Creating a Data Science Centric Organization – Challenges and Opportunities

Canadian Data Science Workshop, April 30th - May 1st 2018

Sallie Keller
Professor of Statistics and Director

BIOCOMPLEXITY INSTITUTE
VIRGINIA TECH

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Biocomplexity Institute of Virginia Tech

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

“From molecules to policy”

Problem-Driven Science

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

Our Evolution

BIOCOMPLEXITY INSTITUTE
VIRGINIA TECH
2015

"Resetting Bioinformatics"

SDAL SOCIAL & DECISION ANALYTICS LABORATORY
2013

"From Molecules to Policy"

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH
VBI
2000

Social and Decision Analytics Lab

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

S. Keller, Koonin, S. E., & Shipp, S. (2012). Big data and city living-what can it do for us?. *Significance*, *9*(4), 4-7.

Social and Decision Analytics Lab

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

The Science of *ALL* data

Why Now?

ALL data revolution – new lens for social observing

Infrastructure

- Condition
- Operations
- Resilience
- Sustainability

Environment

- Climate
- Pollution
- Noise
- Flora/ Fauna

People

- Relationships
- Location
- Economic Condition
- Communication
- Health

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

Gaining insights through *ALL* data sources

Local, State/Provence, and Federal

Designed Data

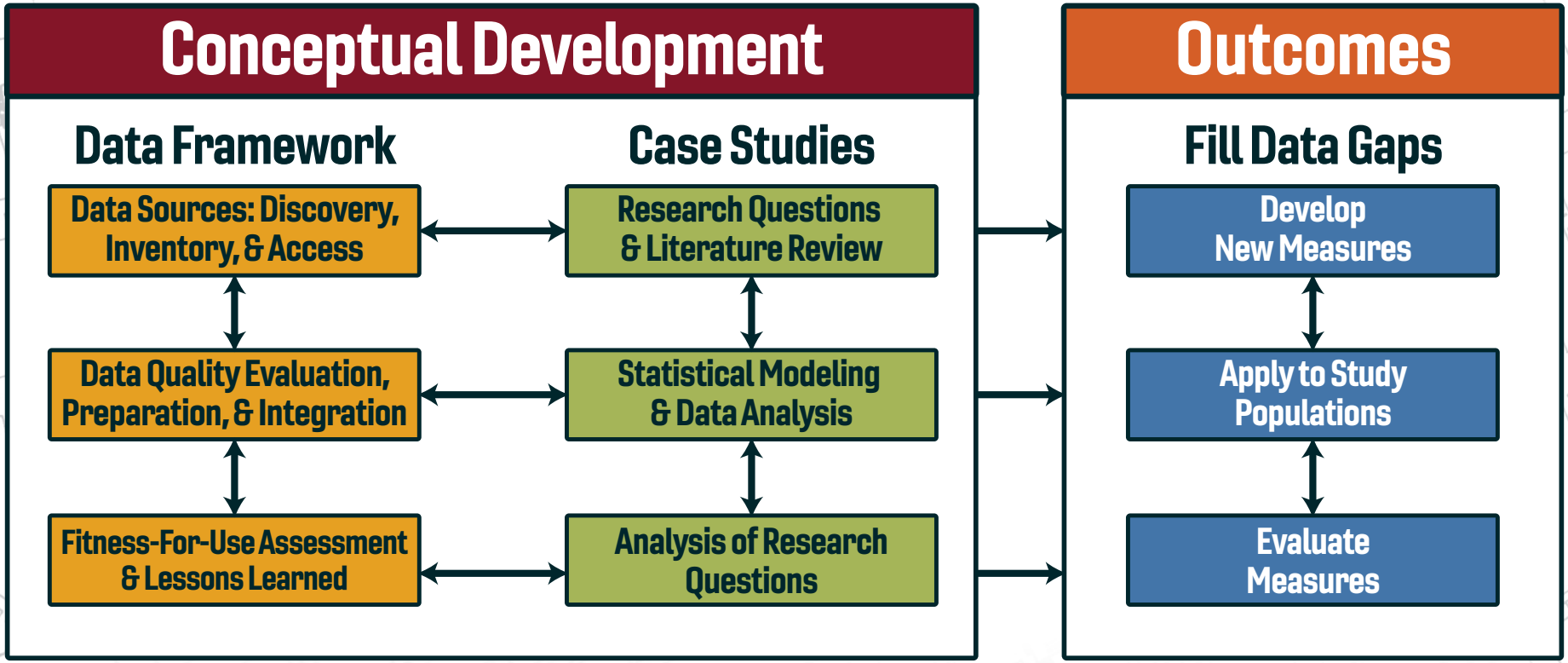
Administrative Data

Opportunity Data

Procedural Data

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

Our *Science of All Data* research model



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

Local / State Government

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

Federal Statistical Agencies

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

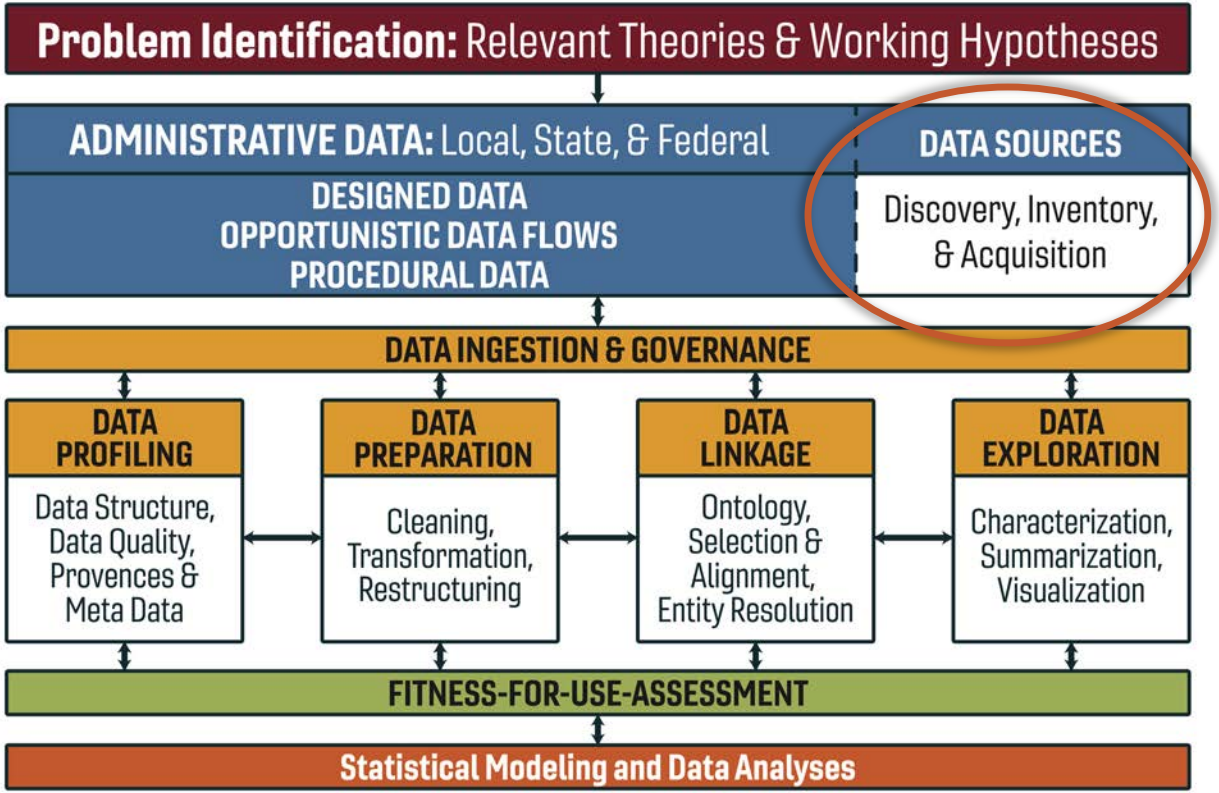
Department of Defense

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

Industry

MITRE Corporation
Proctor & Gamble

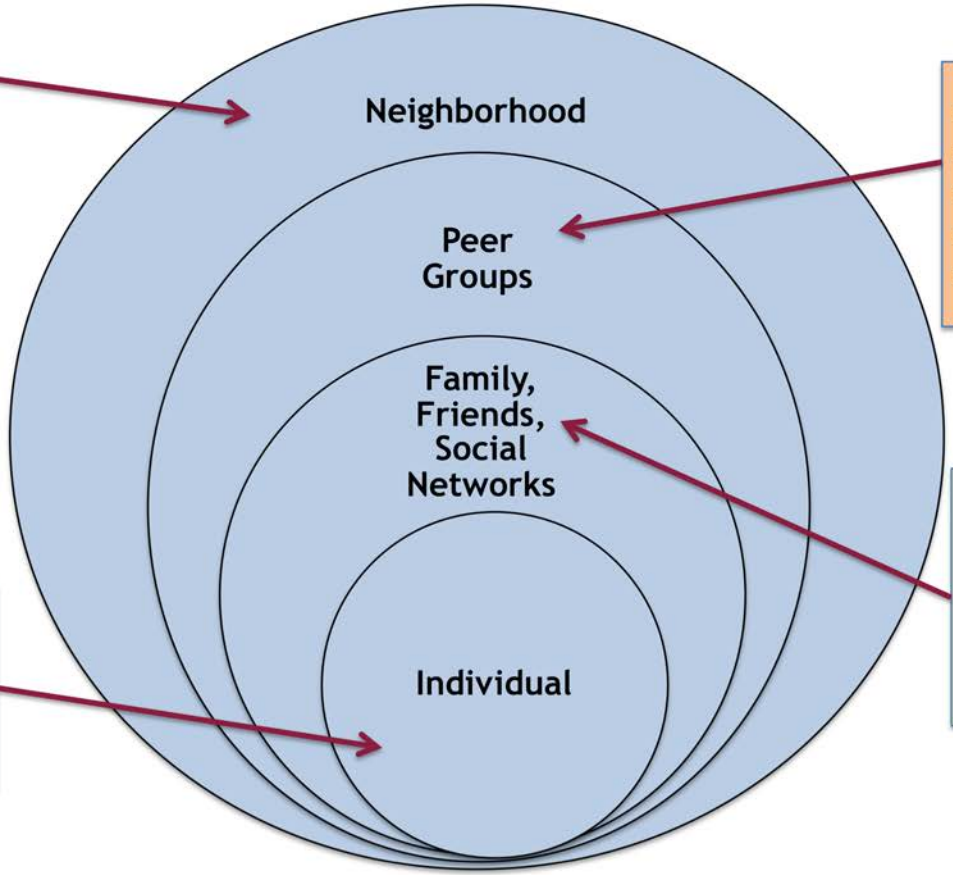
Our emerging *Data Science Framework*



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

Local community Data Map

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



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

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

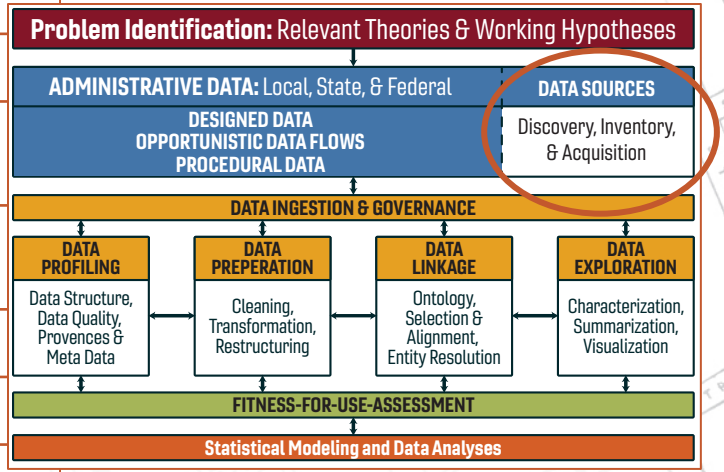
- Family Stability
- Income Stability
- Living Conditions
- Health Literacy
- Support Networks

Data Discovery, Inventory & Acquisition

Data Source	Geography
American Community Survey data (Census), 2011-2015 **(updating now to 2012-2016)**	Census Tracts and Block Groups
American Time Use Survey (BLS), 2017	National
Youth Risk Behavior Surveillance System, 2015	State
County Health Rankings, 2017	County
Built Environment, e.g., Grocery stores, SNAP retailers, recreation centers, community gardens	Address Level
Fairfax real estate tax assessment data	Address Level
Fairfax Open data: Zoning, Environment, water, Parks, Roads	Shapefiles
Fairfax County Youth Survey, 2016 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

