



WEST NILE VIRUS PREDICTION

Presented by:

Sheng Jun Ang | Russell Quah | Jamie Pok | Peggy Man



TABLE OF CONTENTS

01

PROBLEM STATEMENT

Context and Problem
Definition

02

DATASETS

Traps (Train & Test), Weather,
Spray

03

DATA CLEANING & EDA

Mosquitoes species, map for
spray & trap, time series

04

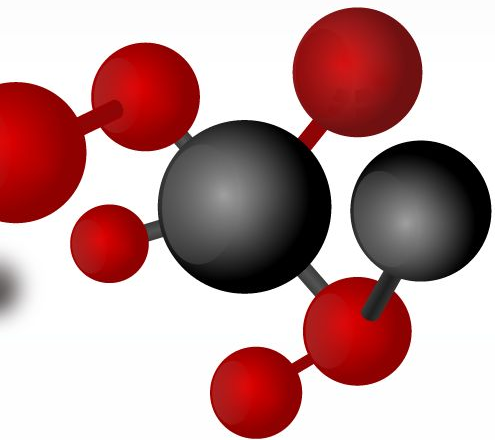
MODELLING

Classifiers
HP tuning
Model Evaluation

05

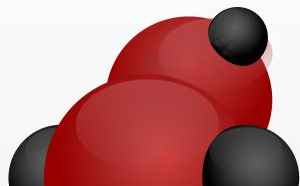
RECOMMENDATIONS

Cost-benefit analysis,
Way forward



01

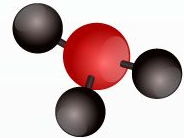
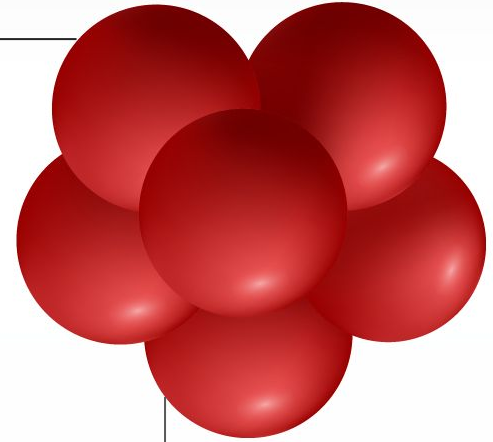
PROBLEM STATEMENT

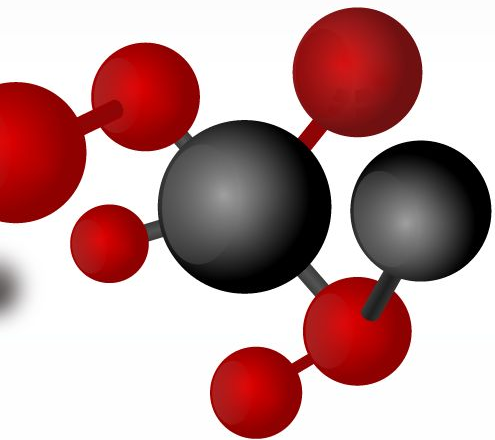


PROBLEM STATEMENT

To predict the probability of West Nile virus presence for a given location, date, and mosquitoes species

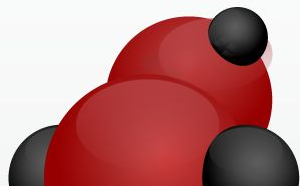
The findings will provide insights on where and when to spray airborne pesticides throughout the Chicago city, to optimize pesticide effectiveness with minimum cost.





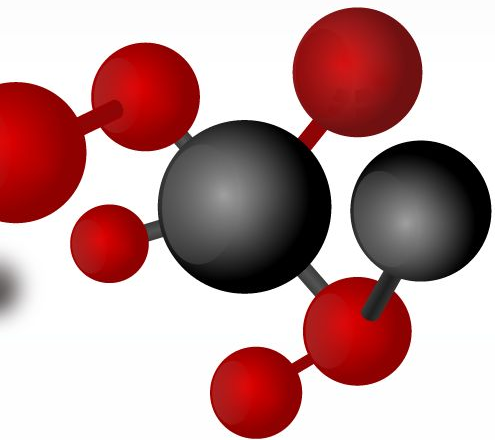
02

DATASETS



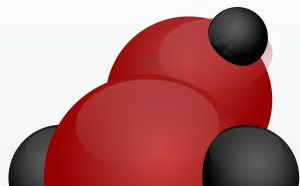
DATASETS

DATASET	Shape	Description	Recorded Date
Trap (Train Set)	10506 X 12	Date & location Test on trapped mosquitoes Virus carrier No. of mosquitoes	May-Oct 2007 May-Oct 2009 Jun-Sept 2011 Jun-Sept 2013
Trap (Test Set)	116293 X 11	Date & location Test on trapped mosquitoes	Jun-Sept 2008 Jun-Oct 2010 Jun-Sept 2012 Jun-Oct 2014
Weather	2944 X 22	Weather conditions from TWO weather stations	Jan-Dec 2007-14
Spray	14835 X 4	Date & location of spray of airborne pesticides	Jun-Sept 2011 Aug-Sept 2013



03

DATA CLEANING & EDA





DATA CLEANING CHALLENGES

TRAIN



813 of duplicate(s)?
Not so.. New entry where mosquitos
exceed 50

SOLUTION: Groupby date, trap, and
species, lat, long, sum the mosquitos.

