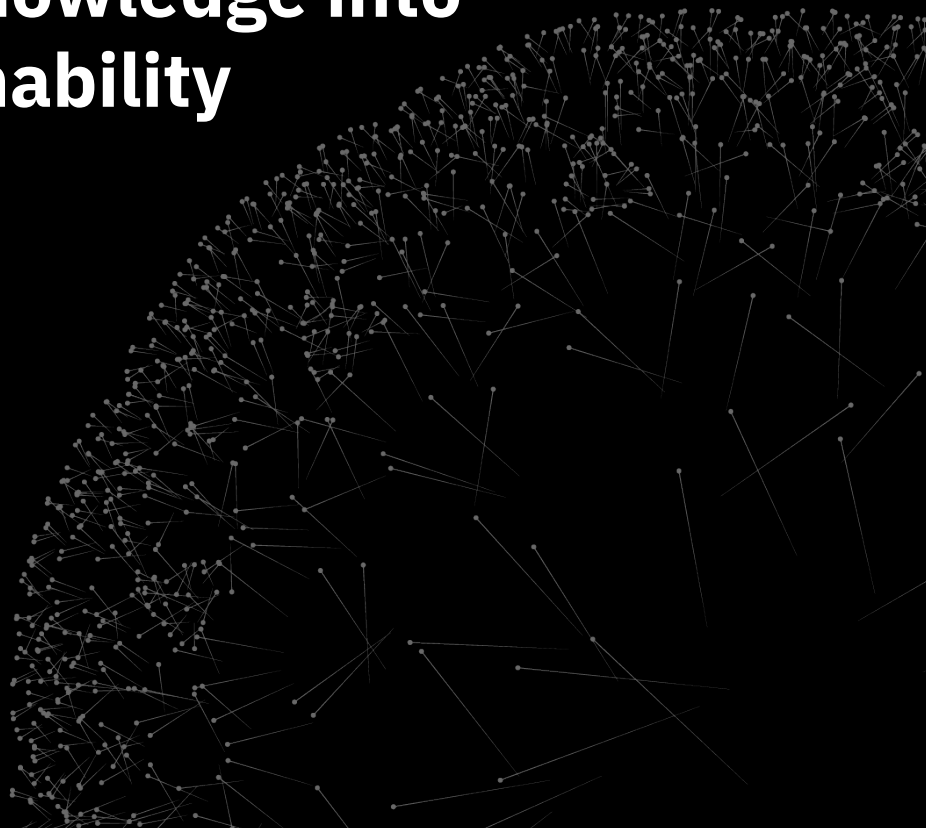
Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability

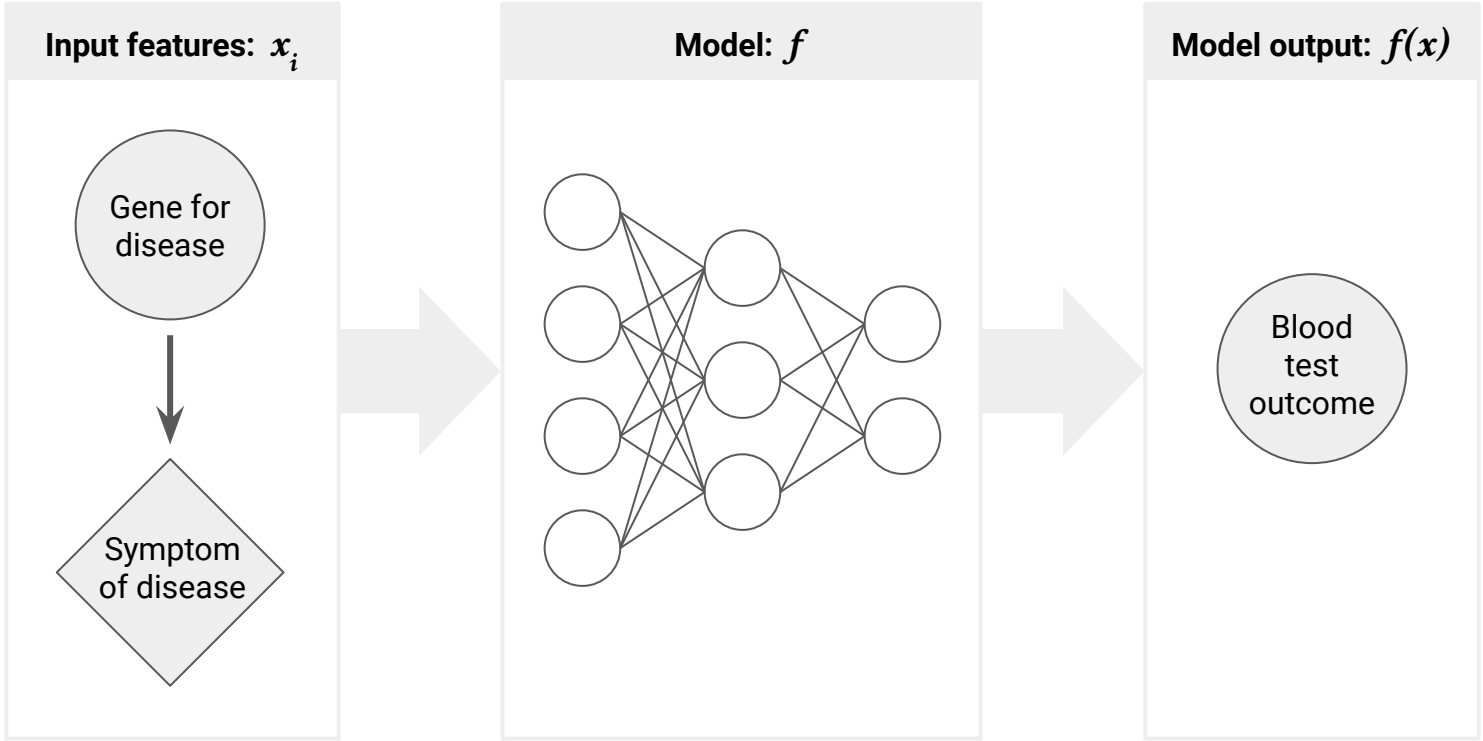
Christopher Frye
Faculty

Colin Rowat
University of Birmingham

Ilya Feige
Faculty



Simple example of *causal structure* underlying a prediction problem



Motivation for model explainability using *asymmetric* Shapley values

Shapley values are the unique attribution method satisfying 4 axioms:

