Decision Tree

Questions to be answered

- What is decision tree?
- How do we use them to help us classify?
- How can I grow my decision tree?

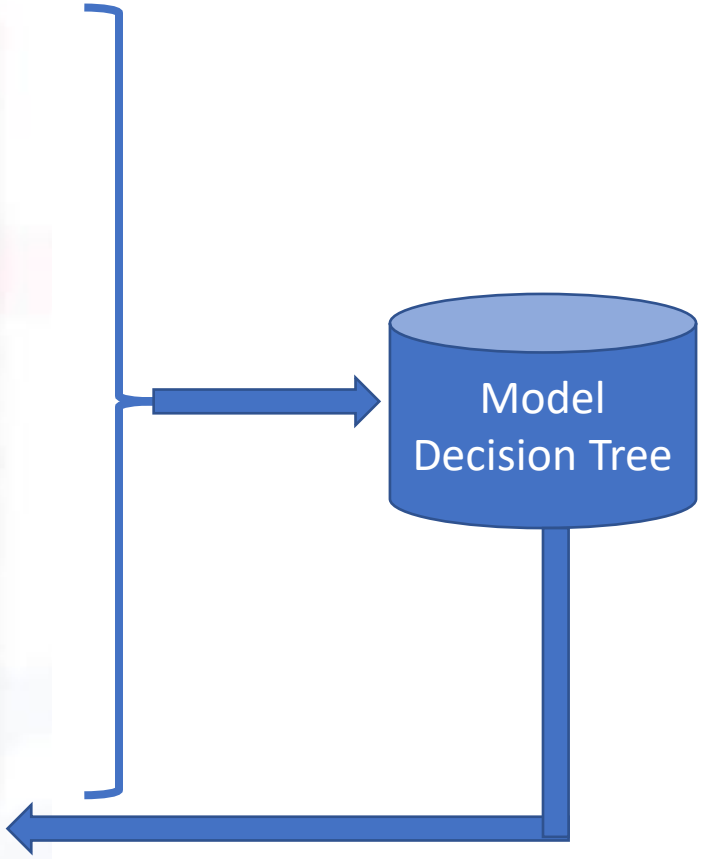
“The basic intuition behind a decision tree is to map out all possible decision paths in the form of a tree. “
Narendra Nath Joshi

How to build a decision tree?

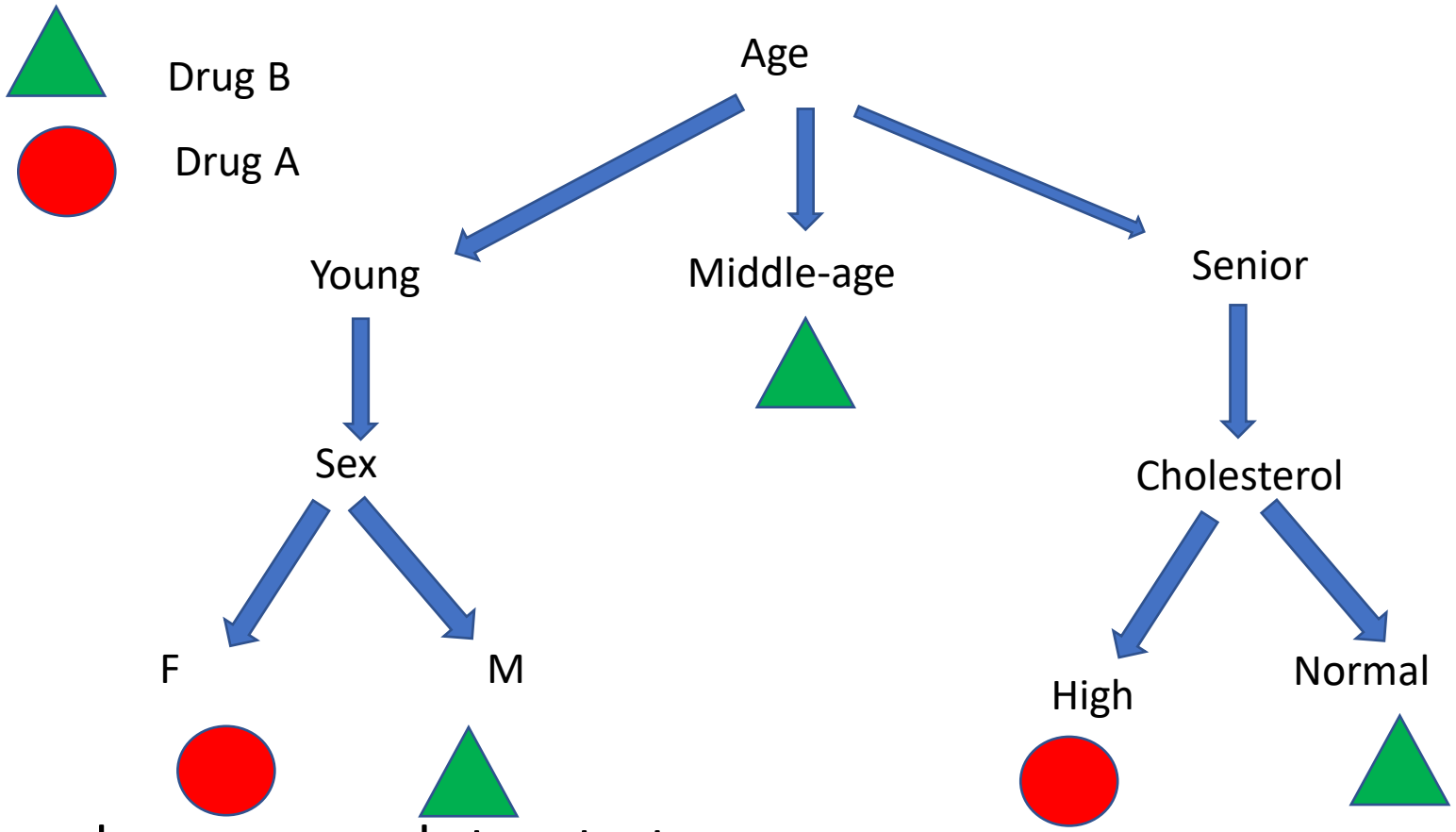
- Imagine that you are a medical researcher compiling data for a study. You already collected the data all of whom suffer from the same illness. During the treatment, each patient responded to one of two medications.
- What drug might be appropriate for the future patient with the same illness?

