National Institutes of Health

ScHARe

Science Collaborative for Health disparities and Artificial intelligence bias REduction

NIH Endorsed Common Data Elements

Think-a-Thons

National Institute on Minority Health and Health Disparities + Office of Data Science Strategy + National Institute of Nursing Research + National Institute of Dental and Craniofacial Research + National Institutes of Health Office of Research on Women's Health + Office of Minority Health

NIMHD

Dr. Eliseo Perez-Stable

ODSS

Dr. Susan Gregurick

NIH/OD

Dr. Larry Tabak

NINR

Dr. Shannon Zenk

NINR

Rebecca Hawes
Micheal Steele
John Grason

NIDCR

ORWH

OMH

NIMHD OCPL

Kelli Carrington
Thoko Kachipande
Corinne Baker

BioTeam

STRIDES

Terra

SIDEM

RLA

Broad Institute

CDE Working Group

Deborah Duran
Luca Calzoni
Rebecca Hawes
Micheal Steele
Kelvin Choi
Paula Strassle
Deborah Linares
Crystal Barksdale
Gneisha Dinwiddie
Jennifer Alvidrez
Matthew McAuliffe
Carolina Mendoza-Puccini
Simrann Sidhu
Tu Le

What is ScHARe?

ScHARe

ScHARe is a cloud-based population science data platform designed to accelerate research in health disparities, health and healthcare delivery outcomes, and artificial intelligence (AI) bias mitigation strategies

ScHARe aims to fill four critical gaps:

- Increase participation of **women & underrepresented populations with health disparities** in data science through data science skills training, cross-discipline mentoring, and multi-career level collaborating on research
- Leverage population science, SDoH, and behavioral Big Data and cloud computing tools to foster a **paradigm shift** in healthy disparity, and health and healthcare delivery outcomes research
- **Advance AI bias mitigation and ethical inquiry** by developing innovative strategies and securing diverse perspectives
- Provide a **data science cloud computing resource** for community colleges and low resource minority serving institutions and organizations

nimhd.nih.gov/schare

ScHARe Data Ecosystem

Researchers can access, link, analyze, and export **a wealth of datasets** within and across platforms relevant to research about health disparities, health care outcomes and bias mitigation, including:

- **Google Cloud Public Datasets:** publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program

  Example: *American Community Survey (ACS)*

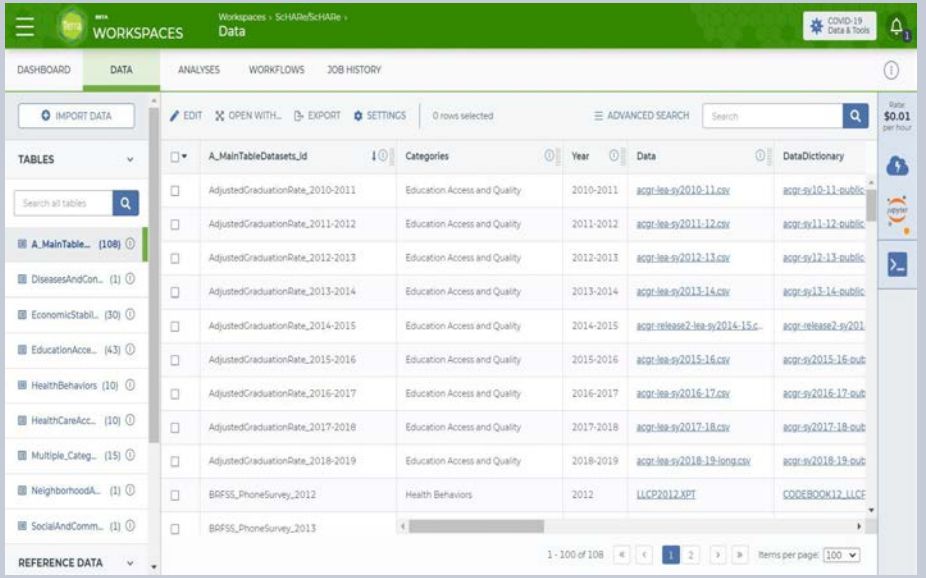
- **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe

  Example: *Behavioral Risk Factor Surveillance System (BRFSS)*

- **Funded Datasets on ScHARe:** publicly accessible and controlled-access, funded program/project datasets using Core Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

  Examples: *Jackson Heart Study (JHS); Extramural Grant Data; Intramural Project Data*

