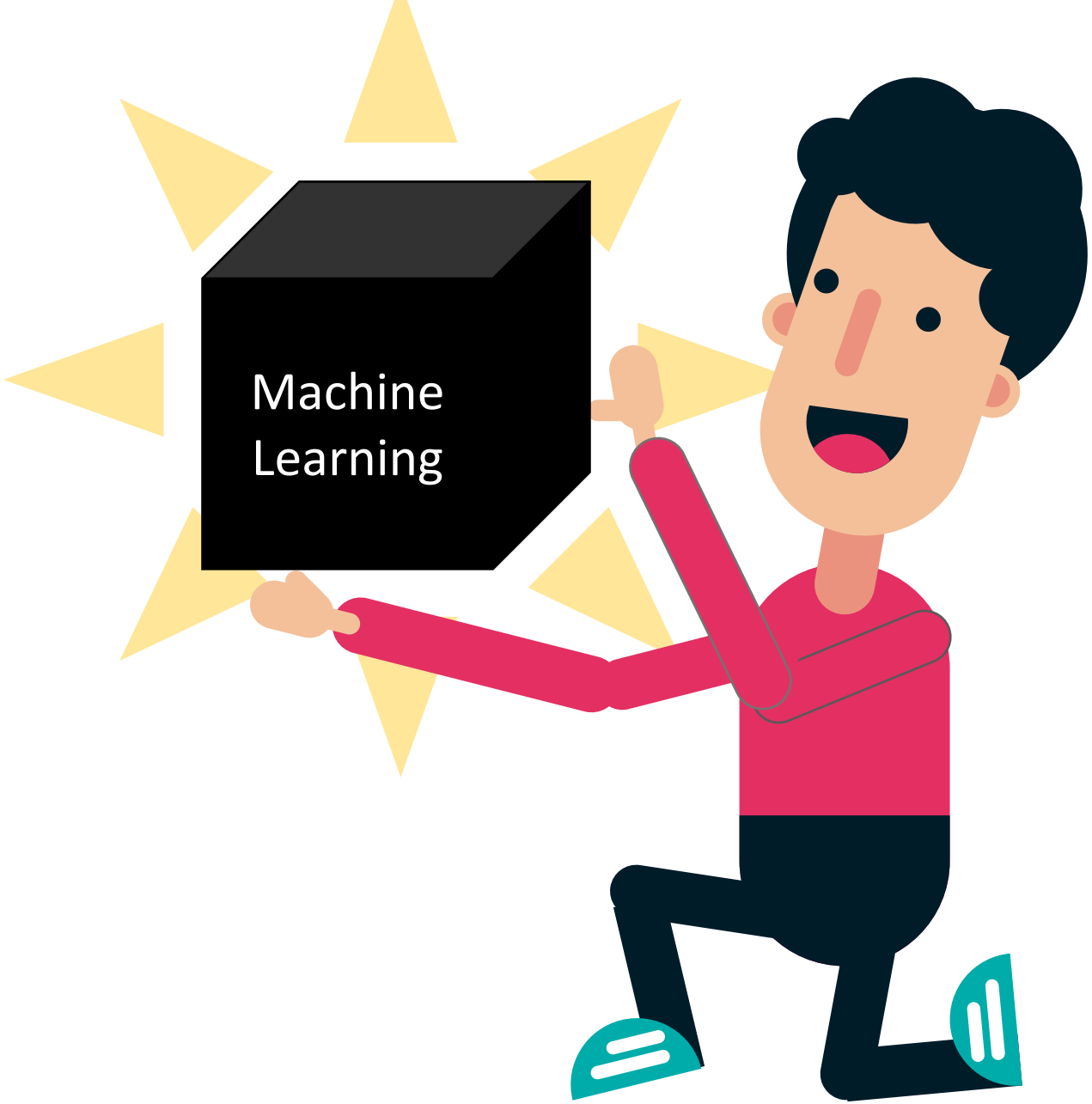
# Visualizing Surrogate Decision Trees of Convolutional Neural Networks

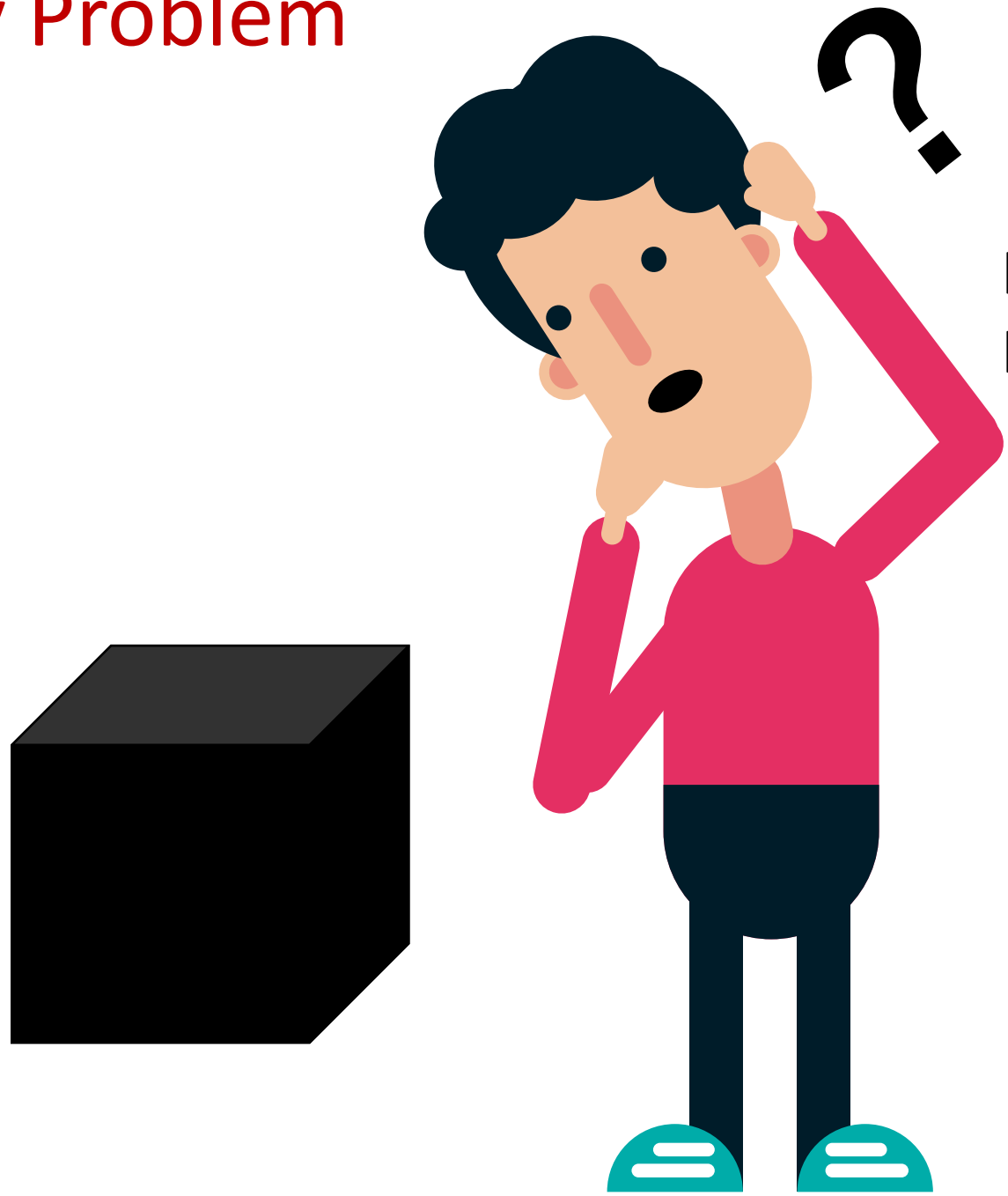
**Shichao Jia**[1], Peiwen Lin[1], Zeyu Li[1], Jiawan Zhang[1] and Shixia Liu[2]

1 


2 

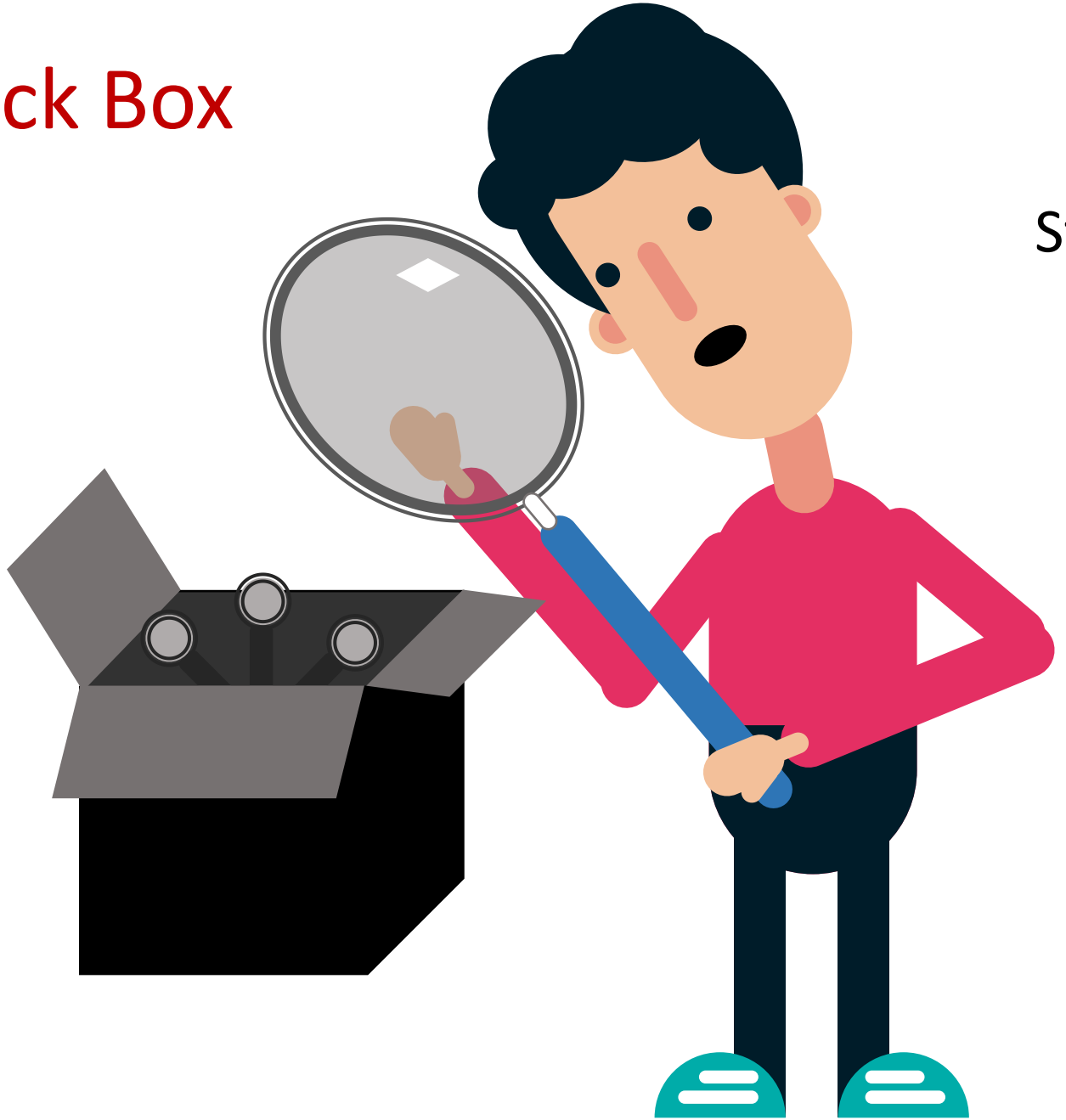

Machine
Learning

# Interpretability Problem

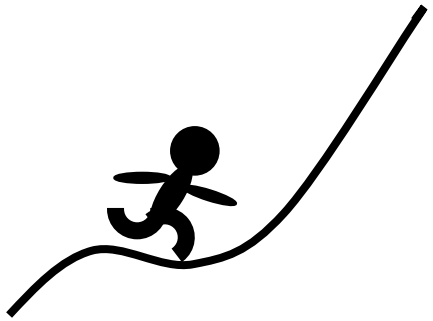


I do not understand !
I can not trust it !

# Strategy 1
# Open the Black Box



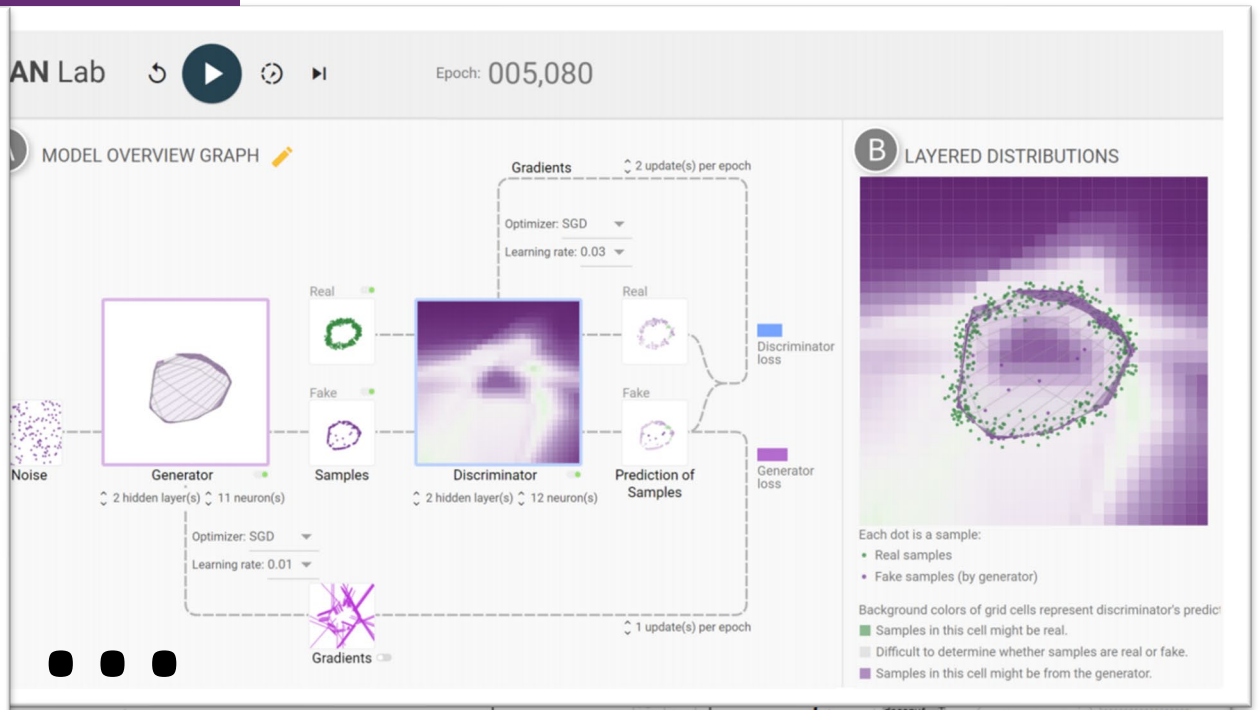
Steep learning curve



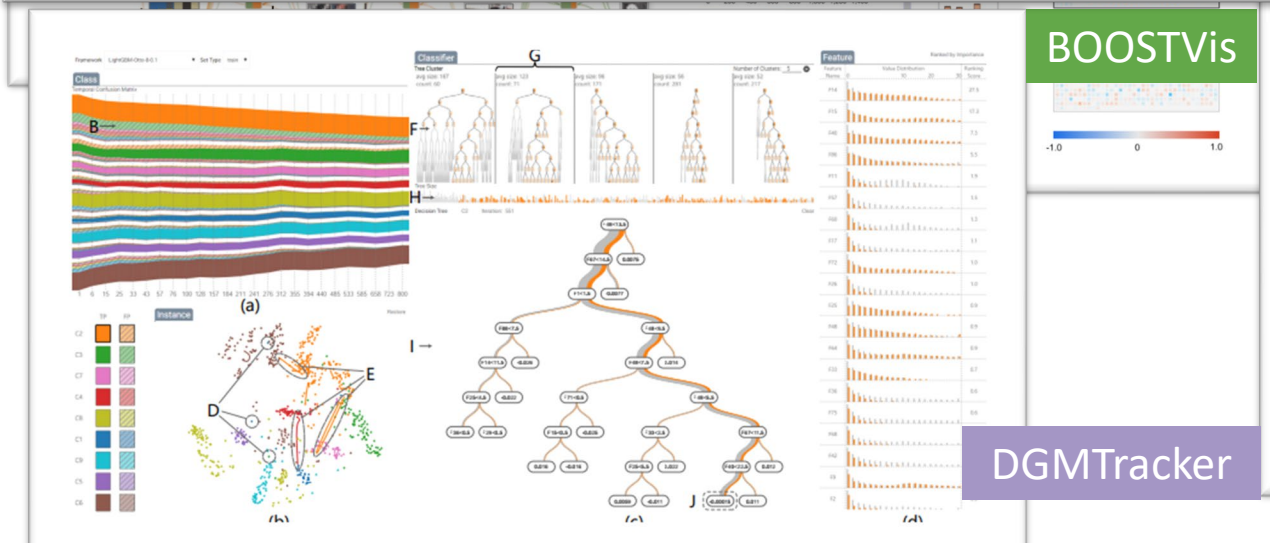
# Related Work

DQNViz

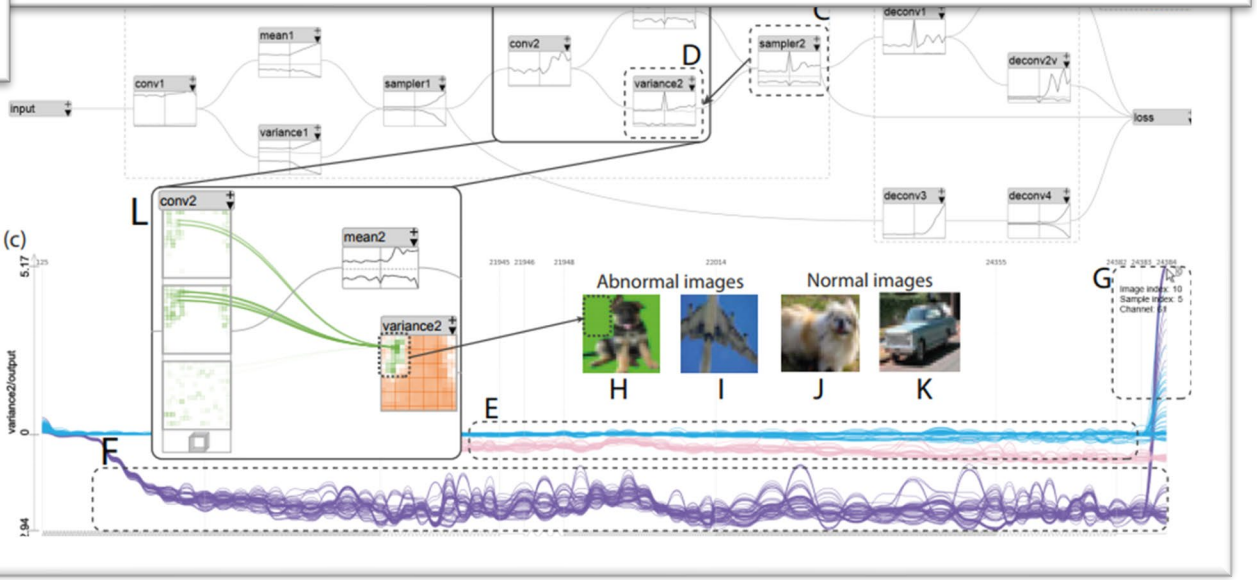

GANLab



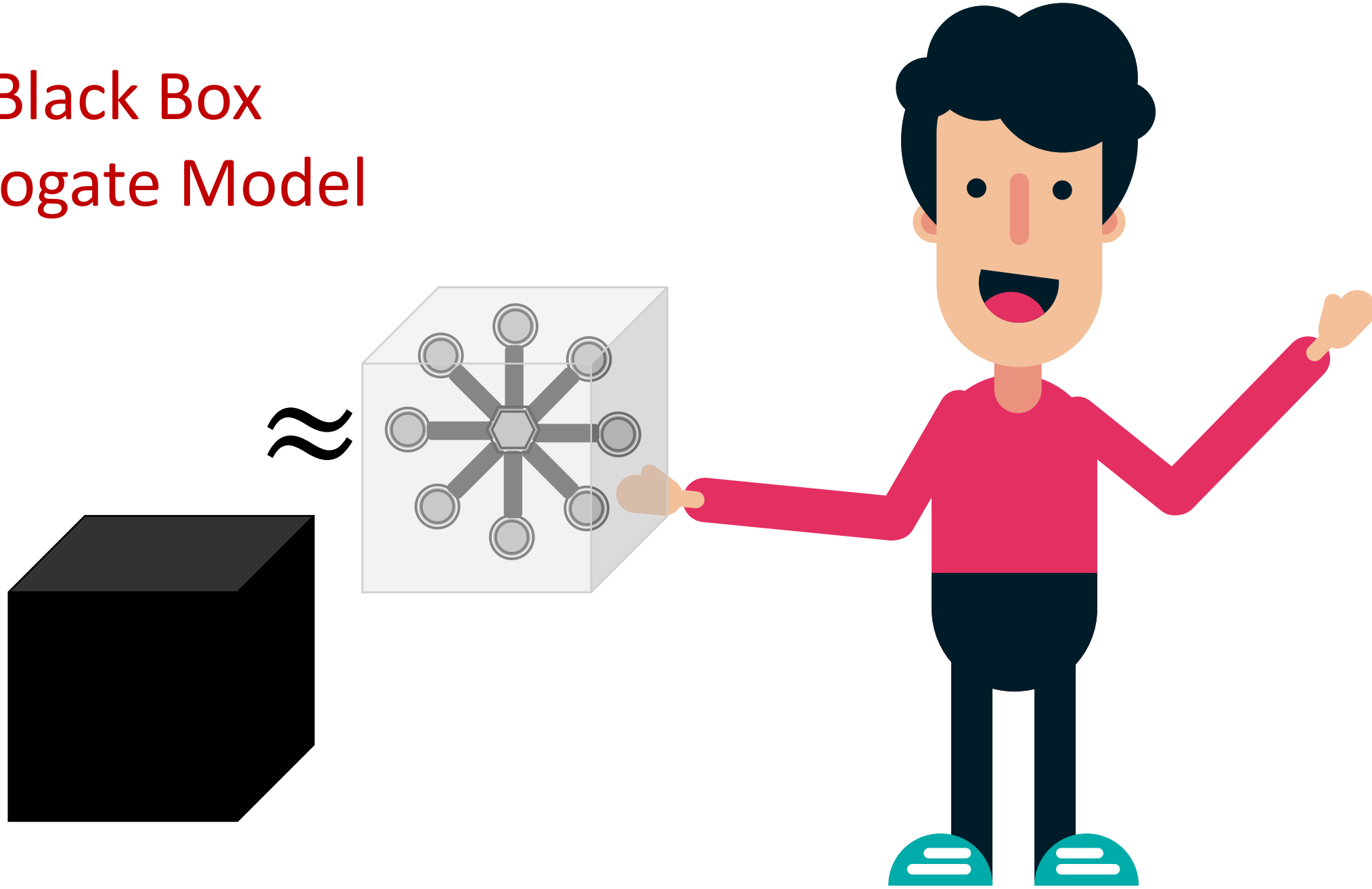

BOOSTVis



DGMTracker

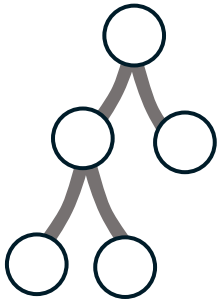


# Strategy 2
# Close the Black Box
# Using Surrogate Model



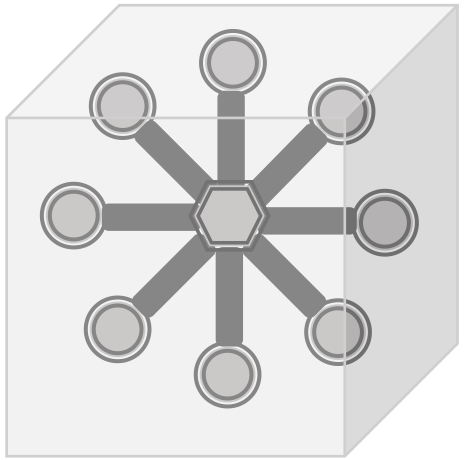
# Strategy 2
# Close the Black Box
# Using Surrogate Model

