Building Resilient Communities: Harnessing the Power of Data

Statistics and Data Science Symposium, May 2018

Sallie Keller
Professor of Statistics and Director
sallie41@vt.edu

BIOCOMPLEXITY INSTITUTE
VIRGINIA TECH

SOCIAL & DECISION ANALYTICS LABORATORY

Biocomplexity Institute of Virginia Tech

- The study of life and environment as a **complex system**
- Understanding biology **in the context of** ecosystems and human-created systems
- **Transdisciplinary** team science

"From molecules to policy"

Problem-Driven Science

Our information biology approach is putting research to work in the real world, breaking down barriers between science and policy.

Social and Decision Analytics Lab

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.

- **Science of *ALL* Data**
- **Community Learning Data Driven Discovery**
  - Defense analytics
  - Education and Labor Force Analytics
  - Health and Well Being Analytics
  - Emergency Management Analytics
  - Industrial Innovation Analytics
- **Information Diffusion Analytics**

Science of *ALL* data is a team sport!

Thanks to my team

- Stephanie Shipp
- Kim Lyman
- Gizem Korkmaz
- Aaron Schroeder
- Bianica Pires
- Dave Higdon
- Joy Tobin
- Vicki Lancaster
- Joshua Goldstein
- Daniel Chen
- Lori Conerly
- Ian Crandell
- Brian Goode
- Cathie Woteki

The Science of *ALL* data

Why Now?

ALL data revolution – new lens for social observing

Infrastructure

- Condition
- Operations
- Resilience
- Sustainability

Environment

- Climate
- Pollution
- Noise
- Flora/ Fauna

People

- Relationships
- Location
- Economic Condition
- Communication
- Health

S. Keller, and S. Shipp. (Forthcoming) "Building Resilient Cities: Harnessing the Power of Urban Analytics" in *The Resilience Challenge: Looking at Resilience through Multiple Lens*, Charles C Thomas Ltd Publishers

Gaining insights through *ALL* data sources

Local, State/Provence, and Federal

Designed Data

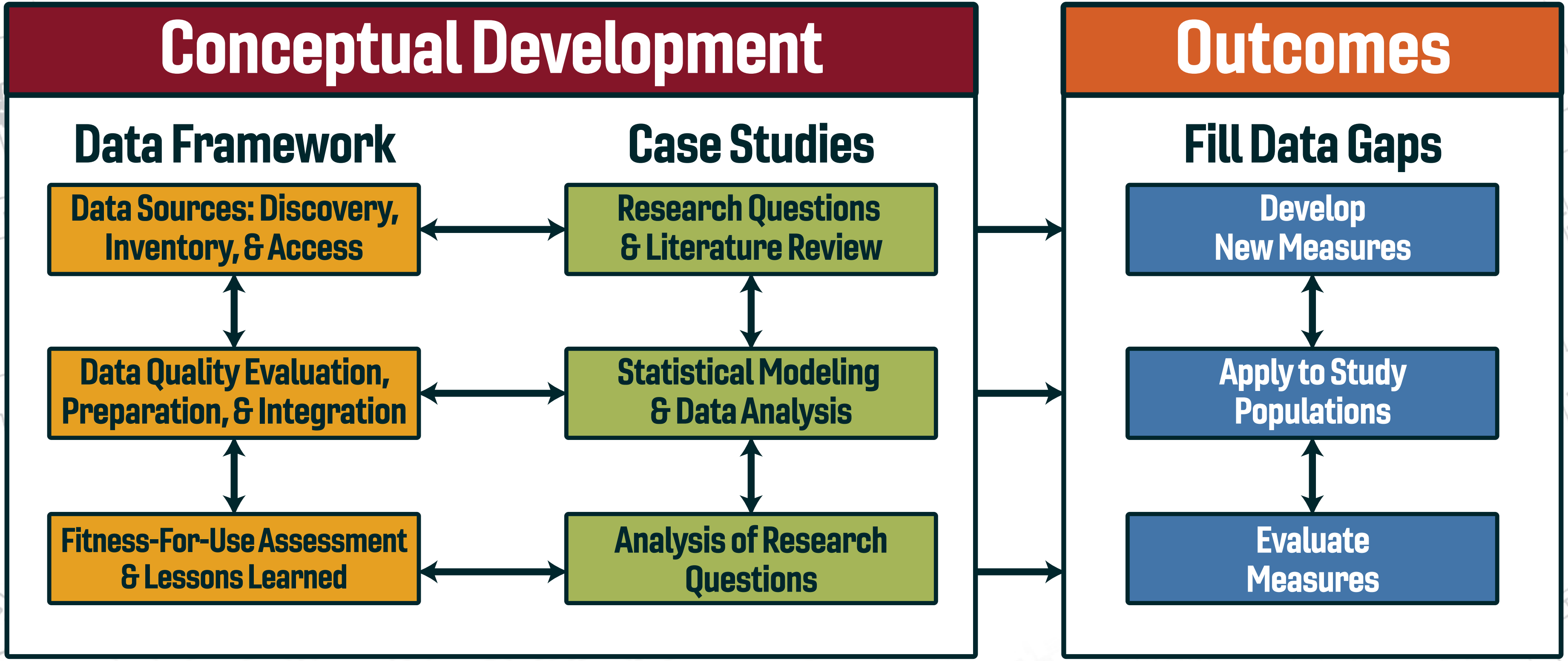
Administrative Data

Opportunity Data

Procedural Data

Keller SA, Shipp S, Schroeder A. (2017). *Does Big Data Change the Privacy Landscape? A Review of the Issues. Annual Reviews of Statistics and its Applications*; 3:161-180.

Our *Science of All Data* research model



Case Studies
Policy focused other people's problems (OPPs)

Local / State Government

Arlington County, Virginia
Fairfax County, Virginia
State Higher Education Council of Virginia
Virginia Department of Emergency Management

Federal Statistical Agencies

U.S. Census Bureau
Housing and Urban Development
National Science Foundation
National Center for Science and Engineering Statistics

Department of Defense

U.S. Army Research Institute
Defense Manpower Data Center
Minerva Research Initiative

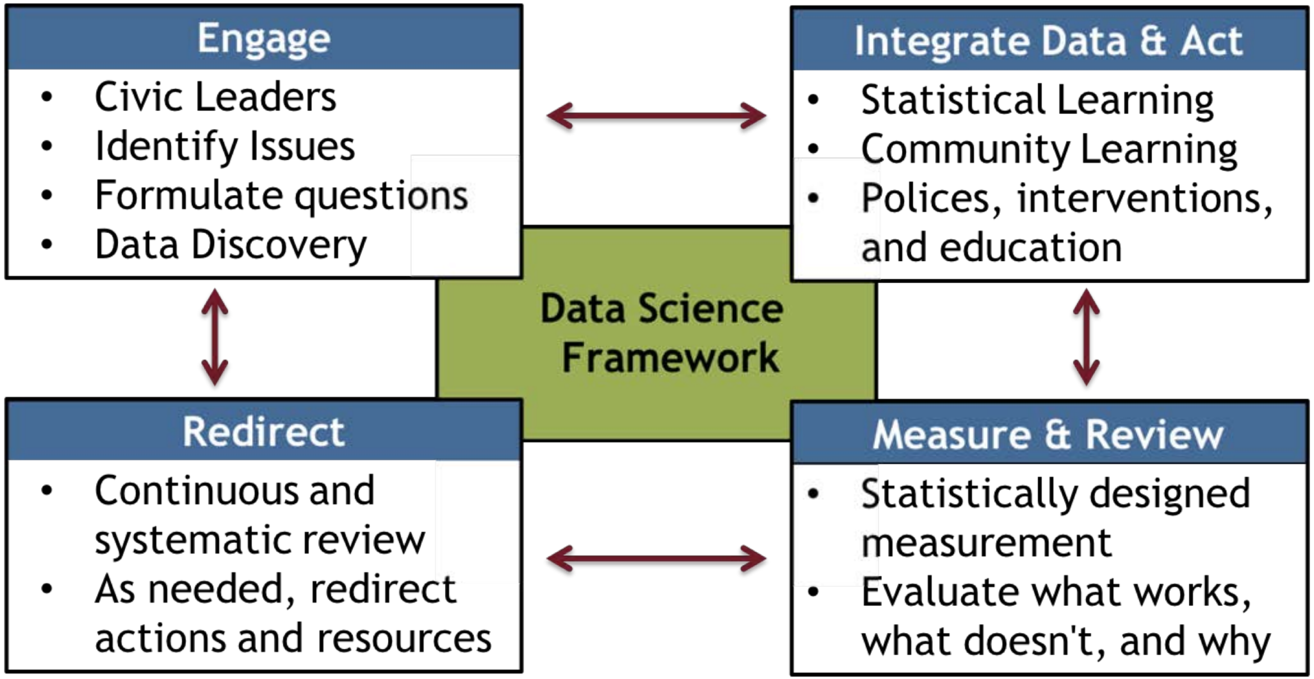
Industry

MITRE Corporation
Proctor & Gamble

Enhancing Prosperity through Data Science

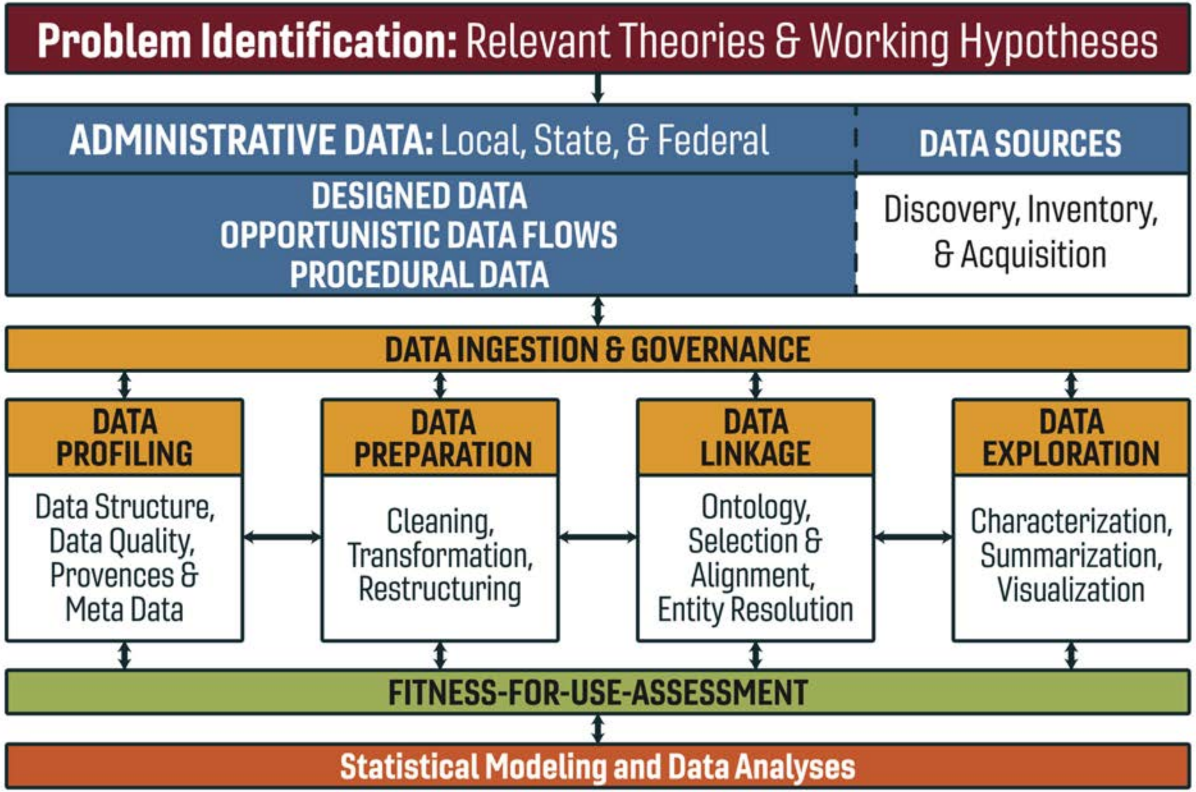
Translating our research model:

Community Learning through Data-Driven Discovery



Keller, S., Lancaster, V., & Shipp, S. (2017). Building capacity for data-driven governance: Creating a new foundation for democracy. *Statistics and Public Policy*, 1-11.

Our emerging *Data Science Framework*



Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Reviews of Statistics and its Applications*, 4:85-108.

