ScHARe

Data Management and Analysis in Python

September 18, 2024

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

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

Outline

5’	Introduction ▪ Experience poll ▪ Interest poll
10’	What is ScHARe?
10’	Workshop setup
5’	Why Python?
10’	Recap from August session
5’	Importance of data cleaning
10’	Tools for data cleaning
10’	How data impacts visualizations
10’	Machine Learning primer
1h10’	Examples of Visualizations, Data Cleaning, Machine Learning
5’	Python tutorials and resources ▪ Evaluation poll

Experience poll

Please check your level of experience with the following:

	None	Some	Proficient	Expert
Python	☐	☐	☐	☐
R	☐	☐	☐	☐
Cloud computing	☐	☐	☐	☐
Terra	☐	☐	☐	☐
Health disparities research	☐	☐	☐	☐
Health outcomes research	☐	☐	☐	☐
Algorithmic bias mitigation	☐	☐	☐	☐

Interest poll

I am interested in (check all that apply):

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

ScHARe

What is ScHARe?

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe

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

ScHARe aims to fill four critical gaps:

- Increase participation of **women & underrepresented populations with health disparities** in data science through data science skills training, cross-discipline mentoring, and multi-career level collaborating on research
- Leverage population science, SDoH, and behavioral Big Data and cloud computing tools to foster a **paradigm shift** in health disparity and healthcare delivery outcomes research
- **Advance AI bias mitigation and ethical inquiry** by developing innovative strategies and securing diverse perspectives
- Provide a **data science cloud computing resource** for community colleges and low resource minority serving institutions and organizations

nimhd.nih.gov/schare

ScHARe

Google Platform Terra Interface

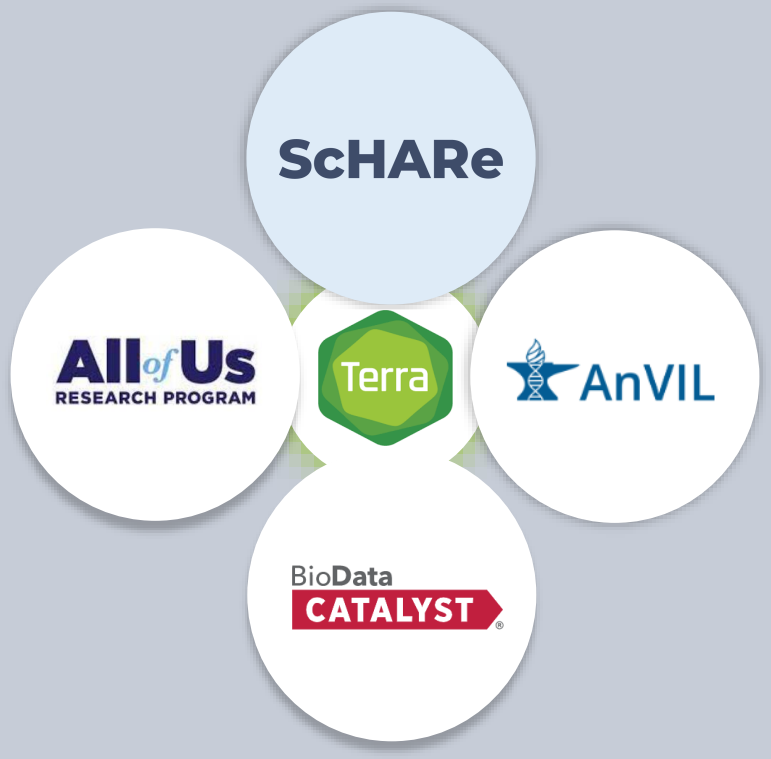
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Copy-and-paste code in Python and R
- Learning Terra on ScHARe prepares you to use other NIH platforms

Terra recommends using **Chrome**

Must have a **Gmail** friendly account

PREPARING FOR AI RESEARCH AND HEALTHCARE USING BIG DATA

Mapping across cloud platforms with Terra interface for collaborative research



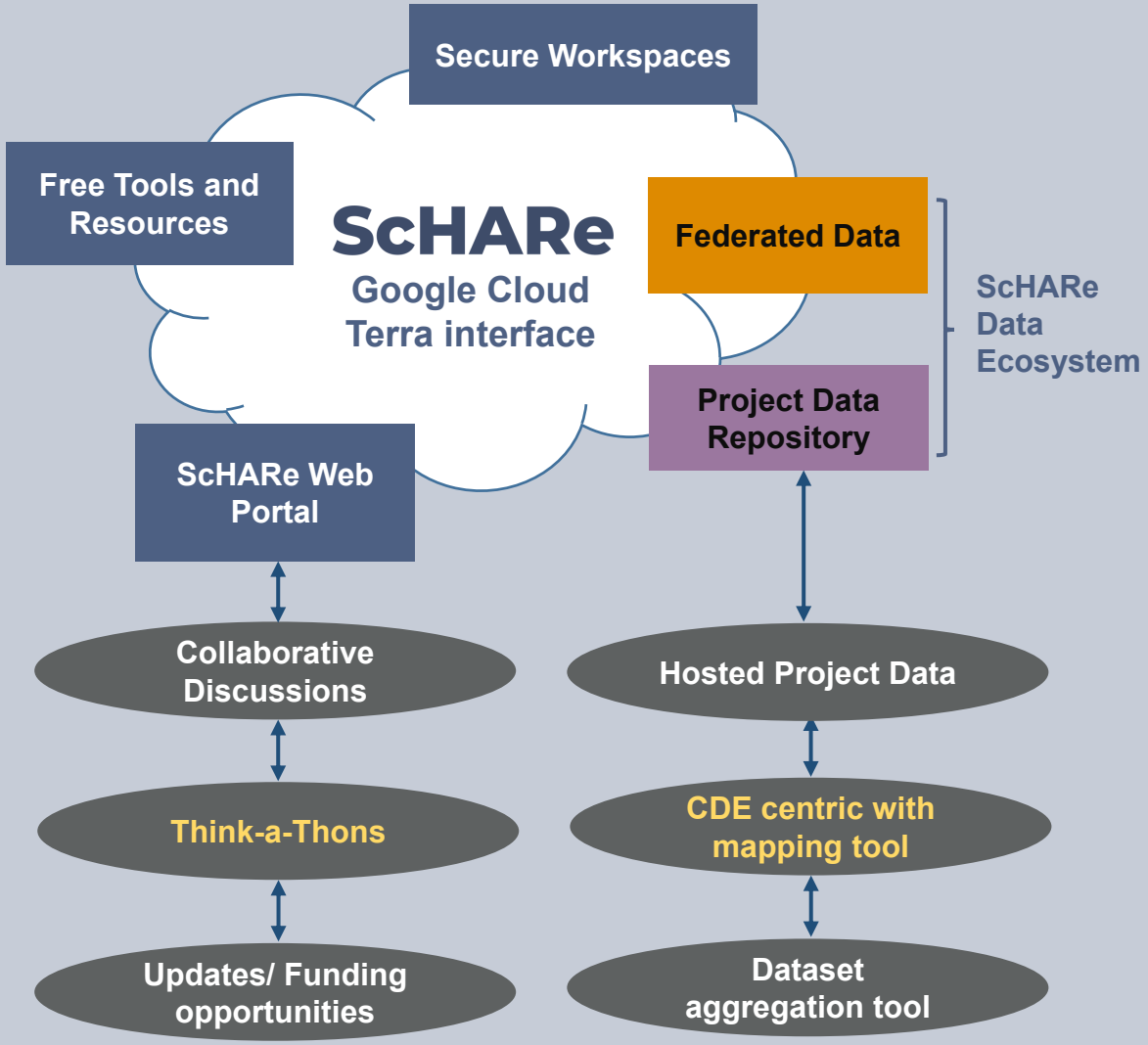
BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe Components

