Where are all the onsite systems in the United States?

Machine Learning-based National Inventory of Buried Wastewater Infrastructure

Nelson da Luz, PhD

Research Assistant Professor

UMass Amherst, CEE

University of Massachusetts Amherst

Disclaimer: The materials being presented represent my own opinions, and do NOT reflect the opinions of NOWRA.

Background

https://www.plconcrete.net/the-benefits-of-precast-concrete-septic-tanks

Safe management of human waste is crucial

~20% of US likely use onsite wastewater treatment systems (OWTS)

1990 - The last census of sanitation systems types serving US

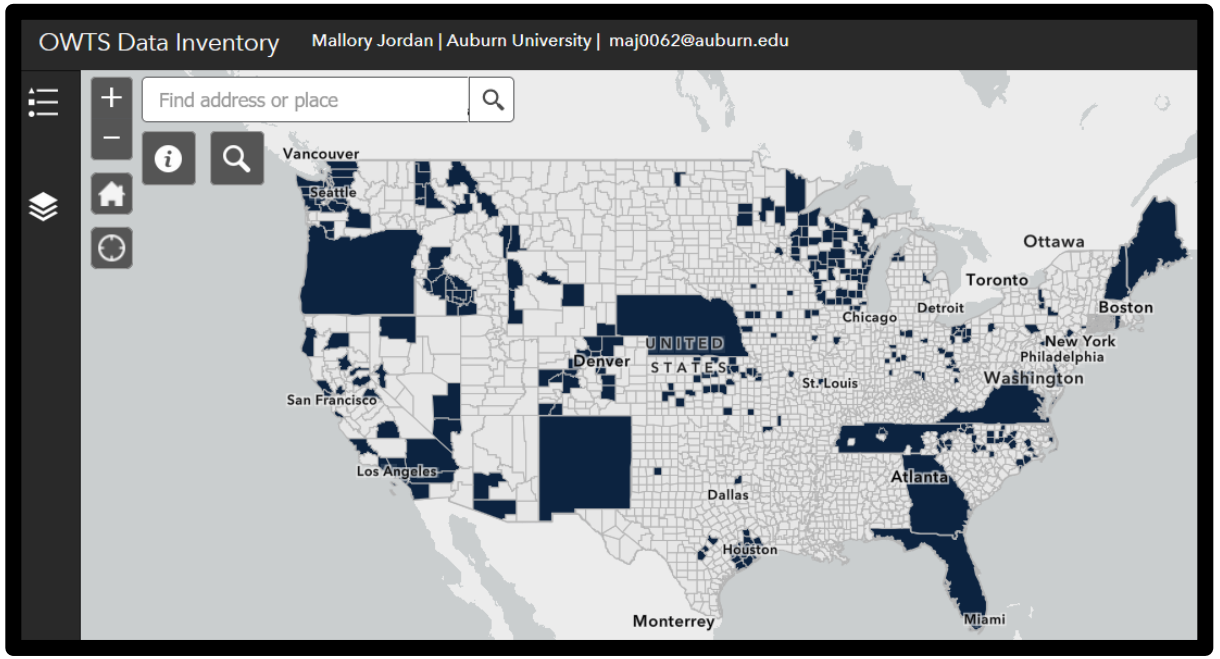
Significant gap in our understanding of the **number, locations, and density of OWTS** across the country

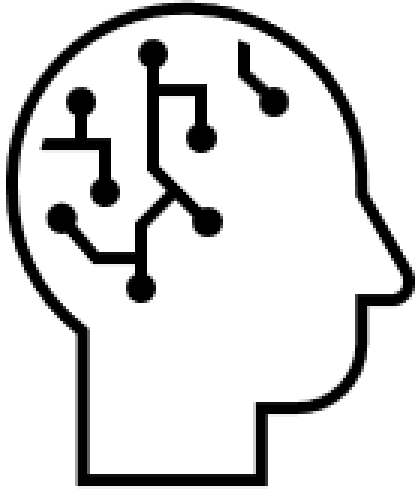
Why does it matter?

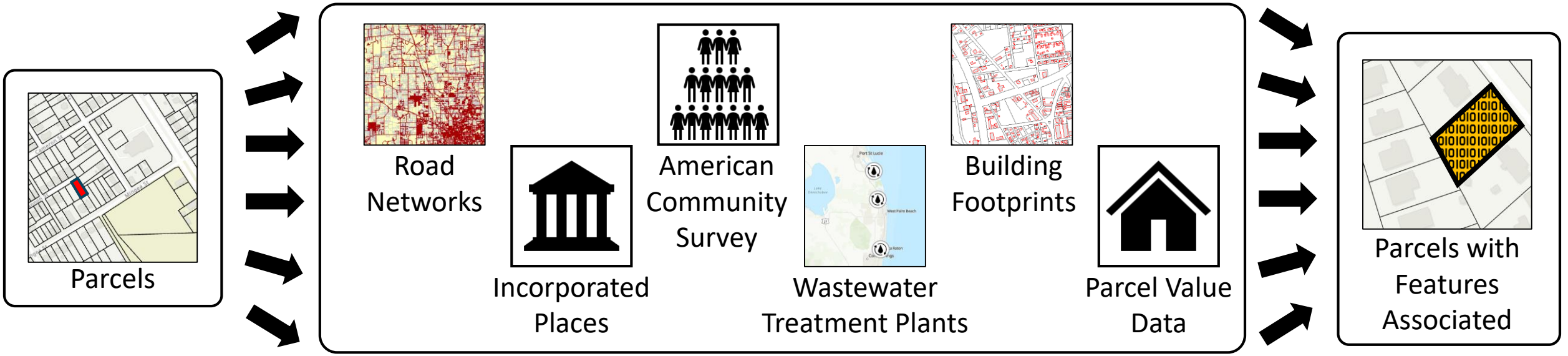
Use Case	Need for Wastewater Inventory	States
Emergency / Disaster Response 	Evaluating prevalence of OWTS in communities affected by natural disasters; impact on drinking water	FL, MA, NC
System Failure and Risk to Water Supply 	Evaluating OWTS prevalence in communities with high groundwater use for drinking	CA, FL, NC, VA
Nutrient Loading / Coastal Concerns 	Quantifying contribution of OWTS to nutrient loading and resulting environmental issues on the coast	FL, NC, VA
Advocacy and Funding 	Directing advocacy and funding to help communities and individuals with maintenance or upgrades	CA, NC, VA
Asset Management, Consolidation, Urban Planning 	Identifying areas for potential consolidation into sewers, growth planning, measuring access	CA, FL, MA, NC, VA
Government Agency Communication 	Streamlining data sources and information about OWTS across agencies	CA, FL, NC, VA

How can we find out where onsite systems are?

