Data

The data we have on hand is quite extensive, providing us with relevant attributes such as the location and time of the accident, as well as the consequent weather and lighting conditions when the event occurred. There are also details indicating how many pedestrians were involved, or if one or more cars that were affected were parked or not.

Breakdown

For this particular case, the Dependent Variable to predict will be the severity of the accident, or **SEVERITYCODE**. Note that the Severity Codes in the provided CSV file are only limited to two values: 1 and 2. According to the supplementary Metadata file, **SEVERITYCODE 1** corresponds to **Property Damage** and **2** corresponds to **Injury**. Thus, the model we will create will be geared towards ***predicting the likelihood of a potential accident to lead to Property Damage or Injury.***

NOTE: There were other Severity Codes disclosed on the Metadata: 0 (Unknown), 2b (Serious Injury), and 3 (Fatality). We believe these were omitted because the dataset already had thousands of rows to work on without these Severity Codes. 0 (Unknown) was most likely omitted because of its potential to confuse the model. Finally, we believe 2 (Injury) would be of utmost importance because we don't want to wait for serious injuries or fatalities to happen before we work on data. In other words if we could prevent injury we could definitely prevent worse things from happening. We have chosen to work with the following Attributes:

- **LOCATION** - Description of the general location of the collision.
- **PERSONCOUNT** - The total number of people involved in the collision.
- **VEHCOUNT** - The number of vehicles involved in the collision.
- **JUNCTIONTYPE** - Category of junction at which collision took place
- **WEATHER** - A description of the weather conditions during the time of the collision.
- **ROADCOND** - The condition of the road during the collision.
- **LIGHTCOND** - The light conditions during the collision.
- **SPEEDING** - Whether or not speeding was a factor in the collision.
  These descriptions were also fathered from the supplementary Metadata file.

We are preserving the following Columns, but may drop them as we continue with the project:

- **X/Y** - Coordinates serving as supplementary to LOCATION
- **INCDATE, INCDTTM** - The date and time of the incident; Could it have happened on, say, a holiday? Or is this relevant?
  Finally, these are the columns we dropped, due to their lesser relevance or because they are details of the accident after the fact: **OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, ADDRTYPE, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, Duplicate SEVERITYCODE, COLLISIONTYPE, PEDCOUNT, PEDCYLCOUNT, SDOT_COLCODE, SDOT_COLDESC, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR, SEVERITYDESC**

We have the original, untouched CSV saved as 'Data-Collisions - RAW.csv' in the repository. Consequently, a CSV with columns dropped is also on the repository, filename 'Data-Collisions - Clean Attributes.csv'.