How to build 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	?



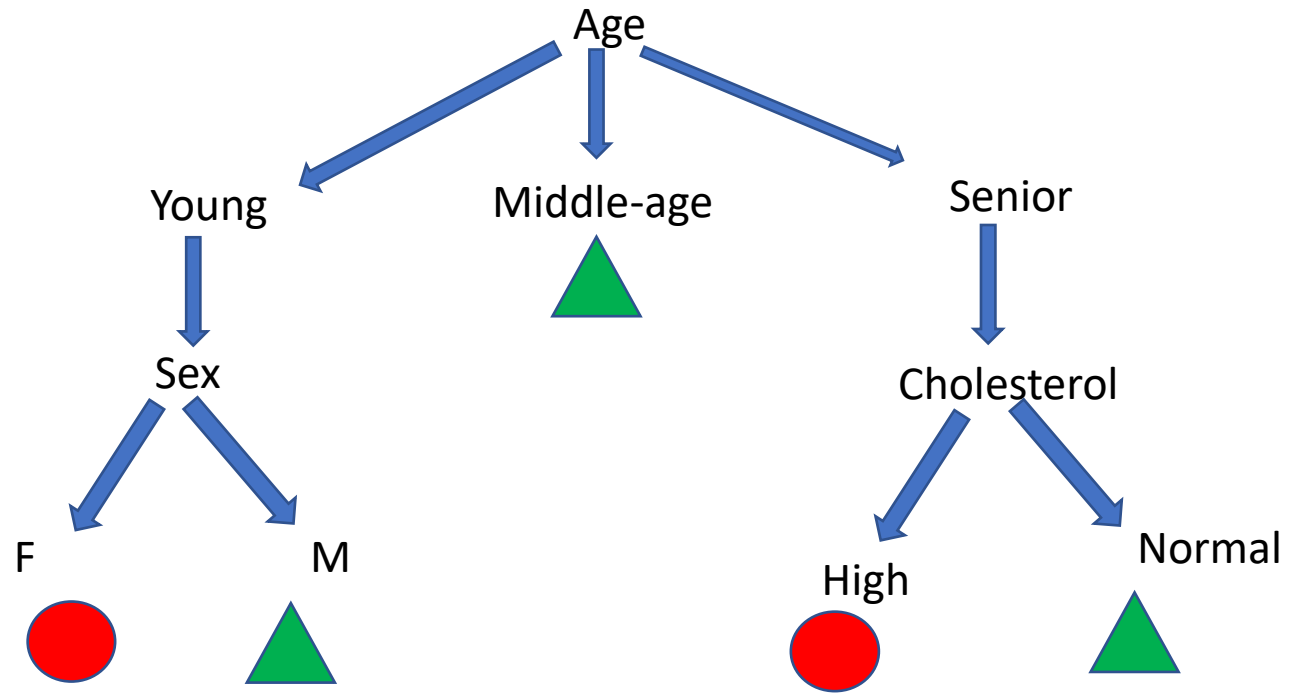
Building a decision tree with the training set



