Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations

Presented By:

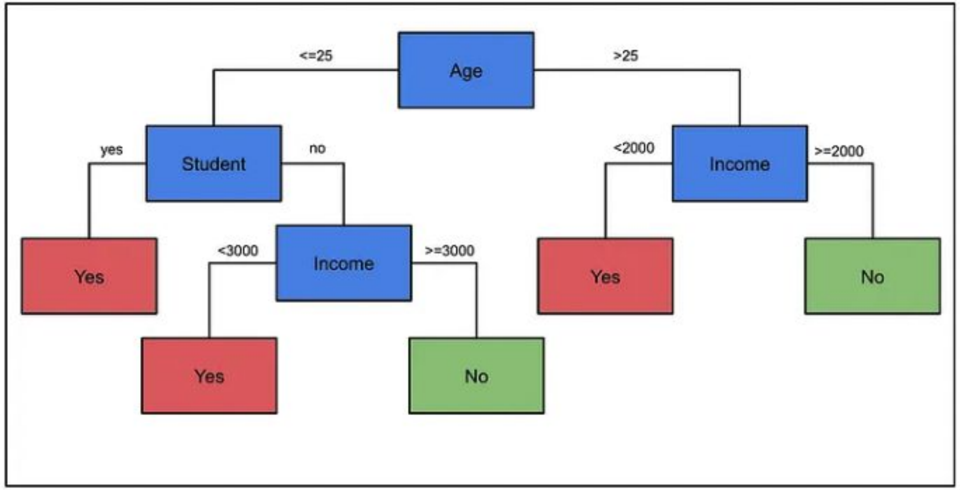
Eric Hansen
Rohan Doshi
Leo Benac

Authored By:

Tessa Han
Suraj Srinivas
Himabindu Lakkaraju

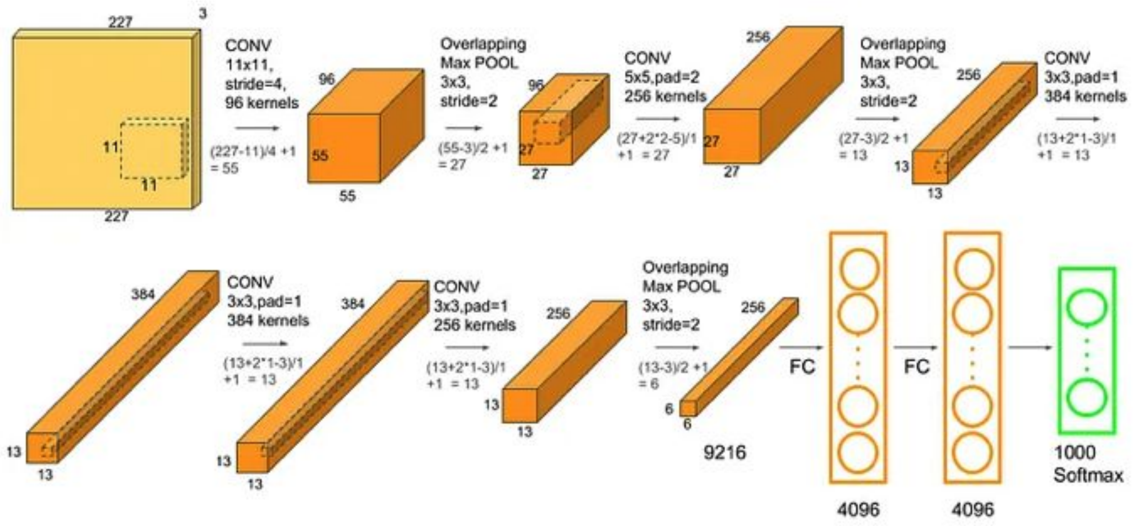
Interpretability

- Machine Learning models makes decision in high stake settings (healthcare, law, finance)
- Growing emphasis on understanding how models make predictions



Explainability

- Dealing with complicated models requires explainability through post hoc explanations



Post Hoc Explanation Methods

- LIME
- C-LIME
- SHAP
- Occlusion
- Vanilla Gradients
- Gradient x Input
- SmoothGrad
- Integrated Gradients

Different Methods Different Explanations

- Inconsistency in their goals

- What is an explanation?

