



# Data Science is for Everyone

Two-Year College Data Science Summit, May 10<sup>th</sup> 2018

**Sallie Keller**  
**Professor of Statistics and Director**



**BIOCOMPLEXITY INSTITUTE**  
VIRGINIA TECH.



**SDAL** SOCIAL &  
DECISION ANALYTICS  
LABORATORY

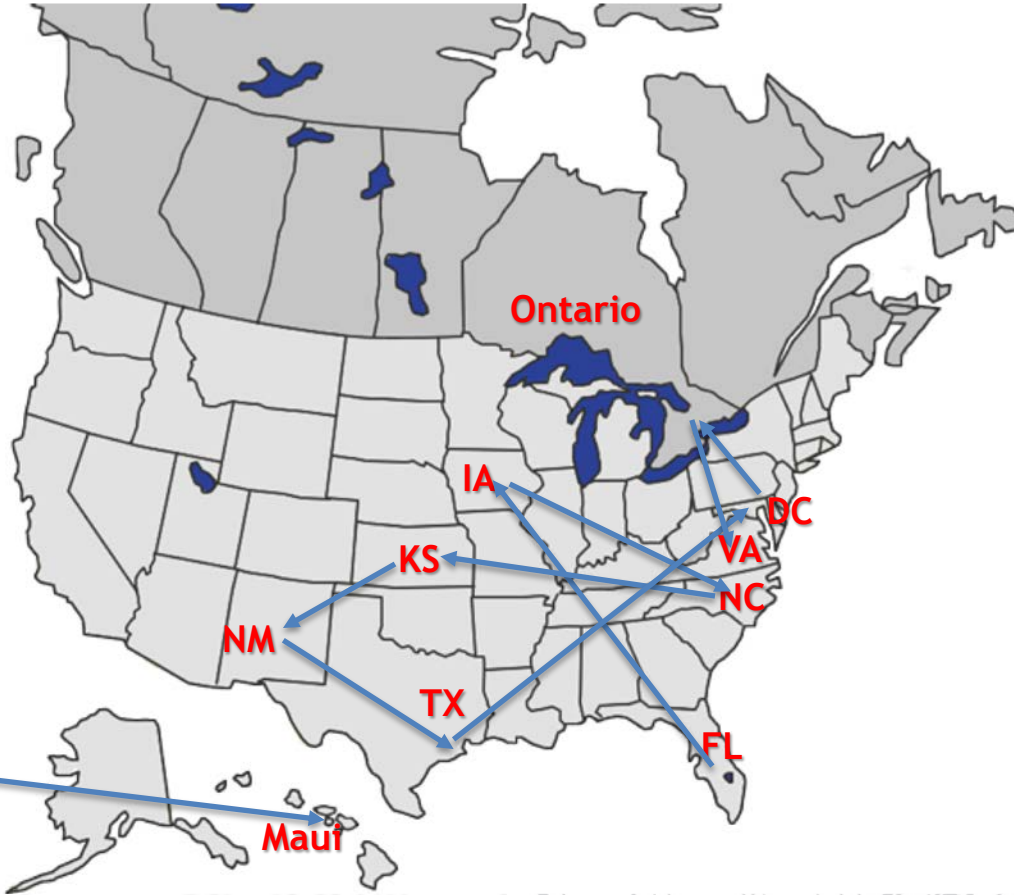
# College



*St. Petersburg Junior College*



# Then what?



# Biocomplexity Institute

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

Understanding biology **in the context of** ecosystems and human-created systems

**Transdisciplinary** team science

**“From molecules to policy”**



## Problem-Driven Science

---

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

# Social and Decision Analytics Lab

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

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

# We are in an *ALL* Data Revolution

## A new lens for social observing

### Infrastructure



- Condition
- Operations
- Resilience
- Sustainability

### Environment



- Climate
- Pollution
- Noise
- Flora/ Fauna

### People

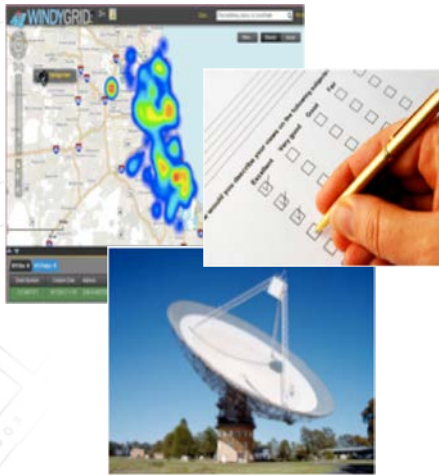


- Relationships
- Location
- Economic Condition
- Communication
- Health

# It is time to leverage *ALL* the data sources

*Local, State/Providence, and Federal*

## Designed Data



## Administrative Data



## Opportunity Data



## Procedural Data

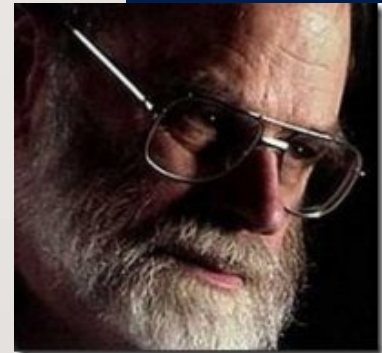
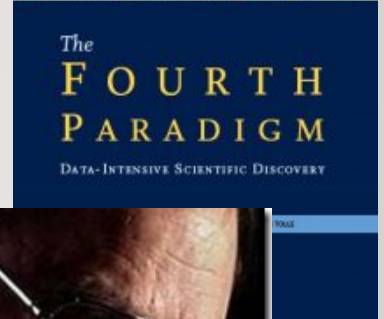


