



Machine learning based estimates of wastewater infrastructure coverage across the United States

Nelson da Luz, PhD

Research Assistant Professor

University of Massachusetts Amherst

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

Background

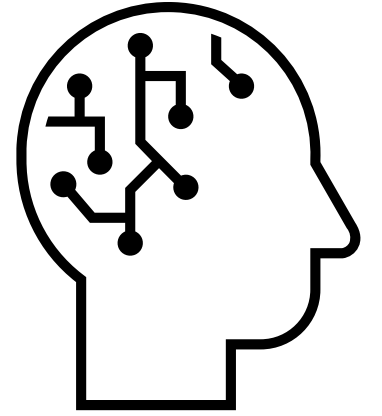
- Safe management of human waste is crucial for ensuring human and environmental health
- At least 20% of the US's population likely use on-site wastewater treatment systems (OWTS)
- The last census of the prevalence of sanitation systems in the US was in 1990
- As a result, there is a significant gap in our understanding of the number, locations, and density of OWTS across the country



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

How can we find out where on-site systems are?

- Our idea: Most sanitation systems are ‘invisible’ (buried) but inferable
- Geospatial on-site and/or sewer data exists in some locales



How can we infer sanitation systems?

- We can fill the gaps by using other indicators that help us infer coverage
- We can leverage these other indicators with machine learning techniques
- Machine learning is part of the AI field and allows analysis of massive quantities of data through processes like pattern matching
- We can use different ML methods depending on the part of the problem we are trying to solve

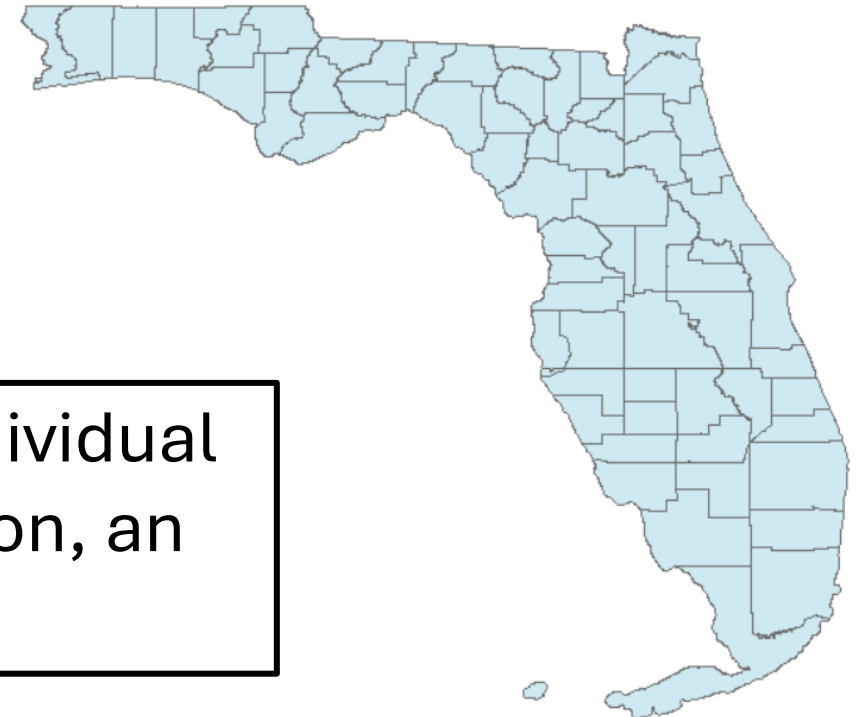
Why do this?

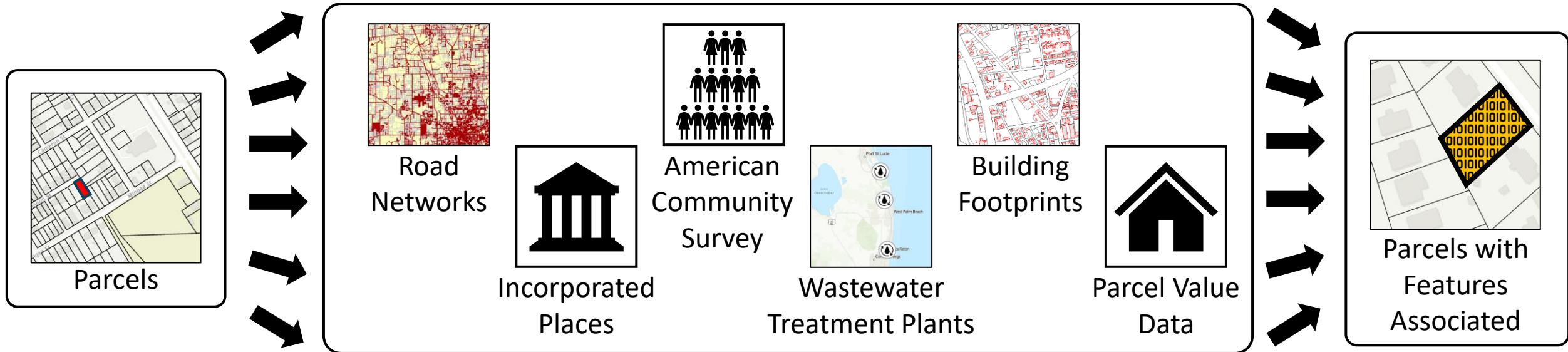
Use Case	Need for Wastewater Inventory	States
Emergency / Disaster Response	Evaluating prevalence of OWTS in communities affected by natural disasters; impact on drinking water	FL, MA, NC
System Failure and Risk to Water Supply	Evaluating OWTS prevalence in communities with high groundwater use for drinking	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

