ScHARe

Think-a-Thons

NIH National Institutes of Health

Preparing for AI-driven Research on ScHARe - Part 2

A Comprehensive Review and Brainstorming Session

Deborah Duran, PhD · NIMHD
Luca Calzoni, MD MS PhD Cand. · NIMHD

March 20, 2024

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

NIH National Institute on Minority Health and Health Disparities + NIH Office of Data Science Strategy + NIH National Institute of Nursing Research

NIH National Institute of Dental and Craniofacial Research

OMH U.S. Department of Health and Human Services Office of Minority Health

NIH National Institutes of Health Office of Research on Women's 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

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	□	□	□	□
Algorithmic bias mitigation	□	□	□	□

Outline

5’ Introduction

- Experience poll

15’ ScHARe overview

- Interest poll

35’ Computational strategies

- Polls

20’ Python data science libraries

15’ Testing and monitoring in algorithm development

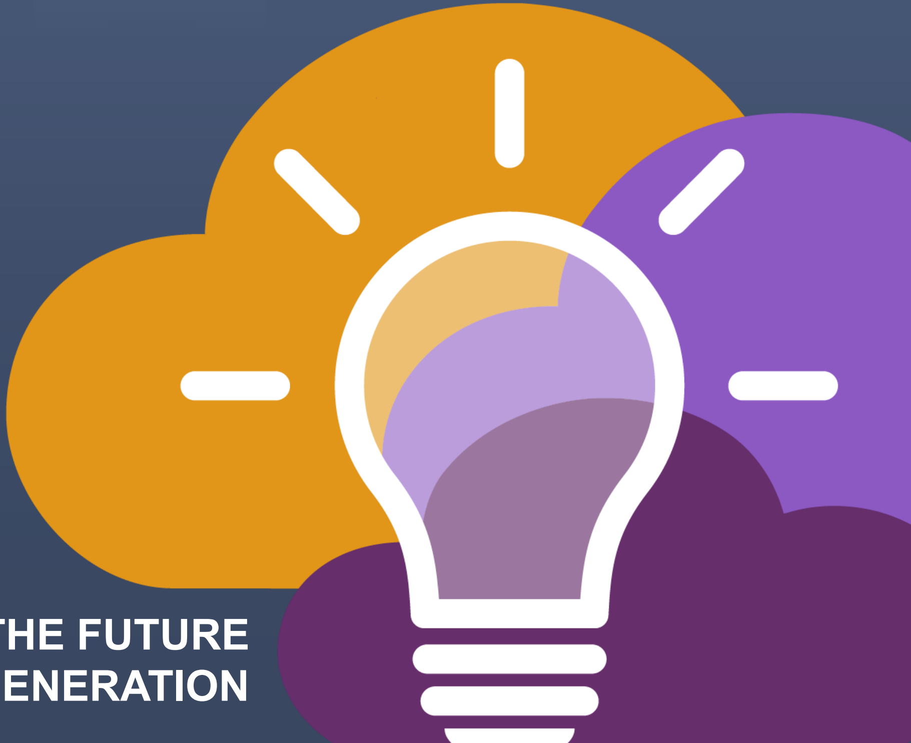
15’ Open science and reproducible research

45’ Research Think-a-Thons brainstorming

- Final poll

ScHARe

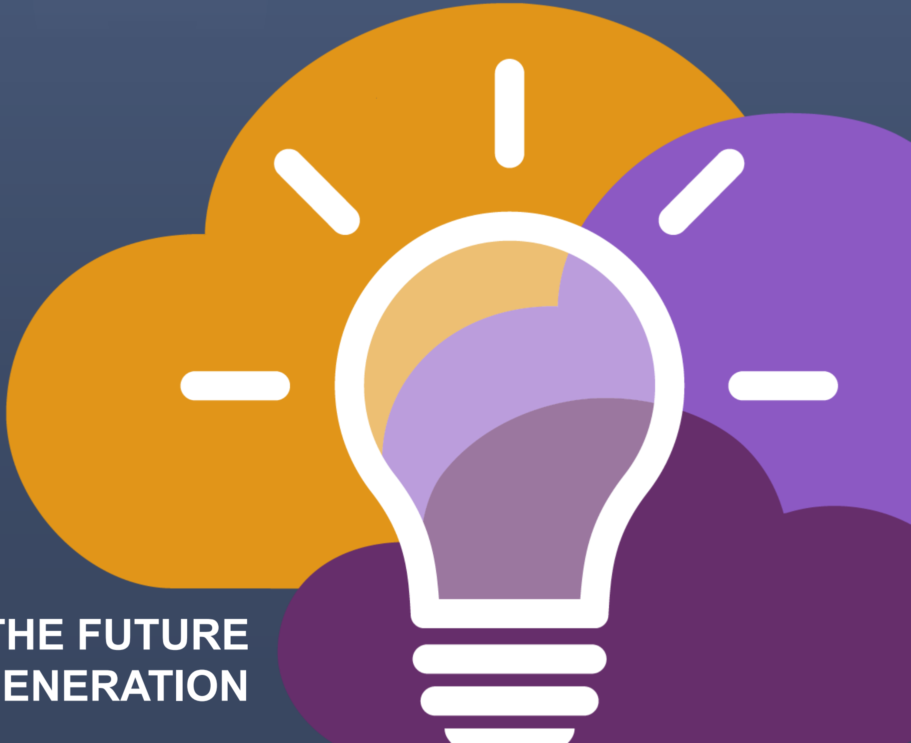
Overview



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe

What is ScHARe?