- **Efficiency:** attribution for the model's output is fully distributed over its input features
- **Linearity:** Shapley values aggregate linearly over members of a linear-ensemble model
- **Nullity:** features that do not influence model predictions receive zero attribution
- **Symmetry:** features that influence model predictions identically get equal Shapley values

$$\phi(i) = \sum_{\pi \in \Pi} \frac{1}{n!} \left(\mathbb{E}[\, f \,|\, x_i \,, x_{\pi < \pi(i)}] - \mathbb{E}[\, f \,|\, x_{\pi < \pi(i)}] \right)$$

Motivation for model explainability using *asymmetric* Shapley values

Shapley values are the unique attribution method satisfying 4 axioms:

- **Efficiency:** attribution for the model's output is fully distributed over its input features
- **Linearity:** Shapley values aggregate linearly over members of a linear-ensemble model
- **Nullity:** features that do not influence model predictions receive zero attribution
- **Symmetry:** features that influence model predictions identically get equal Shapley values

This axiom requires that
gene & symptom have equal Shapley values,
obfuscating the underlying causality

$$\phi(i) = \sum_{\pi \in \Pi} \frac{1}{n!} \left(\mathbb{E}[\, f \,|\, x_i \,, x_{\pi < \pi(i)}] - \mathbb{E}[\, f \,|\, x_{\pi < \pi(i)}] \right)$$

Motivation for model explainability using *asymmetric* Shapley values

Shapley values are the unique attribution method satisfying 4 axioms:

- **Efficiency:** attribution for the model's output is fully distributed over its input features
- **Linearity:** Shapley values aggregate linearly over members of a linear-ensemble model
- **Nullity:** features that do not influence model predictions receive zero attribution
- **Symmetry:** features that influence model predictions identically get equal Shapley values

Dropping this axiom leads to a generalisation of Shapley values that's been explored in the game theory literature:

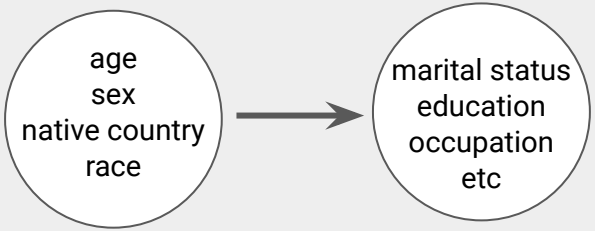
"Asymmetric Shapley values"

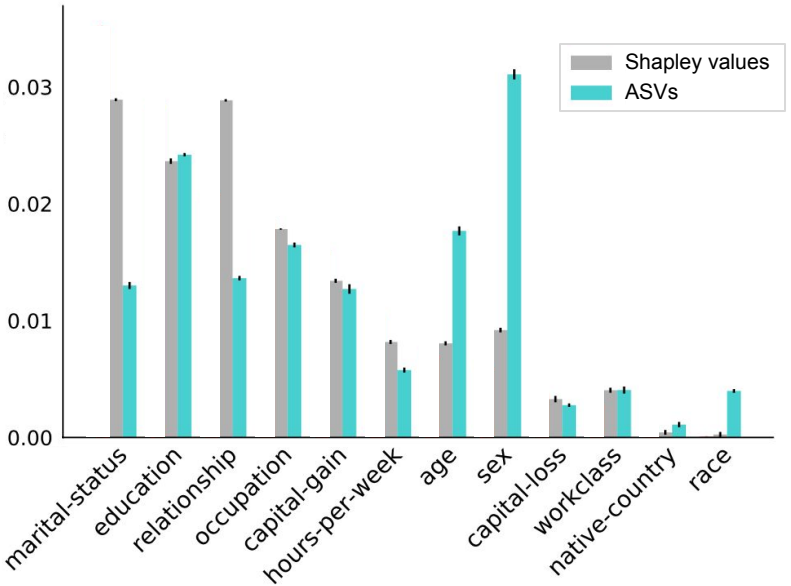
$$\phi(i) = \sum_{\pi \in \Pi} w(\pi) \left(\mathbb{E}[\, f \,|\, x_i \,, x_{\pi<\pi(i)}] - \mathbb{E}[\, f \,|\, x_{\pi<\pi(i)}] \right)$$

Application: explaining income predictions on census data

In the Census Income data set (UCI)
some features are clearly causal ancestors
in an otherwise complex causal process.

With partial causal knowledge, we choose
w(π) to weight those permutations consistent
with our causal understanding.





Additional applications

The paper includes experiments on additional applications:

- **Causal fairness:** ASVs can measure whether a model satisfies notions of fairness defined with respect to a causal graph.
- **Time series:** On data that is intrinsically ordered, ASVs indicate the incremental effect of each time step in the series on the model's output.
- **Feature selection:** ASVs have a direct interpretation as the incremental increase in model accuracy due to each feature.

ASVs are also being used in the real world: to explain COVID-19 forecasting models used by the UK's National Health Service.