OVER 240 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**

Users will be able to **map and link** across datasets

ScHARe Ecosystem: Google hosted 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)
- **Quarterly Census of Employment and Wages** (U.S. Bureau of Labor Statistics)
- **Point-in-Time Homelessness Count** (U.S. Dept. of Housing and Urban Development)
- **Low Income Housing Tax Credit Program** (U.S. Dept. of Housing and Urban Development)
- **US Residential Real Estate Data** (House Canary)
- **Center for Medicare and Medicaid Services - Dual Enrollment** (U.S. Dept. of Health & Human Services)
- **Medicare** (U.S. Dept. of Health & Human Services)
- **Health Professional Shortage Areas** (U.S. Dept. of Health & Human Services)
- **CDC Births Data Summary** (Centers for Disease Control)
- **COVID-19 Data Repository by CSSE at JHU** (Johns Hopkins University)
- **COVID-19 Mobility Impact** (Geotab)
- **COVID-19 Open Data** (Google BigQuery Public Datasets Program)
- **COVID-19 Vaccination Access** (Google BigQuery Public Datasets Program)

Data Ecosystem Structure

FEDERATED
PUBLIC DATA
240+
(Population Science/ SDoH / Behavioral)

Hosted by Google & by ScHARe

REPOSITORY

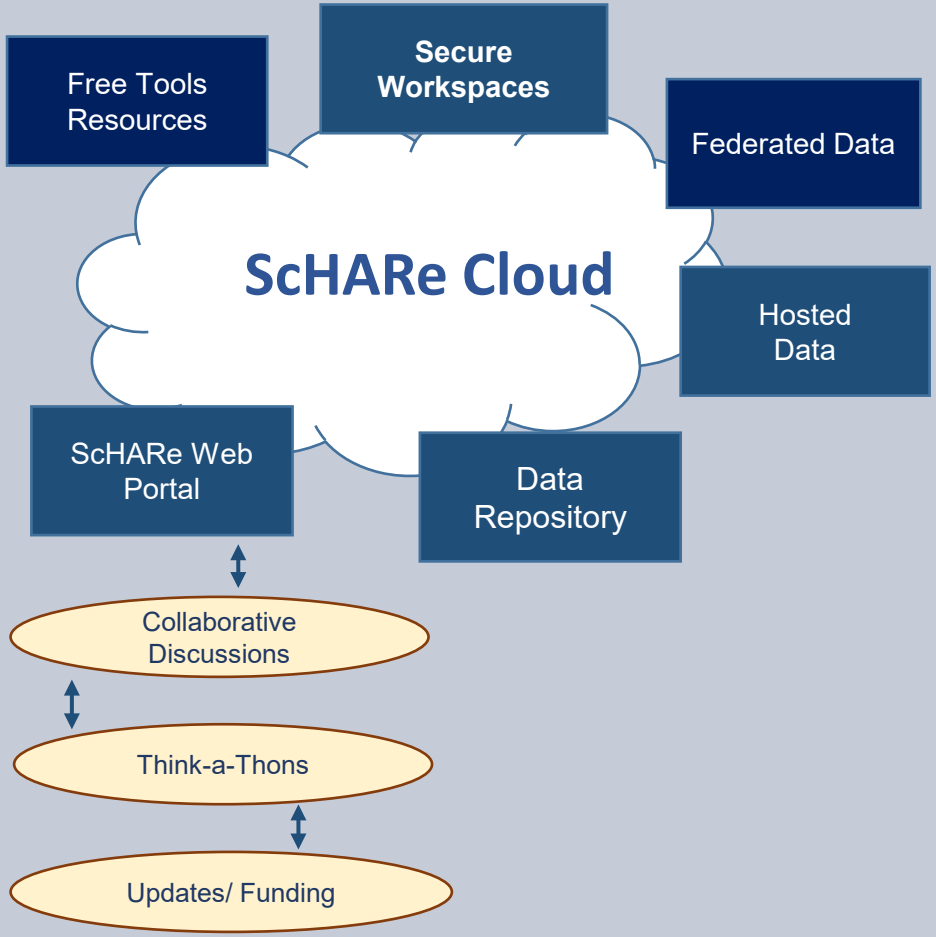
CDE FOCUSED

CDEs enhances Data Interoperability (Aggregation) by using semantic standards and concept codes

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

Components
Terra Interface

Intramural & Extramural Resource



ScHARe REPOSITORY

COMMON DATA ELEMENTS

NLM CDE Repository

Coded NIMHD Common Data Elements

- Labels
- Questions
- Permissible Values

DATA UPLOAD

Acquired **Google and ScHARe Hosted Datasets**

- Overview
- Data Dictionaries
- Data Updates

Project and Key Acquired Datasets

Overview

Description and Links to Overview Material

4-Privacy Levels

COMMON DATA ELEMENTS

Data

Metadata

Data Dictionaries

Analysis Ready

RAS Single Sign-on

ATO

Common Data Elements + **Data**

Data Access

Based On PII Levels and User Needs:

- Public
- Data Use Agreement
- Private

DATA MAPPING, DOWNLOAD AND EXPORT

Other Cloud Platforms
AnVil, BDC, All of Us

DATA MAPPING

ACROSS DATASETS AND PLATFORMS BASED ON CDES

EXAMPLE: CDE linked

ACS **NIMHD Project** **BioData Catalyst**

Aggregated Data Set

CDE Linked Project Data

Data Download in a Variety of Formats

CSV, TSV, XLSX

Data Export to Terra for Analysis

Workspaces

Visualizations Tools

Shiny

URI approach for data interoperability

ScHARe

Project & federated dataset mapping of ScHARe CDEs

- Project Title
- Project Description
- "Core Common Data Elements"
- Other Project Data
- Data Dictionary

\+ AMERICAN COMMUNITY SURVEY

\+ Medical Expenditure Survey (MEPS)

\+

\+ Pharmacy and health insurance databases

PREPARING FOR AI – RESEARCH AND HEALTH CARE USING BIG DATA

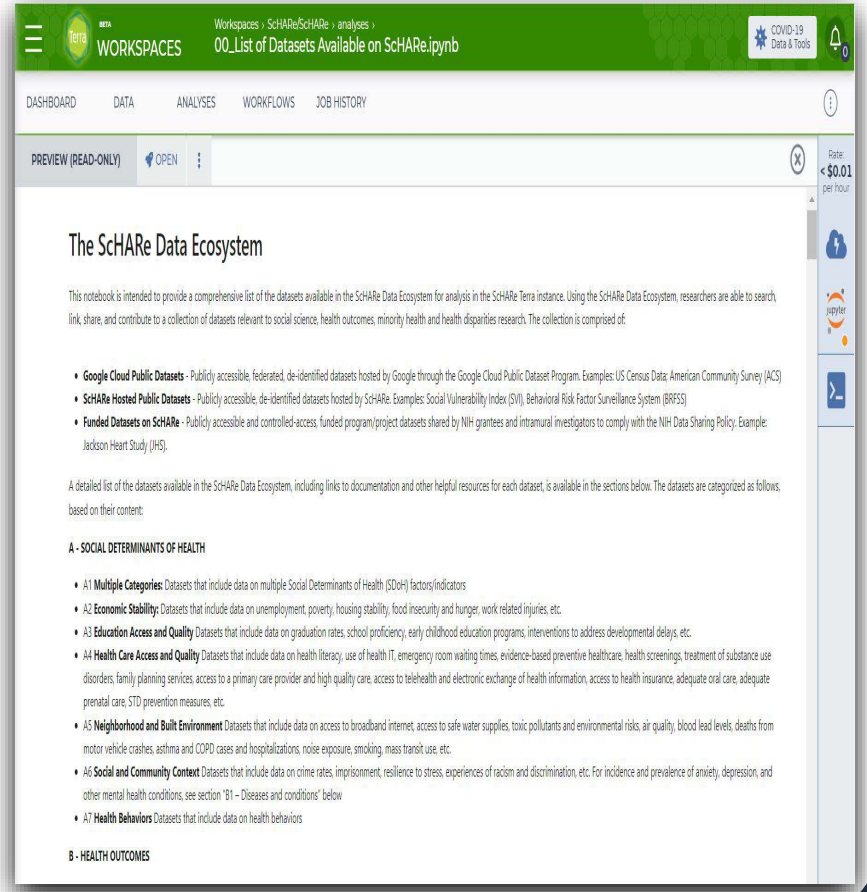
Mapping across cloud platforms with Terra Interface

BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION

Terra Interface: Data Sets and Access to Data

Analyses Tab in ScHARe workspace, the notebook **00_List of Datasets Available on ScHARe** lists all of the datasets available in the ScHARe Datasets collection