- Lack of Mathematical Framework resulting in misunderstanding of their goals and properties

Summary of Key Contributions

1) Formalize the local function approximation (LFA) framework

2) Demonstrate that eight popular explanation methods can be characterized as instances of the LFA framework

3) No Free Lunch Theorem: no single LFA method can perform optimally across all noise neighbourhoods

4) Provide a guiding principle for choosing among LFA explanation methods based on the input domain

Related Work: Connections and Properties among post hoc explanation methods.

- C-LIME and SmoothGrad connections
- Gradient-based explanation methods and when they produce similar explanations.
- Faithfulness to the black-box model, robustness to adversarial attack, and fairness across subgroups.

Local Function Approximation Framework

Definition 1. ***Local function approximation*** *(LFA) of a black-box model* f *on a neighbourhood distribution* $\mathcal{Z}$ *around* $\mathbf{x}_0$ *by an interpretable model class* $\mathcal{G}$ *and a loss function* ℓ *is given by*

$$g^* = \arg\min_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{\xi \sim \mathcal{Z}} \ell(f, g, \mathbf{x}_0, \xi) \tag{1}$$

where a valid loss ℓ *is such that* $\mathbb{E}_{\xi \sim \mathcal{Z}} \ell(f, g, \mathbf{x}_0, \xi) = 0 \iff f(\mathbf{x}_\xi) = g(\mathbf{x}_\xi) \ \ \forall \xi \sim \mathcal{Z}$

Distinction with LIME

- The LFA framework requires that f and g share the same input domain X and output domain Y, suggesting LIME does not fall in LFA
- LFA framework ensures model recovery if $f \in G$ and domain(x) = X
- Optimization is done through splitting the perturbation data into train / validation / test sets. LIME does not, making it possible to overfit a small number of perturbations

Designing Explanations with LFA

LFA guides creation of explanations

LFA requires you define

1. interpretable model class ***G***
2. neighbourhood distribution ***Z***
3. loss function ***l***
4. binary operator $\oplus$ to combine the input and the noise

Correspondence with others explanation methods

Explanation Method	Local Neighbourhood $\mathcal{Z}$ around $\mathbf{x}_0$	Loss Function ℓ
C-LIME	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Squared Error
SmoothGrad	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Gradient Matching
Vanilla Gradients	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2), \sigma \to 0$	Gradient Matching
Integrated Gradients	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(0, 1)$	Gradient Matching
Gradients × Input	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(a, 1), a \to 1$	Gradient Matching
LIME	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0, 1\}^d) \sim \text{Exponential kernel}$	Squared Error
KernelSHAP	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0, 1\}^d) \sim \text{Shapley kernel}$	Squared Error
Occlusion	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0, 1\}^d) \sim \text{Random one-hot vectors}$	Squared Error

LFA with Continuous Noises: Gradient-Based Methods

Gradient Matching Loss:

$$\ell_{gm}(f, g, \mathbf{x}_0, \xi) = \|\nabla_\xi f(\mathbf{x}_0 \oplus \xi) - \nabla_\xi g(\mathbf{x}_0 \oplus \xi)\|_2^2$$

No Free Lunch Theorem for Explanation Methods

Theorem 3 (No Free Lunch for Explanation Methods)**.** *Consider explaining a black-box model f around point $\mathbf{x}_0$ using an interpretable model g from model class $\mathcal{G}$ and a valid loss function ℓ where the distance between f and $\mathcal{G}$ is given by* $d(f, \mathcal{G}) = \min_{g \in \mathcal{G}} \max_{\mathbf{x} \in \mathcal{X}} \ell(f, g, 0, \mathbf{x})$.

