NNSE 784
Advanced Analytics Methods

Instructor: F Doyle (CESTM L210)

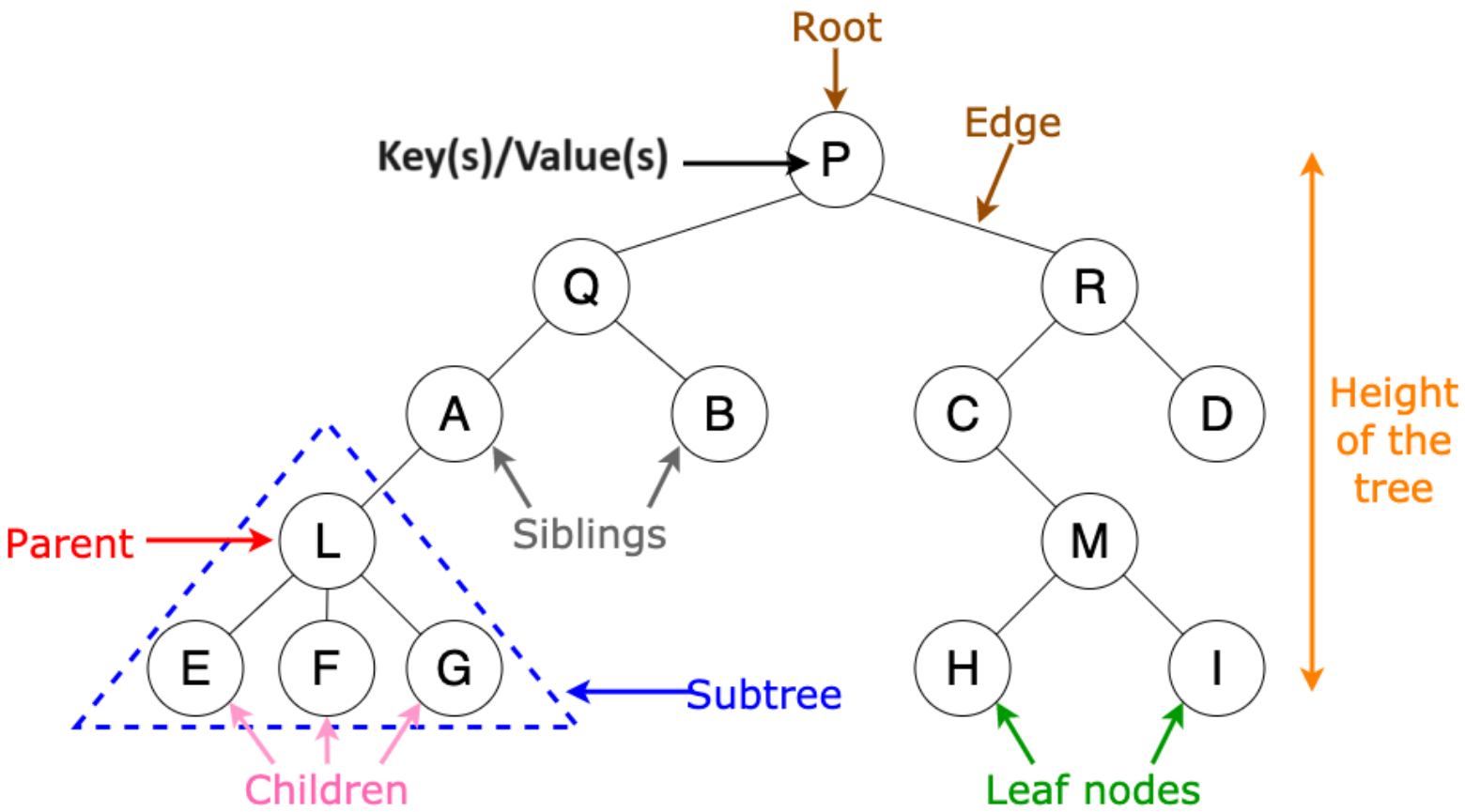
MW 4:30 – 5:50, NFN 203

Slide Set #17
Decision Trees

Lecture Outline

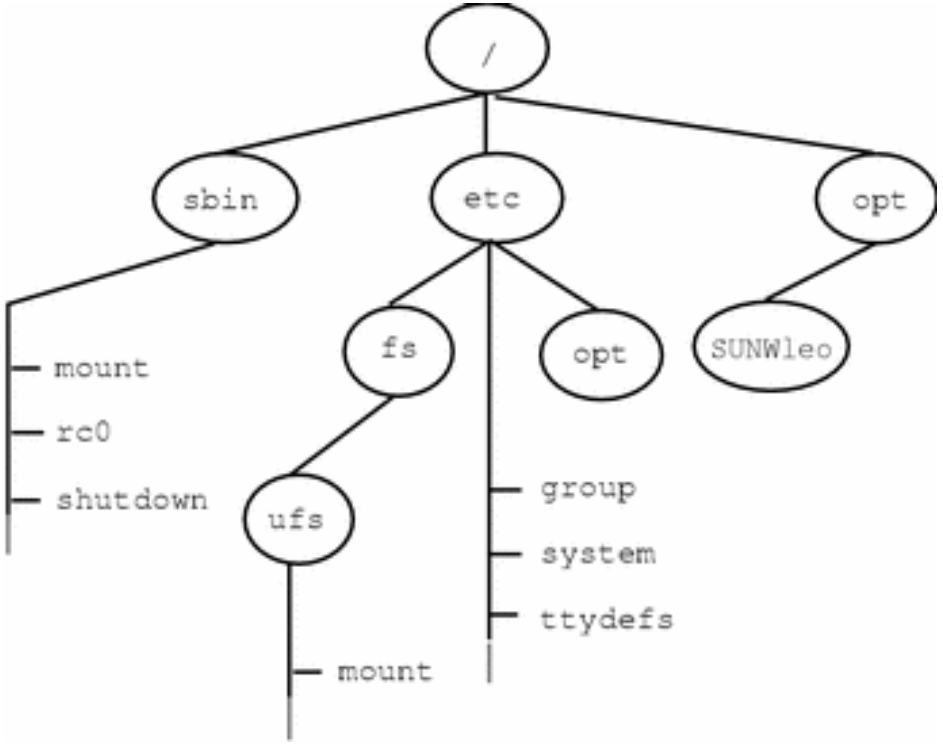
- The “tree” data structure
- Decision Tree concept
- Building a decision tree
  - Information Entropy
  - Attribute Selection
  - Information Gain
- Jupyter Notebook exercises

