Analysis of West Nile Virus in Chicago

Bryan Combs, Sam Lundberg, and Molly Emmett

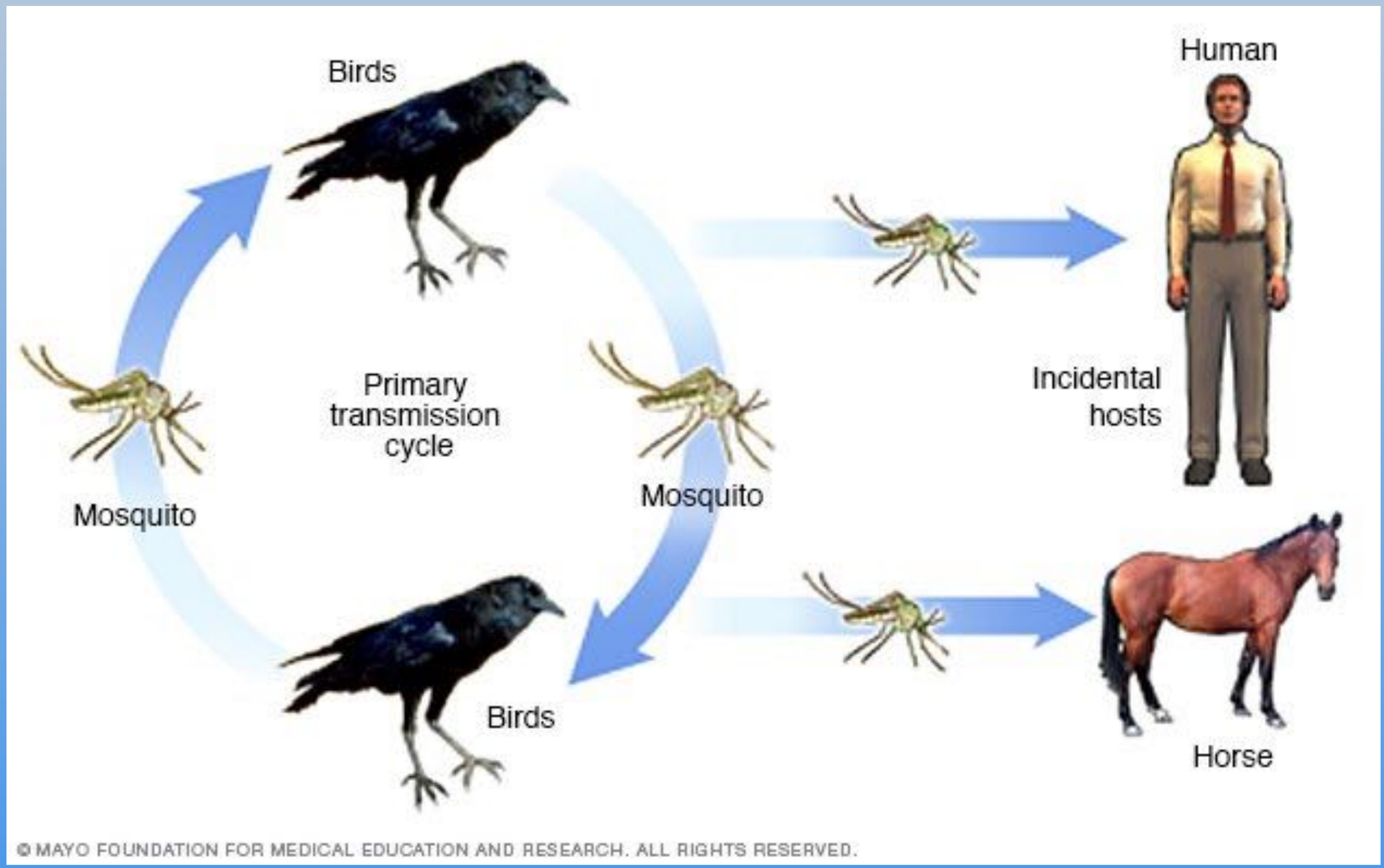
Problem:
Predicting West Nile Virus

Can we predict West Nile Virus such that the city of Chicago can maximize the impact of its integrated pest management system?

