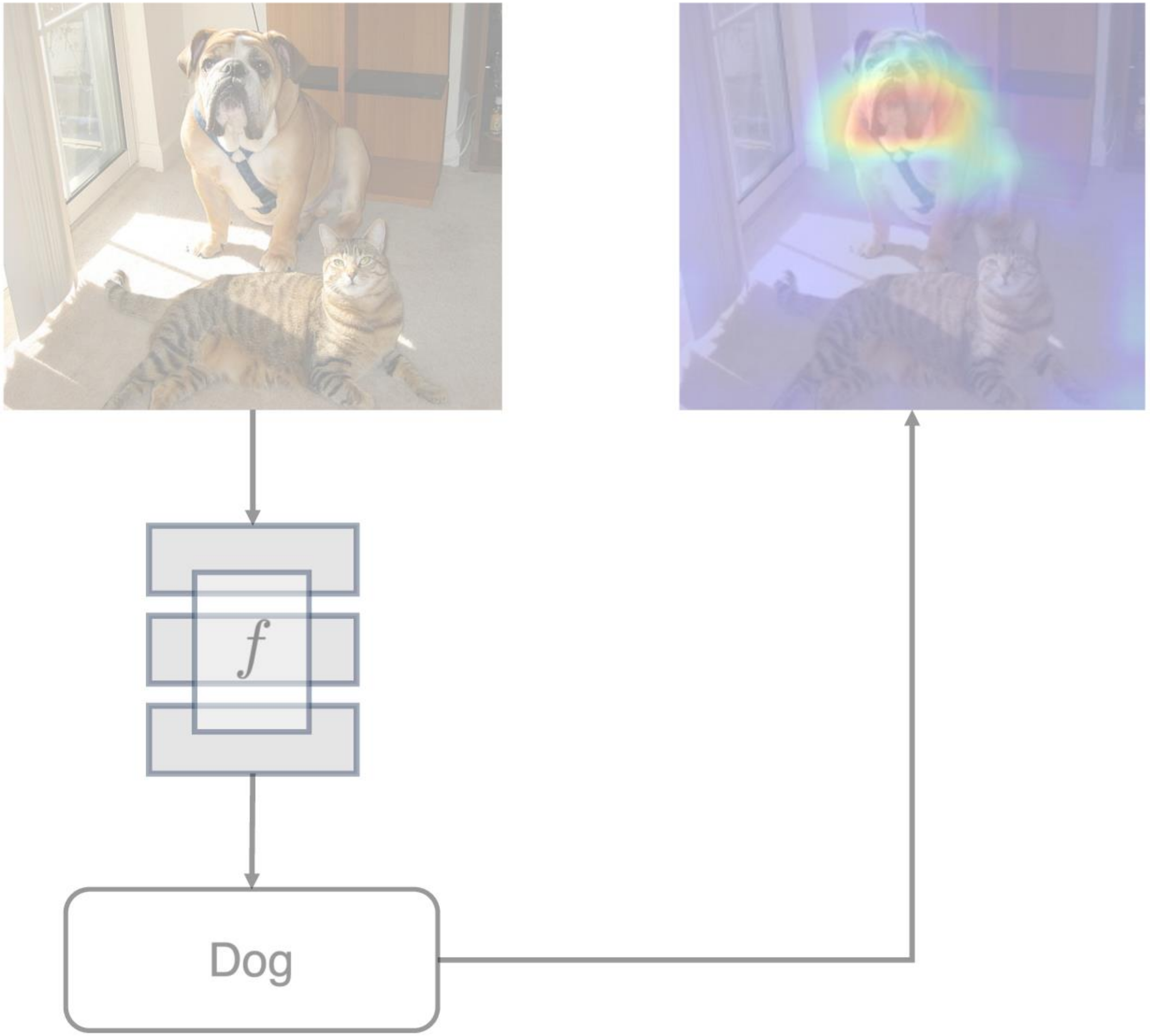
Dividing and Conquering a BlackBox to a Mixture of Interpretable Models