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

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

Google Platform Terra Interface

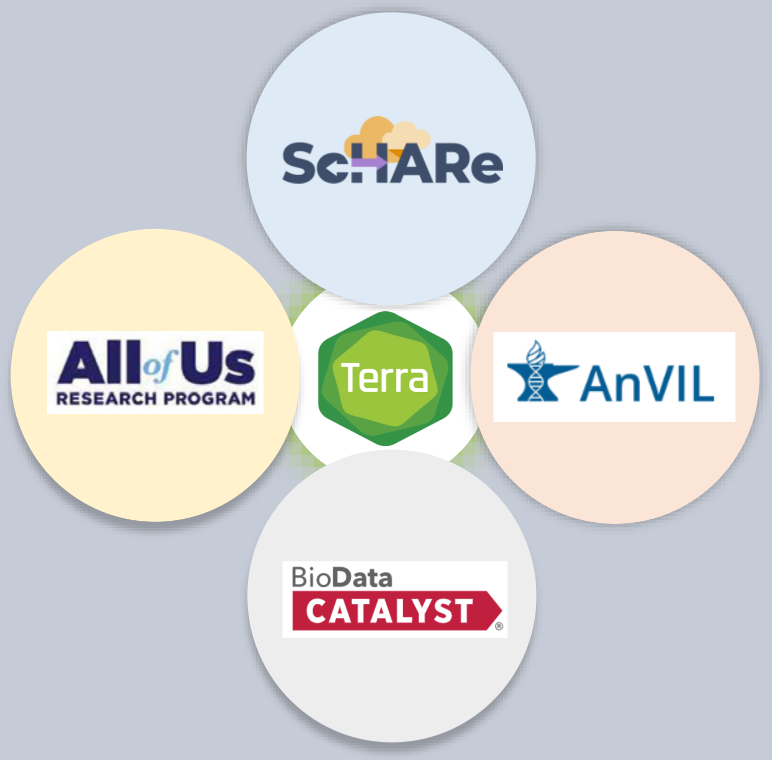
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Cut and paste code in Python and R

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



BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION

Data Ecosystem structure

Population Science/SDoH

240+
FEDERATED PUBLIC DATASETS

- **Population Science / SDoH / Behavioral**
- Hosted by Google & ScHARe

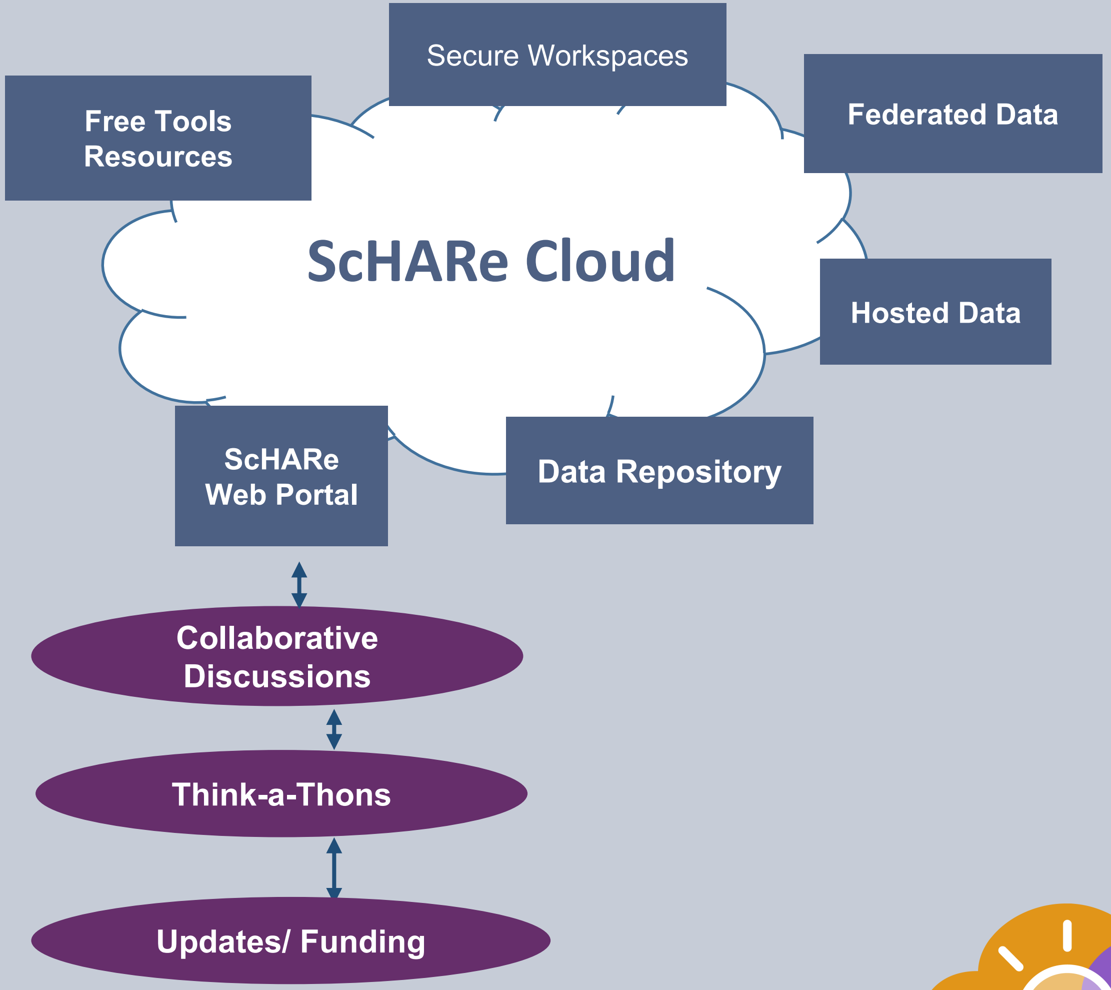
REPOSITORY
CDE FOCUSED

- **CDEs enhance data interoperability** (aggregation) by using semantic standards and concept codes

