# CAPSTONE PROJECT PROPOSAL

**Mentor**: Raj Bandyopadhyay
Christian Farfan Centeno

## Toronto Cycling Collisions

According to the Census, 1.7% of the people in Toronto rode their bikes to work in 2006. This cyclist share is under reported however, as it does not include those who bike for fitness, to run errands, or for leisure. It also may exclude those who use a bike to reach public transportation for their commute, as the Census only allowed one choice for travel method to work. Thus the number of active cyclists in the city is higher than 1.7% of the population. From 2001-2006 there was over 30% increase in the number of cyclists commuting to work (http://goo.gl/sMLsfs) and in 2011 the proportion was reported higher again, despite mode of commuting being removed from the mandatory part of the Census, and becoming voluntary information.

The city of Toronto has developed a Ten Year Plan to develop the area's cycling network and become a more bicycle friendly city. Meant to outline the future investments in infrastructure between 2016-2025, the city has published maps of the proposed expanded cycling network. These maps were produced mostly by data collection on traffic, estimating potential demand by non-cycling trips and non-walking trips less than 5km (http://www.torontocyclingnetwork.info/studying-toronto/). Identifying where to improve or add bike lanes via traffic is only one part of the equation however. A chief issue is cycling collisions and WHY they happen. Identifying which areas or intersections require intervention or perhaps should be avoided by cyclists altogether can be be accomplished by analyzing cycling collision data and applying machine learning methods to make predictions on accident rates.

## Client:

Toronto city council itself. They will be able to reevaluate their cycling network plans, whether by changing the proposed routes themselves, altering the type of routes, or changing the time to construction for various routes. It is quite possible that a simple “downtown raw numbers” versus “outer areas” raw numbers and cyclist anecdotal evidence was used, as suggested by the maps on the Toronto bike plan website. There may be information missed which can be revealed by thinking more statistically about the collisions, and attempting to learn how intersection design, traffic numbers and other variables influence accident severity.

Various bike shares, cycling shops, or simply hobbyist cyclist organizations in the city may be interested as well. Can change their own cycling patterns based on this data (individuals may have their own anecdotal beliefs already, and this can help strengthen such beliefs, or turn them on to overlooked ones). Many of them are probably involved with lobbying or working with those responsible for the ten year bike plan, and as such, this analysis could provide them with tangible statistics based evidence to modify the current plan to better suit their needs. They could also also provide cyclists with the a better idea of what the risks inherent in certain routes and route types are.

## Data:

Bicycle collision data from Toronto Traffic Safety Unit (1986-2010):
https://github.com/farfan92/SpringBoard-/blob/master/cycling_collisions_toronto_1986-2010.xls.csv

Bicycle and vehicular traffic counts at Toronto Open Data:
http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=1a66e03bb8d1e310VgnVCM10000071d60f89RCRD

Collisions from the summer months of 2016, curated by the StruckTOBot twitter account:
https://twitter.com/StruckTOBot?lang=en

# Approach:

I propose to use the Toronto Traffic Safety Unit's cycling collision data between 1986-2010 to identify dangerous intersections for cyclists. The data contains the date/time, severity of injury, as well as GPS and street intersection locations of reported cycling collisions, along with other relevant data. Simply organizing via number of accidents is not viable however, as the intersections with the highest number of accidents will be those with the most traffic. I will attempt to normalize by estimated cycling traffic, via the city's Open data catalog, which contains bicycle counts for 51 different common cycling intersections. From here there are a few approaches:

1. Blanket normalize all the collision data by assuming an intersection will have the same traffic as the closest intersection(s) included in the bicycle count data set.
2. Normalize by some distance-weighted average from counted intersections.
3. Normalize based on the city's vehicular traffic data from 2011, which contains over 2000 intersections.

It is likely some combination of all three options will be necessary. A key component of this project will be in justifying and implementing an appropriate normalization procedure in the downtown core area versus the outer city neighborhoods. A quick study on the correlation between car and bike traffic in different areas of the city will probably be required.

