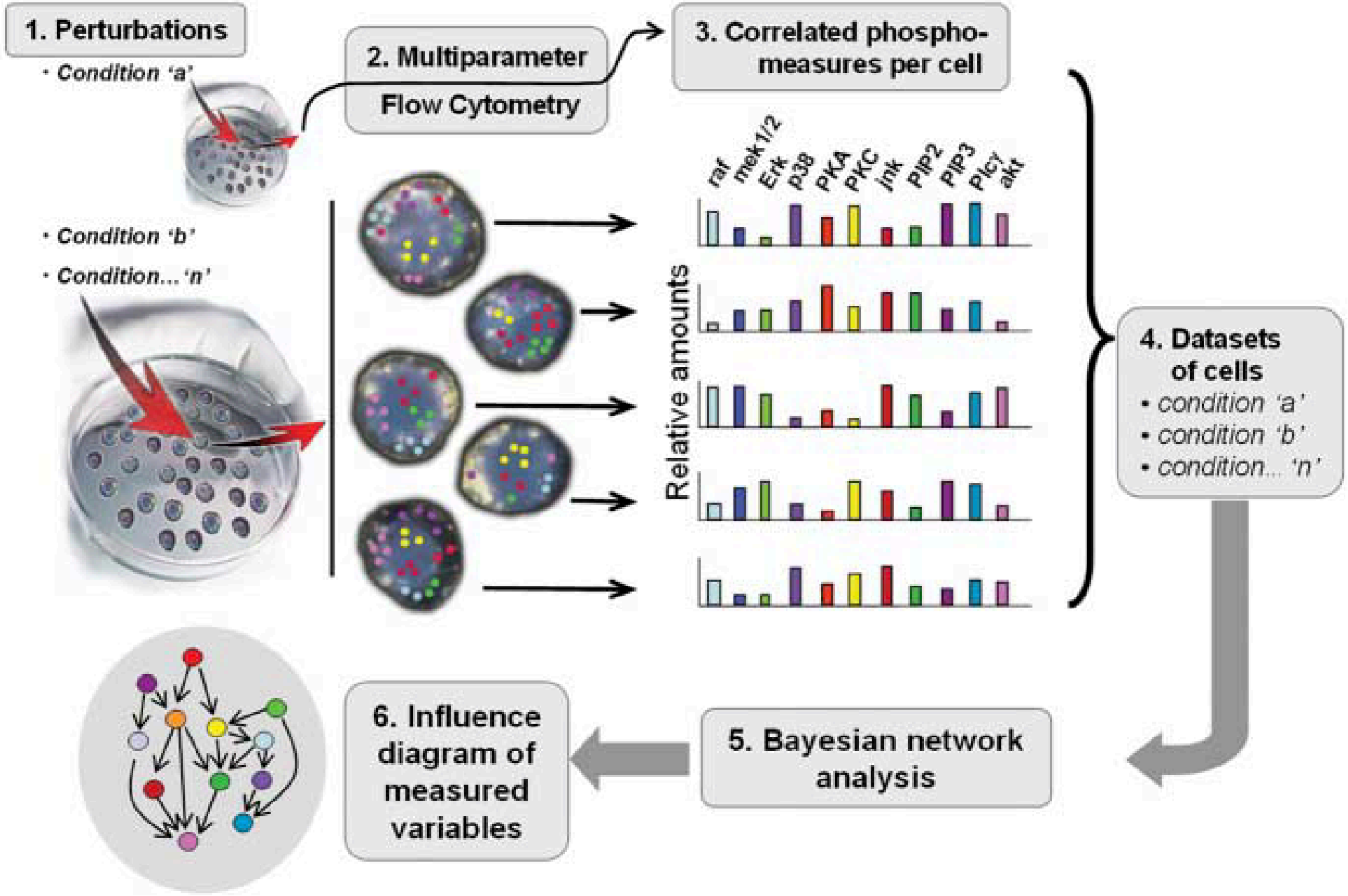
PEW RESEARCH CENTER

Motivating problem

Scientific machine learning (perhaps semi-supervised)

- Use ML to predict Y from $\mathbf{X}$, obtain predictive model $\hat{f}(\mathbf{X})$
- *Hopefully* learn about real world relationships by interpreting $\hat{f}(\mathbf{X})$
  - e.g. how does $\hat{f}(\mathbf{X})$ depend on X_j (one specific variable or “FoI”)
- Optionally(?) use background/domain knowledge

Causal knowledge about $\mathbf{X}$



From *Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data*

Sachs et al (2005)

Why should we care?

