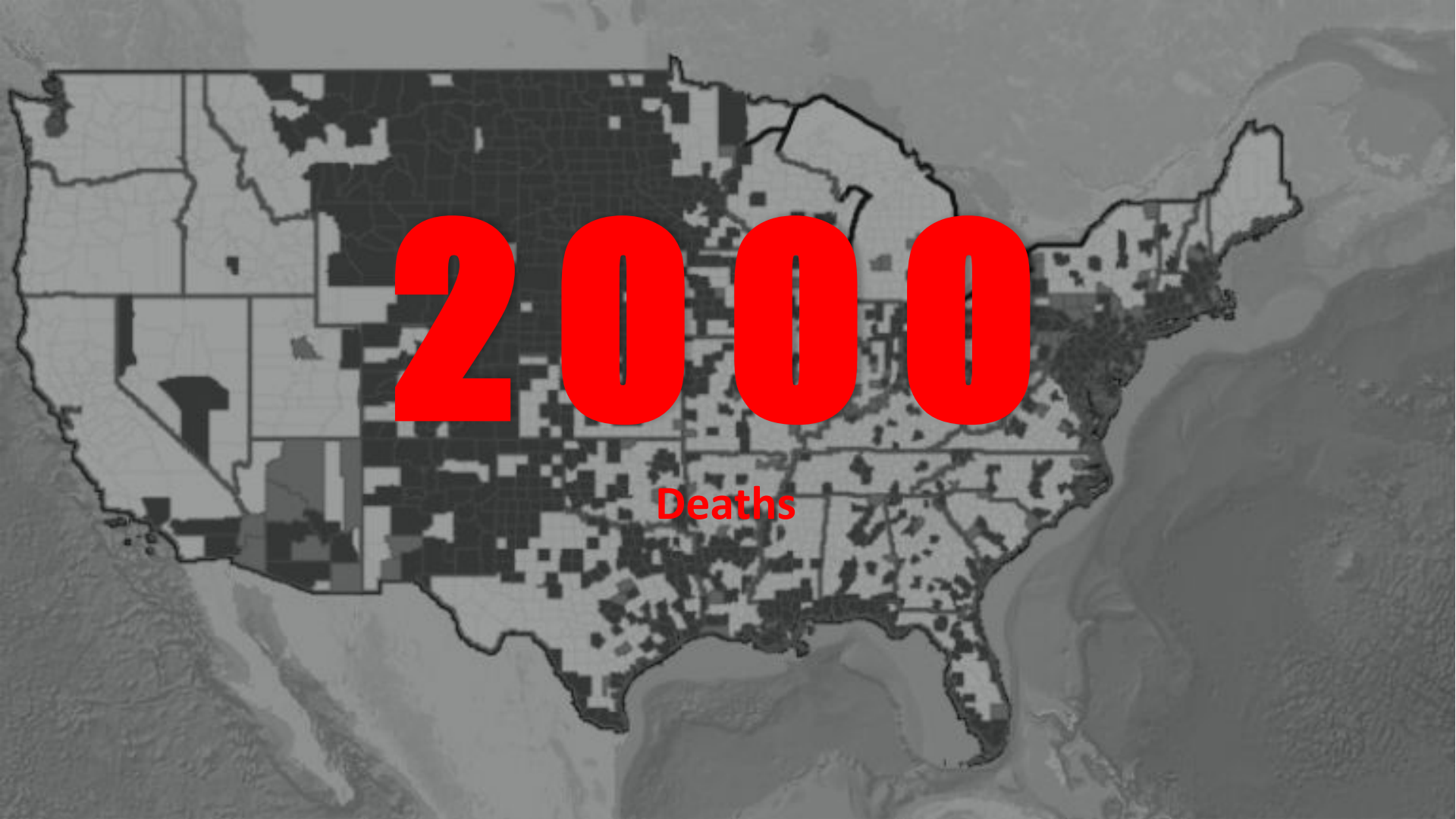
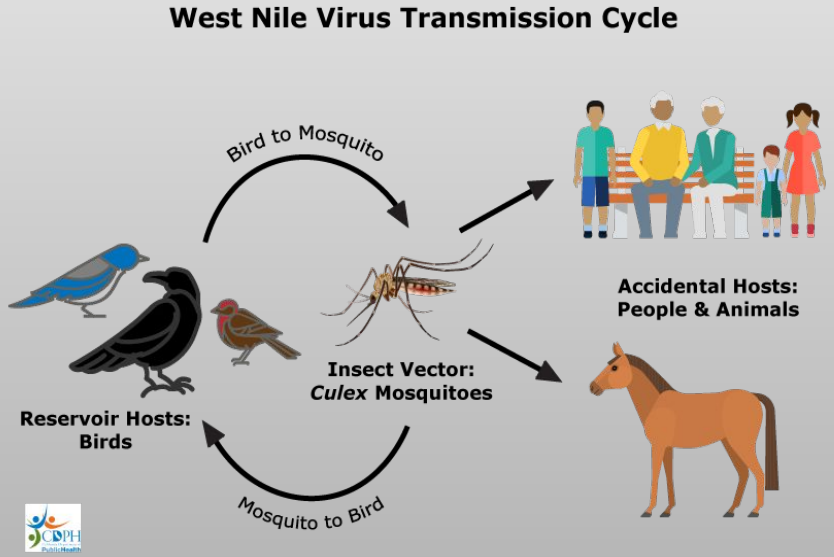
Predicting Chicago West Nile Virus

