Team BlueConduit - AC297R
Statement of Work

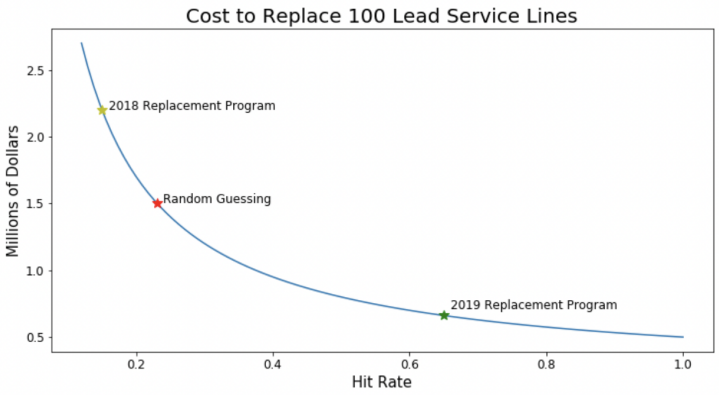
Javiera Astudillo, Max Cembalest, Kevin Hare, Dashiell Young-Saver

Background

BlueConduit believes all cities should be able to replace hazardous infrastructure affordably and quickly. One crucial such task is removing pipes with hazardous materials, notably those made from lead. This is to comply with government regulations, because lead contaminates the water supplied to people's homes. But most cities do not know where their lead service lines are.

BlueConduit's work so far has successfully

a) **Lowered the cost** of replacing service lines
b) **Sped up the process** by increasing the accuracy in predicting lead in water pipes



Problem Statement

Our problem is to improve the way BlueConduit's existing model incorporates spatial information - to help them continue to drive down the cost and time taken to improve water infrastructure.

BlueConduit has already begun to explore how spatial modeling can be used to improve home level service line material predictions. Our team's addition will be to represent the available housing data by situating it on a geospatial graph (as opposed to, say, including

Latitude and Longitude as features in a tabular dataset) to see if this improves our ability to identify the regions that need water lines replaced.

Our primary deliverable is a model with the following basic specifications:

Input: spatial location
Output: probability that the location contains hazardous materials in need of replacing

Goals:

- Implement **physics-based diffusion models** to incorporate the spatial flow of information regarding lead contamination between homes, blocks, etc.
- Experiment with different **resolutions** (e.g. house, block, neighborhood, census tract, etc.) at which the model processes spatial information.
- Model performance is directly tied to improvements in people's health and quality of life by helping them get access to clean water at home. Therefore we are interested in **improving model accuracy even by small amounts.** Based on our initial discussion, improvements may be:
    - Increasing performance on traditional ML classification metrics such as precision and recall.
    - Evaluating model performance against a customized metric: effective cost of replacement.[1]
- *Stretch Goal:* If successful with the Flint dataset, evaluate spatial diffusion approach across non-Flint geographic areas.

Resources

Resources needed:

- Dataset on Flint houses (to be provided by BlueConduit)
- Past model predictions & outputs to train the spatial model (to be provided by BlueConduit)
- Tools to efficiently construct graph of nodes (houses, blocks, etc) and edges (spatial connections between nodes). This will be provided by BlueConduit, and may require signing an NDA.

It seems that for computational resources we might not need much - the Flint dataset only seems to have on the order of 6,000 homes. However, if running a diffusion algorithm is costly, we would be best served by running our work on Harvard's computing cluster.

[1] The effective cost of replacement is essentially a weighed average cost of the successful and unsuccessful digs, and is the core of BlueConduit's economic value proposition to cities. Improving accuracy translates to fewer unsuccessful (i.e. no lead found) digs, and therefore a higher effective cost.

High-Level Project Stages

- Sep 20: retrieve Flint data and begin EDA
- **Sep 23 (in-class deadline): In-class Ignite Talk**
- Oct 1: Finish EDA. Start learning about diffusion models.
- **Oct 6 (11:59pm deadline): Milestone 1**
- Oct 15: Decide how the model will be trained. Choose diffusion algorithm for milestone 2. *Make sure we can reproduce BlueConduit's results on our end!*
- **Oct 28 (in-class deadline): Milestone 2**
- Nov 5: Finalize choice of diffusion algorithm. Focus on improving the way we evaluate our model and present our results.
- **Nov 17 (11:59pm deadline): Milestone 3**
- Dec 1: Finish first draft of paper/blog
- **Dec 15 (11:59pm deadline): Final presentation**