Challenges and Opportunities in Analytics for Human Services

Predictive Analytics for Human Services and Education
National Academies Expert Panel Meeting
September 15, 2017

Sallie Keller
Professor of Statistics and Director

Virginia Tech
Biocomplexity Institute

SDAL
SOCIAL & DECISION ANALYTICS LABORATORY

Biocomplexity Institute of Virginia Tech

The study of life and environment as a **complex system**

The Social and Decision Analytics Laboratory brings together statisticians and social and behavioral scientists to embrace today's data revolution, developing evidence-based research and quantitative methods to inform policy decision-making.

VirginiaTech
Biocomplexity Institute

Charge for the Workshop

Discuss the potential and risks of predictive analytics applied to decision making in health, human services, and education.

- Predictive analytics may offer insights into how to:
  - Tailor educational approaches for individual children
  - Flag individuals in need of mental health intervention
  - Allocate and prioritize early intervention services for children with special needs
  - Better protect children from abuse and neglect
  - Prevent criminal activities and associated arrests

- Distinct Areas to be examined:
  - Methodological issues
  - Ethical considerations
  - Legal concerns

Putting this in Context

- Examine how we got here
- What has changed today, technically and socially and that makes this discussion relevant?
- What is meant by predictive modeling?
- What is the road ahead?

"Everything grows out of what precedes it in a line from beginning to end." Will Safron

Annie E. Casey Foundation

- 1948 established to:
  - Build better futures for disadvantaged children and their families in the U.S.
  - Foster public polices, human service reforms, and community supports that more effectively meet the needs of today's vulnerable children and families
- 1983 the new era: *"What is needed is a renewed determination to **think creatively**, to learn from what has succeeded and what has failed, and, perhaps most important, to foster a sense of **common commitment** among all those concerned with the welfare of children."* Jim Casey

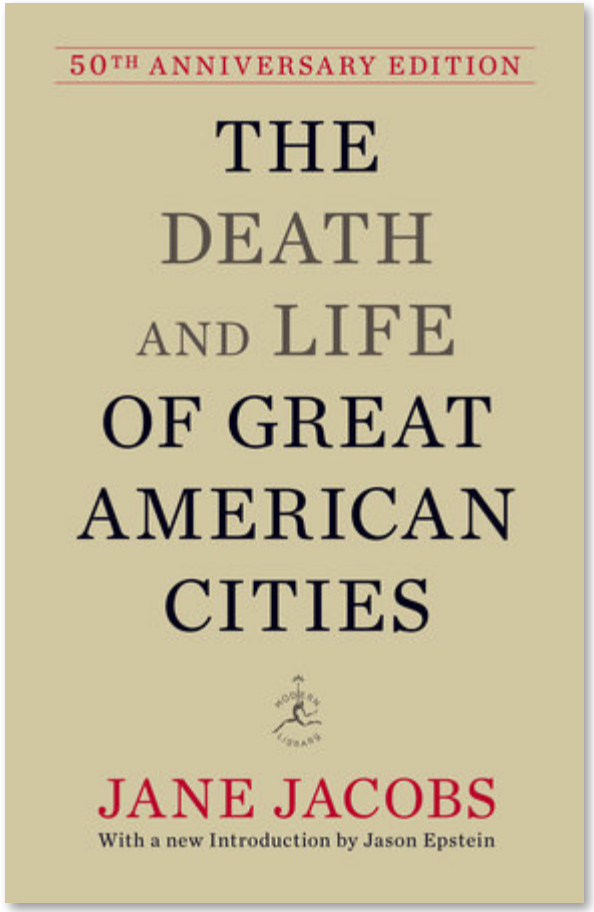
Local Communities Matter

Jane Jacobs, 1961 confronts city planning

- Argued diversity is essential to the lifeblood of a city and that in order to fend off atrophy everyone must have a roll in the decision-making process
- *"Cities have the capability of providing something for everybody, only because, and only when, they **are created by everybody**."*

Her arguments are based on powerful contextual examples and stories, however data and models are absent

Her work provides tremendous insights that have helped set the stage for today's meeting



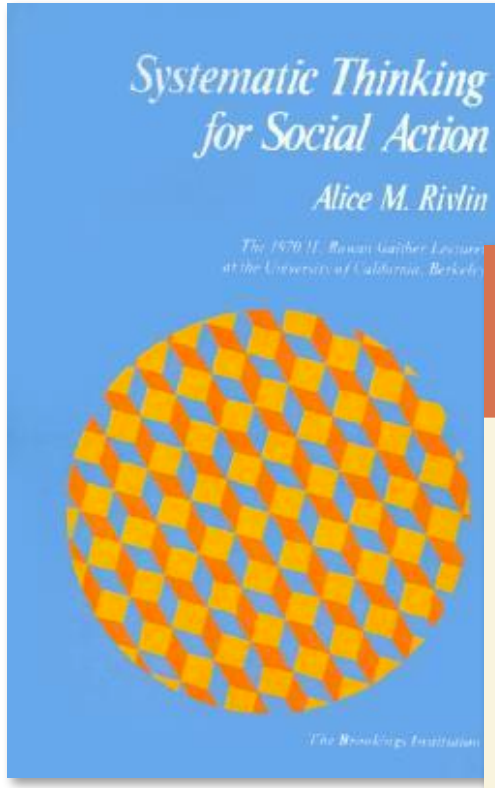
Systematic Analysis of Social Policy

*"Until programs are organized so that **analysts can learn** from them and systematic experimentation is undertaken on a significant scale, prospects seem dim for learning how to produce better social services."* Alice Rivlin

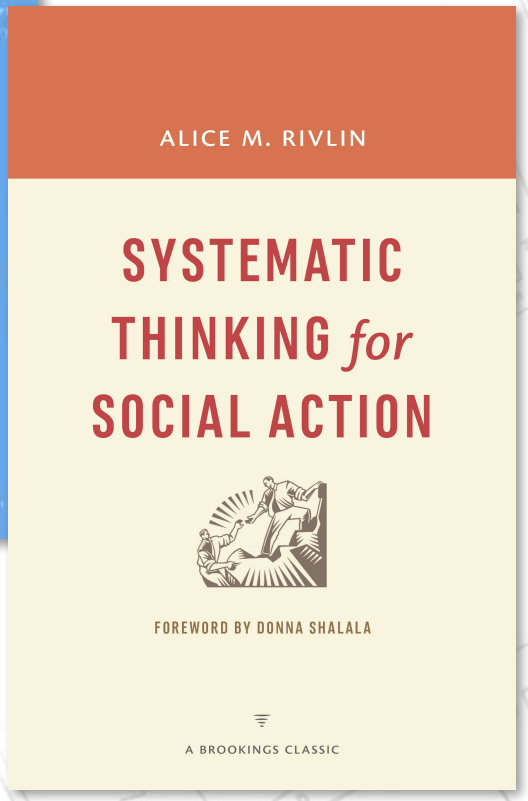
- Define the problem and identify those who suffer from the problem
- State who would win or lose from specific programs and at what costs
- State what programs would do the most good and compare them
- State how these programs can be produced most effectively

"… monstrous gaps in information …."