Key community-based research issues

- Locating and describing a population
- Estimating a statistic and a measure of its variability to evaluate its usefulness for the purpose at hand
- Forecasting future needs
- Evaluating a program, policy, or standard operating procedure

All of this needs to align with spatial scales that matter for decision-making e.g., sub-county/city geographies

S. Keller, S. Shipp, G. Korkmaz, E. Molfino, J. Goldstein, V. Lancaster, B. Pires, D. Higdon, D. Chen, A. Schroeder, 2018. Harnessing the power of data to support community-based research. *WIREs Comp Stat* 2018. doi: 10.1002/wics.1426

Data science innovations to develop *sub-county/city* data-driven insights

- **Synthetic population technology** -**statistically** align data to relevant geographic boundaries
- **Capture housing stock and place-based data** – geocode places and housing units, both owned and rented
- **New sources of data** – obtain local administrative data and local web-scraped data
- **Vulnerability Composite Indicators** – statistically integrate data
- Exploring the data using **visualization tools**

S. Keller, S. Shipp, G. Korkmaz, E. Molfino, J. Goldstein, V. Lancaster, B. Pires, D. Higdon, D. Chen, A. Schroeder, 2018. Harnessing the power of data to support community-based research. *WIREs Comp Stat* 2018. doi: 10.1002/wics.1426

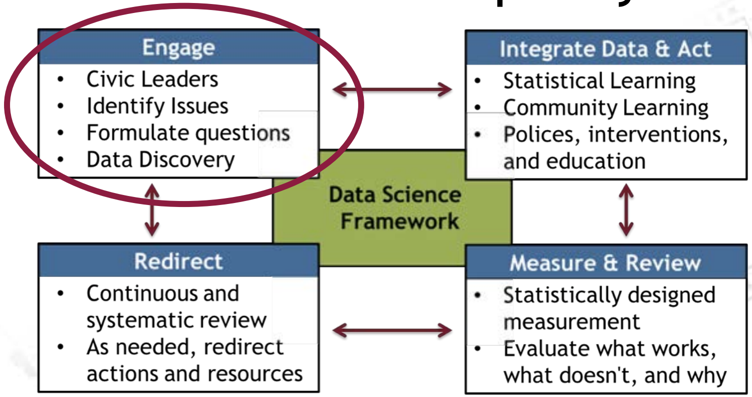
Engagement, Issues, & Questions

Overarching Goal: Develop data-driven insights on current issues and build forecasts to inform future issues

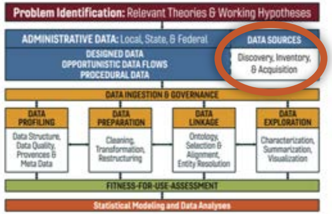
- Expand Fairfax County's capacity to access and integrate county, state, and federal data in useful ways to address critical problems

Project Focus: Identify the trends in obesity and activities related to obesity across geographies of interest for local policy and program development

- Focus on determinants identified in the literature related to obesity – the built environment, nutrition, physical activity, family support, demographic and economic characteristics, etc.

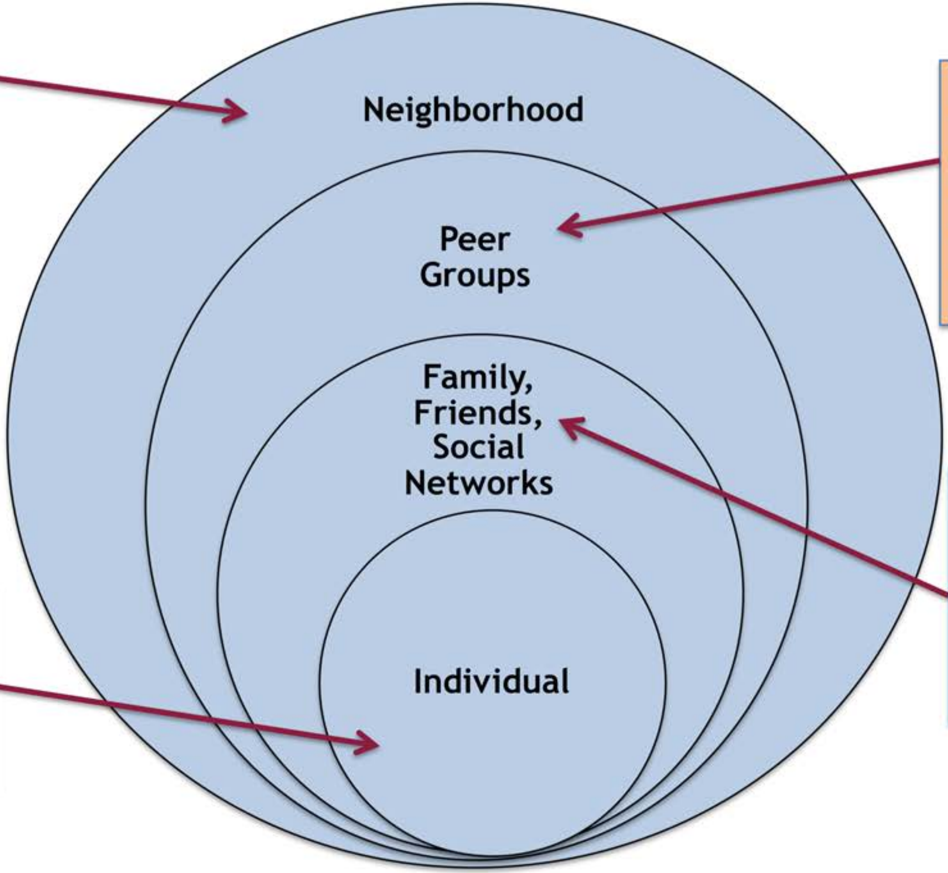


Local community Data Map



- Access to healthy food - grocery stores, community gardens, farmers markets, restaurants (fast food, other)
- Living Conditions
- Personal Safety
- Engagement
- Support Networks

- Behavioral Health
- Physical Health
- Social Wellness
- Support Networks



- Education
- English Literacy
- Health Literacy
- Engagement
- Support Networks

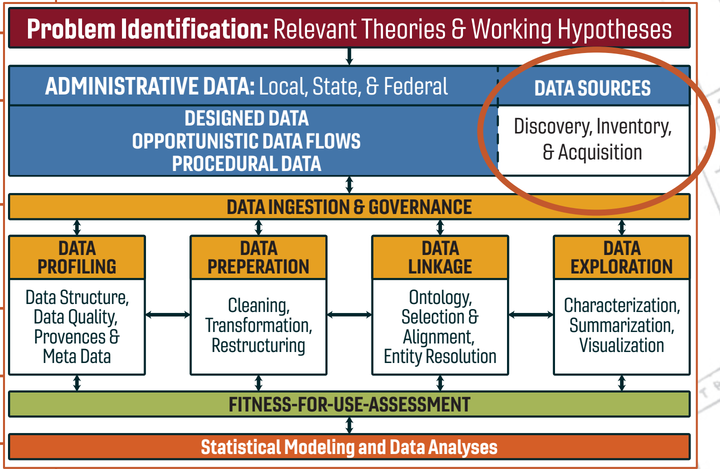
- Family Stability
- Income Stability
- Living Conditions
- Health Literacy
- Support Networks

Data Discovery, Inventory & Acquisition

Data Source	Geography
American Community Survey data (Census), 2011-2015 **(updating now to 2012-2016)**	Census Tracts and Block Groups
American Time Use Survey (BLS), 2017	National
Youth Risk Behavior Surveillance System, 2015	State
County Health Rankings, 2017	County
Built Environment, e.g., Grocery stores, SNAP retailers, recreation centers, community gardens	Address Level
Fairfax real estate tax assessment data	Address Level
Fairfax Open data: Zoning, Environment, water, Parks, Roads	Shapefiles
Fairfax County Youth Survey, 2016 8^{th}, 10^{th}, 12^{th} graders	High School Attendance Area
Virginia Department of Education, 2017	High School
National Center for Education Statistics, 2014-2015	High School
Center for Disease Control, 2014-2015	High School

Initial data sources used with geographic specificity

- **All** are **updated** as new data are available



Re-Distribution of Data and Estimates Across Geographies

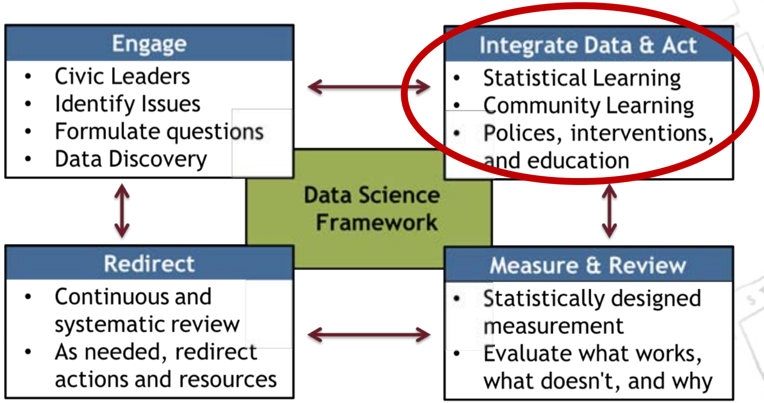
Problem: Data do not align with geographies of interest

- e.g., Supervisor (political) Districts and School Attendance Areas

Solution: Use data **direct aggregation**, if possible, alternatively develop **synthetic populations** based on data and redistribute

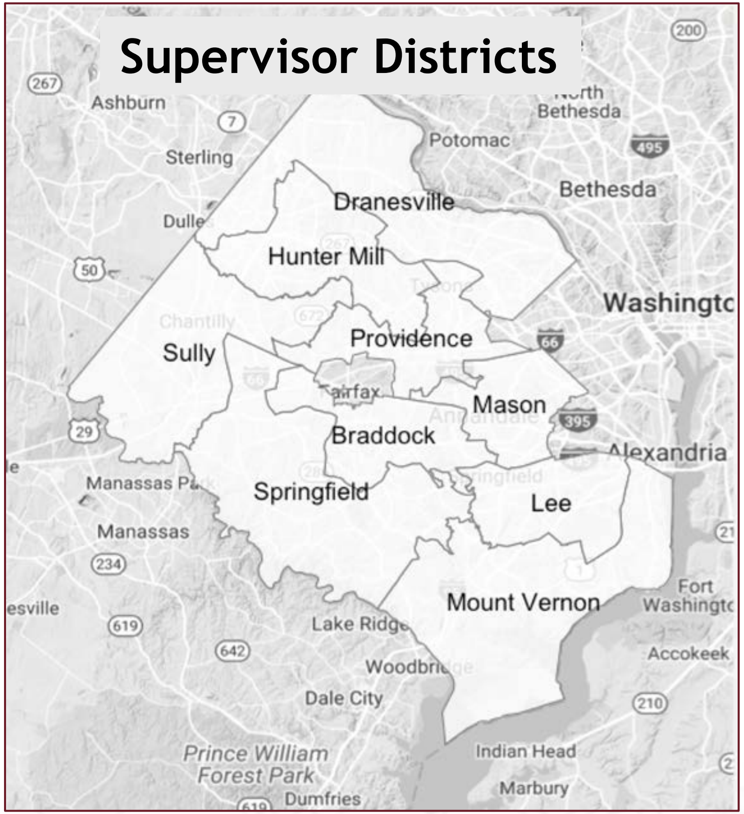
Synthetic re-distribution based on variables of interest

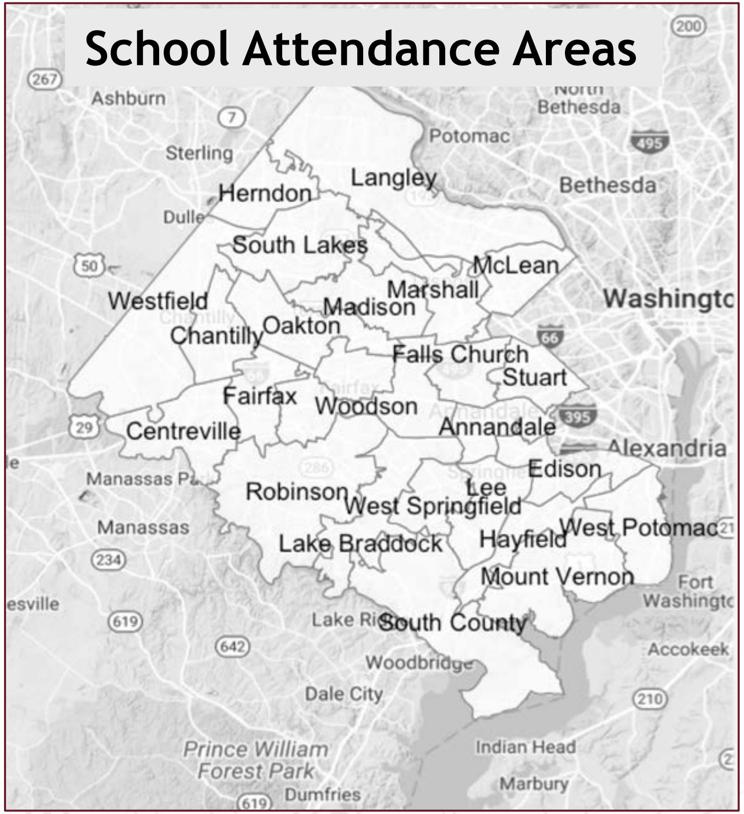
- Multivariate Imputation by Chained Equations (MICE)
- Iterative Proportional Fitting (IPF)



