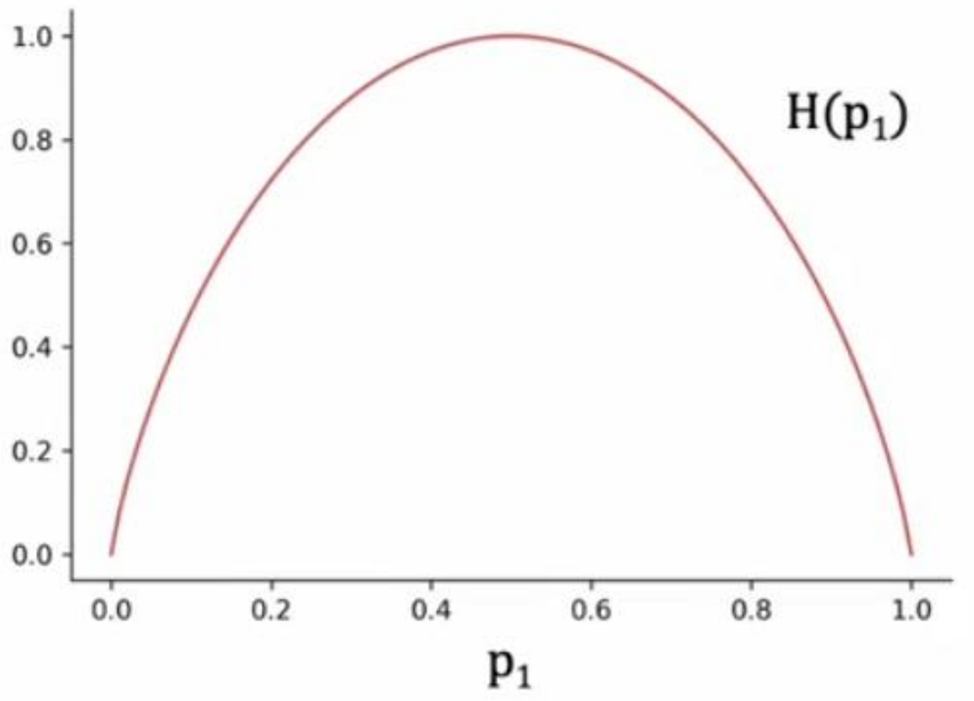
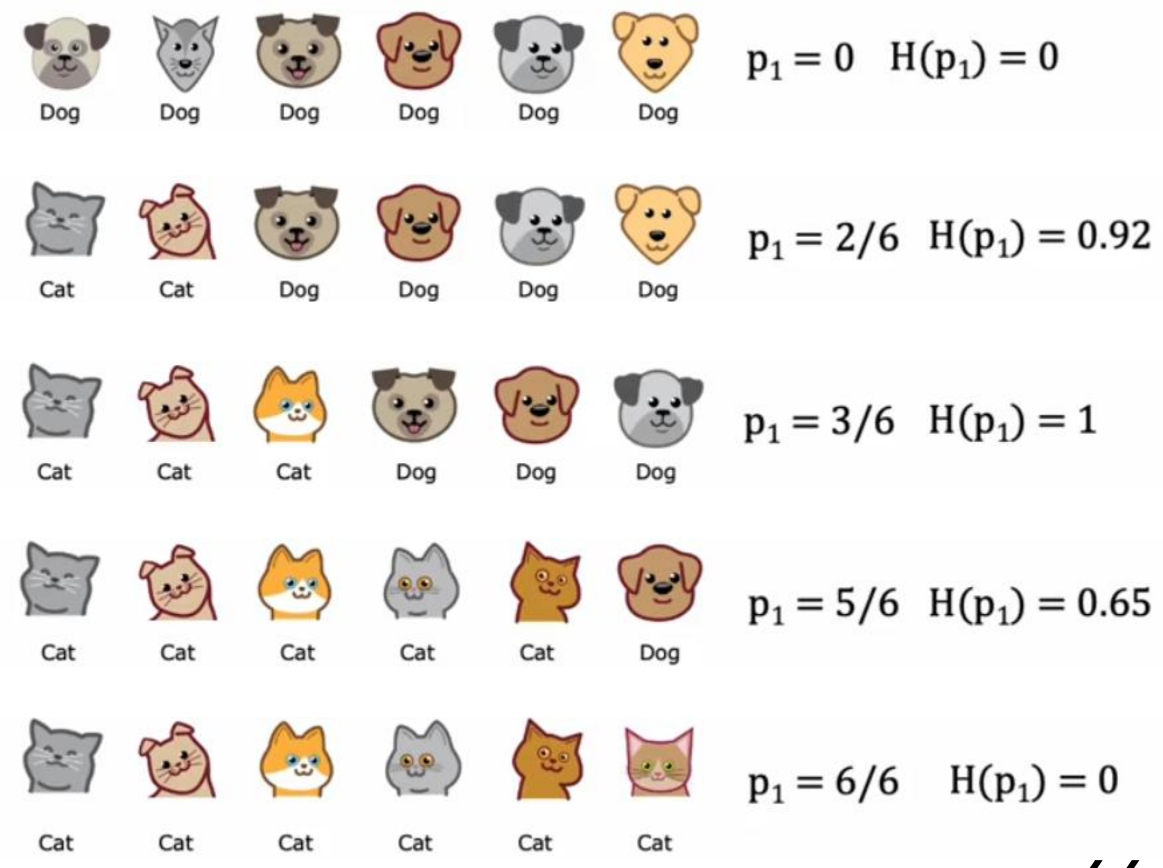
DECISION TREE LEARNING

Entropy (measure of impurity)

