

Right for the Right Reasons:





Avoiding Reasoning Shortcuts via Prototypical Neurosymbolic AI

Luca Andolfi and Eleonora Giunchiglia

University of Rome "La Sapienza" andolfi@diag.uniroma1.it and
Imperial College London e.giunchiglia@imperial.ac.uk

NeurIPS 2025 The Thirty-Ninth Annual Conference on
Neural Information Processing Systems

Reasoning Shortcuts affect Neurosymbolic models.

- **Reasoning Shortcuts** (RS) are spurious associations of unsupervised concepts satisfying constraints expressed as their aggregation.
- *MNIST-Addition*: Consider a dataset comprising the digits pairs ( ) and ( ), each labelled with their sum. The mapping


$$(\textcolor{teal}{0}, \textcolor{teal}{6}) \rightarrow (\textcolor{red}{3}, \textcolor{red}{3}), (\textcolor{teal}{2}, \textcolor{teal}{8}) \rightarrow (\textcolor{red}{4}, \textcolor{red}{4})$$

encodes a RS.

Prototypical Neurosymbolic Models


( Shark,
Fish)

( Dog,
Mammal)

( Duck
Bird)

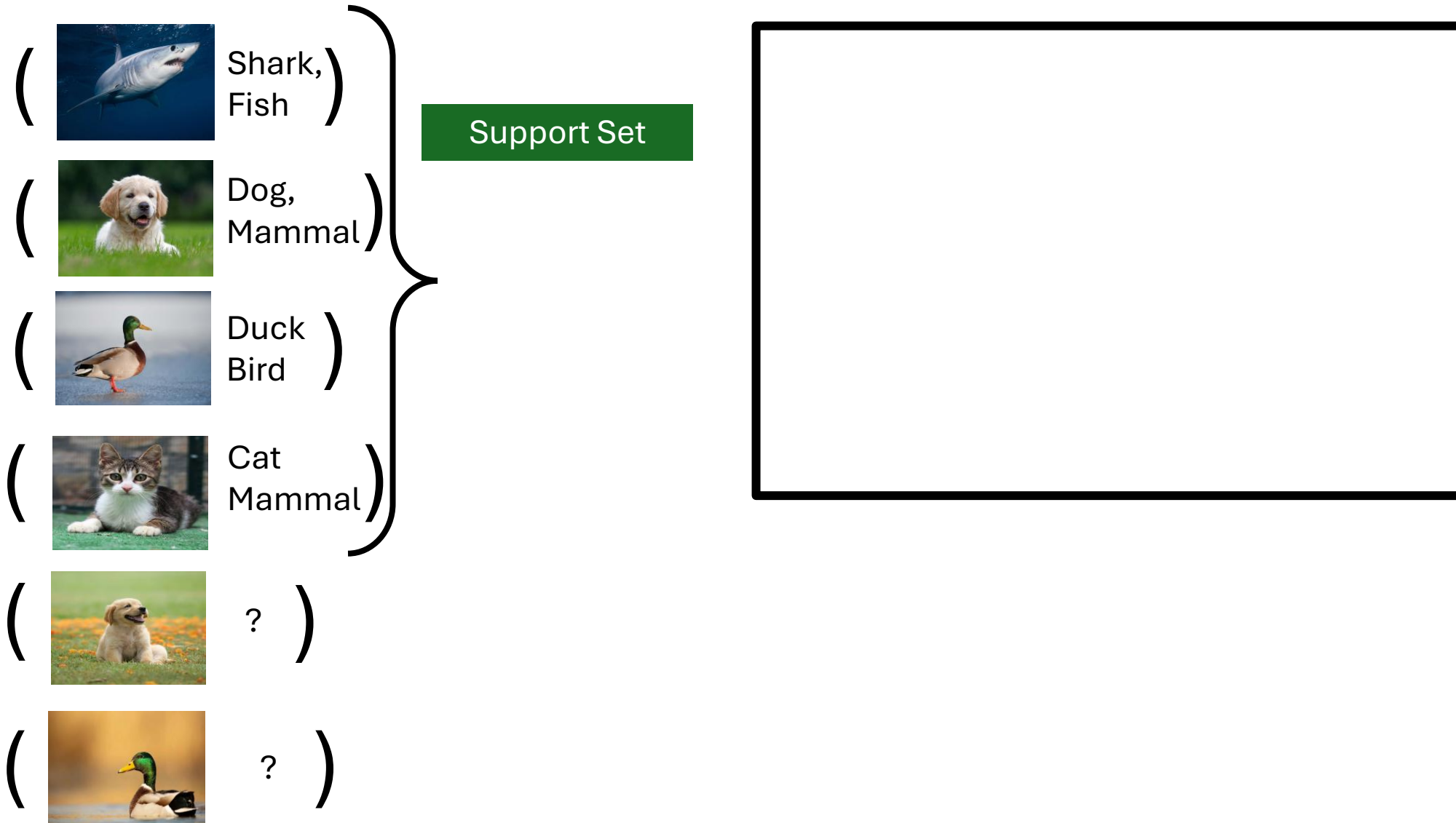
( Cat
Mammal)

( ?)