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

COMPONENTS

Intramural and Extramural Resource



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)*

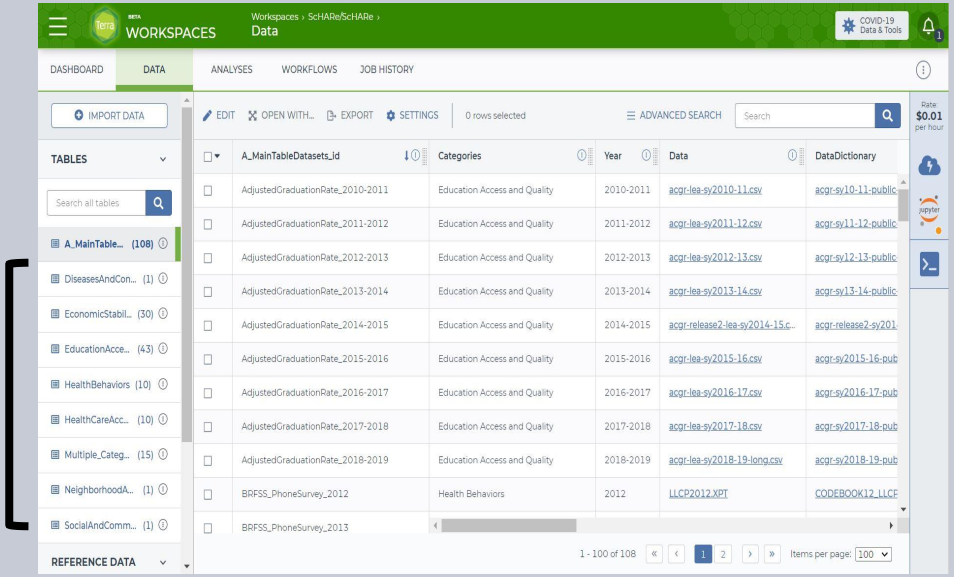
- **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe

  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



TABLES	
A_MainTable...	(108)
DiseasesAndCon...	(1)
EconomicStabil...	(30)
EducationAcce...	(43)
HealthBehaviors	(10)
HealthCareAcc...	(10)
Multiple_Categ...	(15)
NeighborhoodA...	(1)
SocialAndComm...	(1)

A_MainTableDatasets_id	Categories	Year	Data	DataDictionary
AdjustedGraduationRate_2010-2011	Education Access and Quality	2010-2011	acgr-lea-sy2010-11.csv	acgr-sy10-11-public-
AdjustedGraduationRate_2011-2012	Education Access and Quality	2011-2012	acgr-lea-sy2011-12.csv	acgr-sy11-12-public-
AdjustedGraduationRate_2012-2013	Education Access and Quality	2012-2013	acgr-lea-sy2012-13.csv	acgr-sy12-13-public-
AdjustedGraduationRate_2013-2014	Education Access and Quality	2013-2014	acgr-lea-sy2013-14.csv	acgr-sy13-14-public-
AdjustedGraduationRate_2014-2015	Education Access and Quality	2014-2015	acgr-release2-lea-sy2014-15.c...	acgr-release2-sy201
AdjustedGraduationRate_2015-2016	Education Access and Quality	2015-2016	acgr-lea-sy2015-16.csv	acgr-sy2015-16-pub
AdjustedGraduationRate_2016-2017	Education Access and Quality	2016-2017	acgr-lea-sy2016-17.csv	acgr-sy2016-17-pub
AdjustedGraduationRate_2017-2018	Education Access and Quality	2017-2018	acgr-lea-sy2017-18.csv	acgr-sy2017-18-pub
AdjustedGraduationRate_2018-2019	Education Access and Quality	2018-2019	acgr-lea-sy2018-19-long.csv	acgr-sy2018-19-pub
BRFSS_PhoneSurvey_2012	Health Behaviors	2012	LLCP2012.XPT	CODEBOOK12_LLCF
BRFSS_PhoneSurvey_2013				

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)

ScHARe Ecosystem: ScHARe hosted datasets

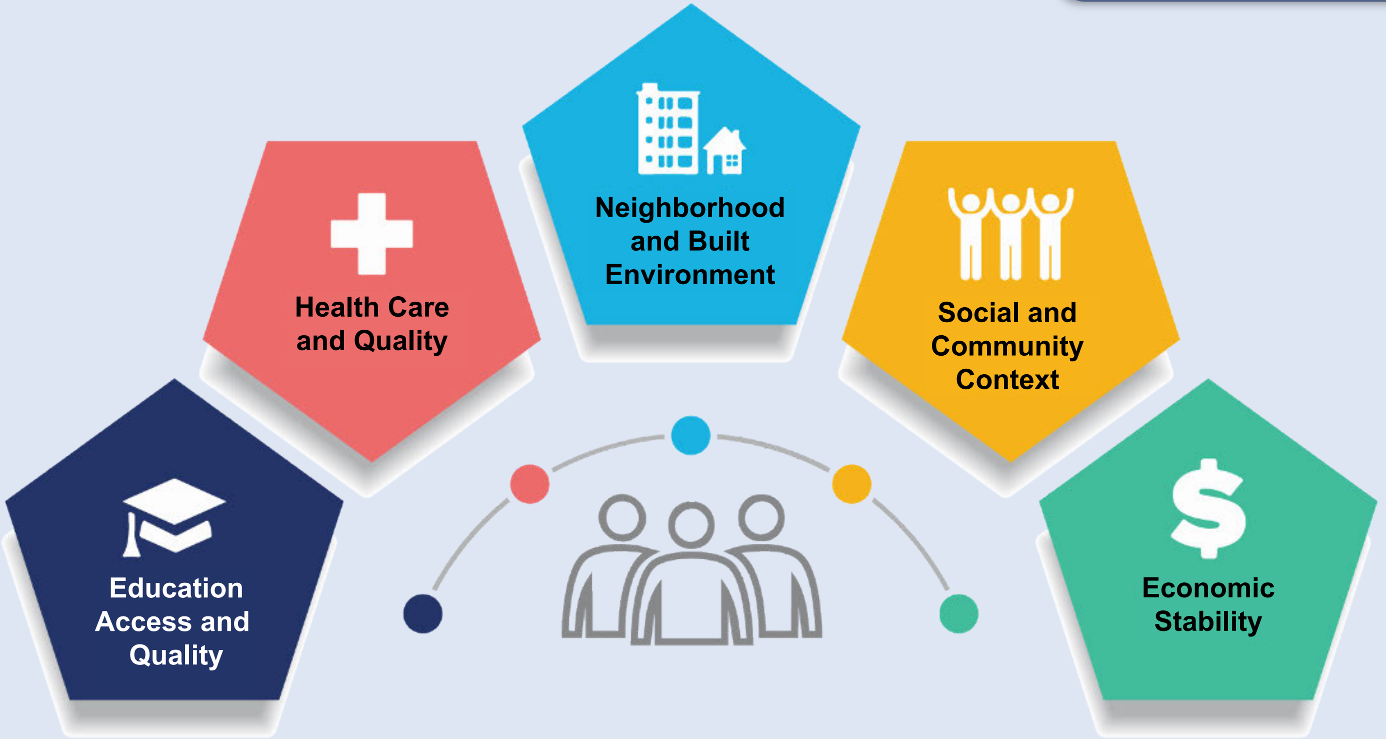
Organized based on the **CDC SDoH categories**, with the addition of *Health Behaviors* and *Diseases and Conditions*:

