

Computation for Public Policy

Course Introduction

Jen Helsby and Eric Potash

Computation for Public Policy

Lecture 1: January 5, 2016

computationforpolicy.github.io

Today

- Course Overview
- Assignment
- Introductions
- Case Study

Lead Poisoning: Impacts

- Lead causes irreversible neurological damage
 - reduced IQ
 - psychiatric disorders
- Hundreds of thousands of infants poisoned each year in the U.S.
 - Billions of dollars in damages.

Lead Poisoning: Causes

- Lead was in paint, pipes, gasoline, etc.
- Banned in the 70's but still found in older buildings.
- Infants are susceptible to poisoning due to
 - physiology
 - “hand-mouth” behavior.

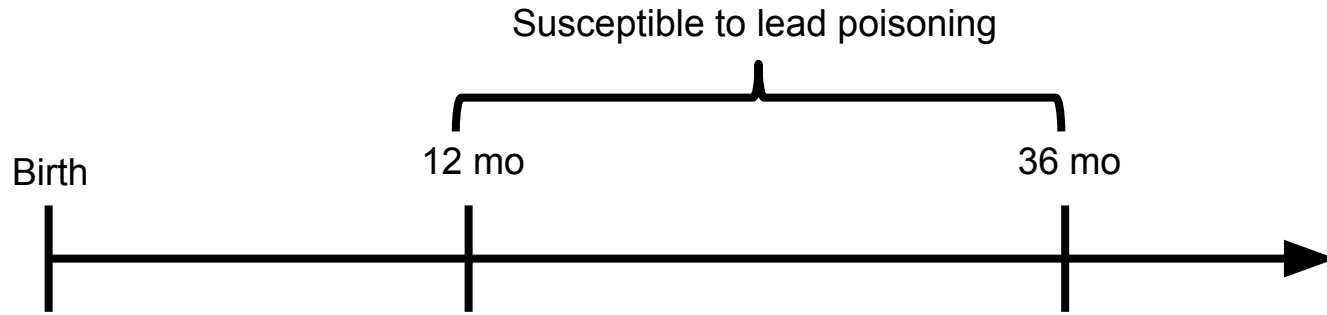
Chicago Lead Protocol

- CDPH has a periodic blood lead level (BLL) testing schedule.
- Children entering Chicago Public Schools are *required* to be screened for lead.

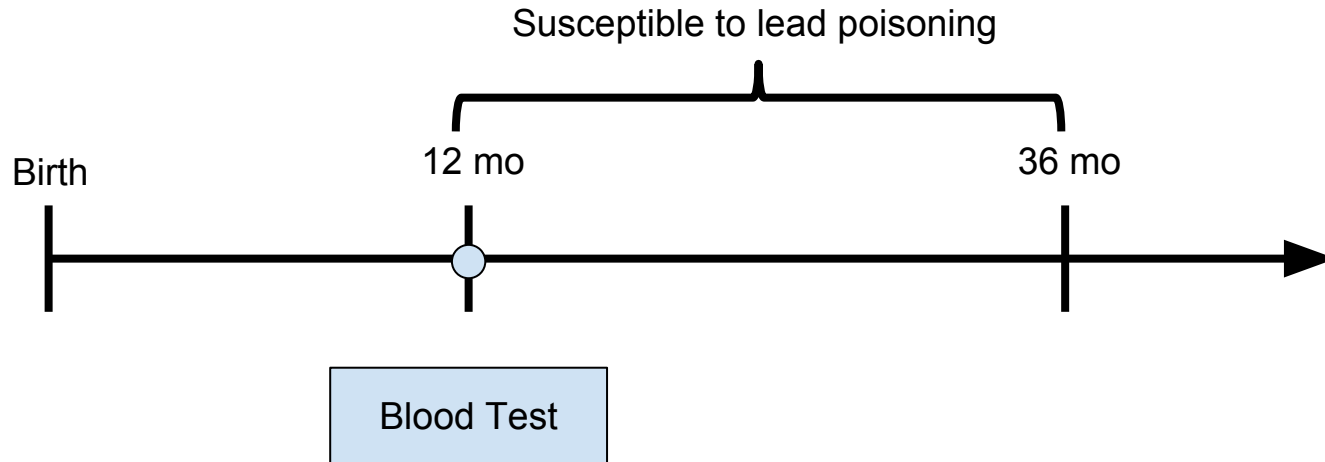
Chicago Lead Protocol

- A test over the threshold (*currently* 5 ug/dL) is considered poisoning.
- Over 10 ug/dL triggers a home investigation.
- Hazards found are legally required to be remediated, but enforcement is scarce.

