Causal interpretability

for human-centered data science

Joshua Loftus

Roadmap

- Our current historical moment, why I care about “interpretability” (~10 min)
- Recent research projects (~35 min)
- Academic vision/philosophy for human-centered data science (~5 min)
- Q&A (~10 min)

Interpretability vs AI
An inflection point (crisis?)

AI Mathematics tutor (Bastani et al, 2024)

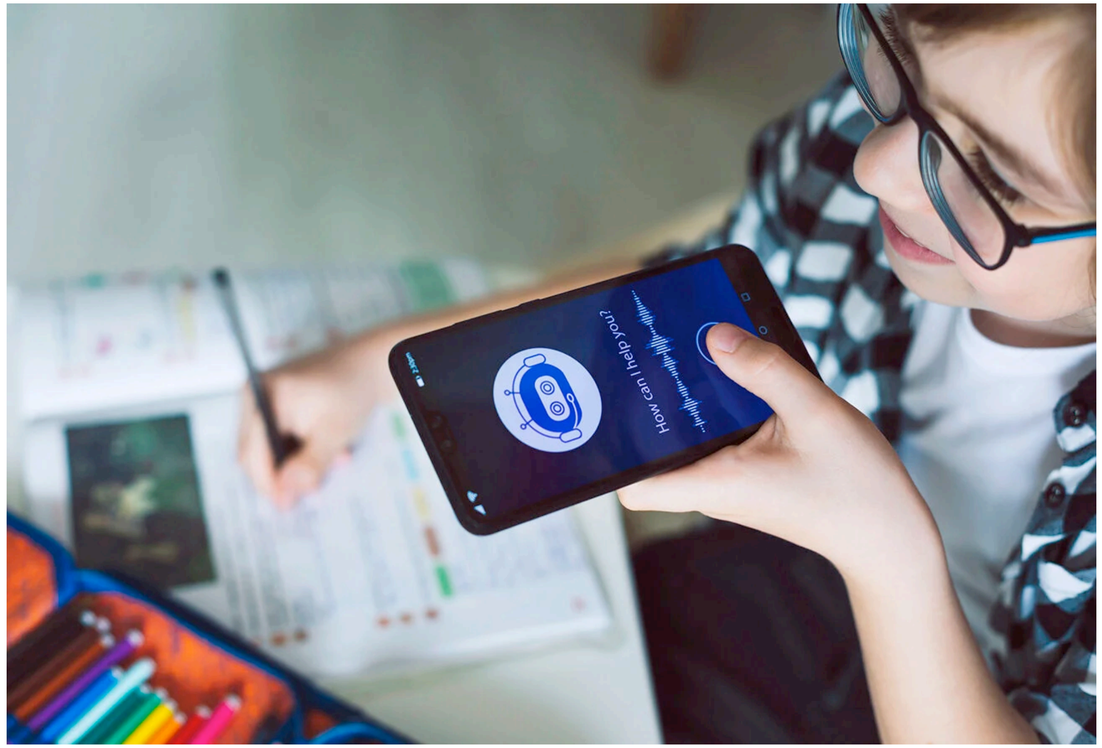
Classrooms randomized to three groups:

control (practice with books)

GPT-4 (standard)

GPT-Tutor (fine-tuned with lesson plan, to avoid giving solutions)

1. Lecture
2. Practice (with or without GPT assistance),
3. Exam (without any assistance).



Results

Source: Generative AI can harm learning (Bastani et al, 2024)

- Students with any "AI" did **better on practice questions**
- Students using **GPT did 17% worse on the exam** (practice sabotaged?)
- Students with **GPT-Tutor did no better on the exam**, but *thought that they had done better*

What’s the point?

- Of math or statistics education, educational technology, data science
- Do we just want a bigger, faster, or fancier model, even if *our own understanding* is left behind?
- What’s happening inside our heads?

[The] Imitation [game] is all you need

Feynman on **cargo cult science**:

“During the war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to imitate things like runways [...] But it doesn't work. No airplanes land. [...] I call these things cargo cult science, [...] they're missing something essential”

Causal understanding

See also: Cargo-cult statistics and scientific crisis (Stark and Saltelli, 2018)

“First we have an observation, then we have numbers that we measure, then we have a law which summarizes all the numbers. But the real *glory* of science is that *we can find a* ***way of thinking*** such that the law is *evident*.”

–Feynman lectures on physics, Vol. 1, 26-3

Model interpretability (or explainability)

Supervised machine learning

Context and notation

- Outcome variable Y, predicted outcome $\hat{Y}$
  - e.g. life expectancy; flood risk; asset risk premium
- Predictor variables $\mathbf{X}$ (possibly high-dimensional)
  - e.g. demographics and behaviors; topography; asset and macro indicators
- Model $\hat{f}(\mathbf{X}) = \hat{Y}$ optimized for predictive accuracy ($\hat{Y}$ is "close" to Y)
  - e.g. linear or additive models, SVM, random forests, neural networks

Interpretable ML (IML) / explainable AI (xAI)

Tools to understand how a given black-box model works

- Predict Y from $\mathbf{X}$ using $\hat{f}(\mathbf{X})$
- Which variables in $\mathbf{X}$ are “important or “contribute” to the prediction?
- How does $\hat{f}(\mathbf{X})$ “depend” on X_j (one specific variable or “F.o.I.”)
- How does an **evaluation metric** vary along some **interpretable dimension**?

Motivation

Why interpret/explain? Why not just use the predictions?

- Model diagnostics / engineering
  - Check that model does not have undesirable properties (e.g. violating regulations like anti-discrimination law)
  - Find issues that could be fixed and design a better model
- Scientific machine learning / causality
  - Does $\hat{Y} = \hat{f}(\mathbf{X})$'s dependence on X_j reveal how Y depends on X_j?
  - Are "important" predictors good candidates for change (intervention)?

Example method: SHAP

A unified approach to interpreting model predictions

SM Lundberg, SI Lee

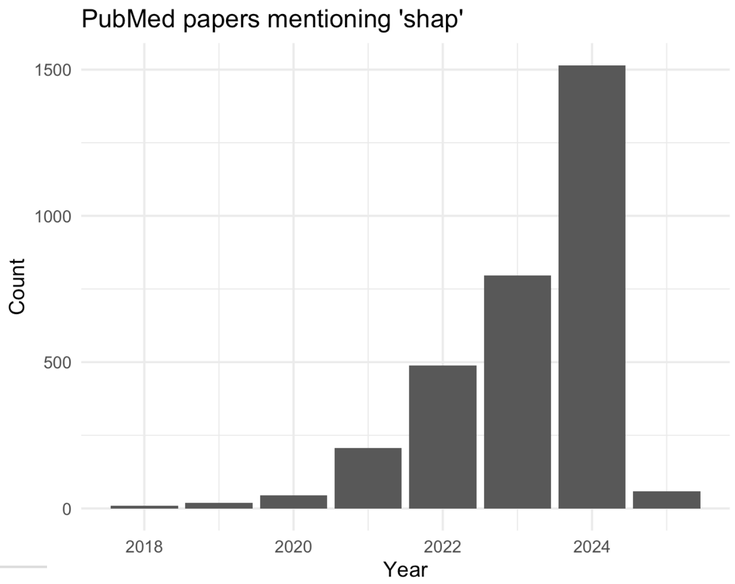
Advances in neural information processing systems, 2017 • proceedings.neurips.cc

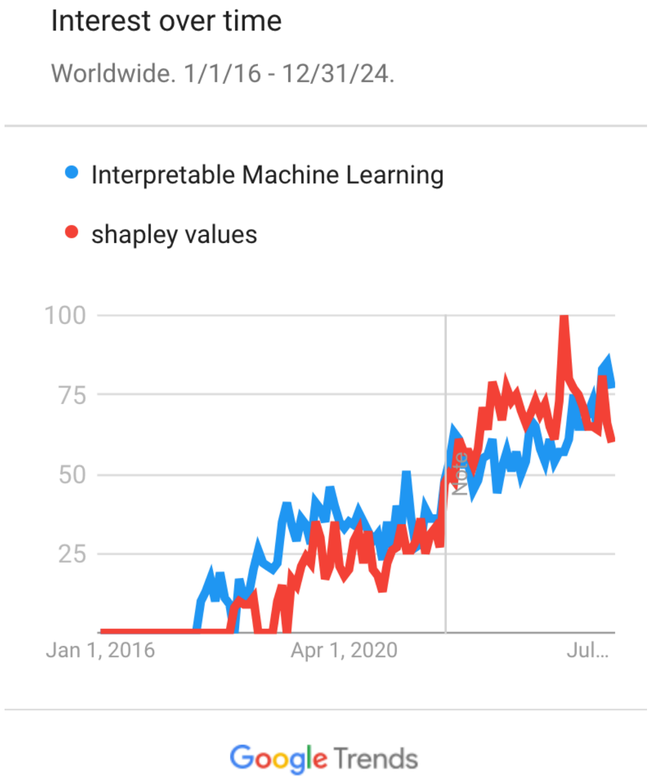
Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between accuracy and interpretability. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these

SHOW MORE

☆ Save Cite Cited by 29767 Related articles All 22 versions





Example application: 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***" [...]

