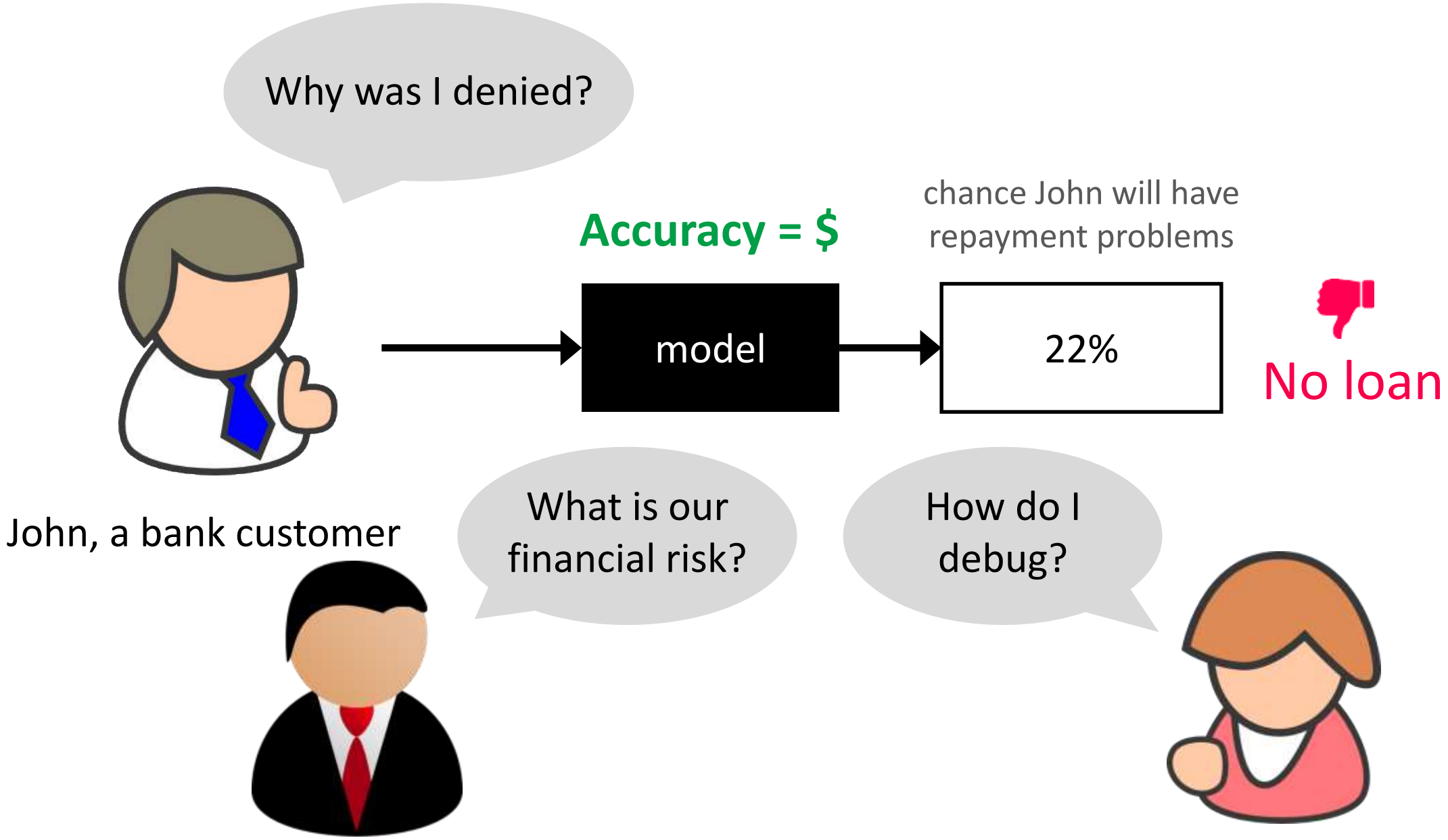
Explainable Machine Learning with Shapley Values

Scott Lundberg
Senior Researcher
Microsoft

Explainable AI in practice

Model development


Why was I denied?
Accuracy = $
chance John will have repayment problems
model
22%
No loan
John, a bank customer
What is our financial risk?
How do I debug?

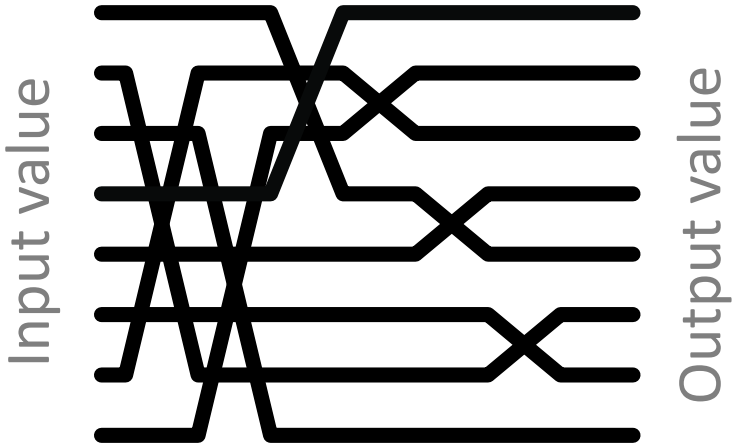
Interpretable **Accurate**

Complex model ← ?

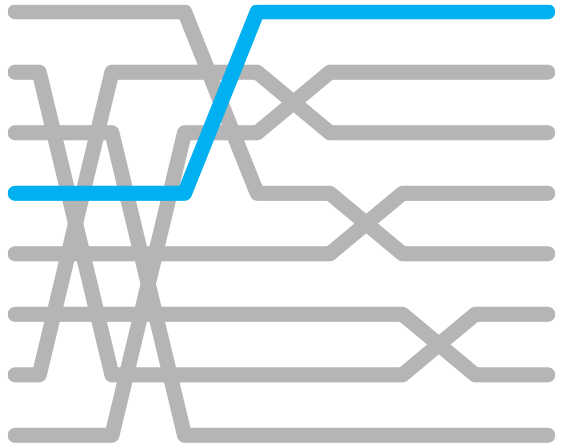
Simple model ? →

Interpretable or accurate: **choose one**.

😁 ⚖ 💰



Complex models are inherently complex!



But a single prediction involves only a small piece of that complexity.

Base rate

16%

$E[f(X)]$

Prediction for John

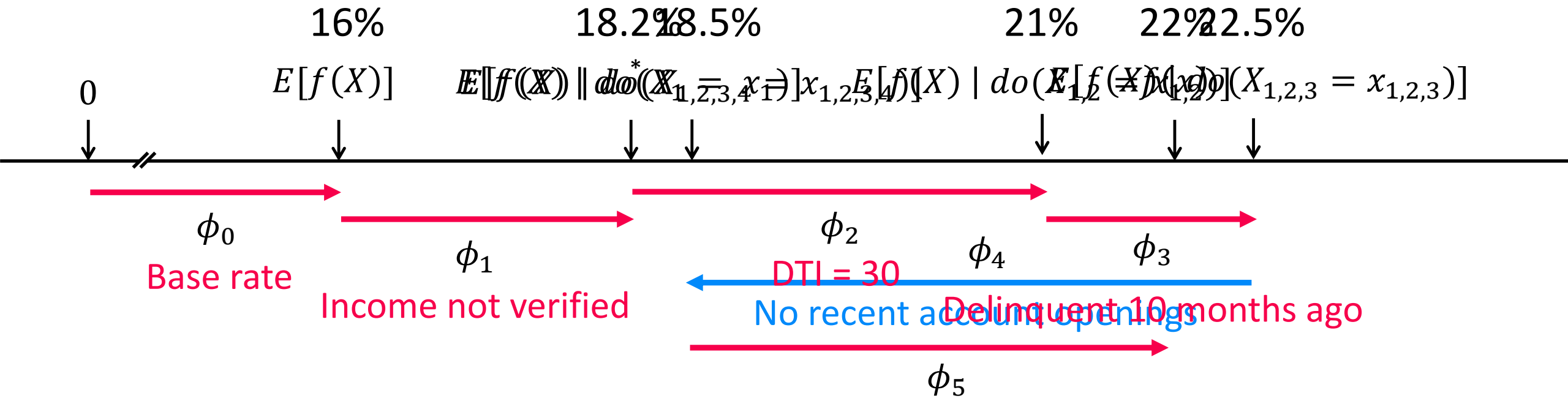
22%

$f(x)$

0

How did we get here?

