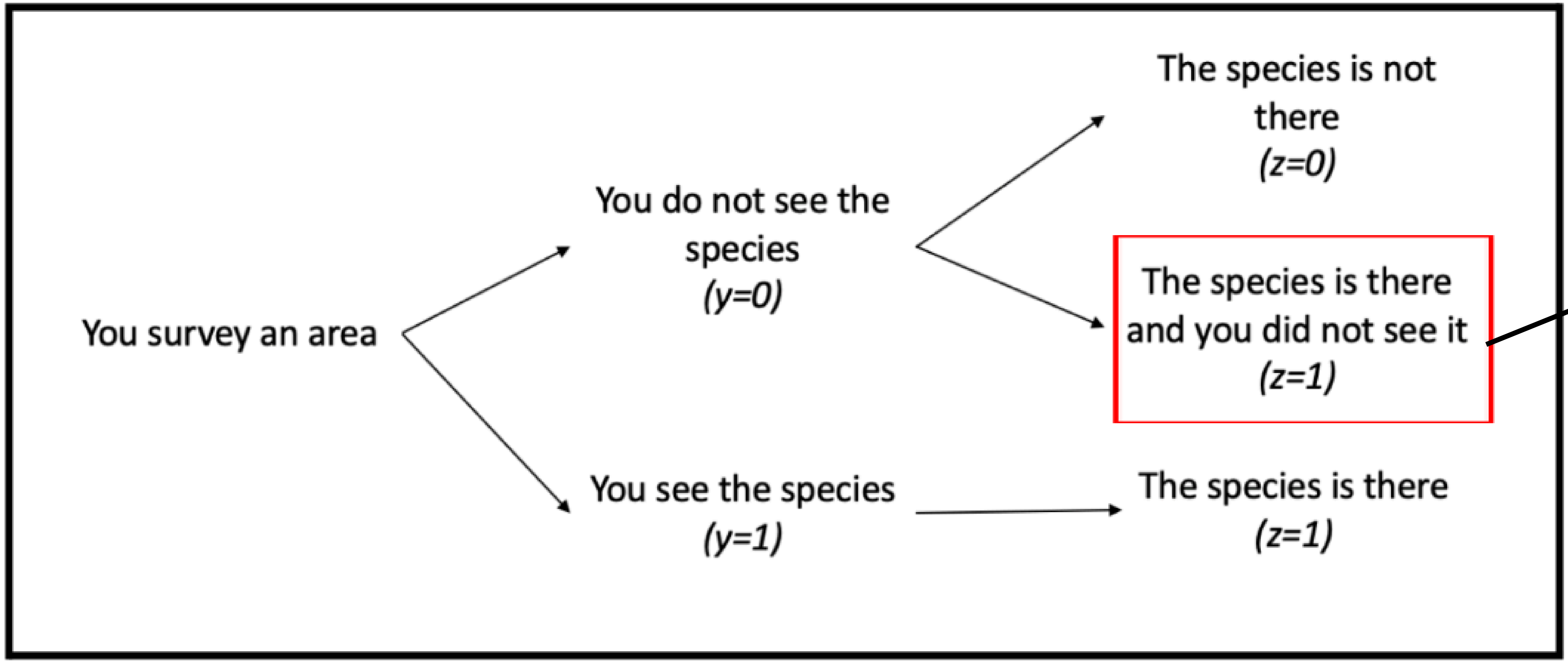
# OCCUPANCY MODELING OF BIRDS IN THE AMAZON RAINFOREST

by Jeremy Borden, Chelsey Hunts, Dawit Mengesha, Yusup Amat, Sriram Raghunath

# BACKGROUND




Observational flow chart relevant for occupancy modeling. The red box contains the possibility of imperfect detection. Variables y and z correspond to detection and occupancy, respectively. Taken from [kevintshoemaker.github.io](kevintshoemaker.github.io), produced by Morgan Byrne, James Golden.

**OCCUPANCY MODELS TRY TO ACCOMMODATE IMPERFECT DETECTION**

# PROJECT GOALS

Has climate change or forest loss affected bird populations in the Amazonas region of Brazil over the last 12 years?