WNV Presence



WNV Class imbalance::
Negative WNV: 8153
Positive WNV: 457

SOLUTION: SMOTE



WEATHER

Weather data incomplete
E.g. Missing value represented by 'M', 'T'

SOLUTION: Requires to study
noaa_weather_qclcd documentation.
Converted 'M' to NaN, filled in 'T' values



SPRAY

Missing value on time column

SOLUTION: Using date as areas sprayed
over time



EDA

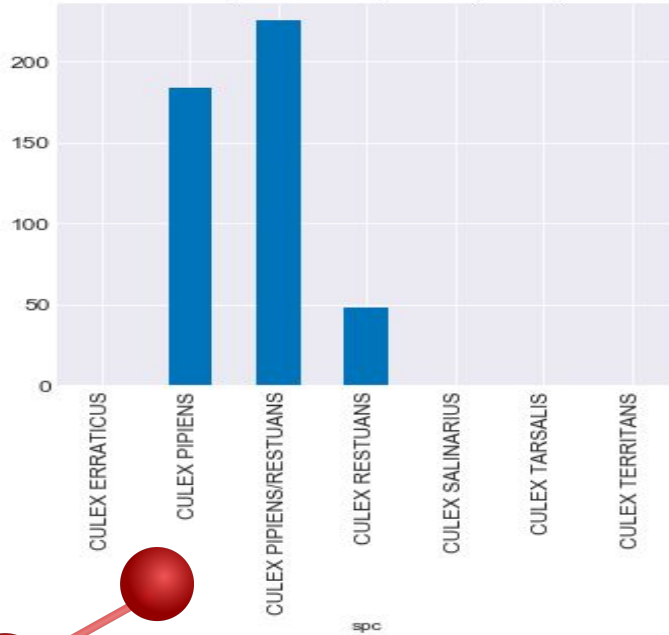


MOSQUITOES SPECIES

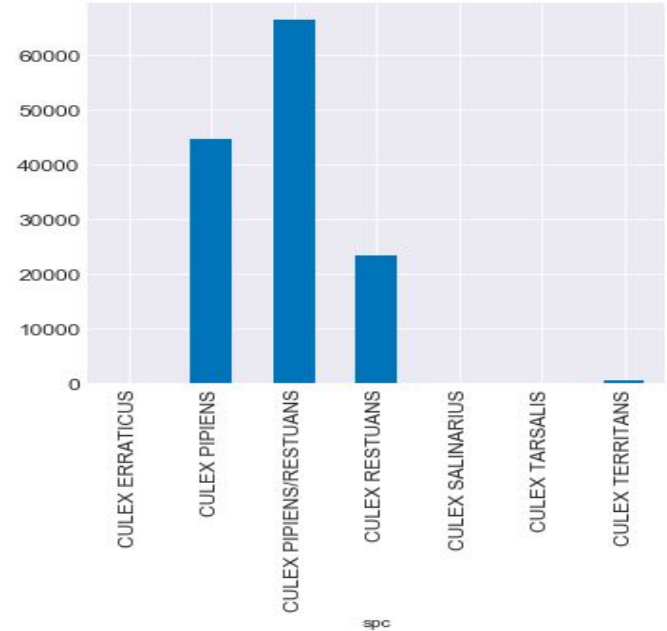
WNV Vectors

PIPENS
RESTUANS

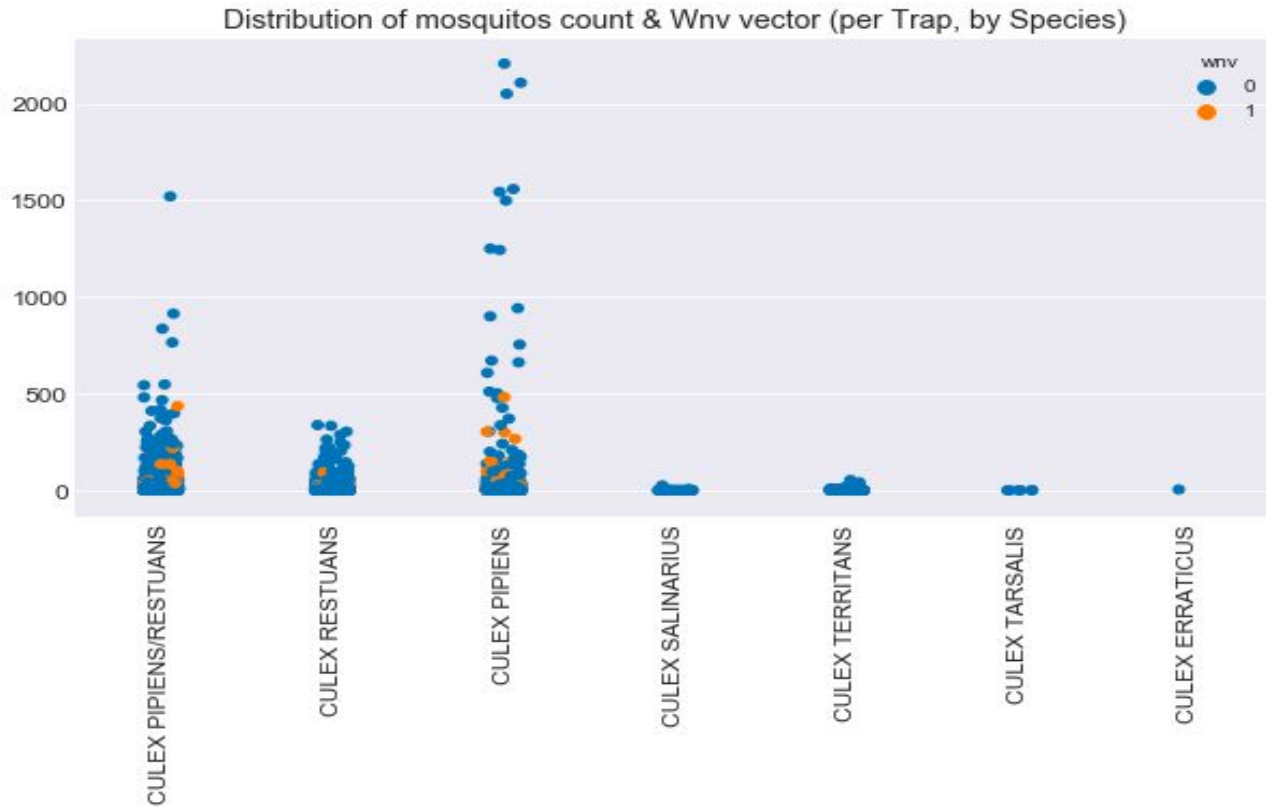
Mosquito count (WnV Species)

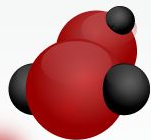


Mosquito count (by Species)