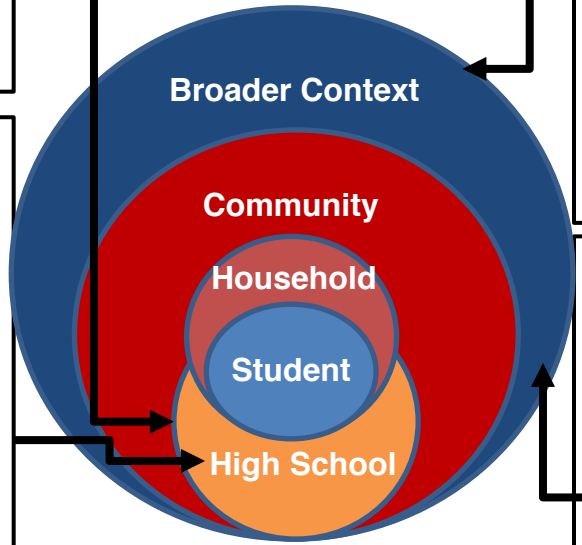
Data Map

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)

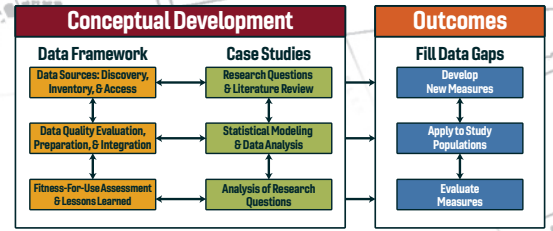
Perception of Postsecondary Availability

- 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)

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

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

Exercising the our full research model

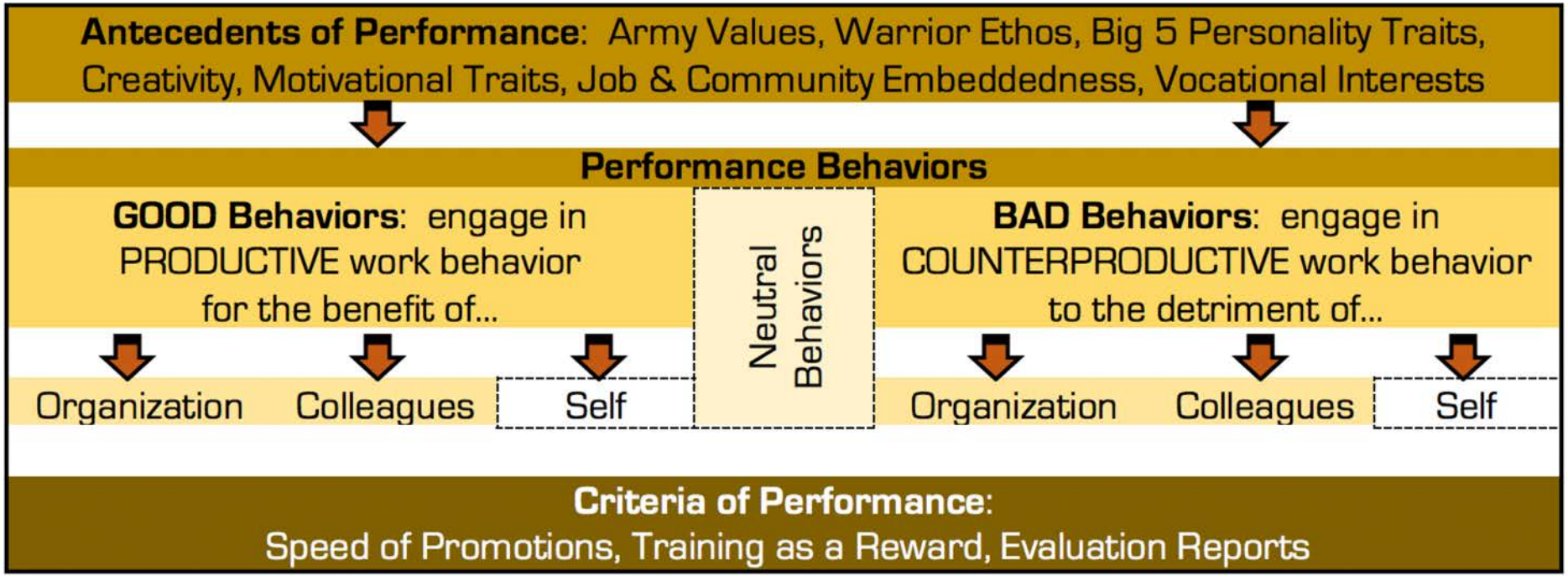


Research Questions:

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

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

Initial Performance Framework

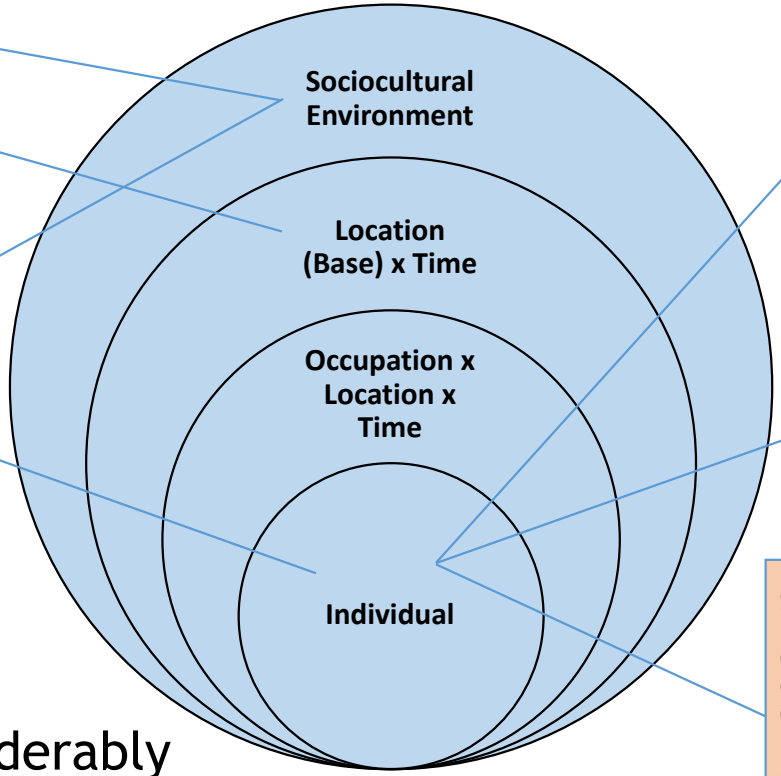


Soldier Data Map

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

Local community (e.g., ACS data)