- Each internal node corresponds to a test
- Each branch corresponds to a result of the test
- Each leaf node assigns a classification

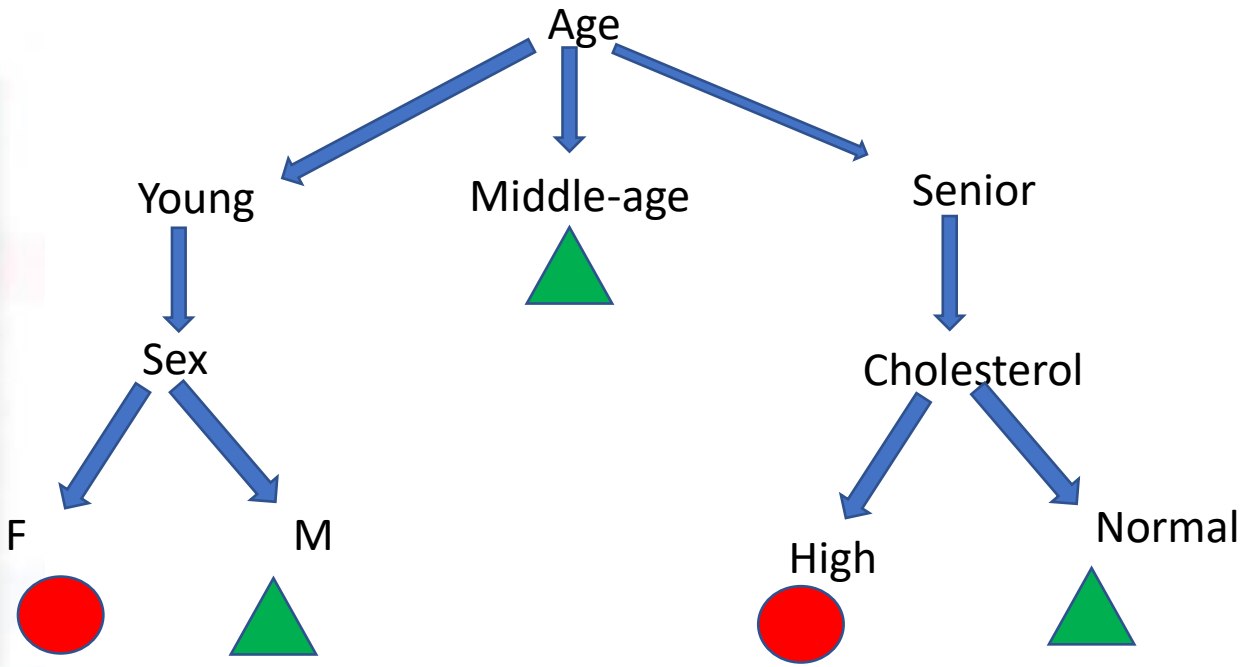
Decision Tree Learning Algorithms

1. Choose the attribute from the dataset.
2. Calculate the significance of attribute in splitting of data
3. Split data based on the value of the best attribute.
4. Go to step 1.