Tested this for two species:
Generalist species: Black vulture (Coragyps atratus)
Forest specialist: Screaming piha (Lipaugus vociferans)

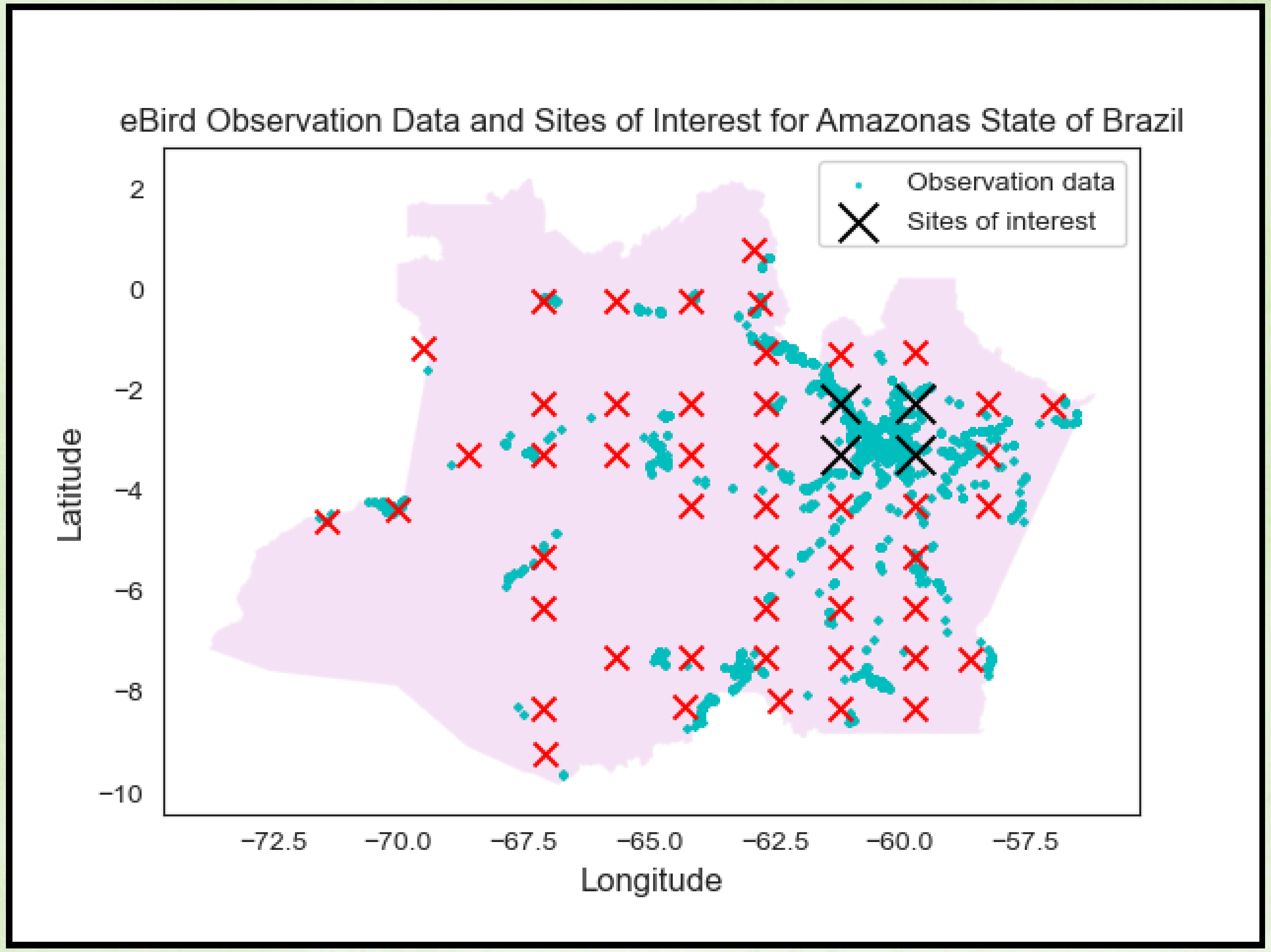
Test performance of various occupancy models

# DATA COLLECTION

EBird Data:

2012-2021

https://ebird.org/home




Climate Data (Worldclim):

Temperature and precipitation

https://www.worldclim.org/data/worldclim21.html

Tree cover and tree cover loss Data:

Enhanced Vegetation Index (EVI), Tree cover loss, Tree cover loss by fires

https://lpdaac.usgs.gov/products/mod13q1v061/; https://www.globalforestwatch.org/dashboards/country/BRA/

# DATA COLLECTION

- Occupancy covariates: Precipitation, Temperature, EVI, Tree Cover Loss, Tree Cover Loss by Fires
- Detection covariates: Year, Day of Year, Time of Day, Number of Observers, Effort Hours, Effort distance

# ML PREPROCESSING

Interested in the effects of environmental covariates over time.