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



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

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

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

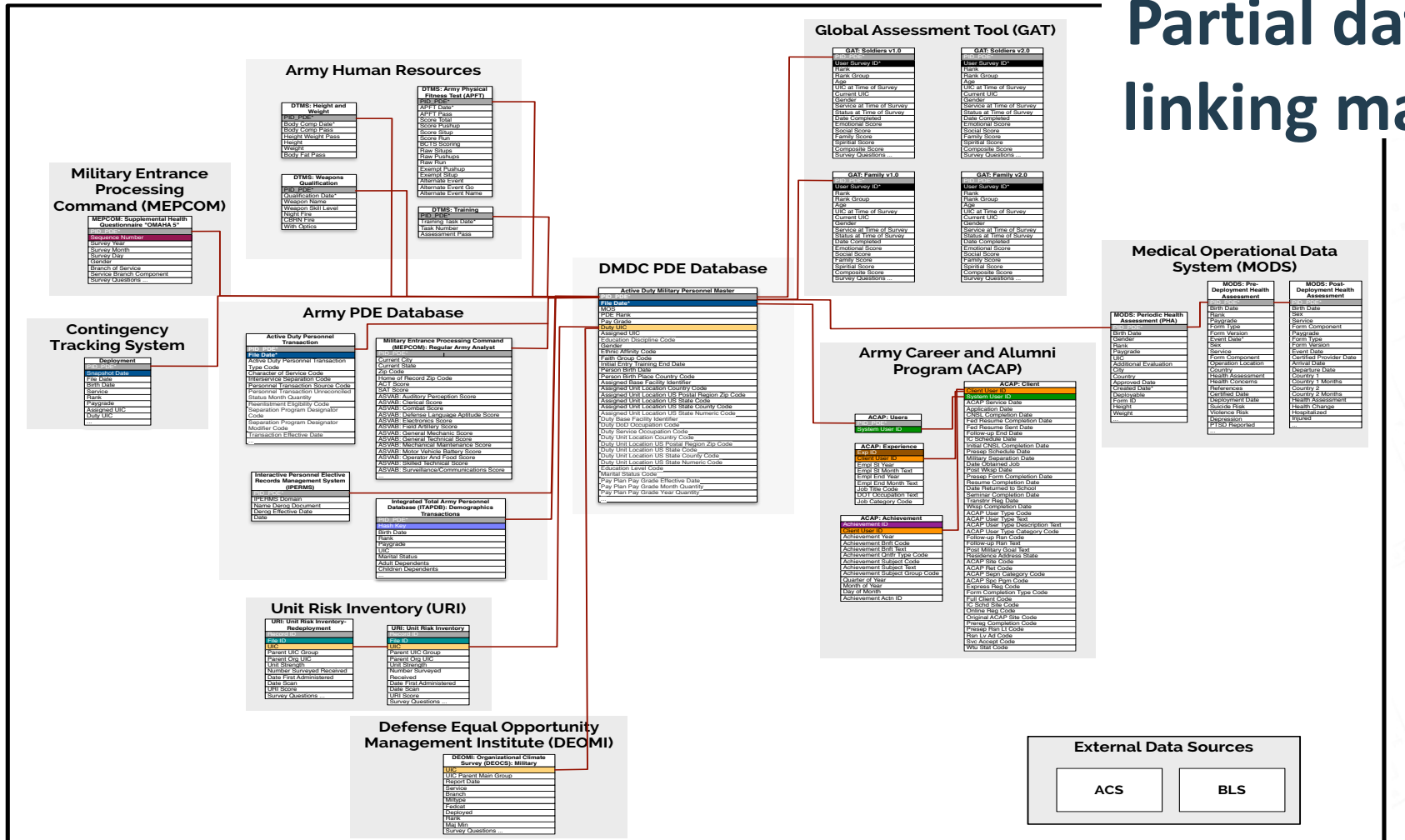
This will grow considerably

Data access

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

Person-Event Data Environment

Partial data linking map



Data pipeline: sharable data products

Demographics Table

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

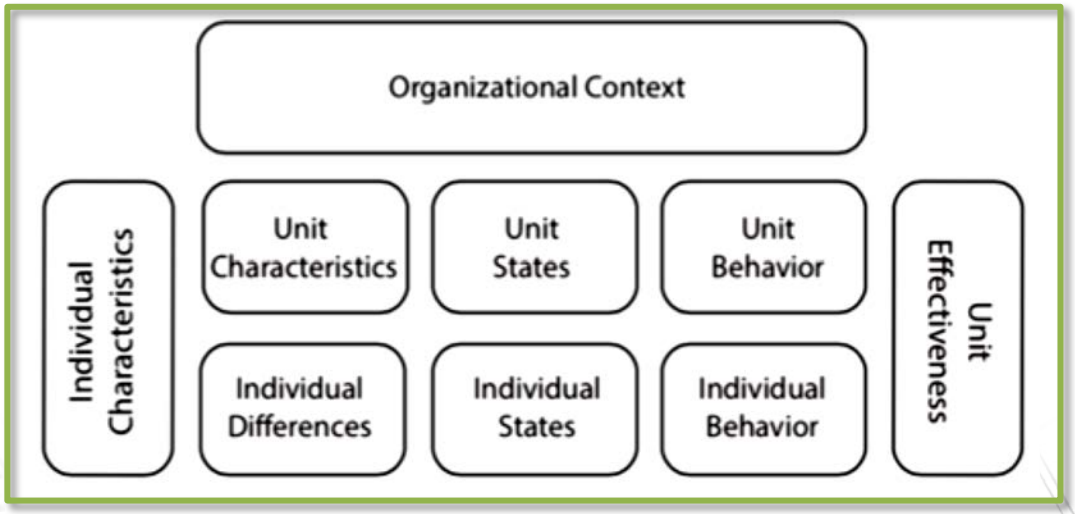
Transaction Table

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

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

Building model complexity

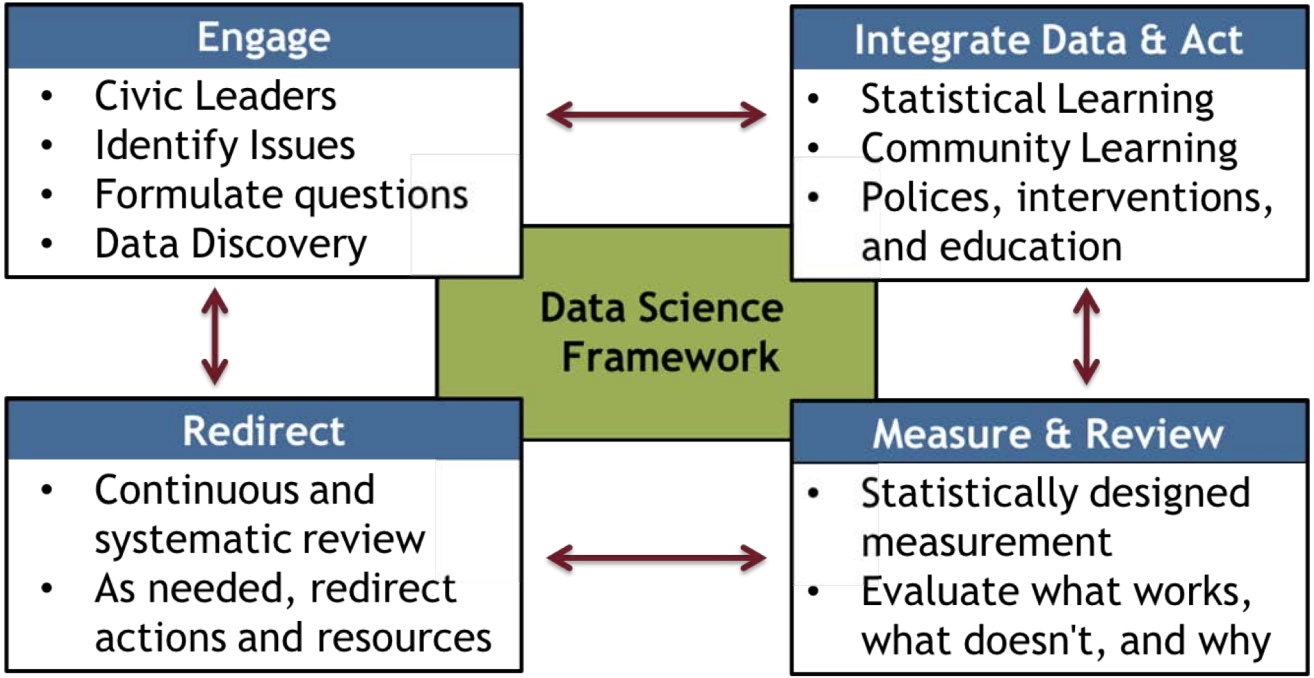
- Model **flexibility** for connecting many data sources and computation
- Need to integrate "external" data sources that **change over time**
- Need to **integrate** person-specific information **in context**
  - Relevant time and activity is with respect to person's term
  - "Exposures" to duties,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



Enhancing Prosperity through Data Science

Translating our research model:

Community Learning through Data-Driven Discovery



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

Observations from our research thus far

Key community-based research issues that have emerged:

- **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 research:

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

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

Re-Distribution of Data and Estimates Across Geographies

Problem: Data do not align with geographies of interest

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

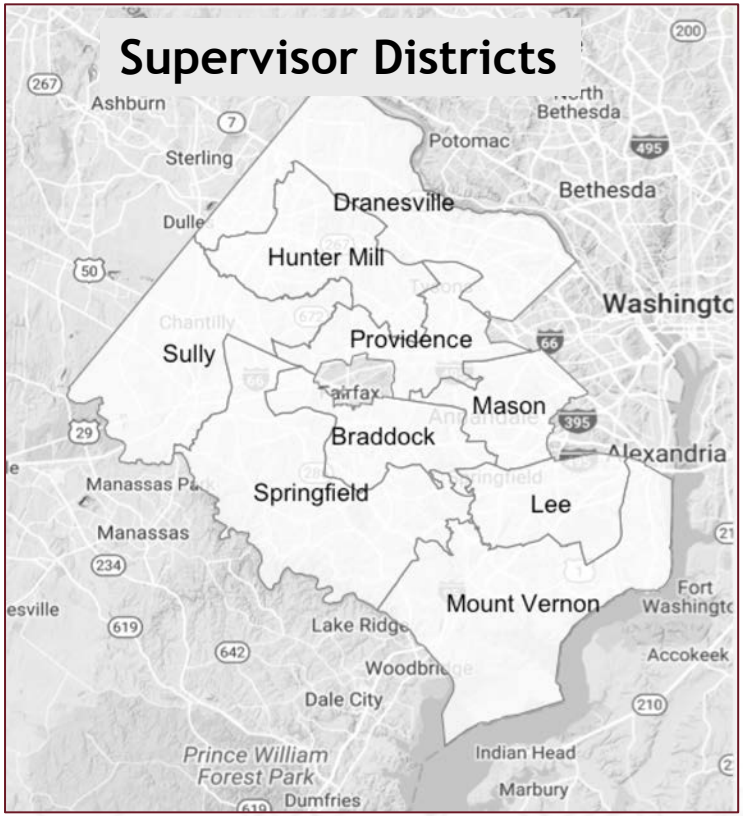
Solution: Use data **direct aggregation**, if possible, alternatively develop **synthetic populations** based on data and redistribute

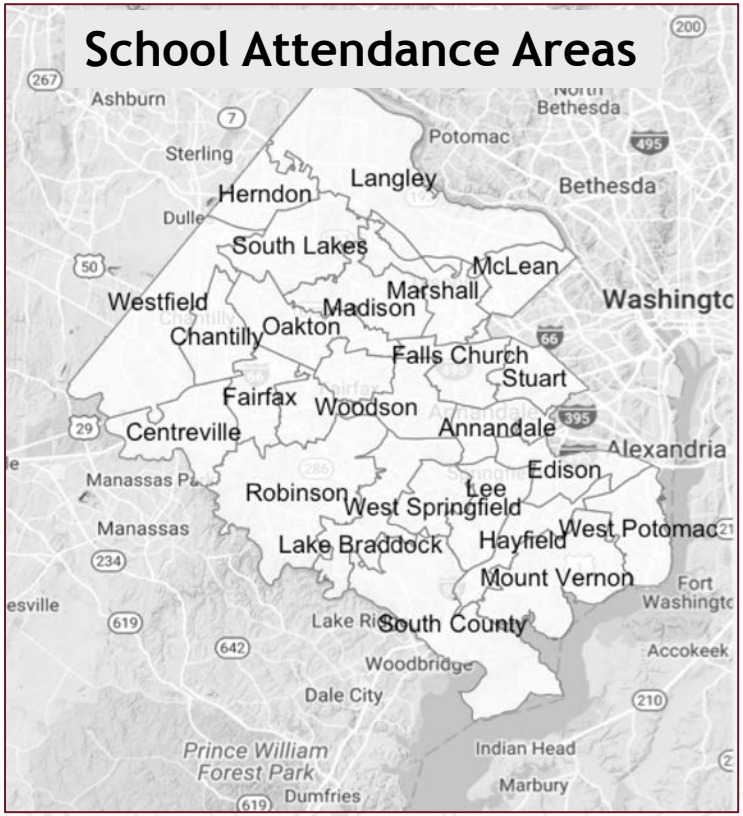
Synthetic re-distribution based on variables of interest

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

Example: Fairfax County, Virginia

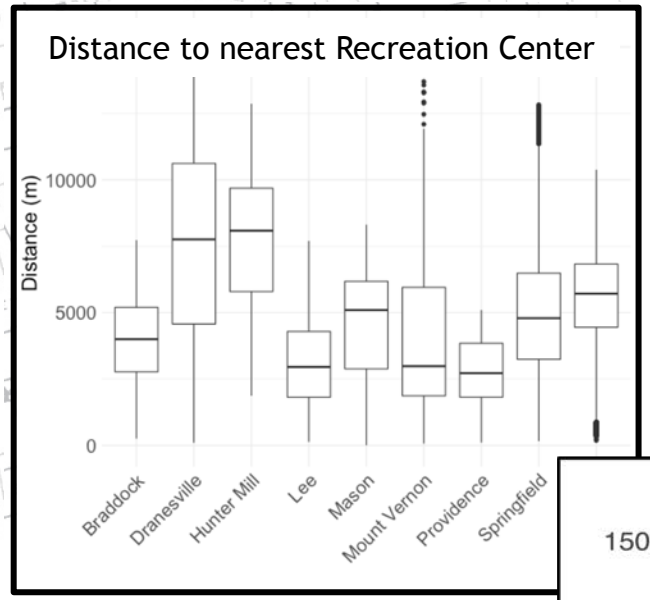
Supervisor Districts and High School Attendance Areas

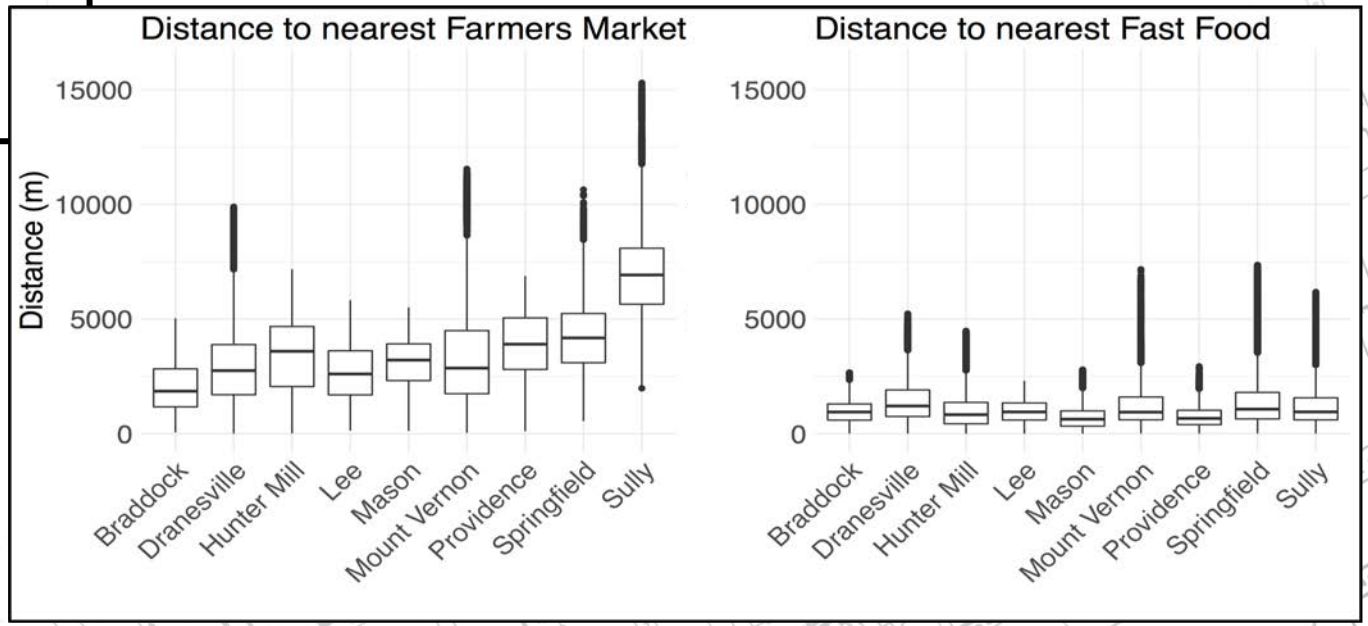




Direct aggregation based on location of housing units

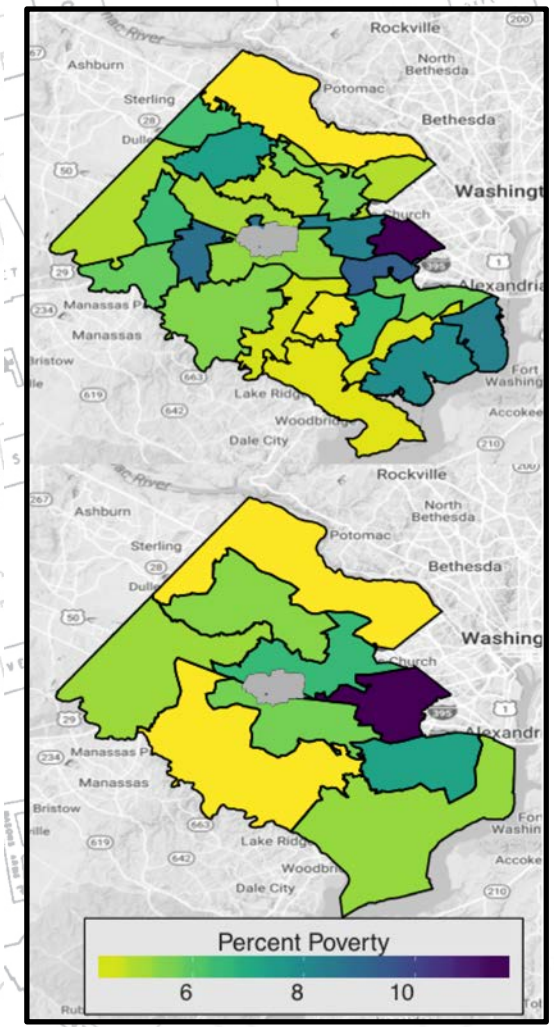
- Geocoding owner-occupied local housing stock
- Adding rental units typically requires imputation





Examples of **place data**:

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

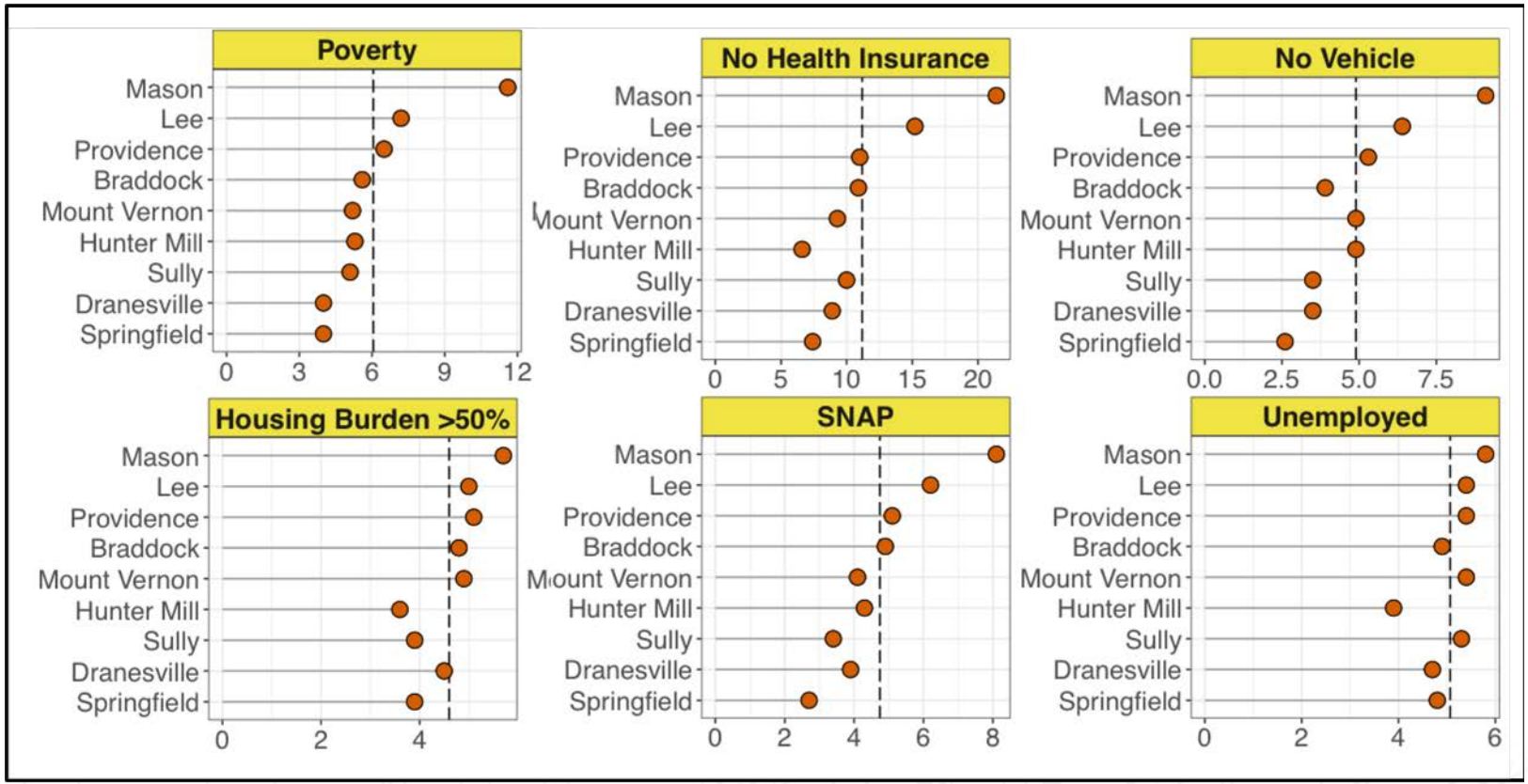


Re-distribution of data based on synthetic populations

- Use American Community Survey (ACS) summaries and PUMS microdata to impute synthetic person data for all people in area of interest
- Re-weight synthetic data according to ACS tables to simultaneously match the relevant distributions, to Census Tracts or Block Groups
  - **Age, income, race, and poverty in this case**
- Aggregate synthetic person data to compute summaries, and margins of error, over the new geographic boundaries

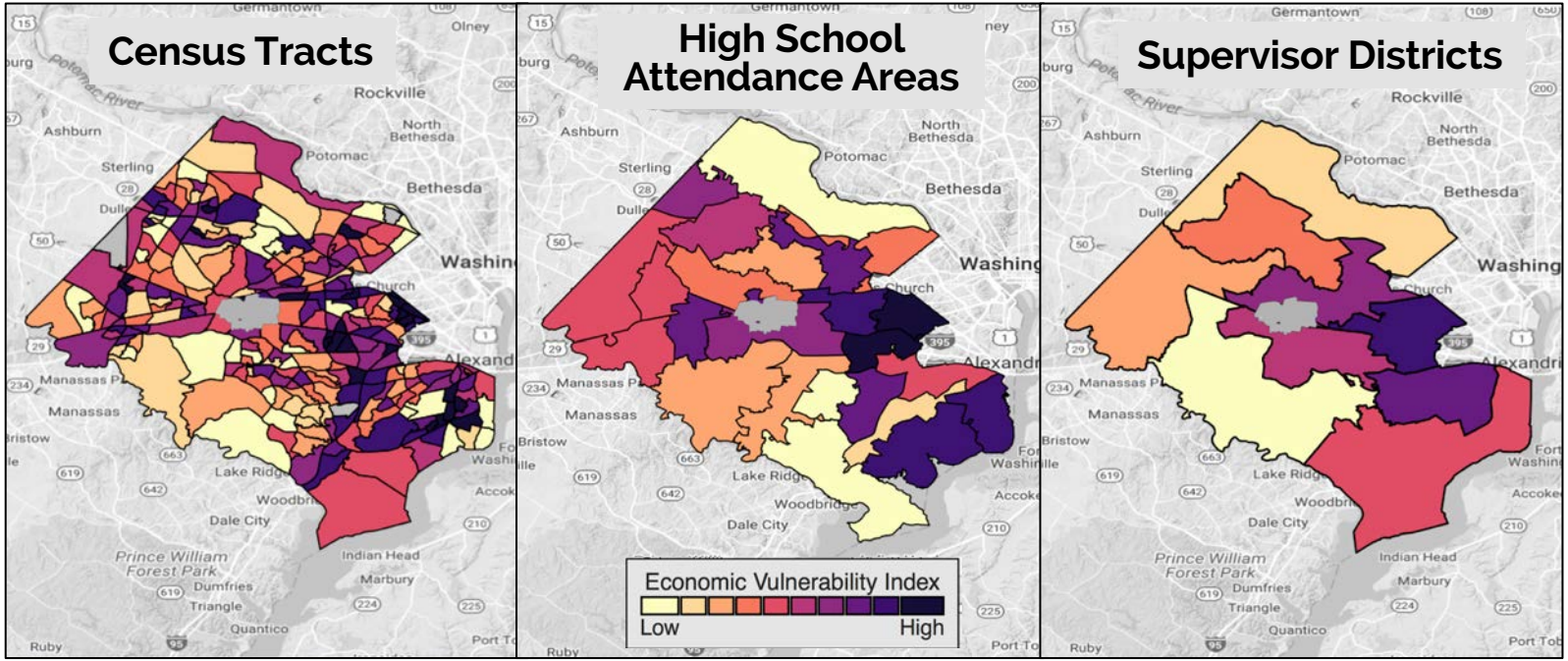
Fairfax Profiles by Supervisor Districts

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



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

Fairfax Sub-County Vulnerability Indicators



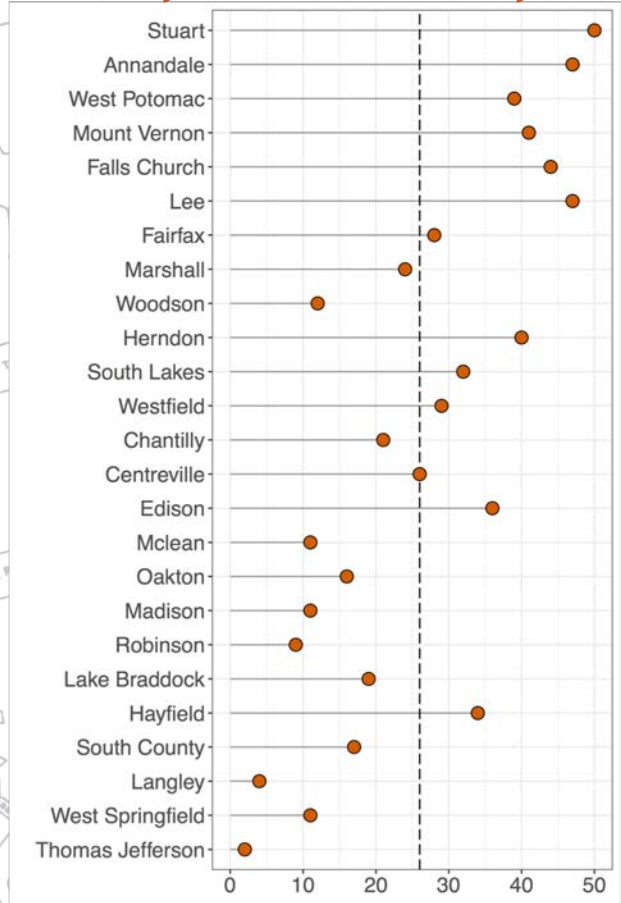
Based on a **statistical combination** of the percentage of Households with:

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

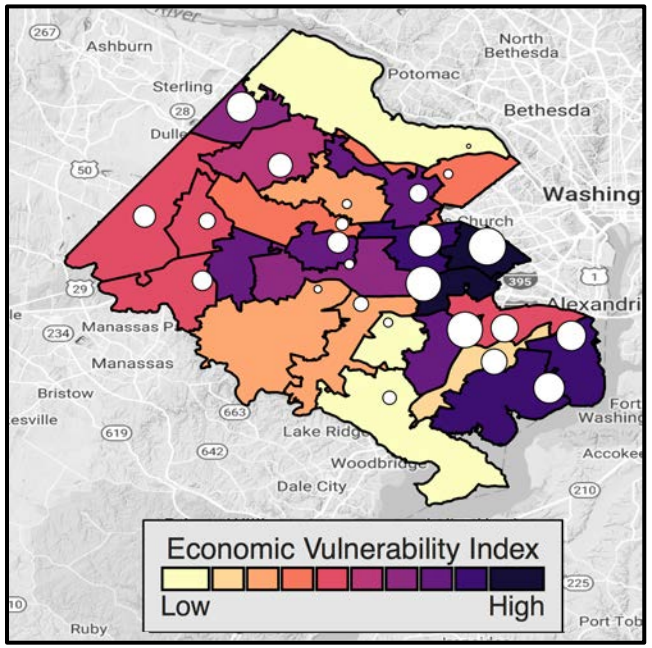
Source: American Community Survey 2011-2015 aligned to Supervisor Districts using **SDAL Synthetic Technology**.

High School Characteristics

High School Vulnerability Index ordered by Economic Vulnerability Index



School Vulnerability Index



Combination of:

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

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

Population Dynamics

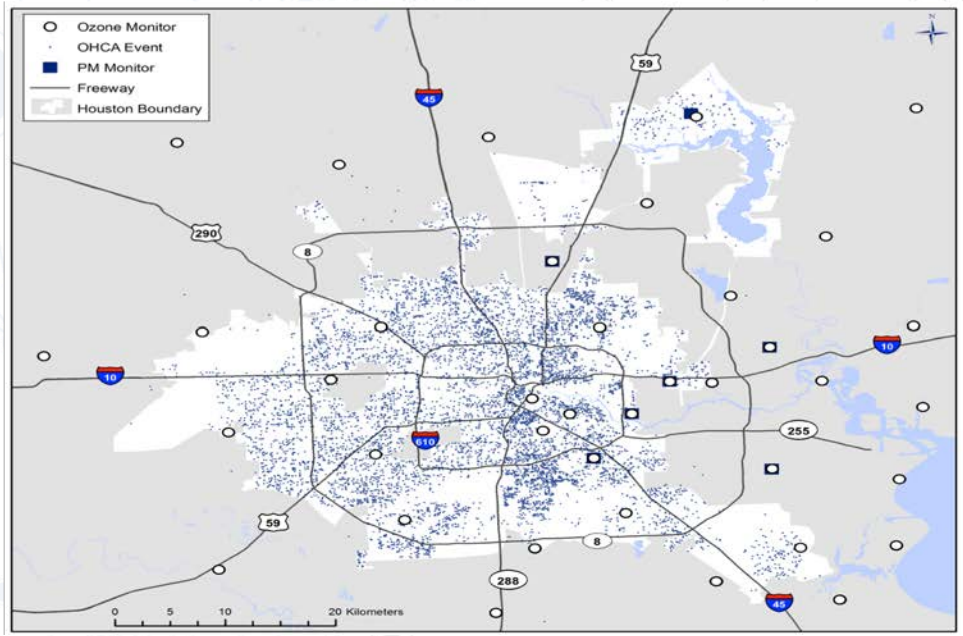
B. Pires, G. Korkmaz, K. Ensor, D. Higdon, S. Keller, B. Lewis, B., and A. Schroeder, 2018. Estimating individualized exposure impacts from ambient ozone levels: A synthetic information approach. *Environmental Modelling & Software*. (Forthcoming)

