ScHARe

Research Think-a-Thons

National Institutes of Health

ScHARe Research Think-a-Thon

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

April 17, 2024

Outline

10’ Introduction and ScHARe overview
- Experience poll
- Interest poll

5’ Research Think-a-Thon expectations

10’ ScHARe platform orientation

15’ Pick a team

1h20’ Brainstorming (Breakout rooms)

30’ Team updates
- Evaluation poll

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	☐	☐	☐	☐

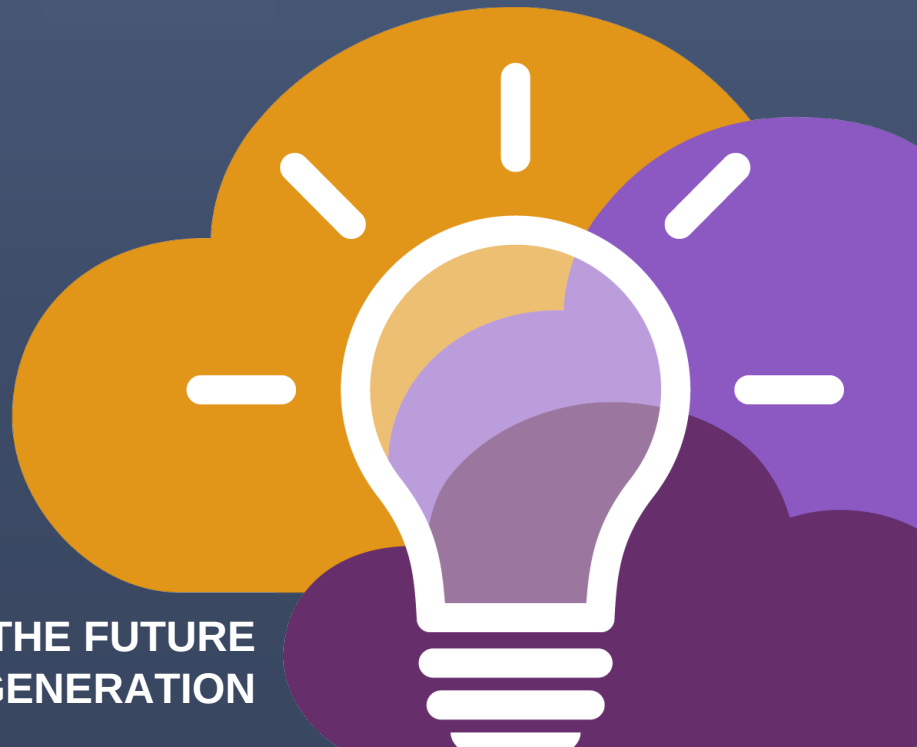
Interest poll

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 bias mitigation and 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 bias
Reduction

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

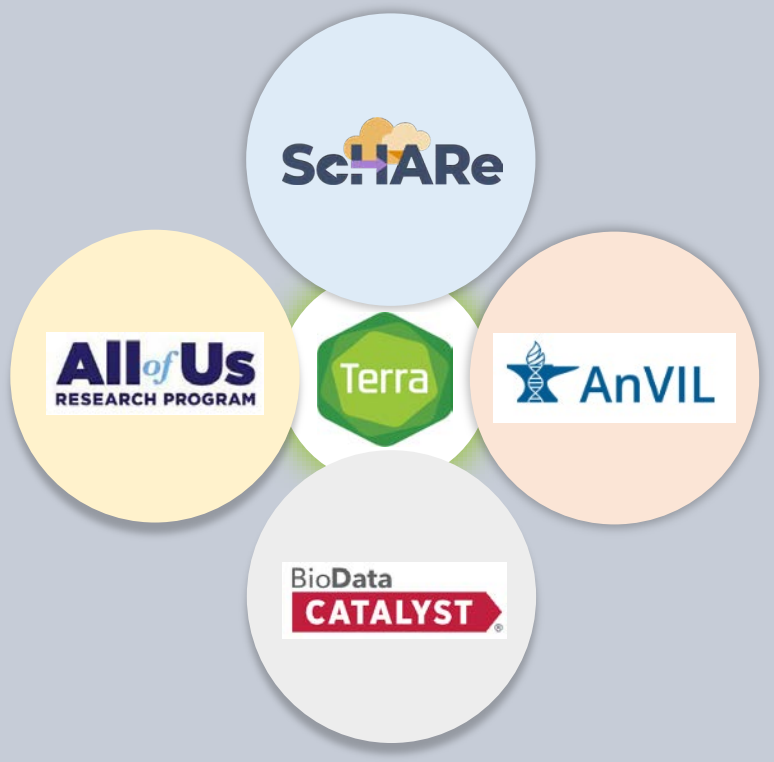
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Cut and paste code in Python and R

