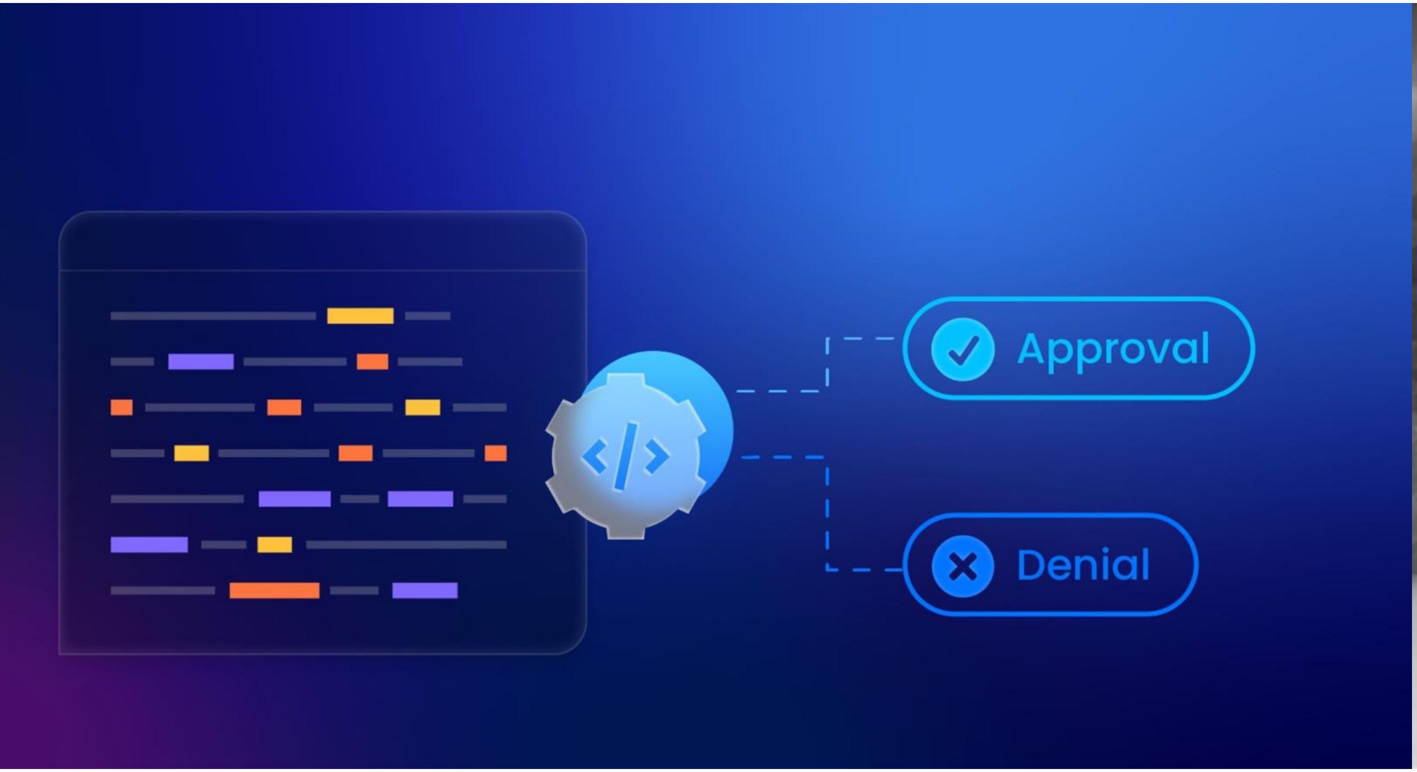
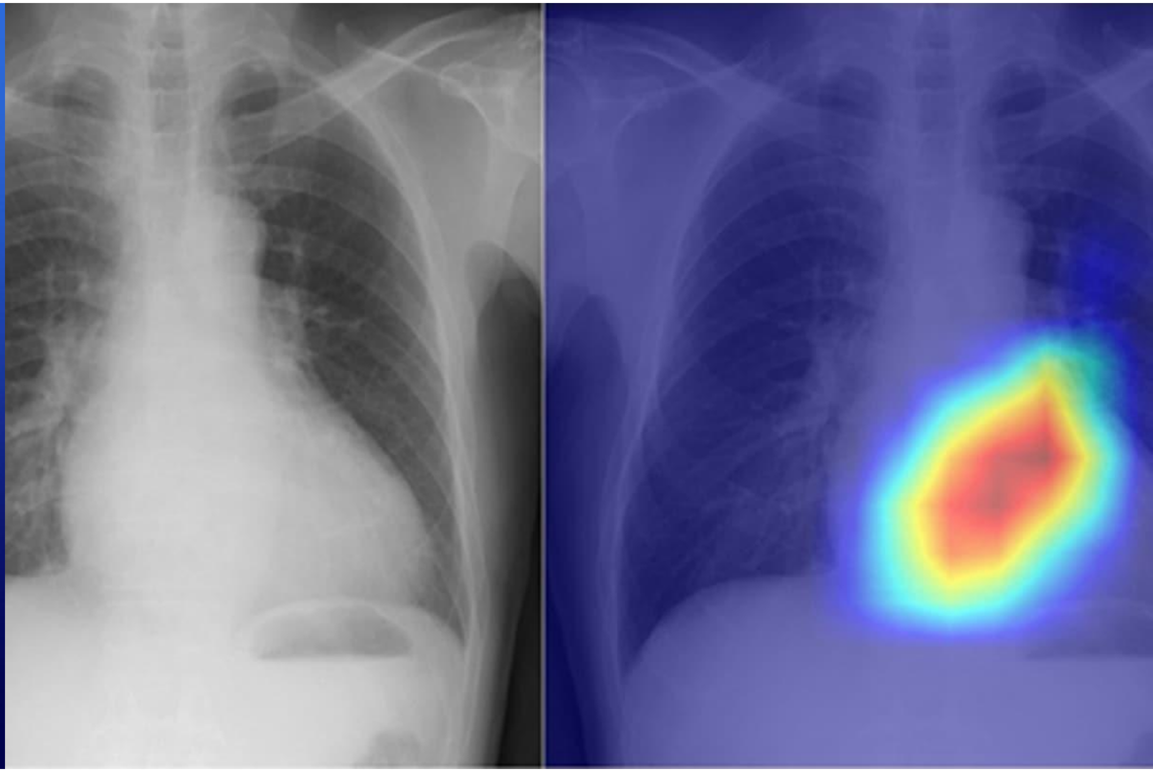
Provably Stable Feature Rankings with SHAP and LIME

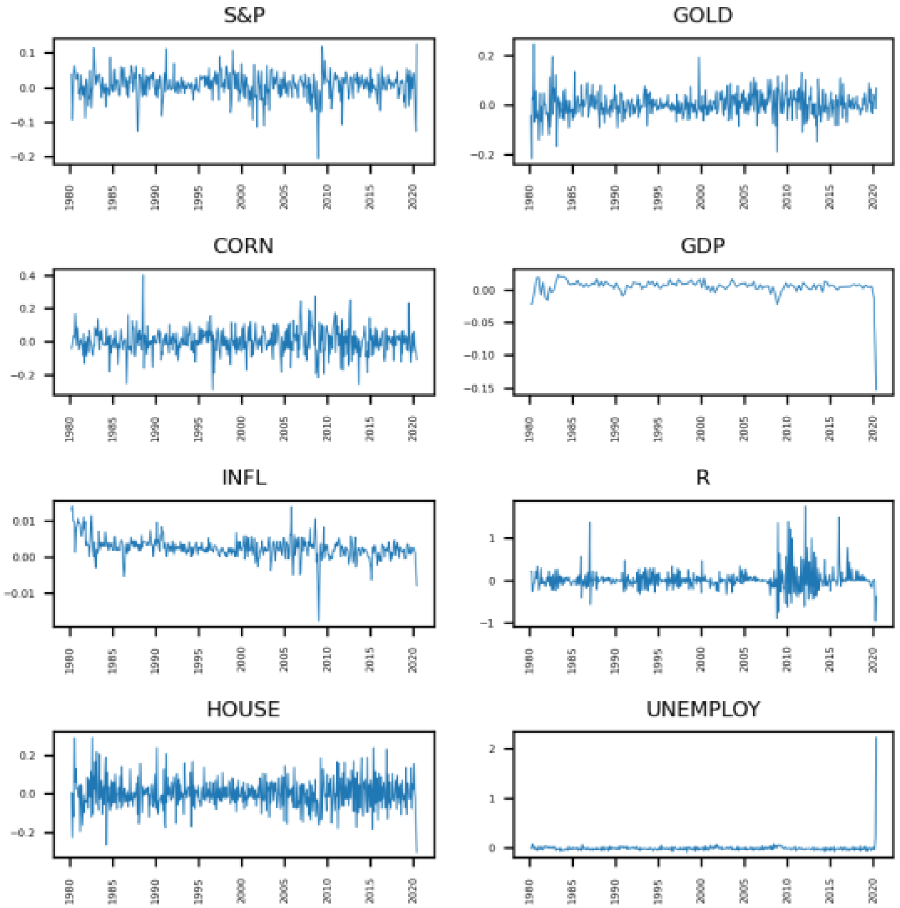
Jeremy Goldwasser

Why did the AI make this decision?