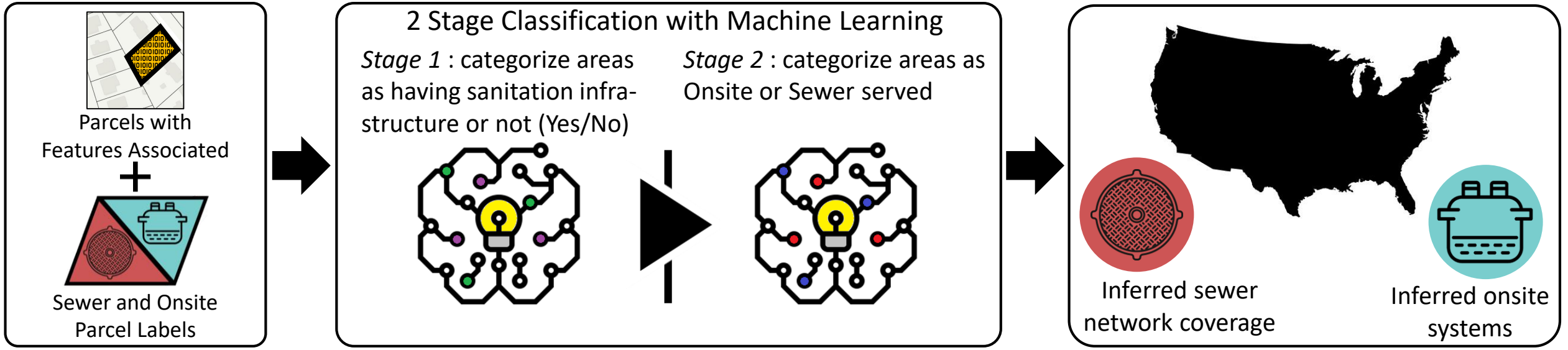
- Our idea: Most sanitation systems are ‘invisible’ (buried) but inferable
- Fill data gaps by using other indicators
- Leverage these other indicators with machine learning
- Machine learning is part of the AI field and allows analysis of massive quantities of data through processes like pattern matching







1. Data Sources and Data Processing

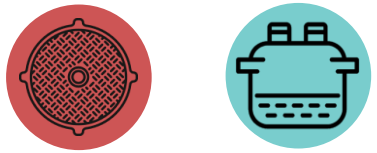


2. Machine Learning (2 Stage Random Forest) Model

3. Predictions Across USA

What does the model output for each parcel?

1. Inferences
   - **Sewer** or **Onsite**



2. Confidence
   - e.g., On a scale of 0 to 100 how confident are we that this parcel is served by **Sewer** or **Onsite**

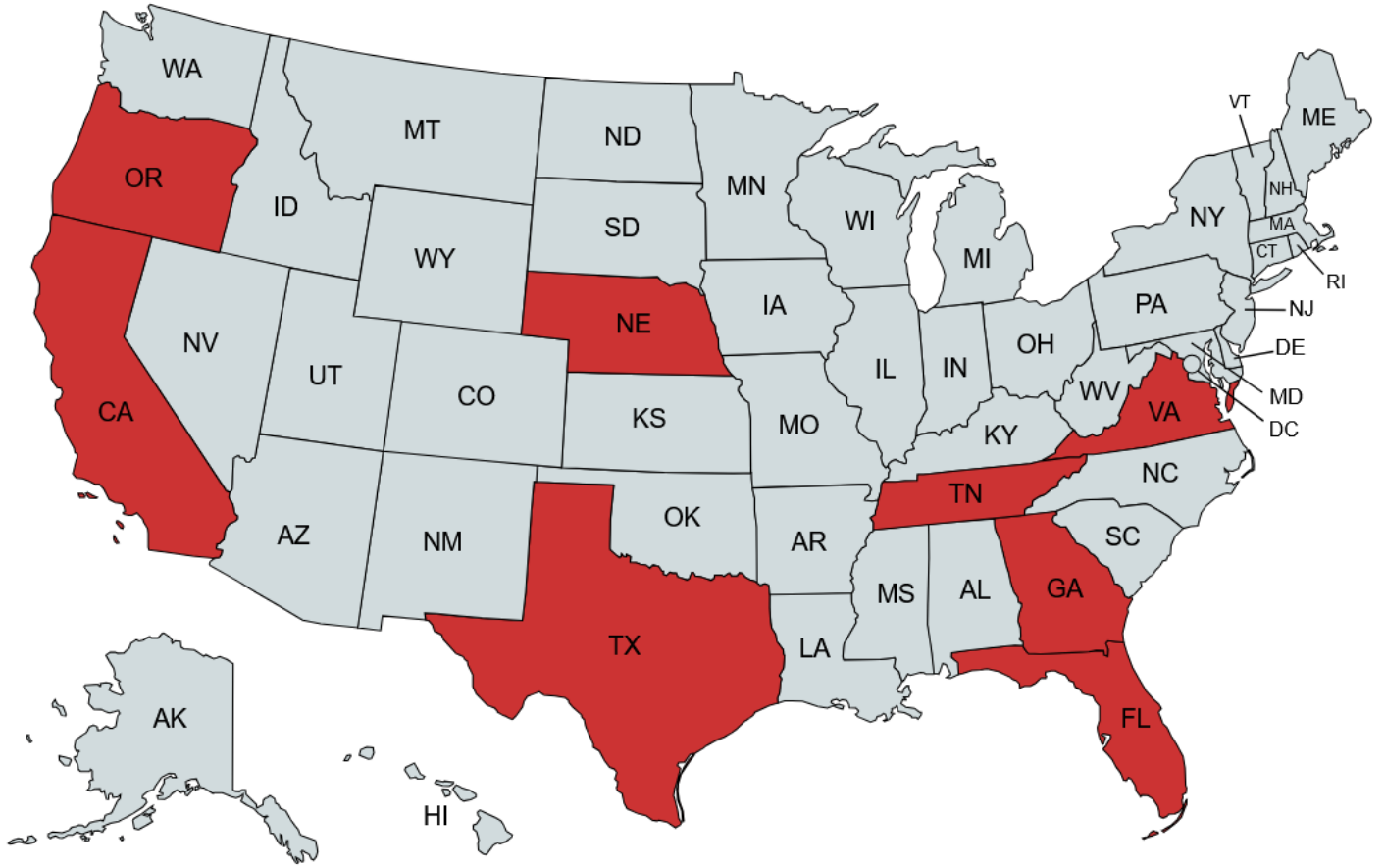
- Accuracy is not a model output, but we can calculate it if we compare to ground truth data
  - Accuracy = #Correct/Total