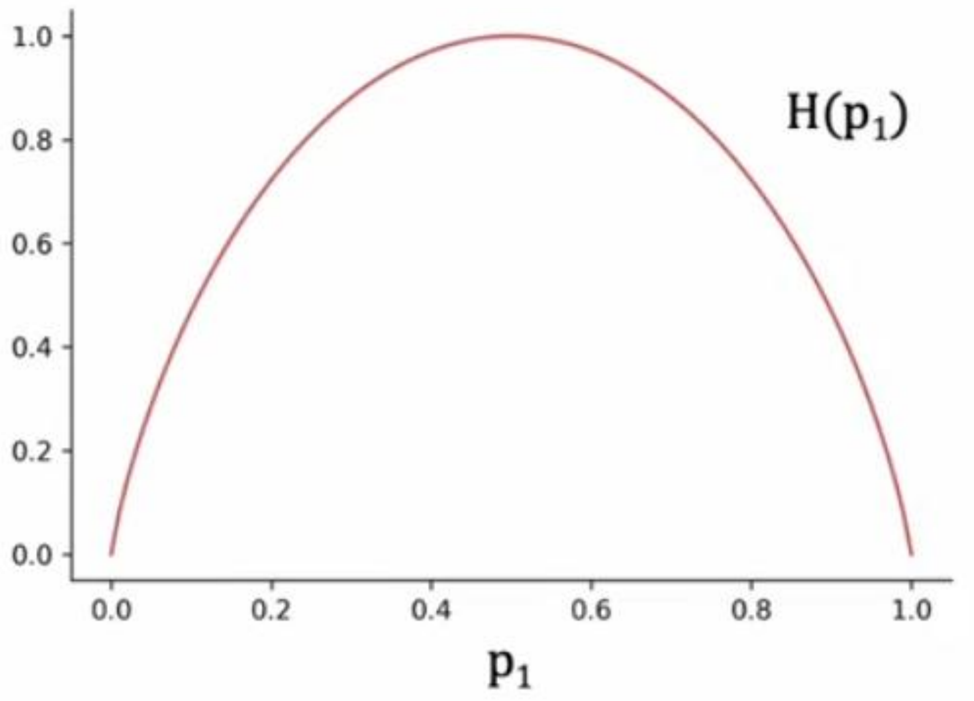
p_1 = fraction of examples that are cats





Entropy (measure of impurity)

p_1 = fraction of examples that are cats

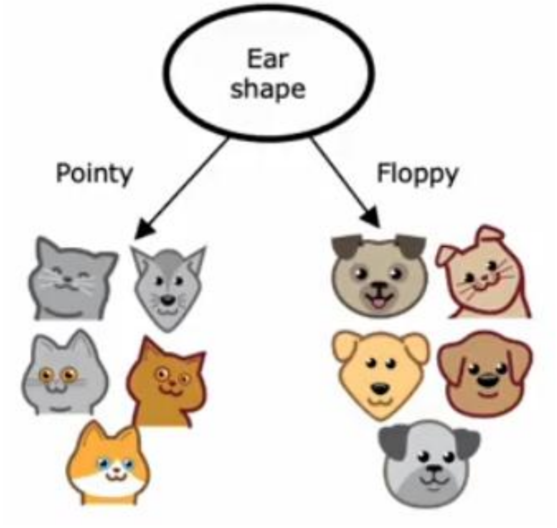


$$p_0 = 1 - p_1$$

$$\begin{aligned} H(p_1) &= -p_1 log_2(p_1) - p_0 log_2(p_0) \\ &= -p_1 \, log_2(p_1) - (1 - p_1) log_2(1 - p_1) \end{aligned}$$

