SCHARE

Tutorial Think-a-Thons

National Institutes of Health

SCHARE

Health Outcome Research Paradigm Shift: Understanding How Big Data Expands Knowledge

April 16, 2025

Deborah Duran, PhD · NIMHD
Elif Dede Yildirim, PhD · NIMHD
Mark Aronson, PhD · NIMHD

Experience poll

Please check your level of experience with the following:

	None	Some	Proficient	Expert
Python	☐	☐	☐	☐
R	☐	☐	☐	☐
Cloud computing	☐	☐	☐	☐
Terra	☐	☐	☐	☐
Health disparities research	☐	☐	☐	☐
Health outcomes research	☐	☐	☐	☐

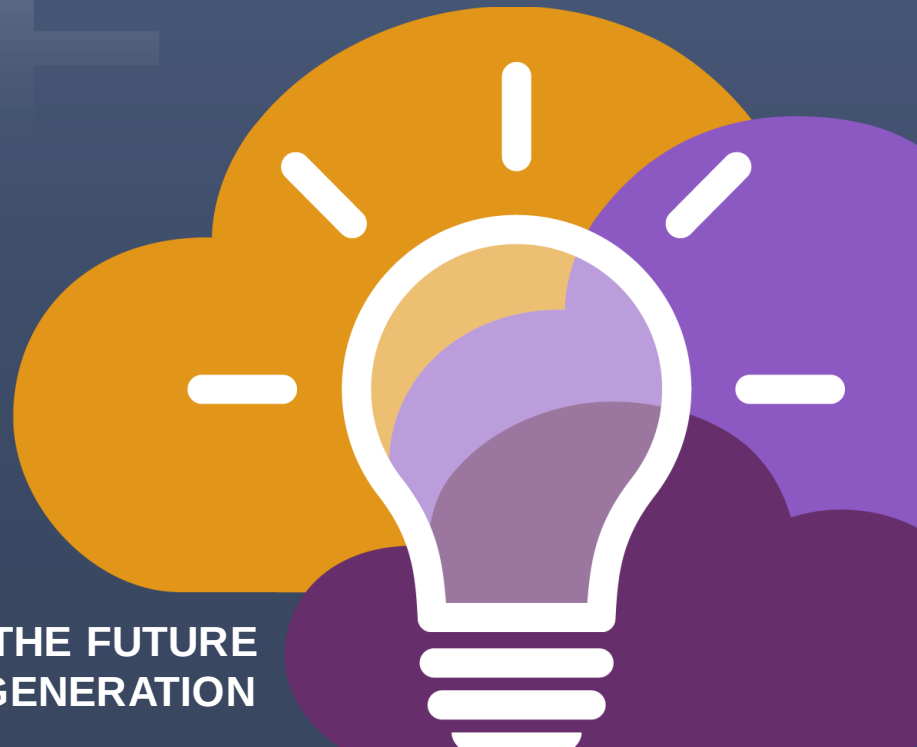
Interest poll

I am interested in (check all that apply):

- ☐ Learning about Health Disparities and Health Outcomes research to apply my data science skills
- ☐ Conducting my own research using AI/cloud computing and publishing papers
- ☐ Connecting with new collaborators to conduct research using AI/cloud computing and publish papers
- ☐ Learning to use AI tools and cloud computing to gain new skills for research using Big Data
- ☐ Learning cloud computing resources to implement my own cloud
- ☐ Developing ethical AI strategies
- ☐ Other

SCHARE

What is SCHARE?



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

SCHARE

Science
Collaborative for
Health disparities and
Artificial intelligence
Reduction of
Errors

Register: nimhd.nih.gov/schare

SCHARE is a cloud-based population science data platform designed to accelerate research population health research, including chronic diseases, health disparities & health outcomes by utilizing transparent artificial intelligence (AI) approaches with a focus on the reduction of errors in the use and reuse of models designed to accelerate innovative research that includes place-based factors and biologics for whole-person health discoveries.

SCHARE aims to fill five critical gaps:

- Leverage population science, place-based, and behavioral Big Data and cloud computing tools to foster a paradigm shift in health disparity and healthcare delivery outcomes research to generate innovative whole-person health discoveries
- Advance use of transparency and sophisticated inquiry to develop innovative strategies and differing perspectives to reproducibility and to reduce AI errors
- Upskill novice untrained users in data science through cloud computing skills training, cross-discipline mentoring, and multi-career level collaborating on research
- Provide a data science cloud computing resource and data center for community colleges, and low resource institutions and organizations
- Offer a project data repository centered on core common data elements for enhanced data interoperability and compliance with NIH Data Management and Sharing Policy

SCHARE

Google Platform
Terra Interface

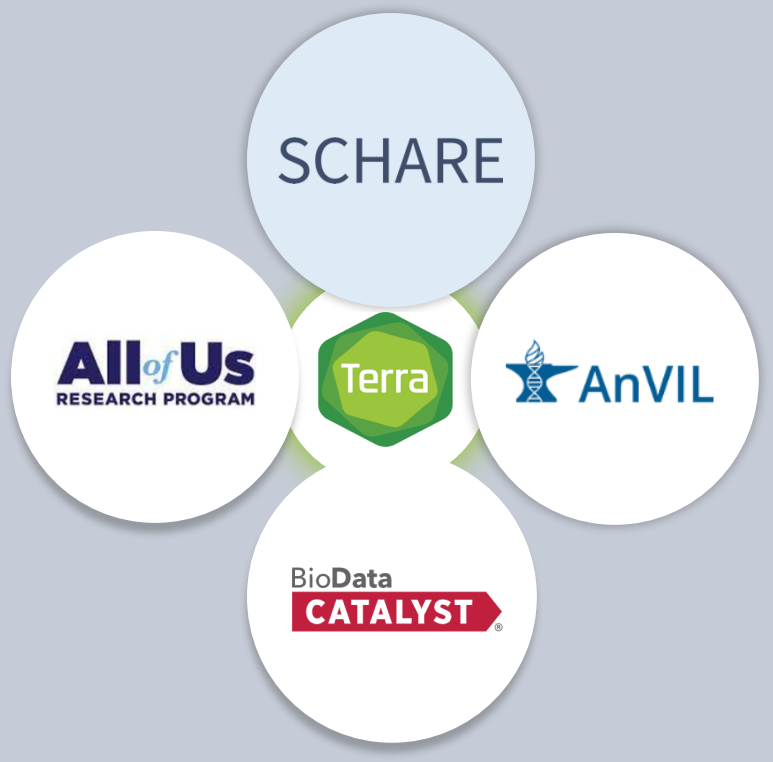
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Copy-and-paste code in Python and R
- Learning Terra on SCHARE prepares you to use other NIH platforms

Terra recommends using **Chrome**

Must have a **Gmail** friendly account

PREPARING FOR AI RESEARCH AND HEALTHCARE USING BIG DATA

Mapping across cloud platforms with Terra interface for collaborative research



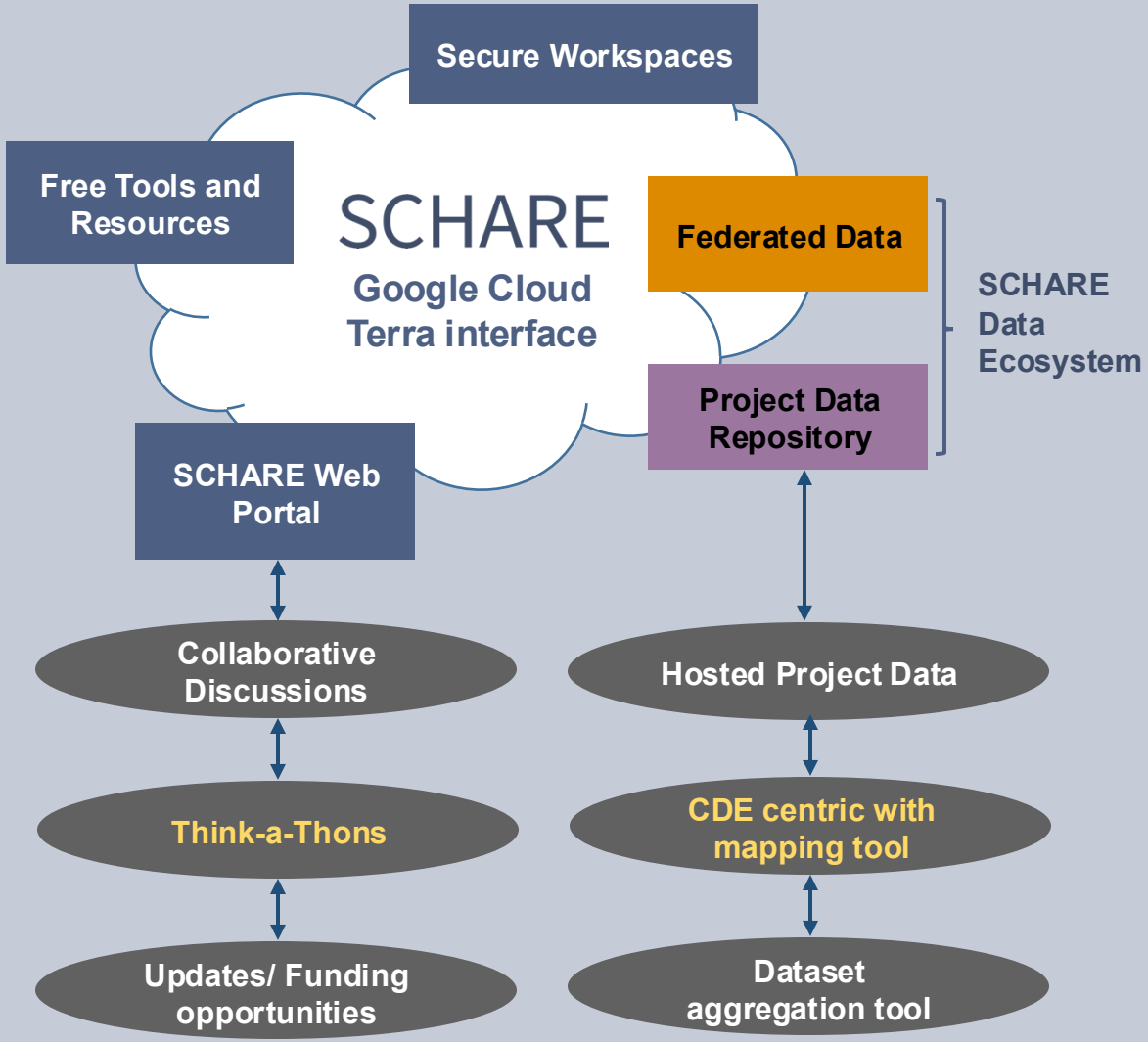
BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION

SCHARE Components

SCHARE co-localizes within the cloud:

1. **Datasets** relevant to health disparities, health care delivery, and health outcomes research, including social determinants of health and other social science behavioral data
2. **A project data repository** for NIH-funded projects centered on Core Common Data Elements for enhanced data interoperability and compliance with NIH Data Management and Sharing policy
3. **Secure, collaborative workspaces** and for researchers and relevant collaborators
4. **Computational capabilities** for collaboratively evaluating designing and assessing fit-for-purpose utilization of datasets and algorithms to generate AI models that are effective and efficient

Intramural and Extramural Resource



SCHARE

SCHARE Terra Platform

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

National/Federated Datasets

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

SCHARE Ecosystem structure

Researchers can access, link, analyze, and export **a wealth of SDoH and population science related datasets** within and across platforms relevant to research about health disparities, health care delivery, and health outcomes, including:

300+
FEDERATED PUBLIC DATASETS

Public datasets

Publicly accessible, federated, de-identified datasets hosted by SCHARE or hosted by Google through the Google Cloud Public Dataset Program

Examples: *Behavioral Risk Factor Surveillance System (BRFSS)*
American Community Survey (ACS)

CDE
FOCUSED REPOSITORY

Project datasets

Publicly accessible and controlled-access, funded program/project datasets using Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

Examples: *Forthcoming datasets such as the Jackson Heart Study (JHS)*

Innovative Approach:
CDE Concept Codes
Uniform Resource Identifier (**URI**)

SCHARE Ecosystem

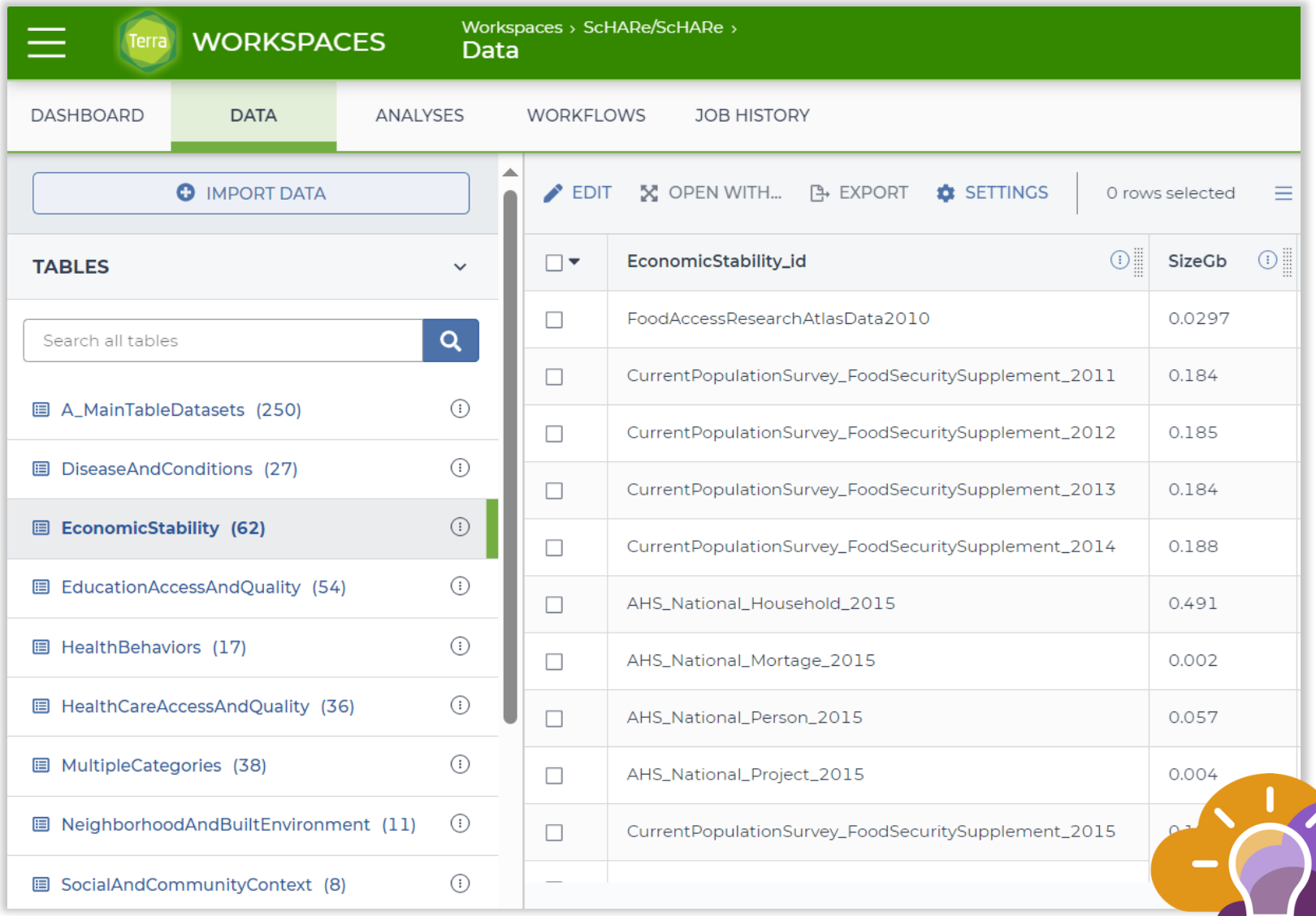
OVER 300 DATA SETS CENTRALIZED

Datasets are categorized by content based on the CDC **Social Determinants of Health categories**:

1. Economic Stability
2. Education Access and Quality
3. Health Care Access and Quality
4. Neighborhood and Built Environment
5. Social and Community Context

with the addition of:

- **Health Behaviors**
- **Diseases and Conditions**



SCHARE Ecosystem: Public datasets

Examples of interesting datasets include:

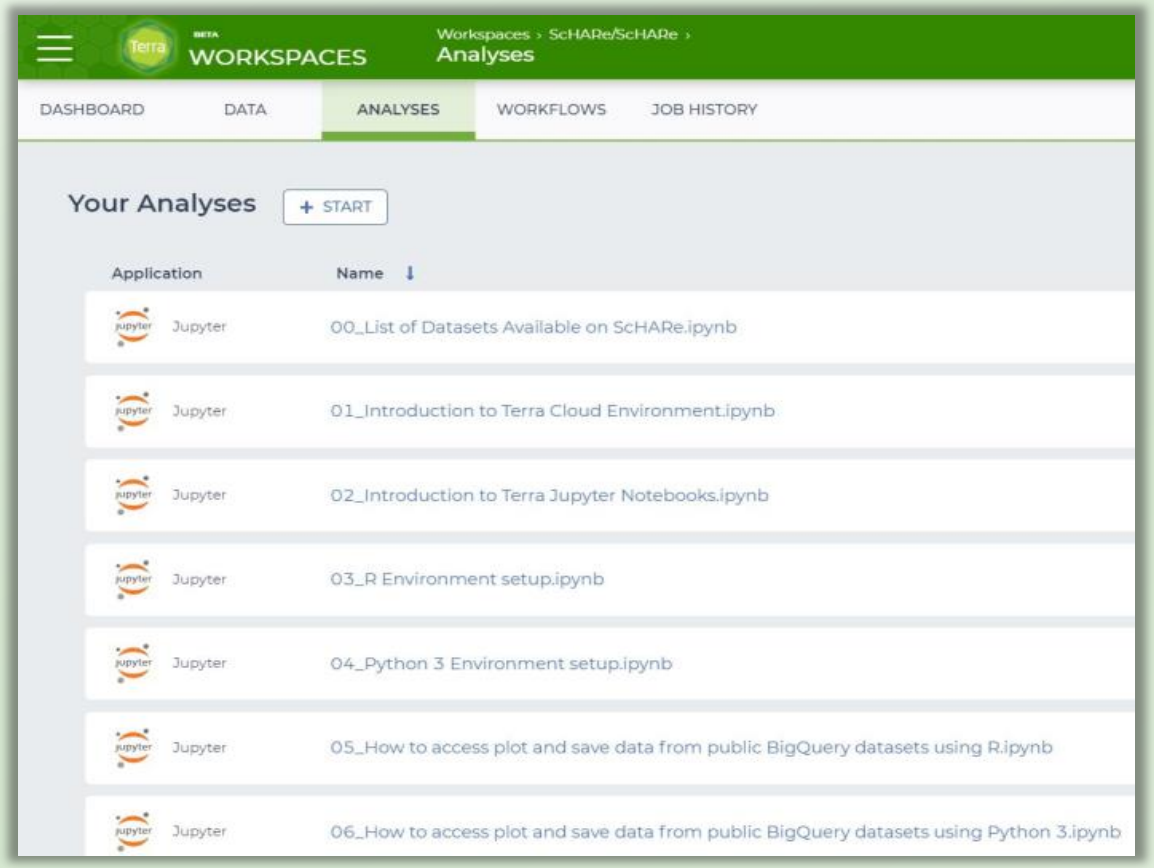
- **American Community Survey** (U.S. Census Bureau)
- **US Census Data** (U.S. Census Bureau)
- **Area Deprivation Index** (BroadStreet)
- **GDP and Income by County** (Bureau of Economic Analysis)
- **US Inflation and Unemployment** (U.S. Bureau of Labor Statistics)
- **U.S. Chronic Disease Indicators** (U.S. Census Bureau)
- **Point-in-Time Homelessness Count** (U.S. Dept. of Housing and Urban Development)
- **National Mental Health** (SAMHSA)
- **US Residential Real Estate Data** (House Canary)
- **Center for Medicare and Medicaid Services - Dual Enrollment** (U.S. Dept. of Health & Human Services)
- **National Mental Health** (SAMHSA)
- **Health Professional Shortage Areas** (U.S. Dept. of Health & Human Services)
- **CDC Births Data Summary** (Centers for Disease Control)
- **BRFSS** Behavioral Risk Factors
- **Community Resilience Estimates:** Community resilience estimates calculated by modeling individual and household characteristics
- **Adult Indicators for Oral Health** (NOHSS)
- **Alzheimer's Disease and Health Aging Data** (NIH)