Tree Structures



Examples of Tree Data Structures

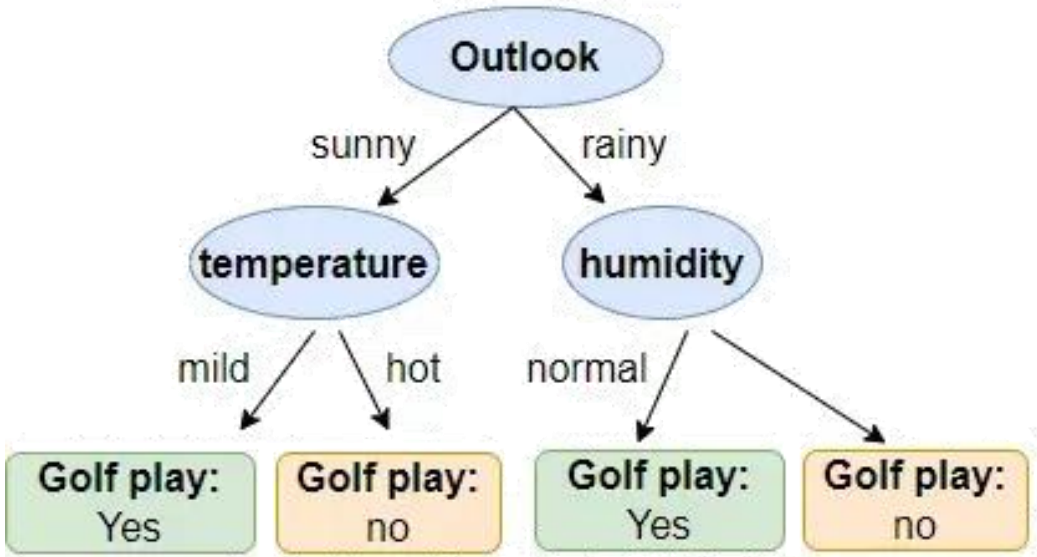
File System

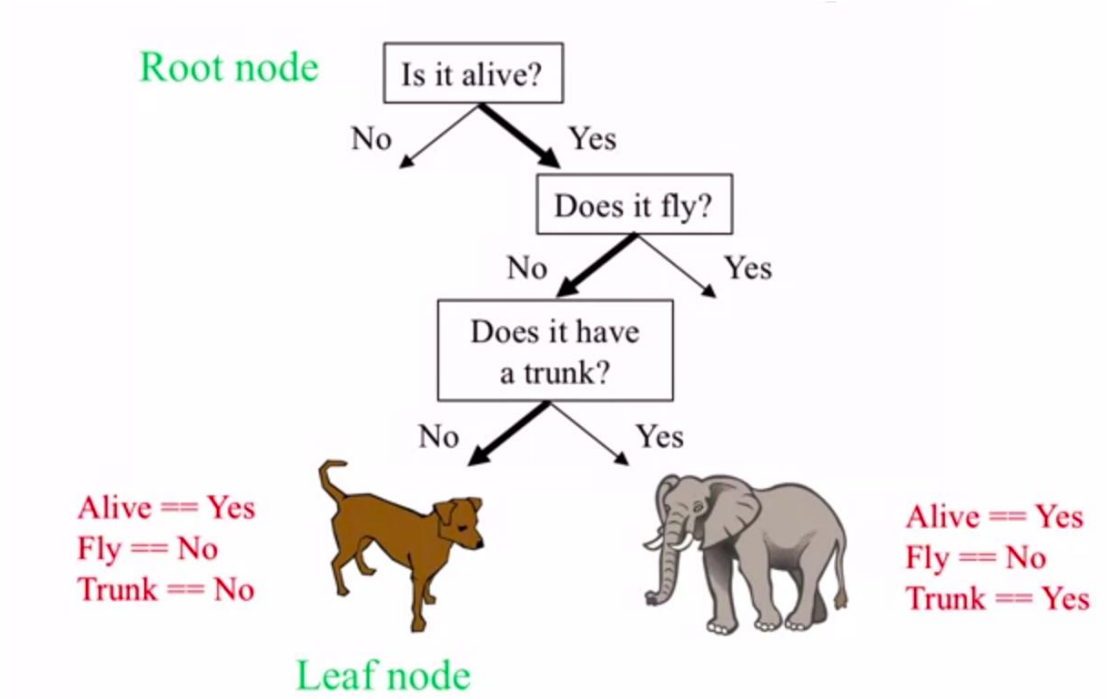


Asset Hierarchy

```
Zero Energy Nanotechnology (ZEN)
|-- Air Handling Unit Group
|    |-- AHU_02A
|    |    |-- Chilled Water Coil
|    |    |    |-- Valve Position
|    |    |-- Demand Control Ventilation
|    |    |    |-- PID Loop Output
|    |    |-- Discharge Air
|    |    |    |-- Air Flow
|    |    |    |-- Static Pressure
|    |    |    |-- Static Pressure Setpoint
|    |    |    |-- Temperature
|    |    |    |-- Temperature Setpoint
|    |    |    |-- VAV Damper Position Avg
|    |    |-- Filter
|    |    |    |-- Differential Pressure
|    |    |-- Mixed Air
|    |    |    |-- Temperature
|    |    |-- Outside Air
```

