Data Science for the Public Good Discussion & Coding

2019 Midwest Big Data Summer School

Aaron D. Schroeder, PhD
Social Decision Analytics Division
Bioinformatics Institute, University of Virginia

University of Virginia

We study policy-focused problems

NCSES National Center for Science and Engineering Statistics

MITRE

P&G

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 & Engineering Statistics

Department of Defense

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

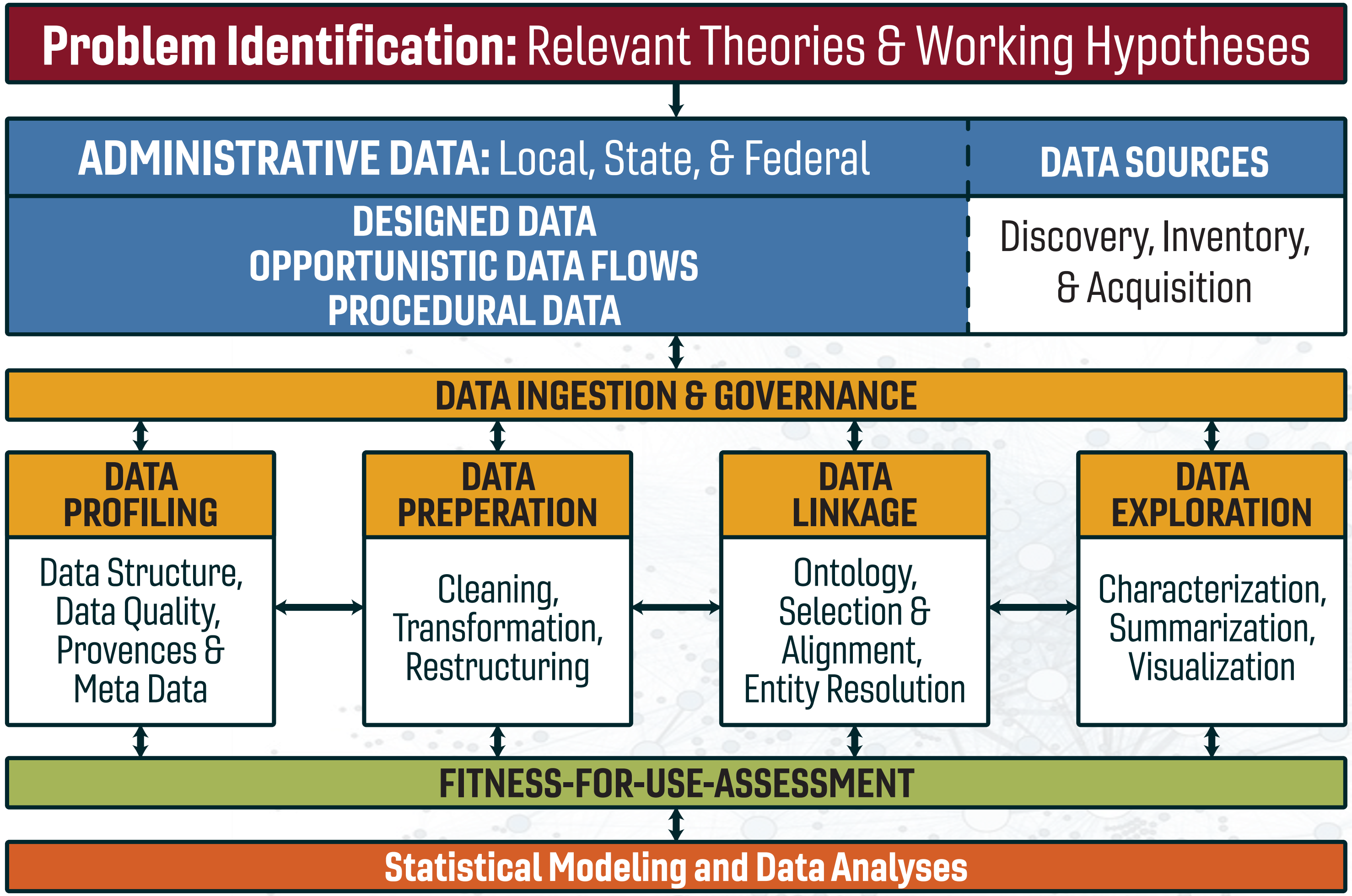
Industry

MITRE Corporation
Procter & Gamble

University of Virginia

- SDAD Web Page

Our Data Science Framework



Keller, S.A., S. Shipp, G. Korkmaz, E. Molfino, J. Goldstein, V. Lancaster, B. Pires, D. Higdon, D. Chen, and A. Schroeder. 2018. "Harnessing the Power of Data to Support Community-Based Research." WIREs Computational Statistics. e1426.

Keller, S.A., G. Korkmaz, M. Orr, A. Schroeder, S. Shipp. 2017. "The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches." Annual Review of Statistics and Its Applications 2017. 4:5.1-5.24.

University of Virginia

Continuous Process of Engagment

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

Integrate Data & Act
- Statistical Learning
- Community Learning
- Polices, interventions, and education

Data Science Framework

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

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

S. Keller, V. Lancaster, S. Shipp. (2017). Building Capacity for Data-Driven Governance: Creating a New Foundation for Democracy. *Statistics and Public Policy*, September.

University of Virginia

Observations from Our Work Thus Far

Issues that have arisen and repeat from our community engagements thus far:

- Locating and describing a population within a community
- 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

Research challenges that are emerging through our work:

- Formalization and automation of Data Science Framework
- Data integration, analysis, and linkage across multiple levels of data support
- Data and corresponding estimation redistribution across multiple geographies
- Composite indices development and alignment with issues

Data Science for the Public Good (DSPG)

- Experiential learning program through which participants learn everything they need to know to get started working as a data scientist in policy-oriented positions
- Creating Public Good-Oriented Data Scientists, or “Policy Scientists”
- Learning not just how to do the work, but also how to work across disciplines on problem-solving teams
- For the graduate fellows, also learning how to run these teams, managing both up an down
- All started when asked to teach ”Research Methods” for Masters of Public Administration students

UNIVERSITY of VIRGINIA

Data Science for the Public Good (DSPG)

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

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

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)

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

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

UNIVERSITY of VIRGINIA

DSPG Experience Goals

Engage in meaningful research focused on real-world problems and addresses social policy

Work in partnership with government sponsors and policy makers

Learn essential tools for scientific and statistical computing, including R, Python, Databases, GIS, and other software tools as needed for projects

Learn the entire Data Science for Policy process, from stakeholder engagement and problem definition to the delivery of policy impacting analyses, tools and data products

Work on truly multi-disciplinary project teams comprising diverse levels of experience (students, post-docs, researchers, and federal leaders) and a broad range of academic perspectives (data science, statistics, economics, sociology, psychology, geography, and others)

Directly interact with policy leaders and government agencies through field trips and regular policy events on Capitol Hill and the surrounding area

Engage with leading policy practitioners and researchers through SDAD's Visiting Scholar program, which includes economists, statisticians, and other social science researchers with diverse public and private sector experience

University of Virginia

DSPG Annual Symposium, Speaker Series

JOIN US

at these Data Science for the Public Good FORUM events

THE UNIVERSITY OF VIRGINIA'S Biocomplexity Institute, Darden School of Business, and Data Science Institute in partnership with the Northern Virginia Technology Council announce the Data Science for the Public Good (DSPG) Forum, an opportunity for civic engagement through two Forum events, including a Distinguished Speaker Series and Annual Symposium.

These events will bring together key public and private stakeholders to discuss how "doing data science" can support evidence-based policymaking and innovation to enhance the quality of life where we all live, learn, work, and play.

LET THE CONVERSATIONS BEGIN: REGISTER NOW

DISTINGUISHED SPEAKER SERIES

Former Governor of Maryland **Martin O'Malley** kicks off the series with his talk, *Smart Government: The Data, the Map, and the Method.*

DATE: June 14, 2019 | **TALK AND Q&A SESSION:** 4:00pm - 5:00pm | **RECEPTION:** 5:00pm - 6:30pm

ANNUAL SYMPOSIUM

The 4th Annual Symposium celebrates the Biocomplexity Institute's 2019 DSPG Young Scholars and their research. Keynote speakers include: **Phil Bourne,** Director of the University of Virginia's Data Science Institute and Acting Dean of the School of Data Science; **Ron Jarmin,** Deputy Director and Chief Operating Officer of the U.S. Census Bureau.

DATE: August 9, 2019 | **TIME:** 1:00pm - 4:30pm

LOCATION OF EVENTS: University of Virginia Darden School of Business Sands Family Grounds, 1100 Wilson Blvd., 30th Floor, Arlington, VA 22209.

REGISTRATION for these events and more information about the Data Science for the Public Good Forum is available at biocomplexity.virginia.edu/events.

nvtc Northern Virginia Technology Council | University of Virginia Biocomplexity Institute | University of Virginia Darden School of Business | University of Virginia Data Science Institute

Data Science for the Public Good Forum

Data Science for the Public Good

The University of Virginia's Biocomplexity Institute, Darden School of Business, and Data Science Institute in partnership with the Northern Virginia Technology Council announce the Data Science for the Public Good Forum, an opportunity for civic engagement, comprised of a Distinguished Speaker Series and Annual Symposium.

The Data Science for the Public Good (DSPG) Forum builds on the Biocomplexity Institute's highly successful and distinctive DSPG research platform that leverages cutting-edge public policy analytics and unprecedented data access, and serves as an honest broker to identify, visualize, and understand the full complement of interests that affect the public good.

As central programs of the Forum, the Distinguished Speaker Series and Annual Symposium will bring together key public and private stakeholders to discuss how "doing data science" can support evidence-based policymaking and innovation to enhance the quality of life where we all live, learn, work, and play.

We can't wait to begin these important conversations with you!

DATA SCIENCE FOR THE PUBLIC GOOD DISTINGUISHED SPEAKER SERIES

June 14, 2019

The former Governor of Maryland, Martin O'Malley, will be our inaugural speaker and kick off the Data Science for the Public Good Distinguished Speaker Series with his talk, *Smart Government: The Data, the Map, and the Method.*

Date: June 14, 2019

Talk and Q&A Session: 4:00-5:00pm
Reception: 5:00-6:30pm

Location: University of Virginia Darden School of Business Sands Family Grounds
1100 Wilson Boulevard, 30th Floor
Arlington, VA 22209

Register for the Distinguished Speaker Series.

DATA SCIENCE FOR THE PUBLIC GOOD ANNUAL SYMPOSIUM

August 9, 2019

In its fourth year, the Data Science for the Public Good Annual Symposium will celebrate the Biocomplexity Institute's 2019 Young Scholars and their research. Speakers will include Phil Bourne, Director of the University of Virginia's Data Science Institute and Acting Dean of the School of Data Science, and Ron Jarmin, Deputy Director and Chief Operating Officer of the U.S. Census Bureau.

Date: August 9, 2019

Symposium: 1:00-4:30pm

UNIVERSITY of VIRGINIA

UNIVERSITY of VIRGINIA

DSPG Syllabus

- [DSPG 2019 Program Schedule]

Projects - WMATA

- **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
- **Analysis Plan:** Using **WMATA administrative data** locate where fare evaders live and **American Community Survey**
- to tell their story at the census
- block group level

UNIVERSITY of VIRGINIA

General Observations about Fare Evasion

- Widespread national and international problem
- Typically estimated using observer surveys and not using administrative data
    - 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

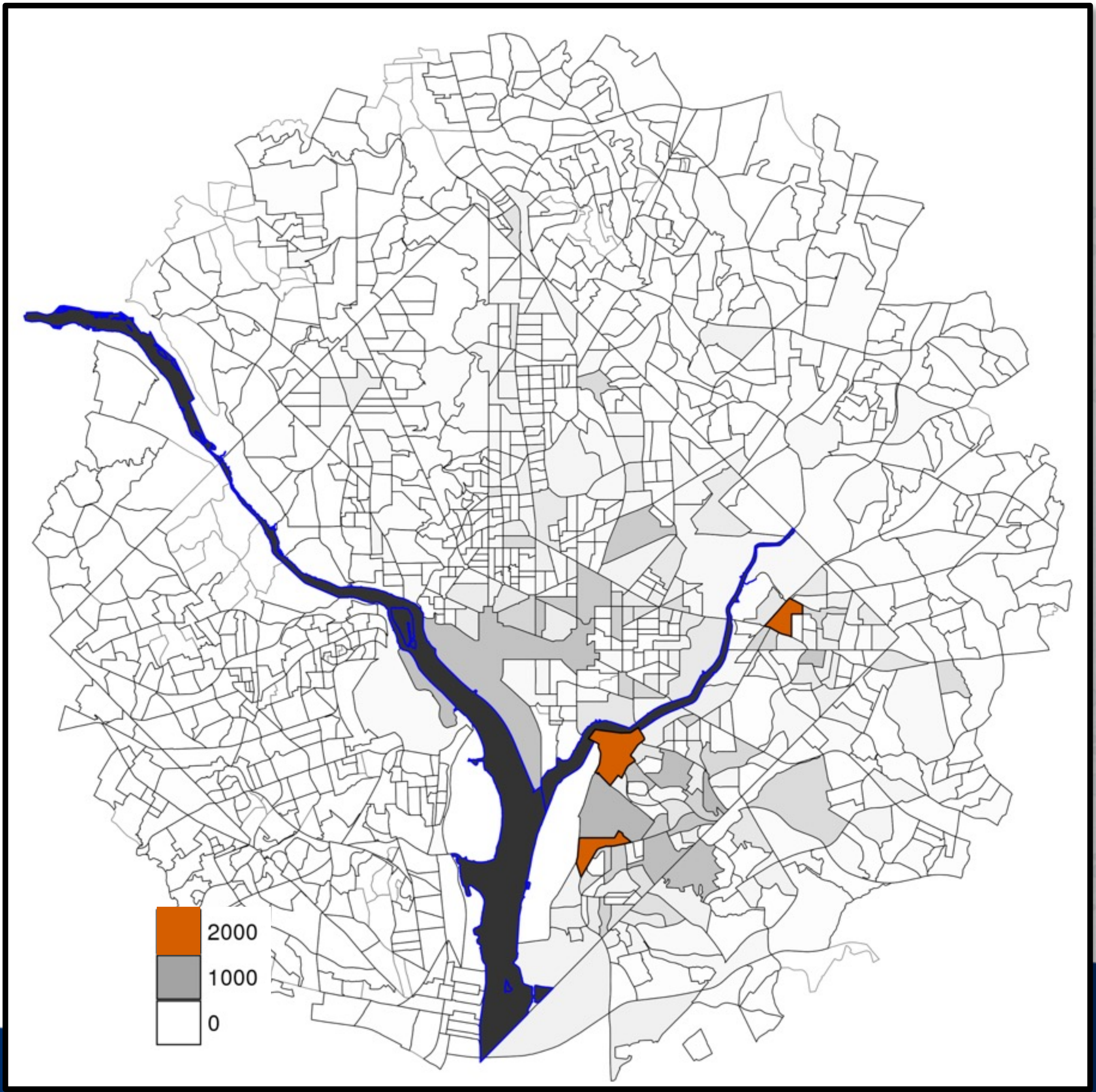
WMATA Administrative Data Sources

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

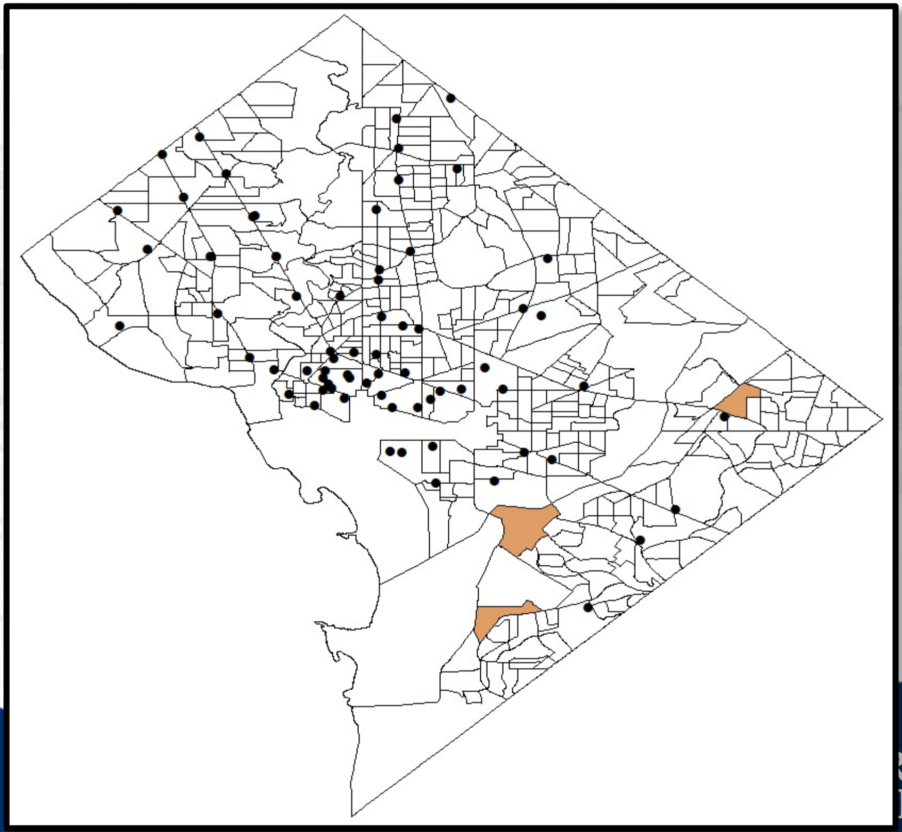
- **Bus Stops (10,988 observations):** Contains the stop ID, stop name, latitude, and longitude
- **Automated Person Counter (APC):** Contains front & back door entries and exits for a bus, route, trip number and bus stop
- **Farebox:** Contains cash & SmarTrip transactions for a bus, trip number, & 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
- Could this contribute to fare evasion?

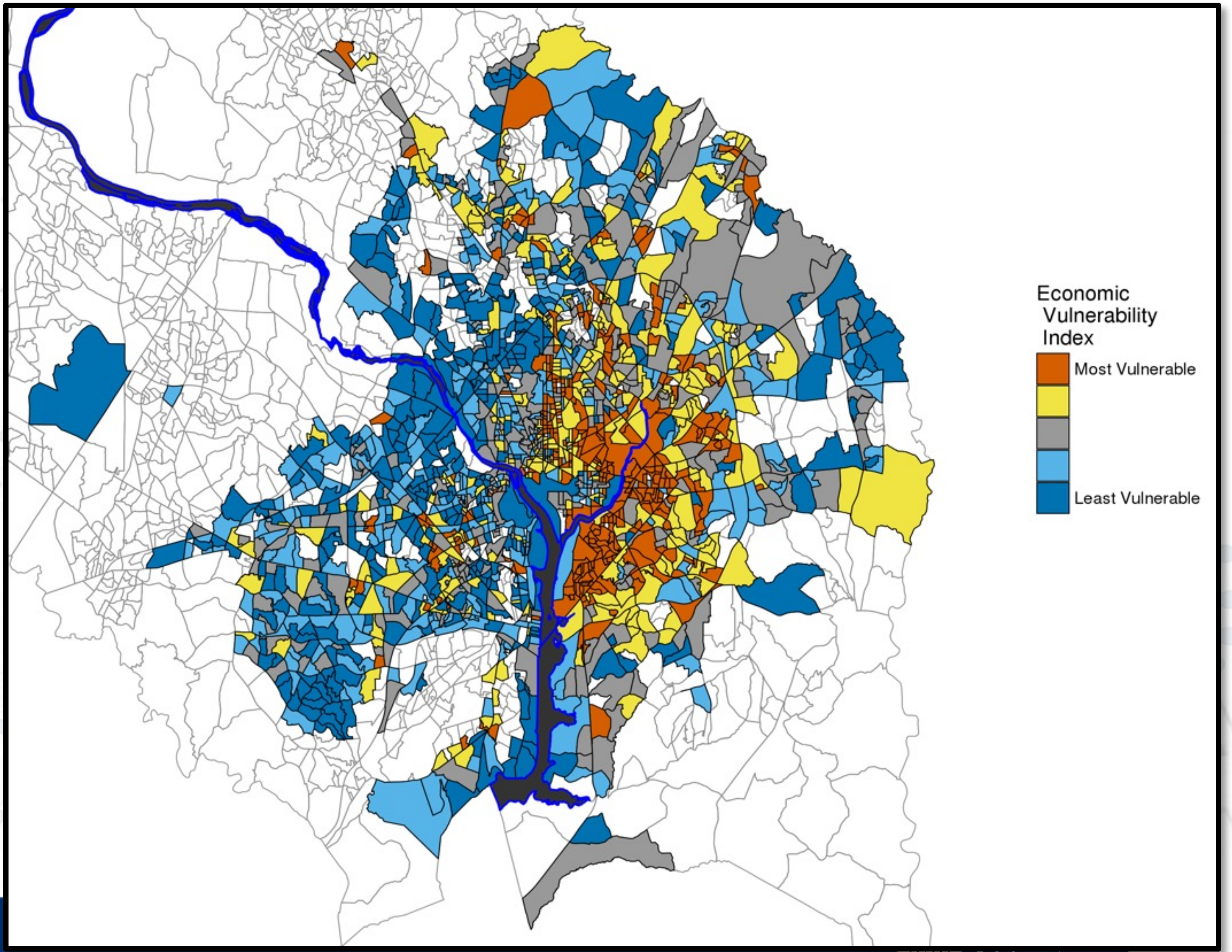


Economic Vulnerability Index

Composite economic vulnerability index by **census block groups** with bus stops in the seven WMATA jurisdictions

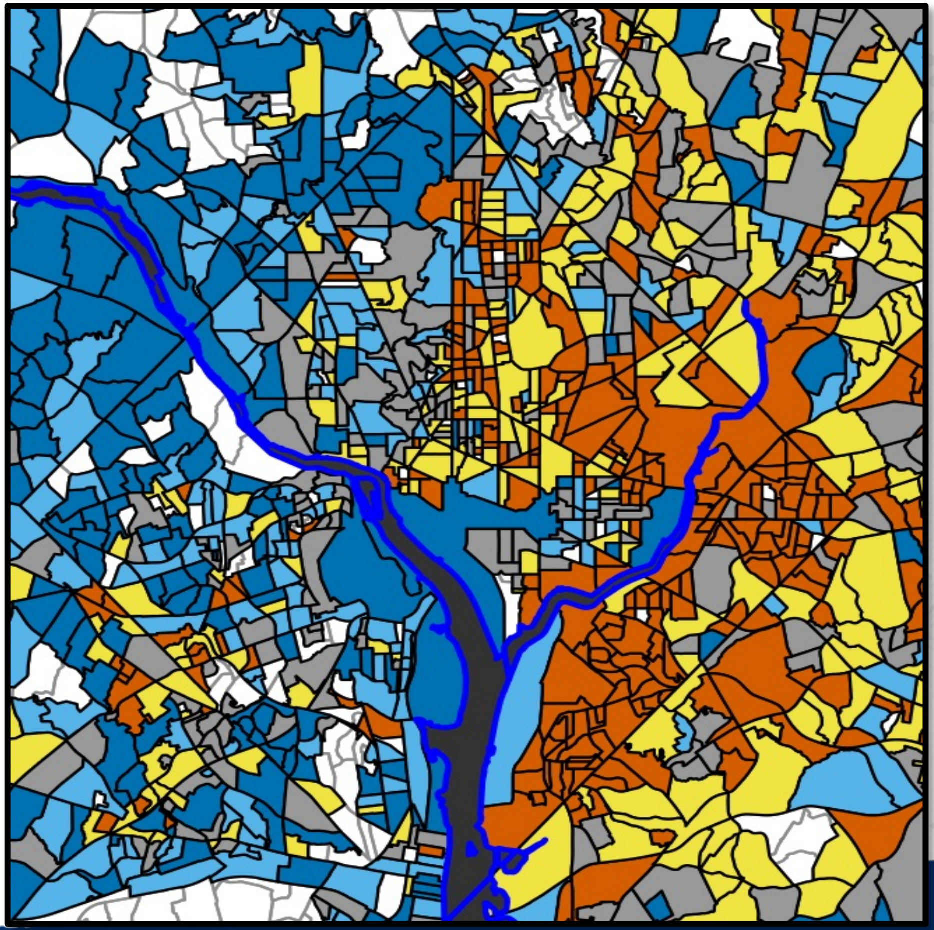
The composite index was constructed using ACS (2015) variables:

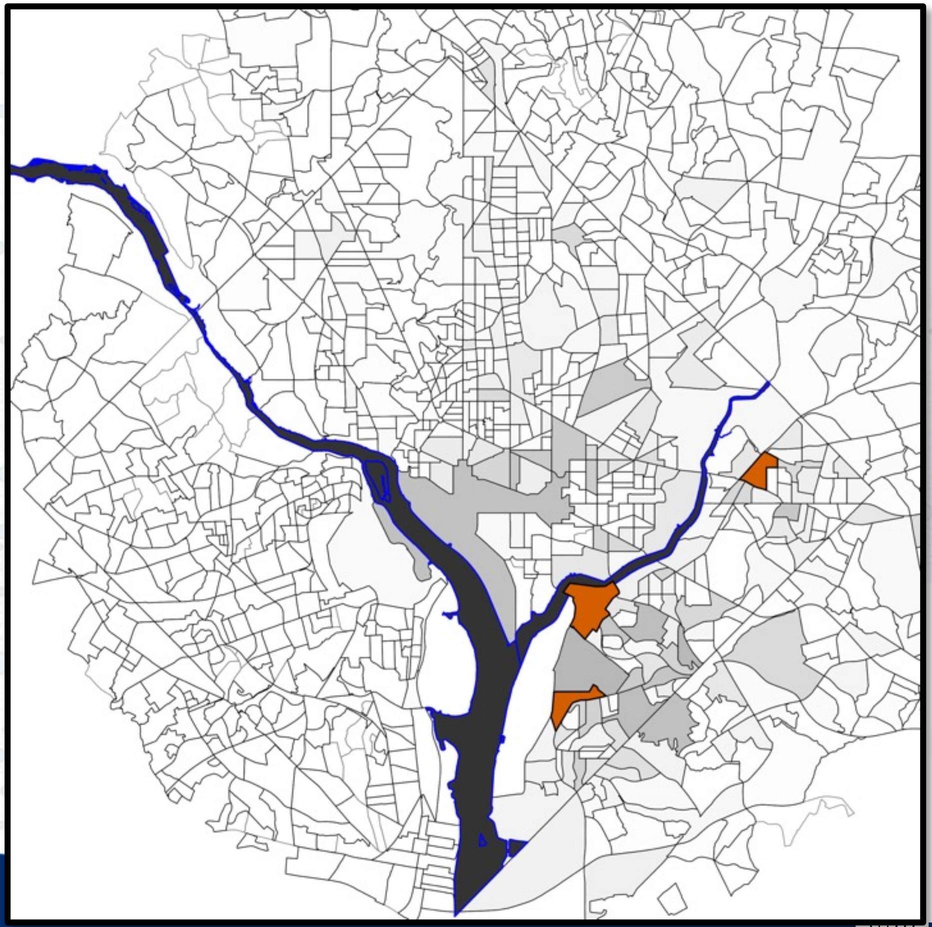
- % households in poverty (Federal)
- % households with no vehicle
- % households qualifying for SNAP
- % households with housing burden > 50%



Insights for Potential Interventions

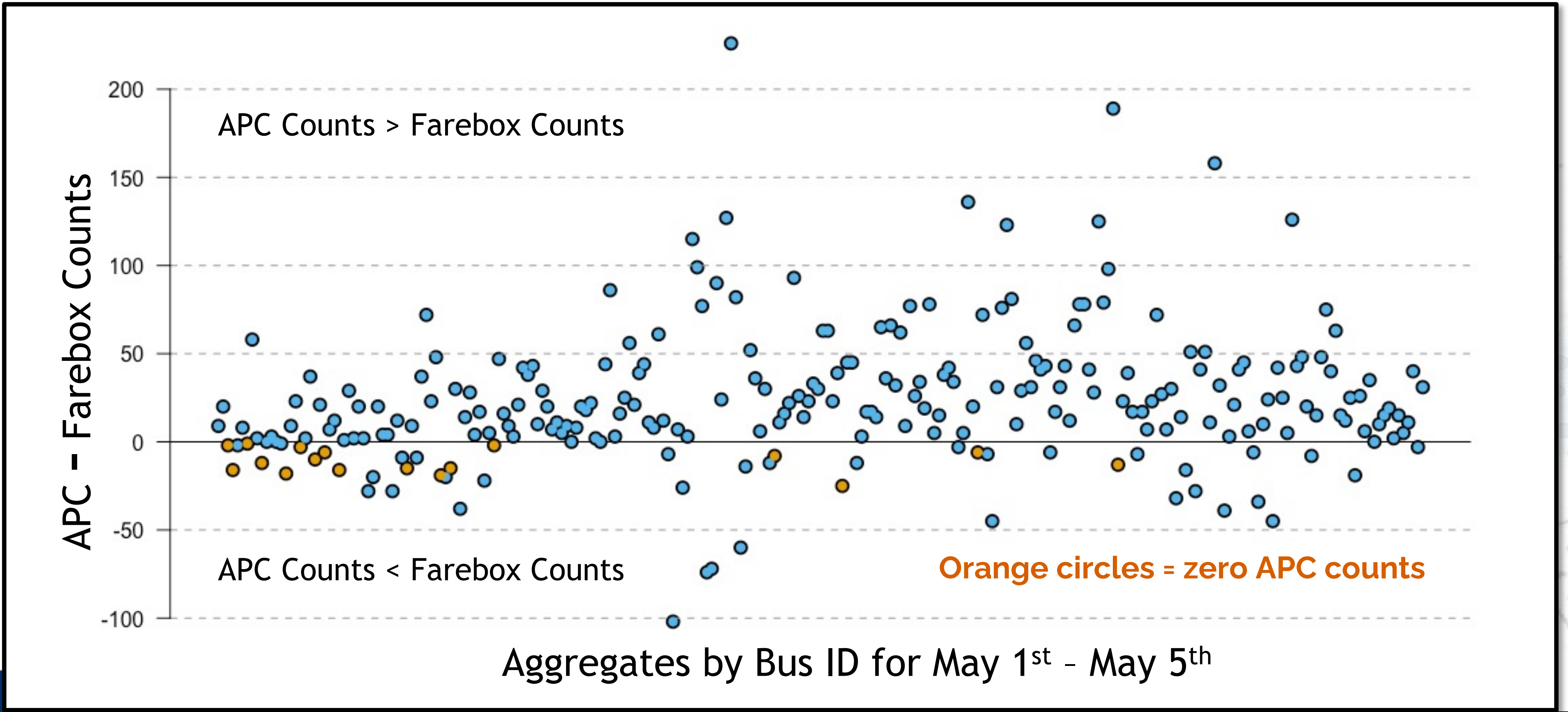
- Not all economically vulnerable Census Block Groups have high numbers of fare evaders, but all Census Block Groups with a high numbers of fare evaders are economically vulnerable.





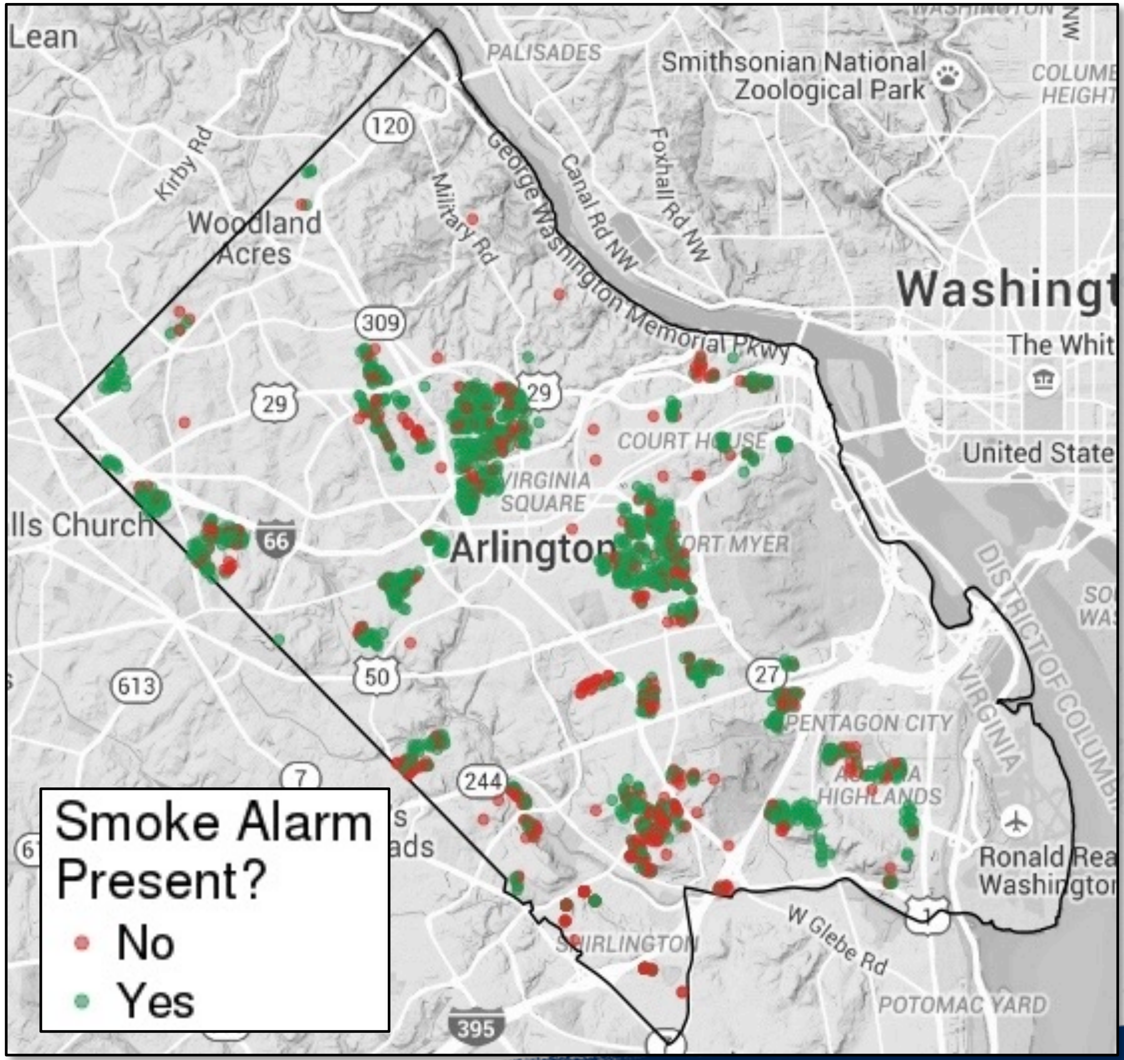
Cautionary Tale: Need Accurate Estimates for Policy Implementation and Evaluation

To test hypotheses or experiments for interventions, need to know how accurately fare evasion can be estimated – what effect sizes can be measured!