Explanation tools have real world impacts

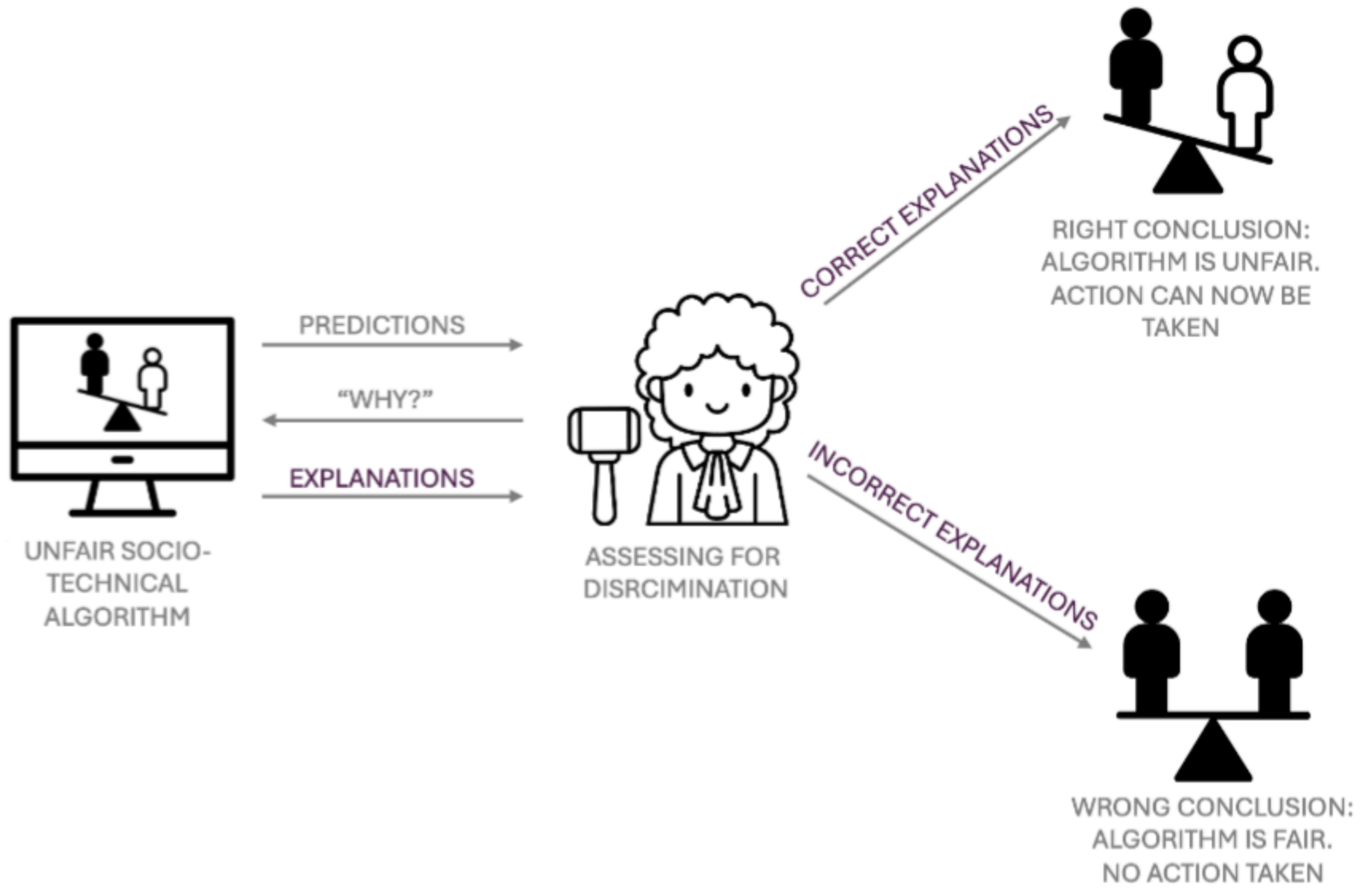

PREDICTIONS
"WHY?"
EXPLANATIONS
UNFAIR SOCIO-TECHNICAL ALGORITHM
ASSESSING FOR DISRCIMINATION
CORRECT EXPLANATIONS
RIGHT CONCLUSION: ALGORITHM IS UNFAIR. ACTION CAN NOW BE TAKEN
INCORRECT EXPLANATIONS
WRONG CONCLUSION: ALGORITHM IS FAIR. NO ACTION TAKEN

Fig. 3. An auditor receiving flawed explanations from an xAI tool may fail to detect unjust discrimination.

