March 15th, 2023 Think-a-Thon

ScHARe I
Terra Accounts and Workspaces

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 about 1 day after the Think-a-Thon

If you want to access work-in-progress from the Think-a-Thon after this time, you will need to set up your own billing and copy any of your workspaces to your own billing

Science collaborative for Health disparities and Artificial intelligence bias Reduction

SciHARe


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

NIMHD OCPL

RLA

BioTeam

STRIDES

Terra

Broad Institute

SIDEM

CCDE Working Group

Outline

5’ Introduction and setup
- Experience poll

15’ ScHARe and Terra overview
- Interest poll

20’ Registration and Account

30’ Billing and Costs

20’ Workspaces and Permissions

30’ Notebooks and Environment

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	☐	☐	☐	☐

Part I
ScHARe and Terra overview

ScHARe

ScHARe is a cloud-based population science data platform designed to accelerate research in health disparities, health care outcomes, and artificial intelligence (AI) bias mitigation strategies

The platform offers researchers at all career levels and disciplines:

- Access to centralized social determinants of health and other social science **datasets**
- The ability to **collaborate** as they apply AI, machine learning, and other advanced analytical techniques to these datasets in a **secure setting**

nimhd.nih.gov/schare

ScHARe

ScHARe aims to fill three critical gaps:

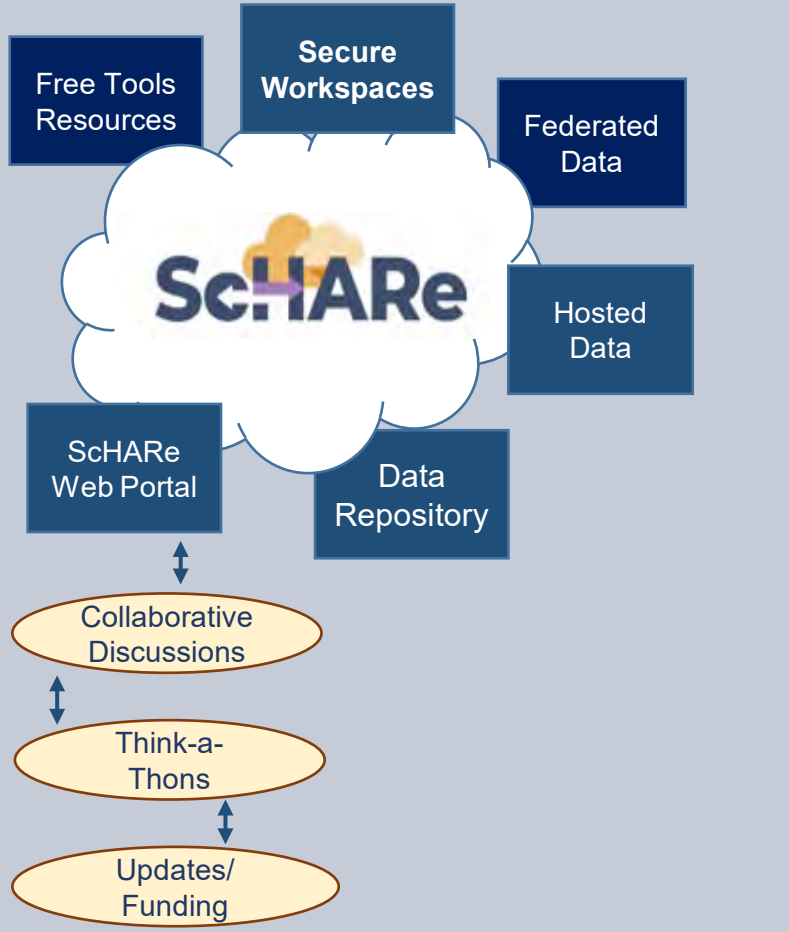
- Foster research collaborations and increase participation of **women and underrepresented populations with health disparities** in data science
- Leverage social science (SDoH) **research opportunities** afforded by Big Data and cloud computing
- Advance **AI bias mitigation** strategies and **ethical inquiry** by increasing the use of diverse eyes and skills

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 Interface, GitHub, NIMHD Web ScHARe Portal



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*

Access to social and behavioral datasets

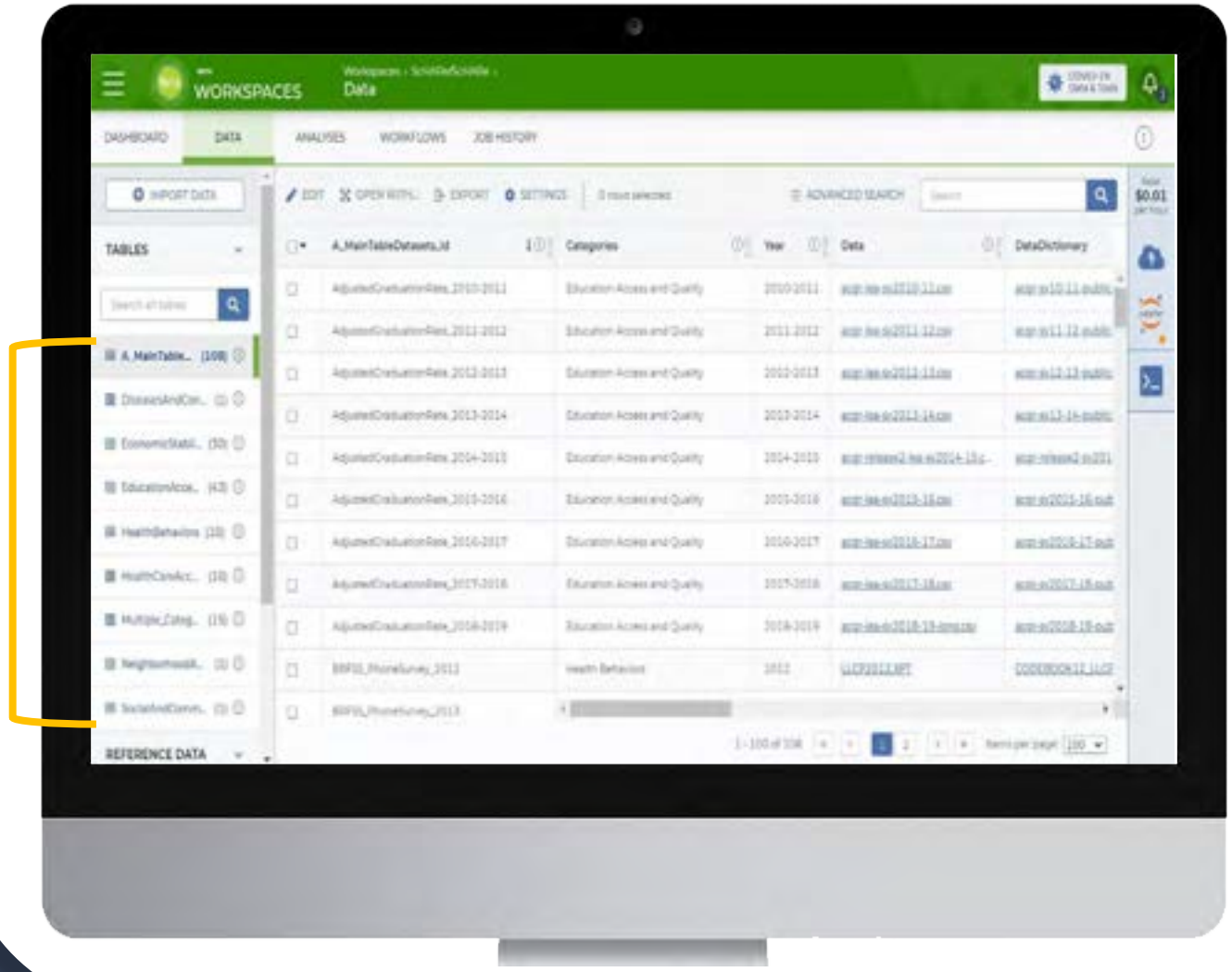
- The cloud offers access to vast repositories of data, and enables mapping and linking across sources
- Extremely large datasets are statistically analyzed to gain detailed insights, often using AI and substantial computer-processing power
- Datasets can be linked together to see how patterns in one domain affect other areas
- Data can be structured into fixed fields or unstructured as free-flowing information
- Increase use of Dark Data

ScHARe

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
- Repository for Funded Datasets on ScHARe – Compliance with NIH Data Sharing Policy

Data organization



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

Cloud computing strategies

- AI/ML can analyze massive amounts of data properly without human intervention
- Terra, standalone or in conjunction with Google Cloud Platform's Vertex AI, can support AL & ML-based analyses
- Workflows (pipelines) are a series of steps performed by a compute engine for bulk analysis
- Computing environments can be customized or standardized (using a custom Docker Image or a startup script)

SciHARe

- Uses **workflows** in Workflow Description Language (WDL), a language easy for humans to read, for batch processing data.
- **Python and R**, including libraries most commonly used
- 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

Cloud security

- The Terra system has been granted Authority to Operate as a FISMA Moderate impact system and is FedRAMP authorized
- NIST framework NIST-800-53 Rev 4 Moderate, to achieve compliance with industry-accepted security standards

SciHARe

- **Secure Workspaces**
- **Duos** used for organization and security
- **RAS** single-sign on in future

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 Data Repository

CORE COMMON DATA ELEMENTS

NOVEL CDE FOCUSED REPOSITORY TO FOSTER INTEROPERABILITY

COMPLY WITH DATA SHARING POLICY - HOST PROJECT DATA

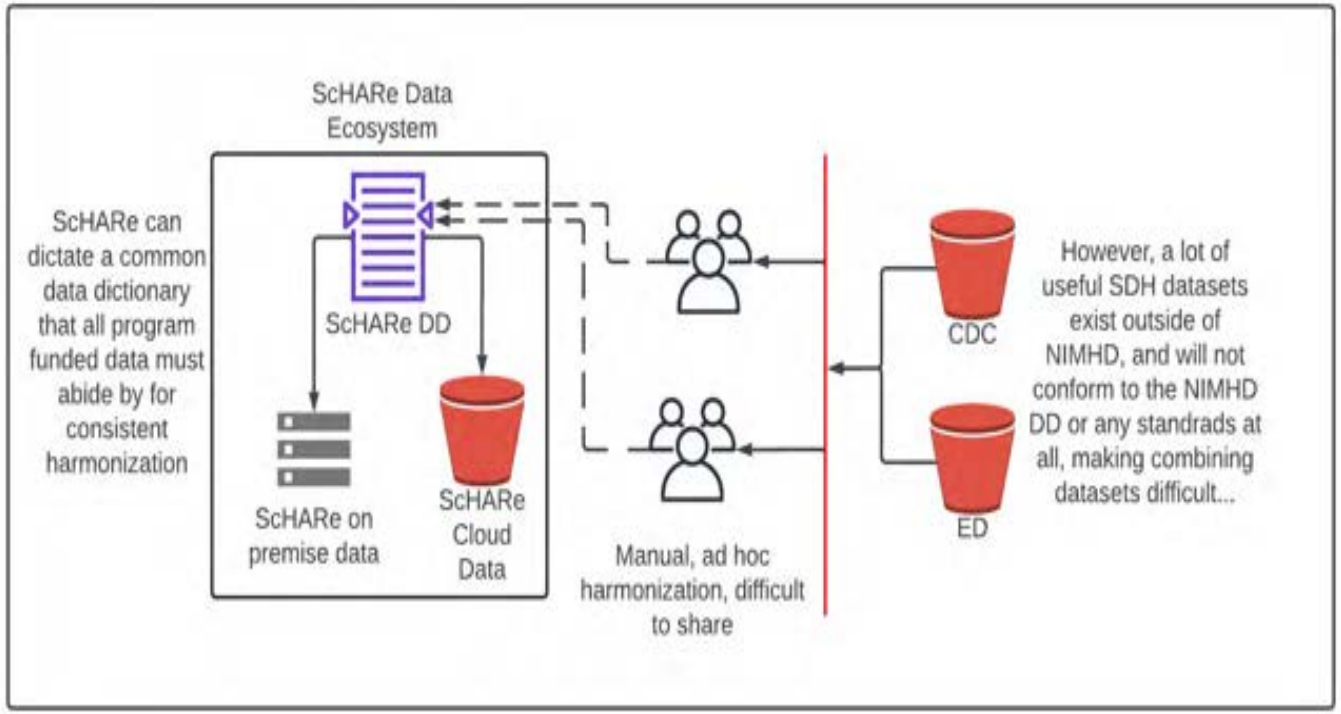
DATA ECOSYSTEM

- Map across datasets
- Map across platforms

UPCOMING

Common Data Elements and Harmonization

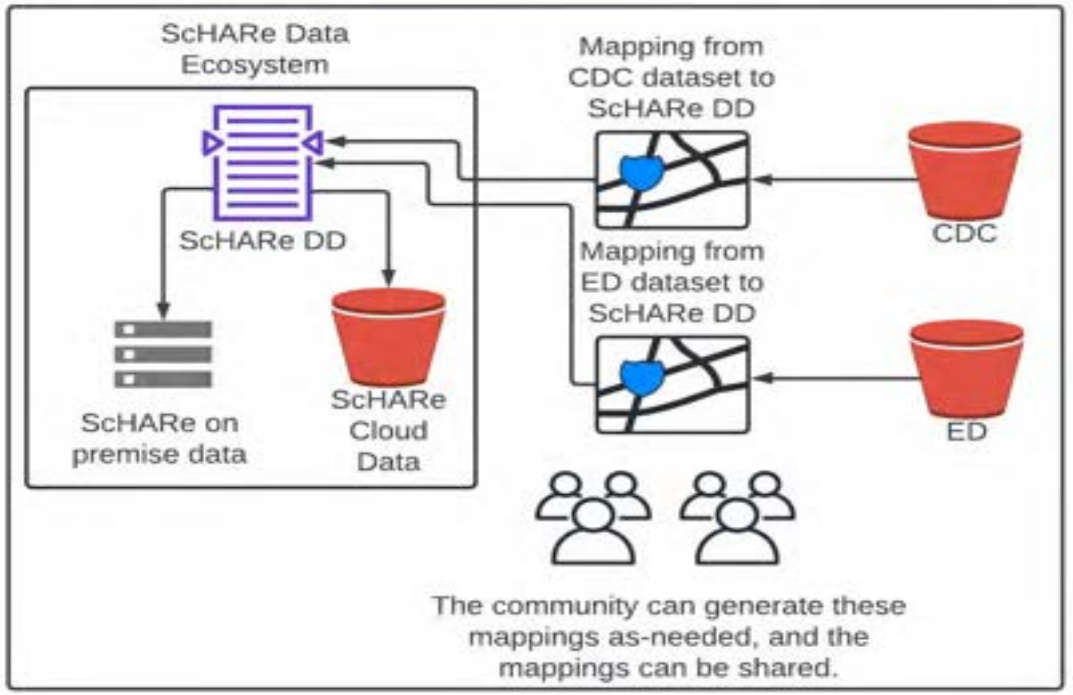
- Through significant efforts, **data can be harmonized** to a particular data dictionary
- To incorporate external data, users must harmonize in an ad-hoc manner that is **specific to a particular use case**, and requires significant manual effort
- It is **difficult to share these mappings** with others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



UPCOMING

Harmonization through mapping

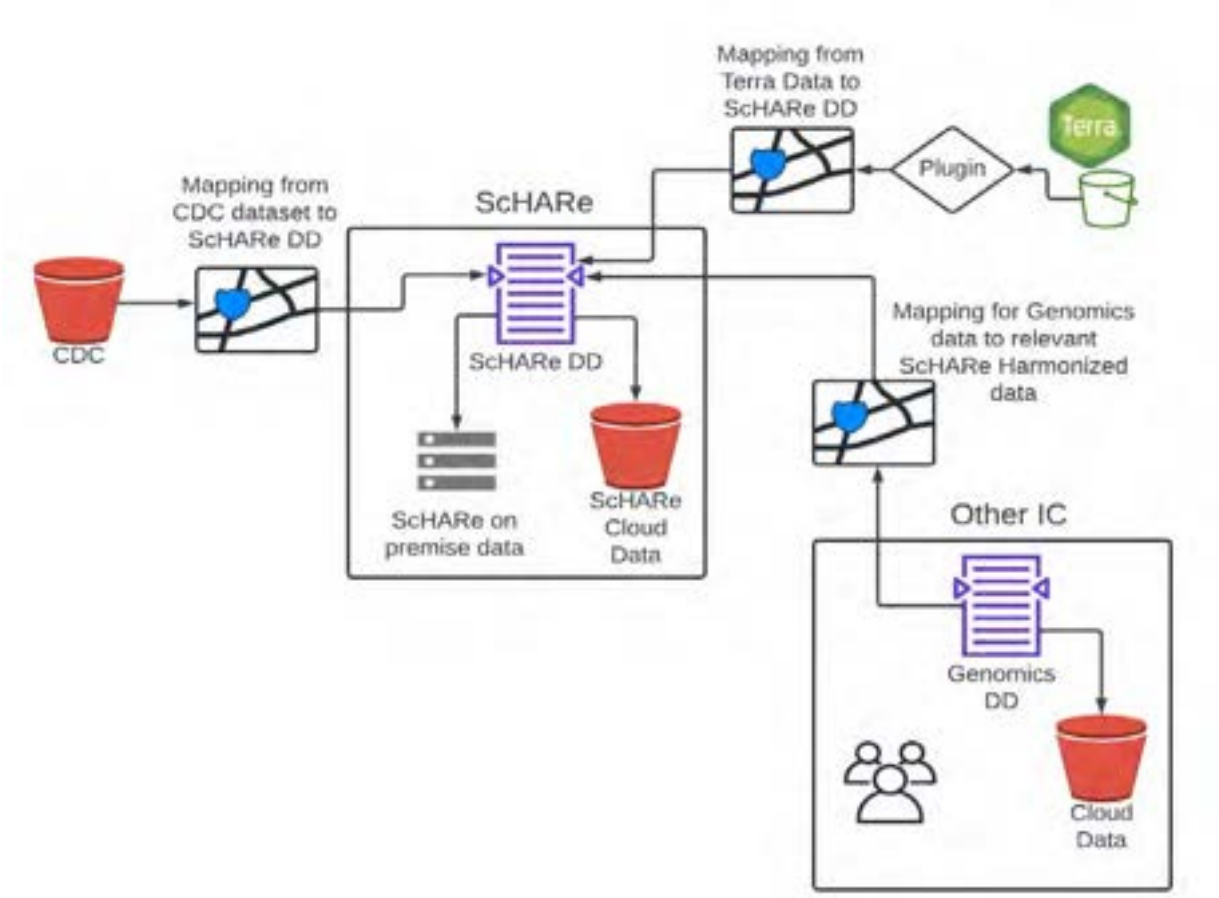
- **Harmonization should be shared**: data standards cannot be enforced on everyone
- We are building a **mapping system** to relate datasets to data dictionaries and transform them on the fly (we call the result of a transform a **dataview**)
- Mappings can easily be **shared amongst the community**: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



UPCOMING

Mapping extends between platforms

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



UPCOMING

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

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

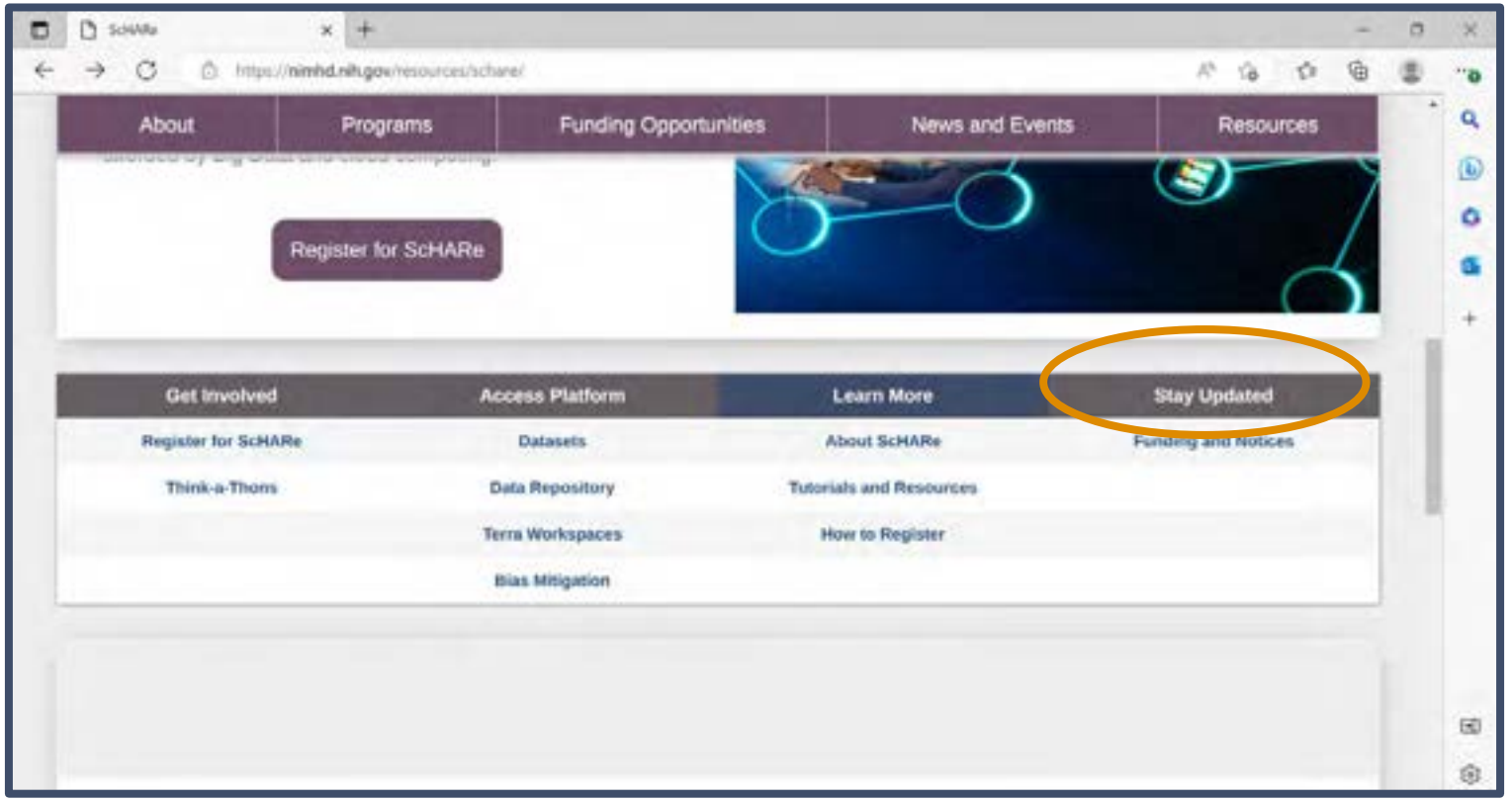
Interest poll

I am interested in (check all that apply):

- ☐ Learning about Health Disparities and Health Outcomes research to apply my data science skills
- ☐ Conducting my own research using AI/cloud computing and publishing papers
- ☐ Connecting with new collaborators to conduct research using AI/cloud computing and publish papers
- ☐ Learning to use AI tools and cloud computing to gain new skills for research using Big Data
- ☐ Learning cloud computing resources to implement my own cloud
- ☐ Developing bias mitigation and ethical AI strategies
- ☐ Other

Two Ways to Sign up for ScHARe News

Scannable from your screen!

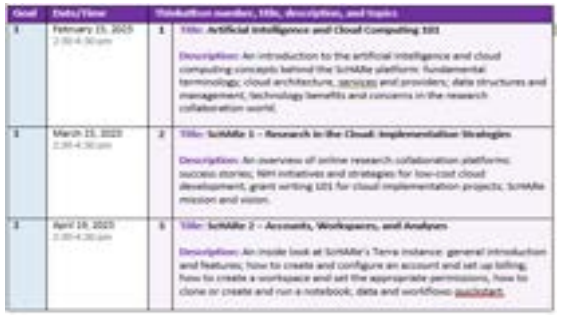


nimhd.nih.gov/schare

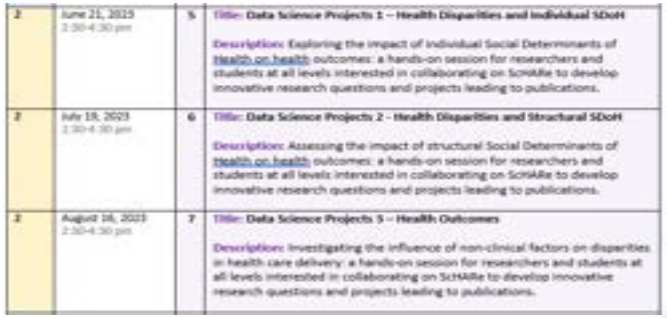
ScHARe Think-a-Thons

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

Instructional



Research teams



Register:

bit.ly/think-a-thons

Part II
Registration and Account

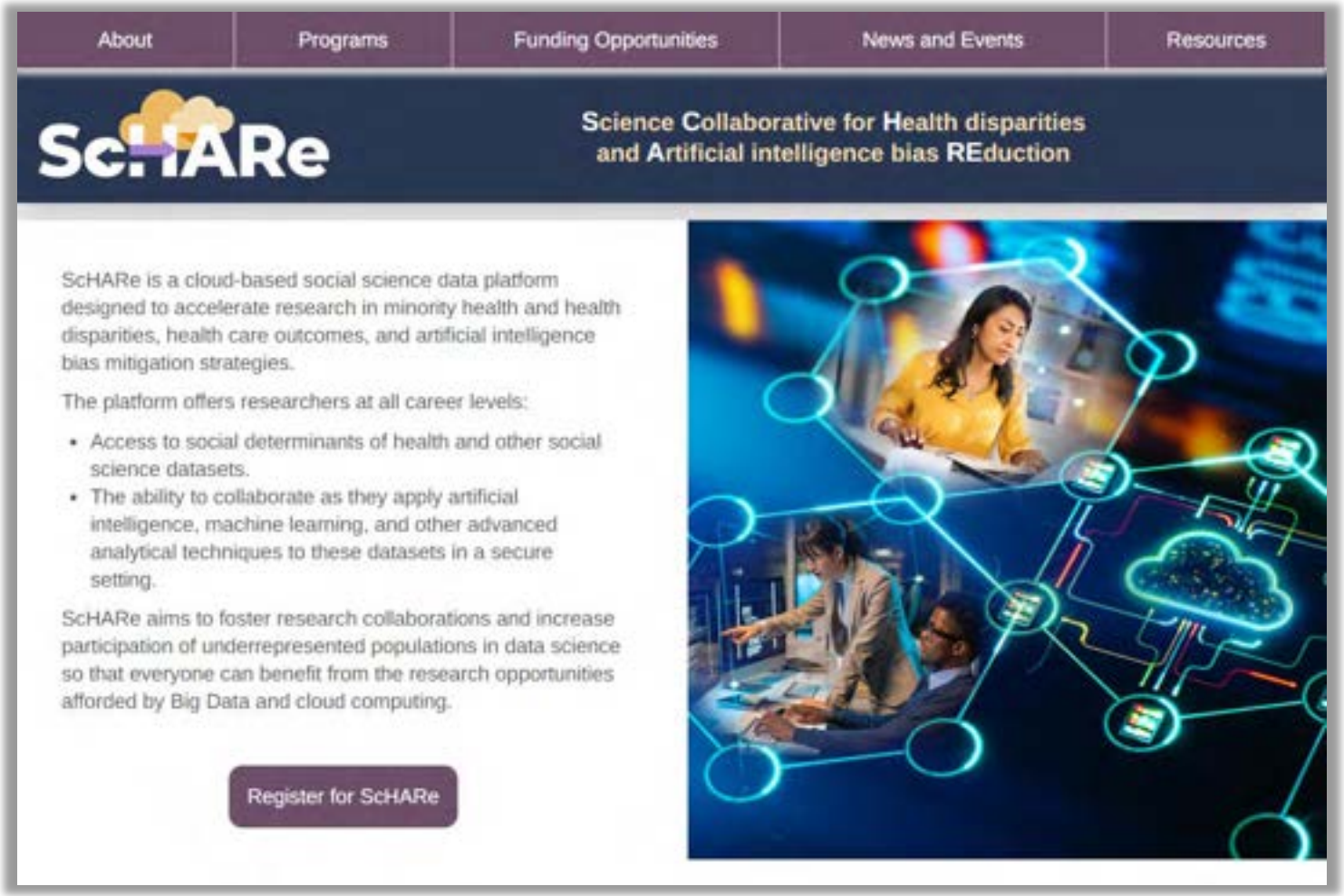
Registering for ScHARe

Complete the following steps to register for ScHARe:

1. Visit the ScHARe portal on the NIMHD website:

   nimhd.nih.gov/schare

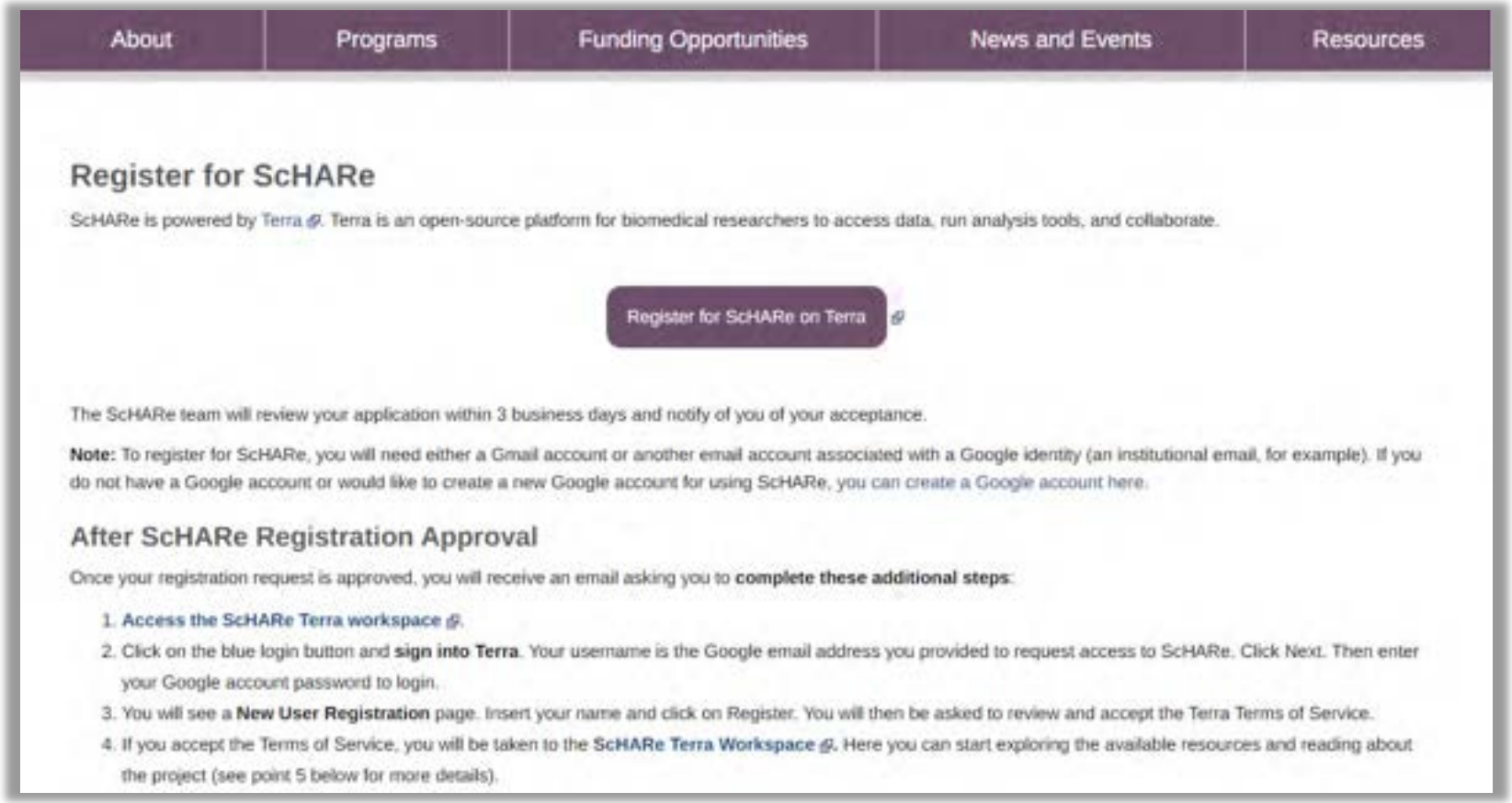
2. Click on the “Register for ScHARe” button



Registering for ScHARe

On the registration page:

Click on the “Register for ScHARe on Terra” button



Registering for ScHARe

Complete the registration form

The ScHARe team will:

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

➢ 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

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

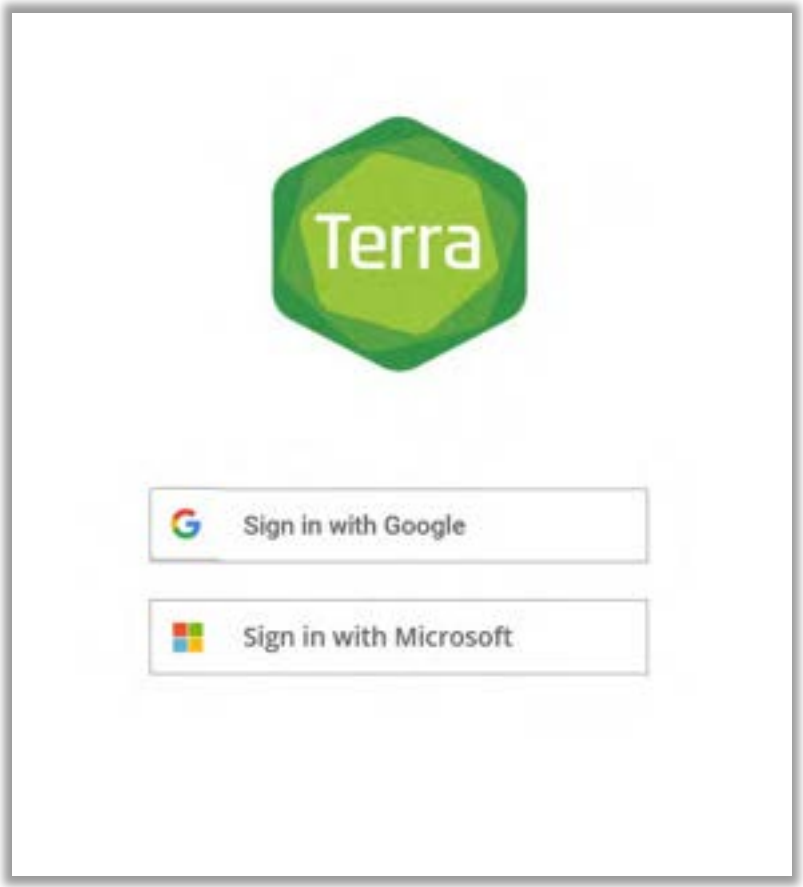
2. Click on the blue “Log in” button

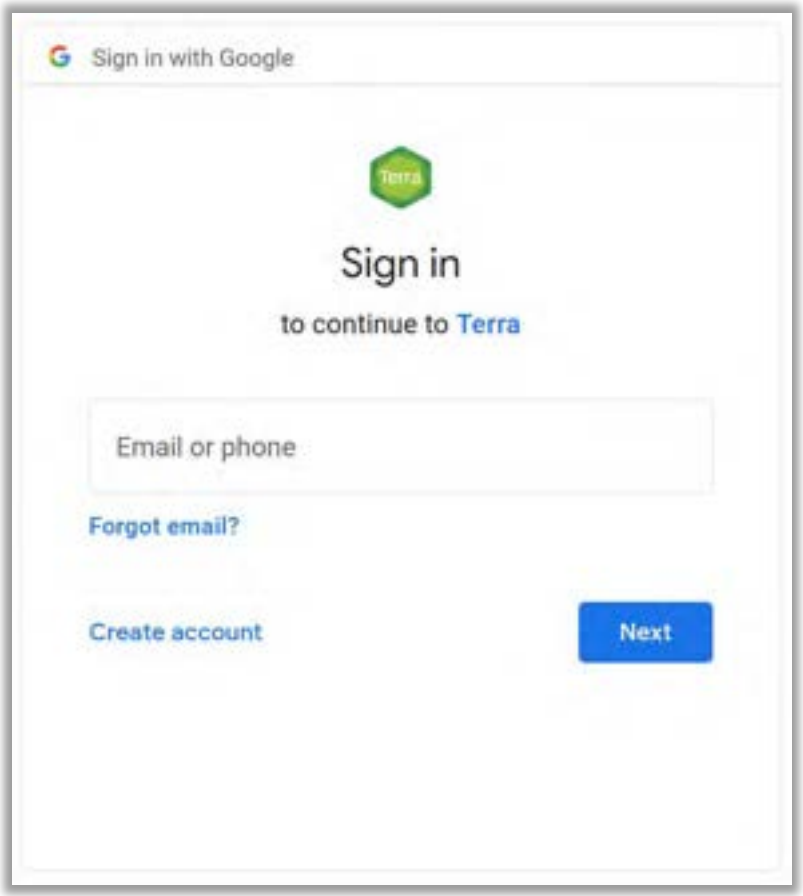
Creating a Terra account

4. Select “Sign in with Google”

5. Sign into Terra. Your username is the Google email address you provided to request access to ScHARe

6. Click “Next” and enter your Google account password to login





Creating a Terra account

7. You will see a New User Registration page. Insert your name and contact email, then click on “Register”

Creating a Terra account

8. Review and accept the Terra Terms of Service

Access to social and behavioral datasets

- The cloud offers access to vast repositories of data, and enables mapping and linking across sources
- Extremely large datasets are statistically analyzed to gain detailed insights, often using AI and substantial computer-processing power
- Datasets can be linked together to see how patterns in one domain affect other areas
- Data can be structured into fixed fields or unstructured as free-flowing information
- Increase use of Dark Data

ScHARe

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
- Repository for Funded Datasets on ScHARe – Compliance with NIH Data Sharing Policy

Part III
Workspaces and Permissions

What is a workspace

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

You can use workspaces to:

- **Link to data in the cloud** for analysis, instead of downloading and storing it yourself
- **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 a programming expert
- **Share reproducible analysis results**
- **Collaborate** while maintaining control of your resources

Dashboard

The landing page (i.e., Dashboard) is your project overview - the questions you're trying to answer, the data and analysis tools you'll use, etc.

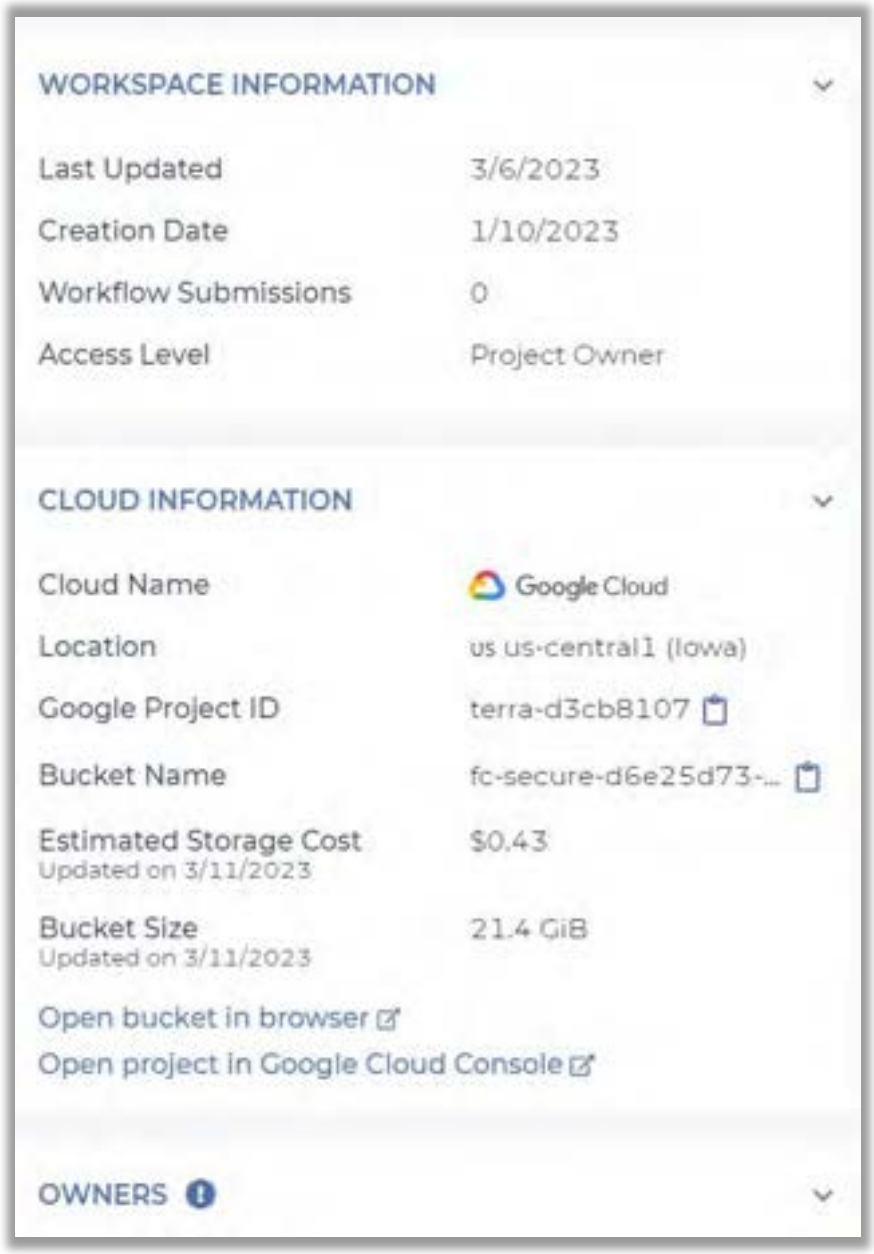
The dashboard uses the **Markdown language**, which lets you organize text with headers and include links and additional references

Dashboard

Workspace details are populated automatically in the right column of the Dashboard

Expandable sections include:

- **Workspace information:** creation date, date last updated, your access level
- **Cloud information:** location of workspace storage, estimated storage cost and size, etc.
- **Workspace owners**
- **Workspace tags:** only visible to owners, tags are useful for searching and indexing

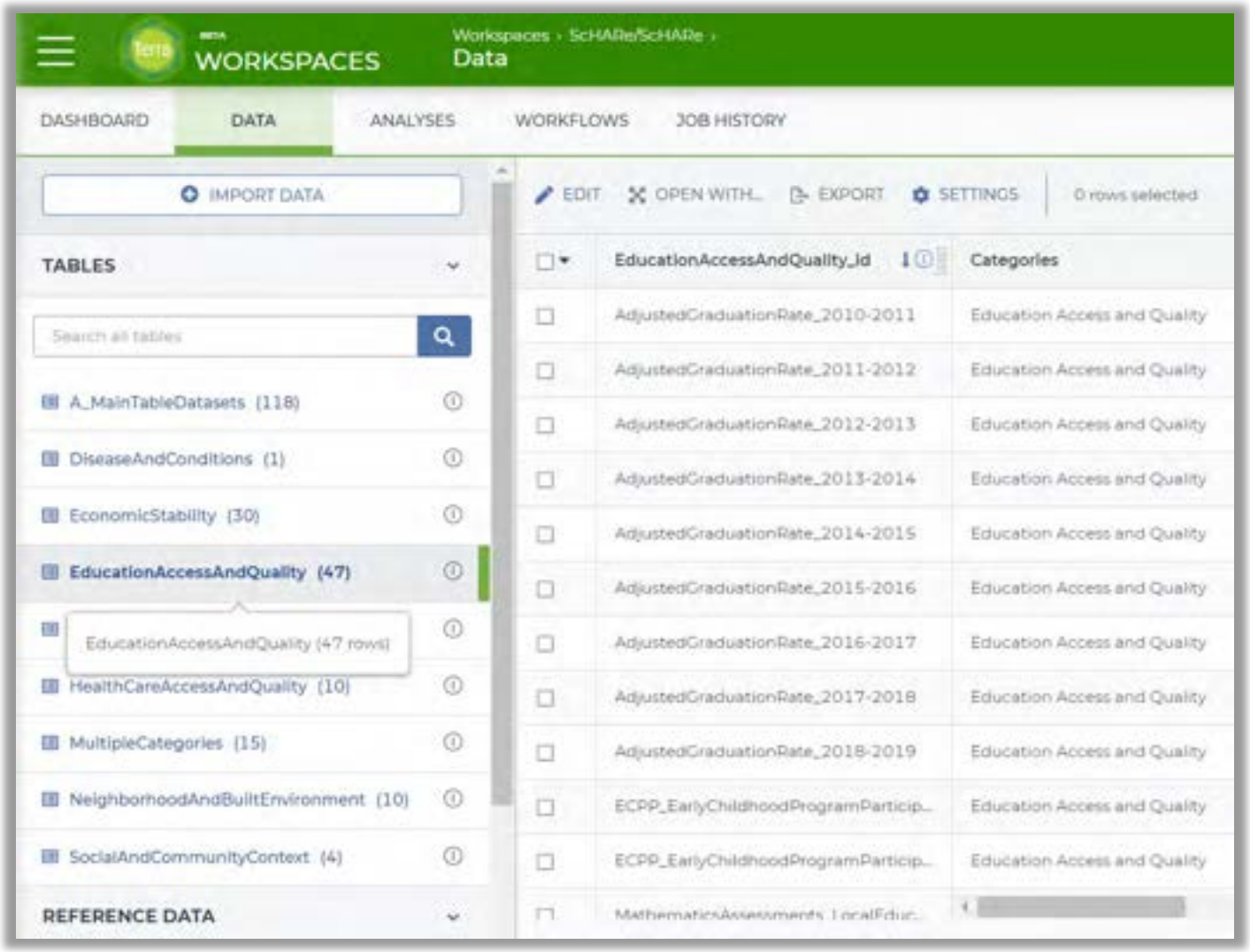


Data

In the **Data** tab, like spreadsheets built right into the workspace, **data tables help keep track of all project data**, no matter where files are stored in the cloud

- 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

More information will be provided in the next Think-a-Thons

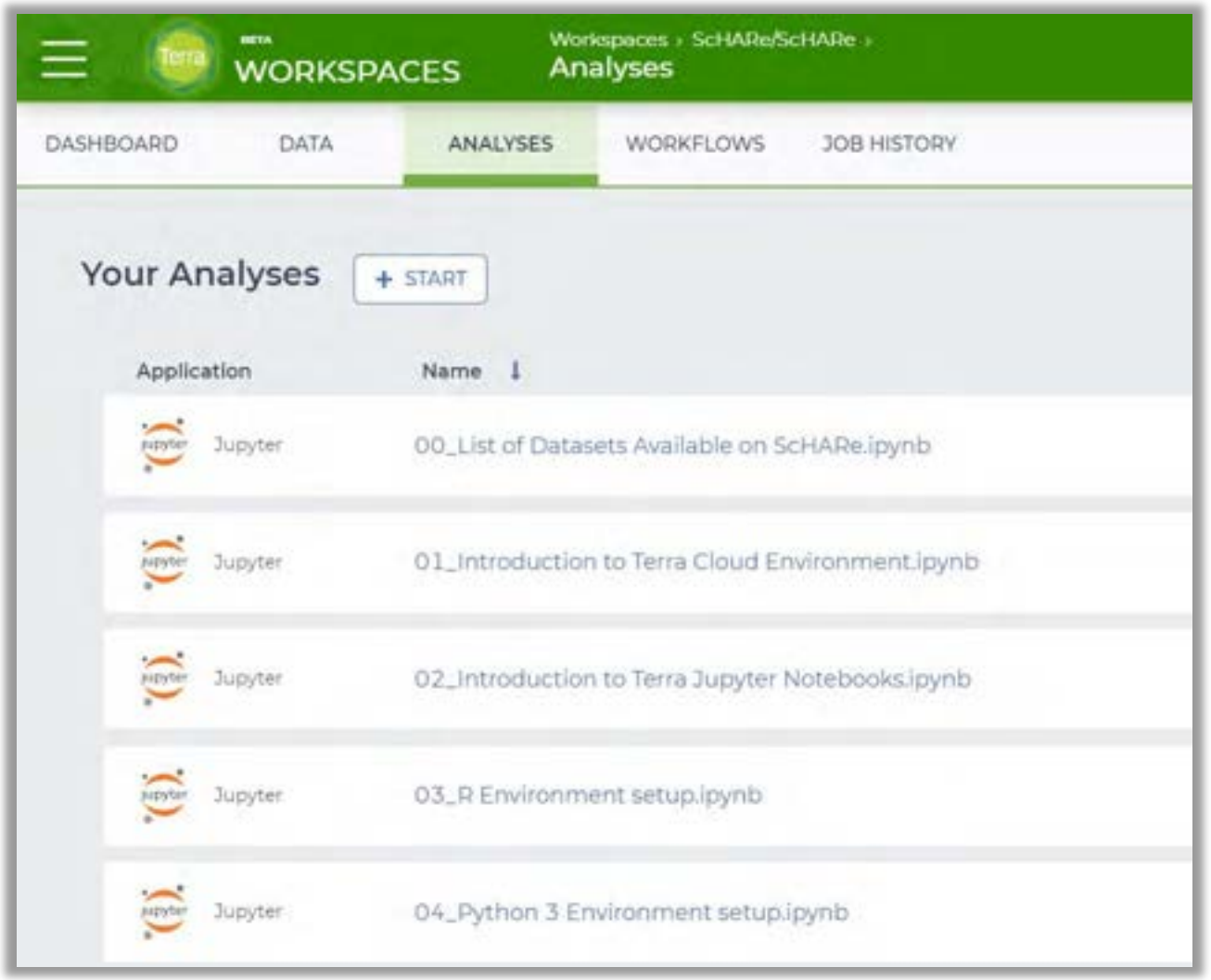


Analyses

In the **Analyses** tab, you can interact with the data and perform/share analyses using Python or R

- In the ScHARe workspace, click on the **Analyses** tab
- This tab contains instructional notebooks and analysis tutorials on how to use the ScHARe resources in the Terra environment

More information on our instructional notebooks will be provided later

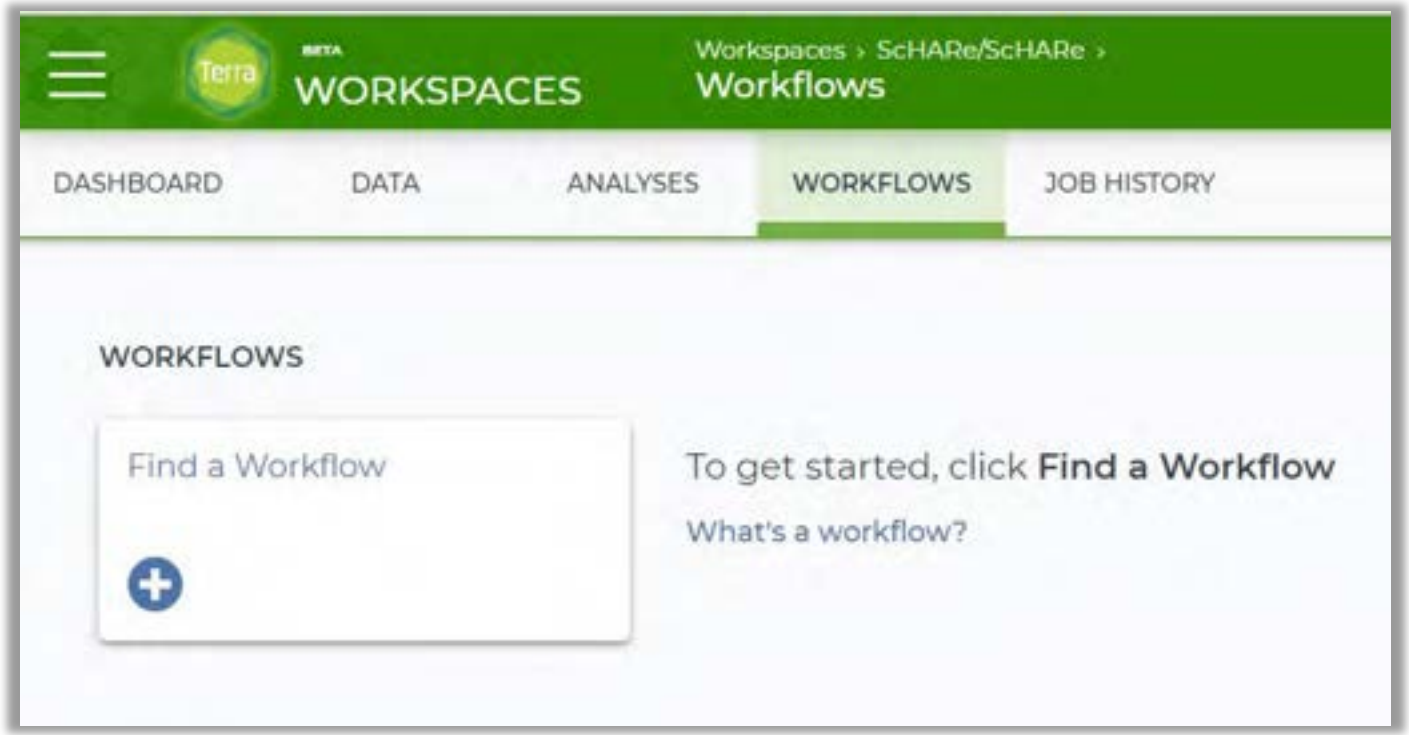


Workflows

In the **Workflows** tab, you will find workflows for bulk analyses

- These are the sorts of repetitive analyses that can be automated
- Workflows in Terra are written in the human-readable Workflow Description Language (WDL)

More information will be provided in the next Think-a-Thons



Creating your first workspace

Let's create your first Terra workspace!