Genny criteria can also be used in place of Entropy

Choosing a Split

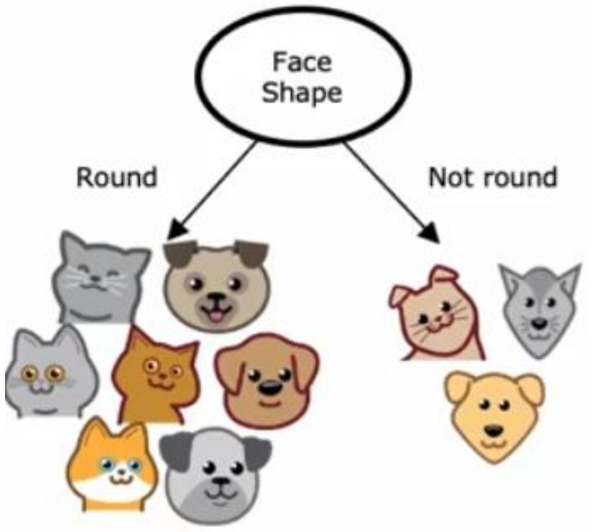


$p_1 = 4/5 = 0.8 \quad p_1 = 1/5 = 0.2$

$H(0.8) = 0.72 \quad H(0.2) = 0.72$

$$H(0.5) - \left(\frac{5}{10}H(0.8) + \frac{5}{10}H(0.2)\right)$$

$= 0.28$

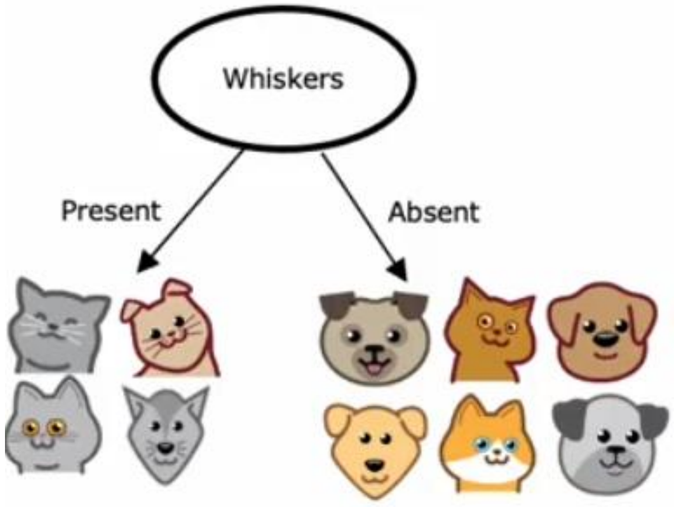


$p_1 = 4/7 = 0.57 \quad p_1 = 1/3 = 0.33$

$H(0.57) = 0.99 \quad H(0.33) = 0.92$

$$H(0.5) - \left(\frac{7}{10}H(0.57) + \frac{3}{10}H(0.33)\right)$$

$= 0.03$



$p_1 = 3/4 = 0.75 \quad p_1 = 2/6 = 0.33$

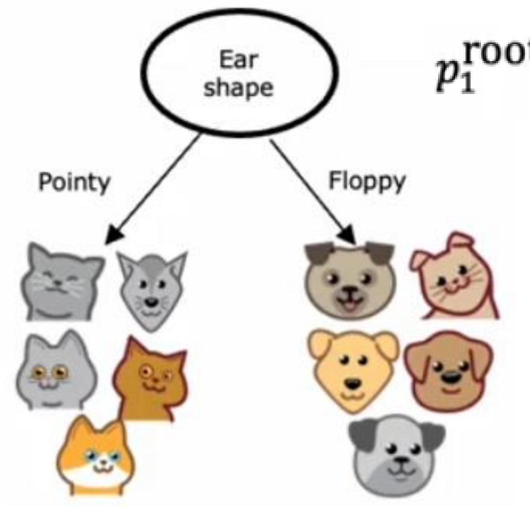
$H(0.75) = 0.81 \quad H(0.33) = 0.92$

$$H(0.5) - \left(\frac{4}{10}H(0.75) + \frac{6}{10}H(0.33)\right)$$

$= 0.12$

Reduction in Entropy = Information Gain

Information Gain



$$p_1^{\text{root}} = {}^{5}/_{10} = 0.5$$

$p_1^{\text{left}} = {}^{4}/_{5}$ $\quad$ $p_1^{\text{right}} = {}^{1}/_{5}$

$w^{\text{left}} = {}^{5}/_{10}$ $\quad$ $w^{\text{right}} = {}^{5}/_{10}$

Information gain

$$= \mathrm{H}(\mathrm{p}_1^{\text{root}}) \; - \left(w^{\text{left}} H\left(p_1^{\text{left}}\right) + w^{\text{right}} H\left(p_1^{\text{right}}\right)\right)$$

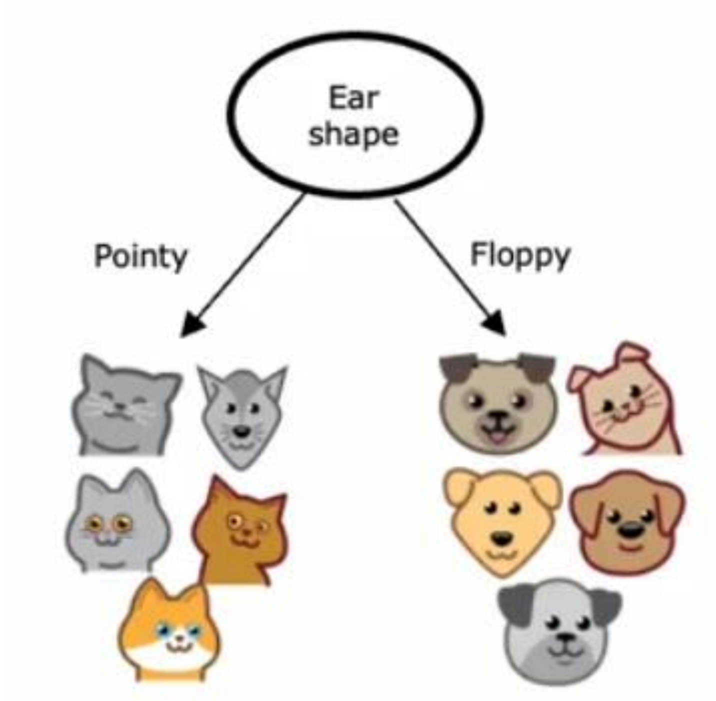
Decision Tree Learning

- Start with all examples at the root node
- Calculate information gain for all possible features, and pick the one with the highest information gain
- Split dataset according to selected feature, and create left and right branches of the tree
- Keep repeating splitting process until stopping criteria is met:
  - When a node is 100% one class
  - When splitting a node will result in the tree exceeding a maximum depth
  - When improvements in purity score are below a threshold
  - When number of examples in a node is below a threshold

