Proceedings of the Biocomplexity Institute
Technical Report 2022-448

Social Impact Data Commons Technical Review

Joel Thurston
Senior Scientist
jt9sz@virginia.edu

Aaron Schroeder
Research Associate Professor
ads7fg@virginia.edu

Kathryn Linehan
Research Scientist
kjl5t@virginia.edu

Micah Iserman
Postdoctoral Scholar
rtb8qy@virginia.edu

Social and Decision Analytics Division
Biocomplexity Institute & Initiative
University of Virginia

22 March 2022

Funding: This project was funded by a grant from the Mastercard Center for Inclusive Growth (#G-202104-02198).

Citation: Thurston, J., Schroeder, A., Linehan, K., & Iserman, M. (2022). Social Impact Data Commons technical review. *Proceedings of the Biocomplexity Institute*, Technical Report. #BI-2022-448. University of Virginia.

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE

Social Impact Data Commons

The University of Virginia's Social and Decision Analytics Division (SDAD) has partnered with the Mastercard Center for Inclusive Growth to build a Social Impact Data Commons capable of tracking social impact across the Washington DC metropolitan region over time. A "data commons" is an open knowledge repository that co-locates data from a variety of sources, builds and curates data insights, and provides tools for tracking issue over time and geography. A data commons enables governments and community stakeholders to continuously learn from and leverage their own data. SDAD and Mastercard have undertaken this project with the shared vision to use data to inform equitable growth and sustainable recovery from social challenges.

SDAD initiated a technical review of the data architecture and data management processes we have developed for the Social Impact Data Commons (see Appendix A for presentation material). The technical review served to advance several key aspects of the project, including furthering researcher engagement in the Social Impact Data Commons and addressing useability, iteration, and refinement.

The University of Virginia's Department of Engineering Systems and Environment (ESE) faculty completed the review (see Appendix B for a list of participating faculty members). ESE is a "leader in the study of human and socio-technical systems [and] create innovations that will help address society's most wicked problems, which are often emergent, large scale, and complex in areas ranging from public health to smart cities to environmental resilience."

Framing questions

The review focused on the project's basic architecture and capacity to accomplish its goals formalized in two questions:

- Where should we be mindful of potential complications or breakdowns in the processes that we have established?
- What suggestions did ESE have for improvements to our architecture and processes?

In addition to written feedback provided in advance, the project team and the ESE reviewers engaged in a robust discussion of the topics. The consensus from the reviewers was that the data commons is coming together nicely, good design choices have been made, and the research is on track to meet the project goals.

General architecture

During the discussion, we clarified that the data being used by the web interface are statically served from GitHub and that all subsequent data processing is done in-browser. A key

recommendation was to conduct rigorous testing to determine the size limits of this approach. As one ESE faculty member described it, we should "push our design until it breaks." This is an important consideration as we anticipate how the project might scale. One reviewer experienced poor performance when trying to select different geographic regions, and they suggested this may be due to the site's high memory usage. Though this also relates to the ongoing optimization of the site's code, we will need to monitor the effects of the data size as new geographic regions and measures are incorporated.

There was consensus that the use of open source components means that the project is well poised for sustainability. No part of the current data distributions or dashboard require ongoing support (such as a maintained server), and we are working to make the data updating process as automated as possible. This would ensure that the current data commons resources remain available even if we were forced to abandon the project due to unforeseen circumstances. The open source framework also maximizes the ability of the data commons to be replicated and deployed by other interested parties.

Community engagement process

In discussing the relationship between our community engagement model and the overall data commons design philosophy, several points were highlighted. While it was clear that the totality of the data commons is more than just a public-facing dashboard – including tools, metrics, training materials – we realized that the entirety of the process could and should be more explicitly denoted in our public-facing materials. For example, our dashboard currently does not highlight the Community Learning through Data Driven Discovery (CLD3) process, which serves as the guiding community engagement model behind the project development.

ESE reviewers asked about how we plan to evaluate community learning and shared insight into their engagement philosophy; they consider both usefulness and usability. While we have evaluation built into the proposal (e.g., this review is one component), we do not currently have plans to complete a learning assessment. We discussed strategies and means of collecting this information, such as:

- Tracking access and use of the public-facing dashboard.
- Using simple surveys to collect data, a strategy that ESE successfully uses.
- Creating a metric that tracks when stakeholders utilize the data commons to perform data-related tasks that they previously requested we do for them or where they were previously not utilizing data to make decisions.

We also discussed creating a series of commonly accessed data scenarios or use cases to display to users, which would provide those new to the data commons with a starting point of how they

can engage with the data. The use cases could include or be built from user testimonials and could be particularly useful to users with limited data acumen.

Dashboard user interface

A portion of the technical review focused on the user interface of the public-facing dashboard that will be used to display and disseminate the data. The questions for this component focused on the ease of finding and downloading data. During the discussion, we clarified that it was never our intention for the web interface to handle massive amounts of data. The primary goal of the dashboard is to help people quickly review data relevant to their topic of interest and decide whether they would like to download it for further use. This is driven by our overriding organizing principle that, while usability is necessary, the data commons must remain scalable and sustainable for expansion to the entire National Capital Region. There was agreement that the dashboard architecture is well poised to accomplish this objective.

Part of the discussion focused on lower-level, practical suggestions regarding the user interface. For example, we discussed the way time range and data selection is handled, which is something that our development team has been continuing to experiment with. The reviewers suggested replacing the individual year minimum and maximum inputs with a slider box or moving away from the year selection option altogether, as it is confusing how this affects the displayed data. One reviewer suggested it may help to move inputs into the main output display area, nearer to the outputs they affect. These suggestions fit with a range of other points on improving the placement and clarity of settings, such as rearranging the settings menu, adding more tooltips, and improving palette names.