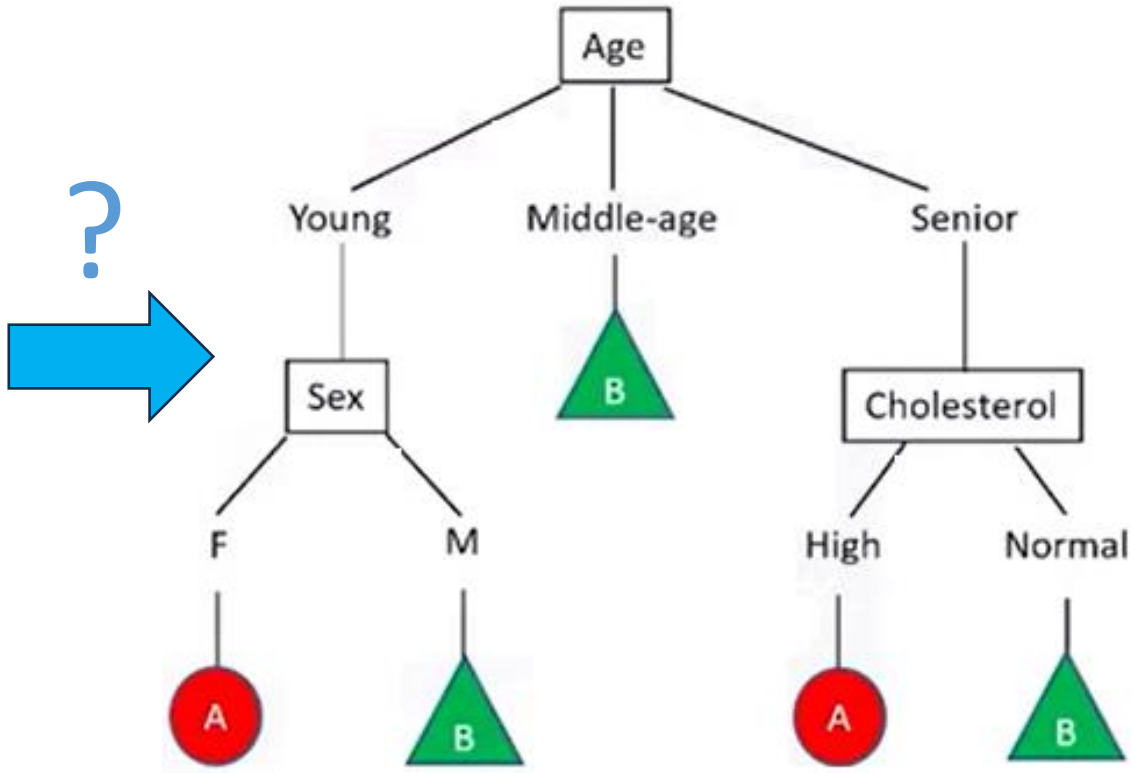
Decision Trees





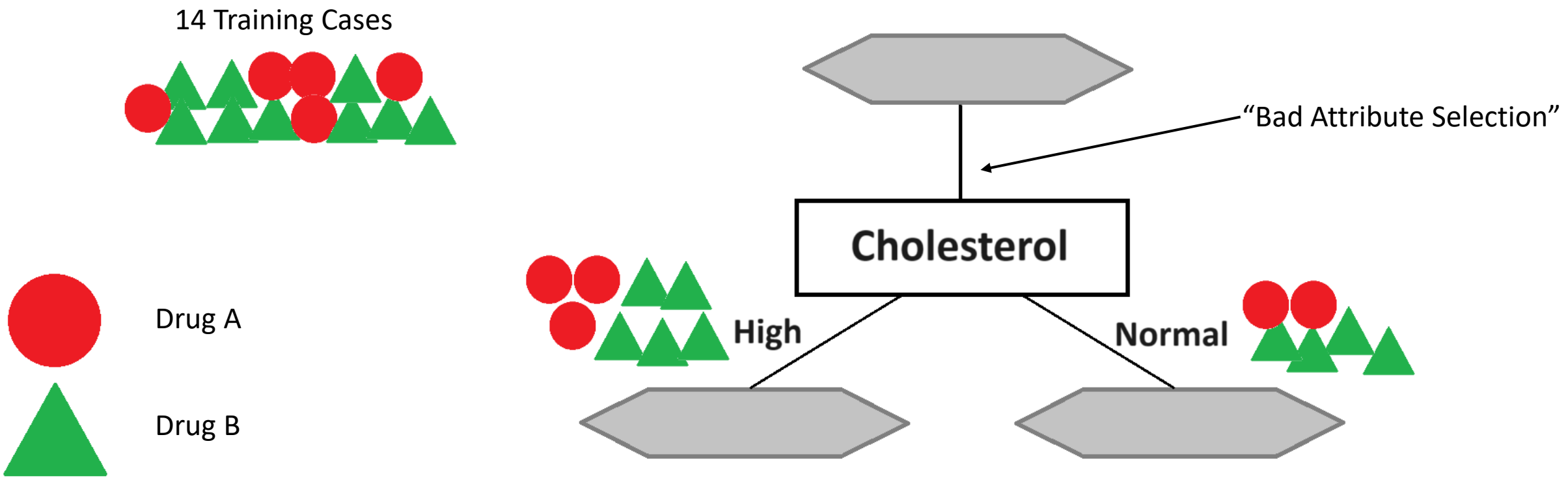
Building a Decision Tree

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?



Building a Decision Tree

- Decision trees are built using recursive partitioning to classify the data
- The algorithm chooses the most predictive feature to split the data



First…What Do We Mean by “Recursion” - Example

```
def factorial(x):
    """This is a recursive function
    to find the factorial of an integer"""

    if x == 1:
        return 1
    else:
        return (x * factorial(x-1))


num = 3
print("The factorial of", num, "is", factorial(num))
```

