STRATEGY FOR A NATIONAL COMMUNITY LEARNING NETWORK

COMMUNITY LEARNING DATA DRIVEN DISCOVERY

Building Capacity for Evidence-Based Decision Making

Principal Investigator:
Sallie Keller, Professor of Statistics and Director
Social and Decision Analytics Laboratory of the Biocomplexity Institute, Virginia Tech

Submission Date:
December 19, 2016

Contractor Name and Address:
Virginia Polytechnic Institute and State University
Biocomplexity Institute of Virginia Tech

1015 Life Science Circle
Blacksburg, Virginia 24061

Submitted to:
Laura and John Arnold Foundation

STRATEGY FOR A NATIONAL COMMUNITY LEARNING NETWORK

COMMUNITY LEARNING DATA DRIVEN DISCOVERY
Building Capacity for Evidence-Based Decision Making

Preface

The Social and Decision Analytics Laboratory (SDAL) of the Biocomplexity Institute of Virginia Tech received a planning grant from the Laura and John Arnold Foundation (LJAF) to develop a strategy for a national community learning network. Through the development of this strategy Virginia Tech formed a partnership with Iowa State University, placing Virginia and Iowa at the center of initiating a national movement that would empower local governments to embrace data-driven governance. The first steps of this national strategy involve massive deployment of community learning across Virginia and Iowa, using these states as exemplars for igniting a full national movement. Based on the research completed in this grant, Virginia Tech and Iowa State University are developing a proposal with LJAF to realize the first steps of this national strategy - a Land Grant Partnership for Data-Driven Governance.

TABLE OF CONTENTS

Acknowledgements 3
Executive Summary 4
Statement of Problem and Approach 5
Overview of the National Strategy 5
CLD3 Engagement across Virginia and Iowa 6
Sustainability Plan for Virginia and Iowa that can be replicated in Other States 8
National Movement Using Virginia and Iowa as the Exemplars 9
Achieving and Maintaining Rigor in CLD3 Research and Methods 10
Resource Estimates to Implement the National Strategy 11
Anticipated Outcomes 12
Conclusions 13
Appendix 1: Community Learning Data Driven Discovery in Arlington County, Virginia 14
Appendix 2: Examples of Community Issues and Questions 35
Appendix 3: Deploying Community Learning across Iowa and Virginia 38
Appendix 4: Data Science Processes and Platforms 47
Appendix 5: Creating the National Conversation through Partnerships 53

Acknowledgements

We acknowledge the leadership of the Virginia Tech project co-PIs: **Sallie Keller,** Professor of Statistics and Director, Social and Decision Analytics Laboratory (SDAL) of the Biocomplexity Institute, Virginia Tech; **Stephanie Shipp**, Deputy Director and Research Economics Professor, Social and Decision Analytics Laboratory of the Biocomplexity Institute, Virginia Tech; **Michael Lambur**, Associate Director, Program Development, Virginia Cooperative Extension, Virginia Tech; and **James Schwartz**, Deputy County Manager, Arlington County, Virginia.

We acknowledge the efforts of our SDAL researchers: **David Higdon**, Professor of Statistics; **Vicki Lancaster**, Research Scientist (statistics); **Mark Orr**, Associate Research Professor (psychology); **Aaron Schroeder**, Research Scientist (data governance and information architecture); **Gizem Korkmaz**, Assistant Research Professor (economics).

We acknowledge the hard work and dedication of the 2016 Data Science for the Social Good Fellows: Undergraduate students **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**.

This vision was developed with substantial input and advice from colleagues at our principal partner institution, Iowa State University. This includes: **Sarah Nusser**, Vice President for Research and Professor of Statistics; **Erin Mullenix**, Director of Data Driven Science, Iowa State University; **Chadwick Higgins**, Senior Director, Iowa State University Extension and Outreach; and **Wolfgang Kliemann**, Associate Vice President for Research.

Strategy for a National Community Learning Network

Executive Summary

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. Virginia Tech developed and piloted a **Community Learning Data Driven Discovery (CLD3)** process that liberates, integrates and makes these data available to **government leaders and researchers** to build an equitable and sustainable social transformation within and across communities to address their most pressing needs. Dr. Sallie Keller, Virginia Tech Director of the Social and Decision Analytics Laboratory will lead the deployment across Virginia and nationally. She is experienced in undertaking and successfully implementing large initiatives at universities, national laboratories, and in the private sector. She has forged a strong partnership with Iowa State University to deploy CLD3 across two states to maximize learning and deployment.

This document presents a national strategy for the massive deployment of the CLD3 process to our 3100 local communities throughout the U.S., providing government leaders with the tools to tackle the issues confronting their communities and to learn "What works, for whom, and in what context?"

The CLD3 process starts with asking local leaders what questions they have which they 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. These steps encompass building capacity of civil servants, Cooperative Extension Service agents, and researchers to work collaboratively to undertake these steps.

The CLD3 deployment is through the Land Grant University, which excels in agriculture, engineering, science, social science, and statistics. This **research is translated to application** through the Agriculture Experimental Stations and Cooperative Extension Service (CES) network, located in every city and county and operated in partnership with federal, state, and local governments. CES is a seasoned nationwide network with a track record of successes focused on addressing societal issues by **disseminating university research findings that provide citizens with the knowledge and skills necessary to solve community problems**. CES agents will work directly with local government leaders and stakeholders to address their issues. They will identify and work with university data scientists and researchers to collaborate with local government leaders to create data learning communities. Together we will develop the workforce through training and embedding data scientists in local settings to enable the full potential of data driven discovery for communities.

CLD3 will reposition our public institutions to focus on communities' quality-of-life and sustainability through data driven problem discovery and solutions. Our goal is to create a national movement with the momentum to deploy CLD3 to all our communities across Virginia and Iowa and then the country. The result will be a new **foundation for democracy** and a deeper partnership between our Public and Land Grant Universities, their communities, and state and federal government.

The creation of Land Grant Universities, which were signed into law by President Lincoln 150 years ago, has benefited communities across America by supporting health, education and agriculture transformation. Cooperative Extension Services have been and continue to be the primary drivers and coordinators of Land Grant Universities and local communities. Our proposal is based on leveraging this existing, successful, federally and state funded program to bring communities into the 21st Century, to raise the skills, capabilities and infrastructure locally to support data driven decision making.

Statement of Problem and Approach

Our vision is to bring the *all data revolution* to **small, medium, and rural** communities across the nation through **Community Learning Data Driven Discovery (CLD3)**. Our CLD3 process liberates, integrates, and makes data available to **government leaders, researchers, and citizens,** enabling them to not only address their most pressing current challenges but also to create an equitable and sustainable social transformation within and across their communities.

Imagine local governments using the massive data flows they collect during the course of conducting their business to inform their decision making. These include data about permits, 911 incidents, snow, housing values and characteristics, bus and metro usage, education, and much more. Now imagine combining these data with social media to understand public sentiment, and with federal and state data to provide demographic and other information at the neighborhood level. These data sources and the statistical learning that follows are instrumental for improving the efficiency and sustainability of local governments. But to realize the enormous potential that lies within these data sources for advancing the public good, local governments need access to a host of research approaches, technical support, training, and tools.

Policy decisions to address the common good are more often informed by political ideology and anecdotes than data-driven evidence. Disparities in healthcare, public health, education, economic, and criminal justice across America have reached a tipping point that require new data-driven approaches to making policy decisions, defining interventions, and evaluating their consequences. This barrier can be overcome with a massive implementation of data-driven community learning (CLD3) across the country through our land grant and public universities, building a sustainable ecosystem able to support evidence-based policymaking at all levels of government.

The proposed first step is to create exemplars across two states, **Iowa and Virginia**, demonstrating how to turn existing local data flows, integrated with state and federal administrative data sources, into knowledge and insights that will support data-driven decisions and policymaking for the improvement of the delivery of services, the equitable access to resources, and for improved quality-of-life for all residents. The second step is **to use this momentum to create a national movement** that will reposition our land grant institutions as the catalysts for communities' quality-of-life and sustainability through data-driven discovery and solutions. The result will be a new **foundation for democracy** and a deeper partnership between our Public and Land Grant Universities, their communities, and state and federal government.

Overview of the National Strategy

CLD3 is a bold, yet attainable, vision to liberate and repurpose local administrative data and integrate it with more traditional state and federal statistical data *and* to unroll this approach across the United States. The challenge is introducing and embedding the CLD3 process in local governments across the country where the skills and resources for data analytics are lacking. This challenge can be accomplished by leveraging our Land Grant Universities whose research excels in agriculture, engineering, science, social science, and statistics. This research is translated into action through the Land Grant University infrastructure and includes the Cooperative Extension Service (CES) network, operated in partnership with federal, state, and local governments.

CES is a nationwide network with a long track record of success in addressing societal issues by disseminating university research findings that provide citizens with the knowledge and skills necessary to solve **community problems**. CES agents will work directly with local government leaders and stakeholders to address their issues. They will identify and work with university data scientists and researchers to create statistical learning communities. Ultimately, well-trained data scientists will be embedded in local settings to enable the full potential of data-driven discovery for communities.

We have demonstrated the CLD3 concept with Arlington, Virginia (and that work will continue, see Appendix 1) and shared our accomplishments with LJAF, who subsequently challenged us to think about how to first scale the CLD3 process across Virginia and Iowa and then nationally.

A national sustainable model and culture of data-driven community learning can be realized through the following three components:

1. Deep CLD3 engagement scaling across Virginia and Iowa through the CES network,
2. A sustainability plan for Virginia and Iowa that can be replicated in other states with other land grant and public university partners, and
3. A plan for engaging new partners using Virginia and Iowa as exemplars.

CLD3 Engagement across Virginia and Iowa

The CLD3 process starts with asking local leaders about their challenges and using data from their communities and other sources to inform their decision-making, and continues through a continuous, sustainable, and controlled feedback loop. Building on the lessons learned from working with Arlington (see Appendix 1), our approach is for CES and university researchers to:

- Expose local leaders to ***data-driven learning*** through a forum across government departments and offices. The forum is structured to bring local program and agency leaders together through a dialogue process that is designed to discuss issues and data sources that span their boundaries.
- Work with local leaders to develop CLD3 governance structure, e.g., relevant data access and sharing agreements, and defining their critical questions and issues facing their community by asking them "What keeps you up at night?"
- Identify and wrangle the data sources that cut across programs to provide the knowledge and insight for building an agile and scalable statewide CLD3 data infrastructure.
- Use statistical and geospatial learning along with the communities' collective knowledge to develop hypotheses to inform policy decisions relevant to issues.
- Develop and deploy intervention strategies and education programs to address issues. Evaluate hypotheses using rigorous experimental design principles. Adapt strategies based on continuous evaluation.
- Train the next generation of local government civil servants to be data savvy.
- Engage entrepreneurs to develop products and tools using local data to support economic development and local, state, and federal government.

Virginia and Iowa are excellent initial exemplars for CLD3 deployment. We are complementary in our style of local governance. Virginia is primarily county-based, governed by boards. Iowa is primarily city-based, governed by mayors and councils. Our first level of engagement will be with the civil servants that support these governing bodies. Common issues include anticipating and preparing for meeting the day-to-day needs of vulnerable populations and the needs of all during emergency situations and other adverse events while understanding how to better serve their communities.

Appendix 1 provides a review of our early and current work with Arlington. This work has been presented at symposia, and papers have been prepared for publication in journals and trade magazines. Also included are three new integrated issues facing the police, parks and recreation, and library departments. These departments are concerned with at-risk youth, barriers to participation in programs, and underserved populations. Appendix 2 provides additional examples of the problems studied or expected to be studied, both existing and emerging, across Iowa and Virginia, each of which involve policy issues that require the development and evaluation of intervention and educational strategies.

CES leadership in our states have embraced the CLD3 vision and recognize that it raises the CES beyond their current approaches and program area boundaries to a more holistic and data-driven engagement with their communities and a more direct engagement with the local government. Communities in Virginia and Iowa will be introduced to CLD3 during the first three years of the project. Local government leaders will be identified within each community and engaged to facilitate the CLD3 implementation processes. Part of the deployment will be locality based and part may be based on problems that cut across more than one community within a state.

In Virginia, work has begun to identify CES units to initiate CLD3. The goal is to work with motivated communities for early CLD3 introduction to maximize the opportunity for quick adoptions of the CLD3 process, resulting in excellent exemplars for other communities. A committee that is a mix of Virginia Cooperative Extension (VCE) extension agents, VCE unit coordinators (i.e., local extension agent leadership), VCE district leaders, and other VCE leadership formed met weekly to guide the CLD3 development. We developed a webinar on the CLD3 process and initiative and held this for VCE unit coordinators. The purpose of the webinar is to create interest in, and provide information about becoming involved. Extension agents throughout these activities have represented approximately 10 percent of the Virginia counties and several agents have participated with local government representatives. Given the strong positive reaction to the webinar, it will be repeated. From these engagements, a list of interested extension units are being identified with the intention of selecting 7-10 units for the first wave of deployments based on criteria below. This selection will be done through a Request for Proposal Process or by some other process that maximizes success. More details of the proposed deployment plan are given Appendix 3.

In Virginia, the community selection criteria are:

- District coverage – include representation from all four districts aligned by region of the state
- Mix of population size and rural/urban designation
- Strength of most recent data-based situation analysis
- Interest in participating from both Virginia Cooperative Extension and local government
- Strength of local government support for Cooperative Extension
- Relationship of agents with local government
- Repeatability and replicability to other units and communities
- Potential issues of high public interest and concerns.

In Iowa, the key community decision structures are the cities, often supported by their surrounding counties. The team at Iowa State University developed criteria to select a set of representative cities to participate in the first phases of the CLD3 project. These criteria are below.

- A mix of urban and rural character of communities
- Population size
- Population diversity
- Base of economic activity: services, manufacturing, agricultural
- Community leadership familiar with data driven approaches
- Community leadership committed to joint project approach
- Engaged Extension office in surrounding county
- Availability of local community college and/or liberal arts college(s) for project collaboration.

Based on these criteria the team has developed a priority list of cities to approach immediately at project start. In addition, Iowa State University has convened a group of interdisciplinary faculty as an advisory

committee and prospective researcher pool who are highly interested in this vision. More details are given in Appendix 3.

Sustainability Plan for Virginia and Iowa that can be replicated in other states

The overarching goal of our approach is to build **local sustainable** data-driven community learning **cultures**, one that can support local government policy development and implementation. This requires **momentum**, a shared and **replicable infrastructure** to support the new data-driven decision making capacity, **sustainable resources** aligned with civil servant and CES skill sets, supplemented by researchers that supports community learning, and a **workforce pipeline**.

Momentum. We will collaborate with a sufficient number of counties and cities within our states in the first three years to create the momentum for CLD3 to spread statewide during years 4 and 5. Successful deployment across Virginia and Iowa will result in a more efficient governance structure and will institutionalize the CLD3 process, providing the momentum for a national deployment.

Replicable Infrastructure – Translating Research to Practice. We will collaborate with faculty, staff, and students across our campuses to develop the underlying CLD3 research and resources to deploy across our states and shared nationally. Working closely with CES leverages the ability to help expose researchers to the important issues and challenges facing communities, thus generating bottom-up rather than top-down research, providing the opportunity for a greater impact on policy and interventions. Key components of the initial shared infrastructure being developed are described below.

- ***Templates (Recipes) and Tools.*** We will develop templates for baseline characterization of communities and for data sharing practices that involve statistical and geospatial integration of local, state, and federal data sources. The development and sharing of research tools will also be included, such as rigorous experimental designs for policy interventions and evaluations, statistical and geospatial indicators, geographic visualizations, and dashboards.
- ***Integrated CLD3 Web Resources.*** We will document best practices and lessons learned across local communities using the web. The webpages will include our vision and overview of CLD3; community pages; a blog for news ideas and best practices with regular updates and insights; information for new cities/counties on how to participate and engage in CLD3; highlights of the impact existing CLD3 efforts towards driving evidence-based decision-making; a running list of successful policy interventions and their contextual constraints that have empirical evidence of success; online forums where government leaders, researchers, and citizens can inquire about the process and discuss their particular challenges and successes with others; and other pages, such as examples of data agreements, statistical tools, and best practices. In year 2 and thereafter, we will maintain and expand CLD3 web-based resources.
- ***CLD3 Data Science Processes and Platforms****.* Starting with deep expertise building federated data systems in Virginia, we will develop and implement statewide-federated data science processes and supporting platforms for Virginia and Iowa local governments, CES agents, and researchers. These processes and platforms are fundamental components of an integrated statewide CLD3 ecosystem and will become part of the foundation of the Universities' infrastructure. The sustainable strategy is for the state Land Grant Universities to become the stewards of this infrastructure ensuring it is accessible by local communities, researchers, and state agencies. This data infrastructure will manage and mediate access to data and data tools, adhering to necessary data privacy, confidentiality, security and data governance constraints. The associated technology will enable greater linkage of data capable of adhering to required privacy and security protections. Researchers across many academic departments, who want to help local or state government solve problems, would be able to find government partners and utilize the coordinated data infrastructure being curated by our institutions to the benefit of their research and the government partner. More of the technical details of CLD3 Data Processes & Platforms are provided in Appendix 4.

- ***Communities of Practice.*** Through focused training, meetings, and workshops where our practitioners, academics, and government colleagues share CLD3 research findings and best practices, we will develop local, regional, and state Communities of Practice, and an overarching national Community of Practice. The use of online forums and integrated website will help to create and maintain these communities of practice.

Sustainable Resources. As the momentum in Virginia and Iowa spreads statewide, Virginia Tech and Iowa State will collaborate with CES leadership to support the realignment of resources to support the CES and CLD3 activities in our communities, including repositioning job skills for both CES and local governments. The continuing education requirements for CES agents will include rigorous data analytics and statistics requirements. This training will vary by state and the roles of the CES agents and faculty. Additional training and partnering with researchers to support this transition for CES agents and their civil servant counterparts will be developed and implemented, such as workshops, data laboratories, and remote learning. Finally, resources currently in the system to support CES, faculty, and local government will need to be redefined. For example, in Virginia, the Virginia Cooperative Extension conducts a programmatic assessment in each locality every three years and in Iowa, they conduct a statewide needs assessment every five years to identify mission-driven, programmatic areas of emphasis. These approaches and timing will be aligned to foster and drive CLD3 sustainability.

Workforce Pipeline. The universities will take responsibility for developing a workforce pipeline of data scientists to join CES and local government. Universities will do this through Fellowships, Honors programs, internships, and research for credit experiences in coordination with their CES. Experiential learning and academic mentoring provide a perfect fit for developing a workforce dedicated to community-based research. One example is Virginia Tech's **Data Science for the Public Good (DSPG)** program (https://www.bi.vt.edu/ndssl/projects/data-science-for-the-public-good-program). DSPG student fellows are members of horizontally and vertically integrated teams; horizontally integrated across the disciplines, (e.g., statistics, data science, social and behavioral sciences, public health), needed to address complex issues; and vertically integrated to collaborate with project stakeholders at all levels, community college, undergraduates, graduates, postdoctoral associates, research faculty, and local, state, and federal agency leadership.

The DSPG model serves as 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. We will formalize this model and explore others that provide students with experiential learning and research opportunities in our local communities with the goal of placing them in local government and the CES workforce after graduation.

National Movement Using Virginia and Iowa as Exemplars

The goal is to institutionalize the CLD3 process, platforms, and resources within Iowa and Virginia and then facilitate the introduction across the country. The groundwork for this introduction is already underway. We have started a national dialogue around the concept of data-driven community learning and the role of the Land Grant Universities. Appendix 5 provides a summary of our conversations with the leadership at universities, the White House, professional societies, other Land Grant Universities, and local, state, and federal agencies. Across the board, the concept is being embraced and a resounding, "This is an idea whose time has come" is being echoed. Now we just need to do it!