Accuracy: We can look at the target and confirm that we hit the target in the bullseye
Confidence: Even though we can't see the target, based on what we know about other targets, we are 85% sure we hit the bullseye

Labels Used to Train the Model

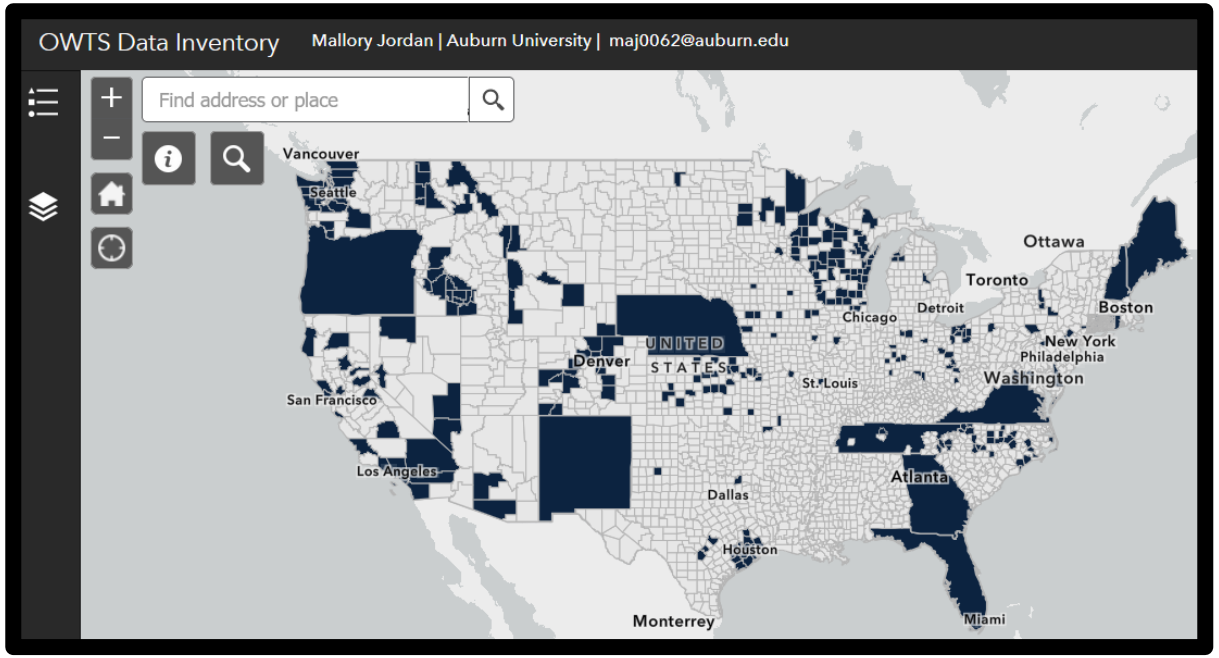
- Label data collection
- 6.7 million labels
  - 1.4 million onsite
  - 5.3 million sewer

Format	States
Readily downloadable	FL
Scraped from database	GA, TN, NE
Data provided by state	VA, CA
Sewer line proximity	OR, TX, CA



WW Infrastructure data is highly decentralized

- Web scraping tables
  - Geocoding addresses
- PDF scraping
- Previously geocoded by state (VA)
- Sewer line proximity (50m)



Training Dataset Construction

- We used data from multiple states to help the model in its ability to generalize over a variety of places
  - Analogy: Raccoons vs. Red Pandas
    - We are investigating the difference between raccoons and red pandas
    - While the animals of each type are generally similar in appearance, the more examples of each type we have, the better we will get at identifying them
- States used for training: FL, VA, GA, NE, TN, CA, TX, OR



