trinnex

Trust in what's next.

Demystify Predictive Modeling

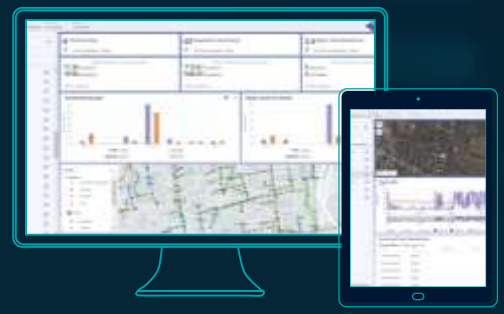
For Service Line Material Prediction

Introductions

Da Yu, P.E.

Account Executive
Trinnex
Cincinnati, OH

trinnex
a CDM Smith company



Digital consulting with domain knowledge

- Risk, Performance & Capital Planning
- Digital Strategy & Transformation
- Decision Analytics & Optimization
- Digital Twin Design & Development

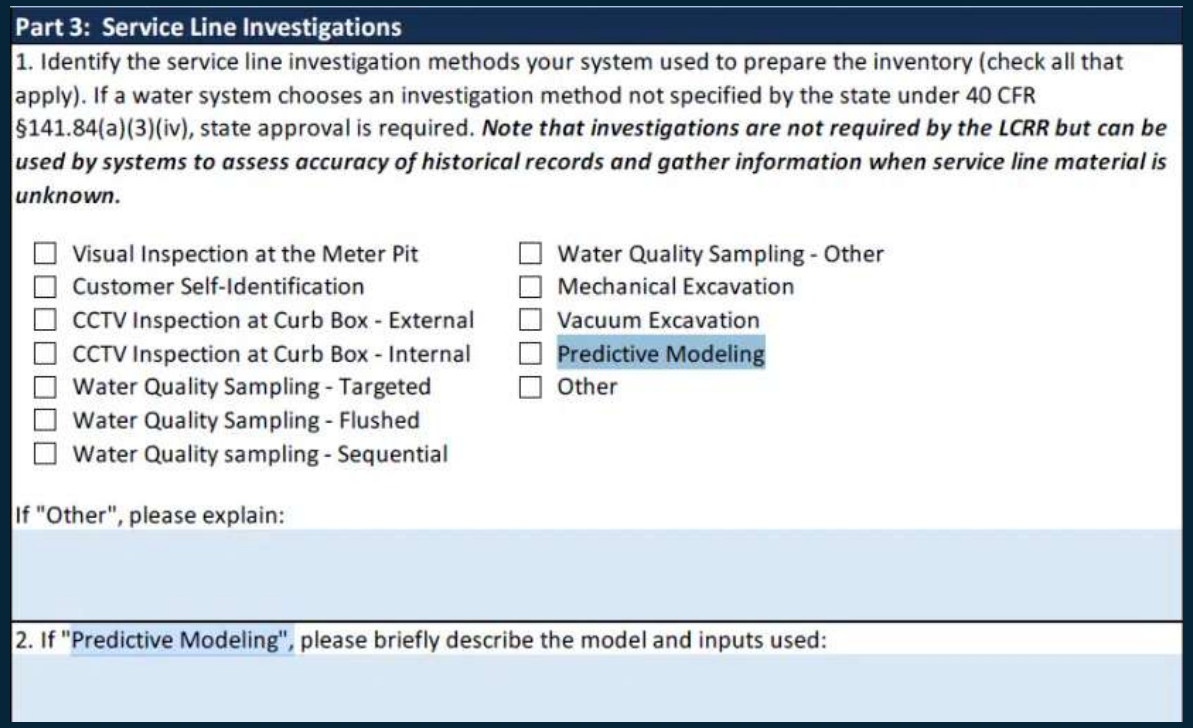
Purpose-built, cost-effective software portfolio

- pipeCAST™
- leadCAST™
- epiCAST™

www.trinnex.io

Machine Learning

- Why?
- What is it?
- How does it work?
- Limitations
- How to use it reliably?



Part 3: Service Line Investigations

1. Identify the service line investigation methods your system used to prepare the inventory (check all that apply). If a water system chooses an investigation method not specified by the state under 40 CFR §141.84(a)(3)(iv), state approval is required. ***Note that investigations are not required by the LCRR but can be used by systems to assess accuracy of historical records and gather information when service line material is unknown.***

- ☐ Visual Inspection at the Meter Pit
- ☐ Customer Self-Identification
- ☐ CCTV Inspection at Curb Box - External
- ☐ CCTV Inspection at Curb Box - Internal
- ☐ Water Quality Sampling - Targeted
- ☐ Water Quality Sampling - Flushed
- ☐ Water Quality sampling - Sequential
- ☐ Water Quality Sampling - Other
- ☐ Mechanical Excavation
- ☐ Vacuum Excavation
- ☐ Predictive Modeling
- ☐ Other

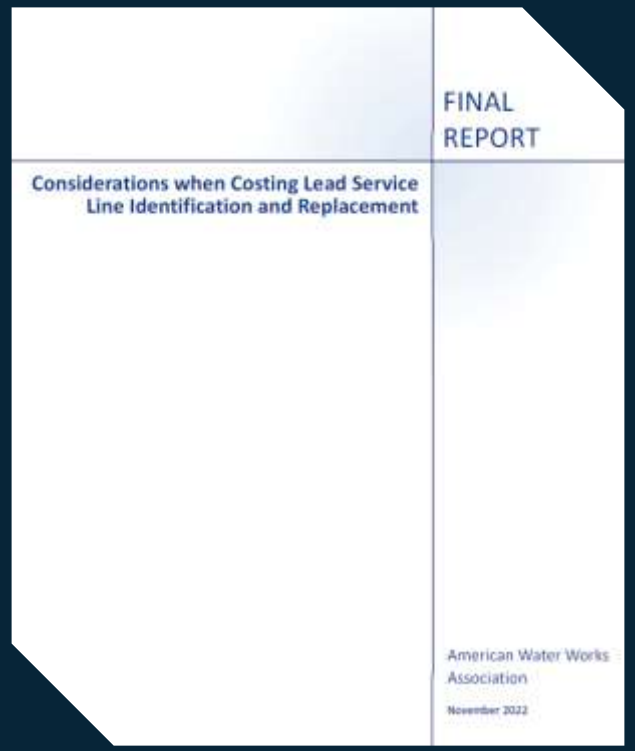
If "Other", please explain:

2. If "Predictive Modeling", please briefly describe the model and inputs used:

Why | Verification Method – Accuracy & Cost

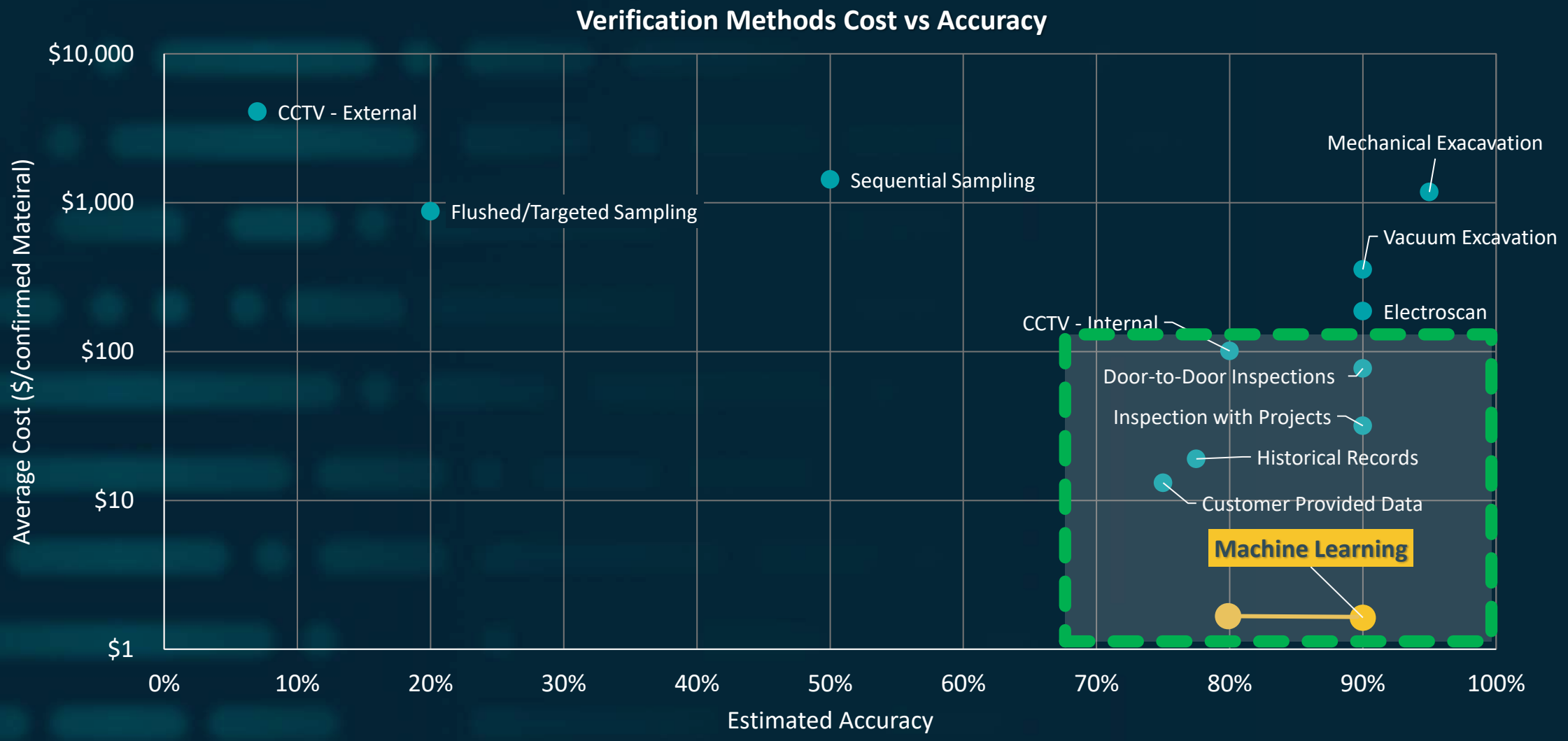
Table 3-10 Average Cost and Accuracy for Service Line Material Verification Methods

Identification Method	Average Cost ($/SL Material Evaluated)	Estimated Accuracy in SL Material Determination	Average Cost ($/Confirmed Material)
Historical Records	$14.13	60%-95%	$14.87-$23.55
Water Quality - Sequential Sampling	$715	Since only used to confirm LSL rather than to confirm non-lead, depends on % of LSLs in system - 50% used for comparison purposes	$1,430
Water Quality - Flushed/Targeted Service Line Sampling	$175	Since only used to confirm LSL rather than to confirm non-lead, depends on % of LSLs in system and effectiveness of corrosion control - 20% used for comparison purposes	$875
Inspections with Past & Current Projects	$28.53	90%	$31.70
Customer-Provided Data	$9.85	75%	$13.13
Door-to-Door Inspections	$69.23	90%	$76.92
Mechanical Excavation	$1,120	95%	$1,179
Vacuum Excavation	$320	90%	$355.56
CCTV/Camera – External	$286	7%	$4,086
CCTV/Camera – Internal	$81	80%	$101
Electro-scan Interior Probe (1,000 inspections)	$168.20	90%	$186.90
Predictive Modeling (after initial ~20% physical confirmation)	$1.30	Highly dependent on input data – possible to get over 90% accuracy (80% accuracy used for analysis purposes)	$1.63



AWWA Report, *Considerations when Costing Lead Service Line Identification and Replacement,* November 2022

