

Open Data, Place-Based Research, and Getting the Measures We Need To Ask the Right Questions

NIH Office of Disease Prevention's
Methods: Mind the Gap Webinar Series

Steve Mooney



DEPARTMENT OF EPIDEMIOLOGY

UNIVERSITY of WASHINGTON
School of Public Health

Bio



1998-2010

Job: Programmer, Microsoft and some startups

Questionable fashion choices: Dale Cooper hair, circular wire-rimmed glasses



2010-2016

Job: Grad Student, Columbia

Questionable fashion choices: thin glasses, sideburns

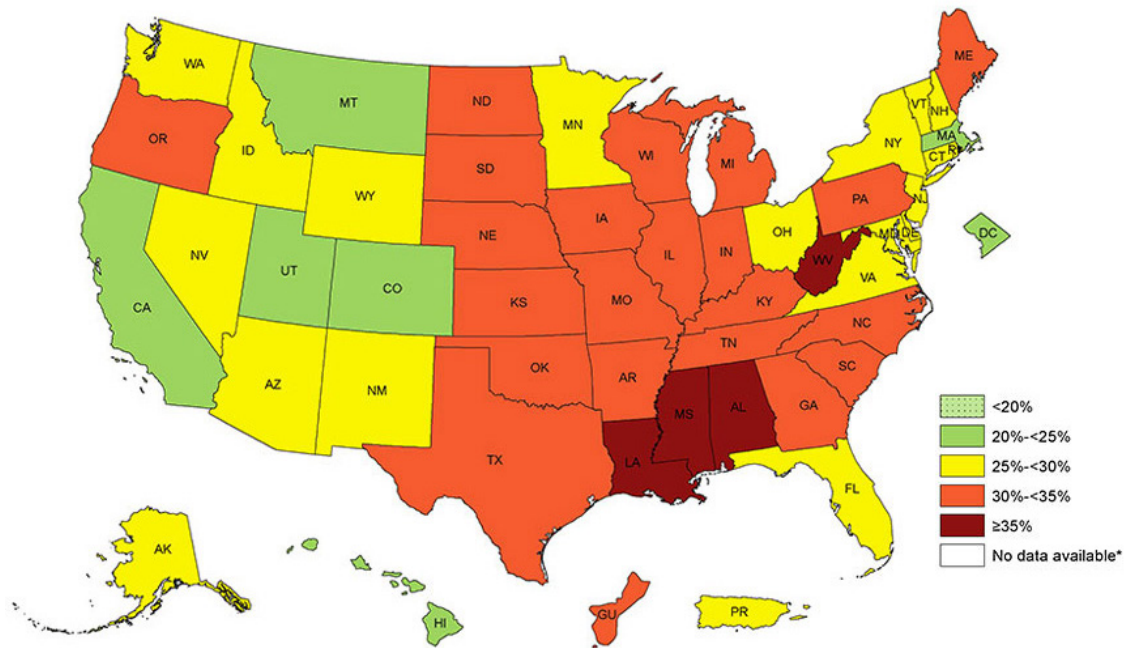


2016 – present

Job: Post doc/Acting Assistant Professor, University of Washington

Questionable fashion choices: thick black glasses frames that yell “I am an intellectual, dammit!”

Place-based effects

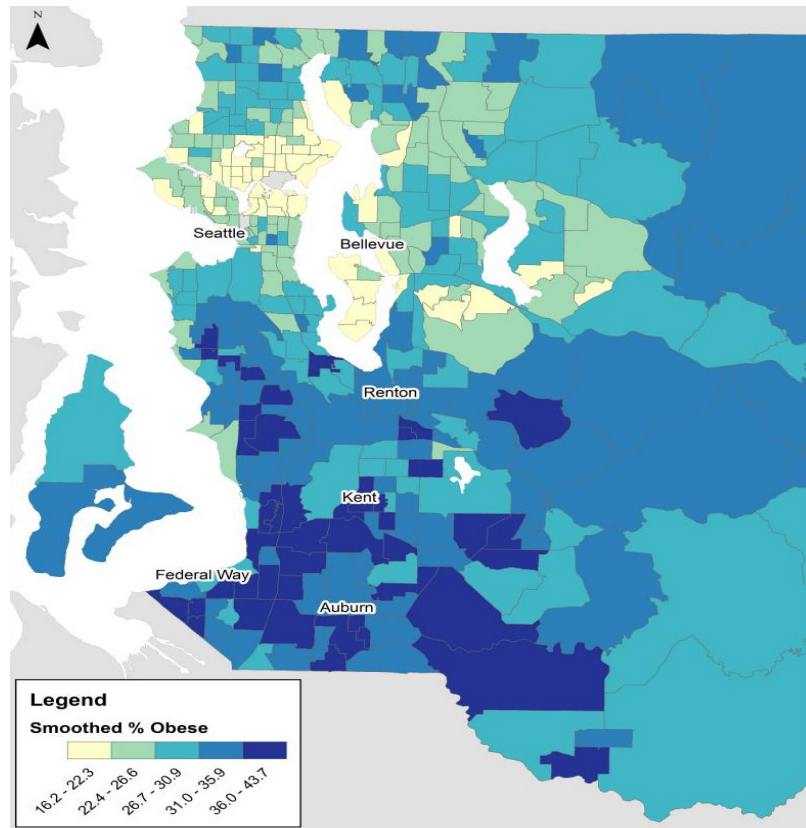


Darker Red = more obesity

Some aspect of each state?

Something aspect of the people in each state?

Place-based effects



Darker **Blue** = more obesity

Some aspect of each tract?
Something aspect of the people in each tract?

Outline

Overture: Informatics – what it means

Act I: Case study in constructing a measure

Act II: Case study in using multiple measures

Curtain Call: Some conclusions and next steps

Informatics

Formally, informatics is the science of information and information processing...

...but what I really mean is: the ways in which using algorithms in research affect both the data we observe and the questions we ask of it

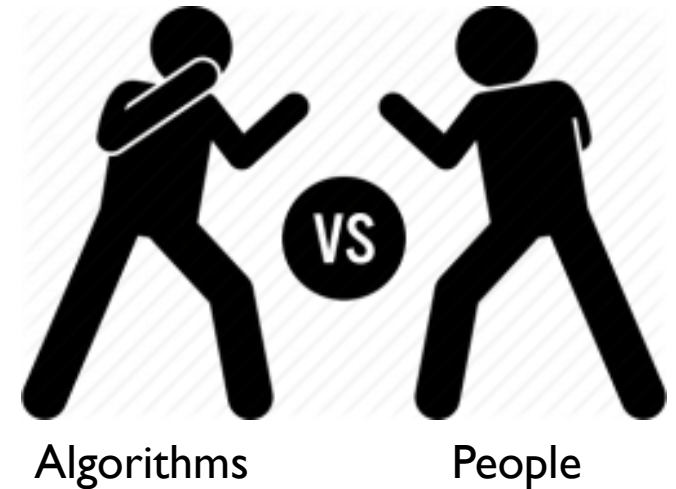
Algorithms vs. People

Compared to people,
algorithms are:

Fast

Scalable

Lacking in common sense



Algorithms vs. People

Compared to people,
algorithms are:

Fast

Scalable

Lacking in common sense



Algorithms vs. People

The key to working with algorithms is to add *your* common sense to *their* speed and scalability



Algorithms vs. People

That means you need to understand them enough to know where they diverge from common sense



Algorithms: why now

It's a lot easier to make more data when you don't have to code it by hand

The growth in algorithmic data collation is, in some sense, what puts us in the 'Big Data' era

Big Data for Epidemiology

Type	Examples
-omic	Whole Exome Sequencing, Metabolomics
Geospatial	Neighborhood Profiling
EHR	Patient records from a health care system
Personal Monitoring	FitBit Readings
Effluent Data	Google Search Trends

(Of course, these can be merged into even bigger data sets)

Big Data for Epidemiology

Type	Examples
-omic	Whole Exome Sequencing, Metabolomics
Geospatial	Neighborhood Profiling
EHR	Patient records from a health care system
Personal Monitoring	FitBit Readings
Effluent Data	Google Search Trends

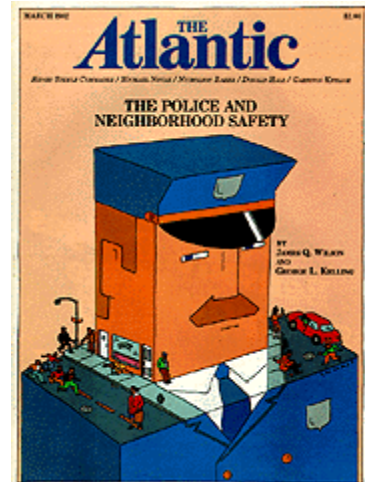
(Of course, these can be merged into even bigger data sets)

Act I:

A Case Study in Constructing a
Neighborhood Measure from 'Effluent
Data'

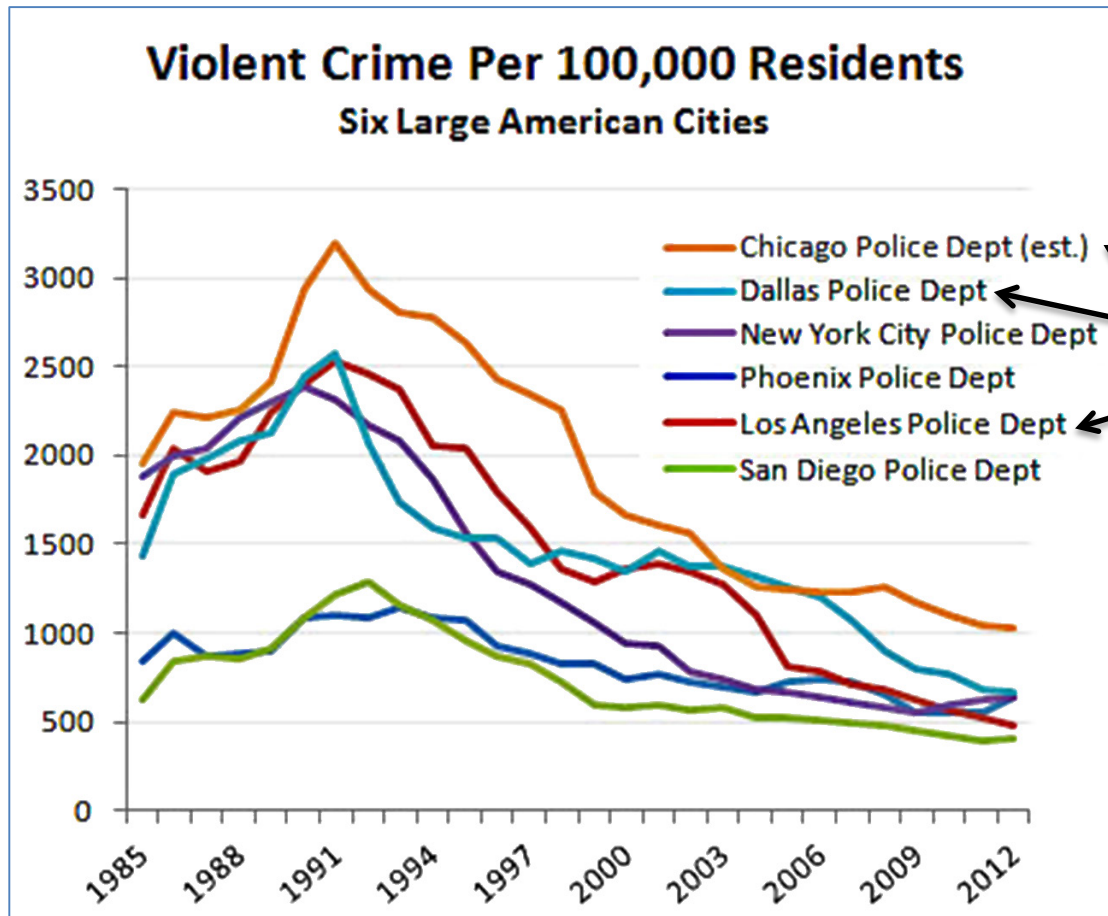
Physical disorder

Hypothesis underlying 'broken windows' policing



Wilson & Kelling wrote an influential popular press article promoting this idea in 1982

Broken windows policing



No broken windows policies

Graph: Mother Jones Magazine

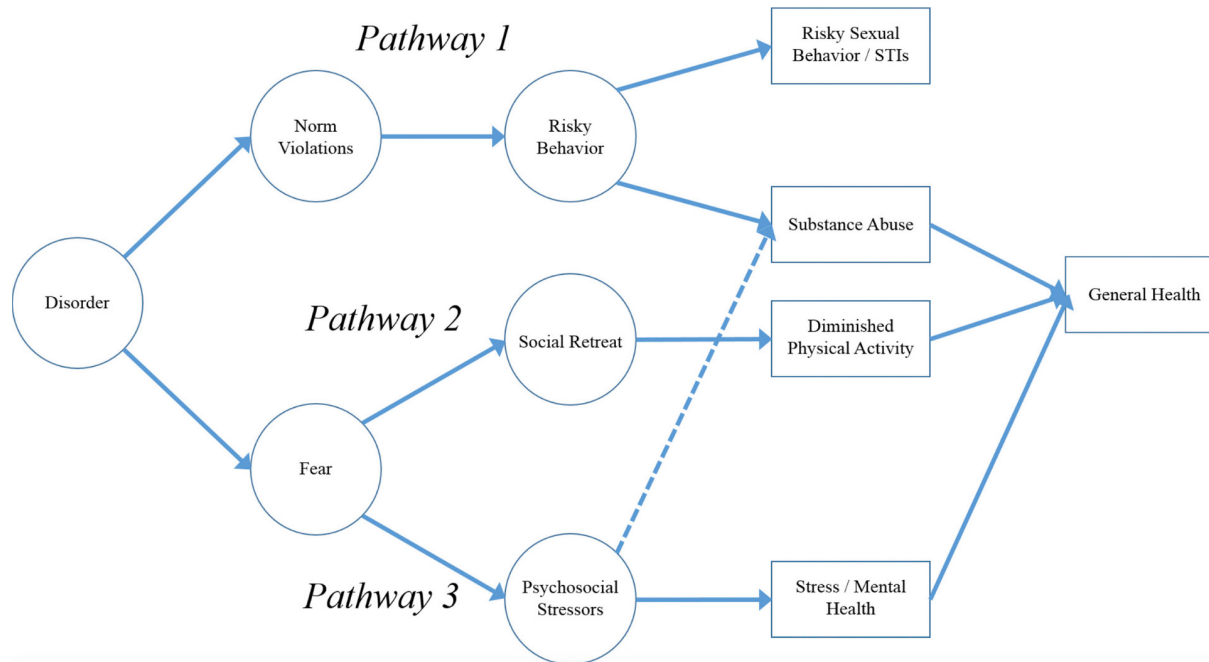
Disorder and walking

"Poor street lighting would prevent me from walking in the evening. Overgrown bushes, shrubs...sometimes you have abandoned homes, and the shrubbery has gotten out of control."



This shrubbery is out of control

Disorder and health more generally



We need measures of disorder to assess these pathways...

Measuring physical disorder

To look more closely at disorder as a theory, we have to *measure* the disorder in a location...

...one way to look at neighborhood infrastructure is on Google Street View

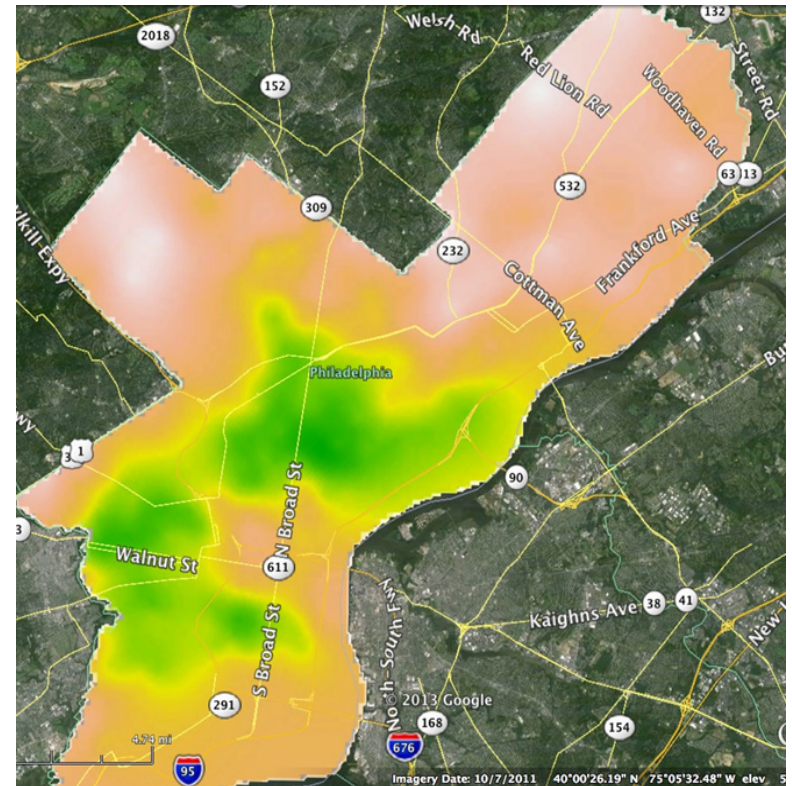
Measurement Example

How do we go from this...



Street imagery

... to this



Physical Disorder Measure

CANVAS



Disorder using CANVAS/Street View

Three step process...

Disorder Indicators

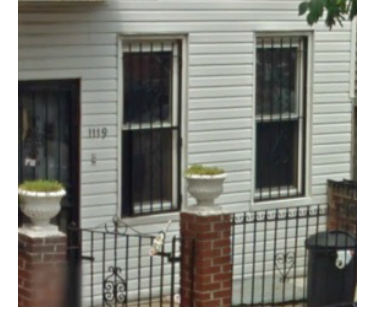
I. Measure indicators of disorder throughout the city



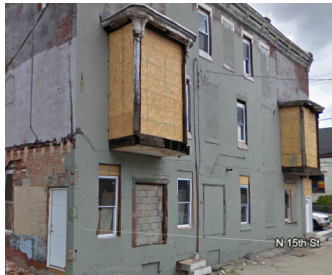
Empty Alcohol Bottles



Poor Maintenance



Bars on Windows



Abandoned Buildings



Abandoned Cars



Vacant Lots



Graffiti



Litter

Collecting data: graffiti

PHDCN.3: Is there graffiti, or evidence of graffiti that has been painted over, on buildings, signs, or walls?

