Building Capacity for Data Driven Governance
Creating a New Foundation for Democracy

Sallie Keller, sallie41@vt.edu

Vicki Lancaster, val20003@vt.edu

Stephanie Shipp, steph19@vt.edu

Social and Decision Analytics Laboratory,
Biocomplexity Institute of Virginia Tech
Arlington, Virginia 22203

Submitted to Statistics and Public Policy Journal
March 2017

Building Capacity for Data Driven Governance
Creating a New Foundation for Democracy

Abstract

Existing data flows at the local level, public and administrative records, geospatial data, social media, surveys, as well as other federal, state, and local databases, are ubiquitous in our everyday life. These data, when integrated, can tell the story of a community. The Community Learning Data Driven Discovery (CLD3) process liberates, integrates and makes these data available to government leaders and researchers to tell their community's story and to use these stories to build an equitable and sustainable social transformation within and across communities to address their most pressing needs.

The CLD3 process starts with asking local leaders what their questions are but cannot currently answer; identifying data sources that can provide insights; wrangling the data (profiling, cleaning, transforming, linking); using statistical and geospatial learning along with the communities' collective knowledge to inform policy decisions; and developing, deploying, and evaluating intervention strategies based on scientifically based principles. CLD3 is a continuous, sustainable and controlled feedback loop. CLD3 is described conceptually and through examples as a process that builds capacity for data driven governance at the local level.

Keywords: Statistical Learning, Geospatial Learning, Administrative Data, Evaluation, Vulnerable Populations

1. INTRODUCTION

Vast amount of data, generated through almost every aspect of living, offer an unprecedented opportunity to improve the health, well-being, and quality of life of our communities. Smart cities, urban analytics, civic analytics, urban informatics, all expressions in today's lexicon that hold the promise the data revolution will democratize data, but has it? The movement to democratize data, to break down the data silos, and to provide access to all people, has been around since at least the 1990s (Sawicki and Craig 1996), However, the capacity at the local community level to store, manage, extract knowledge, and gain insights from data, is still one of our country's great challenges (National Research Council 2013).

Communities, defined geographically or socially, are a complex system of interdependent networks. Policy interventions in one area can ripple through a community with unintended consequences and interactions throughout the network (Pollack 2016). By working across community programs and government agencies, solutions to these complex problems may become more tractable. For example, communities like Arlington, Virginia are concerned about at risk youth. The police department wants to be proactive and reach them before they go down the wrong path, the school system wants them to succeed academically, the libraries want to nurture their reading skills, and the parks and recreation department wants to involve them in activities that will lead to a healthy lifestyle.

Traditionally each department approaches common issues and goals in data isolation. We argue that what is missing in tackling these persistent problems is a defined process that can use the vast amounts of data generated across government departments and programs, nonprofits, businesses, and its citizens to monitor the complex system of networks in a community, and evaluate interventions and interactions. To meet this challenge, our Social and Decision Analytics Laboratory

has developed a novel framework for **community learning** through data driven discovery that enables evidence-driven policy interventions and evaluations.

The cycle of community learning begins with local leaders defining critical questions and issues facing their community; identifying the data sources that provide insights; wrangling the data sources (profiling, cleaning, transforming, linking); using statistical and geospatial learning along with the communities' collective knowledge to inform policy decisions; and developing, deploying, and evaluating intervention strategies based on scientific experimental design principles. Community learning is a **continuous sustainable and adaptive feedback loop**.

We call this community learning cycle, the Community Learning Data Drive Discovery (CLD3) process. It borrows from the features of the Learning Health Systems (JASON 2014) and the systems science approach in public health (Sterman 2006). The learning healthcare system extends the opportunity for discovery throughout the lifecycle of an intervention at the level of the individual; it seeks to determine which intervention works best, for whom, and under what circumstances. The CLD3 process expands this paradigm from the individual to the community level. The community learning cycle is realized from ***direct input*** through interactions with the community leadership and through ***existing data sources*** such as local administrative data, national repositories, and common data frameworks, (e.g., human services, education, and emergency management records; American Community Survey (ACS) and American Housing Survey; and county health rankings (University of Wisconsin Population Health Institute 2017) and evidence for action metrics (National Academies of Sciences 2016), and through ***indirect input*** from the natural opportunistic data flows that occur in a community (e.g., social media, blogs, community websites, and local news).

Community leaders can come from all stakeholder groups, however our initial research focuses on civil servants in local government. Local governments collect massive amounts of data every day to provide services and to plan for the future. Yet, most do not use their own administrative data to inform their decision-making. The CLD3 process can build capacity for data driven governance in these communities by making use of their local administrative data and combining it with state and federal statistical data, social media, and other sources of data to inform a community's questions.

In the sections that follow, the CLD3 process is described and put in context using our collaboration with Arlington County, Virginia. The roles for community learning, statistical learning, and evidence-based policy development are highlighted.

2. DATA TYPES, MODELS, AND STATISTICAL RIGOR

Data is the new language to communicate within and among departments and offices at all levels of government. But what is meant by data is often not well defined. For the purposes of facilitating collaboration and clarity in developing and deploying the CLD3 process, definitions of "data" are categorized in four bins (Keller et al. 2016 and Keller and Shipp 2017):

- **Designed Data** are generated in the pursuit of scientific discovery. Designed data include statistically designed data collections, such as data generated from surveys, experimental designs, registries, and intentional observational collections.

- **Administrative Data,** also referred to as "business practice" data, are collected for the administration of an organization, program, or service process. These data provide an opportunity for gathering information that exists due to normal economic and social activity. Examples of administrative data include Internal Revenue Service data for individuals and businesses, Social Security earnings records, patent and trademark databases, Medicare and Medicaid health utilization data, banking and other financial data, production processes, taxi

trip data, and local data generated from 911 calls and Emergency Management Services (EMS) responses, property assessment and tax data, and data from health and human services, parks & recreation, libraries, and environmental services (e.g. trash & recycling, water & utilities, projects & planning, transportation, and building & permits).

- **Opportunity Data** are data generated on an ongoing basis as society moves through its daily paces. Opportunity data is derived from a variety of sources such as GPS systems and embedded sensors, social media exchanges, mobile and wearable devices, and Internet entries. Captured through a variety of methods including direct flows, Internet searches, web crawling and scraping, these data may exist in a variety of electronic and physical modalities.
- **Procedural Data** are data derived from policies, procedures, and legal requirements; they are the rules and regulations that govern and shape our lives. These policies and procedures affect our work, our personal lives, and society. Examples of procedural data include compensation policies, the Affordable Care Act, the Department of Defense policy 'Don't Ask, Don't Tell', and Supreme Court rulings.