Why | Verification Method – Accuracy & Cost

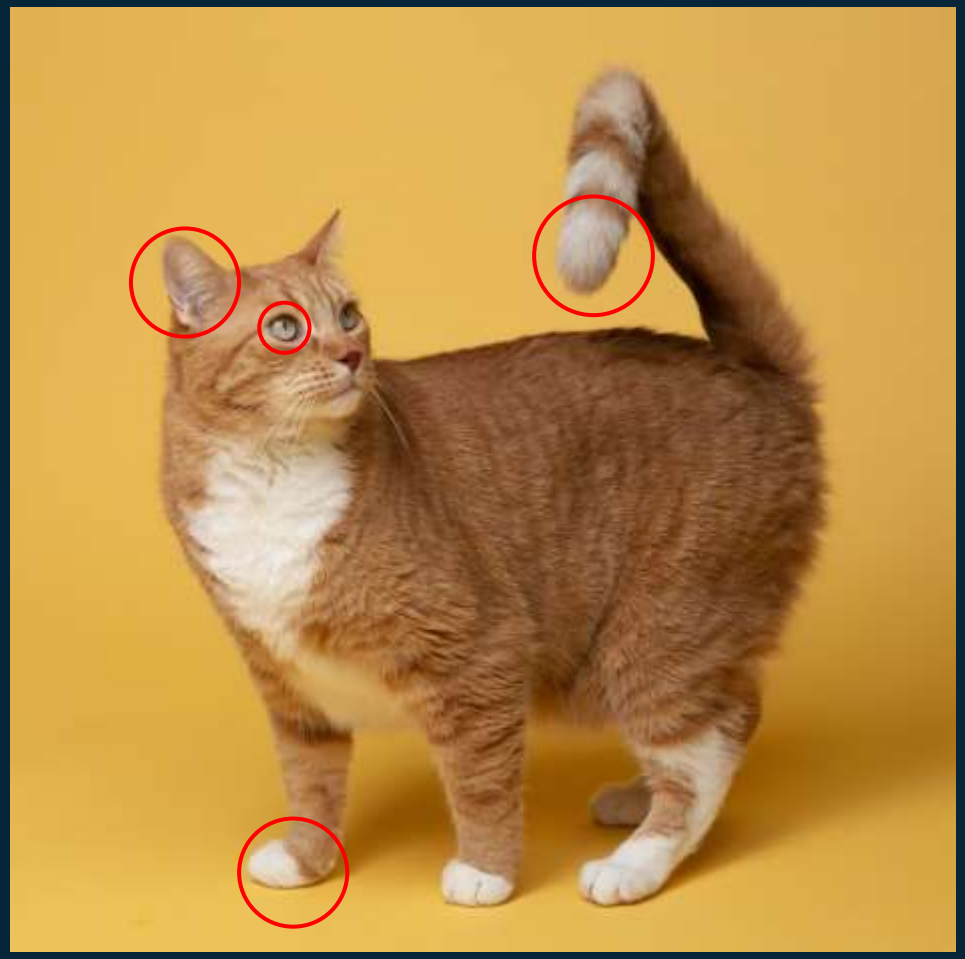


Cats and Dogs

Features

Size
Head
Tail
Ears
Nose
Paws
Color

Cats and Dogs

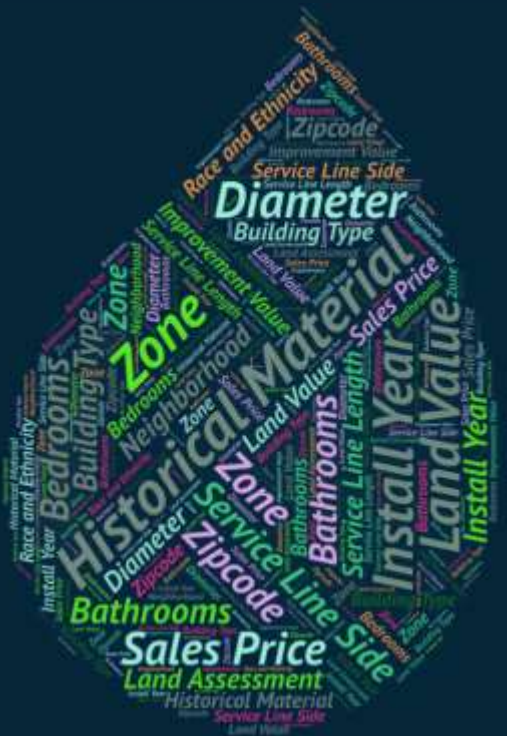


Features

Size
Head
Tail