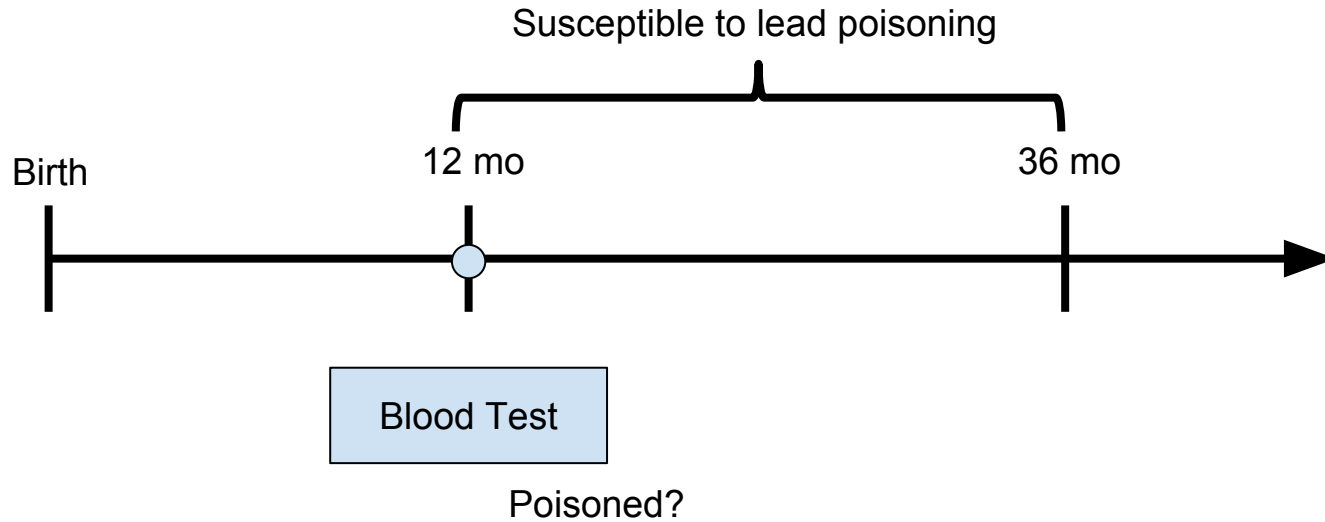
Current Approach



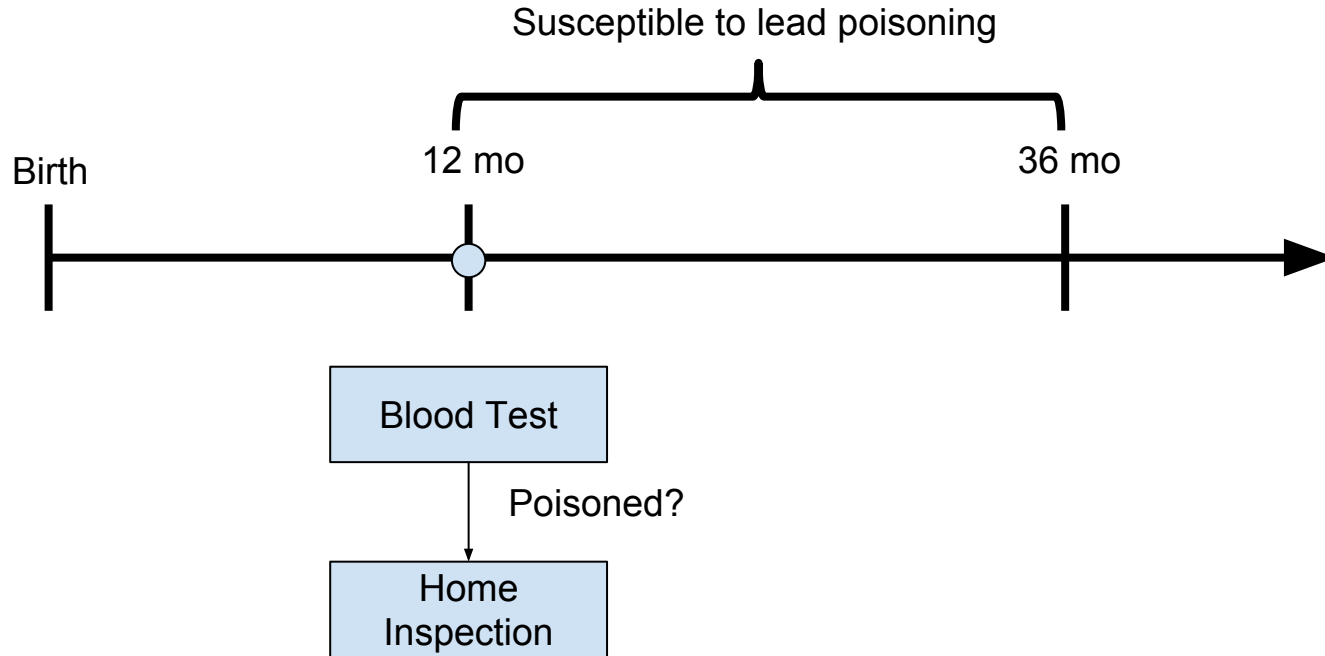
Current Approach



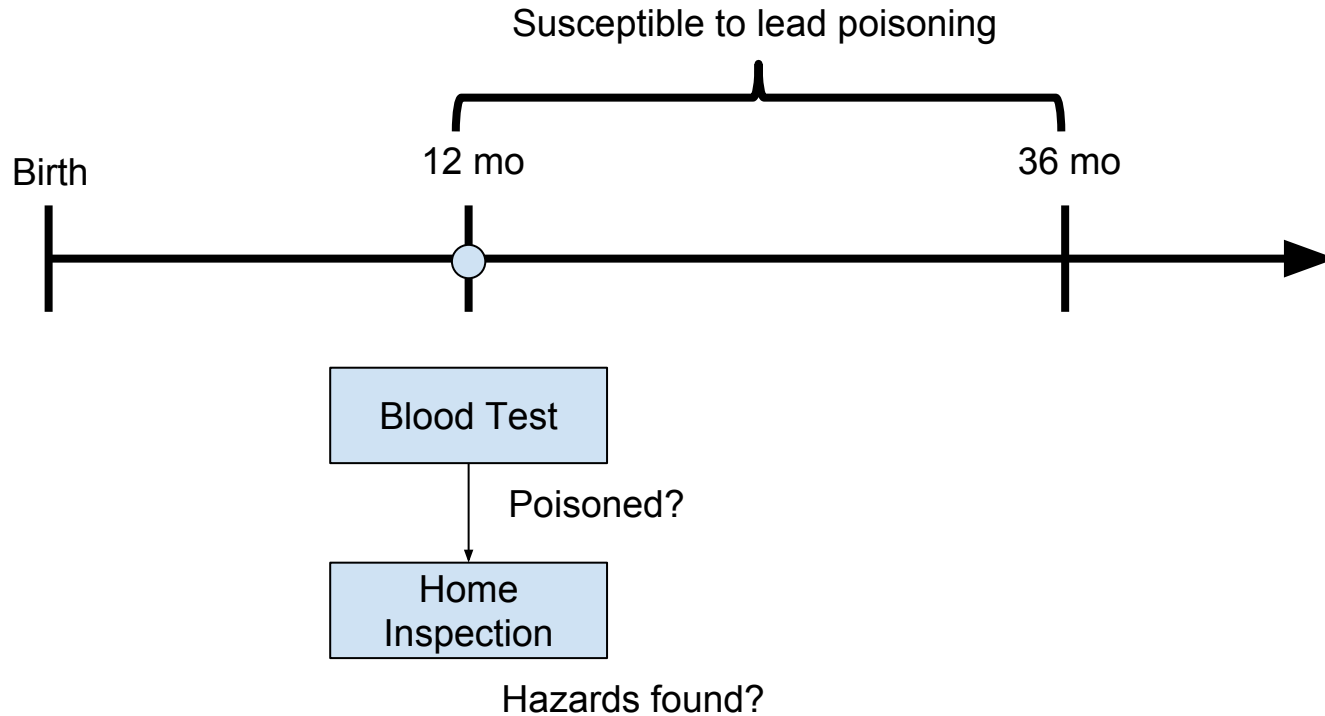
Current Approach



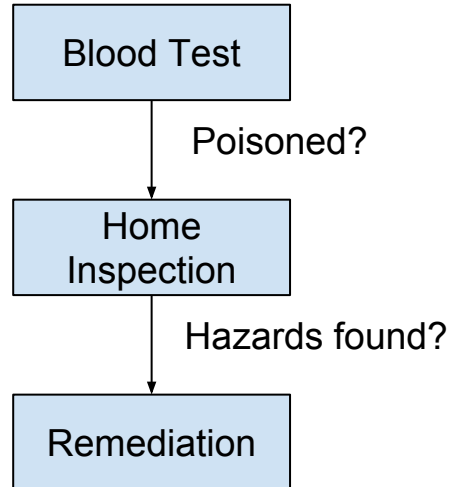
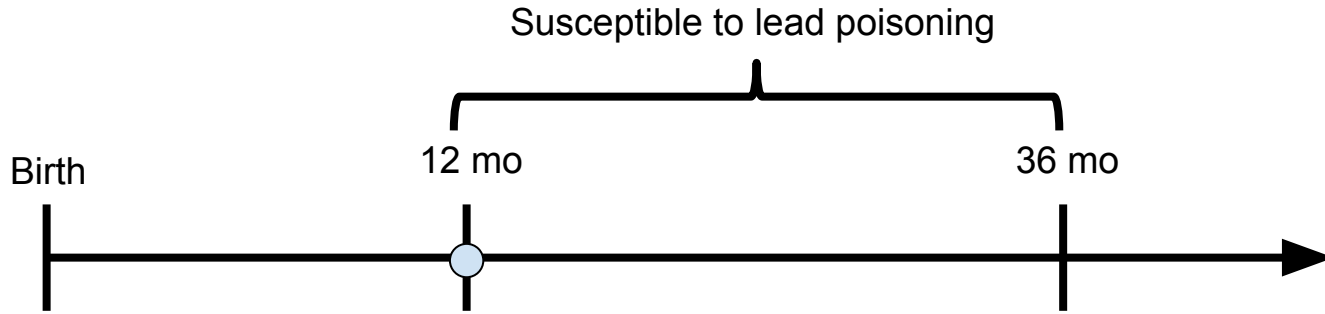
Current Approach



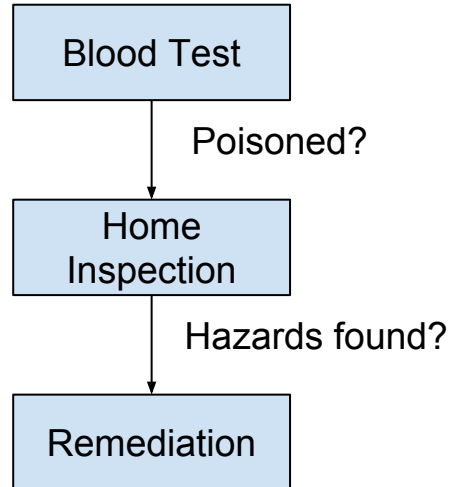
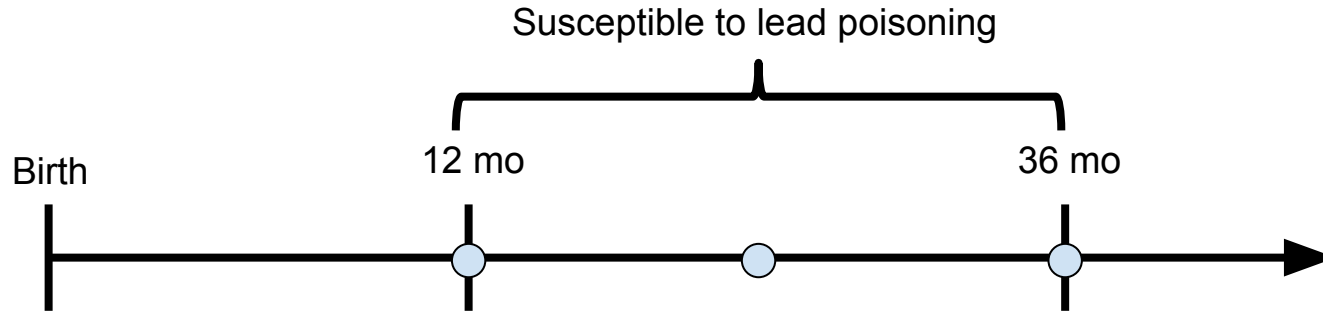
Current Approach