Then, for any explanation g^* over a neighbourhood distribution $\xi_1 \sim \mathcal{Z}_1$ such that $\max_{\xi_1} \ell(f, g^*, \mathbf{x}_0, \xi_1) \leq \epsilon$, *there always exists another neighbourhood $\xi_2 \sim \mathcal{Z}_2$ such that* $\max_{\xi_2} \ell(f, g^*, \mathbf{x}_0, \xi_2) \geq d(f, \mathcal{G})$.

When ***g*** is less expressive than **f**, no single explanation **g*** can perform optimally across all neighborhoods

- Example: **f** is non-linear and **g*** is linear

Characterizing Explanation Methods via Model Recovery

Definition 2 (Model Recovery: Guiding Principle). *Given an instance of the LFA framework with a black-box model f such that $f \in \mathcal{G}$ and a specific noise type (e.g., Gaussian, Uniform), an explanation method performs model recovery if there exists some noise distribution $\mathcal{Z}$ such that LFA returns $g^* = f$.*

How can you evaluate whether an explanation ***g*** works for ***f***?

If ***f*** and ***g**** are the same model class, it should be possible for ***g*** to approximate ("recover") ***f***

Characterizing Explanation Methods via Model Recovery

Let's explore which of the 8 explanation work for various input domains ***X***. Three cases:

1. continuous ***X***
2. binary ***X***
3. discrete ***X***

1. Which Explanation for continuous *X*? *(1/2)*

Assume f and g are linear ($f(\mathbf{x}) = \mathbf{w}_f^\top \mathbf{x}$ *and* $g(\mathbf{x}) = \mathbf{w}_g^\top \mathbf{x}$,)

A) ✅ Additive continuous noise (SmoothGrad, Vanilla Gradients, C-LIME). Why? $w_g = w_f$

$$\ell(f, g, \mathbf{x}_0, \xi) = \|\nabla_\xi f(\mathbf{x}_\xi) - \nabla_\xi g(\mathbf{x}_\xi)\|_2^2$$

B) ❌ Multiplicative continuous noise (Integrated Gradients and Gradient x Input). Why? loss function parameterization

$$\ell(\bar{f}, g, \mathbf{x}_0, \xi) = \|\nabla_\xi f(\mathbf{x}_\xi) - \nabla_\xi g(\xi)\|_2^2.$$

1. Which Explanation for continuous *X*? (2/2)

C) ❌ Multiplicative binary noise (LIME, KernelSHAP, and Occlusion). Why? Consider sinusoidal example

Remark 2. *For* $\mathcal{X} = \mathbb{R}^d$, *periodic functions* f *and* g *where* $f(\mathbf{x}) = \sum_{i=1}^{d} \sin(\mathbf{w}_{f_i} \odot \mathbf{x}_i)$ *and* $g(\mathbf{x}) = \sum_{i=1}^{d} \sin(\mathbf{w}_{g_i} \odot \mathbf{x}_i)$, *and an integer* n, *binary noise methods do not perform model recovery for* $|w_{f_i}| \geq \frac{n\pi}{\mathbf{x}_{0_i}}$.

$$\sin(\mathbf{w}_{f_i}\mathbf{x}_{0_i}) = \sin(\pm n\pi) = \sin(0) = 0$$

$\sin(\mathbf{w}_{f_i}\mathbf{x}_{0_i})$ outputs zero for all binary perturbations

discrete nature of noise makes model recovery impossible

2. Which Explanation for binary *X*?

Consider binary noise methods (continuous noise invalid)

Only Multiplicative binary perturbations methods (LIME, KernelSHAP, and Occlusion) enable ***g*** to recover ***f*** in the binary domain

3. Which Explanation for discrete *X*?

- ❌ continuous noise methods invalid
- ❌ binary noise methods - same sinusoidal logic
- All 8 explanations fail
- Use LFA to design new explanation with a discrete noise type