Such as:

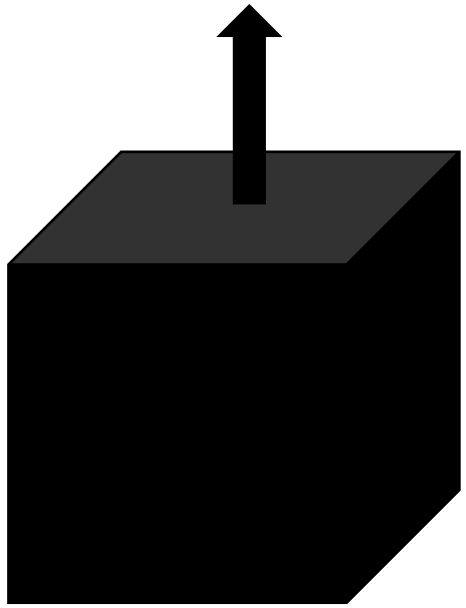


IF A < B Then …
Else IF C > B
Then …

Decision Tree, Rule List …



Same Input

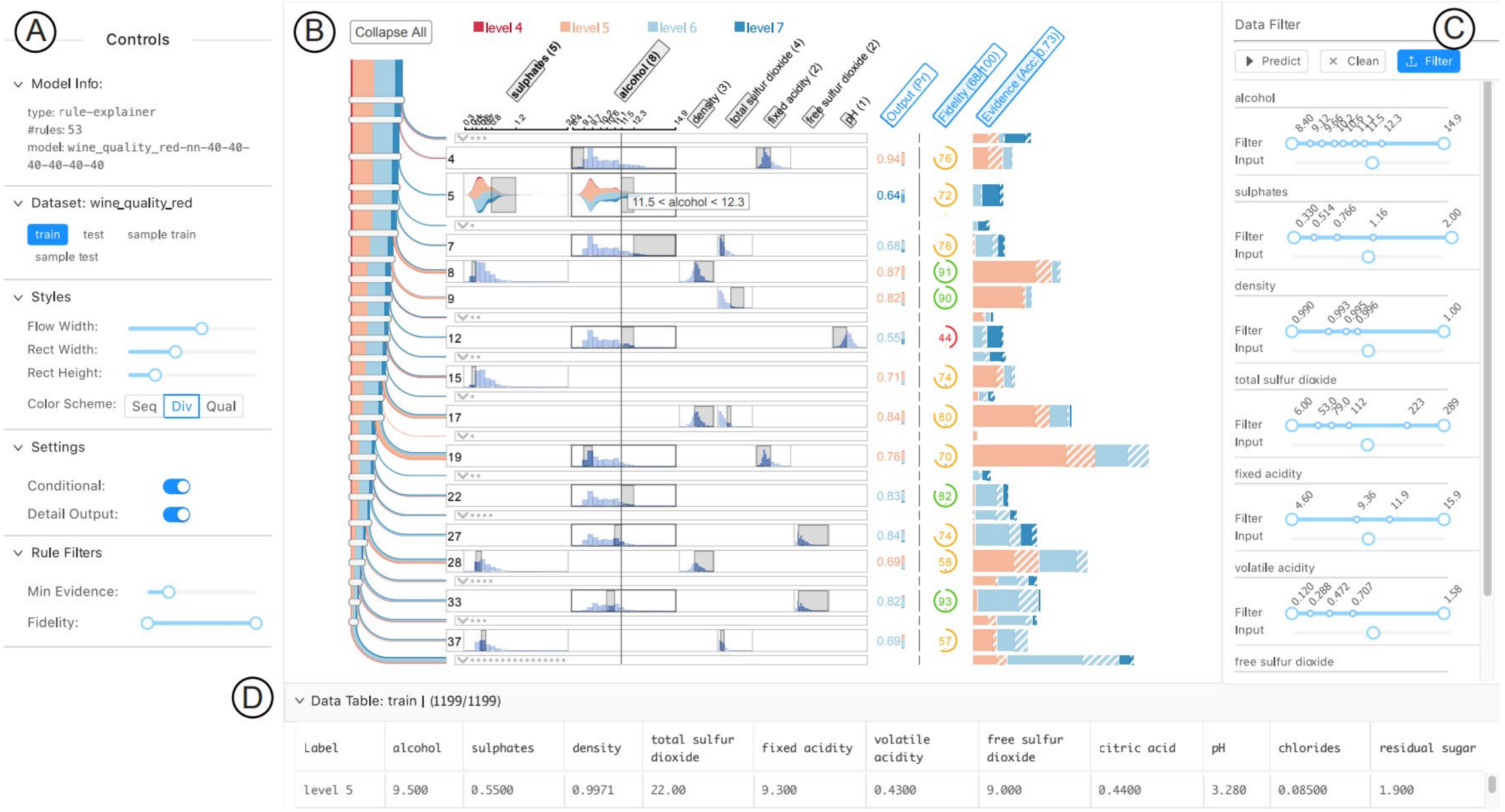


Similar Output

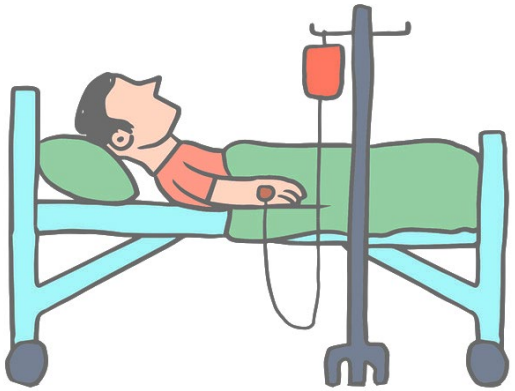
Neural Networks, Random Forest…

# Related Work

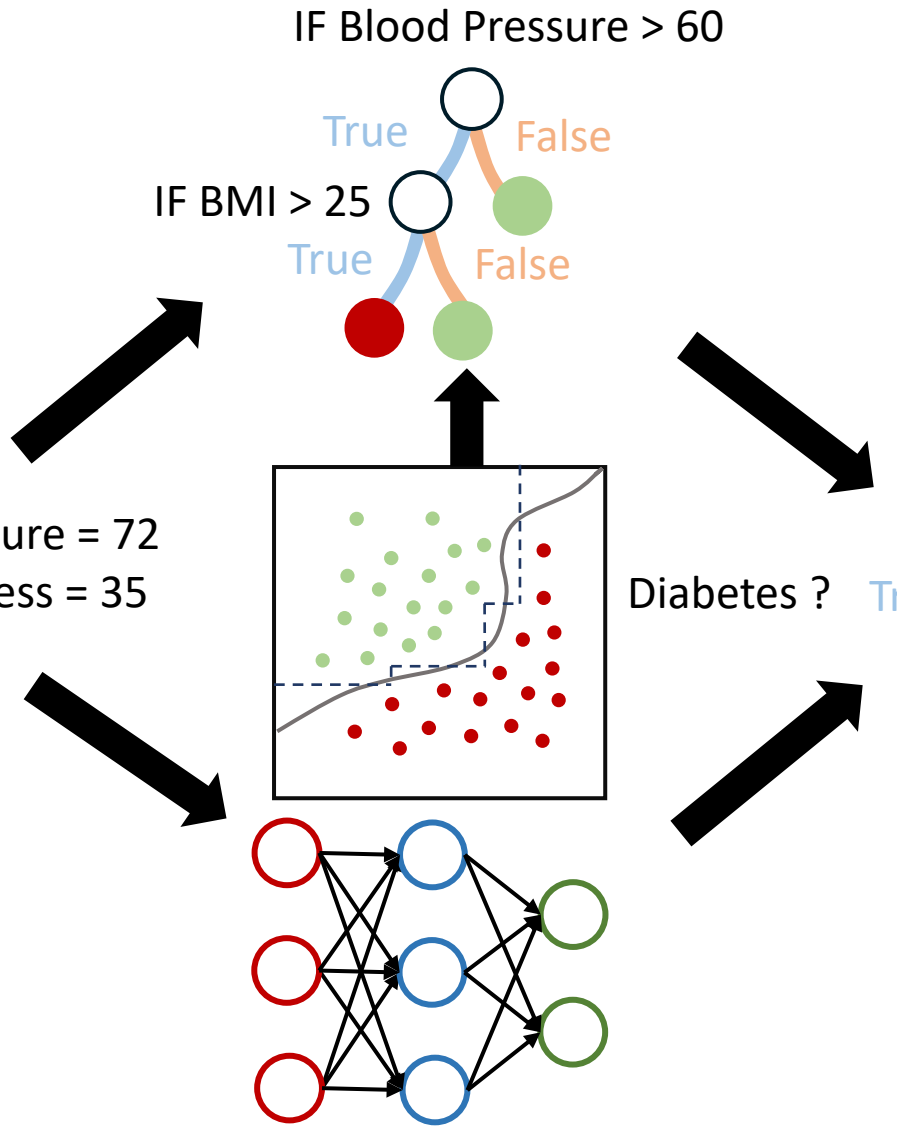
RuleMatrix



# For example



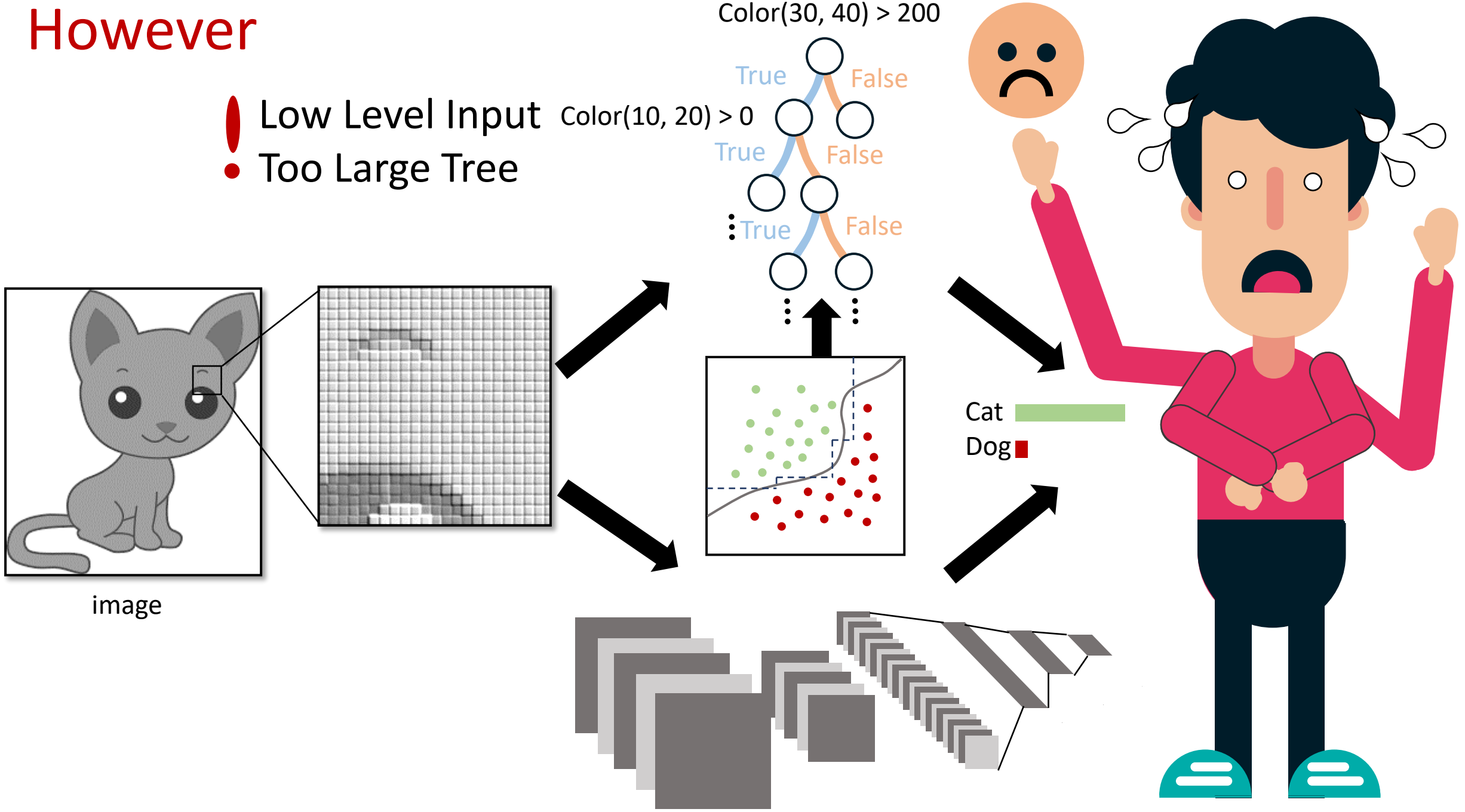
Blood Pressure = 72
Skin Thickness = 35
BMI = 33.6




Diabetes ? True

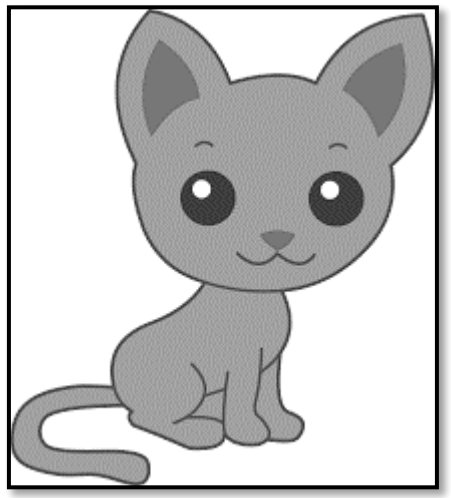
# However

- Low Level Input
- Too Large Tree

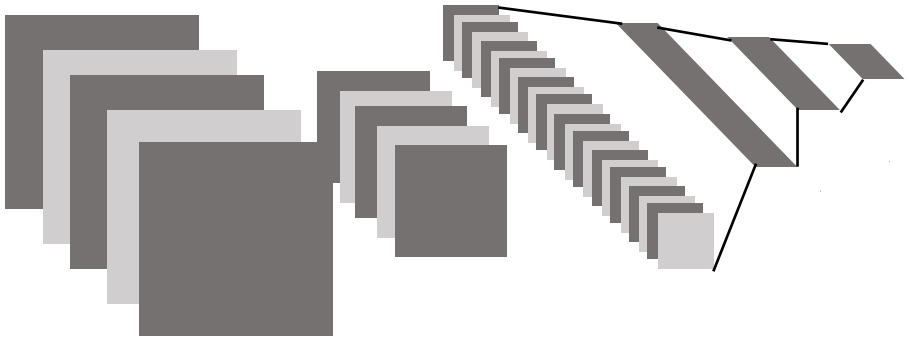


# Our Strategy:

1. Slice the Black Box



image



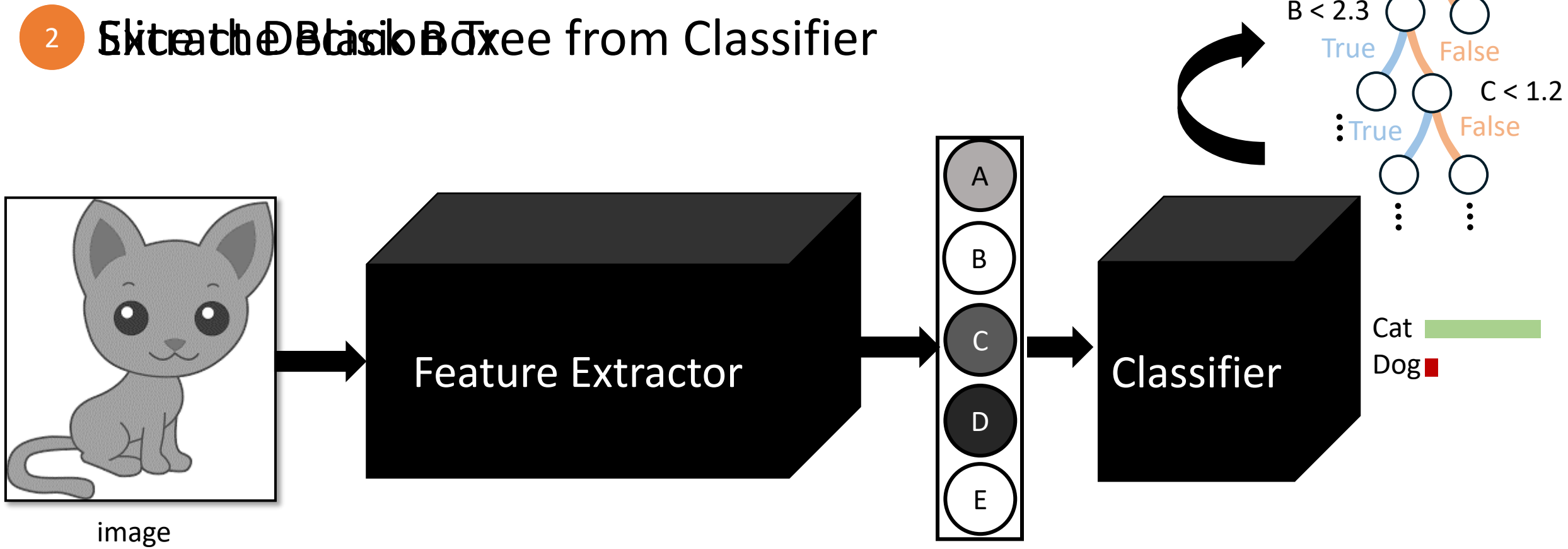
# Our Strategy:

1. Slice the Black Box


image

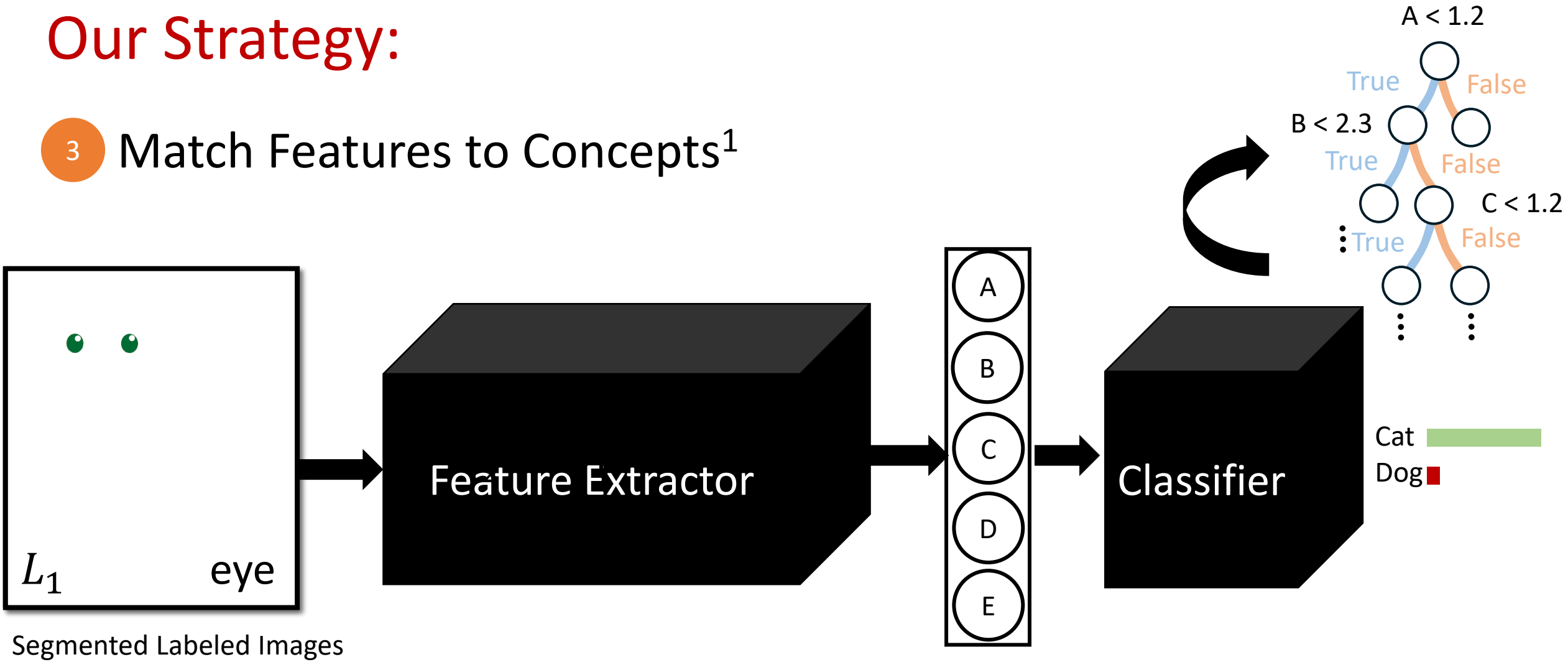
# Our Strategy:

2 Extract Black Box from Classifier



# Our Strategy:

## 3 Match Features to Concepts[1]

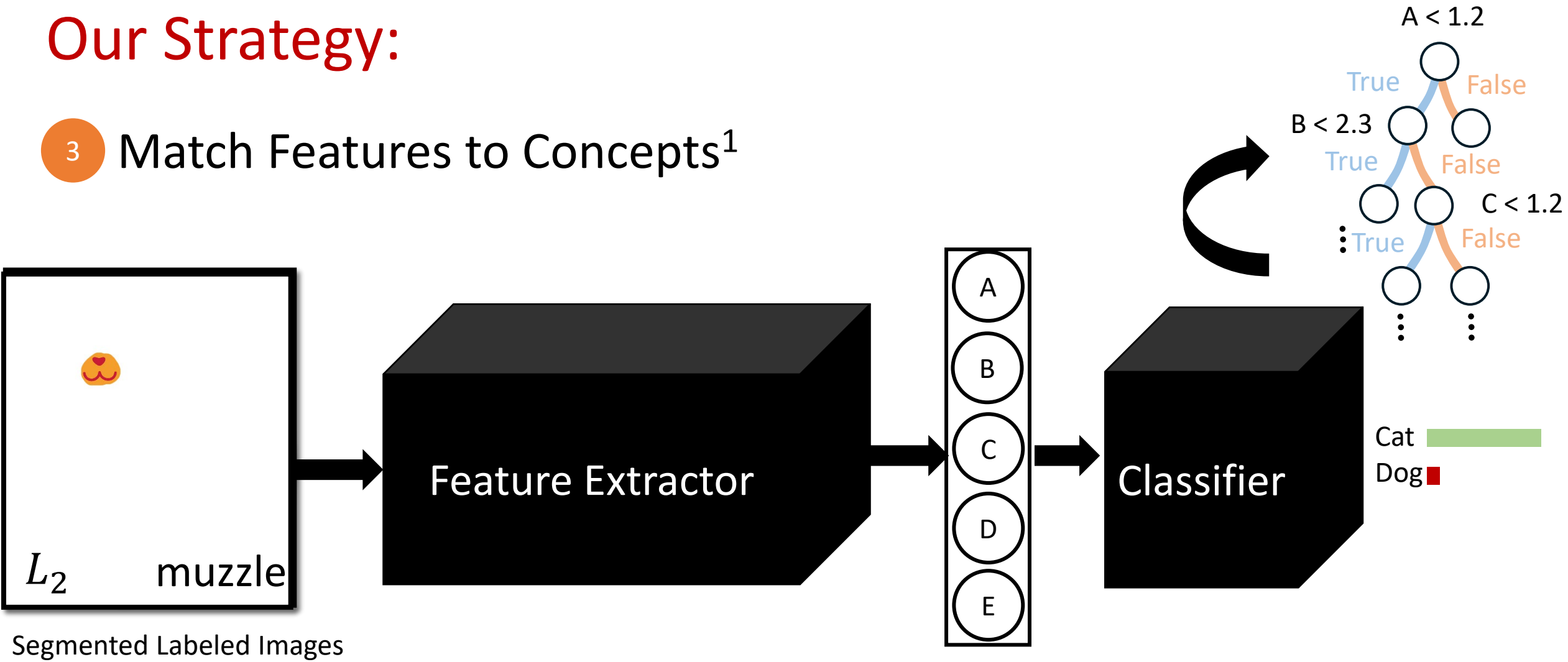



Segmented Labeled Images

1. Network Dissection: Quantifying Interpretability of Deep Visual Representations (CVPR 2017), David Bau, Bolei Zhou, et.al

# Our Strategy:

## 3 Match Features to Concepts[1]

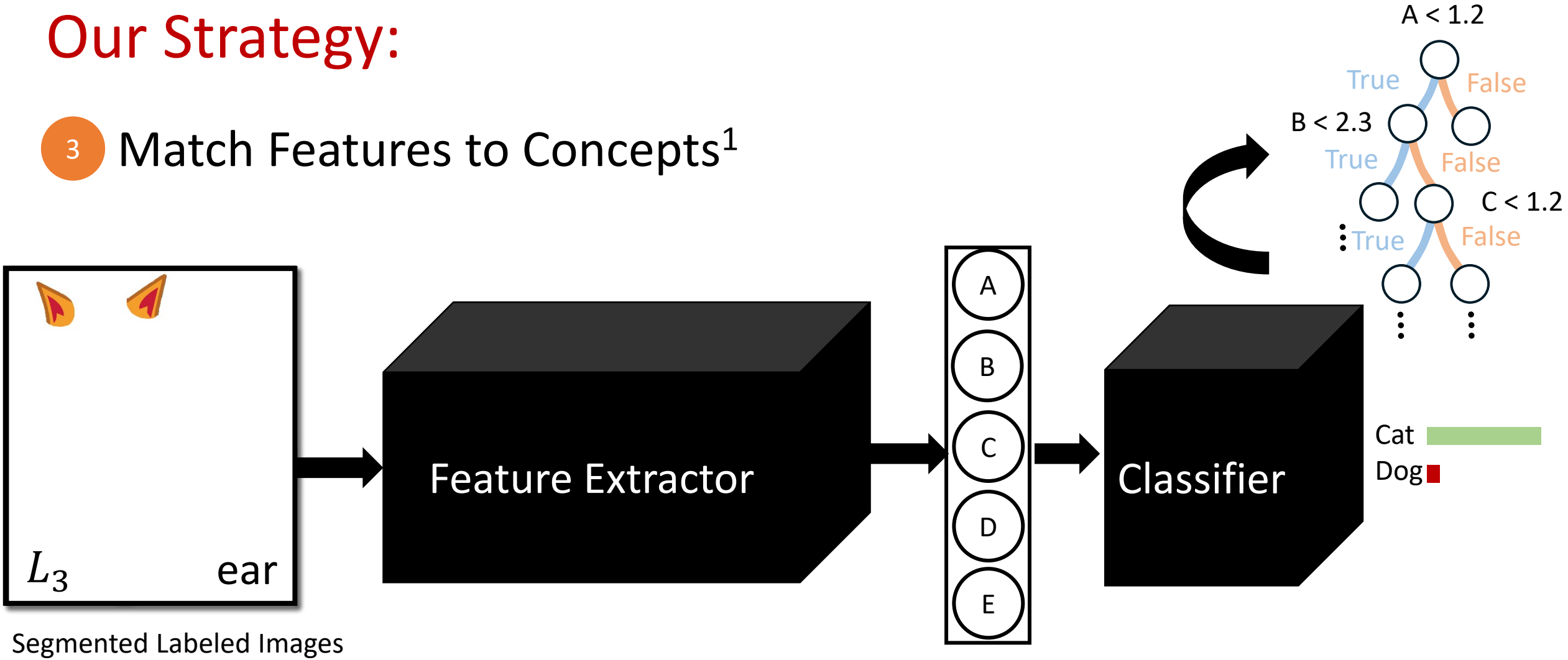



Segmented Labeled Images

1. Network Dissection: Quantifying Interpretability of Deep Visual Representations (CVPR 2017), David Bau, Bolei Zhou, et.al

# Our Strategy:

## 3 Match Features to Concepts[1]




Segmented Labeled Images

1. Network Dissection: Quantifying Interpretability of Deep Visual Representations (CVPR 2017), David Bau, Bolei Zhou, et.al

# Our Strategy:

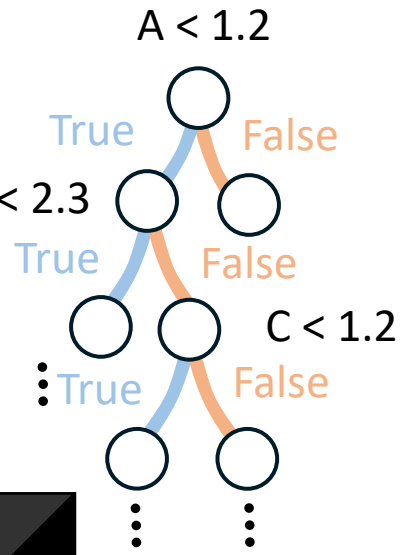
## 3 Match Features to Concepts[1]


eye

Feature Extractor

A
B
C
D
E

Classifier




Cat
Dog

Segmented Labeled Images

$$IoU(M_i, L_j) = \frac{M_i \cap L_j}{M_i \cup L_j}$$

Area of Overlap

Area of Union

$$IoU\ (M_1, L_1) > IoU\ (M_1, L_2) > IoU\ (M_1, L_3)$$

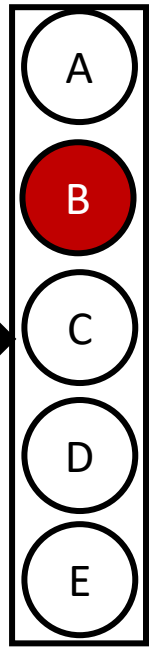
eye > muzzle > ear !

1. Network Dissection: Quantifying Interpretability of Deep Visual Representations (CVPR 2017), David Bau, Bolei Zhou, et.al

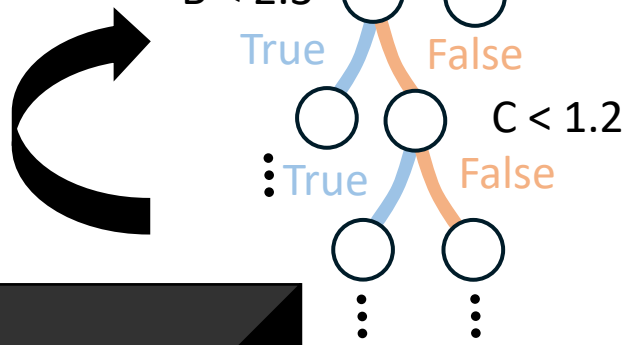
# Our Strategy:

## 3 Match Features to Concepts[1]


Segmented Labeled Images







$$IoU(M_i, L_j) = \frac{M_i \cap L_j}{M_i \cup L_j}$$

Area of Overlap

Area of Union

$$IoU\ (M_2, L_2) > IoU\ (M_2, L_1) > IoU\ (M_2, L_3)$$

muzzle > eye > ear !

1. Network Dissection: Quantifying Interpretability of Deep Visual Representations (CVPR 2017), David Bau, Bolei Zhou, et.al

# Our Strategy:

## 3 Match Features to Concepts[1]


Segmented Labeled Images

muzzle

Feature Extractor

A
B
C
D
E

Classifier

Cat
Dog

eye < 1.2
True
False
muzzle < 2.3
True
False
C < 1.2
True
False

$$IoU(M_i, L_j) = \frac{M_i \cap L_j}{M_i \cup L_j}$$

Area of Overlap

Area of Union

$$IoU\ (M_2, L_2) > IoU\ (M_2, L_1) > IoU\ (M_2, L_3)$$

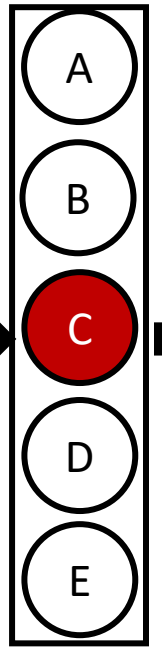
muzzle > eye > ear !

1. Network Dissection: Quantifying Interpretability of Deep Visual Representations (CVPR 2017), David Bau, Bolei Zhou, et.al

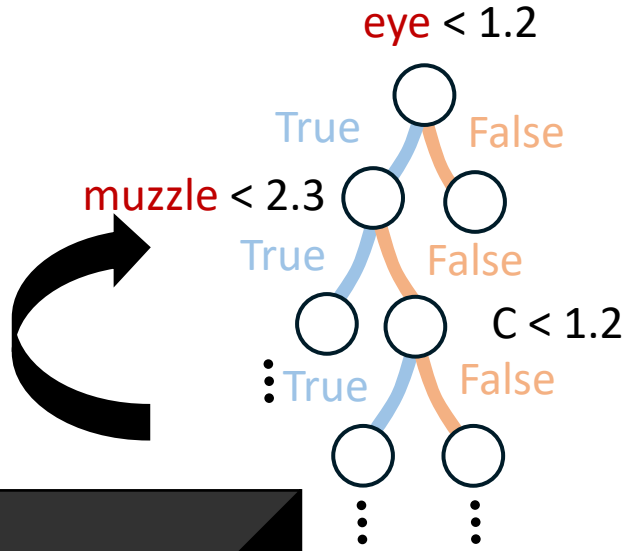
# Our Strategy:

## 3 Match Features to Concepts[1]


Segmented Labeled Images







$$IoU(M_i, L_j) = \frac{M_i \cap L_j}{M_i \cup L_j}$$

Area of Overlap

Area of Union

$$IoU\ (M_3, L_3) > IoU\ (M_3, L_1) > IoU\ (M_3, L_3)$$

ear > eye > muzzle !

1. Network Dissection: Quantifying Interpretability of Deep Visual Representations (CVPR 2017), David Bau, Bolei Zhou, et.al

# Our Strategy:

## 3 Match Features to Concepts[1]


ear

Feature Extractor

A
B
C
D
E

Classifier

Cat
Dog

eye < 1.2
True
False
muzzle < 2.3
True
False
ear < 1.2
True
False

Segmented Labeled Images

$$IoU(M_i, L_j) = \frac{M_i \cap L_j}{M_i \cup L_j}$$

Area of Overlap

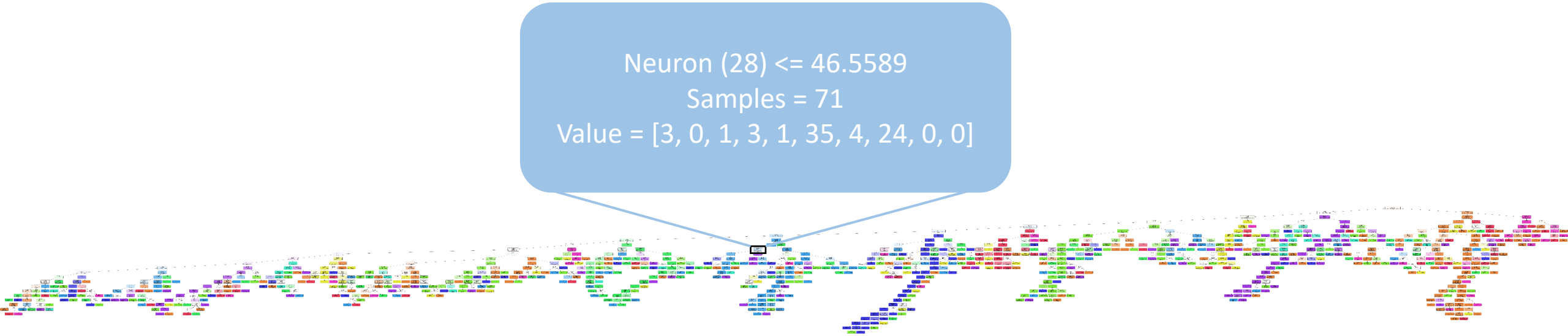
Area of Union

$$IoU\ (M_3, L_3) > IoU\ (M_3, L_1) > IoU\ (M_3, L_3)$$

ear > eye > muzzle !

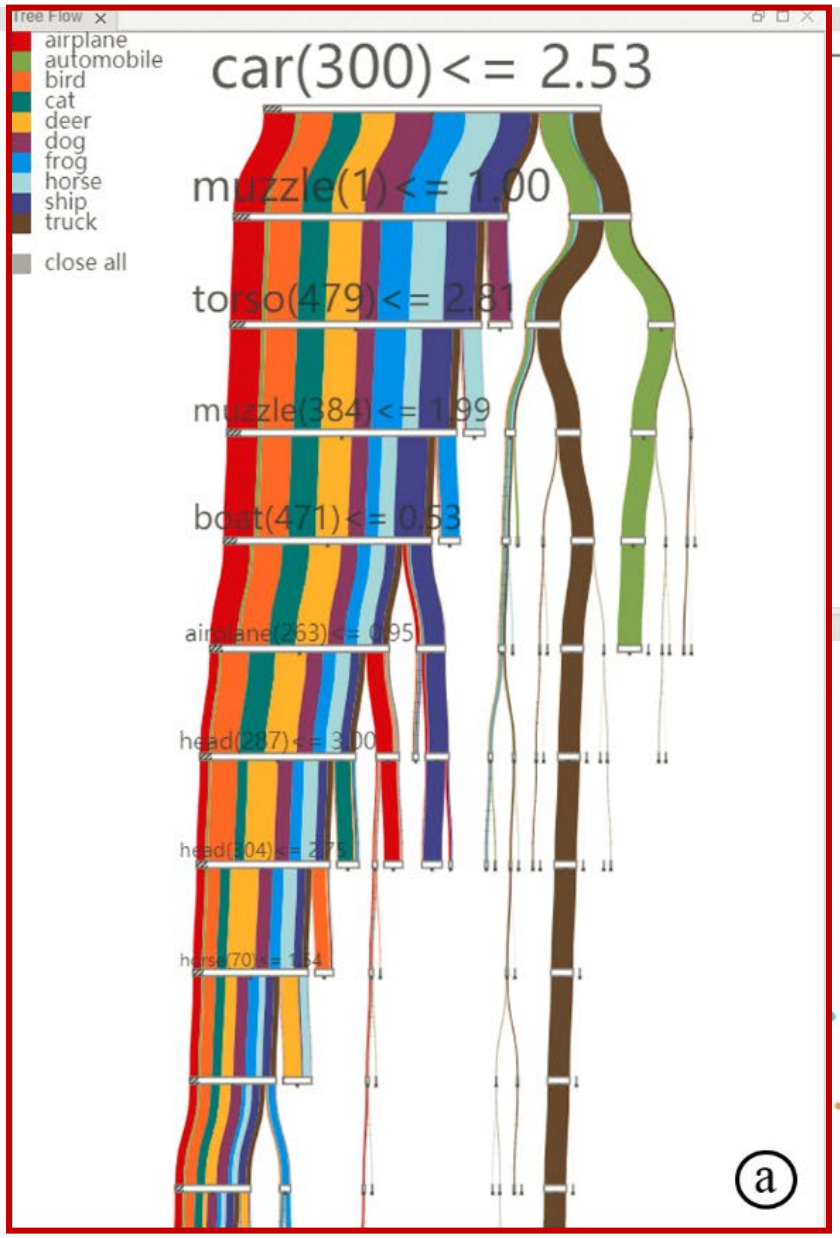
1. Network Dissection: Quantifying Interpretability of Deep Visual Representations (CVPR 2017), David Bau, Bolei Zhou, et.al

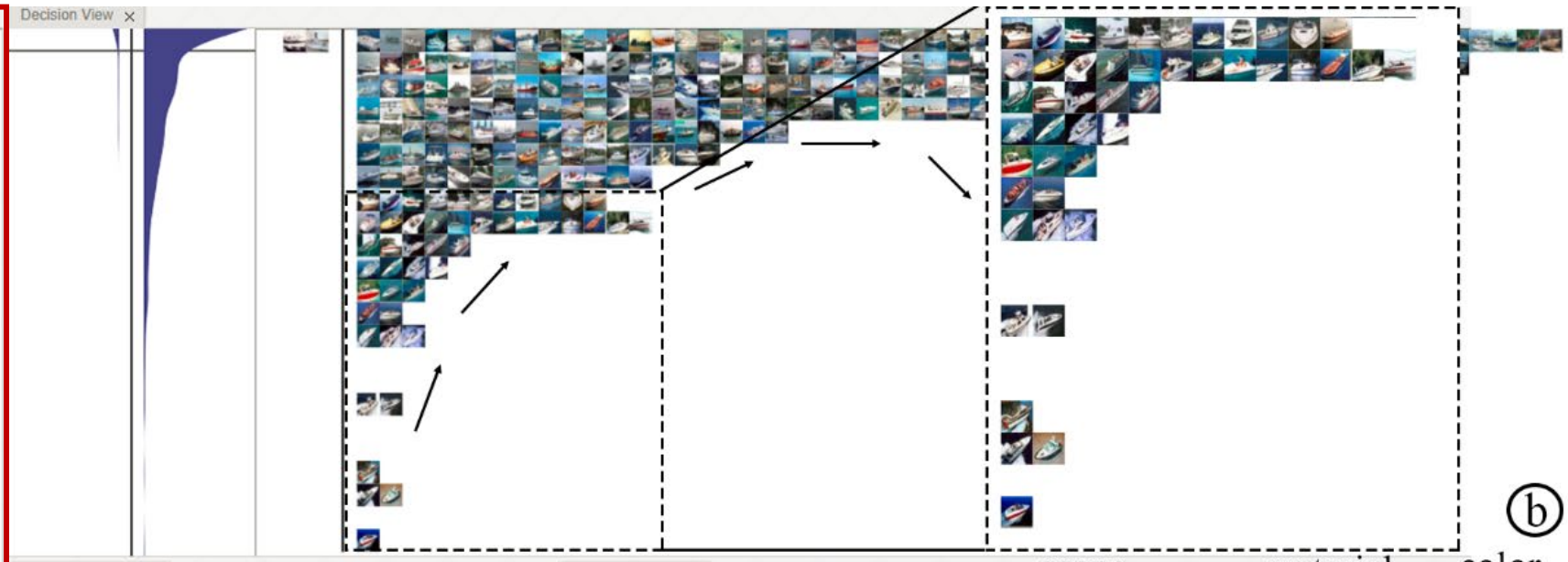
# Challenges:

