TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Informatik

West Nile Virus Group

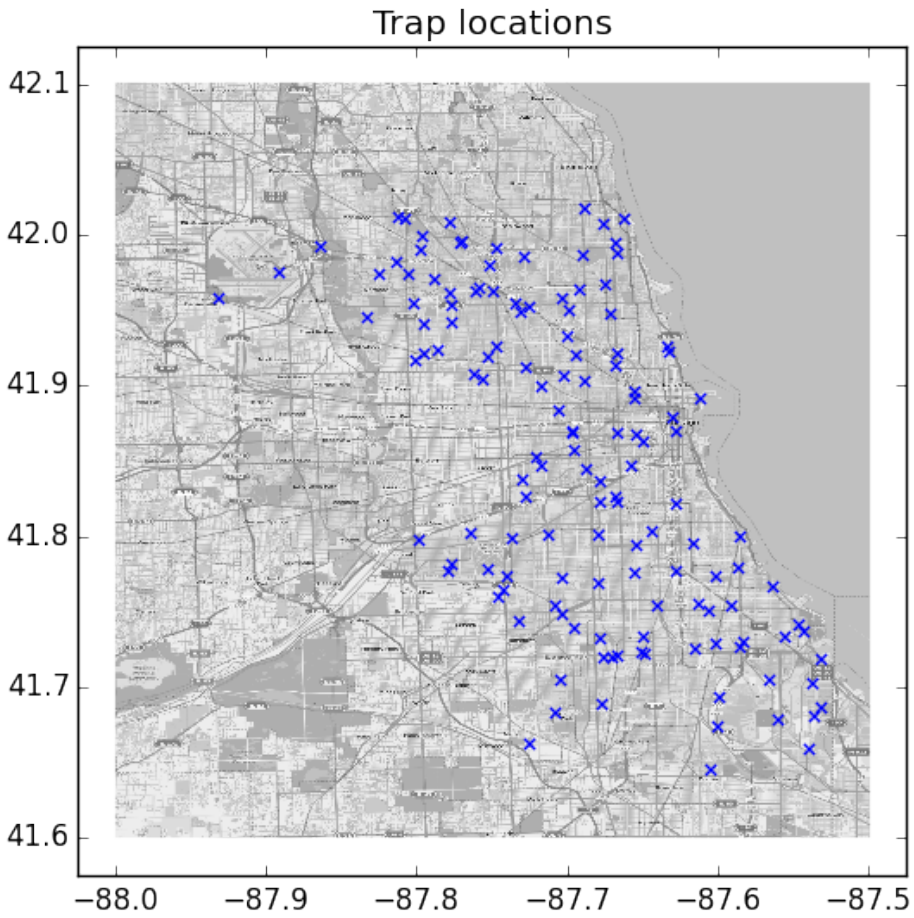
FINAL PRESENTATION

Outline

1. Problem statement
2. Data set description
3. Descriptive mining
4. Predictive mining
5. Conclusions

Problem Statement

- West Nile Virus - an incurable disease, spread by female mosquitoes
- Surveillance and control program established by the City of Chicago and the Chicago Department of Public Health
- 149 traps across the city are observed weekly from late spring through fall
- Data collected contains number of mosquitoes, species type and WNV present or not



Problem Statement

- The results influence when and where the city will spray airborne pesticides

Prediction Goal:

- Given weather, location, testing and spraying data, predict when and where different species will test positive for WNV
- Given data for the years 2007, 2009 , 2011 and 2013 predict the test-results for the years 2008, 2010, 2012 and 2014
- Help the city of Chicago to more effectively and efficiently choose spraying time and locations to prevent virus transmission

Training Set

- 10506 entries with 12 attributes (Date, Trap, Species, Latitude, Longitude, Number of Mosquitoes, WNV present, ...) for years (2007, 2009, 2011 and 2013)
- Traps were observed **irregularly**, some weeks left **unobserved**, **varying** number of traps observed
- If the number of mosquitoes exceeded 50 the rows were split

Date	Address and Street	Trap	Species	NumMosquitoes	WNV Present	Lat, Long, Block, ...
2007-09-19	3700 118th Street	T212	Culex Pipiens	14	0	
2007-09-19	3700 118th Street	T212	Culex Pipiens/Restuans	23	0	
2007-09-19	9100 West Higgins Road	T215	Culex Pipiens	3	0	
2007-09-19	9100 West Higgins Road	T215	Culex Pipiens/Restuans	50	1	
2007-09-19	9100 West Higgins Road	T215	Culex Pipiens/Restuans	43	1	

Test Set

- 116293 entries with 11 attributes (Numbers of Mosquitoes and WNV Present are hidden)
- For **each** observation date, **all** combinations of traps and species are given to prevent data leakage

Date	Address and Street	Trap	Species	Latitude, Longitude, Block, Street, ...
2008-09-15	9100 West Higgins Road	T009	Culex Erraticus	...
2008-09-15	9100 West Higgins Road	T009	Culex Pipiens/Restuans	...
2008-09-15	9100 West Higgins Road	T009	Culex Restuans	...
2008-09-15	9100 West Higgins Road	T009	Culex Pipiens	...
2008-09-15	9100 West Higgins Road	T009	Culex Salinarius	...
2008-09-15	9100 West Higgins Road	T009	Culex Territans	...
2008-09-15	9100 West Higgins Road	T009	Culex Tarsalis	...
2008-09-15	9100 West Higgins Road	T009	Culex Unspecified	...