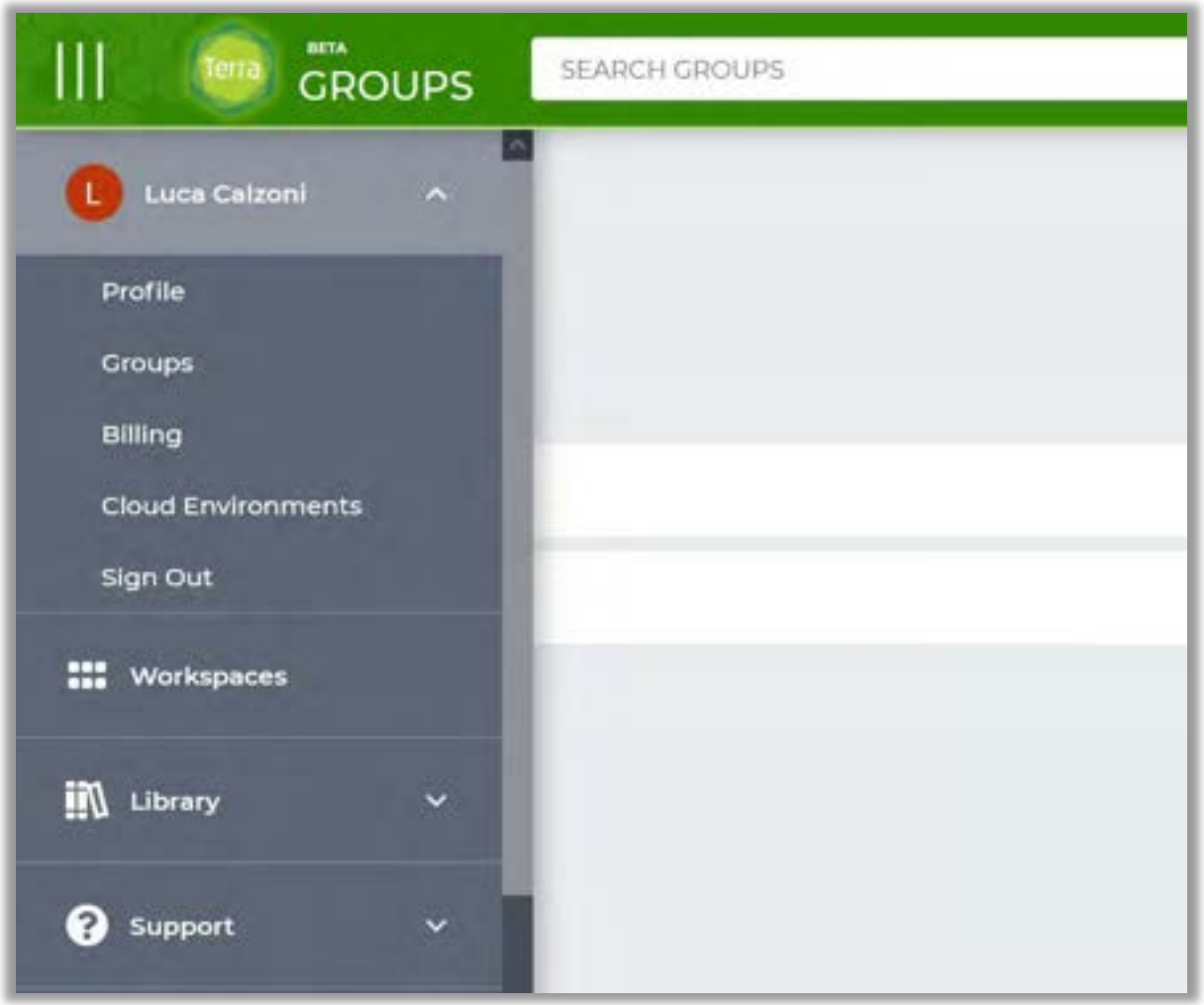
For the purpose of this tutorial, let's assume that you intend to create a workspace that will allow you to work with two groups of collaborators:

- **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 you plan to use to share your results with the scientific community

- **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

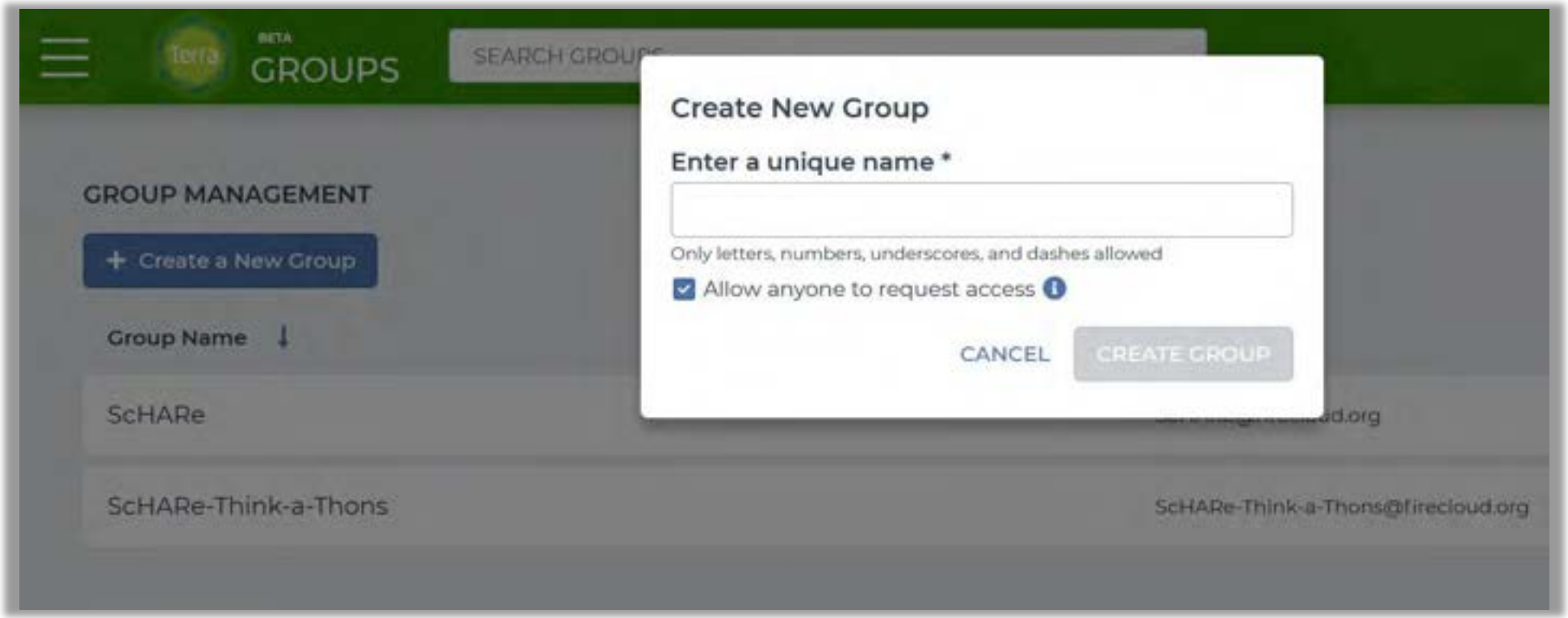
Collaborators and permissions

Click on the menu in the top left corner of the page, then on **“Groups”**



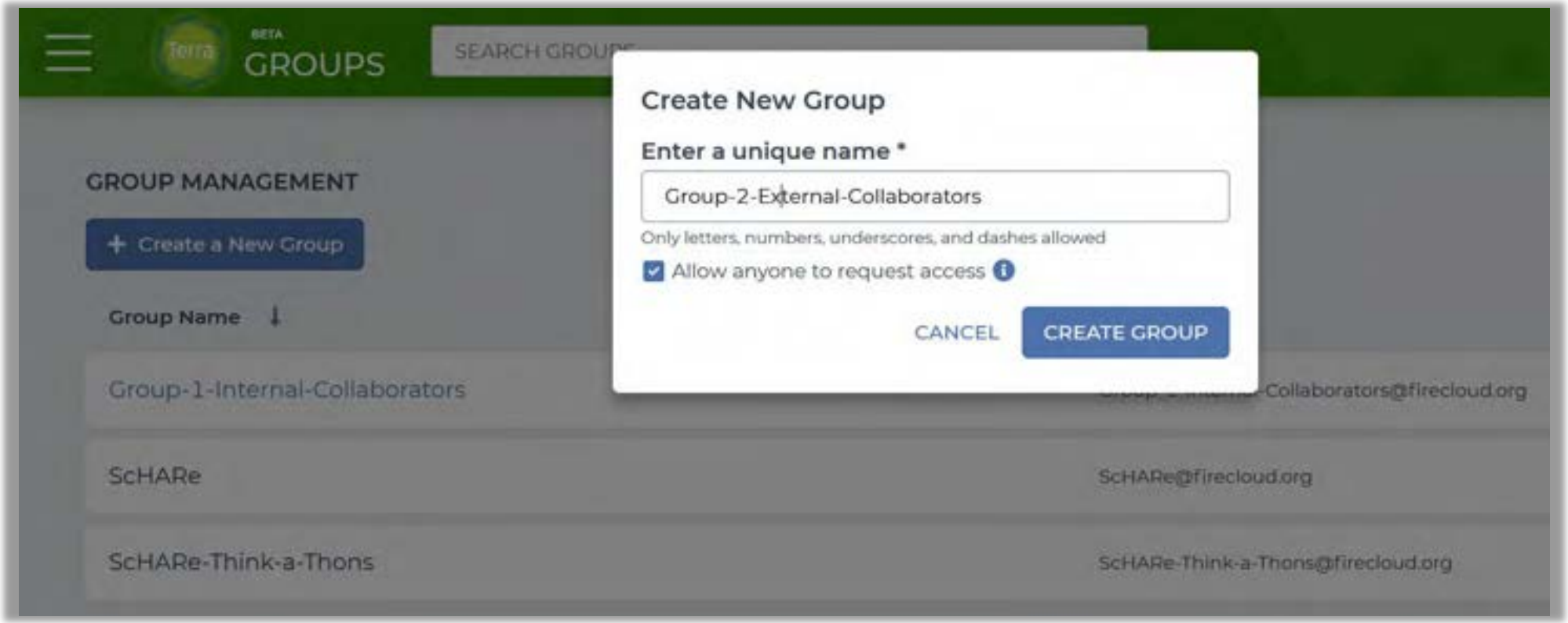
Collaborators and permissions

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



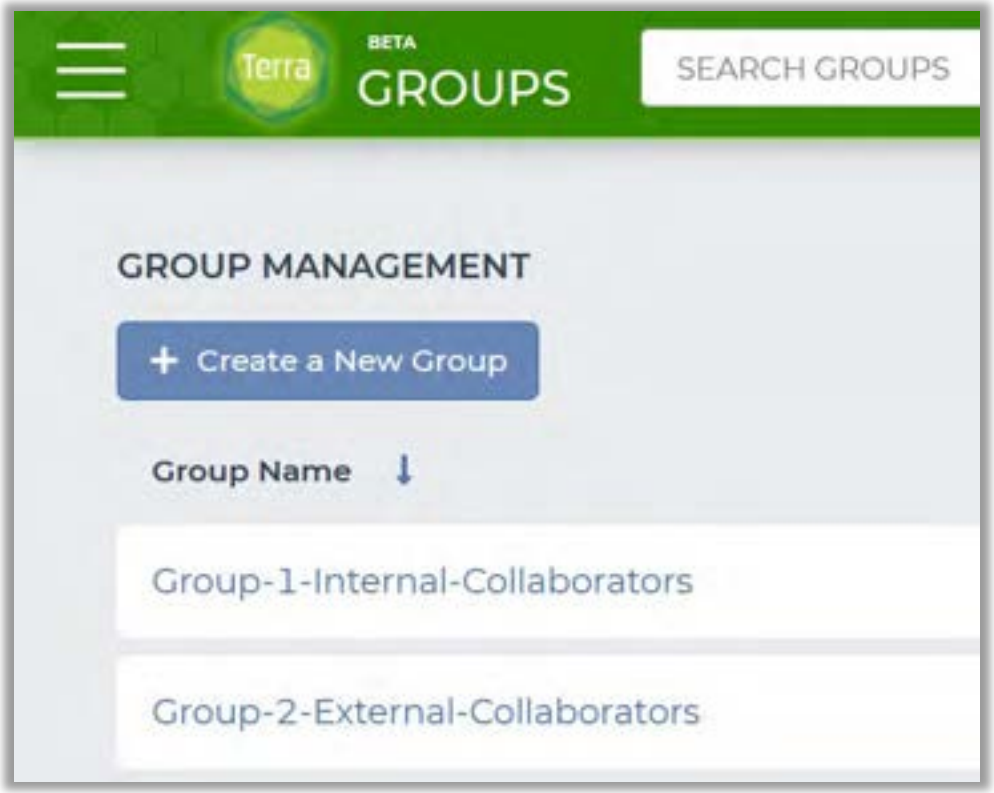
Collaborators and permissions

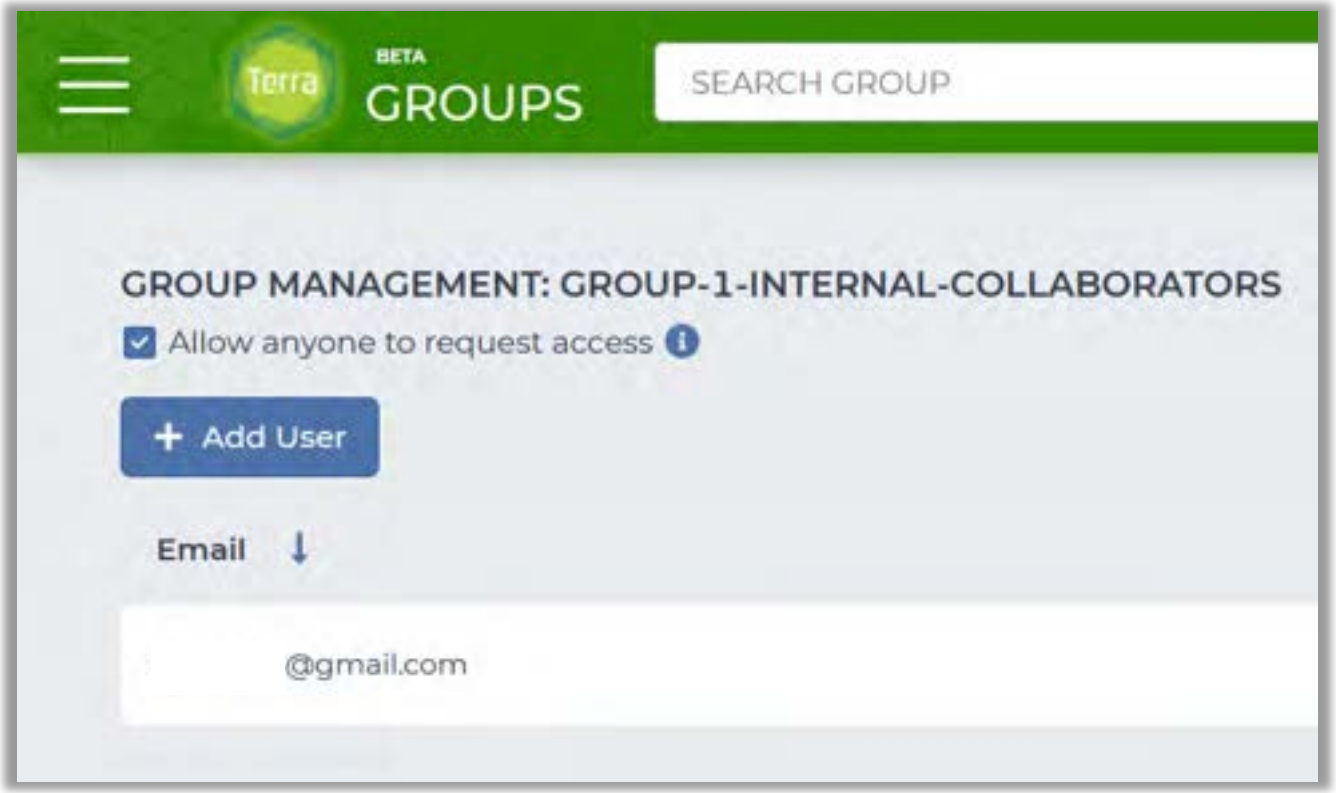
On the Groups page, select “Create a New Group” and proceed to **create two different groups**, one for each of the two groups of collaborators previously identified



Collaborators and permissions

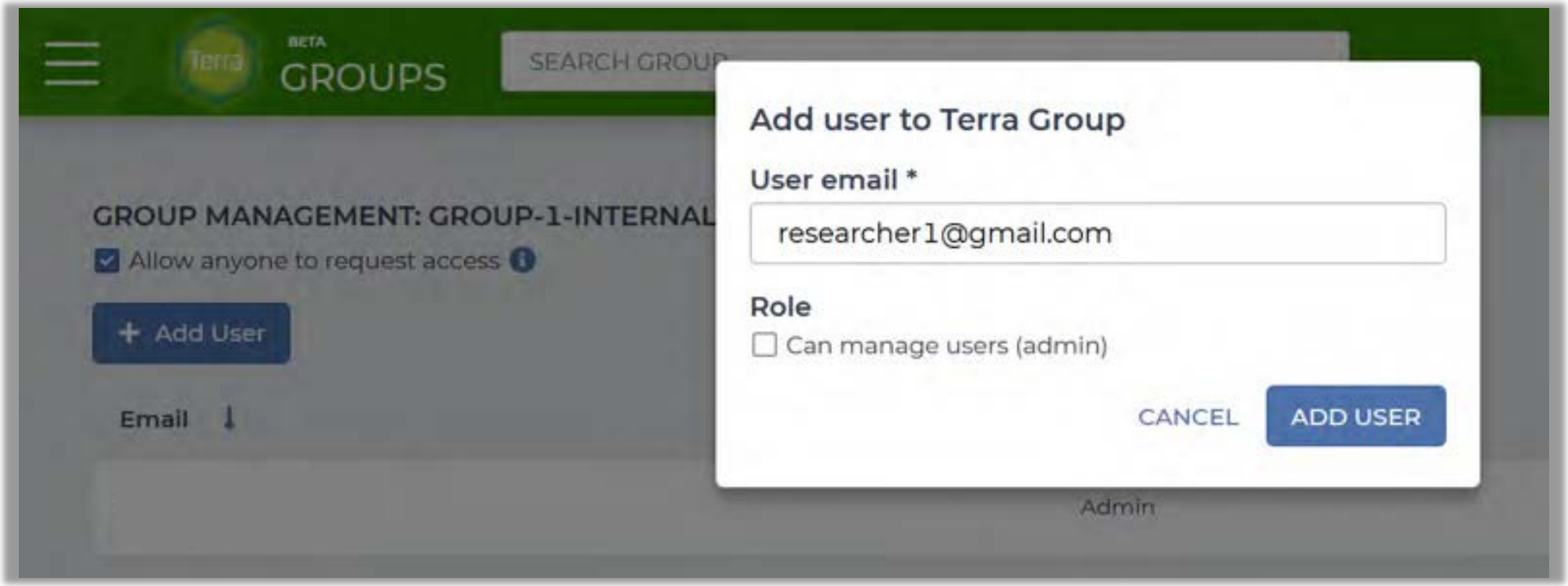
For each group, click on the name of the group and, in the following screen, on “**Add User**”





Collaborators and permissions

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



Authorization domain

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

It's time to create an **authorization domain** for your workspace!

What is an authorization domain (AD)?

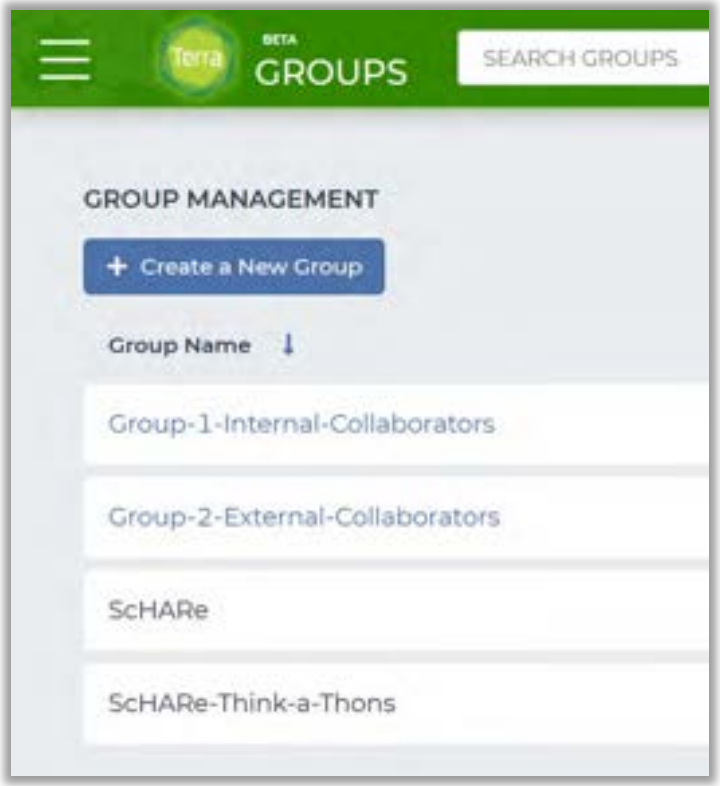
Data in the cloud is more secure than data stored locally. Terra has several layers of protection.

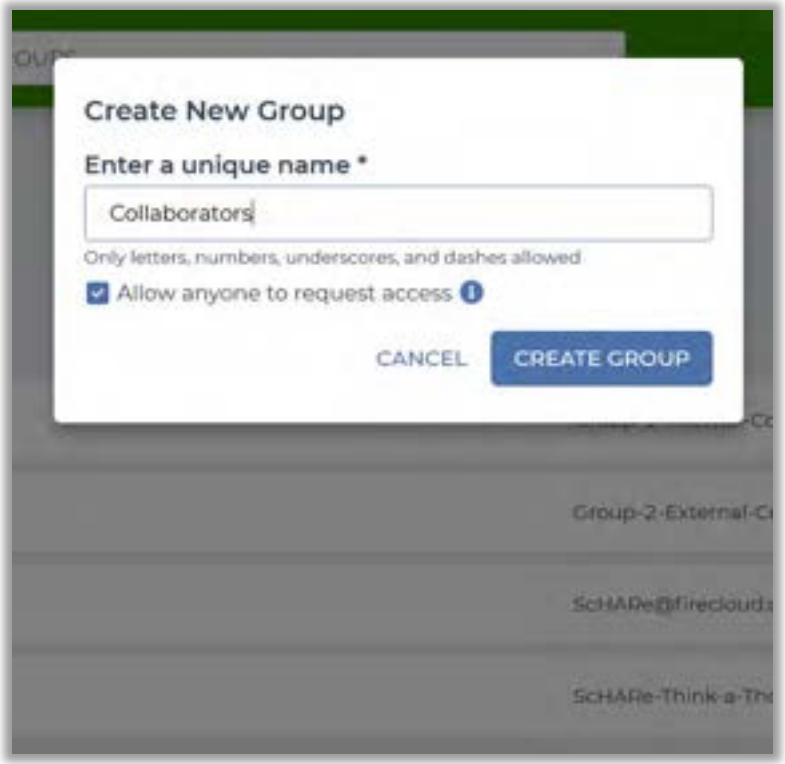
ADs are data protection mechanisms **- managed groups with strictly enforced workspace permissions** that, like a badge, follow the original workspace and all copies of it.

The AD guarantees that only members of the AD can clone (copy) the original workspace. If anyone tries to share a cloned (copied) workspace with a colleague who doesn't have the right badge (doesn't belong to the AD), the recipient won't be able to view the workspace. **Your data and analyses are safe!**

Authorization domain

Create an **authorization domain** for your workspace, by adding a new group called “Collaborators” in the “Groups” section.

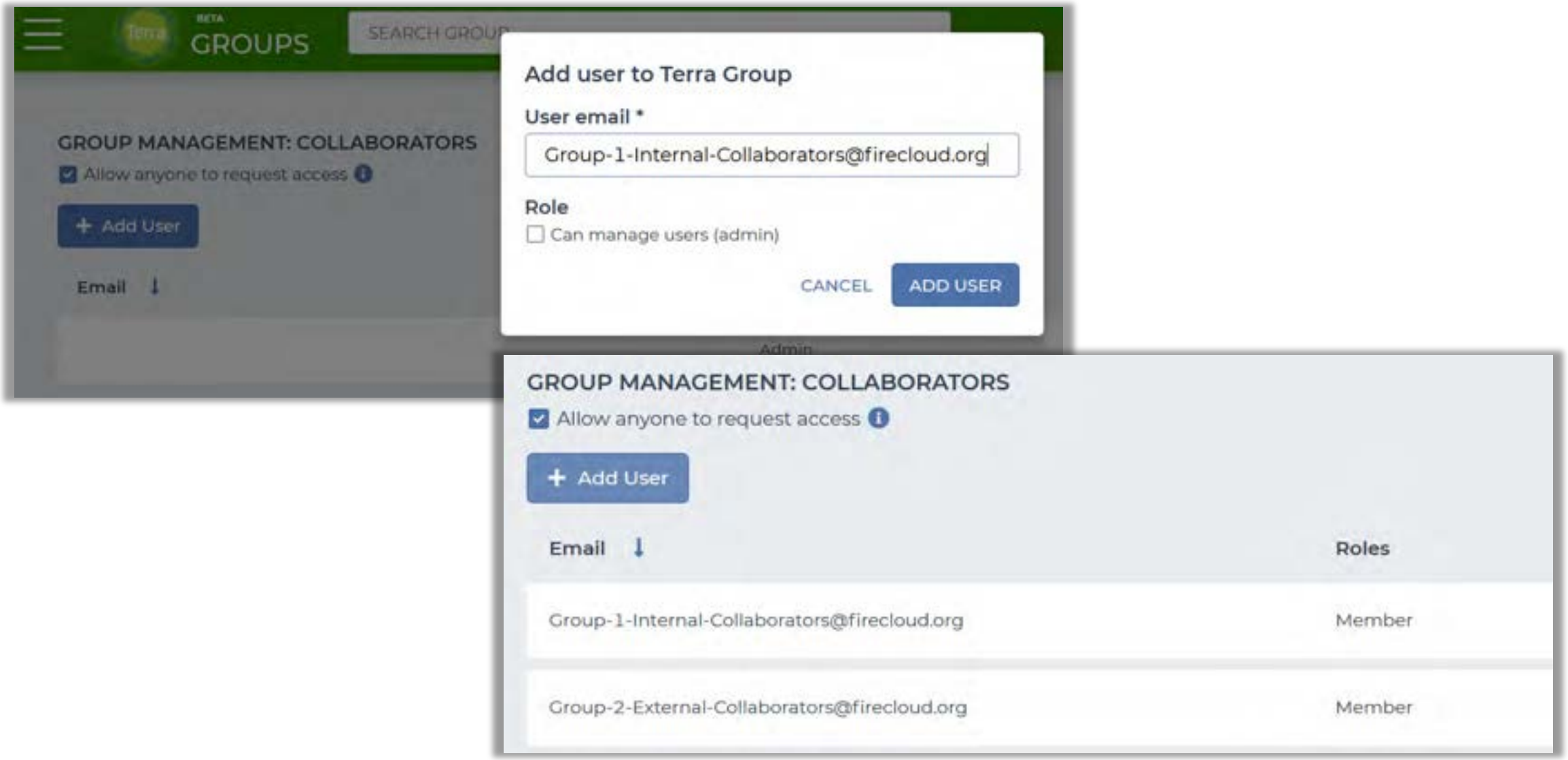




Authorization domain

Both your groups of collaborators will need to **belong to your "Collaborators" authorization domain** to be able to access your workspace. Simply add the group email associated with each group (circled in blue below) to the "Collaborators" group list: every email address in your groups will now belong to the "Collaborators" authorization domain as well. You can add additional users to each group at any time

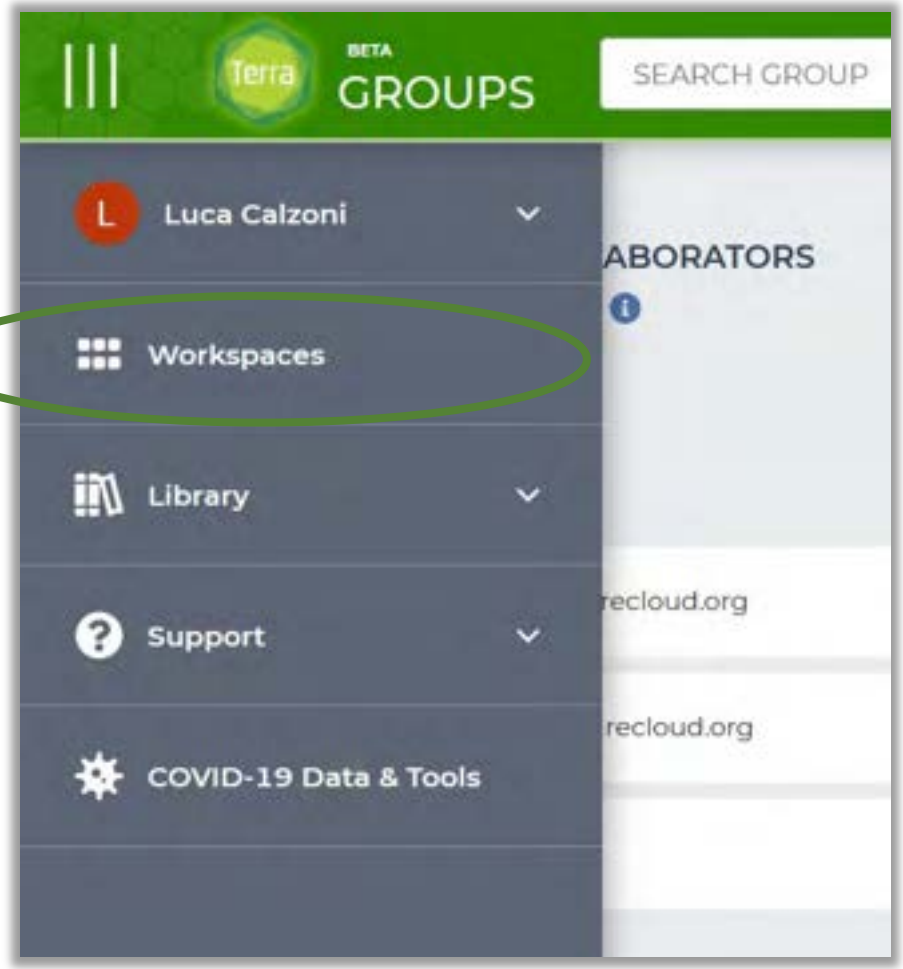
Authorization domain

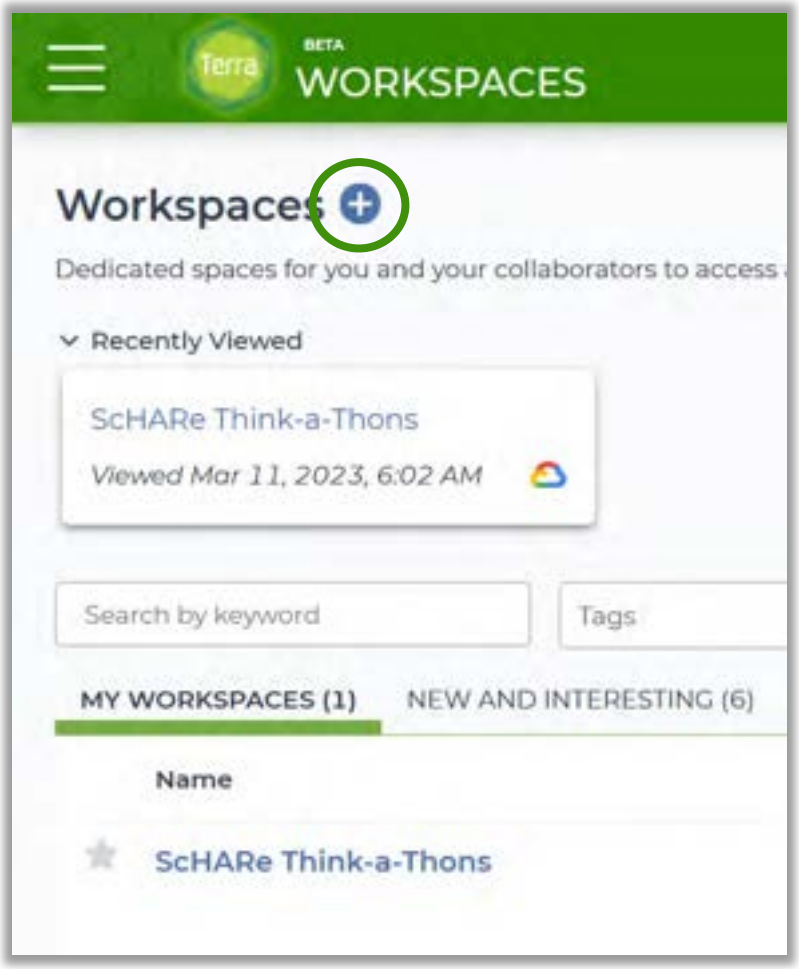


Creating the workspace

It’s now time to create your workspace!