Designing Explanations with LFA

LFA guides creation of explanations

LFA requires you define

1. interpretable model class ***G***
2. neighbourhood distribution ***Z***
3. loss function ***l***
4. binary operator $\oplus$ to combine the input and the noise

Summary of Properties of Existing Explanation Methods

For continuous data

Method	Characteristics of ξ	g recovers f?	Scale of g's weights when $\mathcal{X} \in \mathbb{R}^d$
C-LIME	Continuous, Additive	When $\mathcal{X} \in \mathbb{R}^d$	Gradient
SmoothGrad	Continuous, Additive	When $\mathcal{X} \in \mathbb{R}^d$	Gradient
Vanilla Gradients	Continuous, Additive	When $\mathcal{X} \in \mathbb{R}^d$	Gradient
Integrated Gradients	Continuous, Multiplicative	No	Gradient $\times$ Input
Gradients $\times$ Input	Continuous, Multiplicative	No	Gradient $\times$ Input
LIME	Binary, Multiplicative	When $\mathcal{X} \in \{0,1\}^d$	Gradient $\times$ Input
KernelSHAP	Binary, Multiplicative	When $\mathcal{X} \in \{0,1\}^d$	Gradient $\times$ Input
Occlusion	Binary, Multiplicative	When $\mathcal{X} \in \{0,1\}^d$	Gradient $\times$ Input

Table 2: Summary of properties of existing explanation methods in relation to the LFA framework. In this table, we consider the scale of g's weights when $\mathcal{X} \in \mathbb{R}^d$.

For binary data

Empirical Evaluation - Dataset & Models

1) Common Experimental Goal: Examine different XAI methods' explanations of models' predictions on sample datasets to validate theoretical claims

2) Two Datasets
   a) World Health Organization Life Expectancy dataset (20 features)
   b) Home Equity Line of Credit (HELOC) dataset from FICO (24 features)

3) 4 Models:
   a) 1 Simple Model (generalized linear regression)
   b) 3 Neural Network Models of varying complexity

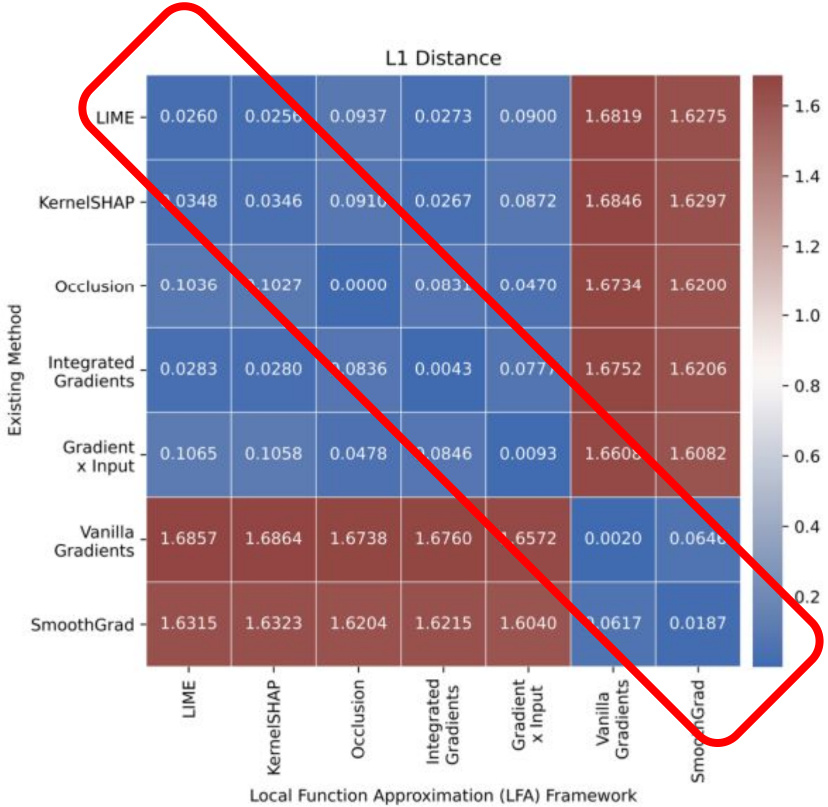
Empirical Evaluation - Experiment #1

1) **Goal:** Show each explanation method fits within the LFA framework by comparing the original method with a LFA re-implementation