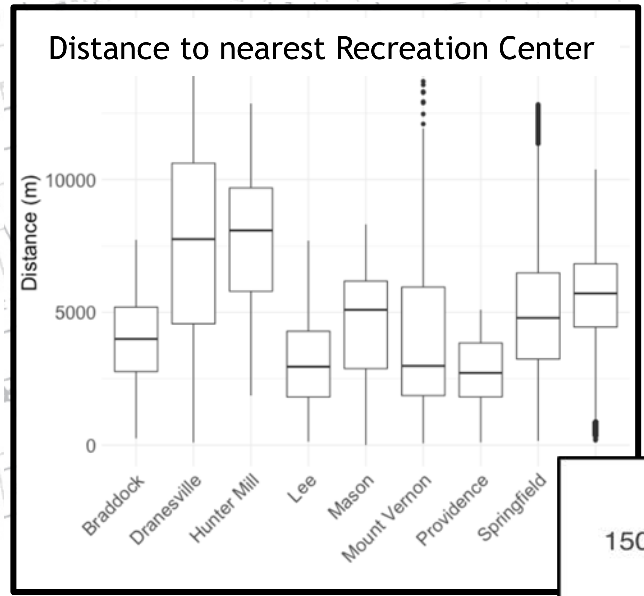
Example: Fairfax County, Virginia

Supervisor Districts and High School Attendance Areas





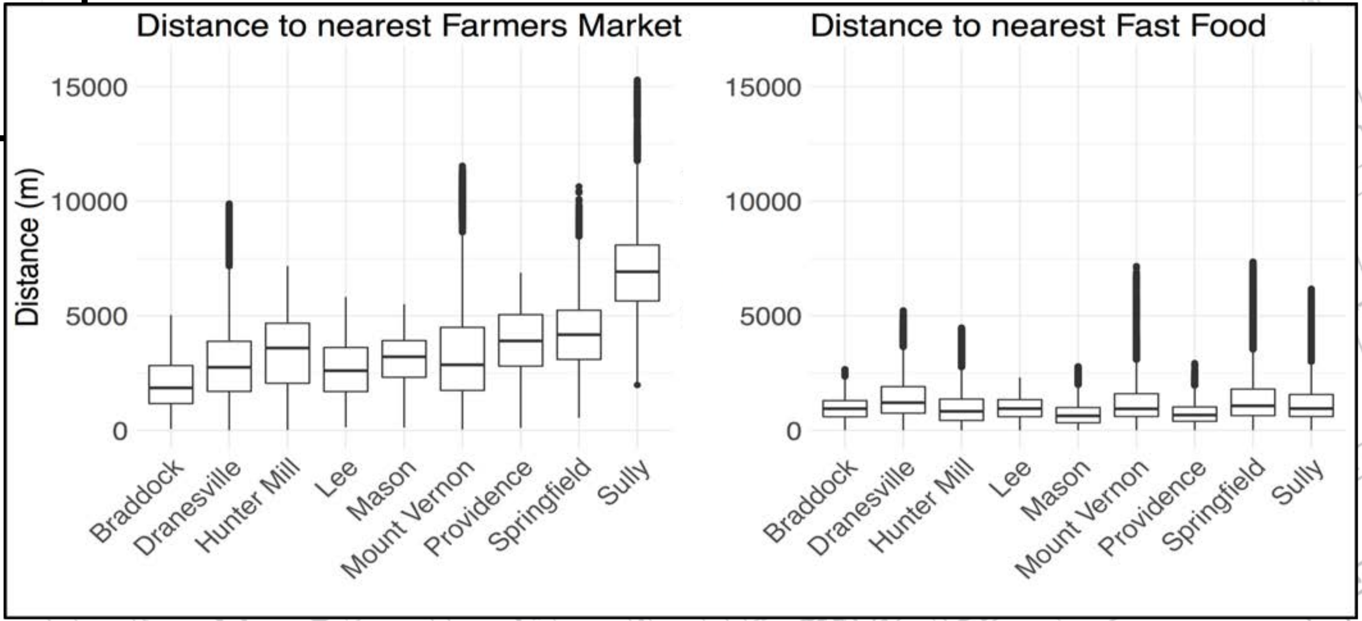
Direct aggregation based on location of housing units



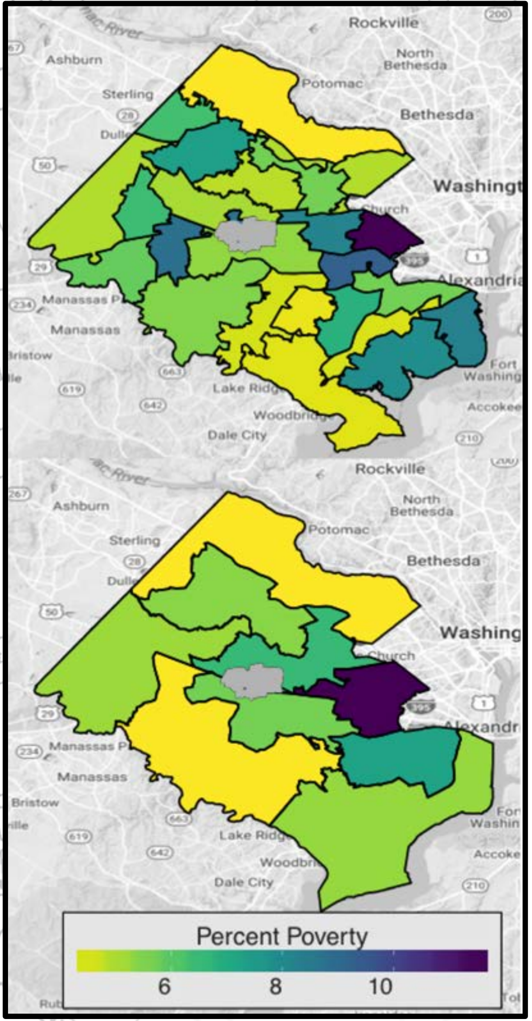
- Geocoding owner-occupied local housing stock
- Adding rental units typically requires imputation

Examples of place data:

- All restaurants
- Fast Food restaurants
- Farmer's Markets
- Community Gardens
- Recreation Centers
- SNAP Retailers
- Parks



Re-distribution of data based on synthetic populations



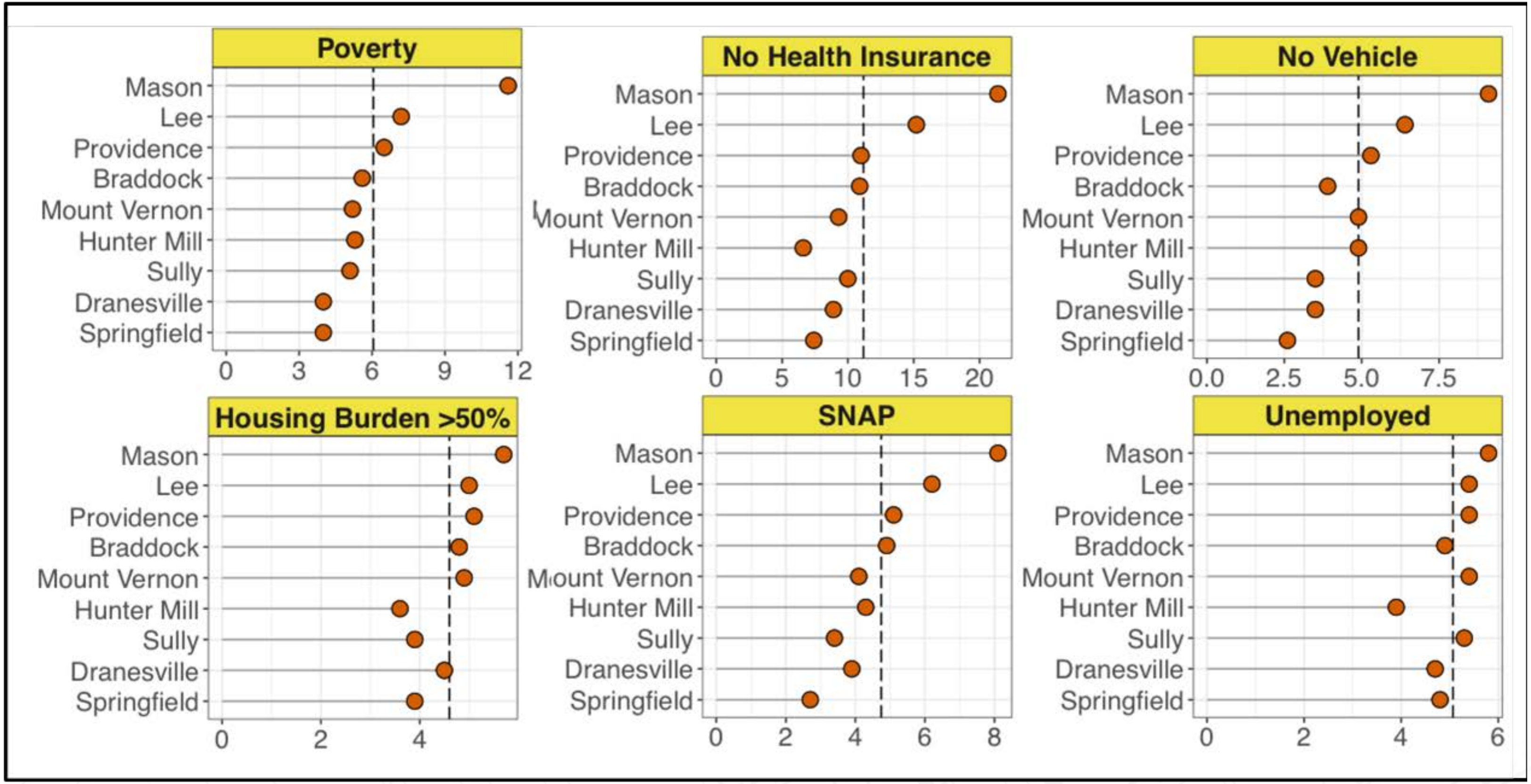
- Use American Community Survey (ACS) summaries and PUMS microdata to impute synthetic person data for all people in area of interest

- Re-weight synthetic data according to ACS tables to simultaneously match the relevant distributions, to Census Tracts or Block Groups
  - Age, income, race, and poverty in this case

- Aggregate synthetic person data to compute summaries, and margins of error, over the new geographic boundaries

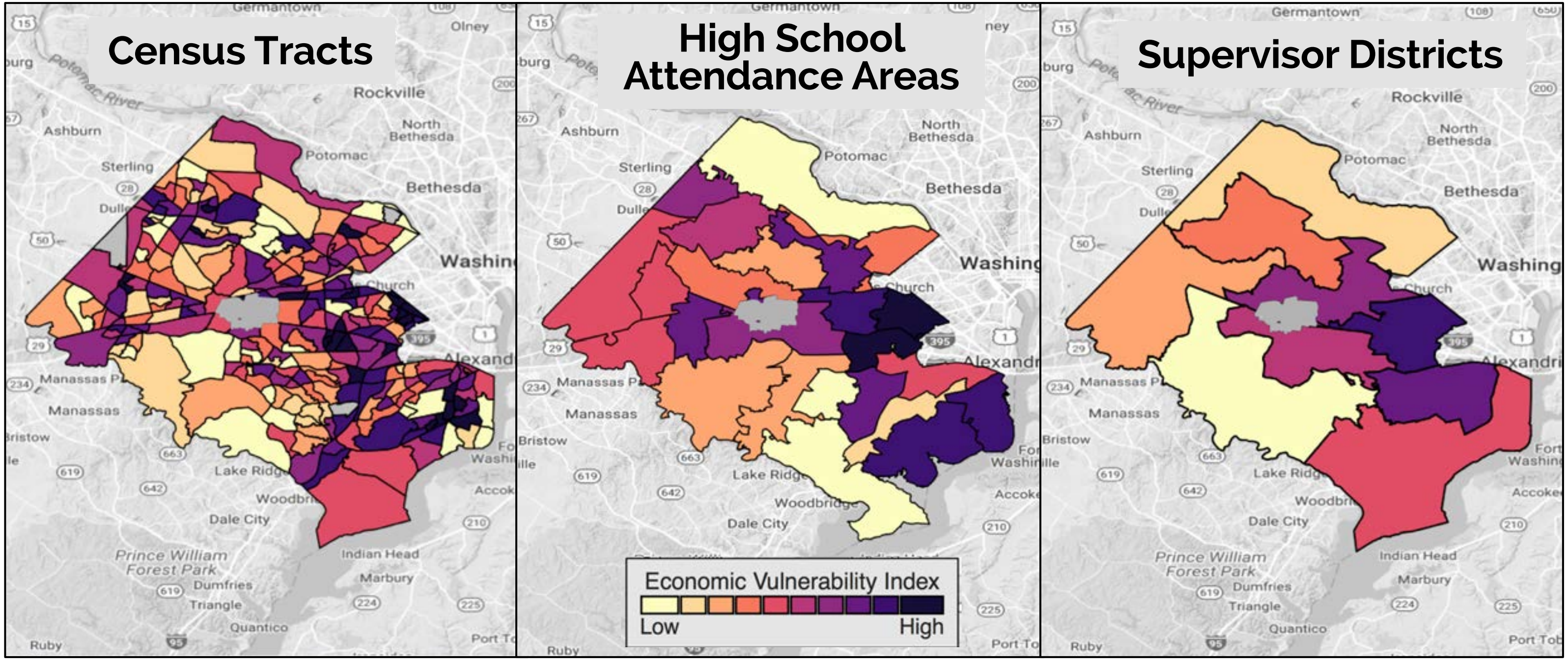
Fairfax Profiles by Supervisor Districts

Dashed lines = Average; Supervisor Districts arranged by Poverty high to low



Source: American Community Survey 2011-2015 aligned to Supervisor Districts using SDAL Synthetic Technology.

