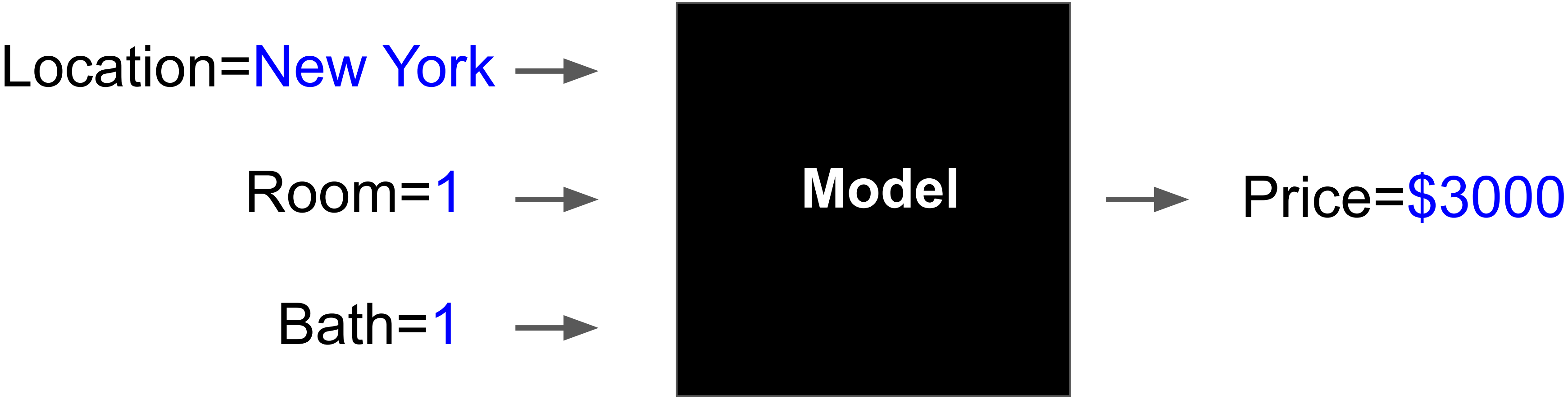
WeightedSHAP: analyzing and improving Shapley based feature attributions

Yongchan Kwon and James Zou

Feature attribution problem



→ **Goal**: quantify the **impact of each feature** for a particular model prediction

Marginal contributions and SHAP

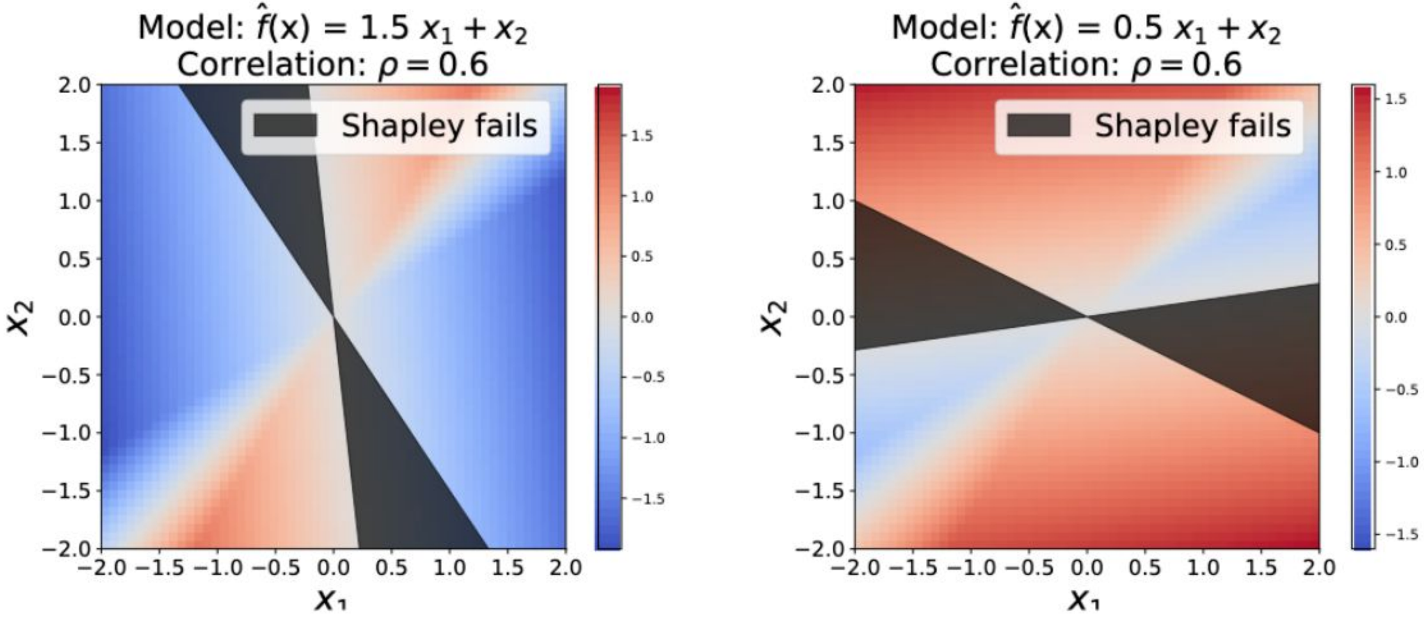
$$\Delta_j(x_i) := \text{Average of} \quad \underbrace{(f(S \cup \{x_i\}) - f(S))}_{\text{difference in model predictions}}$$