Data Analytic and AI Tools

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

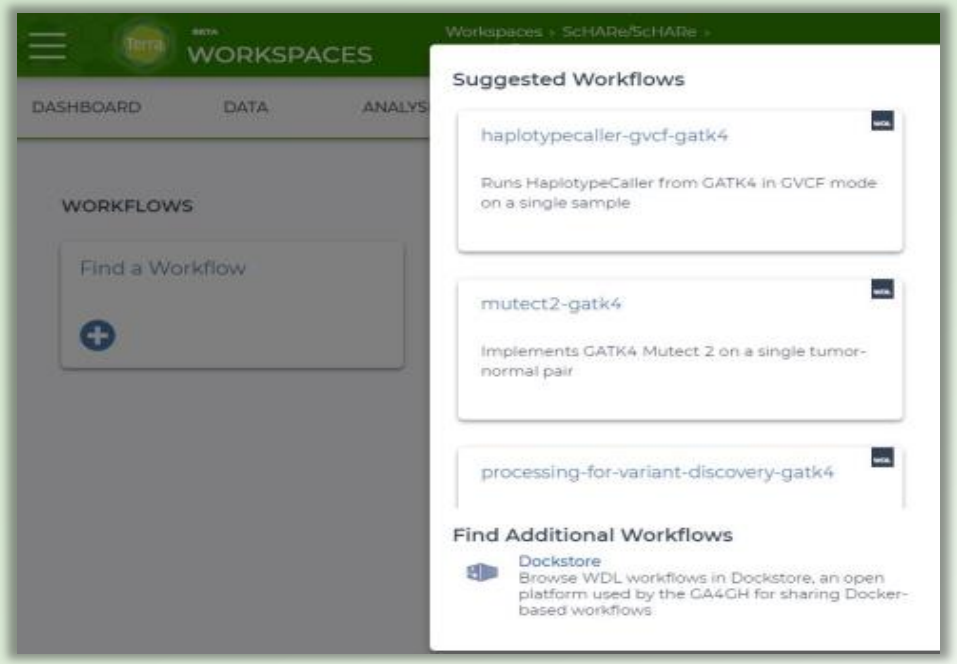
SCHARE Terra interface: Analyses (Notebooks)

Notebooks for analytics and tutorials



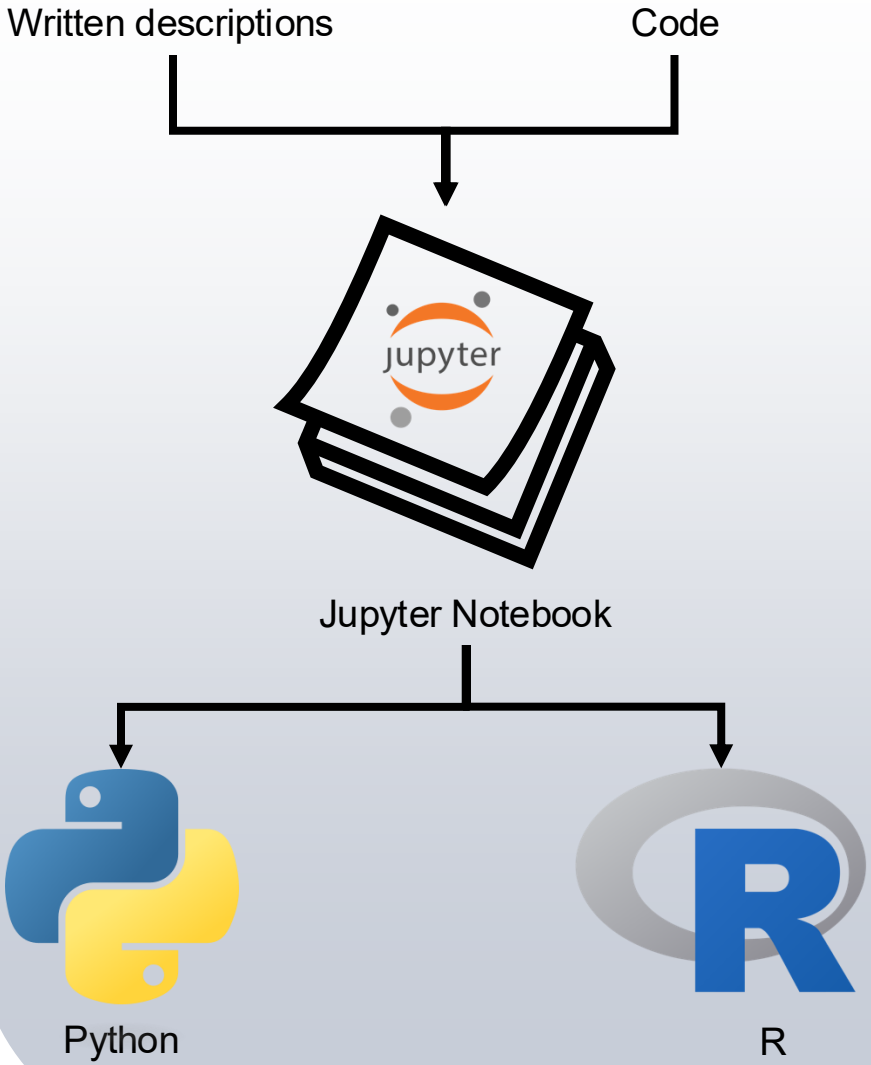
Modular codes

- Easy-to-use copy and paste analytics



- Modular codes developed for reuse

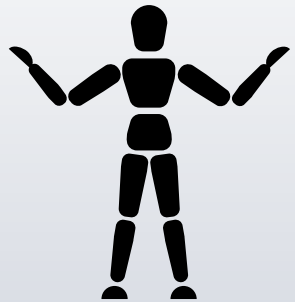
SCHARE Data Analytic Tools



On the SCHARE Workspace

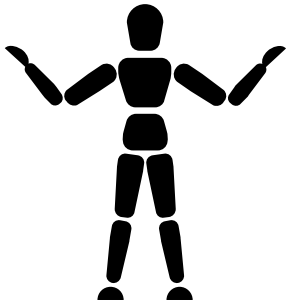
Reference Notebooks

Tutorial Notebooks



Model Notebooks

SCHARE Model Notebooks under Analysis Tab



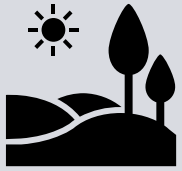
Python code model notebooks
(SCHARE Workspace -> Analyses -> Section B)

b. 00_Introduction to Python.ipynb

"Table of Contents“

Describes the purpose of all other notebooks in this section

b. 01_Python 3 Environment setup.ipynb



Describes what a python environment is and copy & paste code to set yours up

b. 02_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

Copy & paste code for accessing datasets hosted by Google BigQuery

b. 03_How to access plot and save data from ScHARe hosted datasets using Python 3.ipynb