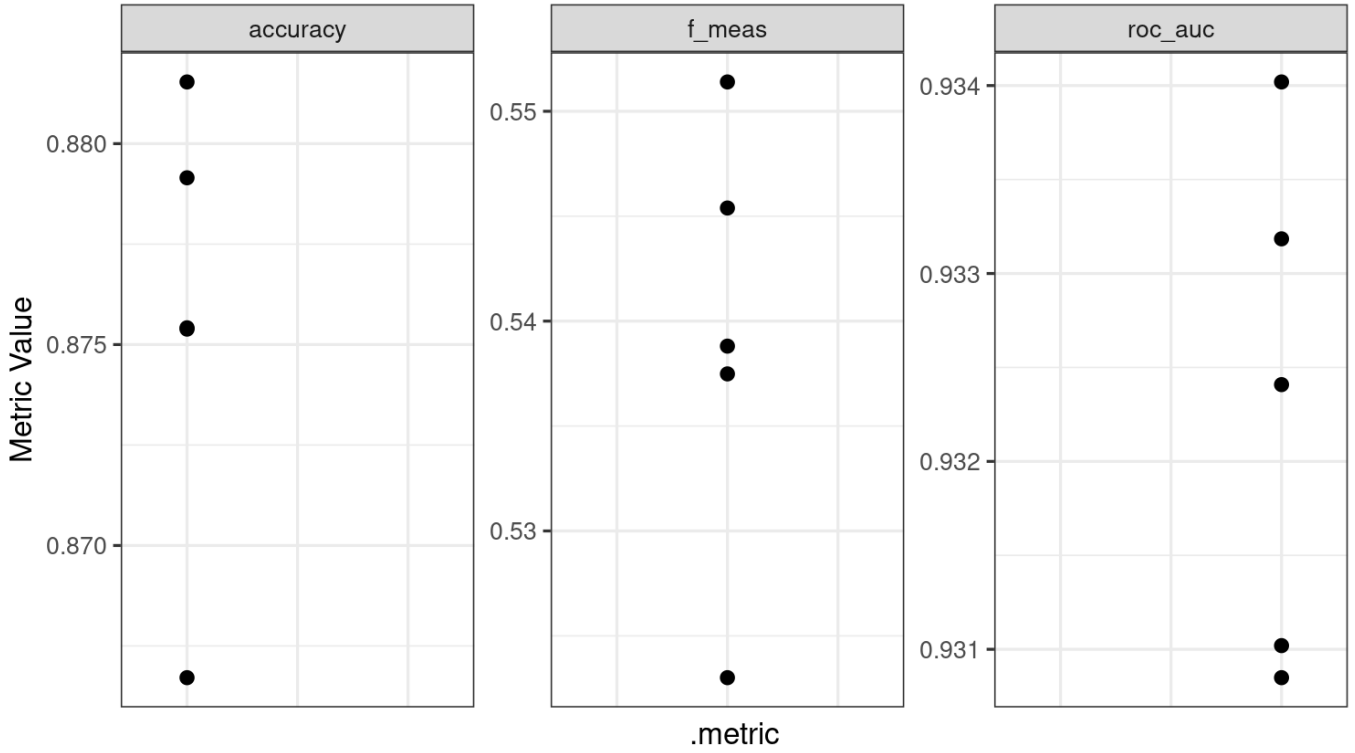


Building the National Model

Priority	Accomplished by
Use a balanced training dataset	equal proportions Onsite and Sewer, requires under-sampling sewer
Reserve some label data for testing	Save 30% of onsite labels for unseen test set
Ensure geographical distinctness by spatial splitting	Separate training samples based on US census tract. If a tract is used for training, it can’t be used for test. Semi-random selection
Multiple sampling folds to ensure reliability of performance metrics	5 sample folds to make sure we didn’t just hit the jackpot on Round 1

Results across 5 sample folds

- Accuracy around 87-88%
- Tight range suggests model robustness
  - Stability
  - Low variance



In the best sample fold

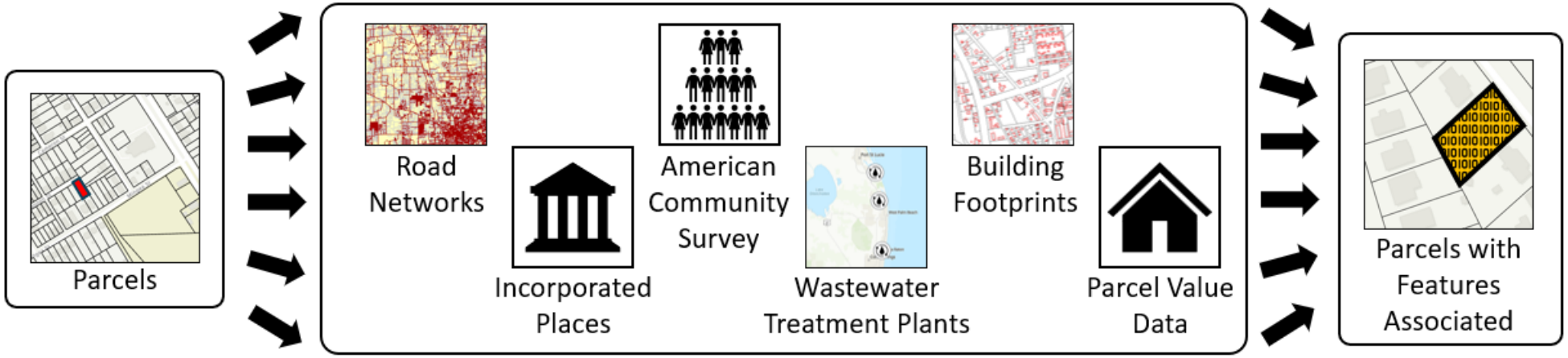
- Range of accuracies across states

State	Accuracy %	Onsite Accuracy %	Sewer Accuracy %
CA	92.3	90.0	92.4
FL	87.9	79.2	88.7
GA	93.4	93.4	NA
NE	98.3	98.3	NA
NC	80.2	99.8	79.4
OR	79.8	NA	79.8
TN	97.1	97.1	NA
TX	81.1	NA	81.1
VA	74.1	98.2	51.7

We have a model, let’s try it nationally

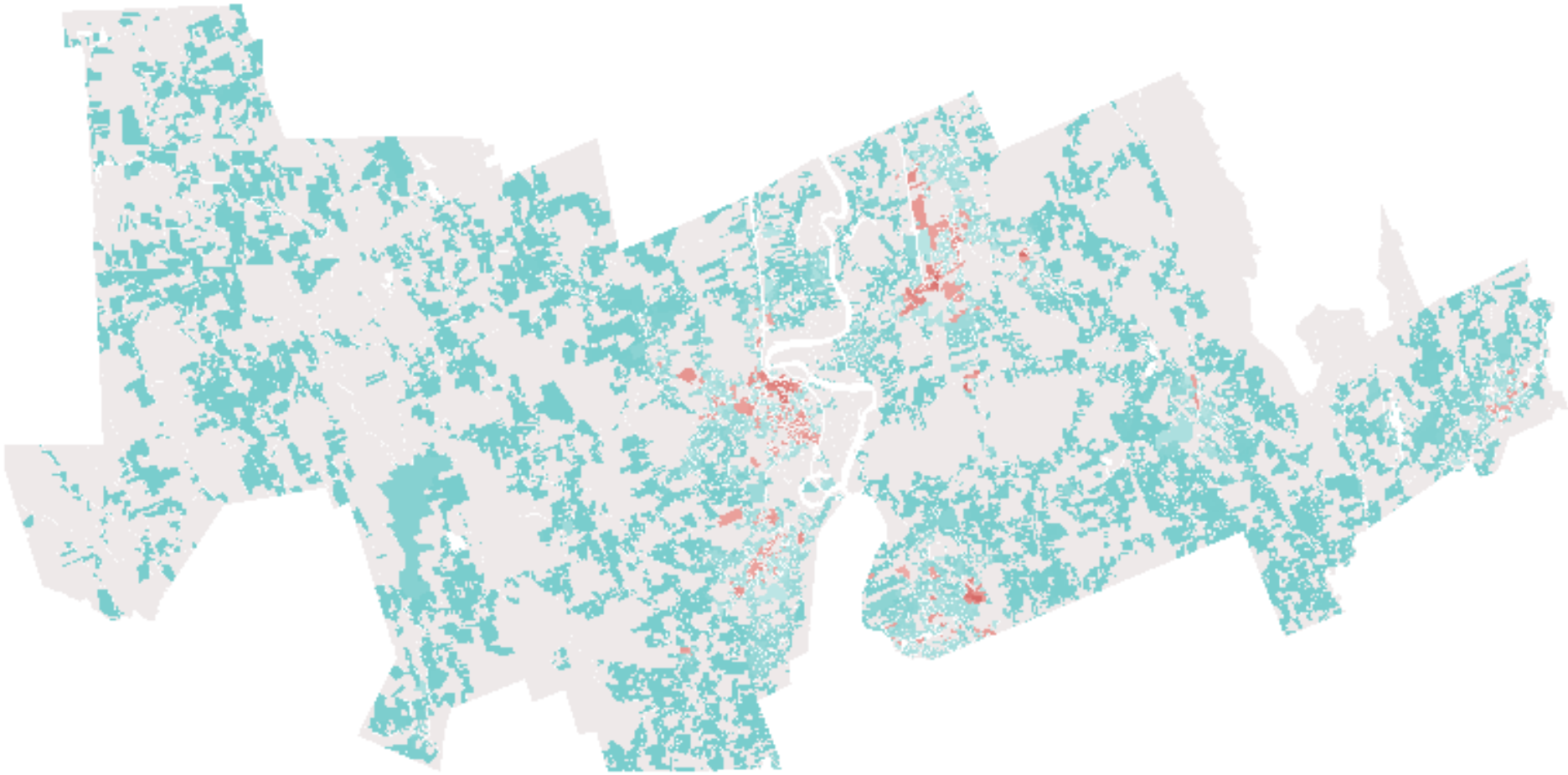
- There are approximately 3150 counties in the US
- We have collected some amount of data from 329 counties
  - Incomplete samples
- Our model is trained on about 60 features
- There are more than 150 million land parcels in the US
- More than 9 billion pieces of information needed to develop a national estimate