- ☐ Yes
- ☐ No
- ☐ Cannot Tell

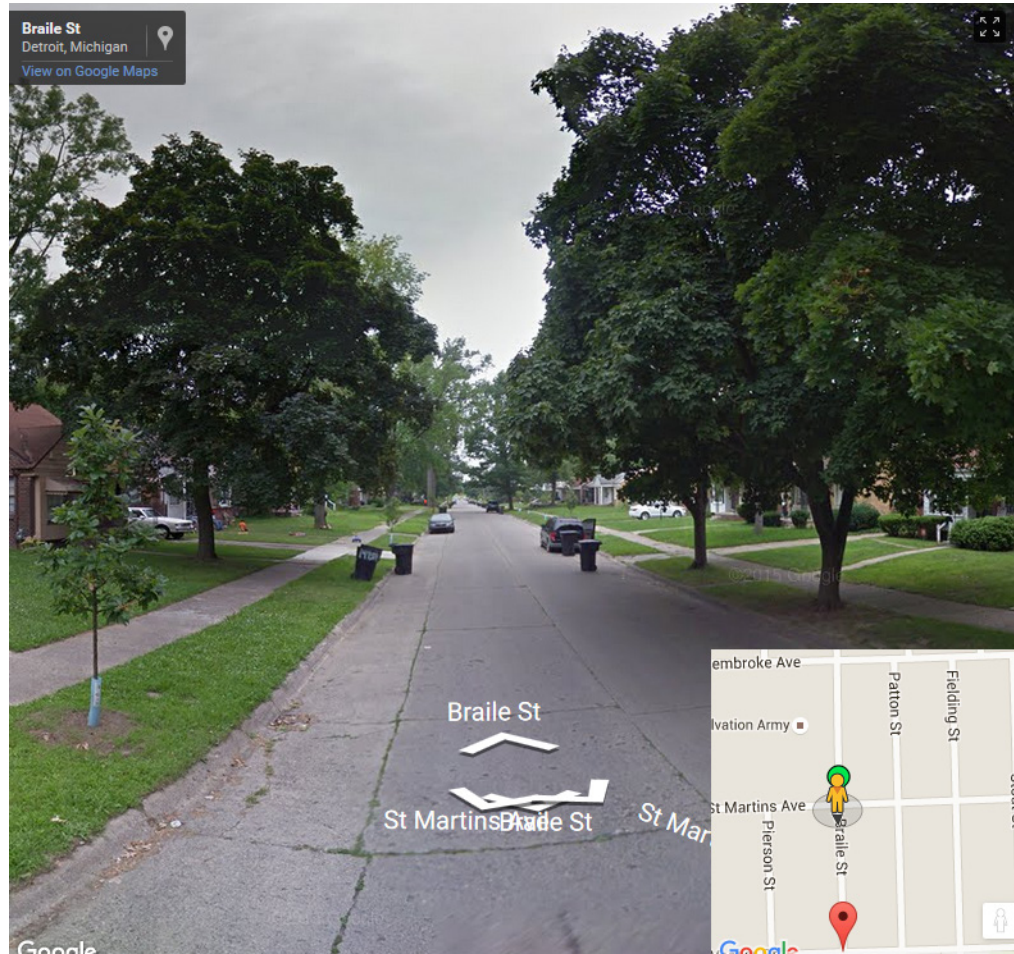


Yes

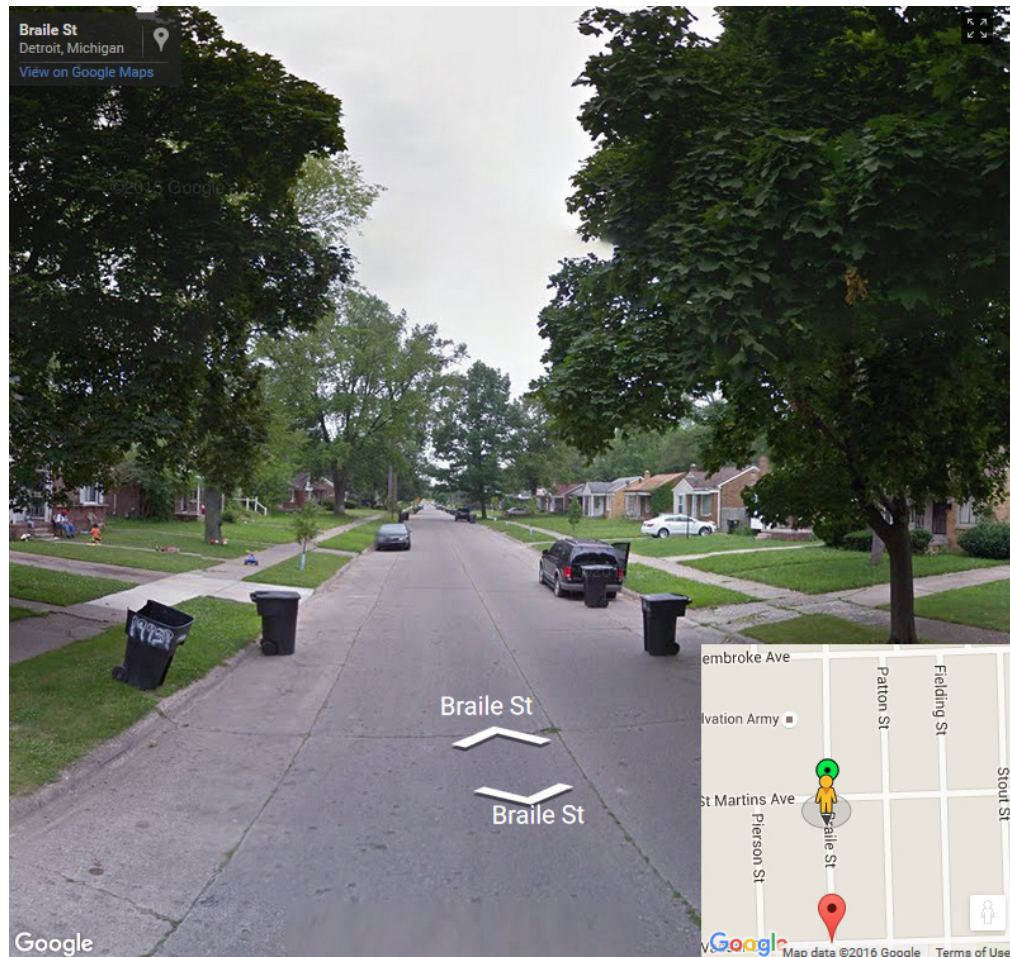


No

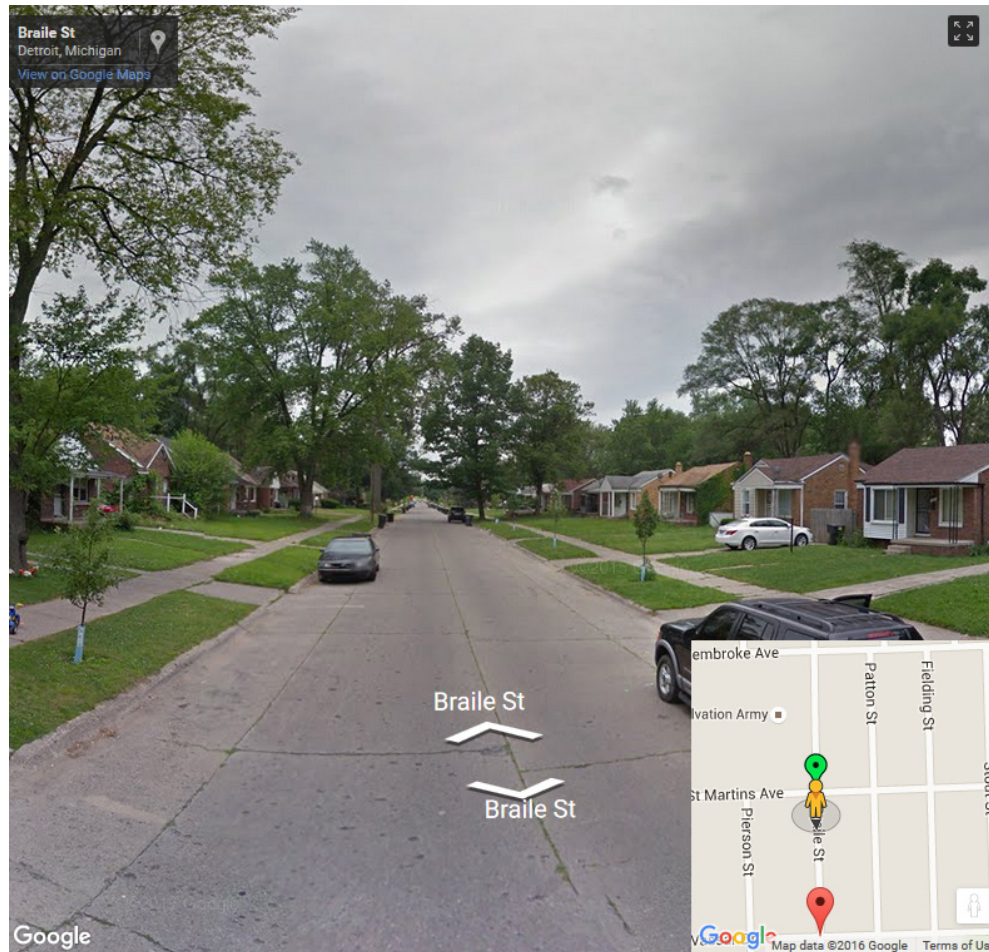
CANVAS/Street View



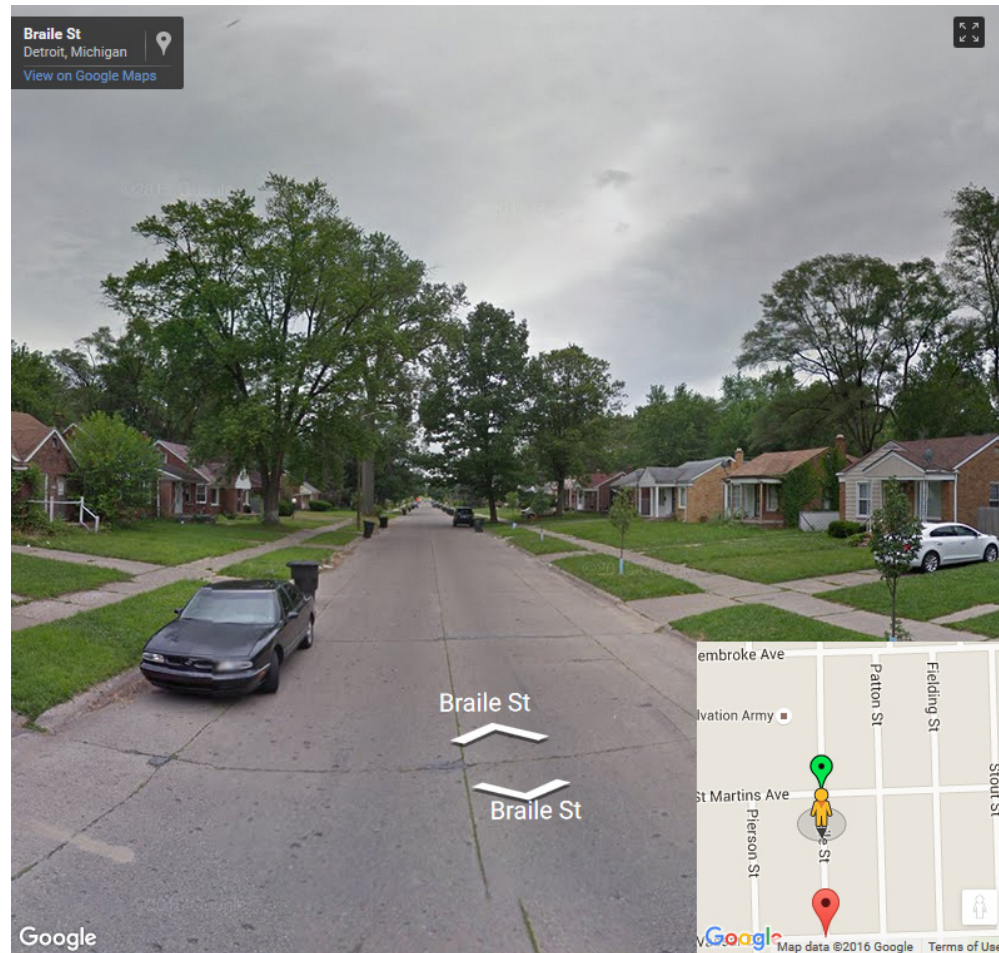
CANVAS/Street View



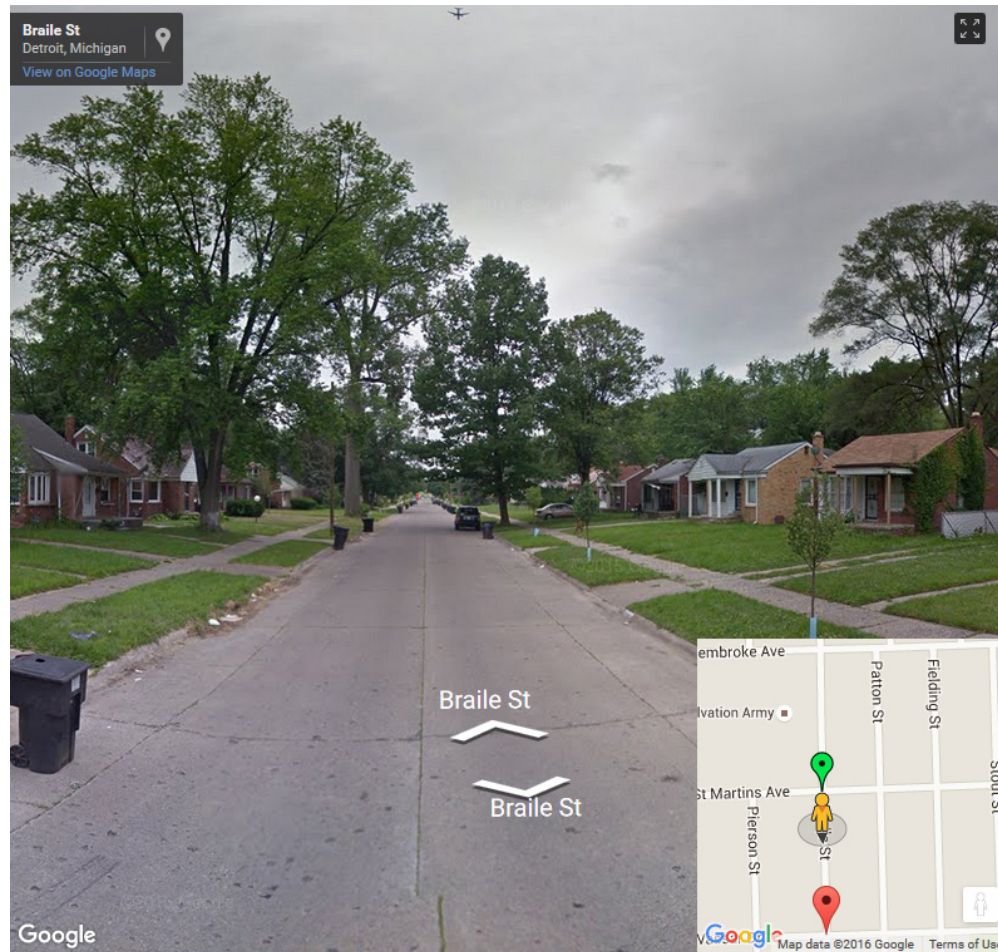
CANVAS/Street View



CANVAS/Street View

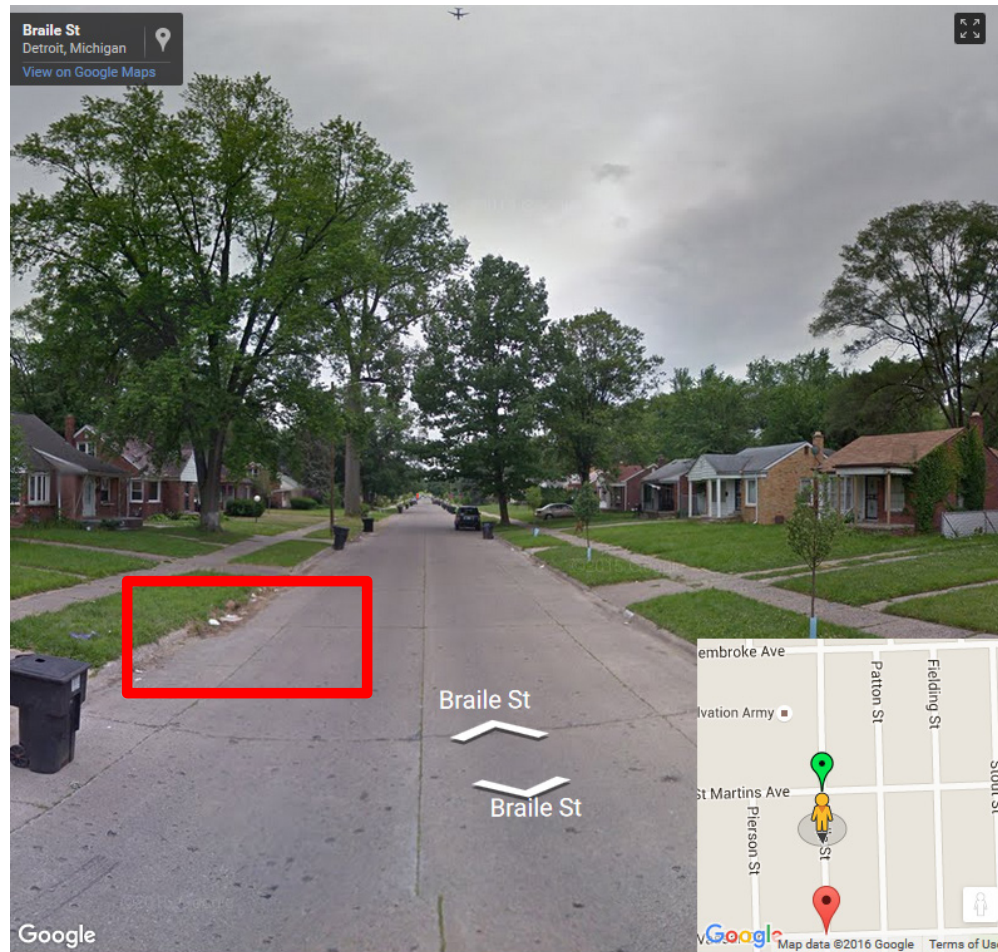


CANVAS/Street View



CANVAS/Street View

Litter



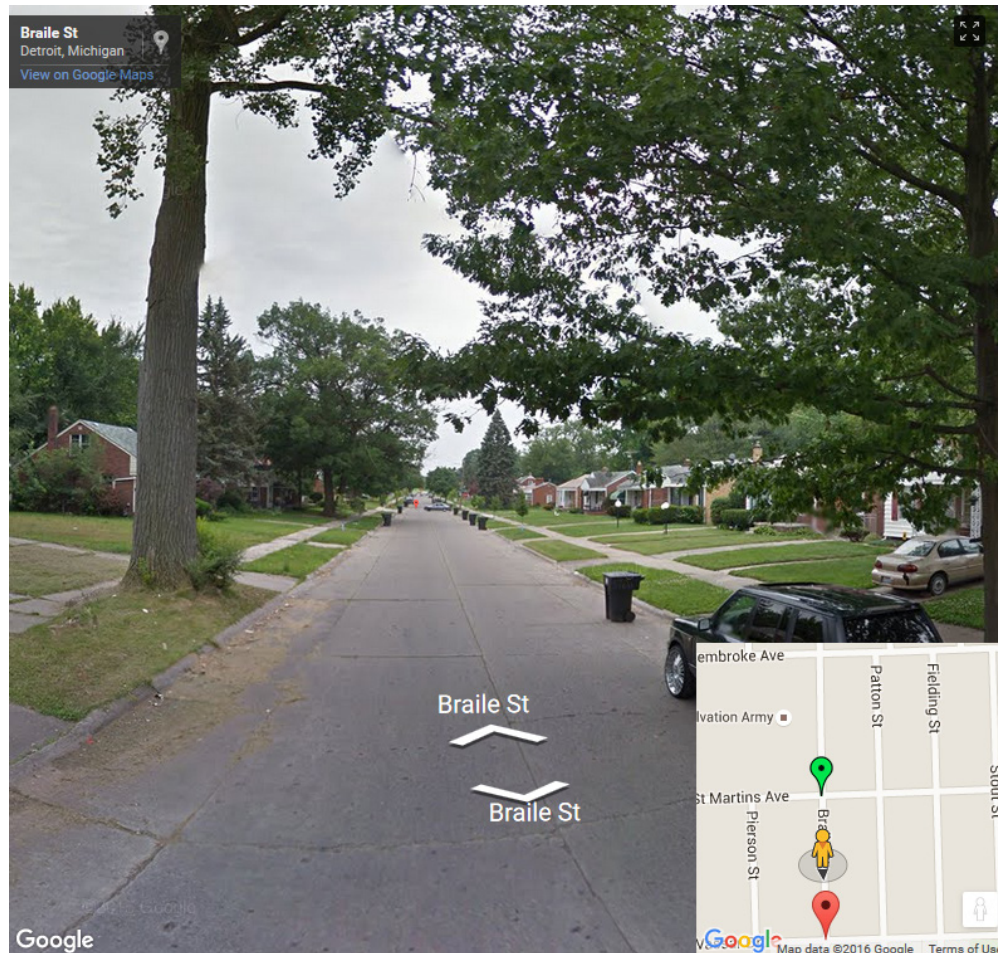
CANVAS/Street View



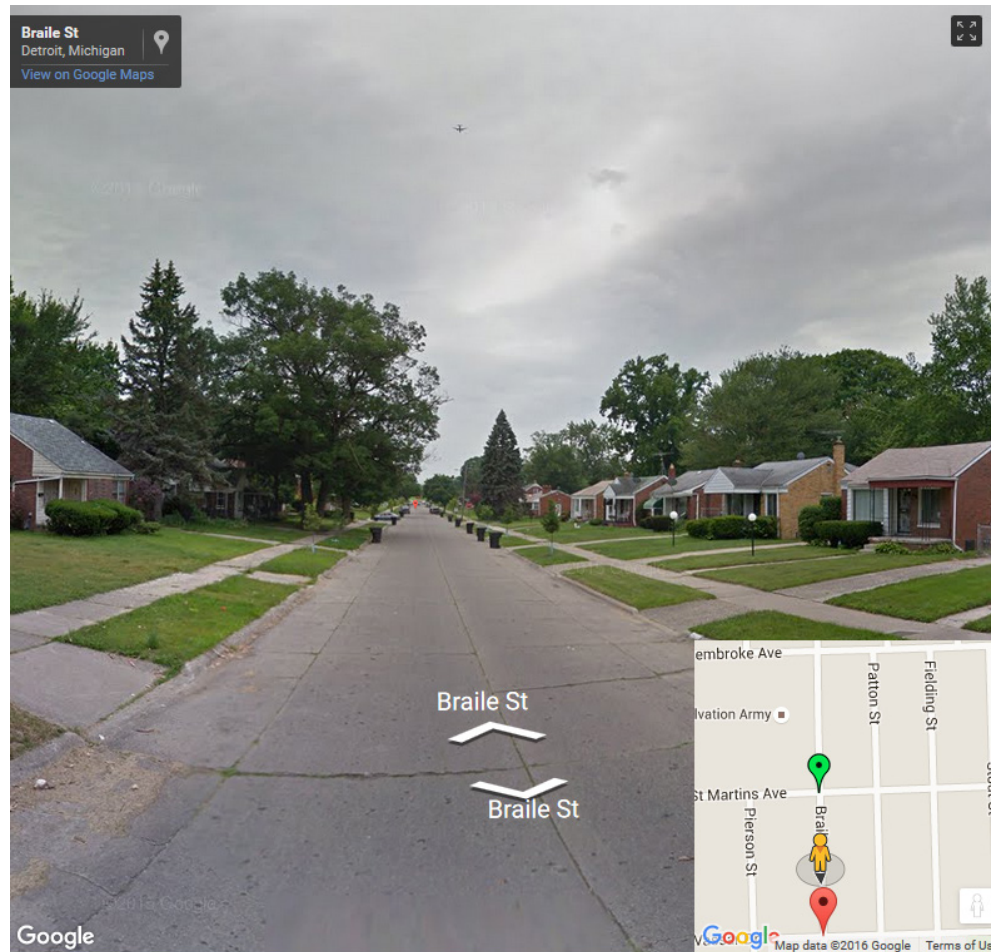
CANVAS/Street View



CANVAS/Street View



CANVAS/Street View



CANVAS/Street View



CANVAS/Street View