The order matters!

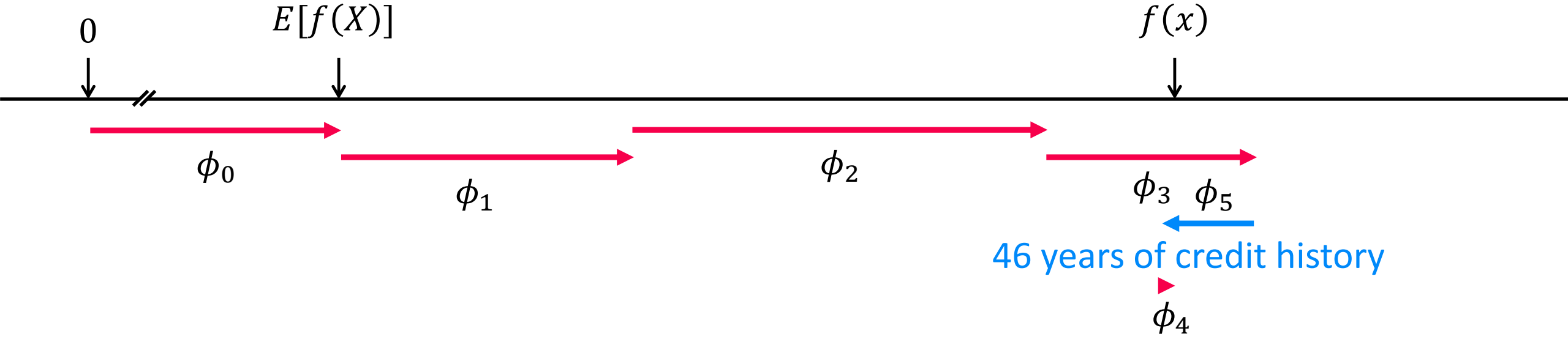


*Janzing et al. 2019

Lloyd Shapley

The order matters!

Nobel Prize in 2012

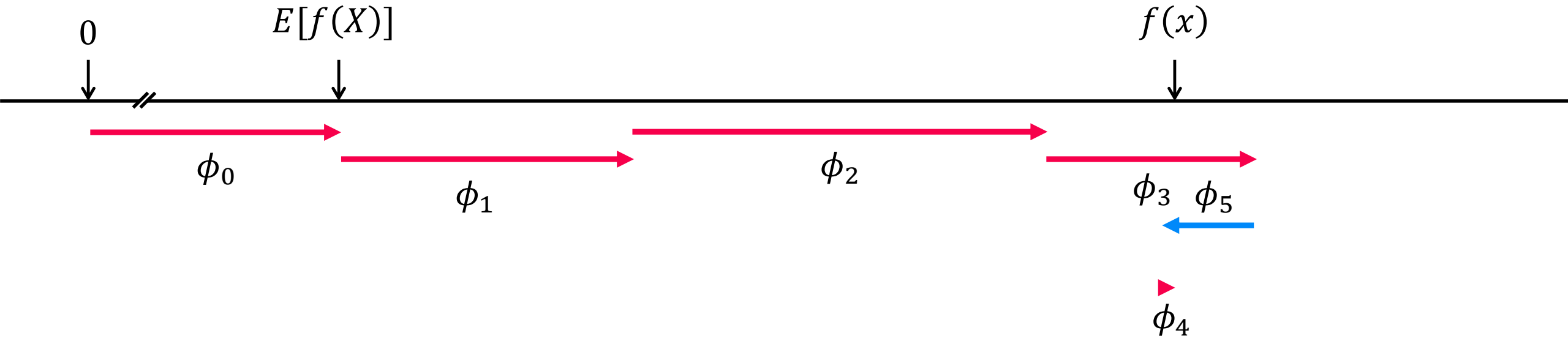


Shapley properties

1

Additivity (local accuracy) – The sum of the local feature attributions equals the difference between rate and the model output.

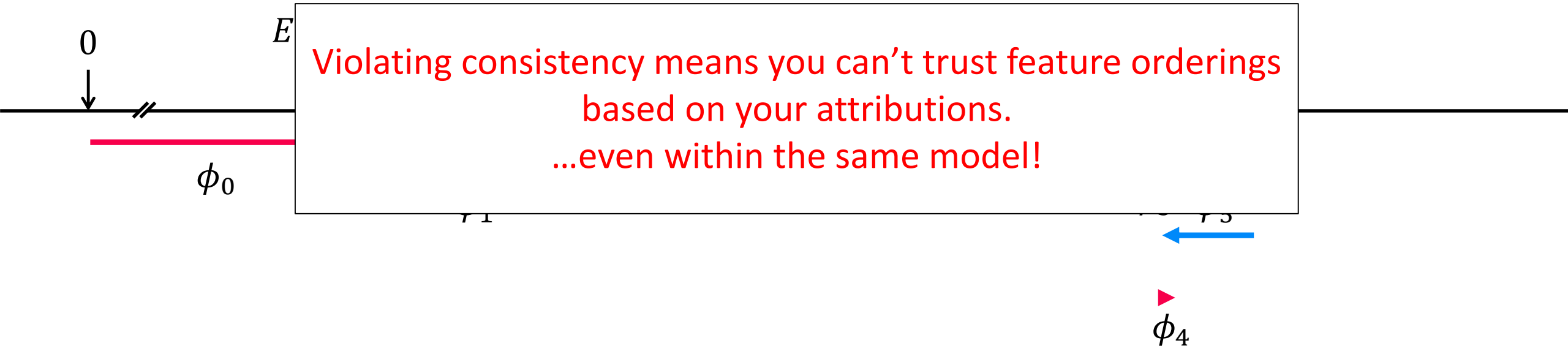
$$E[f(x)] + \sum_{i=1}^{M} \phi_i = f(x)$$



Shapley properties

2

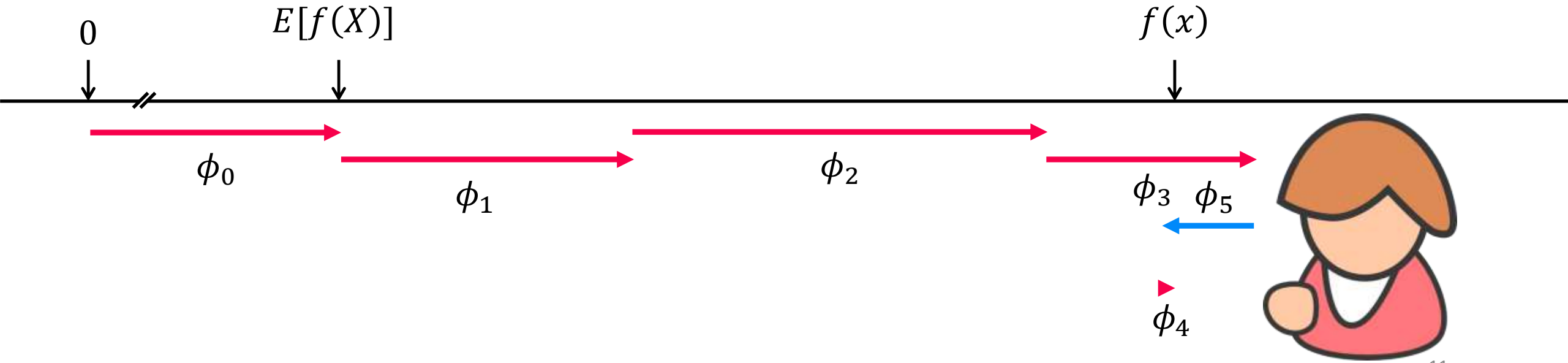
Monotonicity (consistency) – If you change the original model such that a feature has a larger possible ordering, then that input's attribution decrease.



