

Using Digital Data & Deep Learning to Improve Health

Adyasha Maharana
PhD Student, UNC Chapel Hill

Elaine Nsoesie
Assistant Professor of Global Health
Boston University

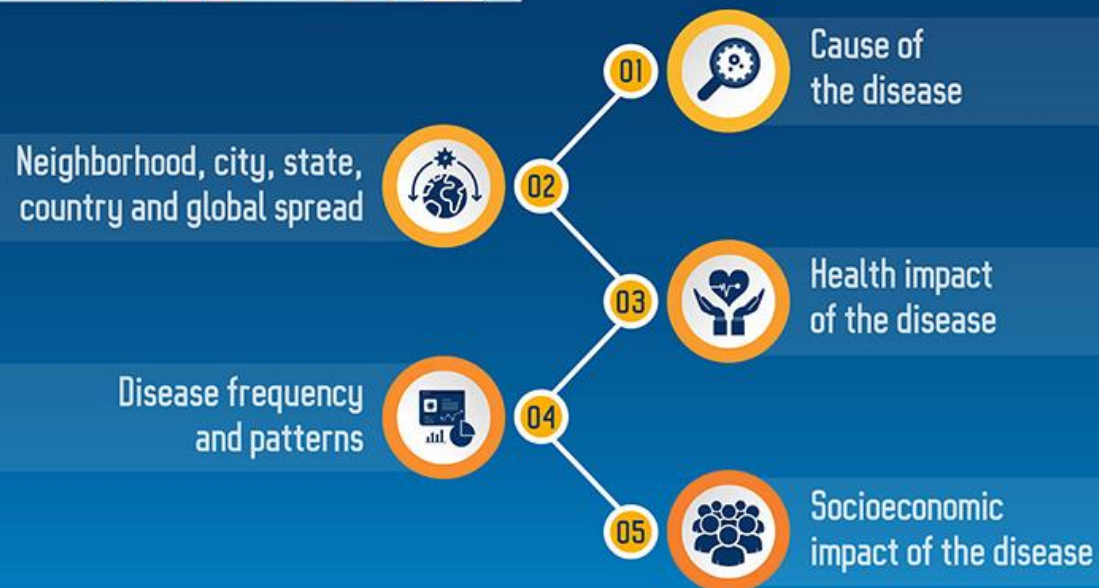


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

WHAT DO EPIDEMIOLOGISTS STUDY?



Epidemiologists are vital because they help paint a picture of what a disease does and how it can be prevented and treated. They do this by studying the following elements:



Source: Centers for Disease Control and Prevention

“The goal of epidemiology, very broadly speaking, is to understand the patterns of disease and health dynamics in populations as well as the causes of these patterns, and to use this understanding to mitigate and prevent disease, and to promote health.”

(Saluthe et al. 2018)

Digital Epidemiology

Broadly speaking, digital epidemiology is epidemiology that uses digital data.

But more importantly, digital epidemiology is epidemiology that uses data that was generated outside the public health system, i.e. *with data that was not generated with the primary purpose of doing epidemiology.*

(Saluthe et al. 2018)

