Identifying Hazardous Waste Sites

•••

Tim Mango
Flatiron Data Science Fellowship
Final Project

Organizations of Interest

Earth Challenge 2020:

A global citizen science initiative that will demonstrate how small digital acts of science can help monitor and improve environmental and human health

Datakind DC:

Project inspiration came from the June Datakind DC datajam

#EC2020

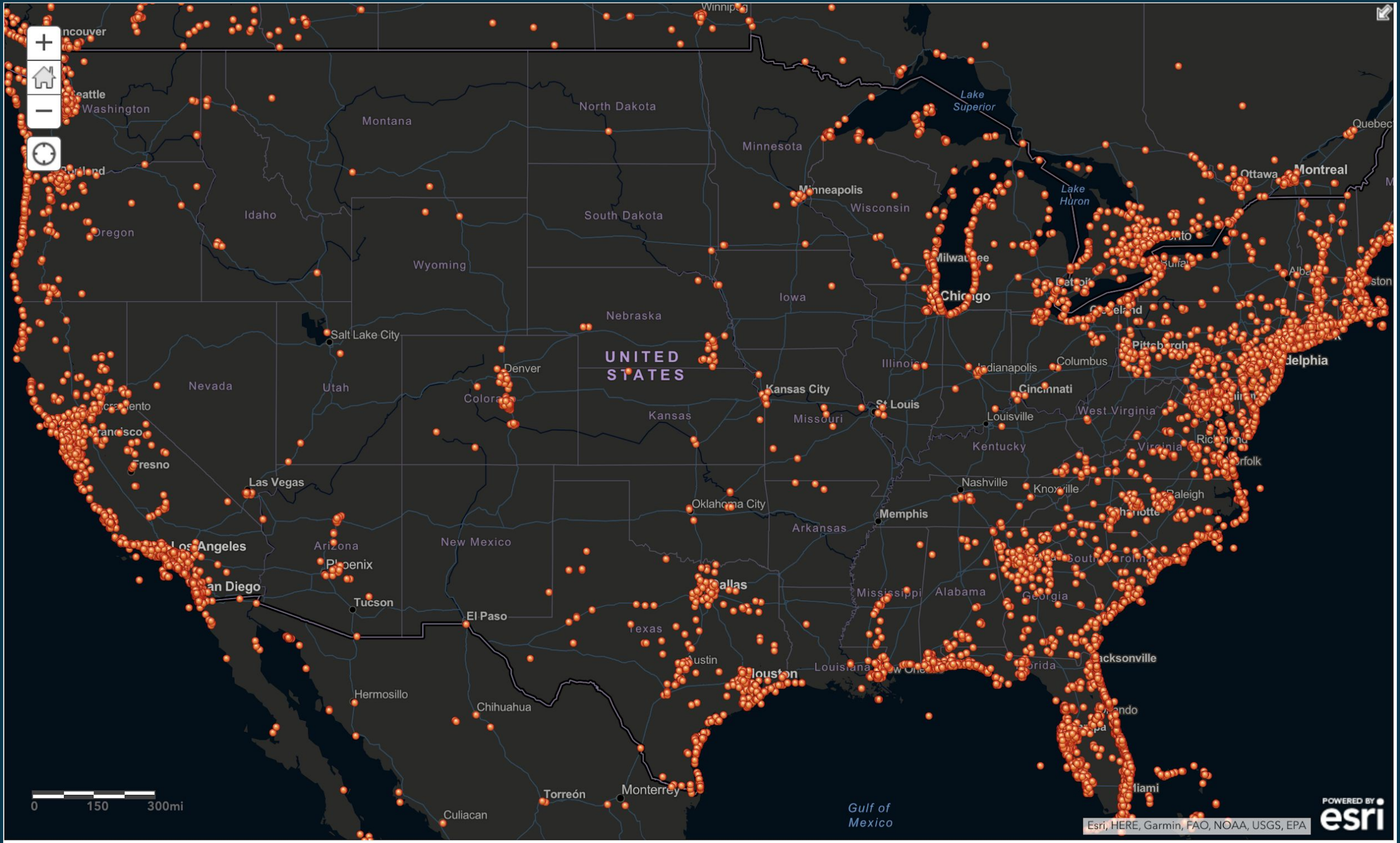
DataKind DC

The Problem

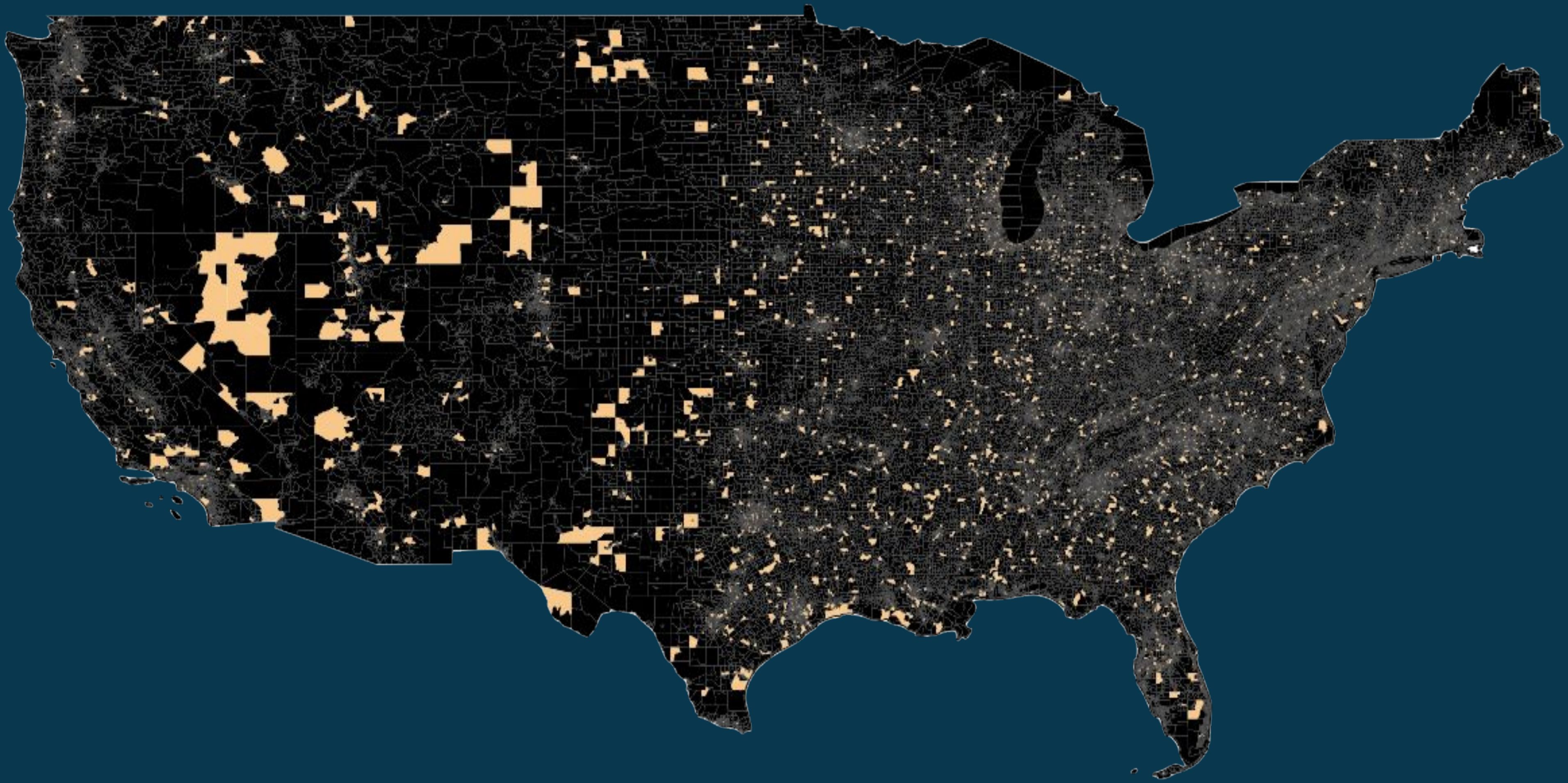
Volunteer Beach Cleanup Team

Hazardous Waste Cleanup Team

Tides Volunteer Waste Cleanup Sites



Toxic Release Inventory Sites 2017



Best Discovered Approach

Hazardous waste sites need to be identified and cleaned. At the same time, volunteers need to be protected from dangerous exposure.

