Smart Metadata in Action
The Social Impact Data Commons

Presented by Joanna Schroeder 0000-0003-1514-5694
Social and Decision Analytics Division

2024 Conference on Smart Metadata for Official Statistics

UVA | BIOCOMPLEXITY INSTITUTE

Center for Inclusive Growth

Outline

- Motivating the Data Commons
- Our Approach
- Project Overview
- Evaluating Metadata Standards
- Case Study: Core Metadata
- Conclusion

Motivating the Data Commons

Informing equitable 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 drive policy and strategic development.

UVA | BIOCOMPLEXITY INSTITUTE

Center for Inclusive Growth

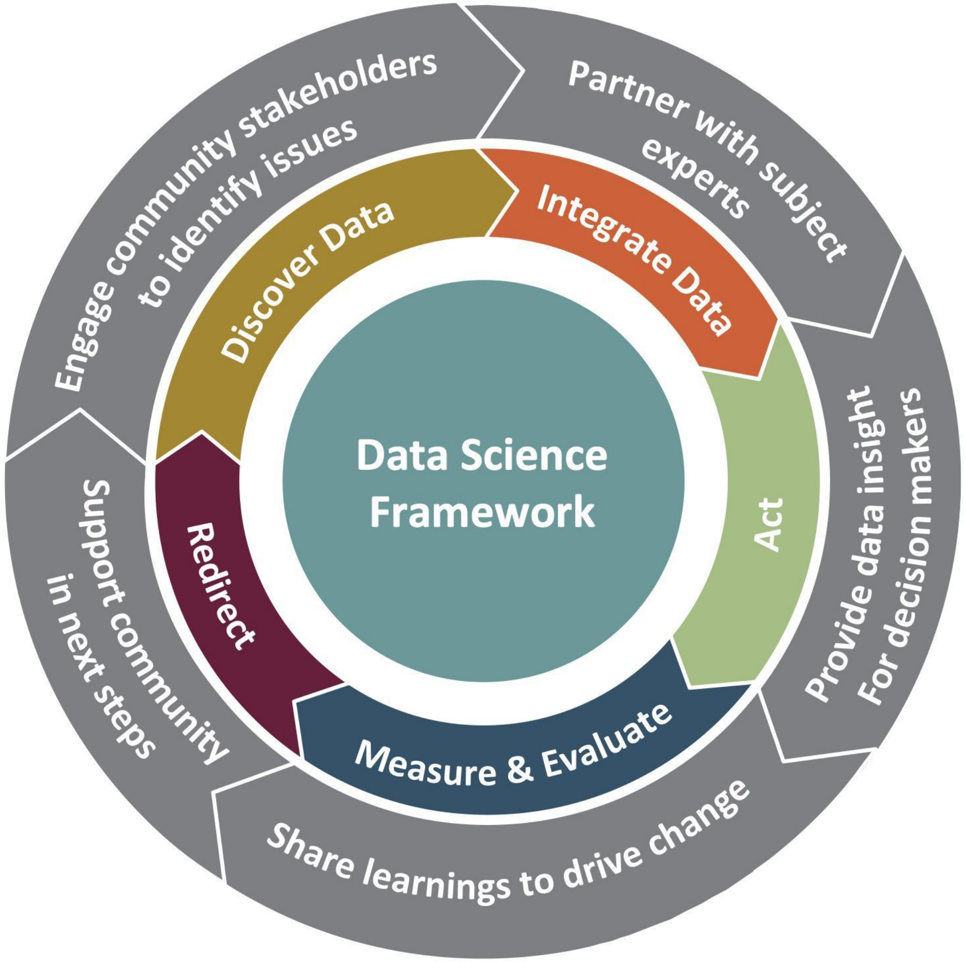
Our approach

Iterative process with decision makers

Community Learning Through Data Driven Discovery (CLD3)

CLD3 goes beyond traditional organizing aspects of collective action programs and supports communities in building capacity for data-informed decision making.

