



April 19, 2023 Think-a-Thon

ScHARe

The logo for ScHARe features the word "ScHARe" in a large, white, bold, sans-serif font. The letter "H" is stylized with a purple arrow pointing to the left and a purple arrow pointing to the right. Above the "H" are two orange and yellow clouds. The entire logo is reflected on a dark blue surface below it.

ScHARe II
Terra Datasets

Sign up for free temporary billing

If you have not filled out the 1-question form on the Think-a-Thon registration confirmation email already, please **provide a Google email address in the chat**

You will be:

- **registered for ScHARe**
 - **added to a free temporary billing project** that will allow you to run all the Think-a-Thon materials with your instructors
-
- You will be active on this billing project for 1 day after the Think-a-Thon
 - If you want to access work-in-progress from the Think-a-Thon after this time, you will need to set up your own billing and copy any of your workspaces to your own billing



Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

Sci!ARe



National Institute
on Minority Health
and Health Disparities



Office of
Data Science Strategy



National Institute
of Nursing Research

Thank you

NIMHD

Dr. Eliseo
Perez-Stable

ODSS

Dr. Susan
Gregurick

NIH/OD

Dr. Larry
Tabak

NINR

Dr. Shannon
Zenk

NIMHD OCPL

Kelli Carrington
Thoko Kachipande
Corinne Baker

NINR

Rebecca Hawes
Micheal Steele
John Grason

BioTeam

STRIDES

Terra

SIDEM

RLA

CCDE Working Group

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

Outline

5' Introduction and setup

- Experience poll

5' ScHARe and Terra overview

- Interest poll

15' Previous Think-a-Thon recap: accounts, workspaces, notebooks

5' ScHARe datasets

- Datasets poll

40' How to work with data uploaded to Terra

40' How to work with Google BigQuery data

10' Billing and costs

- Think-a-Thon poll

Experience poll

Please check your level of experience with the following:

| | None | Some | Proficient | Expert |
|-----------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Python | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| R | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Cloud computing | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Terra | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Health disparities research | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Health outcomes research | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Algorithmic bias mitigation | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |



ScHARe

Part I

ScHARe and Terra Overview

Phase I

**Population Science and SDoH Datasets
Tutorials and Resources
Think-a-Thons**

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 **three 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



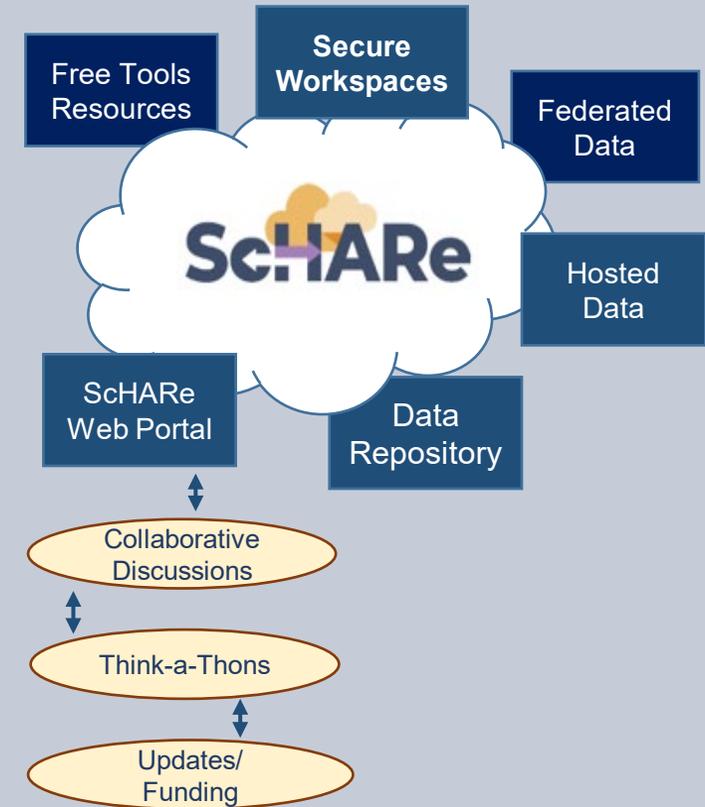
ScHARe Components

ScHARe co-localizes within the cloud:

- **Datasets** (including social determinants of health and social science data) relevant to minority health, health disparities, and health care outcomes research
- **Data repository** to comply with the required hosting, managing, and sharing of data from NIMHD- and NINR-funded research programs
- **Computational capabilities and secure, collaborative workspaces** for students and all career level researchers
- **Tools for collaboratively evaluating and mitigating biases** associated with datasets and algorithms utilized to inform healthcare and policy decisions

Frameworks: Google Platform, Terra, GitHub, NIMHD Web ScHARe Portal

Intramural & Extramural Resource



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

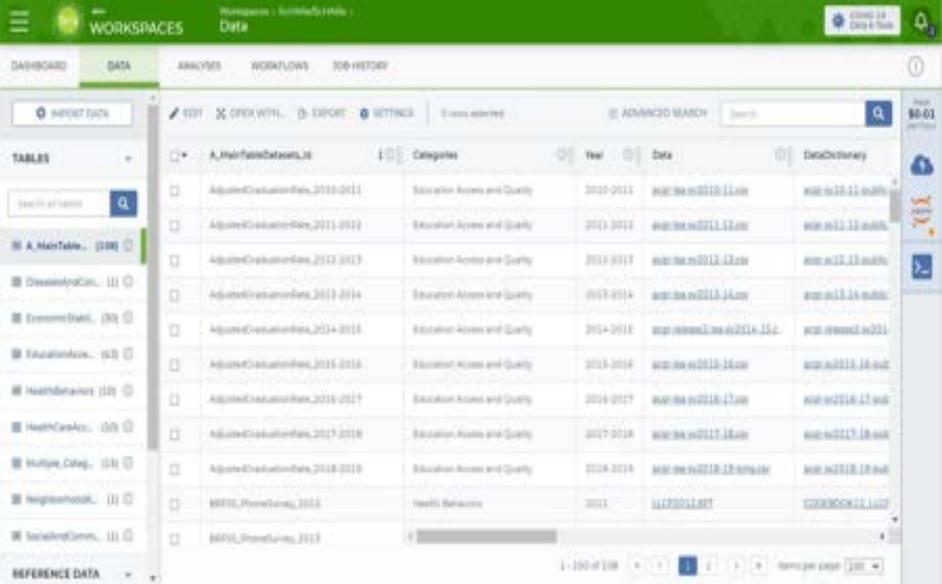
Example: *American Community Survey (ACS)*

- **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*



The screenshot displays the SchARE Data Ecosystem interface. The top navigation bar includes 'WORKSPACES', 'Data', and 'MANAGEMENT'. Below this, there are tabs for 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The main content area shows a table of datasets with columns for 'TABLES', 'A_HairNetDataset_It', 'Categories', 'Year', 'Data', and 'Dictionary'. A yellow box highlights the 'TABLES' column, which lists various datasets such as 'AdjustedCrackUseRate_2010-2011', 'AdjustedCrackUseRate_2011-2012', etc. The table also includes a search bar and a 'REFERENCE DATA' section at the bottom.

On SchARE, 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

Access to Population Science datasets



ScHARe Data Ecosystem will offer access to **300+ datasets**, including:

- Google Cloud Public Datasets
- ScHARe Hosted Public Datasets:
 - American Community Survey
 - U.S. Census
 - Social Vulnerability Index
 - Food Access Research Atlas
 - Medical Expenditure Panel Survey
 - National Environmental Public Health Tracking Network
 - Behavioral Risk Factor Surveillance System
- **Coming Soon:** Repository for Funded Datasets on ScHARe, in compliance with NIH Data Sharing Policy

Cloud computing strategies



- Uses **workflows** in Workflow Description Language (**WDL**), a language easy for humans to read, for batch processing data
- **Python and R**, including most commonly used libraries
- Enables **customization** of computing environments to ensure everyone in your group is using the same software
- **Big Query** and **Tensorflow** access for advanced machine learning
- Enables researchers to create interactive **Jupyter notebooks** (documents that contain live code) and share data, analyses and results with their collaborators in real time
- For novice users, integration with **SAS** is planned

AI bias mitigation strategies

- Widespread use of AI raises a number of ethical, moral, and legal issues – likely not to go away
- Algorithms often are “black boxes”
- **Biases can result from:**
 - social/cultural context not considered
 - design limitations
 - data missingness and quality problems
 - algorithm development and model training
 - Implementation
- If not rectified, biases may result in decisions that lead to discrimination, unequitable healthcare, and/or health disparities
- **Lack of diverse perspectives:** populations with health disparities are underrepresented in data science
- **Guidelines** and recommendations emerging from HHS, NIST, White House, etc.