Although these data sources are ubiquitous in our daily lives, most are <u>not</u> collected with the intention of statistical analysis. For use in a community-learning paradigm, these data sources must be repurposed for statistical learning. This requires a disciplined yet flexible and adaptable data framework to assess data quality and fitness-for-use that is dependent on the intended use of the data (Keller et al. 2016).

Frequently, local problems are being tackled with federal data at an aggregated scale, ignoring the heterogeneity in the smaller spatial scales (e.g., Humes, et al. 2011; Weinberg 2011). In particular, the integration of all data (local, state, and federal administrative data, designed data, procedural data,

and opportunity data), which is necessary for small-scale estimation, is simply absent. For example, in federal, state, and local data portals, data sources are made available but there is no attempt to integrate these sources. Take for example the new GSA Technology Transformation Service, the U.S. Data Federation (2017) provides access to data sources from the federal, state, and local levels (the local data sources are very limited), but does not provide the technology to integrate these sources. Instead, each user must tackle the challenge of integrating these data sources themselves.

Through the development of the of the CLD3 process, we found many organizations using a variety of data sources to support evidence-based policy development. Their uses range from descriptive displays such as dashboards, (e.g., Robert Wood Johnson Foundation 2017), to randomized control trials, (e.g., Abdul Latif Jameel Poverty Action Lab 2017), to draw inferences. Randomized control studies can provide statistical evidence as to whether or not a program works in a specific context, but are unable to explain why (Behn 2015, Deaton et al. 2016). Randomized control studies are valuable but insufficient to explain specific social phenomena. To do this requires an understanding of the context and human interactions (Flyvbjerg 2006).

Statistical reasoning also seems to be absent in many evidence-based applications (National Academies 2017). For example, absent are descriptive statistics that take into account the representativeness of the data or estimate the variability of the statistic; intervention evaluations that acknowledge the interactions (correlations) within the system; models that account for the hierarchical and spatial nature of the data; and experimental designs that account for confounding factors (National Research Council, 2103 and Banerjee et.al 2016).

To address these issues, our research uses statistical learning and develops geospatial indicators to analyze complex social systems. The CLD3 process starts with research questions, discovering data sources to answer these questions, conducting exploratory analysis, and creating statistical models.

The community learning process is designed to enable such research to quantify and understand the association of geospatial indicators pre- and post-policy intervention over time to accurately interpret the impact of an intervention. This involves employing statistical designs that incorporate or account for the myriad of confounders in a complex system. Community learning allows changes to policy and the ability to measure the impact of these changes through statistical analysis.

3. COMMUNITY LEARNING DATA DRIVEN DISCOVERY (CLD3)

Rural and metropolitan communities are fast becoming the engines of social transformation and economic prosperity; they are experimenting, taking risks, and making the hard choices required to address disparities in healthcare and education, and in economic, social, and criminal justice. As a process, CLD3 liberates and integrates existing designed, administrative, opportunity, and procedural data sources and makes them available to civic leaders and researchers to address their most pressing issues. The CLD3 process also provides the tools for rural and urban communities to leverage their assets by incorporating feedback from the community's collective knowledge. The CLD3 is a holistic approach that includes the following steps.

1. *Engages* the community by bringing local government agency and program administrators, civil servants, and community leaders together to breakdown the boundaries among governmental bureaucracies and programs through workshops, forums, and meetings that identify issues and the data sources that can be used to describe and monitor them.
2. *Integrates* data sources to gain usable insights from statistical learning, combines these insights with the knowledge and insights of local government agencies and program leaders to plan and implement policies and interventions to meet the goals of the community within their contextual constraints.

3. *Measures* and *reviews* the intervention pathways to gain a better understanding of what works, what doesn't, and why, by understanding the connections across issues, the timing of problems, and how the data sources fit together.

4. *Redirects* interventions when warranted based on the information from a continuous systematic review.

This CLD3 process is illustrated in **Figure 1**. Data becomes the new language to communicate within and among organizations, departments and programs at all levels of government. Statistics is the tool used to translate the data into a story and statistical learning is the technical process. Statistical learning is composed of methods and techniques for modeling and analyzing complex datasets (James, et al. 2014). It starts with exploratory data analysis methods that are used to isolate patterns and relationships, uncover unexpected behavior, confirm or disprove assumptions, develop explanations and generate hypotheses.

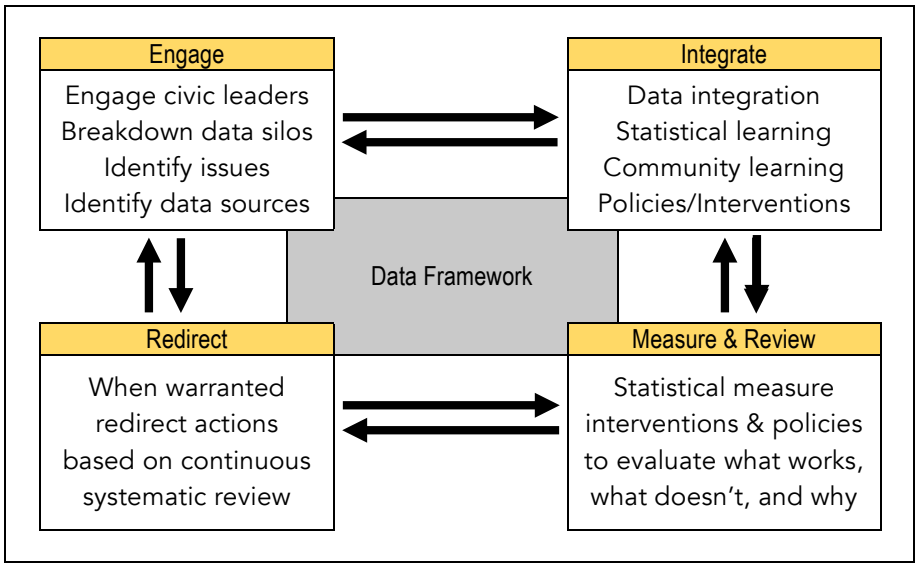


Figure 1. Community Learning Data Driven Discovery (CLD3) Process

Description: The CLD3 process involves four broad steps that are connected through the Data Framework (see Figure 2). Each step is described in more detail in each box – engage with civic leaders, integrate data, measure and explore using local administrative data flows combined with other data sources, and implement and continuously review program and policy changes to redirect resources as needed.

