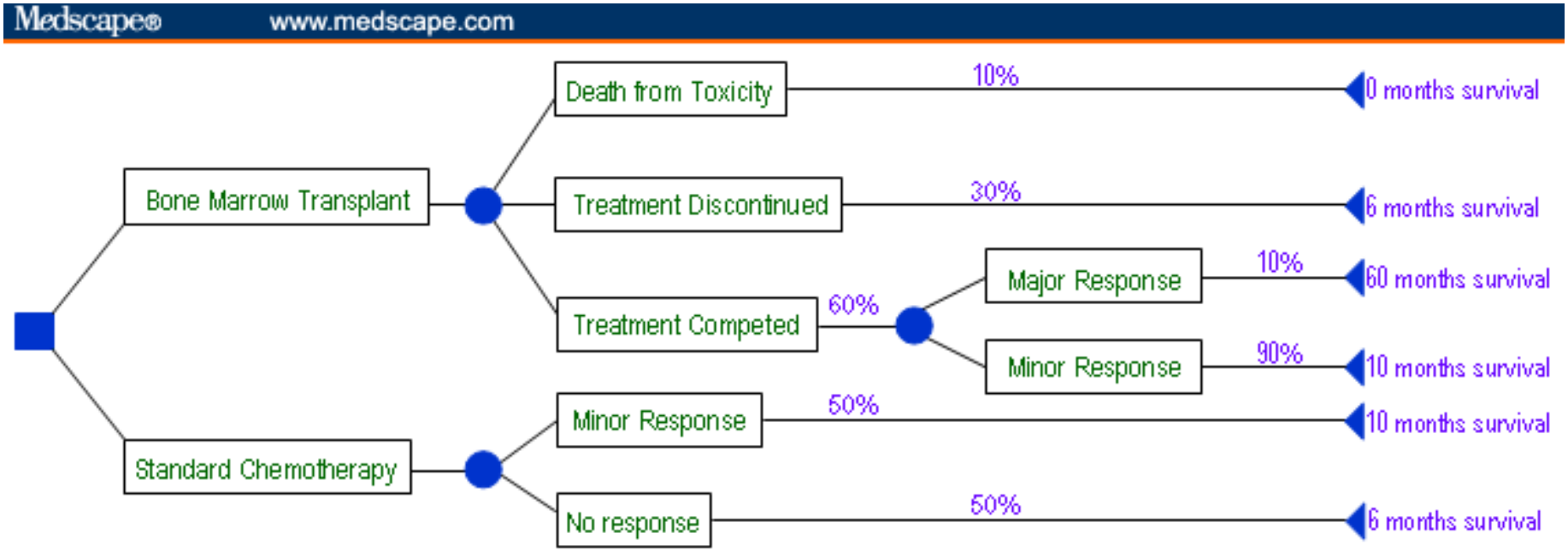
Decision Trees

Credits: Michael Crawford

Decision Trees



Classification by Trees

- Construct a Decision Tree
- ID3 Algorithm
- Shannon Entropy
- Gain of Info

How to Create a Decision Tree

- We first make a list of attributes that we can measure
  - These attributes (for now) must be discrete
- We then choose a *target attribute* that we want to predict
- Then create an *experience table* that lists what we have seen in the past

Sample Experience Table

Example	**Attributes**				**Target**
	Hour	Weather	Accident	Stall	Commute
D1	8 AM	Sunny	No	No	Long
D2	8 AM	Cloudy	No	Yes	Long
D3	10 AM	Sunny	No	No	Short
D4	9 AM	Rainy	Yes	No	Long
D5	9 AM	Sunny	Yes	Yes	Long
D6	10 AM	Sunny	No	No	Short
D7	10 AM	Cloudy	No	No	Short
D8	9 AM	Rainy	No	No	Medium
D9	9 AM	Sunny	Yes	No	Long
D10	10 AM	Cloudy	Yes	Yes	Long
D11	10 AM	Rainy	No	No	Short
D12	8 AM	Cloudy	Yes	No	Long
D13	9 AM	Sunny	No	No	Medium

