

# Data Science Practicum

**(Lecture 7, 30.10.)**

Denisa Šrámková



# eXplainable Artificial Intelligence (XAI)

- Terminology
- Global explainability
  - Feature importance
- Local explainability
  - SHAP

# Motivation

NEWS | 24 October 2019 | Update [26 October 2019](#)

## Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

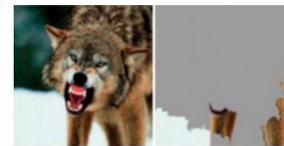
[Heidi Ledford](#)



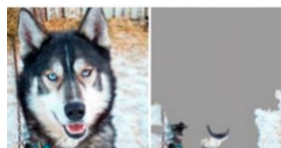
Predicted: **wolf**  
True: **wolf**



Predicted: **husky**  
True: **husky**



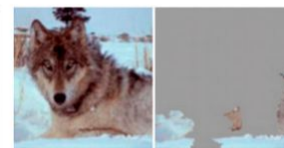
Predicted: **wolf**  
True: **wolf**



Predicted: **wolf**  
True: **husky**



Predicted: **husky**  
True: **husky**



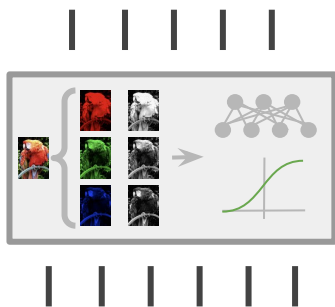
Predicted: **wolf**  
True: **wolf**

<https://www.compact.nl/en/articles/deep-learning-finding-that-perfect-fit/>

# Terminology

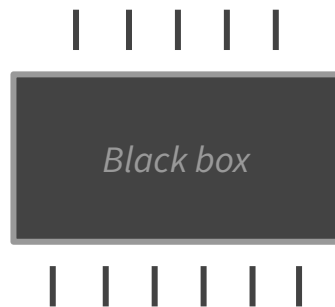
## Interpretability

The model is transparent in its operation and provides information about relationships between inputs and outputs.



## Explainability

The model provides clear and intuitive explanation of the decisions made, enabling us to understand why the model produced a particular result.

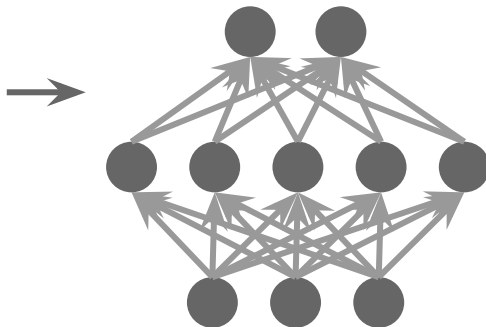


# Example

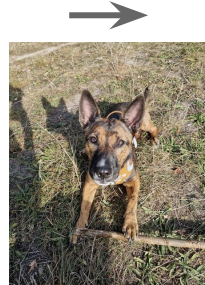
Input data



Model



Output



71.51% “American stafford terrier”

9.01% “Border collie”

1.54% “Wiener dog”

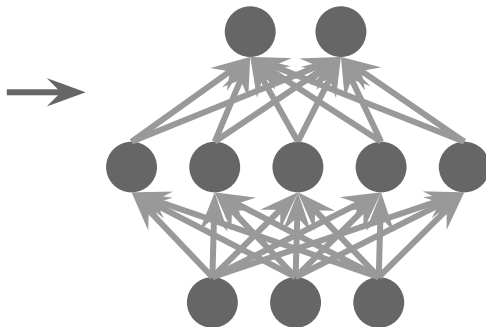
...

# Example

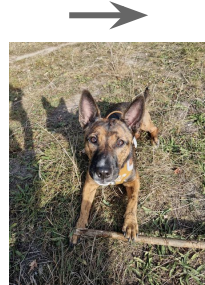
Input data



Model



Output



71.51% “American stafford terrier”

9.01% “Border collie”

1.54% “Wiener dog”

...

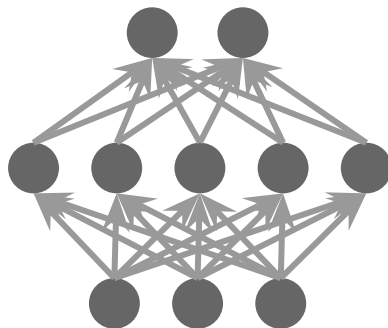
Current explanation: This is an  
American stafford terrier.

# Example

Input data

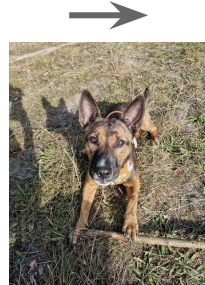


Model



XAI methods  
+ domain  
knowledge

Output



71.51% “American stafford terrier”  
9.01% “Border collie”  
1.54% “Wiener dog”  
...

Current explanation: This is an  
American stafford terrier.

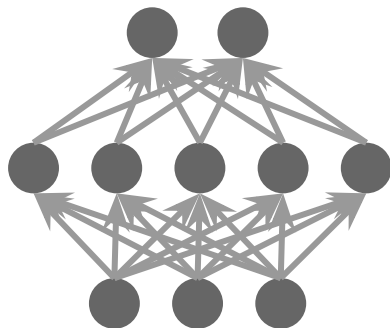


# Example

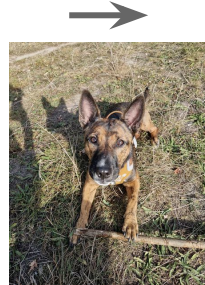
Input data



Model



Output



71.51% “American stafford terrier”

9.01% “Border collie”

1.54% “Wiener dog”

...

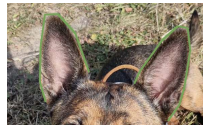
Current explanation: This is an  
American stafford terrier.

+

XAI methods  
+ domain  
knowledge

XAI explanations:

- it has pointy ears, round black nose, tiny paws



- it has this feature



# Types of explainability methods

1. Is my model directly interpretable?

ML model

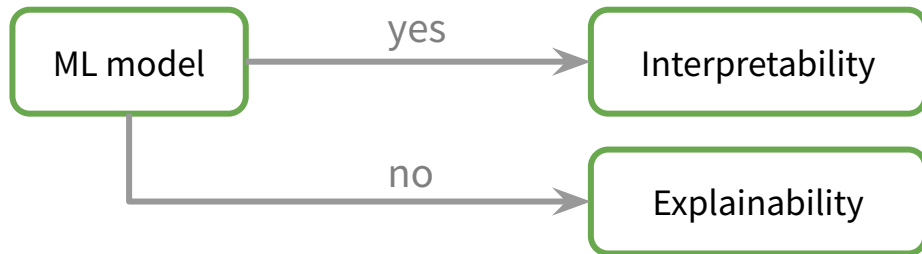
# Types of explainability methods

1. Is my model directly interpretable?



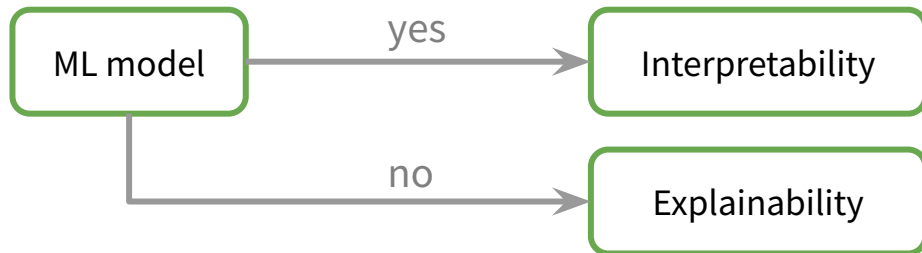
# Types of explainability methods

1. Is my model directly interpretable?



# Types of explainability methods

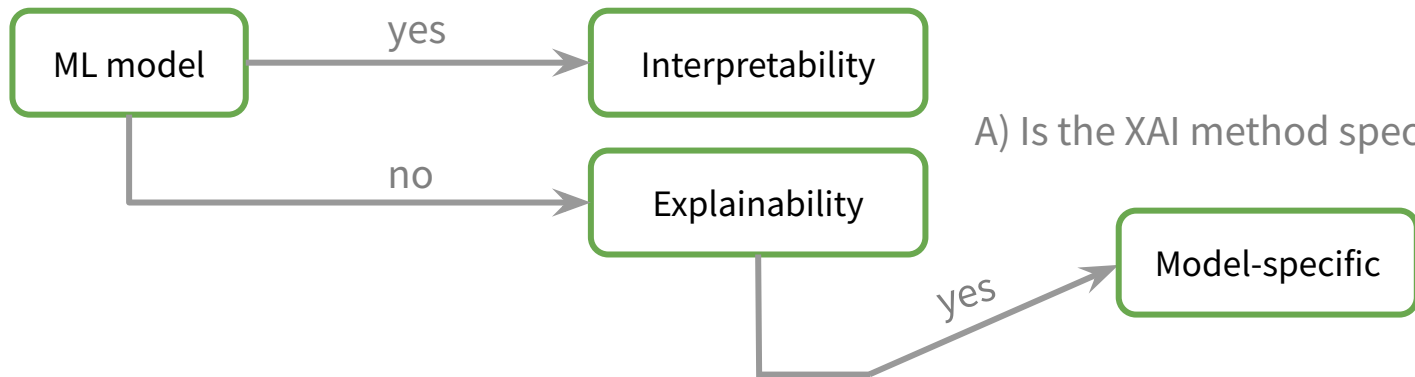
1. Is my model directly interpretable?



A) Is the XAI method specific for the model type?

# Types of explainability methods

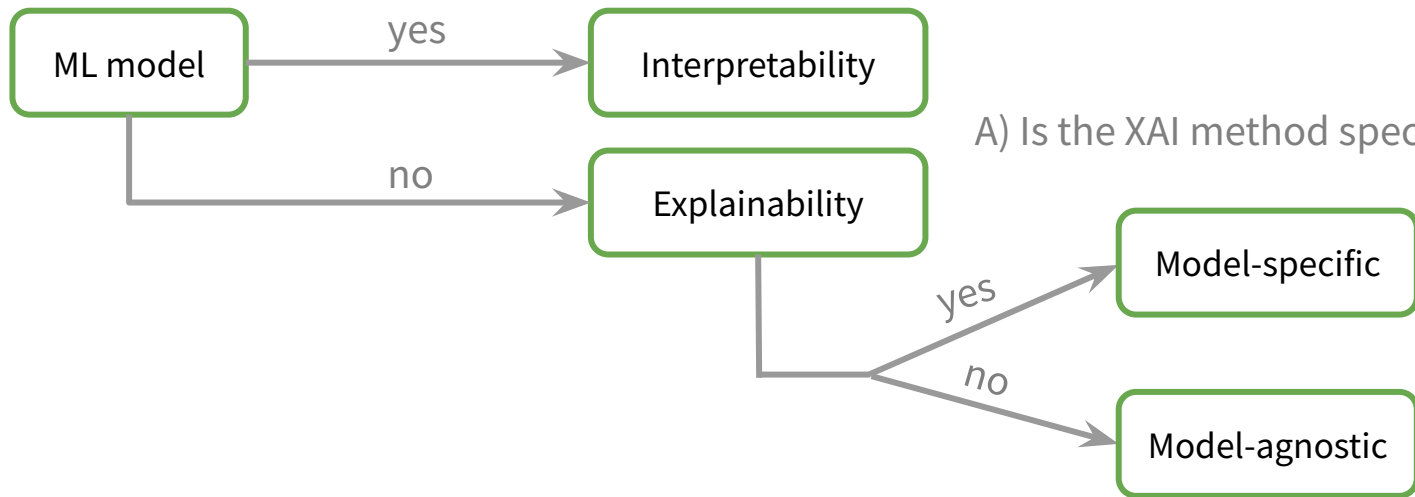
1. Is my model directly interpretable?



A) Is the XAI method specific for the model type?

# Types of explainability methods

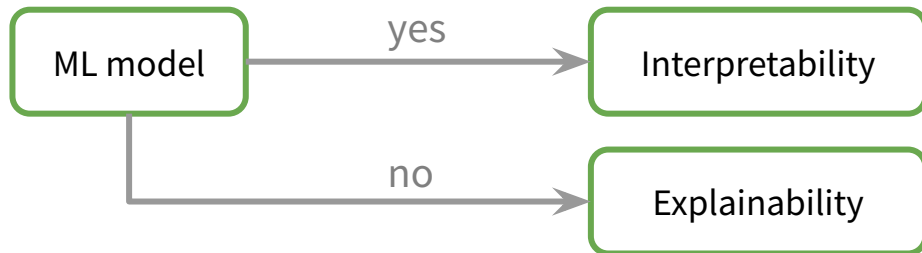
1. Is my model directly interpretable?



A) Is the XAI method specific for the model type?