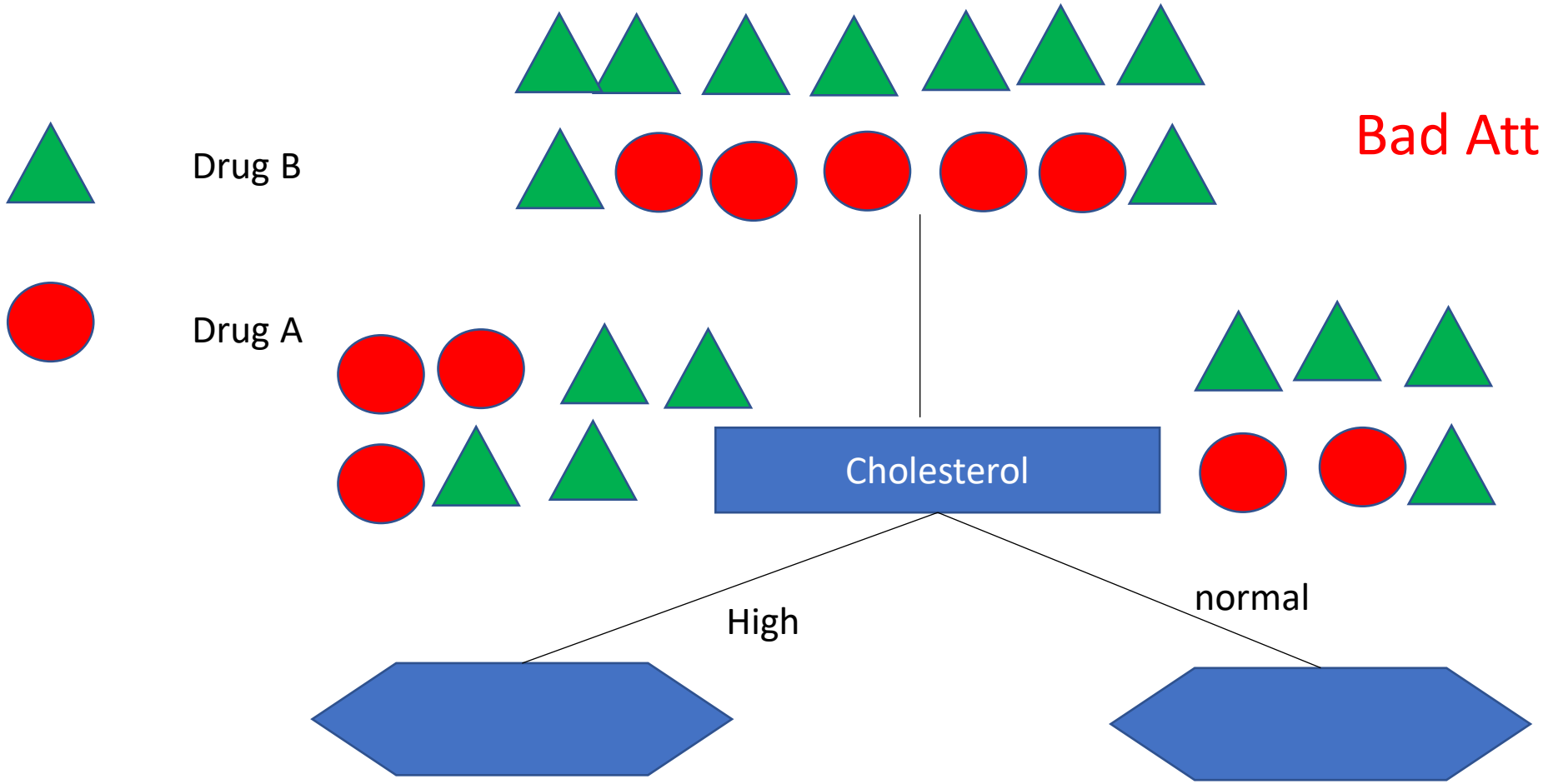


Building the tree
How do we build the decision tree ?

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	?