Decision Tree Algorithms

- The basic idea behind any decision tree algorithm is as follows:
  - Choose the *best* attribute(s) to split the remaining instances and make that attribute a decision node
  - Repeat this process for recursively for each child
  - Stop when:
    - All the instances have the same target attribute value
    - There are no more attributes
    - There are no more instances

ID3 Background

- “Iterative Dichotomizer 3”.
- Invented by Ross Quinlan in 1979.
- Generates Decision Trees using Shannon Entropy

Entropy

- In thermodynamics, entropy is a measure of how ordered or disordered a system is.
- In information theory, entropy is a measure of how certain or uncertain the value of a random variable is (or will be).
- Varying degrees of randomness, depending on the number of possible values and the total size of the set.

Shannon Entropy

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

- Introduced by Claude Shannon in 1948
- Quantifies “randomness”
- **Lower value** implies **less** uncertainty
- **Higher value** implies **more** uncertainty

Information Gain

- Uses Shannon Entropy
- IG calculates effective change in entropy after making a decision based on the value of an attribute.
- For decision trees, it's ideal to base decisions on the attribute that provides the largest change in entropy, the attribute with the highest gain.

Information Gain

$$\textit{For Set } S \textit{, Attribute } A$$
$$\textit{Where } S \textit{ is split into subsets based on values of } A$$
$$\subset_S^A = \text{Subset A of S}$$
$$I_E = \textit{Entropy}, \text{p}(\subset_S^A) = \frac{\text{size}(\subset_S^A)}{\text{size(S)}}$$

$$I_G(S, A) = I_E(S) - \sum^{n} (\text{p}(\subset_S^{A_n}) * I_E(\subset_S^{A_n}))$$

Information Gain

$$I_G(S,A) = I_E(S) - \sum^{n} (\mathrm{p}(\subset_S^{A_n}) * I_E(\subset_S^{A_n}))$$

- Information Gain for attribute A on set S is defined by taking the entropy of S and subtracting from it the summation of the entropy of each subset of S, determined by values of A, multiplied by each subset's proportion of S.

ID3 Algorithm

1. Establish Classification Attribute (in Table R)
2. Compute Classification Entropy.
3. For each attribute in R, calculate Information Gain using classification attribute.
4. Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).
5. Remove Node Attribute, creating reduced table R_S.
6. Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows in the reduced table (e.g., no further division/classification of the data with respect to the Target Attribute is possible).

Example

Model	Engine	SC/Turbo	Weight	Fuel Eco	Fast
Prius	small	no	average	good	no
Civic	small	no	light	average	no
WRX STI	small	yes	average	bad	yes
M3	medium	no	heavy	bad	yes
RS4	large	no	average	bad	yes
GTI	medium	no	light	bad	no
XJR	large	yes	heavy	bad	no
S500	large	no	heavy	bad	no
911	medium	yes	light	bad	yes
Corvette	large	no	average	bad	yes
Insight	small	no	light	good	no
RSX	small	no	average	average	no
IS350	medium	no	heavy	bad	no
MR2	small	yes	average	average	no
E320	medium	no	heavy	bad	no

Example

Model	Engine	SC/Turbo	Weight	Fuel Eco	Fast
Prius	small	no	average	good	no
Civic	small	no	light	average	no
WRX STI	small	yes	average	bad	yes
M3	medium	no	heavy	bad	yes
RS4	large	no	average	bad	yes
GTI	medium	no	light	bad	no
XJR	large	yes	heavy	bad	no
S500	large	no	heavy	bad	no
911	medium	yes	light	bad	yes
Corvette	large	no	average	bad	yes
Insight	small	no	light	good	no
RSX	small	no	average	average	no
IS350	medium	no	heavy	bad	no
MR2	small	yes	average	average	no
E320	medium	no	heavy	bad	no

- Model Attribute can be tossed out, since its always unique, and it doesn't help our result.