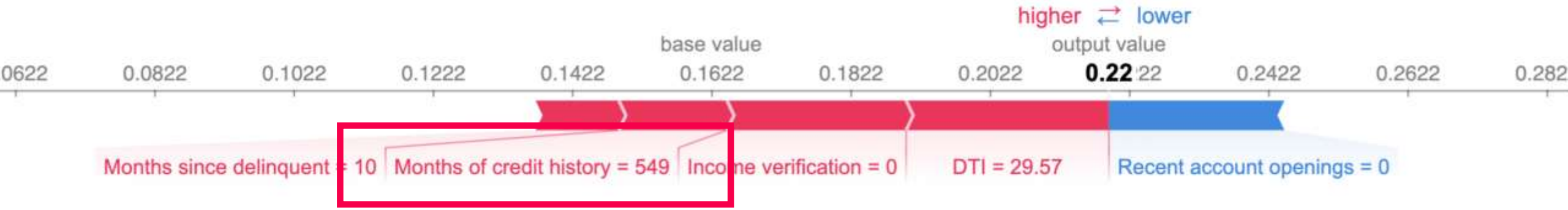
Violating consistency means you can't trust feature orderings based on your attributions.
…even within the same model!

Shapley values result from **averaging over all N! possible orderings**.

(NP-hard)



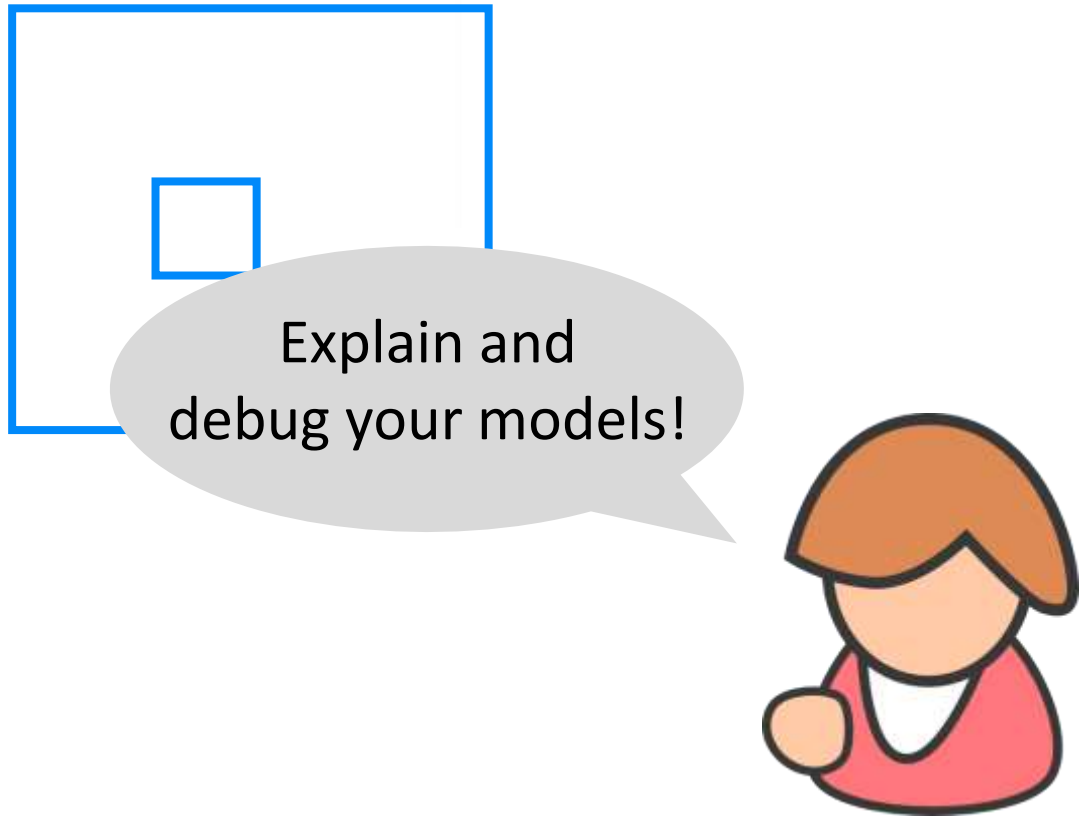
```
ex = shap.TreeExplainer(model, …)
shap_values = ex.shap_values(X)
shap.force_plot(                                                    )
```



Why does 46 years of credit history increase the risk of payment problems?

```
shap.dependence_plot(                                                    )
```

The model is identifying retirement-age individuals based on their long credit histories!