“Justifying” discrimination

- Association between sensitive attribute A and outcome Y
- Is this association “explained” by some other variable(s) $\mathbf{X}$?
- History: multivariate analysis, Udny Yule and “poor laws”
  - Association between welfare and poverty
  - Is this association “explained” by age?
  - Does poverty cause welfare, welfare cause poverty, or age cause both?

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**.”

Yule's unheeded warning

Lots of science is junk

- People (almost always?) interpret regression coefficients as causal effects
- Haber et al. (2022): study of non-RCTs in "1,170 articles from 18 high-profile medical/public health/epidemiology journals (65 per journal) published from 2010–2019"
- "few studies explicitly declared an interest in estimating causal effects"
- "the majority used language that moderately or strongly implied causality"
- "action recommendations were identified in 60.3% [...] of discussion sections, about twice that in abstracts"

The future of science?

~~**Misinterpreted regression coefficients**~~

Misinterpreted ML explanations

Can we do better?

Feature dependence plots

Paper: Causal Dependence Plots (NeurIPS 2024)

Lucius Bynum (NYU) **Sakina Hansen (LSE)**

Baselines: PDP and ICE

Partial Dependence Plot and Individual Conditional Expectation

- Plot “dependence” of $\hat{f}(x_1, x_2, \ldots, x_p)$ on x_j
- Horizontal axis is a grid of possible values of x_j
- At each possible value x^*, and for each observation $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$
  - Compute and store $\hat{f}(x_{i1}, \ldots, x_j \leftarrow x^*, \ldots, x_{ip})$
- For each observation plot the ICE curve $x^* \mapsto \hat{f}(x_{i1}, \ldots, x^*, \ldots, x_{ip})$
- Average ICE curves over sample to get PDP curve

Simple additive example

- Additive model $\hat{f}(\mathbf{x}) = \hat{f}_1(x_1) + \hat{f}_2(x_2) + \cdots + \hat{f}_p(x_p)$
- Superficial/SOTA dependence on x_1: plot the univariate function $\hat{f}_1(\,\cdot\,)$
- Problem: suppose, for example, we know that x_2 is influenced by x_1

  [e.g. if x_1 is age and x_2 is experience, or e.g. if $x_2 = \text{poly}(x_1)$]

  … then what?
- Solution: deeper/causal notion of dependence

“Model-agnostic” explanation tools

Feature dependence plots

- Partial Dependence Plots (**PDP**) and Individual Conditional Expectation (**ICE**). ICE citations > 1,600 since 2015
- Accumulated Local Effects (**ALE**) plots. Citations > 1,300 since 2016.
- SHapley Additive exPlanations (**SHAP**). Citations > 29,000 since 2017.
- Causal Dependence Plots (**CDP**), current work (NeurIPS 2024)

“Model-agnostic” explanation tools

Feature dependence plots

- Partial Dependence Plots (**PDP**) and Individual Conditional Expectation (**ICE**). ICE citations > 1,600 since 2015
- Accumulated Local Effects (**ALE**) plots. Citations > 1,300 since 2016.
- SHapley Additive exPlanations (**SHAP**). Citations > 29,000 since 2017.
- Causal Dependence Plots (**CDP**), current work (NeurIPS 2024)

Do they work?

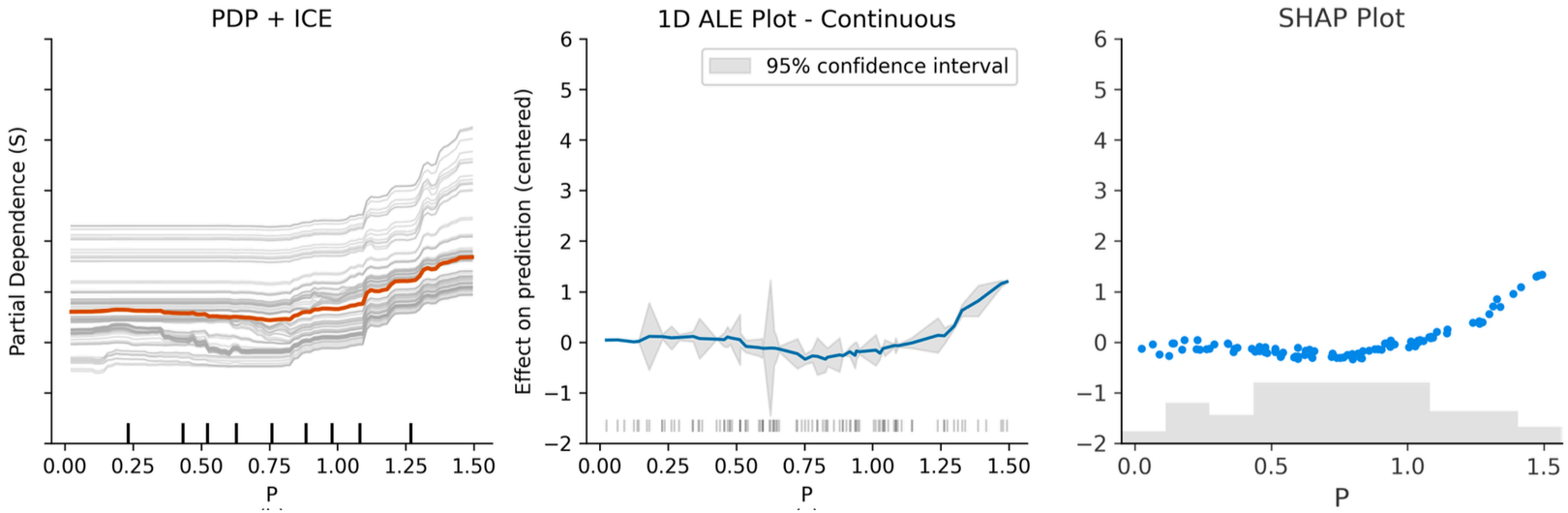
e.g. if an external auditor used them to check for discrimination?

Salary example (simulation)

Parental income P

School funding F

Graduates' salaries S



Fit a random forest (RF) predicting S

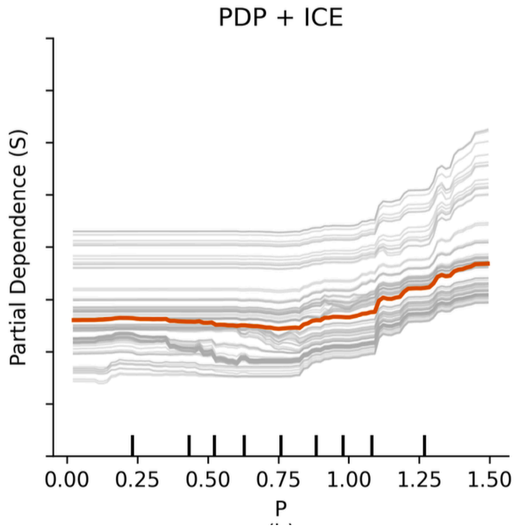
Check model's dependence on P

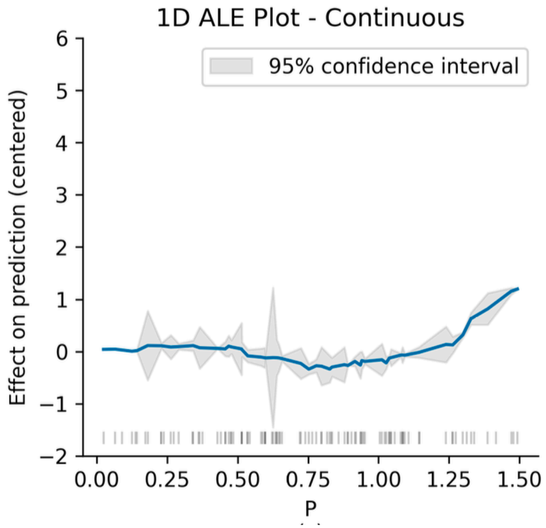
Salary example (simulation)