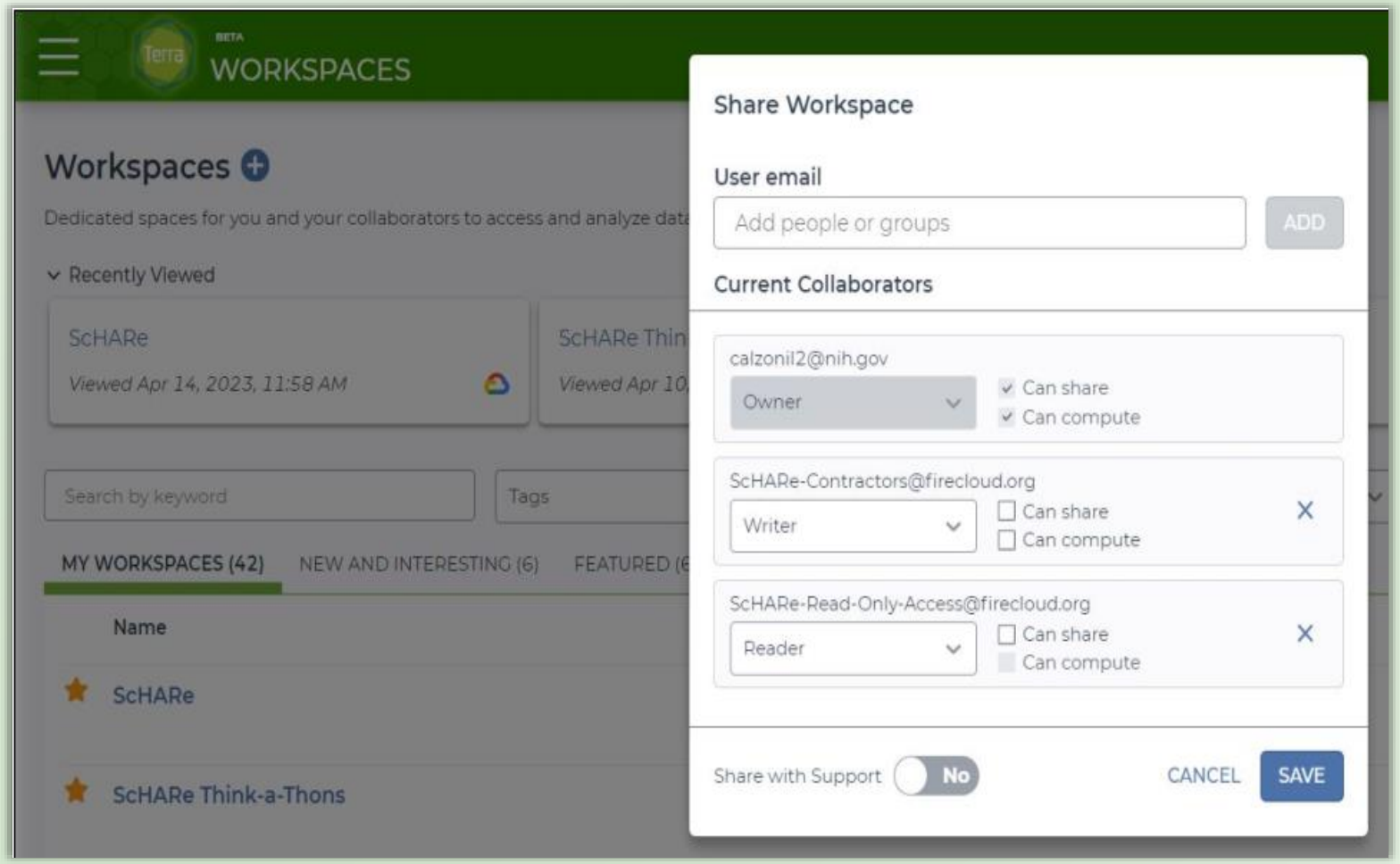
ScHARe co-localizes within the cloud:

1. **Datasets** (including social determinants of health and social science data) relevant to minority health, health disparities, and healthcare outcomes research
2. **CDE-focused data repository** to comply with the required hosting and sharing of data from NIMHD-/NINR-funded programs
3. **User-friendly computational capabilities and secure, collaborative workspaces** for students and all career level researchers
4. **Tools for collaboratively evaluating and mitigating biases** associated with datasets and algorithms utilized to inform healthcare and policy decisions (*upcoming*)

Intramural and Extramural Resource



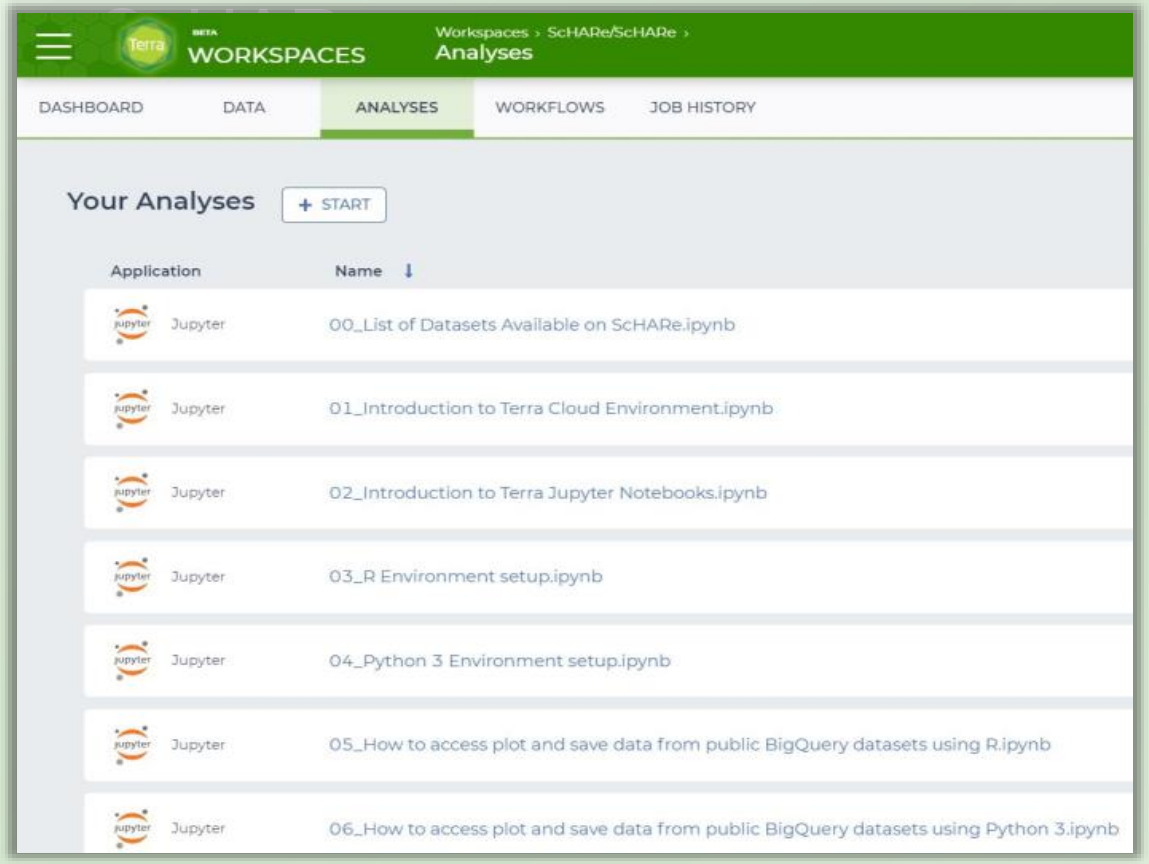
ScHARe Terra interface: secure workspaces



- Secure workspaces **for self or collaborative research**
- **Assign roles**: review or admin
- **Host own data and code**