1970



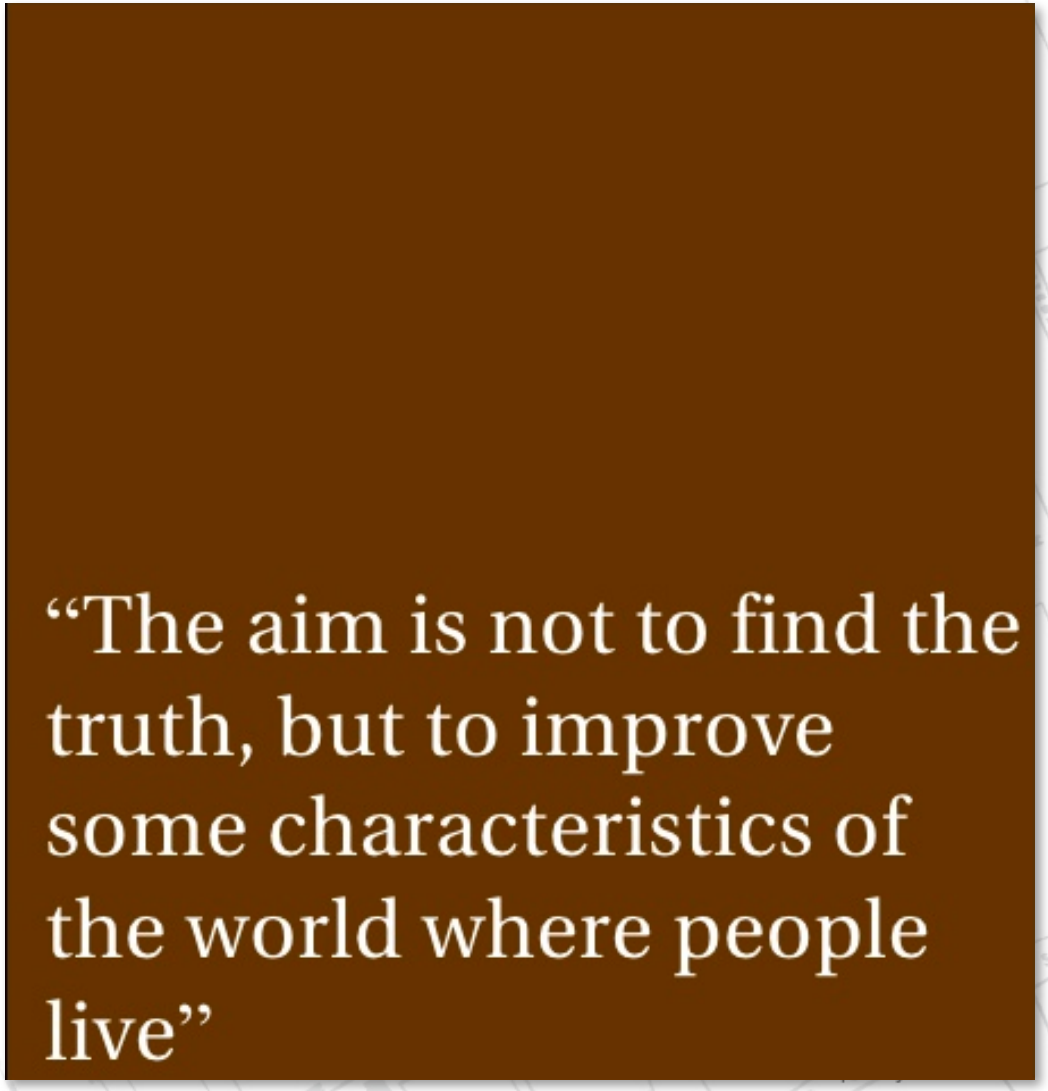
2015



Wicked Problems

"Social problems are never solved. At best they are only resolved – over and over again."

Rittel & Webber, 1973



Fast Forward to 2017

Data are ubiquitous

- Able to fill some of the information gaps

Maturation of:

- Social theory
- Statistical methods
- Mathematical algorithms
- Computational capacity

Filling the Information Gaps: The Science of *ALL* data

Social and Behavioral Data Flows

Infrastructure

- Condition
- Operations
- Resilience
- Sustainability

Environment

- Climate
- Pollution
- Noise
- Flora/ Fauna

People

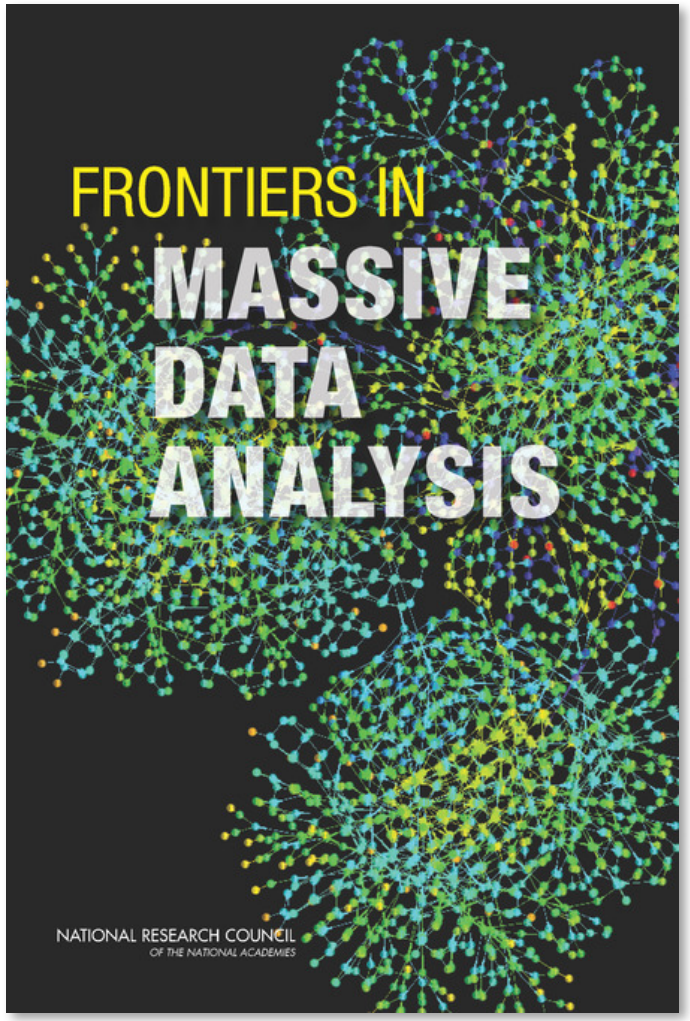
- Relationships
- Location
- Economic Condition
- Communication
- Activities
- Health

VirginiaTech
Biocomplexity Institute

Big Data or Simply *All* Data?

It doesn't matter what it's called,
it only matters what you do with it

- **It is not just about size**
  - Traditional and new sources of data
- **Statistics / analytics**
  - Replication and reproducibility
  - Bias, representativeness, and precision
  - Descriptive, association, causation
  - Change point detection
- **Modeling and simulation**
  - Scenario development
  - **Explanatory analysis vs prediction**

