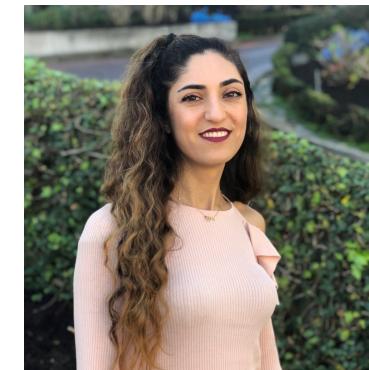
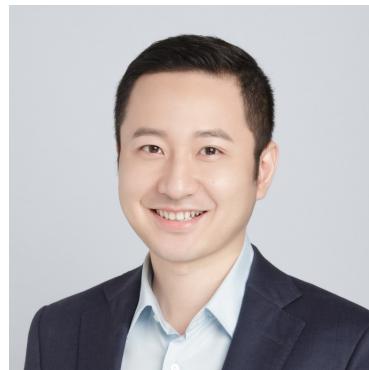


# Dividing & Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat



Shantanu Ghosh<sup>1</sup>, Ke Yu<sup>2</sup>, Forough Arabshahi<sup>3</sup>, Kayhan Batmanghelich<sup>1</sup>

<sup>1</sup>BU ECE, <sup>2</sup>Pitt ISP, <sup>3</sup>Meta



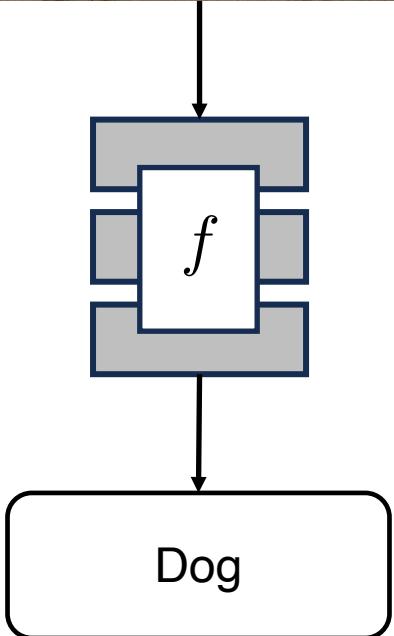
 Meta



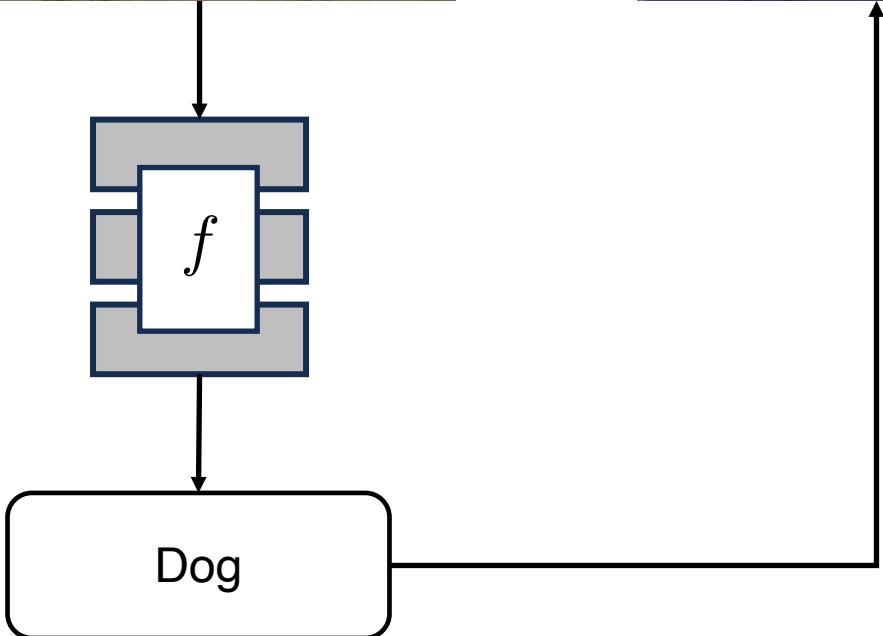
# Post-hoc Explanations



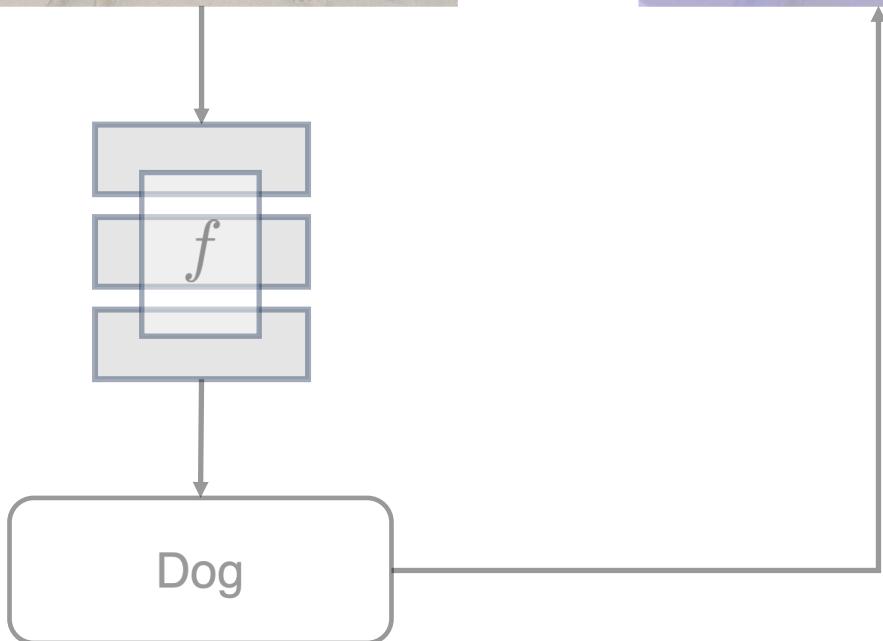
# Post-hoc Explanations



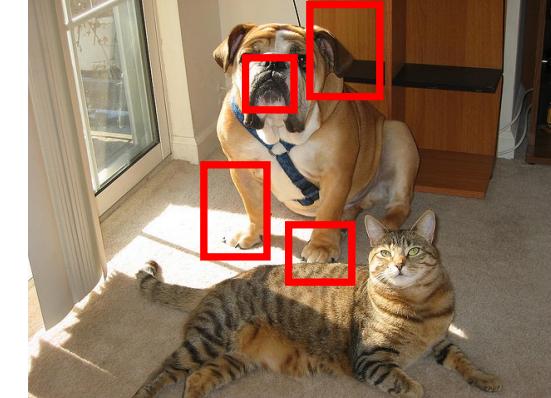
# Post-hoc Explanations



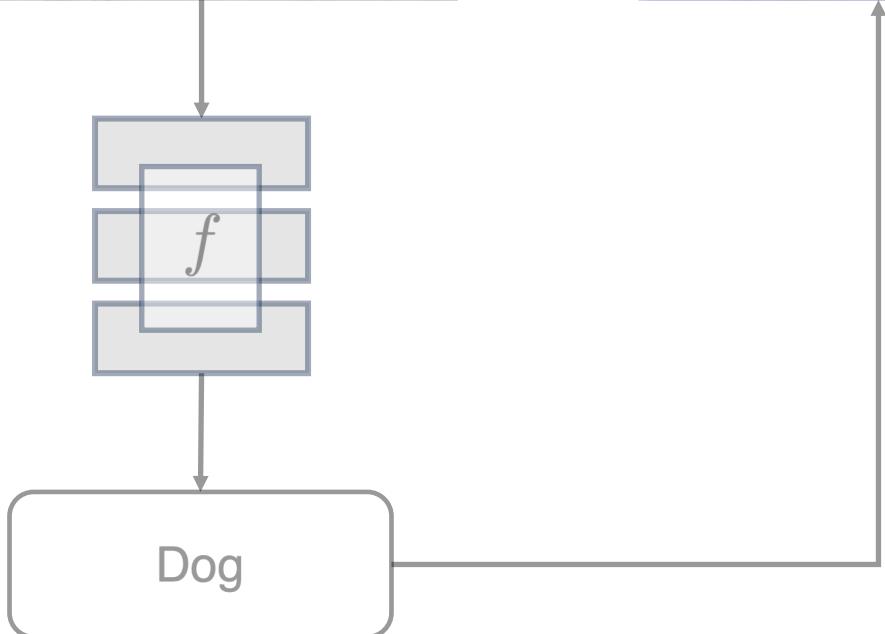
# Post-hoc Explanations Interpretable by design



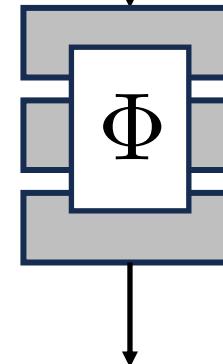
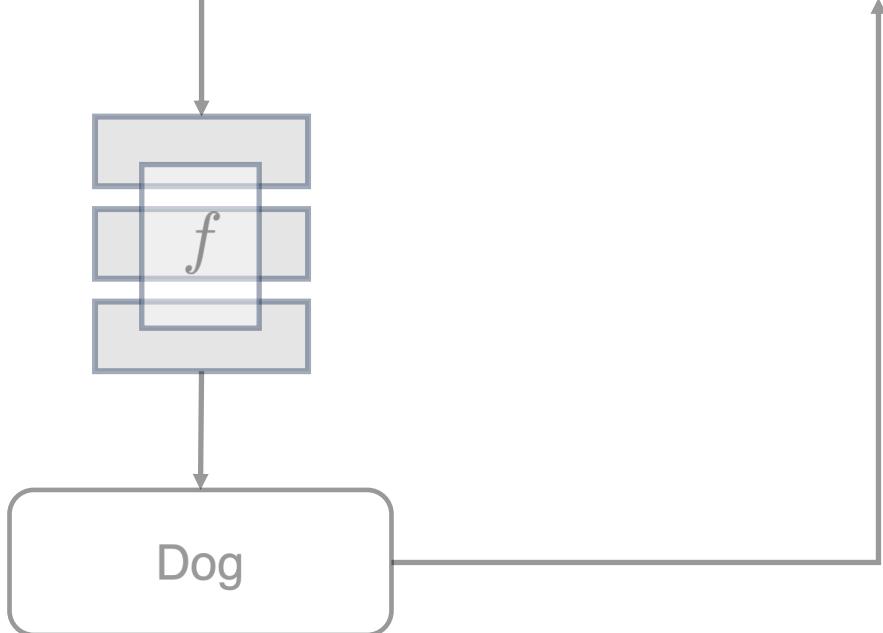
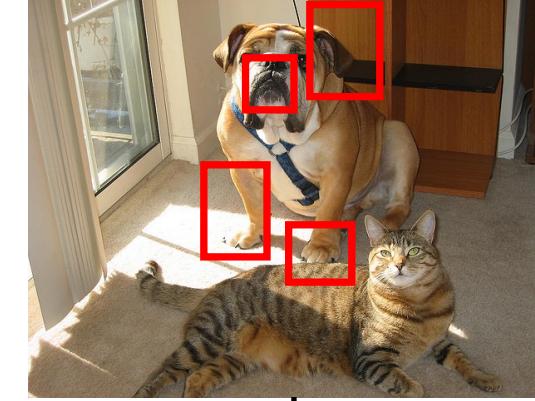
# Post-hoc Explanations Interpretable by design