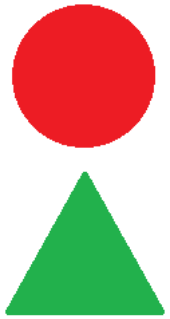
```
The factorial of 3 is 6
```

Remember: 3! = 3*2*1 = 6

Building a Decision Tree

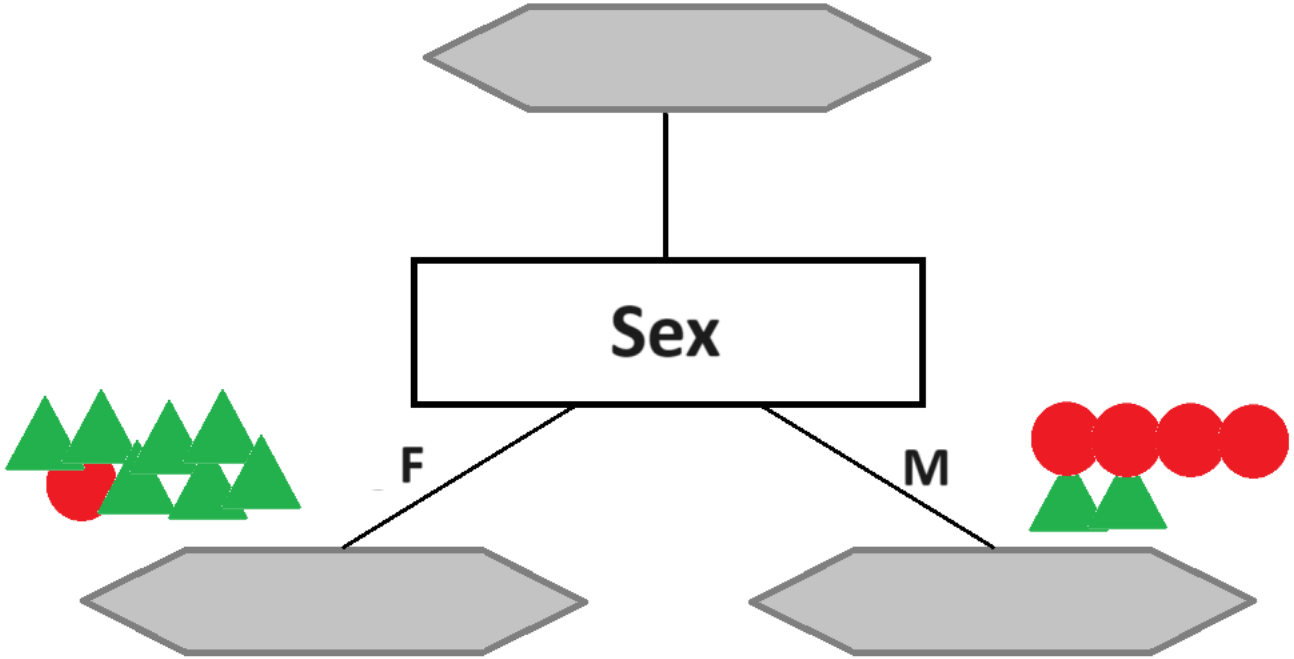
- Let's try a different attribute…

14 Training Cases



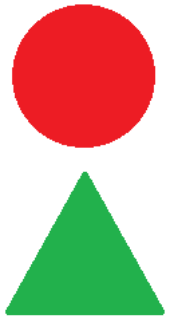
Drug A