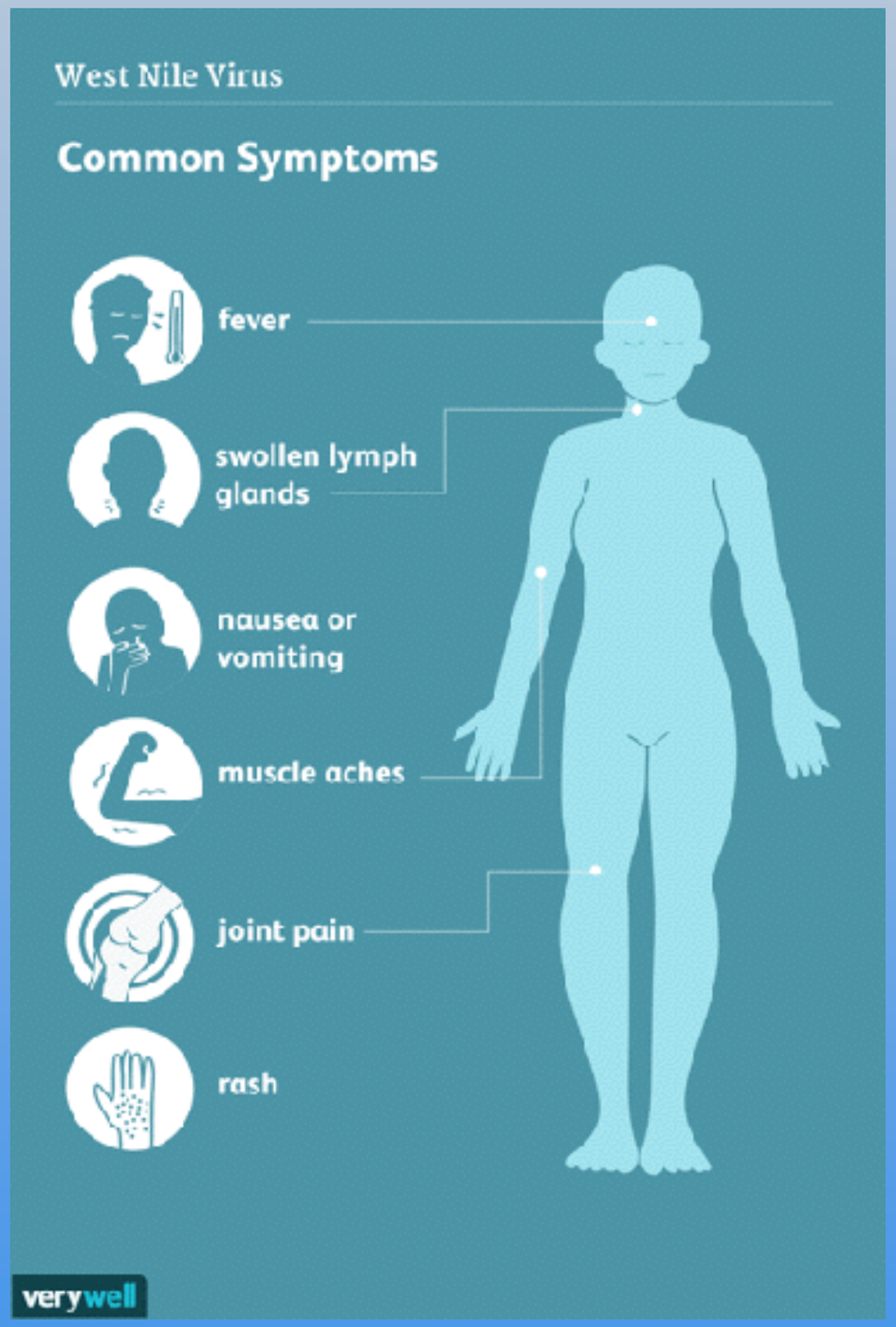
Presentation Overview

- Background: West Nile Virus
- Data Analysis Process and Decisions
- Cost Benefit Consideration
- Outcomes and Recommendations