Example

Engine	SC/Turbo	Weight	Fuel Eco	Fast
small	no	average	good	no
small	no	light	average	no
small	yes	average	bad	yes
medium	no	heavy	bad	yes
large	no	average	bad	yes
medium	no	light	bad	no
large	yes	heavy	bad	no
large	no	heavy	bad	no
medium	yes	light	bad	yes
large	no	average	bad	yes
small	no	light	good	no
small	no	average	average	no
medium	no	heavy	bad	no
small	yes	average	average	no
medium	no	heavy	bad	no

- Establish a target classification
- Is the car fast?
- 5/15 yes, 10/15 no

Example – Classification Entropy

- Calculating for the Classification Entropy

E(S) is calculated based on the attribute of interest (is the car fast?)

$$E(S) = -(5/15)\log_2(5/15)-(10/15)\log_2(10/15) = \sim 0.918$$

- Must calculate Information Gain of remaining attributes to determine the root node.

Note: someone asked in class whether the log is always to the base 2. It is for the ID3 algorithm, as a way to measure information in bits (i.e., in 0's and 1's; how many of these pairs can represent the data). $\text{Log}_2 1=0$ ($2^0 = 1$); $\text{Log}_2 8 = 3$ ($2^3=8$)

Example – Information Gain

- Engine: 6 small, 5 medium, 4 large
- 3 values for attribute engine, so we need 3 entropy calculations

small: 5 no, 1 yes	$I_{small} = -(5/6)\log_2(5/6)-(1/6)\log_2(1/6) = \sim 0.65$
medium: 3 no, 2 yes	$I_{medium} = -(3/5)\log_2(3/5)-(2/5)\log_2(2/5) = \sim 0.97$
large: 2 no, 2 yes	$I_{large} = 1$ (evenly distributed subset)

$I_{Gain}(S, A) = I_{Gain}(fast, engine) = E(S) – [(6/15)*I_{small} + (5/15)*I_{medium} + (4/15)*I_{large}]$

Information gain for engine:

I_{Gain} (fast, engine) = **0.918 – 0.85 = 0.068**

Example – Information Gain

- SC/Turbo: 4 yes, 11 no
- 2 values for attribute SC/Turbo, so we need 2 entropy calculations

yes: 2 yes, 2 no	I_{turbo} = 1 (evenly distributed subset)
no: 3 yes, 8 no	$I_{noturbo} = -(3/11)\log_2(3/11)-(8/11)\log_2(8/11) = \sim 0.84$

$$\mathbf{I_{Gain}(S, A) = I_{Gain}(fast, turbo) = E(S) - [(4/15)*I_{turbo} + (11/15)*I_{noturbo}]}$$

$$\mathbf{I_{Gain}(fast, turbo) = 0.918 - 0.886 = 0.032}$$

Example – Information Gain

- Weight: 6 Average, 4 Light, 5 Heavy
- 3 values for attribute weight, so we need 3 entropy calculations

average: 3 no, 3 yes	$I_{average} = 1$ (evenly distributed subset)
light: 3 no, 1 yes	$I_{light} = -(3/4)\log_2(3/4)-(1/4)\log_2(1/4) = \sim 0.81$
heavy: 4 no, 1 yes	$I_{heavy} = -(4/5)\log_2(4/5)-(1/5)\log_2(1/5) = \sim 0.72$

$$\mathbf{I_{Gain}(fast, weight) = E(S) - [(6/15)*I_{average} + (4/15)*I_{light} + (5/15)*I_{heavy}]}$$

$$\mathbf{I_{Gain}(fast, weight) = 0.918 - 0.856 = 0.062}$$

Example – Information Gain

- Fuel Economy: 2 good, 3 average, 10 bad
- 3 values for attribute Fuel Eco, so we need 3 entropy calculations

good: 0 yes, 2 no	$I_{good} = 0$ (no variability)
average: 0 yes, 3 no	$I_{average} = 0$ (no variability)
bad: 5 yes, 5 no	$I_{bad} = 1$ (evenly distributed subset)

We can omit calculations for good and average since they always end up not fast.

