BOSTON UNIVERSITY | BATMAN LAB

Poster session Exhibit Hall 1 #729
Thu 27 Jul 10:30 a.m. HST — noon HST

ICML International Conference On Machine Learning

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



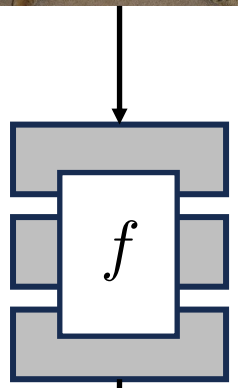

Shantanu Ghosh[1], Ke Yu[2], Forough Arabshahi[3], Kayhan Batmanghelich[1]

[1]BU ECE, [2]Pitt ISP, [3]Meta

Project

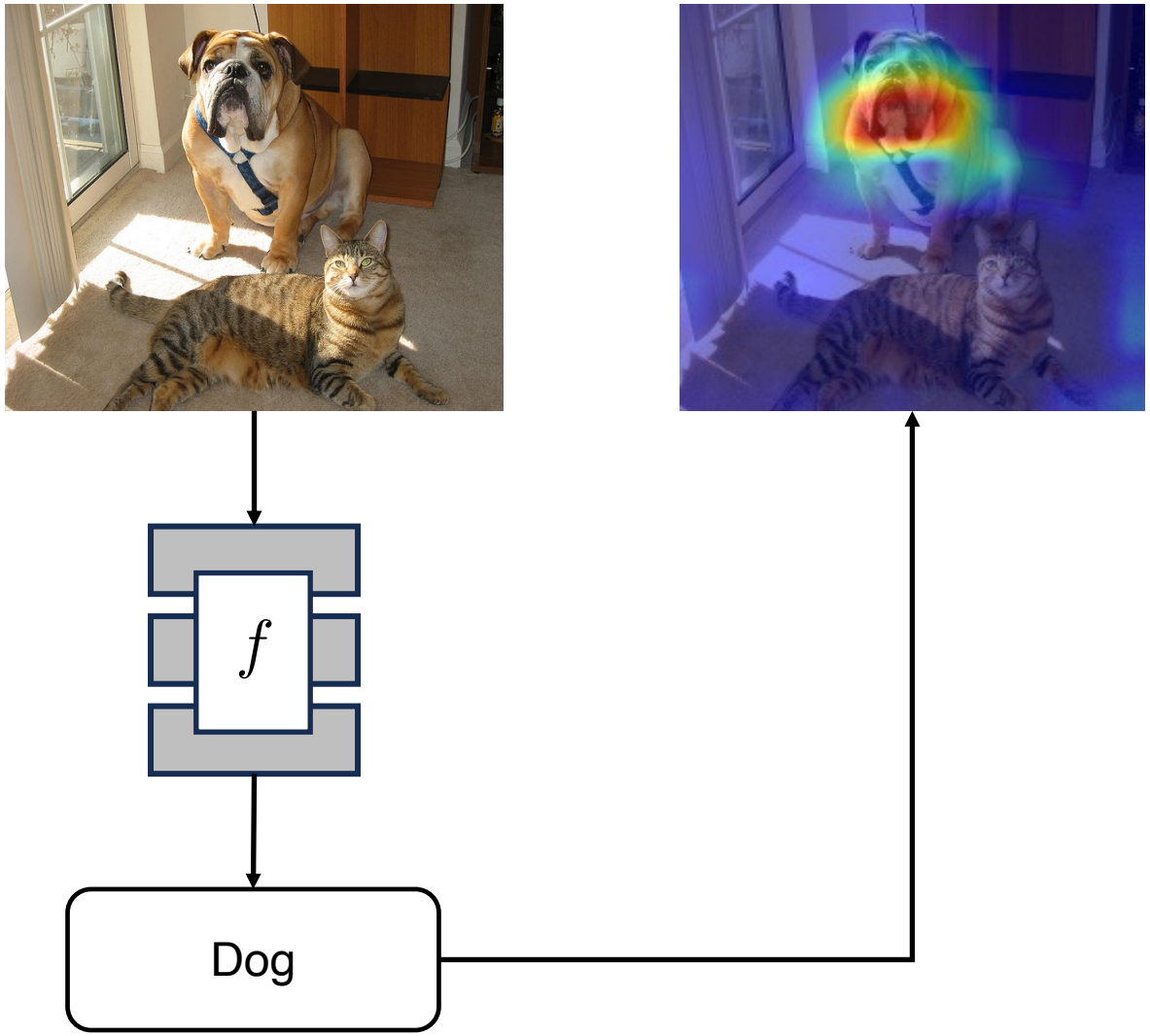
# Post-hoc Explanations

# Post-hoc Explanations



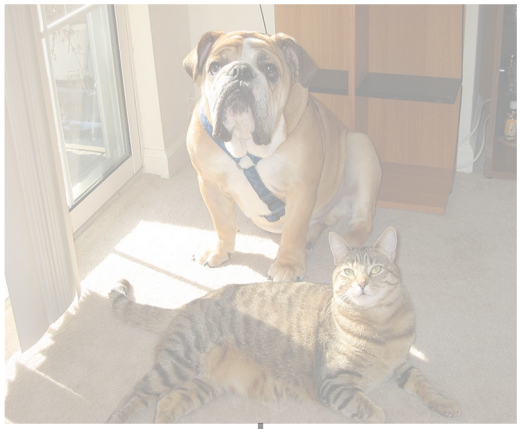

Dog

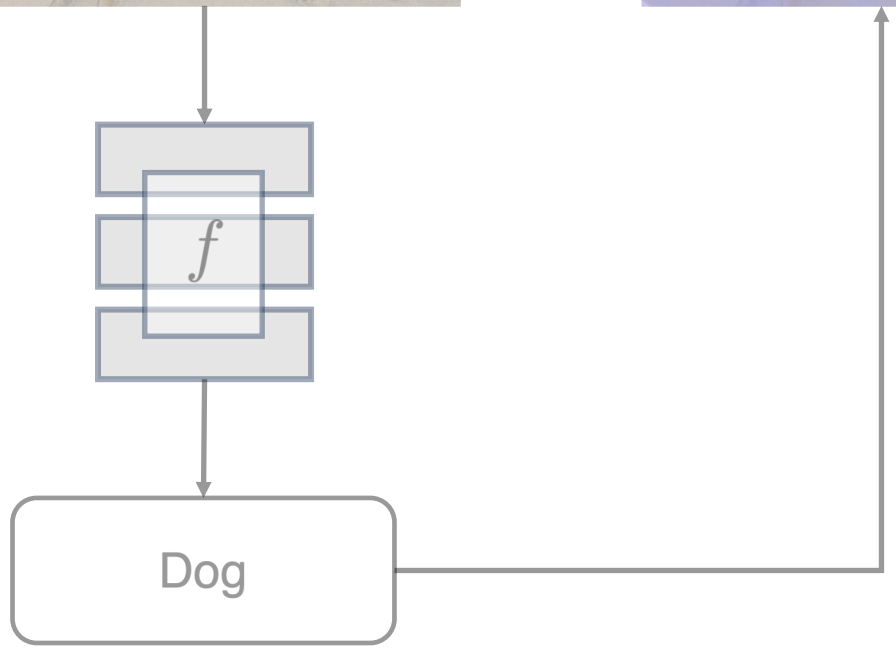
# Post-hoc Explanations




GRAD-CAM. Selvaraju et al. ICCV 2017.

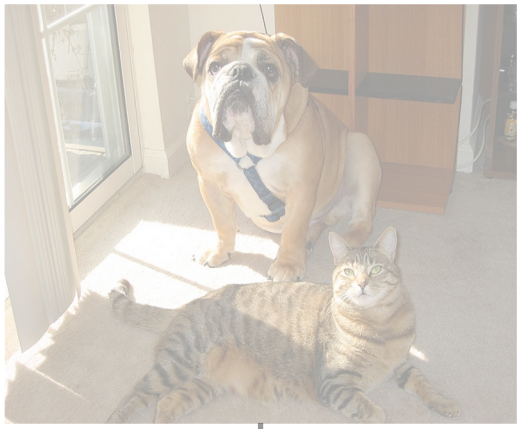
# Post-hoc Explanations Interpretable by design

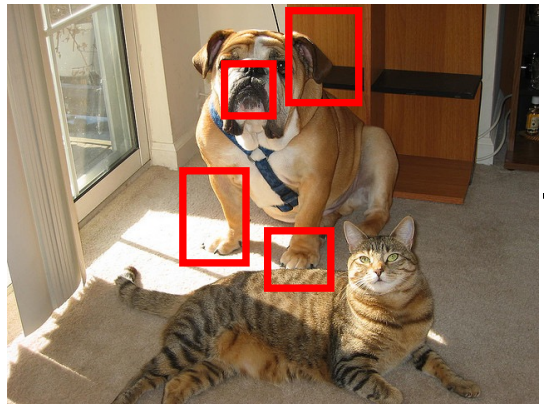




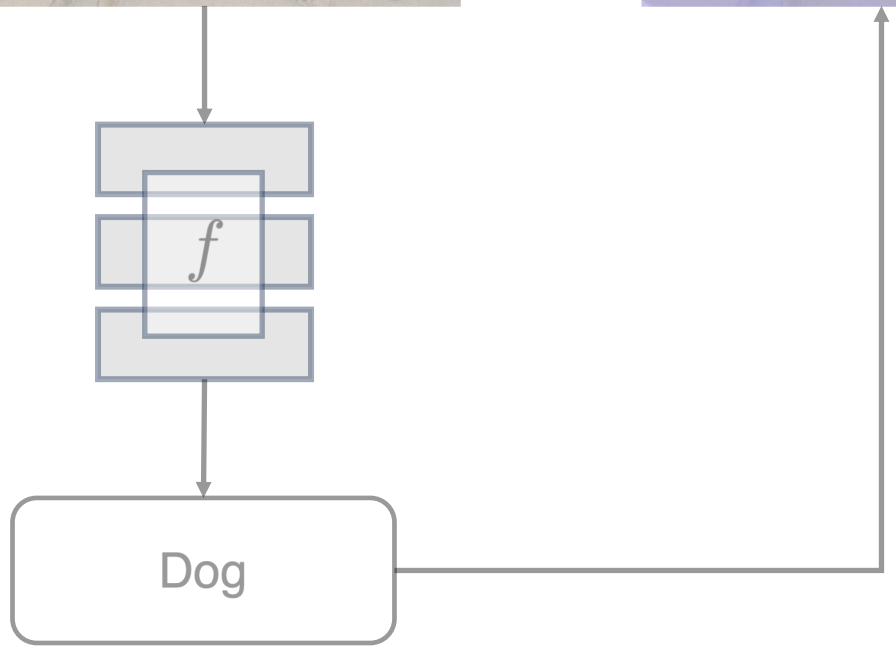

GRAD-CAM. Selvaraju et al. ICCV 2017.

# Post-hoc Explanations Interpretable by design





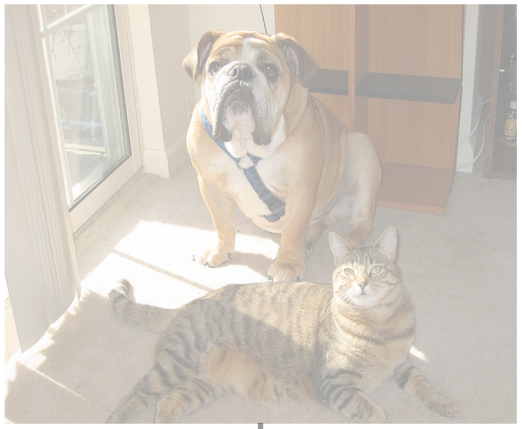
thin ears
shortened muzzle
round feet
....