Another set of suggestions focused on improving the selection process, between (a) allowing for more comparisons (e.g., building complex queries, displaying multiple regions and variables at a time), and (b) making it clearer when data are unavailable (e.g., not allowing selection of regions when their subregions are missing all data). This has also been an ongoing discussion by the development team. We have received similar feedback from stakeholders that have tested the platform. As we evolve the dashboard from a simple data downloading platform into an analysis platform, we will do so with an eye towards maximizing community engagement by providing users with the requested features.

Additional discussion

Coming into the meeting, the reviewers had questions about whether the data file creation process was automated and how we plan to update the dashboard when new data become available. The technical team lead provided details regarding our current data curation process. Although many of these steps were initially performed manually by project team members, we are experimenting

with the most effective approaches for automating data acquisition and integration. Our goal is to automate as many components of the process as possible.

Action items and next steps

Based on the technical review, we have changed or added:

- the selected year option to be more central and easier to adjust;
- the color assigned to NA values and added an NA display to the legend to make it clearer when data are missing;
- options to turn off or adjust map zoom animations; and
- the data format to reduce its size and improve performance.

We are currently in the process of addressing:

- performance, aiming for acceptable performance on systems with limited memory and processing power;
- means to analyze data more flexibly, such as by selecting multiple regions and variables;
- reactivity of the interface, in terms of both screen size (e.g., the menu layout on smaller screens) and data availability (e.g., the set of selectable subregions and variables based on their missingness);
- customizability, such as adjusting the relative size of output elements or turning them on and off;
- means of better contextualizing data and pointing users toward meaningful comparisons or regions of interest; and
- identifying and reducing points of frustration and unnecessary complexity, toward minimizing clicks and time to relevant display.

Future improvements will focus on changing or adding:

- an option to make the settings menu shift the main content, rather than overlay it;
- ways to build complex queries and download data independent of the table output;
- more tooltips and explanations about what settings do and what outputs mean;
- links to additional resources, such as retrieving data directly from GitHub and Dataverse (and future distributions) and mapping polygon data; and
- how selection inputs work so users can select multiple regions for comparison across comparable geographies.

Additionally, per one reviewer's suggestion, we will review comparable sites for inspiration to improve our dashboard. We have also begun construction of a website that encapsulates the entirety of the data commons to include all its components (e.g., tools, training materials, an explanation of the CLD3 process, methodology documentation) along with the public-facing dashboard.

Future considerations

We concluded our discussion with a review about how the data commons might evolve and grow in the future. One of the reviewers suggested building a desktop app that would pull in the data currently being hosted on open source sites. This would minimize backend issues and allow users to interact with data directly on their local machines. A user could customize their download selection of all data or a subset. Because these data would be stored on a user's hard drive rather than in memory (as it is with the browser interface) system performance may be enhanced. Even without a customization option, because we can track access to the online data, we could build a desktop app around the data that people most frequently access. There are modern frameworks available (e.g., Electron) that allow for cross-platform development.

Conclusion

Based on the feedback received from our colleagues in the Engineering Systems and Environment, we are confident that we are on a path to achieving a scalable and replicable technical model for the Social Impact Data Commons that will allow future deployments of similar data commons.

SOCIAL IMPACT DATA COMMONS

an innovative approach to inform equitable growth

UVA BIOCOMPLEXITY INSTITUTE

Center for Inclusive Growth

AGENDA

- Introductions
- Project Overview
- User Interface
- Architecture
- Discussion

AARON SCHROEDER

Research Associate Professor

Social and Decision Analytics

Biocomplexity Institute

University of Virginia

KATHRYN LINEHAN

Research Scientist

Social and Decision Analytics

Biocomplexity Institute

University of Virginia

SOCIAL IMPACT DATA COMMONS

VISION

A Social Impact Data Commons to Inform Equitable Growth

SOCIAL IMPACT DATA COMMONS

VISION

INFORMING EQUITABLE GROWTH

UVA | BIOCOMPLEXITY INSTITUTE

Center for Inclusive Growth

The **University of Virginia** and the **Mastercard Center for Inclusive Growth** have a shared vision to use data to inform equitable growth.

Local communities have data on policies, strategies, events and social behaviors but often lack the analytical tools to use their data to drive policy and strategy development.

Partnering, we can make a difference.

SOCIAL IMPACT DATA COMMONS

VISION

WHAT IS A DATA COMMONS?

An open knowledge repository that co-locates data from a variety of sources, builds and curates data insights, and provides tools designed to track issues over time and geography allowing governments and community stakeholders to learn continuously from their own data.

Key features:

- Data sources, collected and created
- Maps reflecting multiple geographies
- Composite metrics
- Navigation and capability to statistically explore the data
- Data download via web or API
- Metadata

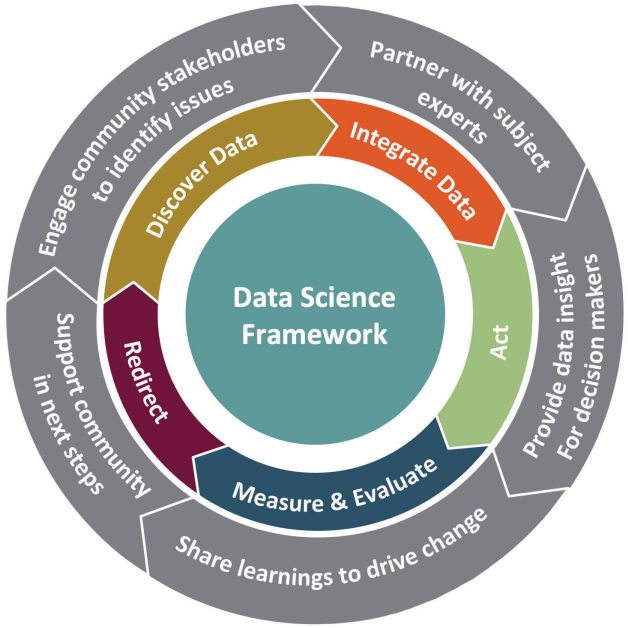
A Data Commons allows multiple audiences to explore issues relevant to their communities.