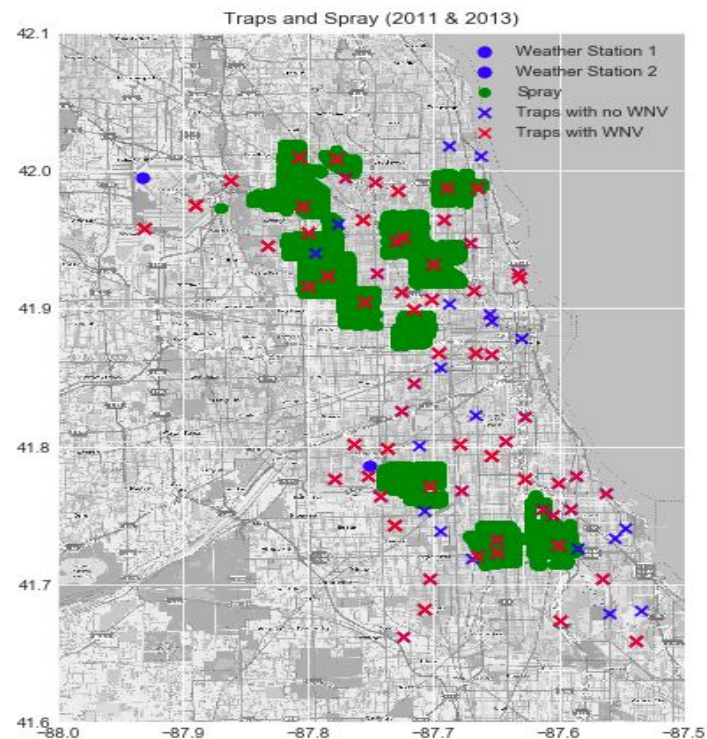
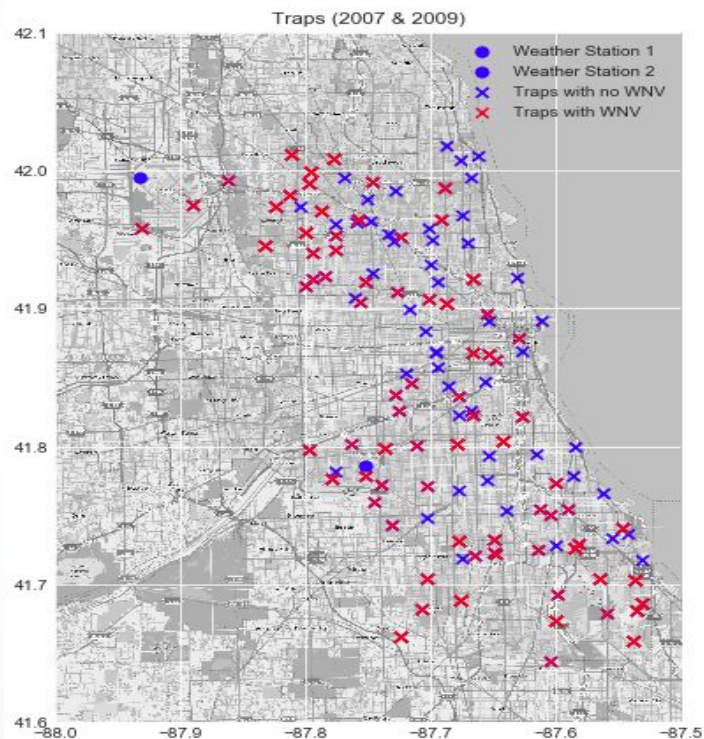