Types of Data

Crowdsourcing



Search



Social Media



Consumer Reviews



Remote Sensing and Place



Google Maps



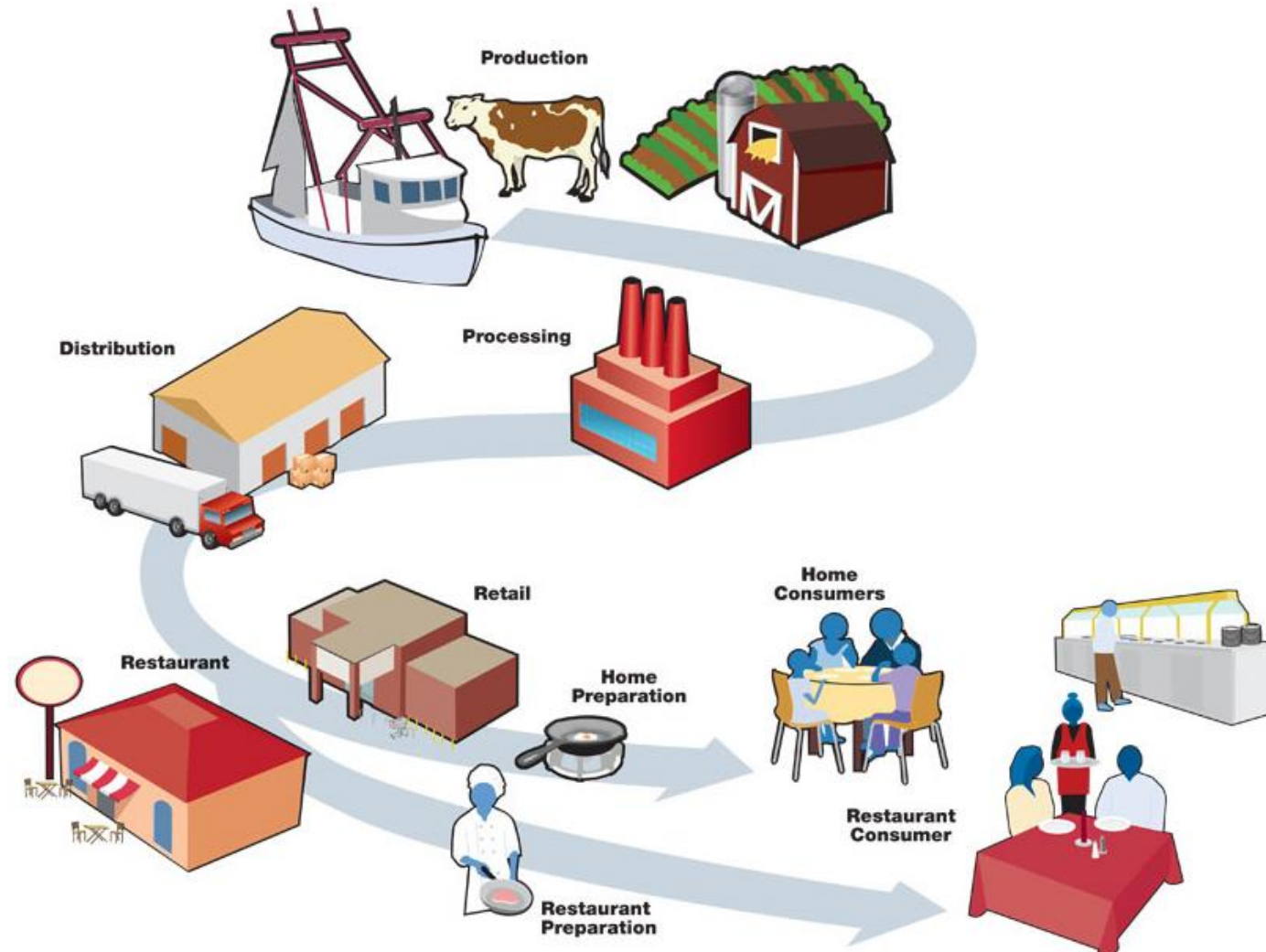
News

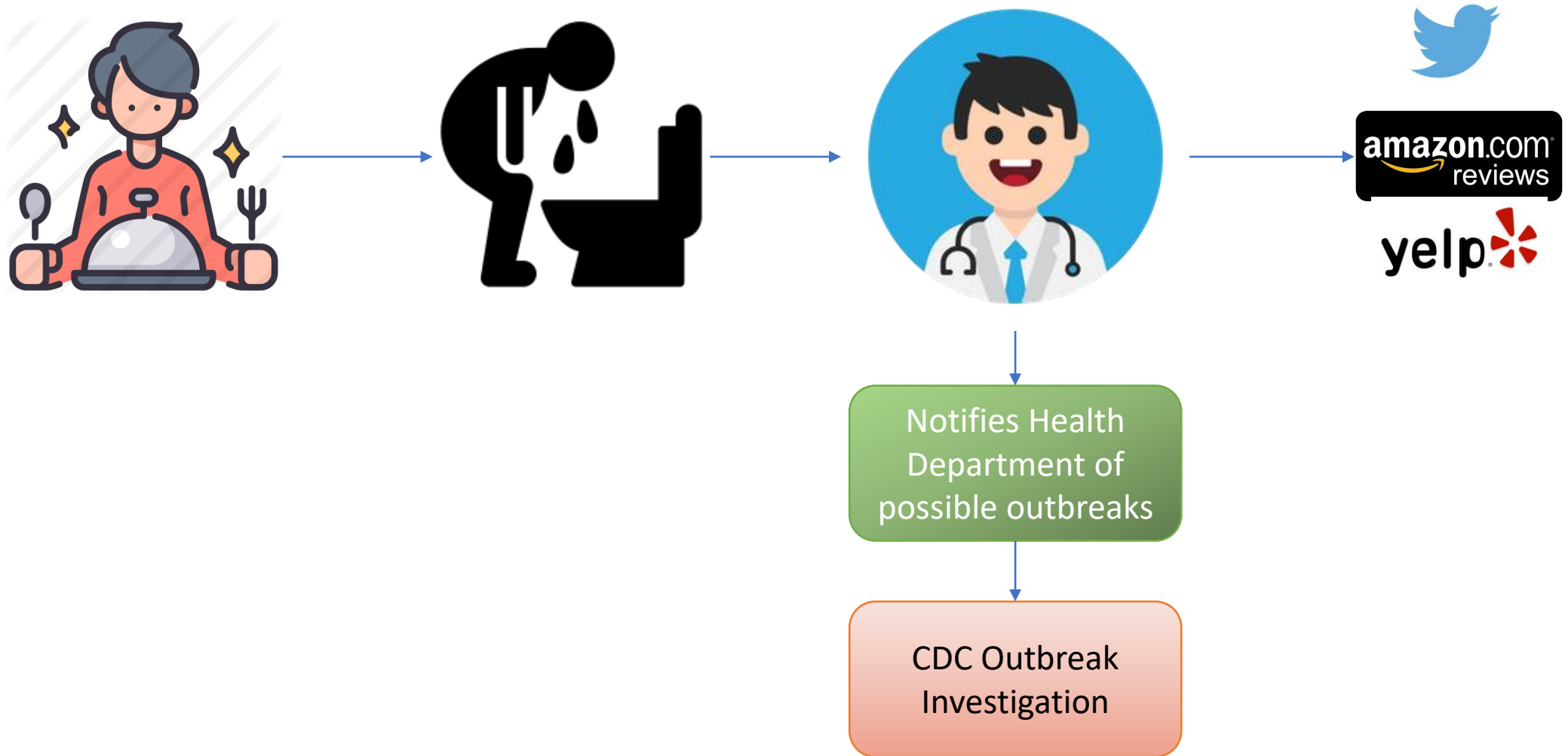


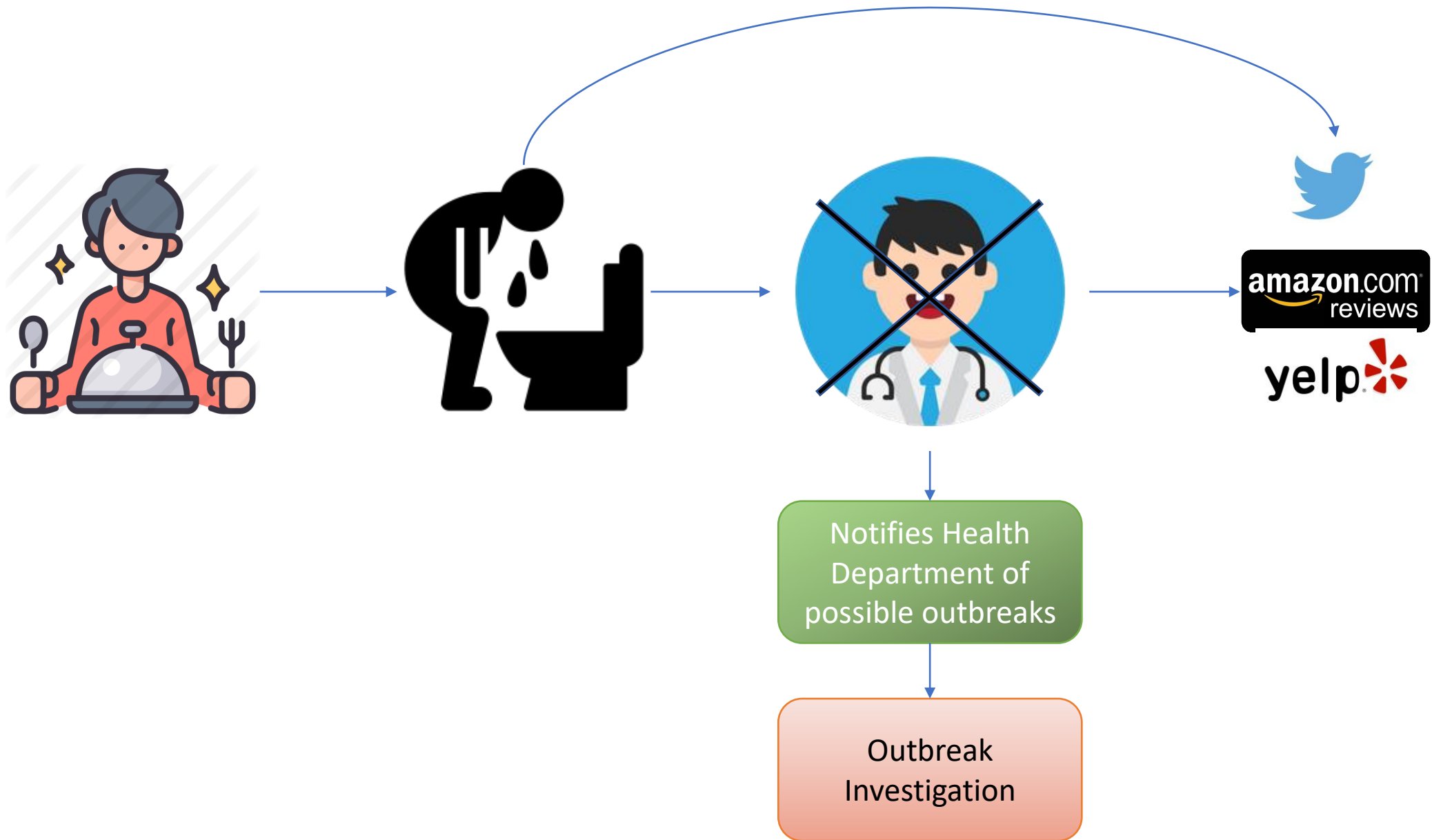
Detecting Reports of Unsafe Foods in Consumer Reviews

Jared B. Hawkins, Gaurav Tuli, Sheryl Kluberg, Jenine Harris, John S. Brownstein, Adyasha Maharana, Kunlin Cai, Joseph Hellerstein, Yulin Hswen, Michael Munsell, Valentina Staneva, Miki Verma, Cynthia Vint, Derry Wijaya, Elaine Nsoesie

Surveillance of Foodborne Illness and Unsafe Foods







Detecting reports of unsafe foods in consumer product reviews

Adyasha Maharana, Kunlin Cai, Joseph Hellerstein, Yulin Hswen, Michael Munsell, Valentina Staneva, Miki Verma, Cynthia Vint, Derry Wijaya, Elaine O Nsoesie 

[Author Notes](#)


JAMIA Open, Volume 2, Issue 3, October 2019, Pages 330–338,


<https://doi.org/10.1093/jamiaopen/ooz030>

Published: 05 August 2019 **Article history** ▼



PDF

 Split View

 Cite



Permissions



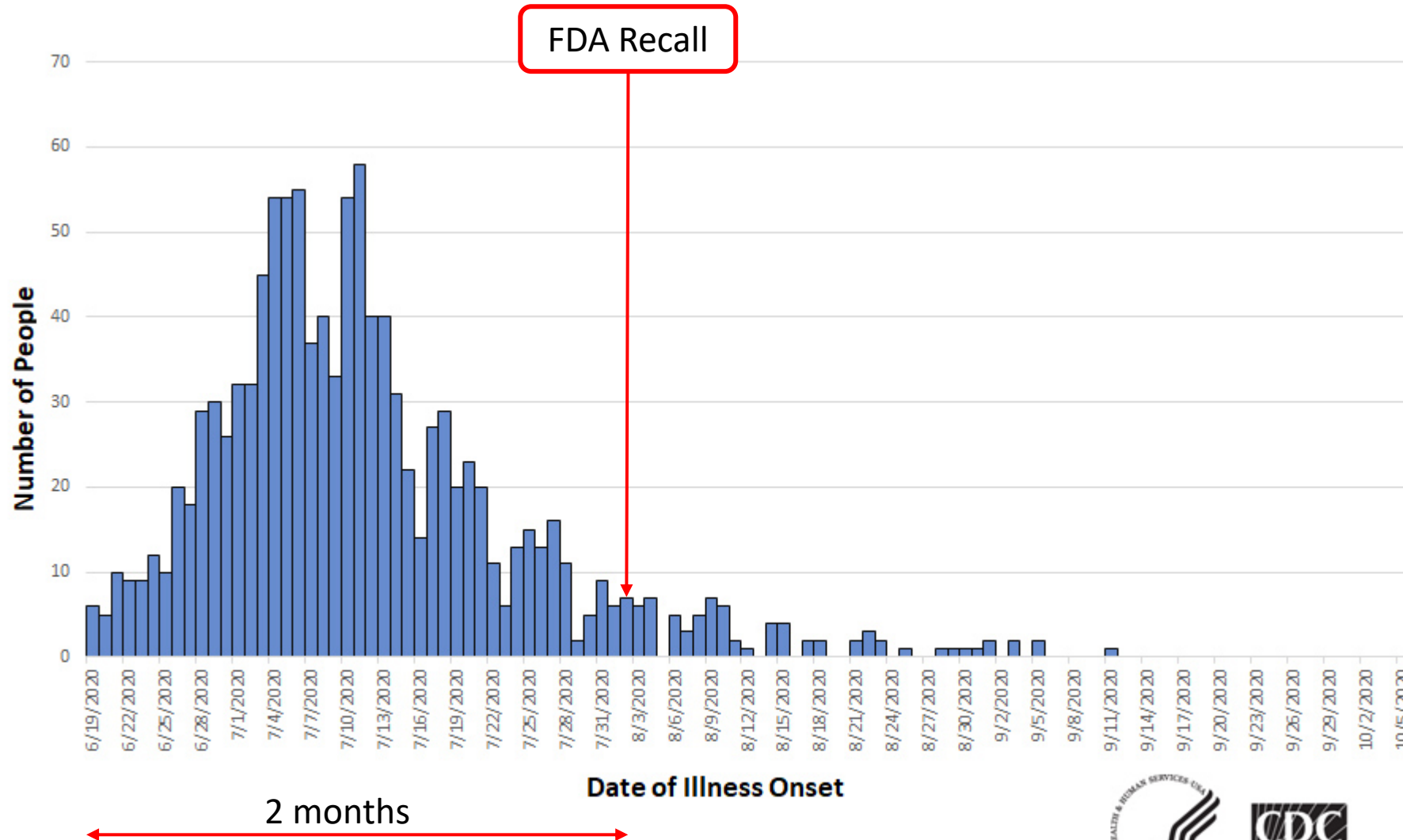
Share ▼

Abstract

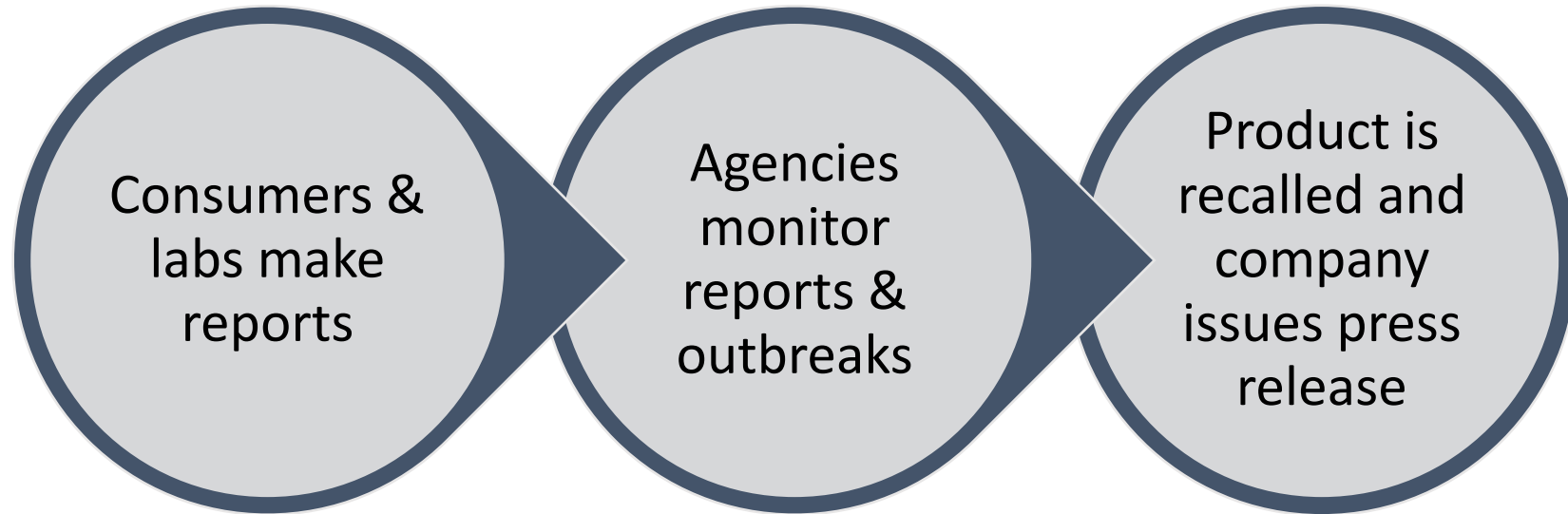
Objectives

Access to safe and nutritious food is essential for good health. However, food can become unsafe due to contamination with pathogens, chemicals or toxins, or mislabeling of allergens. Illness resulting from the consumption of unsafe foods is a global health problem. Here, we develop a machine learning approach

People infected with the outbreak strain of *Salmonella* Newport by date of illness onset*