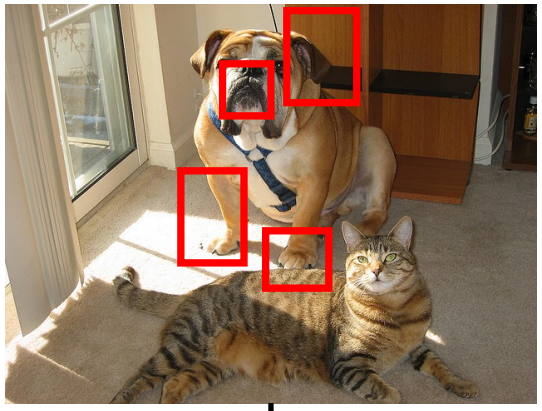


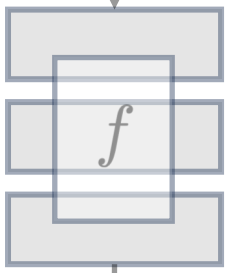

GRAD-CAM. Selvaraju et al. ICCV 2017.

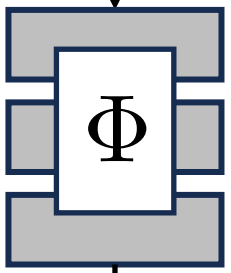
Concept Bottleneck Model. Koh et al. ICML 2020

# Post-hoc Explanations Interpretable by design







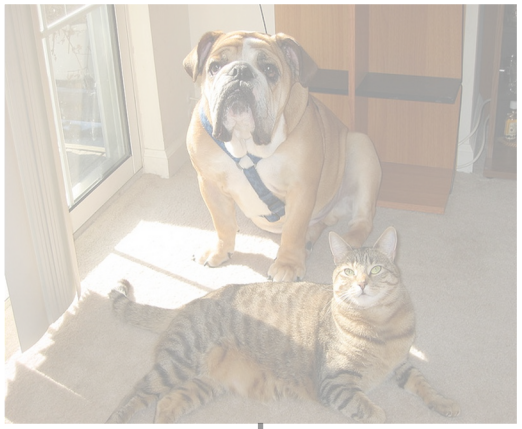

Dog

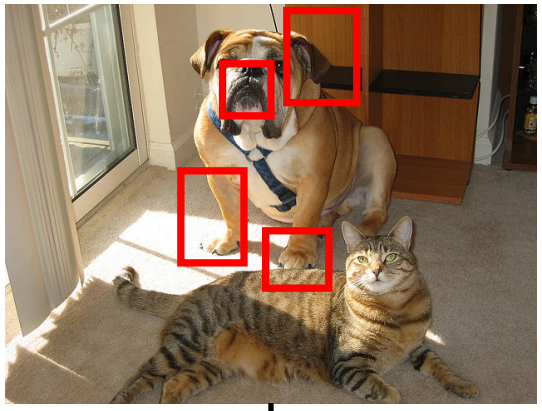
{thin ears, shortened muzzle, round feet….}

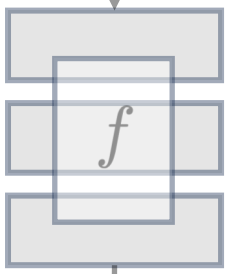
GRAD-CAM. Selvaraju et al. ICCV 2017.

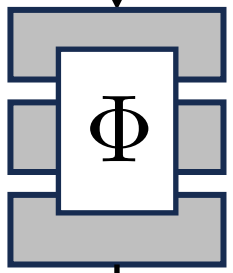
Concept Bottleneck Model. Koh et al. ICML 2020

# Post-hoc Explanations Interpretable by design

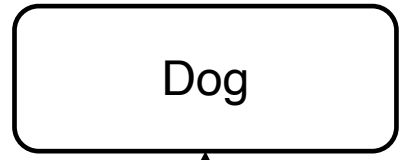


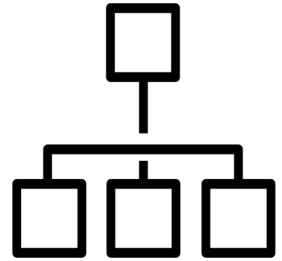











GRAD-CAM. Selvaraju et al. ICCV 2017.

Concept Bottleneck Model. Koh et al. ICML 2020

# Post-hoc explanations

# Post-hoc explanations

Pros:

- Does not alter the Black box

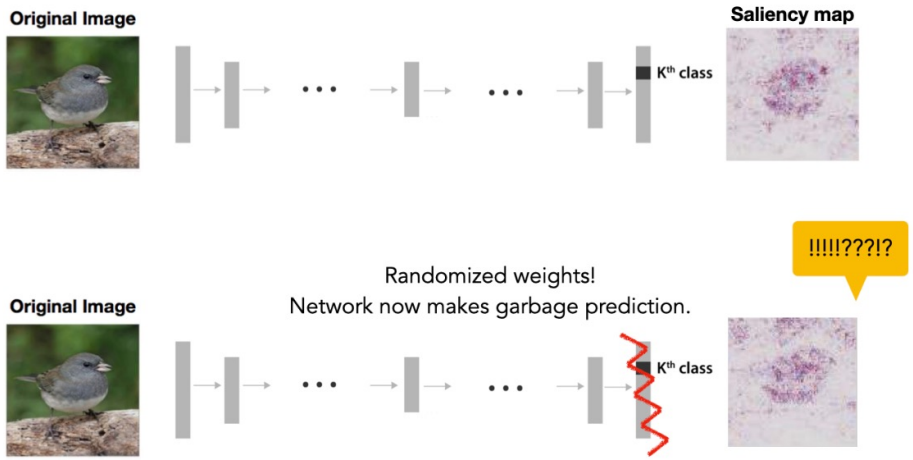
# Post-hoc explanations

Pros:

- Does not alter the Black box

Cons:

- Inconsistent explanations
- No recourse

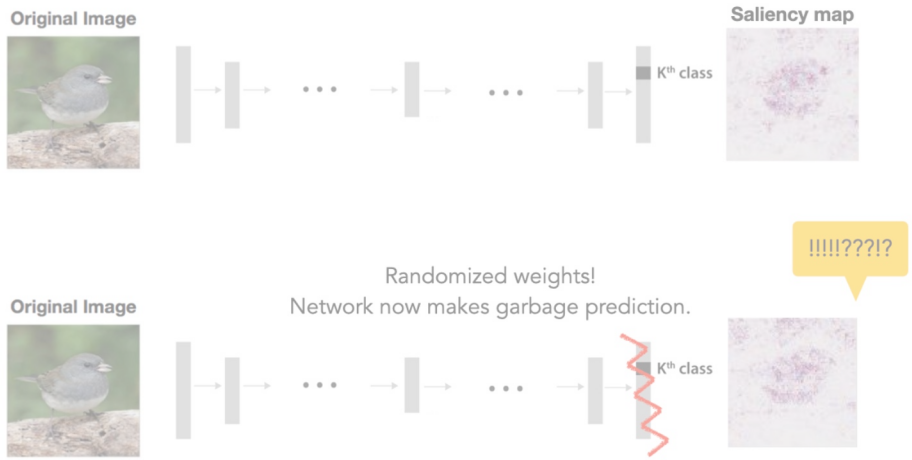



Stop explaining Black Box. Rudin et al. arXiv 2019.
Sanity Checks for Saliency Maps. Adebayo et al. Neurips 2018.

# Post-hoc explanations Interpretable by design

Pros:
- Does not alter the Black box

Cons:
- Inconsistent explanations
- No recourse



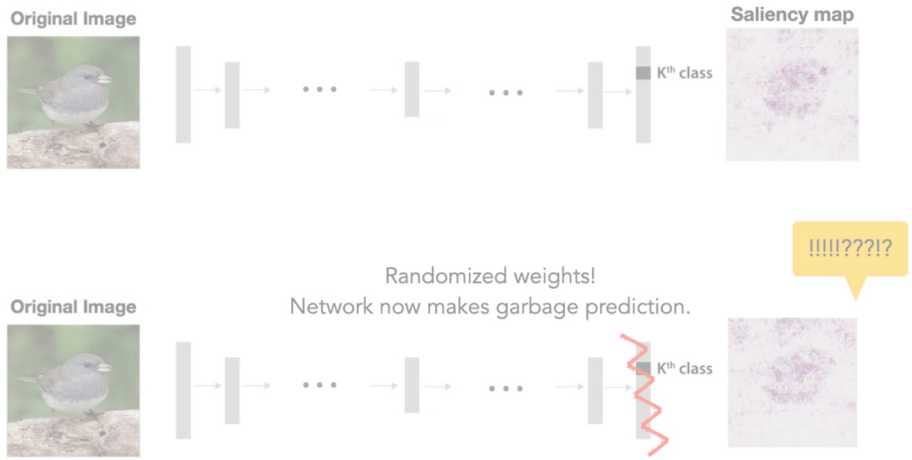

Pros:
- Support concept intervention

Stop explaining Black Box. Rudin et al. arXiv 2019.
Sanity Checks for Saliency Maps. Adebayo et al. Neurips 2018.

# Post-hoc explanations Interpretable by design

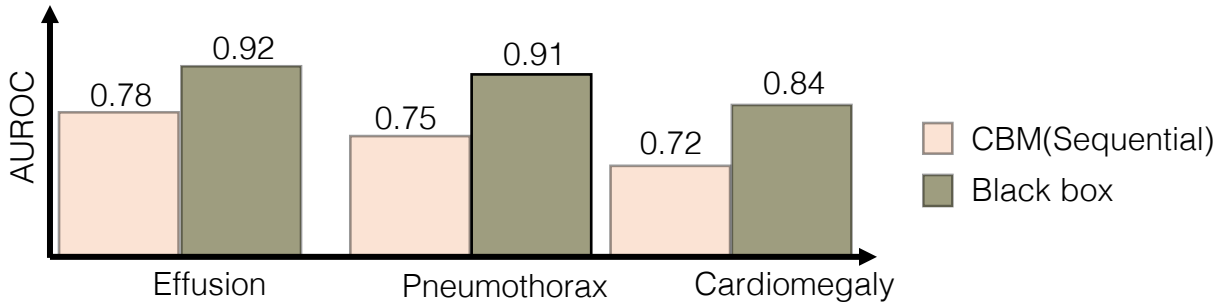
Pros:
- Does not alter the Black box

Cons:
- Inconsistent explanations
- No recourse




Pros:
- Support concept intervention

Cons:
- Harder to train
- Sub par performance




Stop explaining Black Box. Rudin et al. arXiv 2019.
Sanity Checks for Saliency Maps. Adebayo et al. Neurips 2018.

# Post-hoc explanations Interpretable by design

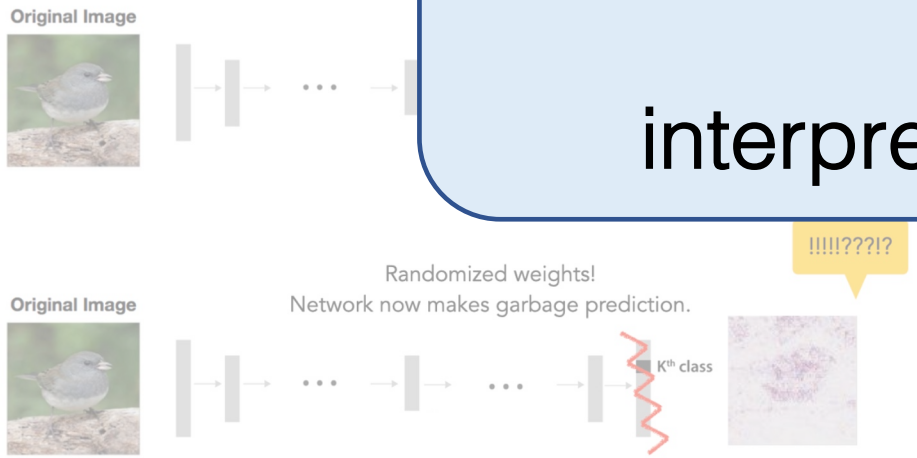
Pros:
- Does not alter the Black box

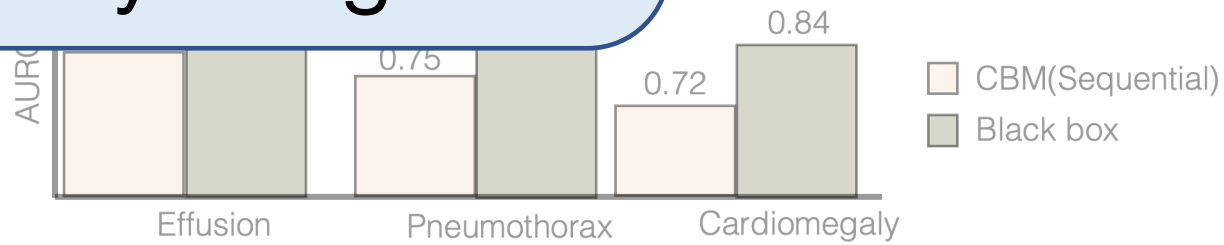
Cons:
- Inconsistent
- No recourse

Pros:
- Support concept intervention

ance

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




Stop explaining Black Box. Rudin et al. arXiv 2019.
Sanity Checks for Saliency Maps. Adebayo et al. Neurips 2018.