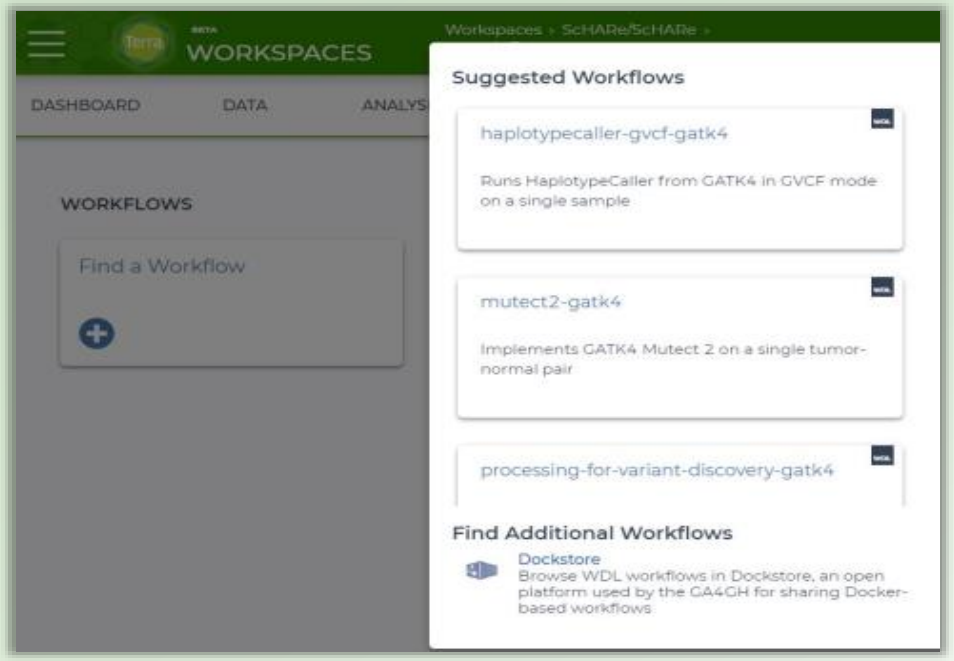
ScHARe Terra interface: analyses

Notebooks for analytics and tutorials



Modular codes

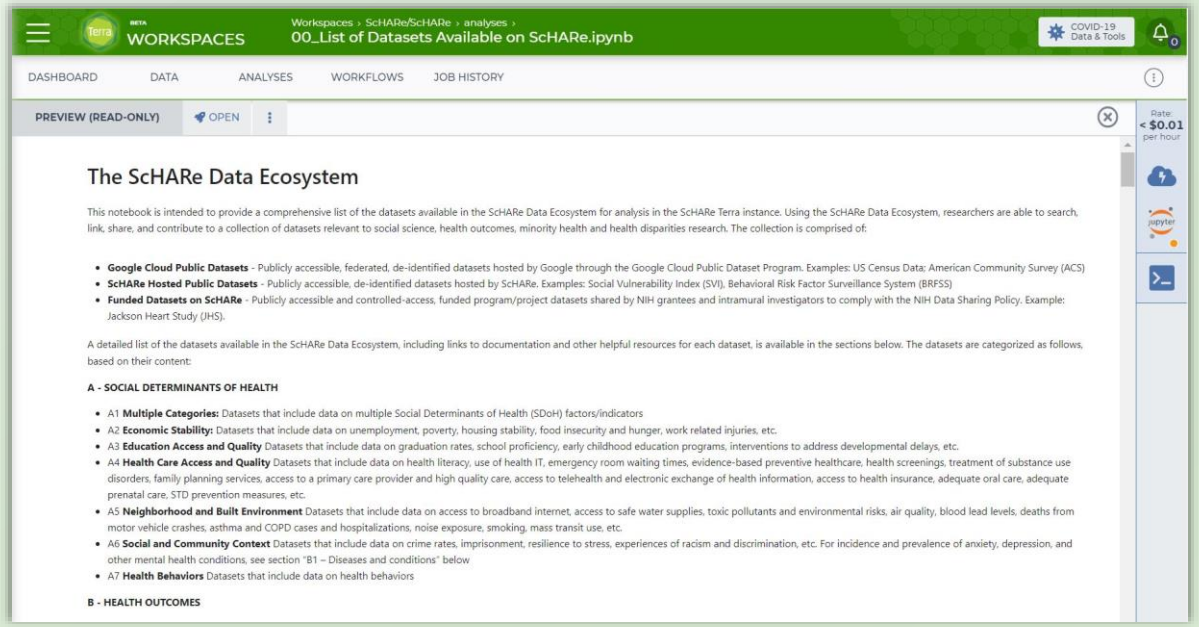
- **Easy-to-use copy-and-paste analytics**



- Modular codes developed for reuse
- **Adding SAS**

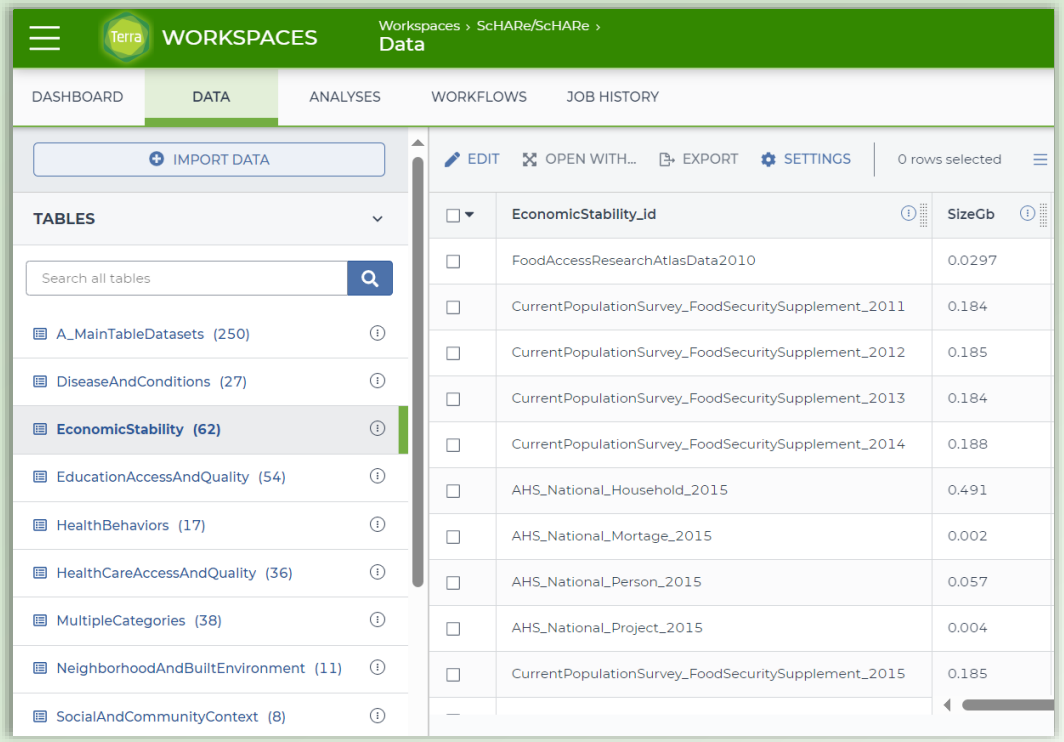
ScHARe Terra interface: access to datasets

What data?



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

Where?



In the **Data tab**, data tables help access data

ScHARe Ecosystem structure

Researchers can access, link, analyze, and export **a wealth of SDoH and population science related datasets** within and across platforms relevant to research about health disparities, health care delivery, health outcomes and bias mitigation, including:

250+
FEDERATED PUBLIC DATASETS

Public datasets

Publicly accessible, federated, de-identified datasets hosted by ScHARe or hosted by Google through the Google Cloud Public Dataset Program

ScHARe **e.g.:** *Behavioral Risk Factor Surveillance System (BRFSS)*

Google **e.g.:** *American Community Survey (ACS)*

CDE
FOCUSED REPOSITORY

Funded datasets

Publicly accessible and controlled-access, funded program/project datasets using Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

e.g.: *Jackson Heart Study (JHS)*
Extramural Grant Data
Intramural Project Data

Innovative Approach:
CDE Concept Codes
Uniform Resource Identifier (**URI**)

ScHARe Ecosystem

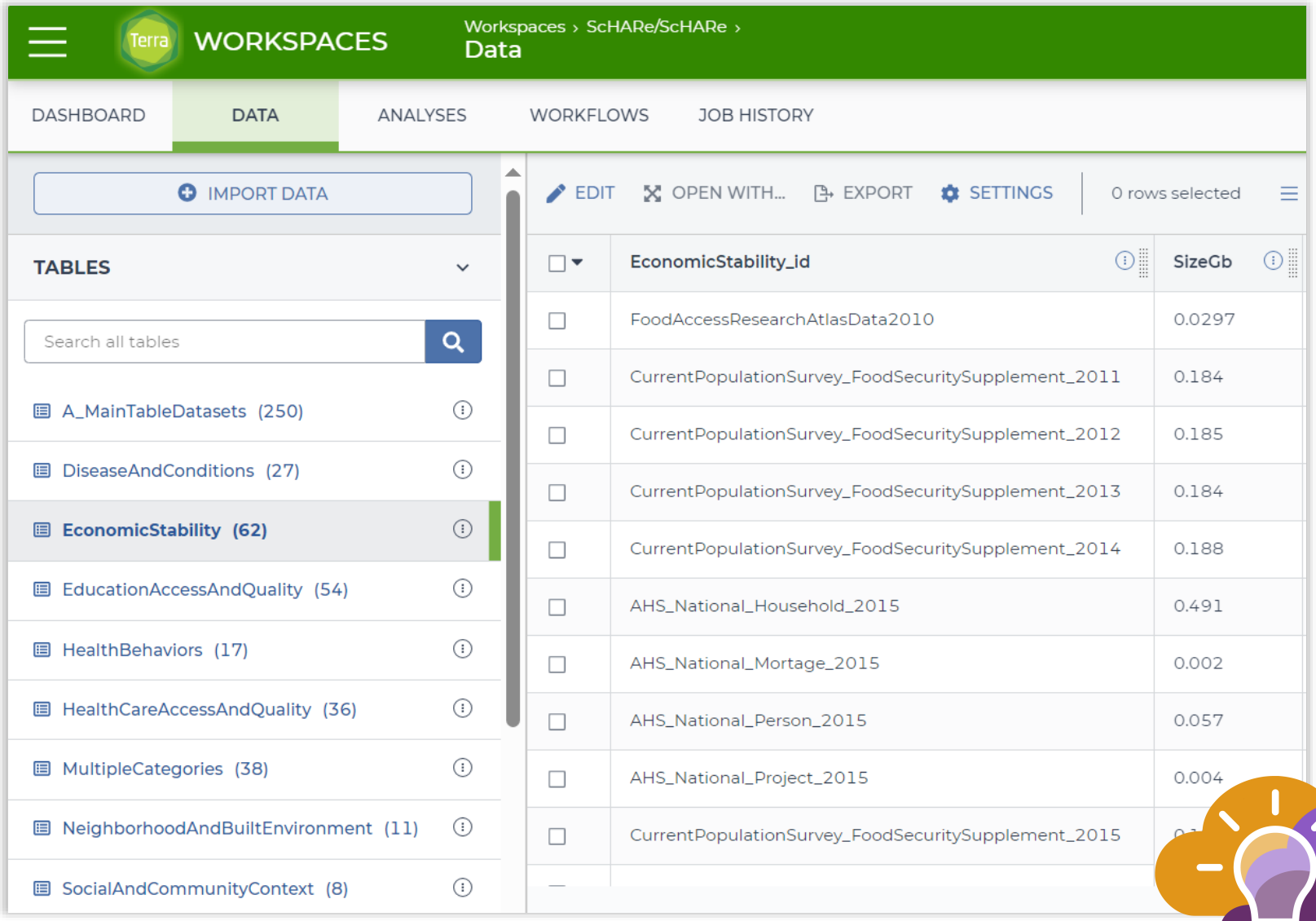
OVER 260 DATA SETS CENTRALIZED

Datasets are categorized by content based on the CDC **Social Determinants of Health categories**:

1. Economic Stability
2. Education Access and Quality
3. Health Care Access and Quality
4. Neighborhood and Built Environment
5. Social and Community Context

with the addition of:

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



ScHARe Ecosystem: ScHARe hosted datasets

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

260+ 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**

https://www.cdc.gov/about/priorities/social-determinants-of-health-at-cdc.html?CDC_AAref_Val=https://www.cdc.gov/about/sdoh/index.html

ScHARe Ecosystem: ScHARe hosted datasets

Education access and quality

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

Economic stability

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

Health care access and quality

Data on health literacy, use of health IT, preventive healthcare, access to health insurance, etc.