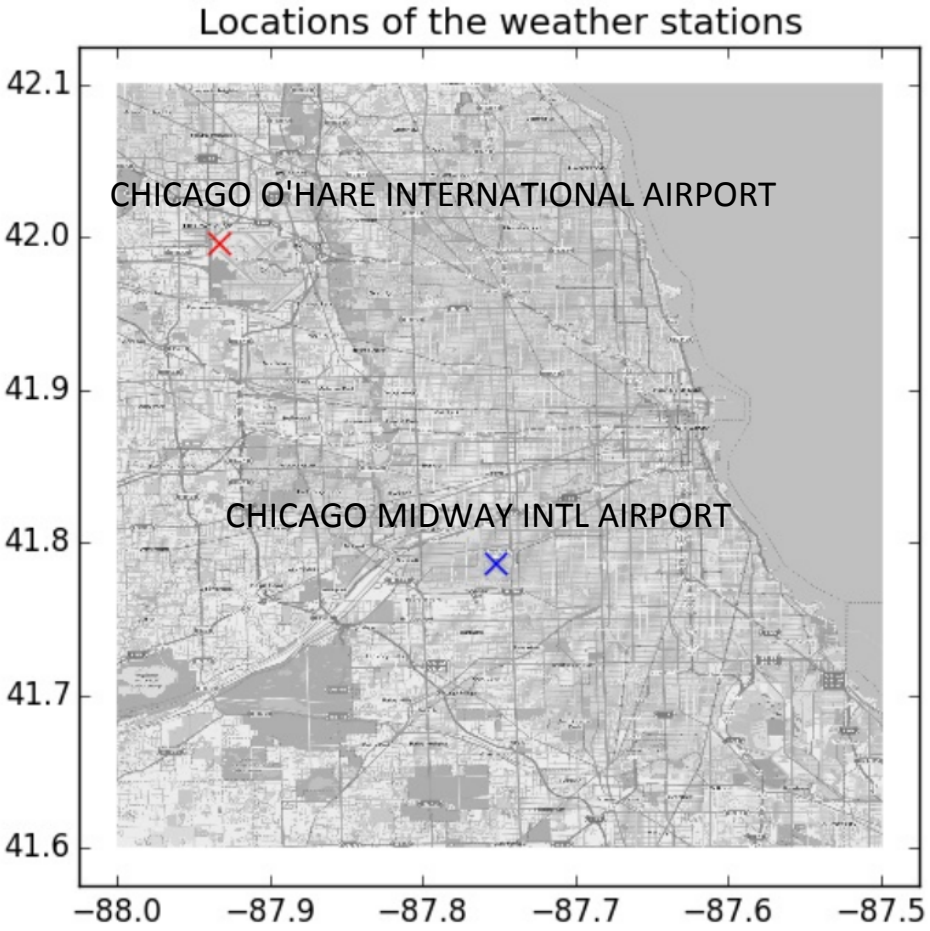
Test Set

- **But** some entries occur more often, for the same date, the same trap and the same species!

Date	Address and Street	Trap	Species	Latitude, Longitude, Block, Street, ...
2008-09-15	ORD Terminal 5, O'Hare International Airport	T009	Culex Erraticus	...
2008-09-15	**ORD Terminal 5, O'Hare International Airport**	**T009**	**Culex Pipiens/Restuans**	**...**
2008-09-15	**ORD Terminal 5, O'Hare International Airport**	**T009**	**Culex Pipiens/Restuans**	**...**
2008-09-15	**ORD Terminal 5, O'Hare International Airport**	**T009**	**Culex Pipiens/Restuans**	**...**
2008-09-15	ORD Terminal 5, O'Hare International Airport	T009	Culex Restuans	...
2008-09-15	ORD Terminal 5, O'Hare International Airport	T009	Culex Pipiens	...
2008-09-15	ORD Terminal 5, O'Hare International Airport	T009	Culex Salinarius	...
2008-09-15	ORD Terminal 5, O'Hare International Airport	T009	Culex Territans	...
2008-09-15	ORD Terminal 5, O'Hare International Airport	T009	Culex Tarsalis	...
2008-09-15	ORD Terminal 5, O'Hare International Airport	T009	Culex Unspecified	...

Weather Data

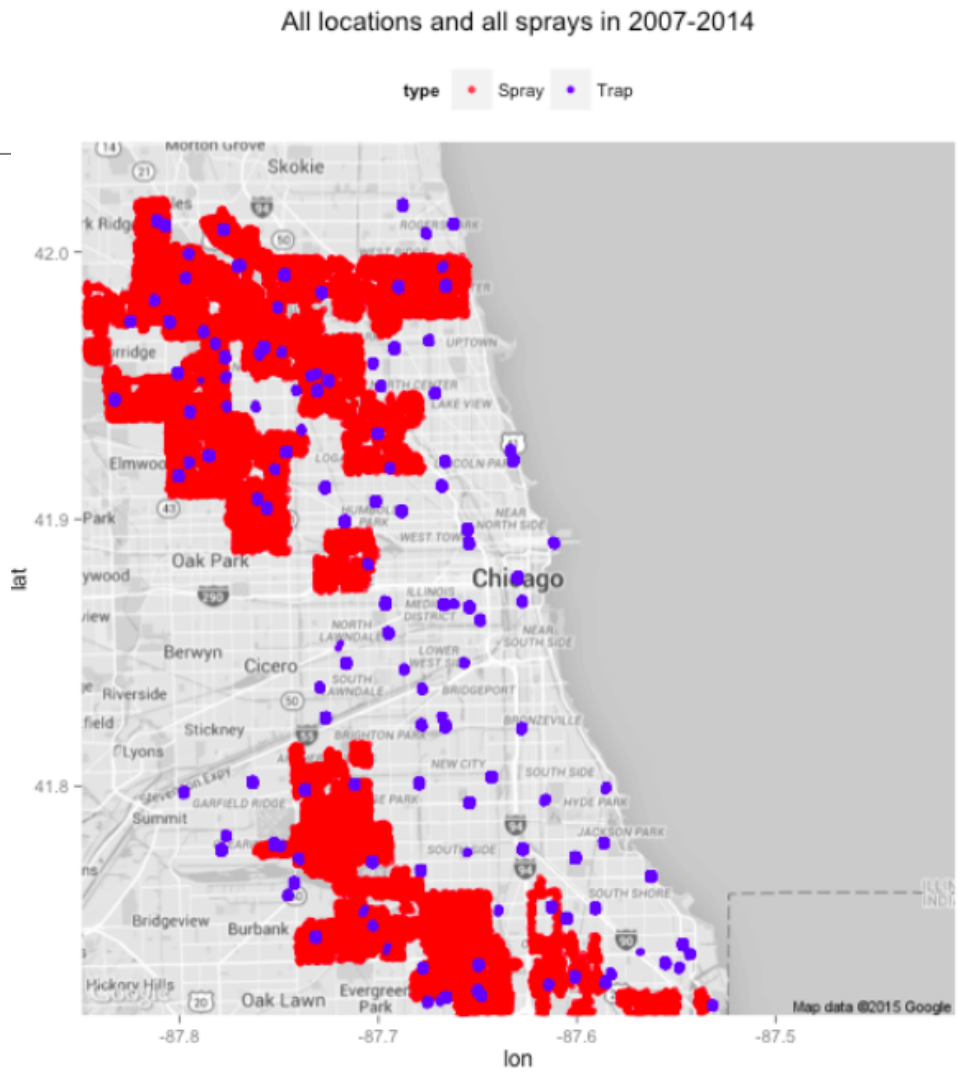
- Two weather stations
- 2944 daily entries with 20 attributes for the years 2007 to 2014
- Relevant attributes
  - Temperature (Max, Min, Avg, Heating, Cooling)
  - Humidity (Precipitation)
  - Day length (Sunrise, Sunset)
  - Wind Speed and direction