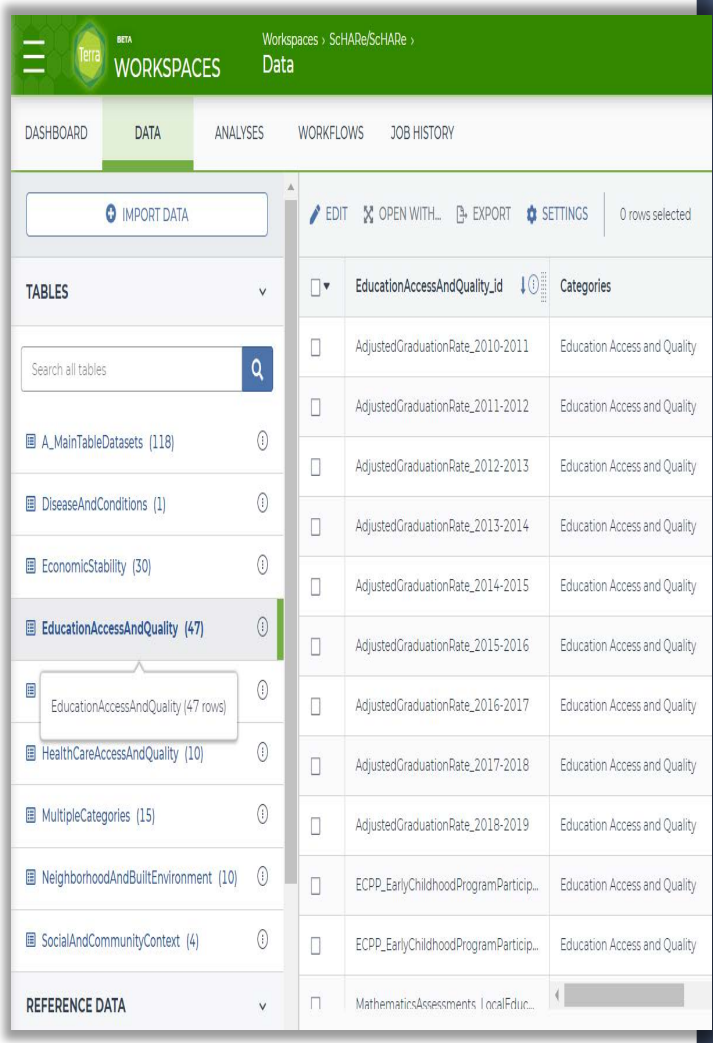
What?



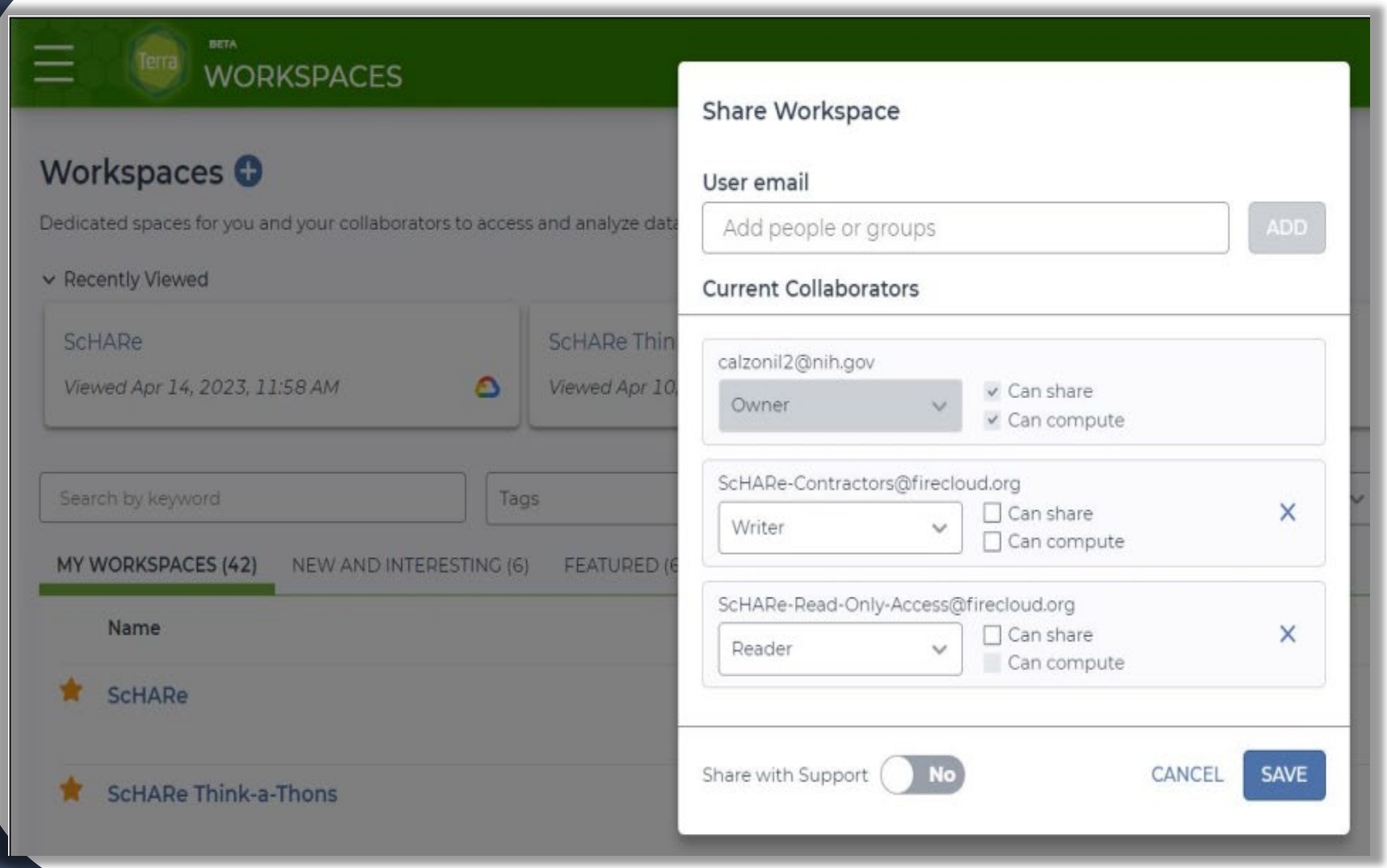
Where?