Critical thinking can rectify AI biases

ScHARe was created to:

- foster participation of **populations with health disparities in data science**
- promote the collaborative identification of **bias mitigation strategies** across the continuum
- create a **culture of ethical inquiry** and critical thinking whenever AI is utilized
- build **community confidence** in implementation approaches
- focus on **implementation of AI bias** guidelines and recommendations

Phase II
(in process)

Repository and Data Ecosystem

SeHARe Data Repository

CORE COMMON DATA ELEMENTS

**NOVEL CDE FOCUSED
REPOSITORY TO FOSTER
INTEROPERABILITY**

**COMPLY WITH DATA SHARING
POLICY - HOST PROJECT DATA**

DATA ECOSYSTEM

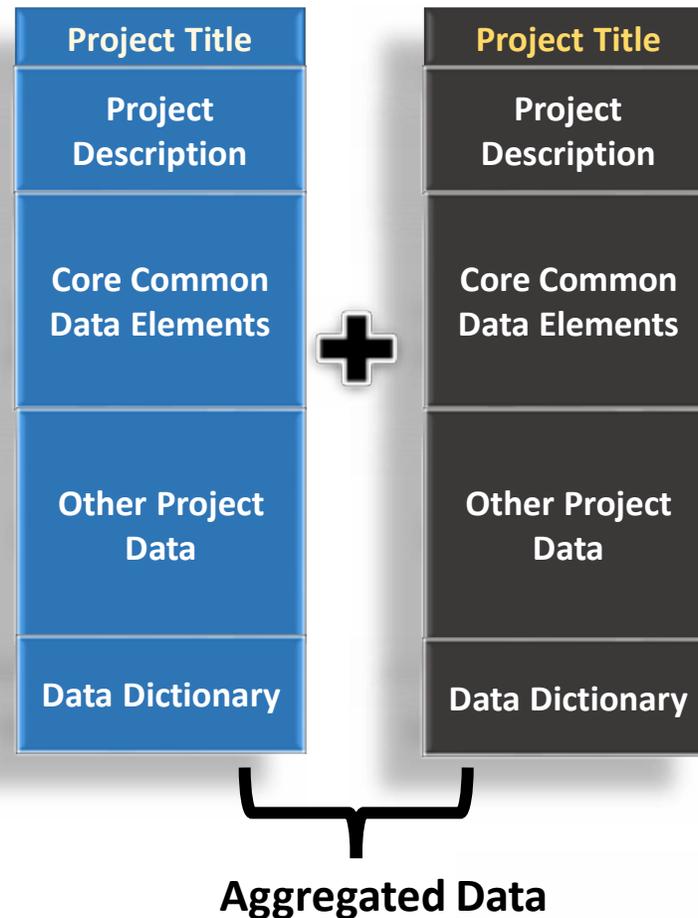
- Map across datasets
- Map across platforms



UPCOMING

Core Common Data Elements Intramural and Extramural Project Repository

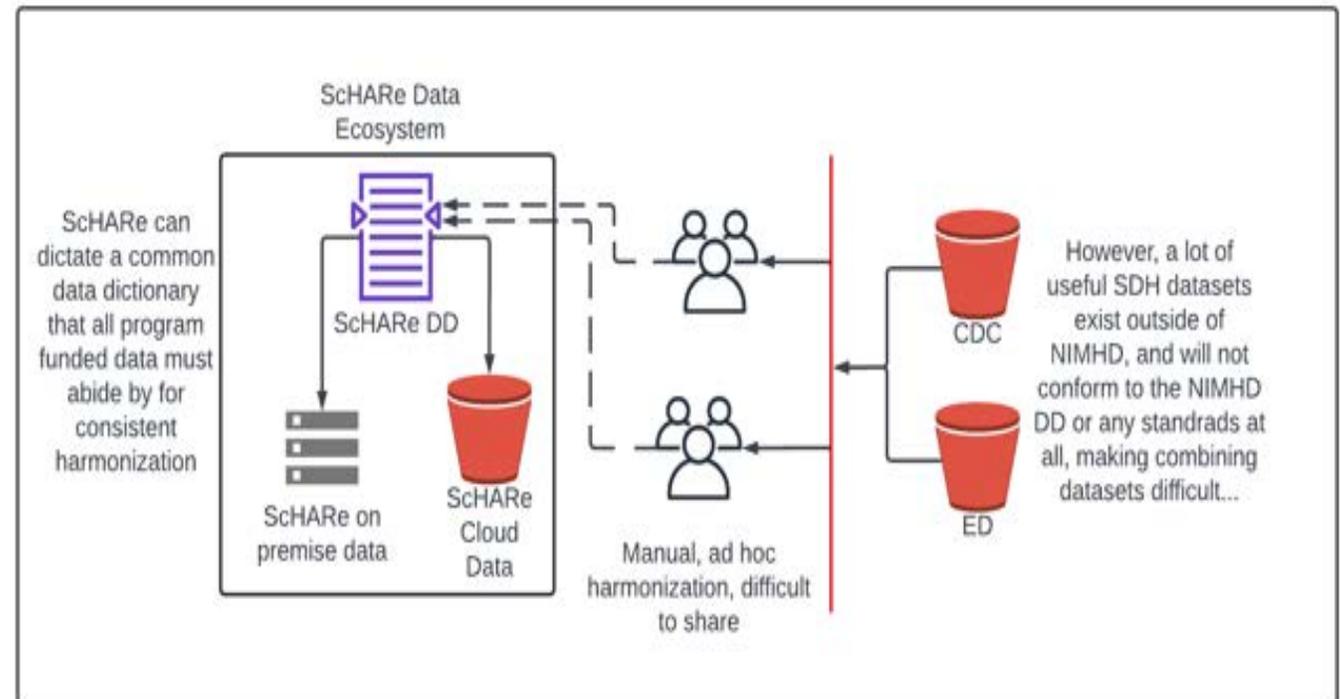
- Complies with **NIH Data Sharing Policy**
- Fosters dataset sharing and interoperability by using or mapping to **Core Common Data Elements**
- Provides resources for **intramural researchers** to work in a secure workspace and host data
- Centralizes **aggregated datasets** for repeat use



UPCOMING FALL

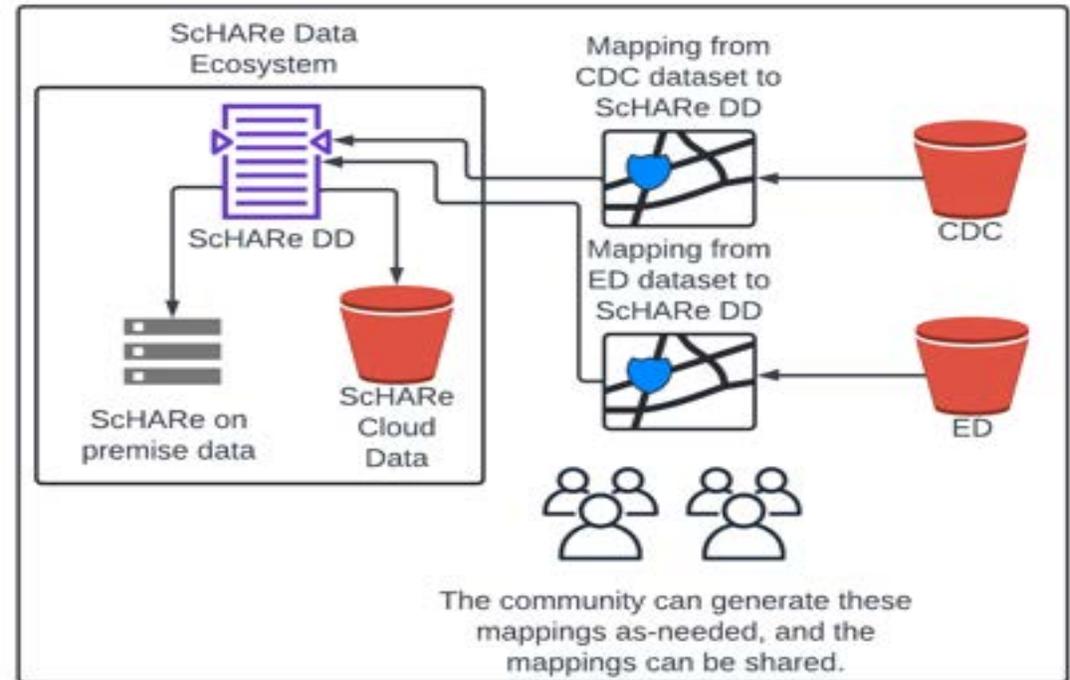
Common Data Elements project mapping

- Through significant efforts, **data can be harmonized** to a particular data dictionary
- To incorporate external data, users must harmonize in an ad-hoc manner that is **specific to a particular use case**, and requires significant manual effort
- It is **difficult to share these mappings** or re-use the ad-hoc harmonization in other projects



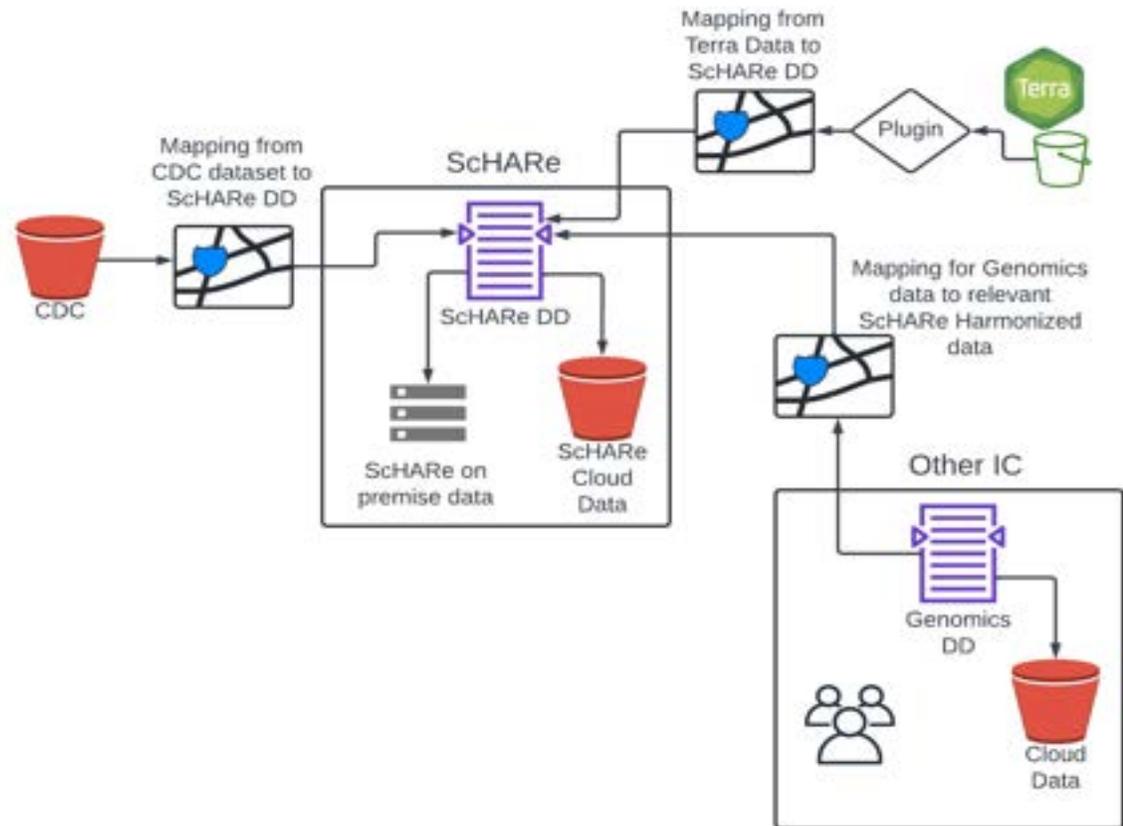
Harmonization through project & federated dataset mapping