200+ datasets

- **What are the Social Determinants of Health?**

Social determinants of health (SDoH) are the **nonmedical factors that influence health outcomes.**

They are the **conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life.**


Health Care and Quality
Neighborhood and Built Environment
Social and Community Context
Education Access and Quality
Economic Stability

www.cdc.gov/about/sdoh/index.html

ScHARe Ecosystem: ScHARe hosted datasets

Examples of datasets for each category include:

Education access and quality

Data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.

Examples:

- **EDFacts Data Files** (U.S. Dept. of Education) - Graduation rates and participation/proficiency assessment
- **NHES - National Household Education Surveys Program** (U.S. Dept. of Education) – Educational activities

ScHARe Ecosystem: ScHARe hosted datasets

Health care access and quality

Data on health literacy, use of health IT, emergency room waiting times, preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.

Example:

- **MEPS - Medical Expenditure Panel Survey** (AHRQ) - Cost and use of healthcare and health insurance coverage
- **Dartmouth Atlas Data -** Selected Primary Care Access and Quality Measures - Measures of primary care utilization, quality of care for diabetes, mammography, leg amputation and preventable hospitalizations

ScHARe Ecosystem: ScHARe hosted datasets

Neighborhood and built environment

Data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.

Examples:

- **National Environmental Public Health Tracking Network** (CDC) - Environmental indicators and health, exposure, and hazard data
- **LATCH - Local Area Transportation Characteristics for Households** (U.S. Dept. of Transportation) – Local transportation characteristics for households

ScHARe Ecosystem: ScHARe hosted datasets

Social and community context

Data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc.

Example:

- **Hate crime statistics** (FBI) - Data on crimes motivated by bias against race, gender identity, religion, disability, sexual orientation, or ethnicity
- **General Social Survey** (GSS) - Data on a wide range of characteristics, attitudes, and behaviors of Americans.

ScHARe Ecosystem: ScHARe hosted datasets

Economic stability

Data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.

Examples:

- **Current Population Survey (CPS) Annual Social and Economic Supplement** (U.S. Bureau of Labor Statistics) - Labor force statistics: annual work activity, income, health insurance, and health
- **Food Access Research Atlas** (U.S. Dept. of Agriculture) – Food access indicators for low-income and other census tracts

ScHARe Ecosystem: ScHARe hosted datasets

Health behaviors

Data on health-related practices that can directly affect health outcomes.

Examples:

- **BRFSS - Behavioral Risk Factor Surveillance System** (CDC) - State-level data on health-related risk behaviors, chronic health conditions, and use of preventive services
- **YRBSS - Youth Risk Behavior Surveillance System** (CDC) – Health behaviors that contribute to the leading causes of death, disability, and social problems among youth and adults

ScHARe Ecosystem: ScHARe hosted datasets

Diseases and conditions

Data on incidence and prevalence of specific diseases and health conditions.

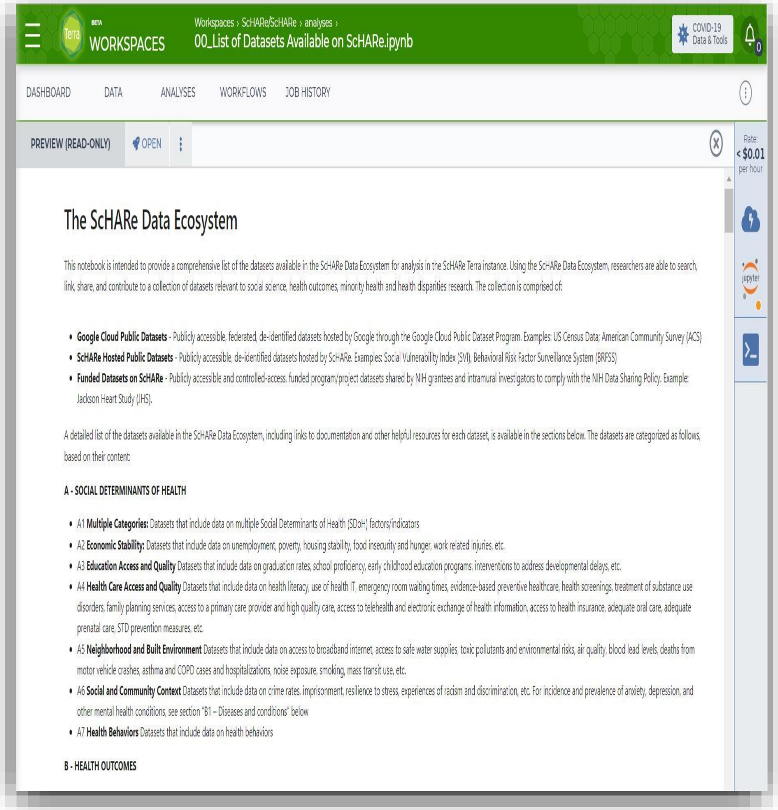
Examples:

- **U.S. CDI - Chronic Disease Indicators** (CDC) - 124 chronic disease indicators important to public health practice
- **UNOS - United Network of Organ Sharing** (Health Resources and Services Administration) – Organ transplantation: cadaveric and living donor characteristics, survival rates, waiting lists and organ disposition

Terra Interface: Datasets 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?