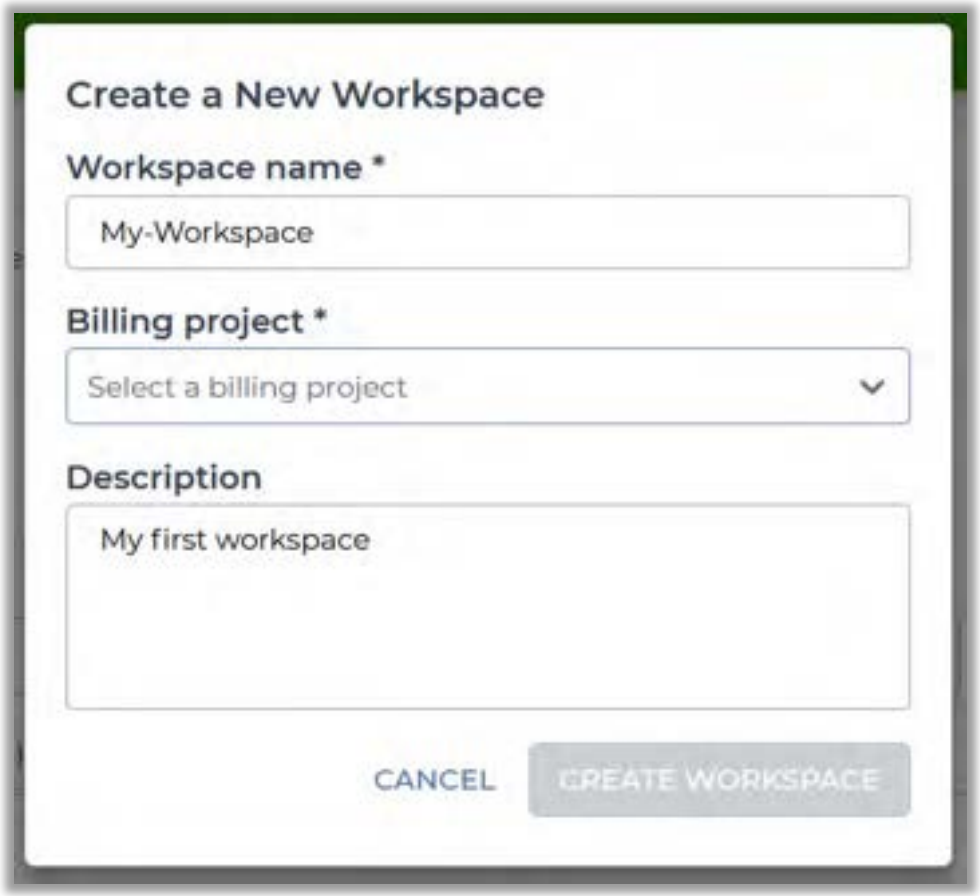
In the “Workspaces” section, click on the “+” sign next to “Workspaces”

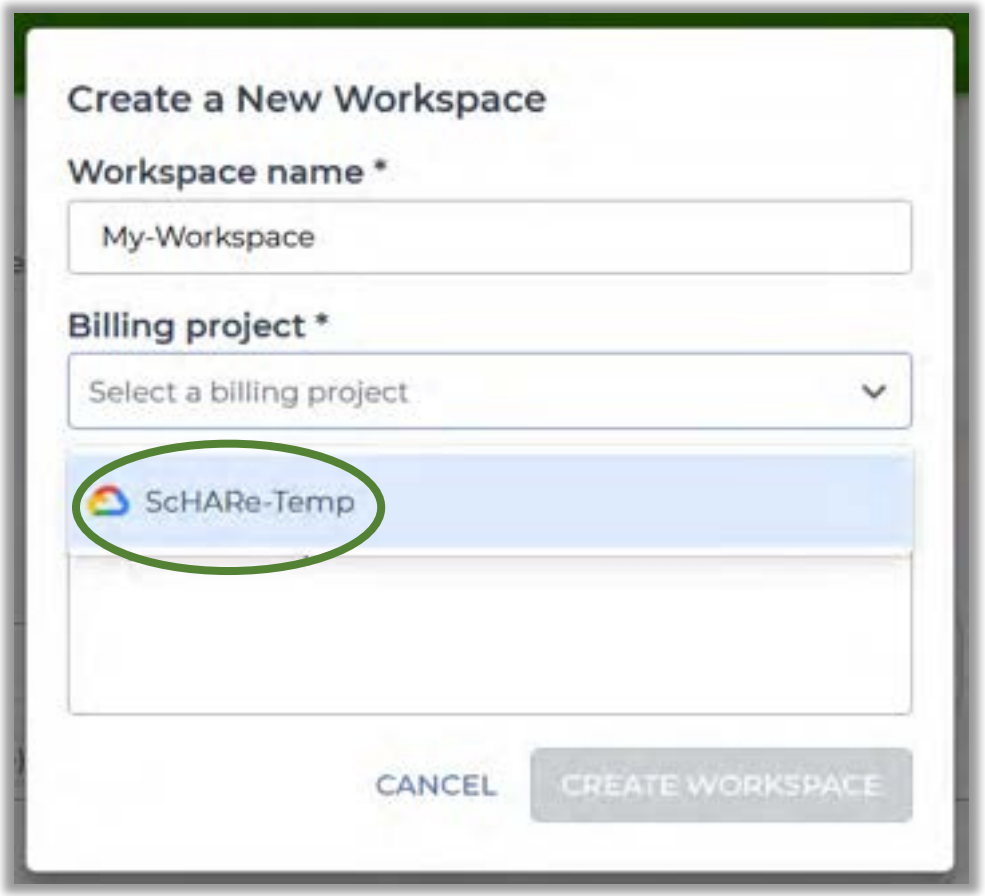




Creating the workspace

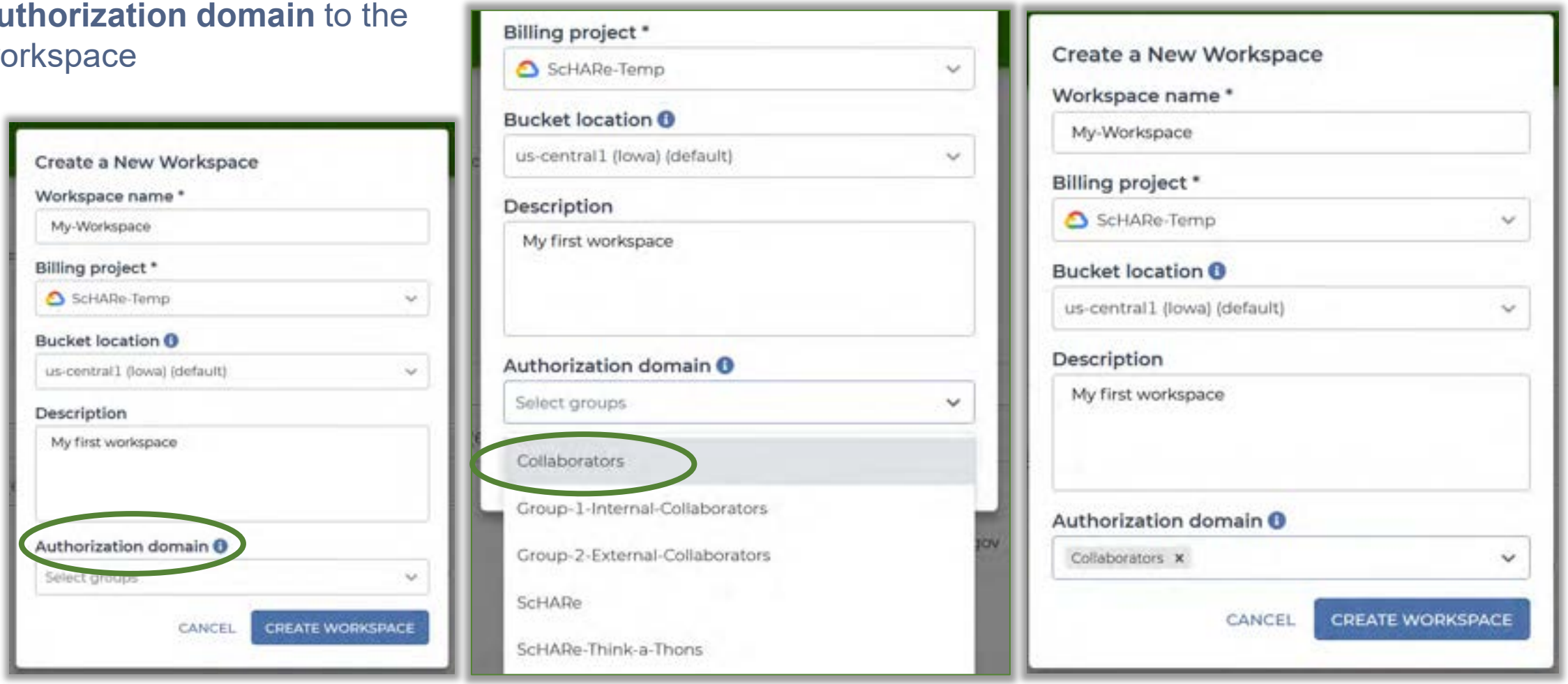
Assign a **name** to your workspace, select our free temporary **billing project "ScHARe-Temp"** (or create one following the instructions provided during this Think-a-Thon), and add a **description**





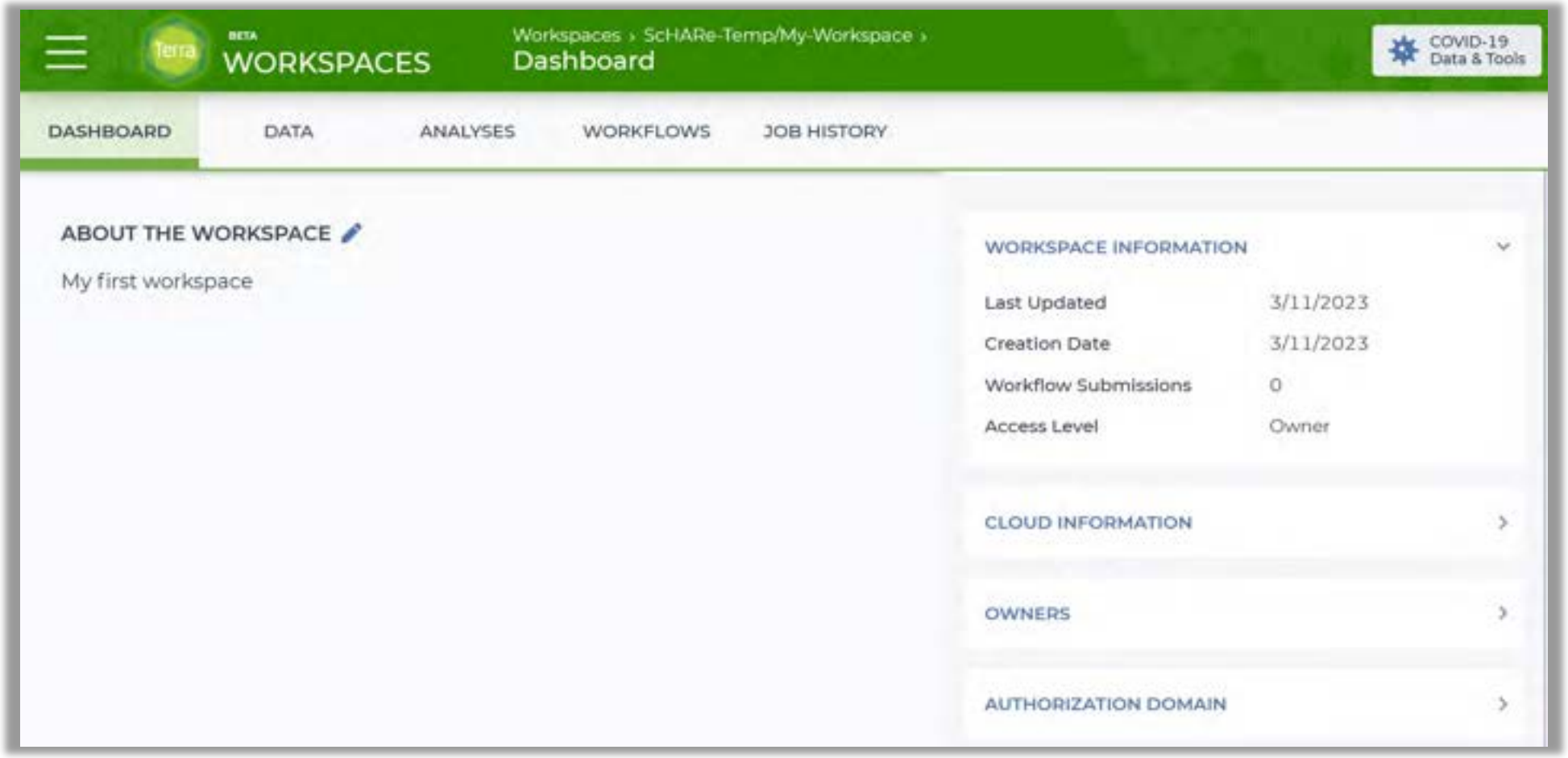
Creating the workspace

Assign the “Collaborators” **authorization domain** to the workspace



Creating the workspace

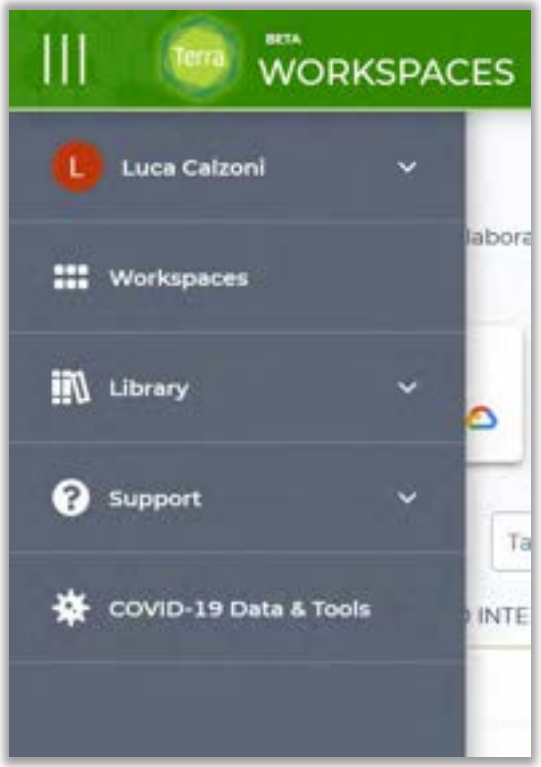
Congratulations, your workspace was successfully created!



Sharing the workspace

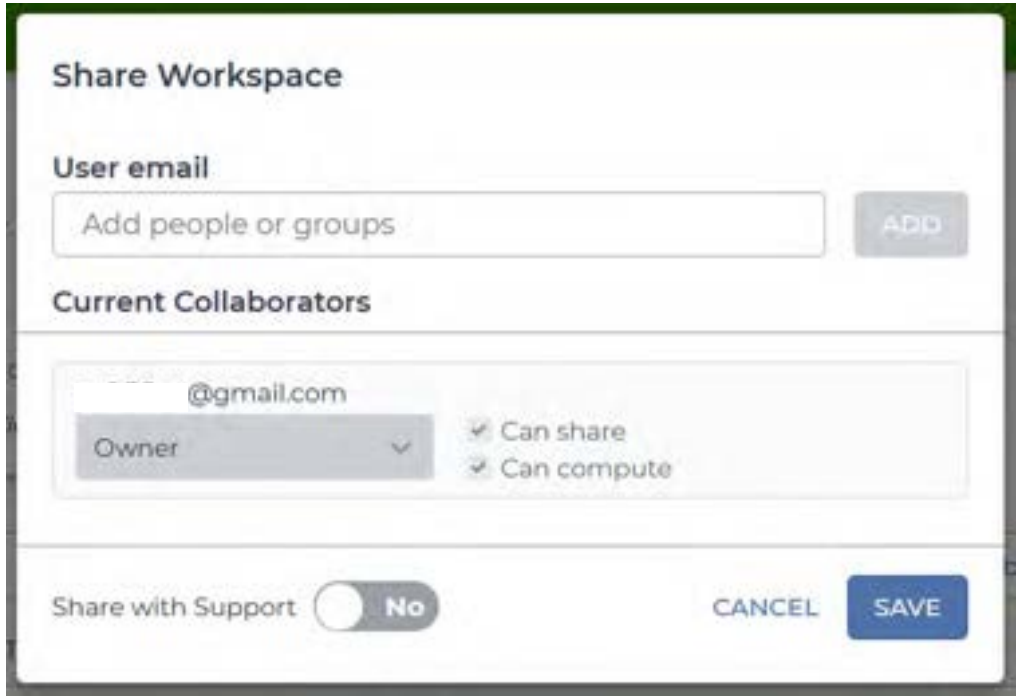
You are now ready to **share it with the two groups of collaborators** you created:

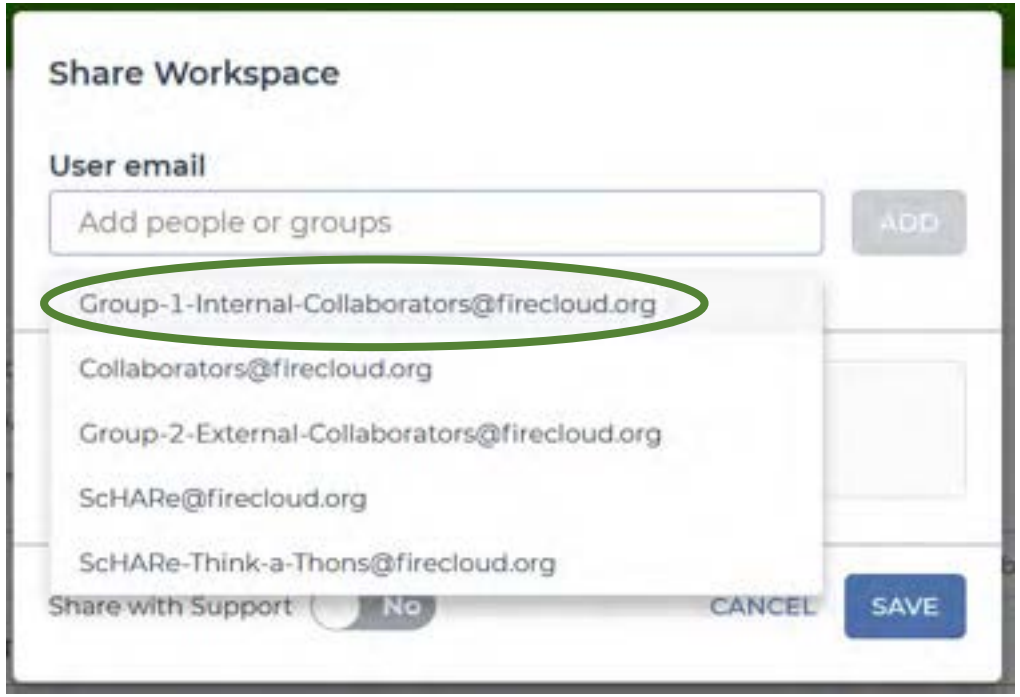
1. **Click on the menu** in the top left corner of the page, then on **“Workspaces”**
2. **Identify your workspace** in the list of workspaces provided on screen and click on the corresponding **vertical three-dot menu** (A), then on **“Share”** (B)



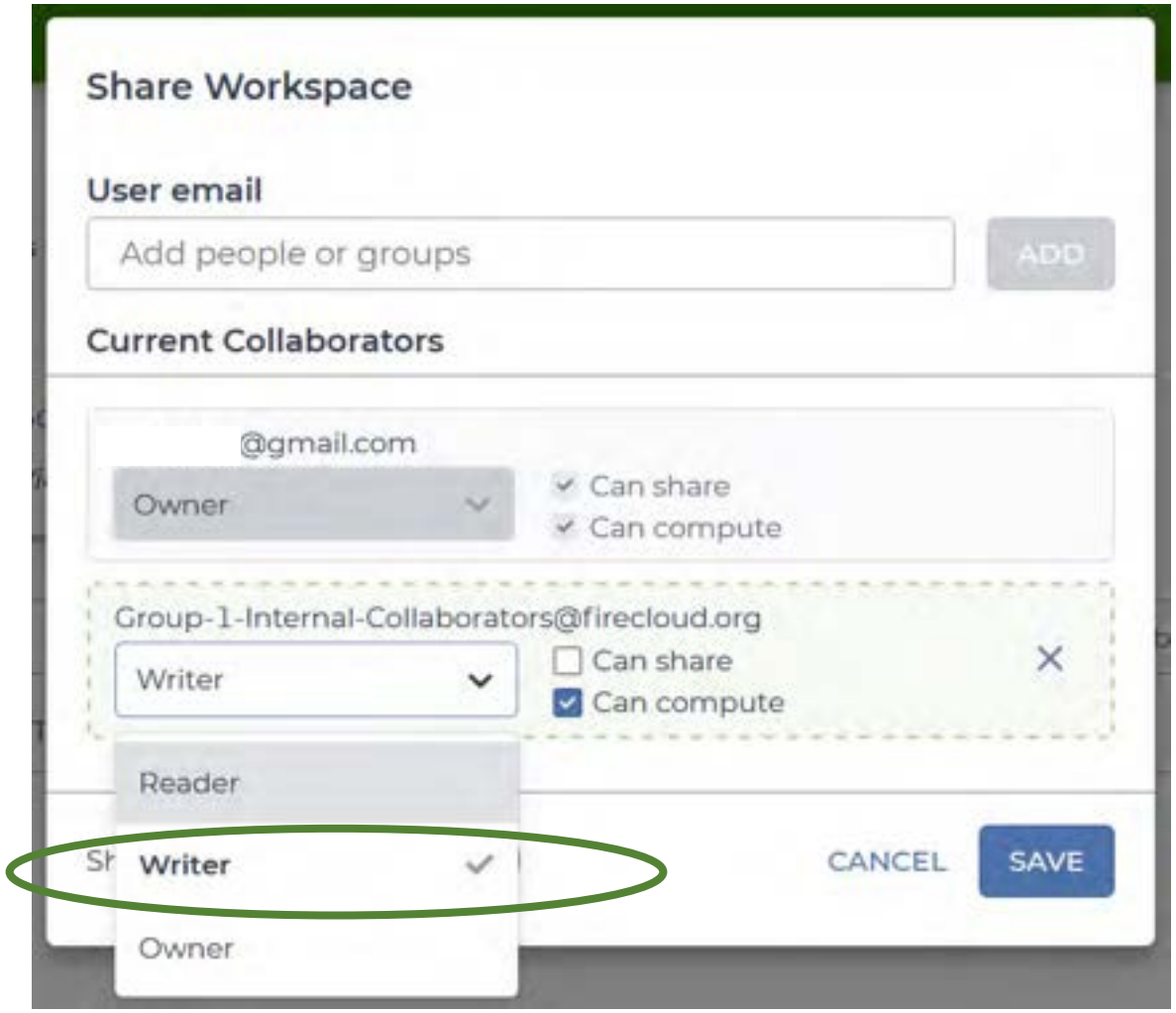
Sharing the workspace

In the drop-down menu, select the group email corresponding to your **first group of internal collaborators**





Sharing the workspace



Assigning permissions – Group 1 (Writer)

Since you want this group to be able to:

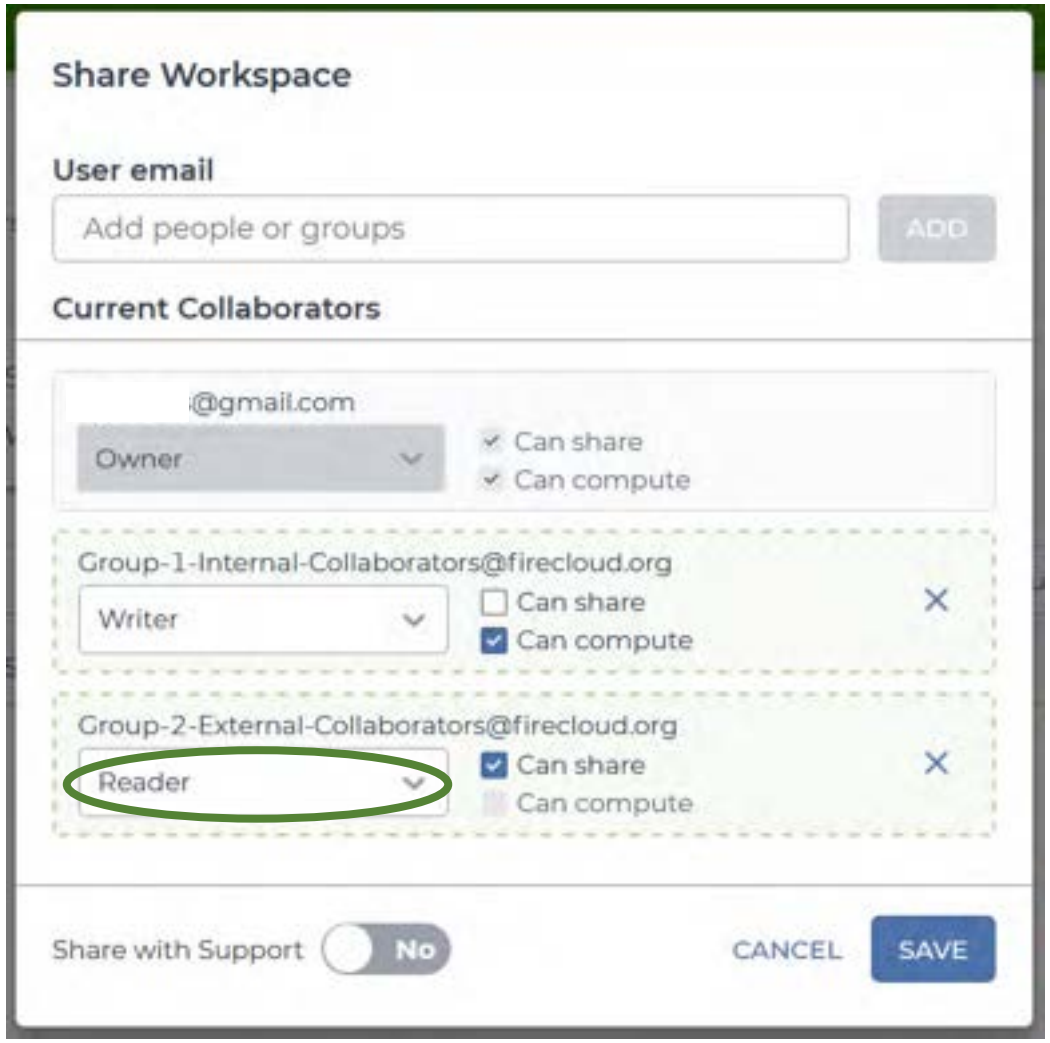
- access your data
- perform computations
- help you to write the notebooks you will use to share results with the scientific community

select the "**Writer**" role for this group in the drop-down menu and check the "**Can compute**" box

Any user belonging to the group will be able to **modify** your workspace and analysis notebooks, **and perform computations** for which you will be billed through your Terra Billing Project

Notice how you are the owner of the workspace, and can add other co-owners if desired

Sharing the workspace



Assigning permissions – Group 2 (Reader)

Since you want this group to be able to see your data, notebooks and analyses, but without the possibility of modifying them, select the "**Reader**" role for this group in the drop-down menu and do not check the "**Can compute**" box.
If you also want the group to be able to share your work, check on the "**Can share**" box

Any user belonging to the group will be able to **see, but not modify** your workspace and analysis notebooks, and they will **not be able to perform computations** – unless they copy your notebooks and create their own Terra Billing Project, sustaining the cost of the computations