- **Harmonization should be shared:** data standards cannot be enforced on everyone
- We are building a **mapping system** to relate datasets to data dictionaries and transform them on the fly (we call the result of a transform a **dataview**)
- Mappings can easily be **shared amongst the community:** a user who mapped CDC data to ScHARe's DD can then share it with the community



Mapping between cloud platforms

- By providing **mappings and simple connector plugins** as needed, we can accommodate new data sources and map between data dictionaries
- This provides the **flexibility to expand the ecosystem** as new standards are developed and adopted
- Current **common data elements** can be incorporated into mappings immediately



UPCOMING

The ScHARe logo features the word "ScHARe" in a dark blue, sans-serif font. The letter "H" is replaced by a stylized orange and yellow cloud. A purple arrow points from the "H" towards the right.

ScHARe

The logo for the All of Us Research Program. "All of Us" is written in a large, bold, blue font, with "of" in a smaller, lighter blue font. Below it, "RESEARCH PROGRAM" is written in a smaller, dark blue, all-caps font.

All of Us
RESEARCH PROGRAM

The Terra logo consists of a green hexagon with a white border. Inside the hexagon, the word "Terra" is written in a white, sans-serif font.

Terra

The AnVIL logo features a blue DNA double helix icon to the left of the word "AnVIL" in a blue, sans-serif font.

AnVIL

The BioData Catalyst logo. "BioData" is written in a dark grey, sans-serif font above a red arrow-shaped bar. Inside the bar, the word "CATALYST" is written in white, all-caps, sans-serif font.

BioData
CATALYST

This creates an extraordinary opportunity for **high-impact collaborations** across platforms

Learning how to use Terra on ScHARe will open up a world of possibilities, giving you access to an interdisciplinary wealth of datasets and resources

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 Think-a-Thons

- Monthly sessions (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 team collaboration
- Networking
- Mentoring and coaching

Instructional

| Order | Date/Time | Subsession number, title, description, and topics |
|-------|-----------------------------------|--|
| 1 | February 15, 2023 2:30-4:30 pm | <p>Title: Artificial Intelligence and Cloud Computing 101</p> <p>Description: An introduction to the artificial intelligence and cloud computing concepts behind the SchARE platform; fundamental terminology; cloud architecture, storage and providers; data structures and management; technology benefits and concerns in the research collaboration world.</p> |
| 1 | March 15, 2023 2:30-4:30 pm | <p>Title: SchARE 1 - Research in the Cloud Implementation Strategies</p> <p>Description: An overview of active research collaboration platforms; success stories; best practices and strategies for low-cost cloud development; grant writing 101 for cloud implementation projects; SchARE mission and vision.</p> |
| 1 | April 18, 2023 2:30-4:30 pm | <p>Title: SchARE 2 - Accounts, Workspaces, and Analysis</p> <p>Description: An inside look at SchARE's Terra instance: general introduction and features; how to create and configure an account and set up billing; how to create a workspace and set the appropriate permissions; how to clone or create and run a notebook; data and workflow analysis.</p> |

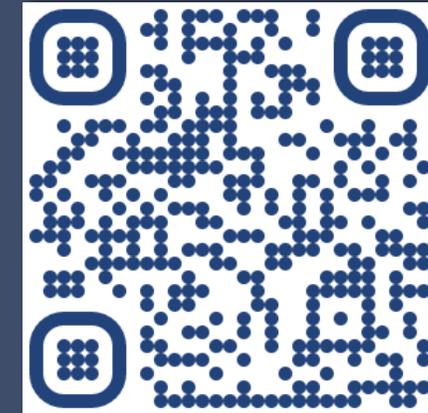
Research teams

| | | |
|---|---------------------------------|--|
| 2 | June 21, 2023 2:30-4:30 pm | <p>Title: Data Science Projects 1 - Health Disparities and Individual SDOH</p> <p>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.</p> |
| 2 | July 19, 2023 2:30-4:30 pm | <p>Title: Data Science Projects 2 - Health Disparities and Structural SDOH</p> <p>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.</p> |
| 2 | August 24, 2023 2:30-4:30 pm | <p>Title: Data Science Projects 3 - Health Outcomes</p> <p>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.</p> |

SchARE
Think-a-Thon



Register:



bit.ly/think-a-thons

The word "SCIENCE" is written in a large, white, sans-serif font. The letters "S", "C", "I", "E", "N", and "C" are in white. The letter "A" is replaced by a stylized orange and yellow cloud. A purple arrow points from the "I" to the "A", and another purple arrow points from the "A" to the "E".

SCIENCE

Part II

Previous Think-a-Thon Recap

Registering for ScHARe

Complete the following steps to register for ScHARe:

1. Visit the ScHARe portal on the NIMHD website:
nimhd.nih.gov/schare
2. Click on the “Register for ScHARe” button
3. On the registration page, click on the “Register for ScHARe on Terra” button
4. Complete the registration form

The ScHARe team will:

- review and approve your application
- send you an email with additional instructions within 3 business days

Complete slides with **step-by-step instructions and screenshots** available at: bit.ly/think-a-thons



Terra recommends using Chrome

- Note: you will need a **Gmail account** or another email account (an institutional email, for example) associated with a Google identity. If you do not have it, you can create one here:

bit.ly/3QeUngh

Creating a Terra account

Complete slides with **step-by-step instructions and screenshots** available at: bit.ly/think-a-thons

The email you will receive after ScHARe registration approval will ask you to **complete the following steps:**

1. Access the ScHARe Terra workspace at:
bit.ly/access-schare
 2. Click on the blue “Log in” button
 3. Select “Sign in with Google”
 4. Sign into Terra. Your username is the Google email address you provided to request access to ScHARe
 5. Click “Next” and enter your Google account password to login
 6. You will see a New User Registration page. Insert your name and contact email, then click on “Register”
 7. Review and accept the Terra Terms of Service
- You will be taken to the ScHARe Terra Workspace: bit.ly/access-schare

Here you can click on the tabs at the top of the page (**Dashboard, Data, Analyses**, etc.) to explore the available resources

Workspaces are the building blocks of Terra - a dedicated space where you and your collaborators can access and organize the same data and tools and run analyses together

They are like **computational sandboxes** with everything you need to complete your project: data, analysis tools, documentation

Workspaces and permissions

Complete slides with **step-by-step instructions and screenshots** available at: bit.ly/think-a-thons

Let's create your first Terra workspace.

Let's assume that you intend to create a workspace that will allow you to work with two groups of collaborators:

- **Group 1 – Internal collaborators:** researchers in your lab, who must be able to access your data, perform computations, and work with you to write the collaborative notebooks used to share results with the public
 - **Group 2 – External collaborators:** researchers at another institution, who you want to be able to see your data, notebooks and analyses, but without the possibility of modifying them
1. **Click on the menu** in the top left corner of the page, then on **“Groups”**
 2. On the Groups page, select **“Create a New Group”** and proceed to **create two different groups**, one for each of the two groups of collaborators previously identified
 3. For each group, click on the name of the group and, in the following screen, on **“Add User”**
 4. **Add the Google email address of at least one researcher** to each group. If you want one or more of your collaborators to be able to manage users and groups, check the **“Can manage users (admin)”** box

You now have two lists of collaborators with whom you can share your workspace, assigning different roles.

Sharing workspaces

Complete slides with **step-by-step instructions and screenshots** available at: bit.ly/think-a-thons

You are now ready to **share the workspace with the two groups of collaborators** you created:

1. **Click on the menu** in the top left corner of the page, then on **“Workspaces”**
2. **Identify your workspace** in the list of workspaces provided on screen and click on the corresponding **vertical three-dot menu**, then on **“Share”**
3. In the drop-down menu, select the group email corresponding to your **first group of internal collaborators**. Since you want this group to be able to access your data, perform computations, and write notebooks, select the **“Writer”** role for this group in the drop-down menu and check the **“Can compute”** box
4. Now, in the drop-down menu, select the group email corresponding to your **second group of internal collaborators**. Since you want this group to be able to see, but not modify your data, notebooks and analyses, select the **“Reader”** role for this group in the drop-down menu and do not check the **“Can compute”** box. If you also want the group to be able to share your work, check on the **“Can share”** box

Billing permissions

To allow collaborators from Group 1 to perform **computations** for which you will sustain the cost, you have to give them **permission to use your Terra Billing Project**

Refer to last month’s slides for complete instructions

Cloning workspaces

If you are interested in using the data resources of a workspace or replicating the analyses showcased in its notebooks, and have the appropriate permissions to do so, you can “clone” (create a copy of) such workspace for your personal use.

You are encouraged to clone the ScHARe workspace and use its resources. Here is how you can do it.

As an example, we will clone the workspace “ScHARe Think-a-Thons”, a ScHARe workspace copy created for this event.

