Towards a Unified Theory of Explainable AI

Student: Veronica Di Gennaro

Mentor: Barbara Di Camillo

Goals and Overview

Explainable AI 1

- **Local Vs. Global XAI**

Existing Methods 2

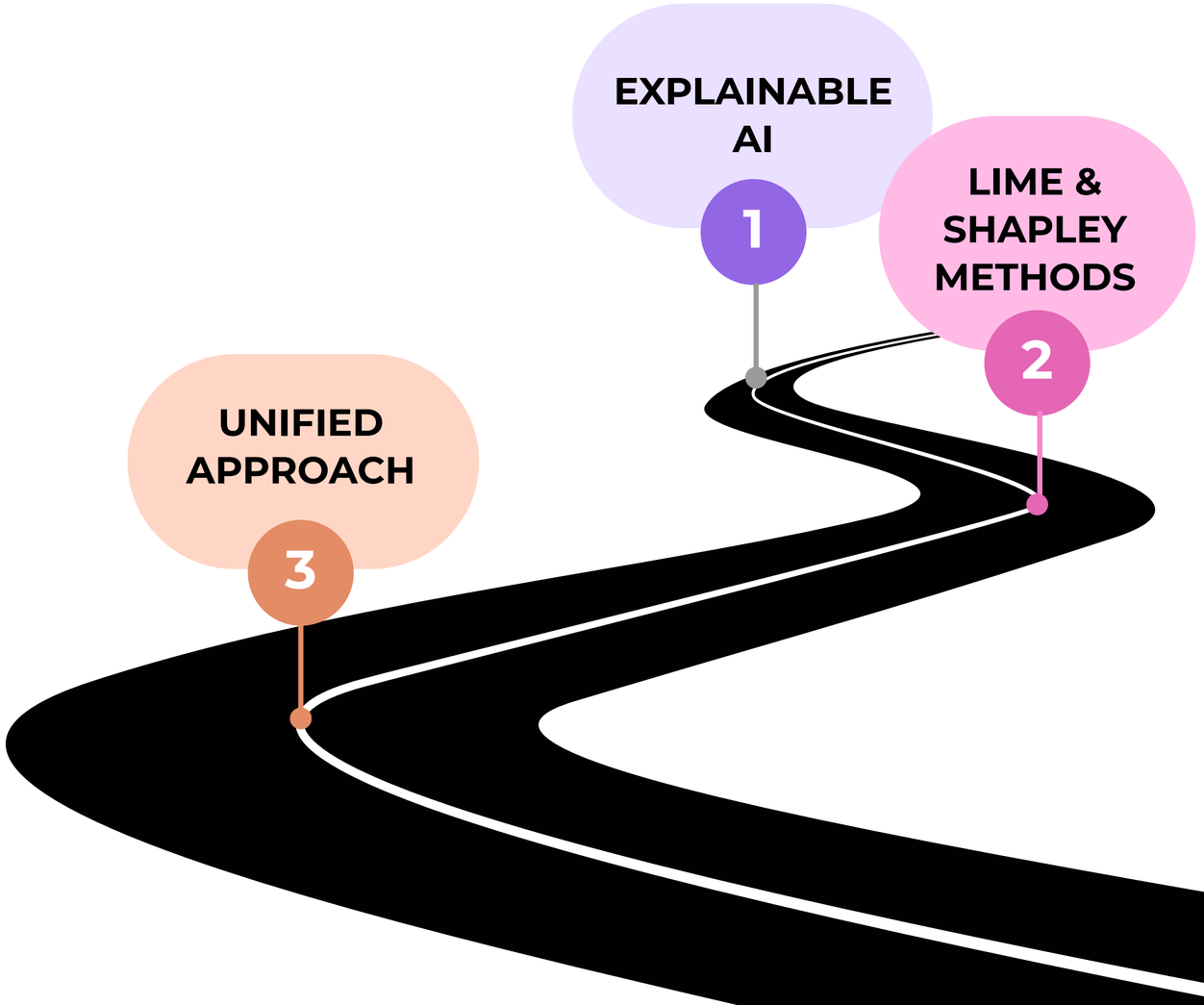
- **active field of research**

LIME 2.1

Shapley 2.2

Unified Approach 3

SHAP 3.1



Motivation

the more complex a machine learning model is...

⇒ the higher its ability to learn complex features'

interaction and **greater representational power**

⇒ BUT the **lower** its **transparency** and **understanding**

of **underlying decision-making mechanisms**.

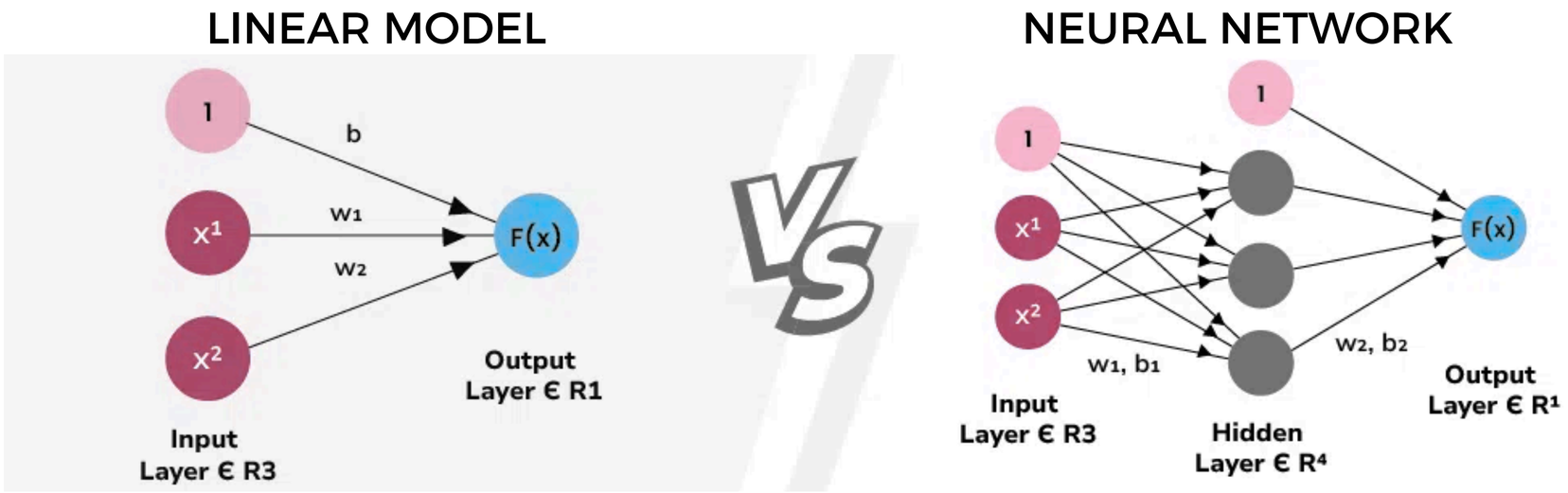
Rise of Explainable AI (XAI)

INTERPRETABILITY = degree to which a human can directly understand the internal workings of an AI model ***without*** the use of ***external explanation tools***.