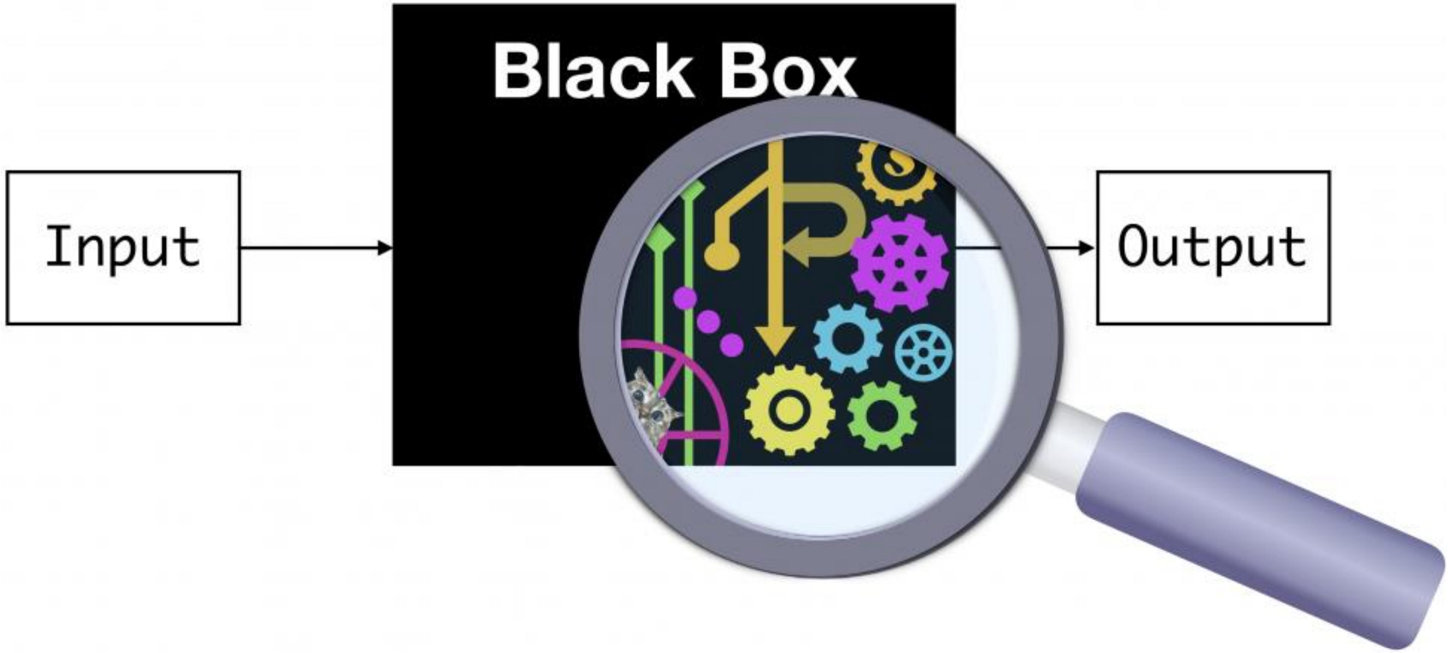




Learning from AI

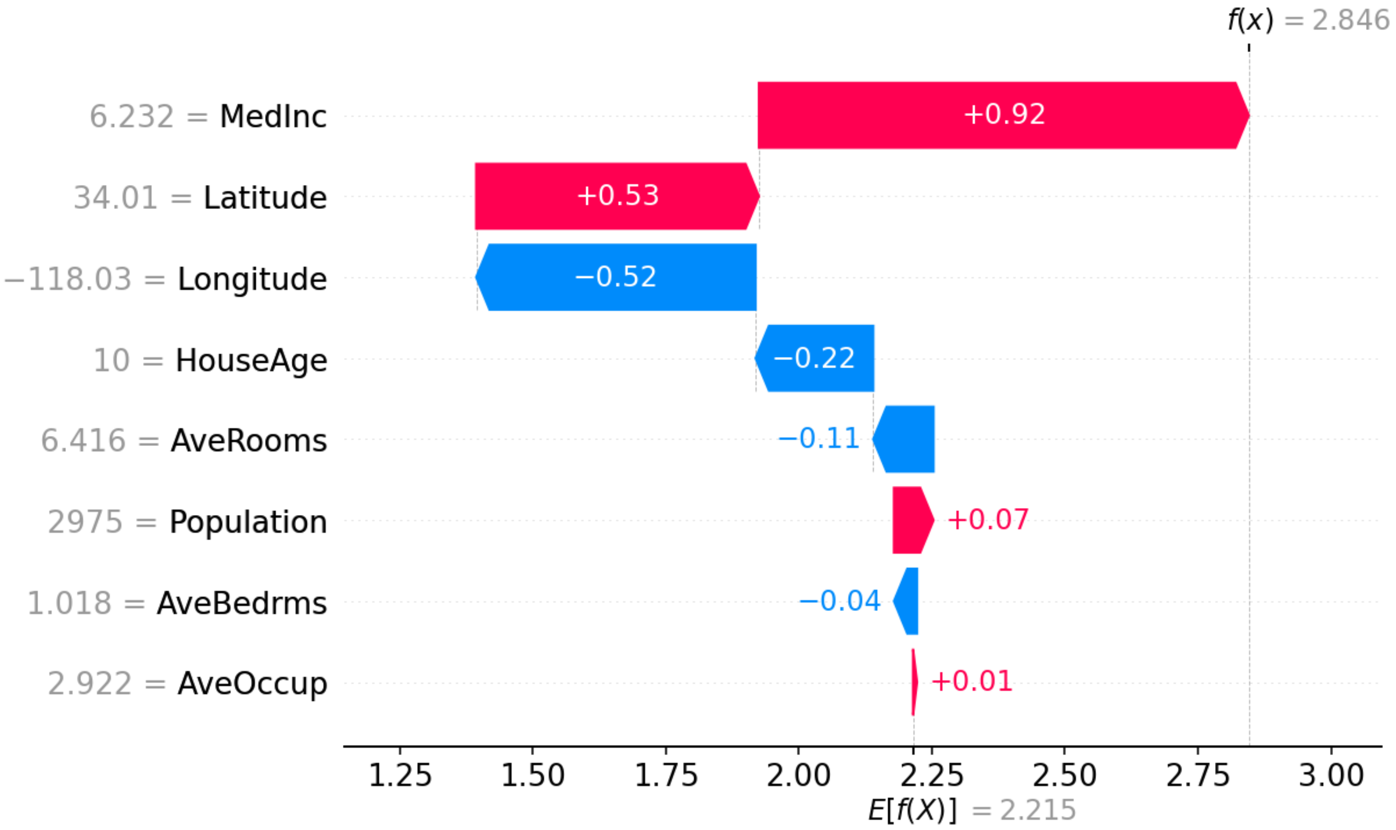


Local Explanations

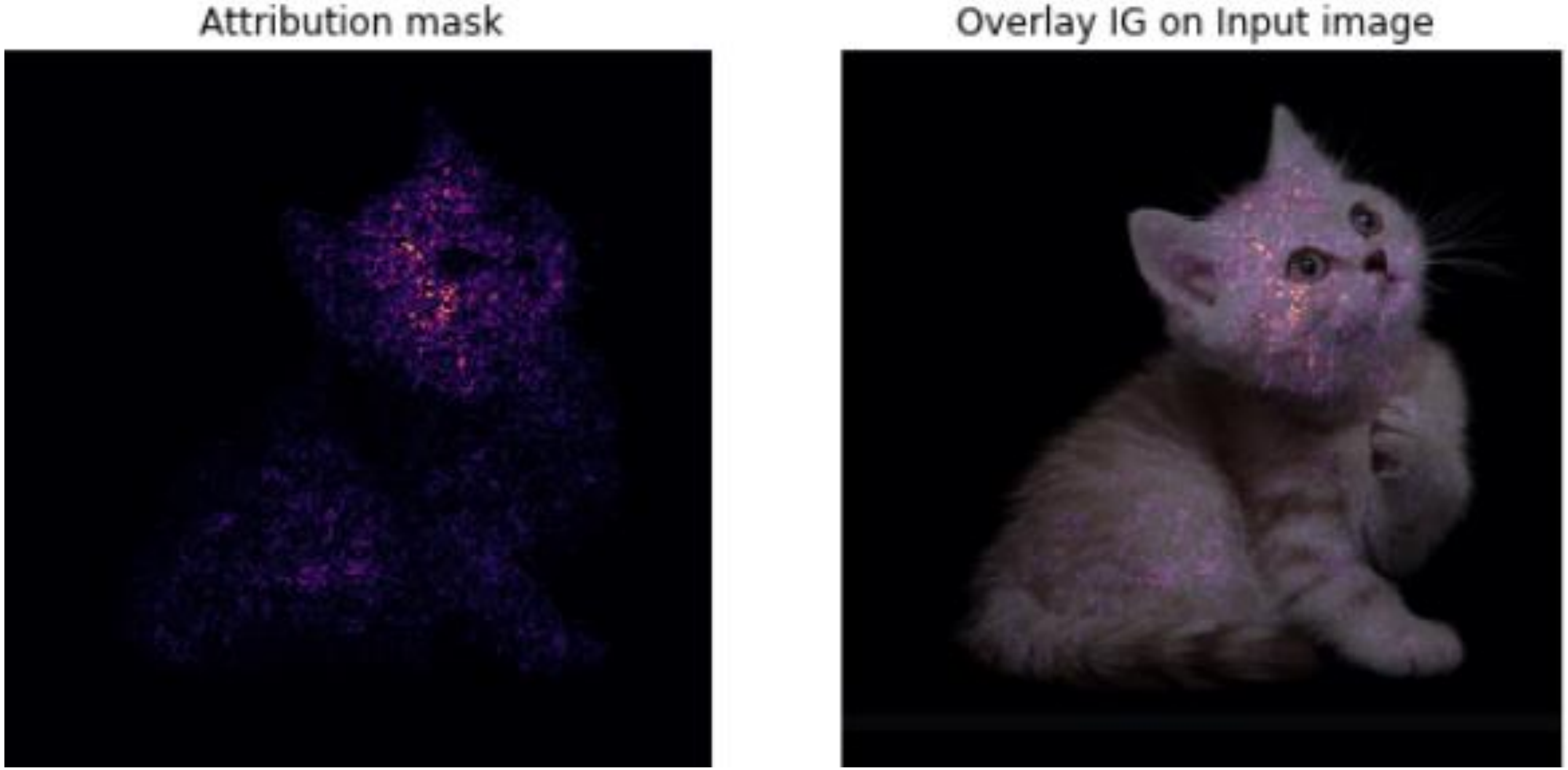


Why did our model predict y on input x?

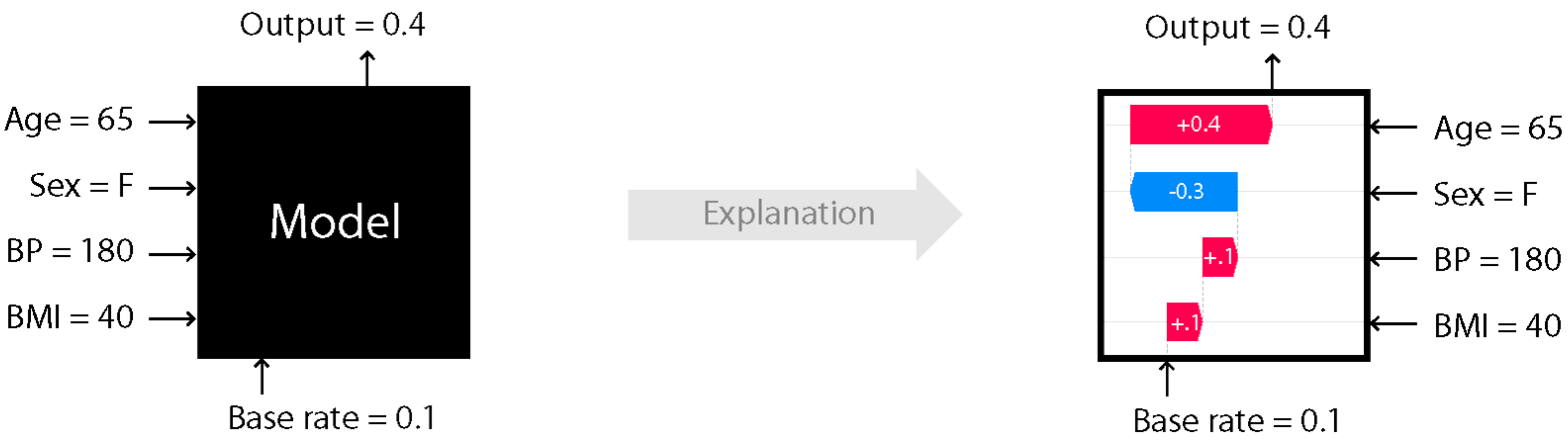
Feature Attributions



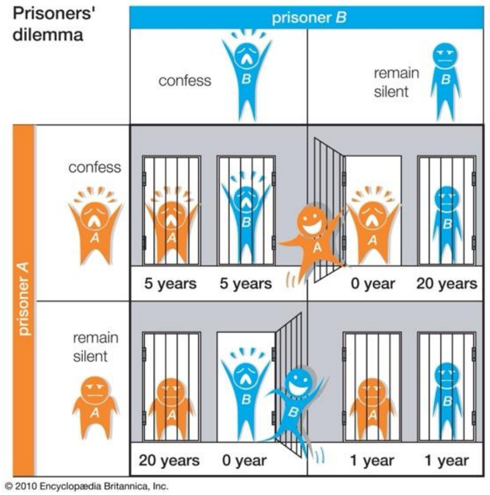
Gradient-Based Methods

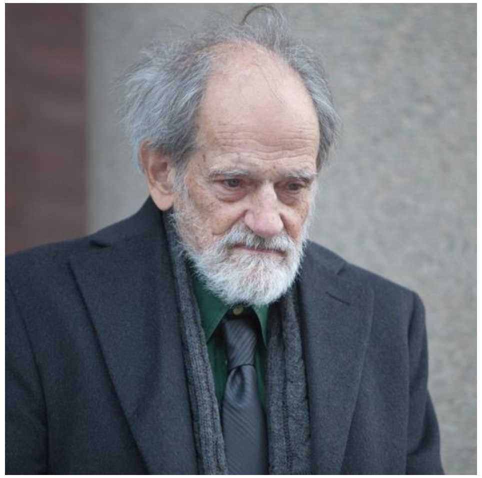


SHAP

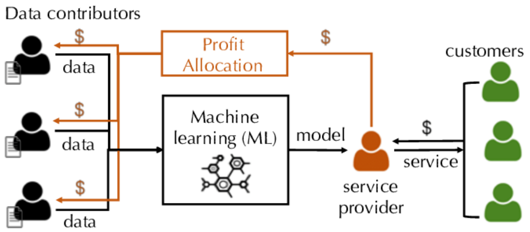


Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Model
Base rate = 0.1
Explanation
Output = 0.4
+0.4
-0.3
+.1
+.1
Age = 65
Sex = F
BP = 180
BMI = 40
Base rate = 0.1





Shapley Values



A Two-Minute Introduction to Shapley Values

How valuable was LeBron James to the 2022-2023 Lakers?


LAKERS
6
bibigo
CLUTCHPOINTS


20
bibigo
LAKERS
3


wish
LAKERS
23
LAKERS


bibigo
LAKERS
6


LAKERS
0

WESTBROOK
0

bibigo
LAKERS
6


LAKERS
6
bibigo
LAKERS
LAKERS

The Shapley Value

Intuition: A player's value depends on their contribution to coalitions of their teammates

$$\phi_j(v) := \sum_{S \in \{1,\ldots,d\} \setminus j} w_S * \{\text{marginal contribution of j to } S\}$$

$$:= \frac{1}{d} \sum_{S \subseteq [d] \setminus \{j\}} \binom{d-1}{|S|}^{-1} \left[v_x(S \cup \{j\}) - v_x(S) \right]$$

- Basketball, evaluate team given players in S
- SHAP, evaluate prediction given features in S: $v_x(S) = \hat{\mathbb{E}}[f(X)|X_S = x_S]$

Shapley Axioms

SHAP uniquely satisfies 3 desiderata for ML feature explanations:

1. **Local Accuracy**. Sum of importance scores is prediction
2. **Missingness**. Missing feature → score is 0
3. **Consistency**. Consider 2 models, A and B. If knowing a feature changes the prediction more on model A, then its score on A is at least as big as on B.

