Welcome Data Science for the Public Good 2018 Cohort

May 29, 2018

Sallie Keller
Professor of Statistics and Director

BIOCOMPLEXITY INSTITUTE
VIRGINIA TECH

SDAL SOCIAL & DECISION ANALYTICS LABORATORY


VirginiaTech
Biocomplexity Institute


BIOCOMPLEXITY
INSTITUTE

We are in an *ALL* Data Revolution

A new lens for social observing

Infrastructure

- Condition
- Operations
- Resilience
- Sustainability

Environment

- Climate
- Pollution
- Noise
- Flora/ Fauna

People

- Relationships
- Location
- Economic Condition
- Communication
- Health

It is time to leverage *ALL* the data sources

Local, State/Provence, and Federal

Designed Data

Administrative Data

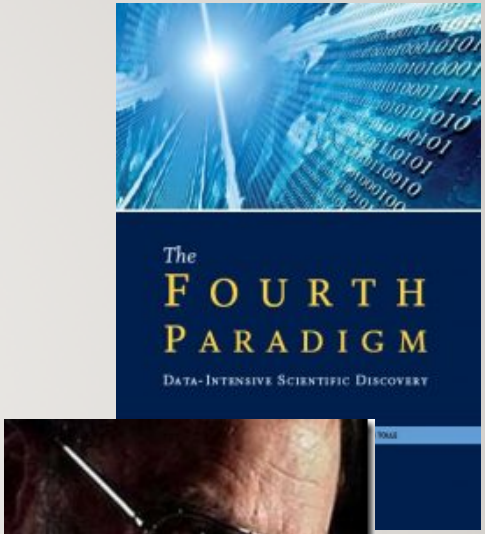
Opportunity Data

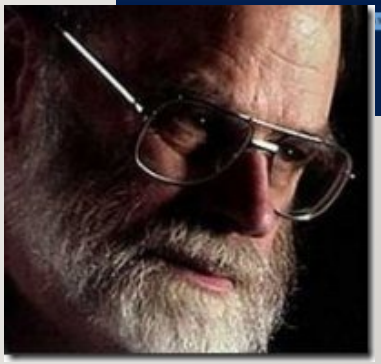
Procedural Data

What is Data Science?

Fourth paradigm

> "… change of all sciences moving from observational, to theoretical, to computational and now to the 4th Paradigm – Data-Intensive Scientific Discovery"



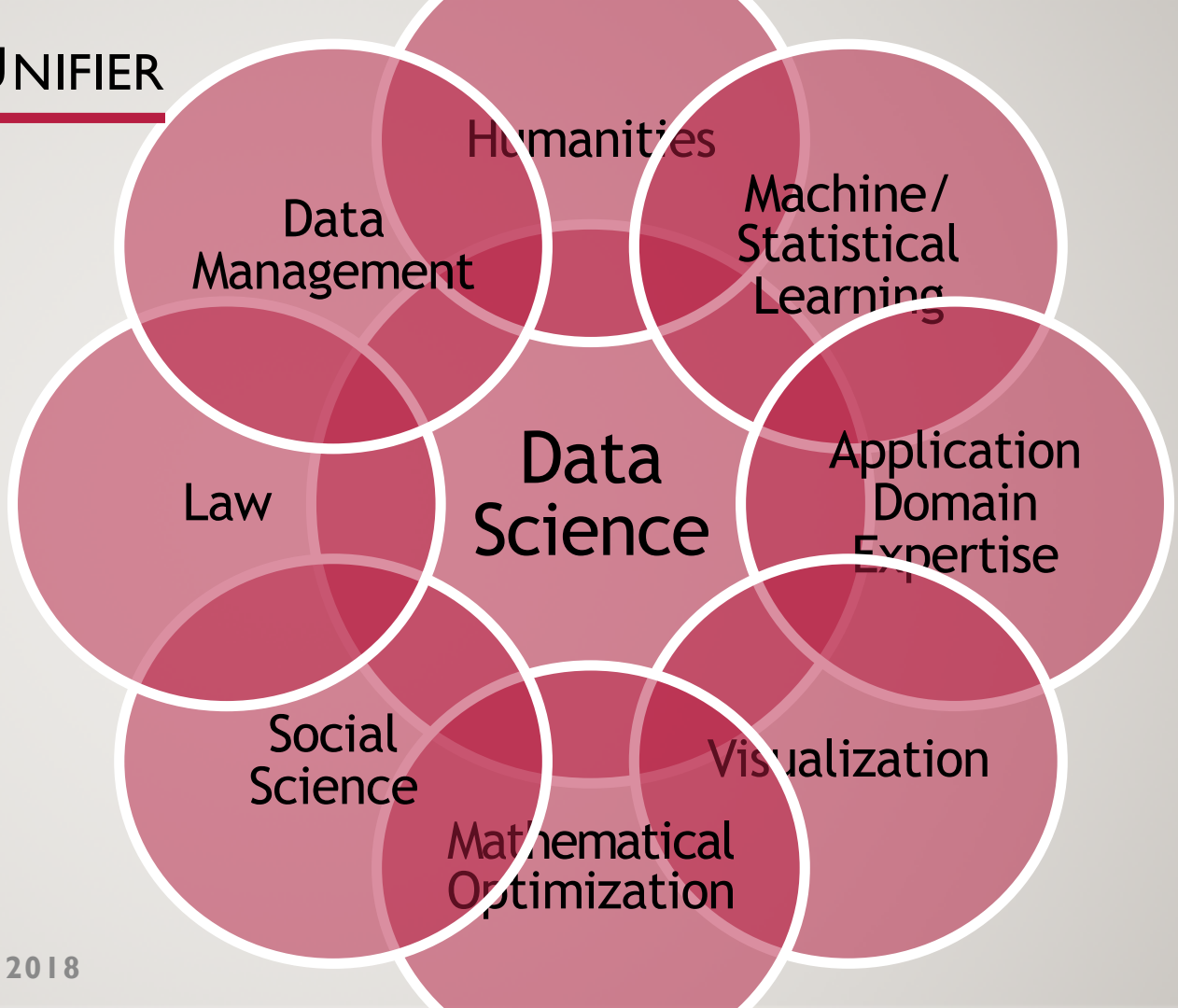


Jim Gray

What is Important?