Fairfax Sub-County Vulnerability Indicators



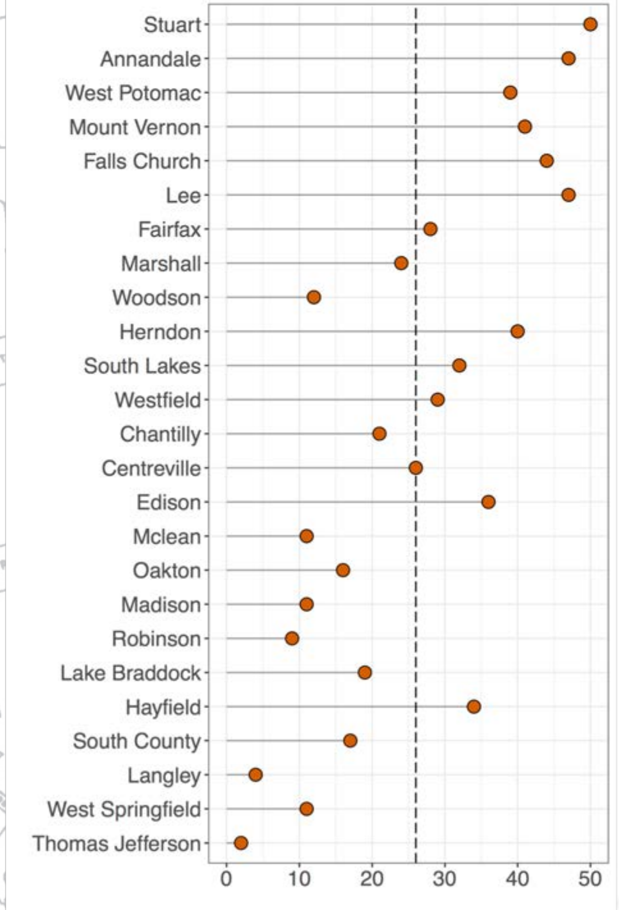
Based on a statistical combination of the percentage of Households with:

- housing burdens > 50% of Household income
- no vehicle
- receiving Supplemental Nutrition Assistance Program (SNAP)
- in poverty

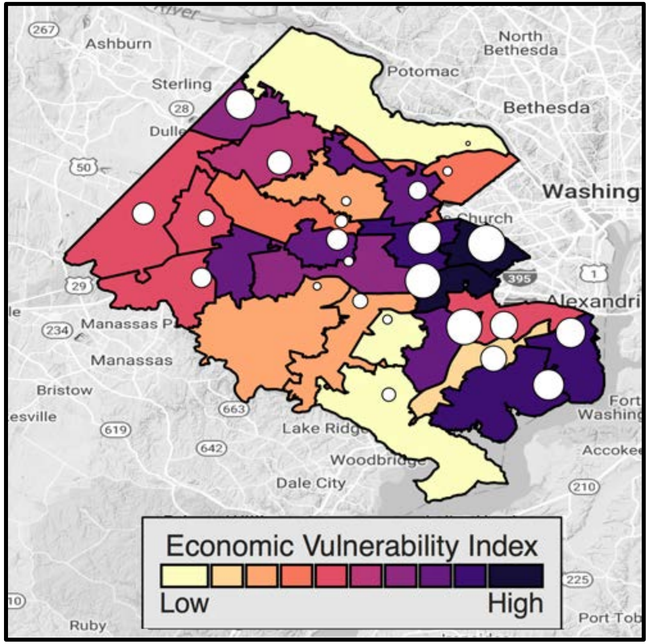
Source: American Community Survey 2011-2015 aligned to Supervisor Districts using SDAL Synthetic Technology.

High School Characteristics

High School Vulnerability Index ordered by Economic Vulnerability Index



School Vulnerability Index



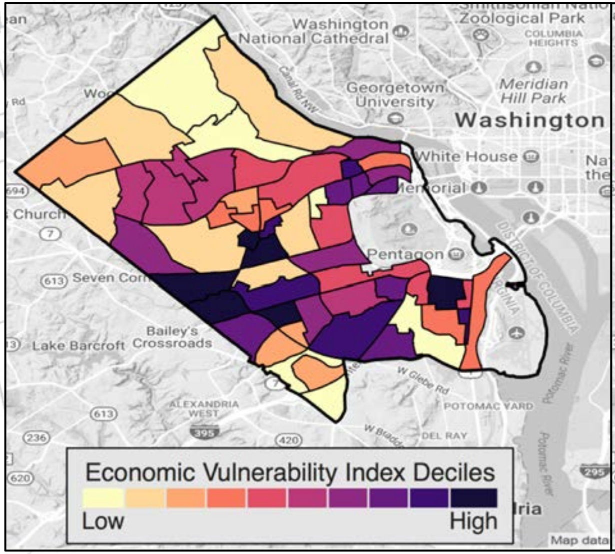
Combination of:

- Percentage of student in LEP classes
- Percentage of students that eligible for **one** of the following:
  - Free/Reduced Meals
  - Medicaid
  - Temporary Assistance for Needy Families
  - Migrant or experiencing Homelessness

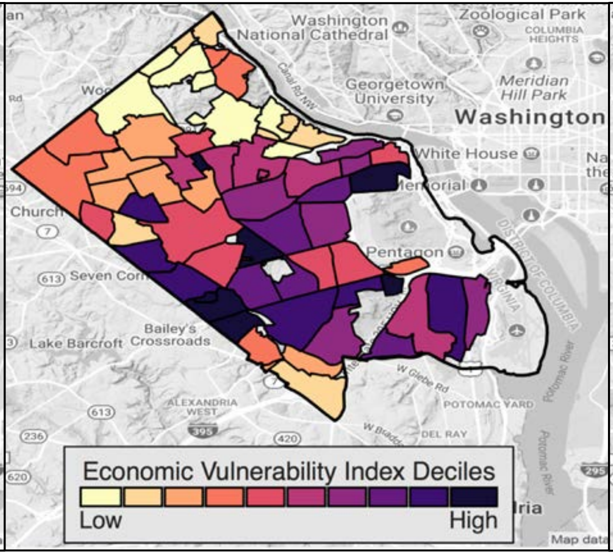
Sources: ACS 2011-2015; NCES, CDC, and VDOE 2014-2015.

Arlington County
Sub-county Vulnerability Indicators

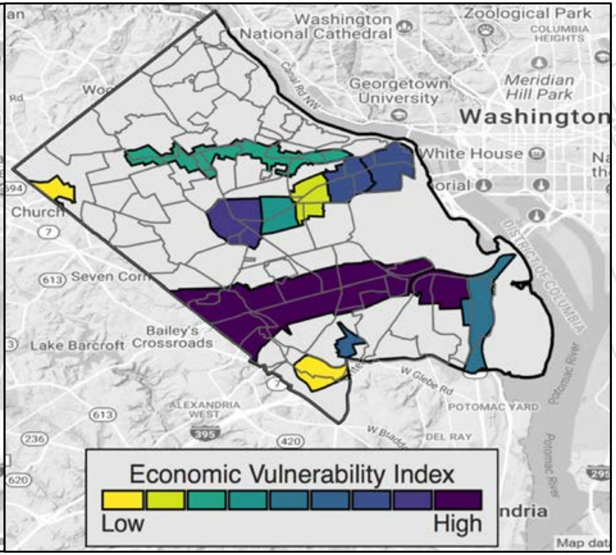
Census Tracts



Arlington Civic Association Neighborhoods



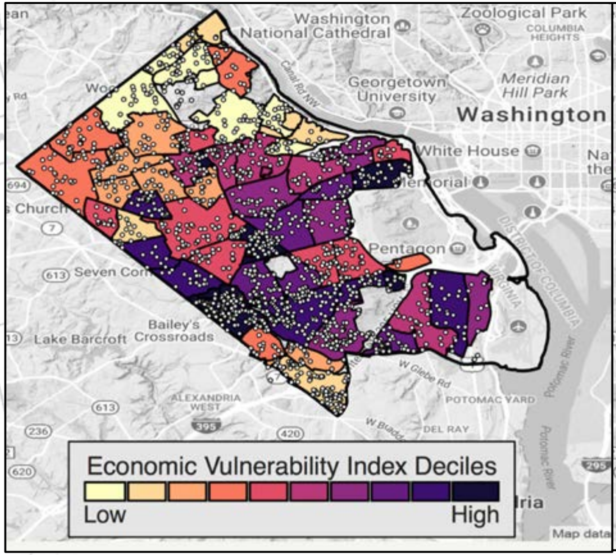
High-Density Planning Regions



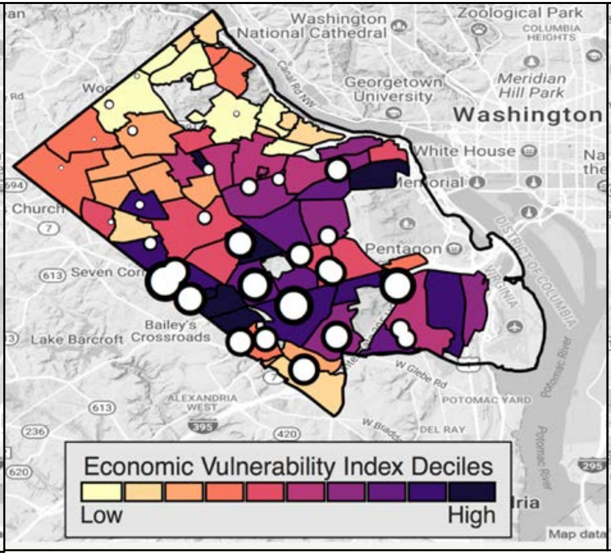
Source: American Community Survey 2012-20156 aligned to geographic areas using SDAL Synthetic Technology.

Arlington County Neighborhood Insights

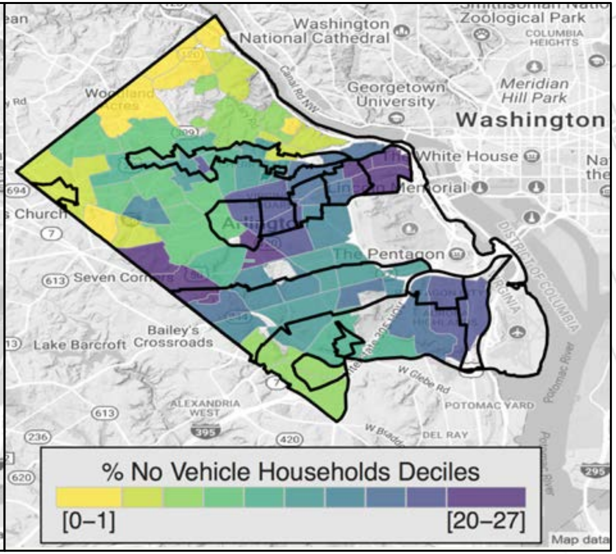
Households **receiving subsidies** from Department of Parks and Recreation



School and neighborhood vulnerability indices



High-Density Planning Regions with % households with no vehicles



Sources: ACS 2012-2016; NCES, CDC, and VDOE 2014-2015; Arlington County Department of Parks & Recreation 2016.

Next Steps

Engage

- Civic Leaders
- Identify Issues
- Formulate questions
- Data Discovery

Integrate Data & Act

- Statistical Learning
- Community Learning
- Polices, interventions, and education

Data Science Framework

Redirect

- Continuous and systematic review
- As needed, redirect actions and resources

Measure & Review

- Statistically designed measurement
- Evaluate what works, what doesn't, and why

Democratization of data across the United States

- Bringing data in service of the public good
- Deepening partnership between communities and Land Grant Universities
- Enabling communities to become *data-driven learning communities*

Virginia Cooperative Extension
Virginia Tech • Virginia State University

Cooperative Extension 1914-2014 100 Smith-Lever Act
Extending Knowledge Changing Lives

Iowa State University Extension and Outreach

S. Keller, S. Nusser, S. Shipp and C. Woteki, (2018). A National Strategy for Community Learning through Data Driven Discovery, *Issues in Science and Technology*, Spring 2018.

