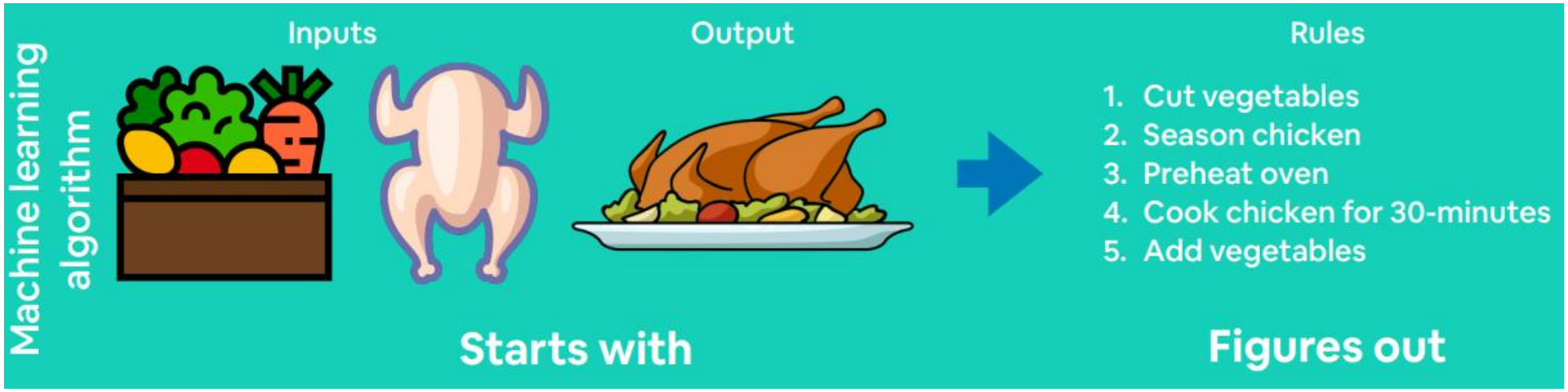
Machine Learning Overview

What is Machine Learning?

“Field of study that gives computers the ability to learn without being explicitly programmed”

Arthur Samuel (1959)

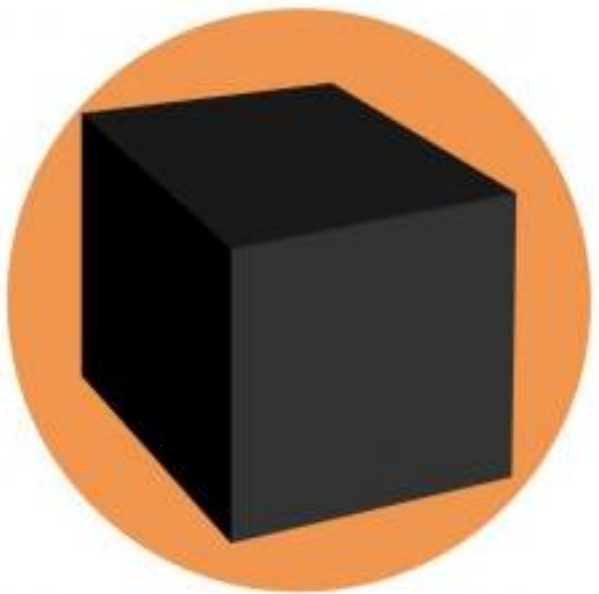
Traditional Programming vs Machine Learning Algorithm



White Box vs Grey Box vs Black Box Model

(Known Internal Code Structure)

(Internal Code Structure Partially Known)

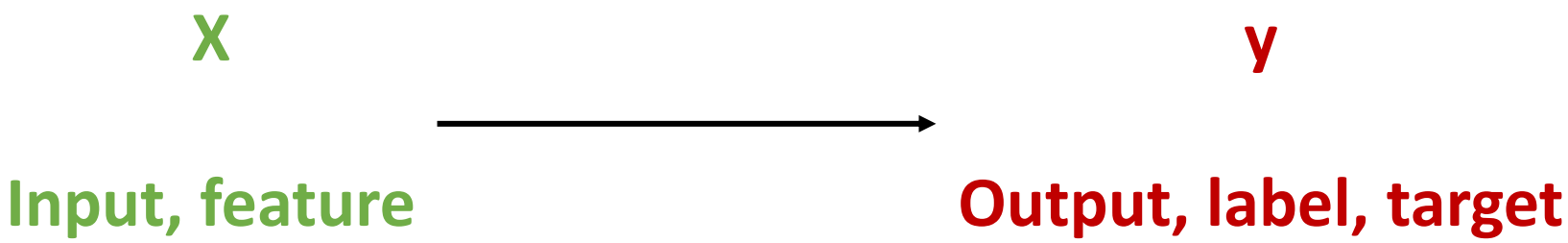


(Unknown Internal Code Structure)

Types of Machine Learning

- **Supervised Learning**
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning



Learns from being given "***right answers***"

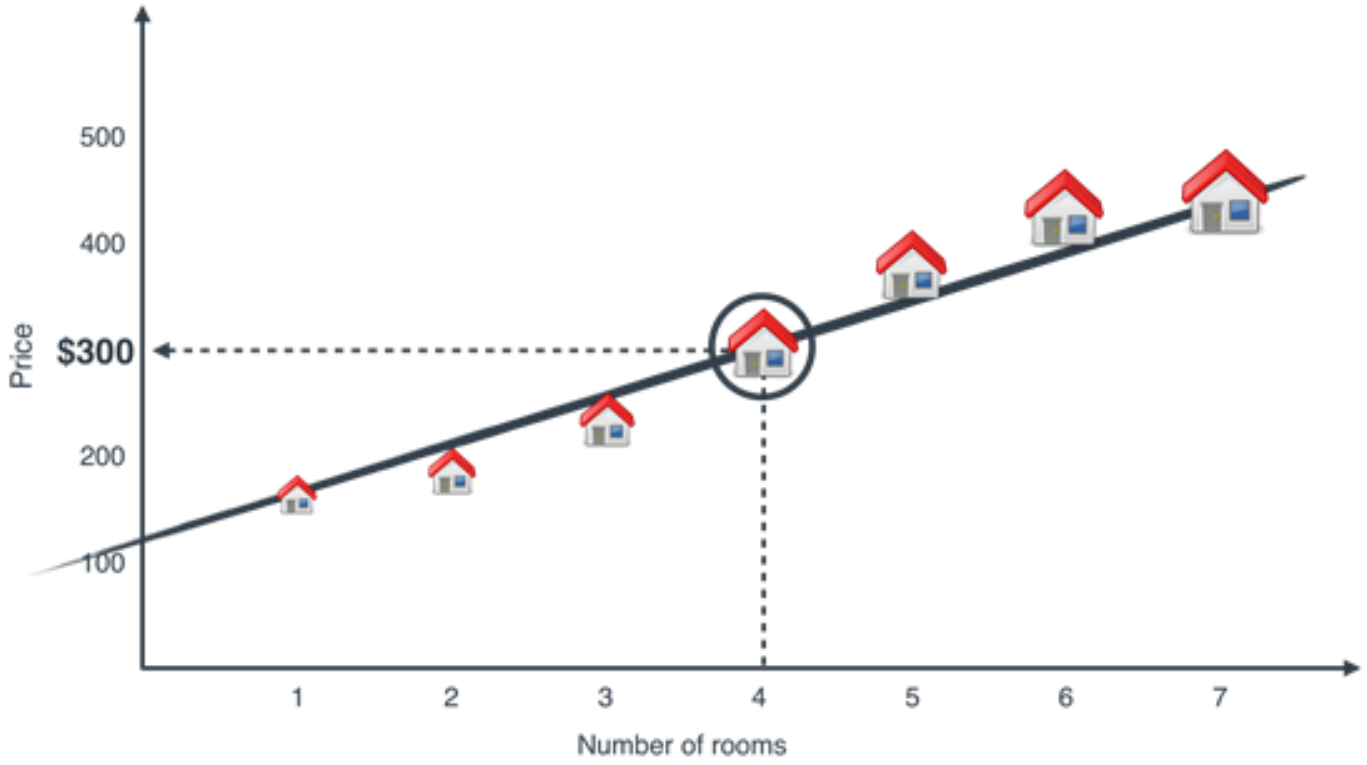
Supervised Learning

Input (X)		Output (y)	Application
Email	→	Spam?	Spam filtering
Audio	→	Text transcripts	Speech recognition
English	→	Spanish	Machine translation
Ad, user info	→	Click? (0/1)	Online advertising
Image, radar info	→	Position of other cars	Self-driving car

