#### S&CC-IRG Preliminary Proposal Track 2: Building Local Government Capacity for Data-Driven Governance and Regional Analysis

#### Aaron Schroeder (PI), Virginia Tech and Alan Dotts (co-PI), Iowa State University

#### Summary

#### Data flows relevant to local 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. Local government decision-making based on these data are instrumental to increasing efficiency and improving outcomes. Despite the enormous potential, many local governments lack the general knowledge or technical and analytical expertise to utilize these data sources to inform their decision-making process. While many issues that are the purview of local government (e.g. educational attainment, opioid addiction, disconnected youth, poverty, homelessness), many policy decisions end up being informed by anecdotes rather than data-derived evidence.

To enhance their capacity, we are working directly with local governments to provide 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 presently being achieved through the development and deployment, and collaborative use with the community of, a set of **data science processes** and **supporting technological platforms**. Capacity is being gained by participating communities as they work collaboratively with issue-area researchers through the processes necessary to prepare these data for use in the decision-making process.

The development of a more formalized, scalable, portable, and inter-jurisdictionally secure version of these processes and platforms and piloting it across Virginia and Iowa is the focus of this proposal. This new **Virtual Analytic Environment** will provide a more **comprehensive framework of** **secure** **data science processes and platforms** that will mediate access to data and data tools, adhere to necessary data privacy and security constraints, and, as it is deployed and used to build capacity in multiple communities, 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 accessibility by local communities, researchers, and state agencies.

**Intellectual Merit**

New institutional priorities across the country are focusing on the need for an increase in the amount and quality of community-based research. New partnership models are viewed as a method to achieve this and to increase the dissemination of such research in peer-reviewed journals. This project will help extend the type and reach of scholars who become involved in community-based research, as well as increasing the capacity for evidence-based decision making at the local level.

**Broader Impacts**

This research will have significant benefit to society through creating a replicable architecture capable of rapid translation of research to practice that will enable data-driven governance across rural and urban communities. This will include a **comprehensive framework of** **secure** **data science processes and 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.

## Project Description (4-page limit)

### Vision & Goals

In this project we are pursuing a vision of local governments acquiring the capacity for data-driven, inter-jurisdictional, evidence-based decision-making in their governance processes.

There is tremendous potential for enhancing the effectiveness of local government decision-making by bringing to bear analyses of all community-relevant data sources, regardless 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 communities 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.

The proposed research is a natural part of the outreach mission of our institutions,. We will work with local governments 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). This vision and 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.

To support this vision, we will achieve the following goals:

* Develop, provide, and pilot the implementation in four communities a new Virtual Analytic Environment, a **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 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 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 (VT/ISU) 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.

### 

### Integrative Research Approach

###### Describe clearly the proposed research including its plan for integrative research and community engagement. Define the community, associated stakeholders, and how the community will engage in meaningful ways with the proposed research. Briefly describe the evaluation approach including initial metrics.

Deploy an Initial Virtual Analytic Environment

To enable community engagement and our integrative research process, we will begin by deploying a supporting set of scalable, transferable, and cost-effective data management, data analytics, and data presentation technology platforms (see figure 1)

* 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.
* 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.
* 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.

Work Collaboratively with Partner Localities to Develop Capacity and help Solve Local Problems

Starting with community-chosen issue areas, VT issue-area researchers will work cooperatively with our partner communities (Arlington, Alexandria, Des Moines, Perry). We will begin building community capacity for managing the data science process by working collaboratively through the processes of community engagement and discovery, data ingestion and management, data fitness analysis, data analysis and hypothesis testing, and data & analysis results presentation.

1. During the Community Engagement & Discovery Process, we will work with the community to:
   1. confirm initial **hypotheses** regarding specific local issues that were already derived from previous work with the partner localities;
   2. 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;
   3. conduct a **data analytics capabilities assessment** to ascertain the community’s current level of data analysis expertise; and,
   4. 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.
   5. deploy necessary data connection technologies as required by an already established **data access plan** to enable the data transfer and management
2. During the **data ingestion & management process**, 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:
   1. 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?
   2. 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)
   3. establishing **data marshalling processes** for: system mediation logic, data pipeline and data transformation, transfer schedule, and data provenance maintenance
   4. 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)
3. During the **data fitness analysis process**, the community will gain capacity in understanding how to assess their data’s “fitness-for-use” by:
   1. applying a disciplined process for profiling data sources in terms of data quality structure and metadata (Keller et al. 2016);
   2. preparing selected data for analysis via cleaning, transformation and restructuring (Schroeder, others);
   3. updating the metadata repository, the Lexicon with the data information specific to the selected data (see discussion of the Lexicon below)
   4. creating linked datasets securely and at the highest possible level of accuracy (Schroeder, others);
   5. testing the prepared dataset(s) as adequate for use in proposed analyses/modelling approaches (Keller census report);
4. During the **data analysis and hypothesis testing process**, the community will gain the capacity to:
   1. create analytic experiments to test their hypotheses
   2. understand or even begin to create statistical models
   3. understand or even begin to perform statistical analyses to test their hypotheses

Continually Update the Lexicon, an Enhanced Metadata Repository,

Mediating the framework’s 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).