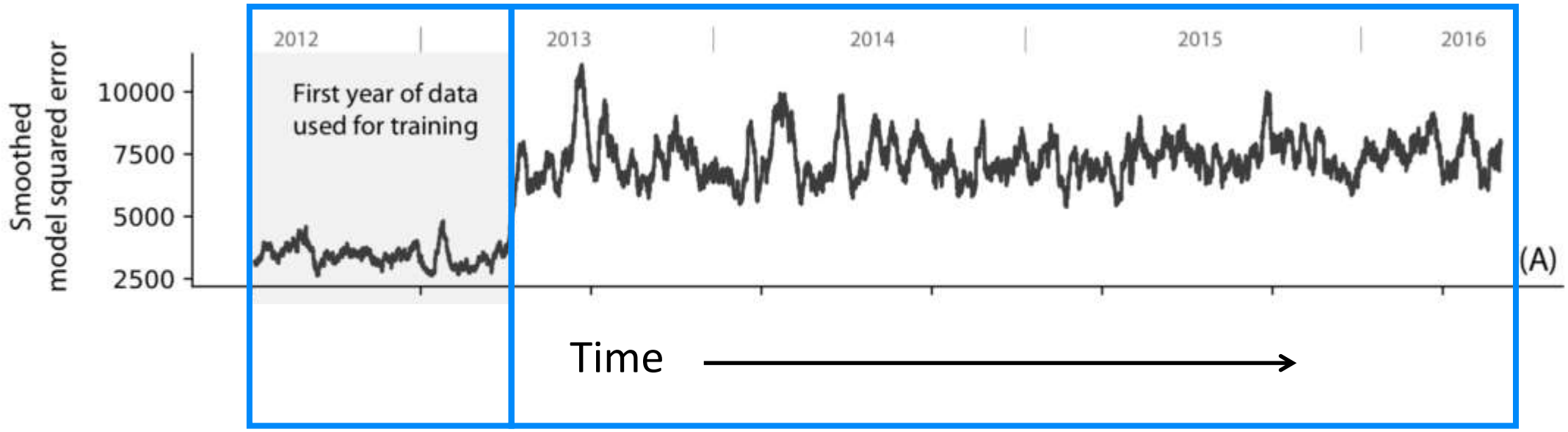
Explainable AI in practice

Model development

Debugging/exploration

Monitoring

Model monitoring

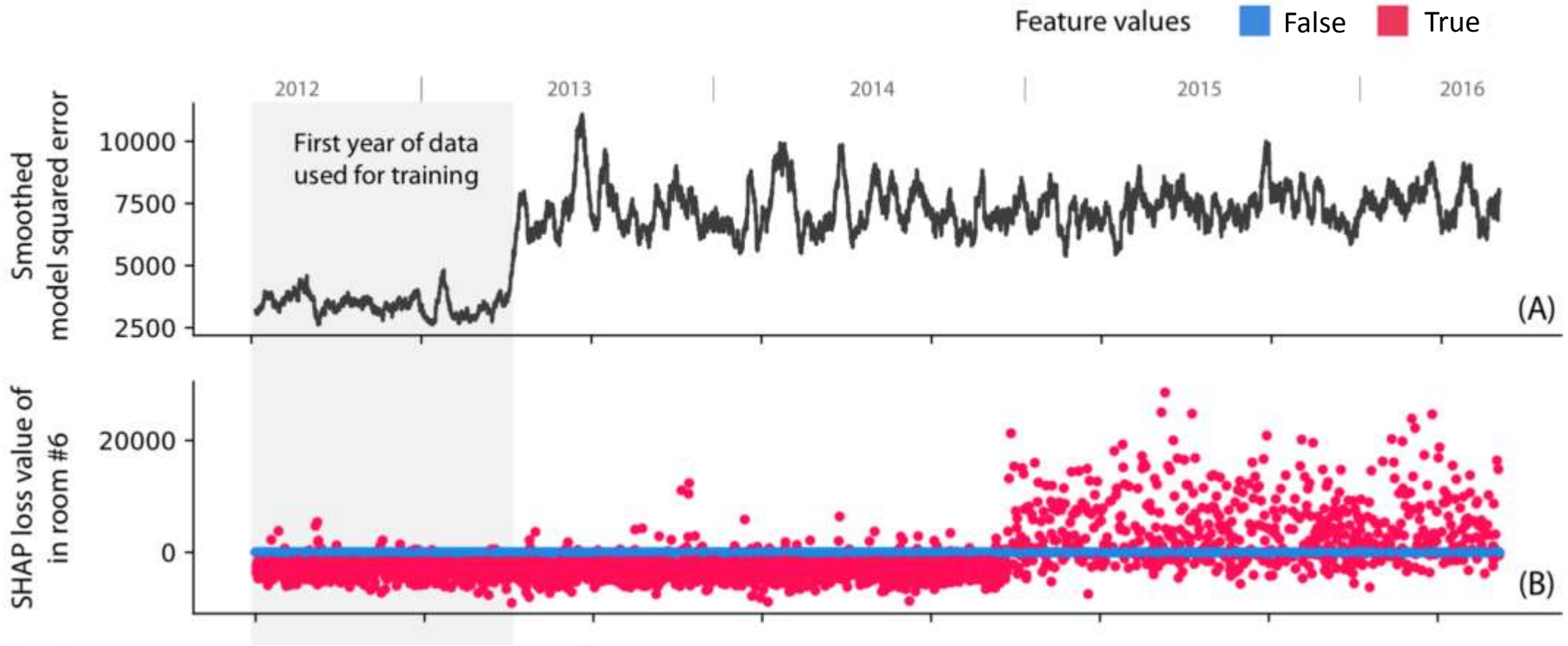


Training performance

Test performance

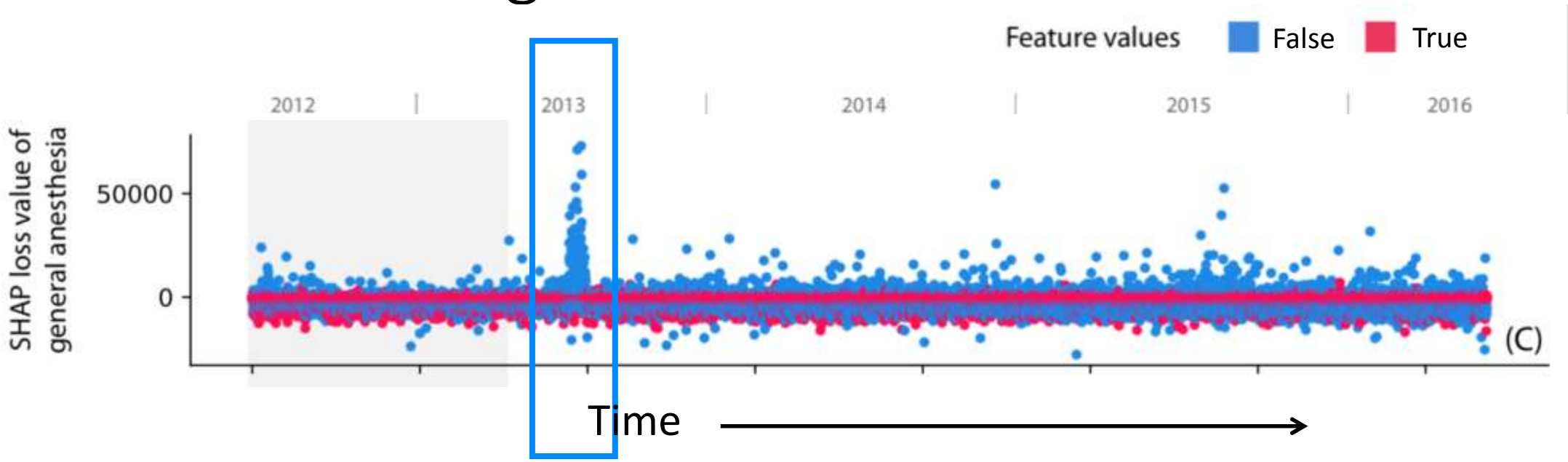
Can you find where we introduced the bug?

Model monitoring



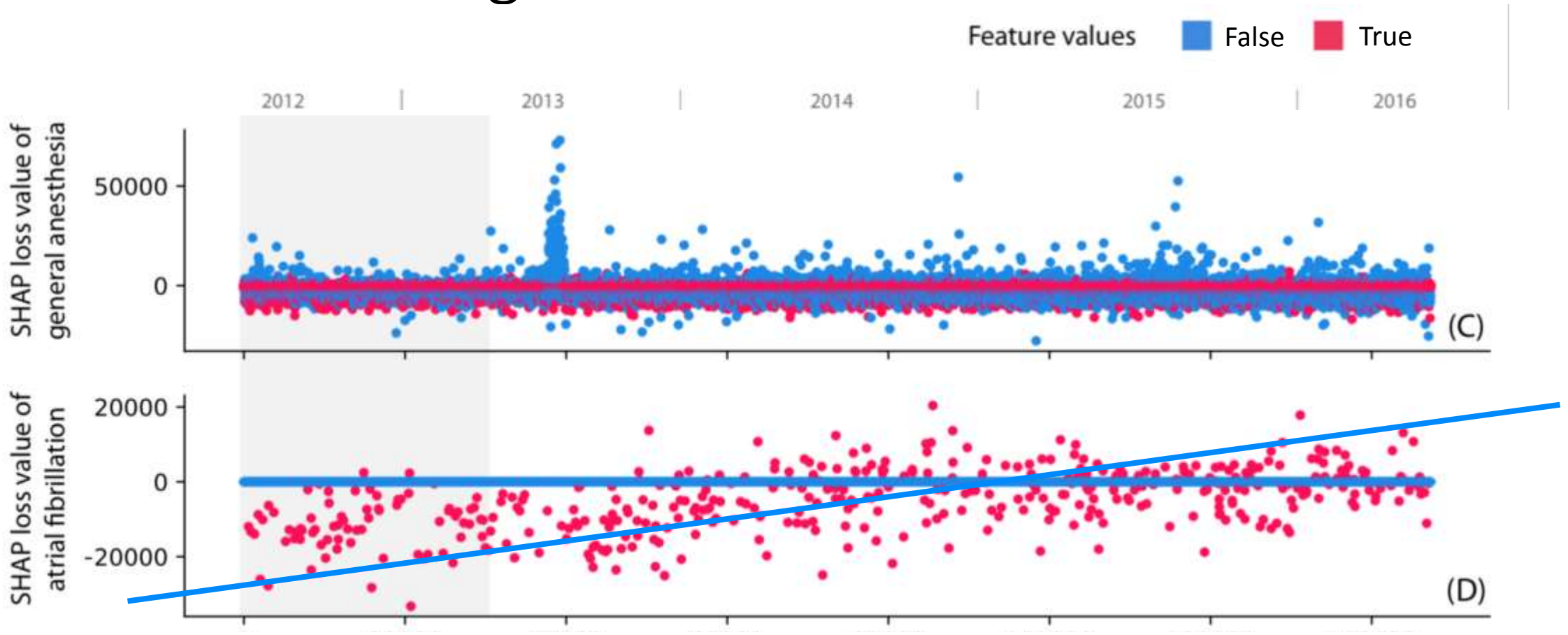
Now can you find where we introduced the bug?