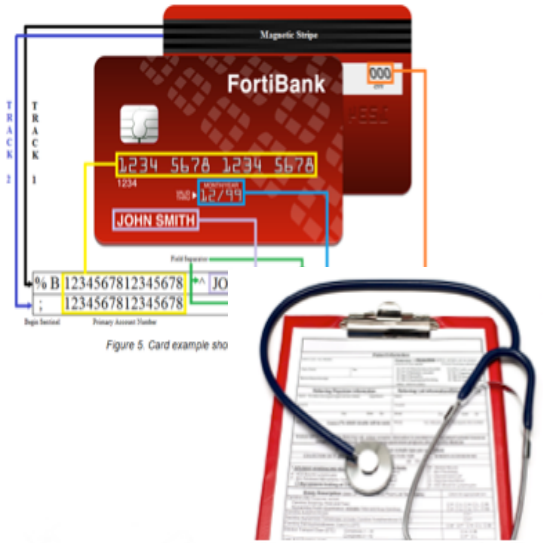


Bring the *ALL* Data Revolution to Communities

State, Federal, Local – Civilian and Defense

Designed Data

Administrative Data

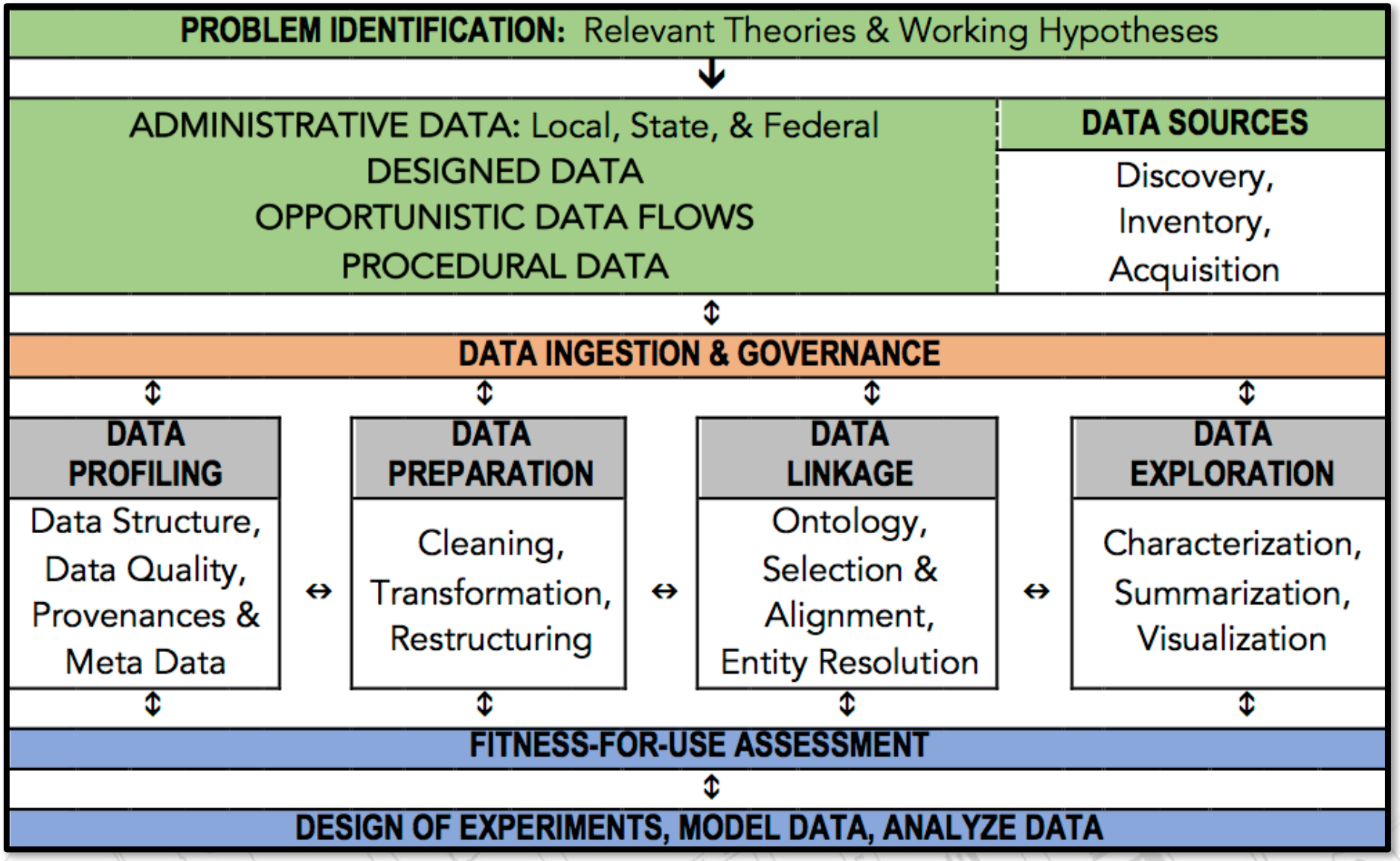


Opportunity Data

Procedural Data

VirginiaTech
Biocomplexity Institute

Repurposing Data needs a Rigorous and Flexible Process : Data Science Framework

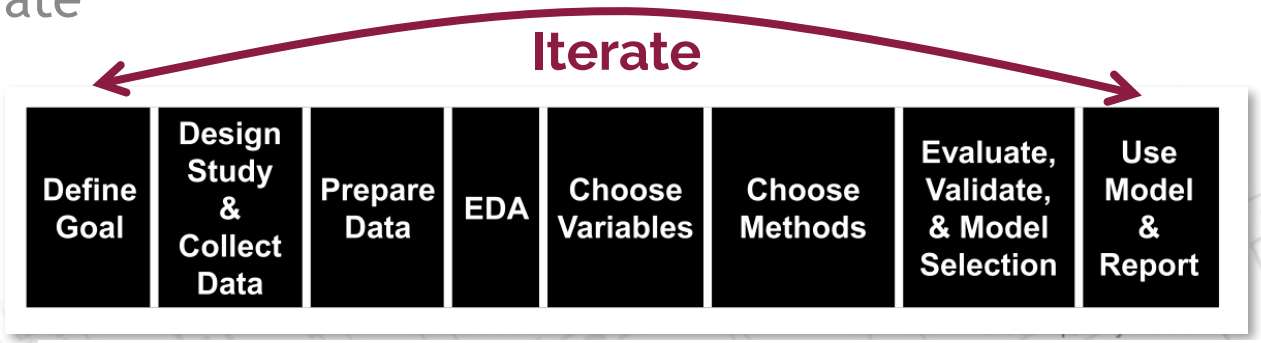


VirginiaTech
Biocomplexity Institute

Today's World of Analytics: Is it more then hype?