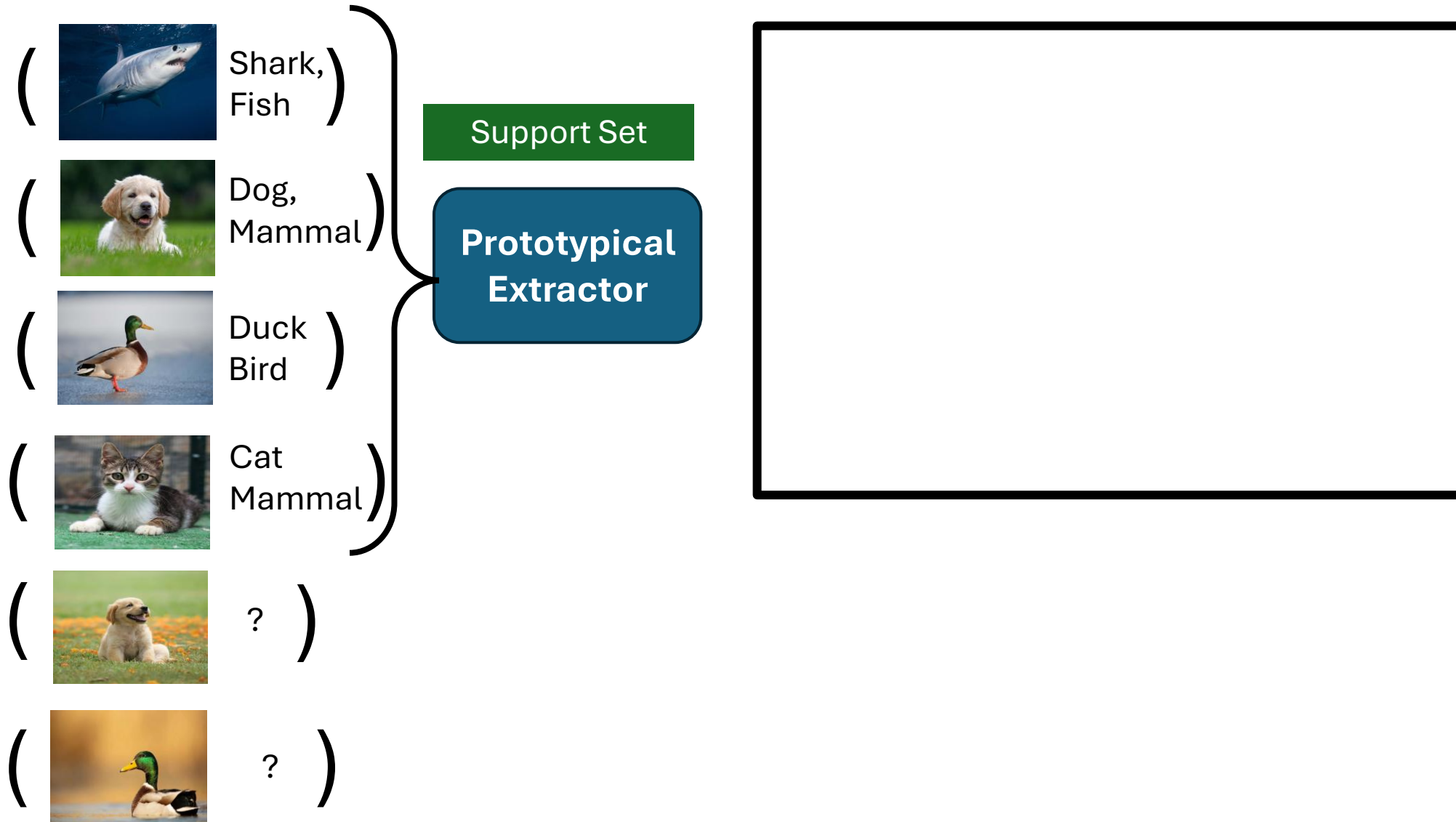
( ?)