Predictive model for Waste Site Prediction: XGBoost

Explanatory Variables: ~200 final variables

Target Variable: ~6,600 of 220,000 Census Block Groups

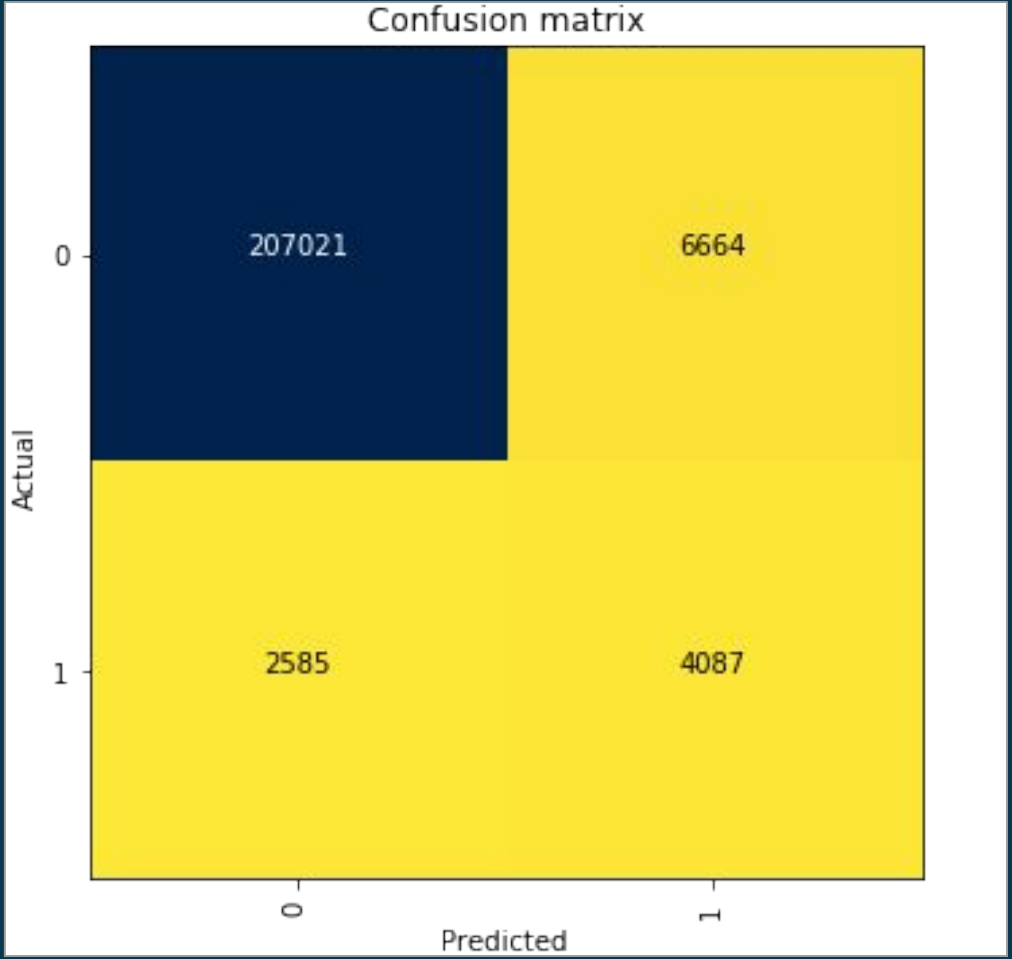
Target Data

Explanatory Data

Most Important Features

1. Land Area
2. Home Value
3. Vacant Houses
4. Percentage Male
5. Education
6. Population Age
7. Transient Population
8. Health Insurance
9. Group housing