Spray Data

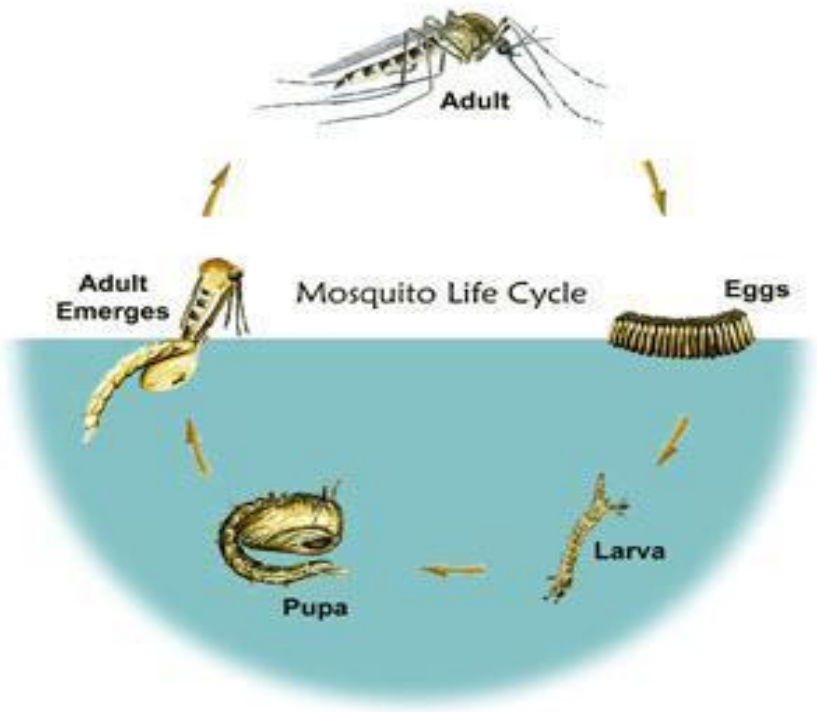
- 14836 entries and 4 attributes (Date, Time, Lat, Long)
- Spraying was done only in 2011 and 2013
- Targets male mosquitoes
- Effectiveness is discussable
- Spraying data was eliminated in the testing set

  -> data leakage



Background Research

- Mosquitoes go through **four stages** in their life cycles – egg, larva, pupa, and adult.
- WNV is primarily associated with the **Culex mosquitoes.**
- Culex mosquitoes are generally **weak fliers**
- Mosquitoes are most **active at high temperature** and become lethargic at low.



Background Research

Positive Correlation with Mosquito Abundance

- **Temperature**: highest correlation 18 days before the capture
- **Day time length**: highest correlation from 5th to 4th week before the capture
- **Precipitation** : highest correlation was found over 10 weeks before the capture

Negative Correlation with Mosquito Abundance

- **Wind speed** : negative correlation with 3 weeks aggregate
- **Humidity**: highest effect was from 15th to 2nd week before the capture.

Tools Used

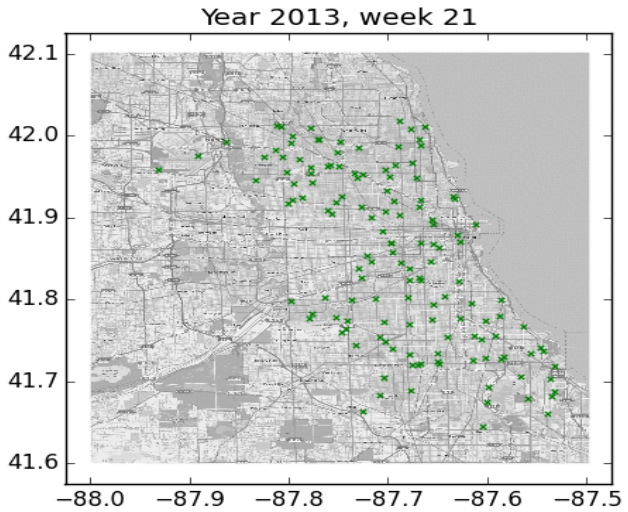
- Python
- RStudio

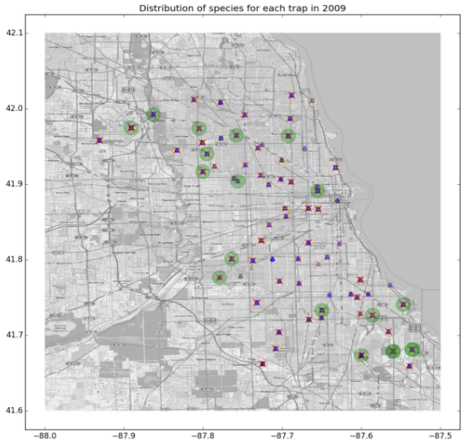
GitHub: https://github.com/shchur/data_mining_lab