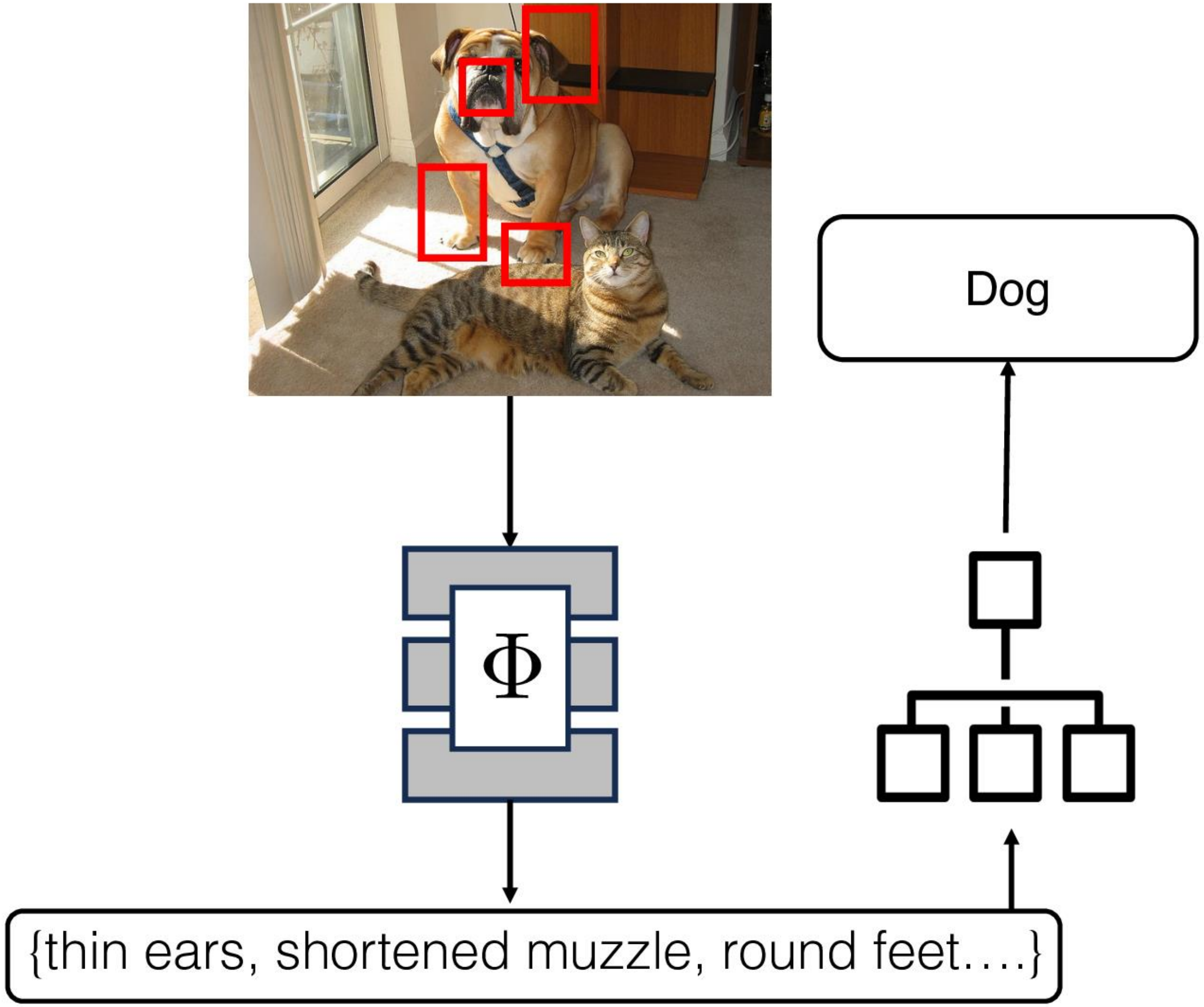
December 12, 2023

Seoul National University

Post-hoc Explanations Interpretable by design



GRAD-CAM. Selvaraju et al. ICCV 2017.



Dog

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

Concept Bottleneck Model. Koh et al. ICML 2020

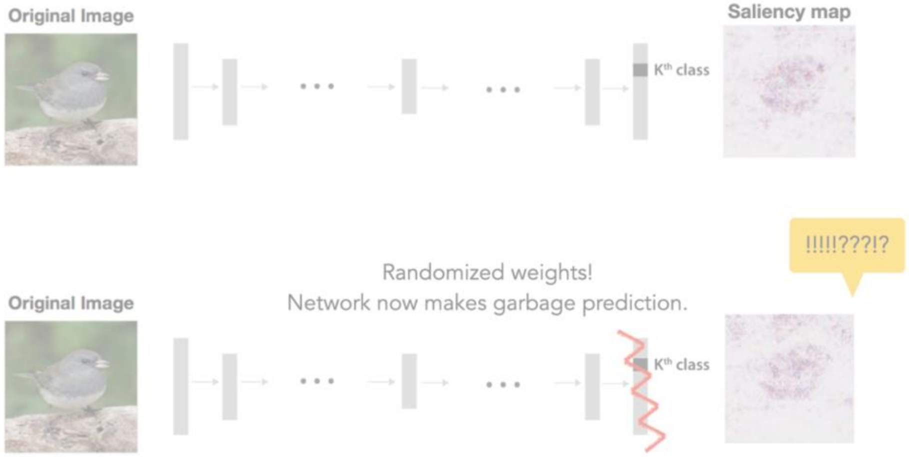
Post-hoc explanations Interpretable by design

Pros:

- Does not alter the Black box

Cons:

- Inconsistent explanations
- No recourse



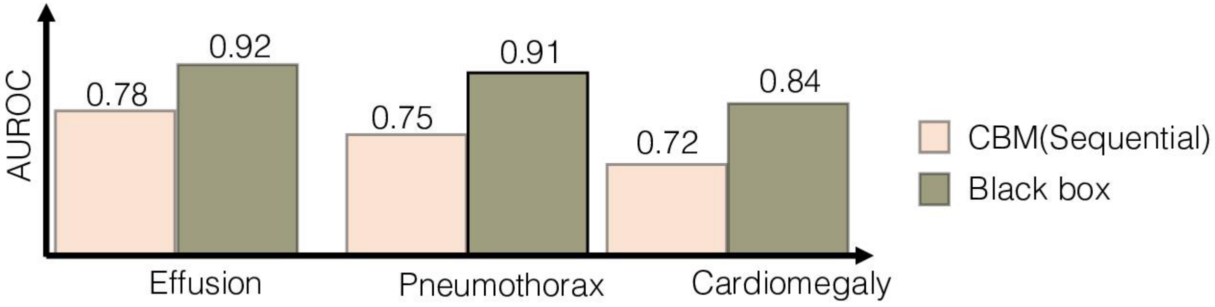
Stop explaining Black Box. Rudin et al. arXiv 2019.
Sanity Checks for Saliency Maps. Adebayo et al. Neurips 2018.

Pros:

- Support concept intervention

Cons:

- Harder to train
- Sub par performance



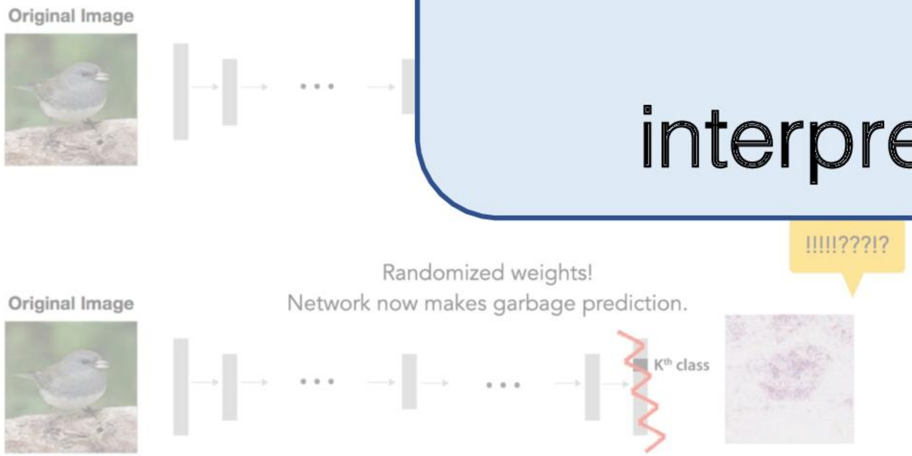
Post-hoc explanations Interpretable by design

Pros:

- Does not alter the Black box

Cons:

- Inconsistent
- No recourse

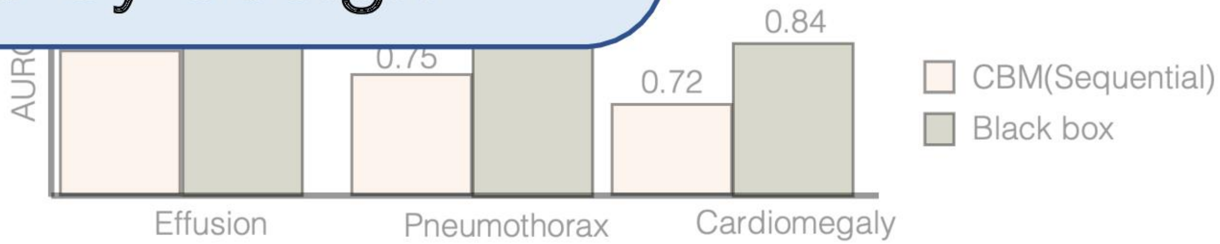


Pros:

- 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.

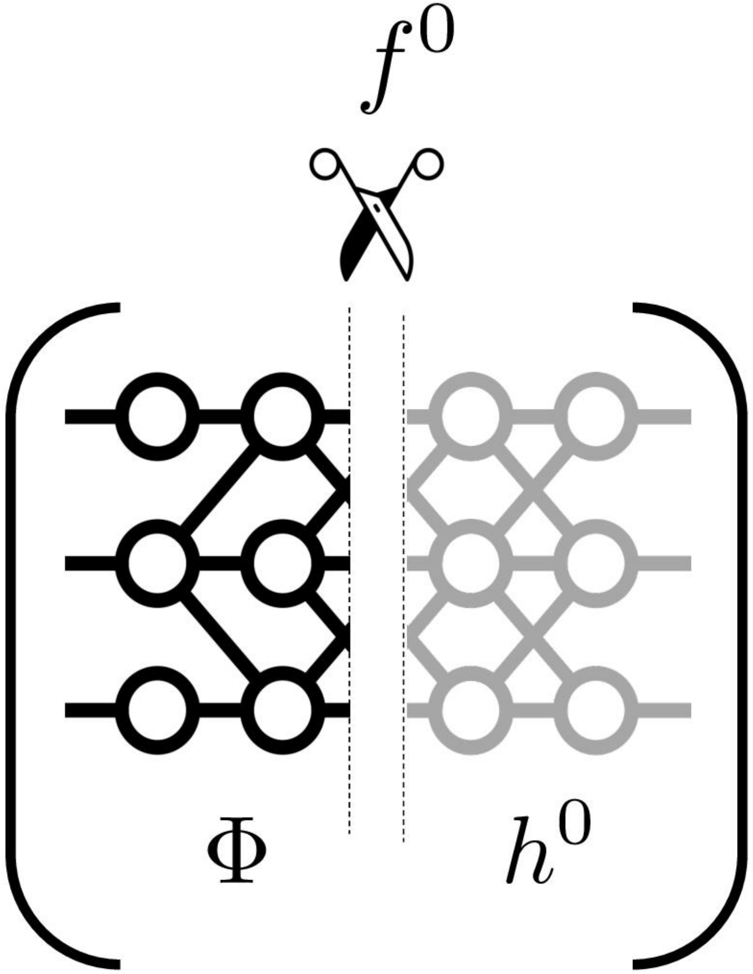
Desirable properties

1. Does not compromise the performance
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



$\mathcal{X}$

$\mathcal{C}$

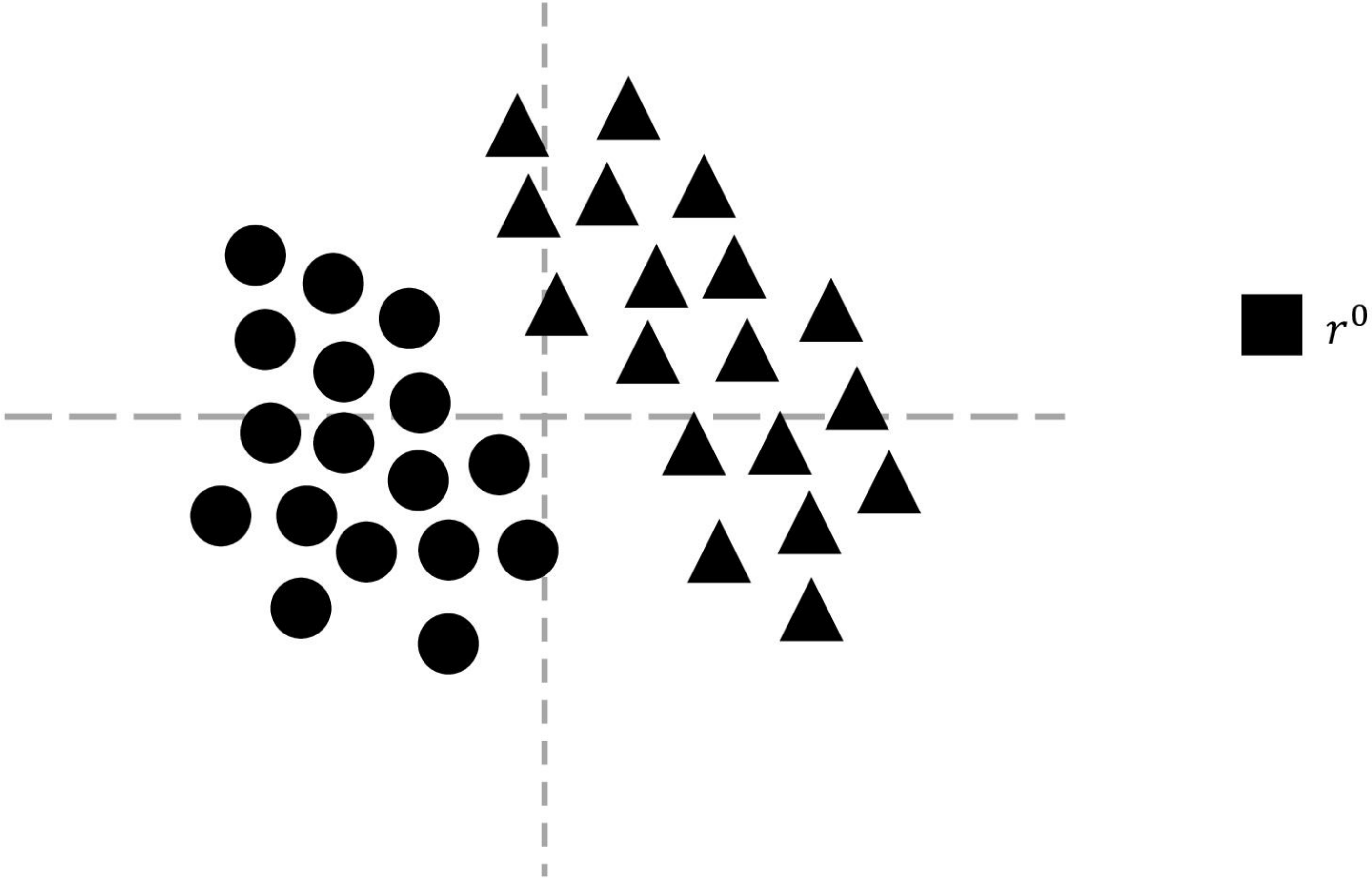
thin ears
shortened muzzle
round feet
....