Ears
Nose
Paws
Color

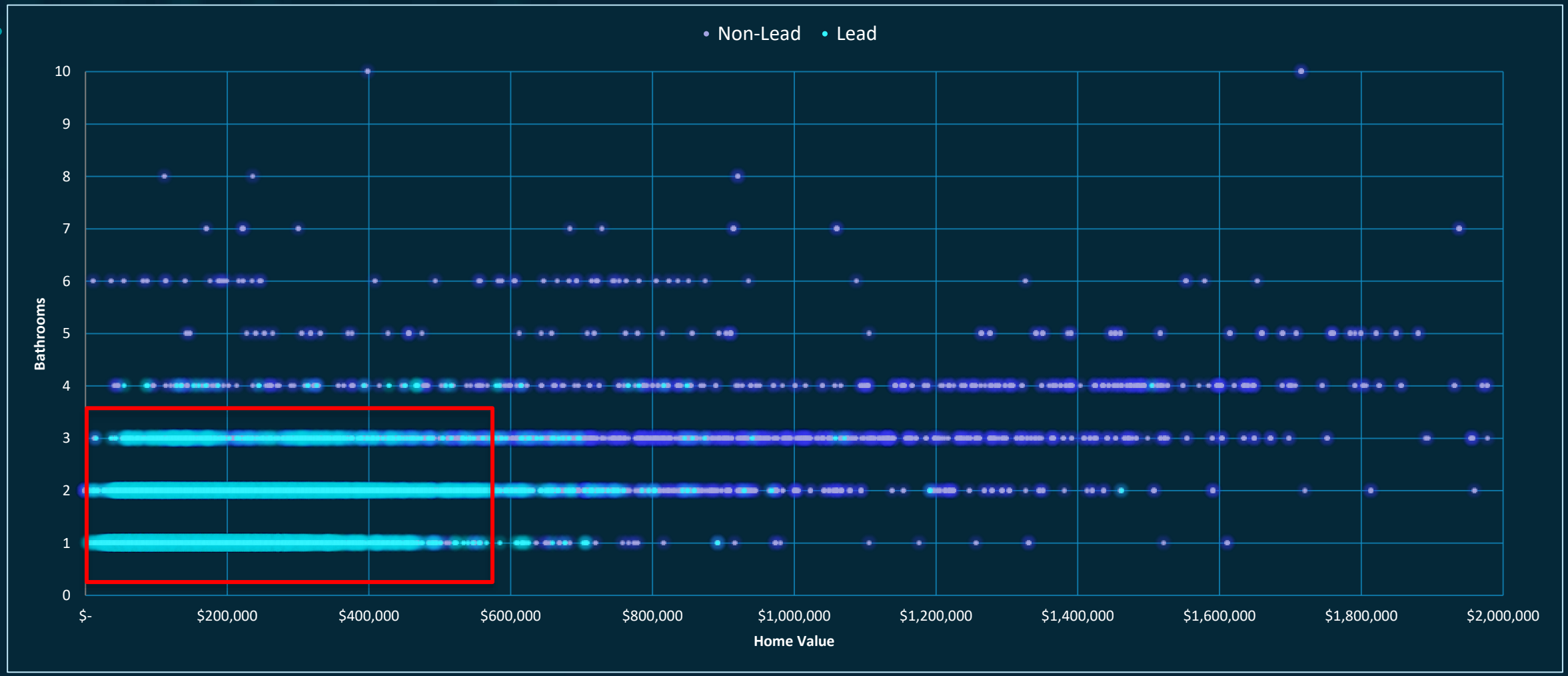
Lead and Non Lead

Lead

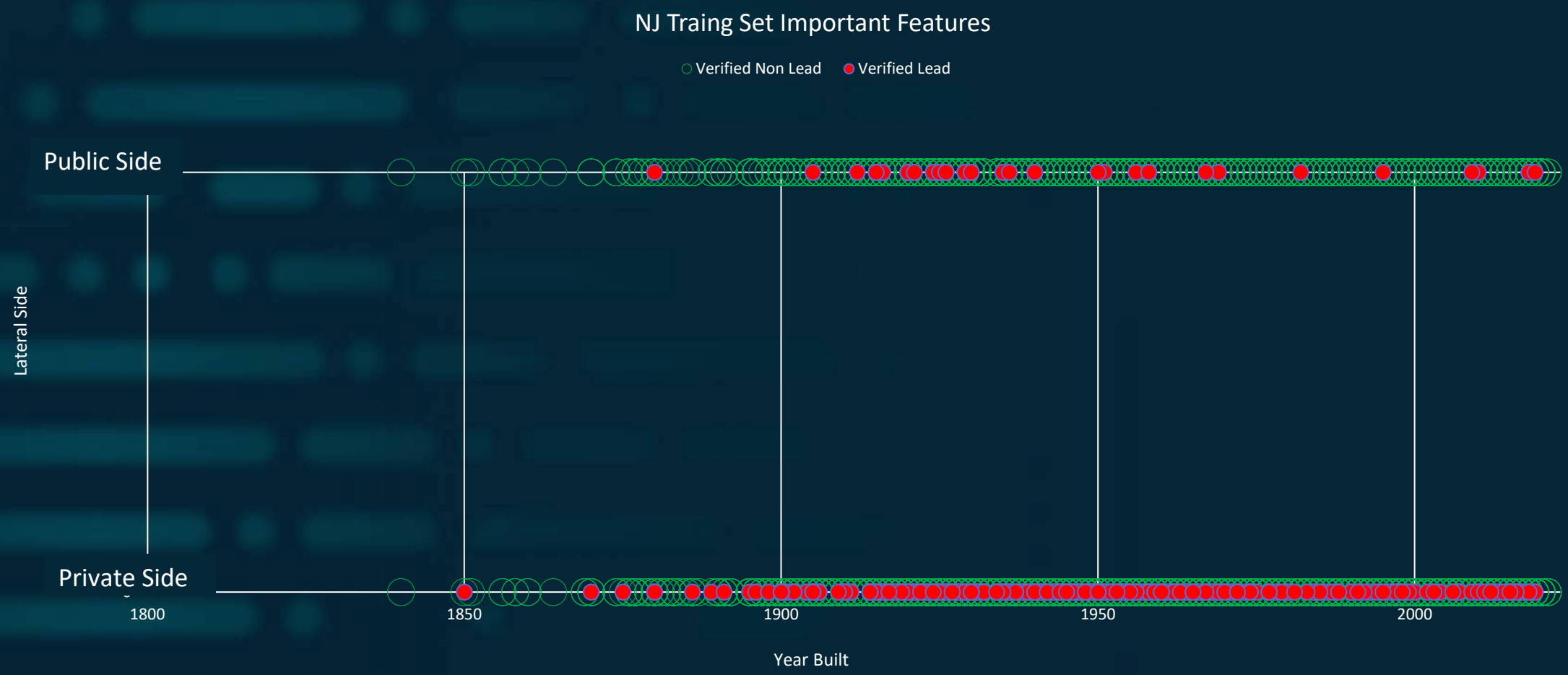


Non Lead

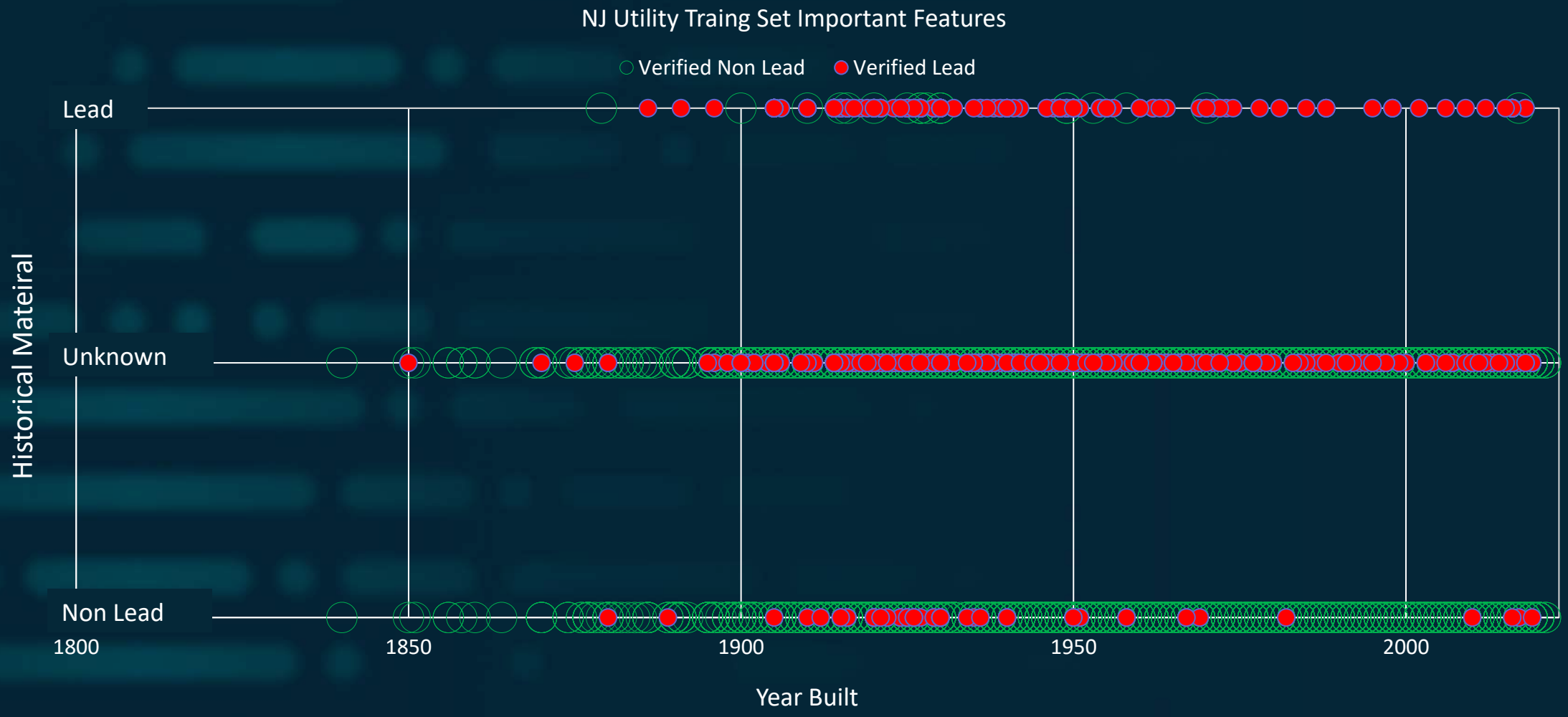
What | Machine Learning Automates the Process

