NIH

May 17, 2023 Think-a-Thon

ScHARe

ScHARe III · Terra Datasets

Deborah Duran, PhD and Luca Calzoni, MD MS PhD Cand. | NIMHD

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 the day of 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

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

ScHARe

NIH National Institute on Minority Health and Health Disparities

+

NIH Office of Data Science Strategy

+

NIH 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

NINR
Rebecca Hawes
Micheal Steele
John Grason

ORWH

OMH

NIMHD OCPL
Kelli Carrington
Thoko Kachipande
Corinne Baker

BioTeam

STRIDES

Terra

SIDEM

RLA

Broad Institute

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 Hosted data
	▪ Python poll
10’	Billing and costs
	▪ Think-a-Thon 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	□	□	□	□

ScHARe

Part I

ScHARe and Terra Overview

ScHARe

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

nimhd.nih.gov/schare

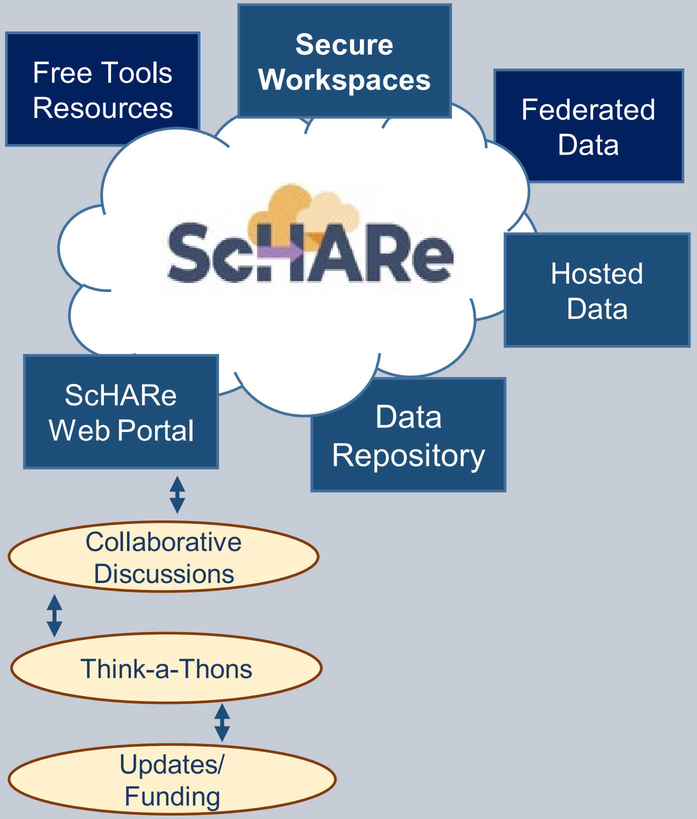
ScHARe Components

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

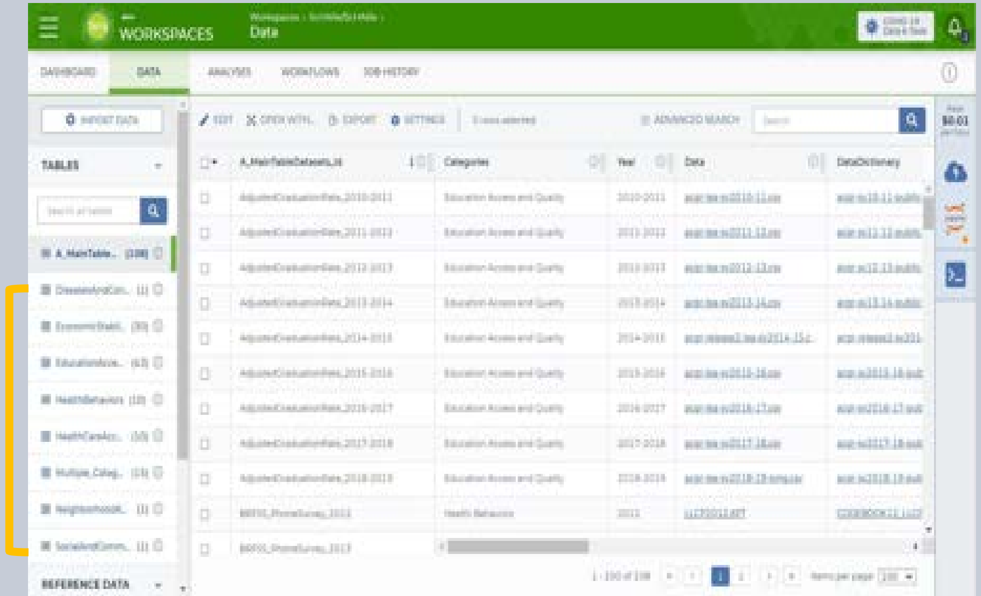
  Example: *American Community Survey (ACS)*

- **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*



A_MainTableDatasets_id	Categories	Year	Data	DataDictionary
AdjustedGraduationRate_2010-2011	Education Access and Quality	2010-2011	[illegible]	[illegible]
AdjustedGraduationRate_2011-2012	Education Access and Quality	2011-2012	[illegible]	[illegible]
AdjustedGraduationRate_2012-2013	Education Access and Quality	2012-2013	[illegible]	[illegible]
AdjustedGraduationRate_2013-2014	Education Access and Quality	2013-2014	[illegible]	[illegible]
AdjustedGraduationRate_2014-2015	Education Access and Quality	2014-2015	[illegible]	[illegible]
AdjustedGraduationRate_2015-2016	Education Access and Quality	2015-2016	[illegible]	[illegible]
AdjustedGraduationRate_2016-2017	Education Access and Quality	2016-2017	[illegible]	[illegible]
AdjustedGraduationRate_2017-2018	Education Access and Quality	2017-2018	[illegible]	[illegible]
AdjustedGraduationRate_2018-2019	Education Access and Quality	2018-2019	[illegible]	[illegible]
BRFSS_PhoneSurvey_2011	Health Behaviors	2011	LLCP2011.XPT	[illegible]
BRFSS_PhoneSurvey_2013				

1 - 100 of 108

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:

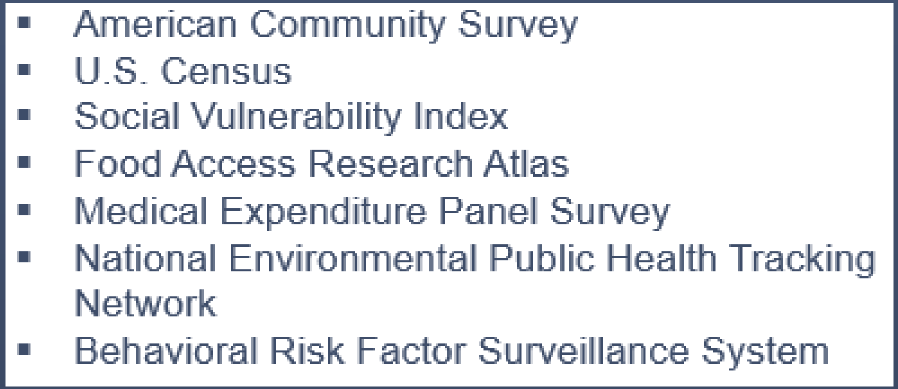
- **Health Behaviors**
- **Diseases and Conditions**

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

ScHARe

- 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.