Copy & paste code for accessing datasets hosted on the SCHARE workspace

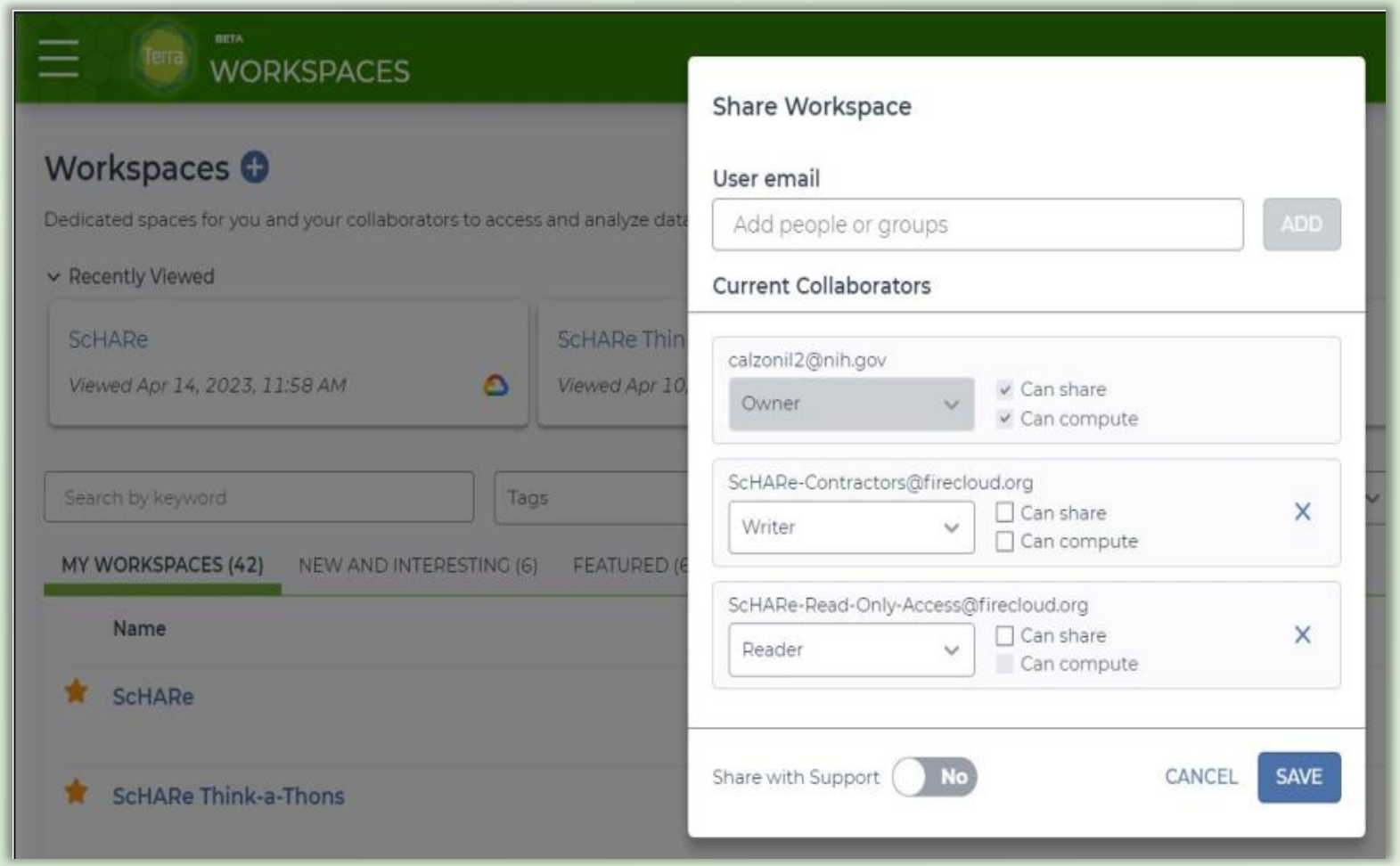
b. 04_How to upload access plot and save data stored locally using Python 3.ipynb

Copy & paste code for accessing data on your local computer

Secure Workspaces for Single and Collaborative Research

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

SCHARE Terra interface: secure workspace



- Secure workspace **for self or collaborative research**
- **Assign roles**: owner, writer, reader
- **Host own data and code**

SCHARE

Data Repository

NIH Data Management and Sharing Policy

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

PUBLICLY AVAILABLE SPRING 2025

SCHARE Data Repository

The Four Data Lifecycle Stages

The SDR is here to support your research and your data throughout all stages of the data lifecycle. Our touchpoints can be contextualized by thinking about your data in these four stages.

Dawn

Dataset Creation

Researchers can choose to store their data themselves, uploading it upon study completion, or use the SDR as a storage interface.

Midday

Dataset Submission

Researchers submit their data for public sharing on the SDR, creating a controlled-access version if the dataset contains sensitive information.

Golden Hour

Dataset Access

Researchers use the public version of datasets on the SDR, or request access to controlled-access data, for secondary studies.

Sunset

Dataset Archival

Once the dataset meets the archival requirements, the dataset collection is removed from the SDR, and the underlying data is archived.

Key Features of the SCHARE Data Repository

Upload your own data

Store collected data and annotate with a data dictionary. Align data to the ScHARe CDEs.

Harmonize data to CDEs

Map uploaded data to CDEs. Join your data with project or federated data via CDEs.

Browse for data

Find relevant federated national datasets or other project data.

Manipulate and aggregate data for analysis

Filter, sort, and select subsets for specific purposes. Link and aggregate datasets.

Control privacy levels and data sharing

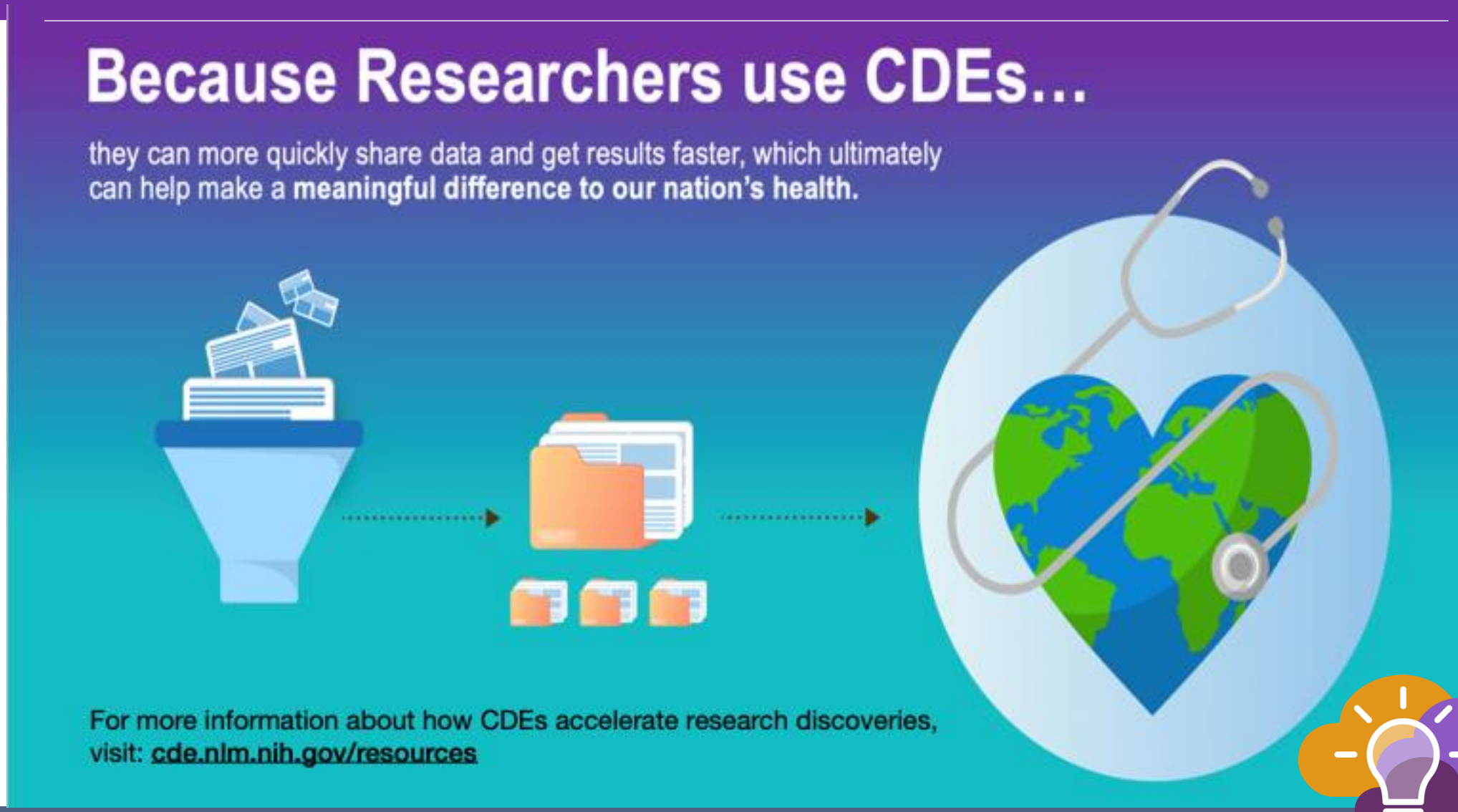
Share confidential data among colleagues. Share public access data with the research community.

Data Analysis via SCHARE Terra or local analysis platform

CDE benefits:

- **Faster start-up for project**
- **Better data aggregation across projects**
- **Shared meaning**
- **Concept-focused to allow questions/answers variations**
- **Coding enables an URI approach for better data interoperability**

A **Common Data Element** (CDE) is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection



SCHARE CDEs Human & Machine Readable

Semantically Defined: (Human Readable)

Each are semantically defined by a standardized coding system for shared meaning

Use of international/national coding systems – LOINC, UMLS, SemNet, FHIR, NCIt

Alcohol: disinfecting or drinking?

Colon: sentence punctuation or biological organ?

Mole: animal, blemish, unit of measure, or spy?

&

Coded (Machine Readable) : Use URI approach of associated codes that can be mapped across coding systems to create data interoperability

Pipes to separate data points (i.e. flower| plant| succulent| grass| tree|)

Human & Machine Readable

SCHARE Core CDEs

PhenX Toolkit

NIH Endorsed

- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex at Birth
- Marital Status
- Education
- Annual Household Income
- Household Size

- English Proficiency
- Disabilities
- Health Insurance
- Employment Status
- Usual Place of Health Care
- Financial Security / Social Needs
- Self-Reported Health
- Health Conditions (and Associated Medications/Treatments)

- **NIMHD Framework***
- **Health Disparity Outcomes***

* Project Level CDEs

Available in Spanish

SCHARE has developed **Common Data Elements** to ensure consistent data collection across studies, facilitate interoperability, and link data from different sources

NIH CDE Repository:
cde.nlm.nih.gov/home
PhenX Toolkit:
www.nimhd.nih.gov/resources/phenx/

SCHARE SDR Collections & Associations

Collections

Each project establishes its COLLECTION:

- Own data (ongoing or final)
- Single or collaborative
- Data Documentation
- Privacy controls
- CDE mapping
- Metadata

Data Submission can be ongoing or at end of project.