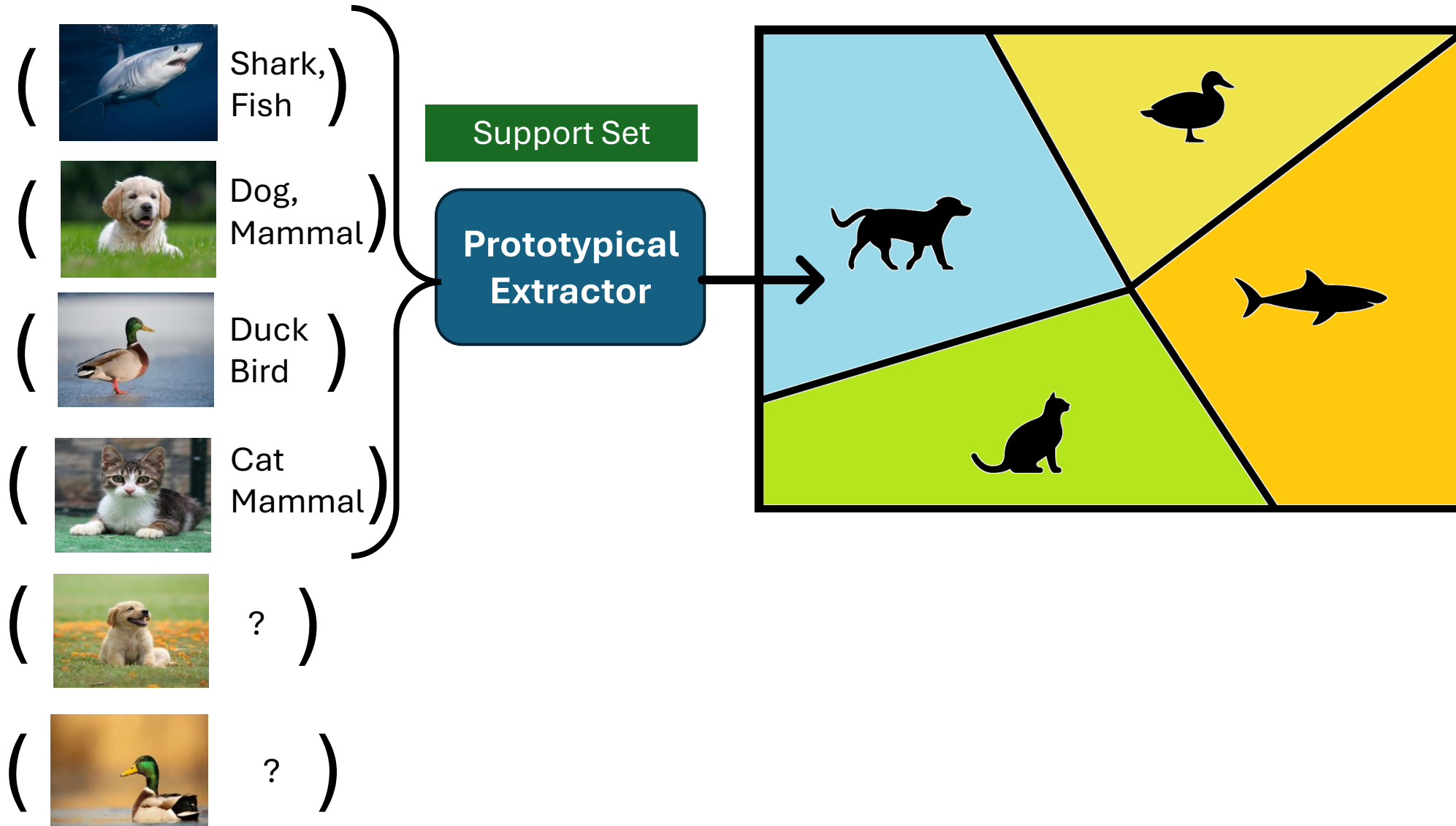
Prototypical Neurosymbolic Models



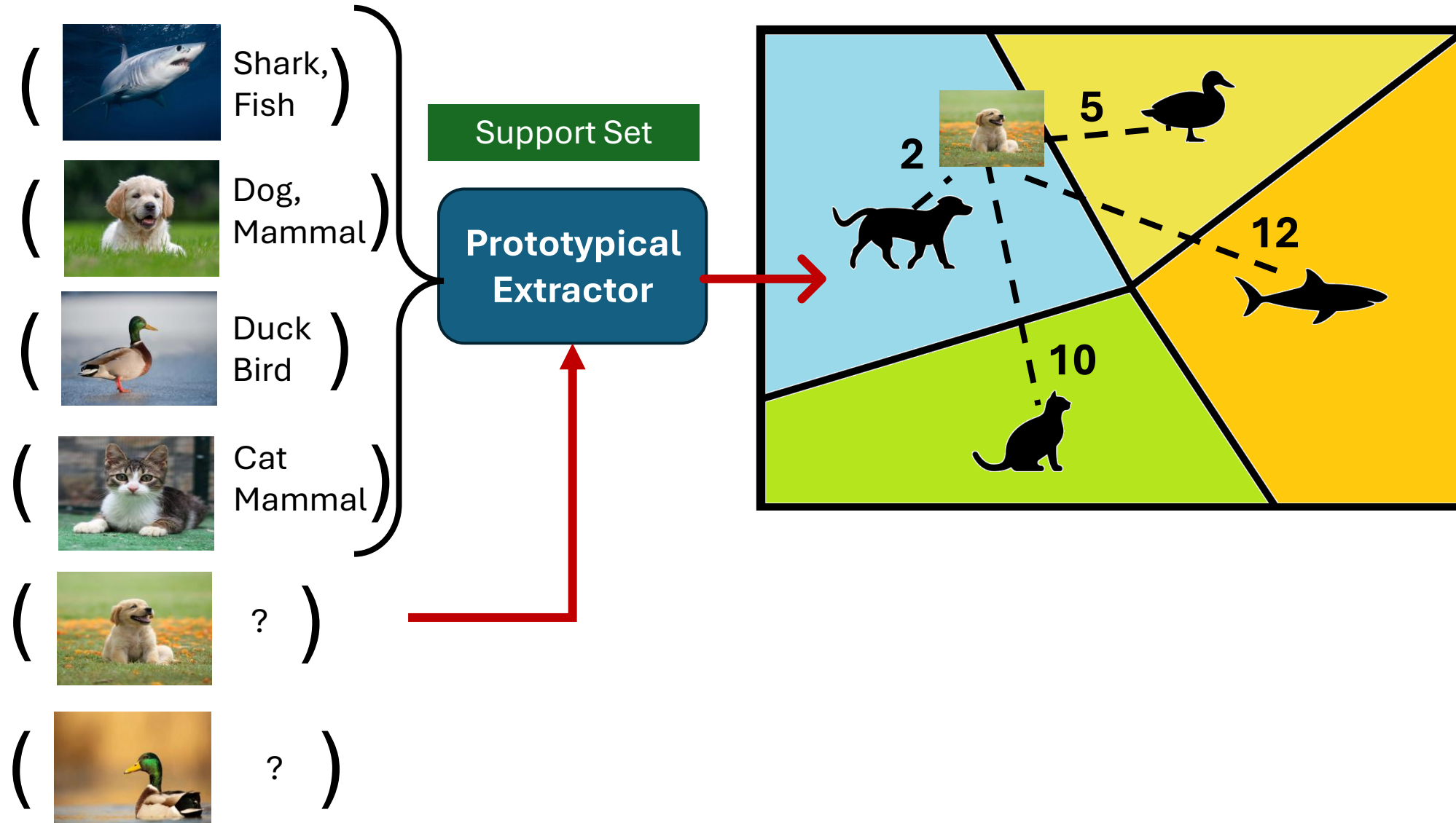
Prototypical Neurosymbolic Models