Billing permissions

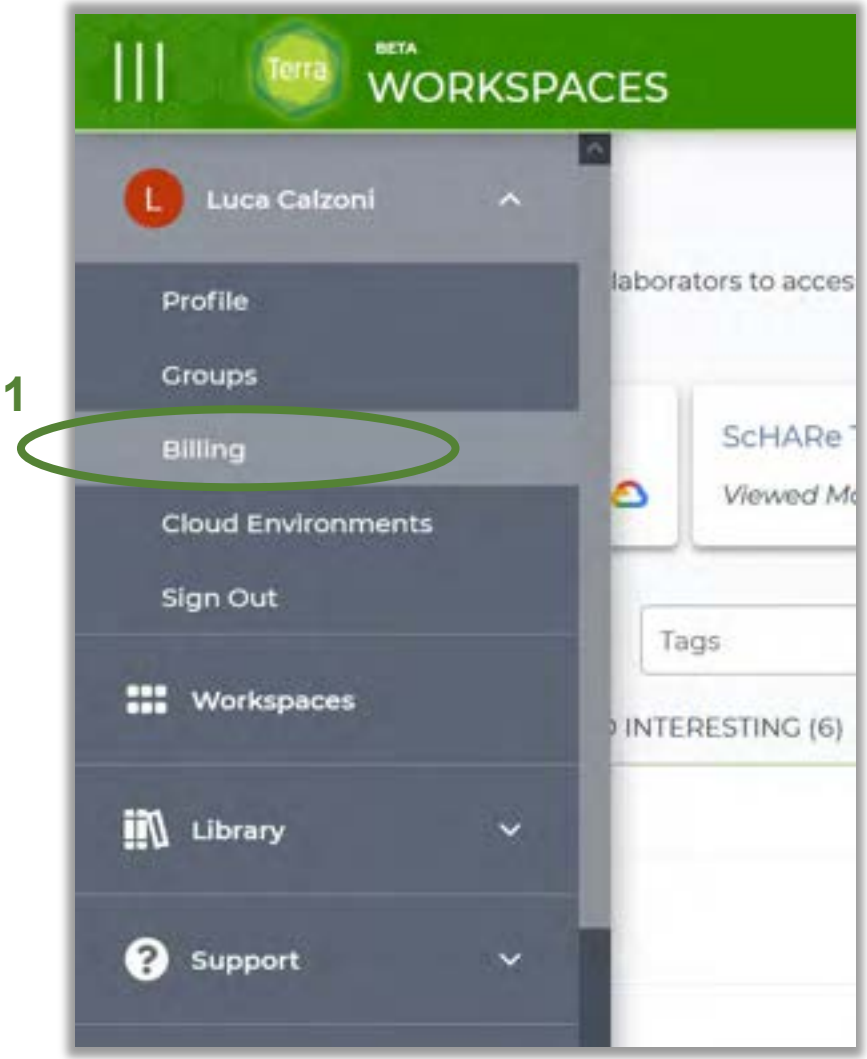
To allow collaborators from Group 1 to perform **computations** for which you will sustain the cost, you have to give them **permission to use your Terra Billing Project**

We will demonstrate how you can do this using the free temporary billing project associated to your workspace

Note - This part of the tutorial is **for demonstration purposes only**. As owners of the free billing project you used to create your demo workspace, we are the only ones able to share it with others. Revisit these slides **after creating your own billing project: you will be able to replicate these steps** using the billing project you own

Instructions:

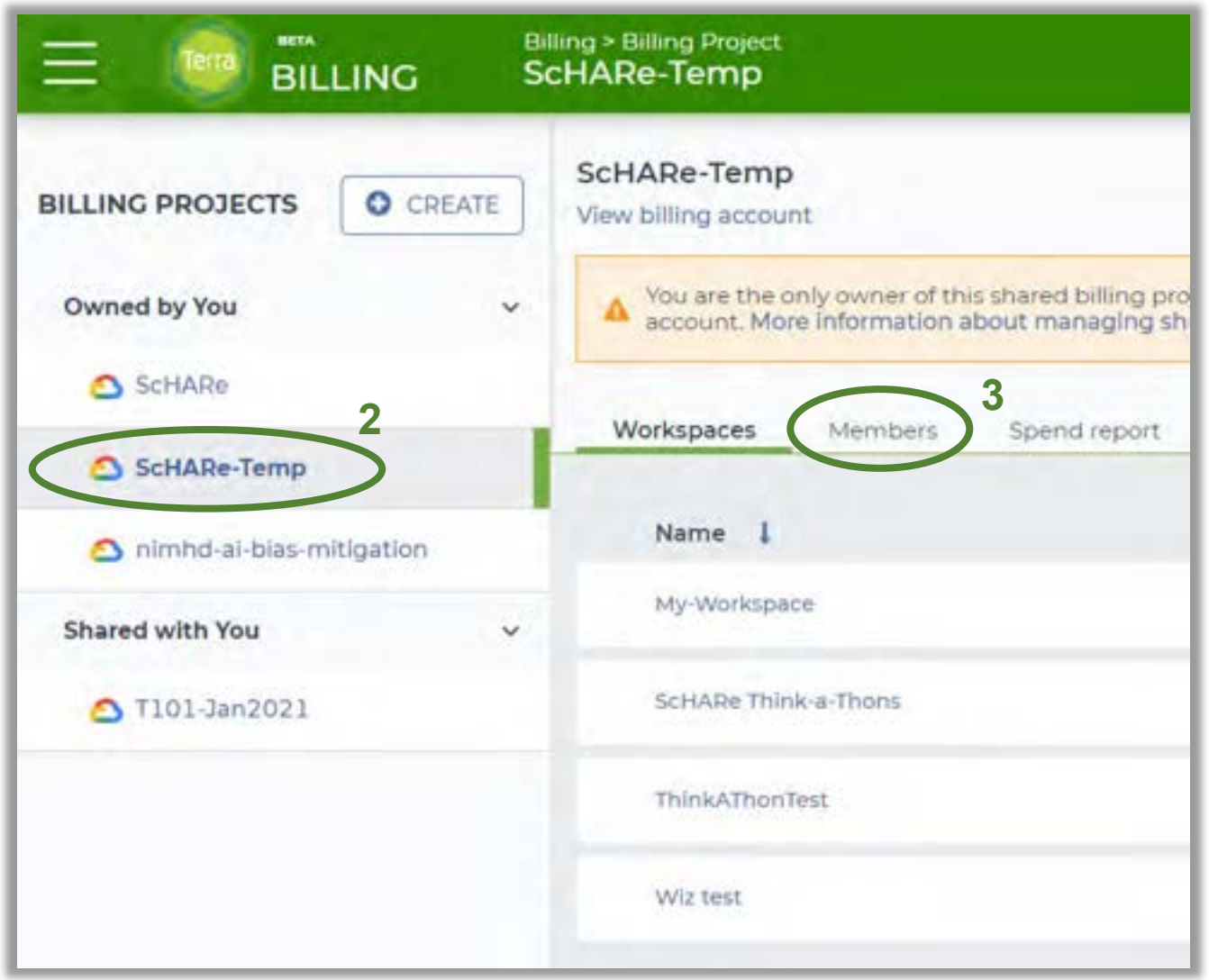
1. **Click on the menu** in the top left corner of the page, then on **"Billing"**



Billing permissions

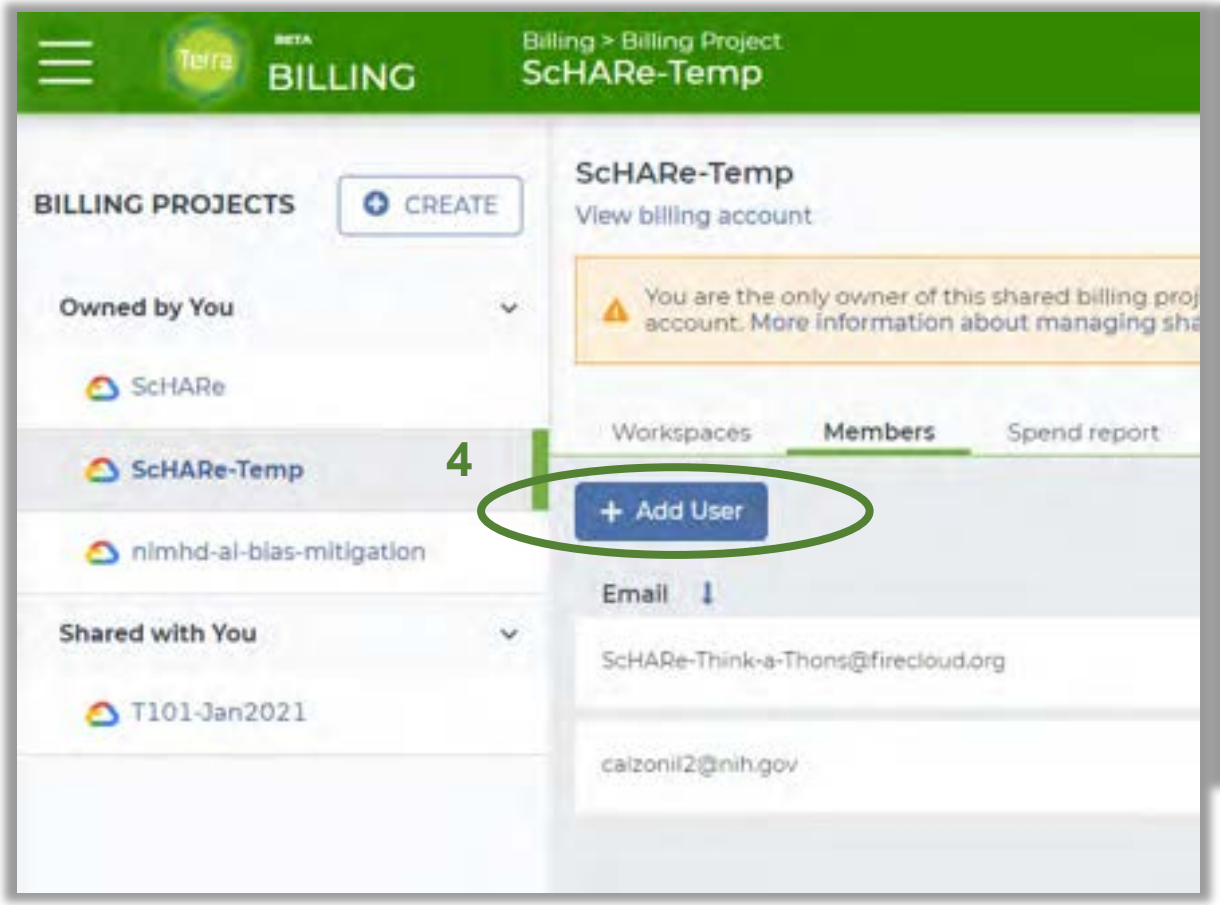
2. Identify the **billing project associated with your workspace** in the list of Billing Projects owned by you, and **click on it**

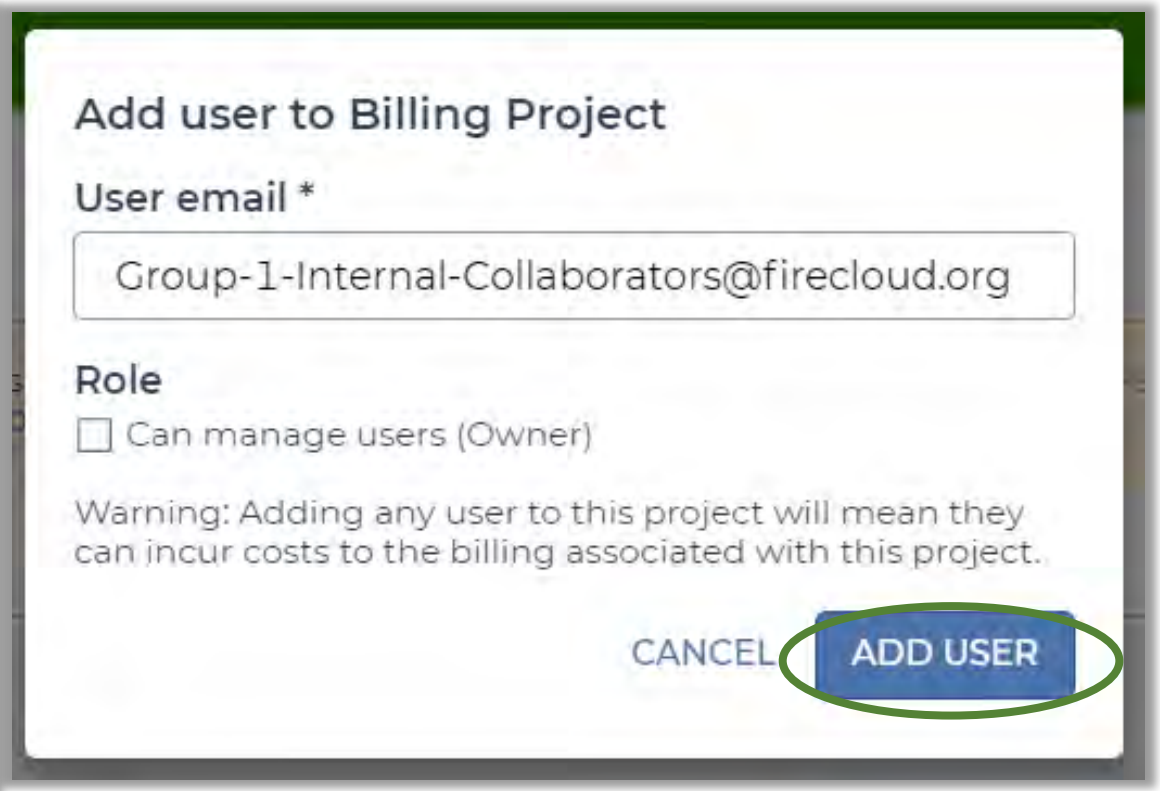
3. Click on the **“Members”** tab



Billing permissions

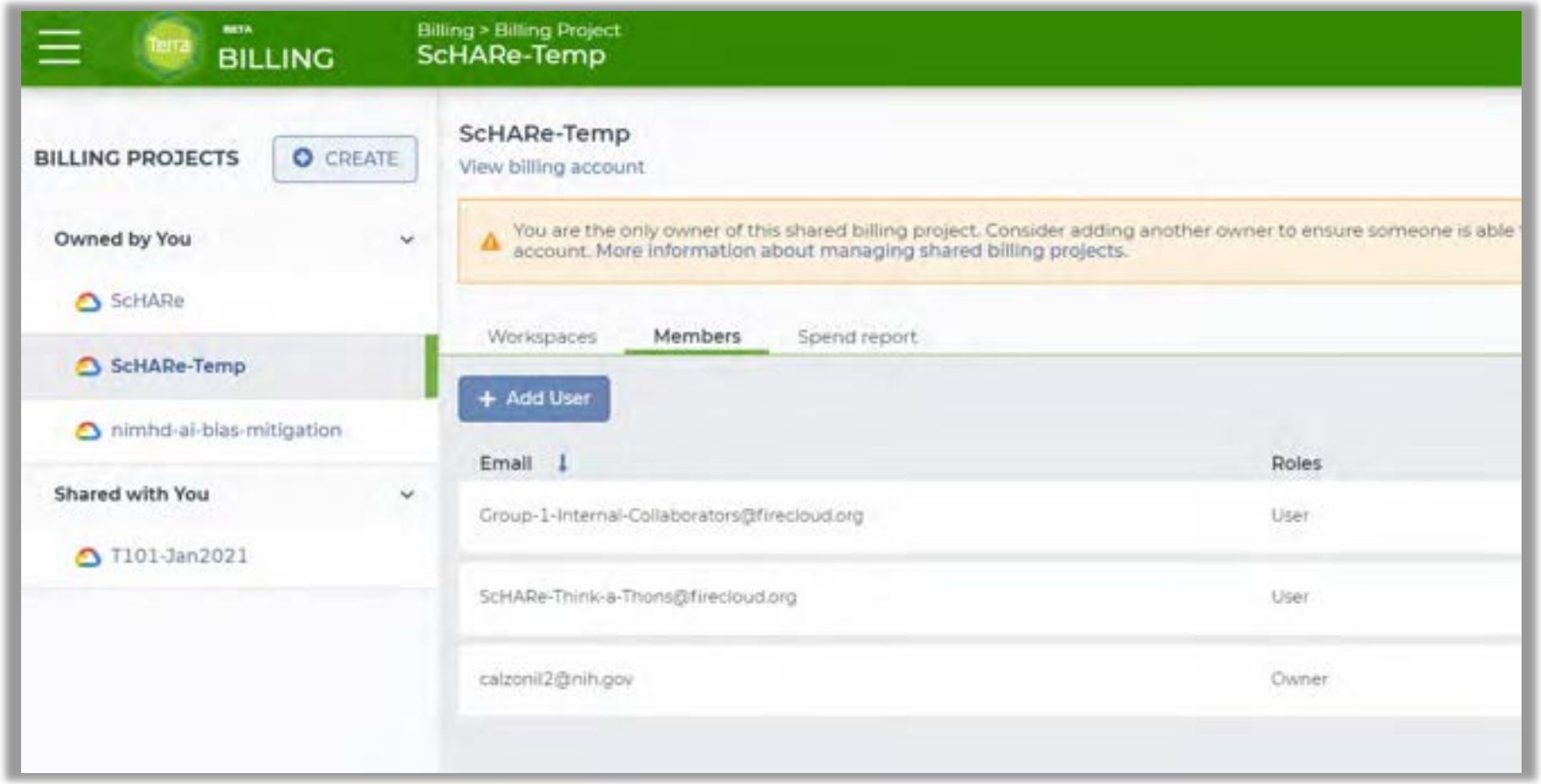
4. Click on “Add User”, add the **group email** associated with collaborators from Group 1 to the Billing Project, and confirm. If you prefer, you can add specific **individual collaborators** instead of the entire group





Billing permissions

Success! Your collaborators can now use your Billing Project to perform computations in the cloud



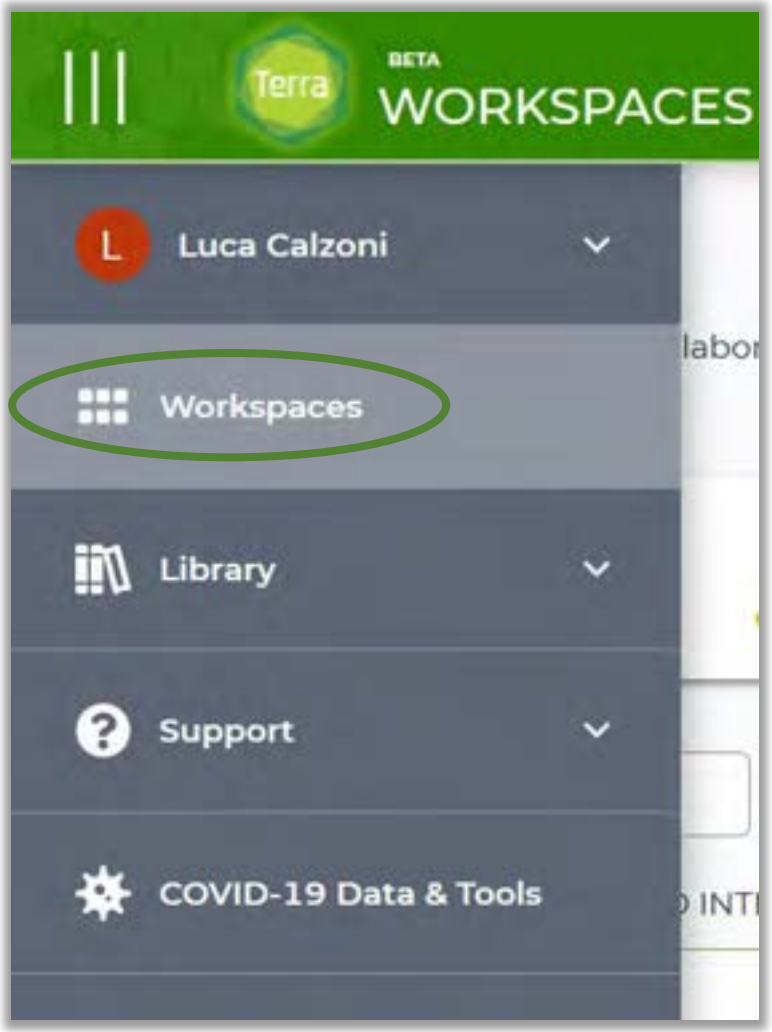
Cloning an existing workspace

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

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

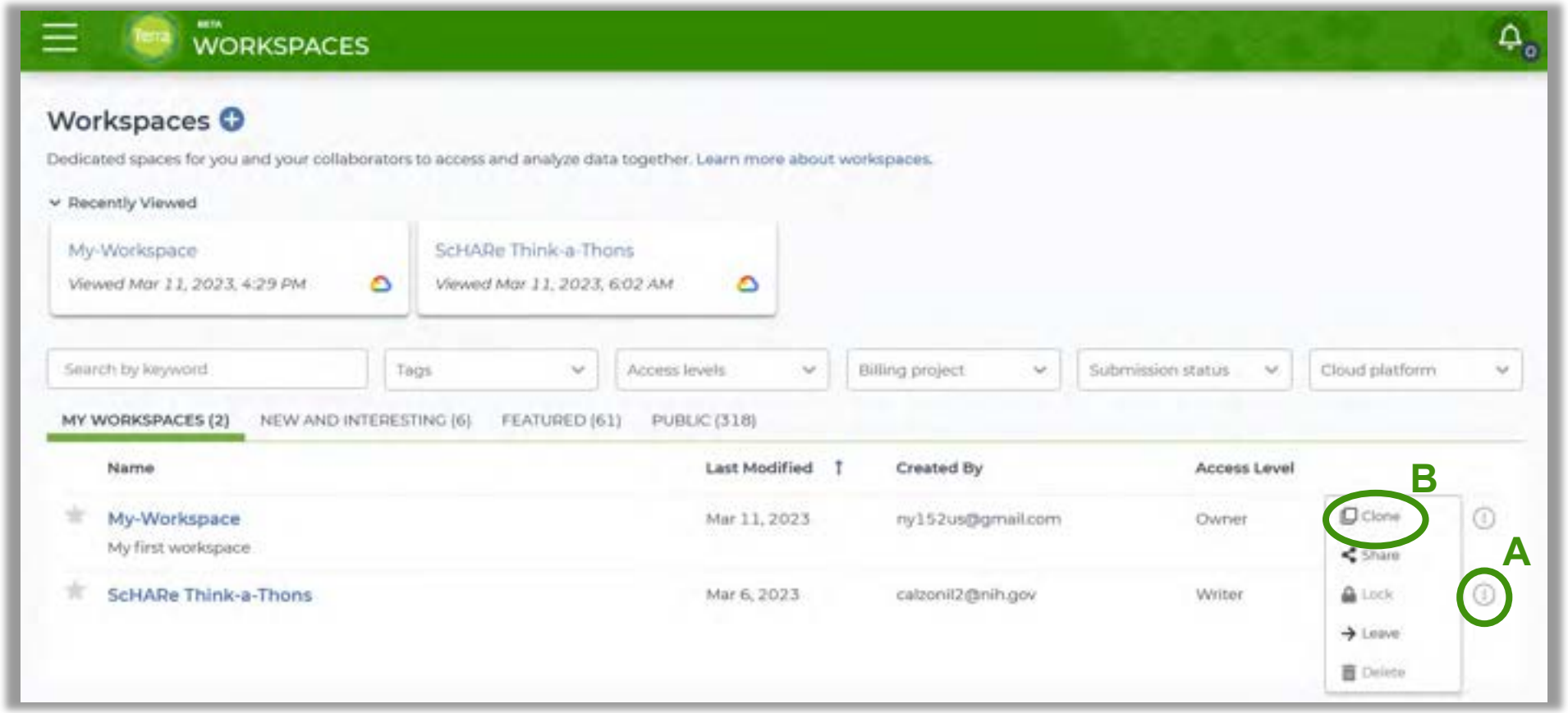
Here is how you can do it. In this example, we will clone the workspace “ScHARe Think-a-Thons”, a copy of the ScHARe workspace we created for this Think-a-Thon

1. **Click on the menu** in the top left corner of the page, then on **“Workspaces”**



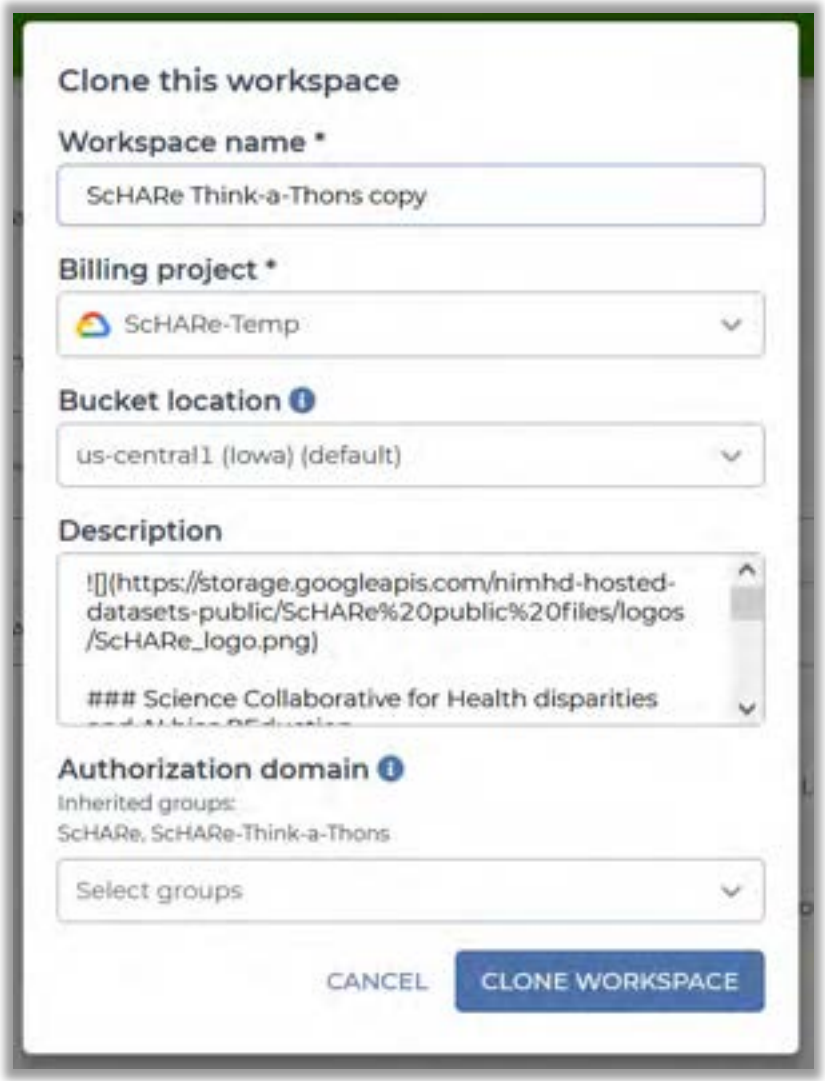
Cloning an existing workspace

2. **Identify the workspace you want to clone** in the list of workspaces provided on screen and click on the corresponding **vertical three-dot menu** (A), then on **“Clone”** (B)



Cloning an existing workspace

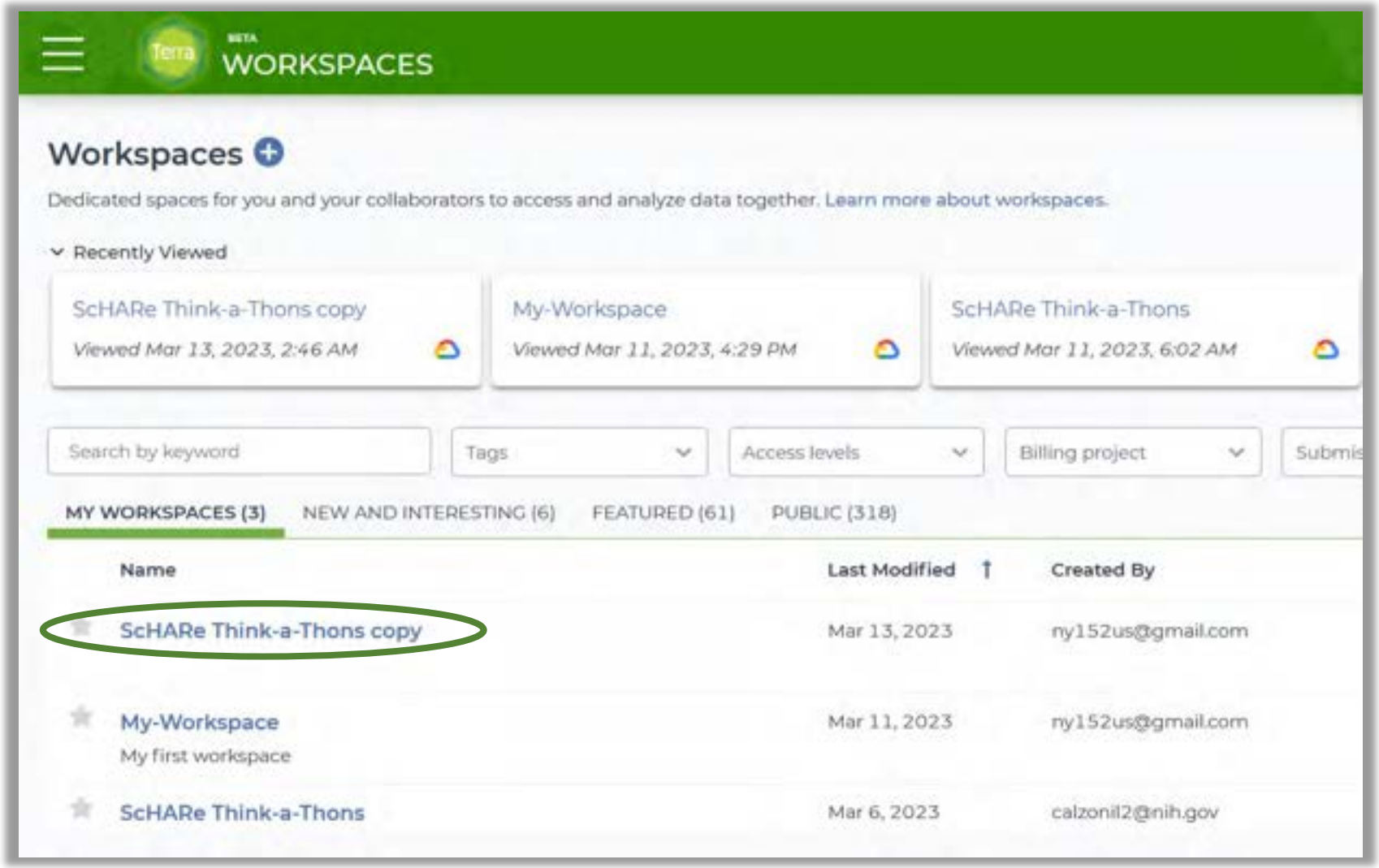
3. Input a **name** for the workspace copy
4. Select the **Billing Project** you want to associate with the workspace. For this example, you can select our free temporary Billing Project “ScHARe-Temp”
5. Select the **bucket location**. A bucket location can only be set when creating a workspace. For 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. Add an additional AD if needed