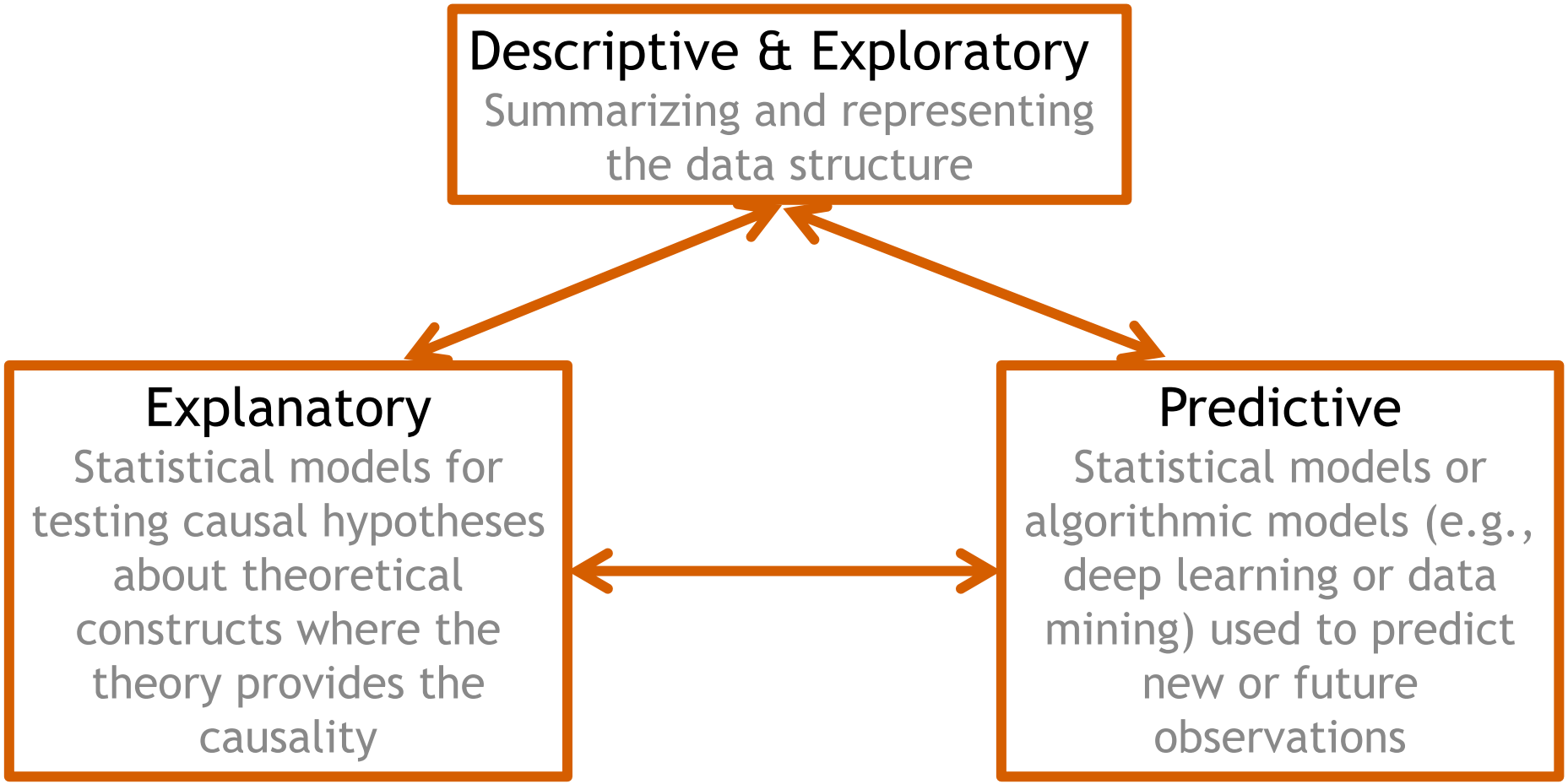
Data and Models

- **Define the problem**
  - Case management? Policy analysis? Programing planning and development?
- Problem definition and lessons learned from the literature drives the **data discovery and acquisition** processes
- Problem type drives **data use concerns**
  - **Ethical biases** due to data availability and acquisition constraints
  - **Privacy** refers to the amount of personal information individuals allow others to access about themselves
  - **Confidentiality** the process that data producers and researchers follow to keep individuals' data private
  - **Security** applies to data storage and transport
- Problem **goals define** type of relevant analyses



Shmueli, Galit. "To explain or to predict?." *Statistical science* 25.3 (2010): 289-310.

Types of Modelling (aka Analytics)



Example: Descriptive → Explanatory Analyses

Metrobus Fare Evasion Project

Problem:

WMATA loses approximately 10-20 million dollars a year due to bus fare evasion on its 1300-1500 daily trips

Research Goal:

Provide insights into the problem of bus fare evasion that can be used to guide fare evasion interventions

Descriptive Analysis Plan:

Use WMATA **administrative data** to locate where fare evaders live and the **American Community Survey** to tell their story at the census block group level

General Observations

- Fare evasion is a **widespread** international problem
- Fare evasion is estimated using **observer surveys** and not using administrative data
- **Issues inherent** in observer surveys include: high costs, missed assignments, difficulty processing large passenger volumes, data interpretation issues, data entry and analysis costs, and potential data collection inconsistencies between observers
- The fare evasion estimates from surveys do not include an estimate of the **variability**

WMATA Administrative Data Sources

Data Sources for the first week of May (5/1 – 5/7/2017):

Bus Stops (10,988 observations)

Contains the stop ID, stop name, latitude, and longitude

Approximate Person Counter (APC)

Contains front & back door entries and exits for a bus, route, trip number and bus stop

Farebox

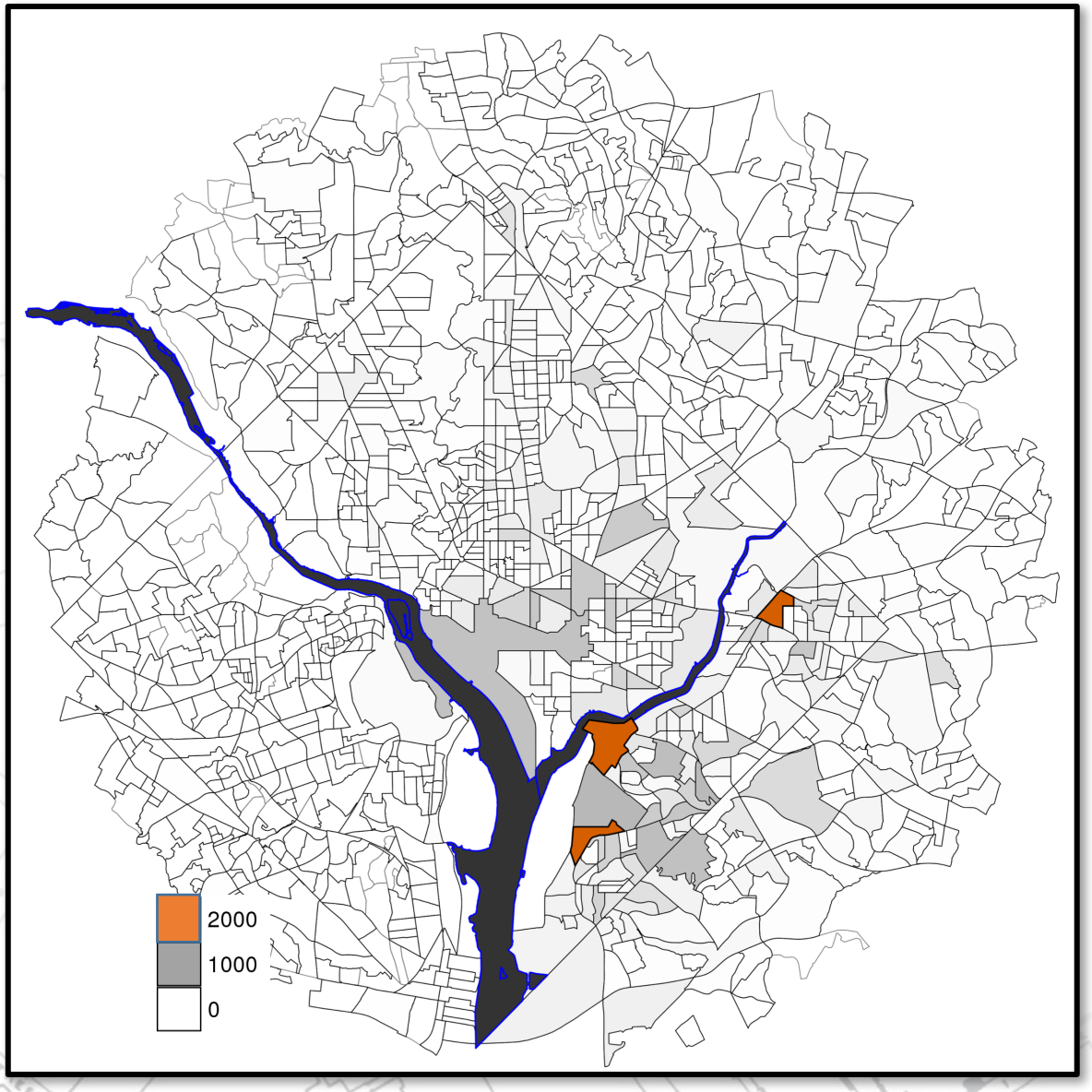
Contains cash & SmarTrip transactions for a bus, trip number, & bus stop