So Much Learning!

Projects – Operation FireSafe



Issue: 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 the probability it has adequate smoke detectors

The DATA

- **Household Level** ***(Administrative Data):*** Operation FireSafe data for the 5,623 single family homes visited
- **Household Level** ***(Administrative Data):*** Real estate tax assessments for 60,343 single family homes including 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

Model to Predict Regions in the County in Need of Smoke Alarms

- Bayesian logistic regression model with conditionally autoregressive spatial effects[1]
- Identify the predictors of housing units in need of a smoke alarm installation or battery replacement

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu_i + \phi_i \qquad \begin{aligned} \mu_i &= \mathbf{X}_i^T \boldsymbol{\beta} \\ \boldsymbol{\beta} &\sim N(\mathbf{0}, \mathbf{I}_{1000}) \end{aligned}$$

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

- Response Y_i=1 if household i needs a smoke alarm intervention.
- The spatial component is given by the weighted neighborhood matrix $\boldsymbol{W}$, where
  - $w_{ij} = 1$ if households *i* and j are in the same block group
  - $w_{ij} = 0.5$ if households *i* and j are in neighboring block groups
  - $w_{ij} = 0$ otherwise

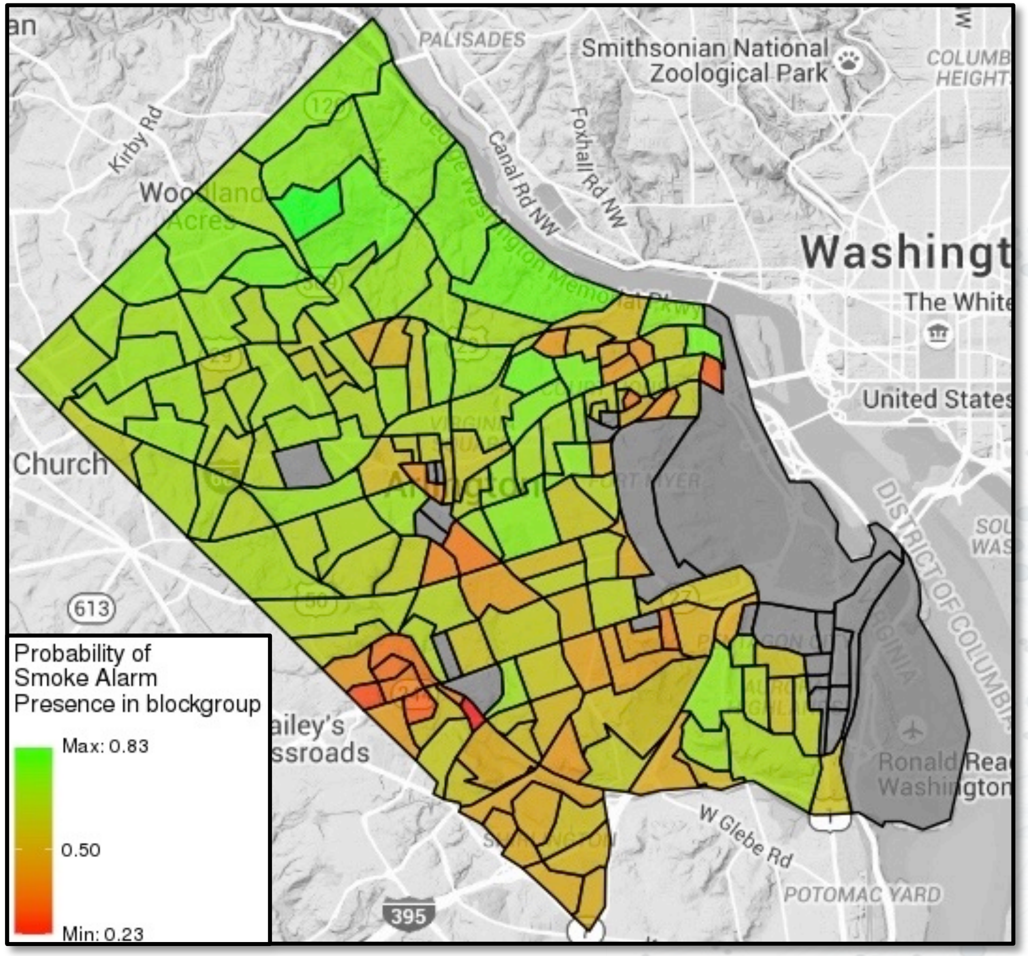
[1]Leroux B, Lei X, Breslow N (1999). Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical Models in Epidemiology, the Environment and Clinical Trials*. Springer-Verlag: New York.

Evaluating Predictive Performance

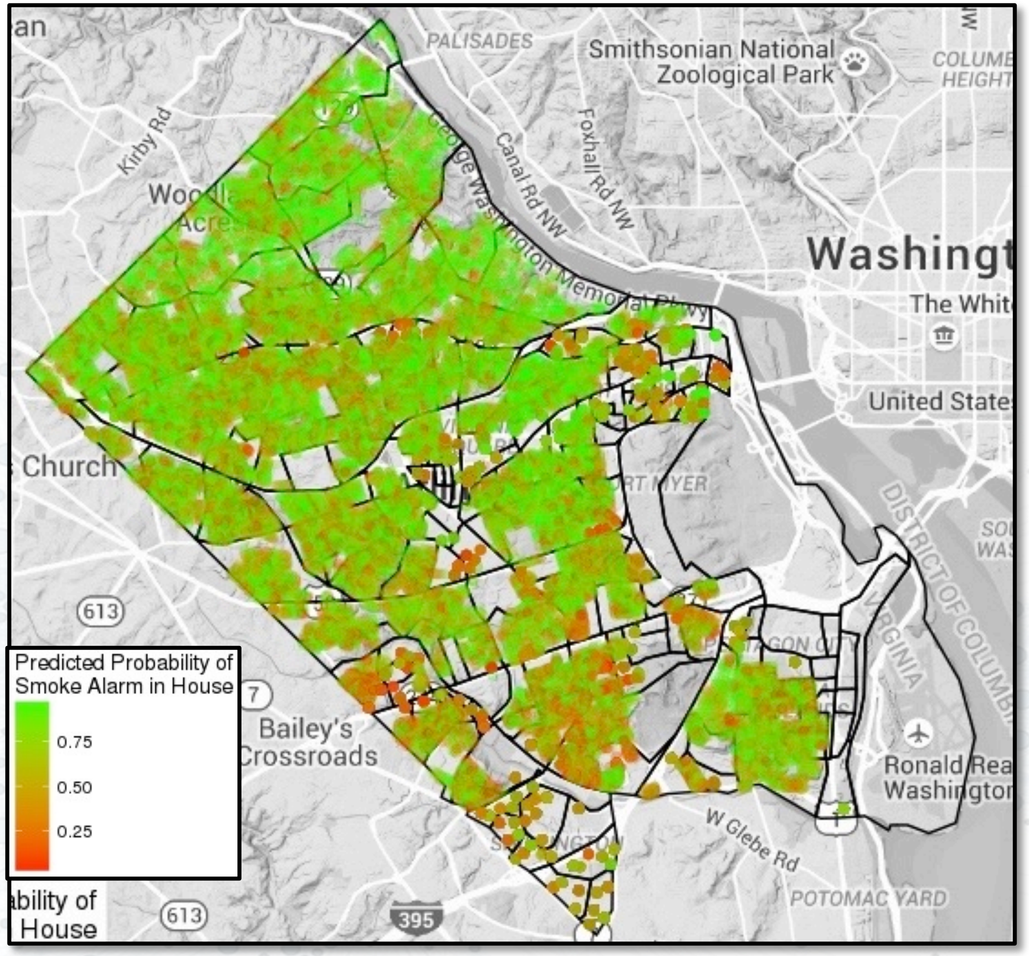
- Metrics
  - **Precision**: When the model suggests the home has a smoke alarm, what percentage of these homes actually have it?
  - **Recall**: What percentage of homes with alarms is the model catching?
- When $p_i < t$ for threshold t, the **model predicts** that household i needs a smoke alarm intervention
  - We find the value of t that maximizes precision and recall
- Models **are trained** using repeated holdout samples of 10%, 15% and 20% of the observations
  - For each holdout sample, the model is fit and predicted probabilities are assigned to the holdout samples
- 2015 Operation FireSafe data used to **estimate the performance** of 2016 data
  - Then both years were used to make final predictions for 2017
- Predictions made at the **household level and aggregated** to Census block groups

Probability of Having a Smoke Alarm

Bayesian logistic regression model with conditionally autoregressive spatial effects



Census Block Level Predictions



Housing Unit Level Predictions

UNIVERSITY of VIRGINIA

Findings

Predictors of smoke detector need

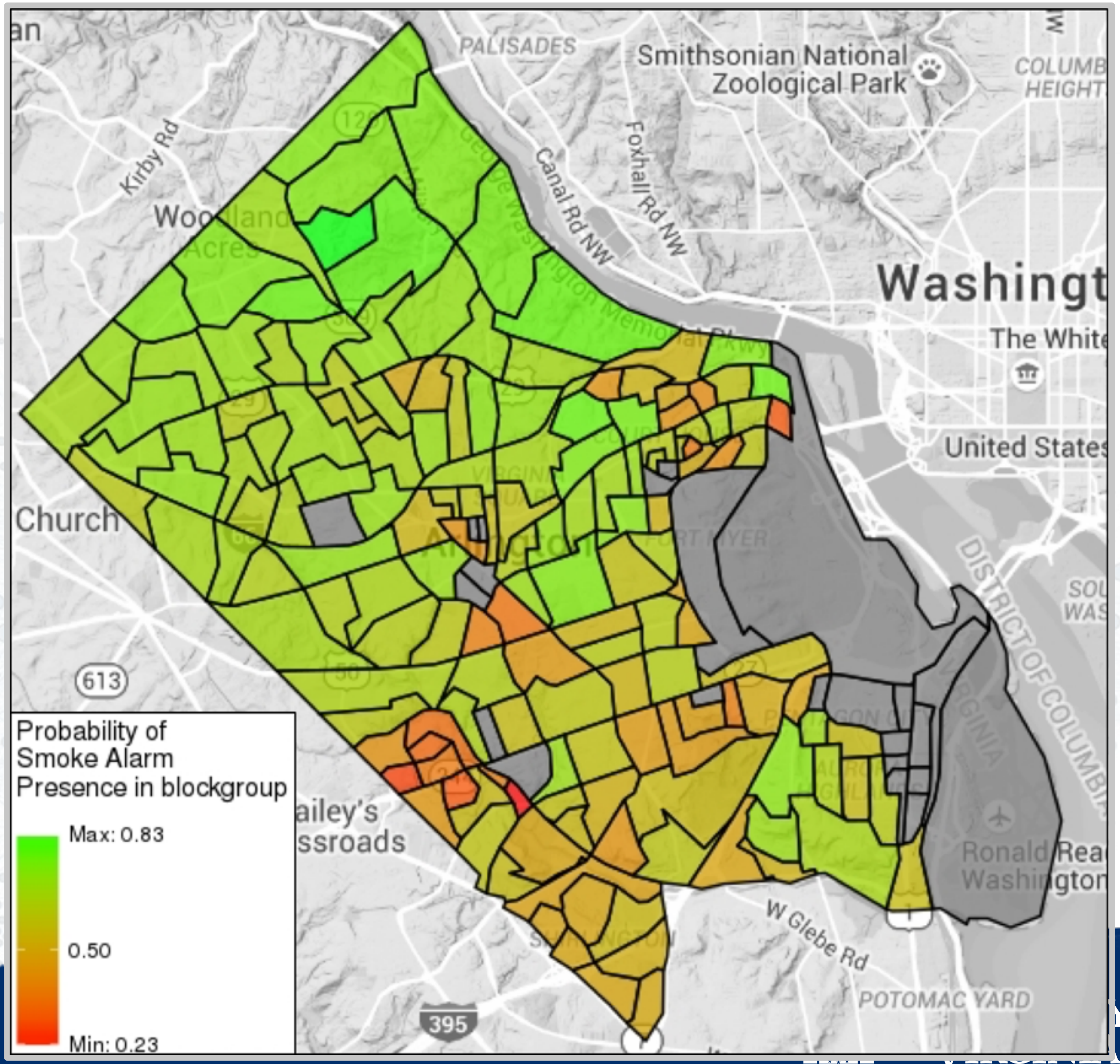
At the household level:

- Home value (-)
- Age of home (+)

At the neighborhood level:

- Median income (-)
- Percent living alone (+)
- Percent family households (-)

ACFD used a list of vulnerable neighborhoods predicted to target their visits for 2017



Projects – Arlington Restaurant Initiative

Measuring the Impact of Alcohol-Related Crime Reduction Strategies for Restaurants and Nightlife in Arlington

- Arlington County features some of the most unique restaurants and nightlife destinations in the Washington D.C. metro region. Areas such as Clarendon, however, with a large number of restaurants have become a difficult issue for police to manage due to alcohol-related crimes such as malicious wounding, sexual assault, public intoxication, assault on police, DUI, disorderly conduct, and rape.

UNIVERSITY of VIRGINIA

Projects – Arlington Restaurant Initiative

Measuring the Impact of Alcohol-Related Crime Reduction Strategies for Restaurants and Nightlife in Arlington

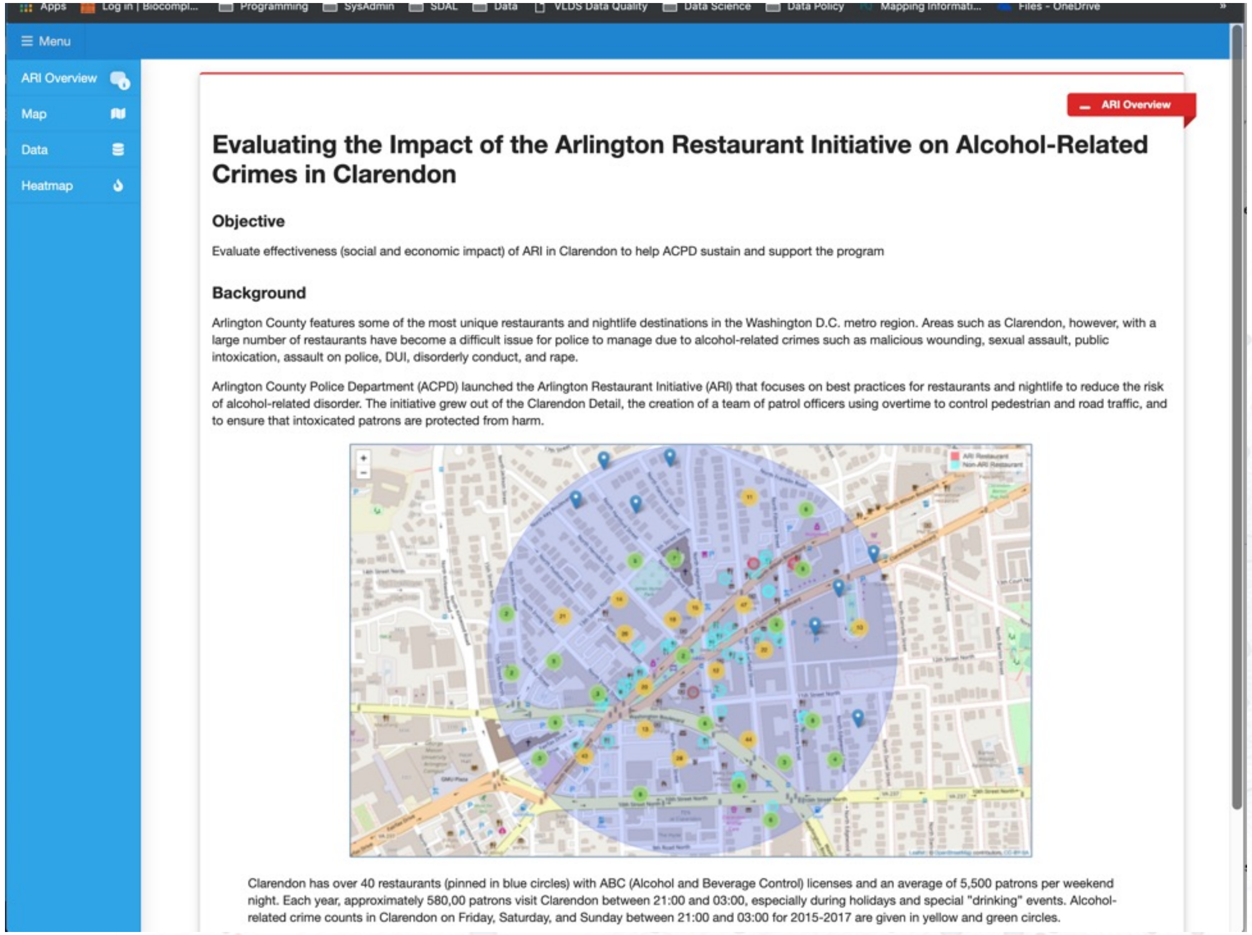
- Arlington County Police Department (ACPD) launched the *Arlington Restaurant Initiative (ARI)* that focuses on best practices for restaurants and nightlife to reduce the risk of alcohol-related disorder. The initiative grew out of the *Clarendon Detail*, the creation of a team of patrol officers using overtime to control pedestrian and road traffic, and to ensure that intoxicated patrons are protected from harm.

Projects – Arlington Restaurant Initiative

Measuring the Impact of Alcohol-Related Crime Reduction Strategies for Restaurants and Nightlife in Arlington

- **Objective:** Evaluate effectiveness (social and economic impact) of ARI in Clarendon to help ACPD **sustain and support program funding**.
- Need a justifiable funding model that goes beyond counting arrest

Projects – Arlington Restaurant Initiative



Menu

ARI Overview

Map

Data

Heatmap

Crime Type

DUI

Arlington Crime Data

Copy CSV Excel Print

Search:

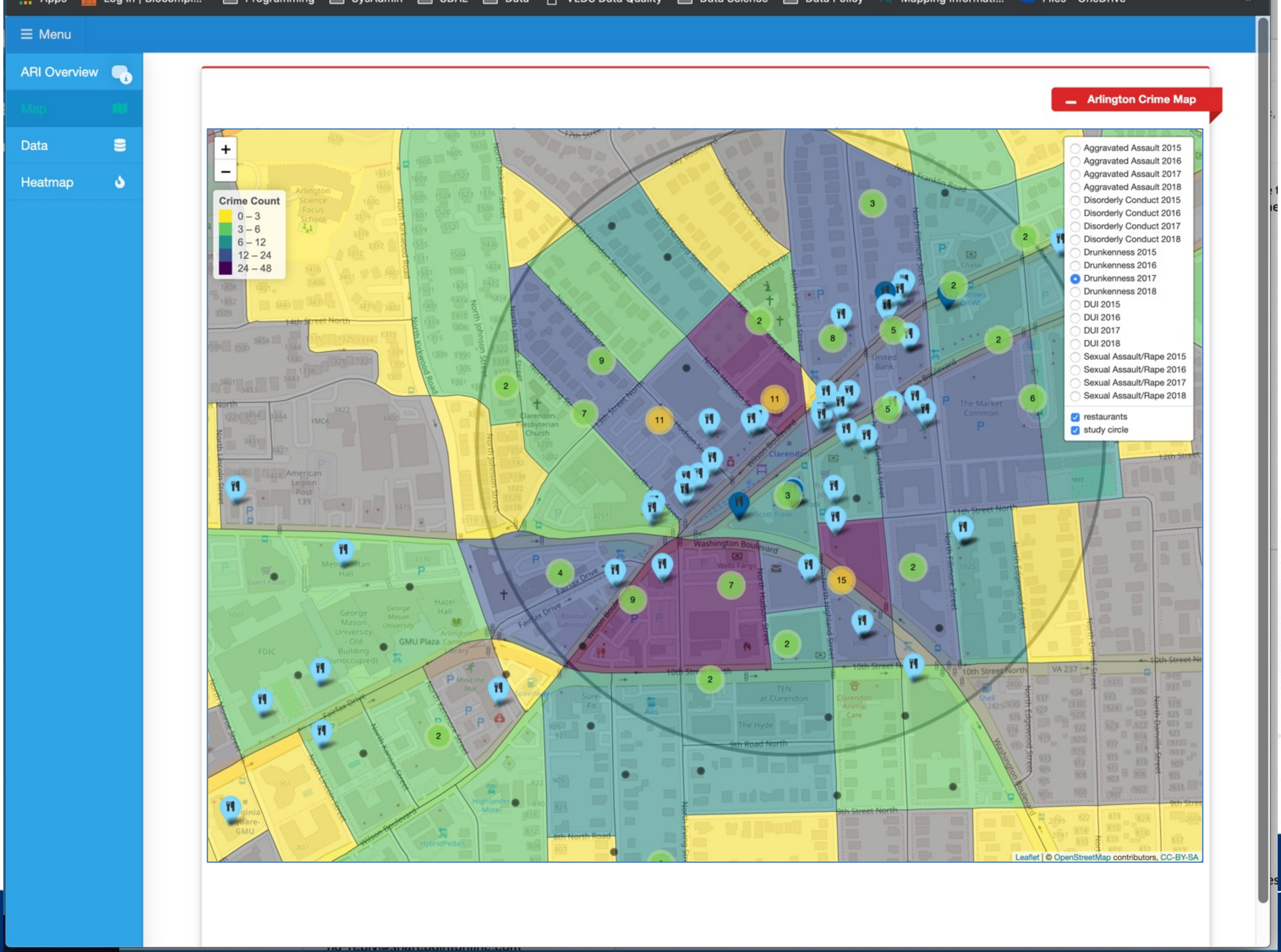
	id	description	location	latitude	longitude	start	end	year	month	day_of_week
1	2018-05310007	DUI	S WALTER REED DR / 18TH ST S	38.85299031	-77.08812649	2018-05-31T00:43:00Z	2018-05-31T00:43:00Z	2018	5	Thursday
2	2018-05280152	DUI 3+ OFFENSE OR 2+ FELONY OFFENSE	1XX N GLEBE RD	38.87262371	-77.10374707	2018-05-28T16:51:00Z	2018-05-28T16:51:00Z	2018	5	Monday
3	2018-05280057	DUI	N LYNN ST / LEE HWY	38.89716894	-77.06996344	2018-05-28T05:00:00Z	2018-05-28T05:00:00Z	2018	5	Monday
4	2018-05270039	DUI	ARLINGTON BLVD / N COLUMBUS ST	38.86547337	-77.11670666	2018-05-27T04:04:00Z	2018-05-27T04:04:00Z	2018	5	Sunday
5	2018-05260260	DUI	XX 8468436440000000	38.84758519	-77.08140523	2018-05-26T23:40:00Z	2018-05-26T23:40:00Z	2018	5	Saturday
6	2018-05260243	DUI	13TH ST S / S GEORGE MASON DR	38.85768855	-77.09867447	2018-05-26T22:37:00Z	2018-05-26T22:37:00Z	2018	5	Saturday
7	2018-05260178	DUI	23XX 25TH ST S	38.84887425	-77.07555001	2018-05-26T17:15:00Z	2018-05-26T17:30:00Z	2018	5	Saturday
8	2018-05260025	DUI	25XX S WALTER REED DR	38.84614347	-77.10075429	2018-05-26T01:30:00Z	2018-05-26T01:30:00Z	2018	5	Saturday
9	2018-05250034	DUI	N HIGHLAND ST / 9TH RD N	38.88215652	-77.09260151	2018-05-25T03:08:00Z	2018-05-25T03:08:00Z	2018	5	Friday
10	2018-05250025	DUI	10TH ST N / FAIRFAX DR	38.88351863	-77.09836161	2018-05-25T02:04:00Z	2018-05-25T02:06:00Z	2018	5	Friday

Showing 1 to 10 of 1,168 entries

Previous 1 2 3 4 5 … 117 Next

UNIVERSITY of VIRGINIA

VERSITY
IRGINIA

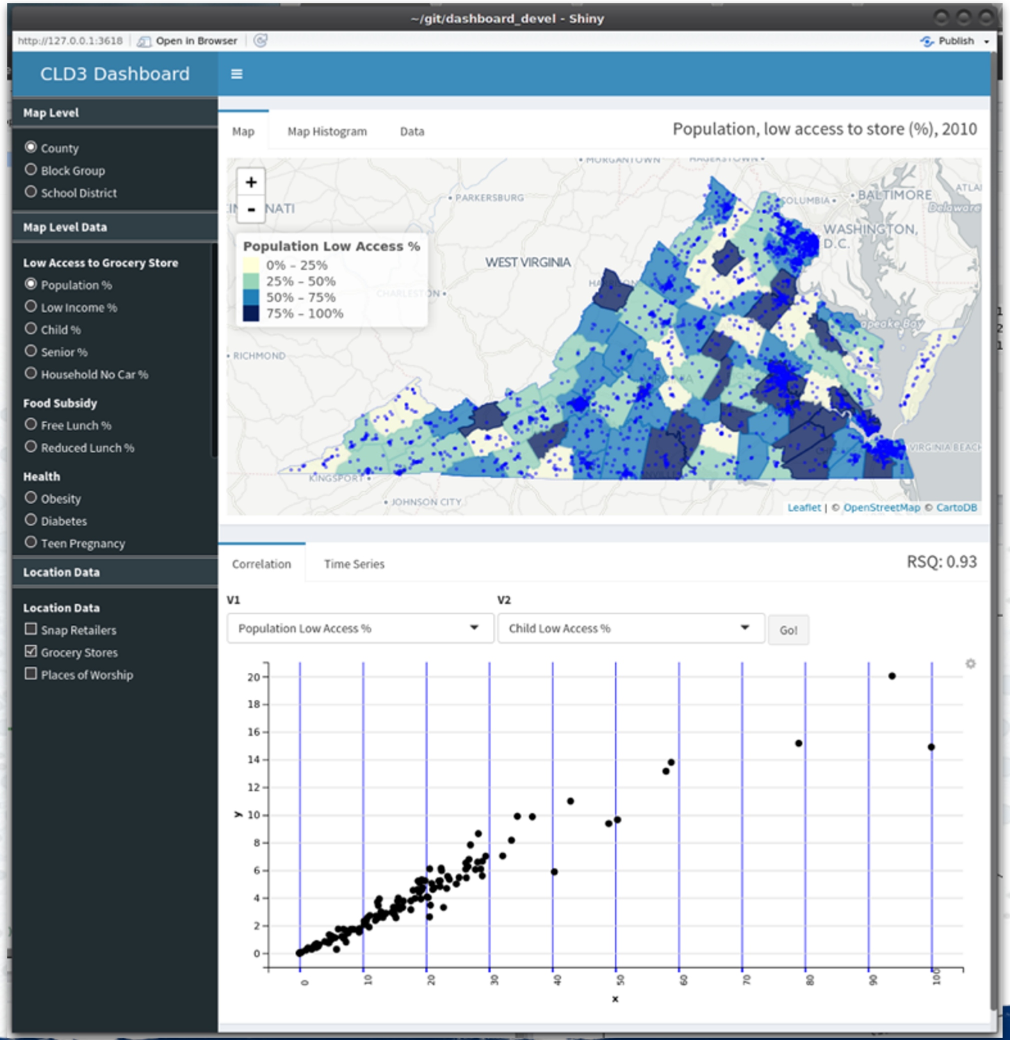


UNIVERSITY of VIRGINIA

New Data Sources for VCE Community Profiles

Project Background

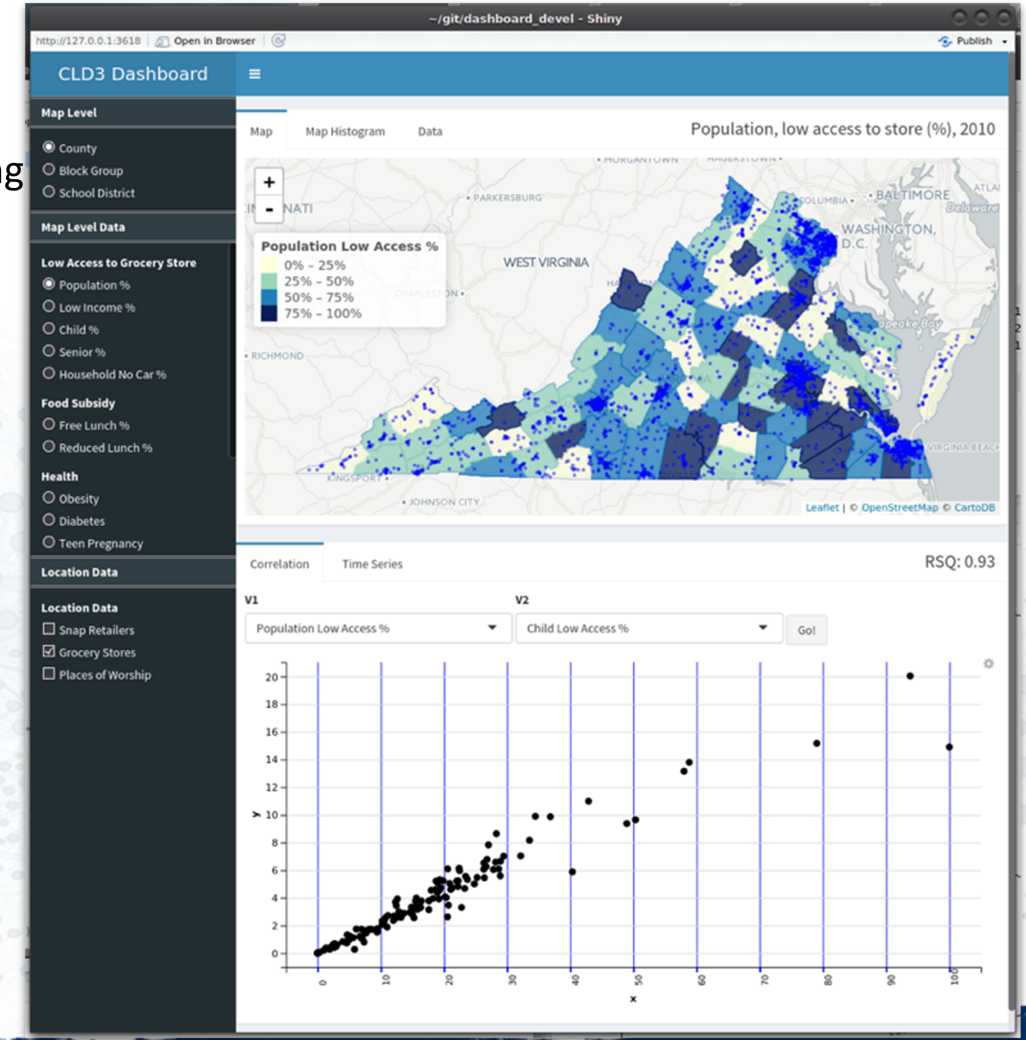
- VCE District Planning includes an process of Community Profiling
- SDAL is currently working with VCE to develop a Dashboard that will allow for a deeper level of analysis than is currently conducted, including:
  - **Location (Place)-Level Data** (Places of Worship, SNAP Providers, Social Service Locations)
  - **Sub-County-Level Data** (Block Groups, Water Districts, School Boundaries)
  - **Tools** for quick descriptive analysis (Maps, Time-Series, Scatterplots