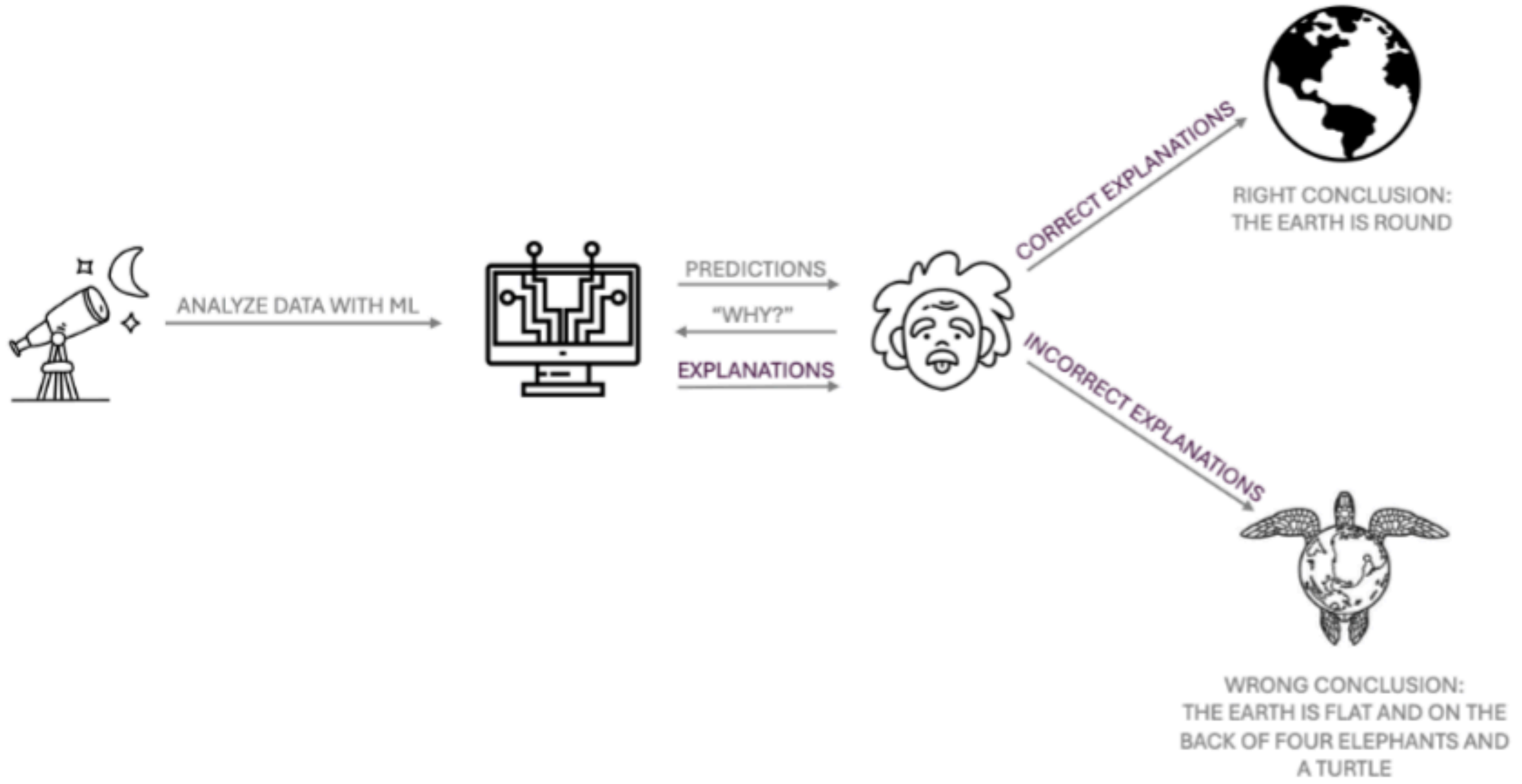
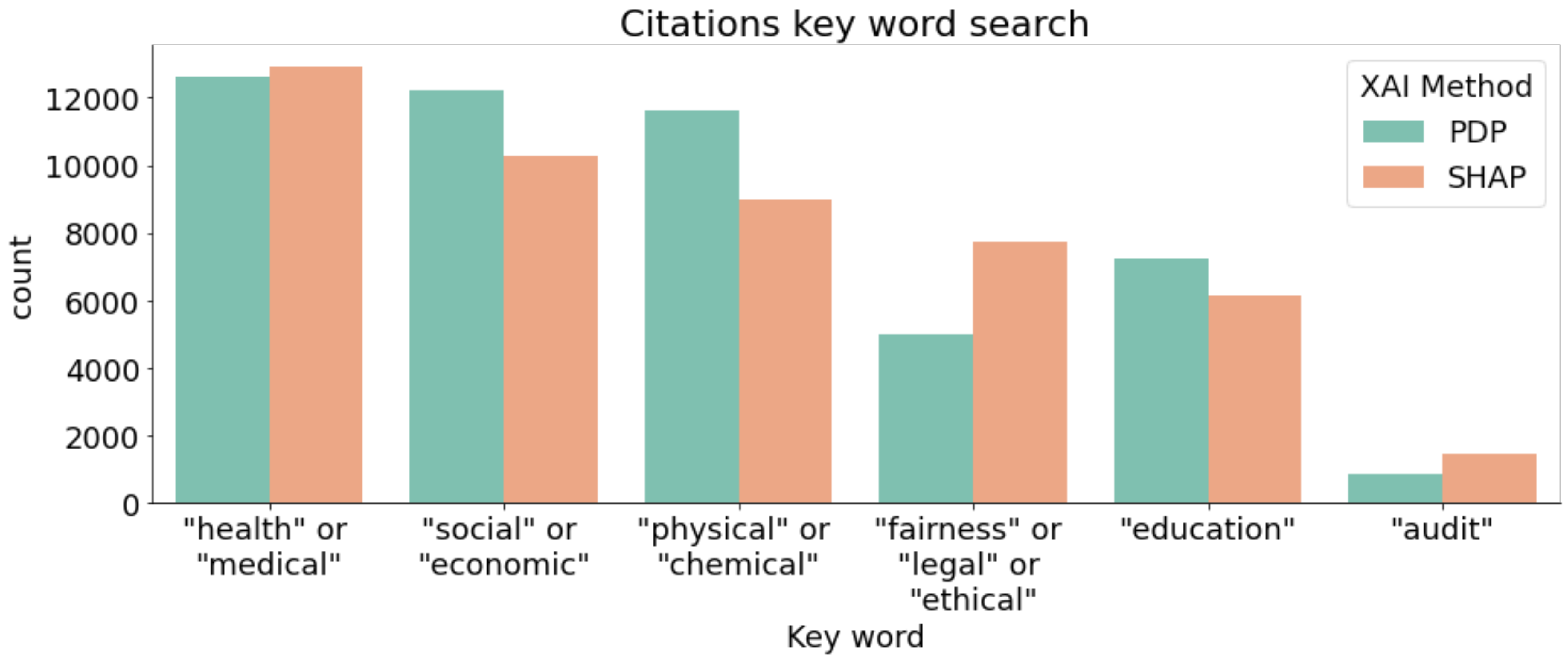
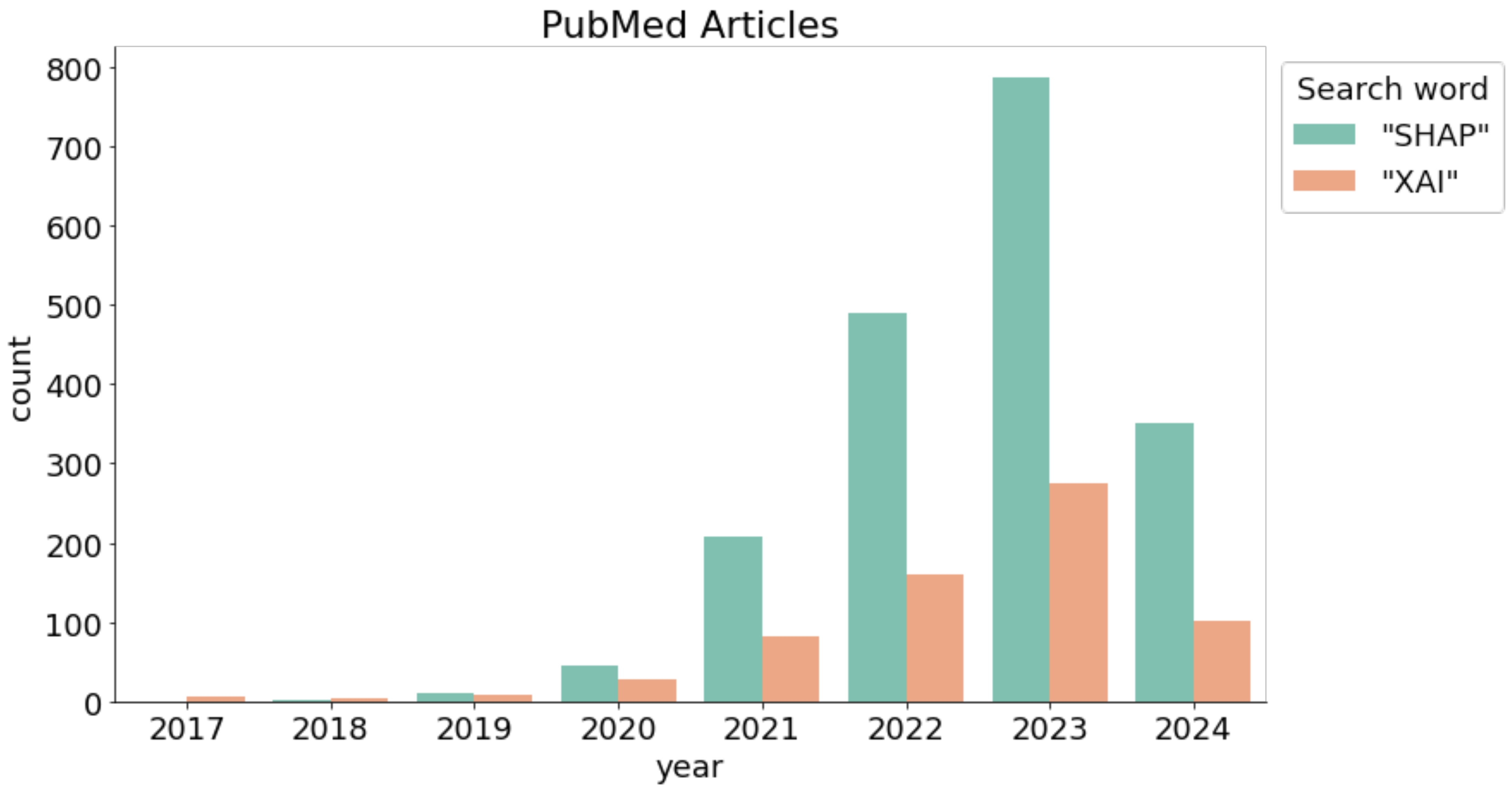