$\mathcal{Y}$

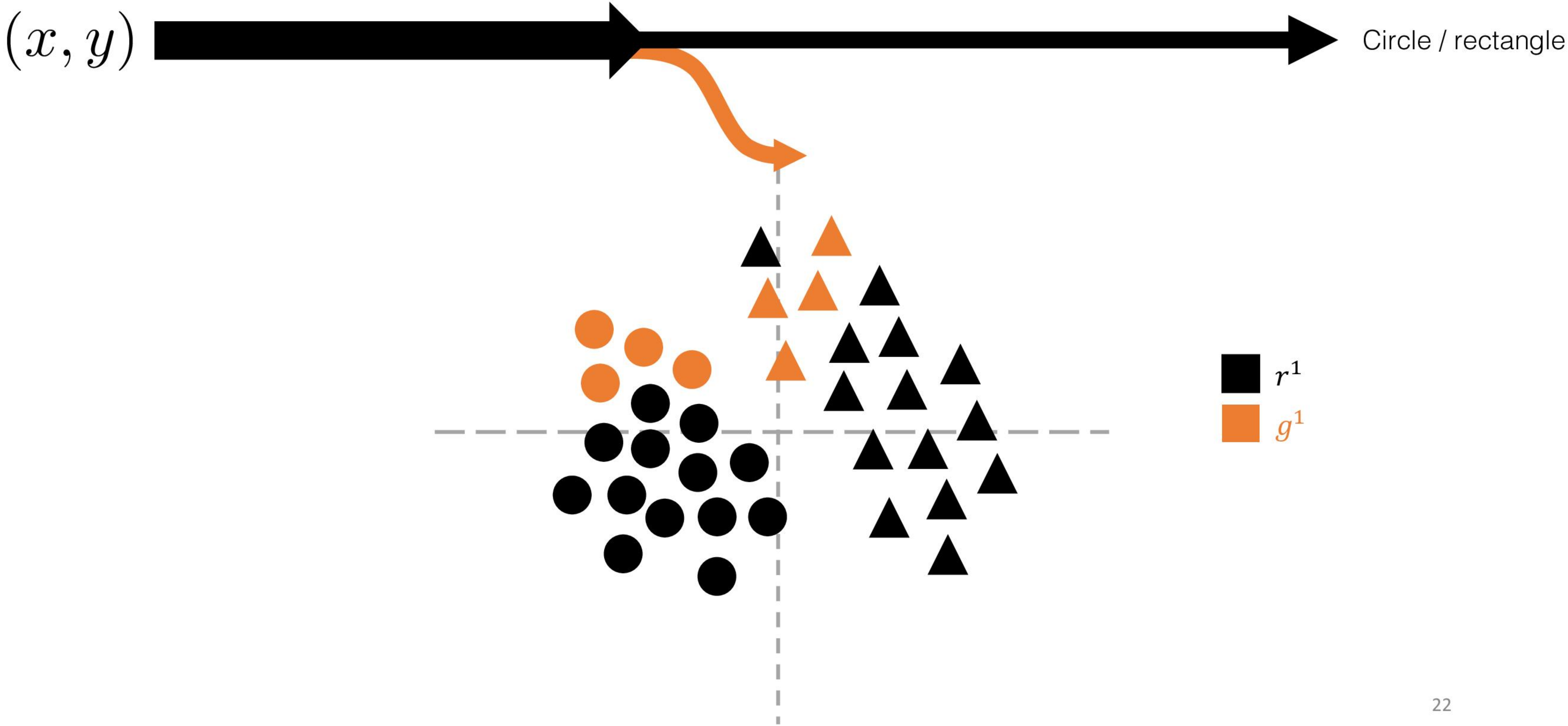
Dog

Iteratively carve out interpretable models

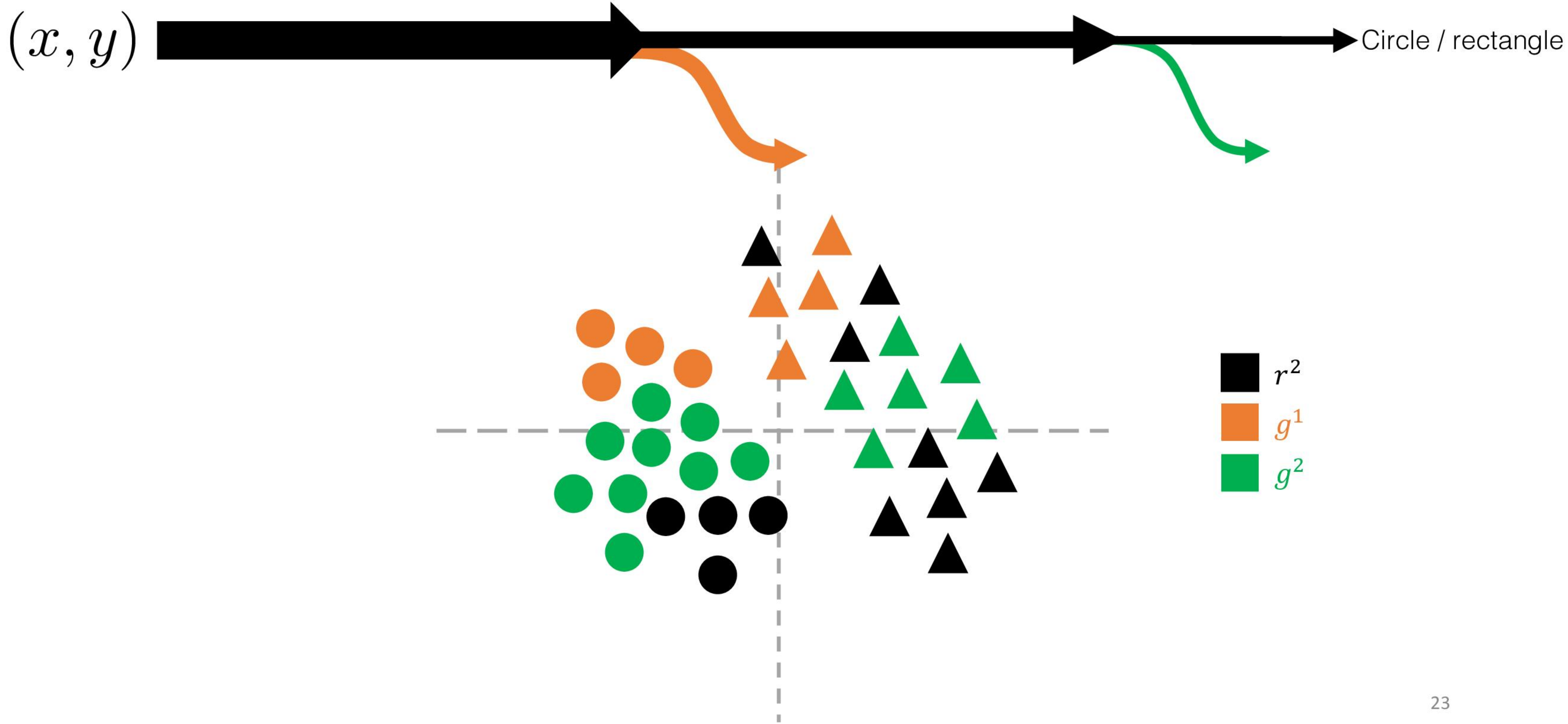
(x, y) → Circle / rectangle



Iteratively carve out interpretable models



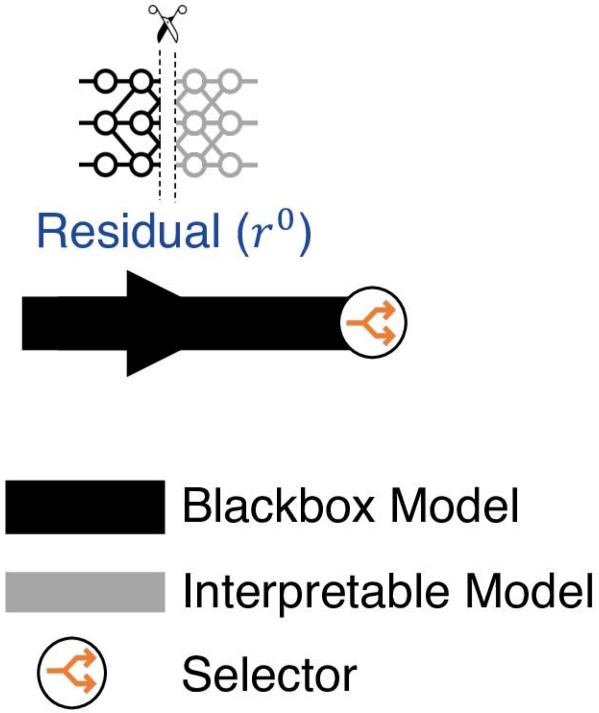
Iteratively carve out interpretable models



(x, y)