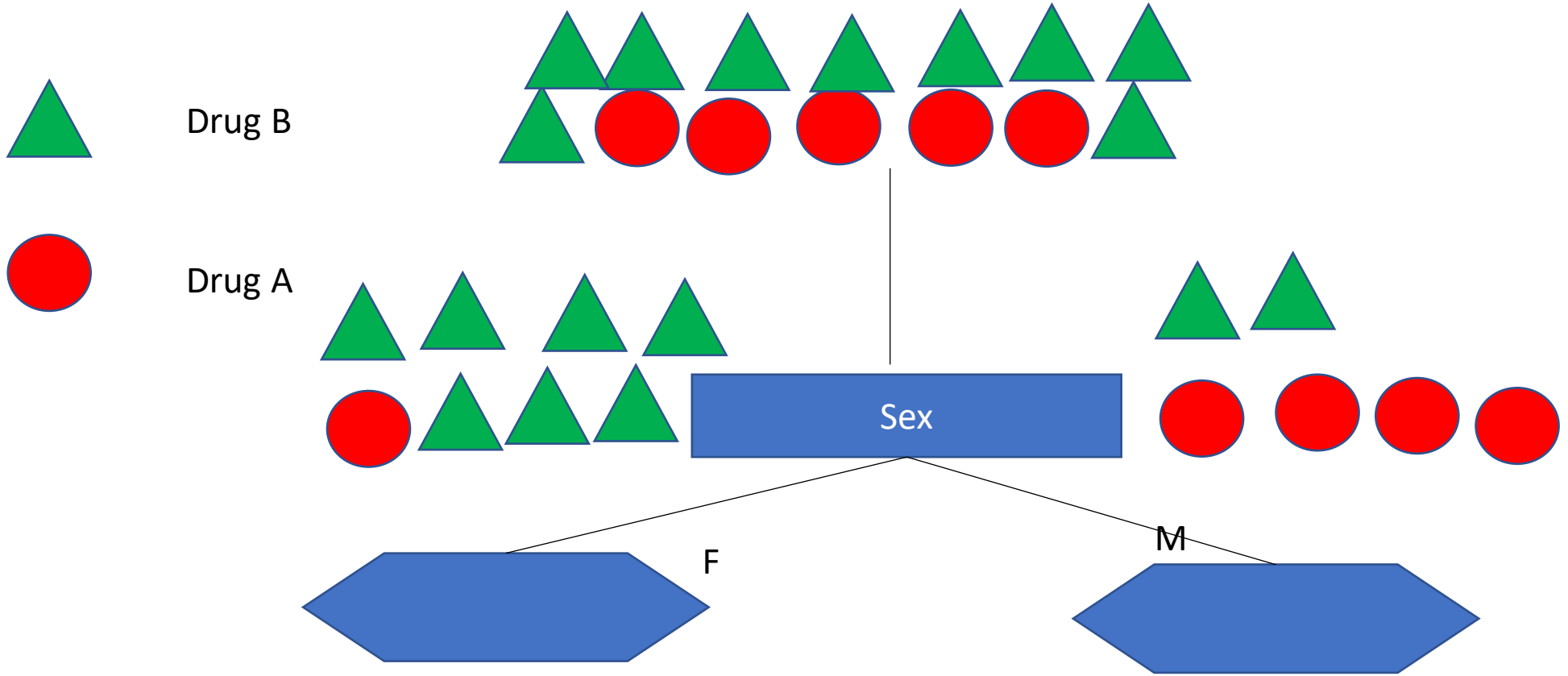


Building the tree
Which attribute is the best?



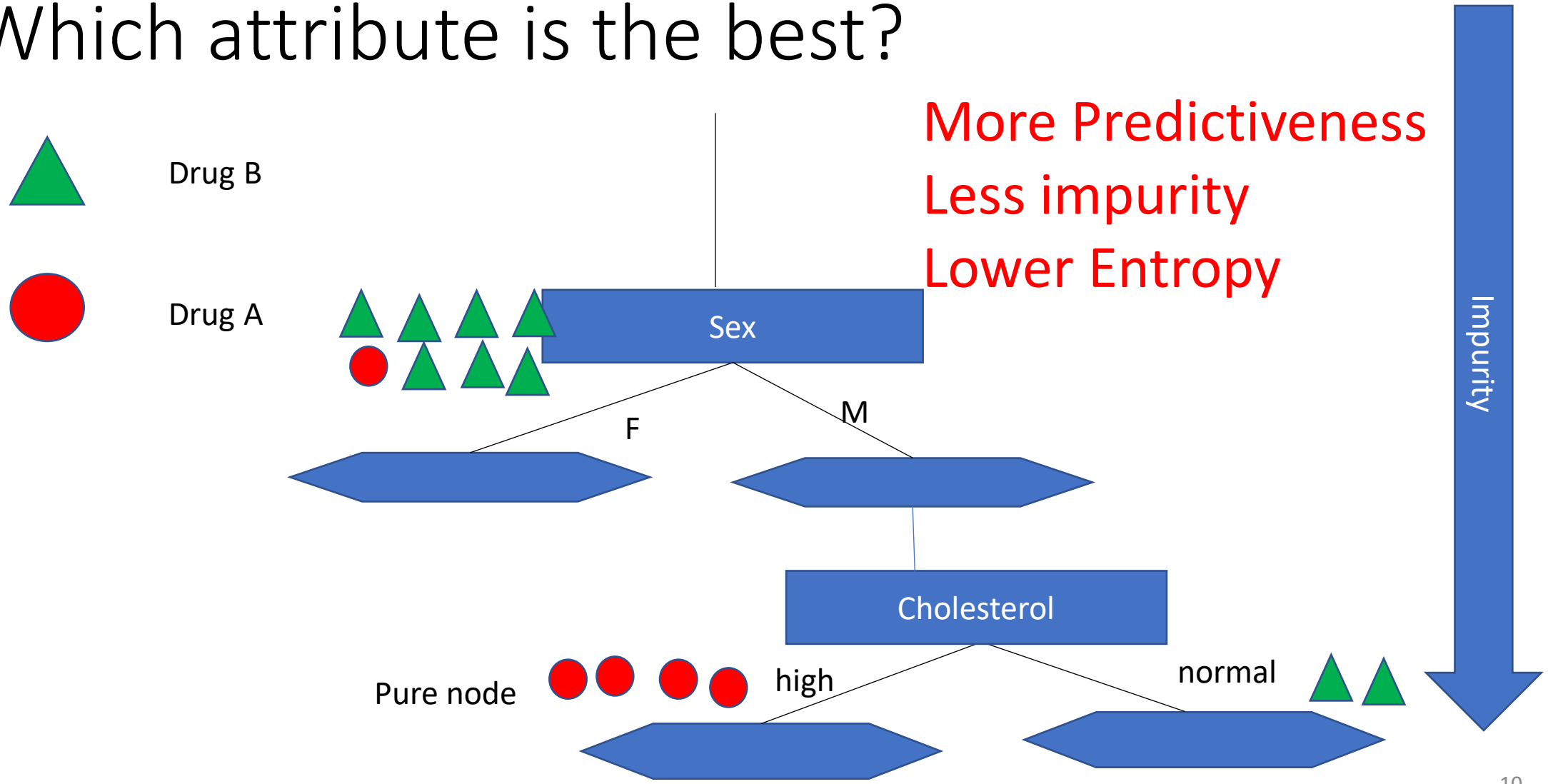
Bad Attribute!