Drug B



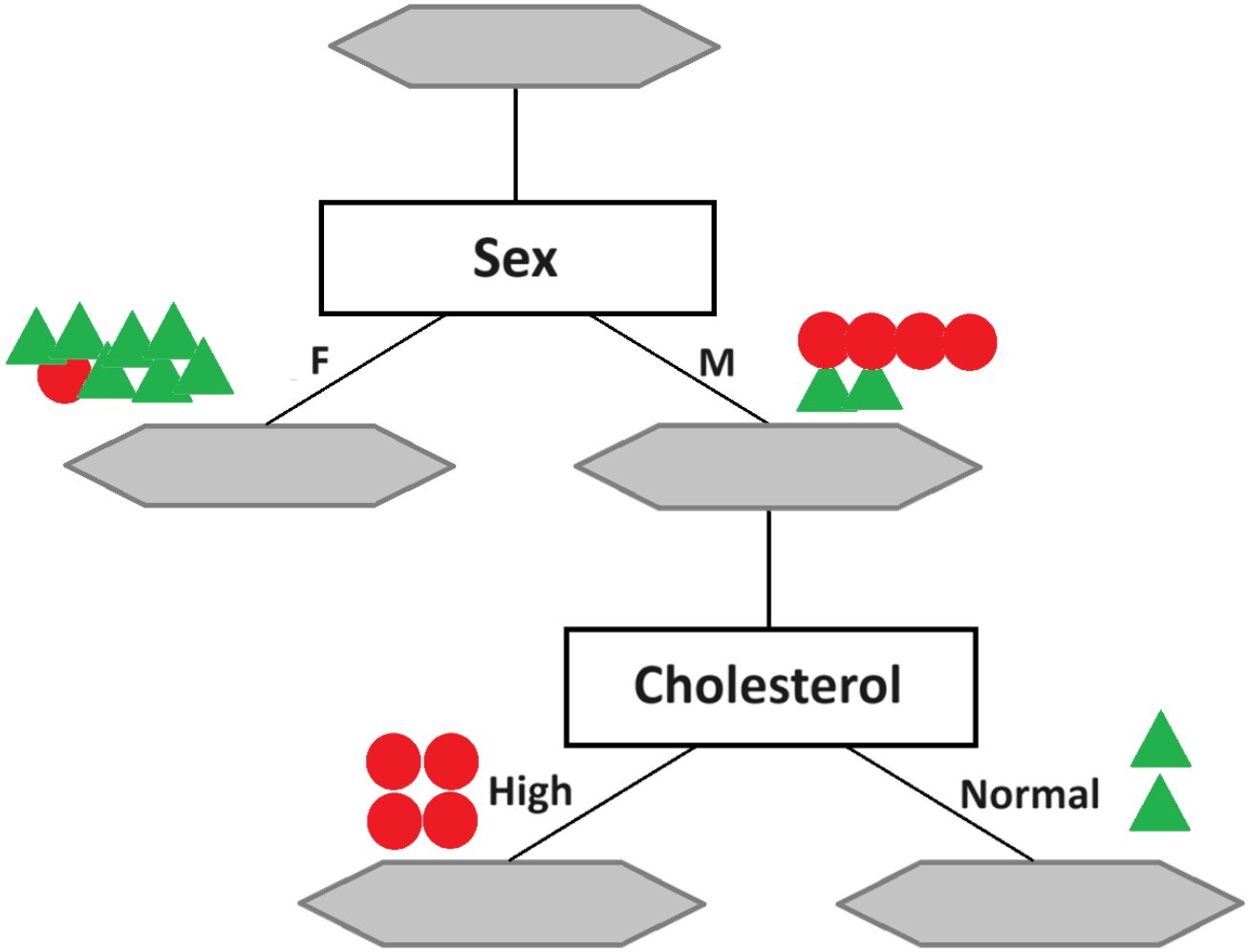
Building a Decision Tree

14 Training Cases



Drug A

Drug B

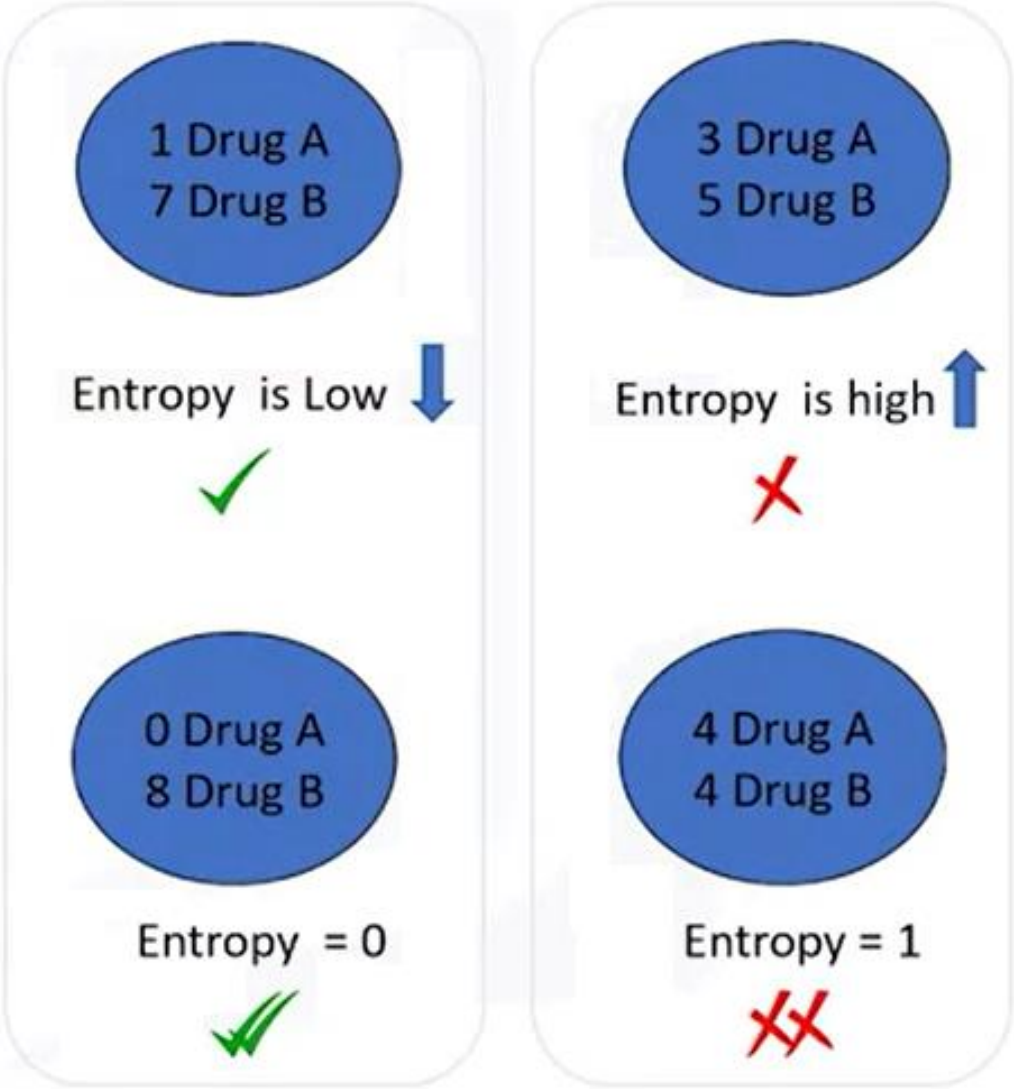


Informational Entropy

- Measure of randomness or uncertainty

$$\text{Entropy} = -p(A)\log_2(p(A)) - p(B)\log_2(p(B))$$