To ignite a national movement will require strong leadership and partnering to obtain the resources and experience to build a national movement (in addition to LJAF). Dr. Sallie Keller, Virginia Tech, Director of the Social and Decision Analytics Laboratory will lead the development and deployment across Virginia and nationally. She is experienced in undertaking and successfully implementing large initiatives at universities, national laboratories, and in the private sector. The national strategy includes planning and implementing a succession plan to an organization to sustain the initiative beyond five years.

We plan to continue discussions to engage agencies at all levels of government and key professional societies, such as the Association of Public and Land Grant Universities (APLU), the National Association of Counties (NACo), National Governors Association, and the National League of Cities (NLC). We will also engage in parallel state-level associations with our respective states. We will work to find other funding partners and have begun discussions with Results for America, which manages the Bloomberg-funded What Works Cities competition, and we are planning on follow-up discussions with Sloan Foundation, MacArthur Foundation, Glaser Foundation, and others. These engagements are critical for the national strategy and to develop the succession plan to ensure sustainability beyond 5 years.

We have begun and plan to continue to engage other Land Grant Universities and other institutions of higher education, community and vocational schools. All resources will be openly available through integrated webpages and websites. These resources will provide the foundation for institutions to expand CLD3 to their communities and contribute to a growing body of community-based research around data-driven community learning. We have relationships with Virginia Commonwealth University and University of Virginia and have begun a relationship with the Northern Virginia Community College, the second largest and most diverse community college in the US. In Iowa, through the three major Regents universities, and extensive network of smaller 4-year institutions, there will be many opportunities to collaborate on community-based research. We will work with these communities and institutions to deploy statewide. These activities will culminate annually in a national workshop focused on CLD3 research, methods, and examples.

Our plan, beyond Iowa and Virginia, is to develop a national competition in year 3 and implement the plan in year 4 to introduce and train communities and Land Grant University extension staff and researchers in additional states. We will collaborate with organizations that have the experience and resources to help implement this process nationally. In parallel, we will support the Land Grant Universities in establishing the necessary infrastructures

As CLD3 becomes institutionalized in local governance and a fundamental part of Land Grant Universities, the billions of federal, state, and local dollars already allocated to support CES will coincide with CLD3 activities. This movement provides a repositioning of how CES engages with communities, providing a deeper partnership between our public and land grant universities, their communities, and state and federal government. In Virginia and Iowa, the CES leadership has embraced this new model for working with communities. The CLD3 model builds on and enhances their work with communities by working with local government leadership, adding local data to their toolbox, widening the reach to research faculty, and introducing flexibility in programming resources.

Virginia Tech and Iowa State will support community-based research through CLD3 combined with other university specific programs. CLD3 will reposition the interactions with communities on a grassroots level. CLD3 will offer the opportunity for faculty to engage our communities with needs-driven scholarship and Extension and Outreach will act as the catalyst and mechanism to create, solidify, and sustain that connection with the community. The realignment will occur in CES and faculty engagement with communities in a needs-driven environment.

Achieving and Maintaining Rigor in CLD3 Research and Methods

Leadership at our institutions is interested in partnership models that increase the level of scholarship in community-based research (i.e., community-based scholarship that is disseminated in peer-reviewed journals and leads to citations for the institution). A clear path for this is the development of teams of extension and non-extension faculty and extension personnel to work with communities to use their data to inform decision-making. This type of dedication to quality community-based research is central to the mission of land grant universities across the country. Working together to embed CLD3 in local governments across our states offers an unprecedented opportunity to directly impact the scientific integrity of community-based research.

Working with Land Grant University CES and researchers in a cohesive way provides communities direct access to data-driven research using rigorous statistical methods and probabilistic evaluations. These are well-trained and dedicated scientists who embrace opportunities to demonstrate the broader impacts of their research. Extension personnel supplement this with local knowledge to isolate those issues that would benefit from scientific scrutiny and the connections to build a bridge between community leaders and the researchers. This unique 3-way partnership between university researchers, extension staff members, and the community is a win-win-win opportunity. It will not only foster cutting-edge community learning but also provide the foundation for world-class research scholarship to be disseminated in peer-reviewed journals, thus sustaining all key partners.

Resource Estimates to Implement the National Strategy

The plan is to roll out Community Learning Data Driven Discovery (CLD3) over five years, building on the proof-of-concept demonstrated in Arlington, Virginia. The plan starts with pilots in Iowa and immediate expansion to new areas in Virginia, with continuing and deepening activity in Arlington. The deployments across Virginia and Iowa would continue in years 2 and 3, reaching and exposing CLD3 to more than 30 communities across these states, creating Iowa and Virginia exemplars. This degree of diffusion across the states will be sufficient for their Cooperative Extension Services to reposition their approach to communities and to realign existing resources such that the CLD3 process will naturally spread and be sustained across each state. Momentum will be maintained through the communities of practice, access to the Data Process and Platforms technology, and the continual updating and provision of informaion on the intgrated websites.

In year 3, plan is to hold a competition to select new states to begin to implement CLD3,. This competition will be in partnership with other organizations like the Association of Public and Land Grant Universities and the National Association of Counties to host and champion the competition.

Resources are estimated to be $10M in years 1-3 to support the spread of CLD3 across Virginia and Iowa and support the development of the sustainable and repeatable infrastructure, as well as igniting the national dialogue. The National competition and expansion is estimated to cost another ~$10M to engage an additional 5 states.

To gain a sense of the resources this strategy would leverage in just Virginia and Iowa, the table below shows the Cooperative Extension Services (CES) investments in Iowa and Virginia for 2015. Investment in this CLD3 initiative has the potential to redirect these CES investments and associated workforce and activities into an ecosystem for evidence-based policy and community decision-making through the integration of local, state, and federal data.

2015 Cooperative Extension Budgets

Source	Iowa		Virginia	
Smith Lever Act Formula Funds	$9,603,568	10%	$7,409,653	11%
Other USDA NIFA fund	$1,095,181	1%	$1,839,573	3%
Total State Funds	$26,080,278	28%	$32,839,573	48%
Local County/City Funds	$34,459,862	37%	$12,835,727	19%
Other USDA Sponsored Research support CES	$13,181,368	14%	$9,958,425	15%
Local Agency Fees (e.g., 4-H Club)	$9,697,721	10%	$2,952,194	4%
Total CES Funds	**$94,152,627**		**$67,835,145**	

USDA: U.S. Department of Agriculture
NIFA: National Instititue of Food and Agriculture
CES: Cooperative Extension Service

Anticipated Outcomes

This strategy addresses how the CLD3 movement starting in Virginia and Iowa can progress across the nation. The strategy includes a plan for sustaining this movement beyond initial investments. Significant and integrated resources are required to create a sustainable national movement. The CLD3 process asks communities to approach decision-making and policy development using data as evidence. We expect Land Grant University and CES to reposition their engagement with the public and to be leveraged as the vehicle for broad-based CLD3 deployment. The initial resource infusion will provide bandwidth for statistical learning approaches to evidence-based policy driven CLD3.

Early successes are needed to form powerful exemplars; this will be Virginia and Iowa. Once CLD3 takes hold in a community, existing local government and CES resources can be realigned to support the transformation, including repositioning the local civil servant and CES workforce and enhancing faculty involvement with our communities to focus on issues driven by communities.

The outcomes we expect include:

- The creation of evidence-based decision-making capacity in local government by bringing data science to small and medium cities and counties and rural areas through a scalable, repeatable, and sustainable process in collaboration with researchers at land grant and public universities. We will start with cities and counties in Virginia and Iowa - creating a sustainable culture where the use of data and statistically rigorous evaluation methods becomes the new norm.
- Opportunities to bring data, research, and analyses to rural areas. This initiative provides the ability to study a network (or networks) of rural communities within and across geographic areas. In Virginia, examples of this research include the transition of traditional agriculture (tobacco and food) to grapes and niche areas supplying unique kinds of produce to markets, e.g. sustainable farming and unusual kinds of produce for new food delivery services. In Iowa, rural research themes include infrastructure, education, public health, economic development, and public safety issues.
- Repositioning of the Virginia Tech and Iowa State land grant universities and their CESs to work with local governments and communities and address city, county, and rural issues by linking their local data with state and federal data sources to create a more complete body of information.
- Creation of a replicable infrastructure (ecosystem) capable of rapid translation of research to practice. This will include open source loosely federated data processes and associated platforms that can be adopted or adapted by other states. Land grant universities will become the stewards of this infrastructure ensuring it is accessible by local communities, researchers, and state agencies.
- Creating and sharing of an integrated system of resources that includes repurposed and integrated data sources, data sharing agreement templates, software and code for statistical and geospatial methods, a data base of interventions that have undergone rigorous evaluation, and other research tools that can be adapted and reused by other communities and researchers. This would capture best practices that could be replicated across communities. These resources will form the foundation for the evolution of a community of practice around community learning and data-driven policy development and evaluation. It will also allow local governments to work with each other and data will become their new language for communicating and sharing ideas, data, and policy changes.
- Engagement of talented researchers across multiple academic departments and disciplines, connecting them to community-based research with local and state government partners and utilizing the coordinated data infrastructure being curated by the land grant universities to the benefit of their research and their government partner while advancing academic curricula.
- Development of programs that serve as an incubator to promote workforce development by educating and training the next generation of public servants and government data scientists and exposing them to data science projects that integrate data from municipal, state, and federal levels of government.

- Opening of entrepreneurial space for the development of new products, processes, and tools based on local community data.
- Local government decision-making will lead to higher quality of life, and efficient and equitable delivery of services in their communities.
- Succession plan to ensure sustainability and growth of CLD3. This will likely be transitioned to a national organization such as the Association of Public and Land Grant Universities (APLU), the National Association of Counties (NACo), National Governors Association, and the National League of Cities (NLC) and potentially funded through state and federal funding.

Conclusions

Our national strategy to create a Community Learning Data Driven Discovery process provides a vision for embedding data driven processes in local governments to better inform decision-making, and drive successful results. The strategy describes how to take the first steps in two states to ultimately realize this vision across the country. As we succeed in Virginia and Iowa, Virginia Tech and Iowa State University will win the respect of their land-grant peers and they will want to join this movement. The development of a national movement capable of leveraging the U.S. Land Grant University infrastructure will provide the opportunity to transform how local governments build evidence into daily decision-making. What could be accomplished with this agenda reflects the origins of the mission of the nation's land grant university network to translate research to application.

Appendix 1: Community Learning Data Driven Discovery in Arlington County, Virginia

1A The Community Learning Proof-of-Concept

This appendix chronicles our work with Arlington County, Virginia. What began in 2014 as a partnership between the Social and Decision Analytics Laboratory (SDAL) of the Biocomplexity Institute at Virginia Tech and Arlington County expanded in 2016 to include Virginia Tech's Cooperative Extension Service. This joint venture strengthens the long-standing mission of the land grant university to enhance the security and social well-being of the citizens it serves and repositions Virginia Tech for the twenty first century.

Introduction

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. To provide the data driven insights and evidence necessary for rural and metropolitan communities to build an equitable and sustainable social transformation, we have developed the **Community Learning Data Driven Discovery (CLD3)** concept. CLD3 is a framework for understanding the causal powers of change. It provides the tools for rural and metropolitan areas to leverage their distinct assets by incorporating feedback from the community's collective knowledge together with existing administrative, designed, procedural, and opportunistic data sources[1] to tackle the question "What works, for whom, and in what context?" CLD3 development is based on direct engagement with community leaders and the use of existing data sources resulting in a process that is transferable and adaptable to the contextual constraints of other rural and metropolitan communities.

CLD3 is a holistic approach that:

- integrates the community's collective knowledge with usable insights gained from statistical learning to provide civic leaders with the information needed to develop mechanisms of intervention and policies that meet the goals of the community within their contextual constraints;
- designs, measures, and evaluates the intervention chain to gain a better understanding of what works, what doesn't, and why, by understanding the correlations between issues, the timing of problems, and how the data sources fit together; and
- redirects interventions when warranted based on the information from a continuous systematic review.

Community learning is achieved by iterating through the following steps:

- Expose local leaders to ***data driven learning*** through a forum across government departments and offices. A single department may be reticent to align their services to tackle the problems facing vulnerable populations due to the complexity of the associated factors such, food insecurity, victimization, homelessness, family, mental health, and socioeconomic barriers. The forum is structured to bring local program and agency leaders together through a dialogue process used to identify issues and data sources that span their boundaries.

[1] *Designed data* are statistically designed data collections, such as surveys or experiments, and intentional observational collections. *Administrative data* are data collected for the administration of an organization, program, or service process. *Opportunity data* are data generated on an ongoing basis as society moves through its daily paces from a variety of sources such as GPS systems and embedded sensors, social media exchanges, mobile and wearable devices, and Internet entries. *Procedural data* represent procedures and policies that can have impact on the trends and associations found in the data analyses.

- Work with local leaders to define critical questions and issues facing their community by asking them "what keeps them up at night?"
- Identify and wrangle the data sources that cut across programs that could provide insights, building an agile and scalable CLD3 data access and process platform (described in detail in Appendix 3).
- Use statistical and geospatial learning along with the communities' collective knowledge to develop hypotheses to inform policy decisions relevant to issues identified.
- Develop and deploy intervention strategies and education programs to address issues. Evaluate hypotheses using experimental design principles. Adapt strategies based on continuous evaluation.

The backbone of the CLD3 concept is the creation of a secure collaborative research environment (ecosystem) to access, link, and analyze administrative data sources from federal, state, and local levels together with opportunistic, designed, and procedural data sources. This requires technology development and implementation that will serve to build capacity for a sustainable culture of collaboration and data driven discovery within and across our local governments. The details of this backbone, the CLD3 Data Access & Process Platform, are provided in Appendix 3.

Part of the CLD3 Data Access & Process Platform is the management of data sharing agreements and processes that respect data governance across agencies. Our experience in Arlington County started by working through the Chief Information Officer's department. SDAL and Arlington County signed a memorandum of understanding that allowed SDAL to access Arlington's data with the commitment to keep the data secure. The management of the data and this commitment was through the implementation of the policy preserving procedures in our CLD3 Data Access & Process Platform. Subsequently, additional data agreements have been put in place to access more sensitive data, such as fire emergency service and police data. Some data sources, such as housing and other property data are publically available and therefore do not require data agreements. State data will require creating data agreements with the corresponding state office, e.g. Clerk of Court data such as marriage licenses and other court related data.

Research in Support of CLD3

In support of the CLD3 concept, we are identifying several areas of research that will benefit not only CLD3 but also support the open data and evidence-based policy movements as well. We often see local problems being tackled with federal data at a gross scale that ignores the heterogeneity in the smaller spatial scales. In particular, we are not seeing 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. 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 (http://federation.data.gov/) provides access to data sources from the federal, state, and local levels (very limited local data sources), but does not take the difficult and time consuming next step and provide the technology to integrate these sources. We are working to make this a reality for leaders and researchers at the community level through the CLD3 Data Access & Process Platform while developing data sharing agreement templates to facilitate the use of the platform across state lines. Putting this in place will help the sustainability of data-driven governance, local communities will not have to expend the time and resources searching and integrating data sources thus knocking down one of the major obstacles to evidence-based policy.

We often see a lack of statistical reasoning when using data to address local issues. For example, spatial statistics that do not take into account or estimate the variability of the statistic, intervention evaluations that do not acknowledge the impact (correlations) within the system, models that do not account for the hierarchical and spatial nature of the data, designs that fail to account for confounding factors. To address these issues, we have begun research into the development of statistical learning and geospatial indicators in complex social systems, identifying the data sources needed to describe them, how to quantify and interpret them, and how to measure the correlation structure between them. Understanding and quantifying the association of geospatial indicators pre- and post-policy intervention is crucial to accurately interpreting the impact of an intervention. Research is also being done in the area of experimental design for policy interventions and evaluations. Statistical designs that incorporate or account for the myriad of confounders in a complex system, that provide the flexibility to adapt to a changing system, and account for the sources of variability either through statistical blocking or in the measurement process are needed. We will also research the applicability of adaptive designs which allow for planned modifications to one or more aspects of the intervention based on analysis of the data from subjects in the intervention. Specifying a proscribed monitoring and data analysis schedule prior to the start of the intervention, allows the researcher to evaluate progress over time and make modifications without undermining the statistical integrity and validity.

1B CLD3 Implementation in Arlington County, Virginia

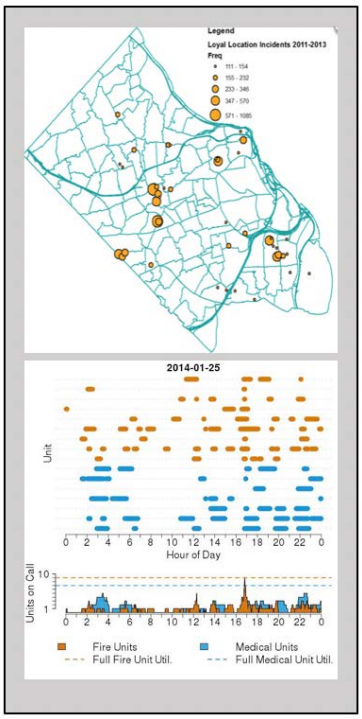
Early Engagement with Arlington Leadership

SDAL is one of four laboratories in the Biocomplexity Institute of Virginia Tech. The Biocomplexity Institute addresses pressing challenges to human health, habitat, and well-being by using an information biology approach to predict, explain, and visualize the behavior of massively interacting systems from the molecular level to societies and populations. To implement this vision, BI decided to add capability and initiated a new laboratory, the Social and Decision Analytics lab (SDAL) three years ago. At that time, SDAL Director, Sallie Keller, and Deputy Director, Stephanie Shipp, assumed the research focus would be on megacities to study societies and populations. To gain experience in urban analytics, they stepped outside the BI front door and met with Arlington leaders. This brought forward the realization that cities and counties of all sizes would deeply benefit by embracing the *all data revolution*. SDAL began to develop the concept of community learning, adapted from health learning systems, to bring data science and the all data revolution to these communities.

Fire Chief Schwartz, now Deputy County Manager, asked us our first question. He wanted to know if was it possible to link together all of the data collected during the course of an emergency to recreate incidents end-to-end through the data. This sounds simple, but it is not. The data being generated during an emergency are used to guide the fire/EMS staff in their response to the emergency and were never intended 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 in the various data sources. In order to address firefighter/EMS questions regarding resource utilization, SDAL cleaned, transformed, and linked the data sources to reconstruct the incident. The data discovery process included:

1. Mapping incidents geospatially to identify hotspots and call volumes.
2. Identifying loyal customers and loyal locations (top right) – serving non-911 level customers leads to stress and anxiety among firefighters and an inefficient use of resources.
3. Visualizing the relationship between weather and incidents (bottom left) – locating resources properly during various weather events would improve response times, e.g., wooded areas during wind events or low income areas during heat events.
4. Calculating the match (or lack of) between call classifications and final incident classifications – leading to better use of resources.

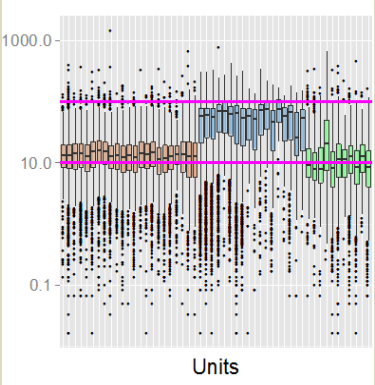
Since then we have engaged with Arlington to take a data-driven approach to improve their emergency alerts program and Operation FireSafe program that installs smoke alarms in homes without them. We are embarking on using administrative data sources from Parks & Recreation, Library, the Police Department, and social media to identify vulnerable populations and barriers to participating in county programs.