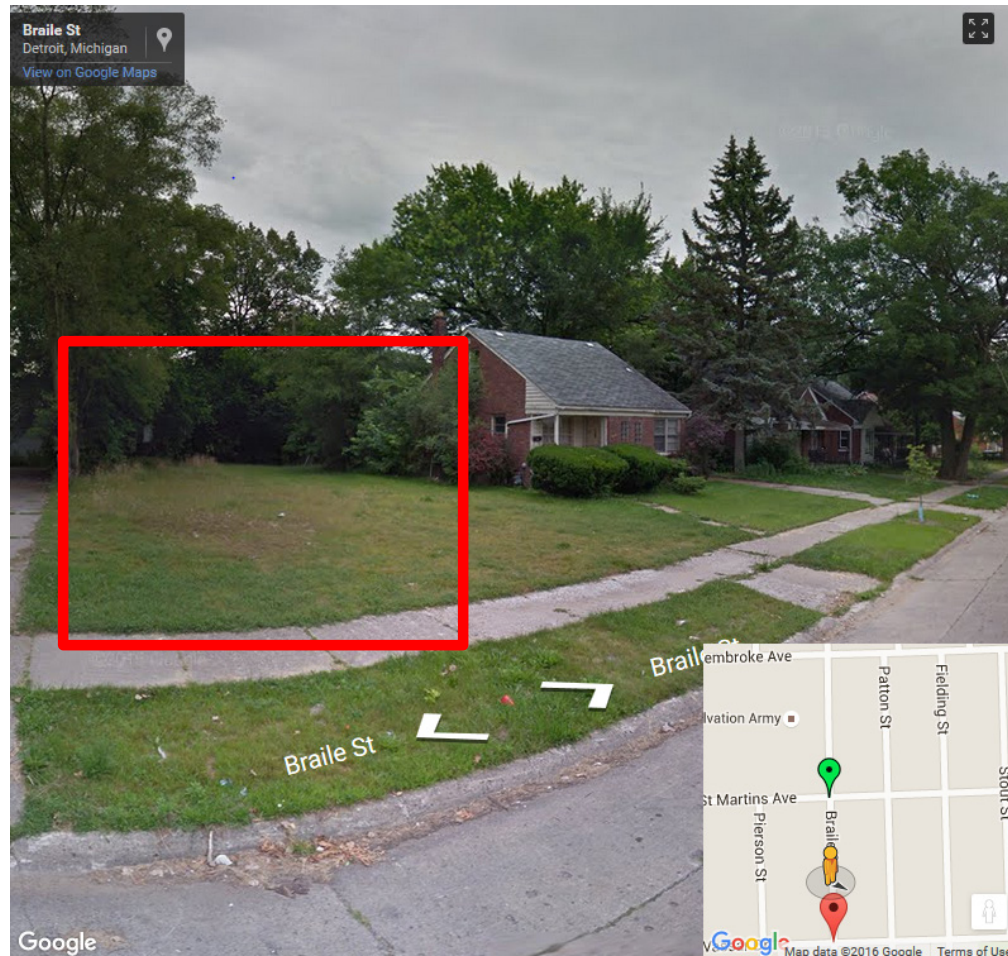


CANVAS/Street View



CANVAS/Street View

Vacant lot



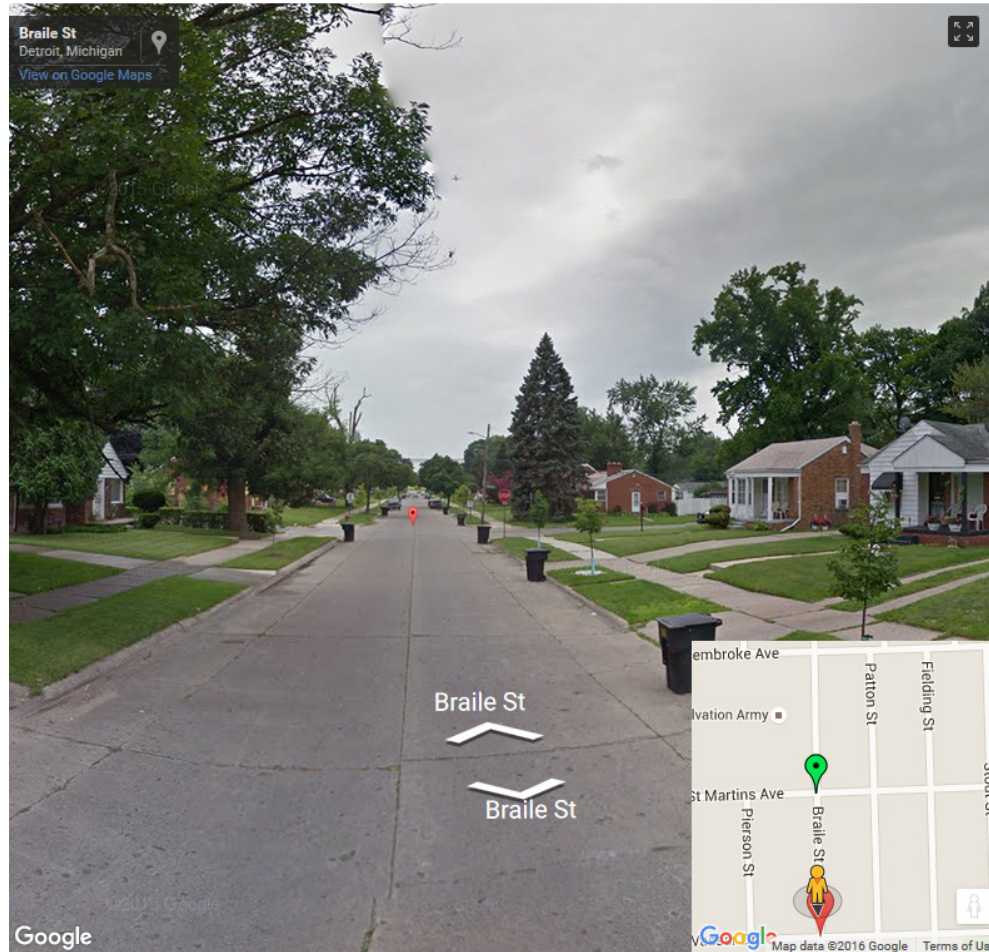
CANVAS/Street View



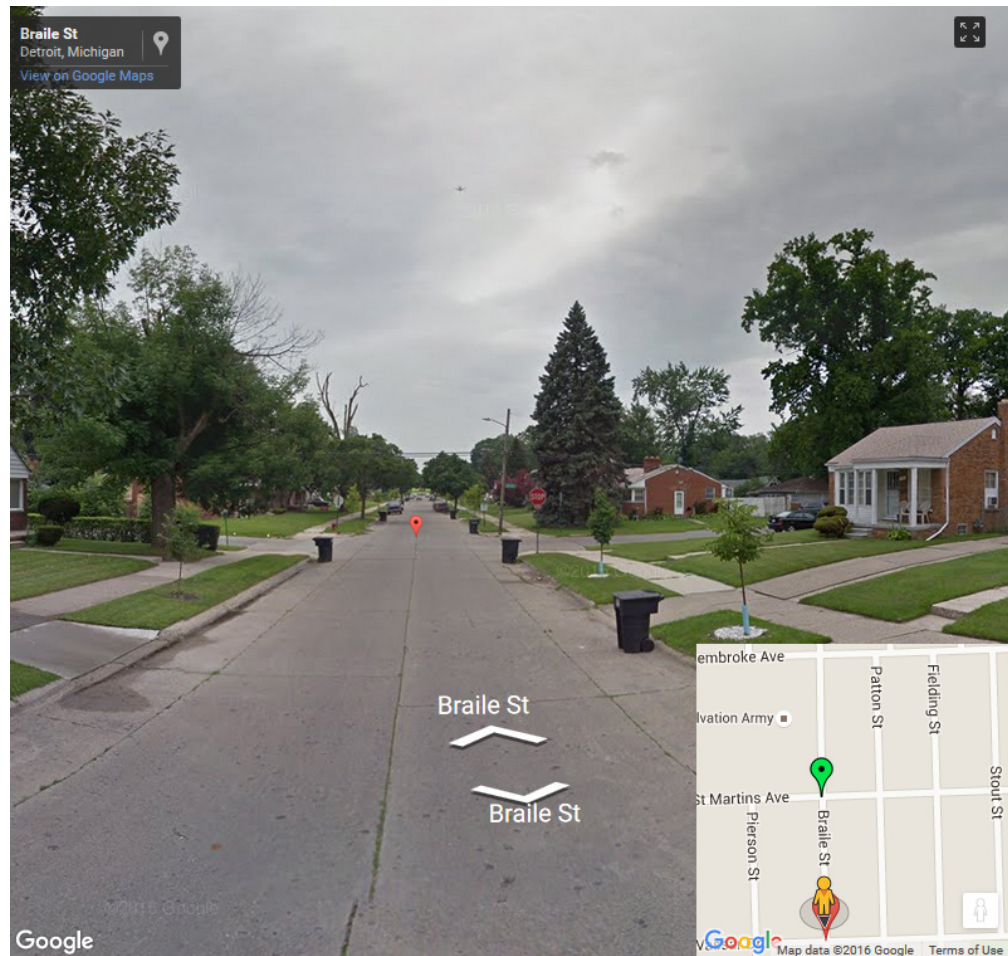
CANVAS/Street View



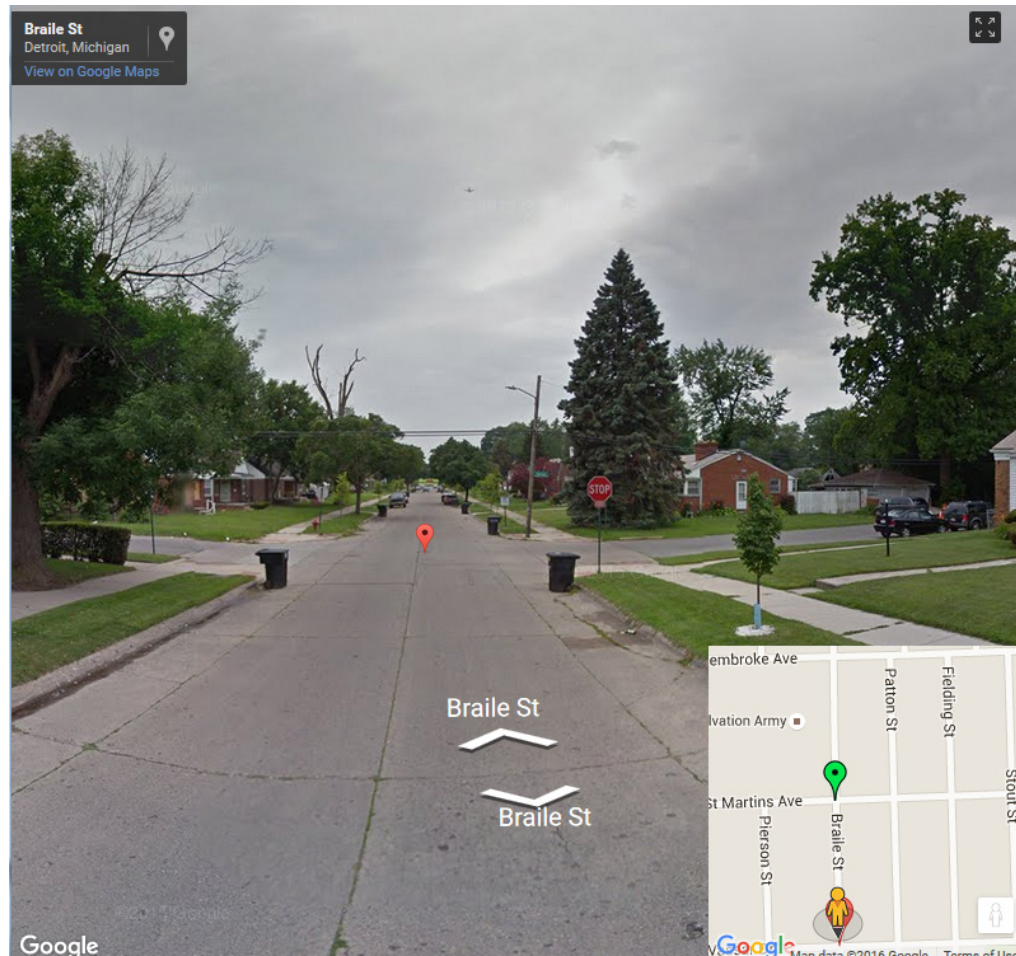
CANVAS/Street View



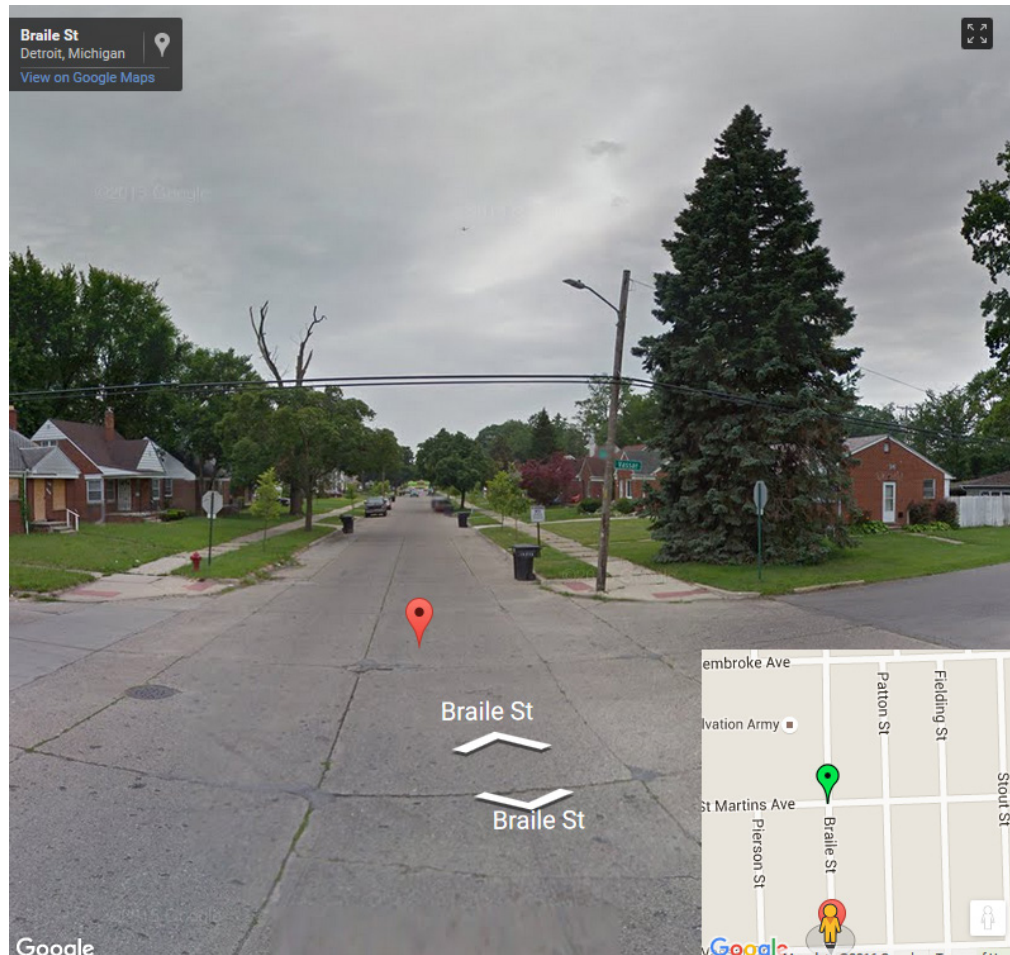
CANVAS/Street View



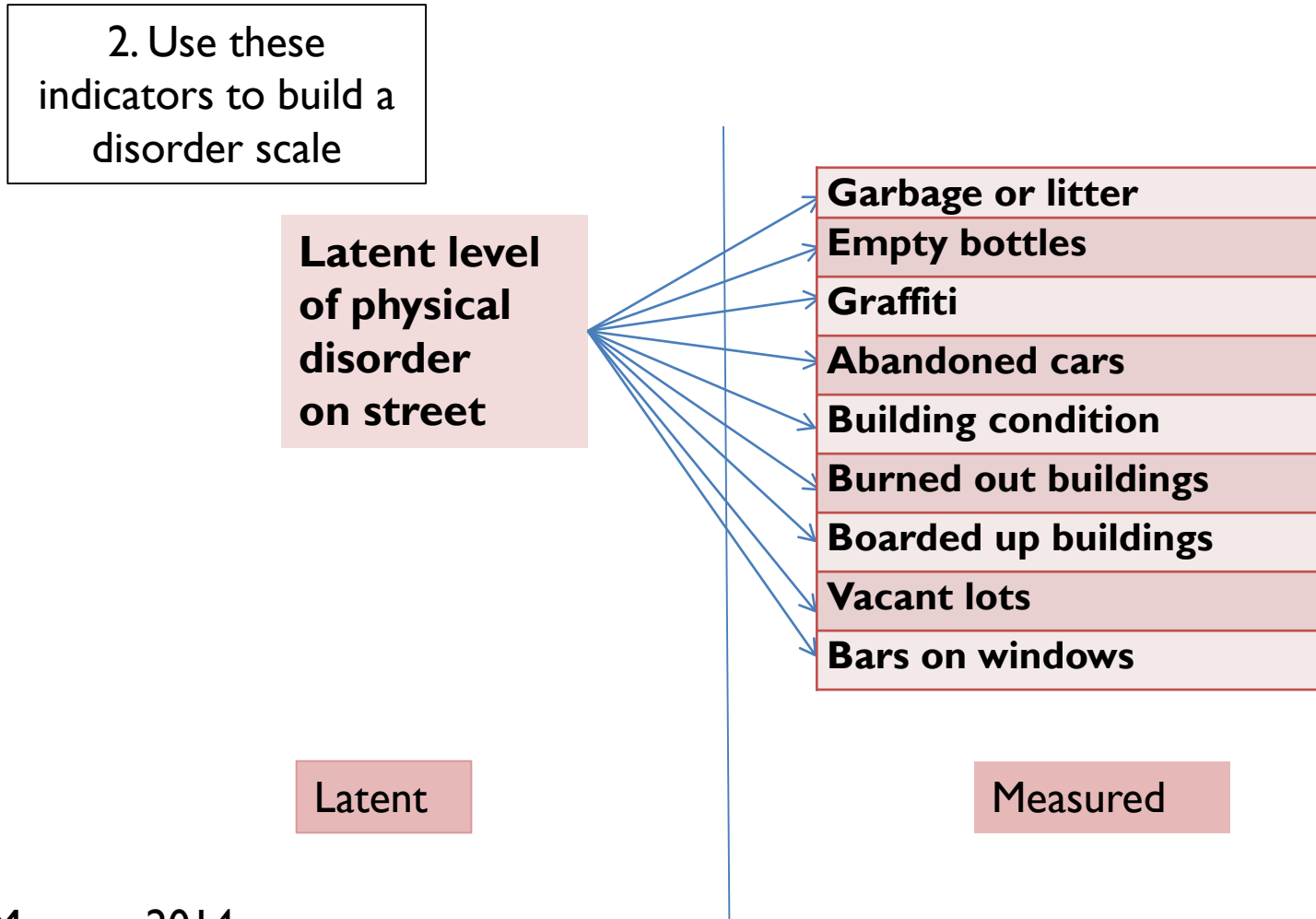
CANVAS/Street View



CANVAS/Street View

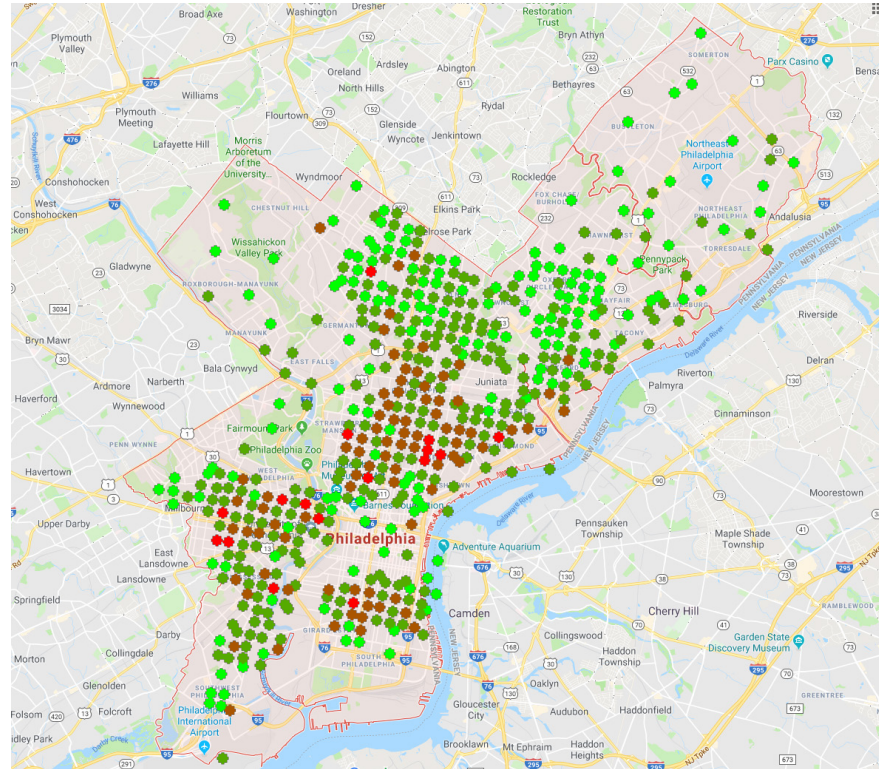


Disorder Scale



Disorder: Raw Data

3. Use Spatial interpolation to estimate what we would have seen in places we did not audit

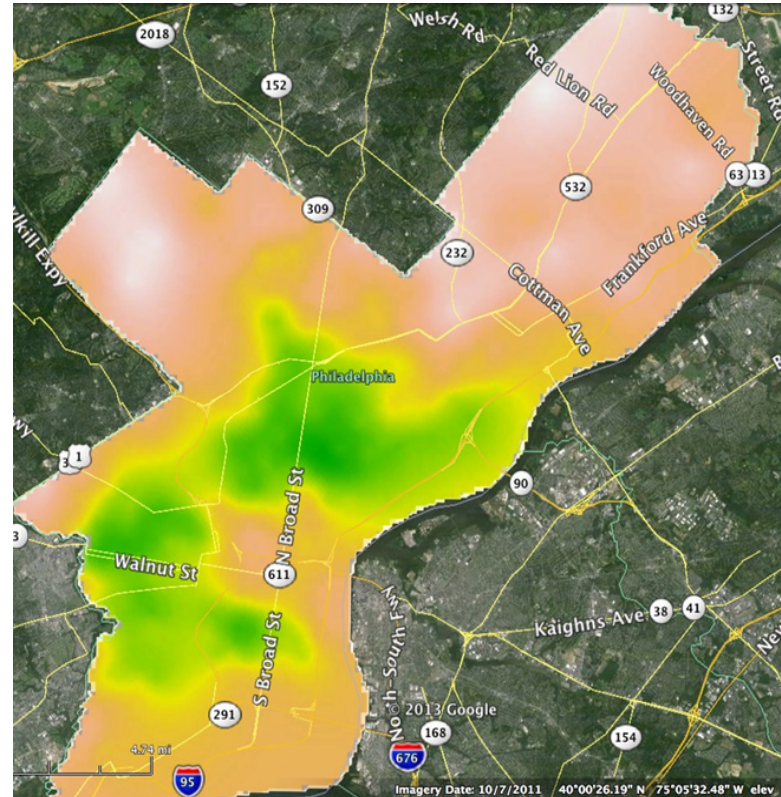