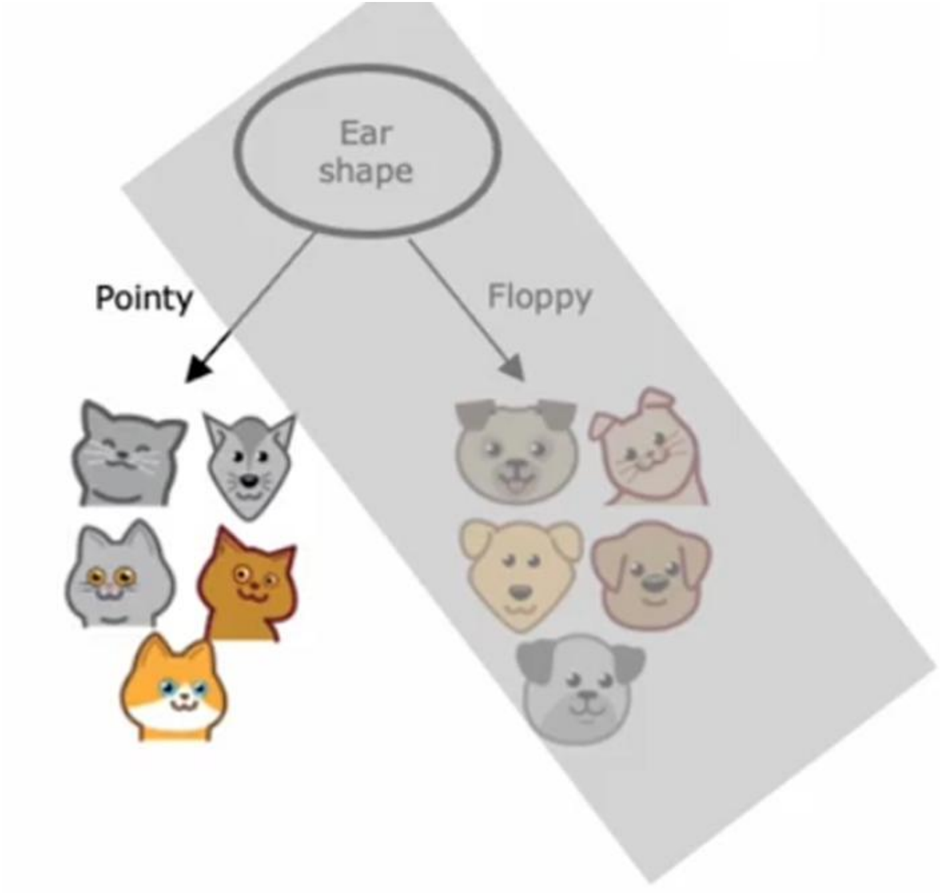
Recursive Splitting

Recursive Splitting

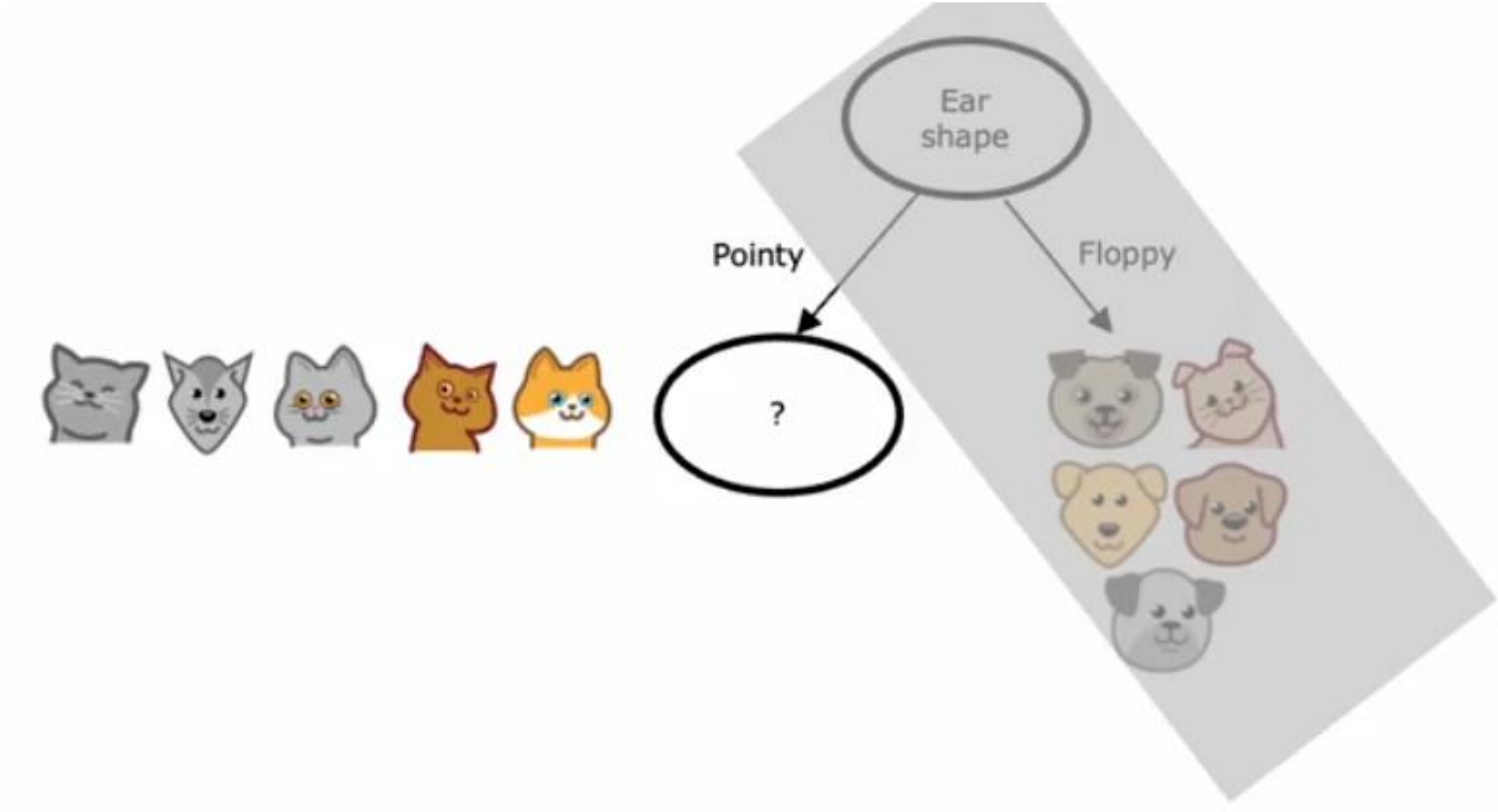