Need to solve a real problem using data…
No applications, no data science.

Data Science as a Unifier



In our lab data science is policy focused on other people's problems

NCSES National Center for Science and Engineering Statistics

Local / State Government

Federal Statistical Agencies

Department of Defense

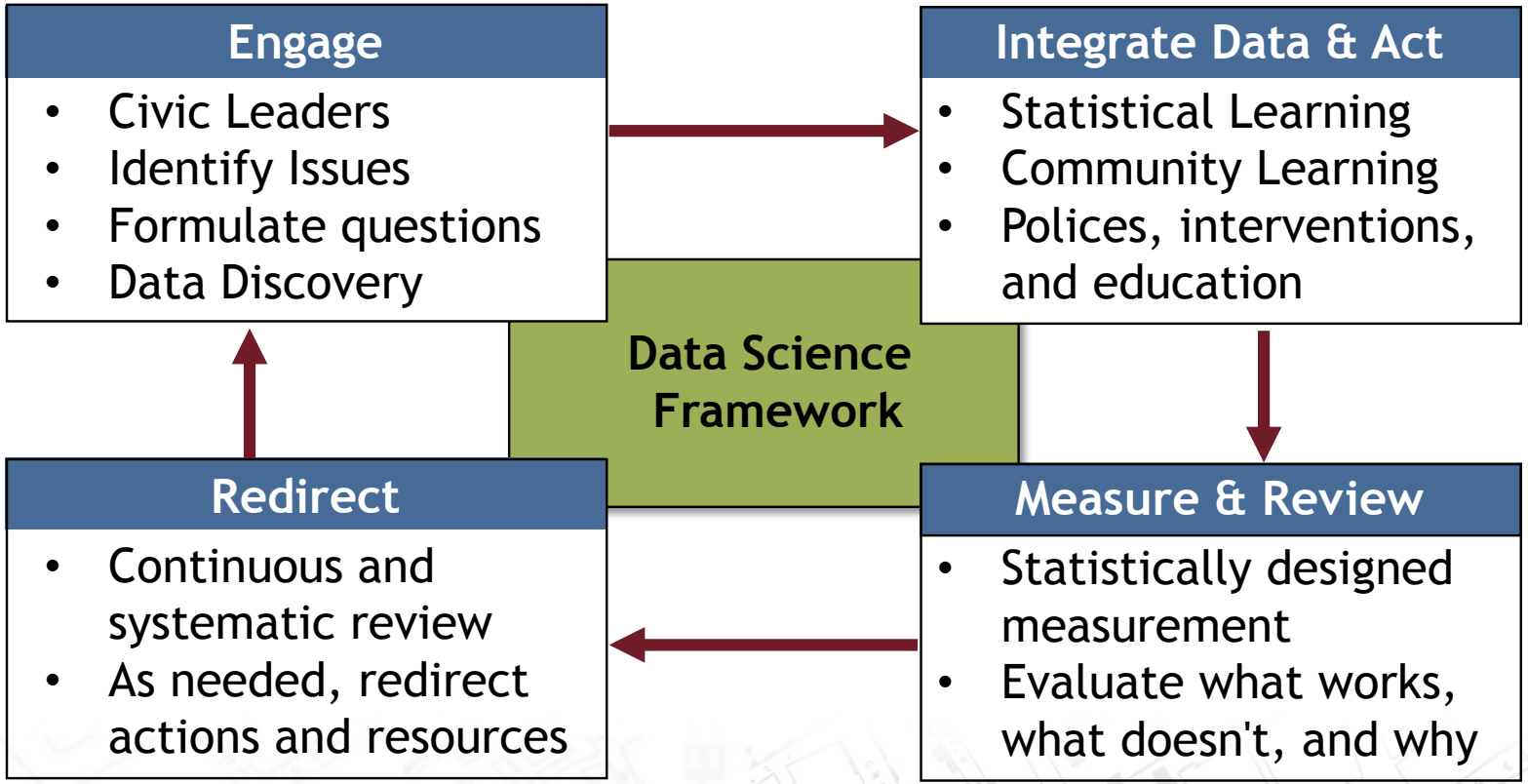
Industry

MITRE

P&G

Enhancing Prosperity through Data Science

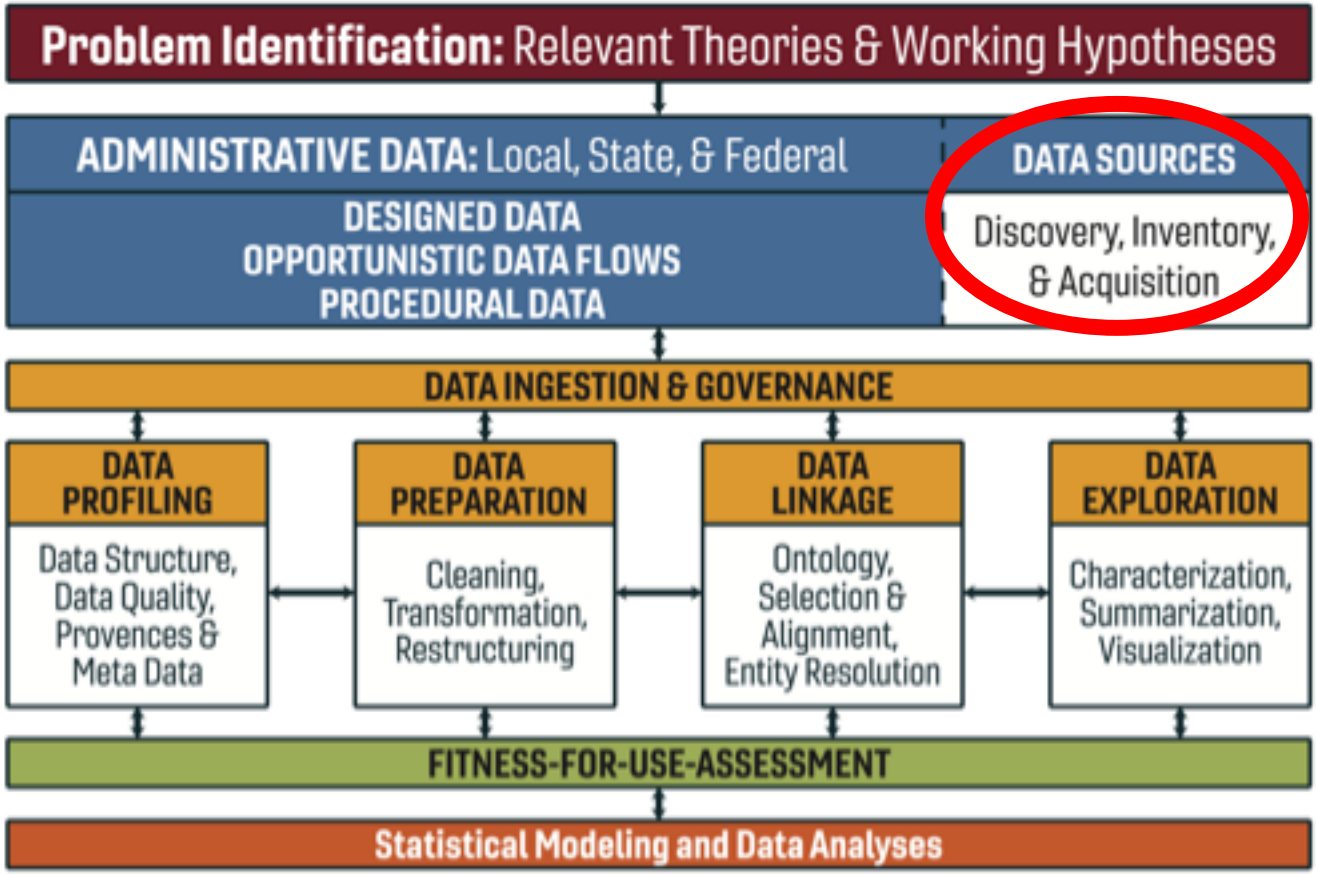
Community Learning through Data-Driven Discovery (CLD3)



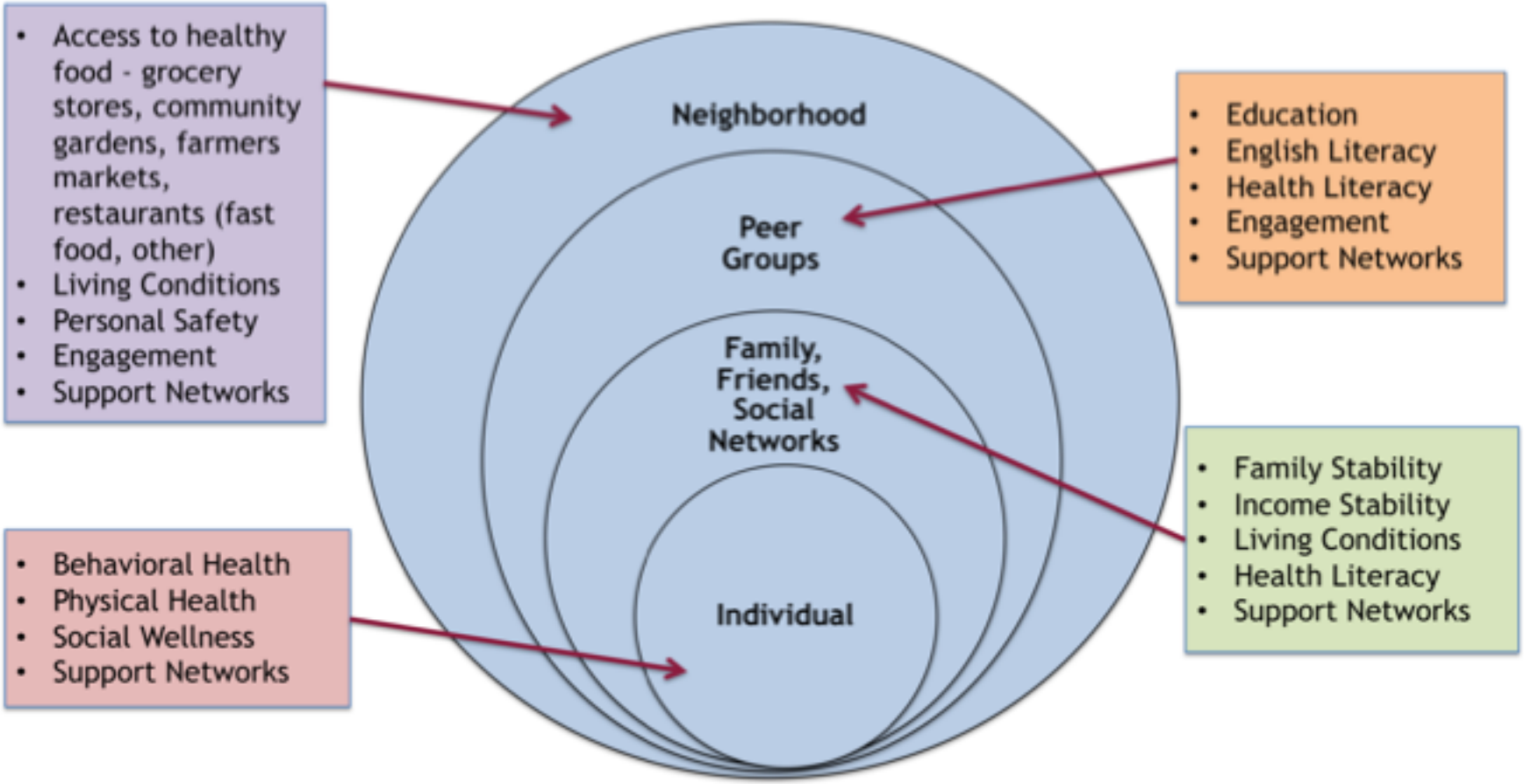
Common CLD3 themes need data science training to address