Example 1. 911 Incidents

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 characteristics are common to people who call the 911 system three or more times in one month?*



Statistical Learning: The medical and fire unit 911 deployment times were calculated and the distributions plotted using box plots. The Medic units were deployed for an average of 100 minutes (blue box plots), thus creating a potential shortage when there are large emergencies. The location and frequency of the 911 calls were mapped. By mapping the location of the 911 calls we identified that over 80 percent of non-emergency calls are from 'loyal' locations, institutions, senior homes, and rehabilitation centers.

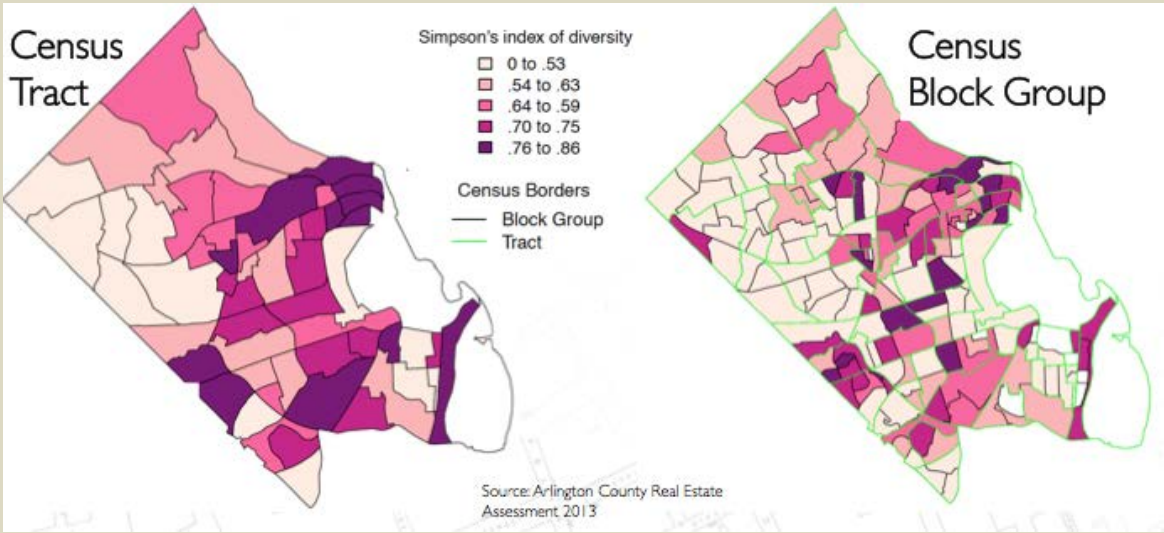
Community Learning: By improving the fire chief's awareness of when and why medic units are out, he can more efficiently anticipate and deploy units and resources. As a consequence, the fire chief reallocated nurse practitioner resources to meet with staff at "loyal" locations to identify what non-emergency 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 a drop-in call volume in the area of the "loyal" locations and a rise in non-911 services.

Example 2. Housing Data

The Census Bureau Director asked us if it is possible to use local data sources to replace or enhance data from their American Community Survey (ACS).

Statistical Learning: SDAL linked the Arlington County real estate assessment records with the ACS. The local housing values were mapped and used as a surrogate for wealth and diversity. This visualization provided a measure of heterogeneity not possible to discern with the ACS data alone.



Community Learning: The local property data provided the geographical granularity needed to address local school and housing issues in an efficient and timely manner. In his letter, Robert Duffy, Planning Director, Arlington County Department of Community Planning, Housing and Development, notes that they would not have undertaken the research using their data if they had not collaborated with SDAL. Duffy stated,

> "We have worked with the Virginia Tech Social and Decision Analytics Lab (SDAL) to provide housing data and insights for their Census Bureau project. We found their results using local housing data to be quite helpful in our work. Their work demonstrated that there is more heterogeneity and diversity in our neighborhoods than can be seen using the American Community Survey. Their efforts also conveyed how bringing data and skillsets from across departments within the organization can yield new tools and insights for solving complex policy problems. Using these results, we are pursuing interdepartmental datasets and collaboration that we would have never been considered without the support of the SDAL."

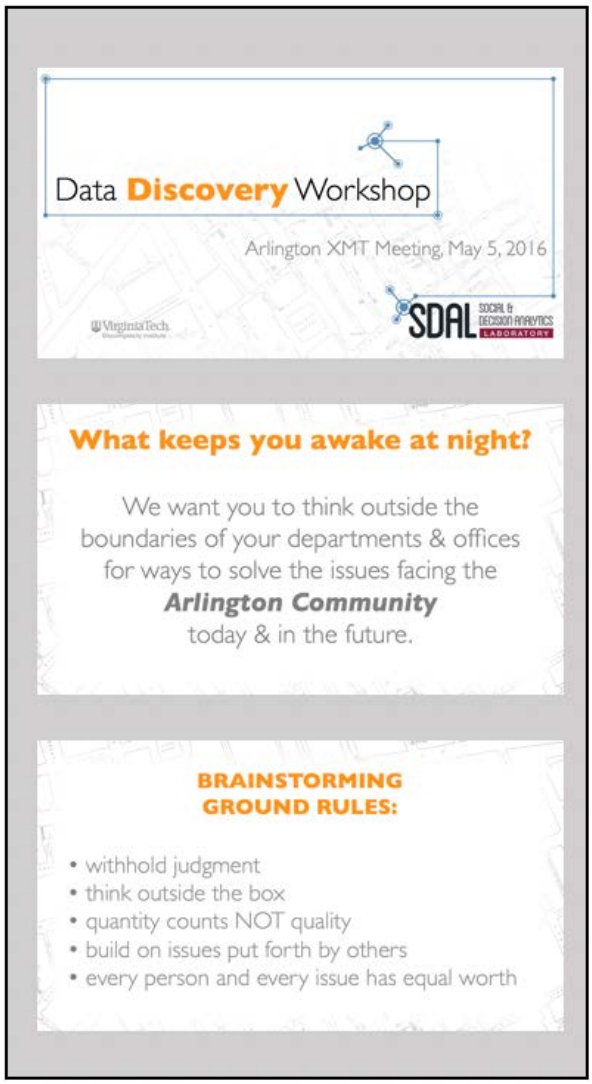
Future Work: These demographic indicators are continually being updated and are being used in projects to identify and describe the underserved residents of Arlington.

Engaging the Community

One of the early steps in the CLD3 process is to bring community leaders together to identify and define those issues that keep them awake at night and to identify the data sources that can provide

insights into these issues. In Arlington, this was accomplished by bringing leaders together for a Data Discovery Workshop. The workshop was structured to breakdown the boundaries among governmental bureaucracies through a dialogue process used to identify issues and data sources that span their boundaries. 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 address the risk factors that align with their mission.

In the Data Discovery Workshop participants were placed into small groups to discuss the local issues that keep them awake at night, no group was allowed to have more than one representative from a department or office. This provided the opportunity to forge new relationships. The issues that were voiced from the groups included:



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?

Potential Data Sources: police, hospital, human services, EMS, homeless survey from MWCOG, community outreach centers, public assistance

Issue 2: What citizens are we reaching, what citizens are we not reaching? How can we identify and reach the underserved population?

Potential Data Sources: library, parks & recreation, citizens at council meeting minutes, community outreach centers, public assistance, civic associations

Issue 3: What kind of facilities do we need?

Potential Data Sources: surveys conducted by the various departments

Issue 4: How do you provide information to diverse populations in their neighborhoods?

Potential Data Sources: civic associations, churches

Issue 5: Is the feedback we are getting representative of the Arlington population? Who are we not hearing from, who are we not reaching?

Potential Data Sources: voting records, citizens at council meeting minutes

Issue 6: How do you consolidate systems? How do you use these systems to identify challenges that may not be apparent?

Potential Data Sources: all Arlington data

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. To address the issues raised in the workshop, their administrative data sources must be liberated and linked across departments and offices.

The action items from the workshop were to create a joint steering committee with Arlington, to identify the existing administrative data sources, and to work with the Arlington lawyer to identify and address the barriers to accessing the administrative data sources. These action items have been completed and CLD3 research is underway.

Data Science for the Public Good Program

With the ever-increasing demand for data scientists in the private sector, attention to the development of a data savvy workforce willing to join civil service is critical. We have taken the initiative to develop a program that exposes college students early on in their education to a career in the public sector through the Data Science for the Public Good (DSPG) program (https://www.bi.vt.edu/sdal/news/student-fellows-in-national-capital-region-learn-to-apply-data-to-solutions). The program serves as 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. The inaugural DSPG was held this summer at the SDAL to teach student fellows from across the U.S. how to sift through vast amounts of information related to public safety, employment, and the provision of services to discover how communities can become more efficient and sustainable. Through the lenses of statistics, social science, and data science research, DSPG students learn to integrate *all* available data resources in order to:

- Identify pressing issues through direct engagement with government and community leaders.
- Develop mechanisms to assist decision-makers in framing their large-scale policy questions and identifying data sources which can be leveraged to address these issues at the local, state, and federal levels.
- Create a two-way data pipeline to give local leaders a direct link to cutting-edge scientific analyses and researchers easier access to federal, state, and local data flows.

The three case studies discussed in the next sections include participation from the 2016 DSPG fellows.

1C Arlington Case Studies

Three CLD3 projects with Arlington are discussed in detail. First is an Evaluation of Open Data Portals to support the launch of Arlington's Open Data Portal. Next is the Emergency Alerts program, whose goal is to better understand the subscriber behavior, and to develop and implement strategies to get more (or the right number of) individuals to subscribe. The final case study is the Operation FireSafe Initiative, whose goal is to ensure all single family homes in Arlington are equipped with smoke alarms.

Evaluation of Open Data Portals

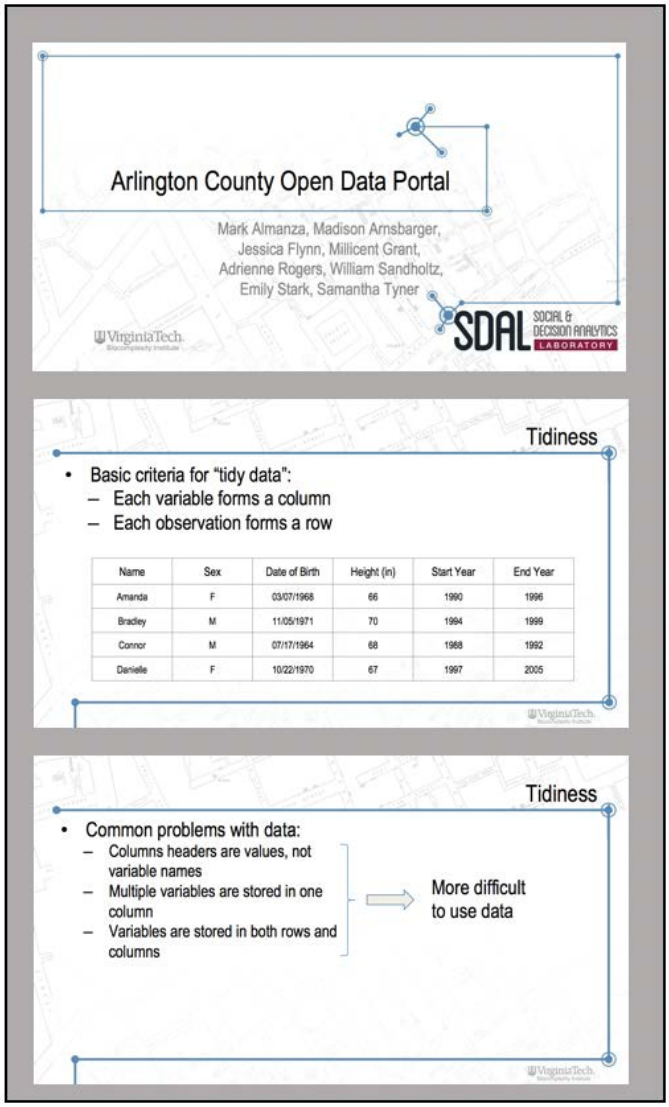
The 2016 DSPG fellows established criteria for a user-friendly open data portal by reviewing literature and industry standards. These criteria suggest a number of best practices for user interface and data structure, which any locality creating and maintaining an open data portal should consider. The criteria are divided into two broad categories with the subcategories listed in parentheses:

1. User Interface (organization, navigation, and design)
2. Data (tidiness, accessibility, documentation, and visualization)

Both categories are evaluated on the basis of accessibility to a broad audience, and applicability to a wide range of uses, including statistical analysis.

They then reviewed Arlington's open data portal and those of San Francisco, Chicago, and Washington, D.C, to find examples that do and do not meet the criteria. The fellows presented their findings to the Arlington chief information officer and deputy.

Their work was described in an article for AMSTAT News (http://magazine.amstat.org/blog/2016/08/01/portals-aug16/).



Emergency Alerts: Leveraging Community Partnerships to Improve Mass Emergency Alert Notification

Introduction

The Arlington County Office of Emergency Management (OEM) runs the "Arlington Alert" system, a public-facing product that provides mass notification of emerging safety hazards and concerns. Their goal is to use the tool effectively, but they struggle to answer questions, such as "How many alerts can be sent before people reach notification fatigue?" and "How can the alerts reach more people to ensure enough people get the information?" At a higher level, they want an overall better understanding of their system, customers, and alert messaging behaviors.

OEM recognized that data analytics was necessary to answer these questions, especially given the County's emphasis to use evidenced-based policy and data-driven decision making. They partnered with SDAL to leverage the lab's social science, statistical expertise, and data analytics skills.

Research Study Design

OEM and SDAL embarked on a 10-week study during which the research team and OEM staff met weekly to discuss the project, provide updates, and address any questions or issues. The process began by formulating specific research questions and defining the scope and approach of the project. There were three main research topics:

1. Understand the demographic characteristics of the people subscribing to their alert system and their geographic distribution across the County. OEM wanted to use this information to target outreach to areas of the County that had low enrollment.
2. Understand whom OEM should target for outreach to maximize the dissemination of information throughout the community. OEM recognized that enrolling all residents and workers in Arlington Alerts was unrealistic, but they wanted to target people who would be most likely to share the alert information with others.

Understand how people respond to alerts and at what point they reach notification fatigue and un-enroll from the system. OEM realized that a significant number of people were starting to opt out of their system and wanted to use this information to prevent over-messaging.

OEM chose these three areas because they felt that this information would allow them to make meaningful adjustments to their system and protocols to improve the overall effectiveness of Arlington Alerts. OEM provided SDAL with data from the Arlington Alert system and municipal demographic data. The lab also utilized American Community Survey data from the Census Bureau, which was merged with the County data. This information was mapped geographically and used to address research topics 1 and 3. SDAL also conducted an in-depth literature review to identify the characteristics of individuals who tend to share information. These characteristics were developed into a psychological profile which was then matched to professional roles using an empirically based and widely researched career tool, the Strong Interest Inventory®. This information was used to address research topic 2.

Results

Mapping the raw distribution of subscribers revealed that the highest concentration of subscribers occurred along one of the densest residential and commercial corridors (Figure 1). This finding was expected as more enrollments occur where there are more people. SDAL then weighted the data by population density (as measured by the American Community Survey) in order to visualize "subscribers per capita" in any given census tract (Figure 2). This revealed much richer information. The County's northern census tract had a much higher concentration of subscribers than the southern census track. By merging the County demographic data and American Community Survey

data with the alert data, SDAL was able to determine the characteristics of the areas with the lowest enrollment (race, income, education level, primary language, etc.). These areas tended to have lower levels of education and income as compared to the rest of the County.

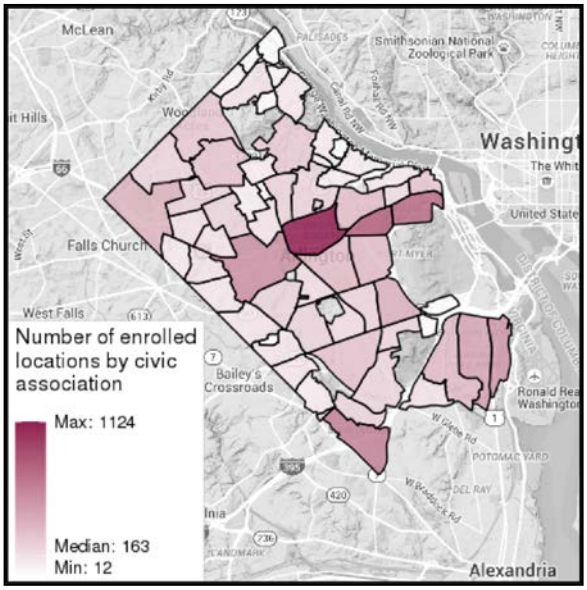


***Figure 1**. Enrollment by civic association. There is a high number of enrollment counts in the Metro Corridor, which has a large population.*

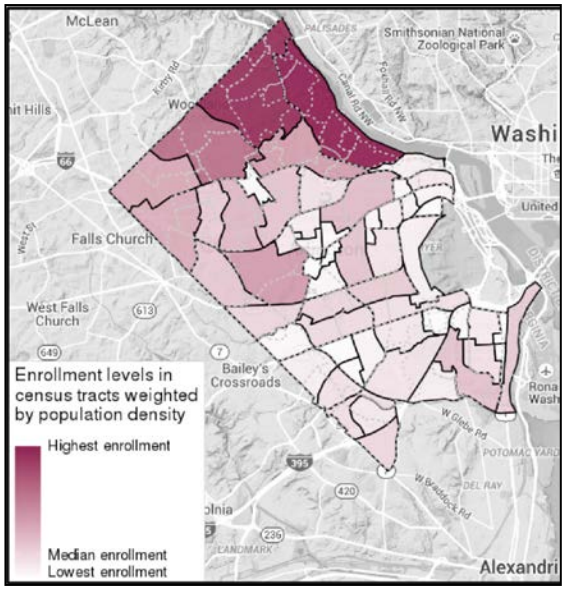


***Figure 2**. Enrollment concentration by census tract accounting for population density. Gray lines are civic associations, black are census tracts. When weighted by population density, North Arlington has higher enrollment than South Arlington.*

The literature review revealed that people with the following traits are more likely to share information: openness to experience, agreeableness, conscientiousness, extroversion, a propensity to trust others, and an emotional commitment to the organization. The Strong Interest Inventory® results indicated that the personality type associated with sharing information tend to work in the following types of careers:

- Teaching & Training
- Restaurant, Entertainment, and Sports Management
- Counseling (school, religious, psychological)
- Facility Management
- Healthcare

SDAL merged multiple Arlington Alerts datasets (timing and content of alert messages, subscribers' account activity, subscribers' text replies to alert messages) to determine how subscribers respond to alerts and to assess notification fatigue. SDAL analyzed the word frequency in all of the alerts sent out in the past two years and found that "warning", "watch", "severe", and "issued" were the most common (Figure 3). These words correspond with National Weather Service's automatic weather alerts, which are automatically sent out through Arlington Alerts. Next, SDAL analyzed the word frequency of the alerts that prompted a subscriber to reply "STOP", which indicates that the individual wanted to un-enroll from the system. The most common words were (Figure 3):

- Traffic
- Open/Reopened/Closed

- Street
- Lanes
- Final

OEM recognized that these words are used in alerts about road closures (usually from accidents, water main breaks, and special events).