EDA

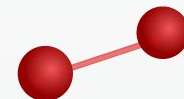




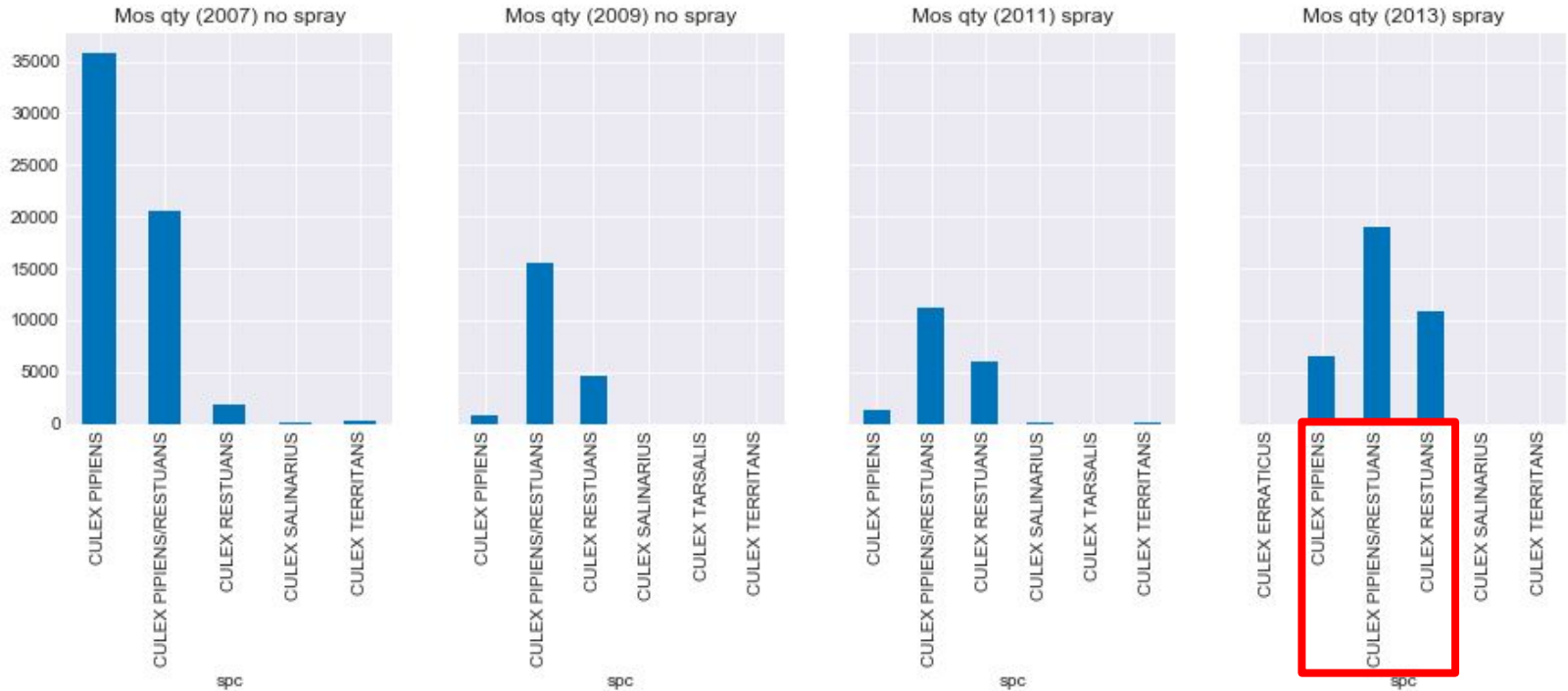
EDA



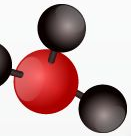
MAP OF TRAPS & SPRAY
COVERAGE



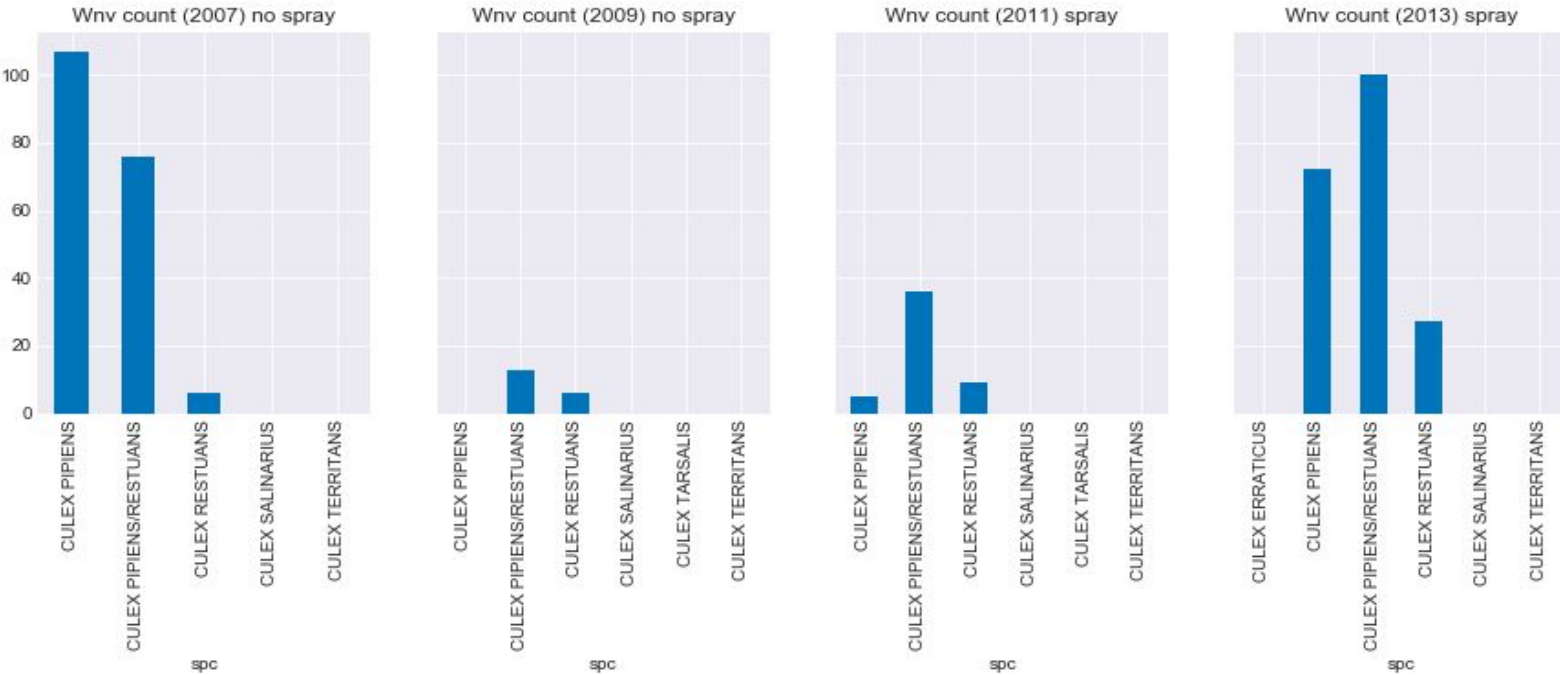
EDA



Mosquito Counts for Each Year



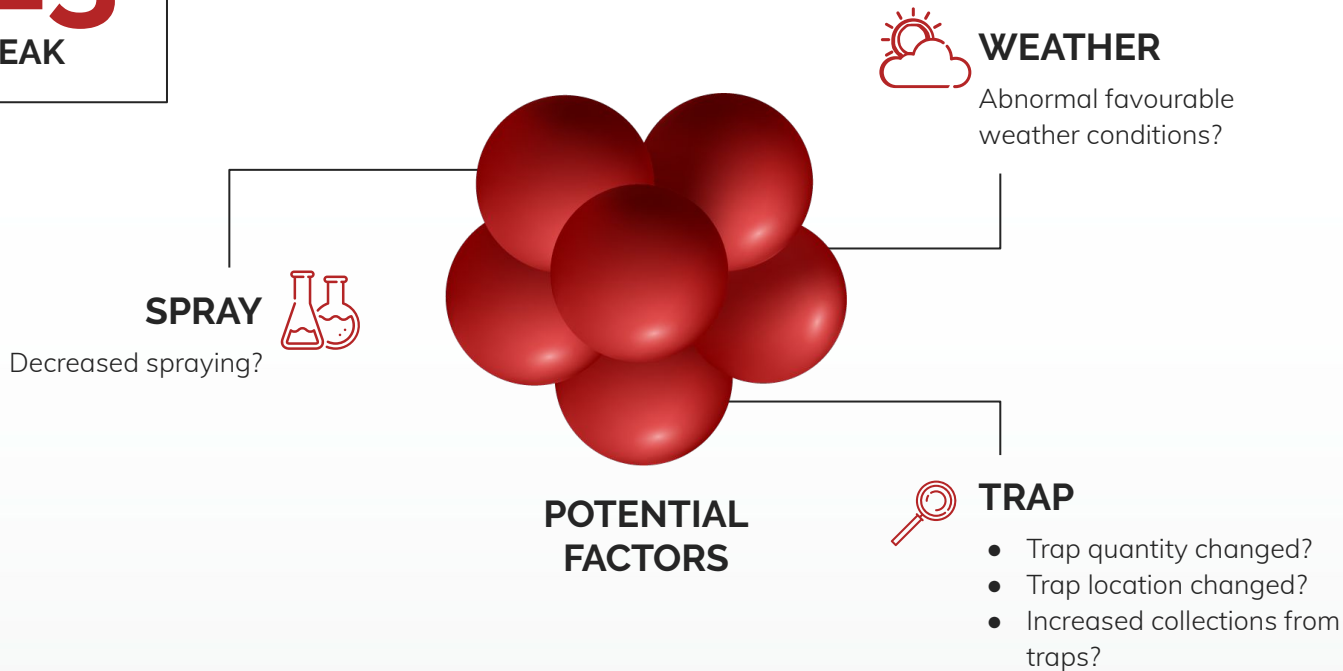
EDA