Data Tab in ScHARe workspace, **data tables help access ScHARe data and keep track of your project data:**

- ScHARe workspace, click on the Data tab
- Under Tables, see a list of dataset categories
- Click on a category, to see a list of relevant datasets
- Scroll to the right to learn more about each dataset

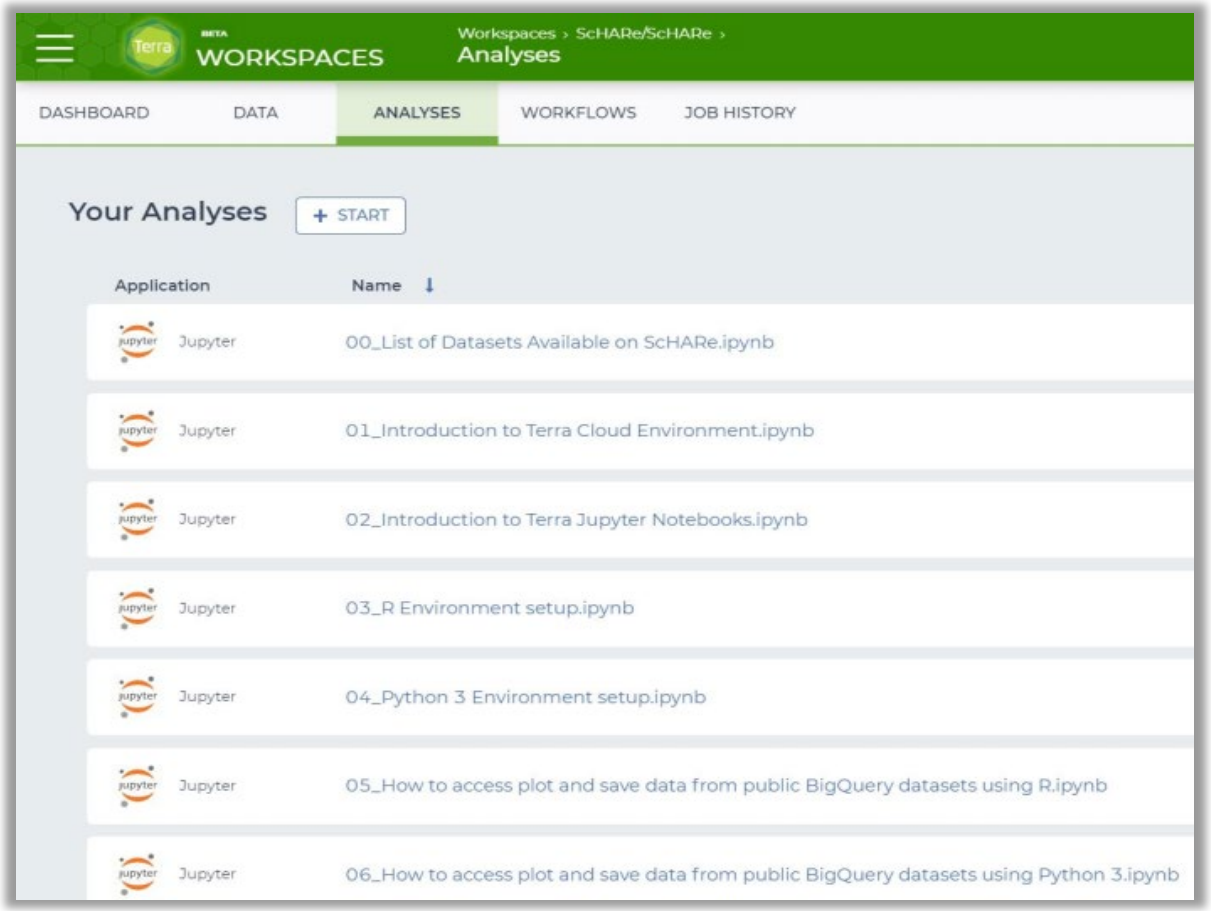


Terra Interface: Secure workspace



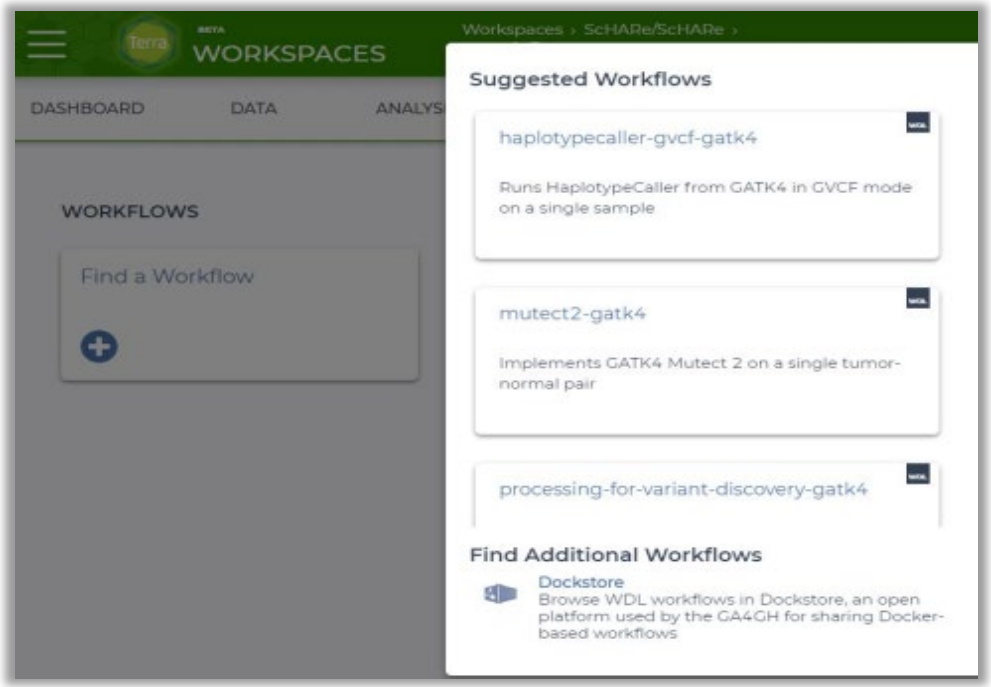
- Secure workspace **for self or collaborative research**
- **Assign roles**: review or admin
- **Host own data and code**

Terra Interface: Notebooks for Analytics & Tutorials



Workflows Modular codes

Easy to Use--Cut and Paste Analytics



- Modular codes developed for reuse
- **Adding SAS**

What are Think-a-Thons?

Think-a-Thons (TaT)

- Monthly sessions (2 1/2 hours)
- Instructional/interactive
- Designed for new and experienced users
- Research & analytic teams to:
  - Conduct health disparities, health outcomes, bias mitigation research
  - Analyze/create tools for bias mitigation
- Publications from research team collaboration
- Networking
- Mentoring and coaching
- Focus:

Types:

- ✓ **Instructional / Tutorial**
- ✓ **Collaborative Research Teams**
- ✓ **Bias mitigation**

Register:

bit.ly/think-a-thons

Think-a-Thon Tutorials