# Types of explainability methods

1. Is my model directly interpretable?

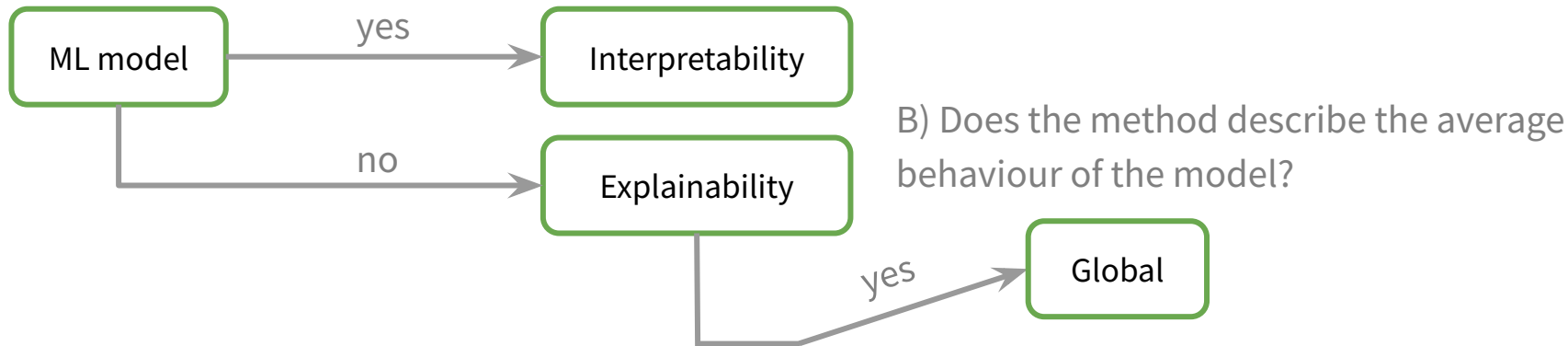


B) Does the method describe the average behaviour of the model?