# WHAT IS DATA SCIENCE?

---

## Fourth paradigm

“... change of all sciences moving from observational, to theoretical, to computational and now to the 4th Paradigm - Data-Intensive Scientific Discovery”



Jim Gray



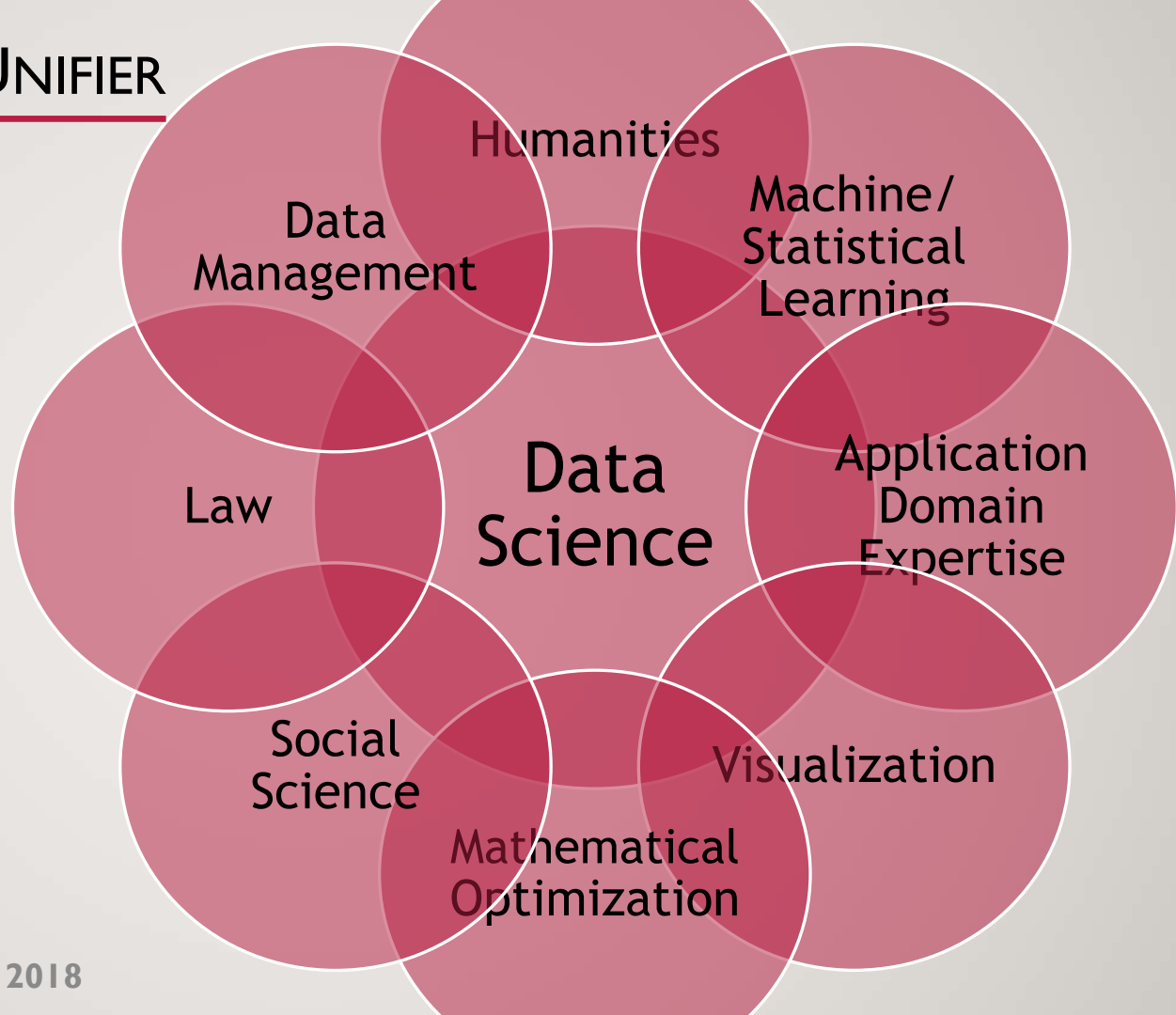
# WHAT IS IMPORTANT?