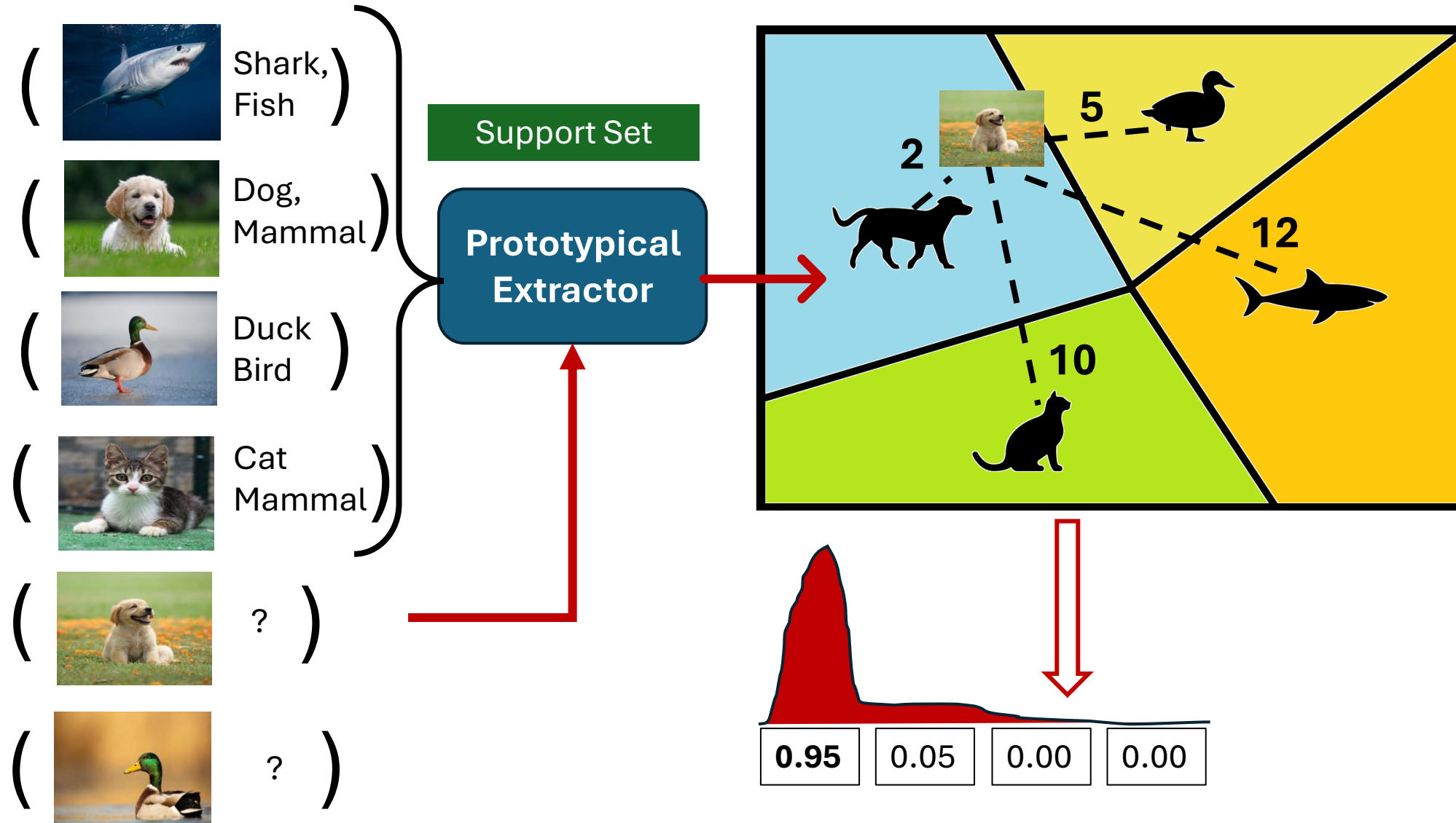
Prototypical Neurosymbolic Models



Prototypical Neurosymbolic Models

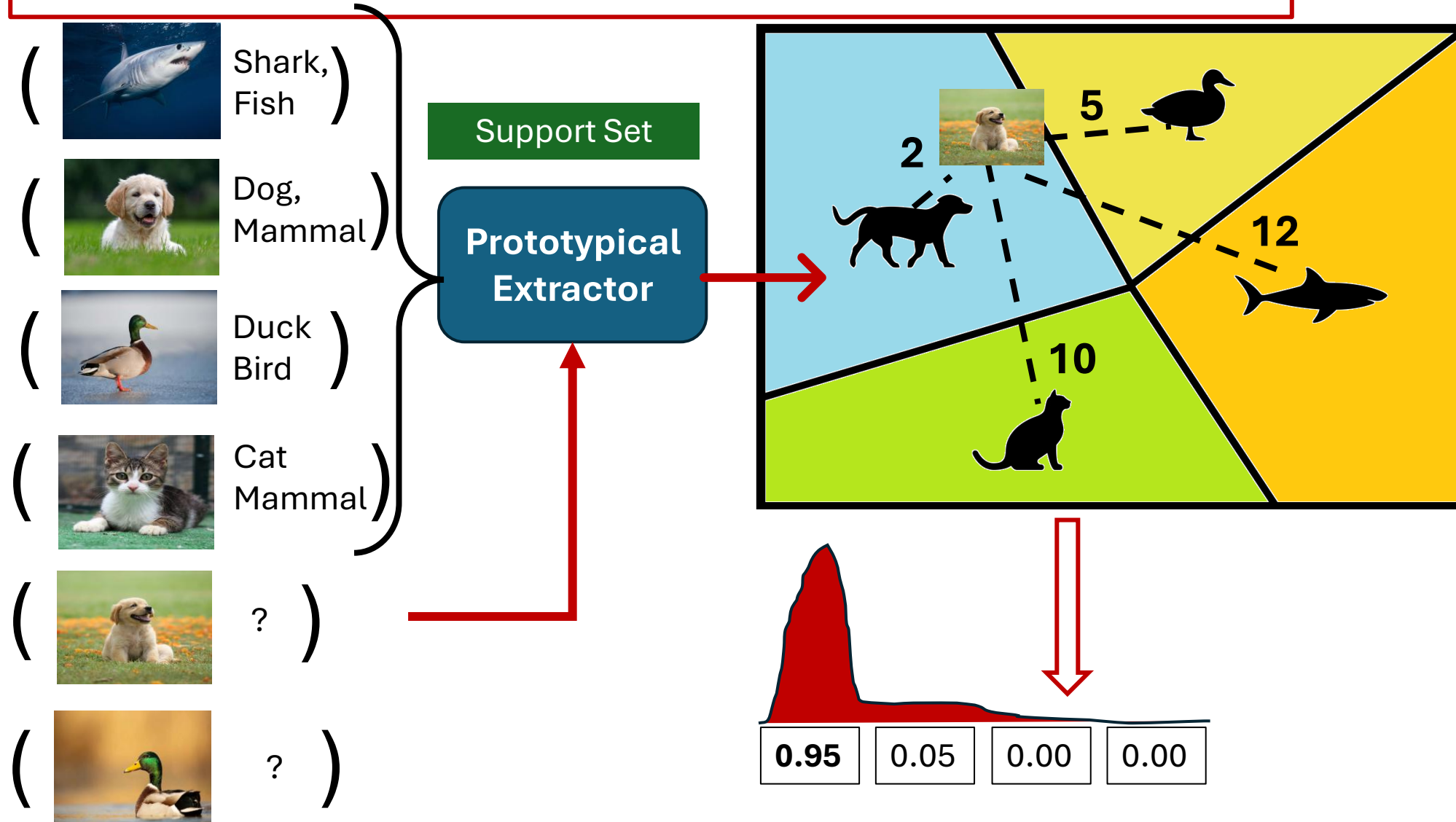