The ScHARe Data Ecosystem

This notebook is intended to provide a comprehensive list of the datasets available in the ScHARe Data Ecosystem for analysis in the ScHARe Terra instance. Using the ScHARe Data Ecosystem, researchers are able to search, link, share, and contribute to a collection of datasets relevant to social science, health outcomes, minority health and health disparities research. The collection is comprised of:

- **Google Cloud Public Datasets** - Publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program. Examples: US Census Data; American Community Survey (ACS)
- **ScHARe Hosted Public Datasets** - Publicly accessible, de-identified datasets hosted by ScHARe. Examples: Social Vulnerability Index (SVI), Behavioral Risk Factor Surveillance System (BRFSS)
- **Funded Datasets on ScHARe** - Publicly accessible and controlled-access, funded program/project datasets shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy. Example: Jackson Heart Study (JHS).

A detailed list of the datasets available in the ScHARe Data Ecosystem, including links to documentation and other helpful resources for each dataset, is available in the sections below. The datasets are categorized as follows, based on their content:

A - SOCIAL DETERMINANTS OF HEALTH

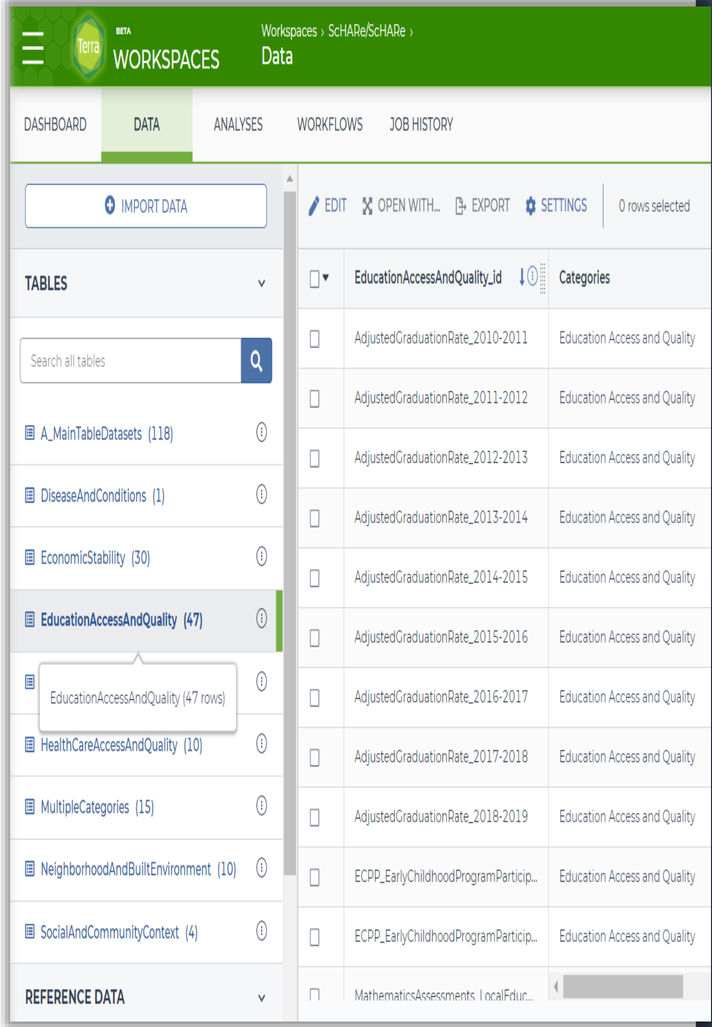
- A1 **Multiple Categories:** Datasets that include data on multiple Social Determinants of Health (SDoH) factors/indicators
- A2 **Economic Stability:** Datasets that include data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.
- A3 **Education Access and Quality** Datasets that include data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.
- A4 **Health Care Access and Quality** Datasets that include data on health literacy, use of health IT, emergency room waiting times, evidence-based preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.
- A5 **Neighborhood and Built Environment** Datasets that include data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.
- A6 **Social and Community Context** Datasets that include data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc. For incidence and prevalence of anxiety, depression, and other mental health conditions, see section "B1 – Diseases and conditions" below
- A7 **Health Behaviors** Datasets that include data on health behaviors

B - HEALTH OUTCOMES

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

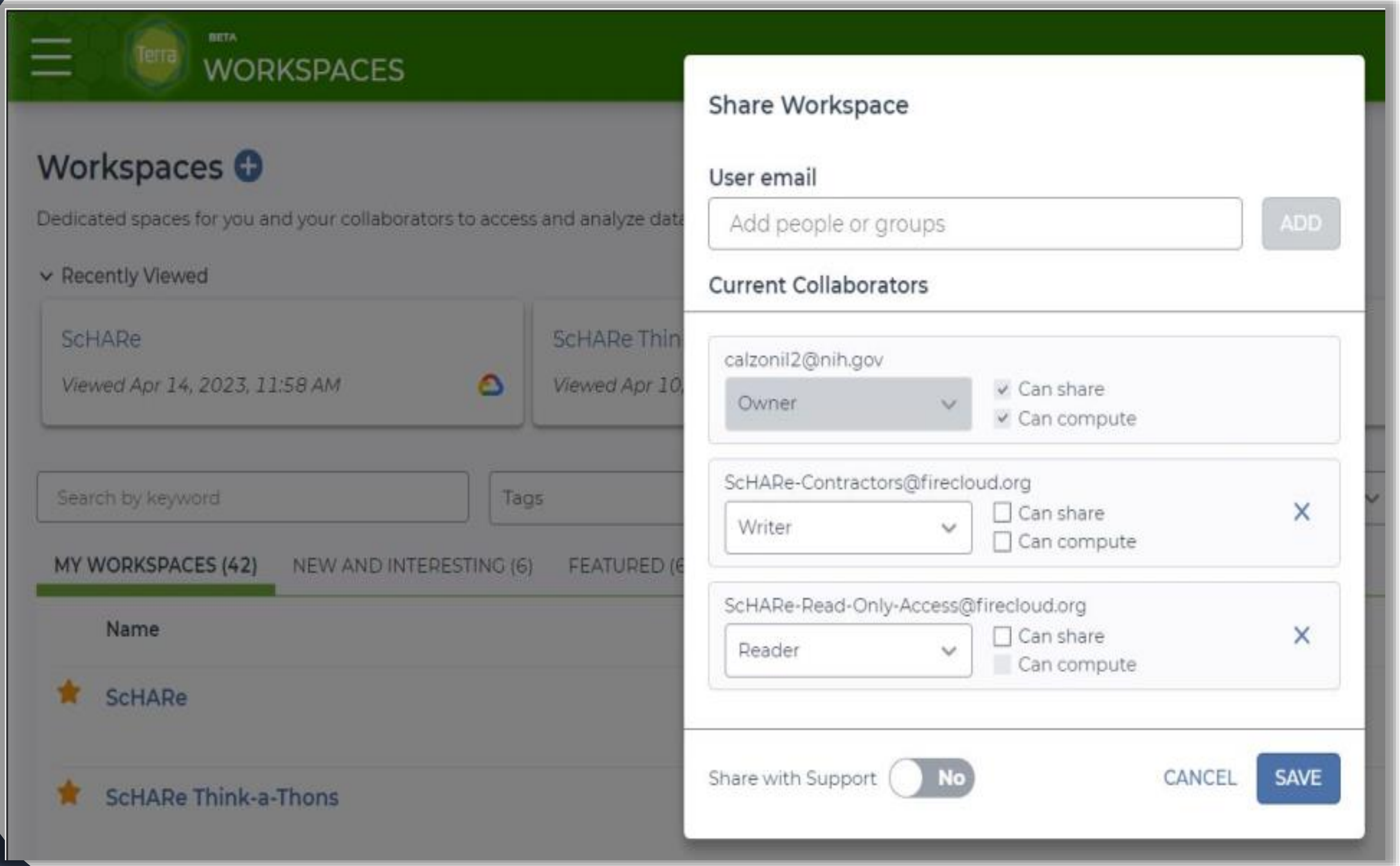
- ScHARe workspace, click on the Data tab
- Under Tables, see a list of dataset categories
- Click on a category, to see a list of relevant datasets
- Scroll to the right to learn more about each dataset