# 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

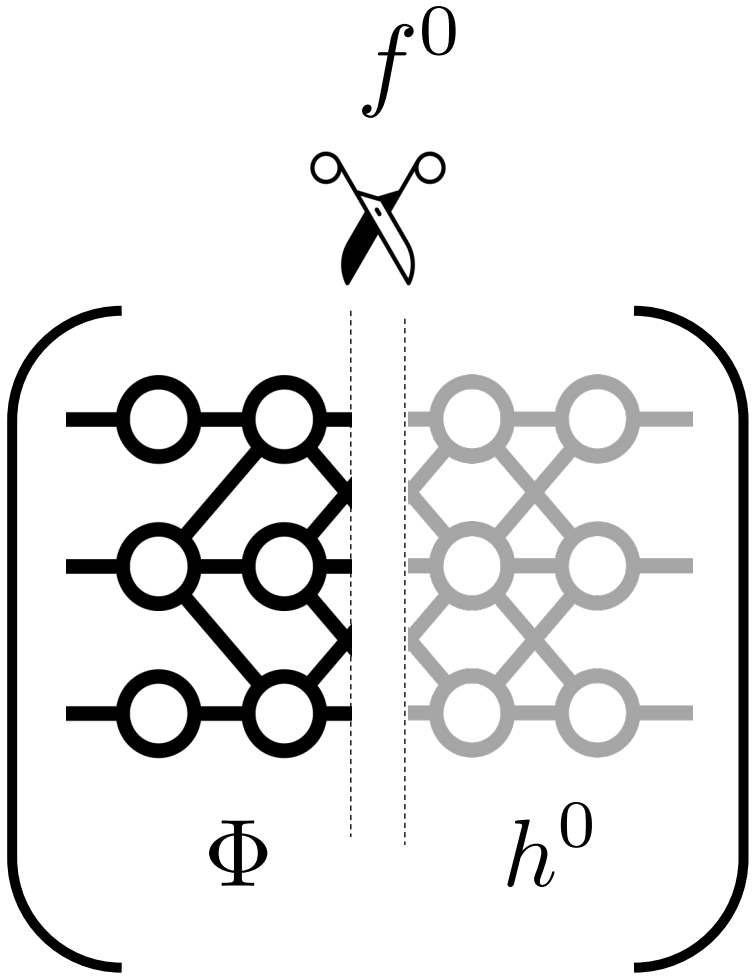
# Desirable properties

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

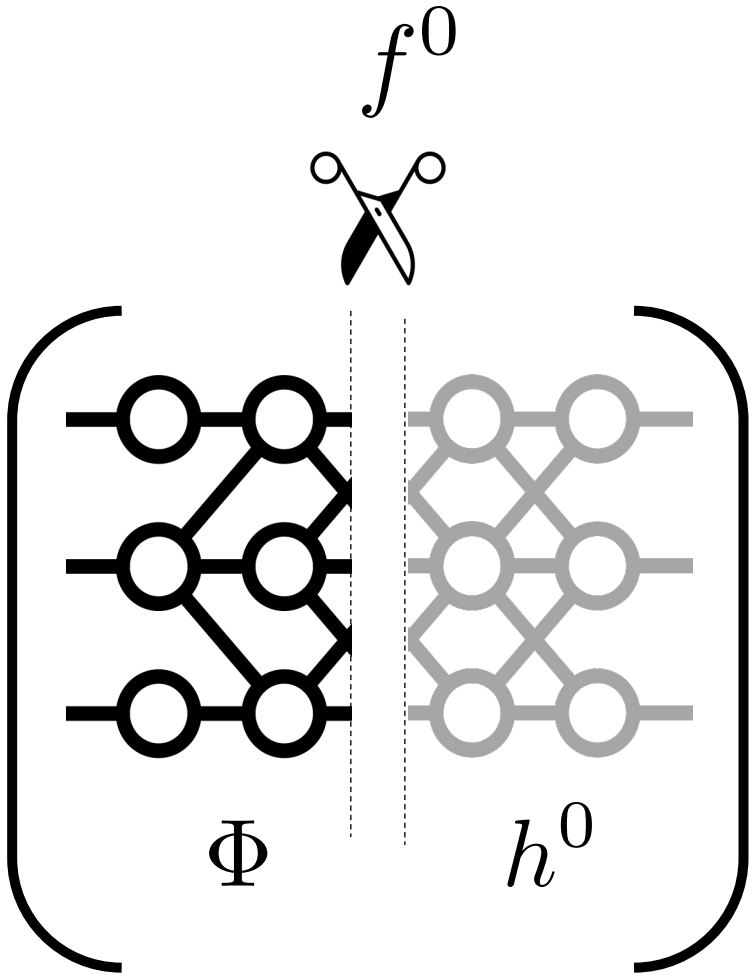
# Design choices

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