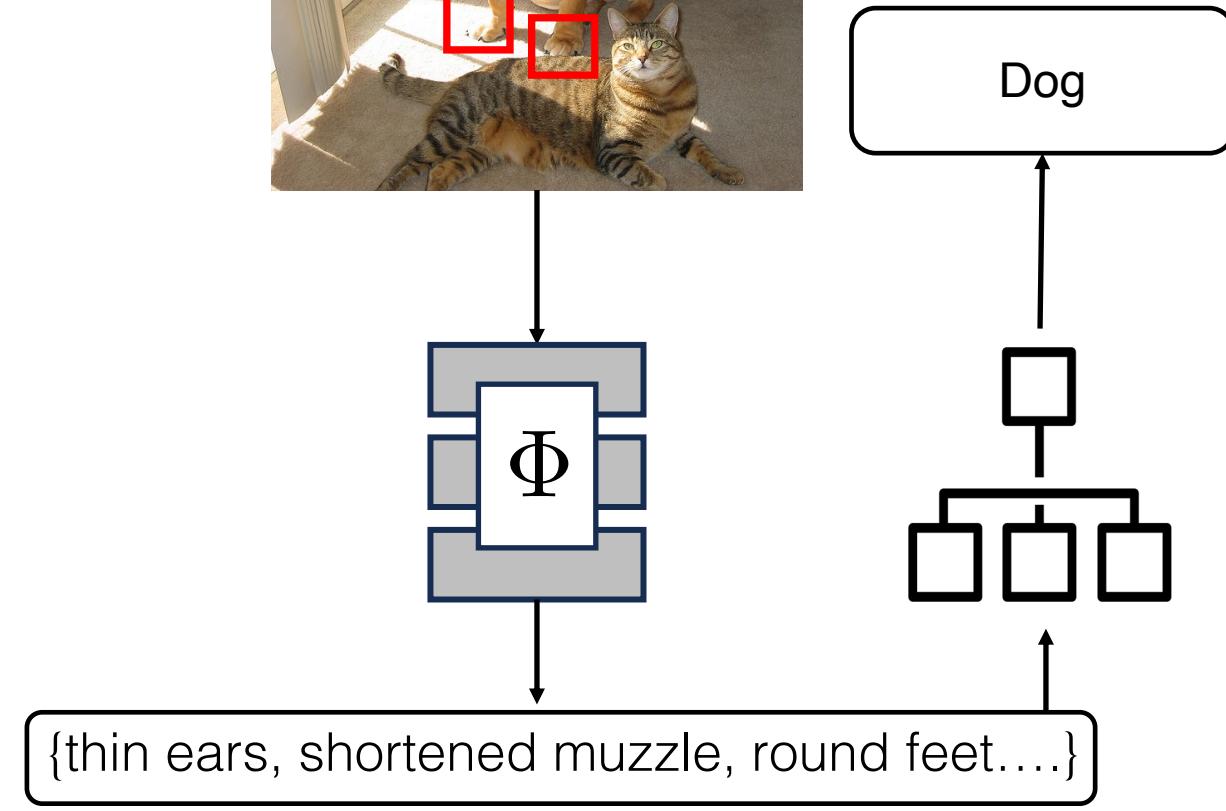
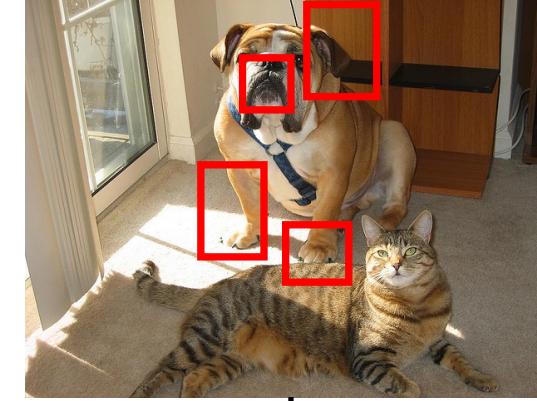
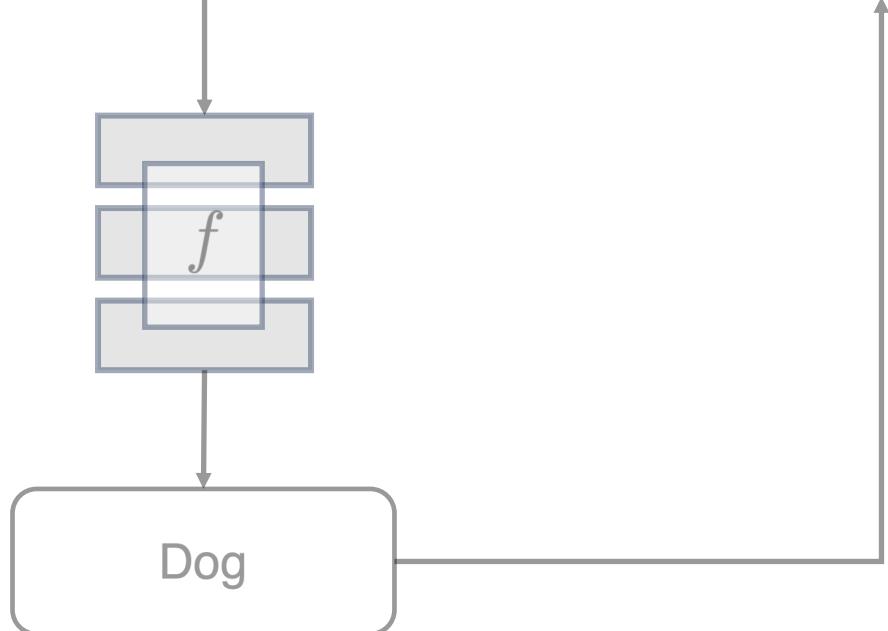
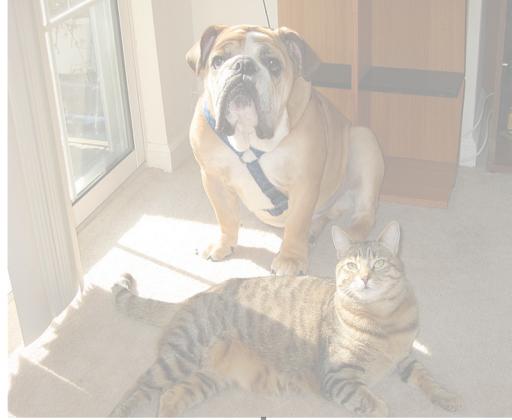
thin ears  
shortened muzzle  
round feet  
....



# Post-hoc Explanations Interpretable by design



# Post-hoc Explanations Interpretable by design



# Post-hoc explanations

# Post-hoc explanations

Pros:

- Does not alter the Black box

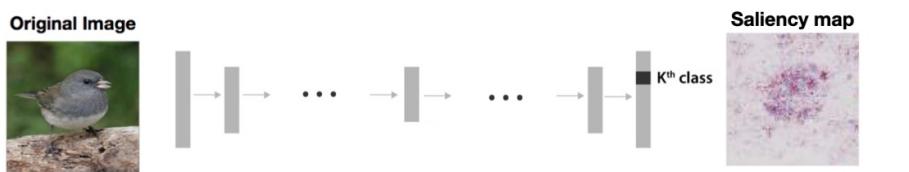
# Post-hoc explanations

## Pros:

- Does not alter the Black box

## Cons:

- Inconsistent explanations
- No recourse



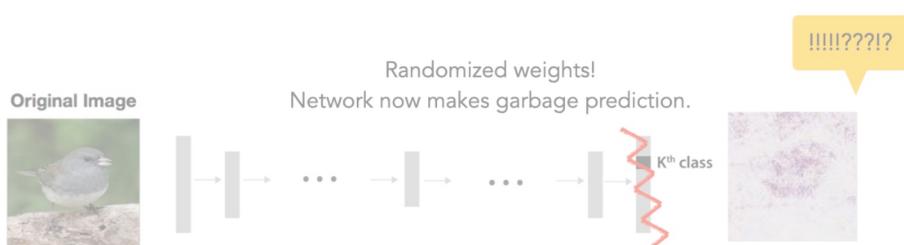
# Post-hoc explanations Interpretable by design

## Pros:

- Does not alter the Black box

## Cons:

- Inconsistent explanations
- No recourse



## Pros:

- Support concept intervention

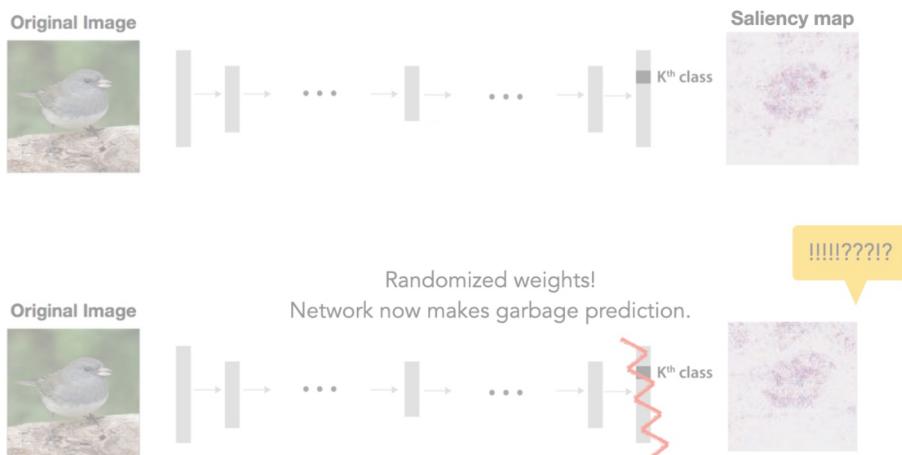
# Post-hoc explanations Interpretable by design

## Pros:

- Does not alter the Black box

## Cons:

- Inconsistent explanations
- No recourse

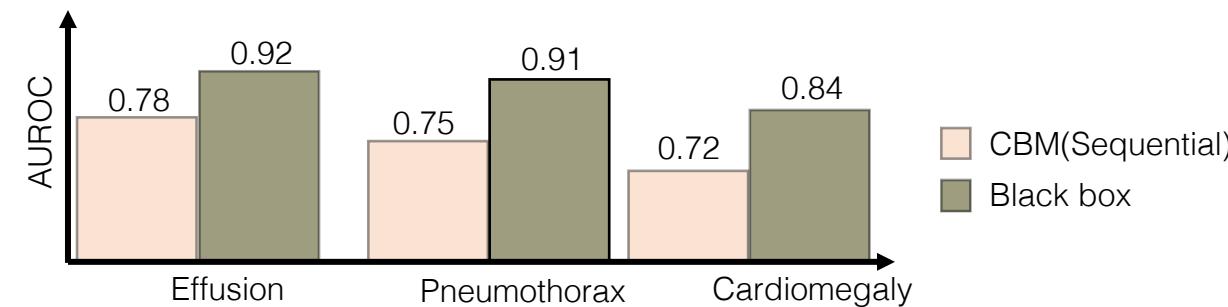


## Pros:

- Support concept intervention

## Cons:

- Harder to train
- Sub par performance



# Post-hoc explanations Interpretable by design

## Pros:

- Does not alter the Black box

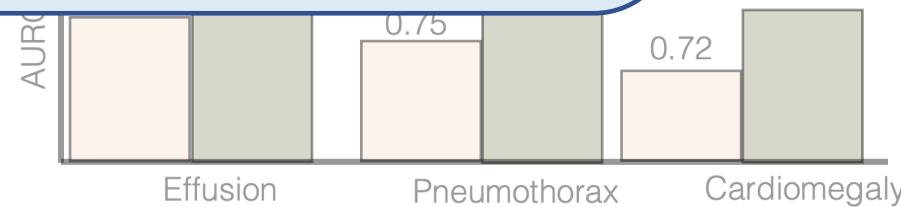
## Pros:

- Support concept intervention

## Cons:

- Inconsistent
- No recourse

Can we blur the line b/w  
post-hoc explanations  
or  
interpretable by design



□ CBM(Sequential)  
■ Black box

# Desirable properties

1. Does not compromise  
the performance

# Desirable properties

1. Does not compromise the performance

# Design choices

1. Iteratively carve out the interpretable models from Black box

# Desirable properties

1. Does not compromise the performance
2. Can be intervened to fix misclassification

# Design choices

1. Iteratively carve out the interpretable models from Black box

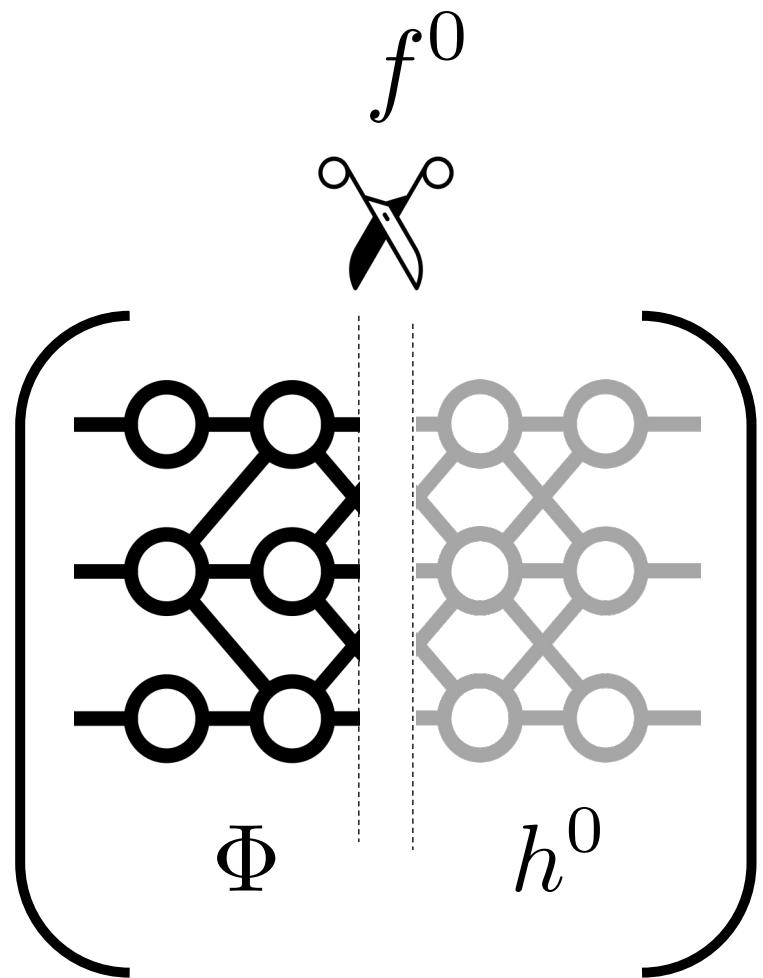
# Desirable properties

1. Does not compromise the performance
2. Can be intervened to fix misclassification

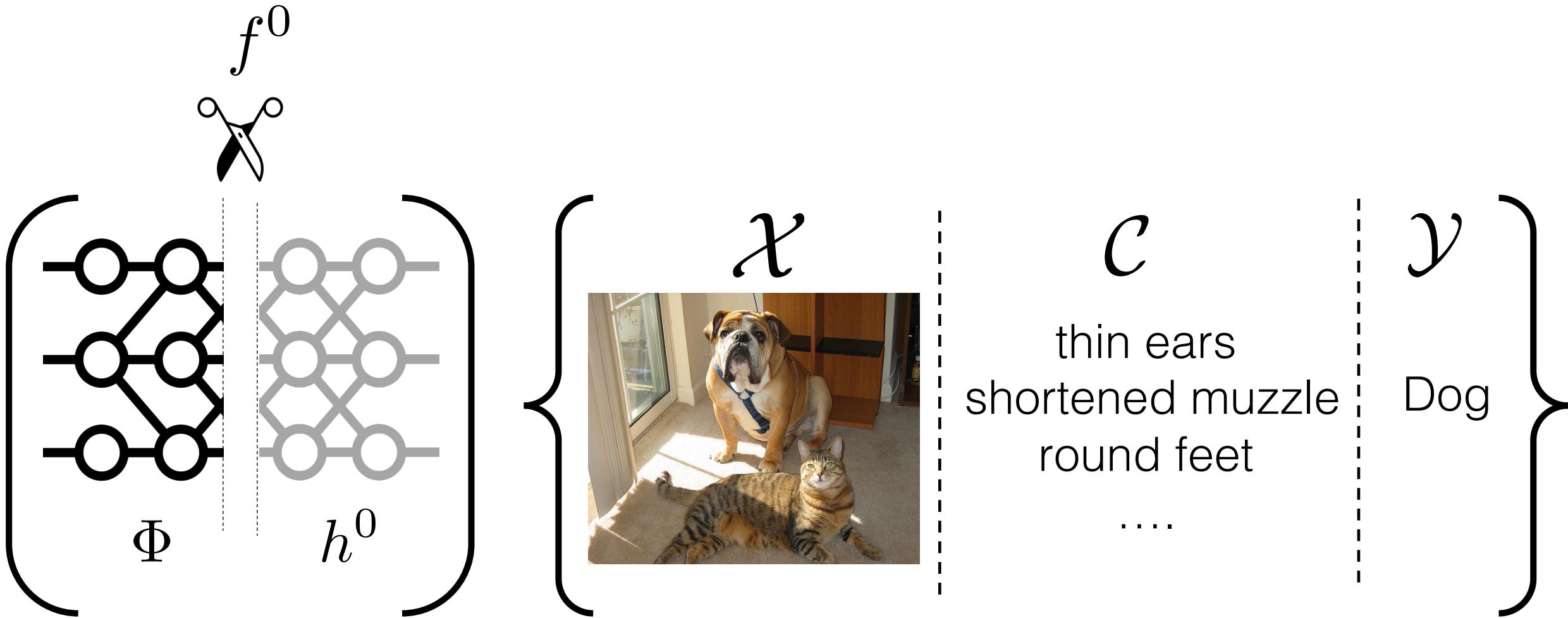
# Design choices

1. Iteratively carve out the interpretable models from Black box
2. Concept based, not pixel based
3. First order logic (FOL) for concept interaction