I_{Gain}(fast, Fuel Eco) = E(S) – [(10/15)*I_{bad}]
I_{Gain}(fast, Fuel Eco) = 0.918 – 0.667 = 0.251

Example – Choosing the Root Node

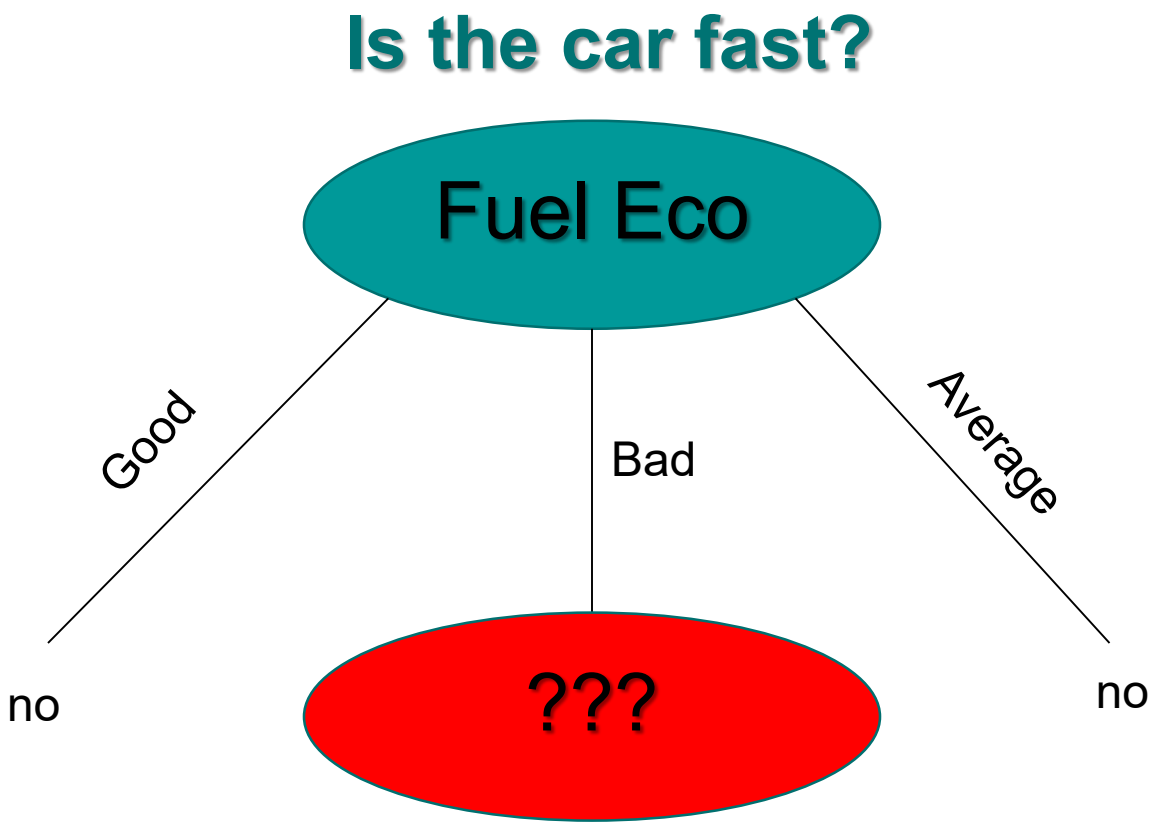
- Recap:

I_{gain} – Engine	**0.068**
I_{gain} – Turbo	**0.032**
I_{gain} – Weight	**0.062**
I_{gain} – Fuel Eco	**0.251**

Our best pick is Fuel Eco – it has the highest information gain

We can immediately predict the car is not fast when fuel economy is good or average.