The democratization of data can contribute to untangling the innumerable causes of social problems through statistical learning. Exploratory data analysis helps to identify how rich an inferential model the data can support or if the data can support inferential analyses at all. No amount of data can supplant the necessity for data analysis grounded in statistical theory to develop, test, and accumulate the causal explanations that can guide policy. Whether the statistical analyses can guide policy depends on whether the story told by the data can be interpreted through the lens of the community and community learning. Does the story provide the relevant information and insights community leaders need to address the issue? Examples of these issues include:

- evaluating an existing program or policy in an effort to make it more resource efficient and cost effective,

- identifying vulnerable populations to assess their needs and provide services, and
- validating and improving the accuracy and precision of estimates calculated from local administrative data that can drive federal funding.

At the heart of this CLD3 process is our data framework **(Figure 2),** which encapsulates a rigorous and flexible approach for re-purposing data, from discovery to analysis to inference (Keller et al. 2016). Statisticians will agree that often the majority of their time is spent preparing the data for analysis. This is especially true when integrating disparate data sources in a community learning context. Even though data quality judgments depend on the data consumer and use of the data, a framework is needed that adheres to scientific principles and provides a systematic approach to use and translate data to a story that is based on statistical and behavioral science principles (Keller et al. 2016). Our data framework provides a disciplined process of identifying data sources, preparing them for use, and then assessing the value of these sources for the intended use(s). When working with local governments in the context of building capacity for data-drive governance, such rigor is necessary particularly for communication across departments and program, transferability, and scalability.

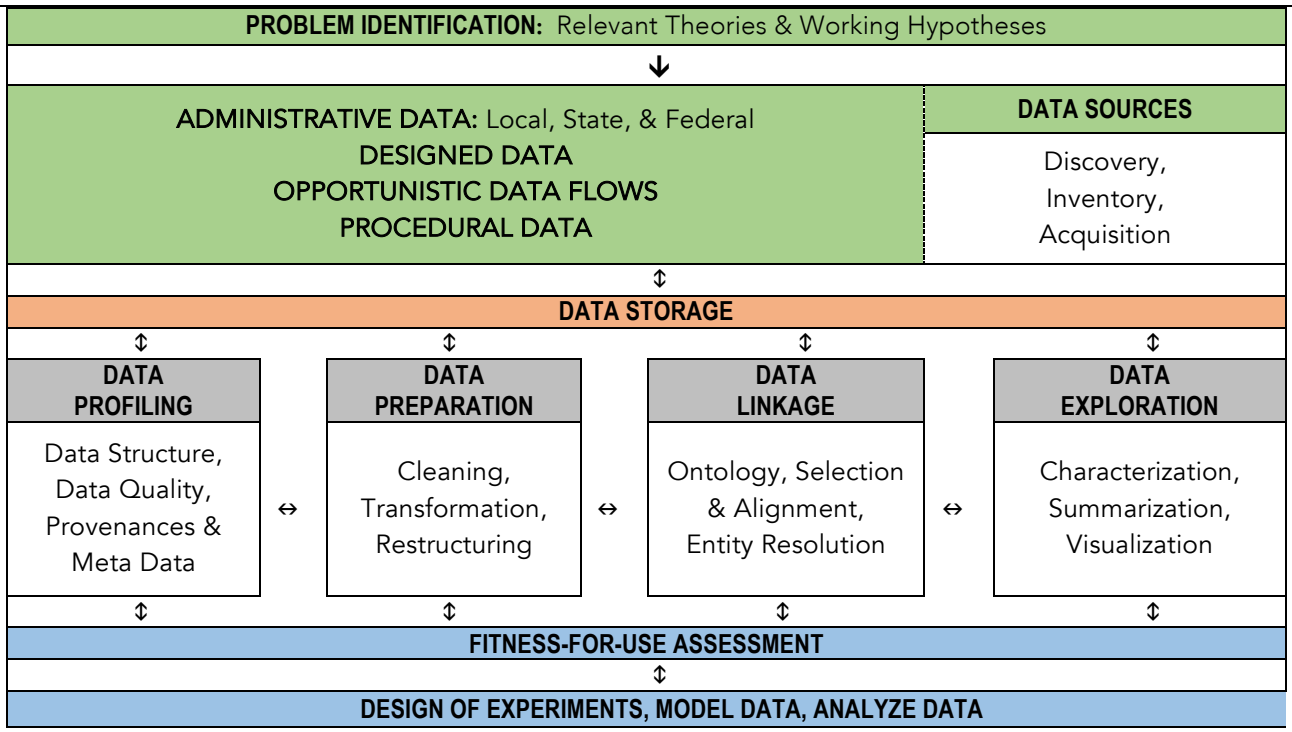


Figure 2. SDAL Data Framework

Description: The SDAL Data Framework provides a rigorous and flexible approach to repurpose local administrative and other data flows, starting with the identification of the problem; discovery and acquisition of data sources; profiling of the data to assess quality; preparing, cleaning, linking, and conducting exploratory analysis. The final steps are examining fitness-for-use in the context of the topic under study; the design of experiments and models; and synthesizing findings and analyzing results.

- Data discovery - the identification of potential data sources that could be related to the specific topic of interest
- Data inventory - the method used to screen and inventory the data sources to determine their value in supporting the research question and if they would be worthwhile to acquire
- Data acquisition – the process of negotiating and acquiring the data, and managing legal, privacy, security, and confidentiality practices
- Data profiling - a determination of both the quality of the data, provenance, and its utility to the project at hand
- Data preparation – the process of cleaning and readying the data for analysis; what is referred to as "wrangling the data"
- Data linkage – the process of building links to ensure compatible meaning, schemas, and ontology for data from multiple sources, resulting from the repurposing of the data
- Data exploration - the analysis of the datasets by summarizing main characteristics, often with visual methods
- Fitness-for-use assessment – the characterization of the information content in the results as a function of the analysis model, e.g., data quality and data coverage (representativeness)
- Design of experiments, modeling and analysis – exploring new uses of the repurposed data to support insights; data driven hypothesis development, and policy; and creating intervention and evaluation strategies

4. COMMUNITY LEARNING TO STATISTICAL LEARNING – A PROOF-OF-CONCEPT

The CLD3 process steps outlined in Figure 1 are described next in the context of a proof-of-concept with examples from our work with Arlington County, Virginia. Arlington County is located in Northern Virginia across the Potomac River from Washington, D.C. It's the geographically smallest self-governing county in the U.S., occupying slightly less than 26 square miles (Arlington County 2017). The Arlington legislative body is composed of five members elected at large. The County Board appoints the County Manager and a variety of citizen boards, commissions and advisory groups to help develop and implement County policies. It is home to many federal employees and federal government headquarters, such as the Department of Defense at the Pentagon, National Science Foundation, and the Transportation Security Administration. Arlington County is a diverse population of approximately 225,000 residents with 16% Hispanic or Latino, 10% Asian, and 9% Black or African American; about 27% of Arlington residents speak a language other than English at home (U.S. Census Bureau 2011-2015 American Community Survey 5-Year Estimates).