Supervised Learning

- Regression
- Classification

Regression



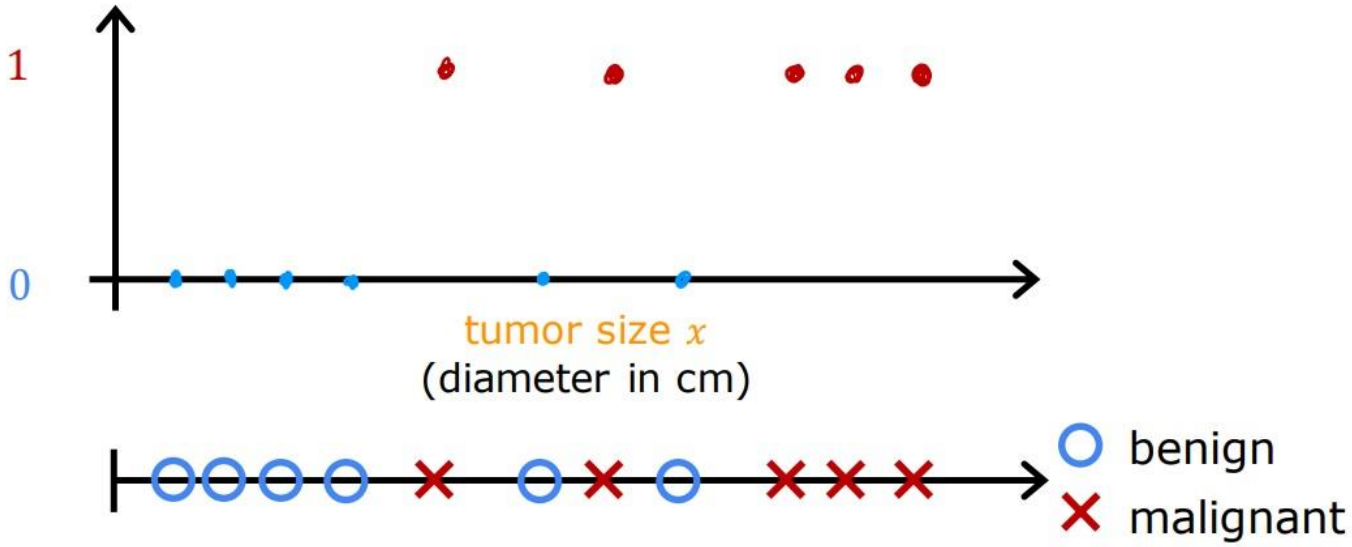
Housing price prediction

Regression

Predict a number

infinitely many possible outputs

Classification



Breast cancer detection

Classification

Predict a categories

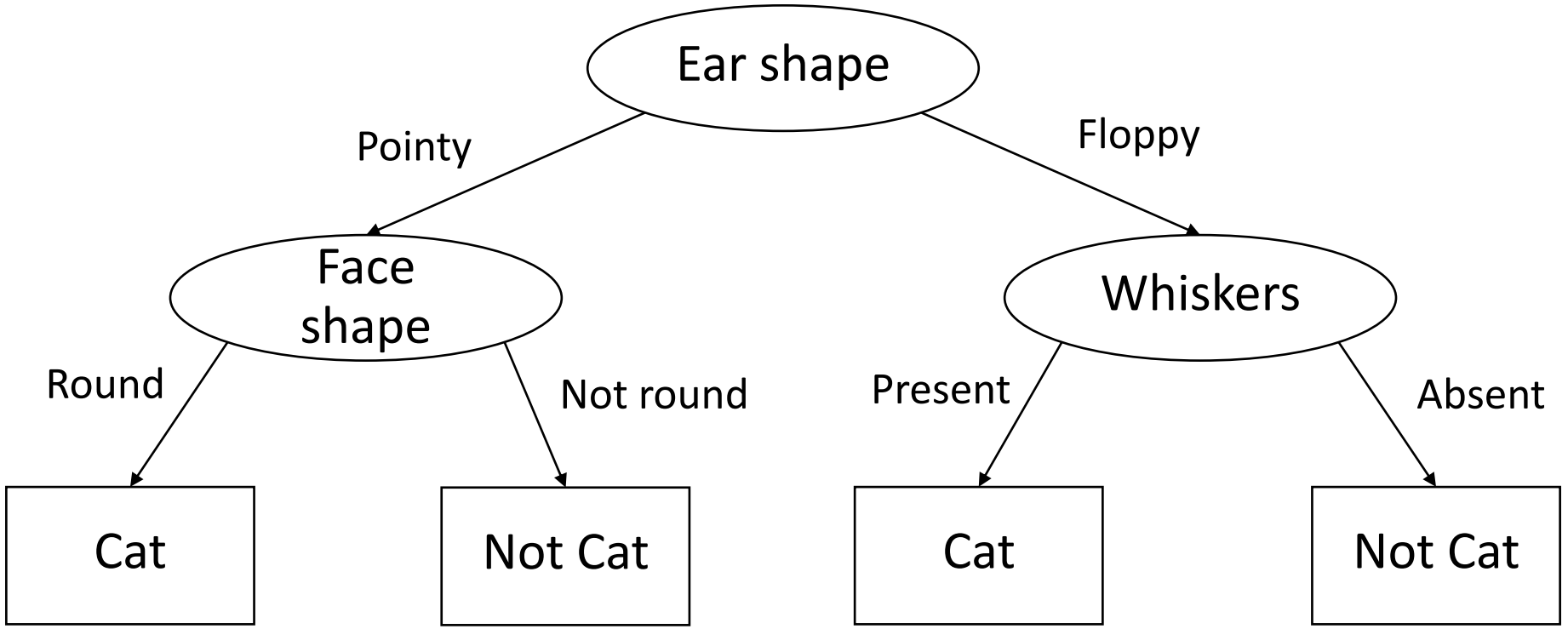
Small number of possible outputs

Decision Tree Model (Part 1)

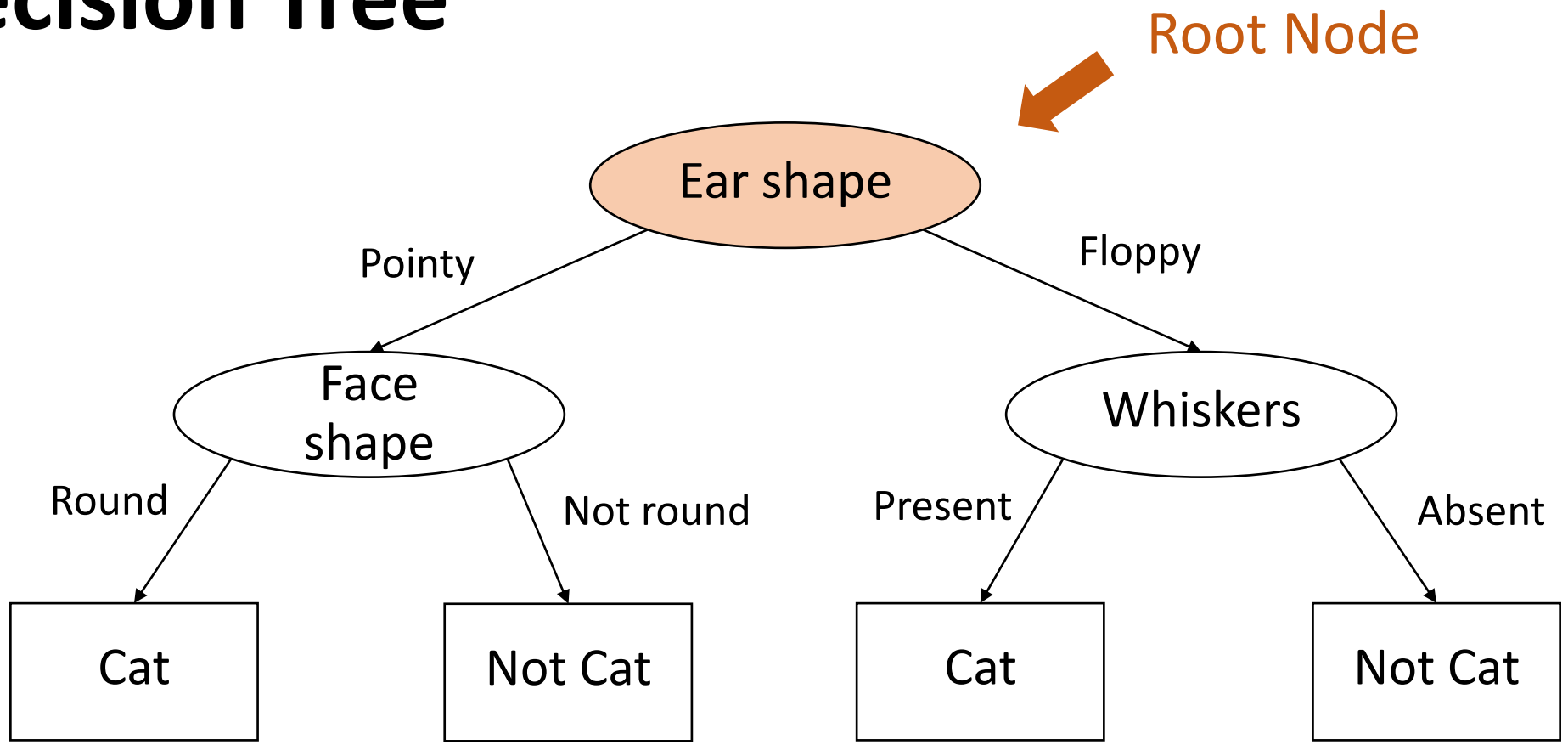
Cat classification example

	Ear shape (x_1)	Face shape (x_2)	Whiskers (x_3)	Cat
	Pointy	Round	Present	1
	Floppy	Not round	Present	1
	Floppy	Round	Absent	0
	Pointy	Not round	Present	0
	Pointy	Round	Present	1
	Pointy	Round	Absent	1
	Floppy	Not round	Absent	0
	Pointy	Round	Absent	1
	Floppy	Round	Absent	0
	Floppy	Round	Absent	0