- Outer wheel: continuous interaction and communication across stakeholders
- Middle wheel: data-driven learning process
- Frontier: between the outer and middle wheels is active collaboration between all partners
- Inner circle: rigorous research framework to guide the data science


Engage community stakeholders to identify issues
Partner with subject experts
Provide data insight For decision makers
Share learnings to drive change
Support community in next steps
Discover Data
Integrate Data
Act
Measure & Evaluate
Redirect
Data Science Framework

Doing Data Science: A Framework and Case Study. Harvard Data Science Review, 2(1). (2020). S.A. Keller, S.S. Shipp, A. D., Schroeder, & G. Korkmaz

Our approach

Grounding in specific local issues

- Virginia Rural Health Data Commons
- Social Impact Data Commons
- Fairfax Women and Girls Data Commons

- Issues were selected on the basis that policy makers can **immediately benefit from the provision of new data and metrics** and will serve as **powerful exemplars** to showcase the impact and value of the Data Commons as the project expands.

Project Overview

Data Commons Components

Data Repositories

- New datasets supporting local decision-making
- Below county geographies
- Open access data & code (Fully reproducible)

Examples:
Broadband, Food, Financial Well-Being

Tools & Methods

- New open-source tools for building dashboards and datasets
- New methods for calculating new measures

Examples:
Demographic Redistribution, Food Insecurity

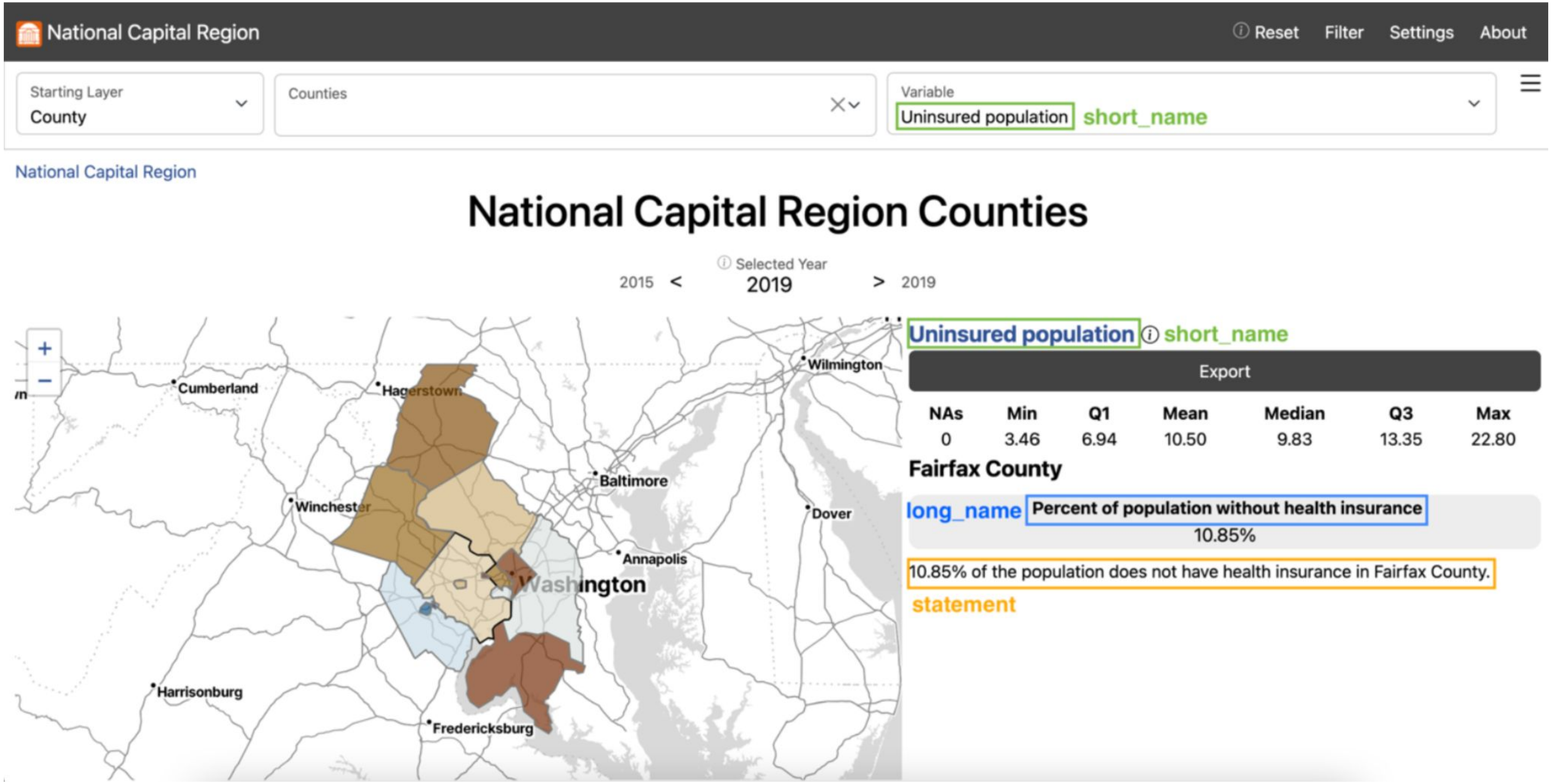
Applications

- Open-source applications & code for assessing and accessing data

Examples:
Dashboard, APIs

Active Core Metadata

Core metadata supports the creation and dissemination of statistical products



Starting Layer: County

Counties

Variable: Uninsured population short_name

National Capital Region

National Capital Region Counties

ⓘ Selected Year: 2019

2015 < > 2019

Uninsured population ⓘ short_name

NAs	Min	Q1	Mean	Median	Q3	Max
0	3.46	6.94	10.50	9.83	13.35	22.80

Fairfax County

long_name **Percent of population without health insurance**

10.85%

10.85% of the population does not have health insurance in Fairfax County.

statement

Core Metadata

Data commons core metadata is a custom standard

- Includes 17 elements
  - Describe datasets accurately and richly
  - Support the creation of the dashboard
- Disseminated in a JSON file on GitHub

- Examples:
  - **category** (internally controlled vocabulary)
  - **long_description** (free text)
  - **aggregation_method** (derived from DDI/OECD AggregationMethod)
  - **citations** (derived from BibTeX/LaTeX)