SOCIAL IMPACT DATA COMMONS

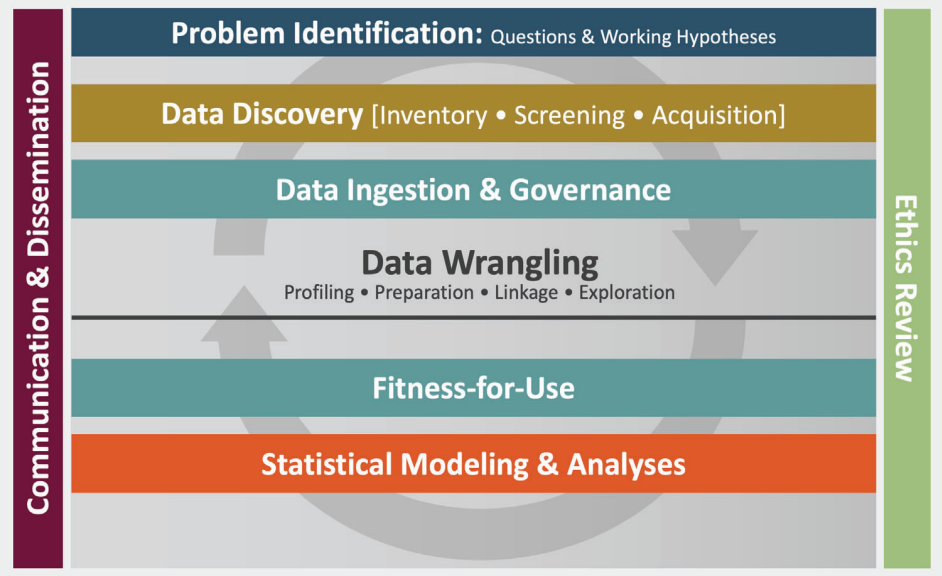
VISION

GUIDING PROCESS

Community Learning Through Data-Driven Discovery



Data Science Framework



SOCIAL IMPACT DATA COMMONS

PROJECT OVERVIEW

KEY ELEMENTS

Government, NGOs, researchers, and the public co-create shared goals, debate issues, formulate questions, ensure data are used ethically.

For the design, construction, deployment of the regional Social Impact Data Commons, and scalability of the data commons to other areas.

Disseminate and bring awareness of the Social Impact Data Commons regionally and beyond, inspiring the transferability of the research.

SOCIAL IMPACT DATA COMMONS

INTENDED IMPACT

- **Design and deploy a Social Impact Data Commons ...**
- **Capable of tracking social impact across the Washington, D.C. metropolitan region over time ...**
- **At the neighborhood and county/city levels.**

Community Learning
through
Data-Driven Discovery
+
Data Commons
=
More Agile, Smarter,
Responsive Governance

SOCIAL IMPACT
DATA COMMONS

VISION

WHY NOW?

Rapid growth expected over the next 25 years in the Washington, D.C. metro region. How this growth evolves will be significantly impacted by business as usual and specific impact activities.

SOCIAL IMPACT DATA COMMONS

PROJECT PHASES

Q1-Q2 Ground design and initial deployment on Arlington County focusing on select issues of interest

Q3-Q4 Expand geographic coverage to City of Alexandria, Washington, D.C., and Fairfax County

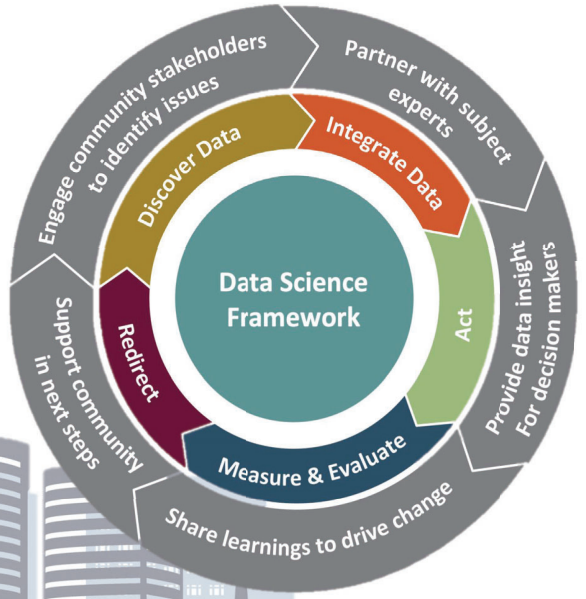
Q5-Q8 Expand to entire Washington, D.C. metro region

SOCIAL IMPACT DATA COMMONS

PROJECT OVERVIEW

ARLINGTON COUNTY ENGAGEMENT

- Identify initial issues worthy of data insights for near-term actions
- Exercise the Community Learning through Data-Driven Discovery (CLD3) process to frame the issues in a data context
- Build proto-type data commons
- Tech review of baseline architecture and data processes



SOCIAL IMPACT DATA COMMONS

TECH REVIEW

SOCIAL IMPACT
DATA COMMONS

TECH REVIEW

USER INTERFACE

Finding Data

Can you use the **top menus** to find the data you want to display?
Can you display it at your chosen **geography level**?
Is it clear what the **menu options** are and how to use them?

Accessing Data

Mouse hover versus mouse click to highlight information?
Is there other information we should add in the **information pane**?
In the **Settings tab**, what is most important?