Model monitoring



Transient electronic medical record

Model monitoring



Gradual change in atrial fibrillation ablation procedure durations

Explainable AI in practice

Model development

Debugging/exploration

Monitoring

Encoding prior beliefs

Human/AI collaboration

Customer retention

Decision support

Human risk oversight

Regulatory compliance

Consumer explanations

Anti-discrimination

Risk management

Scientific discovery

Population subtyping

Pattern discovery

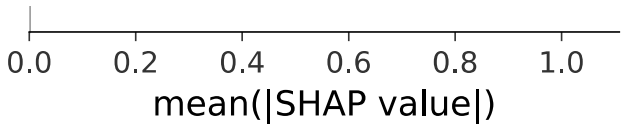
Signal recovery

Thank You

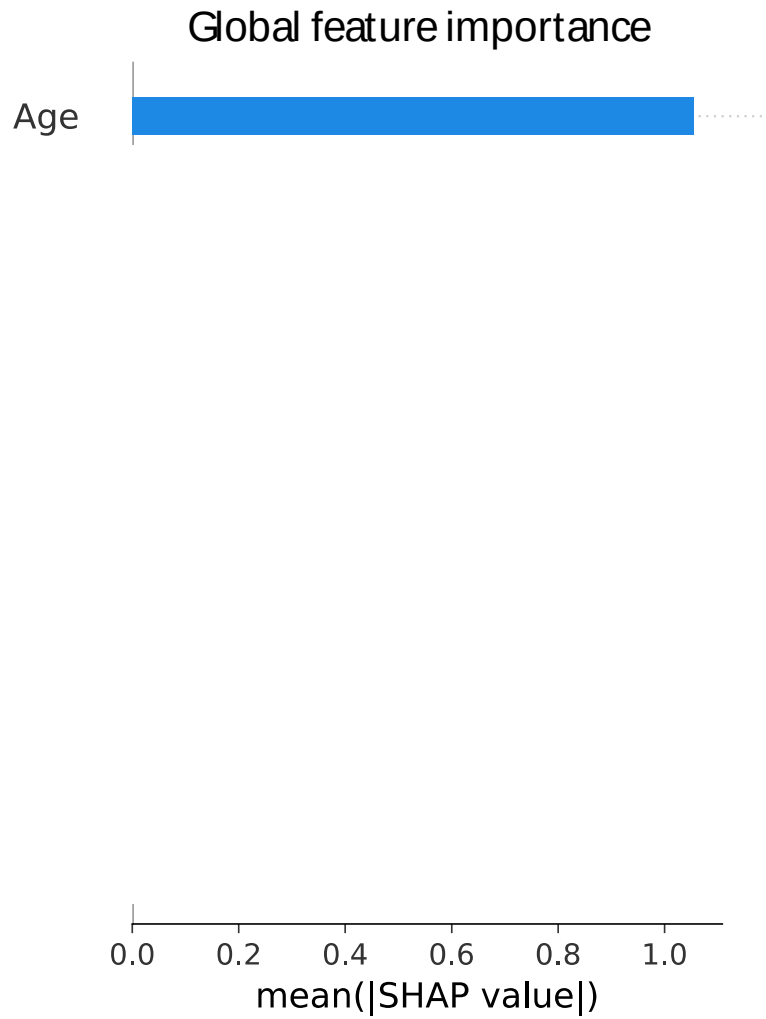
github.com/slundberg/shap

Mortality risk model

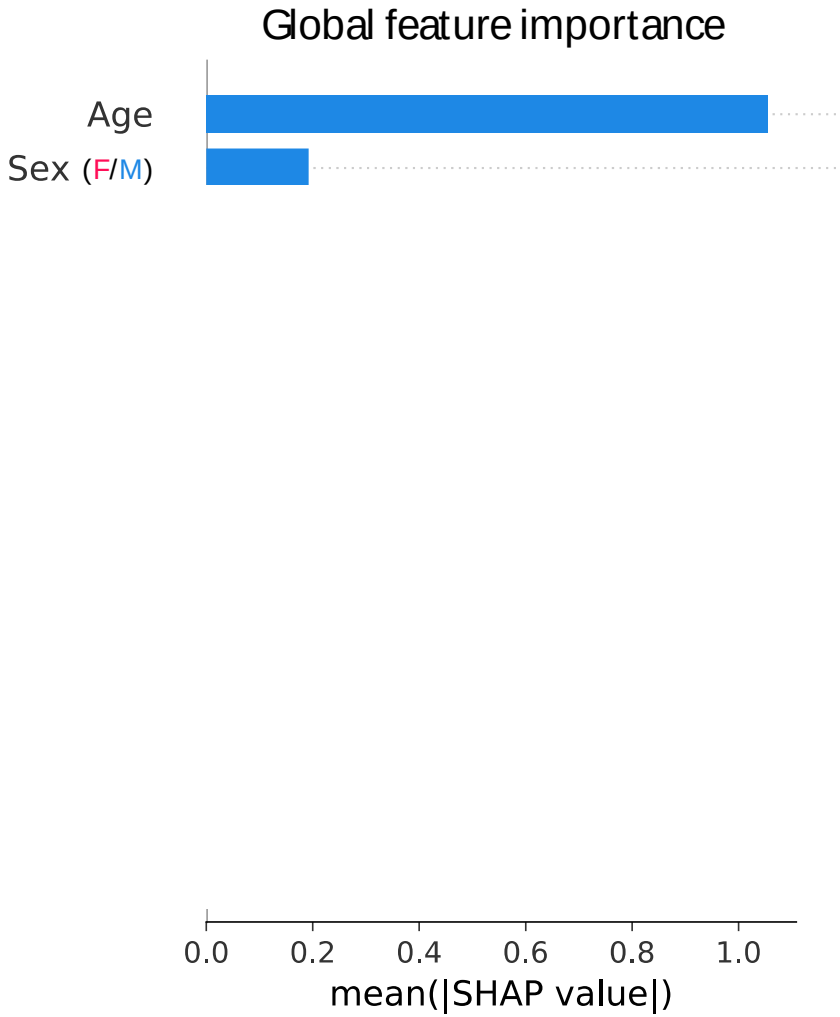
