Capstone Milestone 1: BlueConduit

Machine Learning to Detect Lead Service Lines

What is lead?

- Heavy metal, typically soft and malleable
- One of the first materials used to form pipes & paint (dating back to Rome!)

Lead: Why do we want to remove it?

- When ingested, lead is highly poisonous to humans
  - Young children are particularly vulnerable
- Banned from inclusion in paint in 1978 & all pipes in 1986

2014

Flint changes water source from Detroit to the Flint River

2015

New water corrodes lead pipes

Lead gets into water → Lead **poisoning**

2016

City implements FAST Start to replace pipes across the city

PROBLEMS

- Municipal records are scarce
- Digging pipes to confirm material is expensive

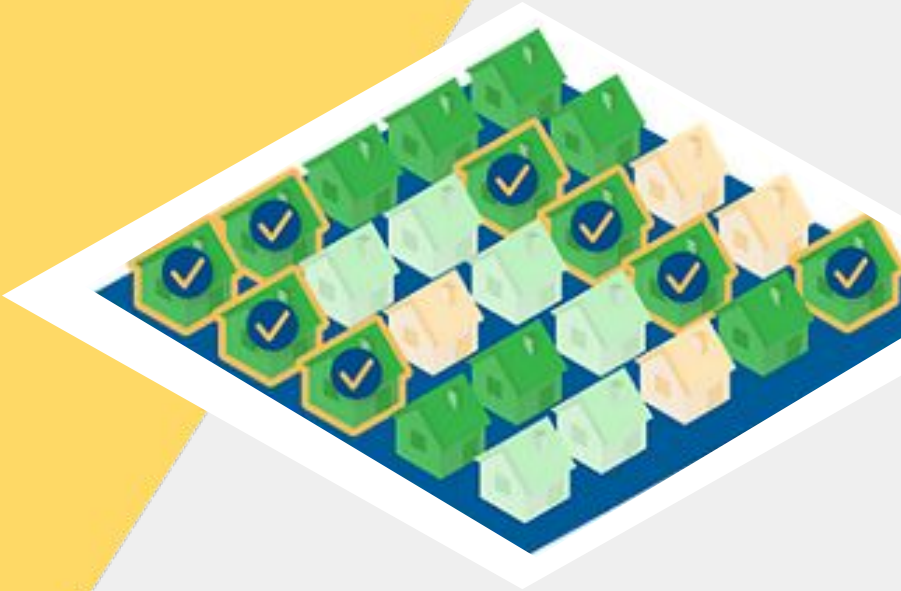
BLUECONDUIT'S MODEL

Uses city data to predict copper/lead.

- Initial program:
  - 15% Hit Rate
- BlueConduit Model:
  - 81% Hit Rate

Scope of Work

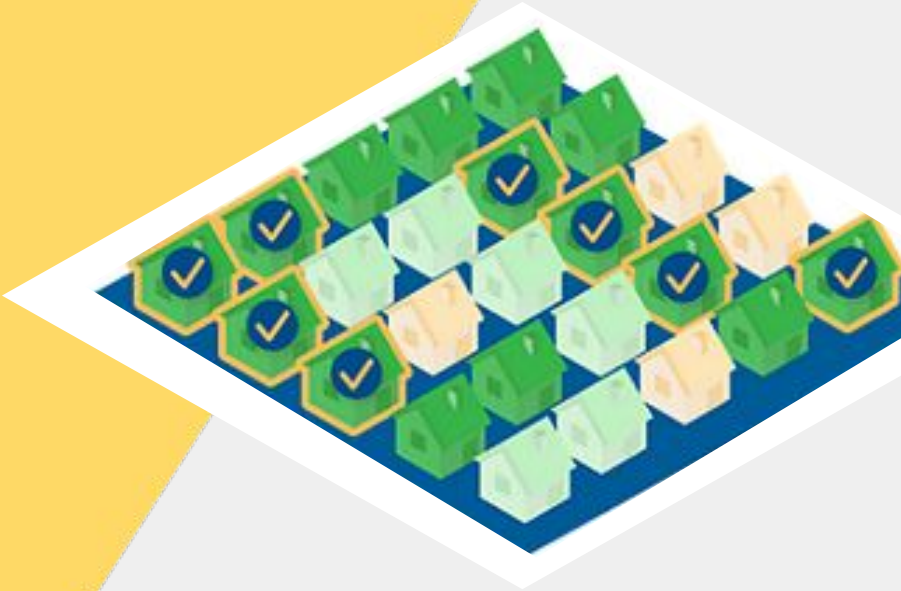
Investigate the utility of including spatial information into existing BlueConduit model



Scope of Work

Experiment with different phases of modeling pipeline

- Spatial features
- Spatial-focused model
- Spatial postprocessing (diffusion)



Interesting topics outside the project scope:

- Improvement of non-spatial features
- Direct tweaks to BlueConduit baseline XGBoost model
- Novel model evaluation metrics
- Alternative modeling paradigms (i.e. computer vision or NLP for info extraction from city records)