# Assumptions

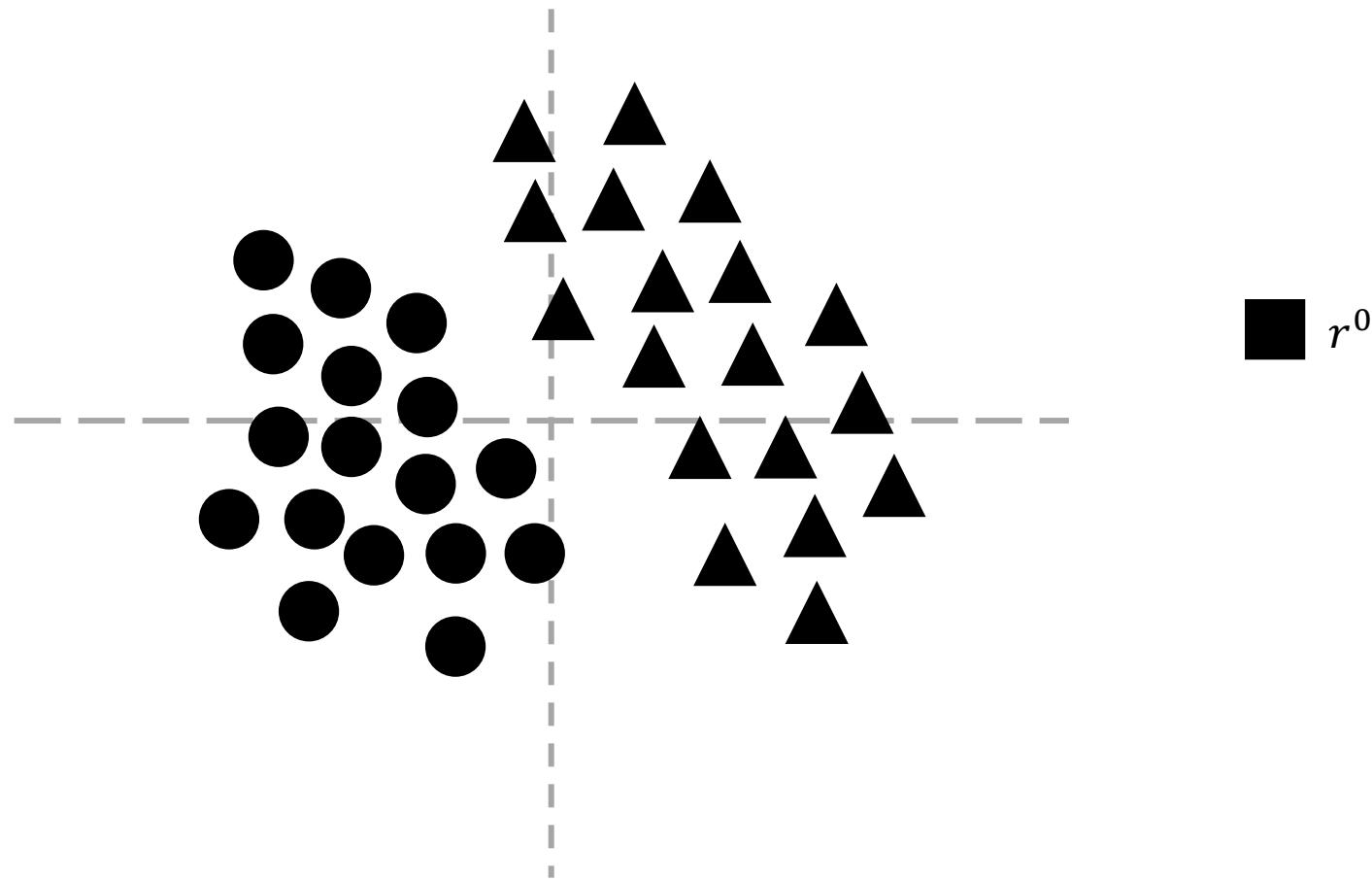


# Assumptions

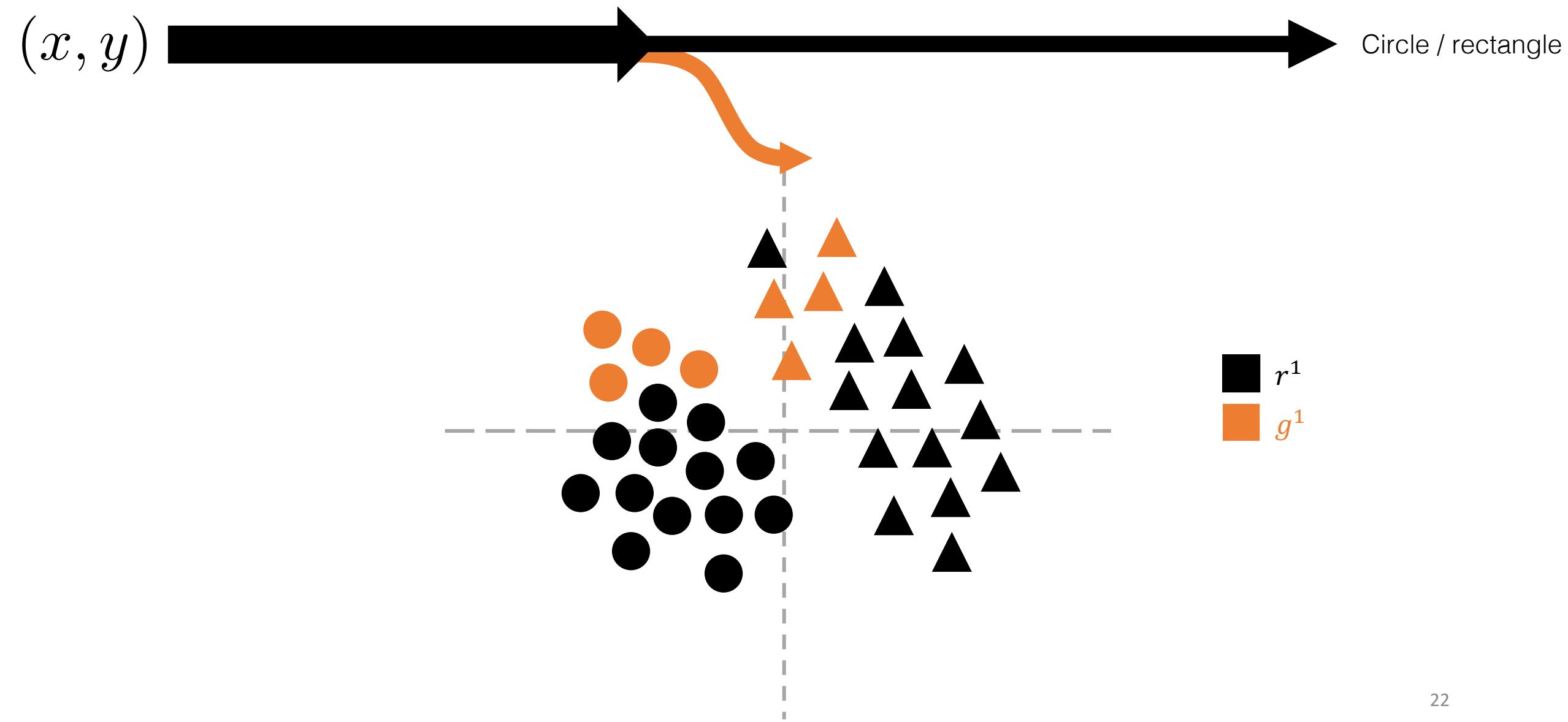


# Iteratively carve out interpretable models

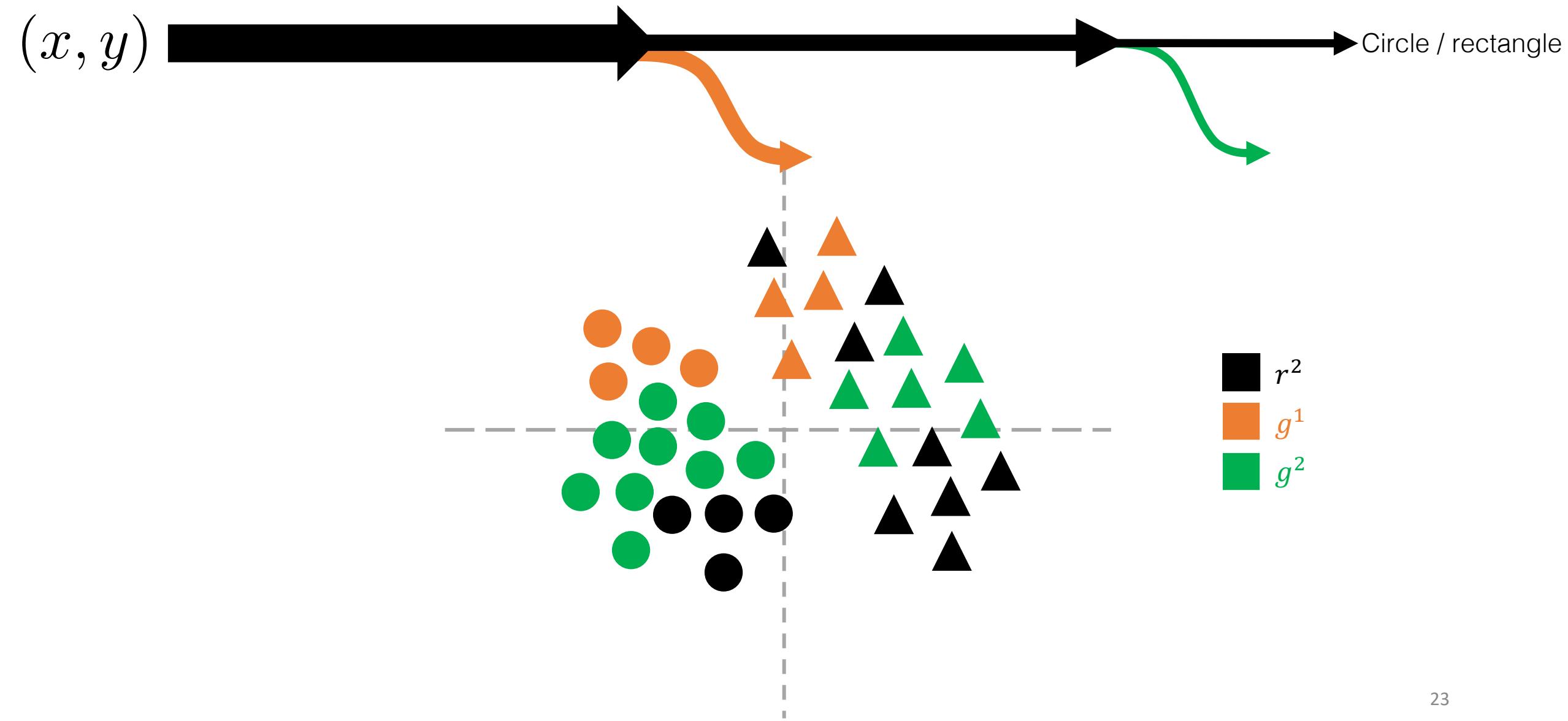
$(x, y)$   Circle / rectangle



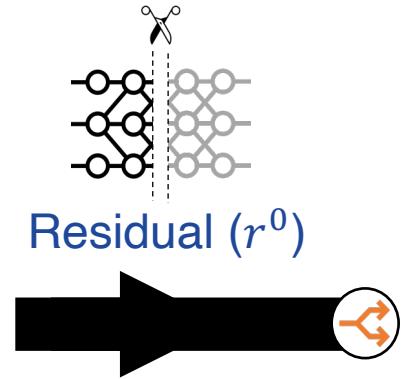
# Iteratively carve out interpretable models



# Iteratively carve out interpretable models



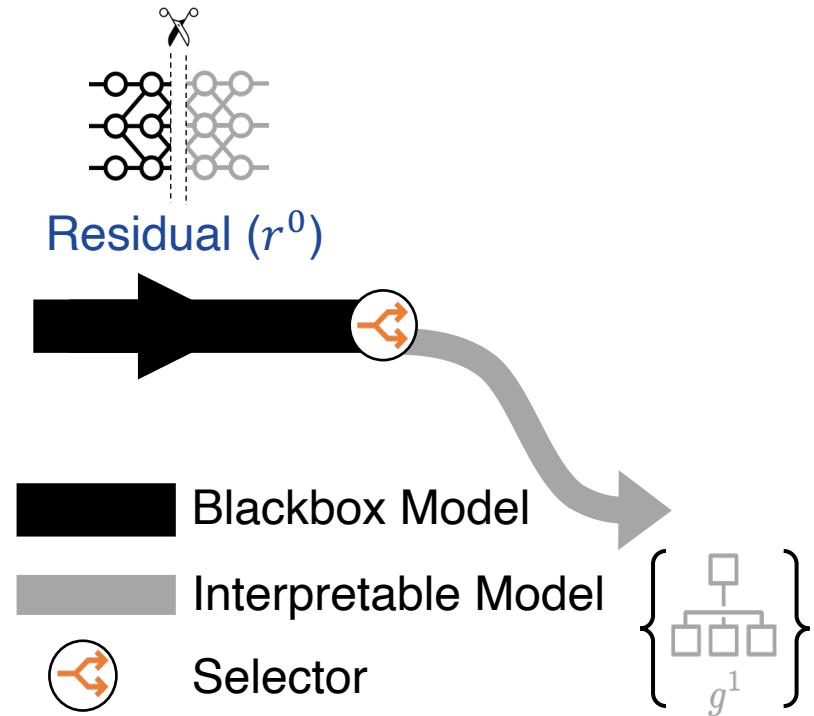
# Iteratively carve out interpretable models



- Blackbox Model
- Interpretable Model
- Selector

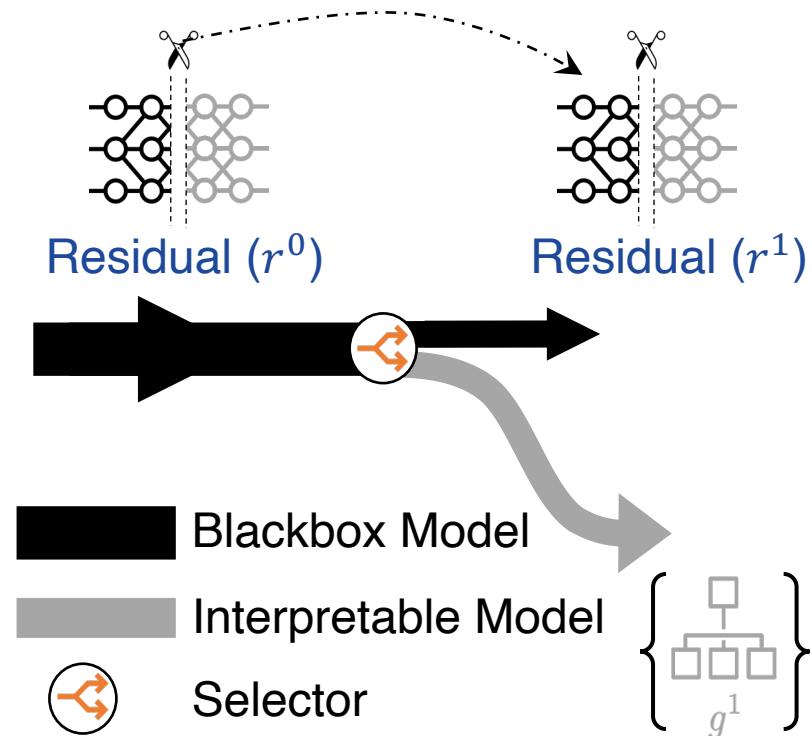
Each  $g$  to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