Considers every possible subset with $|S| = j$

$$\text{SHAP} = \frac{1}{d} \sum_{j=1}^{d} \Delta_j(x_i)$$

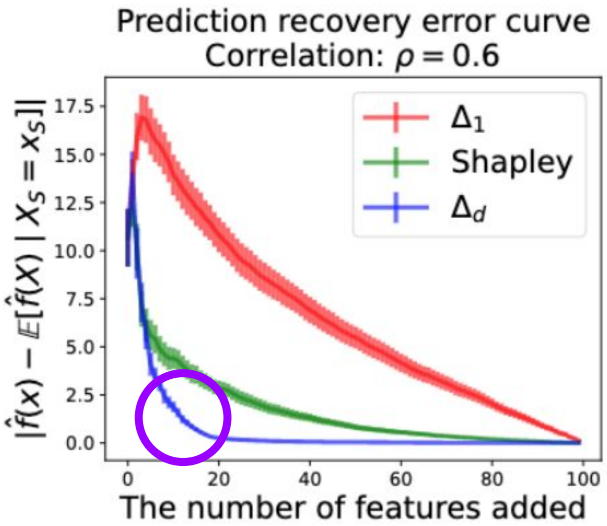
Question:
is **simple mean** optimal?

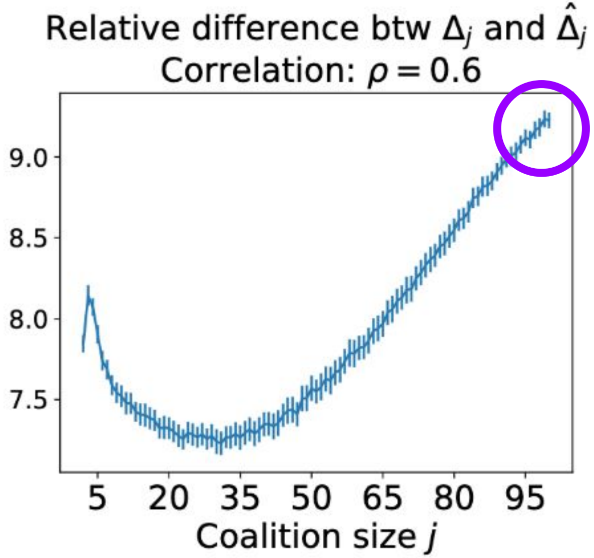
SHAP is *suboptimal*



On non-negligible areas, SHAP **fails** to find more influential features.

Why? Different marginal contributions have different signals and noises.





Marginal contributions with the largest coalition size is the **most effective to capture signals, but having largest estimation errors**.