Team : Amy L, Jennifer S, Matthew L, Manodhar A, Sonal M, Vishnu M

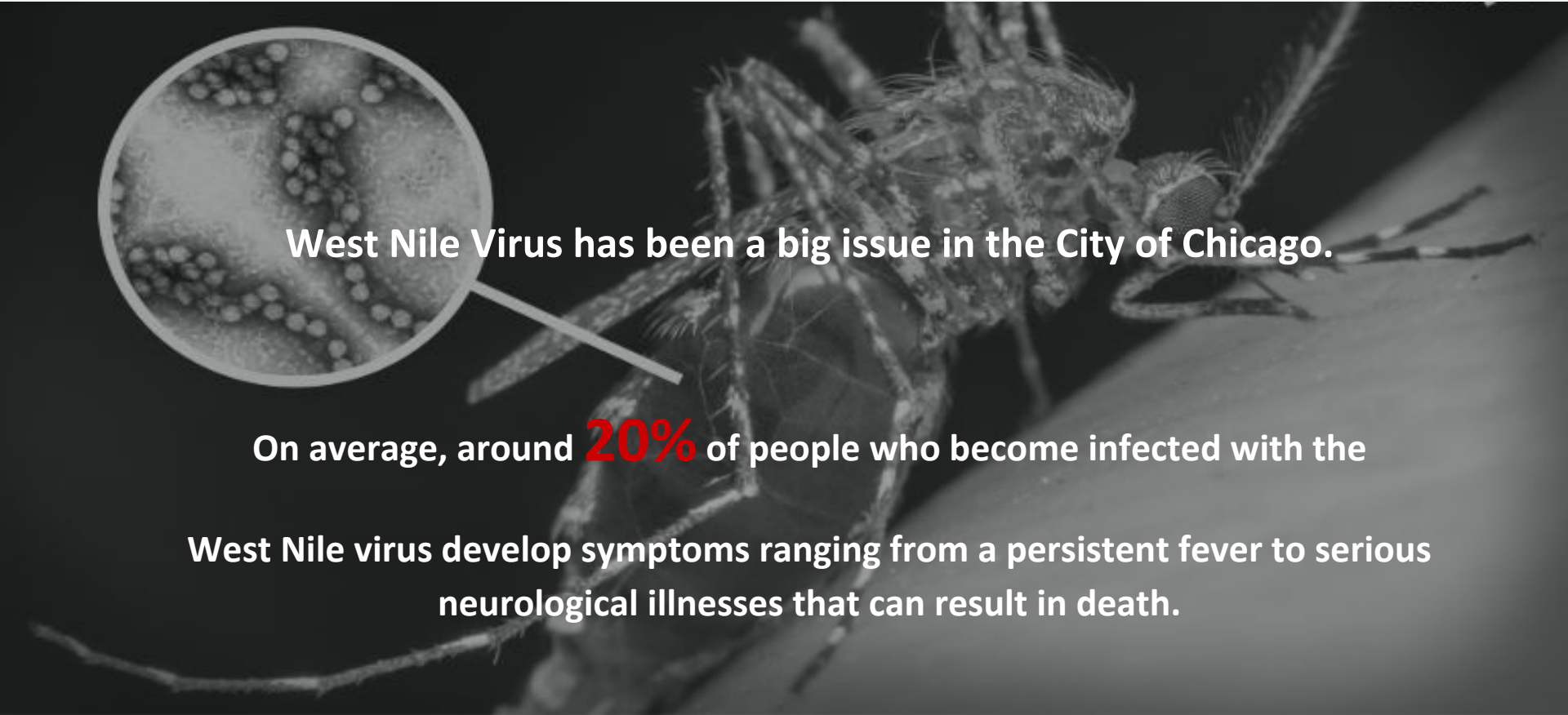

2000
Deaths

What is West Nile Virus (WNV)?

- WNV is a serious disease most commonly spread by infected mosquitoes.
- Infected mosquitoes can then spread the virus to humans and other animals. First discovered in the U.S. in 1999, New York.
- It is believed that hot and dry conditions are more favorable for West Nile virus than cold and wet.



Problem Recognition



West Nile Virus has been a big issue in the City of Chicago.

On average, around 20% of people who become infected with the West Nile virus develop symptoms ranging from a persistent fever to serious neurological illnesses that can result in death.

Data Preparation

Trap data

Features created
- Total mosquitos in a trap, on a given day and location

Data capture
- Trap data is recorded for 2007, 2009, 2011,2013
- Trap data tells us about the trap location, species that are tested, #mosquitos, WNV presence
- Traps samples are tested for WNV presence every week

Weather data

Features created
- Relative humidity
- Length of day
- Heat index (benchmark 65F)
- Moving averages

Data capture
- weather data is captured by 2 stations separated by 17 miles
- Weather data is recorded from 2007 to 2014

Spray data

Features created
- Spray time
- Spray distance

Data capture
- Sprays are done in 2011 and 2013
- sprays are done 2 times in 2011 and 8 times in 2013
- Each spray is separated by a week

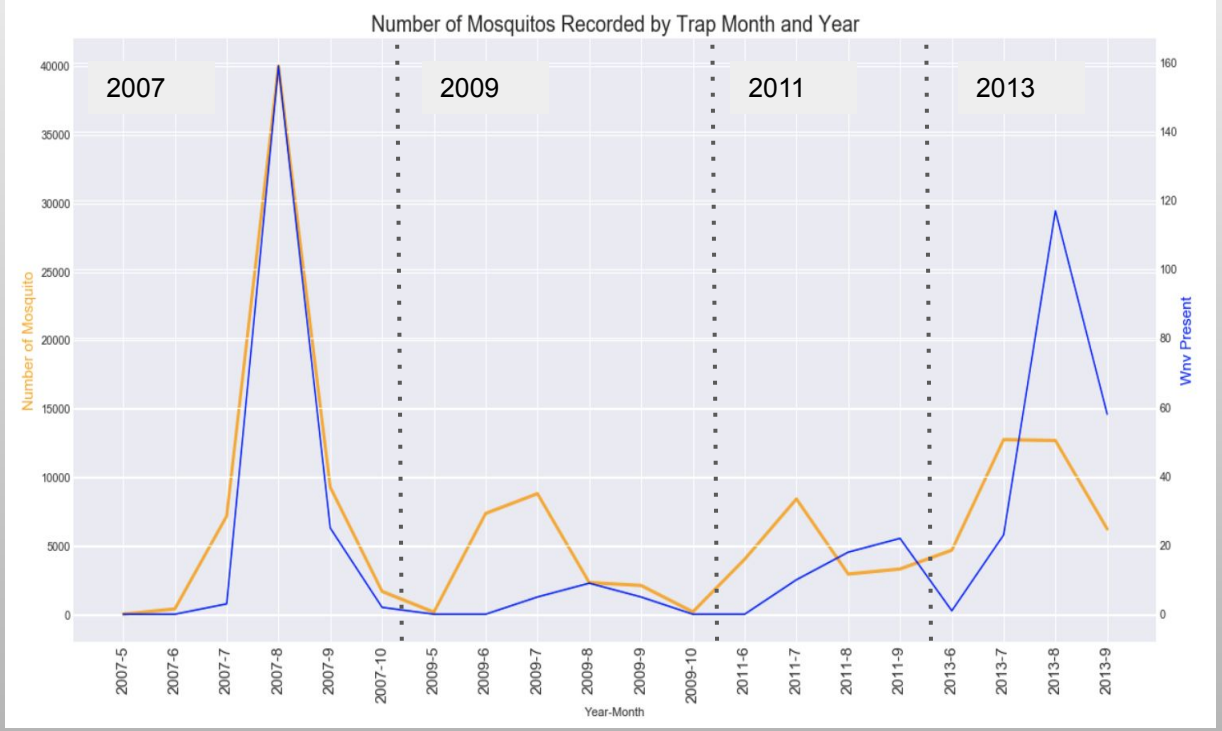
Master Database

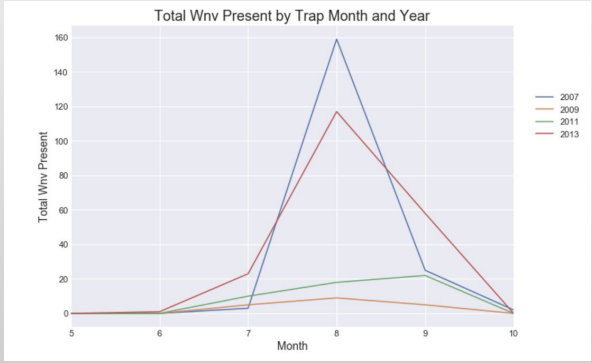
Features created
- Nearest time of spray from trap
- Closest distance from spray to trap
- Distance from trap to weather station

