At long last, we're ready to split our data into a training and testing set, and to actually build a model.

So we'll start by loading the ca tools package, so that we can split our data.

So we'll do library(caTools).

And then, as usual, we're going to set our random seeds so that everybody has the same results.

So use set.seed and we'll pick the number 144.

Again, the number isn't particularly important.

The important thing is that we all use the same one.

So as usual, we're going to obtain the split variable.

We'll call it spl, using the sample.split.

The outcome variable that we pass is labeledTerms$responsive.

And we'll do a 70/30 split.

So we'll pass 0.7 here.

So then train, the training data frame, can be obtained using subset on the labeled terms where spl is true.

And test is the subset when spl is false.

So now we're ready to build the model.

And we'll build a simple cart model using the default parameters.

But a random forest would be another good choice from our toolset.

So we'll start by loading up the packages for the cart model.

We'll do library(rpart).

And we'll also load up the rpart.plot package, so that we can plot the outcome.

So we'll create a model called email cart, using the r part function.

We're predicting responsive.

And we're predicting it using all of the additional variables.

All the frequencies of the terms that are included.

Obviously tilde period is important here, because there are 788 terms.

Way too many to actually type out.

The data that we're using to train the model is just our training dataframe, train.

And then the method is class, since we have a classification problem here.

And once we've trained the cart model, we can plot it out using prp.

There we go.

So we can see at the very top is the word California.

If California appears at least twice in an email, we're going to take the right part over here and predict that a document is responsive.

It's somewhat unsurprising that California shows up, because we know that Enron had a heavy involvement in the California energy markets.

So further down the tree, we see a number of other terms that we could plausibly expect to be related to energy bids and energy scheduling, like system, demand, bid, and gas.

Down here at the bottom is Jeff, which is perhaps a reference to Enron's CEO, Jeff Skillings, who ended up actually being jailed for his involvement in the fraud at the company.