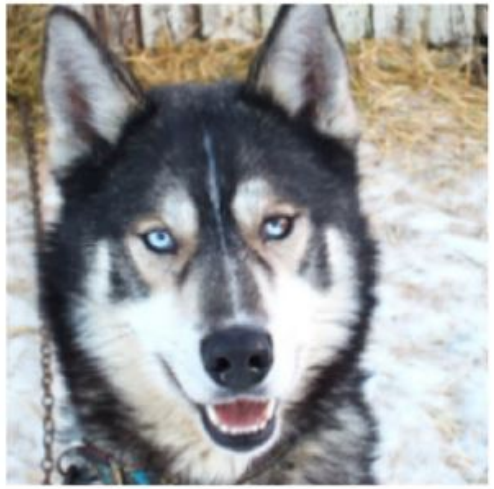
Microsoft

Model Interpretability: Shedding Light on Black Box Techniques

Daniel Schneider
Principal Program Manager
Microsoft Azure ML Platform

Interpretability for Debug

Model performance metrics aren’t enough!



(a) Husky classified as wolf

(b) Explanation

SIGKDD’16 (RSG16)

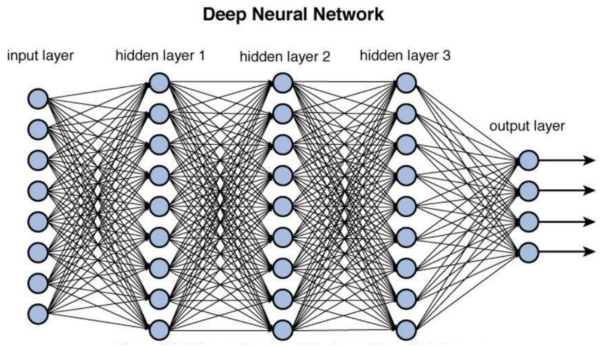
We just trained a perfect snow detector!!!

Interpretability for Regulation

The recent General Data Protection Regulation (GDPR) requirements state a right to explanation for any algorithm whose decision impacts a person's legal status:

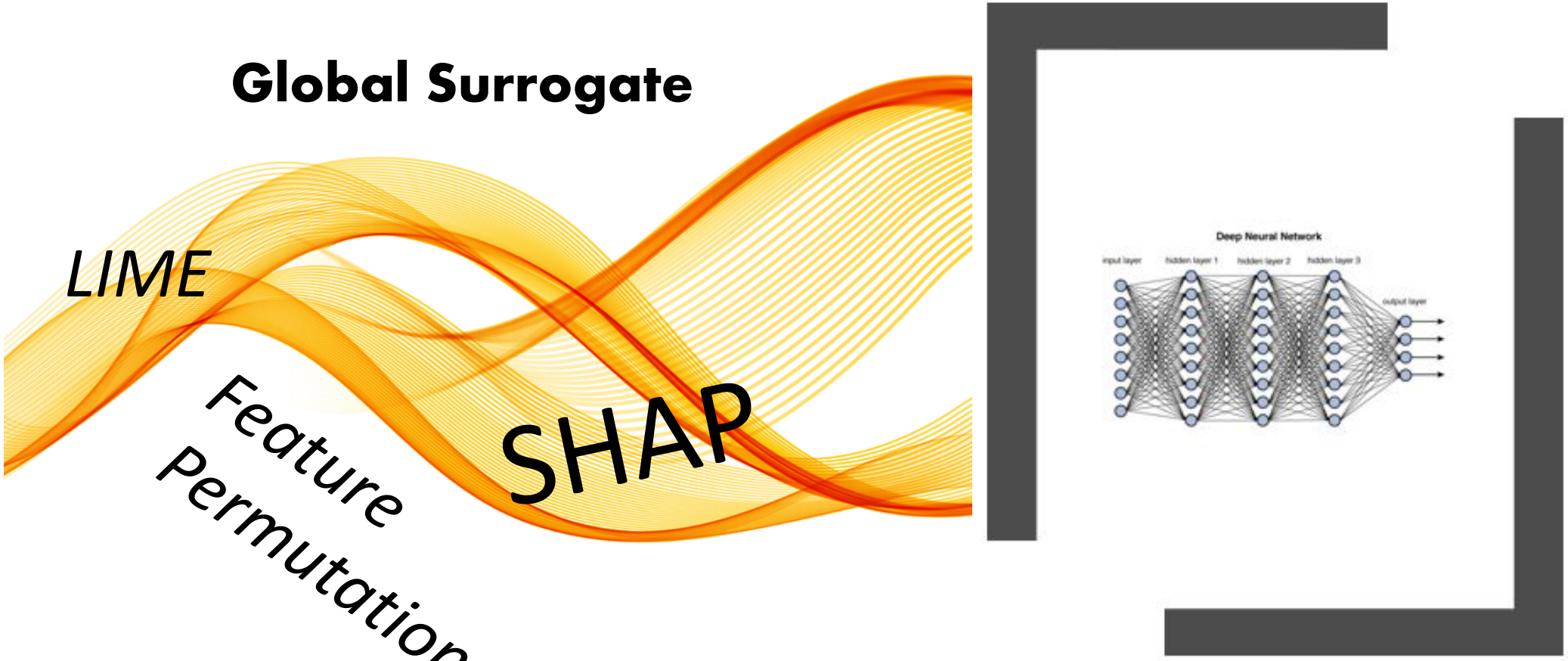
- ✓ Identifying the factors that went into a decision
- ✓ Knowing how varying a factor impacts a decision
- ✓ Comparing similar instances with different outcomes