New Data Sources for VCE Community Profiles

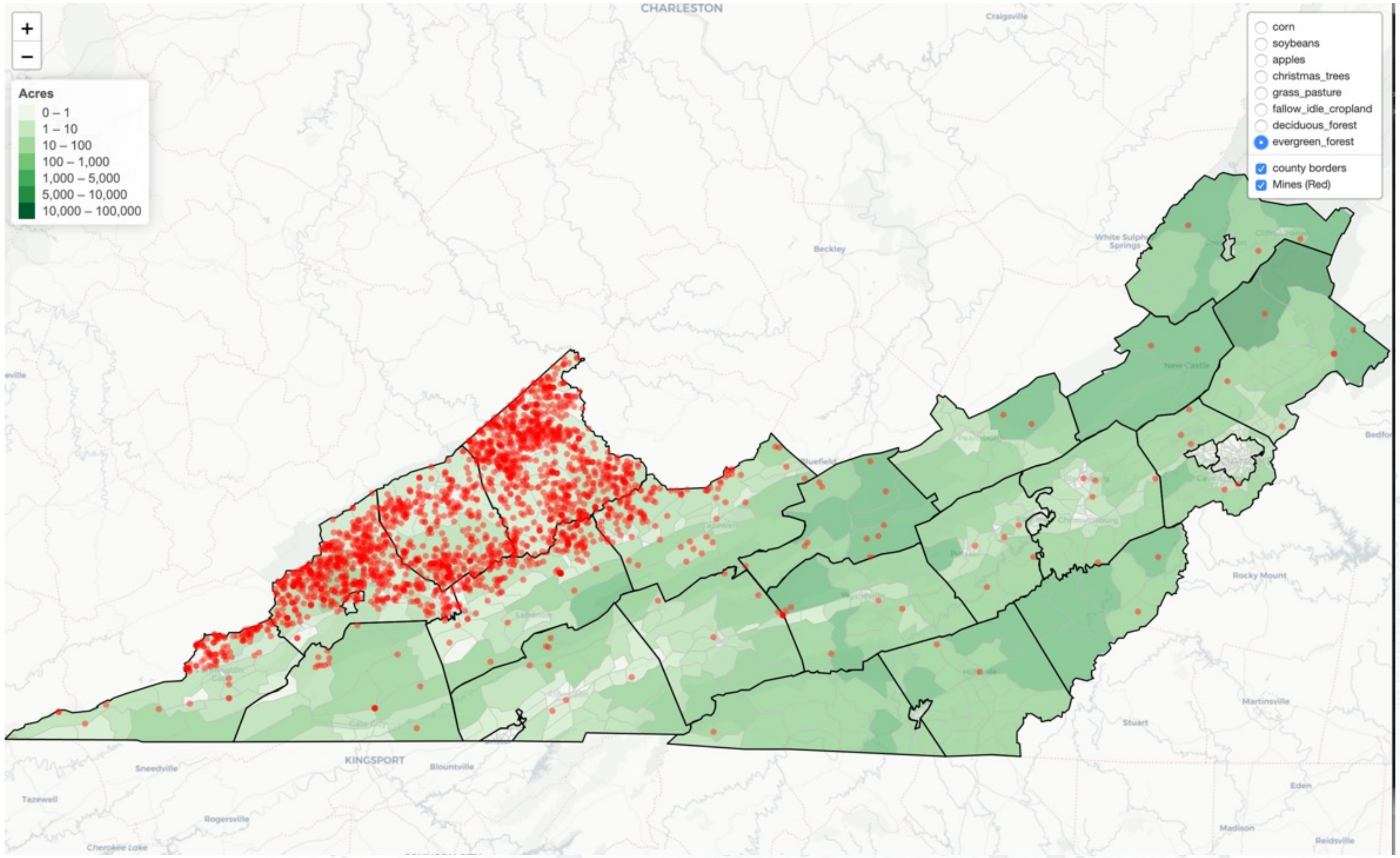
Project Goal(s) / Research Questions

- Discover possible sub-county data sources for use by VCE in creating community profiles, including:
  - **Radon Levels**
  - **Active Coal Mine Locations**
  - **Water Quality (Rivers)**
  - **Water Quality (Treatment Plants)**
  - **Soil Status Quality**
  - **Crop Coverage**
- Profile discovered data sources quality, structure, provenance and metadata
- Determine necessary transformations for data re-purposing and overall data source fitness to support VCE goals
- Incorporate new data into dashboard; create new dashboard tools



UNIVERSITY of VIRGINIA

New Data Sources for VCE Community Profiles



Fairfax Youth Partnership

Describing Populations and Forecasting Future Needs

- **Issue:** Characterize the factors that describe **depression** and **obesity in youth**
- **Goal:** This type of understanding will allow Fairfax County to implement policies targeted to address **youth behaviors**
- **Approach:** Identify access to food and physical activity options, e.g. grocery stores, restaurants, farmers markets, community gardens, parks, and more

University of Virginia

The DATA

Household Level

- Geocoded the household locations
- Supervisor districts

Places of Interest

- Food, parks and recreation centers
- Transportation – routes and modalities
- Clinics

Census Tract Block Group Level

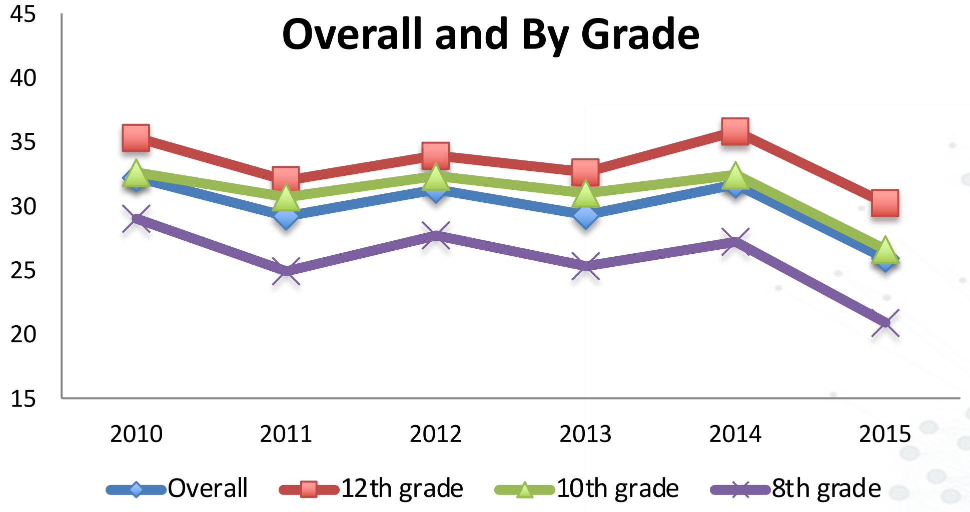
- 5-year 2015 American Community Survey – household level measures of economic vulnerability

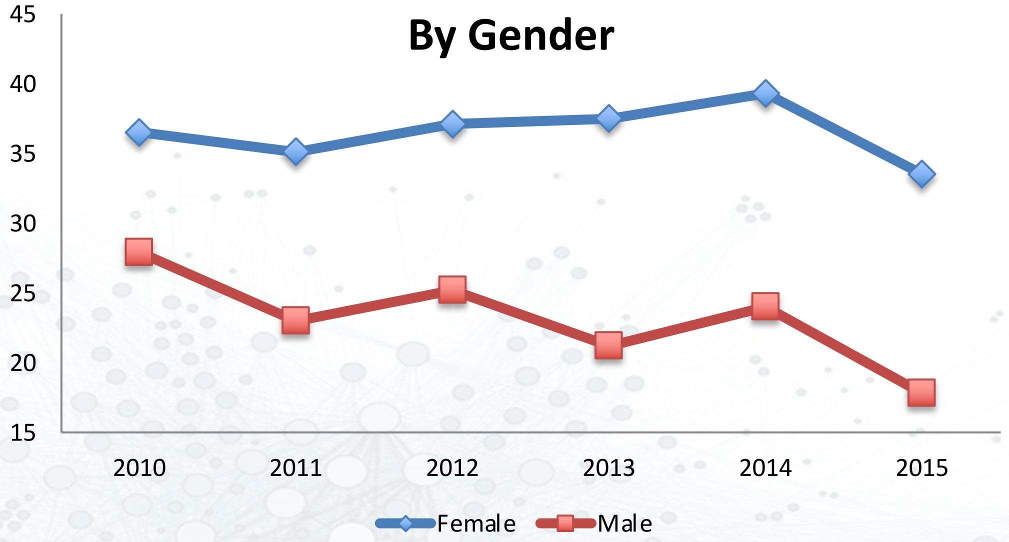
School Level

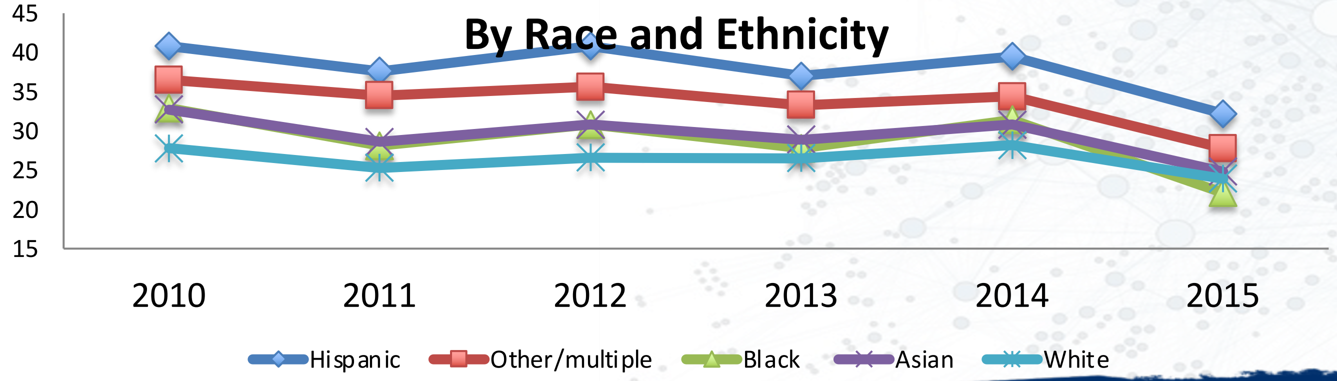
- Locations and boundaries
- Fairfax County Youth Survey, 2010 -2015

Youth Reporting Depressive Symptoms

Percentage of Students Who Felt Sad or Hopeless in the Past Year





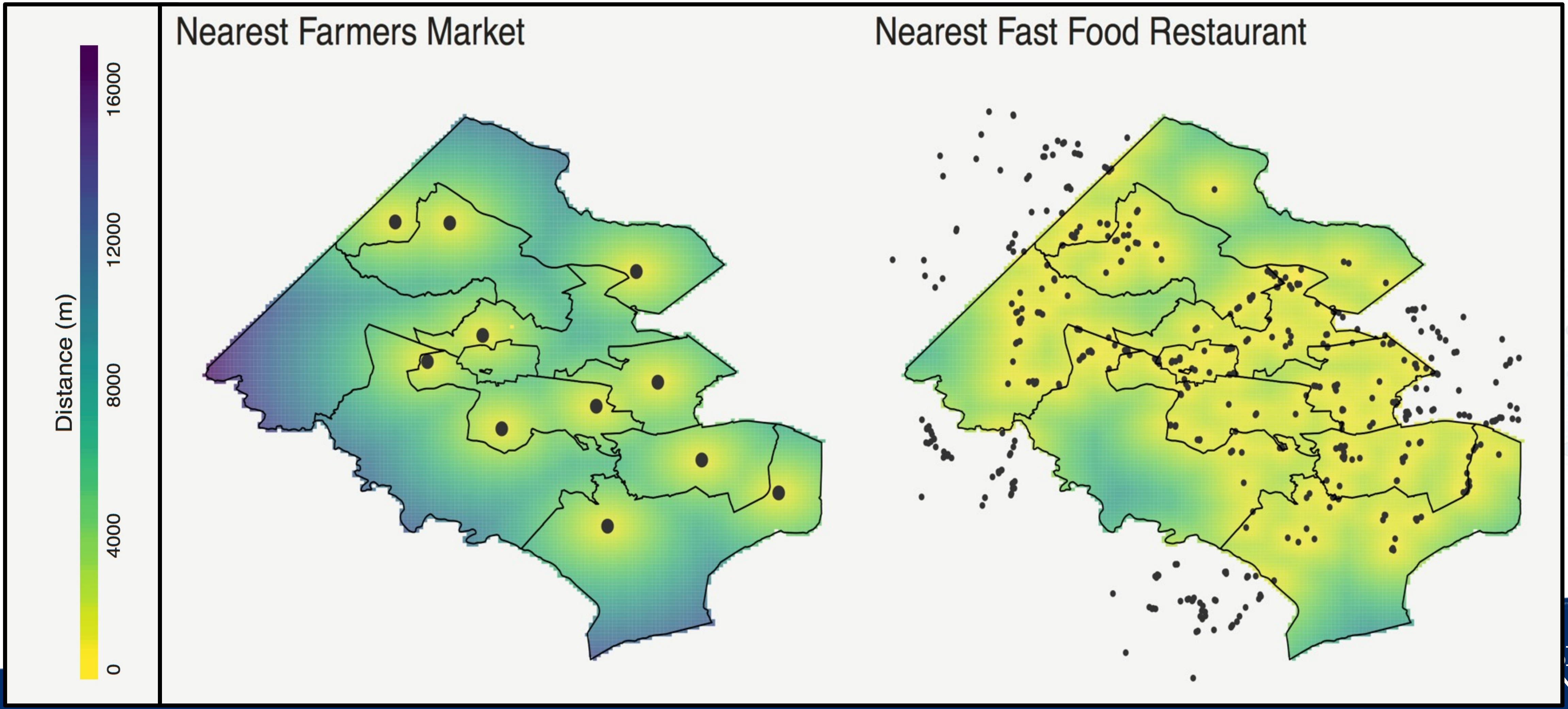


Source: Fairfax County Youth Survey, 2010 - 2015

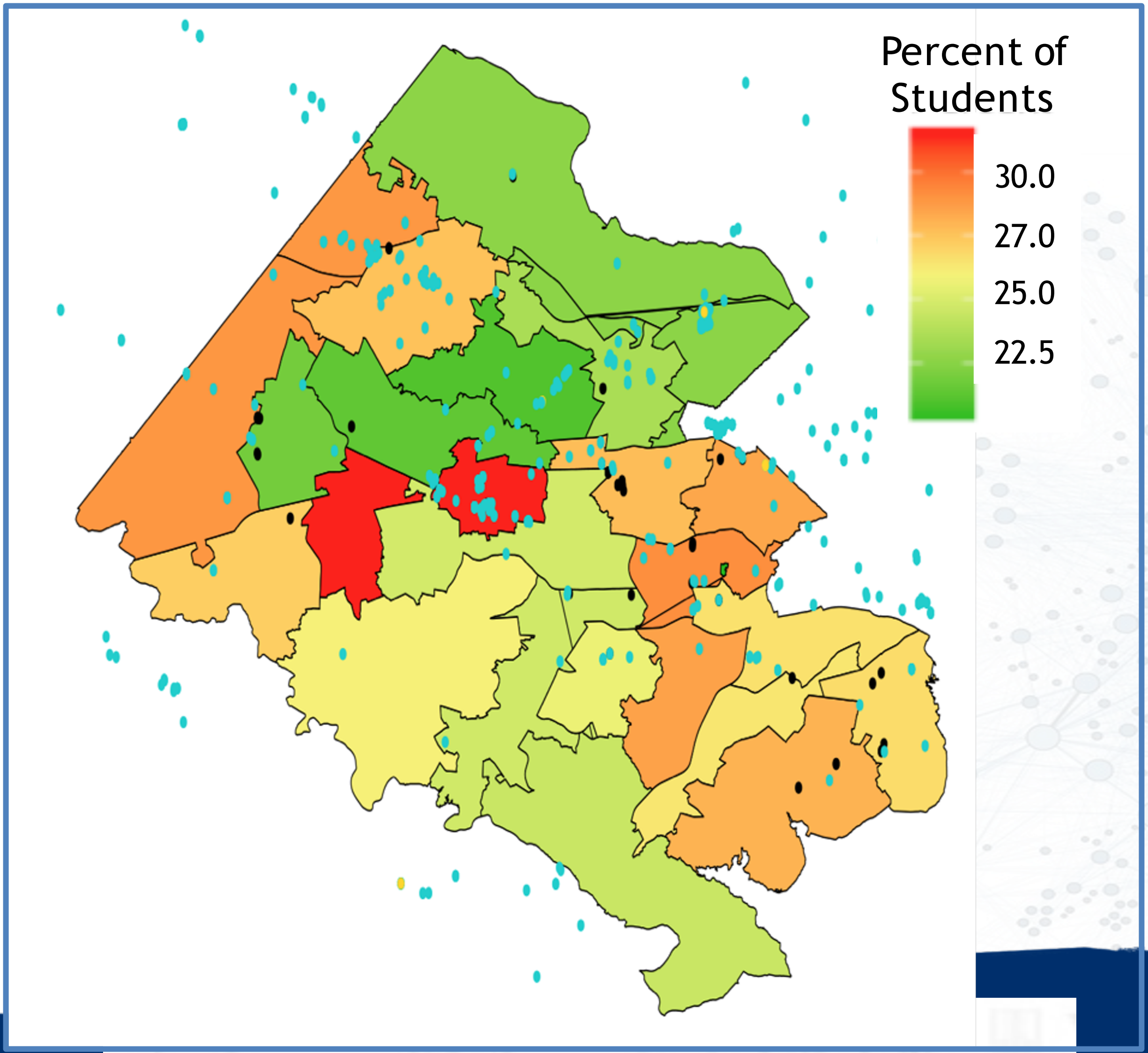
UNIVERSITY of VIRGINIA

Distance to Healthy and Unhealthy Foods based on Location of Housing Units

Supervisor Districts in Fairfax County, Virginia



Students Reporting Depressive Symptoms and Location of Mental Health Providers



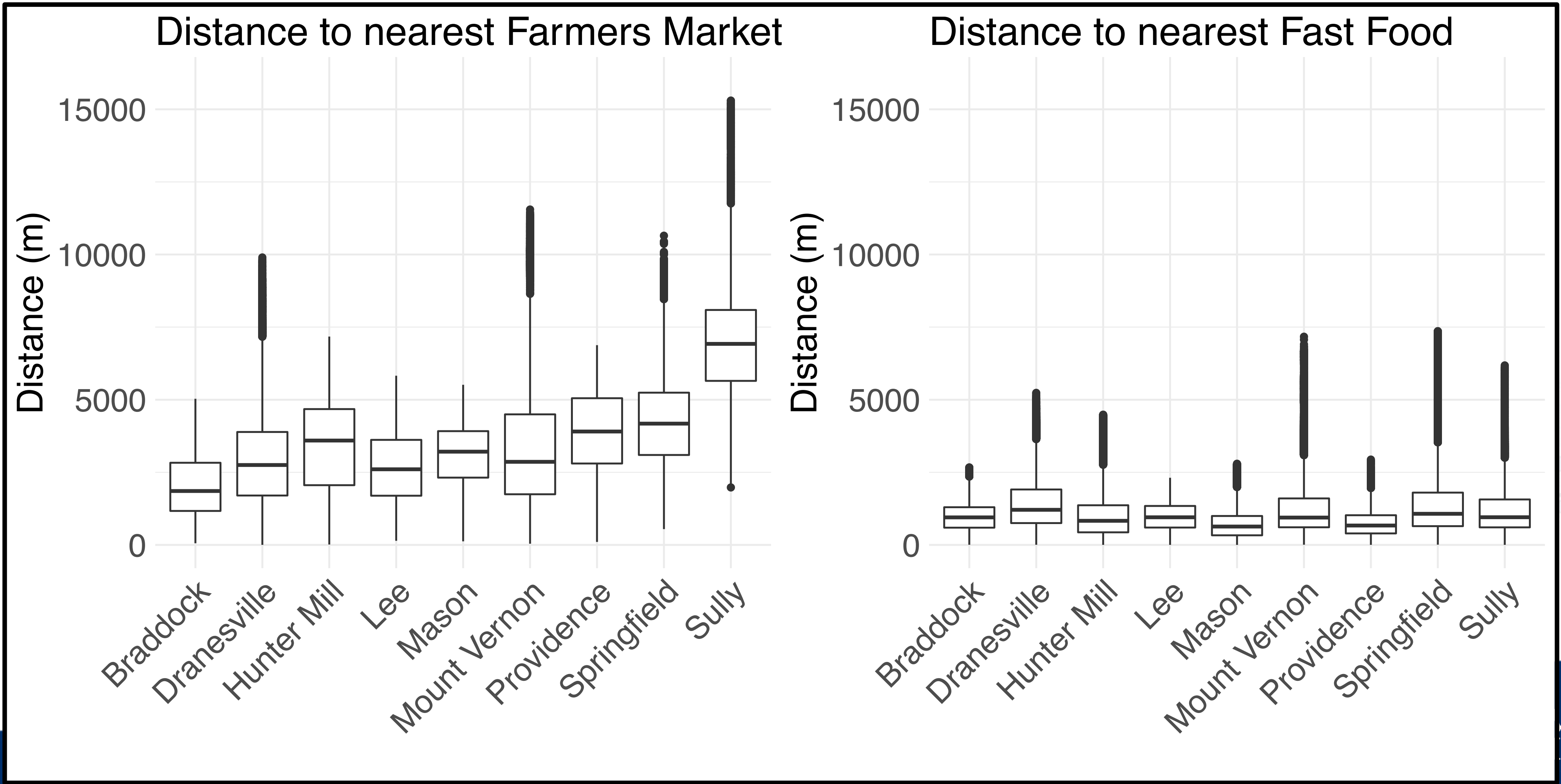
% of Students reporting depressive symptoms on youth survey –those who felt so sad or hopeless almost everyday for two weeks or more in a row during the past 12 months

Point are locations of Mental Health Providers – reported from Psychology Today and SAMHSA

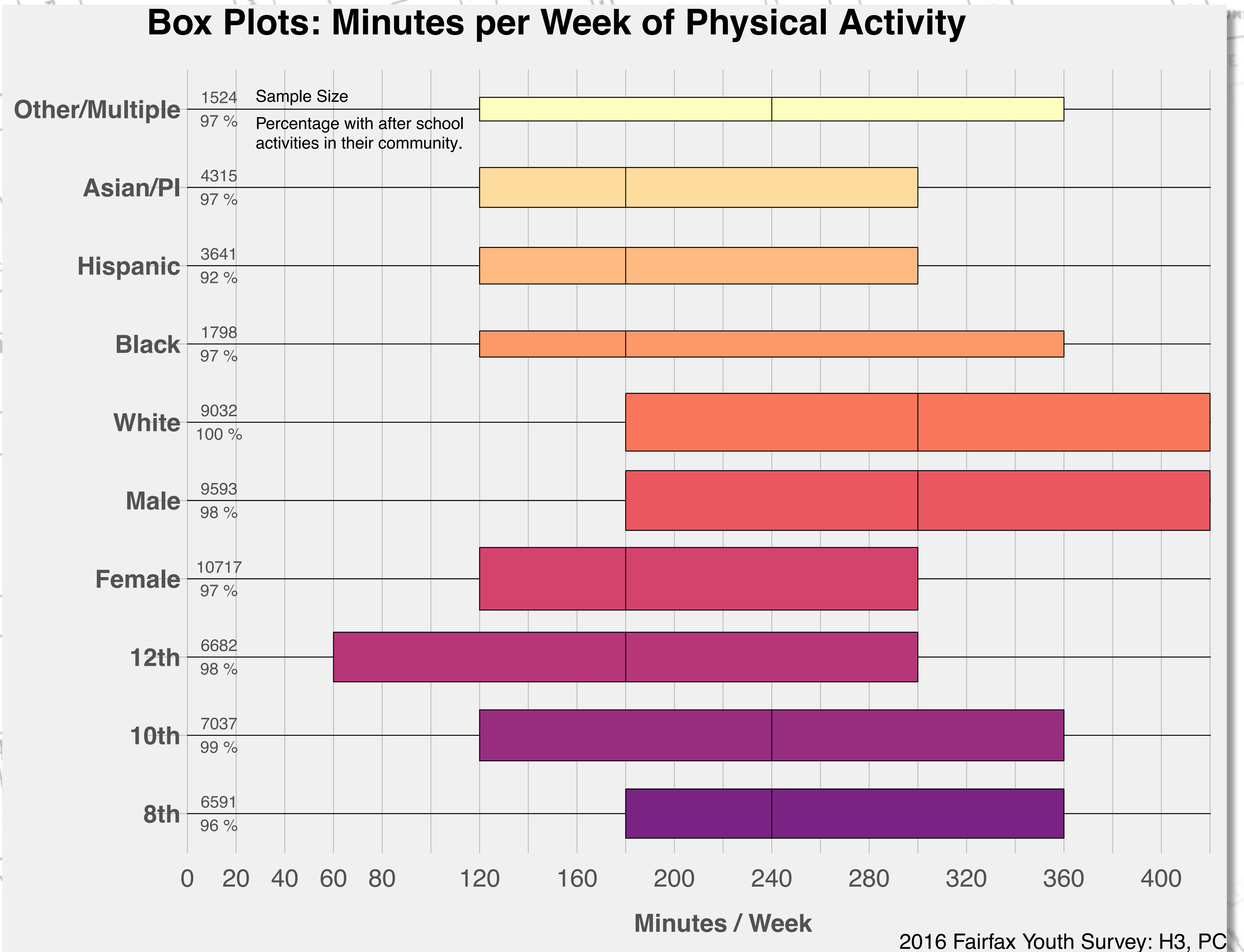
University of Virginia

Distance to Healthy and Unhealthy Foods based on Location of Housing Units

Supervisor Districts in Fairfax County, Virginia

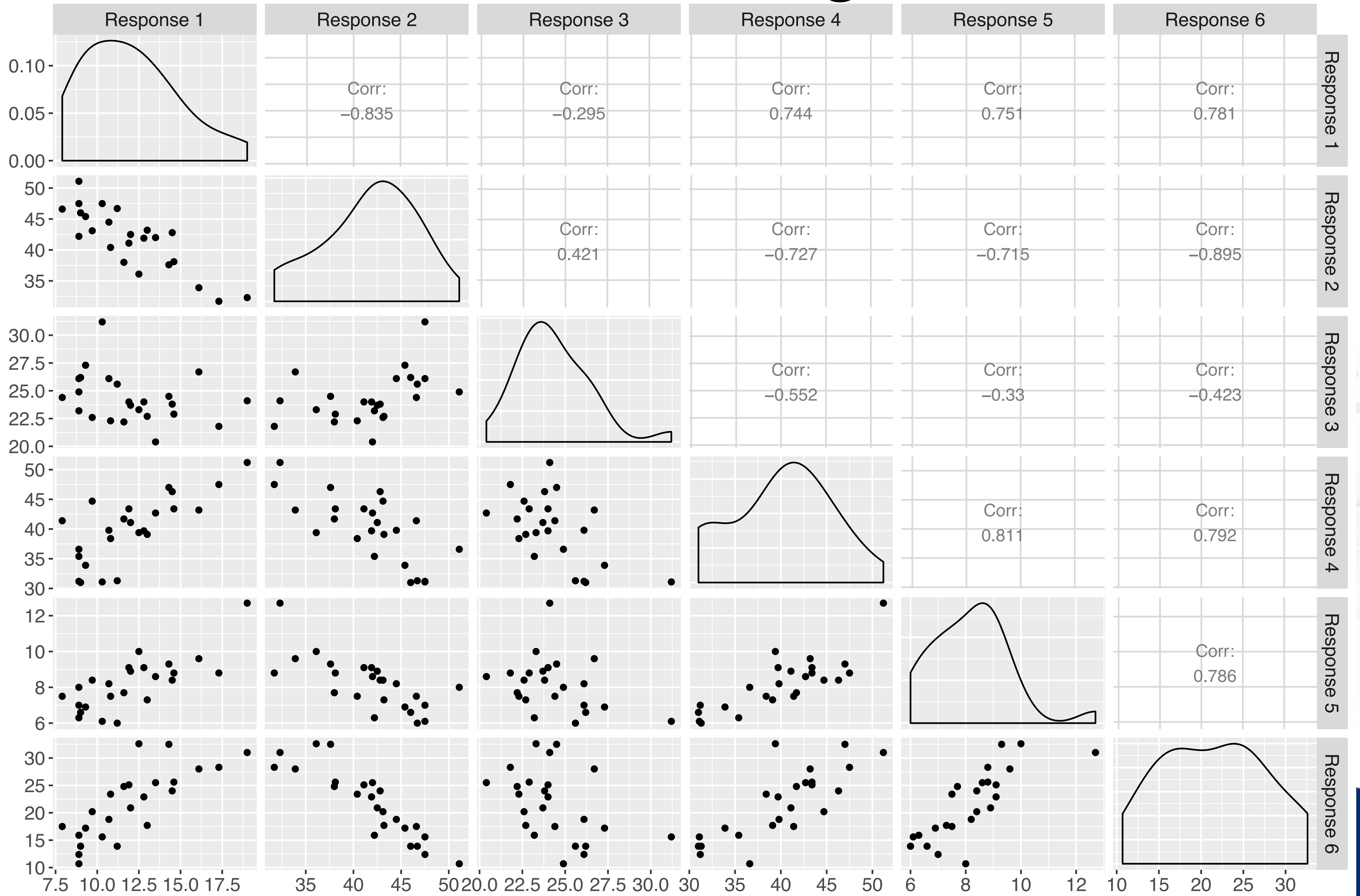


VERSITY
RGINIA



Minutes per Week of Physical Activity by Grade, Gender, Race, and Ethnicity

Factors That Might Affect Obesity



Response Variable	
1	No Physical Activity
2	5+ Days of Physical Activity
3	5+ Servings of Fruit and Vegetables
4	1+sugary drink per day
5	Unhealthy weight loss
6	Food Insecurity

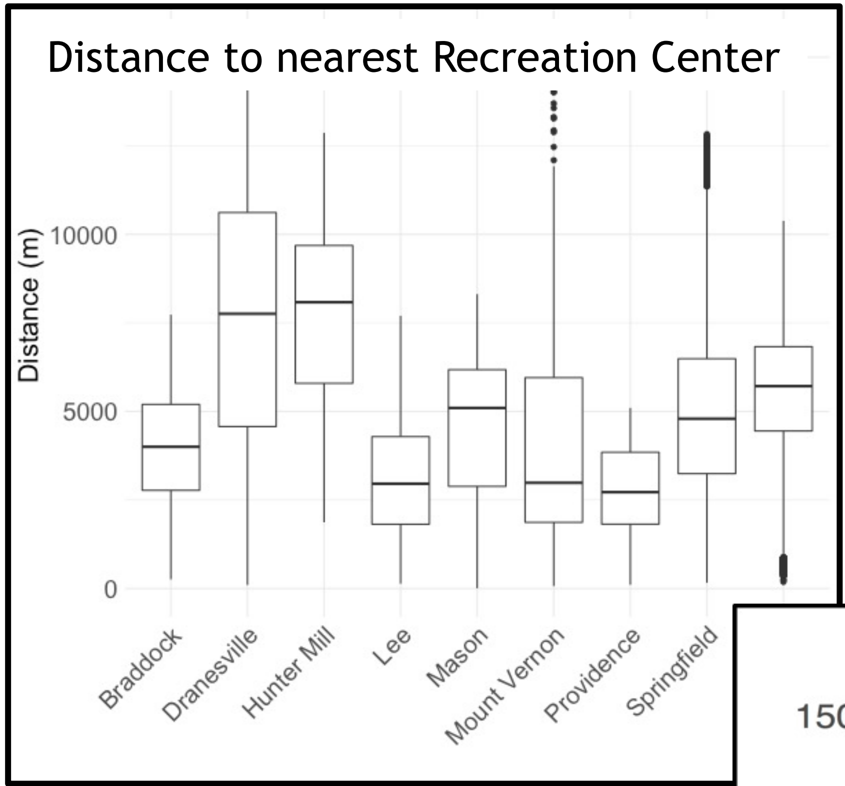
Fairfax Youth Survey, 2016-2017

Statistical Geographic Data Redistribution

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

- **Synthetic population technology** – Creating socioeconomic profiles of Supervisor Districts and High School Attendance Areas by **statistically** aligning American Community Survey data to these new geographic boundaries
- **Geocoding housing units, both owned and rented** – based on tax assessment records and the reconstruction of rental units from files provided by Fairfax demographer
- **New sources of data** – obtained from local administrative data and webscraping, with a focus on access to food and physical activity
- **Vulnerability Composite Indicators** – integrating data
- Exploring the data using **visualization tools**