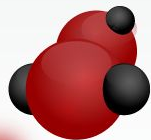


WNV Prevalence for Each Year

Potential contributors of outbreak

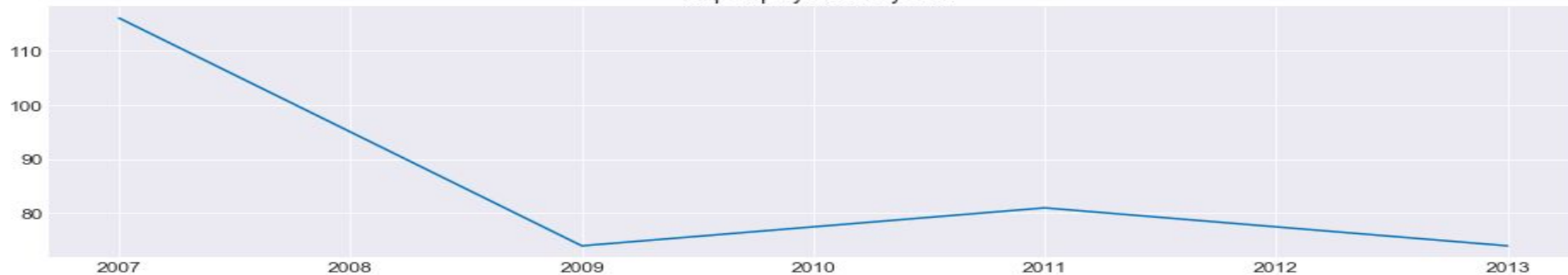
2013
OUTBREAK



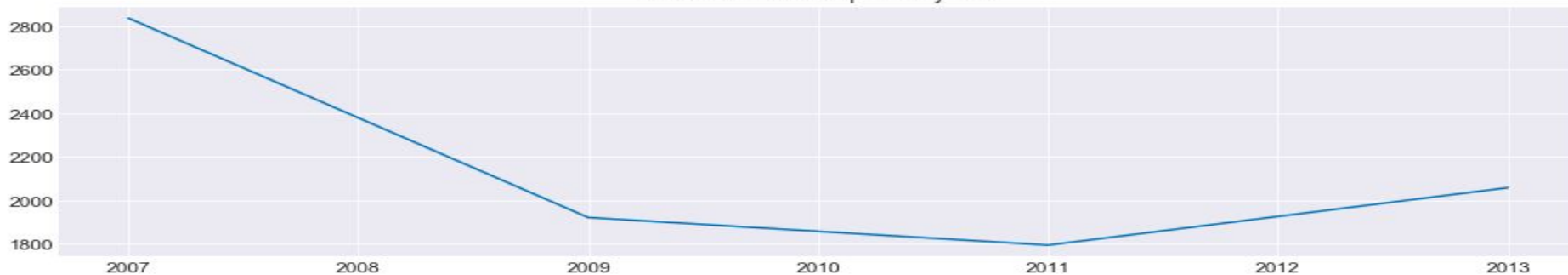


EDA

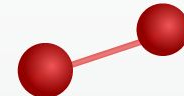
Trap deployed over years



Collections from traps over years