West Nile Transmission

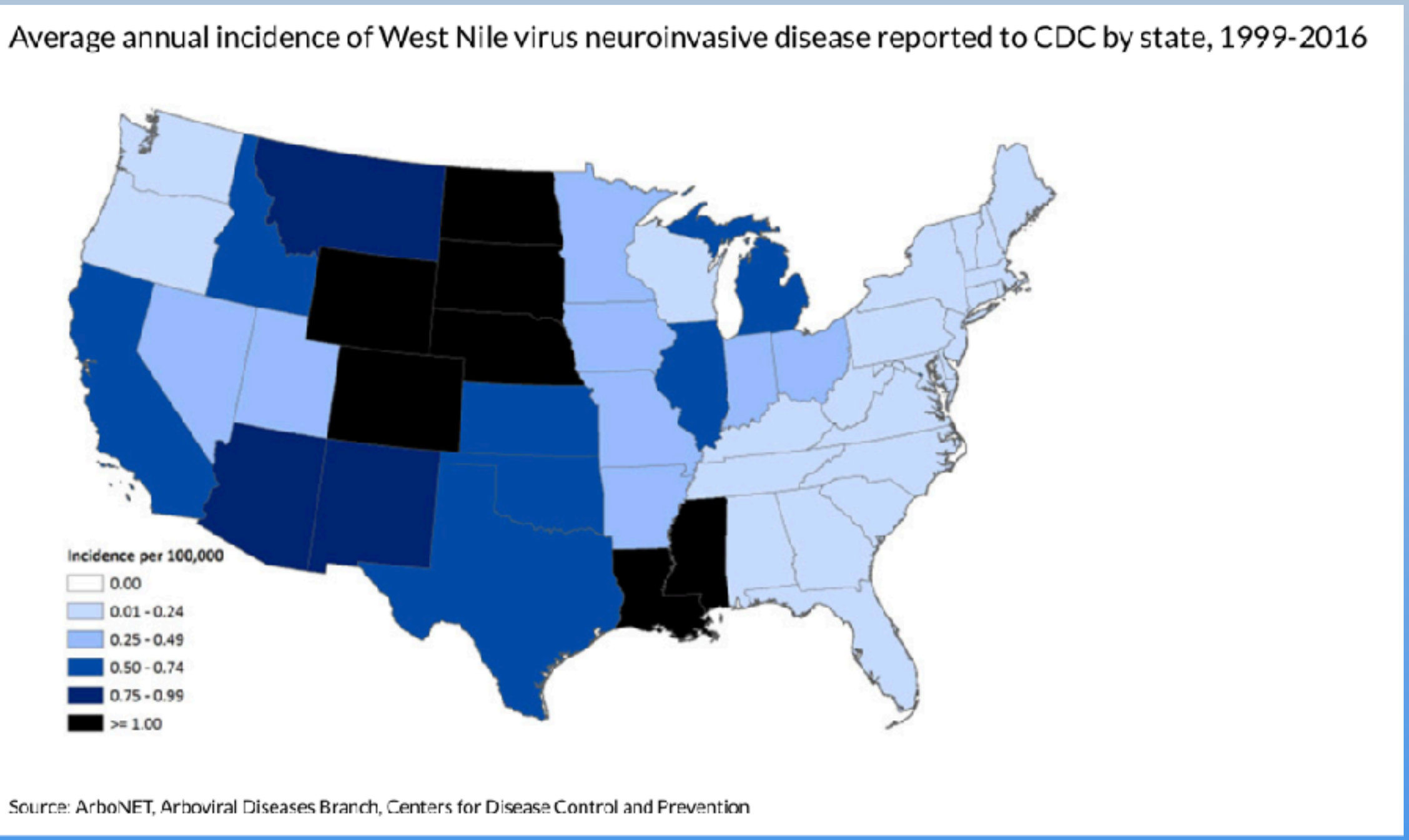


West Nile Impacts



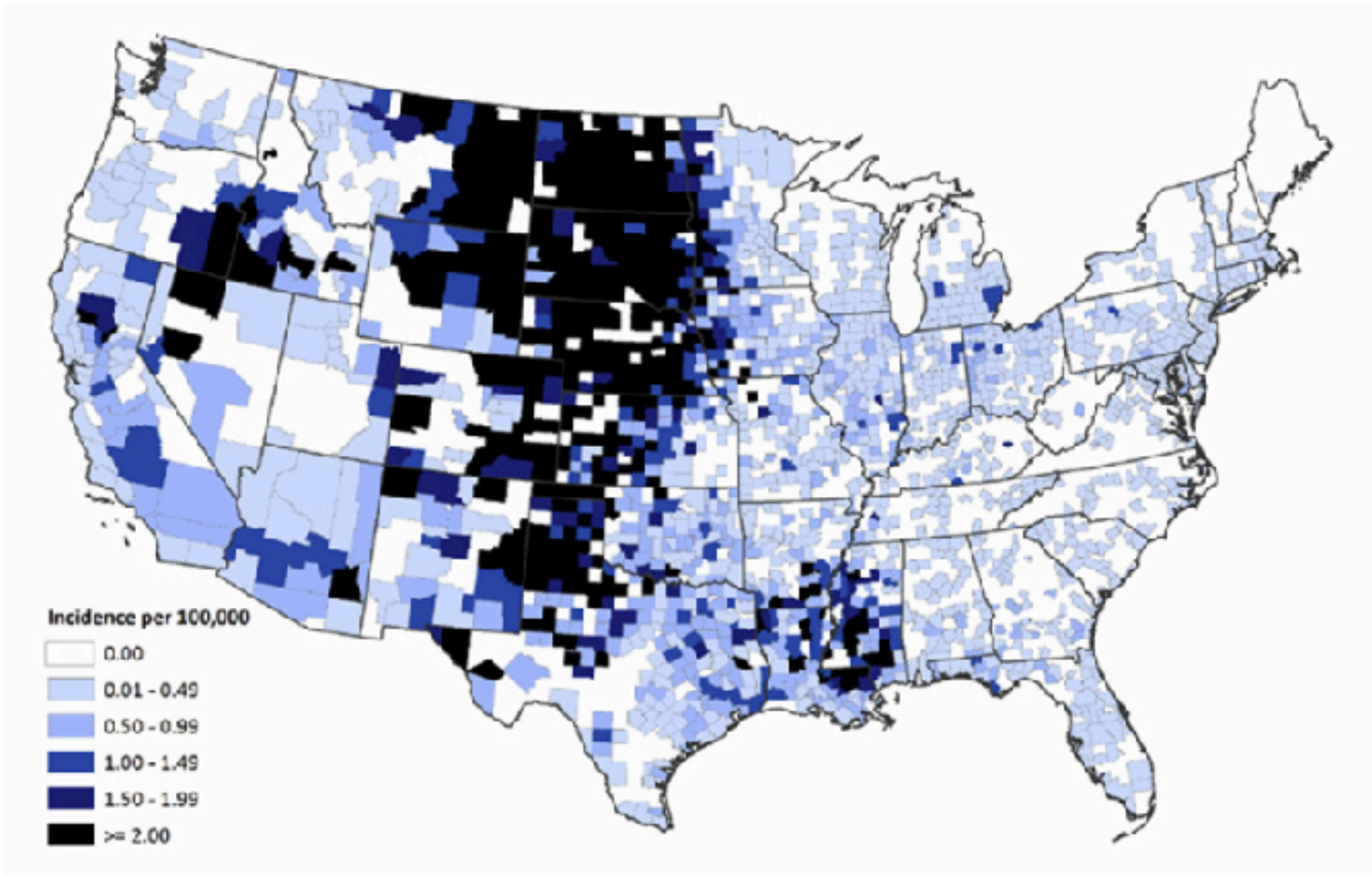
A severe case of West Nile can result in paralysis or death.