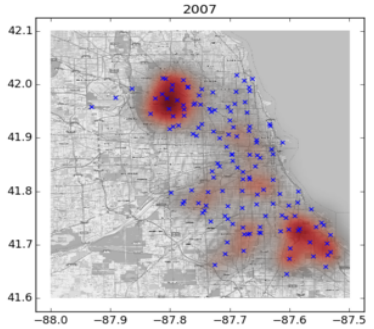
Data cleaning

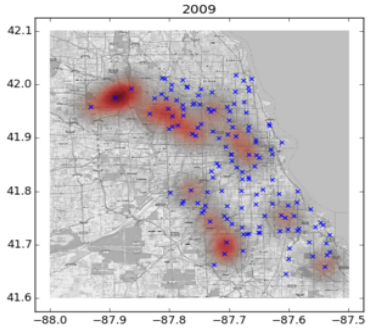
- Missing values in the weather dataset were denoted as M or T
- Missing values were replaced by the averages of the day before and the day after
- For the average temperature the missing values were replaced by the average of the minimum and maximum daily temperature
- Attributes Depth, Water1 and SnowFall only contained zeroes or M's, therefore were removed

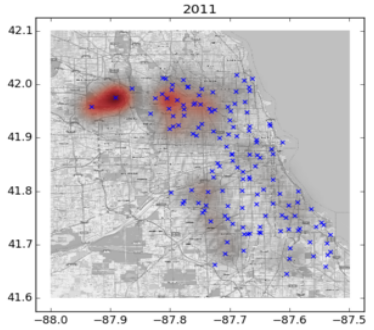
Descriptive Mining - Data Visualization

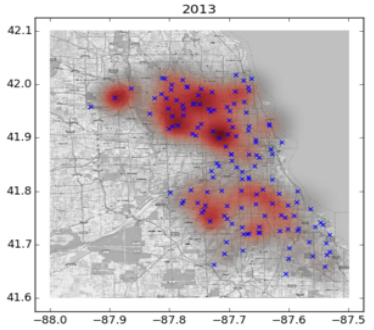






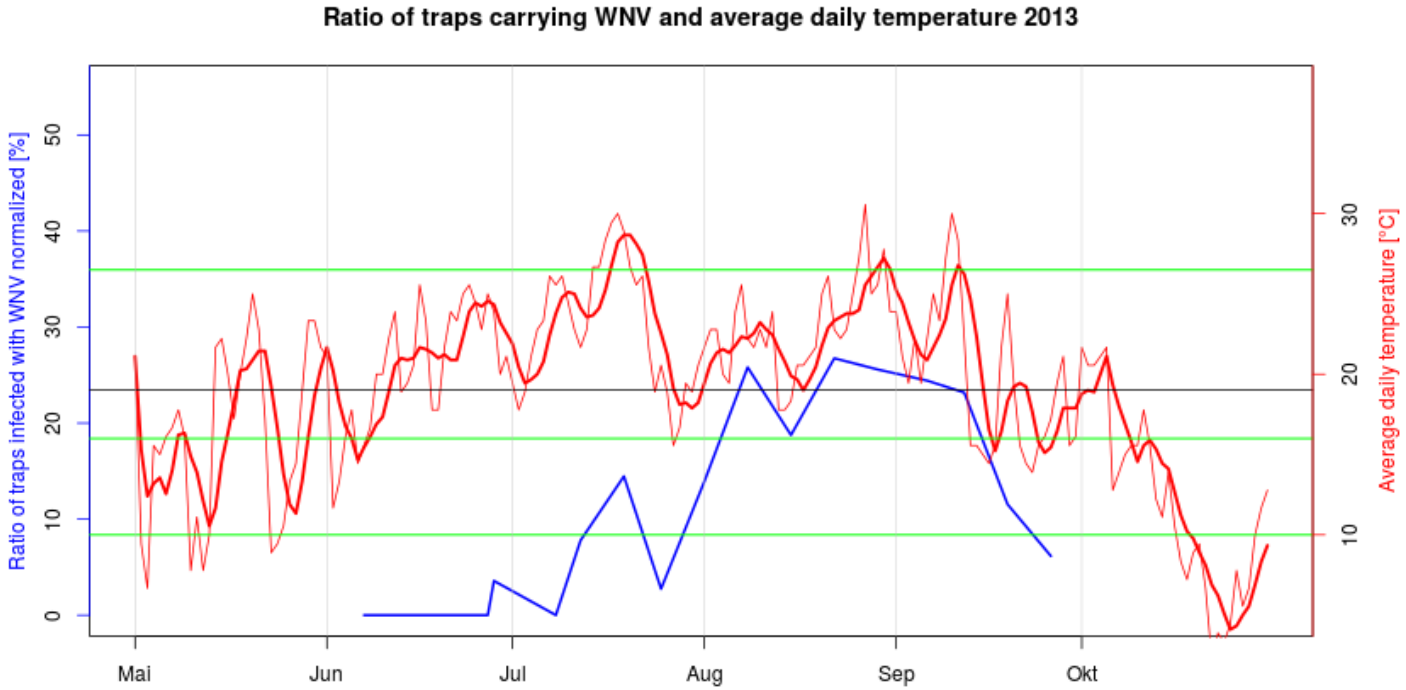






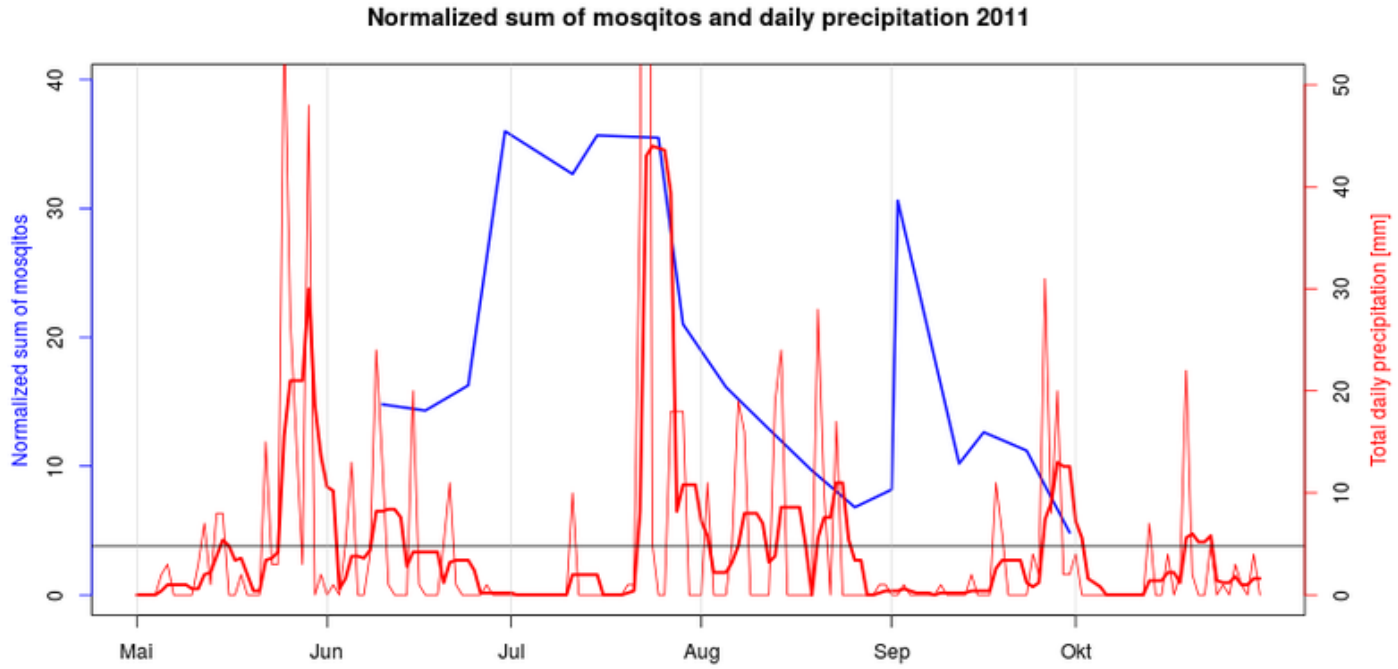
Descriptive Mining - Feature selection

Correlation: Positive.