Recursive Splitting



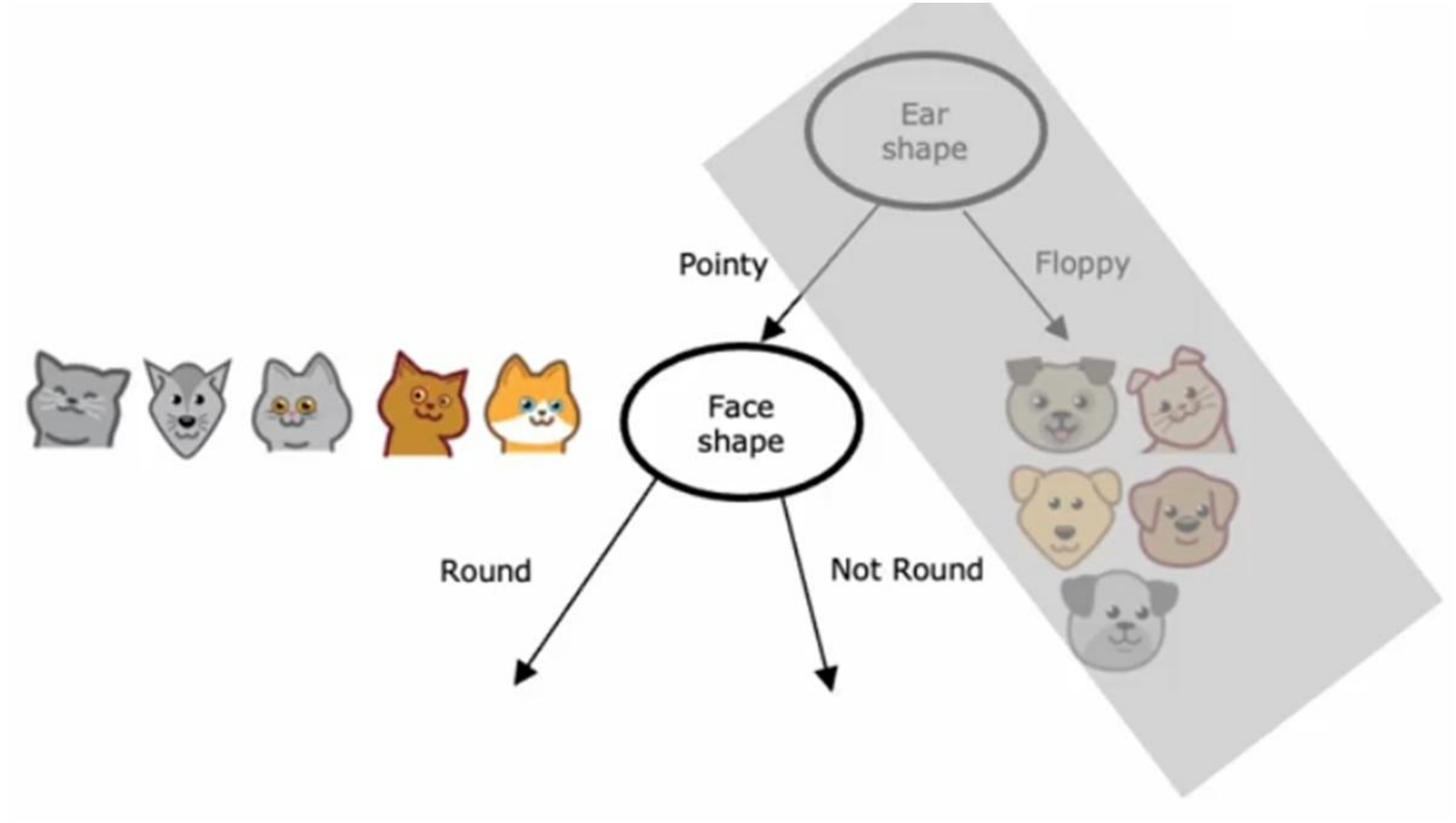
Recursive Splitting



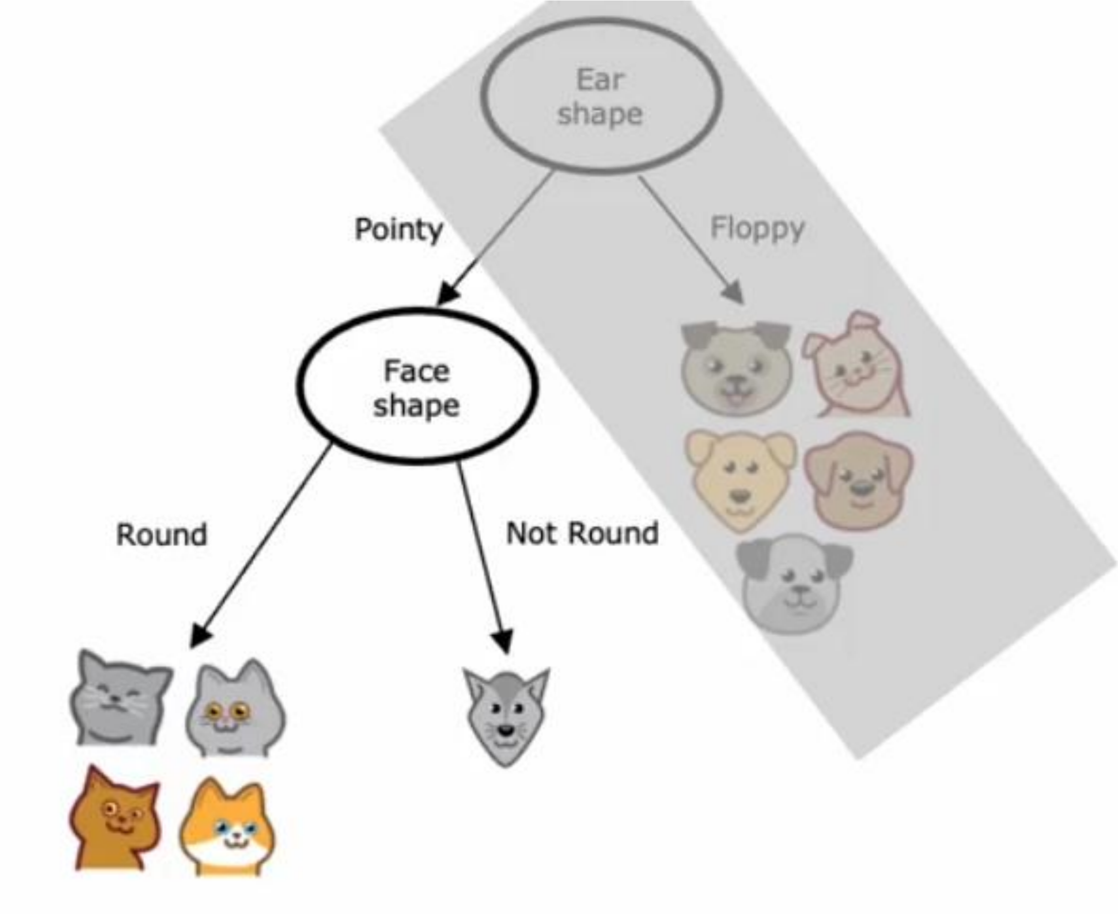
Recursive Splitting