* Health behaviors

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

Neighborhood and built environment

Data on access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, noise exposure, smoking, mass transit use, etc.

* Diseases and conditions

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

Social and community context

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

** Not Social Determinants of Health*

ScHARe Ecosystem: Google hosted datasets

Examples of interesting datasets include:

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

How to access Google hosted datasets

Big Query

The Google public datasets are **available for access on Terra using BigQuery**

- BigQuery is the Google Cloud storage solution for structured data
- It is easy to use, works with large amounts of data and offers fast data retrieval and analysis
- Our **instructional notebooks in the Analyses tab** provide code and instructions on using Big Query to access Google datasets

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

The following Python code will read a BigQuery table into a Pandas dataframe.

From https://cloud.google.com/community/tutorials/bigquery-ibis

Ibis is a Python library for doing data analysis. It offers a Pandas-like environment for executing data analysis composable, and familiar replacement for SQL.

In [9]:
```
# Connect to the dataset
conn = ibis.bigquery.connect(dataset_id='bigquery-public-data.broadstreet_adi')
```

In [10]:
```
# Read table
ADI_table_2 = conn.table('area_deprivation_index_by_census_block_group')
ADI_table_2
```

Out[10]:
```
BigQueryTable[table]
  name: bigquery-public-data.broadstreet_adi.area_deprivation_index_by_census_block_group
  schema:
    geo_id : string
    state_fips_code : string
    county_fips_code : string
    block_group_fips_code : string
    description : string
    county_name : string
    state_name : string
    state : string
    year : int64
    area_deprivation_index_percent : float64
```

ScHARe

The ScHARe Data Ecosystem

This document is intended to provide a comprehensive list of the datasets available in the ScHARe Data Ecosystem for analysis in the ScHARe Terra instance. Using the ScHARe Data Ecosystem, researchers are able to search, link, share, and contribute to a collection of datasets relevant to social science, health outcomes, minority health and health disparities research.

The collection is comprised of:

- **Google-hosted Public Datasets** - Publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program. Examples: US Census Data; American Community Survey (ACS)
- **ScHARe-hosted Public Datasets** - Publicly accessible, de-identified datasets hosted by ScHARe. Examples: Social Vulnerability Index (SVI), Behavioral Risk Factor Surveillance System (BRFSS)
- **ScHARe-hosted Project Datasets** - Publicly accessible and controlled-access, funded program/project datasets shared by NIH grantees and intramural investigators to comply with the

Jackson Heart Study (JHS)

ScHARe
Datasets PDF list

Scan me

bit.ly/ScHARe-datasets

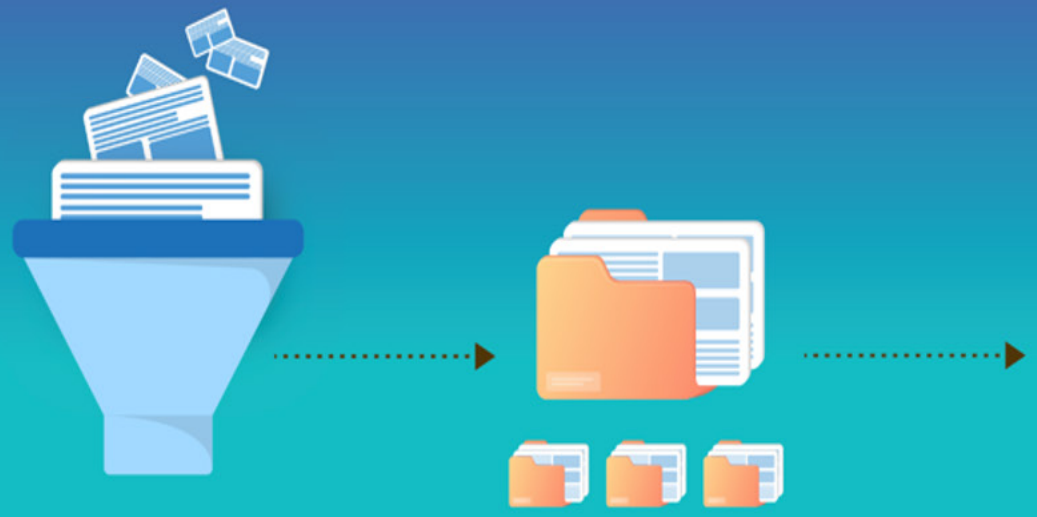
CDE benefits:

- Faster start-up for project
- Better data aggregation across projects
- Shared meaning
- Concept-focused to allow questions/answers variations
- Coding enables an URI approach for better data interoperability

A **Common Data Element** (CDE) is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection

Because Researchers use CDEs...

they can more quickly share data and get results faster, which ultimately can help make a **meaningful difference to our nation's health.**



For more information about how CDEs accelerate research discoveries, visit: cde.nlm.nih.gov/resources

ScHARe Core CDEs

PhenX Toolkit

NIH Endorsed

- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex
- Gender
- Sexual Orientation
- Marital Status
- Education
- Annual Household Income
- Household Size

- English Proficiency
- Disabilities
- Health Insurance
- Employment Status
- Usual Place of Health Care
- Financial Security / Social Needs
- Self-Reported Health
- Health Conditions (and Associated Medications/Treatments)

- **NIMHD Framework***
- **Health Disparity Outcomes***

* Project Level CDEs

ScHARe has developed **Common Data Elements** to ensure consistent data collection across studies, facilitate interoperability, and link data from different sources

NIH CDE Repository:
cde.nlm.nih.gov/home
PhenX Toolkit:
www.nimhd.nih.gov/resources/phenx/

ScHARe REPOSITORY

COMMON DATA ELEMENTS

NLM CDE Repository

Coded NIMHD Common Data Elements

- Labels
- Questions
- Permissible Values

DATA UPLOAD

Acquired **Google and ScHARe Hosted Datasets**

- Overview
- Data Dictionaries
- Data Updates

ATO

Common Data Elements + **Data**

Data Access

Based On PII Levels and User Needs:

- Public
- Data Use Agreement
- Private

Project and Key Acquired Datasets

Overview

Description and Links to Overview Material

4-Privacy Levels

COMMON DATA ELEMENTS

Data

Metadata

Data Dictionaries

Analysis Ready

RAS Single Sign-on

DATA MAPPING, DOWNLOAD AND EXPORT

Other Cloud Platforms
AnVil, BDC, All of Us

DATA MAPPING

ACROSS DATASETS AND PLATFORMS BASED ON CDES

EXAMPLE: CDE linked

ACS **NIMHD Project** **BioData Catalyst**

Aggregated Data Set

CDE Linked Project Data

Data Download in a Variety of Formats

CSV, TSV, XLSX

Data Export to Terra for Analysis

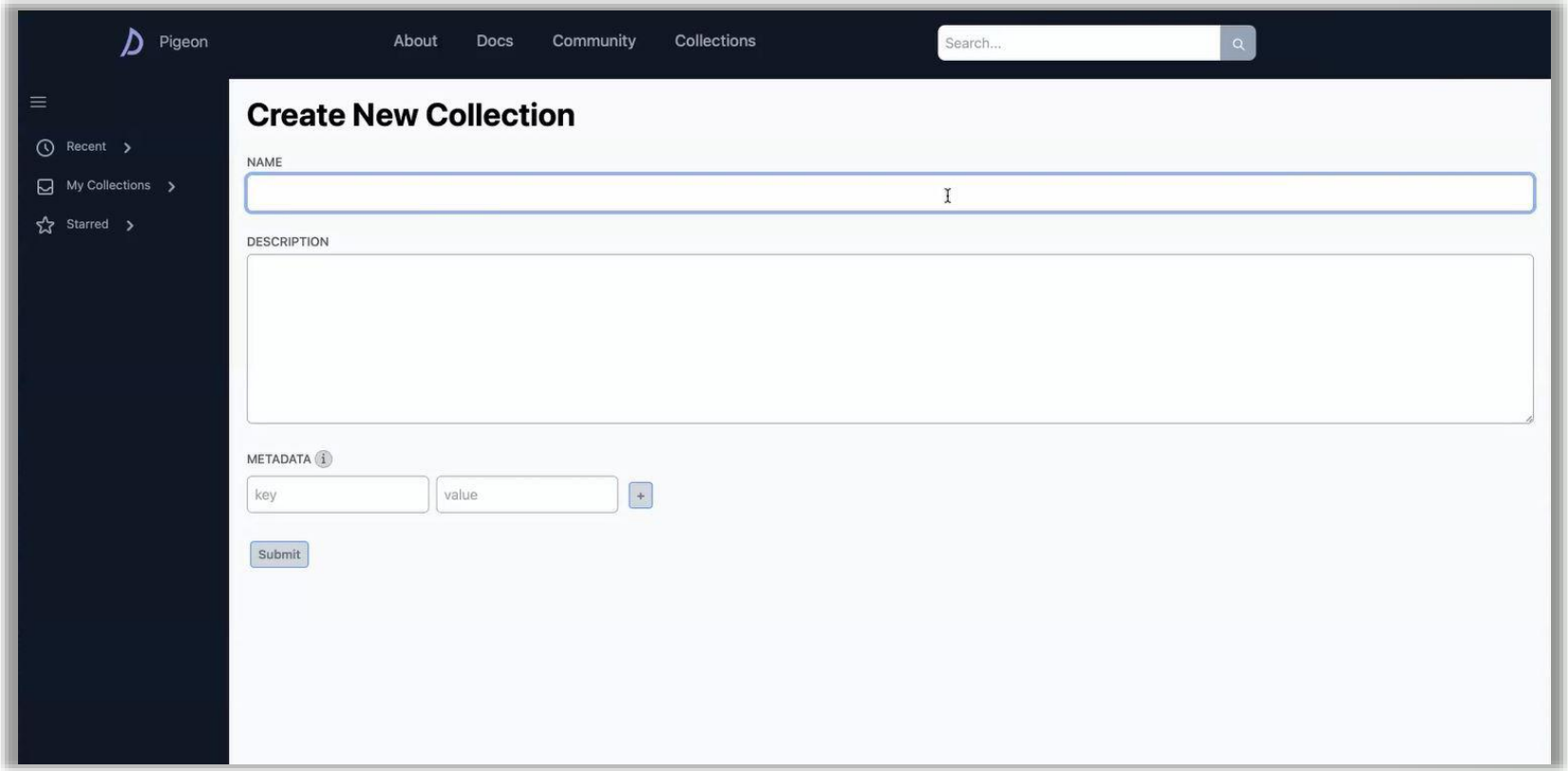
Workspaces

Visualizations Tools

Shiny

ScHARe Repository

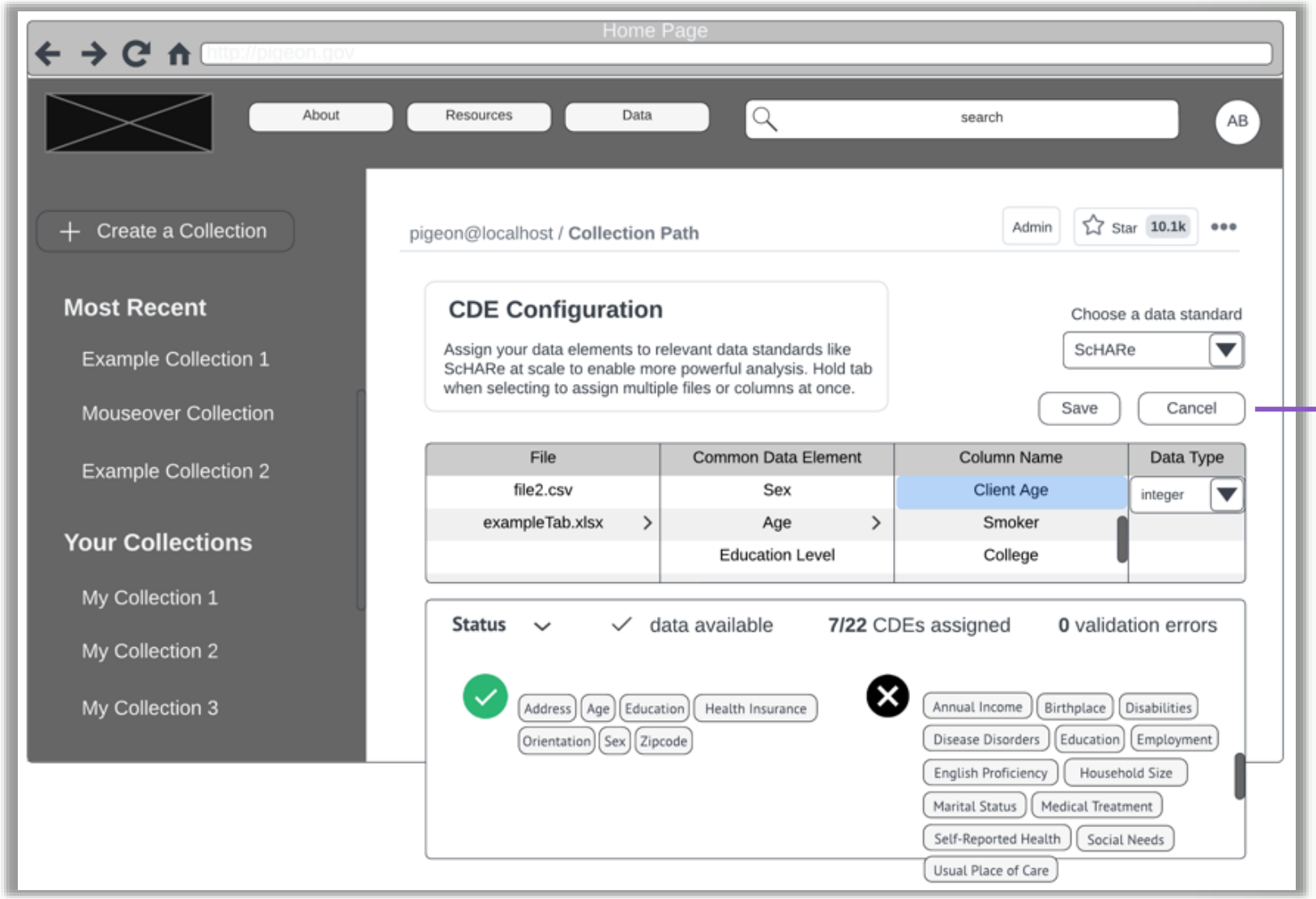
PUBLICLY AVAILABLE FALL 2024



- Host your project data in a **safe space** with privacy levels, secure workspaces, collaboration platform
- **CDE centric**
- **Focus**: Social Science, SDoH, Health Disparities, Health Outcomes Research
- Comply with **NIH Data Management and Data Sharing Policy**
- **Link your data** with others and federated data