Data Issues

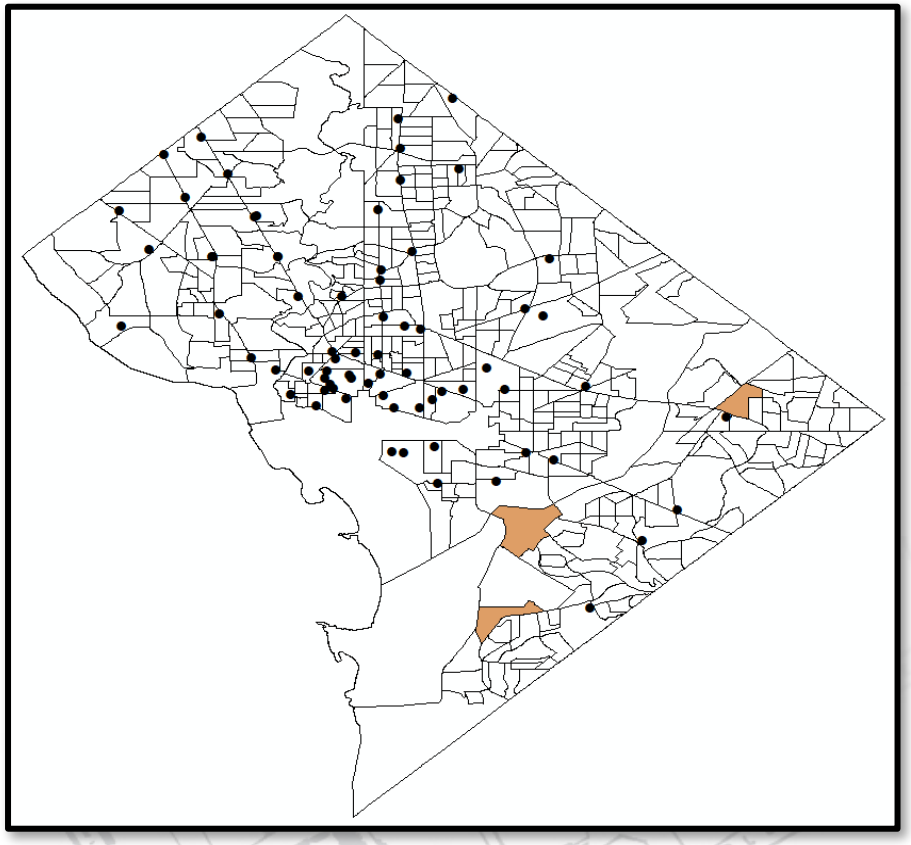
Imprecise latitude and longitude coordinates, missing or mislabeled bus routes and stop IDs in Farebox and APC data, missing trip numbers in Farebox data

Data	Uncleaned	Cleaned	Monday – Friday
APC	3,793,655	3,791,332	3,105,623
Farebox	2,729,668	2,060,055	1,751,335

Fare Evasion in Evenings 2-8 pm



- Locations to add money to SmartTrip cards in DC
- Few and far between in Anacostia
- Could this contribute to fare evasion?



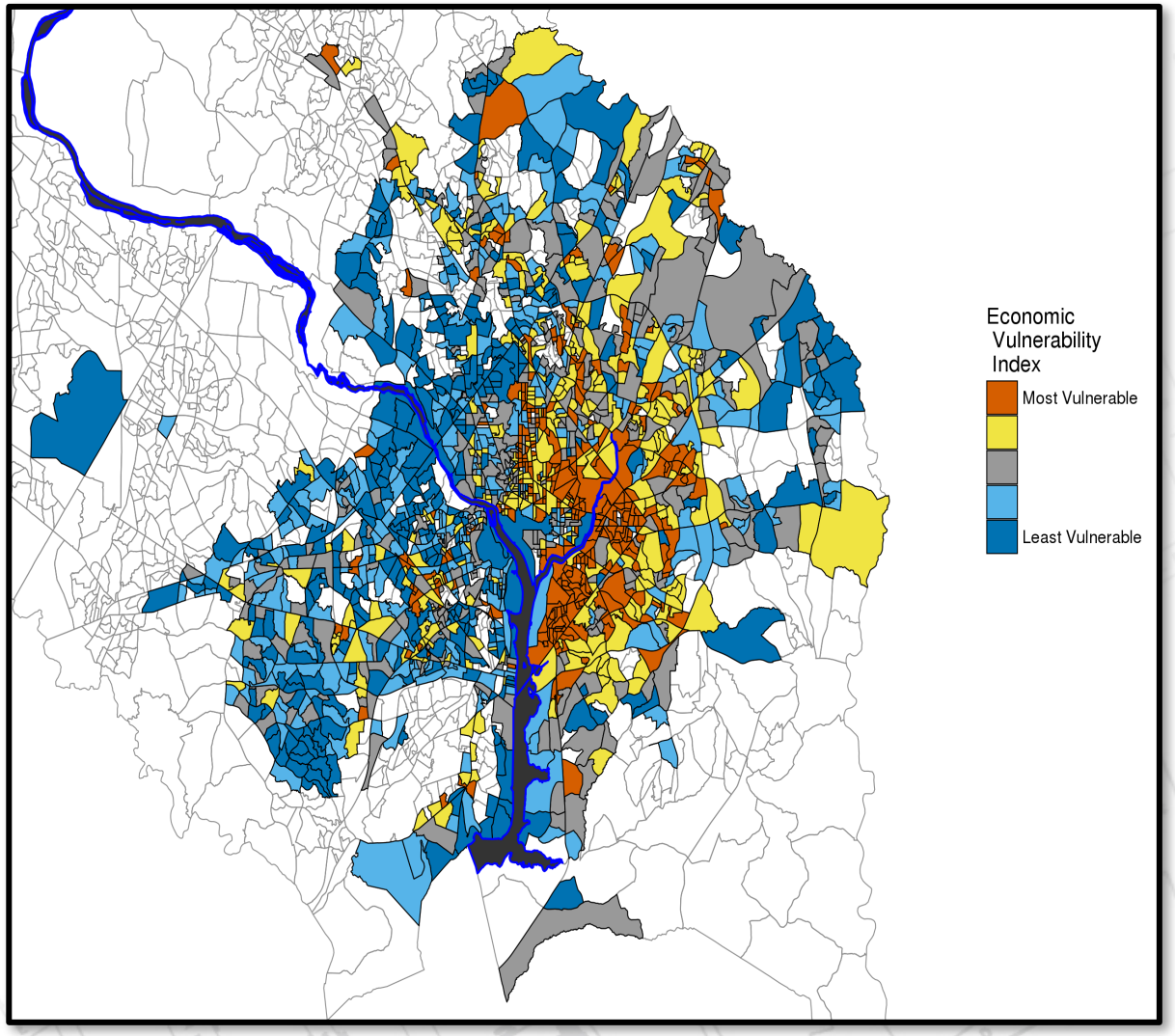
Vulnerability Index: Census Block Groups

Economic Vulnerability Index

Plot of an composite economic vulnerability index by census block groups with bus stops in the seven WMATA jurisdictions