---

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

# DATA SCIENCE AS A UNIFIER

---



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



Local / State Government  
Federal Statistical Agencies  
Department of Defense  
Industry

**NCSES** National Center for Science and Engineering Statistics

**MITRE**



# Enhancing Prosperity through Data Science



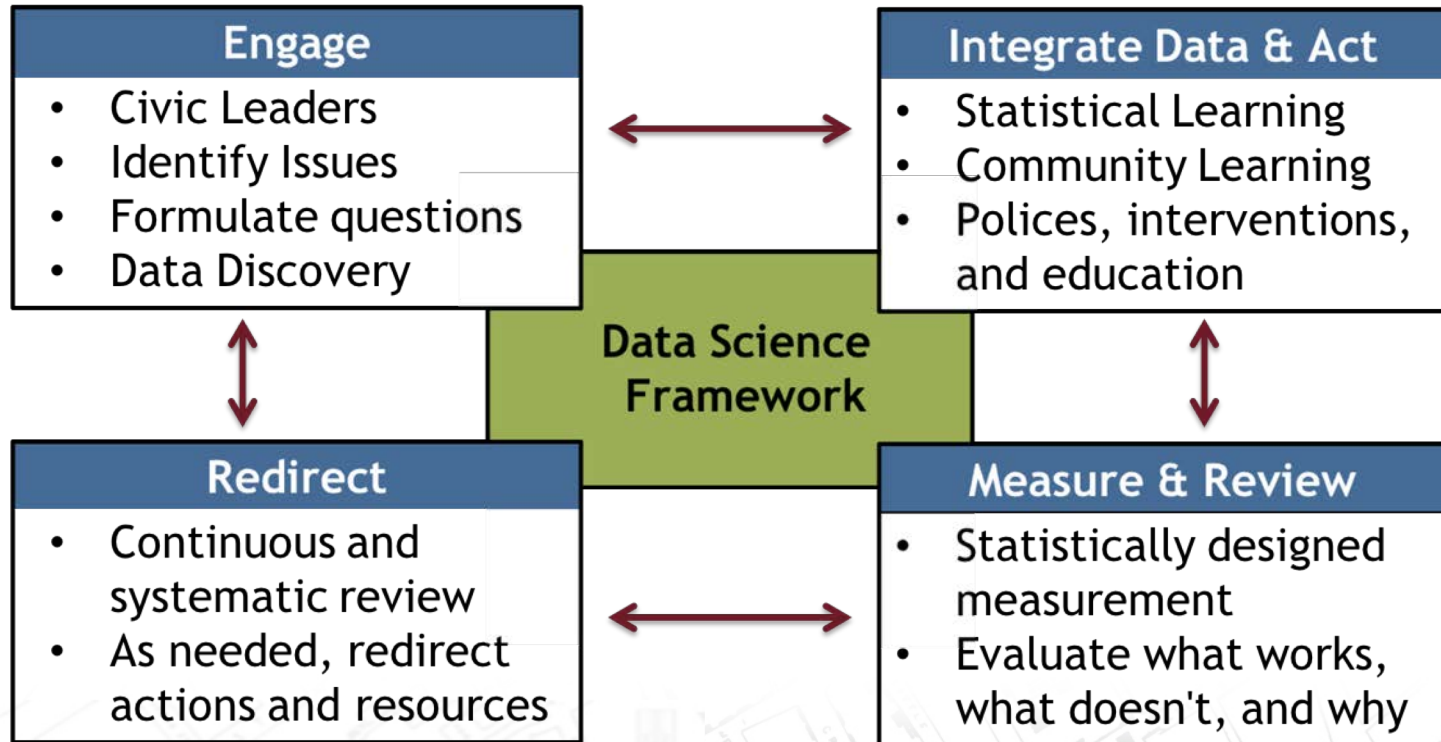
**IOWA STATE UNIVERSITY**  
OF SCIENCE AND TECHNOLOGY

# How have Virginia Tech and Iowa State come together to develop this vision?

- *Big data* initiatives
- Experience with *data science community based research*
- Practicing *engaged scholarship*
- Commitment from leadership to steward *collaborative processes* going forward