Black Box Techniques

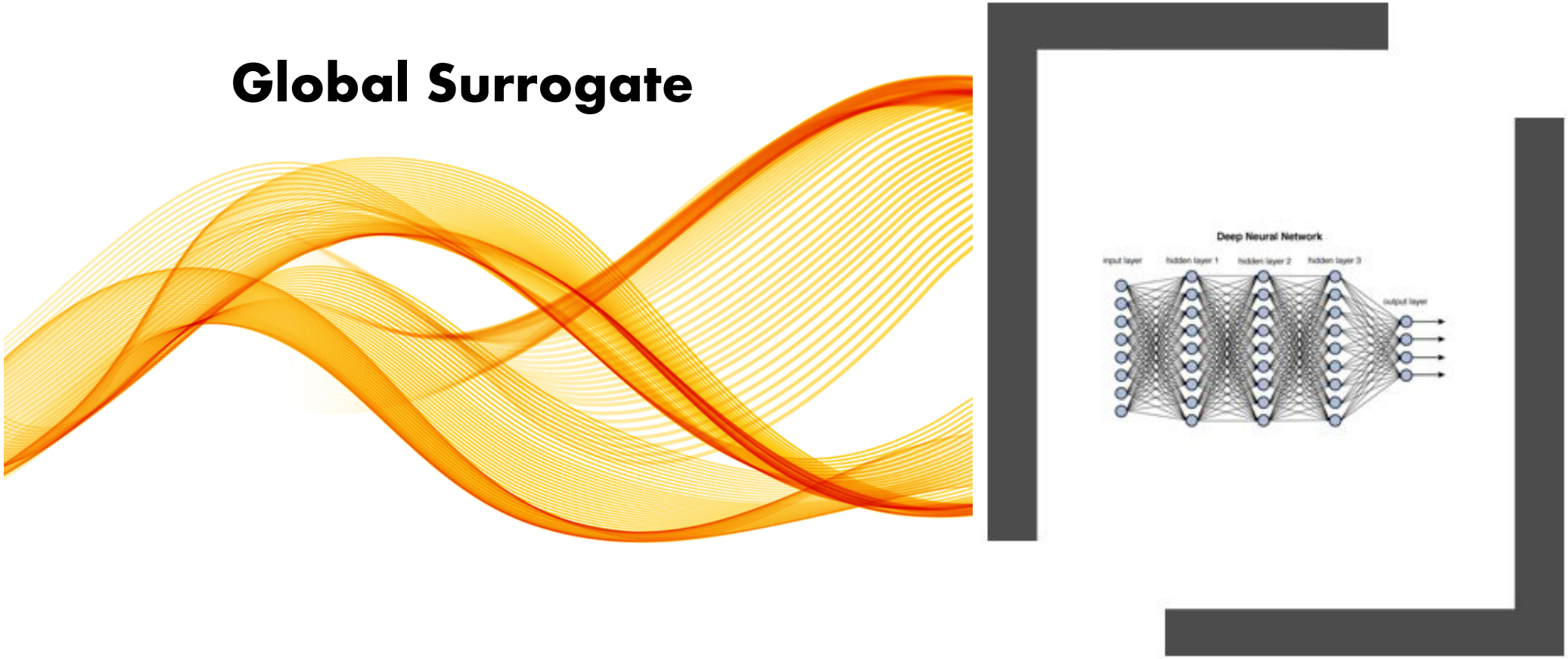


- Increased computational power allows more complex models
- Accuracy goes up and Interpretability goes down dramatically

Shedding Light on Black Box Techniques



Global Surrogate

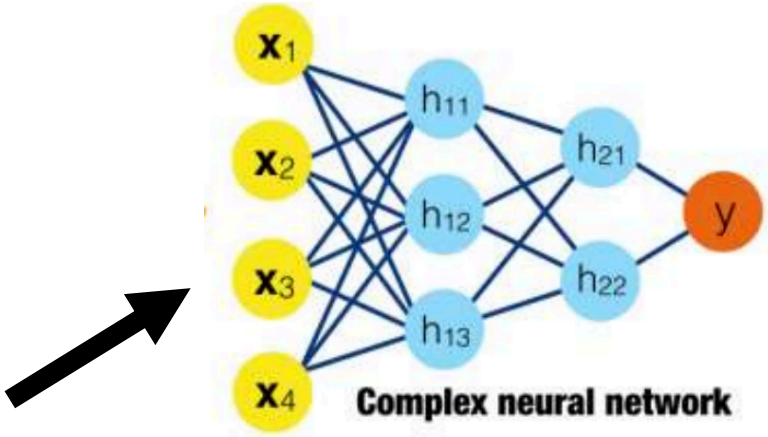


Global Surrogate Models

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33

Global Surrogate Models

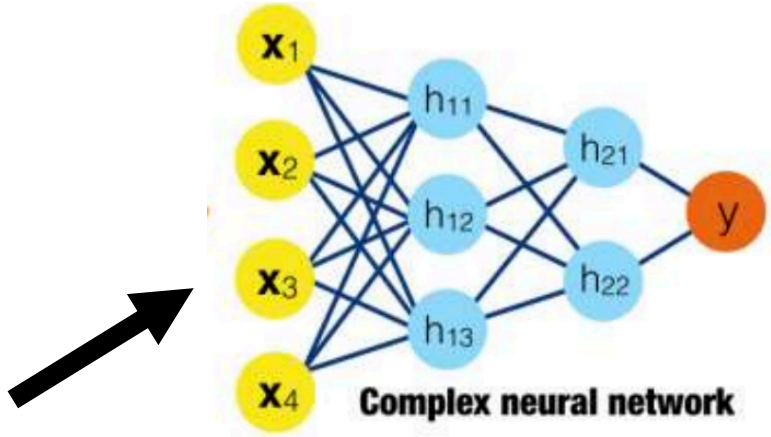
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33



Price
24.0
21.6
34.7
33.4
36.2

Global Surrogate Models

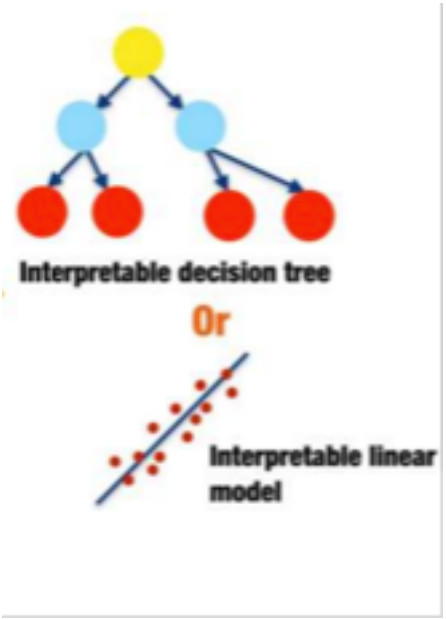
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33