Example – Root of Decision Tree



Example – After Root Node Creation

- Since we selected the Fuel Eco attribute for our Root Node, it is removed from the table for future calculations.

Engine	SC/Turbo	Weight	Fast
small	yes	average	yes
medium	no	heavy	yes
large	no	average	yes
medium	no	light	no
large	yes	heavy	no
large	no	heavy	no
medium	yes	light	yes
large	no	average	yes
medium	no	heavy	no
medium	no	heavy	no

Calculating Entropy E relative to our initial question of whether the car is fast, (given all the remaining values have Fuel Economy being bad) we get 1, since we have 5 yes and 5 no.

$$\mathbf{E(S) = -(5/10)\log_2(5/10)-(5/10)\log_2(5/10) = 1}$$

Example – Information Gain

- Engine: 1 small, 5 medium, 4 large
- 3 values for attribute engine, so we need 3 entropy calculations

small: 1 yes, 0 no	$I_{small} = 0$ (no variability)
medium: 2 yes, 3 no	$I_{medium} = -(2/5)\log_2(2/5)-(3/5)\log_2(3/5) = \sim 0.97$
large: 2 no, 2 yes	$I_{large} = 1$ (evenly distributed subset)

I_{Gain} (fast, engine) **$= E - (5/10)*I_{medium} + (4/10)*I_{large}]$**

I_{Gain} (fast, engine) = **$1 - 0.885 = 0.115$**

Example – Information Gain

- SC/Turbo: 3 yes, 7 no
- 2 values for attribute SC/Turbo, so we need 2 entropy calculations

yes: 2 yes, 1 no	$I_{turbo} = -(2/3)\log_2(2/3)-(1/3)\log_2(1/3) = \sim 0.84$
no: 3 yes, 4 no	$I_{noturbo} = -(3/7)\log_2(3/7)-(4/7)\log_2(4/7) = \sim 0.84$

I_{Gain} (fast, turbo) **$= E - [(3/10)*I_{turbo} + (7/10)*I_{noturbo}]$**

I_{Gain} (fast, turbo) = **$1 - 0.965 = 0.035$**

Example – Information Gain

- Weight: 3 average, 5 heavy, 2 light
- 3 values for attribute weight, so we need 3 entropy calculations

average: 3 yes, 0 no	$I_{average} = 0$ (no variability)
heavy: 1 yes, 4 no	$I_{heavy} = -(1/5)\log_2(1/5)-(4/5)\log_2(4/5) = \sim 0.72$
light: 1 yes, 1 no	$I_{light} = 1$ (evenly distributed subset)