Proof of Concept: Florida

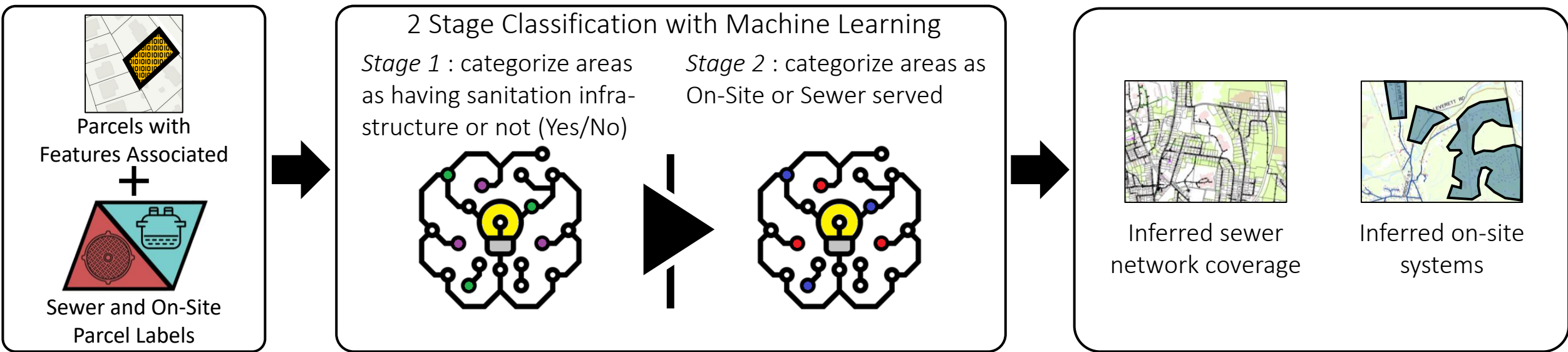
- Florida Water Management Inventory
 - Great source of ground truth data
- Counties of differing population density and sanitation infrastructure types

Goal: Predict whether an individual parcel has a sewer connection, an on-site system, or neither





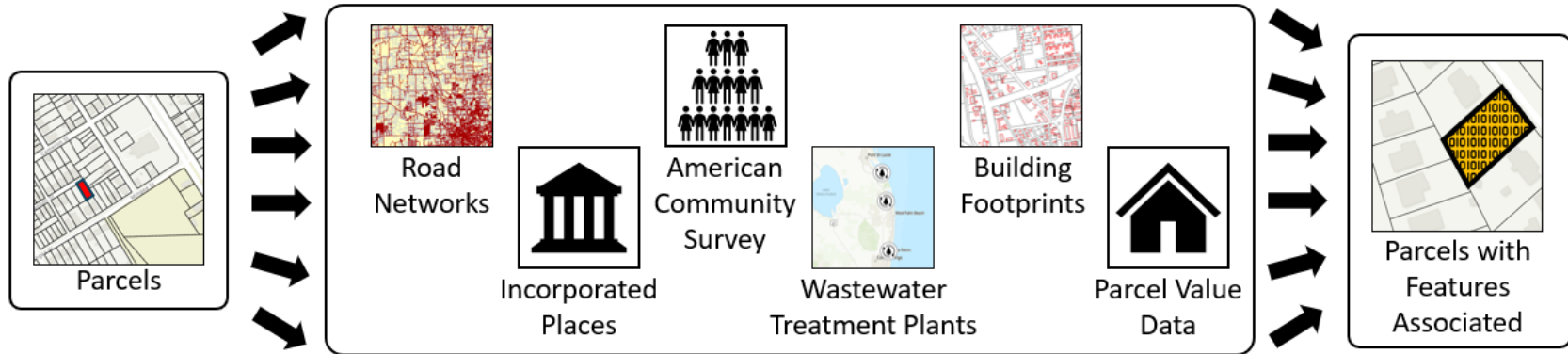
1. We assign features (characteristics) to individual parcels based on these datasets.



2. We train machine learning (Random Forest) models to make the above classifications using the data enriched parcel information.

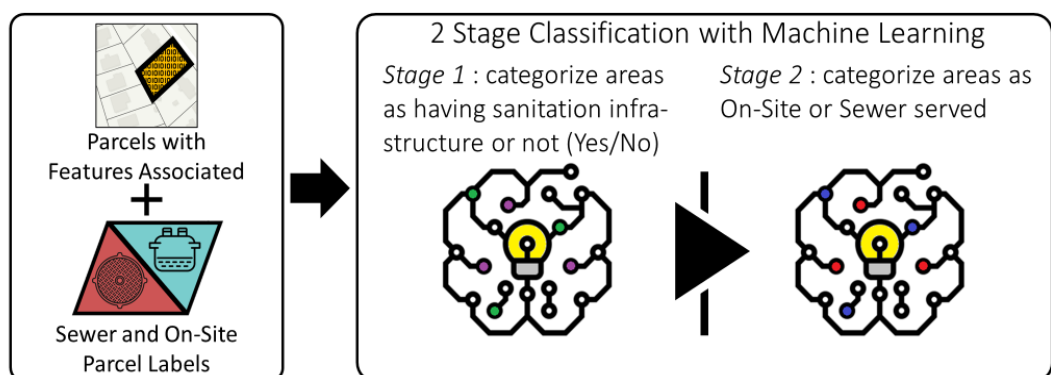
3. Using the trained models we make predictions across areas of interest.

Application in Florida



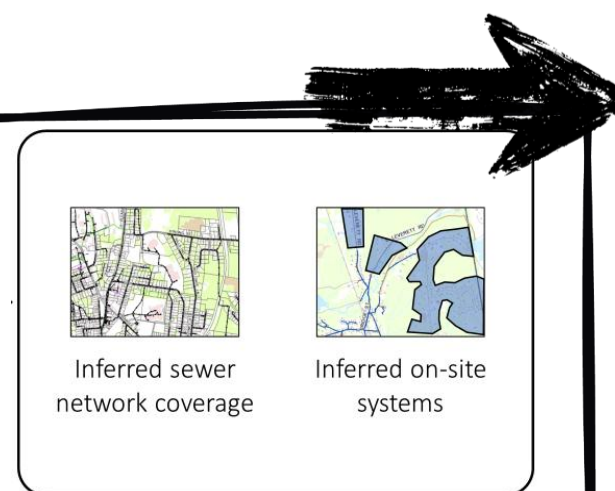
1. We assign features (characteristics) to individual parcels based on these datasets.

- Datasets processed for every county
- We used nationally available datasets to create the characteristics needed



2. We train machine learning (Random Forest) models to make the above classifications using the data enriched parcel information.