1. **Click on the menu** in the top left corner of the page, then on **“Workspaces”**
2. **Identify the workspace you want to clone** in the list of workspaces provided on screen and click on the corresponding **vertical three-dot menu**, then on **“Clone”**
3. Input a **name** for the workspace copy

Complete slides with **step-by-step instructions and screenshots** available at: bit.ly/think-a-thons

4. Select the **Billing Project** you want to associate with the workspace. For this example, you can select our free temporary Billing Project “ScHARe-Temp”
5. Select the **bucket location**. A bucket location can only be set when creating a workspace. For this example, you can leave the default unmodified
6. Change the **Description** if desired
7. A cloned workspace will inherit the **Authorization Domain (AD)** groups of the original workspace. Add an additional AD if needed. More info on ADs [here](#)

Success! The cloned workspace is now listed among your workspaces. You can freely access all of its **resources**

Running and creating notebooks

Complete slides with **step-by-step instructions and screenshots** available at: bit.ly/think-a-thons

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

Let's cover the basics of **creating your first notebook to work with your data**:

1. **Click on the menu** in the top left corner of the page, then on **“Workspaces”**
2. **Click on the new workspace** you created earlier
3. Click on the **“Analyses”** tab
4. Click on the **“Start”** button
5. Select **“Jupyter”**
6. In the next window, **name** to the notebook and choose a language (**“Python 3”**)
7. Click **“Create analysis”**
8. You will now be asked to configure your **Cloud Environment** (the on-demand availability of data storage and computing power needed to perform your computations). You can leave the default values unchanged

Success! Your notebook has been created. **Click on its name** to open it. **Open and run** any **ScHARe instructional notebook** to get a closer look at **how notebooks work**

The logo for ScHARe features the text 'ScHARe' in a white, bold, sans-serif font. The letters 'H' and 'A' are partially obscured by a stylized orange and yellow cloud. A purple arrow points from the 'H' to the 'A', and another purple arrow points from the 'A' to the 'R'.

ScHARe

Part III

ScHARe Datasets

What data can you work with on ScHARe?

Data already in the ScHARe Data Ecosystem:

1. Google Cloud Public Datasets
2. ScHARe Hosted Public Datasets
3. Funded Datasets on ScHARe

Data you upload

to your workspace
(personal or shared with
your collaborators)

ScHARe Ecosystem data

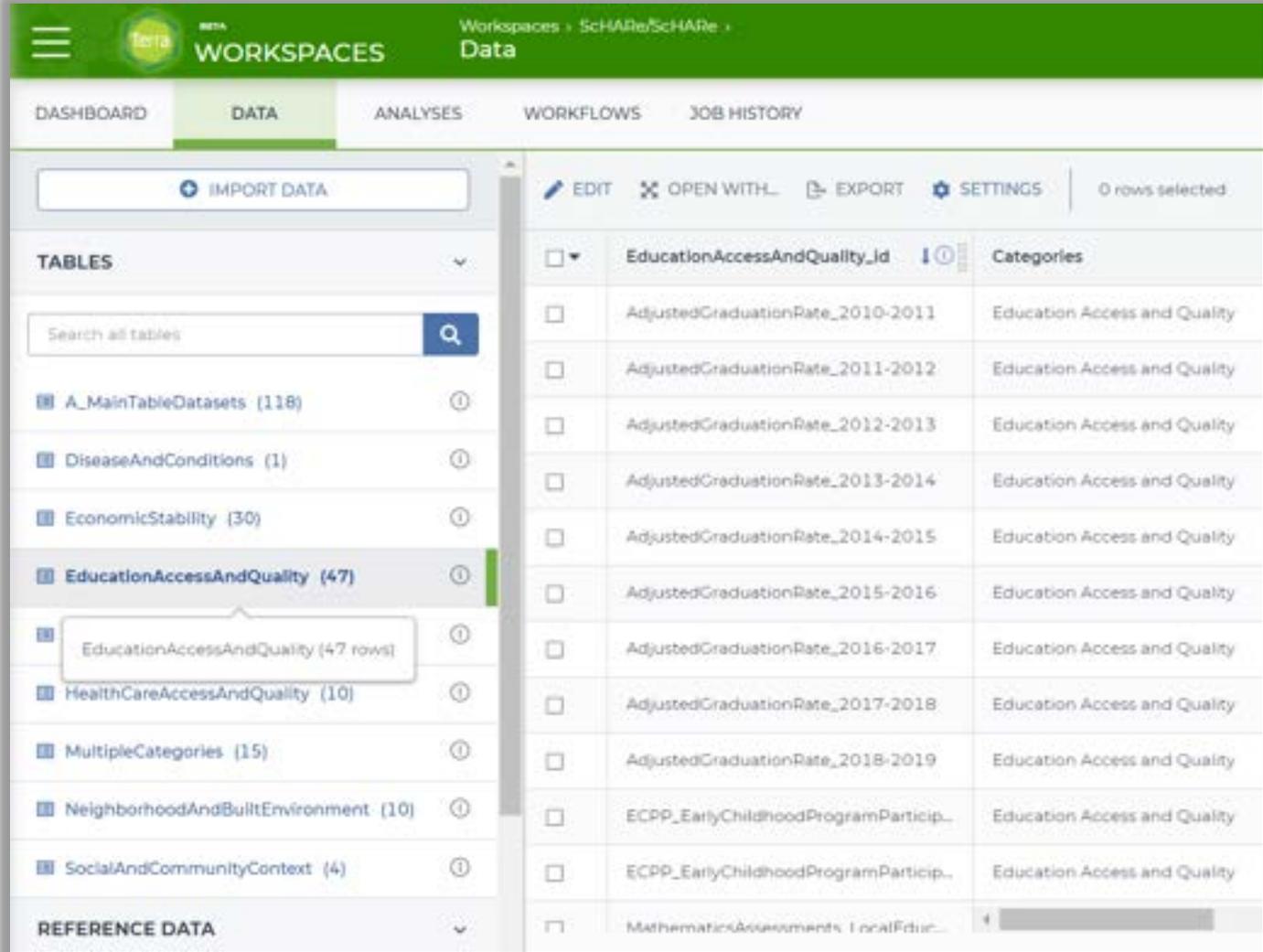
The ScHARe Data Ecosystem is comprised of:

- **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*

Exploring available data

In the **Data** tab, 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

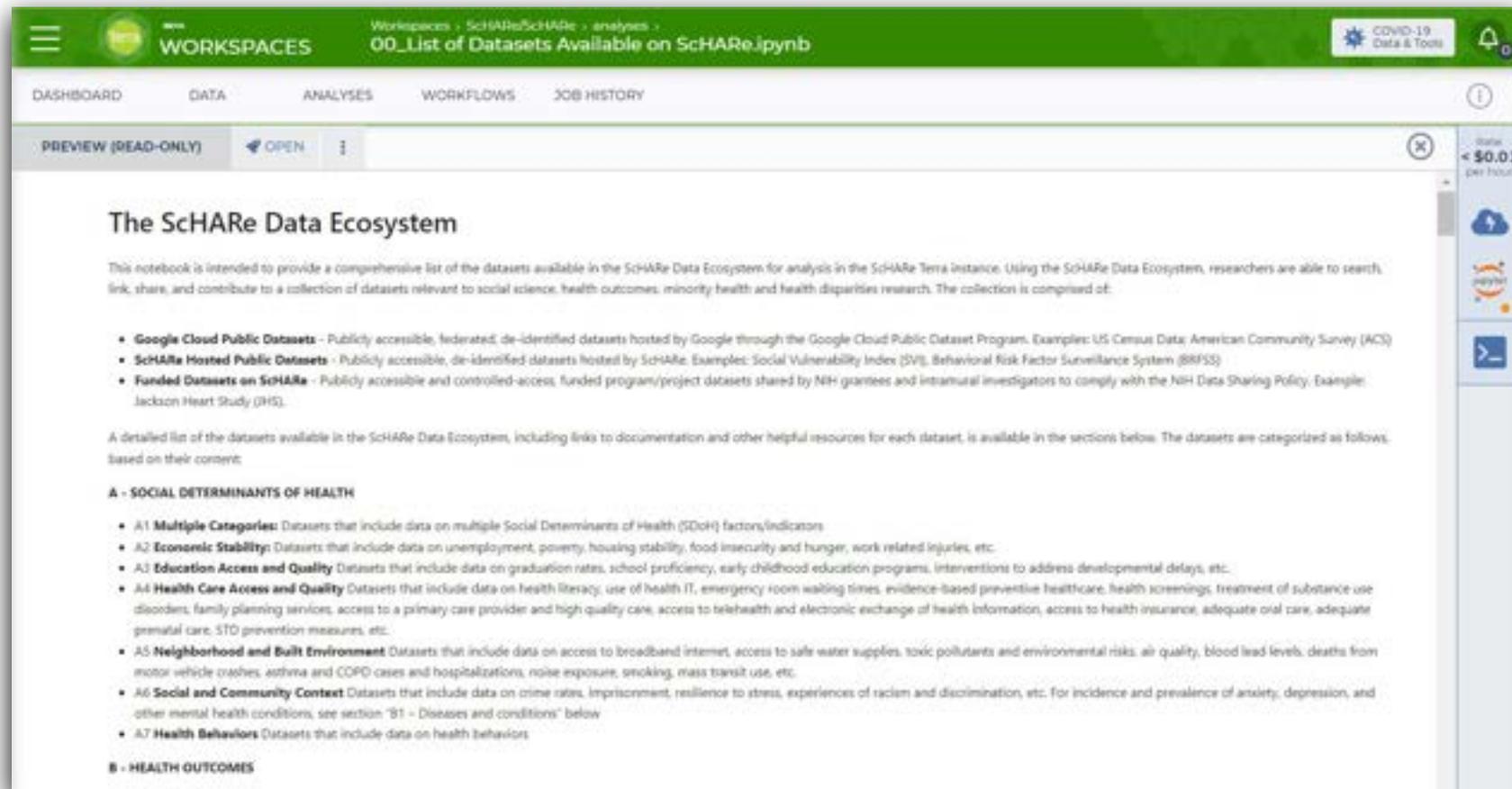


The screenshot displays the Terta WORKSPACES interface, specifically the 'Data' tab. The top navigation bar includes 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. Below the navigation bar, there is an 'IMPORT DATA' button and a search bar for tables. The 'TABLES' section lists various dataset categories, with 'EducationAccessAndQuality (47)' highlighted. A tooltip indicates that this category contains 47 rows. The main table shows a list of datasets with columns for 'EducationAccessAndQuality_Id' and 'Categories'. The 'Categories' column lists 'Education Access and Quality' for all rows. The table also includes an 'EDIT' button, 'OPEN WITH...', 'EXPORT', and 'SETTINGS' options, and a status bar indicating '0 rows selected'.

| 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... | |

Exploring available data

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



The screenshot shows the SchARE Data Ecosystem notebook interface. The top navigation bar includes 'WORKSPACES' and '00_List of Datasets Available on SchARE.ipynb'. The main content area is titled 'The SchARE Data Ecosystem' and contains the following text and list:

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

Datasets poll

1. What other datasets would you like to see?
2. What kind of research questions do you think these datasets will enable for health disparities research?