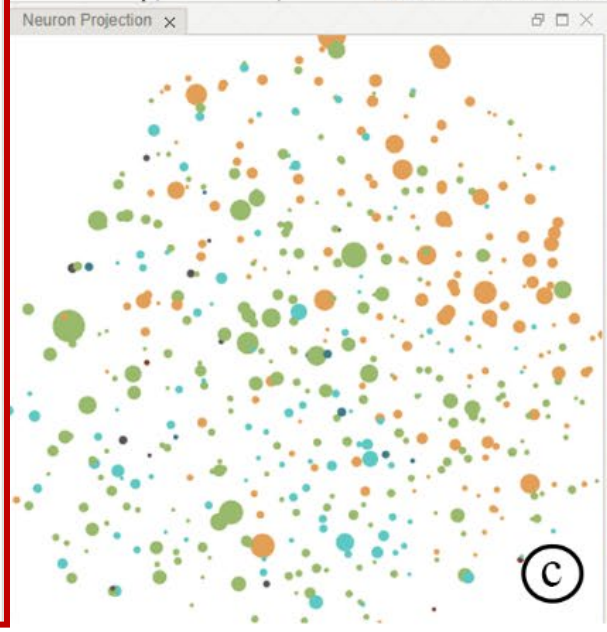


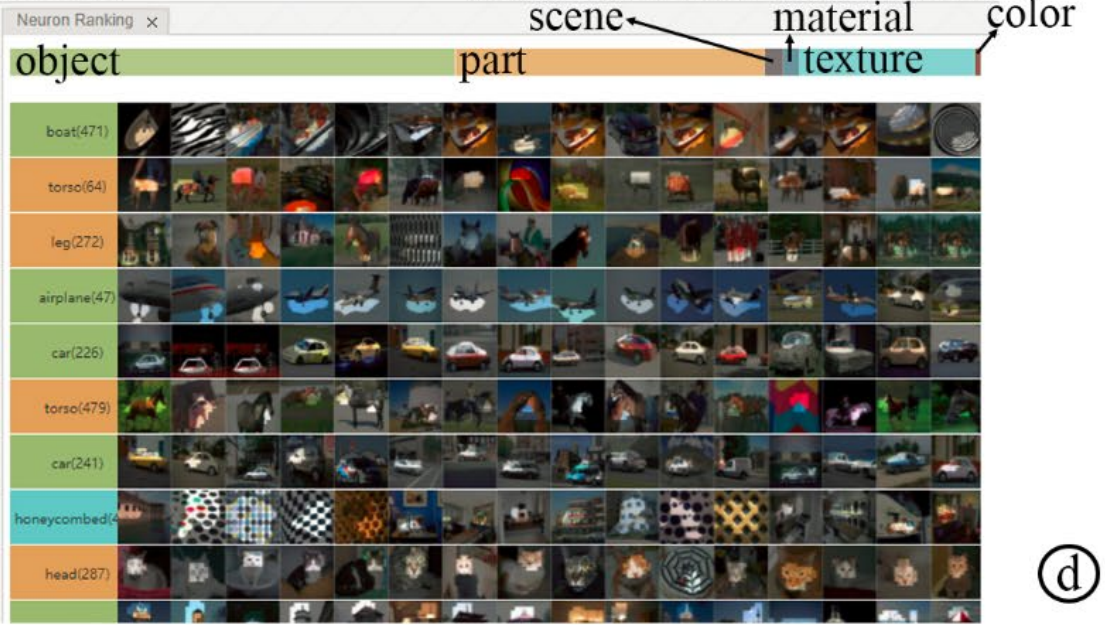

1. Scalability of large decision trees
2. Hard to grasp information of each rule
3. Hard to reason misclassification

# TreeFlow: dataflow along the decision tree

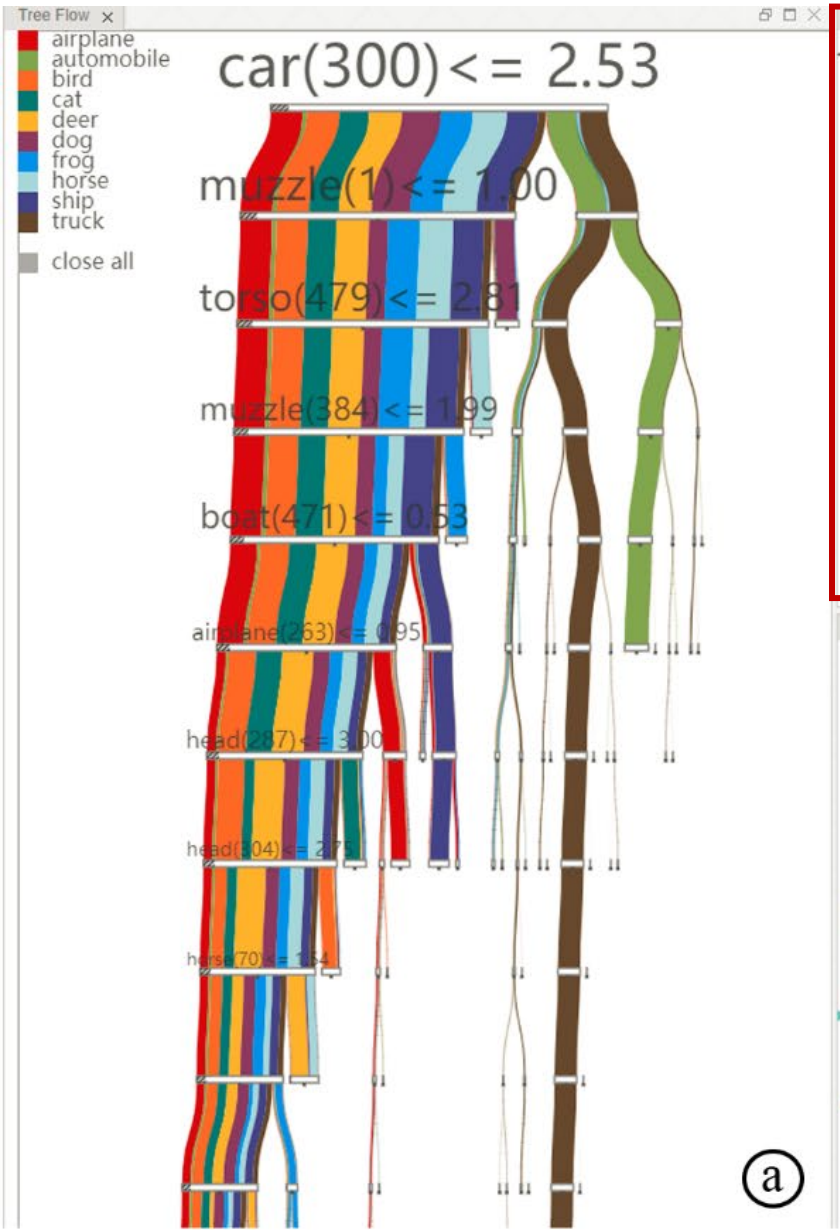


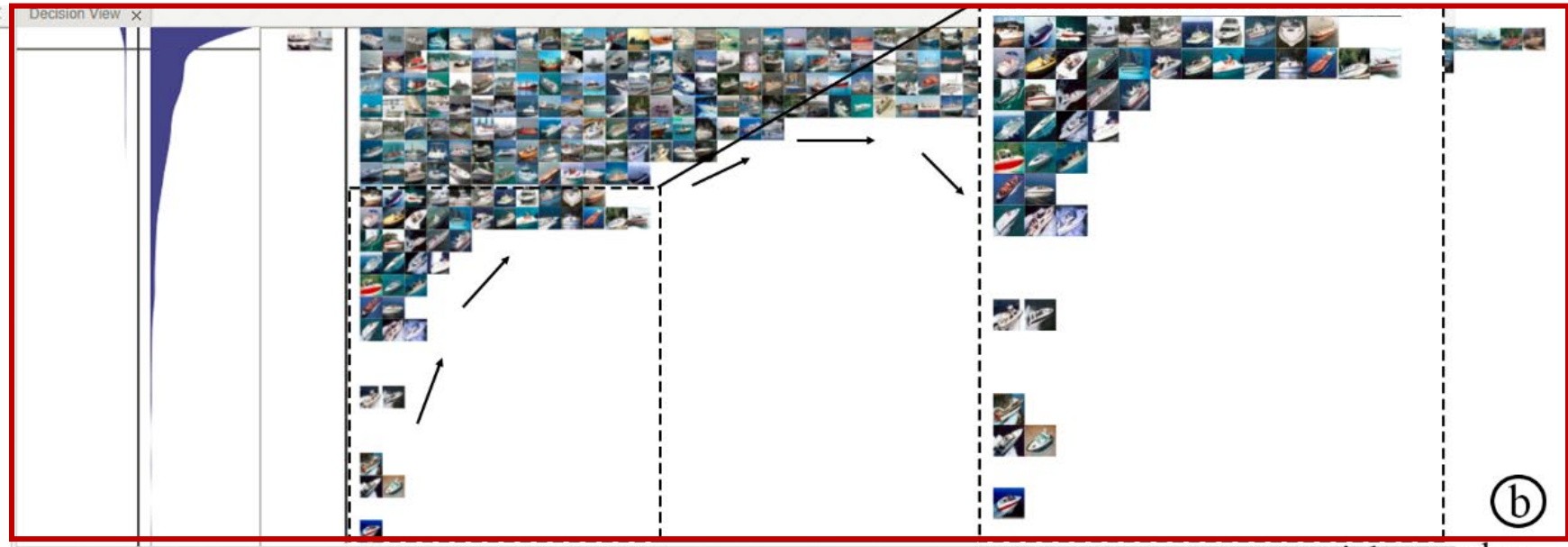



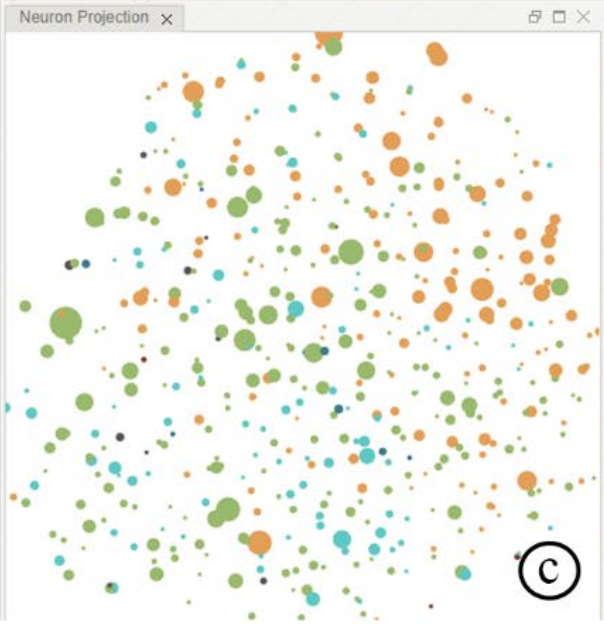



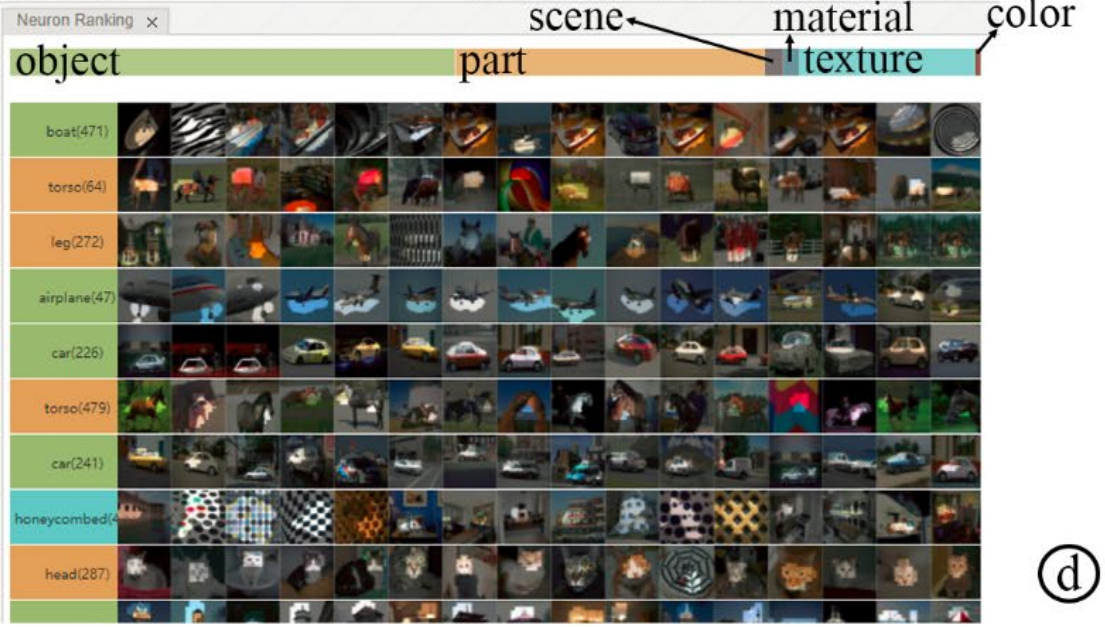



# Decision View: data distribution of decision node

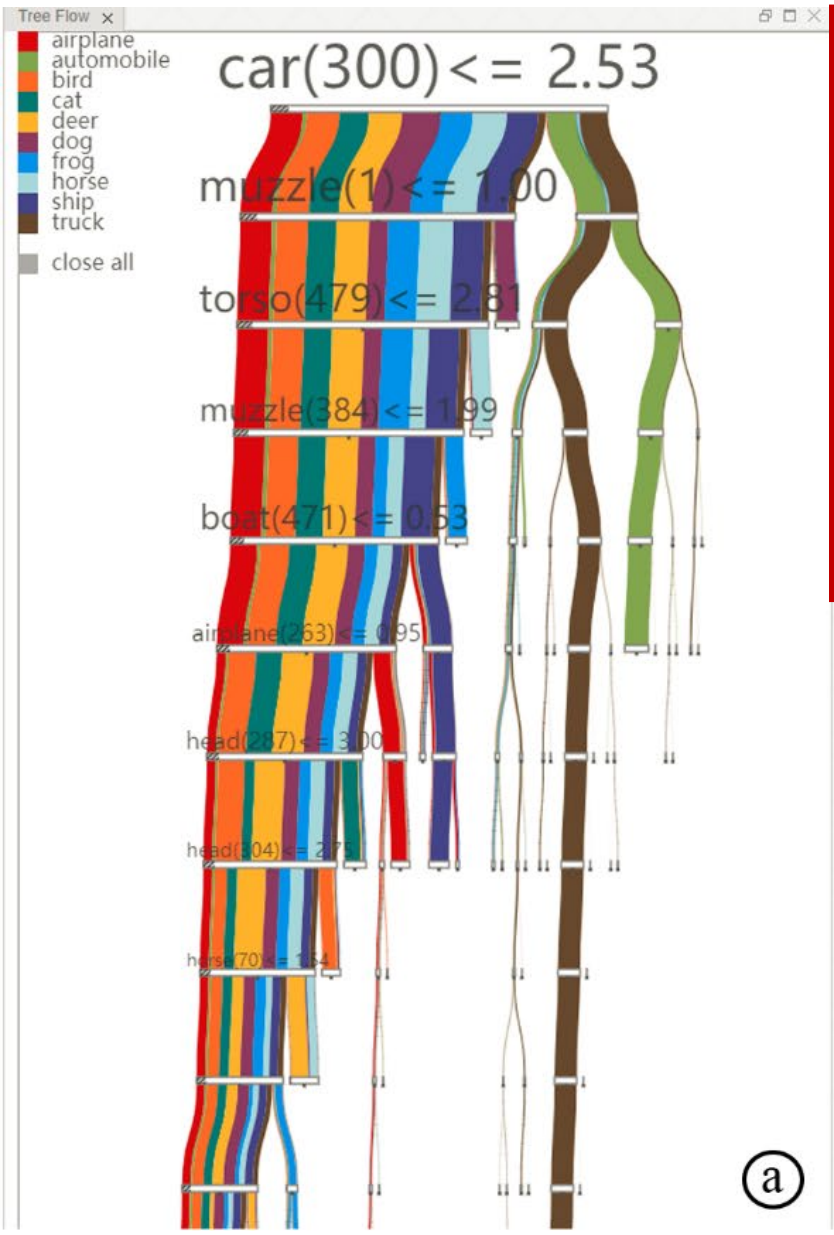


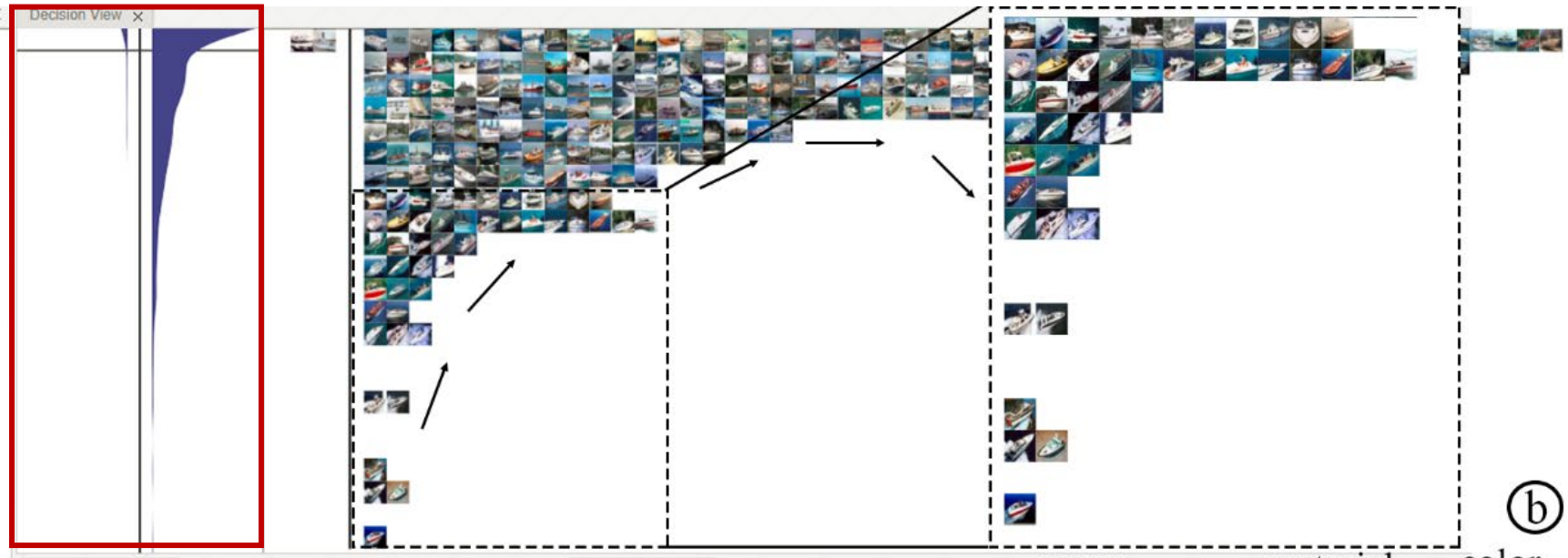