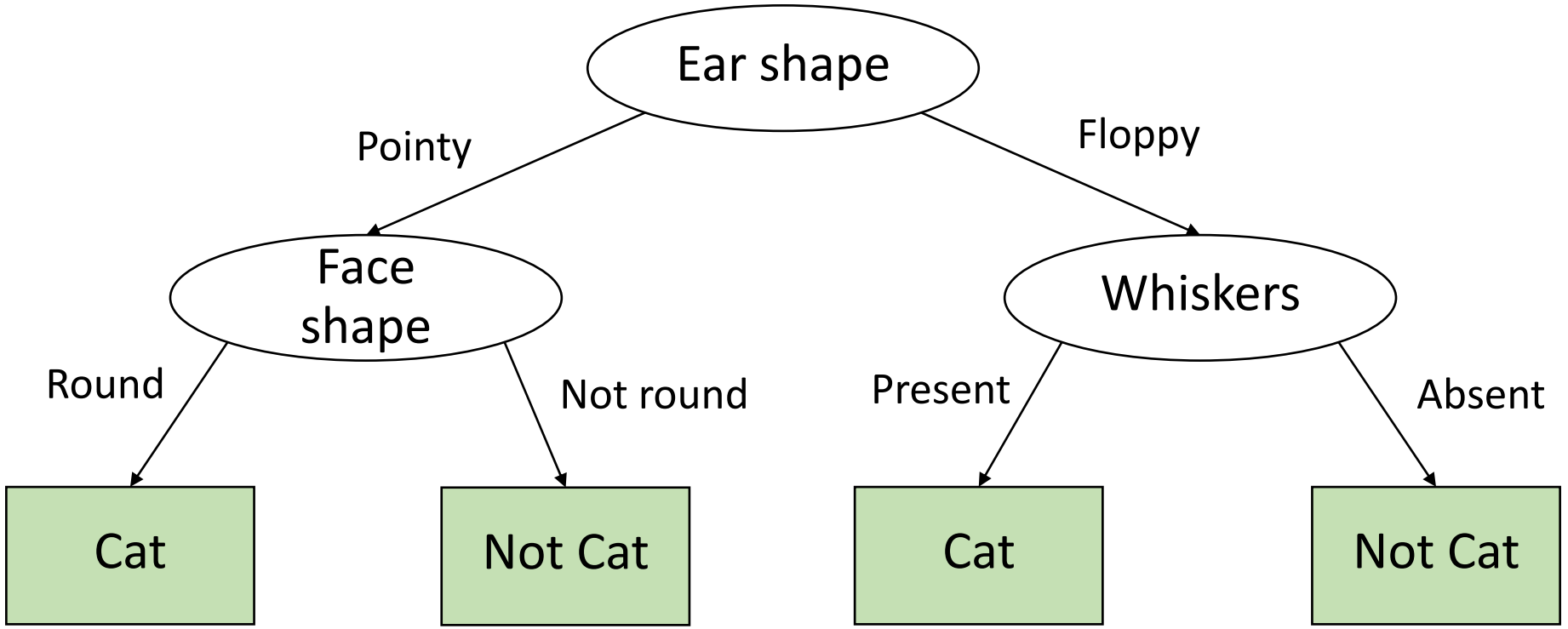
Decision Tree



Decision Tree

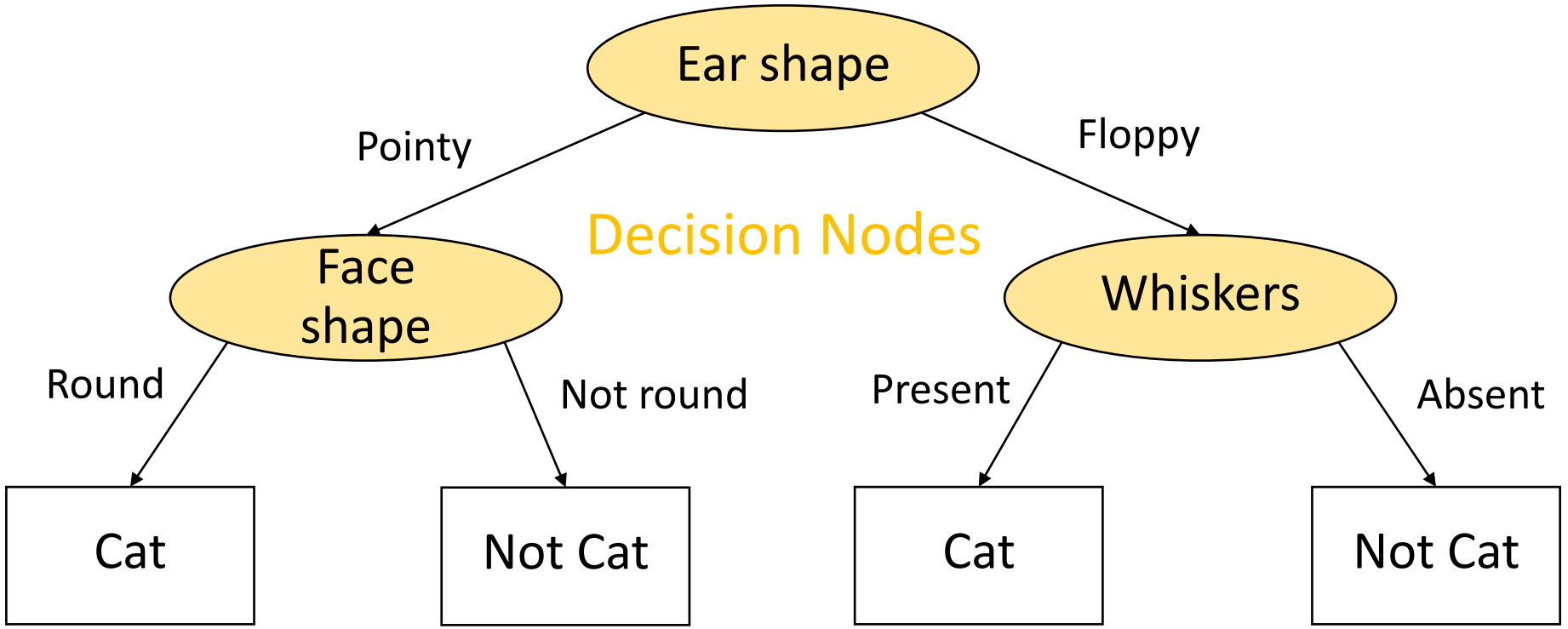


Decision Tree

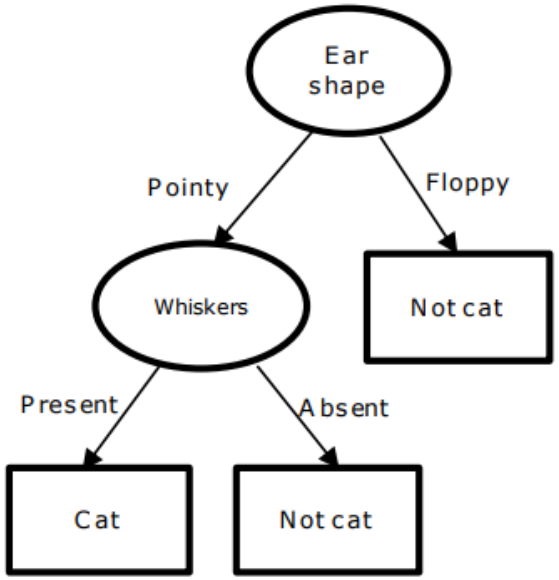


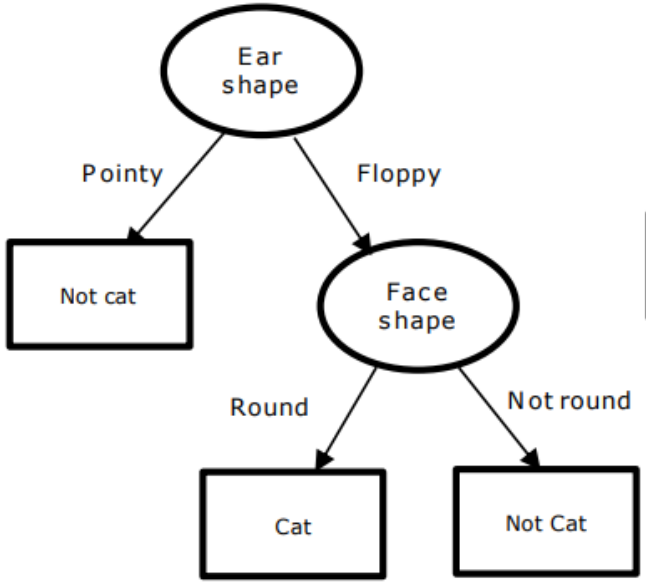
Leaf nodes

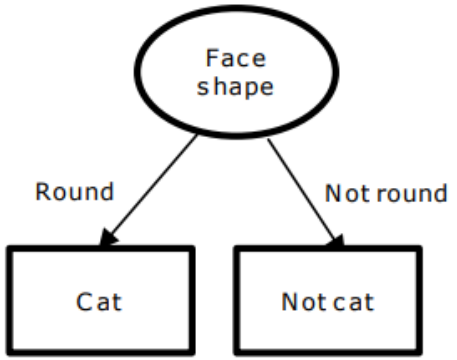
Decision Tree

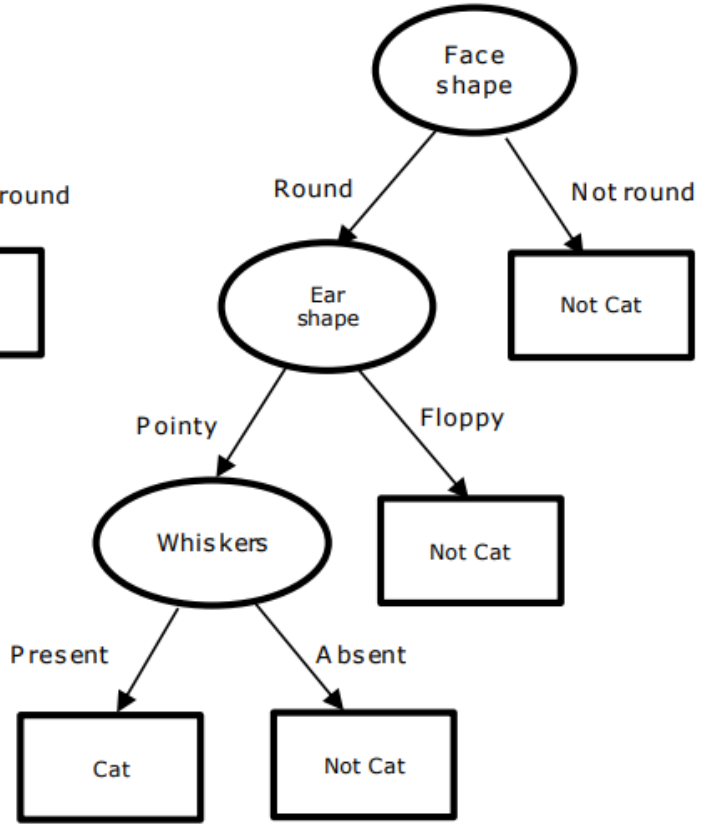


Decision Tree









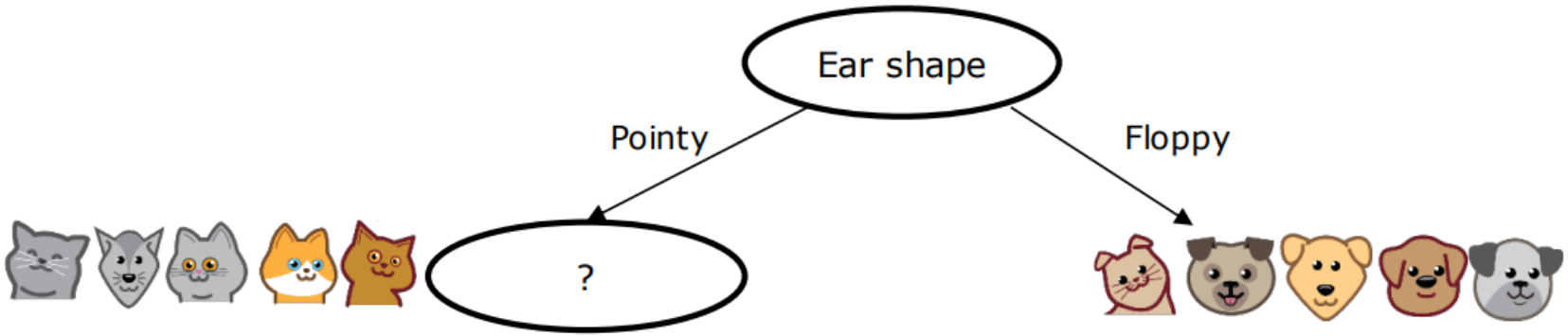
Lab #1

Decision Tree Model (Part 2)