Shapley Estimation

$$\phi_j(v) := \sum_{S \in \{1,\ldots,d\} \setminus j} w_S * \{\text{marginal contribution of j to } S\}$$

$\mathcal{O}(2^d)$ terms may be computationally prohibitive

→ Use sampling-based approximation

Shapley Estimation

Shapley Value

$$\phi_j(v) := \frac{1}{d!} \sum_{\pi \in \Pi(d)} v(S_j^\pi \cup \{j\}) - v(S_j^\pi)$$

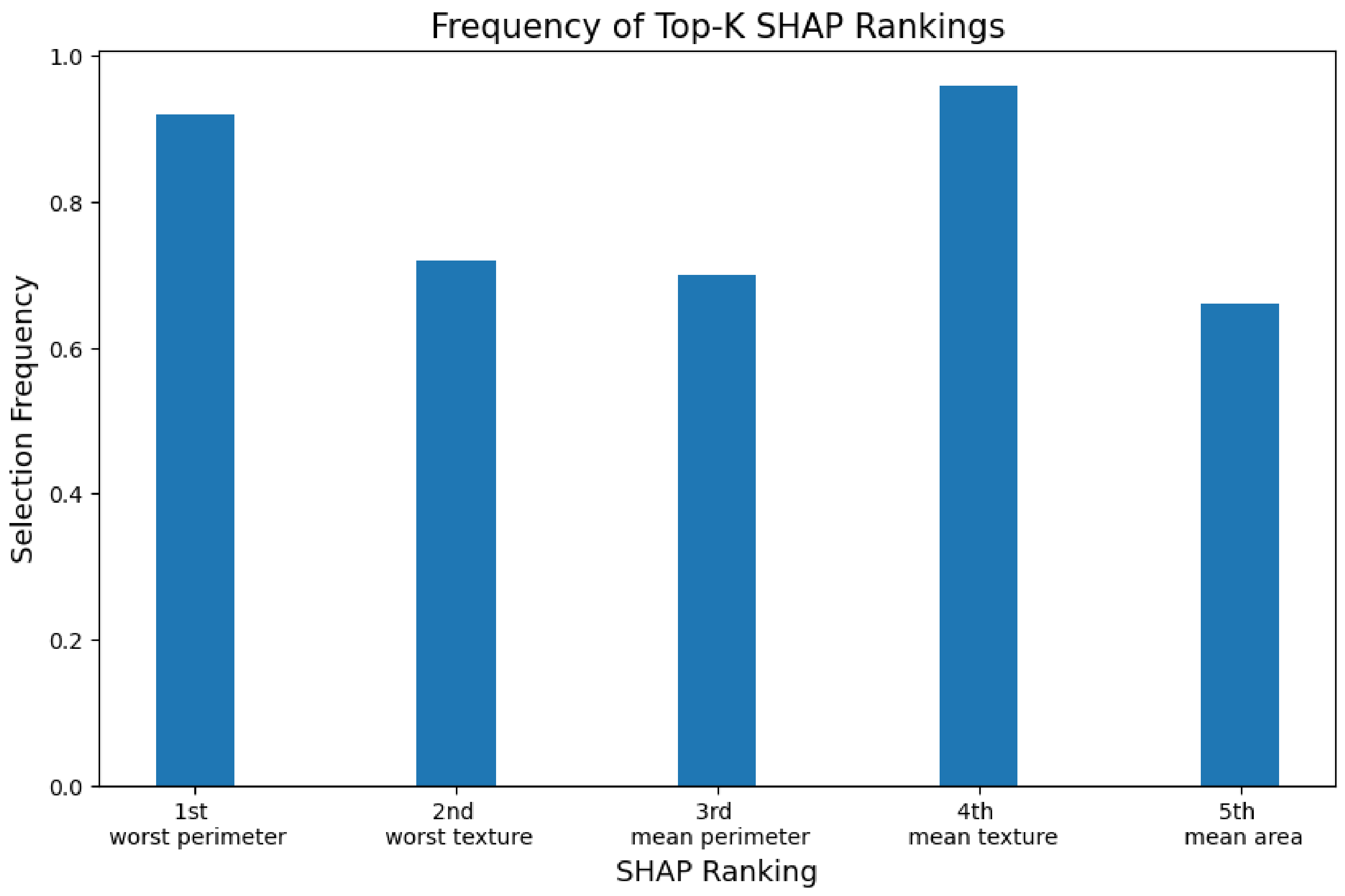
SHAP value function

$$v_x(S) = \hat{\mathbb{E}}[f(X)|X_S = x_S]$$

Shapley Sampling Estimate

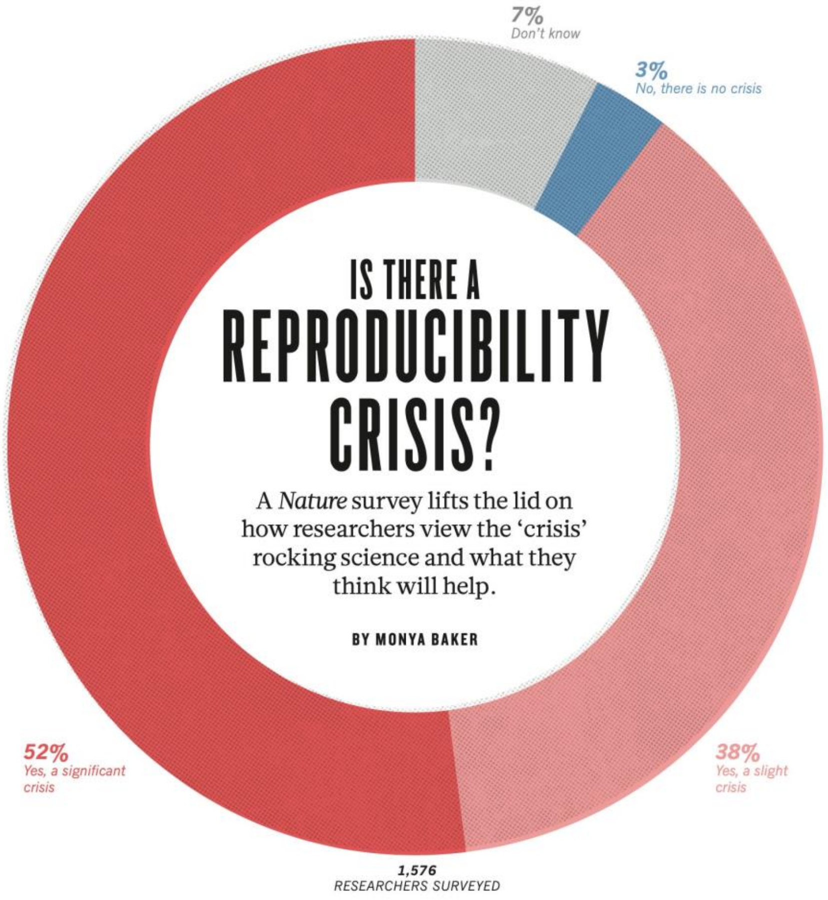
$$\hat{\phi}_j(v) = \frac{1}{n} \sum_{i=1}^{n} v(S_j^i \cup \{j\}) - v(S_j^i).$$

Instability with SHAP



“Stability is a common-sense principle and **a prerequisite for knowledge**. It is related to the notion of scientific reproducibility, which Fisher and Popper argued is **a necessary condition for establishing scientific results**.

Bin Yu and Karl Kumbier, *Veridical Data Science*, (PNAS, 2020)



Desideratum

Estimate an input's Shapley values such that the ranking of the top K attributions are correct with probability exceeding $1 - \alpha$

$$\iff \quad \{\phi_{(\hat{1})} \geq \phi_{(\hat{2})} \geq \ldots \geq \phi_{(\hat{K})} \geq \max_{j>K} \phi_{(\hat{j})}\}$$

$$\iff \quad \left\{ \bigcup_{i=1}^{K} \phi_{(\hat{i})} \geq \max_{j>i} \phi_{(\hat{j})} \right\} \qquad \text{w.p.} \quad \geq 1 - \alpha.$$

where $(\hat{1}), (\hat{2}), \ldots$ denotes the index of the 1^{st}, 2^{nd}, … estimated Shapley value.

e.g. Want to confirm 2 most important features ⇔ Confirm 1^{st} beats 2^{nd}, & 2^{nd} beats 3^{rd}.

Largest Shapley Value

To establish $(\hat{1})$ as the feature with the largest Shapley value, need to show it's bigger than $(\hat{2}), (\hat{3}), \ldots, (\hat{d})$.

- Multiple testing correction → lose power

Rank Verification for Exponential Families

Setting: Observe set of RVs from exponential family; want to rank population params

- To establish winner, only need to do two-tailed test with runner-up!
- Shapley Sampling asymptotically normal

Testing 1st vs 2nd place

Want resampled Shapley estimates to have same order

$$\hat{\phi}^*_{(\hat{1})} \geq \hat{\phi}^*_{(\hat{2})} \iff \hat{\Delta}^*_1 := \hat{\phi}^*_{(\hat{1})} - \hat{\phi}^*_{(\hat{2})} \geq 0$$

Shapley estimates converge to Gaussian

$$\hat{\Delta}_1 := \hat{\phi}_1 - \hat{\phi}_2 \overset{d}{\Rightarrow} \mathcal{N}\left(\phi_1 - \phi_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

$$\hat{\Delta}^*_1 - \hat{\Delta}_1 \overset{d}{\Rightarrow} \mathcal{N}\left(0, 2\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]\right)$$

$$\hat{\Delta}^*_1 | \hat{\Delta}_1 \overset{d}{\Rightarrow} \mathcal{N}\left(\phi_1 - \phi_2, 2\left[\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right]\right)$$

Testing 1st vs 2nd place

Obtain *P(resampled order different)* by integrating over normal tails

Reject if test statistic exceeds critical value. Two-tailed level-α test equivalent to one-tailed level-

Reject if test statistic exceeds critical value. Two-tailed level-α test equivalent to one-tailed level-$\frac{\alpha}{2}$

$$Z_{p_n} := \frac{\hat{\Delta}_1}{\sqrt{2\left[\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right]}} \geq Z_{1-\alpha/2}$$

K Largest Shapley Values

We have a way to establish largest Shapley value. What about 2^{nd}, 3^{rd}, …, K^{th}?

