NIH

September 20, 2023
Think-a-Thon

ScHARe V · Review

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

Science collaborative for Health disparities and Artificial intelligence bias Reduction

SciHARe


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

SciAReS

+

Dr. Deborah Duran NIH/NIMHD

Dr. Luca Calzoni NIH/NIMHD

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
	▪ Experience poll
5’	ScHARe overview
	▪ Interest poll
15’	Account setup
10’	Workspaces and permissions
10’	Notebooks and environment
10’	Datasets
15’	How to upload data to your workspace
25’	How to work with Google hosted data
25’	How to work with ScHARe hosted data
15’	Data Explorer demo
	▪ Data exploration poll
15’	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	☐	☐	☐	☐

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

nimhd.nih.gov/schare

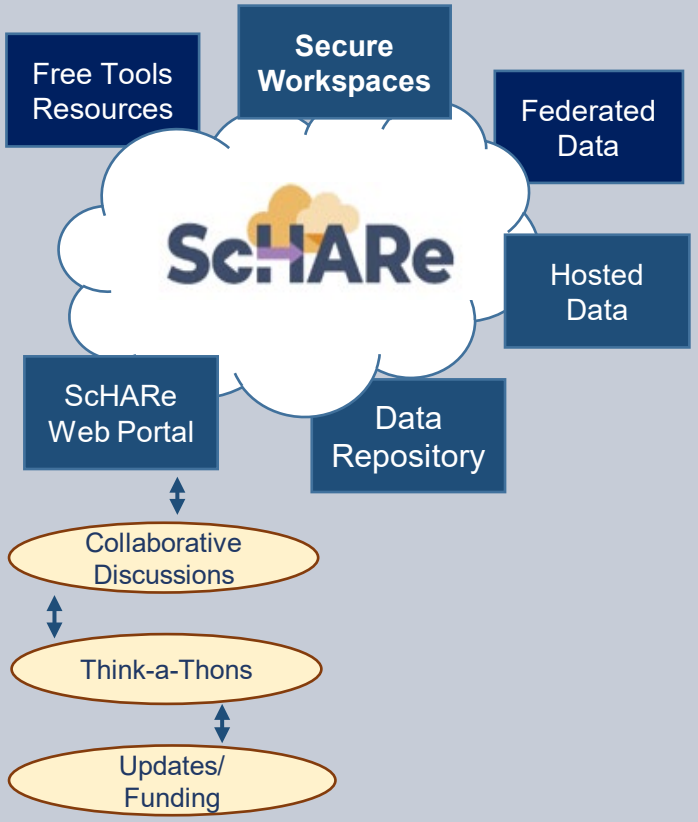
ScHARe Components

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

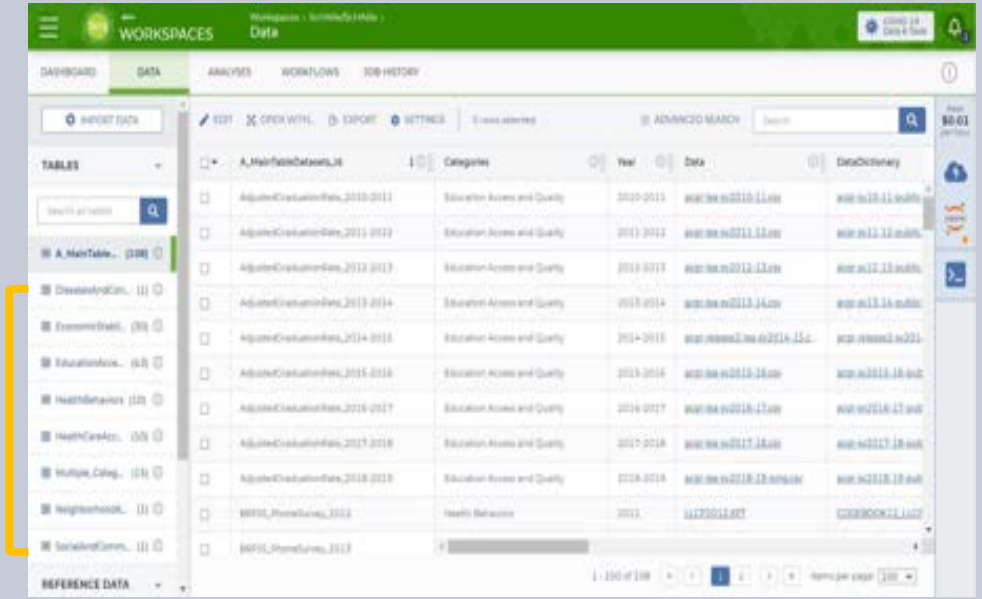
  Example: *American Community Survey (ACS)*

- **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*



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.

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

SciHARe

Phase II

(in process)

Data ecosystem and repository

SCHARe Data Repository

CORE COMMON DATA ELEMENTS

NOVEL CDE FOCUSED REPOSITORY TO FOSTER INTEROPERABILITY

COMPLY WITH DATA SHARING POLICY - HOST PROJECT DATA

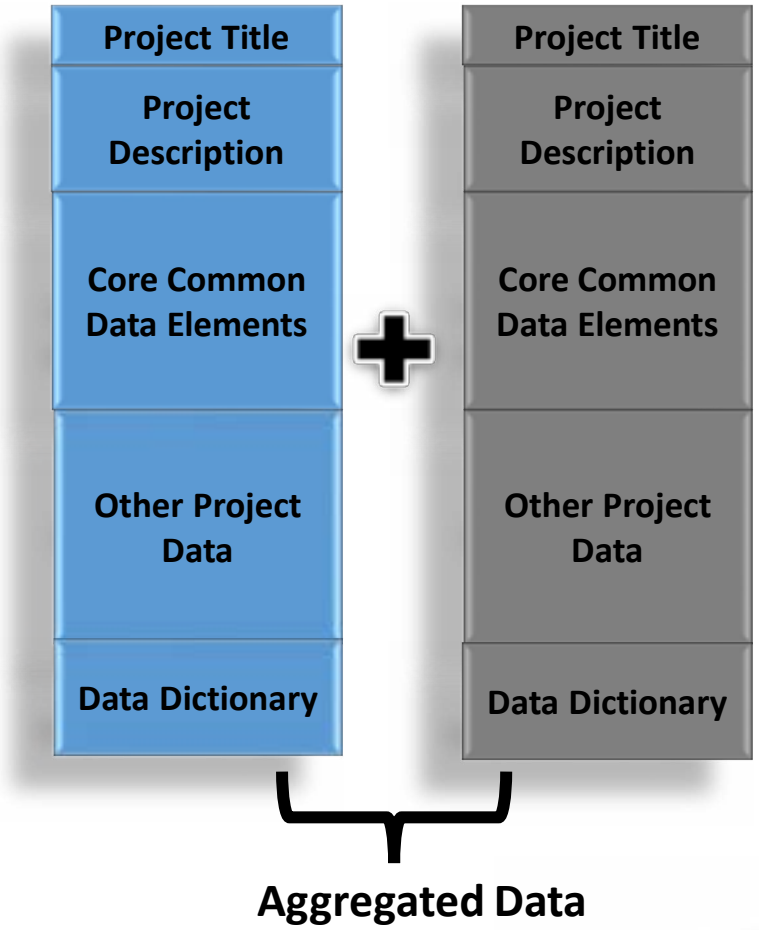
DATA ECOSYSTEM

- Map across datasets
- Map across platforms

UPCOMING

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

Project & federated dataset mapping

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

\+ AMERICAN COMMUNITY SURVEY

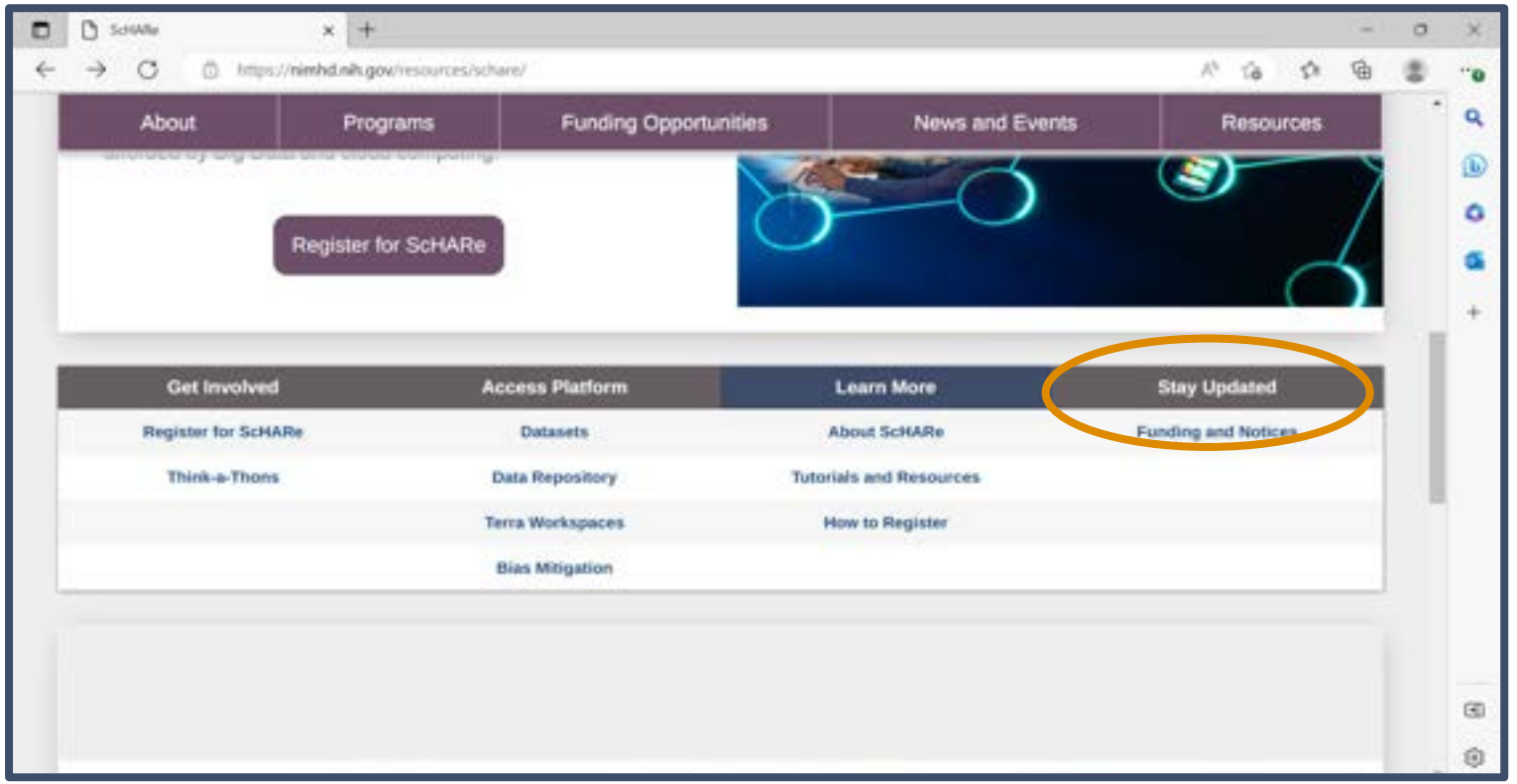
\+ Medical Expenditure Survey (MEPS)

\+ Pharmacy and health insurance databases

Mapping across cloud platforms

UPCOMING

Two ways to sign up for ScHARe news



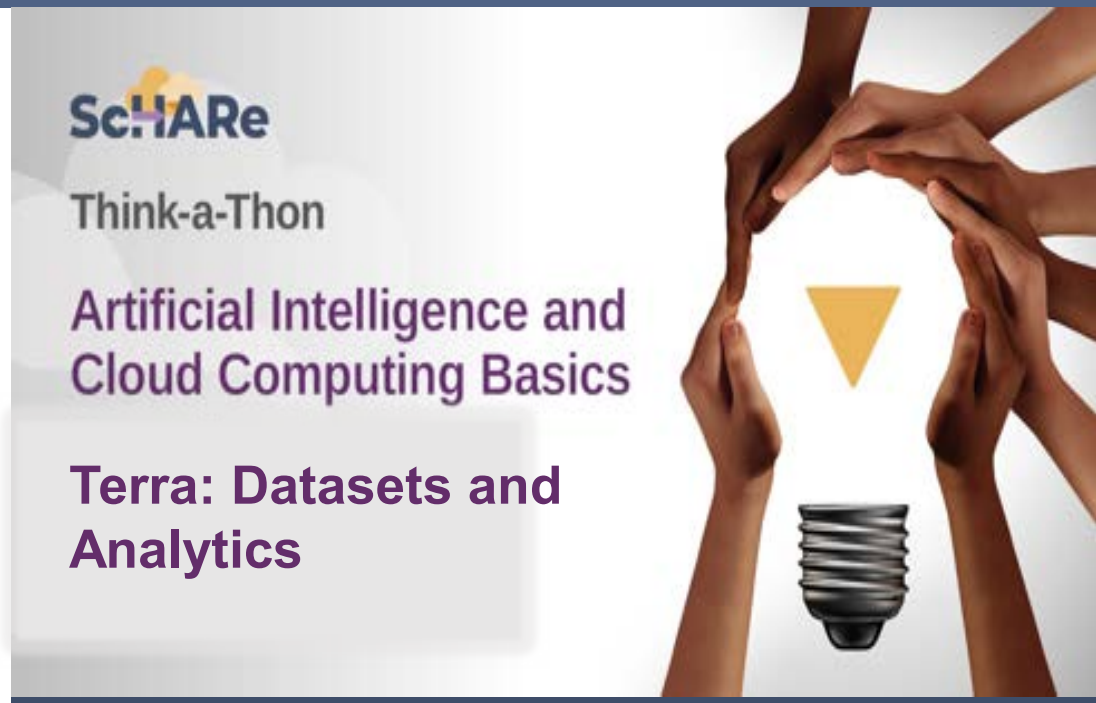
Scannable from your screen!