ScHARe

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

ScHARe

Phase II

(in process)

Repository and Data Ecosystem

ScHARe Data Repository

CORE COMMON DATA ELEMENTS

NOVEL CDE FOCUSED REPOSITORY TO FOSTER INTEROPERABILITY

COMPLY WITH DATA SHARING POLICY - HOST PROJECT DATA

DATA ECOSYSTEM

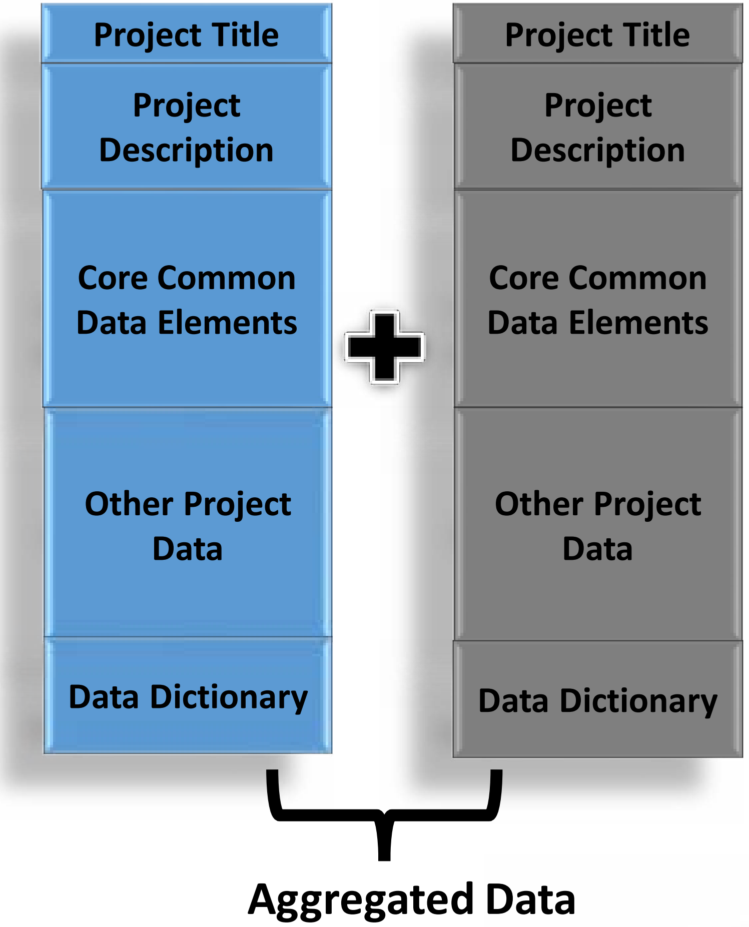
- Map across datasets
- Map across platforms

UPCOMING

ScHARe

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



Project Title

Project Description

Core Common Data Elements

Other Project Data

Data Dictionary

+

Project Title

Project Description

Core Common Data Elements

Other Project Data

Data Dictionary

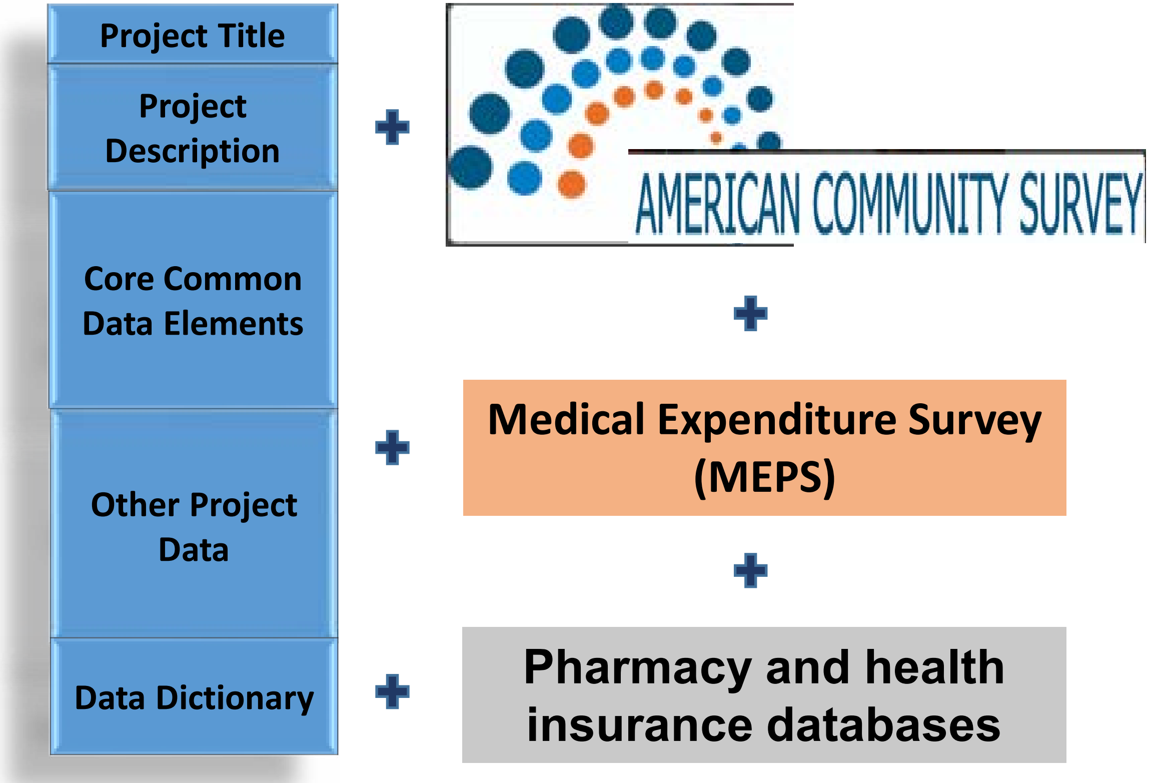
Aggregated Data

smoking
SDoH
geographic
Community
Interventions
Sleep
cancer
smoking
Mental health
geographic
Adherence
Applied Clinical
Mental health
geographic
Decision making
SDoH
Health Care Delivery
Diabetes
Decision making
SDOH
cancer

UPCOMING FALL

ScHARe

Project & federated dataset mapping



Project Title

Project Description

Core Common Data Elements

Other Project Data

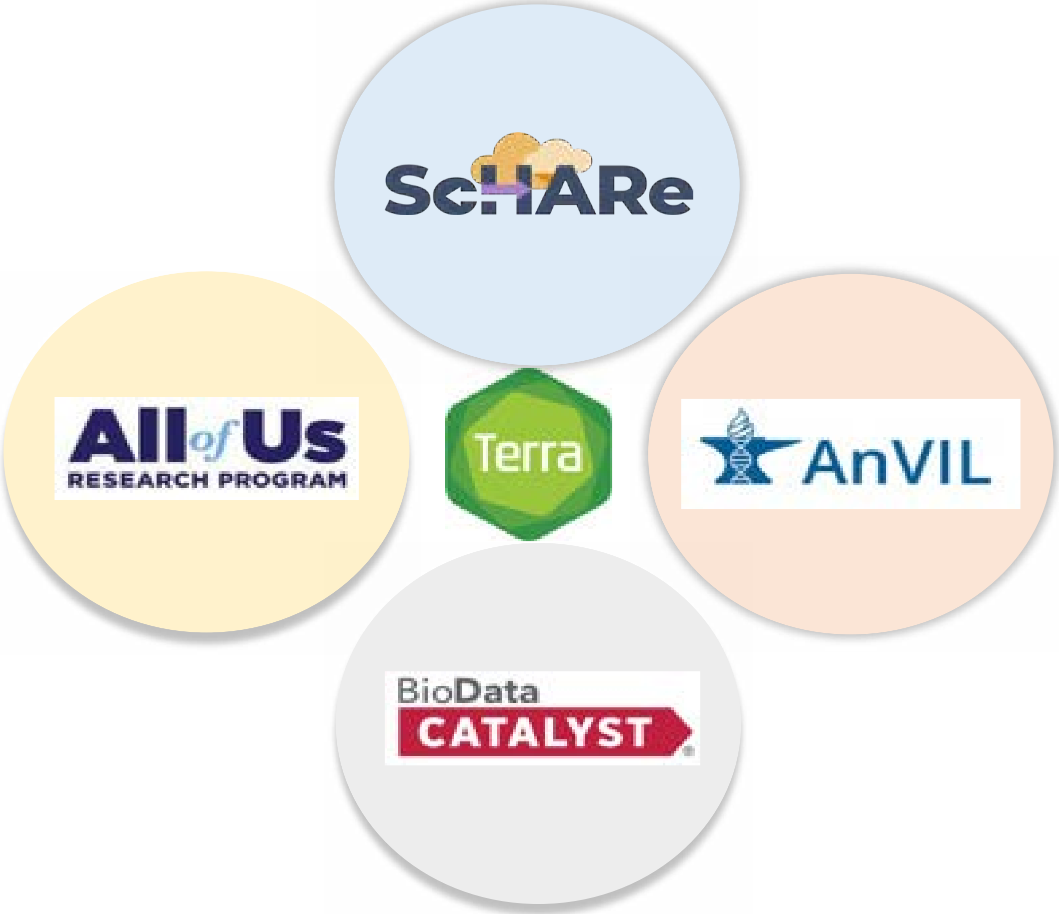
Data Dictionary

+ AMERICAN COMMUNITY SURVEY

+ Medical Expenditure Survey (MEPS)