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

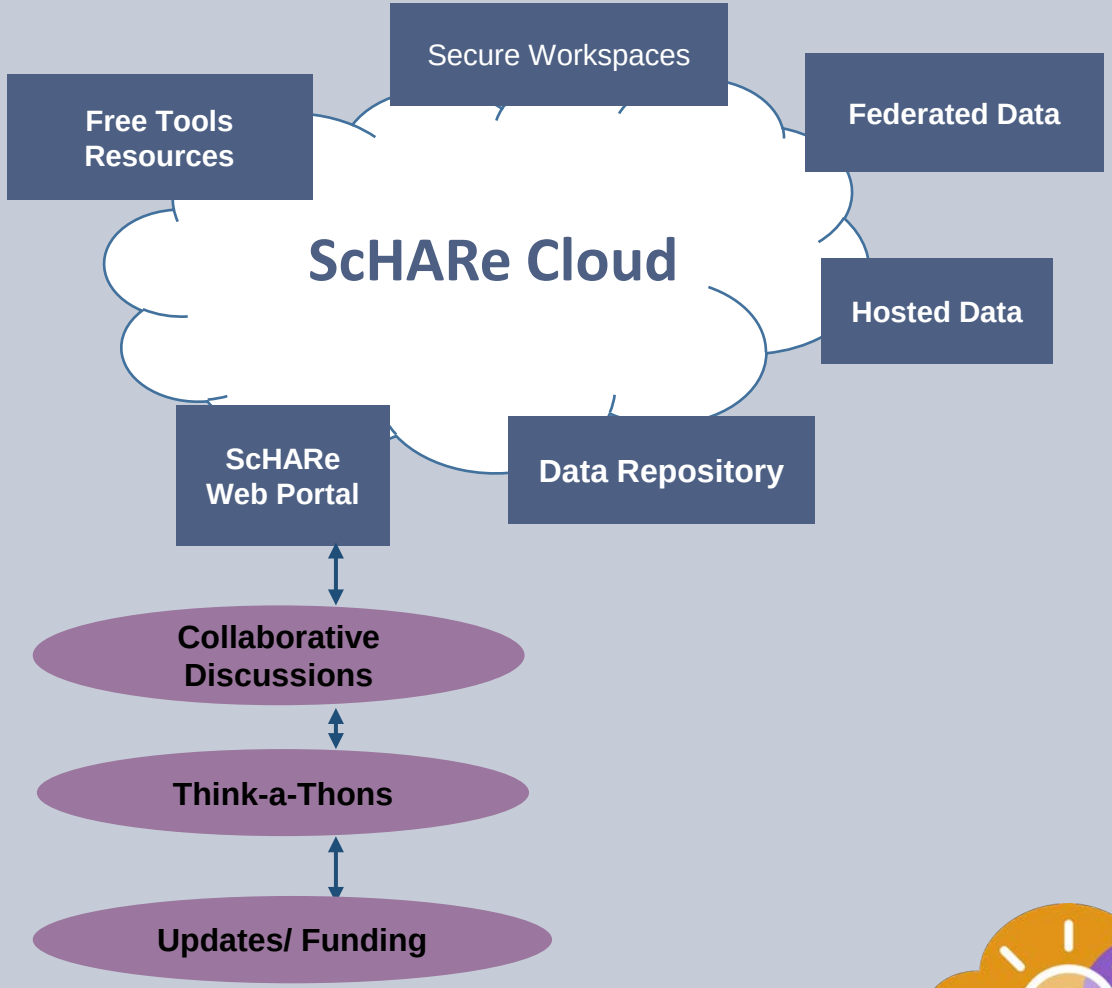
CDE FOCUSED

- **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)*

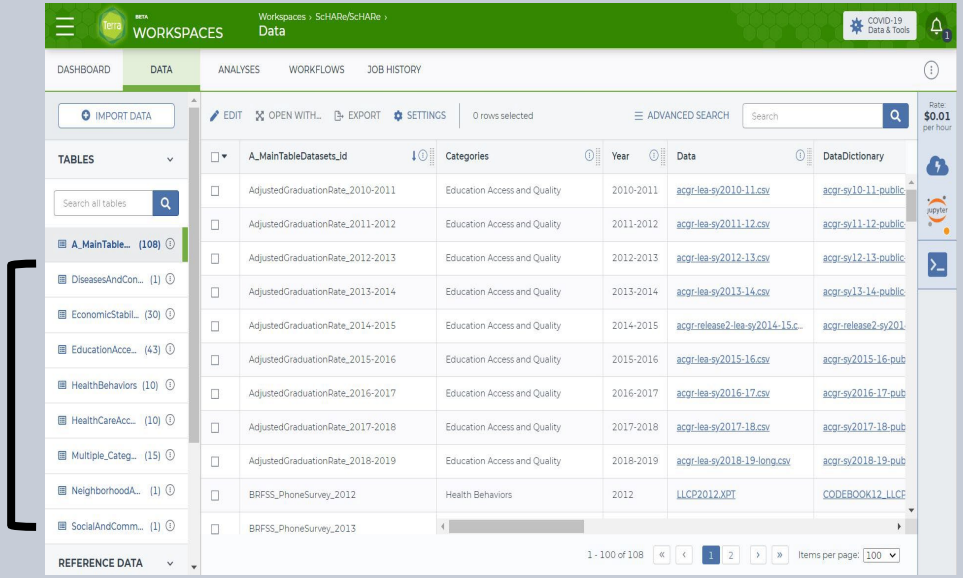
- **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe

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

- **Funded Datasets on ScHARe:** publicly accessible and controlled-access, funded program/project datasets using 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

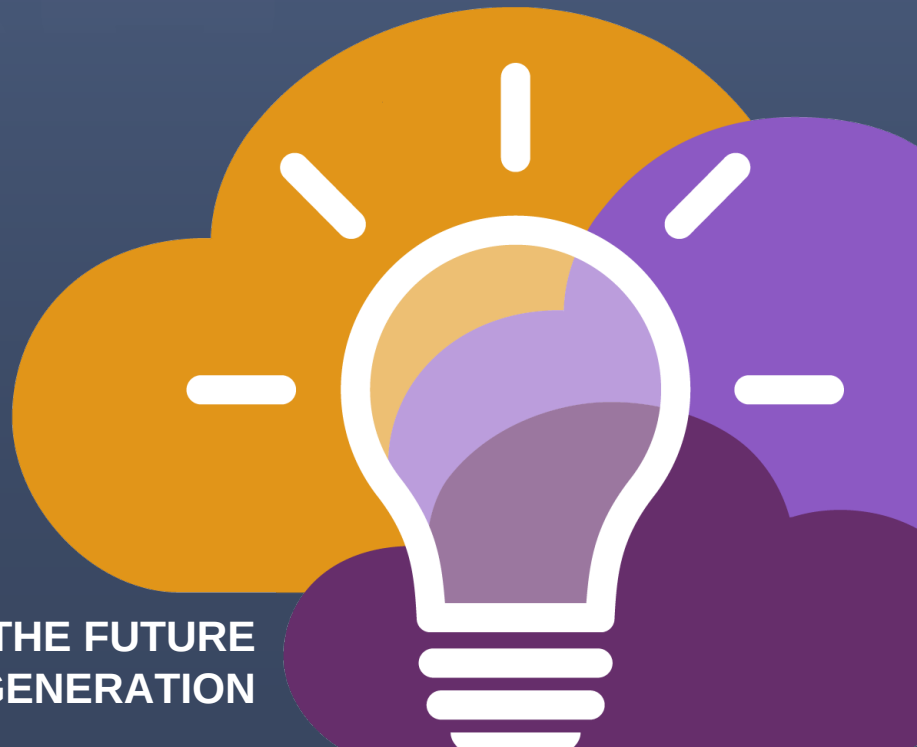
with the addition of:

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

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

ScHARe

Research Think-a-Thon Expectations



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Projects

Four topics, four teams:

- 2 Python
- 1 R
- 1 Statistics

One extra team for other topics

Team leads

Two pre-assigned co-leads:

1. Data science expert
2. Health disparities - healthcare delivery expert

Joining a team

- ✓ Co-leads will present potential research topics
- ✓ Select one of the teams based on interest and analytics used
- ✓ Join the corresponding breakout room
- ✓ At the end of the event, confirm your choice by filling out a form

Note

The teams have limited capacity. You may get reassigned to a different team

Today’s goals

1. Science co-lead will guide the discussion to hone in **research topic and questions**
2. **Datasets** to be used will be identified
3. **Variables** to focus on will be identified (if time allows)