The lower the Entropy, the less uniform the distribution, the purer the node.



Calculating the Starting Entropy

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

S: [9 B, 5 A]

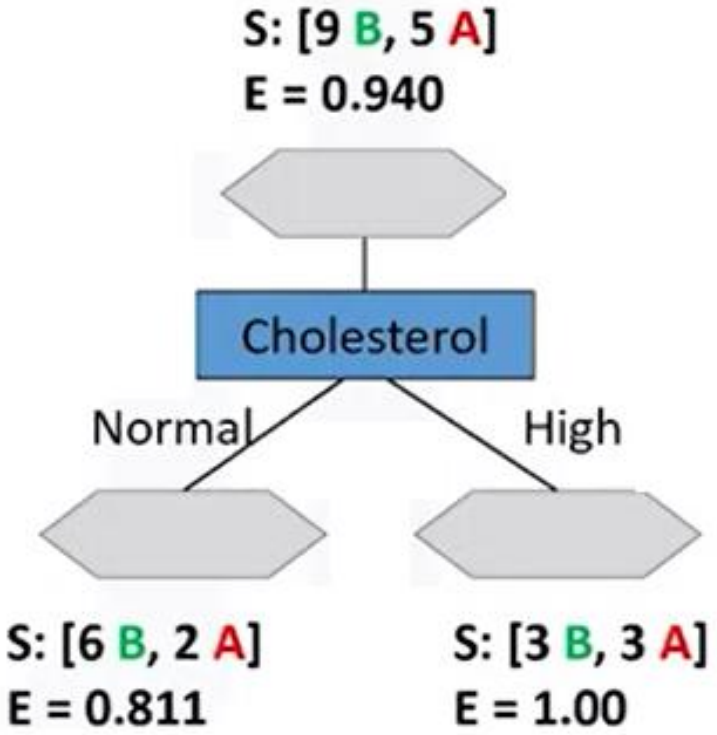
$E = - p(B)\log_2(p(B)) - p(A)\log_2(p(A))$