Downloading Data

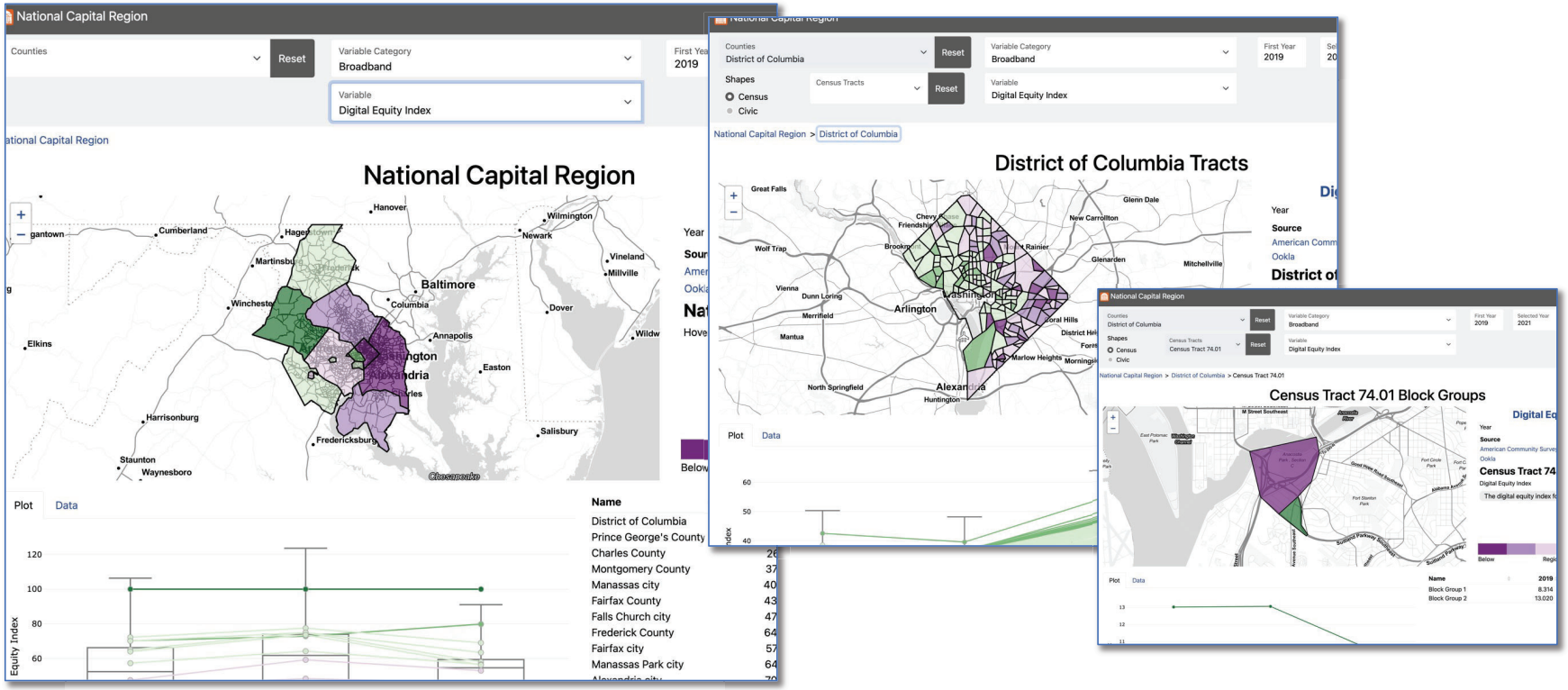
Did you notice where you can **see, print, and download data**?

SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Lightweight JavaScript WebApp, Universally Deployable:

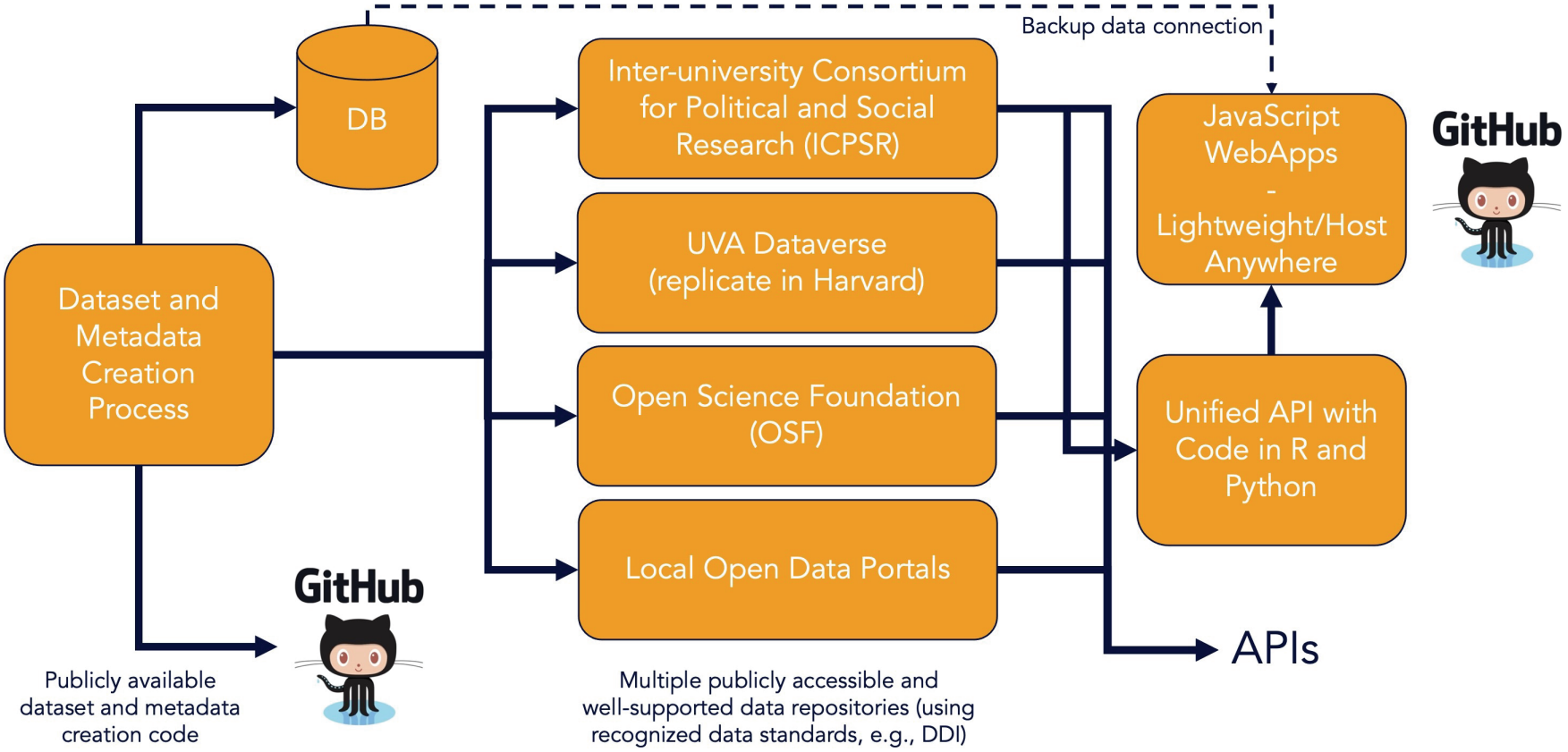
NCR Version (this project: beta): https://uva-bi-sdad.github.io/capital_region/
VDH Version (sister project: deployed, VA only): https://uva-bi-sdad.github.io/vdh_rural_health_site/



SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Modular, Sustainable, Expandable

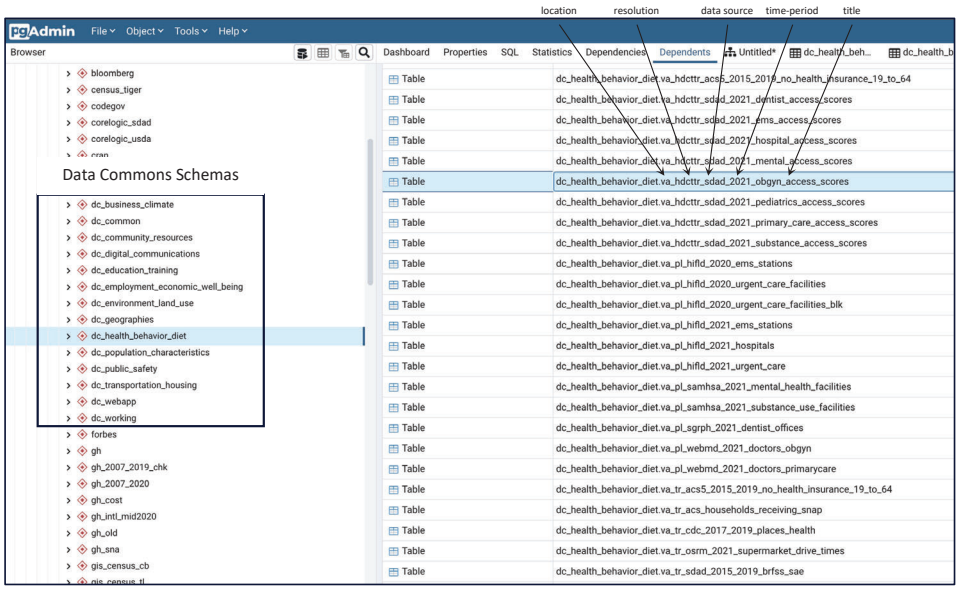


SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Database Structure & Standardization

Standardized Table Name Formatting



Standardized Table Columns

geoid text	region_type text	region_name text	year integer	measure text	value double precision	measure_type text
51059	county	Fairfax	2021	obgyn_fca	0.017075707143007	index
51600	county	Fairfax	2021	obgyn_fca	0.0169828091557208	index
51059	county	Fairfax	2021	obgyn_2sfca	0.000739520762320393	index
51600	county	Fairfax	2021	obgyn_2sfca	0.000748218893939039	index
51059	county	Fairfax	2021	obgyn_e2sfca	0.000883443656865259	index
51600	county	Fairfax	2021	obgyn_e2sfca	0.0009780990462416	index
51059	county	Fairfax	2021	obgyn_3sfca	0.000848524243580422	index
51600	county	Fairfax	2021	obgyn_3sfca	0.000823172803873302	index
51059	county	Fairfax	2021	obgyn_mean10	6.27929244708732	mean
51600	county	Fairfax	2021	obgyn_mean10	4.82797623924621	mean
51059	county	Fairfax	2021	obgyn_median10	6.41865971630388	median
51600	county	Fairfax	2021	obgyn_median10	4.99380888979926	median
51059	county	Fairfax	2021	obgyn_cnt	562	count
51600	county	Fairfax	2021	obgyn_cnt	12	count
51059	county	Fairfax	2021	obgyn_pop_cnt	468177	count
51600	county	Fairfax	2021	obgyn_pop_cnt	9764	count
12	health district	Fairfax	2021	obgyn_fca	0.0170799310241185	index
12	health district	Fairfax	2021	obgyn_2sfca	0.00073979355002781	index
12	health district	Fairfax	2021	obgyn_e2sfca	0.00088888629377427	index
12	health district	Fairfax	2021	obgyn_3sfca	0.000847544852942583	index
12	health district	Fairfax	2021	obgyn_mean10	6.21045700397026	mean
12	health district	Fairfax	2021	obgyn_median10	6.35059801962191	median
12	health district	Fairfax	2021	obgyn_cnt	611	count
12	health district	Fairfax	2021	obgyn_pop_cnt	483847	count
51610	county	Falls Church	2021	obgyn_median10	3.19834067050457	median
51610	county	Falls Church	2021	obgyn_fca	0.0175753292479828	index
51610	county	Falls Church	2021	obgyn_cnt	37	count
51610	county	Falls Church	2021	obgyn_2sfca	0.000747488753642069	index