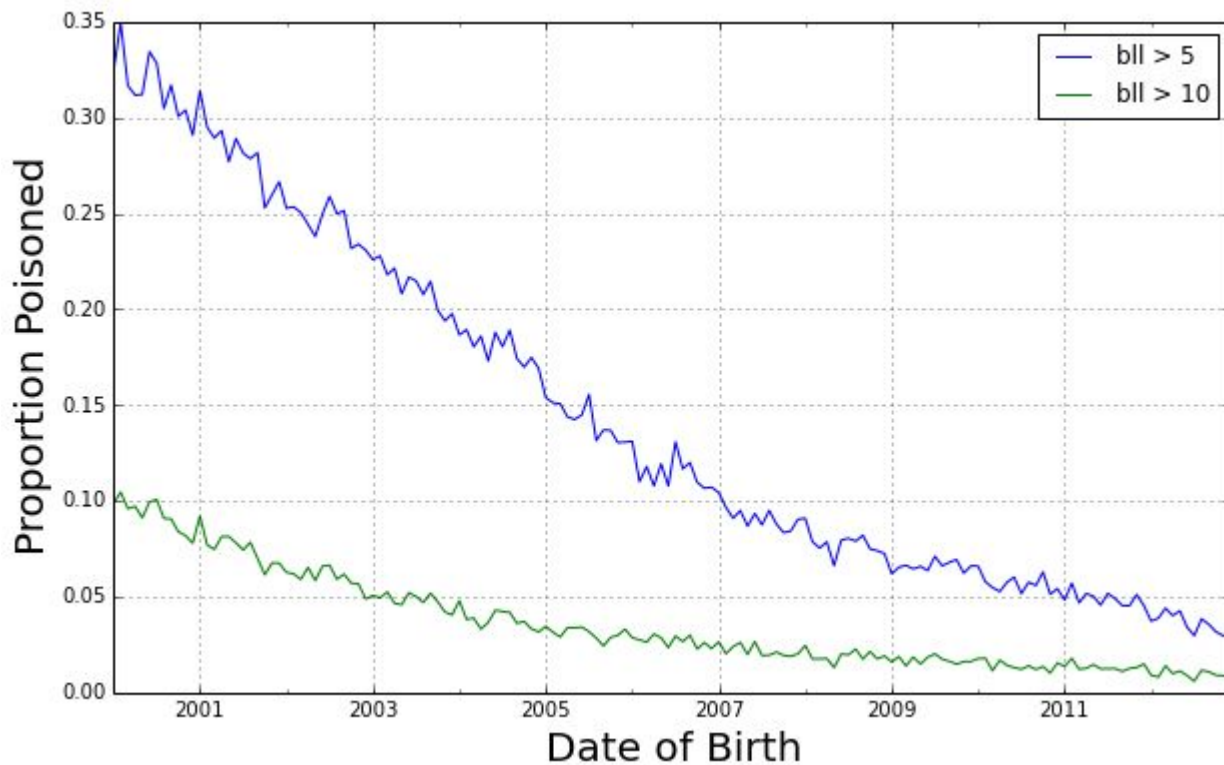
Current Approach



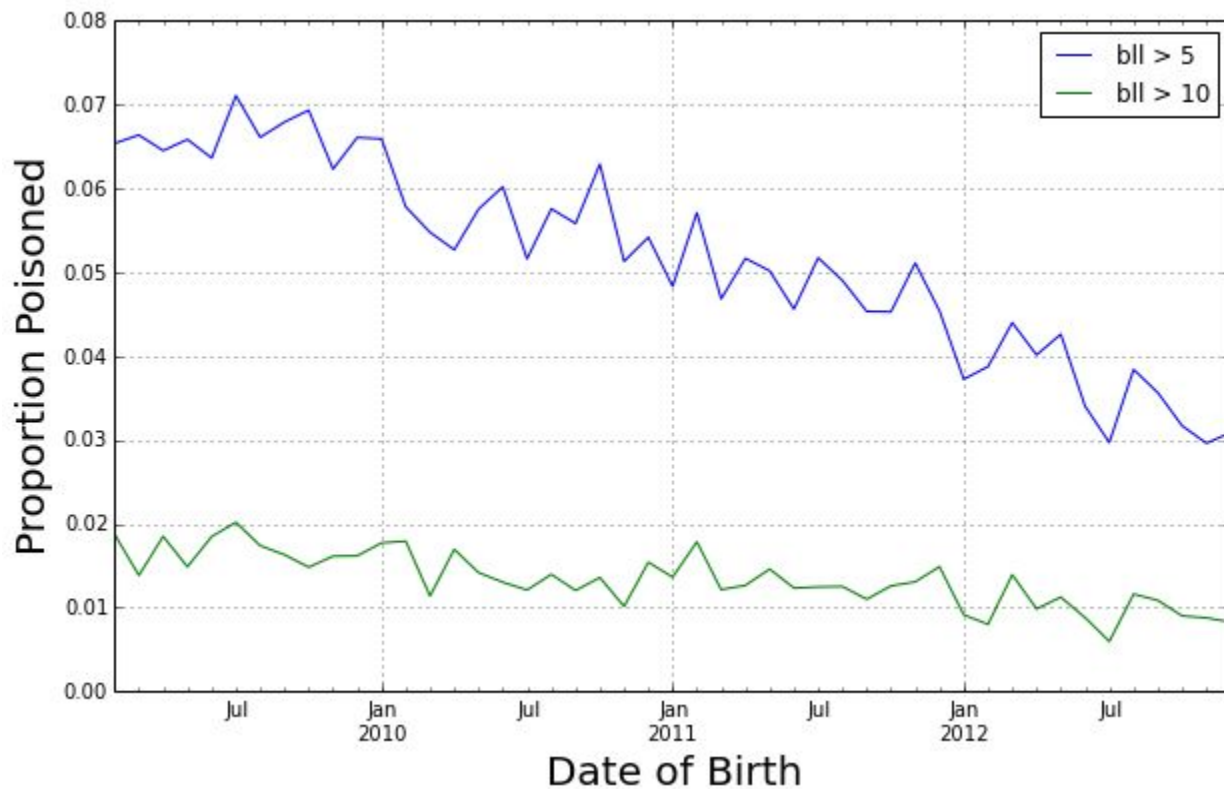
Current Approach



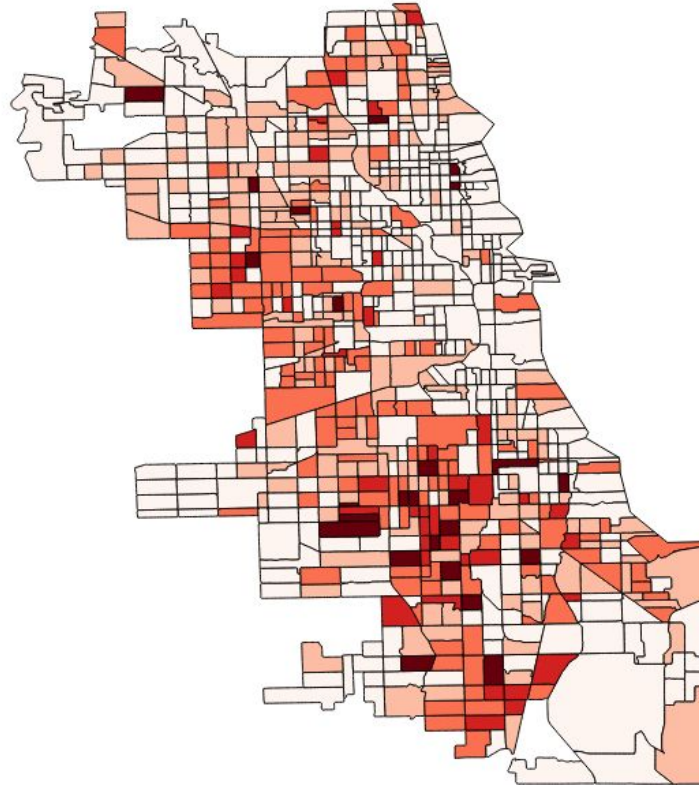
Birth Cohort Incidence








Birth Cohort Incidence



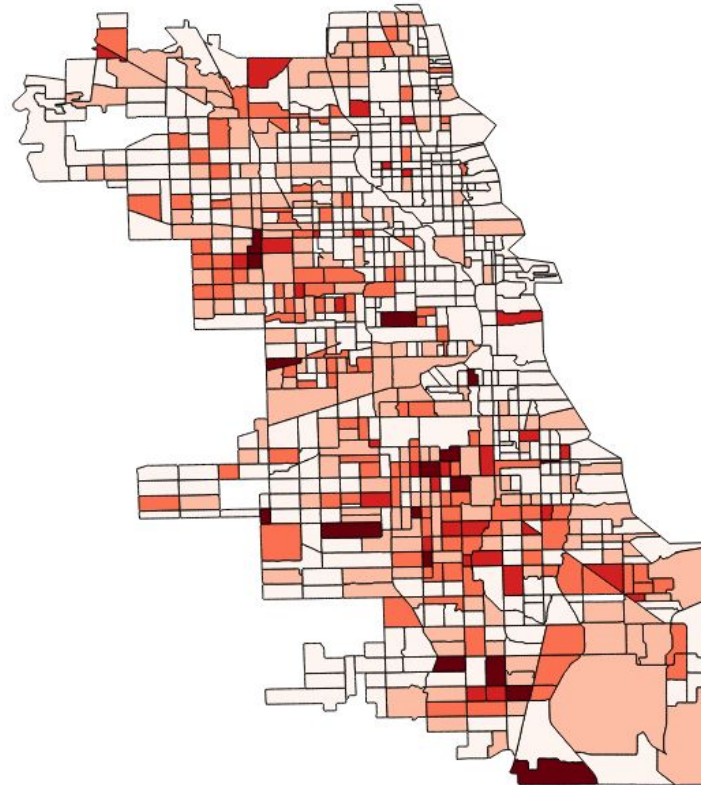
Birth Cohort Incidence by Tract: 2011



Legend

| Incidence Rate | |
|---|-------------|
|  | 0.00 - 0.02 |
|  | 0.02 - 0.05 |
|  | 0.05 - 0.10 |
|  | 0.10 - 0.15 |
|  | 0.15 - 0.30 |