(Simplified) Recall Process



1,297, 156

Amazon reviews for Grocery & Gourmet Food products



5,149

Reviews for recalled products



0.4%

Percentage of reviews

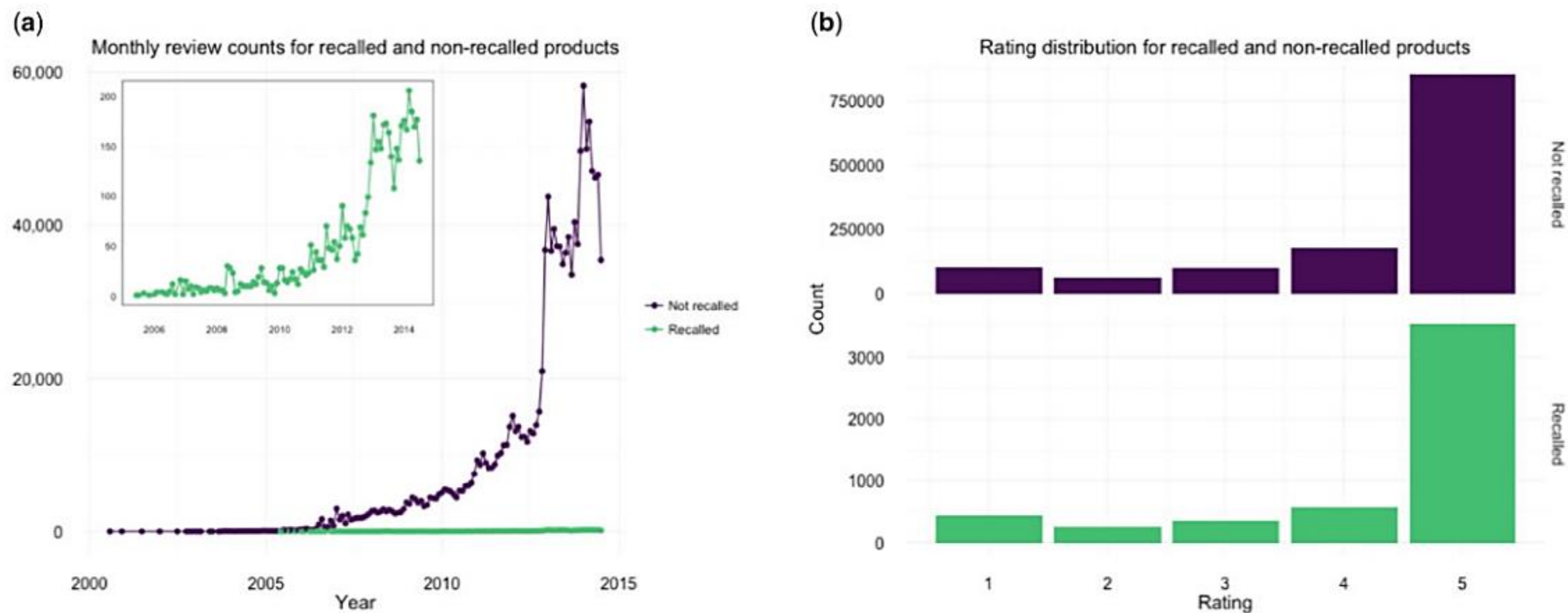
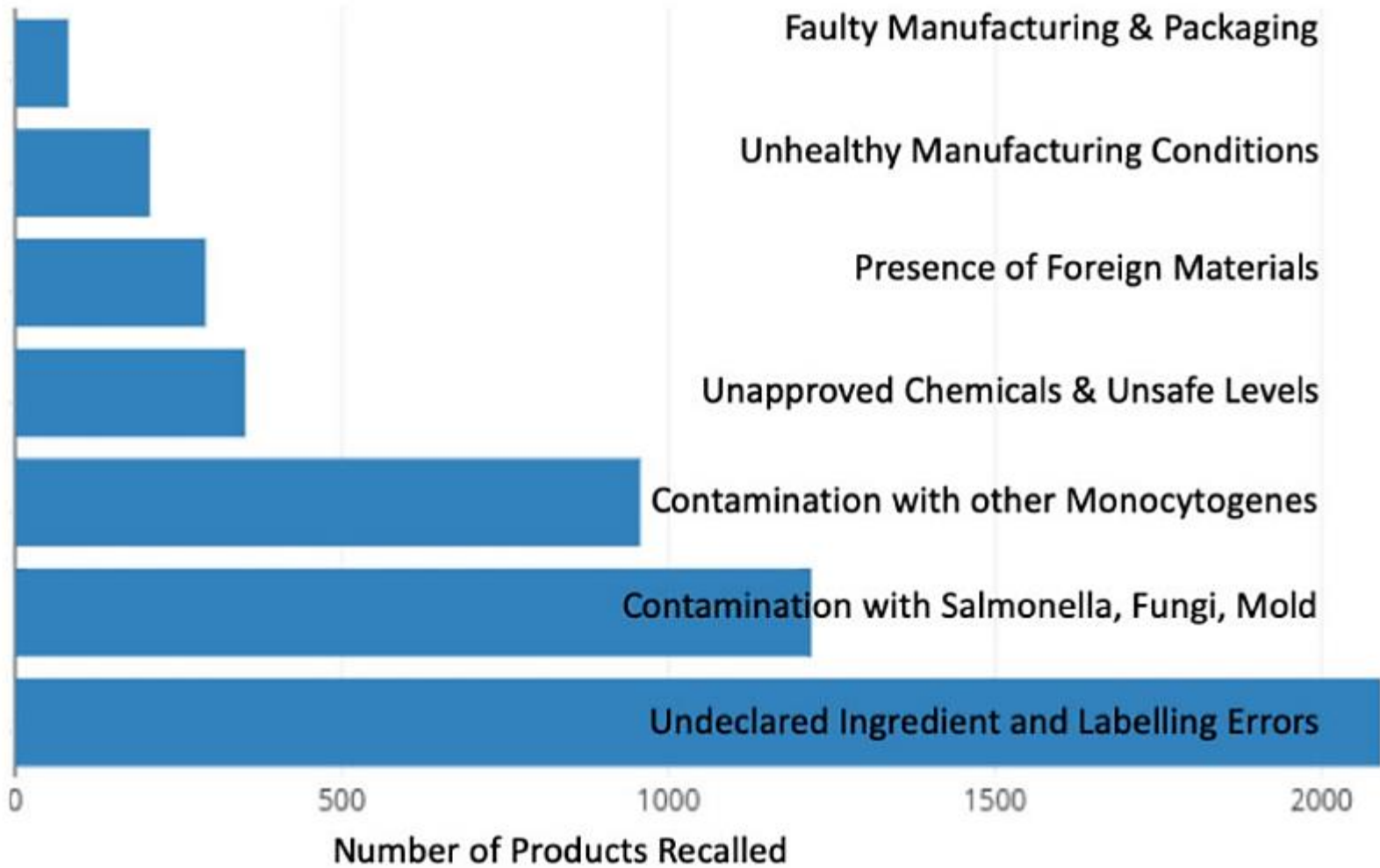


Figure 3. Features of Amazon reviews for the study period. Temporal trends (a) and distribution of customer ratings (b) of Amazon reviews.

Breakdown of Reasons for Recalled Products



6000 Reviews were manually annotated by three labelers

1. Review implies that consumer fell sick/had allergic reactions or has labeling errors
2. Review implies that the product expired or looks/tastes foul and should be inspected
3. Review does not imply that the product is unsafe
4. Review cannot be categorized to the above three categories

352 reviews out of 6000 in Category 1

Example Reviews

“I took the pills as described on the label, but after a few days the pills started upsetting my stomach real bad and made me nauseas”

“I am ANGRY with this company!!!! ... The label does NOT list lemon juice as an ingredient... I contacted the company via Facebook. Their answer, Unfortunately, there has been an error on our packaging which left the lemon juice concentrate off of the ingredient panel on a few pouch production runs...”

Table 2. Performance of the various machine learning approaches employed for identifying unsafe food products

Classifier description	Precision	Recall	F1 score
Linear SVM (Feature selection using Chi^2 $k = 500$)	0.61	0.64	0.62
Multinomial Naive Bayes (Feature selection using Chi^2 , $k = 500$)	0.66	0.66	0.66
Weighted logistic regression (Feature selection using Chi^2 , $k = 500$)	0.58	0.74	0.65
Weighted logistic regression (Feature selection using Chi^2 , $k = 1000$)	0.64	0.71	0.67
Weighted logistic regression (Feature selection using mutual information, $k = 1000$)	0.60	0.68	0.64
Weighted logistic regression with SMOTE (ratio = 1: 5) (tested on real data points only)	0.62	0.68	0.65
Weighted logistic regression with SMOTE (ratio = 1: 3) (tested on real data points only)	0.62	0.71	0.66
Weighted logistic regression with SMOTE (ratio = 1: 2) (tested on real data points only)	0.62	0.70	0.66
Weighted logistic regression with SMOTE (ratio = 1: 1) (tested on real data points only)	0.63	0.66	0.64
BERT (epoch = 10, max sequence length = 128)	0.76	0.67	0.71
BERT (epoch = 10, max sequence length = 128) with focal loss for dealing with imbalanced data ($\alpha = 0.915$, $\gamma = 5$)	0.75	0.74	0.73
BERT (epoch = 20, max sequence length = 256)	0.79	0.67	0.72
BERT (epoch = 30, max sequence length = 256)	0.78	0.71	0.74
BERT (epoch = 30, max sequence length = 256) with focal loss for dealing with imbalanced data ($\alpha = 0.915$, $\gamma = 5$)	0.77	0.71	0.74

BERT is the best performing classifier. Chi^2 refers to Chi-square. The accuracy ($[\text{true positives} + \text{true negative}]/\text{total reviews}$), precision (also known as positive predictive value = $\text{true positives}/\text{predicted positive condition}$), recall (also known as sensitivity = $[\text{true positive}/[\text{true positives} + \text{false negatives}]]$), and F1-score (the harmonic mean of the precision and recall) are discussed.

Implications

- The World Health Organization estimates that in 2010, 600 million people experienced illness due to contaminated food, globally.
- Early identification means regulatory organizations and companies can take appropriate actions to stop the sale of these products.
- Limit the occurrence of large foodborne disease outbreaks.