We trained a model that uses training data from Florida (labels from FLWMI)



3. Using the trained models we make predictions across areas of interest.

We made predictions of sanitation coverage across Florida by inputting the data we calculated in Step 1 into the model from Step 2!

What does the model output for each parcel?

1. Inferences

- **Sewer** or **On-Site**

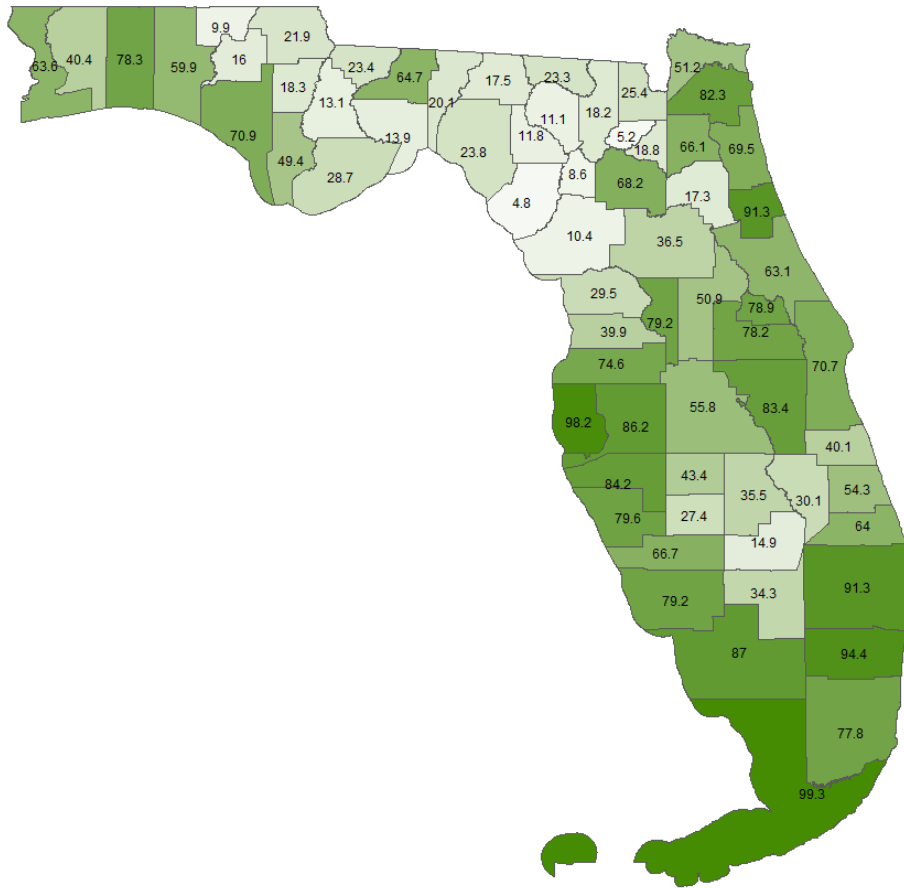
2. Confidence

- How the model expresses certainty about what class the input data belongs to
- e.g., On a scale of 0 to 100 how confident are we that this parcel is served by **Sewer** or **On-Site**

- Accuracy is not a model output, but we can calculate accuracy if we compare to ground truth data
 - Accuracy is a measure of how many inferences are correct out of a total number of inferences that are made

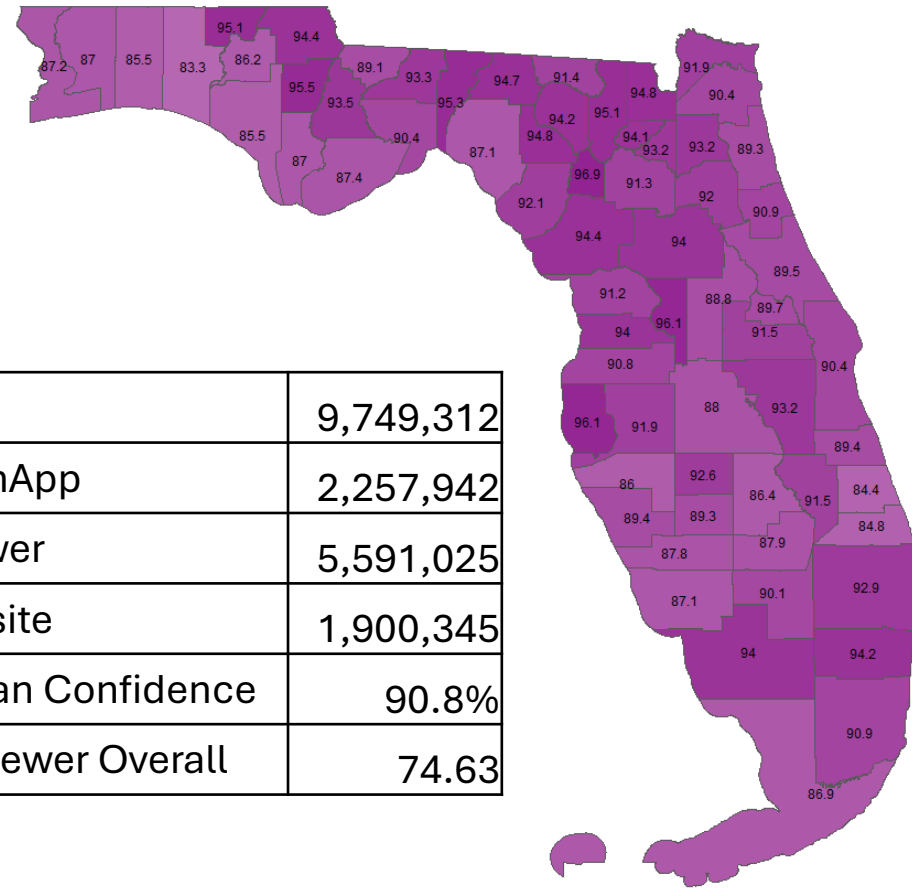


Accuracy: We can look at the target and confirm that we hit the target in the bullseye
Confidence: Even though we can't see the target, based on what we know about other targets, we are 85% sure we hit the bullseye