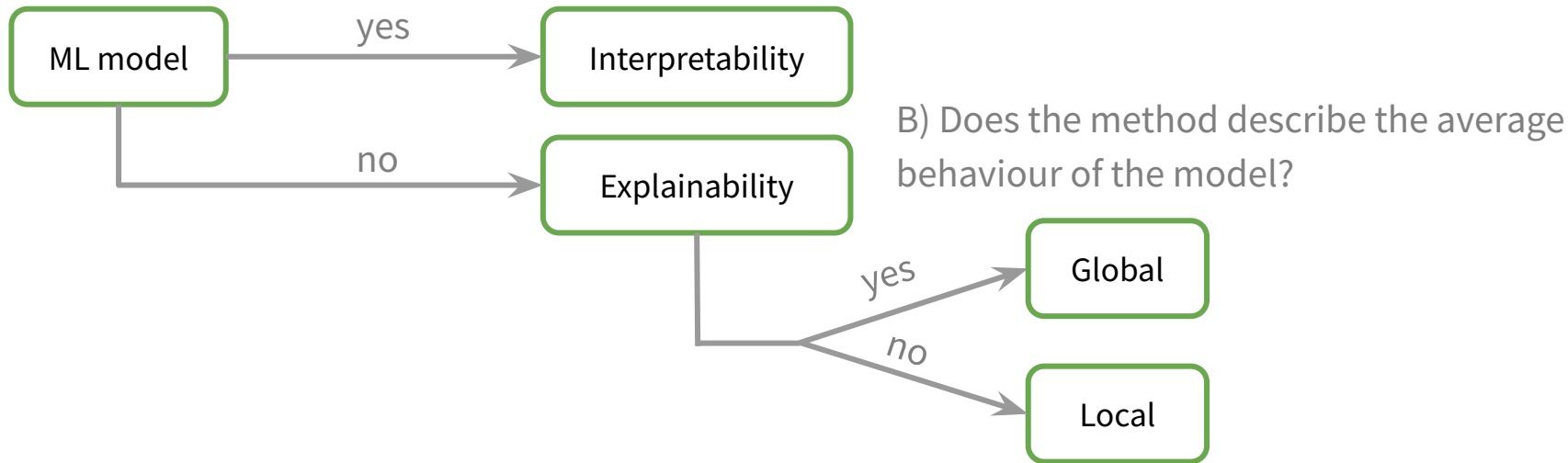
# Types of explainability methods

1. Is my model directly interpretable?

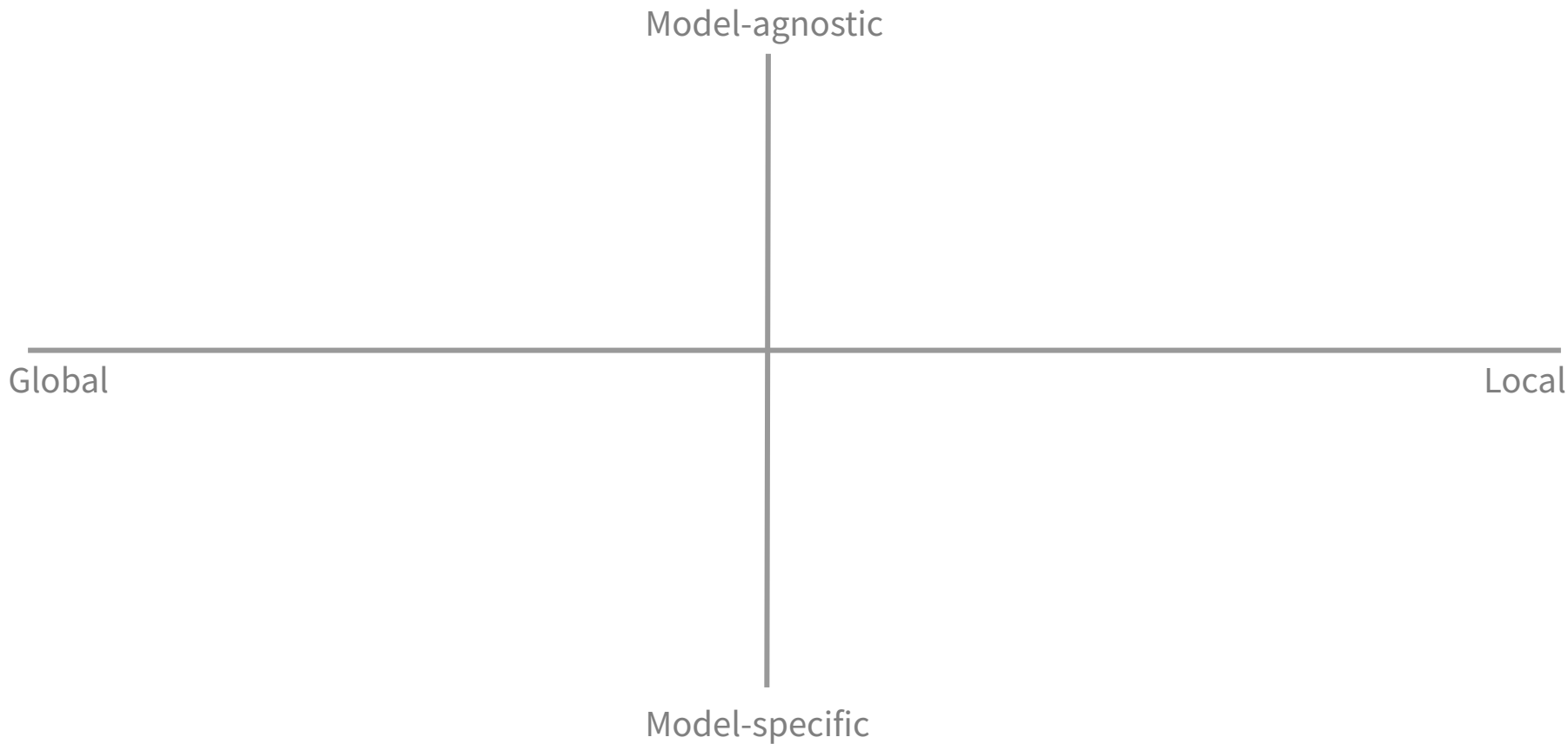


# Types of explainability methods

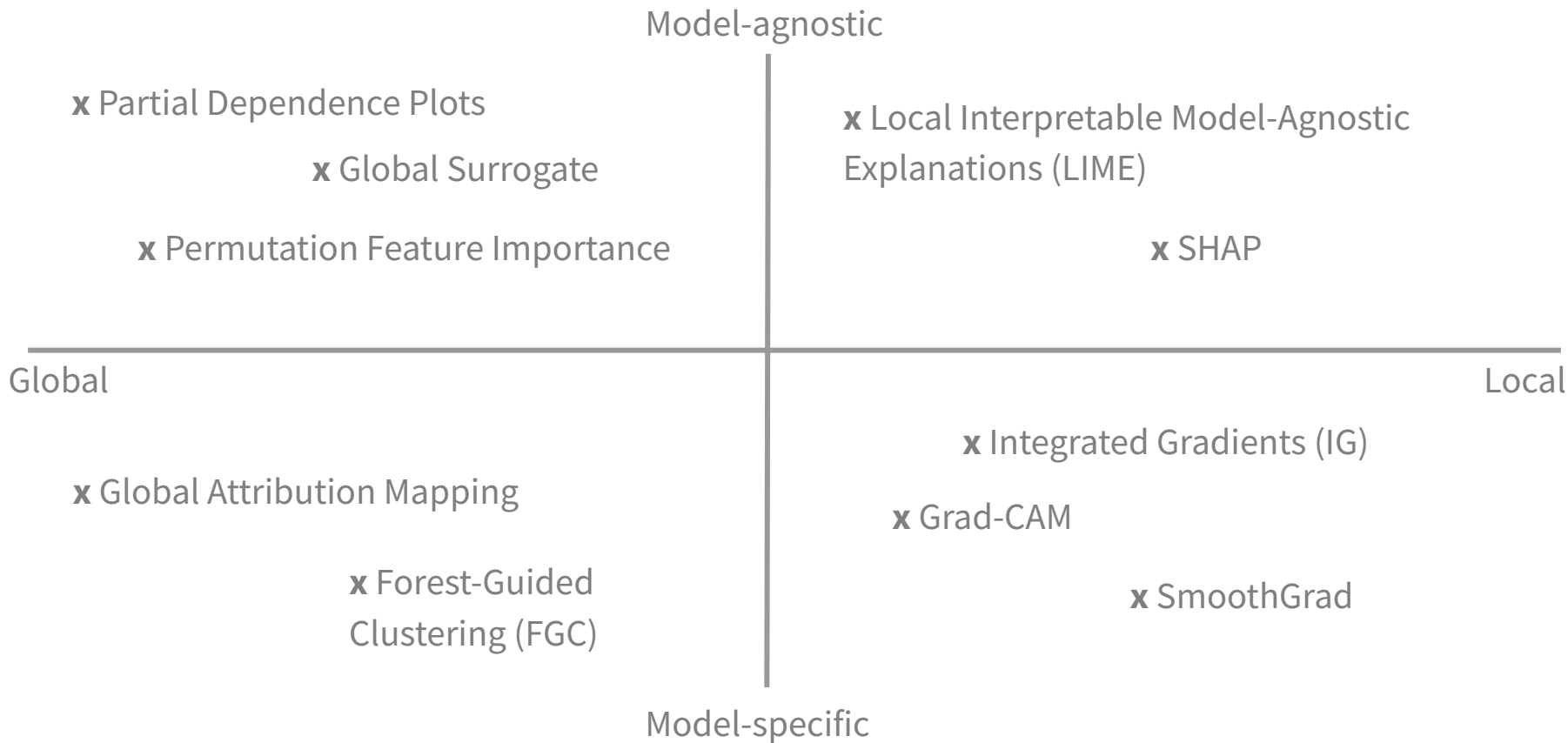
1. Is my model directly interpretable?



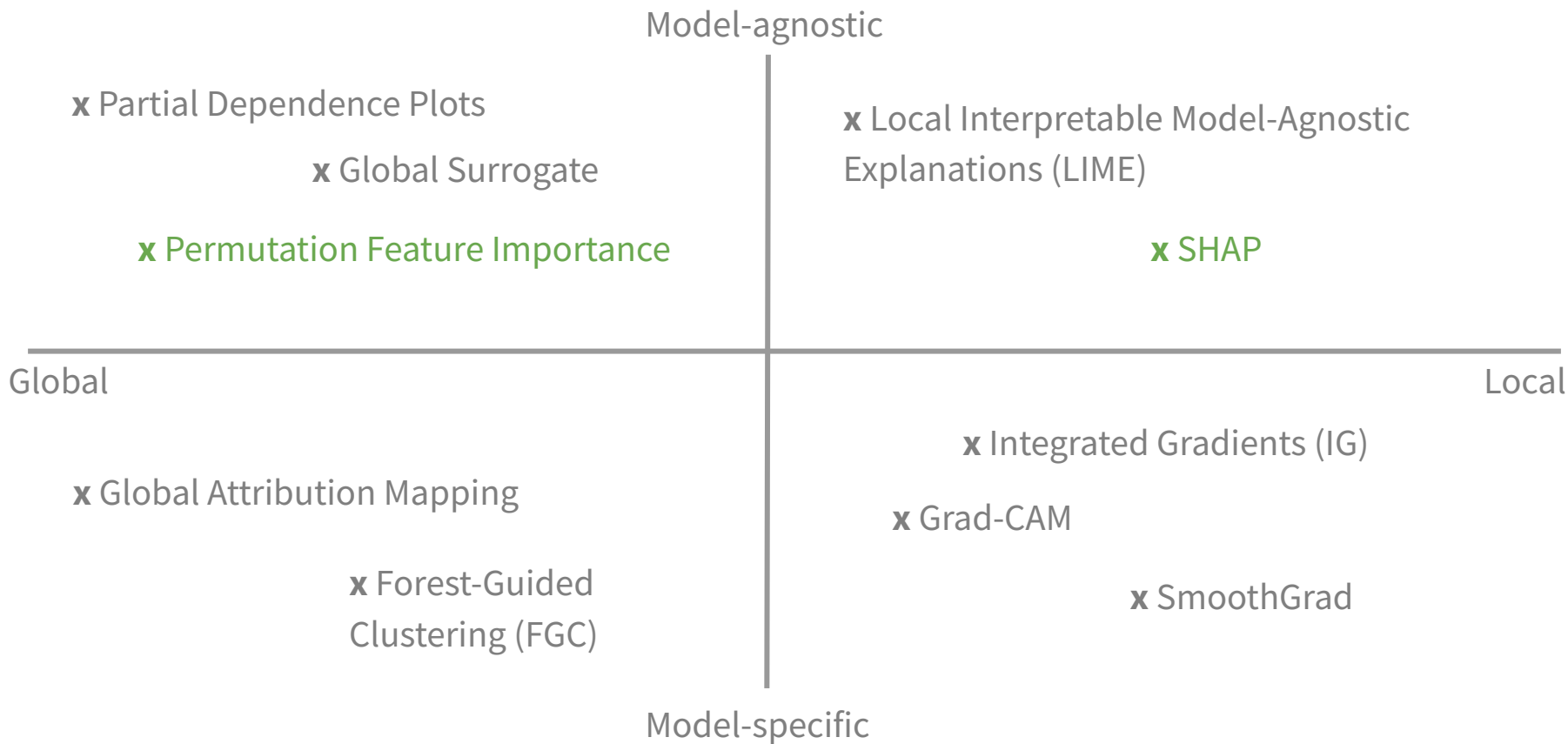
# Concrete methods



# Concrete methods



# Concrete methods















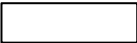




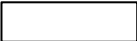




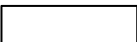







# Permutation Feature Importance

We measure the **importance of a feature as an increase in loss** when the feature is permuted.