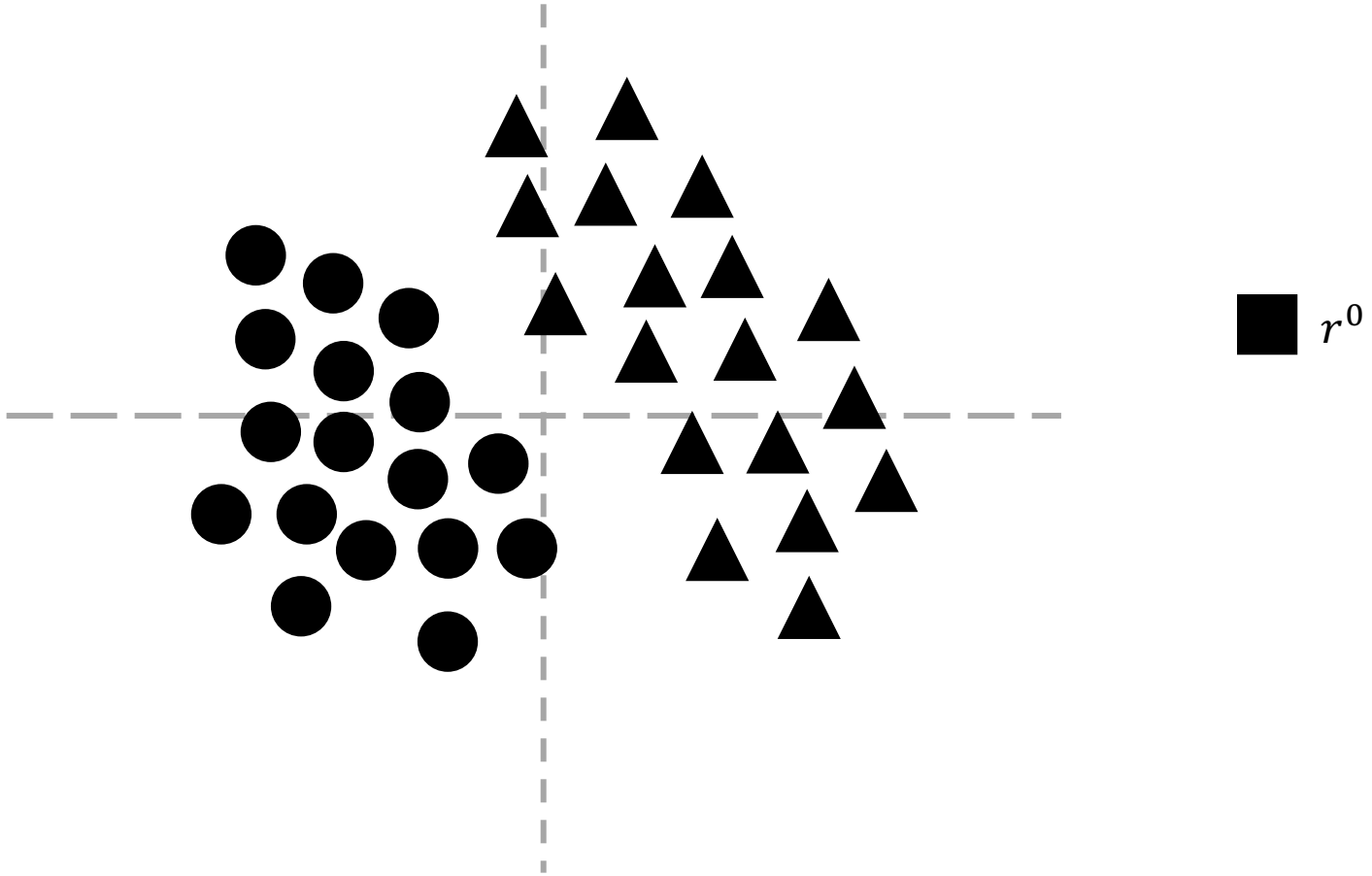


# Assumptions

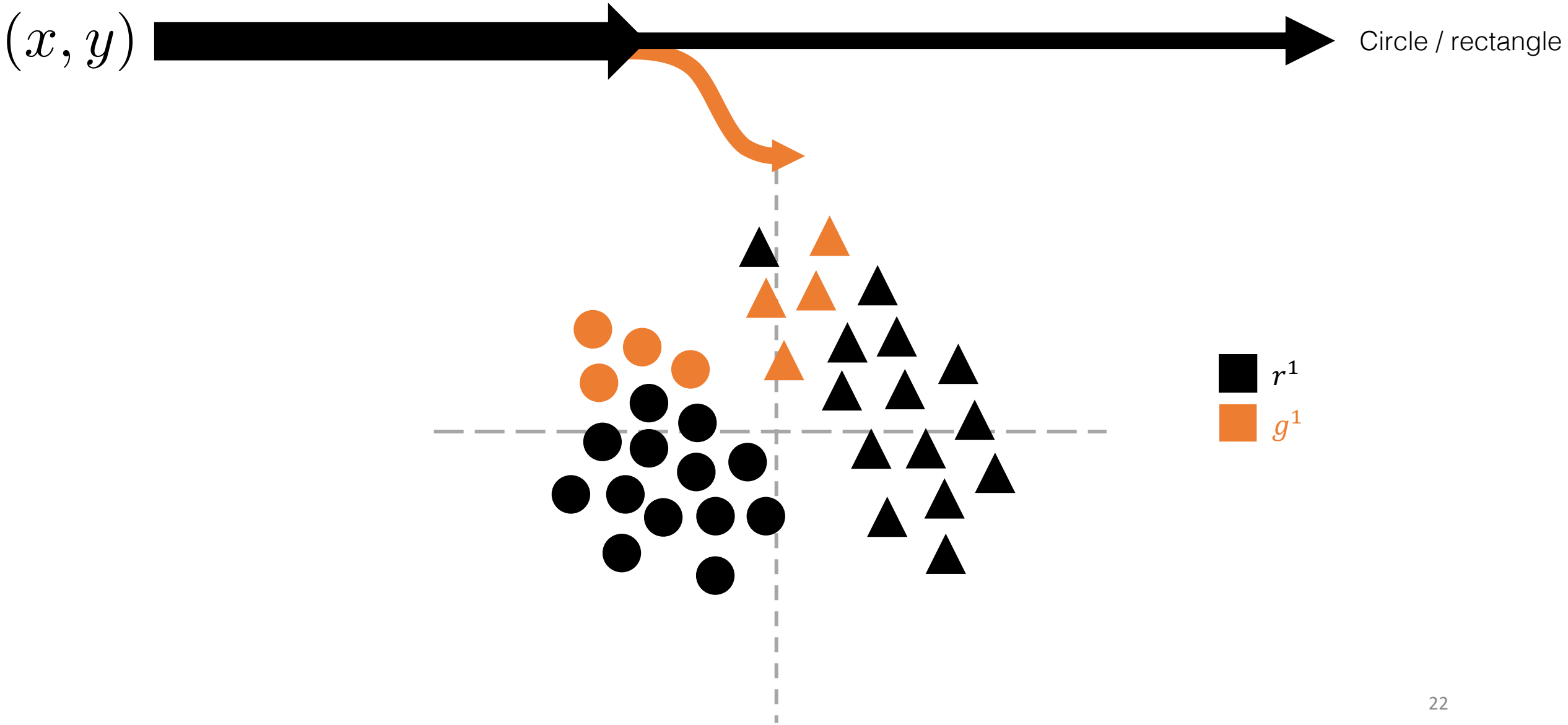



# Iteratively carve out interpretable models

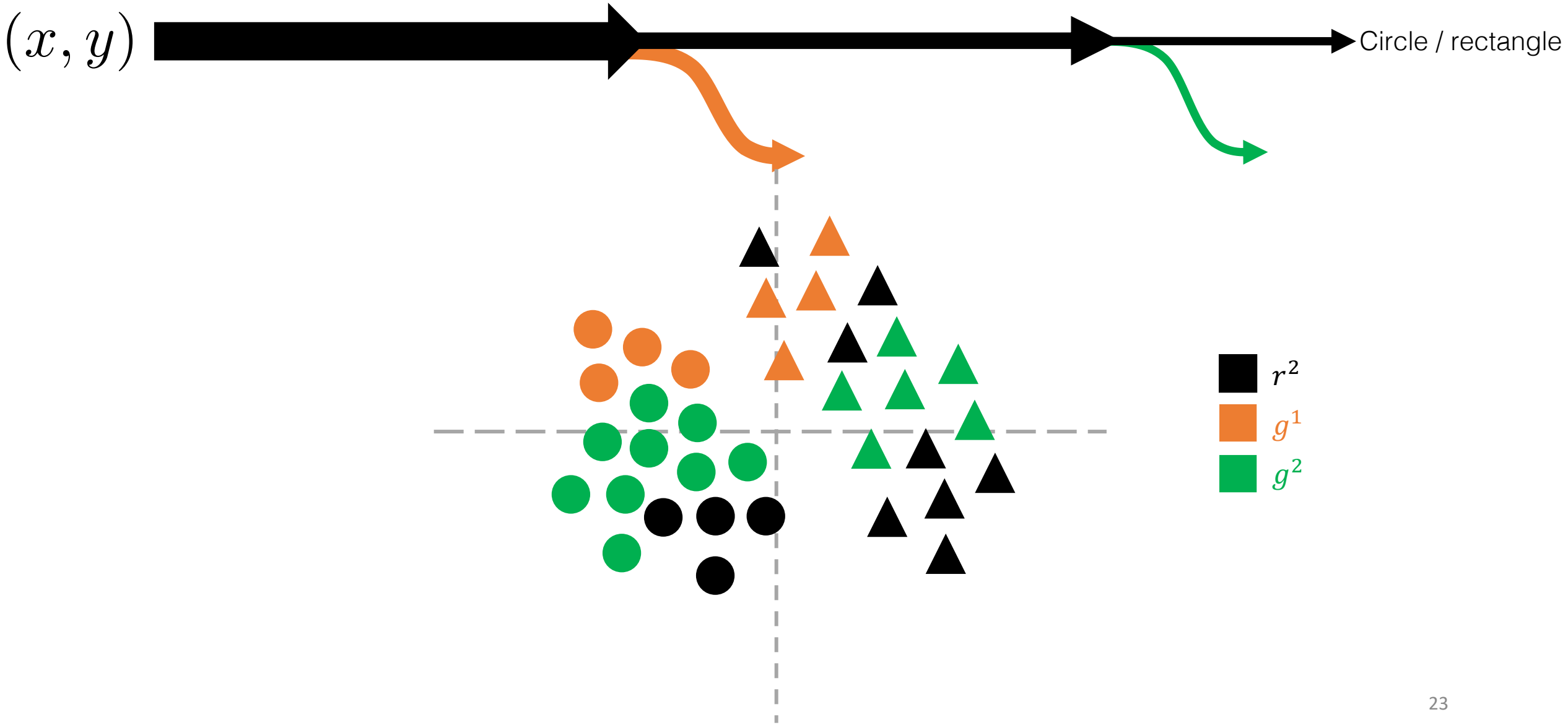
$(x, y)$ → Circle / rectangle



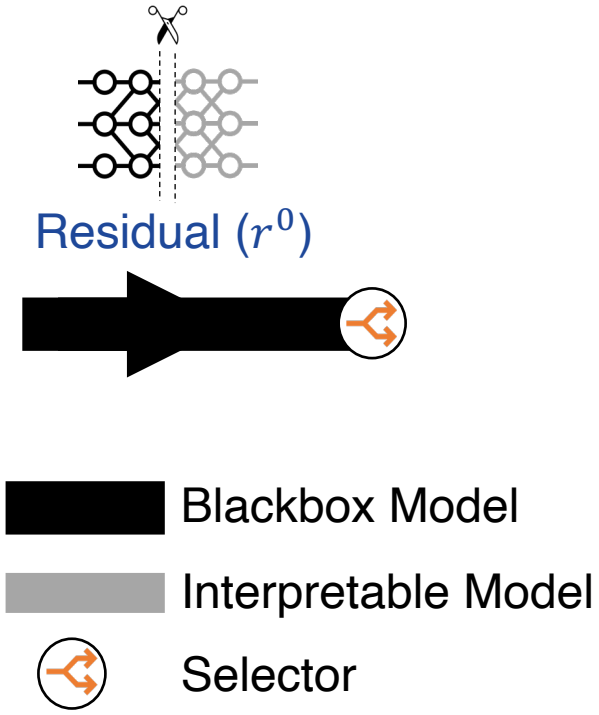
# Iteratively carve out interpretable models