ANALYZE DATA WITH ML
PREDICTIONS
"WHY?"
EXPLANATIONS
CORRECT EXPLANATIONS
RIGHT CONCLUSION: THE EARTH IS ROUND
INCORRECT EXPLANATIONS
WRONG CONCLUSION: THE EARTH IS FLAT AND ON THE BACK OF FOUR ELEPHANTS AND A TURTLE

Fig. 4. A scientist receiving incorrect explanations from an xAI method could reach false conclusions.

Model-agnostic explanation tools are popular!





Google Scholar key word search results among papers citing PDP or SHAP

Feature dependence plots

Interpreting and explaining

Nature, models, and multi-variable questions

- Predictive model: $\hat{y}$ = f(x_1, x_2, …, x_p)
- For each predictor variable/feature x_j we may ask:
- What does this model *do* with x_j?
  - Regression (~1 century), partial dependence plots (~1/4 c.), …
  - “**Holding all other features constant**” / assuming independence between features
- How does this model *depend* on x_j?
  - Indirect discrimination, integrate (real world) dependence *between* features
  - New: causal dependence plots

Partial Dependence Plots

One of the most popular model-agnostic interpretation tools

- First described in Friedman, 2001. Citations > 26,000
- Continuously vary one feature while holding others constant, plot curve of predictions
- Zhao and Hastie, 2021: causal interpretation under some conditions
- Our work generalizes PDPs, containing them as a special case, establishing their general causal interpretation

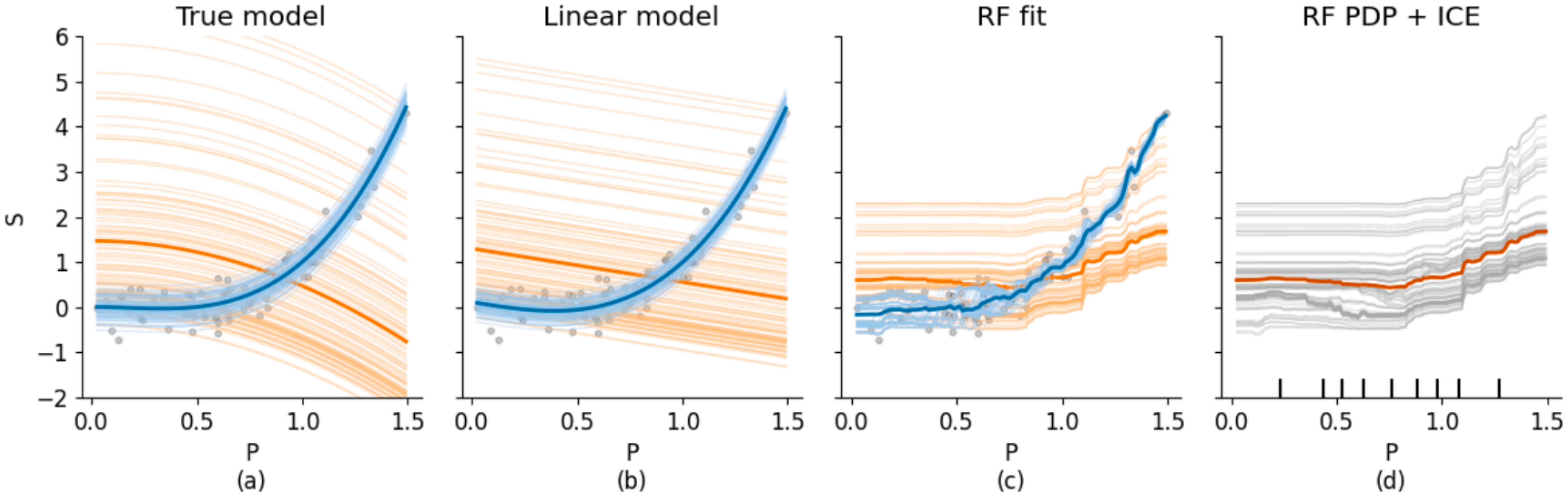
Causal Dependence Plots

Using an explanatory causal model (ECM)

Total Dependence: intervene on a predictor (P), use the ECM to change other predictors (S), then plot the new predictions

The PDP is identical to our NDDP showing that PDPs are a special case of CDPs!