$\neq$

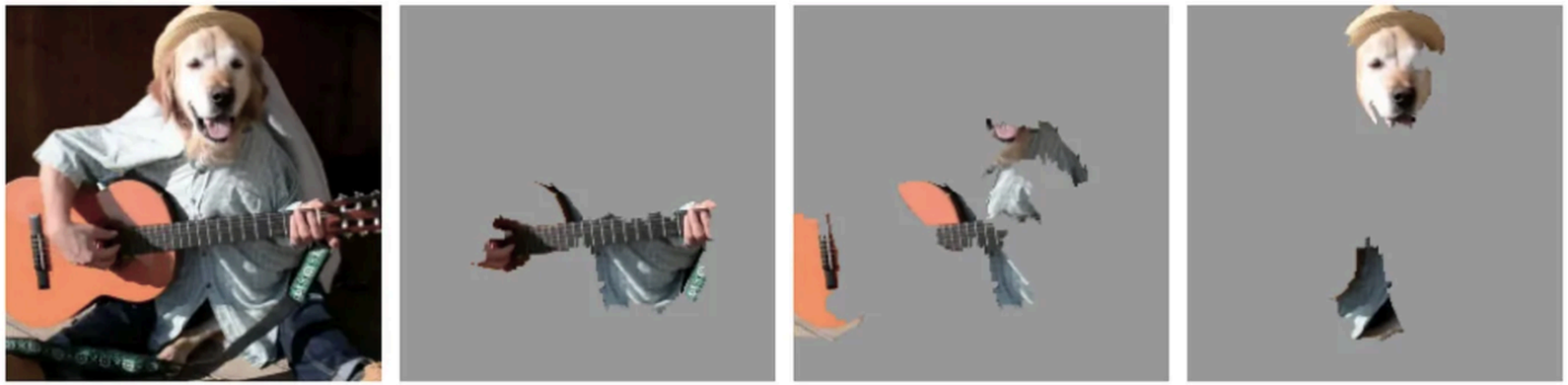
EXPLAINABILITY = property of an AI system that enables it to provide human-understandable explanations regarding ***how*** the model works or the reasons that led to a specific decision.

INTERPRETABILITY → *EXAMPLE:*



GeeksforGeeks. (2025, July 23) https://www.geeksforgeeks.org/machine-learning/linear-regression-vs-neural-networks-understanding-key-differences/

EXPLAINABILITY → *EXAMPLE:*

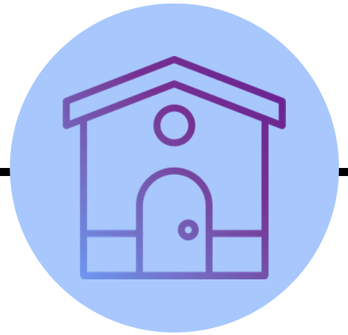


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, February 16). *"Why should I trust you?"*: explaining the predictions of any classifier. arXiv.org. https://arxiv.org/abs/1602.04938

Visual explanations showing exactly which parts of an image influenced a classification prediction.

LOCAL Vs GLOBAL XAI



LOCAL EXPLAINABILITY

answers the question: ***WHY DID THE MODEL MAKE THIS PREDICTION FOR THIS SPECIFIC INSTANCE?***

GLOBAL EXPLAINABILITY

answers the question: ***HOW DOES THE MODEL MAKE PREDICTIONS IN GENERAL? WHAT DOES INFLUENCE MODEL PREDICTION IN GENERAL?***

LOCAL EXPLAINABILITY

set of explanation methods aimed at **describing the behaviour of a machine learning model in the vicinity of a single input instance x**, explaining why the model produced a specific prediction *f(x)*

Formally, a local explanation approximates the model:

$$f(z) \approx g(z) \; \forall z \in \mathcal{N}(x)$$

where:

- *f()* is the original model
- *g()* is an interpretable model
- $\mathcal{N}(x)$ is a local neighborhood of x

GLOBAL EXPLAINABILITY

set of explanation methods aimed at **describing the overall behavior of the model across the entire input space**, providing an interpretable representation of the general regularities, the relationships between variables and the global decision logic of the model

Formally, a global explanation seeks to approximate:

$$f(x) \approx g(x) \; \forall x \in X$$

where g is an interpretable model or representation that captures the average structural behavior of f