# Iteratively carve out interpretable models

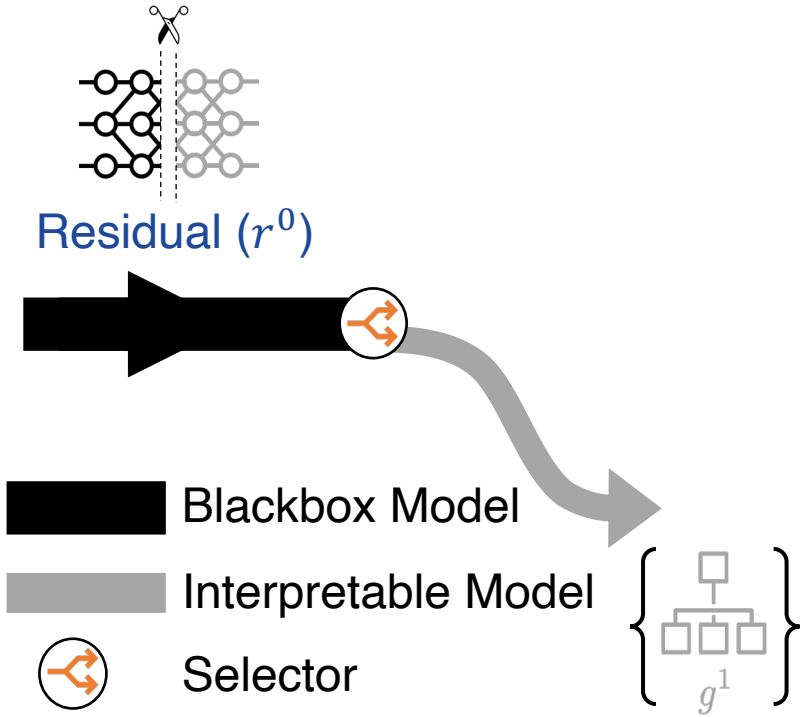


# Iteratively carve out interpretable models



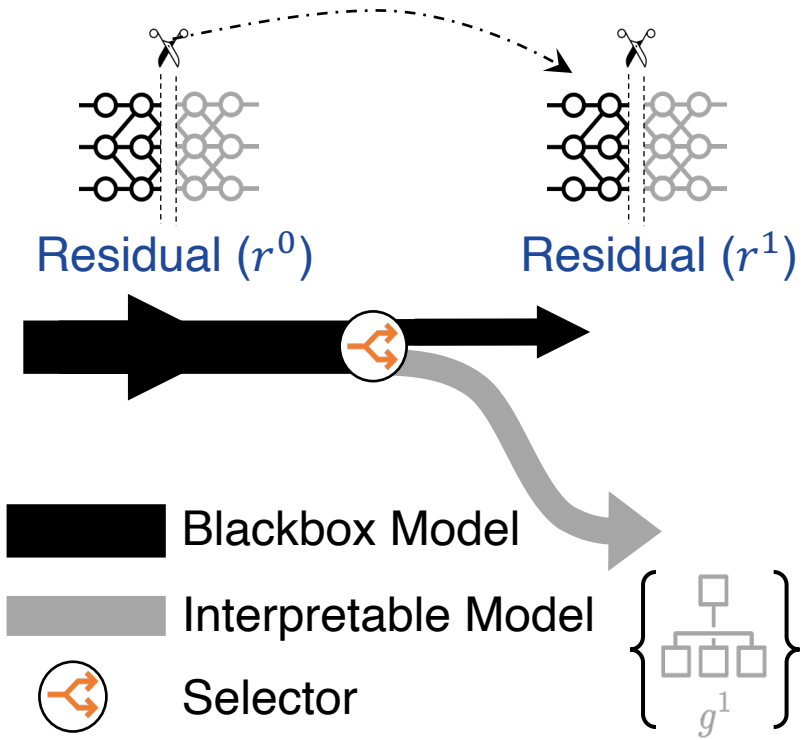

Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



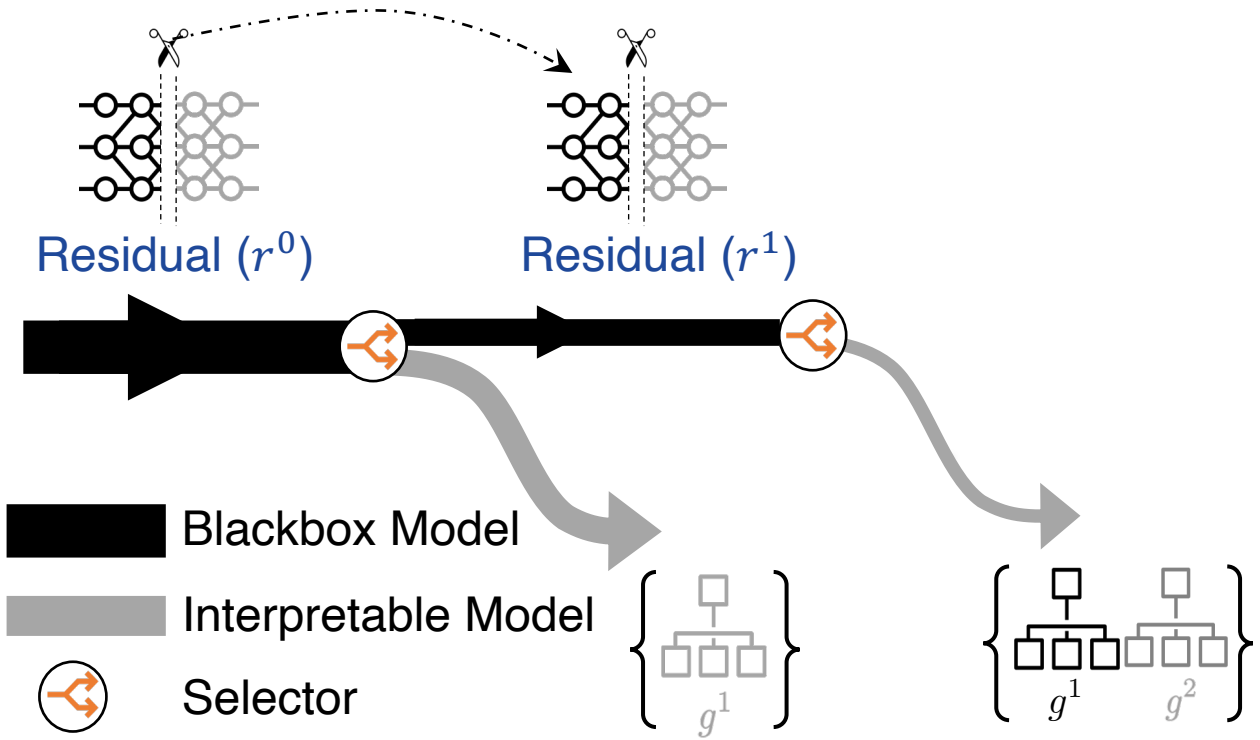

Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



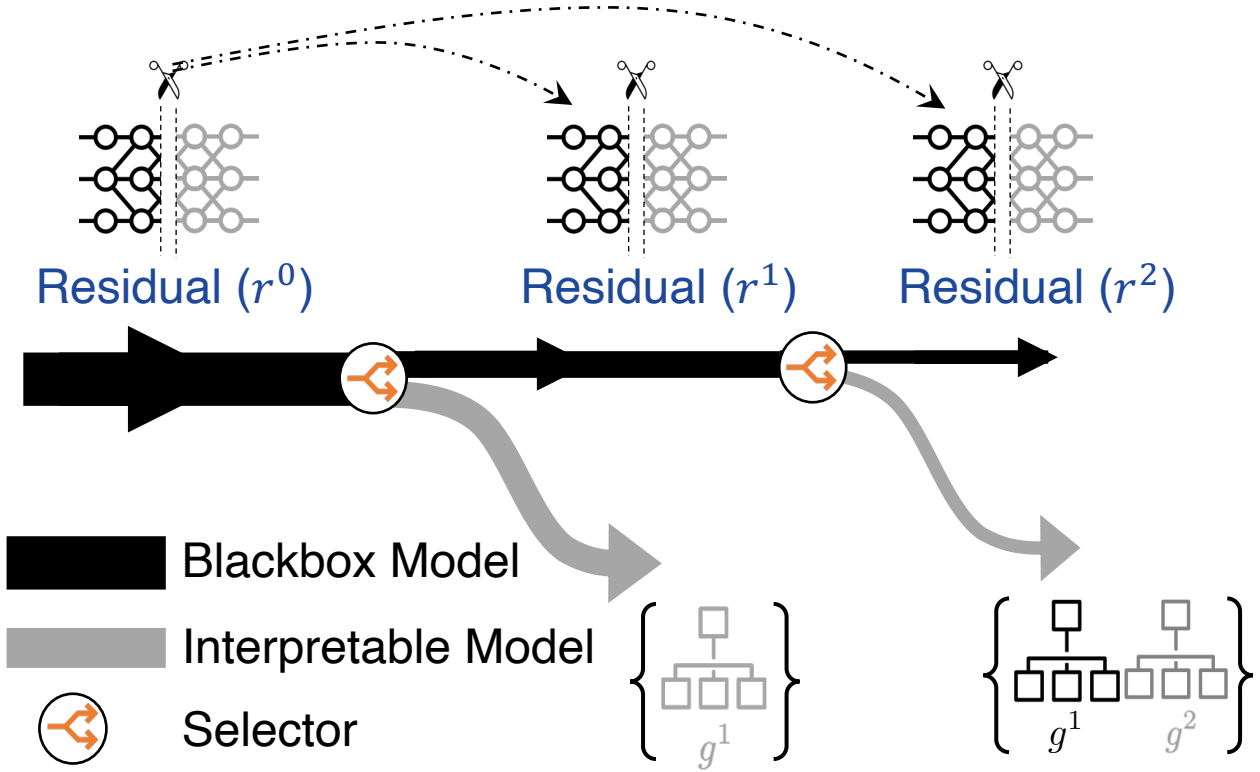

Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



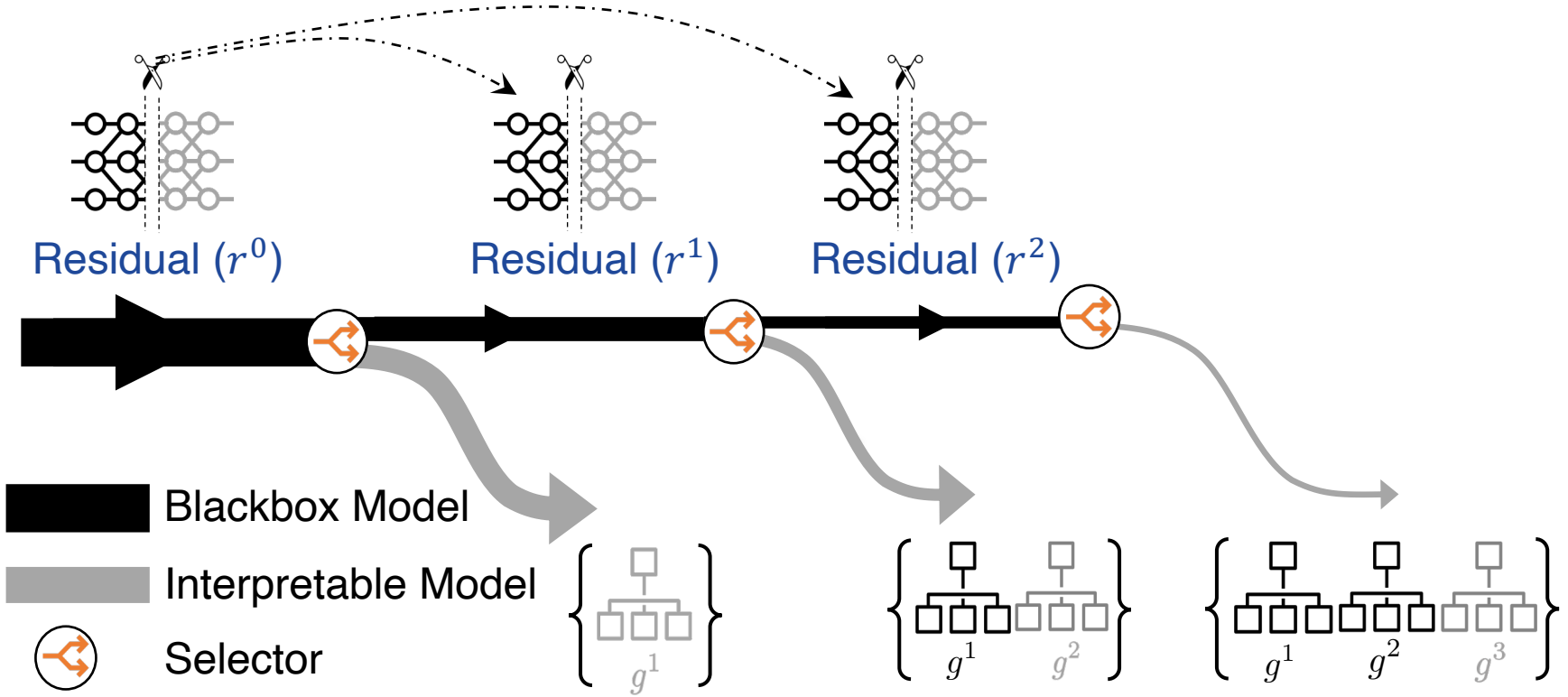

Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



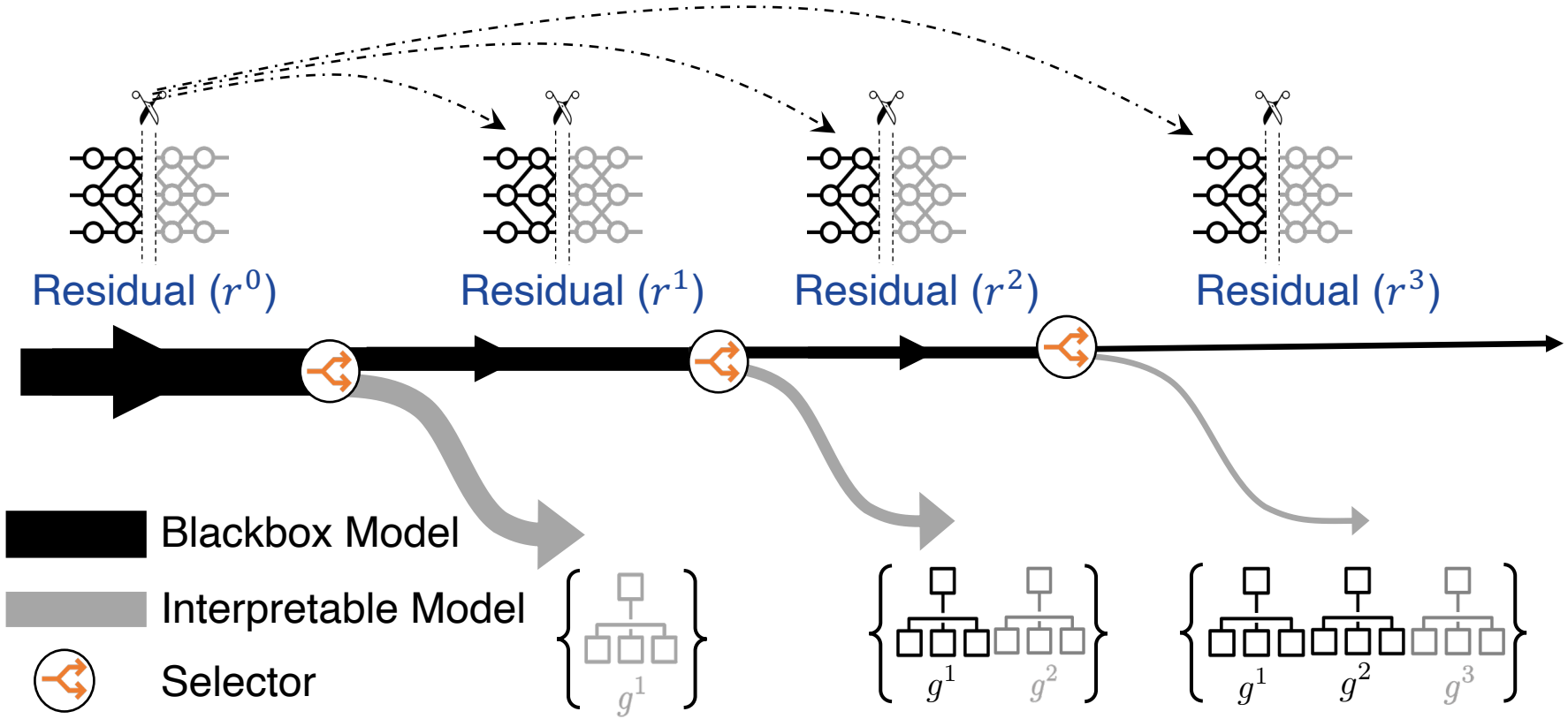

Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# Iteratively carve out interpretable models