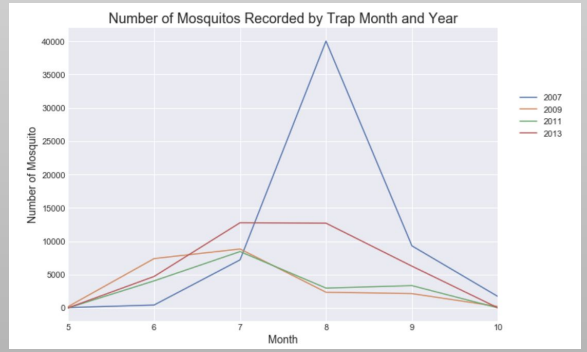
Data capture
- Merged Train data and weather on the dates
- Merged Trap data and spray data on dates and the nearest sprays

Mosquito Distribution by Month

- Throughout 4 years, August showed to be the month where mosquito appeared the most
- After the spray, the Number of Mosquito decreased but spiked up again in 2013

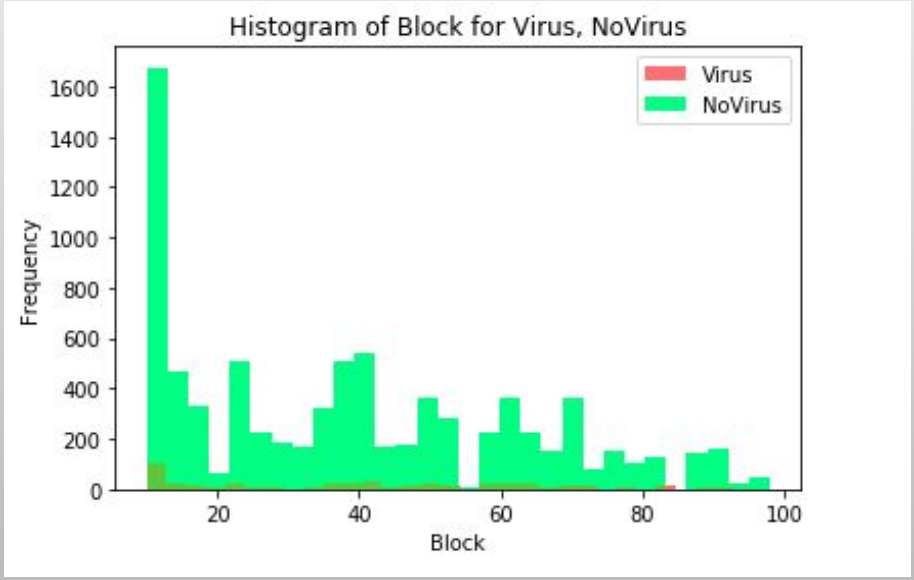


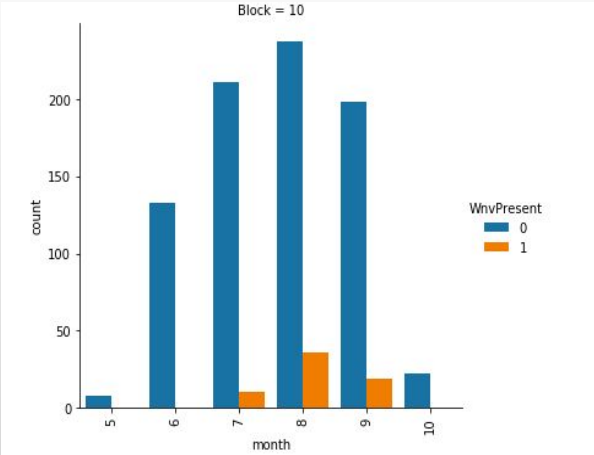


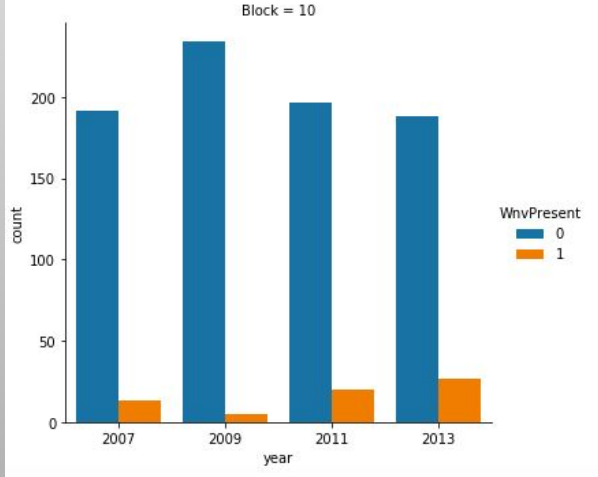


Block Specifics Virus Presence

- The histogram shows the Virus presence was most prominent in block 10
- By month and year, it showed the same trend as overall

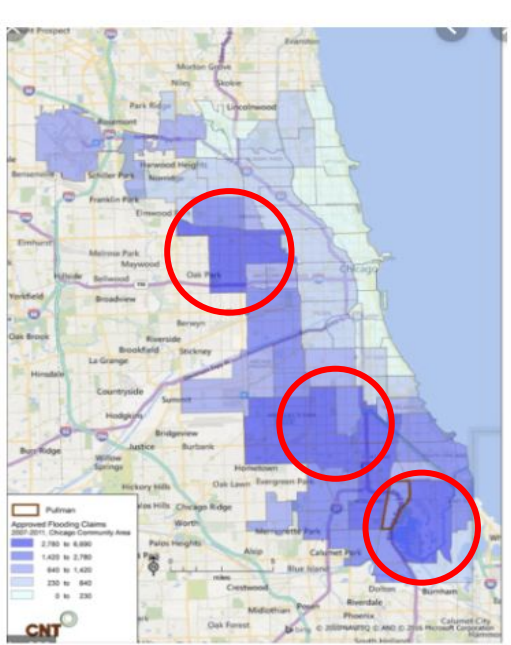


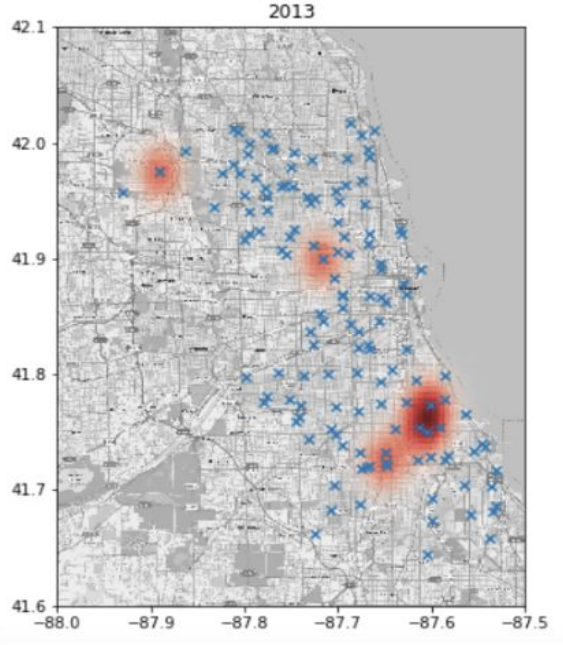




The Increase in Number of Mosquitos in 2013

- There was a clear association between weather **conditions (temperature and precipitation) and mosquito abundance,** which allowed the definition of threshold criteria for temperature and precipitation conditions for mosquito population growth