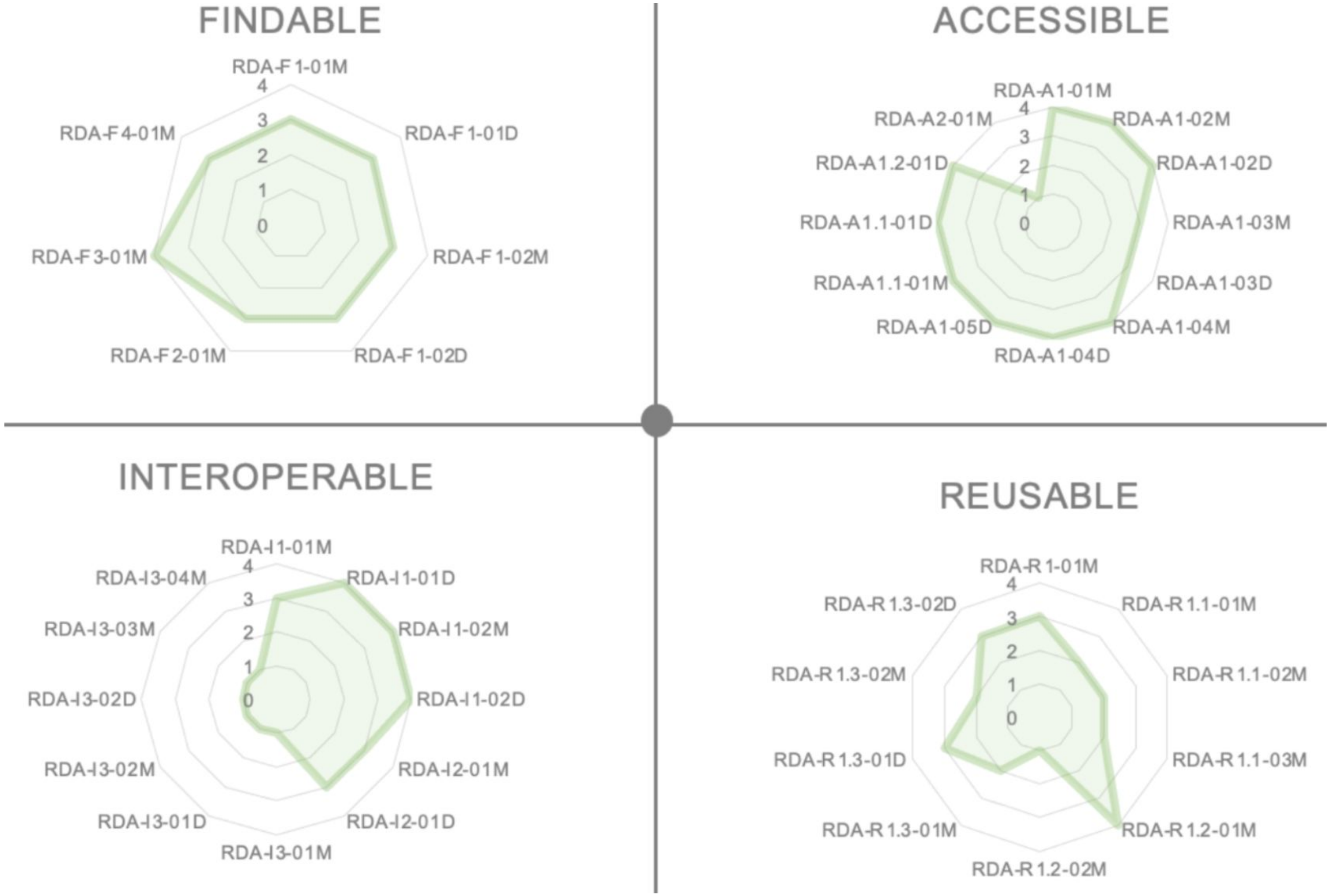
Evaluating Core Metadata

Metadata and standards are evaluable

- We evaluate our evolving metadata and standards against the FAIR guidelines
  - Self-evaluation
  - RDA FAIR Data Maturity Model
- Internal testing infrastructure
  - Evaluates the entire Data Commons system
  - Uses a GitHub runner
  - Each day the GitHub runner clones (refreshes) all linked data repositories
  - Next the runner executes specified tests (python scripts) on the data repositories
  - Example:
    - check_percent_data (Checks if percent data is in the range 0-100)
    - test_measure_info_key (Checks whether measure info files have valid keys for each variable)

Evaluating Core Metadata

RDA FAIR Data Maturity Model



One Essential indicator we have not considered: Metadata is guaranteed to remain available after data is no longer available

Results of our self-evaluation using the RDA FAIR Data Maturity Model

SOCIAL IMPACT
DATA COMMONS

Evaluating Core Metadata

Internal testing schema

- test_measure_info_structure (Checks whether measure_info files have and only have a prescribed list of allowable keys)
    - 60.3% valid
    - Poorly specified test
- test_measure_info_missing_measures (Checks whether measure_info files are missing any measures contained in corresponding data files)
    - 66.7% valid
    - Poorly defined standard for geography data
- test_measure_info_keys (Checks whether measure_info files have valid keys for each variable)
    - 87.5% valid
- test_jsons (Checks whether encountered JSONS are valid JSONS that can be read)
    - 100% valid
- test_measure_info_extra_measures (Checks whether measure_info files have any extra measures not contained in any corresponding data files)
    - 97.4% valid

Results as of 2023-08-03

SOCIAL IMPACT DATA COMMONS

Next Steps

- Accommodate more external collaborators (i.e. dataset producers external to the University of Virginia) and more stakeholders
- Perform user testing
    - Evaluate metadata richness and accuracy for external users.
- Develop a crosswalk of our standard to domain-accepted standards
    - Data Documentation Initiative (DDI)
    - DataCite
    - Data Catalog Vocabulary (DCAT)
    - schema.org
- Push datasets on external repositories will make our datasets more discoverable

Conclusion

- The Data Commons uses actionable and evaluable core metadata
  - to build data products
  - to support the dissemination of statistical products
  - to reduce the documentation burden on researchers
- We are progressing our adherence to FAIR standards
  - Interoperable metadata is the biggest area for improvement
- We have improved the our metadata literacy as a research lab
  - We hope to have instilled an appreciation for metadata within researchers