Project Ideas

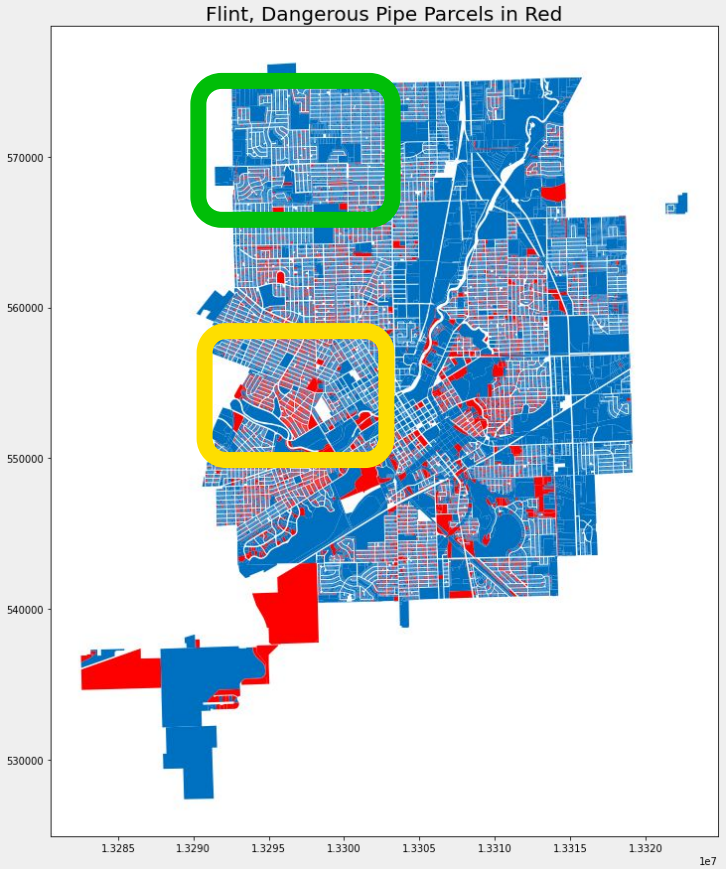
1

NAIVE

Insert longitude & latitude as predictors.

Expected Failure Mode:

- Neighborhoods share longitude and latitude, but not characteristics that predict lead

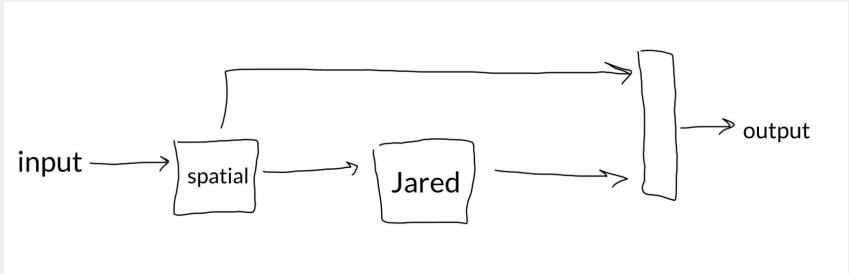


Project Ideas

2

SPATIAL PREDICTORS

Using a combination of Bayesian spatial diffusion (e.g. GPs) and simple models (e.g. KNN), develop set of predictors to be incorporated into baseline model



Potential Failure Modes:

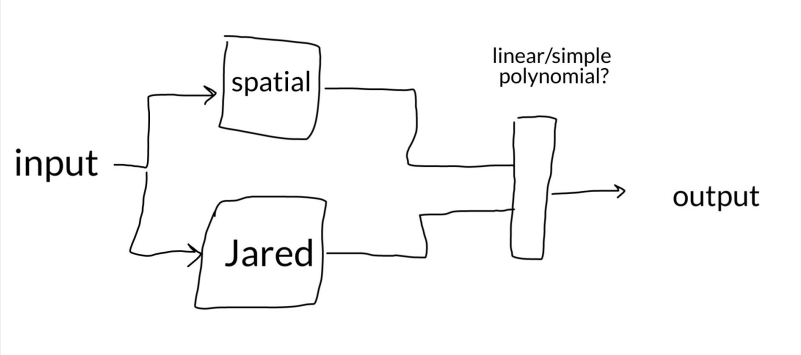
- Data-greedy (will not address low-data environment)
- Computationally costly
- Difficulty defining distance

Project Ideas

3

ENSEMBLE W/BASELINE

Train model using only spatial information (or spatial + small subset of features); ensemble with BC baseline model.



Potential Failure Modes:

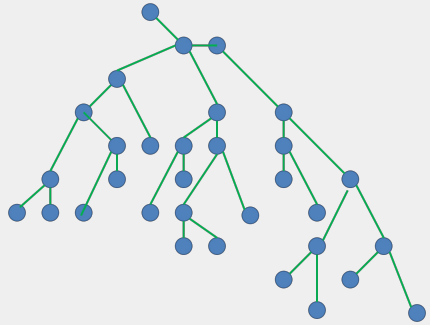
- Gaussian processes do not converge well & scale poorly
- Requires precise understanding of the spatial distances.

Project Ideas

4

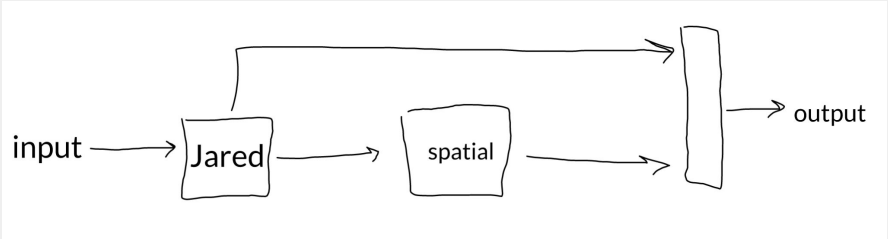
SPATIAL POSTPROCESSING

Utilize graph-based information & opinion dynamic methods to model the flow of info about lead pipes