But still want to model occupancy as with supervised learning (classification) techniques.

Solution: **Create new features shifted by a time step**

| Target | Time Step | Feature (t) | Feature (t-1) |
|---|---|---|---|
| y4 | 4 | | |
| y3 | 3 | | |
| y2 | 2 | | |
| y1 | 1 | | |
| y0 | 0 | | |

Use value of Feature at Time Step t-1 to predict Target at Time Step t

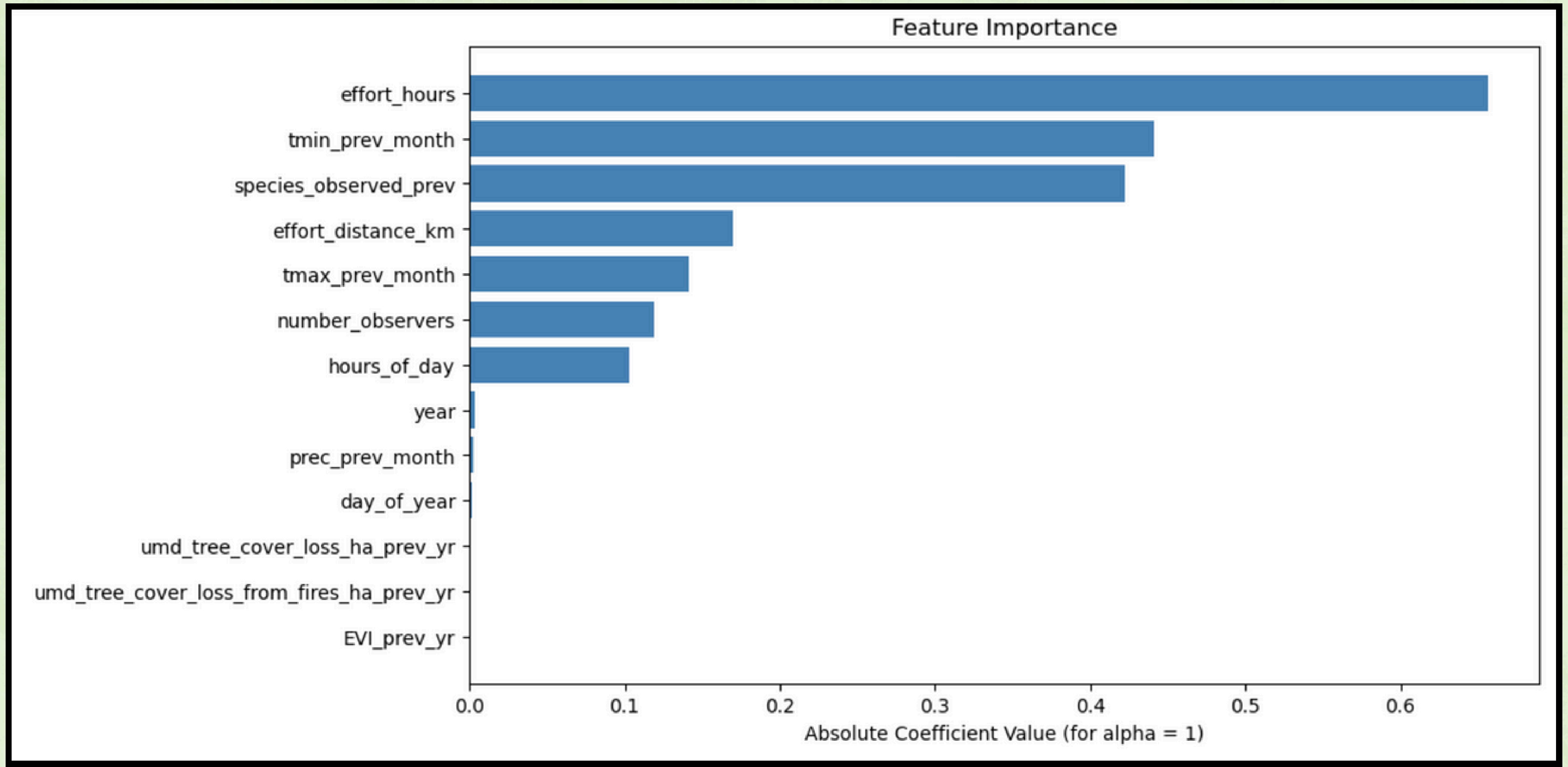
Our shifted features:

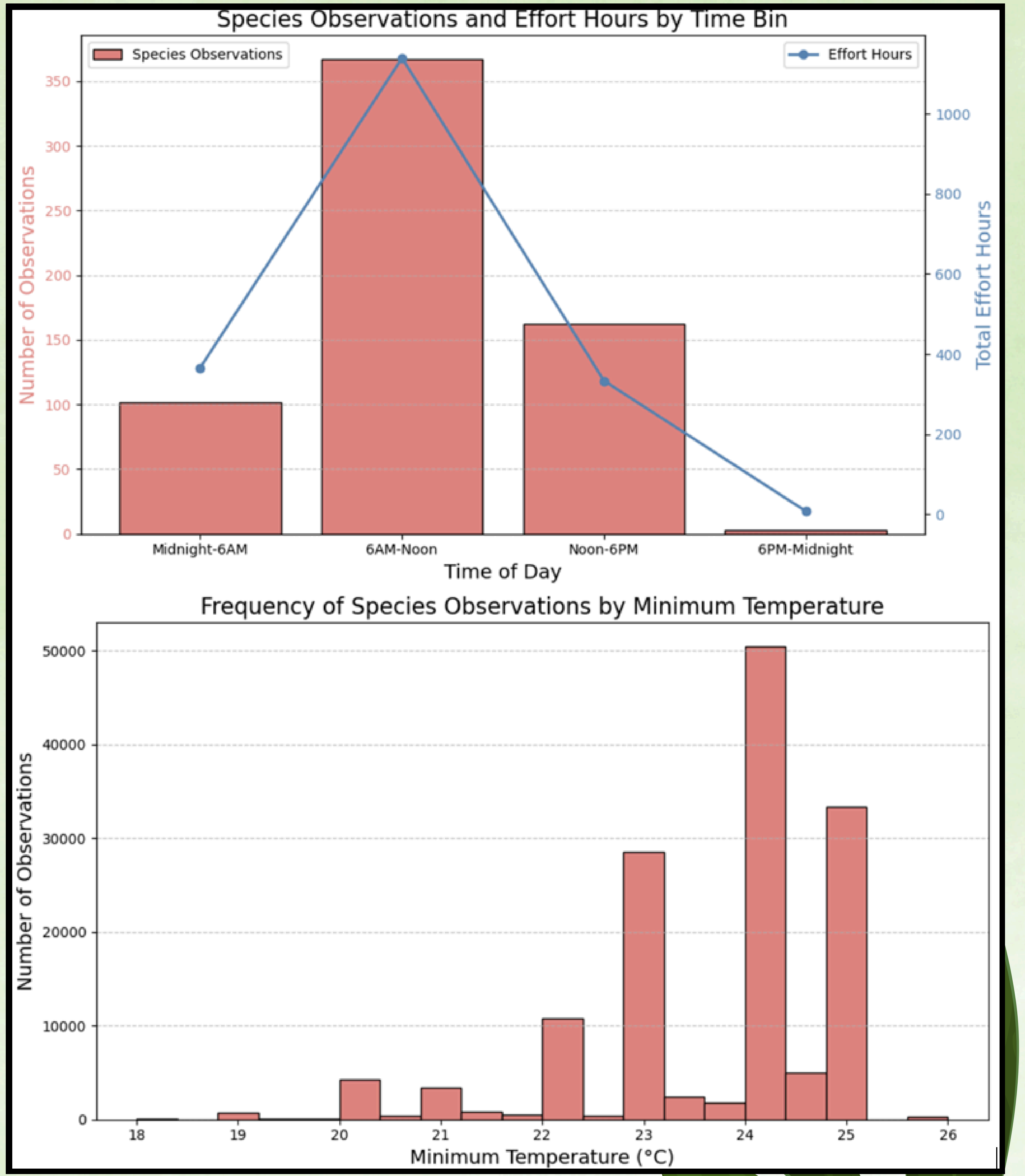
*EVI, tree cover loss, precipitation, temperature, and previous occupancy*