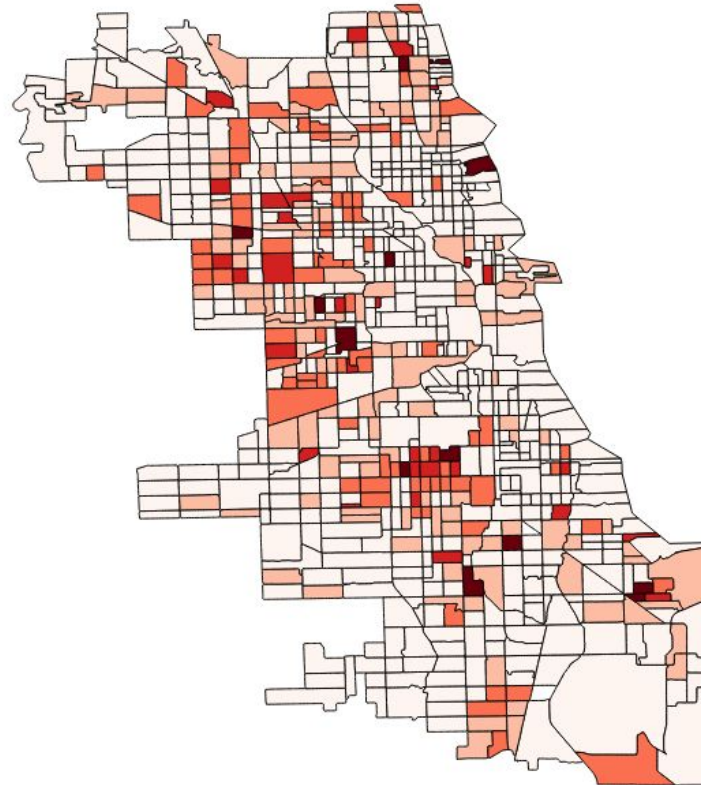
Birth Cohort Incidence by Tract: 2012



Legend

| Incidence Rate | |
|----------------|-------------|
| 0.00 - 0.02 | 0.02 - 0.05 |
| 0.05 - 0.10 | 0.10 - 0.15 |
| 0.15 - 0.30 | |

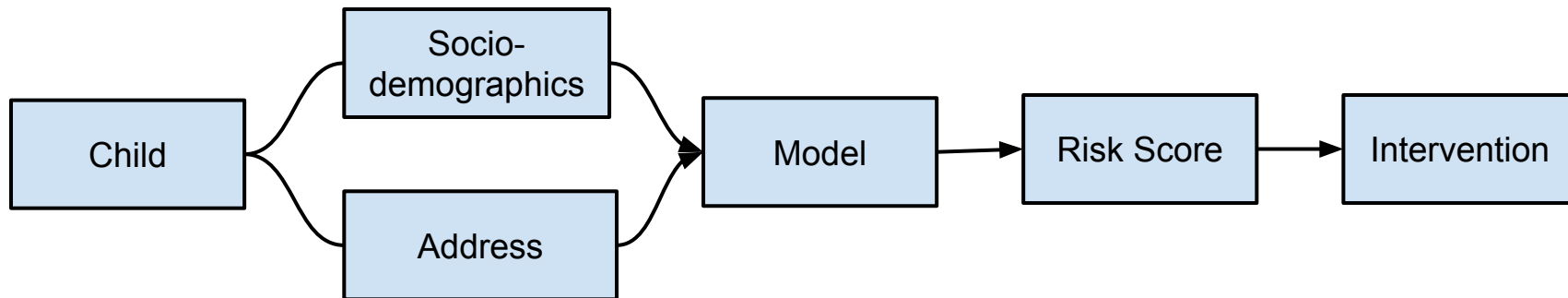
Birth Cohort Incidence by Tract: 2013



Legend

| Incidence Rate | |
|----------------|--|
| 0.00 - 0.02 | |
| 0.02 - 0.05 | |
| 0.05 - 0.10 | |
| 0.10 - 0.15 | |
| 0.15 - 0.30 | |

Predicting Lead Poisoning



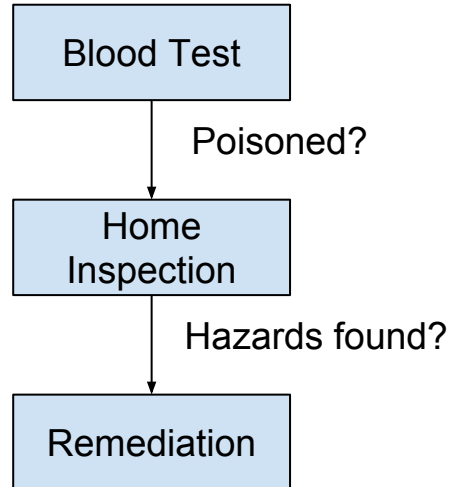
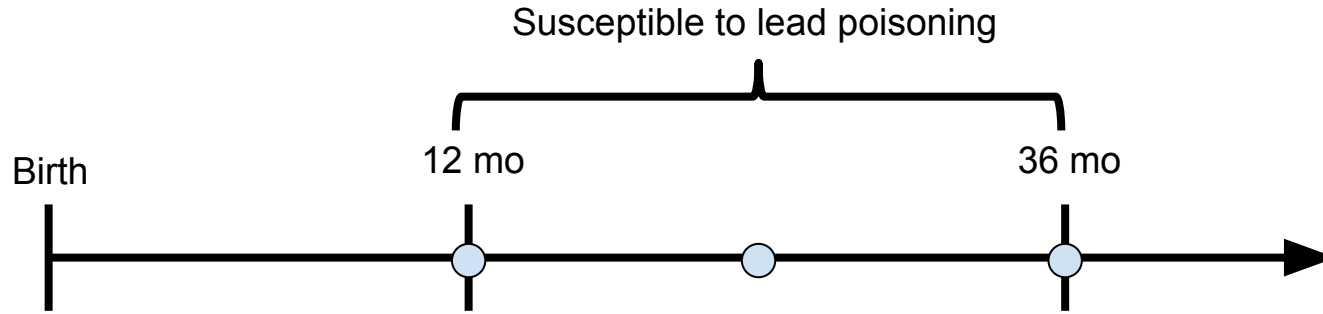
Interventions: Continuous

- Integrate with EMR to give health workers risk information.
- Mailing campaign to provide parents with information about lead hazards, inspections, funds for remediation.

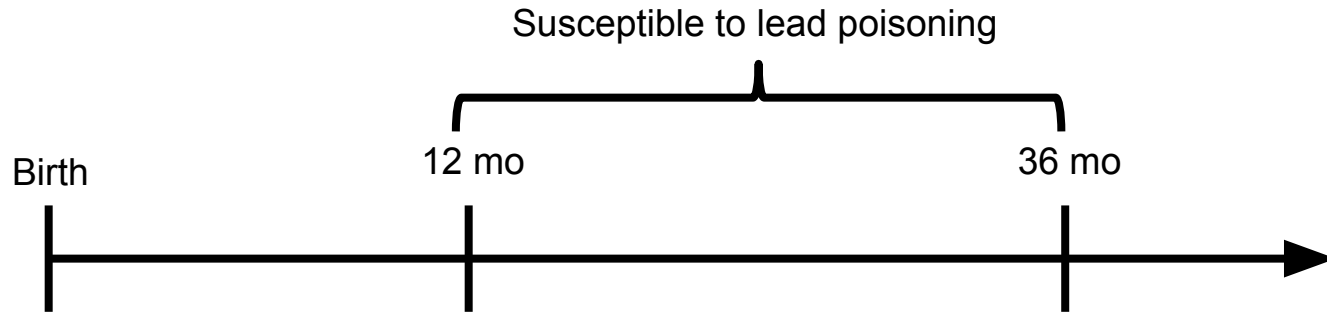
Interventions: Preventative

- Applicable to infants under 12 months, including in utero.
- Investigate (and remediate) home *before* child is susceptible to poisoning.