+ Pharmacy and health insurance databases

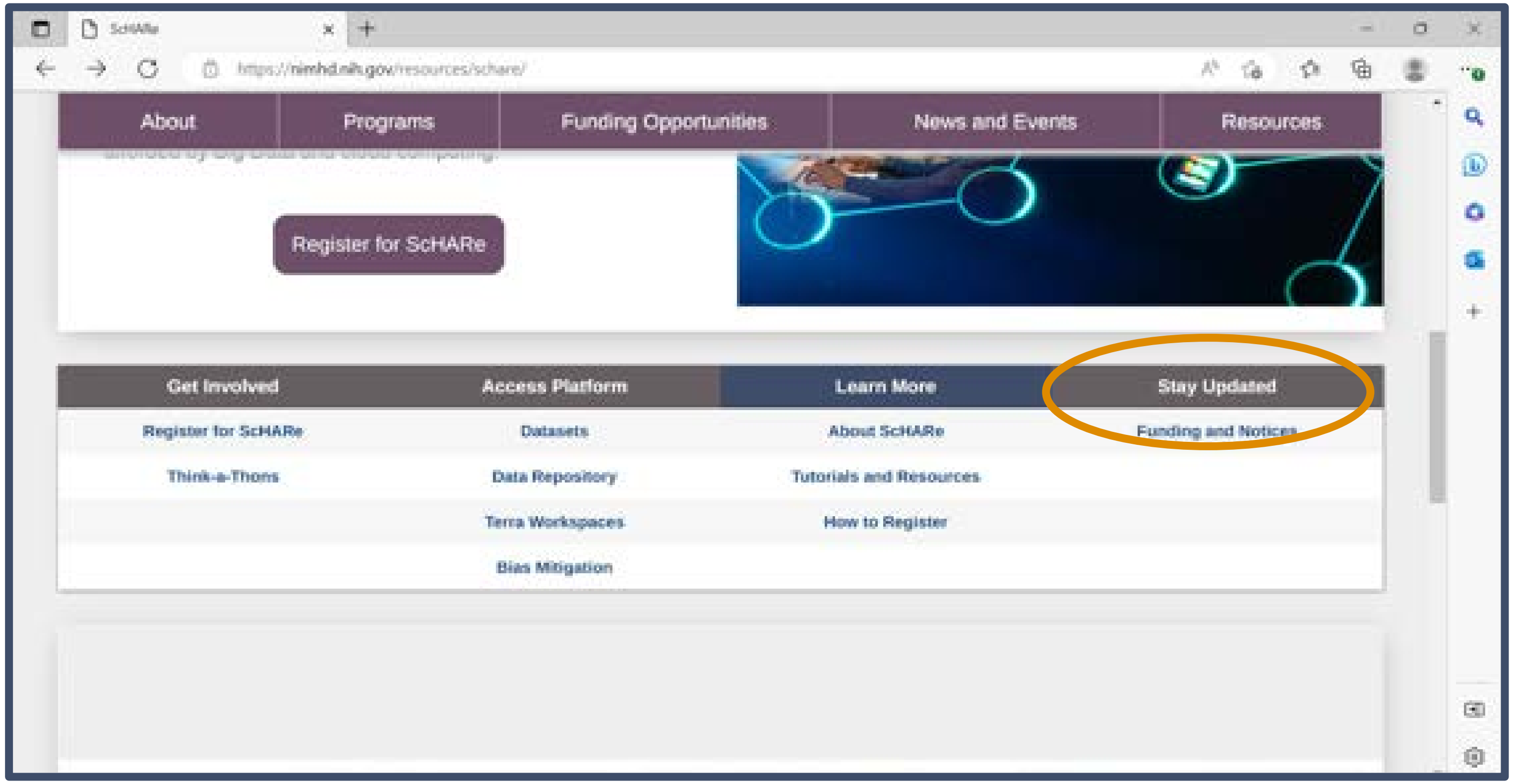
Mapping across cloud platforms



UPCOMING

Two ways to sign up for ScHARe news

Scannable from your screen!



Get Involved	Access Platform	Learn More	Stay Updated
Register for ScHARe	Datasets	About ScHARe	Funding and Notices
Think-a-Thons	Data Repository	Tutorials and Resources	
	Terra Workspaces	How to Register	
	Bias Mitigation		

nimhd.nih.gov/schare

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 (TaT)

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

✓ **Instructional**

✓ **Collaboration research teams**

✓ **Bias mitigation**

ScHARe

Think-a-Thon

Artificial Intelligence and Cloud Computing Basics

Terra: Datasets and Analytics

Register:

bit.ly/think-a-thons

SciHARe

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

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 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 the March Think-a-Thon slides for complete instructions

Copying workspaces

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

Why copy a workspace?

If you are interested in using the data resources of a workspace or replicating the analyses 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 displayed on screen and click on the corresponding **vertical three-dot menu**, then on **“Clone”**
3. Input a **name** for the workspace copy
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. You can disregard this for now. Info on ADs: bit.ly/AutDom

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**

ScHARe

Part III
ScHARe Datasets

ScHARe Data

What data can you work with on ScHARe?

Data you upload

to your workspace

This is your own personal project data, stored on your computer

Data already in the ScHARe Data Ecosystem

1. Google Hosted Public Datasets
2. ScHARe Hosted Public Datasets
3. ScHARe Hosted Project Datasets

ScHARe Ecosystem data

The ScHARe Data Ecosystem is comprised of:

- **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)*

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

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

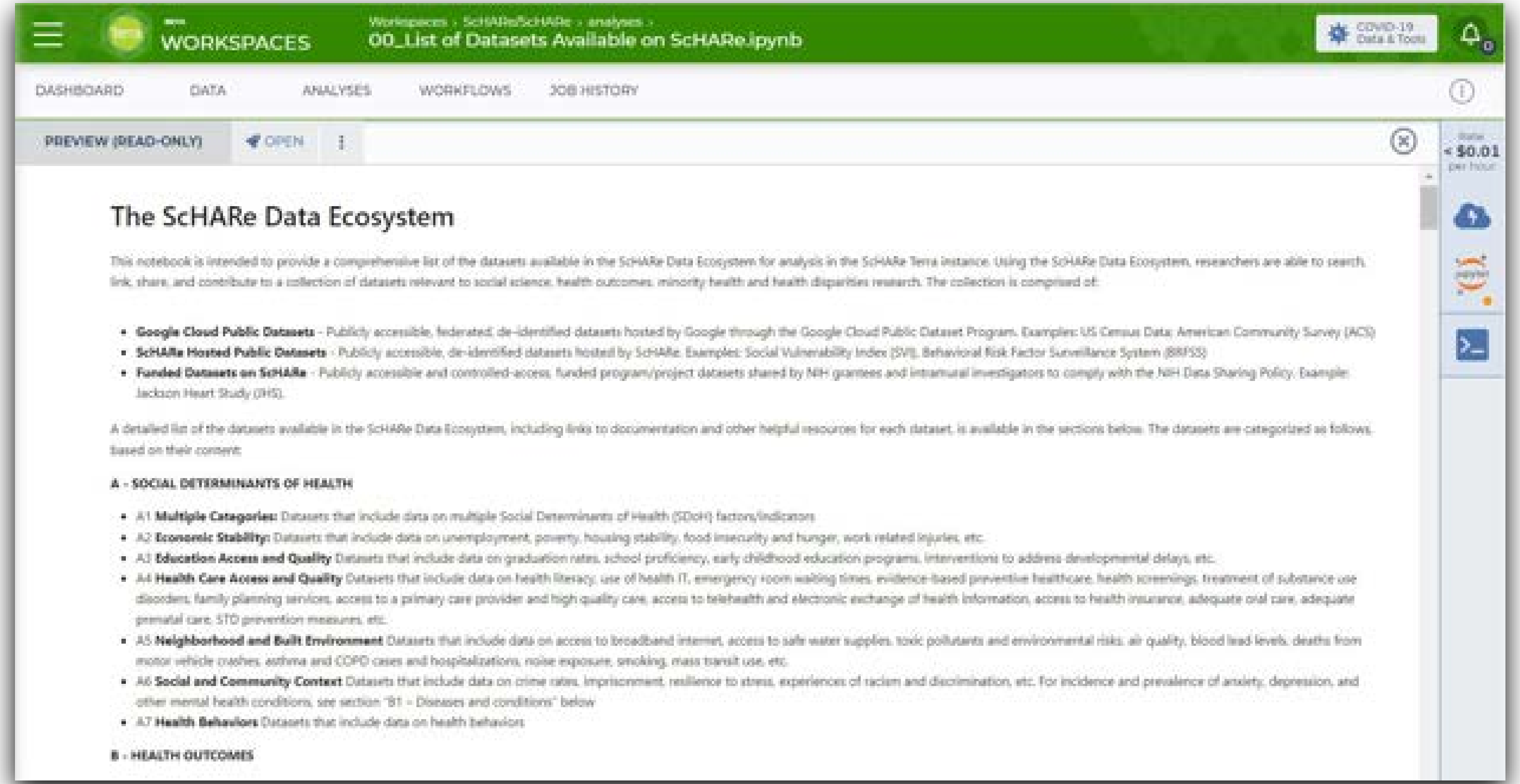
- **ScHARe Hosted Project Datasets:** publicly accessible and controlled-access, funded program/project datasets using Core Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