Red = More Disorder

Latent disorder measure at each audited segment (in Philadelphia)

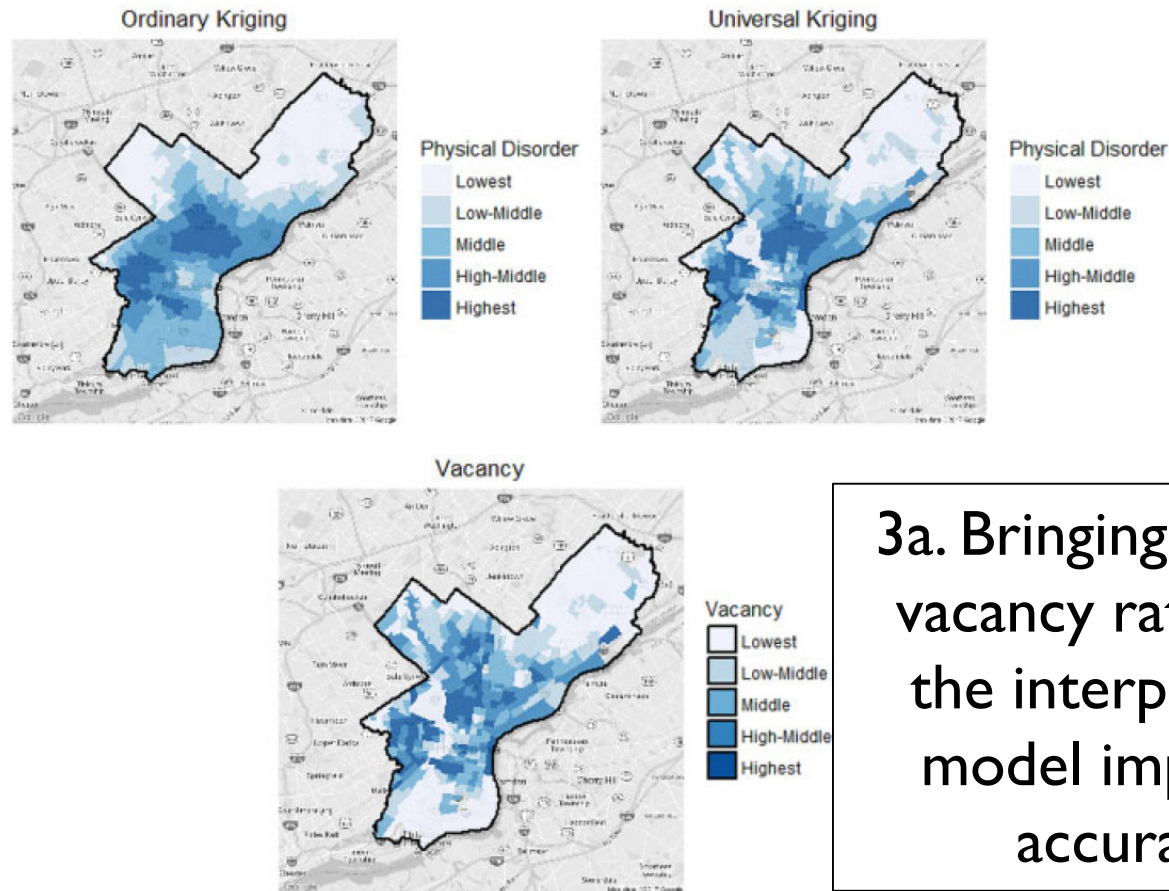
Disorder: Interpolated

3. Use Spatial interpolation to estimate what we would have seen in places we did not audit



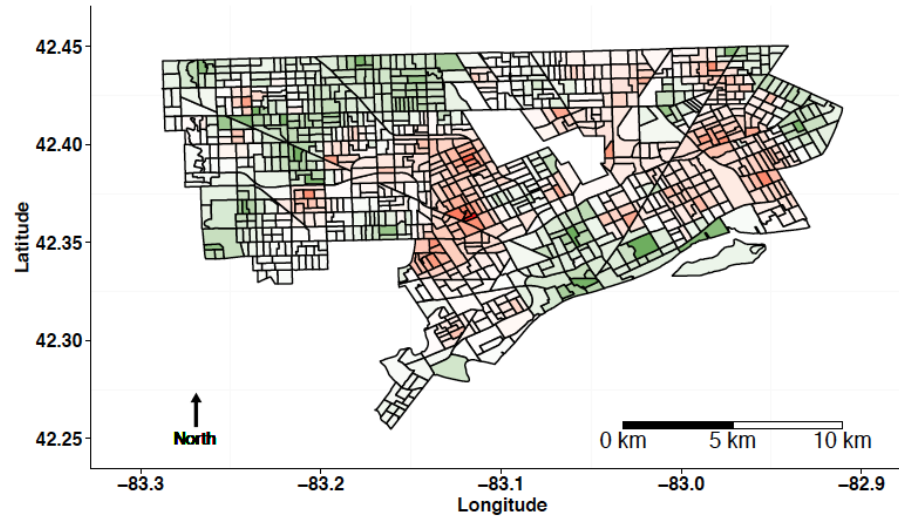
Green = More Disorder

Disorder: with external data

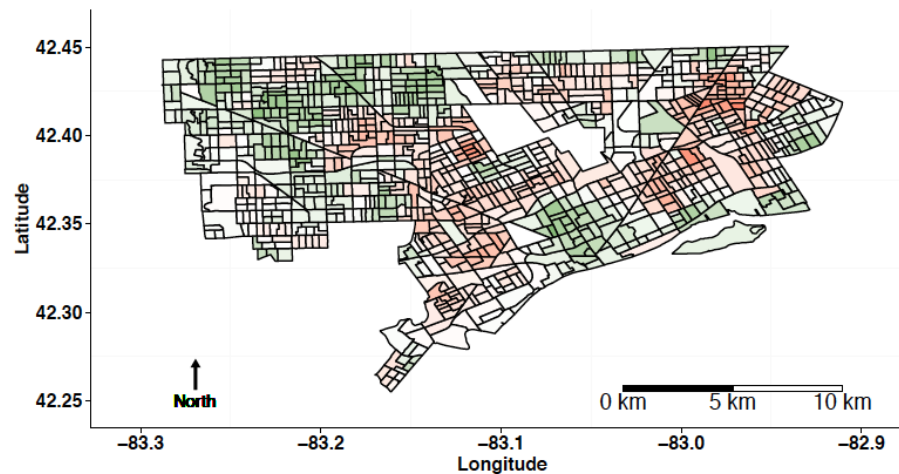


3a. Bringing housing vacancy rates into the interpolation model improves accuracy

Measuring Disorder: Does it work?

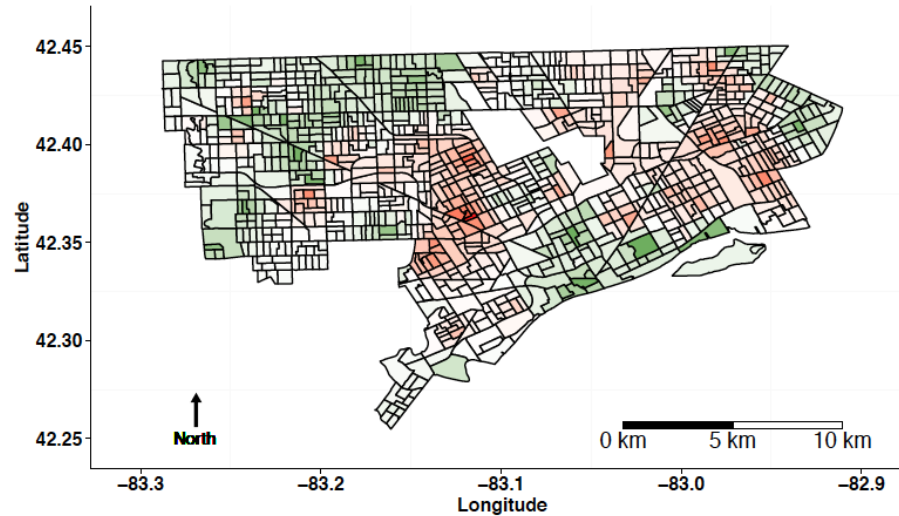


In person audit



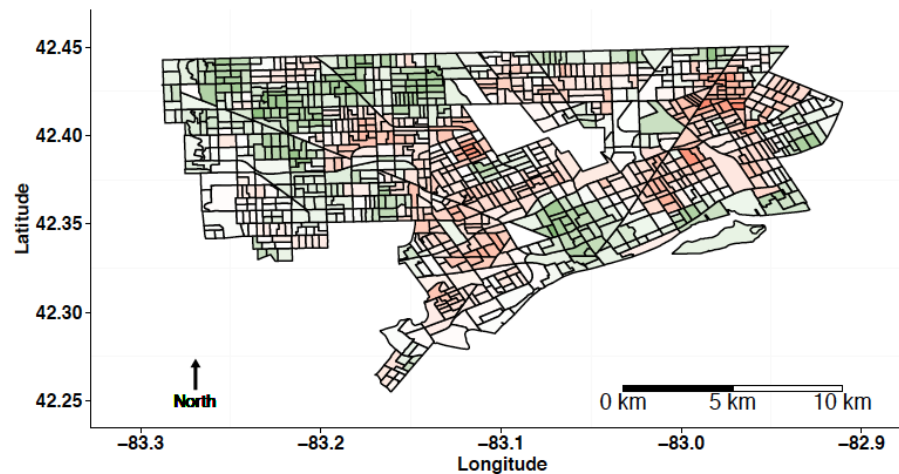
CANVAS audit

Measuring Disorder: Does it work?