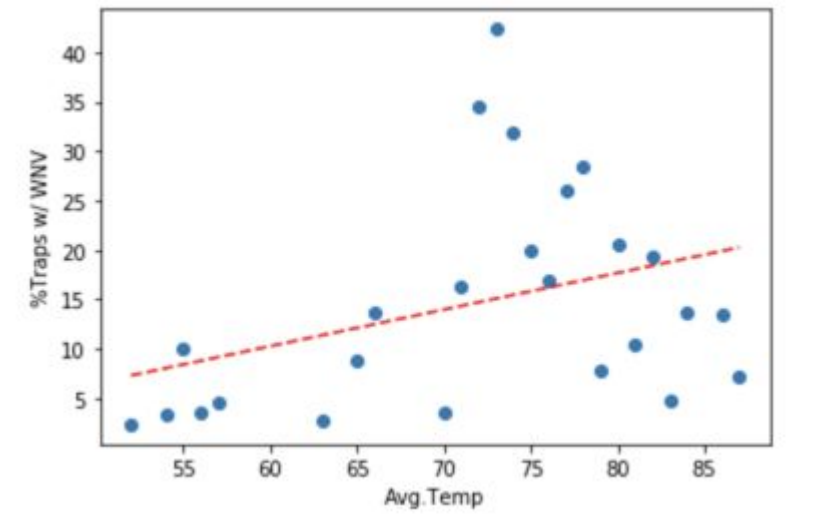




https://academic.oup.com/jme/article-abstract/48/2/468/894988

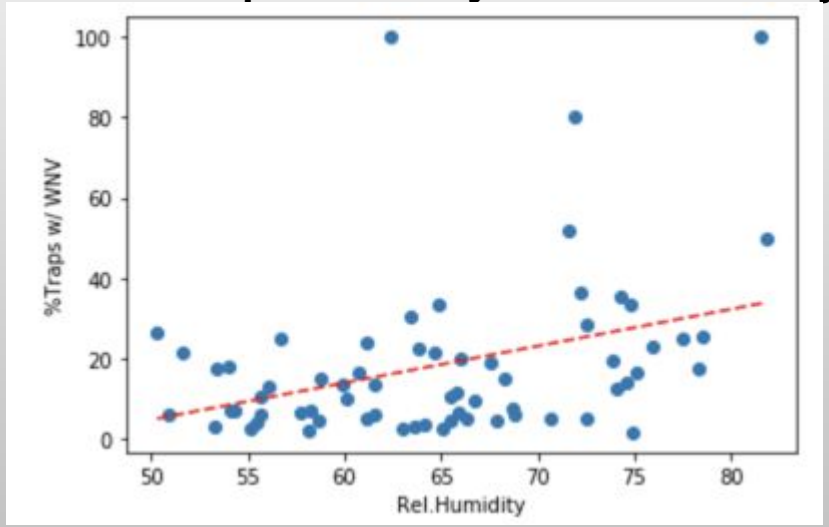
Impact of temperature and relative humidity on Virus

WNV presence by temperature



- The percentage of traps with WNV increases with the increase in temperature
- 30-40% of the traps had WNV presence in the temperature range of 72F-82F