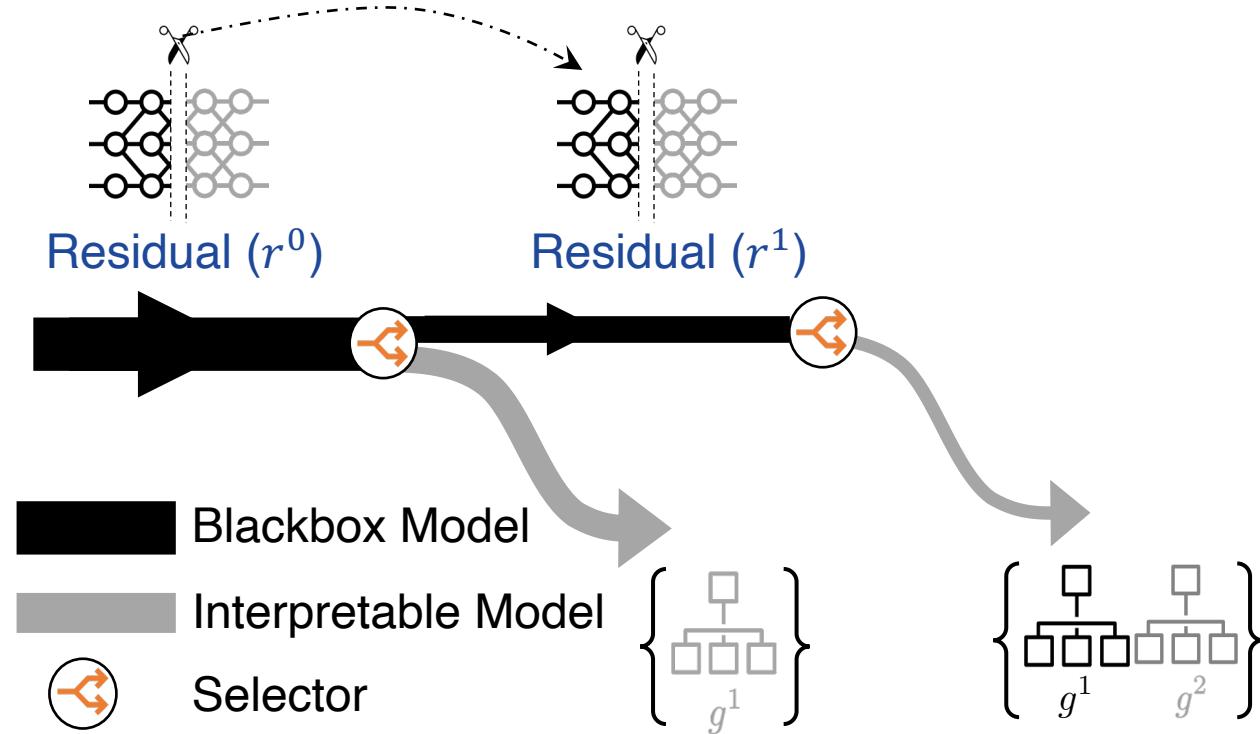
Each  $g$  to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



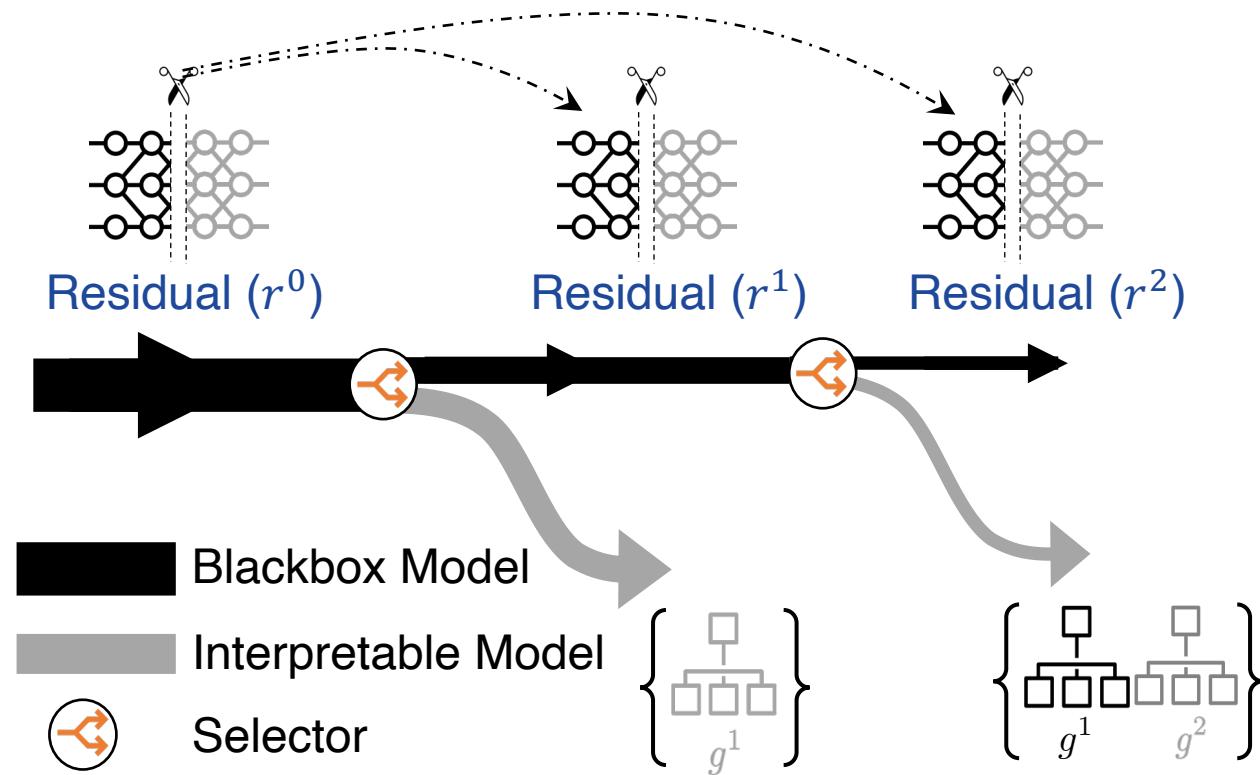
Each  $g$  to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



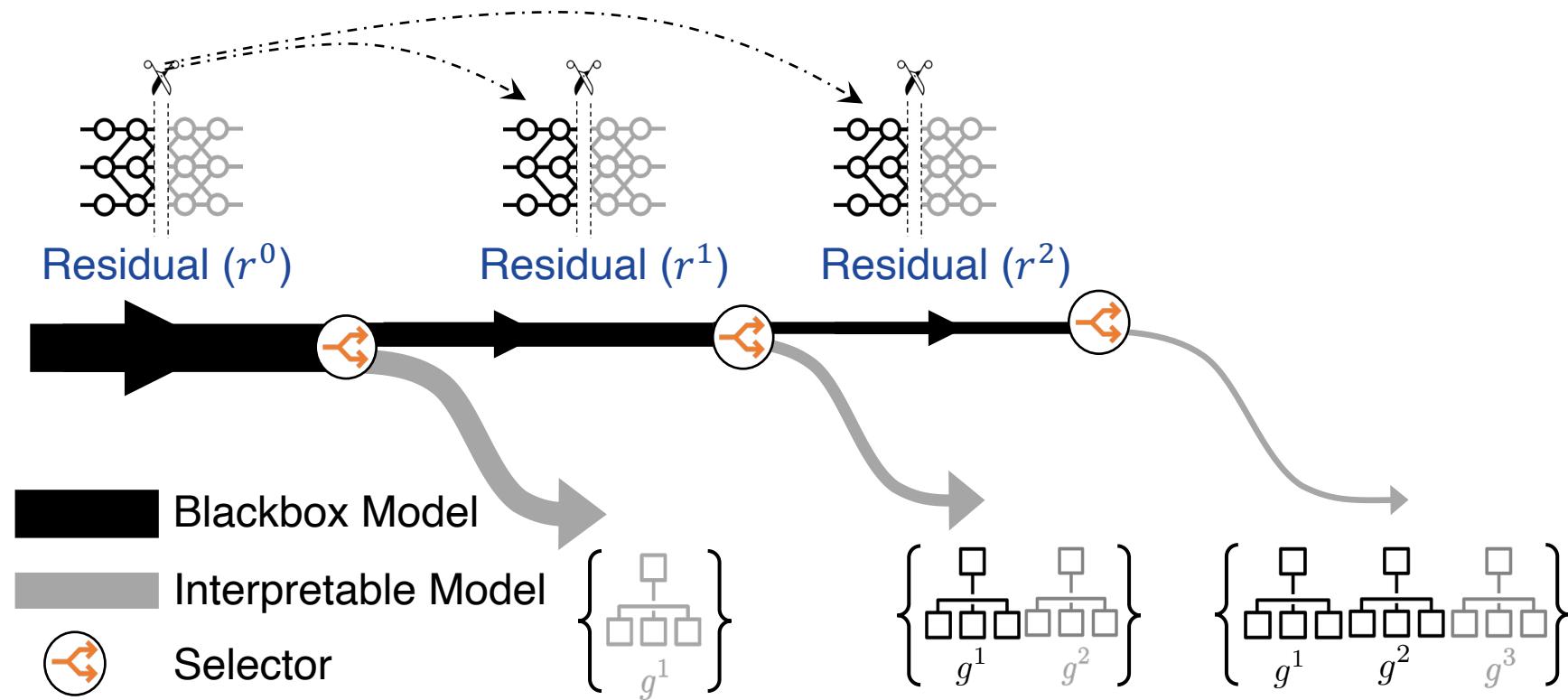
Each  $g$  to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



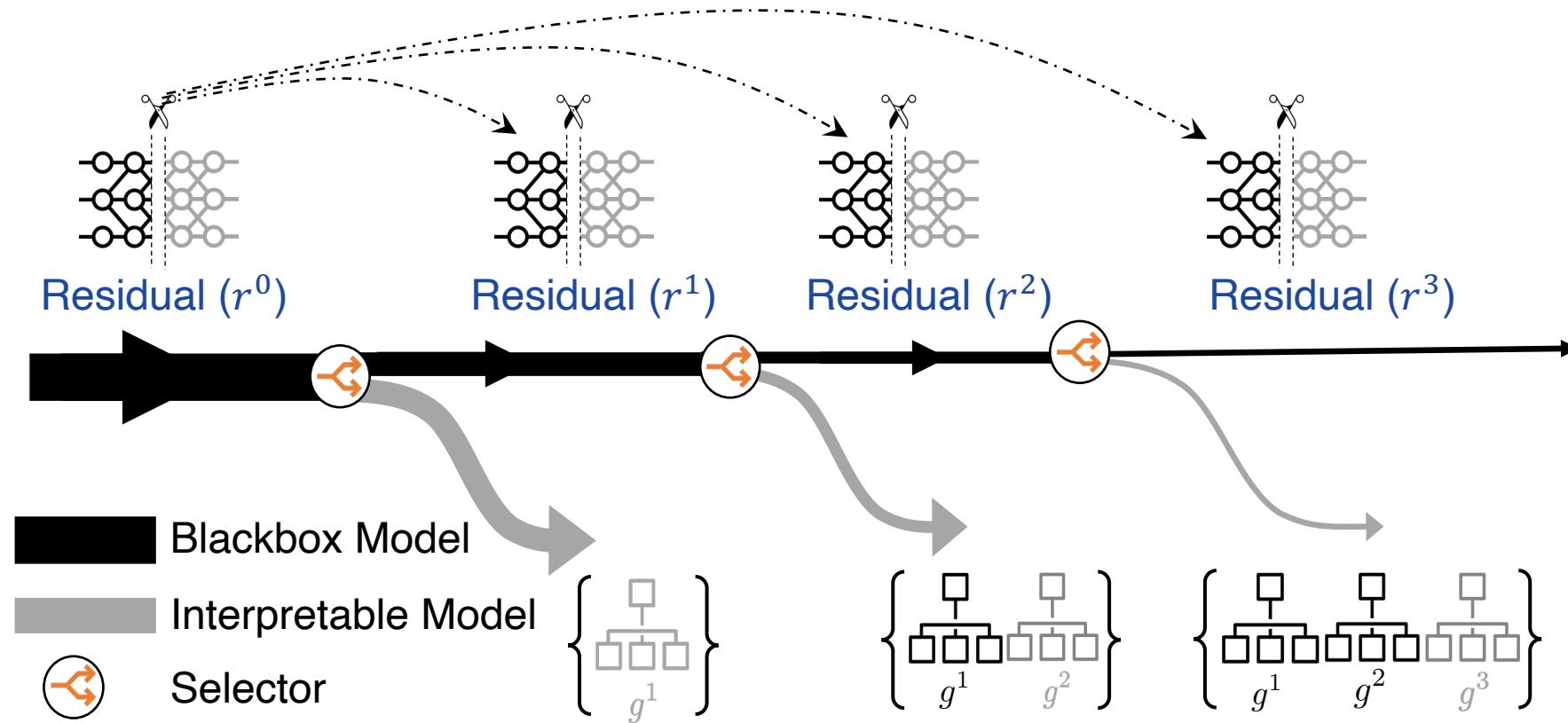
Each  $g$  to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