# ML MODELING APPROACH 1

## 1. BINARY LOGISTIC REGRESSION

Implemented with L1 regularization for feature selection

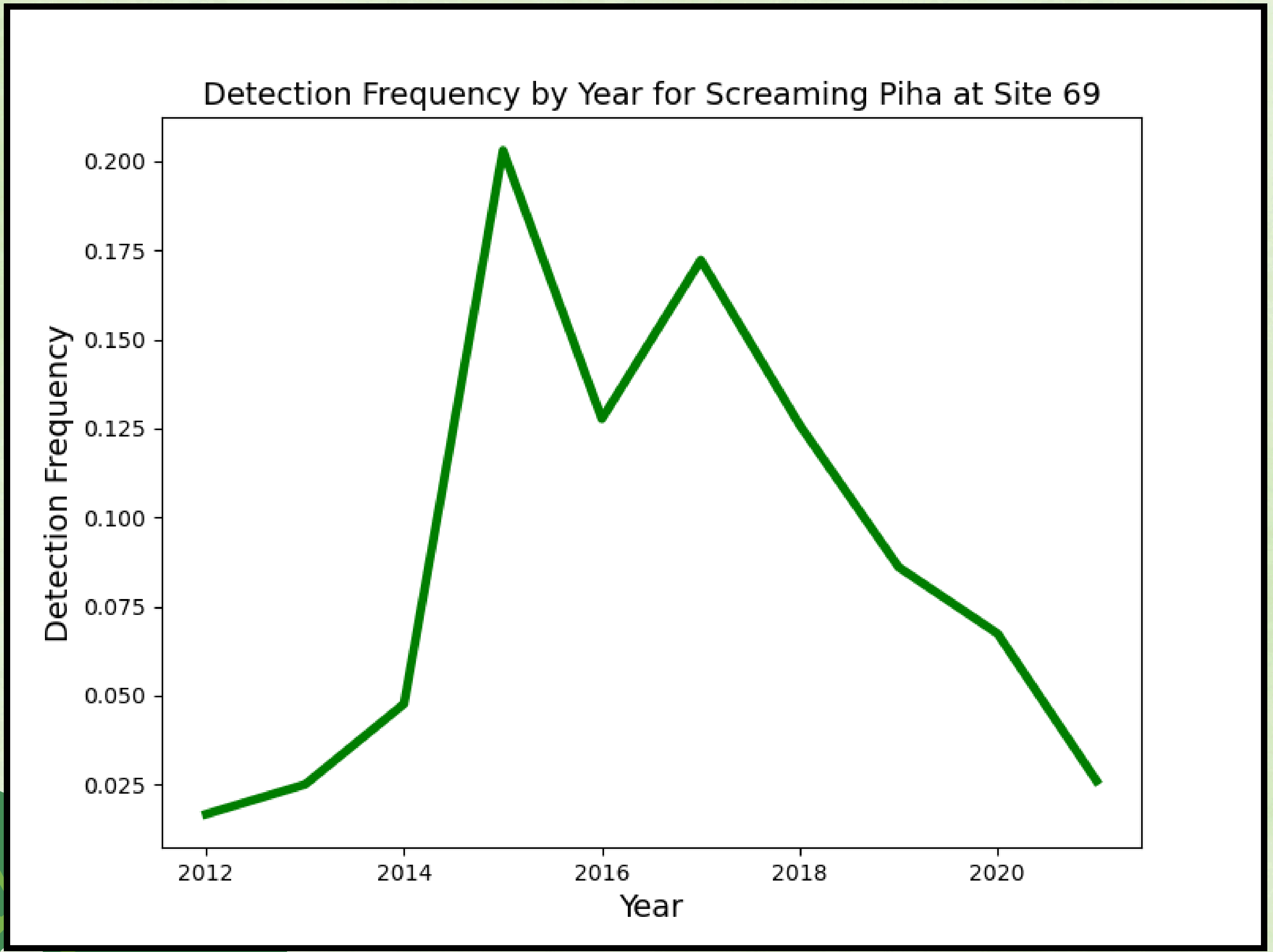



1. Effort hours (Detection covariates)
2. Minimum temperature (Occupancy covariates)
3. Species previously observed

# ML MODELING APPROACH 2

## 2. (BALANCED) RANDOM FOREST




eBird data tends to suffer from <u>class imbalance</u>
To accomodate this, we implement a <u>balanced random forest</u> – like a traditional random forest but draws a bootstrap sample from the minority class and samples the same number from the majority class.

# BEST ML MODEL

Also try augmenting our ML approaches with Synthetic Minority Over-Sampling Technique (SMOTE) - generate synthetic data for minority class to help with class imbalance