Where?



☐	EducationAccessAndQuality_id	Categories
☐	AdjustedGraduationRate_2010-2011	Education Access and Quality
☐	AdjustedGraduationRate_2011-2012	Education Access and Quality
☐	AdjustedGraduationRate_2012-2013	Education Access and Quality
☐	AdjustedGraduationRate_2013-2014	Education Access and Quality
☐	AdjustedGraduationRate_2014-2015	Education Access and Quality
☐	AdjustedGraduationRate_2015-2016	Education Access and Quality
☐	AdjustedGraduationRate_2016-2017	Education Access and Quality
☐	AdjustedGraduationRate_2017-2018	Education Access and Quality
☐	AdjustedGraduationRate_2018-2019	Education Access and Quality
☐	ECPP_EarlyChildhoodProgramParticip...	Education Access and Quality
☐	ECPP_EarlyChildhoodProgramParticip...	Education Access and Quality
☐	MathematicsAssessments_LocalEduc...	

Terra Interface: Secure workspace



BETA WORKSPACES

Workspaces

Dedicated spaces for you and your collaborators to access and analyze dat

Recently Viewed

ScHARe
Viewed Apr 14, 2023, 11:58 AM

ScHARe Thin
Viewed Apr 10,

Search by keyword

Tags

MY WORKSPACES (42) NEW AND INTERESTING (6) FEATURED (6

Name

ScHARe

ScHARe Think-a-Thons

Share Workspace

User email

Add people or groups [ADD]