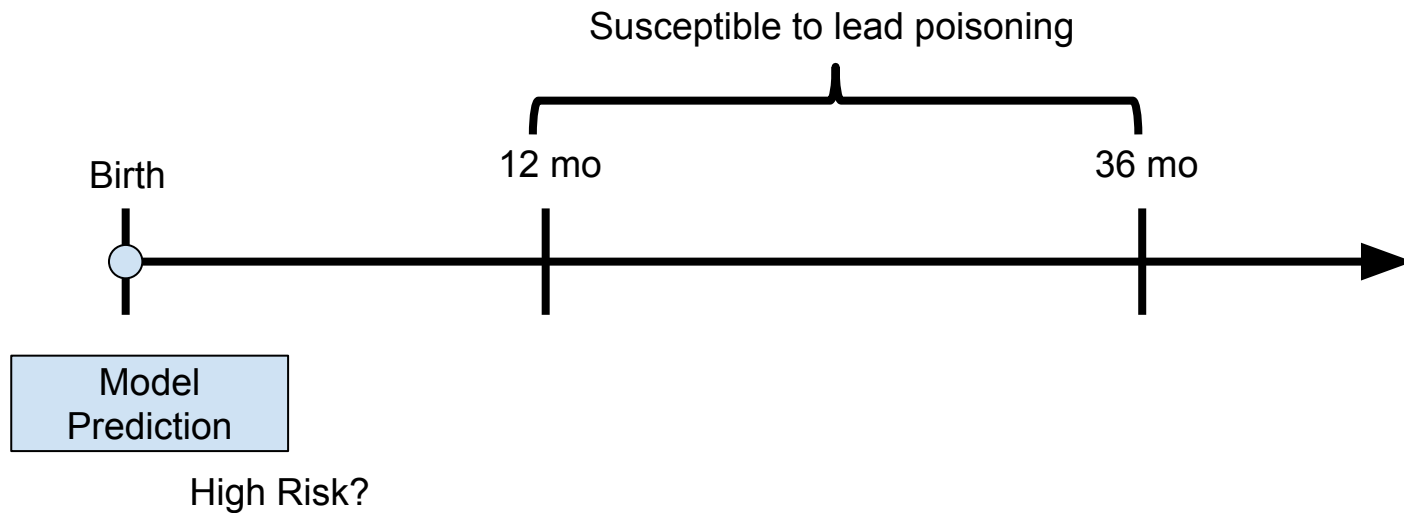
Current Approach



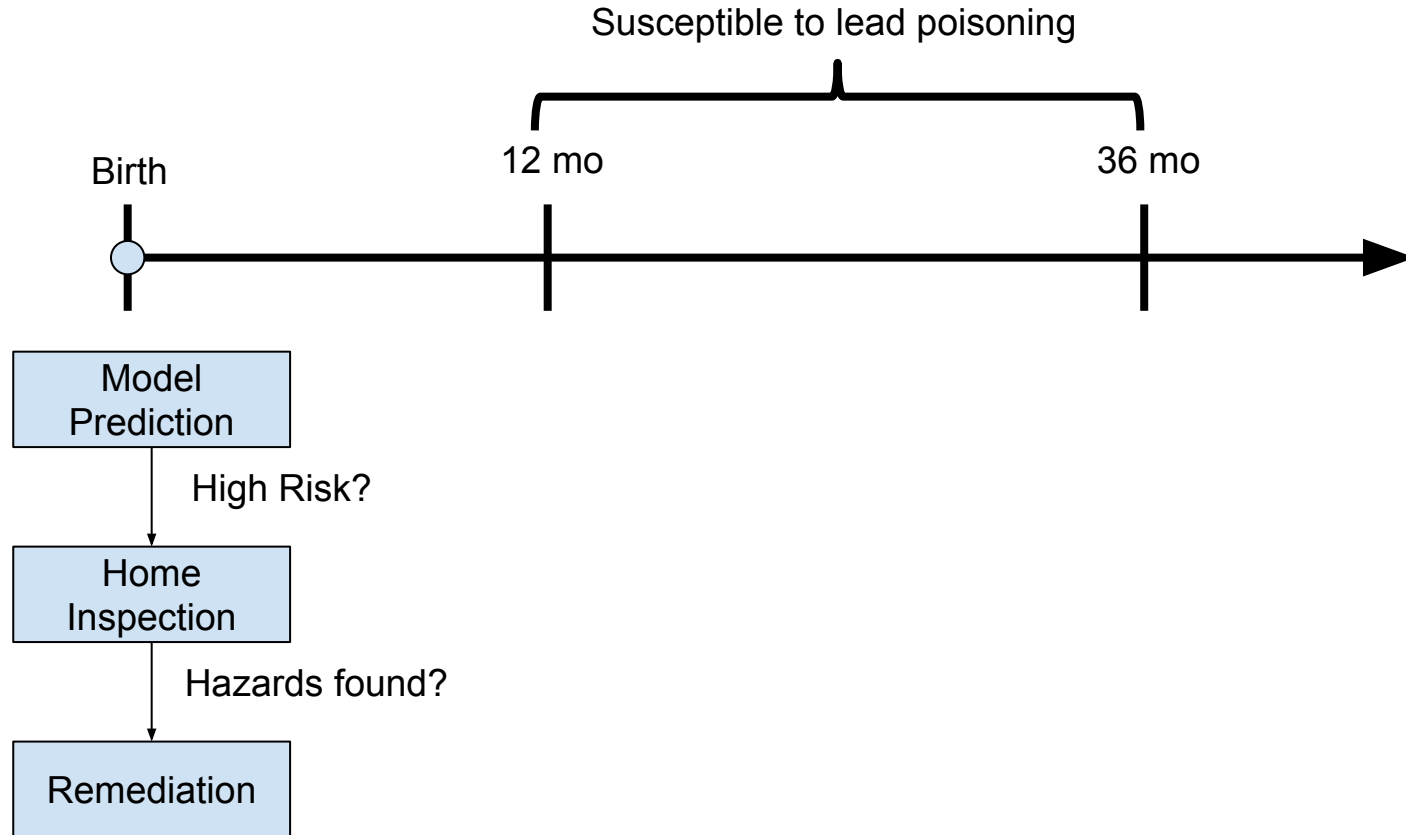
Predictive Approach



Predictive Approach



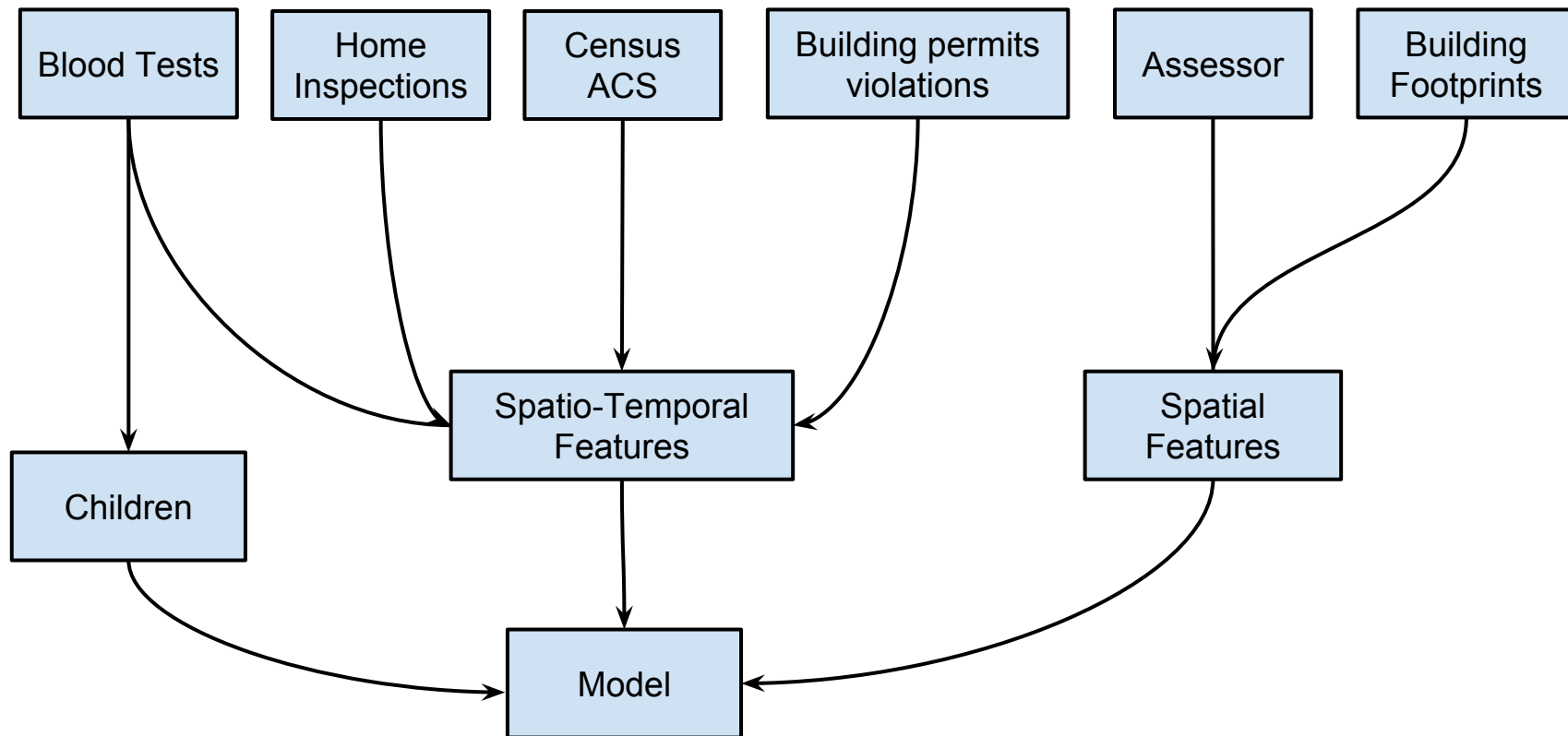
Predictive Approach



What is Machine Learning?

- Models optimized for predictive performance, rather than explanatory power.
- Primarily interested in predictions (y), rather than regression coefficients (β) or variance (Σ).