Rank Verification for Exponential Families

- Repeating procedure until failure to reject controls FWER
- So if 1^{st} beats 2^{nd}, 2^{nd} beats 3^{rd}, …, K^{th} beats $(K+1)^{th}$, then

$$P(\text{not all rankings correct}) \leq \alpha.$$

RankSHAP Algorithm

For each feature:

Get initial Shapley Sampling estimate

While $\geq$ one of the K pairwise comparison tests fails to reject:

Estimate # samples needed for test to reject

Run Shapley Sampling on those features for that many samples

RankSHAP Algorithm

For each feature:

Get initial Shapely Sampling estimate

While $\geq$ one of the K pairwise comparison tests fails to reject:

Estimate # samples needed for test to reject

Run Shapley Sampling on those features for that many samples

Theorem: *P(at least 1 top-K ranking error)* $\leq \alpha$

What if a test fails to reject?

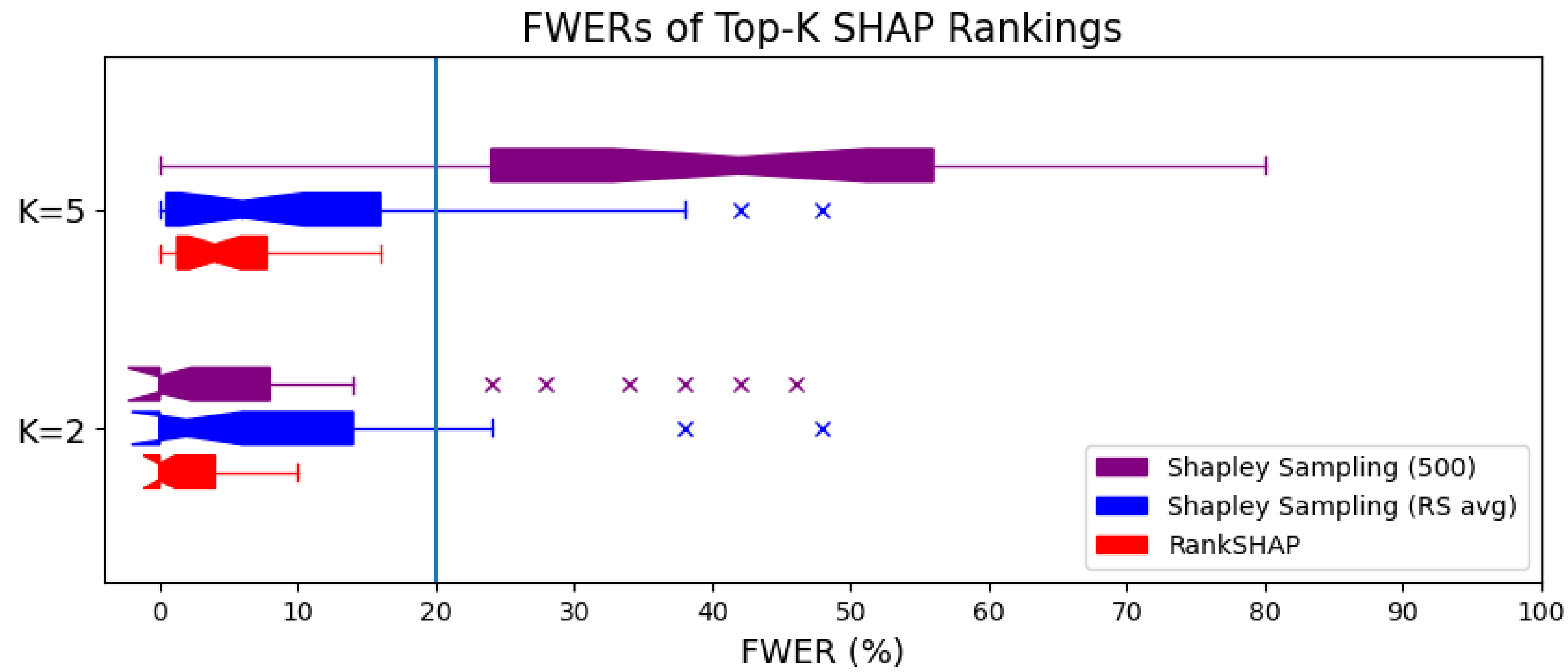
We use the test statistic

$$Z_{p_n} := \frac{\hat{\Delta}_1}{\sqrt{2\left[\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right]}}$$

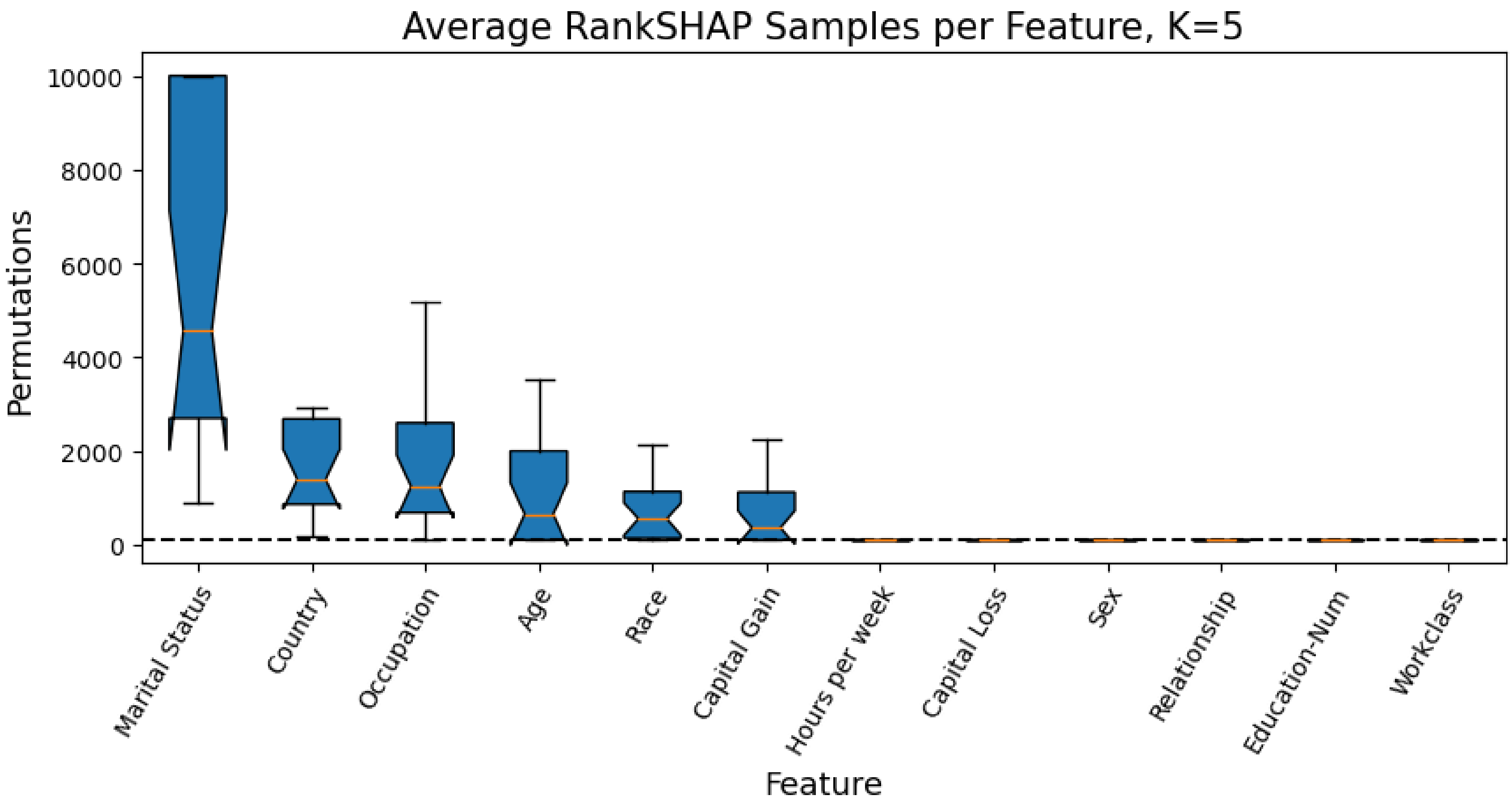
Solve for n_1, n_2 s.t. $Z_{p_n} \geq Z_{1-\alpha/2}$

e.g. forcing $n_1 = n_2 := n'$ yields $n' \geq 2\left(\frac{Z_{1-\alpha/2}}{\hat{\Delta}_k}\right)^2 (\hat{\sigma}^2_{(\hat{k})} + \hat{\sigma}^2_{(k\hat{+}1)}).$