$E = - (9/14)\log_2(9/14) - (5/14)\log_2(5/14)$

E = 0.940

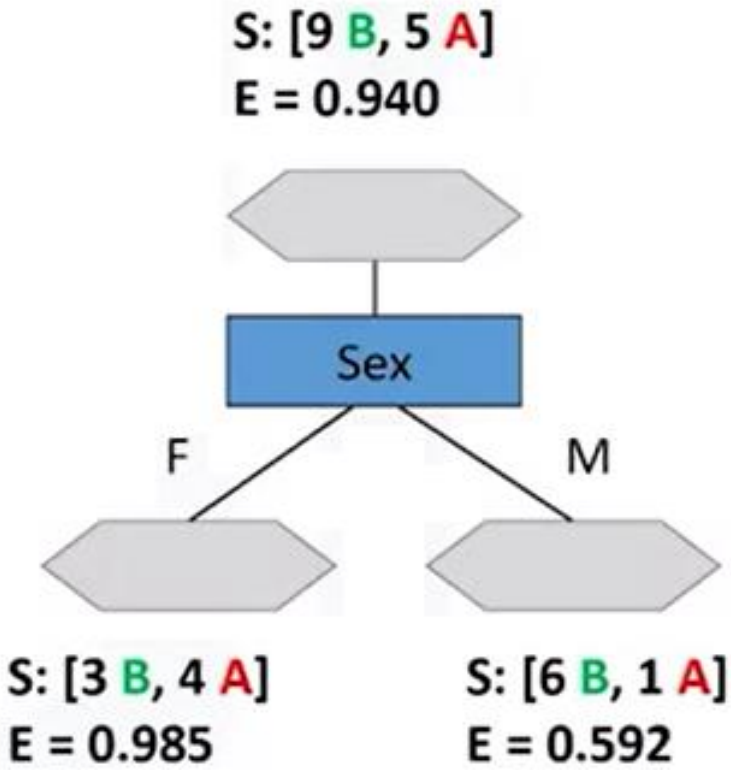
Entropy of a Cholesterol First Split

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

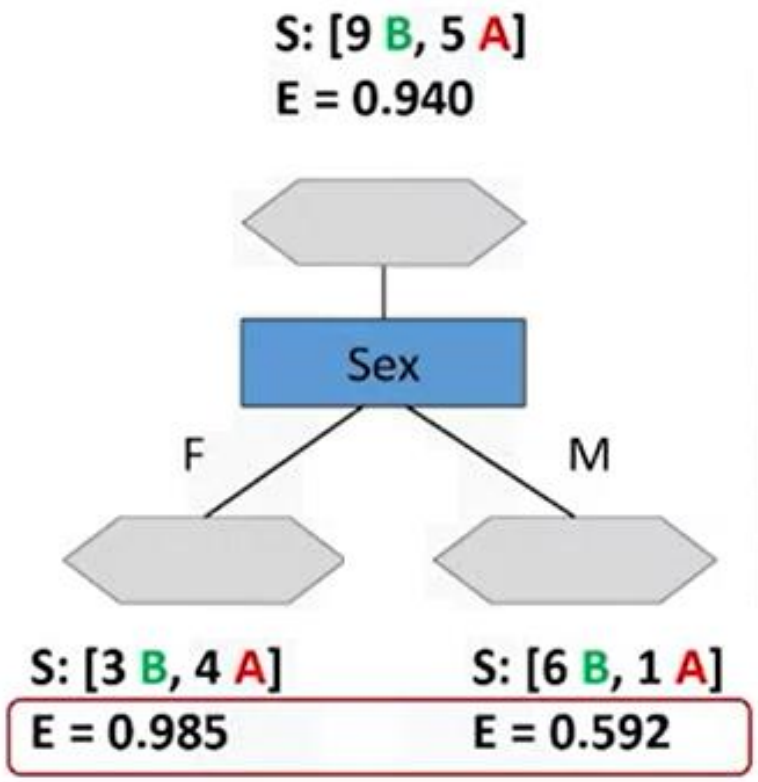


Entropy of a Sex First Split

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

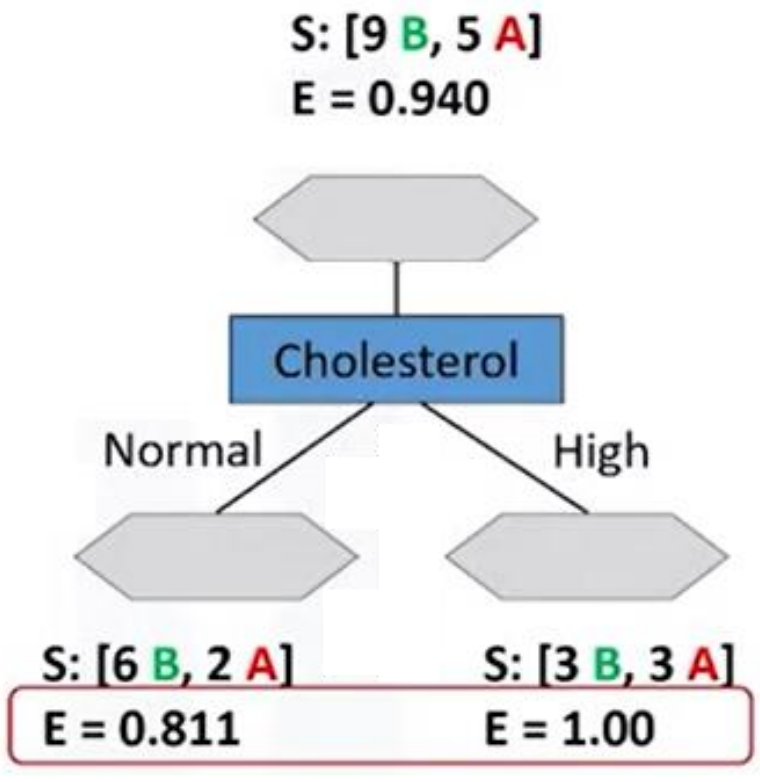


Which Attribute is Best?



Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

?

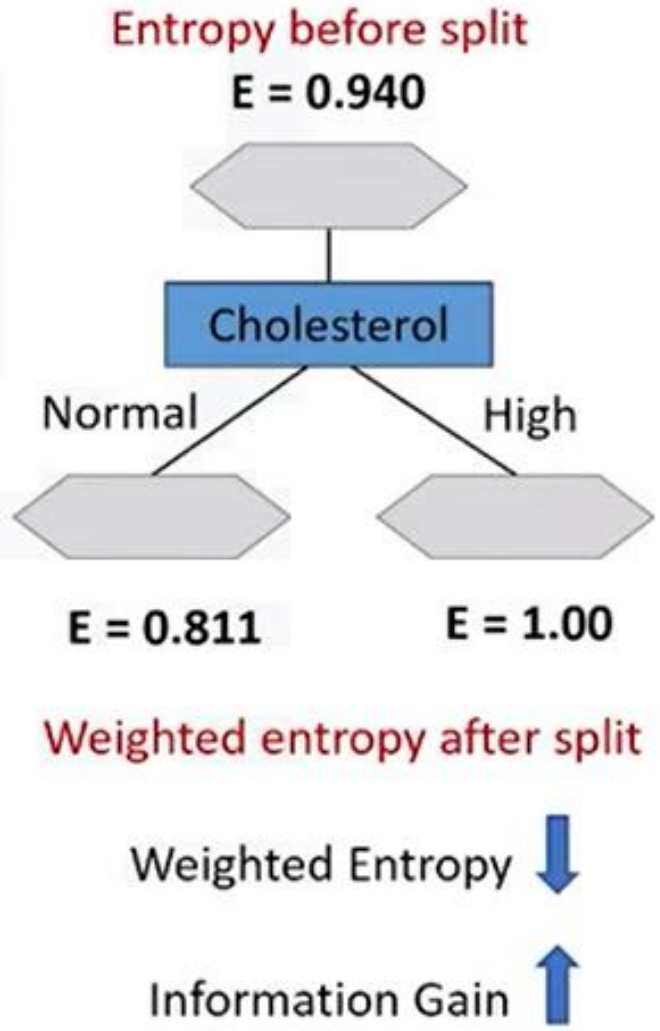


Answer: The tree with the higher Information Gain after splitting.

What is Information Gain?

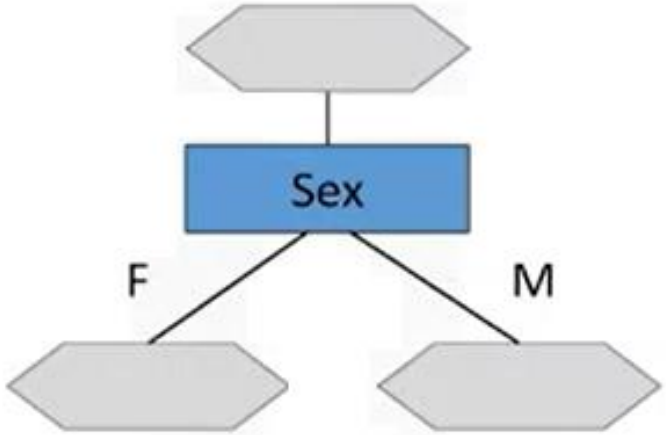
- **Information gain** is the information that can increase the level of certainty after splitting.

Information Gain = (Entropy before split) – (weighted entropy after split)



Which Attribute is Best?

S: [9 B, 5 A]
E = 0.940



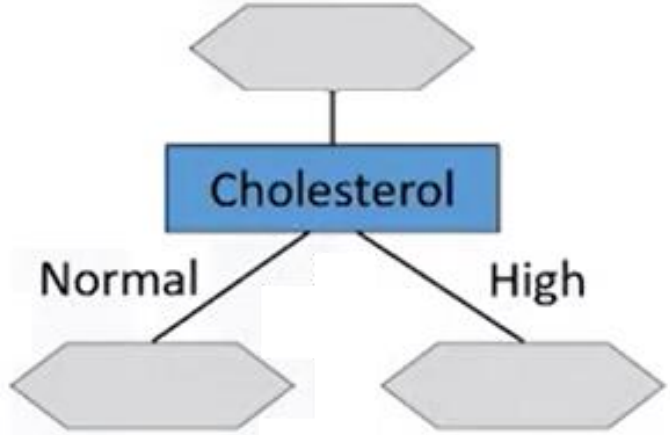
S: [3 B, 4 A]
E = 0.985

S: [6 B, 1 A]
E = 0.592

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

?

S: [9 B, 5 A]
E = 0.940



S: [6 B, 2 A]
E = 0.811

S: [3 B, 3 A]
E = 1.00

Gain (s, Sex)
= 0.940 – [(7/14)0.985 + (7/14)0.592]
=0.151

Gain (s, Cholesterol)
= 0.940 – [(8/14).811 + (6/14)1.0]
=0.048