Model Performance

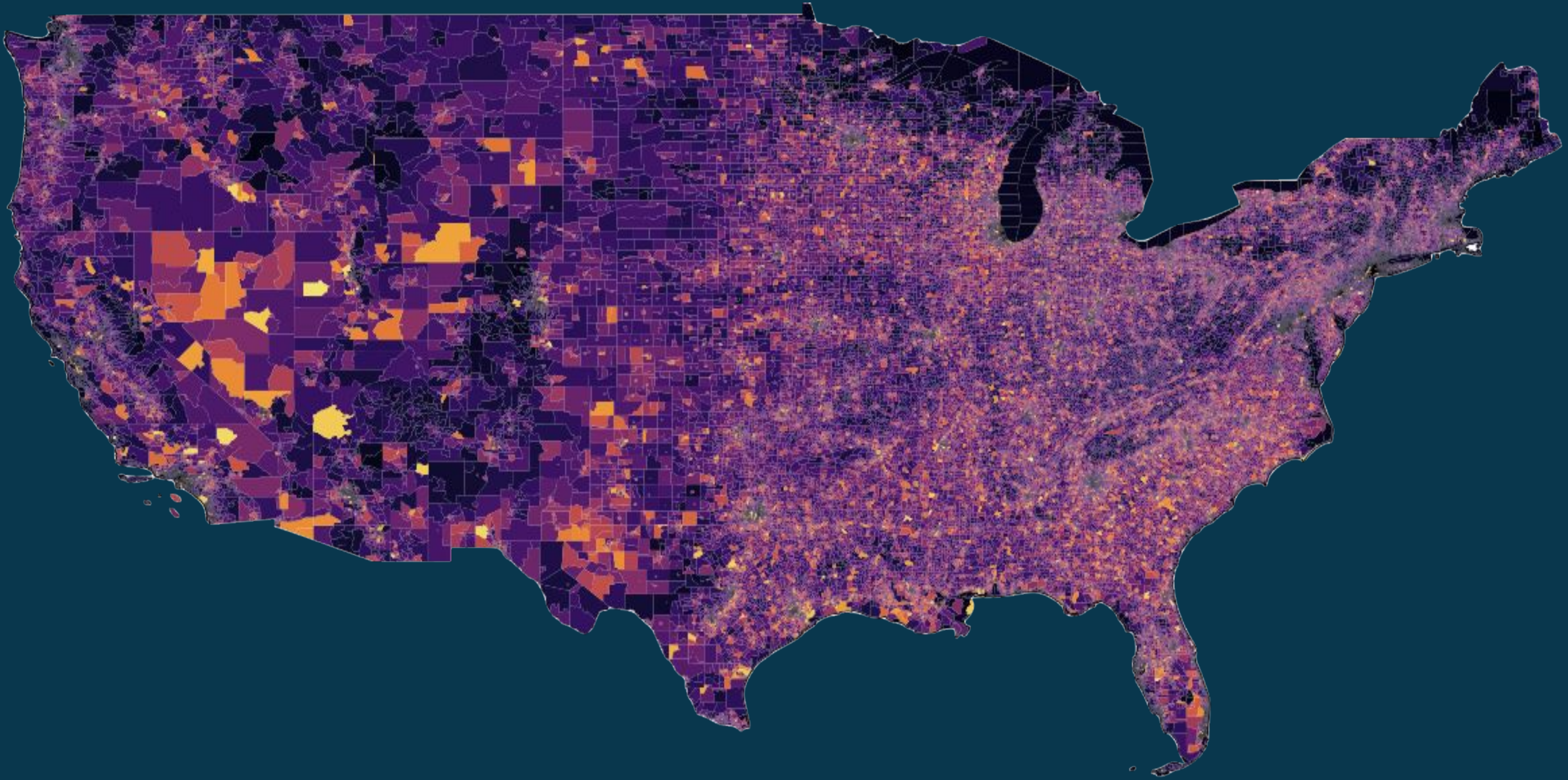


Actual \ Predicted	0	1
0	207021	6664
1	2585	4087

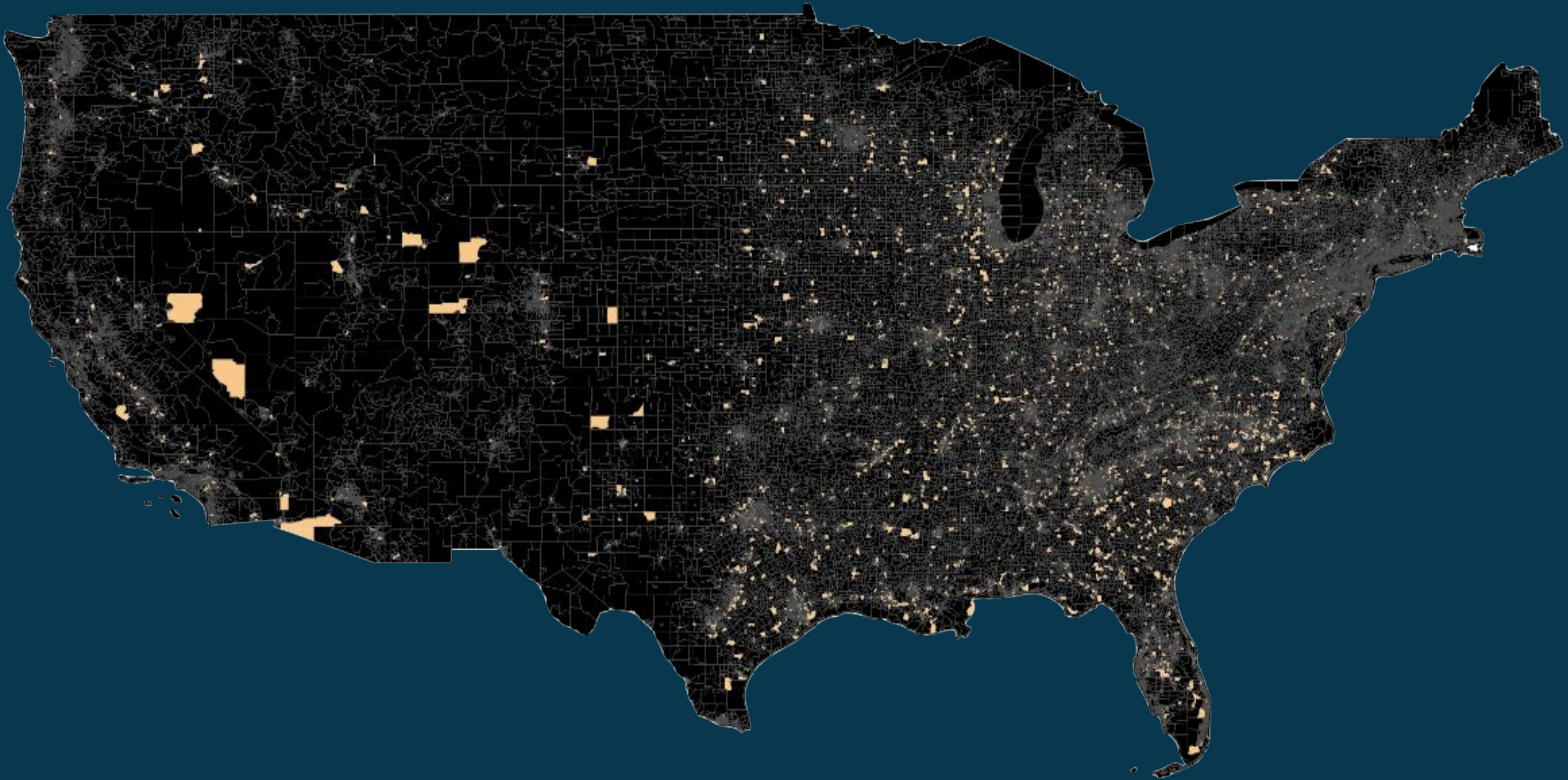
Sensitivity is the metric of interest for the model. Sensitivity measures the probability that the model will correctly identify census block groups that contain TRI sites.