Data Processing Pipeline and Framework

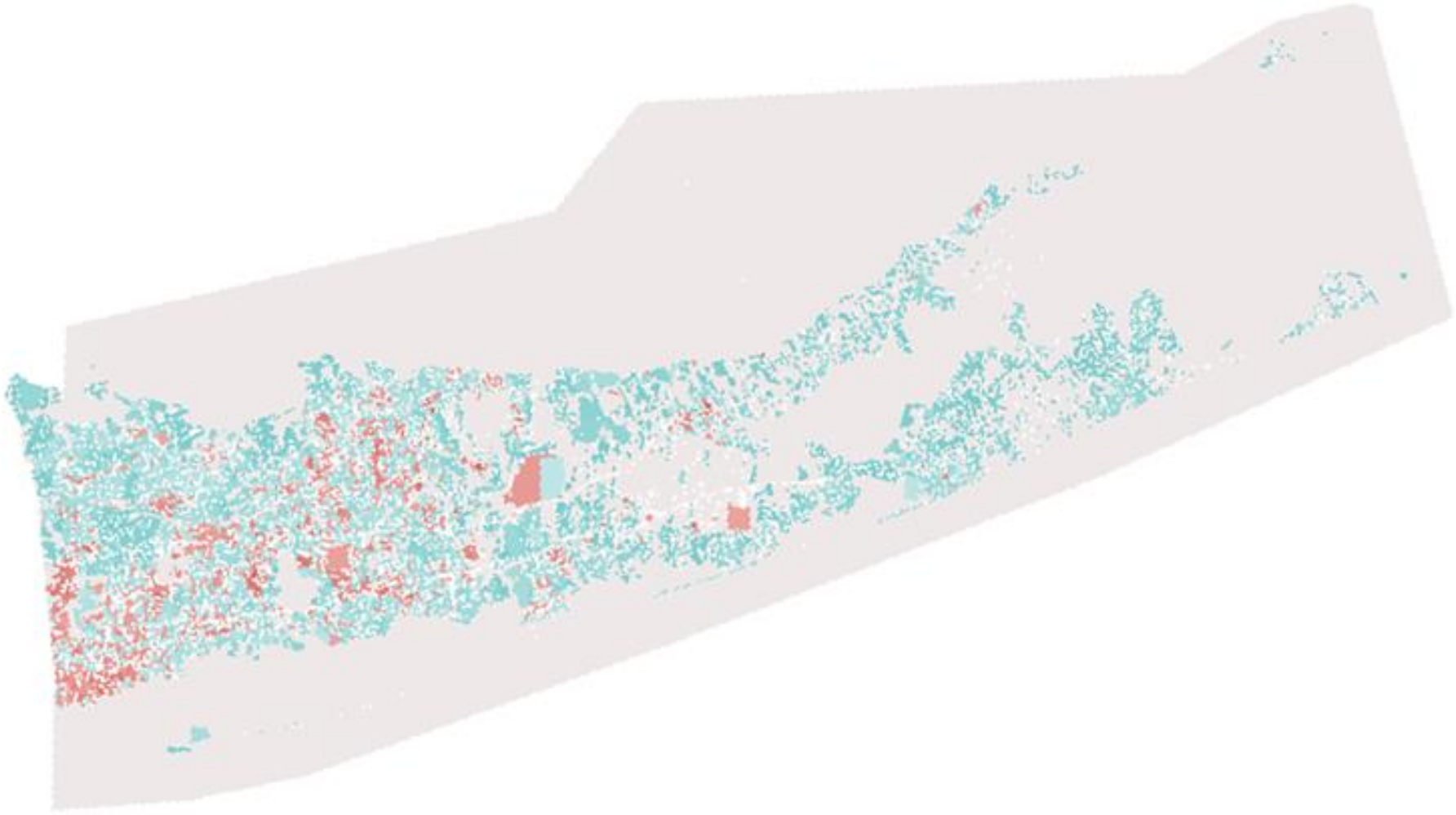


- Each feature must be calculated for every parcel in the US
- Group calculations by county, data storage grouped by state and county FIPS codes
- Pipelines to:
  - Automate data download for input datasets
  - Process each dataset and assign to parcels by entering State FIPS code
  - File check to ensure output files were generated as expected
  - Rerun failed jobs