Proposed idea: WeightedSHAP

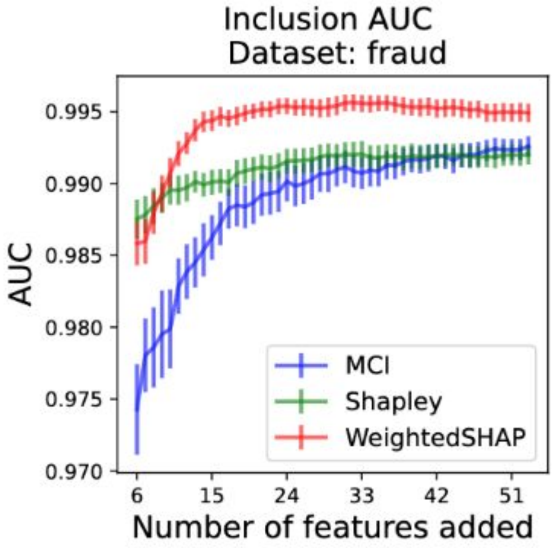
→ Find the **optimal weight** that optimizes some predefined utility function:

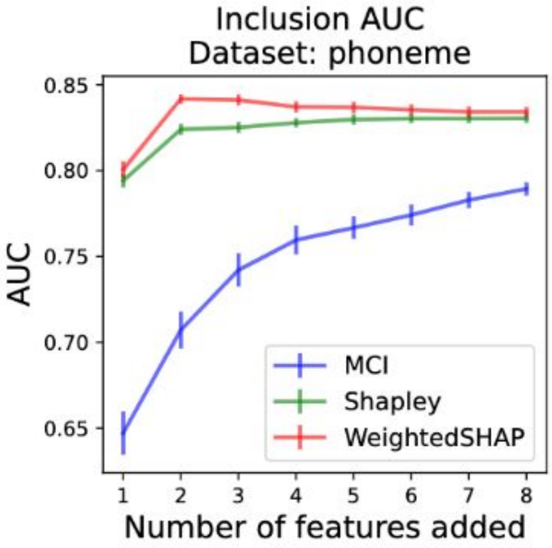
$$\max_{w} \text{Criteria}\left(\sum_{j=1}^{d} w_j \Delta_j(x_i)\right)$$

WeightedSHAP gives larger weights on more important marginal contributions while reducing estimation errors.

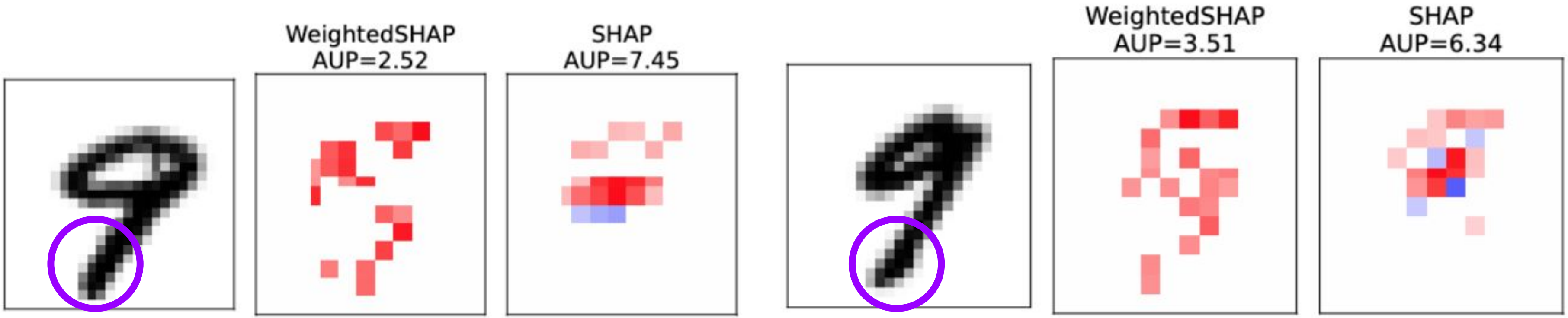
Feature addition experiment

→ WeightedSHAP **identifies influential features** and outperforms SOTA in recovery of the original model prediction.





Illustrative example: WeightedSHAP vs SHAP



Top 10% features selected by WeightedSHAP and SHAP.
→ WeightedSHAP provides **more intuitive explanations.**

Thank you for listening!

Key contributions

- We analyze suboptimality of SHAP
- A generalized attribution method WeightedSHAP

Easy-to-start Jupyter notebooks are available!