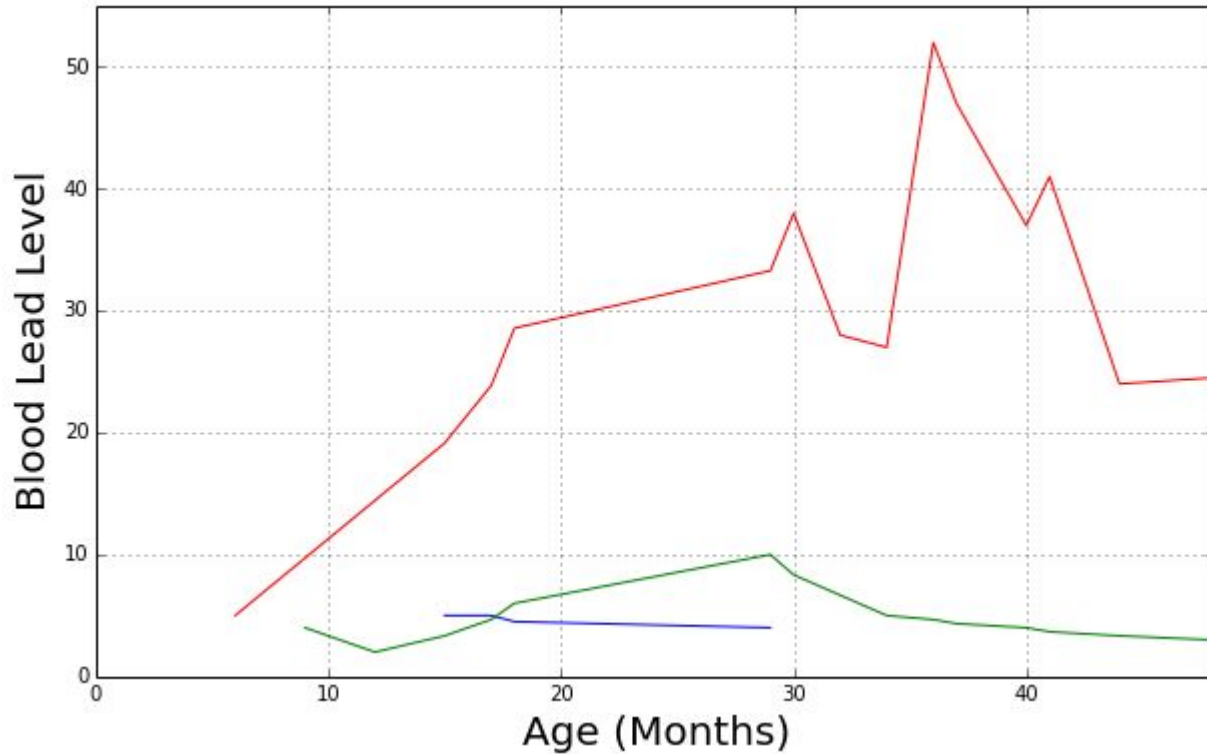
Model Data



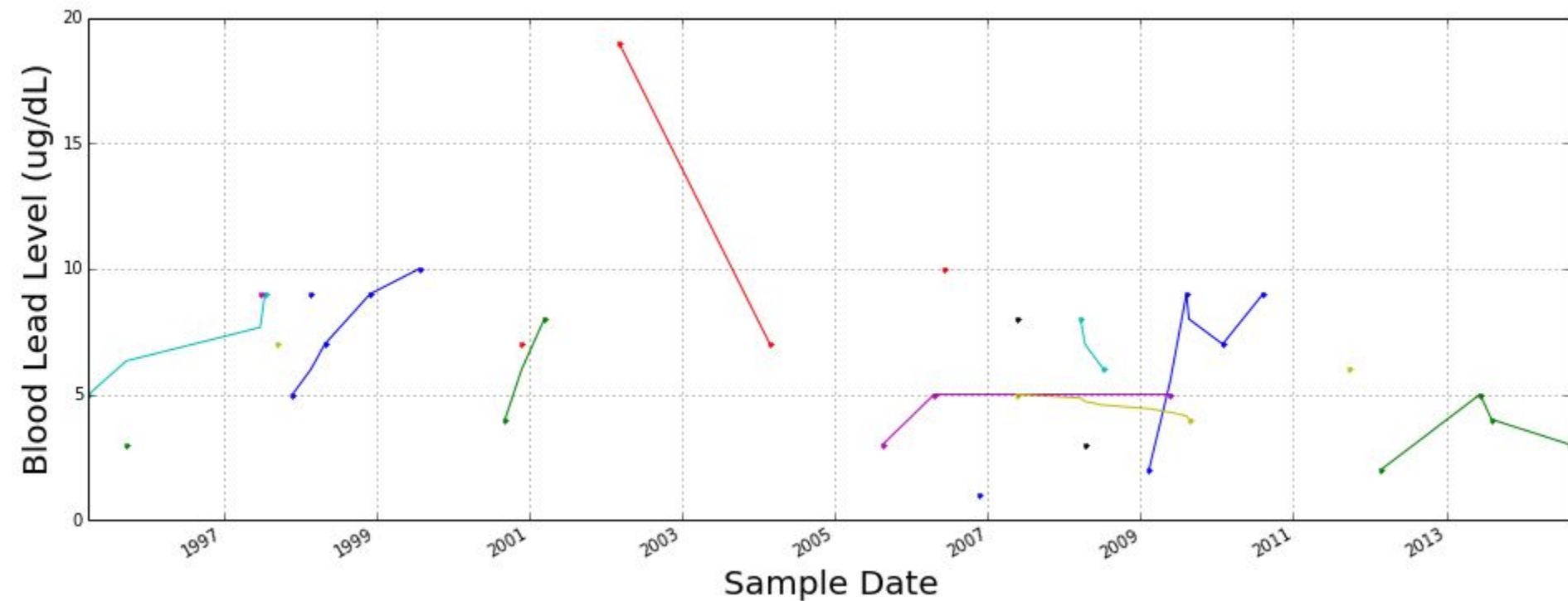
Data Quality: Record Linkage

| First | Last | Date of Birth |
|-------|--------|---------------|
| Eric | Potash | 1/1/2010 |
| Erc | Potash | 2/1/2010 |
| Joe | Walsh | 1/1/2010 |
| Joe | Walsch | 2/1/2010 |