Hampshire County, MA

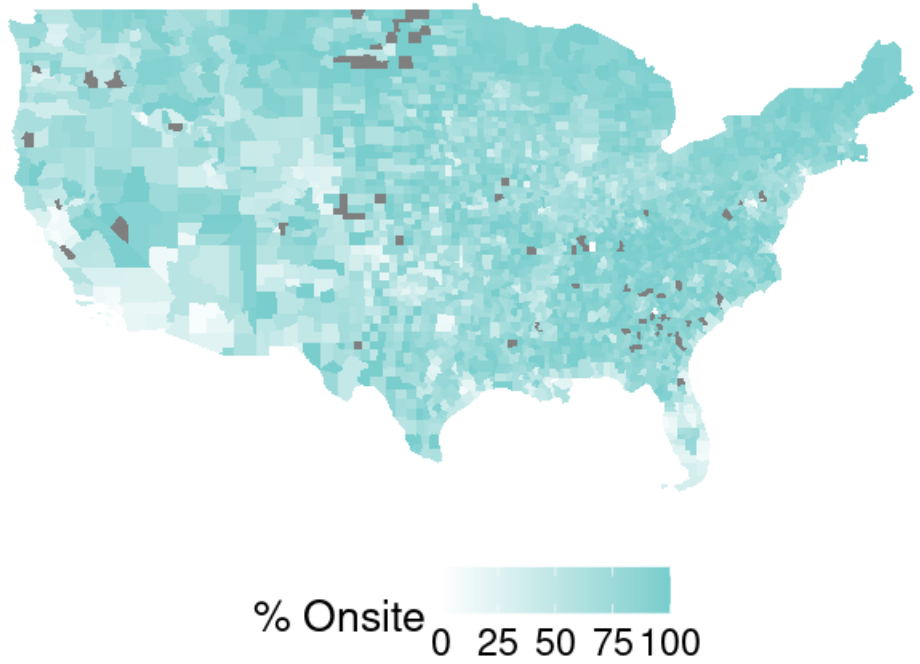


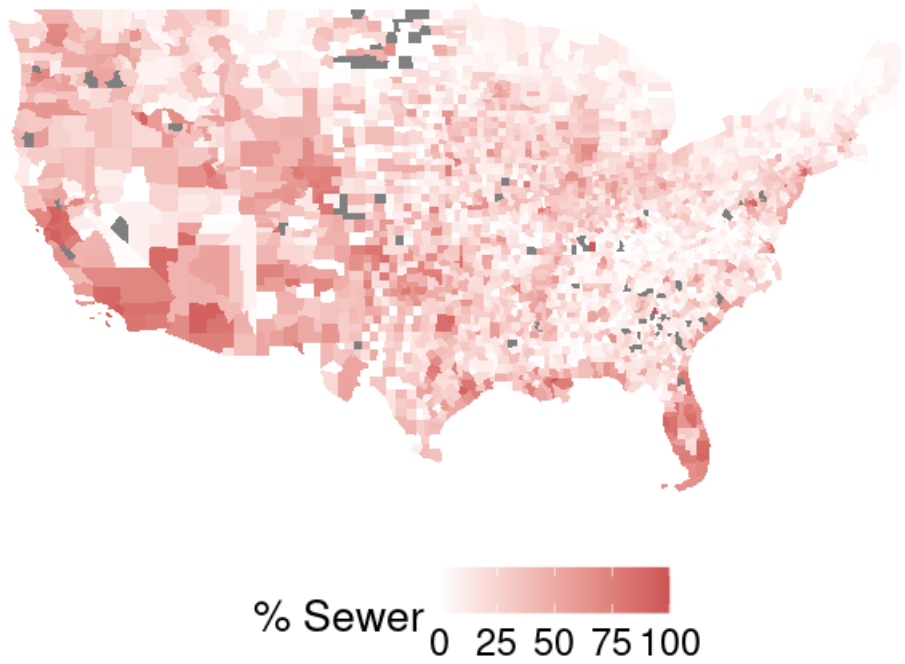
Suffolk County, Long Island, NY



National Results (for now)

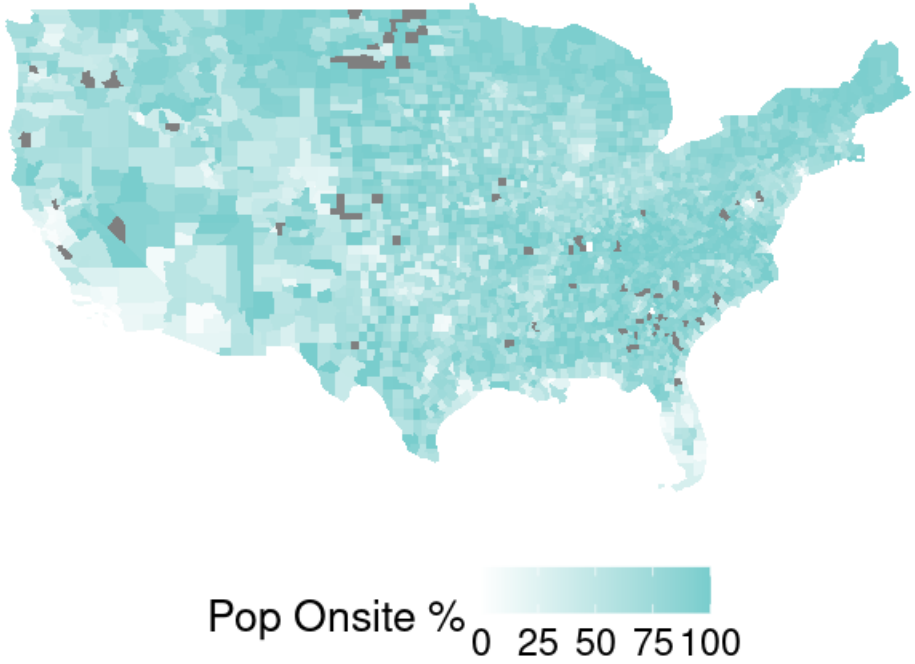
- 52M Onsite served parcels
- 48M Sewer served parcels
- Proportion of Parcels served by Onsite vs Sewer in each county