Deploy and Manage a Policy-Based Access Layer to Mediate All Data Access

Data 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, a data resource broker typically implements a logical namespace (distinct from physical file names) and maintains metadata on data-objects (files), users, groups, resources, collections, and other items in a metadata catalog stored in a relational database management system. This metadata can be queried to locate files based on attributes as well as by name across heterogeneous data resources. A data resource broker will generally have features to support the management and collaborative (and controlled) sharing, publication, replication, transfer, and preservation of distributed data collections.

### Research Capacity-Building

###### Describe activities towards building, developing, and engaging research talent to focus on S&CC integrative research challenges.

This project will help extend the type and reach of scholars who become involved in community-based research, as well as increasing the infrastructure for evidence-based decision making at the local community level. We will collaborate with faculty, staff, and students across our campuses to define the underlying research and resources to be deployed in this Virtual Analytic Environment. This type of dedication to quality community-based research is central to the mission of land grant universities across the country. Working closely with both campus researchers and Cooperative Extension Service specialists deployed in the communities across our states, we will be exposed to the important issues and challenges facing our cities and counties, thus generating bottom-up rather than top-down research, providing the opportunity for a greater impact on realizing data-driven government policy and interventions.

### Integration and Multi-Disciplinary Context

###### Characterize the multidisciplinary, holistic nature of the approach by identifying the disciplines involved in the research, and how the proposed research elements are integrated.

Definition of Transdisciplinary

Public Administration & Policy

Statistics

Economics

Computer Science

Psychology

### Management

###### Describe the participating institution(s) and key personnel briefly (see Project Personnel and Partner Institutions section). Address the roles and responsibilities of the project’s management, and how it will support integration of tasks at a high level and engagement with the community throughout the course of the planned research activities. Describe the approach to collaboration among project members.

I think you need to talk about you, Alan, and supporting postdocs and students to develop and implement the Virtual Analytic Environment. Here is where I would say that you will be leveraging on-going community-based research in the four pilot areas. That will provide for the needs and requirements of the infrastructure. Say that Stephanie Shipp is leading that work and will be your partner. I would list her as a senior researcher. From the Iowa side, Erin Mullenix should probably be listed in the similar role to Steph.

### Budget and Subawardees

###### Provide an estimated aggregate budget request with number of years of support requested. Provide a list of subawardee institutions.

Jim to provide – he probably needs to add 3% for Stephanie into the budget. I can ask.

## Project Personnel and Partner Institutions (1-page limit)

###### Provide current, accurate information for all personnel and institutions involved in the project. Follow the same format as described for Project Personnel and Partner Institutions in the Full Proposal Preparation Instructions below. Submit as a supplementary document.

Here I would just add a para bio for each of the two of you and of Stephanie and Erin.

* **Aaron Schroeder, Ph.D.**, is the Information Architect and Data Scientist the Social and Decision Analytics Laboratory(SDAL)within the Biocomplexity Institute of Virginia Tech. He is responsible for planning and executing major research projects focused on the techniques, methods, and theories related to the governance, integration, storage, retrieval, sharing, and optimal use of policy-relevant data, information, and knowledge for the purposes of policy analysis and program evaluation. A particular focus in this role has been on the integration and analysis of Virginia education, health, social service and non-profit administrative data streams for the purpose of conducting policy analyses and program evaluations impacting a wide range of constituents, including: pre-K child social and health service recipients; child care service operators; primary, secondary, post-secondary and adult education service recipients; state workforce training service recipients; and, U.S. veteran health and social service recipients.
* **Stephanie S. Shipp, Ph.D.,** is the Deputy Director of the Social and Decision Analytics Laboratory(SDAL)within the Biocomplexity Institute of Virginia Tech. Her areas of expertise are economic and demographic analyses of social issues using survey and administrative data, innovation and competitiveness, and evaluation of public programs. Prior to coming to SDAL, she was a senior Research Staff Member at the Institute for Defense Analyses, Science and Technology Policy Institute, where she led several evaluations of federal programs. As a member of the federal Senior Executive Service from 2001-2008, Dr. Shipp was the Director of the Economic Assessment Office that was responsible for evaluating the Advanced Technology Program at the National Institute of Standards and Technology. Prior to that, she led economic and statistical programs at the Census Bureau, the Bureau of Labor Statistics, and Federal Reserve Board. She is fellow of the American Association for the Advancement of Science, fellow of the American Statistical Association, and elected member of the International Statistics Institute.
* **Erin Mullenix, MPA,** is the Director of Data Driven Science at Iowa State University. In her role, Erin develops and directs programs that support campus-wide activities in data-driven science, including Iowa State’s partnership with the Midwest Big Data Hub. Erin has experience across the research and public service spectrum. She was the Research Director at the Iowa League of Cities prior to joining Iowa State this year. In this position, Erin worked closely with local communities on data related issues, providing data-driven support and evidence to local decision-makers. Before this, Erin worked in engineering, process improvement, safety, and state and federal government, providing broad expertise that she applies to program leadership.