Locations (2 characters for state/province or country;
3 fips characters for sub-state/province)
us, United States
va, Virginia
va013, Virginia, Arlington County

Resolutions (2 characters)
bl, census block
bg, census block group
tr, census tract
nb, neighborhood
ct, county
hd, health district
co, country
pl, place locations
pr, person data
bz, business data

Data Sources (up to 5 characters; this list will continually grow)
acs5, American Community Survey 5-Year Data
lodes, LEHD Origin-Destination Employment Statistics
pseo, Post-Secondary Employment Outcomes
qwi, Quarterly Workforce Indicators
mcig, Mastercard Inclusive Growth Score
hifld, Homeland Infrastructure Foundation-Level Data
ookla, OOKLA for Good
webmd, Web MD
sdad, (items that we have calculated)
abc, census address block counts

SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Standardized Geographic Area Names

Kept in schema dc_geographies. Tables, so far, for:

- DC, MD, VA; Counties, Census Tracts, and Block Groups: dc_geographies.ncr_cttrbg_tiger_2010_2020_geo_names
  - Includes all of DC, MD and VA
  - Needs to be filtered to get only NCR counties, tracts, block groups
- DC, MD, VA; School Districts: dc_geographies.ncr_sd_nces_2021_school_district_names
  - Includes all of DC, MD, and VA
  - Needs to be filtered to get only NCR school districts
- Arlington, VA; Civic Associations: dc_geographies.va_013_arl_2020_civic_assoc_geo_names
- VA; health districts: dc_geographies.va_hd_vdh_2021_health_district_geo_names

Each table has three columns: geoid, region_name, and region_type

select * from dc_geographies.ncr_cttrbg_tiger_2010_2020_geo_names

geoid	region_name	region_type
51540	Charlottesville city, Virginia	county
51510	Alexandria city, Virginia	county
51530	Buena Vista city, Virginia	county
51600	Fairfax city, Virginia	county
51610	Falls Church city, Virginia	county
51185	Tazewell County, Virginia	county
51683	Manassas city, Virginia	county
51820	Waynesboro city, Virginia	county
51790	Staunton city, Virginia	county

select * from dc_geographies.ncr_sd_nces_2021_school_district_names

geoid	region_name	region_type
5100033	A. LINWOOD HOLTON GOV SCH	school district
5100060	ACCOMACK CO PBLC SCHS	school district
5100090	ALBEMARLE CO PBLC SCHS	school district
5100120	ALEXANDRIA CITY PBLC SCHS	school district
5100152	ALLEGHANY CO PBLC SCHS	school district
5100078	ALT ED PRGM/BEHAV DISORD YOUTH/MONTGOMERY	school district
5100180	AMELIA CO PBLC SCHS	school district
5100013	AMELIA-NOTTOWAY VOC CTR	school district
5100210	AMHERST CO PBLC SCHS	school district
5100240	APPOMATTOX CO PBLC SCHS	school district
5100032	APPOMATTOX REGIONAL GOV SCH	school district

select * from dc_geographies.va_013_arl_2020_civic_assoc_geo_names

geoid	region_name	region_type
51013_ca_01	Williamsburg	neighborhood
51013_ca_02	Old Dominion	neighborhood
51013_ca_03	Maywood	neighborhood
51013_ca_04	Lyon Park	neighborhood
51013_ca_05	Dominion Hills	neighborhood
51013_ca_06	Alcova Heights	neighborhood
51013_ca_07	Columbia Heights	neighborhood
51013_ca_23	Arlington View	neighborhood
51013_ca_08	Leeway Overlee	neighborhood
51013_ca_09	Clarendon - Courthouse	neighborhood
51013_ca_10	Bluemont	neighborhood

select * from dc_geographies.va_hd_vdh_2021_health_district_geo_names

geoid	region_type	region_name
51_hd_01	health district	Alexandria
51_hd_02	health district	Alleghany
51_hd_03	health district	Arlington
51_hd_04	health district	Central Shenandoah
51_hd_05	health district	Central Virginia
51_hd_06	health district	Chesapeake
51_hd_07	health district	Chesterfield
51_hd_08	health district	Chickahominy
51_hd_09	health district	Crater
51_hd_10	health district	Cumberland Plateau
51_hd_11	health district	Eastern Shore

SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Data Creation, Storage, and Retrieval Process

Data file Creation

- Input: Raw data from sources
- Input: Dataset creation scripts (.R, .Rmd, .py)
- Output: Data Tables written to SDAD Database
- Output: GitHub Data Creation Repo

Metadata Creation

- Input: Metadata from sources
- Input: Relevant literature
- Input: Statistical methods information
- Output: Metadata files (.json)
- Output: GitHub Data Creation Repo

Dataset Storage

- Input: Standard template for Dataset R Packages
- Output: R Packages containing data and data creation scripts (versioned)
- Output: Dataverse Datasets containing standardized, prepared data (versioned)
- Output: Metadata R Package and Dataverse Dataset (versioned)
- Output: Updated GitHub Data Creation Repo