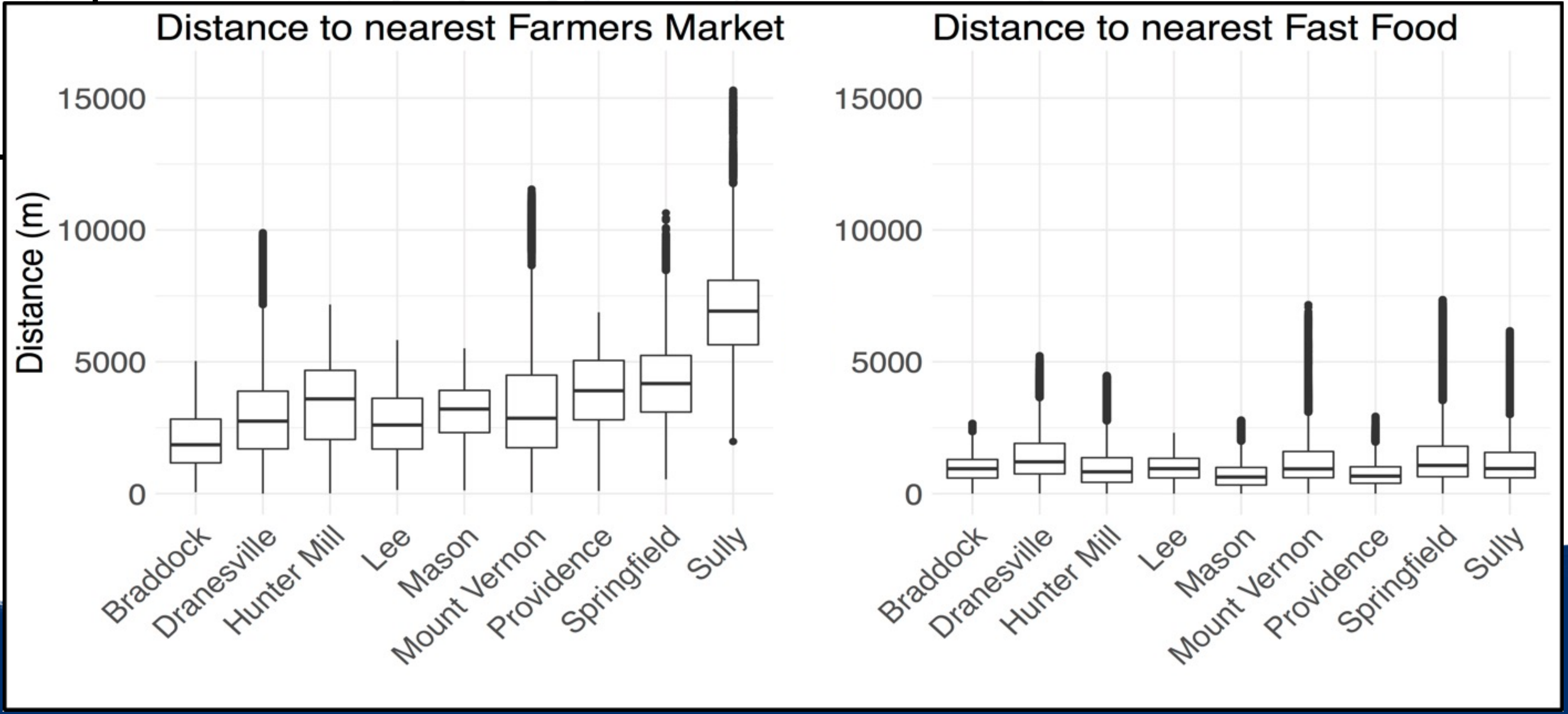
University of Virginia



Direct aggregation based on location of housing units

Geocoding owner-occupied local housing stock

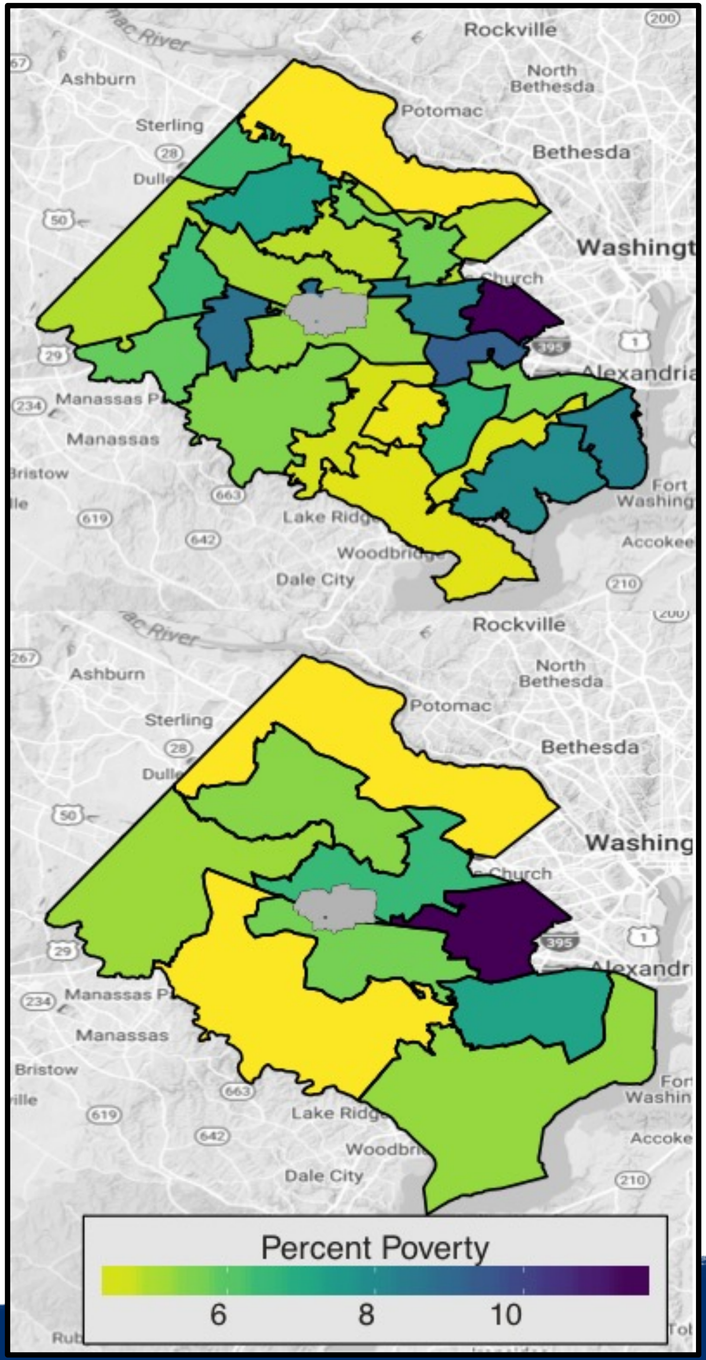
In general, adding rental units can be a challenge and may require imputation



Examples of **place data**:

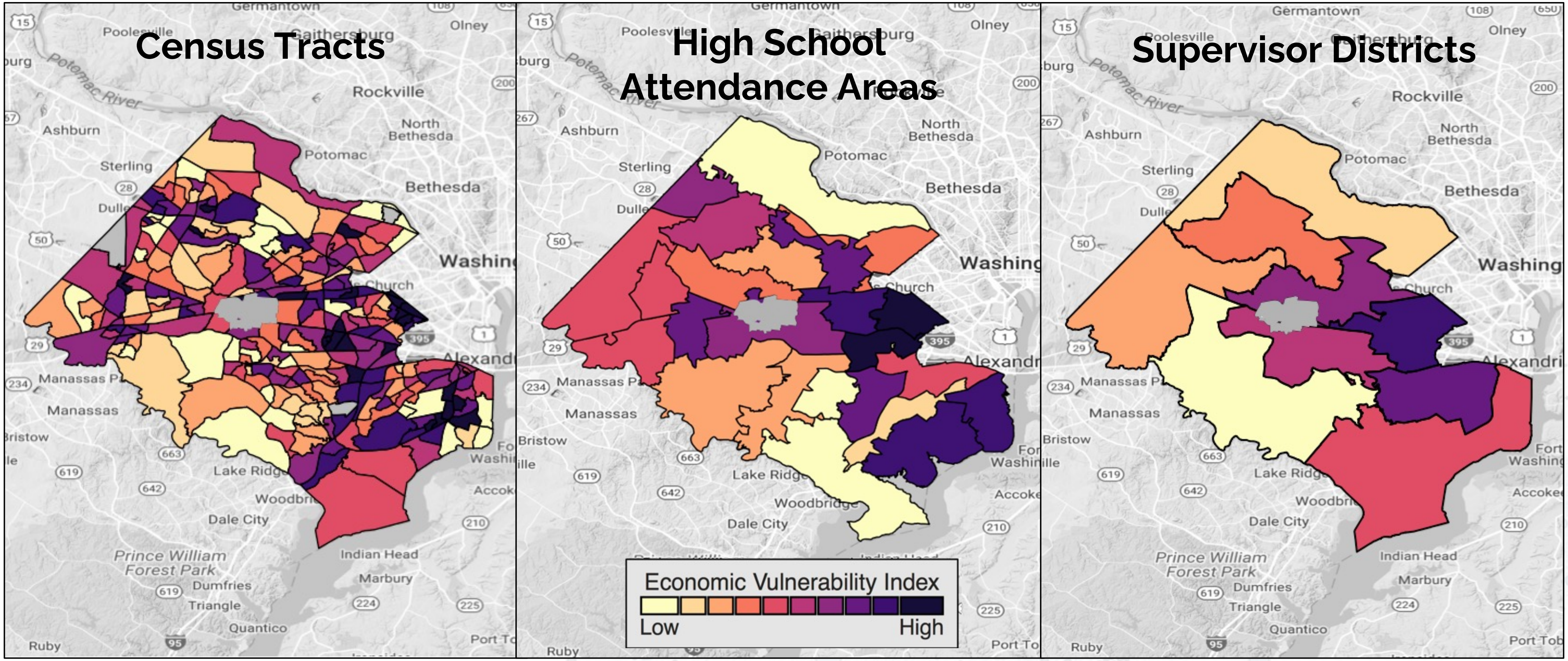
- All restaurants
- Fast Food restaurants
- Farmer's Markets
- Community Gardens
- Recreation Centers
- SNAP Retailers
- Parks

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

Sub-county Vulnerability Indicators

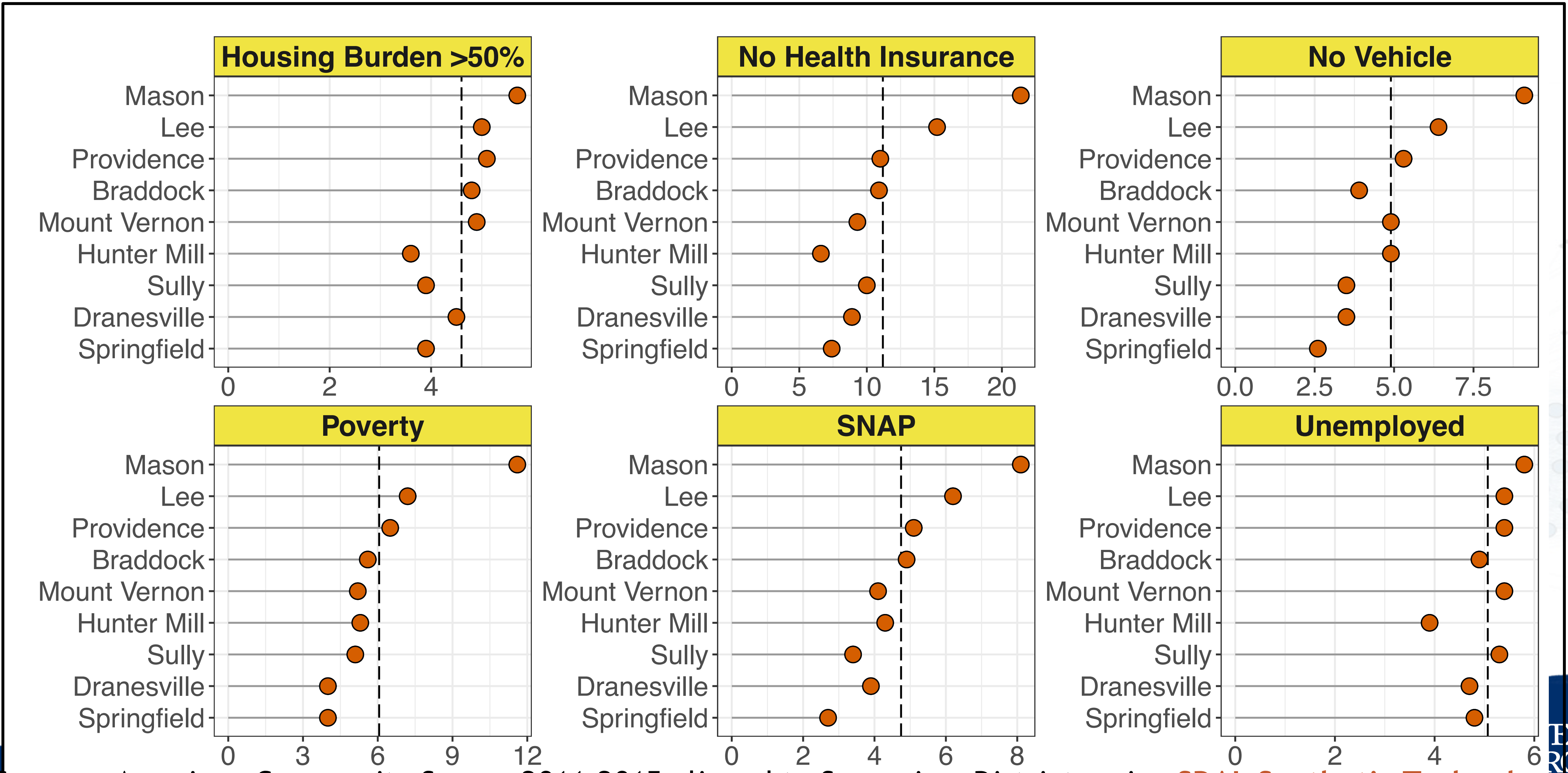


Based on a statistical combination of the percentage of Households with:

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

Fairfax Profiles by Supervisor Districts

Dashed lines = Average; Supervisor Districts arranged by Vulnerability Index from high to low



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

Repurposing Administrative Data for Statistical Purposes

National Academy of Sciences Workshop on Data Privacy for Employment Data, Jun 6 2017

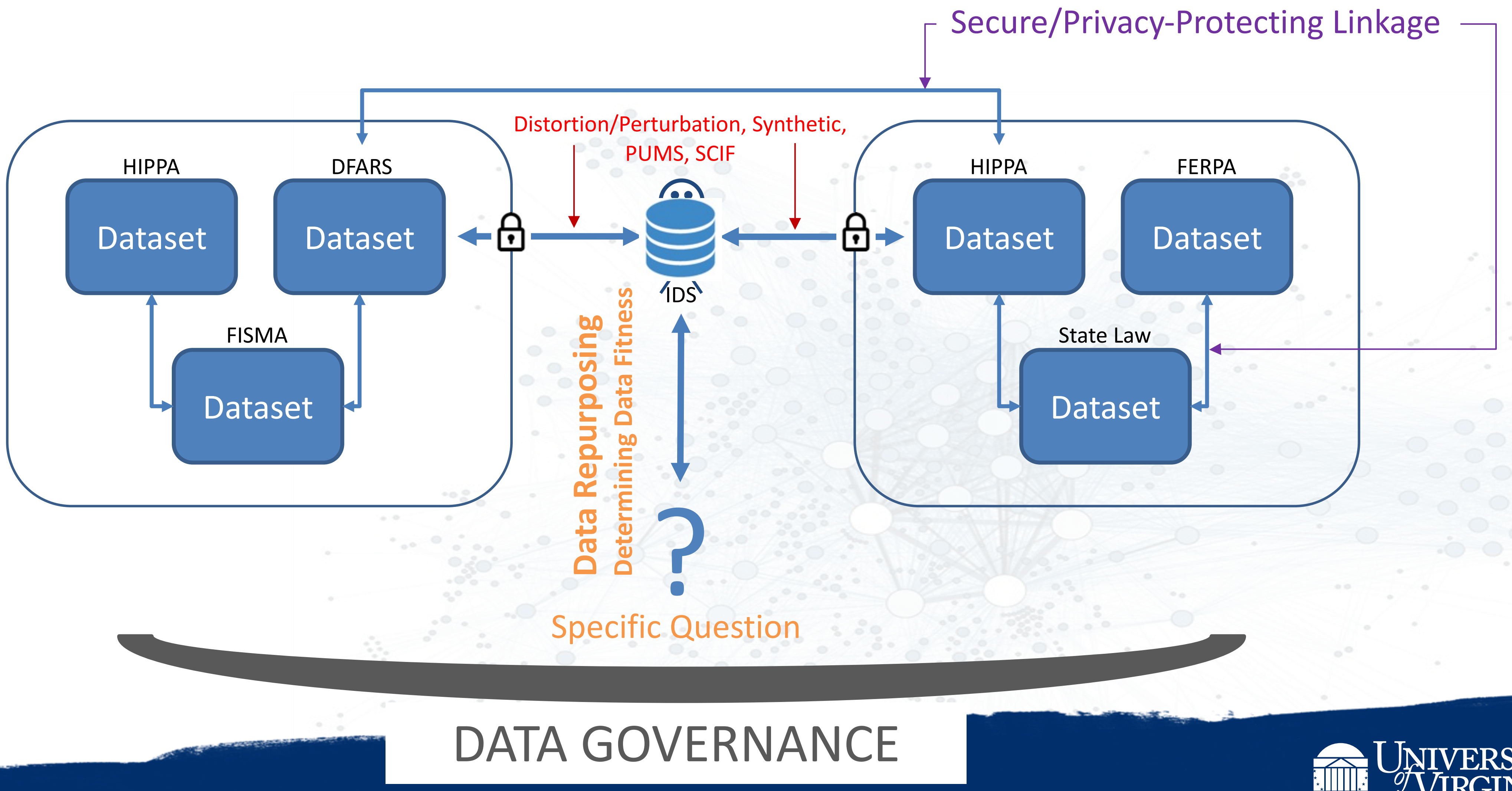
Aaron D. Schroeder, Ph.D.
Senior Data Research Scientist
Social & Decision Analytics Lab
Biocomplexity Institute of Virginia Tech

Every Repurposing Is a New "Investigation"

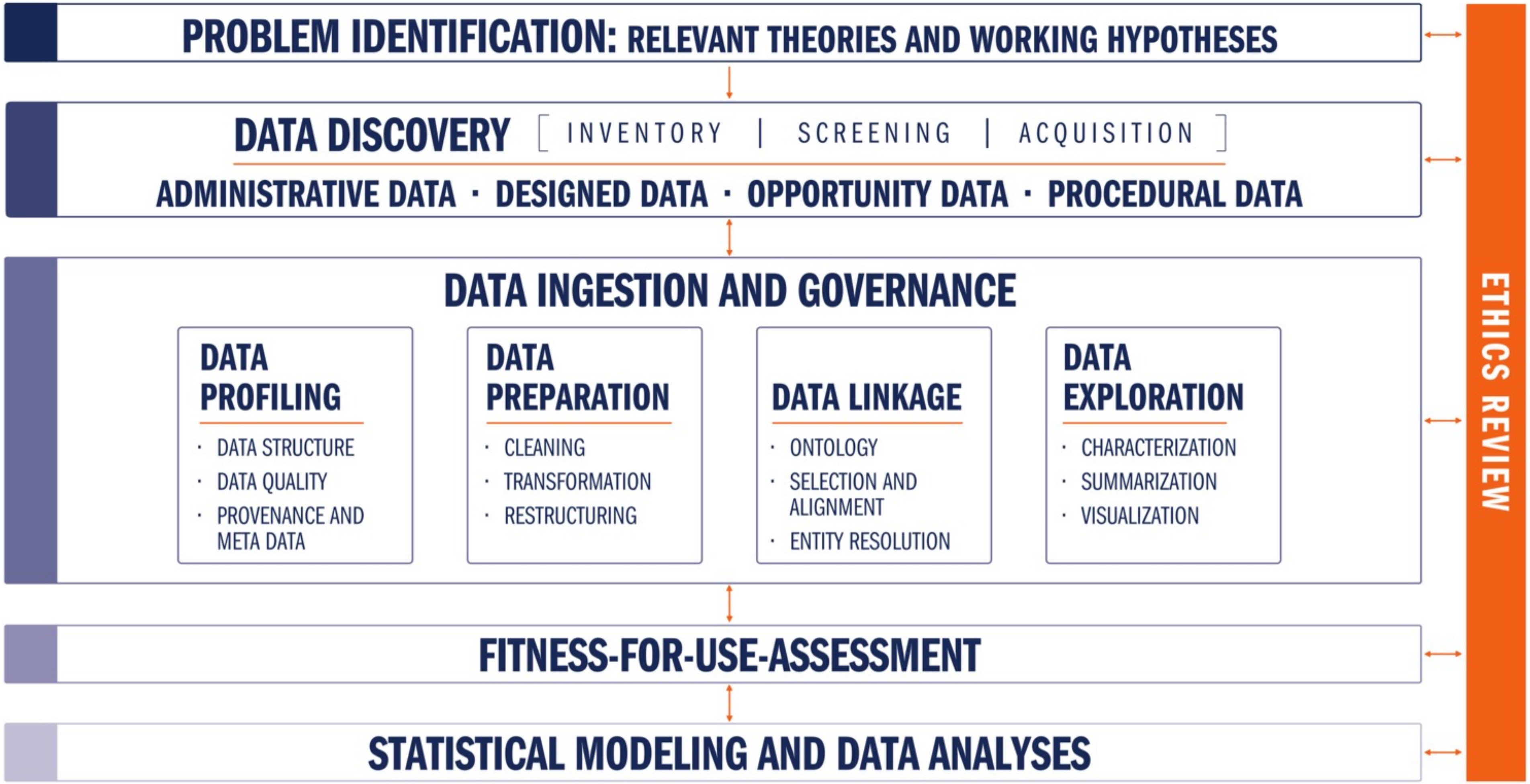
- Locating the Data Repurposing Discussion
- Overview of the SDAD "Investigative Process" for Repurposing Data
- Recommendations for Research-Enabling Standards for Integrated Administrative Data Systems to Aid Future Investigations

Data Repurposing

Locating the Discussion

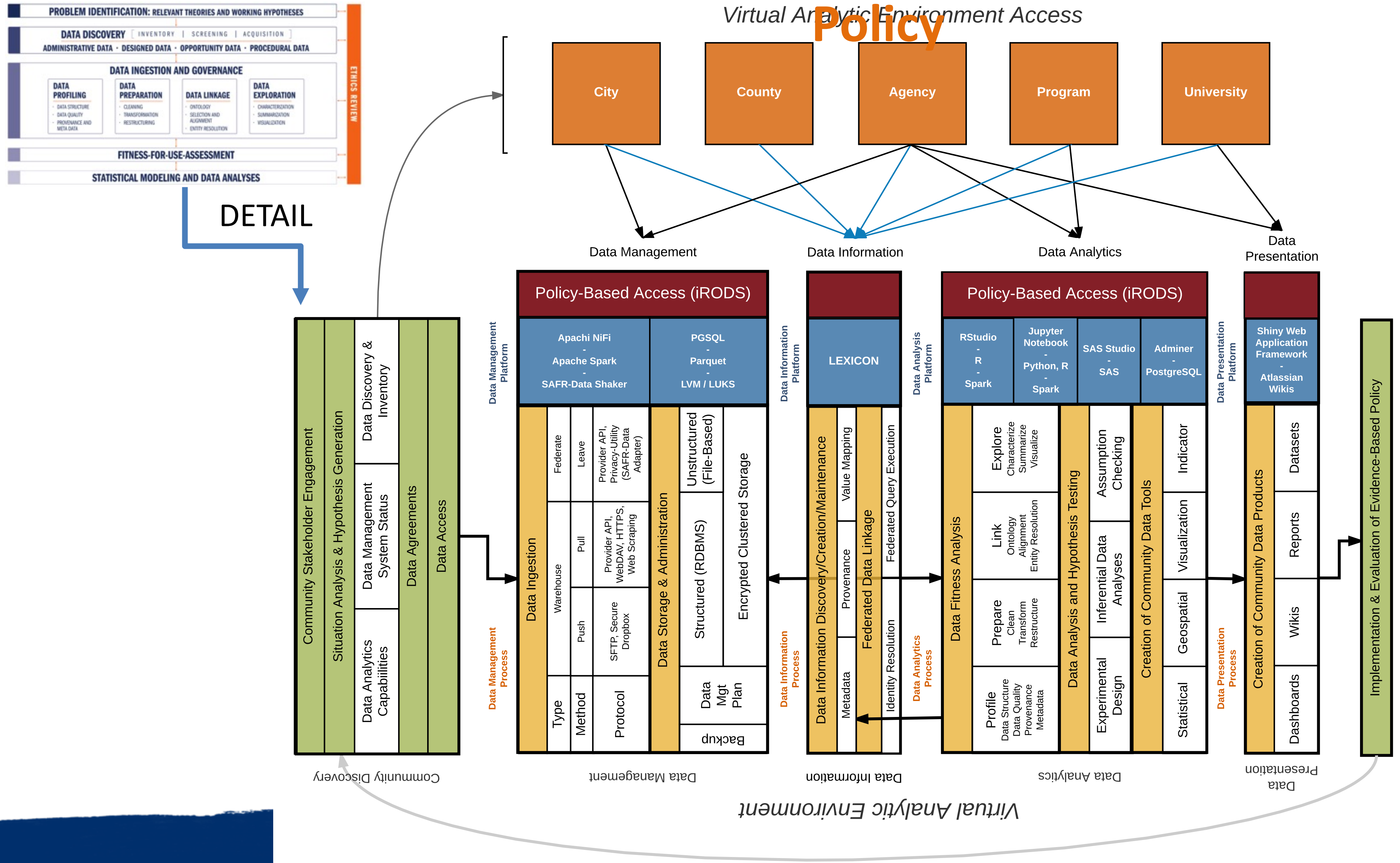


CLD3 - Data Science Processes for Evidence-Based Policy



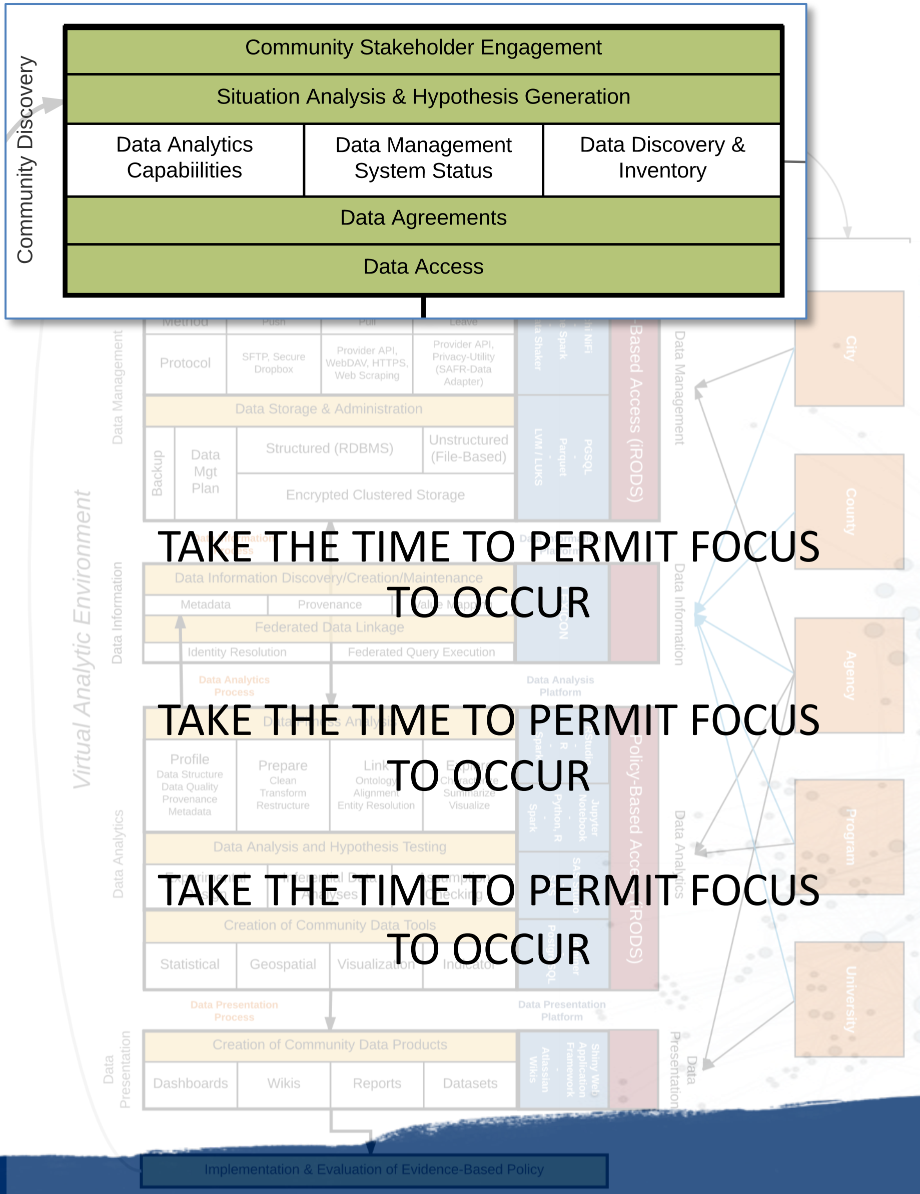
University of Virginia

CLD3 - Data Science Processes & Platforms for Evidence-Based Policy



UNIVERSITY of VIRGINIA

Data Science Processes & Platforms for Evidence-Based Policy



Community Discovery Process

- Facilitate preliminary **problem identification & hypothesis generation** starts with critical community-leader-defined issues and good *contextual assessment*
  - *THIS SHOULD TAKE A LONG TIME*
  - *Biggest BANG for the BUCK for facilitated sessions is here*
- Conduct **data management system status discovery** to ascertain methods and technologies currently employed
- Determine **data storage and management capacity requirements** of the entire process
- Conduct **data analytics capabilities assessment** – **are they going to use a statistical model you build? Or do they just need better counts?**
- Conduct **data discovery and inventory** process to identify potential data sources related to the specific issue areas
- Deploy **data connection technologies** as required by an already established data access plan to enable the data transfer and management

UNIVERSITY of VIRGINIA

Data Science Processes for Evidence-Based Policy

Community Discovery Process

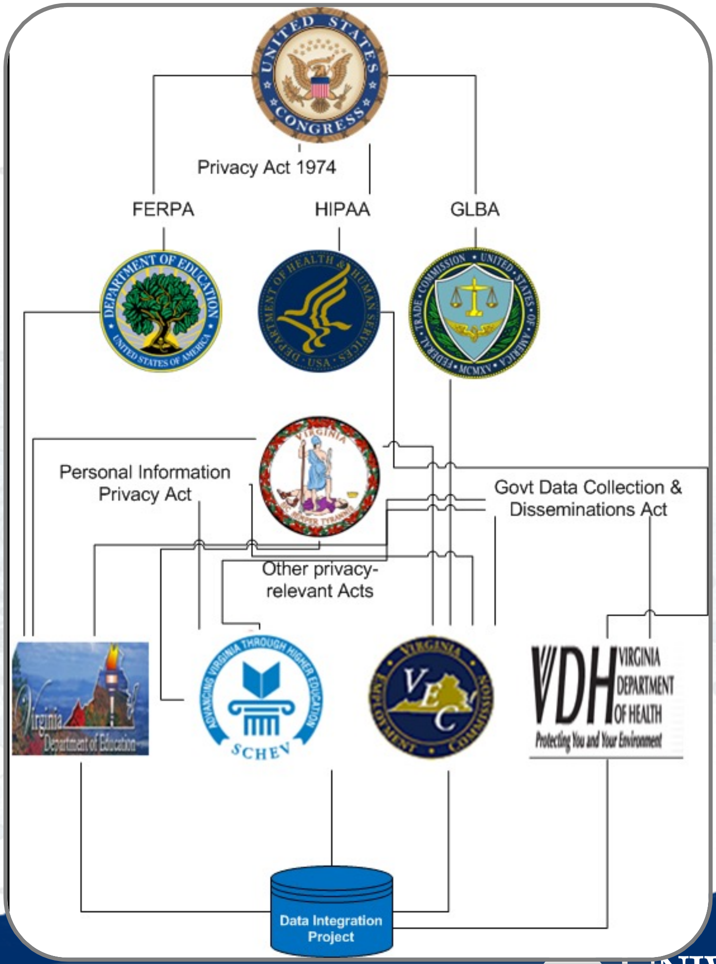
Situation Analysis & Hypothesis Generation: Context, Stakeholders, Joint-Visioning

Example: Implementation Environment of the Virginia Longitudinal Data System

- Multiple levels of statutory law
- Multiple implementations of regulatory law at each level of statutory law
- Most conservative interpretation of regulatory law becomes de facto standard

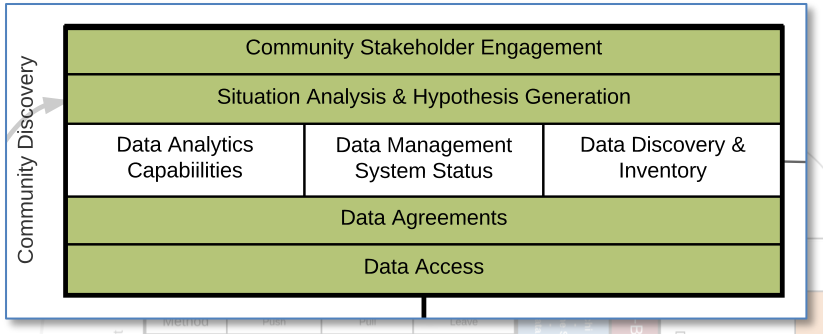
"No one person , inside or outside a government agency, should be able to create a set of identified linked data records between partner agencies"

- Has a direct and significant effect on the potential success of the technical approach chosen – A Centralized, Hierarchical Data Warehouse will likely Fail!
- Easy to see, if you look for it!



- Learn the Language! e.g. Virginia Towns

Data Discovery & Inventory



The first step preliminary inventory

- identify which sources are worthy of a deeper screening
- includes 6 questions and a qualitative evaluation of purpose, data collection method, timeliness, selectivity, accessibility, and description

1. Are the data collected opinion-based, (*e.g., people's attitudes, preferences, etc.*)?
2. Are the data collection recurring, (*i.e., must be collected at least annually*)?
3. Are there data available for 2013?
4. Geographic granularity

   For Education
   - Are the data collected at least the school level?
   - Can the data be linked to other education/workforce datasets, (e.g., K-12, higher education, workforce)?
   - If this is a state dataset, how do they define school districts within this state?
   - If applicable, what types of schools does it cover, (*e.g., public, private, charter*)?

   For Housing
   - Are the data collected at the property or housing unit level?

<u>Additional Screening Information</u>

Purpose:
- What is the purpose of the organization collecting the data, (*e.g., the Virginia Department of Education (VDOE) coordinates education for the state and makes policy recommendations*)?
- Why are the data collected and how does the organization use the data, (*e.g., VDOE collects the data for administrative purposes to assess student and school progress and to inform school policies*)?
- Who else uses these data, (*e.g., businesses, policy-makers, citizens, researchers*)?
- Who do they sell the data to, (*e.g., Zillow for individual homeowners, CoreLogic for multiple uses, business for economic development, Chief Economists at trade associations*)?

Method:
- What is the data collection method, (*e.g., paper questionnaire, operator entry, online survey, interview, sensors, algorithms for creating datasets from twitter feeds*)?
- What is the type of data collected, (*e.g., designed collection, intentional observation, administrative data, digital data*)?
- If designed, who created the questions, (*e.g., government, researchers, private business*)?
- What are the raw sources of the collected data prior to any aggregation, (*e.g., self-report, third party*)?