Building the tree
Which attribute is the best?



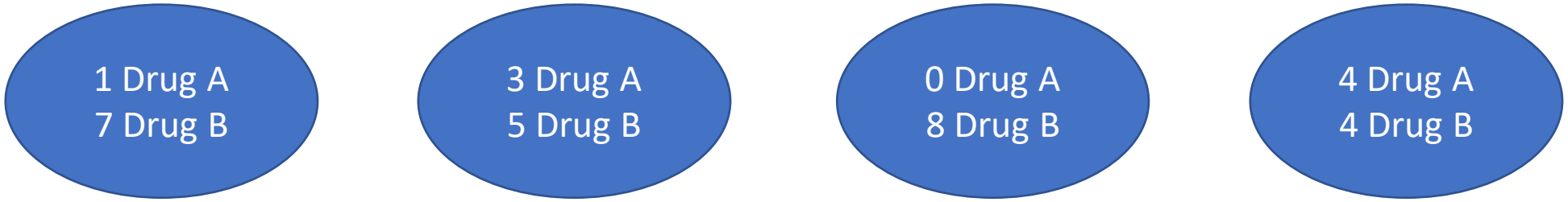
We can say that sex attribute is more predictive than the cholesterol attribute. Predictiveness is based on decreasing the impurity of node.

Building the tree
Which attribute is the best?



Building the Tree
What is entropy?

- Entropy is the amount of information disorder or the amount of the randomness in data.
- It is the measure of the randomness or uncertainty

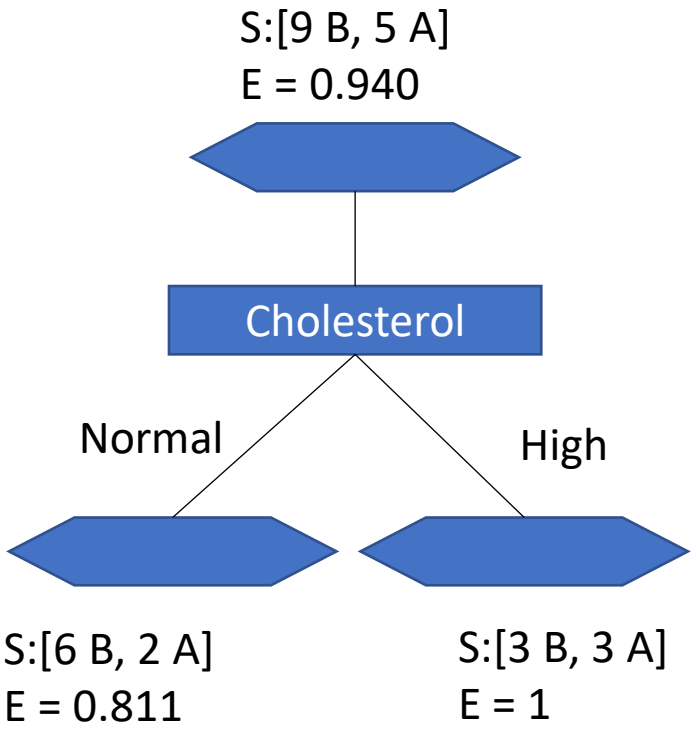


Entropy is low | Entropy is high | Entropy 0 | Entropy = 1

- The lower the entropy , the less uniform the distribution, the purer the node.

Building the tree
Which attribute is the 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