- **Can provide resource as a data center (ongoing)**
- **Fulfills Data Management and Sharing Policy (final)**

Associations

- ASSOCIATIONS comprised of multiple COLLECTIONs:
  - Creates parent collection
  - Own data (ongoing or final)
  - Single or collaborative
  - Data Documentation
  - Privacy controls
  - CDE mapping
  - Metadata
- Adds Collections to the Association

PUBLICLY AVAILABLE SPRING 2025

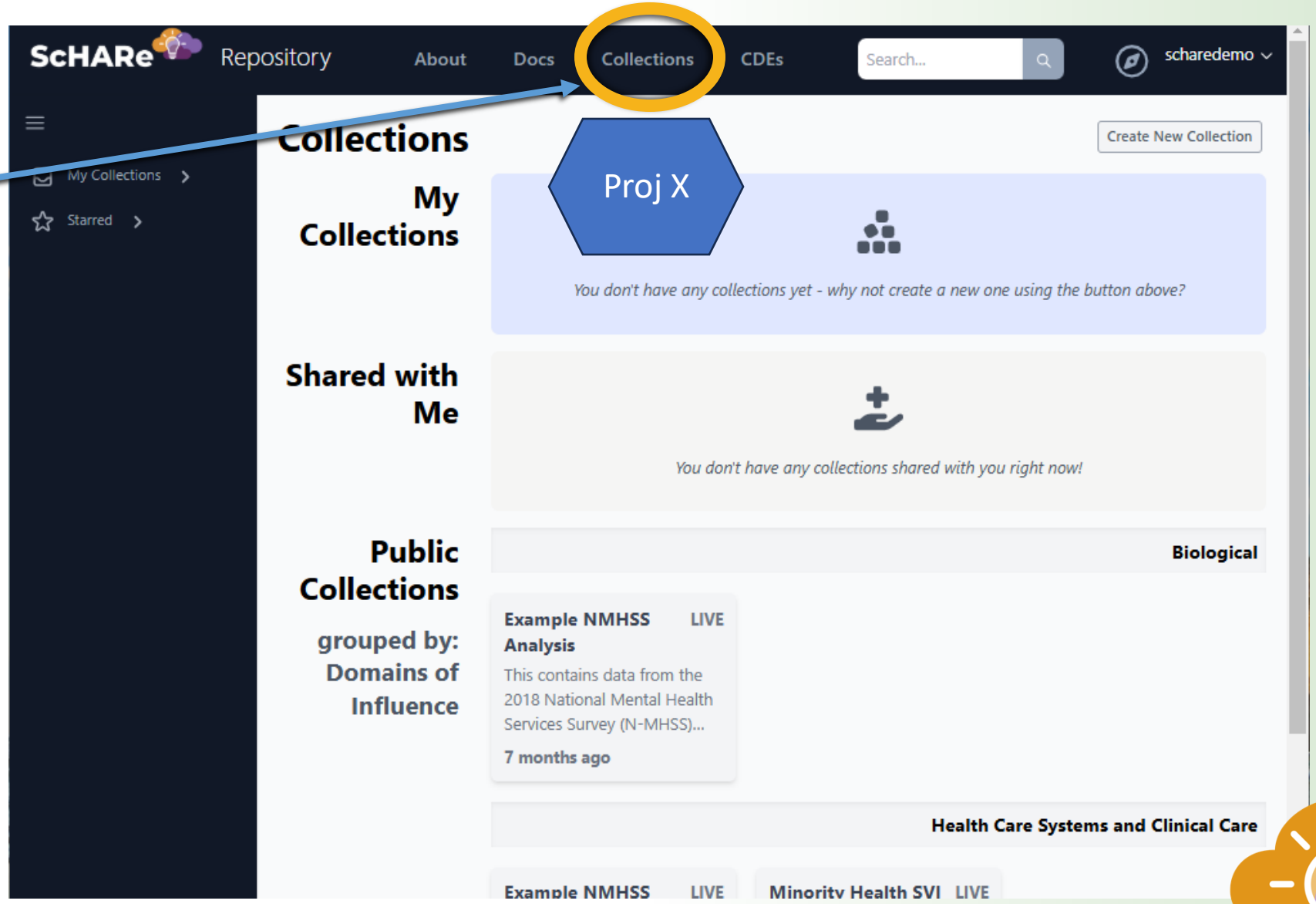
SCHARE Data Repository

Project Data

Collections are a place where you can describe and store your data and any related metadata and federated data.

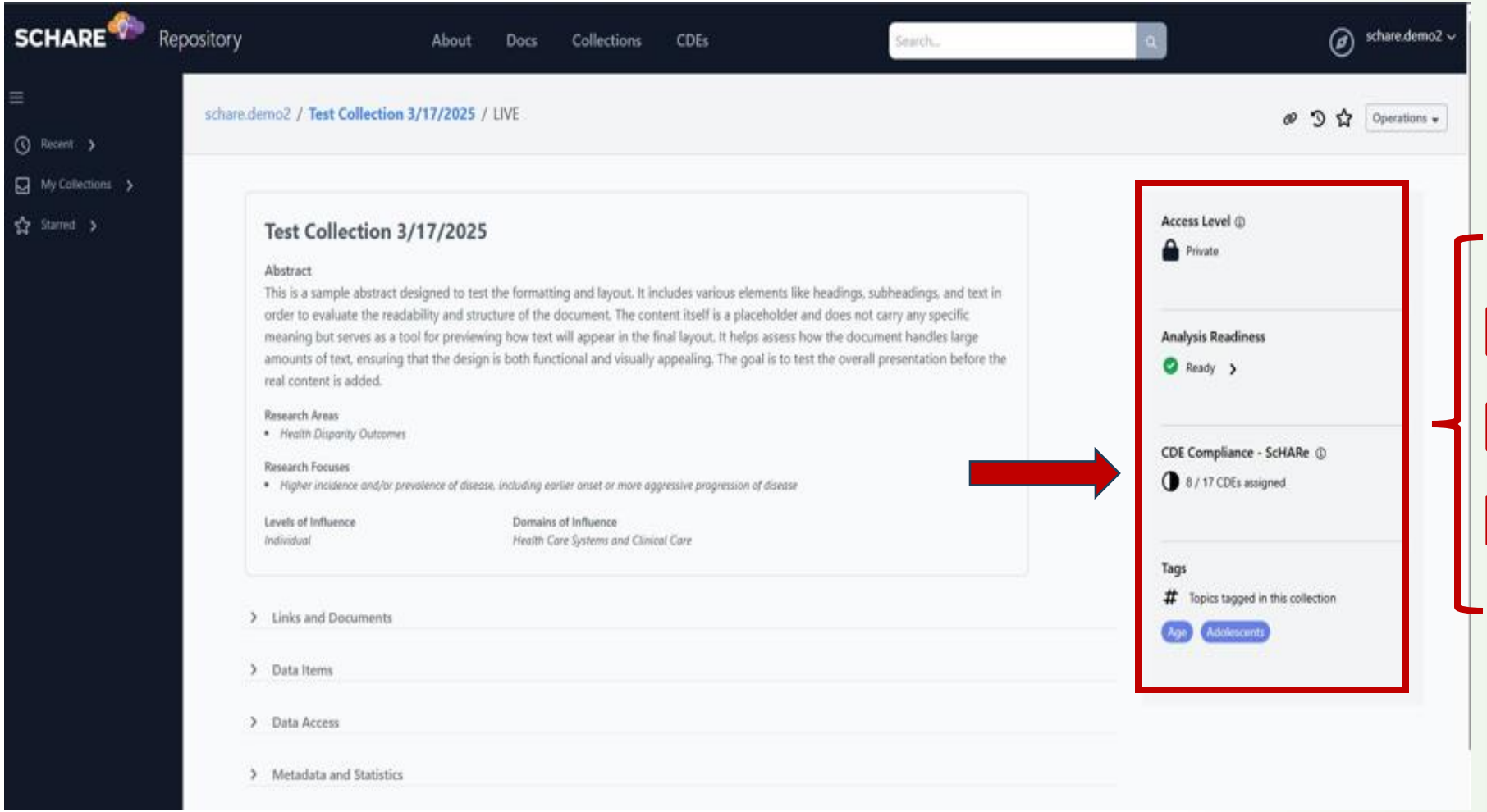
Can be shared with colleagues

Privacy controls & published when you're ready.



SCHARE Data Repository

PUBLICLY AVAILABLE SPRING 2025



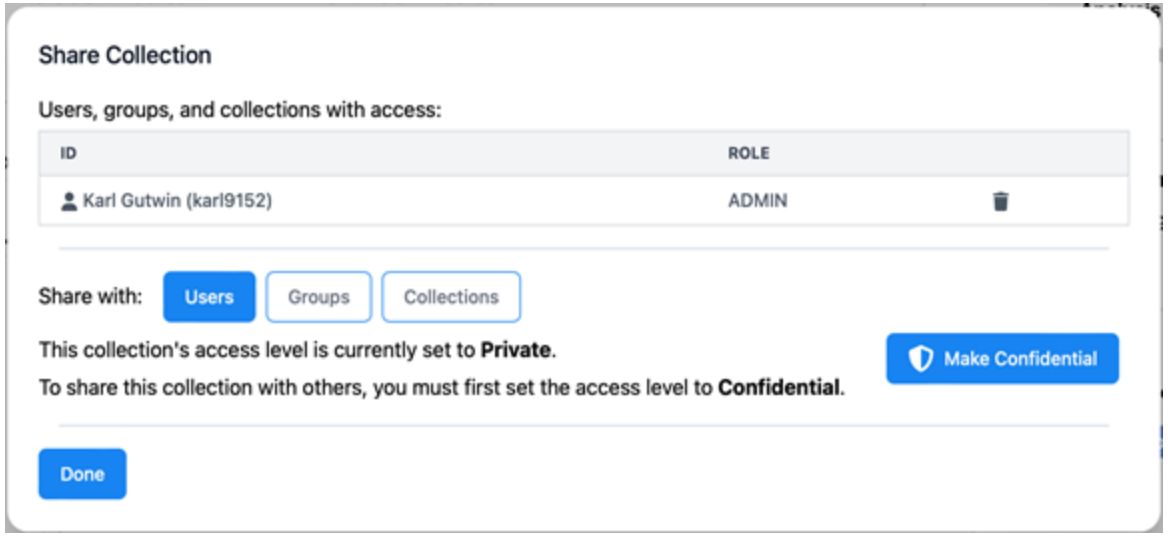
Data Access & Data Interoperability

By default, all collections start out as **Private**.