4.1 Engage with the community

Engagement in the CLD3 process can start in many ways. One approach is to do this through a forum across government departments and offices, or other stakeholder groups, to help community leaders identify and define those issues that cannot currently be answered and to identify the data sources that can provide insights into these issues. This is similar to the collective impact paradigm (Kania and Kramer 2011) that stresses an inclusive and structured approach for identifying and

solving community problems. CLD3 extends the paradigm by adding statistical and social science practices grounded in the scientific method.

An example is our engagement with Arlington County, Virginia. It started with a conversation with local government leaders - the Chief Information Officer, the Fire Chief, Police Chief, and the heads of Economic Development and Planning, Environmental Services, Emergency Management, Community Planning, Housing, & Development, Human Resources, Human Services, Parks & Recreation, and Technology Services. In that meeting, the Fire Chief presented us with the challenge of linking together all of the data collected during the course of an emergency to recreate incidents end-to-end through the data. The Chief Information Officer agreed to work with the Fire Chief to ensure access to data and develop and initial Memorandum of Understanding for data sharing. We acted quickly to gain momentum and developed a first integrated analysis. More details are given in the next section.

The lesson learned in that first meeting was that we had taken it for granted that government leaders and policy makers understood the power their data could bring to actually guide decision making. This should not have been a surprise given how often statisticians have made this same comment in nearly every other application domain (National Research Council 2007). It did, however, lead us to develop a more formal engagement activity.

Working with the Chief Information Officer and Deputy County Manger, we developed a strategy to introduce CLD3 to the county's executive management leadership team. We collaboratively organized a Data Discovery Workshop. This was created to both identify community issues and the data sources that could be used to describe the current situation as a baseline and then to implement and monitor new interventions and policy changes.

The workshop provided a forum for government leaders across departments and offices to envision telling the story of their community through the language of their own program data by combining their data with data sources from other departments. The workshop was structured to breakdown the silos across governmental bureaucracies through a dialogue process to identify issues and data sources that span their boundaries. Our experience confirms that a single department may be reticent to tackle an issue such as childhood obesity due to the complexity of the associated risk factors, diet, lack of exercise, and family psychological and socioeconomic factors. But when approached from a systems perspective, departments across the government can contribute their unique resources and expertise to addressing the risk factors that align with their mission.

The Data Discovery Workshop was held with the Arlington County local government executive management team. The workshop started with a presentation that highlighted examples of integrating multiple data sources to characterize local problems and reveal possible approaches to mitigate issues. Some of the examples came directly from Arlington County and our early work with the Fire/EMS department and demographers at the Department of Community Planning, Housing & Development. Others were carefully chosen from the examples of data driven interventions in other cities that would elicit the response – if City X is doing that, then why aren't we? For example, the City of Charlotte, North Carolina partnered with the University of North Carolina-Charlotte to develop quality of life rankings based on 20 indicators measuring the social, safety, physical, and economic conditions in Charlotte's 173 neighborhood statistical areas. The indicators were used to group the neighborhoods into stable, transitioning, or challenged. The city used this information to direct resources to transitioning and challenged neighborhoods (Charlotte and Mecklenburg County 2017).

Following the overview presentation, participants were placed into small groups to discuss the local issues that keep them awake at night. No group was allowed more than one representative from a

department, program, or office. This provided the opportunity to forge new relationships. The issues that were voiced from the groups are listed in Table 1.

Table 1. Arlington County Issues and Potential Data Sources Identified at Data Discovery Workshop

Arlington County Issues	Potential Data Sources
Issue 1: How do you provide services to those with mental health issues? How do you coordinate services, anticipate needs, end the cycle of imprisonments?	Police, hospital, human services, Emergency Medical Services (EMS), homeless survey from Metropolitan Washington Council of Governments, community outreach centers, public assistance
Issue 2: Which citizens are we reaching? Which citizens are we not reaching? How can we identify and reach the underserved population?	Library, parks & recreation, citizens at council meeting minutes, community outreach centers, public assistance, civic associations
Issue 3: What kind of facilities do we need?	Surveys conducted by the various departments
Issue 4: How do you provide information to diverse populations in their neighborhoods?	Civic associations, churches, nonprofit organizations
Issue 5: Is the feedback to local government representative of the Arlington population? Who are we (local government) not hearing from, who are we not reaching?	Voting records, citizens at county board meetings as evidenced in the minutes
Issue 6: How do you consolidate systems? How do you use these systems to identify challenges that may not be apparent?	All Arlington data, starting with one agency, e.g. the Arlington Department of Human Services has 70 different vendor systems
Issue 7: Does EMS need more medic units? What characteristics are common to residents who call the 911 system three of more times in one month?	911 incident data sources, i.e., initial call report, response time, and final incident report
Issue 8: How many alerts can the Office of Emergency Management send out before people reach notification fatigue? How can more residents be reached?	Office of Emergency Management subscriber database

What became evident at the workshop is that the leaders in Arlington want to be responsive to their residents and to provide an environment in which all residents can thrive. After the Data Discovery Workshop, the county manager created a steering committee to begin to build data-driven learning

collaborations across programs and departments and to select the initial crosscutting issues to address.

The crosscutting issues proposed led to a new discussion about data sharing among departments, between Arlington and Virginia Tech, and between Arlington and data vendors that are under contract to provide systems for the collection of data and production of reports. Many of the vendor contracts do not allow access to the underlying data. This is changing as the Arlington lawyers are now revising the contracts to ensure that they own and have access to their own data.

Many sources of local data are open public records and can be used to start the CLD3 process. Our progress here is described through the examples below, many which use open data.

