**The Data**

* Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “[A Countrywide Traffic Accident Dataset](https://arxiv.org/abs/1906.05409).”, 2019.
* Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. ["Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights."](https://arxiv.org/abs/1909.09638) In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

The data is for the contiguous 49 US States and is continuously being collected starting from February 2016 to June 2020, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.

The dataset consists of 3.5 million accident data points organized into 49 columns. It has features like reporting sources of the accident, state, county, source, severity, sunrise/sunset times, weather conditions, etc.

I will be specifically looking at NY/NJ after cleaning the data since I drive mostly in these two states. The goal is to accurately identify certain factors that contribute to accident severity and to what severity degree.