Introduction

"The objective of your mandate is to provide the city with a ranking of the 1864 intersections in terms of safety (from the most dangerous to the least dangerous), so that it can prioritize the riskiest intersections with the aim of improving infrastructure."

To do so, we need to asses the **dangerousness** of all intersection and rank them. Our modeling needs to be based on the variable accident (`acc`). Thus, let's define the observed **accidents** as a manifestation of the dangerousness in the following way:

$$Accidents = Dangerousness + RandomError \tag{1}$$

Or in terms of variables, let f denote the function of dangerousness that we want to estimate:

$$acc = f + \epsilon \tag{2}$$

The function of dangerousness can be further broken down in two variables which are **risk** and **exposure**.

The **risk** can be defined as the accident probability for a given individual that crosses the intersection or as the accident rate per crossing.

The **exposure** can be defined as the number of people that cross the intersection (i.e. the number of person that are exposed to the risk of crossing).

Thus we get the following decomposition of dangerousness:

$$f = risk * exposure \tag{3}$$

This implies that an intersection characterized by a high probability of accidents but infrequently traversed (low exposure) would result in a low level of dangerousness and a minimal number of accidents. Conversely, a situation with a low risk but high exposure would also lead to relatively low dangerousness and a reduced occurrence of accidents.

As the city of Montreal aims to prioritize infrastructure enhancements, it is logical to adopt the concept of dangerousness, as defined earlier, as a metric for ranking intersections. This stands in contrast to focusing solely on riskiness, represented by accident rates or probabilities of accidents per crossing. This approach aligns with a **utilitarian** perspective that prioritizes reducing the overall number of accidents.

Hence, our objective is to precisely estimate the dangerousness function *f*. As mentioned previously, dangerousness is not directly observed; therefore, our modeling is grounded in accidents. To achieve this, our focus is on minimizing the expected prediction error associated with accidents.

$$EPE(X) = Var(acc) + Bias^2 + Var[\hat{f}(X)] \tag{4}$$

As the variance of accidents is irreducible, representing the random error in Equation 1, our aim is to minimize both the bias and variance in our accident model. This approach is crucial for obtaining the most accurate estimate of the dangerousness function *f*.

Processing

#summary(dat) There are missing values for both variables "ln_distdt" "X"...

#dat<- f_clean_data(dat)