Problem: Each year, reporters and pundits weigh in on which movies they believe will secure academy award nominations for Best Picture. Because of the wealth of data available about movies, we think we can predict whether a movie will be nominated for an academy award more accurately.

Data:

We will collect the following data:

- Meta data about movies, including budget, release dates, genre, director, actors, Oscar nomination status (1=yes, 0=no), etc.
- Reviews of movies from major critics for sentiment analysis

We have researched the following possible sources for data:

- The Movie DB
- List of Oscar Nominees
- IMDB 5000 dataset
- Movie Review dataset
- Cornell Movie Review dataset
- New York Times Movie Review API
- Rotten Tomatoes API

Because so many films are released each year, we will apply a filter criteria to only include movies that are eligible for Oscar contention (e.g., only movies that were reviewed by the New York Times). In addition, we plan to limit our analysis to the last 20 years to ensure that our dataset is complete, and to mitigate the effect of time trends in nominations.

Analytic Techniques:

We plan to conduct this analysis in two stages:

1. For each film, generate a sentiment rating for the film reviews of prominent critics using Natural Language processing. Using similar methods, translate categorical variables like lead actors, directors, production company, etc, into "prestige" ratings. For example, based on a text analysis, lead actor Leonardo DiCaprio would likely get a "5-rating" (on a scale of 1-5) for a "Prestige_of_Lead_Actor" variable.
2. Create an aggregated dataset of structured data (e.g., box office receipts, genre) and our newly created sentiment and prestige ratings. We will split this dataset into a training (first 15-18 years) and testing set (last 2-5 years). We plan to compare several methods for predicting the likelihood that a film will be nominated for a Best Picture academy award:
   - Logistic regression
   - Classification and Regression Trees
   - Random Forests

Impact: The overall goal is to accurately predict Oscar nominees. Additionally, we would like to create a sentiment analysis based on movie review data that can determine whether the review from a reputable source is extra-ordinarily favorable, and if this increases the likelihood that a movie is nominated for an Oscar. Overall, we will be able to determine which variables are significant, and show visually what movie characteristics make it more likely to be nominated for an Oscar.