- **Locating** and **describing** a population within a community
- **Estimating** a statistical summary and its margin of error to evaluate its usefulness **for the purpose at hand**
- **Forecasting** future needs
- **Evaluating** a program, policy, or standard operating procedure

Data Science Framework



Local community Data Map

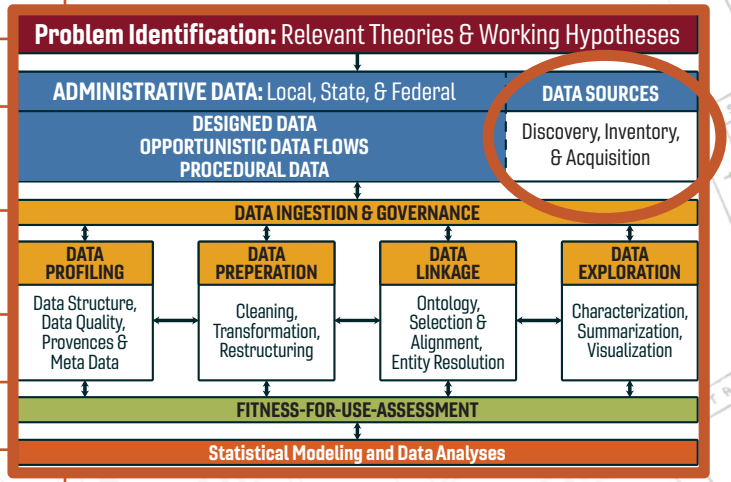


Data Discovery, Inventory & Acquisition

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 8th, 10th, 12th 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



Data Discovery, Inventory, & Acquisition

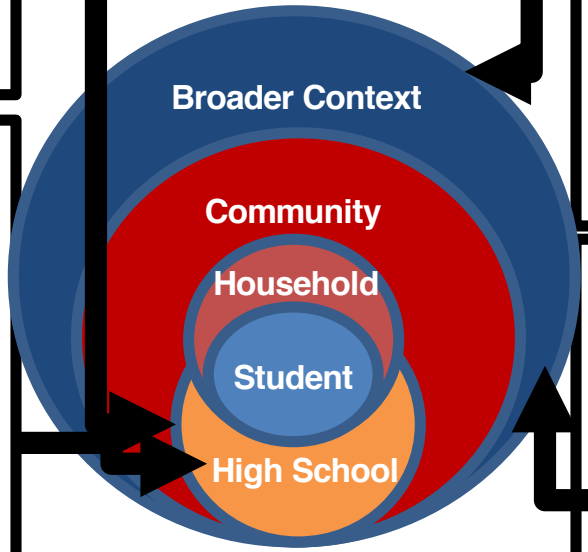
Data Map

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)
- ....



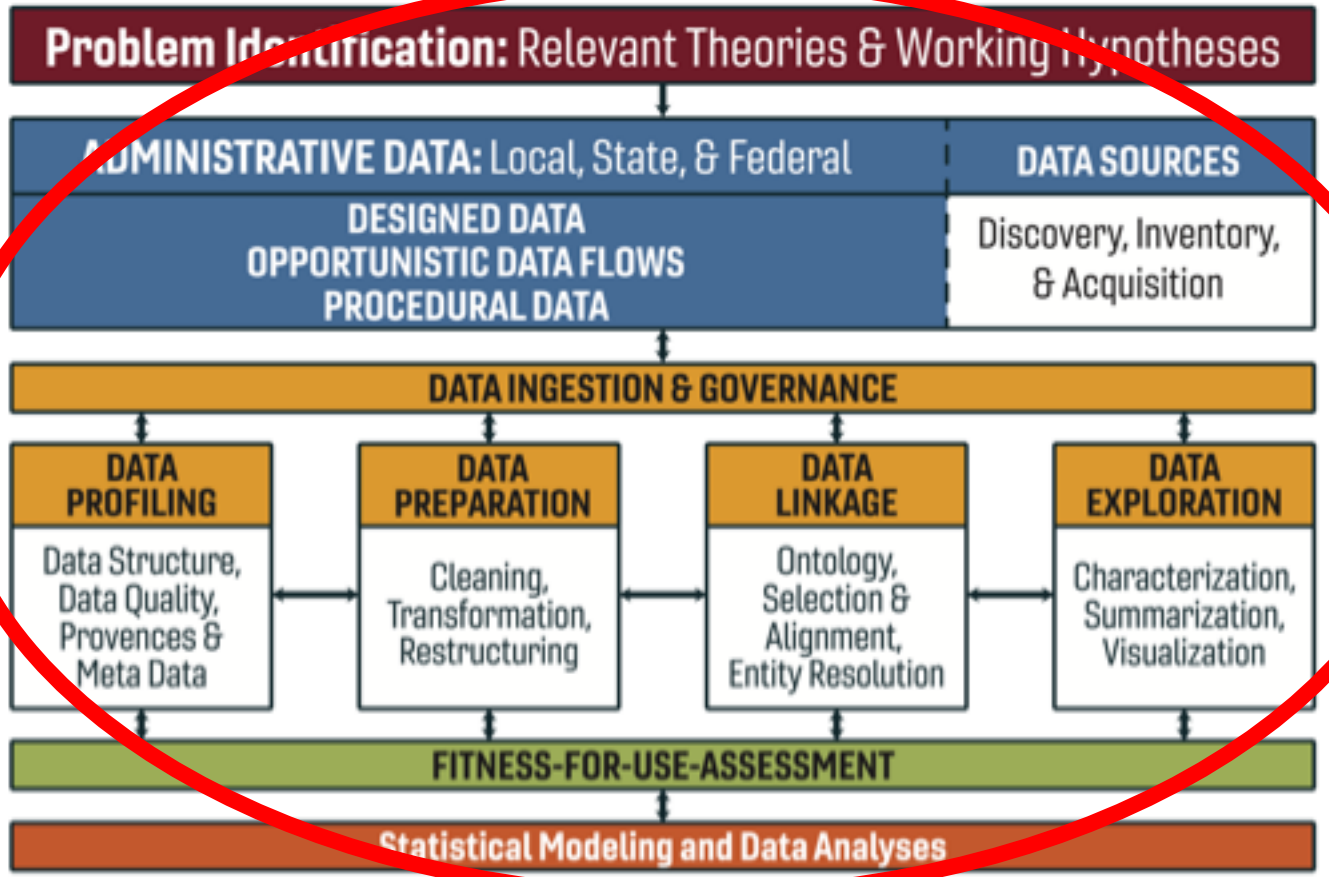
Community Characteristics

- % 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)