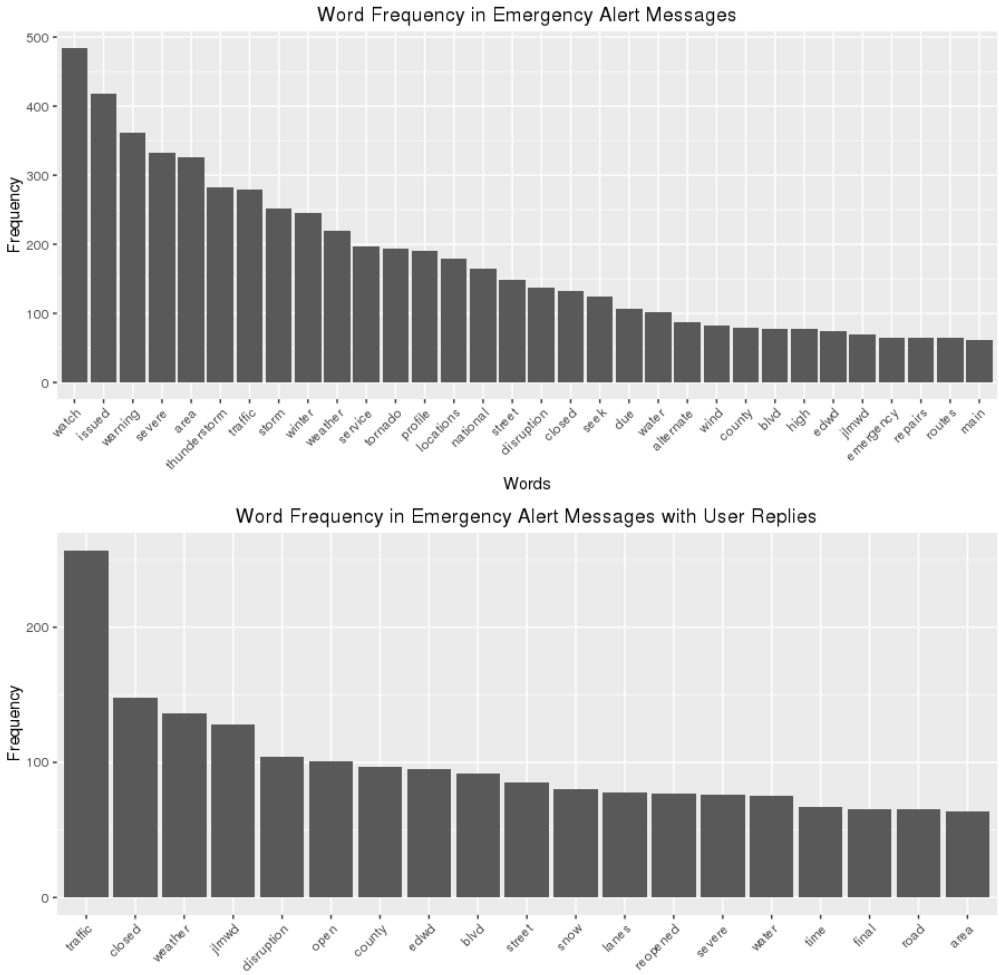


Figure 3. *The top chart shows the word frequency within emergency alert messages. The bottom chart shows the word frequency of messages to which users replied "Stop." Frequencies are only for those words used > 60 times within the past 2 years. The most common words are related to weather and traffic alerts.*

<u>Analysis of Results and Policy Actions</u>

Based on the results from the per-capita subscriber distribution assessment, OEM is now targeting their outreach efforts to the geographic areas that have low enrollment. They are using the demographic characteristics of these areas to help inform the means and methods to encourage enrollment. For example, because these populations are generally lower income and are more likely to rely on their mobile device (rather than a desktop PC) to register, OEM is working with the vendor of their alert system to improve their mobile interface.

OEM is also using the individual characteristics and job roles associated with information sharing behaviors to target outreach efforts. They are partnering with the local school system, local hospital, parks and recreation department, faith based community, and Chamber of Commerce to encourage

teachers, counselors, pastors, healthcare providers, coaches, restaurateurs, and small business owners to enroll in the system. The expectation is that these individuals will be the most likely to share the alerts they receive with their friends, neighbors, family, and coworkers, thereby increasing the dissemination of alerts throughout the community.

Based on the vocabulary assessment related to un-enrolling behaviors, OEM is changing how they send alerts about traffic disruptions. This requires them to balance operational needs with notification fatigue and they are considering several options to limit extraneous and unwanted messages. To do this, they are evaluating the possibility of omitting "final" or "reopening" messages, restricting alerts to primary roadways during rush hours and directing subscribers to local news sources for follow up information.

Evaluations

1. SDAL and OEM will work with the vendor to create and assess a new mobile interface to the alert system based on panel data. Panels of Arlington residents with be constructed based on demographic characteristics such as education, income, English proficiency, and age, to rate the mobile apps on functionality, information quality, and aesthetics. This information will be used to guide the development and selection of the new mobile app. The selection of the mobile app(s) will be based on a statistical analysis of the multivariate panel data.
2. After the development of a new mobile app(s), SDAL and OEM will evaluate enrollment numbers before and after the targeted outreach. The target outreach will be directed at randomly selected census block groups and will highlight the improvements to the alert system. The post-assessment of the outcomes will be conducted every month for six months after the target outreach. By assessing enrollment/un-enrollment behaviors over time – before and after these changes were implemented - it may be possible to determine if the targeted outreach efforts and alert adjustments were effective as a function of the demographics of the census block group. Effectiveness will be measured by an inferential analysis of the enrollment/un-enrollment numbers over time using census block group is a blocking factor.

Conclusion

Overall, the partnership between OEM and SDAL resulted in actionable data analytics that informed several critical adjustments to OEM's policies and procedures. By taking into account the alert types, the differences between them, and the reaction of the enrollees, SDAL helped transform data into knowledge and OEM is now using that knowledge to inform their alert policy and improve the services they deliver to the community.

Future Work

SDAL and OEM are looking into the idea of using simulations based on the principles of herd immunity to better understand the proportion of residents who have to receive the alerts so that all residents are alerted. SDAL is also helping OEM become self-sufficient in conducting data analyses with their alerts data.

FireSafe Project: Residential Smoke Alarm Need in Arlington County

Introduction

In 2014, U.S. fire departments responded to an estimated 367,500 residential fires that resulted in 3,275 civilian deaths and 15,775 injuries and an estimated $6.8 billion in property loss in home fires (Haynes, 2015). According to the same report, there was a civilian fire death every 2 hours and 41 minutes and a civilian fire injury every 33 minutes in 2014. Home fires caused 2,745, or 84%, of the civilian fire deaths. In Virginia, one civilian was killed or injured by fire every 22 hours in 2015, and the estimated total loss associated with fires was $242 million (NFIRS, 2016).

The use of smoke detectors has been shown to be an effective, reliable, and inexpensive method of providing early warning in residential fires (Marshall et al., 1998). It is found that 40% of fires occur in homes without alarms and 70% of home fire deaths occur in homes with either no smoke alarm or homes in which none of the smoke alarms sounded (Ahrens, 2004). Smoke alarms installed and maintained properly can decrease the risk of death by 40%-50% (Ahrens, 2004). In North Carolina, the absence of a smoke detector posed a two- to three-fold risk of death in the event of a fire (Runyan et al., 1992).

In June 2015, Arlington County Fire Department (ACFD) initiated the Operation FireSafe program to increase the number of residences in the county with smoke alarms. The firefighters across the county knock on doors of Arlington residences on Saturdays to offer free smoke alarm inspections and installations. Since the program began, ACFD visited 5,623 single-family homes countywide, 32 percent of the homes had smoke alarms installed. Building a model to identify homes in need of an alarm may increase the efficiency of the program.

In this project, we combine the information collected by the ACFD through the Operation FireSafe program and the housing characteristics of single family homes in Arlington County from the local real estate tax assessment data to develop a model (i) to identify home characteristics that can be used as predictors of smoke alarm absence and (ii) to compute the likelihood of having a smoke alarm for all single family homes in Arlington County to inform future Operation FireSafe visits.

Related Work

Prior studies of smoke detector prevalence rely heavily on survey data or questionnaires for smoke detector and household information. Roberts (1996) conducted interviews with 4043 randomly sampled households in the United Kingdom, gathering data on smoke alarm use and socio-demographic characteristics of the household. Ballesteros and Kresnow (2007) utilized the ICARIS-2 survey from the Center for Disease Control and Prevention's (CDC) National Center for Injury Prevention and Control. Likewise, Forjuoh et al. (1997) analyze questionnaire data from the CDC's Behavioral Risk Factor Surveillance System and the Pennsylvania optional module that covered smoke alarm ownership. Other studies conduct inspections for the smoke detector prevalence data, but rely on resident provided information for housing data. Sidman et al. (2011) combine a home inspection of smoke detectors with a questionnaire on household characteristics. Similarly, McLoughlin et al. (1985) conduct a home inspection of smoke detector presence and rely on a 98-question interview form and tax assessor's lists.

By contrast, our study makes use of the data on smoke detector presence from the Arlington County Fire Department's Operation FireSafe program. During the summer months, Operation FireSafe crews went door to door offering free smoke detectors and fire safety checks. If the resident accepted the services, crew members recorded the number of existing functional smoke detectors in the home. The program generated data on smoke detector presence in 811 single family homes in Arlington County. For data on housing characteristics, we used the 2013 Arlington County local real estate tax assessment data. By merging the Operation FireSafe and local real estate tax assessment

data, we produced a more reliable and comprehensive dataset than is possible through surveys. A similar data-driven effort was implemented in New Orleans (Mayoral Press Release, 2016). Our research is the first to focus on Arlington County.

Much of the literature on household smoke detector prevalence focuses on correlative predictors of smoke detector presence. Ballesteros and Kresnow (2007) use chi-square tests to test for association between fire safety practices and demographic characteristics. In addition, McLaughlin et al. (1985) apply chi-square tests and linear logistic regressions to estimate smoke detector prevalence for households in Montgomery and Fairfax Counties. Roberts (1996) and Forjuoh et al. (1997) use a logistic regression to assign odds ratios to various household characteristics. Sidman et al. (2011) use a Poisson regression with the Huber/White/sandwich variance estimator to estimate risk ratios and 95% confidence intervals for smoke detector coverage.

Our research extends beyond estimating probabilities of smoke detector presence for household factors. Since we merge ACFD smoke detector data and local real estate tax assessment data, we are able to create a model to assign probabilities of smoke alarm presence at the level of the housing unit, whereas previous research only links general household characteristics (age of the home, income and age of the occupant) to smoke alarm prevalence. In addition, this is the first time a Bayesian logistic regression with conditionally autoregressive spatial effects have been used to study this problem.

Data Sources

Operation FireSafe. ACFD provided us with the results of their Operation FireSafe visits to 5,623 single-family homes. The dataset includes time and date of the visit, address, outcome, number of alarms installed, and number of working alarms in the home prior to the visit. 1,733 of the visits resulted in contact with the occupant and the collection of information on the presence of smoke alarms. Almost one-third of these homes did not previously have smoke alarms installed. Figure 1 locates the 1,733 visits made by the ACFD; homes with a smoke detector are identified in green and homes without smoke alarms in red.

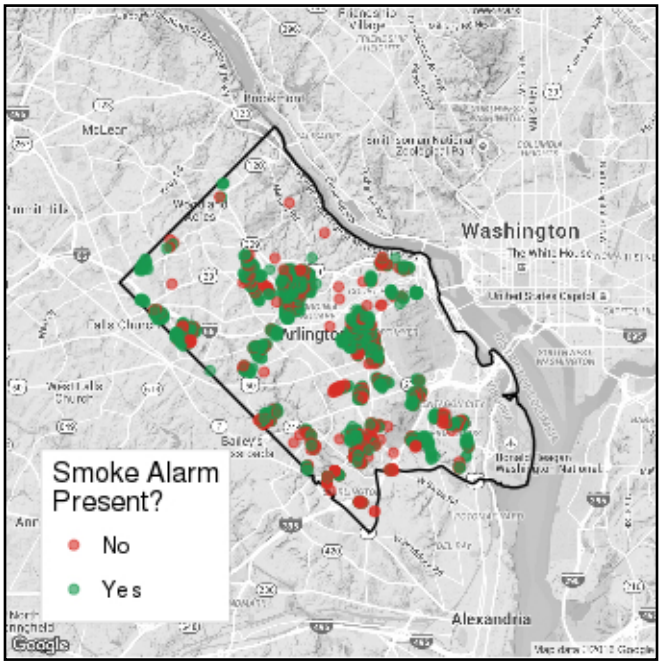


***Figure 2**. Location of the 1,733 Operation FireSafe contacts with residents of single family homes by the Arlington Fire Department.*

Local Real Estate Tax Assessment Data

The local real estate tax assessment data contains information, such as house value, size, age, number of bedrooms, and ownership, for 60,343 single-family housing units in Arlington County.

Combining these data with the ACFD data allows us (i) to identify the characteristics that can be used as predictors of smoke alarm need and (ii) to compute the likelihood of having a smoke alarm for all single-family housing units in Arlington County.

The Model

We constructed a Bayesian logistic regression model (Equation 1) with conditionally autoregressive spatial effects to identify the predictors of smoke alarm need. Formally,

$$
\begin{aligned}
&Y_i \sim \text{Bernoulli}(p_i) \qquad && \mu_i = \mathbf{X}_i^T \boldsymbol{\beta} \\
&\log\left(\frac{p_i}{1-p_i}\right) = \mu_i + \phi_i && \boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{I}_{1000}) \\
&\phi_i | \boldsymbol{\phi}_{-i}, \mathbf{W}, \tau^2, \rho \sim N\left(\frac{\rho \sum_{k=1}^{K} w_{ik}\phi_k}{\rho \sum_{k=1}^{K} w_{ik} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{k=1}^{K} w_{ik} + 1 - \rho}\right) \\
&w_{ik} = \begin{cases} 1 & \text{if houses } i \text{ and } k \text{ are in the same census block group} \\ 0.5 & \text{if houses } i \text{ and } k \text{ are in neighboring census block groups} \\ 0 & \text{otherwise} \end{cases} \\
&\tau^2 \sim \text{Inverse-Gamma}(5, 40) \\
&\rho \sim \text{Uniform}(0, 1)
\end{aligned}
$$

***Equation 1**. Bayesian Logistic Regression Model*

The response variable is binary with $Y_i = 0$ indicating that housing unit *i* has no smoke alarm (ACFD installed one or more alarms) and $Y_i = 1$ if one or more smoke alarms are present. The covariates of interest, X_i, are:

- the natural log of the value of the home (a proxy for income)
- the age of the home
- an indicator of whether or not the residence is a condo
- the number of bedrooms (a proxy for household size)
- an indicator of whether or not the residence is owner occupied.

<u>Findings</u>

Table 1 lists the parameter estimates and credible intervals. We observe that the value and age of the house are significant predictors of whether or not a housing unit has a smoke alarm.

Table 1. The parameter estimates and credible intervals of the Bayesian logistic model.

Parameter Value	Posterior Median	95% Credible Interval
Intercept*	-30.12	(-59.60, -20.39)
log(Home Value)*	2.45	(1.67, 4.83)
Age of the Home*	-0.0109	(-0.0257, -0.0022)
Condo Indicator	-0.48	(-1.58, 0.022)
Number of Bedrooms	-0.12	(-0.39, 0.09)
Owner Occupied Indicator	0.02	(-0.61, 0.65)

*Parameter is statistically significant at the 0.05 level.

Incorporating these values into the model and using the information in the local real estate tax assessment data, we can then estimate the probability that a housing unit has a smoke alarm for all single family homes in Arlington County. The factors that affect the predicted probabilities are (i) the *covariates of interest* and (ii) the location (census block group) of the home.

Because of the spatial dependence included in the model, the predictions are made for each single family residential unit in the local real estate tax assessment data source that are:

- in census block groups that appeared in the original data or in a census block group neighboring those in the original data; and
- with no missing covariate values and/or census block group information.

Consequently, with local administrative data predictions can be made at the level of the 49,178 residential units.

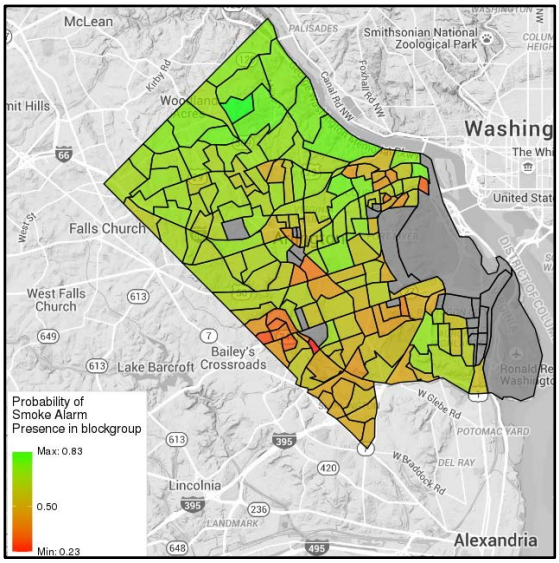


***Figure 2**. Aggregate results by census block group. The figure illustrates the average probability of a home containing a smoke alarm within a census block group in Arlington County.*

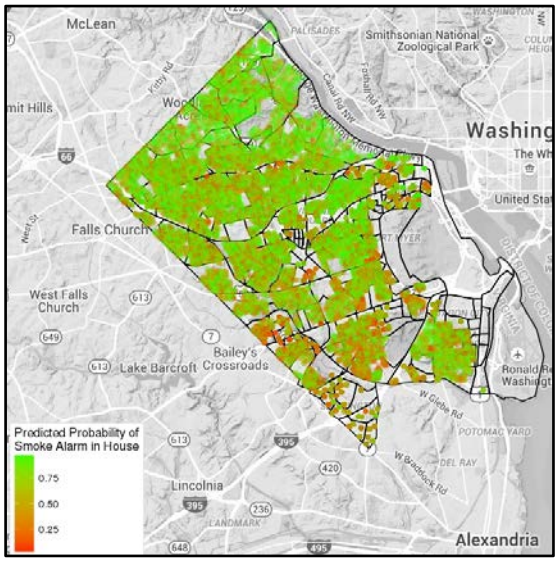


***Figure 3**. The predicted probabilities of smoke alarm presence for 49,178 single family residential units in Arlington County. The lower the predicted probability (red areas) the more likely the unit lacks a smoke alarm.*

Evaluations

1. The ACFD employees will target their visits to the single family homes with the highest probability of having no smoke alarm. The results from these visits will allow SDAL to conduct an assessment of the initial model predictions and re-evaluate the model predictors; additional predictors will be evaluated in light of the new data. The model will be updated and a new set of predictions supplied to the ACFD.
2. The data from future ACFD visits will be used to update the model parameters and develop a new set of probabilities for the 49,178 single family homes. This iterative process will continue until all homes have been visited by the ACFD. A final assessment of the models will include calculating the error rates.

Conclusion and Next Steps

By focusing on the homes with a low probability of an installed smoke alarm, the ACFD can better target future Operation FireSafe visits. The next step is to incorporate other data sources (e.g., American Community Survey) into the model to capture the effects of socio-demographic characteristics.

References

1. Haynes, H. J. (2015). *Fire loss in the United States during 2014*. National Fire Protection Association. Fire Analysis and Research Division.
2. Ahrens, M. (2004). US experience with smoke alarms and other fire detection/alarm equipment. *NFPA, Quincy, MA*.
3. Reported from the National Fire Incident Reporting System (NFIRS) as of April 28, 2016. http://vdfp.virginia.gov/fire_data_statistics/index.htm
4. Marshall, S. W., Runyan, C. W., Bangdiwala, S. I., Linzer, M. A., Sacks, J. J., & Butts, J. D. (1998). Fatal residential fires: who dies and who survives? *JAMA*, *279*(20), 1633-1637.
5. Runyan, C. W., Bangdiwala, S. I., Linzer, M. A., Sacks, J. J., & Butts, J. (1992). Risk factors for fatal residential fires. *New England journal of medicine*, *327*(12), 859-863.
6. "City Launches New Smoke Alarm Initiative Based on Data Analytics", Mayoral Press Release, City of New Orleans http://www.nola.gov/mayor/press-releases/2015/20150331-smoke-alarm-data-campaign/.
7. Roberts, I. (1996). Smoke alarm use: prevalence and household predictors. *Injury prevention*, *2*(4), 263-265.
8. Ballesteros, M. F., & Kresnow, M. J. (2007). Prevalence of residential smoke alarms and fire escape plans in the US: results from the Second Injury Control and Risk Survey (ICARIS-2). *Public health reports*, 224-231.
9. Forjuoh, S. N., Coben, J. H., Dearwater, S. R., & Weiss, H. B. (1997). Identifying homes with inadequate smoke detector protection from residential fires in Pennsylvania. *Journal of Burn Care & Research*, *18*(1), 86-91.
10. Sidman, E. A., Grossman, D. C., & Mueller, B. A. (2011). Comprehensive smoke alarm coverage in lower economic status homes: alarm presence, functionality, and placement. *Journal of community health*, *36*(4), 525-533.
11. McLoughlin, E., Marchone, M., Hanger, L., German, P. S., & Baker, S. P. (1985). Smoke detector legislation: its effect on owner-occupied homes. *American Journal of Public Health*, *75*(8), 858-862.