Co-authors: A. Wang (0000-0001-6926-4336),
K. Linehan (0000-0001-9012-6261),
J. Thurston (0000-0002-3923-9065),
A. Schroeder (0000-0003-4372-2241)

Appendix

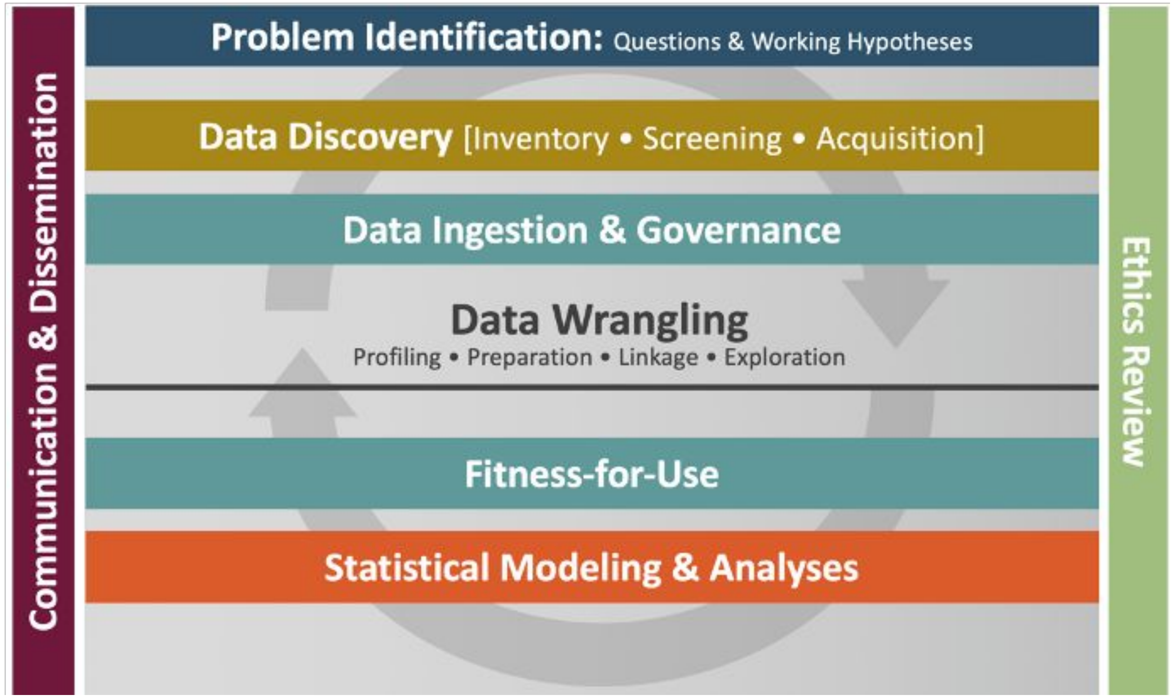
Our approach

Iterative process with local decision makers

The Data Science Framework

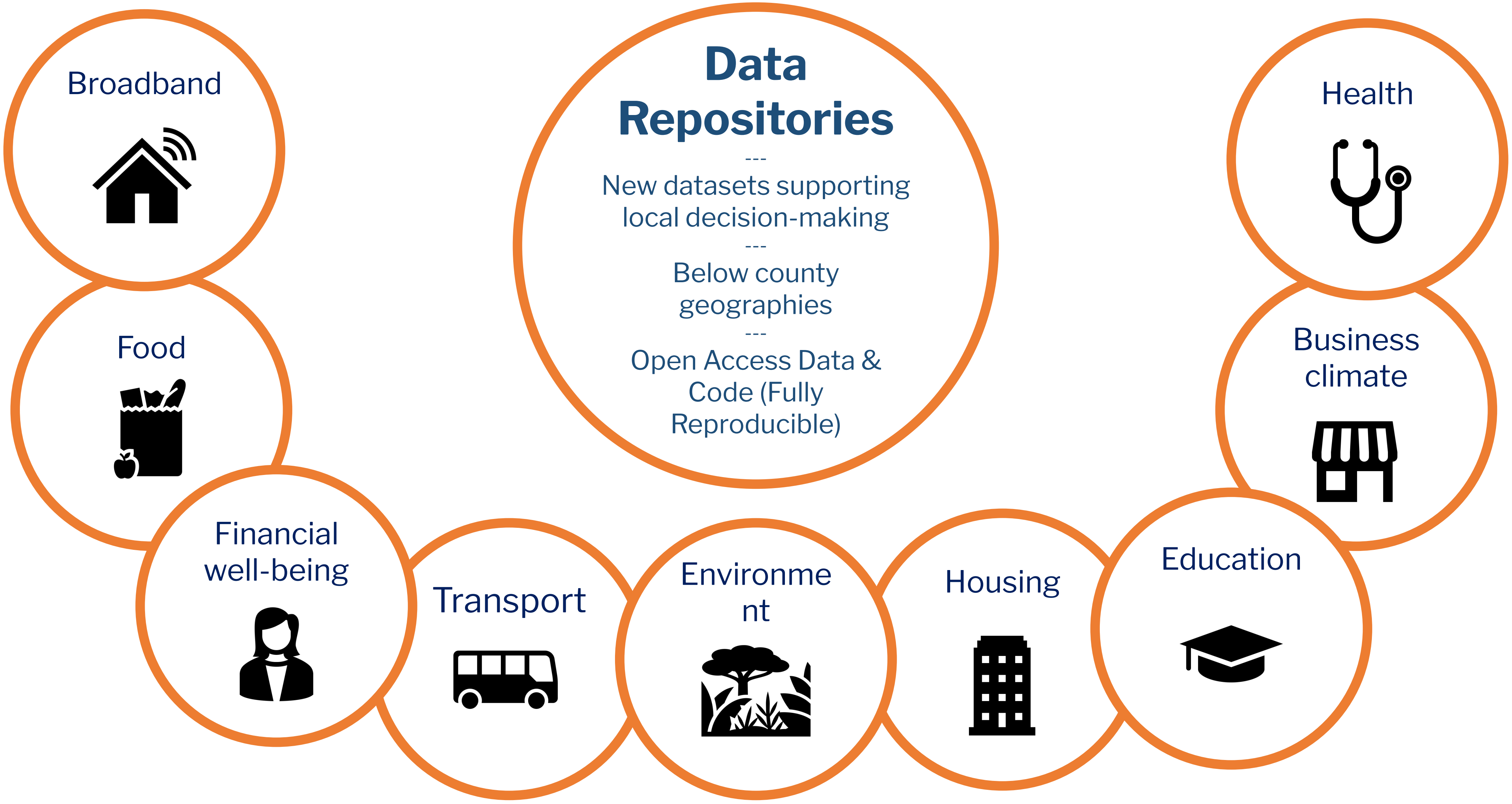
Provides a comprehensive, rigorous, and disciplined approach to problem solving that is:

- At the heart of the Community Learning through Data Driven Discovery (CLD3) process;
- Includes identifying data sources, preparing them for use, and then assessing the value of these sources for the intended use(s); and
- An iterative process.


Problem Identification: Questions & Working Hypotheses
Data Discovery [Inventory • Screening • Acquisition]
Data Ingestion & Governance
Data Wrangling
Profiling • Preparation • Linkage • Exploration
Fitness-for-Use
Statistical Modeling & Analyses
Communication & Dissemination
Ethics Review

Building Capacity for Data Driven Governance - Creating a New Foundation for Democracy Statistics and Public Policy, 4:1-11. (2017) S. A. Keller, V. Lancaster, S. Shipp

Project components



Data Repositories

New datasets supporting local decision-making

Below county geographies

Open Access Data & Code (Fully Reproducible)