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

How to see what data is available

Analyses tab

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 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

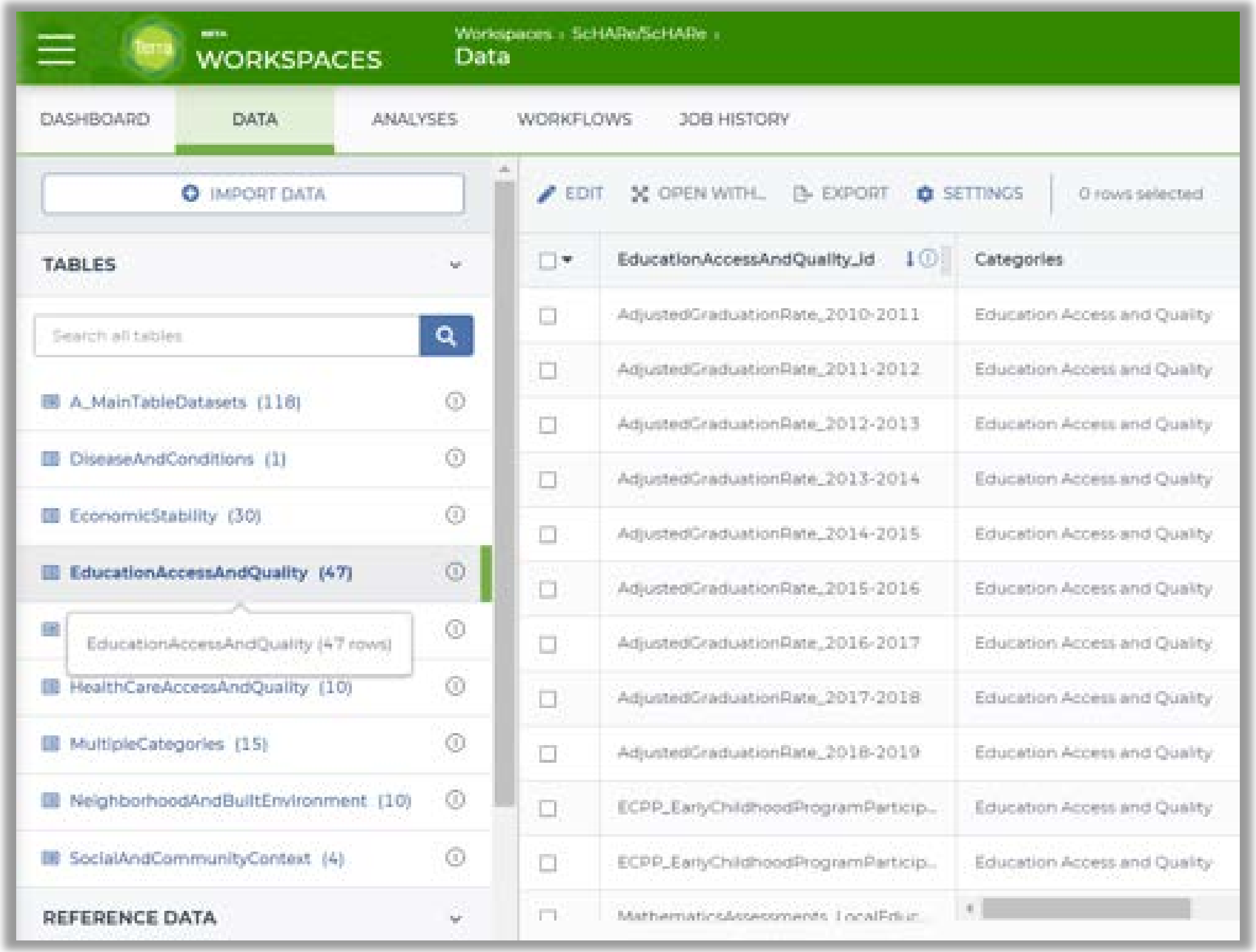
B - HEALTH OUTCOMES

How to access available data

Data tab

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



- A_MainTableDatasets (118)
- DiseaseAndConditions (1)
- EconomicStability (30)
- **EducationAccessAndQuality (47)**
- EducationAccessAndQuality (47 rows)
- HealthCareAccessAndQuality (10)
- MultipleCategories (15)
- NeighborhoodAndBuiltEnvironment (10)
- SocialAndCommunityContext (4)

□	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...	

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

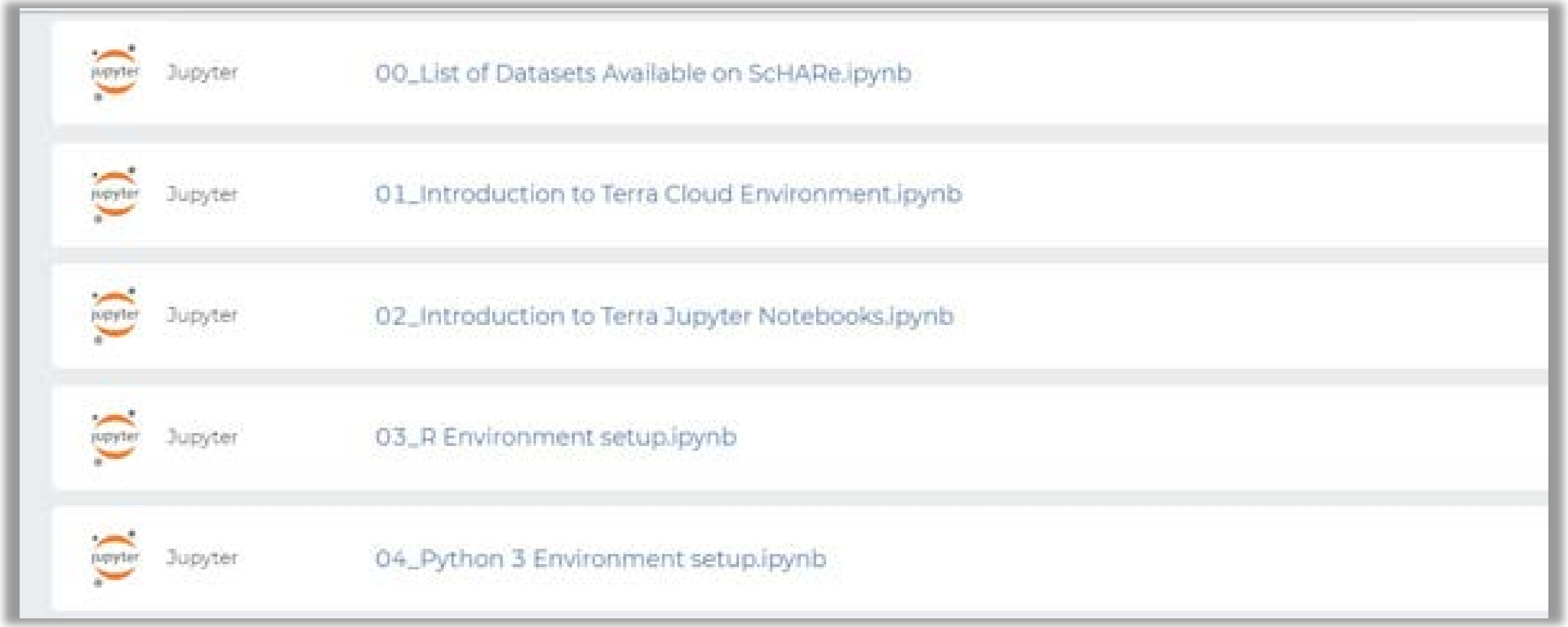
Using notebooks:

- **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

The 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.

What can our notebooks teach you?



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

For the Educators among us:

Notebooks can be great instructional tools:

- they integrate code and explanations into a single document
- they can make your teaching materials more understandable, repeatable, and shareable

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 **using the instructions in our notebooks**

Today’s hands-on tutorials

In this Think-a-Thon, we will focus on the categories highlighted below:

Data you upload

to your workspace

This is your own personal project data, stored on your computer

Data already in the ScHARe Data Ecosystem

1. Google Hosted Public Datasets
2. ScHARe Hosted Public Datasets
3. ScHARe Hosted Project Datasets

SciHARe

Part IV

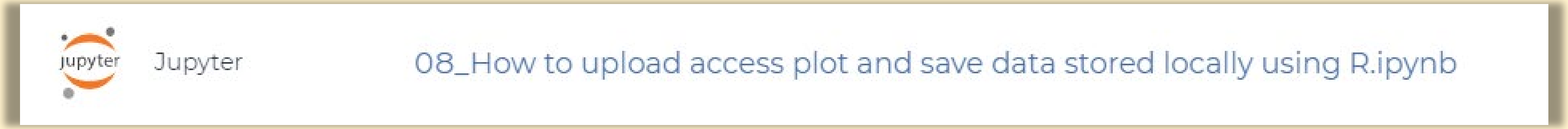
How to Work with Data Uploaded to Terra

How to work with data you upload

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

Instructional materials with step-by-step instructions and videos will be posted online here: **bit.ly/think-a-thons**

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



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

Why 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

How to 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

What data will we work with?

Data you upload

to your workspace

This is your own personal project data, stored on your computer

So that we can all work with the same data today, **we will ask you to:**

1. **download the same dataset** (MHSVI)
2. **upload it** into your Terra workspace

The MHSVI data we will use

- The **Minority Health Social Vulnerability Index dataset** (MHSVI) is a **TSV file**
- **TSV** is an abbreviation for tab-separated values file – a file format commonly used to exchange data between databases
- There are **many other file formats,** each with its own way of separating/storing data. For example: a TSV file uses tabs, while a CSV file uses commas.

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

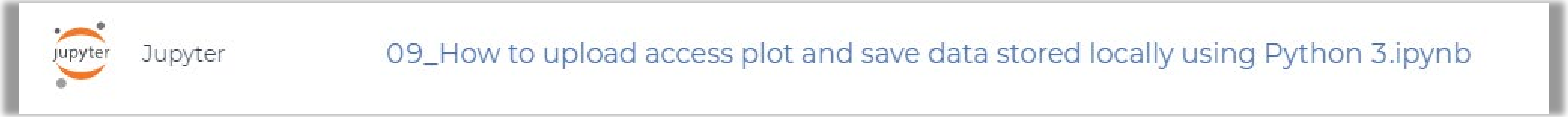
The factors are organized into **six themes**:

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

Let’s start!

To begin:

1. **point your browser to: terra.bio**
2. **log in to Terra**
3. **access the “ScHARe Think-a-Thons” workspace**
4. **go to the Analyses tab**
5. **run the following notebook** and complete the steps illustrated by the instructors:



Jupyter

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

SciHARe

Part V

How to Work with Google Hosted Data

Python poll

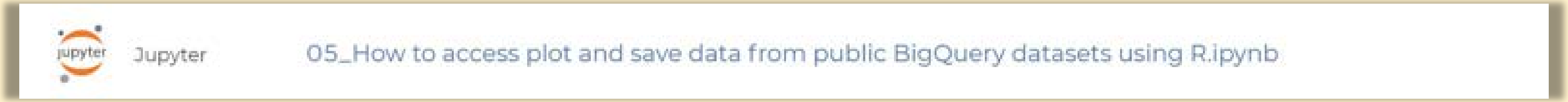
What are the challenges of learning Python in your world today?

How to work with Google hosted data

This tutorial is an introduction to analyzing data **from Google hosted public datasets**

Instructional materials with step-by-step instructions and videos will be posted online here:
bit.ly/think-a-thons

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



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

What are the Google Hosted datasets?

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 access on Terra by using BigQuery**

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

What datasets are available through Google?

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)

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

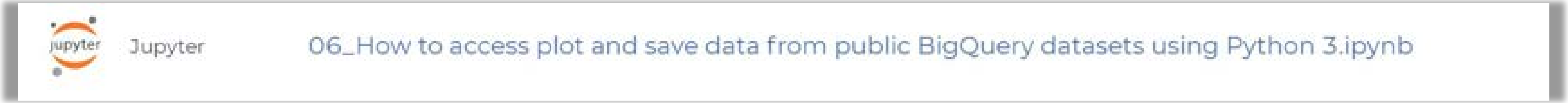
- A **low ADI score** indicates affluence or prosperity
- A **high ADI score** is indicative of high levels of deprivation, which have been **linked to health outcomes**, such as 30-day hospital readmission rates, cardiovascular disease and cancer deaths

➢ **Neighborhood and racial disparities occur when some neighborhoods have high ADI scores** and others have low scores

Let’s start!

To begin:

1. point your browser to: terra.bio
2. log in to Terra
3. access the “ScHARe Think-a-Thons” workspace
4. go to the Analyses tab
5. **run the following notebook** and complete the steps illustrated by the instructors:



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

ScHARe

Part VI

Billing and Costs

What are the cloud costs of working on Terra?