nimhd.nih.gov/schare

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**



Register:

bit.ly/think-a-thons

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

Account setup

We have registered you for ScHARe

You can choose not to use your account. If you prefer to be removed at any time,

email us at

schare@mail.nih.gov

With your consent, you have been:

- **registered for ScHARe**
- **added to a free temporary billing project** that will allow you to run the event materials with your instructors

➢ You will be active on this billing project for the duration of the Think-a-Thon. If you want to access work-in-progress after this time, you will need to set up your own billing and copy your workspaces to it

In preparation for the Think-a-Thon

We want to make sure that everyone:

1. has provided their Gmail address and has been registered for ScHARe
2. can create and set up their Terra account with our help

The next two slides provide instructions on how to do so

for users who could not attend our Think-a-Thon today

Registering for ScHARe

Normally, you would have to 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

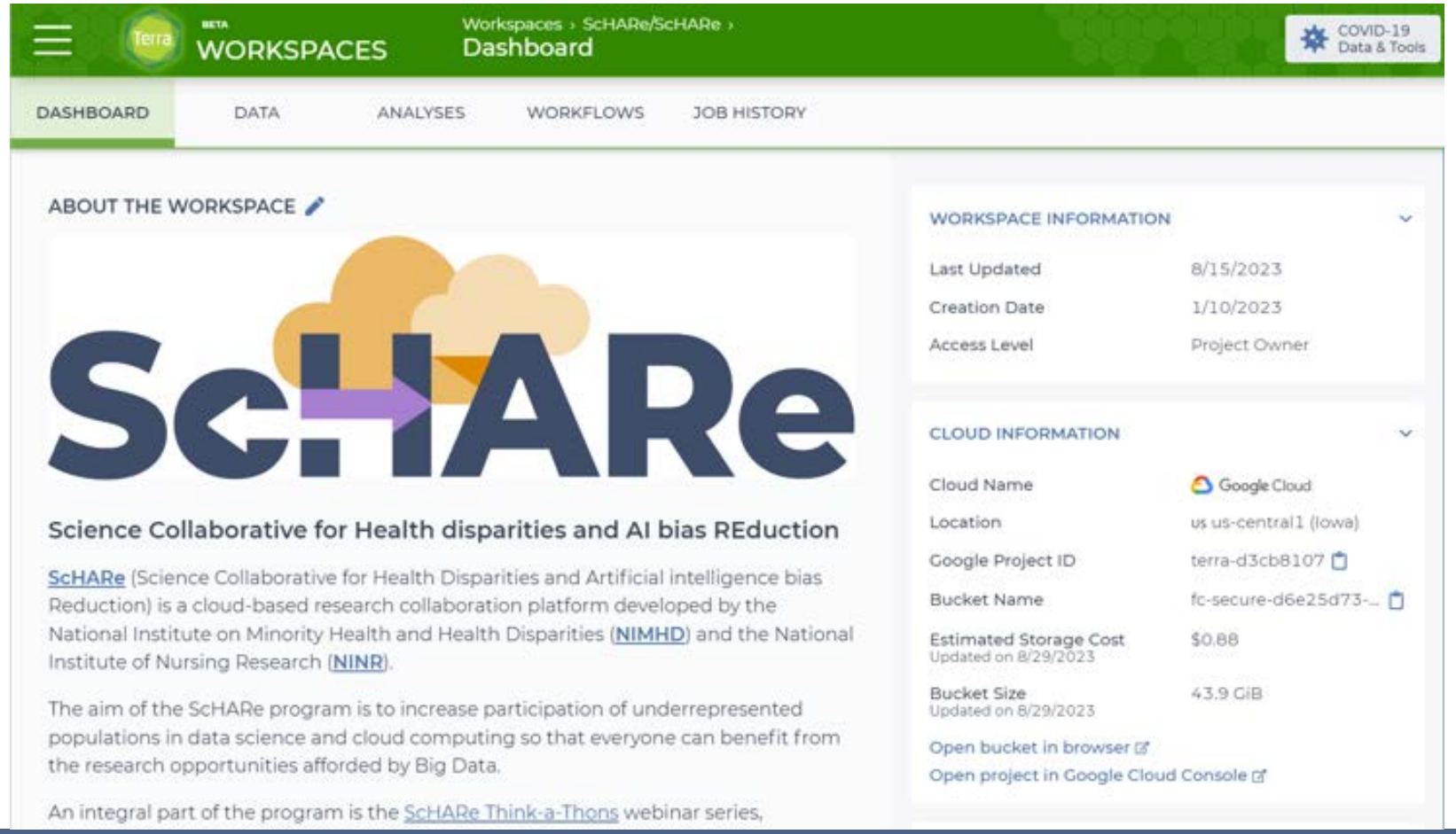
They are like **computational sandboxes** with everything you need to complete your project: data, analysis tools, documentation

Please paste the address below in your browser:

bit.ly/access-schare

If you have already created a Terra account and are logged in, you will see this:

bit.ly/access-schare



If you have not logged in, or have not yet created a Terra account, you will see this:

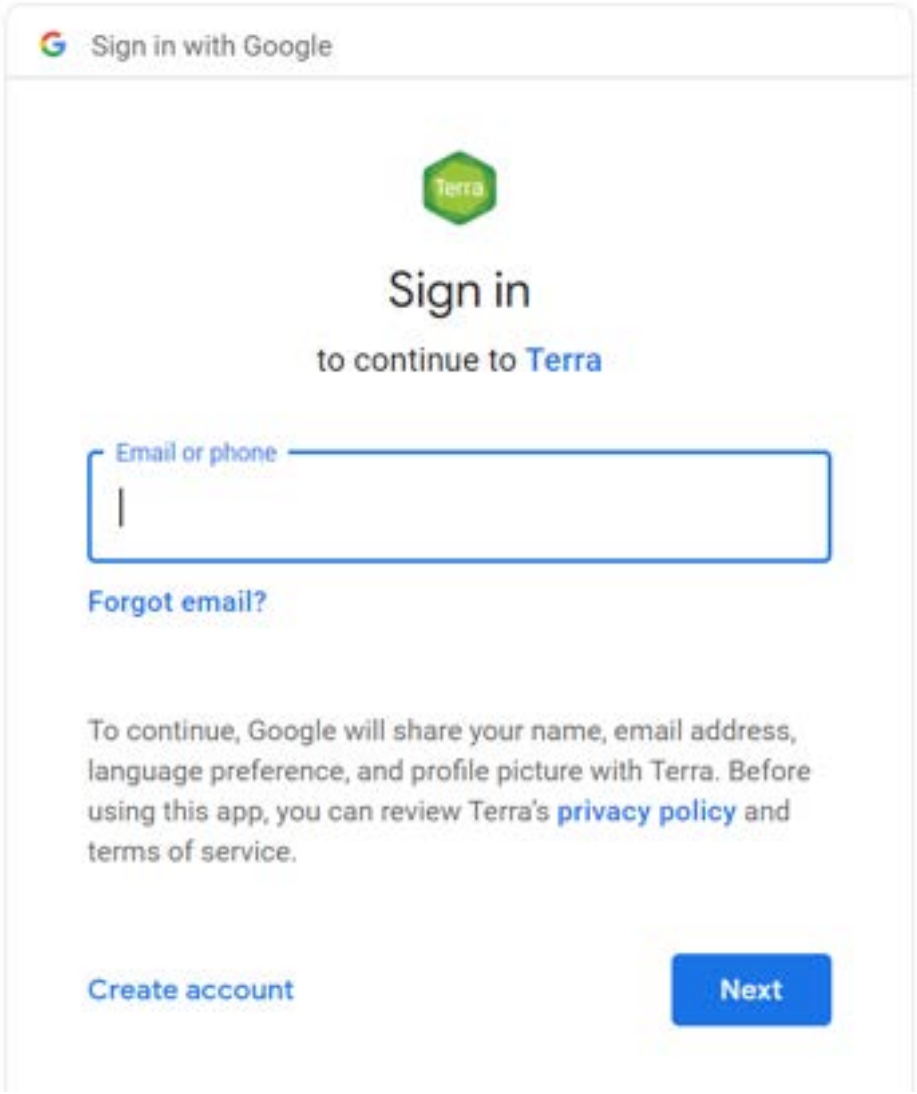
bit.ly/access-schare

Click on the login button:

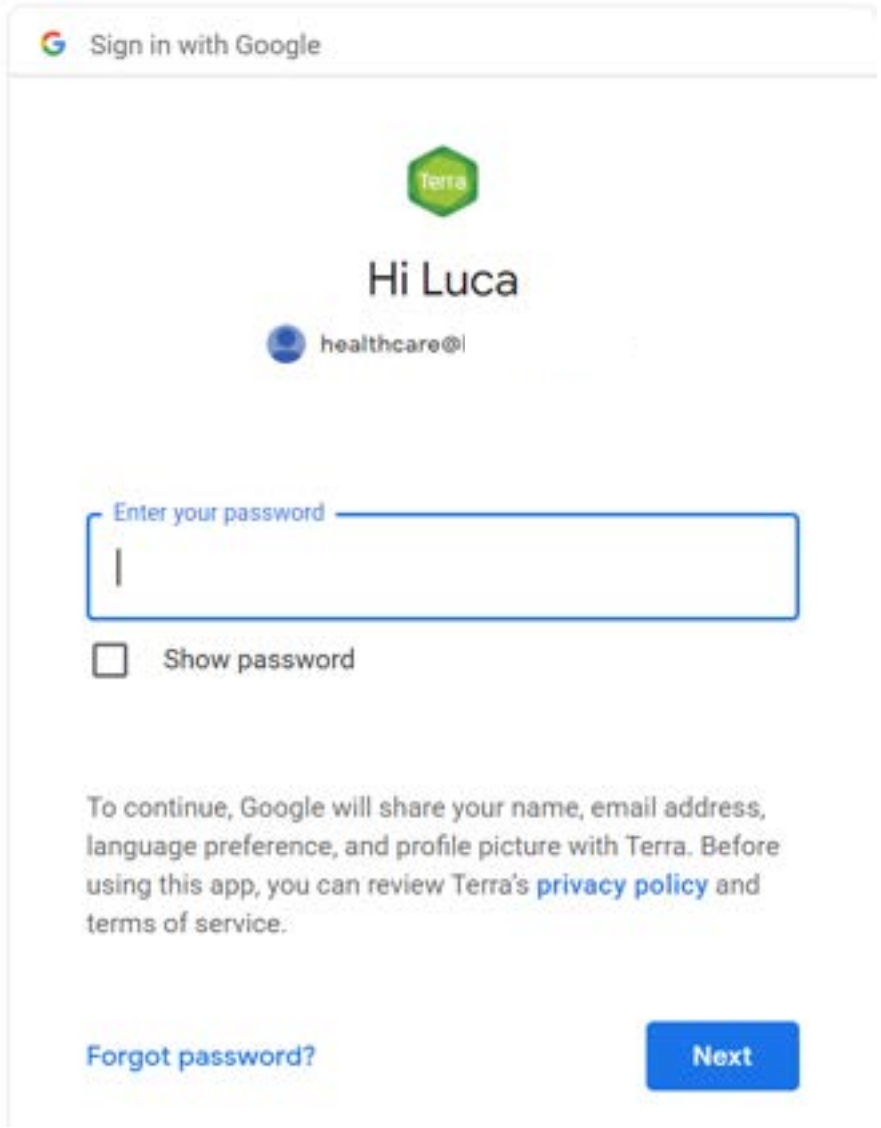
bit.ly/access-schare

Use the Gmail address you provided us with to log in:

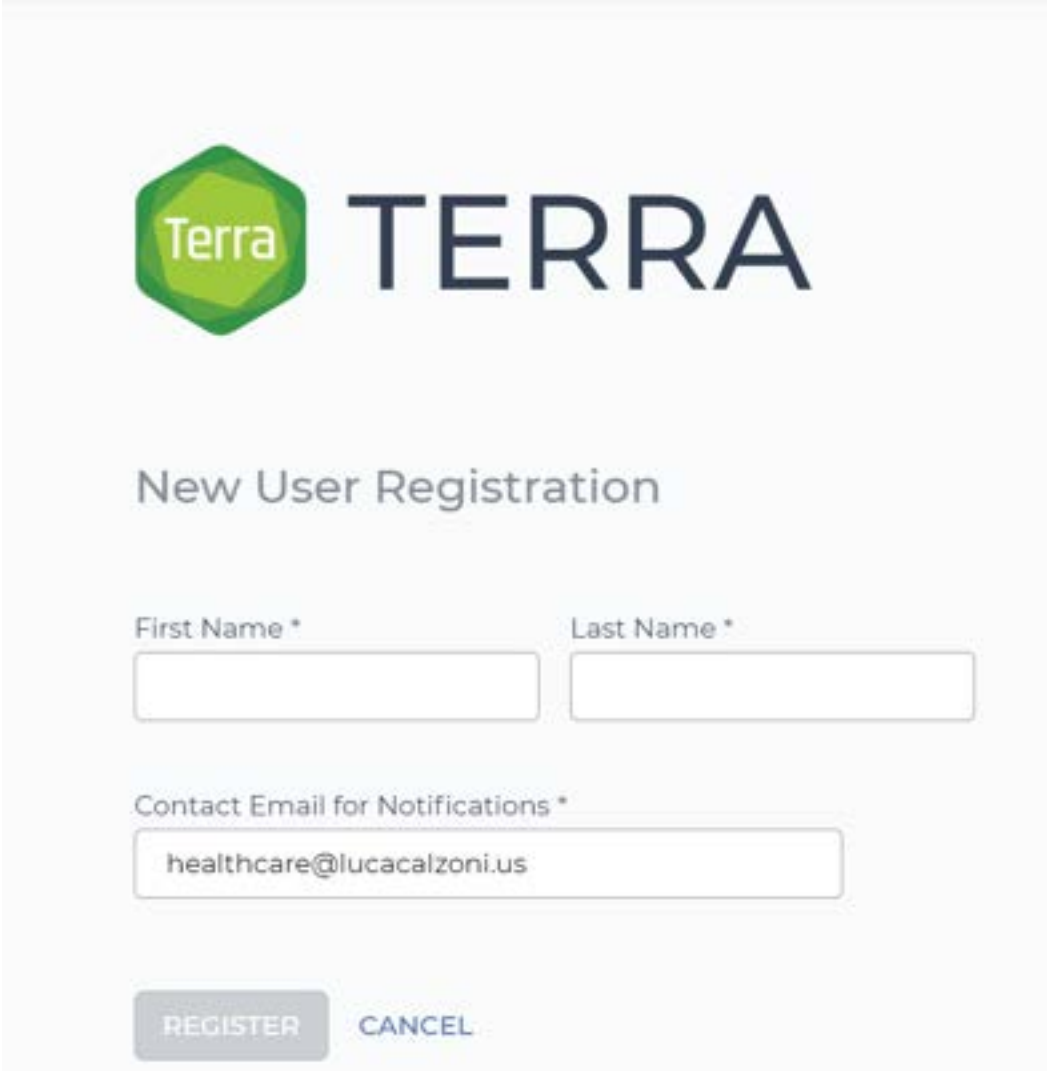
Use the Gmail address you provided us with to log in:



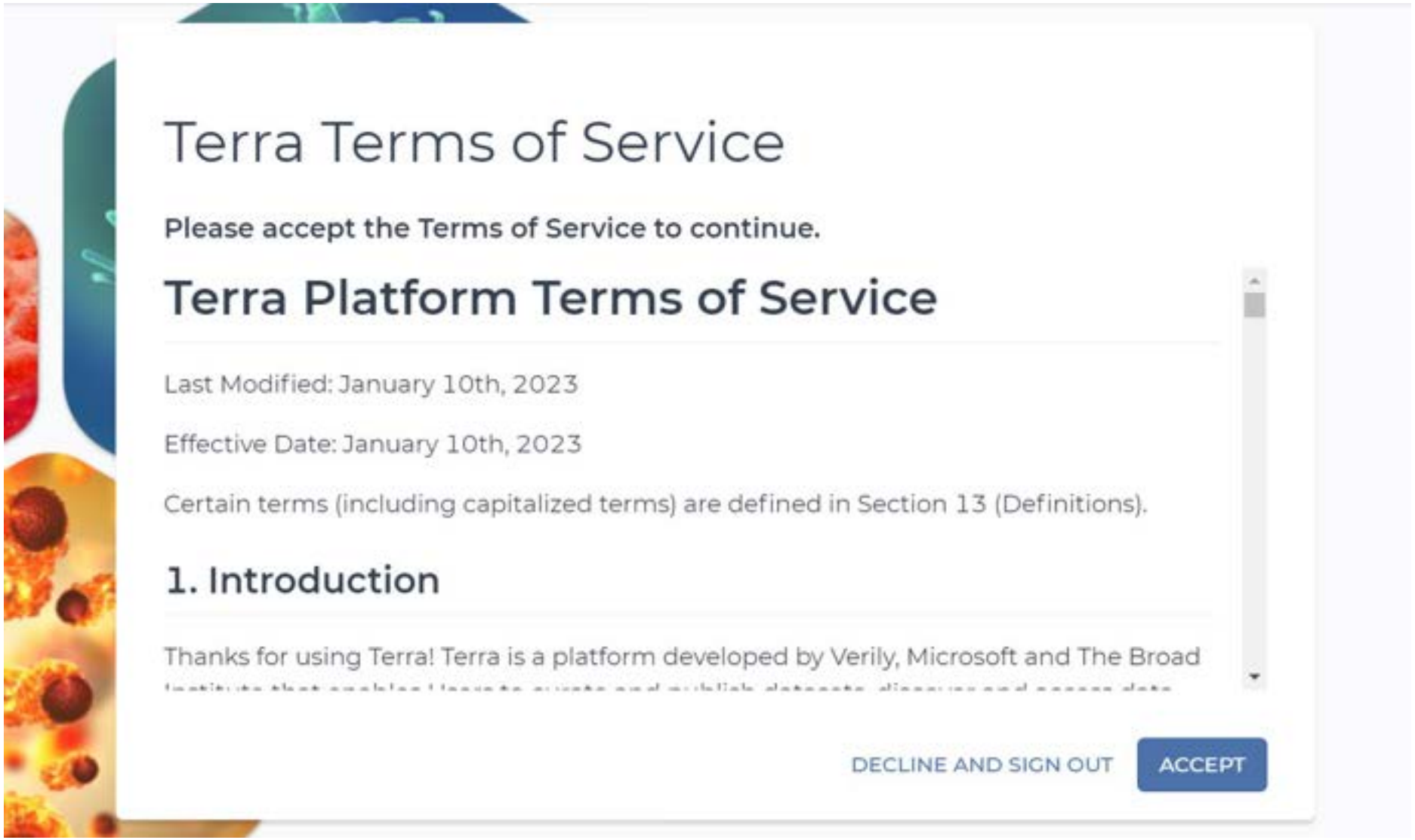
Input the password associated with your Gmail account:



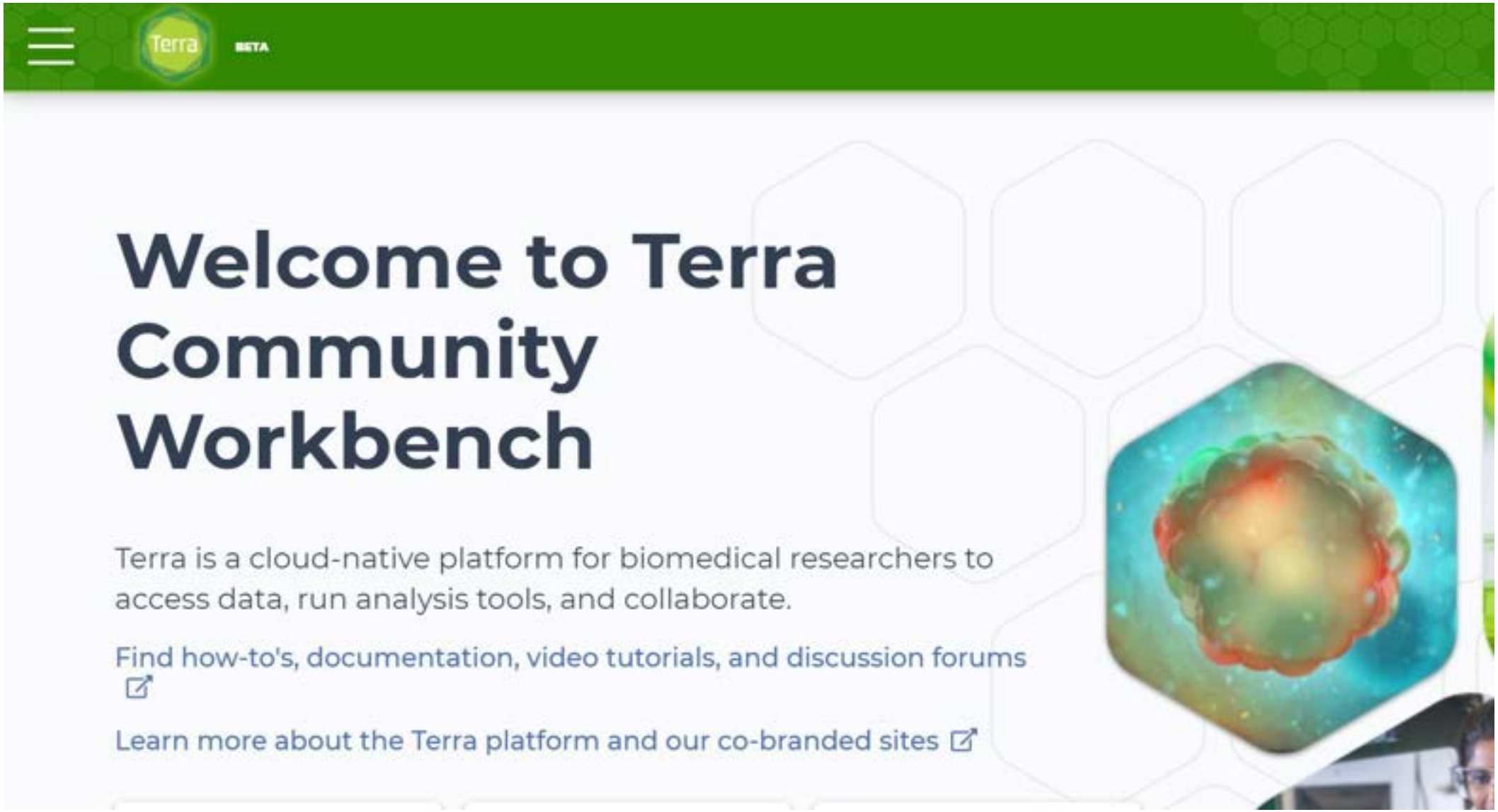
If you are new to Terra, create an account now:



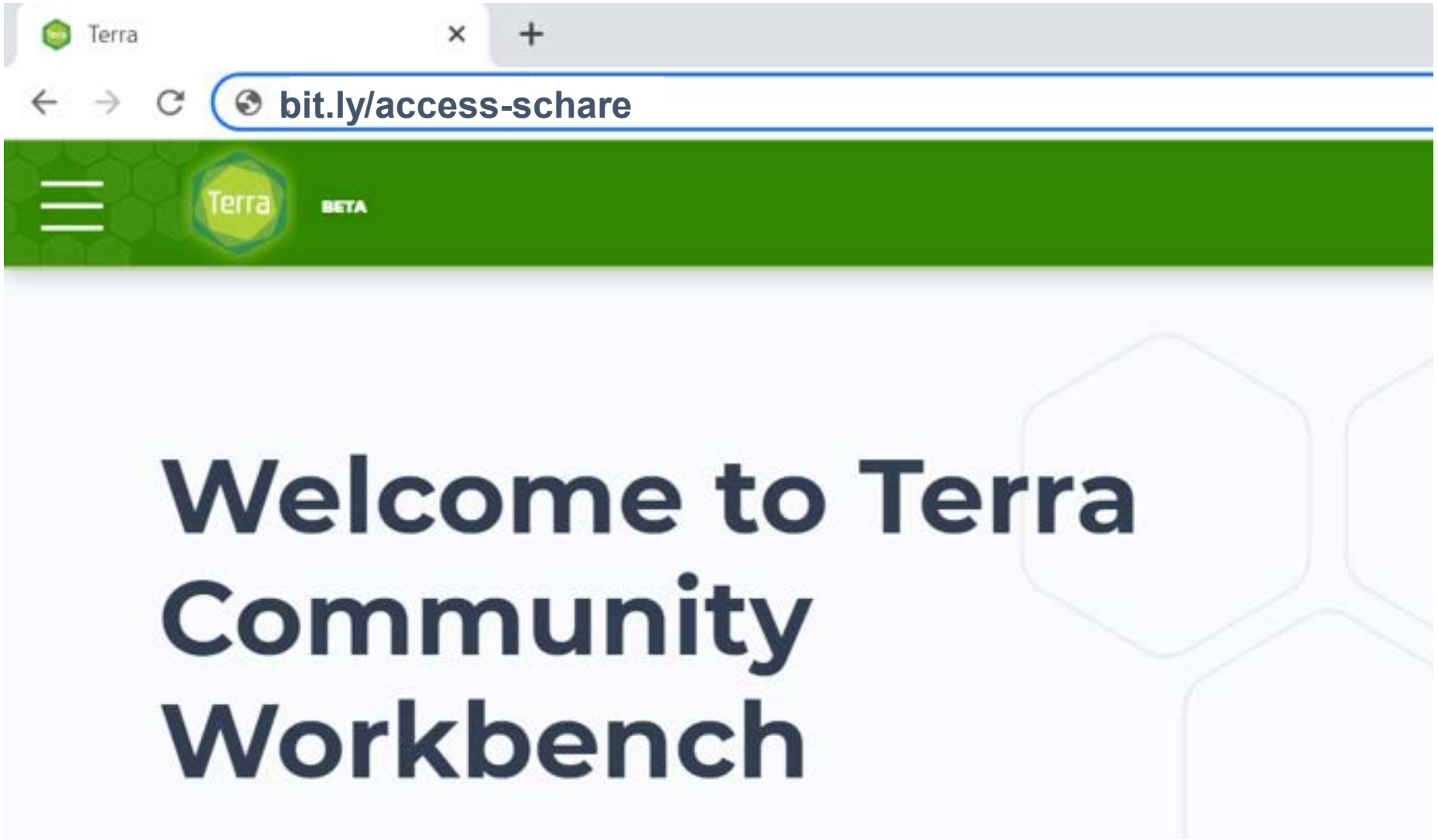
Accept the Terra Terms of Service:



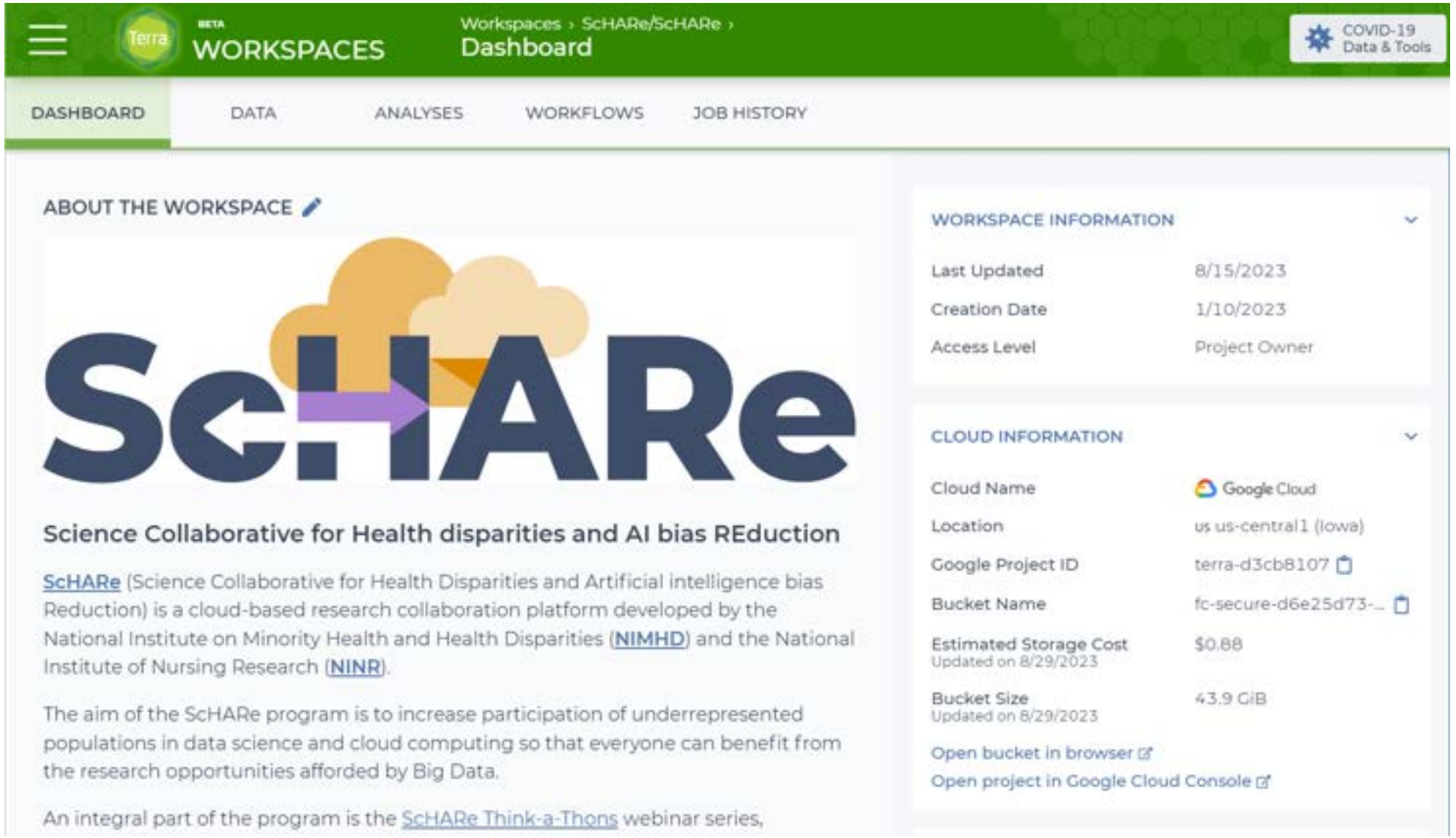
You will see this welcome page:



Paste this address in your browser: bit.ly/access-schare



All users should see this:



Workspaces and permissions

What is a workspace

Workspaces are the building blocks of Terra - a dedicated space where you and your collaborators or students (for the educators among us) 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 research or classroom project: data, analysis tools, documentation

You can use workspaces to:

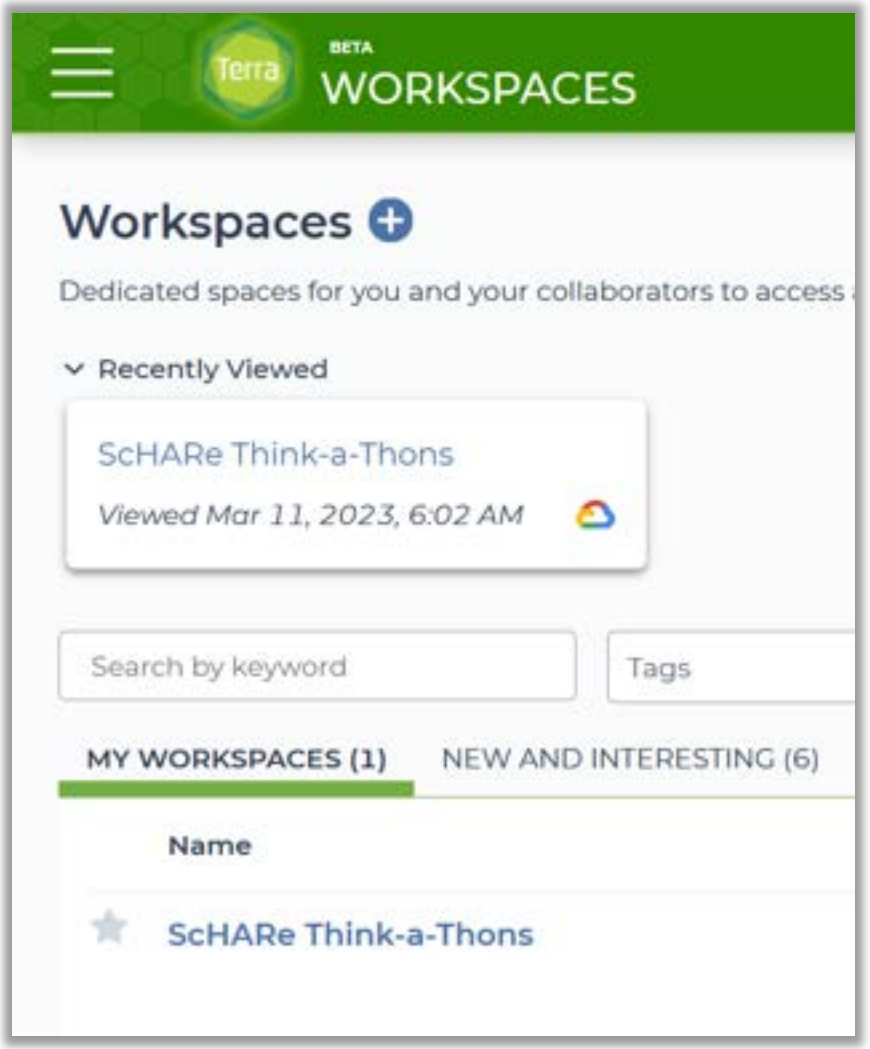
- **Link to data in the cloud** for analysis, instead of downloading and storing it yourselves
- **Combine data** from different sources in a single table for analysis
- **Keep data organized with integrated spreadsheet-like tables** - no matter where in the cloud the data are stored
- **Visualize and analyze data** in real time using Python and R (and soon, SAS)
- **Find and run bulk analysis tools** (workflows) even if you're not programming experts
- **Share reproducible analysis results**
- **Collaborate** while maintaining control of your resources

Creating a workspace

Let's create your first Terra workspace!

For the purpose of this tutorial, we will assume that you intend to create a workspace that will allow you to work with students or researchers on a collaborative project

Watch our demonstration, and replicate our steps later (you will need to set up your own billing, following the instructions provided in the *Billing and costs* section)



Sharing a workspace with collaborators

Let's now see how you can share your workspace with a group of students
(for the educators among us) to work with them on a collaborative project

Requirements

Students must be able to:

- access your data
- perform computations
- work with you to write and edit the interactive notebooks you are using to collaborate

First, let's create a group listing all of our student collaborators.
We will then share the workspace with them and configure billing

Cloning (copying) an existing workspace

Why clone?

- If you are **interested in using the data resources of a workspace, or in replicating the analyses showcased in its notebooks**, and have the appropriate permissions to do so, you can create a copy of such workspace for use by you and your students
- This operation is called **“cloning” the workspace**

You are encouraged to clone the ScHARe workspace and use its resources!

Let’s see how you can do it.

Notebooks and environment

What is a notebook?

A Jupyter Notebook is an interactive analysis tool that includes:

1. **code cells** for analyzing and visualizing data in real time (Terra notebooks support **Python or R**)
2. **documentation** to make it easier to explain, share and reproduce your analyses

If you or your collaborators or students (for the educators among us) are not familiar with **programming**, the code in our notebooks is very easy to understand and reuse. Our tutorials in the notebooks and our Think-a-Thons will also help you understand how notebooks work

We will:

- cover the basics of **creating your first notebook**
- **explore the instructional notebooks** available in the ScHARe workspace and run one of them

Why use notebooks?

A notebook can be a great instructional tool: it integrates code and its output into a single document where you can write and 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 teaching materials **more transparent, understandable, repeatable, and shareable**
- will make it **easier to communicate and share your work** with your students

The ScHARe notebooks

You can see examples of what a notebook can do by checking out the instructional notebooks that **ScHARe offers to help novice users** learn how to use the platform

A list of the available notebooks is provided on the right. **We will also access them online**, as an example

First, let's see how you can create a notebook!

List of ScHARe instructional notebooks

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

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

Sources
www.quanthub.com/python-for-data-science/
coursera.org

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

supports **both Python and R**

Source
en.wikipedia.org

Why Python?

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)**

Source
stackify.com/learn-python-tutorials/

How to learn Python

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**

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 and is very accessible to beginners

Unfamiliar terminology may be a barrier, which today's workshop will hopefully help you overcome

Links to additional **free learning resources** will be provided at the end

Python 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

Python resources

Stackify

30+ Tutorials to Learn Python

Top 30 Python Tutorials

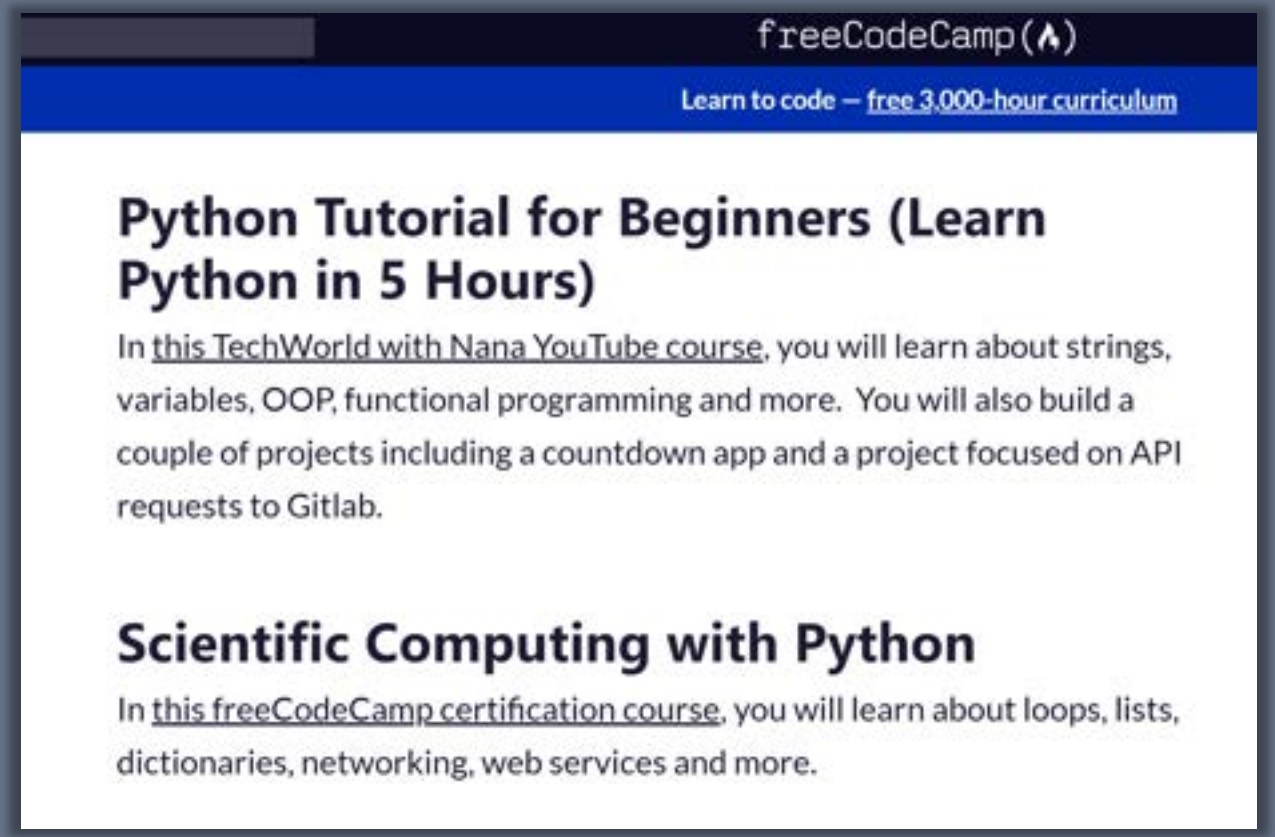
In this article, we will introduce you to some of the best **Python tutorials.** These tutorials are suited for both beginners and advanced programmers. With the help of these tutorials, you can learn and polish your coding skills in Python.