SCHARE Data Repository

Access Levels and Sharing Data

The access level of a collection defines the maximum permissions that can be used to share it with others. You have control over how your data is shared on the ScHARe Data Repository.



By default, all collections start out as **Private**.

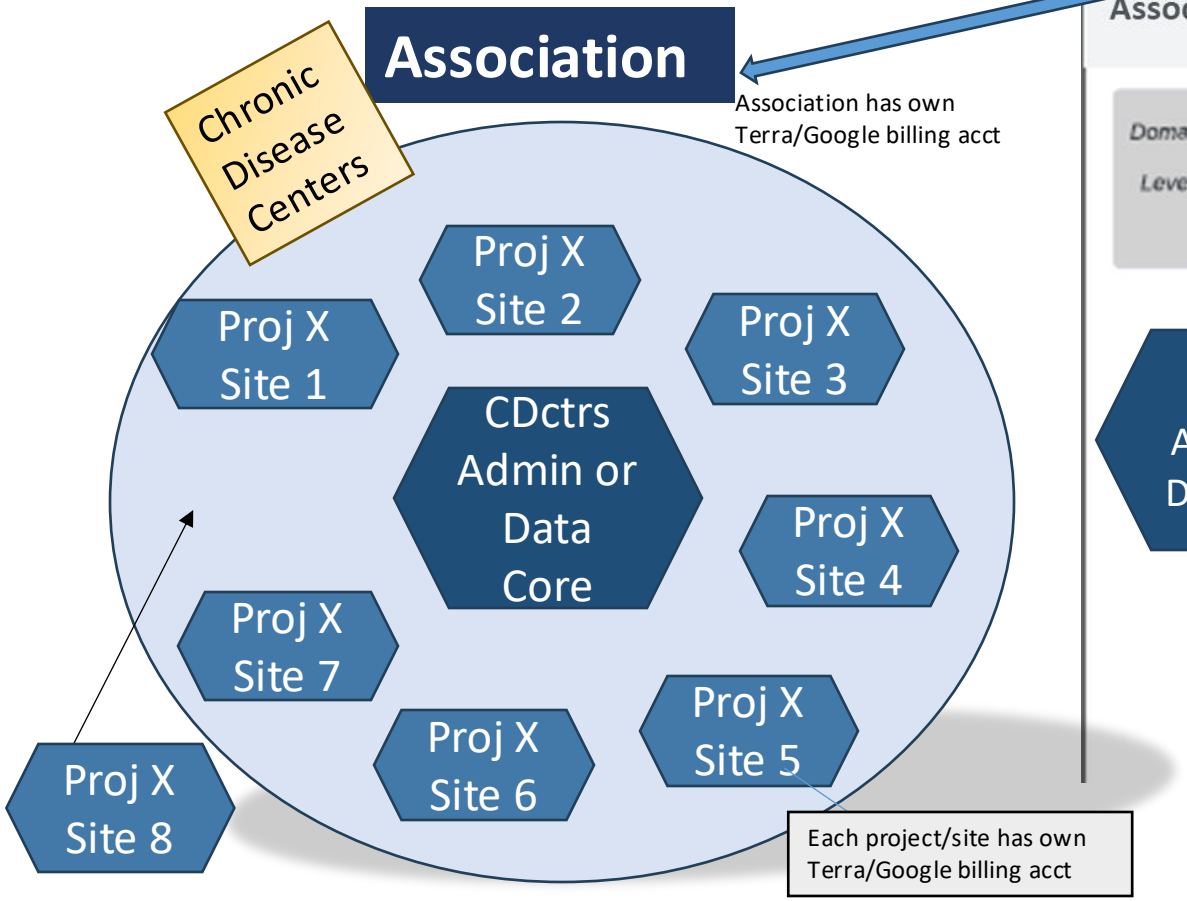
- **Private**: Only the collection's owner can access
- **Confidential**: The collection can be shared with named users
- **Controlled**: The collection can be shared with members of a controlled access group, as well as named users
- **Public**: The collection can be read by any user, including those not logged in; it can also be shared with named users