% Sewer 0 25 50 75 100

1. Inferences: Of all the parcels served by sanitation infrastructure in each county, this % are served by sewer

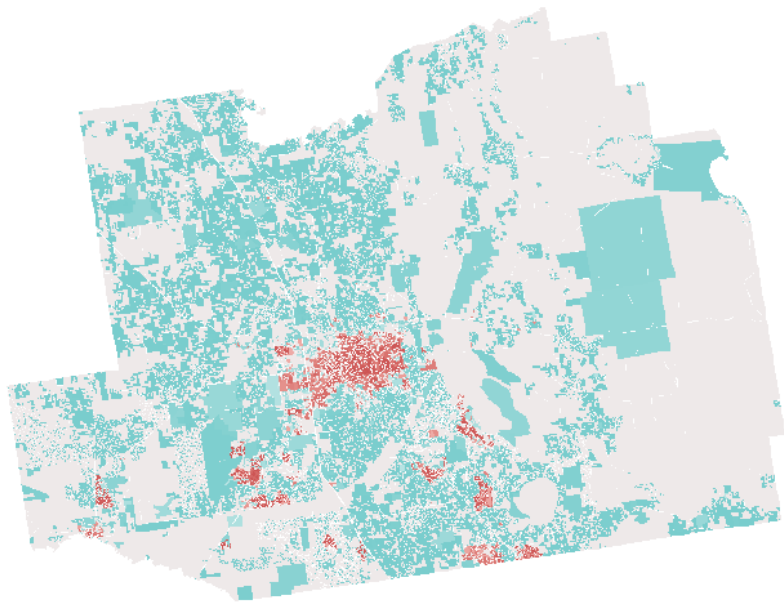


Confidence % 50 60 70 80 90 100

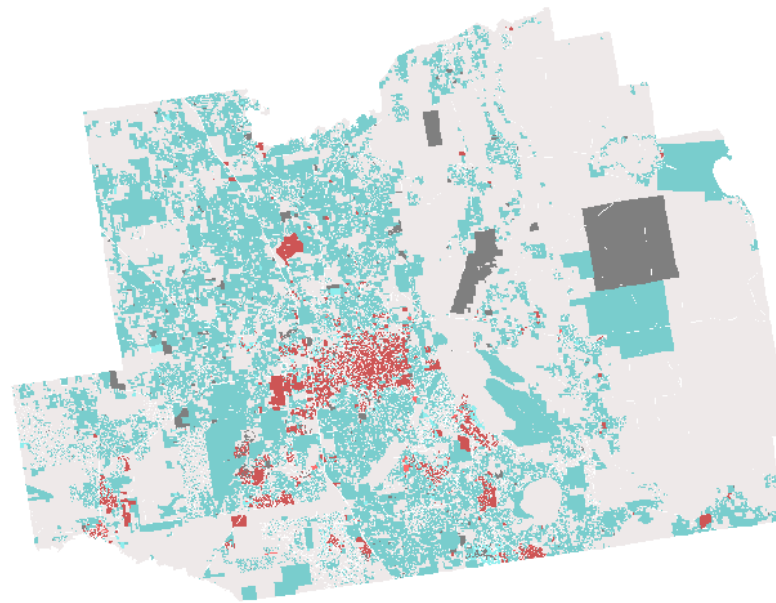
2. Confidence: The average confidence of all the sanitation type inferences that were made in each county

Marion County, FL

Our Prediction
(98.3% Accuracy*)



Ground Truth



- The model can identify the sewer system in Ocala
- It slightly underestimates the total sewer service area
- *comparing against KnownSeptic and KnownSewer labels



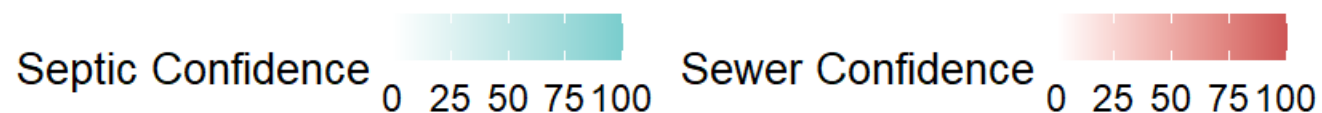
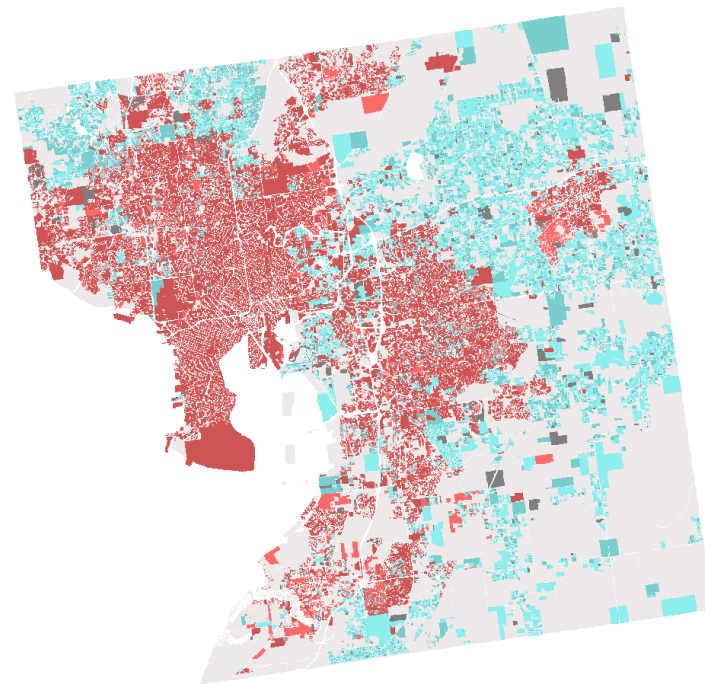
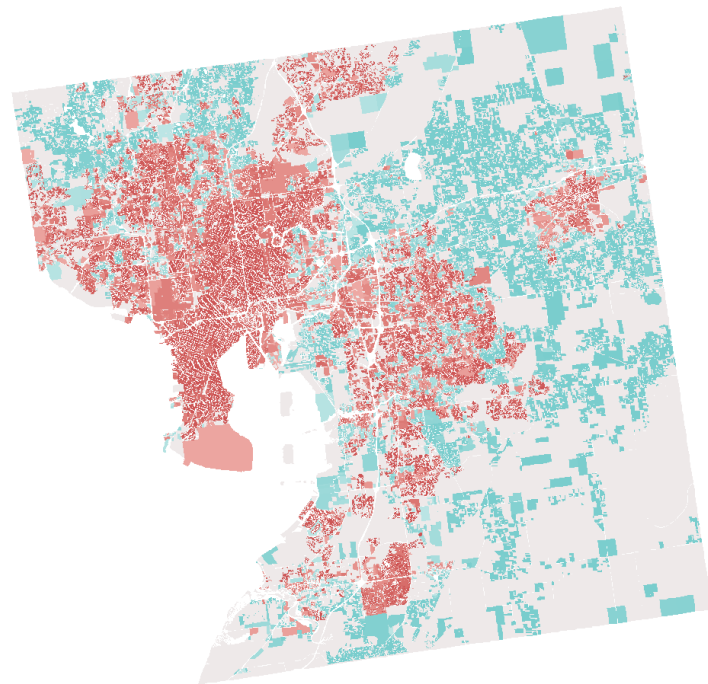
	Known Septic
	Known Sewer
	Likely/SWL Septic
	Likely/SWL Sewer
	Not Applicable
	Unknown/ UNDT