1D Continuing Engagement with Arlington: Future Work

SDAL and VEC are embarking on three new initiatives with Arlington involving the police, parks & recreation, and the library:

- the Police Department wants to implement programs to target at-risk youth;
- the Department of Parks & Recreation would like to implement a policy that ensures program fees are not a barrier to participation; and
- the Library Department is interested in understanding the needs of different underserved populations, such as those aging-in-place and non-English speakers, in order to better serve these populations.

The first task in each initiative is to identify the target population using available data sources. For the Police Department the target population is at-risk-youth; for the Department of Parks & Recreation the target population is residents who do not enroll in park programs or enroll at a reduced fee; and for the Library Department it is those residents who do not use the library, which the library staff hypothesizes could be the elderly and non-English speakers. What these three new initiatives have in common is that they all involve identifying and reaching the underserved residents in Arlington. The data sources and analyses from each department will provide insights into the target populations of the other departments; the resulting interventions have the potential to impact all three departments. For example, if programs that target at-risk youth want to involve parks & recreation programs, they need to be affordable. Similarly, if library programs want to involve at-risk youth, the librarians must understand the needs of this population.

Police Department: At-Risk Youth

The Police Department wants to implement programs to target at-risk youth. The project with the Police Department will involve a review of the literature on how at-risk-youth are defined and identified, followed by data discovery – what are the available data sources that can be used to quantify the at-risk indicators and use them to identify the at-risk youth in Arlington. In general, the at-risk indicators can be grouped into three categories, those pertaining to the individual, the family, and the community. Our challenge is that data at the individual level, such as school records on academic achievement, attendance, and disciplinary infractions, are de-identified before being released to researchers, if they are released at all. We anticipate needing to rely on family indicators such as household income, parental education and employment, family structure, race/ethnicity, welfare receipt, and the number of older siblings; and community indicators such as neighborhoods with a lower percentage of workers holding managerial or professional jobs and higher rates of teenage childbearing. Data from the Department of Parks and Recreation project on reduced fee households will be useful for identifying the location of households that have qualified for a reduction based on WIC, TANF, SNAP, or other DHS programs. Part of the data discovery process will involve the Police Department and their data sources. We will also use data from the American Community Survey, the Arlington ATRACK data on affordable housing, and the local real estate tax assessment data; the data discovery process will identify others. The goal of this first step is to map the at-risk indicators across Arlington and to develop a spatial algorithm that combines the indicators to identify areas with a high probability of youths that are at-risk. The validity and credibility of the resulting algorithm will be evaluated in collaboration with the Captain of the Police Department.

The next steps of the project will involve developing a strategy to identify at-risk youth programs that have a demonstrated success based on the empirical evidence of other researchers and implementing those considered appropriate for Arlington. Identified programs will be shared with the Police Department and other community leaders to select those programs with empirical evidence that are applicable to the contextual constraints of Arlington. The Arlington extension

agents will be consulted regarding their 4-H youth programs[2] and their work with the underserved communities in Arlington.

We anticipate using an adaptive experimental design that will control for the confounders of gender, race/ethnicity, and various family indicators. An adaptive design allows for planned modifications to one or more aspects of the intervention based on analysis of the data from subjects in the intervention. Specifying a proscribed monitoring and data analysis schedule prior to the start of the intervention, allows us to evaluate progress over time and make modifications without undermining the statistical integrity and validity. The specific design will be based on the chosen interventions and our ability to define and access the target population. The response variables will be measured over time and selected based on the literature review and discussions with the Captain of the Police Department as to their applicability for Arlington.

Department of Parks & Recreation and Department of the Library: Users and Nonusers

Both the project with the Department of Parks & Recreation (DPR) and the Library Department (LD) involve identifying the residents who use and do not use their services. In both cases, the goal is to find out the impediments to using their services and to implement changes that will engage the non-users. Within the non-user group, the DPR is concerned with income demographics, whereas the LD is concerned with age and race/ethnicity demographics. It is likely that these are not discrete categories.

The data discovery process with the DPR has been completed and they have provided us with the demographic information of the patron, patron address, location of the program enrolled in, type of program, cost of the program, and the amount of program fee reduction if applicable. We are in the process of geocoding all the patron addresses and will then map the location of the users by level of fee reduction, 25%, 50%, 75%, and no reduction. This information will be integrated with data from the American Community Survey and local real estate tax assessment data, which will provide a more detailed demographic picture across Arlington. The Arlington extension agent, Jennifer Able, will be working with us on this project. Part of her work involves teaching financial workshops in the service enhanced housing programs of the nonprofit, Arlington Partnership with Affordable Housing. Once the geospatial overlays have been constructed and the descriptive statistics calculated, we will meet with the DPR Director of Finance and Budget, and the DPR staff to discuss next steps, including which other Arlington departments and offices should be consulted. In this initial work we are identifying the location of users (reduced fee and full fee) and non-users, and mapping out their demographic profiles in order to target an outreach program(s) aimed at gaining an understanding of the financial constraints to program enrollment. This information will be valuable to other departments and offices that offer services, especially those that are fee-based. We will also evaluate transportation constraints (e.g., public versus private, length of travel time and distances, the synergy of the public transportation schedule and times of the DPR programs, physical barriers to bicycles, etc.) and document the commercial and non-profit competition to the DPR programs. Based on the statistical learning process, DPR may choose to implement an outreach program; in that case we will be able to inferentially evaluate the success of the program by the various demographic confounders. In the future, we will work to integrate the information from other departments and offices to create a more accurate map of resident participation in government services.

The project with the LD will begin with the data discovery process with the Library Director and Deputy Director. The first step is to define the "underserved" target population, which the library staff hypothesizes may be those aging-in-place and non-English speakers. We anticipate integrating

[2] These positive youth development programs have been evaluated using statistically designed studies and inferential analyses by Tufts University (http://ase.tufts.edu/iaryd/researchPositive4HPublications.htm) in cooperation with 21 Land Grant Universities.

the data from the LD with the data and lessons learned from the DPR project in this first step. We will also be reaching out to local organizations and institutions that provide services to aging-in-place adults (e.g., local agencies on aging, senior centers, AARP chapters, faith-based organizations, etc.) and non-English speakers (Just Neighbors, Ethiopian Community Development Council, Inc., Northern Virginia Family Service, faith-based organizations, etc.).

1E What we have learned from our work with Arlington

- Most of the work involved with using local data involves cleaning and linking the data sources. The key to successful linkage is GIS. There is a need to standardize the address fields across data sources; one solution would be to adopt the postal address standards across data sets. See http://pe.usps.gov/text/pub28/.
- Some local government offices and departments contract with vendors to collect and store the data collected during routine government operations. As a consequence, it can be hard for local governments to get their own data from the vendor and even more difficult to change the user interface used to collect the data. In Arlington, when this problem was identified, we worked with an Arlington lawyer who is now changing the agreements to make sure that the Arlington government offices and departments have full access to the underlying data, not just to the reports that the systems produce.
- The public and local employers need to be educated on the benefits of their data. For example, local employers fill out quarterly employment forms required by the state. If they knew these employment numbers were used to allocate transportation funding at the local level, they might take more care in filling out the form and in providing more accurate information.
- Government officials need to be educated as to what "big data" can and cannot do. Big data do not preclude the need for the application of the scientific method, for formulating a hypothesis, understanding the limitations of the data with regard to the representativeness of the population and the validity of the inferences.

1F What we have accomplished so far

Posters from the 2016 Data Science for the Public Good 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
https://www.bi.vt.edu/sdal/careers/call-for-students/student-poster-presentations

- Arlington County Emergency Alerts: Enhancing Enrollment and Effectiveness
- Arlington - Criteria for an Excellent Open Data Portal
- Arlington Fire/EMS 911 - Residential Smoke Alarm Need
- Arlington Fire/EMS 911- Response Times
- Arlington - Tracking Employment Flows Using Non-Survey Data
- DOD Policies and Locations - Characterizing Factors that might Affect Military Enlistment, Retention, and Attrition
- DOD - The Use of Social Media by the Army Population
- HUD/Census - Using Synthetic Populations to Assess Identifiability Risks in the American Housing Survey
- NSF - Open Source Software and Innovation
- NSF - On Variations of Computational Models of Attitude

- SLDS - Assessing the Influence of Career and College Readiness using Kentucky State Longitudinal Data Systems
- SLDS - School Assignment Policies and Race: A Comparison of Jefferson Count, KY and Mecklenburg County, NC School Districts
- VLDS - Characterizing Administrative Data Quality: A New Tool Developed Using Virginia Longitudinal Data Systems

Posters for the 2016 Association of Public Policy and Management Fall Research Conference, Washington, D.C.:

- Using Assessment Data to Characterize Housing Diversity
  SDAL: Gizem Korkmaz, Emily Molfino, Sallie A. Keller, Aaron Schroeder, Daniel H. Weinberg
- An *in-silico* Platform for Environmental Coupling
  SDAL: Bianica Pires, Gizem Korkmaz, Dave Higdon, Sallie Keller, Bryan Lewis, Aaron Schroeder
  Rice University: Katherine Ensor

Relevant Publications by SDAL Researchers:

- Almazna, A., Arnsbarger, M., Flynn, J., Rogers, A., Sandhotz, W., Stark, E., Grant, M., and Tyner, S., 2016, "Criteria for Open Data Portals Offer Best Practices for Websites," *AMSTAT NEWS*, August 1, 2016. http://magazine.amstat.org/blog/2016/08/01/portals-aug16/
- Keller, S., S. Shipp, and A. Schroeder, 2016, "Big Data and the Privacy Landscape" in *Annual Review of Statistics and Its Application*, Vol. 3, pages 161-180.
- Keller, S. and S. Shipp, 2017 (forthcoming), "Building Resilient Cities: Harnessing the Power of Urban Analytics." In J. Bohland and P. Knox (Eds) *The Resilience Challenge: Looking at Resilience through Multiple Lens*, Charles C. Thomas Ltd. Publishers.
- Keller, S., G. Korkmaz, M. Orr, A. Schroeder, and S. Shipp, 2017 (forthcoming), "The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches" in
  Annual Review of Statistics and Its Application, Vol. 4.
- Molfino, E., G. Korkmaz, S. Keller, A. Schroeder, S. Shipp, and D. Weinberg. 2017 (forthcoming), "Can Administrative Housing Data Replace Survey Data?" Cityscape, (https://www.huduser.gov/portal/periodicals/cityscape.html).
- Ziemer, K. and B. Pires. (In Progress). "A New Lens on High School Dropout: Use of Correspondence Analysis and the Statewide Longitudinal Data System."
- Korkmaz, G., J. Goldstein, and S. Tyner. (In Progress). "Residential Smoke Alarm Need in Arlington."

Appendix 2: Examples of Community Issues and Questions

The following provides examples of questions that communities are asking in Iowa and Virginia. We believe they provide insights into why CLD3 is important to communities of all sizes.

Common trends and issues in Iowa communities

- Significant trends of in-migration in localized areas of the state. In some communities, property tax incentives are used to incentivize settling in individual cities, and housing issues are common. Sudden increases in the K-4 ESL student base have resulted in immediate school district needs. Perry, Iowa (population 8,000) and Marshalltown, IA (population 28k) are examples.
- Aging infrastructure is evident in many areas of the state, coupled with challenges in complying with recent environmental mandated regulations. Multiple proposals have surfaced recently in attempt to address transportation infrastructure shortfalls. Water, wastewater and storm water infrastructure is also a critical challenge in Iowa.
- Education challenges exist in school coverage/school consolidations that may coincide with demographic trends in the state. Additionally, Iowa's overall percentage of college-educated populations is comparatively low.
- Healthcare coverage challenges include mental health, rising premiums, availability of specialized services, a generally aging population/demographic, switch to private providers for Medicare, etc.

Common trends and issues in communities in Virginia

- How do you provide services to those with mental health issues, how do you coordinate services, anticipate needs, end the cycle of incarceration?
- What citizens are we reaching, what citizens are we not reaching? How can we identify and reach the underserved population?
- What kind of facilities do we need?
- How do you provide information to diverse populations in their neighborhoods?
- Is the feedback we are getting representative of the Arlington population? Who are we not hearing from, who are we not reaching?

We are just beginning to explore **three new tightly coupled issues with Arlington.** The underlying populations associated with each of these issues are interrelated and value will be achieved by working with these three issues simultaneously – this is the strength of CLD3. Arlington is a community with varied income distribution seeking to assure that all citizens have access to county programs including youth development, home safety and adult education.

- The **Police Department** wants to implement programs to target at at-risk youth – first we need to ***identify the population of at-risk youth and use this to develop interventions***, followed by rigorous evaluation of policy and practice changes.
- The **Department of Parks & Recreation** (DPR) would like to ***ensure that program fees are not a barrier to program participation.*** We are developing the analyses to map out eligibility across the county and then develop, implement, and evaluate a policy to test an approach to reduce barriers to program fees that are equitable and not a barrier to participation.
- The **Library Department** is interested in understanding the trends and needs of different populations, those aging-in-place, non-English speakers, and underserved populations, ***and how they can serve the varying needs of these groups better.***

Arlington Fire/EMS issues related to ***improved response to 911 emergency calls by citizens*** have been successfully addressed by linking data sources across administrative boundaries to recreate incidents end-to-end in time and space. Answers to these issues are leading to changes in standard operating procedures aimed at gaining efficiencies in resources and time. Going forward, statistical methods that can be used to continuously monitor 911 incidents are being developed. For example, working with Arlington's nurse practitioner, document and continuously analyze how policy is implemented (periodic visits to loyal locations) and impacting changes in 911 uses, such as employee turnover at locations. Using statistical time-

series methods we will monitor changes in call volume from frequent or "loyal customers" and "loyal locations" as input for inferential evaluations.

- What are trends in total incidents?
- What are the trends by hour of day, day of week, month, incident type, unit type (e.g., fire or medic units)?
- What is the impact of weather?
- Are the right resources being deployed for incident types?
- When is 'right of refusal' used and what are the policy implications?
- What is utilization by unit?
- What can be gained through the use of social media to deliver improved service?

Arlington housing examples include the integration of local, state, and federal data sources to identify and ***address youth and adult access to schools and recreational programs.*** We are able to address a range of issues including:

- Can families with children afford to live in Arlington? If no, why not and what policies need to be developed?
- Can the number of school age children who will attend Arlington schools in the future be accurately predicted?
- Are families who live in Arlington able to participate in library, parks & recreation activities, and other services? If not, what are the barriers to participation?

Broader integration of housing data with the socio-economic data from 911 incidents and local, state, and federal data sources, allows for the use of statistical methods to ***better understand and provide context around the confounding factors that affect neighborhoods and the ability to access services***.

- Construct and test hypotheses, e.g. is there a relationship between age of the home and the number of 911 incidents in a neighborhood; distance to access health care; participation in county activities; transportation options and employment?
- Based on statistical results, shape ideas for policy interventions and evaluation.

Additional examples of our work with Arlington that involved the 2016 fellows in the **Virginia Tech Data Science for the Public Good** (DSPG) summer program are:

- **Open Data Portal:** DSPG fellows established criteria for a user-friendly open data portal by reviewing literature and industry standards. They then reviewed Arlington's open data portal and those of San Francisco, Chicago, and Washington, D.C, to find examples that do and do not meet the criteria. The fellows presented their findings to the Arlington chief information officer and deputy. The fellows published an article to describe the criteria and findings for *AMSTAT News* (http://magazine.amstat.org/blog/2016/08/01/portals-aug16/).
- **Operation FireSafe:** The Arlington County Fire Department (ACFD) runs the Operation FireSafe Initiative to ***increase the number of residences in the county with smoke alarms***. The DSPG fellows used data from the Arlington fire department, CoreLogic housing data, the American Community Survey, and information from the National Fire Protection Agency to identify the residences most likely to have fire fatalities and least likely to have smoke alarms by constructing a Bayesian logistic regression model with conditionally autoregressive spatial effects. The ACFD will use this information to target households to check and install smoke alarms.
- **Emergency Alerts:** The Arlington Office of Emergency Management (OEM) wanted a better understanding of the subscriber behavior and ***how to get more individuals, or perhaps the right percentage of the population, to subscribe.*** Using the emergency alert the fellows were able to identify the words that correlate with unsubscribing, identify areas of low enrollment, and identify those individuals most likely to share information. OEM used this information to target outreach efforts to low enrollment areas and those individuals most likely to share and to evaluate alerts types that lead to unsubscribing.

- **Employment:** The federal money Arlington County receives for transportation is allocated through the Metropolitan Washington Council of Governments, a consortium of 22 jurisdictions across Virginia, Maryland, and Washington DC. The funding is a function of employment and the ***Arlington County demographers hypothesize that due to the large number of federal civilian, military, and contract employees in Arlington County the number is biased downward.*** The DSPG fellows researched federal, state, and local data sources and documented their limitations. Novel ways to document building occupancy such as water and cell phone data were evaluated along with estimates of their variability. This is an on on-going project.

Questions are emerging from Cooperative Extension Service (CES) Unit Coordinators across Virginia. In addition to the CES network of agents, there are hundreds of volunteers from the community that work with the agents on a variety of topics. Examples follow.

- Topics related to the **use of farmland over time** and the impact of these changes on various aspects of the community and local government:
  - The **increase in grape crops** across the state and potential changes lead to questions regarding how to manage farm revenue from crops, changes in tax revenue and how this could be used in the community (e.g., for education, parks, etc.).
  - The **transition of farmland** to other uses presents challenges to farmers and ranchers, due to their age, as well as changes in revenue from crops, land prices, changes in the tax base, etc. What should be the blend of local, state, and federal policies to support a healthy transition? How can these policies also encourage beginning ranchers and farmers?
- Topics on children and youth:
  - Understanding college readiness to help students make a successful transition to college. Working with local schools, teachers, and regional leaders on data management and analytics and using demographic, academic, behavioral, and socio-economic data from students in elementary and middle schools, it is possible to predict a student's level of college readiness. This information could be used to provide more customized support and interventions while the students are still in elementary and middle school.
  - Targeting youth obesity in rural areas with local data. Understanding the conditions faced by rural families through data that tells the story of factors influencing health outcomes: socioeconomic data, BMI and cardiovascular fitness scores, school districts, public schools and private schools, neighborhood crime rates, healthy food access, distance to fast food, etc. Data from 4-H Healthy Living programs is one source of data to study this issue (http://newsstand.clemson.edu/mediarelations/clemsons-4-h-program-tackling-childhood-obesity/).
  - Many topics around 4-H have been discussed that contribute to positive youth development. **4-H is the largest children and youth organization in the world.** The topics they address span rural and urban living. These positive youth development programs have been evaluated using statistically designed studies and inferential analyses by Tufts University (http://ase.tufts.edu/iaryd/researchPositive4HPublications.htm) in cooperation with 21 Land Grant Universities.