# Permutation Feature Importance















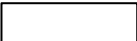




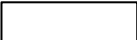




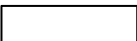





We measure the **importance of a feature as an increase in loss** when the feature is permuted.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					



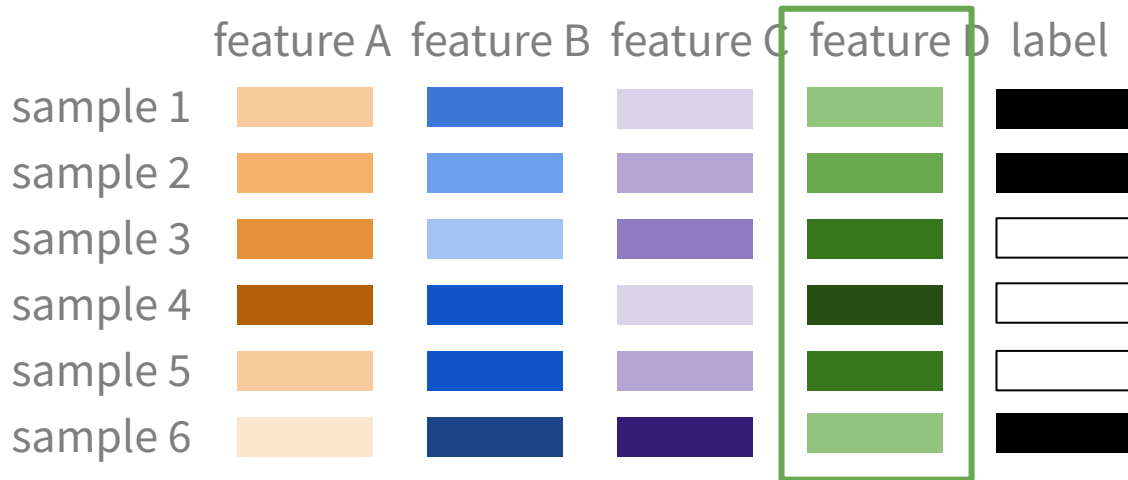
# Permutation Feature Importance

We measure the **importance of a feature as an increase in loss** when the feature is permuted.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					

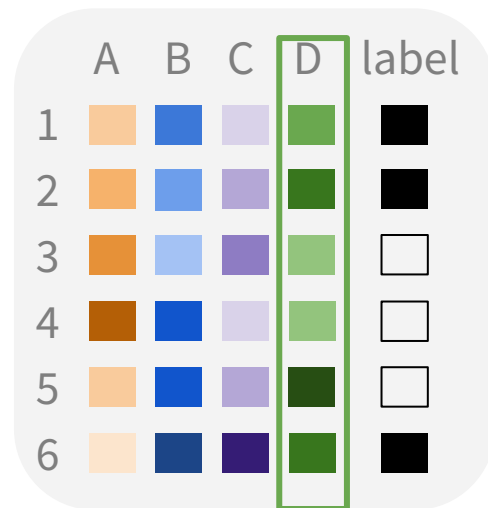
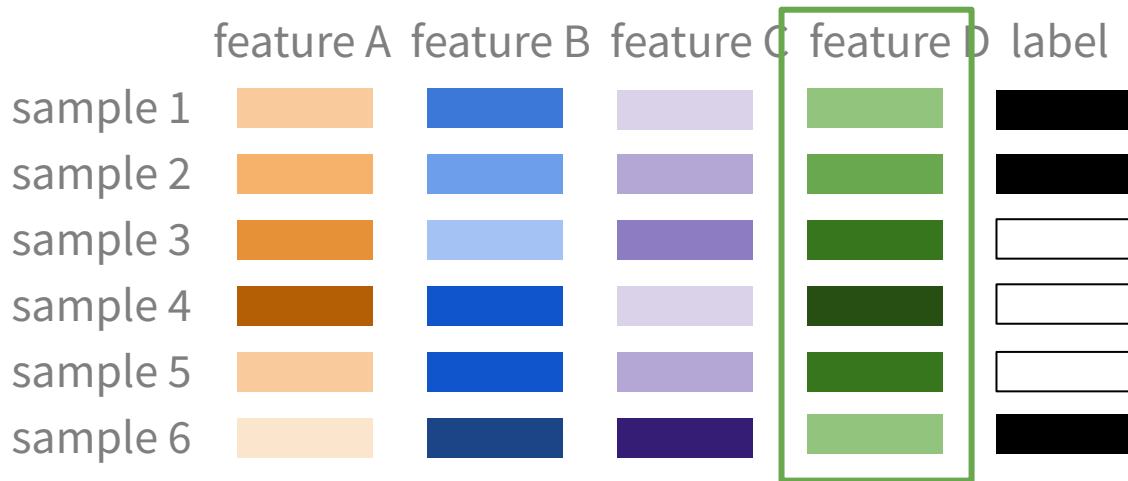
# Permutation Feature Importance

We measure the **importance of a feature as an increase in loss** when the feature is permuted.

































# Permutation Feature Importance































We measure the **importance of a feature as an increase in loss** when the feature is permuted.



# Permutation Feature Importance





































We measure the **importance of a feature as an increase in loss** when the feature is permuted.