Price
24.0
21.6
34.7
33.4
36.2

Predictions of
the black-box model

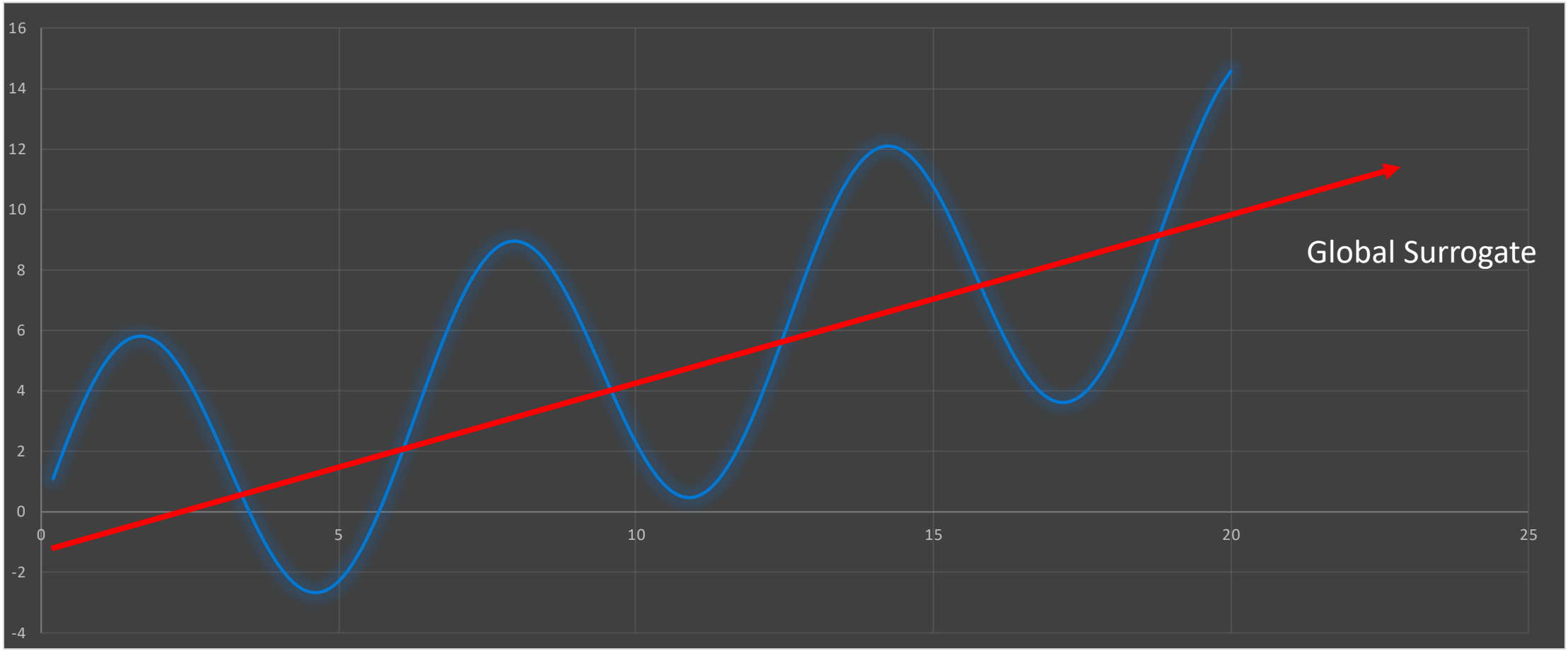
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	Price
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