Explainable AI 1

- Local Vs. Global XAI

Existing Methods 2

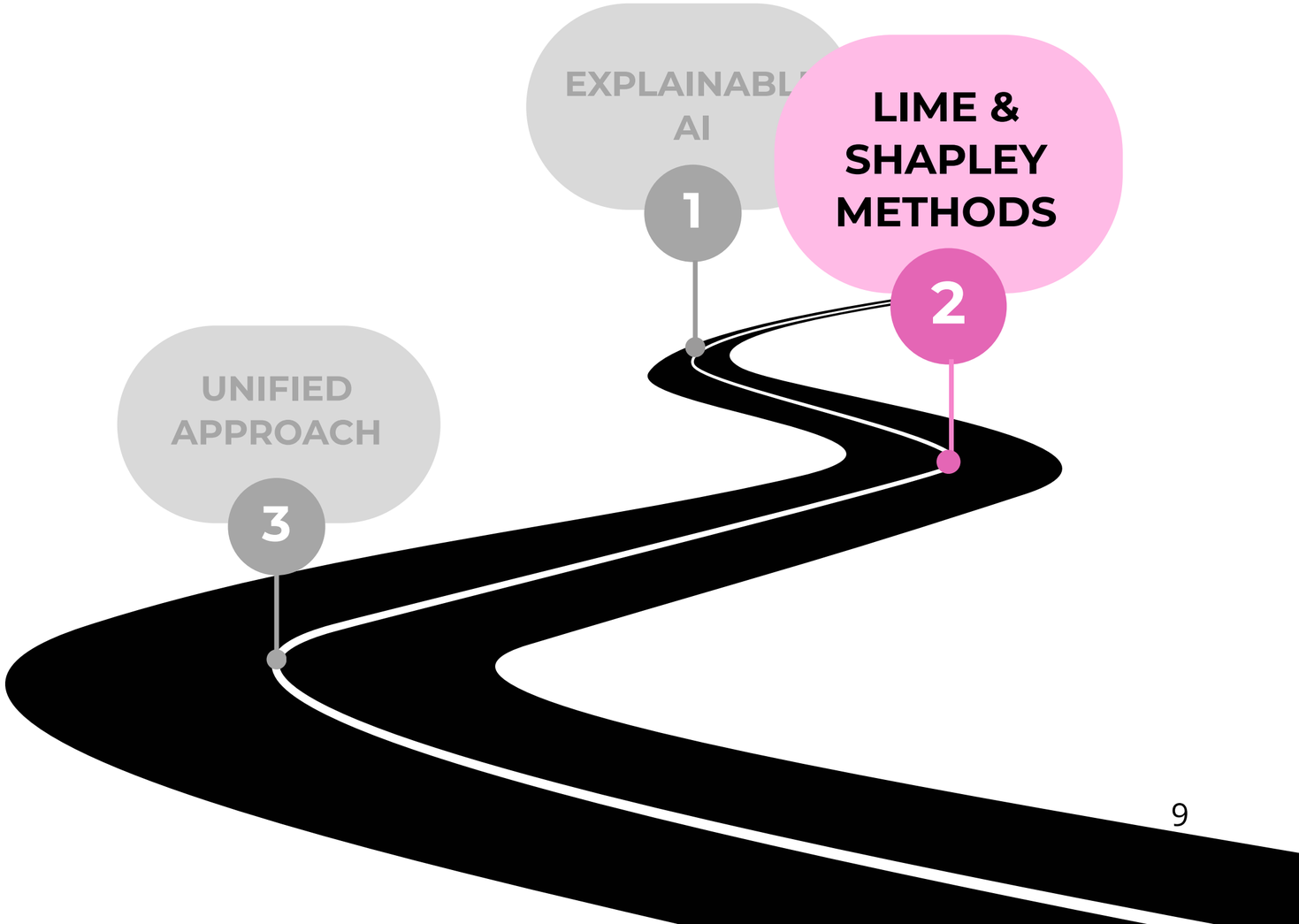
- **active field of research**

LIME 2.1

Shapley 2.2

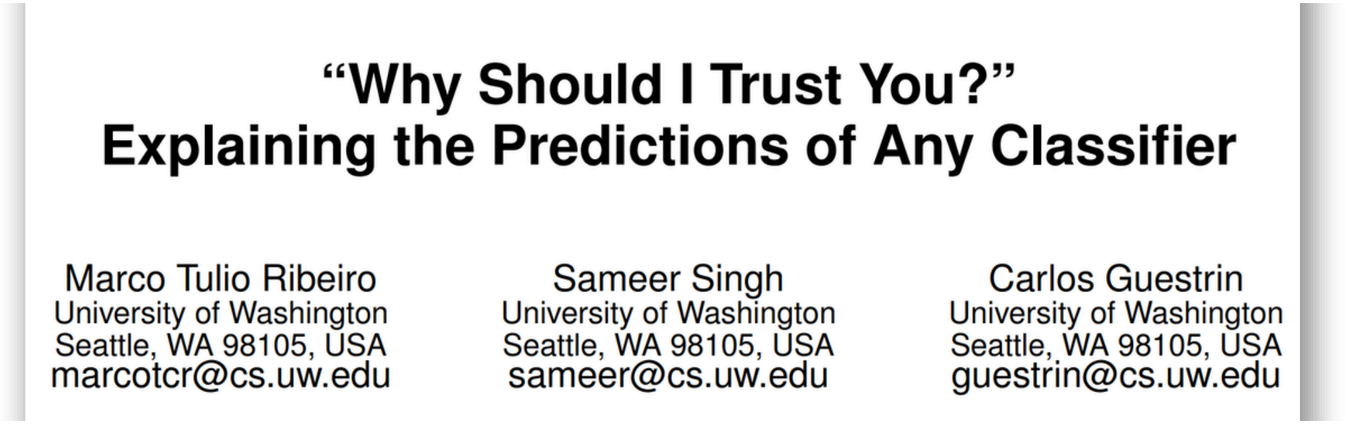
Unified Approach 3

SHAP 3.1



- Explainable AI 1
  - Local Vs. Global XAI
- Existing Methods 2
  - active field of research
  - LIME 2.1
  - Shapley 2.2
- Unified Approach 3
  - SHAP 3.1

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, February 16). "Why should I trust you?": explaining the predictions of any classifier. arXiv.org. https://arxiv.org/abs/1602.04938



"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

LIME

(Local Interpretable Model-agnostic Explanations)

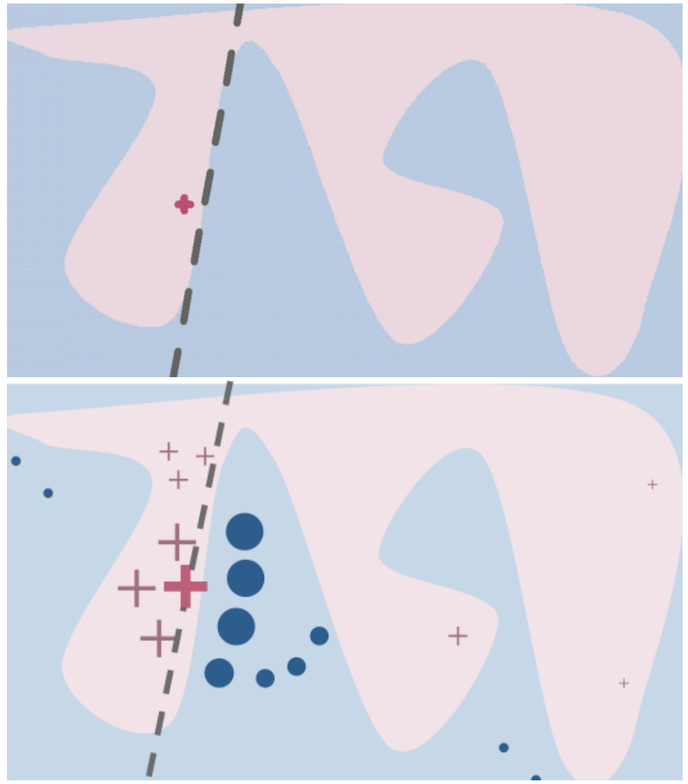
the actual model f can be **complex** and a **black box** (e.g., neural network)

HOWEVER

locally, its behavior can be **approximated** by a simple and interpretable model

LIME:

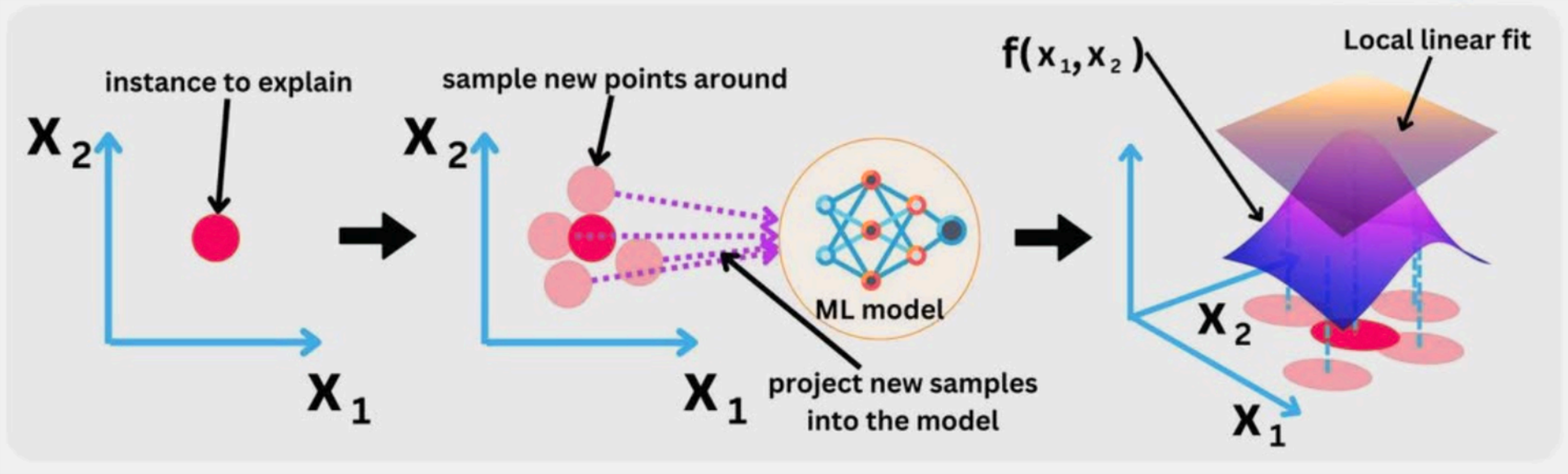
1. takes a **single instance to explain**
2. generates data similar to *x* by applying **perturbations** → z
3. observes model f predictions to *z*
4. trains a simple model to explain the decision on the data *z*



LIME

PERTURBATIONS

A perturbation is the process of mapping an original input x to a "disturbed" version z by removing or modifying randomly some of its components



The method of perturbation depends on the specific data domain **!!**

LIME

SAMPLING STEP FOR LOCAL EXPLORATION

- let $x \in X \subset \mathbb{R}^d$ be the original **instance to be explained**
- we generate a set of ***N* perturbed samples around *x*** by randomly modifying its components (a subset of features)
- let $z_i \in \mathbb{R}^d$ be the i-th perturbed sample