Hillsborough County, FL

Our Prediction
(97.9% Accuracy*)

Ground Truth

- The model can identify the sewer system in Tampa
- It underestimates the total sewer service area
- It assigns confident septic labels in places where FLWMI indicated Likely or SWL Septic
- *comparing against KnownSeptic and KnownSewer labels

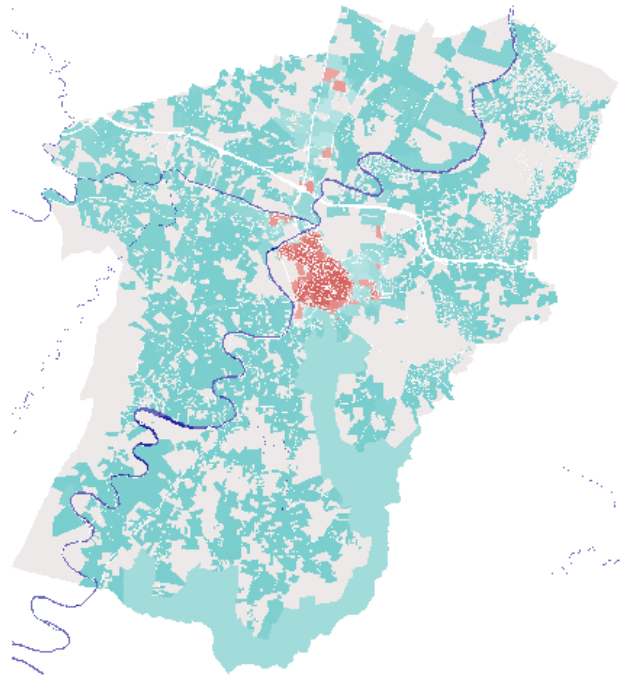


	Known Septic
	Known Sewer
	Not Applicable
	Unknown/ UNDT
	Likely/SWL Septic
	Likely/SWL Sewer

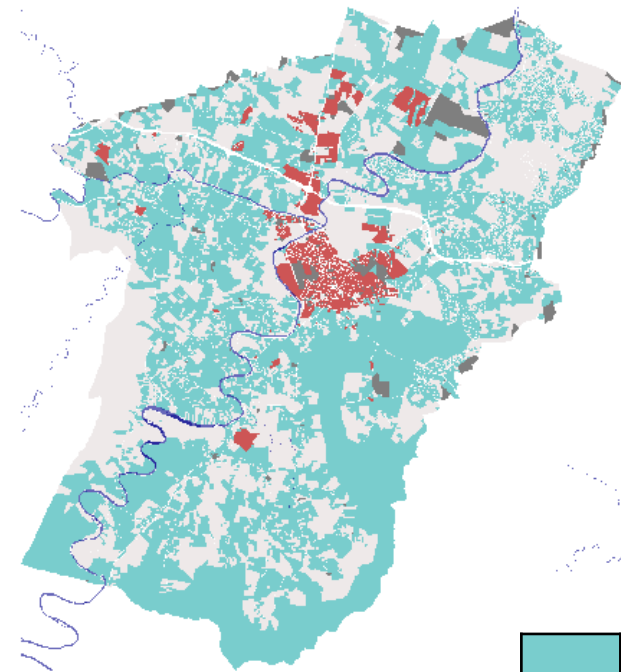
Warren County, VA





- The model can identify the sewer system in Front Royal
- It underestimates the total sewer service area

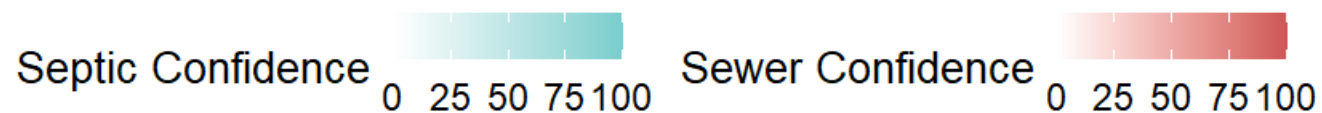
Our Prediction (84.9% Accuracy)