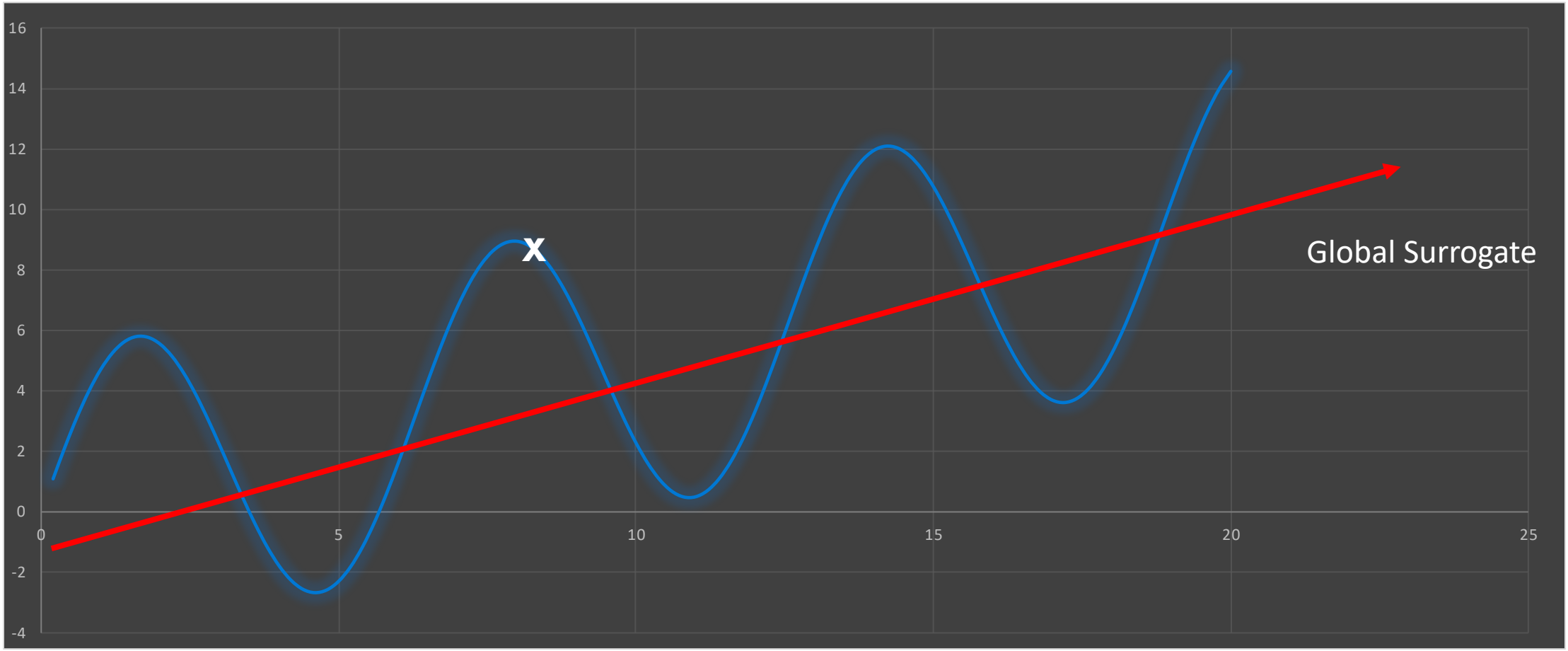


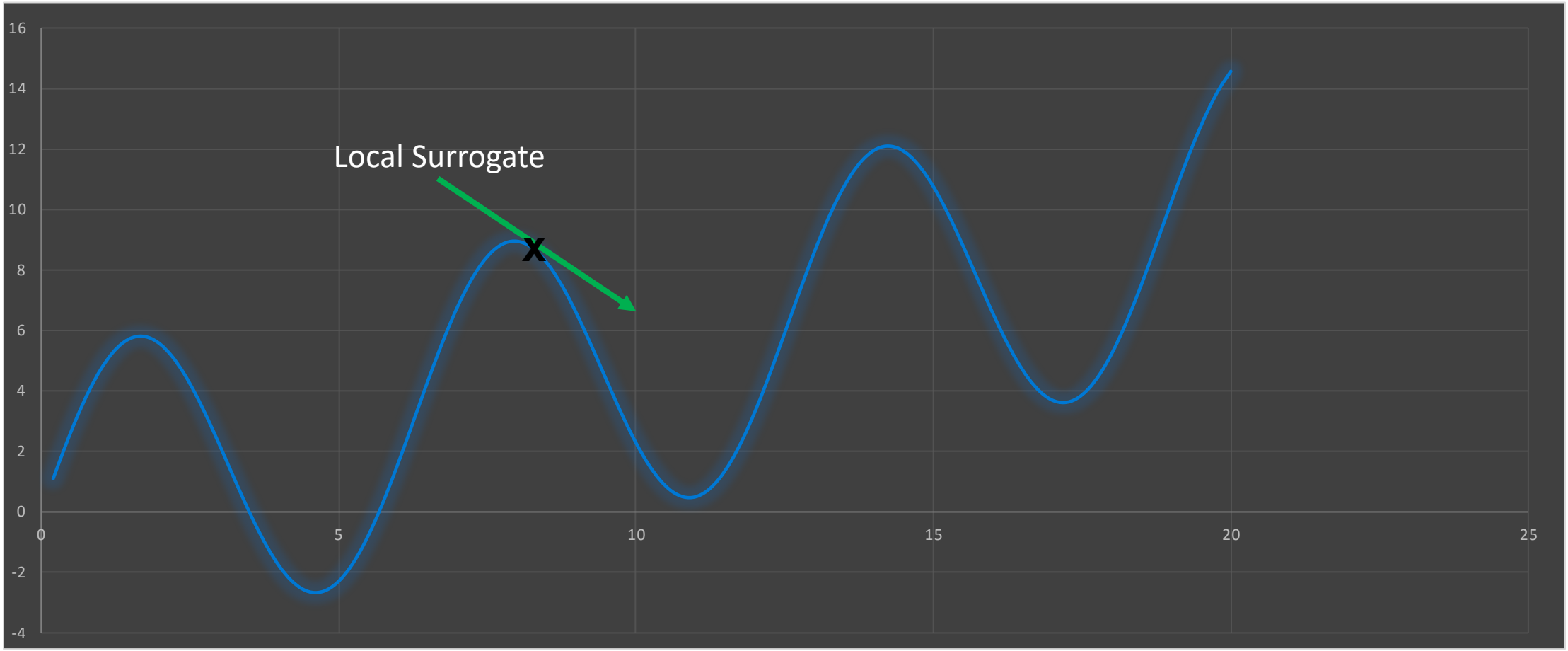
Global Surrogate Models

- Flexibility
- Fidelity measure (e.g., R-squared)
- Easy to explain to non-technical stakeholders

- You draw conclusions about the model and not about the data
- No clear cut-off for R-squared
- The interpretable surrogate model comes with all its +s and –s.

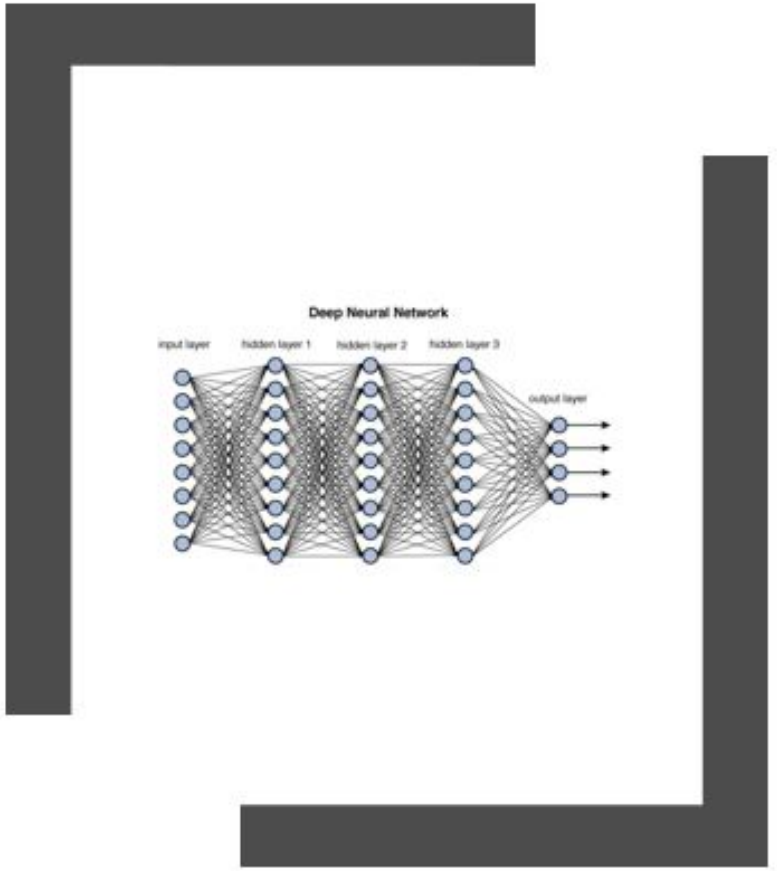

16
14
12
10
8
6
4
2
0
-2
-4
0
5
10
15
20
25
Global Surrogate


16
14
12
10
8
6
4
2
0
-2
-4
0
5
10
15
20
25
X
Global Surrogate


Local Surrogate
X
16
14
12
10
8
6
4
2
0
-2
-4
0
5
10
15
20
25

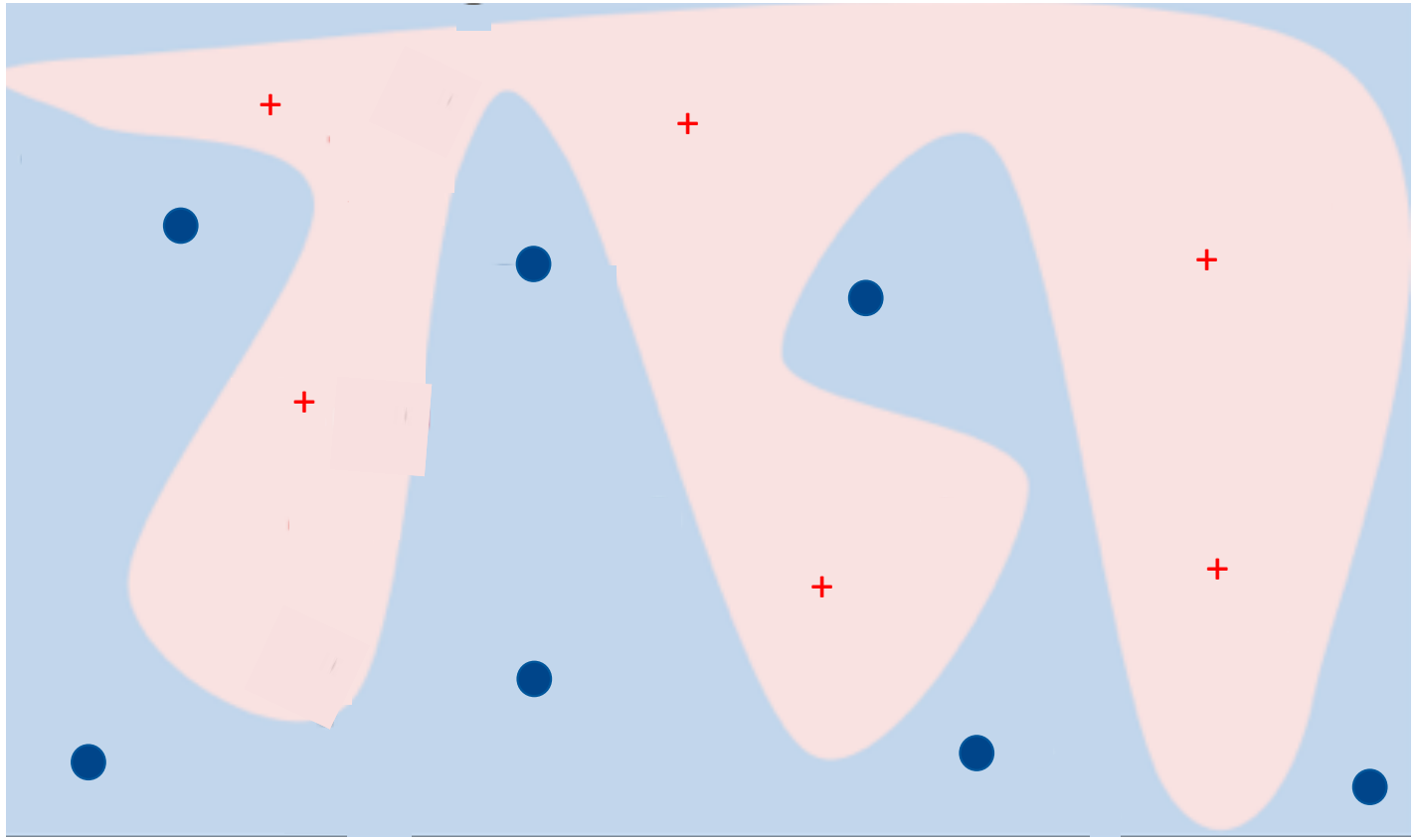
Local Interpretable Model-Agnostic Explanations