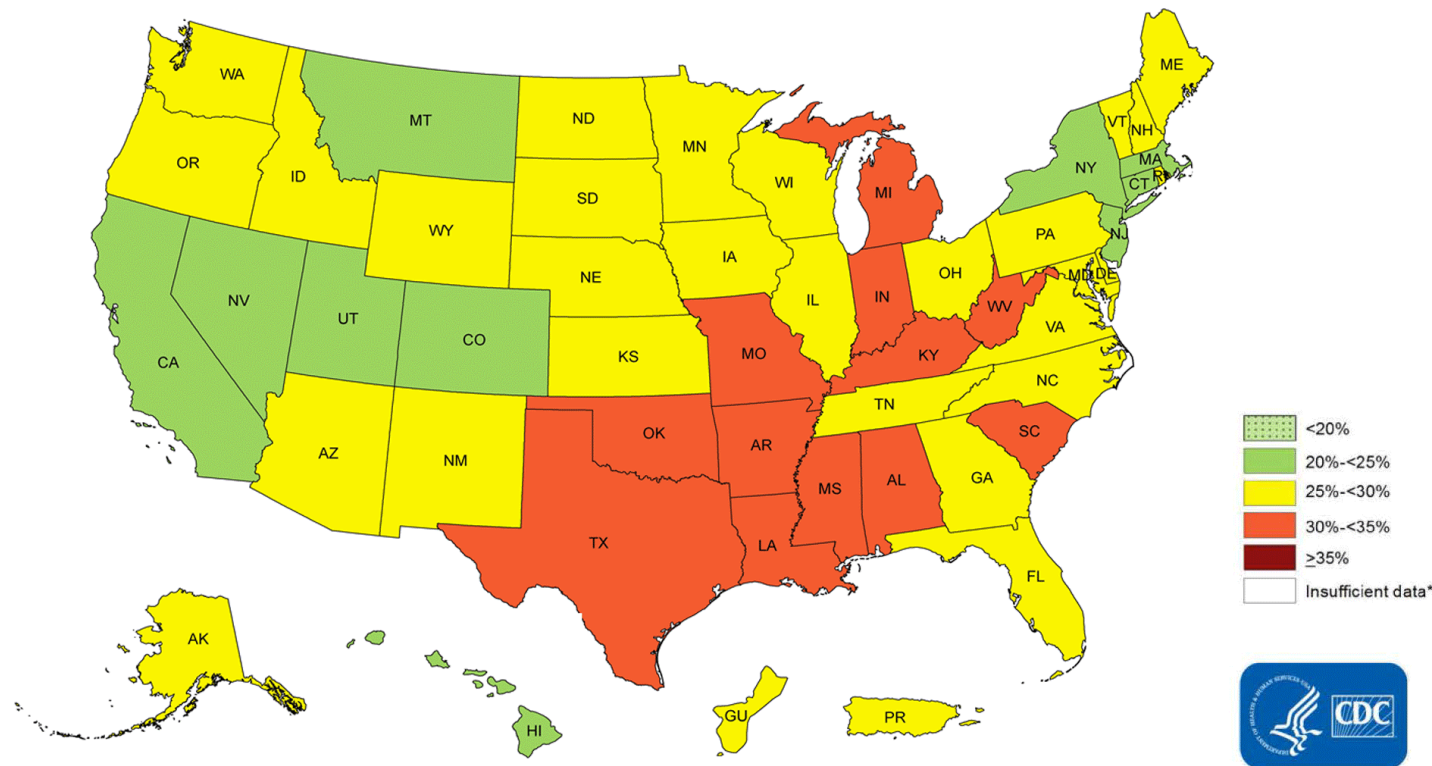
The Built Environment and Obesity in the US

Adyasha Maharana, Elaine Nsoesie

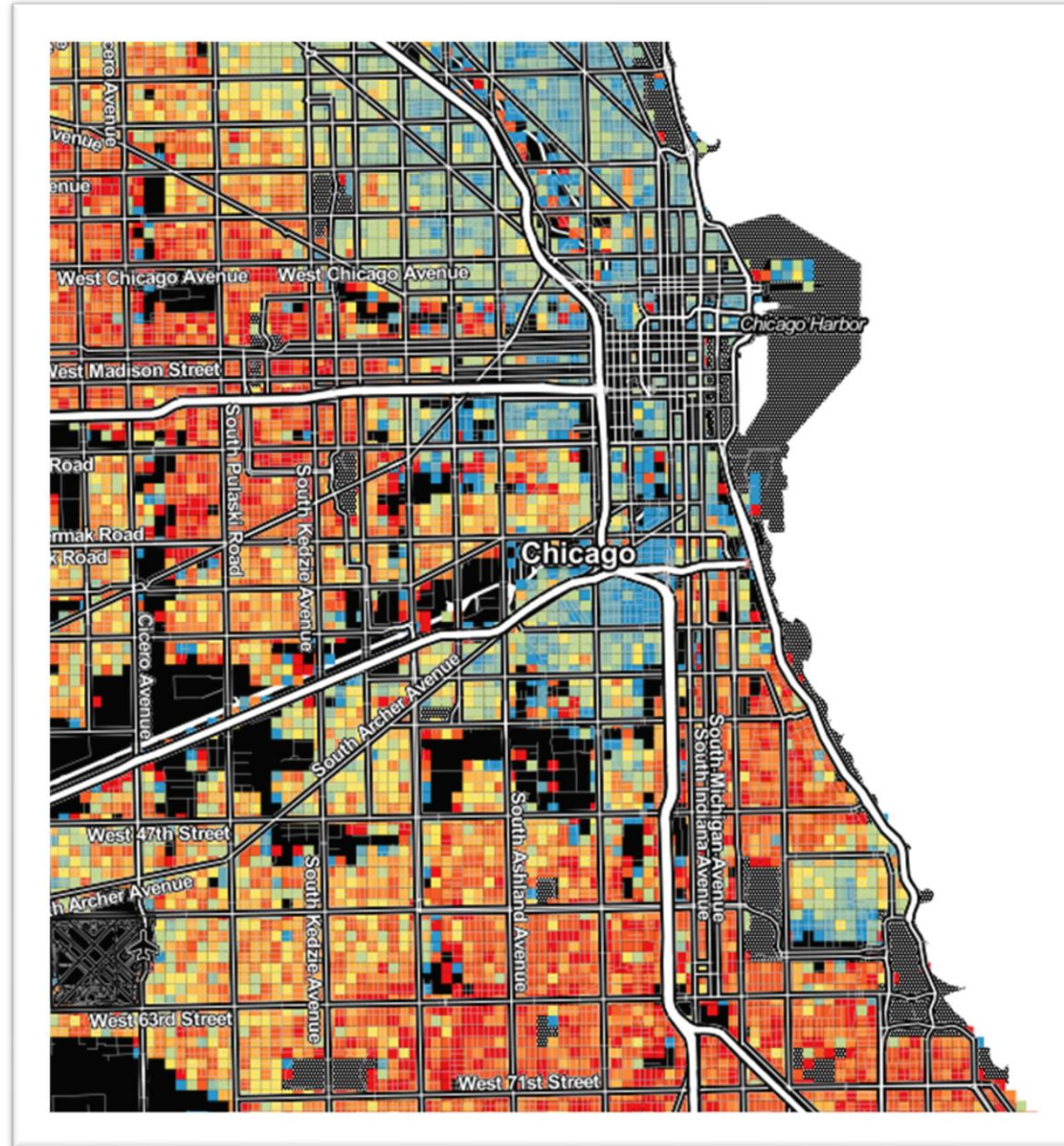
Prevalence[†] of Self-Reported Obesity Among U.S. Adults by State and Territory, BRFSS

[†]Prevalence estimates reflect BRFSS methodological changes started in 2011. These estimates should not be compared to prevalence estimates before 2011.

2011 2012 2013 2014 2015 2016 2017 2018 2019



*Sample size <50 or the relative standard error (dividing the standard error by the prevalence) ≥30%.



Legend

	0.0 - 24.24
	24.25 - 28.57
	28.58 - 32.5
	32.51 - 35.29
	35.30 - 37.5
	37.51 - 40.30
	40.31 - 43.14
	43.15 - 46.51
	46.52 - 52.38
	52.39 - 100.0

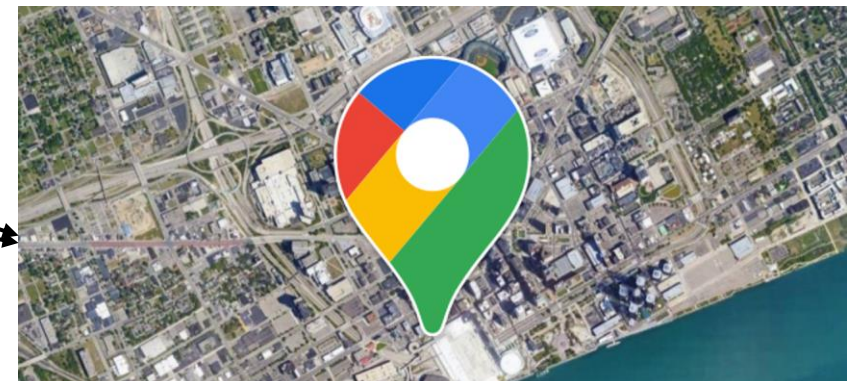
% obese population

There is widespread disparity in obesity prevalence and other health indicators not only at national level but also within a city. Imagine walking 15 minutes in a city and ending up in an area with 20 years lesser life expectancy than you starting point.

Data Sources for Factors Affecting Obesity



Data Sources for Factors Affecting Obesity

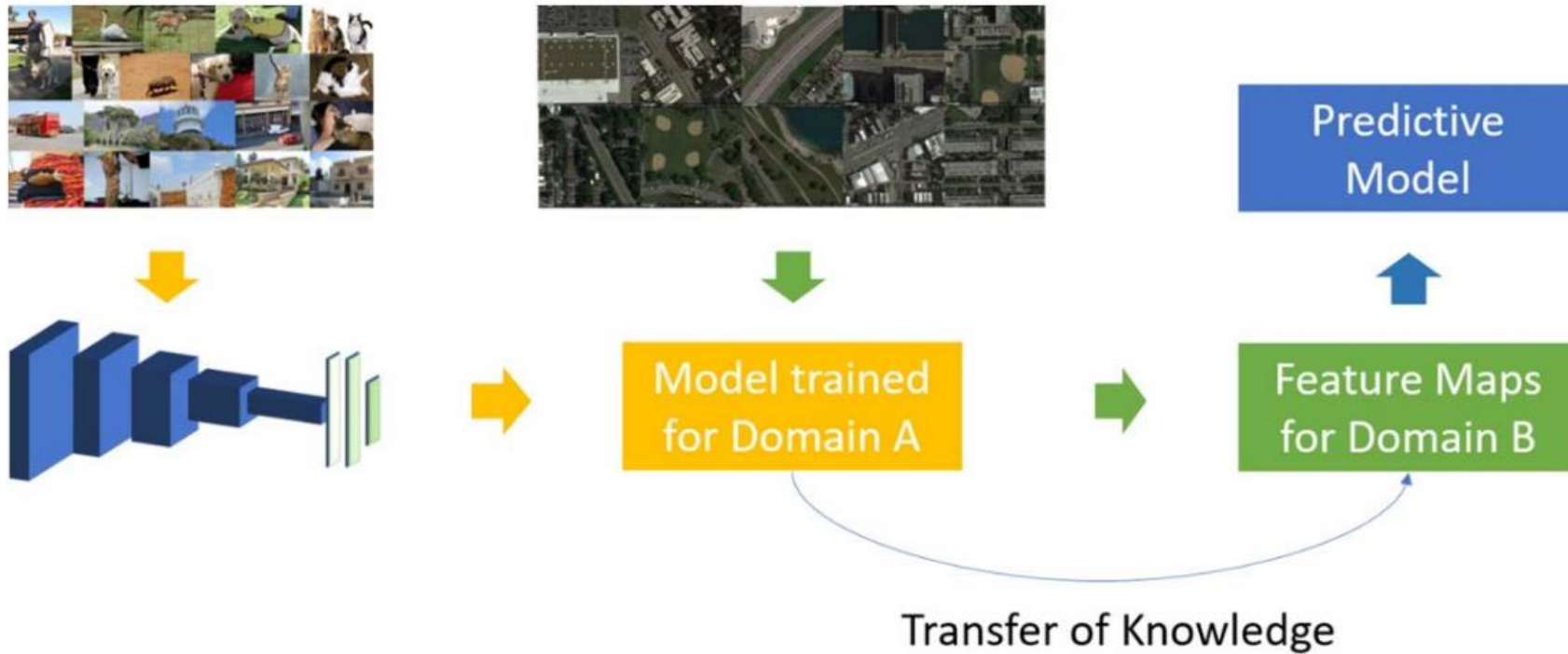


Experimental Settings

- Locations: Los Angeles, California; Memphis, Tennessee; San Antonio, Texas; and Seattle-Tacoma-Bellevue, Washington (1695 census tracts)
- Explanatory Variables:
 - Built Environment: 150000 Google Maps satellite images
 - Places of Interest(POI) : Google POI data at corresponding locations
- Response Variable:
 - Obesity Prevalence: Census-tract level data from 500 Cities Project