	A	B	C	D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					

	A	B	C	D	label
1					
2					
3					
4					
5					
6					

# Permutation Feature Importance

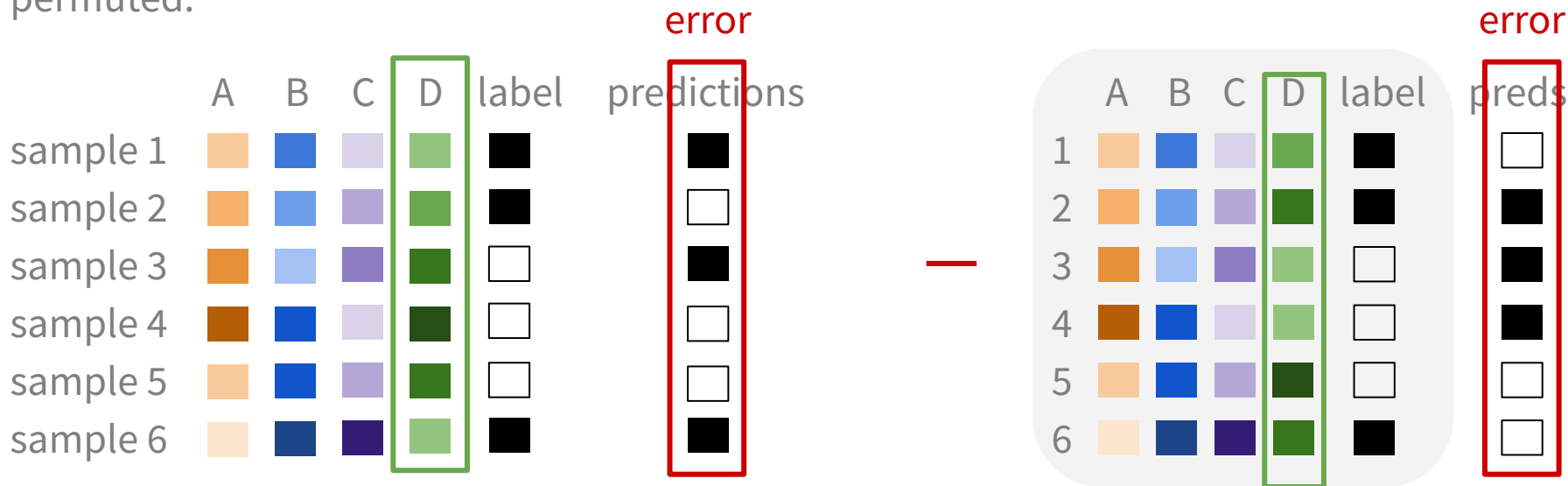
We measure the **importance of a feature as an increase in loss** when the feature is permuted.

	A	B	C	D	label	predictions
sample 1						
sample 2						
sample 3						
sample 4						
sample 5						
sample 6						

	A	B	C	D	label	preds
1						
2						
3						
4						
5						
6						

# Permutation Feature Importance

We measure the **importance of a feature** as an **increase in loss** when the feature is permuted.



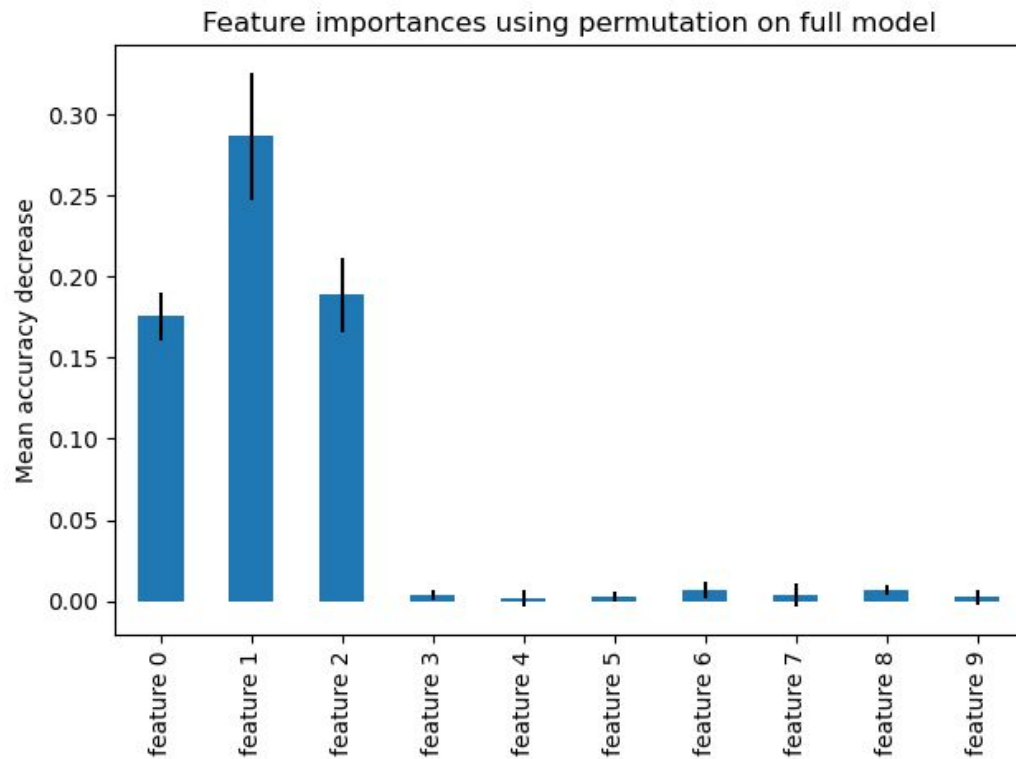
# Permutation Feature Importance

We measure the **importance of a feature as an increase in loss** when the feature is permuted.

Note: Feature independence is assumed (features are not correlated). We might observe that the importance of highly correlated features will be lower (because the importance gets split between correlated features).



# Interpreting Feature Importance results



# Exercise 1

Tabular data:

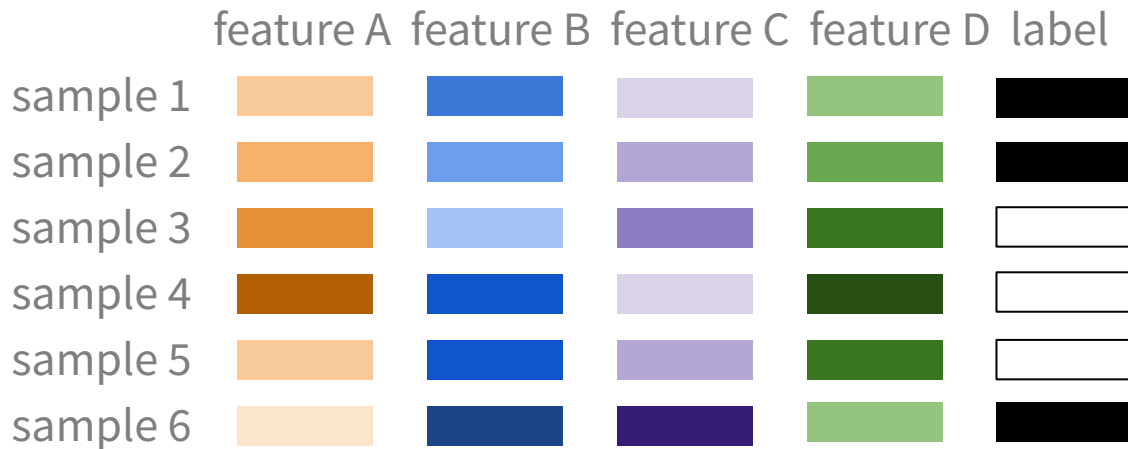
[https://github.com/simecek/dspracticum2023/blob/main/lesson07/ds\\_practicum\\_ex1\\_feature\\_importance.ipynb](https://github.com/simecek/dspracticum2023/blob/main/lesson07/ds_practicum_ex1_feature_importance.ipynb)

# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.