Cloning an existing workspace

Success! The cloned workspace is now listed among your workspaces

You can freely access all of its **resources**



Part IV
Notebooks and Environment

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 this tutorial you should be familiar with **programming**. If you are not, the code in our notebooks is very easy to understand and reuse, and our tutorial will still 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 integrates code and its output into a single document where you can run code, display the output, and also add explanations, formulas, and charts

Using notebooks:

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

ScHARe notebooks

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

A list of the available notebooks is provided on the right. **We will access and run a copy of one of these notebooks**, as an example

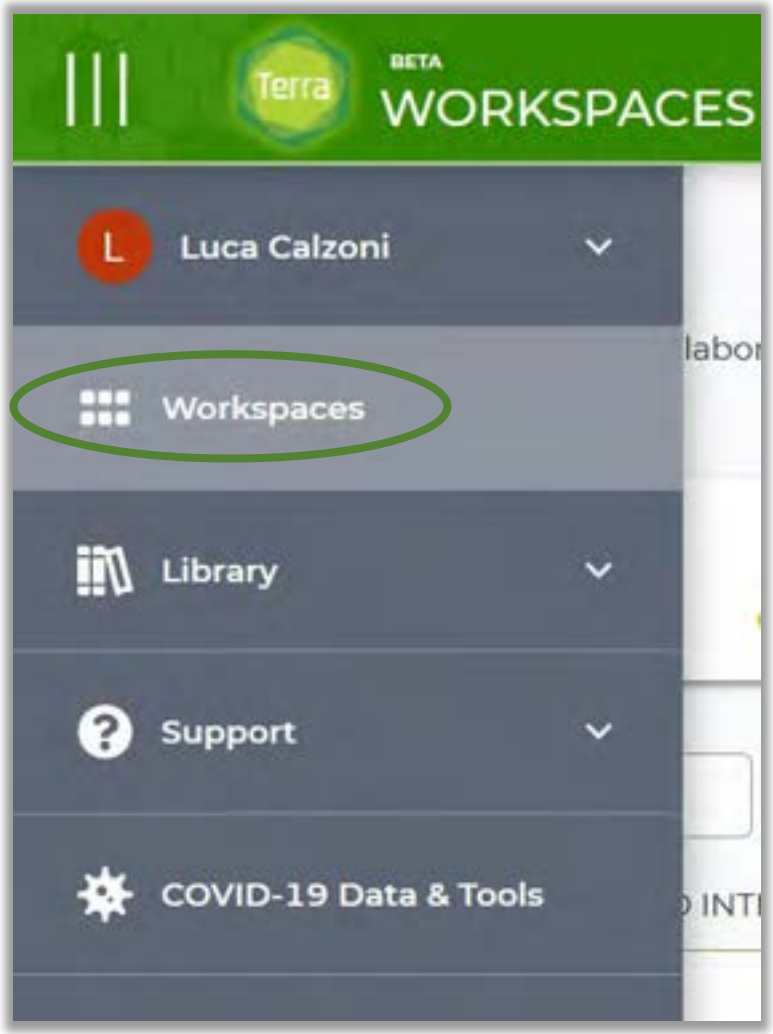
List of ScHARe instructional notebooks

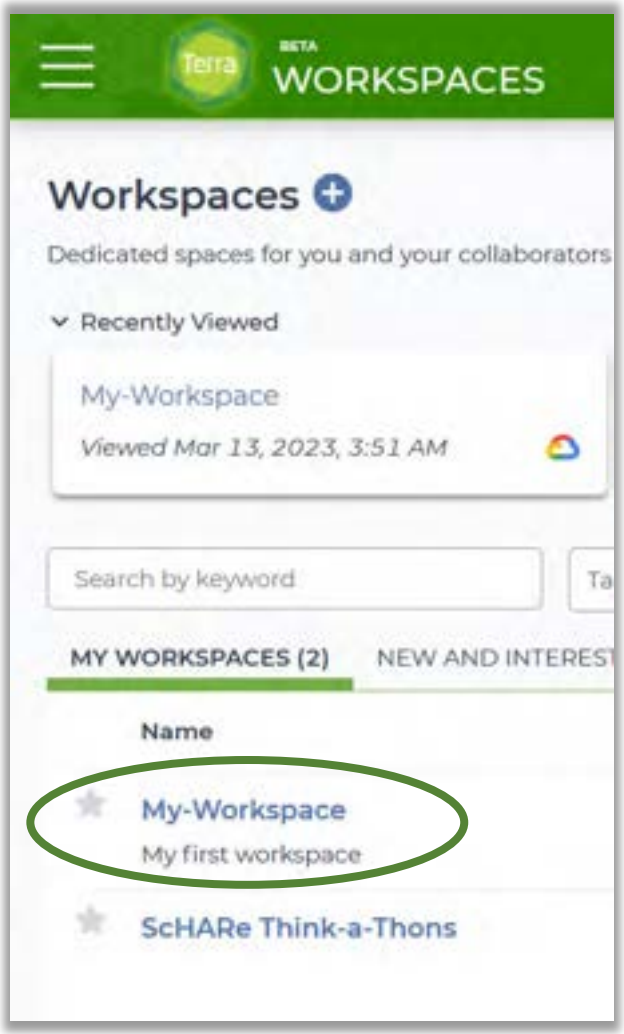
- **00_List of Datasets Available on ScHARe**: a list of the datasets available in the ScHARe Datasets collection.
- **01_Introduction to Terra Cloud Environment**: an introduction to the Terra platform and cloud environment.
- **02_Introduction to Terra Jupyter Notebooks**: an introduction to Jupyter 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.

Creating a notebook

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 today

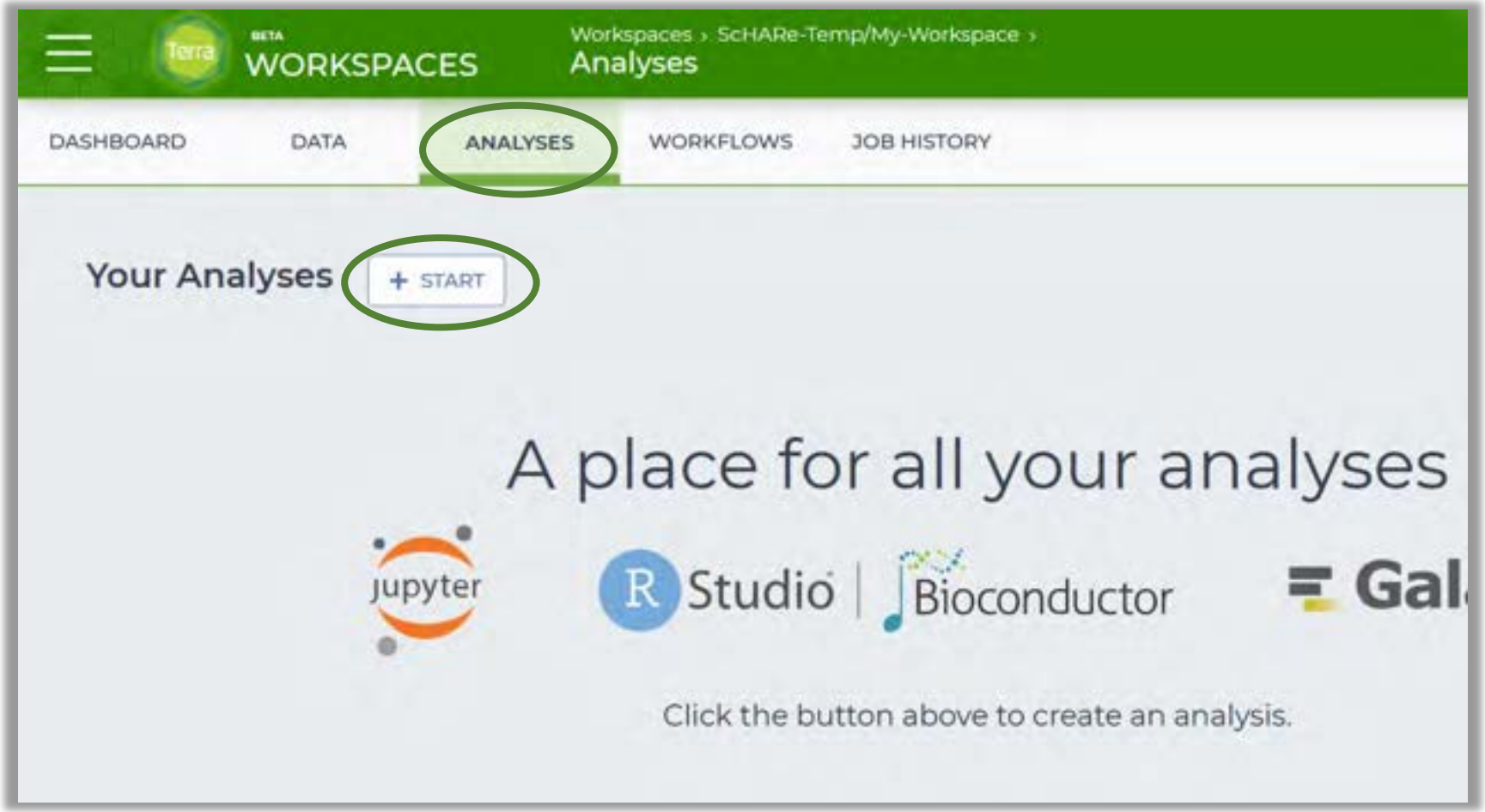




Creating a notebook

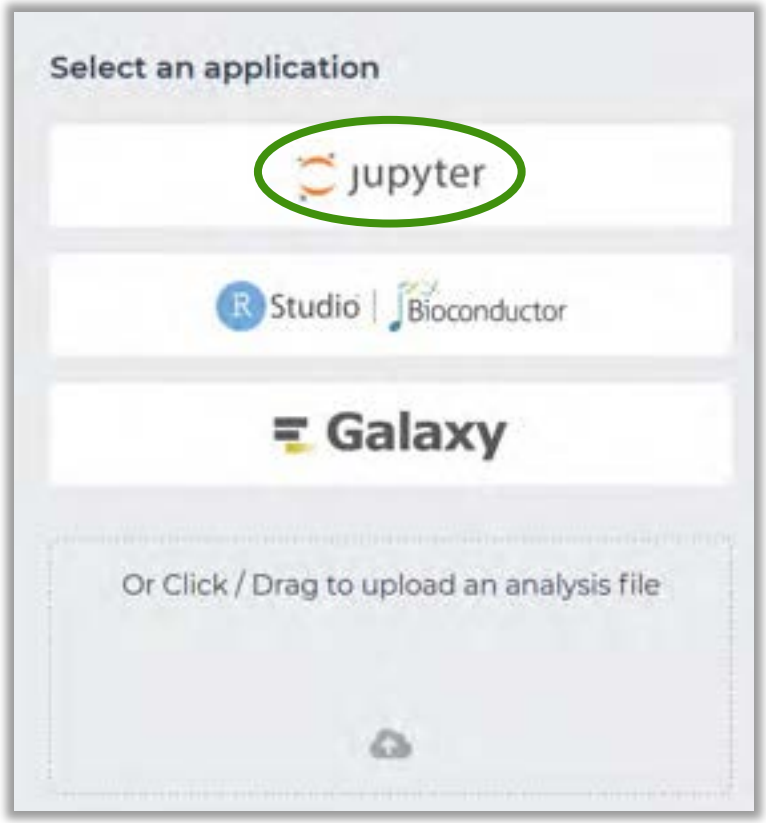
3. Click on the **“Analyses”** tab

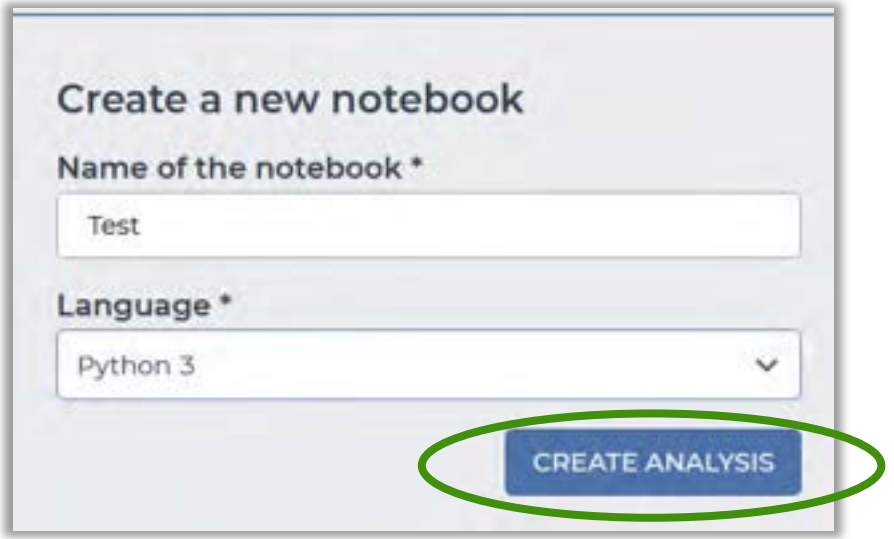
4. Click on the **“Start”** button



Configuring the environment

5. Select “**Jupyter**”
6. In the next window, assign a **name** to the notebook and choose a programming language (“**Python 3**”)
7. Click “**Create analysis**”

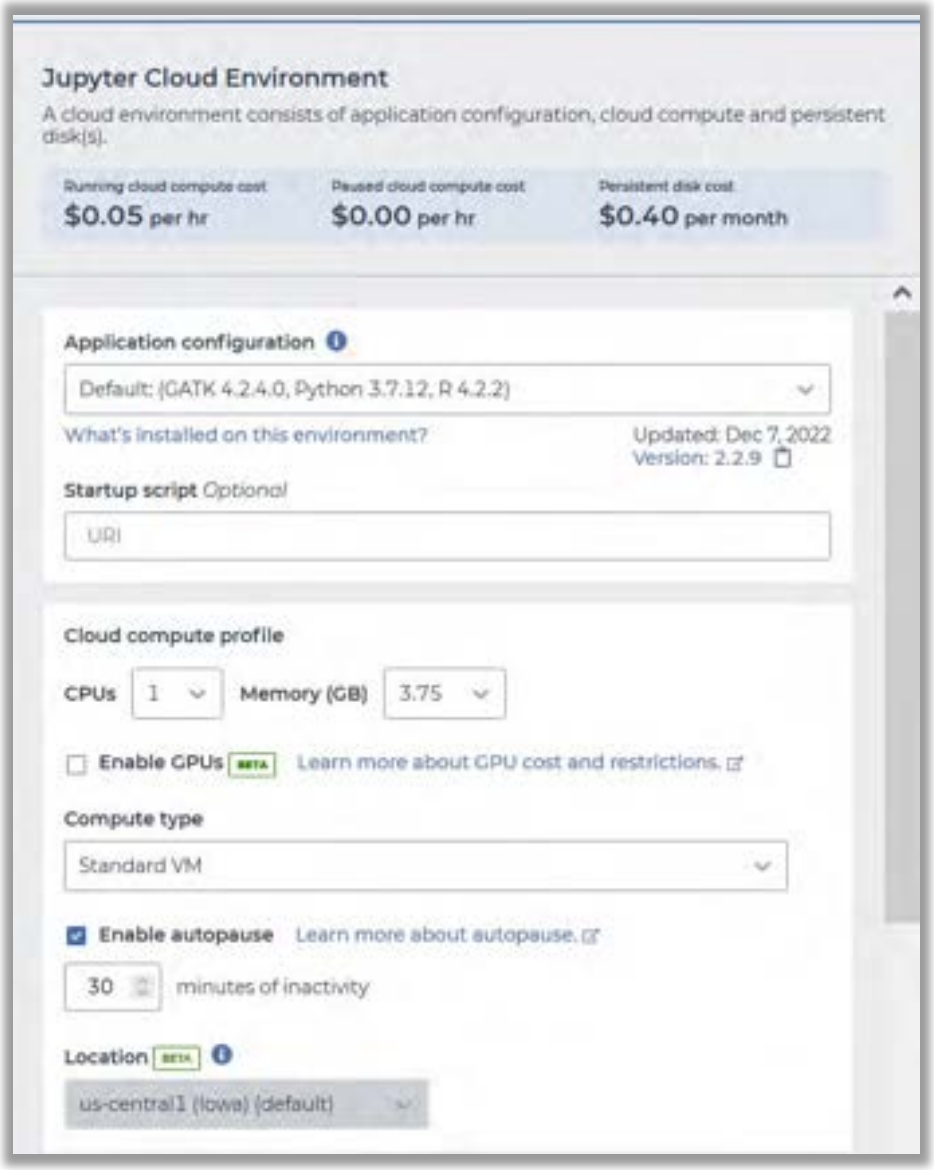




Configuring the environment

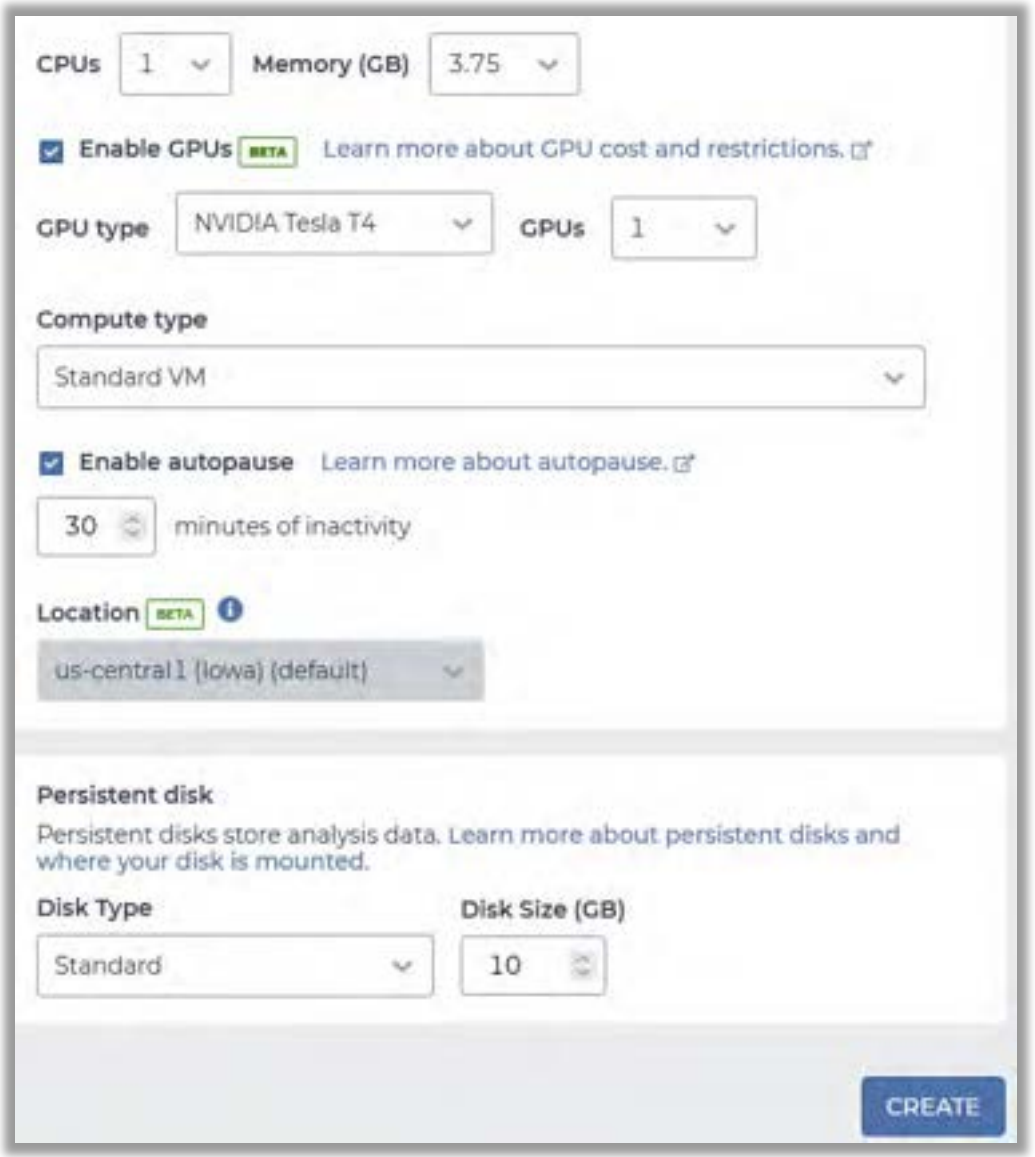
8. You will now be asked to configure your **Cloud Environment.** In this window, you can **configure** the on-demand availability of **computer system resources**, especially data storage and computing power, needed to perform your computations:
   - **Application configuration:** the software application + programming languages + packages used to create your cloud environment. You can leave this unchanged
   - **Cloud compute profile:** the CPU and RAM available to run your application, which determines how much processing can be done at a time.

Notice how changing the parameters changes the estimated costs to run at the top of the window. **Find the balance that's right for you.** More compute power costs more, and you don't want to request (and pay for) significantly more than your computation needs



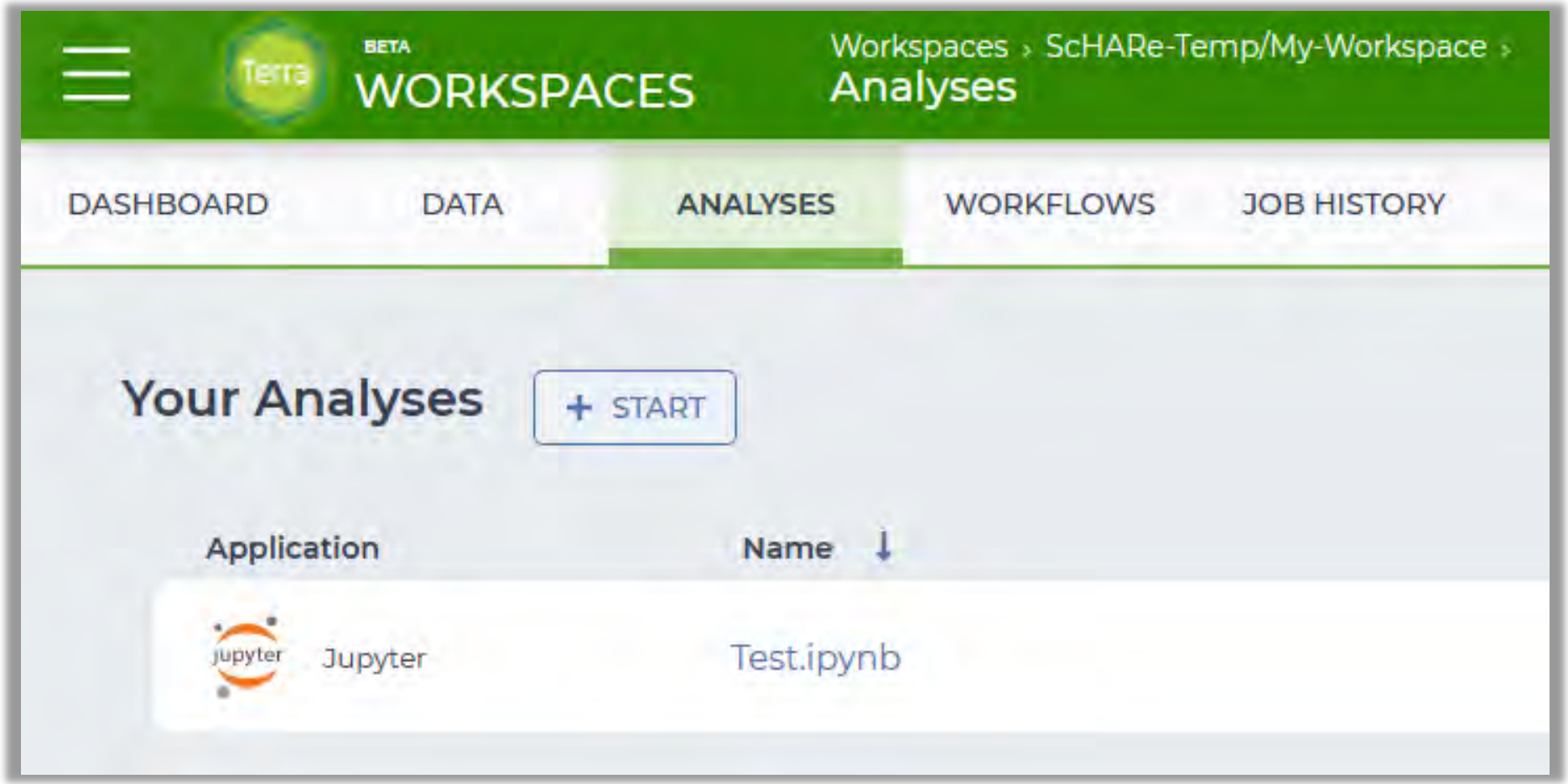
Configuring the environment

- **CPUs, Memory**: leave the default values unchanged
- **Enable GPUs**: checking the box allows you to add graphics processing units (GPUs) to your environment. GPUs are optimized for linear algebra computations, such as matrix multiplication, and for training artificial intelligence and deep learning models. Leave the box unchecked
- **Enable autopause**: recommended. Once your browser becomes inactive and **30** minutes pass, the cloud environment autopauses to save on costs
- **Location**: physical location where the data is going to reside. Leave the default value unchanged
- **Persistent disk:** Terra attaches a persistent disk to your cloud compute to provide an option to keep the data on the disk after you delete your compute. A **minimal cost per hour** is associated with maintaining the disk, even when the cloud compute is paused. Select “Standard” and “10 GB” and click on “**Create**”



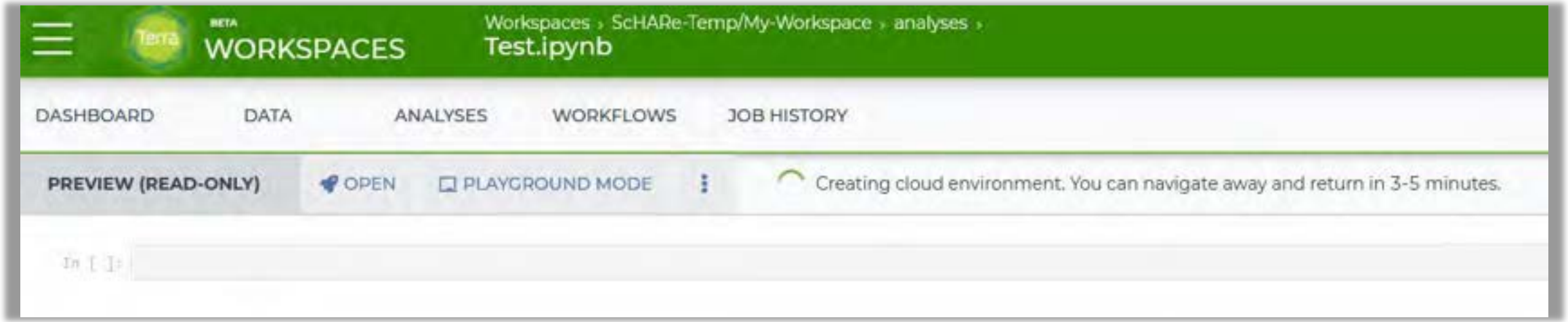
Configuring the environment

- **Success!** Your environment has been configured, and your notebook created. **Click on its name** to open it