Houston EMS Study for Individual Risk

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

Model and Data:

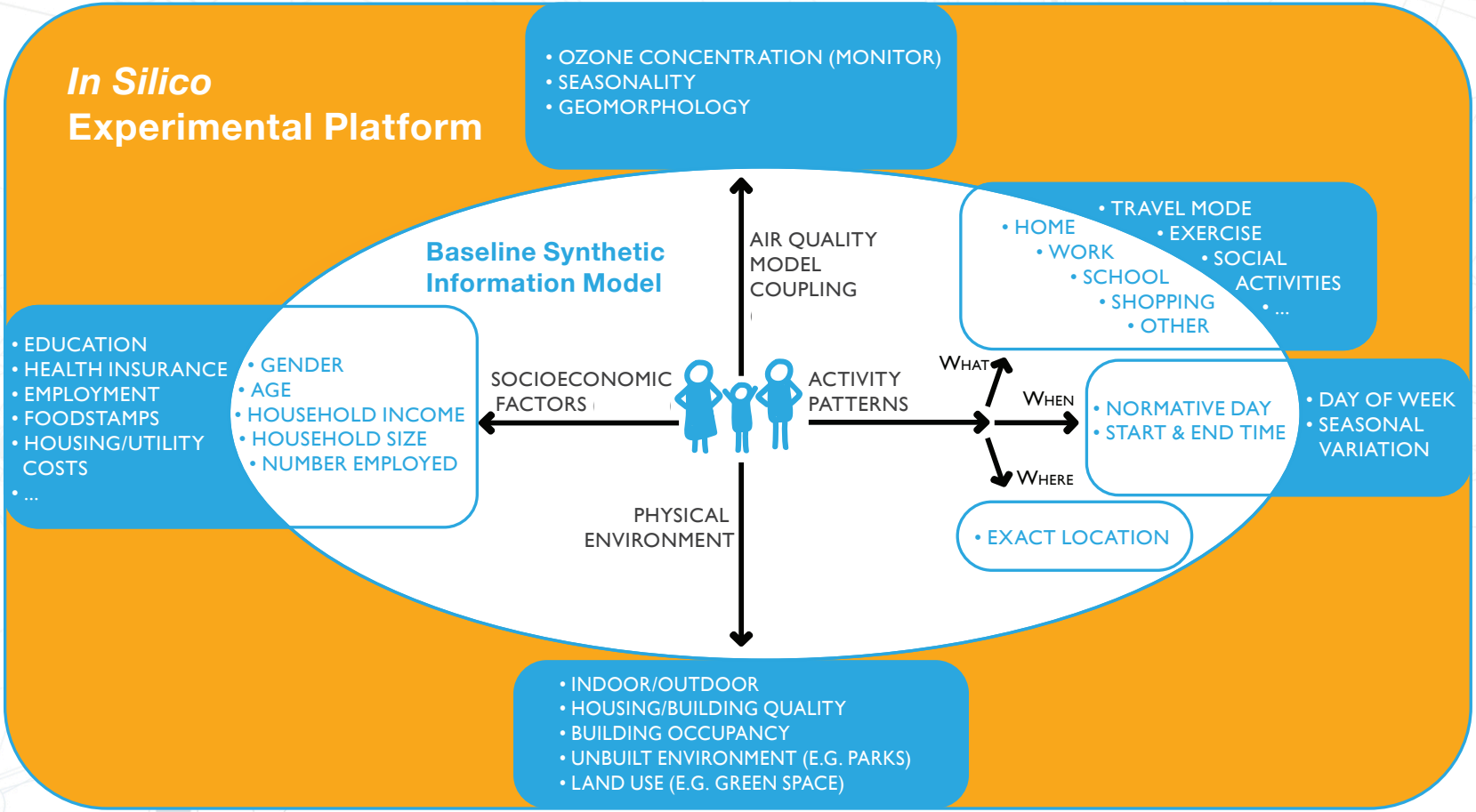
- Pathophysiological link between out-of-hospital cardiac arrest (OHCA) and ozone level
- Case cross-over, time stratified design
  - Houston, 2004-2011
  - EMS data of 11,754 cases
  - Predictor variable is *aggregate ozone over a 3 hour window* leading up to the event



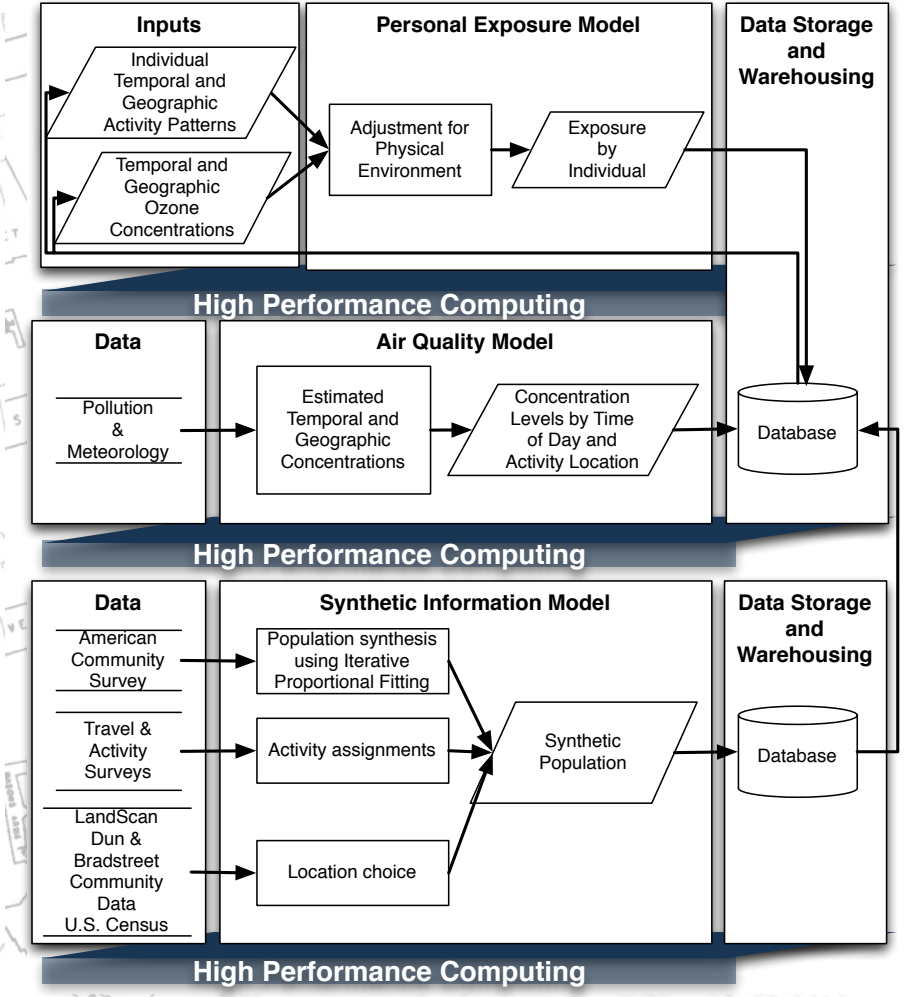
American Heart Association®

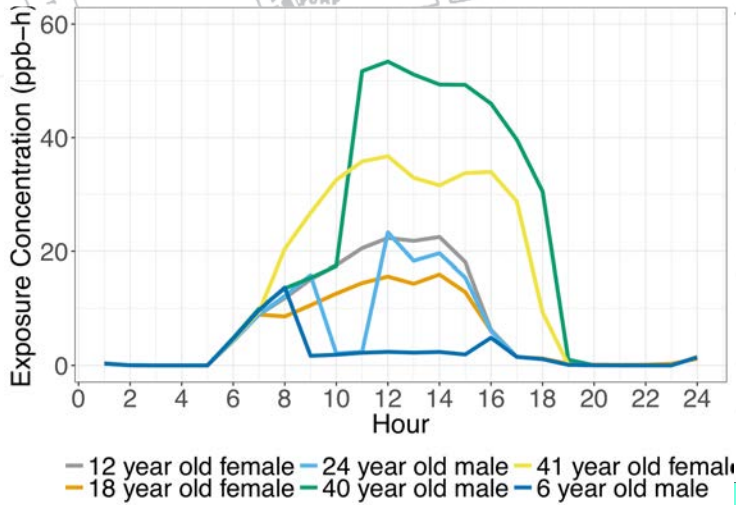
Ensor, et al., Circulation, Volume 127(11):1192-1199

Synthetic Information Platform

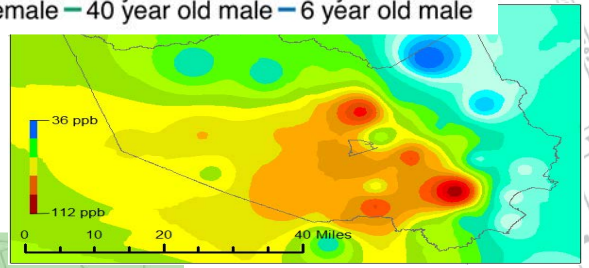


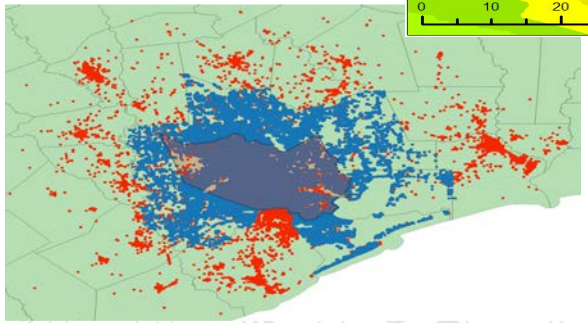
In-Silico Platform for Environmental Coupling