# CLD3 – Community learning through data-driven discovery



# Common CLD3 themes need data science training to address

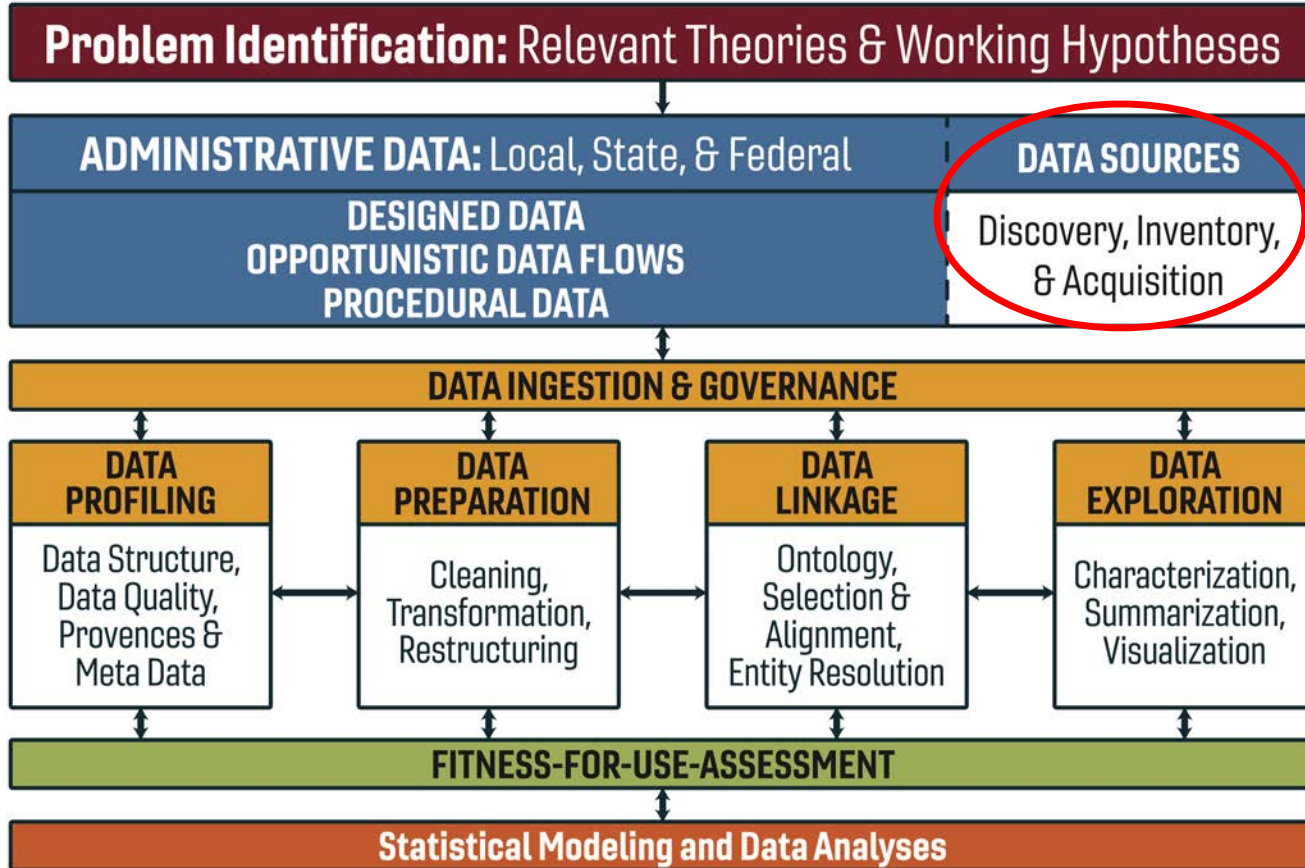
- **Locating** and **describing** a population within a community
- **Estimating** a statistical summary and its margin of error to evaluate its usefulness **for the purpose at hand**
- **Forecasting** future needs
- **Evaluating** a program, policy, or standard operating procedure