Community Learning THROUGH Data Driven Discovery

Meeting Educational Aspirations

Meeting educational aspirations of a state

Issue: Virginia strives to be the "smartest" state by 2030

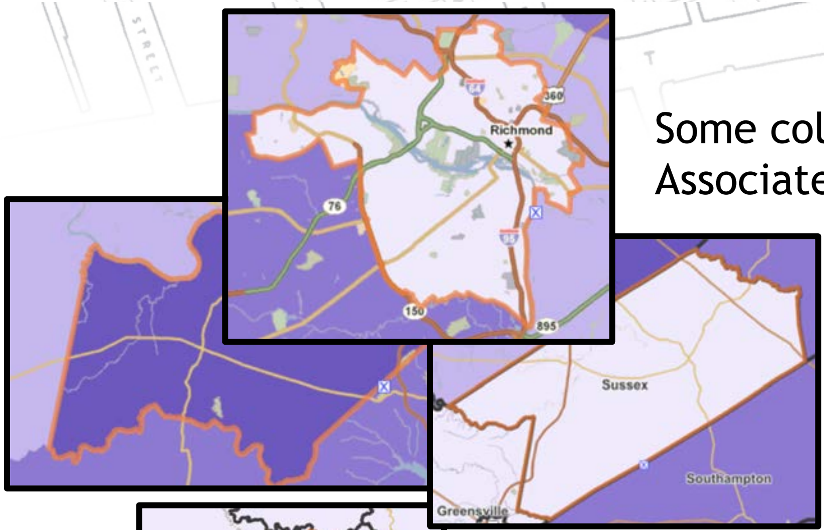
- This will require an increase in post secondary training and education for the 18-65 age group

Goal: To identify subpopulations for outreach and policy development for increasing Virginia's post-secondary education and training levels from 51% in 2016 to 70% by 2030

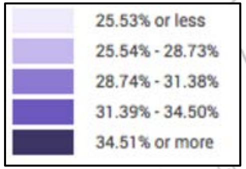


Two Study Areas

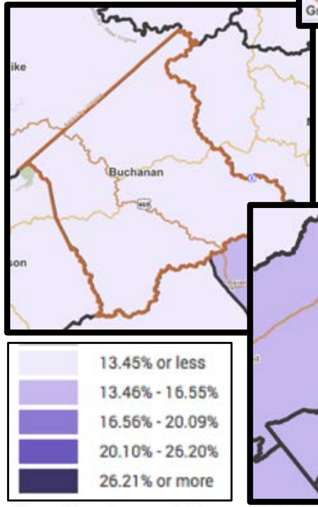
- Richmond Area (Sussex County, Powhatan County, and Richmond City) is demographically diverse with a mix of urban/rural (metro) communities
- Roanoke/Appalachia Area (Buchanan County, Bland County, Roanoke County, and Roanoke City) is a mix of urban/rural (metro/nonmetro), White, and older

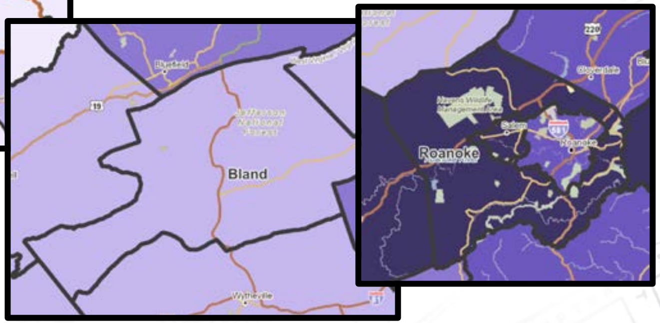


Some college or Associate's degree



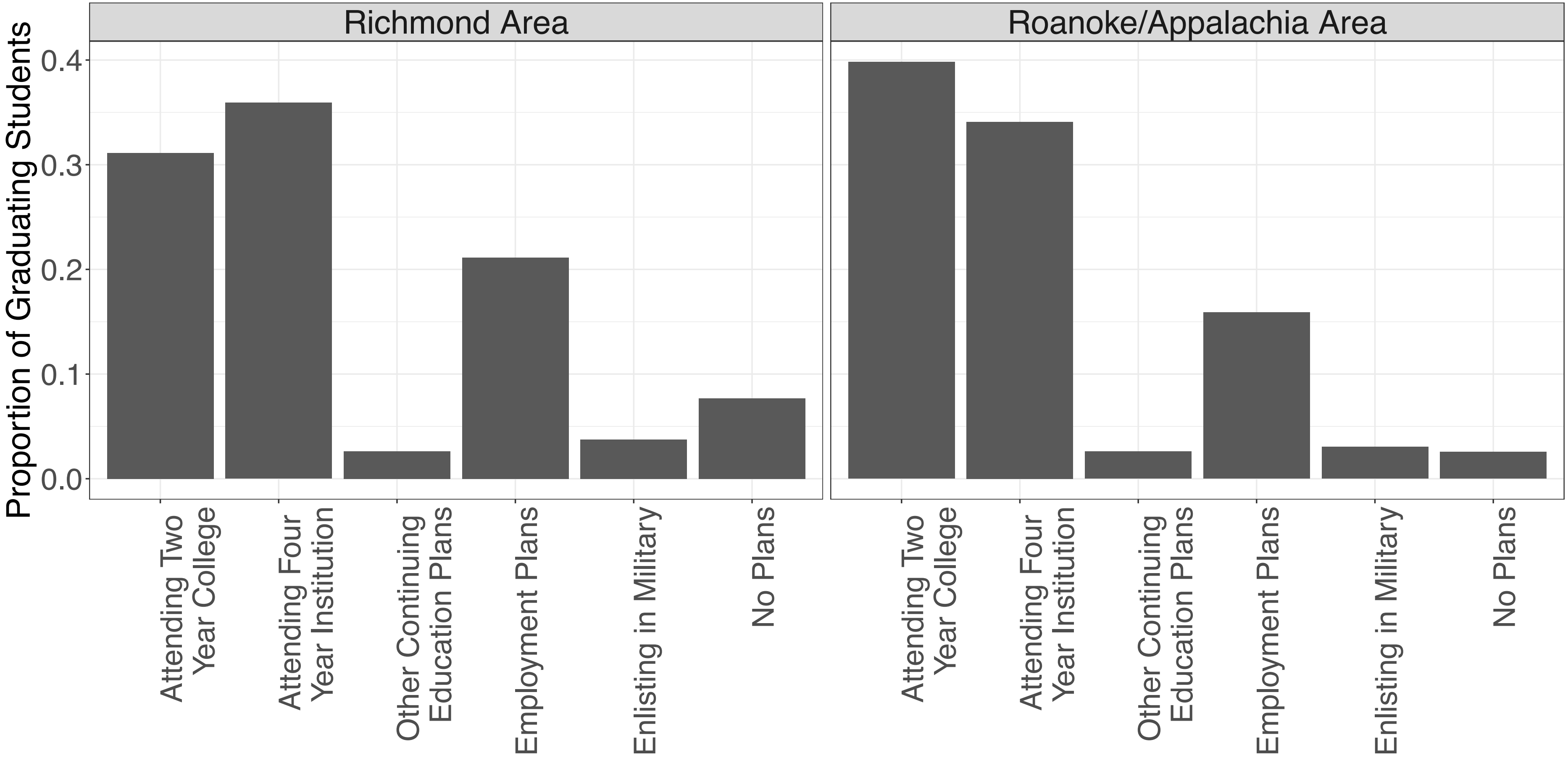
At least a Bachelor's degree





Source: PolicyMap, 2015 American Community Survey 5-year Estimates

Limited insights from post-high school plans



Data Discovery, Inventory, & Acquisition

High School	Postsecondary Education	Credentials and Skill-based Training	Work Experience & STEM Occupations

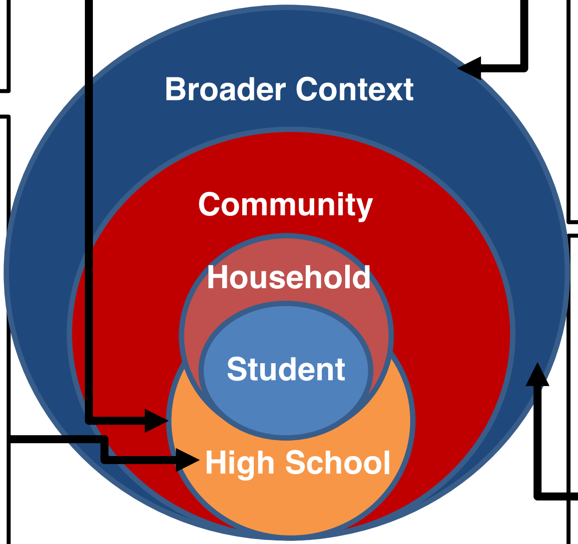
Data Map

High School Student Body Characteristics

- % Students disadvantaged (VDOE)
- % Students by gender (VDOE)
- Student offenses and disciplinary outcomes (VDOE)
- Drop-out rates (VDOE)

High School "Postsecondary-Going" Culture

- Graduation rate (VDOE)
- Advanced/regular degree ratio (VDOE)
- % CTE program graduates (VDOE)
- College application rate (SCHEV)
- College acceptance rate (SCHEV)
- % Enrolled in AP classes (VDOE)
- % Passed AP tests (VDOE)
- % in Dual Enrollment courses (VDOE)
- % Teachers w/ graduate degrees (VDOE)
- % Students took the SAT (College Board)
- Mean SAT scores (College Board)
- ….



Community Characteristics

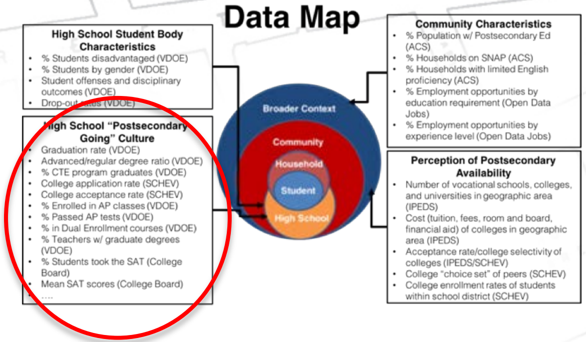
- % Population w/ Postsecondary Ed (ACS)
- % Households on SNAP (ACS)
- % Households with limited English proficiency (ACS)
- % Employment opportunities by education requirement (Open Data Jobs)
- % Employment opportunities by experience level (Open Data Jobs)

Perception of Postsecondary Availability

- Number of vocational schools, colleges, and universities in geographic area (IPEDS)
- Cost (tuition, fees, room and board, financial aid) of colleges in geographic area (IPEDS)
- Acceptance rate/college selectivity of colleges (IPEDS/SCHEV)
- College "choice set" of peers (SCHEV)
- College enrollment rates of students within school district (SCHEV)

Ziemer, K. S., Pires, B., Lancaster, V., Keller, S., Orr, M., & Shipp, S. (2017). A New Lens on High School Dropout: Use of Correspondence Analysis and the Statewide Longitudinal Data System. *The American Statistician*.