Data Science Framework



Parks and Recreation Participation

*Locating populations and **Evaluating a Policy***

Issue: In **Arlington Virginia**, the Department of Parks & Recreation (DPR) policy provides fee reductions to households that meet certain economic criteria

Goal: DPR wants to ensure that **fees are not a barrier** to program participation by targeting outreach

The DATA

Household Level

- Department of Parks & Recreation 2016 enrollment data for all programs
- Geocoded household locations

Census Tract Block Group Level

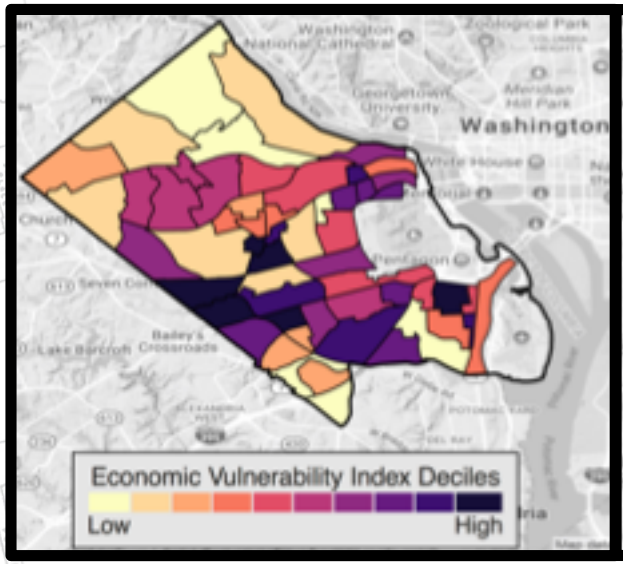
- 5-year 2015 American Community Survey – household level measures of economic vulnerability

School Level

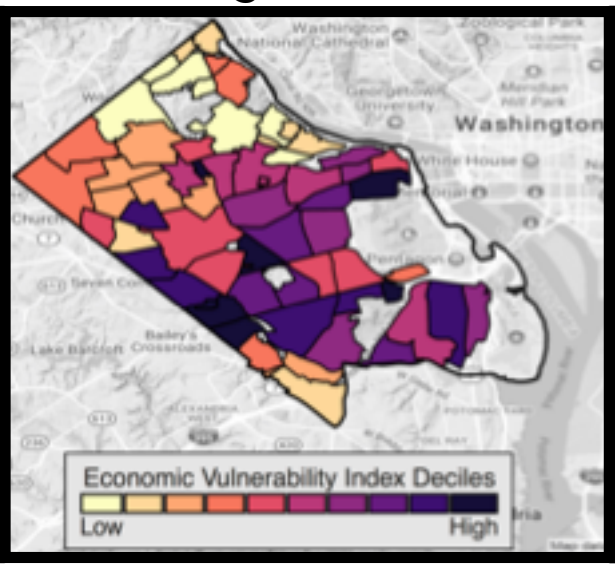
- National Center for Education Statistics
- Center for Disease Control 2014-15
- Virginia Department of Education – student level measures of economic vulnerability aggregated by school