1. Udemy
2. Learn Python the Hard Way
3. Codecademy
4. Python.org
5. Invent with Python
6. Pythonspot
7. AfterHoursProgramming.com
8. Coursera
9. Tutorials Point
10. Codementor
11. Google's Python Class eBook
12. Dive Into Python 3
13. NewCircle Python Fundamentals Training
14. Studytonight
15. Python Tutor
16. Crash into Python
17. Real Python
18. Full Stack Python
19. Python for Beginners
20. Python Course
21. The Hitchhiker's Guide to Python!
22. Python Guru
23. Python for You and Me
24. PythonLearn
25. Learning to Python
26. Interactive Python
27. PythonChallenge.com
28. IntelliPaat
29. Sololearn
30. w3Schools

Python resources

FreeCodeCamp

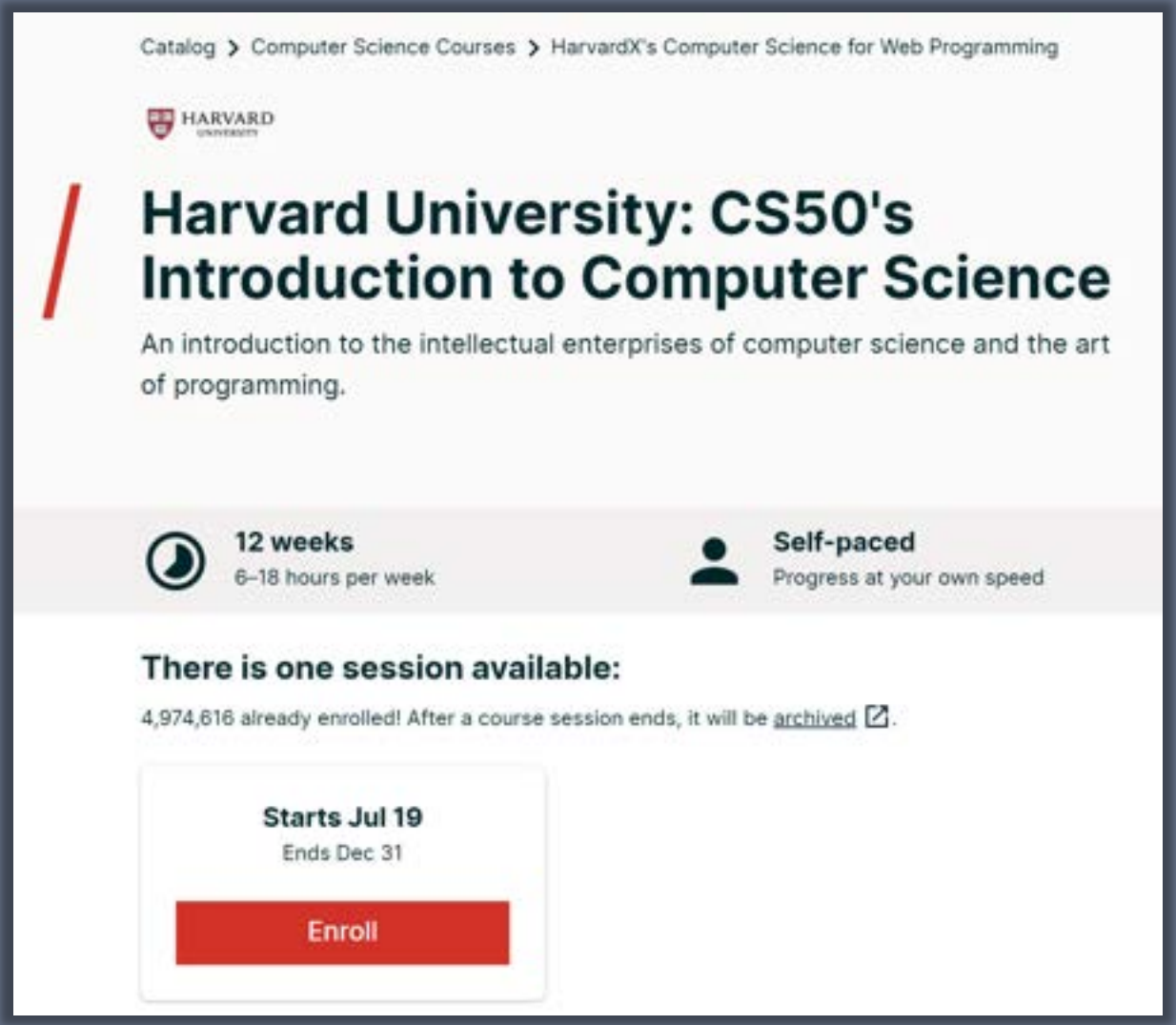
Code Class for Beginners



Python resources

Harvard

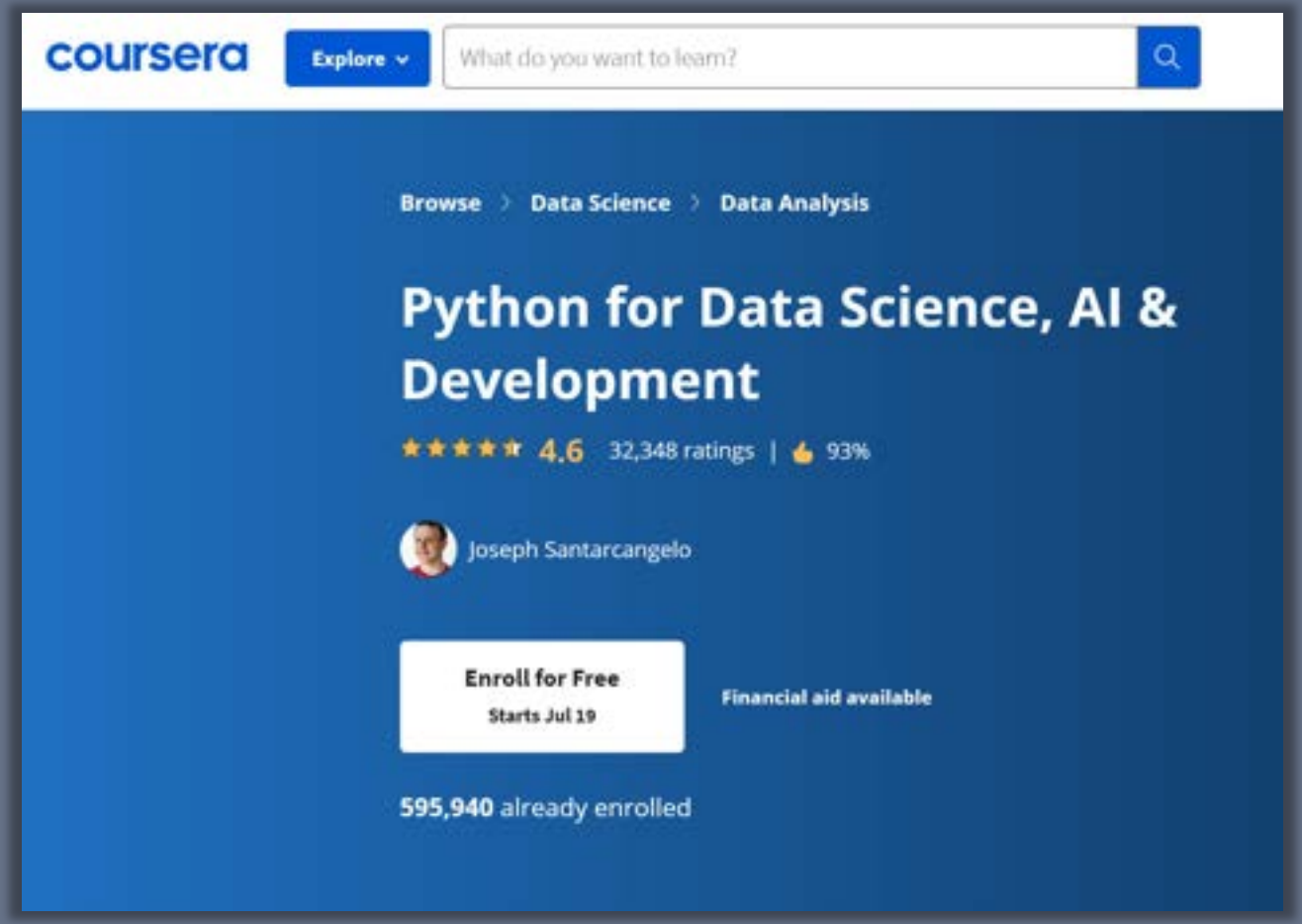
Free Python Course



Python resources

Coursera

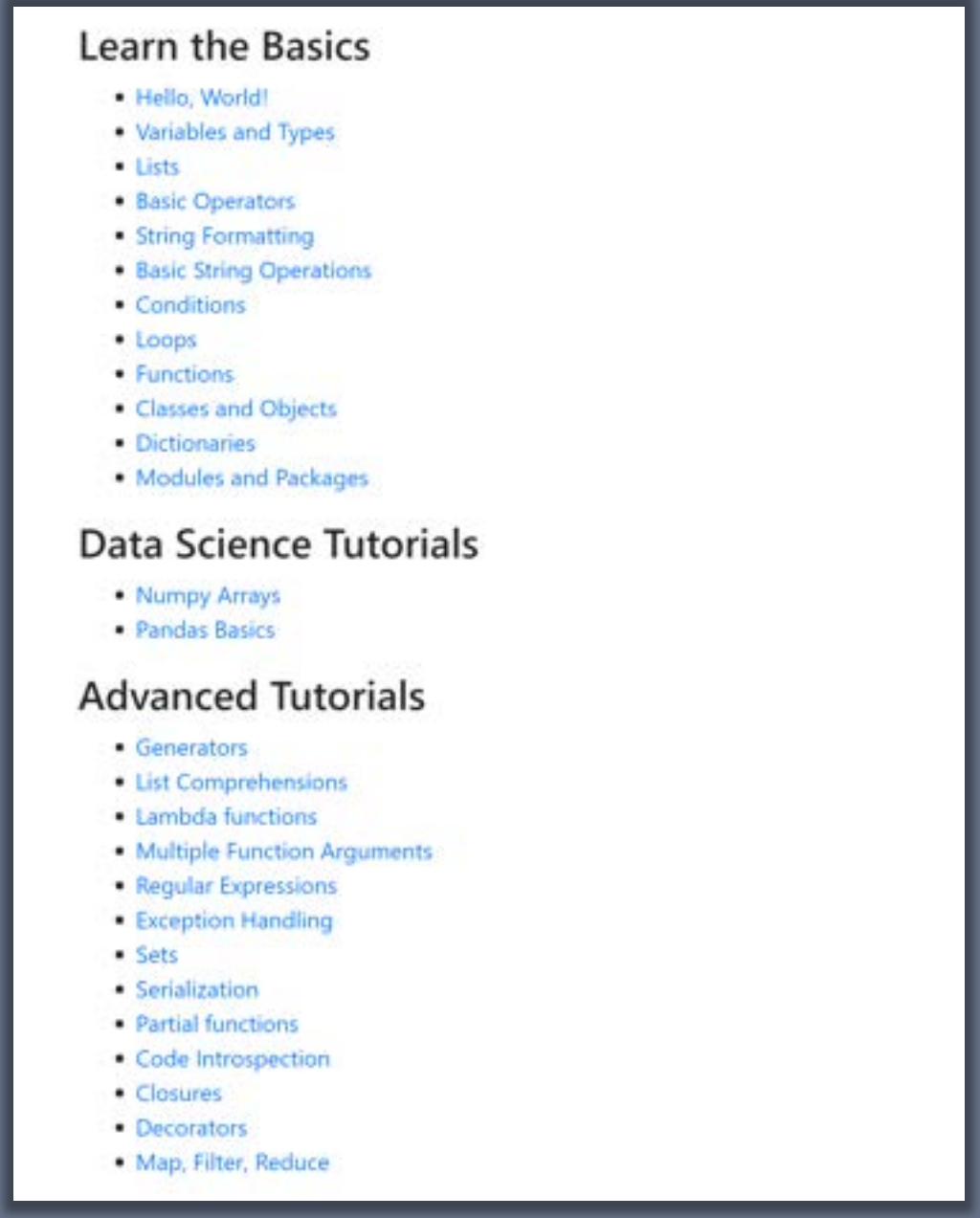
Free and Paid Python Courses



Python resources

LearnPython

Free Interactive Python Tutorials



Learn the Basics

- Hello, World!
- Variables and Types
- Lists
- Basic Operators
- String Formatting
- Basic String Operations
- Conditions
- Loops
- Functions
- Classes and Objects
- Dictionaries
- Modules and Packages

Data Science Tutorials

- Numpy Arrays
- Pandas Basics

Advanced Tutorials

- Generators
- List Comprehensions
- Lambda functions
- Multiple Function Arguments
- Regular Expressions
- Exception Handling
- Sets
- Serialization
- Partial functions
- Code Introspection
- Closures
- Decorators
- Map, Filter, Reduce

Python resources

BestColleges

10 Places to Learn Python for Free

Terra resources

If you are new to Terra, we also 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

Datasets

ScHARe Data

On ScHARe, you can work with:

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

The ScHARe Data Ecosystem is comprised of:

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

   Example: *American Community Survey (ACS)*

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

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

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

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

SciHARe Ecosystem: Google hosted datasets

Examples of interesting datasets include:

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

ScHARe Ecosystem: ScHARe hosted datasets

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

200+ datasets

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

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

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

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

ScHARe Ecosystem: ScHARe hosted datasets

Examples of datasets for each category include:

Education access and quality

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

Examples:

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

ScHARe Ecosystem: ScHARe hosted datasets

Health care access and quality

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

Example:

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

ScHARe Ecosystem: ScHARe hosted datasets

Neighborhood and built environment

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

Examples:

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

ScHARe Ecosystem: ScHARe hosted datasets

Social and community context

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

Example:

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

ScHARe Ecosystem: ScHARe hosted datasets

Economic stability

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

Examples:

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

ScHARe Ecosystem: ScHARe hosted datasets

Health behaviors

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

Examples:

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

ScHARe Ecosystem: ScHARe hosted datasets

Diseases and conditions

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

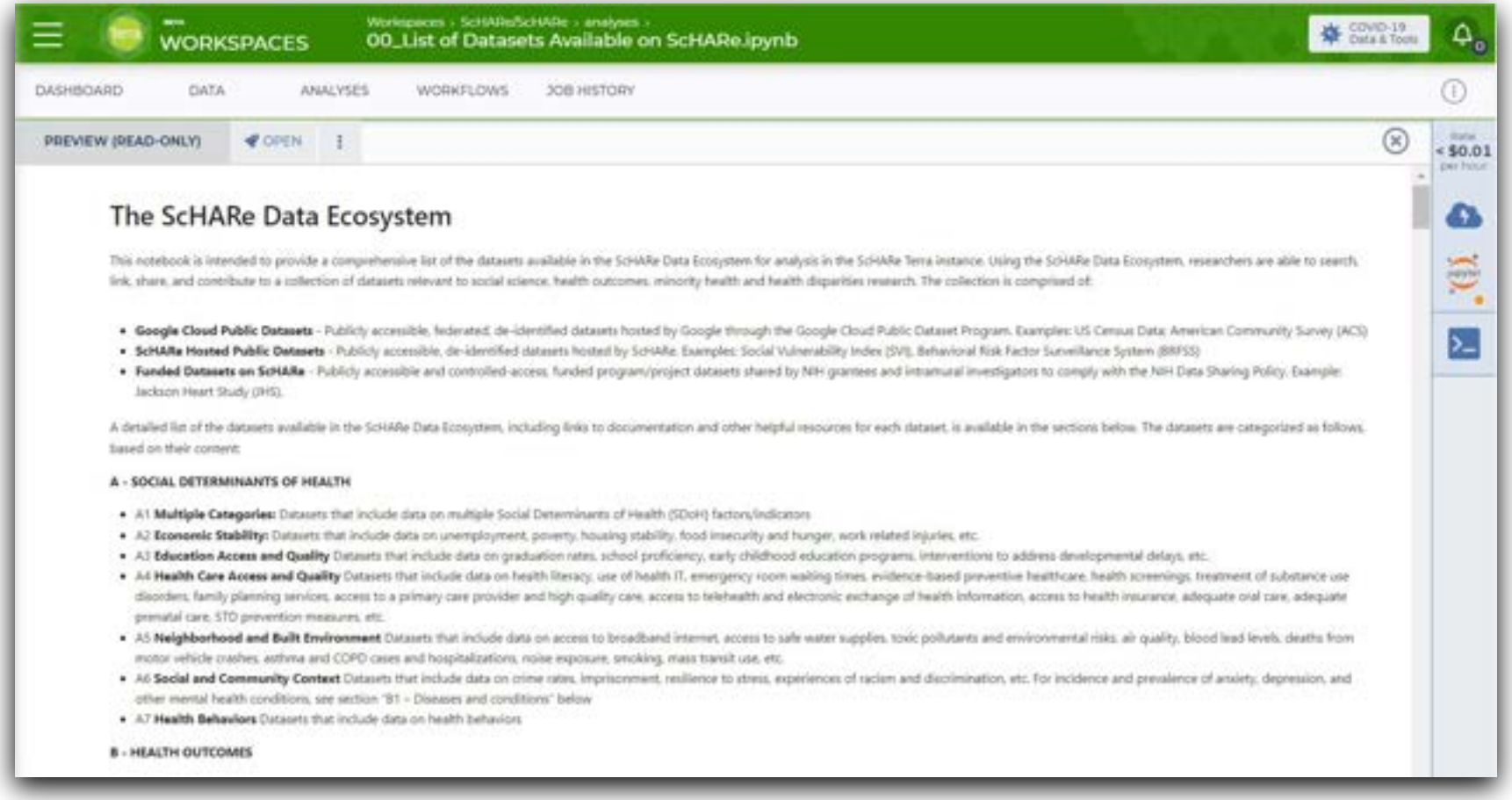
Examples:

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

How to check what data is available

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

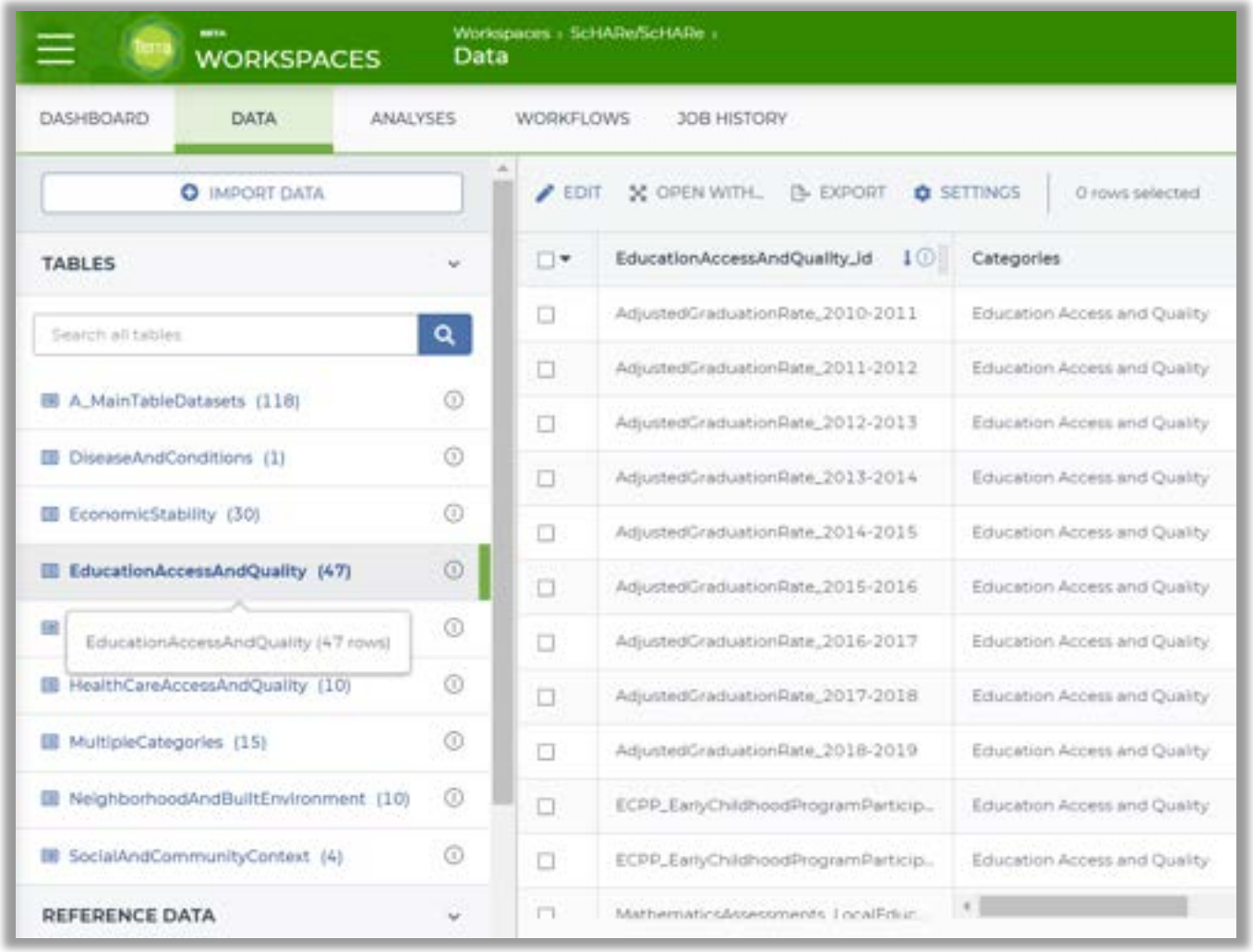


How to access available data

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



How to upload data to your workspace

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**

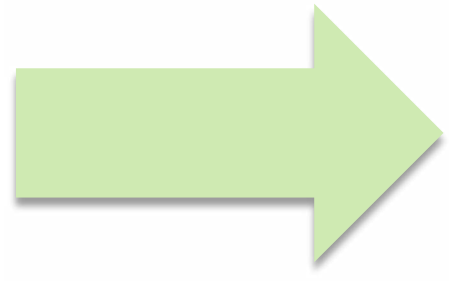
- We will use the **Python** programming language to work with the data
- Notebooks in the **Analyses** section of the ScHARe workspace explain how to use **R** instead

What data will we work with?

Data you upload

to your workspace

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



To give you an example today, **we will:**

1. **download a sample dataset** (MHSVI)
2. **upload it** into a 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

Replicate our steps later

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:

How to work with Google hosted data

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

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

Replicate our steps later

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:

How to work with ScHARe hosted data

ScHARe hosted data tutorial

We will focus on the category of datasets 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

We will work with BRFSS data

BRFSS is the nation's premier system of health-related telephone surveys that collect state-level data about U.S. residents regarding their:

- health-related **risk behaviors**
- chronic **health conditions**
- use of **preventive services**

State health departments use in-house interviewers or contract with telephone call centers or universities to administer the BRFSS surveys **continuously through the year**

BRFSS data is used to:

- help establish and track state and local health objectives
- plan **health programs**
- implement **disease prevention** and **health promotion activities**
- monitor **public health trends**

Replicate our steps later

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. **copy and run the following notebook** and complete the steps illustrated by the instructors:

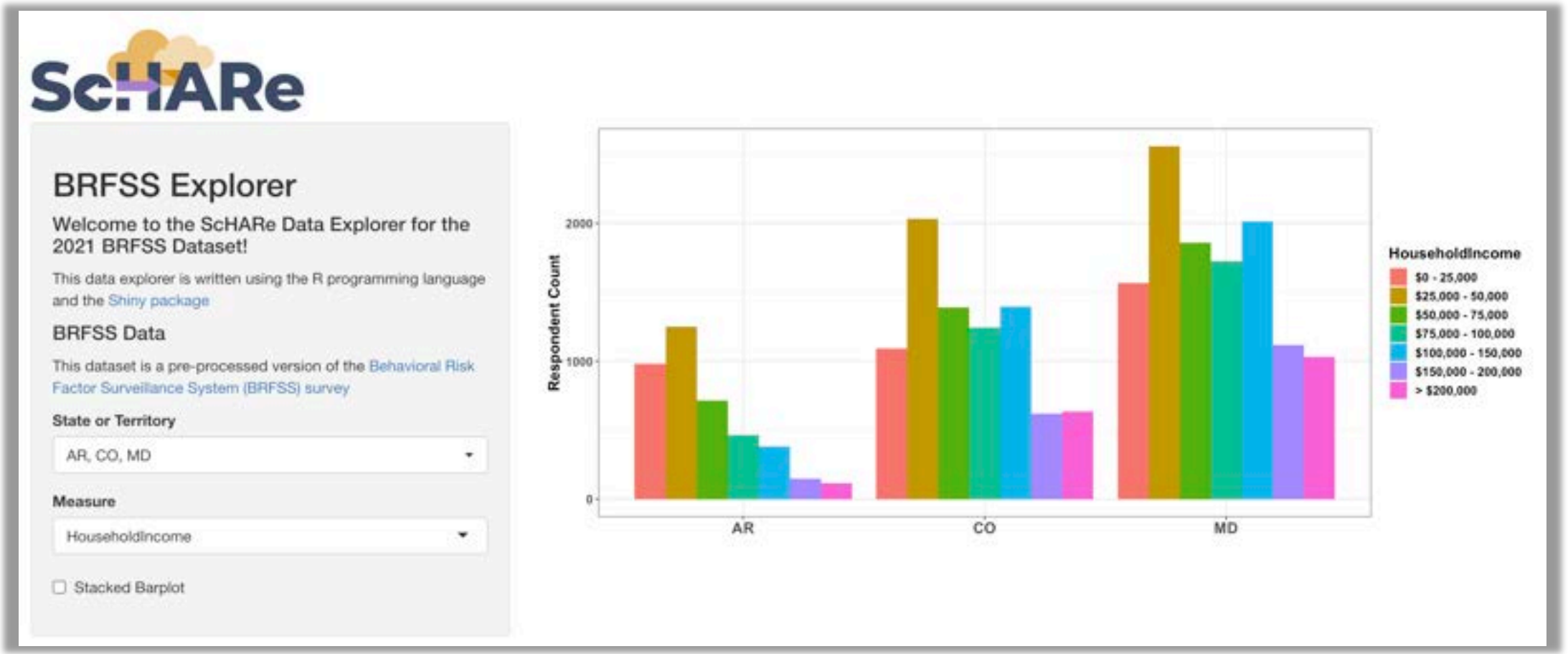
Instructional materials with step-by-step instructions and videos will be posted online here:

bit.ly/think-a-thons

BRFSS Data Explorer demo

Introducing the BRFSS Data Explorer

We leveraged the resources offered by Terra on ScHARe to build a **ScHARe Data Explorer** for the 2021 BRFSS dataset



Introducing the BRFSS Data Explorer

What we used:

- **A package (tool) for R called Shiny:**
  - You can use Shiny to develop **interactive web applications** for data exploration and visualization without any previous experience
  - The applications can be shared as **just code or as pre-built containers**
  - They can be run on a local machine or on a web server as **standalone web pages or dashboards**

R Shiny applications take interactive data visualization to the next level!