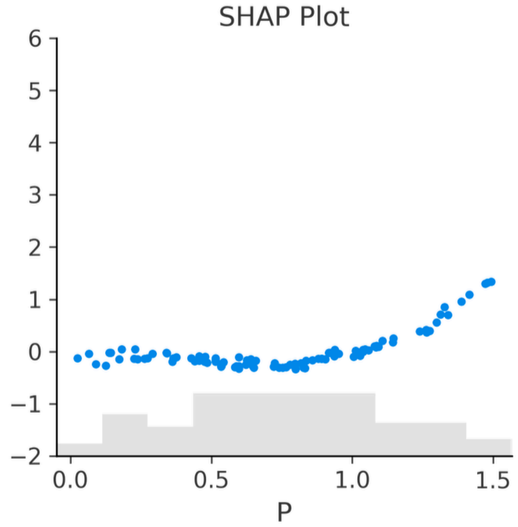
Parental income P

School funding F

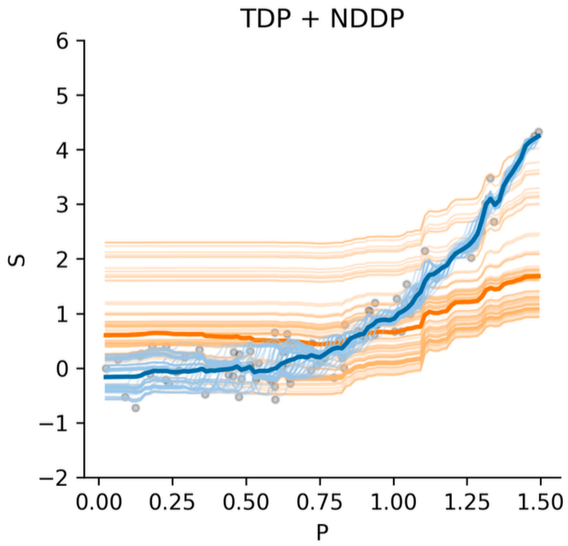
Graduates' salaries S

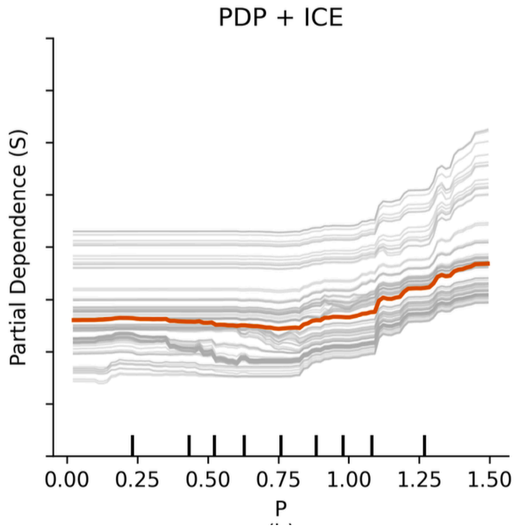


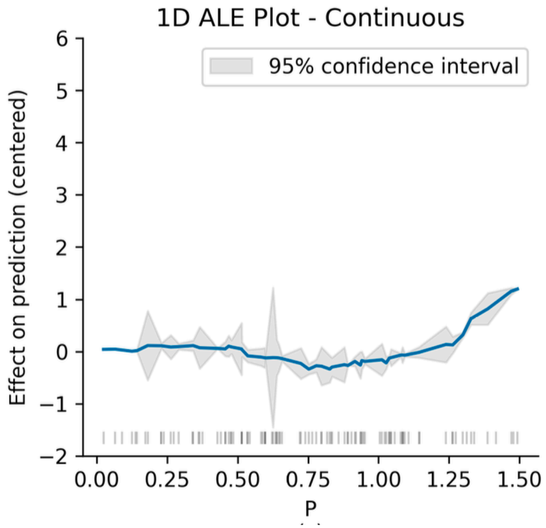


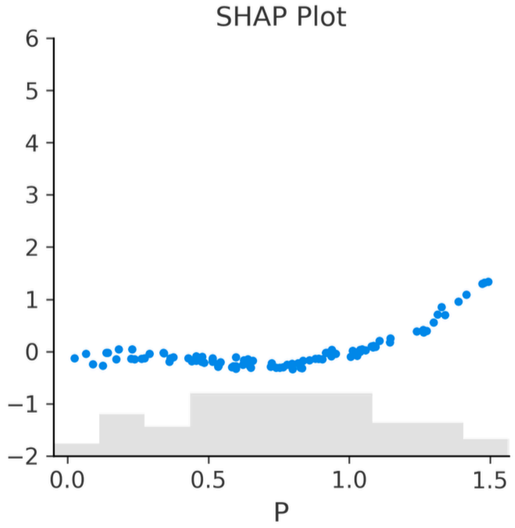


Salary example (simulation)

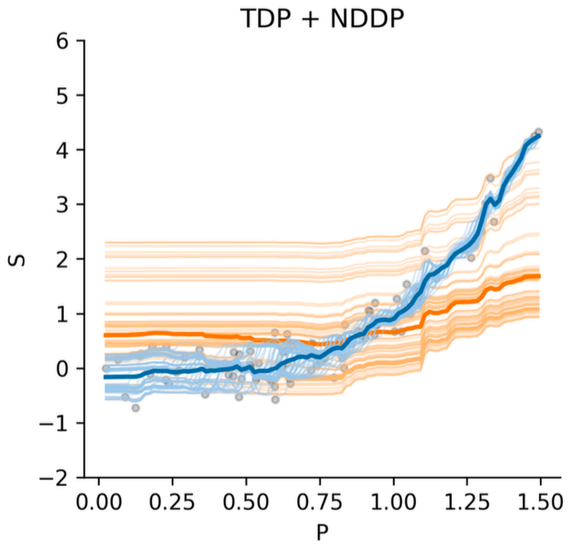




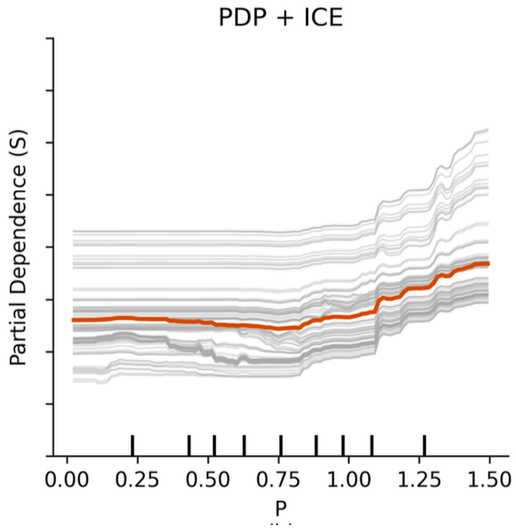


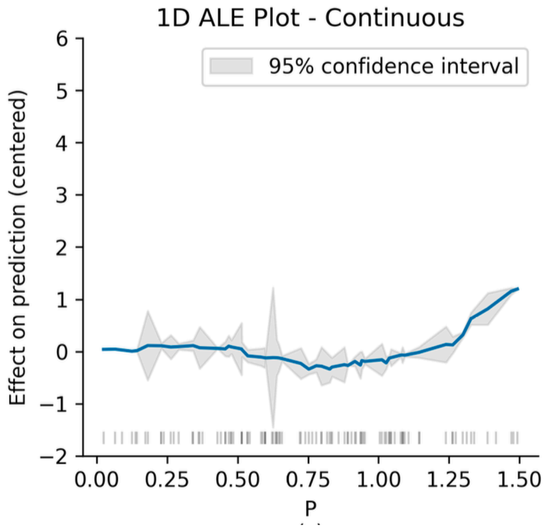


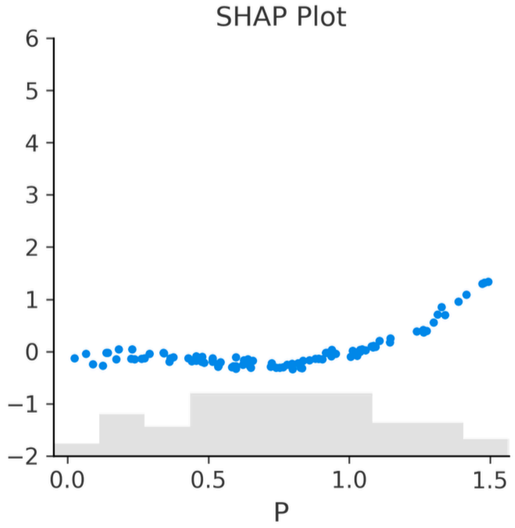
Salary example (simulation)