Arlington County Vulnerability Indicators

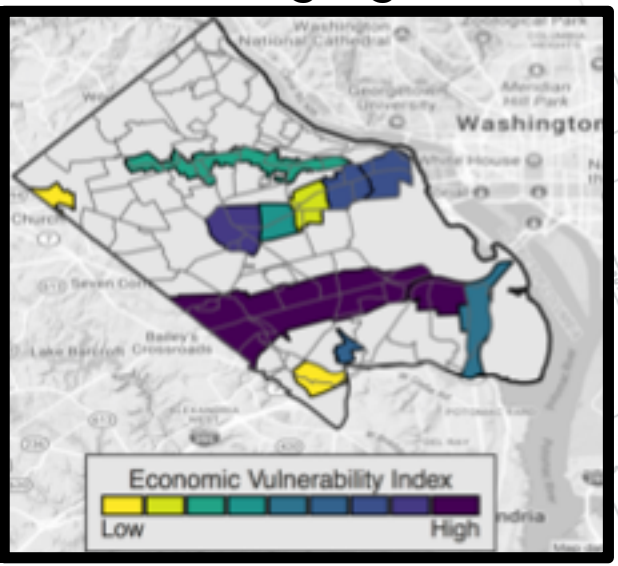
Census Tracts



Civic Association Neighborhoods



High-Density Planning Regions



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

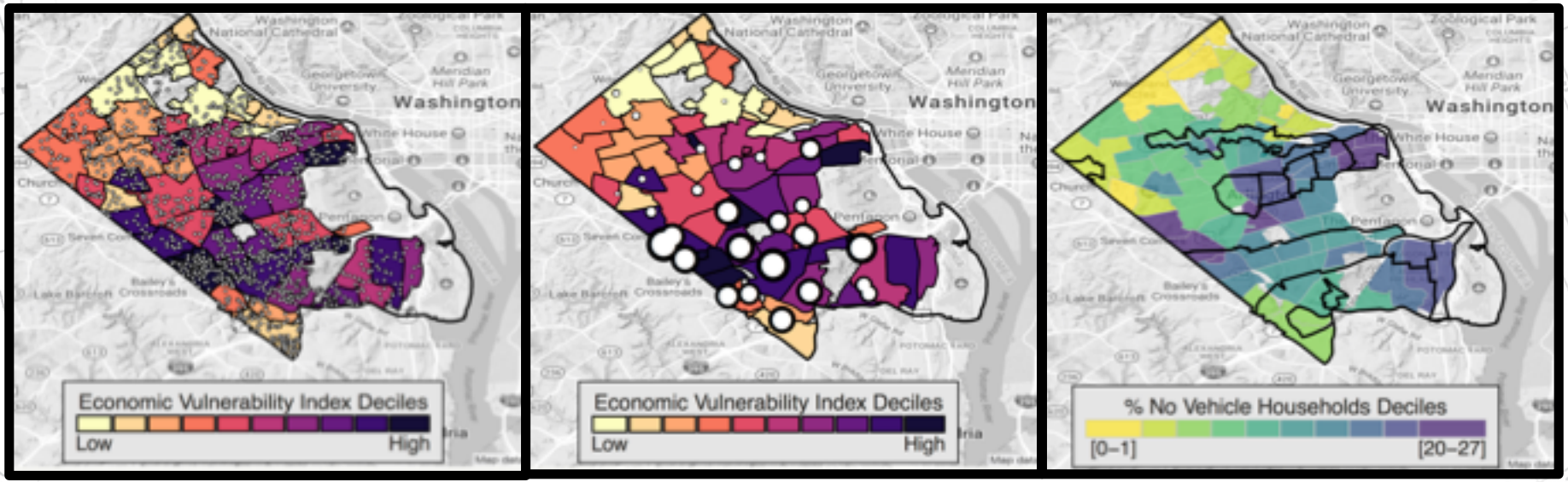
Source: American Community Survey 2012-20156 aligned to geographic areas using SDAL Synthetic Technology.

Arlington County Neighborhood Insights

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

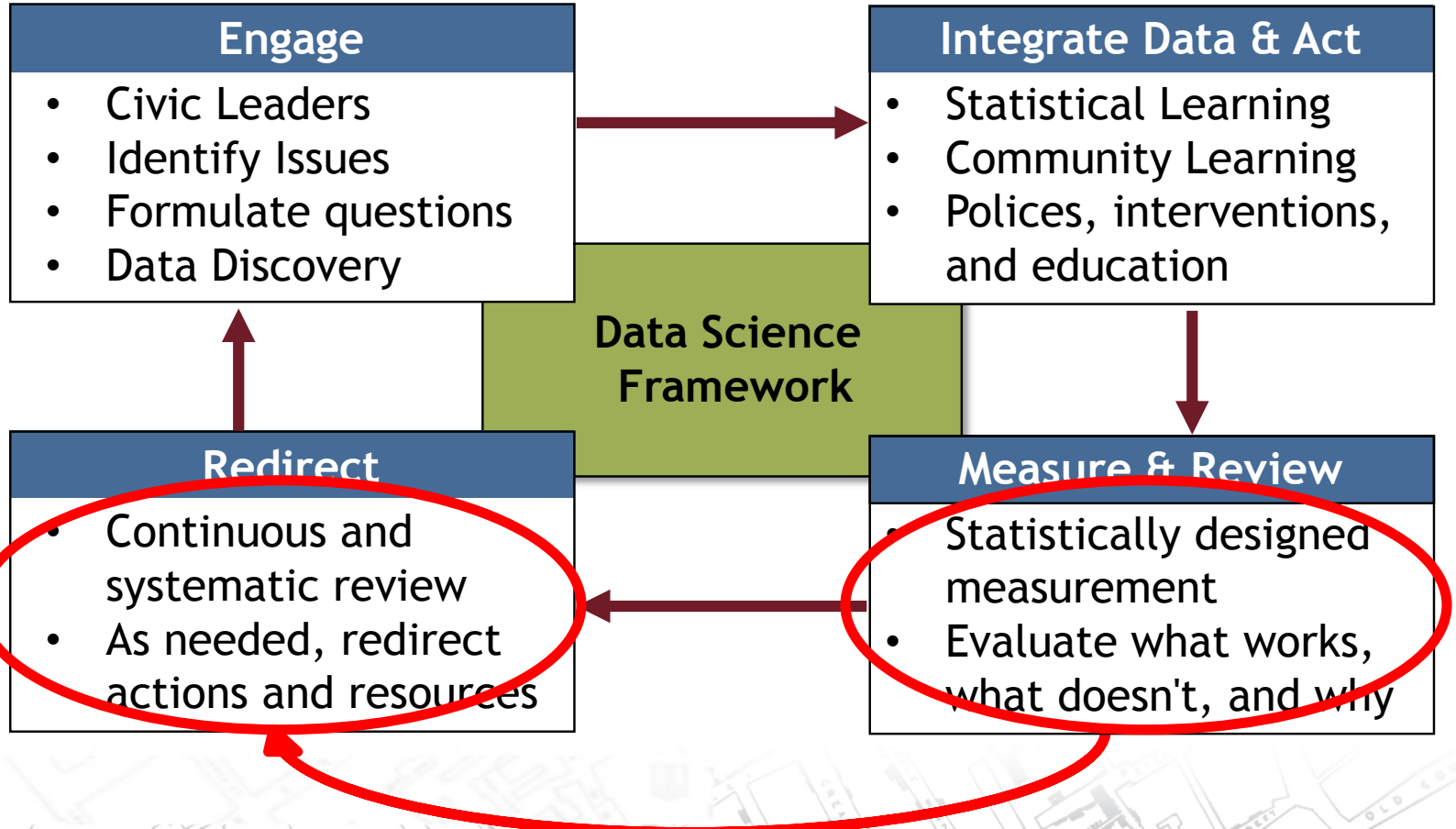
School and neighborhood vulnerability indices