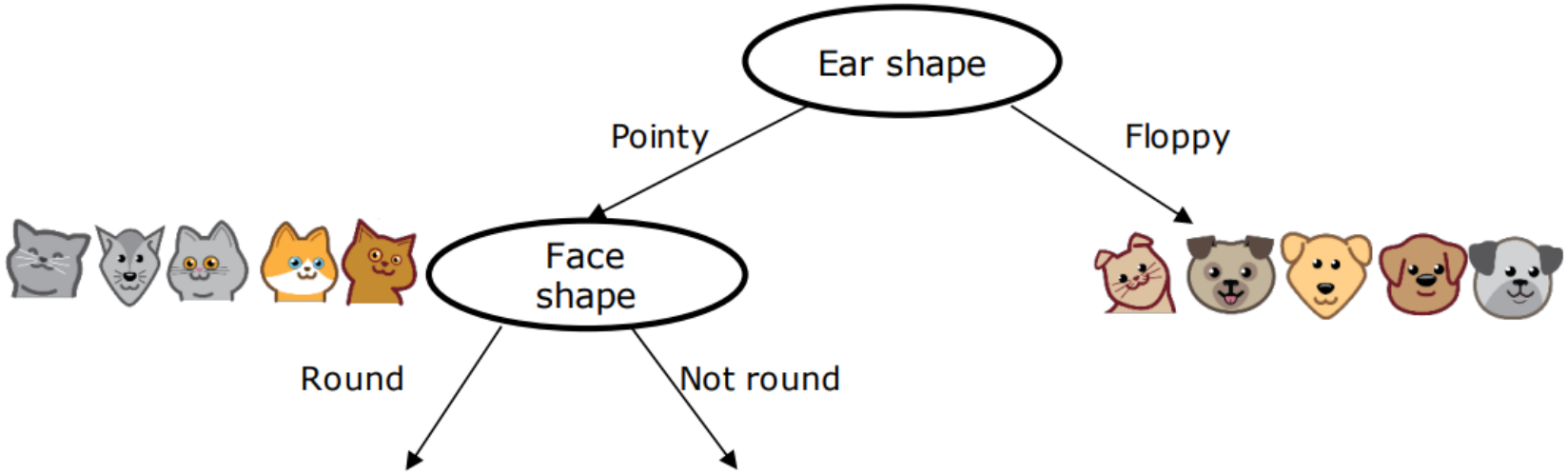
Learning Process

Learning Process

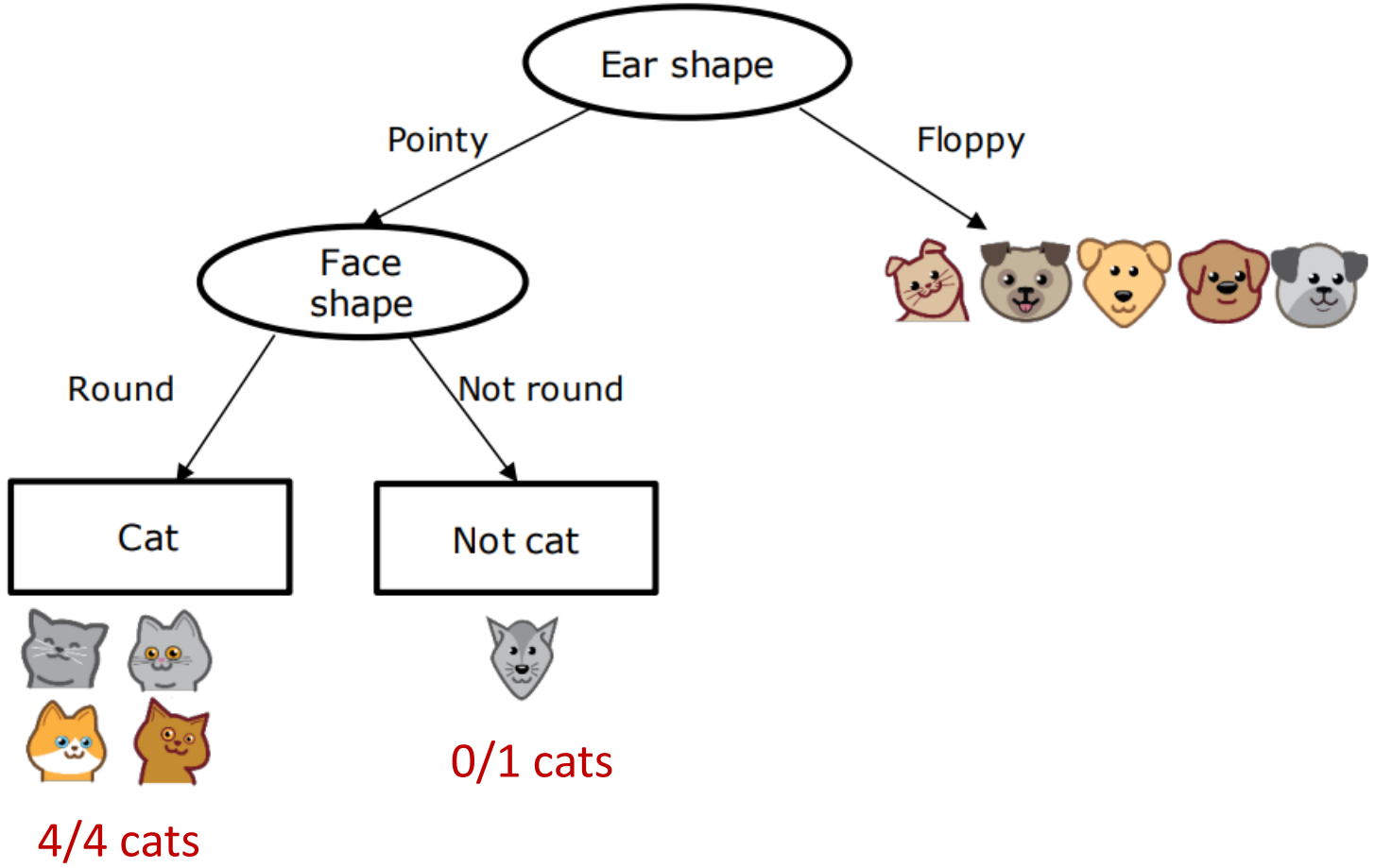
Learning Process



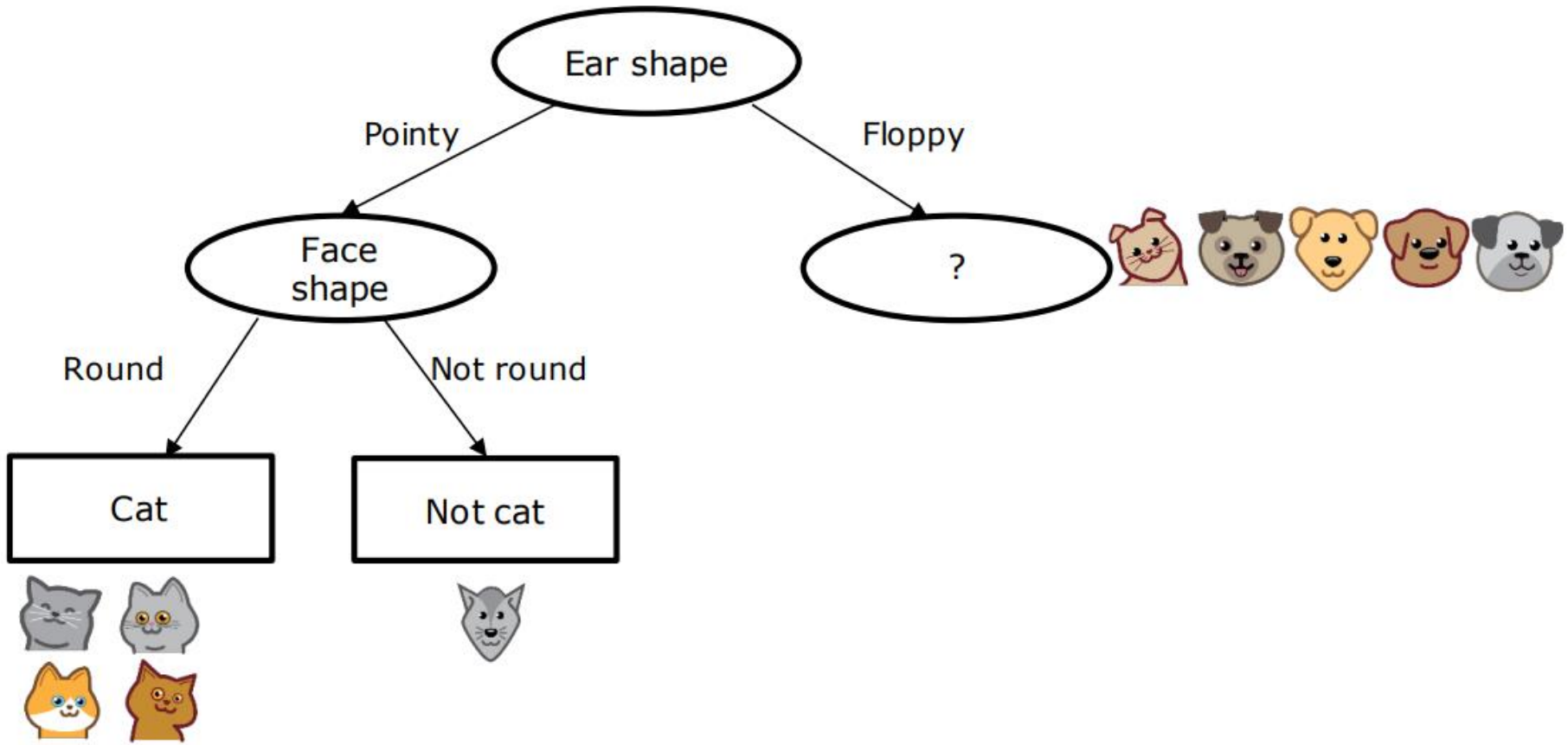
Learning Process



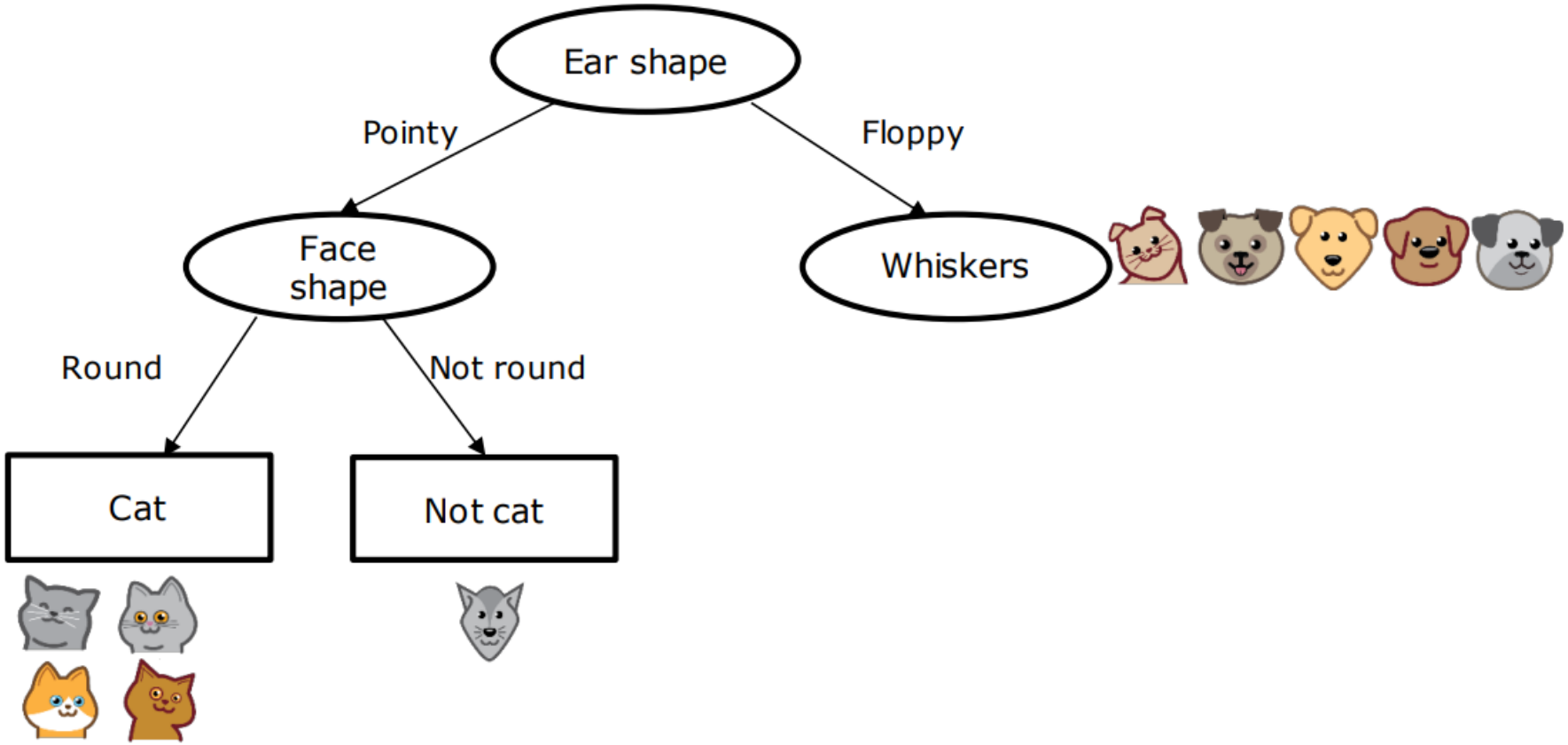
Learning Process



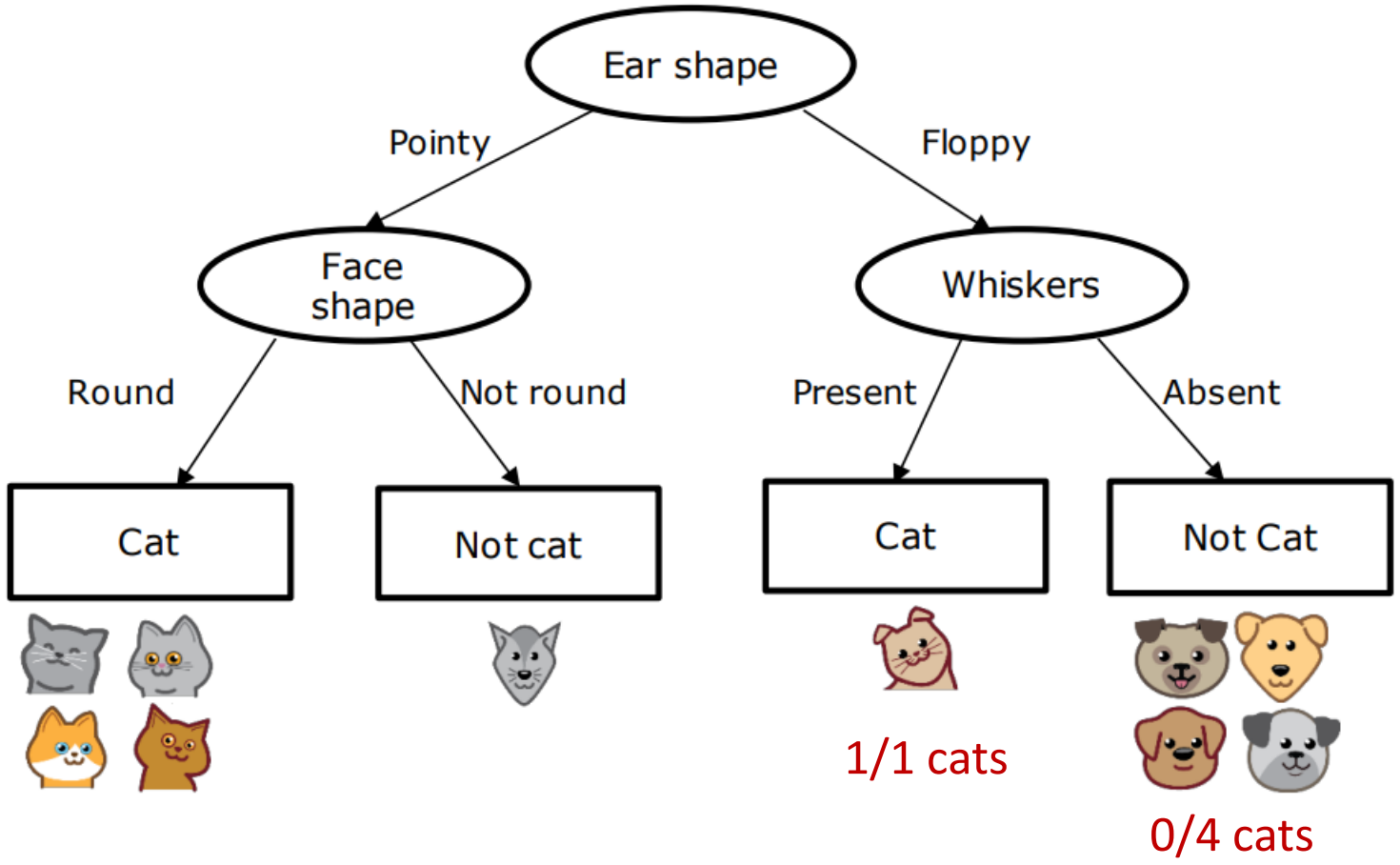
Learning Process