  THEN
- for each generated instance we calculate:
  - the **model's prediction** $f(z_i) =>$ to create a new dataset $\mathcal{Z}$ of perturbed samples with associated labels
  - the **proximity weight** that defines how much that point counts for the explanation (as an exponential kernel defined on some distance function *D*)

$$\pi_x(z_i) = \exp\left(-\frac{D(x, z_i)^2}{\sigma^2}\right)$$

LIME

WEIGHTING STEP

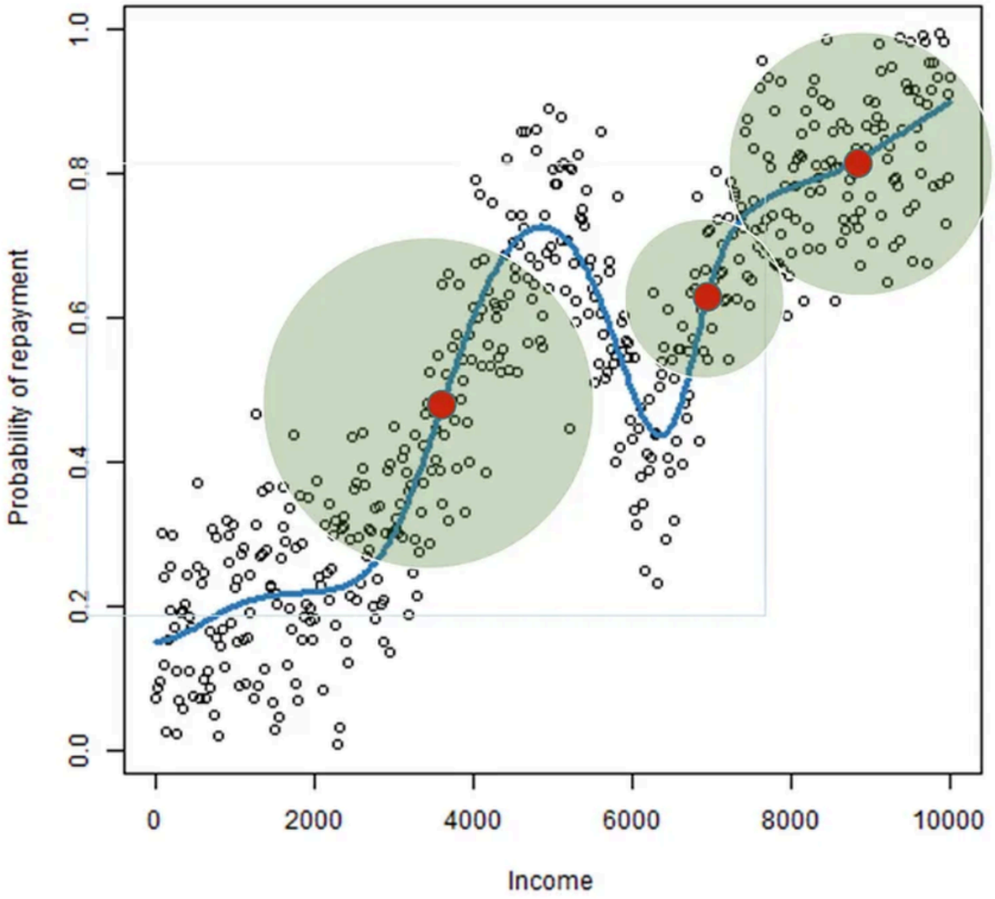
IDEALLY: we should consider only the points in the region of interest (all the linear area of the ML curve around the reference point)

PROBLEM: the **proper size of the region** of interest is not fixed; INSTEAD it **depends on the local curvature of *f(x)***

LIME gives a weight to each generated point

- example: Gaussian (RBF) Kernel:

$$RBF(z_i) = \exp\left(-\frac{\|z_i - x\|^2}{\sigma^2}\right)$$



https://medium.com/data-science/lime-explain-machine-learning-predictions-af8f18189bfe

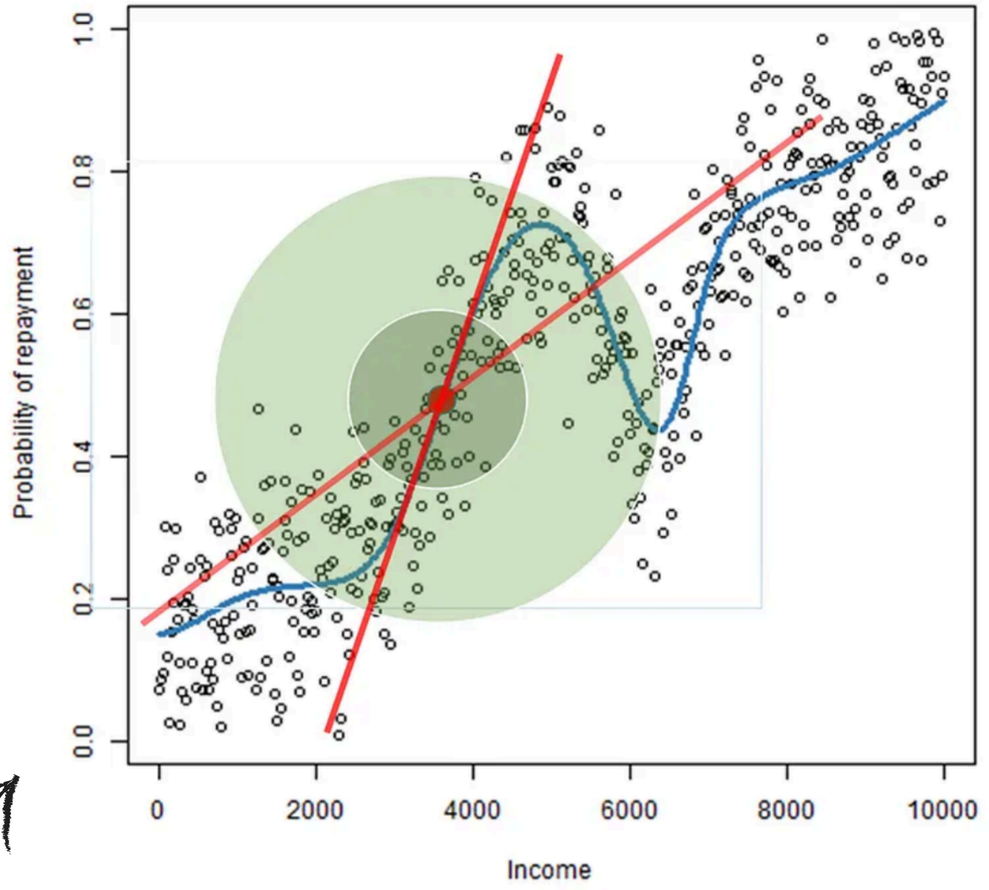
LIME

LOCAL MODEL STEP

- LIME uses an **explainable model to approximate the ML model** in the small region around the reference point.

- Any kind of explainable model for the approximation (Decision Trees, Logistic Regression, ...). Although **Linear Regression is preferred**.

*what happens if the **size of the neighborhood is poorly chosen?***



https://medium.com/data-science/lime-explain-machine-learning-predictions-af8f18189bfe

LIME

INHERENT INTERPRETABILITY OF LINEAR MODELS

EXAMPLE: simple linear regression model

model's prediction is the sum of each feature multiplied by a corresponding weight plus a bias

impact of a feature revealed directly by the model's **mathematical structure**

"best explanation of this simple model is the model itself"

and this is exactly what we expect from the **explanation model *g*:**

$$g(x) = \phi_0 + \sum_{j=0}^{N} \phi_j x_j$$

LIME

OPTIMIZATION

minimization problem formulation:

$$\xi = arg\min_{g \in G} \; L(f, g, \pi_x) + \Omega(g)$$

balancing two competing goals:

Faithfulness (**L**): how closely the explanation model g matches the original model's predictions over a set of samples, weighted by a local kernel