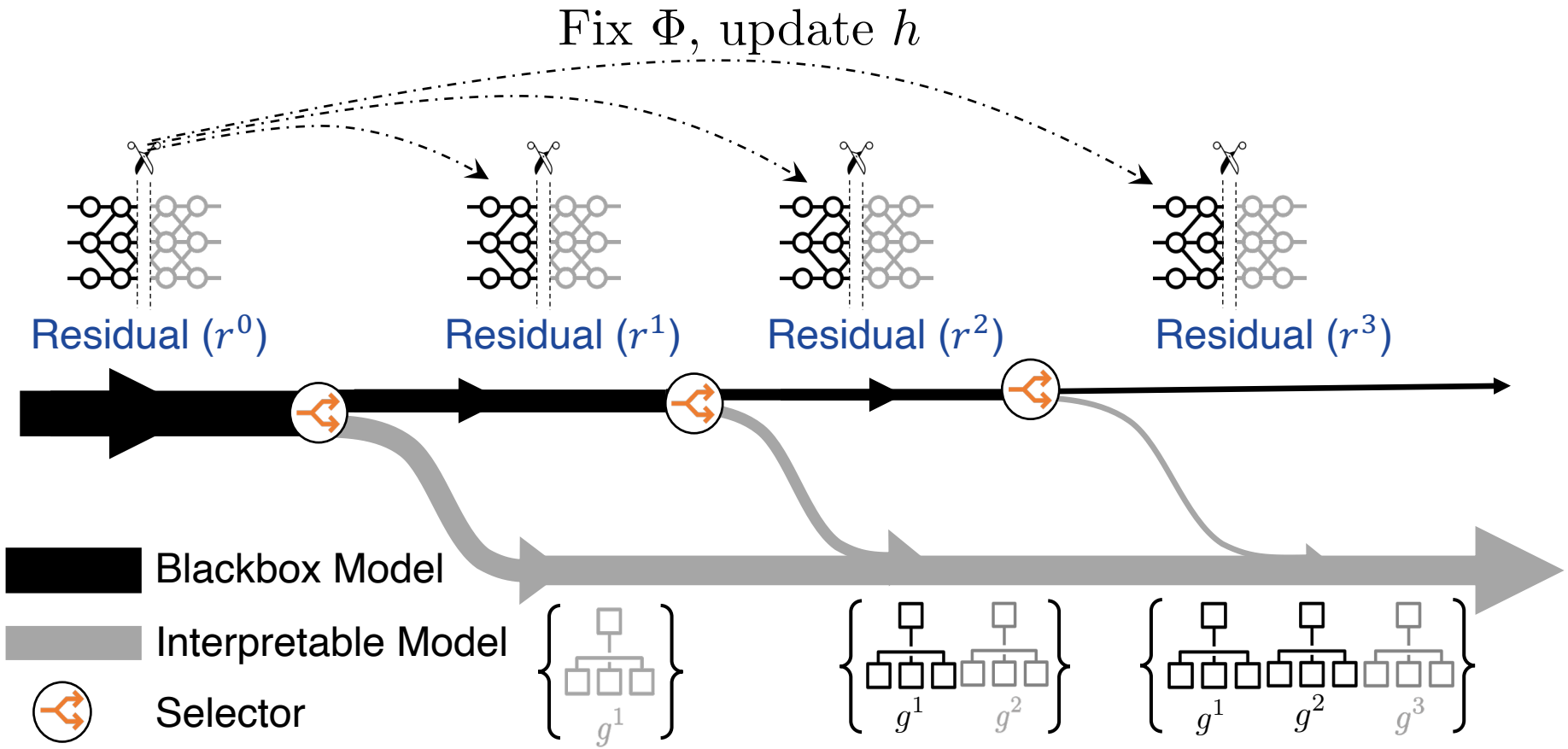

Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

# 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** ↔breast_color_grey∧
tail_pattern_solid

Expert 2

**Olive sided Flycatcher** ↔underparts_color_grey∧
wing_color_grey

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

# Provide the concept-interaction

Expert 1


**Olive sided Flycatcher** ↔breast_color_grey∧
tail_pattern_solid

Expert 2

**Olive sided Flycatcher** ↔underparts_color_grey∧
wing_color_grey

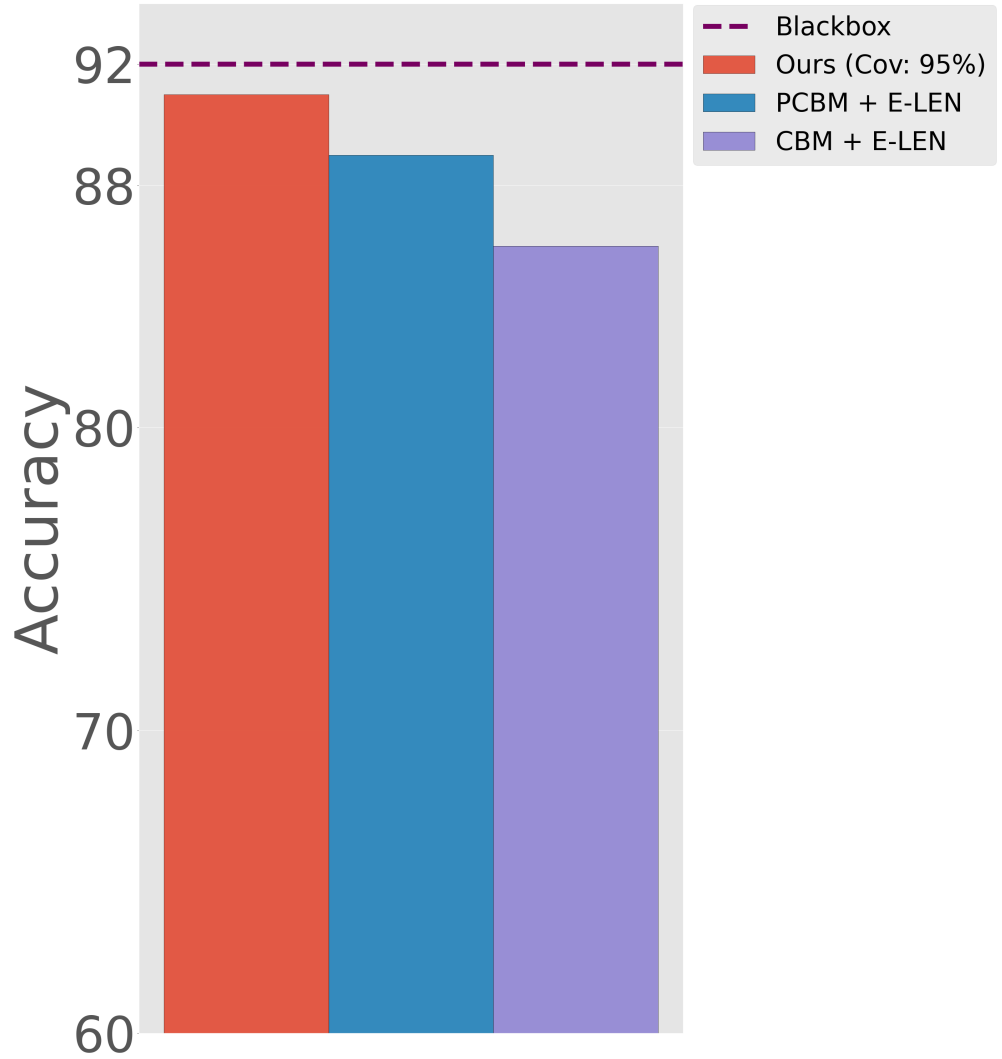
Final Residual (Unexplained)