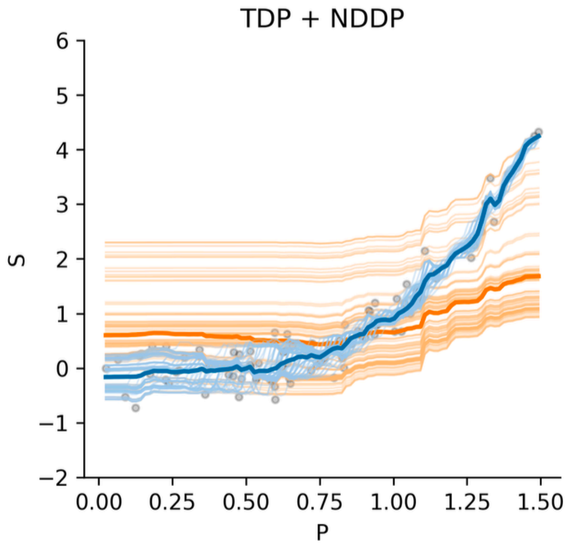
CDP

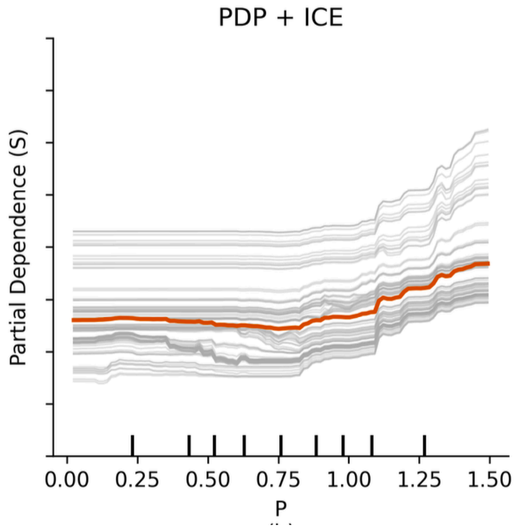


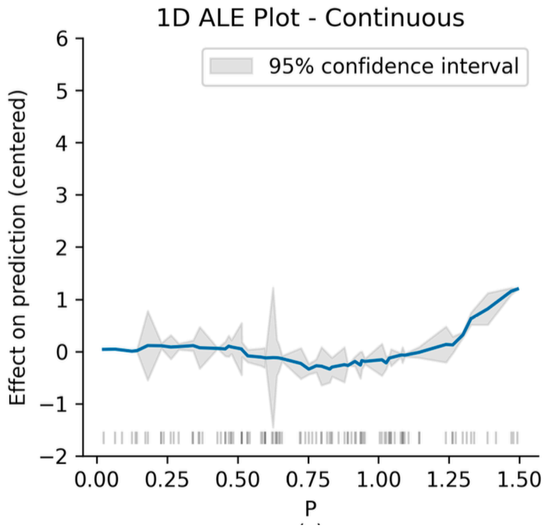


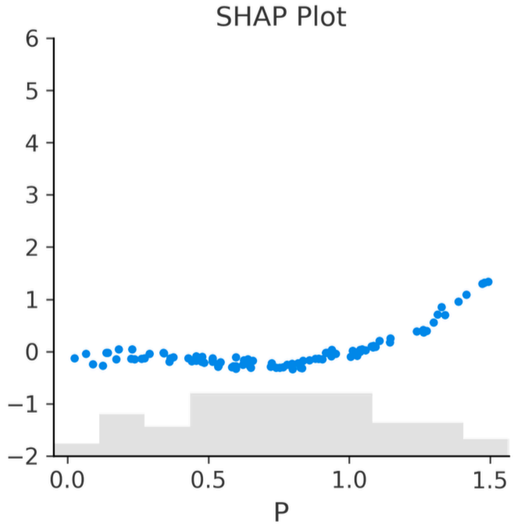


Salary example (simulation)





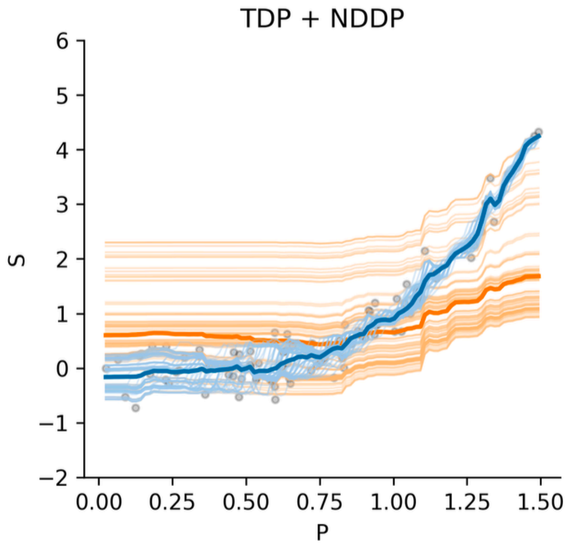


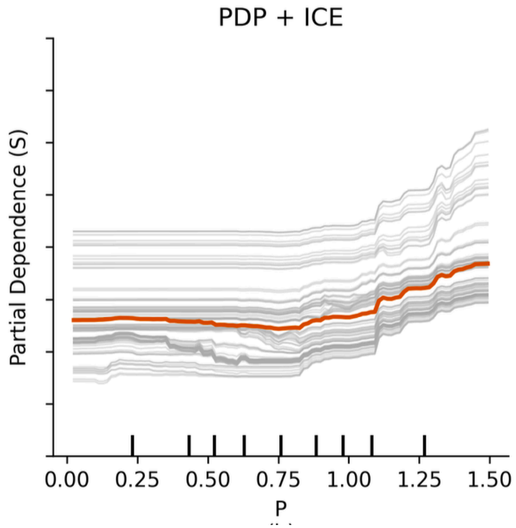


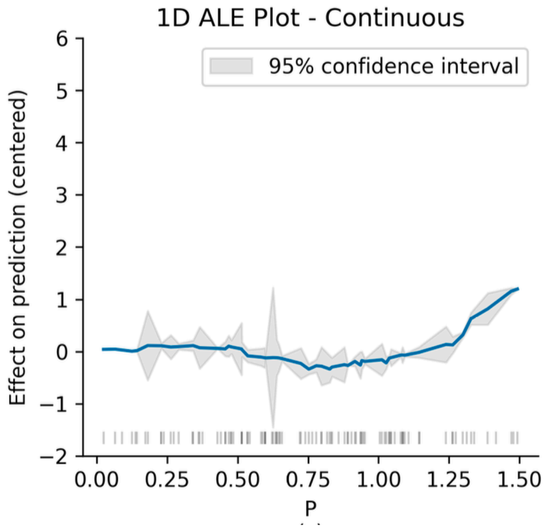
CDP

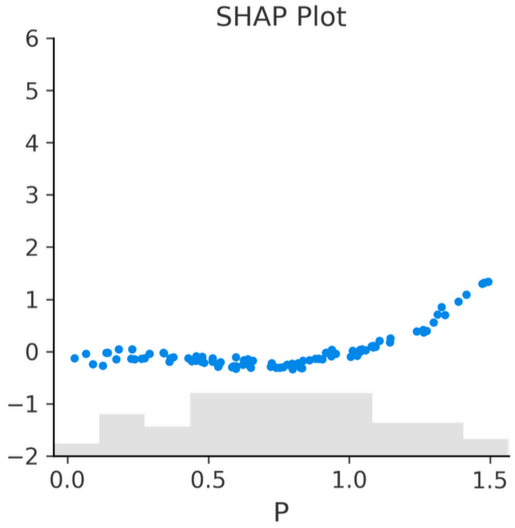
So which of these is a good/correct explanation?

Salary example (simulation)









CDP

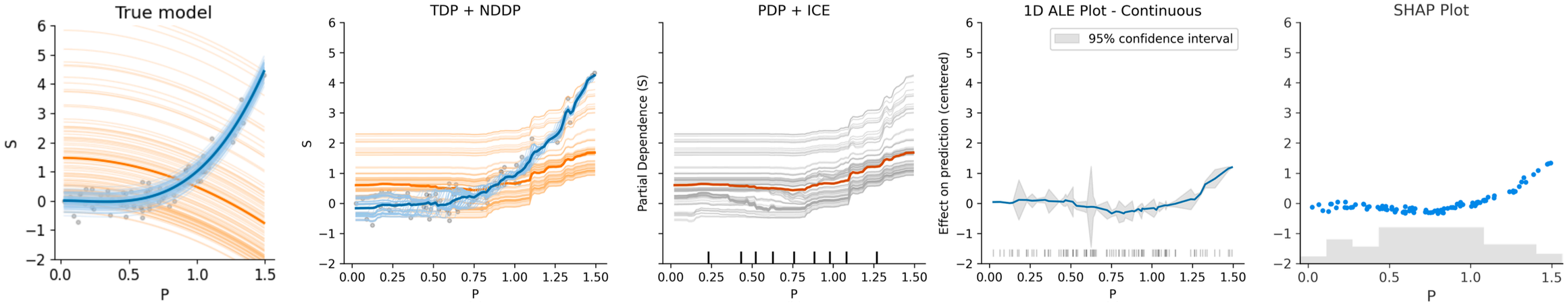
So which of these is a good/correct explanation?

