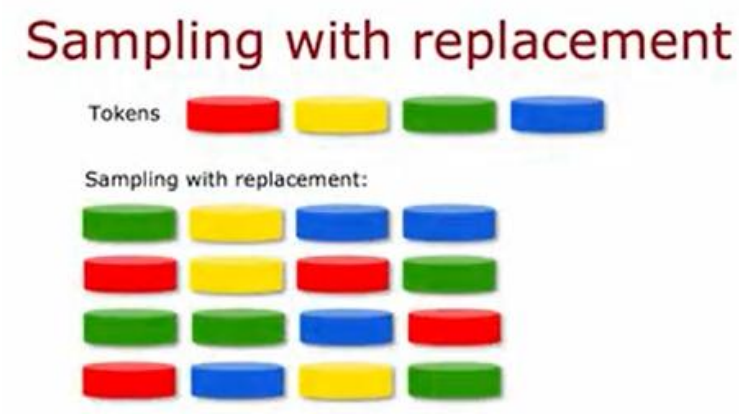
# Sampling with replacement

In order to build a tree together, we're going to need a technique called sampling with replacement. Let's take a look at what that means. In order to illustrate how sampling with replacement works, I'm going to show you a demonstration of sampling with replacement using four tokens that are colored red, yellow, green, and blue. I actually have here with me four tokens of colors, red, yellow, green, and blue. I'm going to demonstrate what sampling with replacement using them look like. Here's a black velvet bag, empty. I'm going to take this example of four tokens, and drop them in.

I'm going to sample four times with replacement out of this bag. What that means, I'm going to shake it up, and can't see when I'm picking, pick out one token, turns out to be green. The term with replacement means that if I take out the next token, I'm going to take this, and put it back in, and shake it up again, and then take on another one, yellow. Replace it. That's a little replacement part. Then go again, blue replace it again, and then pick on one more, which is blue again. That sequence of tokens I got was green, yellow, blue, blue.

Notice that I got blue twice, and didn't get red even a single time. If you were to repeat this sampling with replacement procedure multiple times, if you were to do it again, you might get red, yellow, red, green, or green, green, blue, red. Or you might also get red, blue, yellow, green. Notice that the with replacement part of this is critical because if I were not replacing a token every time I sample, then if I were to pour four tokens from my bag of four, I will always just get the same four tokens.

That's why replacing a token after I pull it out each time, is important to make sure I don't just get the same four tokens every single time. The way that sampling with replacement applies to building an ensemble of trees is as follows. We are going to construct multiple random training sets that are all slightly different from our original training set.




In particular, we're going to take our 10 examples of cats and dogs. We're going to put the 10 training examples in a theoretical bag. Please don't actually put a real cat or dog in a bag. That sounds inhumane, but you can take a training example and put it in a theoretical bag if you want. I'm using this theoretical bag, we're going to create a new random training set of 10 examples of the exact same size as the original data set. The way we'll do so is we're reaching and picking out one random training example. Let's say we get this training example.

Then we put it back into the bag, and then again randomly pick out one training example and so you get that. You pick again and again and again. Notice now this fifth training example is identical to

the second one that we had out there. But that's fine. You keep going and keep going, and we get another repeats the example, and so on and so forth. Until eventually you end up with 10 training examples, some of which are repeats. You also notice that this training set does not contain all 10 of the original training examples, but that's okay.

That is part of the sampling with replacement procedure. The process of sampling with replacement, lets you construct a new training set that's a little bit similar to, but also pretty different from your original training set. It turns out that this would be the key building block for building an ensemble of trees.