Prototypical Neurosymbolic Models



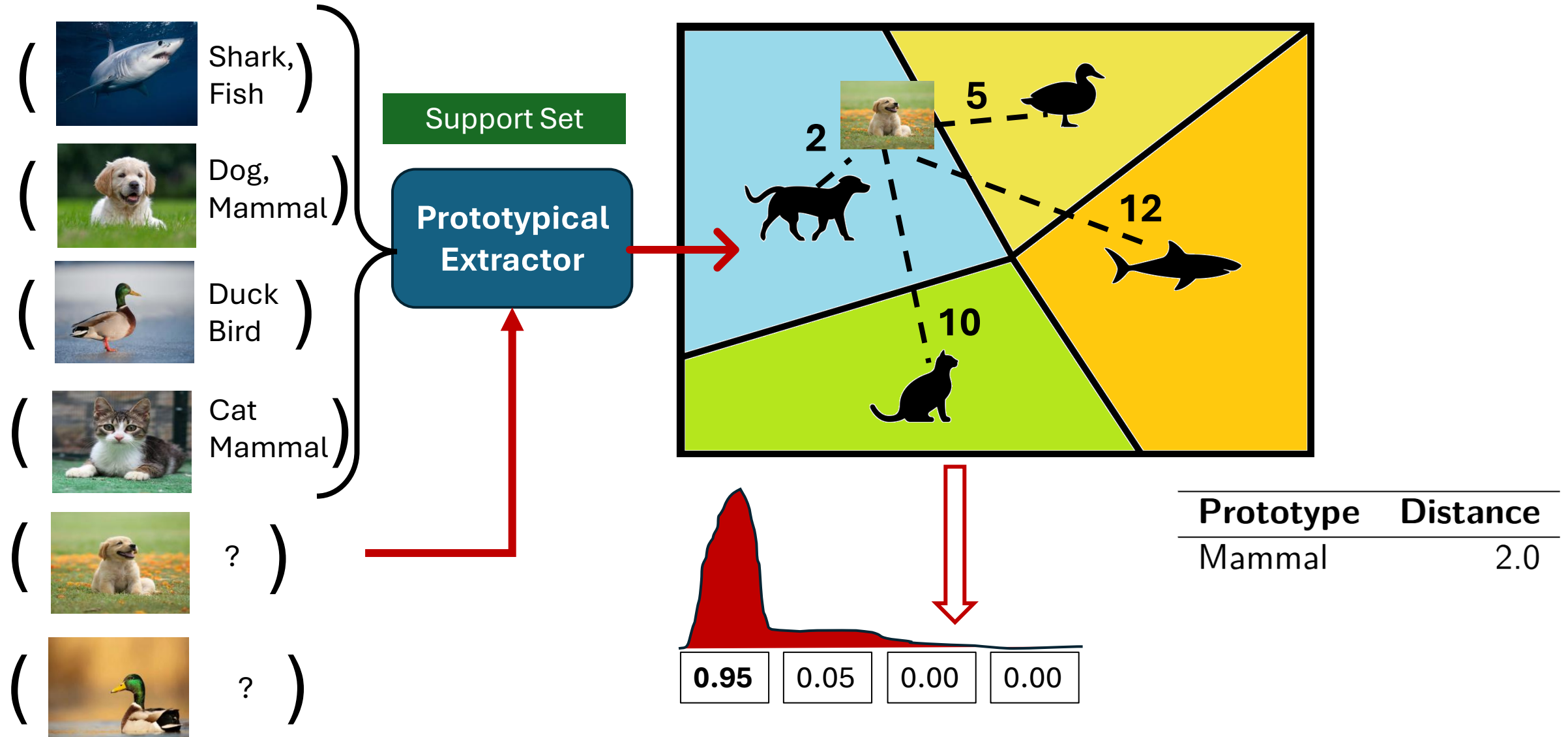
Prototypical Neurosymbolic Models

$$K = (\text{Dog} \vee \text{Cat} \rightarrow \text{Mammal}) \wedge (\text{Duck} \rightarrow \text{Bird}) \wedge (\text{Shark} \rightarrow \text{Fish})$$