West Nile Dispersion by State



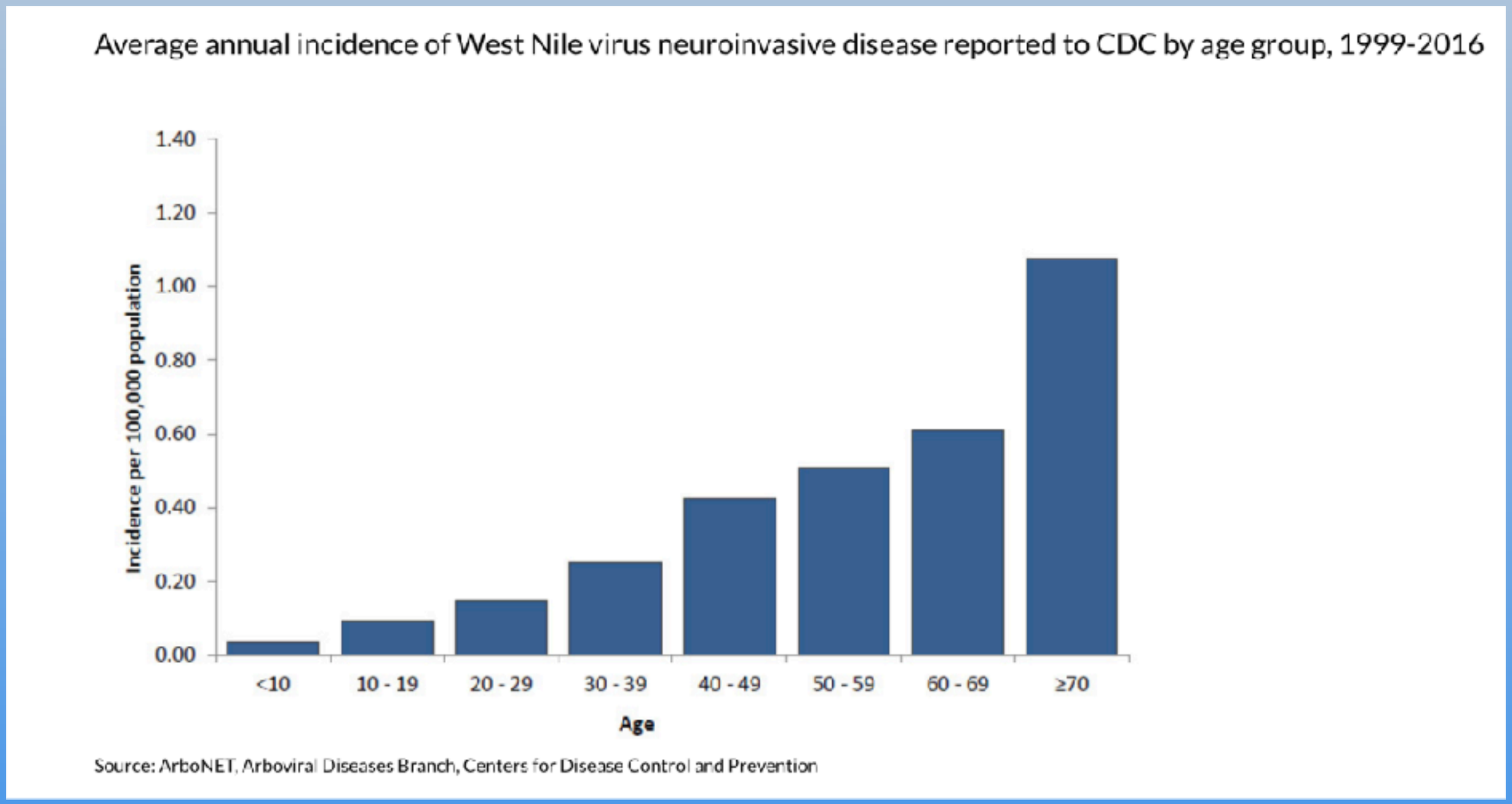
West Nile Dispersion by County

Average annual incidence of West Nile virus neuroinvasive disease reported to CDC by county, 1999-2016