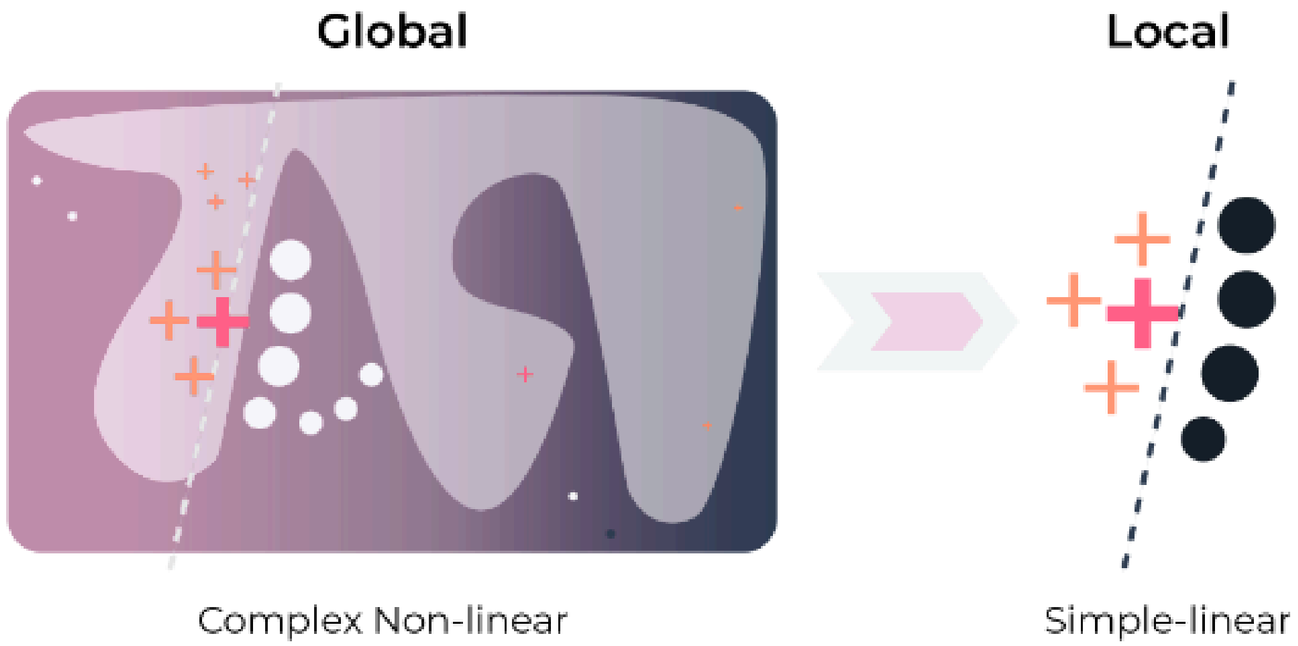
$$L = \sum_{z_i \in \mathcal{Z}} \pi_x(z_i)[f(z_i) - g(z_i)]^2$$

Complexity (**Ω**): A penalty that prevents the explanation from becoming too complex (e.g. too many features)

LIME

ONLY a LOCAL METHOD

NOT **explaining** the entire model BUT **ONLY what happens around the single instance**



Explainable AI 1

- Local Vs. Global XAI

Existing Methods 2

- active field of research

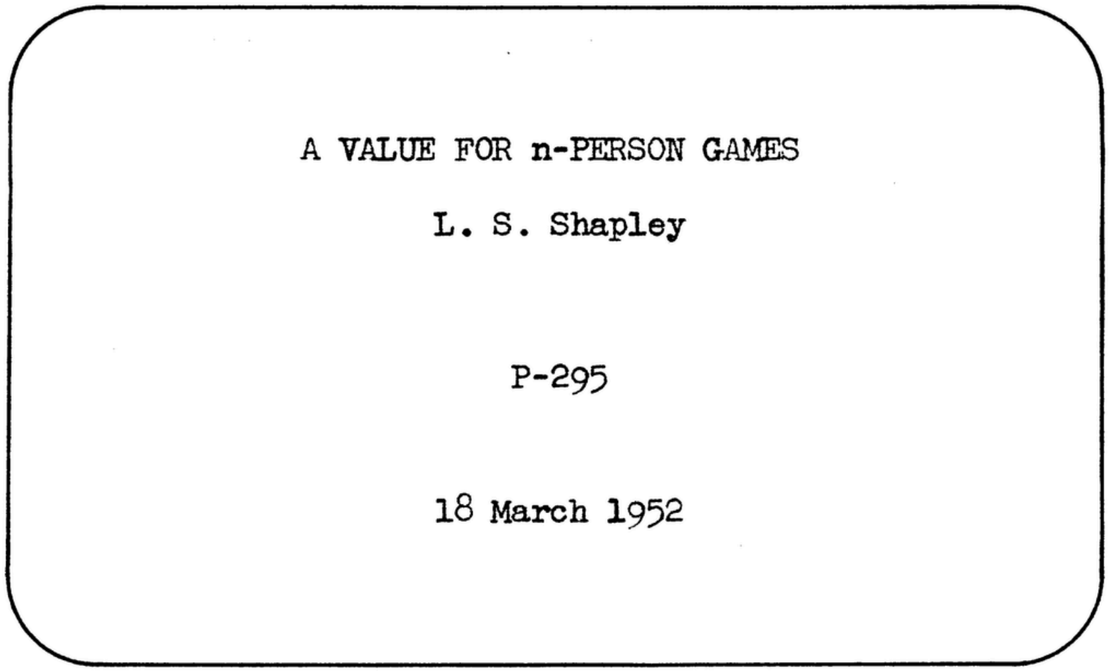
LIME 2.1

Shapley 2.2

Unified Approach 3

SHAP 3.1

Shapley, L.S. (1953) 17. A Value for N-Person Games. In: Kuhn, H.W. and Tucker, A.W., Eds., Contributions to the Theory of Games (AM-28), Volume II, Princeton University Press, 307-318.



A VALUE FOR n-PERSON GAMES

L. S. Shapley

P-295

18 March 1952

SHAPLEY

in game theory

originally introduced by Lloyd S. Shapley in 1952 to find a solution to the following problem:

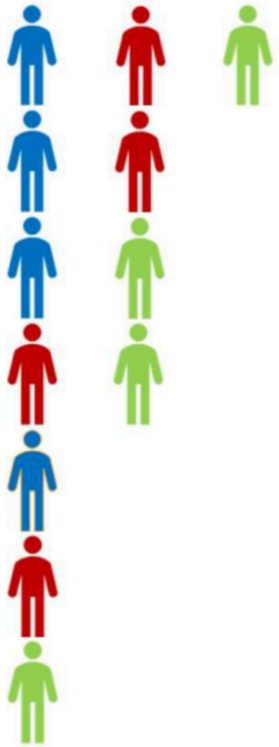
In a group of players, with different skill sets, that worked together to reach a payout, ***what is the fairest way to split the payout*** *among them?*

SOLUTION:

The fair distribution of total gains among players is computed based on their **marginal contributions.**

We take the **weighted average** of all possible marginal contributions of a player **across every possible coalition!!**

EXAMPLE



Price
for different coalitions
(possible coalitions are 8)

$C_{123} = 12000$

$C_{12} = 8000$

$C_{13} = 7000$

$C_{23} = 6000$

$C_1 = 5000$

$C_2 = 5000$

$C_3 = 4000$

$C_0 = 0$

Imagine we can play the same game many times considering different coalitions of players:

What is the contribution of player 1 to the overall prize?

we calculate how the prize changes when player 1 joins each coalition:

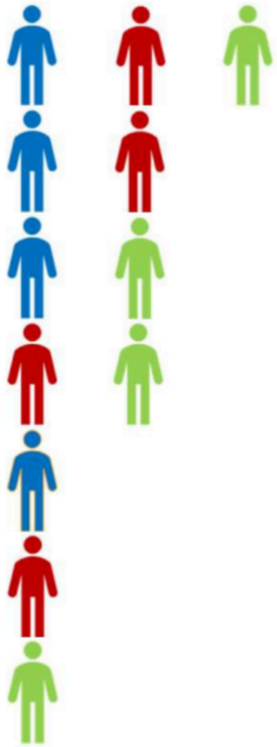
$$C_{231} - C_{23} = 12000 - 6000 = 6000$$
$$C_{21} - C_2 = 8000 - 5000 = 3000$$
$$C_{31} - C_3 = 7000 - 4000 = 3000$$
$$C_1 - C_0 = 5000 - 0 = 5000$$

weighted average:

$$6000(2/6) + 3000(1/6) + 3000(1/6) + 5000(2/6) = 4667$$

corresponding to the contribution of player 1 ✓

EXAMPLE



Price
for different coalitions
(possible coalitions are 8)