Ground Truth



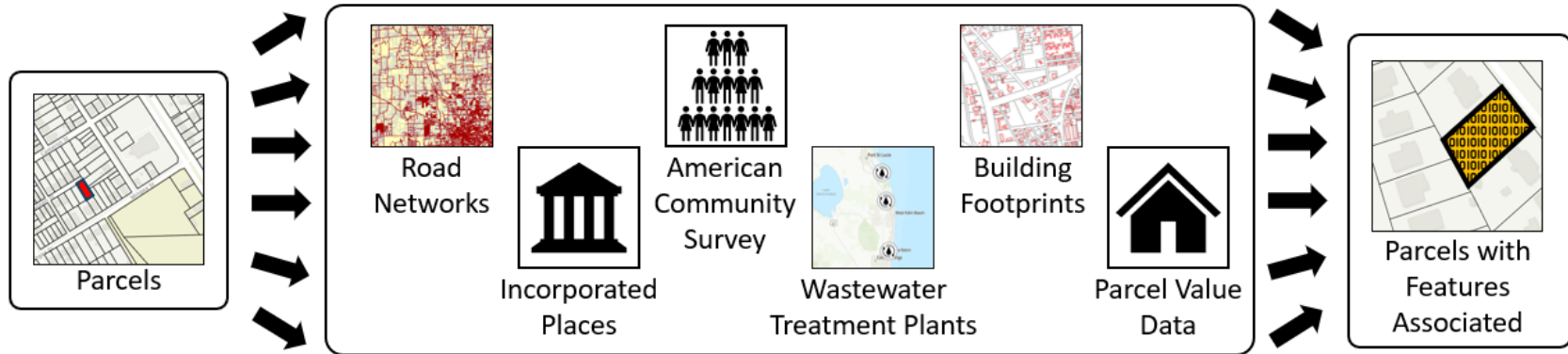
	Septic
	Sewer
	Not Applicable
	Unknown



What about other places?

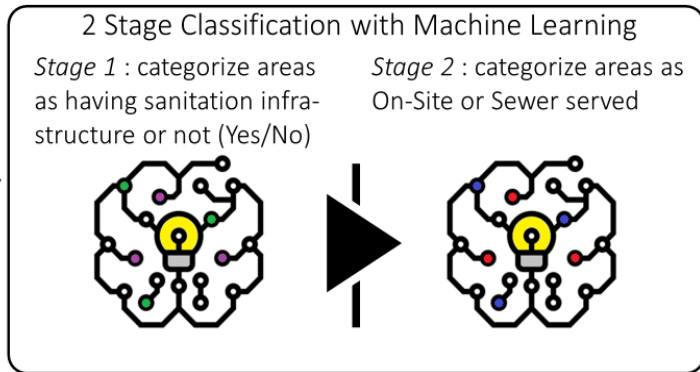
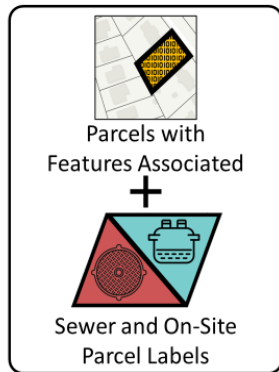
- So far, we have a mix of onsite and sewer data from a few states:
 - Florida
 - Virginia
 - Georgia
 - Nebraska
 - Tennessee
 - California
- How does adding more data help the model in new places?
- What is the value of having data in the place where we are making estimates?

Application in California



1. We assign features (characteristics) to individual parcels based on these datasets.

- Datasets processed for every county
- We primarily used nationally available datasets to create the variables needed*
- *the wastewater treatment plant data comes from CA



2. We train machine learning (Random Forest) models to make the above classifications using the data enriched parcel information.

- We trained three sets of models:
 - using data from FL only
 - using data from FL, VA, NE, TN, and GA
 - Using the above states as well as CA

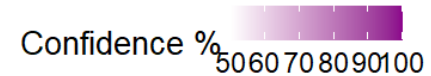
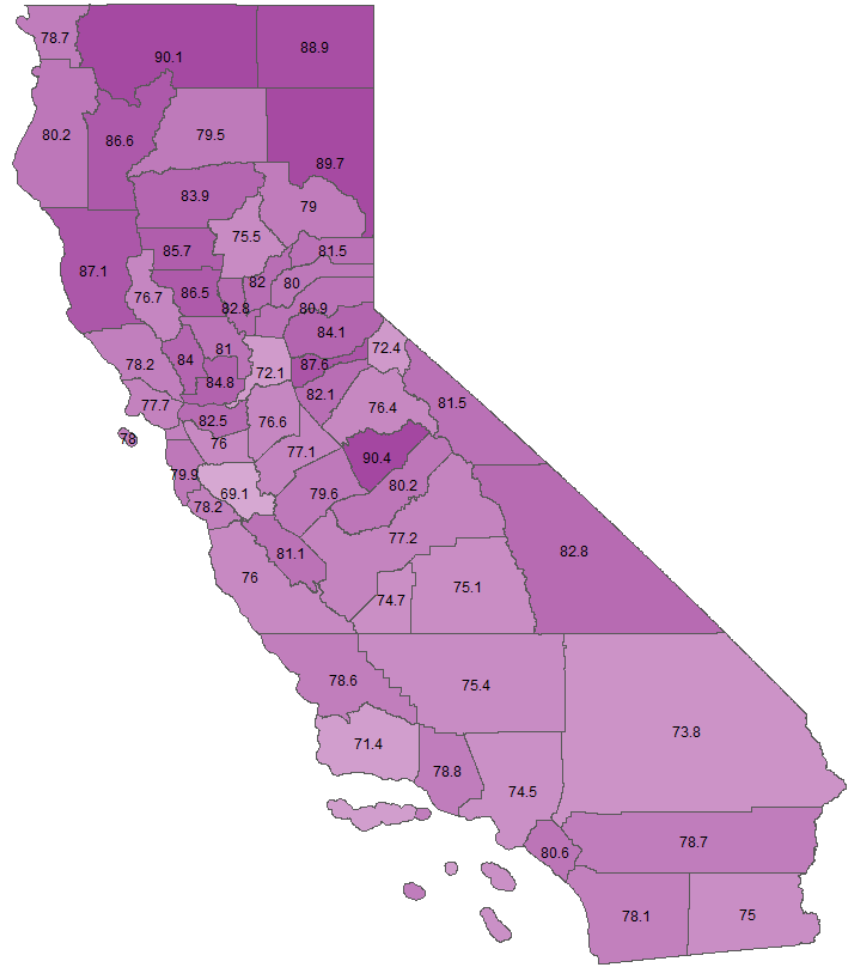
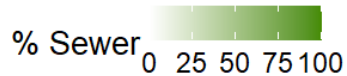
CA Results

- Using CA data improves performance in the completely held out test set of 23k parcels (all from CA)
- A little goes a long way: Only 1.5% of the S1 training data and 2.25% of the S2 training data comes from California, but the overall effect is an increase in accuracy of 16.5% (72% to 88%!)