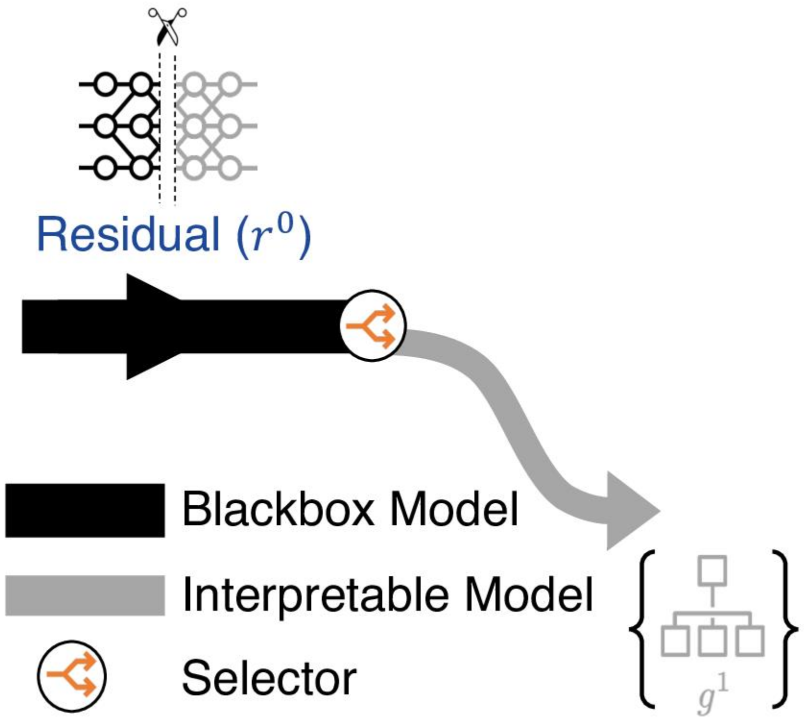
Circle / rectangle

Iteratively carve out interpretable models



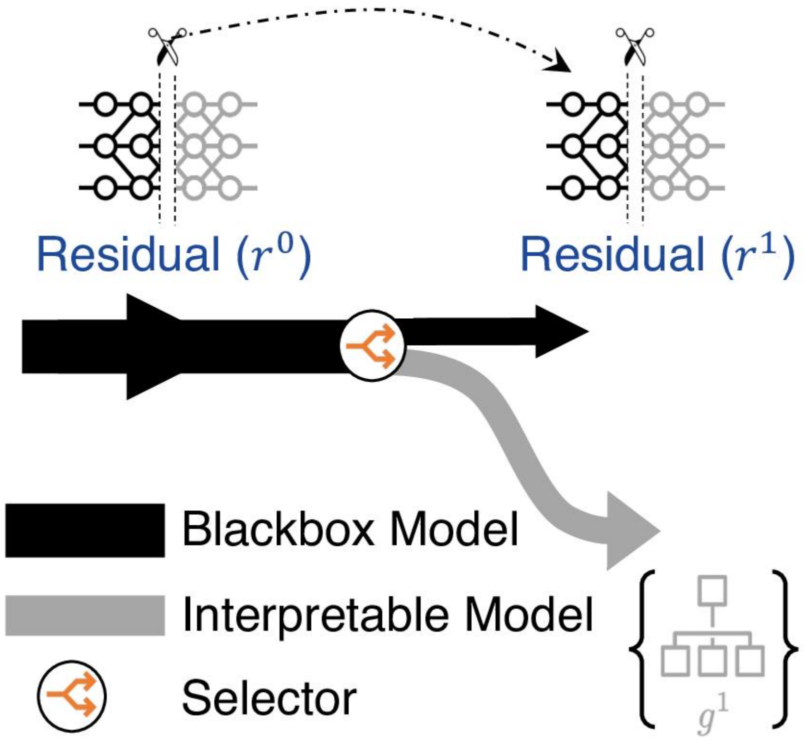
Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

Iteratively carve out interpretable models