What is a notebook?

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

To get the most out of the next tutorials you should be familiar with **programming**. If you are not, the code in our notebooks is very easy to understand and reuse, and our tutorials will still help you learn how to work with data

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

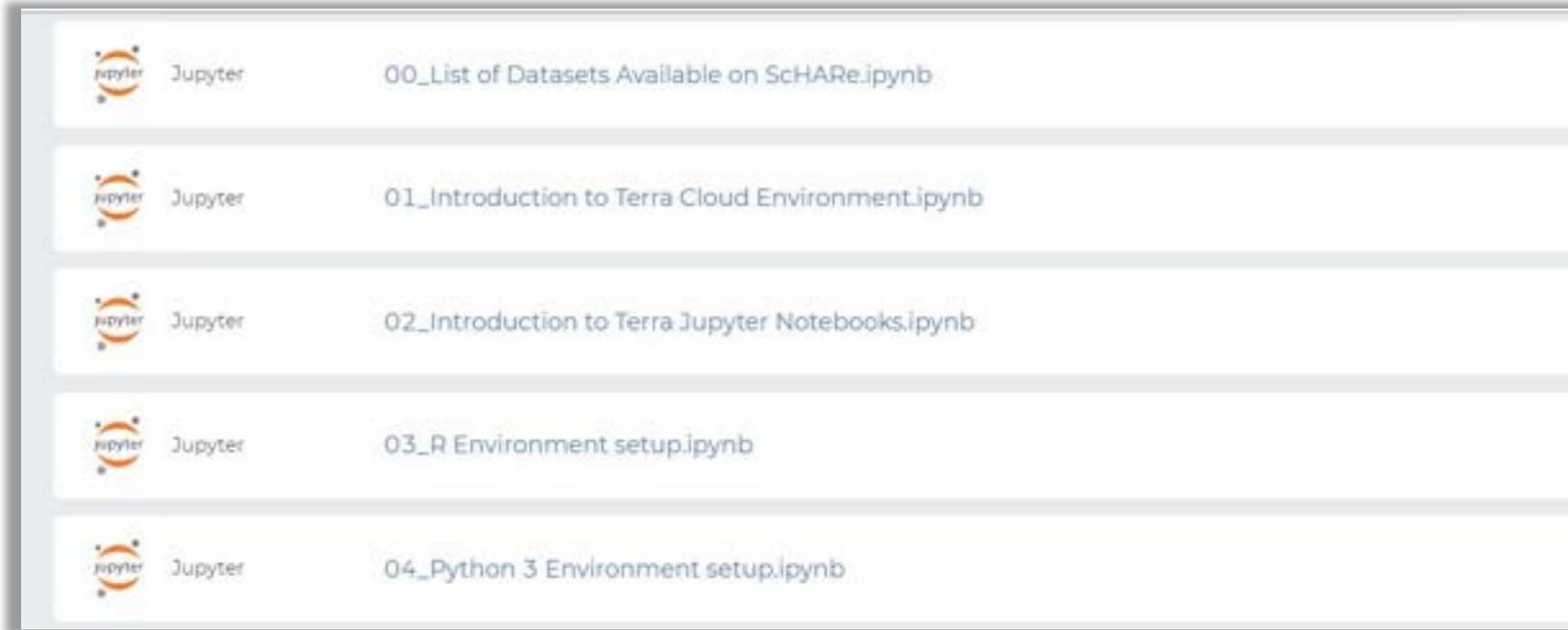
Using notebooks:

- **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

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 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 R:** instructions on how to import to Terra, access, plot, and save data from datasets stored locally on your computer.
- **09_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.

How to access, plot, analyze data and save results



| | | |
|---|---------|--|
|  | 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 |

Besides describing the datasets available on SchARE, our notebooks in the **Analyses** tab also explain how to configure the cloud computing environment and how to **access, plot, and analyze data** that you upload to your workspace or from datasets in the SchARE Data Ecosystem

We will now demonstrate how you can work with data by **using the instructions in our notebooks**

The logo for SCIARe features the word in a white, bold, sans-serif font. The letters 'C' and 'A' are stylized with a purple arrow pointing left and a purple arrow pointing right, respectively. Above the letters 'C' and 'A' is a stylized orange and yellow cloud. The entire logo is set against a dark blue background and has a faint, semi-transparent reflection below it.

SCIARe

Part IV

How to Work with Data Uploaded to Terra

How to work with data uploaded to Terra using Python

Please use the following notebook, accessible from the **Analyses** tab in the ScHARe Think-a-Thons workspace, as a reference as you complete the steps illustrated by the instructors in this tutorial:



Jupyter

09_How to upload access plot and save data stored locally using Python 3.ipynb

This tutorial is a guided introduction to analyzing data **stored on your computer and uploaded to your Terra workspace**

- We will use the **Python** programming language to work with the data
- Notebook 8 in the **Analyses** section of the ScHARe workspace explains how to use **R** instead



Jupyter

08_How to upload access plot and save data stored locally using R.ipynb

What is Python?

Python is a **computer programming language** used in data science to:

- manipulate and analyze data and conduct statistical calculations
- create data visualizations
- build machine learning algorithms

Python's **data science libraries** are powerful. Examples include:

- **Numpy** - for linear algebra and high-level mathematical functions
- **Pandas** - for handling data structures and manipulating tables
- **SciPy** - for data science tasks like interpolation and signal processing
- **Scikit-learn** - a machine learning library that is useful for classification, regression, and clustering algorithms
- **PyBrain** - for machine learning tasks and to test and compare algorithms

According to SlashData:

- there are 8.2 million Python users
- **69%** of machine learning developers and data scientists **use Python (vs. 24% of them using R)**

Sources

www.quanthub.com/python-for-data-science/
[coursera.org](https://www.coursera.org)

How can you learn Python?

Can you learn Python with no experience?

Python is the **perfect** programming language **for people without any coding experience**, as it has a simple syntax, which makes it very accessible to beginners

How long does it take to learn Python?

It can take **2 to 5 months**, but you can write your first short program in **minutes**

Online resources

You can take advantage of the dozens of “**Python for data science**” **online tutorials** for beginners and advanced programmers listed here:

- [Stackify - 30+ Tutorials to Learn Python](#)
- [FreeCodeCamp - Code Class for Beginners](#)
- [Harvard – Free Python Course](#)
- [Coursera – Free and Paid Python Courses](#)
- [LearnPython – Free Interactive Python Tutorials](#)
- [BestColleges – 10 Places to Learn Python for Free](#)

What is R?

- R is a programming language for statistical computing and graphics
- It is used by data miners, bioinformaticians and statisticians for data analysis
- Users have created **packages** to augment its functions
- Third-party **graphical user interfaces** are also available, such as **RStudio**

Sources

stackify.com/learn-python-tutorials/
en.wikipedia.org

The MHSVI data we will work with

- We will use a TSV file stored on ScHARe as a proxy for data stored on your PC
- The TSV file (a tab-separated values file commonly used to exchange data between databases) contains example data from the Minority Health Social Vulnerability Index dataset

Commonly used file formats in Data Science:

CSV, TSV, XLSX, ZIP, TXT, JSON, HTML, PDF

Python code to read them: [here](#)

What is the Minority Health Social Vulnerability Index (MHSVI)?

- MHSVI is a 2021 extension by the Office of Minority Health (OMH) of the original Social Vulnerability Index (SVI) launched by the CDC in 2011
- The dataset uses U.S. Census data to help plan **support for communities in public health emergencies**
- It combines the 15 original SVI **social factors** with additional factors known to be associated with COVID-19 outcomes

The factors are organized into **six themes**:

- Socioeconomic Status
- Household Composition and Disability
- Minority Status and Language
- Housing Type and Transportation
- Health Care Infrastructure and Access
- Medical Vulnerability

The logo features the word 'SCHEMA' in a bold, white, sans-serif font. The letters 'S', 'C', 'H', 'A', 'R', and 'E' are white. The letter 'M' is replaced by a stylized orange and yellow cloud. A purple arrow points from the 'M' towards the 'A', and another purple arrow points from the 'C' towards the 'M'.

SCHEMA

Part V

How to Work with Google BigQuery Data

What is the Google Cloud Public Datasets program?

The Google Cloud public datasets are **datasets that Google hosts** for researchers to access using the Cloud

- Google pays for the storage and public access of these datasets
- **Users pay only for the queries** they perform on the data

The Google public datasets are **available for query in BigQuery using Terra**

What is BigQuery?

- BigQuery is the Google Cloud **storage solution for structured data** (like a spreadsheet optimized for quick retrieval of particular sections that you access with a "query")
- It is easy to use, works with large amounts of data and offers **fast data retrieval and analysis**
- Many datasets, including the *Area Deprivation Index (ADI)*, are stored in BigQuery

Sources

cloud.google.com/bigquery
en.ryte.com/wiki/BigQuery

The ADI data we will work with

- We will access and use data from the **Area Deprivation Index** dataset

What is the Area Deprivation Index (ADI)?