Project expectations

- ✓ Literature review
- ✓ Dataset assessment for AI readiness (i.e., missing variables, fair representation of populations, etc.)
- ✓ Data dictionary, data sheet and dataset facts
- ✓ Decision on analytics to be used
- ✓ Document the types of biases encountered and how each was addressed
- ✓ Draft publication

Focus on health disparities/healthcare delivery…

A health disparity is a health difference that adversely affects disadvantaged **populations** in comparison to a reference population, based on one or more **health outcomes**

Health Disparity Outcomes

The health outcomes are categorized as:

- Higher incidence and/or prevalence of disease, including earlier onset or more aggressive progression of disease.
- Premature or excessive mortality from specific health conditions.
- Greater global burden of disease, such as Disability Adjusted Life Years (DALY), as measured by population health metrics.
- Poorer health behaviors and clinical outcomes related to the aforementioned.
- Worse outcomes on validated self-reported measures that reflect daily functioning or symptoms from specific conditions.

Populations with Health Disparities

Populations that experience health disparities include:

- Racial and ethnic minority groups
- People with lower socioeconomic status (SES)
- Underserved rural communities
- Sexual and gender minority (SGM) groups
- People with disabilities

...and 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**

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

If certain communities have less access to education, jobs, fresh food or healthcare, they might face **more challenges in staying healthy** or may not have the same **opportunities to make healthy choices**

Meetings and time commitment

3 months to complete the project in preparation for publication

The co-leads will **assign tasks** to the participants

Meetings other than Think-a-Thons to:

- review progress of tasks
- help/teach others what each participant is contributing
- assessing what else needs to be completed

Experience conducting ethical AI

Transparency

Public perception and understanding of how AI works

- Technical documentation for duplication/re-use
- Tools:
  - Data dictionary
  - Health sheet (Data sheet)
  - Model cards (capabilities and purpose of algorithms are openly and clearly communicated to relevant stakeholders)

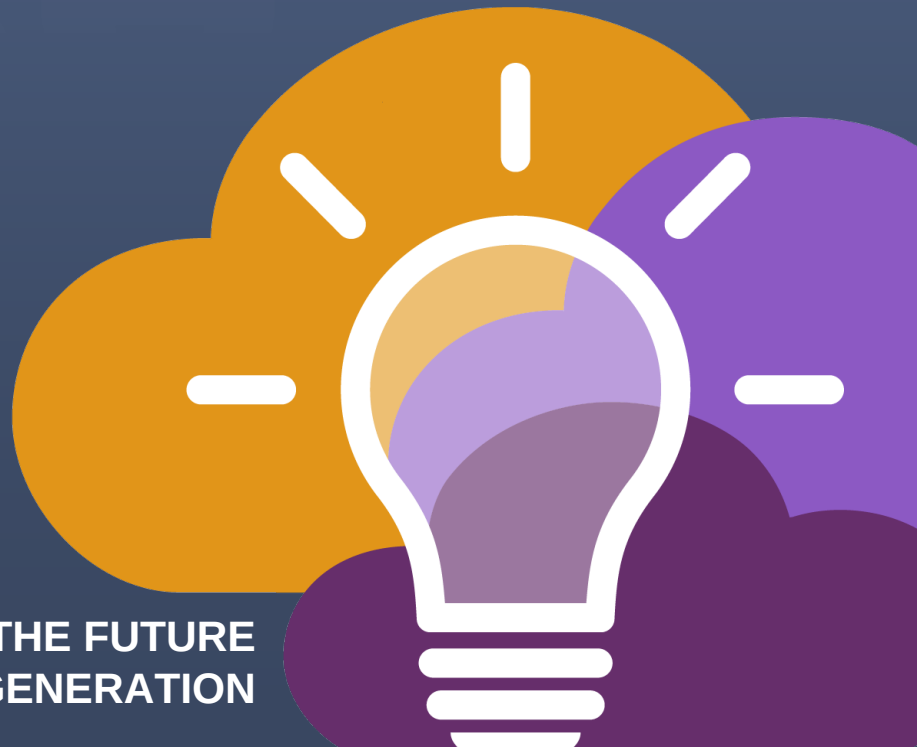
Fairness

***F*indable:** *providing metadata, documentation, and clear identifiers*
***A*ccessible:** *wide audience*
***I*nteroperable:** *standardized formats and APIs enable seamless integration*
***R*eusable:** *clear documentation, licensing, reduce redundancy*