Results


FWERs of Top-K SHAP Rankings
K=5
K=2
0
10
20
30
40
50
60
70
80
90
100
FWER (%)
Shapley Sampling (500)
Shapley Sampling (RS avg)
RankSHAP

	N	D	SHAP			
			K=3, Avg FWER	K=7, Avg FWER	K=3, Prop $<\alpha$	K=7, Prop $<\alpha$
`Adult`	32,561	12	2%	14%	1	0.8
`Bank`	45,211	16	6%	3%	1	1
`BRCA`	572	20	1%	7%	1	0.9
`WBC`	569	30	3%	10%	1	0.8
`Credit`	1,000	20	1%	6%	1	1



Average RankSHAP Samples per Feature, K=5
Permutations
10000
8000
6000
4000
2000
0
Feature
Marital Status
Country
Occupation
Age
Race
Capital Gain
Hours per week
Capital Loss
Sex
Relationship
Education-Num
Workclass

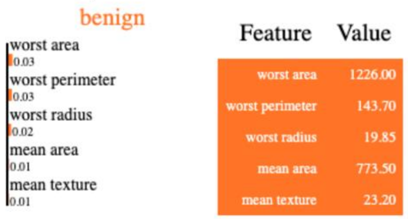
LIME

LIME & S-LIME

- To explain an input, LIME fits an interpretable model on samples randomly generated around it
  - K-LASSO selects K features in order
- At each step, S-LIME generates enough samples s.t. highest-“scoring” selection beats runner-up w.p. $\geq 1 - \alpha$

(a) Iteration 1 of LIME

(b) Iteration 2 of LIME



(c) Iteration 3 of LIME

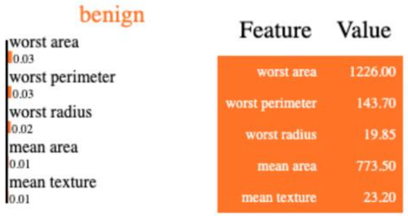
(d) Iteration 4 of LIME

(a) Iteration 1 of S-LIME

(b) Iteration 2 of S-LIME

(a) Iteration 1 of LIME

(b) Iteration 2 of LIME



(c) Iteration 3 of LIME

(d) Iteration 4 of LIME

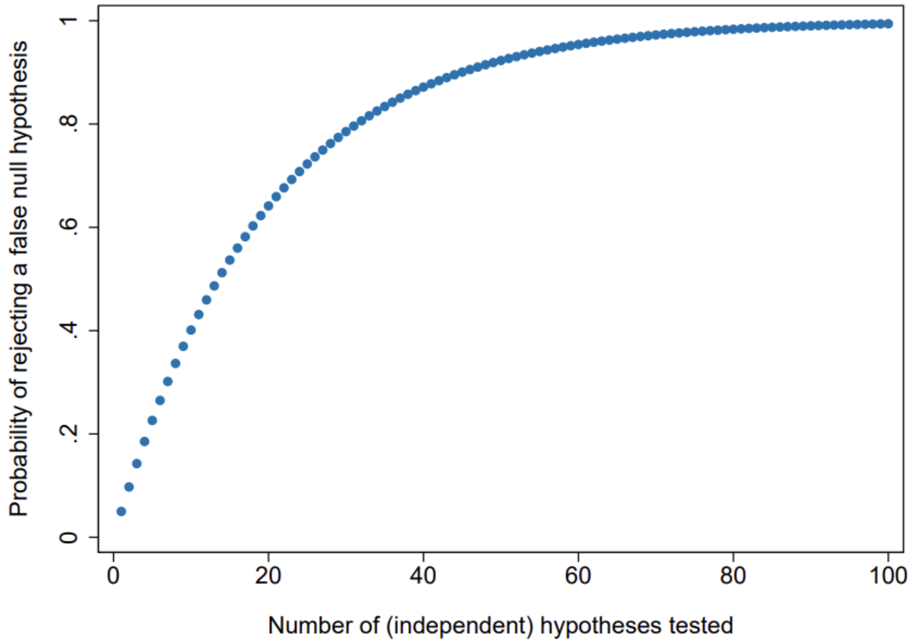
(a) Iteration 1 of S-LIME

(b) Iteration 2 of S-LIME

[LIME]() & [S-LIME]()

- To explain an input, LIME fits an interpretable model on samples randomly generated around it
  - K-LASSO selects K features in order
- At each step, S-LIME generates enough samples s.t. highest-"scoring" selection beats runner-up w.p. $\geq 1 - \alpha$
  - *This does not control FWER!*

Multiple Hypothesis Testing: The Problem



Under the null, probability of rejecting at least on hypothesis increases rapidly with number of independent hypothesis tests

Bonferroni Correction

FWER = P(at least 1 false rejection)

Bonferroni lowers significance threshold by a factor of # tests

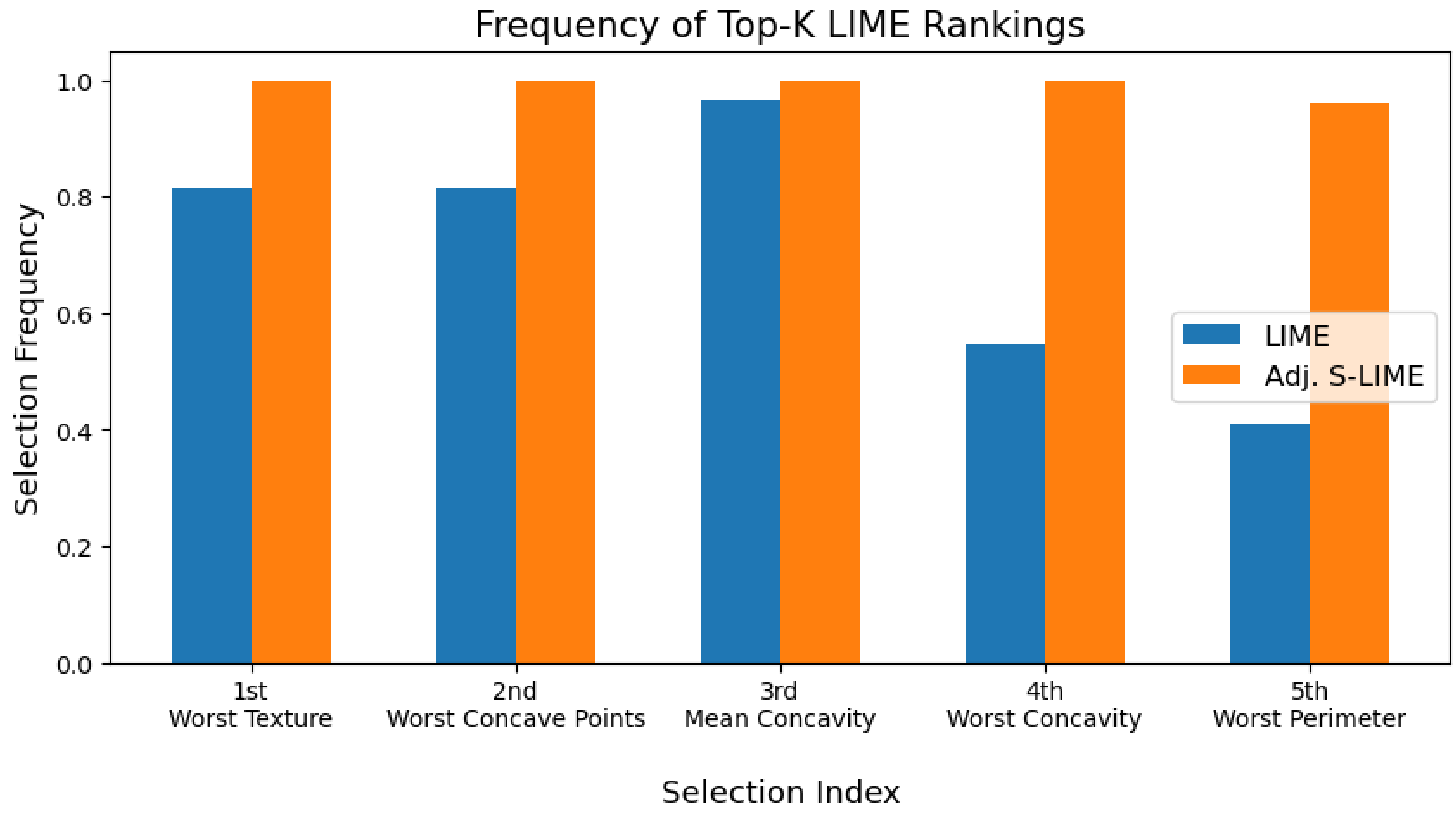
Bonferroni Correction

FWER = P(at least 1 false rejection)