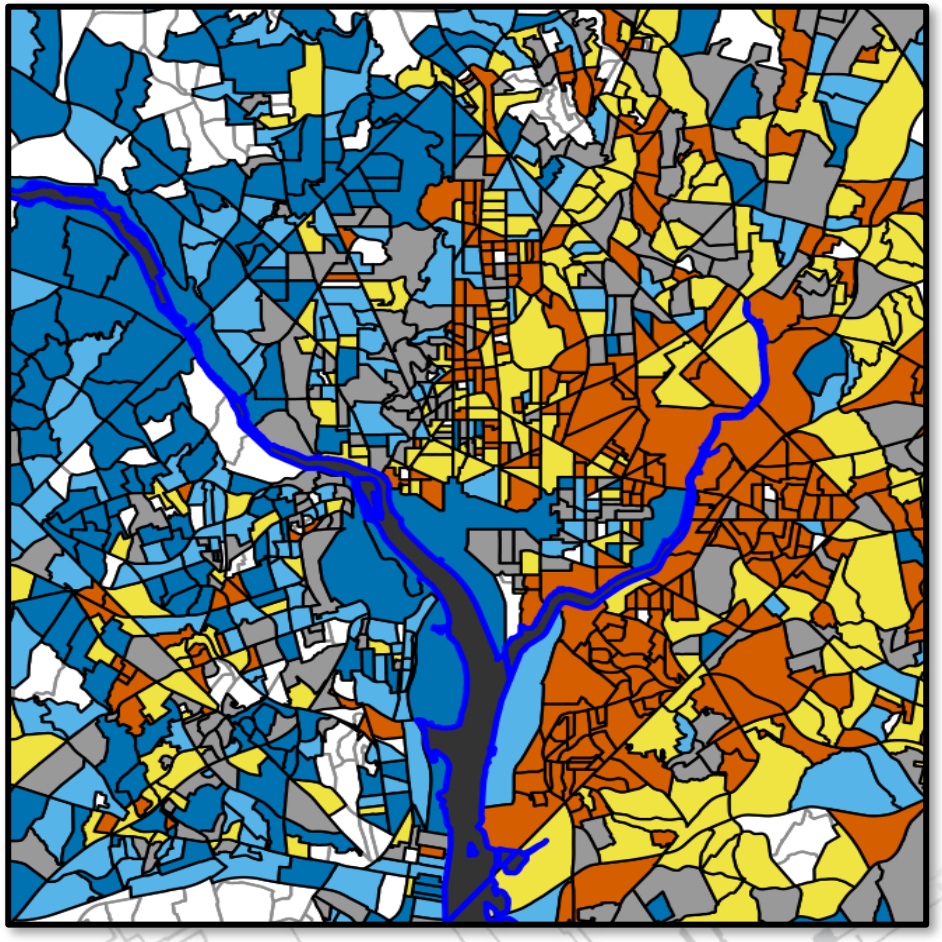
The composite index was constructed using ACS (2015) variables:

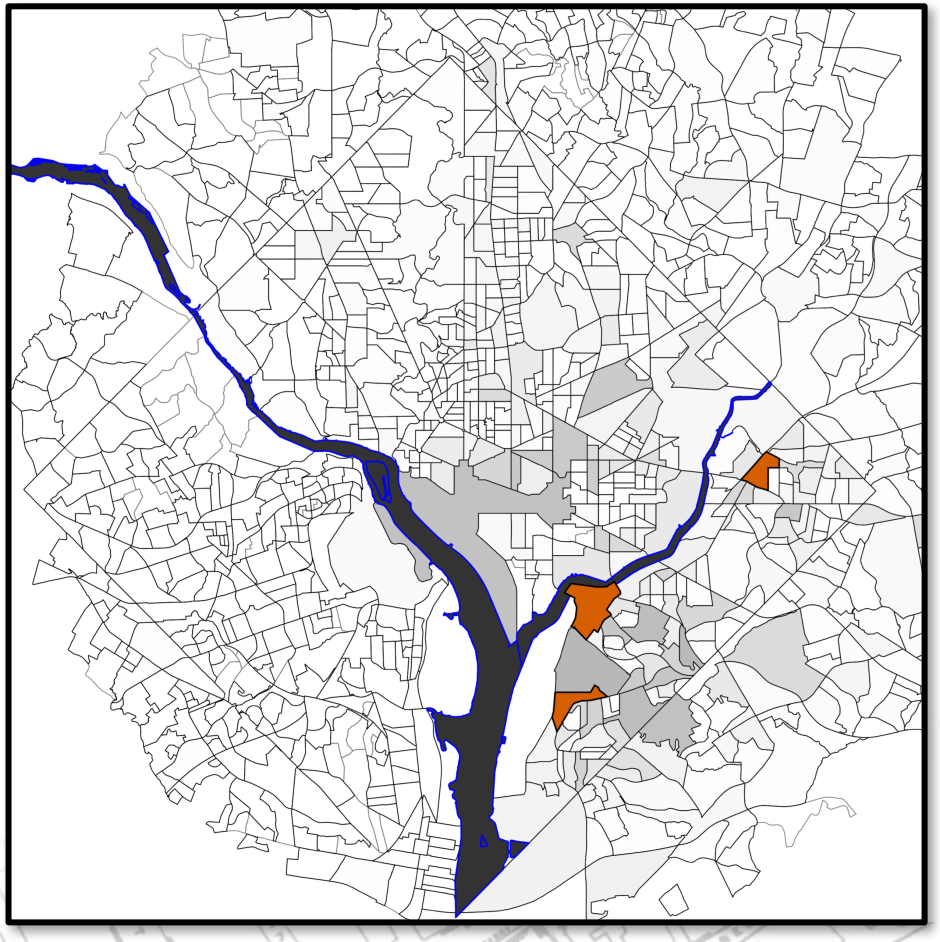
- % households in poverty (Federal)
- % households with no vehicle
- % households qualifying for SNAP
- % households with housing burden > 50%



Insights: Hypotheses and Interventions to Test

Not all economically vulnerable Census Block Groups have high numbers of fare evaders, but all Census Block Groups with a high numbers of fare evaders are economically vulnerable.





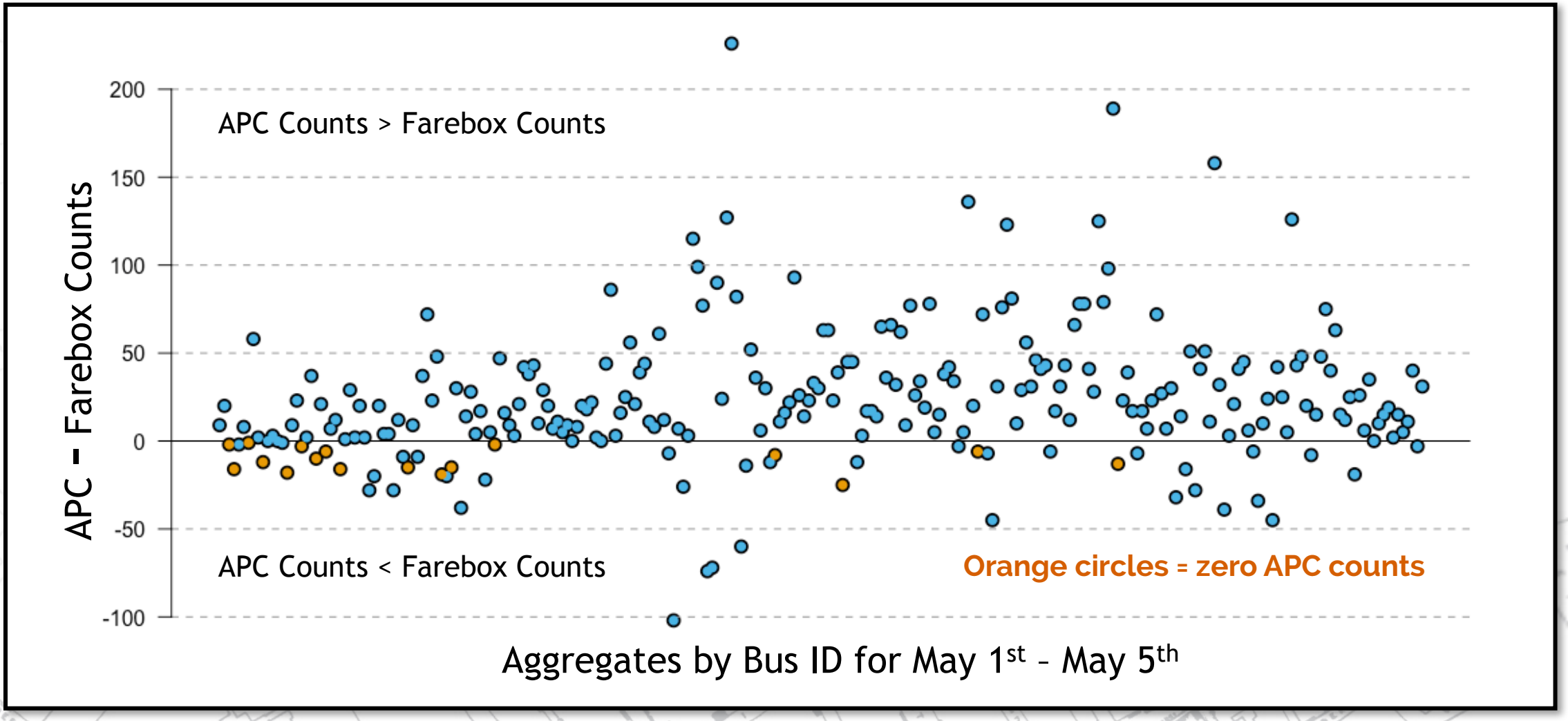
Descriptive Analyses: Allow strategies and hypotheses to Present Themselves