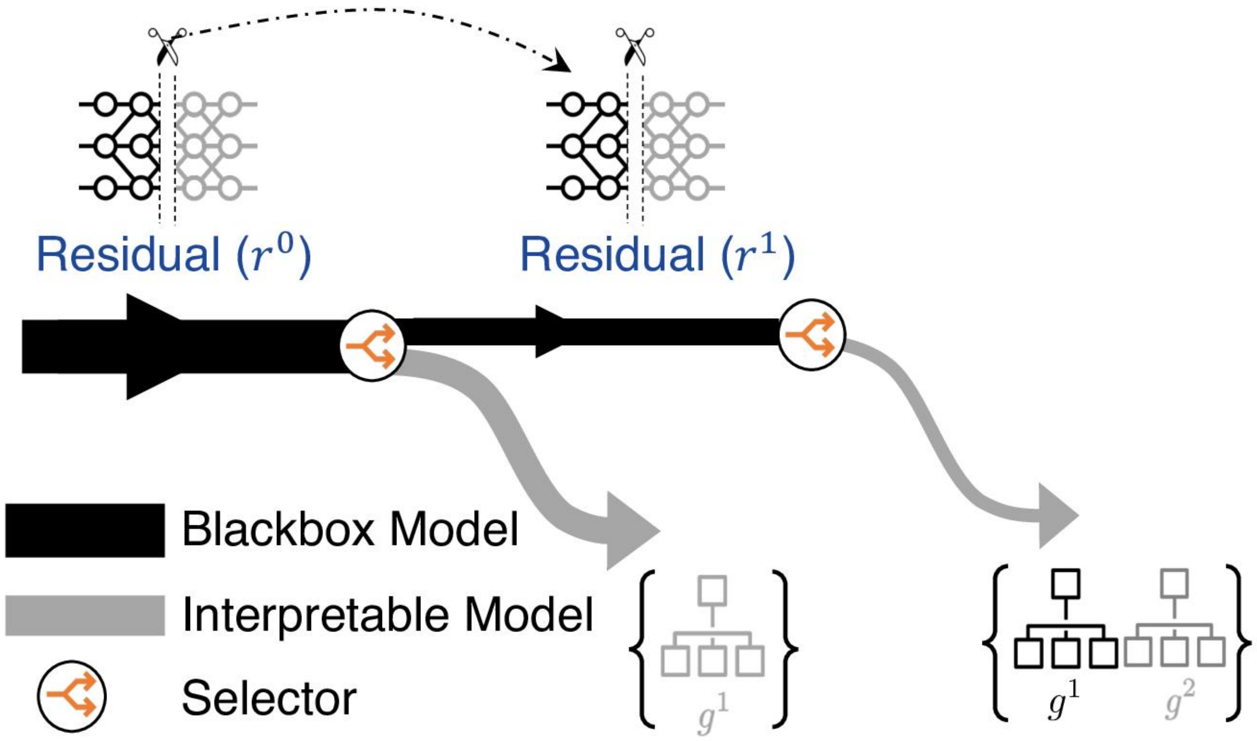
Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

Iteratively carve out interpretable models



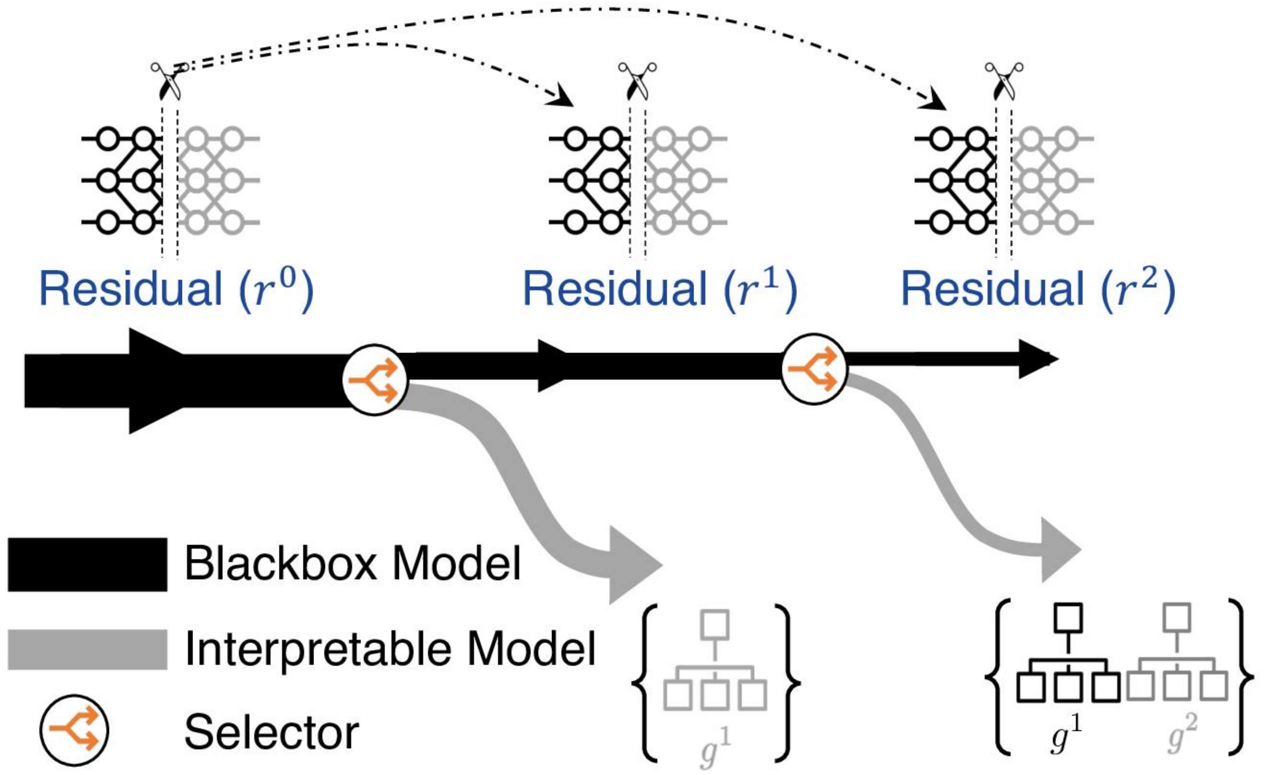
Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