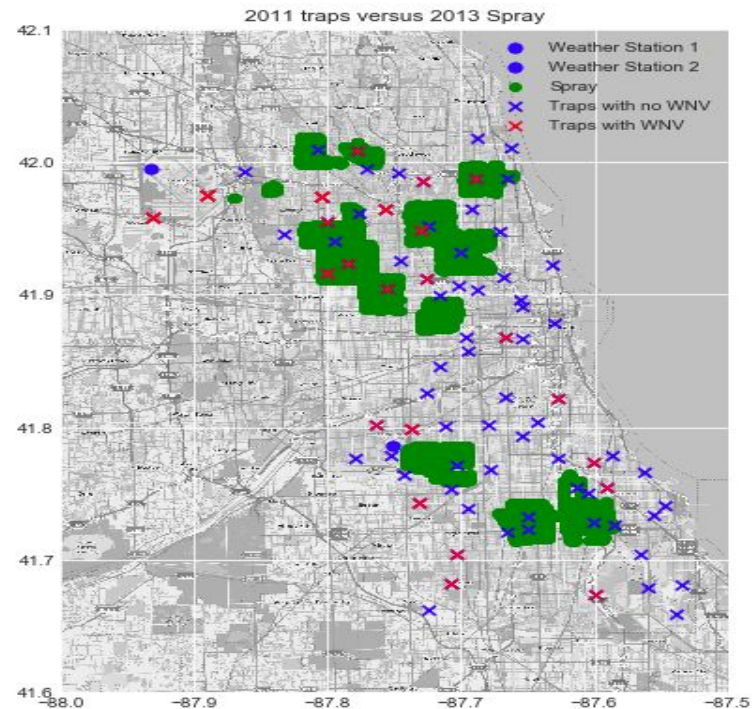
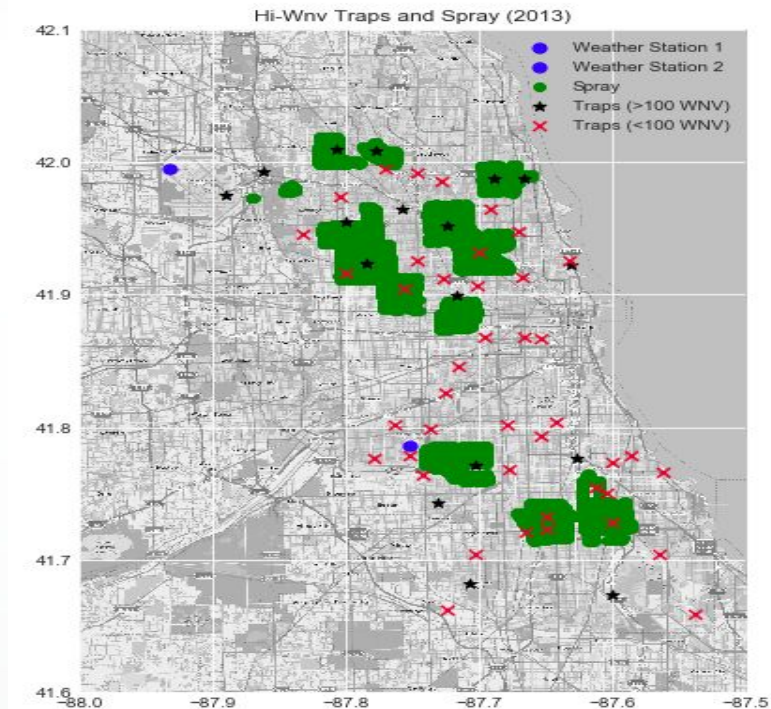


Examine Potential Factors: Change in Trap Quantity

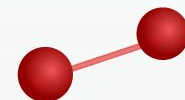




EDA



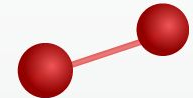
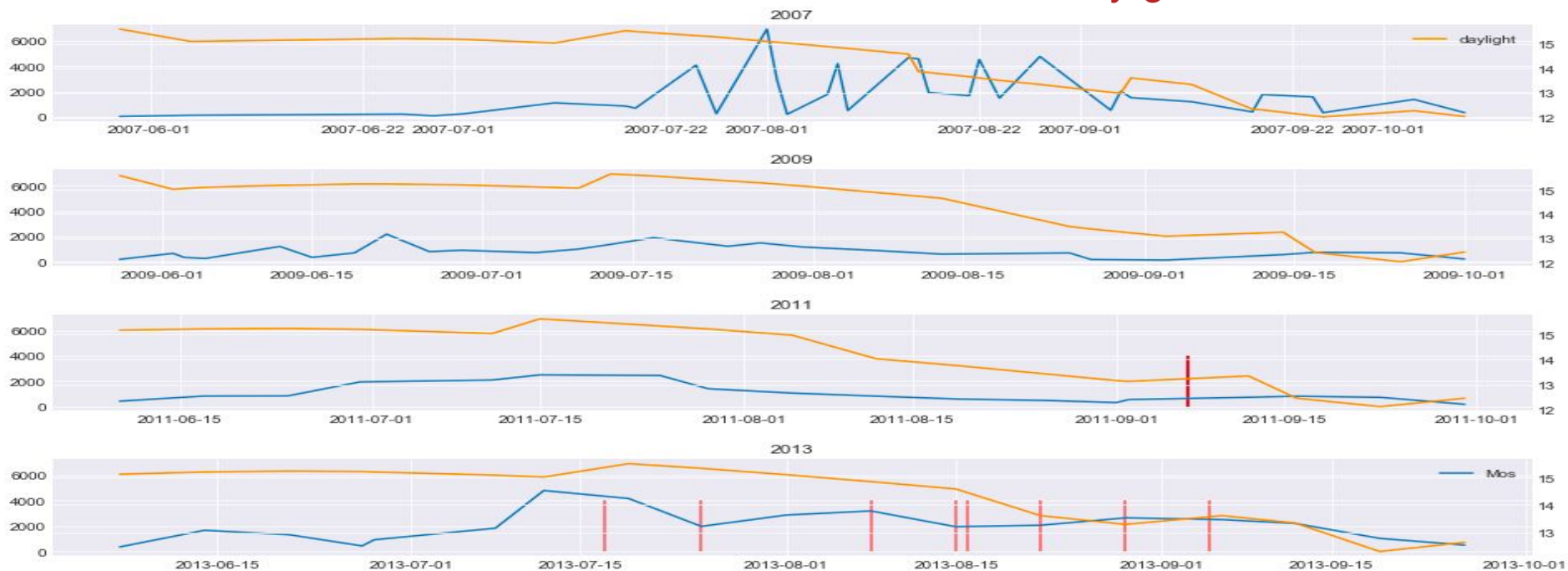
Examine Potential Factors: Trap Locations Changed