Web: bit.ly/think-a-thons

February	**Artificial Intelligence and Cloud Computing 101**
March	**ScHARe 1 – Accounts and Workspaces**
April	**ScHARe 2 – Terra Datasets**
May	**ScHARe 3 – Terra Google-hosted Datasets**
June	**ScHARe 4 – Terra ScHARe-hosted Datasets**
July	**An Introduction to Python for Data Science – Part 1**
August	**An Introduction to Python for Data Science – Part 2**
September	**ScHARe 5: A Review of the ScHARe Platform and Data Ecosystem**
October	**Preparing for AI 1: Common Data Elements and Data Aggregation**
November	**Preparing for AI 2: An Introduction to FAIR Data and AI-ready Datasets**
January	**Preparing for AI 3: Computational Data Science Strategies 101**
February	**Preparing for AI 4: Overview Prep for AI Summary with Transparency, Privacy, Ethics**

ScHARe for Educators (Community Colleges & Low Resource MSIs)
ScHARe for American Indian / Alaska Native Researchers
ScHARe for Coders and Programmers to conduct Research

Generational Career & Discipline Exchange

Think-a-Thons Training/Mentoring Pipeline

NLM
OIC Experts
Fellows

Think-a-Thons

✓ **Instructional**
✓ **Research**

N3C
Aim AHEAD
BioData Catalyst
All of US
AnVil
HEAL

Using experts + *To train and mentor novice users* + *To increase diverse perspectives in biomedical research*