Sensitivity: **61.26%**

Industrial Waste Prediction Probability Map

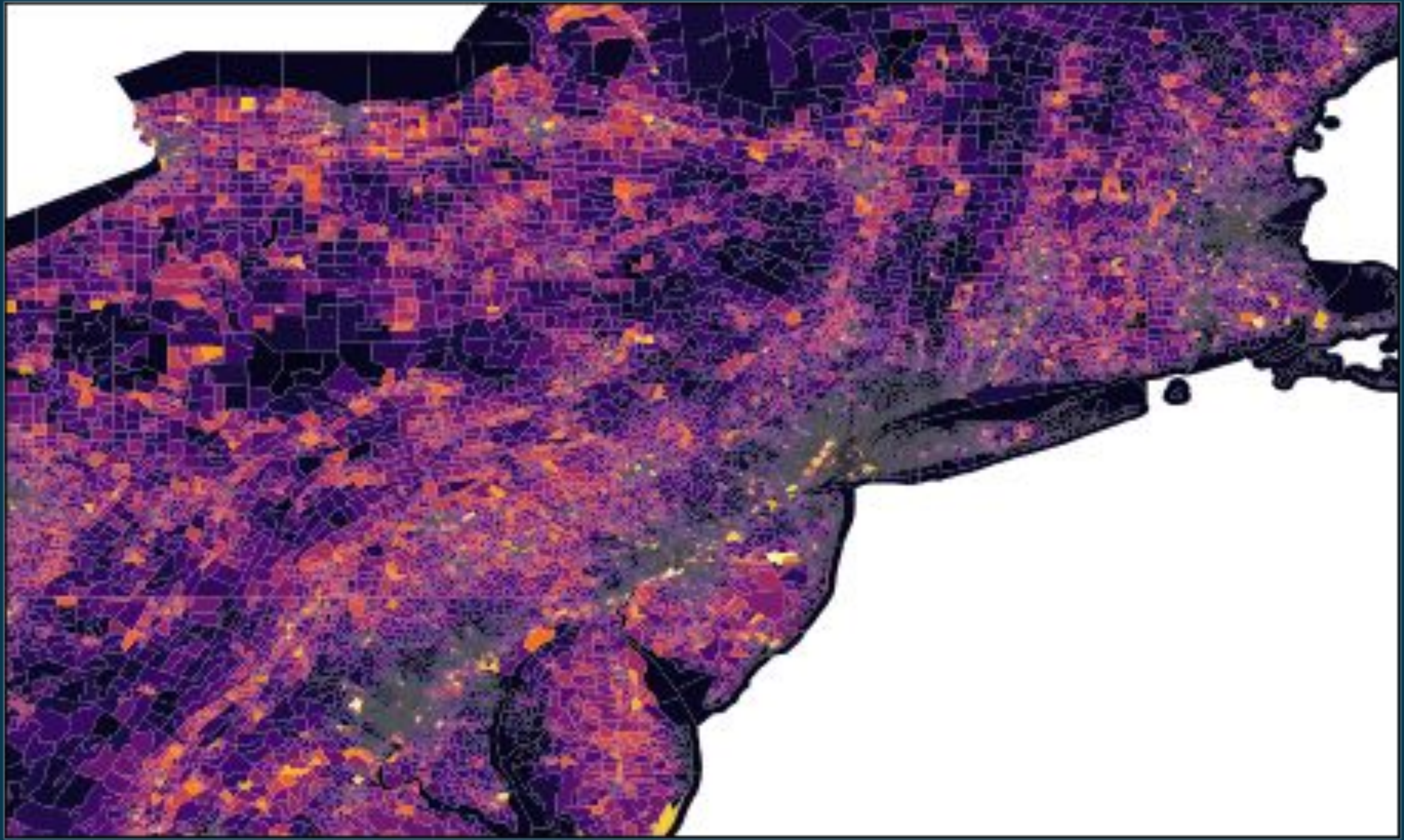