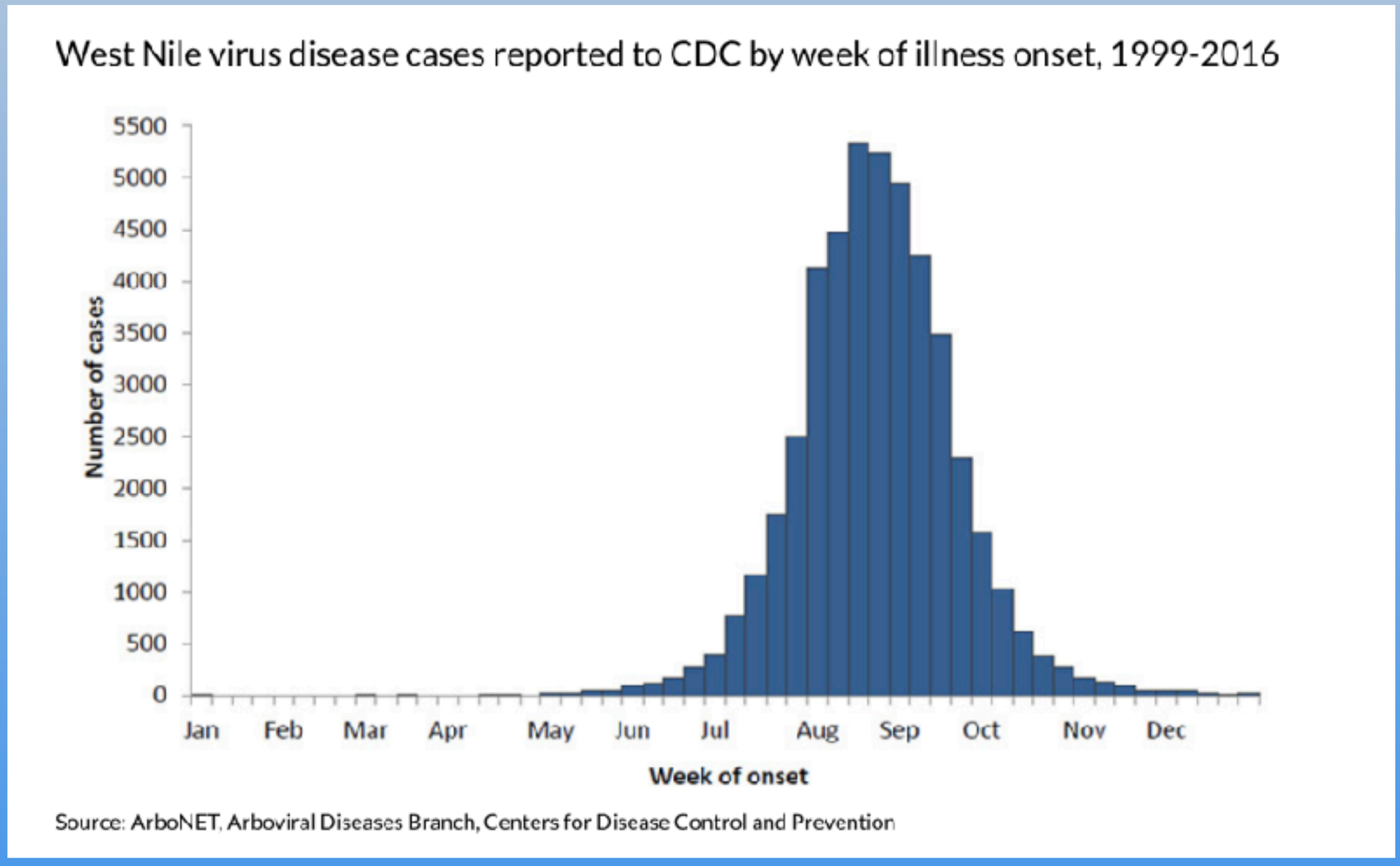


Source: ArboNET, Arboviral Diseases Branch, Centers for Disease Control and Prevention

West Nile Dispersion by Age



West Nile Dispersion by Week



Data Analysis Process

We iterated through the following major subsections of analysis:

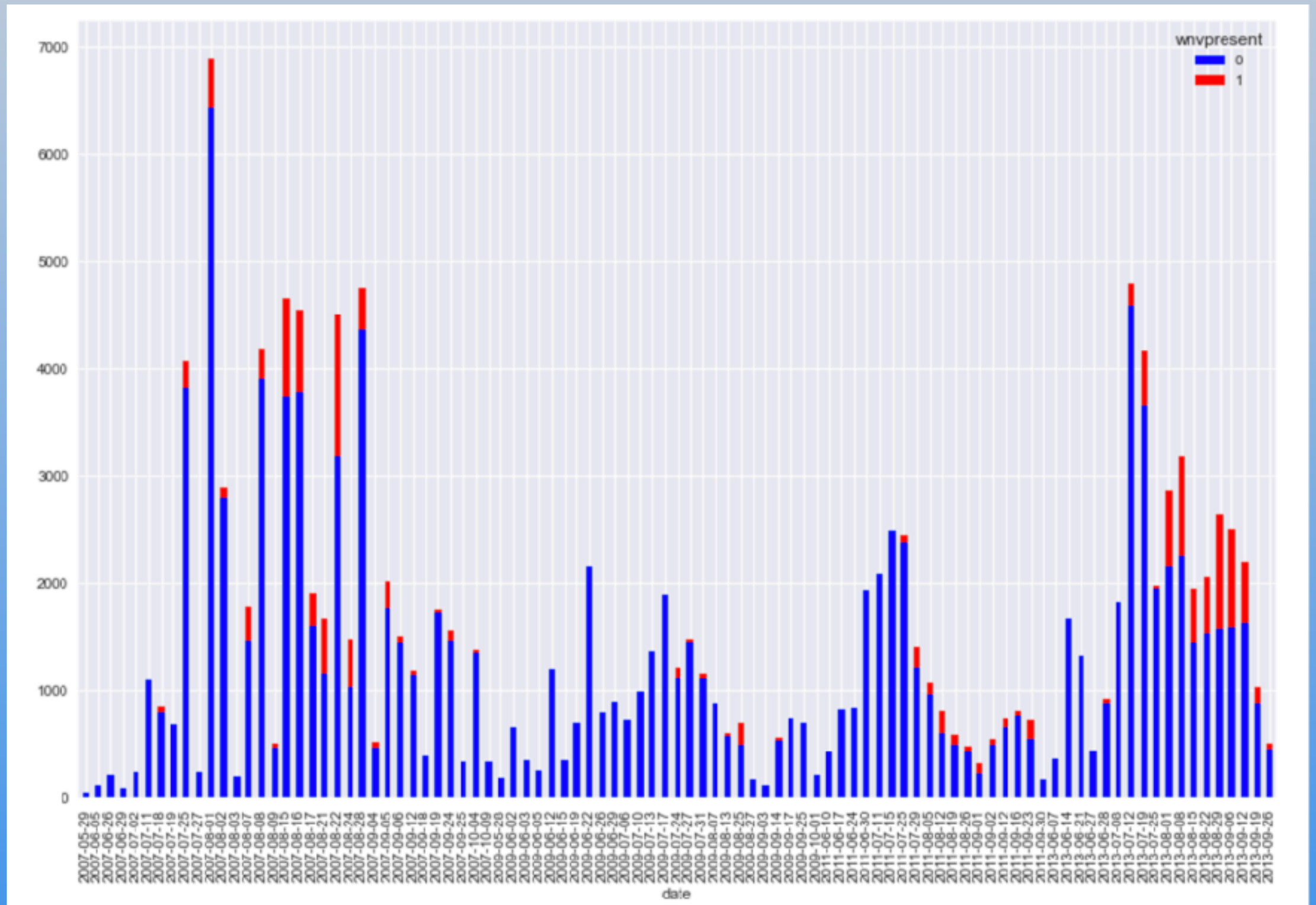
- Exploratory Data Analysis
- Feature Engineering
- Modeling