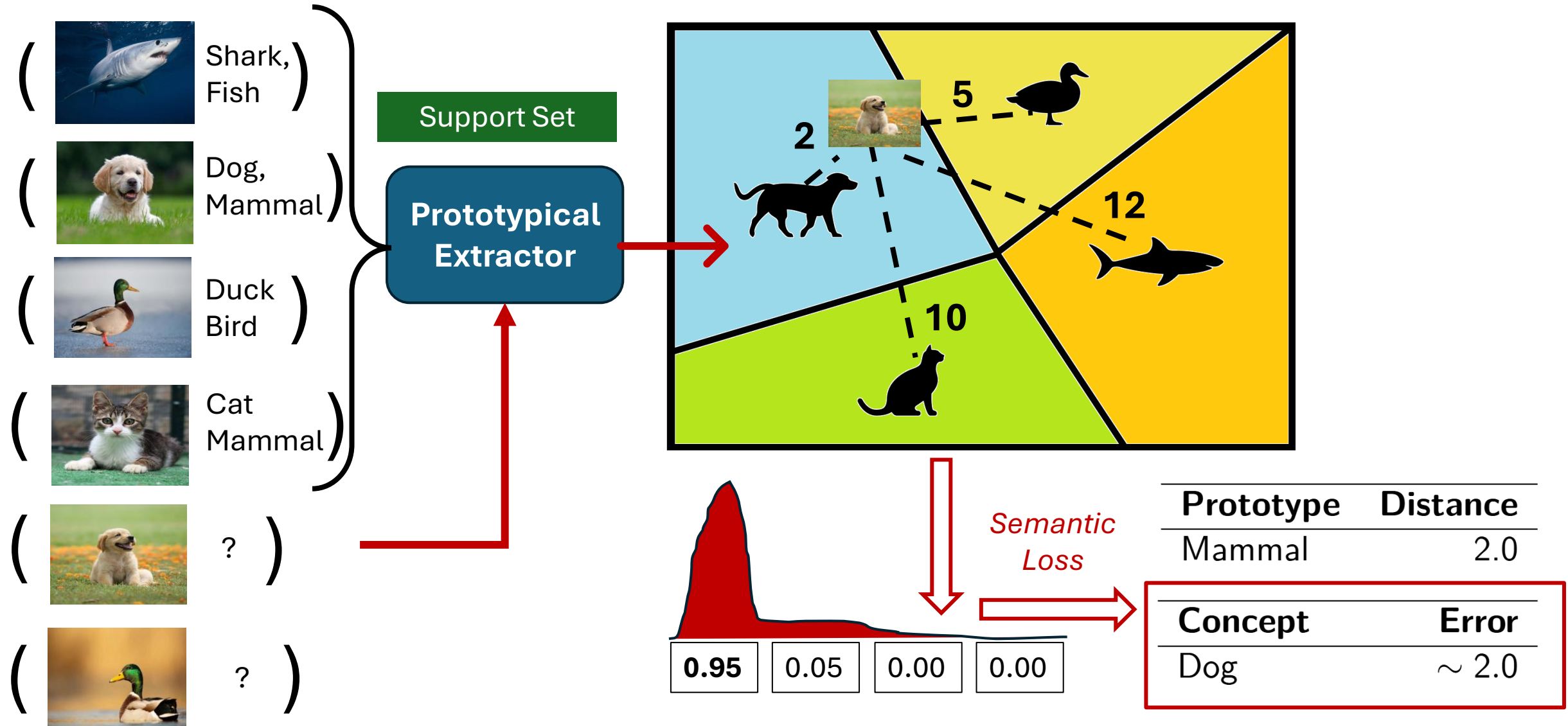
Prototypical Neurosymbolic Models

$$K = (\text{Dog} \vee \text{Cat} \rightarrow \text{Mammal}) \wedge (\text{Duck} \rightarrow \text{Bird}) \wedge (\text{Shark} \rightarrow \text{Fish})$$

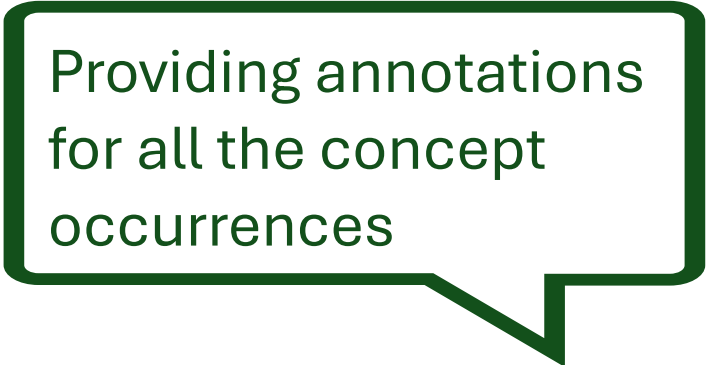


Prototypical Neurosymbolic Models

$$K = (\text{Dog} \vee \text{Cat} \rightarrow \text{Mammal}) \wedge (\text{Duck} \rightarrow \text{Bird}) \wedge (\text{Shark} \rightarrow \text{Fish})$$