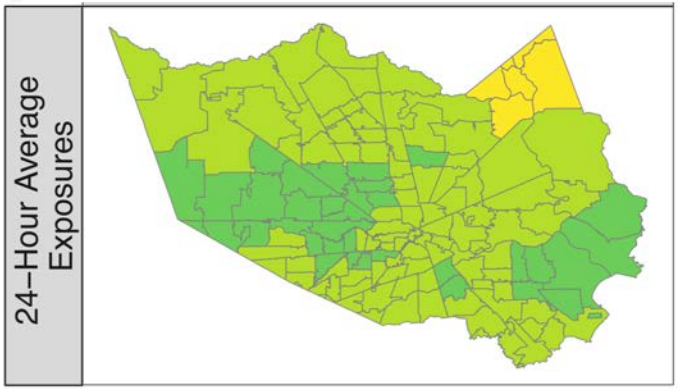
10:00 am
August 26, 2008

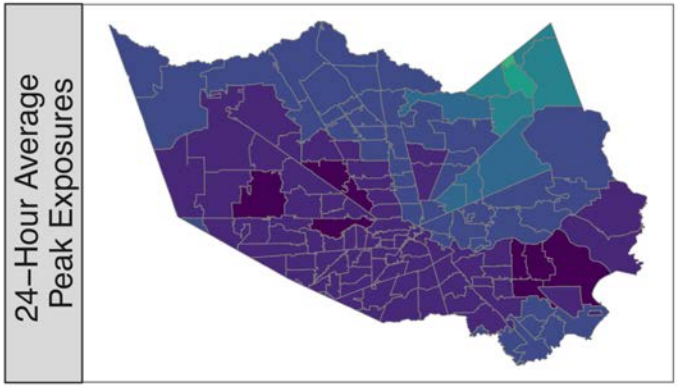




4.9M people
1.8M Households
1.2M Locations

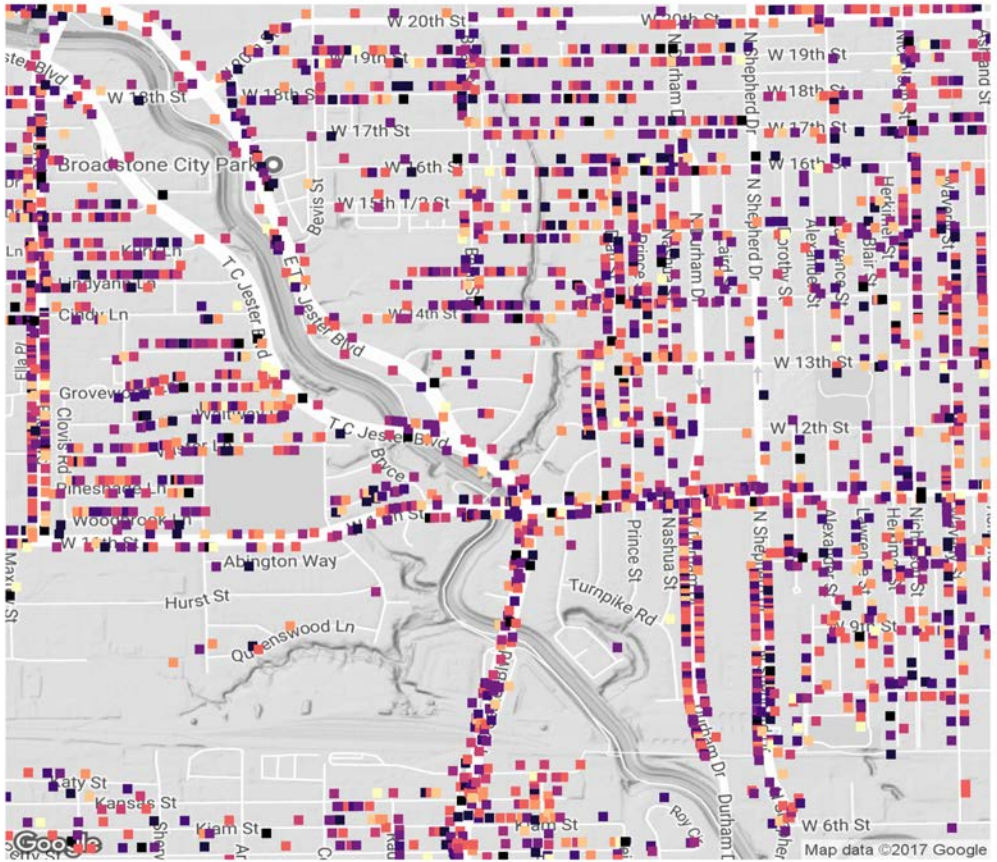
Location and movement matter



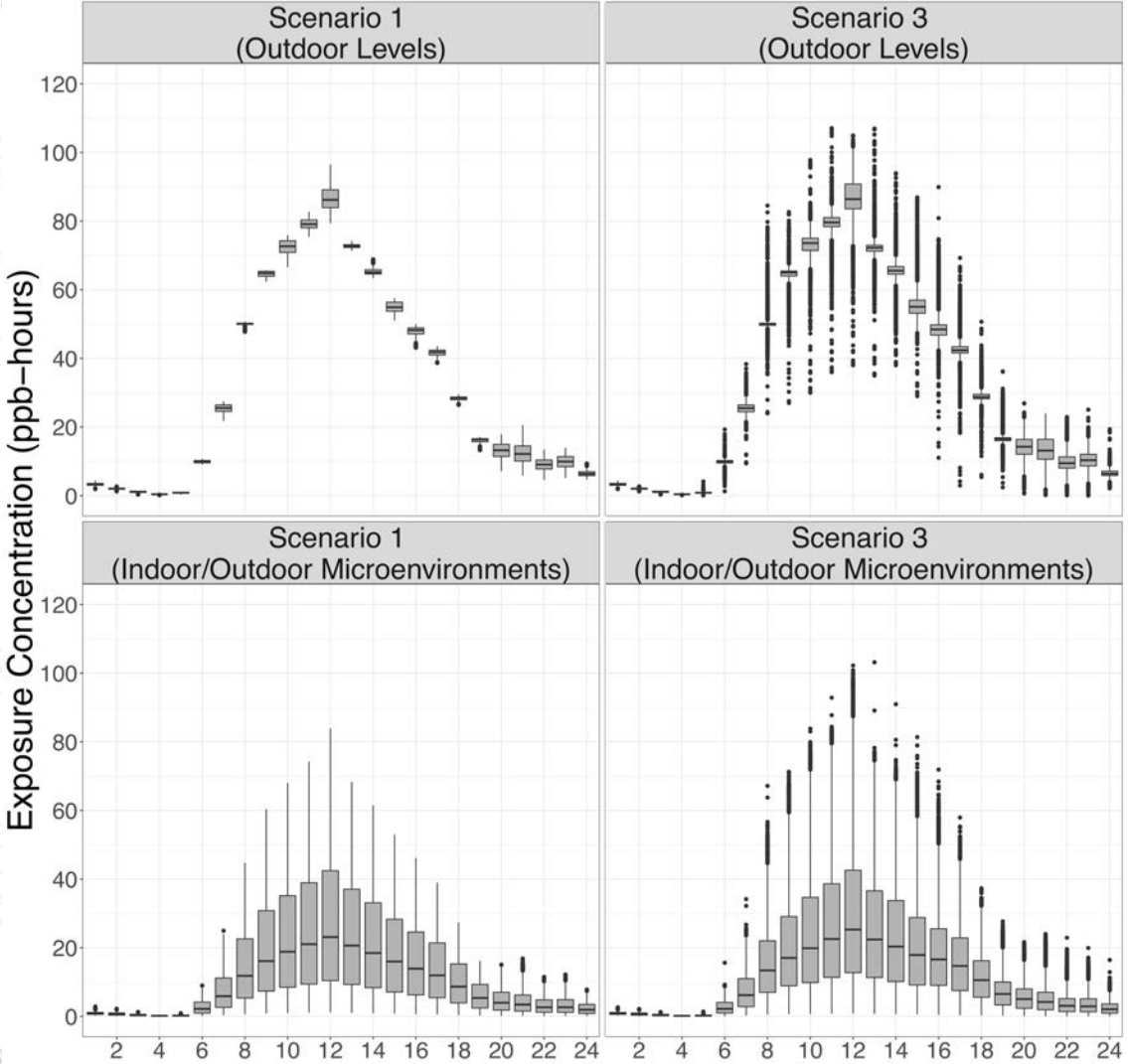


Exposure Concentrations (ppb) by Zip Code

8 18 27 36



24–Hour Household Average Exposure Concentration (ppb) 0 35



Exercising the platform

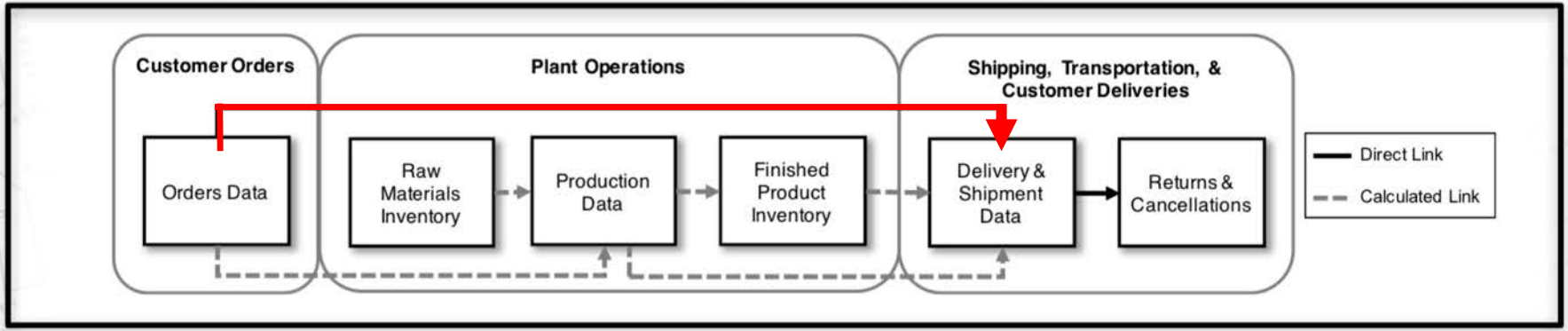
Scenario 1:
Population stays home

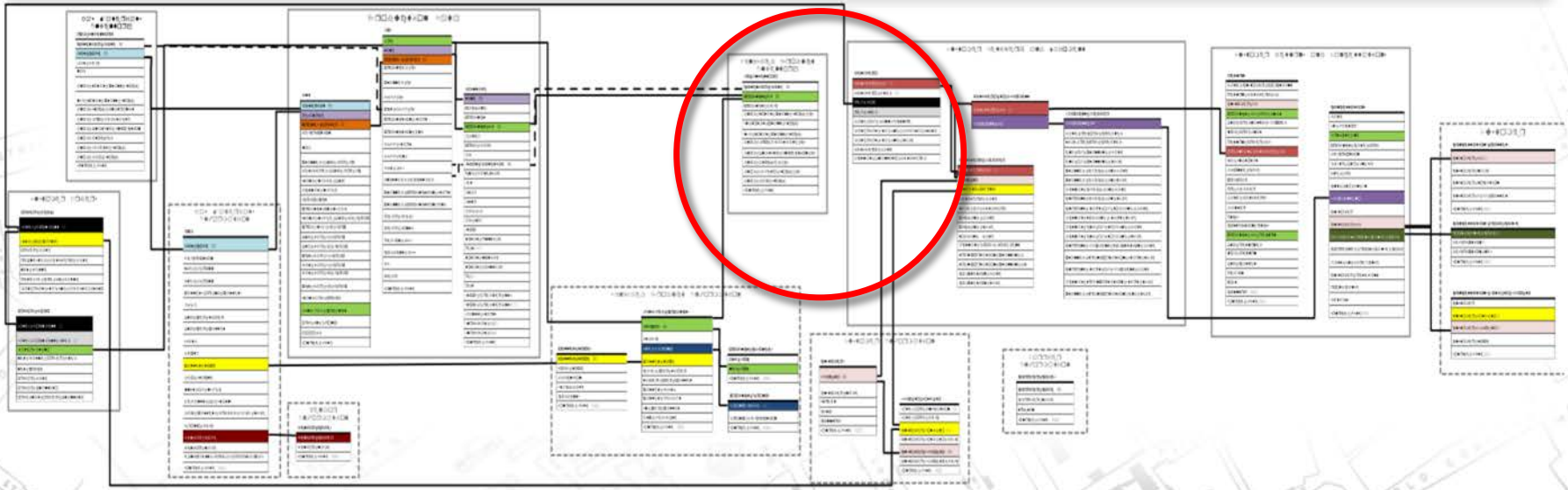
Scenario 3:
Population moves

Fortune 500 Company

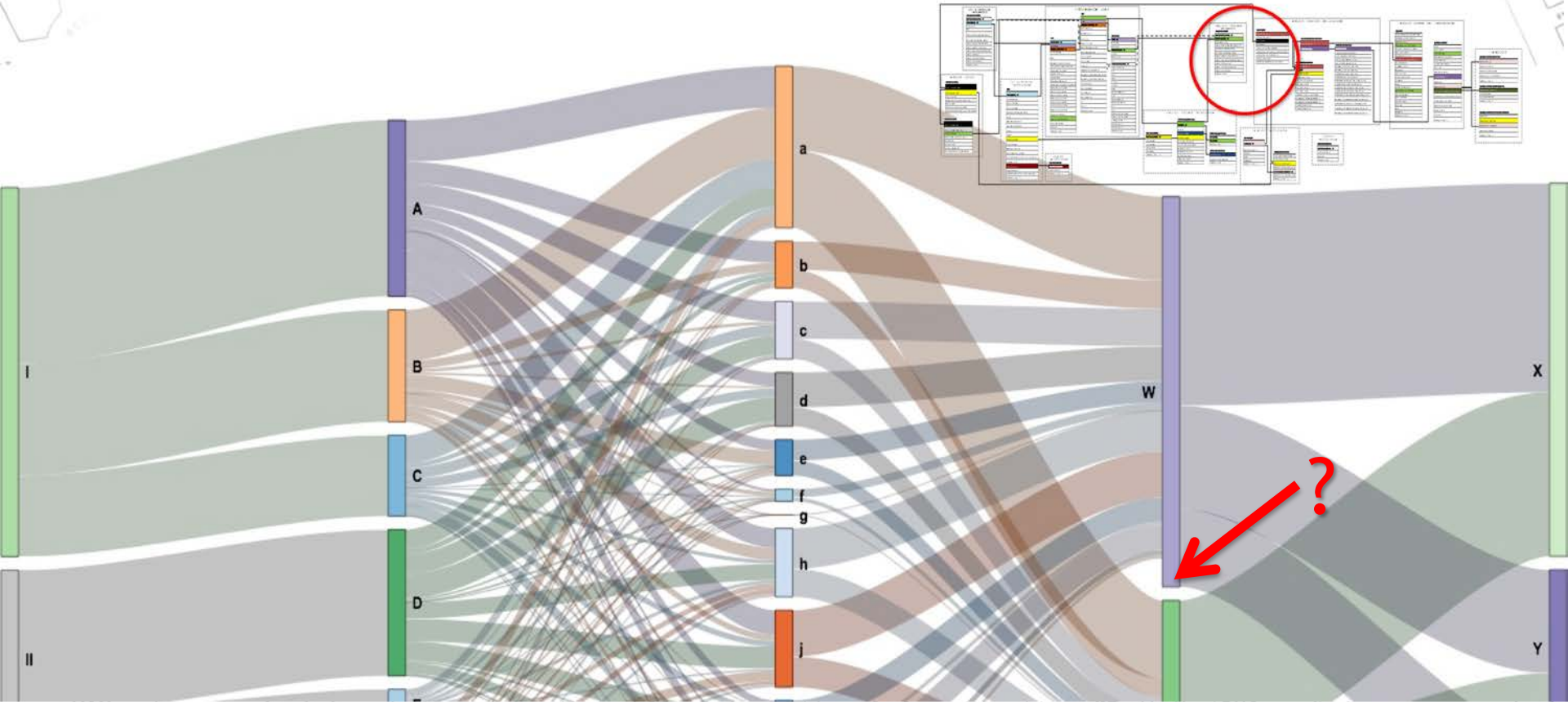
B. Pires, J. Goldstein, D. Higdon, S. Reese, P. Sabin, G. Korkmaz, S. Ba, K. Hamall, A. Koehler, S. Shipp, S., and S. Keller, 2017, A Bayesian Simulation Approach for Supply Chain Synchronization, in the *Post-Proceedings of the 2017 Winter Simulation Conference* (WSC), 3rd – 6th December, Las Vegas, NV.