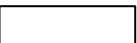




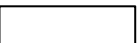




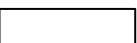





# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.









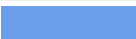







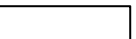




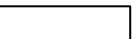




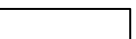





# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					

# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.







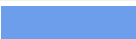







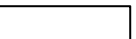




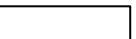




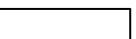





	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					



marginal contribution  $\phi(A)$   
for different feature sets  $s$

# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					







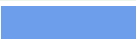







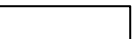




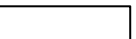




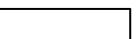







marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$

# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					



marginal contribution  $\Phi(A)$   
for different feature sets  $s$







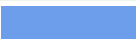







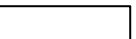




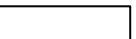




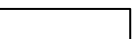





$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$

~~A~~ B C D | ~~B~~ C D




# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					







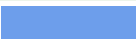







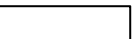




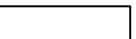




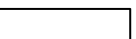







marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$



# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					







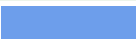







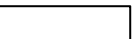




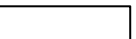




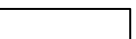







marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$



# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					







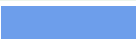







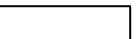




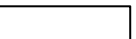




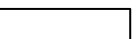







marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$



# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					







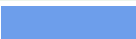







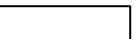




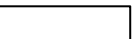




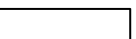







marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$



# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					







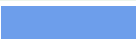







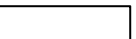




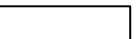




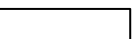







marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$



# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					







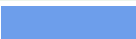







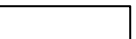




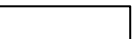




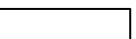







marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$



# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					









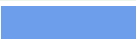







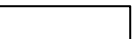




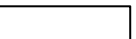




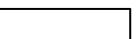





marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$


The diagram shows the equation  $\Phi(A) = f(s_{+A}) - f(s_{-A})$ . Below the terms, there are dashed lines representing the feature sets. For  $s_{+A}$ , a line labeled 'A' connects to the set, indicating feature A is present. For  $s_{-A}$ , a line connects to the set, indicating feature A is absent.

# SHapley Additive exPlanations (SHAP)

We compute the **contribution of each feature** to the model prediction **for some specific input**.

	feature A	feature B	feature C	feature D	label
sample 1					
sample 2					
sample 3					
sample 4					
sample 5					
sample 6					



marginal contribution  $\Phi(A)$   
for different feature sets  $s$

$$\Phi(A) = f(s_{+A}) - f(s_{-A})$$

Shapley value (contribution):  $\Phi(A) = \sum_{s \in S} w_s \Phi_s(A)$



# SHAP paper

## 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

### Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

### 1 Introduction

The ability to correctly interpret a prediction model's output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled. In some applications, simple models (e.g., linear models) are often preferred for their ease of interpretation, even if they may be less accurate than complex ones.

### 3 Simple Properties Uniquely Determine Additive Feature Attributions

A surprising attribute of the class of additive feature attribution methods is the presence of a single unique solution in this class with three desirable properties (described below). While these properties are familiar to the classical Shapley value estimation methods, they were previously unknown for other additive feature attribution methods.

The first desirable property is *local accuracy*. When approximating the original model  $f$  for a specific input  $x$ , local accuracy requires the explanation model to at least match the output of  $f$  for the simplified input  $x'$  (which corresponds to the original input  $x$ ).

#### Property 1 (Local accuracy)

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^M \phi_i x'_i \quad (5)$$

The explanation model  $g(x')$  matches the original model  $f(x)$  when  $x = h_x(x')$ , where  $\phi_0 = f(h_x(0))$  represents the model output with all simplified inputs toggled off (i.e. missing).

The second property is *missingness*. If the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact. All of the methods described in Section 2 obey the missingness property.

#### Property 2 (Missingness)

$$x'_i = 0 \implies \phi_i = 0 \quad (6)$$

Missingness constrains features where  $x'_i = 0$  to have no attributed impact.

The third property is *consistency*. Consistency states that if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.

**Property 3 (Consistency)** Let  $f_x(z') = f(h_x(z'))$  and  $z' \setminus i$  denote setting  $z'_i = 0$ . For any two models  $f$  and  $f'$ , if

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \quad (7)$$

for all inputs  $z' \in \{0, 1\}^M$ , then  $\phi_i(f', x) \geq \phi_i(f, x)$ .

**Theorem 1** Only one possible explanation model  $g$  follows Definition 1 and satisfies Properties 1, 2, and 3:

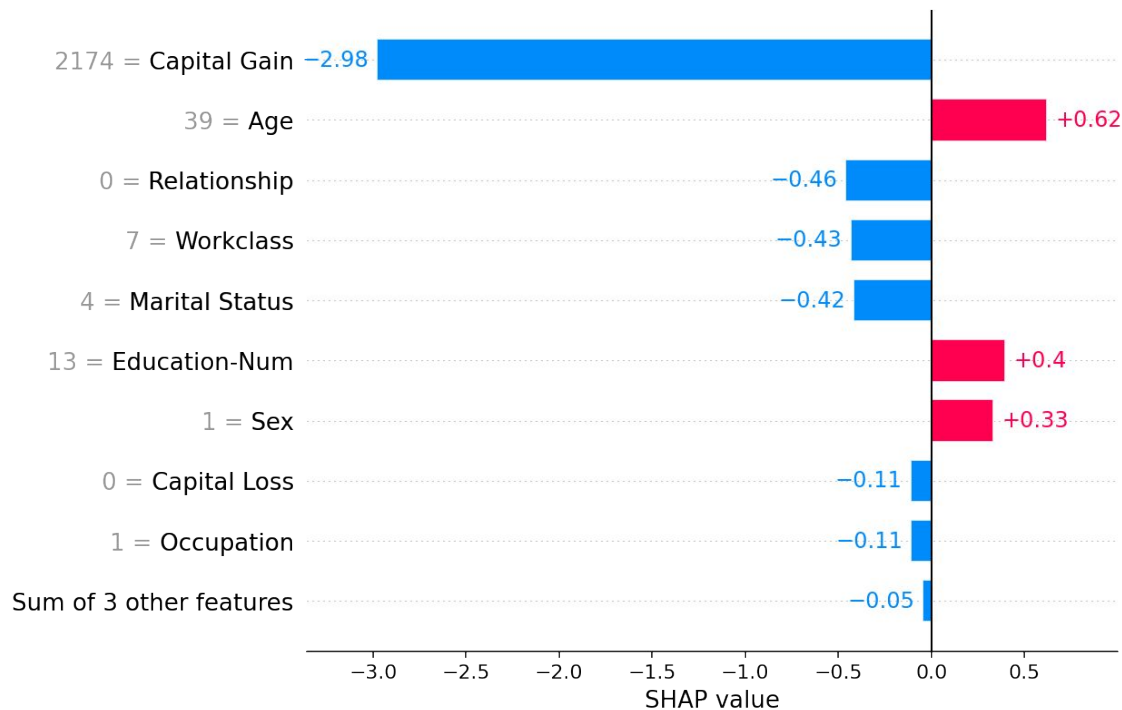
$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \quad (8)$$

where  $|z'|$  is the number of non-zero entries in  $z'$ , and  $z' \subseteq x'$  represents all  $z'$  vectors where the non-zero entries are a subset of the non-zero entries in  $x'$ .