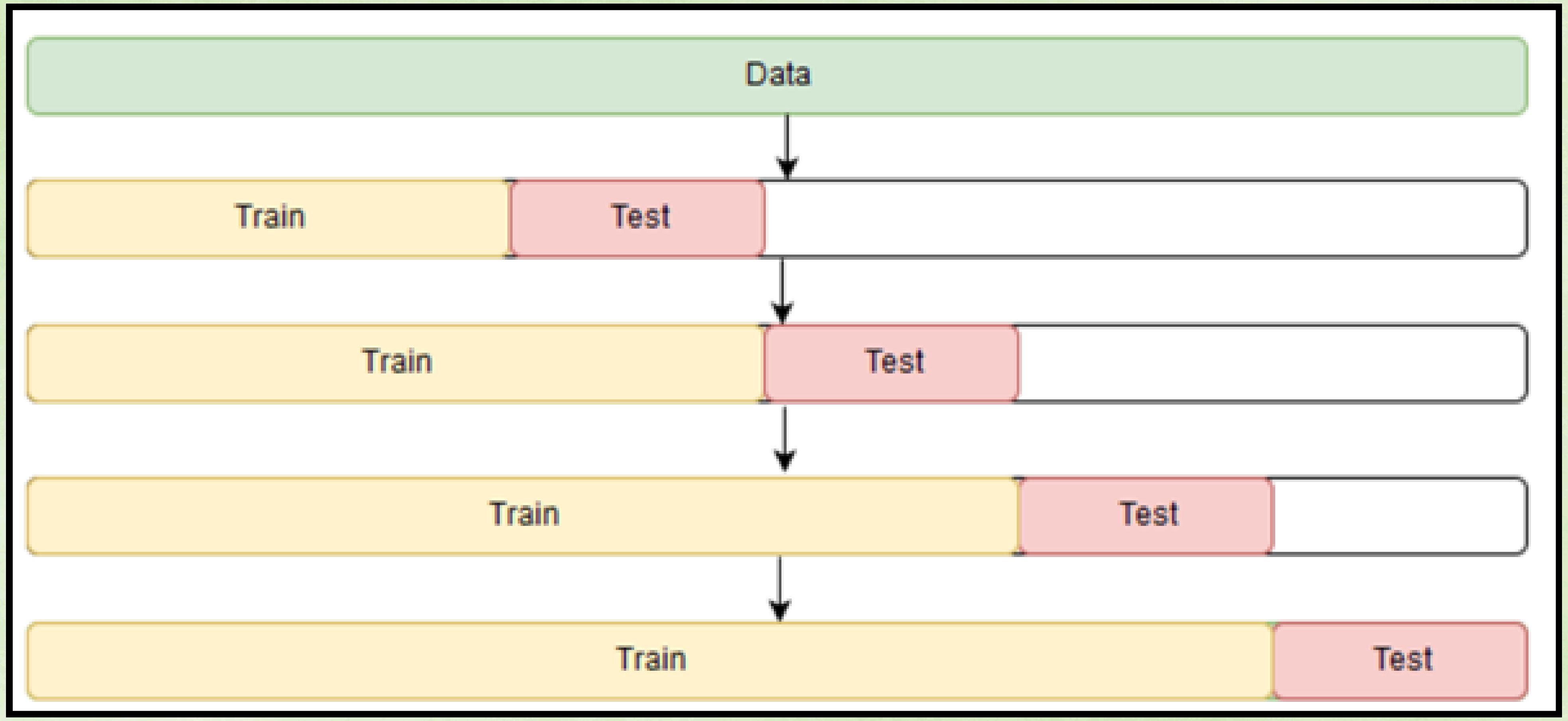



Image from: ‘Cross Validation in Time Series,’ Soumya Shrivastava, *Medium*

Compare our ML models with rolling cross-validation

**Best ML classification model - Binary logistic regression with L1 regularization and SMOTE (based on F1 score)**

# MODELING APPROACH

## 3. SPOCCUPANCY MODELS IN RSTUDIO

We use spOccupancy library in R to fit a spatial occupancy model.

This will allow us to accommodate imperfect detection.

Our occupancy covariates are temperature, precipitation, tree cover loss, EVI

While our detection covariates are day-of-the-year, time-of-day, effort, and number of observation

## BASIC MODEL STATEMENT

$$y_i|z_i \sim \text{Bernoulli}\,(p \cdot z)$$
$$z_i \sim \text{Bernoulli}\,(\psi)$$
$$\text{logit}\,(p) = \alpha_0 + \sum_j \alpha_j \cdot A_j$$
$$\text{logit}\,(\psi) = \beta_0 + \sum_j \beta_j \cdot B_j$$

with

$y_i =$ data at site i

$p =$ detection probability

$z_i =$ true occupancy state at site i

$\psi =$ occupancy probability

$\alpha_j =$ model parameters relating detection probability and detection covariates $A_j$

$\beta_j =$ model parameters relating occupancy probability and occupancy covariates $B_j$