- Metadata and data should be easy to find for both humans and computers
- Ensure that data represents relevant populations

ScHARe

Platform Orientation

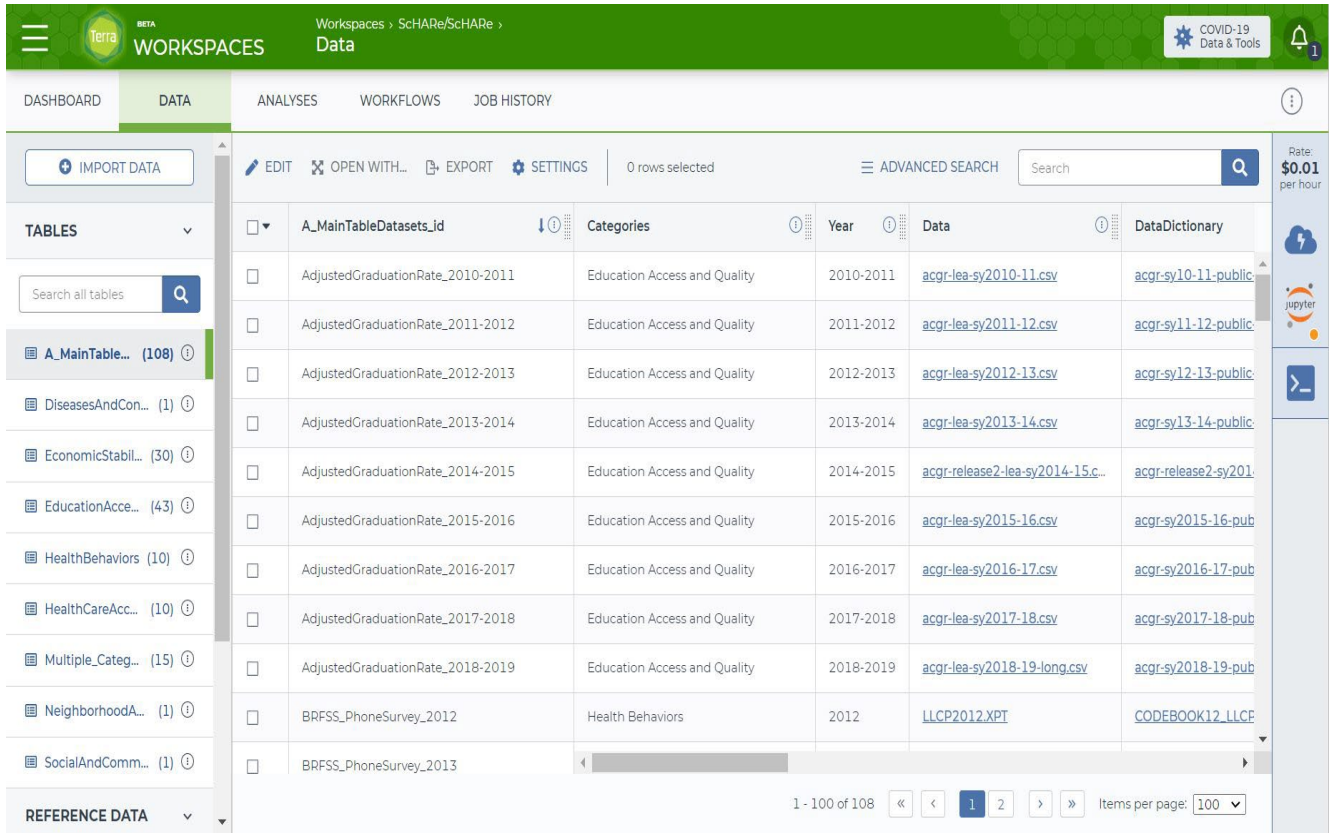


BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

SDoH-related Datasets Available on ScHARe: A Valuable Resource

ScHARe provides a valuable platform for researchers seeking **SDoH-related data**

Explore the available datasets to identify potential resources that align with your research interests in social determinants of health and their impact on various health outcomes



ScHARe Ecosystem

The ScHARe Data Ecosystem is comprised of:

1. **Google Hosted Public Datasets:** publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program
   Example: *American Community Survey (ACS)*

2. **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe
   Example: *Behavioral Risk Factor Surveillance System (BRFSS)*

3. **ScHARe Hosted Project Datasets:** 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*

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*:

240+ 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**

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.