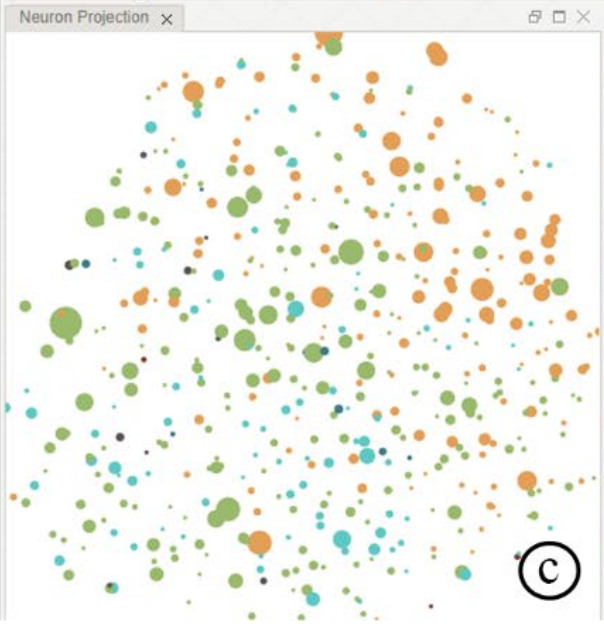



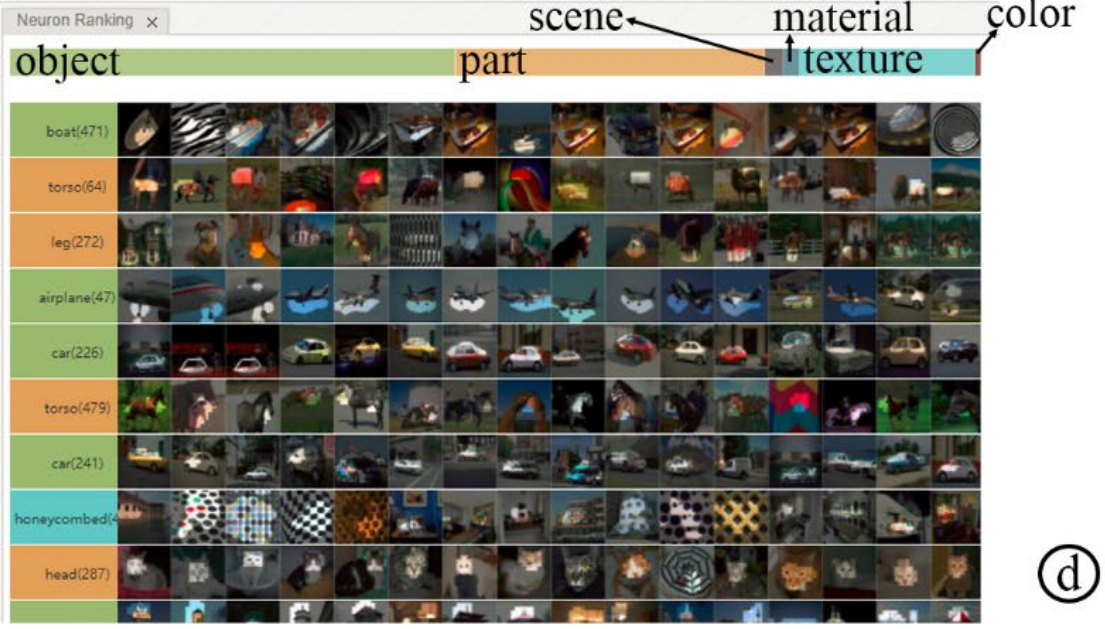



# Decision View: data distribution of decision node

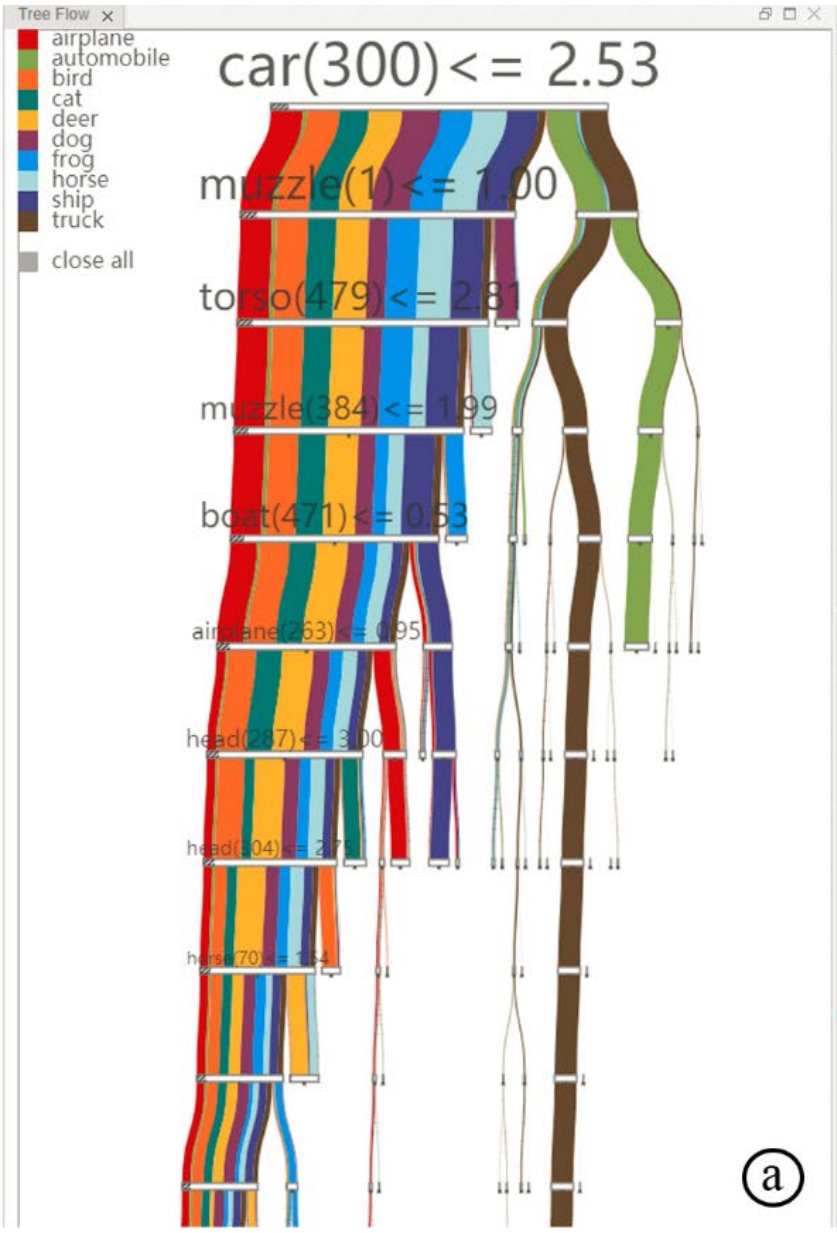


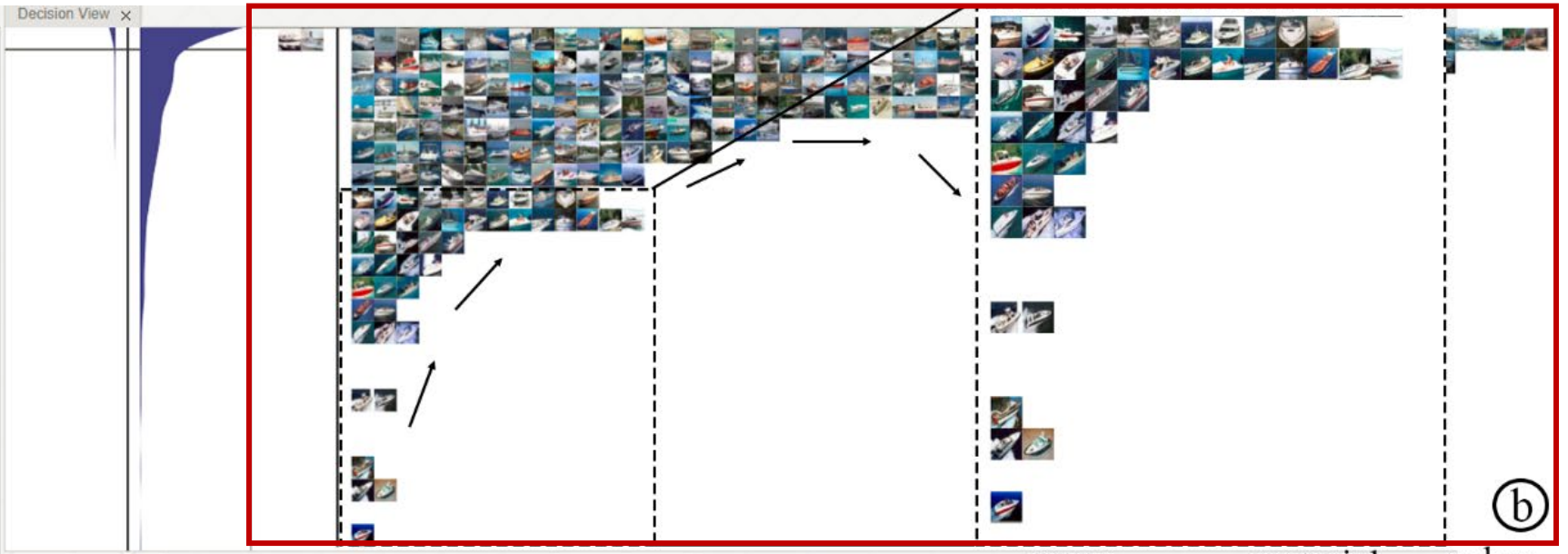



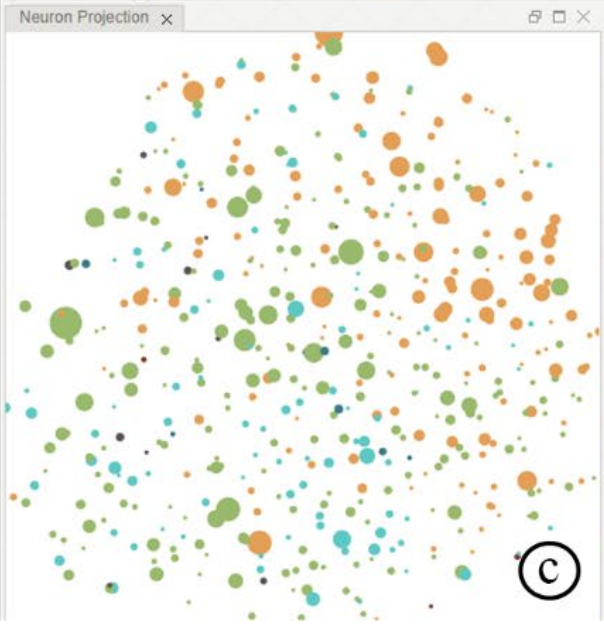



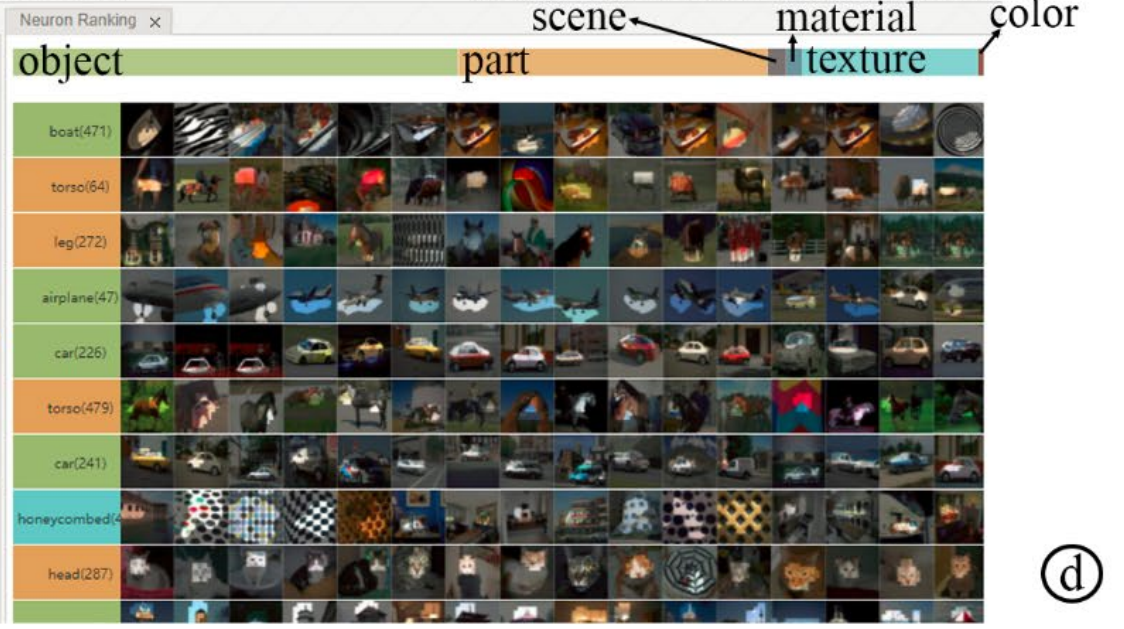



# Decision View: data distribution of decision node

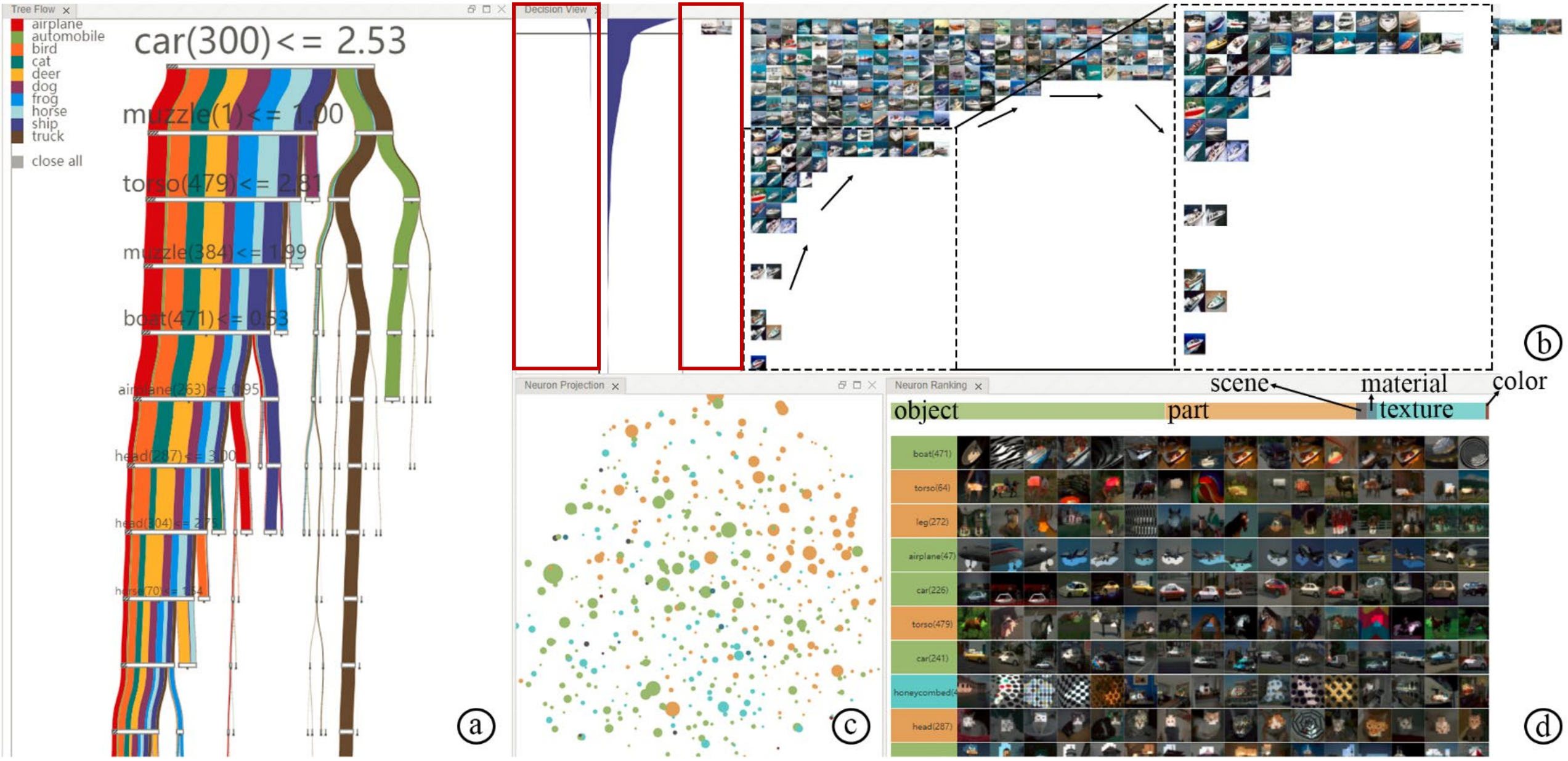


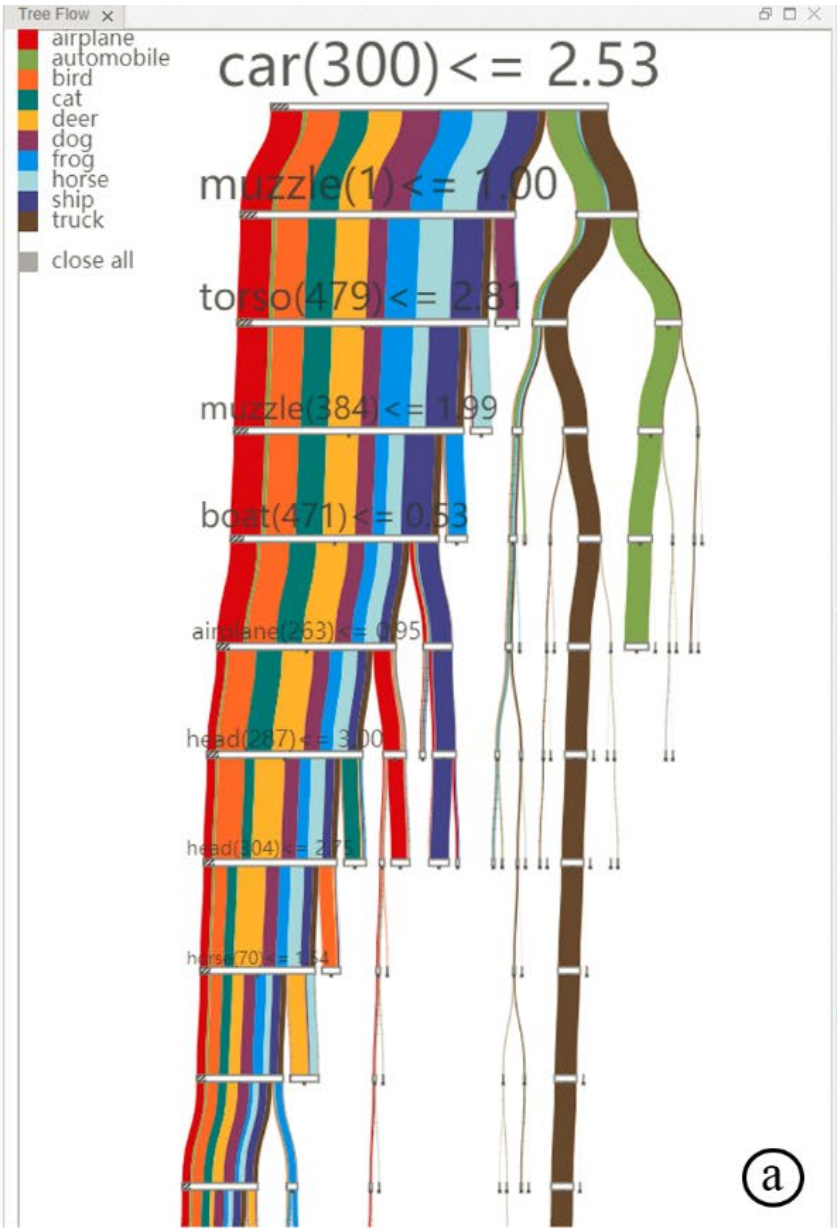



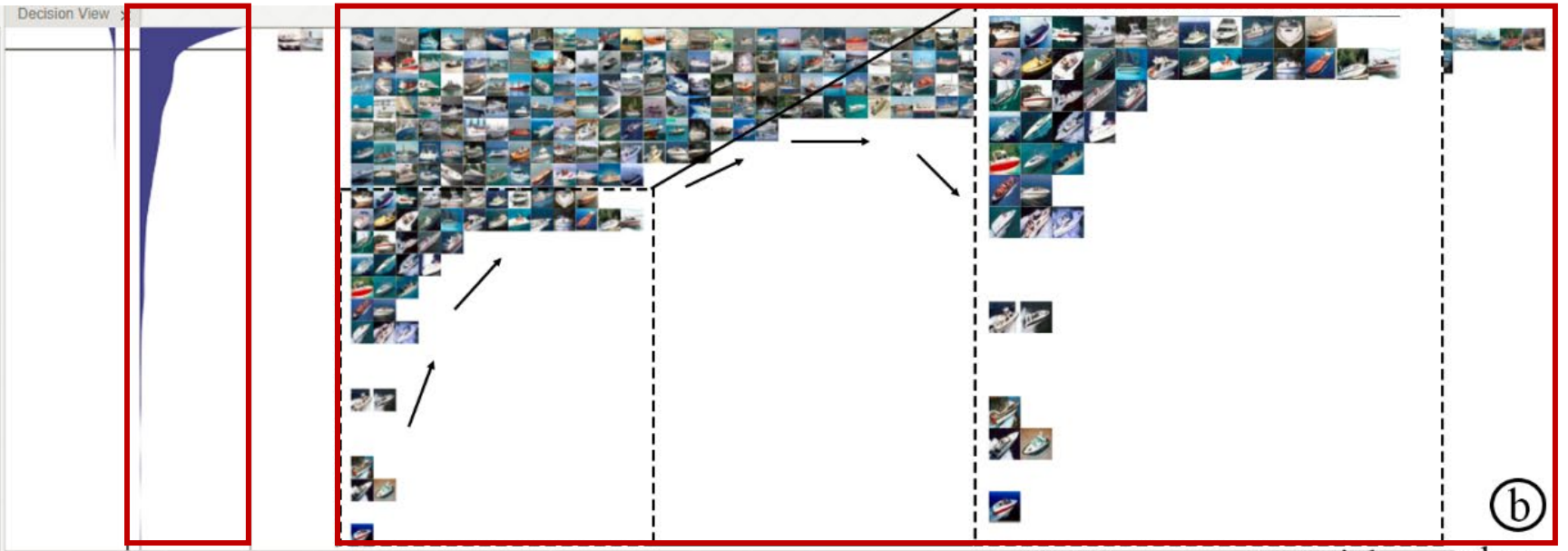



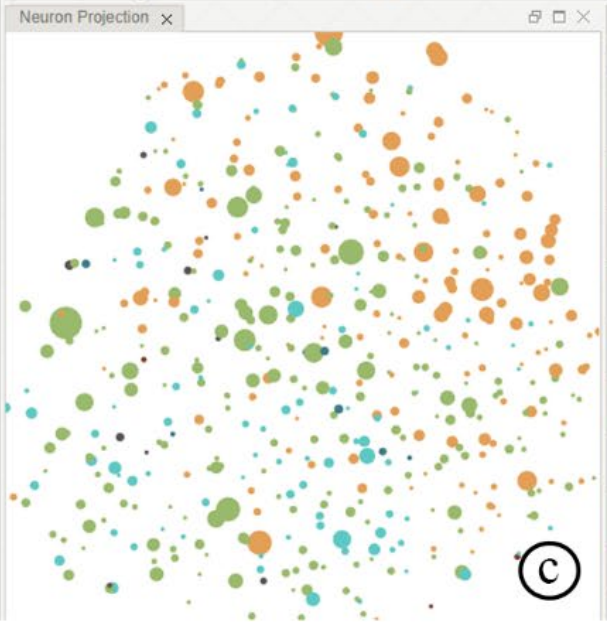
# Decision View: data distribution of decision node