4.2 Integrate, Measure, Review, Redirect

Following the initial engagements with the civic leaders, some compelling and successful demonstrations of the community learning cycle are necessary to move toward the adoption of this approach. Also, a forum like the Data Discovery Workshop raises expectations among the participants so efforts should be made to meet some of these expectations to maintain interest and momentum. To meet the expectations raised in the workshop, two actions were taken. First, local government demonstrated commitment to the next steps by establishing a data steering group to guide the development of collaborative projects between government programs, departments, and our laboratory researchers. Second, we made sure that the timing of the workshop could leverage a program in our laboratory that would accelerate our ability to work on some of the issues raised by the leadership. These projects became a central focus of our Data Science for the Public Good (DSPG) summer student fellow program.

The DSPG program is an incubator to educate and train the next generation of government data scientists by exposing them to data science projects that integrate data from the municipal, state, and federal levels of government. This includes engaging students and postdoctoral fellows in the data discovery learning cycle giving them formal, hands-on, training of data discovery, access, cleaning, preparation, and exploratory analysis. Civic leaders participate in team meetings and begin to see their data and programs with a new lens. Arlington now has a full community-based research model running, and these collaborations are starting to drive the planning directions of the steering committee. This process started slowly but gained momentum as agency and program leaders heard about this work (SDAL 2016).

To achieve the full community learning cycle, the steps taken need to be demonstrated clearly and repeatedly. The three examples described below demonstrate how this can be accomplished and how community learning through statistical learning is achieved. Although the process may appear linear, it is not; it involves many steps, iterations, and deep collaboration with the stakeholders.

.

Example 1. Arlington County Fire/Emergency Medical Services (EMS)

As noted earlier, the Arlington Fire Chief posed the first question. He asked us if it was possible to link 911 incidents through the data to improve the fire department's situational awareness. The data being generated during an emergency are used to guide the Fire/EMS staff in their response to an emergency and not designed to be used in a retrospective analysis of the incident. What we discovered were silos of data, starting with the call data coming into the center, response times for engine and medic units, and final incident reports for each medic and engine unit. There was no unique identifier that could be used to link an incident across the various data sources. By linking 911 incident data sources with EMS health records and geographic data (following the processes

outlined in the data framework in **Figure 2**), the data were repurposed to address several Fire/EMS questions.

Many people call 911 when there is not a true emergency. Serving non-emergency customers is stressful for firefighters and an inefficient use of resources. By linking calls with their geographic location, it was discovered that what seemed like frequent non-emergency calls from the same individual were actually frequent calls from different individuals at the same location, such as senior assisted living centers and rehabilitation centers **(see Figure 3, top).** Data discovery and statistical learning changed the perception of the problems (through community learning) leading administrators to change their approach from an emergency response to one that focused on addressing the needs of each center through education and engaging the help of the Department of Human Services. The fire chief worked with the head of the human services department to arrange visits to these centers to identify what services were needed and to provide training on the appropriate use of 911 calls. This intervention resulted in an immediate decline in 911 calls that eventually increased, prompting new visits to the centers. The decline in 911 calls meant fewer fire and medic units on the road and a more efficient allocation of resources for both the Fire/EMS and Human Services departments.