2) **Methodology:**
   a) For each XAI method (e.g. LIME), randomly select 100 test set points
   b) Use the LFA and original (Meta) implementations to explain the predictions of black-box models
   c) Evaluate the similarity of the two implementations' explanations (L1)

3) **Results:**
   a) All 7 methods tested (excl. C-LIME) display near-zero L1 distances
   b) Implies each method fits within the LFA framework

Empirical Evaluation - Experiment #1



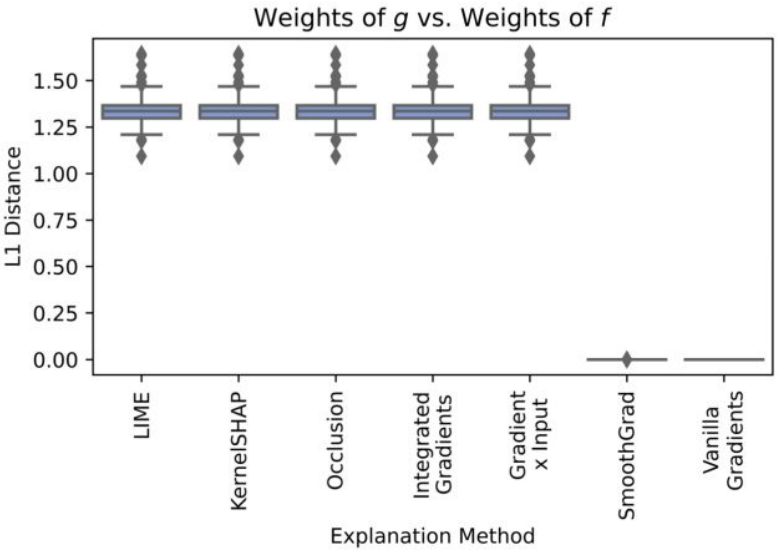
Empirical Evaluation - Experiment #2

1) **Goal:** Confirm that all XAI methods within the LFA achieve “model recovery”
   a) Show that LFA methods recover the black box model “f” when f is an interpretable model

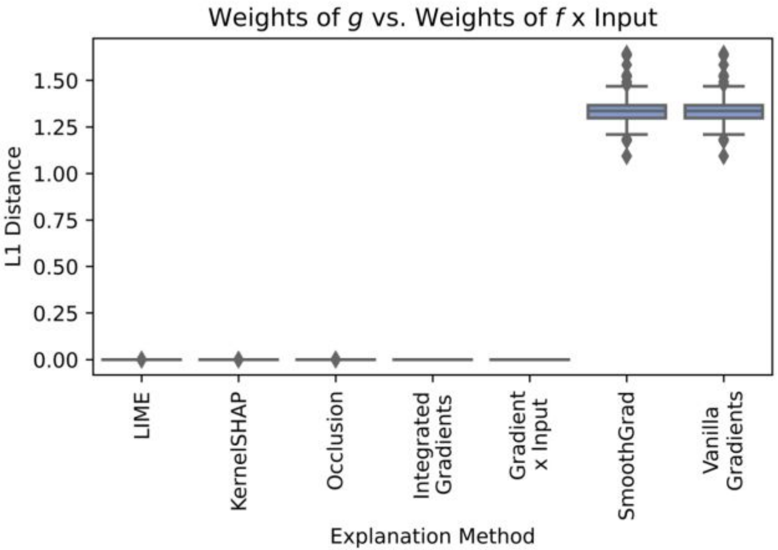
2) **Methodology:**
   a) Set f to be the trained linear regression model on each dataset
   b) Generate LFA explanations (g*), where model class G is linear models
   c) Compare weights of g* with f