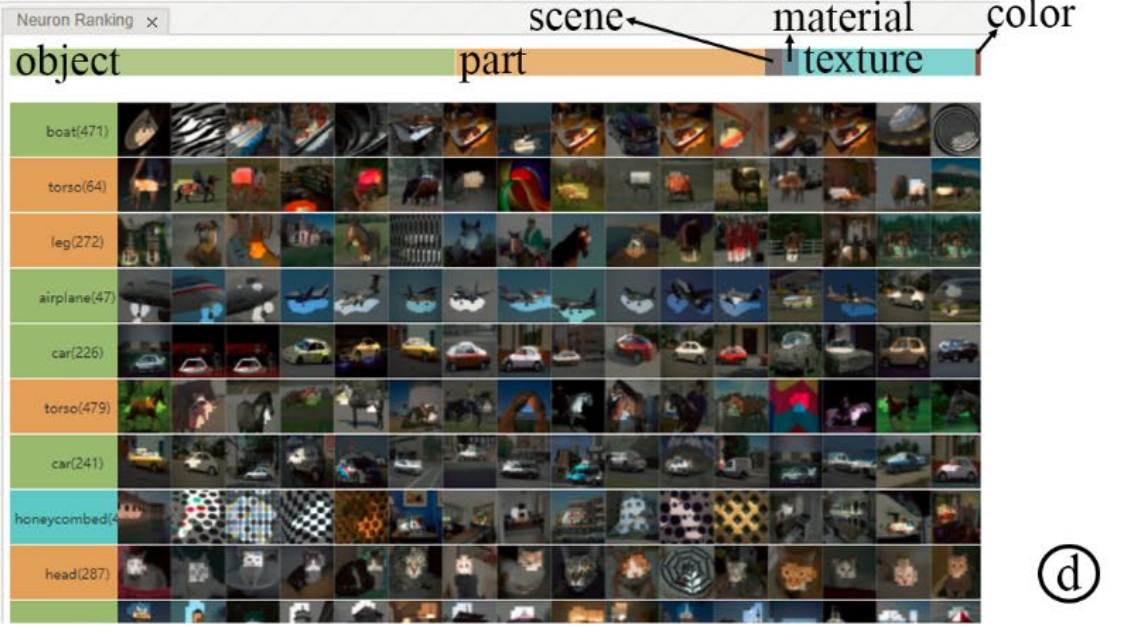


# Decision View: data distribution of decision node

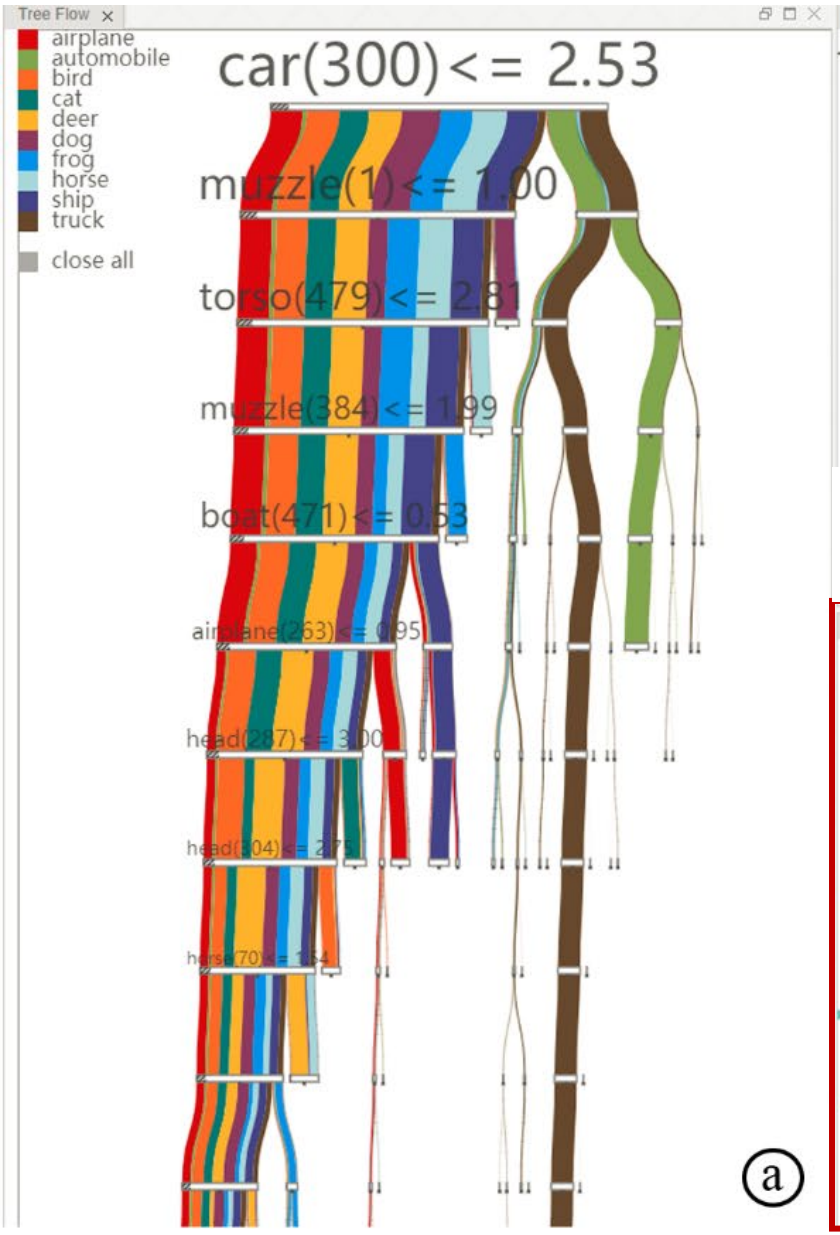


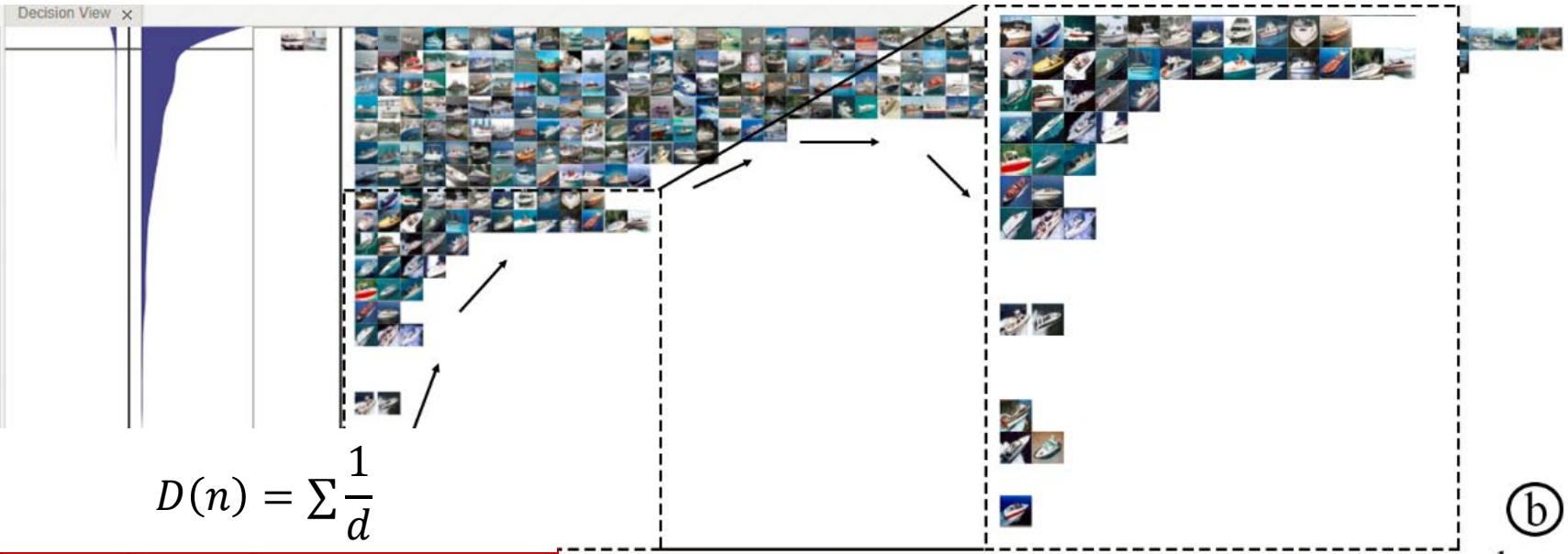



# Neuron View: t-SNE projection of neurons based on semantics similarity

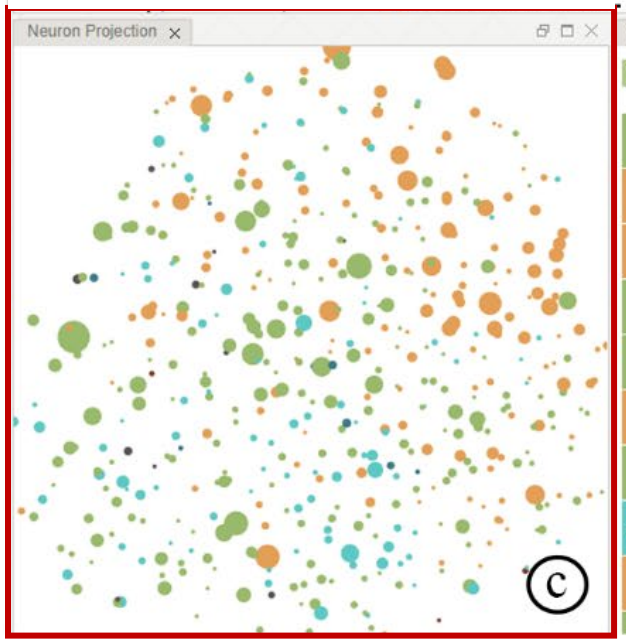



a

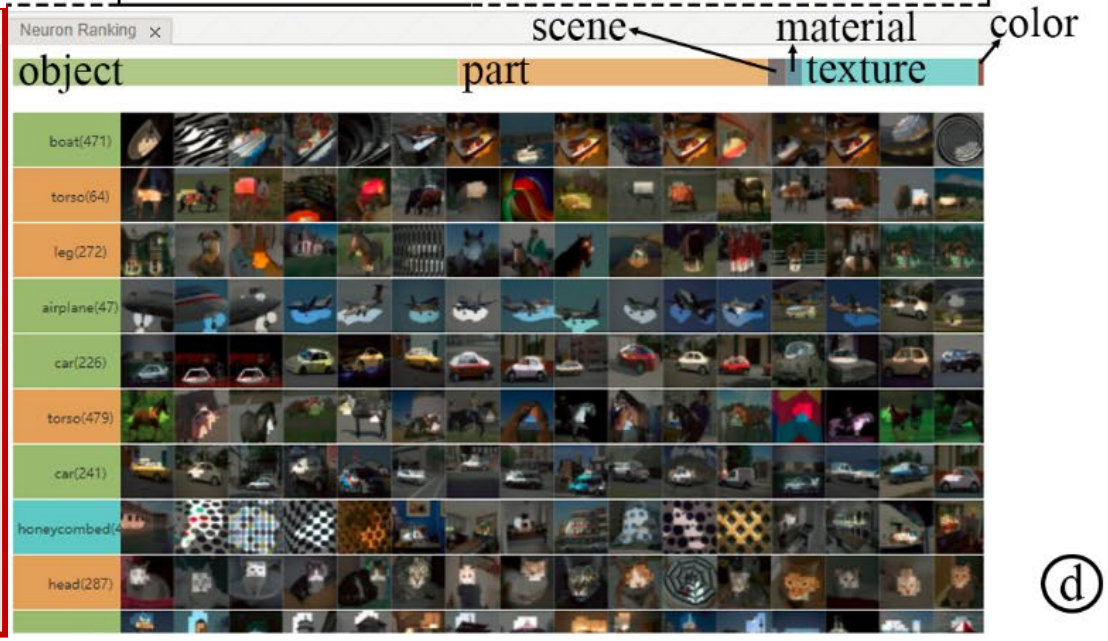



$$D(n) = \sum \frac{1}{d}$$

b

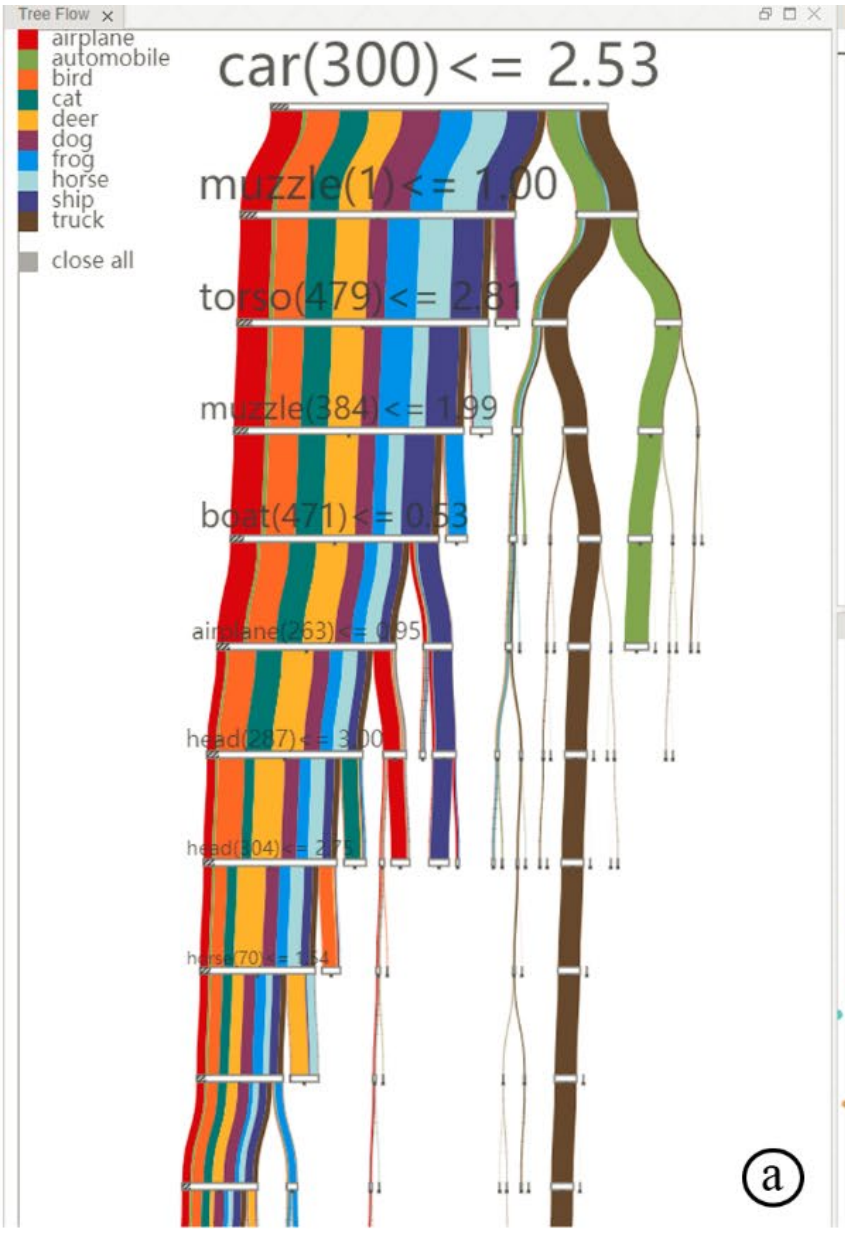


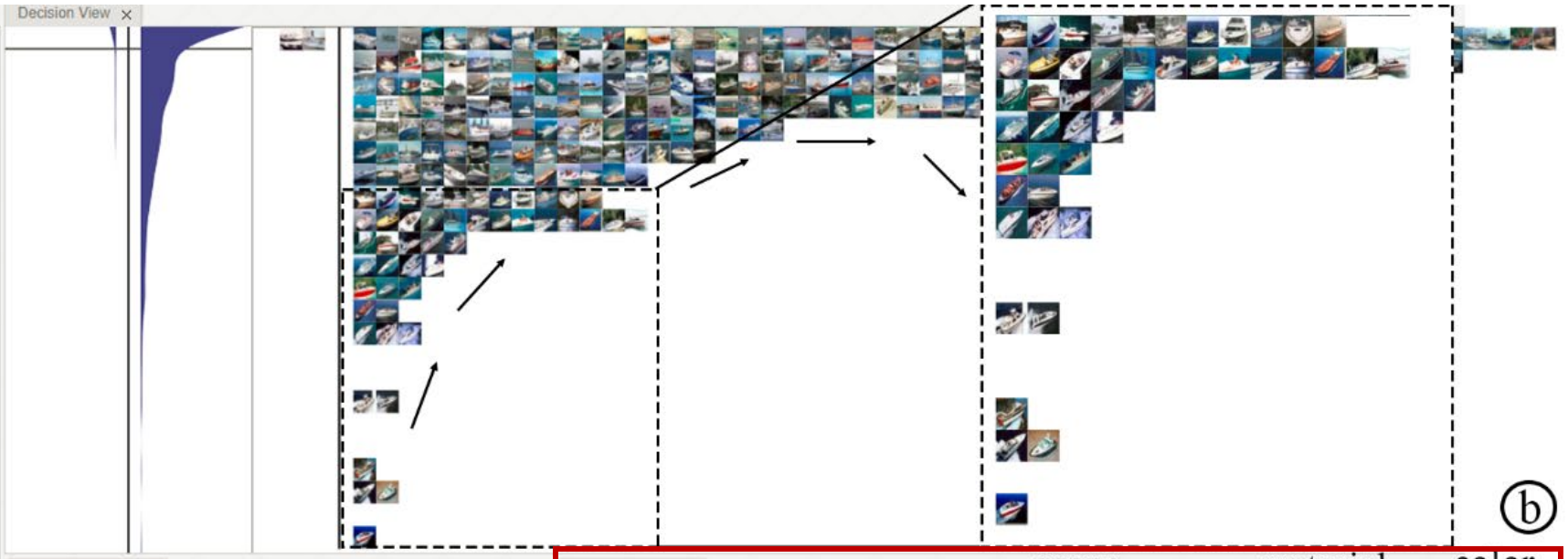

c

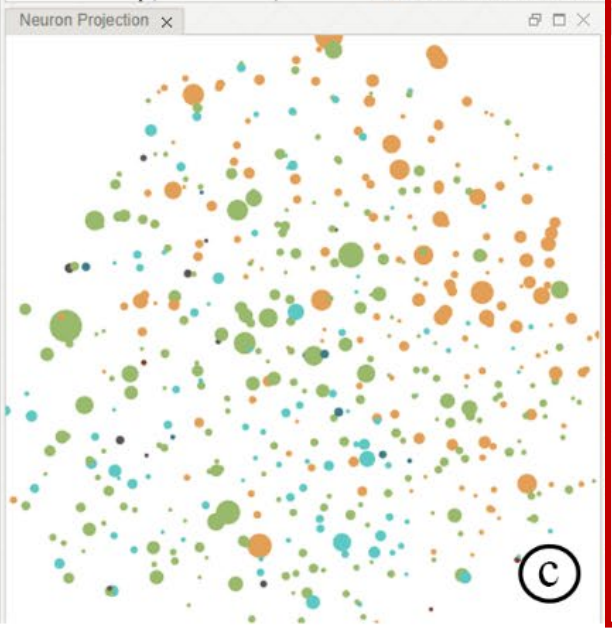


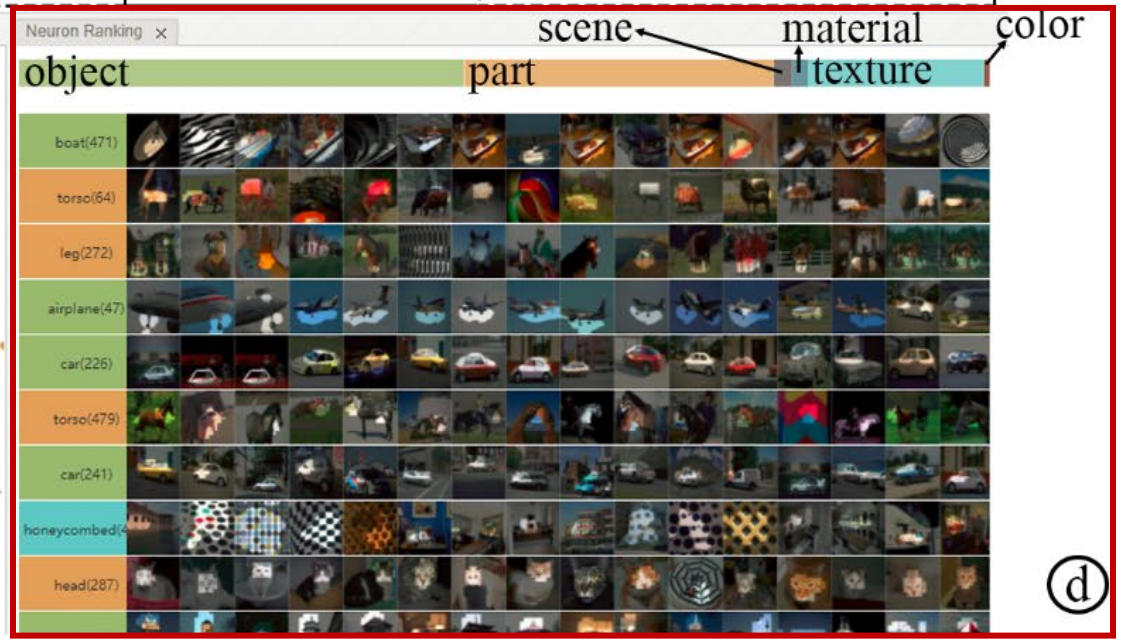

d

# Neuron Ranking: a collection of semantic labels and salience maps



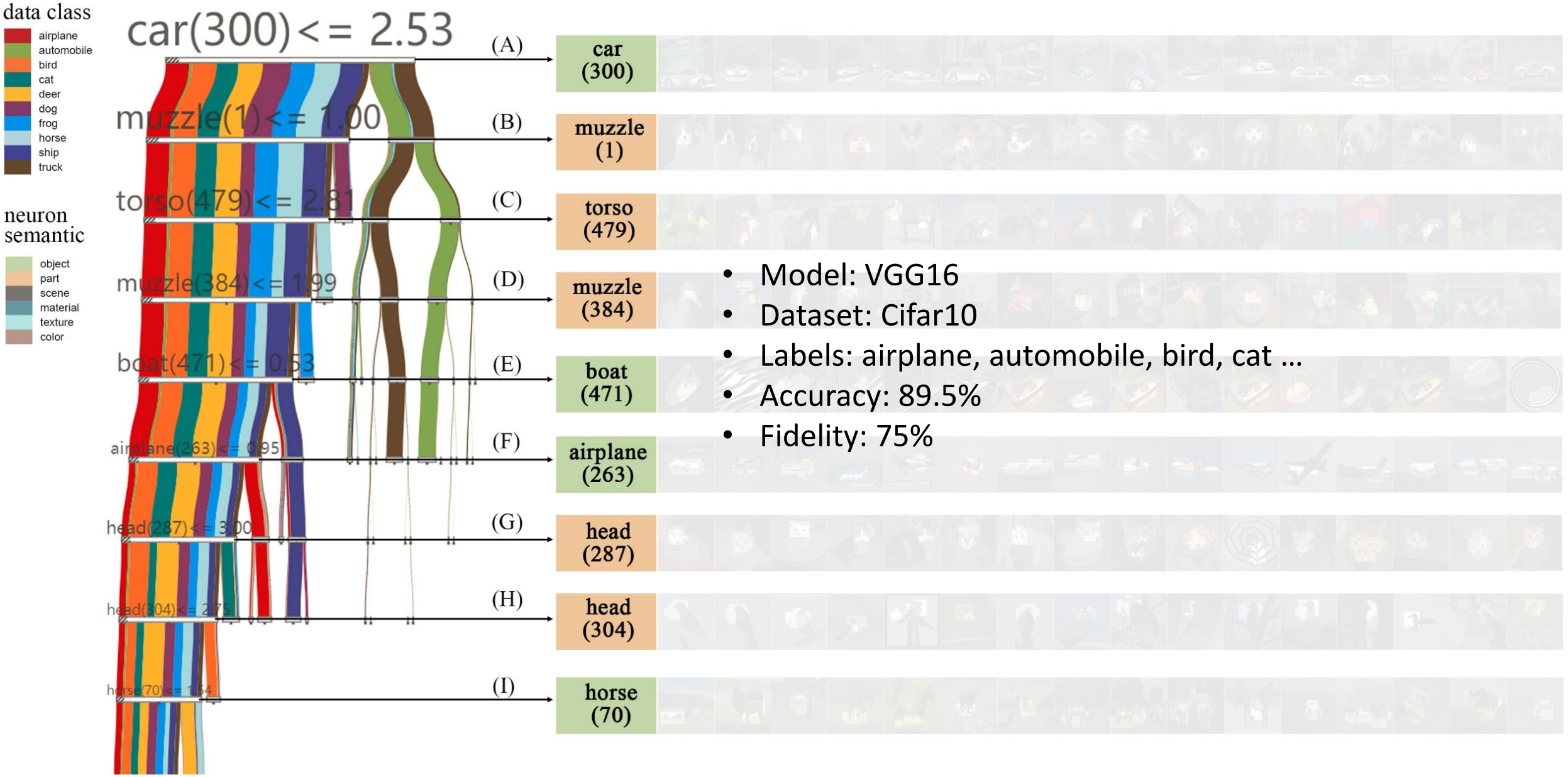