ScHARe Repository

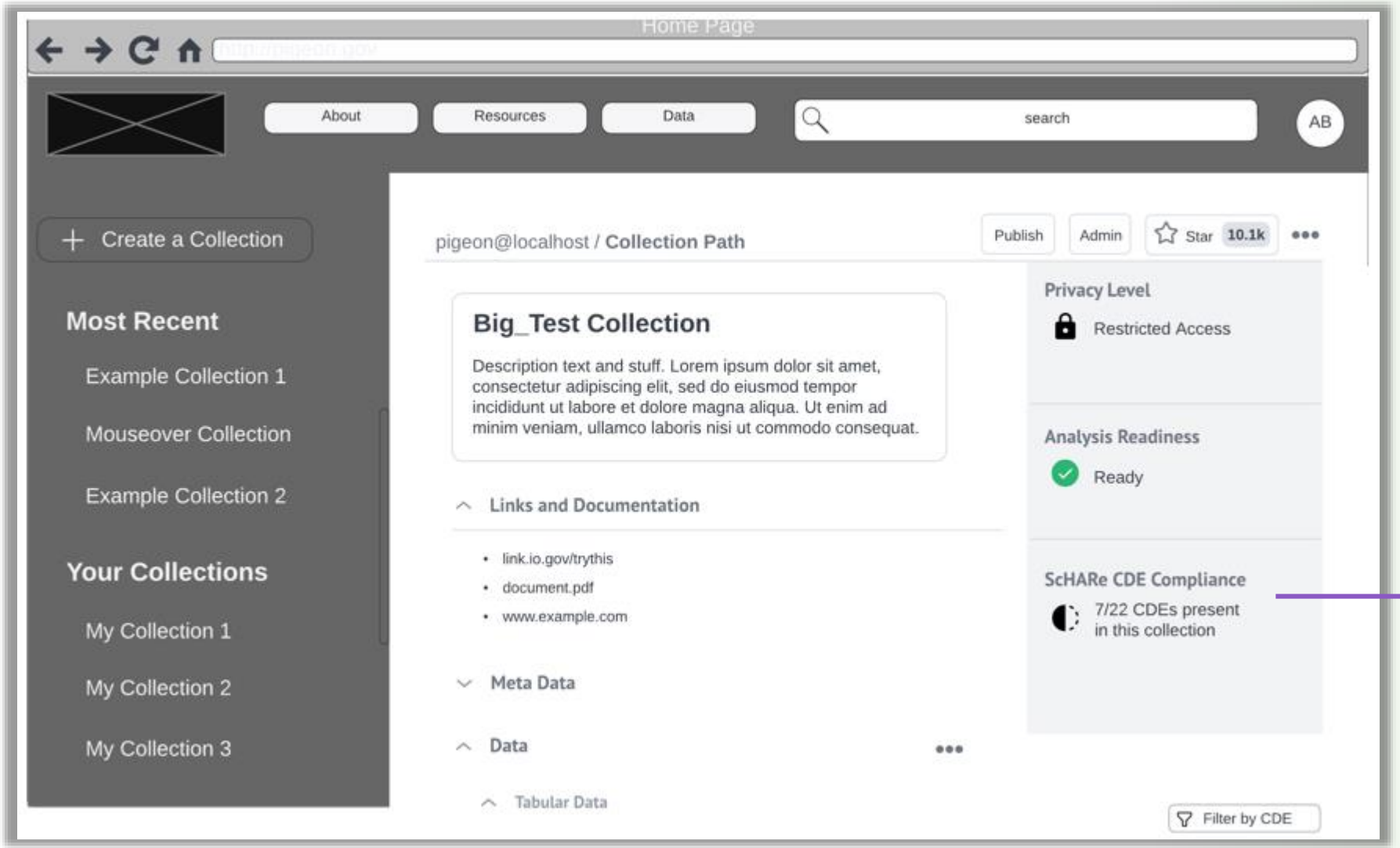
PUBLICLY AVAILABLE FALL 2024



Map project CDEs or variables to ScHARe-PhenX CDEs

ScHARe Repository

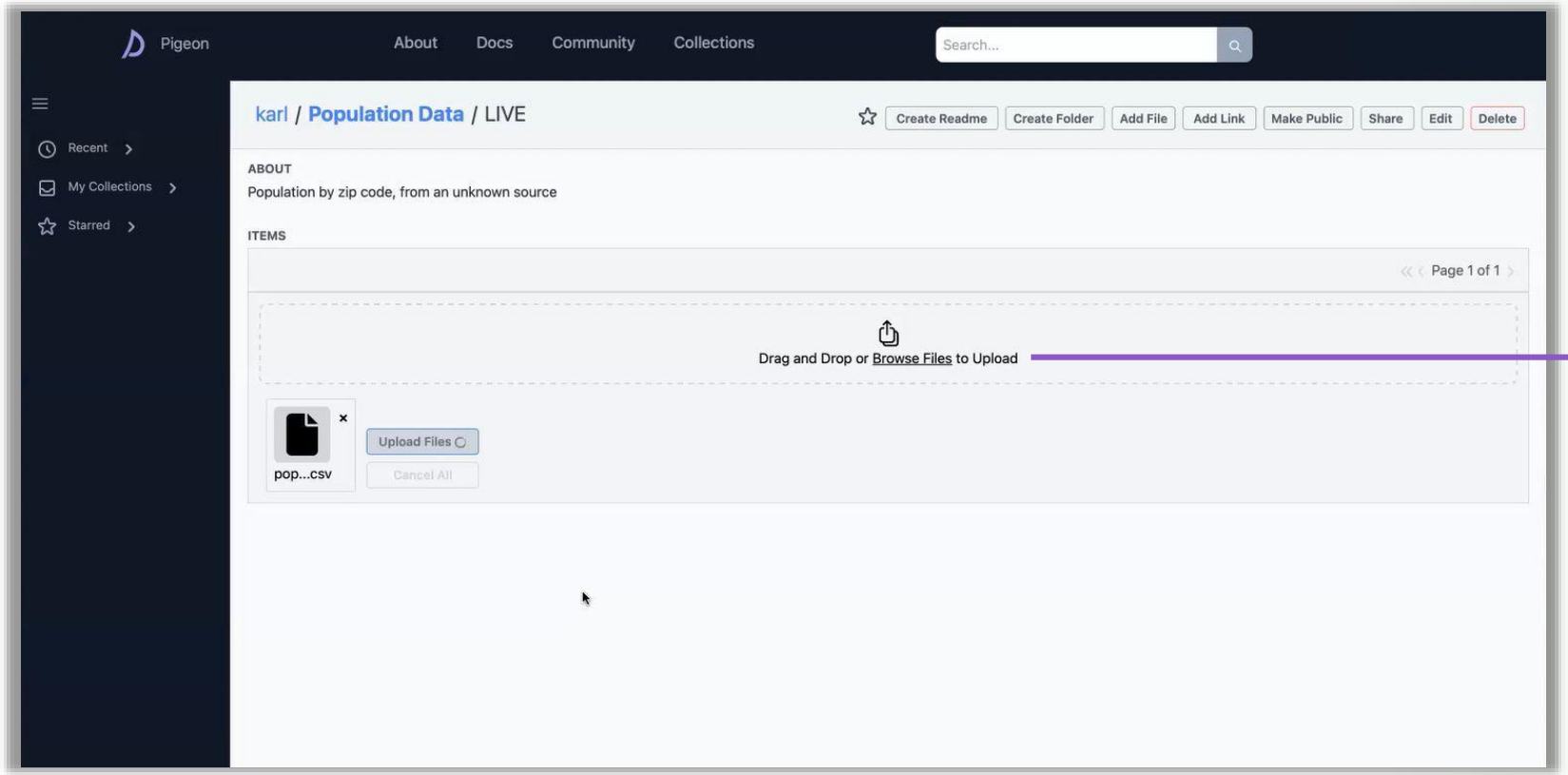
PUBLICLY AVAILABLE FALL 2024



Shows number of project CDEs that match or can map to ScHARe-PhenX CDEs

ScHARe Repository

PUBLICLY AVAILABLE FALL 2024



Aggregate datasets with drag-and-drop features

ScHARe Repository

PUBLICLY AVAILABLE FALL 2024

Pigeon

About Docs Community Collections

Search...

Recent

My Collections

Starred

karl / Population Data / LIVE / population_by_zip_2010.csv

Save Cancel

Parser Type

csv

Options

Select one to add...

OPTION	SETTING

Columns

minimum_age	Integer	Add →
maximum_age	Integer	Add →
gender	String	Gender fMCdgD9I:0001
zipcode	String	nihcde:7kijL9i3sx
geo_id	String	Add →

Results ✓ Data available ✓ 0 parsing errors ✗ 5 validation errors

Table Preview

population	minimum_age	maximum_age	gender	zipcode	geo_id
50	30	34	female	61747	8600000US61747
5	85		male	64120	8600000US64120
1389	30	34	male	95117	8600000US95117
231	60	61	female	74074	8600000US74074
56	0	4	female	58042	8600000US58042

View aggregated dataset

ScHARe

Research Think-a-Thons

- Novice **training webinars** for data science, cloud computing and research using Big Data
- **Target:** underrepresented populations, women, racial/ethnic and sexual gender minorities, rural and poor populations

Generational career & discipline exchange

Think-a-Thons

3rd Wednesday of every month 2 pm

Goals:

- Upskill underrepresented populations in data science and cloud computing
- Foster a research paradigm shift to use Big Data in health disparities/health outcomes research
- Promote use of Dark Data

1. TUTORIAL AND TARGETED THINK-A-THONS

Launched April 2024

- Monthly sessions (2 1/2 hours)
- Instructional/interactive
- Designed for new/experienced users
- Networking
- Mentoring and coaching
- Topics include:
  - Data Science 101
  - Terra
  - Social Determinants of Health analytics
  - Common Data Elements
  - AI readiness
  - Ethical and transparent AI
  - Bias mitigation

2. RESEARCH THINK-A-THONS

- Multi-career (students to senior investigators)
- Multi-discipline (data scientists and researchers)
- Featured datasets with guest experts leads
- Guest experts in topic areas, analytics, data sources etc. to provide guidance
- Generate research idea - decide design, datasets and analytics
- Learn Ethical AI
- Publications

Register:

bit.ly/think-a-thons

Think-a-Thon tutorials

bit.ly/think-a-thons

February	Artificial Intelligence and Cloud Computing 101
March	ScHARe 1 – Accounts and Workspaces
April	ScHARe 2 – Terra Datasets
May	ScHARe 3 – Terra Google-hosted Datasets
June	ScHARe 4 – Terra ScHARe-hosted Datasets
July	An Introduction to Python for Data Science – Part 1
August	An Introduction to Python for Data Science – Part 2
September	ScHARe 5: A Review of the ScHARe Platform and Data Ecosystem
October	Preparing for AI 1: Common Data Elements and Data Aggregation
November	Preparing for AI 2: An Introduction to FAIR Data and AI-ready Datasets
January	Preparing for AI 3: Computational Data Science Strategies 101
February/March	Preparing for AI 4: Overview Prep for AI Summary with Transparency, Privacy, Ethics
April	Research Teams – SDoH and Health Disparities
May	Be a Part of the Future of Knowledge Generation 1: AI/Cloud Computing Basics and CDEs
July	Be a Part of the Future of Knowledge Generation 2: AI-Ready Datasets and Computations