Top Unidentified Industrial Waste Dumping Locations

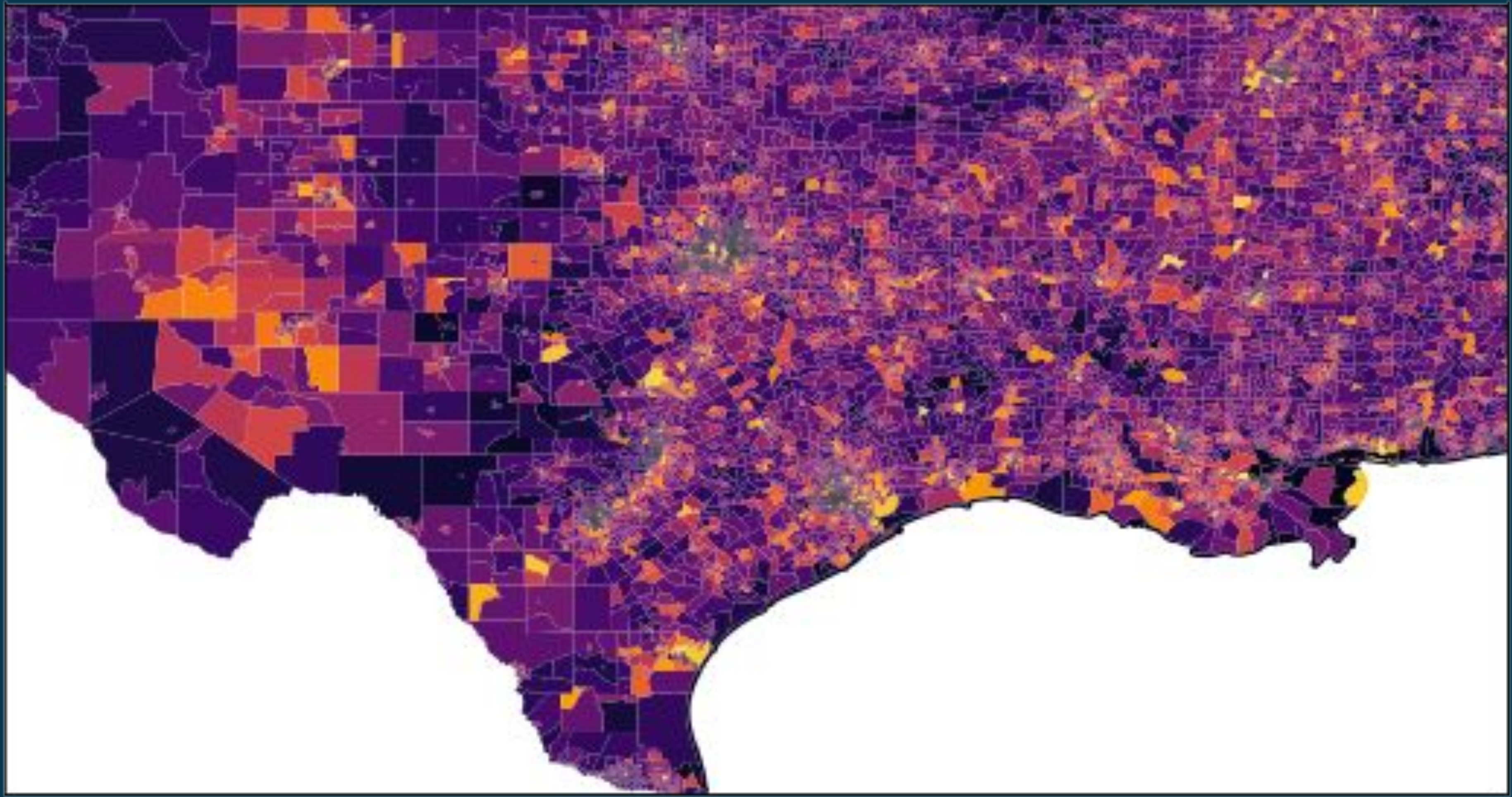


Thank you for your time!