Global feature importance



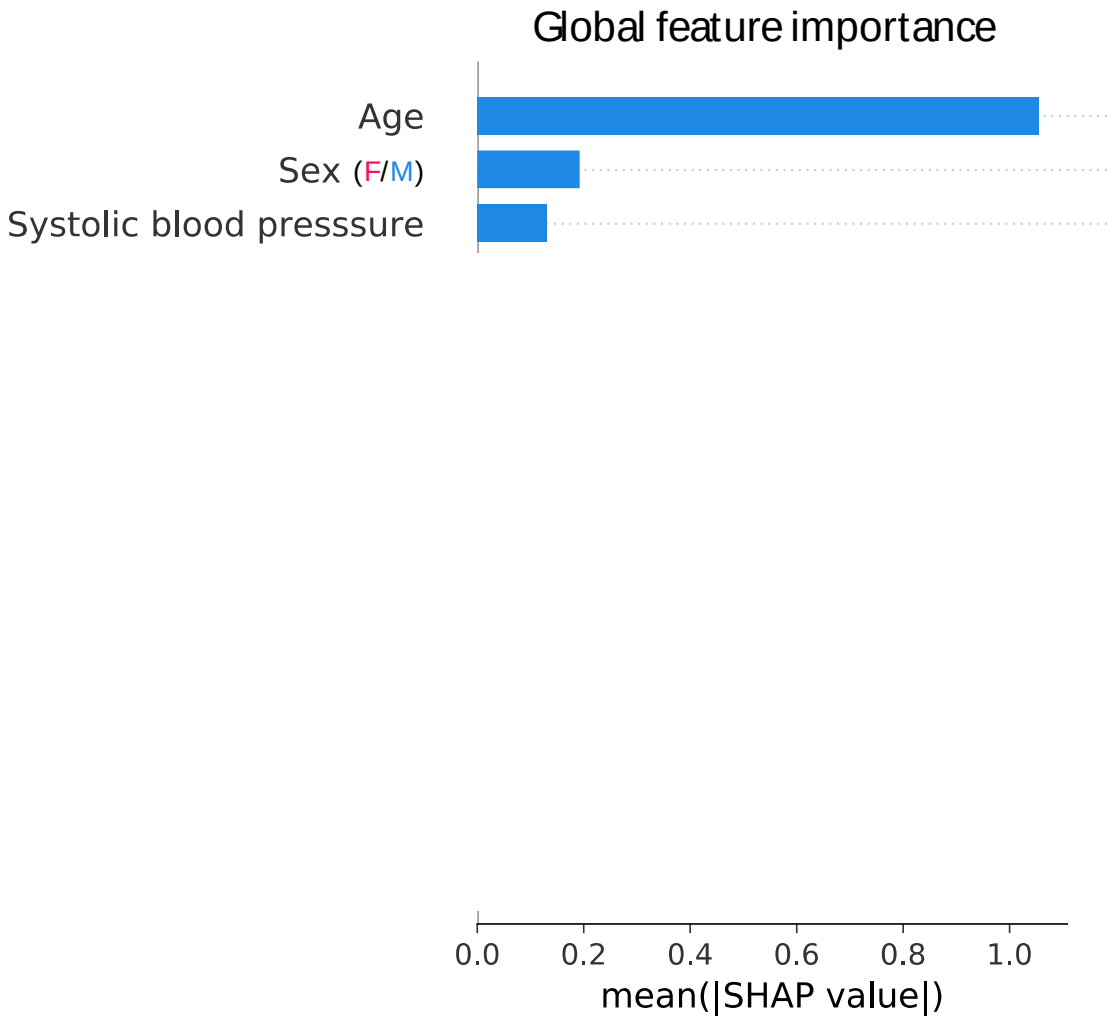
Mortality risk model



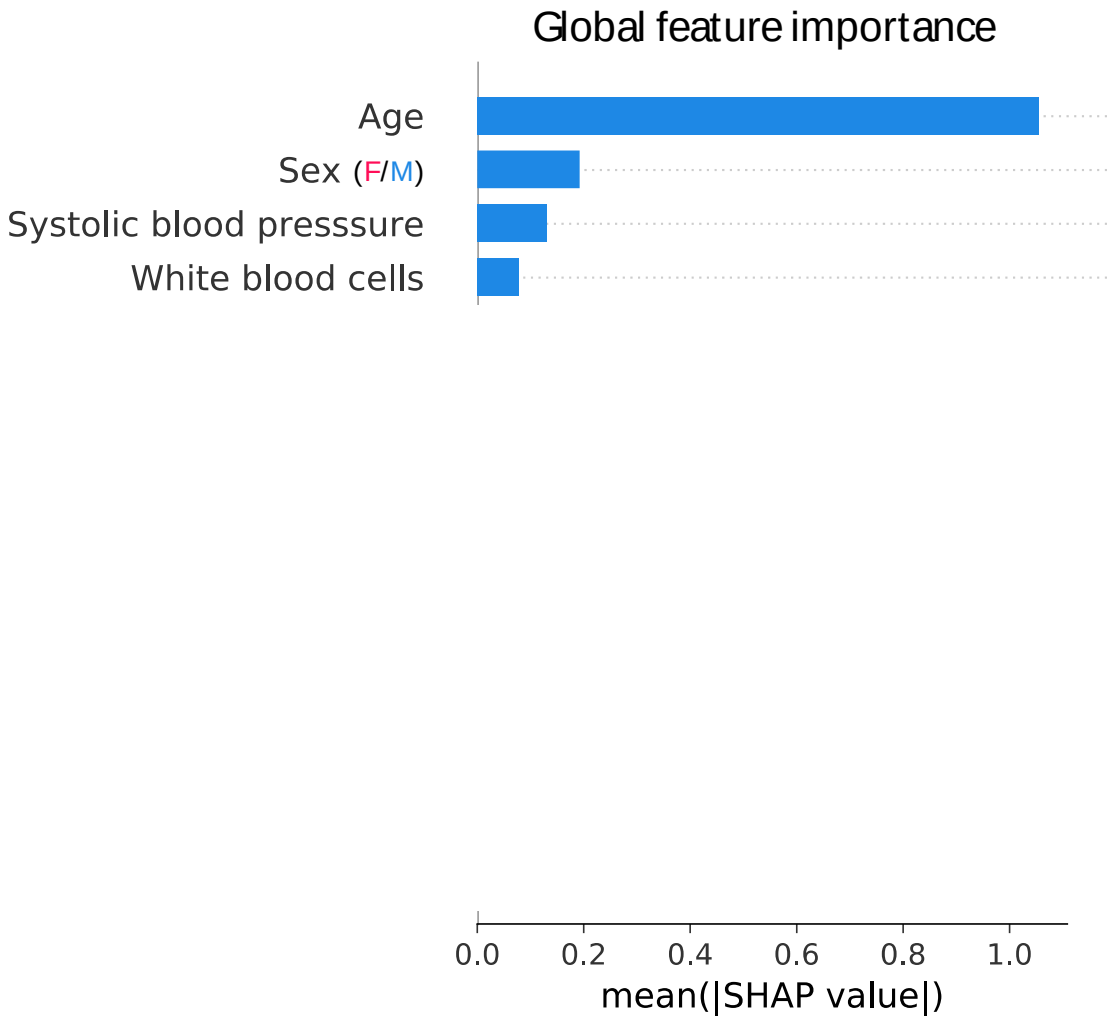
Mortality risk model



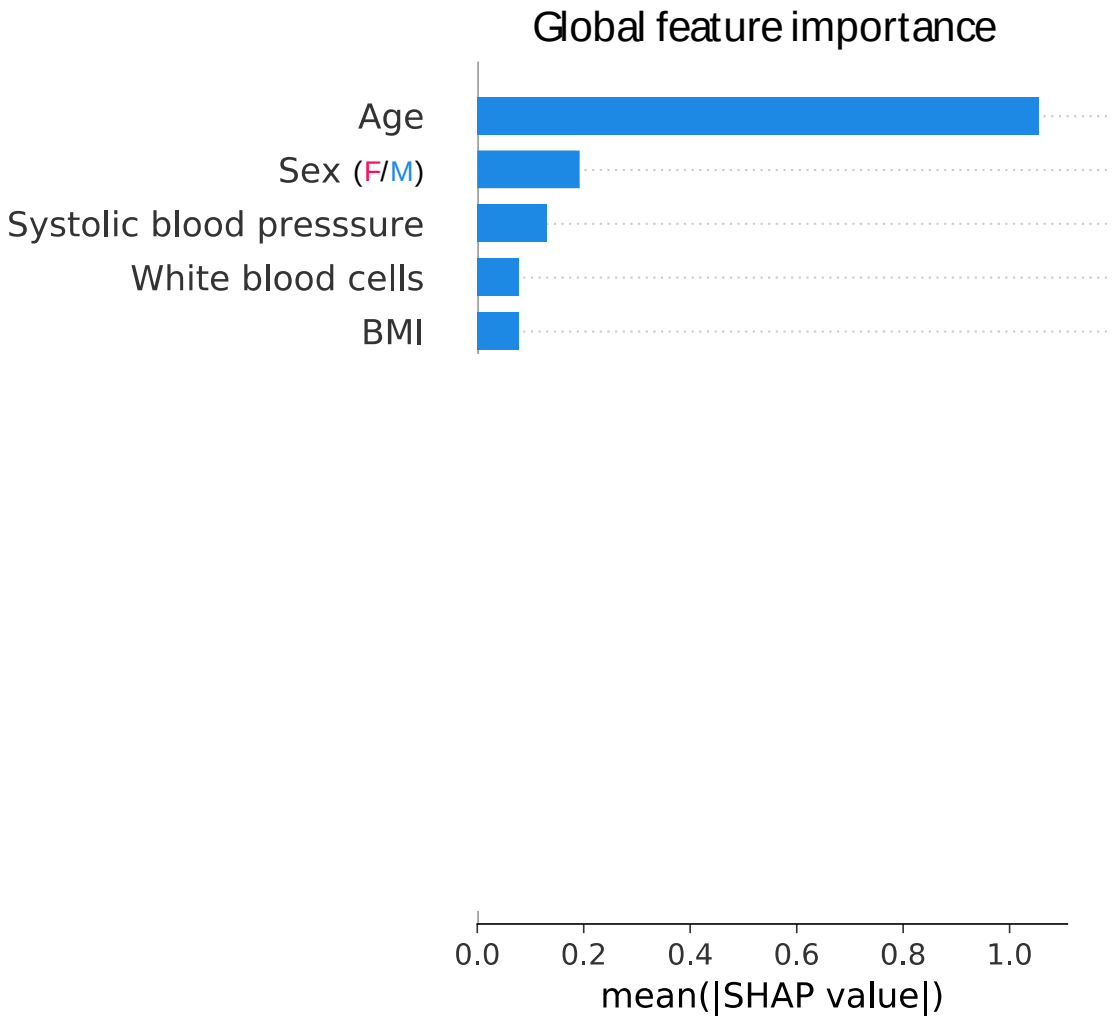
Mortality risk model



Mortality risk model

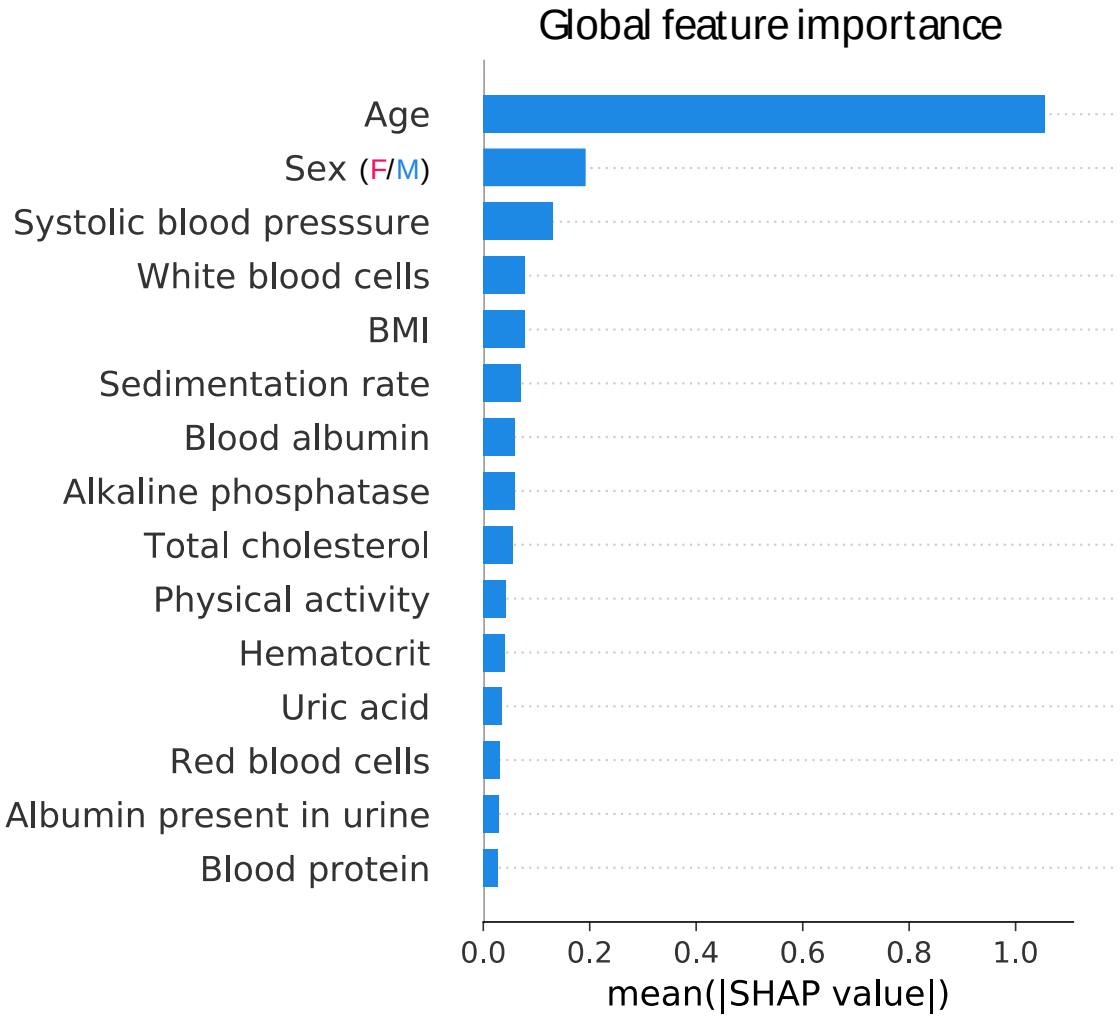


Mortality risk model



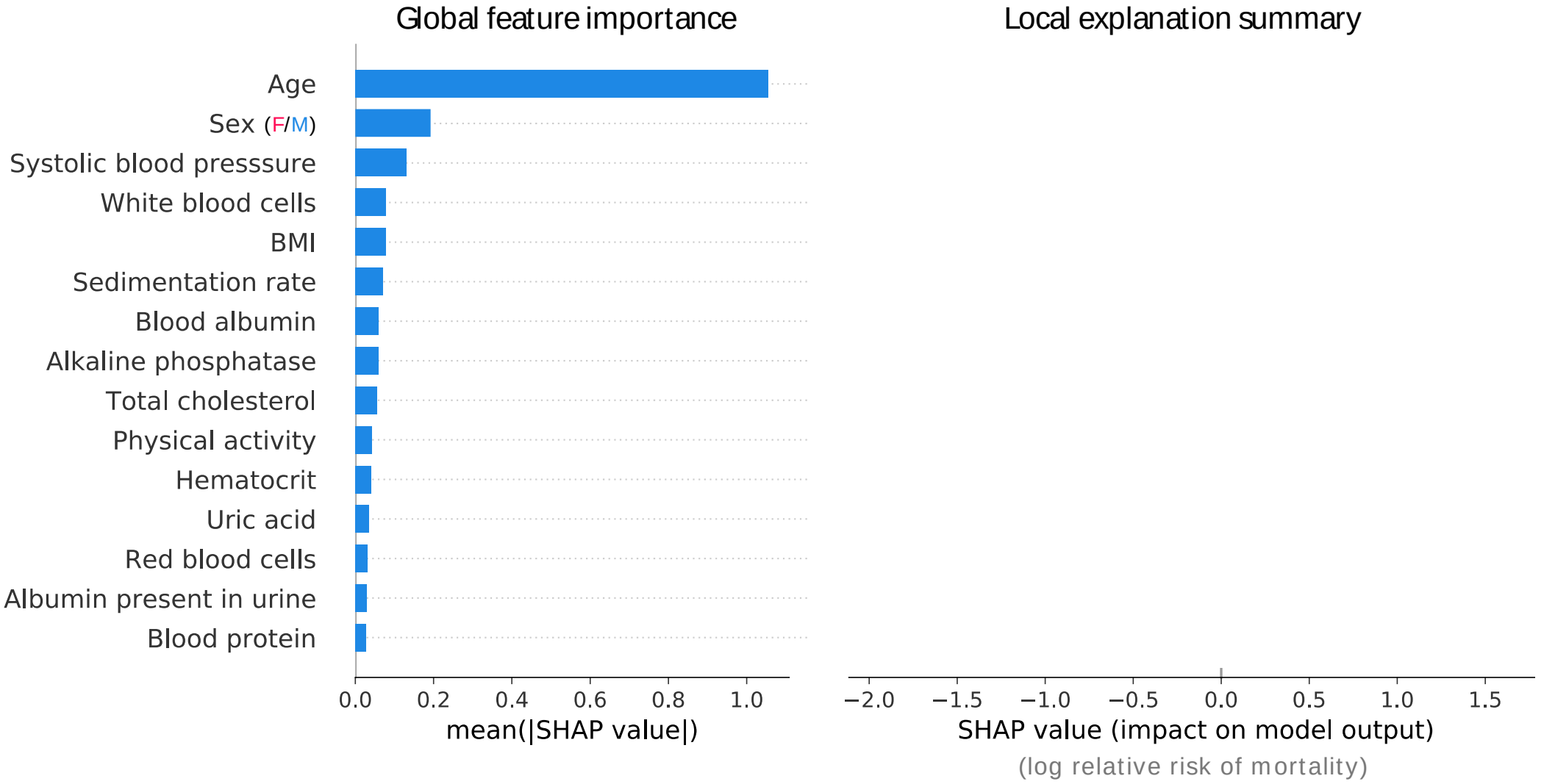