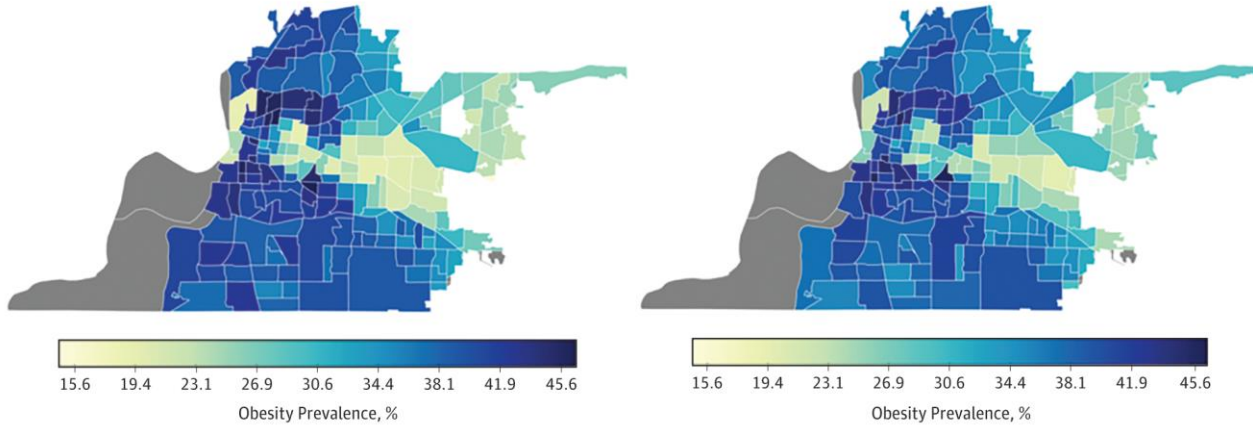
Modelling Approach



Predictive Modelling

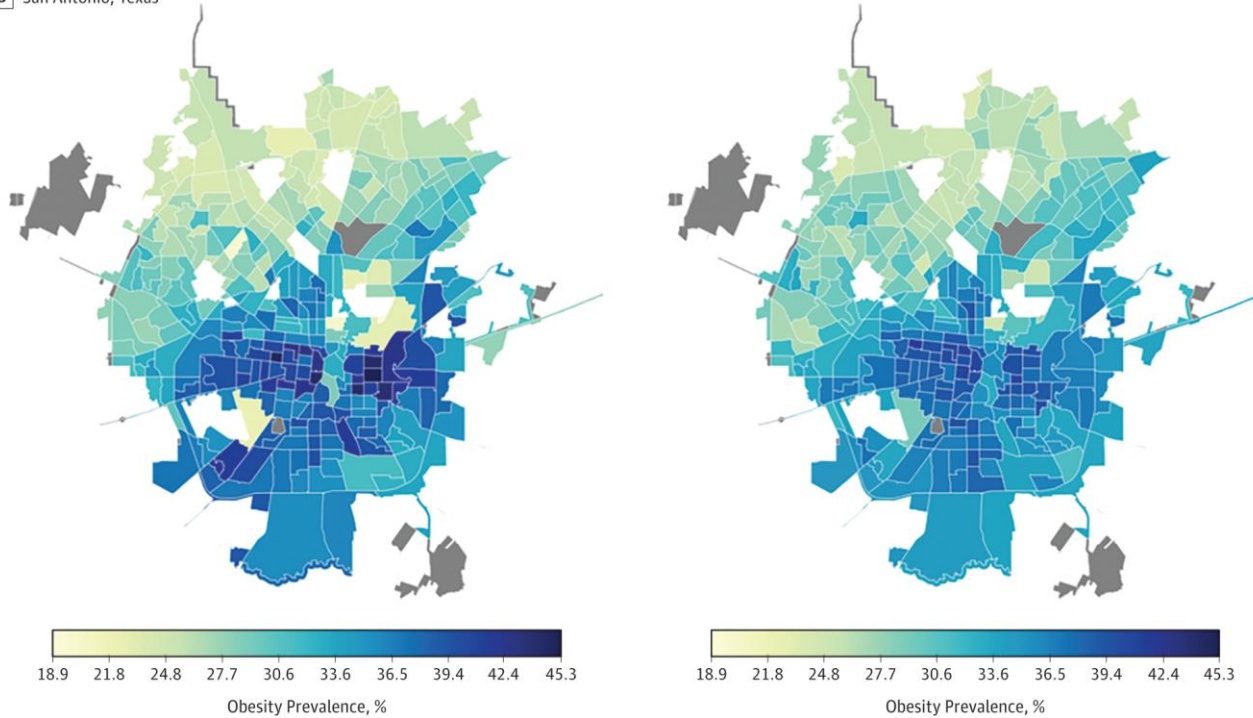
- We applied Elastic Net a regularized regression method that eliminates insignificant covariates, preserves correlated variables, and is well suited to the high-dimensional ($n = 4096$) feature vectors.
- After regularization, we retain 125 features in the model.
- We perform 5-fold cross-validation as well as out-of-sample regression analysis

A Memphis, Tennessee



Built Environment features explain 73.3% and 61.5% variation in obesity prevalence for Memphis and San Antonio respectively in out-of-sample estimates.

B San Antonio, Texas

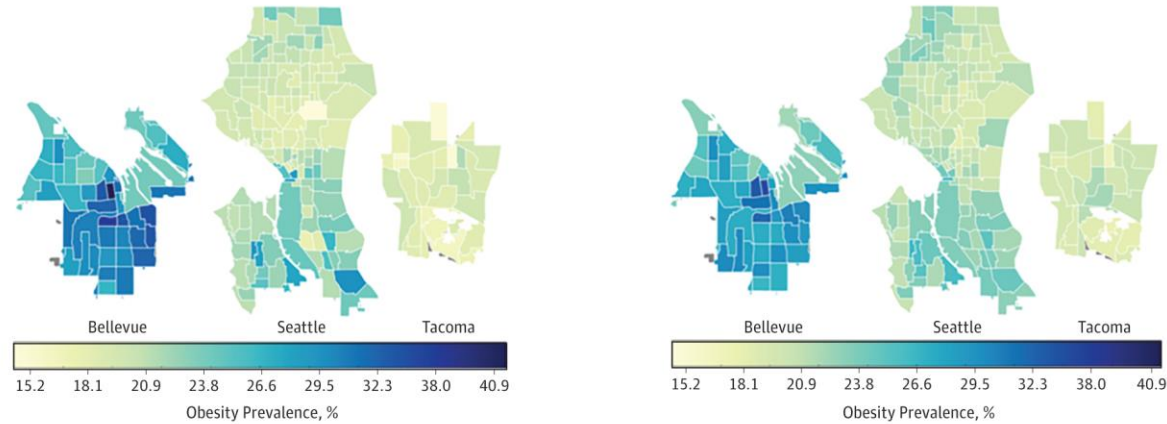


Places of Interest features explain 43.2% and 43.0% variation in obesity prevalence for Memphis and San Antonio respectively in out-of-sample estimates.

Actual Obesity Prevalence

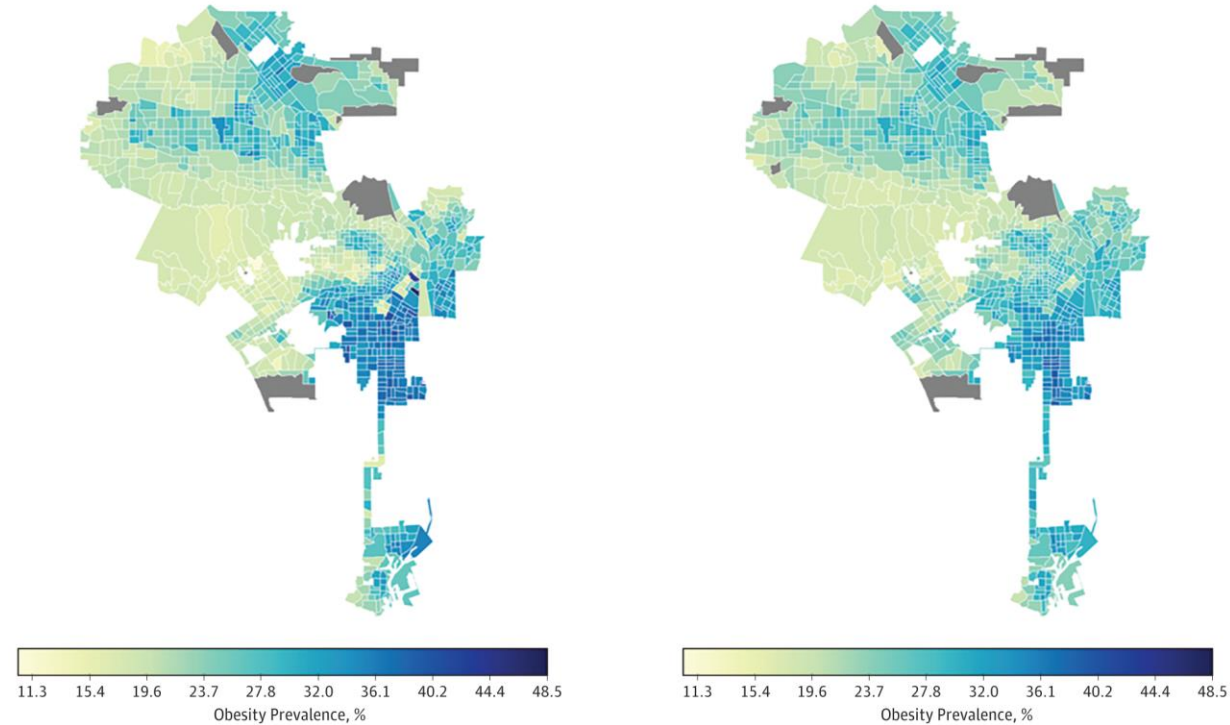
Cross-validation estimates

A Seattle, Washington



Built Environment features explain 55.8% and 56.1% variation in obesity prevalence for Sea-Tac-Bel and Los Angeles respectively in out-of-sample estimates.