- The Area Deprivation Index can show where **areas of deprivation and affluence** exist within a community
- It is calculated with **17 indicators from the U.S. Census American Community Survey (ACS)**, which encompass income, education, employment, and housing conditions at the Census Block level
- The ADI is available on BigQuery for release years 2018-2020 and is **reported as a percentile that is 0-100%**, with 50% indicating a "middle of the nation" percentile

- A **low ADI score** indicates affluence or prosperity
- A **high ADI score** is indicative of high levels of deprivation, which have been **linked to health outcomes**, such as 30-day hospital readmission rates, cardiovascular disease and cancer deaths
- **Neighborhood and racial disparities occur when some neighborhoods have high ADI scores** and others have low scores

How to work with Google Big Query data using Python

Please use the following notebook, accessible from the **Analyses** tab in the ScHARe Think-a-Thons workspace, as a reference as you complete the steps illustrated by the instructors in this tutorial:



Jupyter

06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

This tutorial is a guided introduction to analyzing data **stored by Google** as part of the Google Cloud Public Datasets Program **and accessible from your Terra workspace**

- We will use the **Python** programming language to work with the data
- Notebook 8 in the **Analyses** section of the ScHARe workspace explains how to use **R** instead



Jupyter

05_How to access plot and save data from public BigQuery datasets using R.ipynb



SCIARe

Part VI

Billing and Costs

What are the cloud costs of working on Terra?

The Terra platform infrastructure is **free to use**

However, the following operations in Terra **may incur charges**:

1. Virtual Machine compute costs

In cloud computing, a **virtual machine** is an emulation of a computer system that provides the functionality of a physical computer

Terra allows you to **customize** the characteristics of your virtual machine based on your computation needs (more on this later)

- A **high-performance machine costs more**
- You will be charged for the **time you use the machine**

The image shows a screenshot of the Google Cloud Platform (GCP) console's VM configuration page. The 'Cloud compute profile' section is visible, showing 'CPUs' set to 1 and 'Memory (GB)' set to 3.75. There is an unchecked checkbox for 'Enable GPUs' with a 'BETA' label and a link to learn more. The 'Compute type' is set to 'Standard VM'. The 'Enable autopause' checkbox is checked, with a '15' minute inactivity timer. The 'Location' is set to 'us-central1 (Iowa) (default)'. The 'Persistent disk' section shows 'Disk Type' as 'Standard' and 'Disk Size (GB)' as 10.

| Setting | Value |
|------------------|--|
| CPUs | 1 |
| Memory (GB) | 3.75 |
| Enable GPUs | <input type="checkbox"/> BETA |
| Compute type | Standard VM |
| Enable autopause | <input checked="" type="checkbox"/> 15 minutes of inactivity |
| Location | us-central1 (Iowa) (default) |
| Disk Type | Standard |
| Disk Size (GB) | 10 |

What are the cloud costs of working on Terra?

2. Data storage

- You will be charged for any data stored in the storage spaces (“**buckets**”) associated with your account

3. Data egress (i.e. moving data) costs

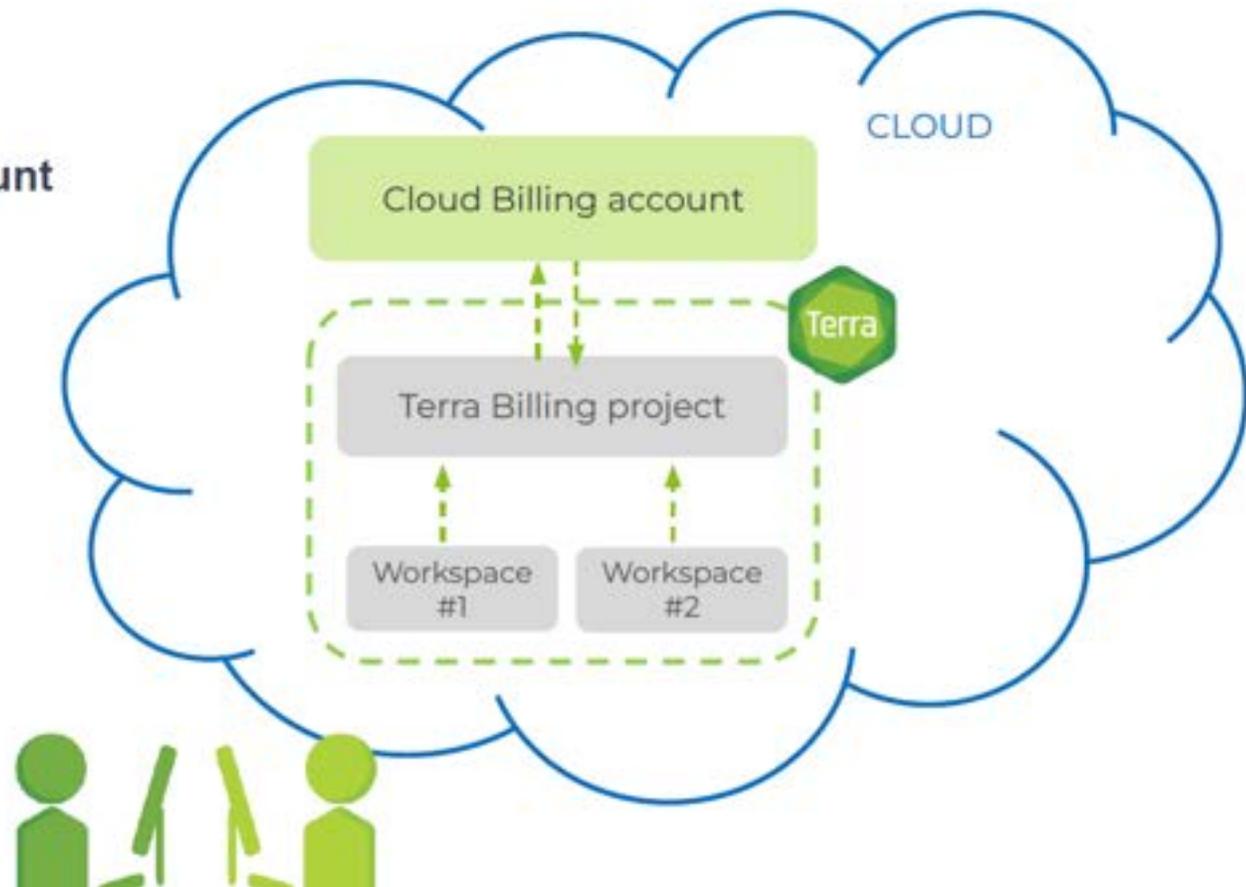
- When creating a bucket to store data, you are asked to set its location. This is because the data are going to be stored in data warehouses located in physical places (“**regions**” – more info [here](#)). Regions exist, among other reasons, to accommodate the need of certain users to keep their data in defined regions.

You will pay to **move stored data between regions**

How will I be charged for these costs?

Terra runs on Google Cloud Platform (GCP). All Terra costs are GCP fees that are ultimately paid for by a **Google Cloud Billing account** linked to Terra – specifically, to a **Terra Billing project**

- ▶ Each Billing project is linked to an umbrella Google **Cloud Billing account**
- ▶ A **Terra Billing project** is a pass-through assigned to a workspace when you create it
- ▶ All GCP fees (storage, compute, egress) are charged **per workspace** - *regardless of who does the analysis or whether they have access to a billing project.*



How will I be charged for these costs?

Will I incur any costs today?

Today and for one day after the Think-a-Thon, **access to a free temporary billing project** will allow you to run all the materials with your instructors

What happens after tomorrow?

You will no longer have access to the free temporary billing project. If you want to access work-in-progress from the Think-a-Thon, you will need to **set up your own billing** and copy any of your workspaces to your own billing

Next, we will show you how to set up your own billing

Get \$300 in free Google Cloud credits

If you've never used Google Cloud before, **you are eligible for \$300 in free Google Cloud credits** you can use for working in Terra

Conditions for Google Cloud credits eligibility

- You haven't previously signed up for the Free Trial
- You've never been a paying customer of Google Cloud, Google Maps Platform, or Firebase
- If you're part of an organization that uses Google Cloud, your email will likely not be eligible



What can I do with my credits in Terra?

The credits will cover anything that has a cost in Terra - such as storing data and running analyses. You can't use credits to add GPUs to your computing resources, and you are limited to 4 workspaces at a time

How long will my \$300 credits be available?