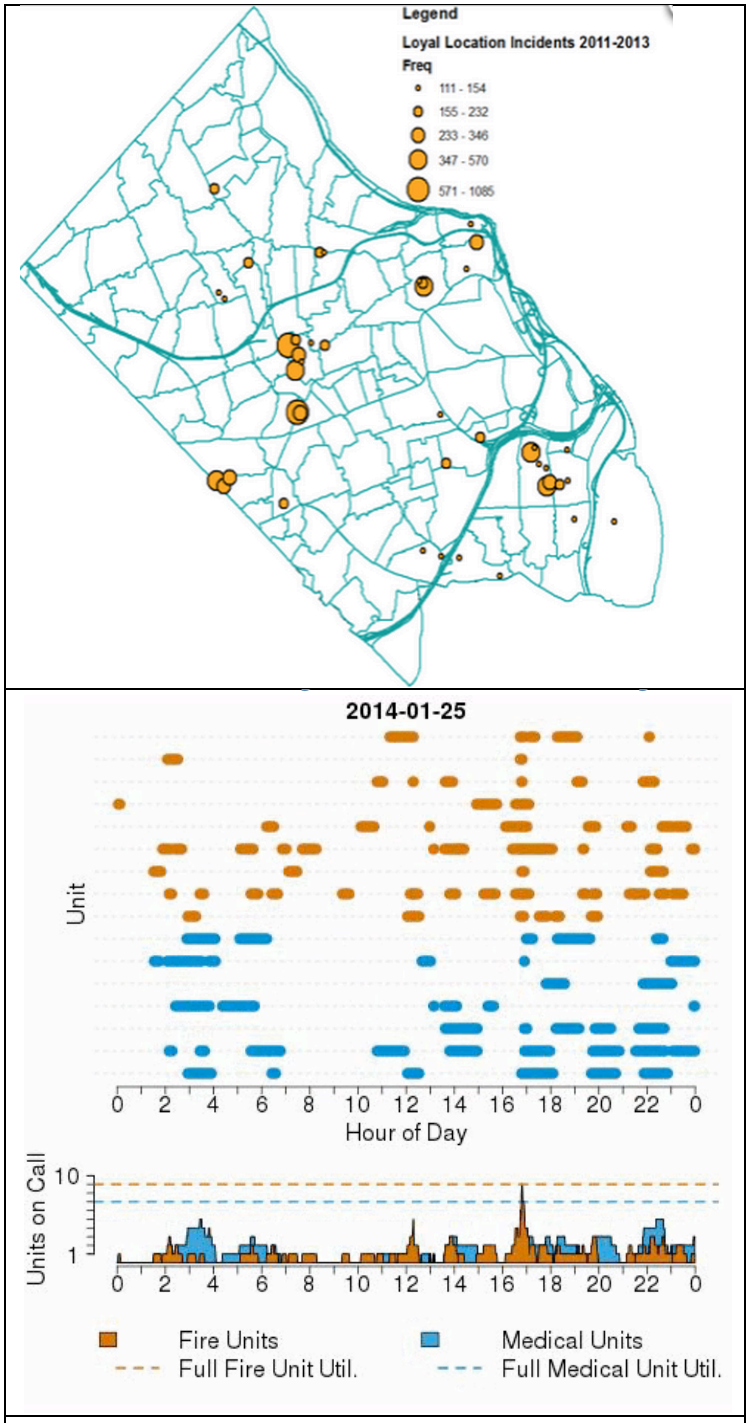


Figure 3. Improving Fire/EMS Situational Awareness

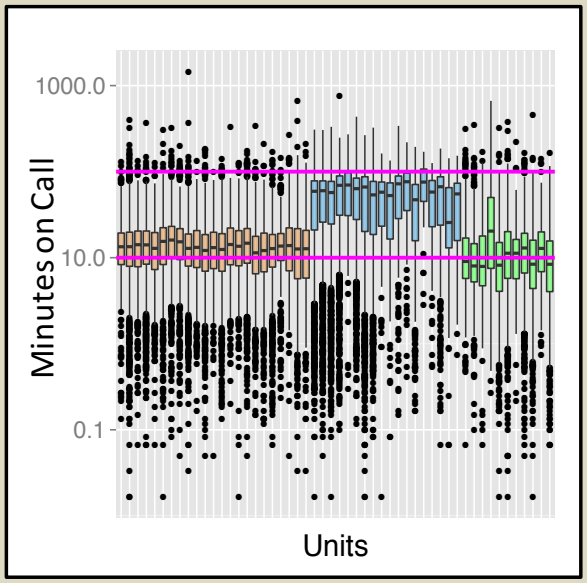
Description: Top - Locations of high non-emergency call volumes are mapped by geography. Highest volumes are from senior centers and rehabilitation centers. Bottom - Fire engine (gold) and medic units (blue) are mapped to show utilization by hour of day and day of week over a 3-year time frame. This is a snapshot for one 24-hour period.

The ability to recreate Fire/EMS incidents through the data provided situational awareness and new insights. For example, by visualizing the relationship between weather and incident volume, resources can be located during weather events to improve response times, e.g., wooded areas during wind events or low-income areas during heat events. General unit utilization over time of day and day of week can help determine if new resources are needed **(see Figure 3 bottom).** In addition, calculating the match (or lack of) between call classification and final incident classification leads to better use of resources. The statistical analyses of the linked data sources and the lessons learned are highlighted in **Figure 4** that presents the usage of engine, medic, and other units by day aggregated over three years of data.

Research Questions: The Fire/EMS Department in Arlington asked SDAL to link their 911 incident data end-to-end to improve situational awareness. The fire chief asked two questions:

1. Do they need more medic units; and
2. What are the characteristics of those who call the 911 system three or more times in one month?

Data Discovery and exploration: Fire/EMS 911 incidents are linked to geospatial locations of incidents over time (see bottom of Figure 3).



Statistical Learning: The medical (blue) and fire (gold) units for Fire/EMS 911 deployment times were calculated from the data linkage and the distributions plotted using box plots. The green unit are other command vehicles. Minutes on call are on a log scale. The Medic units were deployed for an average of 100 minutes, thus creating a potential shortage when there are large emergencies. The location and frequency of the 911 calls were mapped (Figure 3). By mapping the location of the 911 calls, it was discovered that over 80 percent of non-emergency calls are from 'loyal' locations.

Community Learning: Based on the statistical learning, the fire chief reallocated nurse practitioner resources to meet with staff at "loyal" locations to identify what non-emergency services are needed and to instruct the staff on when to call 911. By improving the fire chief's awareness of when medic units are out, where, and why, he is able to more efficiently anticipate and deploy units and resources.

Future Work: We are now monitoring the 911 calls to measure the impact of this intervention. We are looking to see whether there is a drop in call volume in the area of the "loyal" locations and a rise in non-911 services.

Figure 4. 911 Incidents – Improving Situational Awareness

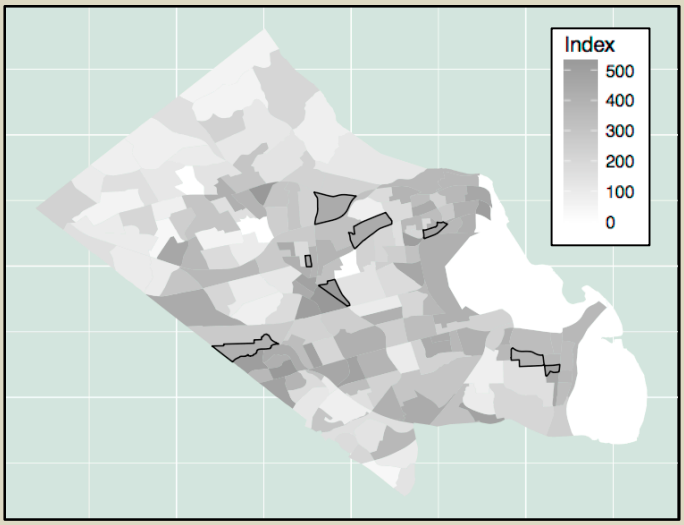
Example 2. Multi-Department Question Regarding Reaching Residents with Services

Many departments during the Data Discovery Workshop expressed the desire to understand why residents use or do not use their services. As a first step to addressing this issue, we developed a county map to identify neighborhood block groups based on their economic vulnerability. Our initial analysis developed a simple economic vulnerability index at the census block-group level using American Community Survey (ACS) data. The block-groups were ranked based on: the percentage of Households (HHs) spending more than 50% of HH income on housing, percentage of HHs with no vehicle, percentage of HHs receiving Supplemental Nutrition Assistance Program (SNAP), and percentage of HHs in poverty. These ranks were then summed to create an overall index. Albeit it is simple, this has provided some valuable insights. For example, overlaying the locations of the households receiving Department of Parks & Recreation program fee discounts, there are a disproportionate number of census block groups in the top quartile of the Vulnerability Index with less than 10 HHs receiving program fee discounts. This has presented the department with some guidance on where they could target outreach. It also begins to lay the foundation for the statistical evaluation of new policies. **Figure 5** summarizes these findings and next steps.

Arguably, many different indices can be developed. However, little research has been done to create indices at the sub-county or sub-metropolitan area geographic detail. This research is needed as local governments begin to govern based on data-driven evidence versus anecdote.

Research Questions: Who is eligible for county services and subsidies but does not using them? Why not?

Data Discovery and Exploration: Using American Community Survey data, a Vulnerability Index is created and mapped to identify Census Block Groups most and least eligible for reduced fees for county programs. Dark grey are regions with no Households (HHs).



Statistical Learning: The 181 Arlington County block-groups are ranked based on the percentage of HHs spending ≥ 50% of HH income on housing, percentage of HHs with no vehicle, percentage of HHs receiving Supplemental Nutrition Assistance Program (SNAP), and percentage of HHs in poverty. The index is a simple sum of the ranks. The dark gray outlined areas outlined have a vulnerability index in the top quartile. Less than 10 of the 2317 households receiving reduce fee discounts are located in those neighborhoods.