LIME

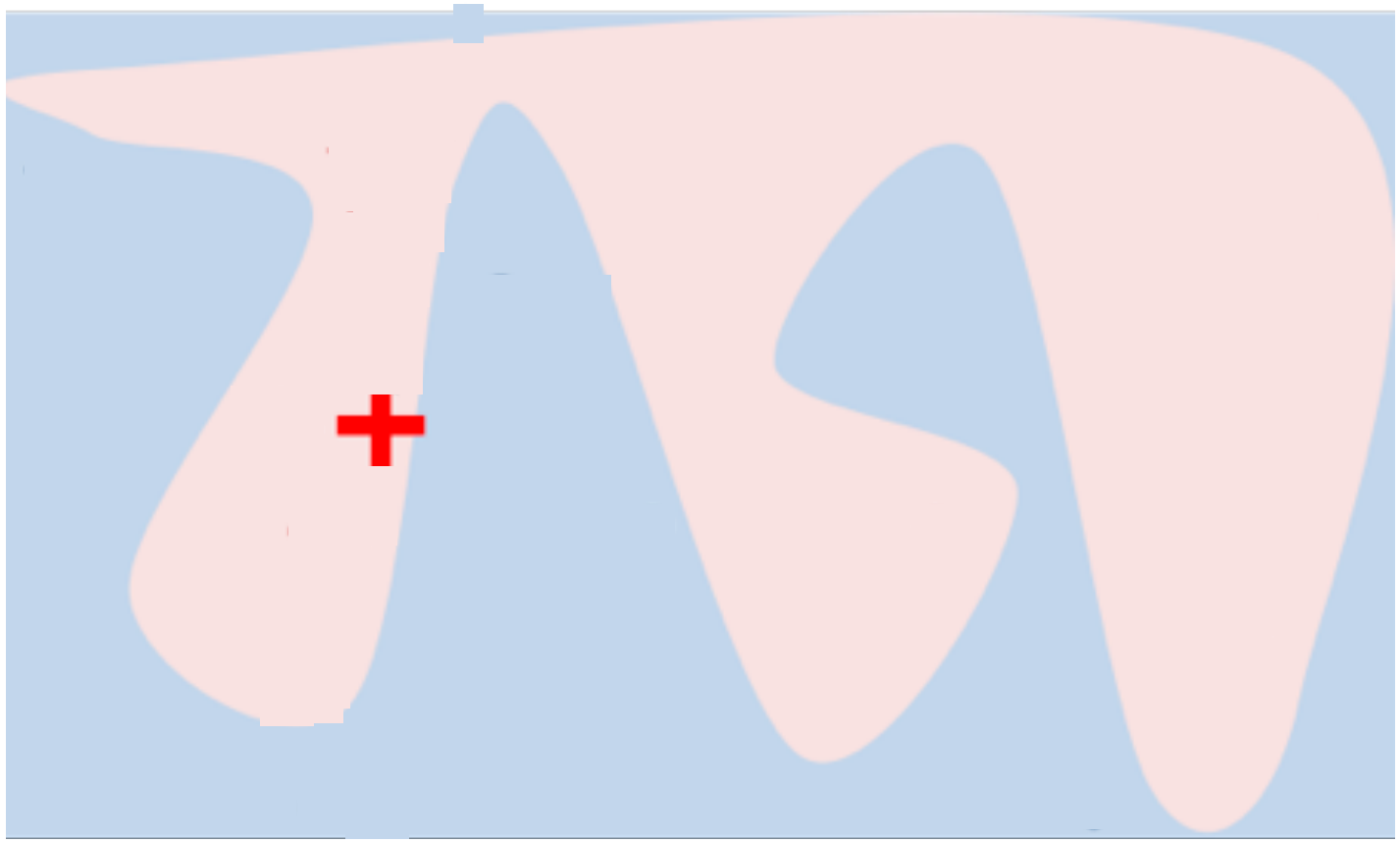


Local Interpretable Model-agnostic Explanations (LIME)

- LIME focuses on training local surrogate models to explain individual predictions.