	Accuracy (%)		
	Only FL data	Without CA Data	With CA Data
Overall	57.4%	71.8%	88.3%
States used for training	FL	FL GA TN NE VA	FL GA TN NE VA CA
States used for testing	CA	CA	CA






n	12,740,568
NonApp	1,101,208
Sewer	7,636,690
On-Site	4,002,670
Confidence	79.9%
% Sewer	65.6



1. Inferences: Of all the parcels served by sanitation infrastructure in each county, this % are served by sewer

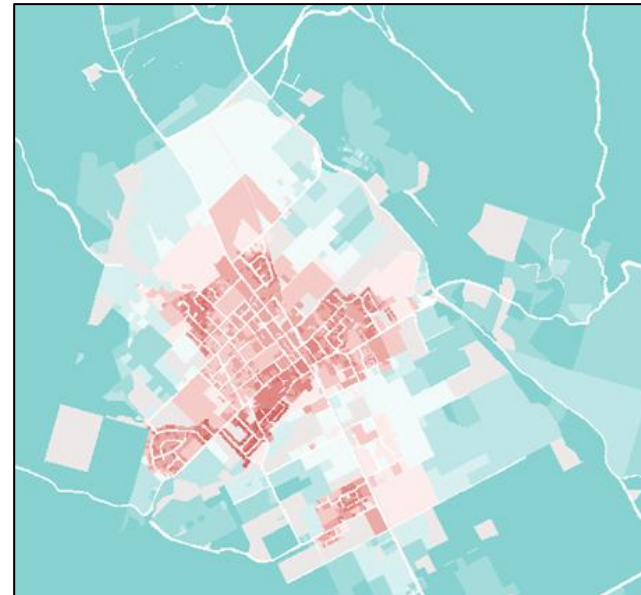
2. Confidence: The average confidence of all the sanitation type inferences that were made in each county

Napa County

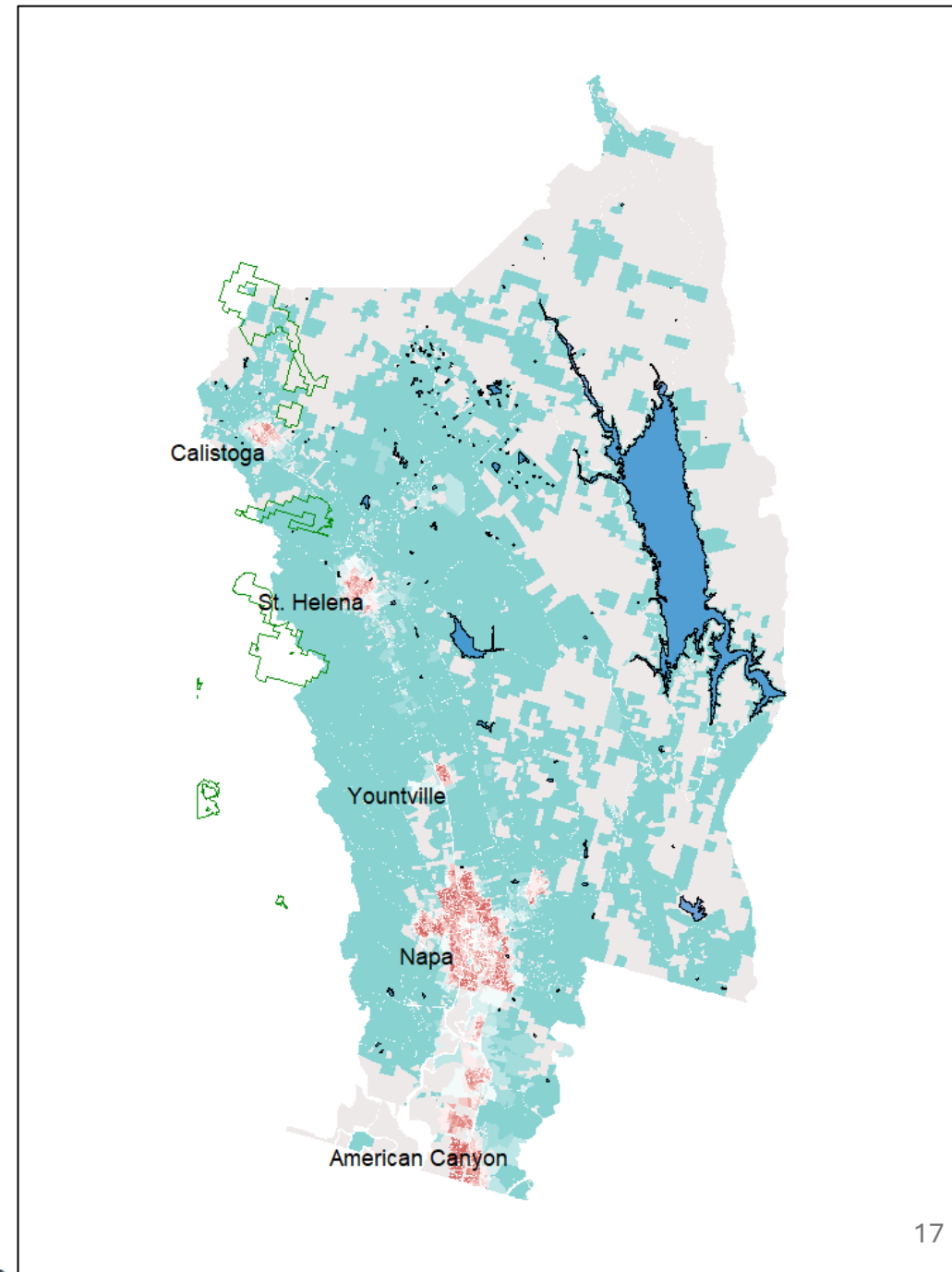
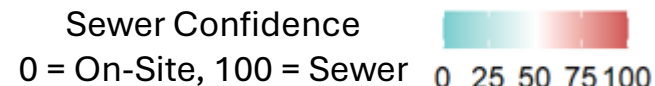
	On-Site
	Sewer
	Not applicable