Learning Process



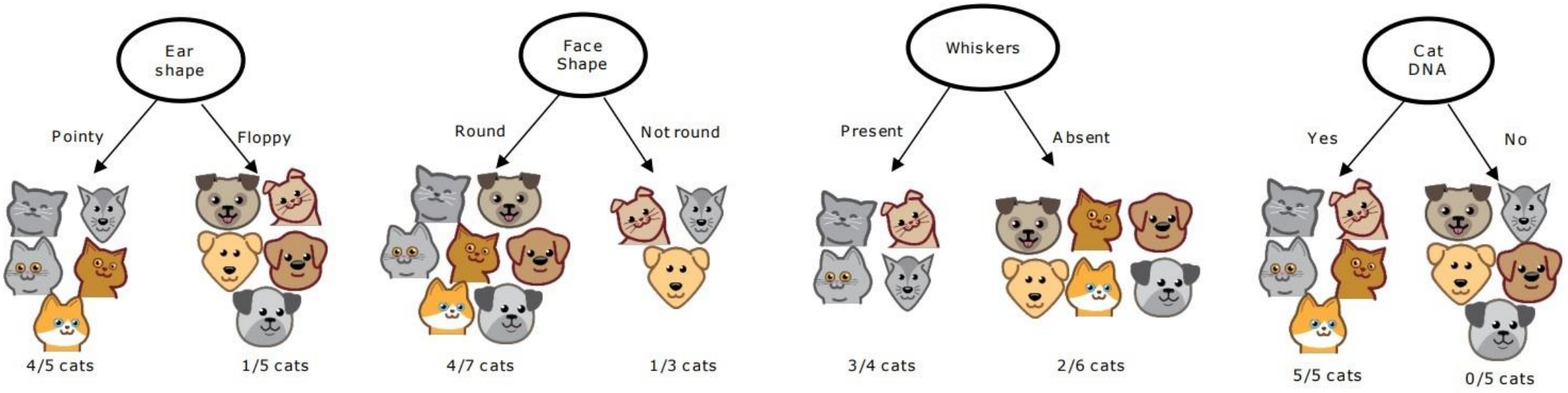
Learning Process



Learning Process

Decision 1: How to choose what feature to split on at each node?

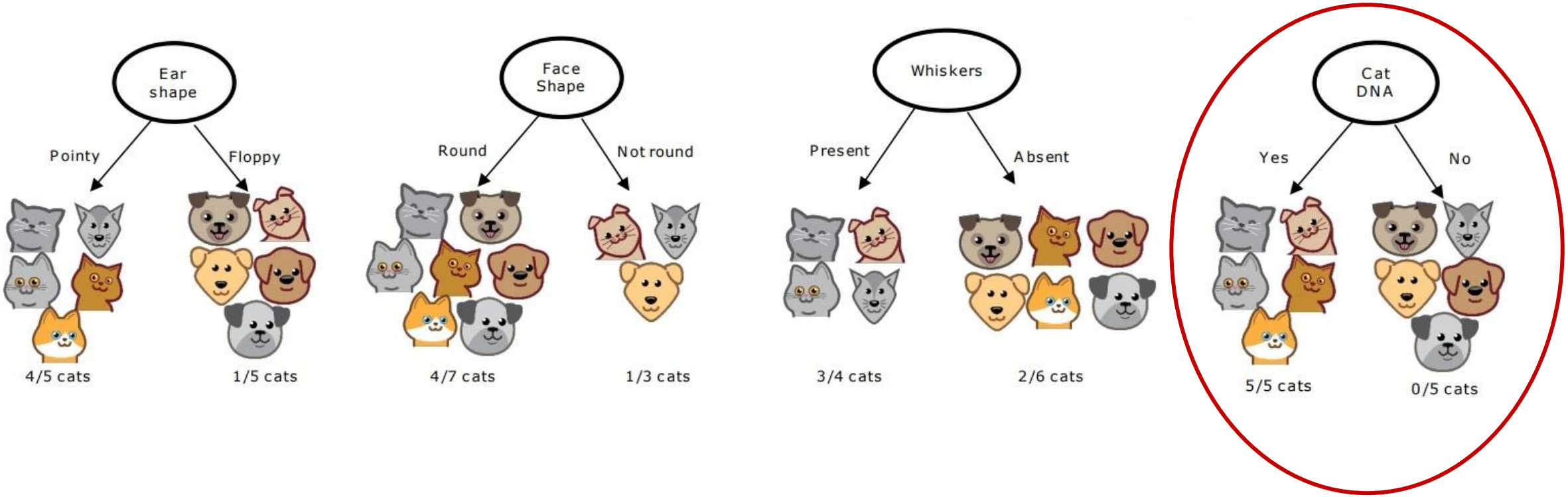
Maximize purity (or minimize impurity)



Learning Process

Decision 1: How to choose what feature to split on at each node?