Iteratively carve out interpretable models



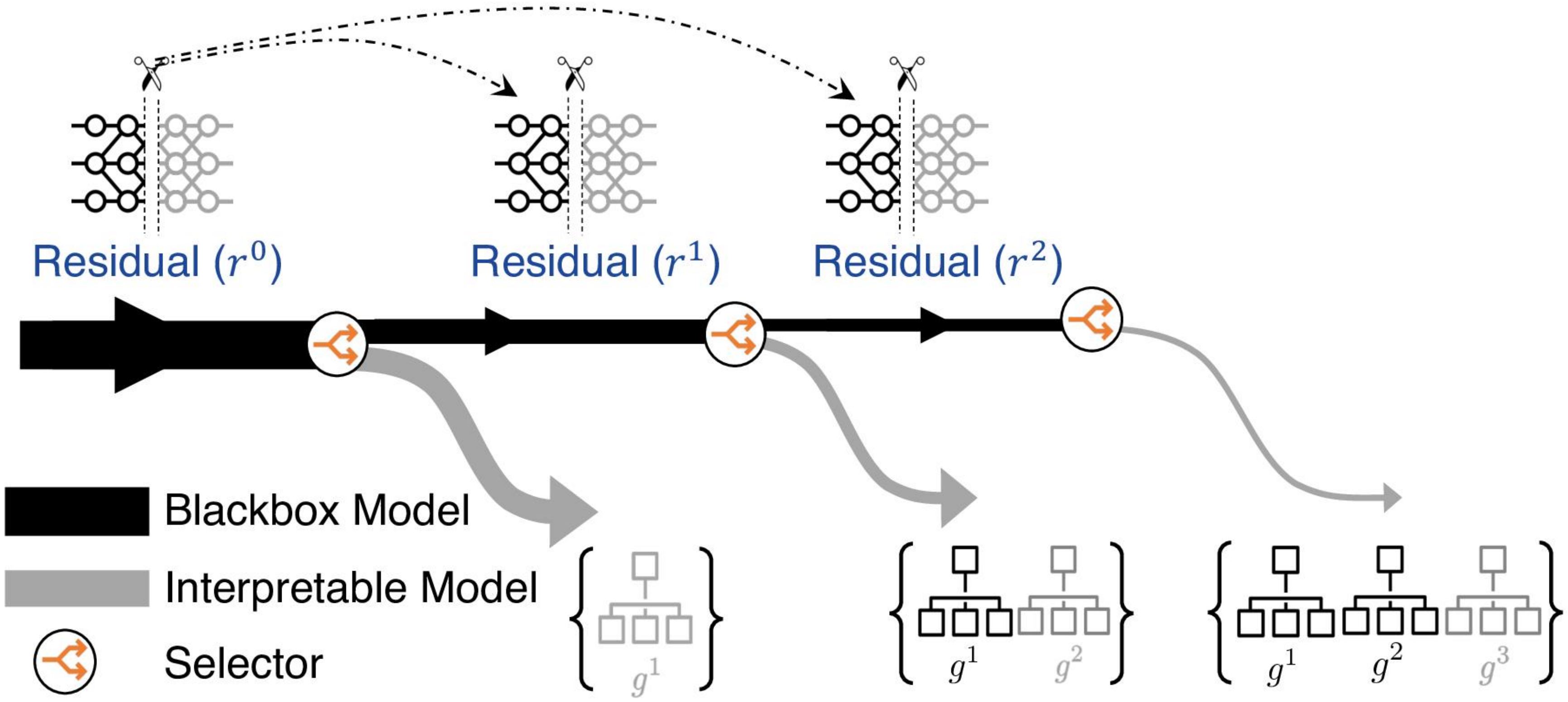
Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

