Predicting West Nile Virus in Chicago

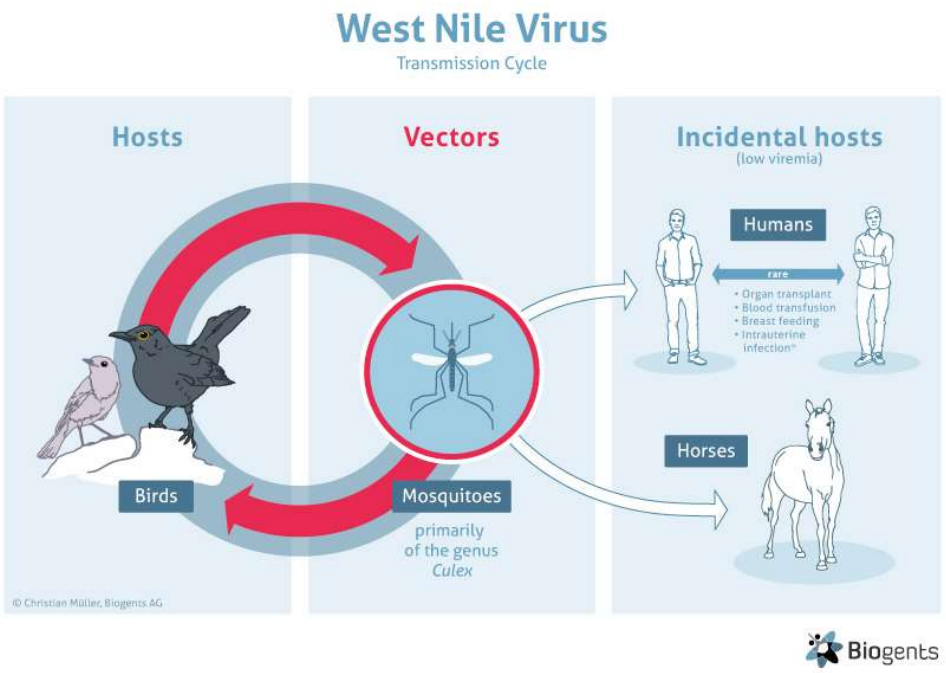
John Han Wei Tan

Christopher Gozali

Angeline Chandraatmadja

PROBLEM STATEMENT

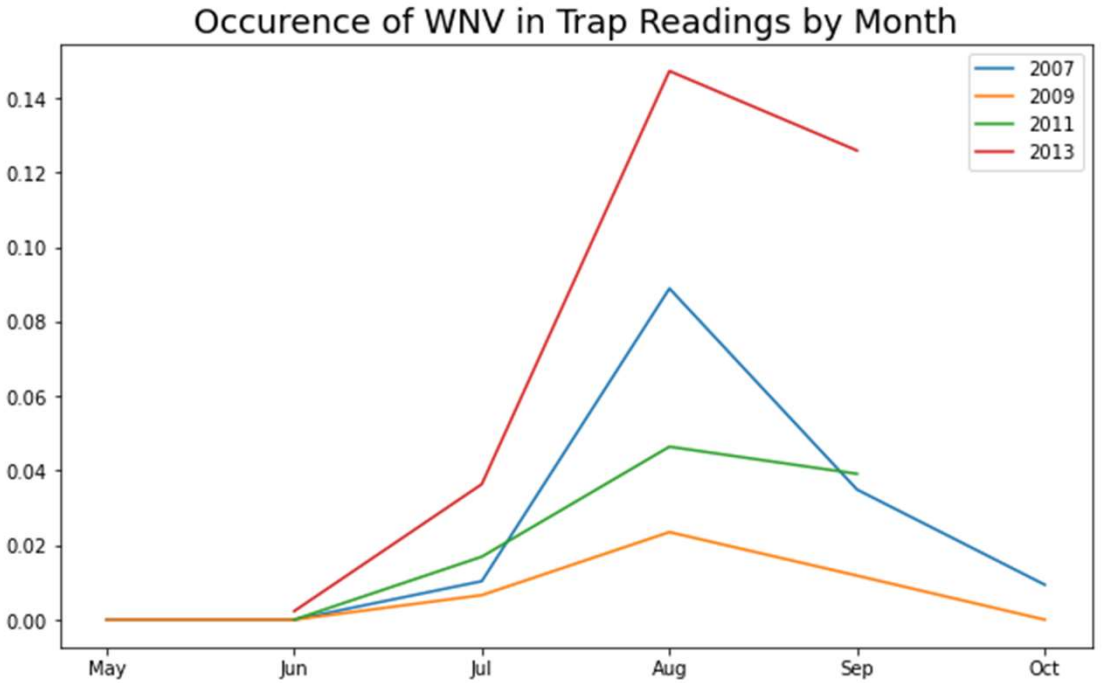
- There is growing concern in the recent spread of West Nile Virus (WNV) in Chicago
- DATA-SCIENCE has collaborated with the Public Health Department of Chicago to develop a predictive model for WNV occurrence.
- Model optimised for screening purposes