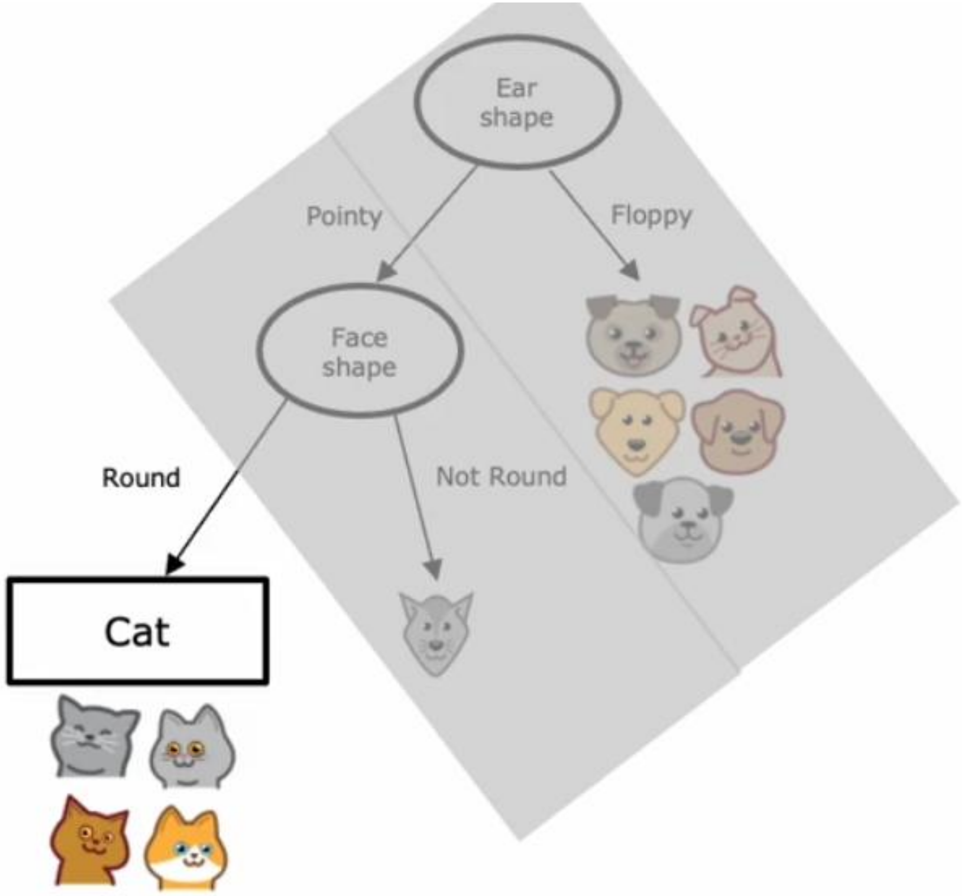
Recursive Splitting



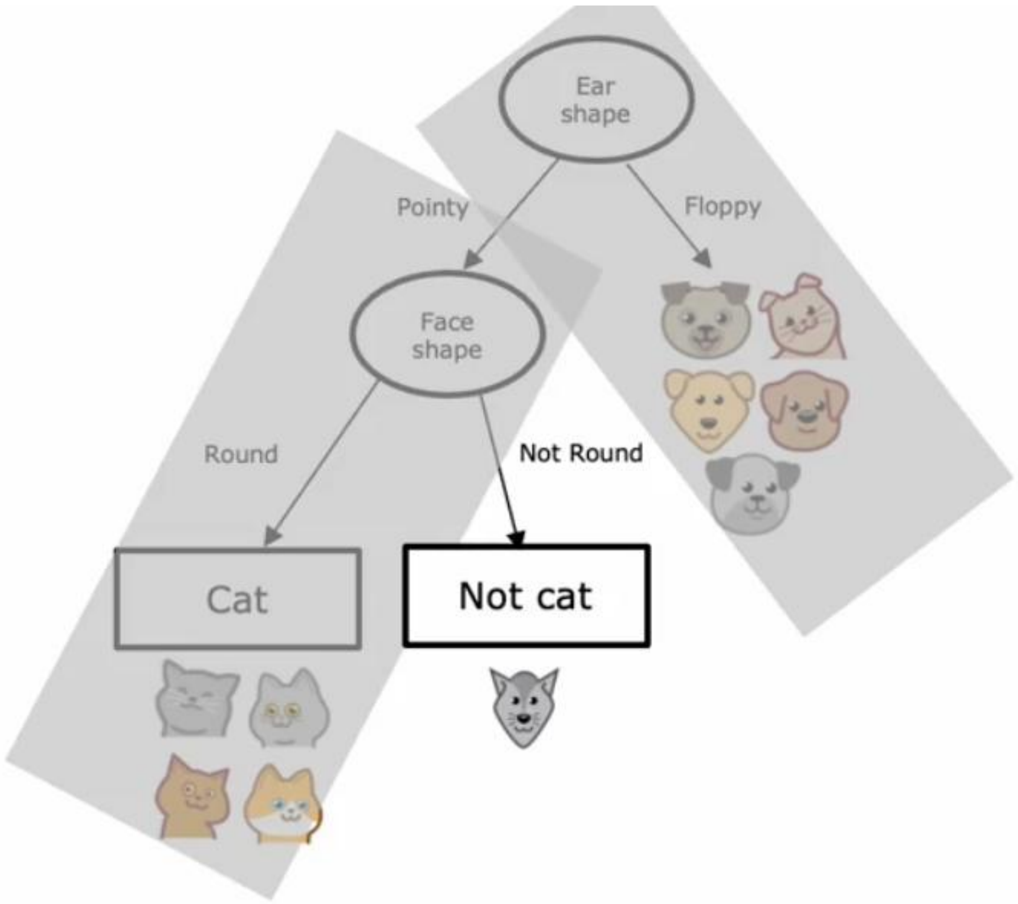
Recursive Splitting