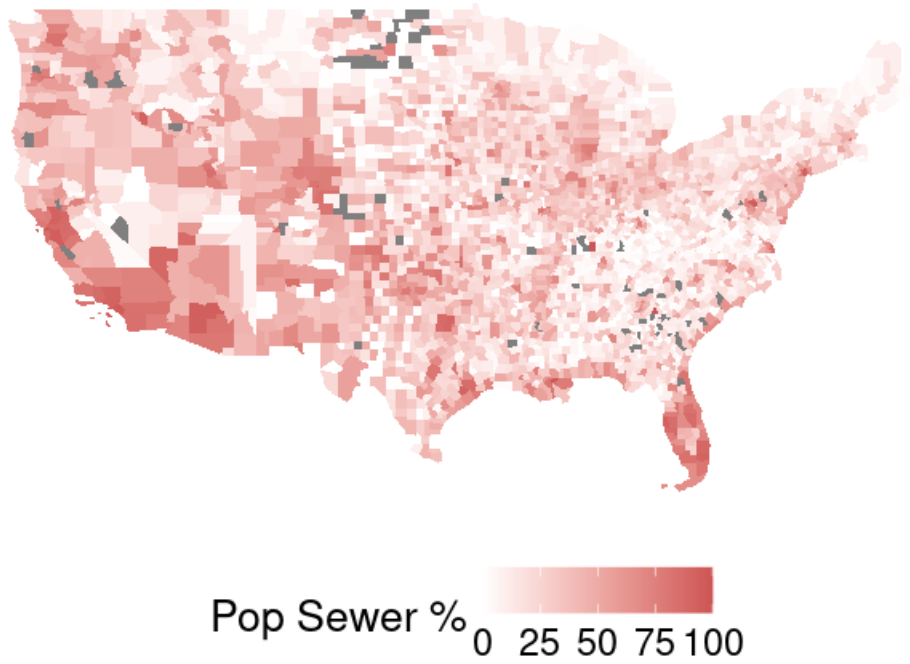




National Results for Population (for now)

- 147M Onsite served population (42% US population)
- 192M Sewer served population (56% US population)
- Proportion of Population served by Onsite vs Sewer in each county



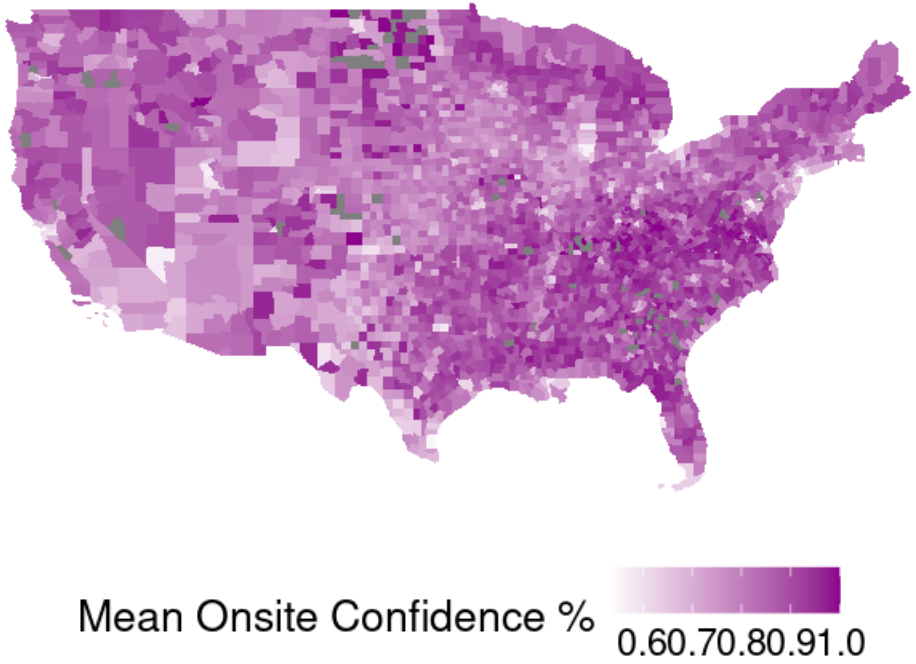


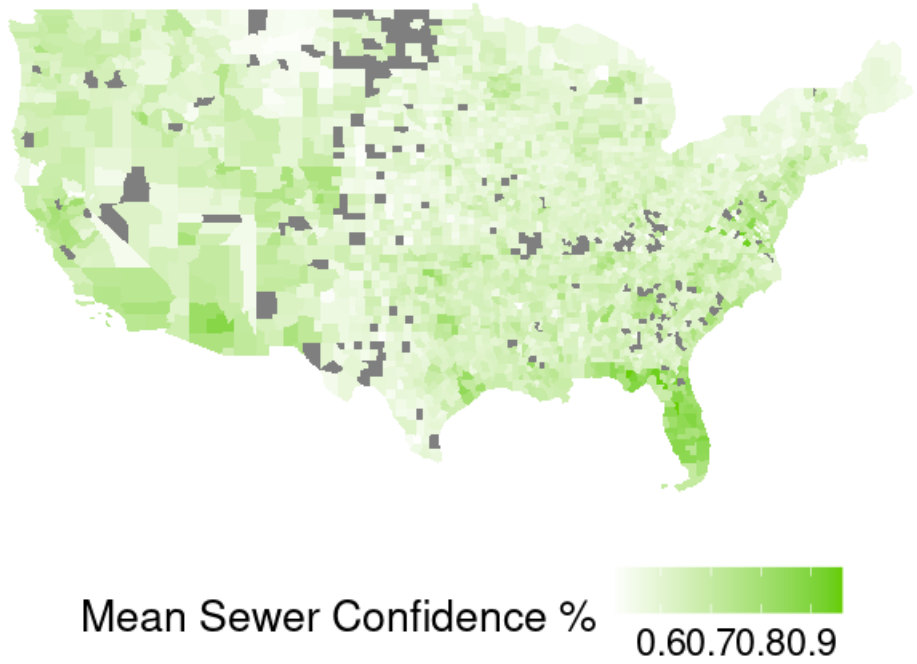
What do these results suggest

- Conventional estimates of the Onsite served US population are between 15-25%
  - We estimated 42%
- We are overestimating # of Onsite served locations
  - Residential vs non-residential parcels
    - 137M (-5M) residential Onsite served population
    - 173M (-19M) residential Sewer served population
  - Need more refinement
- Still reveals wastewater system type similar in principle to Transient Non-Community Water Systems (TNCs)

National Results for Model Confidence

- The model is generally much more confident in its Onsite predictions than its Sewer predictions