Potential Interventions that could be tested through statistically designed studies:

- Reduce barriers to paying fares
- Turn criminals into customers with low-income fare products
- Make fare evasion harder and enforcement easier
- Change narrative about fare evasion – address cultural reactions to paying full fare through outreach and marketing campaigns

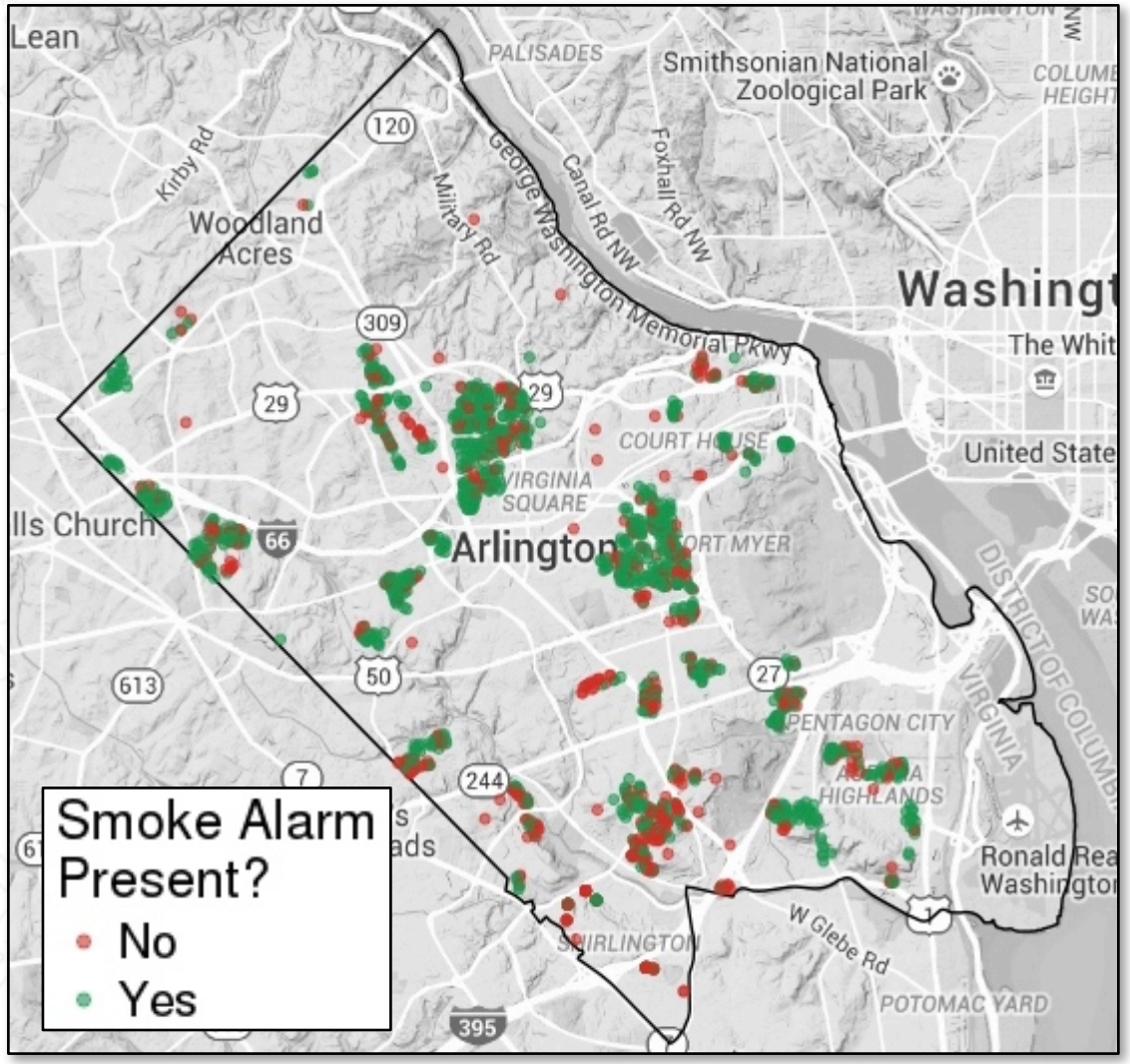
Cautionary Tale: Need Accurate Estimates

Explanatory Models: For studies to test hypotheses or experiments for interventions, need to know how accurately fare evasion can be estimated – what effect sizes can be measured!



Example: Predictive → Program Planning

Arlington Operation FireSafe



Issue: The Fire Department wants to improve the efficiency of their Operation FireSafe program

- Out of 5,623 visits to single family homes only 1,799 had an adequate number of working smoke detectors

Goal: Construct models to predict for each single family home (and each block group) the probability it has adequate smoke detectors

VirginiaTech
Biocomplexity Institute

The DATA

Household Level (*Administrative Data*): Operation FireSafe data on the location of the 5,623 single family homes visited, time and date of the visit, and the outcome

Household Level (*Administrative Data*): Real estate tax assessments for 60,343 single family homes which includes tenure, home age, value, size, and number of bedrooms

Household Level (*Opportunity Data*): Geocoded the single family home locations

Census Tract Block Group Level (*Designed Data*): 5-year 2015 American Community Survey – household level demographic and socioeconomic data

Predicting the Probability a single family home has adequate smoke detection

Bayesian logistic regression model with conditionally auto-regressive spatial effects

- Predict probability of having adequate working smoke alarms
- Block group and household predictions
- Able to Identify potential drivers of smoke alarm needs