In person audit

~3350 person-hours



CANVAS audit

~105 person-hours

Act II:

A Case Study in Working with Multiple
Neighborhood Measures

Measuring neighborhood context

To compute a location-specific measure:

1. Geocode the street address
2. Identify the neighborhood around it
3. Compute the variable

Step 1: Geocoding

- Turn address into latitude/longitude

6100 Executive Boulevard, Rockville, MD

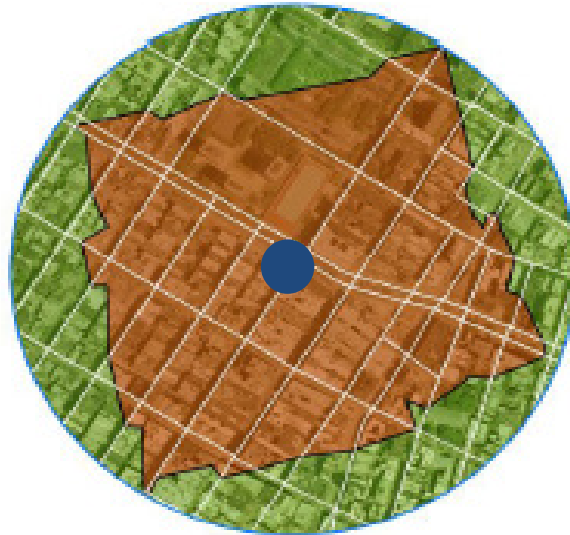


(39.048035,-77.126166)

- There is software to do this

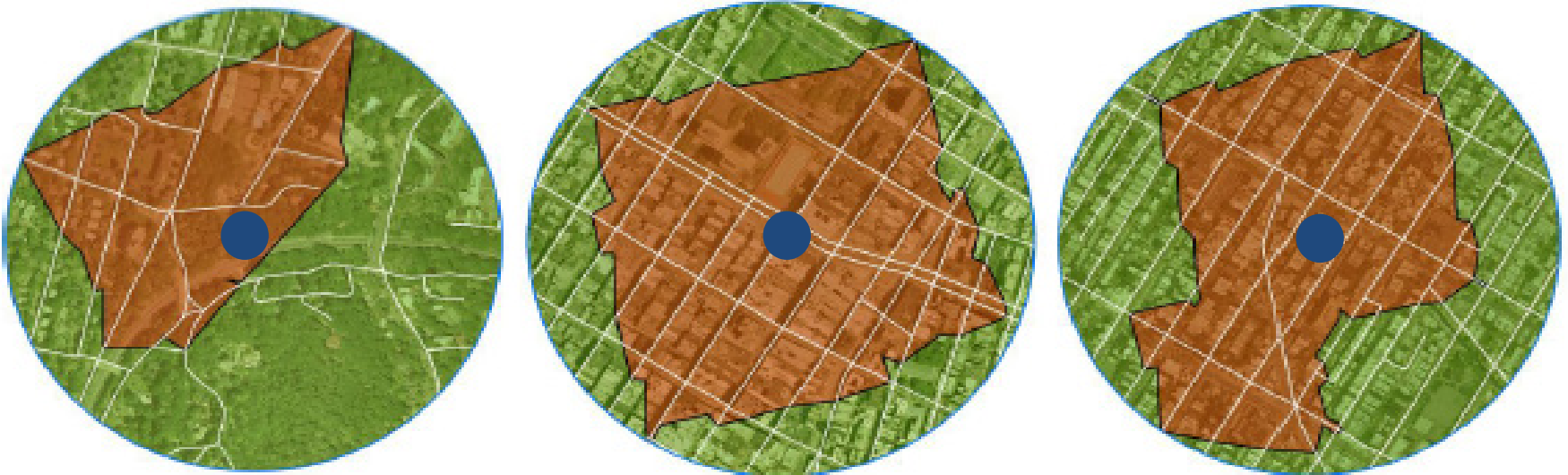
Step 2: Identify the neighborhood

Define an area around the latitude/longitude as the neighborhood



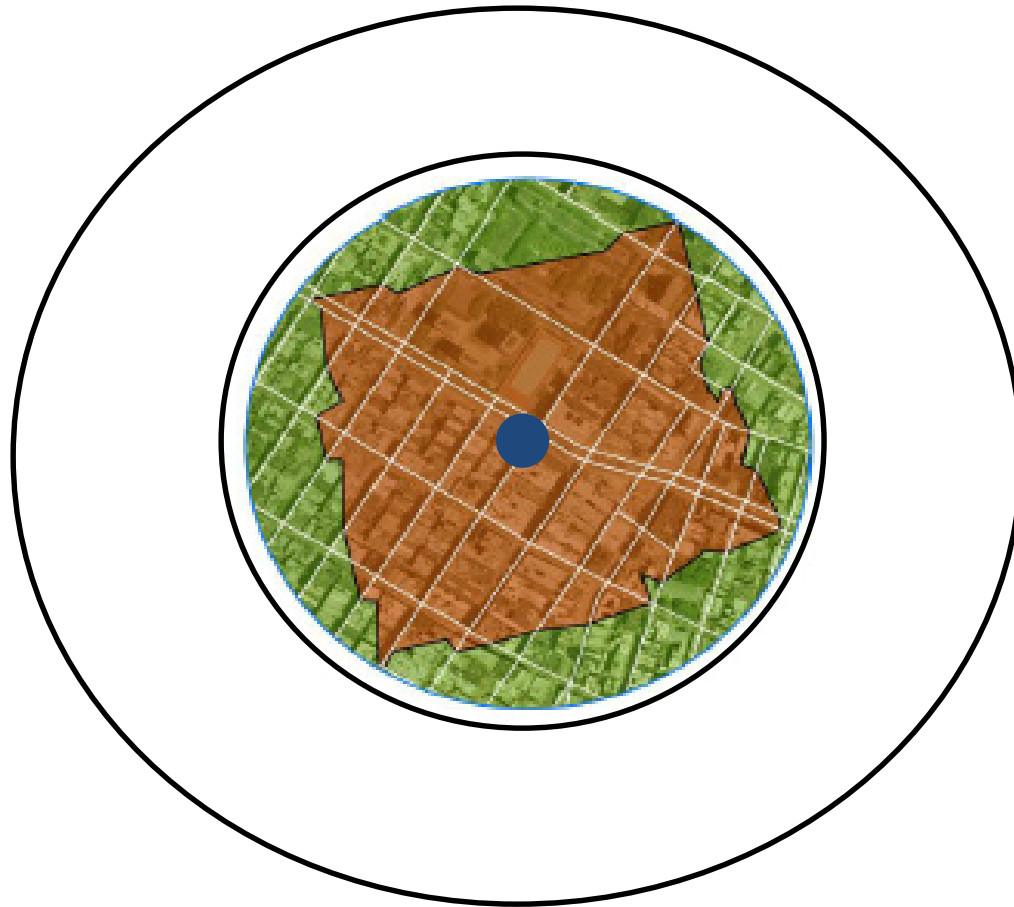
Step 2: Identifying the neighborhood

Define an area around the latitude/longitude as the neighborhood



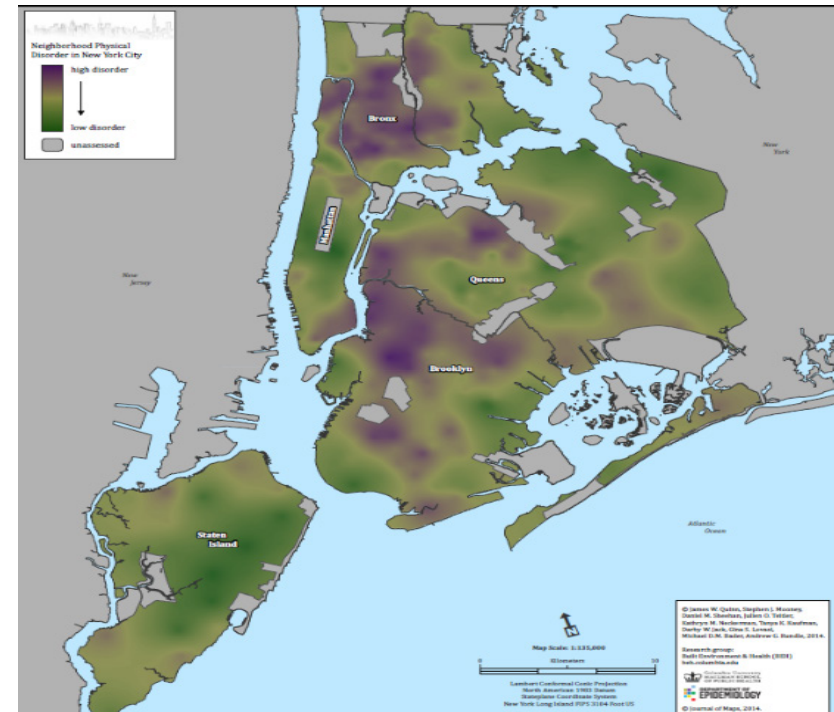
Network buffers (brown) vary in size whereas radial buffers (green circles) do not

Step 2: Identifying the neighborhood



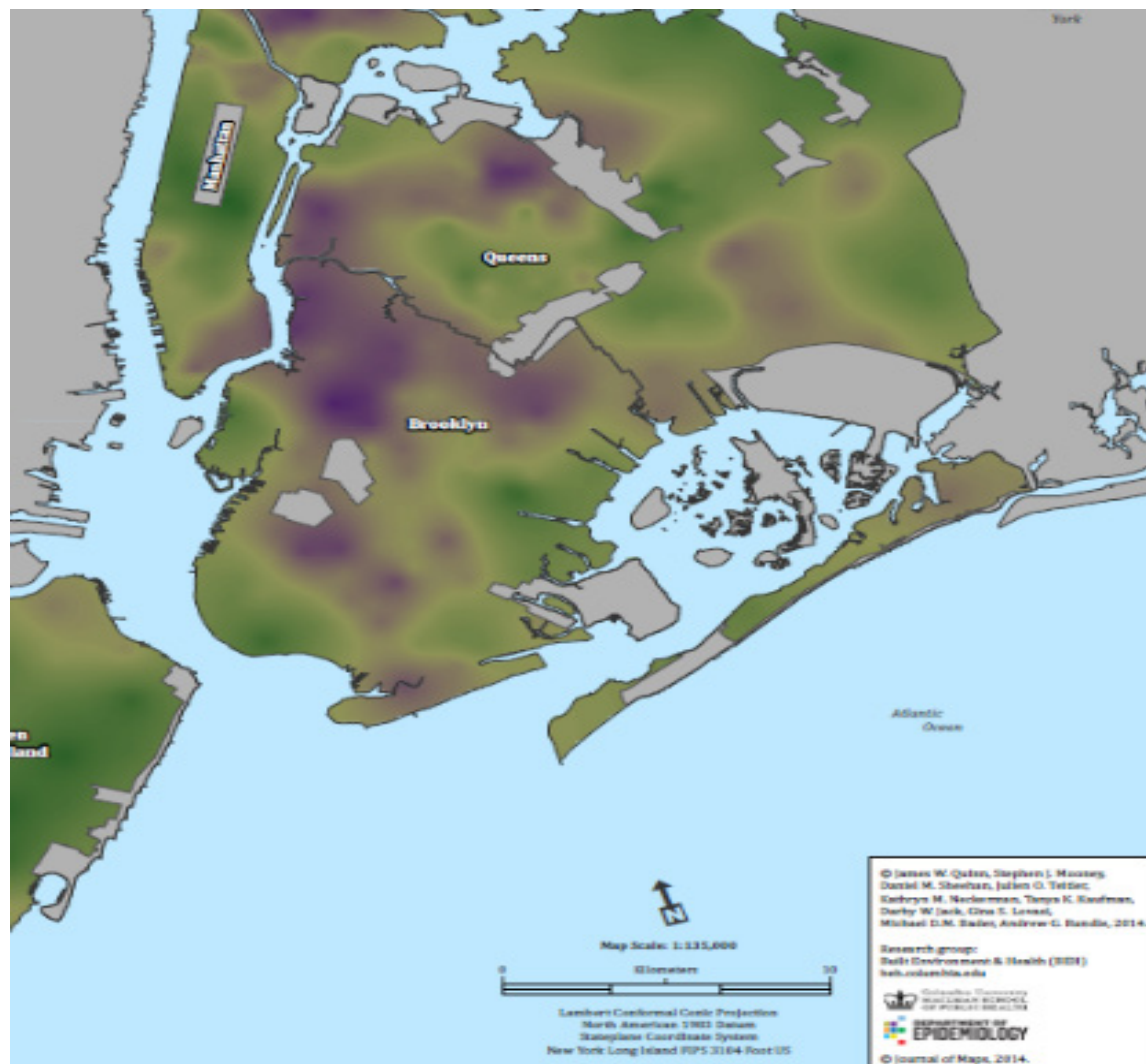
Step 3: Compute

- Use geographic information system (GIS) software to compute mean value for some measure in neighborhood area

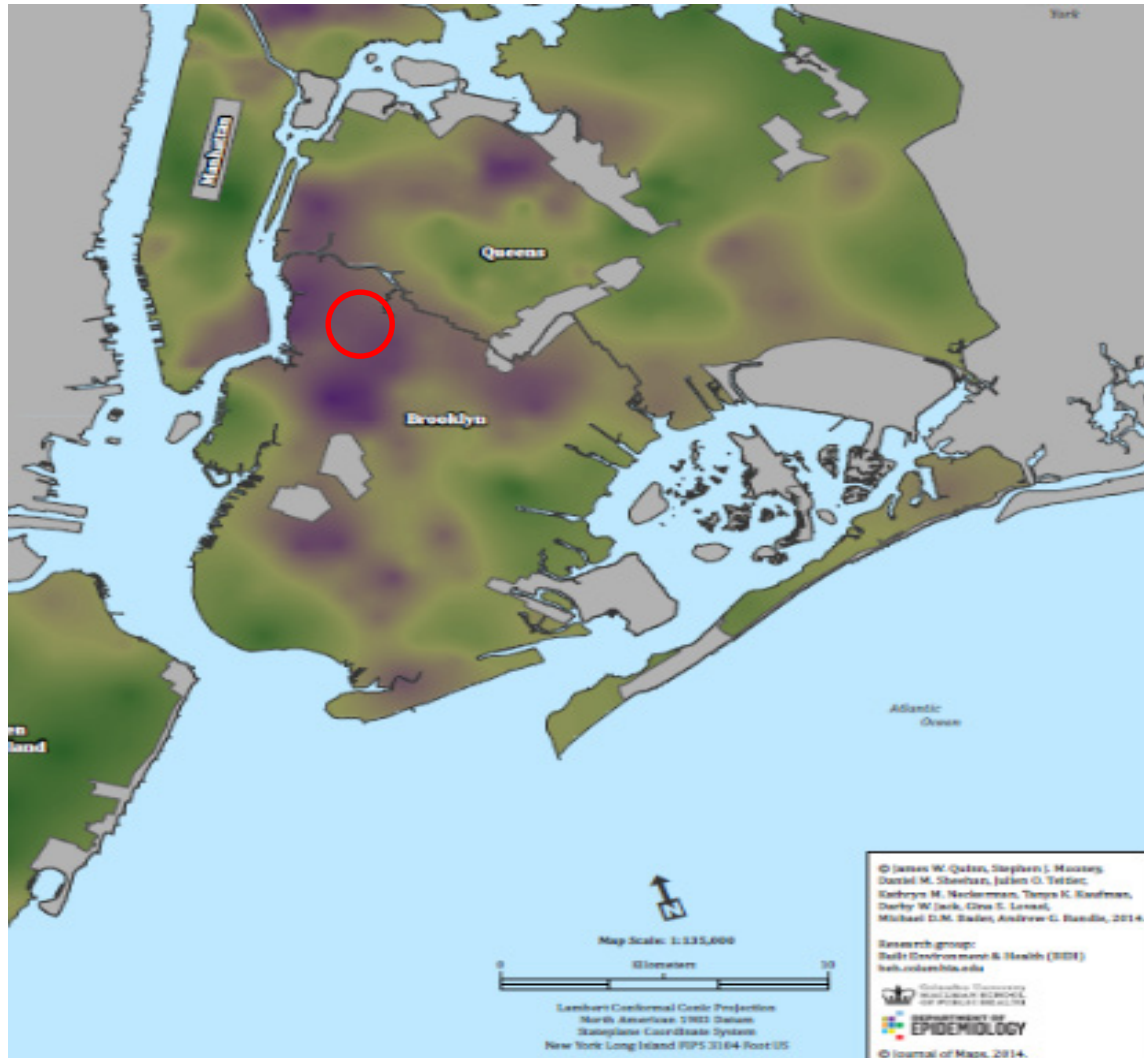


Physical disorder ('broken windows') in New York City

Step 3: Compute



Step 3: Compute



Measuring neighborhood context

So, what should we measure...

Context and Physical Activity

- Many, many contextual factors have been studied in reference to physical activity:
 - Socioeconomics
 - Distance to nearest retail and/or transit
 - Distance to parks
 - Crime/disorder
 - Racial/ethnic composition
 - Immigrant composition
 - Traffic
 - Safety from traffic
 - Tree canopy
 - Lot size/garden space
 - Air pollution
 - Etc.

Context and Physical Activity

- Many, many contextual factors have been studied in reference to physical activity:
 - Socioeconomics
 - Distance to nearest retail and/or transit
 - Distance to parks
 - Crime/disorder
 - Racial/ethnic composition
 - Immigrant composition
 - Traffic
 - Safety from traffic
 - Tree canopy
 - Lot size/garden space
 - Air pollution
 - Etc.

Context and Physical Activity

- Many, many contextual factors have been studied in reference to physical activity:
 - Socioeconomics
 - Median household income
 - Proportion of households living below poverty line
 - Maximum educational attainment
 - Mean years of schooling
 - Property values
 - Age-adjusted median household income
 - Etc.
 - Safety from traffic
 - Tree canopy
 - Lot size/garden space
 - Air pollution
 - Etc.

Introducing the...

Neighborhood Environment-Wide Association Study (NE-WAS)

- Taking explicit inspiration from Genome-Wide Association Studies (GWAS)
- Briefly:
 1. Measure all the things!
 2. Identify the measures most correlated with the outcome

(Reminder: correlation is not causation...but causation causes correlation ... so they're correlated)

Pilot NE-WAS: Cohort

- 3,497 adults aged 65-75 living in NYC in 2011
 - Sample weights available to generalize to older adult population of the city
 - Physical activity using the Physical Activity Scale for the Elderly (PASE)
 - Self-reported home addresses
 - 3 waves of data, but looking at baseline only here

Pilot NE-WAS: Outcome measure

- Physical Activity Scale for the Elderly (PASE)
 - Twelve item self-report scale resulting in score representing past-week energy expenditure
 - Widely used in research with older adults
 - Previously validated ($r=0.68$ against gold standard)

Pilot NE-WAS: Analytic outcomes

- Analyzed four outcomes from the PASE:
 - PASE score (total activity)

Pilot NE-WAS: Analytic outcomes

- Analyzed four outcomes from the PASE:
 - PASE score (total activity)
 - Gardening (dichotomous)

Gardening in NYC? Really?