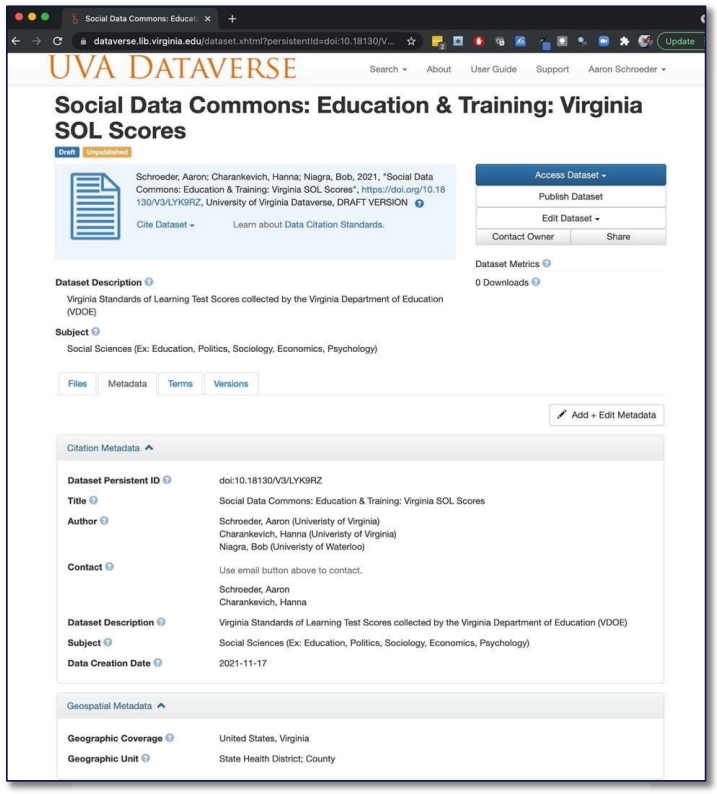
Dataset Retrieval

- Input: Dataverse API calls to access data and metadata files from Dataverse
- Reconcile data with latest website deployment
  - e.g. Checking for standards adherence, Checking for changes made

SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Dataverse Storage: Leveraging free, well-supported, data and metadata storage and download



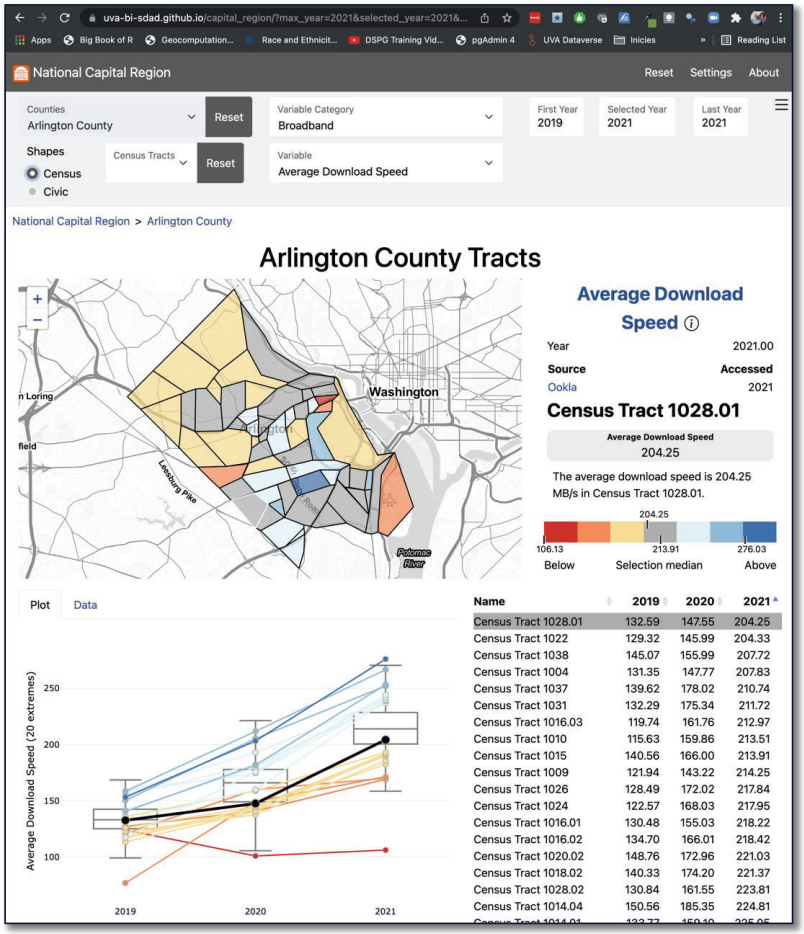
SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Equity of Broadband Example

Multiple Measures to Tell the Story

- Average download speeds (from Ookla) are relatively high across Arlington with the slowest average still above 100Mb (the newer standard for "broadband")



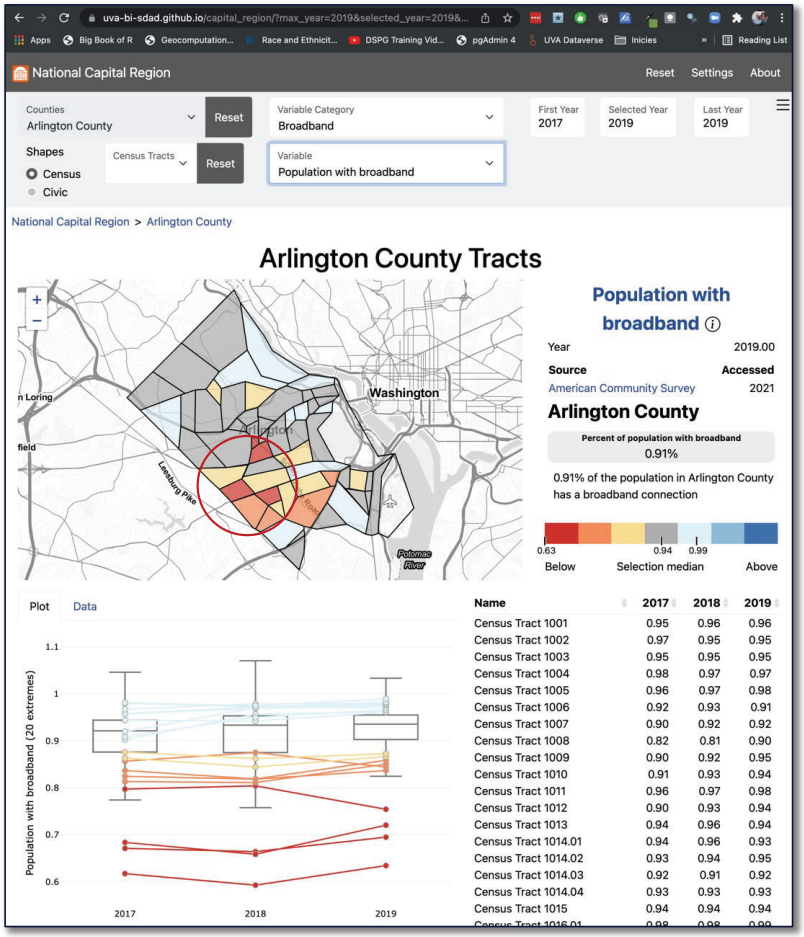
SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Equity of Broadband Example