Description:
- What is the general topic of the data, (*e.g., student learning, housing quality*)?
- What are the earliest and latest dates for which data are available, (*e.g., 1995-2005*)?

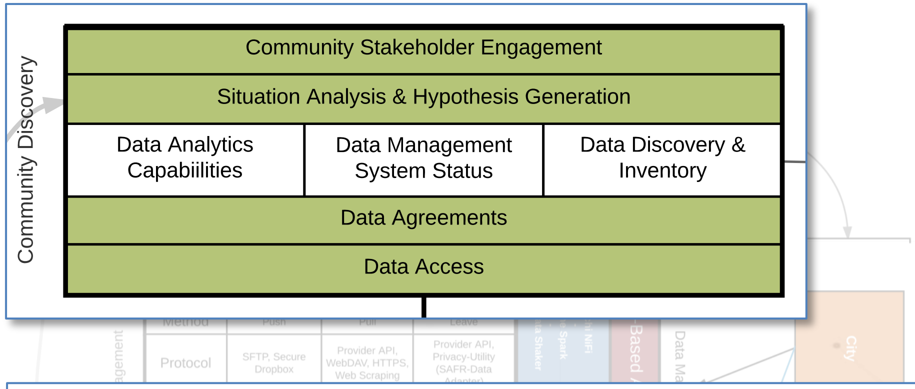
Timeliness:
- Are the data collected and available periodically, (*e.g, every year or decade*)?
- How soon after a reference period ends can a data source be prepared and provided, (*e.g., one year*)?

Selectivity:
- What is the universe (*e.g., population*) that the data represents (*e.g., students who attended public school in Virginia in 1995*)?

Accessibility:
- How are the data accessed, (*e.g., API, downloaded - csv, txt, etc.*)?
  - Are they open data?
  - Any legal, regulatory, or administrative restrictions on accessing the data source?
  - Cost? Is it one-time or annual or project-based payment?
- Describe any gaps/concerns you see with this dataset

Does this dataset appear to meet for the needs for the Census Bureau study? Yes/No

Data Discovery & Inventory



- Conducted on a selected subset from inventory
- Much deeper dive
- Output is a subset of data sets selected for acquisition and initial analysis of their fitness for analysis

Description/Features

- What is the temporal nature of the data: longitudinal, time-series, or one time point?
- Are the data geospatial? If Yes, at what level, (*e.g. census tracts, coordinates*)?

Metadata

- Is there information available to assess the transparency and soundness of the methods to gather the data for our purposes, (*i.e., supplementing the census*)?
- Is there a description of each variable in the source along with their valid values?
- Are there unique IDs for unique elements that can be used for linking data?
- Is there a data dictionary or codebook?

Selectivity

- What unit is represented at the record level of the data source, (*e.g., person, household, family, housing unit, property*)?
- Does this universe match the stated intentions for the data collection? If not, what has been included or excluded and why (*e.g., do the data exclude certain individuals due to the way the data are collected*)?
- What is the sampling technique used (if applicable, *e.g., convenience, snowball, random*)?
- What is the coverage, (*e.g. response rate*)?

Stability/Coherence

- Were there any changes to the universe of data being captured (including geographical areas covered) and if so what were they, (*e.g., changed the geographical boundaries of census tracts*)?
- Were there any changes in the data capture method and if so what were they, (*e.g., revised questions, data collection mode, classification categories, algorithms for social media data*)?
- Were there any changes in the sources of data and if so what were they, (*e.g., data were reported by teachers in 2010 and reported by principals in 2011; used Current Population Survey in 2011 and American Community Survey in 2012*)?

Accuracy

- Are there any known sources of error, (*e.g., missing records, missing values, duplications, erroneous inclusions*)?
- Describe any quality control checks performed by the data's owner, (*e.g., deleted duplicates, checked for recording errors*).

Accessibility

- Are any records or fields collected, but not included in data source, such as for confidentiality reasons, (*e.g., does not include any student files in which there are less the 5 students in a category*)?
- Is there a subset of variables and/or data that must be obtained through a separate process, (*e.g. state level data openly available, but one must apply to get census tract*)?
- If yes, is there a separate legal, regulatory, or administrative restrictions on accessing the data source?
- Cost? Is it a one time, annual, or project-based payment?

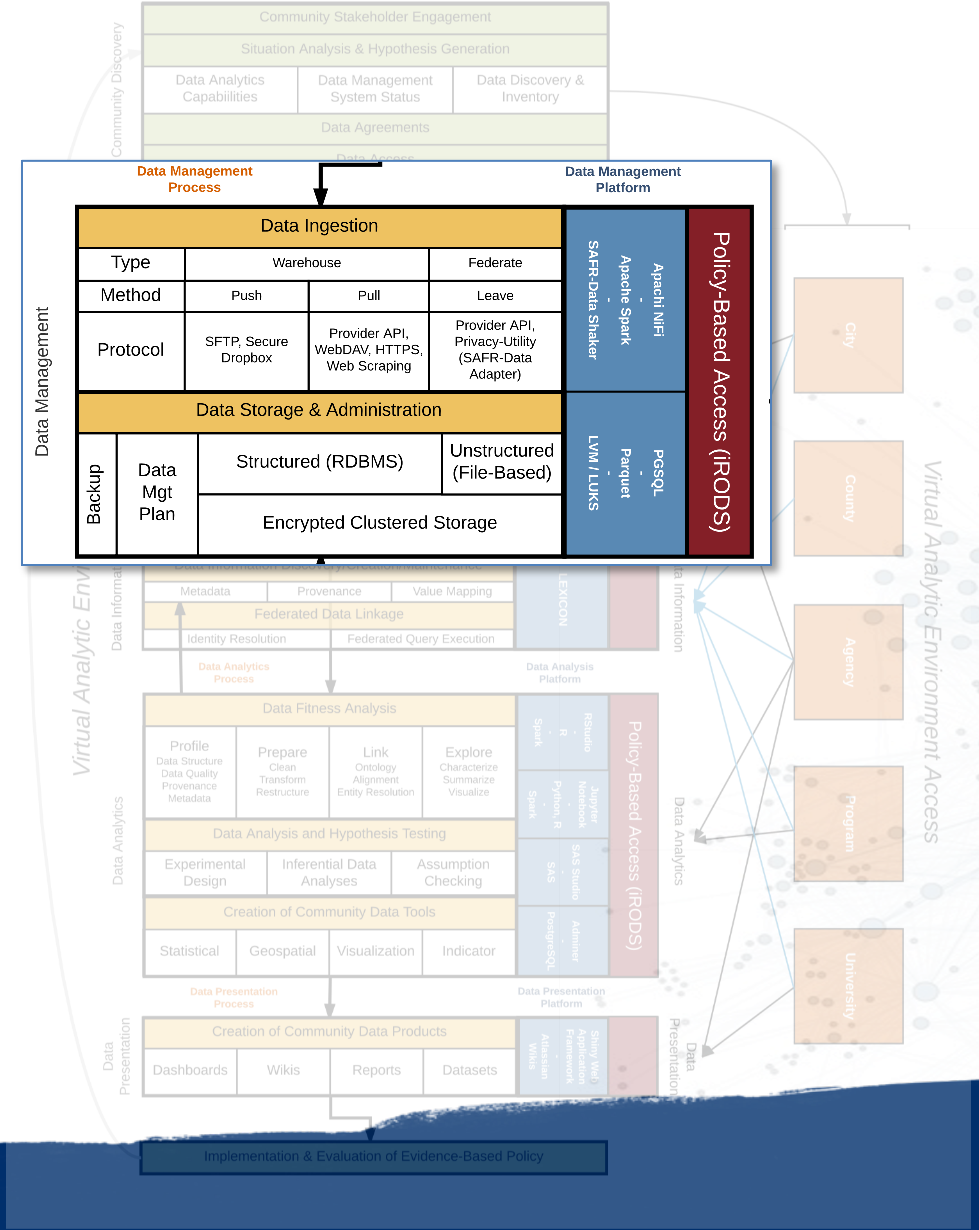
Privacy and security

- Was consent given by participant? If so, how was consent given, (*e.g., online form, in-person discussion*)?
- Are there legal limitations or restrictions on the use of the data, (*e.g., Family Educational Rights and Privacy Act -FERPA*)?
- What confidentiality policies are in place, (*e.g., cannot share data outside of requesting institution; does not include personally identifiable information*)?

Research

- What research has been done with this dataset, (*e.g., impact of policies, predictors of student success, housing stock inventory assessment*)?
- Include any links to research if provided.
- List any other data use notes provided by the supplier.

Data Science Processes for Evidence-Based Policy

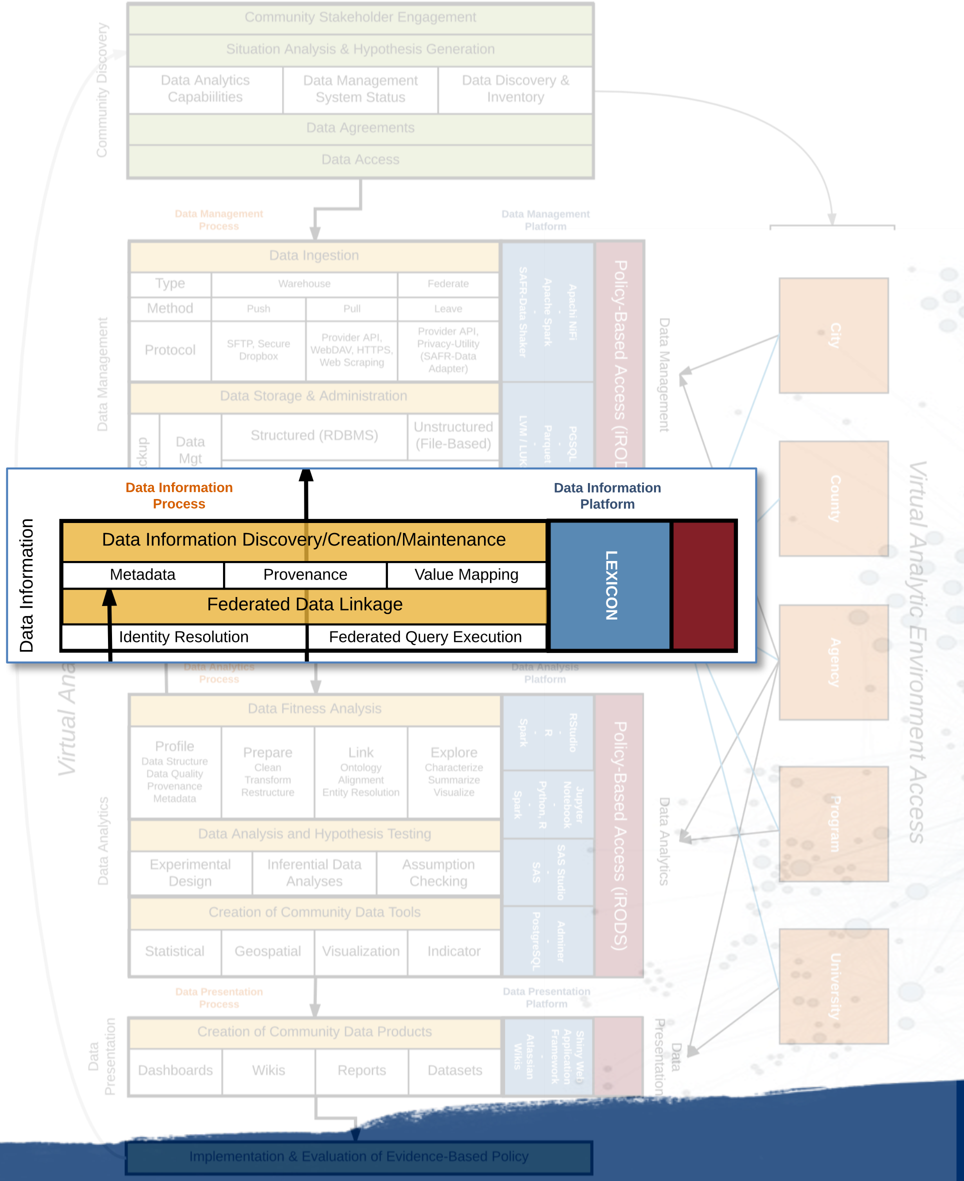


Data Management Process & Platform

- Establish **type and method of data transfer**
  - pushed to or pulled into the cooperative platform?
  - staying where it is and being dynamically queried in a federated manner as needed?
- Establish the **best transfer protocol(s)** to use given the types and method of transfer
  - e.g. SFTP, secure Dropbox, secured REST API, VT SAFR-Data Adapter for secure federated queries
  - Establish designed collection systems (e.g. behavioral experiments)
- Establish **data marshaling processes**
  - system mediation logic, data pipeline and data transformation, transfer schedule, and data provenance maintenance
- Establish **secure data storage procedures**
  - e.g. each project being stored on a new project-dedicated encrypted partition, original data being stored as non-removable and non-editable

UNIVERSITY of VIRGINIA

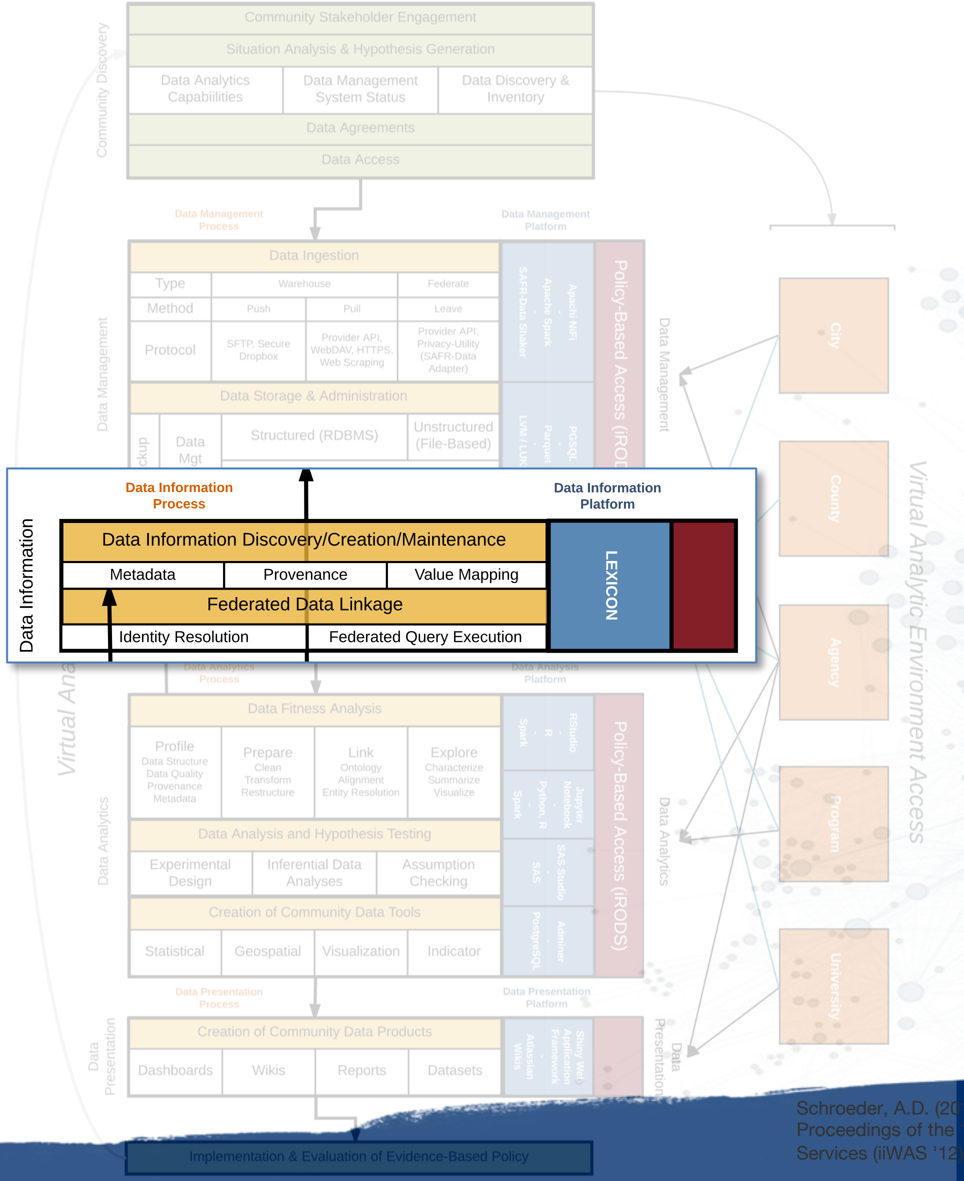
Data Science Processes for Evidence-Based Policy



Data Information Process & Platform

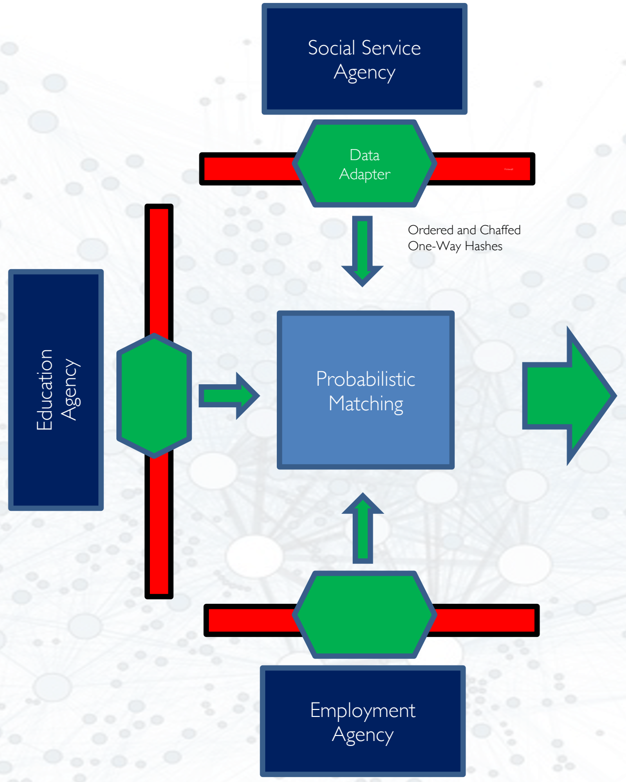
- The **Lexicon**: an inventory of and history of changes to:
  - every available data field in every available data source
  - the structure of their storage
  - possible values and meanings of the information
  - possible transformations of each set of field values from one data source to another another data source
  - methods of data source access
  - matching algorithms and how they are to be used in conjunction with possible field value transformations
- Provides fundamental functions for the operation of the framework and is a **requirement** that the data information be collected from all partner communities
- Enables removal of much complexity required for high quality data linkage
  - i.e. No enforcing data standardization schemes on data partners

Data Science Processes for Evidence-Based Policy



Data Information Process & Platform

Federated Data Linkage



Social Service Data			
NewID	Gen	FIPS	SES
7	F	51059	Med
2	F	51059	Low
9	M	51059	High

Education Data			
NewID	Attend	ACT	SAT
7	.97	32	1580
2	.71	18	2110
9	.86	22	1360

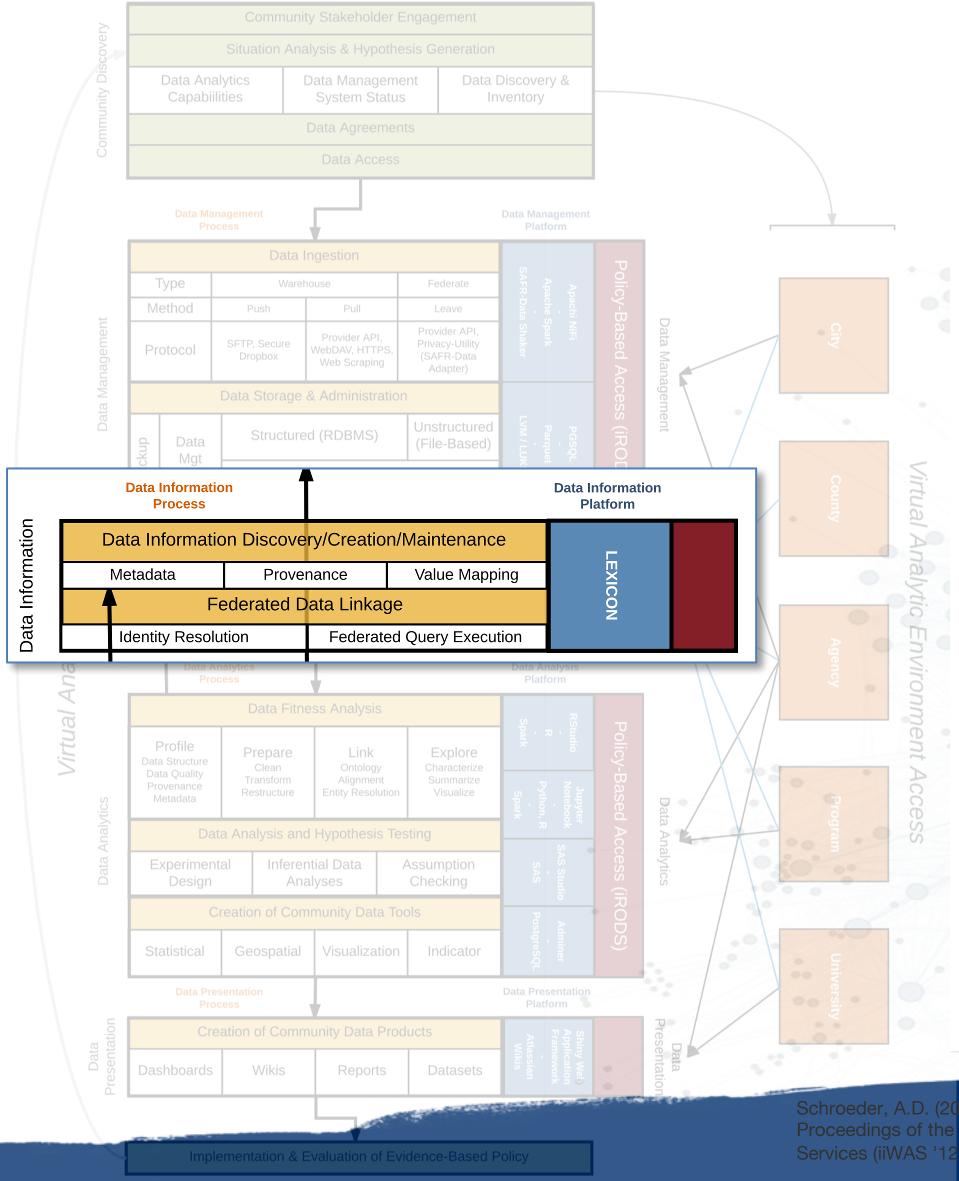
Employment Data			
NewID	Type	Length	perHr
7	0310	36	16.25
2	0340	6	22.50
9	7800	21	13.75

Schroeder, A.D. (2012). Pad and Chaff: Secure Approximate String Matching in Private Record Linkage. Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services (iiWAS '12), pp. 121-125. DOI=10.1145/2428736.2428757. ACM, New York, NY, USA.

https://github.com/dads2busy/SAFRLink-DataAdapter
https://github.com/dads2busy/SAFRLink-Identity
https://github.com/dads2busy/SAFRLink-Crypto

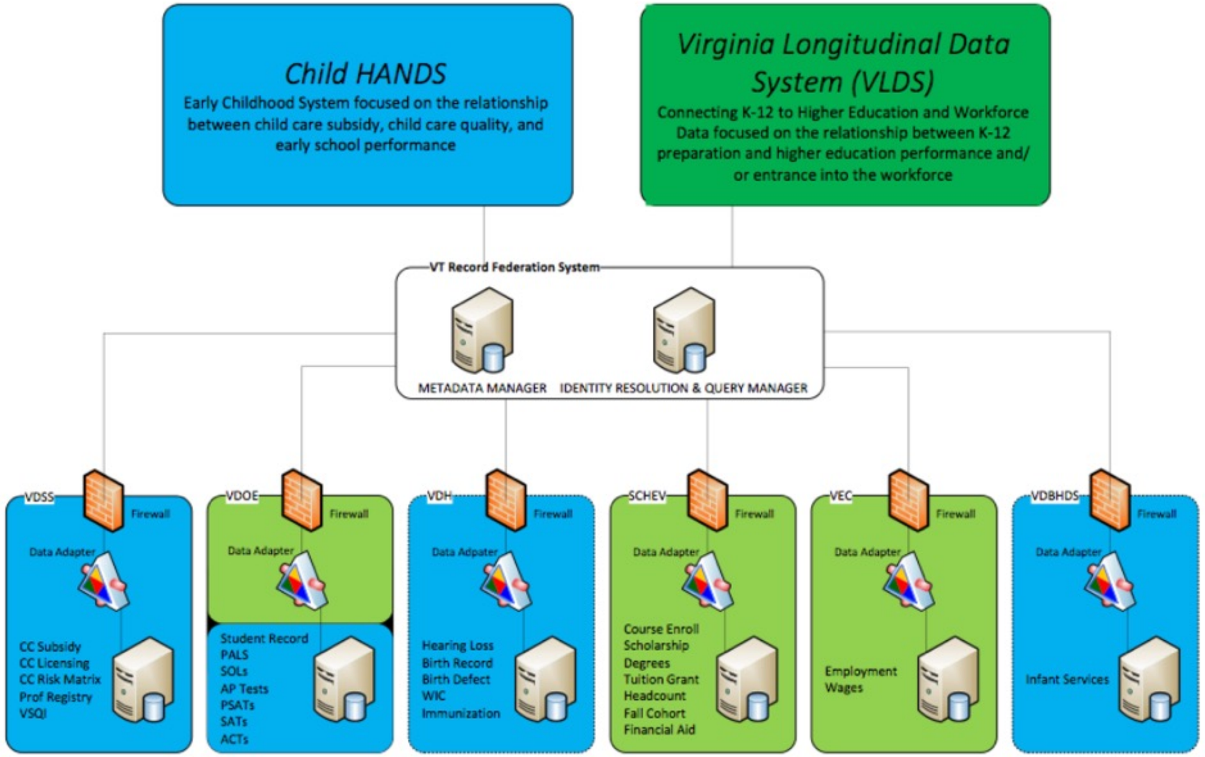
UNIVERSITY of VIRGINIA

Data Science Processes for Evidence-Based Policy



Data Information Process & Platform

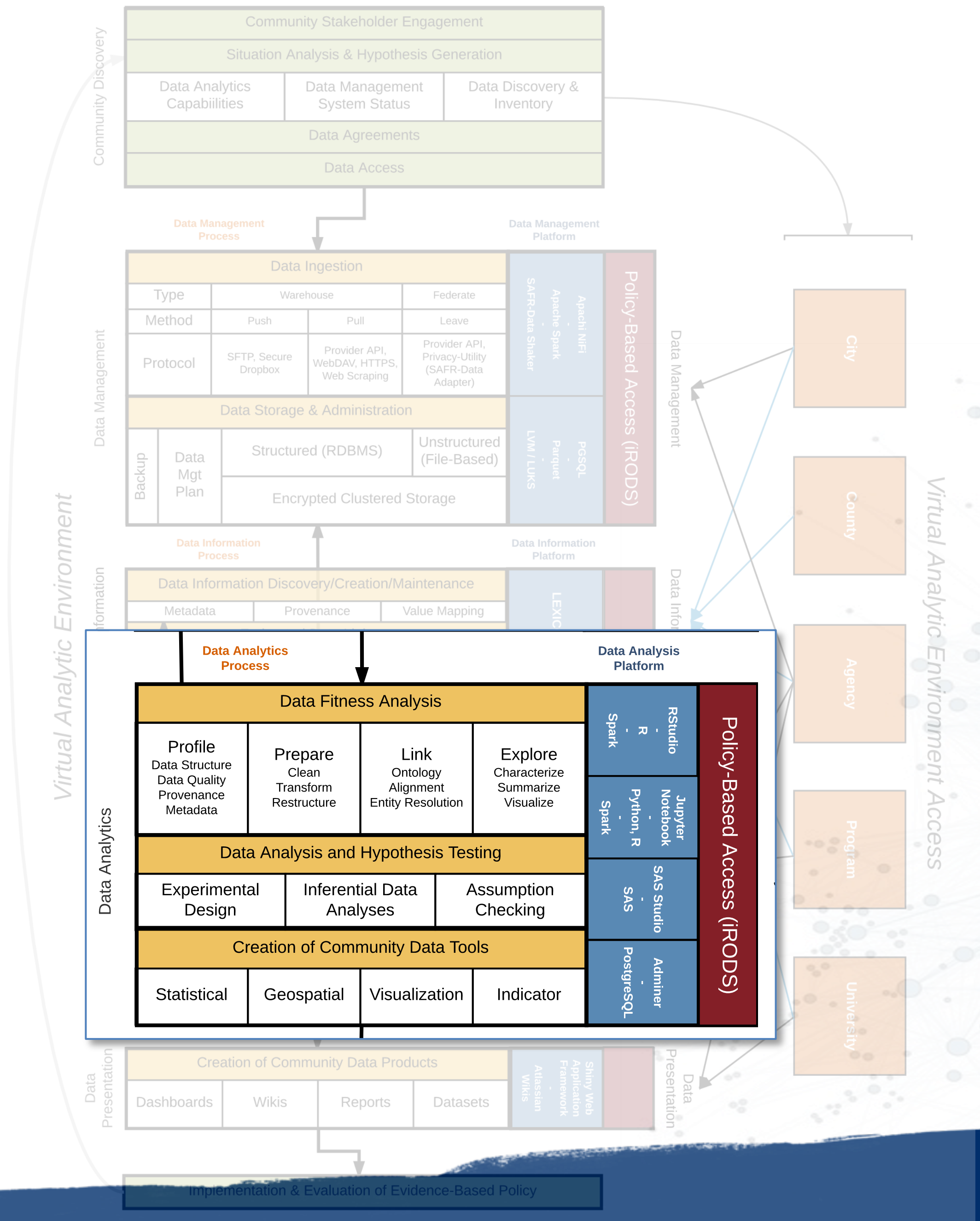
Federated Data Linkage



Schroeder, A.D. (2012). Pad and Chaff: Secure Approximate String Matching in Private Record Linkage. Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services (iiWAS '12), pp. 121-125. DOI=10.1145/2428736.2428757. ACM, New York, NY, USA.

https://github.com/dads2busy/SAFRLink-DataAdapter
https://github.com/dads2busy/SAFRLink-Identity
https://github.com/dads2busy/SAFRLink-Crypto

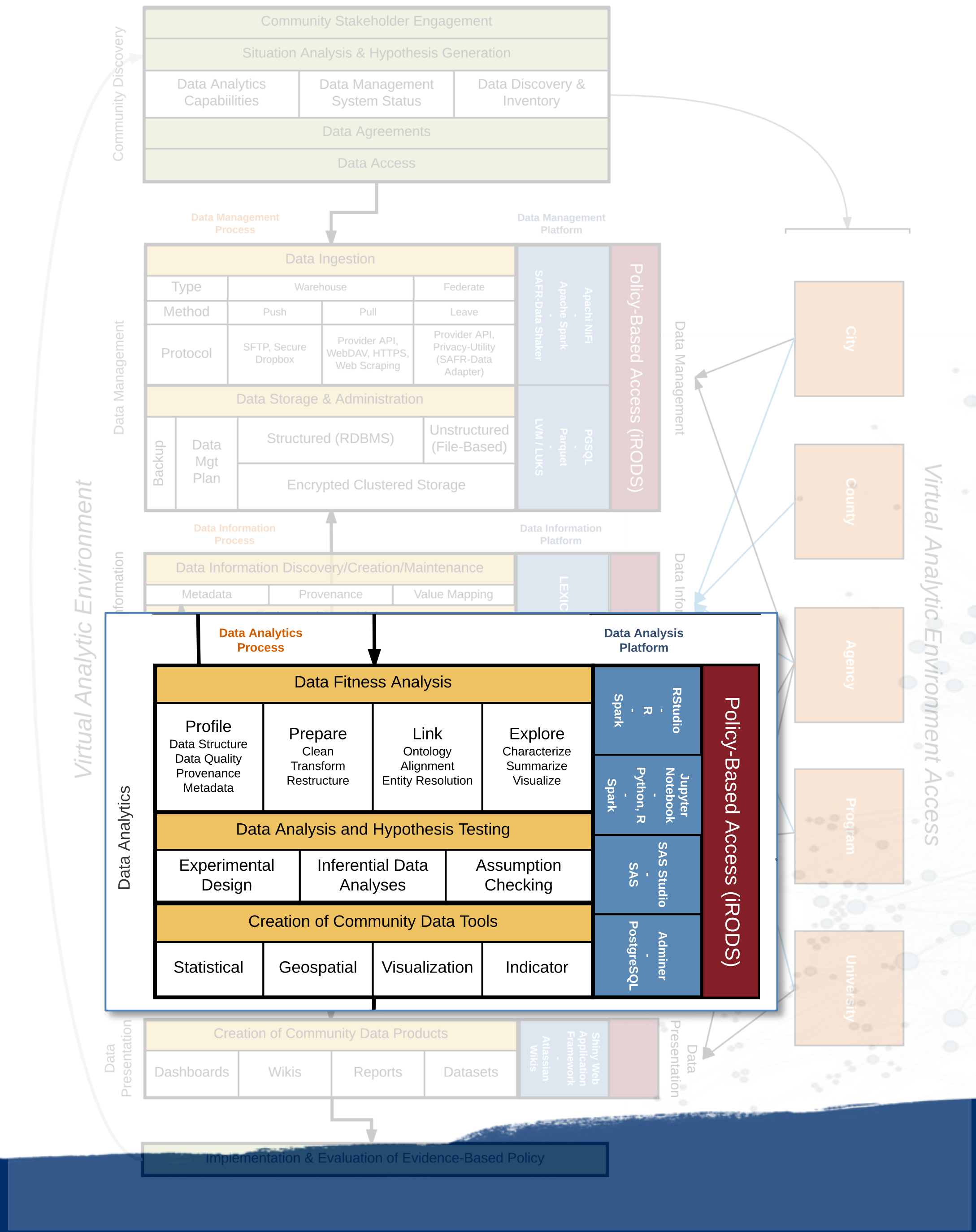
UNIVERSITY of VIRGINIA



Data Science Processes for Evidence-Based Policy

Data Analytics Process

- Data Fitness Analysis (Data Re-Purposing)
  - **Modeling is a function of the research question (the use) – drives all data actions**
  - Fitness assessment is about fitness of the data for the model
  - Fitness is a function of the model, data quality needs of the model, and data coverage (representativeness) needs of the model
  - When using multiple data sources, fitness will need to assess linking across data sources
  - Fitness must characterize information content in the results
    - Accuracy and precision