$C_{123} = 12000$

$C_{12} = 8000$

$C_{13} = 7000$

$C_{23} = 6000$

$C_1 = 5000$

$C_2 = 5000$

$C_3 = 4000$

$C_0 = 0$

probability that player 1 joins a coalition of two players

given by:

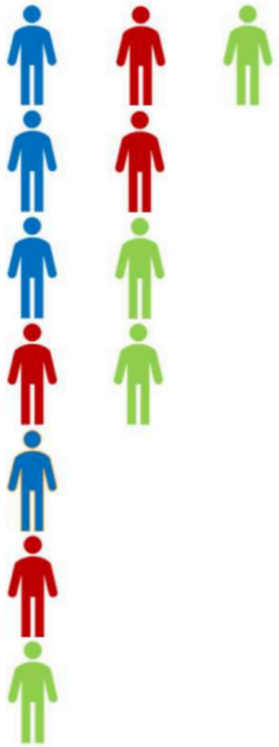
- 3! = 6 possible ways of forming an ordered coalition of 3 people (123, 132, 213, **231**, 312, **321**)
- 2 possible ways in which player 1 can join a group of already 2 players

weighted average:

6000(2/6) + 3000(1/6) + 3000(1/6) + 5000(2/6) = 4667

corresponding to the contribution of player 1 ✓

EXAMPLE



Price
for different coalitions
(possible coalitions are 8)

C_{123} = 12000

C_{12} = 8000

C_{13} = 7000

C_{23} = 6000

C_1 = 5000

C_2 = 5000

C_3 = 4000

C_0 = 0

What is the contribution of player 2 to the overall prize?

5000(2/6) + 3000(1/6) + 2000(1/6) + 5000(2/6) = 4167

What is the contribution of player 3 to the overall prize?

4000(2/6) + 2000(1/6) +1000(1/6) + 4000(2/6) = 3166

NOTICE: If we sum the contributions from all three players we retrieve the total prize in the case all players play the game (=12000)

SHAPLEY

in machine learning

- this problem can be translated **from GAME THEORY to MACHINE LEARNING:**

  - game → prediction task in the single instance
  - players → feature values of the specific instance
  - payout → prediction/model output

marginal contribution:

$$f_{S\cup i}(x_{S\cup i}) - f_S(x_S)$$

where:

- $f_{S\cup i}(\cdot)$ = prediction of the model *knowing* the feature value
- $f_S(\cdot)$ = prediction of the model *without knowing* the feature value

SHAPLEY

in machine learning

taking the **weighted average** of all possible marginal contributions of a **feature** across **every possible subset**