I_{Gain} (fast, weight) **$= E - [(5/10)*I_{heavy}+(2/10)*I_{light}]$**

I_{Gain} (fast, weigt) = **$1 - 0.561 = 0.439$**

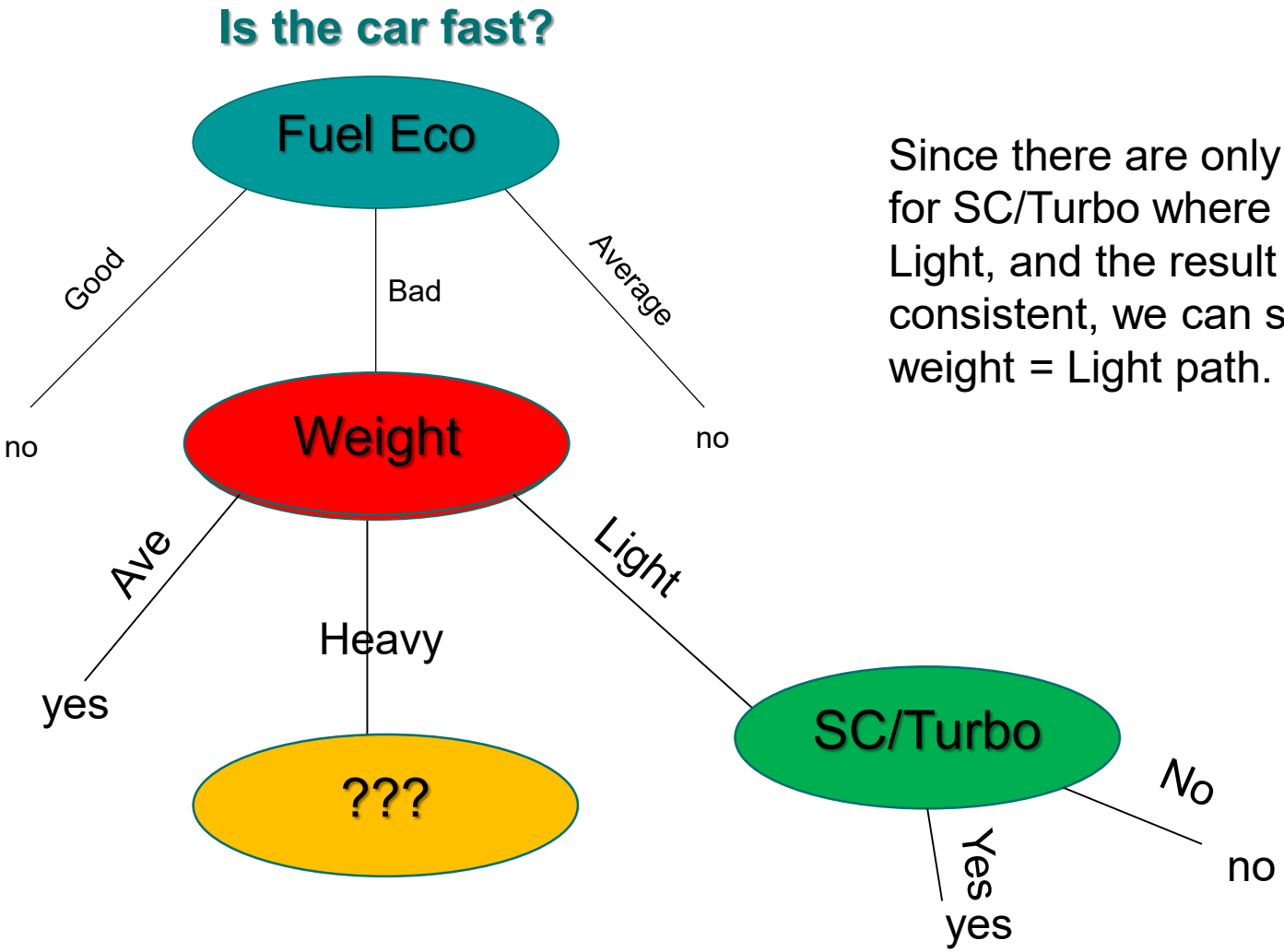
Example – Choosing the Level 2 Node

- Recap:

I_{gain} – Engine	**0.115**
I_{gain} – Turbo	**0.035**
I_{gain} – Weight	**0.439**

Weight has the highest gain, and is thus the best choice.

Example – Decision Tree



Since there are only two items for SC/Turbo where Weight = Light, and the result is consistent, we can simplify the weight = Light path.

Example – Updated Table All Heavy Weight

Engine	SC/Turbo	Fast
medium	no	yes
large	yes	no
large	no	no
medium	no	no
medium	no	no