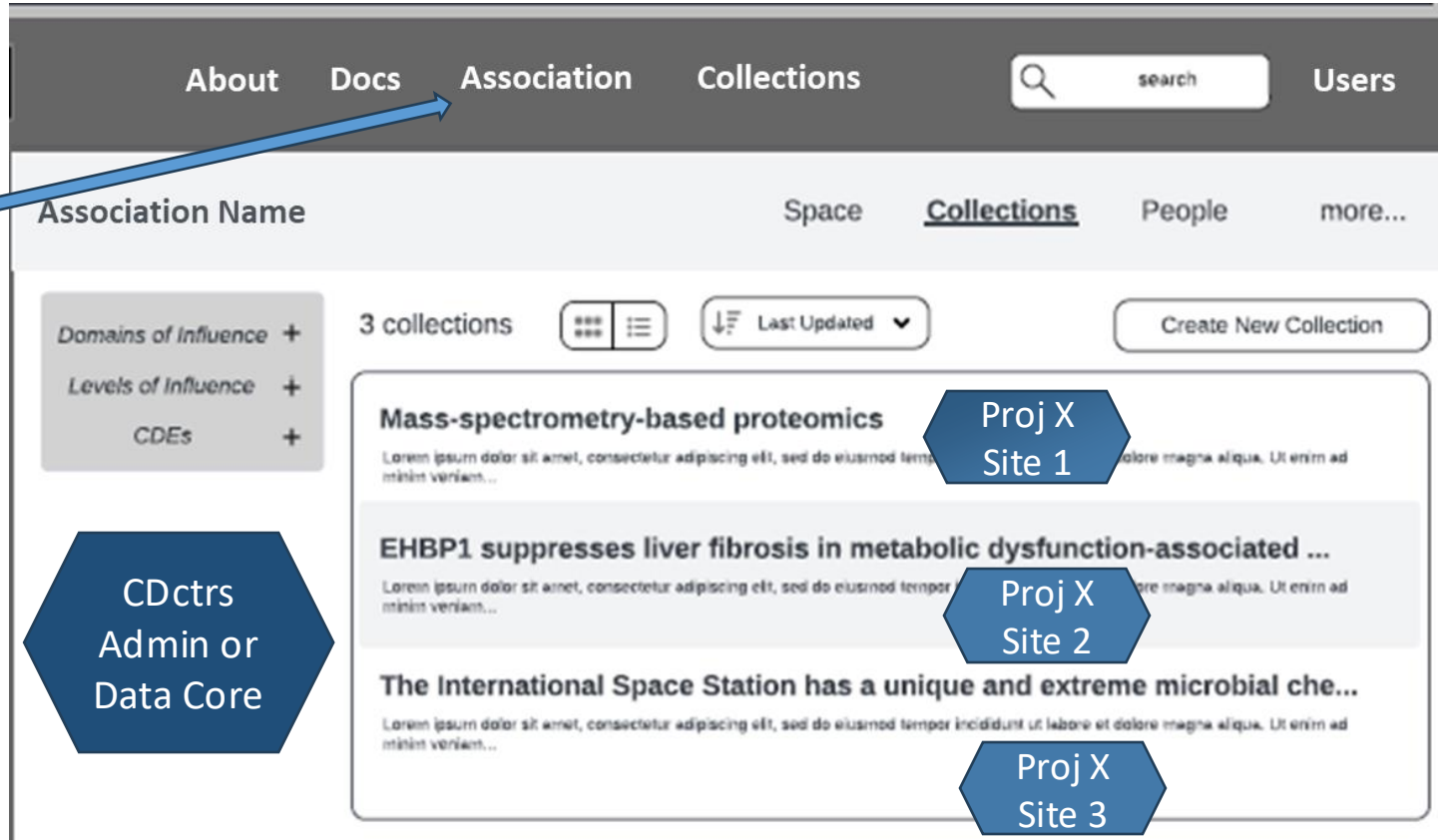
PUBLICLY AVAILABLE SPRING 2025

SCHARE Data Repository

Multi-Site Data

Project/Collections are integrated into the Association



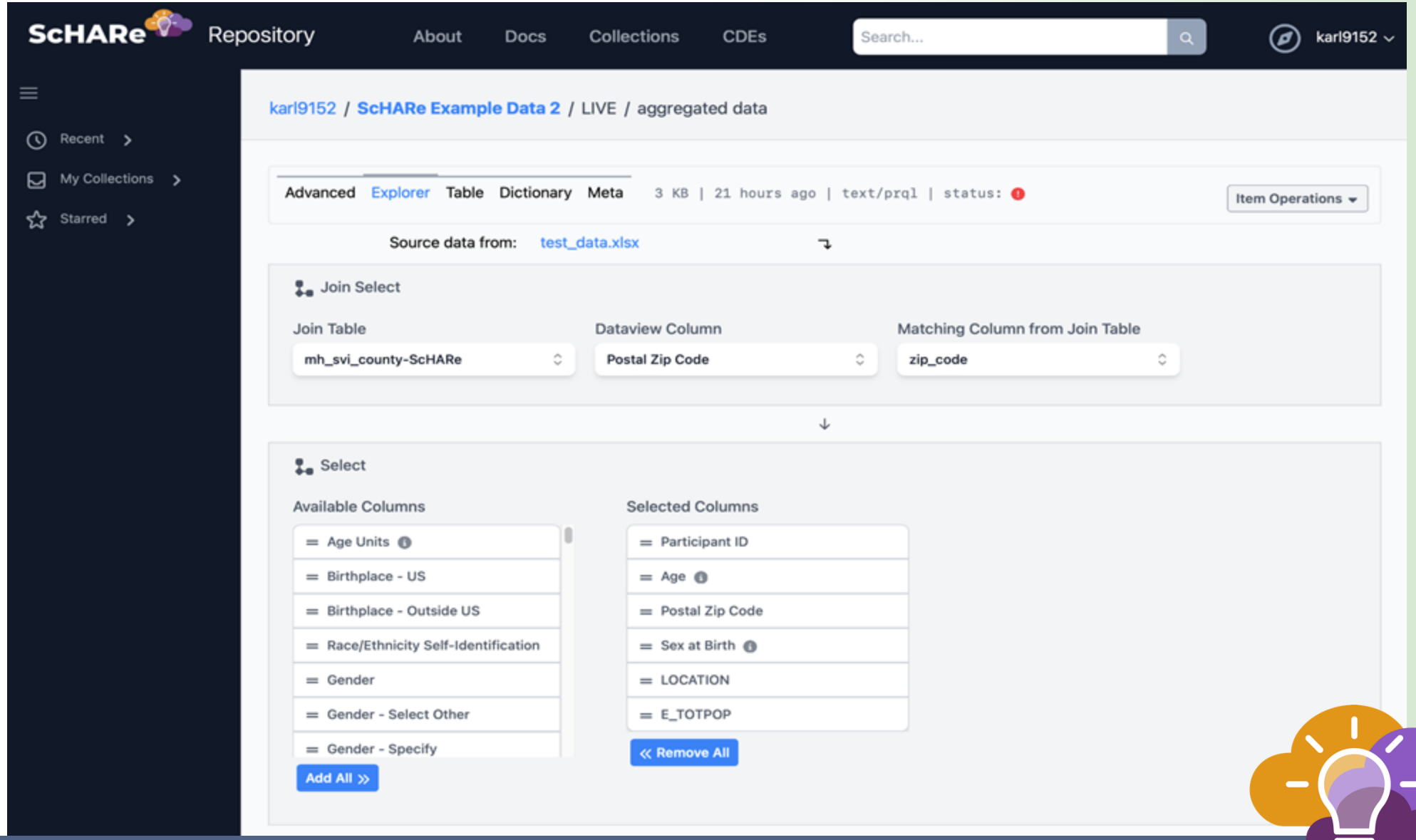


Each Association and Collection has its own space and metadata

SCHARE Data Repository

PUBLICLY AVAILABLE SPRING 2025

Data Aggregation Tool



SCHARE AI Tools

Coming soon!

Authenticate

Gemini Assistant

Use **Gemini Assistant** to launch a simple Q&A chat window to get assistance with writing your data analysis code. The chat interface is powered by the Gemini model and is designed to answer questions related to assisting novice coders with writing analysis code. Type your question in the box and click the **Generate** button to call the model and generate an output.

Note: while the data you send through this tool and data sent back are protected under Terra's Enterprise Google Cloud permissions, and are not reused by Google for future model training, we advise not sending any sensitive information (e.g. PII or PHI) through the model. Sticking to general questions or inserting dummy variable names to your questions are good practices to ensure the privacy of your data.

Select Model Gemini 1.5 Flash

Select Location us-central1

Question: Type your coding question here...

Generate

SCHARE AI Tools

User friendly metadata, data and model documentation tools

```
import pyschare as sc
labels= sc.data_labels
```

Enter the project title and a brief description or abstract in the provided text boxes. Once done, press the 'Save' button to generate the dataset facts summary.

Project Title:

Project Description:

Save

Fill in the metadata fields with the relevant information about your dataset, such as filename, format, URL, and domain. After completing the fields, click the "Save" button to save and display the metadata table.

Filename:

Format:

URL:

Domain:

Keywords:

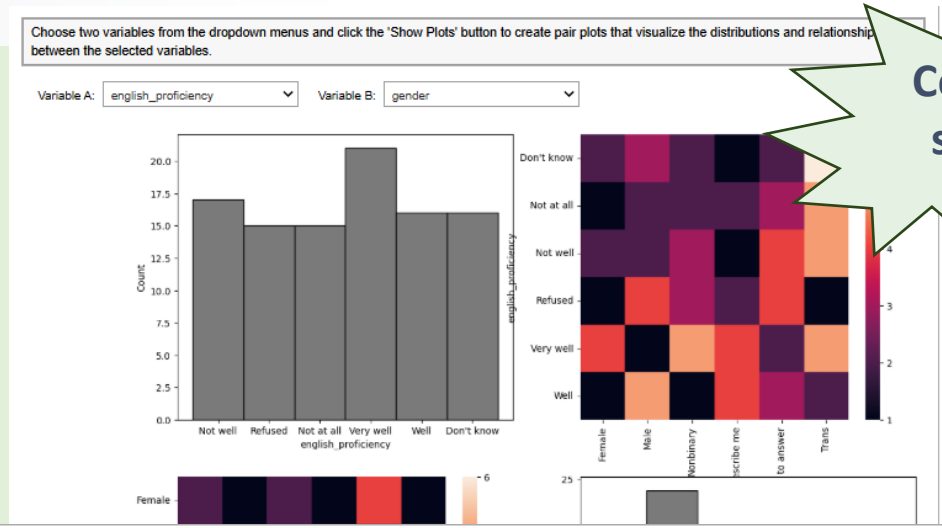
Type:

Geography:

Data Collection Method:

Time Method:

Rows:



Coming soon!

Upload your data dictionary file (CSV format) and select the appropriate columns for variable names and descriptions. The variables table will be generated based on your selections. Click the "Upload Data Dictionary" button to get started.

Upload Data Dictionary (0)

Variable Names:

Variable Descriptions:

Upload your dataset (CSV format), and click the 'Show Data' button to view the dataset. Using the dataset's variable names, input the names of ordinal, nominal, continuous, and discrete variables, separated by commas (e.g., Variable_a, Variable_b, Variable_c). After entering the variables, press the 'Show Statistics Table' button to generate and view the statistical summaries.

Upload Data (1)