Appendix 3: Deploying Community Learning Across Iowa and Virginia

The Community Learning Data Drive Discovery (CLD3) deployment across Virginia and Iowa includes two great Land Grant Universities, Virginia Tech and Iowa State University. The CLD3 leadership is **Sallie Keller,** Professor of Statistics and Director, Social and Decision Analytics Laboratory of the Biocomplexity Institute, Virginia Tech; **Sarah Nusser**, Vice President for Research and Professor of Statistics, Iowa State University; **Stephanie Shipp**, **Deputy** Director and Research Economics Professor, Social and Decision Analytics Laboratory of the Biocomplexity Institute, Virginia Tech; **Erin Mullenix**, Director of Data Driven Science, Iowa State University; **Michael Lambur**, Associate Director, Program Development, Virginia Cooperative Extension, Virginia Tech; and **Chadwick Higgins**, Senior Director, Iowa State University Extension and Outreach, Iowa State University.

Transforming Communities across America

Transforming local communities using evidence-based decision-making will be accomplished through the existing Cooperative Extension Service (CES) associated with the Land Grant Universities. With close community relationships, an understanding of local issues, access to government leaders, and a history of addressing local needs, CES agents are uniquely prepared to be the arms and legs of Community Learning Data Driven Discovery (CLD3). With 75 Land Grant Universities across the country, CLD3 will not only leverage an existing infrastructure with a unique plan for long-term sustainability, and will also facilitate the evolution of the 102 year old CES organization to the next generation.

Deployment Strategy

The CLD3 Deployment Strategy possesses three broad goals:

- Expand and enhance the role of Cooperative Extension across the country.
- Collaborate with local government officials in an unprecedented manner bringing new technologies to local decision-making along with analytics to measure progress and effectiveness of programs.
- Transition the CLD3 process from a philanthropic funded initiative to a locally and federally sustainable program.

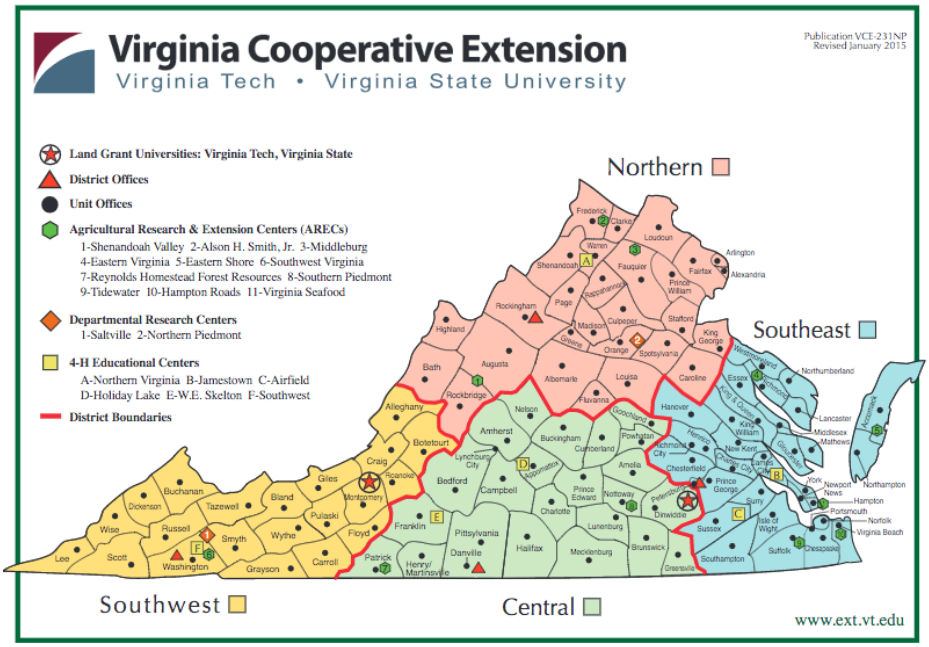
Educating CES agents and local government staff will occur directly by researchers and CLD3 leadership at Virginia Tech and Iowa State University. Training the first wave of CLD3 'transformation agents' will occur in large part during face-to-face meetings and working sessions. There will be close involvement with CLD3 leadership as we go through the CLD3 process, including collaboration with other local leaders, forums across government departments and offices, identifying and retrieving local data, data analytics followed by sharing visualizations with local leaders. Subsequent phases of deployment, both within and across states, will occur in a more scalable manner with education materials developed and refined during the Virginia and Iowa deployments.

Virginia and Iowa Deployment

Virginia

Virginia local governance is based on counties and independent cities. The county governing body in Virginia is known as 'the board of supervisors' where most counties elect their supervisors. In most counties, the board of supervisors appoints a county administrator who is the chief executive of the county government and has the responsibility of overseeing all administrative matters not assigned elsewhere by the state constitution to someone else.

As part of the national Cooperative Extension Service (CES), Virginia Cooperative Extension (VCE) maintains local offices in four districts containing 95 counties and 12 cities across the Commonwealth. These 107 local offices, called units, are divided into four geographic districts (Central, Northern, Southeast, Southwest) depicted in the figure below.



Each unit office is staffed by Extension Agents working in one of three program areas - Youth Development, Family and Consumer Services, and Agriculture and Natural Resources. Agents are supervised by a District Director. While units operate independently to address issues and needs within their communities, they also do some cross-unit programming on broad topics such as identifying food desserts or understanding the impact of the transition of farmers to other occupations

Recruiting VCE Units

Our early planning has included preliminary outreach across Virginia with a webinar to introduce CLD3 to VCE agents across the state and obtain buy-in with an eye to identify candidate communities for early inclusion in our first wave of seven units in year one of the project. We anticipate a degree of self-selection is likely to occur based on the interest level of VCE agents and the appetite of community leaders to engage in the emerging program. We will continue to get buy-in and educate the VCE agents as we identify the additional units and begin staging our second wave of eight units for deployment in year two. These onboarding plans will be followed by educating local government leaders and getting them on board with our mission.

Considerations for our selection of counties or cities will include:

- District coverage – include representation from all four districts aligned by region of the state
- Mix of population size and rural/urban designation
- Evaluate strength of most recent data-based situation analysis
- Interest in participating from both Virginia Cooperative Extension and local government
- Strength of local government support for Cooperative Extension
- Relationship of agents with local government
- Repeatability and replicability to other units and communities
- Potential issues of high public interest and concerns

During the first wave, processes associated with obtaining, accessing, and fusing data to begin the analysis process and add to our catalogue of recipes will be refined by our researchers and the CLD3 leadership.

These onboarding plans will be followed by educating community leaders and getting them on board with our mission. With these partners we will define 'what keeps them up at night' topics to study based on their

local issues and needs. Once the communities are up and running, we will expand to a second wave where subsequent communities are brought into CLD3.

Timeline for a Three-Year Rollout

Early in 2017, CLD3 researchers and leadership will continue to work shoulder to shoulder with interested units in informing and gauging interest from local government. Key local government leaders to engage will be county administrators and city managers. VCE program management, (i.e., VCE directors), will work with VCE Unit Coordinators to provide sufficient information to local government leaders to allow them to take the appropriate steps in securing local government approval should they be interested in participating. The final selection of units will be made by a committee of the four District Directors, Senior District Director, appropriate Associate Directors, and CLD3 leadership using the consideration identified previously.

After selections are made, a face-to-face kick-off meeting of all partners will be held. The purpose of the meeting will be to provide an overview of the project over the three-years and to establish initial expectations for VCE and local government staff. Initial expectations might include:

- Determine who will directly support the local deployment from VCE, resource needs, and how these resources will be allocated, (it is anticipated the equivalent of 1 FTE equivalent of support will be needed in the first year).
- Appoint appropriate local government staff to directly work with their implementation of CLD3 and identify any resource needs, (it is anticipated resources may be needed to support data extractions and the development of data sharing agreements).
- Identify university researchers to work with Extension and local government on the early issues requiring data science skills. Researchers will maintain involvement in the project to understand process and outcomes.
- In partnership with local government, establish and facilitate a local CLD3 steering committee to manage and sustain the implementation of CLD3
- Work with local government to identify needed data collection needs associated with issues being studied, e.g., community development and the provision of services
- Conduct statistical analyses, generate initial hypotheses, and work collaboratively between VCE and local government to create a response plan, i.e. how the work of Extension and the county/city will change as a result of lessons learned from the data.
- Develop and deploy of intervention strategies and education programs to address issues. Evaluate hypotheses using rigorous experimental design principles and Adapt strategies based on continuous evaluation.
- Look at existing Extension programs in the county/city and identify data needs that could allow for the improvement and expansion of that programming.
- Work with local government to secure funding to continue data analysis and community learning research.

The length of the project in each unit will be three years. Funding will be provided to local VCE offices and government for three years on a descending scale (100% first year, 50% second year, 25% third year). Correspondingly, experiential, hands-on training and materials development will escalate over the three years. Ultimately, at the end of the third year, local VCE and government staff will have acquired the skills and built infrastructure to continue CLD3 as the new standard for operations with ongoing support through community-based research by the academic researchers, communities of practice, annual meetings, and web resources. Anticipated activities and resources for each unit over the three years are presented below.

- **Each year,** CES agents will work with local government civil servants to conduct data discovery and analysis, coordinate data discovery forums and issue identification, identify hypothesis and questions

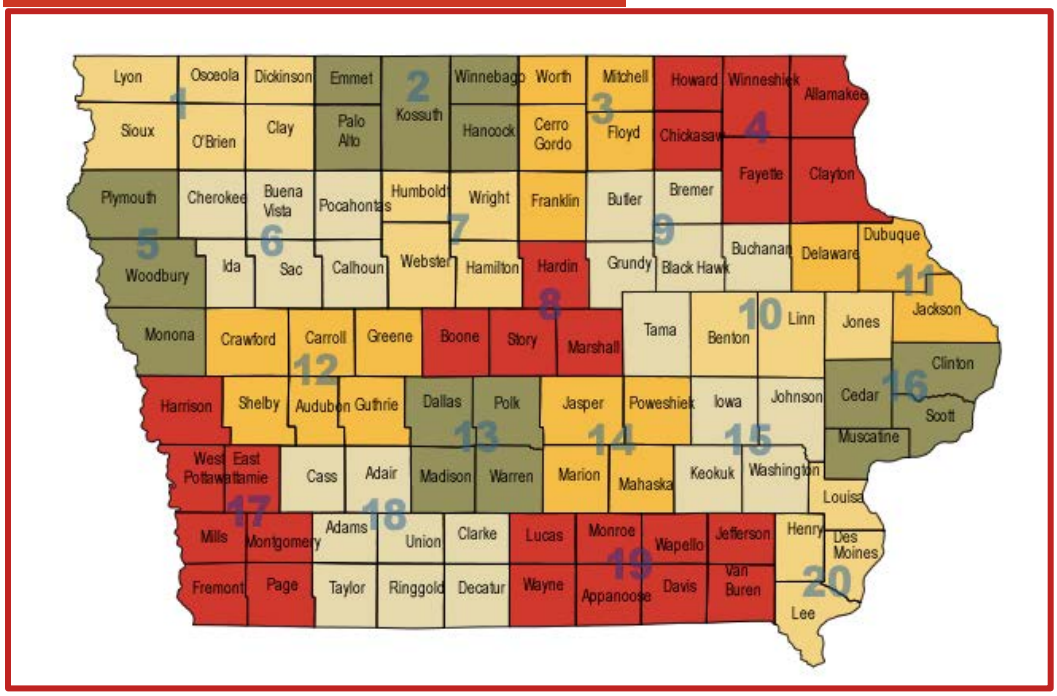
using statistical and geospatial learning and insights from local government sponsor, and complete situational analysis for each issue identified. Document successes, challenges and how they were addressed, outputs, and process outcomes. In addition, in each year, the following additional activities will be undertaken:

- **Year 1** – Work with local government to create a local steering committee and creation of data sharing agreements will be initiated.
- **Year 2** – Continue data discovery and analysis. Develop evaluation plans using experimental design principles, as appropriate. Begin executing evaluation plans, as appropriate. Make modifications of interventions, policies, and evaluation strategies, as appropriate. Identify key people in VCE, Virginia Tech researchers, and local government to participate in data discovery, cleaning and transforming data, and analysis. Begin process of training these individuals to take on these roles as the project support ends.
- **Year 3** – Continue data discovery and analysis, strategy development and implementation, and evaluation planning and execution as appropriate from years one and two with local steering committee taking the lead. Develop an exit plan so that the process will continue to be implemented in the following years through the Cooperative Extension service and partnerships with researchers.

<u>Iowa Deployment</u>

There are 944 incorporated cities in Iowa, and 99 counties. Community populations in Iowa range widely from 15 to over 200,000. Over half of city communities in Iowa have less than 500 in population. Iowa State University Extension and Outreach has 100 local offices spanning all 99 counties in the state. The local offices act as launching pads for CLD3 engagement. The CLD3 Team will guide all project functions and will utilize the 2014 County Needs Assessment data to help inform community interests and concerns.

IOWA STATE UNIVERSITY
Extension and Outreach



Iowa State University has developed a set of criteria to select an initial set of representative cities to participate in the first phases of the CLD3 project. These criteria include:

- Urban versus rural character of communities
- Population size
- Population diversity
- Base of economic activity: services, manufacturing, agricultural
- Community leadership familiar with data driven approaches
- Community leadership committed to joint project approach
- Engaged Extension office in surrounding county
- Availability of local community college and/or liberal arts college(s) for project collaboration

Iowa State University has begun the site selection process, and has initiated conversations with two local communities that appear to be an excellent fit. Meetings and site visits are being scheduled to begin the process. During these meetings, the Iowa State University CLD3 Team will introduce local officials and interested local stakeholders to the project.

On campus, Iowa State University has convened a group of interdisciplinary faculty as an advisory committee who are highly interested in this proposal. With great enthusiasm and motivation toward this project, this group of faculty looks forward to engaging in CLD3. We have also engaged information technology/data platform experts, in conversation with counterparts at Virginia Tech to begin considerations for the technology and data framework.

Iowa State University has had meetings with and plans to schedule additional meetings with local organizations with a public focus that they may interact with throughout the CLD3 implementation. Discussions have already taken place with the Iowa League of Cities. Stakeholder engagement will be sought from the Iowa State Association of Counties, the National League of Cities, the National Association of Counties, the Greater Des Moines Partnership (who maintains a Des Moines Regional Research, Stats and Data Hub), and the Des Moines Metropolitan Planning Organization (along with its affiliated Capitol Crossroads program), and the Midwest Big Data Hub.

Meetings with state agencies will be scheduled at the beginning of the local engagement process, and will include: Iowa Departments of Management, Public Health, Workforce Development, and Education; the Iowa Economic Development Authority, Iowa Homeland Security and Emergency Management, local and county emergency management offices; in addition to local city officials and department leadership.

In the short and long-term, Iowa State University is organizing a group of local government officials from multiple communities around the state that will serve as a “sounding board” to local government issues. This group of roughly 25 officials from a diverse cross-section of local communities in Iowa will participate in regular calls, and where appropriate, can comment on and engage in issues raised by the CLD3 project.

The Iowa State University CLD3 team also plans to travel to Virginia to meet with the Virginia Tech SDAL and VCE CLD3 participants to explore best practices and successful projects to guide the Iowa implementation and long-term deployment planning process.

Demonstrating the Art of the Possible

The efforts carried out to date in support of bringing CLD3 to communities in Virginia are described next (also see Arlington pilots studies described in Appendix 1). These include examples of questions that might be asked by local government civil servants and possible ways to address their questions. These efforts serve as exemplars for new efforts, may help communicate key ideas and concepts of CLD3 to CSE agents and local government officials, and will serve as prototype tools to build into the CLD3 Data Science Processes and Platforms described in Appendix 4.

Prototype platform for data access and analysis

The CLD3 team developed an initial prototype of a web-based platform to access, link, analyze, and view relevant data (Figure A3.1). This prototype currently includes data from the American Community Survey, Healthy Communities data from the Robert Wood Johnson Foundation, data from the U. S. Department of Agriculture, and local geospatial locations scraped from the Google Maps API.

As a demonstration, the relationship between the proportion of adults with diabetes and median income by county is shown in the lower scatterplot, showing the general trend that counties with lower median income tend to have higher adult diabetes rates. There are a number of counties (circled in the lower left of the plot) that defy this general trend. Upon using the linkage capabilities in this prototype, it was discovered that these counties are all homes to large universities.

The prototype will continue to expand its capability; it can already handle time series, and carry out rudimentary analyses. As efforts begin with local communities, additional local data sources will be added to this prototype. This centralized capability for managing data and analyses will allow recipes, analysis approaches, and data sources used in one investigation to be adapted to aid other investigations. Maintaining this capability for all local governments in the state will facilitate the use of data and analytical approaches for community learning.

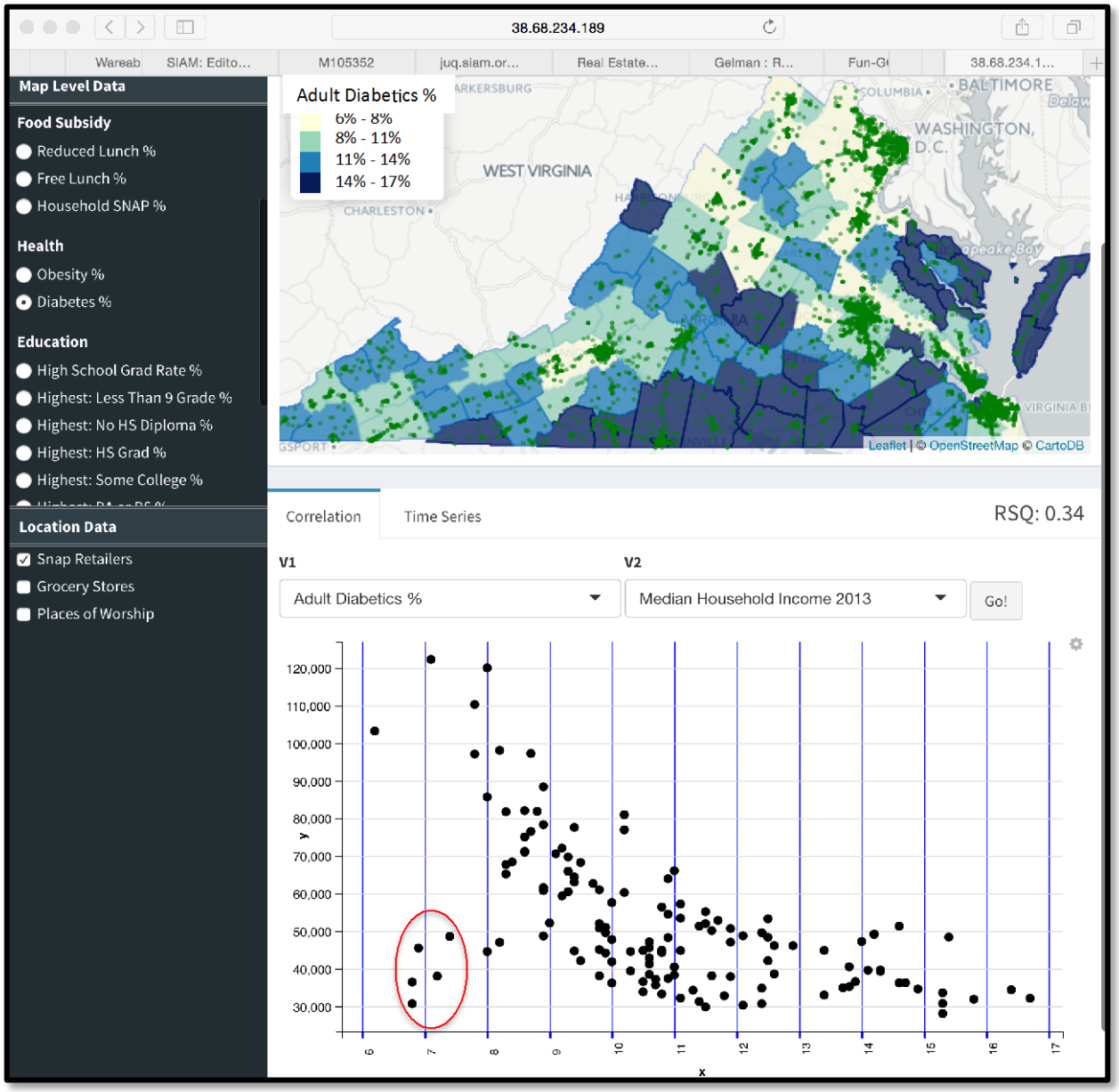


Figure A3.1. Screenshot from prototype for accessing, combining, viewing, and analyzing different data sources. Here the prototype is linking data from Google Maps, the American Community Survey, RWJF County Health Rankings, and USDA. Top: Diabetes rates by county, overlaid with the locations of SNAP retailers. Bottom: Scatterplot of % of Adult Diabetics and Median Household income by county. The pattern of the scatter plot illustrates a higher incidence of diabetes in counties with lower incomes. Interestingly, the 5-circled counties do not follow this general pattern, but rather show low income as well as lower incidence of diabetes. Further exploration is called for to understand the characteristics of these communities and whether learnings could be shared with other low income, high incidence of diabetes counties.