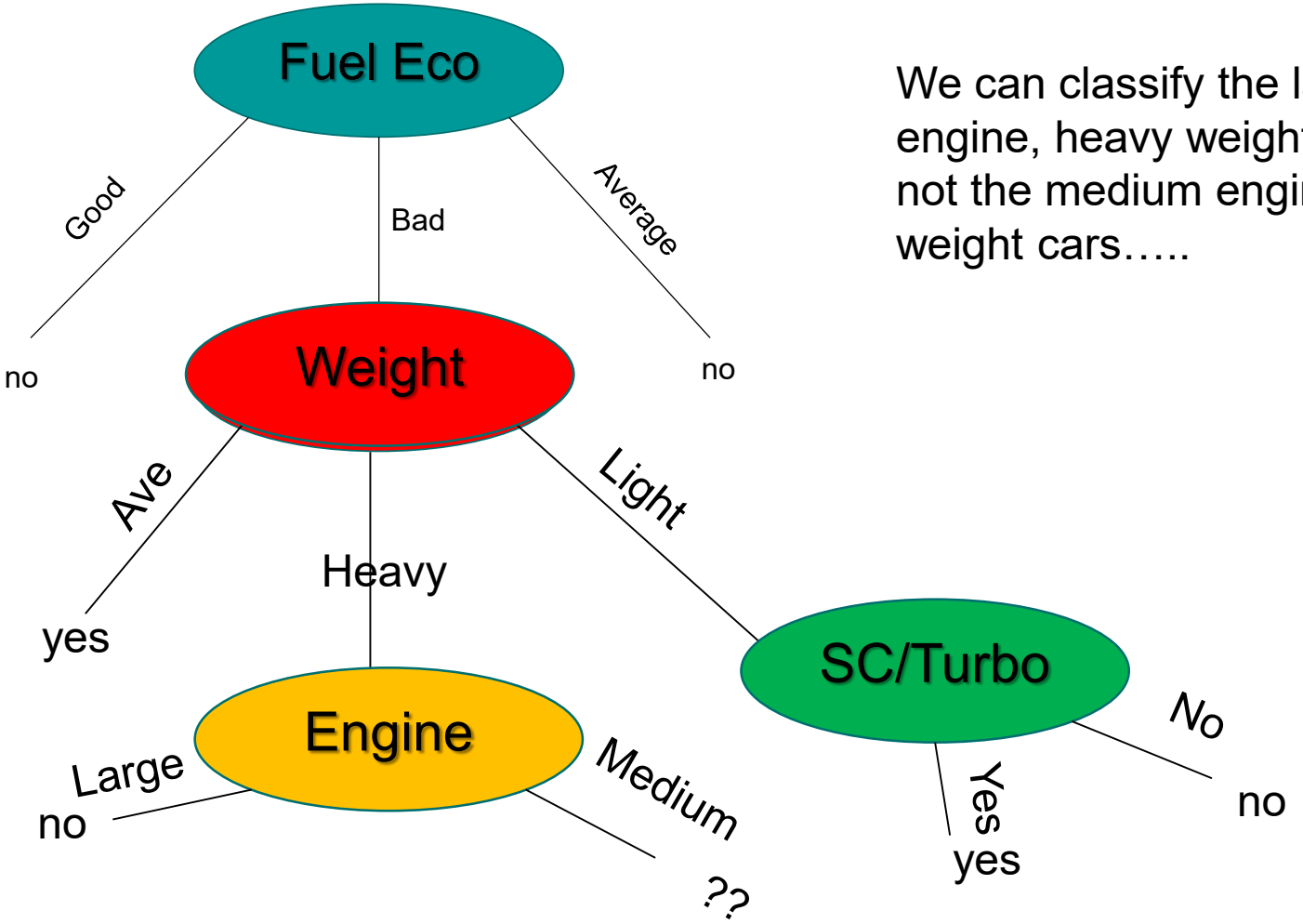
- All cars with a medium engine are not fast.
- All medium engine cars with no turbo are not fast.

Due to inconsistent patterns in the data (no pattern to fast or slow cars anymore), that is the end of useful classification.

This is a limit of the ID3 algorithm & discrete classification decisions in general.

Example – Decision Tree

Is the car fast?



We can classify the large engine, heavy weight cars, but not the medium engine heavy weight cars…..

Notes on ID3 Algorithm for Decision Trees

- ID3 attempts to make the shortest decision tree out of a set of learning data, shortest is not always the best classification.

- Requires learning data to have completely consistent patterns with no uncertainty.