The topic of this recitation is election forecasting, which is the art and science of predicting the winner of an election before any votes are actually cast using polling data from likely voters.

In this recitation, we are going to look at the United States presidential election.

In the United States, a president is elected every four years.

And while there are a number of different political parties in the US, generally there are only two competitive candidates.

There's the Republican candidate, who tends to be more conservative, and the Democratic candidate, who's more liberal.

So for instance a recent Republican president was George W. Bush, and a recent Democratic president was Barack Obama.

Now while in many countries the leader of the country is elected using the simple candidate who receives the largest number of votes across the entire country is elected, in the United States it's significantly more complicated.

There are 50 states in the United States, and each is assigned a number of electoral votes based on its population.

So for instance, the most populous state, California, in 2012 had nearly 20 times the number of electoral votes as the least populous states.

And these number of electoral votes are reassigned periodically based on changes of populations between states.

Within a given state in general, the system is winner take all in the sense that the candidate who receives the most vote in that state gets all of its electoral votes.

And then across the entire country, the candidate who receives the most electoral votes wins the entire presidential election.

Now while it seems like a somewhat subtle distinction, the electoral college versus the simple popular vote model, it can have very significant consequences on the outcome of the election.

As an example, let's look at the 2000 presidential election between George W. Bush and Al Gore.

As we can see on the right here, Al Gore received more than 500,000 more votes across the entire country than

George W. Bush in terms of the popular vote.

But in terms of the electoral vote, because of how those votes were distributed, George Bush actually won the election because he received five more electoral votes than Al Gore.

So our goal will be to use polling data that's collected from likely voters before the election to predict the winner in each state, and therefore to enable us to predict the winner of the entire election in the electoral college system.

While election prediction has long attracted some attention, there's been a particular interest in the problem for the 2012 presidential election, when then-New York Times columnist Nate Silver took on the task of predicting the winner in each state.

To carry out this prediction task, we're going to use some data from RealClearPolitics.com that basically represents polling data that was collected in the months leading up to the 2004, 2008, and 2012 US presidential elections.

Each row in the data set represents a state in a particular election year.

And the dependent variable, which is called Republican, is a binary outcome.

It's 1 if the Republican won that state in that particular election year, and a 0 if a Democrat won.

The independent variables, again, are related to polling data in that state.

So for instance, the Rasmussen and SurveyUSA variables are related to two major polls that are assigned across many different states in the United States.

And it represents the percentage of voters who said they were likely to vote Republican minus the percentage who said they were likely to vote Democrat.

So for instance, if the variable SurveyUSA in our data set has value -6, it means that 6% more voters said they were likely to vote Democrat than said they were likely to vote Republican in that state.

We have two additional variables that capture polling data from a wider range of polls.

Rasmussen and SurveyUSA are definitely not the only polls that are run on a state by state basis.

DiffCount counts the number of all the polls leading up to the election that predicted a Republican winner in the state, minus the number of polls that predicted a Democratic winner.

And PropR, or proportion Republican, has the proportion of all those polls leading up to the election that predicted a Republican winner.