Lead Level Trajectories



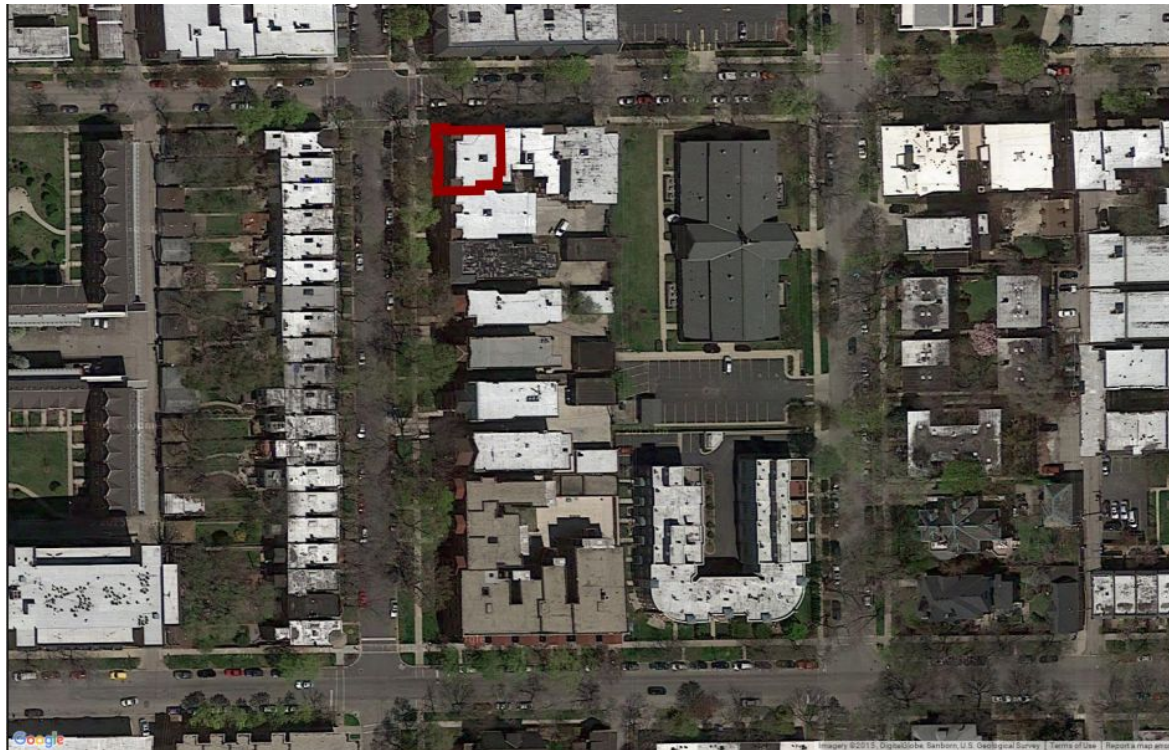
Lead Levels at a Single Address



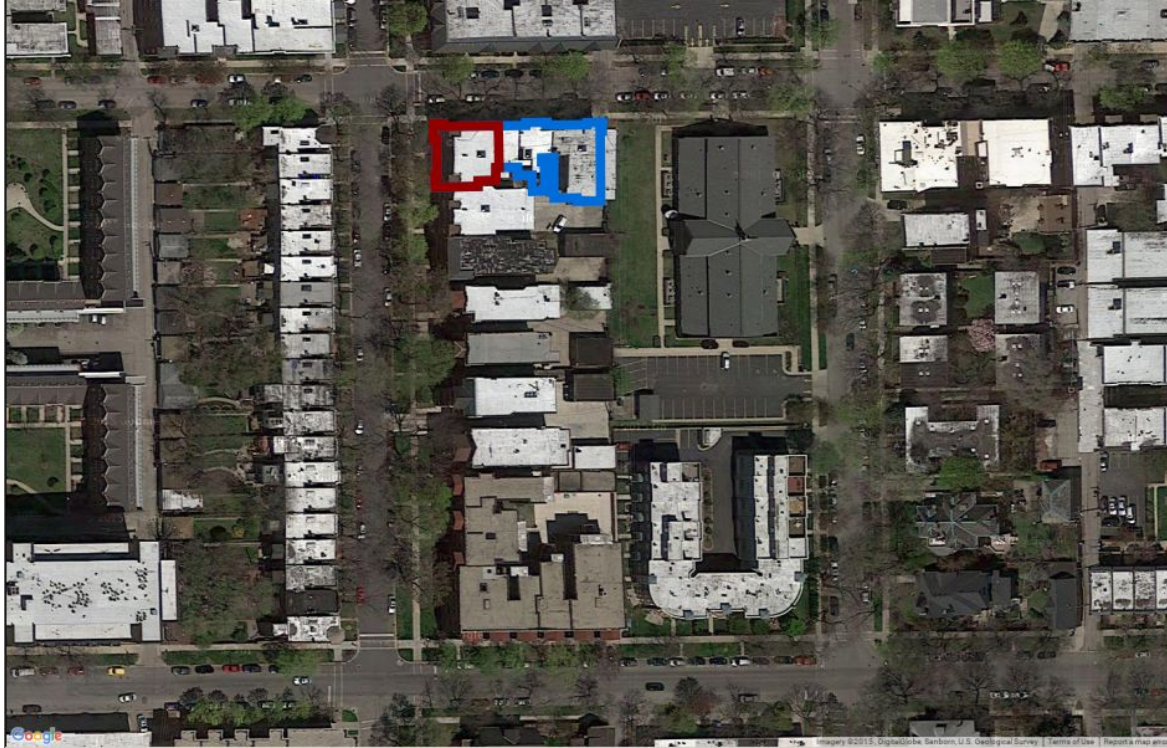
Spatial Levels



Spatial Levels: ~600k addresses



Spatial Levels: ~500k complexes

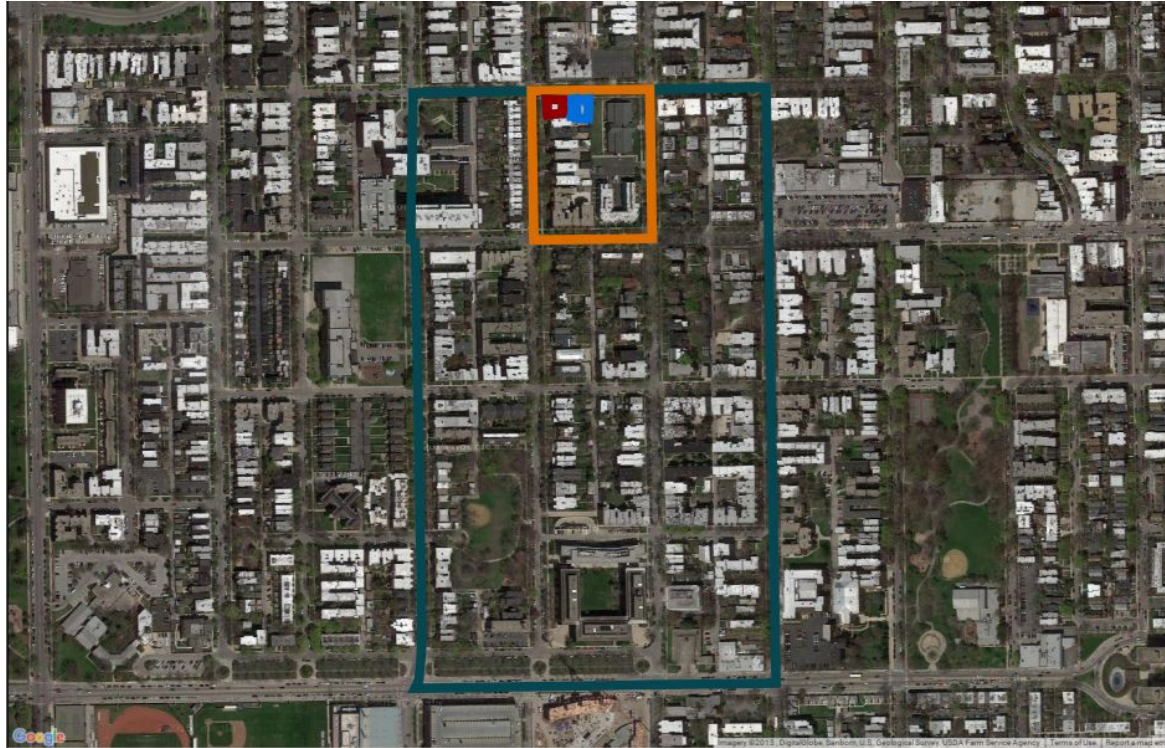


Spatial Levels: ~40k census blocks

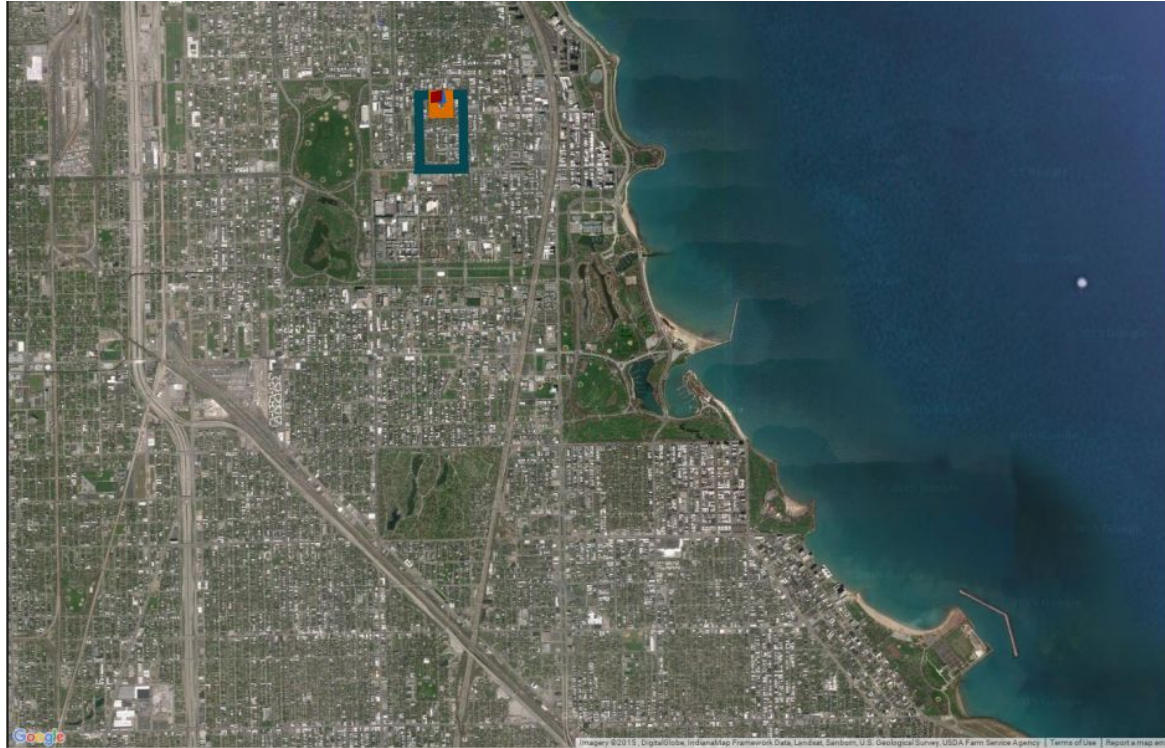


An aerial photograph of a city grid, likely New York City, showing a dense arrangement of buildings and streets. A red and blue location pin is placed on a building in the upper-middle section of the image, which is also highlighted by an orange rectangular box. The surrounding area includes various commercial and residential structures, green spaces, and a river or canal on the right side. The Google Maps logo is visible in the bottom-left corner.

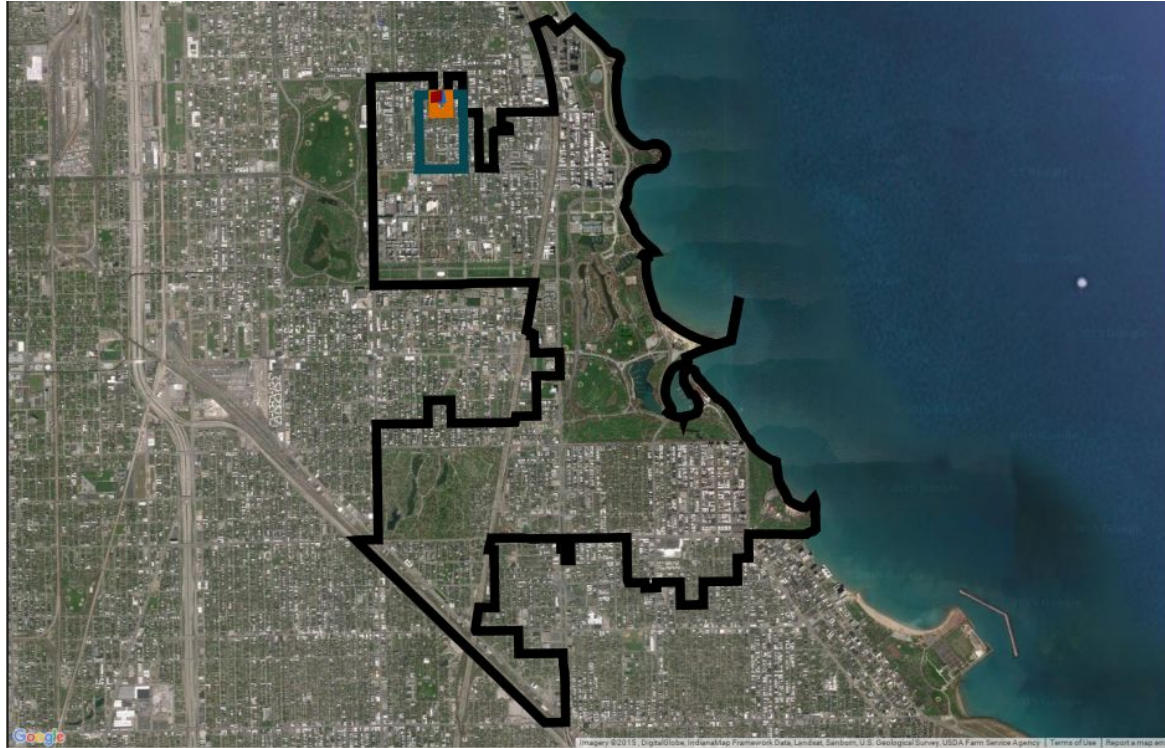
Spatial Levels: 800 census tracts



Spatial Levels: 800 census tracts



Spatial Levels: 50 wards



Spatio-temporal Aggregation

- 5 spatial levels
 - Address, Complex, Block, Tract, Ward
- 6 time periods
 - 1,3,5,7,9 years and all time
- $5*6= 30$ spacetime aggregations levels