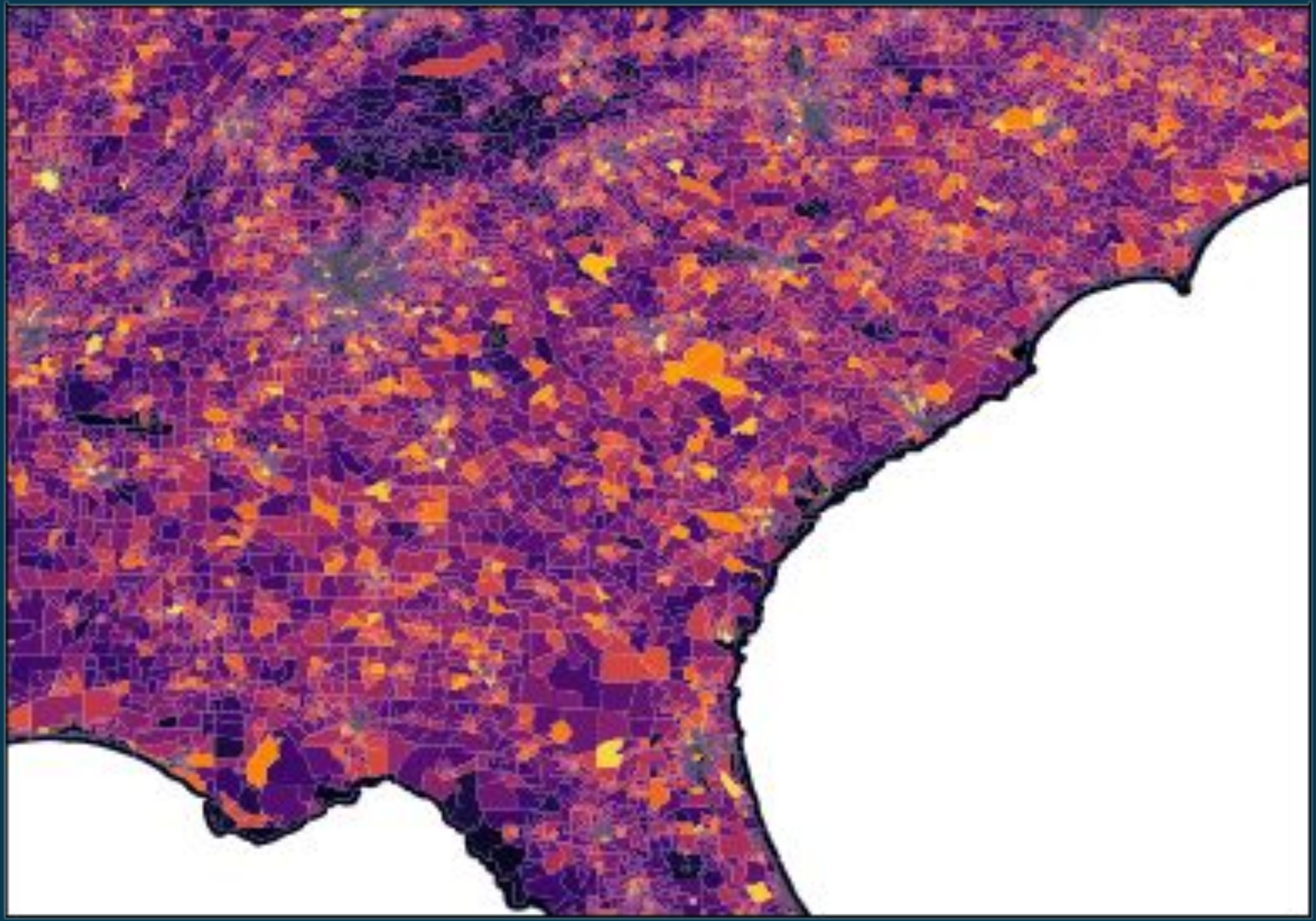
Northeast Prediction Probability Map



Southwest Prediction Probability Map



Southeast Prediction Probability Map



Industrial Waste Prediction Probability Map