- NYC has five boroughs
- Only one of them is Manhattan



Queens: lawn, driveway, no sidewalk

Pilot NE-WAS: Analytic outcomes

- Analyzed four outcomes from the PASE:
 - PASE score (total activity)
 - Gardening (dichotomous)
 - Walking (dichotomous)

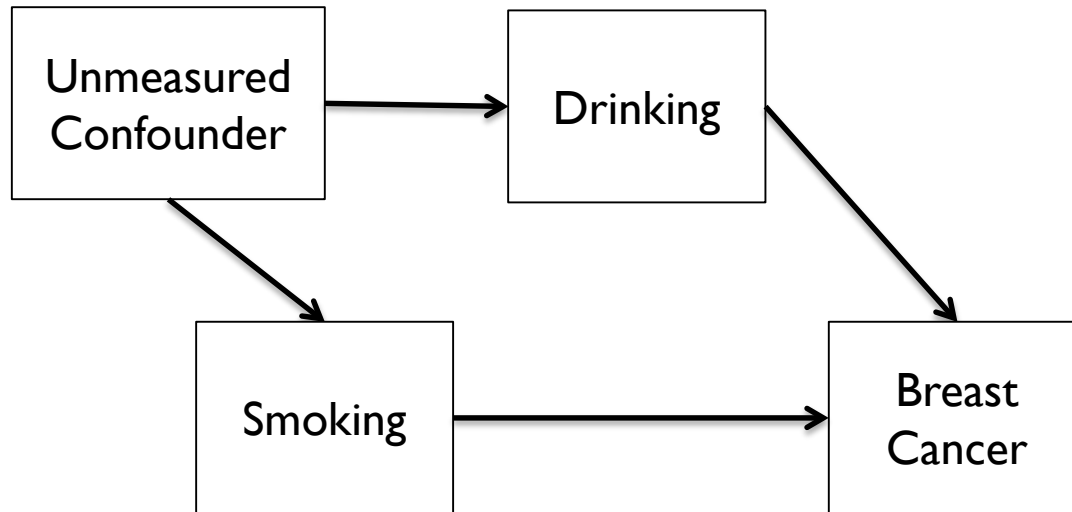
Pilot NE-WAS: Analytic outcomes

- Analyzed four outcomes from the PASE:
 - PASE score (total activity)
 - Gardening (dichotomous)
 - Walking (dichotomous)
 - Heavy Housework (dichotomous, negative control)

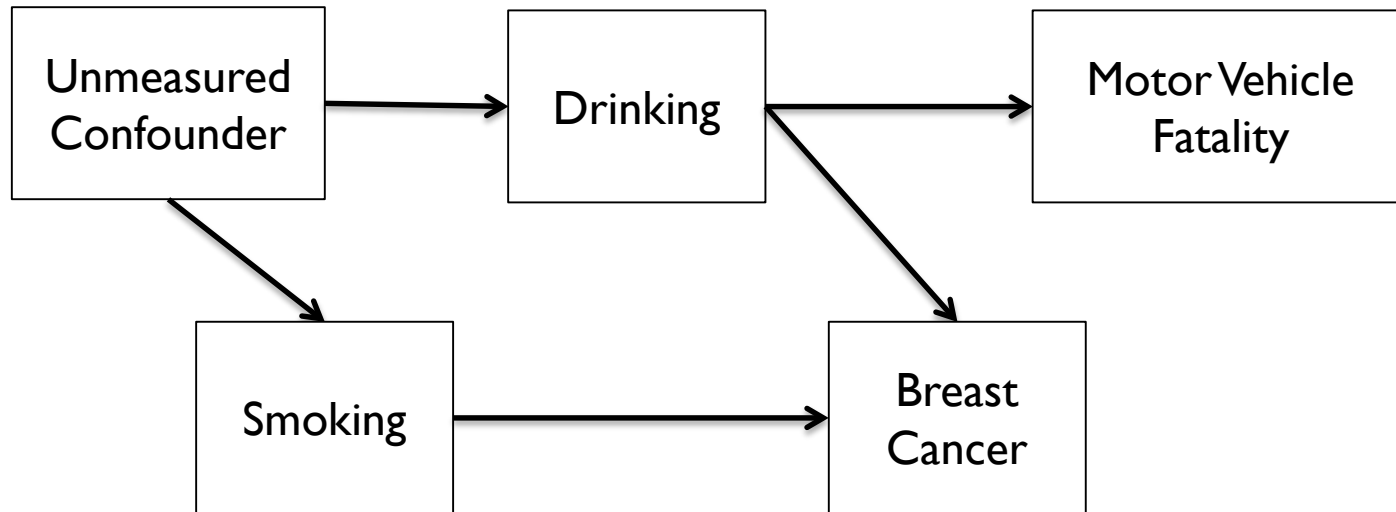
Negative controls



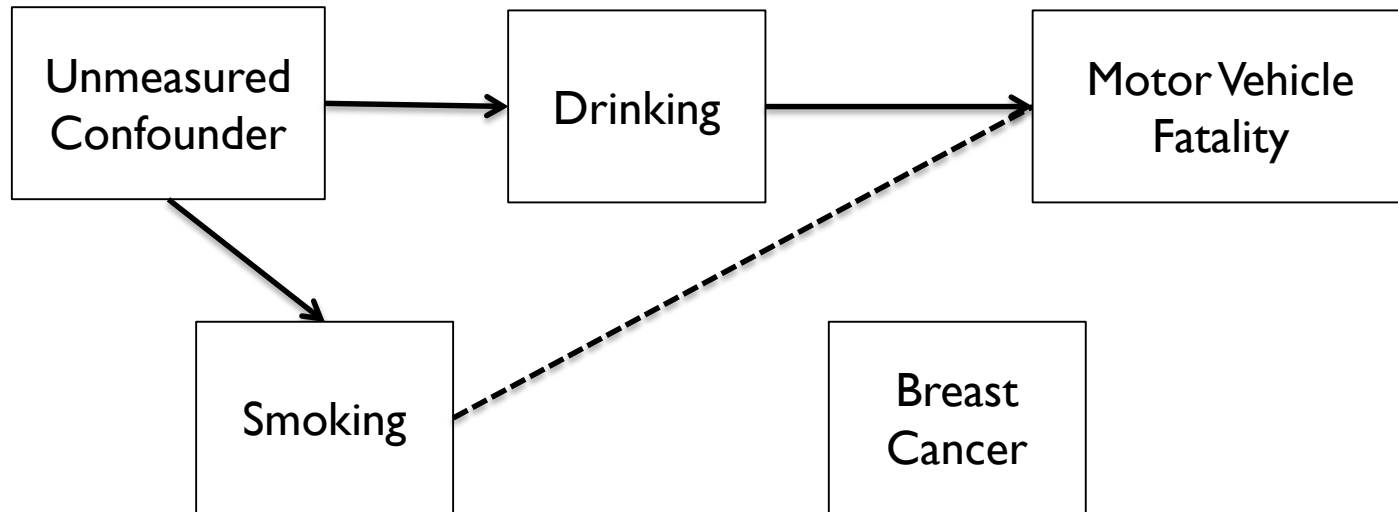
Negative controls



Negative controls



Negative controls



Should see no association between smoking and motor vehicle fatality
after controlling for drinking!

Pilot NE-WAS: Neighborhood measures

Category	Number of Measures	Examples
Demographics/Housing	121	Population density, % of total population aged 10-14
Education/Employment/Income	102	% college grad, % in labor force
Urban Form	50	% walk to work, bus stop density
Crime/Disorder	35	Mean neighborhood disorder, % of streets rated as filthy
Parks	5	% of land area dedicated to large parks
Pedestrian Safety	24	Pedestrian injury count from 2000- 2009
Total	337	

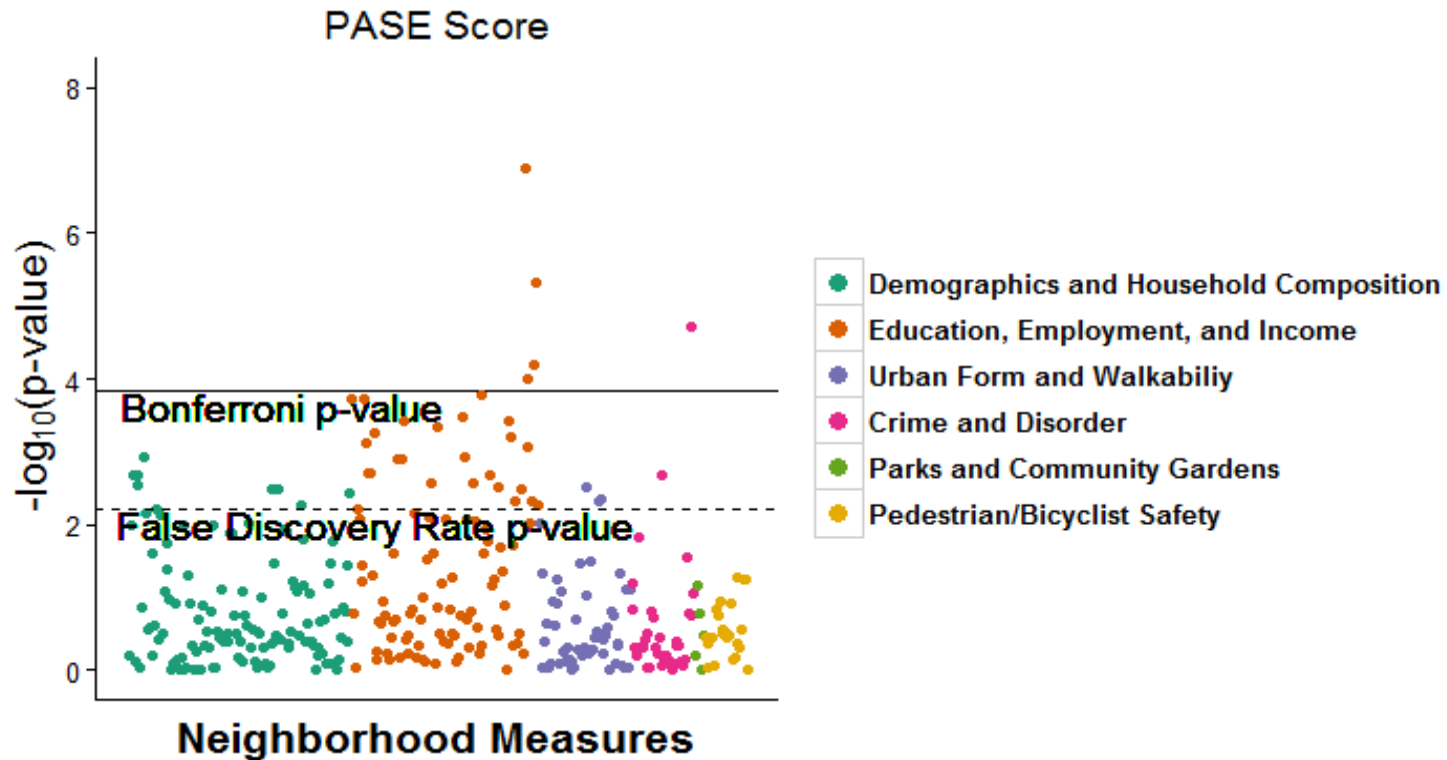
Pilot NE-WAS: 2 Analyses

1. Initial analysis (following GWAS paradigm): bivariate regression for each measure independently controlling for:
 - Age
 - Race/ethnicity
 - Educational attainment
 - Income
 - Housing type (apartment/low-rise apartment/high-rise apartment)
2. (LASSO regression to select final neighborhood measures)

Results

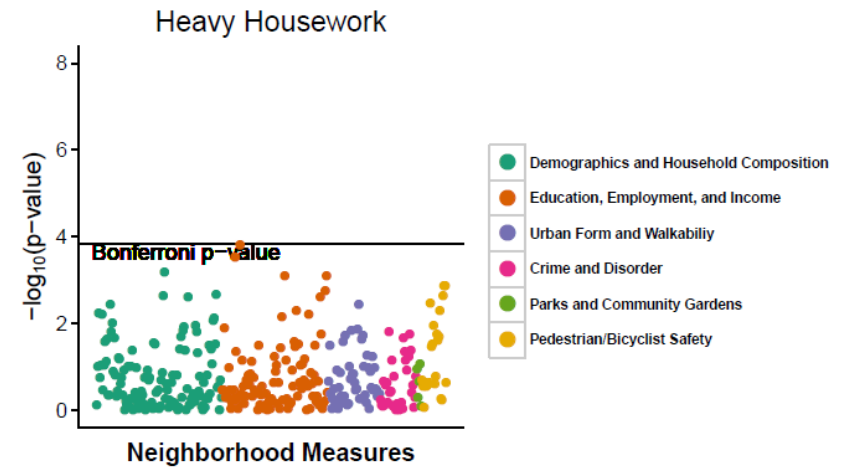
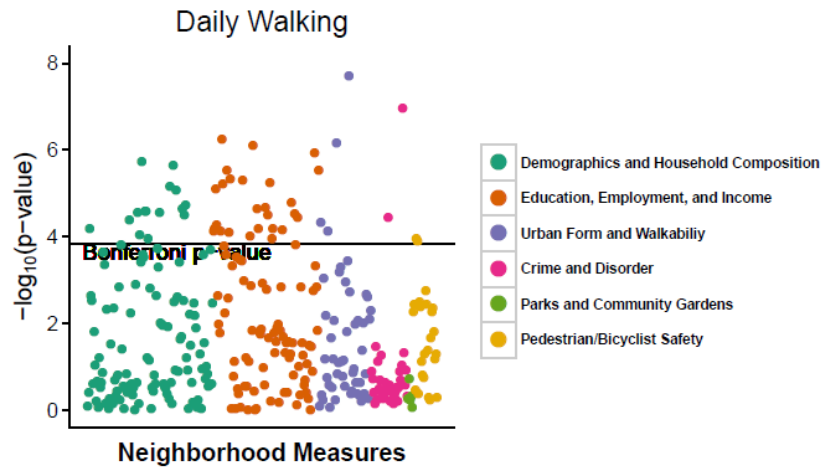
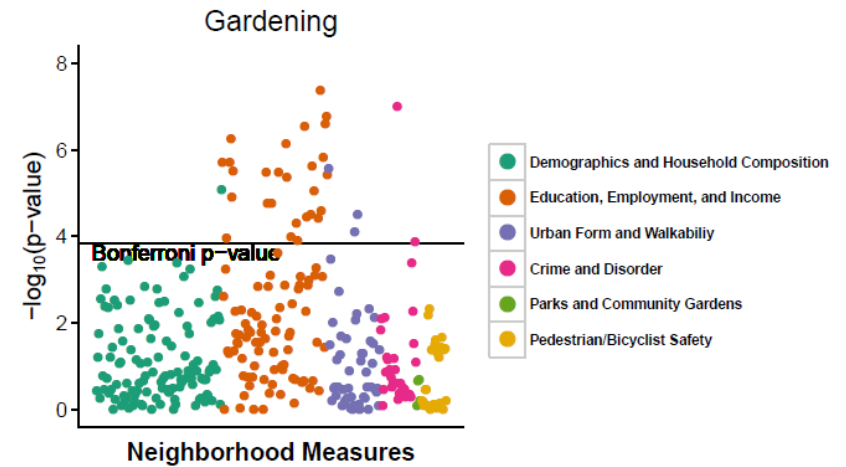
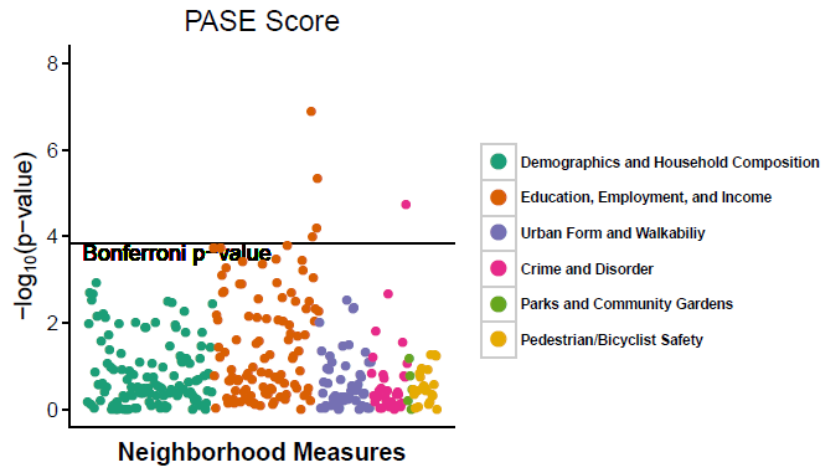
Did we learn anything?

Bivariate analysis



Most neighborhood variables were not significantly associated with total activity after Bonferroni correction

Bivariate analysis



Bivariate analysis: Top measures

	PASE Score	Gardening	Walking Daily	Heavy Housework
Count of significantly associated measures	5 (1.5%)	33 (9.8%)	49 (14.4%)	0 (0.0%)
Top 5 statistically significant measures (by p-value of coefficient)	Incomes below half the poverty level (-)	Incomes below half the poverty level (-)	60-90 minute travel time to work (-)	--
	Incomes below the poverty line (-)	Physical disorder (-)	Broken windows (-)	--
	Well-maintained windows (+)	Incomes below the poverty line (-)	Some college or more (+)	--
	Incomes above twice the poverty level (+)	Incomes above twice the poverty level (+)	Commuting by car, truck, or van (-)	--
	Incomes between half and three-quarters of the poverty level (-)	Any interest, dividend, or rental income (+)	Workers in professional or management industries (+)	--

Substantive Conclusions

1. Socio-economic status related to all activity
2. Urban form more strongly related to walking

Methodological Conclusions

1. Results were more interpretable in measure domains than as specific measures
 - Dimensionality reduction prior to analysis (e.g. Principal Component Analysis) probably makes sense
2. Naïve machine learning was uninterpretable
 - BUT: blame the craftsman, not the tool!

Curtain Call: Conclusions and Next Steps

Informatics in neighborhood measurement

Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity

Adyasha Maharana, MS¹; Elaine Okanyene Nsoesie, PhD²

Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States



Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei

A Neighborhood-Wide Association Study (Nवास): Example of prostate cancer aggressiveness

Shannon M. Lynch , Nandita Mitra, Michelle Ross, Craig Newcomb, Karl Dailey, Tara Jackson, Charnita M. Zeigler-Johnson, Harold Riethman, Charles C. Branas, Timothy R. Rebbeck

Maharana, 2018

Gebru, 2017

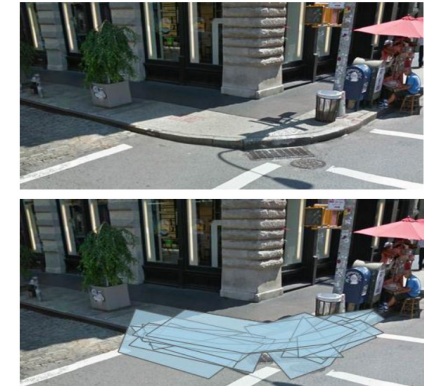
Lynch, 2017

Informatics in place-based health research



Figure 4.

Detecting near-misses in
real-time video



(b) No Curb Ramp

Identifying accessibility issues
from Google Street View

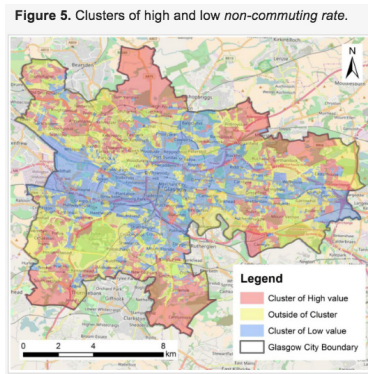


Figure 5. Clusters of high and low non-commuting rate.