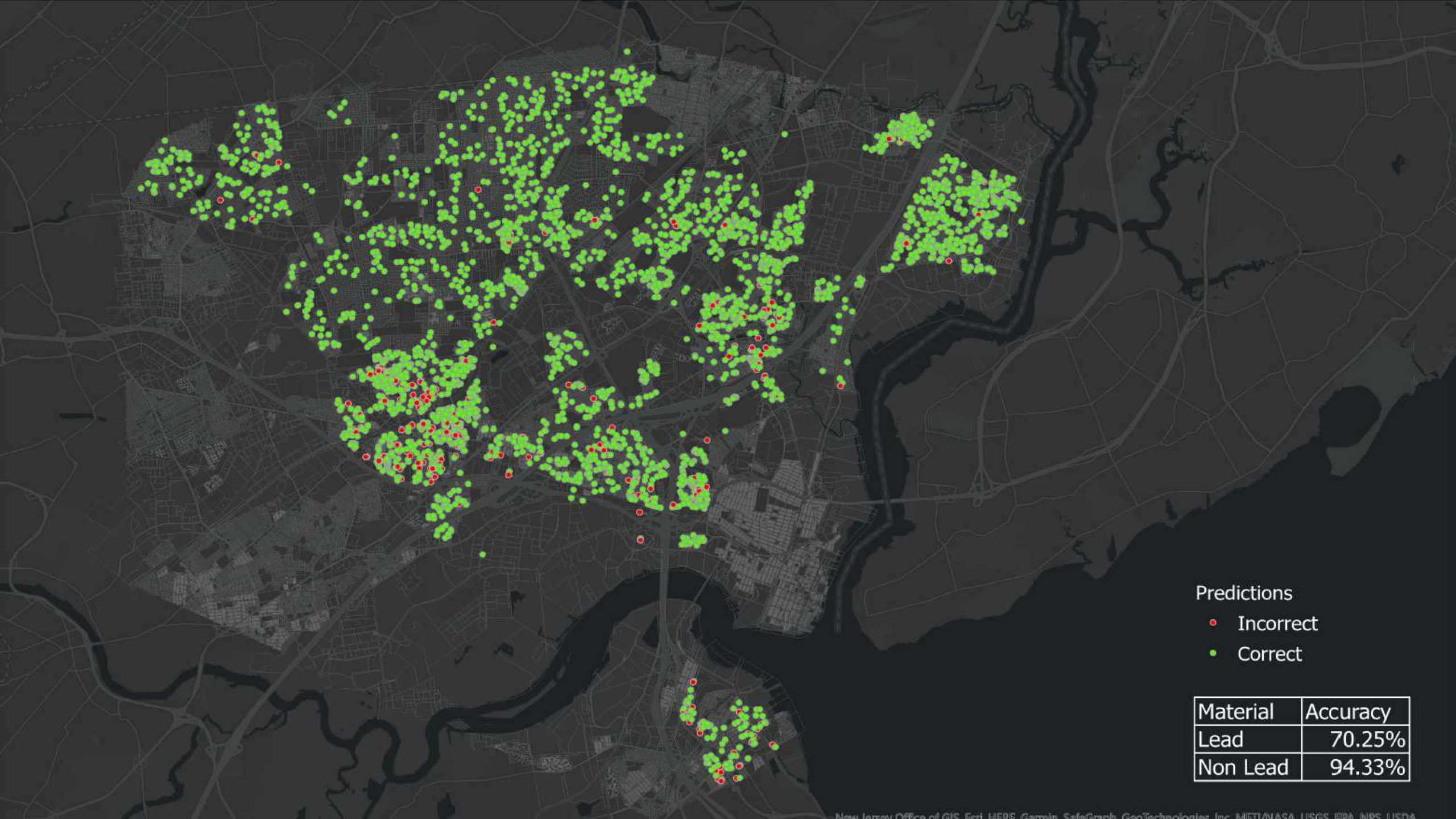


What | Machine Learning



What | Machine Learning

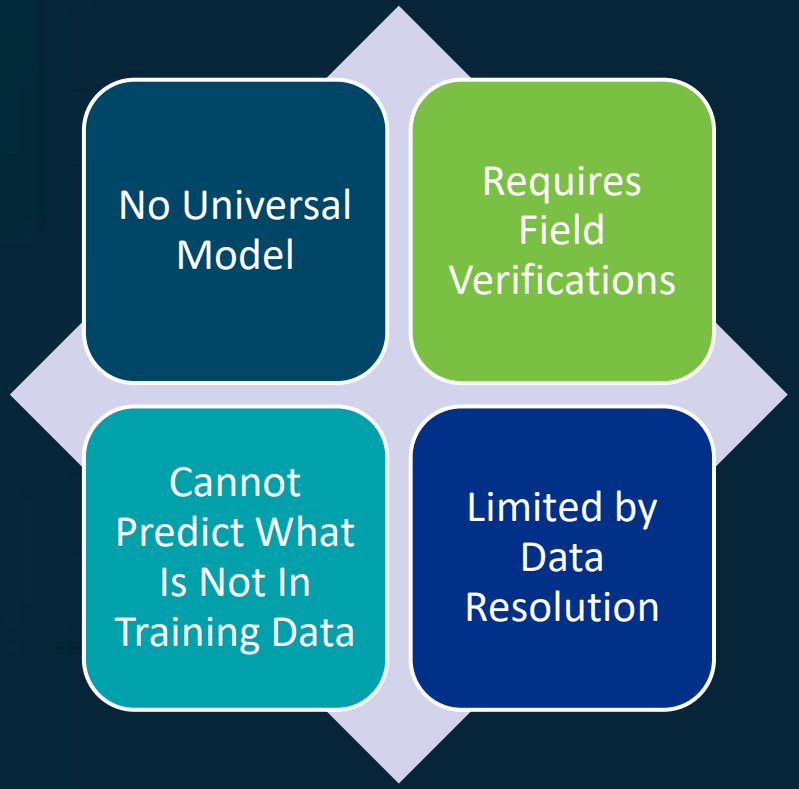



Predictions
Incorrect
Correct
Material | Accuracy
Lead | 70.25%
Non Lead | 94.33%
New Jersey Office of GIS, Esri, HERE, Garmin, SafeGraph, GeoTechnologies, Inc, METI/NASA, USGS, EPA, NPS, USDA

Limitations

Utility A banned lead in 1950
Utility B enacted federal ban in 1988

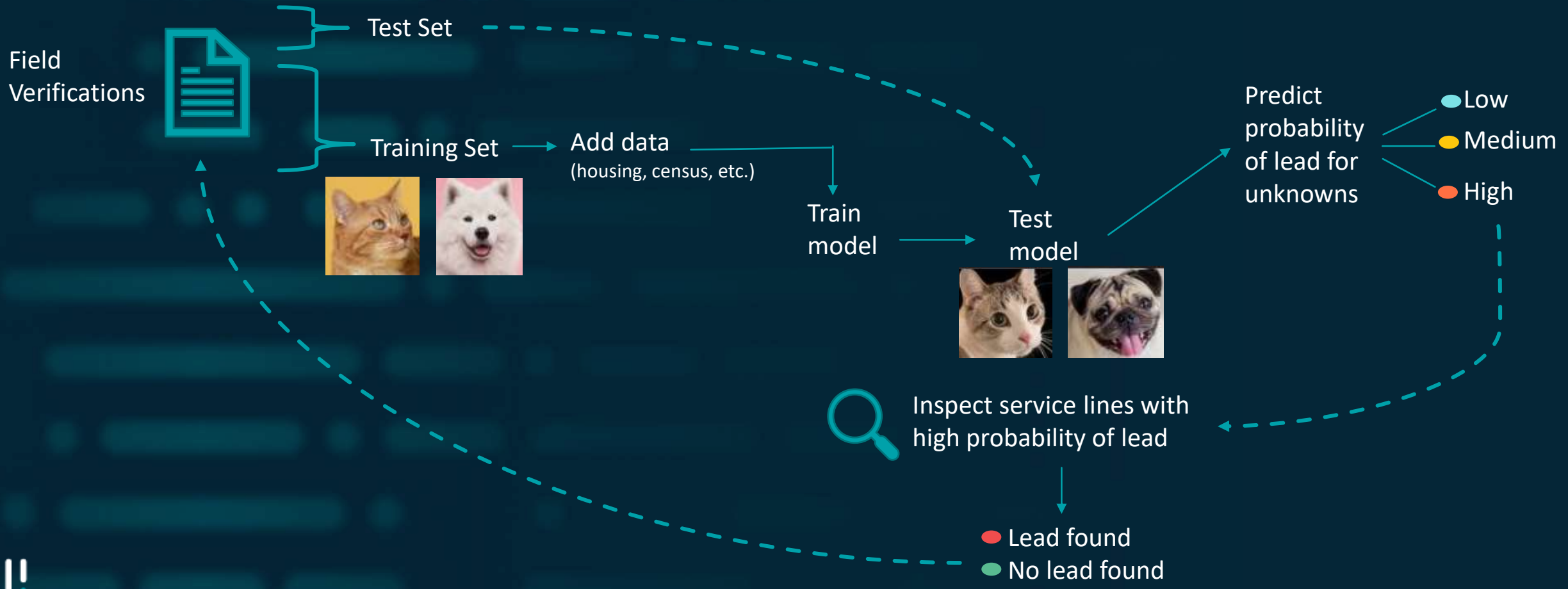
If training/test dataset is 100% copper, model will only predict copper