EDA

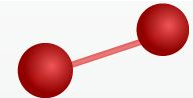
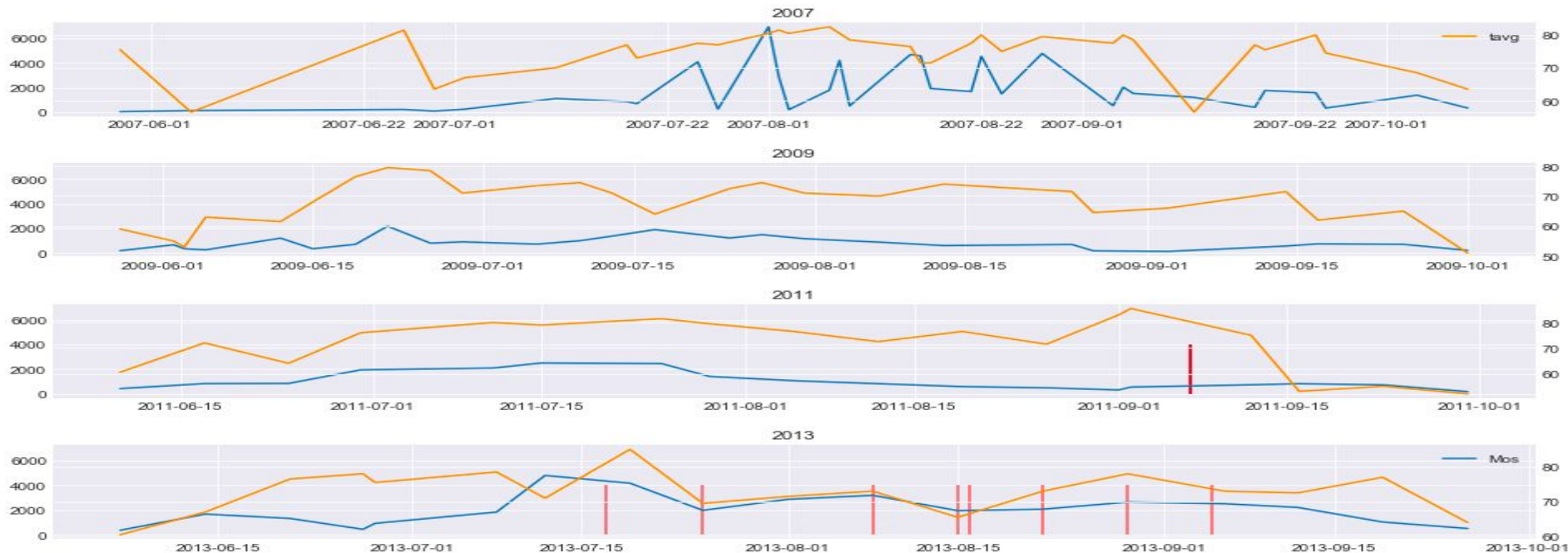
Examine Potential Factors: Weather Conditions - **Daylight**





EDA

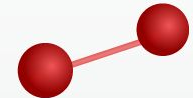
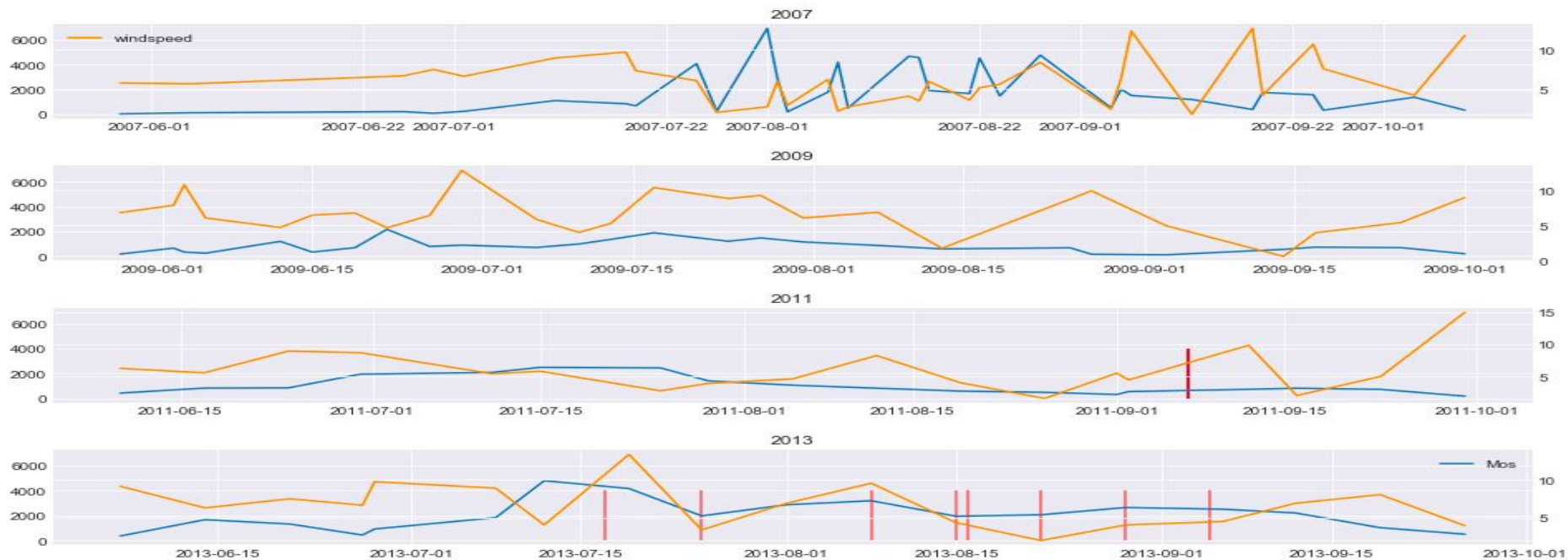
Examine Potential Factors: Weather Conditions - **Average Temperature**





EDA

Examine Potential Factors: Weather Conditions - **Wind Speed**



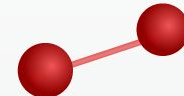


EDA

Examine Potential Factors: Weather Conditions - **Rain Fall**



High rainfall, not experienced in years above



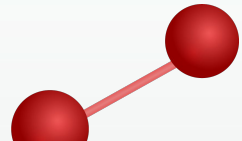


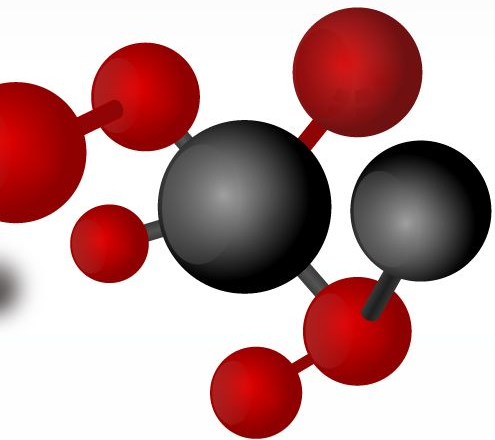
EDA



SUMMARY: FACTORS LEADING TO 2013 OUTBREAK

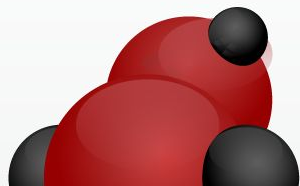


- Heavy rainfall from mid-June to mid-July supported by relatively high temperatures and wind speeds (< 10 mph)
 - No spraying in months prior to July
 - Spraying if done, missed areas of traps with high wnv (more than 100).
- 