Data Science Processes for Evidence-Based Policy

Data Analytics Process

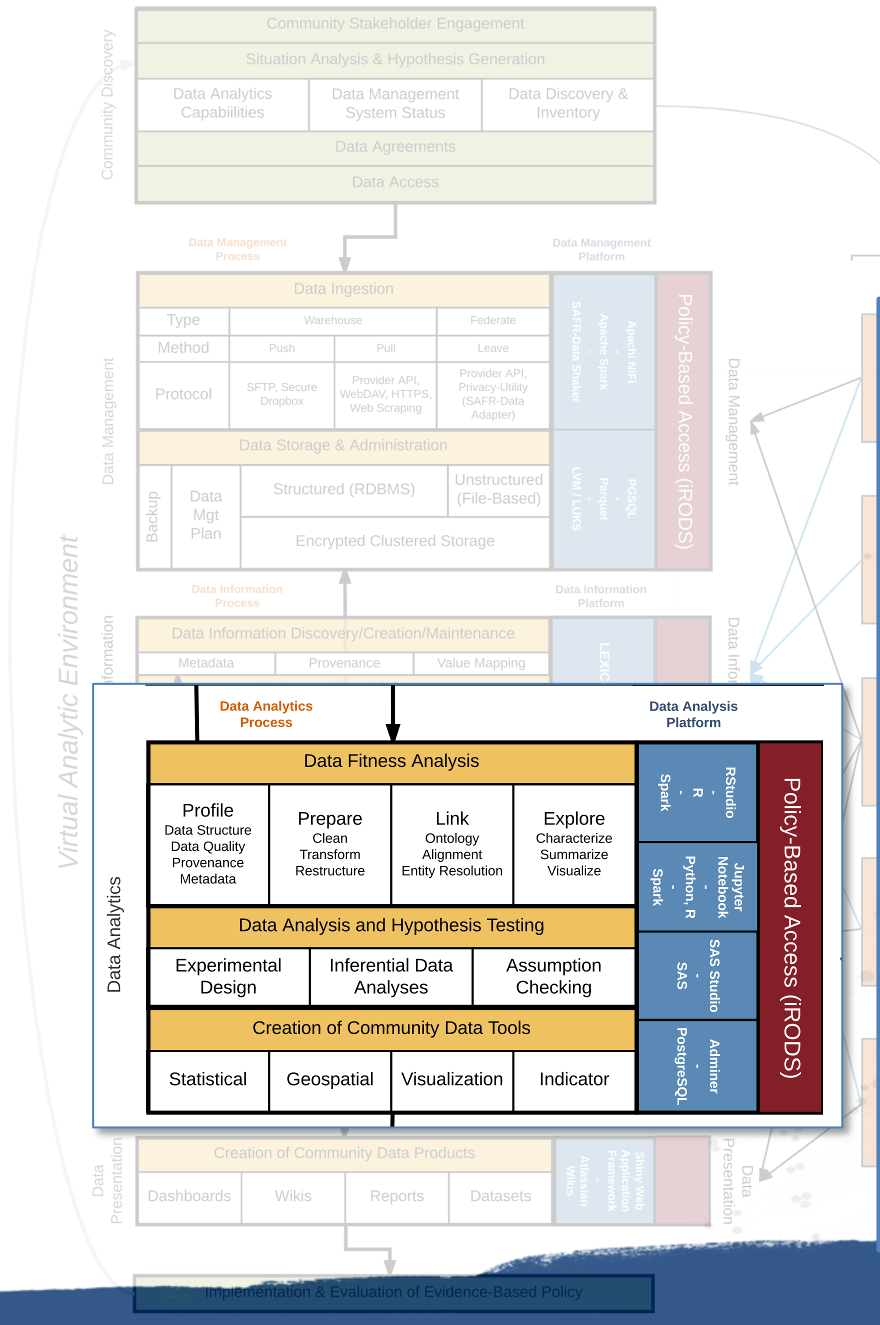
- Data Fitness Analysis
  - **Data Profiling**
    - **Structure**
    - **Quality**
    - **Provenance & Metadata**
  - Data Preparation
  - Data Linkage
  - Data Exploration
- Data Analysis & Hypothesis Testing
- Creation of Community Data Tools

University of Virginia

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance



Missing Variables

values in column headers instead of variable names

e.g. Value-ranges being used as column headers (0-9|10-19|20-29|…)

Combined Variables

more than one variable represented in a attribute (column) value

e.g. An attribute combining gender and age (m25, f32,…)

Multiple Observation Directions

variables in both columns and rows

e.g. A dataset with an element(column) for each day of the month (horizontal) and an element(column) for 'month' (vertical)

note. the messiest and can be dealt with multiple ways according to the needs of the specific analysis

Combined Observation Unit Types

more than one observation unit type per table

e.g. A table containing both individual demographic data and a periodic measurement like weekly attendance where demographic data and weekly attendance are separate observational units and need to be in separate datasets.

Divided Observation Unit Type

observation unit type is split among multiple tables

e.g. Individual demographic information split among several datasets; for example, separate tables for gender, ethnicity, and surname.

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Missing Variables

Emergency Services Data Processing Example

PRID: 12144089 **Incident Number:** 110010048 **CAD Call Number:** 110000053

Service: Arlington County Fire Department (State ID: 00071)
Base: Station 05
Unit: [redacted] (Transport)
Shift: A Shift
EMD: Not Available
Dispatched As: Fall injury(s)
Mass Casualty: No
Vehc. Disp. GPS: 38.85722,-77.056908
Type of Svc: Scene Unscheduled
Response Code: Priority 01
Mode to Ref: Lights / Sirens
Moved Via: Stretcher
Position: Semi-Fowlers
Outcome: Treated, Transported by ACFD

Date: January 1, 2011
Team: ALS
Crew 1: *Primary Caregiver*
[redacted]
EMT-P
Crew 2: *Driver*
[redacted]
EMT-I

** designates an ALS Provider*

Mode to Rec: No Lights/Sirens
Moved From: Stretcher
Pt. Condition: Improved
CMS Service Level: ALS, Level 1 Emergency

Ref Other Type: *Public Building*
Location: 1000 S HAYES ST
ARLINGTON, VA 22202-4901
Requester: [redacted]-LOSS PREVENTION
Ref. GPS: 38.865246,-77.058892

Receiving: *Hospital*
Virginia Hospital Center (Arlington)
Emergency Department
1625 North George Mason Drive
Arlington, VA 22205-3698
(703)558-5000
Dest. GPS: 38.888911,-77.128434
Rec. RN: [redacted], PA
Destination Basis: Protocol

Last Name: [redacted] **First:** [redacted]
Address: [redacted]
City: [redacted] **ST:** MD **Zip:** 20748
County: Prince Georges
Citizenship: United States
Phone: [redacted]
DOB: [redacted] **SSN:** [redacted]
Age: 44y **Sex:** F **Weight:** 140 lb
Height:
Subscriber: No
Race: Black, non-Hispanic
Barriers to Care: None Noted

Times	
Received:	11:22:30
Notified:	11:23:27
Dispatch:	11:23:32
EnRoute:	11:25:03
At Ref:	11:26:55
At Patient:	12:00:15
Leave Ref:	11:30:00
At Rec:	12:12:00
Transfer Care Dest:	12:20:00
Available:	12:50:00

Scene Information

Description: Pt was sitting in a chair in the security office of Macy's being tended to by security staff.
Num. Patients On Scene: 1

Chief Complaint (Category: Fall injury(s))

Laceration to R knee

Duration: 20 Minutes
Anatomic Location: Extremity - Lower

Secondary Complaint

Hypertension

History of Present Illness

M105 AOSTF a 44 YOF sitting on a chair in the Macy's security office holding a 4x4 on her right knee. Pt is CA&Ox3 w/a patent airway. Visual examination reveals an approx 2-inch laceration through the sub-cutaneous tissue. Pt denies neck or back pain and LOC. Applied steri-strips to close wound, then applied cold pack and wrapped with cling. Pt states that she tripped over an expansion joint in the floor in the mall, landing on her R knee. +PMS in the affected extremity. Pt initially indicated she would have her boy friend drive her to the hospital in MD, however, VS reveal relative hypertension that does not decrease with time. Pt agreed to be transported by medic unit. Assisted pt onto stretcher and moved her to the medic unit. In the medic unit, established IV access via 20g angiocath in the RAC w/NS Lock. BGL=84. Obtained 4-lead ECG: NSR. Initiated transport. Notified hospital via radio. Monitored pt's BP enroute to hospital. Transported pt to VHC-Arlington where pt was placed in Express Care Exam 4 and care was transferred to [redacted], PA, w/out incident.

Medical History	Current Medications	Allergies
Hypertension Obtained From: Patient	None	None

Neurological Exam

Level of Consciousness: Alert **Loss of Consciousness:** No
Chemically Paralyzed: No
Neurological Present: Normal
Mental Present: Normal

Glasgow Coma Scale	E	V	M	Tot
Int:	4	5	6	= 15

Pupils	Left	Right
Size:	Normal	Normal

	Motor	Sensory
LA:	Normal	Normal
RA:	Normal	Normal
LL:	Normal	Normal
RL:	Normal	Normal

Airway

Status: Patent

Respiratory

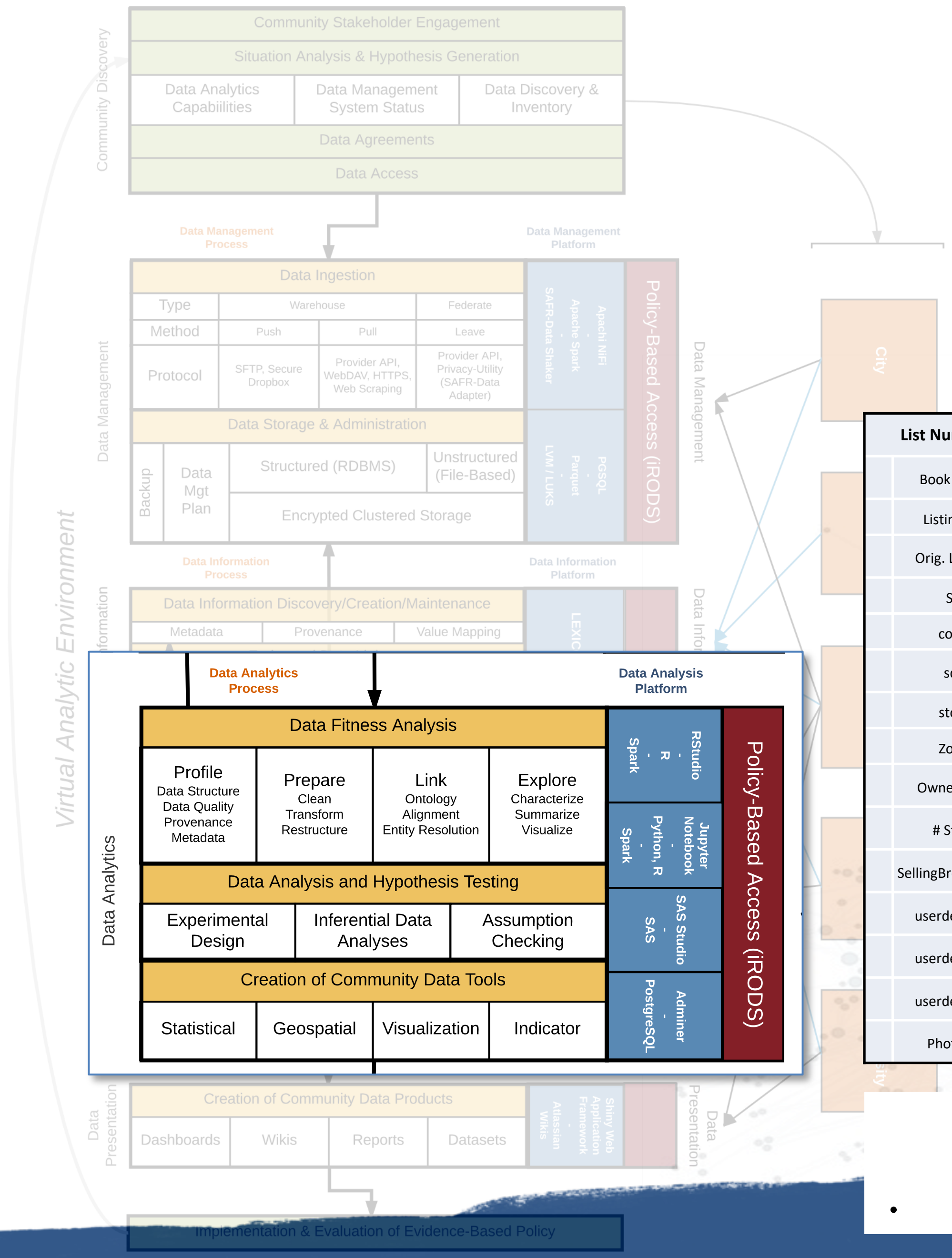
Effort: Normal
Sounds: **L:** Clear **R:** Clear

Cardiovascular

JVD: Not Appreciated **Cap. Refill:** Less than 2 Seconds
Edema: Not Appreciated

Pulses	Left	Right
Carotid:	Not Checked	Not Checked
Radial:	Strong	Not Checked
Femoral:	Not Checked	Not Checked
Dorsalis:	Not Checked	Not Checked

Values of a variables used for column headers



Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Combined Observation Unit Types

Current Structure of Williamsburg MLS Data

List Number	Agency Name	Agency Phone	Agency Email	Listing Agent	Listing Agent Phone	Listing Agent Email	Co-Listing Agent	Property Type	Card Format
Book Section	Selling Agency	Selling Agency Phone	Selling Agency Email	Selling Agent	Selling Agent Phone	Selling Agent Email	Co-Selling Agent	End Date	book_sec
Listing Date	Sold Date	Under Cont. Date	Fall-thru Date	Status	Status Change	Withdraw Date	Cancel Date	Contingent	Cont. Remarks
Orig. List Price	Price	Sold Price	high_price	Low Price	assessed_val	Partial Tax Assmnt	financing	Area	Relocation
St. #	box_nbr	St. Dir.	Street Name	Address 2	streetdirsuffix	Street Suffix	carrier_route	City	State
county	country	Zip Code	geo_county	Taxes	geo_lat	geo_lon	Est. Fin. SqFt	sqft1	sqft2
sqft3	sqft4	Year Built	2+ Bdroms on 1st Flr	Realtor.com Type	lot_size	Total Acres	Condo Level	sell_broker_comm	Variable Commission
stories	Total Rooms	Total Bedrooms	total_bath	Baths - Full	Baths - Half	baths_3_4	Garage Type	garage_stall	Water Frontage
Zoning	taxes	Tax Year	Subdivision	Public Remarks	Agent Remarks	**Parcel ID**	Legal Description	Directions	Foreclosure
Owner Phone	Owner Name	Neighborhood	mod_timestamp	Ltd Service Agent	Occupied By	Owner/Agent	Mster Bdrm 1st Floor	SqFt Source	Listing Type
# Stories	# Fireplaces	Golf Frontage	IDX Y/N	Supplement Attached	Seller Concession(s)	Special Assmnts	Type	Rollback Taxes	userdefined16
SellingBroker Incent	Ownership	Describe Concession	How Sold	Selling Broker Comp	userdefined22	Assessed Value	Est.Unfinished Sq Ft	Tax Rate	Garage Bays
userdefined27	userdefined28	userdefined29	userdefined30	Est. Closing Date	userdefined32	userdefined33	Lot Description	Short/CompromiseSale	userdefined36
userdefined37	userdefined38	userdefined39	userdefined40	userdefined41	userdefined42	userdefined43	userdefined44	userdefined45	userdefined46
userdefined47	userdefined48	userdefined49	userdefined50	userdefined51	userdefined52	userdefined53	userdefined54	userdefined55	userdefined56
Photo URL	Days on Market	Rooms	Features						

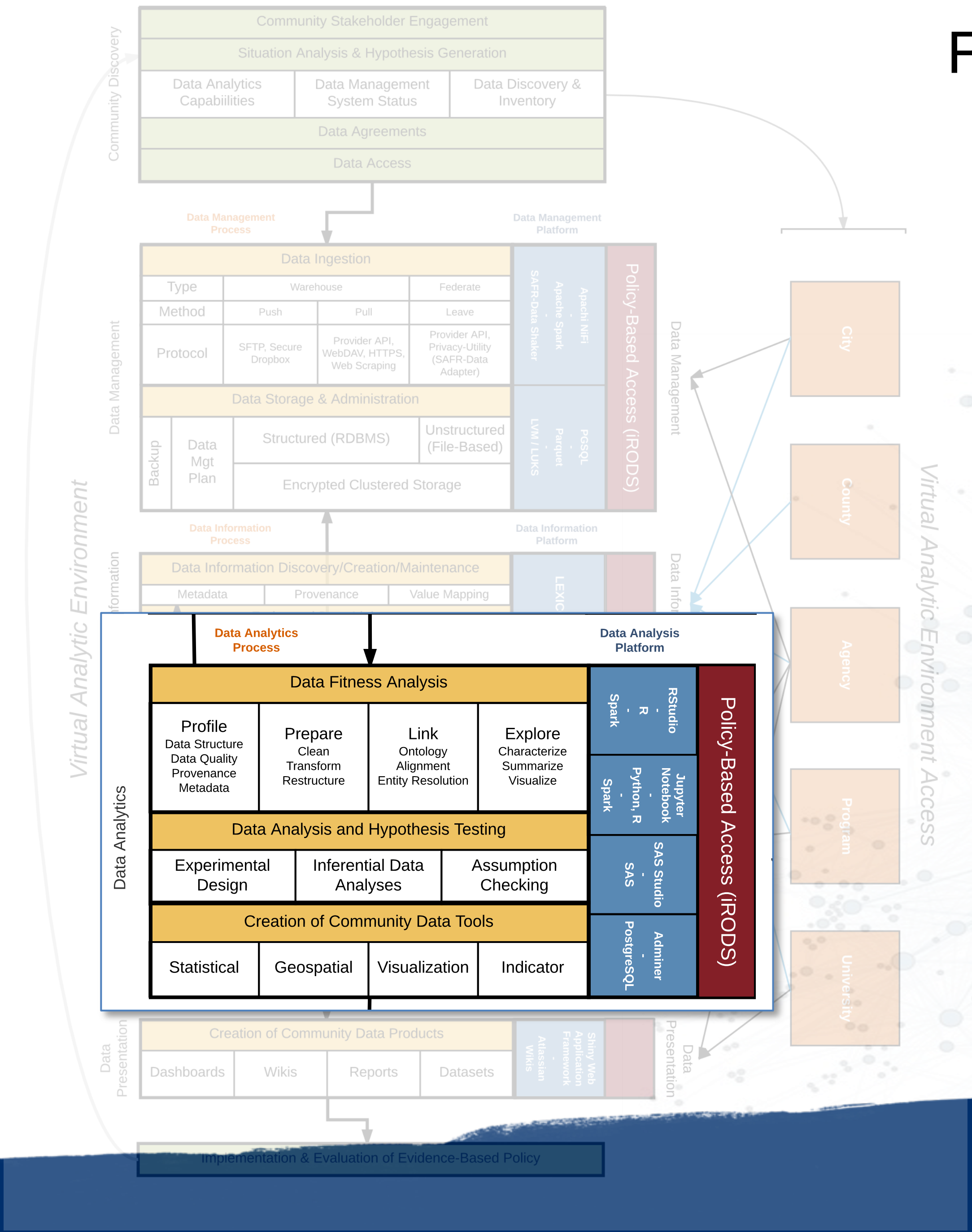
- This is a single record with 128 fields all keyed to the variable “List Number”
- Structured this way, it is not possible to analyze property changes over time
- Pulling out a definitive list of unique properties using “Parcel ID” seems like a possibility
- However, “Parcel ID” is left blank in over 7% of entries – extra work required – perhaps including address, but address is not standardized

Repurposing Data for Statistical Purposes

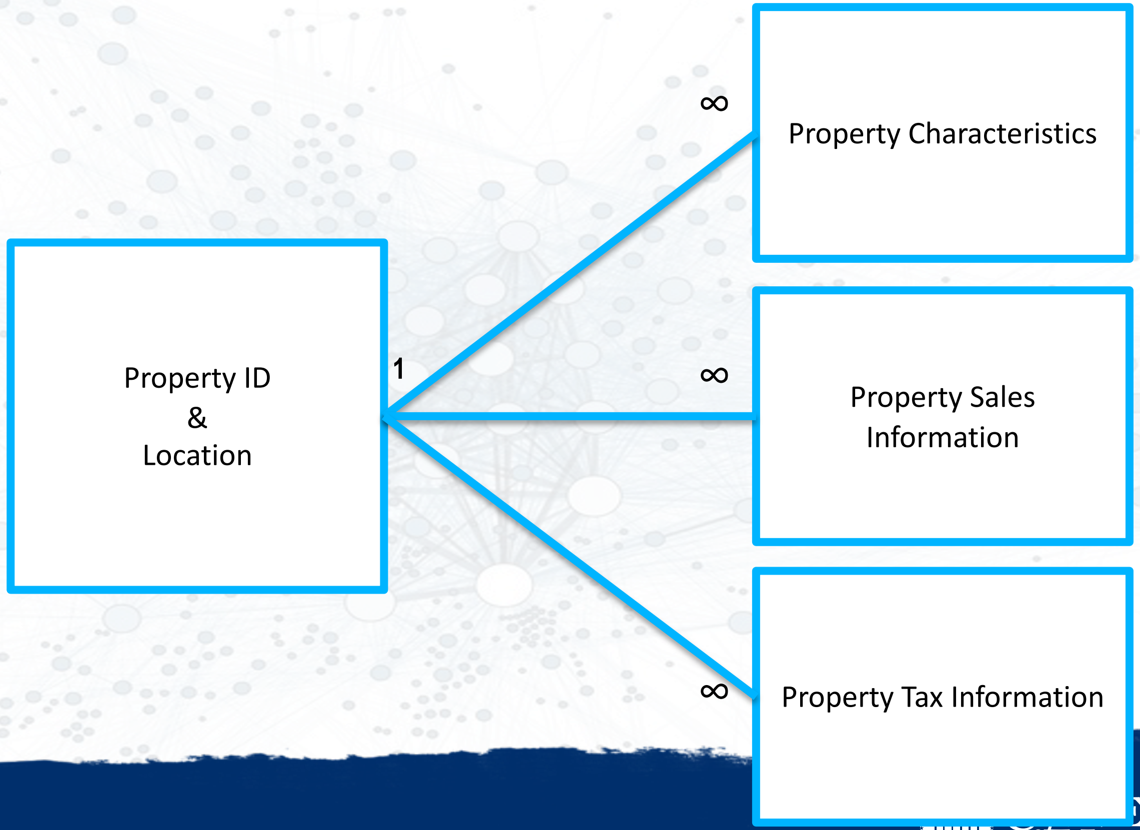
Data Fitness Analysis: Profiling

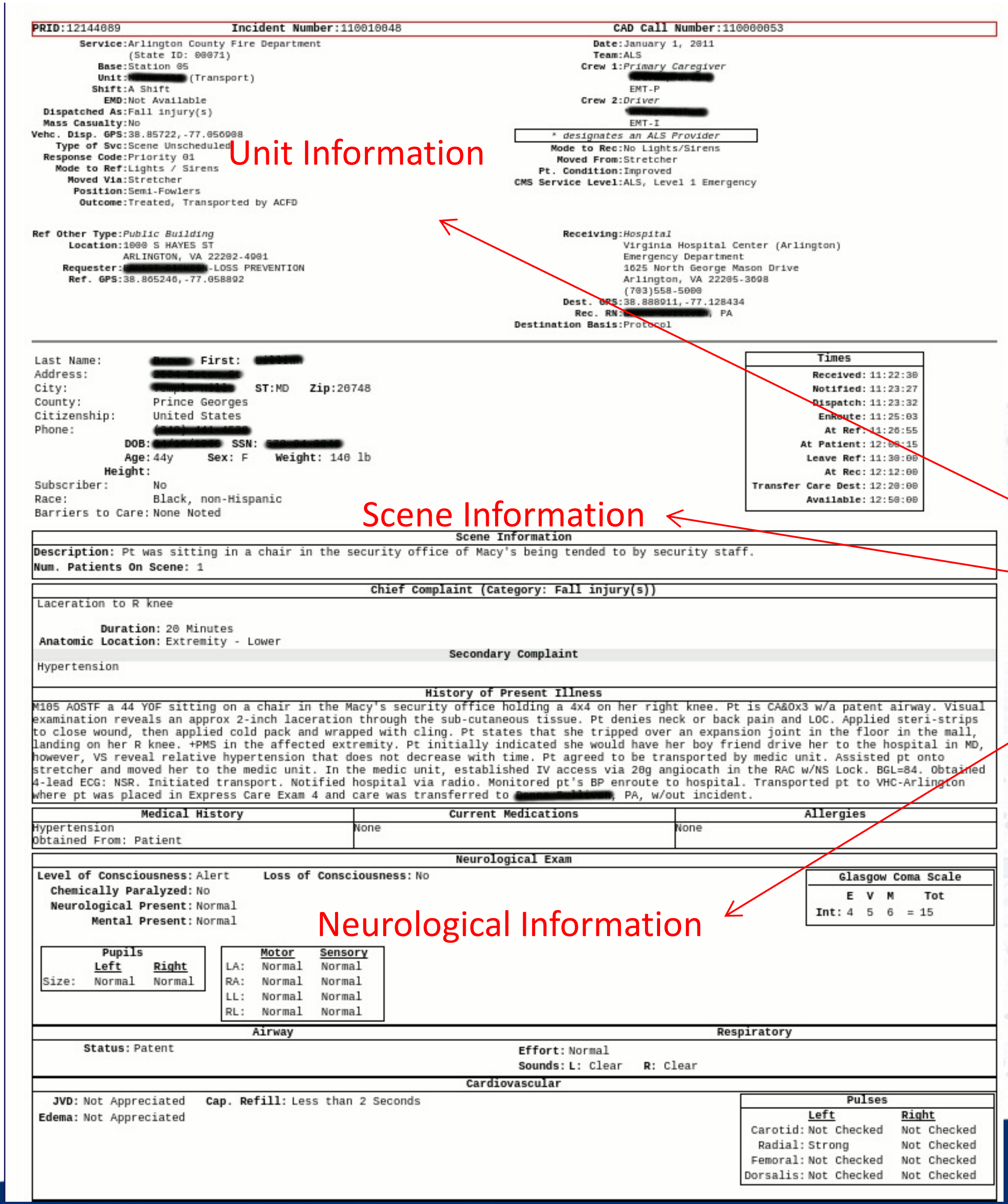
Structure, Quality, Metadata & Provenance

Combined Observation Unit Types



Ideal Restructuring of MLS Data





PRID:12144089 Incident Number:110010048 CAD Call Number:110000053

Service:Arlington County Fire Department (State ID: 00071)
Base:Station 05
Unit:[redacted] (Transport)
Shift:A Shift
EMD:Not Available
Dispatched As:Fall injury(s)
Mass Casualty:No
Vehc. Disp. GPS:38.85722,-77.056908
Type of Svc:Scene Unscheduled
Response Code:Priority 01
Mode to Ref:Lights / Sirens
Moved Via:Stretcher
Position:Semi-Fowlers
Outcome:Treated, Transported by ACFD

Date:January 1, 2011
Team:ALS
Crew 1:*Primary Caregiver*
[redacted]
EMT-P
Crew 2:*Driver*
[redacted]
EMT-I
** designates an ALS Provider*
Mode to Rec:No Lights/Sirens
Moved From:Stretcher
Pt. Condition:Improved
CMS Service Level:ALS, Level 1 Emergency

Ref Other Type:*Public Building*
Location:1000 S HAYES ST
ARLINGTON, VA 22202-4901
Requester:[redacted]-LOSS PREVENTION
Ref. GPS:38.865246,-77.058892

Receiving:*Hospital*
Virginia Hospital Center (Arlington)
Emergency Department
1625 North George Mason Drive
Arlington, VA 22205-3698
(703)558-5000
Dest. GPS:38.888911,-77.128434
Rec. RN:[redacted], PA
Destination Basis:Protocol

Last Name: [redacted] First: [redacted]
Address: [redacted]
City: [redacted] ST:MD Zip:20748
County: Prince Georges
Citizenship: United States
Phone: [redacted]
DOB:[redacted] SSN: [redacted]
Age:44y Sex: F Weight: 140 lb
Height:
Subscriber: No
Race: Black, non-Hispanic
Barriers to Care: None Noted

Times	
Received:	11:22:30
Notified:	11:23:27
Dispatch:	11:23:32
EnRoute:	11:25:03
At Ref:	11:26:55
At Patient:	12:00:15
Leave Ref:	11:30:00
At Rec:	12:12:00
Transfer Care Dest:	12:20:00
Available:	12:50:00

Scene Information
Description: Pt was sitting in a chair in the security office of Macy's being tended to by security staff.
Num. Patients On Scene: 1

Chief Complaint (Category: Fall injury(s))
Laceration to R knee
Duration: 20 Minutes
Anatomic Location: Extremity - Lower

Secondary Complaint
Hypertension

History of Present Illness
M105 AOSTF a 44 YOF sitting on a chair in the Macy's security office holding a 4x4 on her right knee. Pt is CA&Ox3 w/a patent airway. Visual examination reveals an approx 2-inch laceration through the sub-cutaneous tissue. Pt denies neck or back pain and LOC. Applied steri-strips to close wound, then applied cold pack and wrapped with cling. Pt states that she tripped over an expansion joint in the floor in the mall, landing on her R knee. +PMS in the affected extremity. Pt initially indicated she would have her boy friend drive her to the hospital in MD, however, VS reveal relative hypertension that does not decrease with time. Pt agreed to be transported by medic unit. Assisted pt onto stretcher and moved her to the medic unit. In the medic unit, established IV access via 20g angiocath in the RAC w/NS Lock. BGL=84. Obtained 4-lead ECG: NSR. Initiated transport. Notified hospital via radio. Monitored pt's BP enroute to hospital. Transported pt to VHC-Arlington where pt was placed in Express Care Exam 4 and care was transferred to [redacted], PA, w/out incident.