Reveal rare high-magnitude mortality effects

Mortality risk model

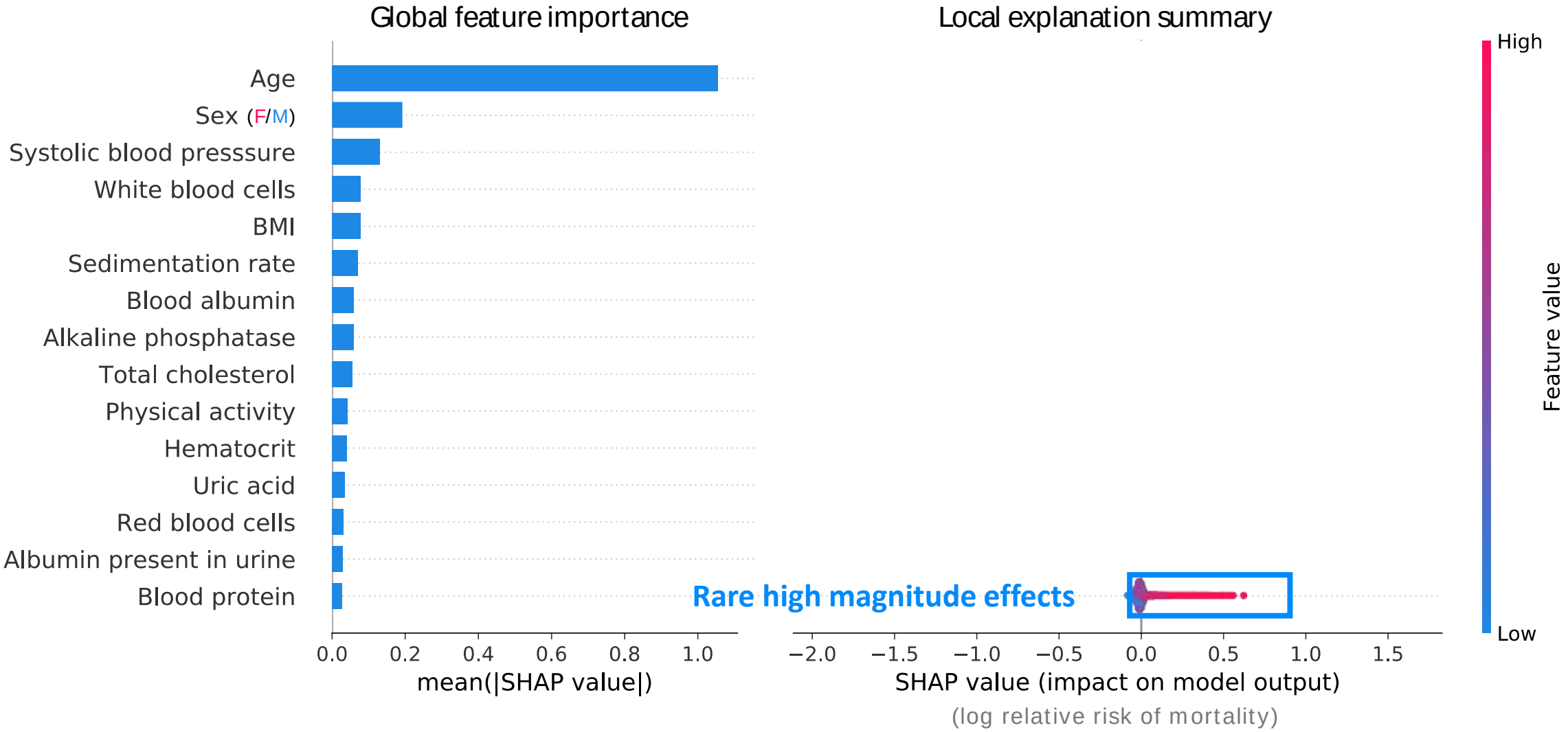


Conflates the prevalence of an effect with the magnitude of an effect

Reveal rare high-magnitude mortality effects



Reveal rare high-magnitude mortality effects



Reveal rare high-magnitude mortality effects