Indicator of Postsecondary-Going Culture

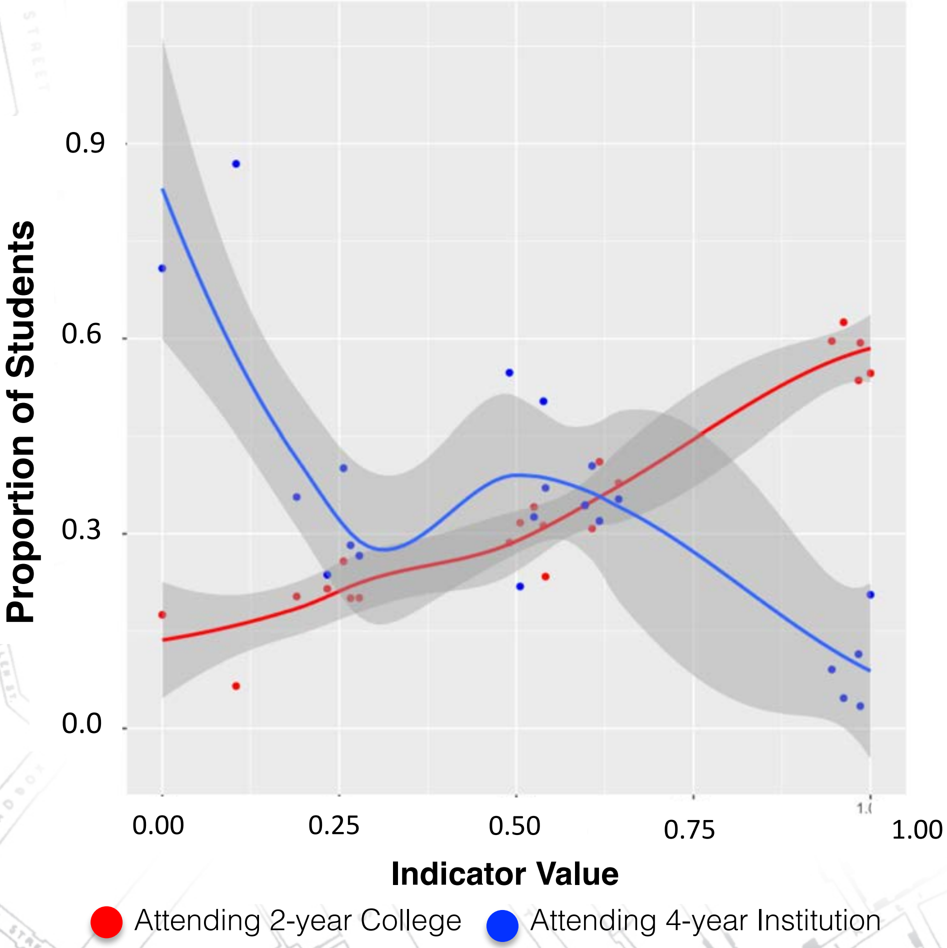


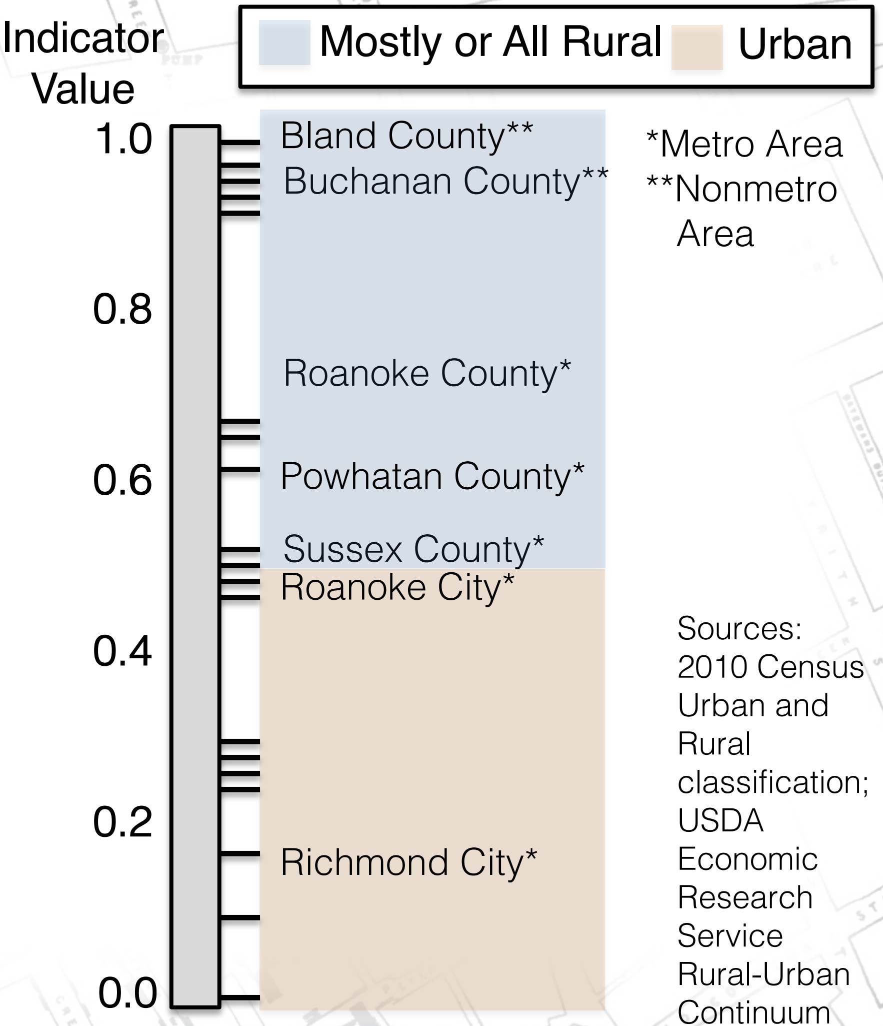
- Can we measure/quantify postsecondary-going culture in high schools?
- Variable selection based on literature in college-going culture and feedback from experts
- Principle components analysis to understand the underlying interrelationships of the data, assign weights to variables, and assign indicator values to each high school

High School "Postsecondary-Going" Culture

- Graduation rate (VDOE)
- Advanced/regular degree ratio (VDOE)
- % CTE program graduates (VDOE)
- College application rate (SCHEV)
- College acceptance rate (SCHEV)
- % Enrolled in AP classes (VDOE)
- % Passed AP tests (VDOE)
- % in Dual Enrollment courses (VDOE)
- Student/Teacher ratio
- % Teachers w/ undergraduate or graduate degrees (VDOE)
- % Students took the SAT (College Board)
- Mean SAT scores (College Board)
- ....

Indicator Values



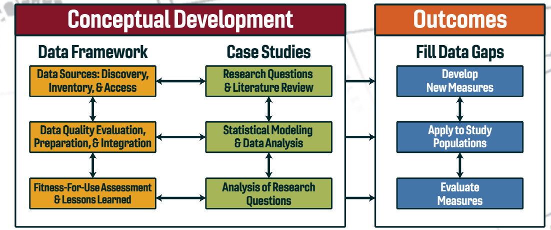


*Metro Area
**Nonmetro Area

Sources: 2010 Census Urban and Rural classification; USDA Economic Research Service Rural-Urban Continuum

U.S. Army Research Institute for the Behavioral and Social Sciences

Exercising the our full research model



Research Questions:

- What is the **value of combining** DoD, civilian, and non-federally collected data sources to enhance or complement a representative use of PDE and other DOD and non-DOD data sources?
- How does this help capture and model individual, unit, and organizational characteristics and non-military **contexts** that affect important questions?
- Explore these questions in the context of a specific **case studies**
- Use outcomes to **drive new measurement to fill data gaps**

Case Studies: Army attrition and performance are being examined using longitudinal data at the level of the Soldier and the Team/Unit

Initial Performance Framework

Antecedents of Performance: Army Values, Warrior Ethos, Big 5 Personality Traits, Creativity, Motivational Traits, Job & Community Embeddedness, Vocational Interests

Direct Determinants

Cognitive Component	Social Component	Physical Component

Behaviors of Performance

Productive Behaviors		**Counterproductive Behavior**
Actions/behaviors that contribute to the organization's goals anchored at the floor by achieving the minimum standard needed to meet the contractual obligations of the job.	Impact Organization Impact Colleagues Impact Self	Actions/behaviors that impede progress towards an organization's goals and can also harm self and colleagues.

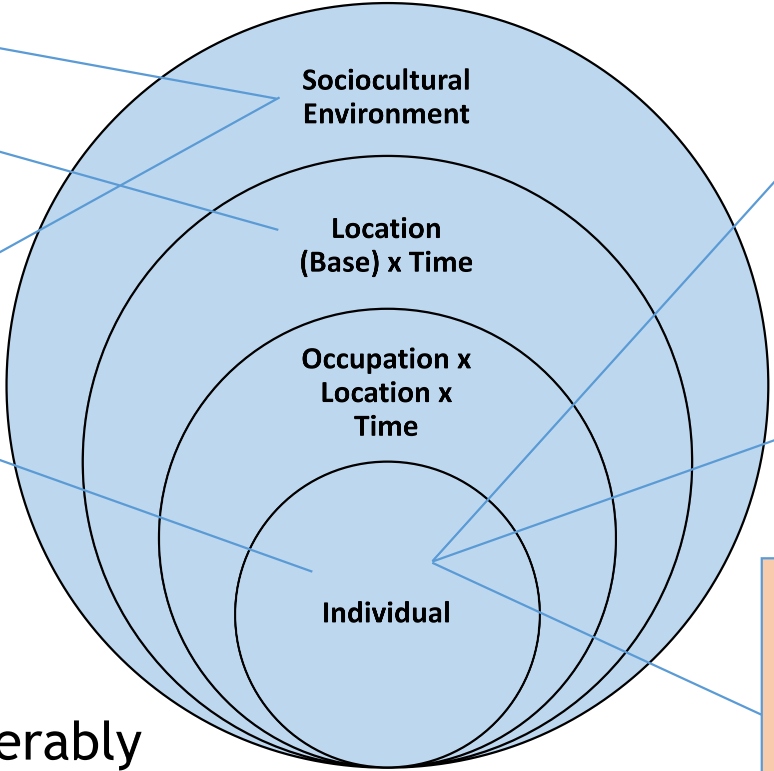
Criteria of Performance (Outcomes)

Soldier Data Map

Policy changes (e.g., peacetime vs. war)
Non-personal shock events (e.g., 9/11)
Job alternatives (e.g., ACS employment)

Local community (e.g., ACS data)

Constructs to be Modeled
National Army prestige/support
Cohesion
Job satisfaction
Job investment
Commitment norms



Demographics
Race
Ethnicity
Sex
Birthdate/Age
Faith group
Education level and discipline
Marital status
Spouse in military indicator
Number and type of dependents
ASVAB score
State/country of residence before entry

Service Dates and Locations
Length of time in service
Length of service agreement
Location (base) over time
Obligation begin and end dates
Term of service
Date of initial entry
Date of end of initial training

Military-Specific Characteristics/Incentives
Security clearance
Education incentive indicator
Career status bonus program indicator
Object of mission (e.g., advanced cruise missile)
Occupation group (primary and secondary)
Re-enlistment eligibility
Aeronautical rating code (e.g., astronaut)
Flying status indicator
Pay grade (e.g., E-3) and length of time in grade
Character of service (e.g., honorable)

This will grow considerably

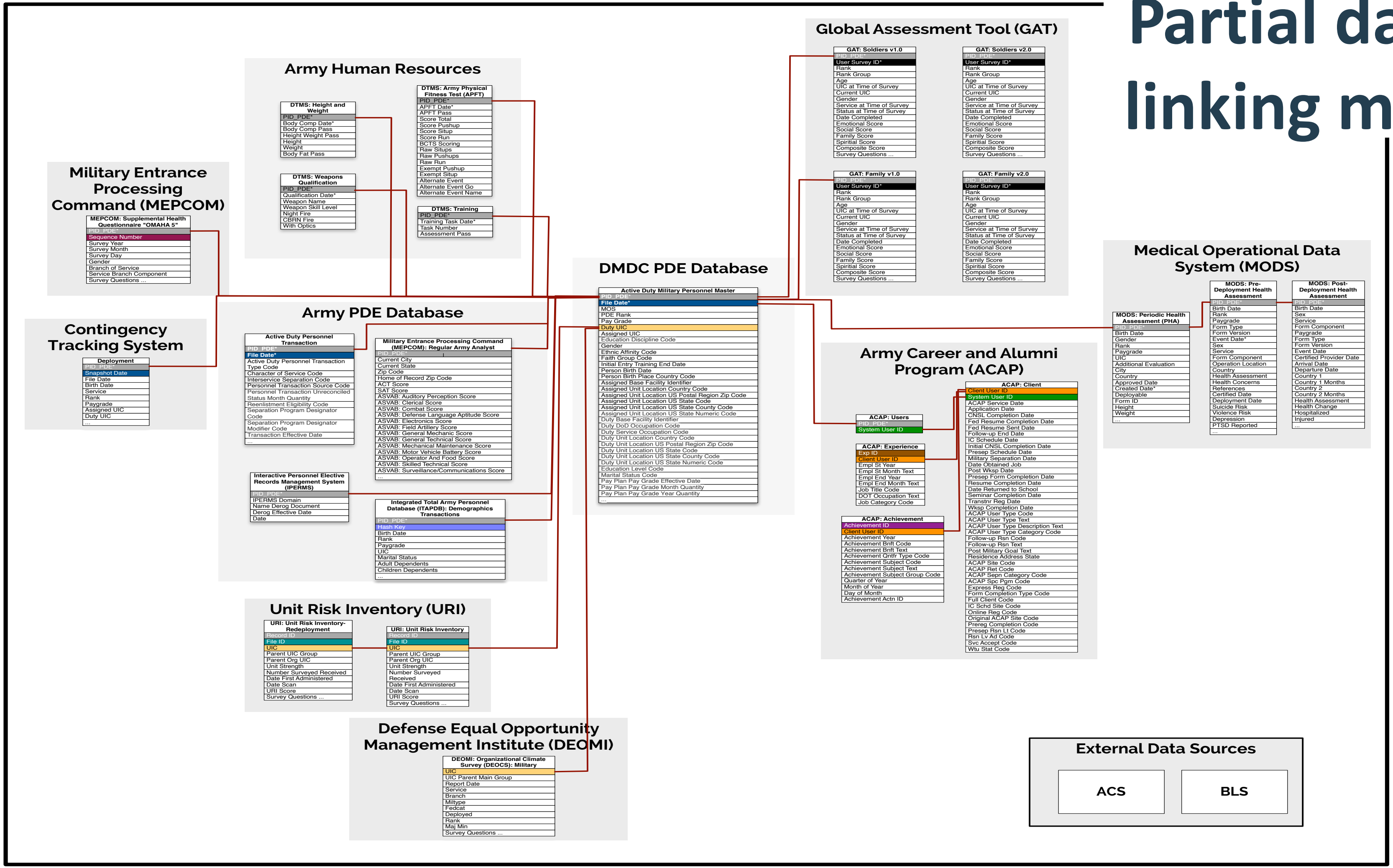
Data access

- Common Access Cards
- IRB processes integrated and updated to accommodate anticipated data needs for social construct development
- Access to Person Data Environment (PDE)
- Building data environment in PDE, e.g., Rstudio, R Markdown for profiling, Oracle to manage metadata
  - Requesting data
  - Importing data
  - Exercising data profiling, preparation, linkage, and exploration
  - Running models and exporting model results

Person-Event Data Environment

Research Facilitation Laboratory

Partial data linking map



Data pipeline: sharable data products

Demographics Table

- Information about the enlistee that typically remains static over time, e.g., gender, race, ethnicity, entry test scores
- Simple rules are applied to resolve duplicates and entries with multiple values
- Contains one row per PID