Number of RSs in Prototypical Neurosymbolic Models



Providing annotations
for all the concept
occurrences

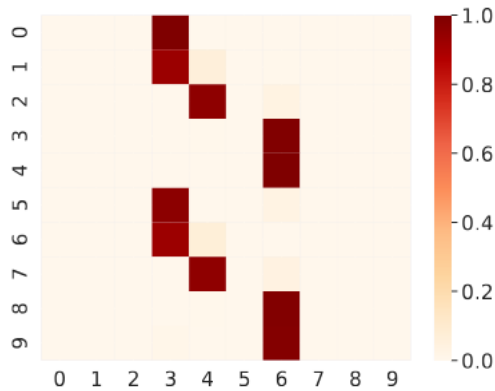
- Same **number** of RSs of **dense** annotation mitigation strategies using **just one** annotation per concept under **separability** in the embedding space.

Experimental Results (1)

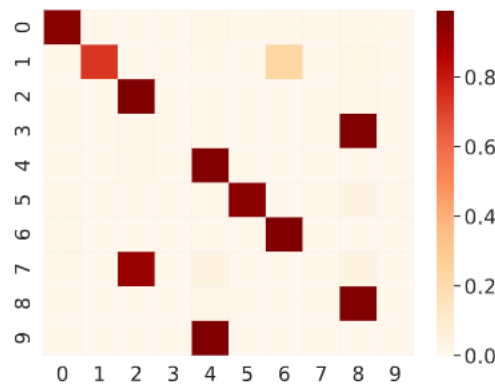
- Improved RS mitigation compared to prior approaches.

Experimental Results (1)

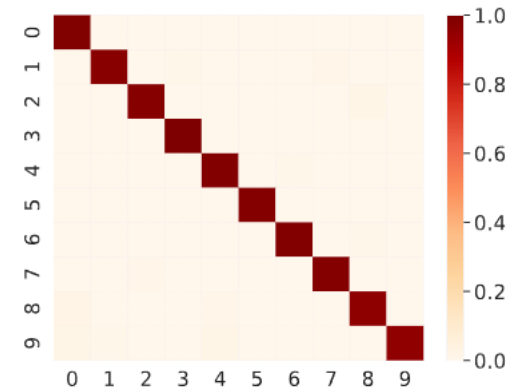
- Improved RS mitigation compared to prior approaches.



(a) SL



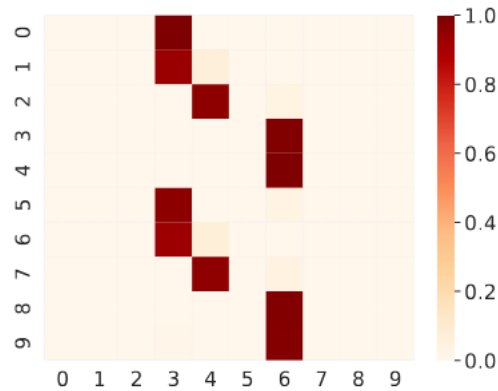
(b) SL+Pre^{*}



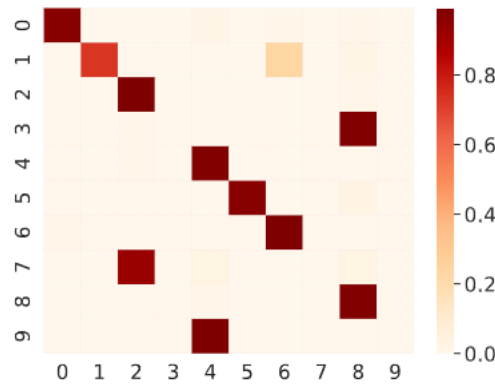
(c) SL+PNet

Experimental Results (1)

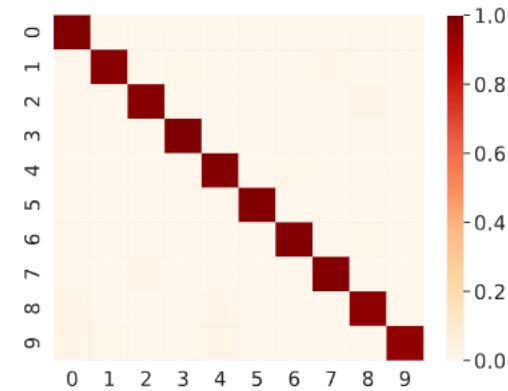
- Improved RS mitigation compared to prior approaches.



(a) SL



(b) SL+Pre^{*}



(c) SL+PNet

- Strong performance even with just few concept annotations.

Experimental Results (2)

- Stable results across varying amounts of unlabelled data.

Experimental Results (2)

- Stable results across varying amounts of unlabelled data.