Blue: Total Dependence
Orange: Natural Direct Dependence



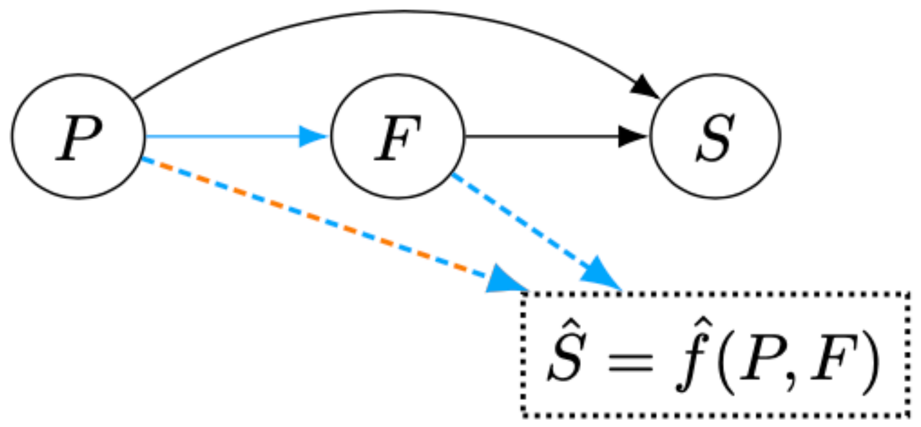
Counterfactual curves for individual points are shown as thin, light lines, with averages displayed as thick, dark lines

(a) True relationships

(b) Linear model. Note the *non-linear* Total Dependence

(c-d) Random forest model

(d) Partial Dependence Plot



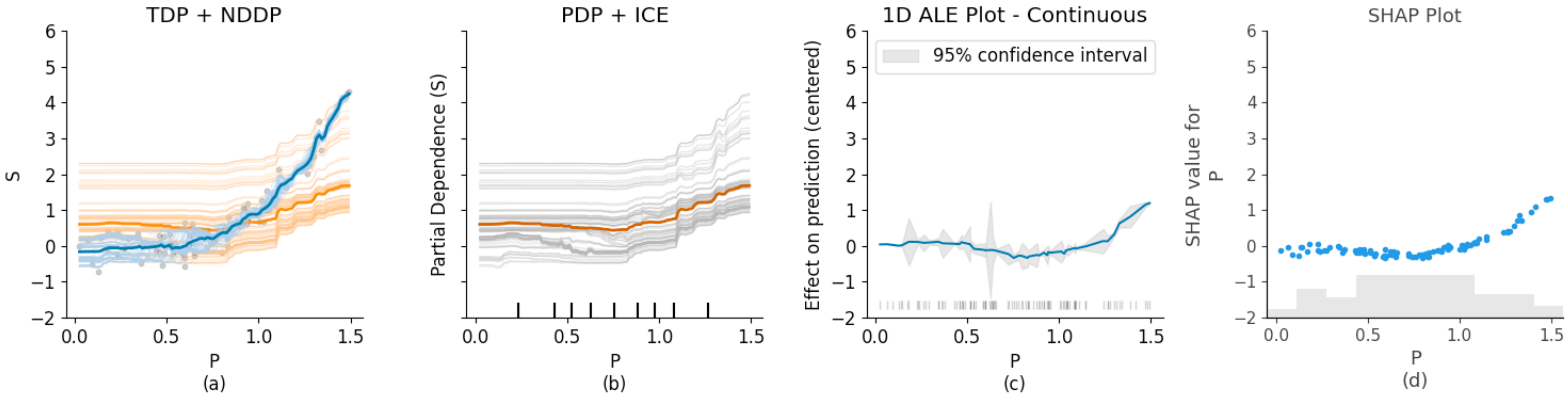
$$\begin{cases} P \sim \mathcal{U}[0, 1.5], \\ F = 2P^3 + \mathcal{N}(0, 0.2^2), \\ S = F - P^2 + \mathcal{N}(0, 0.2^2) \\ \hat{S} = \hat{f}(P, F) \end{cases}$$

ECM

Theorem:
PDP (+ ICE) = NDDP

Valid causal interpretation of PDPs

Other comparisons



(a) (b) (c) (d)

Total effect appears attenuated

both by marginalizing (global: / PDP)

and by conditioning (local: ALE / SHAP)

But…

How do we get an ECM?