Exploratory Data Analysis

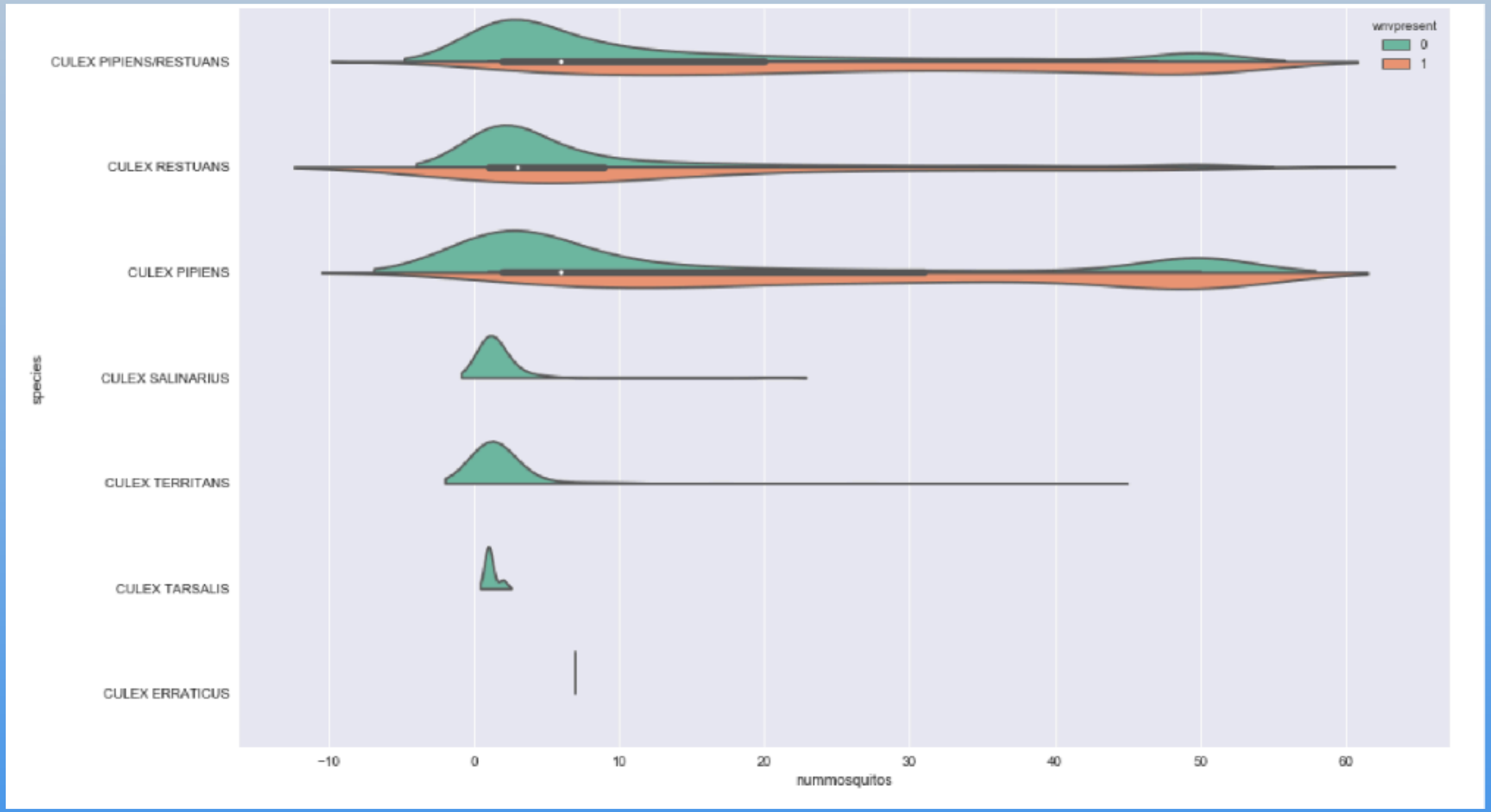
Missing Data:

- Weather
  - Replaced missing values M (missing) and T (trace amounts) with -1.
  - This allows the iterative model to treat everything with a missing or trace value the same.

Exploratory Data Analysis



Exploratory Data Analysis



Feature Engineering

Distance

- Filtering by latitude, longitude, and year, we created a function that calculates:
  - 1) minimum distance and
  - 2) average distance between mosquitoes that tested positive for WNV vs. those that did not.