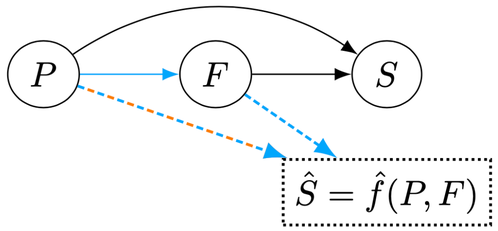
Causal Dependence Plots

Using an explanatory causal model (ECM)

Blue: Total Dependence
Orange: Natural Direct Dependence

CDP





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

ECM

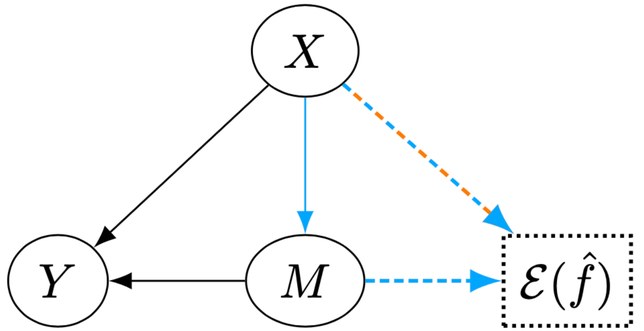
Points: training data

Thin lines: counterfactual curves for individual points

Thick, dark lines: empirical averages

Zhao and Hastie, 2021:

causal interpretation *under some conditions*

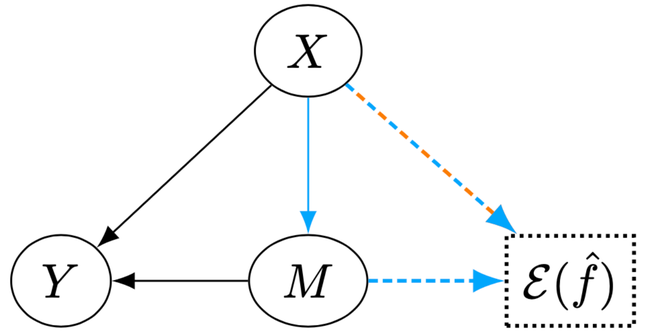


Theorem:
PDP (+ ICE) = NDDP

Universally valid causal interpretation of PDPs

Zhao and Hastie, 2021:

causal interpretation *under some conditions*



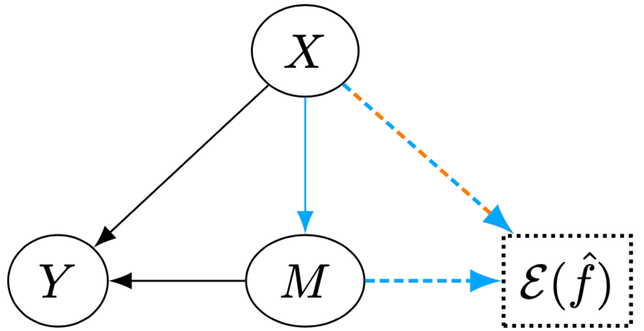
Theorem:
PDP (+ ICE) = NDDP

Universally valid causal interpretation of PDPs

Direct effects are agnostic to causal structure…

Zhao and Hastie, 2021:

causal interpretation *under some conditions*



Theorem:
PDP (+ ICE) = NDDP

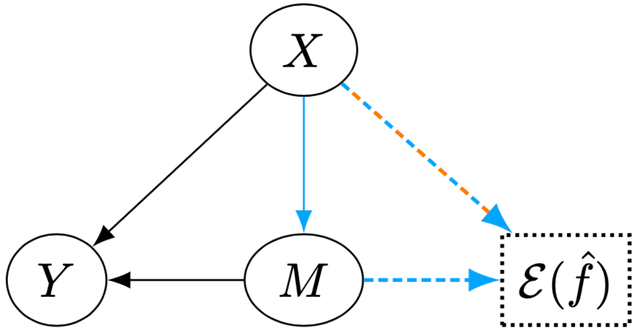
Universally valid causal interpretation of PDPs

Direct effects are agnostic to causal structure…

Wanted: IML/xAI methods capable of showing more than only direct effects!

The CDP framework

- As we vary the plot axis variable $x_j \leftarrow x^*$
- Instead of holding other predictors fixed (assuming independence)
- We assume a causal model (ECM)
- The ECM tells us how other predictors change in response to $x_j \leftarrow x^*$
- Various CDPs: Total, Direct, Indirect, Partially Controlled (see paper)



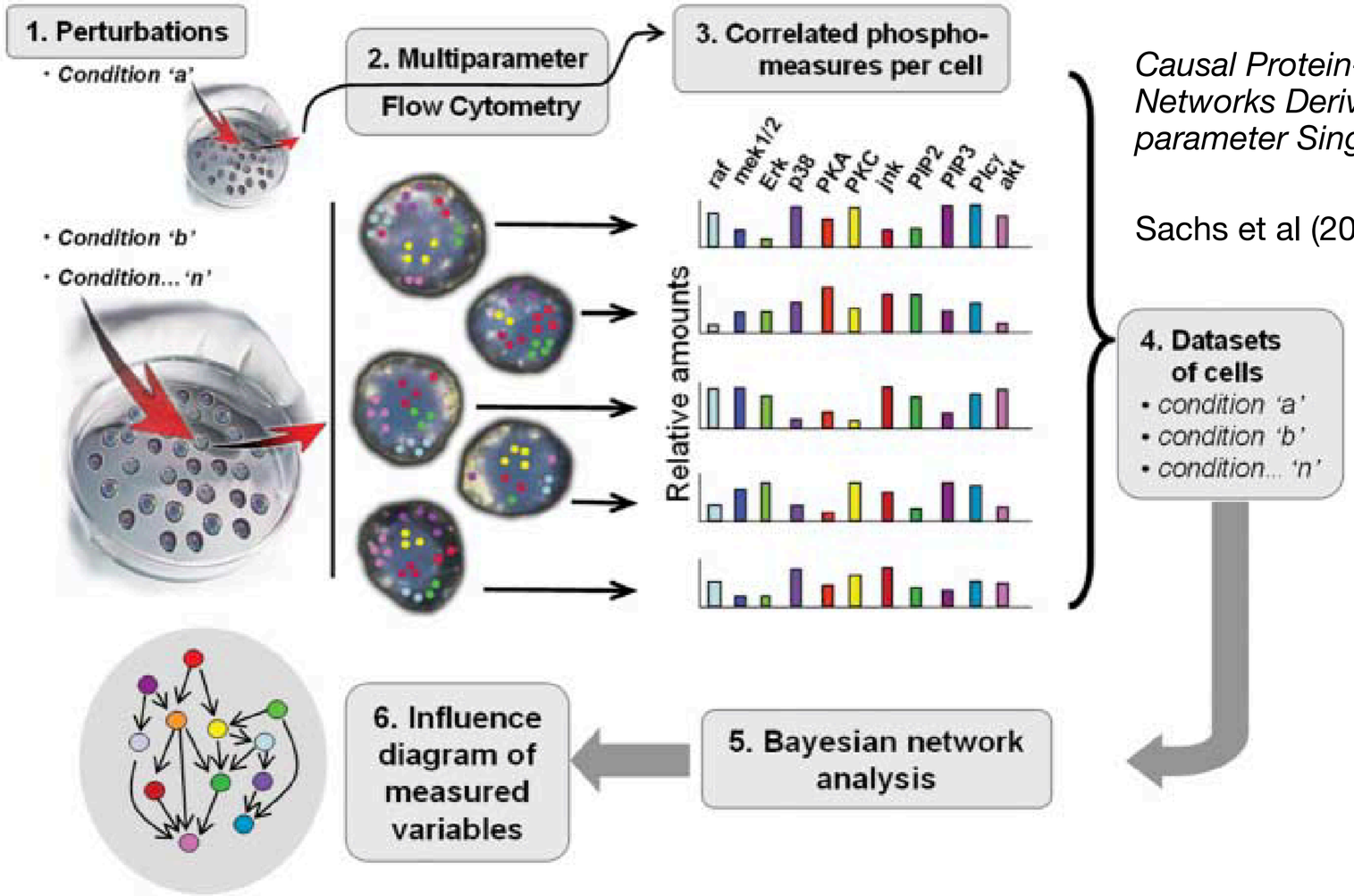
Broad idea: use an Explanatory Causal Model (ECM) as a bridge

Between output (or evals) and input (or controllable parameters)

Leverage causal ideas/methods: mediation, confounder sensitivity, etc.

How do we get an ECM?