Use domain knowledge

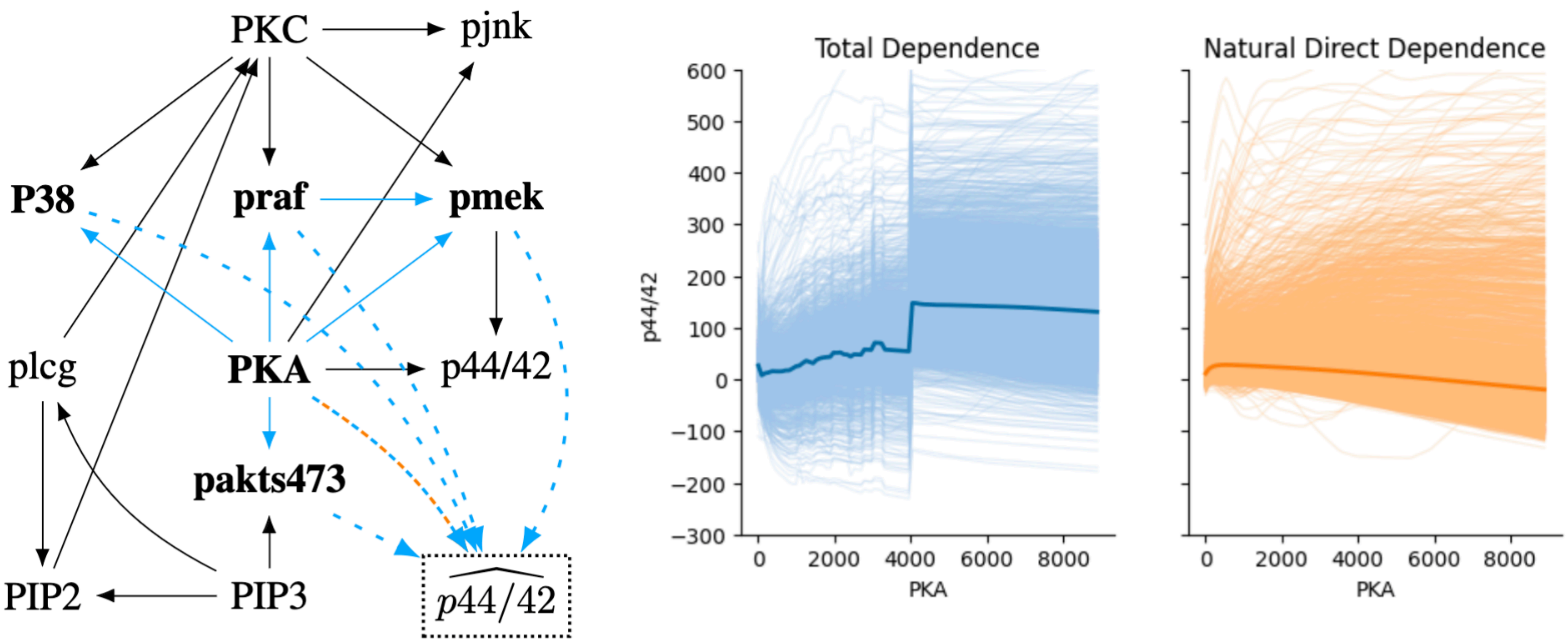
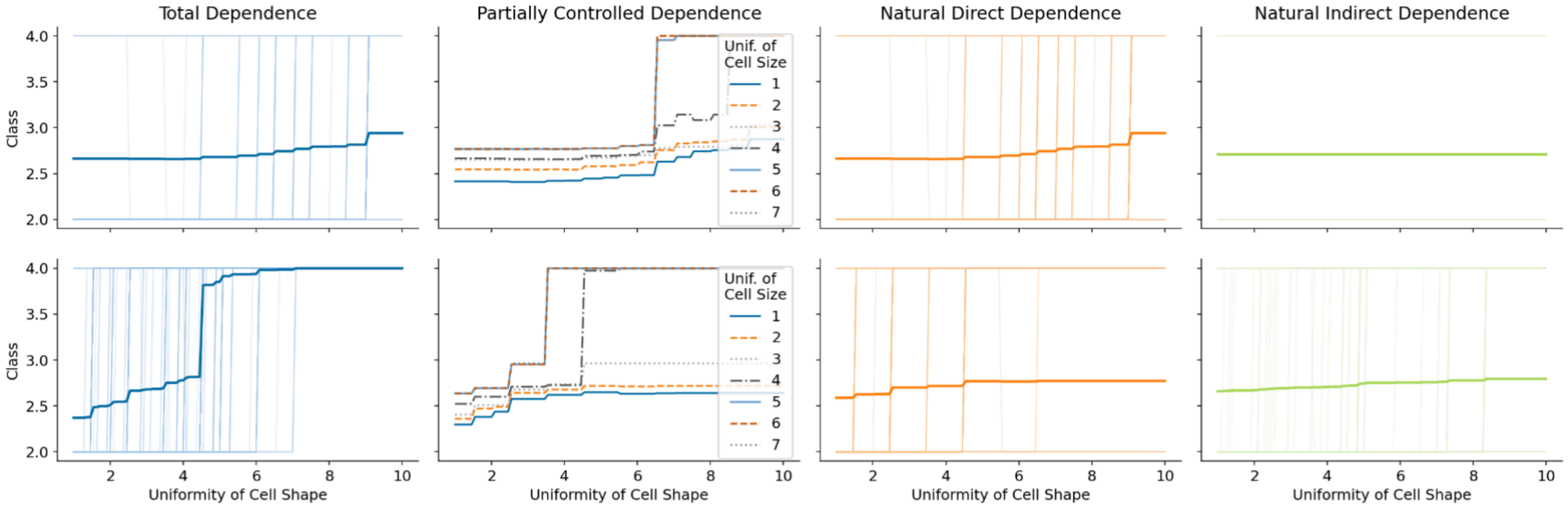
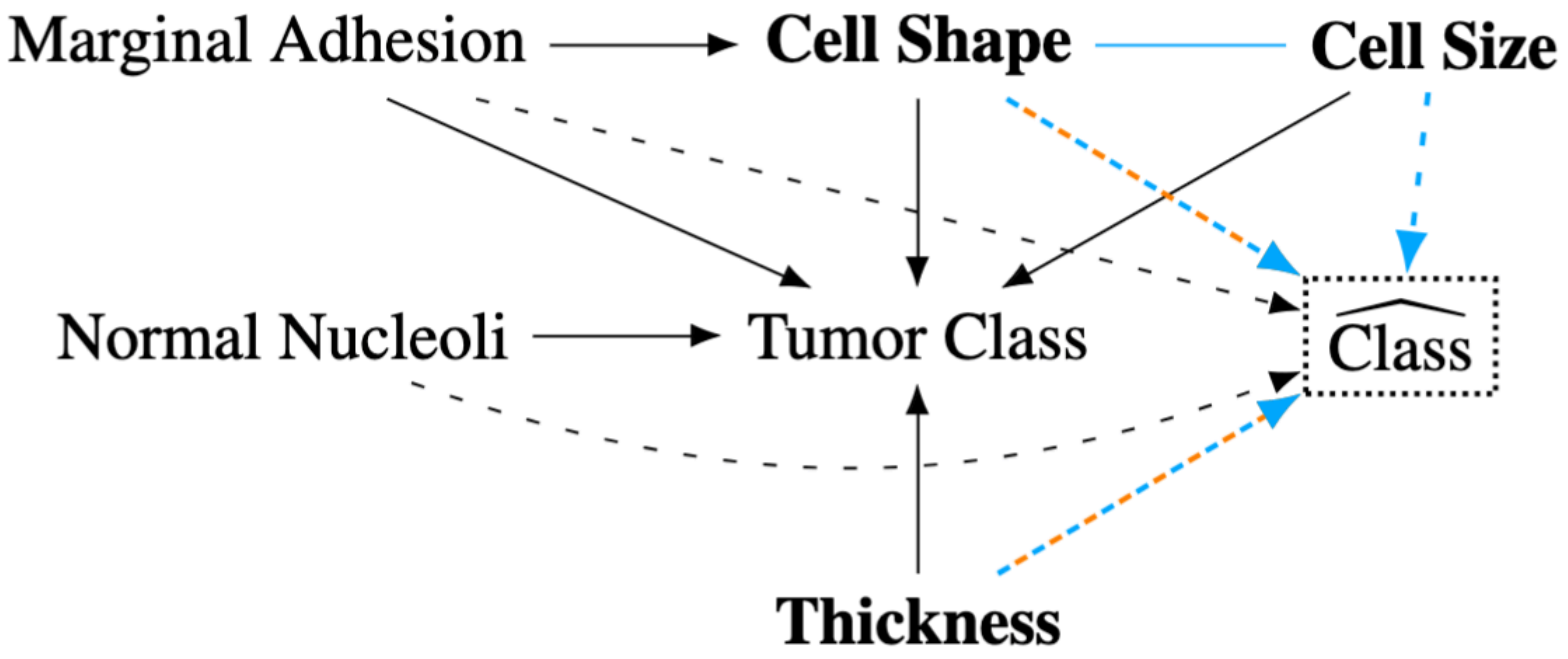