- The model predicted sewer in the major incorporated cities in Napa County

NAME	Napa
Total	50,191
NonApp	3,910
Sewer	32,992
On-Site	13,289
Confidence	84%
% Sewer	65.7
% On-Site	34.3

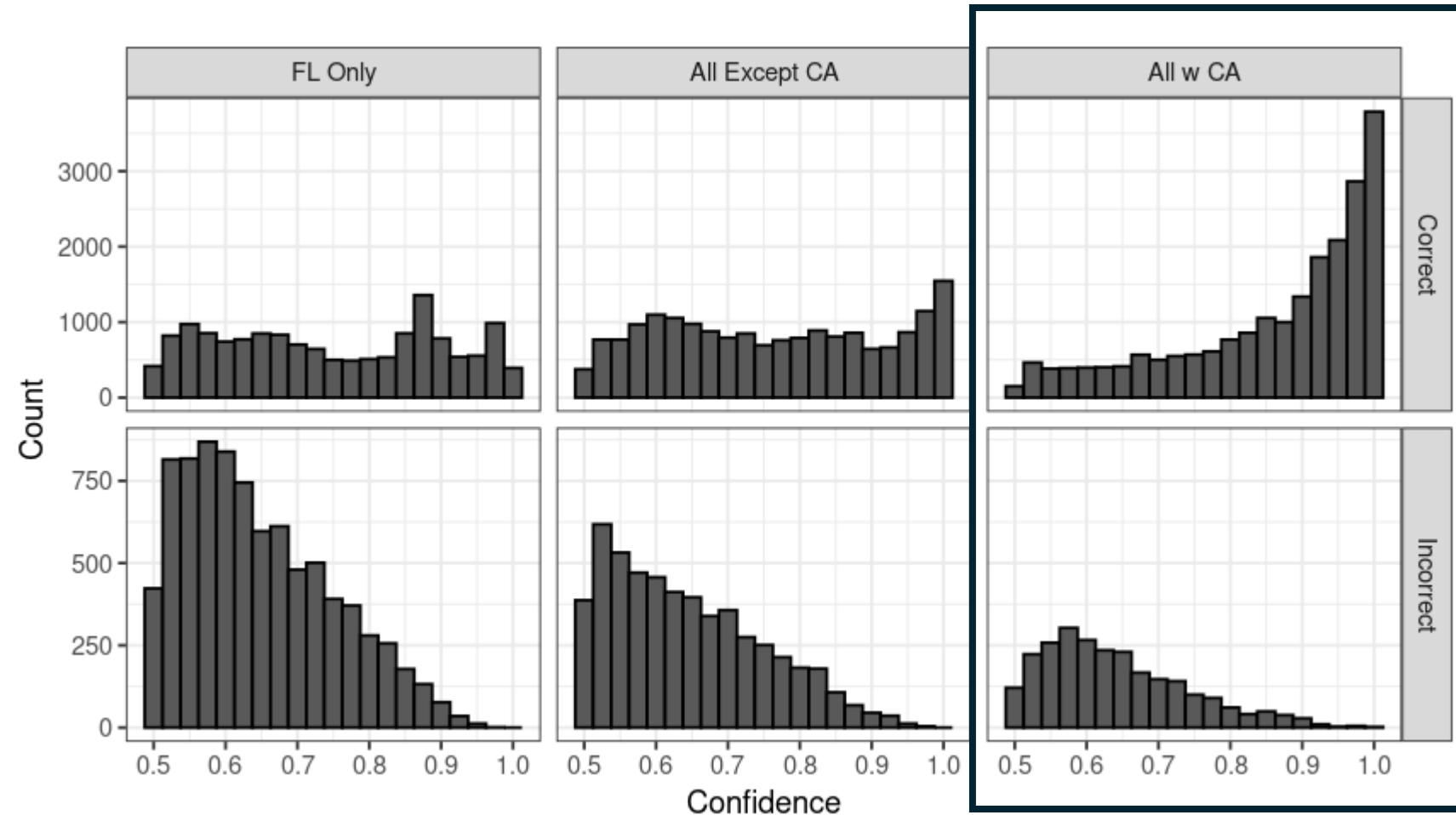


Around St. Helena



Usefulness of Model Confidence Metric

- Distributions of model confidence in the test set
- Focusing on the model trained with CA data (boxed right)
- When the model is confident, it is often correct (& vice versa)
- This suggests that it would be useful to collect data from places with lower confidence to improve the model



Harnessing *Street View Imagery* and *Computer Vision* for Utility Mapping



- **Why do this:** If we produce labels inferred through computer vision, we can use them to train the 2-Stage model and improve results in targeted places where there is no sanitation system data otherwise and the 2-Stage model has lower confidence
- **Street View as a Resource:** We can use Google Street View imagery for large-scale urban data gathering
- **Computer Vision Capabilities:** We employ sophisticated algorithms for the detection and classification of utility hole covers
- **Sewer System Mapping:** We use detected locations to pinpoint and map out sewer networks and on-site service
- **Classification Efficacy:** Achieves greater than **80%** accuracy rate in correct parcel classification as sewer or on-site.

Conclusions

- We developed a 2-stage machine learning model for estimating coverage of wastewater infrastructure
- The model generates inferences and confidence values
- Results are promising in states like Florida, Virginia, and California
- Model performance improves as you add more data from a variety of places
- Inventories of wastewater infrastructure can be useful for a variety of purposes including asset management, risk assessment and disaster planning, and evaluating nutrient loads in groundwater

Next Steps

- Incorporating more 'labels' from different states
- Developing system for sharing model predictions with end-users
 - Active collaboration/project with CA Wastewater Needs Assessment
- Looking to build capacity to support federal agencies, state agencies, and non-profit organizations interested in leveraging this type of data