Supply Chain flows are complicated by humans





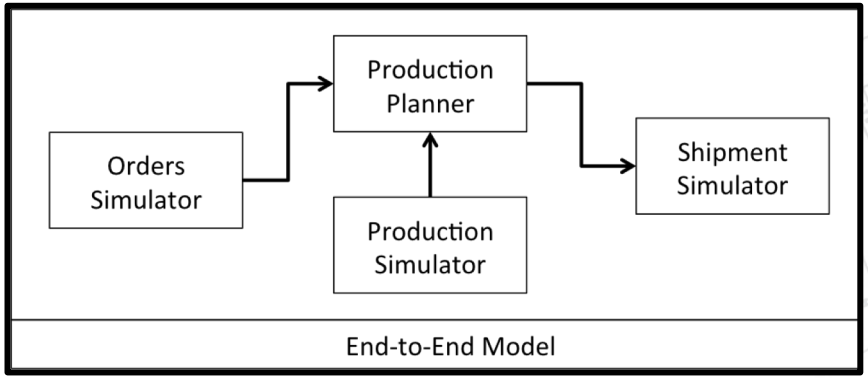
Unravel transactional information flows and link to supply chain production activities



Modeling the Supply Chain End-to-End (E2E)

Combine a Bayesian approach with discrete-event simulation

- Inform using **current transactional data**
  - Orders
  - Line specific production dynamics
  - Raw materials
  - Current inventories
  - Shipments

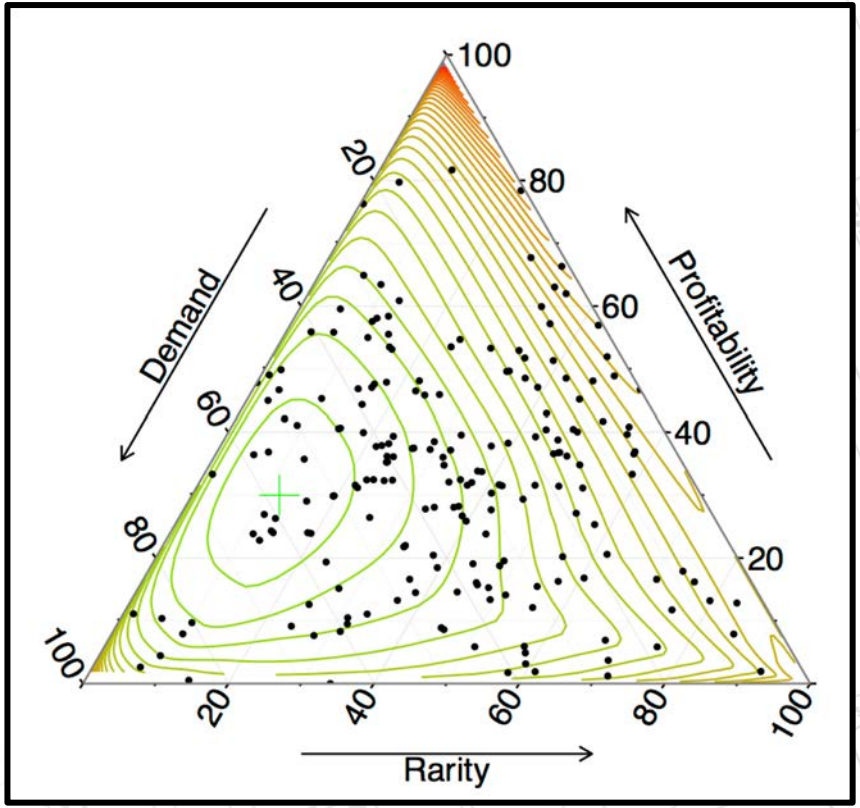


- Four **integrated** simulators:
  - Orders Simulator matches order quantity to empirical patterns from customers
  - Production Simulator estimates the rate for production runs
  - Production Planner produces production schedules
  - Shipment Simulator models the loading, shipment, and delivery of finished products to customers

Resulting in a framework for a data-driven understanding of supply chain dynamics

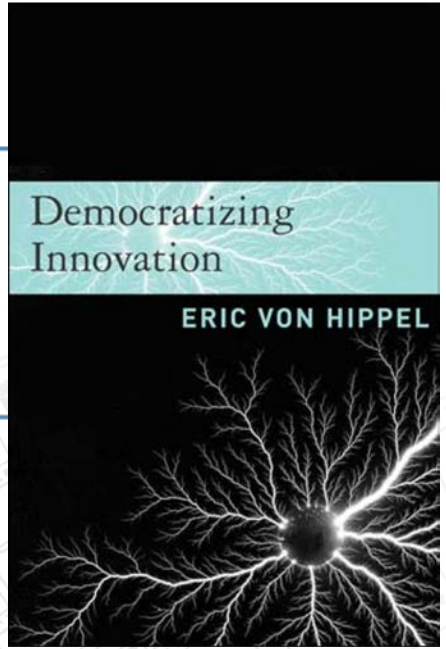
Simulation-based investigation

- Maximizing (profit, on-time delivery)
- Carried out a sequence of runs uniformly over the SKU space of:
  - Demand
  - Profitability
  - Rarity
  - Production times
  - Safety stock
- Fit a response surface using a Gaussian process emulator to seek out an optimal settings for supply chain synchronization



NCSES National Center for Science and Engineering Statistics

Measuring Innovation: Open Source Software



S. Keller, G. Korkmaz, C. Robbins, and S. Shipp, 2018. Modeling, Infrastructures, and Standards: New Opportunities to Observe and Measure Innovation. *Proceedings of the National Academy of Sciences,* (in-revision).

Why Care?

Open Source Software **(OSS)** are digital products, including those provided without direct payment

- OSS is **used across fields**; e.g., Google Chrome, Linux, R, Python, Wikipedia...
- OSS **supports research outputs**; e.g., peer reviewed publications, patents, startups, licenses …

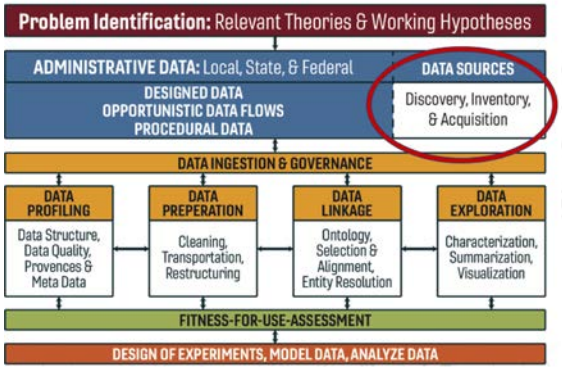
- Innovation that is being **created outside of the business sector**
- **"The Open Source World is Worth Billions." (Redman 2015)**
  - Could be **missing a major contribution** to economic growth

Challenge: Can the scope and impact of OSS be measured using publicly available data?

Can the scope and impact of OSS be measured using publicly available data?

Desirable data dimensions for measuring OSS:

- **Stock Measures:** How much open source software is in use?
- **Flow Measures:** How much is created each year?
- **Categories:** What types can be identified?
- **Sectors and Collaborators:** Who creates it?
- **Users:** Who benefits for its development?



Discovered Sources

stackoverflow

GitHub

osalt.com

SourceForge

depsy

BLACKDUCK Open Hub

ggplot2

A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the ...

Tags

graphics phylogenetics

103 contributors

- *RStudio*
- *Hadley Wickham*
- *Winston Chang*
- *kohske takahashi*
- *tidyverse*

+98 more

View in API

Get badge

100 percentile impact overall

Compared to all research software on CRAN , based on relative downloads, software reuse, and citation.

Downloads

6.3M 100 percentile

Based on latest downloads stats from CRAN.

Citations

1.7k 100 percentile

Based on term searches in ADS (0) and Europe PMC (1702)

Read more about how we got this number.

Dependency PageRank

10.00 100 percentile

Measures how often this package is imported by CRAN and GitHub projects, based on its PageRank in the dependency network.

Read more about what this number means.

depsy

Reused by 9019 projects

Hmisc

Contains many functions useful for data analysis, high-level graphics, utility operations, functions...

ggmap

A collection of functions to visualize spatial data and models on top of static maps from various on...

rstan

User-facing R functions are provided by this package to parse, compile, test, estimate, and analyze ...

GGally

The R package 'ggplot2' is a plotting system based on the grammar of graphics. 'GGally' extends 'gg...

lmerTest

Different kinds of tests for linear mixed effects models as implemented in 'lme4' package are provid...

1859 ML_for_Hackers

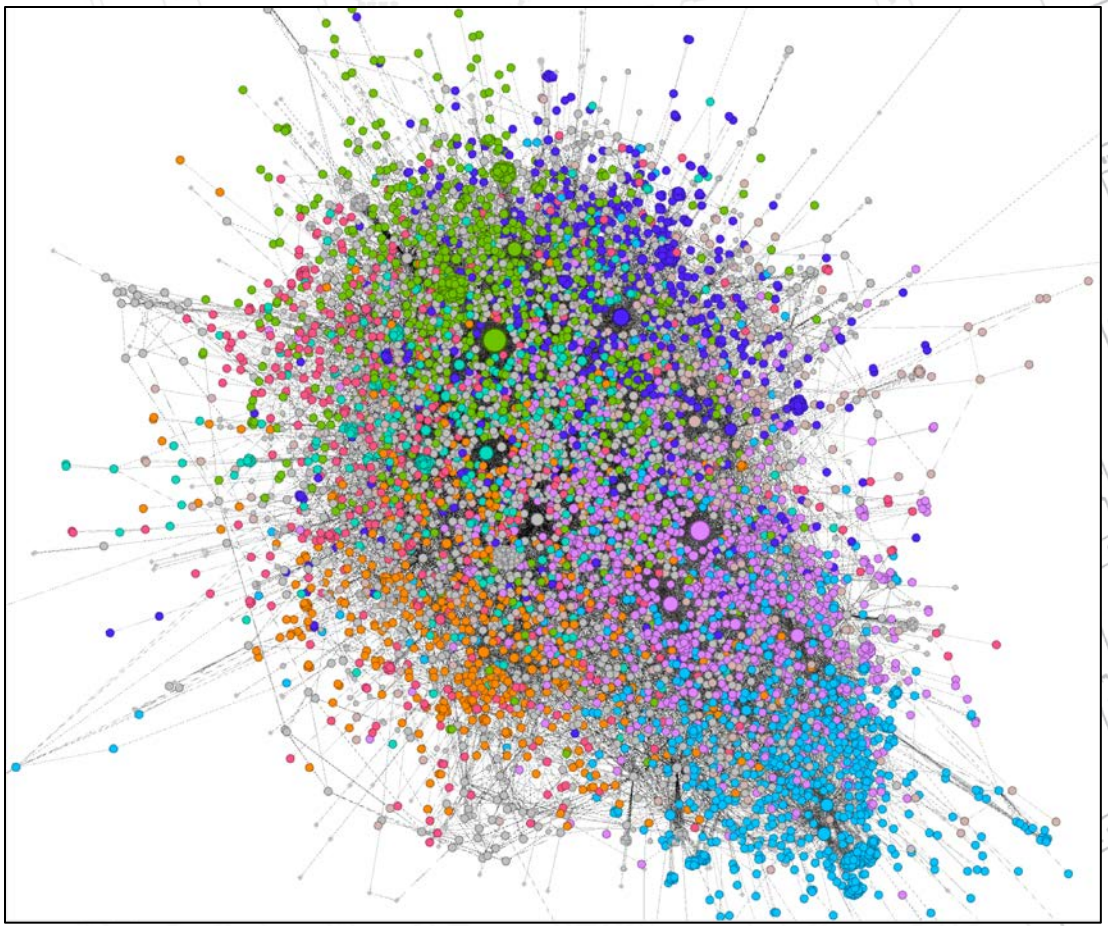
Code accompanying the book "Machine Learning for Hackers"