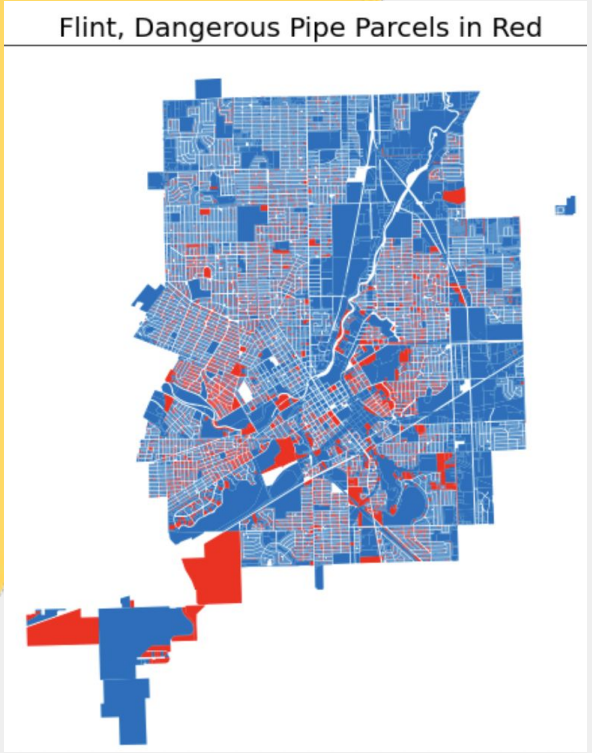
Potential Failure Modes:

- **Information loss** with probabilities from Jared model not tied to their idiosyncrasies
- Novel territory, not clear that information will be useful
- May not sufficiently improve baseline model



WHAT IS SUCCESS?

- Understand whether spatial information can improve the BlueConduit model.
- Spatial info improves current model's "hit rate".



Flint, MI

Team and Collaboration Infrastructure

- Code organization & infrastructure:
  - GitHub
    - All code refactored into internal package for reproducibility
  - Colab (Compute-intensive tasks)
  - AWS (Hosting of OSRM)

Team and Collaboration Infrastructure

- Tasks:
  - Individually assigned during weekly meetings.
  - Track: Slack and on the following meeting.

Team and Collaboration Infrastructure

- Communication channel:
  - *Internal*: Slack
  - *Partner*: email + Slack
- Weekly regular meetings:
  - *Internal + course staff*: Tuesdays
  - *Partner*: Wednesdays

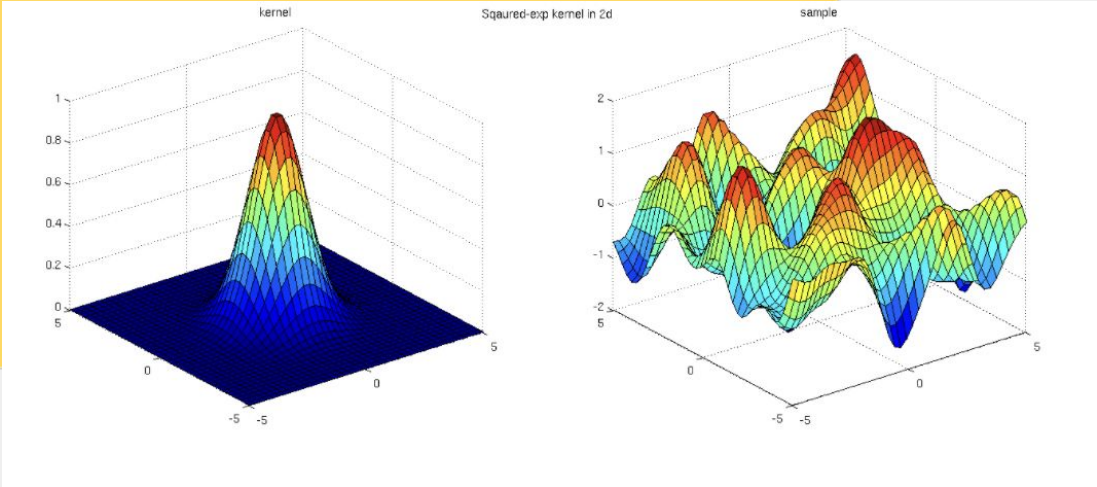
Learning Goals

- Technical
  - Gaussian processes/ diffusion models
  - Transform locations into block distances
  - Use geospatial libraries
  - Learn historical context and technical jargon
- Managing
  - Organize and prioritize tasks towards a goal
  - Anticipate different scenarios
  - Translate partner needs into project constraints

Relevant Knowledge/Literature Review

Kernel Cookbook - University of Toronto (due to suggestion of Gaussian Process as spatial model)

https://www.cs.toronto.edu/~duvenaud/cookbook/



Relevant Knowledge/Literature Review

Data Dependent Distance Metric for Efficient Gaussian Processes Classification (due to the pipes being structured under the city's road network)

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.719.8106&rep=rep1&type=pdf

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(-\frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2}{2\lambda^2}\right) \quad (14)$$

Using equation 9, the kernel in equation 14 can be generalized as:

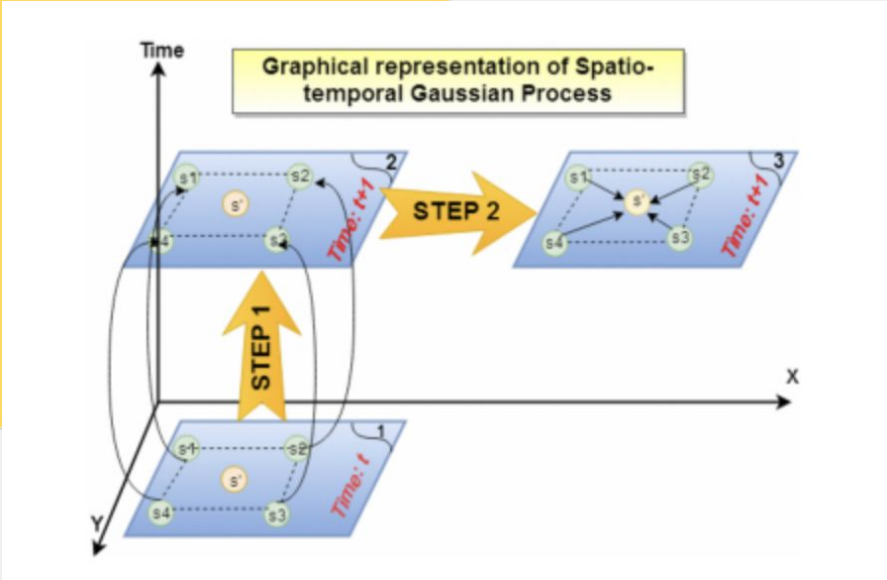
$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(-(\boldsymbol{x_i} - \boldsymbol{x_j})^T A(\boldsymbol{x_i} - \boldsymbol{x_j})\right) \quad (15)$$

This suggests that, as for any other kernel-based learning algorithm, the problem of learning with GP is actually finding the right specification of matrix A in equation 15.

Relevant Knowledge/Literature Review

Bayesian Spatio-Temporal Gaussian Process (due to importance of YEAR BUILT as a feature)

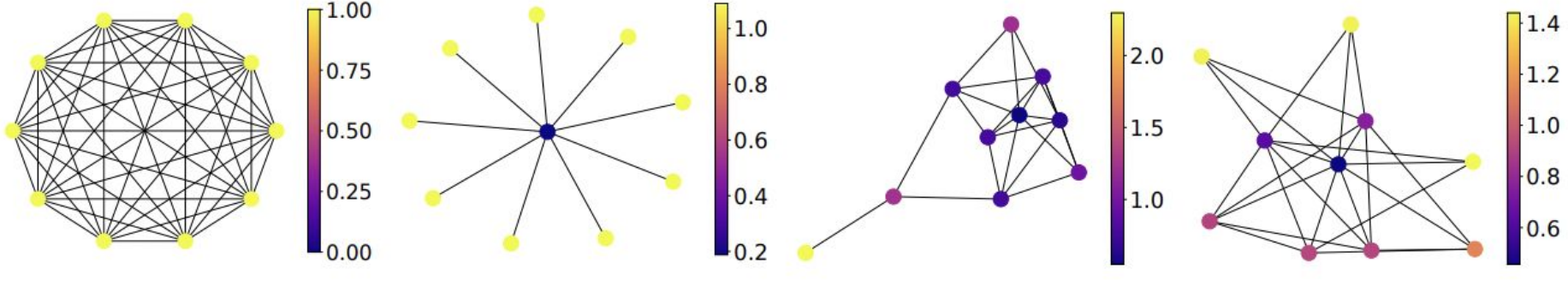
https://dl.acm.org/doi/fullHtml/10.1145/3300185



Relevant Knowledge/Literature Review

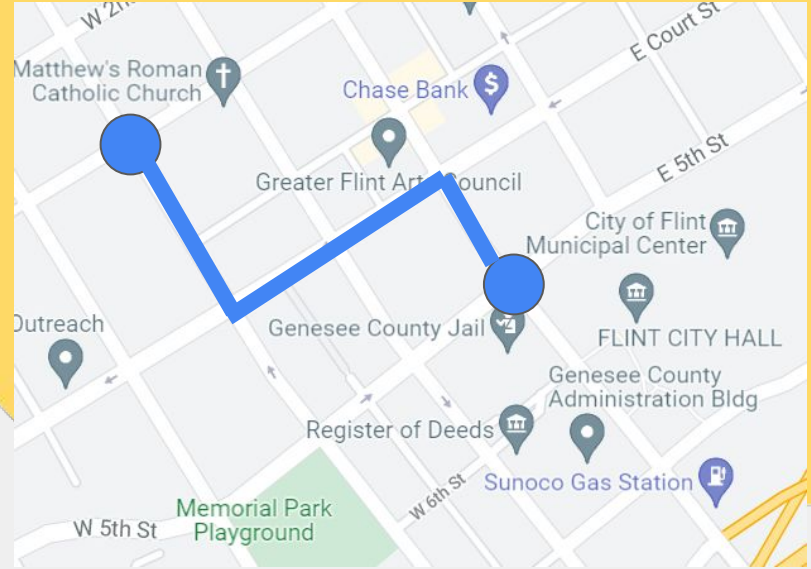
Matern Kernels on Graphs (suggested by cross validation)