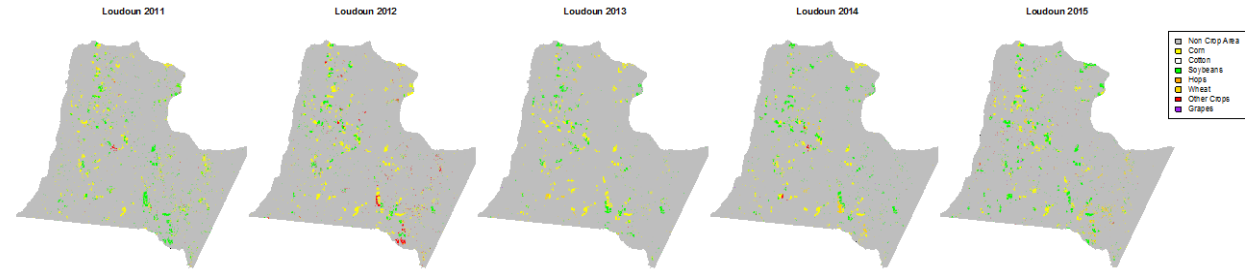
Webinar with VCE Unit Coordinators to Create Interest in CLD3

As part of the outreach and education process, communications materials including a webinar and web-based CLD3 prototype were shared with a preliminary set of unit coordinators and city officials.

In late September, the team shared a one-page information description of CLD3 with VCE Unit Coordinators (UCs) and local government officials. A CLD3 webinar designed to create interest in, and provide information about, becoming involved in the project was conducted on October 27. Prior to the webinar, nine units indicated interest and seven UCs participated in the webinar (Greensville/Emporia, Lynchburg, Rockbridge, Frederick, New Kent, Arlington, and Alexandria), as well as two county administrators from New Kent and Rockbridge. During the webinar, we provided three data examples to demonstrate examples of repurposing local administrative data flows to answer questions asked by local government civil servants (illustrated in this appendix.) The webinar was well received. It lasted one-hour and can be viewed at: https://virginiatech.webex.com/virginiatech/ldr.php?RCID=703ac072a04832cd1cbd1e34203f0034

Crop Coverage over time for Loudoun County

To demonstrate how satellite data provided by USDA could be used to give crop coverage at high spatial resolution, we extracted data from the USDA's NASS CropScape website. Data from 2011 to 2015 are shown in Figure A3.2 below. The data shows how farmland has changed over time for Loudoun County, which can be aggregated to show a time series as shown below. The point of this demonstration is to convey how data compiled for one purpose (e.g. measuring agricultural production) might be repurposed to serve another (e.g. managing farm transition). This data has been ingested into the prototype tool, for all counties, so that it can be compared other variables that have been recorded over the same time periods.



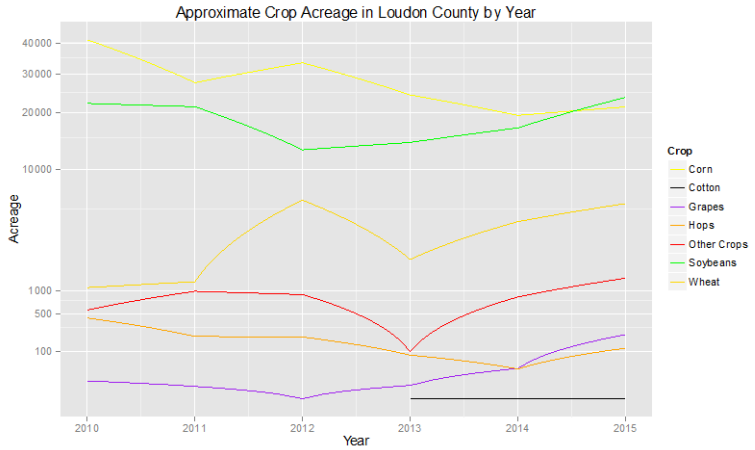


Figure A3.2. Crop usage in Loudoun County 2010 – 2015. The maps show the evolution of cropland at a high spatial resolution.

Estimates of crop acreage aggregated for each crop is shown in the left hand figure. This can support planning and farm transition. For example, the green and yellow trend lines show that most acreage is devoted to corn and soybeans with some fluctuation between the two. The purple trend line shows the increase in grape crops especially since 2012.

Initial looks at teen birth rate trends for Virginia counties

Finally, Figure A3.3 shows teen birth rates over the past 6 years for counties in Virginia. These data and plots were produced using the web-based prototype. Note that the median income for each county is conveyed using color – blue for counties with high median income, red for counties with low median income. The actual income scale is given in the lower right portion of the plot. Some trends can be observed from this plot:

- Counties with higher median income tend to have lower teen birth rates.
- Trends can be observed over time for each county including while most teen birth rates are holding constant, there are some counties with decreasing rates (e.g. Charles City, Roanoke), and there are some whose rates seem to be increasing (e.g. Bland, Craig, Dickenson).

It is important to determine if these perceived changes could be explained by statistical variation. Seeing many counties at once makes it easier to identify counties with similar challenges, so that strategies for improving health outcomes could be devised collaboratively. Such comparisons will also be needed to assess and compare the efficacy of different strategies to improve outcomes. Note that with this data alone it impossible to determine the causes of any observed trends. However this data visualization highlights issues that would be difficult to identify in a standard report and points out areas requiring further exploration using in-depth local data and information sources.

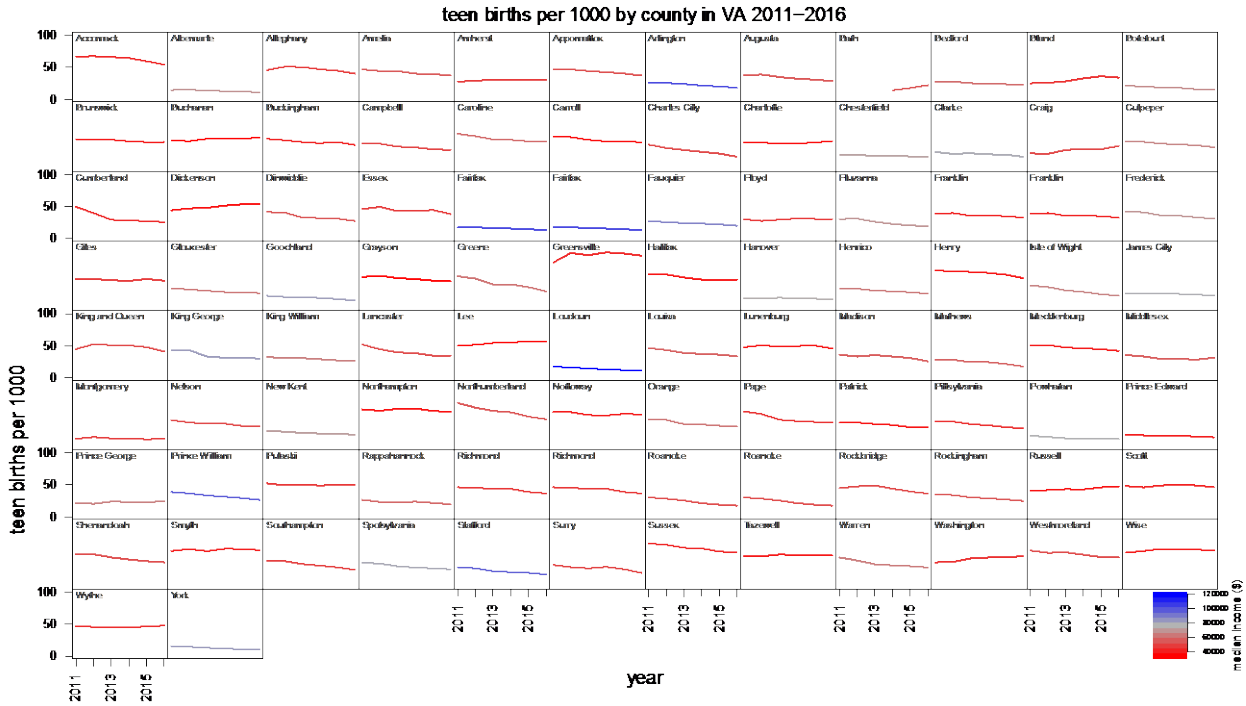


Figure 3A.3. Teen births per 1000 female teenagers are plotted for each county in Virginia. Each trend line shows the teen birth rate by county. The color of the line corresponds to the median income – blue is high, gray is medium, red is low (see scale in lower right). Teen births tend to be higher in lower income counties. Most counties show a decreasing trend of teen births, however some show an increase, suggesting a topic for further study.

Appendix 4: Data Science Processes and Platforms

Summary

We are pursuing a vision of governmental agencies and programs having the capacity for data-driven, inter-jurisdictional, evidence-based decision-making in their governance processes. Data flows relevant to government decision-making, such as local administrative records, surveys, geospatial data, and sensor data, as well as federal, state, and social media data sources, are ubiquitous. Government decision-making based on these data is instrumental to increasing efficiency and improving outcomes. Despite the enormous potential of this, there remains a lack of capacity, be it in technical, analytical, or just general knowledge, to utilize these data sources to inform their decision-making processes. To enhance their capacity, we are working directly with local governments to develop experience in using local, state, federal, and other data sources, and in working collaboratively with issue-area researchers, to support a data-driven model of governance.

This capacity building effort is being achieved through the development, deployment, and collaborative use of a set of **data science processes** and **supporting technological platforms**. Building on continuing work with local government partners in Virginia, these processes and platforms are being tied together in a more comprehensive manner that is scalable, portable, and inter-jurisdictionally secure. This new **Virtual Analytic Environment** is being piloted with additional local governments in Virginia and Iowa and will mediate access to data and data tools, adhere to necessary data privacy and security constraints, and, allow for the creation of a **loosely federated data infrastructure** across an entire state.

Our Land Grant Universities will serve as the stewards of this architecture, ensuring continuous accessibility by local communities, researchers, and state agencies.

An Opportunity for Data-Driven Governance

There is tremendous potential for enhancing the effectiveness of local government decision-making by liberating and repurposing community-relevant data sources for use in analyses, regardless of their origination. Local government decision-making based on these data and the associated research are instrumental to improving performance, cost, and function. Local governments themselves often have significant stores of historic and current electronic data at their disposal. However, the insights and patterns that analysis of that data might reveal are often obscured from use by siloed information technology systems that have been developed solely for the purpose of processing isolated transactions for the administration of specific programs (Goldsmith, 2016). This results in local government missing the opportunity to re-purpose their administrative data sources and combine them across departments and programs to help gain valuable insights about their community.

Unfortunately, many local governments and programs lack both the analytic and technological capacity to carry out just the first steps of re-purposing their administrative datasets (i.e. profiling, cleaning, transforming) to make them ready for analysis in combination with other datasets (e.g. state, federal, other). In turn, this inability to provide datasets suitable for analysis can make it exceedingly difficult to work cooperatively with policy-area researchers who can bring invaluable insights to the decision-making process.

Vision for a Virtual Analytic Environment to Support Data-Driven Governance at Scale

Working with local governments and programs we plan to enhance their capability for data-driven decision making by developing, implementing, and providing a Virtual Analytic Environment where government analysts and academic researchers can work cooperatively on community-relevant issues using all available community-relevant datasets, including locally-derived data sources (e.g. administrative data, sensor data), sources derived by neighboring communities, state and federal data sources, and data provided by non-governmental entities (e.g. community-oriented non-profit organizations). The initial design of the Virtual

Analytic Environment is based on Virginia Tech's current work with Virginia counties and cities using a set of data science processes and supporting technological platforms designed to support cooperative community research. The following goals will be achieved:

- Develop, provide, and implement a statewide **sustainable comprehensive framework of secure data science processes and platforms**, including:
    - processes of data ingestion and management, data analytics, and analysis presentation that will support local government evidence-based decision making and researchers engaged in community-based research,
    - a set of actively managed technology platforms providing the latest in open-source database, GIS, data analytic, and data presentation technologies, as well as enabling secure and policy-based sharing of data across jurisdictions.
- These data science processes and platforms will be developed in a manner to make them easily replicated and curated beyond their development stage to create a statewide and ultimately national ecosystem.
- Establish a data analytic-centric community-engagement model that keeps barriers to participation as low as possible. For example, there will be no expectation of any significant modification to existing local government data systems, such as data standards, as a prerequisite of participation. Instead, the system host (Virginia Tech/Iowa State) takes on the responsibility for:
    - maintaining a comprehensive database of metadata of all data sources being provided by participating localities, including mappings between data sets using different data standards,
    - providing to the locality, with support, the requisite technologies needed to securely connect their existing data resources to the venue.
- Create and implement a sustainability plan for maintaining the Virtual Analytic Environment.

Overview of the Secure Virtual Analytic Environment

Figure 1 provides an overview of the Data Science Processes and Platforms Framework that includes a **Virtual Analytic Environment**. This is designed to enable data-driven governance by deploying a set of scalable, secure, transferable, and cost-effective data management, data analytics, and data presentation technology platforms.

- Built using the latest **open-source data technologies**, individual technology platforms can be hosted locally or in the cloud. A set of Virtual Private Servers (VPSs) is dedicated to each participating community.
- **Virtual Analytic Environment Transferability:** As the data management and analysis expertise of a community grows, it may become desirable for that community to have a more direct controlling role over the technology platforms. In fact, such developments are a sure sign of success in capacity-building efforts. The technology platforms are purposely designed and constructed for ease of transferability for just such a circumstance.
- **Policy-Based Access:** Secure access of all community data will be mediated by a **rule-based access system** to enforce the data access restrictions and requirements dictated by applicable federal, state and local rules and regulations. Within the Virtual Analytic Environment, this "policy-based access engine" functions as what may be described as a "data resource broker". A data resource broker provides a uniform interface to heterogeneous computer data storage resources over a network or networks. As part of this, the data resource broker will implement a logical namespace (distinct from physical file names) and maintain metadata on data-objects (files), users, groups, resources, collections, and other items in a metadata catalog. This metadata can be queried to locate files based on attributes as well as by name across heterogeneous data resources.

- The computer code that underlies the technology platforms is open-source and will be made available to any community under GNU Public License along with documentation.

Details of the Data Science Processes

We are building community capacity for managing the data science process by working collaboratively through the processes of community stakeholder engagement & discovery, data ingestion and management, data fitness analysis & hypothesis testing, metadata management, and data & analysis results presentation.

- **Community Discovery:** During **community engagement & discovery processes**, work with the community to:
    - (with new partner communities) conduct preliminary hypothesis generation where we start with critical community-defined issues and work with community leaders to elicit an initial set of variables suspected to be causatively related to each issue; (with existing partner communities) confirm initial **hypotheses** regarding specific local issues that were already derived from previous work;
    - conduct a **data management system status discovery process** to ascertain the methods and technologies currently employed for data management, as well as their capacity to handle the data storage and management requirements of the entire process;
    - conduct a **data analytics capabilities assessment** to ascertain the community's current level of data analysis expertise; and,
    - conduct a **data discovery and inventory process** where first, potential data sources that could be related to the specific issue areas are identified and screened to determine their potential usefulness in supporting the research questions, and second, additional details are inventoried for those that are deemed worthwhile and in need of additional data profiling.
    - deploy necessary data connection technologies as required by an already established **data access plan** to enable the data transfer and management

- **Data Management**: During **management processes** the community will gain the capacity to execute the processes necessary to configure their datasets for secure transfer and/or secure remote dynamic-access. This process entails:
    - establishing the **type and method of data transfer**: is the data being pushed to or pulled into the cooperative platform, or is the data staying where it is and being dynamically queried in a federated manner as needed?
    - establishing the **best protocol(s)** to use given the types and method of transfer (e.g. SFTP, secure dropbox, secured REST API, VT SAFR-Data Adapter for secure federated queries)
    - establishing **data marshaling processes** for: system mediation logic, data pipeline and data transformation, transfer schedule, and data provenance maintenance
    - establishing **secure data storage procedures** (e.g. each project being stored on a new project-dedicated encrypted partition, original data being stored as non-removable and non-editable)

- **Data Analytics**: During **data analytics processes** the community will gain capacity in understanding how to (Keller et al. 2016):
    - apply a disciplined process for assessing data fitness for use through profiling data sources in terms of data quality structure and metadata;
    - prepare selected data for analysis via cleaning, transformation and restructuring (Schroeder, others);
    - correct and continually update metadata (see discussion of the Lexicon below)
    - creating linked datasets securely and at the highest possible level of accuracy (Schroeder, others);

    - testing the prepared dataset(s) as adequate for use in proposed analyses/modeling approaches;
    - create analytic experiments to test their hypotheses
    - understand or even begin to create statistical models
    - understand or even begin to perform statistical analyses to test their hypotheses

- **Data Lexicon**: During the **data information processes**, which runs in parallel with the data fitness process, the community will gain capacity in understanding how to maintain the information necessary to enable high-quality linkage of datasets across jurisdictional boundaries and levels of government. Mediating this ability to support the linkage and use of datasets from different sources is an enhanced metadata repository referred to as the Lexicon. At its base, the **Lexicon** serves the function of a metadata repository - a database created to store metadata from various systems. Metadata is information about the structures that contain the actual data. Metadata is often said to be "data about data", however, the Lexicon goes far beyond this definition, proving a centralized node of data source information that can be used for provenance tracking and data linkage within a heterogeneous network of data sources (A. D. Schroeder 2013b). Specifically, the Lexicon is an inventory of and history of changes to:
    - every available data field in every available data source
    - the structure of their storage
    - possible values and meanings of the information stored
    - possible transformations of each set of field values from one data source to another set of field values from another data source
    - methods of data source access
    - matching algorithms and how they are to be used in conjunction with possible field value transformations

  The Lexicon provides fundamental functions for the operation of the framework and, therefore, it is a requirement that the data information necessary for its operation be collected from all partner communities. With this information, the Lexicon enables the removal of much of the complexity required for high quality data linkage from all data partners (i.e. no enforcing data standardization schemes on data partners). The Lexicon is housed and maintained in an RDBMS by staff of the participating land grant university (e.g. Virginia Tech, Iowa State).

- **Data Presentation**: During the **data presentation process** the community will gain capacity in using the latest open-source technologies to enable multiple methods of communication (e.g. within research teams, with policy decision makers, with the public), including:
    - Statistical reports and visualizations
    - Interactive data dashboards
    - Interactive project wikis
    - Public-facing interactive wikis

<u>Enabling a Data Driven Future for our Localities</u>

In pursuit of our vision of governmental agencies and programs having the capacity for data-driven, inter-jurisdictional, evidence-based decision-making, we are developing, deploying, and analyzing the utility of a Virtual Analytic Environment. Building on continuing work with local government partners, this Virtual Analytic Environment will provide to localities and researchers a collaborative analytic environment designed purposefully to support a set of data science processes and will be built upon a set of open-source technological platforms that are scalable, portable, and inter-jurisdictionally secure. This set of processes and supporting platforms will not only enable an enhanced analytic capability of a locality's own data, but will also provide a platform for the secure sharing and analysis of data across jurisdictional boundaries. Our Land Grant Universities, as a recognized and legitimate responsibility of their outreach missions, will serve as the stewards of the Virtual Analytic Environments in their respective states.