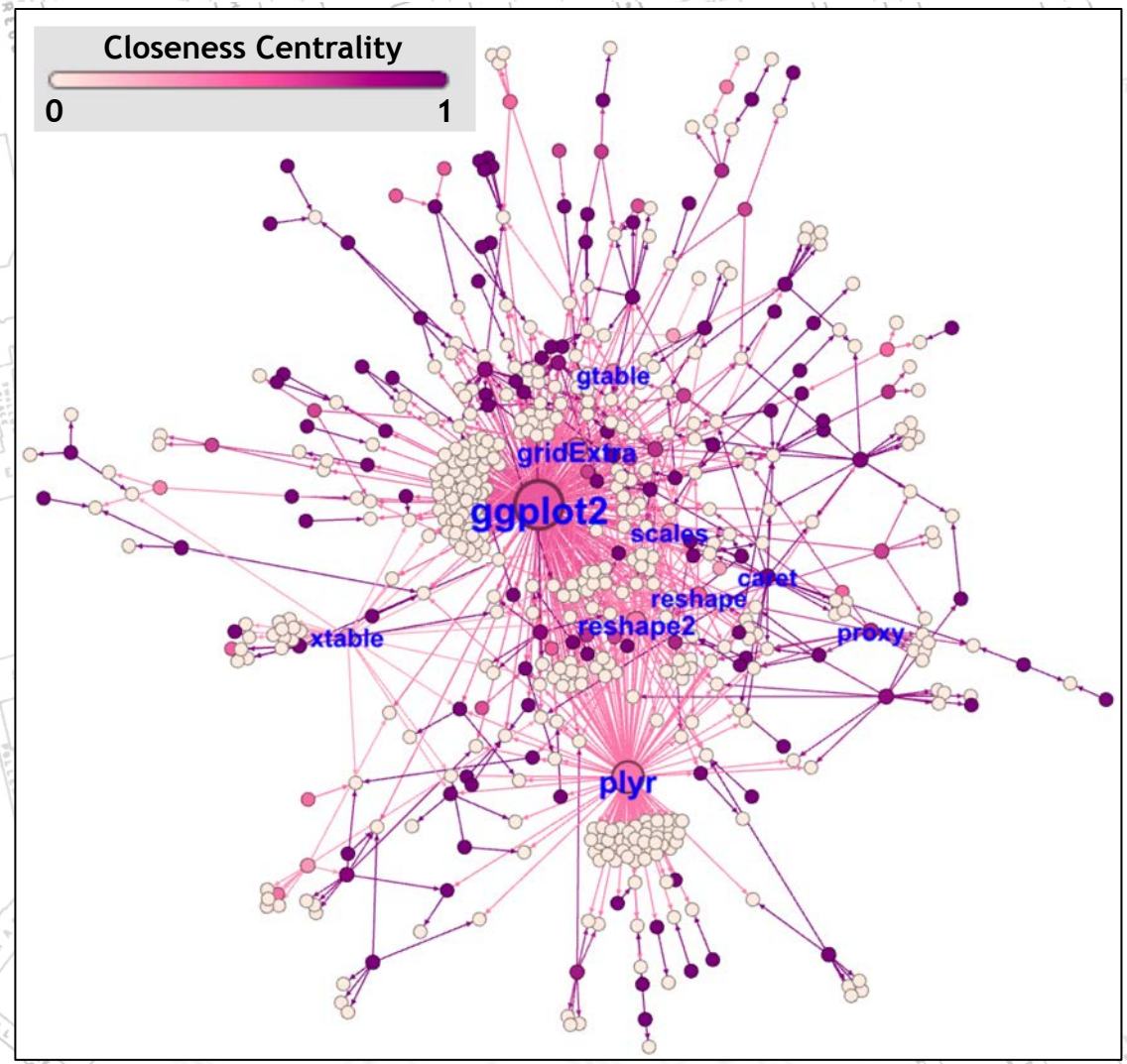
Dependency network of R packages

depsy

Identified communities (modularity class)

- Data wrangling, exploration & visualization
- Statistical analysis packages
- Web-based data/API processing
- Packages for matrix operations
- Spatial data analysis
- Time series analysis





Subgraph of dependency network of R Packages

Community with the largest number of nodes is illustrated (8%)

- Node size indicates the out-degree
- Node color represents the closeness centrality

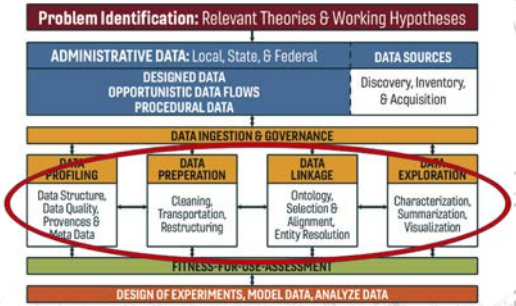
depsy

Returning to our research question: measuring scope and impact of OSS

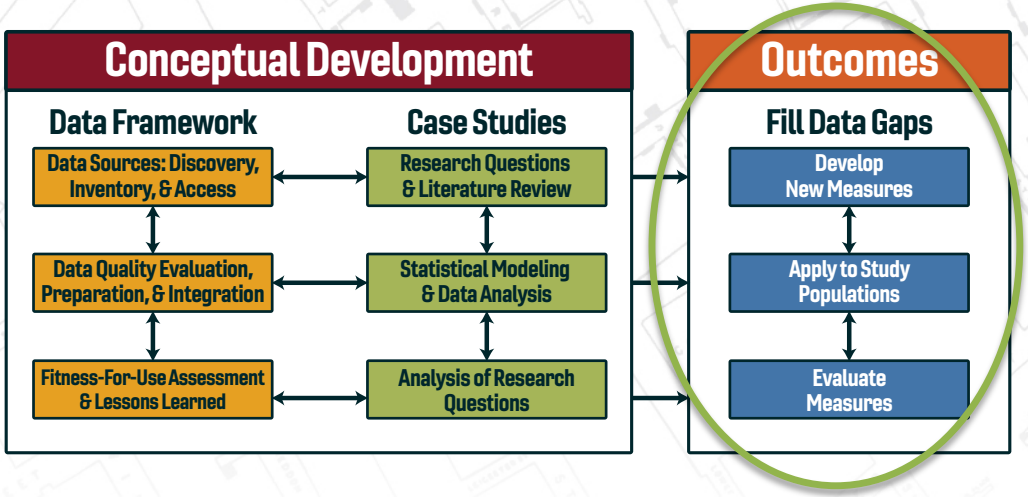
Next steps: Build accurate and repeatable models to predict costs to produce OSS

- Cost estimation models are mathematical algorithms or parametric equations used to estimate the costs of a product or project
- Common attributes in software development cost models:
  - **Product attributes** (reliability, complexity, reusability)
  - **Platform attributes** (execution time, storage constraints, volatility)
  - **Personnel attributes** (capabilities of analysts and programmers application, platform, language and toolset experiences)
  - **Project attributes** (use of software tools, multi-site development, required development schedule

Fitness-for-Use: Evaluate data quality and utility for capturing these attributes



New Basic Research



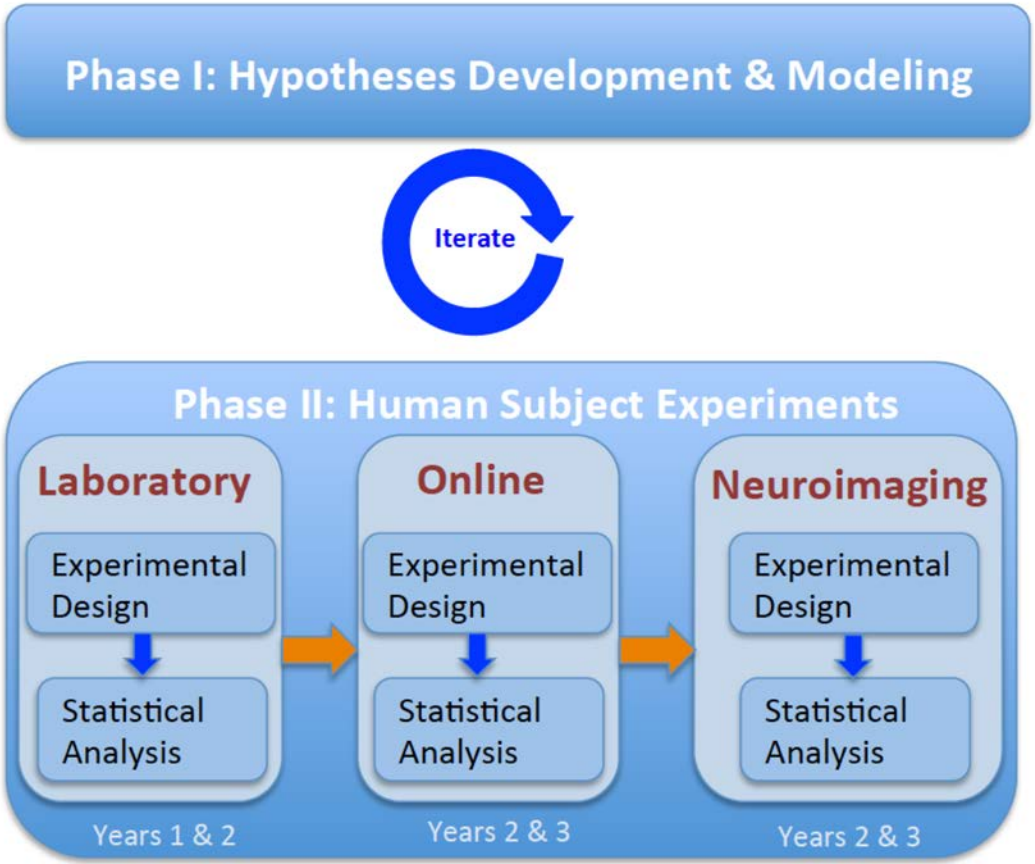
Collective Action and Coordination

- The use of *social networking sites* was a distinctive feature of uprisings
- Social media help to reach a *critical mass* of participants

  - Collective action problem: join only if joined by "enough" others
  - Coordination game: Two or more people each make a decision to participate with the potential to achieve shared *mutual benefits*
  - Coordination requires that people know about each other and that this information is common knowledge

Korkmaz, G., C.J. Kuhlman, S.S. Ravi, F. Vega-Redondo. 2017. "Spreading of Social Contagions without Key Players." World Wide Web. pp. 1-35.

Experimental Framework



- How do social networks facilitate actionable common knowledge?
- What is the role of network topology on coordination?
- How does it spread through the network?

Korkmaz, G., M. Capra, A. Kraig, C.J. Kuhlman, K. Lakkaraju, and F. Vega-Redondo. "Coordination and Collective Action on Communication Networks." *forthcoming* In Proceedings of the 17th ACM International Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2018).

Building Capacity

Scaling data science activities and influence

Local / State Government

- Practice community-based participatory

Federal Statistical Agencies

- Be explicit in cooperative agreement language

Department of Defense

- Expand researcher access to data

Industry

- Hands-on tech transfer models

Democratization of data across the United States

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

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

Workforce Development

Data Science for the Public Good (DSPG)

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

IDENTIFYING STEM EDUCATION PATHWAYS

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

EXPLORING MENTAL HEALTH SERVICES FOR FAIRFAX COUNTY YOUTH

Sponsor: Michelle Gregory, Sophia Dutton, and Linda Hoffman, Fairfax Health and Human Services

RESIDENTIAL SMOKE ALARM NEED IN ARLINGTON COUNTY

Sponsor: Battalion Chief Mike Gower, Arlington County Fire Department

HOW DO EVENTS AFFECT CRIME?

Sponsor: Captain Bruce Benson and Niki Levy, Arlington County Police Department

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

Sponsor: Carol Robbins, The National Center for Science & Engineering Statistics at the National Science Foundation

MODELING RESPONSE TIME FOR STRUCTURE FIRES

Sponsor: Battalion Chief Mike Gower, Arlington County Fire Department

PROFILE OF NEW KENT, VA

David Park, Joseph Kim, David Hinkle, Lata Kodali (Virginia Tech) with Di
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 for (SDAL)
Sponsor: Tod Massa (SCHEV – State Council for Higher Education in Virginia)

DEFINING AND MEASURING EQUITY IN ALEXANDRIA, VA

Sponsor: Emily Maslino, 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: Craig Russell, Andrew Slaughter, US Army Research Institute for Behavioral & Social Science Research

DISCOVERING NON-TRADITIONAL DATA SOURCES FOR BUSINESS INNOVATION

Harsimrat Pandher (VT), David Park (VT), Daniel Wilkin (VT), Joseph Kim (VT)
Claire Kelling (PSU) with Gizem Korkmaz and Stephanie Shipp (SDAL)
Sponsor: Gary Anderson, The National Center for Science & Engineering Statistics

A STUDY ON WMATA BUS FARE EVASION

Sponsor: Jayme M. Johnson, Catherine Vanderwaart, Washington Metropolitan Area Transit Authority

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

Claire Kelling (PSU), Kyle Morgan (VT), Craig Morton (VT), Hannah Brinkley (VT), Adrienne Rogers (VT, with Mark Orr, Stephanie Shipp, and Bianica Pires (SDAL)

Challenges

Cultural, Technical, and Infrastructure

- **Cultural Challenges**
  - Team activity
  - OPP life cycle
  - Patience to let significant research challenges emerge
  - Nature of publications
- **Technical Challenges**
  - Data sharing
  - Data and code pipeline development
  - Federated and sharable processes and platforms
  - Data engineers
- **Infrastructure Challenges**
  - Computational and HPC access and storage
  - Funding

Thank You