http://proceedings.mlr.press/v130/borovitskiy21a/borovitskiy21a.pdf

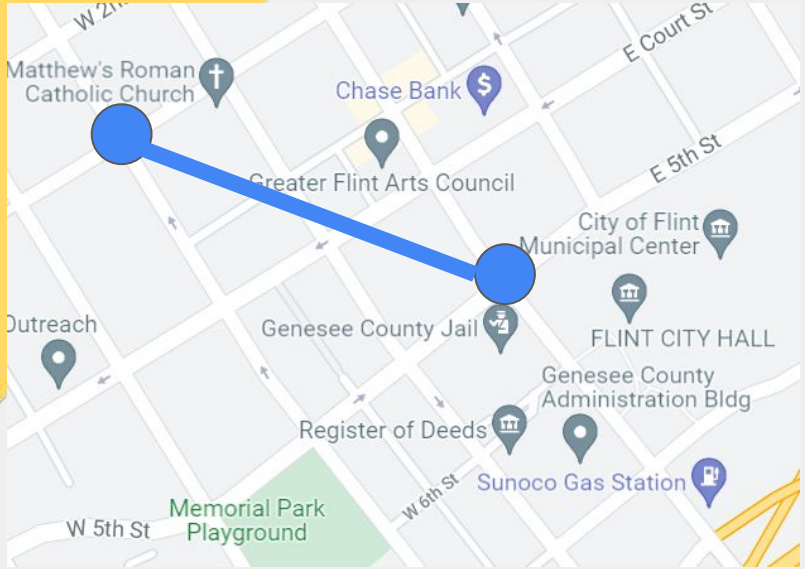


Current Work

Street vs. Euclidean Distances

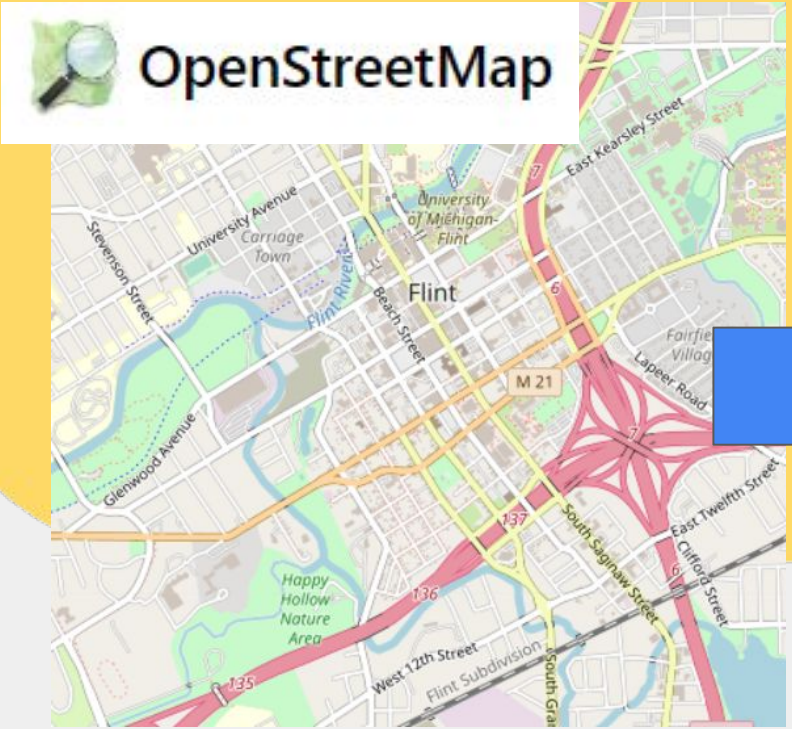


vs.



Current Work

Street vs. Euclidean Distances



Parcels

Parcels	1	2	3	…	n
1	t_{11}	t_{12}	t_{13}	…	t_{1n}
2	t_{21}	t_{22}	t_{23}	…	t_{2n}
3	t_{31}	t_{32}	t_{33}	…	t_{3n}
…	…	…	…	…	…
n	t_{n1}	t_{n2}	t_{n3}	…	t_{nn}

t_{ab} = walking time from a to b

EDA - Flint data

About 26,000 parcels (dataset rows)
74 features

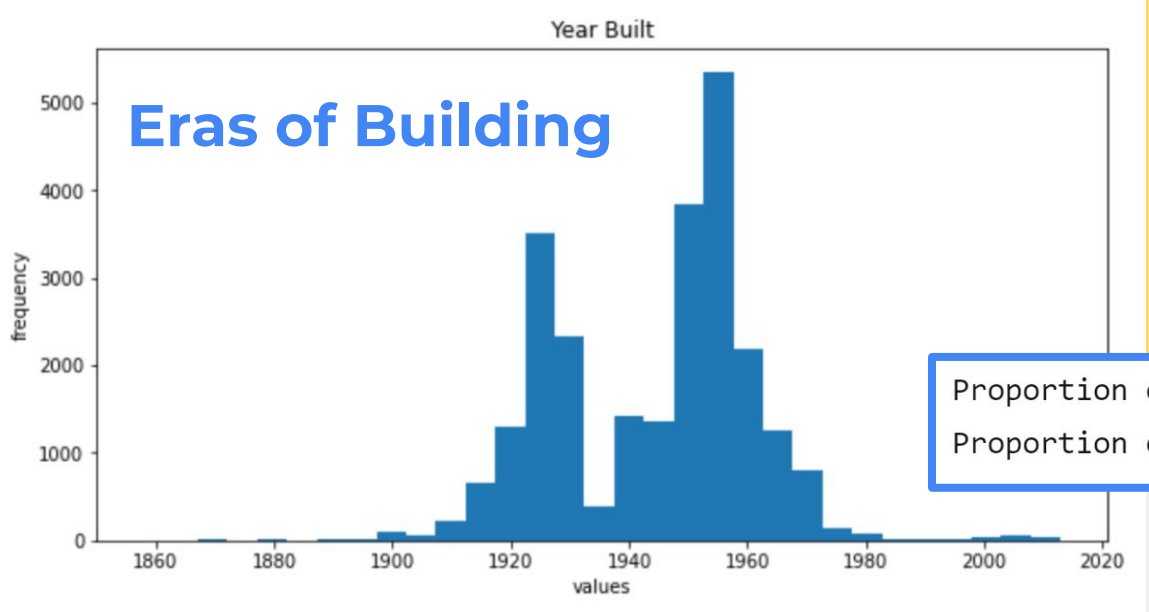
Missingness

Commercial Condition 2013	26760
B_aggregate_income	4666
B_imputed_value	1500
B_imputed_rent	1500
B_hispanic_household	1500
Housing Condition 2012	400
Residential Building Style	114
Housing Condition 2014	109
USPS Vacancy	59
Owner State	26

