Example

“Hours studied” as continuous variable and “A on the test” as classifier. The data looks like this(pre-sorted for your convenience):

HOURS STUDIED	A ON TEST
4	N
5	Y
8	N
12	Y
15	Y

In this case, we have data relating the number of hours various students studied in an attempt to determine its effect on their test performance. We want to discretize this data, so let's start by calculating entropy of the data set itself:

	A ON TEST	LOWER THAN A
Overall	3	2

$$Entropy(D) = -(\frac{3}{5}log_2(\frac{3}{5}) + \frac{2}{5}log_2(\frac{2}{5})) = .529 + .442 = .971$$

For all of our samples, the entropy to beat is .971! Now, let's iterate through and see which splits give us the maximum entropy gain. To find a split, we average two neighboring values in the list.

Split 1: 4.5

Starting off, we split at 4.5 ((5+4)/2). Now we get two bins, as follows:

	A ON TEST	LOWER THAN A
<=4.5	0	1
>4.5	3	1

Now, we calculate entropy for each bin and find the information gain of this split:

$$Entropy(D_{<=4.5}) = -(\frac{1}{1}log_2(1) + 0log_2(0)) = 0 + 0 = 0$$

$$Entropy(D_{>4.5}) = -(\frac{3}{4}log_2(\frac{3}{4}) + \frac{1}{4}log_2(\frac{1}{4})) = .311 + .5 = .811$$

Now net entropy is:

$$Info_a(D_{new}) = \frac{1}{5}(0) + \frac{4}{5}(.811) = .6488$$

And our gain is:

$$Gain(D_{new}) = .971 - .6488 = .322$$

Looking good! Our maximum entropy gain is now .322. That's good, but we still need to go through the rest.

Split 2: 6.5

Average our next two values, and we get 6.5. Now we repeat the process for this split:

	A ON TEST	LOWER THAN A
<=6.5	1	1
>6.5	2	1

Again, calculate the entropy for each bin:

$$Entropy(D_1) = -(.5log_2(.5) + .5log_2(.5)) = 1$$

$$Entropy(D_1) = -(\frac{2}{3}log_2(\frac{2}{3}) + \frac{1}{3}log_2(\frac{1}{3})) = .389 + .528 = .917$$

Wow! This one is looking bad already, but let's finish it.

$$info_a = \frac{1}{3}(1) + \frac{2}{3}(.917) = .944$$

$$Gain(D_{new}) = .971 - .944 = .027$$

This is less gain than we had before, so our best split point is still at 4.5. Let's check the next one now.

Split 3: 10

Average our next two values, and we get 10. Our split now looks like this:

	A ON TEST	LOWER THAN A
<=10	1	2
>10	2	0

As before, we calculate entropy for each bin and determine information:

$$Entropy(D_{<=10}) = \frac{1}{3}log_2(\frac{1}{3}) + \frac{2}{3}log_2(\frac{2}{3}) = .917$$

$$Entropy(D_{>10}) = -(1log_2(1) + 0log_2(0)) = 0$$

$$info_a = \frac{2}{5}(0) + \frac{3}{5}(.917) = .55$$

Finally, we calculate gain once more.

$$Gain(D_{new}) = .971 - .55 = .421$$

This is the clear winner at this point!

Split 4: 13.5

And for our final potential split, we split at 13.5. Now our splits look like this! For the sake of brevity, I won't calculate this one, but we can infer based on the data below that it would have poor entropy gain because the first bin has a 50/50 distribution.

	A ON TEST	LOWER THAN A
<=13.5	2	2
>13.5	1	0

Choose the split

After calculating everything, we find that our best split is split 3, which gives us the best information gain of .421. We will partition the data there!

According to the algorithm, we now can further bin our attributes in the bins we just created. This process will continue until we satisfy a termination criteria.

When to Terminate

There are two popular options for stopping the algorithm:

1. **Terminate when a specified number of bins has been reached.** This makes sense when you value the degree to which you understand your data. A data set with 3 bins is much easier to think about than one with 12.
2. **Terminate when information gain falls below a certain threshold.**This is another common method. If information gain drops below a certain value, the algorithm can terminate early. There are a number of ways to calculate what the minimum threshold should be, and it can also be determined empirically through testing.

Oftentimes, both of these criteria will be used in conjunction to yield optimal results.