# West Nile Virus Project Proposal

## 7/6/2021

## Blaine Murphy

### Problem statement

In less than one month I will use the West Nile virus data collected by the city of Chicago in conjunction with weather and insect spraying data to create a ML model capable of accurately predicting (AUC >0.8) the presence of mosquitos carrying the virus at various locations across the city.

### Proposal

The city of Chicago had its first recorded case of West Nile virus in 2002 and in response the Chicago Department of Public Health (CDPH) set up a weekly program to collect mosquito samples and test for the virus across the city. These data are then used to decide where in the city to spray for mosquitos. I hope to create a predictive model using this data along with weather and spraying data to create a predictive model foretelling where and when mosquitos carrying the virus will appear.

The West Nile virus (WNV) is a potentially deadly virus that is primarily transmitted to humans from the bites of infected mosquitos. According to the CDC, the WNV is the leading cause of mosquito born disease in the United States and cases primarily occur in summer through fall. Additionally, the CDC says “1 in 5 people who are infected develop a fever and other symptoms,” and “1 out of 150 infected people develop a serious, sometimes fatal, illness.” Given the risk to citizens of Chicago, the CDPH seek to mitigate that risk by employing an extremely accurate predictive model. This model will advise where to spray for mosquitos in the future.

Model accuracy is the measure of success for this project. The data set is heavily imbalanced, 95% of the samples test negative for WNV. Given this imbalance I will use AUC as my accuracy score and the goal is greater than 0.80.

The stakeholders for this project are the City of Chicago and the CDPH. My analysis, modelling and recommendations will be presented to them for approval. Other stakeholders are virologists and entomologists that will have input into the modelling process.

The data available for this project are ~10,000 records of mosquito sampling and testing in the city of Chicago. These data are from 136 different mosquito traps located across the city from 2007 to 2013 and only consist of the summer months when the mosquitos are more prevalent. The locations of traps are recorded by their address and their latitude and longitude, and the species of mosquitos is also recorded. Also available for this effort is data detailing the mosquito spraying efforts of the City of Chicago. These data include the date, time, latitude, and longitude of a number of mosquito spraying routes through the city. Finally, weather data from the NOAA is available for 2 weather stations in Chicago from May 2007 through October 2013. These data consist of daily measurements of temperature, precipitation, and wind as well as other measurements. I will use all of these data in my analysis and predictive modeling efforts.

Feature engineering will be key to the accuracy of the model. A few preliminary ideas for calculated features are:

* Rainfall total over the last week
* Average wind speed over the last week
* Average temperature of the last week
* Proximity to mosquito spray and time since last spray
* Positive test last week
* Trap Proximity to positive test
* Mosquito species prevalence in the local area

This is an incomplete list that I will be adding to once I begin my exploratory data analysis.

I foresee some constraints that might be encountered in this project. This is a project that would greatly benefit from subject matter expertise that may not be available or I may not be able to gather during the course of this project. In addition, there are only 2 weather stations available for analyzing the weather. This may not be sufficient if there are localized precipitation events that are missed by these two weather stations. Finally, the mosquito traps may be inconsistent across the city and through time for any number of reasons, resulting in a sampling error that cannot be removed.

I will complete this project in less than one month. Upon completion of this project, I will deliver a pipeline to the CDPH that will clean input data for any missing values, calculate new features, and model the likelihood of finding West Nile virus carrying mosquitos at a given trap location.

## Sources

Centers for Disease Control and Prevention (USA), <https://www.cdc.gov/westnile/index.html>