Your credits will be available for 3 months, or until you have used up all \$300. Once your credits run out or expire, you can upgrade to a paid account

3 easy steps to set up billing

1. Sign in to the [Google Cloud Console](#) with your Terra user ID and **set up a Google Cloud Billing account**

You'll be invited to activate your free trial: **you won't be billed until the credits expire**

2. In the [Google Cloud Console Billing page](#), **link your Google Cloud Billing and Terra accounts**

Add terra-billing@terra.bio as a Principal, with Billing Account User role

Use the same Google ID for both the Cloud Billing account and your Terra user name

3. In the [Terra Billing page](#), **create a Terra Billing project**

Select the previously created Google Cloud Billing account to fund it

For detailed instructions, see [this Terra page](#)

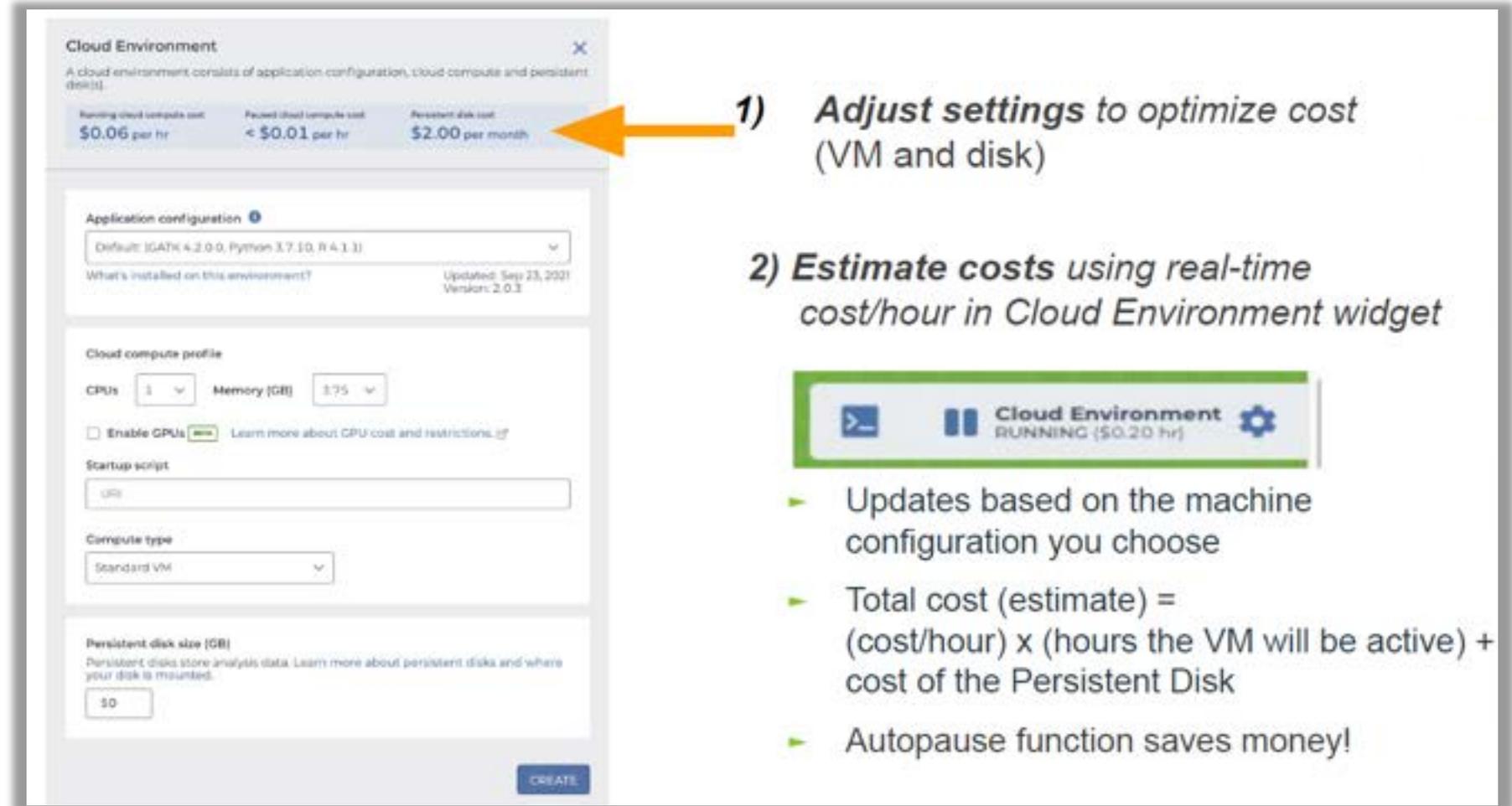
Understanding and monitoring costs

You can **ESTIMATE COSTS**:

1. **analysis costs**
2. cloud storage costs
3. egress (i.e., data moving) costs

You can **CHECK ACTUAL COSTS** in the Google Cloud Platform Console

You can **REDUCE COSTS** in several ways (for advanced users)



1) Adjust settings to optimize cost (VM and disk)

2) Estimate costs using real-time cost/hour in Cloud Environment widget

- ▶ Updates based on the machine configuration you choose
- ▶ Total cost (estimate) = (cost/hour) x (hours the VM will be active) + cost of the Persistent Disk
- ▶ Autopause function saves money!

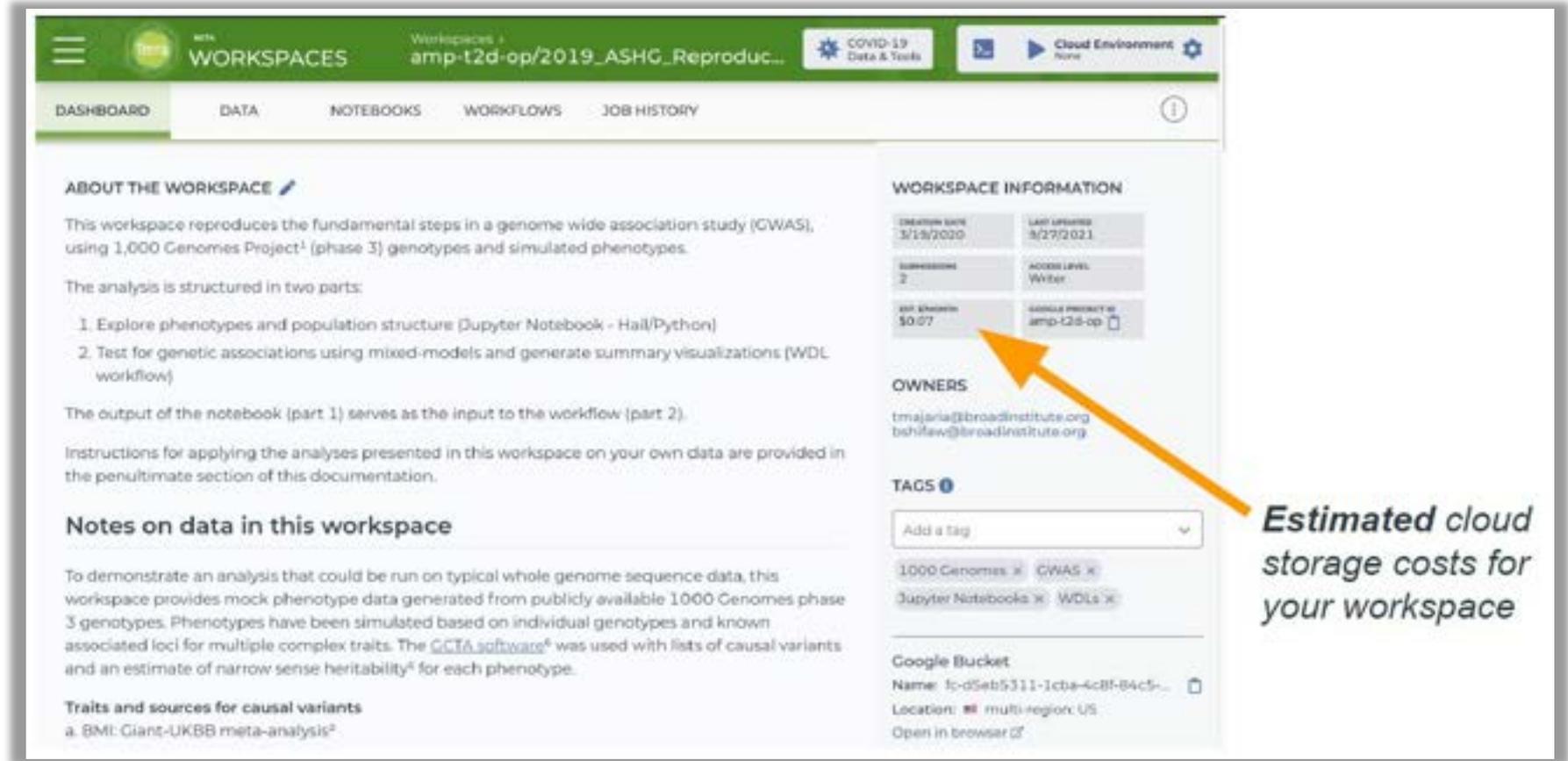
Understanding and monitoring costs

You can **ESTIMATE COSTS**:

1. analysis costs
2. **cloud storage costs**
3. egress (i.e., data moving) costs

You can **CHECK ACTUAL COSTS** in the Google Cloud Platform Console

You can **REDUCE COSTS** in several ways (for advanced users)



The screenshot displays the Google Cloud Workspaces interface for a workspace named "amp-t2d-op/2019_ASHG_Reproduc...". The interface includes a navigation menu with options like DASHBOARD, DATA, NOTEBOOKS, WORKFLOWS, and JOB HISTORY. The main content area is divided into two columns. The left column, titled "ABOUT THE WORKSPACE", provides details about the workspace's purpose (reproducing GWAS steps) and the analysis structure. The right column, titled "WORKSPACE INFORMATION", contains a table with the following data:

| CREATION DATE | LAST UPDATED |
|---------------|-------------------|
| 3/19/2020 | 9/27/2021 |
| SUBSCRIPTIONS | ACCESS LEVEL |
| 2 | Writer |
| EST. STORAGE | SAMPLE PROJECT ID |
| \$0.07 | amp-t2d-op |

An orange arrow points from the text "Estimated cloud storage costs for your workspace" to the "\$0.07" value in the "EST. STORAGE" row. Below the table, the "OWNERS" section lists "tmajaria@broadinstitute.org" and "bshilaw@broadinstitute.org". The "TAGS" section includes a search box and tags for "1000 Genomes", "GWAS", "Jupyter Notebooks", and "WDLs". At the bottom, the "Google Bucket" section shows the name "fc-d5eb5311-1cbe-4c8f-b4c5-...", location "multi-region: US", and a link to "Open in browser".

Understanding and monitoring costs

You can **ESTIMATE COSTS**:

1. analysis costs
2. cloud storage costs
3. egress (i.e., data moving) costs

You can **CHECK ACTUAL COSTS** in the Google Cloud Platform Console

You can **REDUCE COSTS** in several ways (for advanced users)

File Details

Filename
HG03642.final.cram.crai

DOS uri's can't be previewed

File size
1.31 MB

View this file in the Google Cloud Storage Browser

DOWNLOAD FOR < \$0.01*

Terminal download command
`gsutil cp gs://nih-nhlbi-biodata-catalys`

> More Information

* Estimated. Download cost may be higher in China or Australia.