Configuring the environment

- Creating the cloud environment **will take 3-5 minutes**

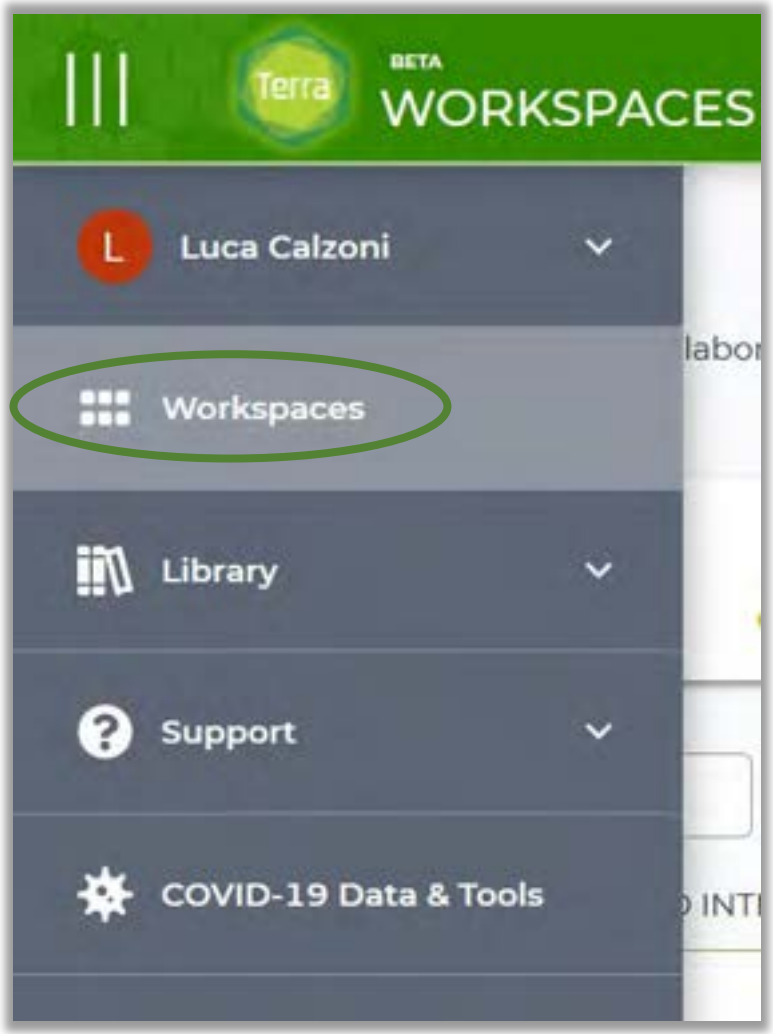


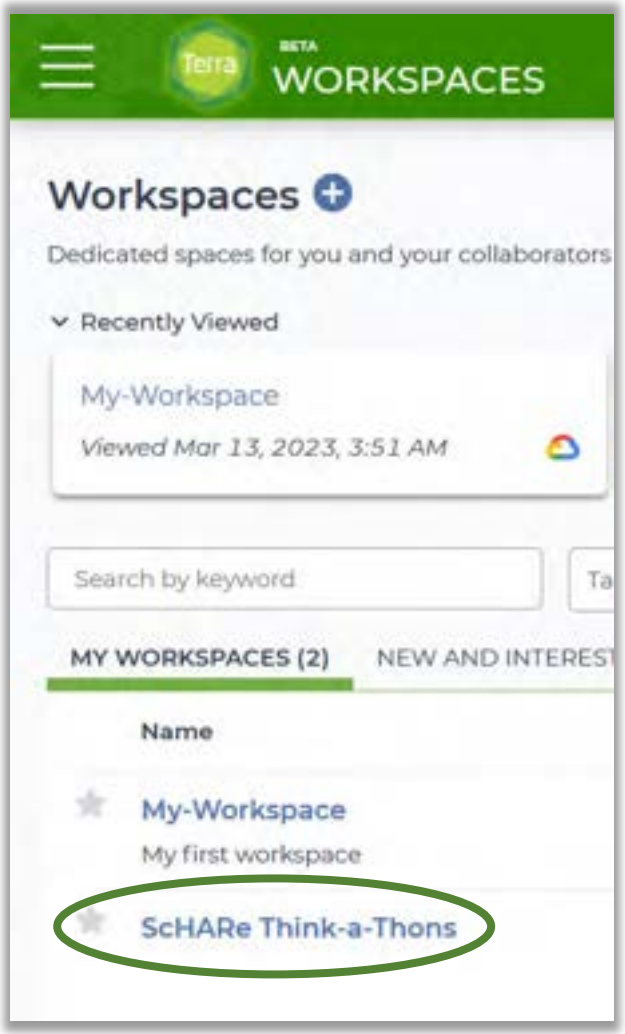
Since this newly-created notebook is empty, we will open and run one of the **ScHARe instructional notebooks** instead, to give you a closer look at **how notebooks work.**

Running a notebook

1. **Click on the menu** in the top left corner of the page, then on **“Workspaces”**

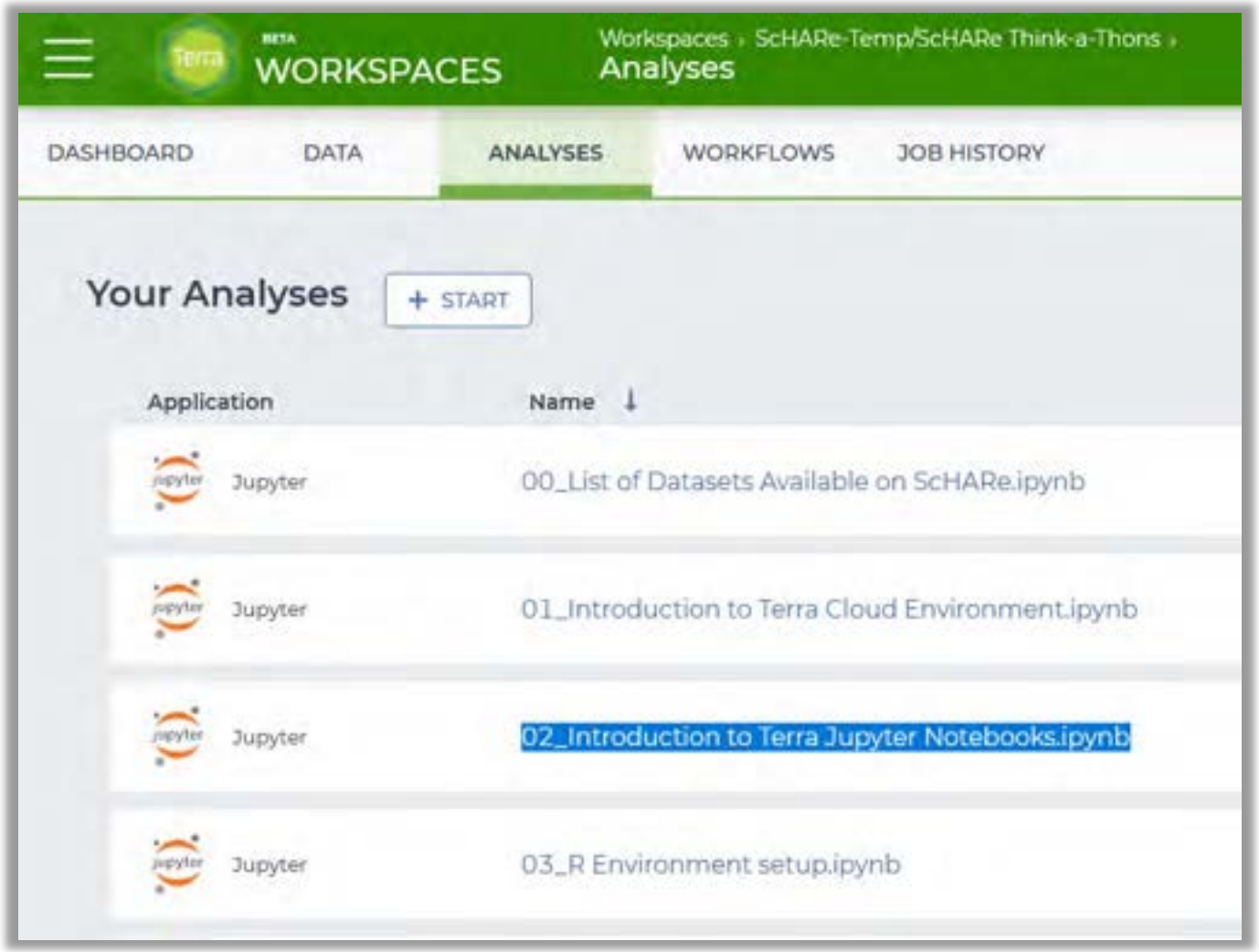
2. **Click on the ScHARe Think-a-Thons workspace** that was shared with you by the instructors





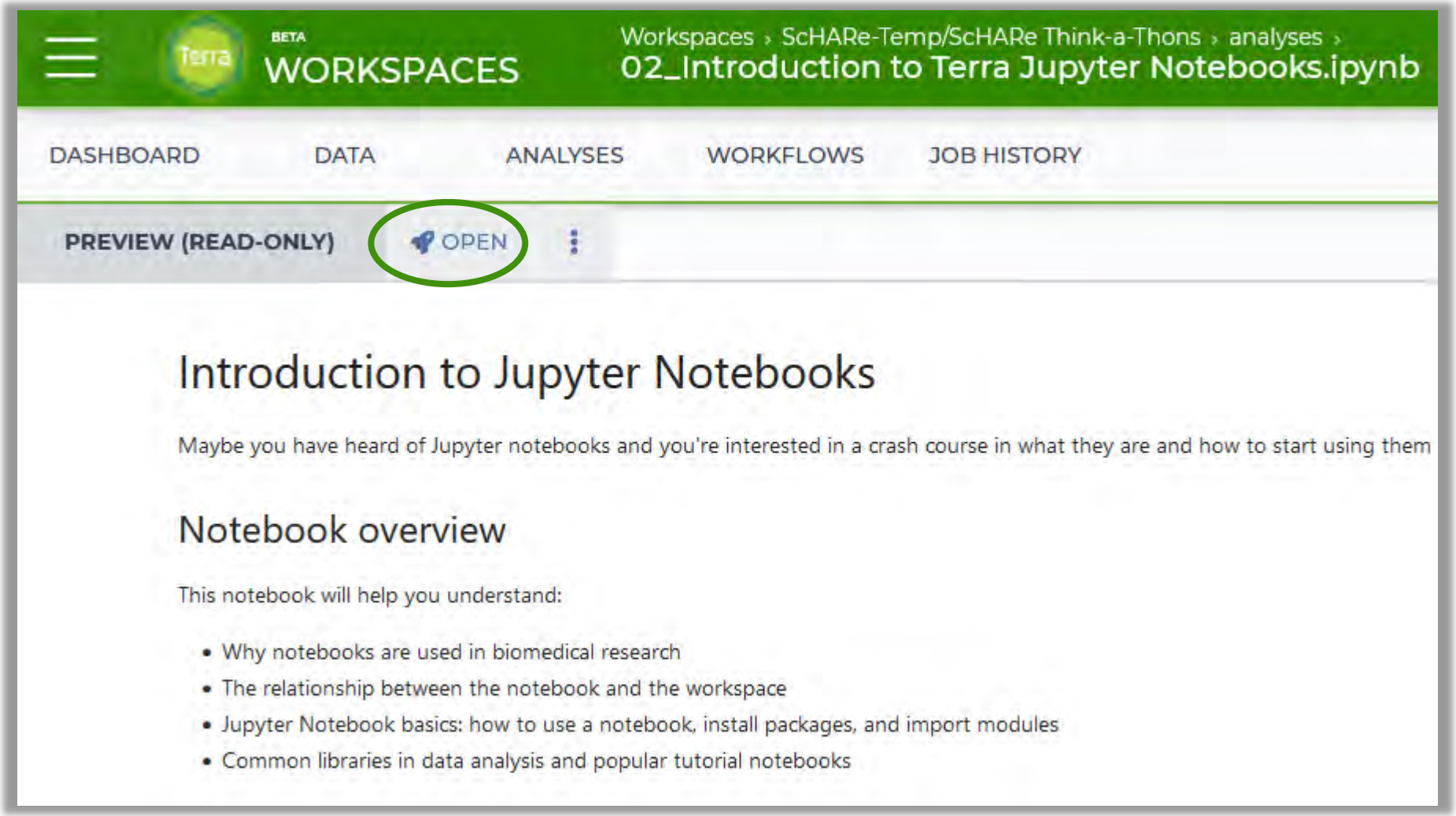
Running a notebook

3. Click on the **“02_Introduction to Terra Jupyter Notebooks”** notebook



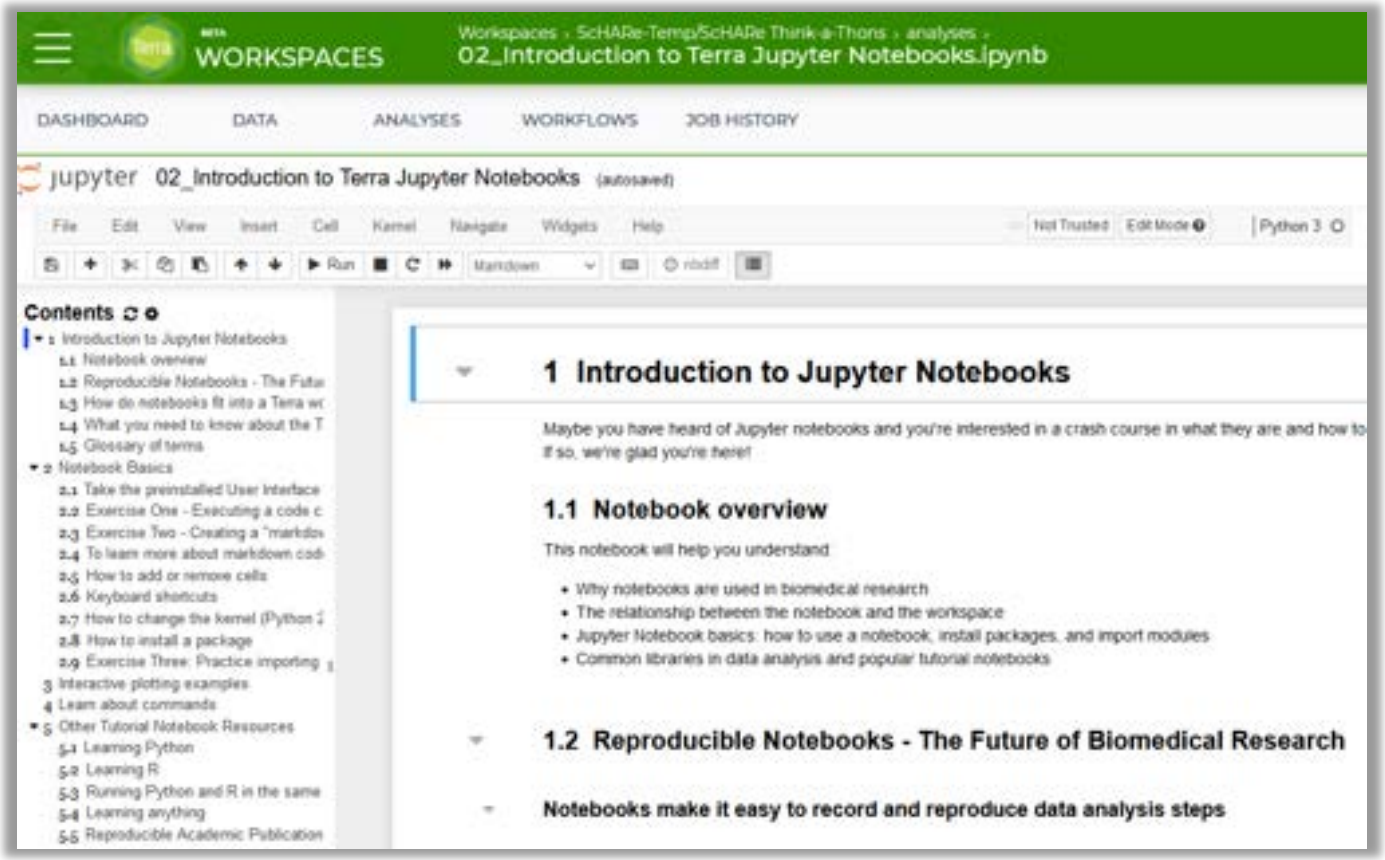
Running a notebook

4. Click on "**Open**"



Running a notebook

5. Configure the environment as shown before, then click on "**Create**"
6. **Follow the step-by-step instructions in the notebook**, which will help you understand:
   - why notebooks are used in research
   - the relationship between the notebook and the workspace
   - how to use a notebook, install packages, and import modules
   - common libraries in data analysis



Part V
Billing and Costs

What are the cloud costs of working on Terra?

The Terra platform is **free to use**

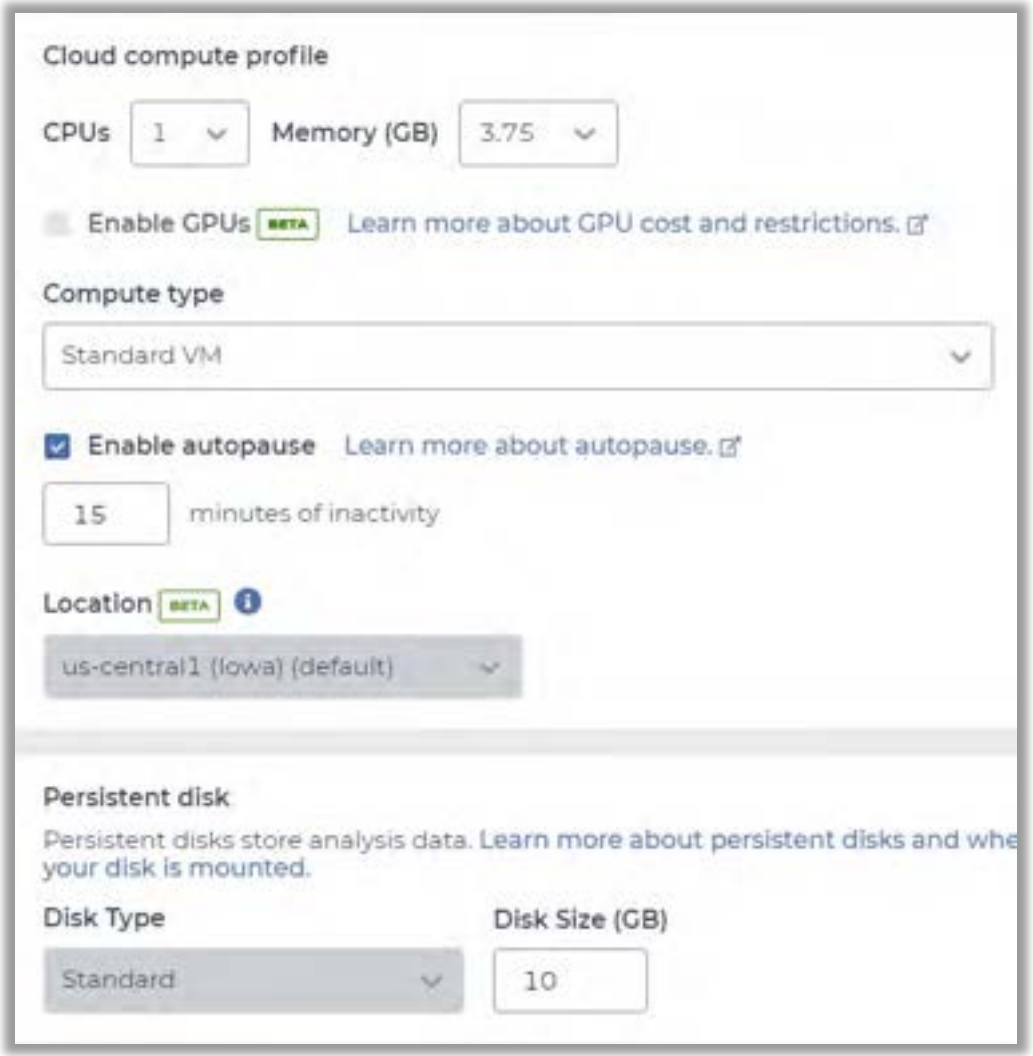
However, the following operations in Terra **may incur charges:**

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**

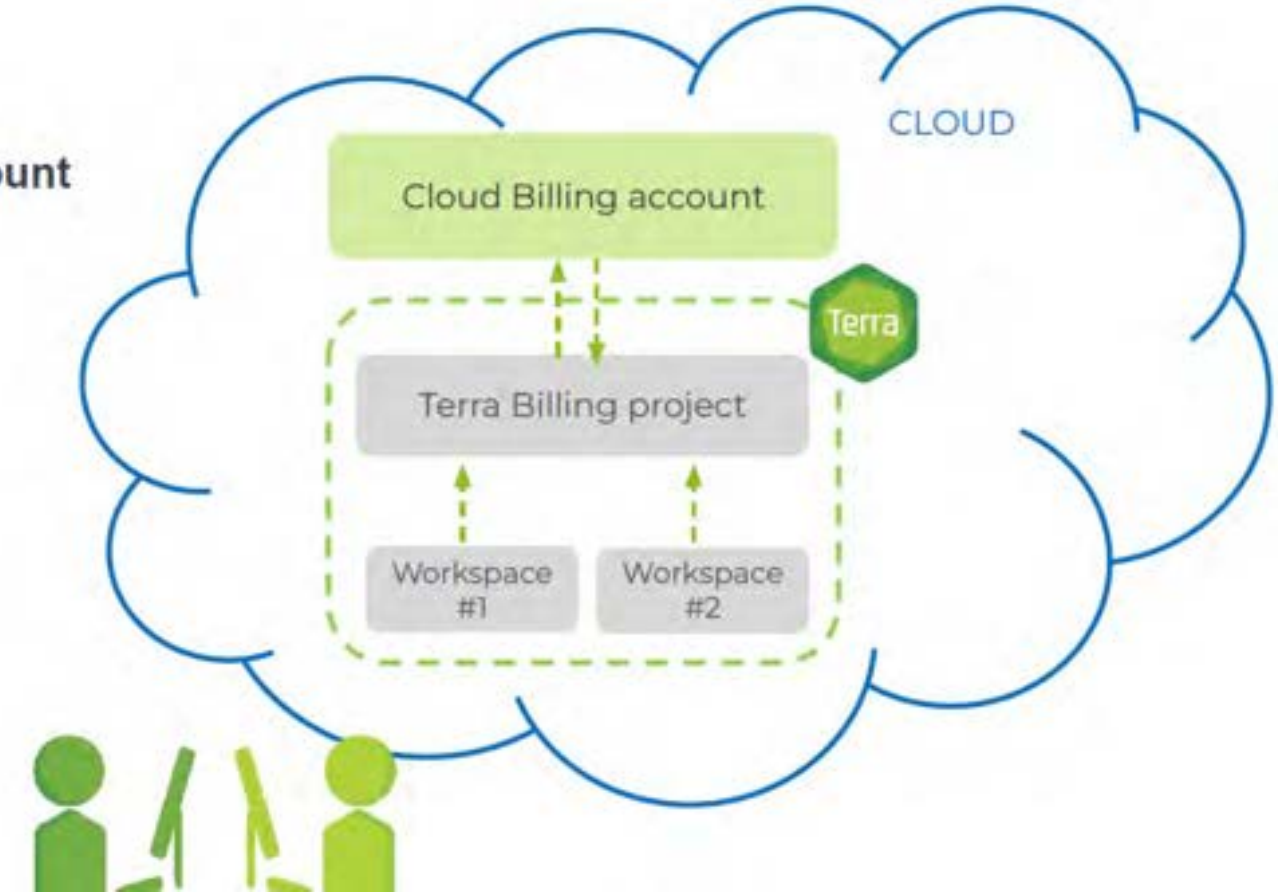
- As we will see later, when creating a bucket to store data you are asked to set its location ("**region**"), i.e. the physical place where the data is going to reside (e.g.: "US-CENTRAL1 – Iowa", or "US-EAST4 - Northern Virginia"). More info on regions here. You will pay to **move stored data to a bucket in a different region**

- You will pay to **download data from a bucket that is configured as a Requester Pays bucket.** With such buckets, the data requester pays the cost of the data download instead of the bucket owner

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?

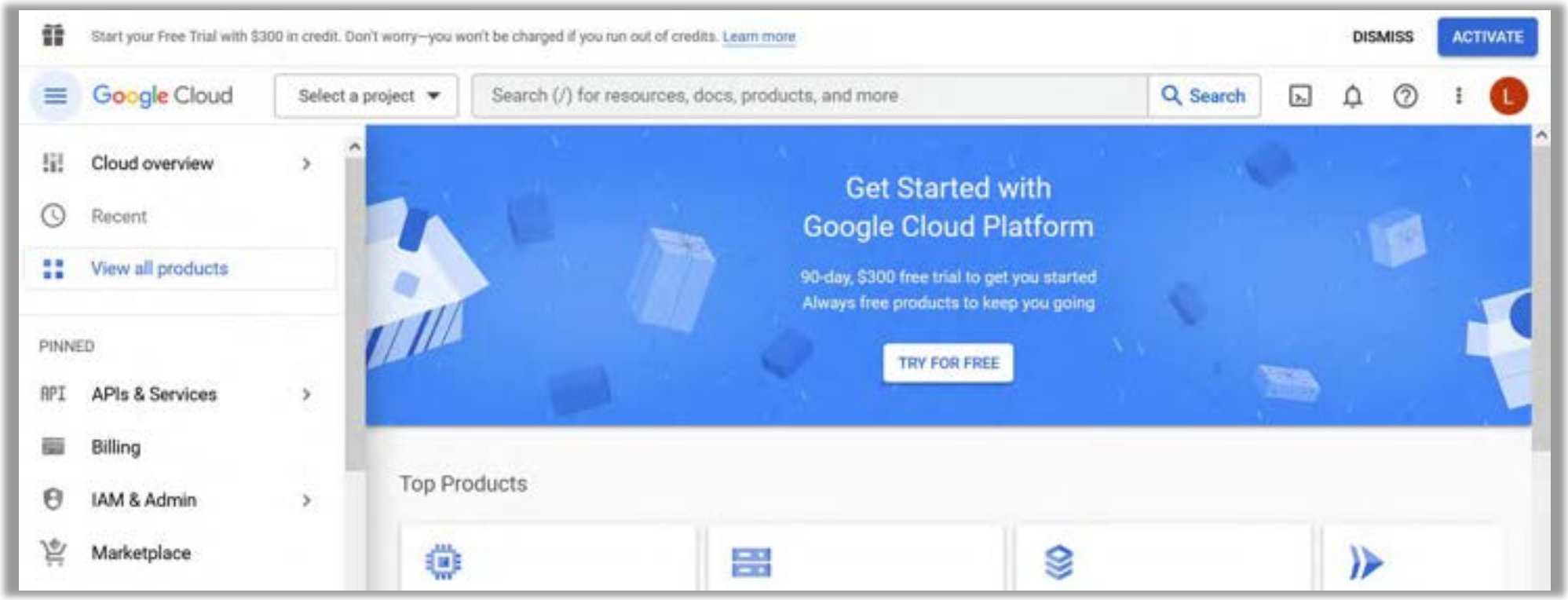
The credits will cover anything that has a cost in Terra - such as storing data and running analyses. You can’t use credits to add GPUs to your computing resources, and you are limited to 4 workspaces at a time

How long will my $300 credits be available?