ABOUT THE WNV

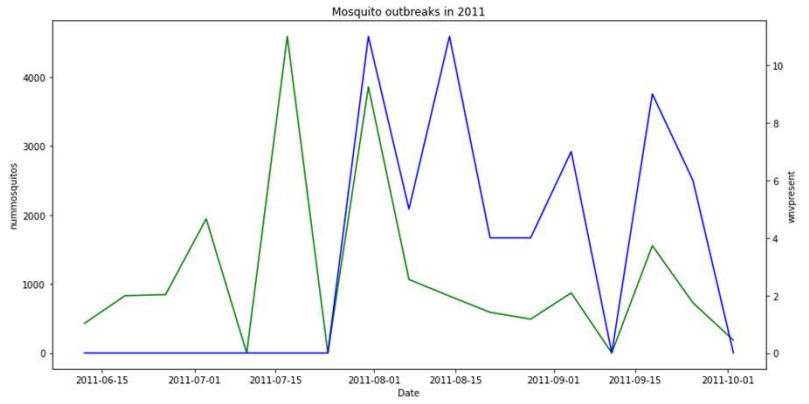
- 1/5 of infected humans experience symptoms of WNV
- Symptoms include fevers, body aches
- Serious symptoms occur when virus attacks central nervous system
- 1/1500 fatality rate in humans.

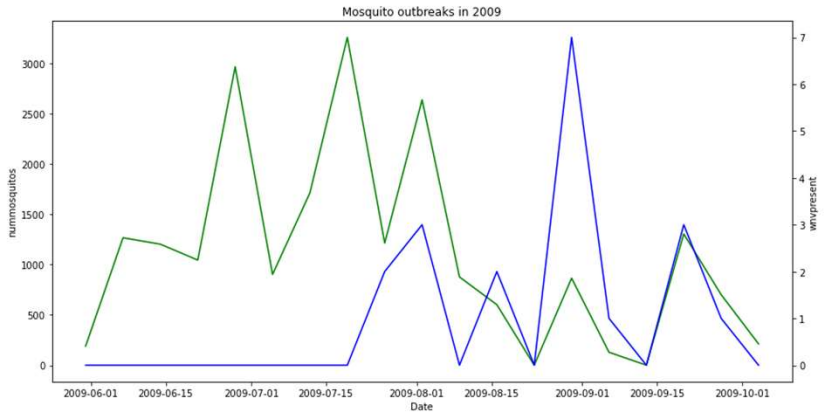
West Nile Virus peaks in August (summer months)



Trends

- Spikes in the number of mosquitos coincides with WNV detection
- Usually a time lag of between 2 – 4 weeks between peaks
- Feature engineering on features related to mosquito breeding cycles and activity