DONE

Find price to egress

Will you download derived data to save locally or elsewhere?

| TABLES | DATA | NOTEBOOKS |
|-------------------------------|------|-----------|
| aligned_reads_index (2504) | | |
| aliquot (2504) | | |
| germline_variation_... (2504) | | |
| pheno-data (2504) | | |
| program (1) | | |
| project (1) | | |
| read_group (2504) | | |
| reference_file (44) | | |
| sample (2504) | | |
| simple_germline_vs_... (2504) | | |
| study (1) | | |

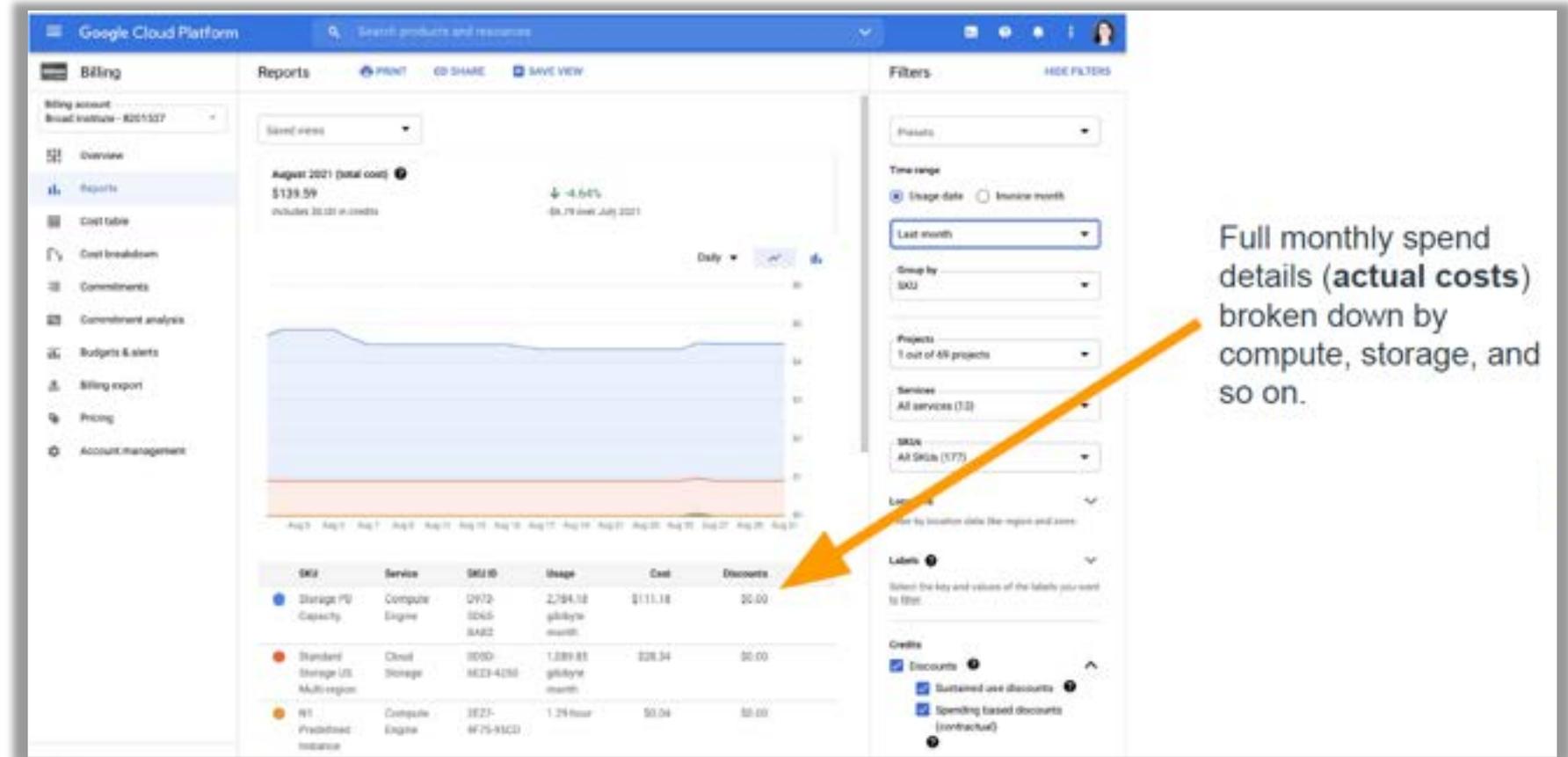
Understanding and monitoring costs

You can **ESTIMATE COSTS**:

1. analysis costs
2. cloud storage costs
3. egress (i.e., data moving) costs

You can **CHECK ACTUAL COSTS** in the Google Cloud Platform Console

You can **REDUCE COSTS** in several ways (for advanced users)



Understanding and monitoring costs

You can **ESTIMATE COSTS**:

1. analysis costs
2. cloud storage costs
3. egress (i.e., data moving) costs

You can **CHECK ACTUAL COSTS** in the Google Cloud Platform Console

You can **REDUCE COSTS** in several ways ([guides are for advanced users](#))

Terra allows you to find the right balance between cost and time

Saving on workflow costs

- ▶ Delete intermediate files: [guide](#)
- ▶ Call-caching: [guide](#)
- ▶ Checkpointing: [guide](#)
- ▶ Preemptible VMs: [guide](#)

Saving Cloud Environment costs

- ▶ Size application compute appropriately: [guide](#)
- ▶ Move generated data to regional or nearline storage: [guide](#)
- ▶ Autopause: [guide](#)

Saving on storage costs

- ▶ Ask how much are you storing, where are you storing it, and how frequently will you access it?
- ▶ Move data to regional or nearline storage: [guide](#)

SciARe

The logo for SciARe features the word "SciARe" in a white, bold, sans-serif font. The letters "i" and "A" are partially obscured by a stylized orange and yellow cloud. A purple arrow points from the "i" towards the "A", and another purple arrow points from the "A" towards the "i", creating a circular flow.

Thank you

Think-a-Thon poll

1. Rate how useful this session was:

- Very useful
- Useful
- Somewhat useful
- Not at all useful

Think-a-Thon poll

2. Rate the pace of the instruction for yourself:

- Too fast
- Adequate for me
- Too slow

Think-a-Thon poll

3. How likely will you participate in the next Think-a-Thon?

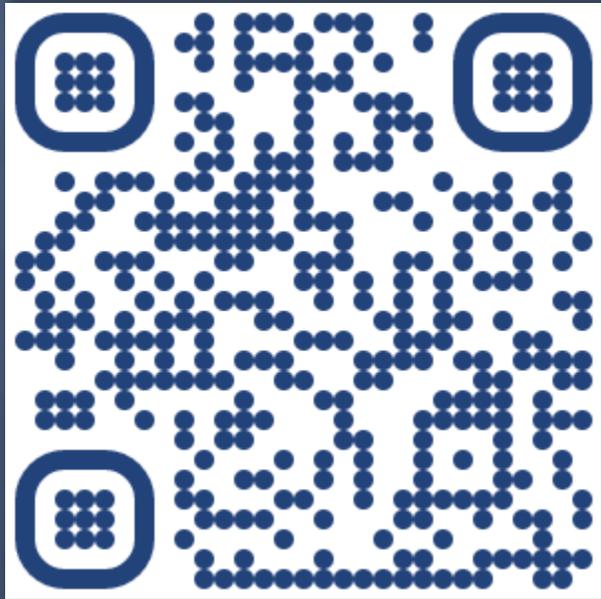
- Very interested, will definitely attend
- Interested, likely will attend
- Interested, but not available
- Not interested in attending any others

Terra tutorials and resources

If you are new to Terra, we recommend exploring the following resources:

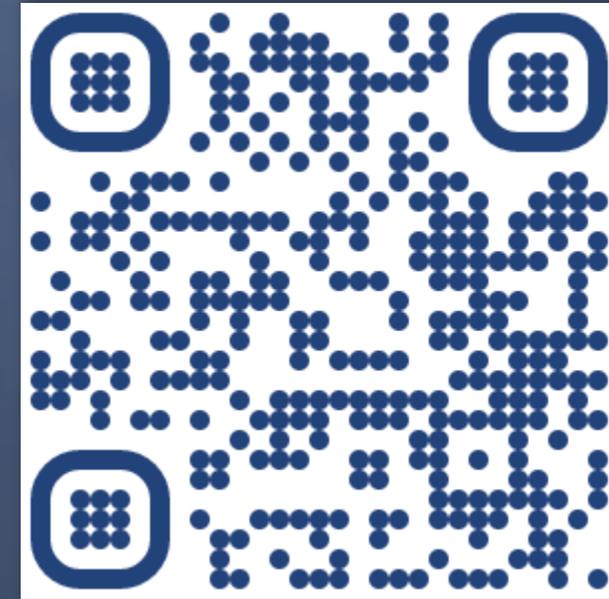
- [Overview Articles](#): Review high-level docs that outline what you can do in Terra, how to set up an account and account billing, and how to access, manage, and analyze data in the cloud
- [Video Guides](#): Watch live demos of the Terra platform's useful features
- [Terra Courses](#): Learn about Terra with free modules on the Leanpub online learning platform
- [Data Tables QuickStart Tutorial](#): Learn what data tables are and how to create, modify, and use them in analyses
- [Notebooks QuickStart Tutorial](#): Learn how to access and visualize data using a notebook
- [Machine Learning Advanced Tutorial](#): Learn how Terra can support machine learning-based analysis

Next Think-a-Thons:



bit.ly/think-a-thons

Register for ScHARe:



bit.ly/join-schare

✉ schare@mail.nih.gov

References and credits

- **Tutorials and notebooks:** The Broad Institute, Inc., Verily Life Sciences, LLC