Goal: "Upskilling"

- ✓ Data science specialist into health disparities and health outcomes research
- ✓ Health Disparity/Outcomes researchers into using big data and cloud computing

Target Audience:

- ✓ Underrepresented populations (women, race/ethnic) users not trained in data scien
- ✓ Data scientist with no or little research experience.
- ✓ Resource & Tool for Community Colleges and Low Resource MSIs and Organizations

Upcoming

ScHARe

Think-a-Thons (TaT)

Research Teams

Title: Data Science Projects 1 – Health Disparities and Individual SDoH **Description:** Exploring the impact of individual Social Determinants of Health on health outcomes: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.
Title: Data Science Projects 2 - Health Disparities and Structural SDoH **Description:** Assessing the impact of structural Social Determinants of Health on health outcomes: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.
Title: Data Science Projects 3 – Health Outcomes **Description:** Investigating the influence of non-clinical factors on disparities in health care delivery: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.

- Foster a research paradigm shift to use Big Data
- Promote use of Dark Data

- Multi-career (students to sr. investigators)
- Multi-discipline (data scientist & researchers)
- Feature Datasets with Guest Expert Leads
- Secure experts in topic area, analytics, data sources etc. to provide guidance
- Generate research idea - decide potential design, datasets & analytics
- Select co-leads to coordinate completion outside of TaT
- Publications

Register:

bit.ly/think-a-thons

ScHARe CDE Adoption:

Making Data Interoperable (URI Approach)
Concept Code Mapping (Harmonization)

ScHARe

Adopted CDEs to:

- Standardize data for people & computers (human and machine readable)
- Enable data sharing across studies (data interoperability)
- Enhance data interpretation & analysis (semantically defined and standardized coded)
- Simplify collaboration
- Reduces project start-up & results time

***BIG DATA AND AI*: REQUIRES NEW APPROACHES FOR COLLECTION, MANAGEMENT, ANALYSIS**

Covid revealed the need to have real time data

ScHARe

Core Common Data Elements Working Group

NIMHD Extramural
- Barksdale, Crystal
- Calzoni, Luca*
- Dinwiddie, Gniesha
- Doose, Michelle
- Duran, Deborah*
- Linares, Deborah
- Le, Phuong-Tu
- Sidhu, Simrann

NIMHD Intramural
- Choi, Kelvin
- Strassle, Paula

NINR Extramural
- Grason, John
- Hawes, Rebecca

NINR Intramural
- Steele, Michael

CDE and/or Health Disparities Experts:
- McAuliffe, Matthew (CIT)
- Mendoza-Puccini, Carolina (NINDS)
- Alvidrez, Jennifer (NIH/OD)

ScHARe "CORE" CDE Development

Core Set:

- Few critical questions required from all studies/sites
- Minimal burden
- Allows for questions to be asked in any way, but reported in a standardized format
- Allows for any number of other questions to be collected as collector chooses

Criteria:

- PhenX Toolkit first
- Validated source
- Adaptation of a validated source
- Generate new gap area CDE

ScHARe CDEs Defined and Coded – URI Approach

Education

What is the highest level of education you have completed?

Shared Semantics and Concept Code:
An indication of the years of schooling completed in graded public, private, or parochial schools, and in colleges, universities, or professional schools. **C17953**

URI approach in data repository uses codes to harmonize data rather than semantics. This improves data interoperability.

***Survey Questions become CDEs when they are semantically defined by a standardized coding system for shared meaning and in a format that is human and machine readable for ease of reuse**

How a Survey Question Became a CDE

Survey Questions become CDEs when they are:

- **semantically defined by a standardized coding system for shared meaning**
- **in a format that is human and machine readable for ease of reuse**

Making of a CDE from a Protocol/Question

Need a standardized defined concept and related code.
Source: NCI Thesaurus

Race/Ethnicity Self-Identification

A textual description of a person's race. C17049 |The ethnicity of a person. C16564 | An individual's perspective or subjective interpretation of an event or information. C74528

American Indian or Alaska Native |
Asian or Asian American |
Black of African American |
Hispanic, Latino, or Spanish |
Native Hawaiian or Other Pacific Islander |
Middle Eastern or North African |
White

URI approach in data repository uses codes to harmonize data rather than semantics. This improves data interoperability.