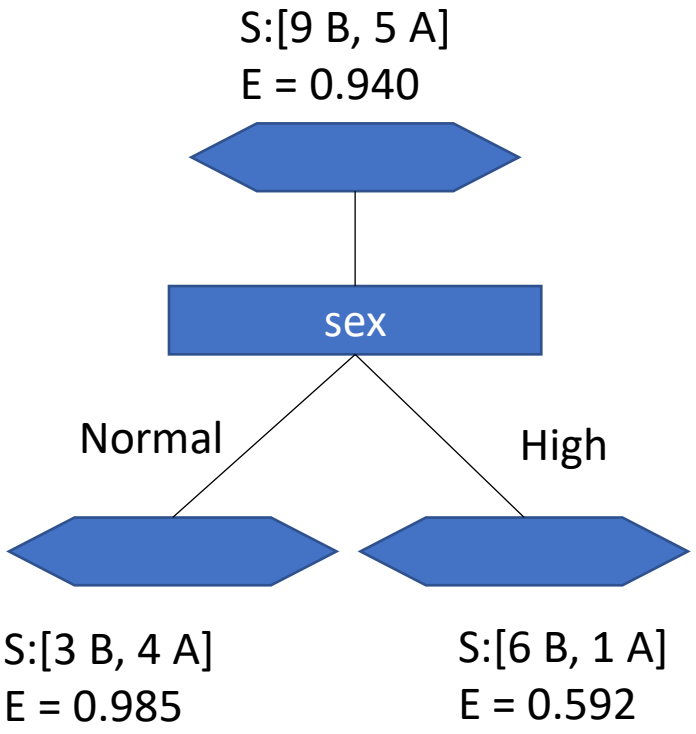


S:[9 B, 5 A]
E = -p(B)log(p(B))-P(A)log(P(A))
E =-(9/14)log(9/14)-(5/14)log(5/14)
E = 0.940

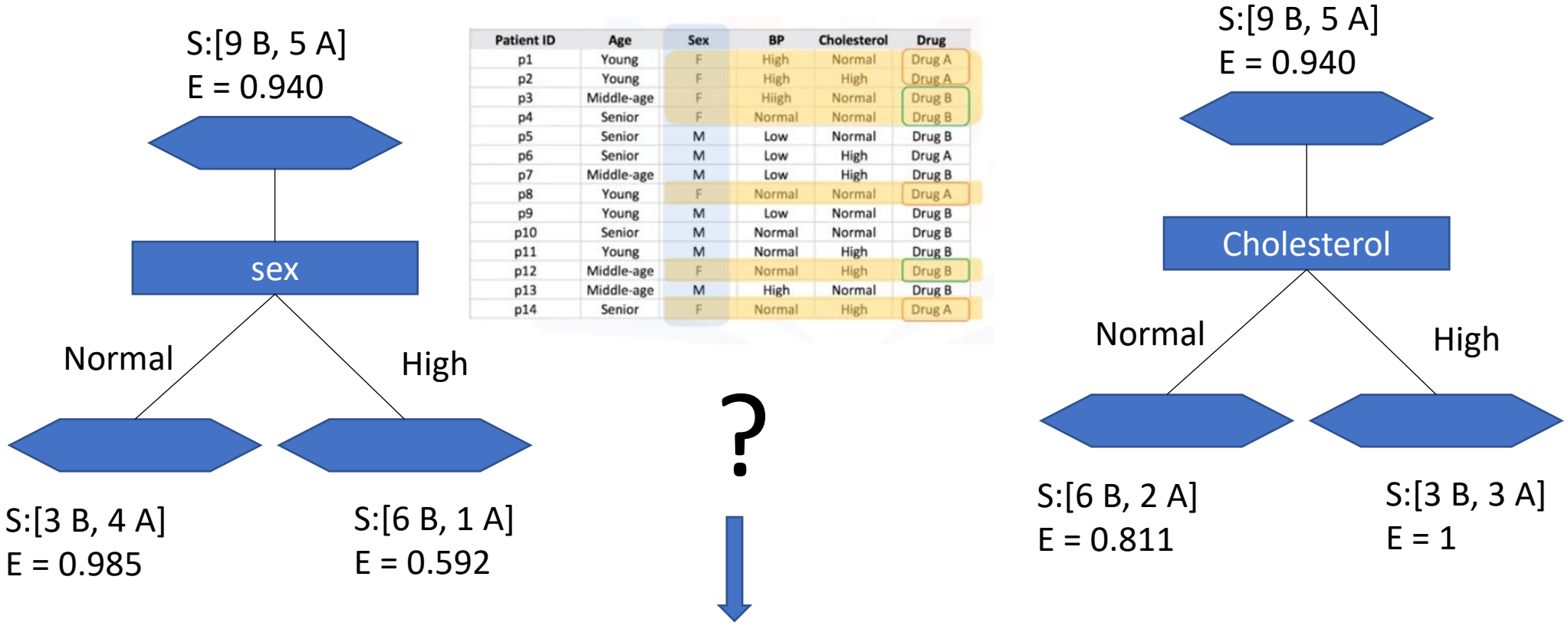
You should always to through all the attribute to calculate the entropy, then choose the best

Building the tree
Which attribute is the 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



Building the Tree
Which one is better ?