Target Distribution

```
Number of dangerous parcels:  10266
Proportion of parcels that are dangerous:  0.382
```

EDA - year built



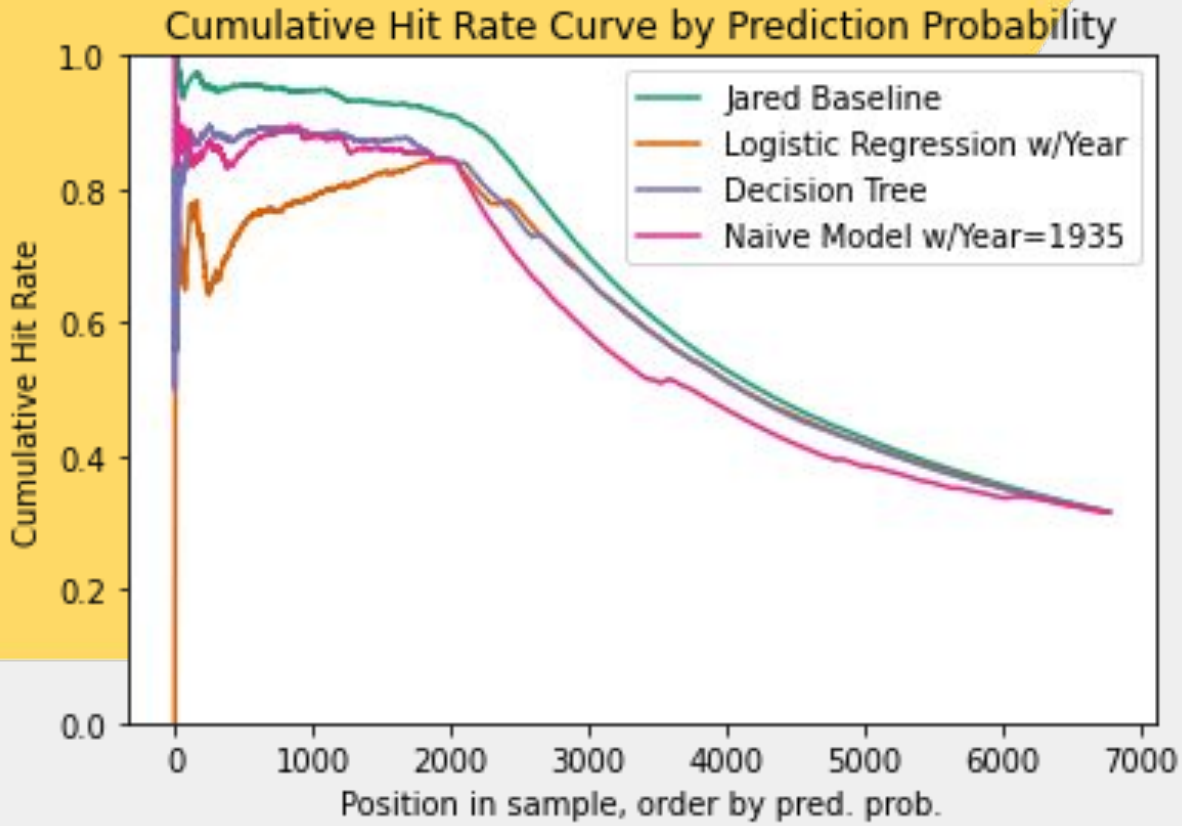
Predictiveness

```
Proportion of Pre Depression homes dangerous:  0.885
Proportion of Post WW2 homes dangerous:  0.072
```