04

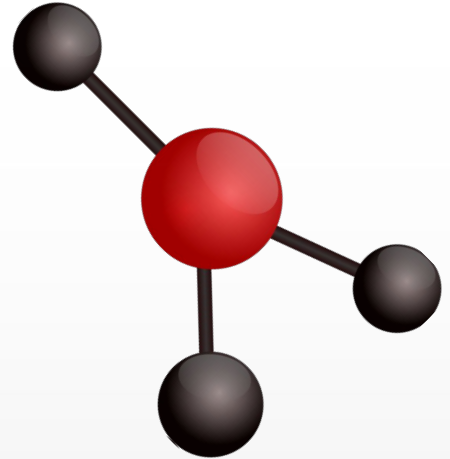
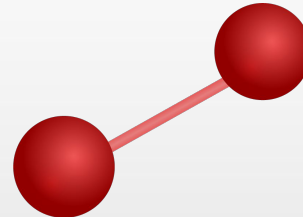
MODELLING



MODELLING

- Resampling (SMOTE Upsample)
 - Std Scale (where relevant)
-
- Logistic Regression
 - Support Vector Machine (SVM)
 - K Nearest Neighbours (KNN)
 - Decision Tree Classifier
 - Random Forest*
 - Gradient Boosting Classifier
 - XGBoost Classifier

*Downsampling + Upsampling explored



SUMMARY

	LogReg	SVM	KNN	DT	RF(Smote O&U)	GBc	XGBc
accuracy(val)	0.945	0.918	0.934	0.814	0.847	0.91	0.914
sensitivity	0	0.226	0.066	0.65	0.664	0.226	0.285
precision	0	0.226	0.173	0.171	0.212	0.196	0.241
F1	NaN	0.226	0.095	0.271	0.322	0.21	0.261
roc_auc	0.791	0.795	0.721	0.804	0.866	0.847	0.849

- Picked both Random Forest and XGBoost classifier as the best model, based on F1 score and roc_auc, further explored (Smote O&U) on XGBoost.
- Kaggle ROC_AUC Public:
- XGBoost (Smote U): 0.74243
- RandForest (Smote O&U): 0.75238
- XGBoost (Smote O&U): 0.77011



05

RECOMMENDATIONS



COST-BENEFIT ANALYSIS (Year 2018)

Zenivex coverage (acre) per barrel

= $30(\text{gal barrel}) \times 1.48(\text{AI per barrel}) / 0.007 (\text{appln rate}) = 6343 \text{ acre}$

Recommended area coverage = $128,000 + 35,200 = 163,200 \text{ acre}$

(area 1 Lat 42.05 to 41.9 Lon -87.9 to -87.62 , area 2 Lat 41.79 to 41.68 Lon -87.75 to -87.6)

Budget allocated for Public Health = 5,235,000 USD

Expected cost per spray run = 277,881 USD

Recommended Sprays (2 per month, Jun to Aug) = 6

Potential Losses = $176(\text{Wnv cases}) \times 38,000(\text{Mean cost per case}) = 6,688,000 \text{ USD}$


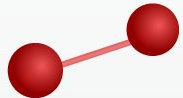
Cost-Benefit Analysis = $277,881 \times 6 / 6688000 = 0.25$ case reduction to be C/B neutral






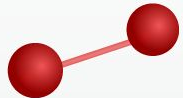
RECOMMENDATIONS

To improve cost-effectiveness of pesticide deployment, the proposed recommendations relies on timing and coverage area:

- Spraying should be focused in the months of Jun to Jul (periods of high rainfall), and targeted at region of traps with high wnv as a start.
 - Future deployments should be tailored accordingly to match rainfall patterns; mosquito population generally spike 2 weeks after heavy rainfall.
 - 0.25 WNV case reduction to be considered C/B neutral.
 - Data on the type of pesticide and cost per spray could be collected to provide more precise cost benefit spray recommendations
 - Alternatives of control: larvicides, natural predators (guppies, koi) to further check mosquito population
- 
- 



RECOMMENDATIONS

- The classifier model could be used to provide insights to areas for targeted spraying in the longer term, as new data on wnv clusters, and weather data is available.
 - Other sources of data could further improve analytics and model prediction: residential areas, schools, and nursing homes. Potential insights to
 - types and general state of the residential areas; to better inform alternative mosquito control programs through public outreach campaigns,
 - identify potential sites (e.g. work areas where rainwater may pool unnoticed) for mosquito breeding for early prevention efforts
 - areas of higher risk (i.e. children and older folks) that could influence spraying times
- 
- 

The slide features several decorative molecular models. In the top left, there is a small molecule with two red spheres connected by a thin red line, and another red sphere below it. In the top center, there is a large, dense cluster of red spheres. To the right of this cluster, there is a single red sphere. On the far right, there is a molecule with two black spheres and one red sphere. The word "THANKS" is centered in the middle of the slide in a large, bold, red font.

THANKS

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.
Please keep this slide for attribution.



LITERATURE REVIEW

Cases in Chicago	https://www.cdc.gov/westnile/statsmaps/cumMapsData.html#one
Mean cost per case averted for year 2018 (USD)	“Cost effectiveness of a targeted age-based West Nile virus vaccination program” by Shankar, et al. Elsevier public health journal, May 2017
Cost of pesticide (USD) per 30 gallon barrel	https://www.dnainfo.com/chicago/20160907/north-park/city-s-pray-north-park-for-mosquitoes-wednesday https://www.forestrydistributing.com/aqua-zenivex-e20-ulv-in-secticide-zeocon
	1.48 lbs Etofenprox per gallon, 0.007Al/acre appln rate
Budget allocated for Public Health	https://www.chicago.gov/content/dam/city/depts/obm/supp_info/2018Budget/2018_Budget_Overview.pdf