Medical History	Current Medications	Allergies
Hypertension Obtained From: Patient	None	None

Neurological Exam
Level of Consciousness: Alert Loss of Consciousness: No
Chemically Paralyzed: No
Neurological Present: Normal
Mental Present: Normal

Glasgow Coma Scale	E	V	M	Tot
Int:	4	5	6	= 15

Pupils	Left	Right
Size:	Normal	Normal

	Motor	Sensory
LA:	Normal	Normal
RA:	Normal	Normal
LL:	Normal	Normal
RL:	Normal	Normal

Airway
Status: Patent

Respiratory
Effort: Normal
Sounds: L: Clear R: Clear

Cardiovascular
JVD: Not Appreciated Cap. Refill: Less than 2 Seconds
Edema: Not Appreciated

Pulses	Left	Right
Carotid:	Not Checked	Not Checked
Radial:	Strong	Not Checked
Femoral:	Not Checked	Not Checked
Dorsalis:	Not Checked	Not Checked

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Combined Observation Unit Types

Emergency Services Data Processing Example

Multiple Types of Observation on Single Page

Additionally:
Forms made available online as nested HTML Tables - each needing separate extraction

Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**

Structure, Quality, Metadata & Provenance

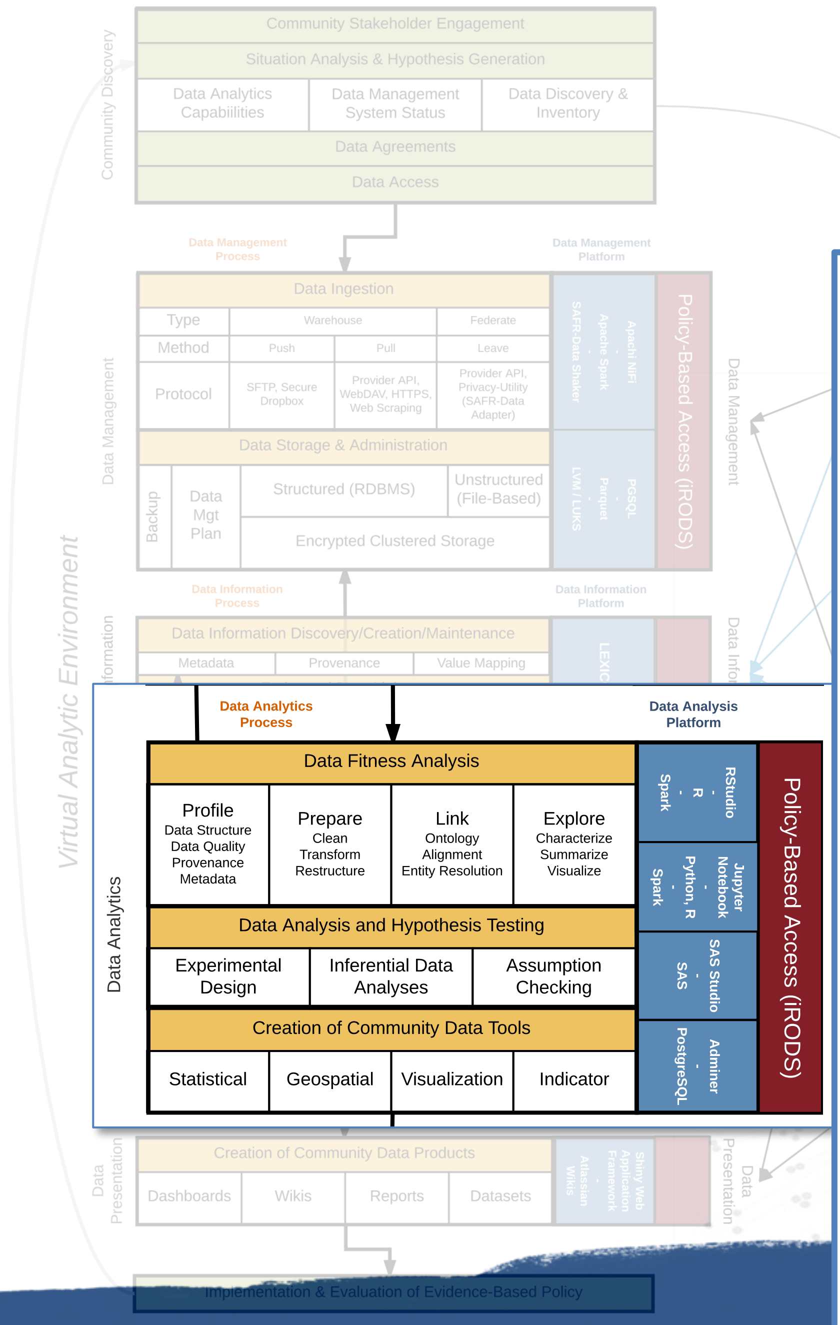
Divided Observation Unit Types

NC Student Data

Demographics Recorded in Multiple Tables

- Actual 2011 data from different tables linked via unique ID
- Many more tables with apparently separately collected demographics
- Derivation of Demographic Truth is now Probabilistic

gender1	id	gender2
F	43XXX13	M
F	43XXX14	M
M	76XXX46	F
F	74XXX98	M
F	76XXX23	M
F	77XXX40	M
M	74XXX98	F
M	78XXX73	F
F	78XXX74	M
M	77XXX84	F
F	79XXX87	M
M	71XXX95	F
M	21XXX96	F
M	71XXX54	F
F	71XXX55	M
F	77XXX86	M
F	80XXX24	M
M	76XXX79	F



Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**

Structure, **Quality**, Metadata & Provenance

Completeness

percentage of elements properly populated

e.g. Testing for NULLs and empty strings where not appropriate

Value Validity

percentage of elements whose attributes possess meaningful values

e.g. A comparison constraint like {male; female} or an interval constraint like age = [0,110]

Consistency

a measure of the degree to which two or more data attributes satisfy a well-defined dependency constraint – relationship validation

e.g. Zip-code – state consistency or gender – pregnancy consistency

Uniqueness

the number of unique values taken by an attribute, or a combination of attributes in a dataset

e.g. Frequency distribution of an element

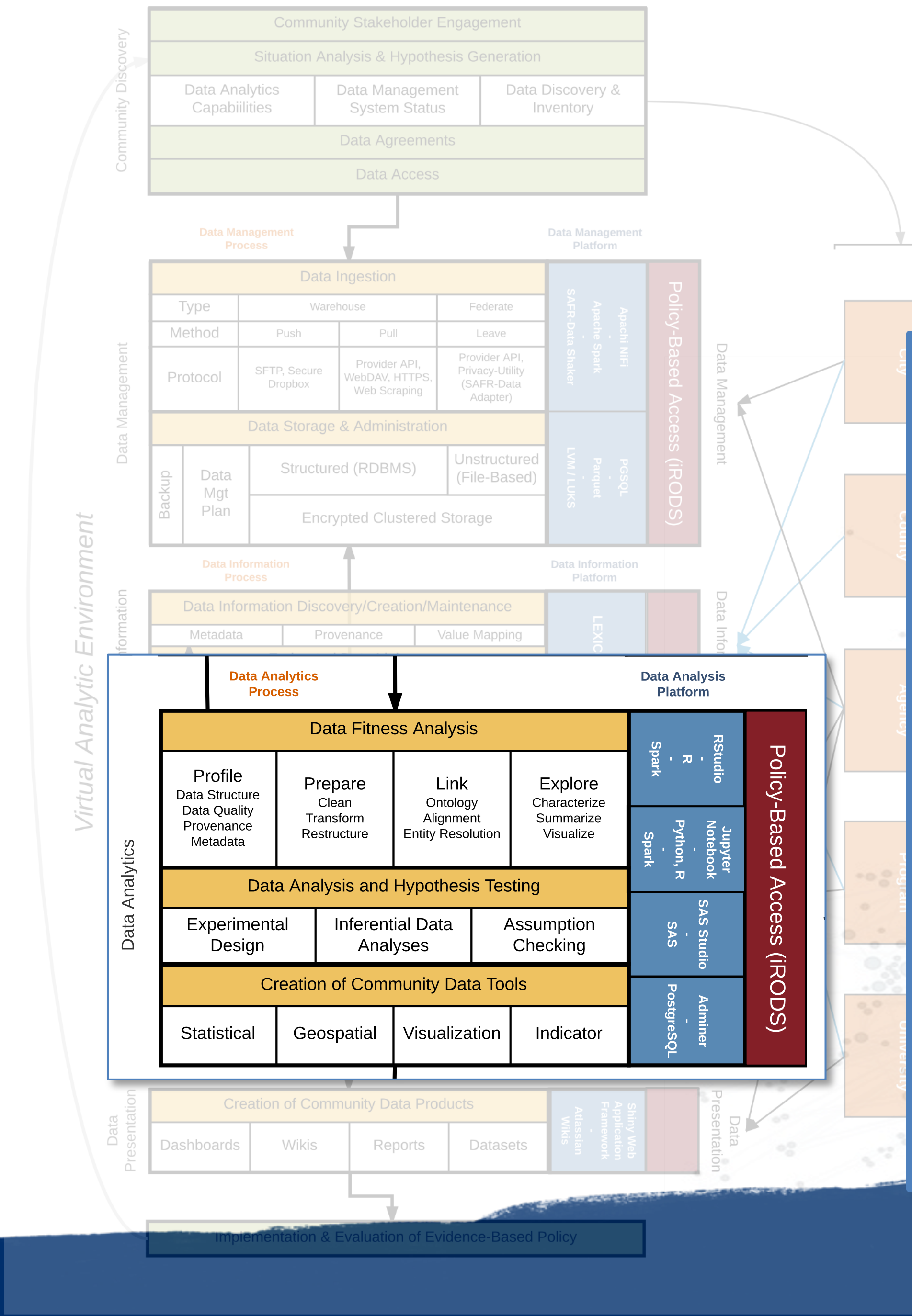
note. The more homogeneous the data values of an element, the less useful the element is for analysis

Duplication

a measure of the degree of replication of distinct observations per observation unit type

e.g. Greater than 1 registration per student per official reporting period

note. Duplication occurs as a result of choice of level of aggregation



Repurposing Data for Statistical Purposes

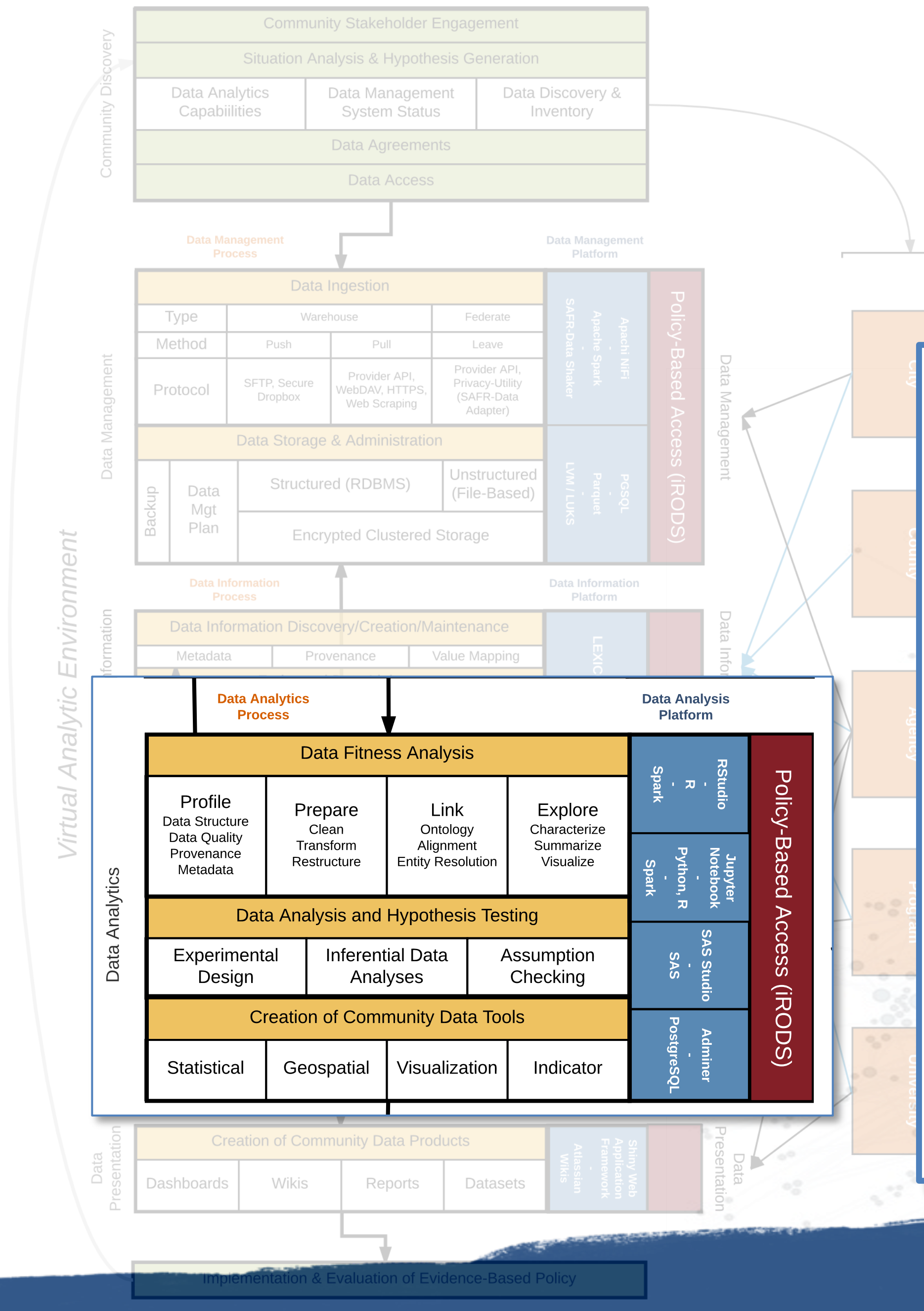
Data Fitness Analysis: **Profiling**

Structure, **Quality**, Metadata & Provenance

Completeness

- Seems straight-forward -- Nope
- A set of data is complete with respect to a *given purpose* if the set contains all the relevant data for that purpose
- A common measure is the proportion of data that has values to the proportion that "should" have values.
  - Completeness is *application-specific*
  - Incorrect to simply measure number of missing field values in a record without considering which fields are necessary
    - MLS Data had MANY highly incomplete fields that were not necessary for the study at hand
- Data that are missing can be categorized as:
  - record fields not containing data
  - records not containing necessary fields
  - datasets not containing the requisite records

UNIVERSITY of VIRGINIA



Repurposing Data for Statistical Purposes

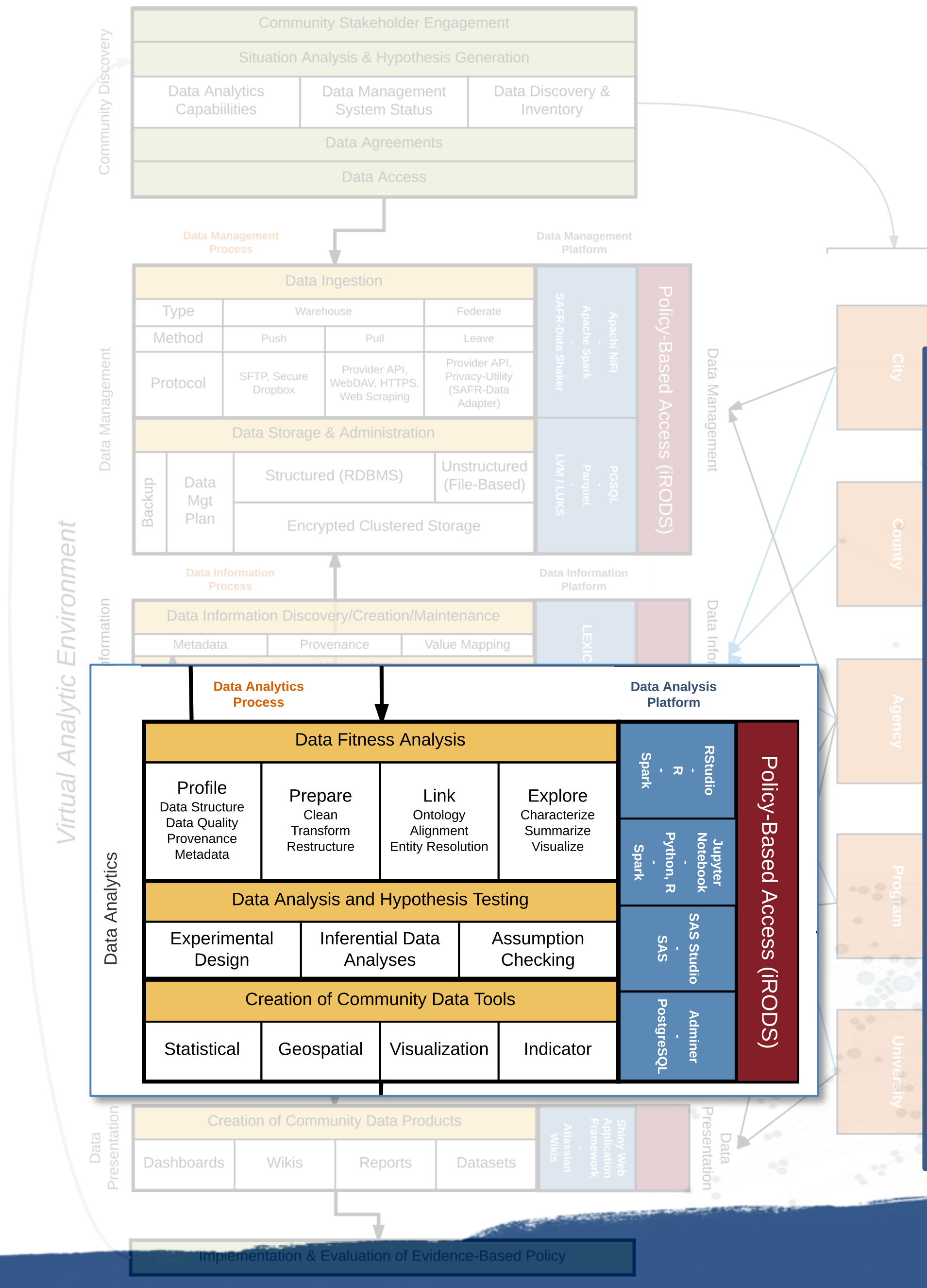
Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Value Validity

- Data elements with proper values have ***value validity***
- The percentage of data elements whose attributes possess values within the range expected for a legitimate entry is a measure of value validity
- Checking for value validity generally comes in the form of straight-forward domain constraint rules
  - How many entries contain non-valid values for a non-empty text field representing gender?
    - < *count gender where gender is not* (*male, female*) >
  - How many entries contain non-valid values for a non-empty integer field representing age?
    - < *count age where age is not between* [0, 110] >

University of Virginia



Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**

Structure, **Quality**, Metadata & Provenance

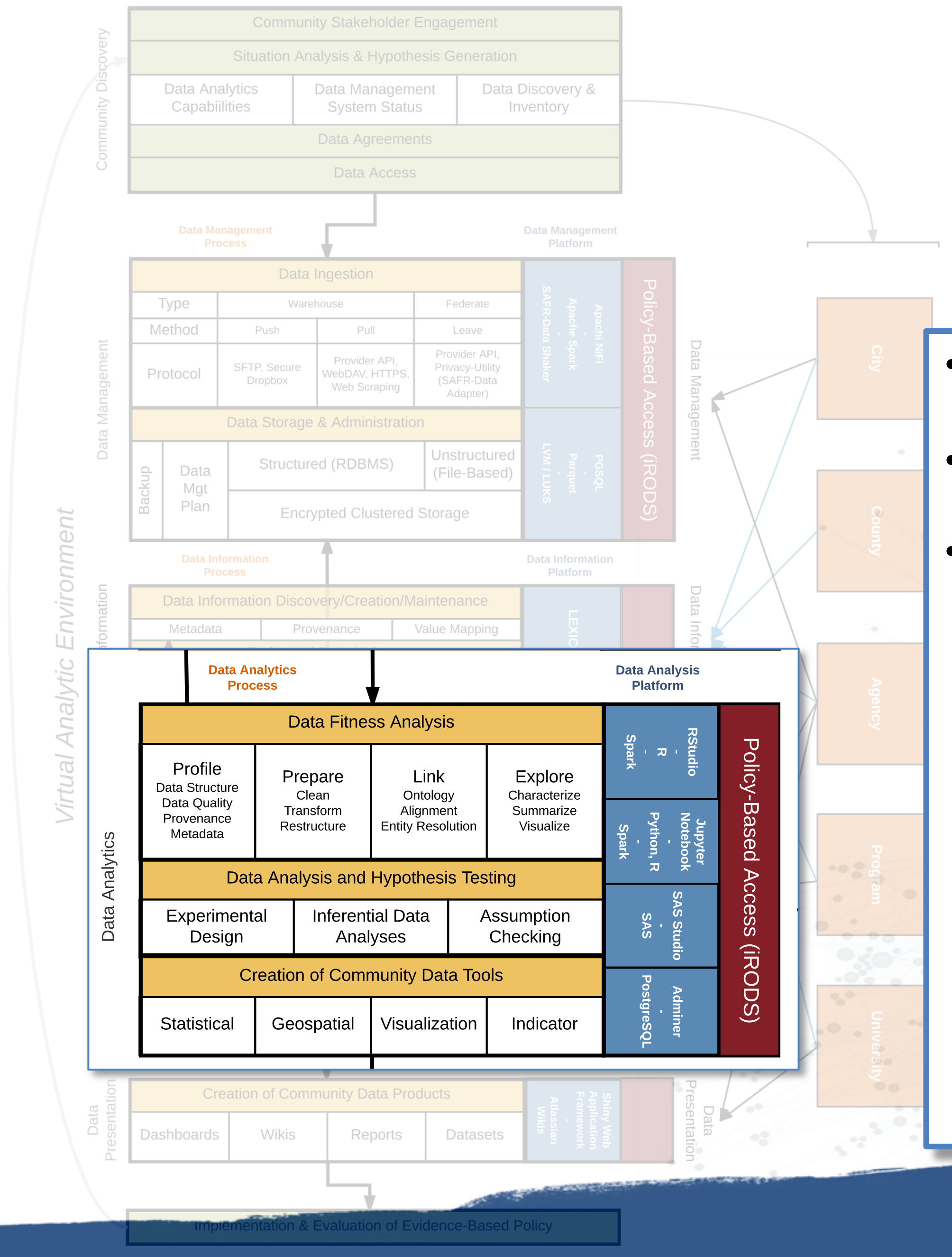
Value Validity

Pulled from current James City County MLS Data

zip_code	area	subdivision	neighborhood	zoning	parcel_id
23185	JCC	Governors Land	River Reach	R-4	4511000022
23188	JCC	Wellington		RESIDENT	1330800178
23188	JCC	Powhatan Secondary		RES	3741600013
23185	JCC	Kingsmill	Padgetts Ordinary	R 4	5041100213
23185	JCC	Pointe @ Jamestown		RES	4640600108
23185	JCC	Paddock Green	Paddock Green	R1	

Comparison constraint: **zoning 2015, James City County** = {A-1, R-1, R-2, R-3, R-4, R-5, R-6, R-7, R-8, LB, B-1, M-1, M-2, RT, PUD, MU, PL, EO}

- During Data Profiling issues are described, not “fixed”
- The appropriate fix depends upon the needs of the research
- It may be appropriate to simply normalize all zoning entries to the five major categories of zoning: Residential, Mixed Residential-Commercial, Commercial, Industrial, and Special



Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Consistency

- The Degree to Which Two or More Attributes Satisfy a Dependency Constraint
- Simple example
  - Location disagreements like zip and state (**Record-Level**)
- More complex example (**Longitudinal**)
  - Consistency with locally derived “truth”
  - VDOE Student Record, no definitive list of student demographics
  - Truth must be derived from multiple observations
    - Student Record has multiple observations per school year
    - Query here shows disagreement on gender for some of the observations when Student Record is matched to itself
      - **select count(distinct a.internal_id)**
        from vdoe.student_record a
        join vdoe.student_record b
        on a.internal_id = b.internal_id
        and a.gender <> b.gender
      - **16,310 / 2,346,058 individuals have more than one value for gender**

University of Virginia

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Record Consistency

Emergency Services Data Processing Example

PRID:12144089 **Incident Number:**110010048 **CAD Call Number:**110000053

Service: Arlington County Fire Department
(State ID: 00071)
Base: Station 05
Unit:[redacted] (Transport)
Shift:A Shift
EMD: Not Available
Dispatched As: Fall injury(s)
Mass Casualty: No
Vehc. Disp. GPS:38.85722,-77.056008
Type of Svc: Scene Unscheduled
Response Code: Priority 01
Mode to Ref: Lights / Sirens
Moved Via: Stretcher
Position: Semi-Fowlers
Outcome: Treated, Transported by ACFD

Date: January 1, 2011
Team:ALS
Crew 1:*Primary Caregiver*
[redacted]
EMT-P
Crew 2:*Driver*
[redacted]
EMT-I
** designates an ALS Provider*
Mode to Rec: No Lights/Sirens
Moved From: Stretcher
Pt. Condition: Improved
CMS Service Level:ALS, Level 1 Emergency

Ref Other Type:*Public Building*
Location:1000 S HAYES ST
ARLINGTON, VA 22202-4901
Requester:[redacted]-LOSS PREVENTION
Ref. GPS:38.865246,-77.058892

Receiving:*Hospital*
Virginia Hospital Center (Arlington)
Emergency Department
1625 North George Mason Drive
Arlington, VA 22205-3698
(703)558-5000
Dest. GPS:38.888911,-77.128434
Rec. RN:[redacted], PA
Destination Basis: Protocol

Last Name: [redacted] **First:** [redacted]
Address: [redacted]
City: [redacted] **ST:**MD **Zip:**20748
County: Prince Georges
Citizenship: United States
Phone: [redacted]
DOB:[redacted] **SSN:** [redacted]
Age:44y **Sex:** F **Weight:** 140 lb
Height:
Subscriber: No
Race: Black, non-Hispanic
Barriers to Care: None Noted

Times	
Received:	11:22:30
Notified:	11:23:27
Dispatch:	11:23:32
EnRoute:	11:25:03
At Ref:	11:26:55
At Patient:	12:00:15
Leave Ref:	11:30:00
At Rec:	12:12:00
Transfer Care Dest:	12:20:00
Available:	12:50:00

Consistency Issue:
Violates time dependency constraint
Leaves scene before arriving

Scene Information

Description: Pt was sitting in a chair in the security office of Macy's being tended to by security staff.
Num. Patients On Scene: 1

Chief Complaint (Category: Fall injury(s))

Laceration to R knee

Duration: 20 Minutes
Anatomic Location: Extremity - Lower

Secondary Complaint

Hypertension

History of Present Illness

M105 AOSTF a 44 YOF sitting on a chair in the Macy's security office holding a 4x4 on her right knee. Pt is CA&Ox3 w/a patent airway. Visual examination reveals an approx 2-inch laceration through the sub-cutaneous tissue. Pt denies neck or back pain and LOC. Applied steri-strips to close wound, then applied cold pack and wrapped with cling. Pt states that she tripped over an expansion joint in the floor in the mall, landing on her R knee. +PMS in the affected extremity. Pt initially indicated she would have her boy friend drive her to the hospital in MD, however, VS reveal relative hypertension that does not decrease with time. Pt agreed to be transported by medic unit. Assisted pt onto stretcher and moved her to the medic unit. In the medic unit, established IV access via 20g angiocath in the RAC w/NS Lock. BGL=84. Obtained 4-lead ECG: NSR. Initiated transport. Notified hospital via radio. Monitored pt's BP enroute to hospital. Transported pt to VHC-Arlington where pt was placed in Express Care Exam 4 and care was transferred to [redacted], PA, w/out incident.

Medical History	Current Medications	Allergies
Hypertension Obtained From: Patient	None	None

Neurological Exam

Level of Consciousness: Alert **Loss of Consciousness:** No
Chemically Paralyzed: No
Neurological Present: Normal
Mental Present: Normal

Glasgow Coma Scale	E	V	M	Tot
Int:	4	5	6	= 15

Pupils	Left	Right
Size:	Normal	Normal

	Motor	Sensory
LA:	Normal	Normal
RA:	Normal	Normal
LL:	Normal	Normal
RL:	Normal	Normal

Airway

Status: Patent

Respiratory

Effort: Normal
Sounds: **L:** Clear **R:** Clear

Cardiovascular

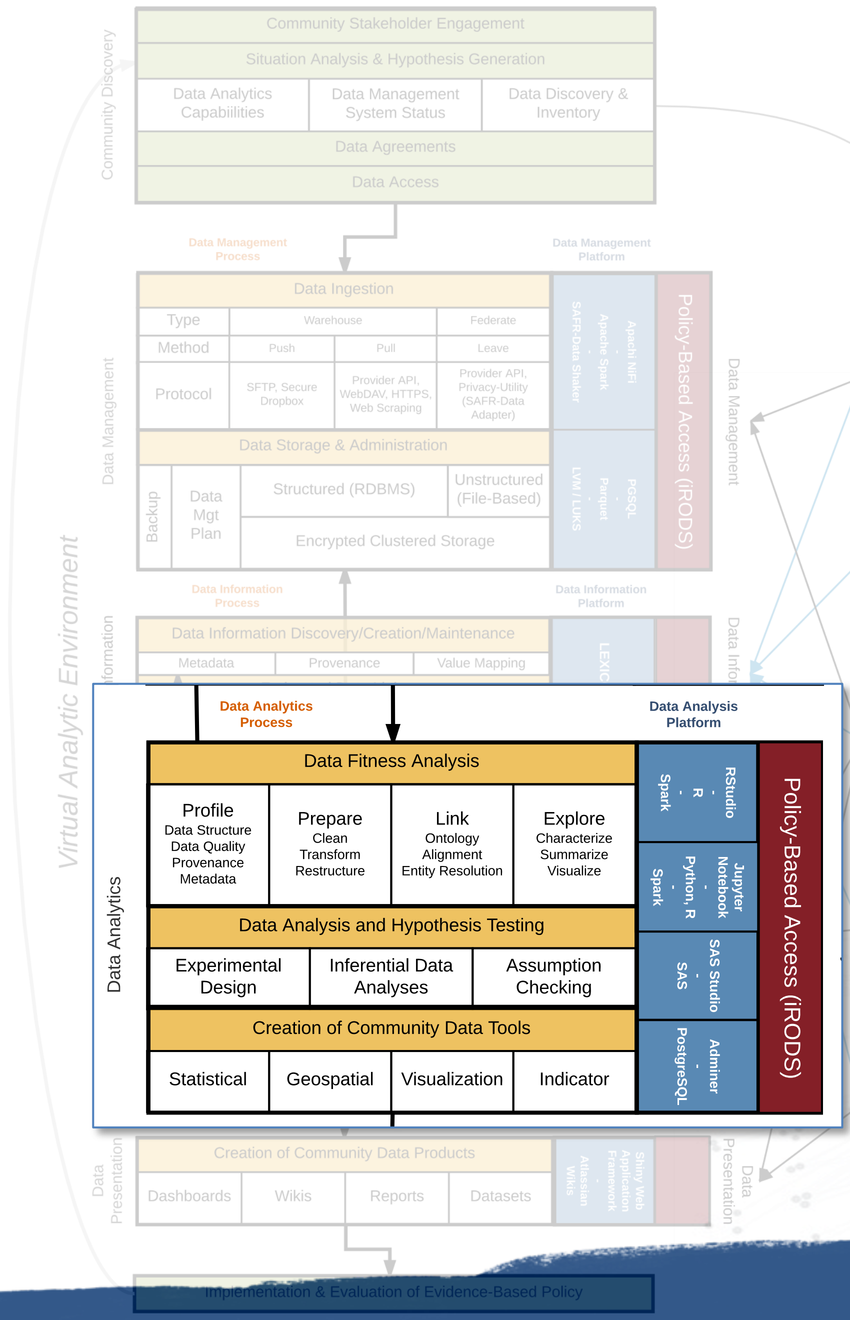
JVD: Not Appreciated **Cap. Refill:** Less than 2 Seconds
Edema: Not Appreciated

Pulses	Left	Right
Carotid:	Not Checked	Not Checked
Radial:	Strong	Not Checked
Femoral:	Not Checked	Not Checked
Dorsalis:	Not Checked	Not Checked

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance



Observation Unit Definition
Datasets (tables) without definition and/or non-meaningful/confusing naming

Observation Unit Attributes Definition
Attributes (columns) without definition and/or non-meaningful/confusing naming

Semantic Confusion
Attributes with the same name but different definitions
e.g. An attribute named “Grade” can refer to both a 'score' for a test or the 'level/year’

Multiple Attribute Names
Attributes with different names but the same definition
e.g. Attributes name “Grade” and “Year” both referring to 'level/year' of schooling

Inconsistent Attribute Formats
Attributes of the same type that are formatted differently
e.g. Most commonly an issue when dealing with dates and times

Data Process History
Attributes collected at different locations, with different tools

System of Origin
Where was this data originally collected?

Intermediate Storage Systems
Chain of Custody

Contact Information
Who can I contact with my questions?

Transformation
What happened to the data since collection and why?

Getting this stuff in order is a BIG part of Data Repurposing!

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Observation Unit Attributes Definition

Emergency Services Data Processing Example

PRID:12144089 **Incident Number:**110010048 **CAD Call Number:**110000053

Service: Arlington County Fire Department (State ID: 00071)
Base: Station 05
Unit:[redacted] (Transport)
Shift:A Shift
EMD: Not Available
Dispatched As: Fall injury(s)
Mass Casualty: No
Vehc. Disp. GPS:38.85722,-77.056908
Type of Svc: Scene Unscheduled
Response Code: Priority 01
Mode to Ref: Lights / Sirens
Moved Via: Stretcher
Position: Semi-Fowlers
Outcome: Treated, Transported by ACFD

Date: January 1, 2011
Team:ALS
Crew 1:*Primary Caregiver*
[redacted]
EMT-P
Crew 2:*Driver*
[redacted]
EMT-I
** designates an ALS Provider*
Mode to Rec: No Lights/Sirens
Moved From: Stretcher
Pt. Condition: Improved
CMS Service Level:ALS, Level 1 Emergency

Ref Other Type:*Public Building*
Location:1000 S HAYES ST
ARLINGTON, VA 22202-4901
Requester:[redacted]-LOSS PREVENTION
Ref. GPS:38.865246,-77.058892

Receiving:*Hospital*
Virginia Hospital Center (Arlington)
Emergency Department
1625 North George Mason Drive
Arlington, VA 22205-3698
(703)558-5000
Dest. GPS:38.888911,-77.128434
Rec. RN:[redacted], PA
Destination Basis: Protocol

Last Name: [redacted] **First:** [redacted]
Address: [redacted]
City: [redacted] **ST:**MD **Zip:**20748
County: Prince Georges
Citizenship: United States
Phone: [redacted]
DOB:[redacted] **SSN:** [redacted]
Age: 44y **Sex:** F **Weight:** 140 lb
Height:
Subscriber: No
Race: Black, non-Hispanic
Barriers to Care: None Noted

Times	
Received:	11:22:30
Notified:	11:23:27
Dispatch:	11:23:32
EnRoute:	11:25:03
At Ref:	11:2[illegible]:55
At Patient:	12:00:15
Leave Ref:	11:30:00
At Rec:	12:12:00
Transfer Care Dest:	12:20:00
Available:	12:50:00

Scene Information
Description: Pt was sitting in a chair in the security office of Macy's being tended to by security staff.
Num. Patients On Scene: 1

Chief Complaint (Category: Fall injury(s))
Laceration to R knee
Duration: 20 Minutes
Anatomic Location: Extremity - Lower

Secondary Complaint
Hypertension

History of Present Illness
M105 AOSTF a 44 YOF sitting on a chair in the Macy's security office holding a 4x4 on her right knee. Pt is CA&Ox3 w/a patent airway. Visual examination reveals an approx 2-inch laceration through the sub-cutaneous tissue. Pt denies neck or back pain and LOC. Applied steri-strips to close wound, then applied cold pack and wrapped with cling. Pt states that she tripped over an expansion joint in the floor in the mall, landing on her R knee. +PMS in the affected extremity. Pt initially indicated she would have her boy friend drive her to the hospital in MD, however, VS reveal relative hypertension that does not decrease with time. Pt agreed to be transported by medic unit. Assisted pt onto stretcher and moved her to the medic unit. In the medic unit, established IV access via 20g angiocath in the RAC w/NS Lock. BGL=84. Obtained 4-lead ECG: NSR. Initiated transport. Notified hospital via radio. Monitored pt's BP enroute to hospital. Transported pt to VHC-Arlington where pt was placed in Express Care Exam 4 and care was transferred to [redacted], PA, w/out incident.