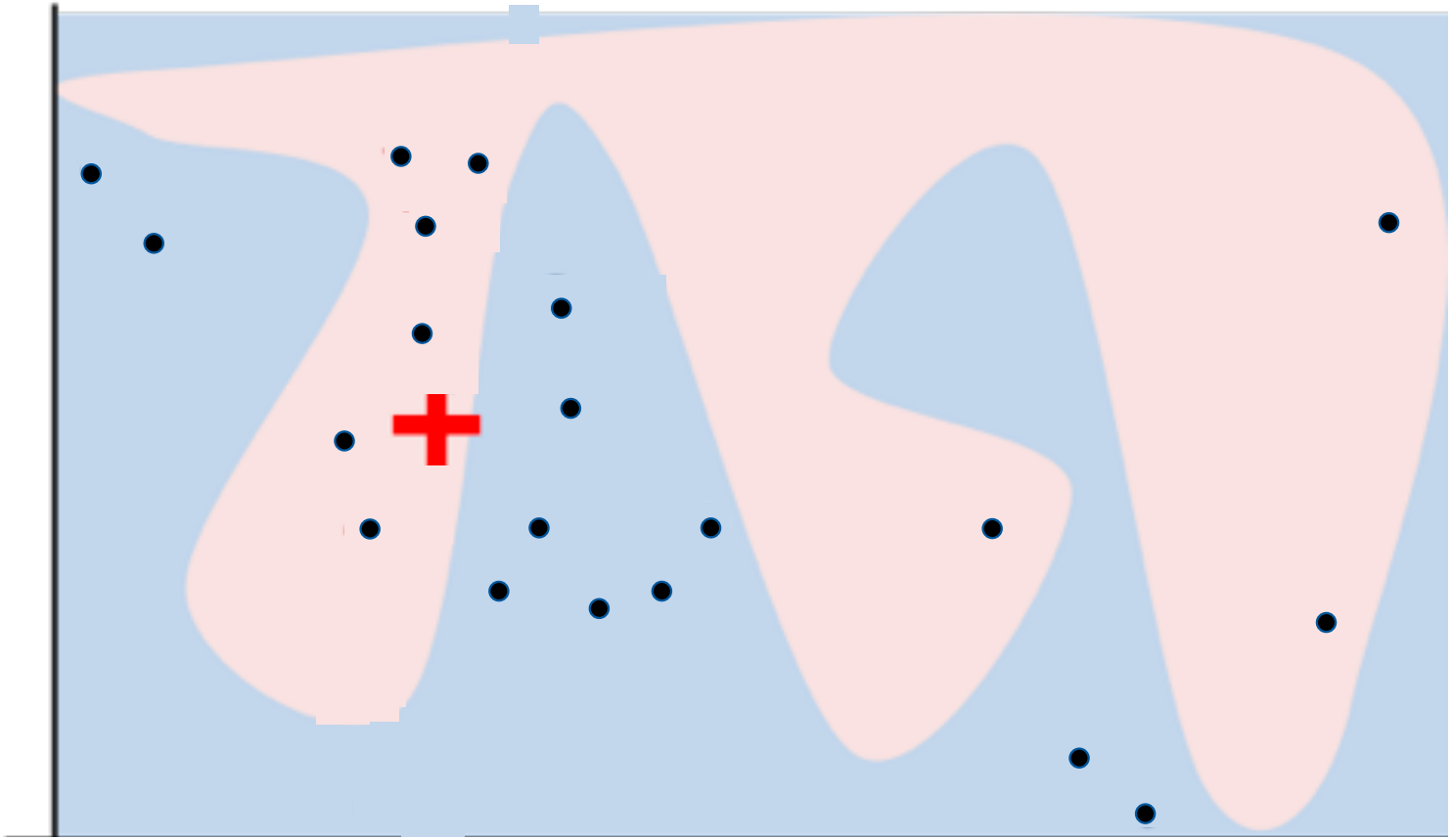


Local Interpretable Model-agnostic Explanations (LIME)



Local Interpretable Model-agnostic Explanations (LIME)

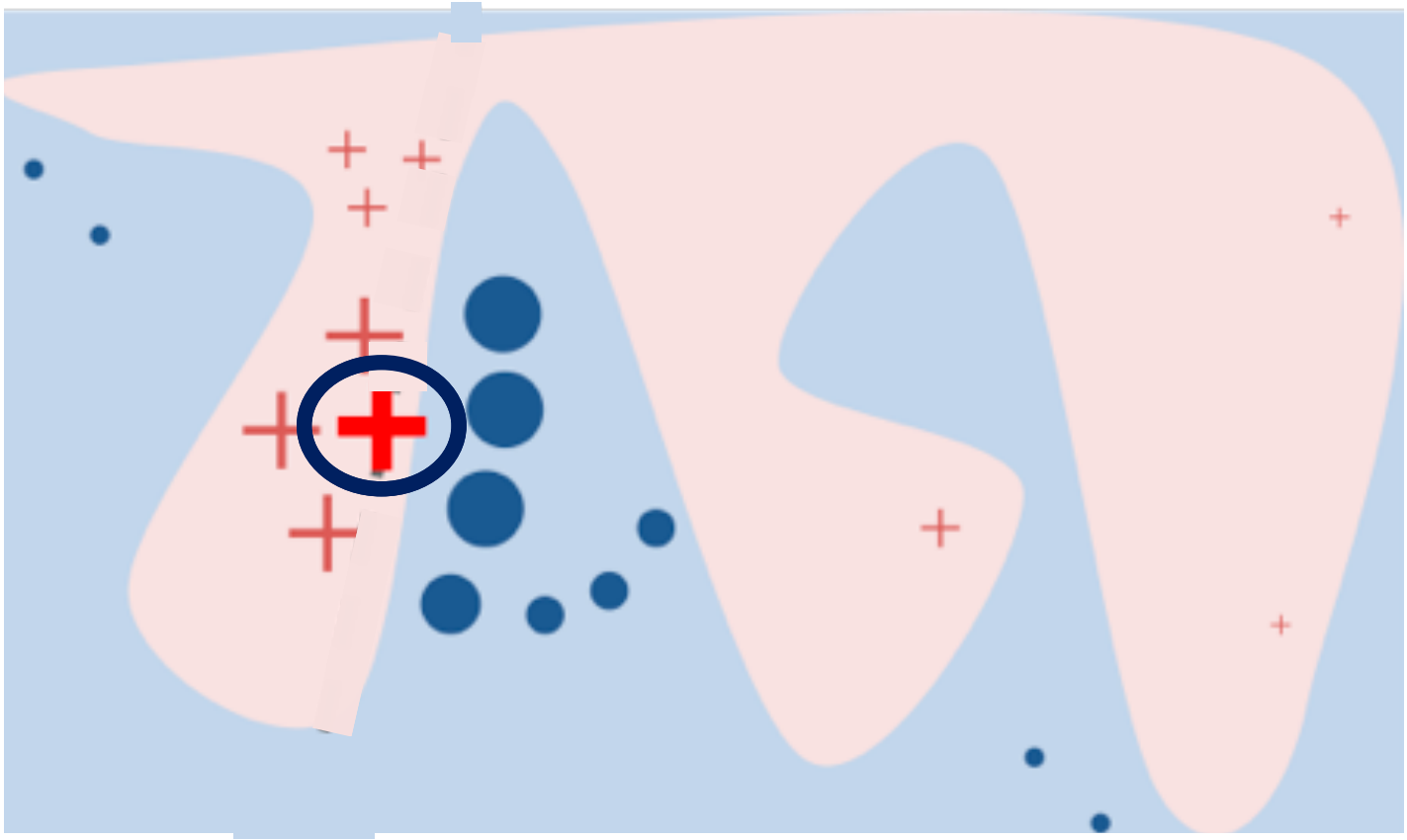
- Permute data (take the point of interest and create fake data around it)