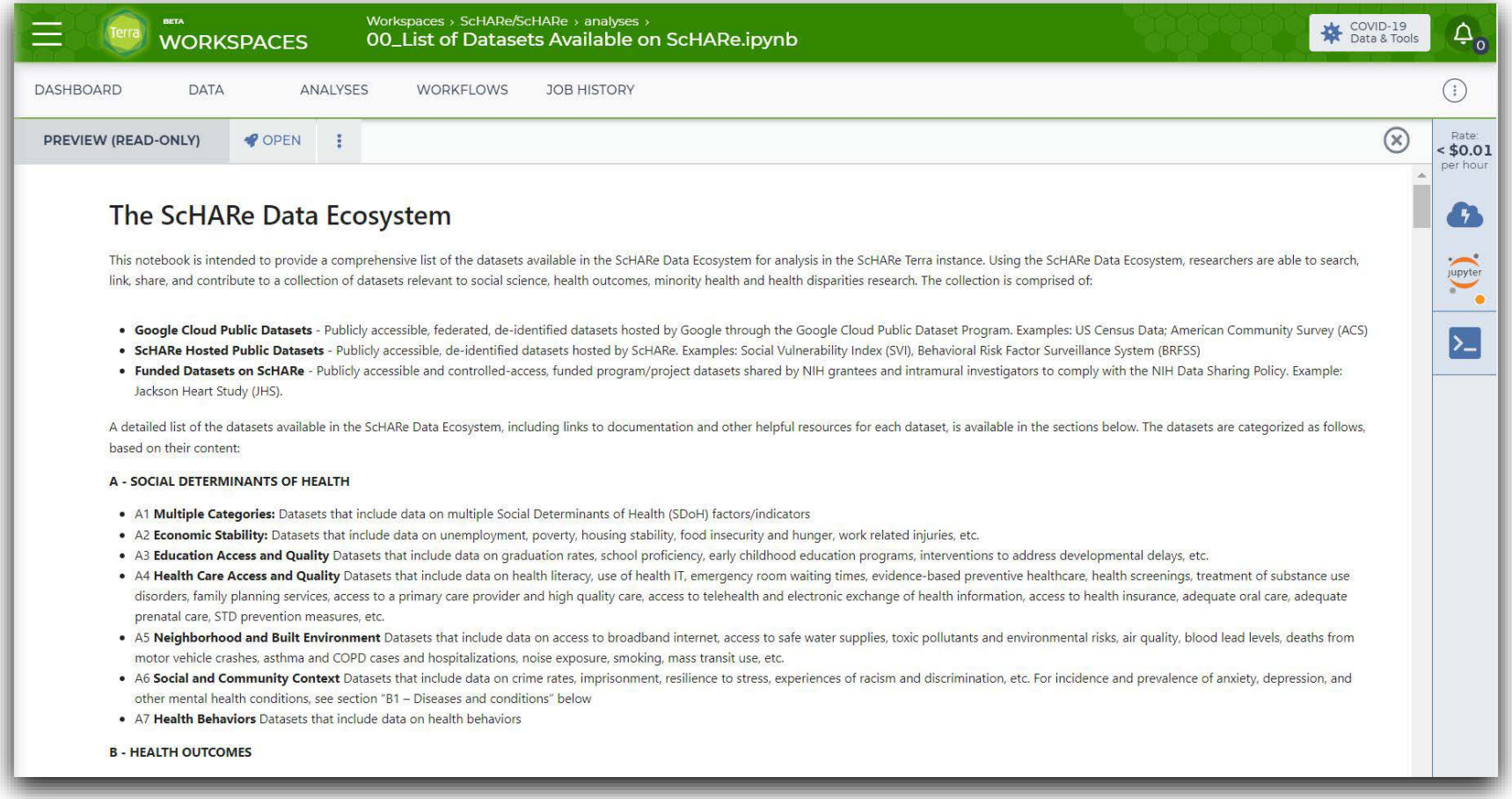
Examples:

- **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

How to check what data is available on ScHARe

Analyses tab

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



ScHARe datasets

ScHARe

The ScHARe Data Ecosystem

Last updated: November 27, 2023

This document 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-hosted 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)
- **ScHARe-hosted Project Datasets** - Publicly accessible and controlled-access, funded program/project datasets shared by NIH grantees and intramural investigators to comply with the

Jackson Heart Study (JHS)