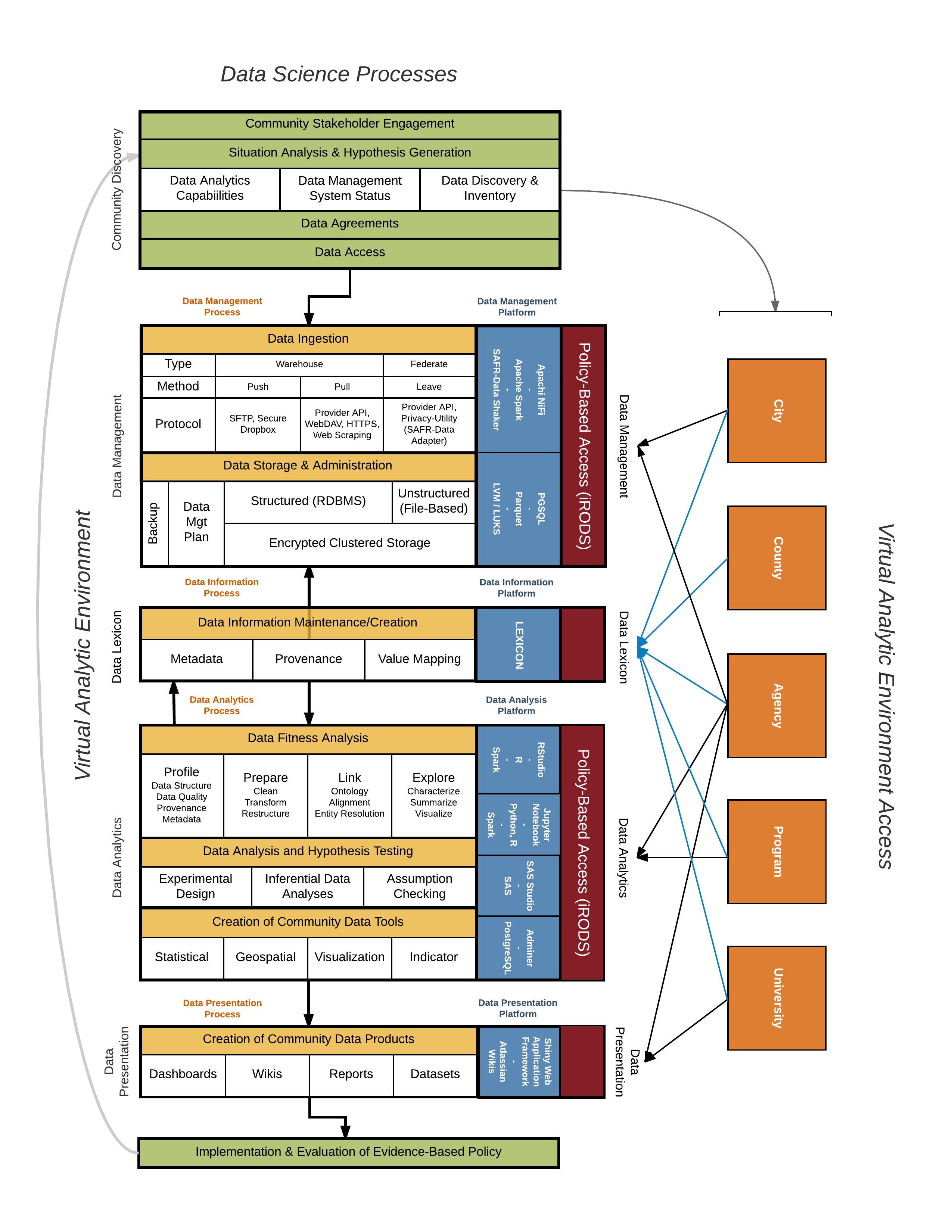


Figure 1: CLD3 Processes and Platforms Framework

References

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.

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.

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.

———. 2012b. “VT-VLDS Data Adapter User Guide.” Virginia Information Technologies Agency.

———. 2013a. “VT-VLDS Identity Resolution and Query Execution Process Overview.” Virginia Information Technologies Agency.

———. 2013b. “VT-VLDS Shaker Specifications.” Virginia Information Technologies Agency.