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

The tree with the higher information gain after splitting

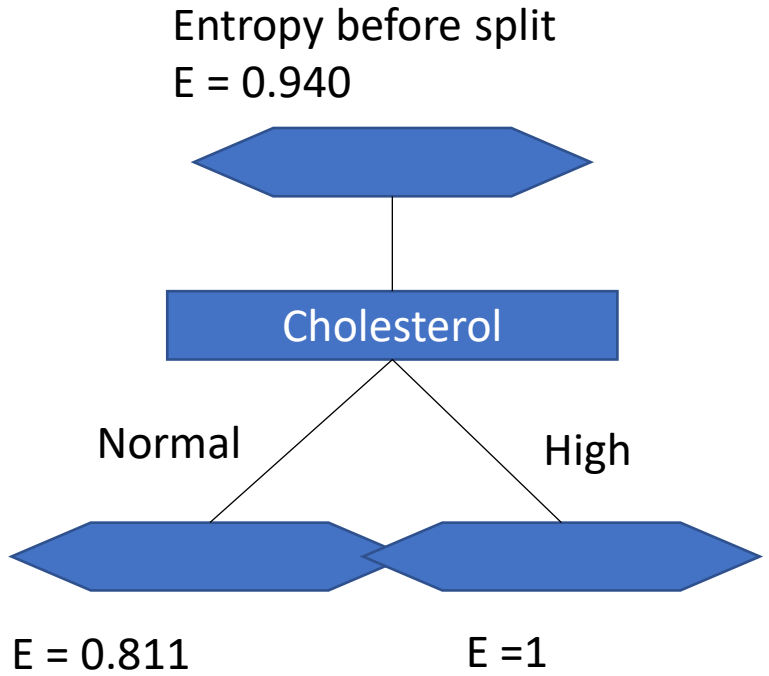
Building the tree
What is information gain?

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

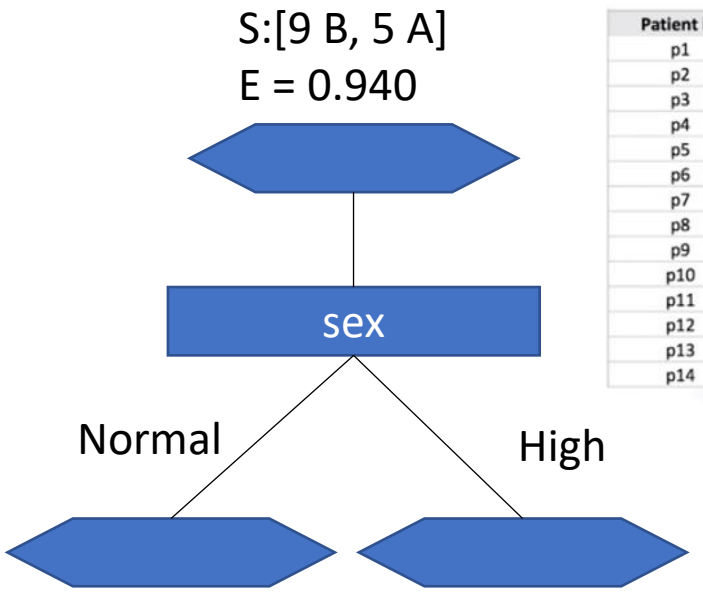
Information Gain = (Entropy before split)- (weighted entropy after splitting)

Weighted Entropy

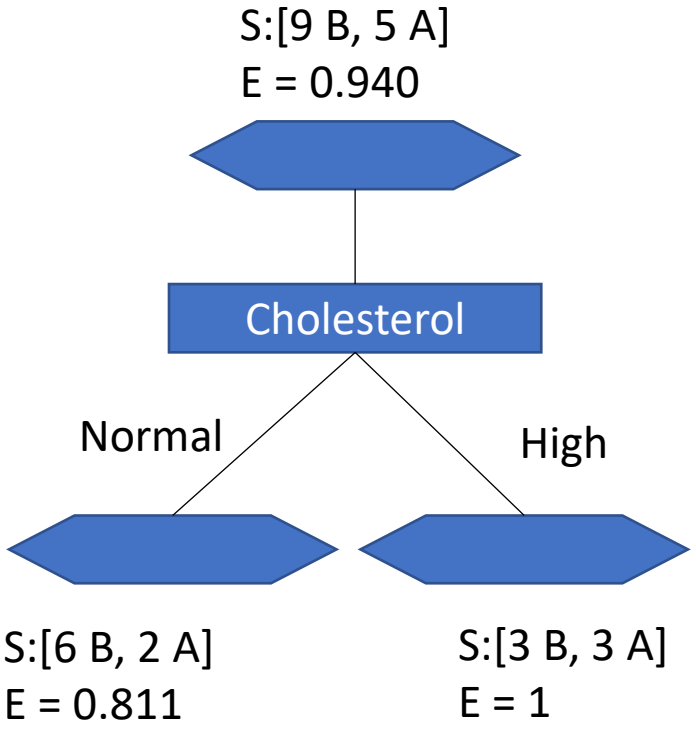
Information Gain



Building the Tree
Which one is better ?



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



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

Gain(s,Cholesterol=0.048