Where do we go from here

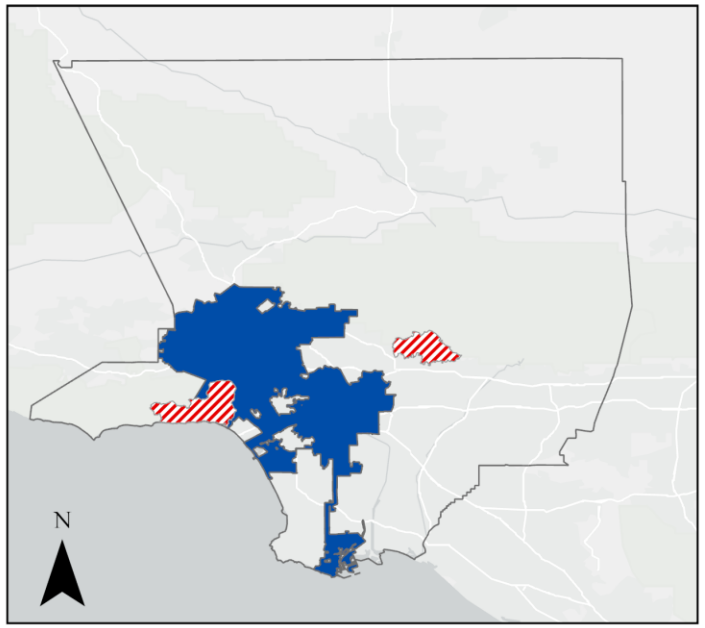
- Continue to collect label data
- Experiment with unbalanced training data that reflects more sewer coverage or better represents small sewer systems
- Experiment with selecting stratified ‘representative’ census tract samples for training instead of mostly random census tracts
  - e.g., different degrees of urban-ness, estimated proportions by Census

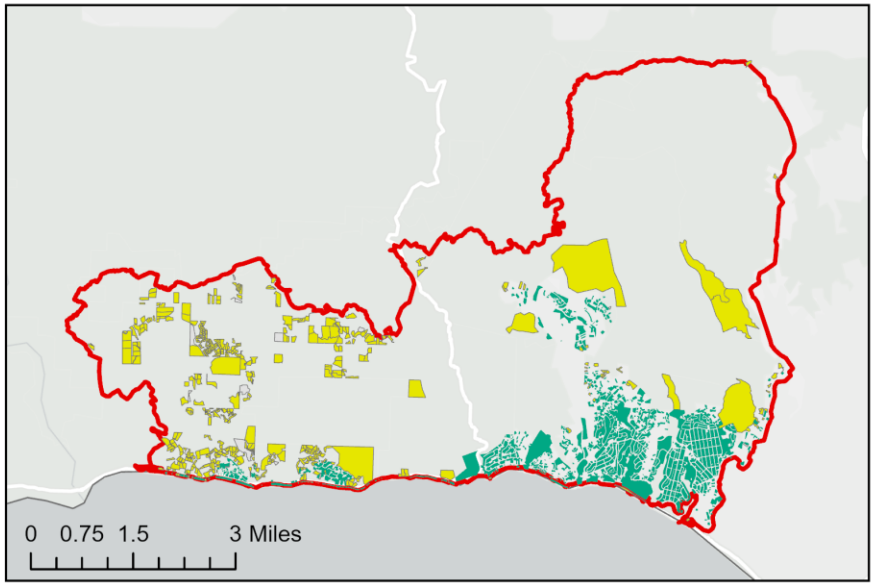
Why does it matter?

Use Case	Need for Wastewater Inventory	States
Emergency / Disaster Response 	Evaluating prevalence of OWTS in communities affected by natural disasters; impact on drinking water	FL, MA, NC
System Failure and Risk to Water Supply 	Evaluating OWTS prevalence in communities with high groundwater use for drinking	CA, FL, NC, VA
Nutrient Loading / Coastal Concerns 	Quantifying contribution of OWTS to nutrient loading and resulting environmental issues on the coast	FL, NC, VA
Advocacy and Funding 	Directing advocacy and funding to help communities and individuals with maintenance or upgrades	CA, NC, VA
Asset Management, Consolidation, Urban Planning 	Identifying areas for potential consolidation into sewers, growth planning, measuring access	CA, FL, MA, NC, VA
Government Agency Communication 	Streamlining data sources and information about OWTS across agencies	CA, FL, NC, VA

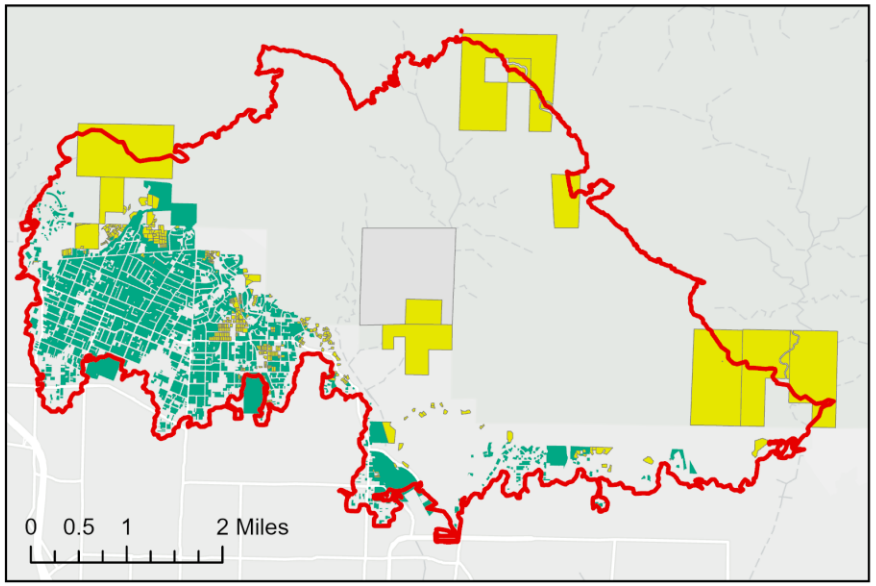
Fire Impacts on Wastewater Systems in Los Angeles County

Wastewater Infrastructure Type
- Non Applicable
- Onsite
- Sewer
- Fire Affected Boundary
- LA City Boundary
- LA County Boundary





Palisades Fire
496 Onsite Served Parcels with Damaged Structures
6,510 Sewer Served Parcels with Damaged Structures



Eaton Fire
385 Onsite Served Parcels with Damaged Structures
6,391 Sewer Served Parcels with Damaged Structures

Quantifying Risks to New England Aquaculture

Conclusion

- It’s been a long journey on this research topic
- We’ve built a robust modeling and data processing infrastructure
- There is room for improvement in the national inventory
- The dataset has many potential uses

Acknowledgments

- California State Water Resources Control Board

- Florida Department of Health
- Virginia Department of Health

- Collaborators: Emily Kumpel, Jay Taneja

Questions