They gained popularity as tools to make custom data visualizations with the dashboards tracking the COVID-19 pandemic

R Shiny apps can be launched on ScHARe **from Terra's built-in RStudio environment**

Let's now see a demo!

Data exploration poll

With the BRFSS Data Explorer in mind, what other features would be helpful in your day-to-day use of such a data visualization tool?

Billing and costs

What are the cloud costs of working on Terra?

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

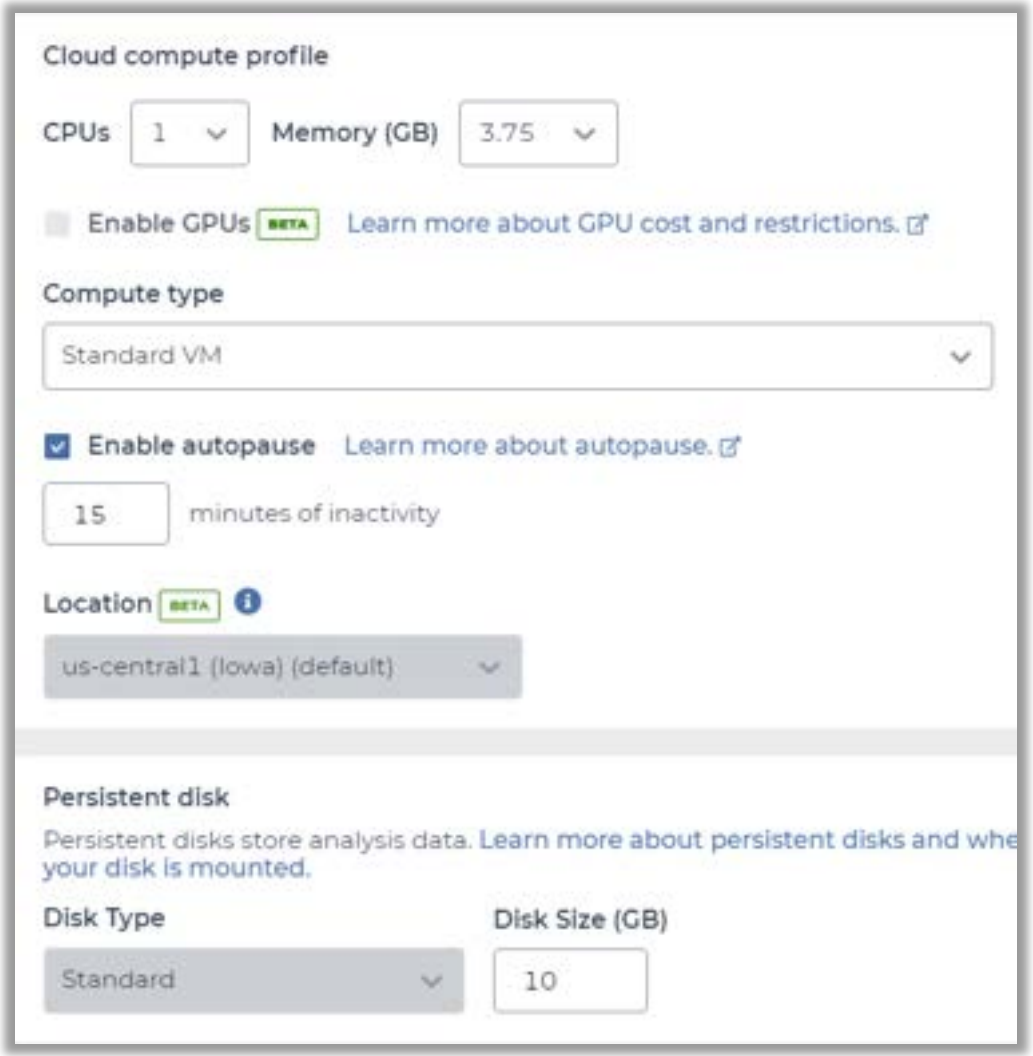
However, the following operations in Terra **may incur charges:**

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



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

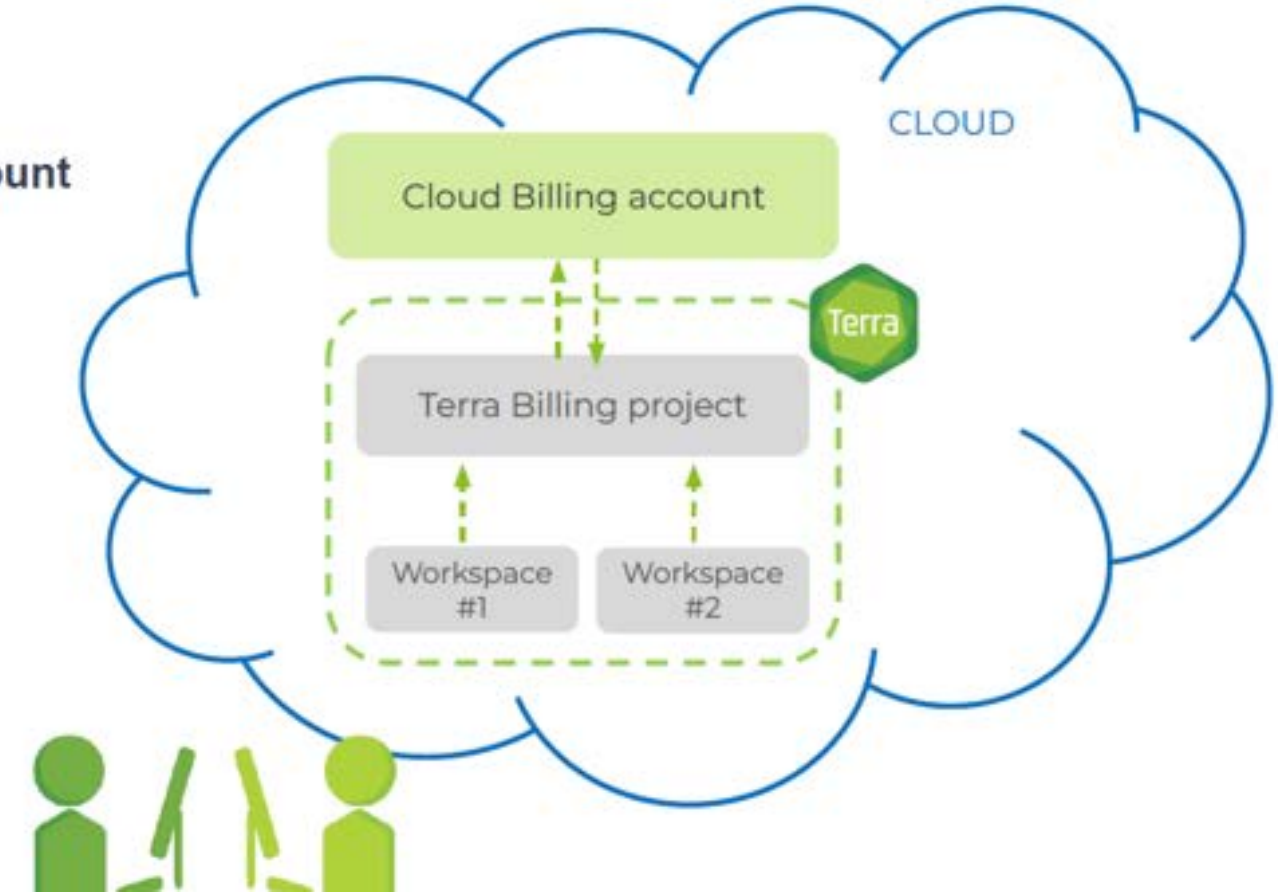
- When creating a bucket to store data, you are asked to set its location. This is because the data are going to be stored in data warehouses located in physical places ("**regions**" – more info here). Regions exist, among other reasons, to accommodate the need of certain users to keep their data in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

How will I be charged for these costs?

Since Terra runs on Google Cloud Platform (GCP), all Terra costs are GCP fees. You pay these fees through a **Google Cloud Billing account** linked to a **Terra Billing project**

- **You are only charged once and by Google directly, not twice by Terra and Google**

- 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, **access to a free temporary billing project** will allow you to run all the materials with your instructors

What happens after today?

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

How do I set up billing?

For the Educators among us

In a scenario where you plan to **work with the students in your class** on a collaborative project (approximate cost: $2 to $10/analysis/student), billing can be set up in **two ways**:

1

You as **the teacher (or your institution)** set up a billing account and share it with the students

2

Each student sets up their own billing account and is eligible for **$300 in free Google Cloud credits**

Next, we will guide you through the **steps needed to set up billing on Google Cloud first and then Terra**

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** and the next 3 slides

Step 1. Set up a Google Cloud Billing account

1. Go to the **Google Cloud console** at https://console.cloud.google.com/ and sign in with your Terra user ID. If you haven't already set up a billing account, you'll be invited to activate your free Google Cloud credits
2. Click the **activate** button and follow the instructions
3. You'll need to verify your identity with a one-time verification sent to a cell phone, and give a credit card, PayPal account or bank account. **You won't be billed until the free credits expire**
4. **Verify** the Google Cloud Billing account in the Billing page. You should see **My Billing Account** in the top left
5. Google will create a project, **My First Project**, funded by your free credits, in the **My Projects** tab

Step 2. Link the Cloud Billing account to Terra

The next step is to link the Google Cloud Billing Account to your Terra account, so that **Terra and Google can communicate** about cost and billing

You must use the same Google ID for both the Cloud Billing account and your Terra user name

1. When logged into Google with your Terra user ID, go to the **Google Cloud Console Billing page**.
2. Select the **checkbox beside the Google Cloud billing account** you will use for Terra.
3. On the right panel, below **Permissions**, select the **Add Principal** button.
4. Add "**terra-billing@terra.bio**" under **New Principal** in the form.
5. In the dropdown, select the role **Billing > Billing Account User.**
6. Click **Add**.
7. Click on the **Save** button

Note: "terra-billing@terra.bio" will appear in the list as "terra-billing@firecloud.org." This is expected.

Step 3. Create a Terra Billing project

Once Terra is linked to a Google Cloud Billing account, you can create a Billing project on Terra, which will **allow you to create a workspace to store and analyze data**

1. Go to the **Billing page** from the main navigation (click on **your name** to expand the drop-down, and select **Billing**)
2. Click on the **"+ Create"** button at the top left
3. If prompted to **Enable Billing Permissions**, select the **Google identity** of the Google Cloud Billing account, and click **Allow**. This lets Terra access Cloud Billing accounts associated with your Terra user name (Google ID).
4. Enter a **unique name** for your Terra Billing project
5. Select the **Google Cloud Billing account** that will fund the Billing project

You may see multiple Cloud Billing accounts that you can select for this Terra Billing project. If you need to locate a Billing account ID, navigate to the **Google Developers Console** and click on **Billing**. Look for the number below **Billing account ID**

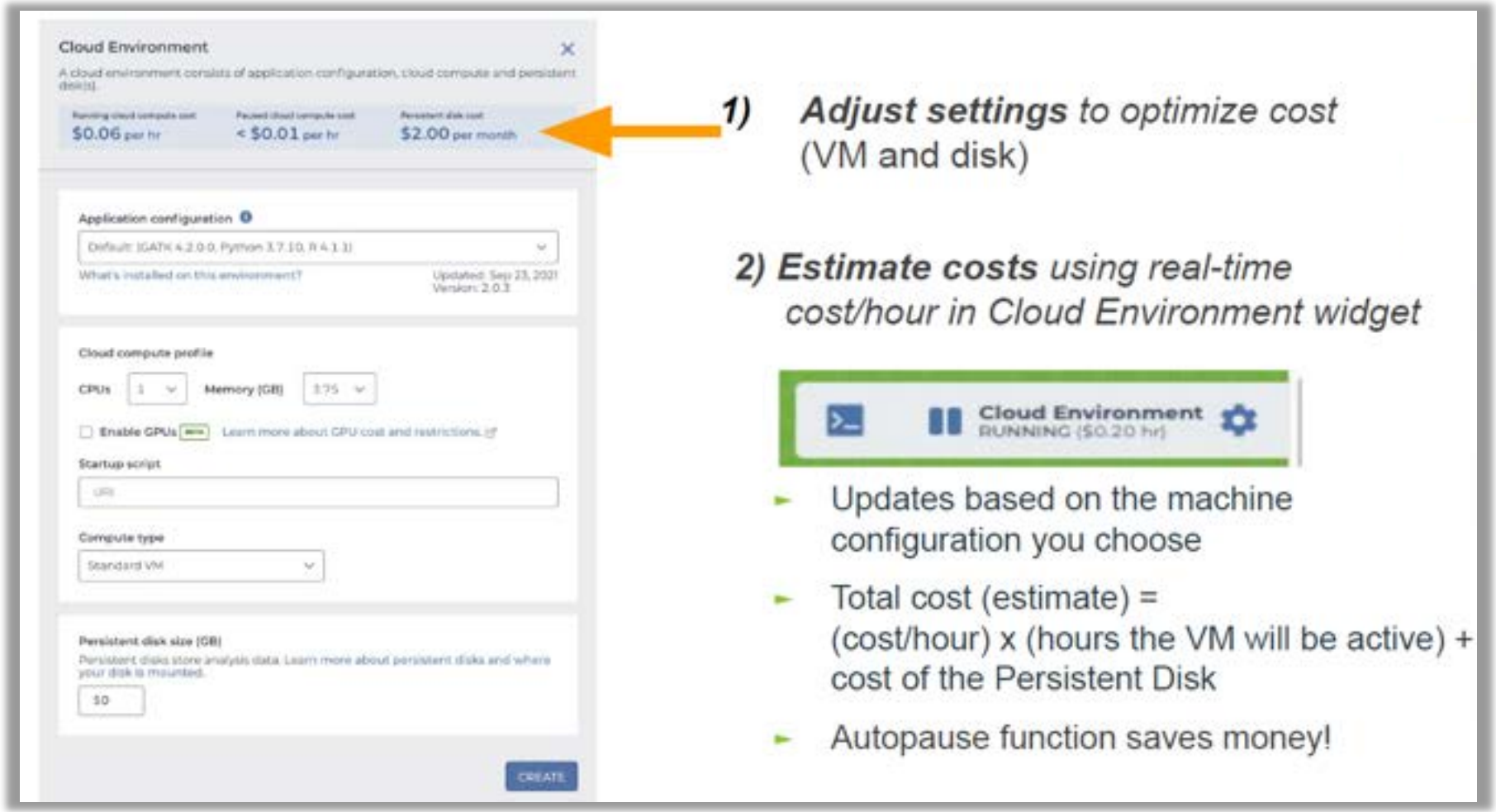
Understanding and monitoring costs

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)