3) **Results:**
   a) All 7 models satisfy the LFA guiding principle of “model recovery”
   b) Weights of g* = (gradient of f) or (gradient of f) x (input) for each method

Empirical Evaluation - Experiment #2



(a)



(b)

Empirical Evaluation - Experiment #3

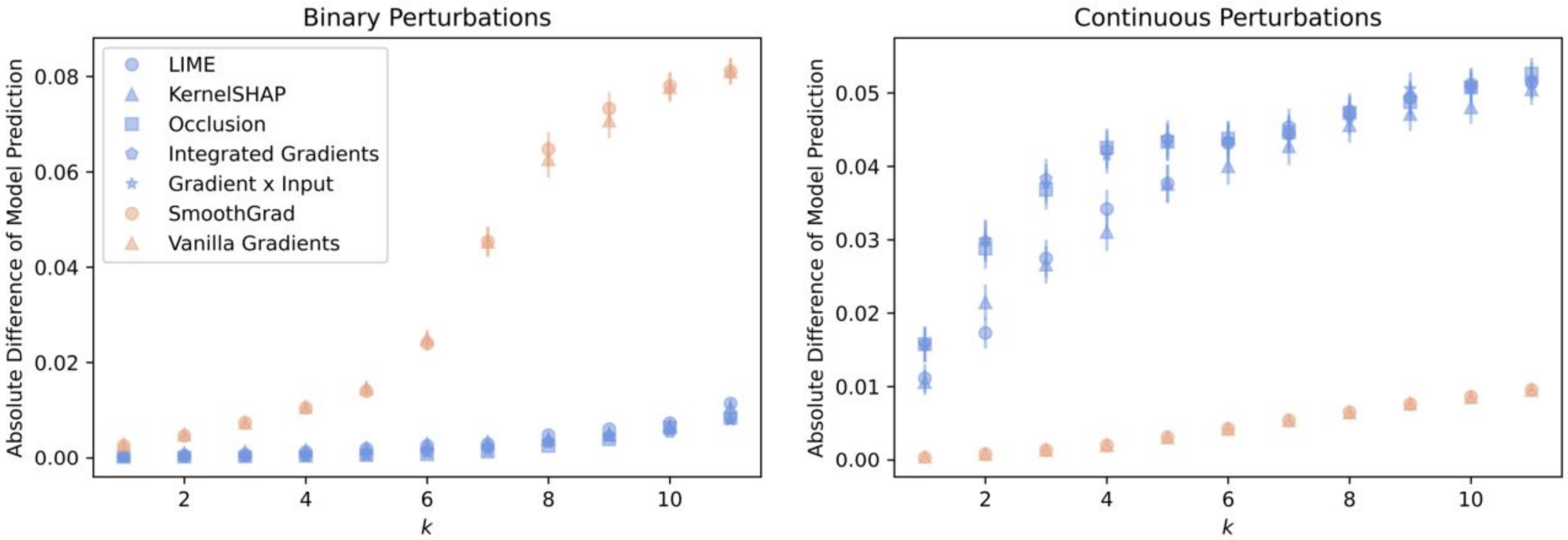
1) **Goal:** Illustrate the LFA No Free Lunch Theorem for common XAI methods

2) **Methodology:**
   a) For each method, generate explanations for 100 random test points
   b) For k = (1, 12), evaluate each explanation by
      i) replace the top/bottom-k features with 0 (binary perturbation)
      ii) addinging Gaussian noise to the top/bottom-k features (continuous perturbation)
   c) Calculate the absolute change in model prediction after perturbation

3) **Results:**
   a) SmoothGrad & Vanilla Gradients perform less well for binary perturbations
   b) 5 other methods perform less well for continuous perturbations