After normalization we will have the intersections with the smallest amount of collisions dominate our data due to statistical noise. Thus I will need to put some sort of prior on our data, so that the prior dominates at these noisy intersections (http://goo.gl/obrgxF). An empirical Bayesian method can be used to estimate the prior distribution based on the data.

Once completed, we would have a list of the most “dangerous” intersections based on number of collisions and number of traffic. This can be extended further however, by grouping data based on severity of injury (fatal, major, minor, minimal, none), and producing plots/maps in this manner. A weighting system can also be assigned, where a fatal or major accident is worth more than a minimal one, and a map of the most dangerous intersections would then carry some idea of the relative risk conditional on the accident type. By having a map/list of dangerous intersections under various criteria, one can compare this to the proposed cycling network expansion map. This will allow me to see if not considering the potential risk in various intersections has caused certain areas to be overlooked in the City's ten year plan, or if certain routes should be given higher priority in terms of implementation date.

By taking into account the type of intersection, street types, time of day/year, and resulting severity of injury one can attempt to train a learning model on the data. The goal would be to predict the collision rates for the 2016 year. By scrapping the collision reporting tweets from Twitter, and matching up the relevant features, I could then compare the predicted collision rates to those observed since November 2015 (when the StruckTObot account was launched). Particularily intresting would be comparing the accident rates during the summer months where a stretch of Bloor Street (a major downtown road) has been trialling a seperated bike corridor, as well as accident rates on Richmond Street and Adelaide street, which have had bike corridors installed as a pilot since 2014. A well-

designed and trained model will hopefully be able to predict the collision rates under these new conditions. Thus I'd have some confirmation that recommendations for cycling network expansion could be accompanied with believable predictions.

Output will obviously be tabular list of dangerous intersections, with columns for total collisions, collisions by accident severity, and various possible weighted “risk” variables.

A  vidualization of this deliverable would be a stylized “map” of Toronto (or various sub areas), with data points sized by normalized collision rate and possible colored via a gradient that takes into account accident severity. Over laying this over a simple map of the TTC system and existing/proposed bike network could provide some visually intuitive means of conveying information to client.

The main deliverable and goal is to have a weighted collision rate prediction for the Bloor Street, Adelaide and Richmond street sections with and without bike corridors. The prediction WITH bike corridors should hopefully match the new data from the past year, while that without bike corridors will provide an estimate for what the collision rates would have been otherwise.

# Thoughts and concerns:

Obviously, the normalization procedure for intersections which are far from a counted bicycle intersection will heavily distort the data based on my assumptions. I'll have to be careful with this.

May be difficult to keep map visualization of collisions and relative risk from looking too cluttered or too “simple”, but feel like it would make excellent practice for striking the right balance when presenting to Clients.

Traffic count data is ongoing but also relatively new. It only has commenced since 2010, and each intersection was counted at wildly varying dates. Thus there is an implicit variability in this traffic data. It is also not clear how bike and vehicular traffic may have changed from 1986 to 2010 and from then to now. Thus we are bound to compare newer traffic numbers with older collision data. The open data sets don't go far back, and thus I may need to use my best judgement in applying scaling traffic counts for older data. May need to draw inferences this way. Could possibly contact Traffic department and request if this data (if it exists) be provided.

Time/date, age of cyclist, and 'driver action' (ran red light, was impaired etc.) are contained in the collision data. There are many extensions based on this that can be made, i.e. are their locations where cyclists are more likely to be at fault than others?However, the accident data I have since Nov 2015 from the Twitterbot does not generally have many details and as such my predictive model would be trained on a more feature rich dataset than I can compare to. Perhaps will want to split the large data set into training and prediction sets for this purpose.

Can attempt to contact whatever committee/council is responsible for the ten year bike plan, share findings, ask what sort of analysis or alterations they may be interested in. A future extension may involve comparing collision data on Bloor Street at the end of the summer bike lane trial period to collision data over summers from the previous five or so summers, if someone can expediate the Toronto OpenData request (currently takes months). A positive effect on cyclist safety could be valuable ammunition in city council debate against those opposed to the bike lanes based on the possibility of increased congestion or an increase in accidents via more cyclists behaving improperly.