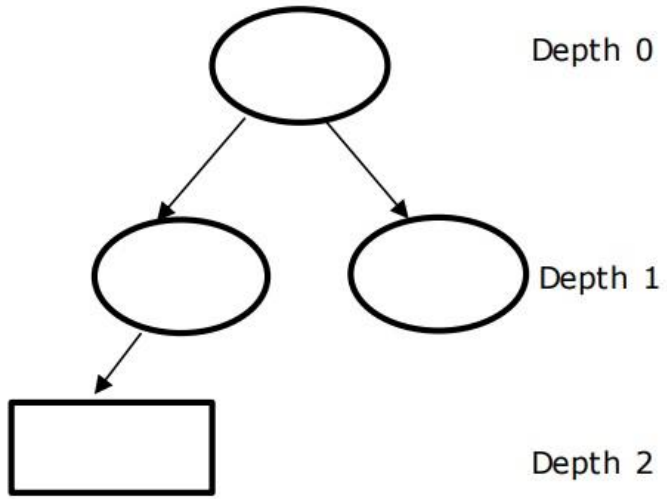
Maximize purity (or minimize impurity)



Learning Process

Decision 2: When do you stop splitting?

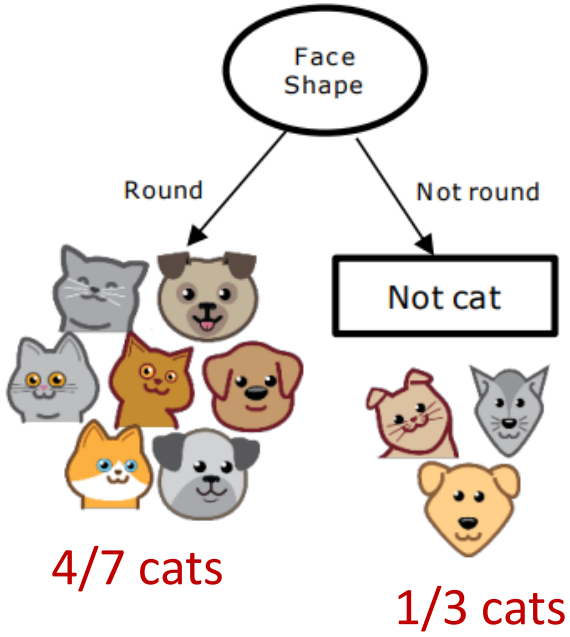
- When a node is 100% one class
- When splitting a node will result in the tree exceeding a maximum depth



Learning Process

Decision 2: When do you stop splitting?

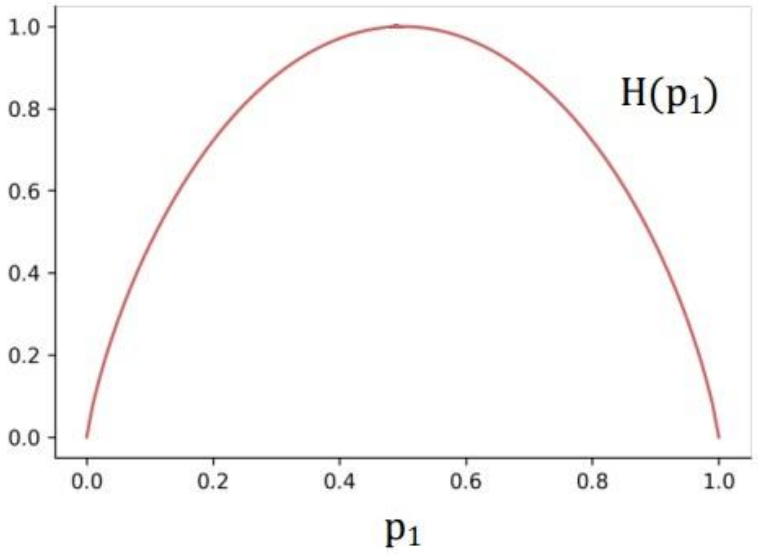
- When a node is 100% one class
- When splitting a node will result in the tree exceeding a maximum depth
- When improvements in purity score are below a threshold
- When number of examples in a node is below a threshold



Measuring Purity

Entropy as a measure of impurity

p_1 = fraction of examples that are cats



$p_1 = 0 \quad H(p_1) = 0$

$p_1 = 2/6 \quad H(p_1) = 0.92$

$p_1 = 3/6 \quad H(p_1) = 1$

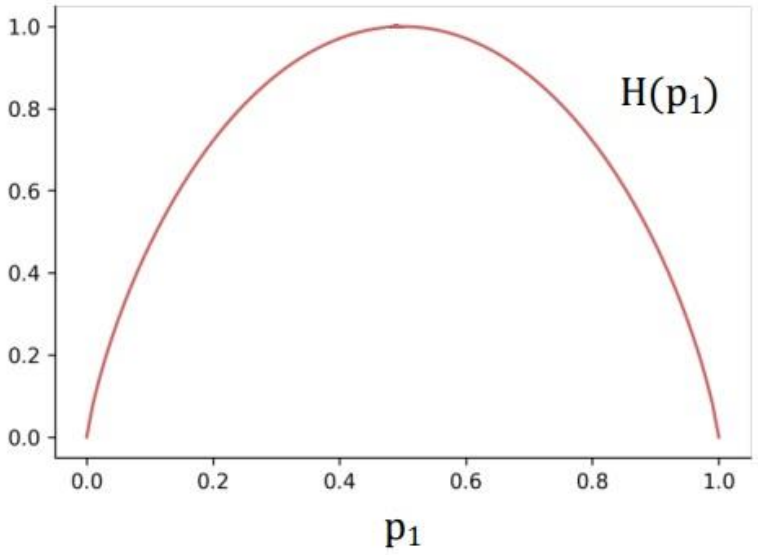
$p_1 = 5/6 \quad H(p_1) = 0.65$

$p_1 = 6/6 \quad H(p_1) = 0$

Measuring Purity

Entropy as a measure of impurity

p_1 = fraction of examples that are cats



$$p_0 = 1 - p_1$$

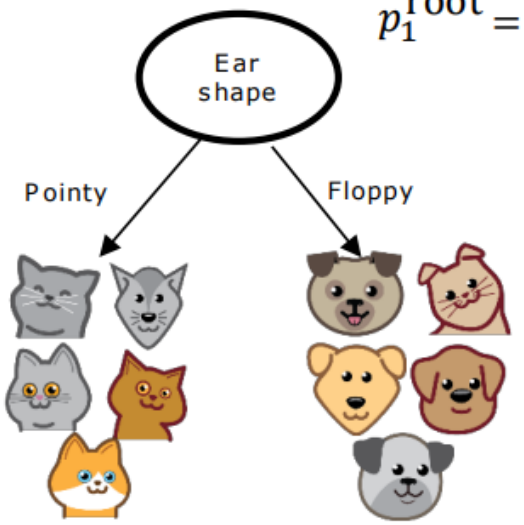
$$\mathrm{H}(p_1) = -p_1 log_2(p_1) - p_0 log_2(p_0)$$

$$= -p_1 log_2(p_1) - (1 - p_1) log_2(1 - p_1)$$

Note: $log_2(0) = 0$(-inf), $log_2(0.5) = -1$, $log_2(1) = 0$

Choosing a split => Information Gain

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



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

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

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

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

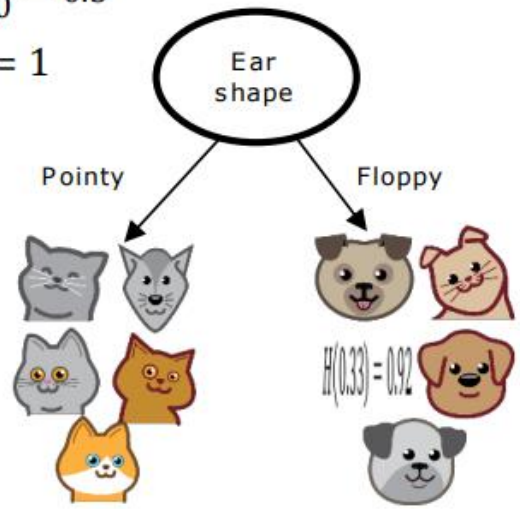
Information gain

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

Choosing a split => Information Gain

$p_1 = {}^{5}/_{10} = 0.5$

$H(0.5) = 1$



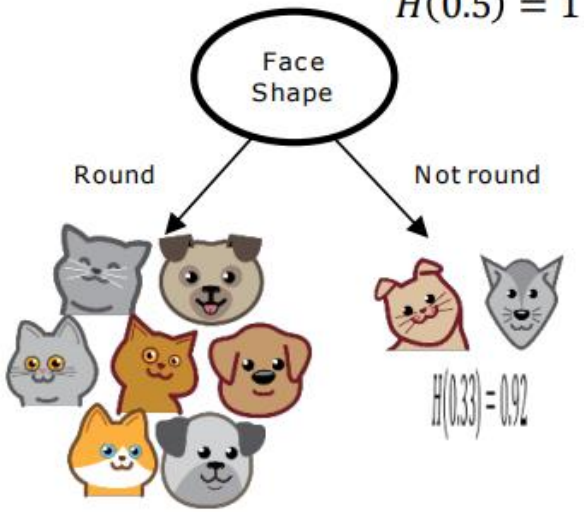
$p_1 = {}^{4}/_{5} = 0.8$ $p_1 = {}^{1}/_{5} = 0.2$