Current Collaborators

calzonil2@nih.gov
Owner
☑ Can share
☑ Can compute

ScHARe-Contractors@firecloud.org
Writer
□ Can share
□ Can compute

ScHARe-Read-Only-Access@firecloud.org
Reader
□ Can share
□ Can compute

Share with Support: No

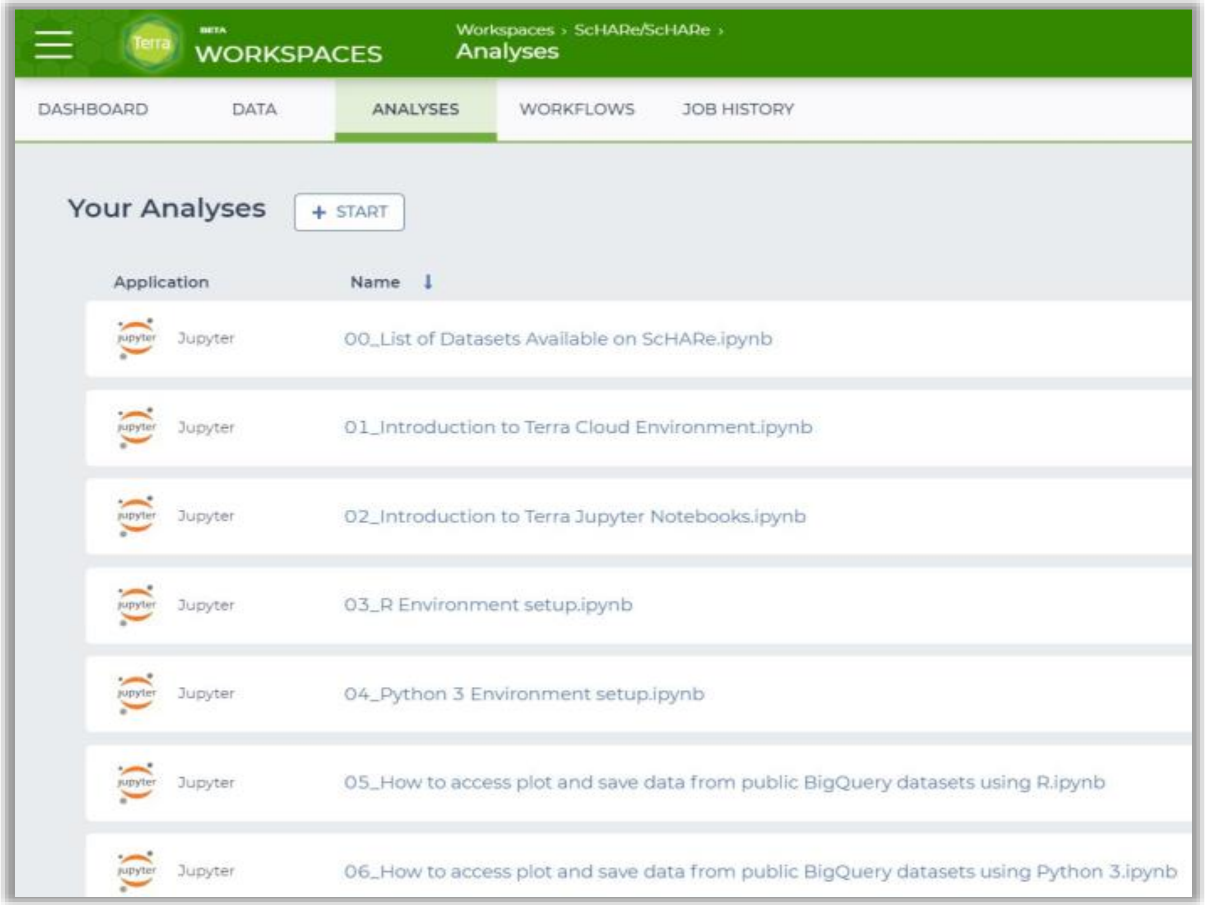
CANCEL SAVE

- 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

A notebook integrates code and its output into a single document where you can run code, display the output, and also add explanations, formulas, and charts



BETA WORKSPACES

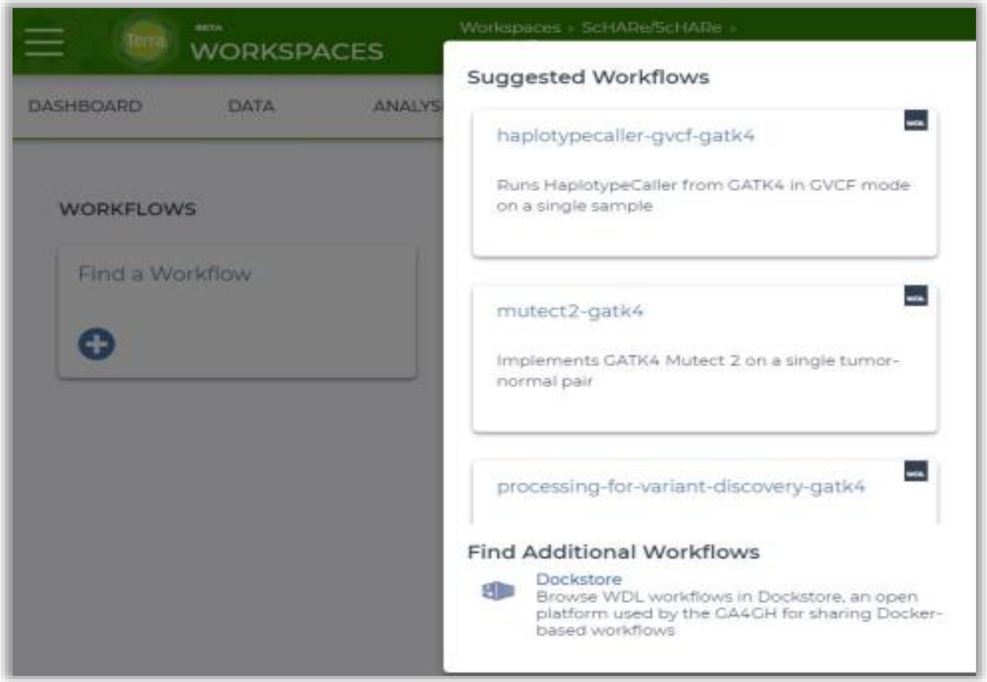
Workspaces › ScHARe/ScHARe ›
Analyses

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

Your Analyses + START

Application	Name ↓
Jupyter	00_List of Datasets Available on ScHARe.ipynb
Jupyter	01_Introduction to Terra Cloud Environment.ipynb
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb
Jupyter	03_R Environment setup.ipynb
Jupyter	04_Python 3 Environment setup.ipynb
Jupyter	05_How to access plot and save data from public BigQuery datasets using R.ipynb
Jupyter	06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

Easy to Use--Cut and Paste Analytics



BETA WORKSPACES

Workspaces › ScHARe/ScHARe ›

DASHBOARD DATA ANALYS

WORKFLOWS

Find a Workflow

Suggested Workflows

haplotypecaller-gvcf-gatk4

Runs HaplotypeCaller from GATK4 in GVCF mode on a single sample

mutect2-gatk4

Implements GATK4 Mutect 2 on a single tumor-normal pair

processing-for-variant-discovery-gatk4

Find Additional Workflows

Dockstore
Browse WDL workflows in Dockstore, an open platform used by the GA4GH for sharing Docker-based workflows

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

ScHARe

PREPARING FOR CLOUD COMPUTING AI – HEALTH DISPARITY AND CARE RESEARCH USING BIG DATA

AGGRAGATING DATA SETS TOOL: Variables & CDEs

CDE Mapping Project & Federated Data

Project Title

Project Description

“Core Common Data Elements”