MODEL FEATURES

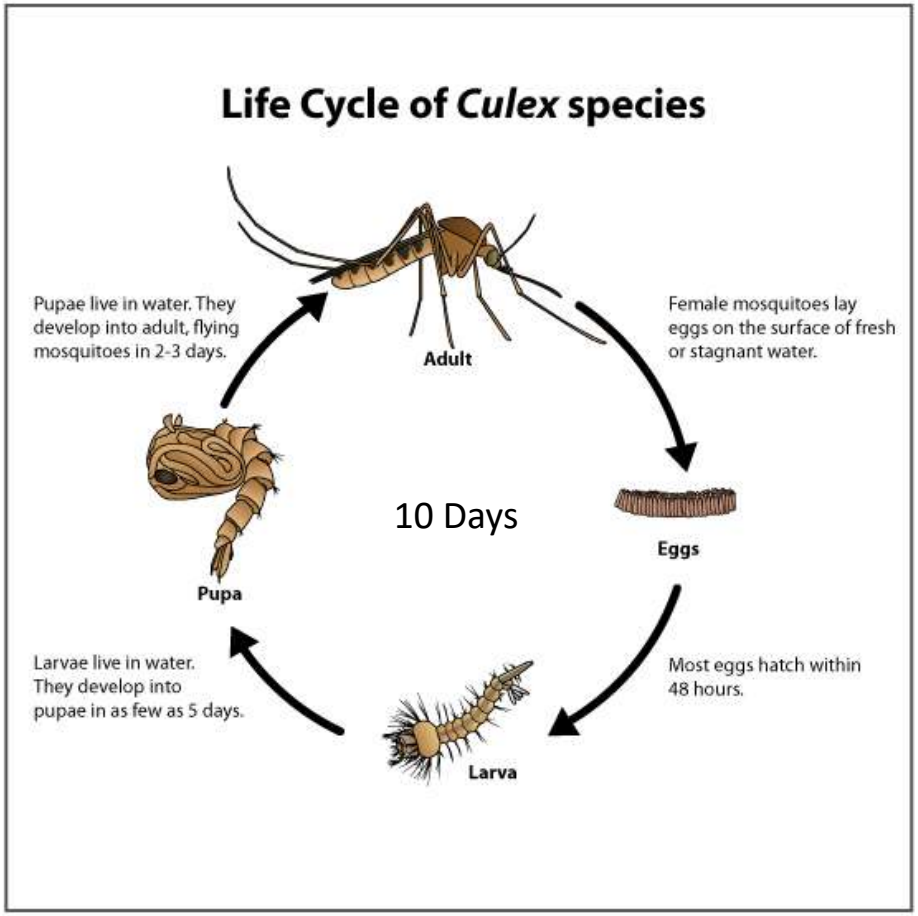
Feature name	Feature Type	Rationale
species	Nominal (pipiens, pipiens/restuans, restuans)	Only two species of the Culex mosquito are carriers of WNV
tavg	Continuous	Related to mosquito activity
depart	Continuous	Related to mosquito activity
dewpoint	Continuous	Related to mosquito activity
cool	Continuous	Related to mosquito activity
daylight hours	Continuous (Engineered)	Related to mosquito activity
relative humidity	Continuous (Engineered)	Related to mosquito activity
weathertype	Nominal (ra, hz, br, fg, ts, vc, dz)	Related to mosquito activity
preciptotal	Continuous (Engineered)	Related to mosquito breeding
stnpressure	Continuous (Engineered)	Related to rainy season
sealevel	Continuous (Engineered)	Related to rainy season
resultspeed	Continuous (Engineered)	Related to rainy season

Species

species	nummosquitos	wnvpresent
CULEX ERRATICUS	7	0
CULEX PIPIENS	44671	240
CULEX PIPIENS/RESTUANS	66268	262
CULEX RESTUANS	23431	49
CULEX SALINARIUS	145	0
CULEX TARSALIS	7	0
CULEX TERRITANS	510	0

Culex restuans:

Culex pipiens:



Engineered Features

- Daylight Hours
- Relative Humidity
- Lagged and Rolled features
  - *Preciptotal*
  - *Stnpressure*
  - *Sealevel*
  - *Resultspeed*