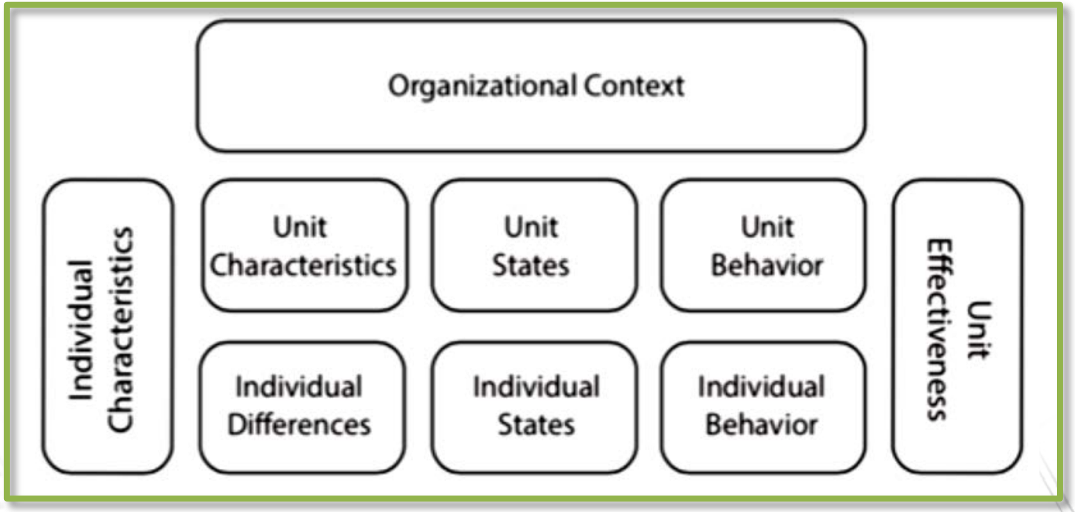
Transaction Table

- Events or enlistee information that can change periodically, e.g., duty station, rank, pay grade, interservice separation code
- Contains multiple rows per PID

Column Name	Description	Original Table
PID_PDE	Enlistee's Unique ID	Master
PN_SEX_CD	Gender	Master
RACE_CD	Race Code	Master
INIT_ENT_TRN_END_DT	Initial Entry Training End Date	Master
DATE_BIRTH_PDE	Person Birth Date	Master
PN_BIRTH_PLC_CTRY_CD	Person Birth Place Country Code	Master
HOR_ZIP_CODE_PDE	Home of Record Zip Code	Analyst
ACT_SCORE	ACT Score	Analyst
SAT_SCORE	SAT Score	Analyst
AP	ASVAB: Auditory Perception Score	Analyst
CO	ASVAB: Combat Score	Analyst
. . .	. . .	. . .

Building model complexity

- Model flexibility for connecting many data sources and computation
- Need to integrate "external" data sources that change over time
- Need to integrate person-specific information in context
  - Relevant time and activity is with respect to person's term
  - "Exposures" to duties, leaders, training, ...
  - Unit, duty locations, commitment, …



Population Dynamics

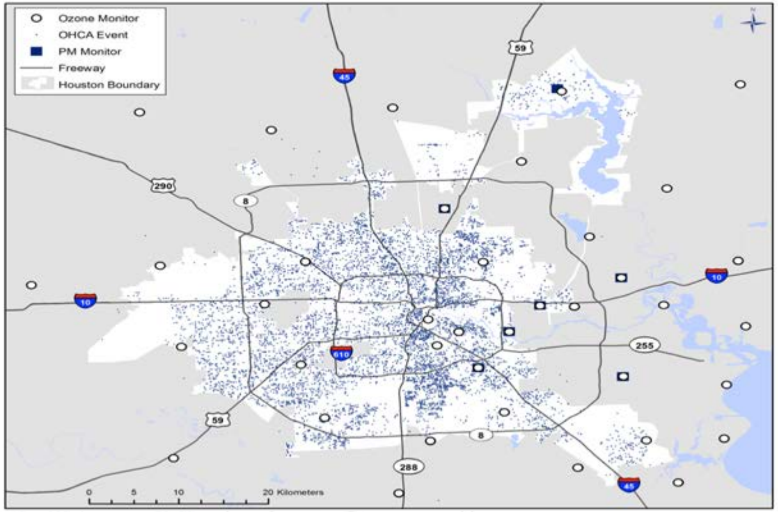
B. Pires, G. Korkmaz, K. Ensor, D. Higdon, S. Keller, B. Lewis, B., and A. Schroeder, 2018. Estimating individualized exposure impacts from ambient ozone levels: A synthetic information approach. *Environmental Modelling & Software*. (Forthcoming)

Houston EMS Study for Individual Risk

Goal: Identify links between air pollution and acute health events at community level

Model and Data:

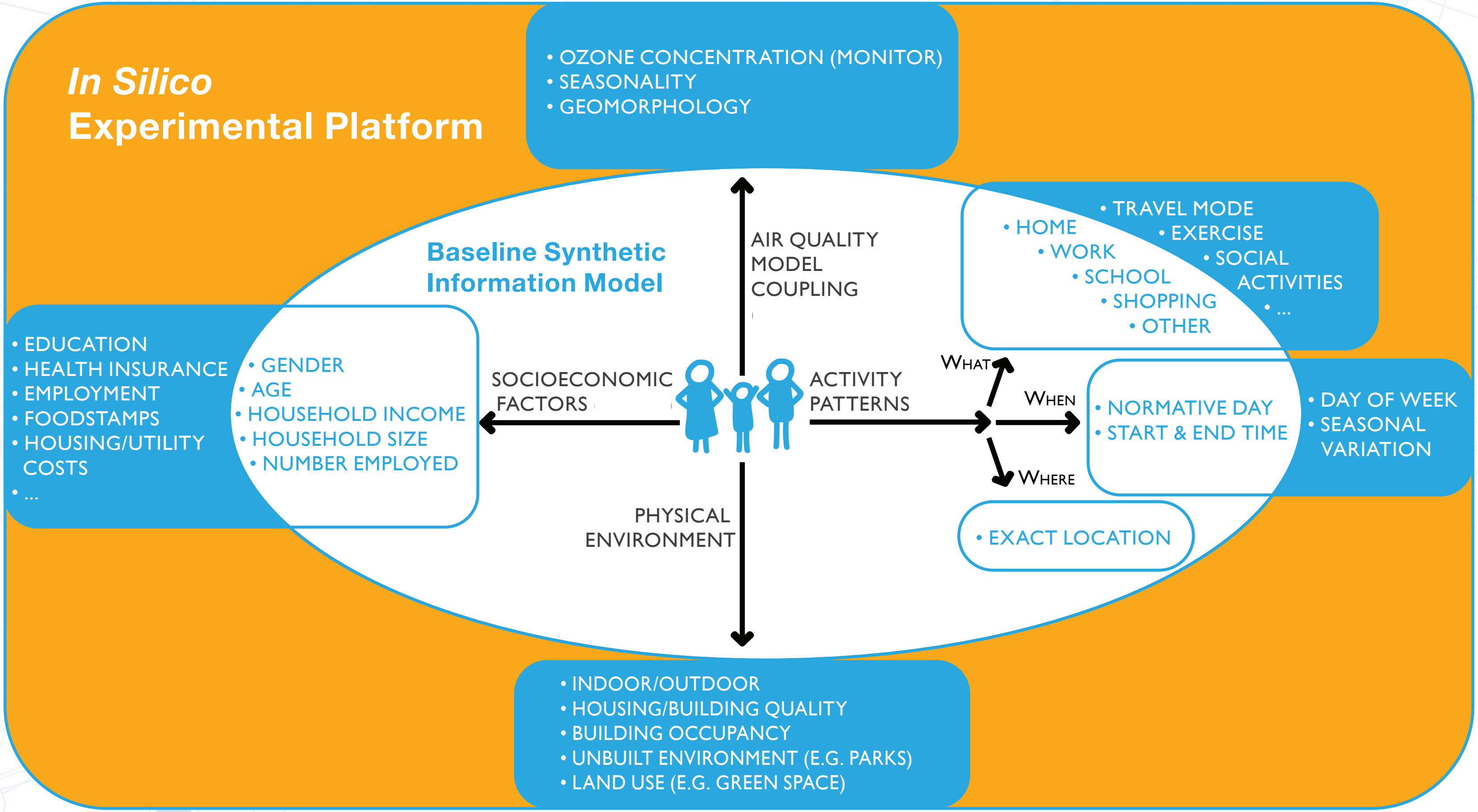
- Pathophysiological link between out-of-hospital cardiac arrest (OHCA) and ozone level
- Case cross-over, time stratified design
  - Houston, 2004-2011
  - EMS data of 11,754 cases
  - Predictor variable is *aggregate ozone over a 3 hour window* leading up to event



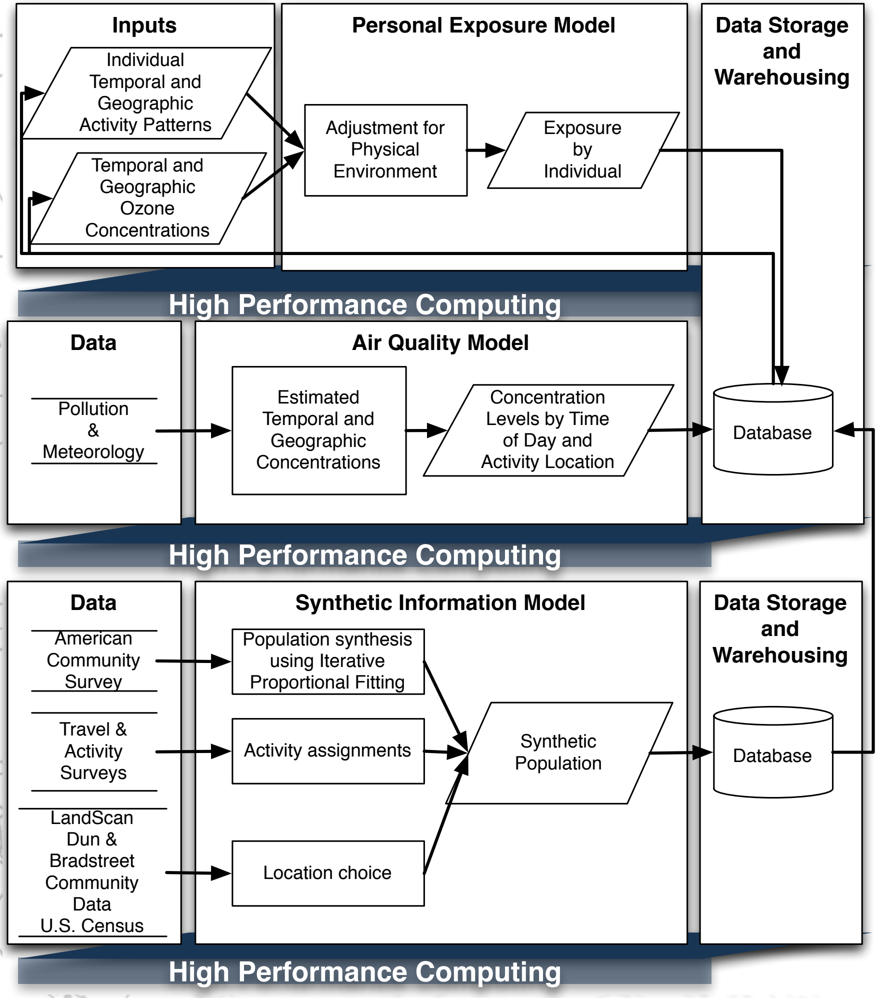
Ensor, et al., Circulation, Volume 127(11):1192-1199

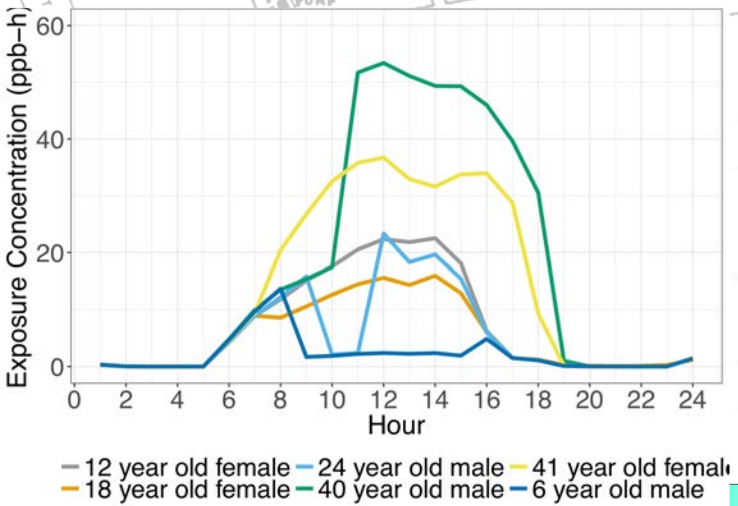
Results: 20 ppbv increase in ozone 1 to 3 hours previous of event was associated with a 4.4% increased risk

Synthetic Information Platform

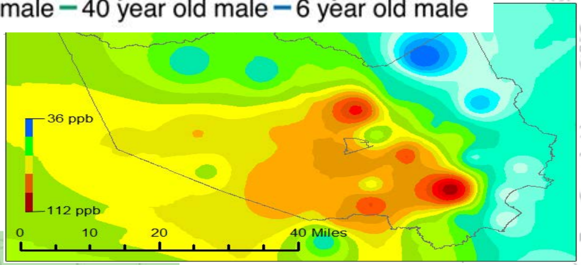


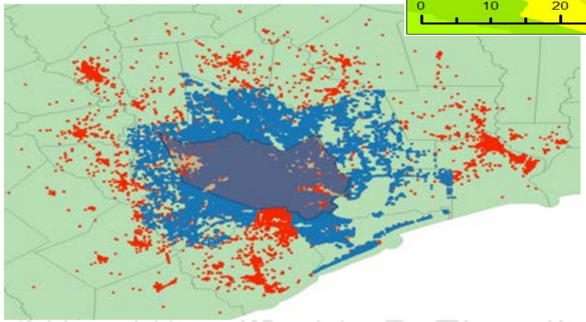
In-Silico Platform for Environmental Coupling