Bonferroni lowers significance threshold by a factor of # tests

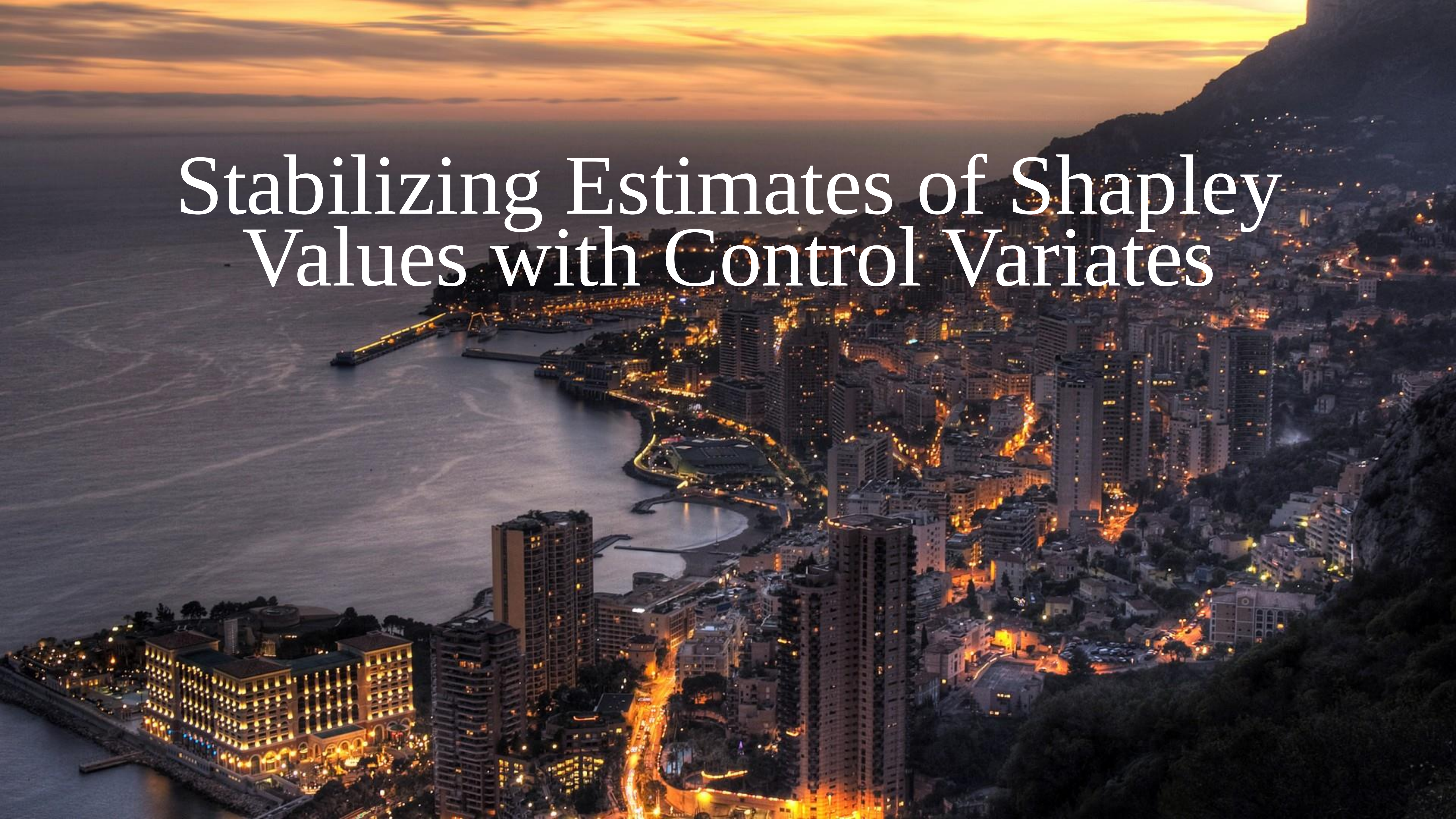
Running S-LIME w/ $\frac{\alpha}{2K}$ controls FWER = P(at least 1 incorrect ranking) $\leq \alpha$

Factor of 2 ensures selected feature beats all, not just runner-up


Frequency of Top-K LIME Rankings
Selection Frequency
1.0
0.8
0.6
0.4
0.2
0.0
LIME
Adj. S-LIME
1st
Worst Texture
2nd
Worst Concave Points
3rd
Mean Concavity
4th
Worst Concavity
5th
Worst Perimeter
Selection Index

	N	D	LIME			
			K=3, Avg FWER	K=7, Avg FWER	K=3, Prop $<\alpha$	K=7, Prop $<\alpha$
Adult	32,561	12	0%	NA	1	NA
Bank	45,211	16	0%	8%	1	0.9
BRCA	572	20	0%	12%	1	0.8
WBC	569	30	0%	NA	1	NA
Credit	1,000	20	0%	0%	1	1

Stabilizing Estimates of Shapley Values with Control Variates



Control Variates

Control Variates

- Given an estimator $\hat{A}$ that is unbiased for unknown target A^*
- Choose a correlated estimator $\hat{B}$, unbiased for known estimand B^*
- Define the control variates estimator $\tilde{A} \coloneqq \hat{A} - c(\hat{B} - B^*)$
  - For any c, $\tilde{A}$ is unbiased
- For $c^* = Cov(\hat{A}, \hat{B})/Var(\hat{B})$, $Var(\tilde{A})$ is reduced by a factor of $(1 - \rho^2_{\{\hat{A},\hat{B}\}})$ over $Var(\hat{A})$.
  - Here, variance = MSE

Shapley Estimation

Shapley Value

$$\phi_j(v) := \frac{1}{d!} \sum_{\pi \in \Pi(d)} v(S_j^\pi \cup \{j\}) - v(S_j^\pi)$$

Shapley Sampling Estimate

$$\hat{\phi}_j(v) = \frac{1}{n} \sum_{i=1}^{n} v(S_j^i \cup \{j\}) - v(S_j^i).$$

KernelSHAP Estimate

$$\hat{\phi}(x) = argmin_{\phi \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(\phi^T z^i - (v_x(z^i) - v_x(\mathbf{0}))\right)^2$$

Constructing a Control Variate

SHAP value function

$$v_x(S) = \hat{\mathbb{E}}[f(X)|X_S = x_S]$$

- We need an estimator that 1) is positively correlated with the Shapley estimate, and 2) has known estimand
- Each of the n samples indexes a random subset of features. The control variate will use these same subsets – but for a different predictor: the **Taylor Expansion** of *f* around *x*
- What order expansion it is depends on how $X_{-S}|X_S$ is sampled in the value function

A Control Variate for the Shapley Value (1/2)

Theorem 1. *Define* $\mu := \mathbb{E}[X]$ *and* $\Sigma_{jk} := Cov(X_j, X_k)$. *Let the value function* $v_x(S)$ *compute conditional means by sampling each feature from its marginal distribution. The* j^{th} *Shapley value of the second-order Taylor approximation of* f *around* x *is*

$$\begin{aligned}\phi_j^{approx}(x) = \frac{\partial f}{\partial x_j}(x_j - \mu_j) \\ - \frac{1}{2}\left[\sum_{k=1}^{d}(x_k - \mu_k)\frac{\partial^2 f}{\partial x_j x_k}\right](x_j - \mu_j) \\ - \frac{1}{2}\sum_{k=1}^{d}\Sigma_{jk}\frac{\partial^2 f}{\partial x_j x_k}.\end{aligned}$$