Community Learning: The county program leaders learned about which neighborhoods (block groups) have the highest vulnerability so they could target outreach to these areas and consider other policy changes, e.g. increasing the level of subsidies.

Future Work: Explore construction of Vulnerability Indexes using local data and by new boundaries, e.g., define a neighborhood by its elementary school boundaries.

Figure 5. Vulnerability Index with Respect to Subsidy Eligibility

Example 3. Using Local Property Data for Planning

Arlington County demographers are concerned about the availability of affordable housing and in knowing how long a household has lived in Arlington. This information is used in their projections of the number of school age children. Since Arlington's housing capacity is limited by the supply of suitable land for residential development, it is important to monitor shifts in the type of residential development from single-family to multi-family housing. They asked the following questions:

- Where are the high-value (and the low-value) single-family homes in the county? Do these cluster and cross census tract boundaries?
- Which census tracts show diversity in value? Which census tracts show homogeneity in value?
- Which census tracts show diversity in other housing characteristics, (e.g., where are the new units, the large units, the multifamily units)?

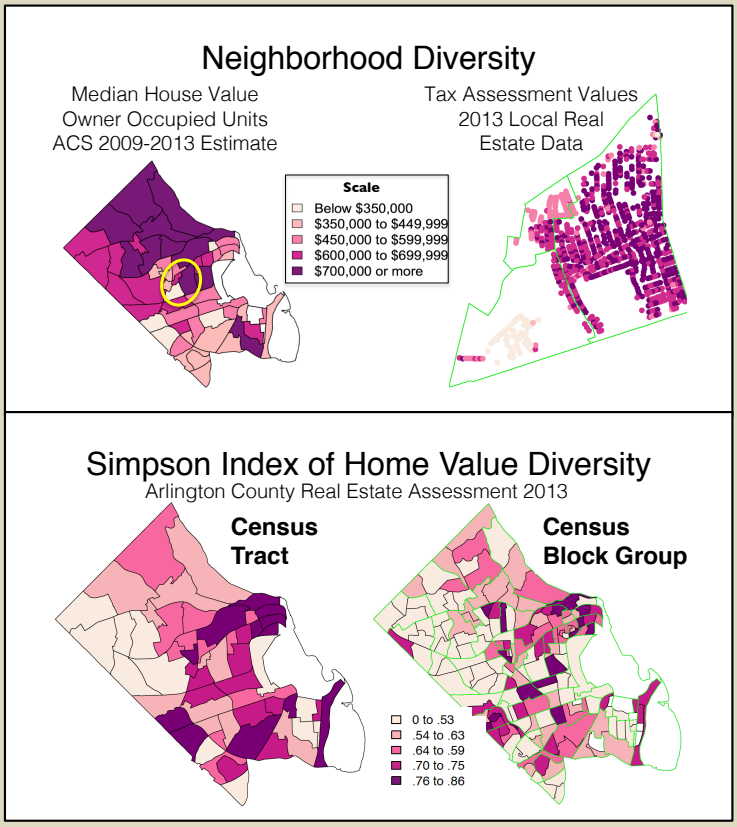
The charts presented in Figure 6 demonstrate the usefulness of local property data to answer the demographers' questions. Local real estate tax assessment data include the value of the home and characteristics of the structure and land on which the structure is built. The Arlington County real estate assessment data included 59,289 single-family housing units with assessed values and geocoded addresses placing them within the Arlington County census tracts. These data provide a detailed view of a geographic area that is not possible using similar data from federal surveys, such as the American Community Survey (ACS). The top chart in Figure 6 compares housing value aggregated by census tract using the ACS on the left hand side and then shows the area highlighted by the yellow circle (four census tracts) on the right hand side using the local real estate tax assessment data. The data displayed on the right provides a more detailed picture of the housing

stock since it reflects the heterogeneity within a census tract by using non-aggregated data at the level of the housing unit.

In addition, local real estate tax assessment data can be used as a proxy for wealth and to calculate measures of diversity (Simpson, 1949). The Simpson indices of diversity were estimated using the local data, house value, year built, property type, and number of bedrooms, for each census tract and census block-group (following Narwold and Sandy 2010) and displayed in the bottom of Figure 6. Not displayed is the same index calculated using the 2009-2013 ACS data, which is a sample of 450 households representing the approximately 60,000 single-family housing units contained in the local tax assessment data. The estimates using the local data are based on the population of single-family homes, whereas the ACS estimates are based on a sample that have less geographic detail. The Arlington demographers used the findings based on the local data in their long-run planning analyses.

Research Question: Is sufficient affordable housing available to encourage younger families to live in Arlington County? How long have families lived in Arlington?

Data Discovery and Exploration: Arlington County local real estate tax assessment data is used and compared to the American Community Survey (ACS) housing value estimates for 2009-2013.



Statistical Learning: Housing value data are used as a proxy for diversity. Comparing ACS data at the Census block level to the individual tax assessed values of the houses shows substantial heterogeneity within Census block groups not captured in the Census aggregates. The area highlighted by the yellow circle on the top map is now blown up on the right-hand side. This map shows local county administrative property data for housing values mapped at the housing unit level. The more detailed level of the county property assessment data allows one to observe the heterogeneity within a census tract that could not be seen using ACS data alone.

Measures of diversity, such as Simpson indexes, can be computed using local property data, in this case housing values. The bottom panel shows Simpson indices using local property data at the Census tract and Census block-group levels. Using 2009-2013 ACS data for the analogous index would be based on fewer than 450 observations (versus ~60,000 housing units) and would only present two geographic areas.

Community Learning: Arlington demographers found the mapping of home values, shown on the top map, and as an index shown on the bottom map, to be very informative for assessing the availability of a variety of housing opportunities for families and for planning school enrollment by age.

Future Work: Build on work in Figure 5 (constructing Vulnerability Indexes) and combine using local property data to expand understanding of Arlington at the neighborhood level to support planning and economic development.

Figure 6. Using Local Property Data to Identify Diversity of Housing Stock

5. CONCLUSIONS - IMPLEMENTING AND MATURING COMMUNITY LEARNING DATA DRIVEN DISCOVERY (CLD3)

Data-driven evidence is a necessary tool for community leaders to make the hard choices required to address disparities in healthcare and education, and in economic, social, and criminal justice. We have developed and piloted the CLD3 process that combines the promise of the data revolution with a scientifically rigorous approach to evidence-based policymaking and evaluation. Insights from the CLD3 process provide the necessary information for local governments to make decisions about alignment of services and funding.