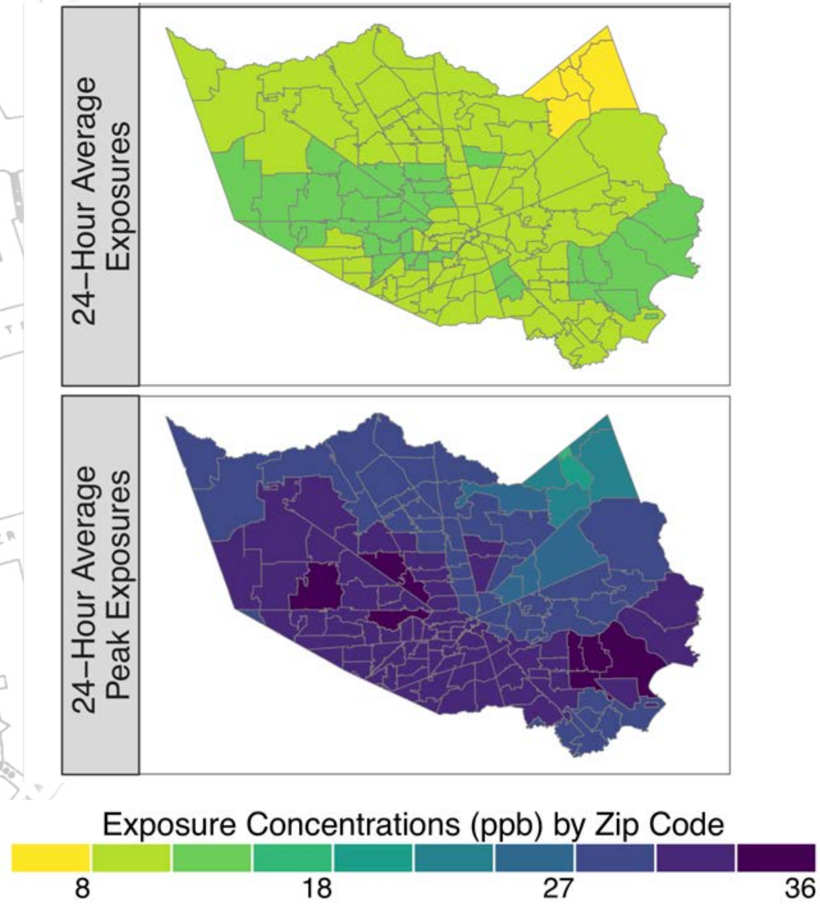
10:00 am
August 26, 2008

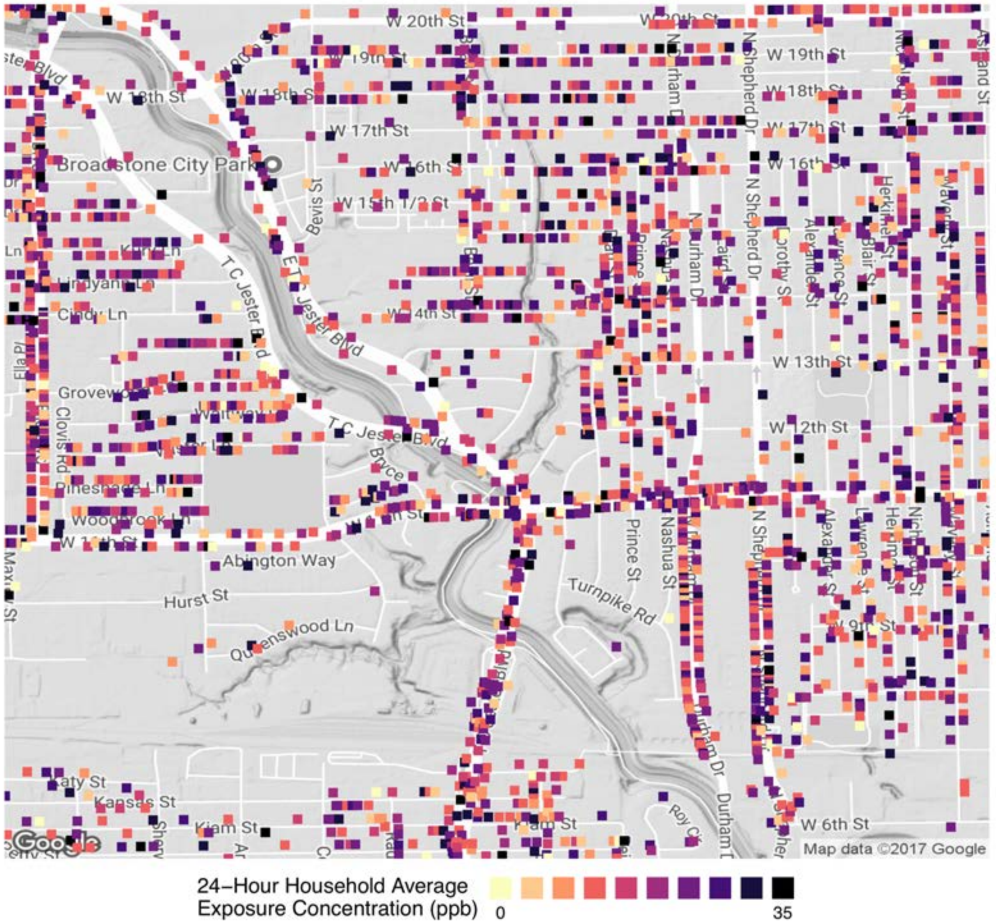


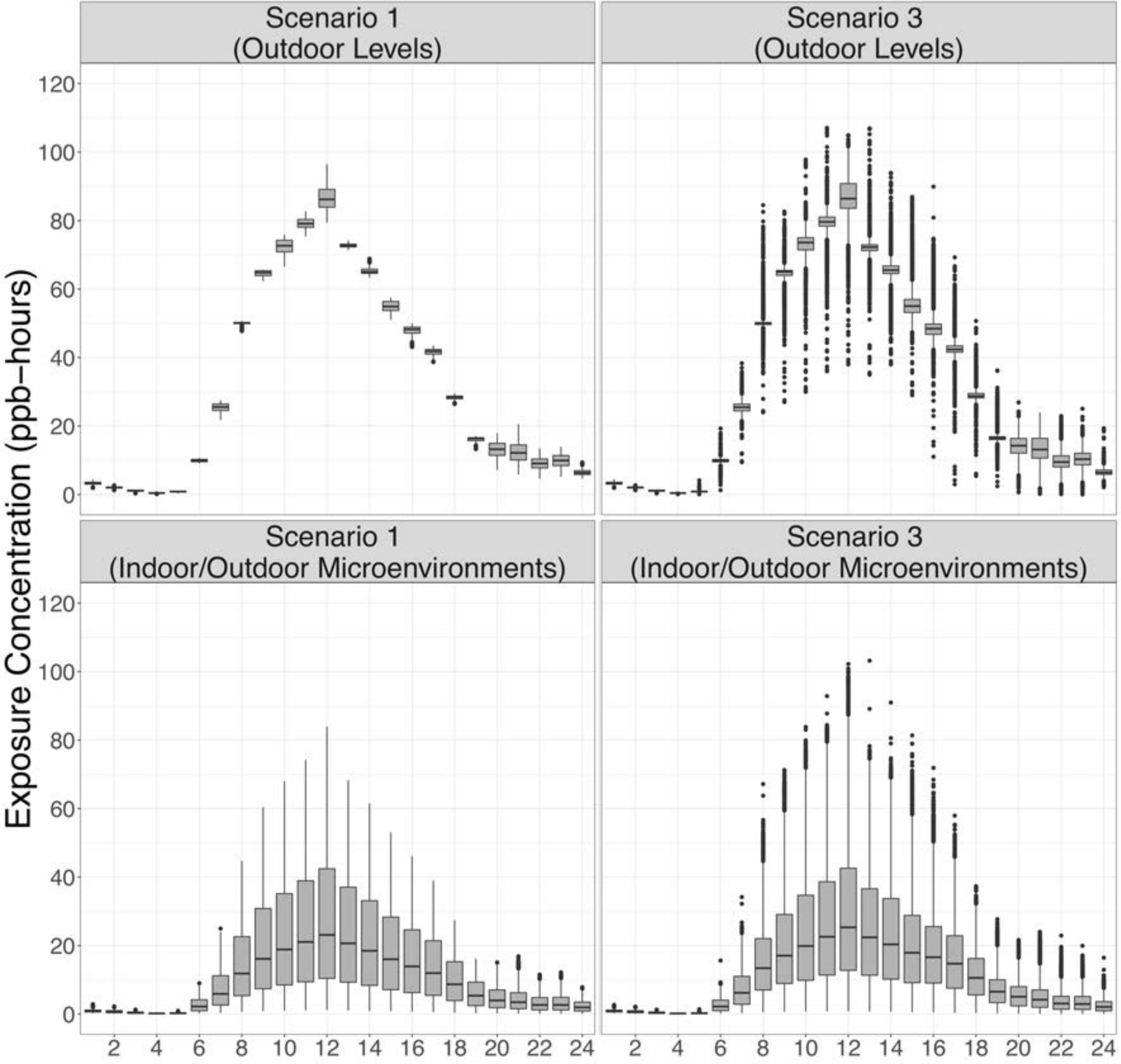


4.9M people
1.8M Households
1.2M Locations

Location and movement matter







Exercising the platform

Scenario 1:
Population stays home

Scenario 3:
Population moves

Workforce Development

Data Science for the Public Good (DSPG)

https://www.bi.vt.edu/sdal/projects/data-science-for-the-public-good-program

IDENTIFYING STEM EDUCATION PATHWAYS
Sponsor: Pat Ruggles, The National Center for Science & Engineering Statistics at the National Science Foundation (NSF)

EXPLORING MENTAL HEALTH SERVICES FOR FAIRFAX COUNTY YOUTH
Sponsor: Michelle Gregory, Sophia Dutton, and Linda Hoffman, Fairfax Health and Human Services

RESIDENTIAL SMOKE ALARM NEED IN ARLINGTON COUNTY
Sponsor: Battalion Chief Mike Gowen, Arlington County Fire Department

HOW DO EVENTS AFFECT CRIME?
Sponsor: Captain Bruce Benson and Niki Levy, Arlington County Police Department

MODELING THE IMPACT OF OPEN SOURCE SOFTWARE: NETWORK OF R PACKAGES
Sponsor: Carol Robbins, The National Center for Science & Engineering Statistics at the National Science Foundation

MODELING RESPONSE TIME FOR STRUCTURE FIRES
Sponsor: Battalion Chief Mike Gowen, Arlington County Fire Department

PROFILE OF NEW KENT, VA
David Park, Joseph Kim, David Hinkle, Lata Kodali (Virginia Tech) with Da
Sponsor: Carl Frick, Virginia Corporate Extension (VCE) representative

CREATING SYNTHETIC DATA FOR VIRGINIA LONGITUDINAL DATA SYSTEM
Sean Pili, Kyle Morgan, Ronnie Fesco, and Lata Kodali (Virginia Tech) with Aar… der (SDAL)
Sponsor: Tod Massa (SCHEV – State Council for Higher Education in Virginia)

DEFINING AND MEASURING EQUITY IN ALEXANDRIA, VA
Sponsor: Emily Molfino, City of Alexandria

PROFILING ARMY BASES
Goal: Identify publicly available data sources (e.g. Census and BLS data) to create social, demographic, economic, and other quantitative profiles of Army bases and their surrounding areas. Identify relevant variables for use in statistical models.
Sponsors: Greg Ruark, Andrew Slaughter, US Army Research Institute for Behavioral & Social Science Research

DISCOVERING NON-TRADITIONAL DATA SOURCES FOR BUSINESS INNOVATION
Harsimrat Pandher (VT), David Park (VT), Daniel Wilkin (VT), Joseph Kim (VT)
Claire Kelling (PSU) with Gizem Korkmaz and Stephanie Shipp (SDAL)
Sponsor: Gary Anderson, The National Center for Science & Engineering Statistics

A STUDY ON WMATA BUS FARE EVASION
Sponsor: Jayme M. Johnson, Catherine Vanderwaart
Washington Metropolitan Area Transit Authority

ANALYZING THE ECONOMIC IMPACT AND SOCIAL INTEGRATION OF REFUGEES IN ROANOKE, VIRGINIA
Claire Kelling (PSU), Kyle Morgan (VT), Craig Morton (VT), Hannah Brinkley (VT), Adrienne Rogers (VT), with Mark Orr, Stephanie Shipp, and Bianica Pires (SDAL)

Concluding Remarks

- We are at the forefront of creating the Science of ***All*** Data
- Without applications (problems) data science would not exist
- Our research is driven by Other People's Problems
- Our vision is to bring the ***All*** Data Revolution to ***all*** organizations –local, state, federal government, industry, and non-profit organizations

Thank You