Scan me

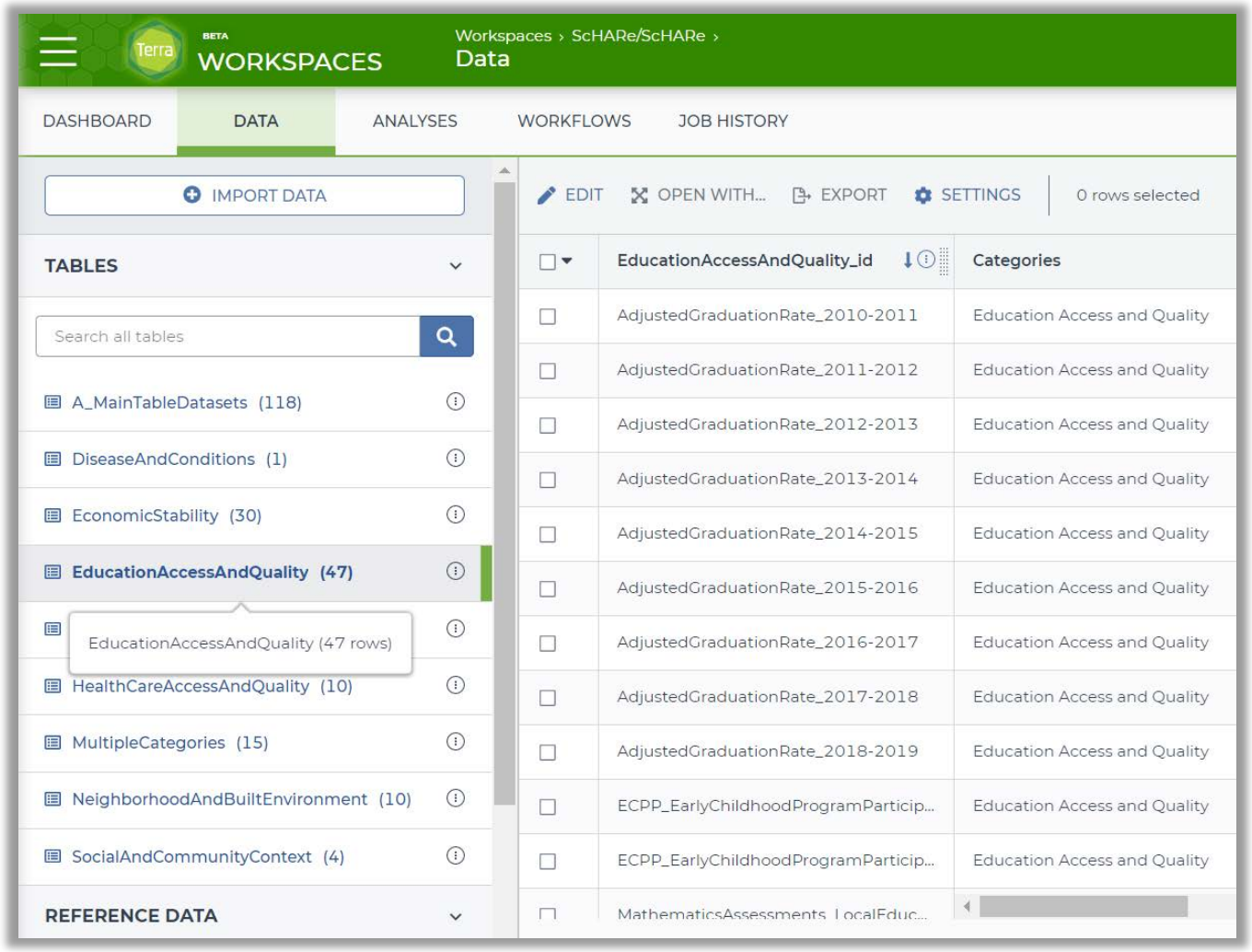
bit.ly/ScHARe-datasets

How to access available data on ScHARe

Data tab

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

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



Notebooks

A **Jupyter Notebook** is an interactive analysis tool that includes:

- **code cells** for manipulating and visualizing data in real time (Terra notebooks support **Python or R**)
- **documentation** to make it easier to share and reproduce your analysis

In past Think-a-Thons, we:

- covered the basics of **creating your first notebook**
- **explored the instructional notebooks** available in the ScHARe workspace

If you are not familiar with **programming**, the code in our notebooks is very easy to understand and reuse, and our tutorials will help you understand how notebooks work.

Why use notebooks?