# CLD3 provides inputs to data science training





# Data Science Framework



# Local community Data Map

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

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

Neighborhood

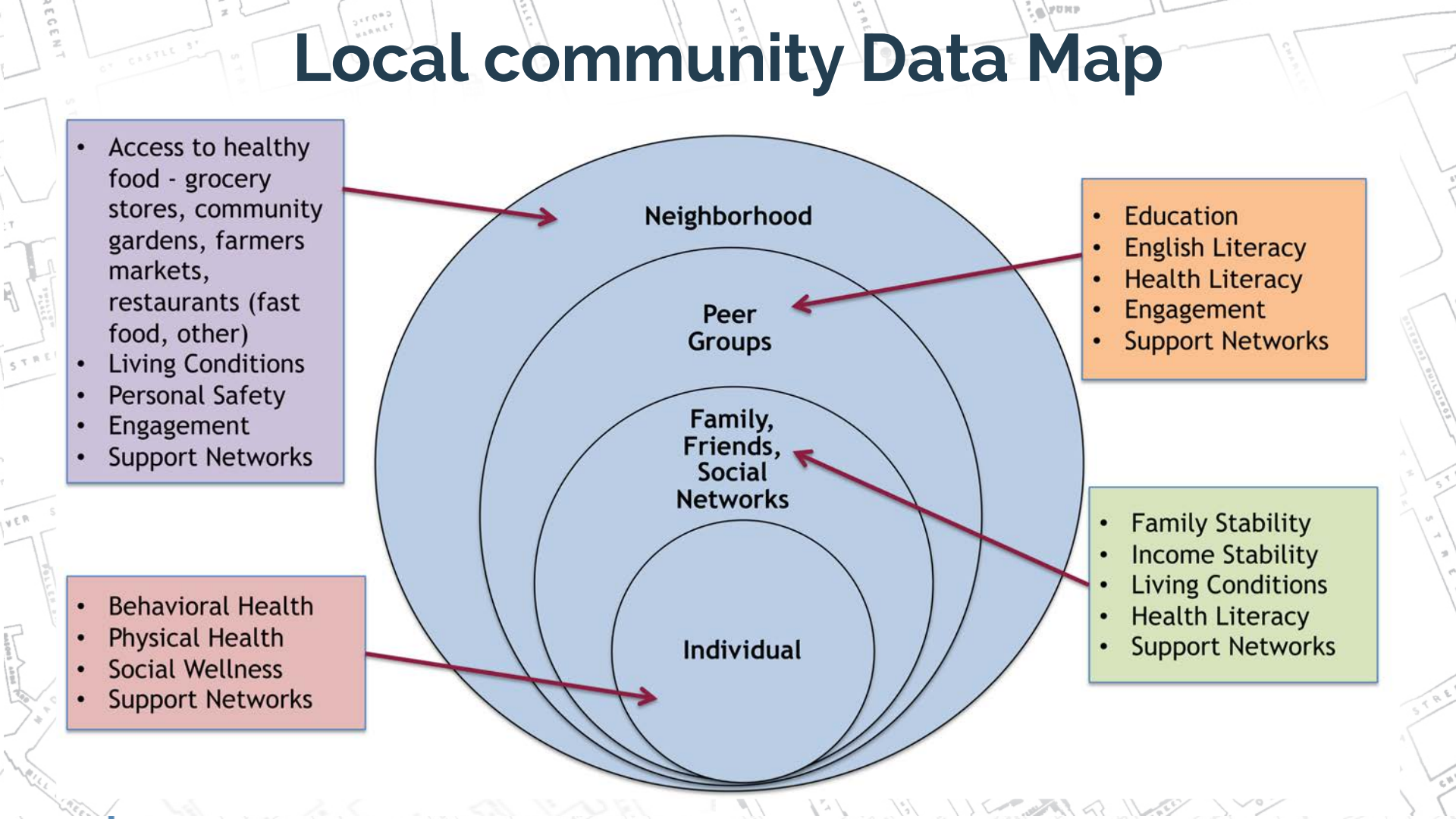
Peer Groups

Family, Friends, Social Networks

Individual

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

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

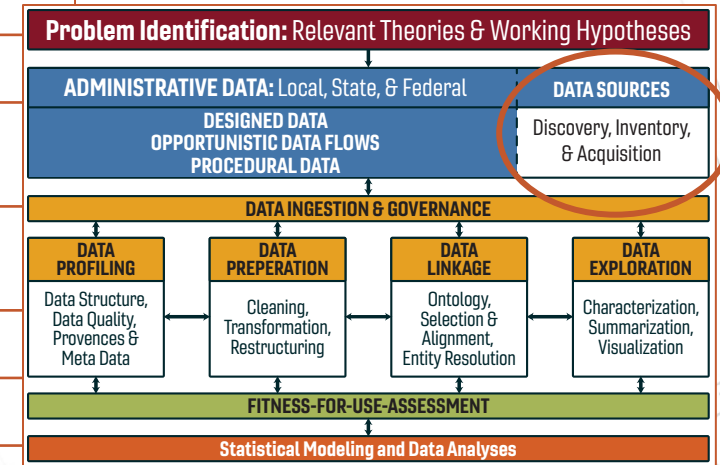


# Data Discovery, Inventory & Acquisition

Data Source	Geography
American Community Survey data (Census), 2011-2015 ( <b>updating now to 2012-2016</b> )	Census Tracts and Block Groups
American Time Use Survey (BLS), 2017	National
Youth Risk Behavior Surveillance System, 2015	State
County Health Rankings, 2017	County
Built Environment, e.g., Grocery stores, SNAP retailers, recreation centers, community gardens	Address Level
Fairfax real estate tax assessment data	Address Level
Fairfax Open data: Zoning, Environment, water, Parks, Roads	Shapefiles
Fairfax County Youth Survey, 2016 8 <sup>th</sup> , 10 <sup>th</sup> , 12 <sup>th</sup> 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

High School

Postsecondary Education

Credentials and Skill-based Training

Work Experience & STEM Occupations

Formal Education

Credentials & Skill-based Training

Job Postings & Resumes

CollegeBoard

IPEDS  
Integrated Postsecondary Education Data System

HarvardX

o.net

DATA AT WORK  
THE UNIVERSITY OF CHICAGO

MONSTER

burningglass®  
TECHNOLOGIES

VIRGINIA DEPARTMENT OF EDUCATION

SCHEN

coursera

MITx

opendata  
.cs.vt.edu

Virginia Workforce Connection

IPUMS  
USA

Credential Engine™  
Moving Credentialing Forward

CAREERBUILDER®

indeed

AMERICAN COMMUNITY SURVEY  
U.S. CENSUS BUREAU

Community

County Health Rankings & Roadmaps  
Building a Culture of Health, County by County