Causal knowledge about X



Causal Protein-Signaling Networks Derived from Multi-parameter Single-Cell Data

Sachs et al (2005)

Supervenience: "low level" predictors of "higher level" outcomes

e.g. genomic predictors of phenotypes/health

Use domain knowledge

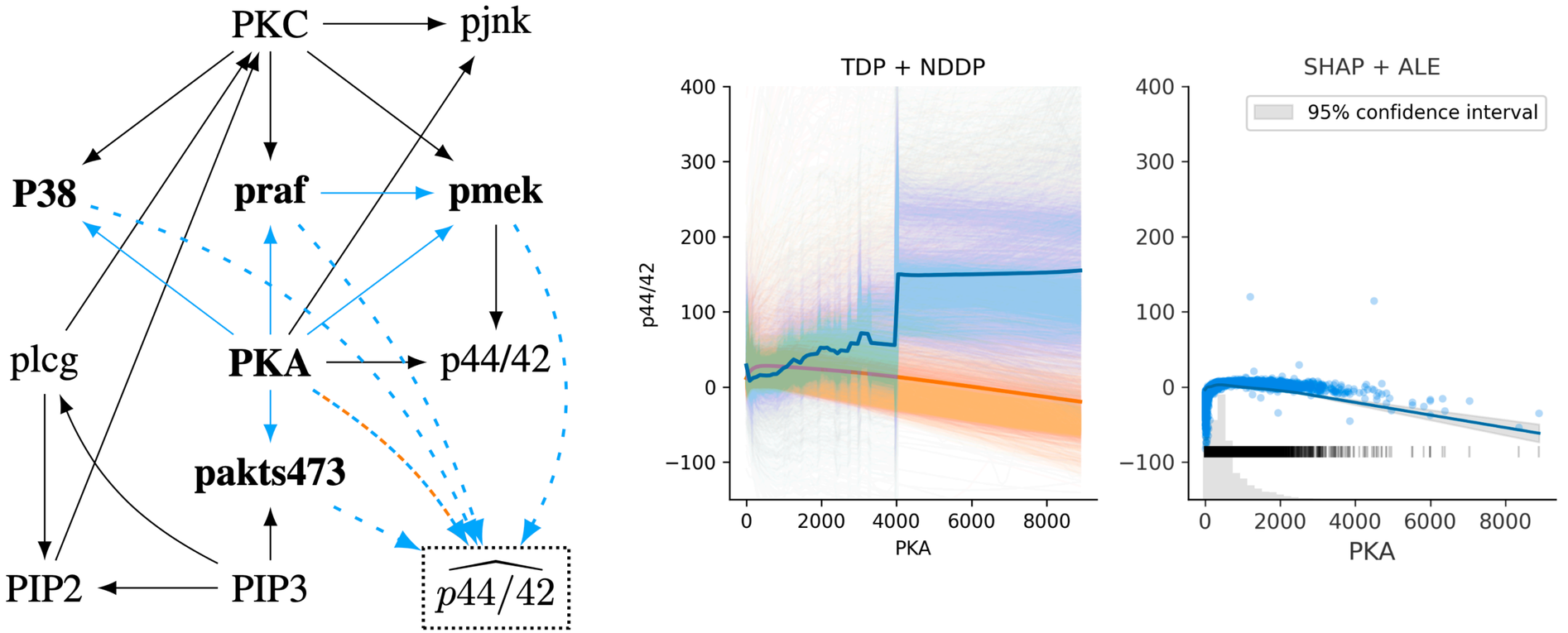
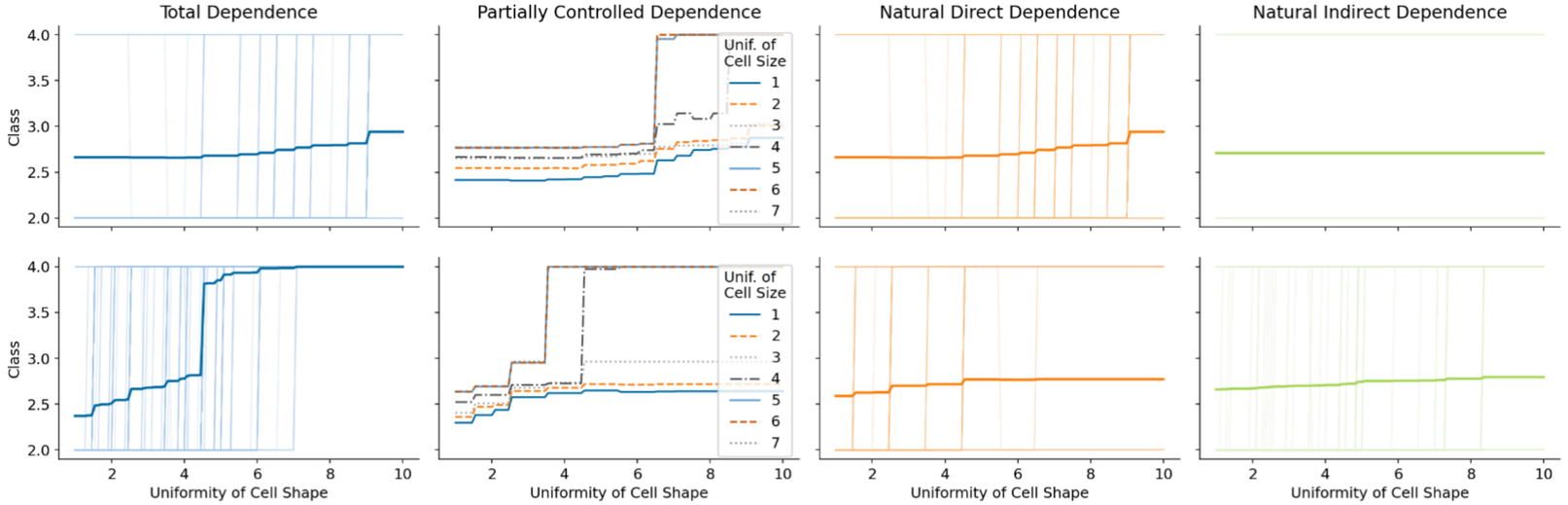
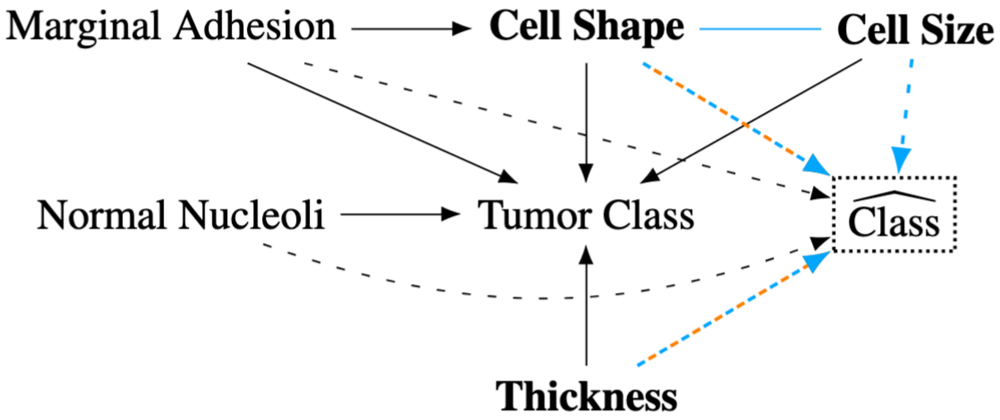


Figure 5: ECM for the Sachs et al. [49] dataset (left), CDPs for an MLP predictive model (center), ALE (line) and SHAP (points) plots (right). All plots visualize the effect of PKA on predicted p44/42. PKA and its descendants are bolded. The NDDP (i.e. PDP + ICE), ALE, and SHAP all show an overall decrease, while 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





Hypothesize an ECM

Model diagnostics or OOD generalization

- Exploratory analysis
- Distribution shift
- Stress testing models

"All models are wrong, but some are useful"
- George Box

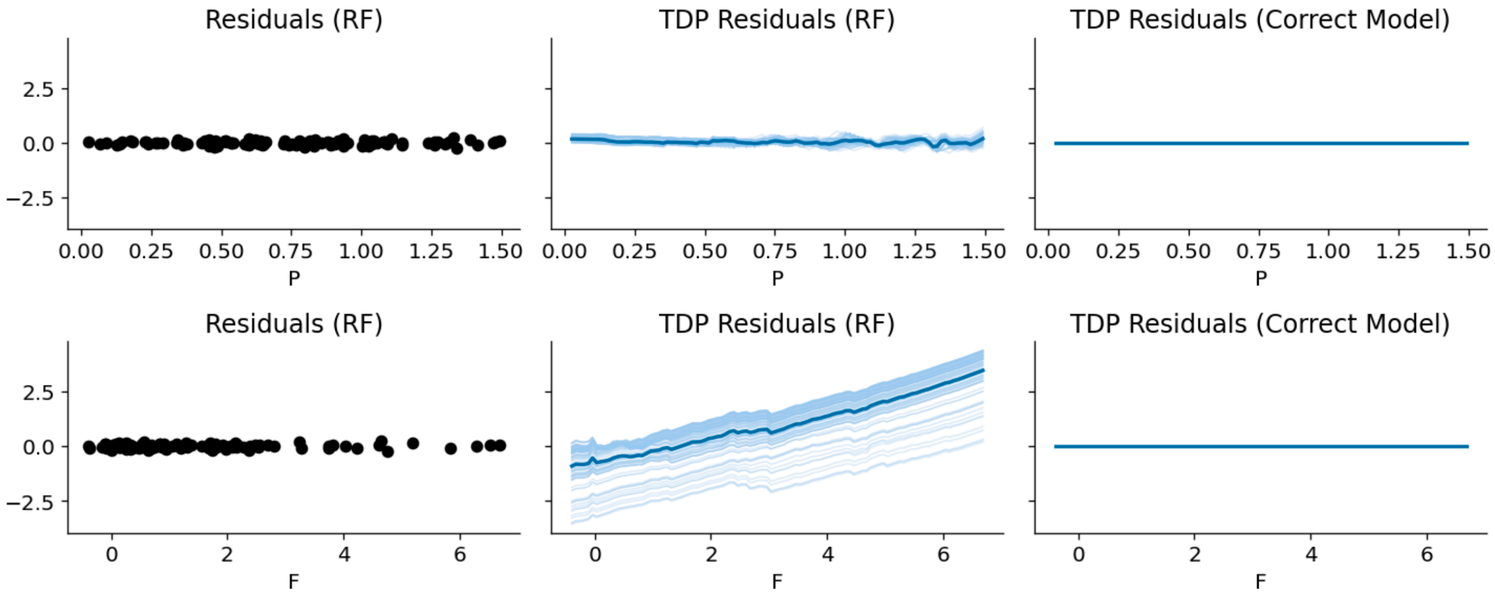


Figure 11: Regular versus CDP residuals for the example in Figure 1, plotted against feature P in the top row and F in the bottom row. Model multiplicity means two models can produce nearly the same predictions, with high accuracy, while using different functional relationships. Accuracy can only show if the model is "observationally correct," (left column) while CDPs can help determine if the model is also "causally correct" (middle vs right columns)