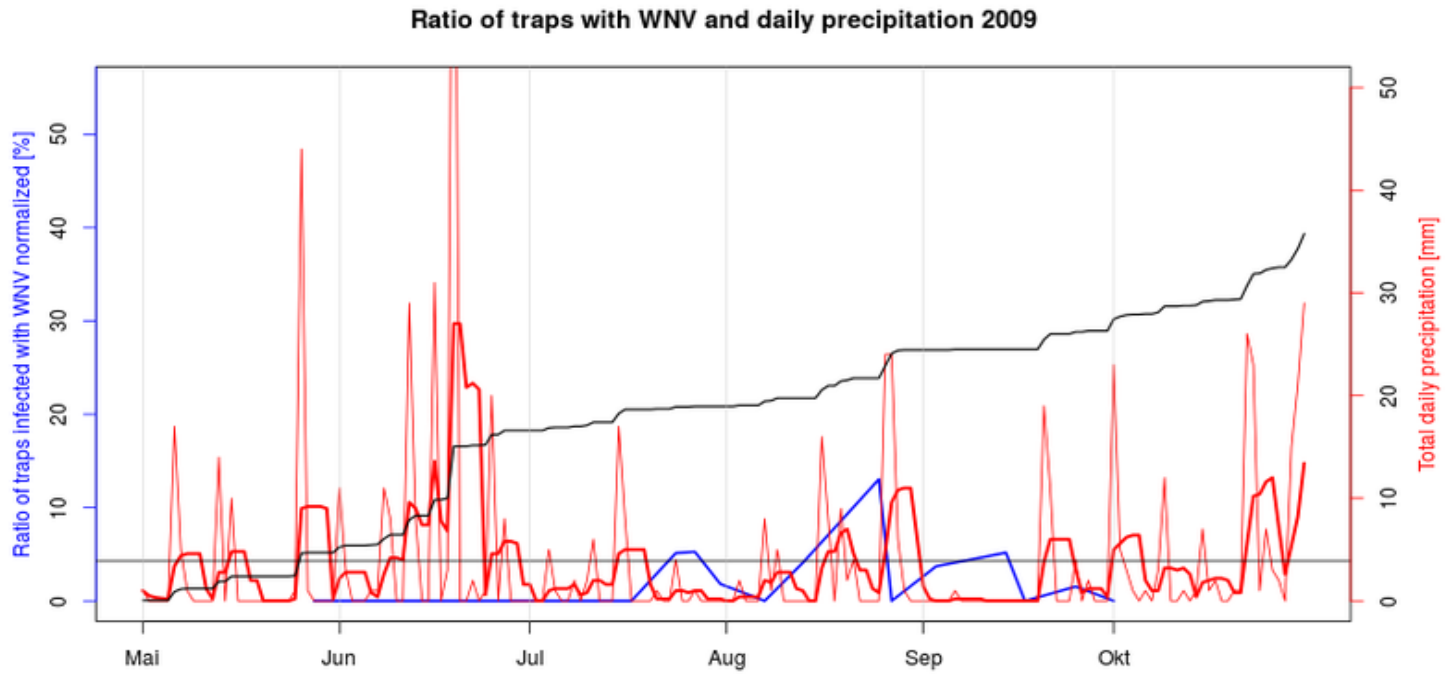
Descriptive Mining - Feature selection

Correlation: Uncorrelated.



Descriptive Mining - Feature selection

Correlation: Medium.



Descriptive Mining - Feature selection

- **The following table summarizes the correlation of WNV occurrence w.r.t weather attributes.**
- **Legend:** 

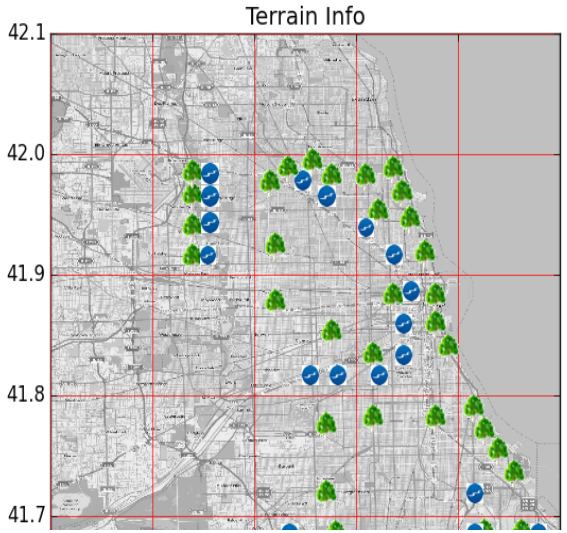
Years	Normalized sum of mosquitoes and avg. Temperature	Ratio of traps carrying WNV and avg. Temperature	Normalized sum of mosquitoes and avg. Precipitation	Ratio of traps carrying WNV and avg. Precipitation	Trap infection rate and average daily wind speed
2007					
2009					
2011					
2013					

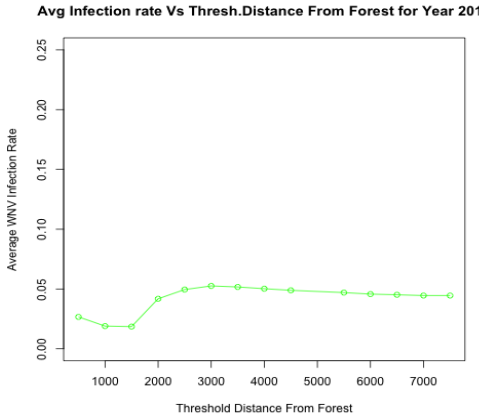
Descriptive Mining - Feature selection

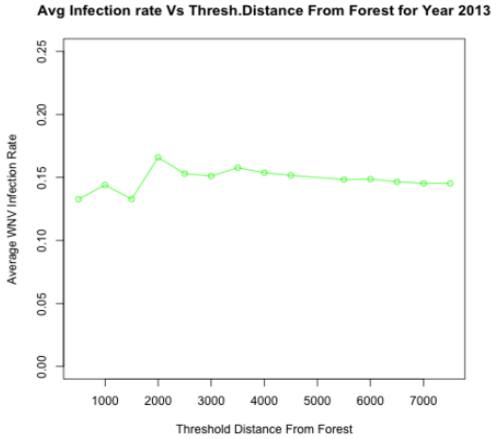
- Avg. Temperature vs Normalized sum = Correlation found only in 2011 and 2013.
- Avg. Temperature vs Ratio of traps = Correlation found only in 2007 and 2013.
- Avg. Precipitation vs Normalized sum = No Correlation found for any year.
- Avg. Precipitation vs Ratio of traps = Correlation found only in 2007.
- Avg. Wind Speed vs Trap infection rate = Correlation found in 2007, 2011, 2013.