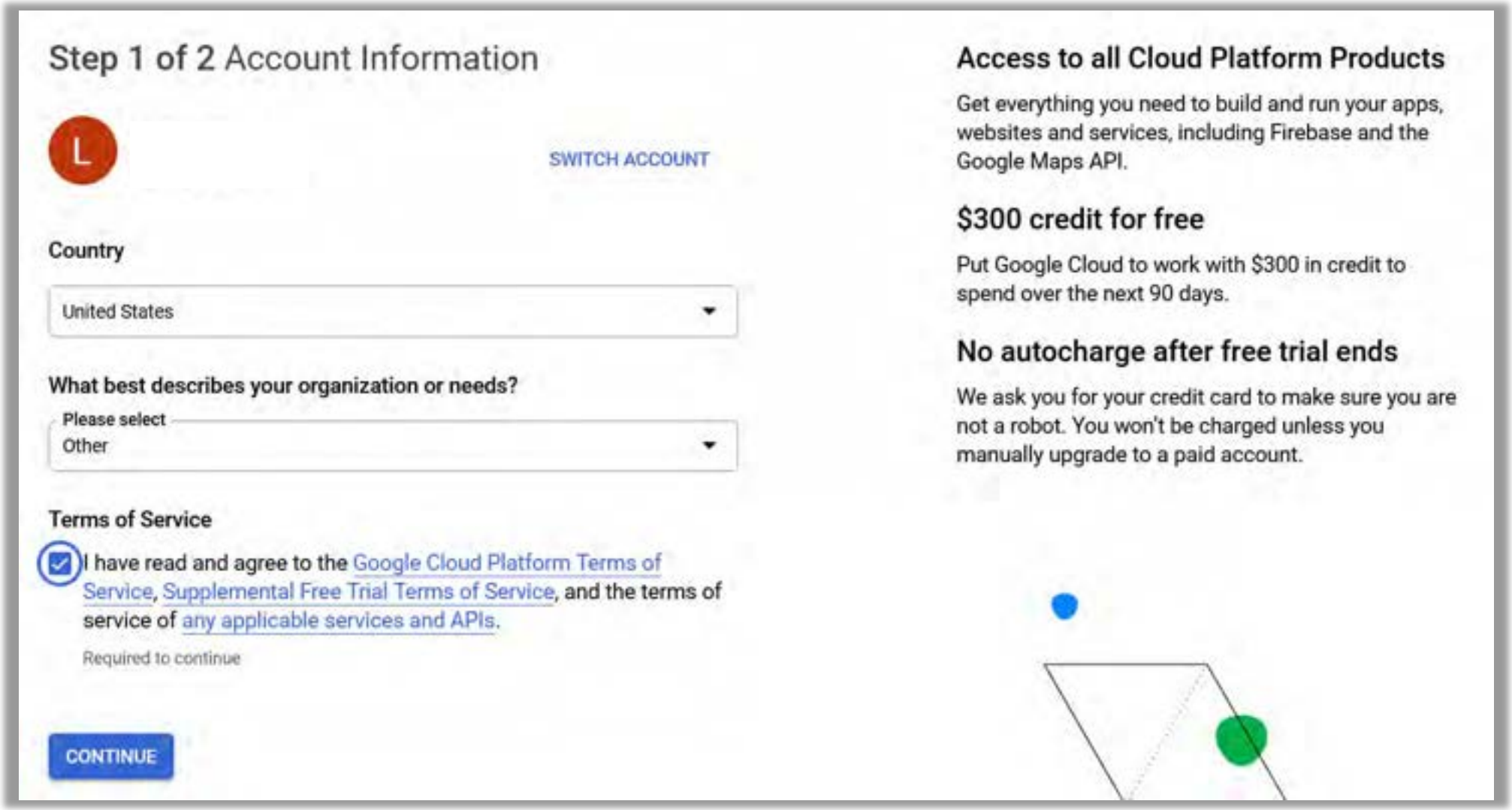
Your credits will be available for 3 months, or until you have used up all $300. Once your credits run out or expire, you can upgrade to a paid account

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 invitated to activate your free trial
2. Click the **activate** button and follow the instructions



Step 1. Set up a Google Cloud Billing account

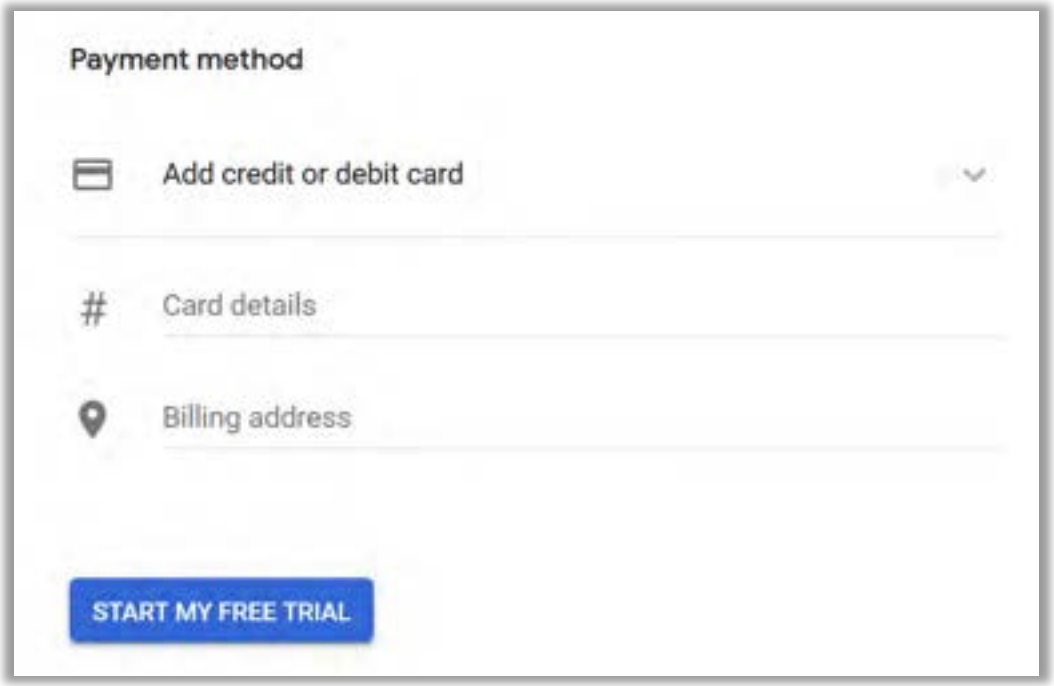


Step 1. Set up a Google Cloud Billing account

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 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 Cloud Billing account, you can create a Terra Billing project that 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**

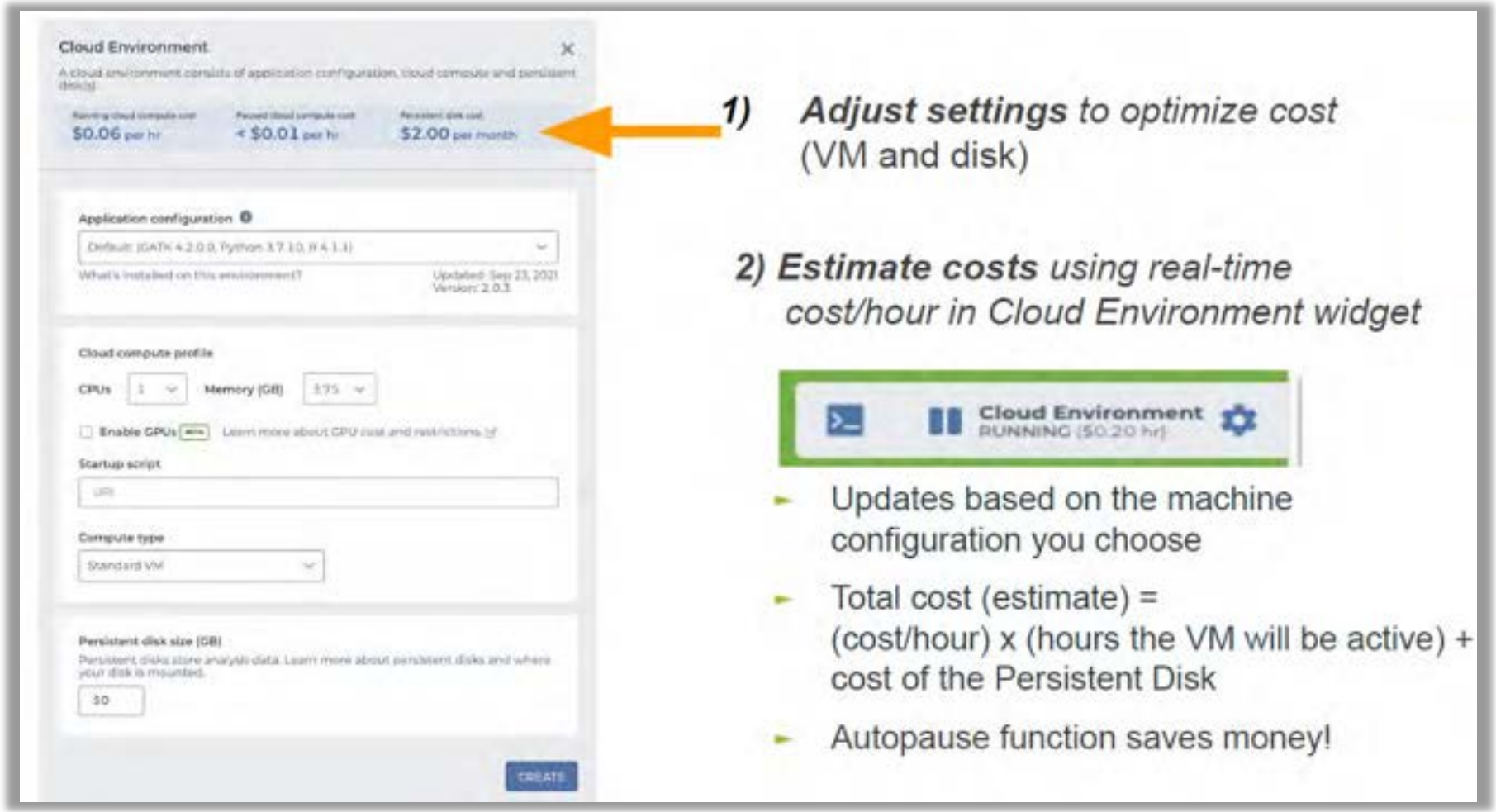
Understanding and monitoring costs

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)

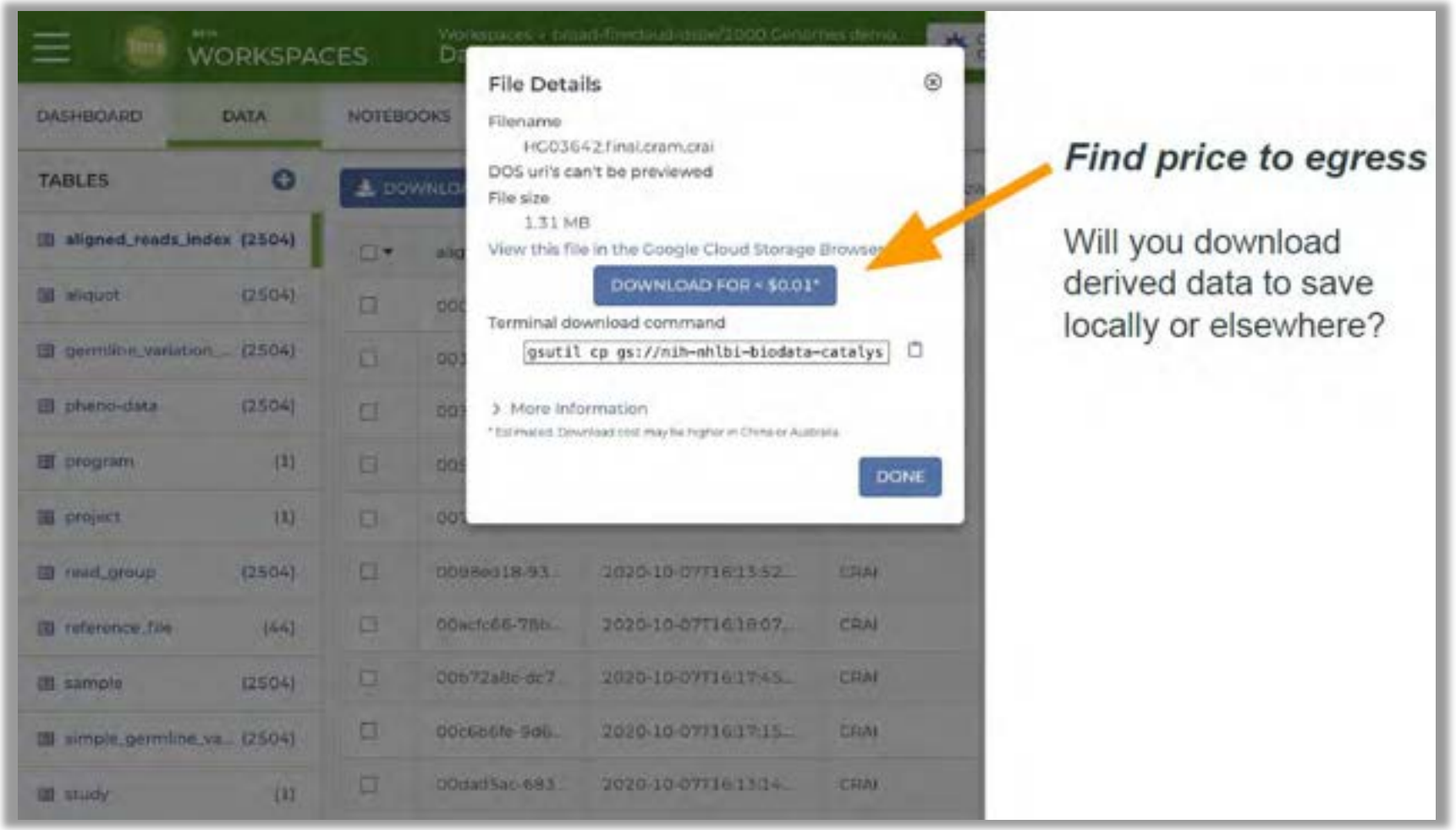
Understanding and monitoring costs

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)



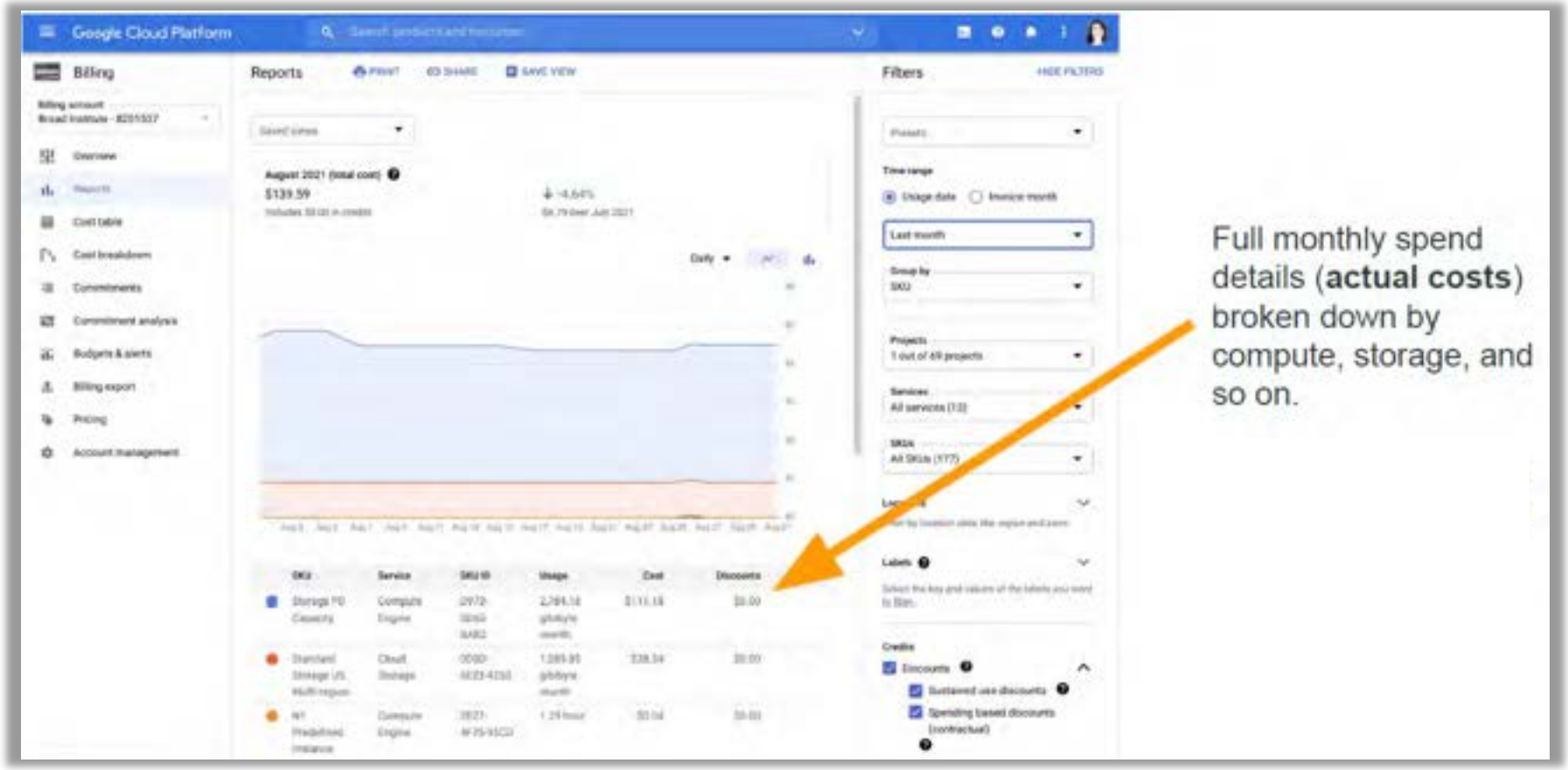
Understanding and monitoring costs

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

Next Think-a-Thons:
Data Sets Review and Analyzing Data

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

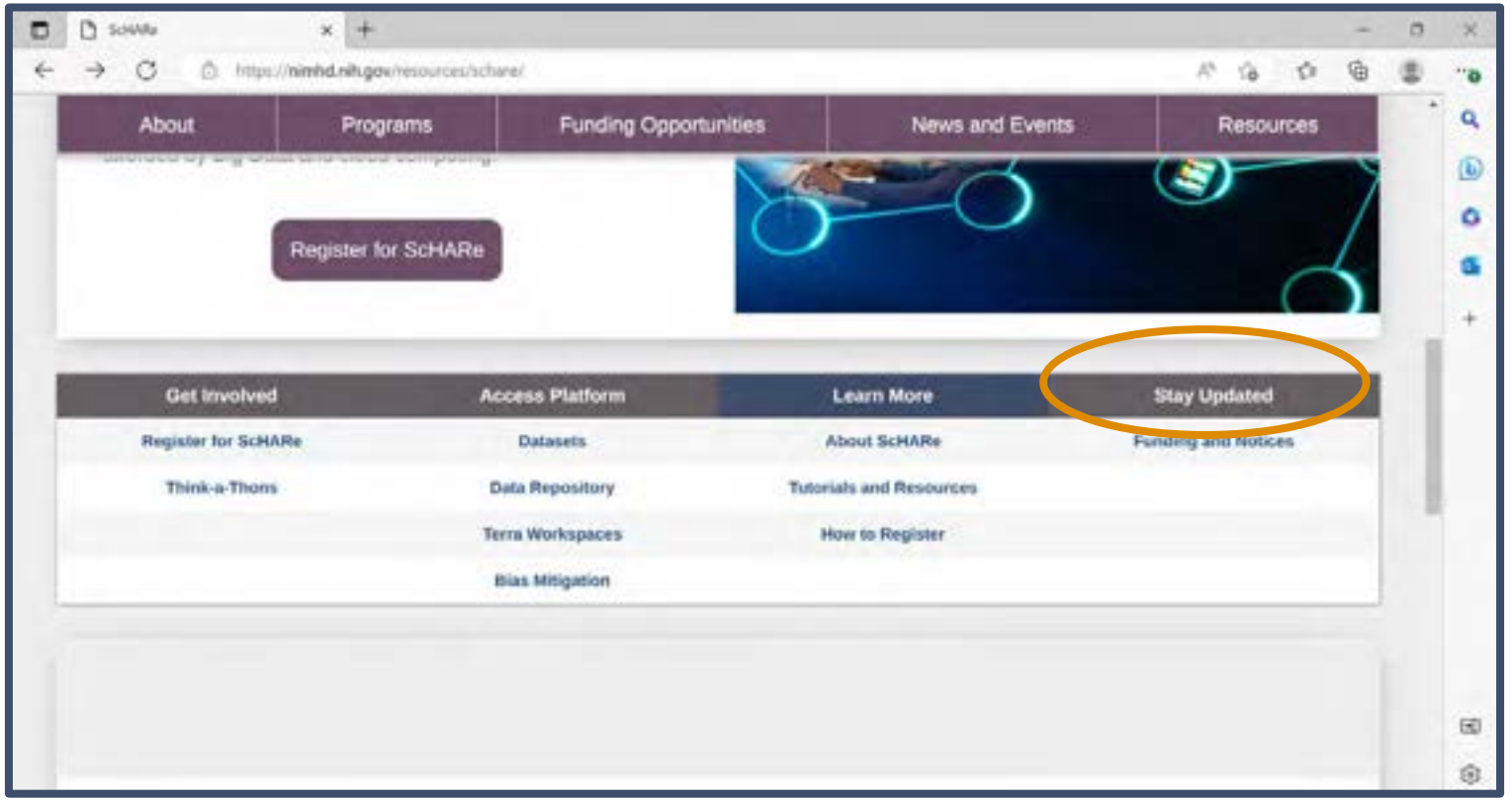
Register for ScHARe:

bit.ly/join-schare

schare@mail.nih.gov

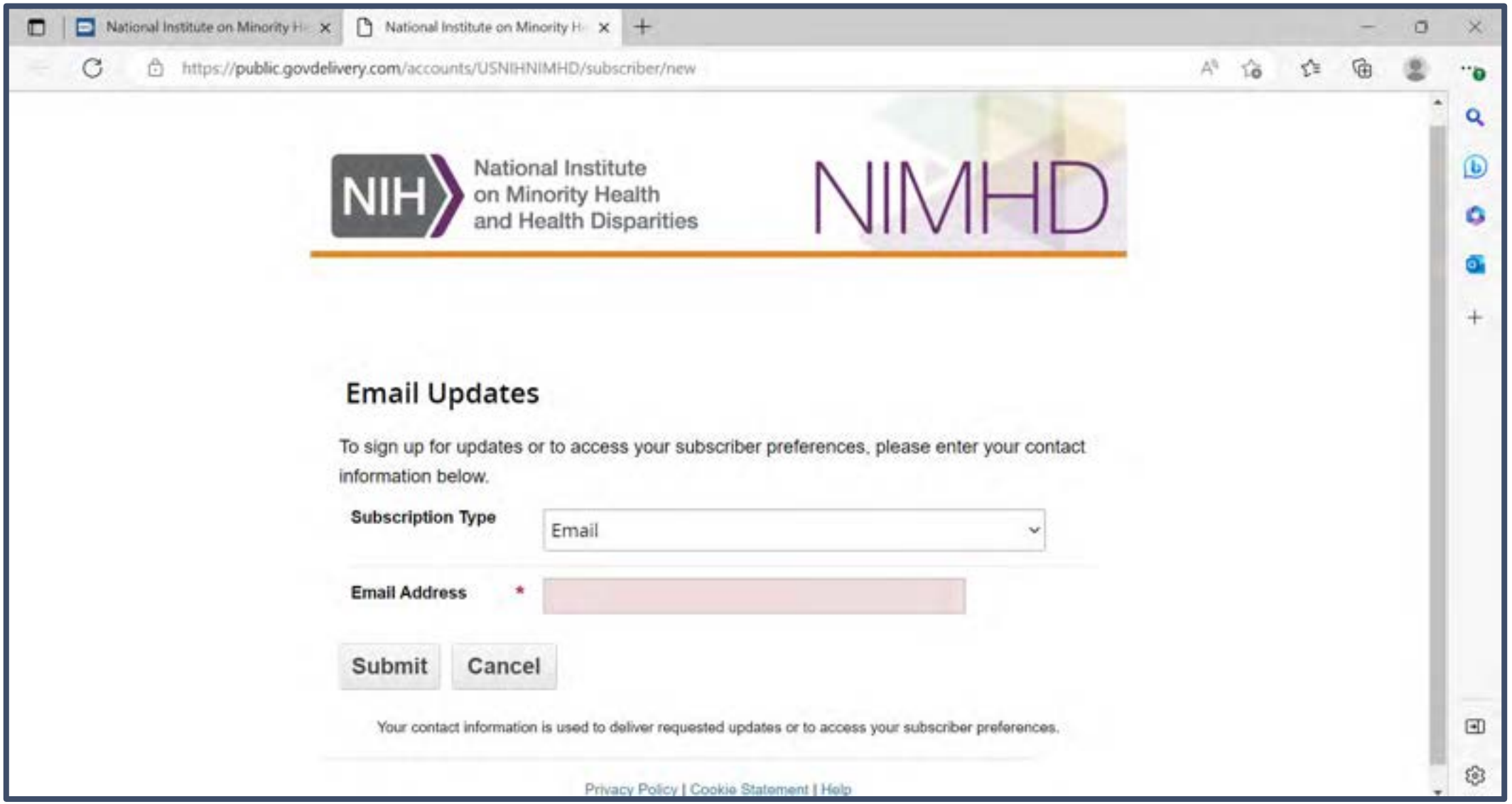
Two Ways to Sign up for ScHARe News

Scannable from your screen!

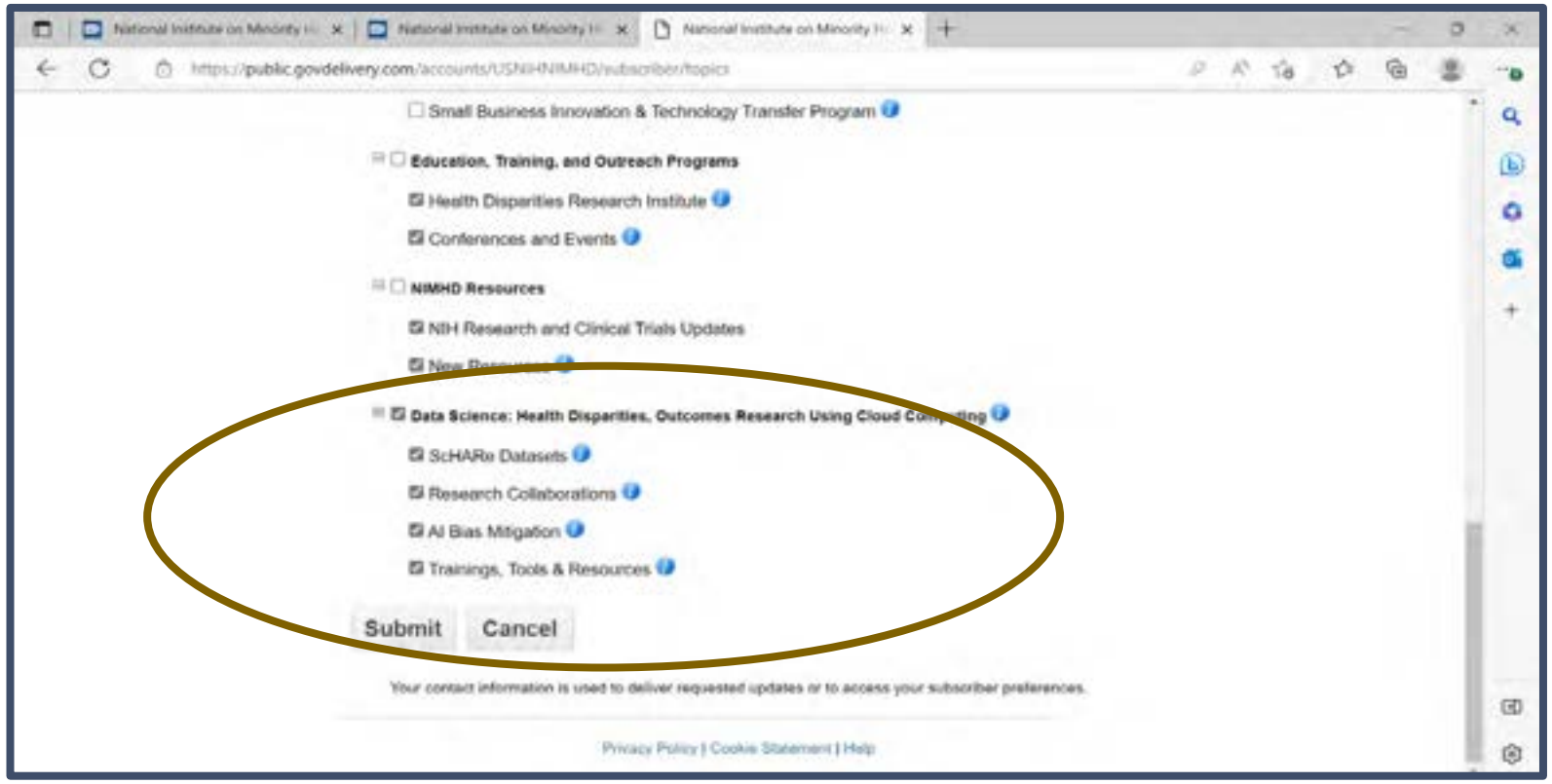


nimhd.nih.gov/schare

Step 1: Email Address



Step 2: Listserv Selection



ScHARe

Scroll down for ScHARe listserv options to sign up for news about:

- **ScHARe Datasets** – new datasets coming regularly
- **Research Collaborations** – opportunities to join multidisciplinary ScHARe-based research teams
- **AI Bias Mitigation** – bias mitigation tools and research updates
- **Trainings, Tools, & Resources** – Think-a-Thons, teaching resources, & more!

References and credits

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