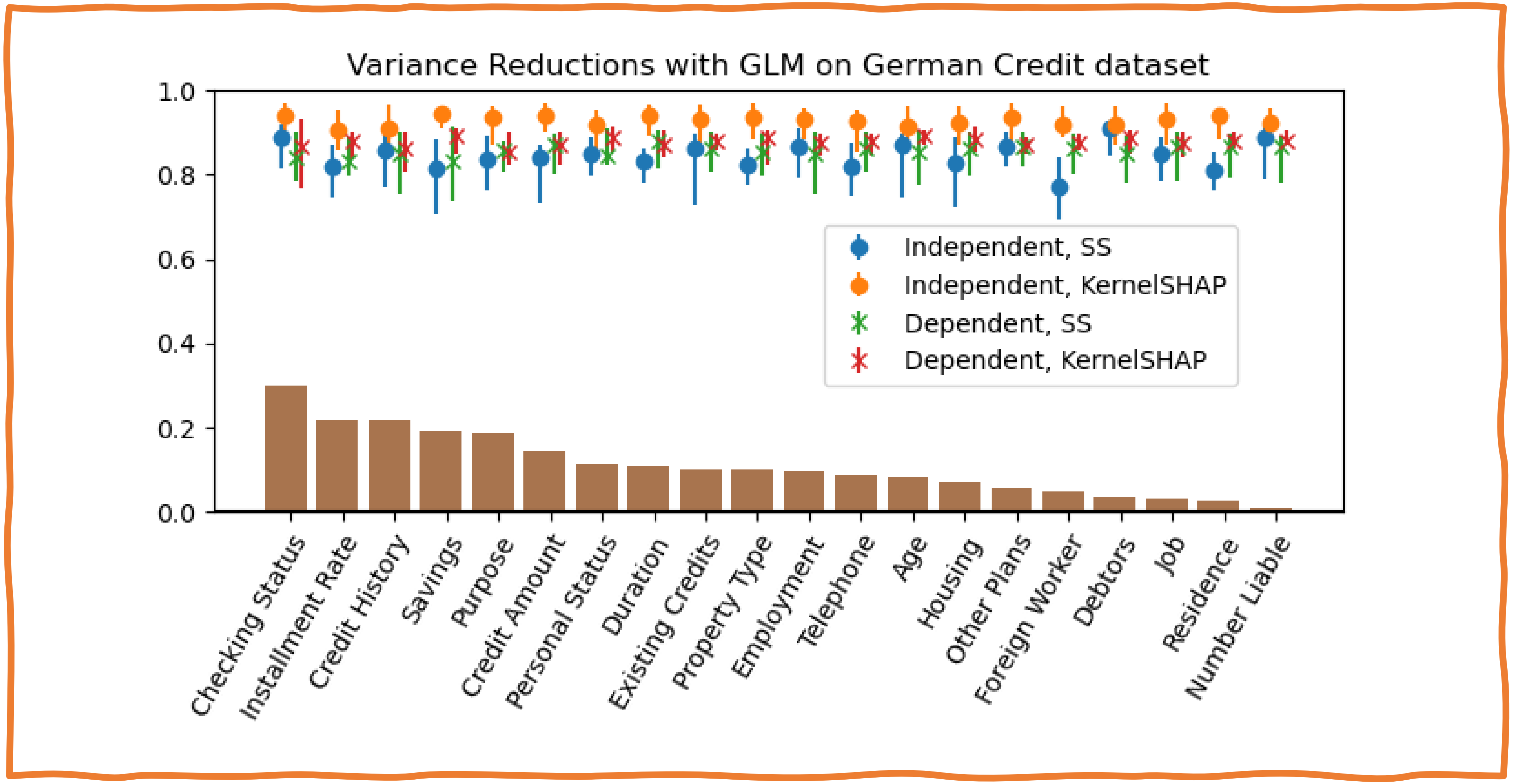
A Control Variate for the Shapley Value (2/2)

Theorem 2. *Define the value function to take the conditional expectation of* $f(X)|X_S$ *over a multivariate normal distribution. Let* $S^m \subseteq [d] \backslash \{j\}$ *be the subset of features appearing before* j *in the* m^{th} *permutation of* $[d]$*, and let* Q_S *and* R_S *be matrices defined in Section 1.2 of the supplementary material that do not depend on* x*. Define*

$$D_j := \frac{1}{d!} \sum_{m=1}^{d!} \left([Q_{S_j^m \cup j} + R_{S_j^m \cup j}] - [Q_{S_j^m} + R_{S_j^m}] \right).$$

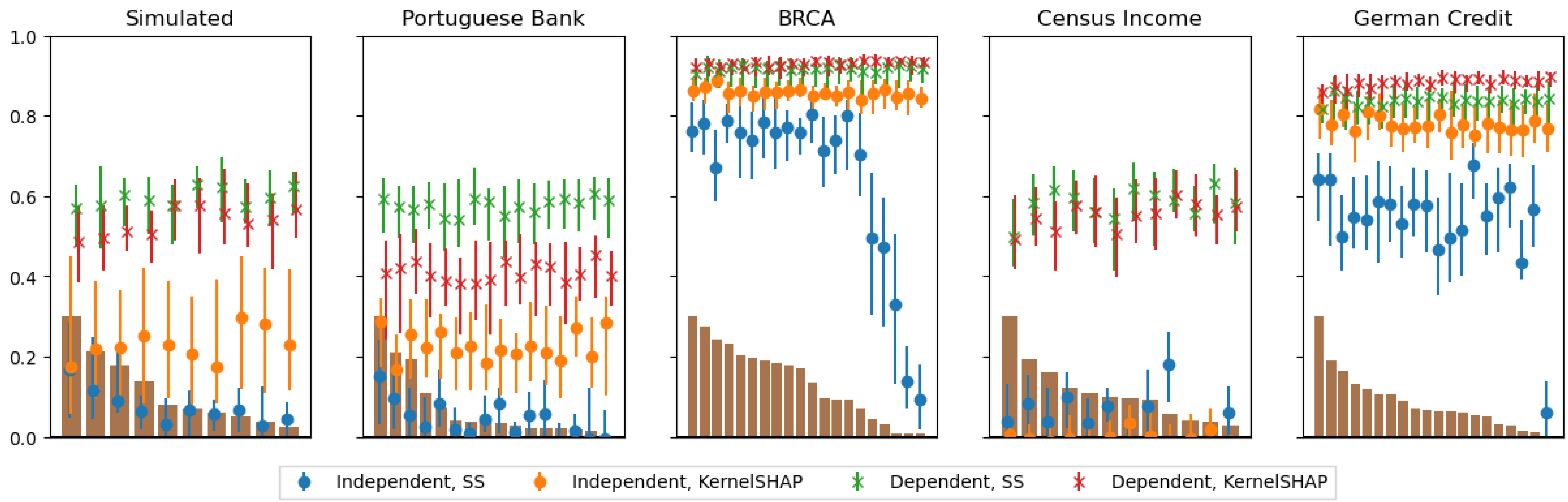
The j^{th} *Shapley value of the first-order Taylor approximation of* f *around* x *is*

$$\phi_j^{approx}(x) = \nabla f(x)^T D_j (x - \mu).$$

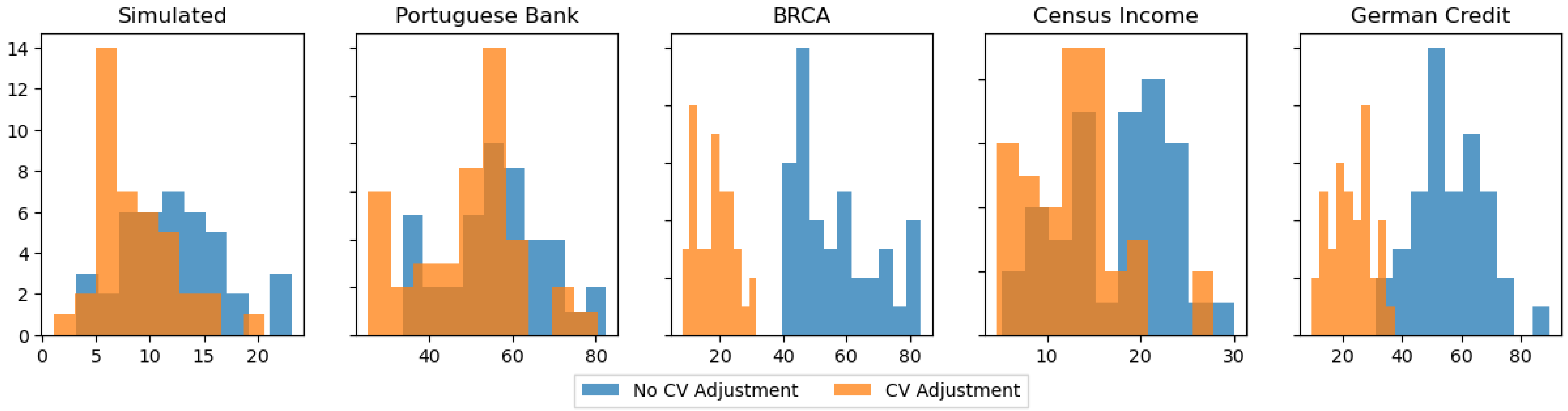


Variance Reductions with GLM on German Credit dataset
1.0
0.8
0.6
0.4
0.2
0.0
Independent, SS
Independent, KernelSHAP
Dependent, SS
Dependent, KernelSHAP
Checking Status
Installment Rate
Credit History
Savings
Purpose
Credit Amount
Personal Status
Duration
Existing Credits
Property Type
Employment
Telephone
Age
Housing
Other Plans
Foreign Worker
Debtors
Job
Residence
Number Liable

ControlSHAP Variance Reductions, Neural Net



Importance Rankings More Stable



Conclusion

- ML explanations must be trustworthy to be useful
- Identify most important features with statistical guarantees for SHAP & LIME

Next Steps

- Keep computational work with sequential testing
- Local → Global Shapley feature importances

Thanks for your attention!