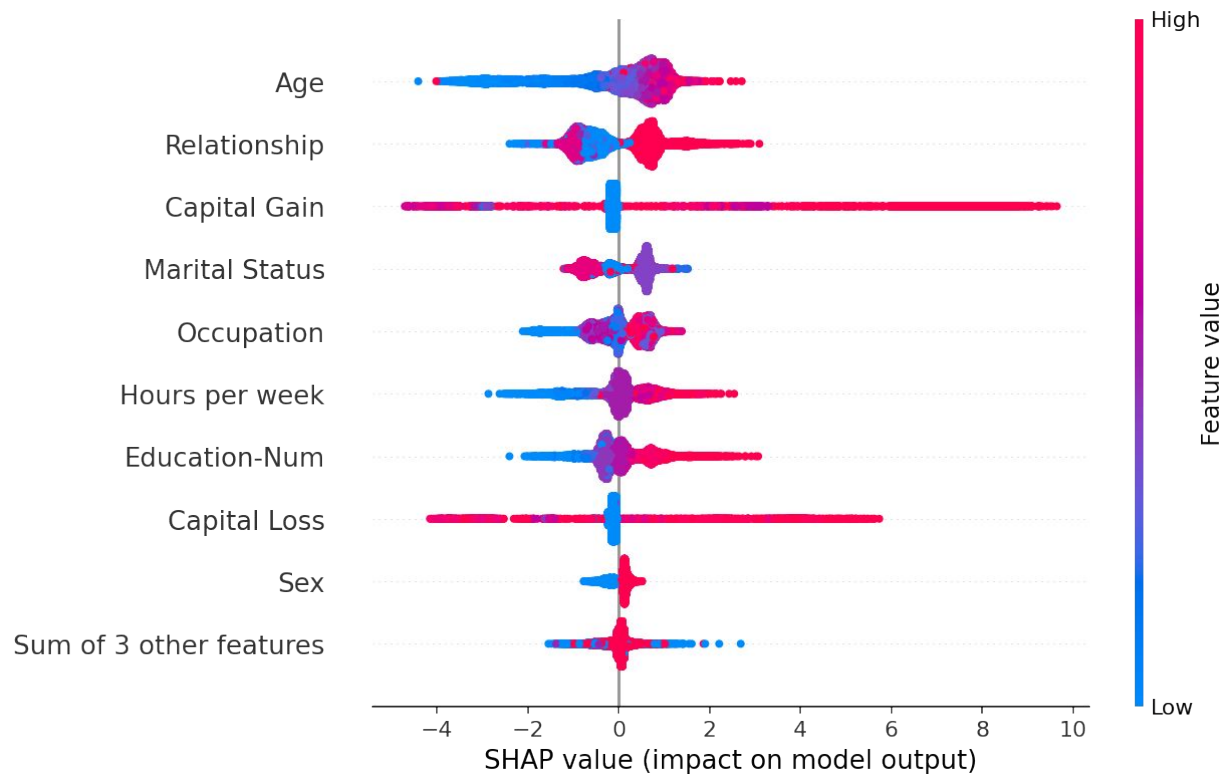
Theorem 1 follows from combined cooperative game theory results, where the values  $\phi_i$  are known as Shapley values [9]. Young (1985) demonstrated that Shapley values are the only set of values that satisfy three axioms similar to Property 1, Property 3, and a final property that we show to be redundant in this setting (see Supplementary Material). Property 2 is required to adapt the Shapley proofs to the class of additive feature attribution methods.

Under Properties 1-3, for a given simplified input mapping  $h_x$ , Theorem 1 shows that there is only one possible additive feature attribution method. This result implies that methods not based on Shapley values violate local accuracy and/or consistency (methods in Section 2 already respect missingness).

# Interpreting SHAP results

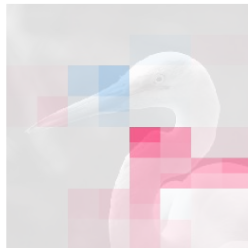


# Interpreting SHAP results

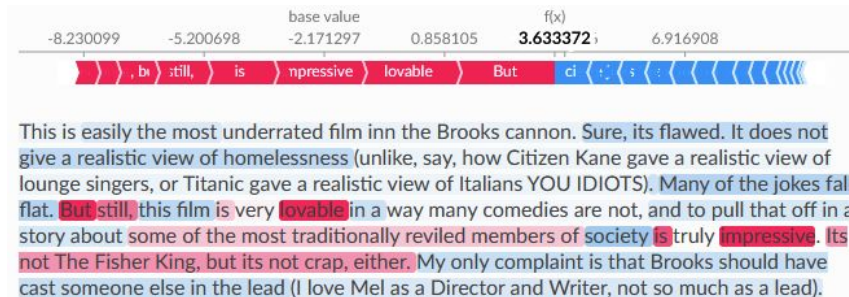
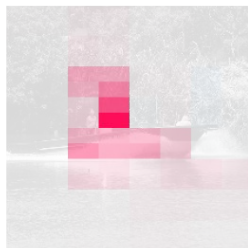


# Interpreting SHAP results

American\_egret



speedboat



# Exercise 2

Tabular data:

[https://github.com/simecek/dspracticum2023/blob/main/lesson07/ds\\_practicum\\_ex2\\_shap\\_tabular.ipynb](https://github.com/simecek/dspracticum2023/blob/main/lesson07/ds_practicum_ex2_shap_tabular.ipynb)

Text:

[https://github.com/simecek/dspracticum2023/blob/main/lesson07/ds\\_practicum\\_ex2\\_shap\\_text.ipynb](https://github.com/simecek/dspracticum2023/blob/main/lesson07/ds_practicum_ex2_shap_text.ipynb)

Images:

[https://github.com/simecek/dspracticum2023/blob/main/lesson07/ds\\_practicum\\_exercise2\\_shap\\_images.ipynb](https://github.com/simecek/dspracticum2023/blob/main/lesson07/ds_practicum_exercise2_shap_images.ipynb)

# Homework

- 1) Choose model and dataset and apply Feature Importance or SHAP on it.
- 2) Have a look at few examples and decide whether the model's decision are correct - it is not making it's prediction based on some artefact from the data.
- 3) Think about how 2) could be done automatically without manually looking at the specific cases (*can it be done for your use case?*).