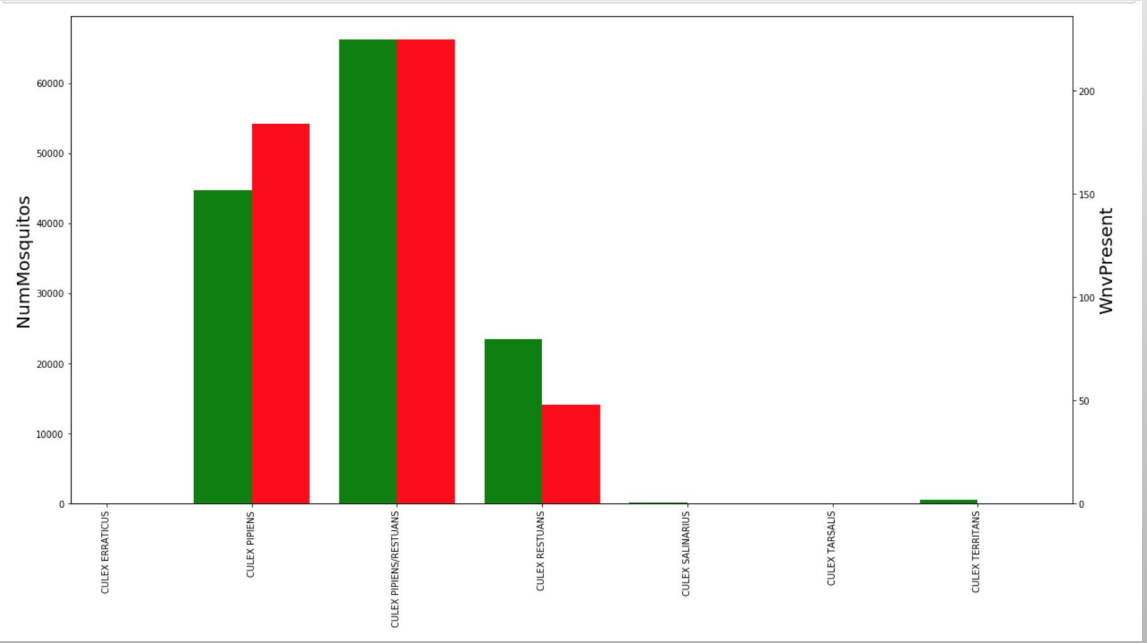
WNV presence by relative humidity



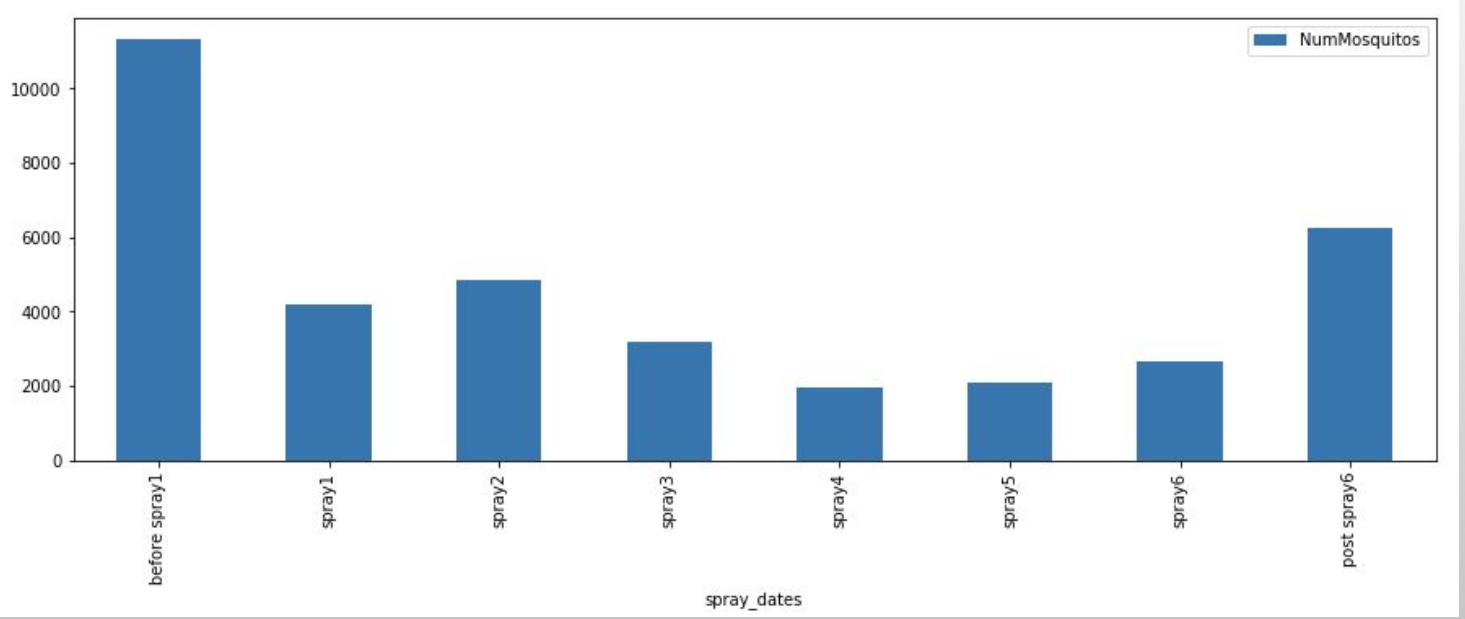
- The percentage of traps with WNV increases with the increase in relative humidity

Virus Presence by Mosquito Species

- Among those 3 species that spread the virus, we confirmed CULEX PIPIENS/RESTAUANS is the main carrier associated with the presence of the virus



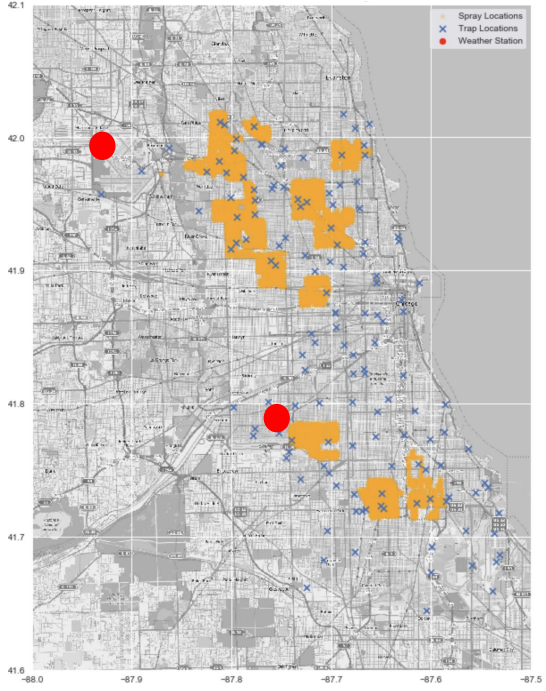
Effect of sprays on number of mosquitos



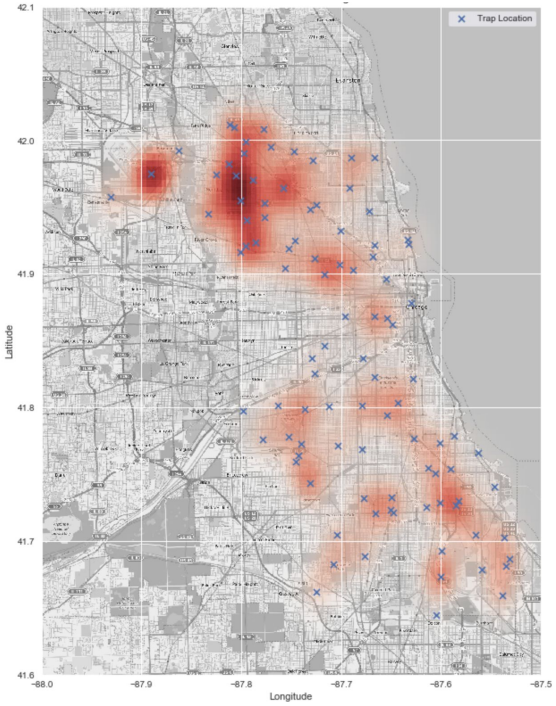
- Sprays are deployed every week starting from 17th July till 5th Sep 2013
- In 2013 the number of mosquitoes decreased only during the spray time when compared with number of mosquitos before and after spray.