To provide the data driven evidence necessary for local governments at all levels (e.g., cities, counties, towns) to build sustainable equity into its governance, it is time to move away from one-off studies and reports on isolated issues and bring the methods and tools of statistical sciences forward to ignite a culture of Community Learning. The CLD3 is natural for statisticians and social scientists to embrace since it mirrors the research process - data integration and exploration driving hypothesis development, development intervention strategies based on statistical and community learning, and a continuous evaluation of the intervention based on data integrated across departments and agencies. It is simply a restating of the scientific method that undergirds our research process on a daily basis. However, it is not necessarily so obvious to civic leaders in their policy development. This is why researcher engagement is needed to begin a sustainable movement.

Acknowledgements

We would like to thank the following people: Social and Decision Analytics Laboratory (SDAL) researchers David Higdon, Joshua Goldstein, Gizem Korkmaz, Emily Molfino, Mark Orr, Bianica Pires, Aaron Schroeder, and Katie Ziemer; 2016 SDAL Data Science for the Public Good student

fellows Mark Almanza, Virginia Tech; Madison Arnsbarger, Virginia Tech; Jessica Flynn, Cornell University; Adrienne Rogers, Virginia Tech; Will Sandholtz, University of California Berkeley; Emily Stark, Austin Peay State University; and Iowa State University Graduate Students Millicent Grant and Samantha Tyner; 2014-2015 summer SDAL student 2014-2015 Kyle Angelotti, University of Virginia; and Virginia Cooperative Extension (VCE) Edwin Jones and Michael Lambur. We would also like to thank Arlington County leadership and their staff, especially James Schwartz, Deputy County Manager and formerly the Fire Chief; Fire/EMS (Michael Gowen); Community Planning, Housing and Development (Elizabeth Hardy and Andrew D'Huyvetter); Parks and Recreation (Michael Peter); Economic Development (Andrea Morris); and the Executive Management Team who participated in the Data Discovery Workshop.

References

Abdul Latif Jameel Poverty Action Lab (2017), https://www.povertyactionlab.org/

Arlington County (2017), Arlington Fast Facts, https://projects.arlingtonva.us/data-research/fast-facts/

Banerjee, S. (2016), Spatial Data Analysis. *Annual review of public health*, *37*, 47-60.

Behn, B. (2015), The Black Box of Randomized Controlled Trials. Driving Government Performance, go to: http://hks.harvard.edu/EE/BehnReport.

Charlotte and Mecklenburg County (2017), Charlotte and Mecklenburg County Quality of Life Explorer, https://mcmap.org/qol/

Deaton, A., & Cartwright, N. (2016), *Understanding and misunderstanding randomized controlled trials* (No. w22595), National Bureau of Economic Research.

Flyvbjerg, B. (2005), Social science that matters. Foresight Europe, 2, 38-42.

Humes, K., Jones, N. A., & Ramirez, R. R. (2011), *Overview of race and Hispanic origin, 2010*. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

James, G., Witten, D., & Hastie, T. (2014), An Introduction to Statistical Learning: With Applications in R. Springer, New York.

JASON (2014), Data for Individual Health. Report prepared for Prepared for: Agency for Healthcare Research and Quality. The MITRE Corporation. McLean, VA, https://healthit.ahrq.gov/sites/default/files/docs/publication/2014-jason-data-for-individual-health.pdf

Kania, J. and M. Kramer (Winter 2011), Collective Impact. Stanford Social Innovation Review. https://ssir.org/articles/entry/collective_impact.

Keller, S. and Shipp, S. (2017, forthcoming) "Building Resilient Cities: Harnessing the Power Of Urban Analytics." In J. Bohland, P. Knox (Eds.) *The Resilience Challenge: Looking at Resilience through Multiple Lens*, Charles C Thomas Ltd Publishers

Keller, S. Shipp, S., Orr, M. Higdon, D. Korkmaz, G. Schroeder, A. Molfino, E. Pires, B. Ziemer, K., and D. Weinberg. (2016), *Leveraging External Data Sources to Enhance Official Statistics and Products*. Report prepared for the U.S. Census Bureau. Social and Decision Analytics Laboratory (SDAL), Biocomplexity Institute of Virginia Tech.

Kellogg Commission on the Future of State and Land-Grant Universities. (January 2001), Returning to Our Roots: Executive Summaries of the Reports of the Kellogg Commission on the Future of State and Land Grant Universities. Washington, D.C.: National Association of State Universities and Land-Grant Colleges.'

Narwold, A. and J. Sandy (2010), Valuing housing stock diversity. International Journal of Housing Markets and Analysis 3(1), 53–59.

National Academies of Sciences, Engineering, and Medicine. (2016), Metrics That Matter for Population Health Action: Workshop Summary. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. 2017. Refining the Concept of Scientific Inference When Working with Big Data: Proceedings of a Workshop. Washington, DC: The National Academies Press.

National Research Council. (2013), Frontiers in Massive Data Analysis. National Academies Press.

National Research Council. (2007), State and Local Government Statistics at a Crossroads. Panel on Research and Development Priorities for the U.S. Census Bureau's State and Local Government Statistics Program. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Pollock, K. (2016), "Policy: Urban Physics." *Nature* 531.7594: S64-S66.

Robert Wood Johnson (2017), http://www.rwjf.org/en/library/grants/2015/12/planning-a-city-health-dashboard-and-a-workplan-for-scaling-the-.html

Rittel, H., and Webber, M. (1973), "Dilemmas in a General Theory of Planning". Policy Sciences, 4(2): 155-169.

Sawicki, D. and W. Craig (1996), "The democratization of data: bridging the gap for community groups". Journal of the American Planning Association, 62(4): 512–523.

SDAL (2016), Student Fellows in National Capital Region Learn to Apply Data to Solutions for the Public Good, https://www.bi.vt.edu/sdal/news/student-fellows-in-national-capital-region-learn-to-apply-data-to-solutions

Simpson, E. H. (1949), Measurement of diversity. Nature.

Sterman, J. D. (2006), Learning from evidence in a complex world. American journal of public health, 96(3), 505-514.

University of Wisconsin Population Health Institute (2017), County Health Rankings Key Findings 2017.

U.S. Data Federation (2017), http://federation.data.gov/

Weinberg, D. H. (2011), US neighborhood income inequality in the 2005–2009 period. American Community Survey report. Washington, DC: US Census Bureau.