Daylight Hours

Culex species are not active during the day

Greater mosquito activity on longer nights

Daylight hours engineered by calculating time between *sunset* and *sunrise*

Relative Humidity

- Relative humidity estimated by using dewpoint and tavg to calculate vapour pressure at atmospheric conditions.

$$e = 6.11 \times 10^{\left(\frac{7.5 \times T_d}{237.3 + T_d}\right)} \qquad e_s = 6.11 \times 10^{\left(\frac{7.5 \times T}{237.3 + T}\right)}$$

$$RH = \frac{e}{e_s}$$

Symbol	Meaning
e, e_s	Actual vapour pressure, Saturated vapour pressure
T, T_d	Temperature, Dewpoint Temperature

Lag and roll

- Preciptotal:
  - Rain forms pools of stagnant water which is deal for mosquito breeding
  - 10 day lag to account for Culex lifecycle
- Stnpressure, Sealevel, Resultspeed:
  - Features describe atmospheric pressure and wind vector sum (speed and direction)
  - Related to wet/dry seasonality
  - 28 day lag
  - 7 day rolling average

MODELLING

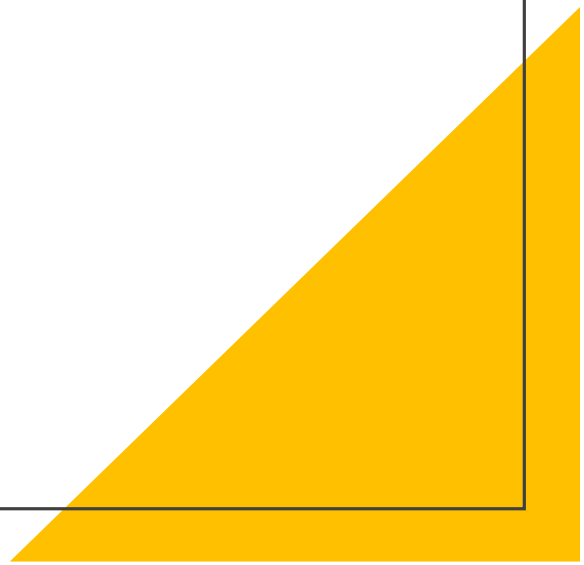
Image: zifphoto

GRIDSEARCH

Models chosen:

- Logistic Regression
- Random Forest
- Ada Boost
- Gradient Boost
- Neural Network

The models are optimized over AUCROC score



SELECTION CRITERIA

- **RECALL**
- We are interested in the positive class (WNV present), recall captures the true positive rate.

- **Less Overfitting**
- Small difference between train and validation AUCROC score

MODEL SELECTION & EVALUATION

SELECTION

	Logistic Regression	Random Forest	ADA Boost	Gradient Boost	Neural Network
Train AUC ROC	81.90%	89.56%	74.36%	76.22%	87.43%
Holdout AUC ROC	80.86%	85.12%	72.28%	73.60%	86.13%
Recall	79.71%	87.68%	97.10%	97.10%	88.41%
Precision	12.29%	13.12%	9.65%	9.96%	13.77%

MODEL EVALUATION

- Precision is low for all methods because the presence of WNV is generally impacted with changes in weather
- ROCAUC scores for Kaggle unseen dataset is lower than the validation data
- Models are overfitted on 2007, 2009, 2011, 2013 data (train dataset)