Making of a CDE from a Protocol/Question

- A person having origins in any of the original peoples of North and South America (including Central America) and who maintains tribal affiliation or community attachment. (OMB) C41259 |
- A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent, including for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. (OMB) C41260 |
- A person having origins in any of the Black racial groups of Africa. Terms such as "Haitian" or "Negro" can be used in addition to "Black or African American". (OMB) C16352 |
- A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race. The term, "Spanish origin" can be used in addition to "Hispanic or Latino". (OMB) C17459 |
- A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands. (OMB) C41219 |
- Denotes a person having origins in the region of southwest Asia, between the India subcontinent and Europe, including Kuwait, Turkey, Lebanon, Israel, Iraq, Iran, Jordan, Saudi Arabia, lands east of Pakistan or the other countries of the Arabian Peninsula. Also includes people of Jewish ethnicity including Sephardic and Ashkenazic. C77820 :
- Denotes a person whose ancestry is in any of the countries of the northern part of the African continent: Algeria, Egypt, Libya, Morocco, Sudan, Tunisia, and Western Sahara. C126529 |
- A person having origins in any of the original peoples of Europe, the Middle East, or North Africa. (OMB) C41261

Making of a CDE from a Protocol/Question

Need a standardized defined concept and related code. Source: NCI Thesaurus

Code Mapping

	NCIT	Loinc	UMLS CUI
American Indian or Alaska Native	C41259	LA10608-0	C0282204
Asian or Asian American	C41260	LA6156-9	C0003988
Black of African American	C16352	LA10610-6	C0085756
Hispanic, Latino, or Spanish	C17459	LA6214-6	C0086409
Native Hawaiian or Other Pacific Islander	C41219	LA10611-4	C1513907
Middle Eastern or North African	C43866	Mena no loinc	C1553353
White	C41261	LA4457-3	C0043157

Matched CDE

Income (Project 1)		Income (Project 2)
Less than $10,000		Less than $10,000
$10,000-$24,999		$10,000-$24,999
$25,000-$34,999		$25,000-$34,999
$35,000-$49,999		$35,000-$49,999
$50,000-$74,999		$50,000-$74,999
$75,000-$99,999		$75,000-$99,999
$100,000-$149,999		$100,000-$149,999
$150,000-$199,999		$150,000-$199,999
$200,000 or more		$200,000 or more

Reported this way

Collected this way

Mappable CDE

Income (Project 1)

Less than $10,000 |
$10,000-$24,999 |
$25,000-$34,999 |
$35,000-$49,999 |
$50,000-$74,999 |
$75,000-$99,999 |
$100,000-$149,999 |
$150,000-$199,999 |
$200,000 or more

Reported this way

Income (Project 2)

Less than $10,000 |
$10,000-$19,999 |
$20,000-$29,999 |
$30,000-$39,999 |
$40,000-$49,999 |
$50,000-$59,999 |
$60,000-$69,999 |
$70,000-$79,999 |
$80,000-$89,999 |
$90,000-$99,999 |
…….
$200,000 or more

Collected this way

Mapped using algorithms

ScHARe CDEs in Survey Format

ScHARe CDEs Labels

For FUNDED PROJECT DATA – Common Data Elements Centralized for Interoperability and Data Sharing

- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex
- Gender
- Sexual Orientation
- 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 (Associated Medications/Treatments)

NIH Endorsed

**NIMHD Framework
**Health Disparity Outcomes

(** project level CDE)

NIH CDE Repository: https://cde.nlm.nih.gov/home

Cross-walked with PhenX SDoH

NIH-endorsed CDEs have been reviewed and approved by an expert panel, and meet established criteria. They are designated with a gold ribbon.

ScHARe

COMMON DATA ELEMENTS

NLM CDE Repository

Coded NIMHD Common Data Elements

- Labels
- Questions
- Permissible Values

DATA UPLOAD

Acquired **Google and ScHARe Hosted Datasets**

- Overview
- Data Dictionaries
- Data Updates

REPOSITORY

Project and Key Acquired Datasets

Overview

Description and Links to Overview Material

4-Privacy Levels

COMMON DATA ELEMENTS

Data

Metadata

Data Dictionaries

Analysis Ready

RAS Single Sign-on

DATA MAPPING, DOWNLOAD AND EXPORT

Other Cloud Platforms
AnVil, BDC, All of Us

DATA MAPPING

ACROSS DATASETS AND PLATFORMS BASED ON CDES

EXAMPLE: CDE linked

ACS **NIMHD Project** **BioData Catalyst**

Aggregated Data Set

CDE Linked Project Data

Data Download in a Variety of Formats

CSV, TSV, XLSX

Data Export to Terra for Analysis

Workspaces

Visualizations Tools

Shiny

ATO

Common Data Elements + **Data**

Data Access

Based On PII Levels and User Needs:

- Public
- Data Use Agreement
- Private

URI approach for data interoperability

ScHARe

Thank You

Next Think-a-Thons:

bit.ly/think-a-thons

Register for ScHARe:

bit.ly/join-schare

schare@mail.nih.gov

durande@mail.nih.gov or luca.calzoni@nih.gov