# MODELING APPROACH

## SPOCCUPANCY MODEL ANALYSIS

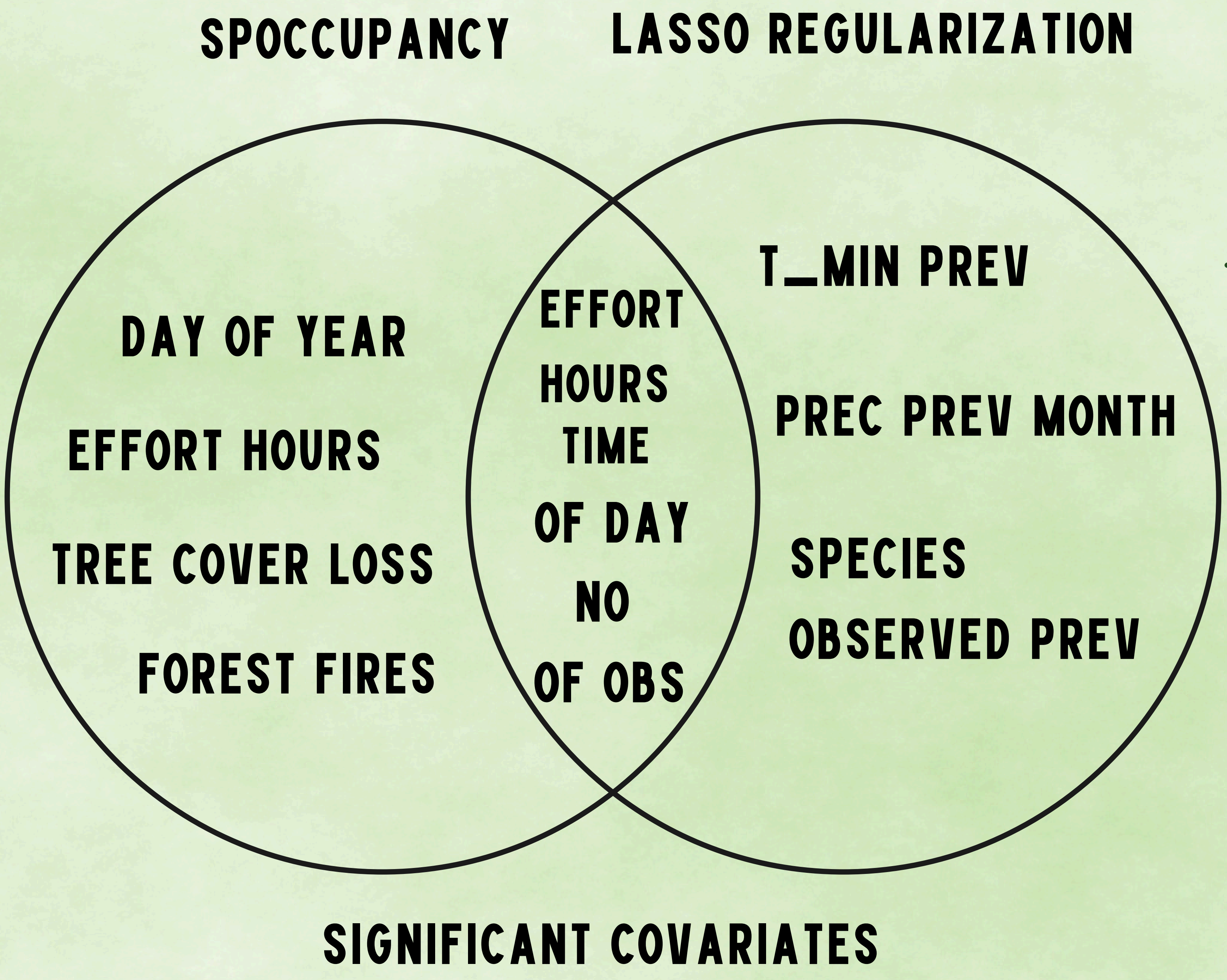
```
Occurrence (logit scale):
               Mean     SD    2.5%     50%   97.5%    Rhat  ESS
(Intercept)  0.0109 1.6356 -3.1908  0.0053  3.2795  1.0001 6261
precip       0.0080 1.6405 -3.1507  0.0099  3.1966  1.0003 6000
tmin         0.0277 1.6488 -3.2144  0.0509  3.3188  1.0002 6000
tmax        -0.0076 1.6594 -3.2778 -0.0372  3.2979  1.0003 6000
EVI          0.0144 1.6522 -3.2616  0.0334  3.2487  1.0001 6580
umdha        1.7232 0.4624  1.1325  1.7046  2.3602 47.1902    2
umdfire     -1.9763 0.2963 -2.5103 -1.9171 -1.5225 14.0234    3

Detection (logit scale):
               Mean     SD    2.5%     50%   97.5%   Rhat  ESS
(Intercept) -2.7726 0.2205 -3.2018 -2.7711 -2.3404 1.0008 5593
doy          0.0027 0.0006  0.0015  0.0027  0.0040 1.0010 5456
tod         -0.0378 0.0142 -0.0658 -0.0376 -0.0096 1.0016 5808
n.obs        0.0764 0.0250  0.0269  0.0763  0.1256 1.0016 6000
effort       0.4051 0.0369  0.3334  0.4050  0.4781 1.0005 5696

Spatial Covariance:
          Mean     SD   2.5%    50%  97.5%   Rhat  ESS
sigma.sq 0.997 3.4121 0.1793 0.5864 3.8768 1.2514 4414
phi      0.000 0.0000 0.0000 0.0000 0.0000 1.0064 4146
```

- Modelled with both spatial and temporal correlation effects
- Analysis shows that the detection covariates such as day of year, time of day, number of effort hours, and occupancy covariates such as tree cover loss, and forest fire loss are significant
- Spatial correlation between sites is very high
- Performed posterior predictive checks to compute Bayesian p-value, WAIC
- **Bayesian p-value = 0.7525**
- **WAIC = 2661.56**

# RESULTS AND CONCLUSIONS




- Among the ML models, Binary Logistic Regression with SMOTE performed better than the others
- Among the occupancy models, SpOccupancy models tPGOcc and stPGOcc performed similarly well

## Limitations

- Sites clustered in the same region -- low variance in occupancy covariates between sites.
- Variable and sometimes strong class imbalance
- Environmental data collected at different frequencies in time.

## Future directions

- Model with more species across multiple sites in the Amazon rainforest
- Implement spatial correlation with ML models

THANK YOU!