High-Density Planning Regions with % households with no vehicles



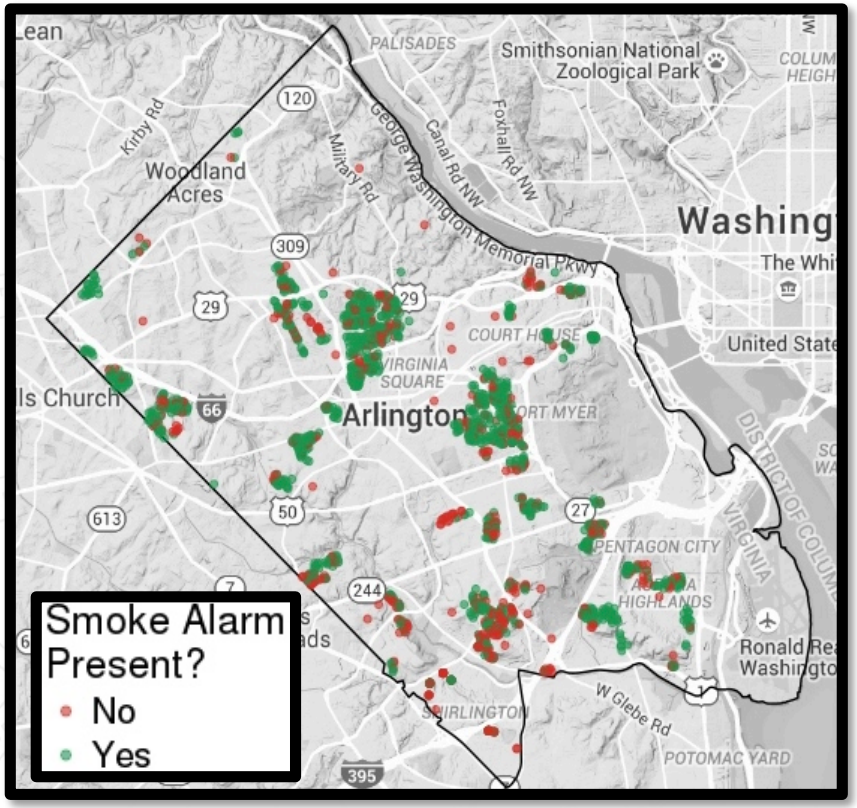
Sources: ACS 2012-2016; NCES, CDC, and VDOE 2014-2015; Arlington County Department of Parks & Recreation 2016.

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

Arlington Operation FireSafe

Evaluating and Informing a Program



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 working smoke detector

Goal: Construct a model to estimate for each single family home the probability it has adequate smoke detectors

The DATA

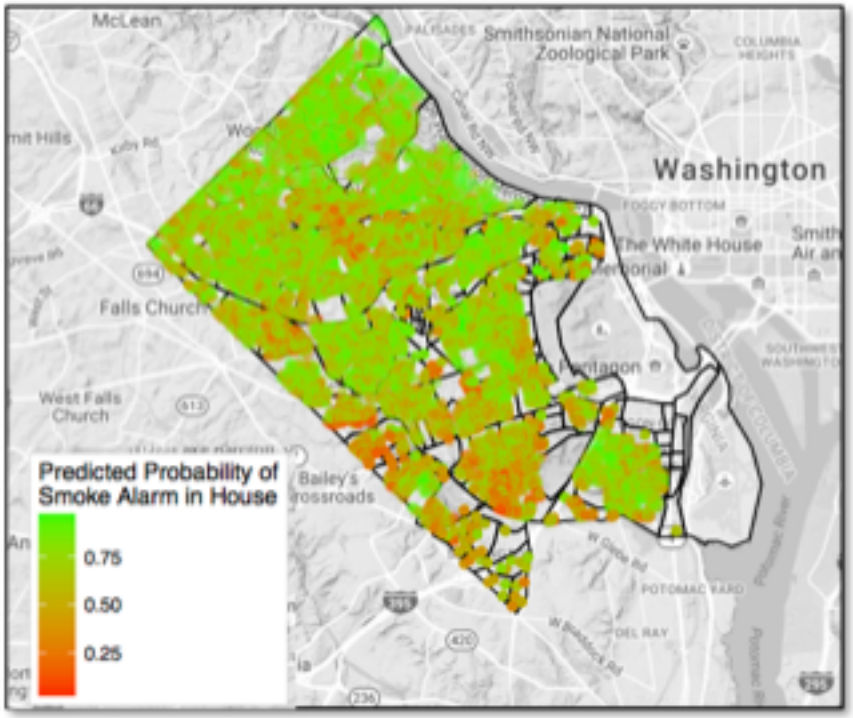
Household Level

- Operation FireSafe data on the location of the 5,623 single family homes visited, time and date of the visit, and the outcome installed smoke detector; working smoke detector in place
- Real estate tax assessments for 60,343 single family homes which includes tenure, home age, value, size, and number of bedrooms
- Geocoded the single family home locations