Local Interpretable Model-agnostic Explanations (LIME)

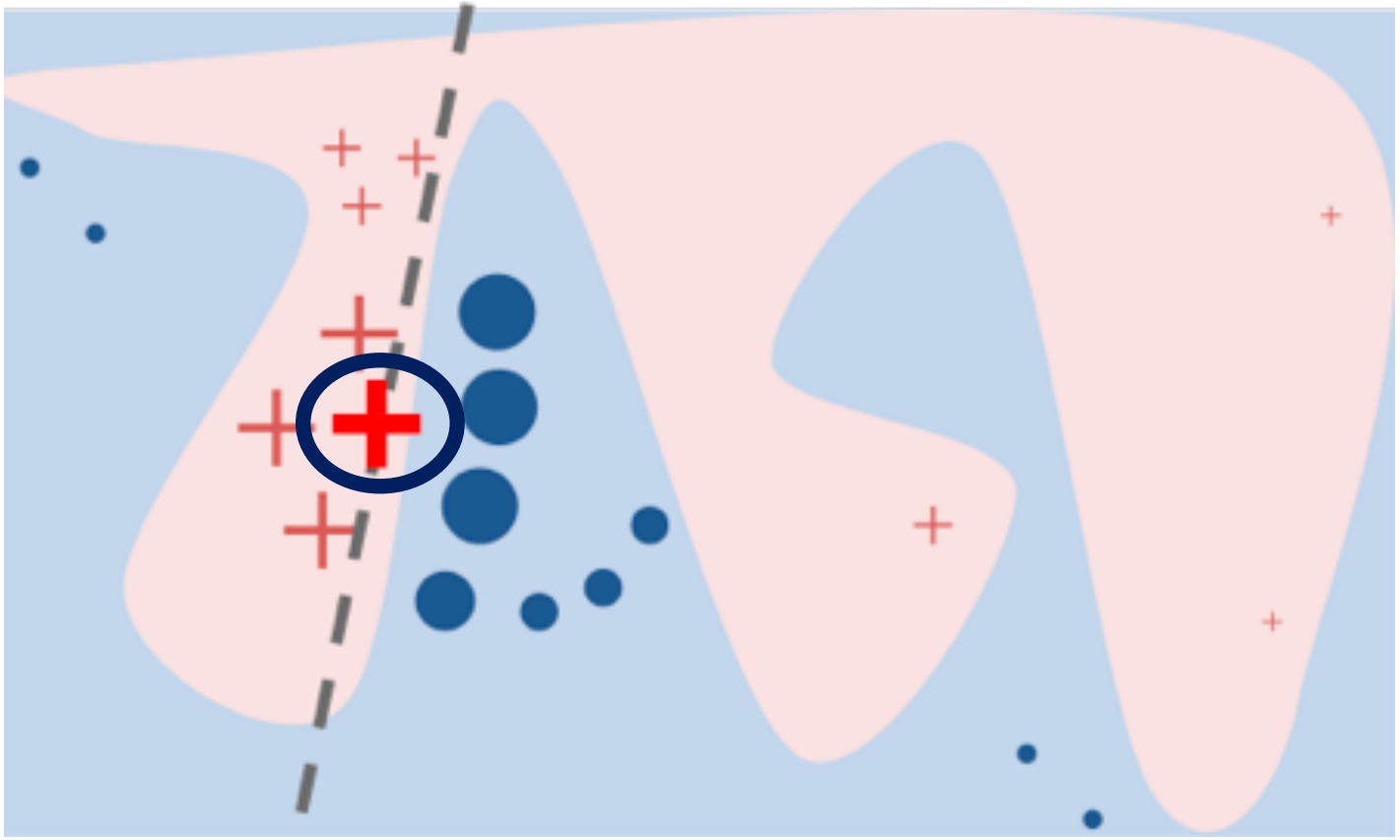
- Take the black box model and make predictions on this new dataset (shown by signs)

- Calculate distance between fake and original data (shown by size)



Local Interpretable Model-agnostic Explanations (LIME)

- Fit an interpretable model on the permuted data with similarity scores

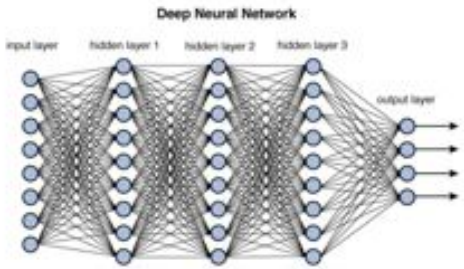


Local Interpretable Model-agnostic Explanations

- Flexibility
- Fidelity measure
- Works for tabular data, text, and image

- Perturbation Stage: **unlikely data points** which can then be used to learn local explanation models.
- **Instability of the explanations**: if you repeat the sampling process, then the explanations that come out can be different.

