**Who to send this to right away?**

Gizmodo - done

Tech Crunch reporter

<http://theconversation.com/us>

Ann Verhey Henke, Alison Miller, Marc Zimmerman (SPH) - done

Rebecca Cunningham (UM Med),

Marty Kaufman (UM-Flint GIS) - done

Ben Geiger, Jim, and State people

Bryce (DEQ) -

Analysis and writing ideas

Consider real estate variables: Zillow - time of last sale?

Splines and smoothing - http://www.r-bloggers.com/thats-smooth/

Quantile Regression

* Quantile Regression Forests <http://www.jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf>
* Non parametric Regression Quantile <https://cs.stanford.edu/~quocle/jmlrq.pdf>
* Bootstrapping Quantile Regression for standard errors <http://www.arts.uwaterloo.ca/~pchausse/UWERG/Doc/2013-11-13.pdf>
* Boostrap for QR for standard errors: R function in package <http://www.inside-r.org/packages/cran/quantreg/docs/boot.rq>
* ME: Bootstrap the sample 100 times. But append each one into a large dataset 100x the size of the current one. Place an variable for ‘Bootstrap\_Sample’. Then do geom\_quant(group=Bootstrap\_Samp) to get 100 lines

Missing Data R packages - [datascienceplus.com/imputing-missing-data-with-r-mice-package/](http://datascienceplus.com/imputing-missing-data-with-r-mice-package/)

### Conversation with Braun and Thoughts - 2016-05-03 7pm

## What models to run?

\* lm v fixed effects v random effects

\* include or exclude Year\_Built\_NewCat

\* all data v at least 1 repeat v at least 2 repeat

\* all dates v only 2016 Feb and later

\* 90th v mean

# What makes teh house get tested 2 times vs 1 time?

\* MCAR -

\* MAR - Missing at Random: Based on observables, Age of home? Neighborhood?

\* MNAR - Missing Non-Random: based on unobservable, e.g., trust in government

\* - Propensity score - probability that we have that observation or not? P

\* Poisson regression for number of repeated samples. Modeling the probability of having x number of observations.

\* or just binomial

\* - PRopensity to join in is a function of time also!

\* Cant ignore the zeros.

# Lead over time, split into group by age of property

\* when mixed together it's hard to see the pattern (or believe it!)

\* when when you split pre/post 1950 houses, you can see one of them has the pattern clearly

\* facet = pre/post 1950. x = date of sample. y = lead

# UPDATED Residential Testing Data

# Residential testing over time vs sentinel over time

# Sentinel only over time with error bars

# better histogram of lead[ INSERT HISTOGRAM OF LEAD LEVELS HERE]

# Visualize model results

# service line lead

# check for any senesitive informaiton

Done

# follium map for lead only

Summary Notes

There is lead in Flint’s water. And we know that leads to more questions than answers: Where it is? Which homes are most at risk? When will the lead levels decrease?

We want to shed light on these questions with data. Instead of heavy-handed headlines, we will rely on hard data. Using data from diverse sources, we use cutting edge-methods in data science and statistics.

The crisis is also one of transparency of information. We’d like to bring the key information to the citizens of Flint as clearly as possible.

What we want to do in this short writeup is give some early results that help to understand the lead level readings that are being continuously collected in Flint.

For questions about health and getting obtaining lead test kits your home, visit [Michigan.gov/flintwater/](http://www.michigan.gov/flintwater/).

## Diverse data sources

The figures and data below are based on several datasets. Instead of only using one or two sources of data, we are connecting the dots between many.

\* *Residential Testing data*: Flint residents continue to submit water samples ot the Department of Environmental Quality (DEQ), which tests their water and posts results to the [Michigan.gov/FlintWater/ website.](http://www.michigan.gov/flintwater/)

\* *Sentinel Sites Testing data*: Available at [Michigan.gov/FlintWater/ website](http://www.michigan.gov/flintwater/).