Hara, 2014
Loewenherz, 2017
Sun, 2017

Assessing air pollution
exposure from Strava data

Algorithms vs. People

That means **you need to understand them enough to know where they diverge from common sense**



Bottom Line

Programming can dramatically increase our ability to measure aspects of environment and analyze those environmental measures...

Bottom Line

Programming can dramatically increase our ability to measure aspects of environment and analyze those environmental measures...

...but we still need to be smart about what those analyses mean!

Questions/Discussion

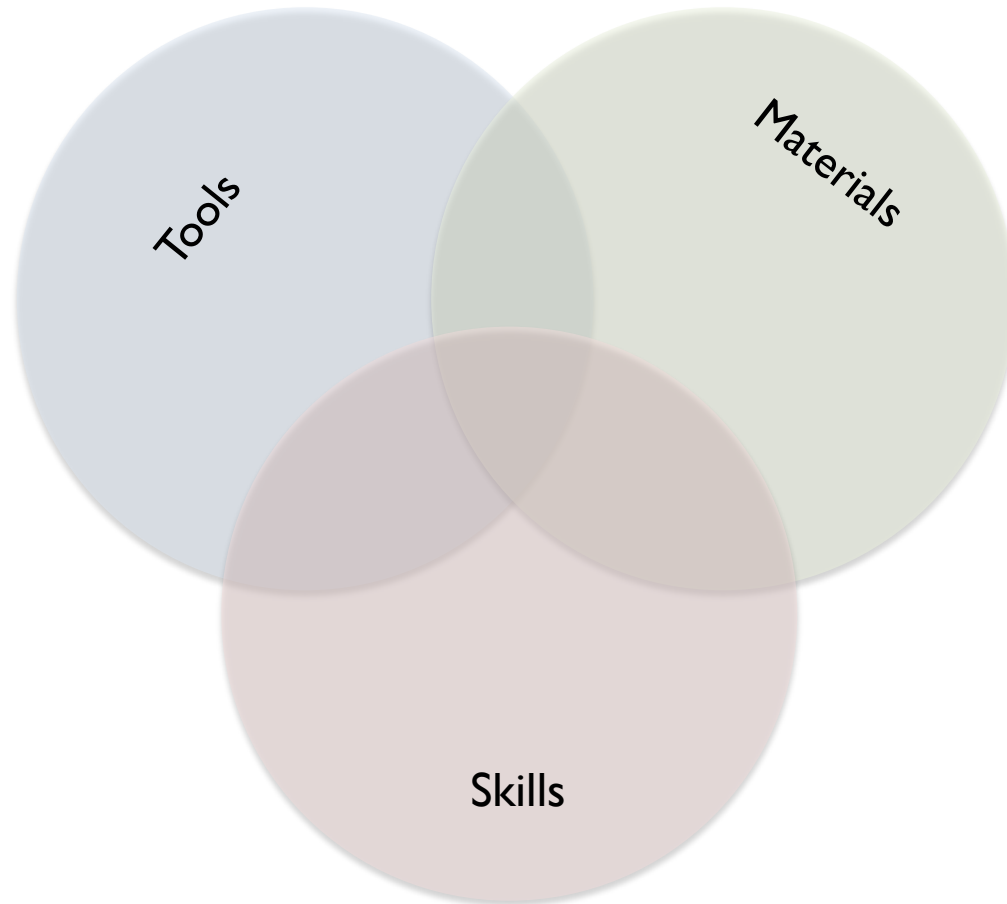
sjm2186@uw.edu

HARBORVIEW
INJURY PREVENTION
& RESEARCH CENTER

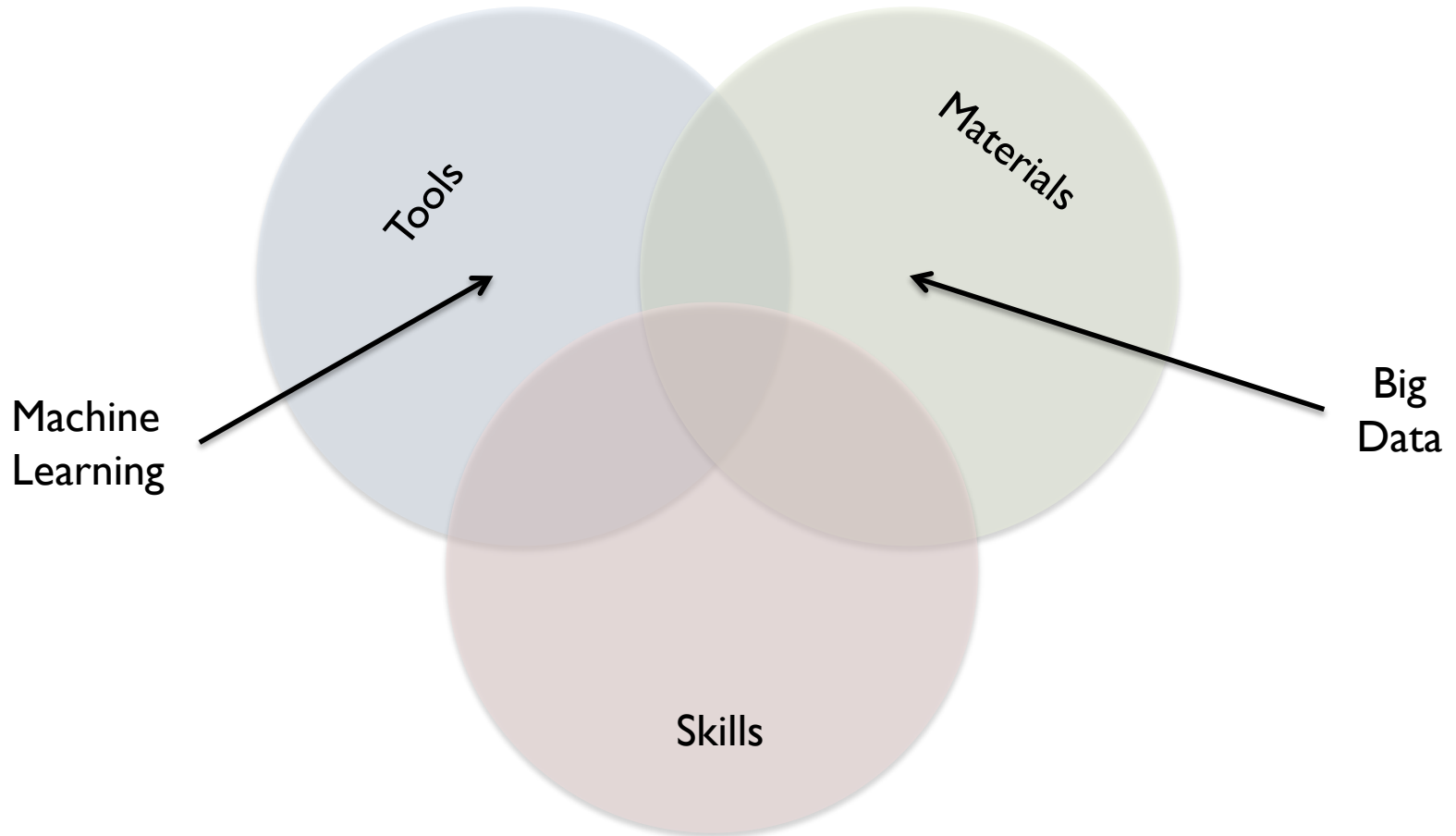


DEPARTMENT OF EPIDEMIOLOGY
UNIVERSITY of WASHINGTON
School of Public Health

Components of a Project



Components of a Project



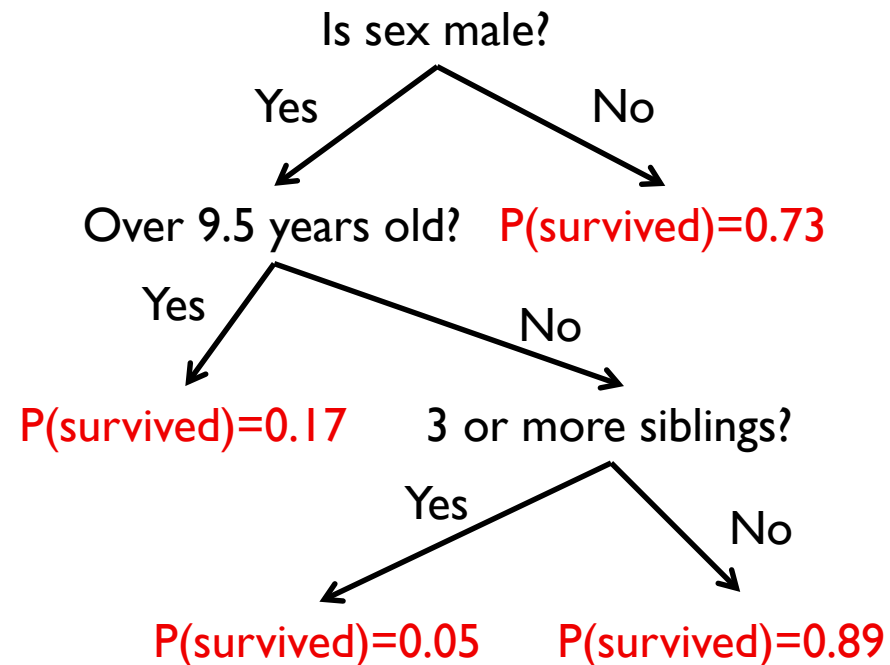
Types of Machine Learning

Source	Examples
Supervised	Logistic Regression
Semi-supervised	(Supervised, but allowing for missing outcome data)
Unsupervised	K-means clustering
Reinforcement	Game playing

Machine learning: Random forest

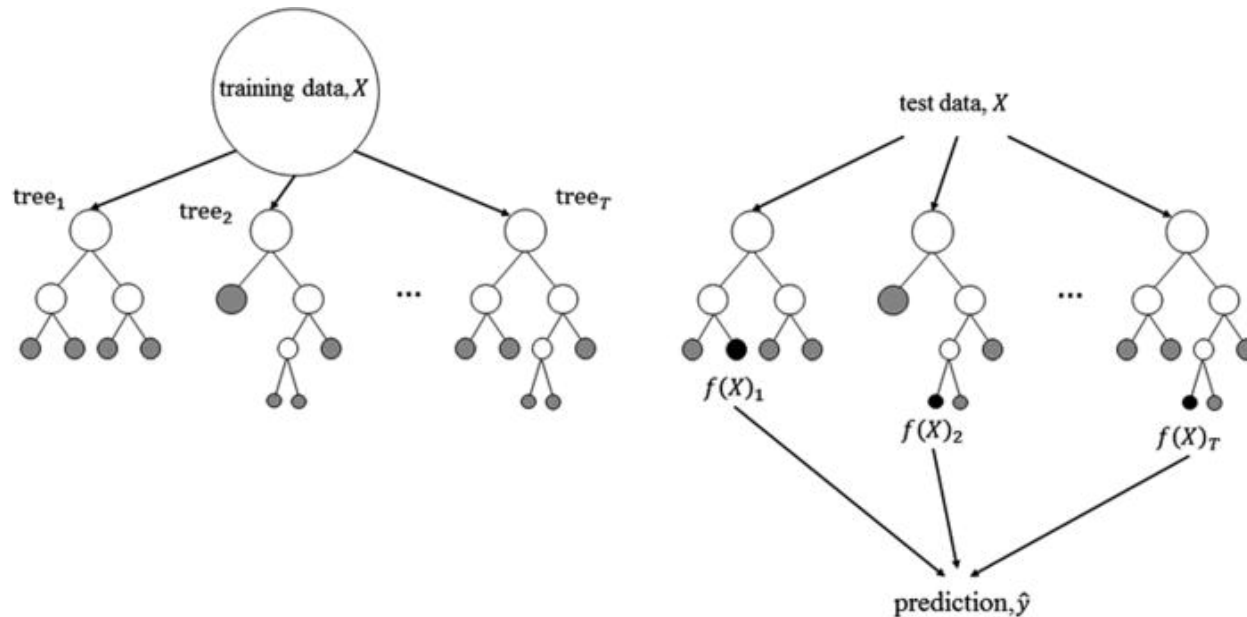
- Random forest is a data-driven statistical model based on 'forests' of decision or regression trees
- Decision tree:
 - Select a cut-point in a given variable that best classifies outcome.
 - Then repeat with another variable, hierarchically.
 - Select tree that best predicts overall

Decision tree predicting children's survival after Titanic started to sink



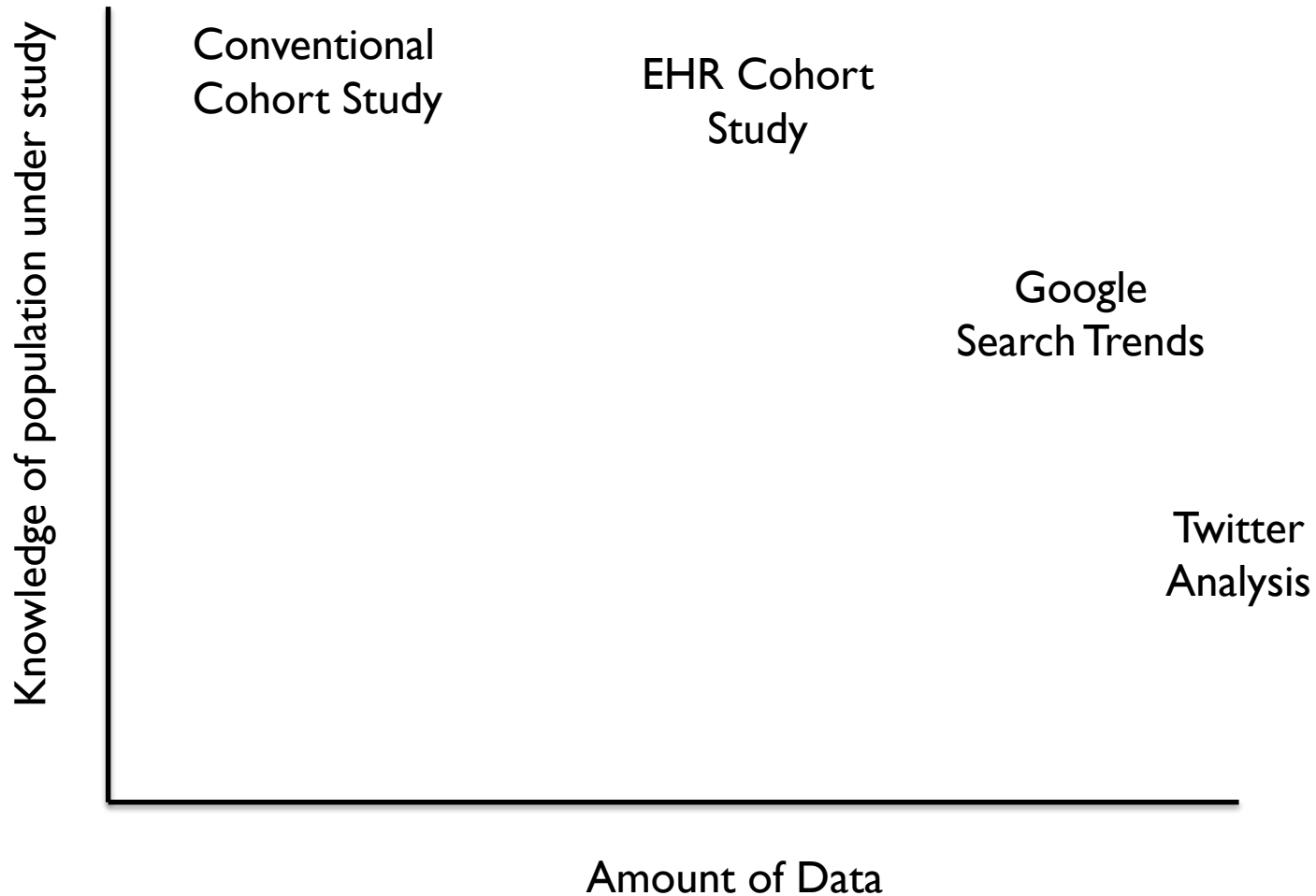
Adapted from Wikipedia

Machine learning: Random forest

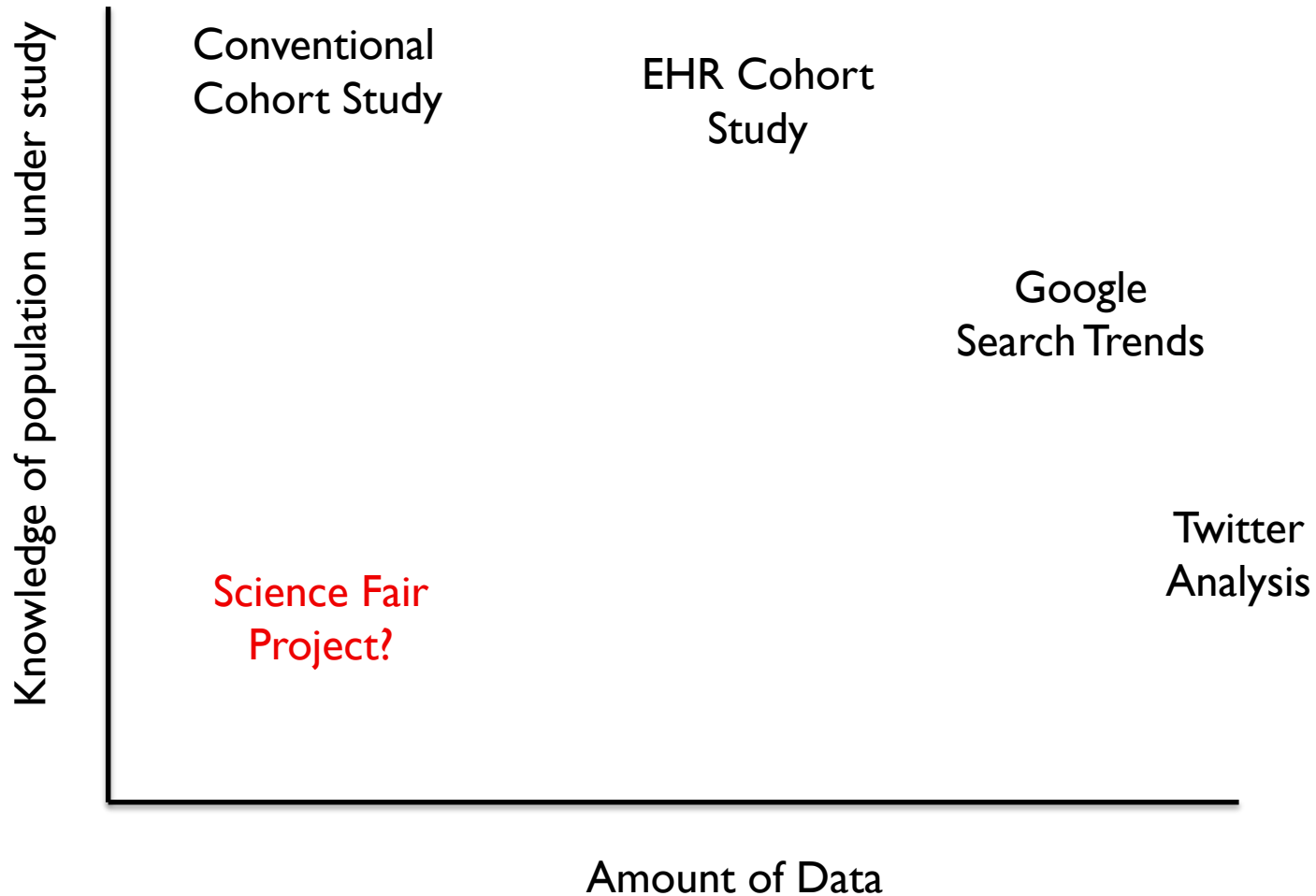


Key idea: Randomly subset dataset to fit tons of regression trees, then combine each tree's prediction into a final predictive model