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

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

$= 0.28$

$H(0.5) = 1$



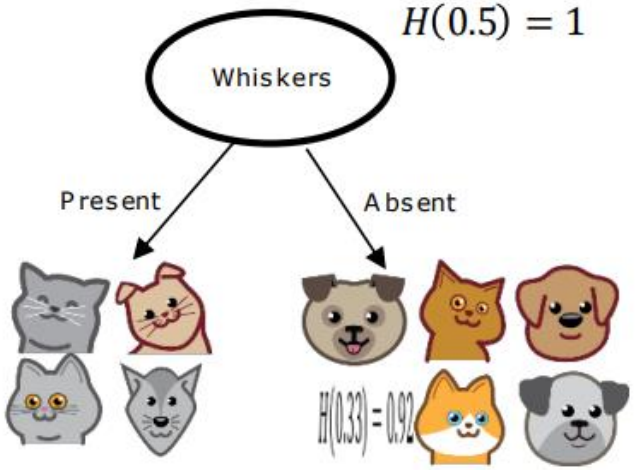
$p_1 = {}^{4}/_{7} = 0.57$ $p_1 = {}^{1}/_{3} = 0.33$

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

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

$= 0.03$

$H(0.5) = 1$



$p_1 = {}^{3}/_{4} = 0.75$ $p_1 = {}^{2}/_{6} = 0.33$

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

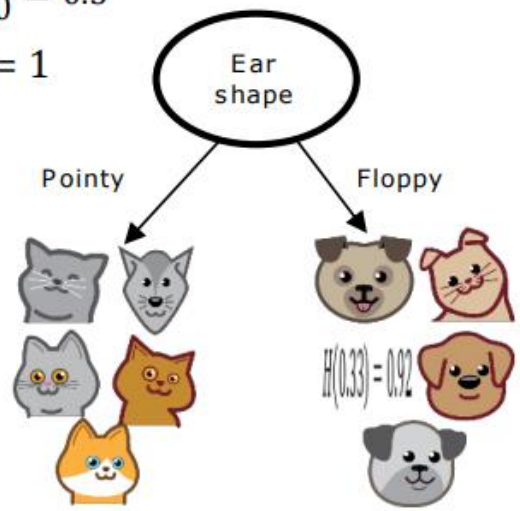
$$H(0.5) - \left(\frac{4}{10}H(0.75) + \frac{6}{10}H(0.33)\right)$$

$= 0.12$

Choosing a split => Information Gain

$p_1 = {}^{5}/_{10} = 0.5$

$H(0.5) = 1$



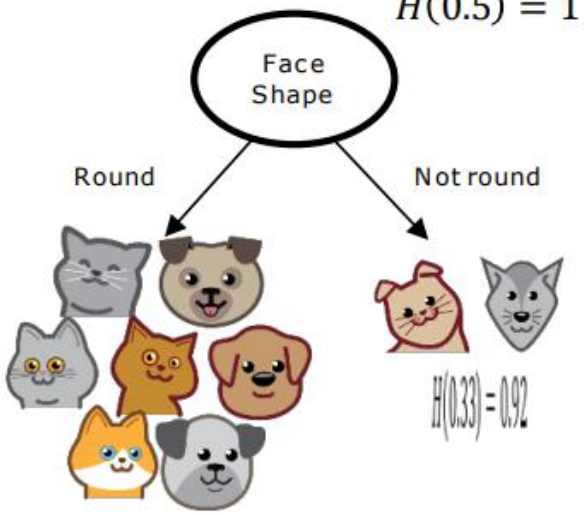
$p_1 = {}^{4}/_{5} = 0.8$ $p_1 = {}^{1}/_{5} = 0.2$

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

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

$= 0.28$

$H(0.5) = 1$



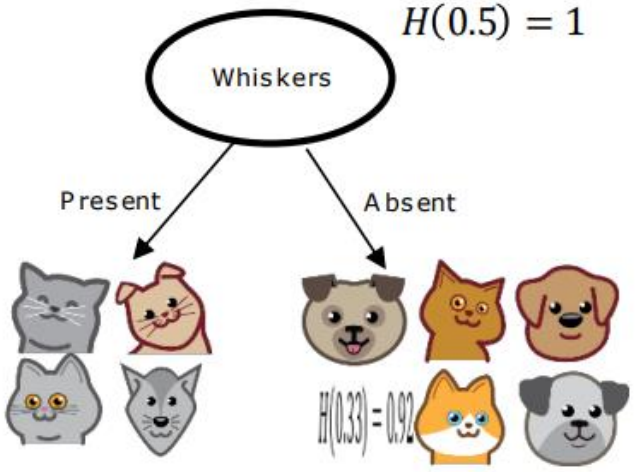
$p_1 = {}^{4}/_{7} = 0.57$ $p_1 = {}^{1}/_{3} = 0.33$

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

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

$= 0.03$

$H(0.5) = 1$



$p_1 = {}^{3}/_{4} = 0.75$ $p_1 = {}^{2}/_{6} = 0.33$

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

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

$= 0.12$

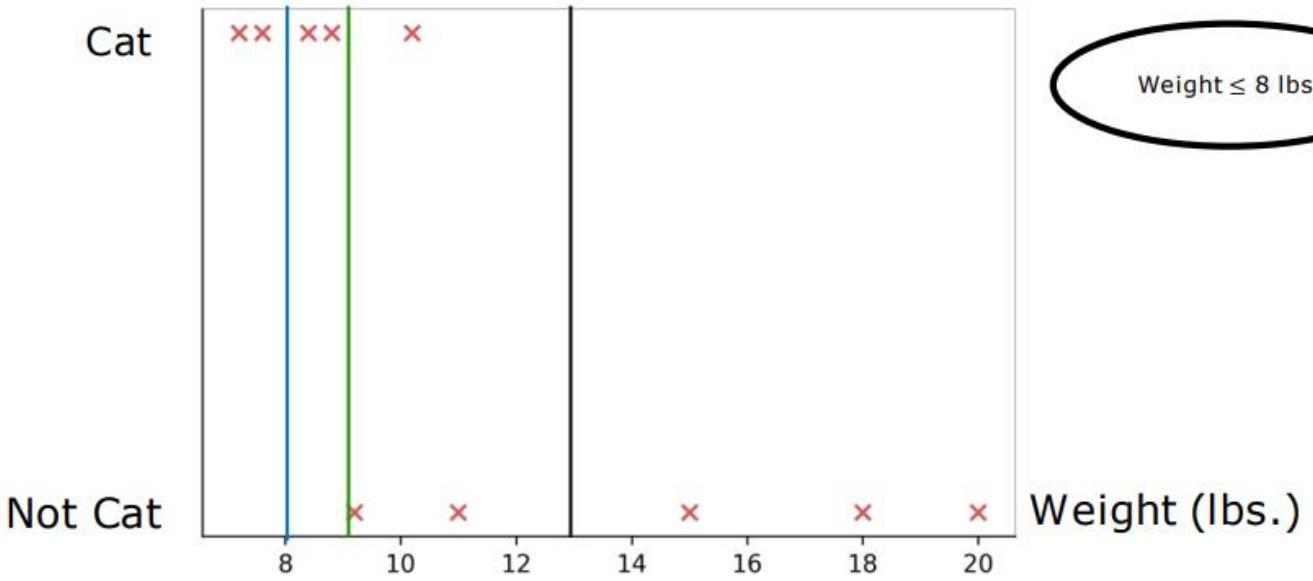
Putting it together

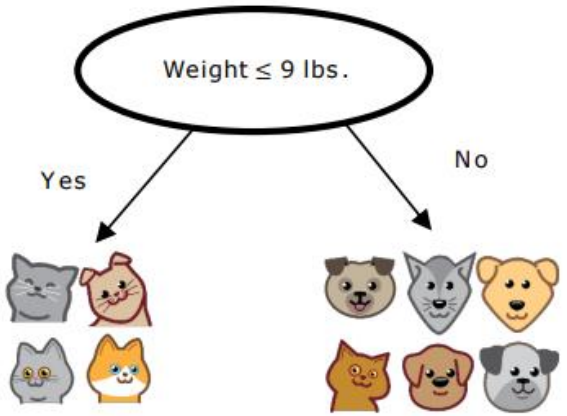
- Start with all examples at the root node
- Calculate information gain for all possible features, and pick the one with highest information gain
- Split dataset according to selected feature, and create left and right branches of the tree
- Keep repeating splitting process until stopping criteria is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

Continuous Valued Features

	Ear shape	Face shape	Whiskers	Weight (lbs.)	Cat
	Pointy	Round	Present	7.2	1
	Floppy	Not round	Present	8.8	1
	Floppy	Round	Absent	15	0
	Pointy	Not round	Present	9.2	0
	Pointy	Round	Present	8.4	1
	Pointy	Round	Absent	7.6	1
	Floppy	Not round	Absent	11	0
	Pointy	Round	Absent	10.2	1
	Floppy	Round	Absent	18	0
	Floppy	Round	Absent	20	0