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

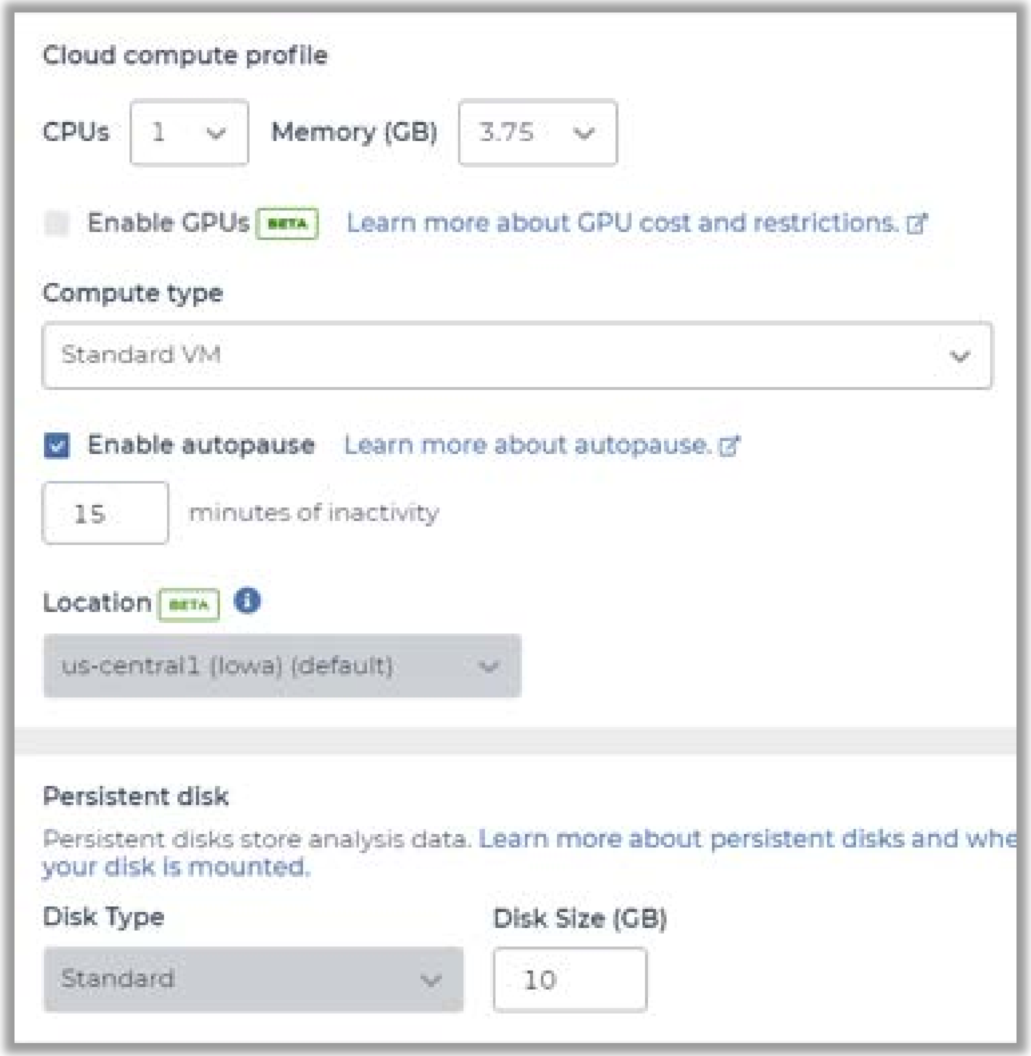
However, the following operations in Terra **may incur charges:**

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



Cloud compute profile

CPUs 1 Memory (GB) 3.75

□ Enable GPUs BETA Learn more about GPU cost and restrictions.

Compute type

Standard VM

☑ Enable autopause Learn more about autopause.

15 minutes of inactivity

Location BETA

us-central1 (Iowa) (default)

Persistent disk

Persistent disks store analysis data. Learn more about persistent disks and whe
your disk is mounted.

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

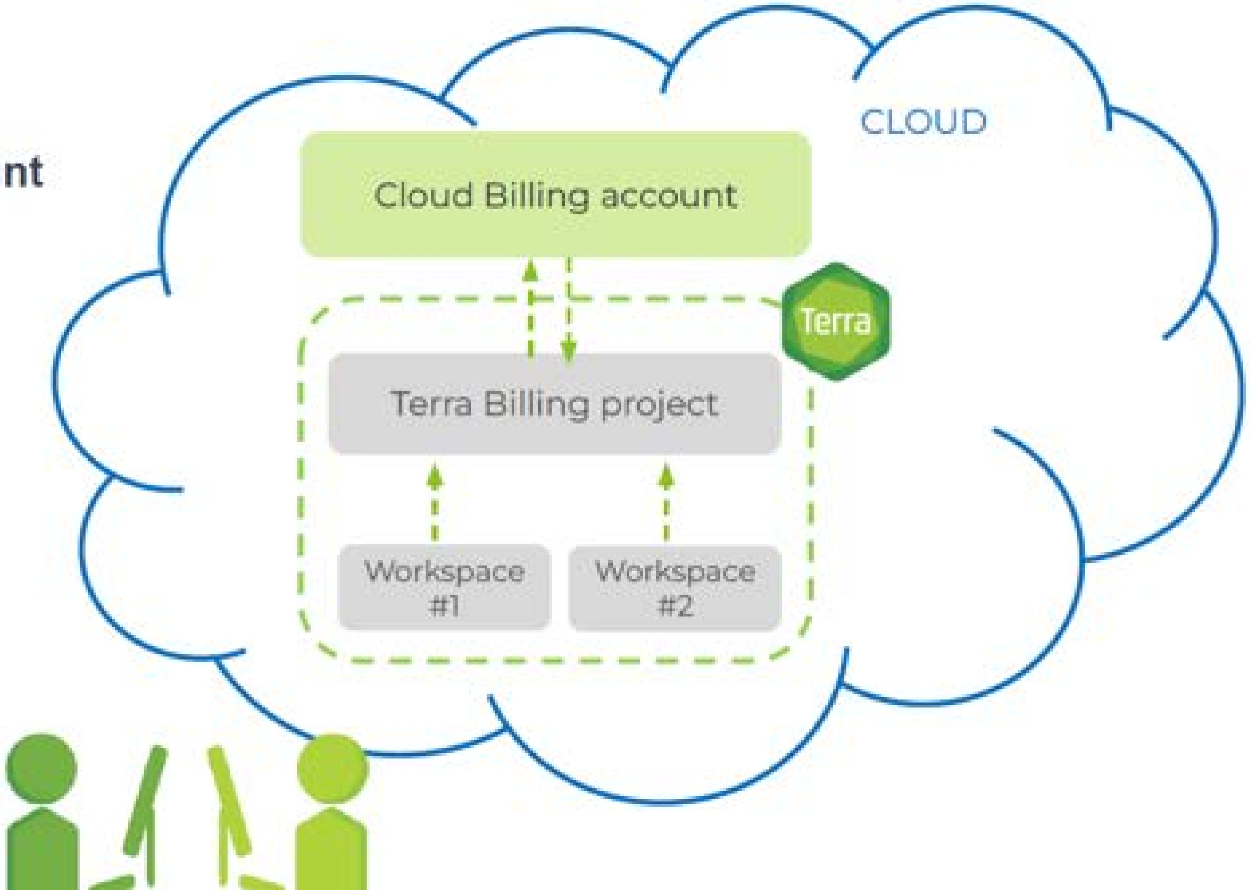
3. Data egress (i.e. moving data) costs

- 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 your Terra Billing project

For detailed instructions, see **this Terra page**

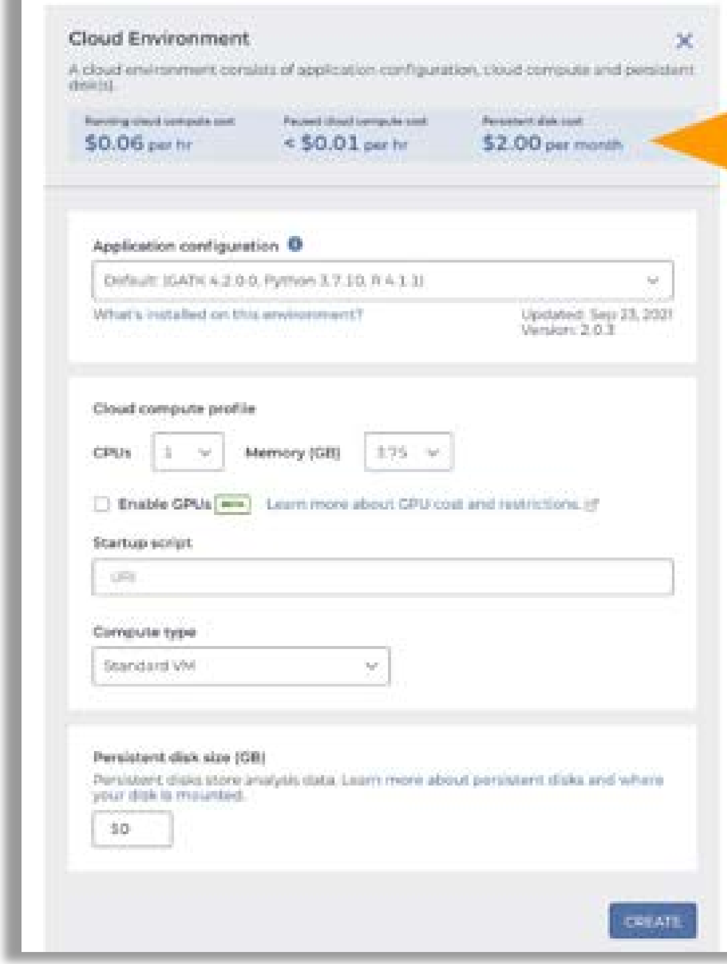
Understanding and monitoring costs

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)



Cloud Environment

A cloud environment consists of application configuration, cloud compute and persistent disk(s).