Iteratively carve out interpretable models



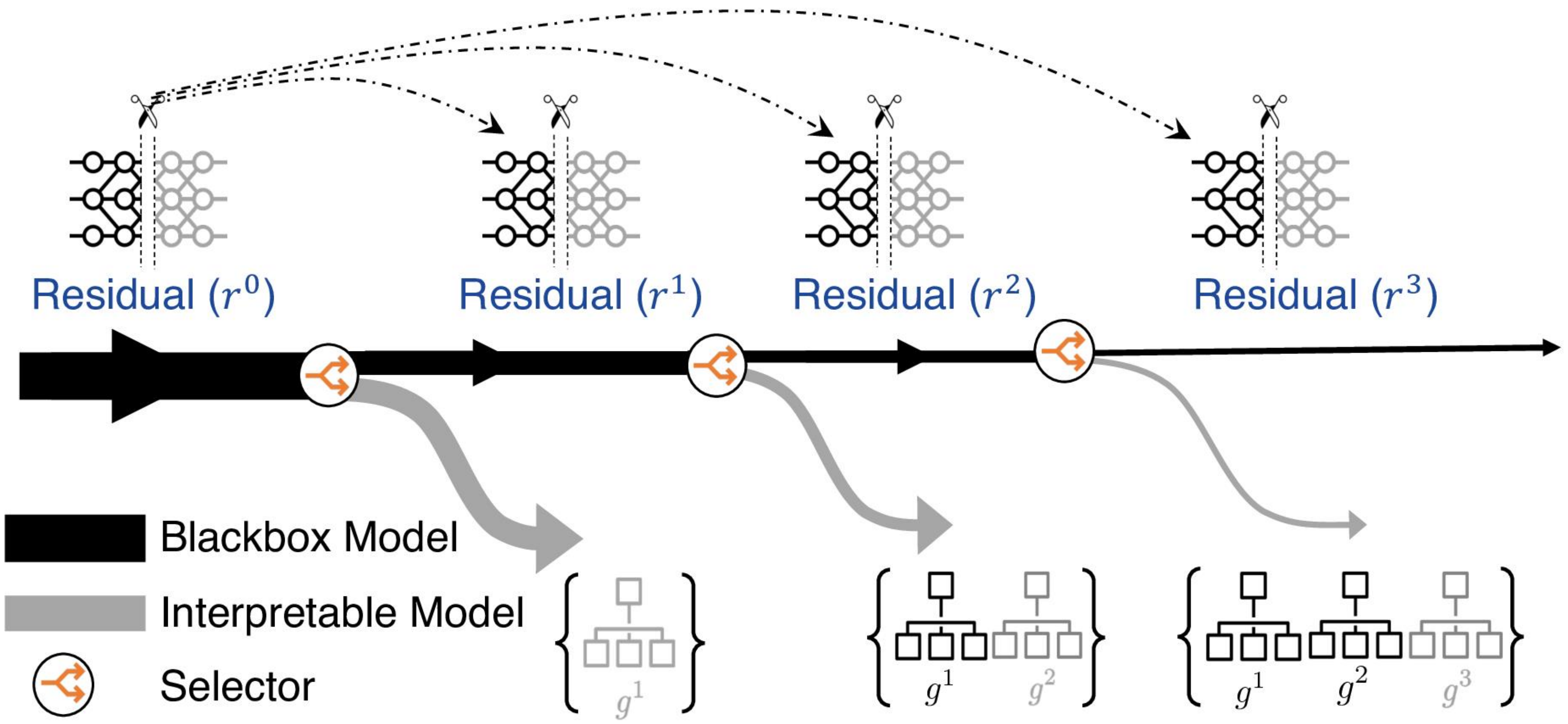
Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

Iteratively carve out interpretable models



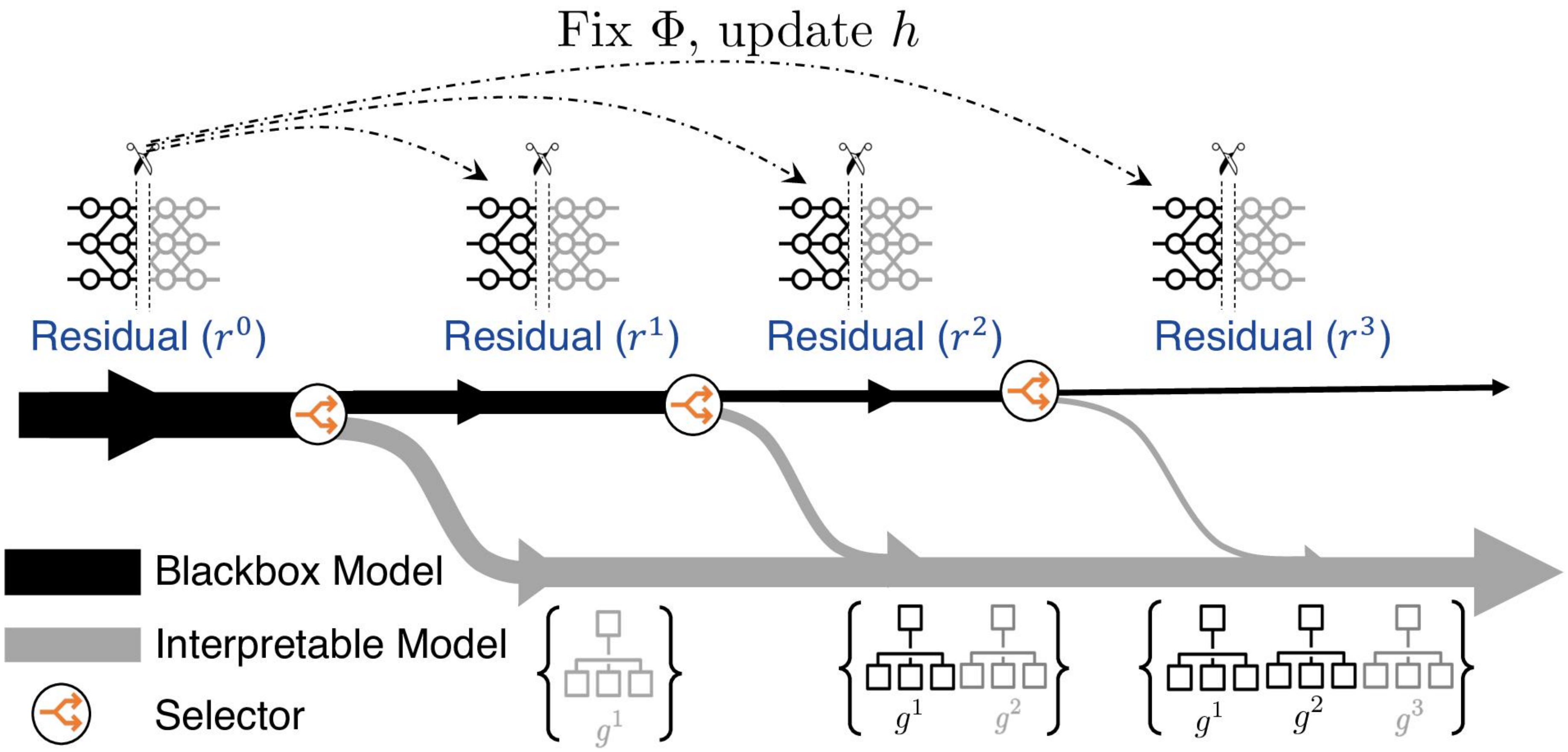
Each g to produce sample specific FOLs (Barberio et al. AAAI 2022) .

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) .