Virus Presence vs Spray, Trap & Weather Stations

Spray, Trap and Weather Stations



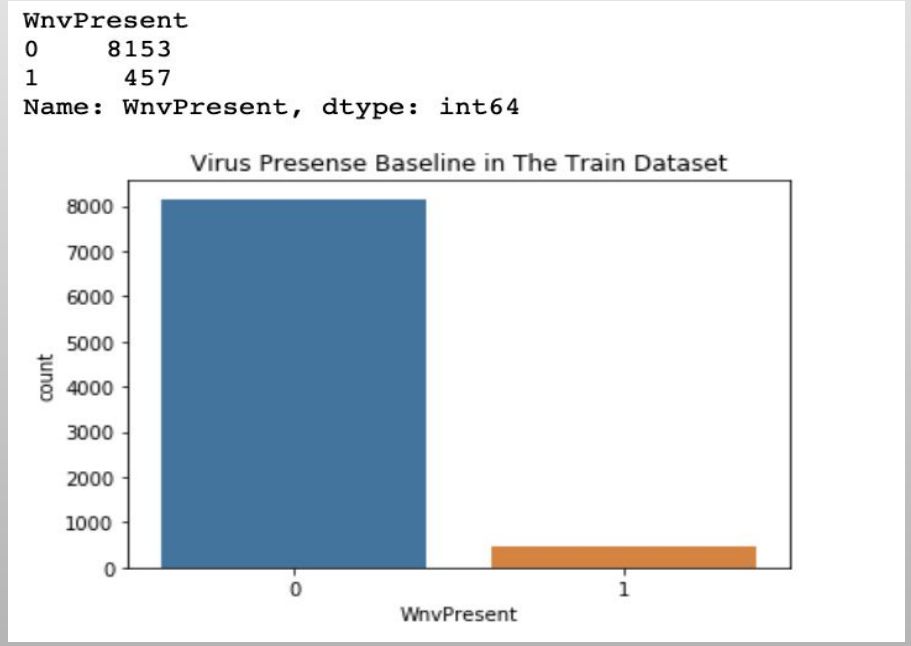
Presence of West Nile Virus



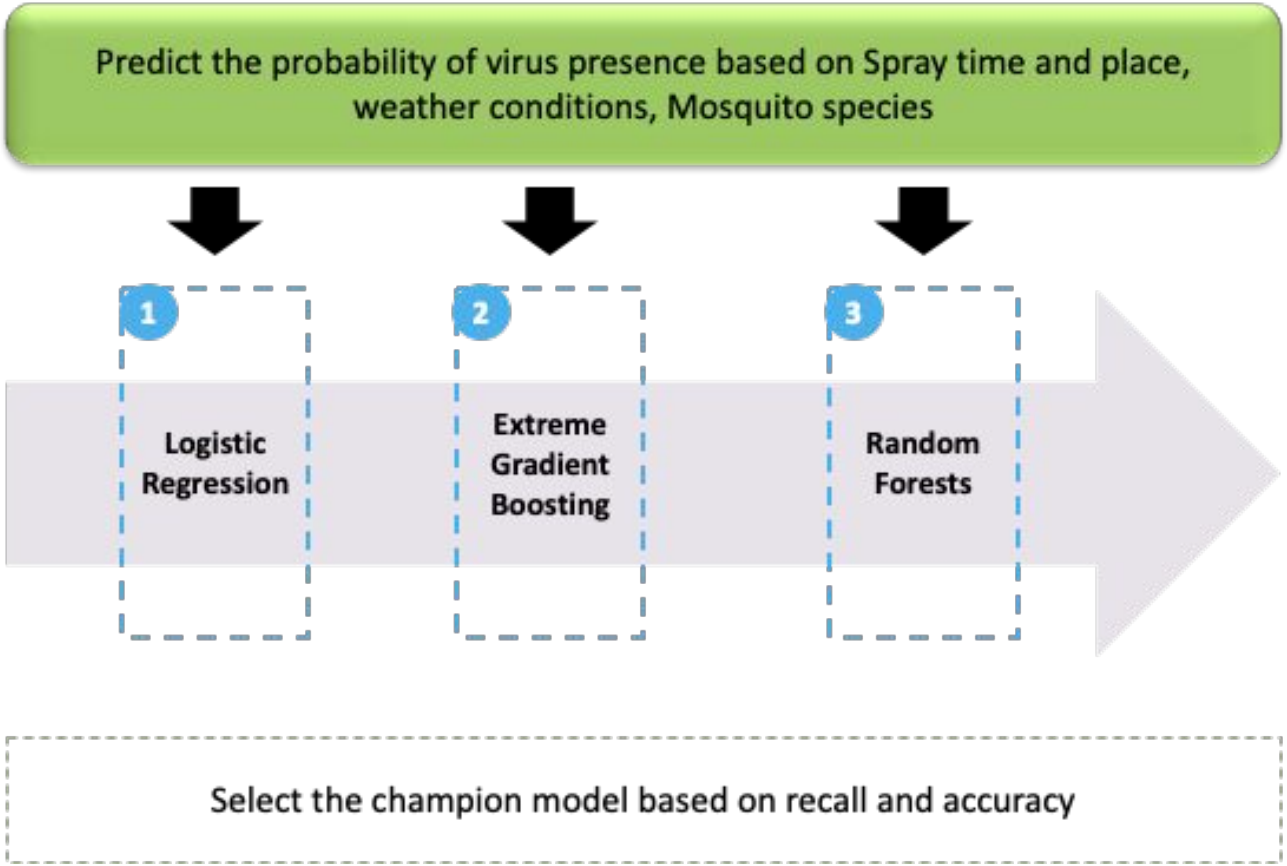
Baseline Model

We noticed that the dataset was highly imbalanced with ~95% of observations for Wnv Not Present and the remaining ~5% for Wnv Present.

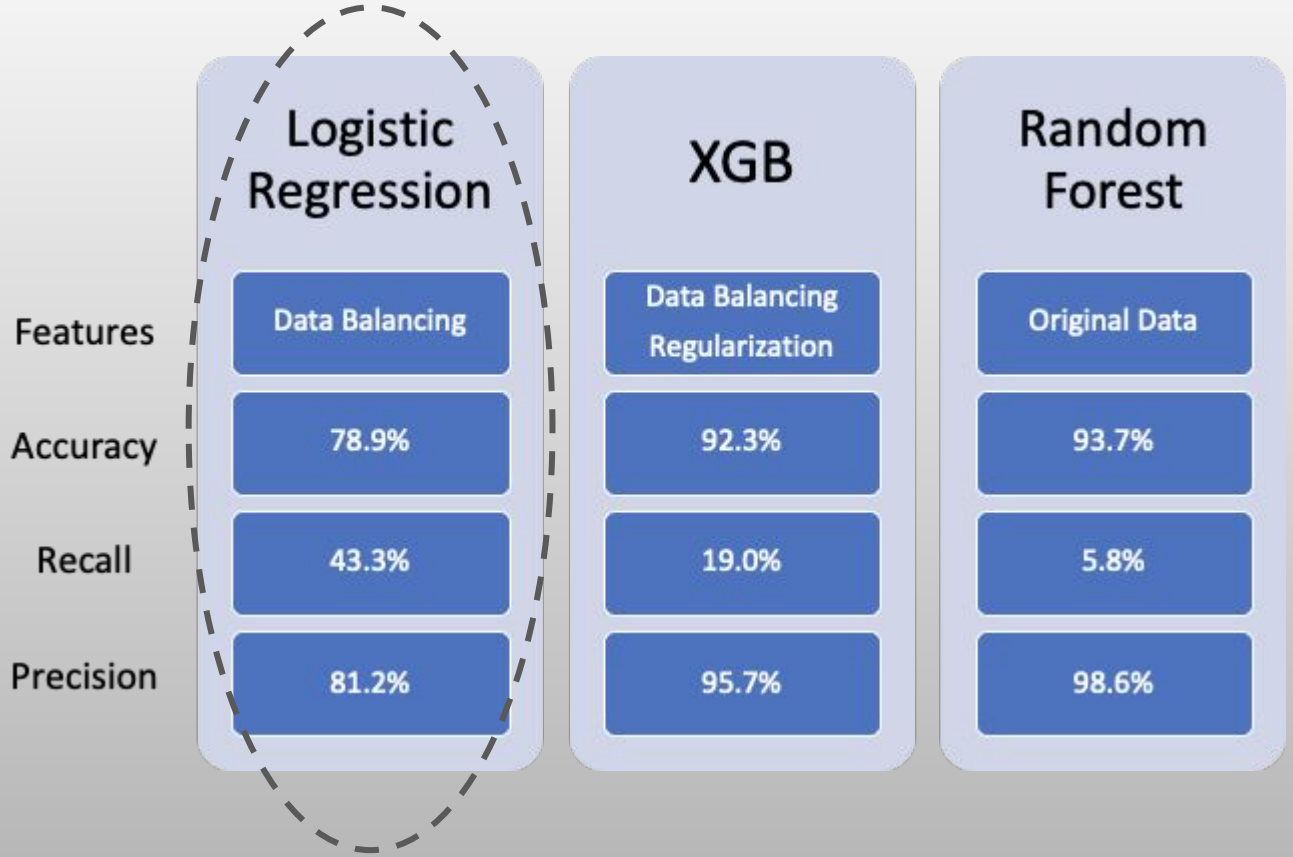
```
WnvPresent
0     8153
1      457
Name: WnvPresent, dtype: int64
```