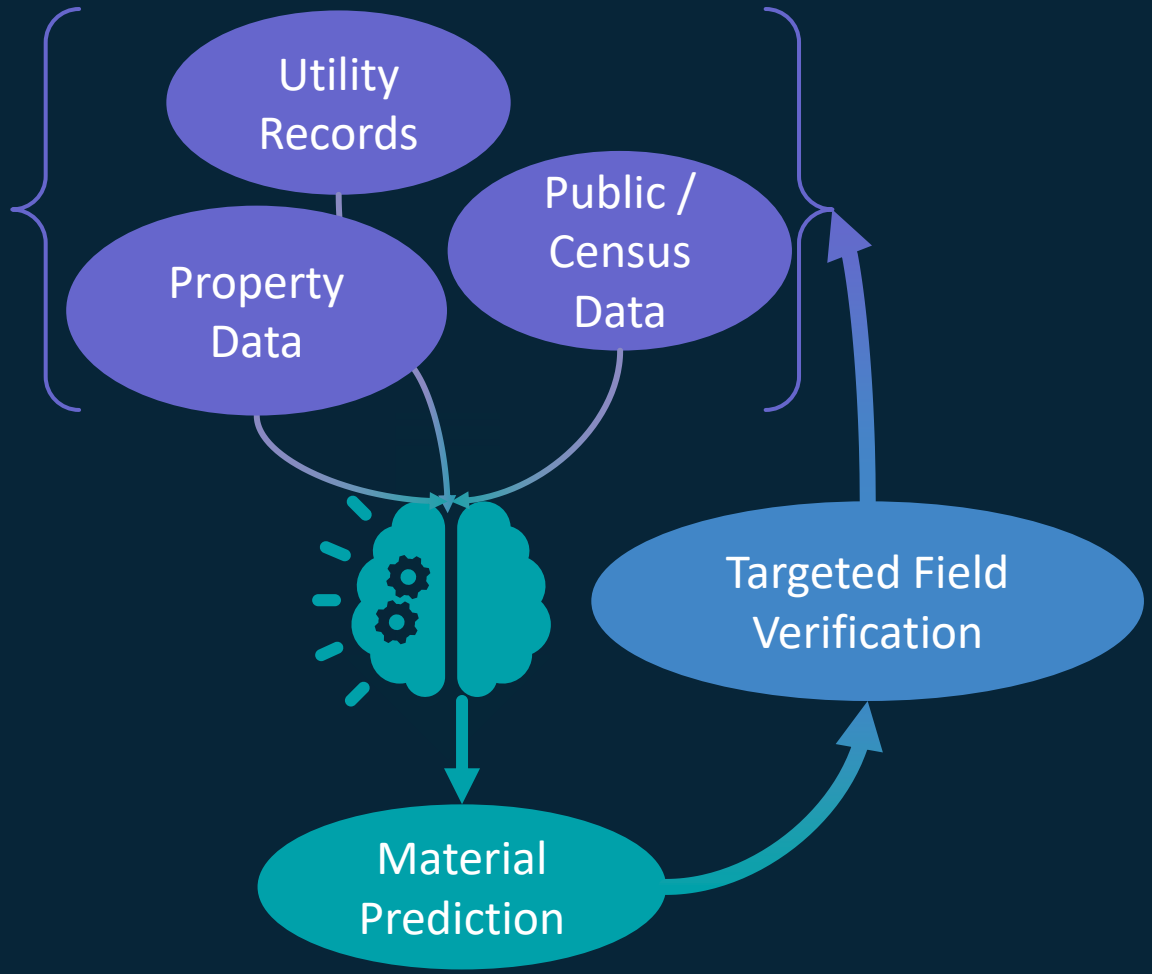
States accepting ML require multi-point field verifications

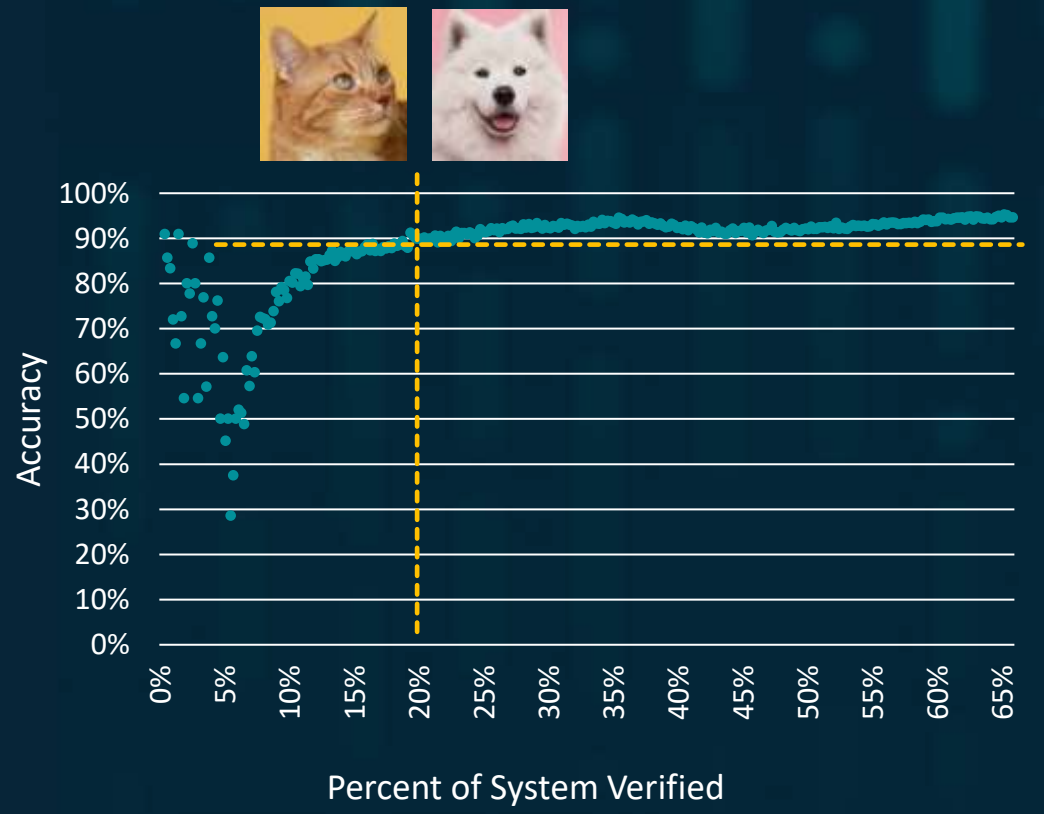
Parameters based on census block are not property specific