HOKIES 4 HIRE  
Career and Professional Development

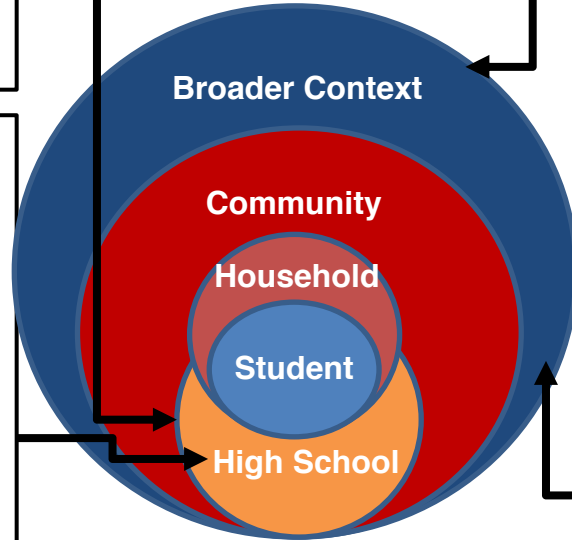
# Data Map

## High School Student Body Characteristics

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

## High School “Postsecondary-Going” Culture

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



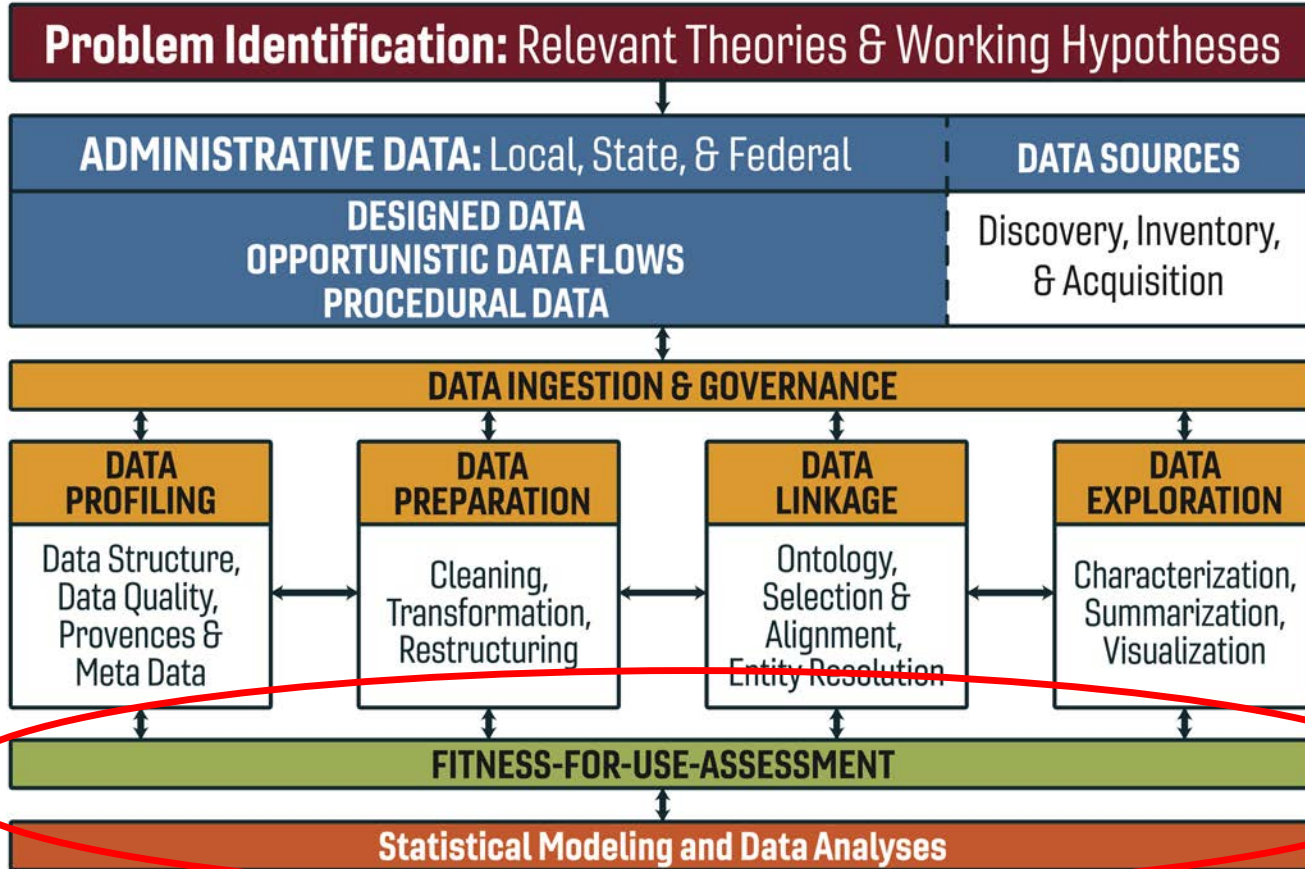
## Community Characteristics

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

## Perception of Postsecondary Availability

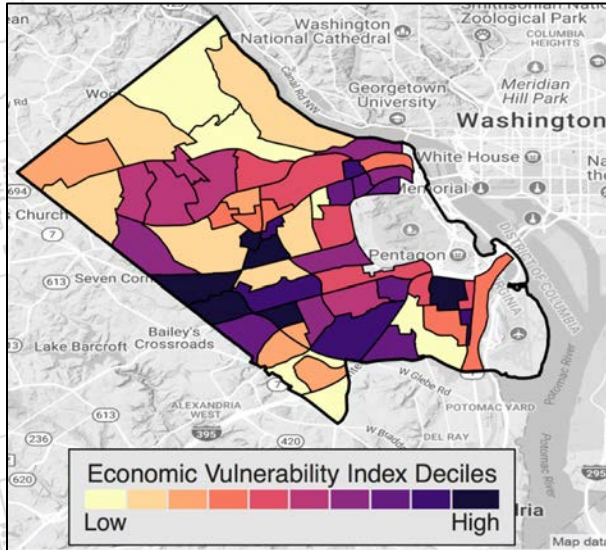
- Number of vocational schools, colleges, and universities in geographic area (IPEDS)
