**2. Data acquisition and cleaning**

**2.1 Data sources**

Data on weather conditions and the severity of co-occurring traffic incidents can be found in a Kaggle dataset located ​[here](http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv)​. The original dataset originally had 221,226 features. Because the data set has so many features, we could safely drop the features with missing weather data; the resulting sample still had 194,310 attributes, a very robust n sample.

**2.2 Data cleaning**

There are numerous problems with these data. First, some of the attributes had missing data. Because the dataset was so large, I decided to drop the missing data from the set. Even after dropping the missing data, there were still over 190,000 features, which left me with a more than adequate dataset. Second, various attributes were over-weighted to certain outcomes. For example, ‘Light conditions’ was over-weighted to ‘daylight’, ‘Road Conditions’ was overweighted to ‘dry’, and ‘Weather’ was ‘overweighted’ to ‘clear’. To fix the issue, I decreased the weighting of those particular outcomes and normalized the dataset. I did not discover a significant number of outliers.

**2.3 Feature selection**

The ‘severity code’ attribute, which incorporates numerous other attributes within the dataset to holistically estimate accident severity, was determined to be the target variable. The numerous features that feed into the ‘severity code’ attribute were not kept, as they are already reflected in the resultant ‘severity code’.

The data contained numerous administrative attributes, which would solely be used to help police keep administrative track of the data. Because these attributes are arbitrary administrative descriptors, these attributes were also dropped.

Location data was dropped, as it does not directly contribute to our problem statement.

The remaining attributes are weather conditions, road conditions, and light conditions. These weather attributes will be used to predict the target variable of ‘accident severity’.

After data cleaning and features selection, there are 194,310 features and four attributes.