How | Machine Learning Process



How | Machine learning requires field data



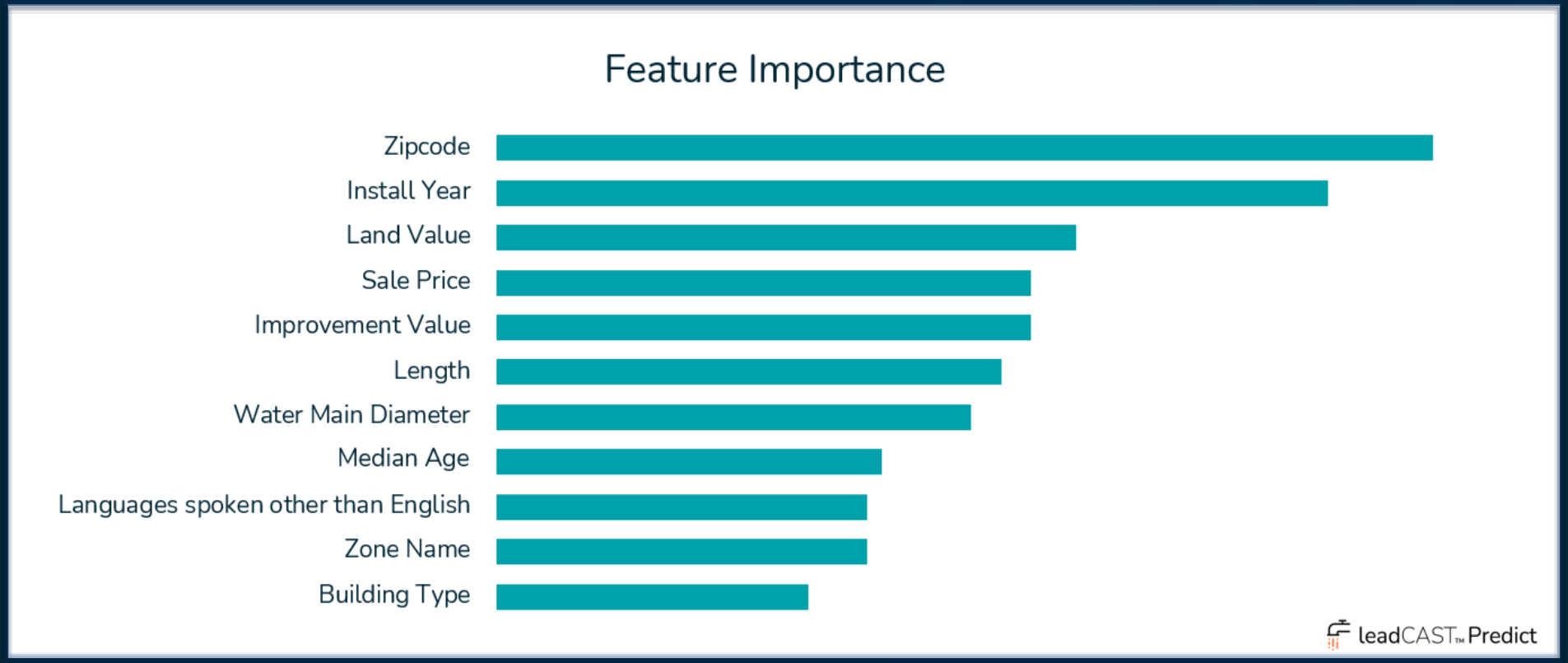


Model improves with more material verifications

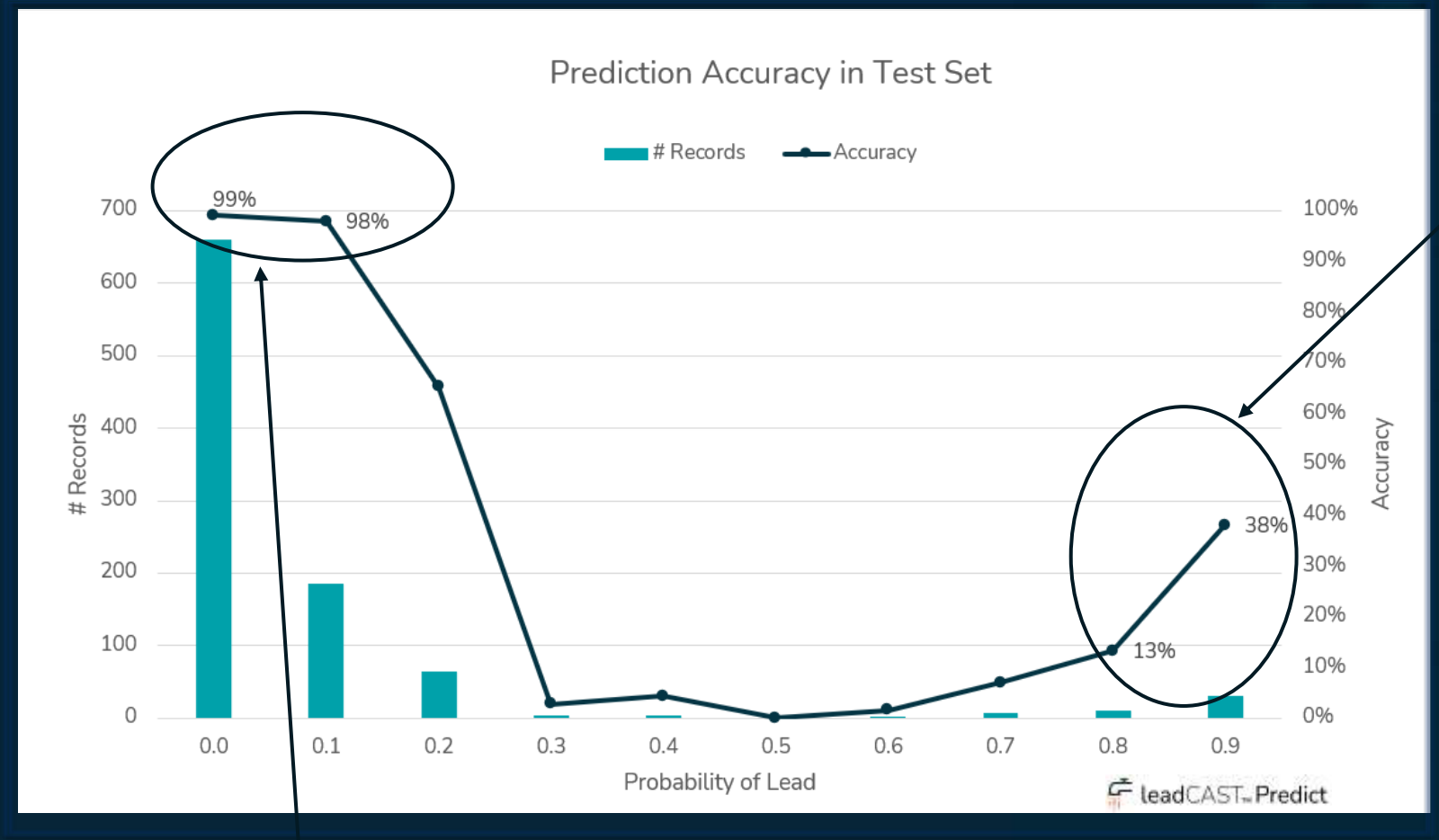
Important Predictor Variables

- The most influential predictor variables (features) vary by community, but there are some common themes.