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

# Does not compromise performance

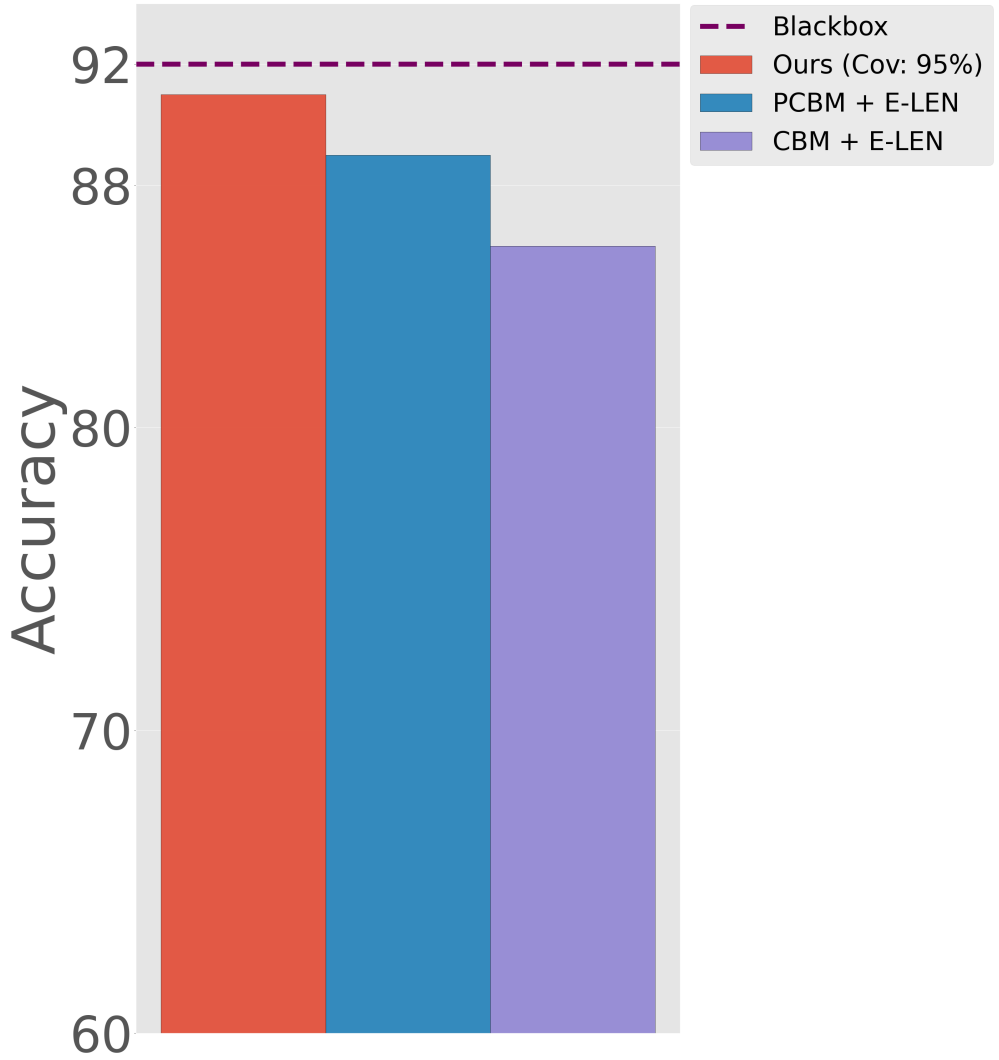
# Does not compromise performance

## CUB-200 with ViT

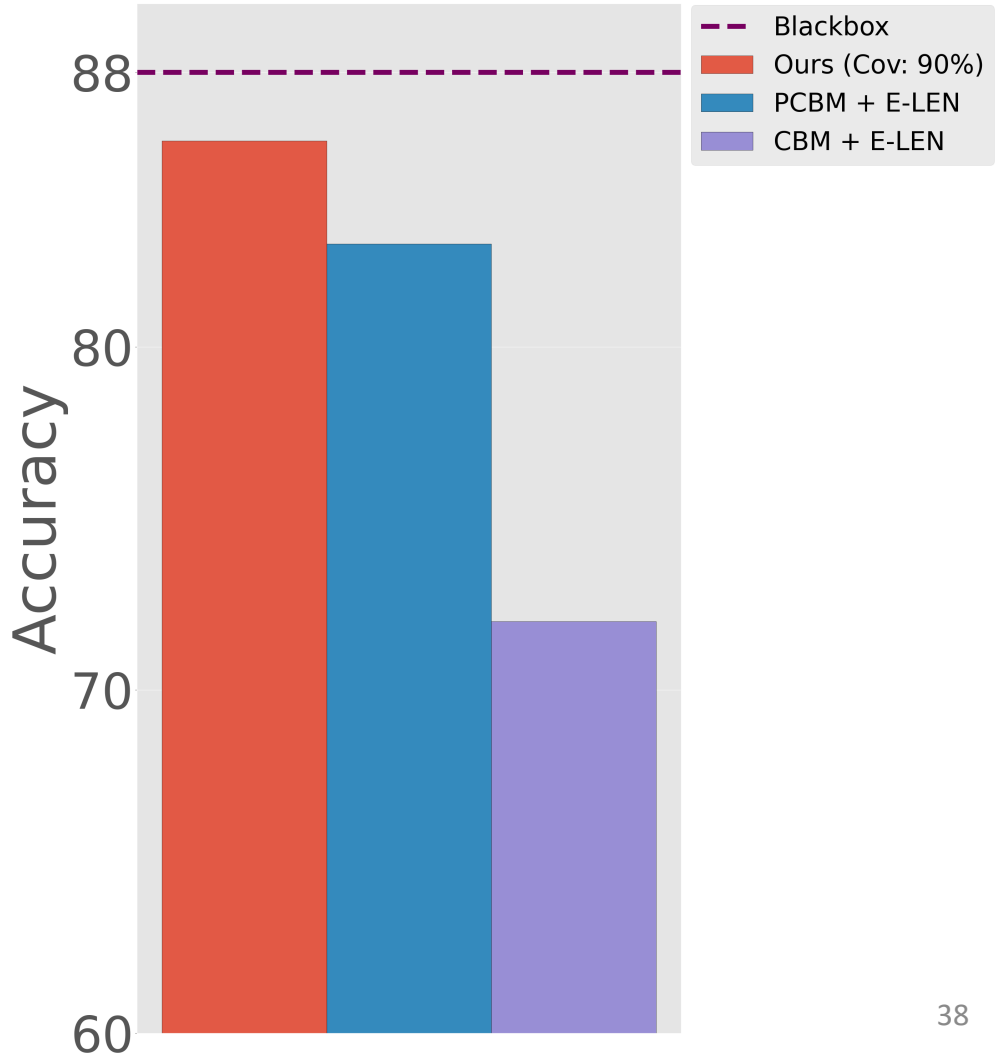


# Does not compromise performance

## CUB-200 with ViT

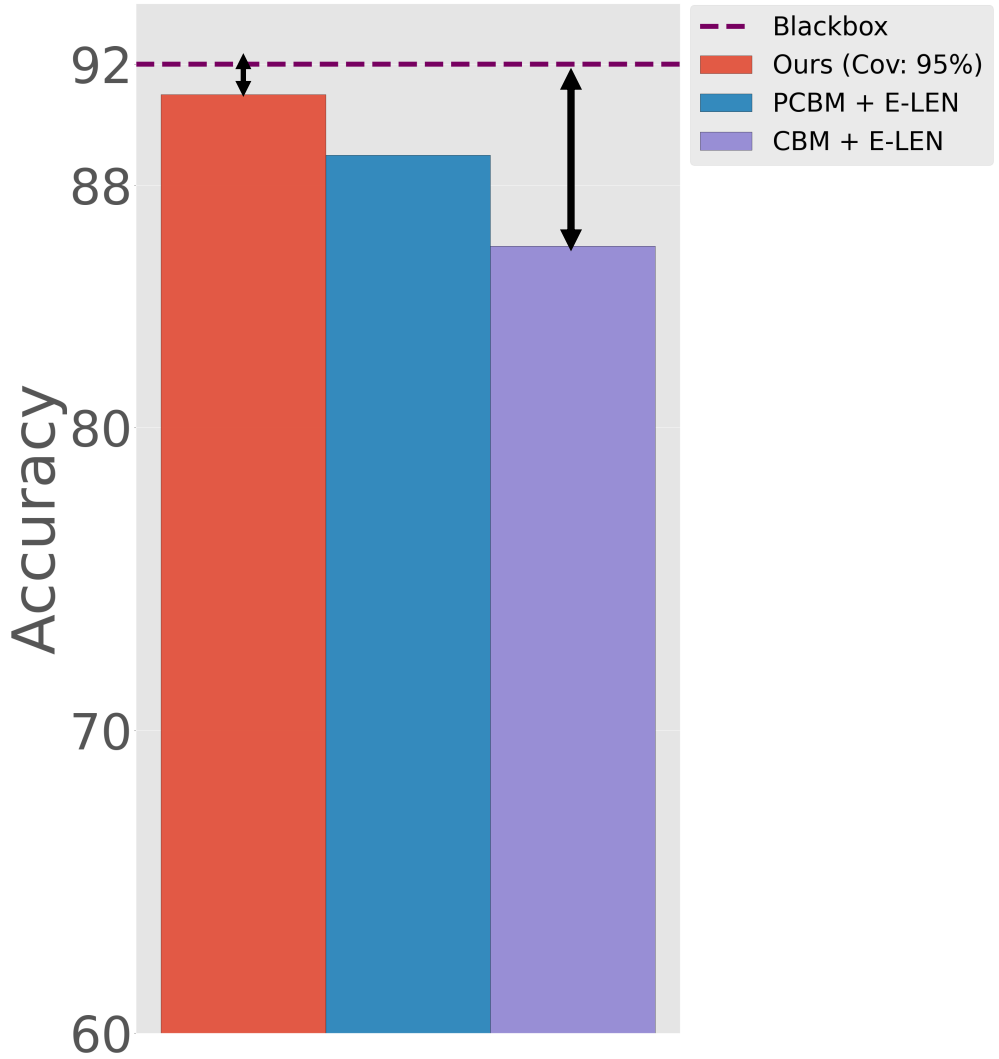



## CUB-200 with ResNet101

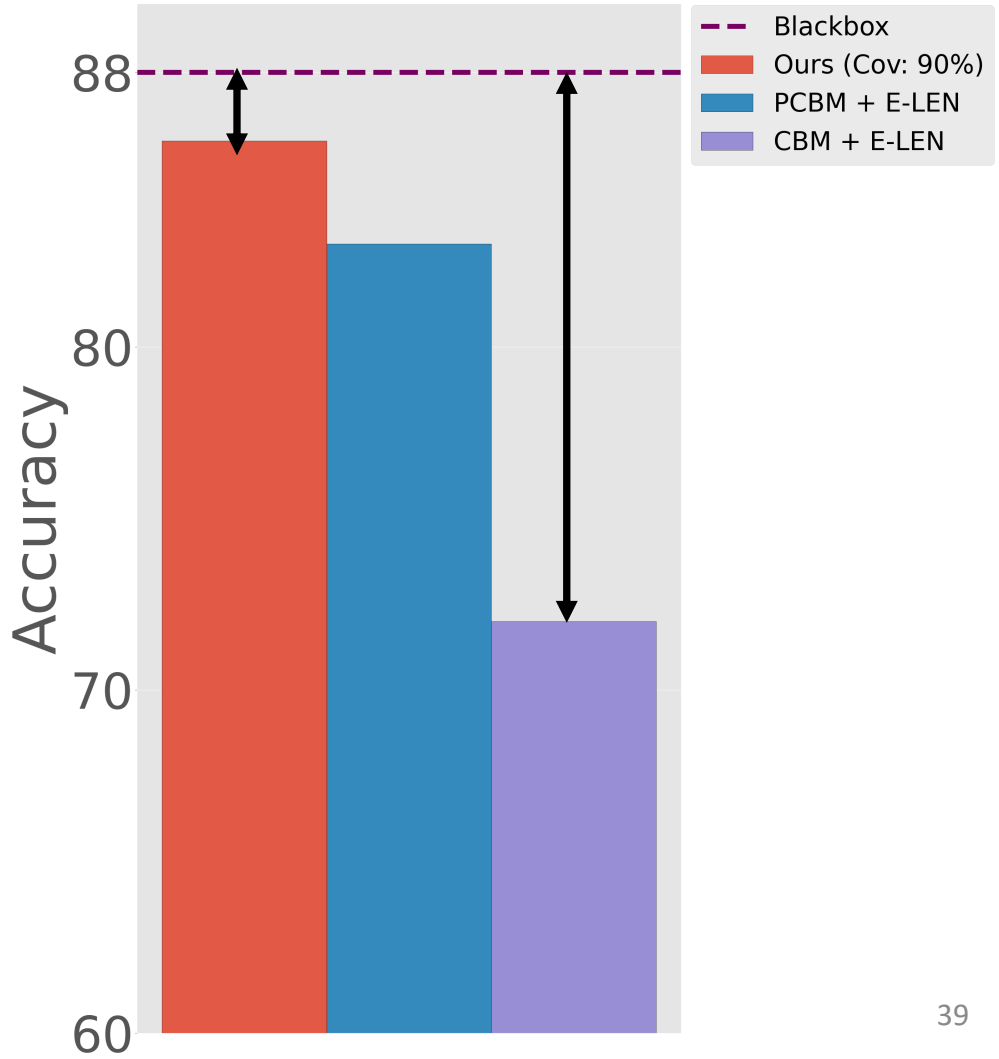


# Does not compromise performance

## CUB-200 with ViT



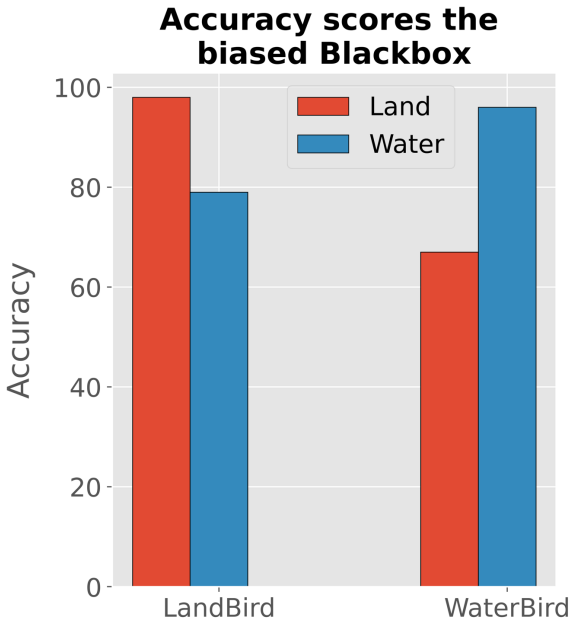

## CUB-200 with ResNet101



# Application on shortcut

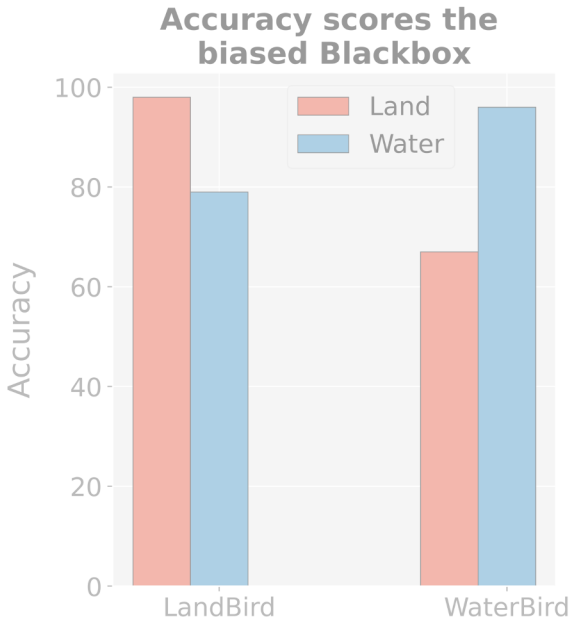
Tackling Shortcut Learning in Deep Neural Networks. Ghosh et al. SCIS Workshop, ICML 2023.