Medical History	Current Medications	Allergies
Hypertension Obtained From: Patient	None	None

Neurological Exam
Level of Consciousness: Alert **Loss of Consciousness:** No
Chemically Paralyzed: No
Neurological Present: Normal
Mental Present: Normal

Glasgow Coma Scale	E	V	M	Tot
Int:	4	5	6	= 15

Pupils	Left	Right
Size:	Normal	Normal

	Motor	Sensory
LA:	Normal	Normal
RA:	Normal	Normal
LL:	Normal	Normal
RL:	Normal	Normal

Airway
Status: Patent

Respiratory
Effort: Normal
Sounds: **L:** Clear **R:** Clear

Cardiovascular
JVD: Not Appreciated **Cap. Refill:** Less than 2 Seconds
Edema: Not Appreciated

Pulses	Left	Right
Carotid:	Not Checked	Not Checked
Radial:	Strong	Not Checked
Femoral:	Not Checked	Not Checked
Dorsalis:	Not Checked	Not Checked

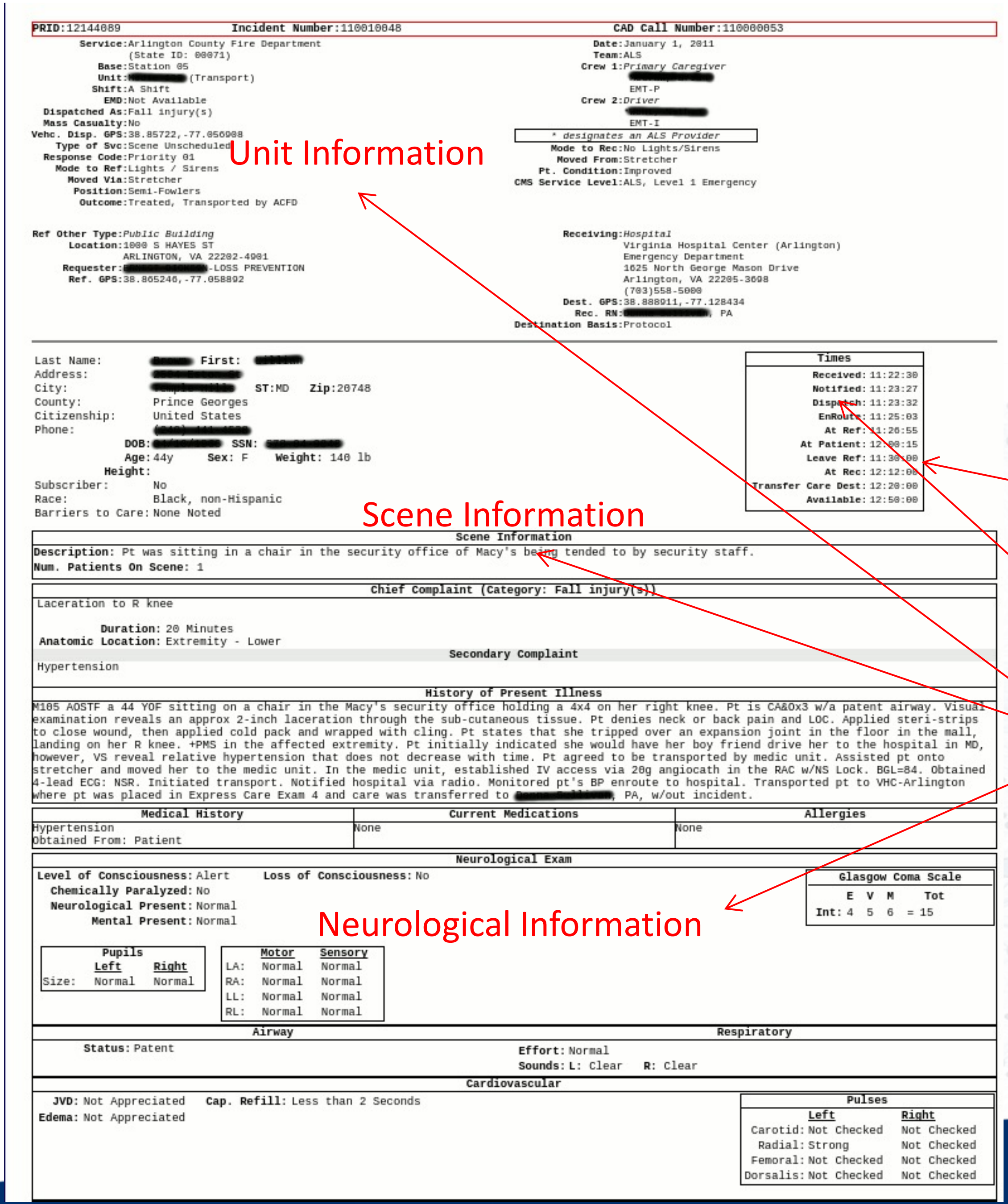
Metadata needed:
Time definitions unclear

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Emergency Services Data Processing Example



PRID:12144089 Incident Number:110010048 CAD Call Number:110000053

Service:Arlington County Fire Department (State ID: 00071)
Base:Station 05
Unit:[redacted] (Transport)
Shift:A Shift
EMD:Not Available
Dispatched As:Fall injury(s)
Mass Casualty:No
Vehc. Disp. GPS:38.85722,-77.056008
Type of Svc:Scene Unscheduled
Response Code:Priority 01
Mode to Ref:Lights / Sirens
Moved Via:Stretcher
Position:Semi-Fowlers
Outcome:Treated, Transported by ACFD

Date:January 1, 2011
Team:ALS
Crew 1:*Primary Caregiver*
[redacted]
EMT-P
Crew 2:*Driver*
[redacted]
EMT-I
** designates an ALS Provider*
Mode to Rec:No Lights/Sirens
Moved From:Stretcher
Pt. Condition:Improved
CMS Service Level:ALS, Level 1 Emergency

Unit Information

Ref Other Type:*Public Building*
Location:1000 S HAYES ST
ARLINGTON, VA 22202-4901
Requester:[redacted]-LOSS PREVENTION
Ref. GPS:38.865246,-77.058892

Receiving:*Hospital*
Virginia Hospital Center (Arlington)
Emergency Department
1625 North George Mason Drive
Arlington, VA 22205-3698
(703)558-5000
Dest. GPS:38.888911,-77.128434
Rec. RN:[redacted], PA
Destination Basis:Protocol

Last Name: [redacted] First: [redacted]
Address: [redacted]
City: [redacted] ST:MD Zip:20748
County: Prince Georges
Citizenship: United States
Phone: [redacted]
DOB:[redacted] SSN: [redacted]
Age:44y Sex: F Weight: 140 lb
Height:
Subscriber: No
Race: Black, non-Hispanic
Barriers to Care: None Noted

Times	
Received:	11:22:30
Notified:	11:23:27
Dispatch:	11:23:32
EnRoute:	11:25:03
At Ref:	11:26:55
At Patient:	12:00:15
Leave Ref:	11:30:00
At Rec:	12:12:00
Transfer Care Dest:	12:20:00
Available:	12:50:00

Scene Information

Scene Information

Description: Pt was sitting in a chair in the security office of Macy's being tended to by security staff.
Num. Patients On Scene: 1

Chief Complaint (Category: Fall injury(s))

Laceration to R knee

Duration: 20 Minutes
Anatomic Location: Extremity - Lower

Secondary Complaint

Hypertension

History of Present Illness

M105 AOSTF a 44 YOF sitting on a chair in the Macy's security office holding a 4x4 on her right knee. Pt is CA&Ox3 w/a patent airway. Visual examination reveals an approx 2-inch laceration through the sub-cutaneous tissue. Pt denies neck or back pain and LOC. Applied steri-strips to close wound, then applied cold pack and wrapped with cling. Pt states that she tripped over an expansion joint in the floor in the mall, landing on her R knee. +PMS in the affected extremity. Pt initially indicated she would have her boy friend drive her to the hospital in MD, however, VS reveal relative hypertension that does not decrease with time. Pt agreed to be transported by medic unit. Assisted pt onto stretcher and moved her to the medic unit. In the medic unit, established IV access via 20g angiocath in the RAC w/NS Lock. BGL=84. Obtained 4-lead ECG: NSR. Initiated transport. Notified hospital via radio. Monitored pt's BP enroute to hospital. Transported pt to VHC-Arlington where pt was placed in Express Care Exam 4 and care was transferred to [redacted], PA, w/out incident.

Medical History	Current Medications	Allergies
Hypertension Obtained From: Patient	None	None

Neurological Exam

Level of Consciousness: Alert Loss of Consciousness: No
Chemically Paralyzed: No
Neurological Present: Normal
Mental Present: Normal

Glasgow Coma Scale	E	V	M	Tot
Int:	4	5	6	= 15

Neurological Information

Pupils	Left	Right
Size:	Normal	Normal

	Motor	Sensory
LA:	Normal	Normal
RA:	Normal	Normal
LL:	Normal	Normal
RL:	Normal	Normal

Airway

Status: Patent

Respiratory

Effort: Normal
Sounds: L: Clear R: Clear

Cardiovascular

JVD: Not Appreciated Cap. Refill: Less than 2 Seconds
Edema: Not Appreciated

Pulses	Left	Right
Carotid:	Not Checked	Not Checked
Radial:	Strong	Not Checked
Femoral:	Not Checked	Not Checked
Dorsalis:	Not Checked	Not Checked

(Quality) Consistency Issue:
Violates time dependency constraint
Leaves scene before arriving

(Metadata) Metadata needed:
Time definitions unclear

(Structure) Multiple Types of Observation on Single Page

Additionally:
(Structure) Forms made available online as nested HTML Tables - each needing separate extraction

Example: Consistency

Quality Analysis can only be Semi-Automated

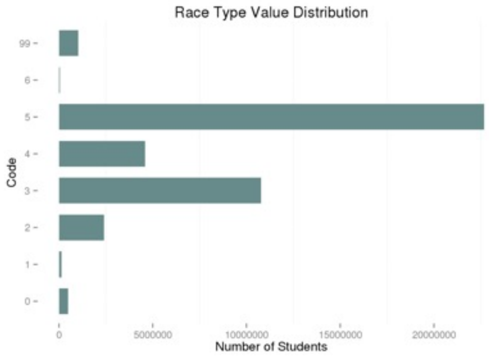
Race Type Report

SUMMARY OF DATA ELEMENT

Race Type is recorded for each student as the one or more races they identify with. This set was collected from students in the school system during the school years of 2005-2015. During our collection period, starting with the 2010-2011 school year, the race type definition expanded to allow students to choose multiple races and separately specify Hispanic ethnicity. In this process categories were eliminated. Eliminated categories included the 'Unspecified or Unknown' category and the 'Hispanic' category. A student's Hispanic ethnicity is now captured under an ethnic flag.

Number of Unique Values: 8

value	value_description
99	Multiple race types reported
1	American Indian or Alaskan Native
2	Asian
3	Black or African American
4	Hispanic (valid before 2010 school year)
5	White
6	Native Hawaiian or Other Pacific Islander
0	Unspecified or Unknown (valid before 2010 school year)



DATA ANALYTICS SUMMARY OF RESULTS

Test	Measurementvalues	Value
Completeness	Number of missing values	0
	Percent of complete values	100 %
Validity	Number of invalid values	0
	Percent of valid values	100 %
Uniqueness	Number of unique values	8
Record Consistency I	Number of inconsistent records	0
	Percent consistent records	100 %
Record Consistency II	Number of inconsistent records	2830444
	Percent consistent records	93.28 %
Longitudinal Inconsistency I	Individuals with inconsistent records	160055

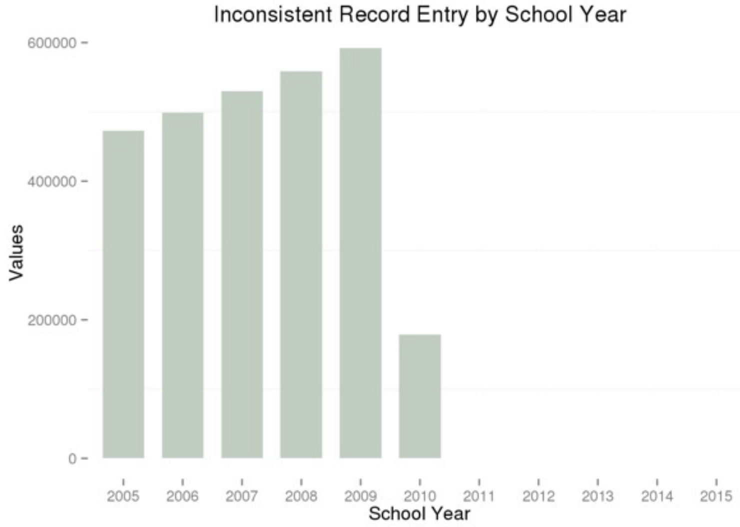
RECORD CONSISTENCY II

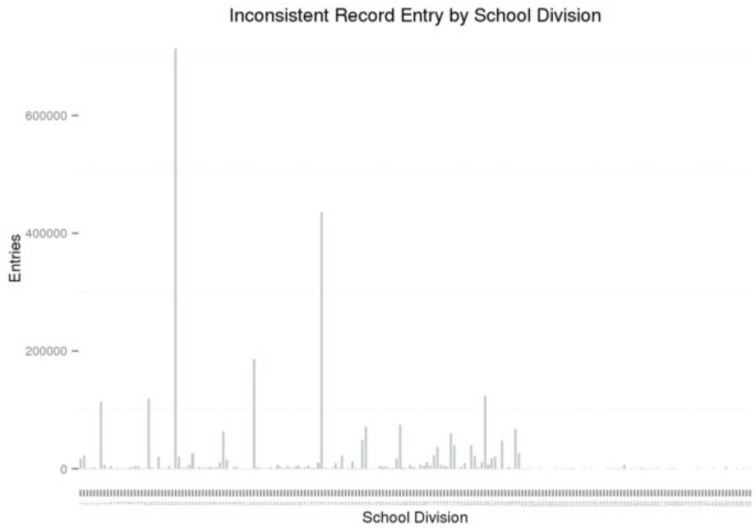
Find Records with an inconsistent relationship between Race Type and School Year .

Check if there any records that have a race_type of '4' after the 2009 school year.

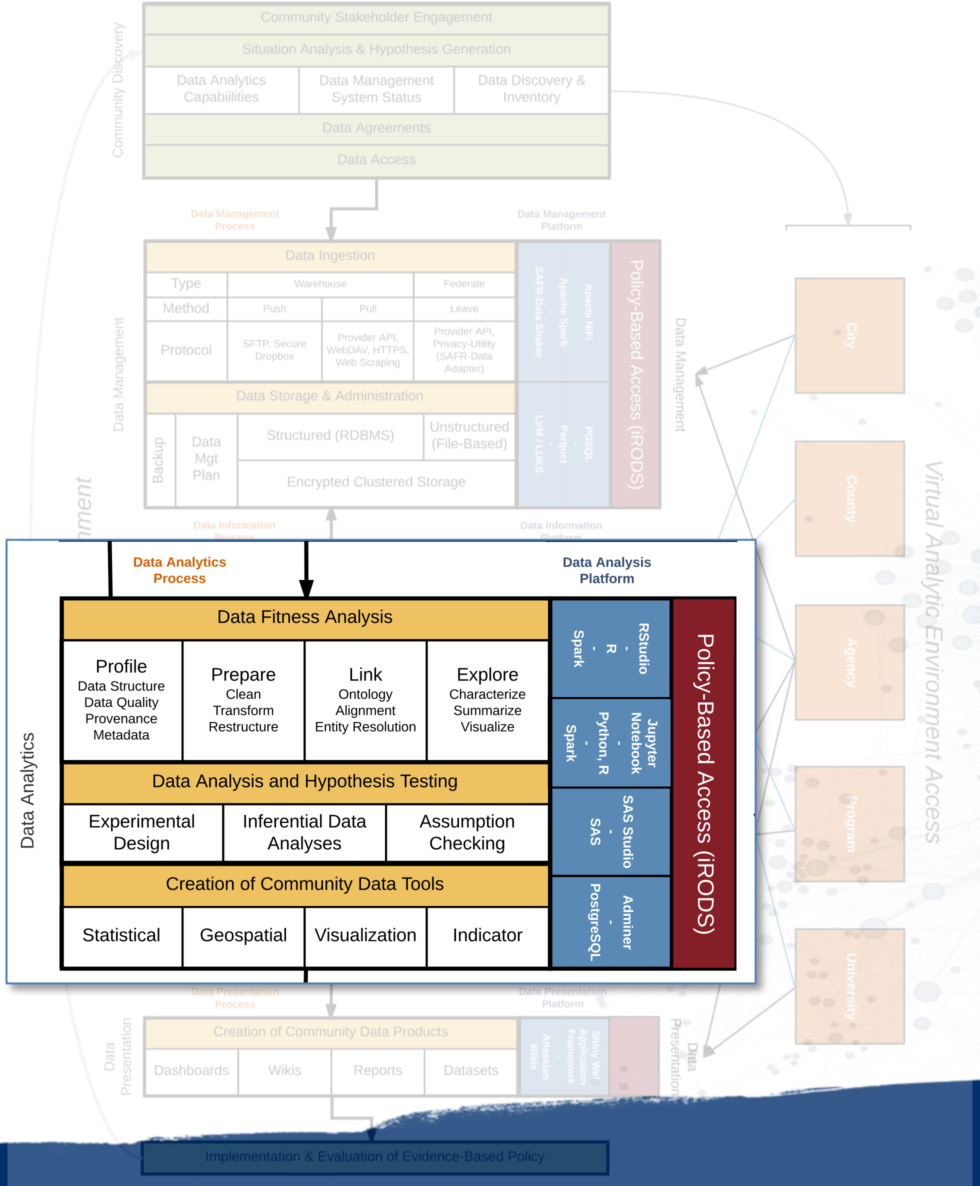
Number Inconsistent: 2830444

Percent Consistent: 93.28 %





Data Science Processes for Evidence-Based Policy



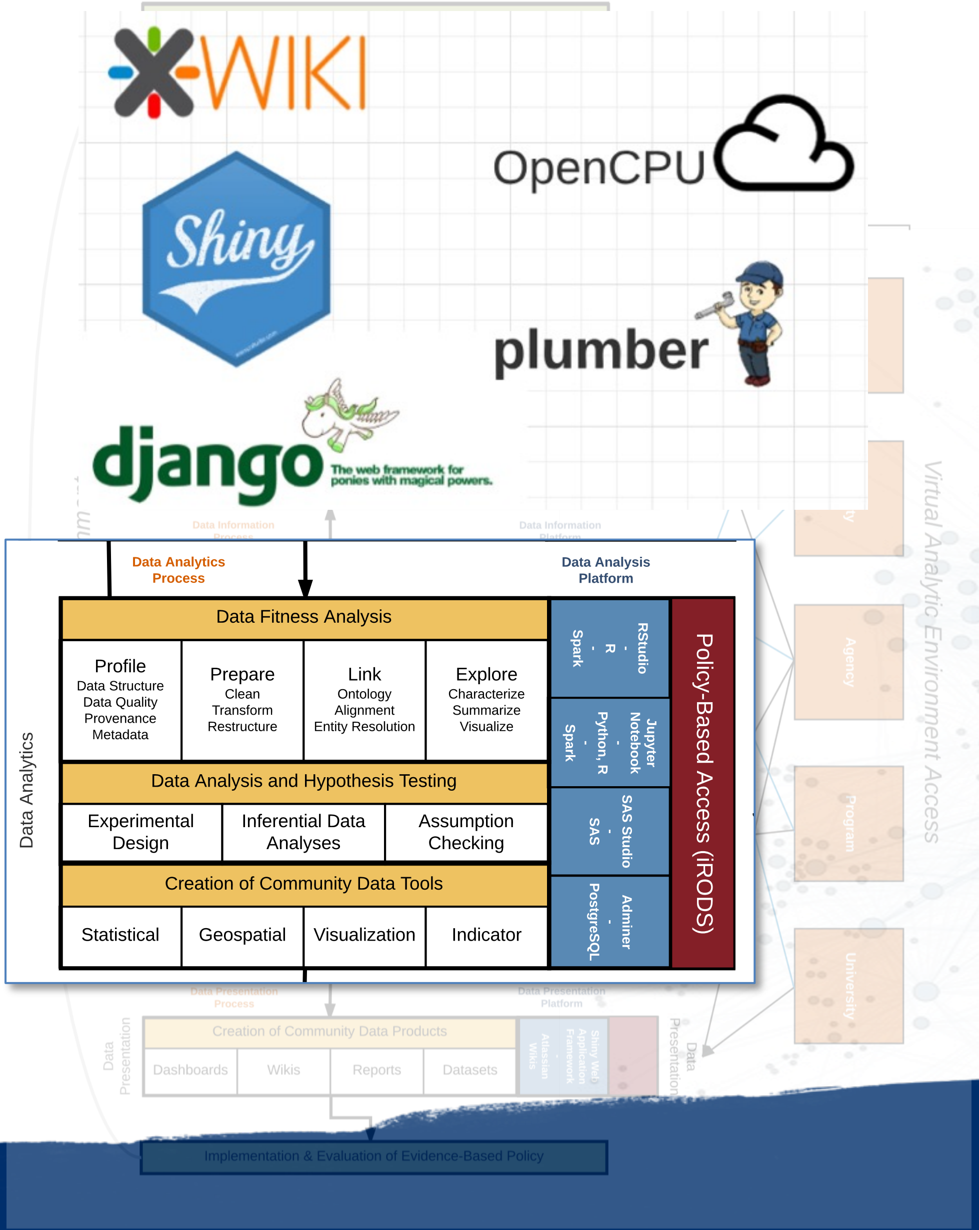
Data Analytics Process

- Data Fitness Analysis
  - **Data Preparation**

The Data Preparation Phase includes the activities necessary to “fix” the issues of Quality, Structure, and Metadata discovered during Data Profiling – activities can include:

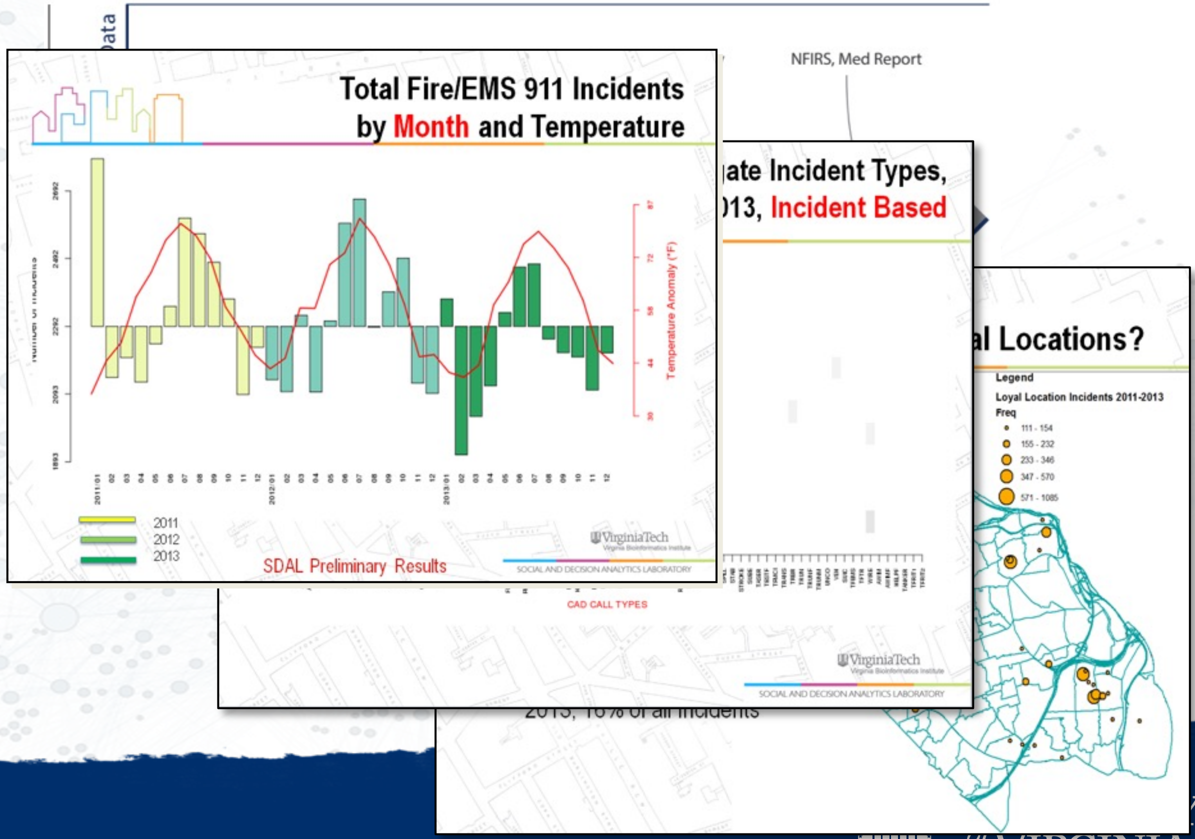
- Cleansing
  - Missing Values
  - Date Formats
  - Nominal => numeric
  - Outliers
  - Inconsistent Data
  - De-duplication
- Transformation
  - Aggregation
  - Normalization
  - Smoothing
  - Winsorization
  - Feature Construction
- Restructuring

Data Science Processes & Platforms for Evidence-Based Policy

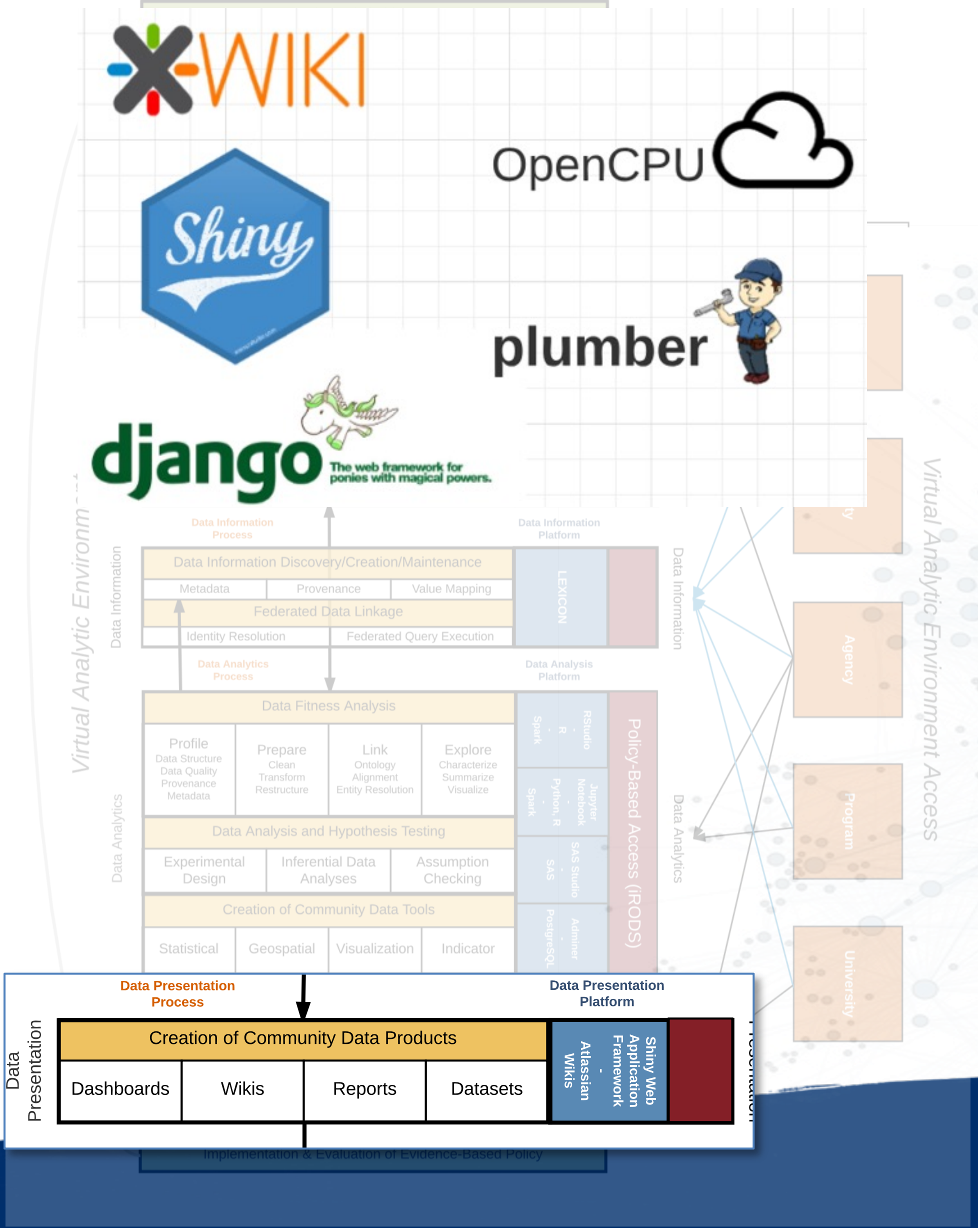


Data Analytics Process

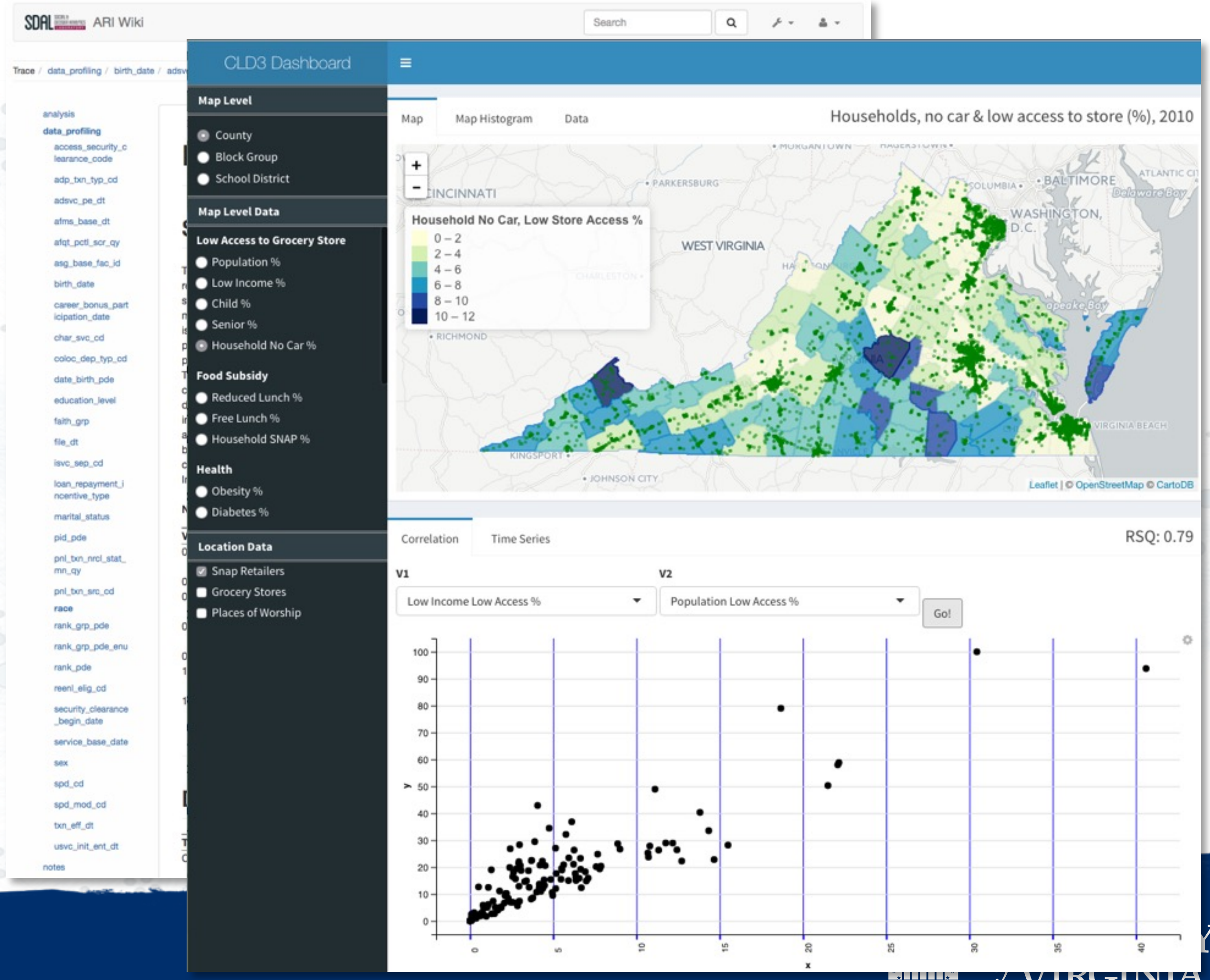
- Data Fitness Analysis
  - **Data Analysis & Hypothesis Testing**



Data Science Processes & Platforms for Evidence-Based Policy



Data Presentation Process



CLD3 - Data Science Processes & Platforms for Evidence-Based Policy