Sample feature importance plot:



Example Model Training Strategy – Conservative

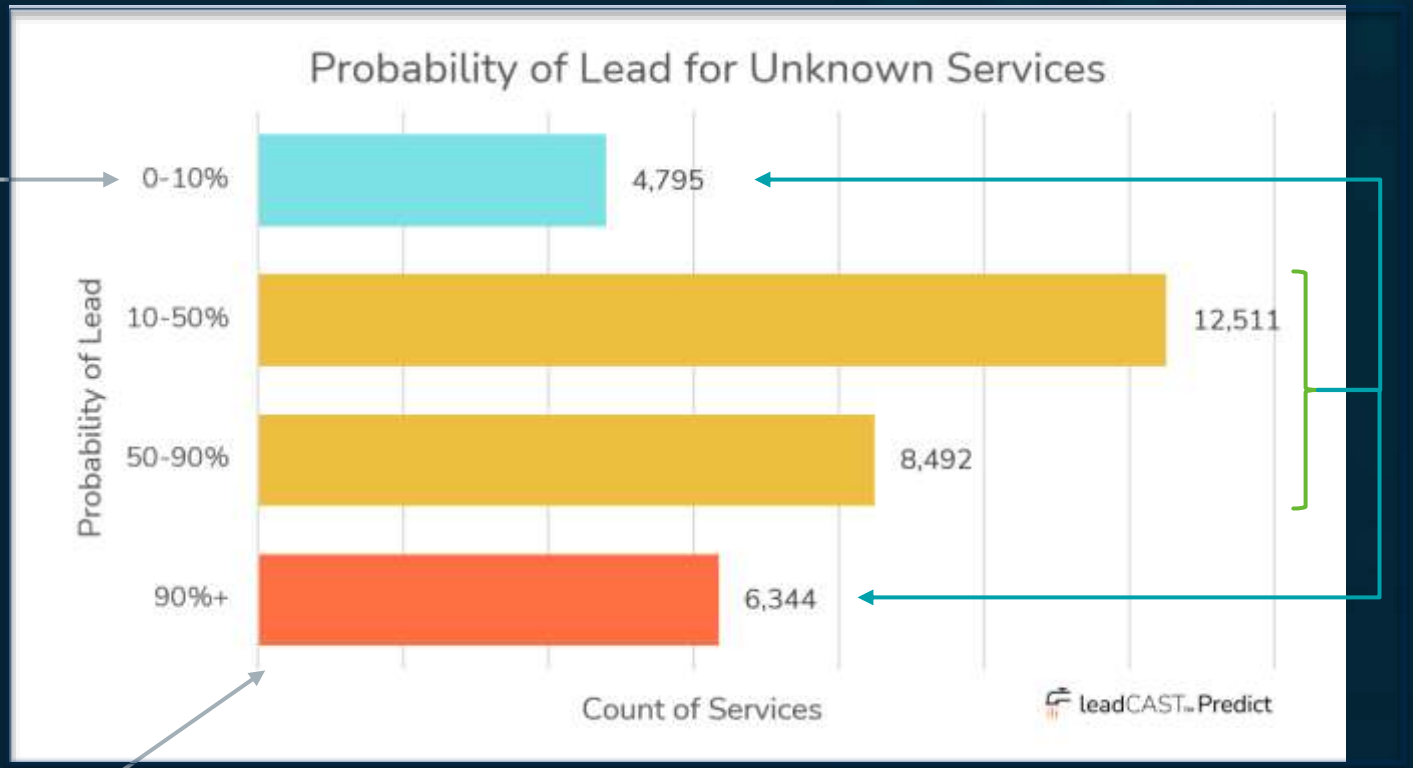


This model is flagging more lead than there actually is in the test set (conservative approach).

This model is optimized to minimize false negatives (conservative approach).

Machine Learning Cycle

In test set, predictions in this range are 99% accurate.



Over time, as model learns from new inspections, more services should move from middle range to high or low probability of lead and hit rate will improve with targeted field verifications.

In the test set, predictions in this range are less accurate because the initial model has been conservatively calibrated to over-predict lead.

trinnex™ | leadCAST™ Predict

Service Line Material Prediction Model Report

Service line material predictions are powered by leadCAST Predict. This report provides a look behind the scenes at the process by which leadCAST Predict has ingested and analyzed your data in order to provide service line material predictions based on machine learning. It also provides performance metrics, which are helpful for understanding how accurate and reliable you can expect service line material predictions to be, given the data currently available for modeling.

Current Model Accuracy: 93.45 % +- 0.01 at 95% confidence interval

Note: Accuracy is the percentage of correct predictions out of all predictions. See Model Evaluation for additional performance metrics that provide a deeper understanding of the model's ability to correctly predict service line material and minimize false negatives.

Table of Contents:

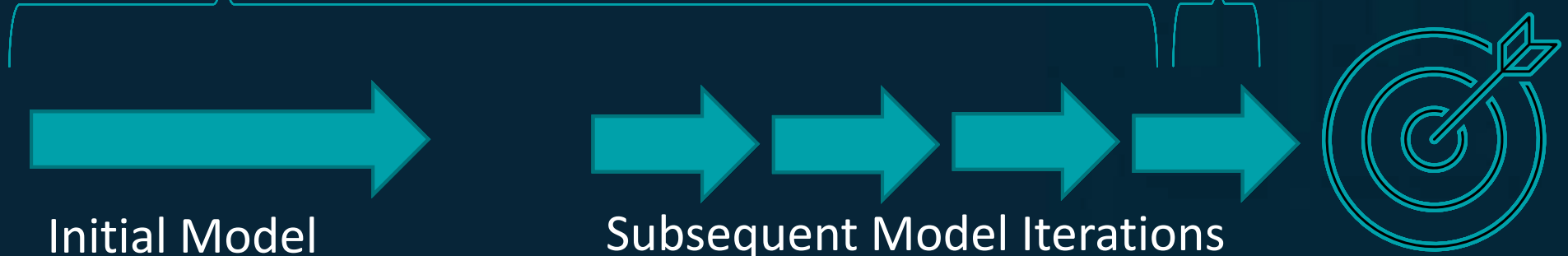
Two primary uses for machine learning

Use #1: Planning and Prioritization

- Prioritize field inspections
- Estimate replacement costs
- Increase (or decrease) confidence in historical records

Use #2: Inventory

Appropriate if/when model reaches acceptable level of accuracy



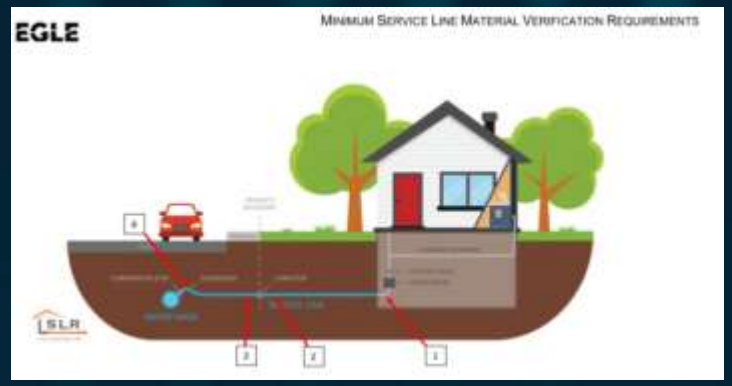
Initial Model

Developed from initial inventory, historical records.

Subsequent Model Iterations

Improved with field verification data.

States with Machine Learning Guidance



Michigan EGLE

- 3-4 point verification method
- Random, unbiased data set
- Use to assess reliability of existing records and/or predict materials at other locations

New Jersey NJDEP

- ML is more accurate than using historical data alone
- Develop a “Historical Records Materials Confusion Matrix”
- Annual predictive modeling report required
- Examples include model accuracies from 80% to 92%
- Sample size must be large enough
    - 8,100 training data points in Pittsburgh (12% of system)
    - 15,447 training data points in Flint (36% of system)
- Need to continue running model until can classify with strong probabilities (i.e. >90% and <10% probability of lead) while middle probabilities remain unknown

THANKS!

Do you have any questions?

Da Yu, P.E.
da.yu@trinnex.io
(407) 760-5807

leadCAST™ Predict