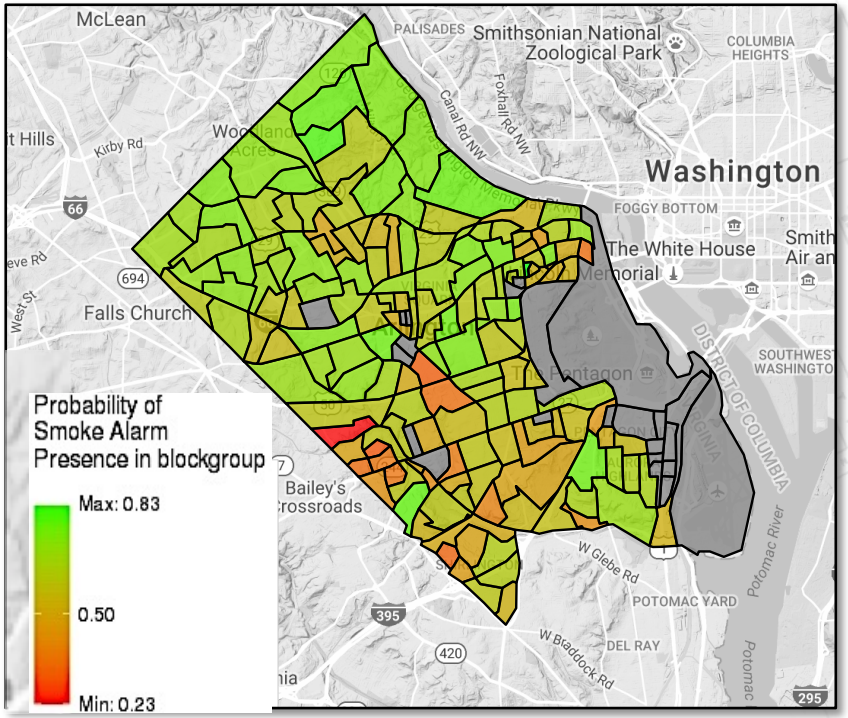
Census Tract Block Group

- 5-year 2015 American Community Survey – household level demographic and socioeconomic data

Probability of Having a Home Smoke Detector

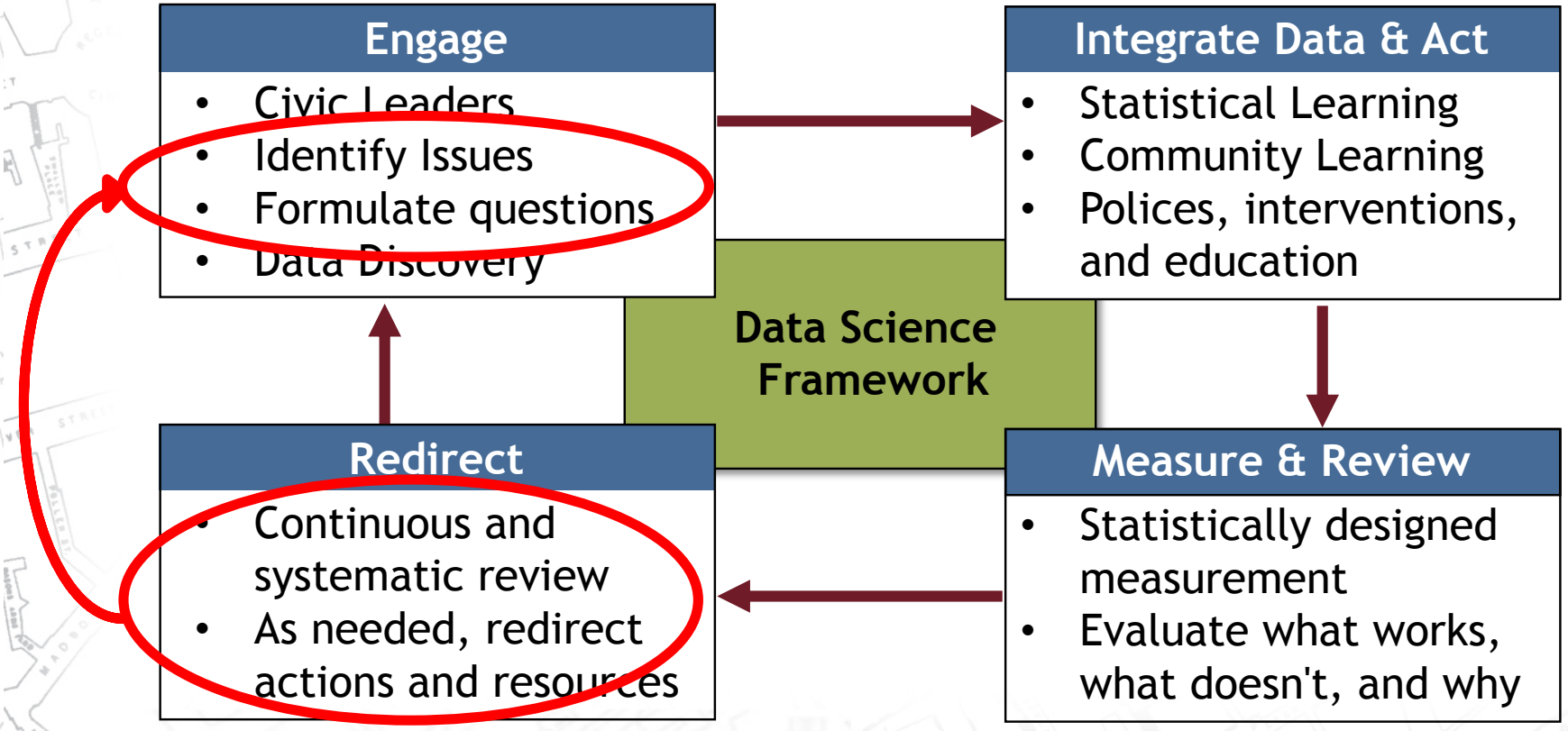


Housing Unit Level Predictions



Census Block Level Predictions

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

Data Science for the Public Good 2017

IDENTIFYING STEM EDUCATION PATHWAYS

EXPLORING MENTAL HEALTH SERVICES FOR FAIRFAX COUNTY YOUTH

RESIDENTIAL SMOKE ALARM NEED IN ARLINGTON COUNTY

HOW DO EVENTS AFFECT CRIME?

MODELING THE IMPACT OF OPEN SOURCE SOFTWARE: NETWORK OF A PACKAGES

MODELING RESPONSE TIME FOR STRUCTURE FIRES

PROFILE OF NEW KENT, VA

CREATING SYNTHETIC DATA FOR VIRGINIA LONGITUDINAL DATA SYSTEM

DEFINING AND MEASURING EQUITY IN ALEXANDRIA, VA

PROFILING ARMY BASES

DISCOVERING NON-TRADITIONAL DATA SOURCES FOR BUSINESS INNOVATION

A STUDY ON WMATA BUS FARE EVASION

ANALYZING THE ECONOMIC IMPACT AND SOCIAL INTEGRATION OF REFUGEES IN ROANOKE, VIRGINIA

Thank You