Spatio-temporal Variables

- ~2k potential features, use ~1k features

| feature | description |
|----------------------|---|
| count | number of tests |
| tested | whether there has been a test |
| poisoned | whether there has been a poisoned test |
| ebll_count | number of poisoned tests |
| ebll_prop | proportion of poisoned tests |
| avg_bll | average blood lead level |
| median_bll | median blood lead level |
| max_bll | maximum blood lead level |
| min_bll | minimum blood lead level |
| std_bll | standard deviation of blood lead level |
| kid_count | number of children tested |
| kid_ebll_here_count | number of children with poisoned tests |
| kid_ebll_here_prop | proportion of children with poisoned tests |
| kid_ebll_first_count | number of children with first poisoned test |
| kid_ebll_first_prop | proportion of children with first poisoned test |

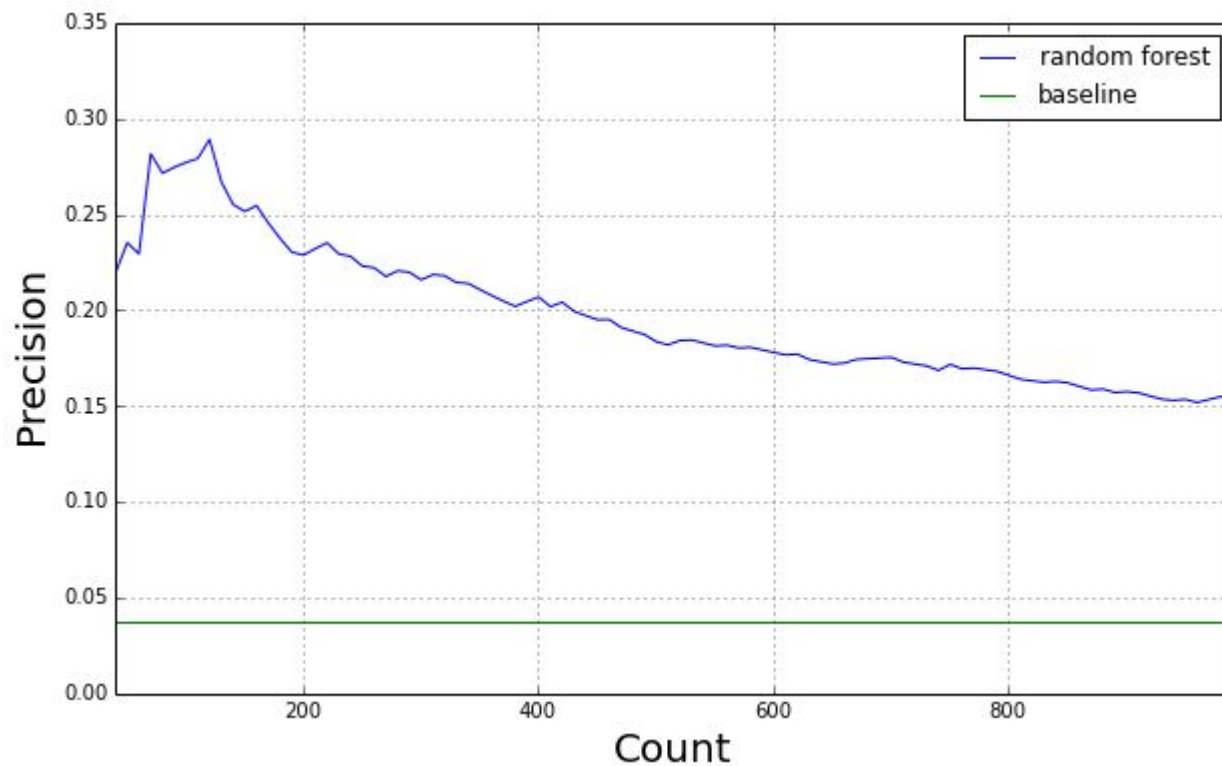
Cross Validation

- Use a temporal cross validation.
 - Historical reenactment
- Train the model on data until “today”, e.g. 2013-01-01.
- Test on kids who are as of 2013-01-01 under 12 months

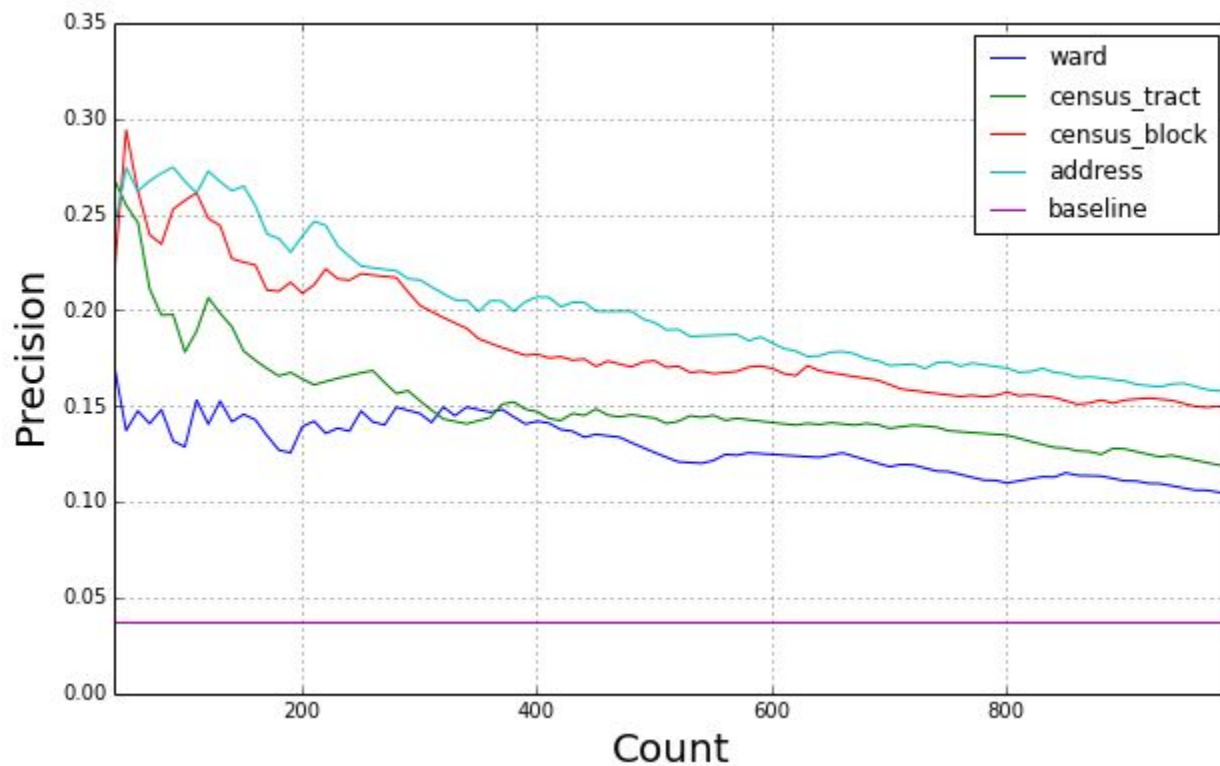
Evaluation Metrics

- Our partners have limited resources for interventions.
- A useful metric is the “precision at k”
 - e.g. we take our $k=1000$ highest risk scores and predict them as poisoned and calculate precision.
- Compare model’s performance with a random baseline
 - e.g. predict random 1000 children as poisoned.

Model Evaluation



Model Evaluation



Appendix: Feature Selection

| | c |
|--|----------|
| name | |
| tract_acs_health_pct_Insured_public | 0.054714 |
| tract_acs_health_pct_unInsured | 0.057941 |
| kld_sex_M | 0.064409 |
| tract_cumulative_test_count | 0.075659 |
| address_Inspection_Init_days | 0.080799 |
| address_test_ebll_kld_ratio | 0.169231 |

| | c |
|--|-----------|
| name | |
| kld_blrth_days | -0.932934 |
| address_building_year | -0.198676 |
| tract_cumulative_ebll_kld_count | -0.084628 |
| tract_buildings_avg_year_built | -0.073164 |
| tract_acs_edu_pct_advanced | -0.065776 |

| | Importance |
|--|------------|
| feature | |
| kld_blrth_days | 0.016776 |
| kld_test_number | 0.012587 |
| kld_date_of_blrth_month | 0.007908 |
| census_tract_tests_3y_kld_ebll_ever_prop | 0.007549 |
| census_tract_tests_3y_kld_ebll_here_prop | 0.006996 |
| census_tract_tests_1y_kld_ebll_ever_prop | 0.006376 |
| census_tract_tests_3y_ebll_prop | 0.006236 |
| census_tract_tests_3y_kld_ebll_future_prop | 0.006179 |
| census_tract_tests_3y_kld_ebll_first_prop | 0.006031 |
| census_tract_tests_3y_address_ebll_prop | 0.005950 |
| census_block_tests_5y_kld_max_bll_avg | 0.005675 |
| census_tract_tests_3y_kld_max_bll_avg | 0.005451 |
| building_assessor_age_min | 0.005109 |
| building_assessor_age | 0.005091 |
| building_assessor_age_max | 0.005080 |