	Holdout AUCROC	Kaggle AUCROC
ADA Boost:	72.28%	66.84%

STRONG FEATURES

Strong Logistic Regression coefficient or Random Forest high feature importance

Mosquito Activity

dewpoint – Log reg
tavg – Log reg
daylighthours – Random Forest, Log Reg

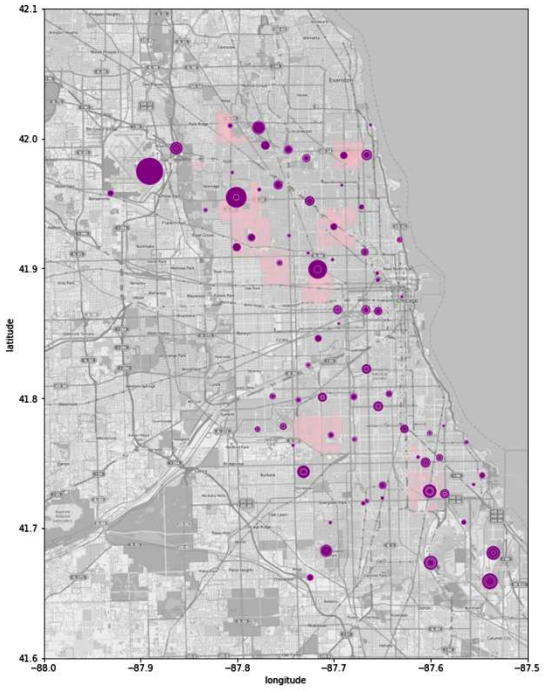
Rainy Season

resultspeed_roll7_lag28 – Random Forest
stnpressure_roll7_lag28 – Log Reg

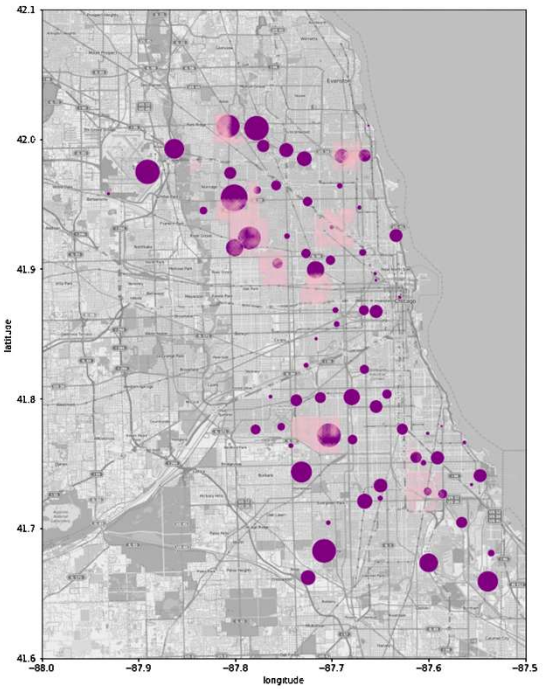
SPRAY RECOMMENDATIONS

Exploring the effects of spraying

1 week before spraying starts (Aug 8):