References

1. Goldsmith, S., Crawford, S., & Weinryb Grohsgal, B. 2016. Innovations in Public Service Delivery: Issue No. 4: Predictive Analytics: Driving Improvements Using Data. Inter-American Development Bank.
2. Keller, Sallie Ann, Gizem Korkmaz, Mark Orr, Aaron Schroeder, and Stephanie Shipp. 2016. "The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches." *Annual Review of Statistics and Its Application* 4 (1). Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, California 94303-0139, USA.
3. Schroeder, Aaron D. 2012a. "Pad and Chaff: Secure Approximate String Matching in Private Record Linkage." In *Proceedings of the 14th International Conference on Information Integration and Web-Based Applications & Services*, 121–25. ACM.
4. ———. 2012b. "VT-VLDS Data Adapter User Guide." Virginia Information Technologies Agency.
5. ———. 2013a. "VT-VLDS Identity Resolution and Query Execution Process Overview." Virginia Information Technologies Agency.
6. ———. 2013b. "VT-VLDS Shaker Specifications." Virginia Information Technologies Agency.

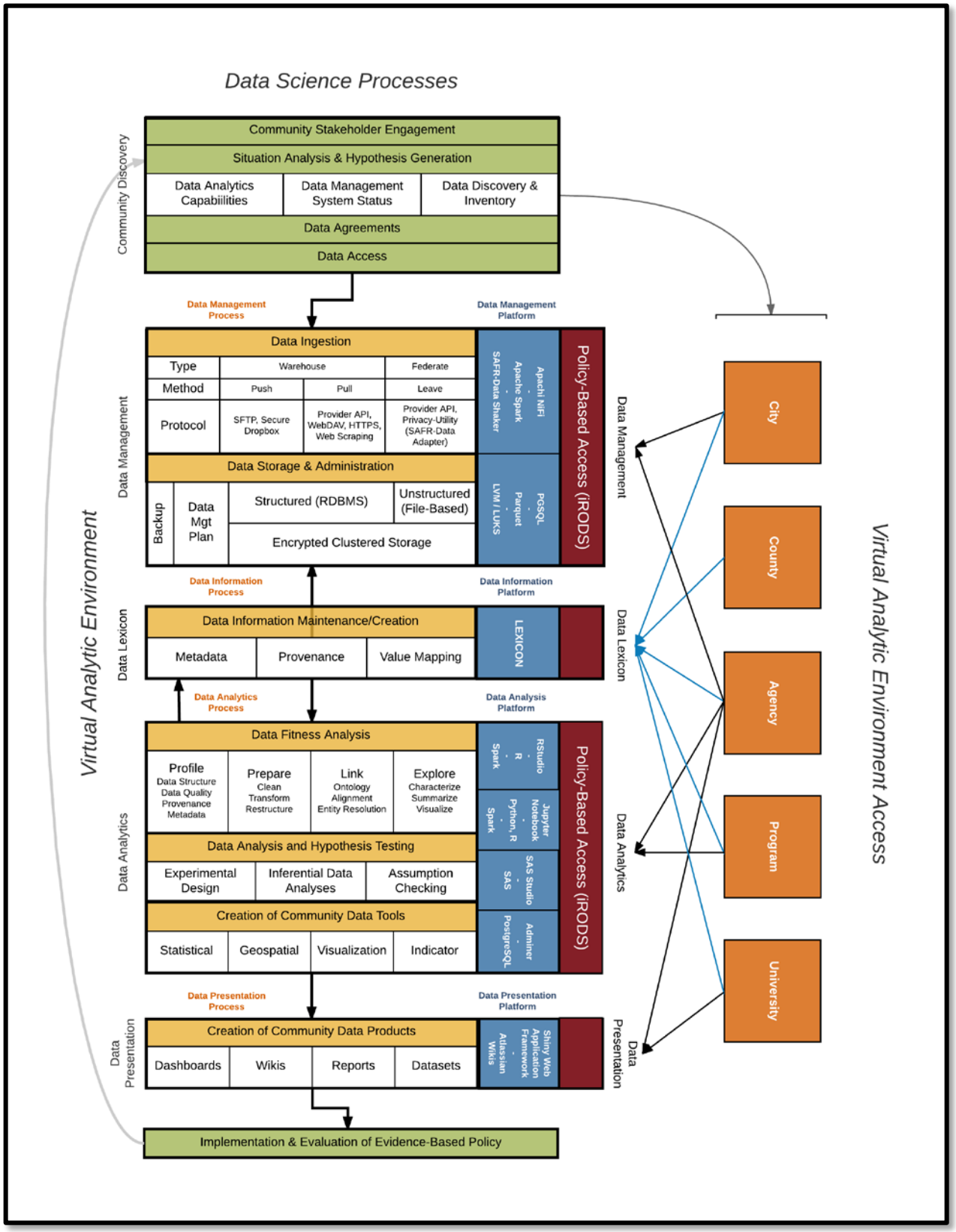


Figure 4-1: Data Science Processes and Platforms Framework

Appendix 5: Creating the National Conversation through Partnerships

CLD3 will fundamentally change how cities and counties shape policies. Data analytics will inform decision-making and become the new normal. The primary purpose of the discussions and presentations that the Virginia Tech and Iowa State teams have undertaken over the past 3 months was to socialize our CLD3 vision, seek feedback on the CLD3 concept, gain buy-in, and identify natural ways to collaborate and partner. Our discussions and interactions are described below:

University Interactions

Virginia Tech

The **Board of Visitors (BOV) and Virginia Tech leadership** enthusiastically support the creation of the CLD3 process throughout the Land Grant University network. Dr. Sallie Keller gave a presentation to the Virginia Tech BOV on August 28, 2016.

- Virginia Tech President Sands said we should embrace CLD3 at part of the future of what land grants universities do and that this is the future of Land Grand Universities.
- BOV Research Chair, Mahmoud Kami, said that CLD3 must be part of rebuilding the U.S. infrastructure.
- Virginia Tech Provost Rikakis, as well as Deans and Institute Directors have all acknowledged that the CLD3 initiative is central to the Virginia Tech Beyond Boundaries strategic planning and it has been incorporated into the Data and Decision Sciences Destination area.
- The Vice President for the Northern Capital Region, Steve McKnight, the Vice President for Research and Innovation, Theresa Mayer, the Vice President for Advancement, Charles Phlegar, and Associate Vice President for Principal Gifts, Monecia Taylor are delivering the message across the state and to potential supporters and funders.
- Dean Paul Knox, founding Dean of the Virginia Tech Honors College is creating a signature experiential learning program with SDAL based on our Data Science for the Public Good program.

Seeking partnerships with other public universities, we have also engaged with other universities:

- *Virginia Commonwealth University* - Steve Woolf, Director, Center on Society and Health, for his public health expertise.
- *University of Virginia* - Robert Pianta, Dean of the Curry School of Education, Novartis Professor of Education, and Founding Director of the Center for Advanced Study of Teaching and Learning, for his experience in studying education issues.
- *Johns Hopkins University* - Beth Blauer, Director, Center for Government Excellence, to share best practices and explore developing the data processes and platform infrastructure. She is leading the deployment of the What Works Cities, a national initiative to build local government capacity to use data for decision-making and policy development.

Iowa State University

The introduction of CLD3 was quickly embraced by Iowa State University President Leath. The Vice President for Research, Sarah Nusser, is leading this initiative. She immediately put together an outstanding team and hired the former research director of the Iowa League of Cities to develop and deploy CLD3 across Iowa.

- President Leath
- Senior Vice President and Provost, Jonathan Wickert
- Vice President Research Nusser involved as site PI, will speak on 10/24 at Extension and Outreach convening event for Iowa State faculty across the university
- Vice President Extension and Outreach Cathann Kress
- Senior Director for Extension and Outreach Chad Higgins
- Iowa State Foundation President Larissa Holtmyer Jones

- Iowa State Foundation Vice President for Development Jeremy Galvin (ISUF research liaison)
- Vice President for Research's Research Leadership Council, which includes College Associate Deans for Research, Extension Research director (Higgins), Vice President Economic Development and Business Engagement

State Interactions

Commonwealth of Virginia

- The team has had several productive discussions with Executive-level leadership across many state agencies in the Commonwealth of Virginia. They are very interested in and actively support CLD3.
- *Virginia Department of Health and Human Services* - Bill Hazel, Secretary
  - Discussions have focused on how to apply the CLD3 process across the state starting with using data to understand and shape policy about the prescription drug addiction crisis, developing standard data sharing agreements, and convening Virginia's public universities to work on these issues.
  - We have been asked to develop a business plan for a Public Private Partnership around the development of a federated statewide Data Access and Process Platform with Virginia Tech as the stewards and curators.
- *Virginia Center for Innovation Technology* - Anthony Fung, Deputy Secretary of Technology. We were part of his advisors as he prepared the response to the Governor's Directive 7 on Leveraging the Use of Shared Data and Analytics (https://governor.virginia.gov/media/6007/ed-07-leveraging-the-use-of-shared-data-and-analytics.pdf)
- *Virginia Health Workforce Development Authority* - Marissa Levine, Health Commissioner, expressed early support for our work with Arlington County and interest in similar activities across the state.
- *Virginia Department of Emergency Management* (VDEM) - Jeffery Stern, State Coordinator
  - We did an evaluation of VDEM data access and data flows and provided guidance on how to enhance data informed communication.
  - Stern has interest in identifying vulnerable neighborhoods to ensure they receive assistance and support during emergency events such as snowstorms, flooding, and other events.
- *Virginia Department of Education* - Jennifer Piver-Renna, Senior Executive Director for Research

Iowa State Agencies

- Iowa Economic Development Authority: Gail Kotval, Iowa Innovation Council liaison (10/3/2016)
- Iowa Department of Management: Ted Nellesen, City Budget Director, David Roederer, Chief of Staff, being scheduled
- Iowa Workforce Development: Marketa Oliver, Workforce Services Administration, former city administrator, being scheduled
- Iowa Emergency Management Association: AJ Mumm, Polk County Emergency Management Coordinator, being scheduled
- Iowa Homeland Security and Emergency Management Division: Angela Chen, Bureau Chief, being scheduled

Federal Agencies and Offices Interactions

Leaders of federal programs understand the power of collaborative and flexible relationships that unite all levels of government. The benefits flow both ways, federal programming funds are important to the administration of programs at the state and local levels and administrative data sources at the state and local levels are needed at the federal level. Federal leaders that we have met have stated that CLD3 is visionary and is the way forward to bring evidence-based data-driven decision-making to all levels of government.

- *Department of Health and Human Services* - Karen DeSalvo, Acting Assistant Secretary for Health
    - "CLD3 is creating a national resource, not a report."
    - CLD3 could be a significant complementary capability and capacity to support Public Health 3.0.
- *Department of Agriculture* - Cathi Woteki, Under Secretary for Research, Education & Economics
    - Echoed Virginia Tech President Sands regarding the powerful impact CLD3 could have on the future vision and mission of Land Grant Universities. Noted the Association of Public and Land Grant Universities is grappling with this now and she will help connect us to that activity (http://www.aplu.org/).
    - Introduced us to the director of the National Institute of Food and Agriculture, the primary federal funders of the Cooperative Extension Service.
    - Suggested that we provide language for the 2018 Farm Bill that could accelerate the CLD3 adoption.
- *Census Bureau* - John Thompson, Director
    - Said that the CLD3 will bring the Land Grant Universities into the 21st century and the Census Bureau will partner to provide federal statistics to CLD3.
- *Department of Homeland Security* - Reginald Brothers, Under Secretary for Science and Technology is interested in CLD3 to advance data analytics to create more resilient communities.
- *The White House Office of Science and Technology* Policy - Tom Kalil, Deputy Director for Technology and Innovation
    - Stated that we have created a process that overcomes the barriers to achieving data-driven communities; CLD3 highlights what is possible and provides a path forward for overcoming the barriers commonly cited by others such as data sharing agreements, lack of workforce capacity, interoperability, privacy, meta data, and stovepipes of data.
- *The White House Office of Social Innovation and Civic Participation* - Dave Wilkinson, Director. He's been following the Administration's "place-based" initiatives and has first-hand knowledge of the challenges communities face in harnessing data for better decision-making. David quickly embraced the idea for CLD3 and invited Sallie Keller to speak at the White House workshop on December 6, 2016 on "Using Innovation, Evidence, and Behavioral Science – Common Ground Solutions to Build Economic Opportunity and Create Stronger Communities."
- *National Science Foundation* - Fen Zhao, Program Officer for the Big Data Hubs & Spokes program
    - Iowa State is a co-PI in the Midwest Big Data Hub. Iowa VPR Sarah Nusser has discussed CLD3 with Zhao.
    - Virginia Tech is part of the Southern Big Data Hub and we have discussed CLD3 with their leadership.

We have met with other federal leaders as well. Each of the following agencies has agreed to support CLD3 in some way and are interested in the ability to access local data flows for federal statistics.

- *Department of Justice* - Jeri Mulrow, Bureau of Justice Statistics, Acting Director;
- *Department of Transportation* - Carlos Monje, Acting Under Secretary of Transportation for Policy and Assistant Secretary for Transportation Policy, and Jason Broehm, Transportation Analyst;
- *Federal Reserve Board* - Barbara Robes, Principal Economist, Consumer and Community Development Research Section, Consumer and Community Affairs;
- *Department of Housing and Urban Development* - Shawn Bucholtz, Director, Housing and Demographic Analysis Division, Office of Policy Development and Research;
- *Department of Agriculture* - Sonny Ramaswamy, Director, and Meryl Broussard, Associate Director, USDA National Institute of Food and Agriculture (NIFA).

Interactions with Other Organizations

Results for America

Results for America is working with local elected officials and policymakers to improve outcomes for young people, their families, and communities all across the United States by shifting local public resources toward evidence-based, results-driven solutions. Results for America is interested in the CLD3 process to scale and sustain the work that they are doing with Bloomberg's What Works Cities. **Nicole Dunne, Vice President, Innovation and Community Impact**, described the Results for America and CLD3 goals as the same, that is, to shift local public resources toward evidence-based, results-driven solutions to create the new normal.

Results for America is the campaign manager for What Works Cities, a national initiative designed to help 100 mid-size U.S. cities enhance their use of data and evidence to improve the lives of their residents. These cities have publicly committed to enhance their use of data and evidence to improve services, inform local decision-making and engage residents. They also launched the Local Fellowship Program in which fellows participate in a 12-month initiative designed to help their local governments better invest in what works. They have enlisted the support of local, state, and national leaders as Moneyball for Government "All-Stars" as part of the Results for America Moneyball for Government campaign. We are developing a relationship with Results for America and a partnership seems likely. **Maia Jachimowicz, Vice President, Evidence-Based Policy Implementation**, expressed support for working with CLD3. She has introduced us to the Local Government Fellows across six cities to explore opportunities for Virginia Tech to bring statistical and social and behavioral expertise to bear on their issues. We have met with the following fellows: Jennifer Reed, Chief Performance Officer at Government of the District of Columbia, Gilbert Montano, Chief of Staff, City of Albuquerque, and David Gottesman, CountyStat Manager, Office of the Montgomery County, Maryland Executive ,

Actionable Intelligence for Social Policy (AISP) at University of Pennsylvania

Dennis Culhane co-founded **Actionable Intelligence for Social Policy (AISP**) at the University of Pennsylvania. AISP is an initiative that focuses on the development, use, and innovation of integrated data systems for policy analysis and program reform (via 13 sites across the US). . They have 13 sites across the United States to implement their mission. **Iowa State and Virginia Tech are partnering with AISP on multiple initiatives** that will inform CLD3 and have been invited by AISP to contribute to the following AISP projects:

- NSF Midwest Big Data Hub; Sarah Nusser is co-PI;
- LJAF funded Public-Academic Research Colloquium on November 29-30, 2016 – Virginia Tech Social and Decision Analytics Lab is a co-organizer;
- Project to instruct counties/cities on how to create an Integrated Data System. Virginia Tech Social and Decision Analytics Lab is participating on three of the four teams (Policies and Practices, Research and Data Framework, and Innovation Technology).

MetroLab Network

Virginia Tech in collaboration with Arlington County is a member of the **MetroLab Network.** The collaboration allows for networking opportunities to share best practices as well as collaborate on important policy and planning research in Arlington. An example of best practices is their primer that presents 10 best practices for city-university partnership that closely align with our principals; for example, find the interaction between city (county) priorities and university expertise (see http://metrolab.heinz.cmu.edu/wp-content/uploads/2016/07/06-City-University-Partnership-Principals-6-22-2016.pdf)

Lab @ DC

David Yokum, the new Director of the Lab @ DC noted that CLD3 will provide the framework to connect the local and state programs together, on a regional basis as well as a state basis. The Lab @ DC emerges from a vision to systematically incorporate the scientific method into day-to-day governance. It is a team of applied scientists, working in the Office of the City Administrator and in partnership with a network of universities

and research centers, who conduct empirical projects to generate timely, relevant, and high-quality evidence that informs the District's most important decisions.

- David Yokum, Senior Policy Advisor, Office of the City and Director of the Lab @ DC
- Sam Quinney, Applied Research Analyst, Lab @ DC, Office of the City Administrator, Washington DC
- Jennifer Reed, Chief Performance Officer at Government of the District of Columbia, and Results for America Local Government Fellow.

National Association of Counties

The **National Association of Counties, (NACo)** work with elected government officials at the local level. This is complementary to the emphasis of working with civil servants in CLD3. NACO represents all counties and provides data through a County Explorer that has over 900 indicators. See http://explorer.naco.org/

- Emilia Istrate, NACo's Director of Research and Outreach, is responsible for developing NACo's research priorities that support NACo's strategic plan and mission. http://www.naco.org/people/emilia-istrate
- Maeghan Gilmore is the program director for health, human services and justice issues in the County Solutions and Innovation (CSI) department. In this capacity, she develops and directs the association's programs that provide information, training and assistance to county officials to help meet the emerging challenges they face in their work.

Arthur Scott, Associate Legislative Director - Agriculture & Rural Affairs; Rural Action Caucus Liaison will introduce CLD3 to Senator Kaine. He invited Vicki Lancaster, Virginia Tech SDAL statistician, to speak about CLD3 at the annual NACo Rural Action Caucus meeting on December 8, 2016. The CLD3 ideas were enthusiastically embraced.

Interactions with Local Public Health and Data Science Offices

We have discussed CLD3 with local officials who are both curious and interested in learning more about CLD3 and engaging in the journey to deploy it.

- David Lessinger, Chief of Staff to the Chief Administrative Officer at the City of New Orleans
- Denice Ross, OMB, who was in New Orleans before joining OMB. She is focused on the nexus of open data and solution-development at the local level.
- Abdul El-Sayed, MD, DPhil, Executive Director & Health Officer, and Eric Kessel, Public Health Division Administrator, Health Policy and Planning at City of Detroit, Detroit Health Department. https://www.statnews.com/2016/05/03/detroit-health-director/
- Karen Hacker, M.D., M.P.H., Director, Allegheny County Health Department. http://www.achd.net/admin/director-hacker.html

Interactions with Other Land Grant Universities

We have discussed CLD3 and the national vision with several other Land Grant Universities to access their interest in joining the movement. All were excited about the vision and would like to be included in the next wave of state CLD3 deployments. These include:

- Purdue University, Suresh Garimella, Vice President for Research
- Colorado State University, Bruno Sobral, Director of the One Health Institute
- University of District of Columbia, Sabine O'Hara, Dean of Agriculture

In addition, we have been approached by local and state government officials involved with Cooperative Extension Services in Minnesota, Delaware, and New Jersey asking for help to be considered in future partnerships.