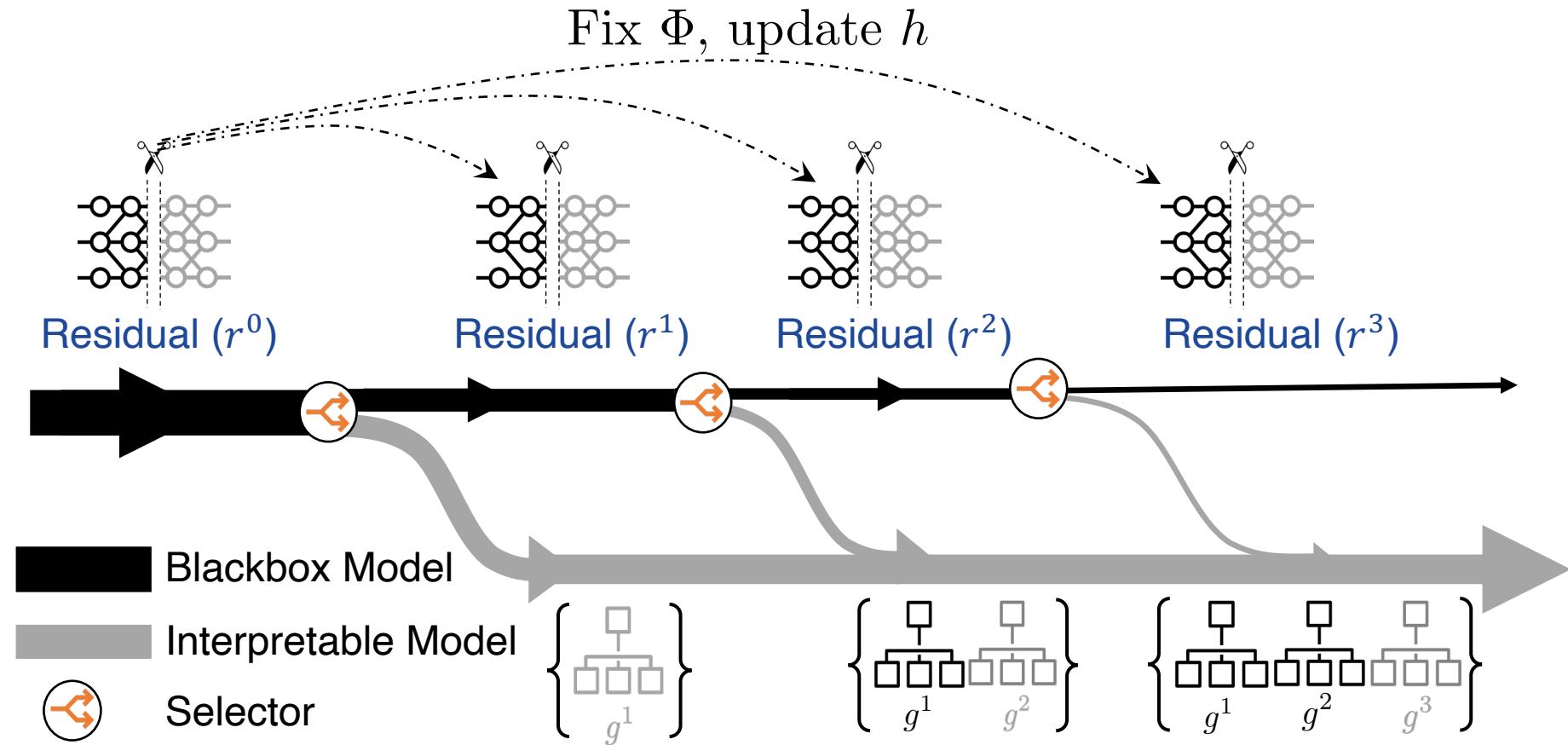
Each  $g$  to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



Each  $g$  to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



Each  $g$  to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Provide the concept-interaction

# Provide the concept-interaction

Expert 1



Olive sided Flycatcher  $\leftrightarrow$ breast\_color\_grey $\wedge$   
tail\_pattern\_solid

These experts are derived from CUB-200 dataset with ViT as Black box

# Provide the concept-interaction

Expert 1



Olive sided Flycatcher  $\leftrightarrow$ breast\_color\_grey $\wedge$   
tail\_pattern\_solid

Expert 2



Olive sided Flycatcher  $\leftrightarrow$ underparts\_color\_grey $\wedge$   
wing\_color\_grey

# Provide the concept-interaction

Expert 1



Olive sided Flycatcher  $\leftrightarrow$  breast\_color\_grey  $\wedge$   
tail\_pattern.Solid

Expert 2



Olive sided Flycatcher  $\leftrightarrow$  underparts\_color\_grey  $\wedge$   
wing\_color\_grey

Final Residual (Unexplained)