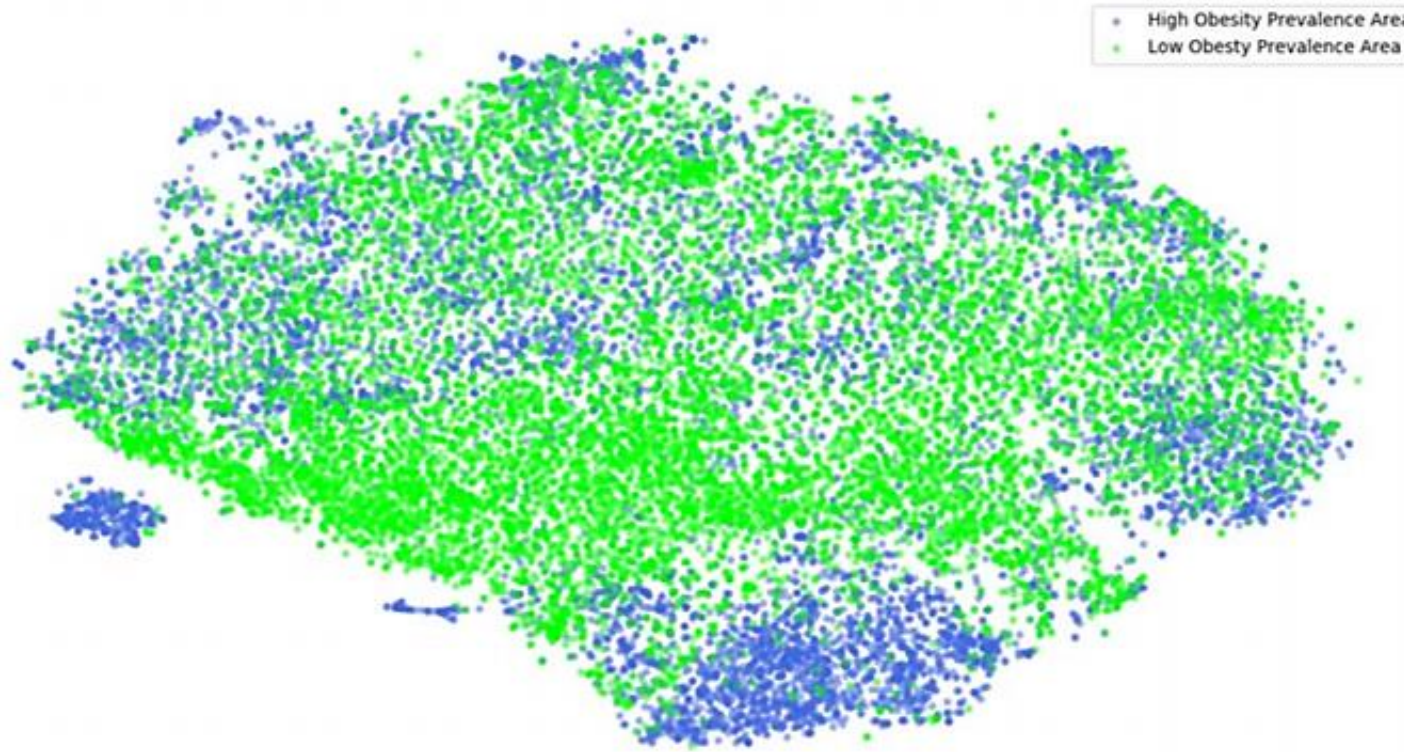
B Los Angeles, California



Places of Interest features explain 14.0% and 29.2% variation in obesity prevalence for Memphis and San Antonio respectively in out-of-sample estimates.

Actual Obesity Prevalence

Cross-validation estimates



t-SNE Visualization for High and Low Obesity Areas in San Antonio, Texas



Green Spaces in Low-Obesity Clusters

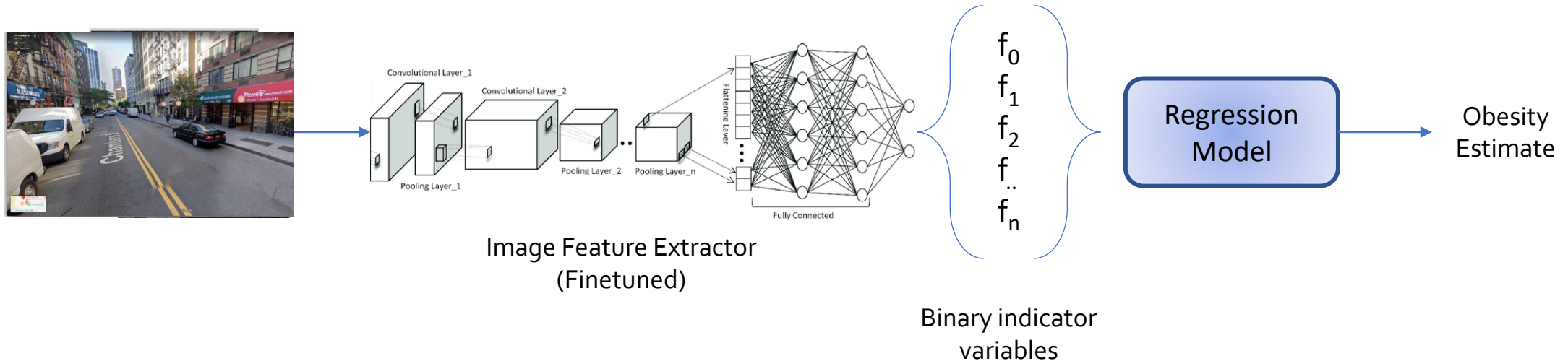


Sparse Greenery in High-Obesity Clusters

Neighborhood Looking Glass: Estimation of Health Indicators from Google Street View

(Nguyen et al. 2018, Javanmardi et al. 2020)

Interpretable Obesity Estimation from Built Environment



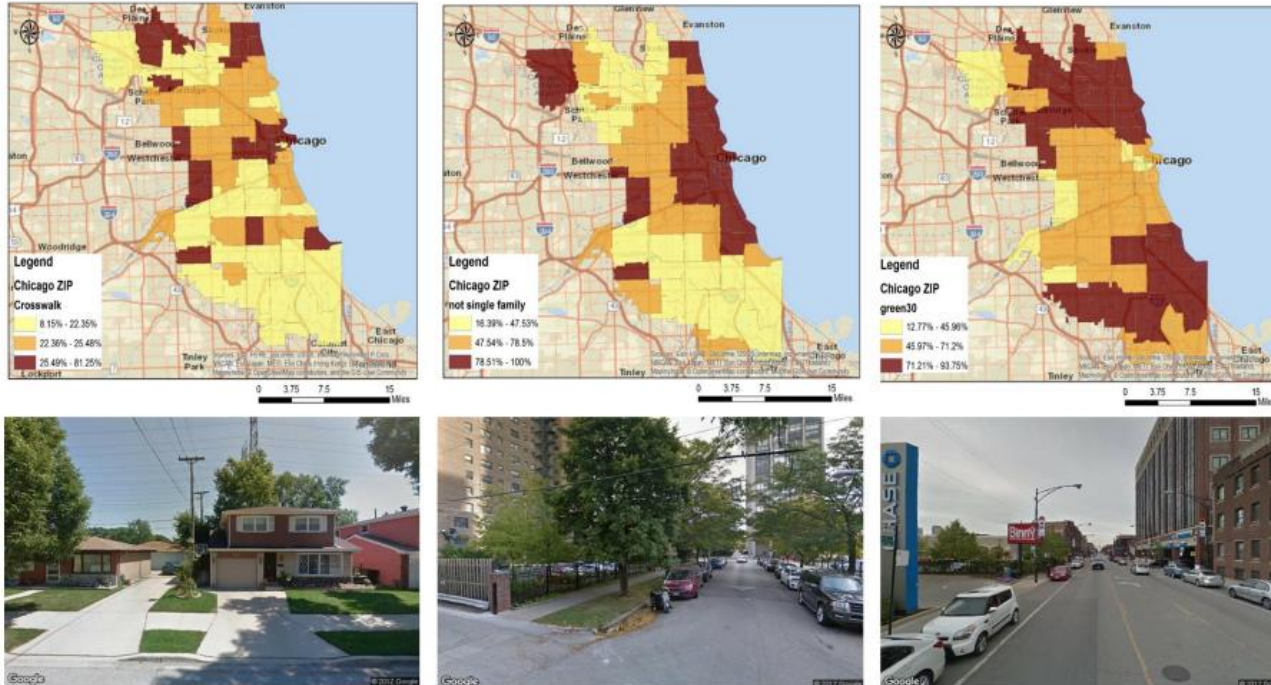
Finetuned deep learning models are used to label Google Street View images with built environment features which are used for obesity estimate.

This allows for interpretable estimation and is helpful for public policy-making.

Experimental Settings

- Locations: Chicago, Salt Lake City, Charleston
- Nguyen et al. 2018 use following binary features:
 1. Green Street (Trees/landscaping comprises at least 30% of the image)
 2. Building Type (Single-family detached house vs. other)
 3. Presence of Crosswalk (Yes vs. No)
- VGG-16 was finetuned for classification on 14000 annotated images.
 - Achieved 85.40%, 84.59% and 93.03% accuracy on 1, 2, 3 respectively
- Sociodemographic characteristics of each neighborhood were added to adjust for potential confounding

Results



Crosswalks



Commercial Buildings



Green Streets

Zip Code Distribution of Built Environment Characteristics in Chicago, Illinois.

Table 1 Descriptive characteristics of neighbourhood characteristics

	Green streets	Crosswalk present	Commercial/apartment building
	Mean (SD)	Mean (SD)	Mean (SD)
Salt Lake City	59.0 (49.2)	8.0 (27.1)	38.5 (48.7)
Chicago, Illinois	71.2 (45.3)	22.5 (41.8)	55.8 (49.7)
Charleston, West Virginia	78.6 (41.0)	3.4 (18.1)	44.9 (49.7)
N	2 26 875	1 50 300	53 360

Neighbourhood characteristics derived from street images collected between December 2016 and February 2017 from Google's Street View Image API.

Charleston had highest % of green streets (79%).

Chicago had most commercial buildings (56%) and streets with crosswalks (23%).

Percentage of blacks was correlated with *more* crosswalks and *more* commercial buildings, converse for Hispanics.

Table 2 Built environment predictors of adult obesity and diabetes,*
Salt Lake City

Built environment characteristics	Obese	Diabetes
	Prevalence ratio (95% CI)†	Prevalence ratio (95% CI)†
Green streets		
Third tertile (highest)	0.73 (0.63 to 0.85)	0.86 (0.77 to 0.96)
Second tertile	0.99 (0.92 to 1.06)	1.03 (0.97 to 1.08)
Crosswalks		
Third tertile (highest)	0.76 (0.69 to 0.85)	0.87 (0.80 to 0.95)
Second tertile	1.02 (0.97 to 1.07)	1.01 (0.95 to 1.06)
Commercial buildings/apartments		
Third tertile (highest)	0.79 (0.67 to 0.94)	0.81 (0.67 to 0.98)
Second tertile	0.93 (0.86 to 1.01)	0.91 (0.84 to 0.99)
N	727 737	736 218

*Data source for health outcomes: Utah Population database.

†Adjusted Poisson models were run for each outcome separately. Models controlled for individual-level age, sex, race, ethnicity, education and marital status as well as zip code-level population density, percentage of the population 65 years and older, percentage of Hispanics, percentage of blacks, median household income and percentage of householder living in current residence for 5 years or more. Built environment characteristics were categorised into tertiles, with the lowest tertile serving as the referent group. SEs were adjusted for clustering of values at the zip code level.