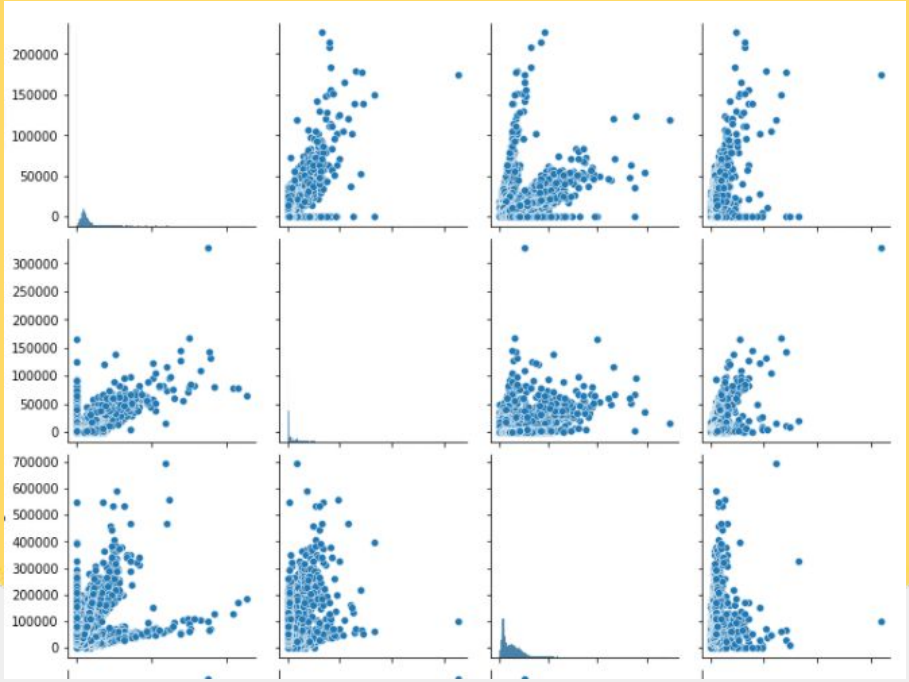
Missingness

1,629 homes with unreasonable values

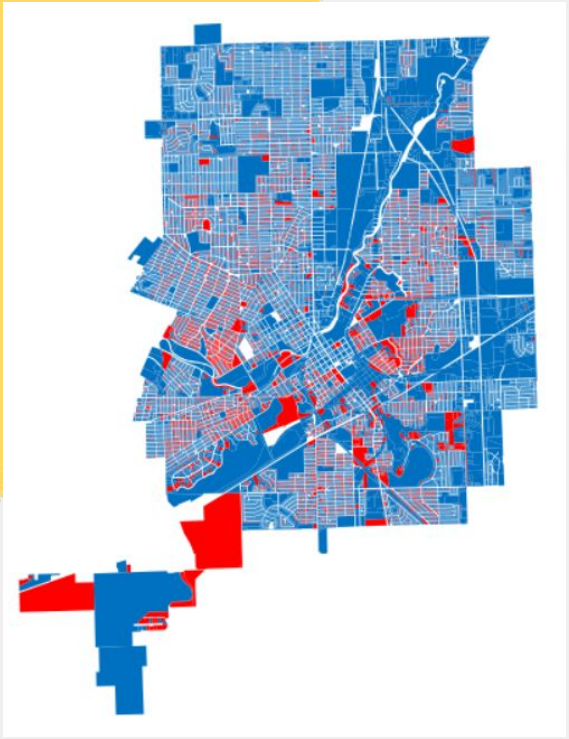
EDA - naive models with just Year Built



EDA - Relevant visualizations



Home value & size highly correlated

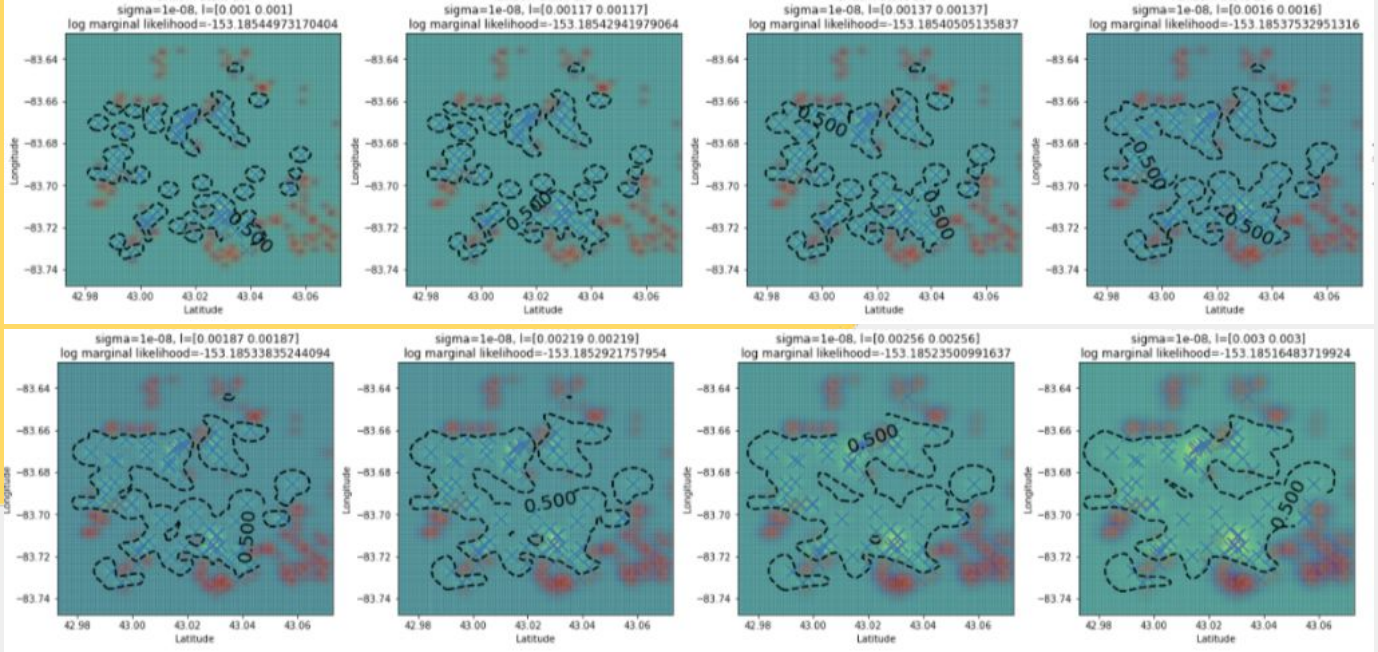


Lead pipe geospatial distribution

Spatial Modeling - Gaussian process

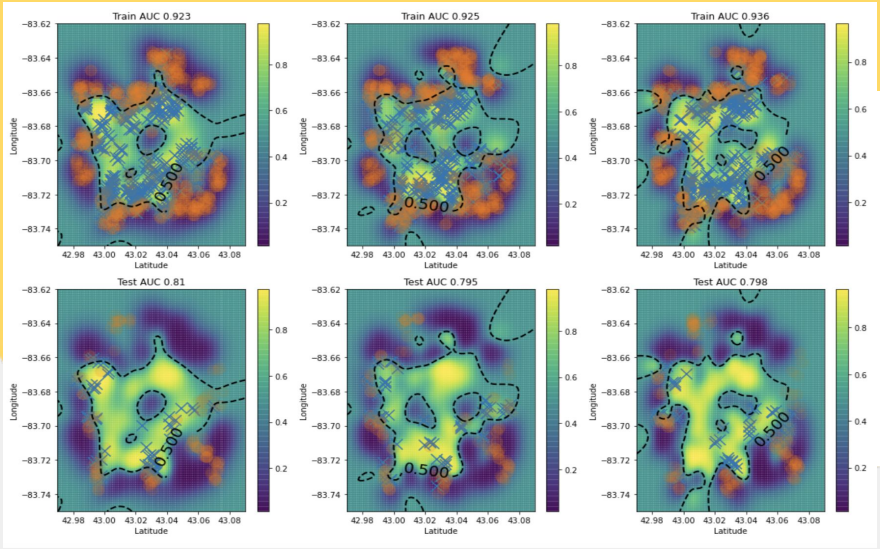
L parameter in RBF kernel particularly important to tune