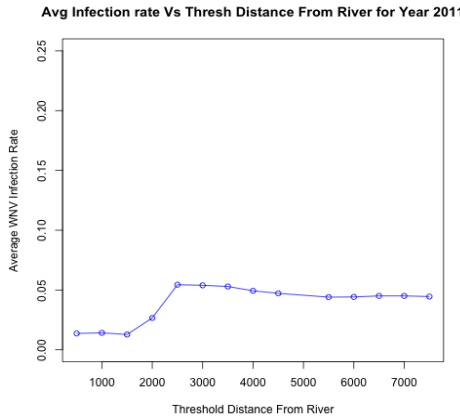
Descriptive Mining - Feature selection

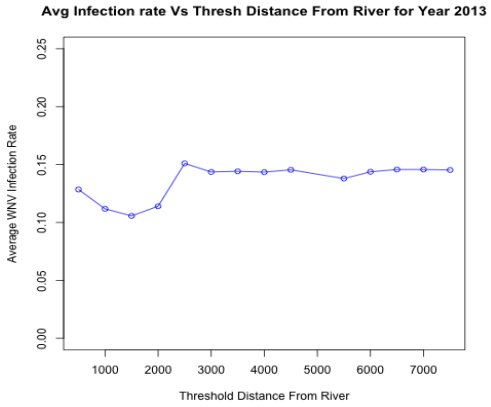
Terrain Analysis of Chicago









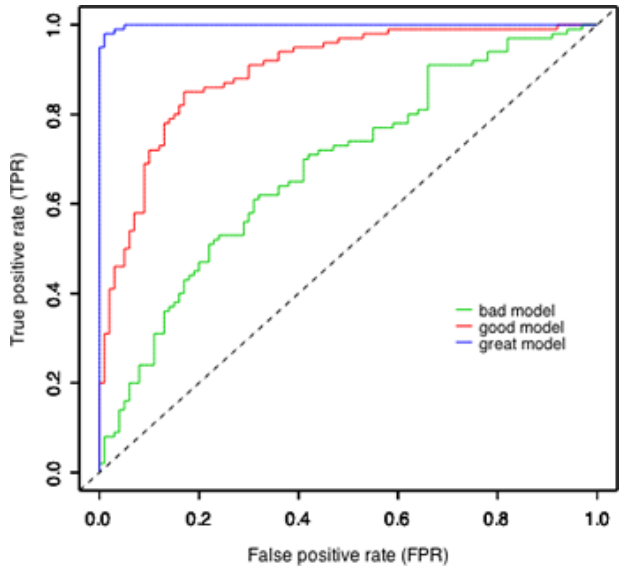


Descriptive Mining - Conclusions

- Visualization gave us a better understanding of the overall problem
- There was irregular data in the training set that could not be compensated for
- No patterns persistent across all the years have been observed
- Relationships between different features too complex and are hard to analyze visually

Predictive Mining

- Predictive goal - predict probability of WNV presence in mosquitoes collected in 2008, 2010, 2012, 2014
- Prediction quality is evaluated as the Area under the Receiver Operating Characteristic Curve
- The only factor influencing AUC is how well the classes are separated insensible to skewness or unnormalized probabilities
- Random guessing gives ~ 50% AUC



Predictive Mining

- The test set is split into 2 parts - public (30%) and private (70%)
- The number of submissions to Kaggle is limited, therefore need to evaluate the prediction locally (without knowing the correct labels for the test set)
- Solution - 4-fold cross-validation - leave out one of the years (2007/2009/2011/2013), train on the others, and predict for the left-out year
- Predictions only submitted to the leaderboards if CV score is high enough
- The prediction quality for obtained predictions will be shown as

Cross validation AUC	Total test set AUC 0.3 * Public + 0.7 * Private	Private leaderboard position

Initial attempt

- Simply concatenate weather and training/test data for each row
- Encode nominal features as integers (species, address, address accuracy)
- Encode date as *month* and *day*
- Use random forest classifier (decision trees)

Cross validation AUC	Total test set AUC	Private leaderboard position
0.712	0.681	967

Trying different classifiers

- We also tried kNN, SVM and Logistic regression
- All of the algorithms provided results significantly worse to those of Random forests
  - kNN is working on the assumption that all features are equally important
  - SVM requires a lengthy process of hyperparameter tuning, and feature engineering
  - Logistic regression depends on a careful choice of features, which is hard to do.
- Used boosting classifiers like Adaboost, but it did not provide any improvement either
- CRF not possible - Continuous features, Infeasible to construct feature functions.

Training with new features

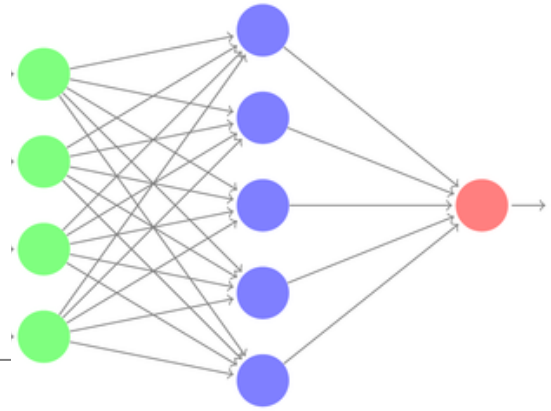
- New feature added that expresses the deviation of the night time temperature from the average night time or minimum temperature. [4]
- The squared difference of the daily maximum temperature and the daily average temperature from a temperature threshold of 25°C. [5]
- The predictions obtained by adding these features were less than the initial attempt so these were dropped.

Training with new features

- Truncate the features that were suggested to be irrelevant by [1], [2]
- Encode mosquito species using “one hot” approach
- Transform date to number of days since 1/06 of the given year
- New derived features - *CumulativeHeat, CumulativePrecip, TavgOver14Days*
- Again, use random forest classifier

Cross validation AUC	Total test set AUC	Private leaderboard position
0.751	0.7695	385

Neural Networks



- Best results but don’t have sound theoretical basis
- Configurations
  - Varying Number of Hidden Layers / Neurons
  - Autoencoders and dropout
- Deep? Neural Networks
- Autoencoders can be used for any ML algorithm

Layers	No of Neurons	AUC	Rank
1	[400]	0.7736	325
2	[400 400]	0.7791	287
3	[400 200 400]	0.7834	258
1	[1000]	0.7833	259
4	[400 200 100 50]	0.7438	471
2	[400 200]	0.7852	205
81 > 27	Autoencoder	0.7921	232
81 > 243	Autoencoder	0.7933	204

“Unfair” prediction 1 - Utilizing the data leakage

- Number of mosquitoes not explicitly given in the test set
- Records are split, if number of mosquitoes exceeds 50
- Therefore, it’s possible to infer number of mosquitoes for the test set
- test_reduced contains the number of rows in the test set for each day
-

```
P(WNV | day, ...) = (test_reduced[day] - min(test_reduced)) /

(max(test_reduced) - min(test_reduced))
```

Cross validation AUC	Total test set AUC	Private leaderboard position
---	0.744	487