Broadband

Food

Financial well-being

Transport

Environme nt

Housing

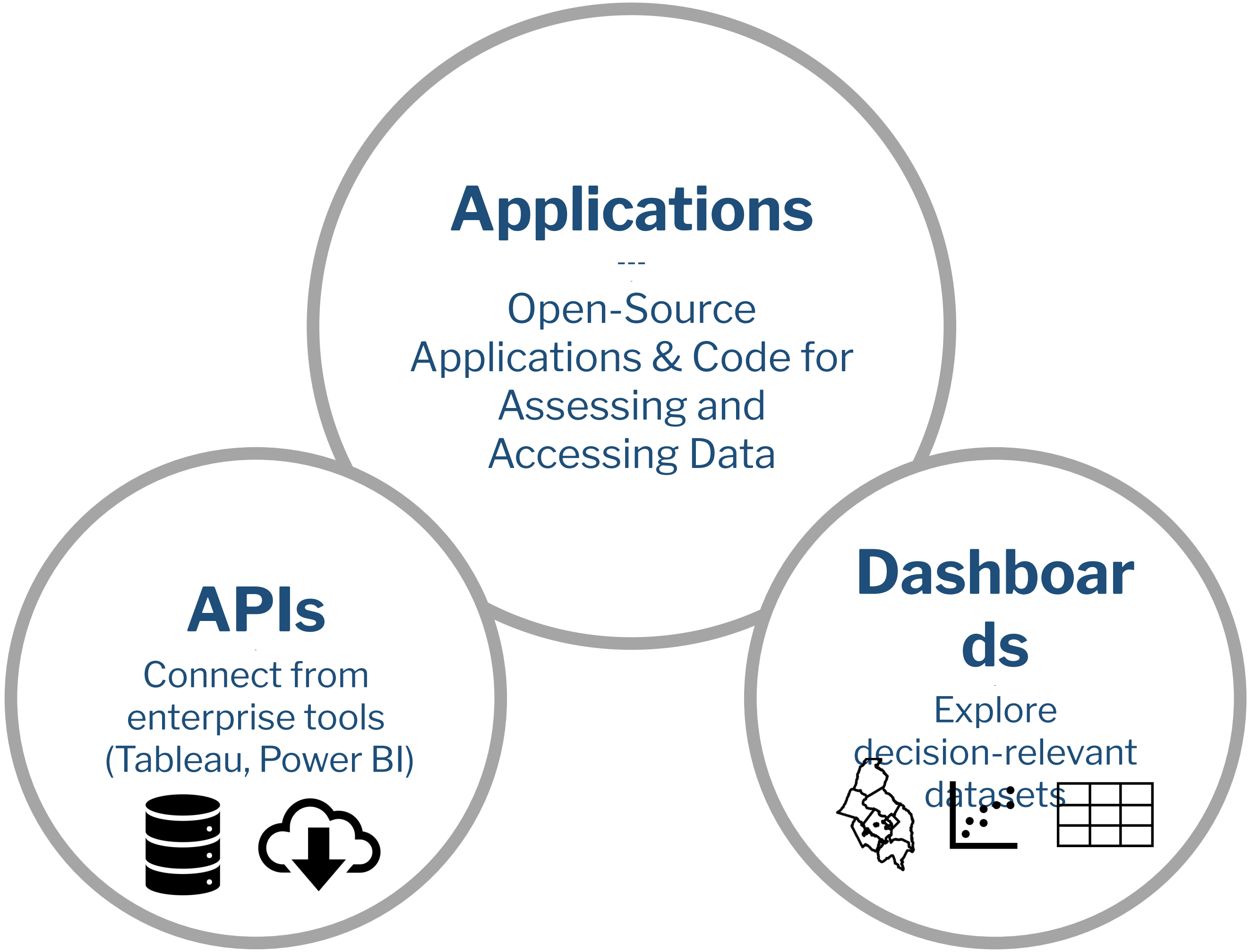
Education

Business climate

Health

SOCIAL IMPACT
DATA COMMONS

Project components



Applications

Open-Source Applications & Code for Assessing and Accessing Data

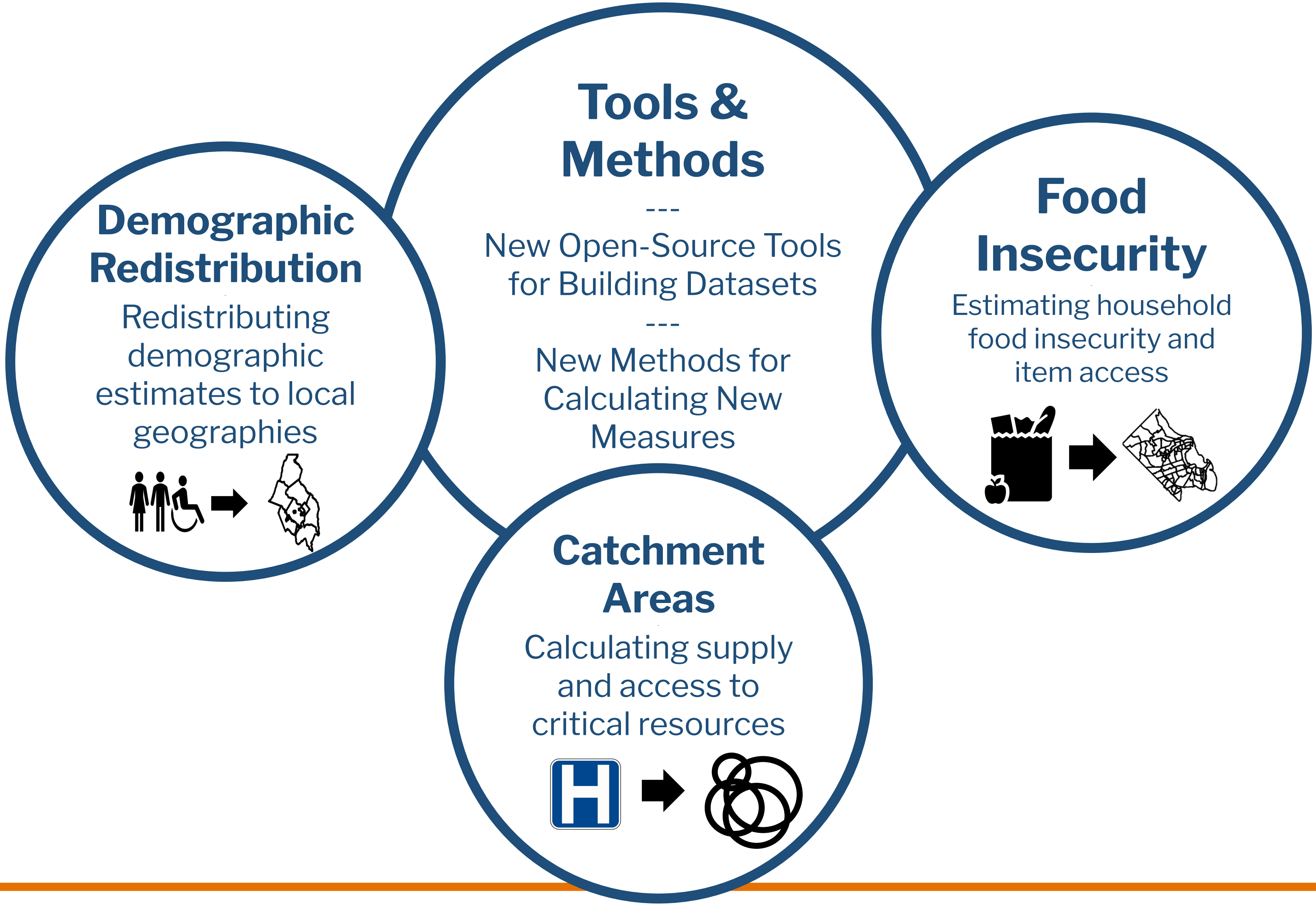
APIs

Connect from enterprise tools (Tableau, Power BI)

Dashboards

Explore decision-relevant datasets

Project components



Tools & Methods

New Open-Source Tools for Building Datasets

New Methods for Calculating New Measures

Demographic Redistribution

Redistributing demographic estimates to local geographies

Food Insecurity

Estimating household food insecurity and item access

Catchment Areas

Calculating supply and access to critical resources

Evaluating Core Metadata

FAIR Standards

- We defined three categories for progress:
  - Achieving: Our system meets all or nearly all of the guidance
  - Working Towards: Our system implements some of the guidance
  - Not Addressing: Our system meets none or very little of the guidance

- We find that we begun to address most of the principles with the implementation of our metadata systems.
  - Strongest in the principles of findability and accessibility (GitHub).
- Weakest in interoperability
  - Only 4/17 core metadata elements are derived from a widely-adopted metadata schema