SHAP


Deep Neural Network
input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer

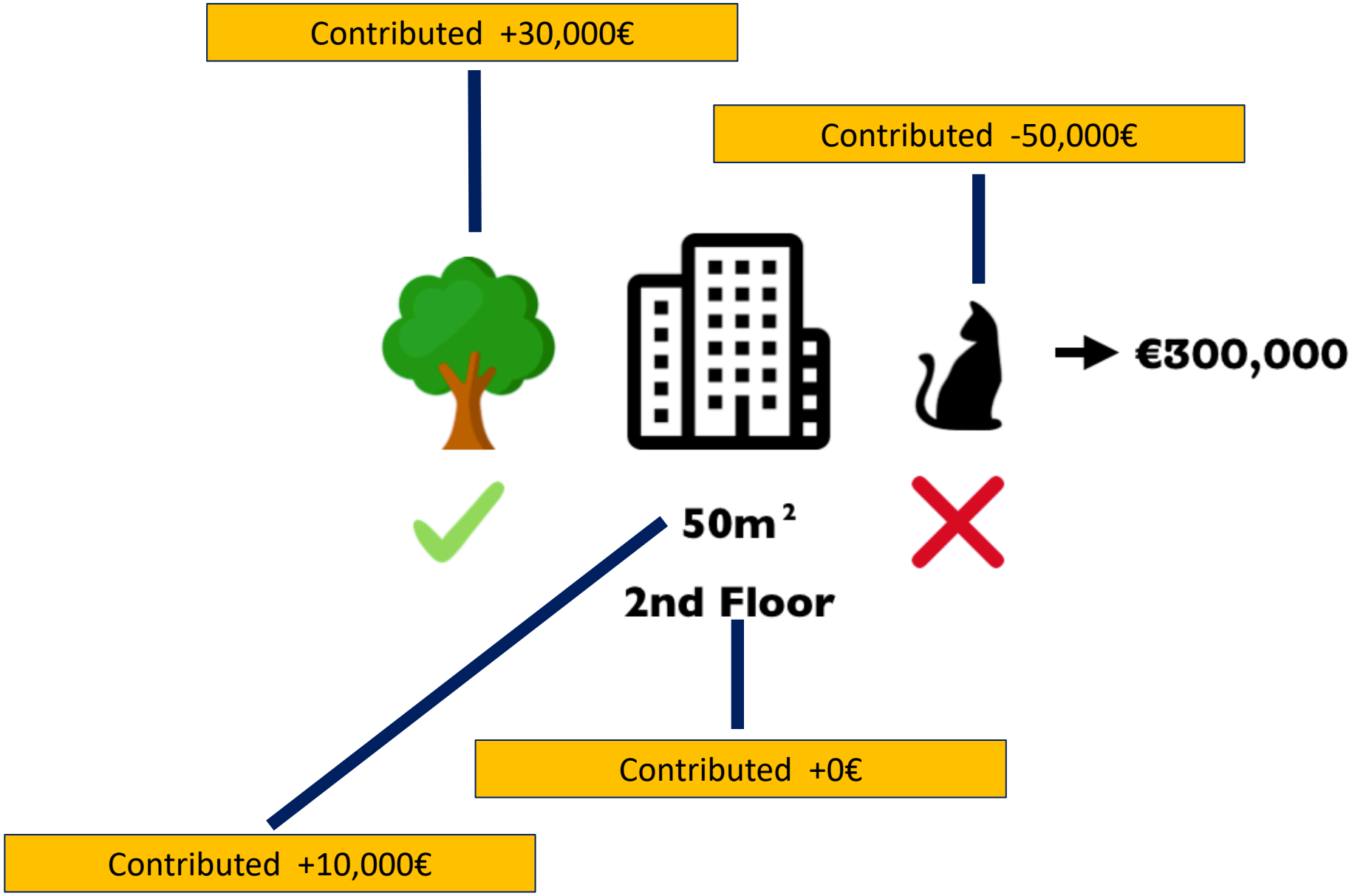
SHAP

- Not a new concept
- Concept based on game theory
- Mathematically sound
- Application in MI relatively new

SHAP

How much has each feature contributed to the prediction compared to the average prediction?

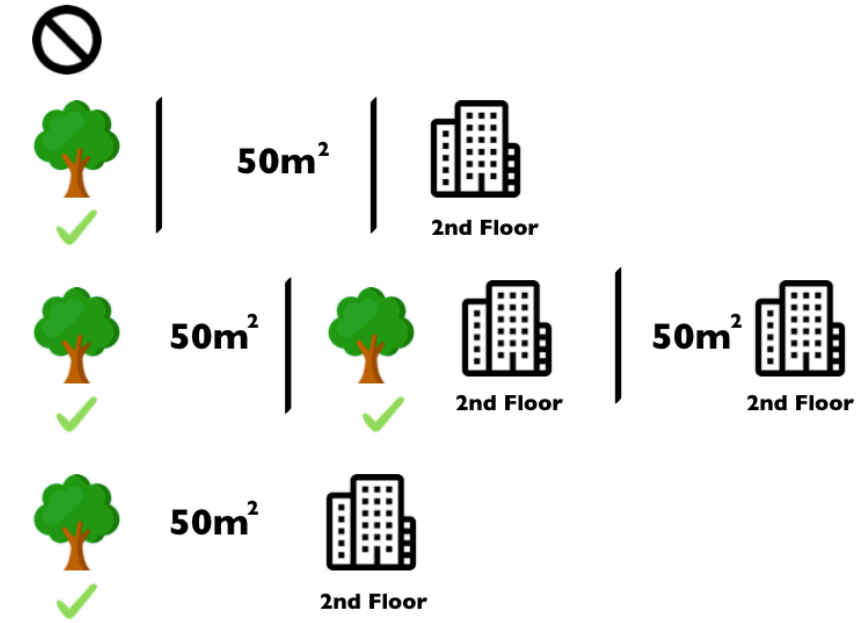
- House price prediction: 300,000€
- Average house price prediction for all apartments is 310,000€
- Delta here is -10,000€

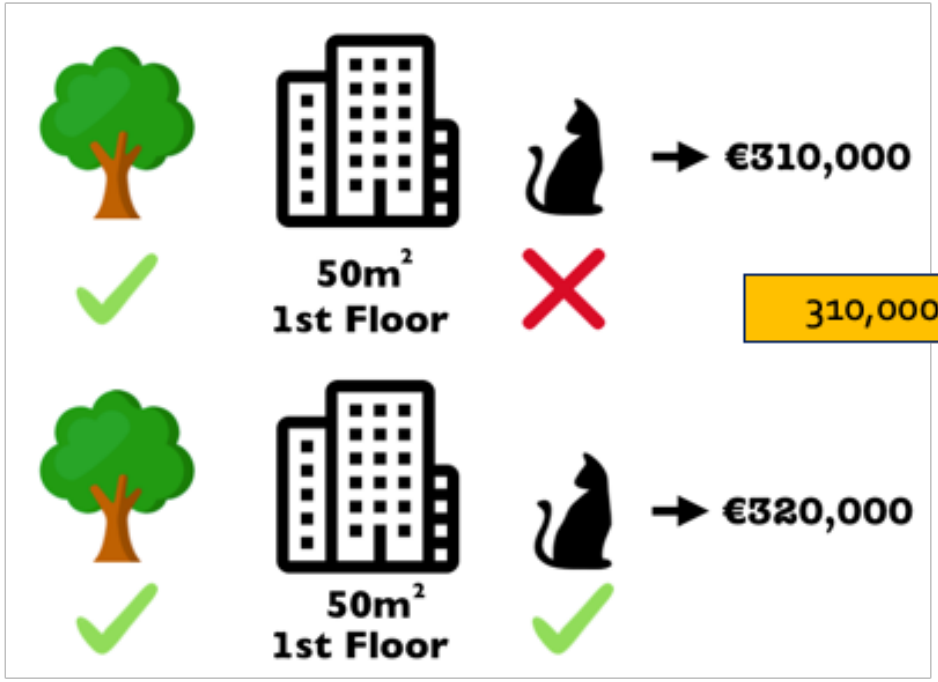

Contributed +30,000€
Contributed -50,000€
€300,000
50m²
2nd Floor
Contributed +0€
Contributed +10,000€

SHAP

How do we calculate Shapley values?

- Take your feature of interest (e.g., cats-banned) and remove it from the feature set
- Take the remaining features and generate all possible coalitions
- Add and remove your feature of interested to each of the coalitions and calculate the difference it makes





310,000 – 320,000 = -10,000€

SHAP

- Based on a solid theory and distributes the effects fairly
- **Contrastive explanations**. Instead of comparing a prediction to the average prediction of the entire dataset, you could compare it to a subset or even to a single data point.

- Computation time: 2^k possible coalitions of the feature values for k features
- **Can be misinterpreted**
- **Inclusion of unrealistic data instances** when features are correlated.

Machine Learning Interpretability