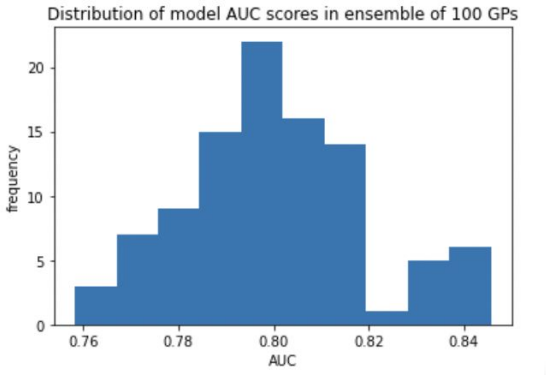
$$k_{\mathrm{SE}}(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$$



Spatial Modeling - Gaussian Process Ensemble

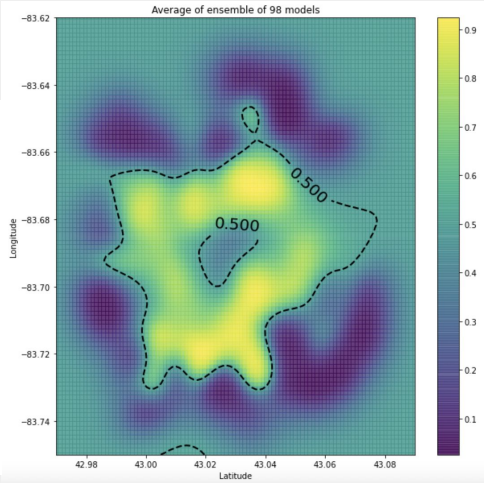
Sample of 2,000 homes each
(here showing 3 on Train and Test)
X's are homes with lead, O's are homes without





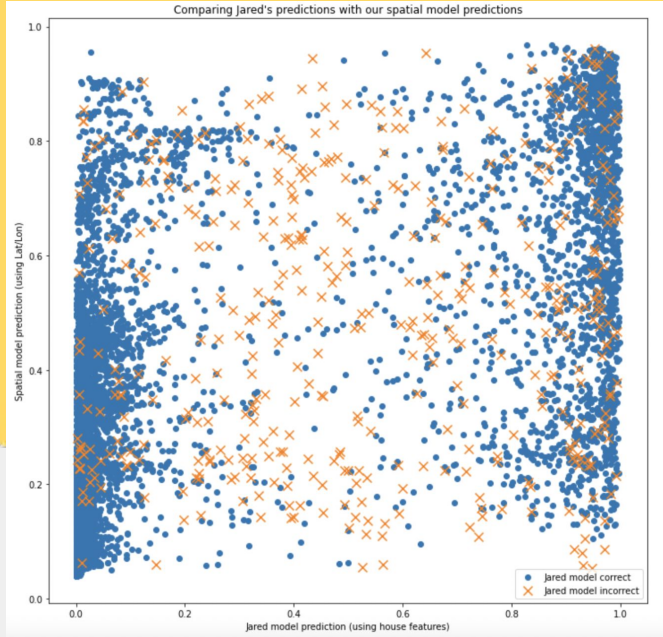
Mean of Ensemble

AUC: 0.8288

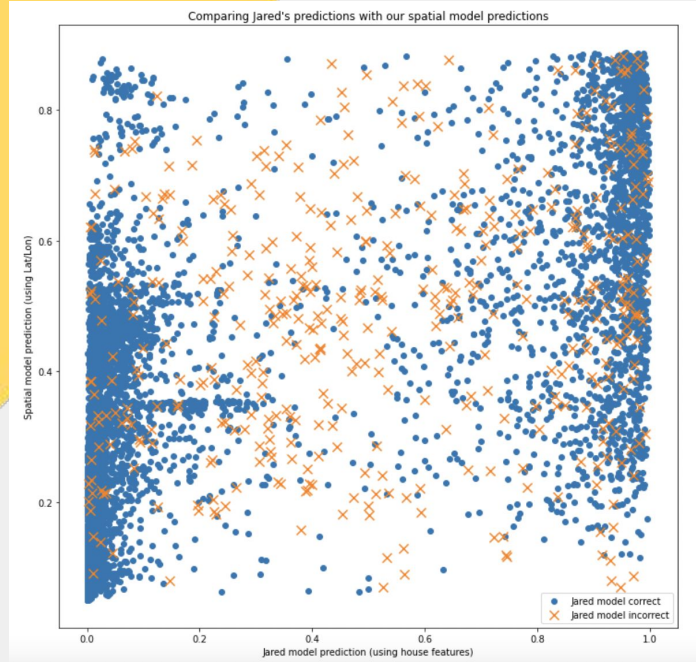


Spatial Modeling - Gaussian process Ensemble

Individual GP



Ensemble of 100 GPs



What's At Stake?

Save lives in Flint

Save $$$ in Flint

Save lives and $$$ elsewhere

Thank you!

Kevin, Javiera, Max, Dash