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

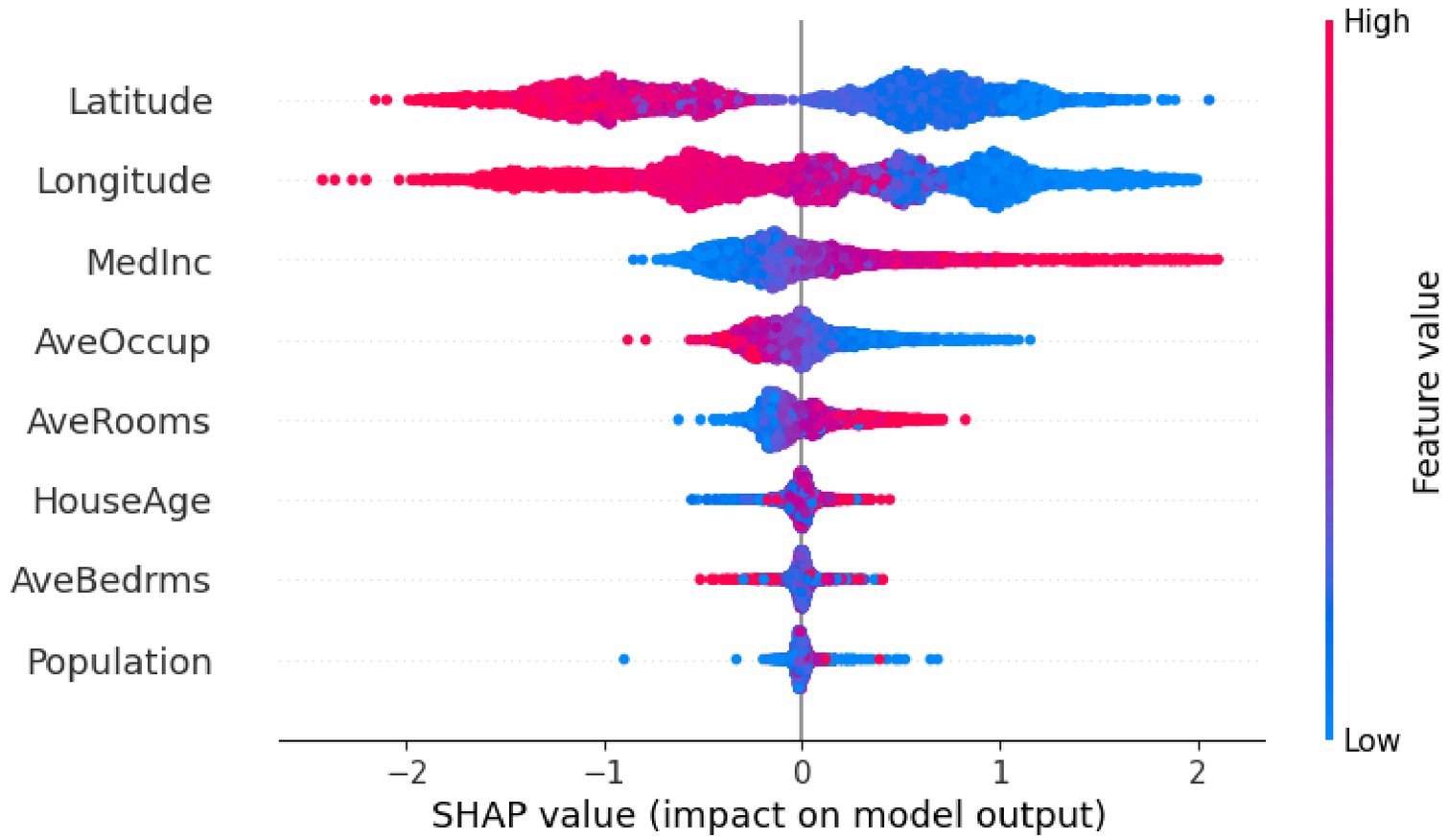
SHAPLEY VALUES

SHAPLEY

BOTH LOCAL and GLOBAL METHOD

Google Colab
google.com

example: SUMMARY (or BEESWARM) PLOT



Explainable AI 1

- Local Vs. Global XAI

Existing Methods 2

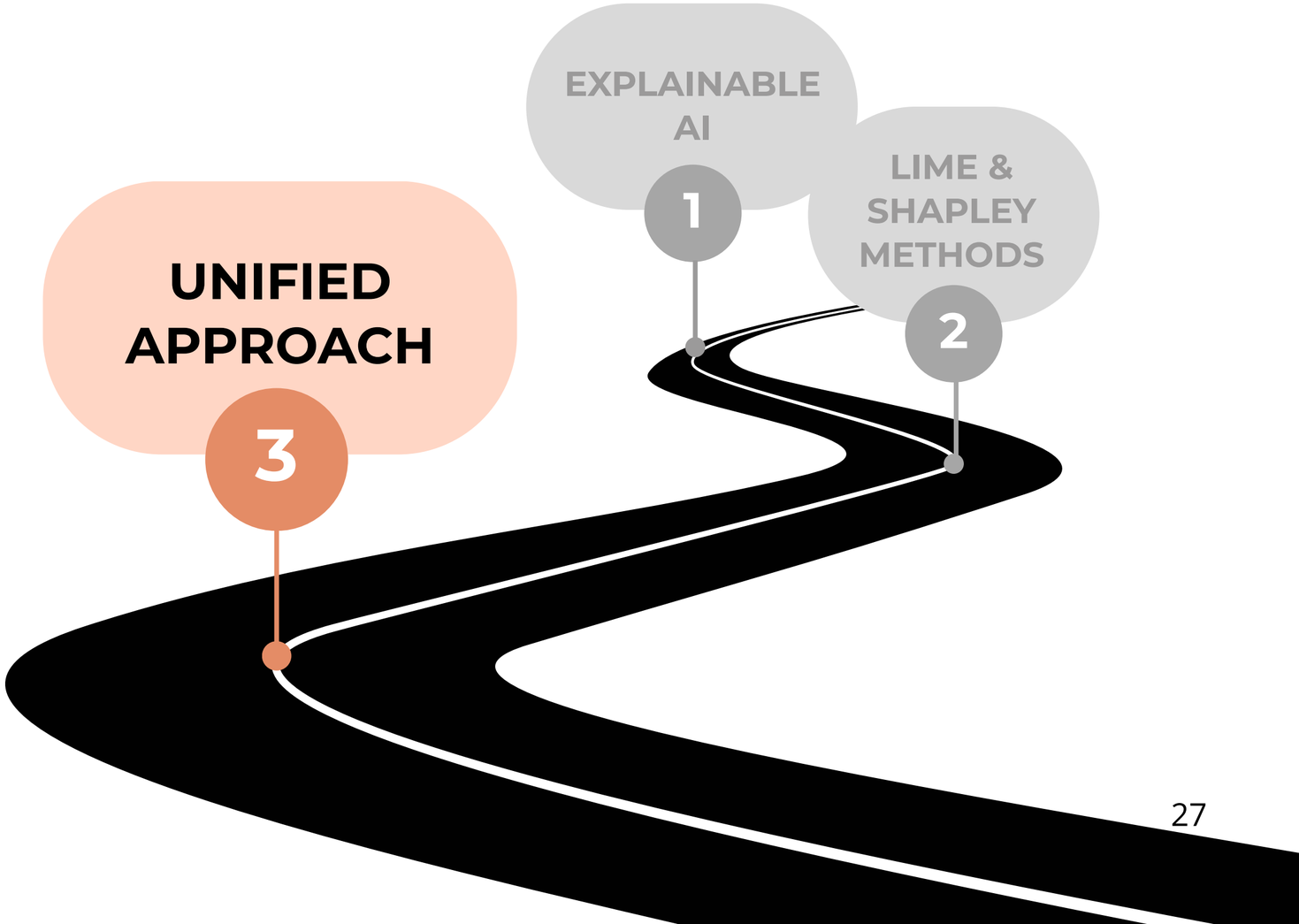
- active field of research

LIME 2.1

Shapley 2.2

Unified Approach 3

SHAP 3.1



Explainable AI 1

- Local Vs. Global XAI

Existing Methods 2

- active field of research

LIME 2.1

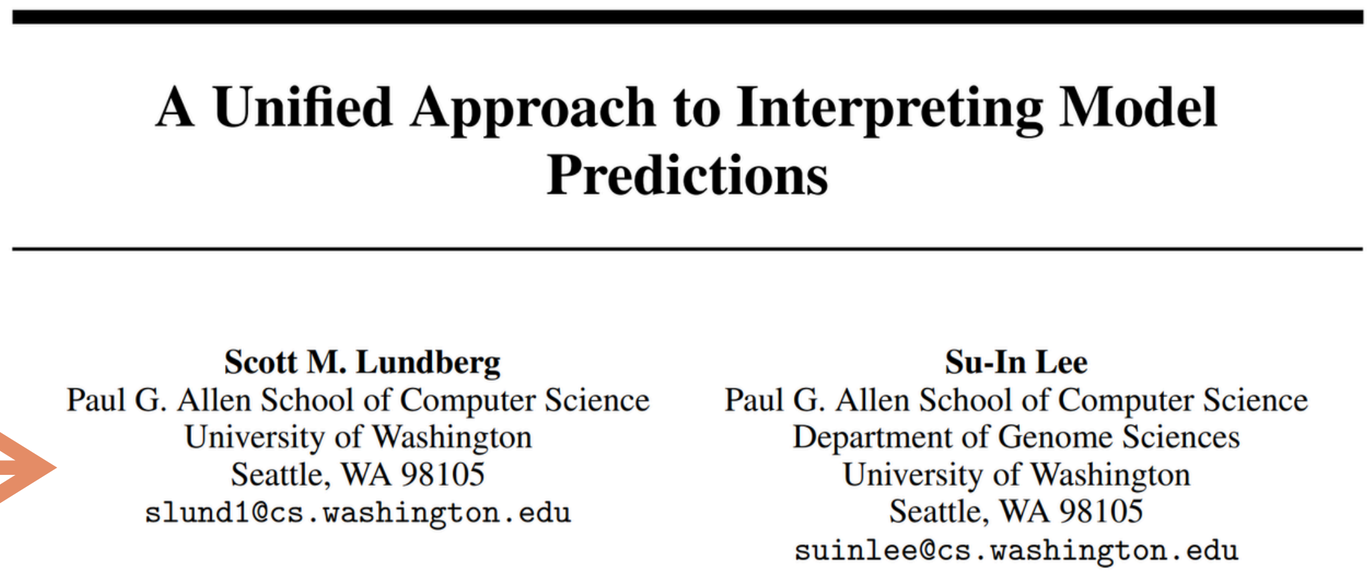
Shapley 2.2

Unified Approach 3

SHAP 3.1

Shapley and Lime in a unified framework...

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions https://doi.org/10.48550/arxiv.1705.07874



A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

SHAP

(SHapley Additive exPlanations)

the paper introduces the **desirable properties for any XAI method**:

- BOTH **local** and **global** explanations
- mathematical and theoretical foundation
  - properties of:
    - **SYMMETRY**,
    - **EFFICIENCY**
    - **NULLITY**
    - **ADDITIVITY**
- **feasibility** in terms of computational complexity

1 **SYMMETRY** The XAI values of two features are the same IF they contribute equally to all possible subsets of features (aka IF they are symmetric)

$$\forall S \subset F \setminus \{i, j\} : f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) = f_{S\cup\{j\}}\left(x_{S\cup\{j\}}\right) \implies \phi_i(x) = \phi_j(x)$$

2 **EFFICIENCY** The sum of the XAI values for all features for a specific instance equals the model prediction (on the specific instance) minus the average prediction

$$\sum_{i}^{F} \phi_i(x) = f(x) - \mathbb{E}[f(X)]$$

3 **NULLITY** A feature that does not change the predicted value (dummy player in game theory) regardless of which coalition of feature values it is added to should have a XAI value of `0`.

$$\forall S \subset F \setminus \{i\} : f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) = f_S(x_S) \implies \phi_i(x) = 0$$

4 **ADDITIVITY** If two models described by the prediction functions *f* and *g* are combined, the total XAI value for a feature should correspond to the sum between the contributions derived from *f* and *g*

$$\phi_i^{f+g}(x) = \phi_i^f(x) + \phi_i^g(x)$$

SHAP

(SHapley Additive exPlanations)

- The paper shows that, as of today, **Shapley** is the only methods to **satisfy all mathematical properties** of symmetry, efficiency, nullity and additivity

- Shapley is also **both** a **local** and **global** method

- BUT it has a **HUGE LIMITATION**:

  In order to *exactly* compute Shapley Values we must evaluate all possible coalitions of features:

  ➔ **computational cost grows exponentially with the number of features**

SHAP

(SHapley Additive exPlanations)

- It identifies a unique class of methods called **additive feature attribution methods** that approximate the model locally (exactly like LIME!)