Other Project Data

Data Dictionary

+ AMERICAN COMMUNITY SURVEY

+

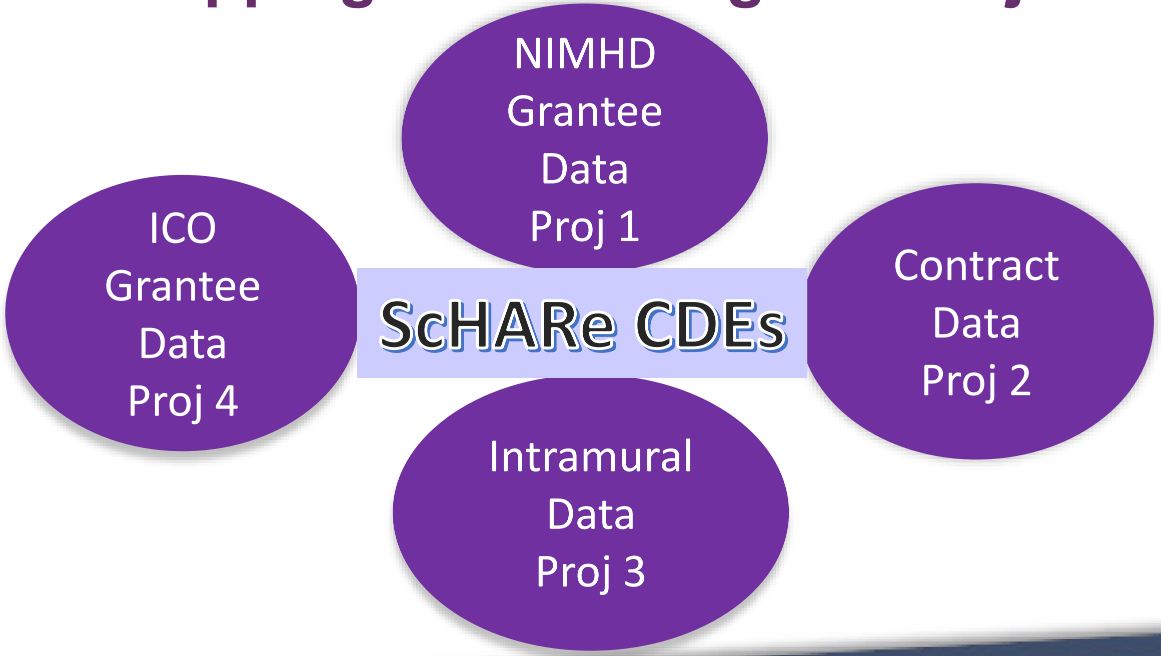
+ BRFSS

+

+ Medical Expenditure Survey (MEPS)

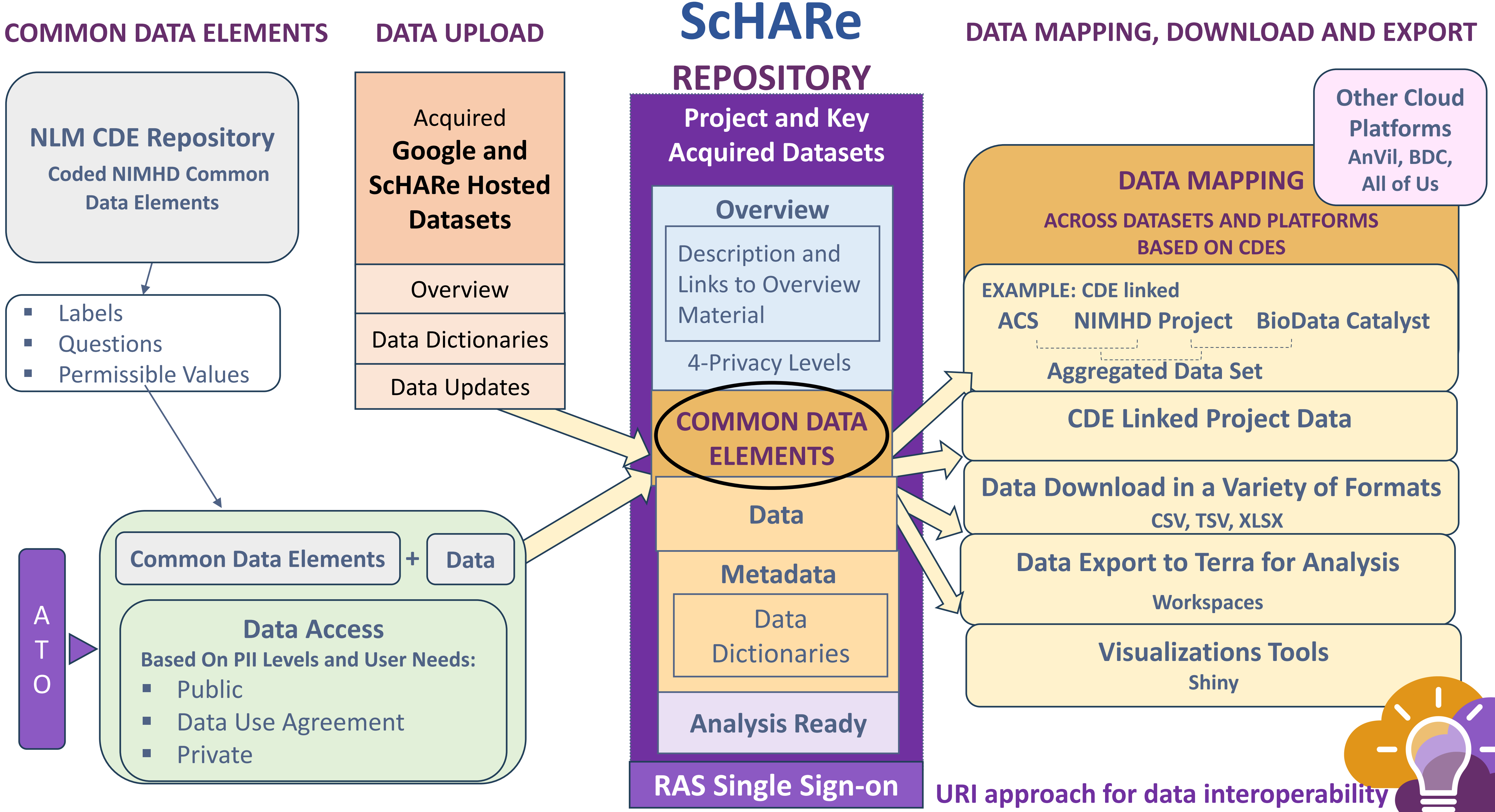
***Assign DOIs to aggregated data sets*

CDE Mapping across Program Projects



BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION

ScHARe



COMMON DATA ELEMENTS

NLM CDE Repository

Coded NIMHD Common Data Elements

- Labels
- Questions
- Permissible Values

Common Data Elements + Data

A T O

Data Access

Based On PII Levels and User Needs:

- Public
- Data Use Agreement
- Private

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

URI approach for data interoperability