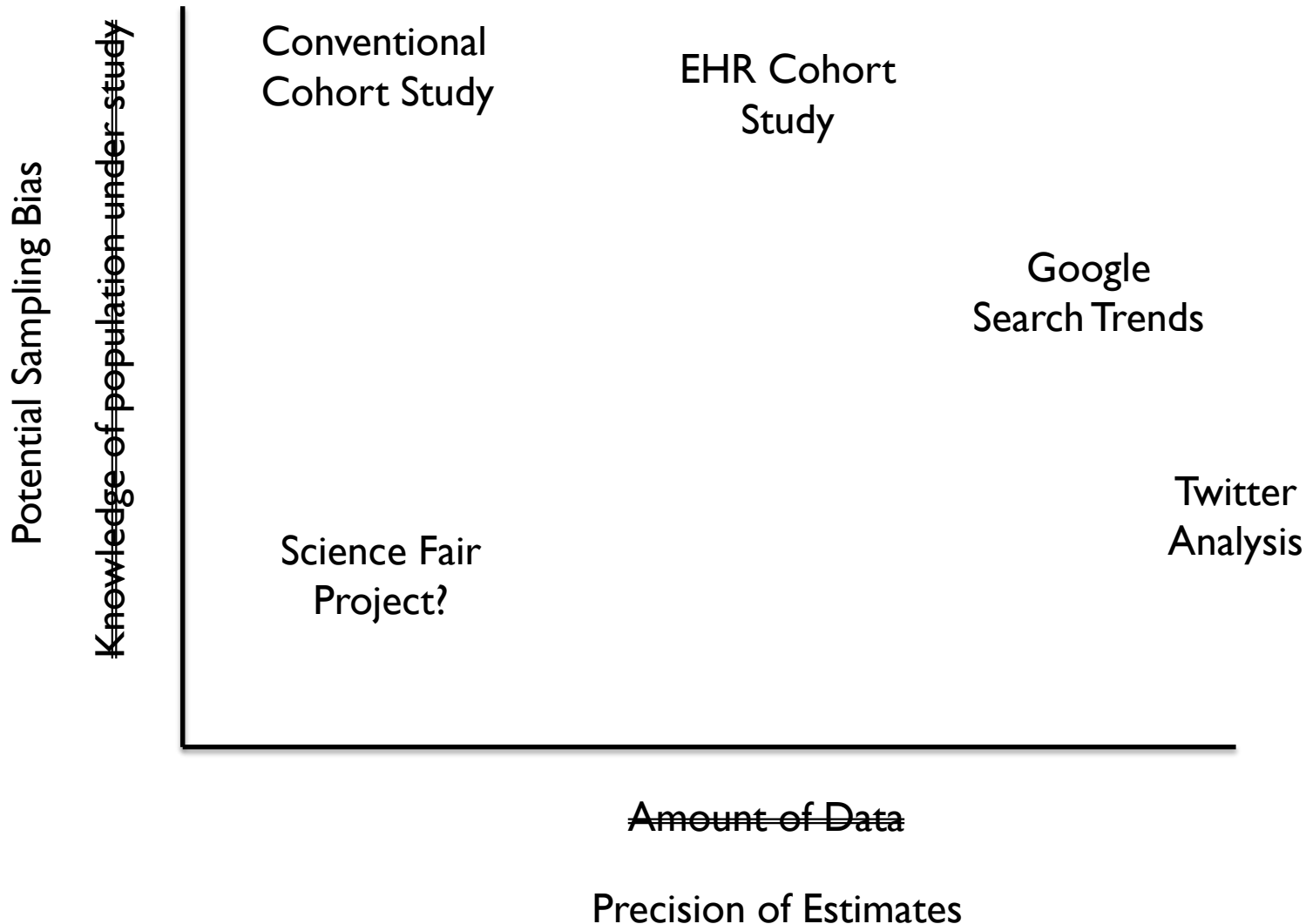
Uses of Big Data



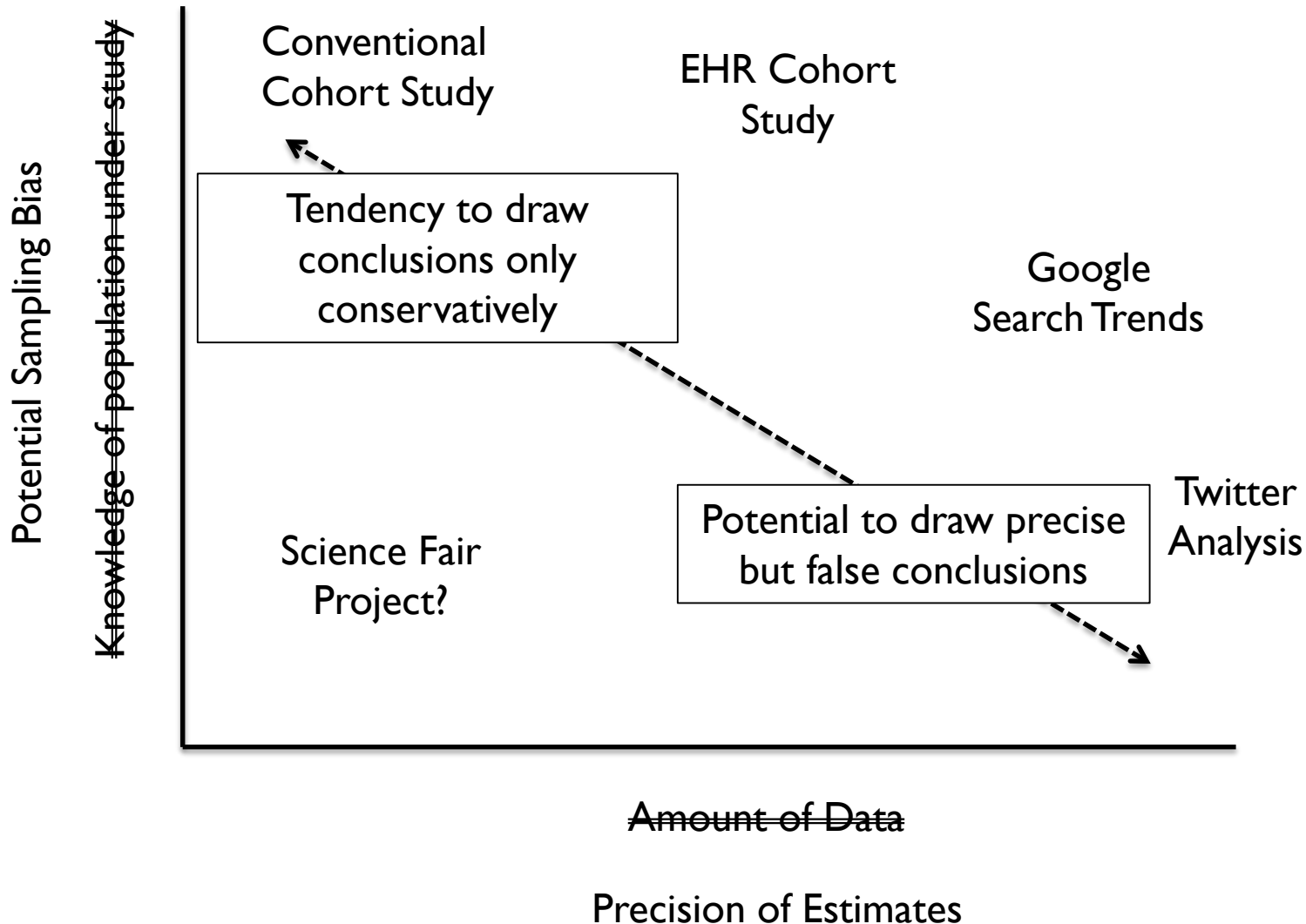
Uses of Big Data



Uses of Big Data

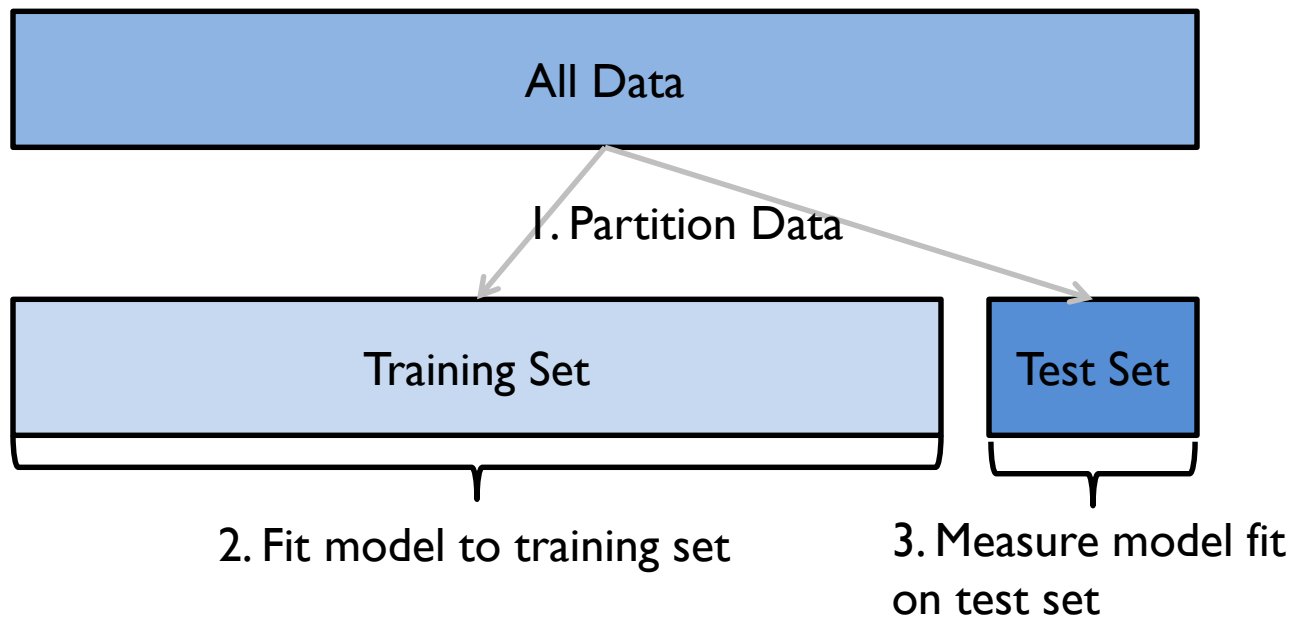


Uses of Big Data



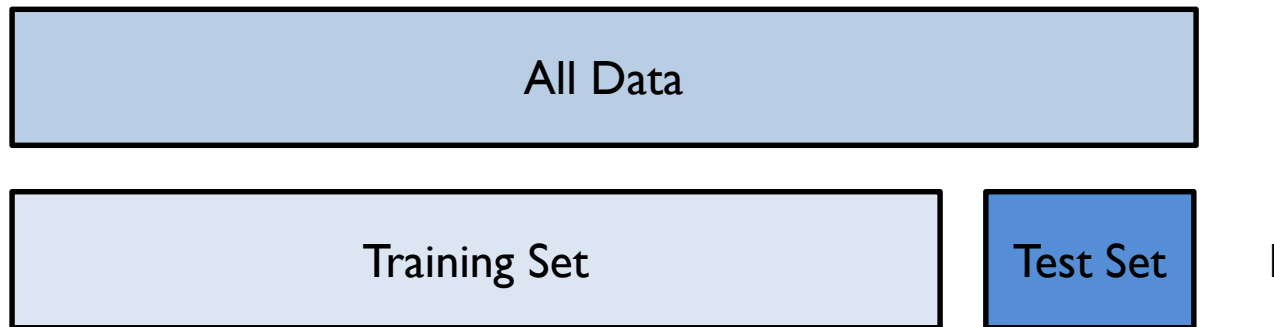
Machine learning: Cross-validation

- A method to assess model fit, primarily to minimize overfitting (for LASSO, by tuning the penalty)



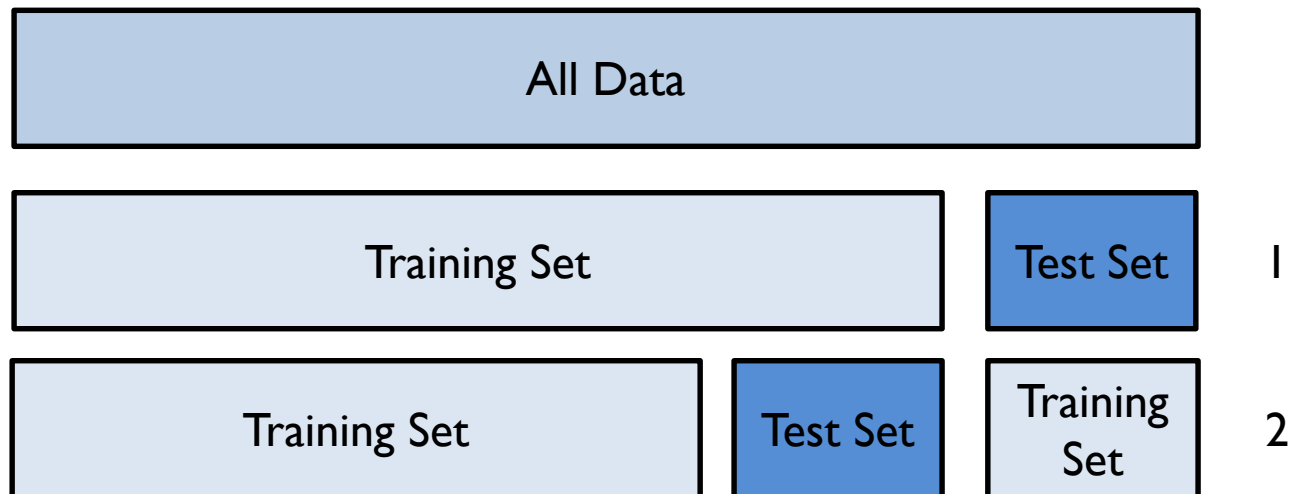
Machine learning: Cross-validation

- A method to assess model fit, primarily to minimize overfitting (for LASSO, by tuning the penalty)



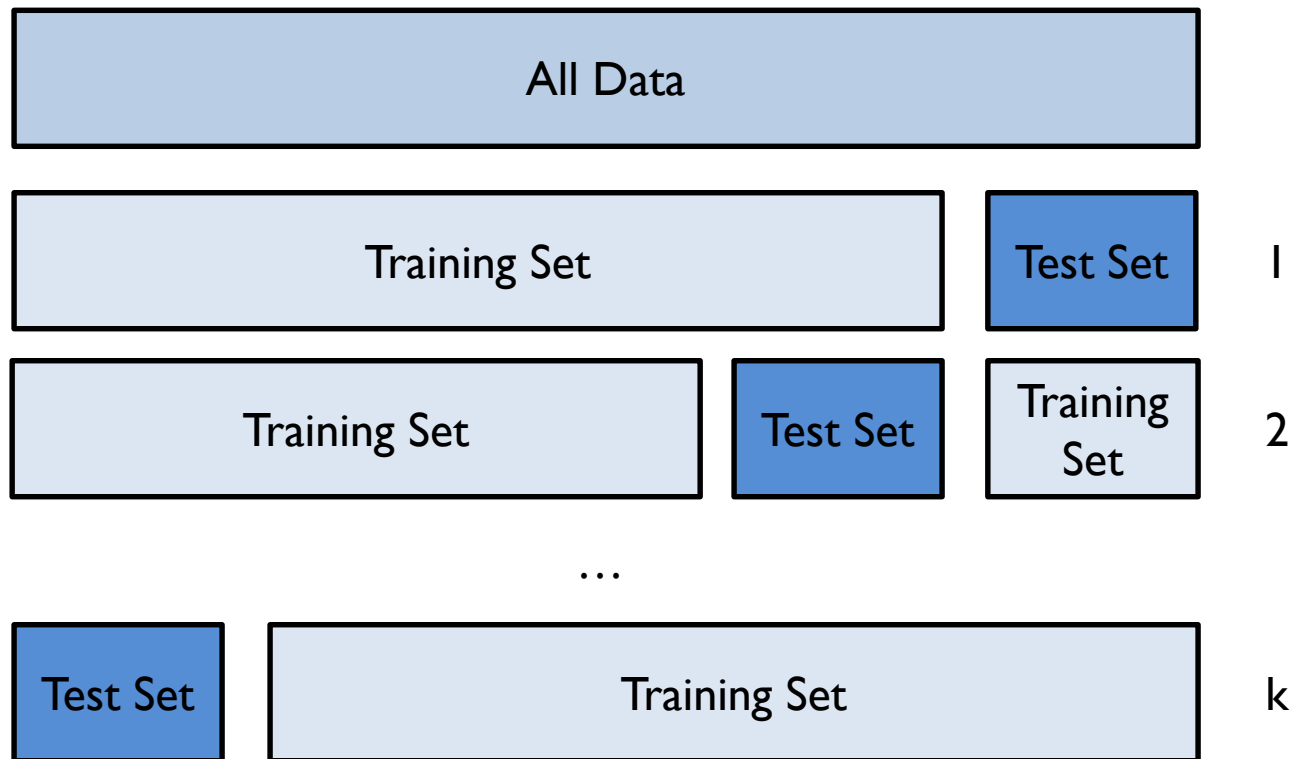
Machine learning: Cross-validation

- A method to assess model fit, primarily to minimize overfitting (for LASSO, by tuning the penalty)



Machine learning: K-fold cross-validation

- A method to assess model fit, primarily to minimize overfitting (for LASSO, by tuning the penalty)



LASSO results (Imputation I)

	PASE Score	Gardening	Walking Daily
Count of selected variables	3 (0.8%)	45 (12.7%)	22 (6.3%)
Top 5 variables	Income below half the poverty level (-)	Incomes under poverty, no kids < 18 (-)	Group quarters (+)
	Sales and related occupations (-)	Cul-de-sac intersections (+)	5-9 year old males (-)
	Incomes between 1/2 and 3/4 of poverty (-)	No issues with windows (+)	Hispanic or Latino householders (-)
		Farming, fishing, or forestry occupations(-)	Incomes below poverty, no kids< 18 (-)
		Commute times of 20 to 29 minutes (-)	Women aged 35-44 (-)

LASSO results (Imputation 2)

	PASE Score	Gardening	Walking Daily
Count of selected variables	23 (6.8%)	9 (2.6%)	23 (6.8%)
Top 5 variables	Income below half the poverty level (-)	Incomes just above the poverty level (-)	Broken windows (-)
	3-way Intersections (+)	Incomes below half the poverty level (-)	Households with a Hispanic householder (-)
	Incomes just below the poverty level (-)	Physical Disorder (-)	Residents who commute by car, truck, or van (-)
	30-39 minute commutes (+)	Incomes just below the poverty level (-)	Non-family households with non-relatives (+)
	Non-family households not living alone (+)	Vacant housing units for rent (-)	Household incomes 35-40K (-)