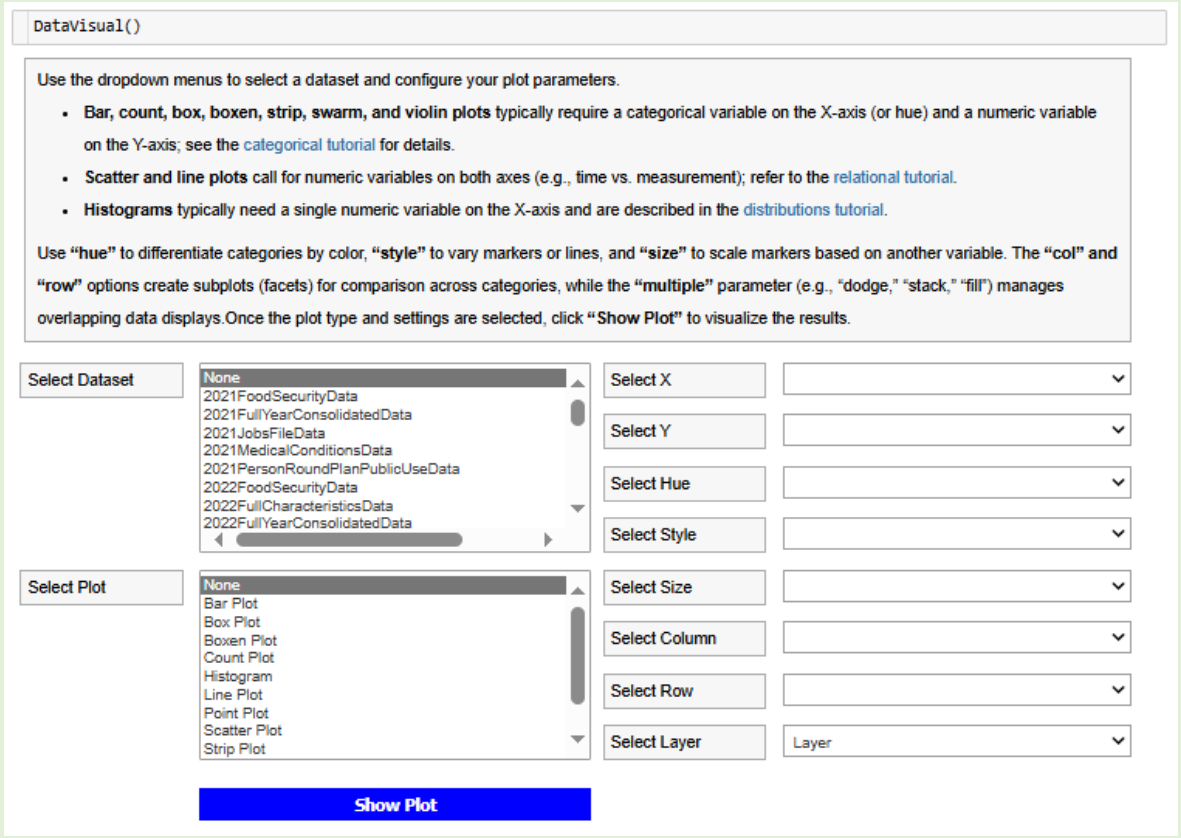
Show Data

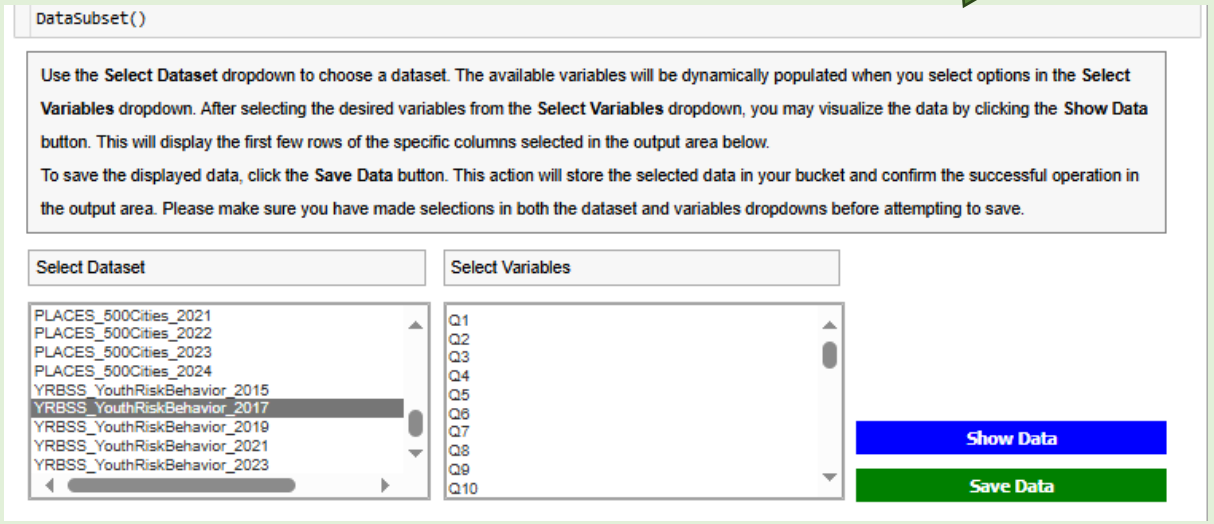
	Column1	entity:A_MainTableDatasets_id	Categories	Data	DataDictionary	FileFormat	Homepage	SizeGb	
0	0	2021FoodSecurityData	Health Care Access and Quality	gs://fc-secure-d6e25d73-4b50-4dbc-ac10-ec68998...	gs://fc-secure-d6e25d73-4b50-4dbc-ac10-ec68998...	XLSX	https://www.meps.ahrq.gov/mepsweb/data_stats/d...	0.806	Exp
1	1	2021FullYearConsolidatedData	Health Care Access and Quality	gs://fc-secure-d6e25d73-4b50-4dbc-ac10-ec68998...	gs://fc-secure-d6e25d73-4b50-4dbc-ac10-ec68998...	XLSX	https://www.meps.ahrq.gov/mepsweb/data_stats/d...	0.118	Exp

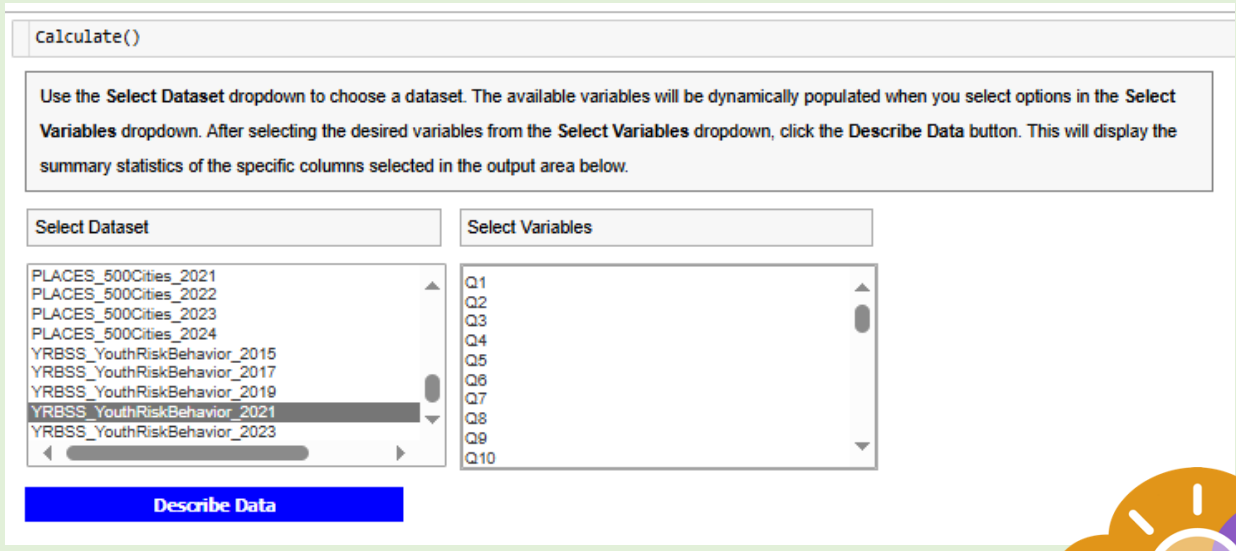
SCHARE PySCHARE Python Package

Coming soon!

PySCHARE package to search datasets and variables, subset, save, and visualize datasets



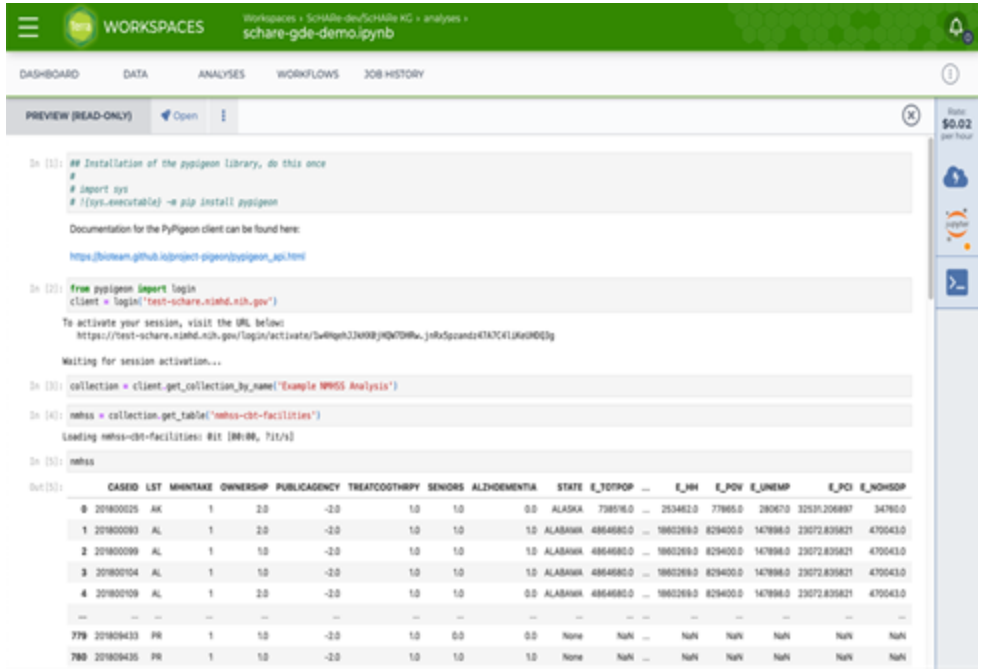




SCHARE Data Repository

PUBLICLY AVAILABLE SPRING 2025

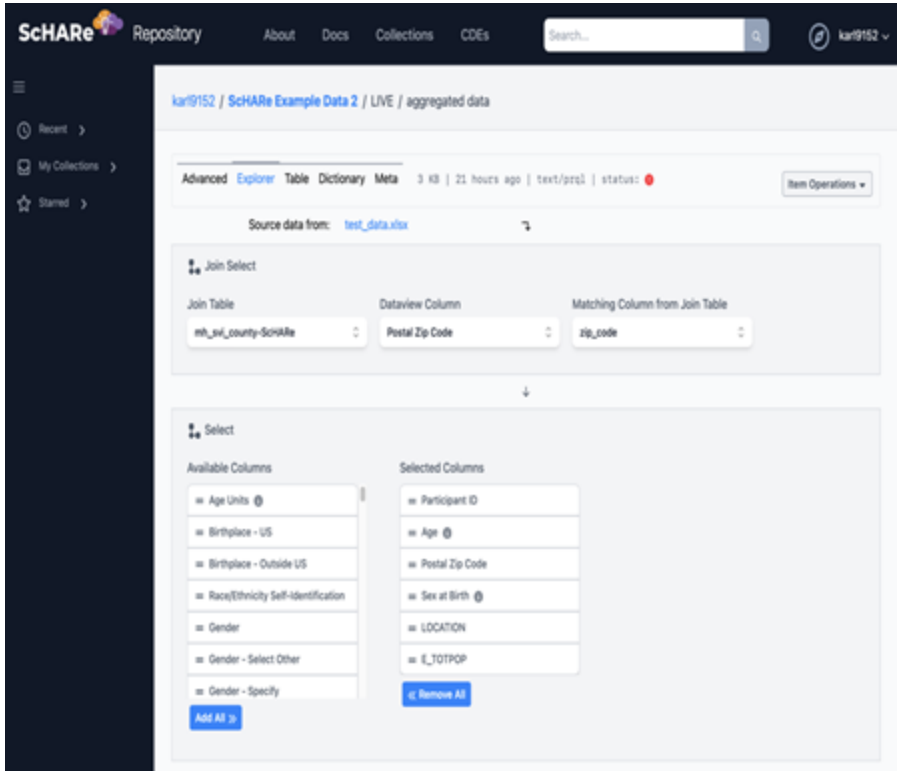
Connect to a Terra Jupyter notebook



For Analytics:

Download to Terra Account or your own computer

Use Dataviews on the Repository



How do I analyze or aggregate data from the ScHARe Data Repository?