Related: Robust agents learn causal models (Richens et al, ICLR 2024)

Visualize uncertainty

Plot the “envelope” of a set of ECMs

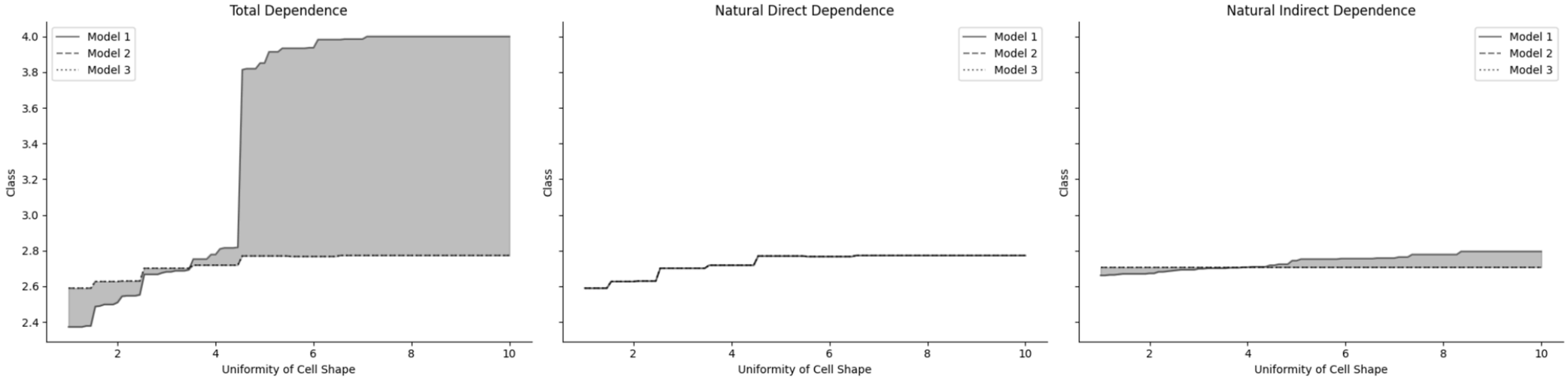


Figure 6: Total Dependence Plots, Natural Direct Dependence Plots and Natural Indirect Dependence Plots for the Breast Cancer Wisconsin dataset under three possible DAGs found by the PC algorithm: (1) $\mathcal{G}_B$ with the edge Cell Shape $\rightarrow$ Cell Size, (2) $\mathcal{G}_B$ with the edge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

Future: combine with various methods, e.g. counterfactual conformal inference

Limitations

Limitations of model-agnostic explanations

- Same model, different explanations

Limitations of model-agnostic explanations

- Same model, different explanations
- “Good” explanations of a model may be “bad” explanations of the world

Limitations of model-agnostic explanations

- Same model, different explanations
- “Good” explanations of a model may be “bad” explanations of the world
- This holds for CDPs as well

Limitations of model-agnostic explanations

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

Limitation of CDPs

Bad ECMs can lead to bad explanations

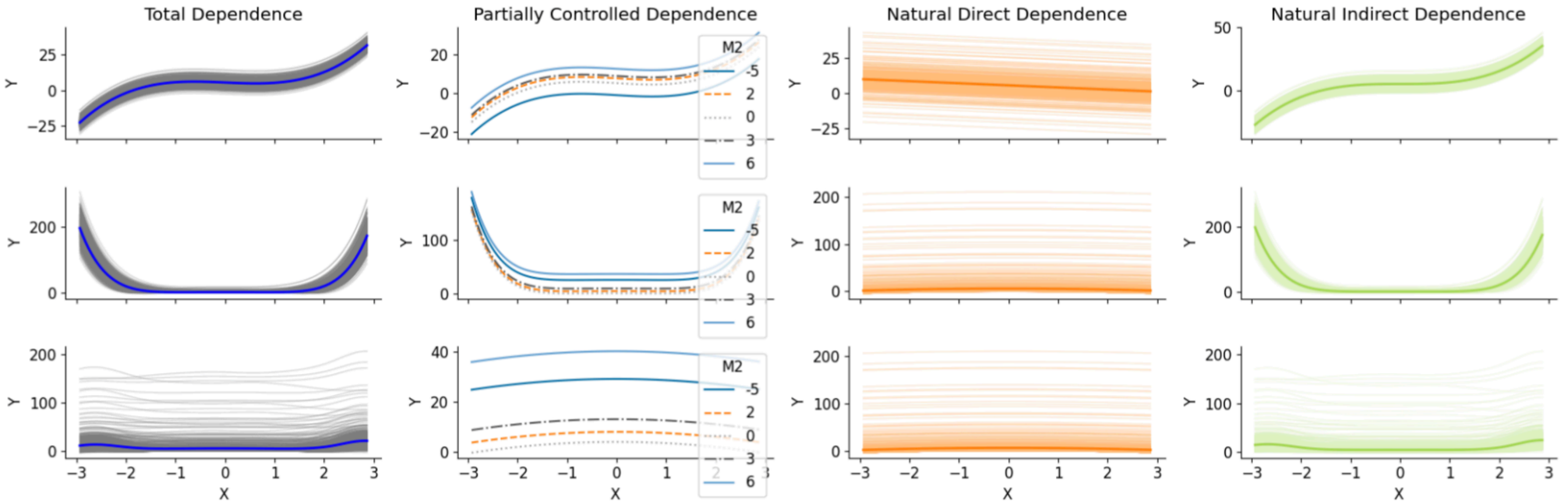


Figure 6: CDPs for the simulation example in Section B.1, shown for a 'good' black-box and correct ECM (top row), a 'bad' black-box model and correct ECM (middle row), and a 'good' black-box and misspecified ECM (bottom row).

Limitations of non-causal explanations

- i.e. potential relative strengths of CDPs
- They also make strong assumptions (but tacitly), e.g. independence
- ECMs show assumptions explicitly!

Limitations of non-causal explanations

- i.e. potential relative strengths of CDPs
- They also make strong assumptions (but tacitly), e.g. independence
- ECMs show assumptions explicitly!
- Attenuation: more automatic methods, like SHAP and PDPs, usually only show “direct effects”

Limitations of non-causal explanations

- i.e. potential relative strengths of CDPs
- They also make strong assumptions (but tacitly), e.g. independence
- ECMs show assumptions explicitly!
- Attenuation: more automatic methods, like SHAP and PDPs, usually only show “direct effects”
  - i.e. only detect “direct discrimination”

Limitations of non-causal explanations

- i.e. potential relative strengths of CDPs
- They also make strong assumptions (but tacitly), e.g. independence
- ECMs show assumptions explicitly!
- 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 show 0 effect

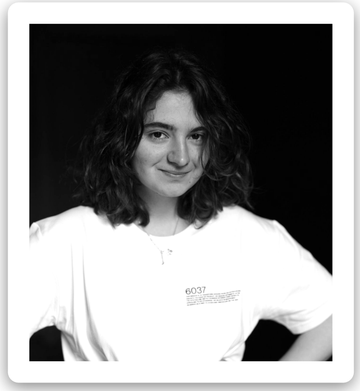
Newer work

Shapley values and causal interpretability

Sakina Hansen,
LSE, **UK**

Maksym Buleshnyi,
Ukrainian Catholic University, **Ukraine**

Heorhii Chyzhmak,
Kremenchuk Mykhailo Ostrohradskyi National University, **Ukraine**



Svitlana Hovorova,
Ukrainian Catholic University, **Ukraine**

PROGRAM GOALS

- Part of the RAI for Ukraine program
- 72 students from 11 Ukrainian universities
- 47 mentors from 23 academic institutions in 8 countries
- Provide a **sense of normalcy**, and high-quality **research opportunities**, to students in Ukraine