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

Using notebooks:

- **is now a major part of the data science workflow** at research institutions across the globe
- can make your work **more transparent, understandable, repeatable, and shareable**
- will **speed up your workflow** and make it easier to communicate and share your results

ScHARe notebooks

Take a look at what a notebook can do by checking out the instructional notebooks that **ScHARe offers to help novice users** learn how to use the workspace and its resources

A list of the available notebooks is provided on the right.

List of ScHARe instructional notebooks

- **00_List of Datasets Available on ScHARe**: a list of the datasets available in the ScHARe Datasets collection.
- **01_Introduction to Terra Cloud Environment**: an introduction to the Terra platform and cloud environment.
- **02_Introduction to Terra Jupyter Notebooks**: an introduction to Jupyter Notebooks on the Terra platform.
- **03_R Environment setup**: instructions on how to setup your cloud environment for R-based notebooks.
- **04_Python 3 Environment setup**: instructions on how to setup your cloud environment for Python 3-based notebooks.
- **05_How to access plot and save data from public BigQuery datasets using R**: instructions on how to access, plot, and save data from datasets available on the cloud through the Google Cloud Public Datasets Program, using R.
- **06_How to access plot and save data from public BigQuery datasets using Python 3**: instructions on how to access, plot, and save data from datasets available on the cloud through the Google Cloud Public Datasets Program, using Python 3.
- **07_How to access plot and save data from ScHARe hosted datasets using Python 3:** instructions on how to access, plot, and save data from datasets hosted by ScHARe in this workspace.
- **08_How to upload access plot and save data stored locally using Python 3:** instructions on how to import to Terra, access, plot, and save data from datasets stored locally on your computer.

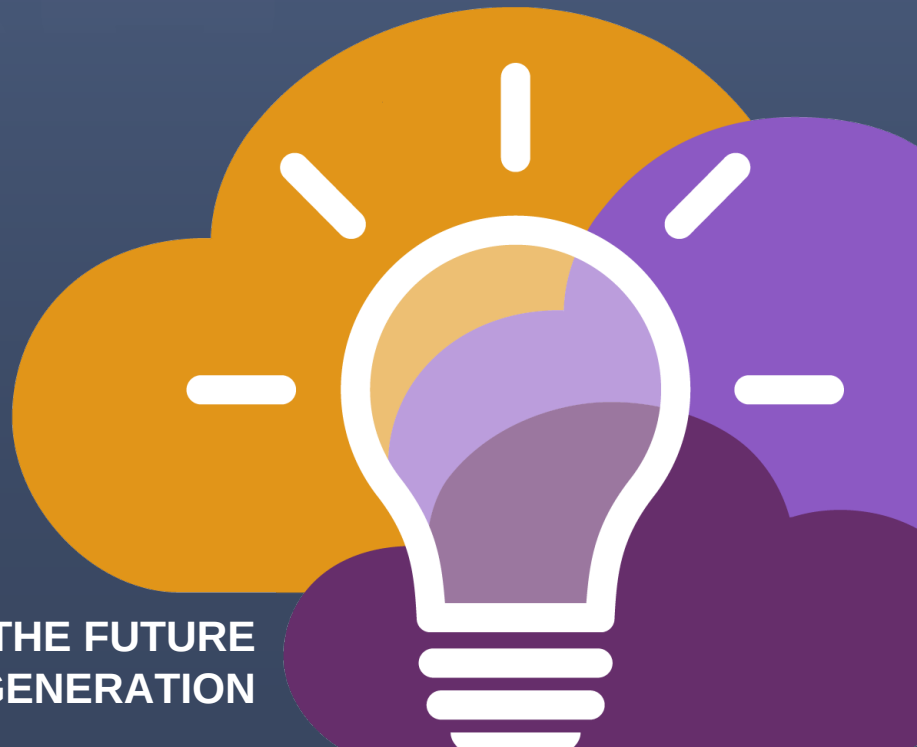
ScHARe training

bit.ly/think-a-thons

September 20, 2023	2 hours	View video: **ScHARe 5: A Review of the ScHARe Platform and Data Ecosystem** Toward Goal 1: • Create and configure an account and workspace and set the appropriate permissions • Create and run a notebook • Set up billing • Add data to a workspace • Access, visualize, and analyze data from Google-hosted federated/national datasets (e.g., the American Community Survey) and ScHARe-hosted public datasets (e.g., the Behavioral Risk Factor Surveillance System)	View slides (PDF, 3 MB)

ScHARe

Brainstorming



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Let’s brainstorm research ideas

Let’s consider:

- innovative approaches and methodologies, such as AI
- datasets publicly available on ScHARe

ScHARe

Thank you