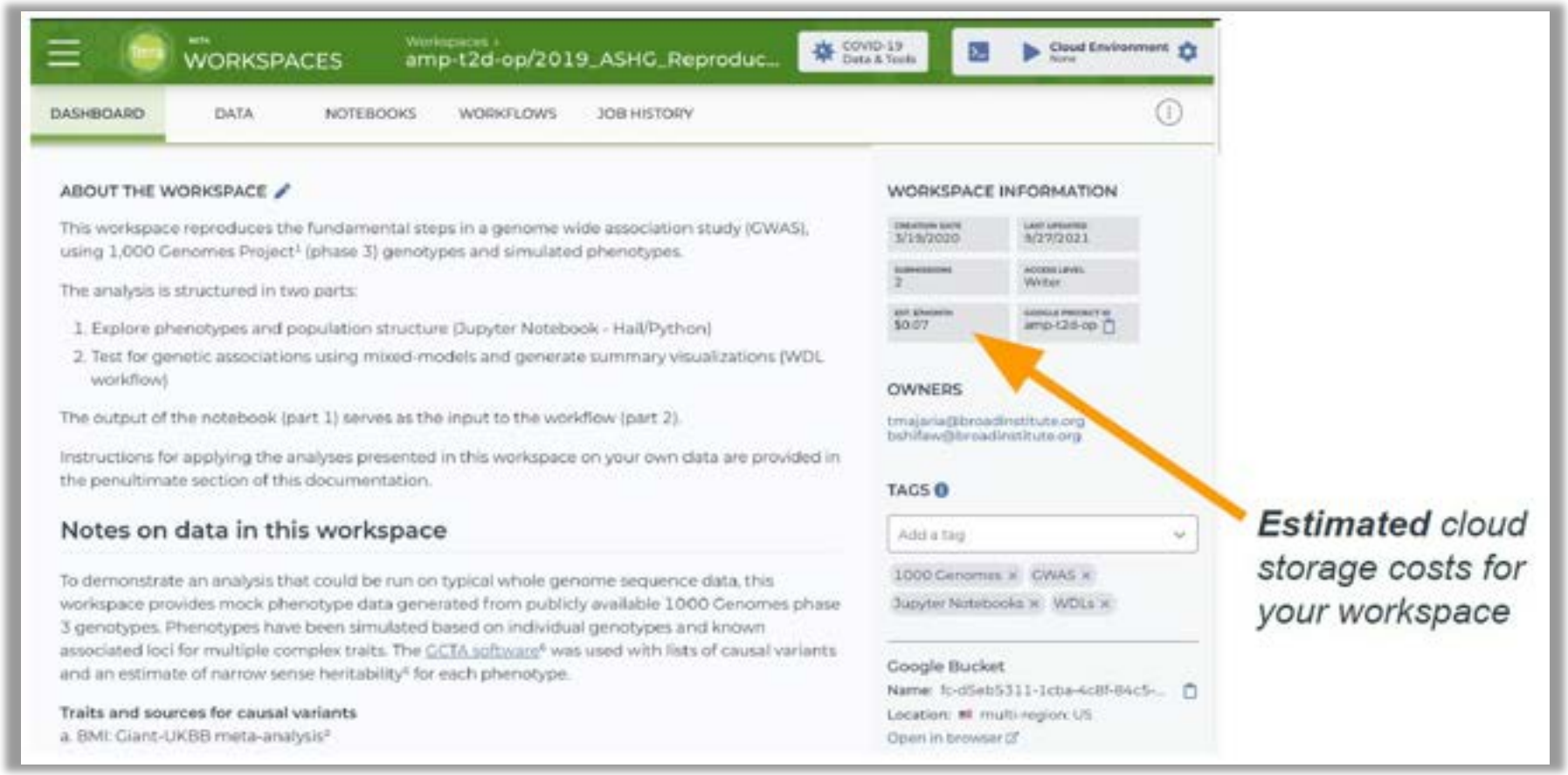
Understanding and monitoring costs

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 (for advanced users)



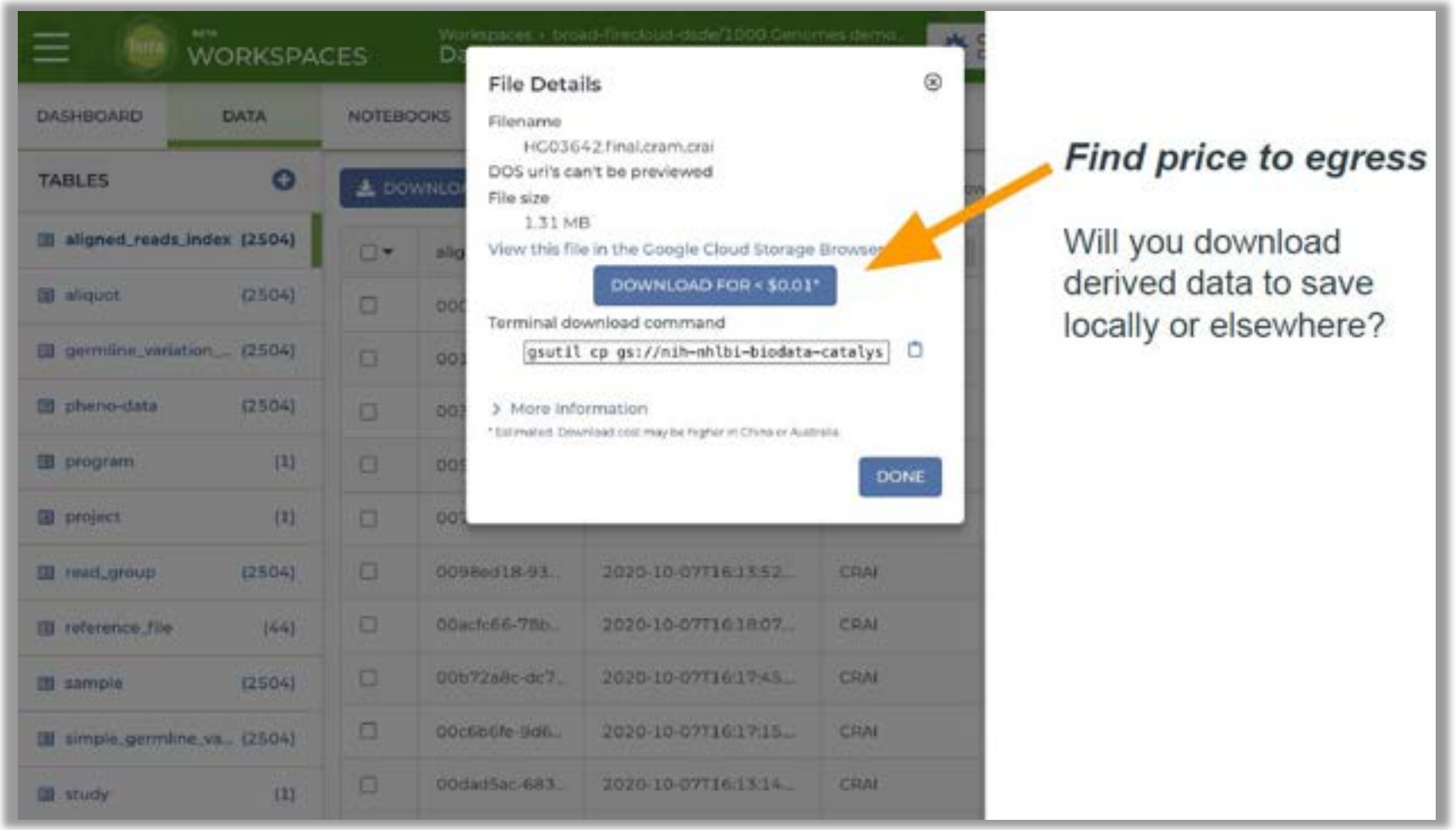
Understanding and monitoring costs

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 (for advanced users)



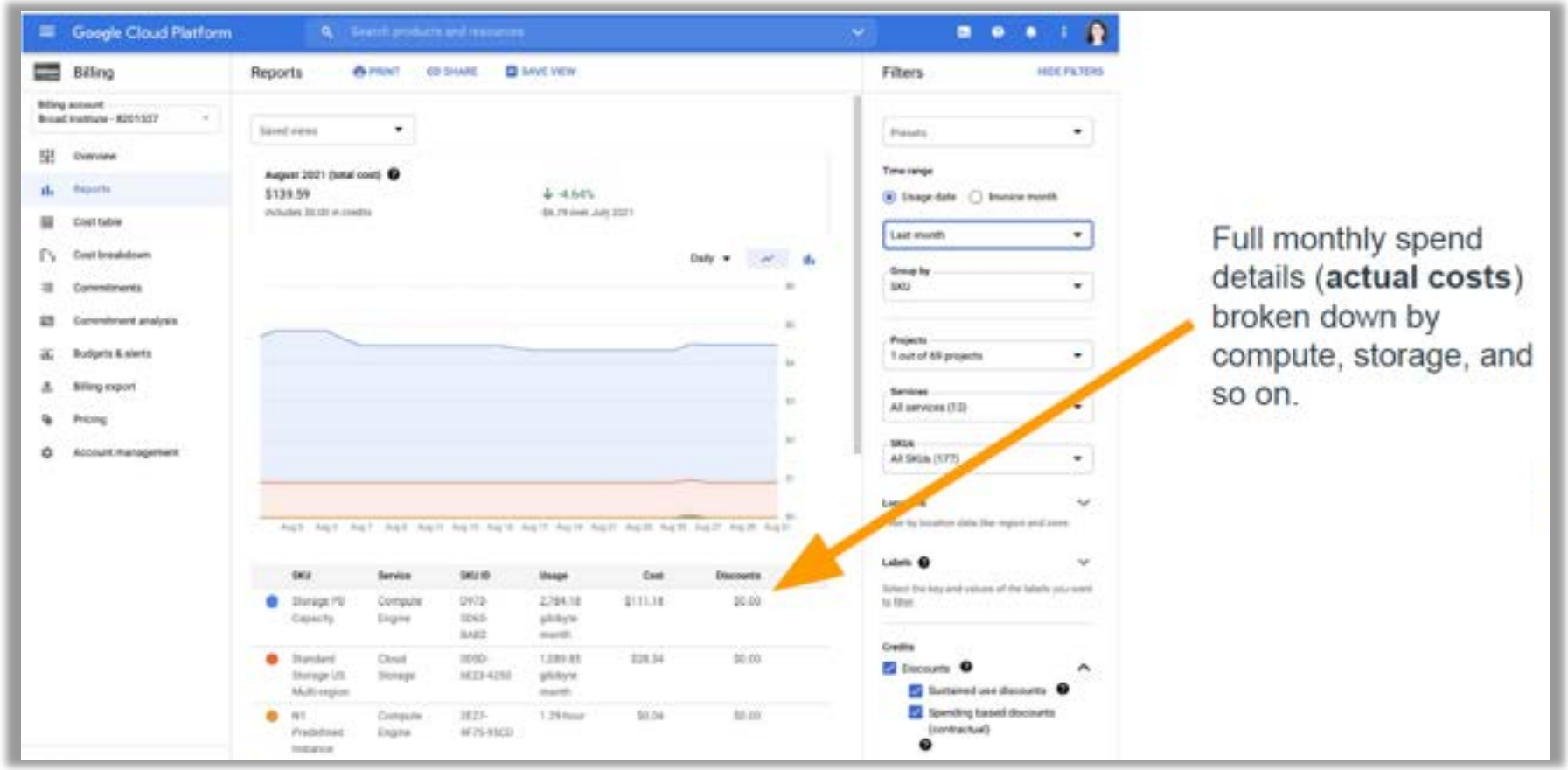
Understanding and monitoring costs

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 (for advanced users)



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

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