Tertile = Any of the three groups containing a third of the population

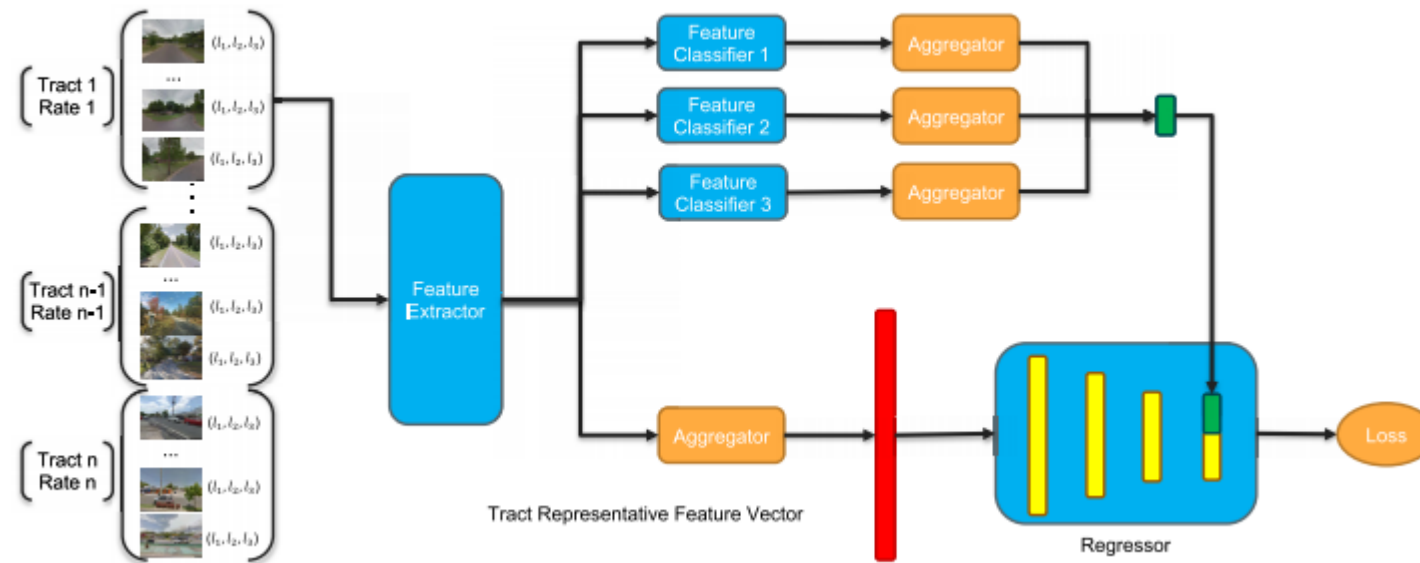
Lowest tertiles serve as referent group.

More green streets, crosswalks and presence of commercial buildings were associated with lower individual obesity prevalence.

Similarly for diabetes prevalence.

Findings are consistent with prior literature. Mix land uses, connected streets, higher residential density and attractive scenery promote physical activity.

Multi-tasking for Classification & Estimation



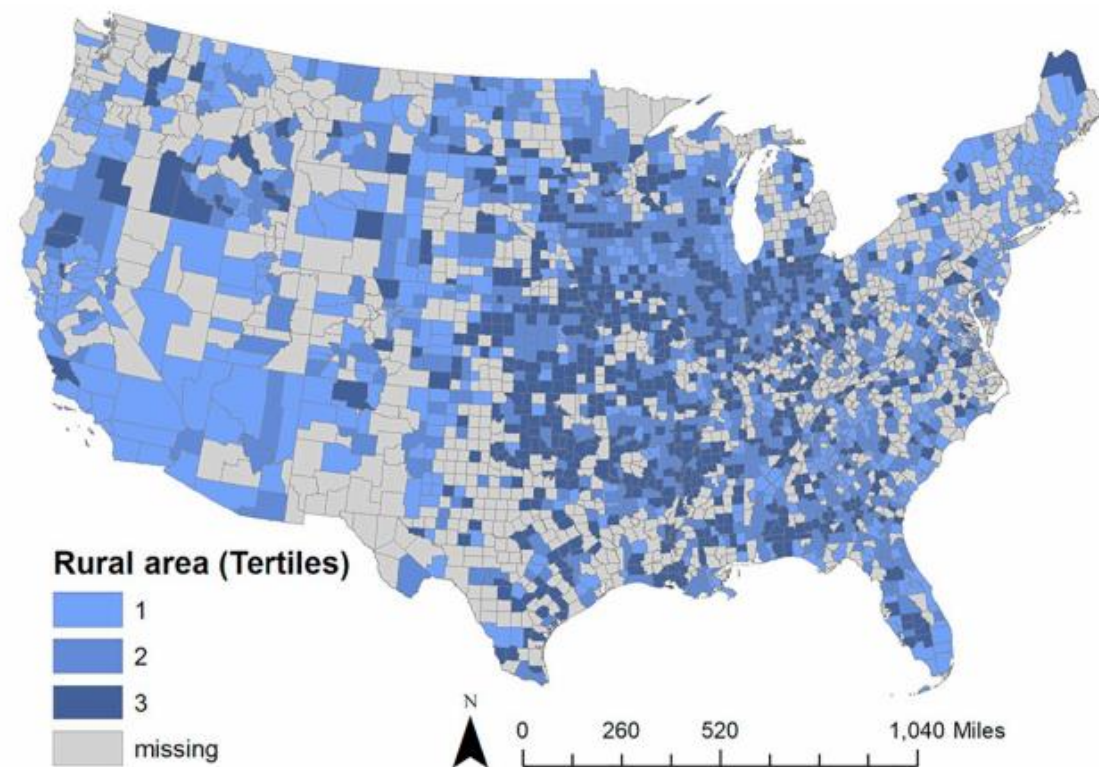
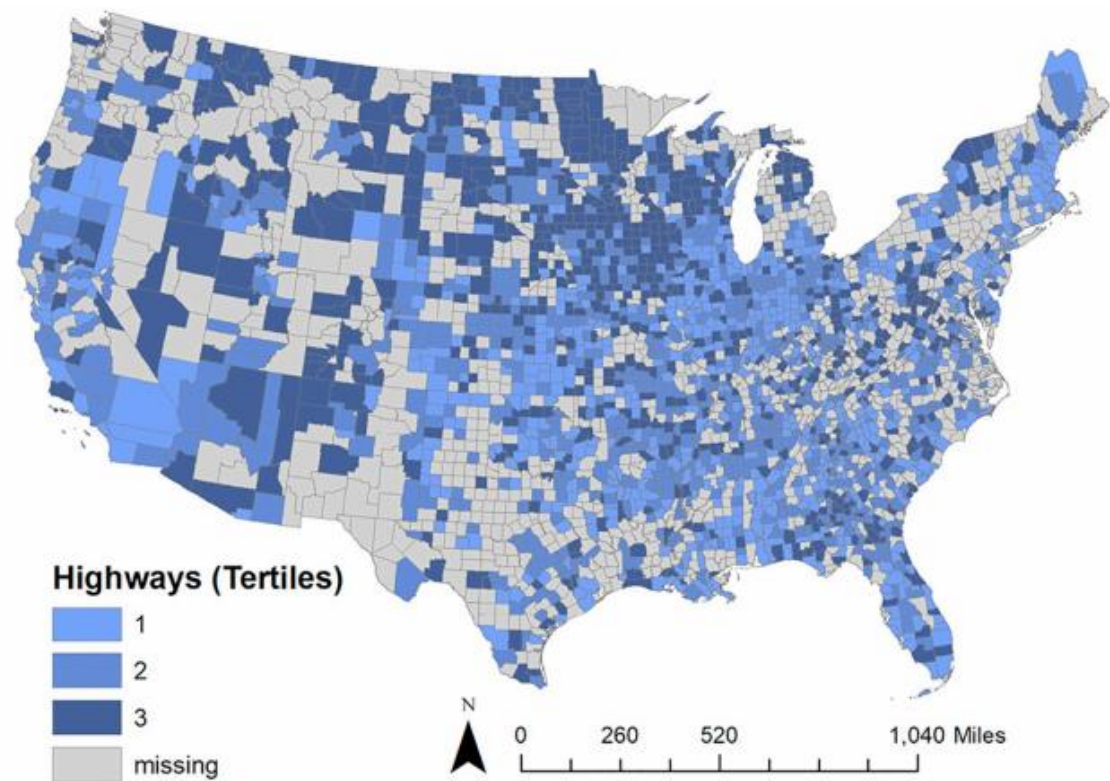
(Javanmardi et al. 2020)

In a similar study, the image feature extractor is jointly finetuned for classification of binary features as well as estimation of obesity prevalence.

This approach explains up to 70.40% variation in prevalence as compared to the previous approach which loses information when aggregating over all locations in a census tract and explains only 6%.

Scaling to entire country

- Following pilot studies, these approaches were scaled to locations in the entire country for more built environment features.
- Relied on Google Vision API to gather labels.
 - Approximately 15 days to label 16 million Google Street View images
- Built environment features:
 - Presence of highland
 - Rural vs. Urban area
 - Presence of grasslands



Main Outcomes

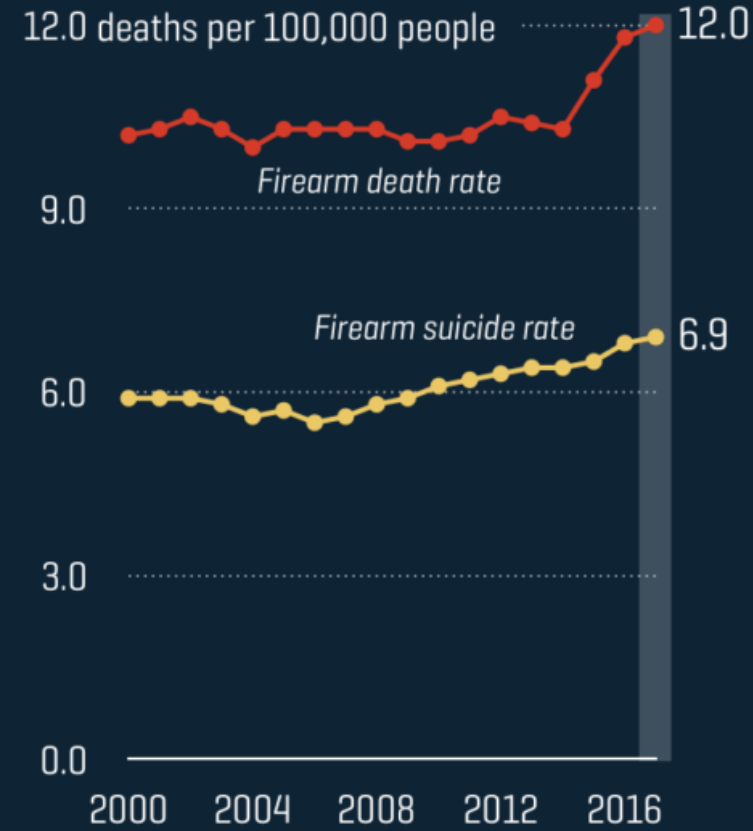
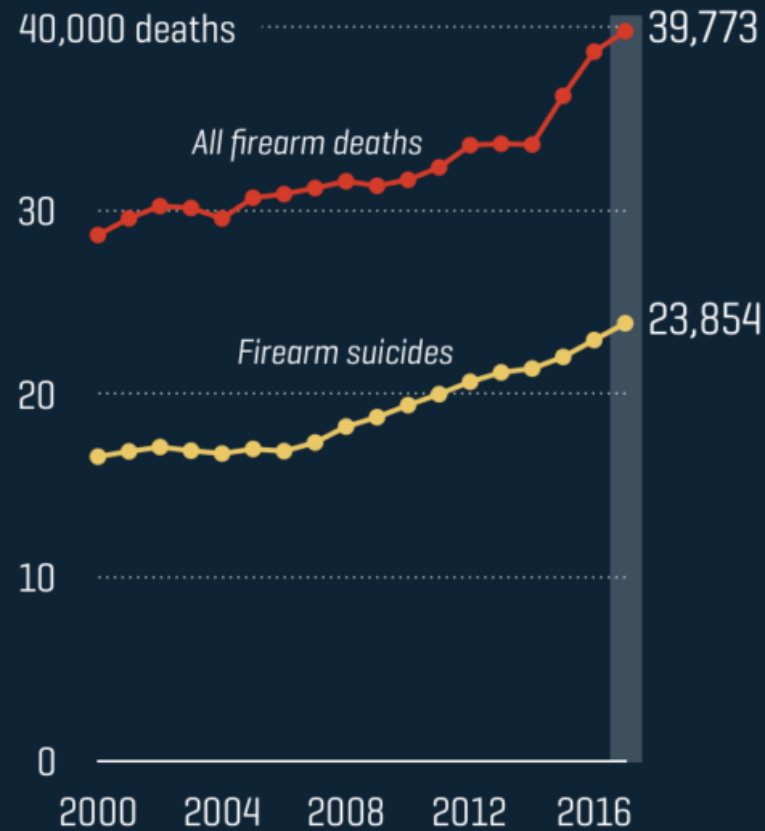
- Economic disadvantage was related to fewer highways, more rural areas, and fewer grasslands.
- Presence of highways was beneficial for self-rated health, diabetes, premature mortality, physical distress, mental distress, physical inactivity, and teen births but was non-significant for obesity.
- Counties with higher percentages of rural areas had worse health in terms of higher obesity, diabetes, fair/poor self-rated health, premature mortality, physical distress, physical inactivity and teen birth rates but had lower rates of excessive drinking.