# Application on shortcut




Tackling Shortcut Learning in Deep Neural Networks. Ghosh et al. SCIS Workshop, ICML 2023.

# Application on shortcut




Biased Blackbox

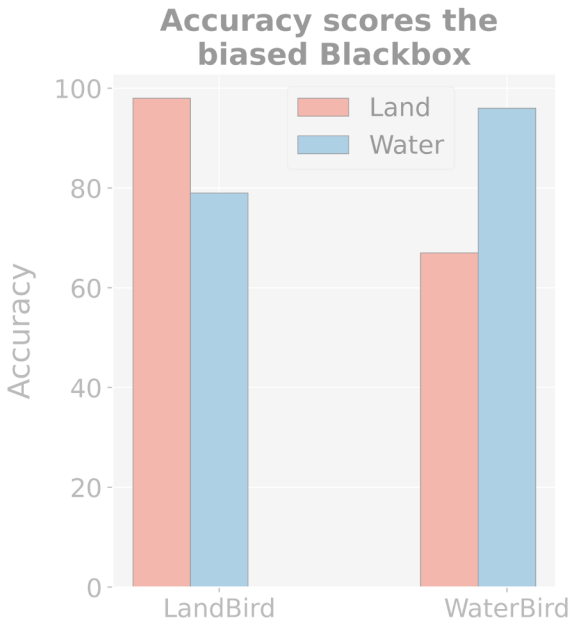
Groundtruth: WaterBird

Prediction : LandBird

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

Tackling Shortcut Learning in Deep Neural Networks. Ghosh et al. SCIS Workshop, ICML 2023.

# Application on shortcut

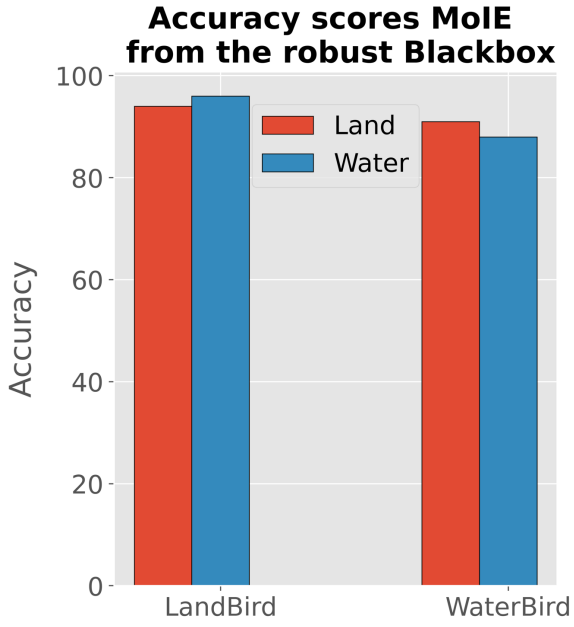



Biased Blackbox

Groundtruth: WaterBird

Prediction : LandBird

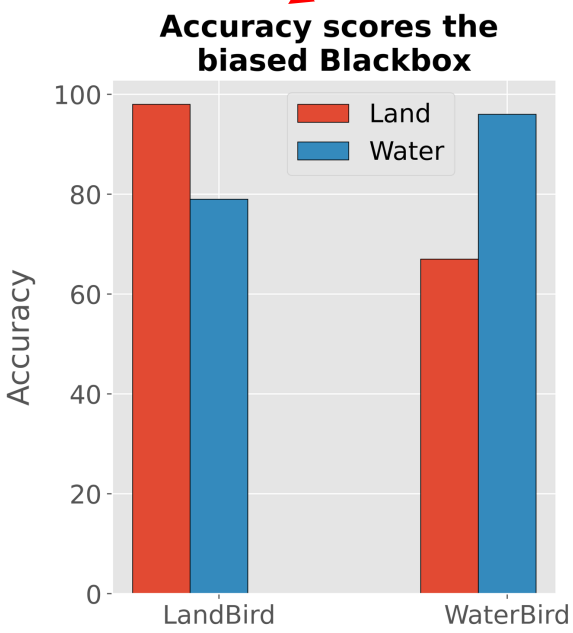
Explanation : LandBird $\leftrightarrow$ WingShapeRoundedwings $\wedge$ Forest






Tackling Shortcut Learning in Deep Neural Networks. Ghosh et al. SCIS Workshop, ICML 2023.

# Application on shortcut

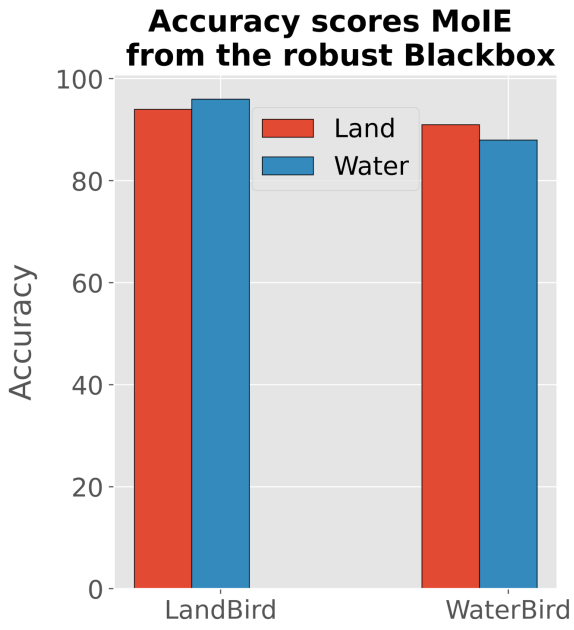



Biased Blackbox

Groundtruth: WaterBird

Prediction : LandBird

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






Tackling Shortcut Learning in Deep Neural Networks. Ghosh et al. SCIS Workshop, ICML 2023.

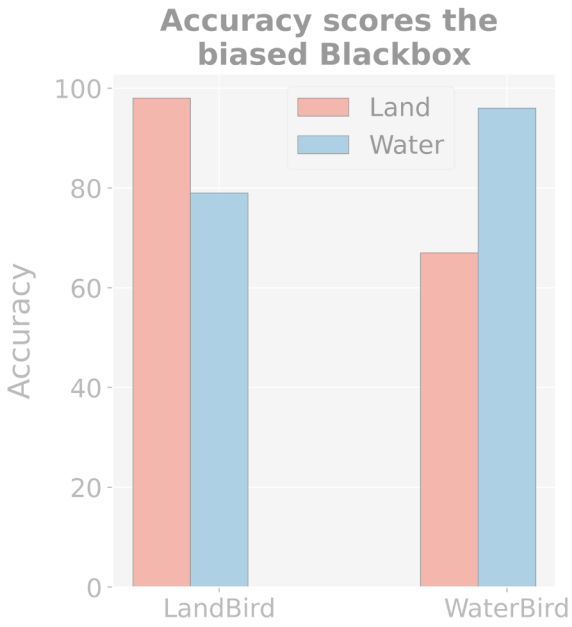
# Application on shortcut

Robust Blackbox

Groundtruth: WaterBird

Prediction : WaterBird

Explanation : WaterBird $\leftrightarrow$ BillLengthAboutTheSameAsHead $\wedge \neg$BillLengthShorterThanHead $\wedge \neg$SizeSmall5__9in $\wedge \neg$ShapePerchingLike $\wedge$ CrownColorWhite

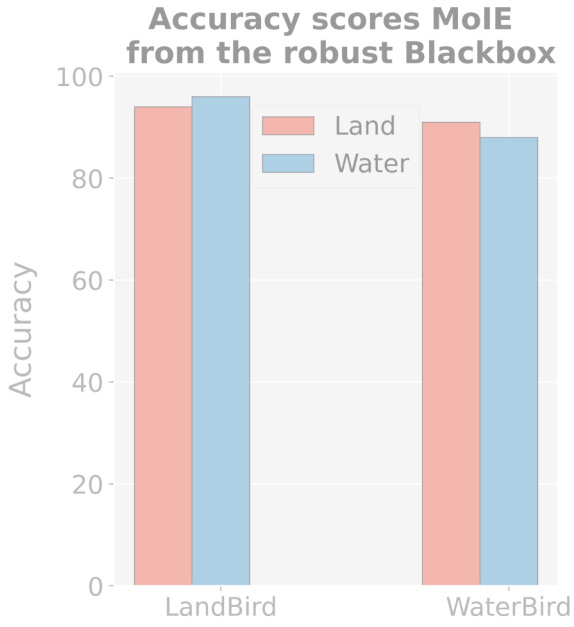



Biased Blackbox

Groundtruth: WaterBird

Prediction : LandBird

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



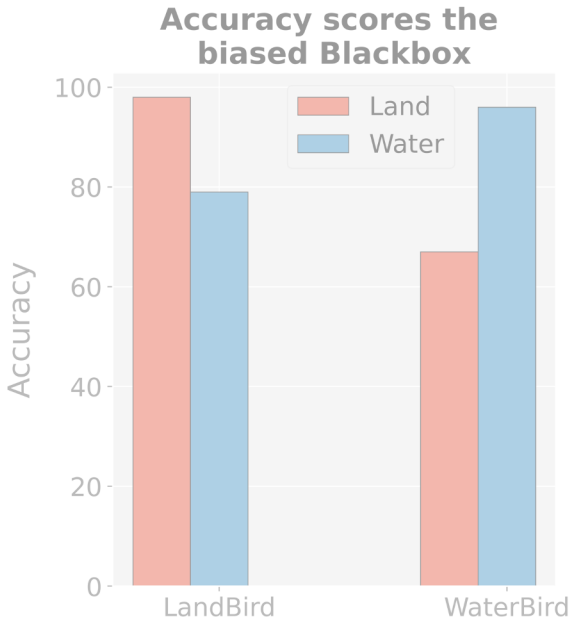



Tackling Shortcut Learning in Deep Neural Networks. Ghosh et al. SCIS Workshop, ICML 2023.

# Application on shortcut

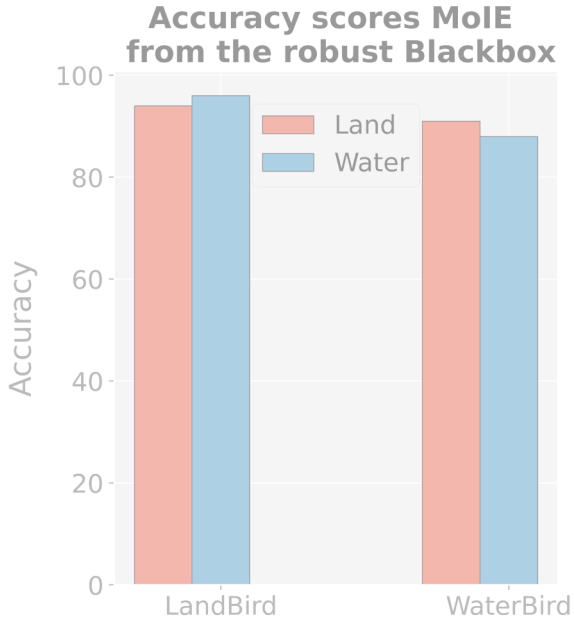
Robust Blackbox

Groundtruth: WaterBird
Prediction : WaterBird
Explanation : WaterBird $\leftrightarrow$ BillLengthAboutTheSameAsHead
$\wedge \neg$BillLengthShorterThanHead $\wedge \neg$SizeSmall5__9in
$\wedge \neg$ShapePerchingLike $\wedge$ CrownColorWhite




Biased Blackbox

Groundtruth: WaterBird
Prediction : LandBird
Explanation : LandBird $\leftrightarrow$ WingShapeRoundedwings $\wedge$ **Forest**






Tackling Shortcut Learning in Deep Neural Networks. Ghosh et al. SCIS Workshop, ICML 2023.

# We have more results

- We achieve higher completeness scores
- We performs better in test-time interventions
- We update to MoIE-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