Multiple Measures to Tell the Story

- However, specific areas can be identified that have a significantly lower level of broadband adoption the the rest of Arlington



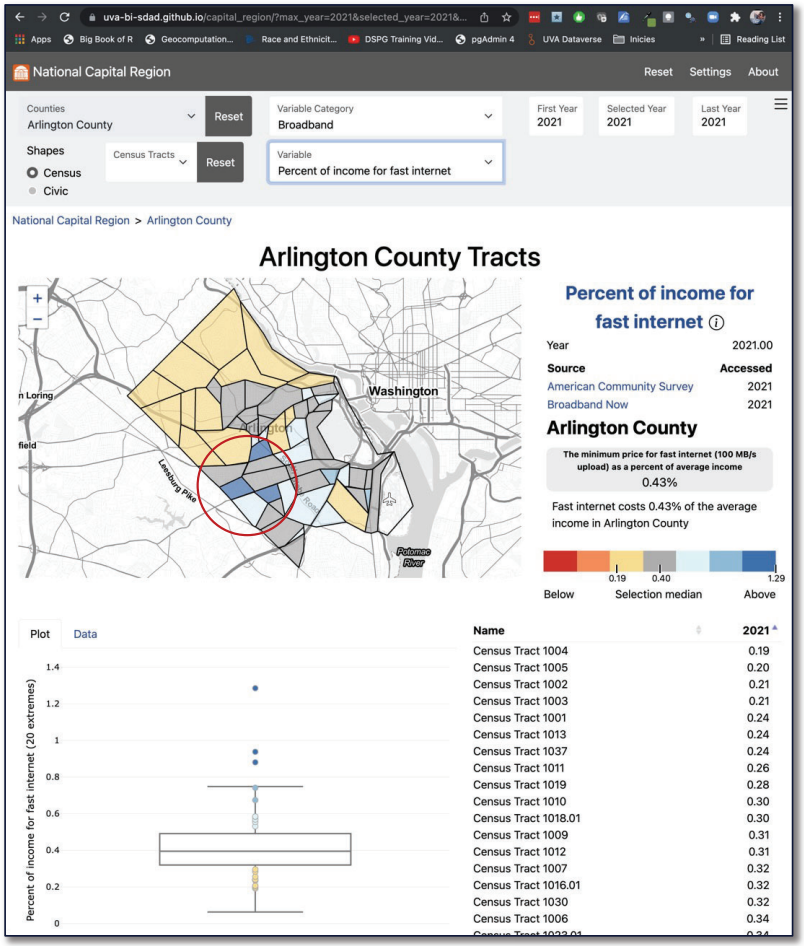
SOCIAL IMPACT DATA COMMONS

DATA COMMONS GENERALIZED ARCHITECTURE

Equity of Broadband Example

Multiple Measures to Tell the Story

- These areas of lowest broadband adoption appear to directly correlate with the areas having the highest ratio of household income to the cost of broadband, indicating an economic issue, as opposed to an issue of availability.



SOCIAL IMPACT DATA COMMONS

SUSTAINABILITY

SOCIAL IMPACT
DATA COMMONS

SUSTAINABILITY

VALUE PROPOSITION

- Create a capability that is used and relied upon by governments, NGOs, business and industry, and the public
- Become the authoritative *Regional Gold Standard* of data and metrics that transcend jurisdictions and neighborhoods, allowing transparent community comparisons
- Close the geographic and time gaps in current data sources

UVA BIOCOMPLEXITY INSTITUTE

Center for Inclusive Growth

Partnering, we can make a difference.

SOCIAL IMPACT DATA COMMONS

Appendix B

Brian L. Smith, PE
Professor and Chair
https://engineering.virginia.edu/faculty/brian-l-smith-pe

Devin K. Harris
Professor
Faculty Director - Clark Scholars Program
Director - Center for Transportation Studies
https://engineering.virginia.edu/faculty/devin-k-harris

William T. Scherer
Professor
Editor-in-Chief Systems
Associate Chair ESE, Academic Programs
Director, Accelerated Masters Degree Program (AMP
https://engineering.virginia.edu/faculty/william-t-scherer

Laura Barnes
Associate Professor
https://engineering.virginia.edu/faculty/laura-barnes

Mehdi Boukhechba
Assistant Professor, Academic General Faculty, Research Track
https://engineering.virginia.edu/faculty/mehdi-boukhechba

Afsaneh Doryab
Assistant Professor, Systems Engineering
Assistant Professor, Computer Science
https://engineering.virginia.edu/faculty/afsaneh-doryab

Arsalan Heydarian
Assistant Professor
https://engineering.virginia.edu/faculty/arsalan-heydarian

Majid Shafiee-Jood
Assistant Professor, Academic General Faculty, Research Track
https://engineering.virginia.edu/faculty/majid-shafiee-jood