- Cost (tuition, fees, room and board, financial aid) of colleges in geographic area (IPEDS)
- Acceptance rate/college selectivity of colleges (IPEDS/SCHEV)
- College “choice set” of peers (SCHEV)
- College enrollment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

# Data Science Framework

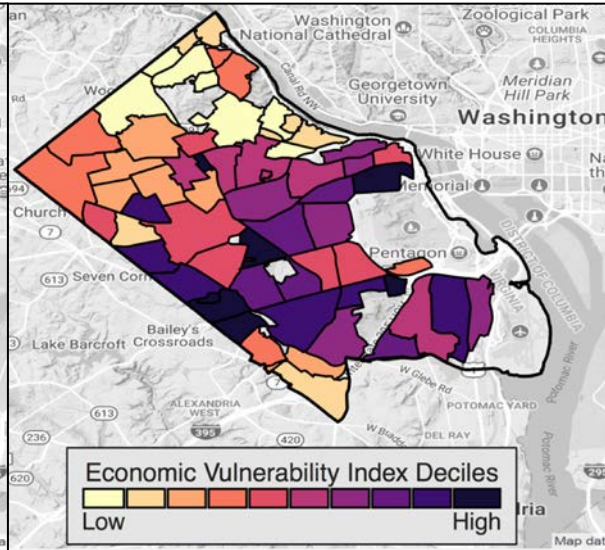


# Arlington County Vulnerability Indicators

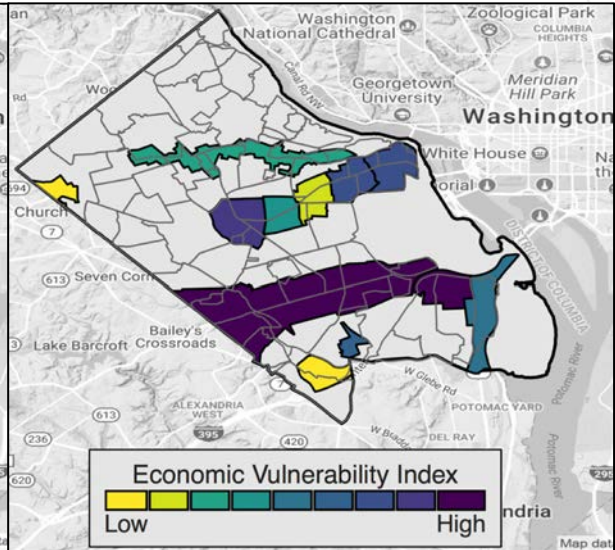
## Census Tracts



## Civic Association Neighborhoods



## High-Density Planning Regions

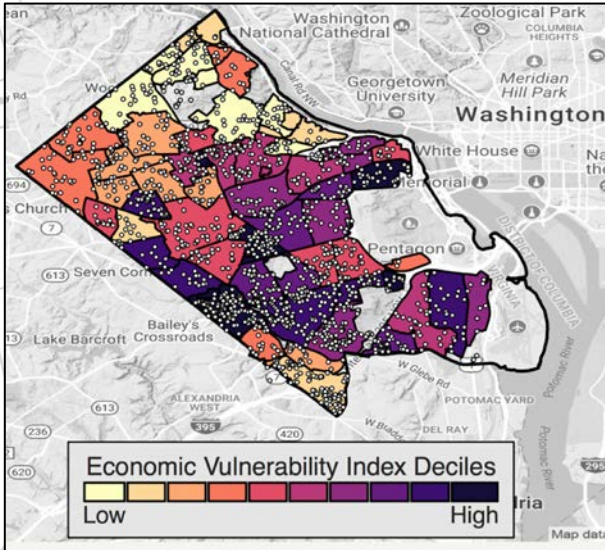


Based on a **statistical combination** of the percentage of Households with:

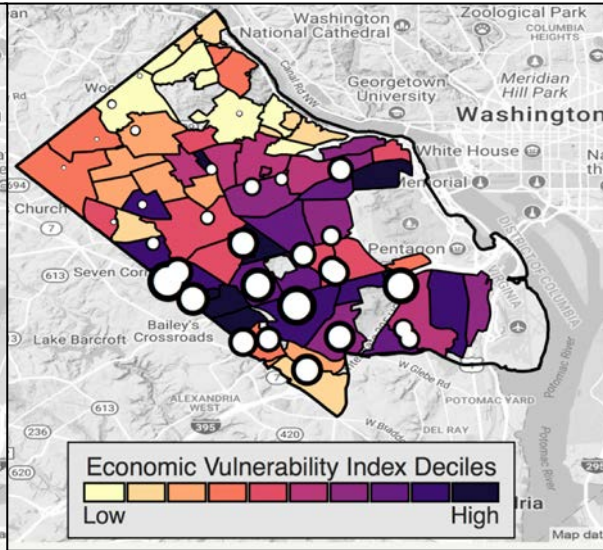
- housing burdens > 50% of Household income
- no vehicle
- receiving Supplemental Nutrition Assistance Program (SNAP)
- in poverty

# Arlington County Neighborhood Insights

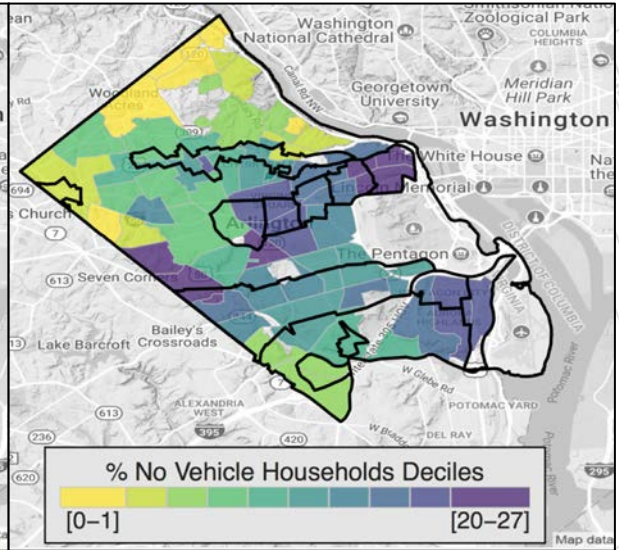
Households receiving subsidies from Department of Parks and Recreation



School and neighborhood vulnerability indices



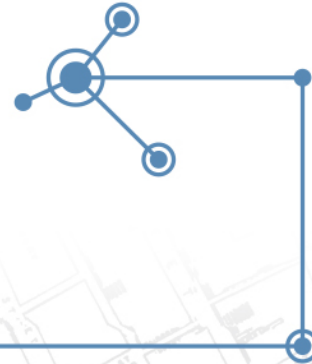
High-Density Planning Regions with % households with no vehicles



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



# Building Data Science Capacity



# Engage partners that can help scale the benefits of data science

—COOPERATIVE—  
EXTENSION

*Extending Knowledge, Changing Lives*

- Mission of *Land Grant Universities* is to enhance the security and *social well-being of its communities*
- Cooperative Extension professionals know how to involve university researchers in *community based research through engaged scholarship*

# Data Science for the Public Good (DSPG)

## Build momentum through purposeful experiential learning



**IDENTIFYING STEM EDUCATION PATHWAYS**  
Sponsor: Fali Sogutoglu, 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 Hoffmann, Fairfax Health and Human Services

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

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

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

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

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

**CREATING SYNTHETIC DATA FOR VIRGINIA LONGITUDINAL DATA SYSTEM**  
Seun Doh, Kyle Morgan, Ronnie Ferizo, and Lata Kodali (Virginia Tech) with Ash  
 Sponsor: Todd Maxia (SCHEV - State Council for Higher Education in Virginia)

**DEFINING AND MEASURING EQUITY IN ALEXANDRIA, VA**  
Sponsor: Emily Williams (City of Alexandria)

**PROFILING ARMY BASES**  
Cable, identify publicly available data sources (e.g., Census and USGS) data to create social, demographic, economic, and other quantitative profiles of Army bases and the surrounding areas. Identify relevant variables for use in statistical models.  
 Sponsor: Greg Davis, Andrew Slaughter, US Army Research Institute for Behavioral & Social Science Research

**DISCOVERING NON-TRADITIONAL DATA SOURCES FOR BUSINESS INNOVATION**  
Hamsirat Pancher (VT), David Park (VT), Daniel Wilkin (VT), Joseph Kim (VT)  
 Claire Koling (PSU) with Goren Kozmar and Stephanie Shipp (GDAL)  
 Sponsor: Gary Anderson, The National Center for Science & Engineering Statistics

**A STUDY ON WMATA BUS FARE EVASION**  
Sponsor: Jayna M. Johnson, Catherine Vanderveert  
 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 Bianca Pires (GDAL)

**Thank You**