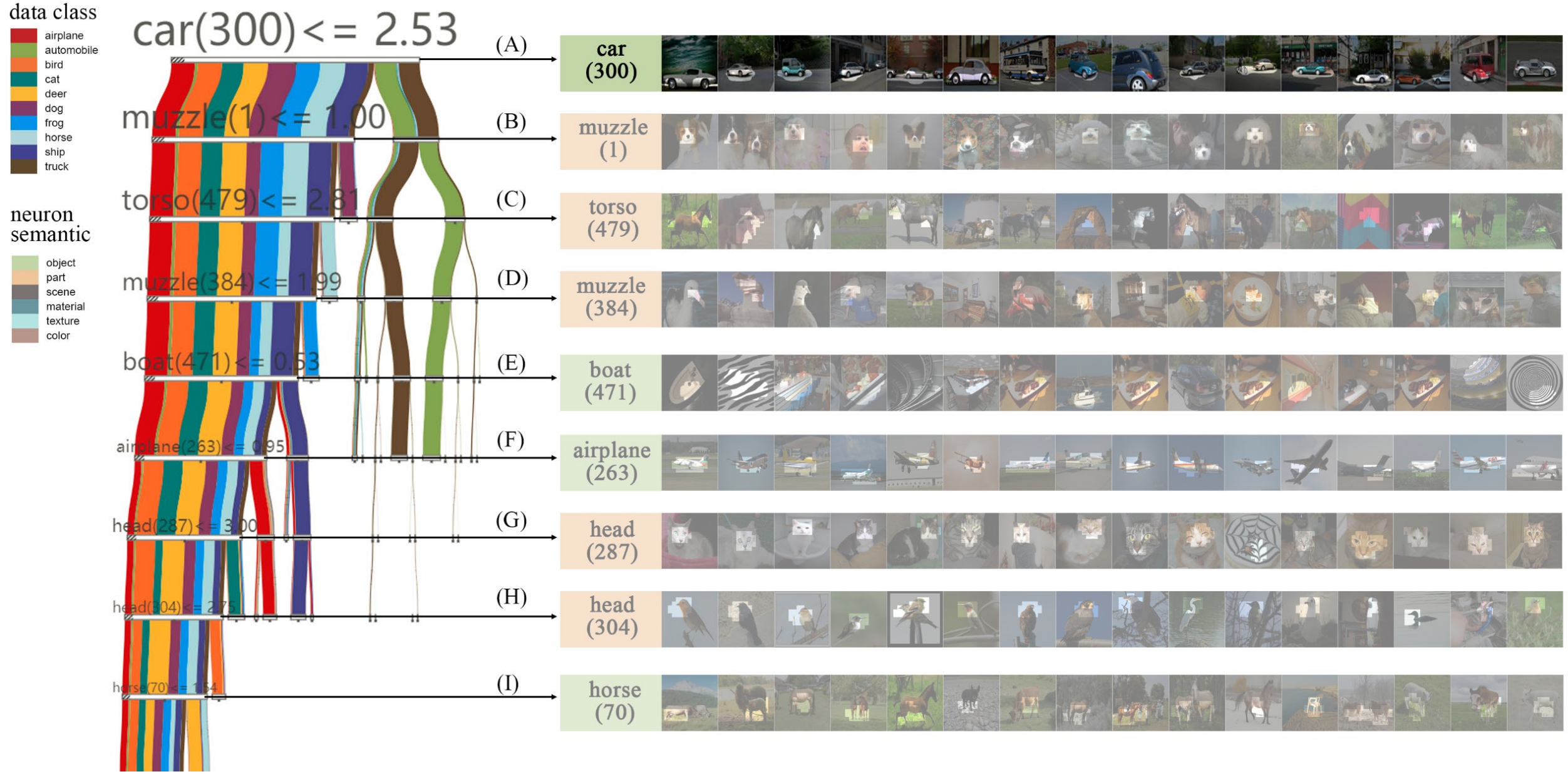
# Basic Interactions

# Use Case 1: Interpreting Surrogate Decision Tree

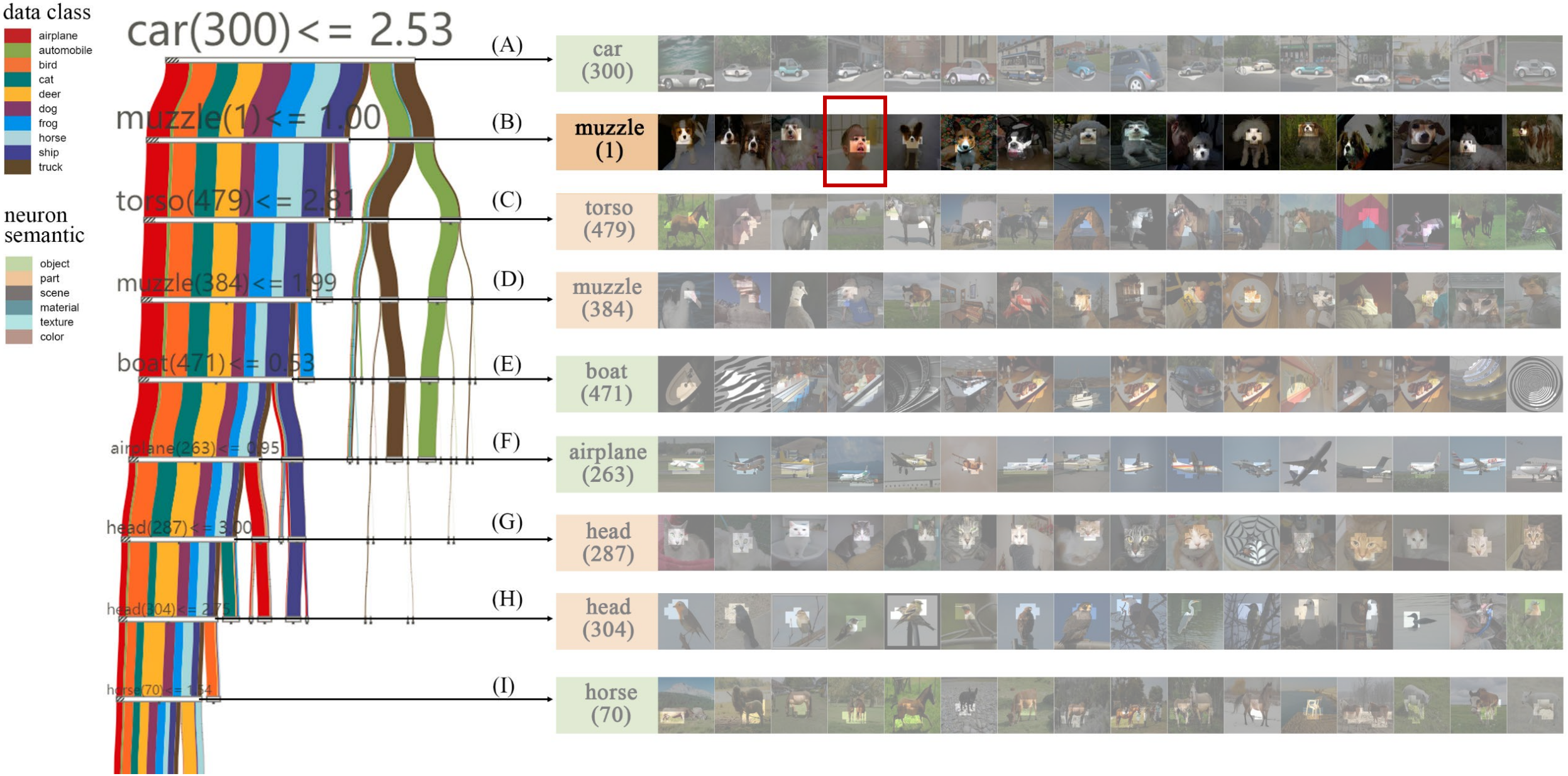



- Model: VGG16
- Dataset: Cifar10
- Labels: airplane, automobile, bird, cat …
- Accuracy: 89.5%
- Fidelity: 75%

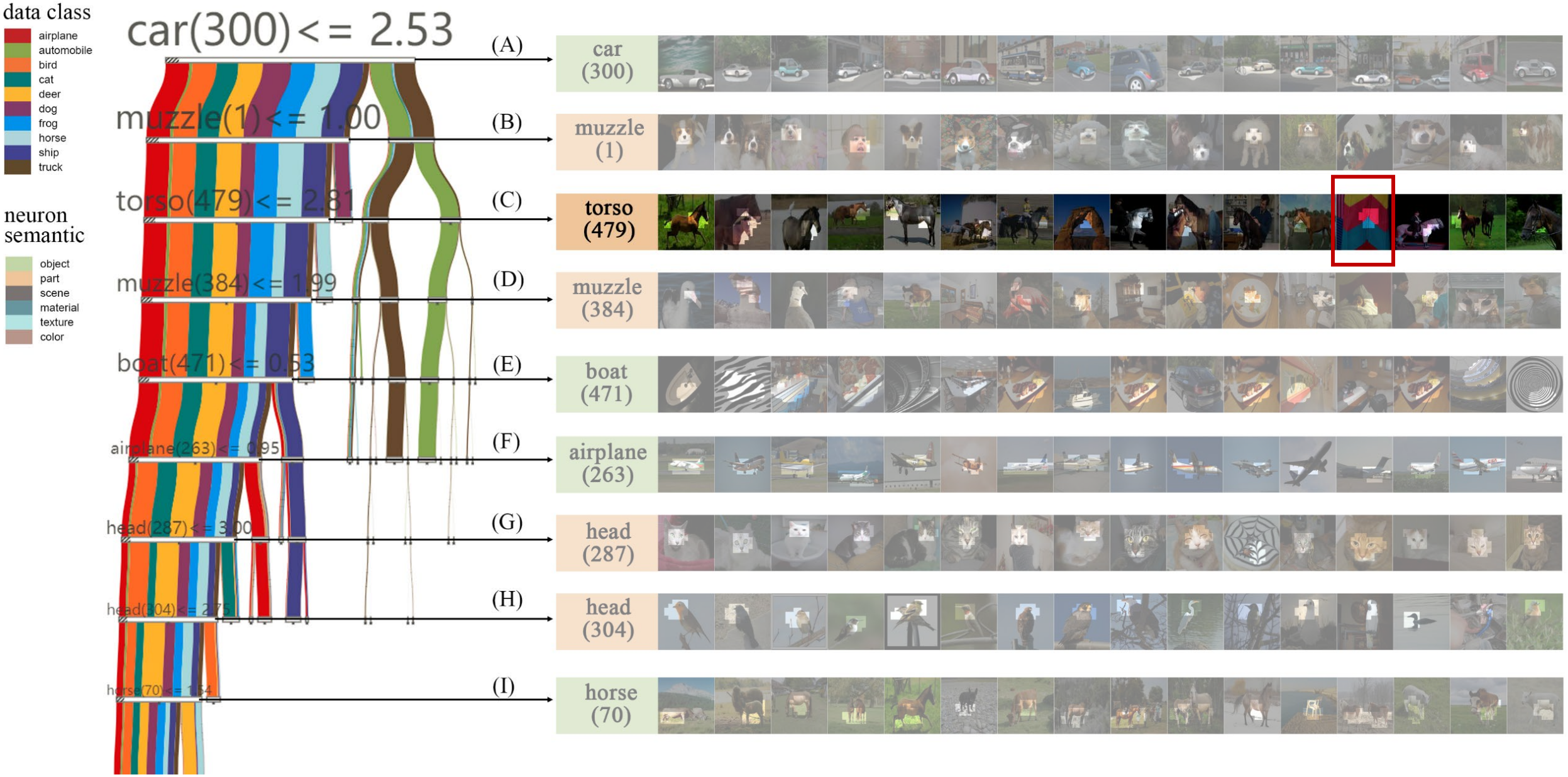
# Use Case 1: Interpreting Surrogate Decision Tree



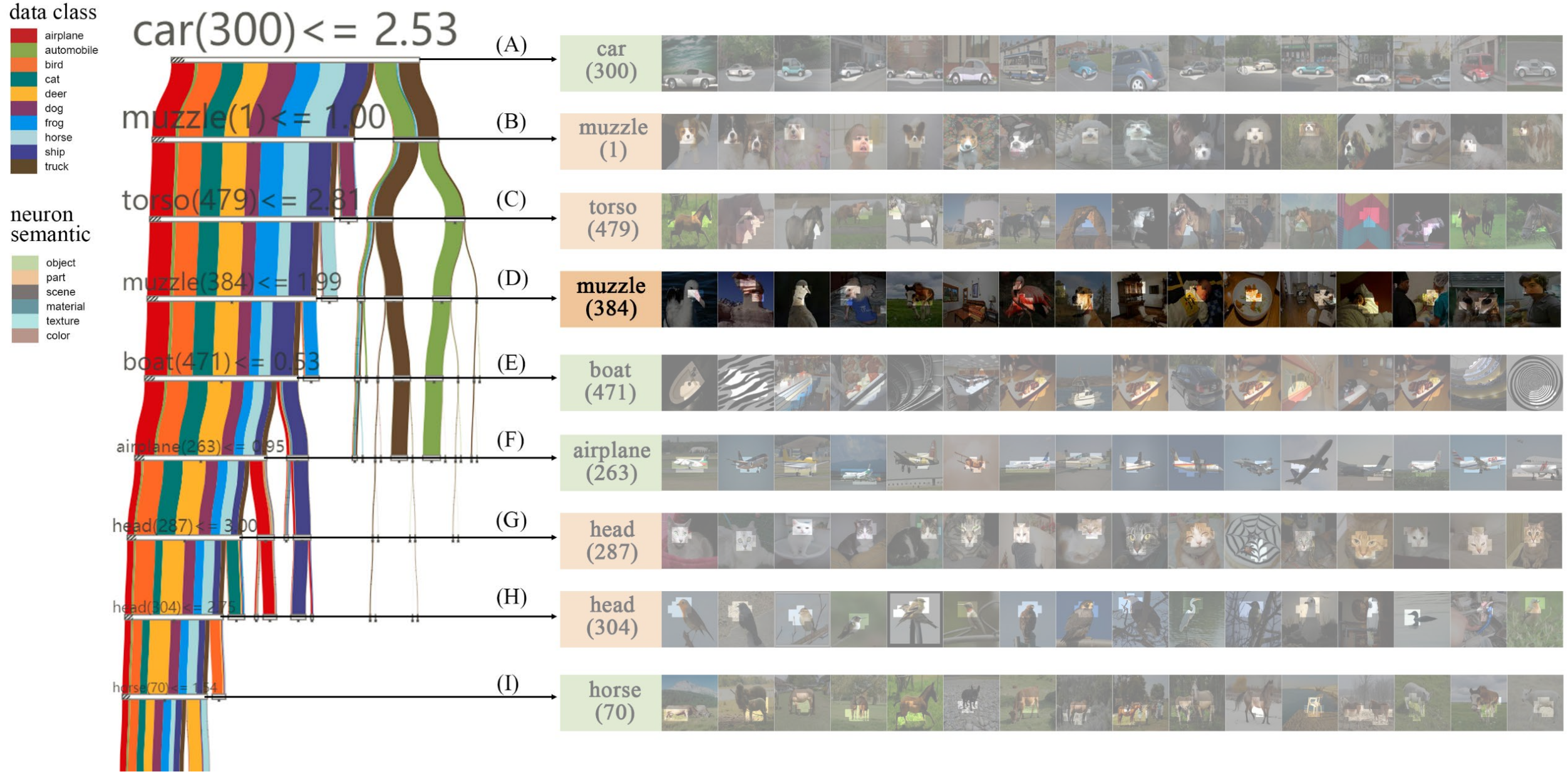
# Use Case 1: Interpreting Surrogate Decision Tree



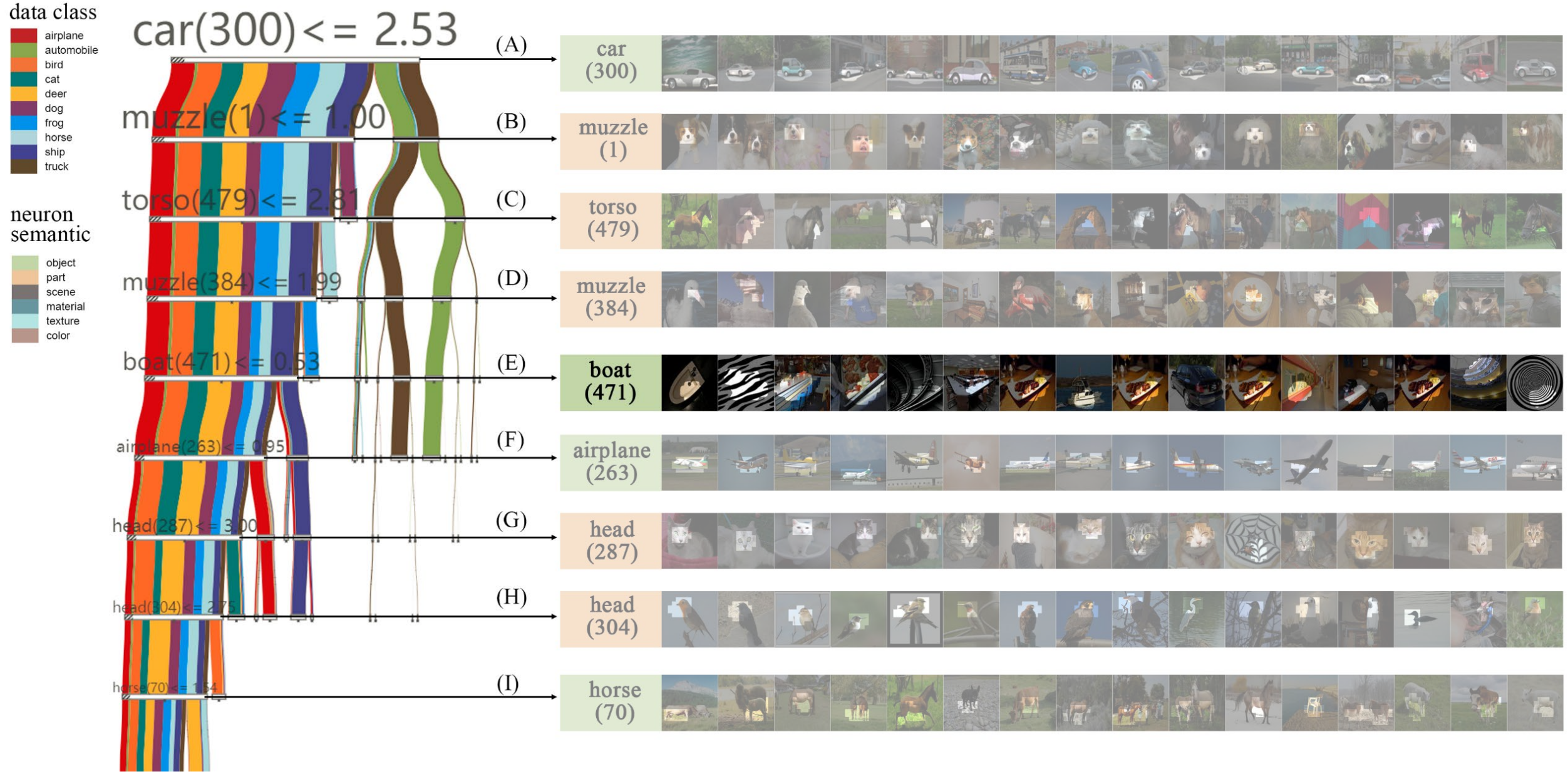
# Use Case 1: Interpreting Surrogate Decision Tree