Continuous Valued Features



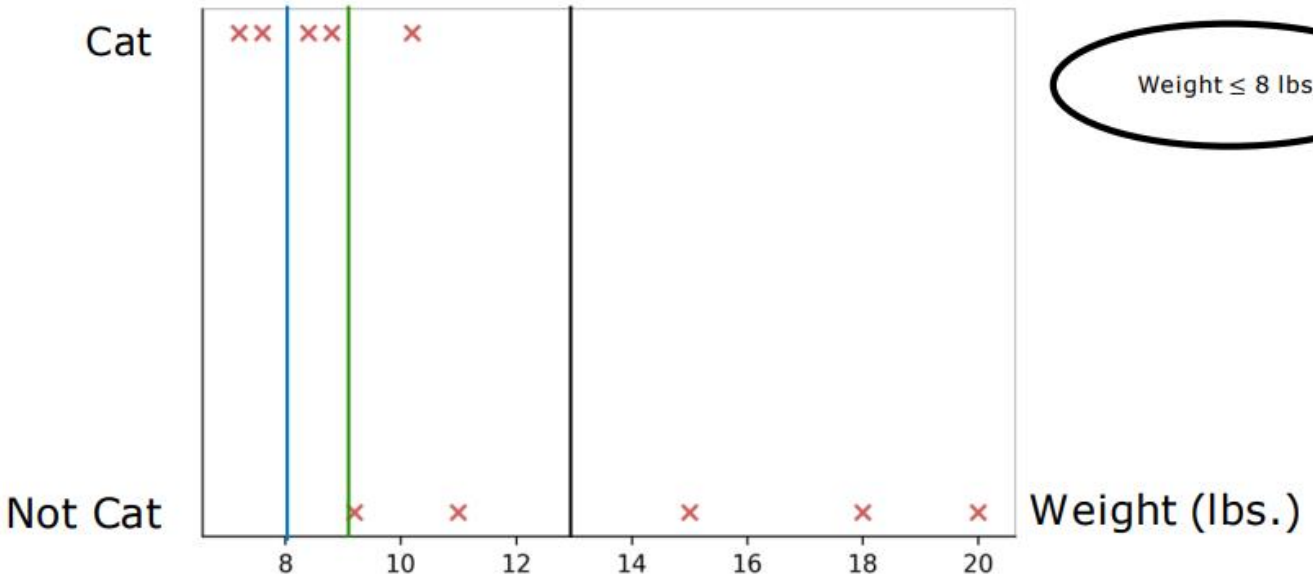


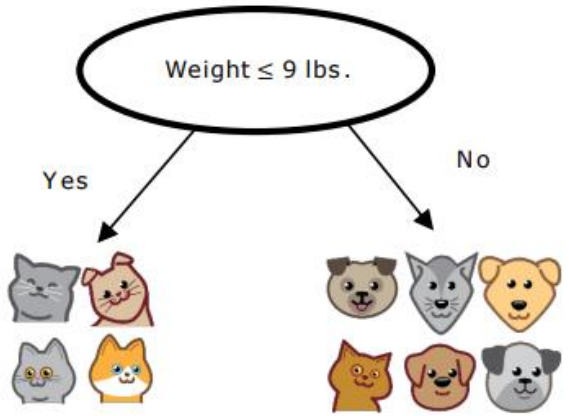
$$\mathrm{H}(0.5) - \left(\frac{2}{10}H\left(\frac{2}{2}\right) + \frac{8}{10}H\left(\frac{3}{8}\right)\right) = 0.24$$

$$\mathrm{H}(0.5) - \left(\frac{4}{10}H\left(\frac{4}{4}\right) + \frac{6}{10}H\left(\frac{1}{6}\right)\right) = 0.61$$

$$\mathrm{H}(0.5) - \left(\frac{7}{10}H\left(\frac{5}{7}\right) + \frac{3}{10}H\left(\frac{0}{3}\right)\right) = 0.40$$

Continuous Valued Features



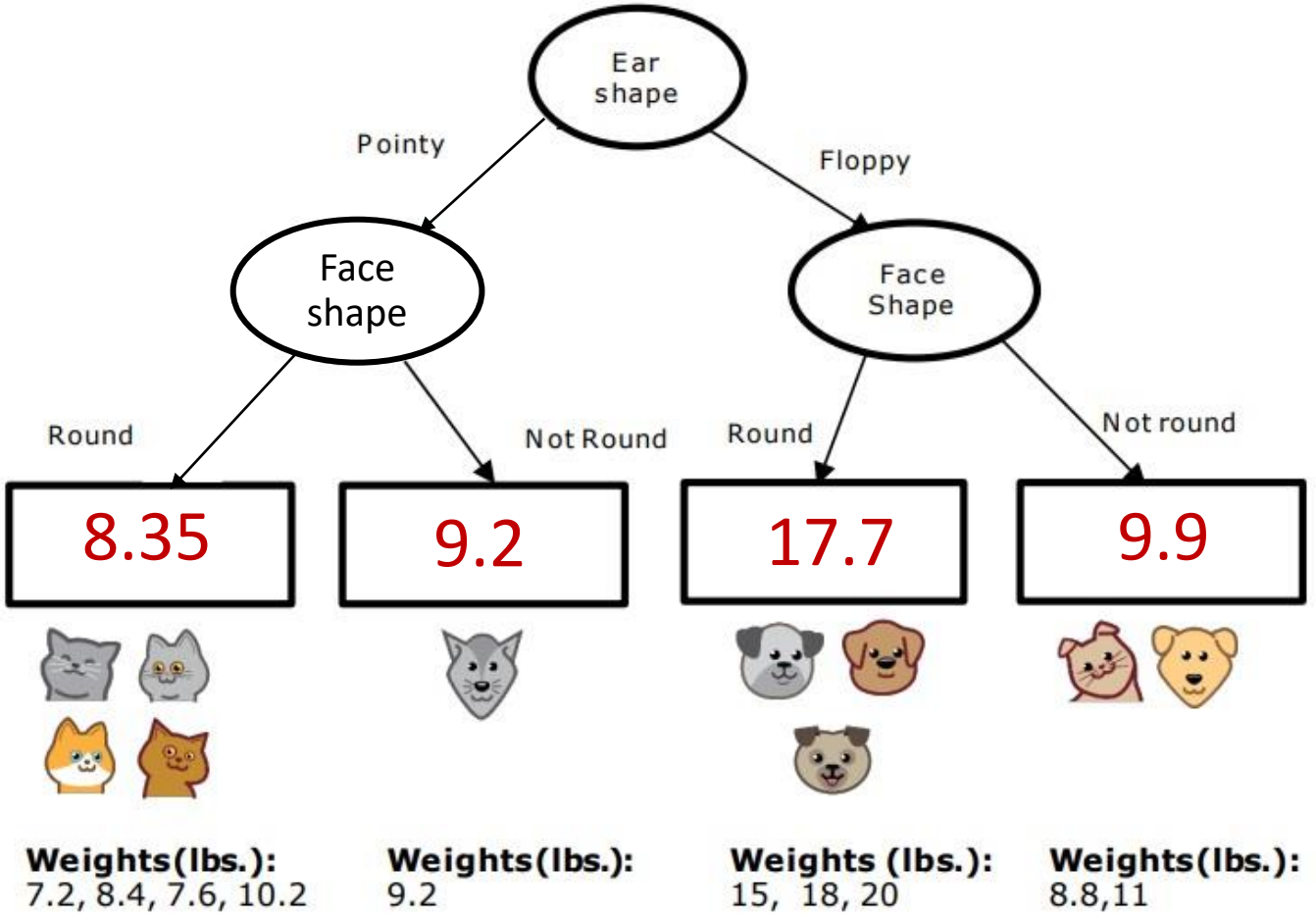


$$\mathrm{H}(0.5) - \left(\frac{2}{10}H\left(\frac{2}{2}\right) + \frac{8}{10}H\left(\frac{3}{8}\right)\right) = 0.24$$

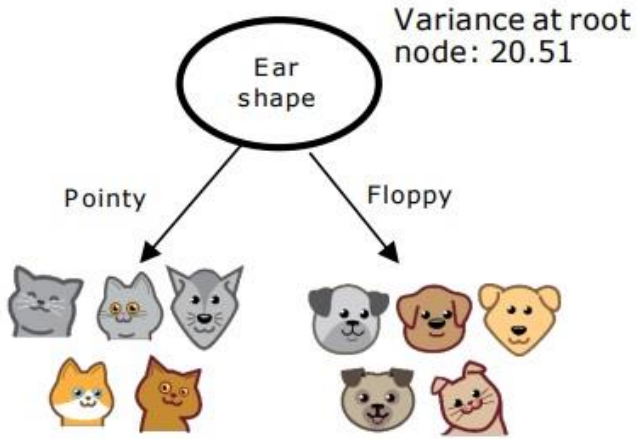
$$\mathrm{H}(0.5) - \left(\frac{4}{10}H\left(\frac{4}{4}\right) + \frac{6}{10}H\left(\frac{1}{6}\right)\right) = 0.61$$

$$\mathrm{H}(0.5) - \left(\frac{7}{10}H\left(\frac{5}{7}\right) + \frac{3}{10}H\left(\frac{0}{3}\right)\right) = 0.40$$

Regression Trees



Regression Trees



Variance at root node: 20.51

Weights: 7.2, 9.2, 8.4,7.6, 10.2

Variance: 1.47

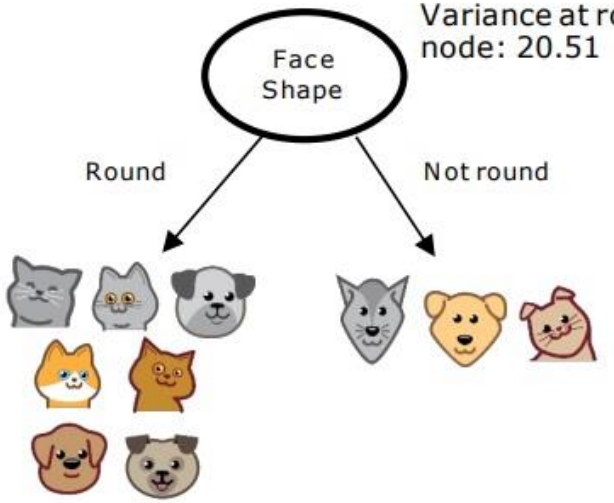
$w^{\text{left}} = 5/10$

Weights: 8.8, 15, 11, 18, 20

Variance: 21.87

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

$$20.51 - \left(\frac{5}{10} * 1.47 + \frac{5}{10} * 21.87\right) = 8.84$$



Variance at root node: 20.51

Weights: 7.2, 15, 8.4, 7.6,10.2, 18, 20

Variance: 27.80

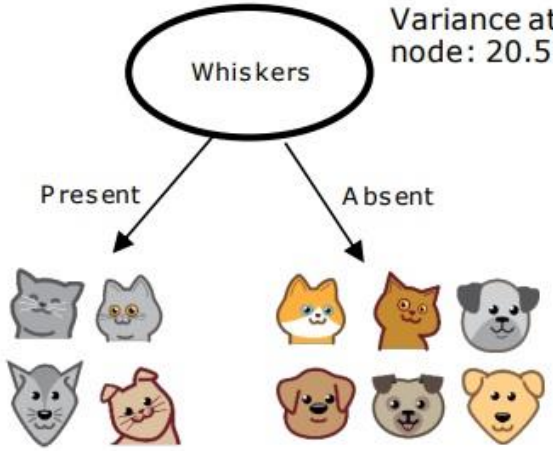
$w^{\text{left}} = 7/10$

Weights: 8.8,9.2,11

Variance: 1.37

$w^{\text{right}} = 3/10$

$$20.51 - \left(\frac{7}{10} * 27.80 + \frac{3}{10} * 1.37\right) = 0.64$$



Variance at root node: 20.51

Weights: 7.2, 8.8, 9.2, 8.4

Variance: 0.75

$w^{\text{left}} = 4/10$

Weights: 15, 7.6, 11, 10.2, 18, 20

Variance: 23.32

$w^{\text{right}} = 6/10$

$$20.51 - \left(\frac{4}{10} * 0.75 + \frac{6}{10} * 23.32\right) = 6.22$$

Lab #2

Thank you for your attention