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i) = \mu_i + \phi_i$$

$$\mu_i = \mathbf{X}_i^T \boldsymbol{\beta}$$

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{I}_{1000})$$

$$\phi_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\theta))$$

$$\phi_i \setminus \boldsymbol{\phi}_{-i}, \boldsymbol{W}, \tau^2, \rho \sim N\left(\frac{\rho \sum_{k=1}^{K} \omega_{ik}\phi_k}{\rho \sum_{k=1}^{K} \omega_{ik} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{k=1}^{K} \omega_{ik} + 1 - \rho}\right)$$

$$\tau^2 \sim Inverse\ Gamma(5, 40)$$

$$\rho \sim Uniform(0, 1)$$

$Y_i = 0$, not adequate coverage

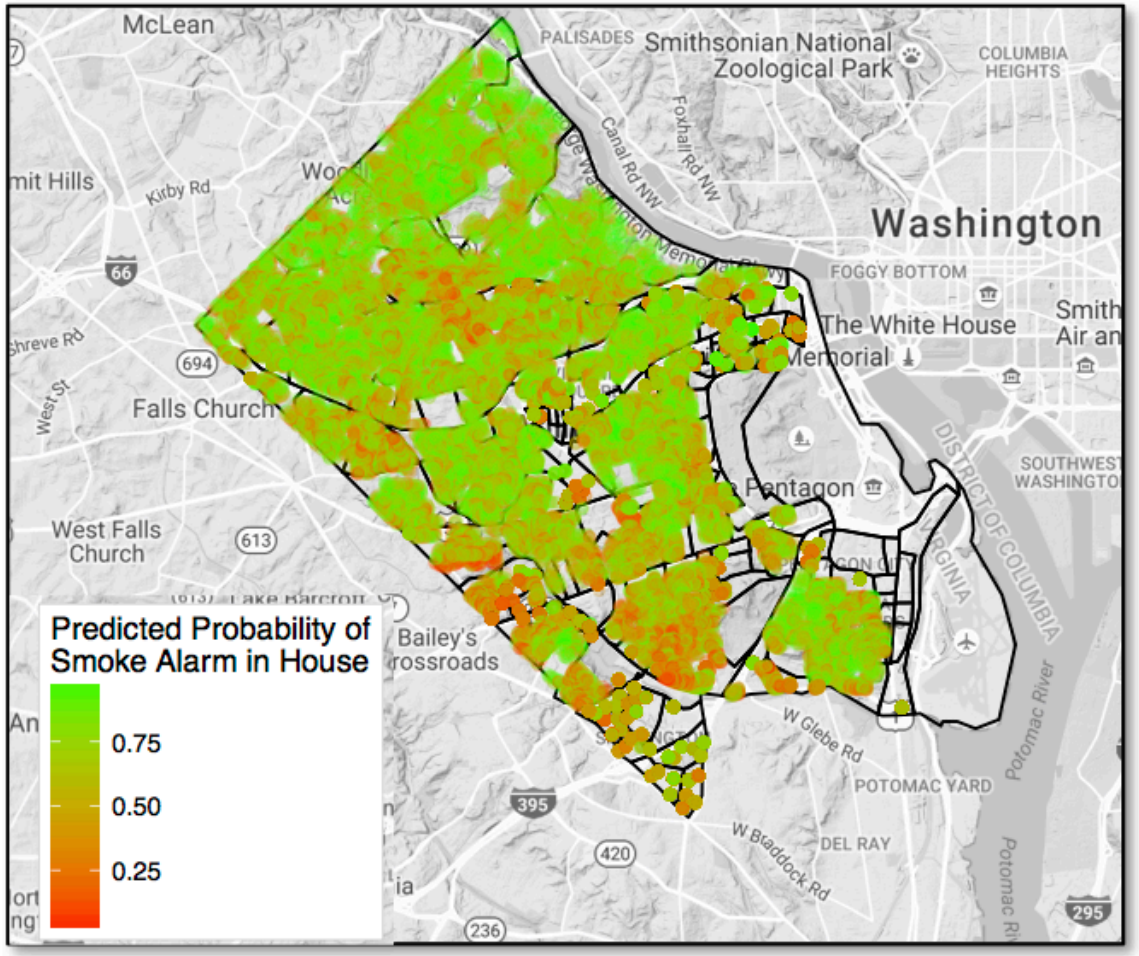
$Y_i = 1$, adequate working smoke alarm(s) present

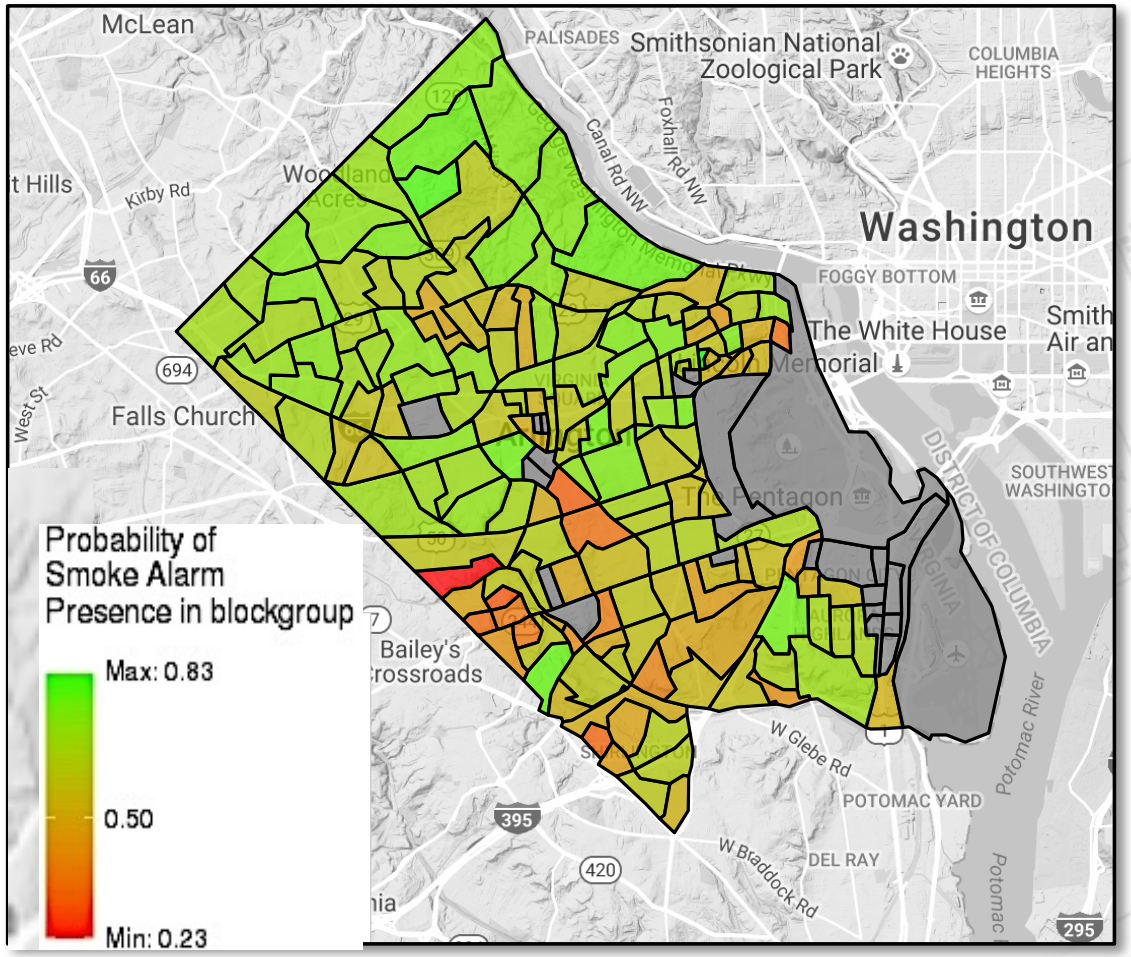
Prediction Uncertainty: Precision/Recall Area Under the Curve (AUC)

VirginiaTech Biocomplexity Institute

Probability of Home Smoke Detection

Bayesian logistic regression model with conditionally autoregressive spatial effects





Housing Unit Level Predictions

Census Block Level Predictions

The Road Ahead

Data and Models hold Great Promise, but …

Data, analytics, and data scientists alone will not make the difference

Today's **wicked problems** need *transdisciplinary* teams

- **Individuals** living the experience
- **Community** with mutual concerns
- **Subject Matter** Experts to include both researchers and professionals, technical and programmatic
- **Organizations** that can take action, create precedents, influence policies and regulations
- **Creative Thinking** from all members of the team to find new and unexpected solutions and supporting data and models

Use of data and models must be **transparent!**

Discussion