Running cloud compute cost: $0.06 per hr

Paused cloud compute cost: < $0.01 per hr

Persistent disk cost: $2.00 per month

Application configuration

Default: (GATK 4.2.0.0, Python 3.7.10, R 4.1.1)

What's installed on this environment?

Updated: Sep 23, 2021
Version: 2.0.3

Cloud compute profile

CPUs: 1 Memory (GB): 3.75

□ Enable GPUs beta Learn more about GPU cost and restrictions.

Startup script: URI

Compute type: Standard VM

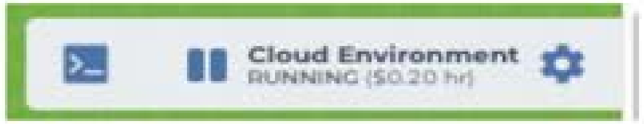
Persistent disk size (GB)

Persistent disks store analysis data. Learn more about persistent disks and where your disk is mounted.

50

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

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



Cloud Environment
RUNNING ($0.20 hr)

- 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!

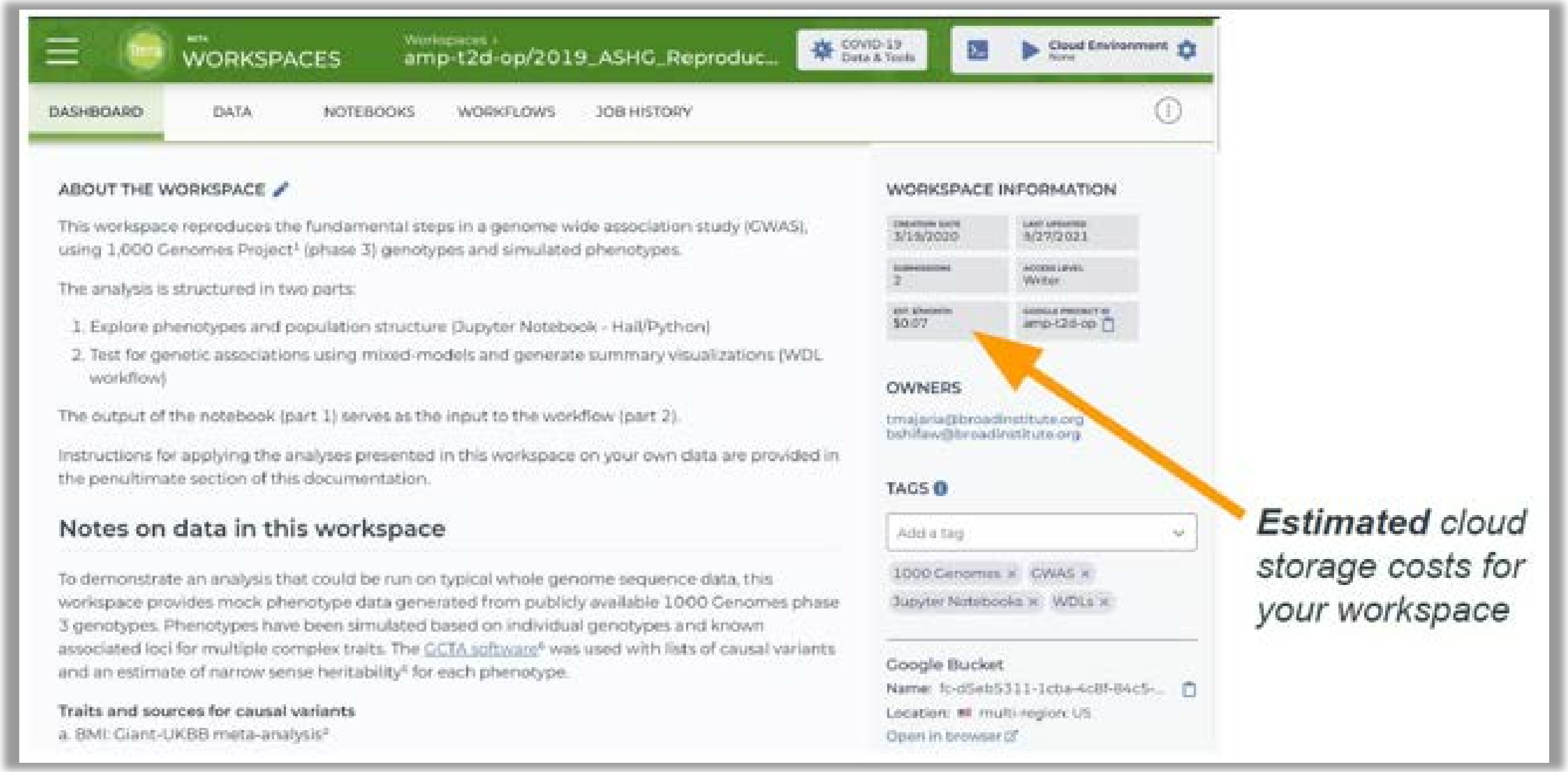
Understanding and monitoring costs

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)



ABOUT THE WORKSPACE

This workspace reproduces the fundamental steps in a genome wide association study (GWAS), using 1,000 Genomes Project[1] (phase 3) genotypes and simulated phenotypes.

The analysis is structured in two parts:

1. Explore phenotypes and population structure (Jupyter Notebook - Hail/Python)
2. Test for genetic associations using mixed-models and generate summary visualizations (WDL workflow)

The output of the notebook (part 1) serves as the input to the workflow (part 2).

Instructions for applying the analyses presented in this workspace on your own data are provided in the penultimate section of this documentation.

Notes on data in this workspace

To demonstrate an analysis that could be run on typical whole genome sequence data, this workspace provides mock phenotype data generated from publicly available 1000 Genomes phase 3 genotypes. Phenotypes have been simulated based on individual genotypes and known associated loci for multiple complex traits. The GCTA software[6] was used with lists of causal variants and an estimate of narrow sense heritability[6] for each phenotype.

Traits and sources for causal variants

a. BMI: Giant-UKBB meta-analysis[2]

WORKSPACE INFORMATION

CREATION DATE 3/19/2020	LAST UPDATED 9/27/2021
SUBMISSIONS 2	ACCESS LEVEL Writer
EST. $/MONTH $0.07	GOOGLE PROJECT ID amp-t2d-op

OWNERS

tmajaria@broadinstitute.org
bshifaw@broadinstitute.org

TAGS

Add a tag

1000 Genomes × GWAS × Jupyter Notebooks × WDLs ×

Google Bucket

Name: fc-d5eb5311-1cba-4c8f-84c5-...