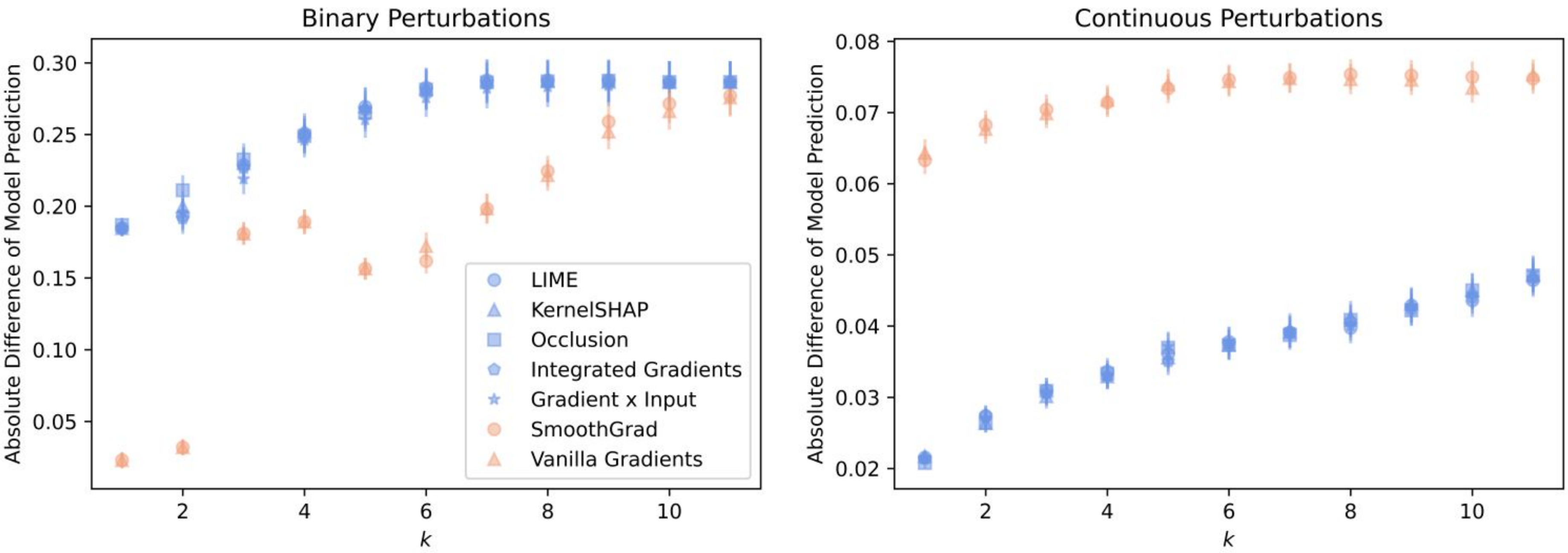
Empirical Evaluation - Experiment #3

Perturbation of Bottom-K Features



Empirical Evaluation - Experiment #3

Perturbation of Top-K Features



Summary of Key Contributions

1) Formalize the local function approximation (LFA) framework

2) Demonstrate that eight popular explanation methods can be characterized as instances of the LFA framework

3) No Free Lunch Theorem: no single LFA method can perform optimally across all neighbourhoods

4) Provide a guiding principle for choosing among LFA explanation methods based on the input domain

Future Work

1) Extend LFA analysis to additional post-hoc explanation methods

2) Develop a similar unifying conceptual framework for the *interpretability* of different model explanations
    a) This work develops a unifying framework for evaluating *faithfulness* of explanations
    b) May require more than a theoretical examination;
        i) Human-Computer Interaction research such as user studies

Discussion Questions

1) Do you agree that the Local Function Approximation framework is a useful conceptual tool for understanding & comparing explanation methods?

2) If no explainability method can perform optimally across all perturbation distributions, as implied by the No Free Lunch Theorem, does that mean explainability should be defined relative to a perturbation neighborhood?
   a) How does that complicate interpretability, especially for practitioners without ML expertise?

3) More broadly, what do you think about papers that attempt to create conceptual coherence & clarity across the field of explainability? Should this be a higher priority area of research?