**Data 606 Capstone: Predicting Accident Severity from Crash Data**

Every year car accidents cause thousands of deaths, millions of injuries, and billions of dollars of expense in this country and around the world. America’s roadways are a huge source of human and financial loss, one that we have become unfortunately accustomed to and regard as an externality of our lifestyles and the current transportation methods we undertake. The fact is, car accidents cause far too much damage to ignore, and efforts should and are being made to engineer our cars and roads to be safer. For my part, I’m seeking to work with an extensive collection of accident data to be able to predict accident severity based on available information and uncover the biggest risk factors that lead to accidents.

The specific question I am asking has two parts. First, I want to predict, based on geospatial, time, and weather data, the severity of an accident as soon as it is reported. The data I am using rates severity on a 1-4 scale, determined by how much of a delay the accident causes. It is apparent shortly after an accident takes place how severe it is based on the ensuing backup, but it is not often immediately apparent. I am seeking to be able to predict, based on information that would be known from the very first accident report, how severe an accident is so that an accurate estimation of delays can be immediately reported to drivers on the road. There will be an emphasis on being able to accurately predict the most severe accidents, as these cause the most severe delays. In application, this model could be possibly embedded within mapping software. Secondarily, I want to determine which factors are the biggest determinants of these severe accidents. The importance of this is self-evident. Cars could be engineered more safely, road conditions could be improved, more streetlights could be built, etc. There exist many potential possibilities to increase safety based on whatever is discovered.

For my methodology, I want to first conduct a thorough exploratory data analysis after the data is cleaned and ready for use. There is a great deal of very informative visualization potential given the geographic nature of this data, as well as the potential to conduct some spatial statistical tests on certain locales of interest. I intend to use factor analysis methods to uncover the biggest determinants of severe crashes and fit the best multiclassification model I can use to predict severity.