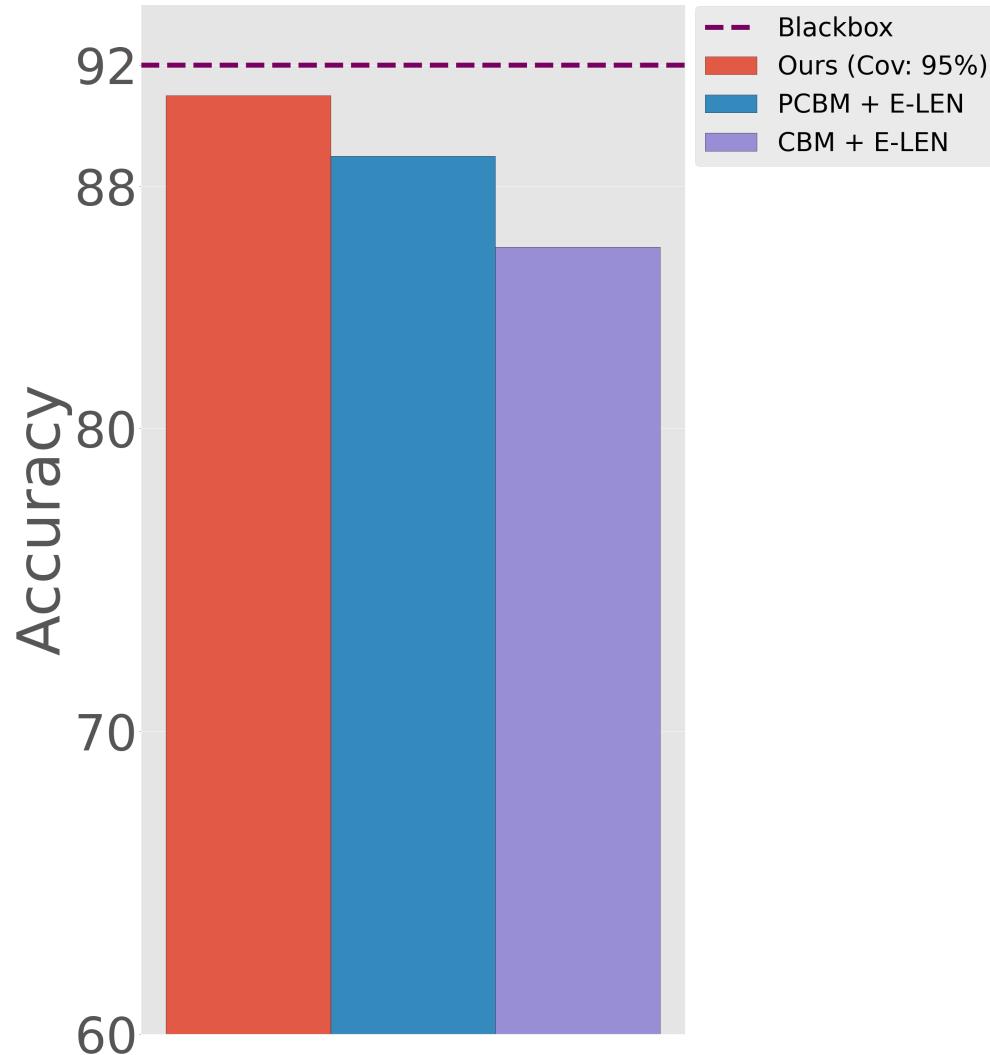


These experts are derived from CUB-200 dataset with ViT as Black box

Does not compromise performance

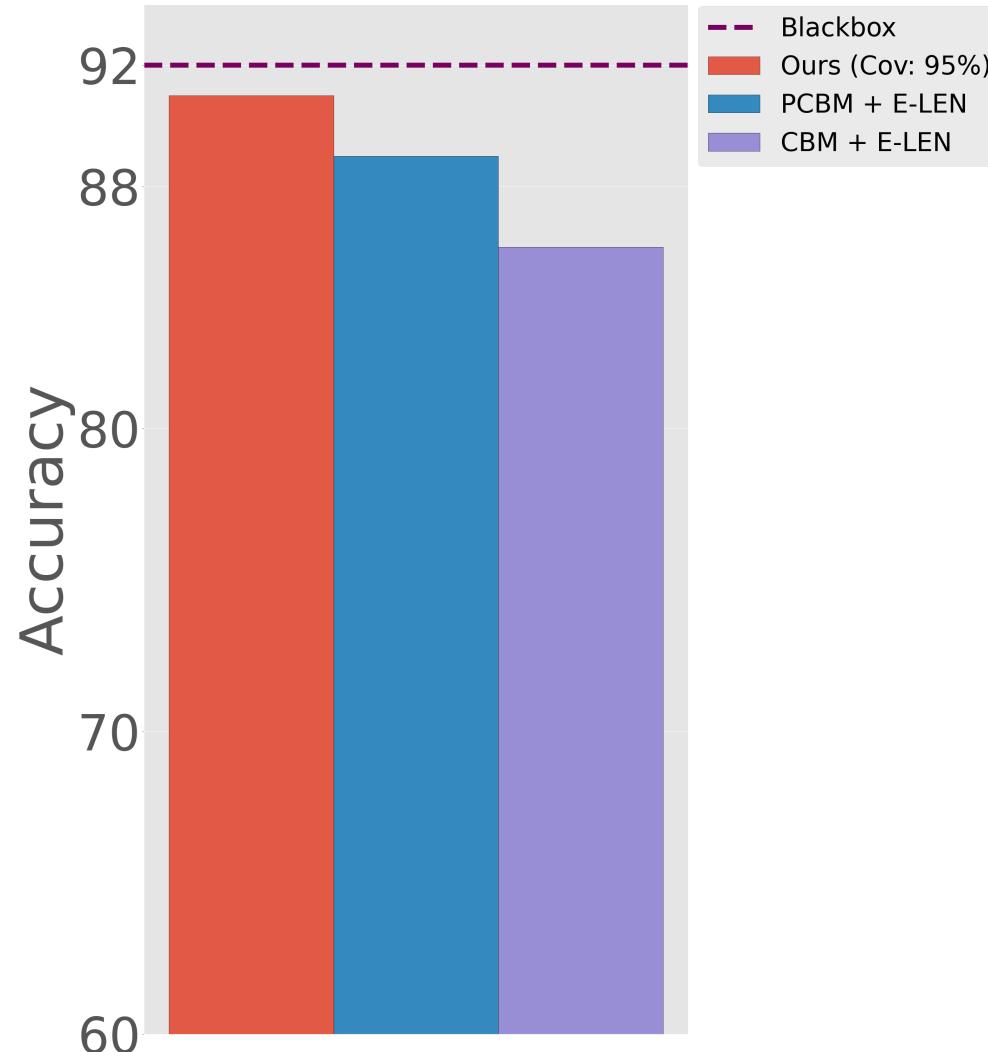
# Does not compromise performance

## CUB-200 with ViT

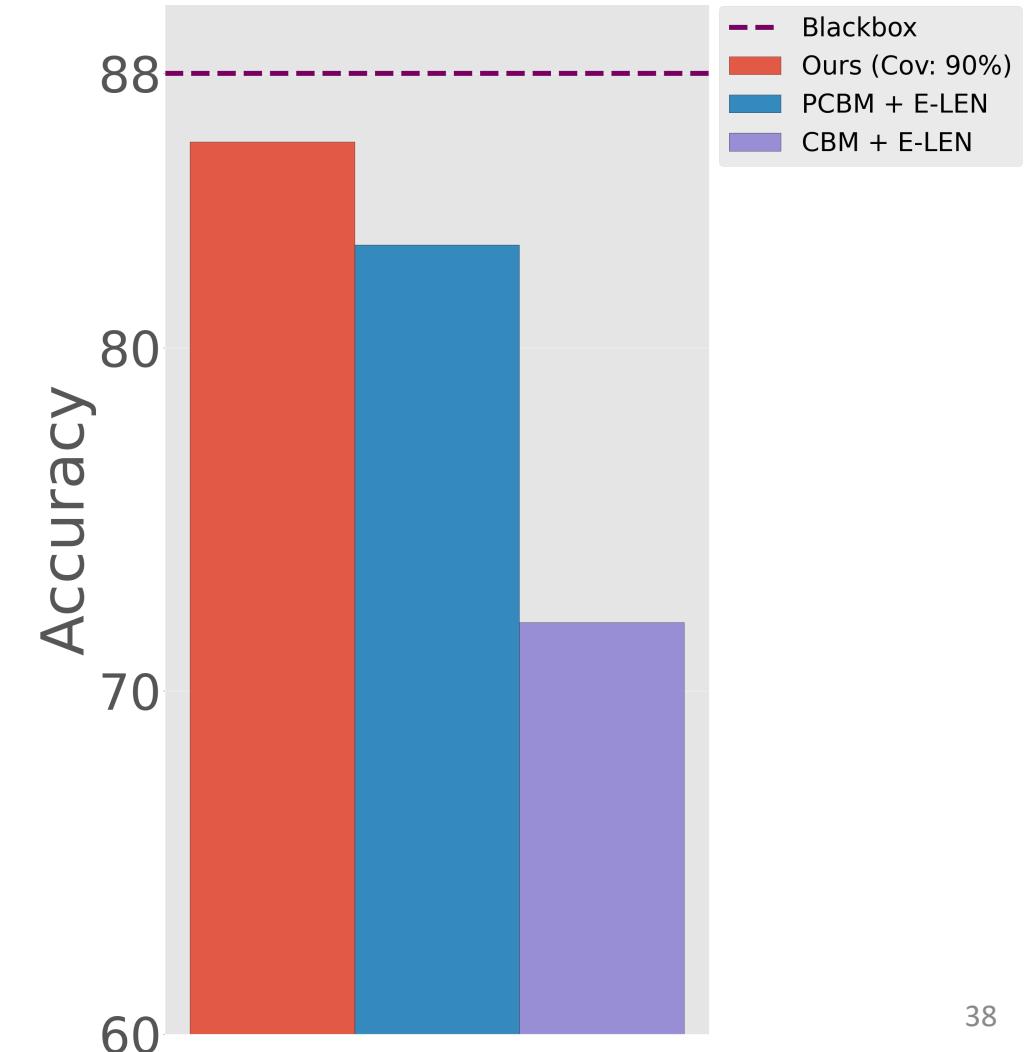


# Does not compromise performance

## CUB-200 with ViT

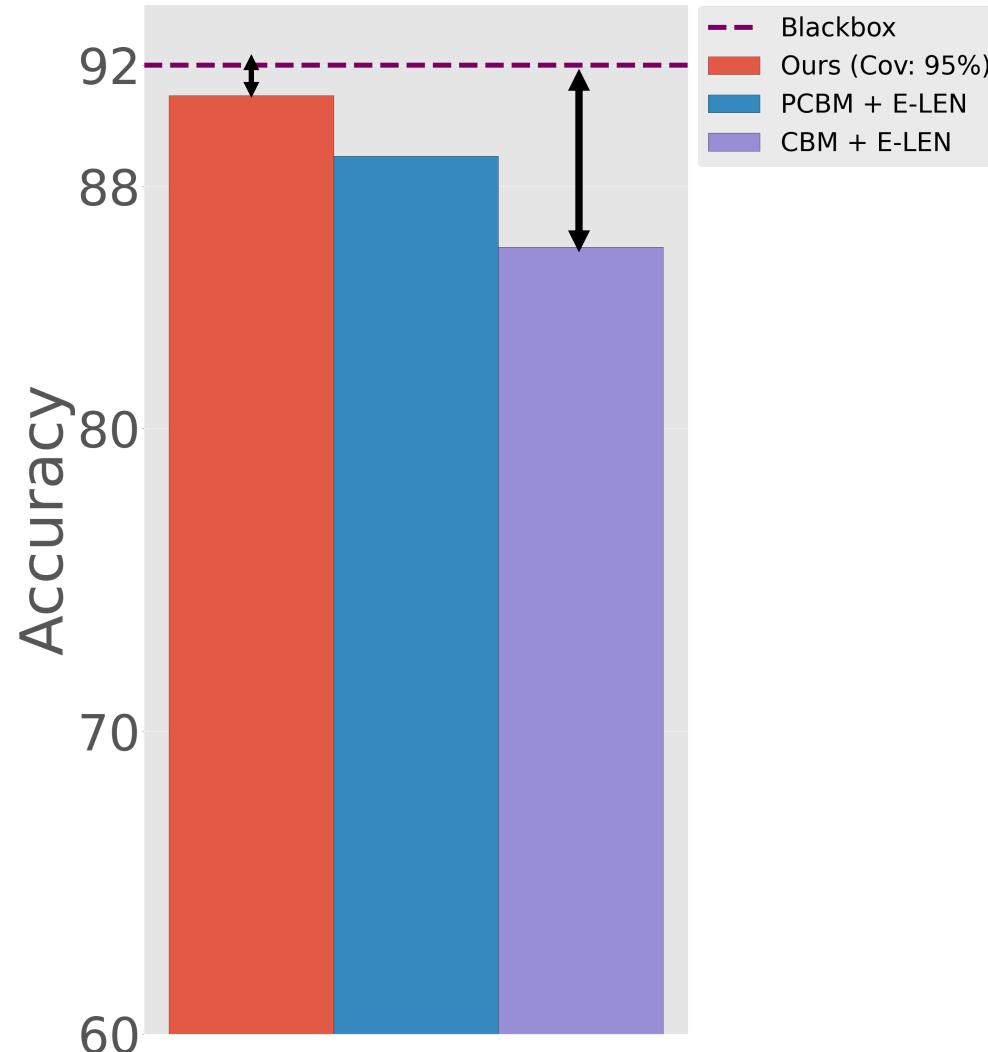


## CUB-200 with ResNet101

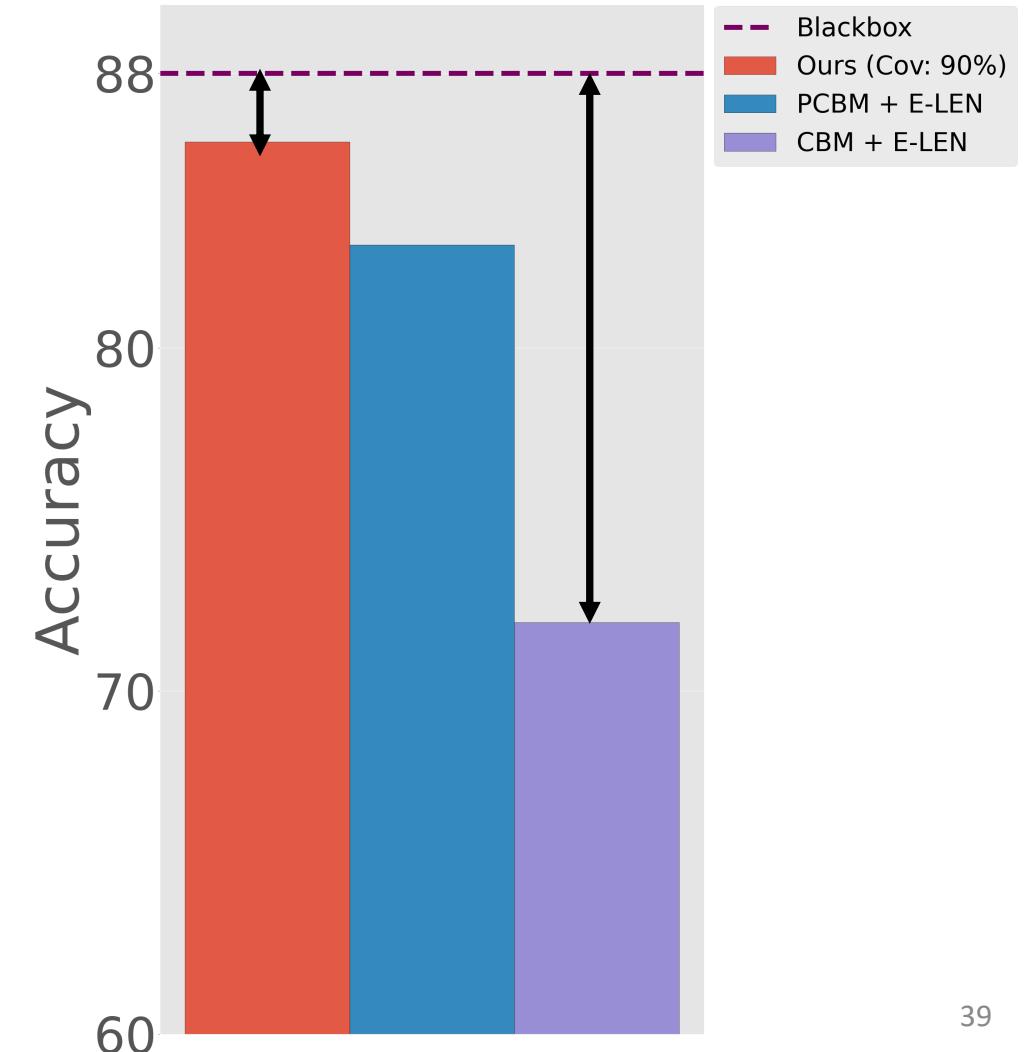


# Does not compromise performance

## CUB-200 with ViT



## CUB-200 with ResNet101



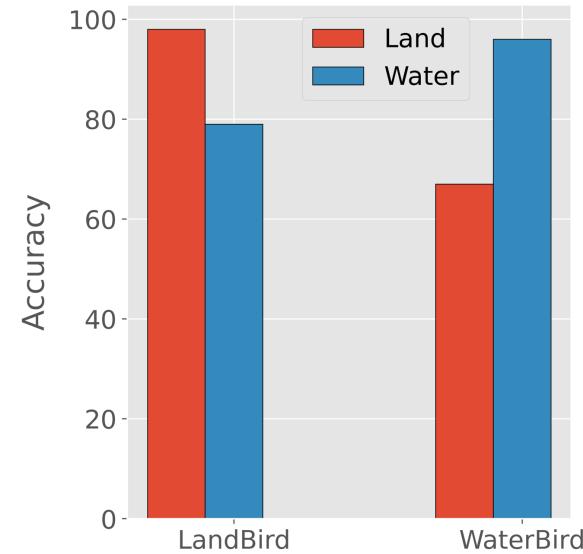
# Application on shortcut



# Application on shortcut



**Accuracy scores the  
biased Blackbox**



# Application on shortcut



Accuracy scores the  
biased Blackbox



## Biased Blackbox

Groundtruth: WaterBird

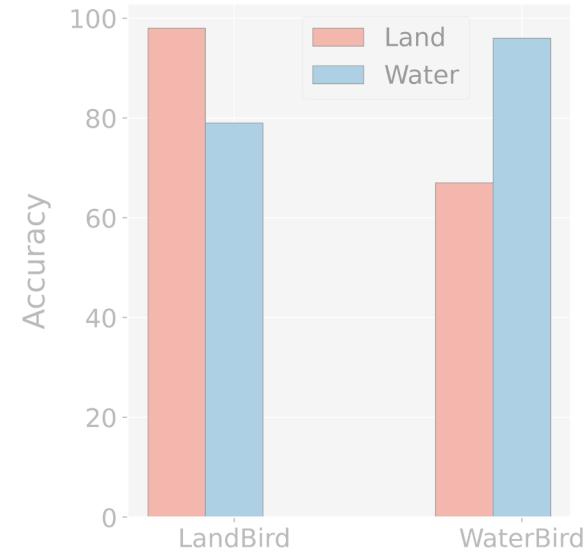
Prediction : LandBird

Explanation : LandBird  $\leftrightarrow$  WingShapeRoundedwings  $\wedge$  Forest

# Application on shortcut



Accuracy scores the  
biased Blackbox



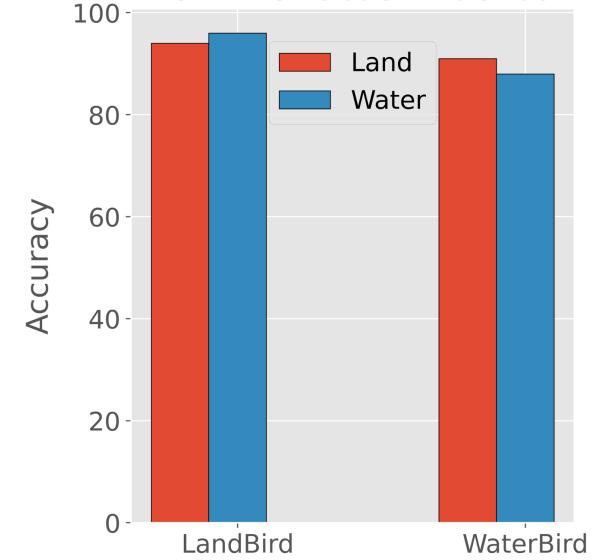
Biased Blackbox

Groundtruth: WaterBird

Prediction : LandBird

Explanation : LandBird  $\leftrightarrow$  WingShapeRoundedwings  $\wedge$  Forest

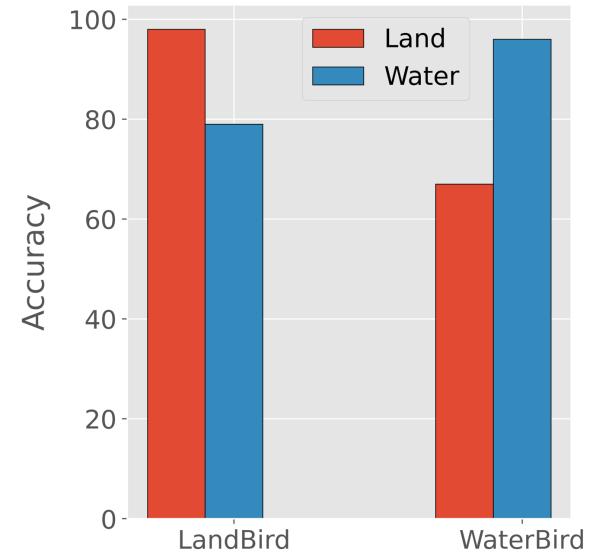
Accuracy scores MoIE  
from the robust Blackbox



# Application on shortcut



Accuracy scores the  
biased Blackbox



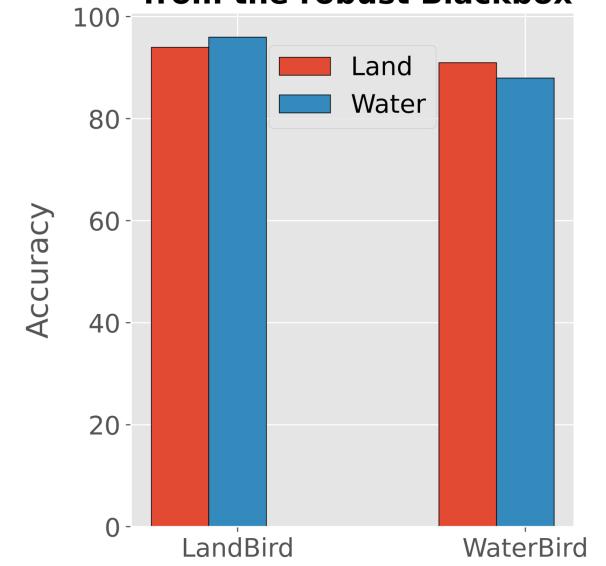
Biased Blackbox

Groundtruth: WaterBird

Prediction : LandBird

Explanation : LandBird  $\leftrightarrow$  WingShapeRoundedwings  $\wedge$  Forest

Accuracy scores MoIE  
from the robust Blackbox



# Application on shortcut



## Robust Blackbox

Groundtruth: WaterBird