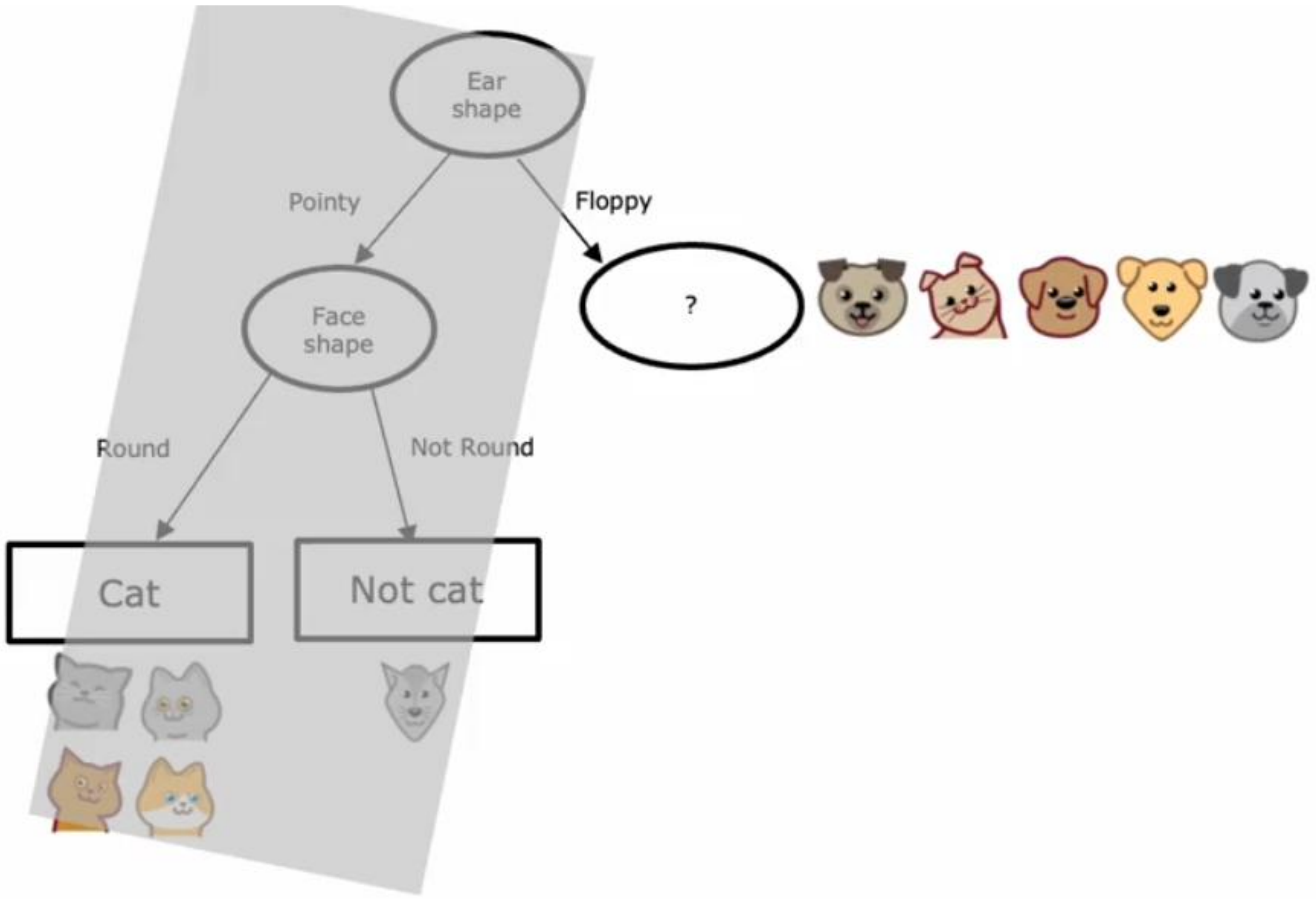
Recursive Splitting



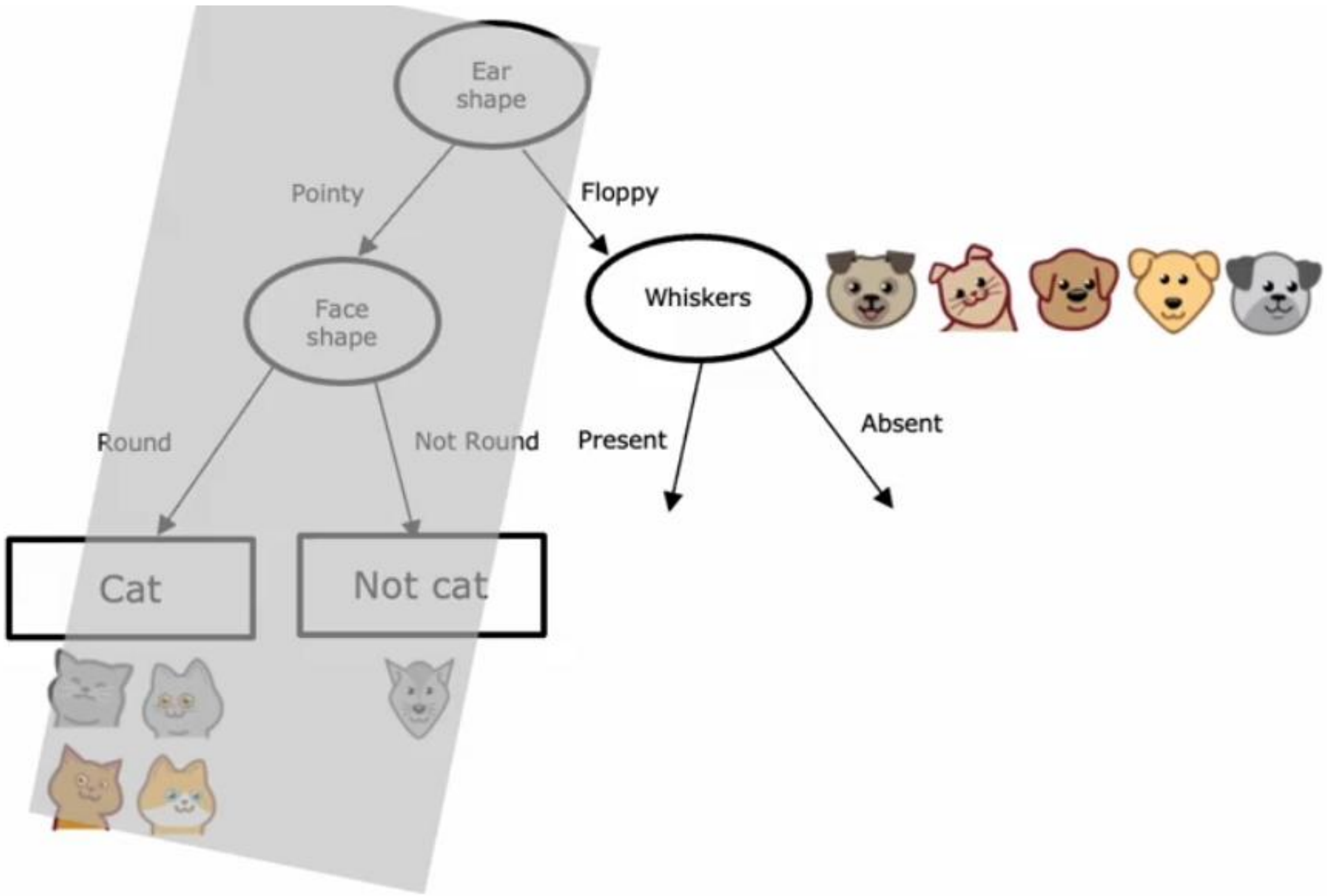
Recursive Splitting