Feature Engineering

Time

- We mapped temperatures over a fourteen day period to incorporate temporal causation that did not occur on the day of measurements.

Feature Engineering

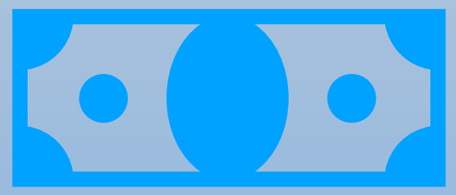
Clustering

We clustered our data by longitude and latitude in an attempt to identify possible groupings of West Nile Virus, but it didn't return significant results.

Sea Level

We created a binary feature separated by the lowest standard deviation of sea level, but this also didn't return significant results.

Considering Costs & Benefits



Monetary

Temporal

Medical

Environmental

Considering Costs & Benefits

Monetary and Temporal

Planned Spray

9.9 million people in Chicago metropolitan area

$3,288,527 in mosquito spraying in 2016

$.33/person

Emergency Spray

450,381 people in Sacramento

$2.98 million spraying in 2005

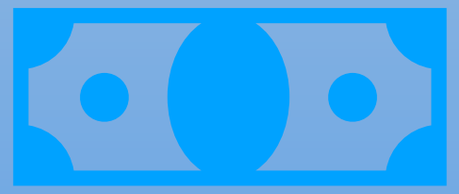
$6.62/person

Considering Costs & Benefits

Environmental

- Clarke: largest mosquito control company in the United States
- Sprays Anvil at .62 ounces/acre
- External criticism: small amounts can kill fish and bees
- Harm to humans unknown

Considering Costs & Benefits

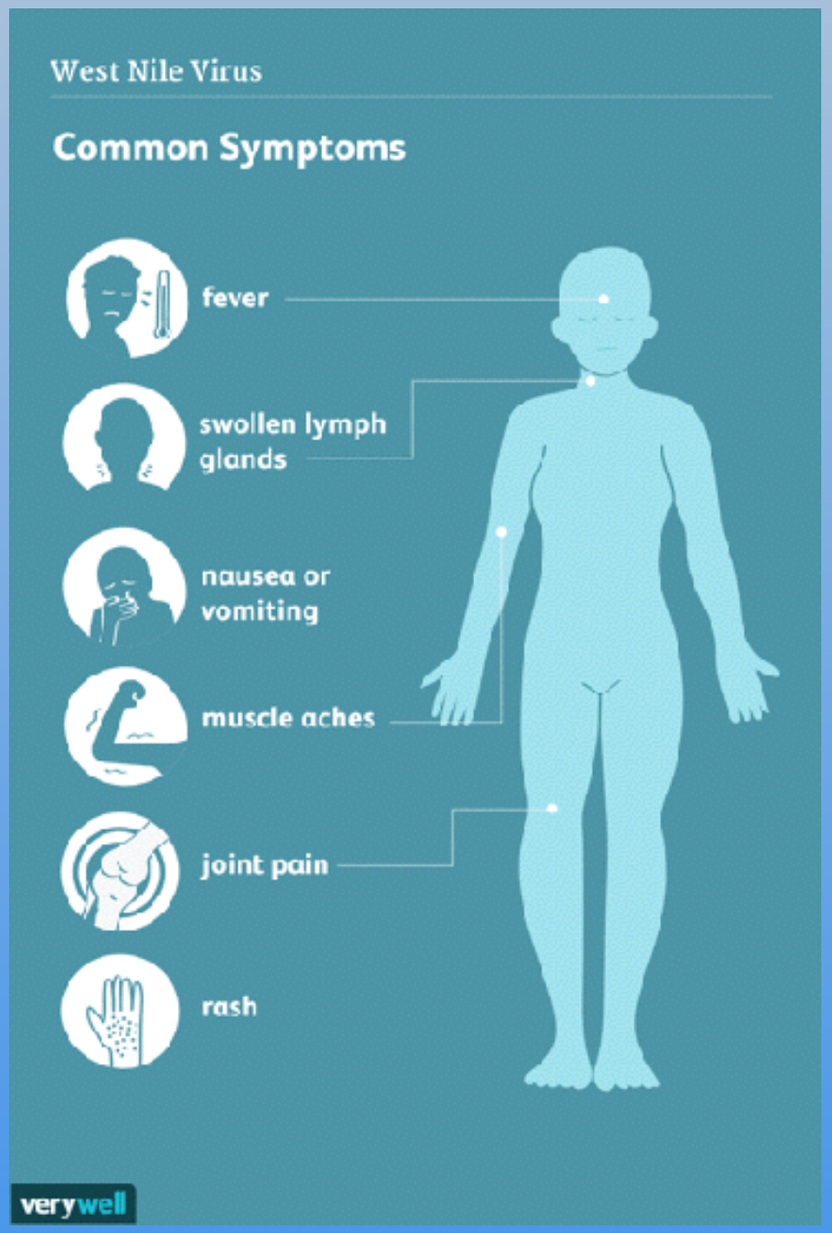


Considering Costs & Benefits

Medical

Senior population.
The older you are, the more risk you have of developing illness.

Health challenged population.
Cancer, diabetes, hypertension, kidney disease, and receiving an organ transplant increase your risk of developing illness.



Modeling

The Question of Sensitivity vs. Accuracy

Modeling

Model Selection

We chose **Adaboost**.

In our EDA, we saw non-normality in the distribution of our data.

We had large categorical variables (e.g. trap) and didn't want to dummy that many variables.