Combining with the last best prediction

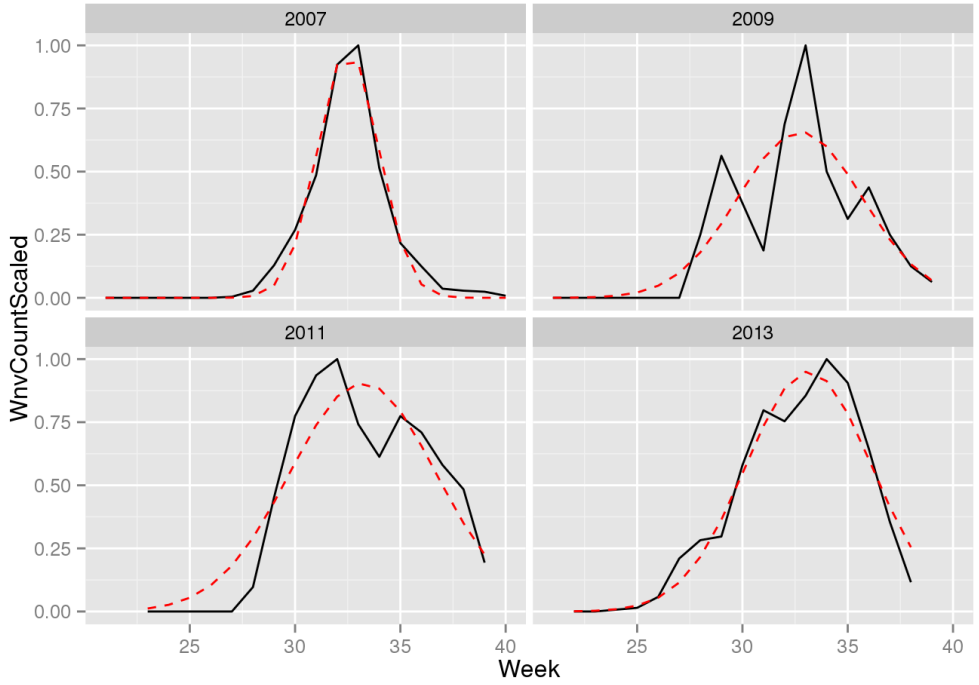
- For each test file row take the average between the last best prediction with random forest and derived features and the one using row counting
- Set the probabilities for species, that never carried the virus to 0

Cross validation AUC	Total test set AUC	Private leaderboard position
---	0.795	180

“Unfair” prediction 2
Curve fitting

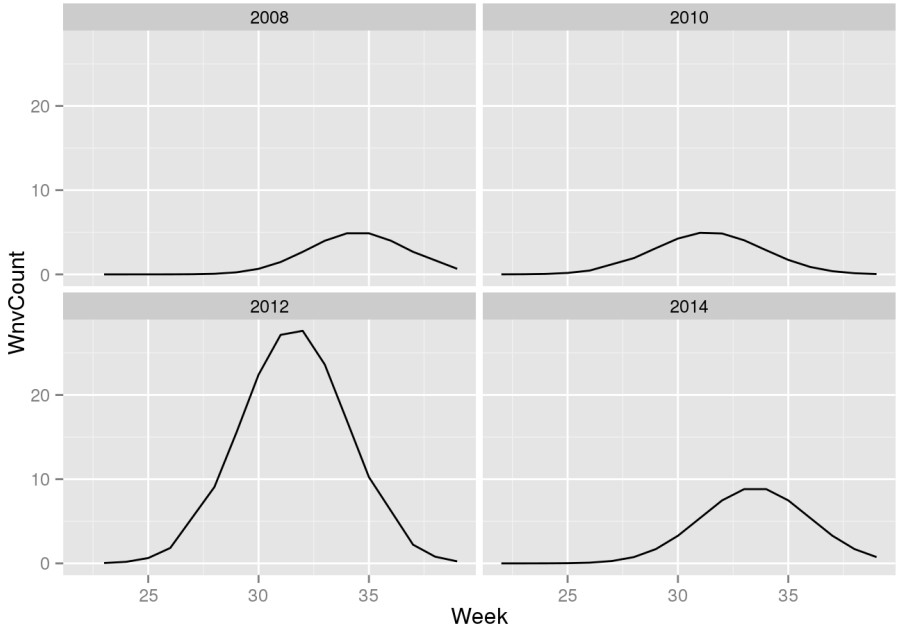
- The population of WNV across different years can be modelled well with a Gaussian [3]

$$P = height * e^{-\frac{(week-center)^2}{width}}$$



“Unfair” prediction 2
Curve fitting

- Submit the prediction files only with data for a chosen year
- Tune the parameters *height, center, width* trying to maximize the AUC
- Resulting predictions are overfitted and can’t be applied to make predictions in real world setting
- However, this method allows to get a great score and leaderboard position



Total test set AUC	Private leaderboard position
0.835	25

Predictive Mining - Conclusion

- Usage of derived features is crucial to obtaining good predictions
- Various classifiers were used, but most require “hand engineering” of features and intensive hyperparameter tuning for good performance
- The only classifiers that do not need this (Random Forest and Neural Networks) “hand engineering” don’t have sound theoretical explanation, however provide the best results
- It is possible to obtain very good results by using “unfair” methods, which, however, have no practical use

Kaggle Competitions for Praktikum

- Merits
  - Sense of Competition
  - The task/metric is very well defined
  - Progress is easy to see/compare
- Demerits
  - Focus shifts on prediction early on, without going through descriptive mining
  - Usually, data is already cleaned, which rarely happens in the real world

References

[1] - http://www.parasitesandvectors.com/content/6/1/129

[2] - http://www.parasitesandvectors.com/content/3/1/19

[3] - https://www.kaggle.com/oconnoda/predict-west-nile-virus/population-model

[4] - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2858365/

[5] - http://www.sacbee.com/news/local/health-and-medicine/article23397396.html

Questions