Aug 8 to September 14



- Mosquito Density in Traps
- Areas sprayed on Aug 8 once week to September 5

Recommendations

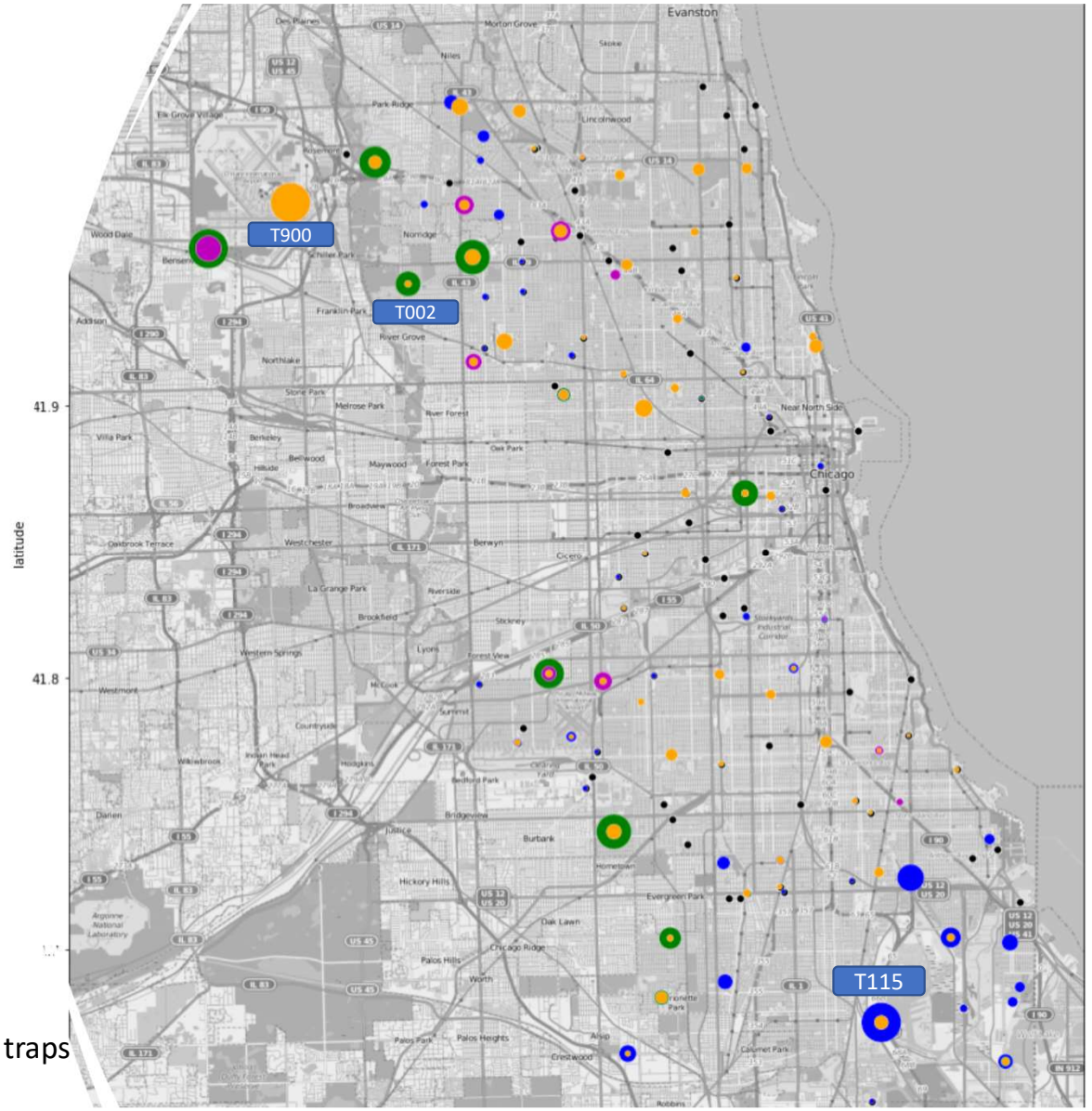
- Additional pesticide application targeting mosquito larvae/pupae
  - Apply onto areas which may accumulate rainwater
  - Apply when rainy/wet to prevent mosquito numbers from spiking 2 weeks later.
- Conduct spraying during summer months (July – September)

Project Limitations

- Model has low precision as we have over relied on city wide weather conditions
- Difficult to predict presence of WNV in specific areas of interest

WNV Hotspots

trap	latitude	longitude	nummosquitos	wnvpresent
T900	41.974689	-87.890615	15386	66
T115	41.673408	-87.599862	21668	41
T002	41.954690	-87.800991	3710	18
T138	41.726465	-87.585413	9936	16
T003	41.964242	-87.757639	1346	14
T011	41.944869	-87.832763	1311	11
T225	41.743402	-87.731435	2014	11



* Color represents year,
* Dot size represents population size of mosquito found in traps

T115
WNV Occurrence: 41
Nummo Mos: 21,668

T900
WNV Occurrence: 66
Nummo Mos: 15,386

T002
WNV Occurrence: 18
Nummo Mos: 3,710

Future Work

- Include localized environmental data besides weather
  - Current features make it difficult to predict local clusters
  - E.g. extent of urbanization, population density
- Include more spray data
  - More data will help in conducting a more rigorous analysis on spray effectiveness
- Combine different models to create ensemble model
  - 40% Logistic Regression and 60% ADA boost improved Kaggle score to 0.706

Thanks!