Since we're working with unbalanced classes, we chose an iterative model that would progressively train each weak learner.

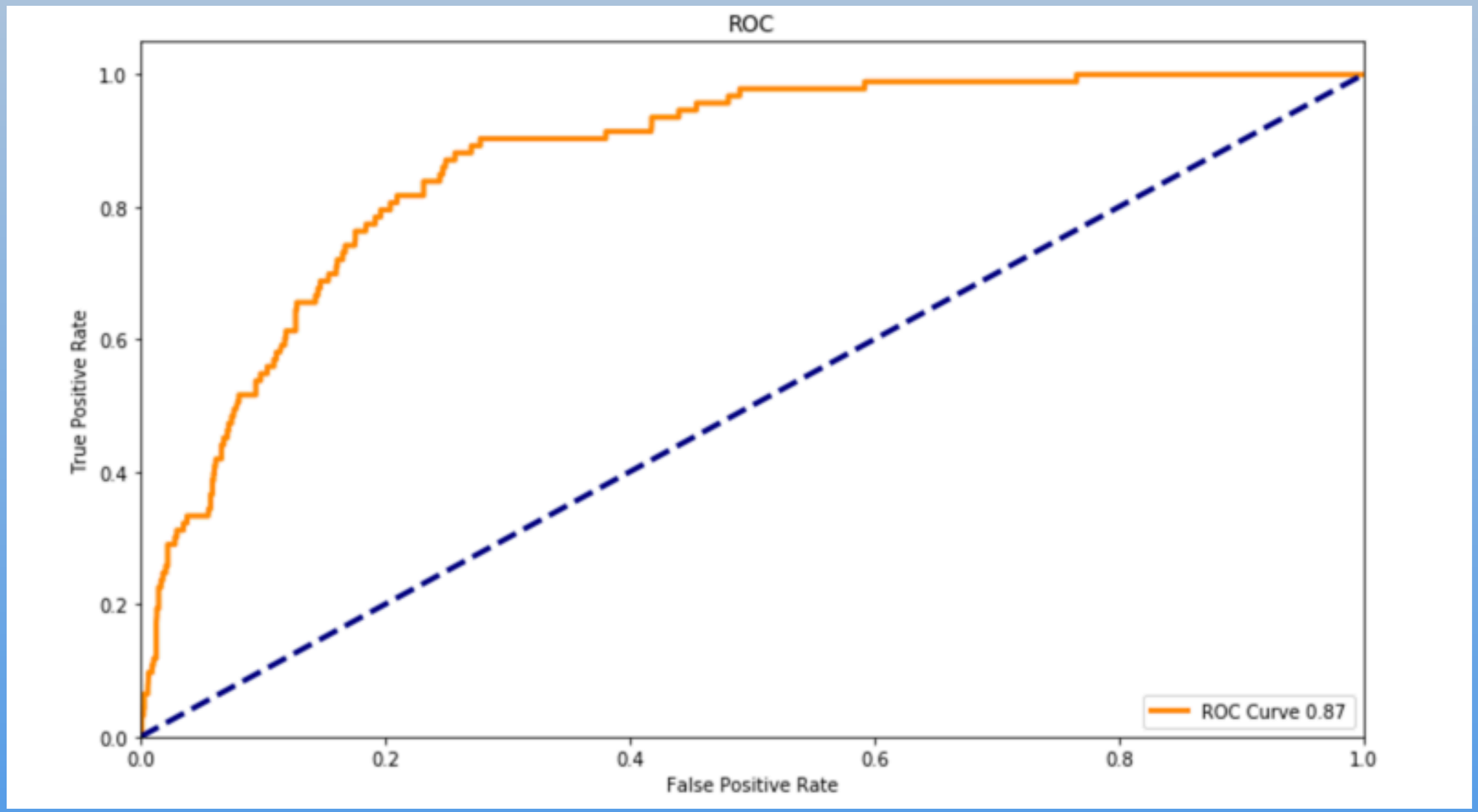
Modeling

Threshold Particularity

We created a function that iterates through thresholds in order to determine the optimal confusion matrix balance for our model.

We visualized our Roc Curve so we can adjust our classification threshold to provide an optimal balance between recall and specificity.

Outcomes


ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve 0.87

Outcomes

Optimized Model Train and Test Score on Sensitivity

Train Cross Validation:
.804

Test:
.796

Outcomes

Sensitivity

True Positives / (True Positives + False Negatives)

Baseline for Validation Set
93/(93 + 2,026) = .04

Model Outcome
74/(74 + 19) = .80

Recommendations

1. Use predictions to identify high volume West Nile areas for spraying when trap data is not available.

2. Use risk factors identified by prominent features to predict when spraying season would be more prevalent.

3. Use predictions of West Nile to inform medical providers of presence of West Nile in the area.

Recommendations

Spray Target

$8, 075/sq mile: Carol Stream

$**19,334.45/sq mile**: Wheaton Mosquito Abatement District

Recommendations

Consider the following predictors of WNV when forming the integrated pest management system.

- Sunrise
- Distance_5
- Month x Distance x Temperature from Station 1
- Latitude
- Longitude
- Distance x Species Culex Pipiens/Restuans

Next Steps

- Pursue misclassifications.
- Push for uniformity in station data collection.
- Record neuroinvasive impact of virus in areas with vulnerable populations, such as senior living centers, hospitals, and other renderings of health centers.
- Evaluate if spraying decreases occurrences in subsequent seasons.