- Can be formulated as a **minimization problem**:

$$\xi = arg\min_{g \in G} \; L(f, g, \pi_x) + \Omega(g)$$

with **proximity weigth term** carefully chosen and **regularization** equal to zero in order **recover an approximation of Shapley values**

(by theorem 2 in the paper)

$$\pi_x(z) = \frac{(M-1)}{(M \text{ choose } |z|) \cdot |z| \cdot (M - |z|)}$$

$$\Omega(g) = 0$$

$$L(f, g, \pi_x) = \sum_{z \in Z} [f(z) - g(z)]^2 \cdot \pi_x(z)$$

Advantages

- The choices of terms in the minimization ensure that the **explanations satisfy the theoretical properties** associated with Shapley values.

- **Less computationally costly** than original and exact Shapley values computation.

- Retrieving Shapley Value we obtain **BOTH local** and **global** explanations

Conclusions

- two major approaches: LIME, which focuses on local model approximation, and Shapley values, which provide a theoretically grounded way to attribute feature importance

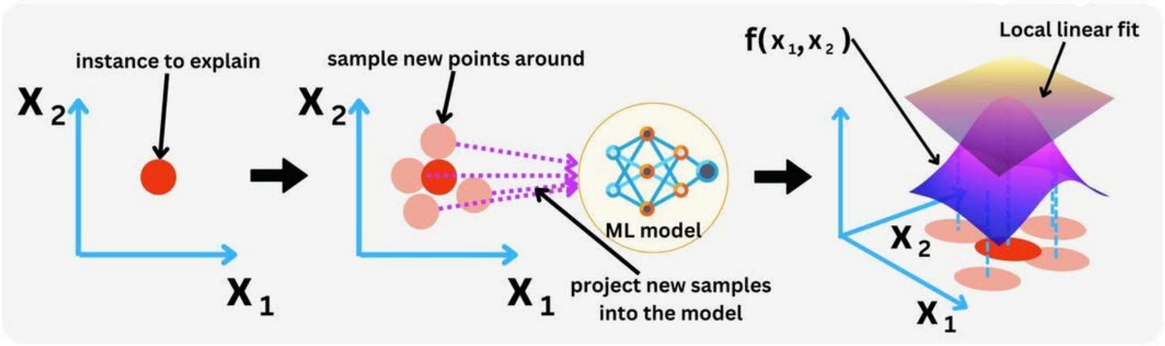
- The **SHAP framework unifies these ideas** and demonstrates that Shapley values provide the unique mathematically consistent solution for additive feature attribution methods.

Thank you for your attention!

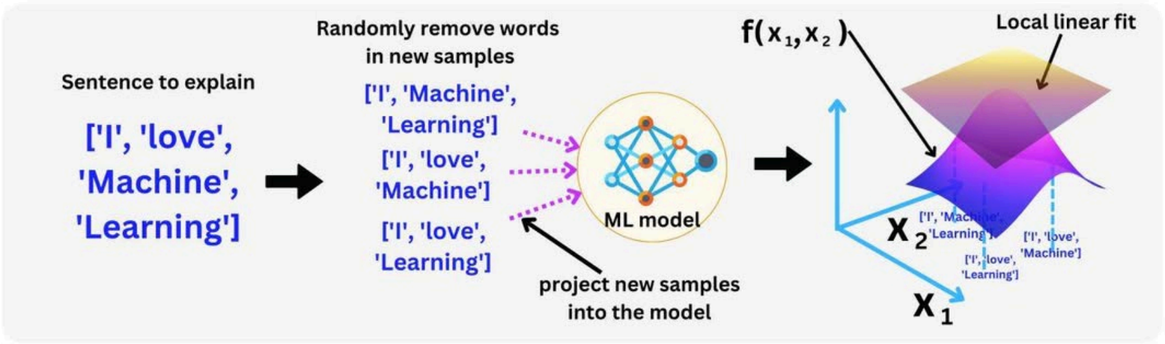
LIME

PERTURBATIONS

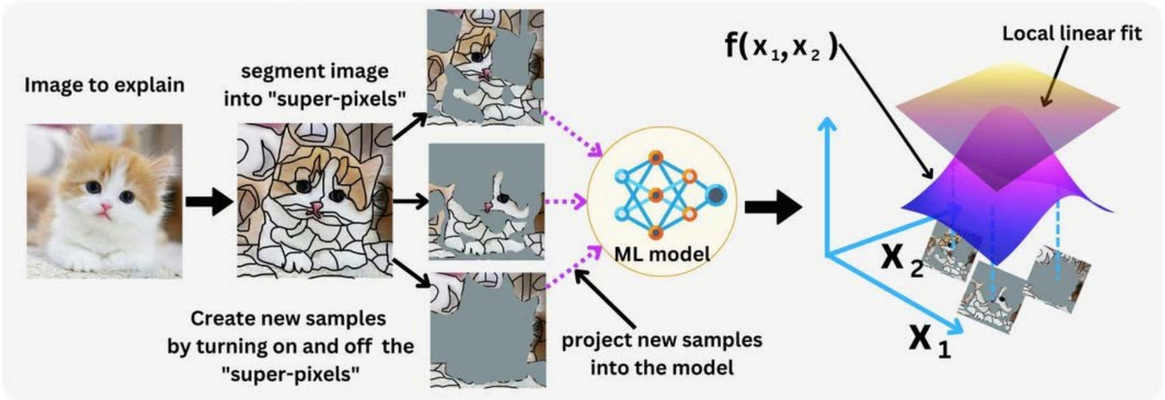
LIME with Tabular data



LIME with Text data



LIME with Image data



LIME

PERTURBATIONS

EXAMPLE in the image domain

- x → original image
- x' → interpretable representation of x (e.g. a binary vector indicating presence or absence of a "superpixel")

superpixel = contiguous patch of similar pixels

Original Image

Interpretable Components

alteration of image data then consists in randomly hiding some of these superpixels

LIME

PERTURBATIONS

EXAMPLE for tabular data:

scenario: the model is given a set of symptoms of a patient and the model outputs a sickness score

- x → original input → e.g. **(fever=38.8°C; cough=True)**
- x' → interpretable representation of x → e.g. a binary vector (high=1, low=0, cough=1)

  identifying 3 bins for the continuous variable:
  - low: <35°C
  - mid: 35-37°C
  - high: >37°C

LIME perturbs this sample by modifying one or both variables:

- for the discrete variable cough → it can flip it → (high=1, low=0, cough=0)

  mapping it back to original representation → **(fever=38.8°C; cough=False)**
- for the continuous variable → it can substitute it with a value from other bins

  → (high=0, low=0, cough=1)

  mapping it back to original representation → **(fever=36.5°C; cough=True)**

SHAPLEY

in machine learning *how do we "remove" a feature?*

model prediction "*without knowing* the feature value":

$$f_S(x_S) = \int f(x_1, x_2, \ldots, x_p) dP_{x_i \notin S}$$

we marginalize the model predictions across all the feature values that do not belong to the subset S

EXAMPLE:

For example, suppose we trained a model with 2 features:

- Age x1 ∈ [15, 95]
- Smoking x2 ∈ {0, 1}

and we want to calculate the Shapley value for feature x2 for the example Jane: Age = 80, Smoke = 1

We need to calculate:

$$f_{S \cup i}(x_{S \cup i}) = f(x_1 = 80, x_2 = 1)$$

$$f_S(x_S) = \int f(80, x_2) dx_2 = f(80, 0) \cdot Prob(x_2 = 0) + f(80, 1) \cdot Prob(x_2 = 1)$$