Current Status:

The data for this project was provided by the National Oceanographic and Atmospheric Administration (NOAA) National Geophysical Data Center (NGDC). The Solar-Terrestrial Physics (STP) division at NGDC has made Mount Wilson Observatory sunspot group data and Geostationary Operational Environmental Satellite (GOES) solar flare event data available on their ftp site. A python script (goes_sunspot_parser.py) was used to gather data for both measurements from 1982 to 2014. Only class M and X flares were written to the data frame, as these types of flares pose the most severe threat. Two separate data frames containing sunspot group data and x-ray event data were created using this script, and then written to csv files so that the data could be cleaned up without an internet connection.

The raw x-ray event and sunspot group data were cleaned using a Python script entitled sun_x_ray.py. The bulk of preparing the x-ray event data revolved around slicing and manipulating the measurement date and start time columns so that a final column of datetime data could be created using the Datetime package. A similar process was done to the sunspot group data.

These two data frames, labeled clean_x_ray and clean_sunspot, are then used to determine if an observed sunspot group is associated with a flare event. The data frames have two feature columns in common, NOAA sunspot group number and the time of observation in UTC. The script matches sunspot groups and flare events using the NOAA sunspot group number. If a flare and a sunspot group share a common NOAA number, the time difference between the flare event and each sunspot measurement in the group is calculated. Sunspots that are observed within six hours of a flare event are considered to be associated with the flare. If multiple sunspot observations were made within six hours of a flare measurement, only the observation with the minimum time difference is considered to be associated. The associated sunspot group observations are written to a new data frame and then merged with the appropriate solar flare observation. The final product is a merged data frame of solar flare events and their associated sunspot observations.

Future Work:

A data frame of unassociated sunspot groups needs to be created. The response and feature columns of the associated and unassociated data frames will need to be selected in order to concatenate the associated and unassociated data frames into the final data set. For this work there will be three features (zurich, penumbra, compactness) and the response will be associated/unassociated. Several groups have applied computationally expensive techniques to similar datasets (cascade-correlation neural nets), but I'd like to explore less exotic machine learning methods such as logistic regression.