Model Building

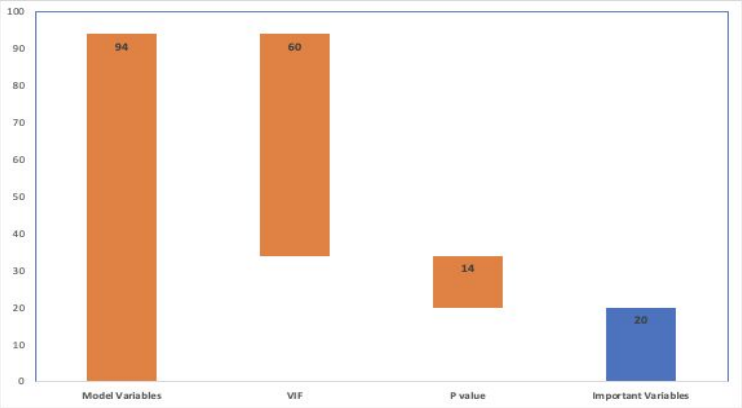


Machine Learning Models

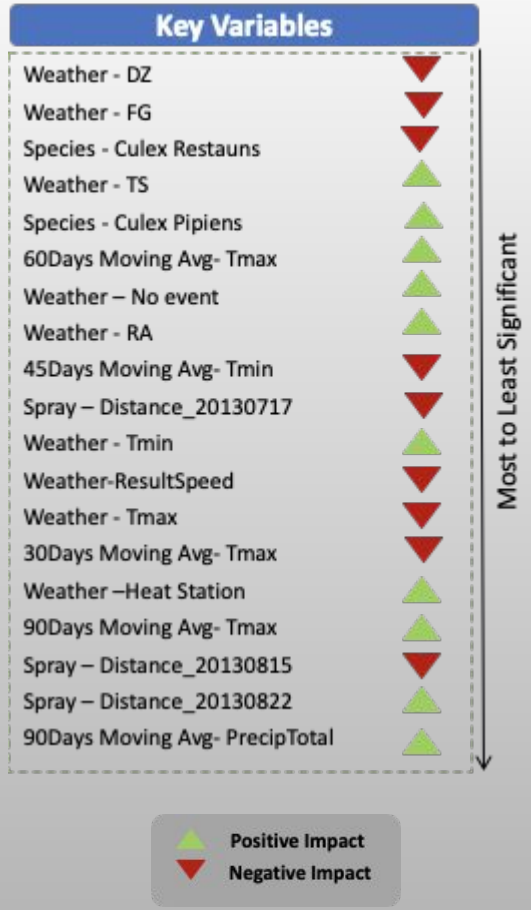


Best Model - Logistic Regression

- SMOTE for balance data
- Included Moving average variables for Weather
- VIF and p-value - variable removal



	Train	Test
Accuracy	81.4%	78.9%
Recall	81.6%	43.3%



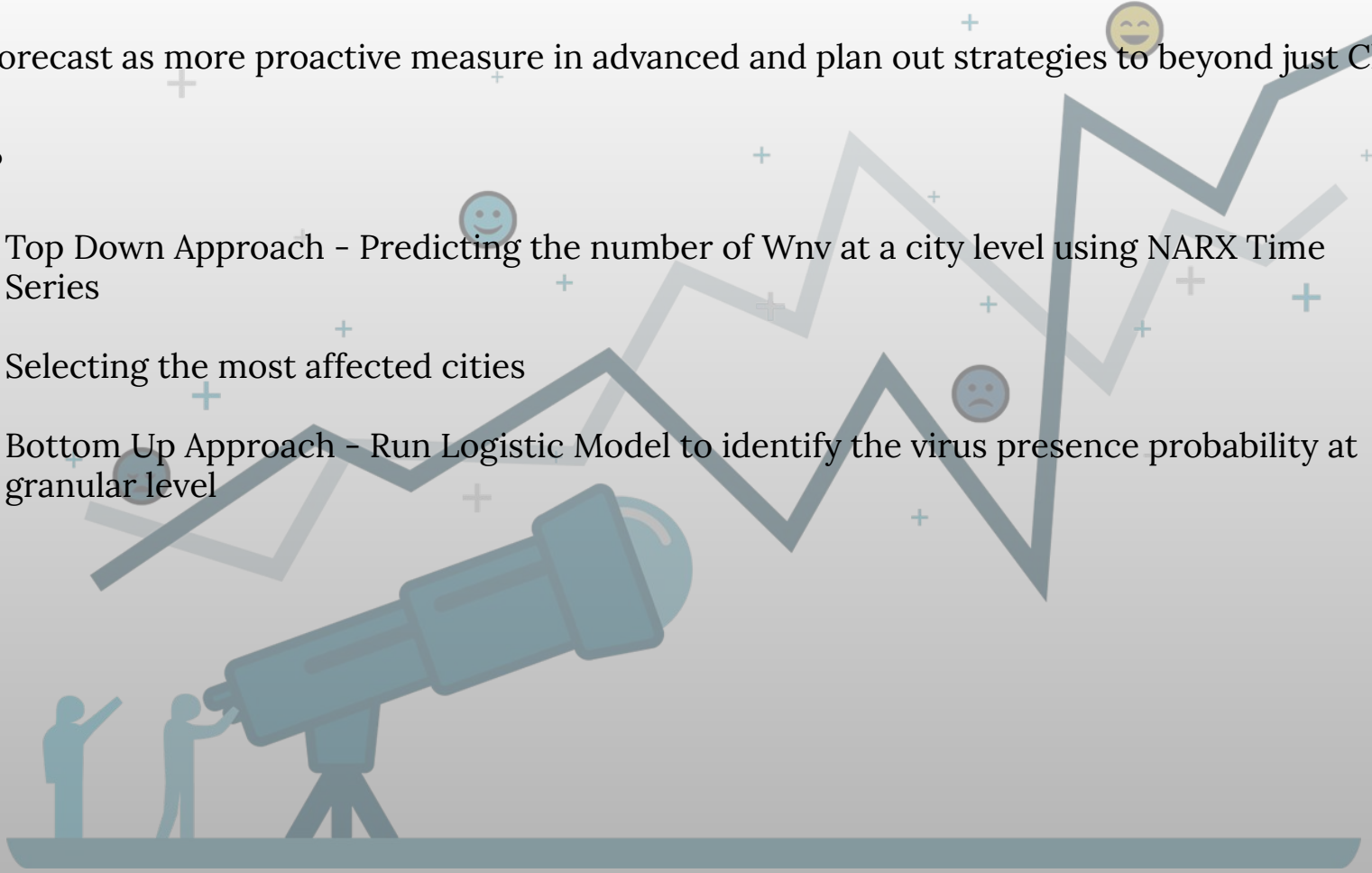
What’s Next?

Why?