Figure 5: ECM for the Sachs et al. [43] dataset and corresponding CDPs for the effect of PKA on predicted p44/42. PKA and its descendants are bolded. While the NDDP (i.e. PDP + ICE) shows an overall decrease, the TDP shows an increase. *Conclusions depend strongly, qualitatively, on the specific interpretive question we ask, and causal modeling allows us to formulate questions precisely*.

Learn from data

CDPs after causal discovery algorithms





Marginal Adhesion

Cell Shape

Cell Size

Normal Nucleoli

Tumor Class

$\widehat{\text{Class}}$

Thickness

Help!
I’m uncertain about the ECM

Visualize uncertainty

Plot the “envelope” of a set of ECMs

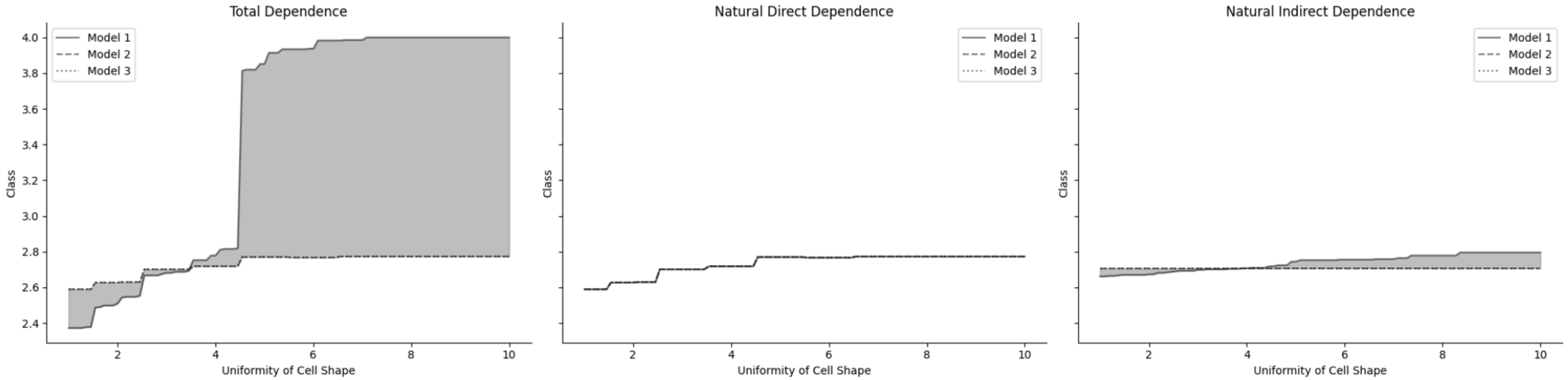


Figure 6: Total Dependence Plots, Natural Direct Dependence Plots and Natural Indirect Dependence Plots for the Breast Cancer Wisconsin dataset under three possible DAGs found by the PC algorithm: (1) $\mathcal{G}_B$ with the edge Cell Shape → Cell Size, (2) $\mathcal{G}_B$ with the edge Cell Size → Cell Shape, and (3) $\mathcal{G}_B$ with no edge between Cell Size and Cell Shape.

Limitations

(everybody to the limit)