Firearm Violence: Case-control Study using Satellite Images

(Jay et al. 2020)

U.S. GUN DEATHS CONTINUED TO CLIMB IN 2017

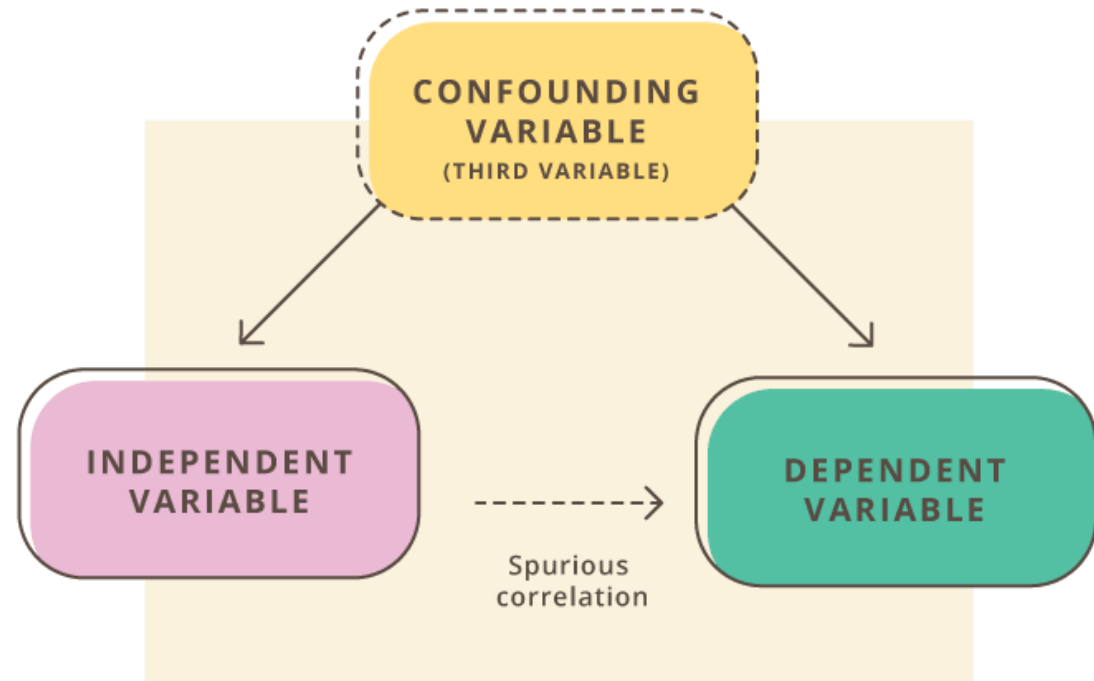
The United States saw nearly 40,000 firearm fatalities last year, an uptick driven largely by an increase in suicides.



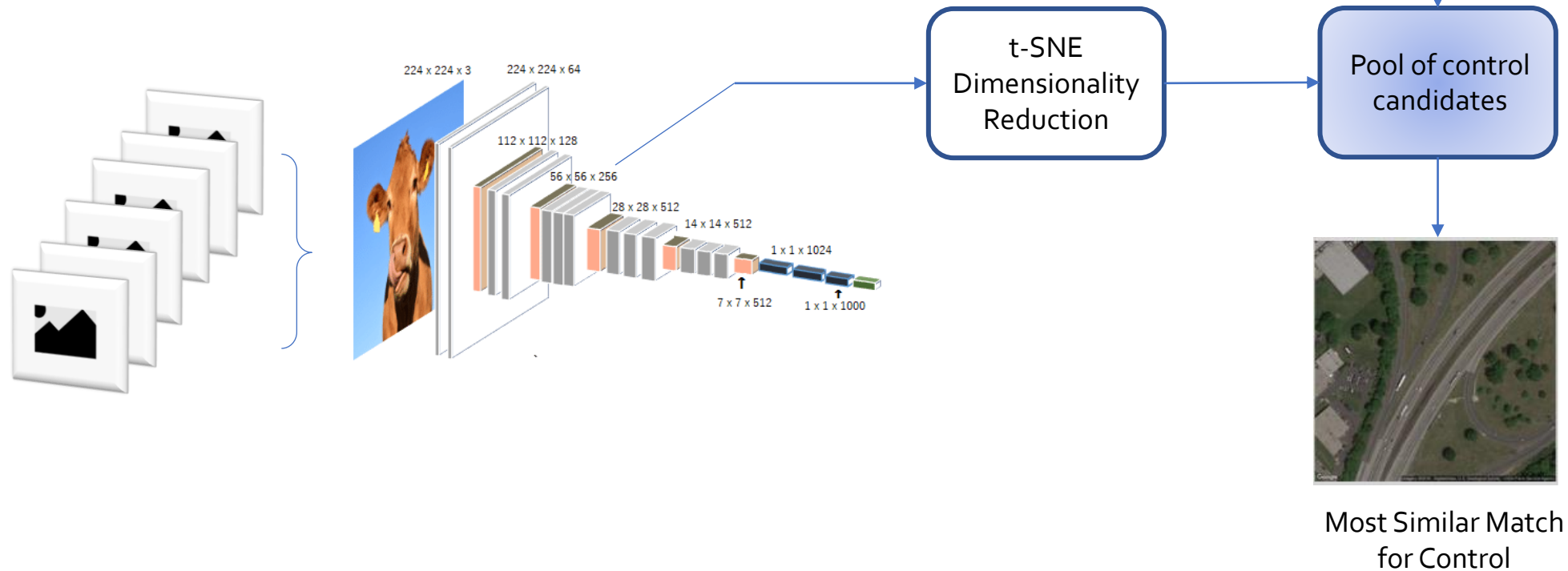
Source: CDC WONDER Underlying Cause of Death counts and age-adjusted rates

Alcohol Outlets and Firearm Violence

- Alcohol outlets (AOs) = Bars, Restaurants and Beer stores
- Many studies have found positive association between the density of AOs and rates of violence at the neighborhood level
- Place/Built environment is a potential confounder.
- Researchers must control for potential confounders.



Matching Cases and Controls on Visual Appearance



Index case

Matched control



Map of Case (n=1609) and Control (n=1609) locations

Results

Table 1 Characteristics of case locations, matched controls and unmatched locations

Variable	Mean (SD*)		
	Cases (n=1609)	Controls (n=1609)	Unmatched (n=21 190)
Vacant lots (n)	2.5 (4.5)	2.1 (4.5)	1.1 (3.3)
Street trees (n)	3.6 (4.9)	4.0 (5.2)	4.6 (6.1)
Commercial land use (parcels/block)			
<i>Same block</i>	2.2 (4.9)	1.4 (2.8)	0.9 (2.7)
<i>Within one block, mean</i>	2.1 (2.3)	1.8 (2.3)	1.5 (2.3)
<i>Within two blocks, mean</i>	1.7 (1.4)	1.5 (1.2)	1.2 (1.2)

*SD reported for continuous variables only.

†Calculated using inverse distance weighting from block group centroids, following Branas *et al.*⁹

‡Calculated as kernel density of drug-related police incident reports, mean-centred and scaled by SD over the full dataset.

- The case–control matching process substantially improved balance on each of the potential confounders, compared with case locations to matched controls than to unmatched units.
- These findings strengthen the argument for a causal association between AOs and violence, even after accounting for differences in the physical and social environment surrounding those institutions.

References

- Salathe, Marcel, et al. "Digital epidemiology." *PLoS Comput Biol* 8.7 (2012): e1002616.
- Maharana, Adyasha, and Elaine Okanyene Nsoesie. "Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity." *JAMA network open* 1.4 (2018): e181535-e181535.
- Maharana, Adyasha, et al. "Detecting reports of unsafe foods in consumer product reviews." *JAMIA open* 2.3 (2019): 330-338.
- Javanmardi, Mehran, et al. "Analyzing associations between chronic disease prevalence and neighborhood quality through Google Street View images." *IEEE Access* 8 (2019): 6407-6416.
- Jay, Jonathan. "Alcohol outlets and firearm violence: a place-based case–control study using satellite imagery and machine learning." *Injury prevention* 26.1 (2020): 61-66.
- Nguyen, Quynh C., et al. "Neighbourhood looking glass: 360° automated characterisation of the built environment for neighbourhood effects research." *J Epidemiol Community Health* 72.3 (2018): 260-266.

Thank You



Original Investigation | Health Informatics

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

Adyasha Maharana, MS; Elaine Okanyene Nsoesie, PhD

Abstract

IMPORTANCE More than one-third of the adult population in the United States is obese. Obesity has been linked to factors such as genetics, diet, physical activity, and the environment. However, evidence indicating associations between the built environment and obesity has varied across studies and geographical contexts.

OBJECTIVE To propose an approach for consistent measurement of the features of the built environment (ie, both natural and modified elements of the physical environment) and its association with obesity prevalence to allow for comparison across studies.

DESIGN The cross-sectional study was conducted from February 14 through October 31, 2017. A convolutional neural network, a deep learning approach, was applied to approximately 150 000 high-resolution satellite images from Google Static Maps API (application programming interface) to extract features of the built environment in Los Angeles, California; Memphis, Tennessee; San Antonio, Texas; and Seattle (representing Seattle, Tacoma, and Bellevue), Washington. Data on adult

Key Points

Question How can convolutional neural networks assist in the study of the association between the built environment and obesity prevalence?

Findings In this cross-sectional modeling study of 4 US urban areas, extraction of built environment (ie, both natural and modified elements of the physical environment) information from images using convolutional neural networks and use of that information to assess associations between the built environment and obesity prevalence showed that physical characteristics of a