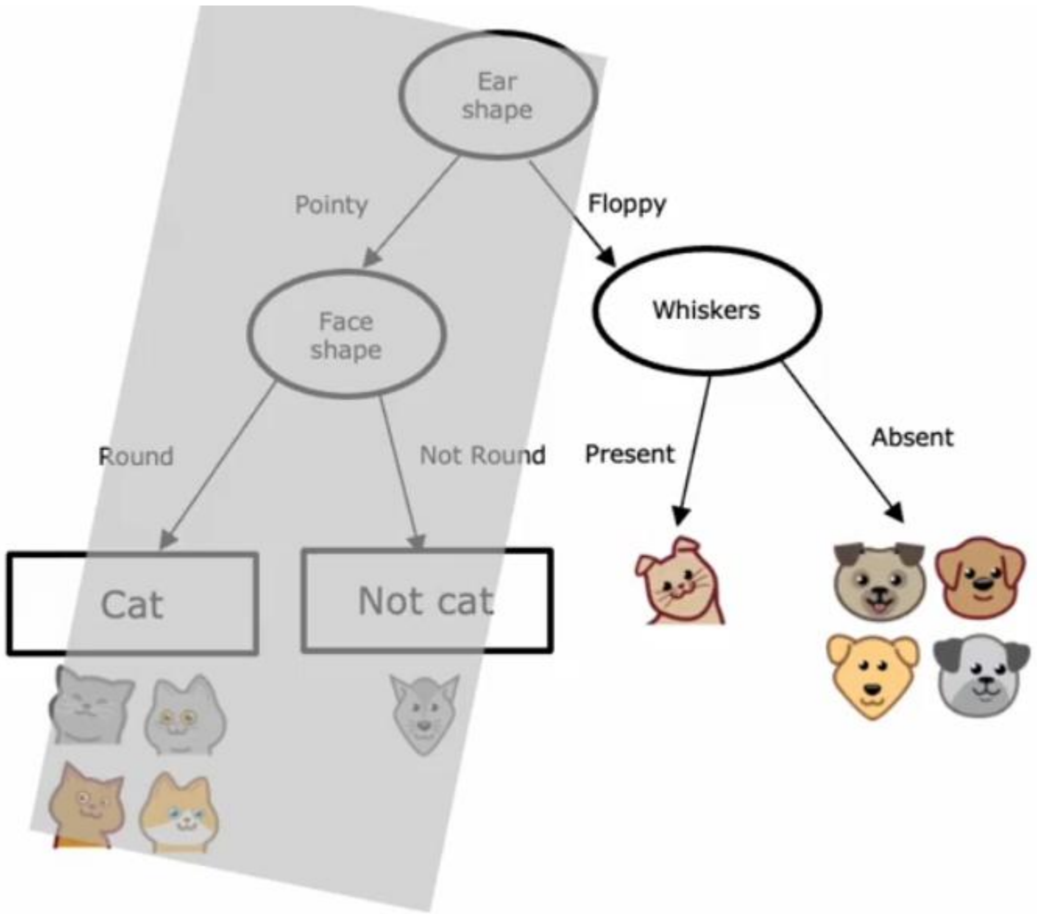
Recursive Splitting



Recursive Splitting



Recursive Splitting



Recursive Splitting