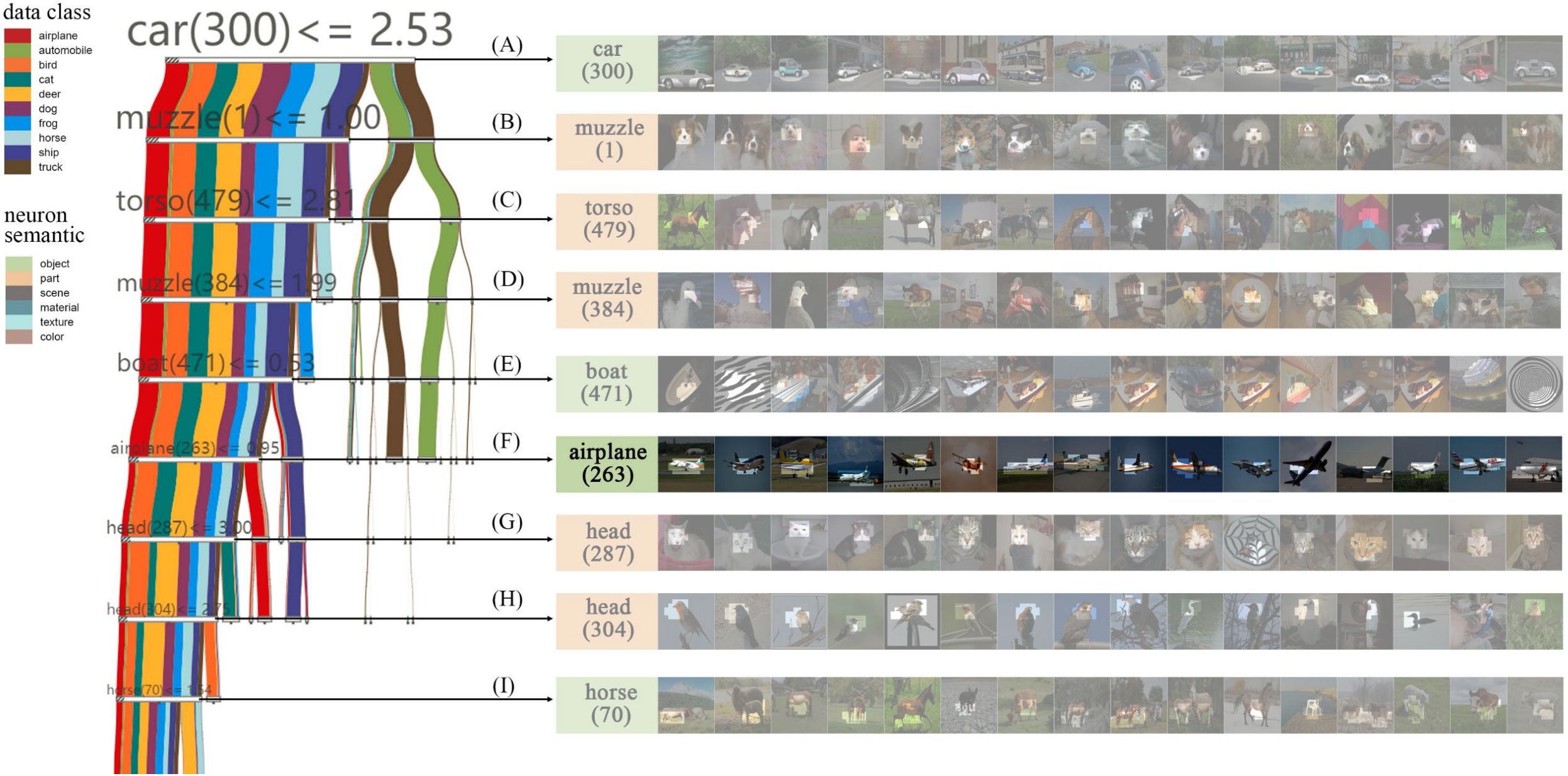
# Use Case 1: Interpreting Surrogate Decision Tree



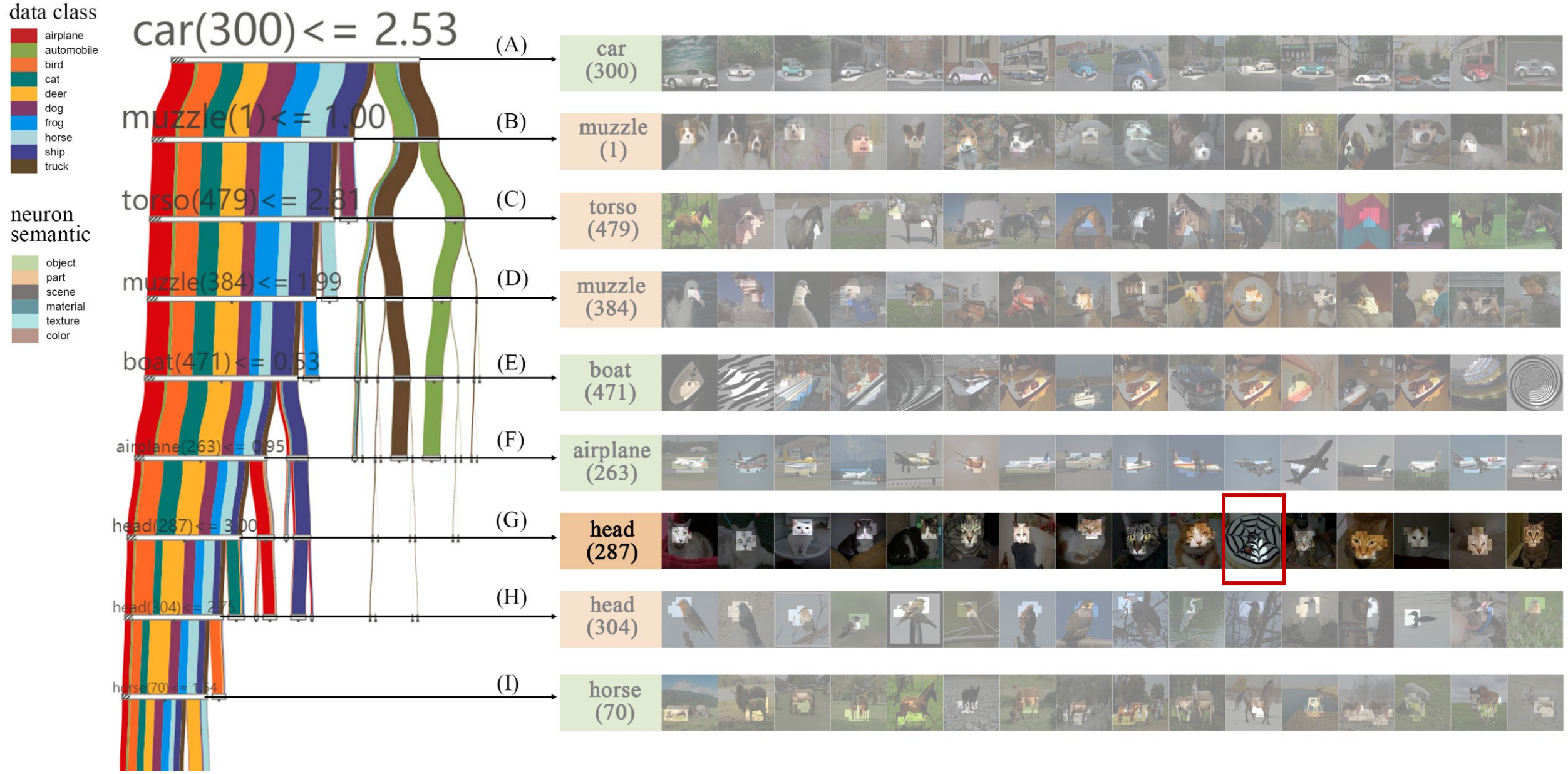
# Use Case 1: Interpreting Surrogate Decision Tree



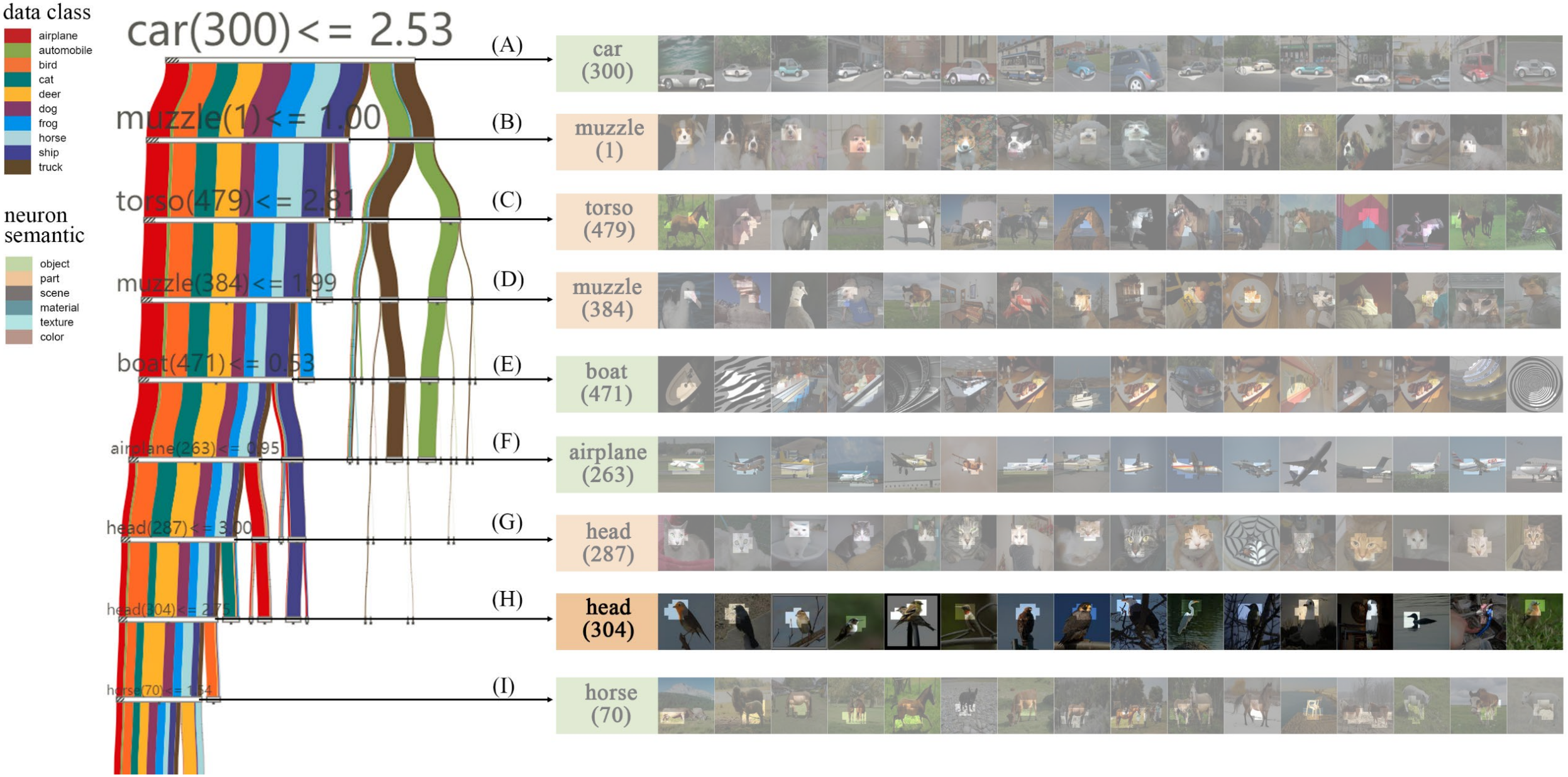
# Use Case 1: Interpreting Surrogate Decision Tree



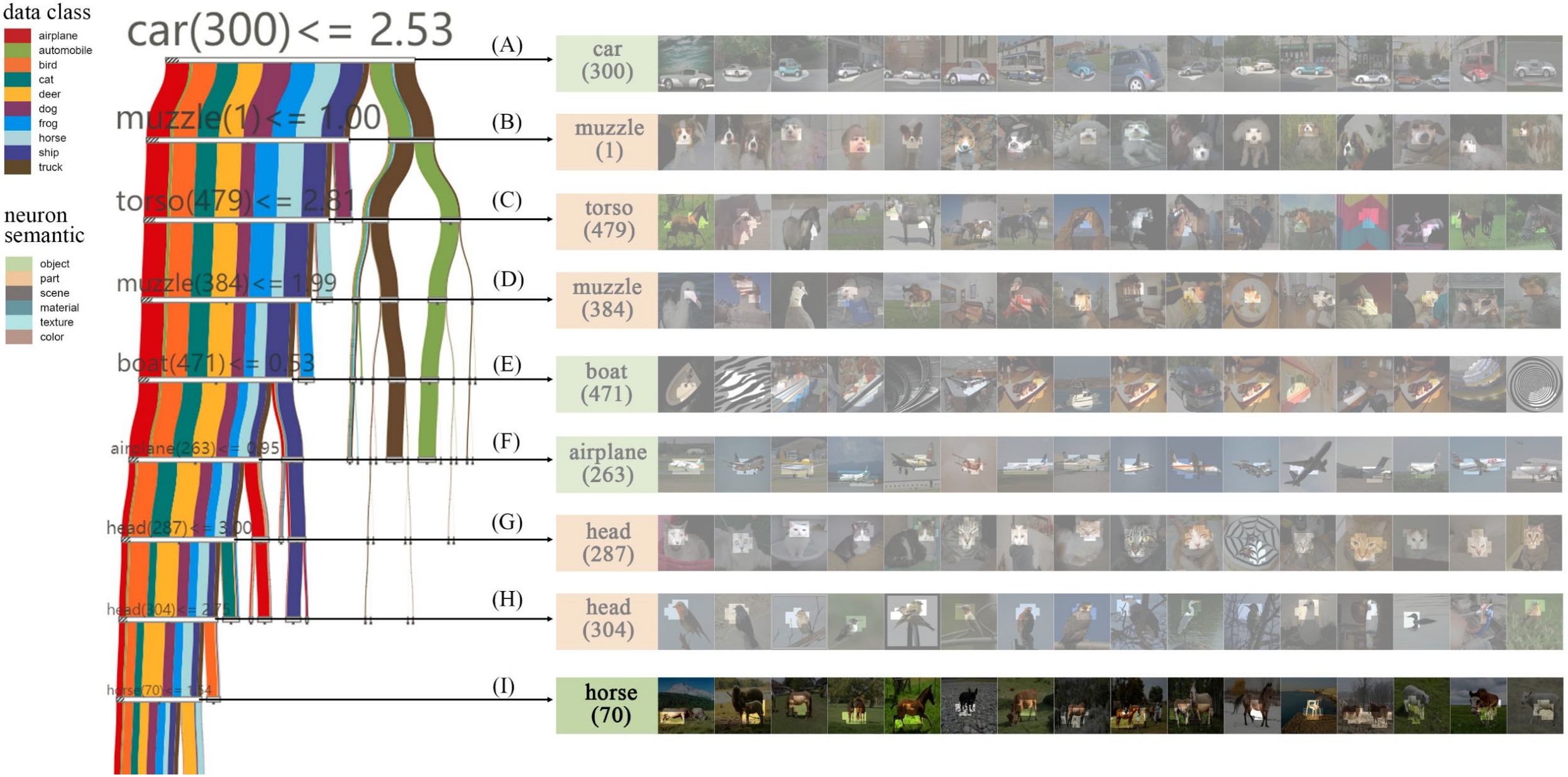
# Use Case 1: Interpreting Surrogate Decision Tree

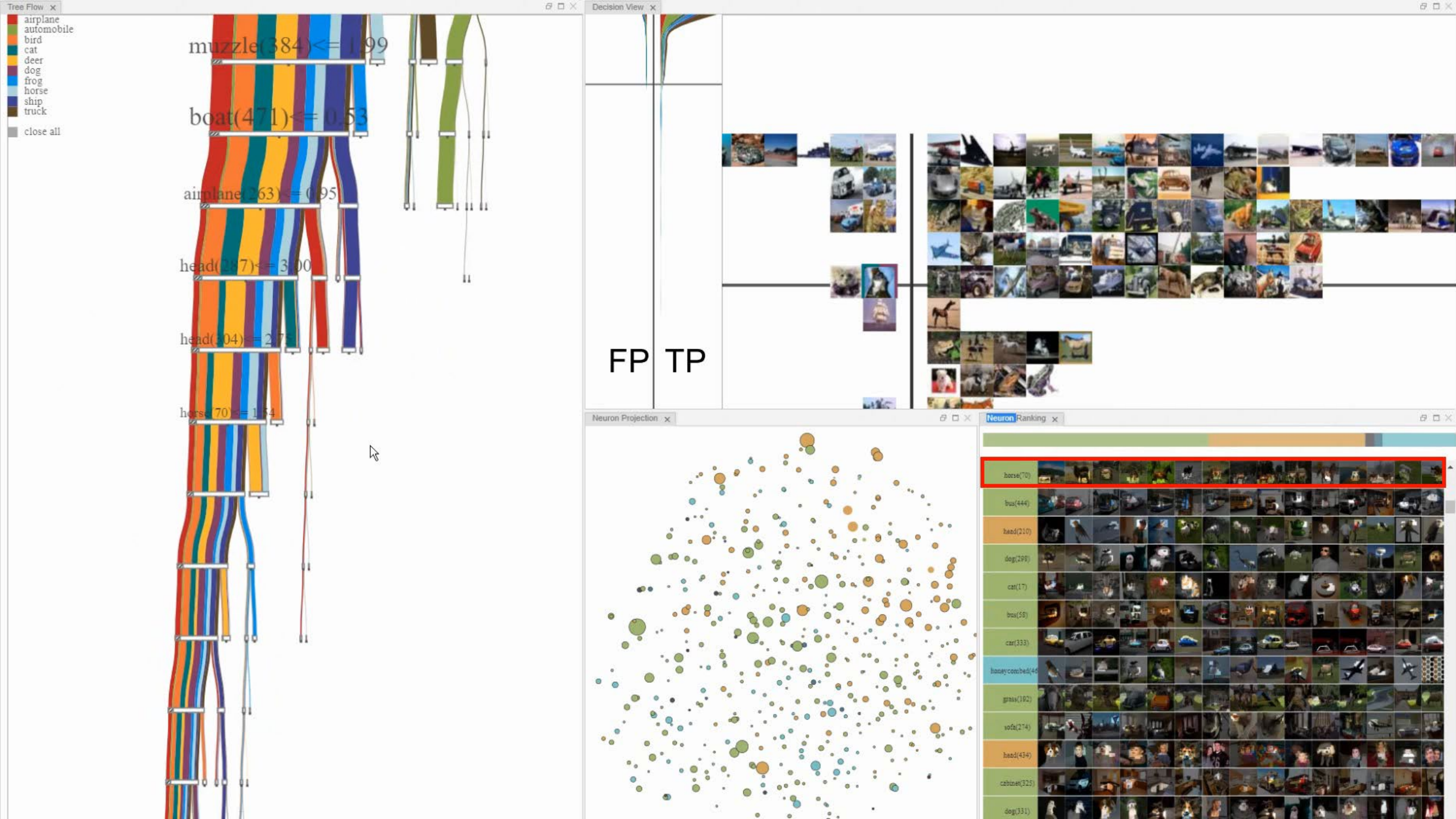


# Use Case 1: Interpreting Surrogate Decision Tree

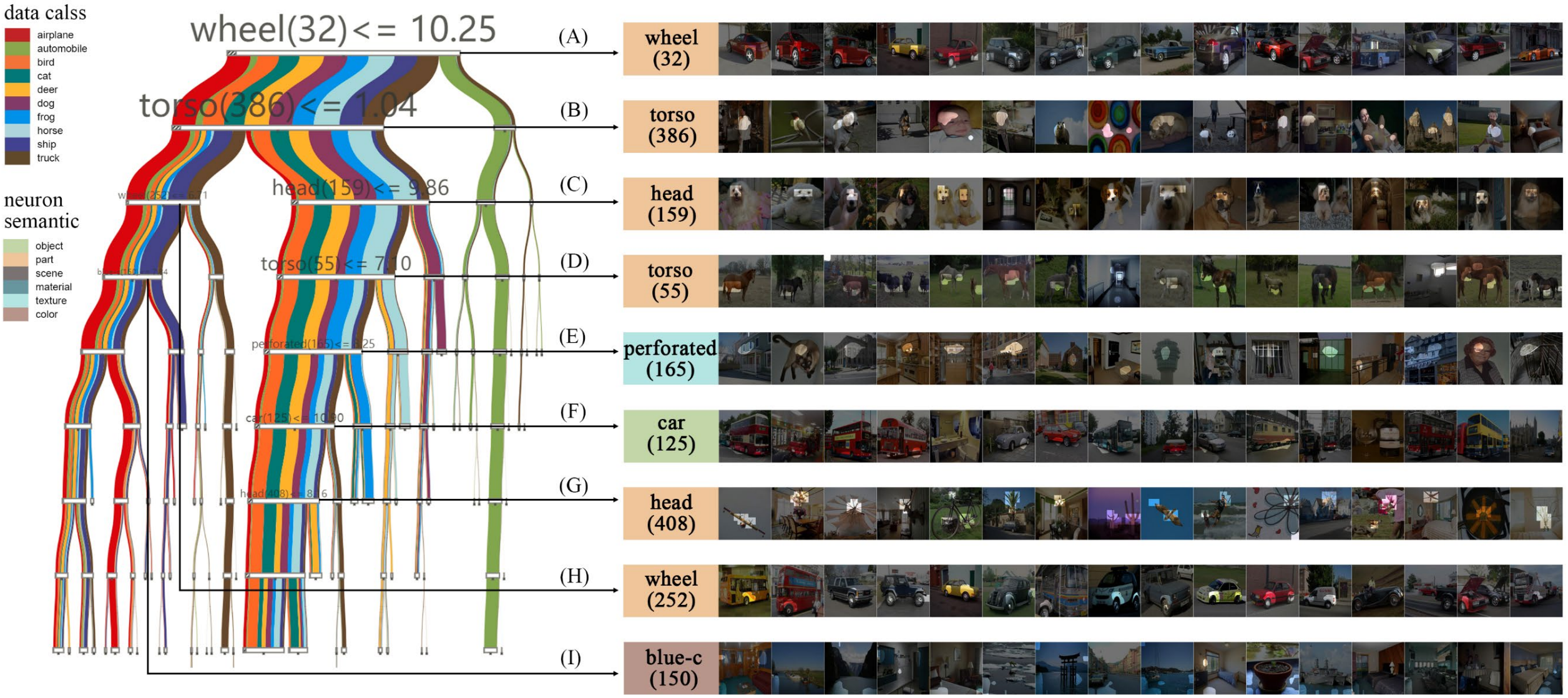


# Use Case 1: Interpreting Surrogate Decision Tree




Tree Flow
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck
close all
muzzle(384)<= 1.99
boat(471)<= 0.53
airplane(263)<= 0.95
head(287)<= 3.00
head(304)<= 2.75
horse(70)<= 1.54
Decision View
FP
TP
Neuron Projection
Neuron Ranking
horse(70)
bus(444)
head(210)
dog(298)
cat(17)
bus(58)
car(333)
honeycombed(46
grass(192)
sofa(274)
head(434)
cabinet(325)
dog(331)

# Use Case 2: Comparing Surrogate Decision Trees



# Conclusion & Discussions

## 1 Conclusion

- New strategy to convert CNNs to surrogate decision trees
- CNN2DT, a visual analytics system
- Use cases and user study

## 2 Potential Users

- Non-DL experts & ML practitioners

## 3 Limitations & Future Work

- Scalability for color encoding, number of classes
- Dependent on semantic database
- Multi-semantics of each neuron | Non-interpretable neurons
- Further evaluations on other architectures

Thanks