**Data Science Processes and Platforms Framework**

**Building Capacity for Data-Driven Governance**

#### 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 are instrumental to increasing efficiency and improving outcomes. This is especially true for many local governments where, despite the enormous potential, 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 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 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, 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 continuous accessibility by local communities, researchers, and state agencies.

**Description of the Data Science Processes and Platforms Framework**

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 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 (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.

**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 marshalling 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:
  + - apply a disciplined process for assessing data fitness for use through profiling data sources in terms of data quality structure and metadata (Keller et al. 2016);
    - 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/modelling approaches (Keller census report);
    - 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 to tools to create multiple forms of presentation for communicating their results and analyses, including:
  + - Statistical reports and visualizations
    - Interactive data dashboards
    - Interactive wikis

**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.

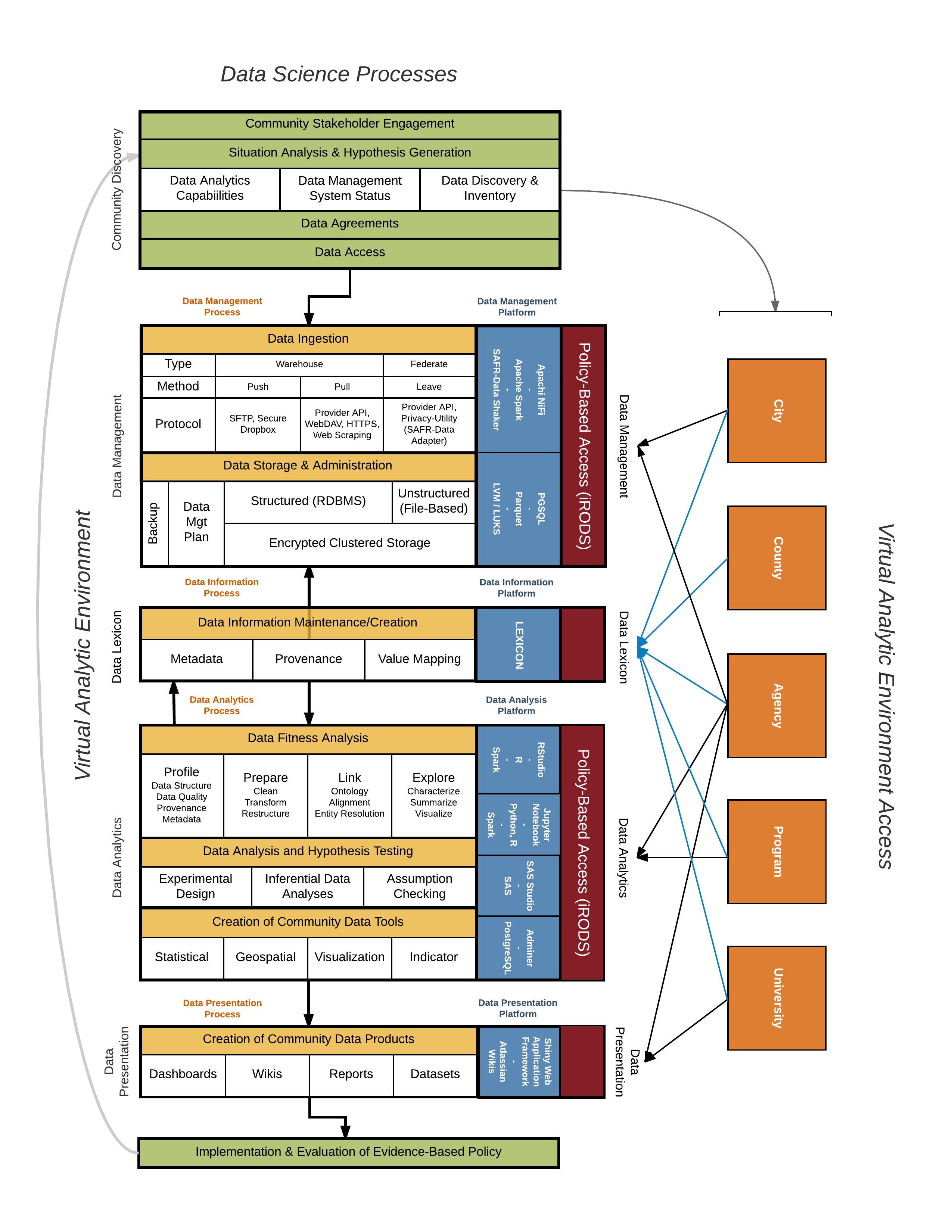


Figure 1: Data Science Processes and Platforms Framework