Prediction : WaterBird

Explanation : WaterBird  $\leftrightarrow$  BillLengthAboutTheSameAsHead  
 $\wedge \neg$ BillLengthShorterThanHead  $\wedge \neg$ SizeSmall5\_9in  
 $\wedge \neg$ ShapePerchingLike  $\wedge$  CrownColorWhite

Accuracy scores the biased Blackbox



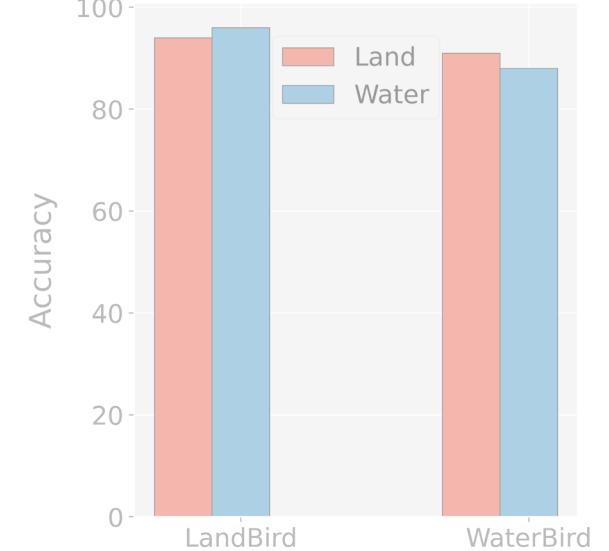
## Biased Blackbox

Groundtruth: WaterBird

Prediction : LandBird

Explanation : LandBird  $\leftrightarrow$  WingShapeRoundedwings  $\wedge$  Forest

Accuracy scores MoIE from the robust Blackbox



# Application on shortcut



Robust Blackbox

Groundtruth: WaterBird

Prediction : WaterBird

Explanation : WaterBird  $\leftrightarrow$  BillLengthAboutTheSameAsHead  
 $\wedge \neg$ BillLengthShorterThanHead  $\wedge \neg$ SizeSmall5\_9in  
 $\wedge \neg$ ShapePerchingLike  $\wedge$  CrownColorWhite

Accuracy scores the biased Blackbox



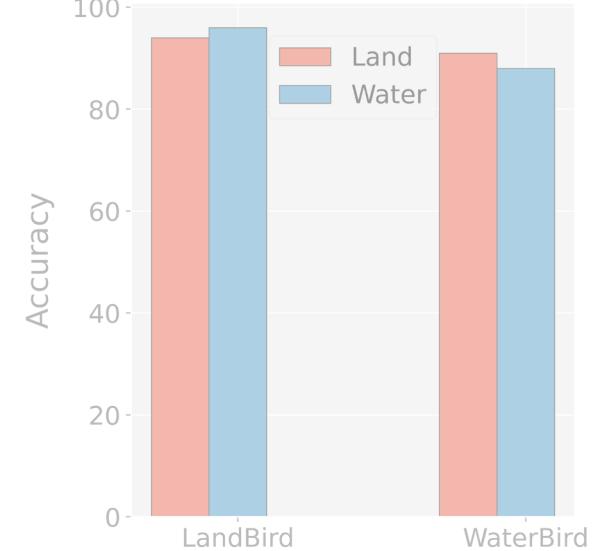
Biased Blackbox

Groundtruth: WaterBird

Prediction : LandBird

Explanation : LandBird  $\leftrightarrow$  WingShapeRoundedwings  $\wedge$  Forest

Accuracy scores MoIE from the robust Blackbox



# We have more results

- We achieve higher completeness scores
- We performs better in test-time interventions
- We update to MolE-CXR to perform efficient transfer learning for CXRs (In [IMLH](#) workshop at [ICML 2023](#) and [MICCAI 2023](#))

Project website



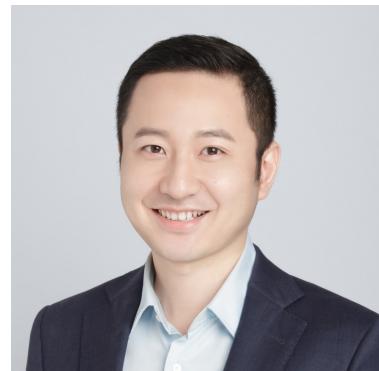
Poster session Exhibit Hall 1 #729

Thu 27 Jul 4:30 p.m. EDT — 6 p.m. EDT

# Project website



## Thank you



Shantanu Ghosh, Ke Yu, Forough Arabshahi, Kayhan Batmanghelich

Poster session Exhibit Hall 1 #729  
Thu 27 Jul 10.30 a.m. HST — noon HST