Limitation of CDPs

Bad ECMs can lead to bad explanations

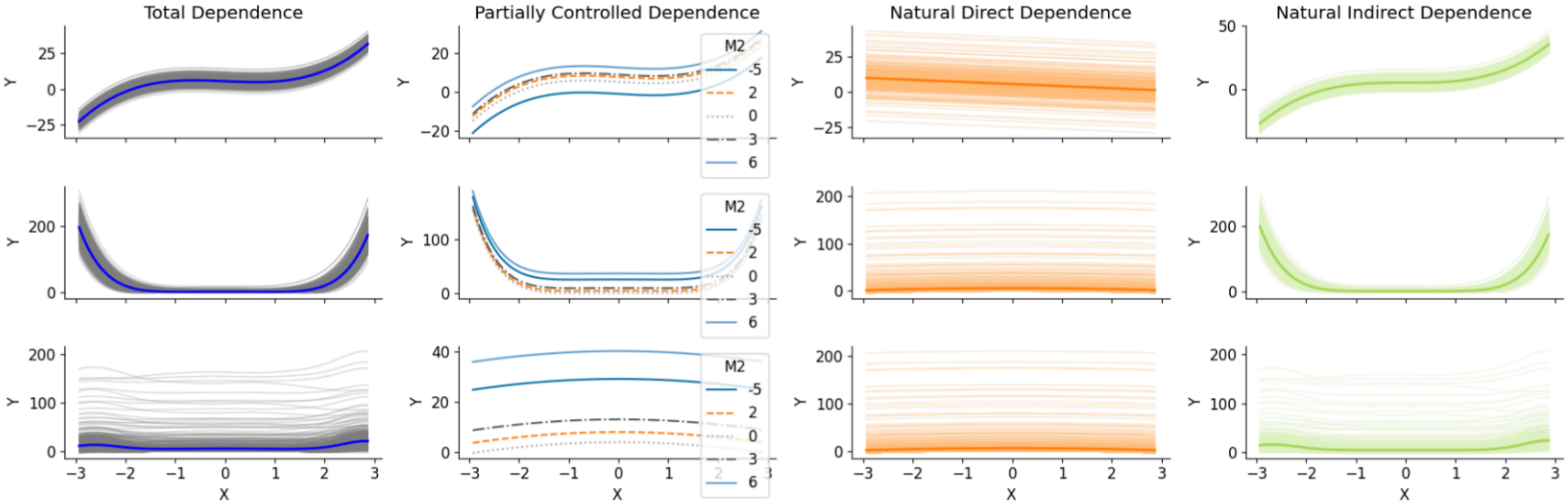


Figure 6: CDPs for the simulation example in Section B.1, shown for a ‘good’ black-box and correct ECM (top row), a ‘bad’ black-box model and correct ECM (middle row), and a ‘good’ black-box and misspecified ECM (bottom row).

Limitations of non-causal explanations

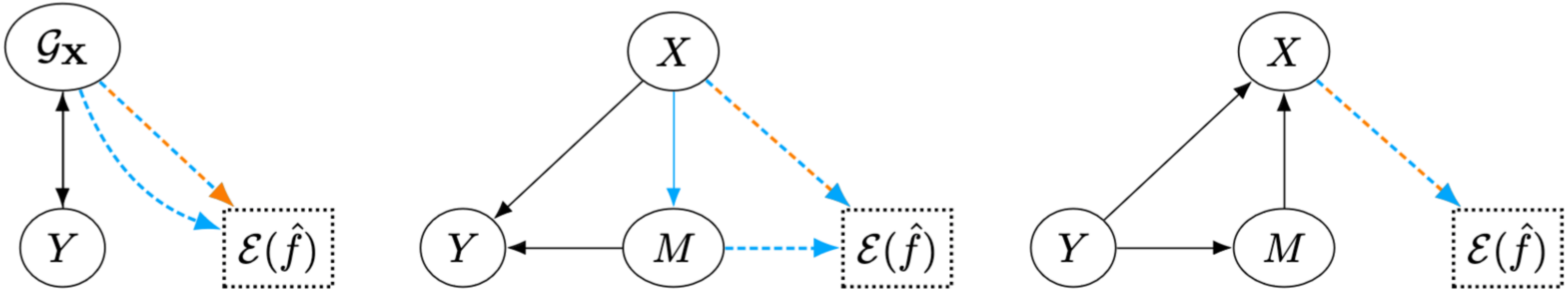
- Who care?
- They are wrong
- They are not even wrong
- Attenuation: more automatic methods, like SHAP and PDPs, usually only show “direct effects”
    - i.e. only detect “direct discrimination”
    - If model does not take S as an input, PDP / SHAP always show 0 effect

Limitations of model-agnostic explanations

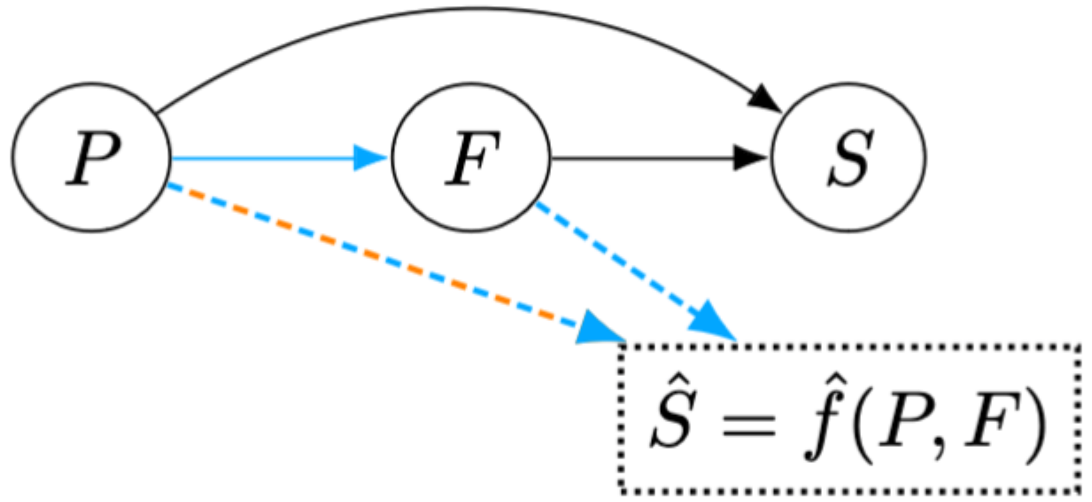
- Same model, different explanations
- Good explanations of a model may be different from good explanations of the world (!!!)
- This holds for CDPs as well
    - Causal interpretation of PDP: natural direct effect of X on $\hat{Y}$

A point worth repeating

Explaining a model is not the same as explaining the world



(a) (b) (c)



$$Y \neq \hat{Y} \text{ (!!!)}$$

Applications and future directions

- Easy to use software
- Visualize/infer uncertainty, relax assumptions
- Model-agnostic auditing or diagnostics:
  - Fairness, robustness, distribution shift, future, …
  - *Hypothesize* ECMs to ask specific questions!
- Human-guided exploration/explanation of large (e.g. “foundation”) models

“*All models are wrong, but some are useful*”
- George Box

Yule again:

> Measurement does not necessarily mean progress. Failing the possibility of **measuring that which you desire**, the lust for measurement may, for example, merely result in your **measuring something else** - and perhaps **forgetting the difference** - or in your **ignoring some things because they cannot be measured.**

Thanks for listening