\* *Parcel data* obtained from the City of Flint

\* *Service Line data* provided by City of Flint and UM-Flint GIS Center https://www.umflint.edu/gis

\* *Fire Hydrant data* provided by City of Flint

## Value of Residential Testing Data

Most of the media attention is based on data with over 600 houses samples repeatedly (Sentinel Site data). But we are using more than 8,000 unique houses contributing over 15,000 total samples.

There’s more value in that data. That’s what we will explore here.

## **## Where is there high lead?**

Elevated lead readings are occurring throughout the city. They appear to be quite geographically diverse. A location is determined to have *elevated lead* if the DEQ recorded an amount of **15 parts per billion** in a water sample (using EPA standards).

The map shows all the parcels with great than 15ppb, which represent 8% of all samples.

<img src="../Images/lead\_readings.png" width=800px/>

## 

## **Elevated lead levels are less than 10% of readings**

* High lead level readings (i.e. greater than 15 parts per billion) are only about 8.3% of the dataset
* Dangerous lead levels (greater than 50 ppb) make up 3.1% of the data.
* Very dangerous lead levels (greater than 150 ppb) make up 1.2% of samples.
* The vast majority ( 74% ) of the readings are measuring less than 4 pbb
* The 90% of the readings are 12 ppb or less.

[ INSERT HISTOGRAM OF LEAD LEVELS HERE]

## **## Note about data sources**

## Residents provide sample on a volunteer basis. The City and State encouraged them to do so.

But the residential testing data provides a wide range of types of properties and areas and lead levels. The data can help us understand

## 

## **## What helps us predict to lead?**

[avoid CAUSAL language]

The lead readings are known to be highly variable and depend on a number of factors including the way the test was conducted, the time of the day, and the number of hours during which water sat idle in the pipes. The factors of interest are the attributes of the property, including the age of construction, condition of the property, when in 2015-16 the sample was taken, material of in-house plumbing, material of service line pipe connecting house plumbing to street pipes).

There seem to be lots of relevant factors. The \*Property Age\* seems to be very important.

<img src="../Images/lead\_by\_yearbuilt\_residential\_tests\_annotate.png" width=600px>

### 

### **## Important variable: the age of the property**

We observed that one attribute of the parcel that is **very** correlated with lead levels is the **year during which the property was built**. We grouped the properties together into six categories, based on the date of construction. We plot the mean of the logarithm of the lead level for each epoch. You'll notice there is a sharp decline for more buildings built after 1950: for those built in 1950 or older, 10% vs 6% of the younger properties of readings are above 15 ppb.

### **## How do lead service lines affect elevated lead readings?**

[## Are the elevated lead readings caused by lead service lines?] WATCH OUT FOR CAUSAL LANGUAGE

What percent are lead? 8%

23% are unknown.

The lead service lines play a role, but not as much as you would think. We still see high lead readings even with service lines are made of copper, zinc, and other materials.

**##** Interactive map of high lead readings for homes with various service lines

IFrame('http://web.eecs.umich.edu/~jabernet/FlintWater/service\_line.html', width=700, height=700)

### **## Can we predict high lead levels?**

In our data exploration, we were able to apply various machine learning algorithms to the data to predict where we think elevated lead levels might be found. We use the 15000 readings for 8000 properties to predict which of the 55000 properties might also have high led.

Here are the locations of those properties with highly-probable elevated lead (> 15 parts ber billion).

### Interactive map of \*\*predicted\*\* locations of elevated lead

How well can we predict lead lead levels? Confusion matrix: Of all

**Make sure to include something on each of these:**

Map of existing high lead

Map with Predict which houses have high lead. From mode.

Map with Service line and Lead

Summarize other key variables (house age, prop condition, service line material) on Map by different level (parcel, census block, ward, zip)

Plot Res vs Sent samples Over time.

Compare residential to sentinel by each car distn and joint distn (e.g., age of house)

Lead by House year built for Res and Sent samples

Percentage of ‘old houses’ sampled through 2016