WNV Forecast as more proactive measure in advanced and plan out strategies to beyond just Chicago

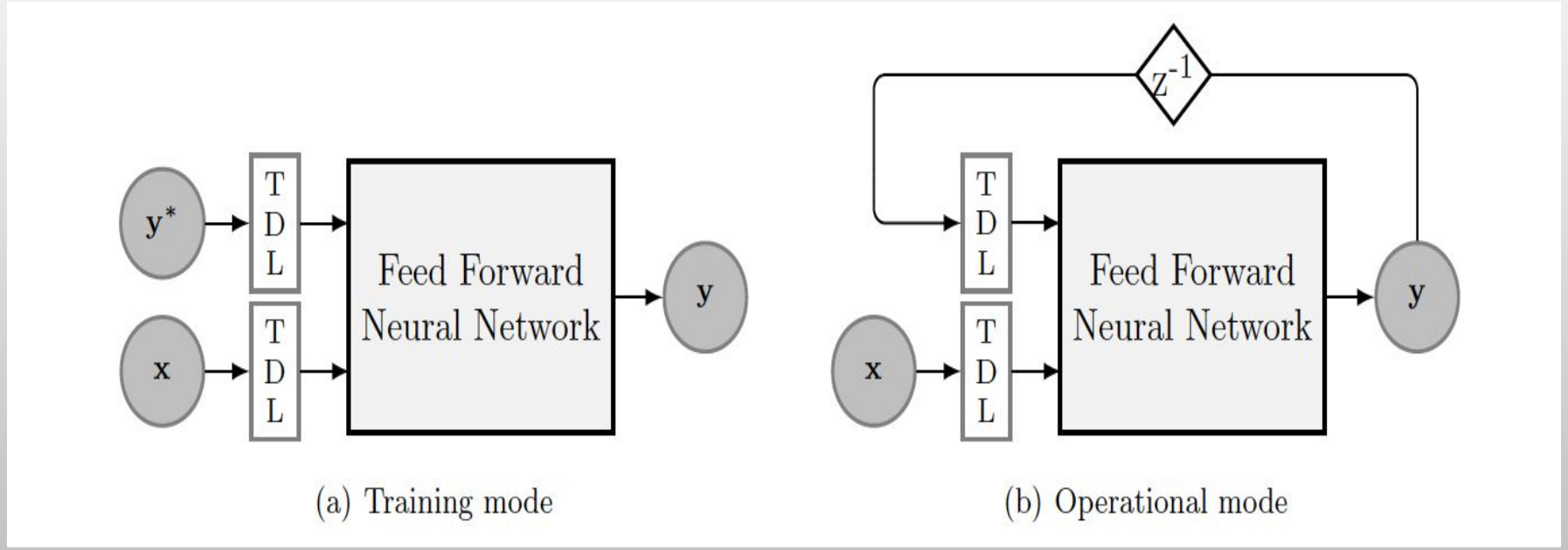
How?

- ➔ Top Down Approach - Predicting the number of Wnv at a city level using NARX Time Series
- ➔ Selecting the most affected cities
- ➔ Bottom Up Approach - Run Logistic Model to identify the virus presence probability at granular level



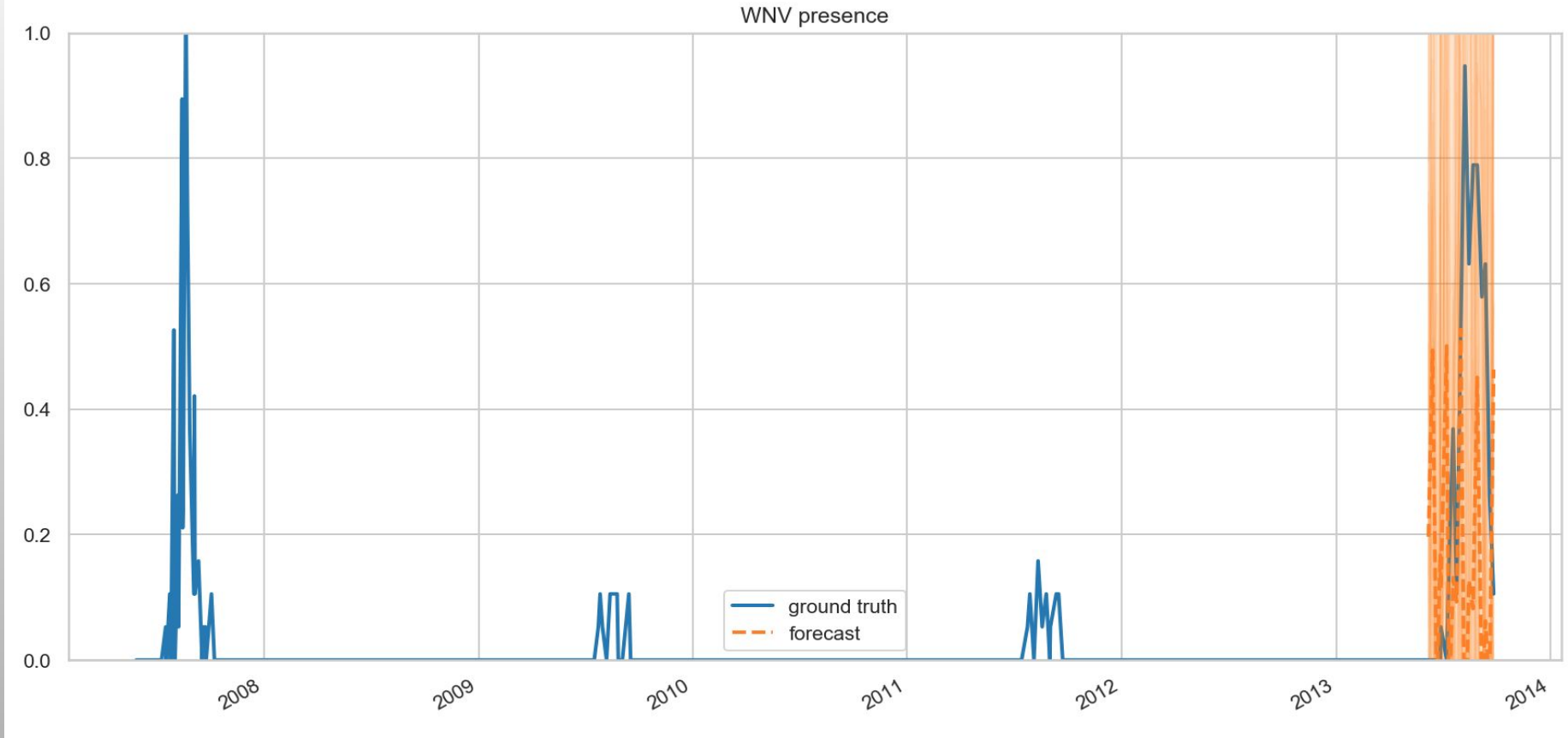
Approach

Nonlinear AutoRegressive Neural Net with Exogenous Variables (NARX)



Station 1: NRMSE = 0.3932827
Station 2: NRMSE = 0.40466395

WNV Forecast



Key Learnings and Takeaways

1. Accuracy is not the best measure
2. Having context background is key
3. Assess your prediction based on the context knowledge
4. Look at the big picture to impact bigger scope
5. Most importantly, understand the objective