RAI for Ukraine

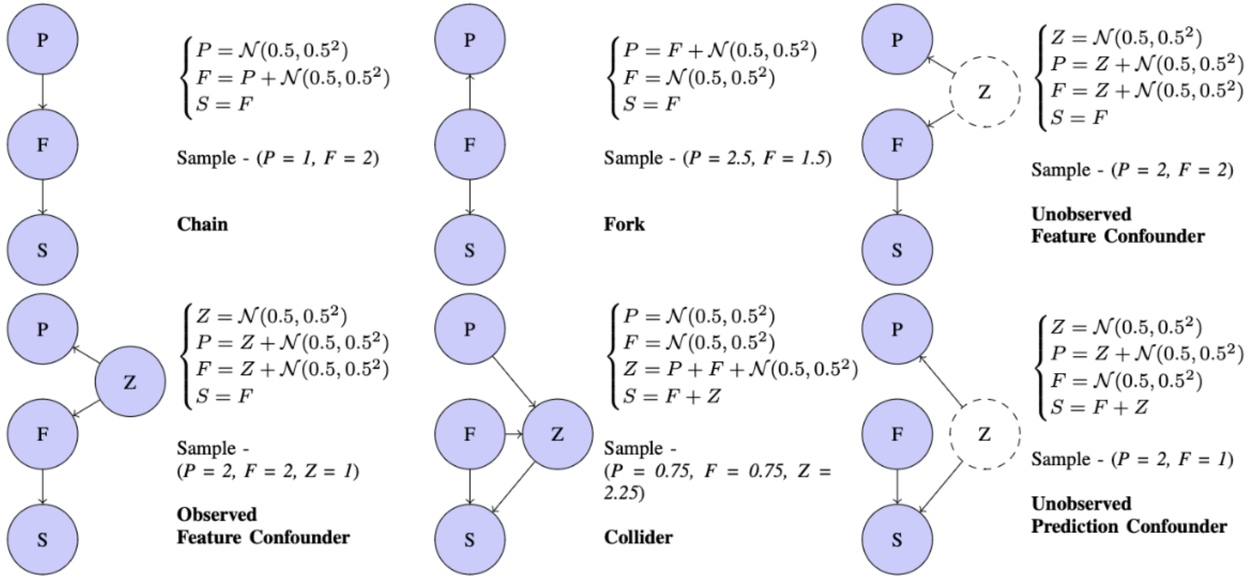
Responsible AI Research for Ukrainian Scholars

Many varieties of SHAP methods

- Standard (observational)
  - Marginal vs conditional (Janzing et al., AISTATS 2020)
- Asymmetric (Frye et al., NeurIPS 2020)
  - Distal vs proximate
  - Conditional vs interventional/causal (Heskes et al., NeurIPS 2020)
- do-Shapley (Jung et al., ICML 2022)
- PW-SHAP (Ter-Minassian et al. ICML 2023)
- … (ongoing)

Some initial results

Validity of interpretation depends on underlying causal structure



Structure	Marginal	Conditional		Causal	
		Symmetric	Asymmetric	Symmetric	Asymmetric
Chain	$\phi_P \approx 0, \phi_F \approx 1$	$\phi_P \approx 0.25, \phi_F \approx 0.75$	$\phi_P \approx 0.5, \phi_F \approx 0.5$	$\phi_P \approx 0.25, \phi_F \approx 0.75$	$\phi_P \approx 0.5, \phi_F \approx 0.5$
Fork	$\phi_P \approx 0, \phi_F \approx 1$	$\phi_P \approx 0.5, \phi_F \approx 0.5$	$\phi_P \approx 0, \phi_F \approx 1$	$\phi_P \approx 0, \phi_F \approx 1$	$\phi_P \approx 0, \phi_F \approx 1$
Observed Feature Confounder	$\phi_P \approx 0, \phi_F \approx 1, \phi_Z \approx 0$	$\phi_P \approx 0.2, \phi_F \approx 0.7, \phi_Z \approx 0.2$	$\phi_P \approx 0, \phi_F \approx 0.5, \phi_Z \approx 0.5$	$\phi_P \approx 0, \phi_F \approx 0.7, \phi_Z \approx 0.3$	$\phi_P \approx 0, \phi_F \approx 0.5, \phi_Z \approx 0.5$
Collider	$\phi_P \approx 0, \phi_F \approx 0.25, \phi_Z \approx 0.75$	$\phi_P \approx 0.2, \phi_F \approx 0.3, \phi_Z \approx 0.5$	$\phi_P \approx 0.25, \phi_F \approx 0.75, \phi_Z \approx 0$	$\phi_P \approx 0.2, \phi_F \approx 0.5, \phi_Z \approx 0.3$	$\phi_P \approx 0.25, \phi_F \approx 0.5, \phi_Z \approx 0.25$
Unobserved Feature Confounder	$\phi_P \approx 0, \phi_F \approx 1$	$\phi_P \approx 0.25, \phi_F \approx 0.75$	$\phi_P \approx 0.25, \phi_F \approx 0.75$	$\phi_P \approx 0.25, \phi_F \approx 0.75$	$\phi_P \approx 0.25, \phi_F \approx 0.75$
Unobserved Prediction Confounder	$\phi_P \approx 0.5, \phi_F \approx 0.5$	$\phi_P \approx 0.5, \phi_F \approx 0.5$	$\phi_P \approx 0.5, \phi_F \approx 0.5$	$\phi_P \approx 0.5, \phi_F \approx 0.5$	$\phi_P \approx 0.5, \phi_F \approx 0.5$

Applications and future directions

- Expand on CDP framework (journal version)
- User-friendly software libraries
- Visualize/infer uncertainty, relax assumptions (partial ECMs)
- Survey/tutorial on SHAP variants & causality (guidelines on which to use)
- Applications: Fairness, robustness, distribution shift, scientific ML…
- Human-guided exploration/explanation of large (e.g. “foundation”) model

I just woke up, what did I miss?

Summary

Interpret/understand “black-box” prediction models

Visualize how predictions “depend" on input features

New: “Dependence” includes causal relationships between input features

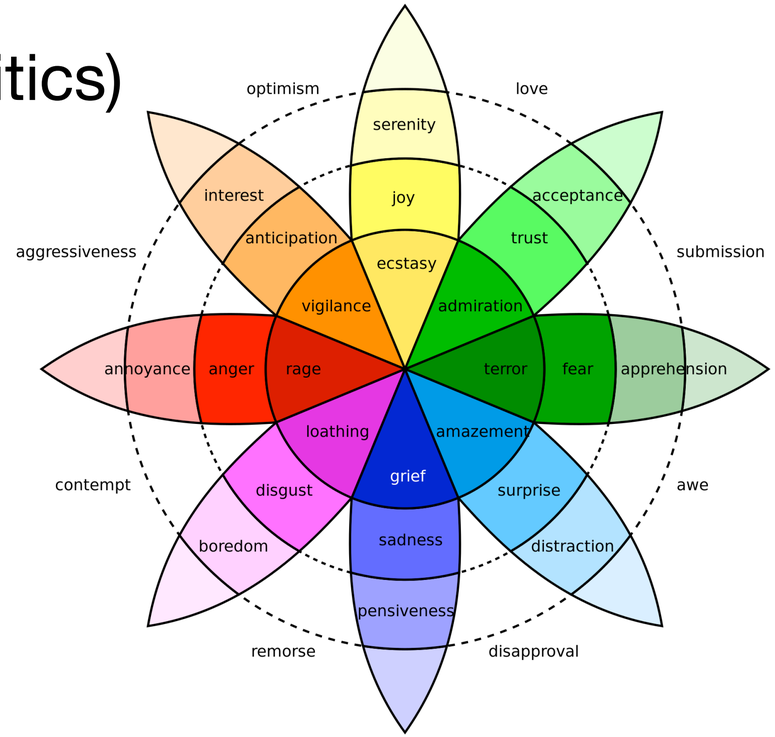
Concluding remarks (provocations?)

For more, see my paper
Position: The Causal Revolution Needs Scientific Pragmatism (ICML 2024)

Pluralism (and values)

Internal conflict

- Humans (and other animals) have different values, emotions, affective states
- We are conflicted even at the individual level, internally
- And socially (e.g. politics)



Whose values?

Promoted and scaled up by automation and technology

A. N. Whitehead: “It is a profoundly erroneous truism, repeated by all copy-books and by people when they are making speeches, that we should cultivate the habit of *thinking of what we are doing*.

The precise opposite is the case. Civilization advances by extending the number of **important operations** which we can perform *without thinking about them*.”

Feynman vs Whitehead?

Ways of thinking or machines that “think”?

- “Civilization advances”
  - In what direction? Is that good? For who?
- Automating “important operations”
  - Who decides what’s important? How do they know if it’s working?

Human-centered data science: models and tools
that work for (all of) us, helpful ways of thinking

For more, see my paper
Position: The Causal Revolution Needs Scientific Pragmatism (ICML 2024)

Thanks for your attention!

Q&A time