Location: multi-region: US

Open in browser

Estimated cloud storage costs for your workspace

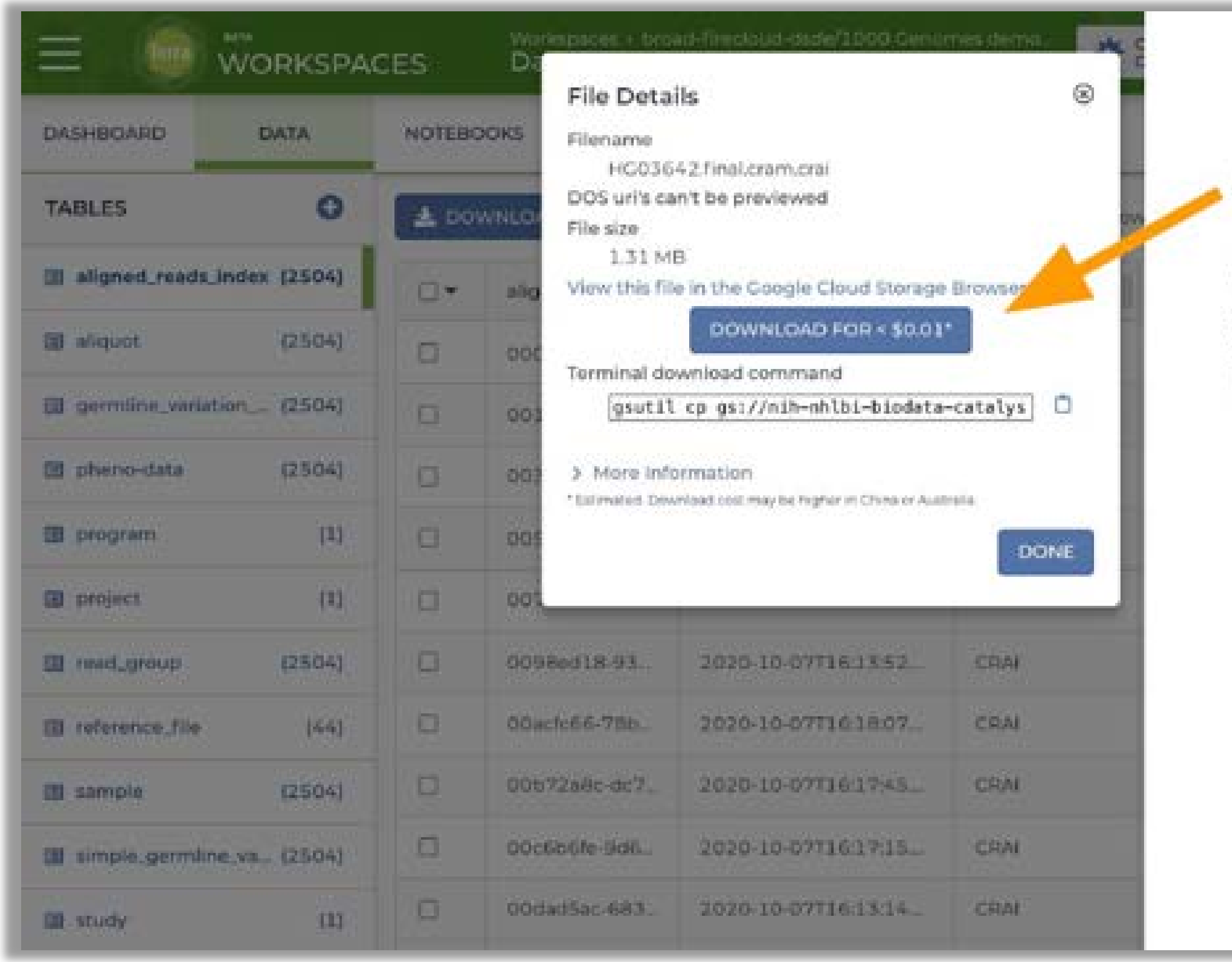
Understanding and monitoring costs

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)



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_va...	(2504)
study	(1)

□	0098ed18-93...	2020-10-07T16:13:52...	CRAI
□	00acfc66-78b...	2020-10-07T16:18:07...	CRAI
□	00b72a8c-dc7...	2020-10-07T16:17:45...	CRAI
□	00c6b6fe-9d6...	2020-10-07T16:17:15...	CRAI
□	00dad5ac-683...	2020-10-07T16:13:14...	CRAI

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?

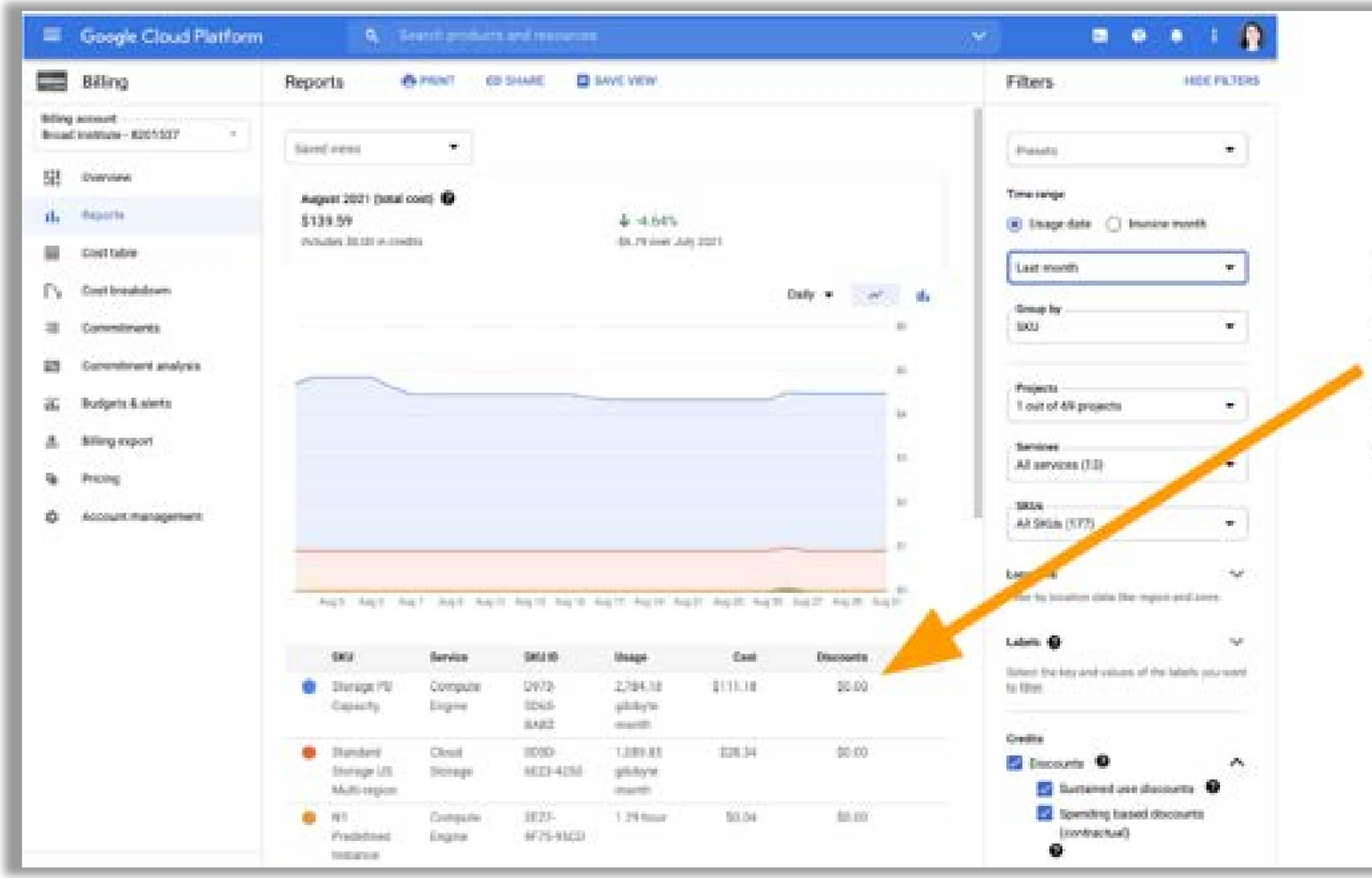
Understanding and monitoring costs

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)



August 2021 (total cost)
$139.59
Includes $0.00 in credits

-4.64%
$6.79 over July 2021

SKU	Service	SKU ID	Usage	Cost	Discounts
Storage PD Capacity	Compute Engine	D973-5D65-BAB2	2,784.18 gibibyte month	$111.18	$0.00
Standard Storage US Multi-region	Cloud Storage	0D5D-6E23-4250	1,089.83 gibibyte month	$28.34	$0.00
N1 Predefined Instance	Compute Engine	2E27-4F75-95CD	1.79 hour	$0.04	$0.00

Full monthly spend details (**actual costs**) broken down by compute, storage, and so on.

console.cloud.google.com

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

ScHARe

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