SPECIAL EVENTS

- ScHARe for **Educators** (Community Colleges and low-resource MSIs)
- ScHARe for **American Indian/ Alaska Native Researchers**
- ScHARe for **Coders and Programmers** to conduct research

Experience conducting ethical AI

Transparency

Public perception and understanding of how AI works

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

Fairness

Findable: *providing metadata, documentation, and clear identifiers*
Accessible: *wide audience*
Interoperable: *standardized formats and APIs enable seamless integration*
Reusable: *clear documentation, licensing, reduce redundancy*

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

ScHARe

Training pipeline

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Think-a-Thons training/mentoring pipeline

NLM
OIC Experts
Fellows

+

Think-a-Thons

- ✓ Instructional
- ✓ Research

+

AIM-AHEAD

+

AnVil | **HEAL**
N3C | **All of US**
BioData Catalyst

Using AI experts

to train and mentor novice AI users

to upskill and mentor diverse perspectives in AI

to increase diverse perspectives in biomedical research

Goal: "Upskilling"

- ✓ Data science specialists into health disparities and health outcomes research
- ✓ Health disparities/outcomes researchers into using big data and cloud computing

Target Audience:

- ✓ Underrepresented populations (women, race/ethnic) users not trained in data science
- ✓ Data scientists with no or little research experience
- ✓ Resource and tool for Community Colleges and low-resource MSIs and organizations

ScHARe

Data Management and Analysis in Python

September 18, 2024

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

ScHARe

Workshop setup

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

We have registered you for ScHARe

To opt out,

email us at

schare@mail.nih.gov

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

Let's make sure that everyone:

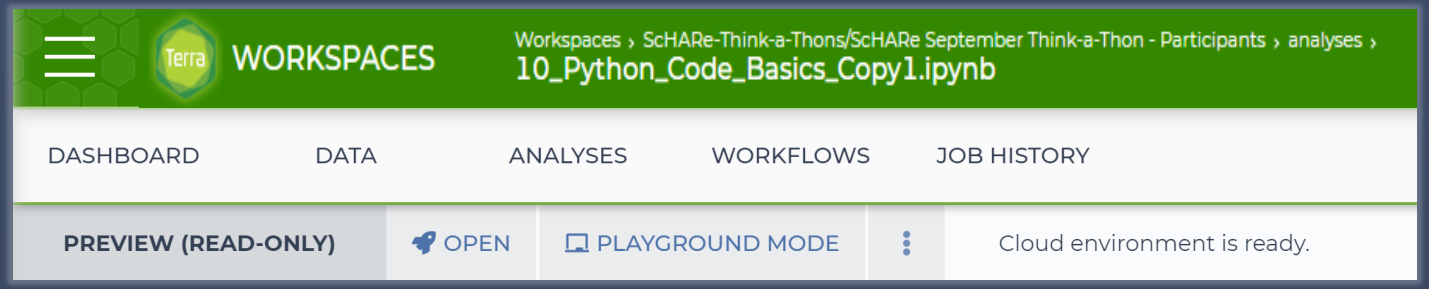
✓ 1. has provided their Gmail address and has been registered for ScHARe

2. has created a Terra account

3. can access the tutorial we will be using today at: bit.ly/schare-python-notebooks

4. has configured their cloud environment

5. can run the tutorial in playground mode:

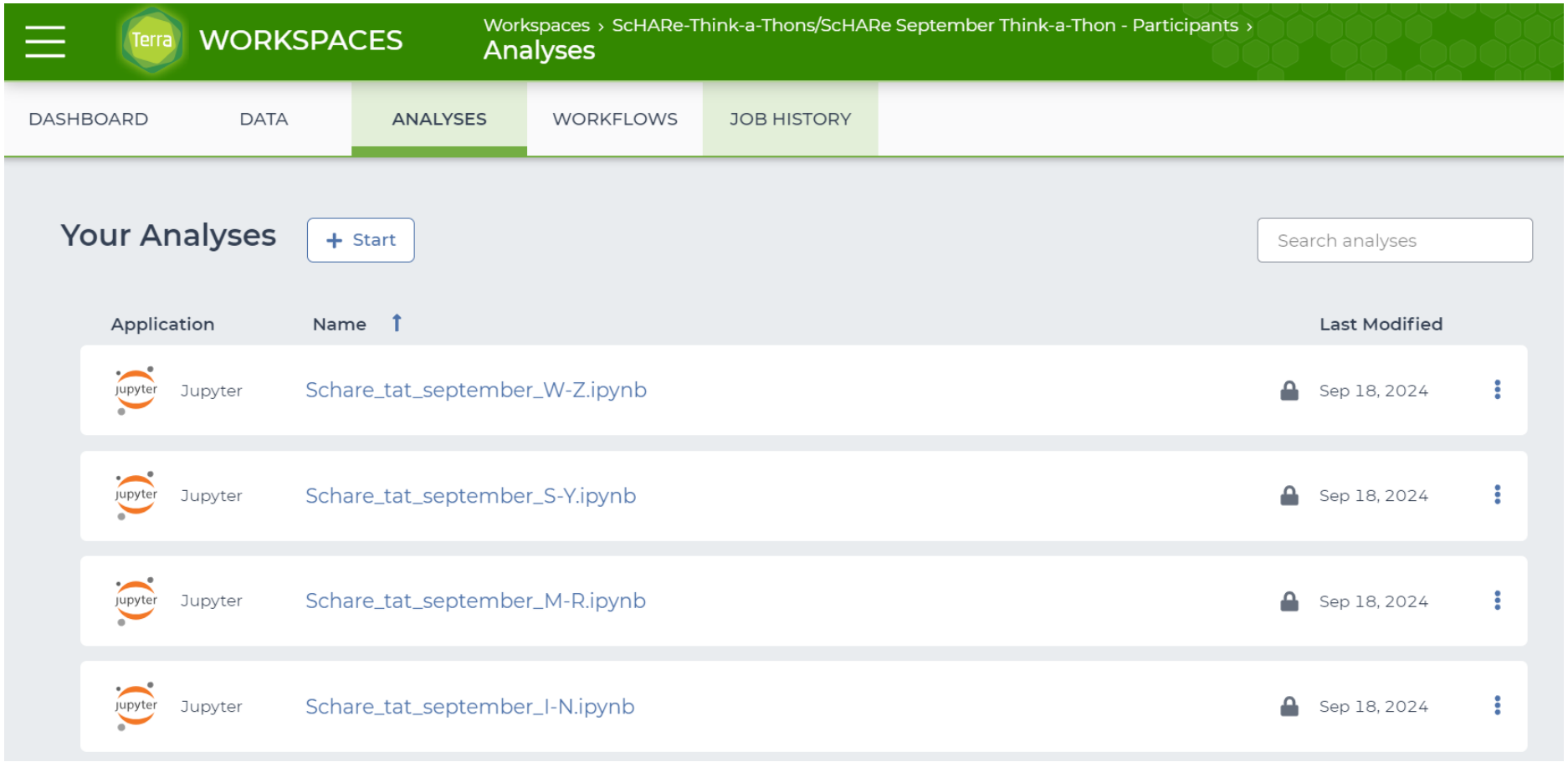


Please paste the address below in your browser:

bit.ly/schare-python-notebooks-2

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

bit.ly/schare-python-notebooks-2



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

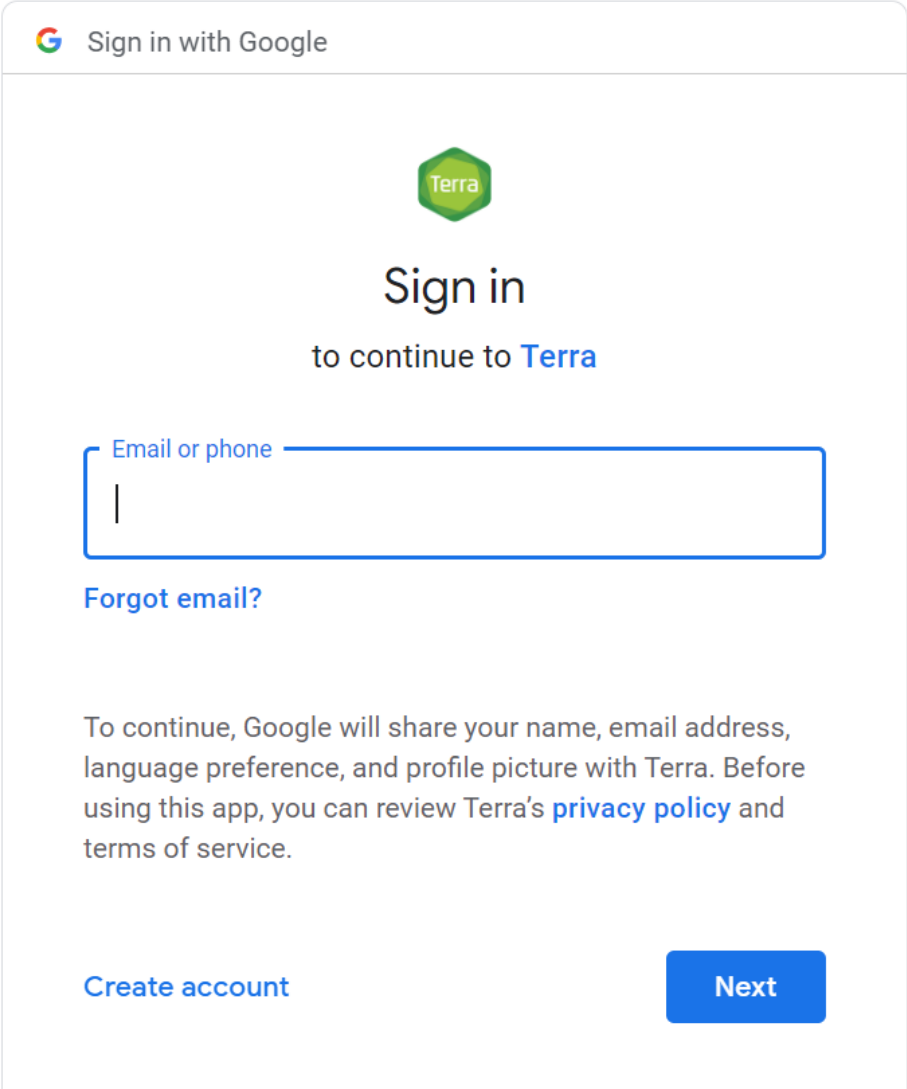
bit.ly/schare-python-notebooks-2

Click on the login button:

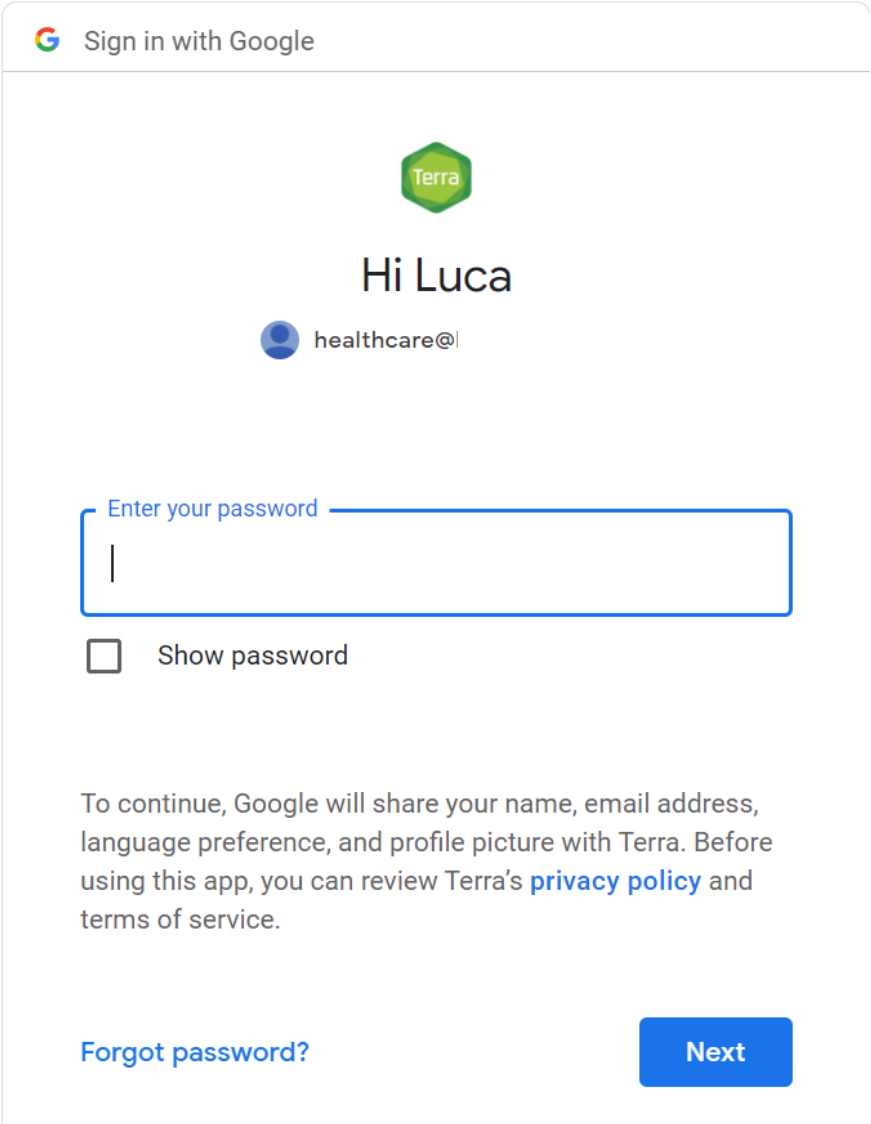
bit.ly/schare-python-notebooks-2

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

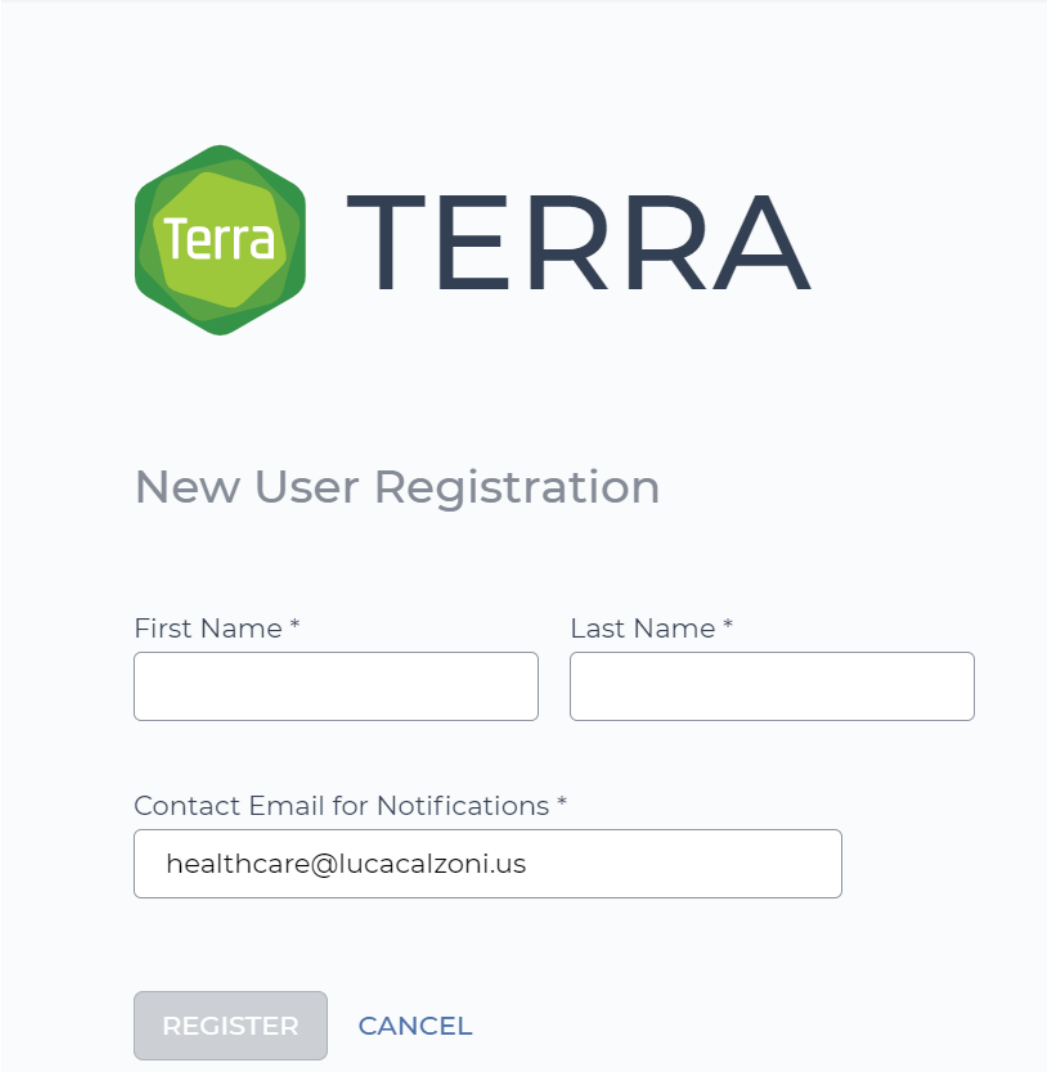
Use the Gmail address you provided us with to log in:



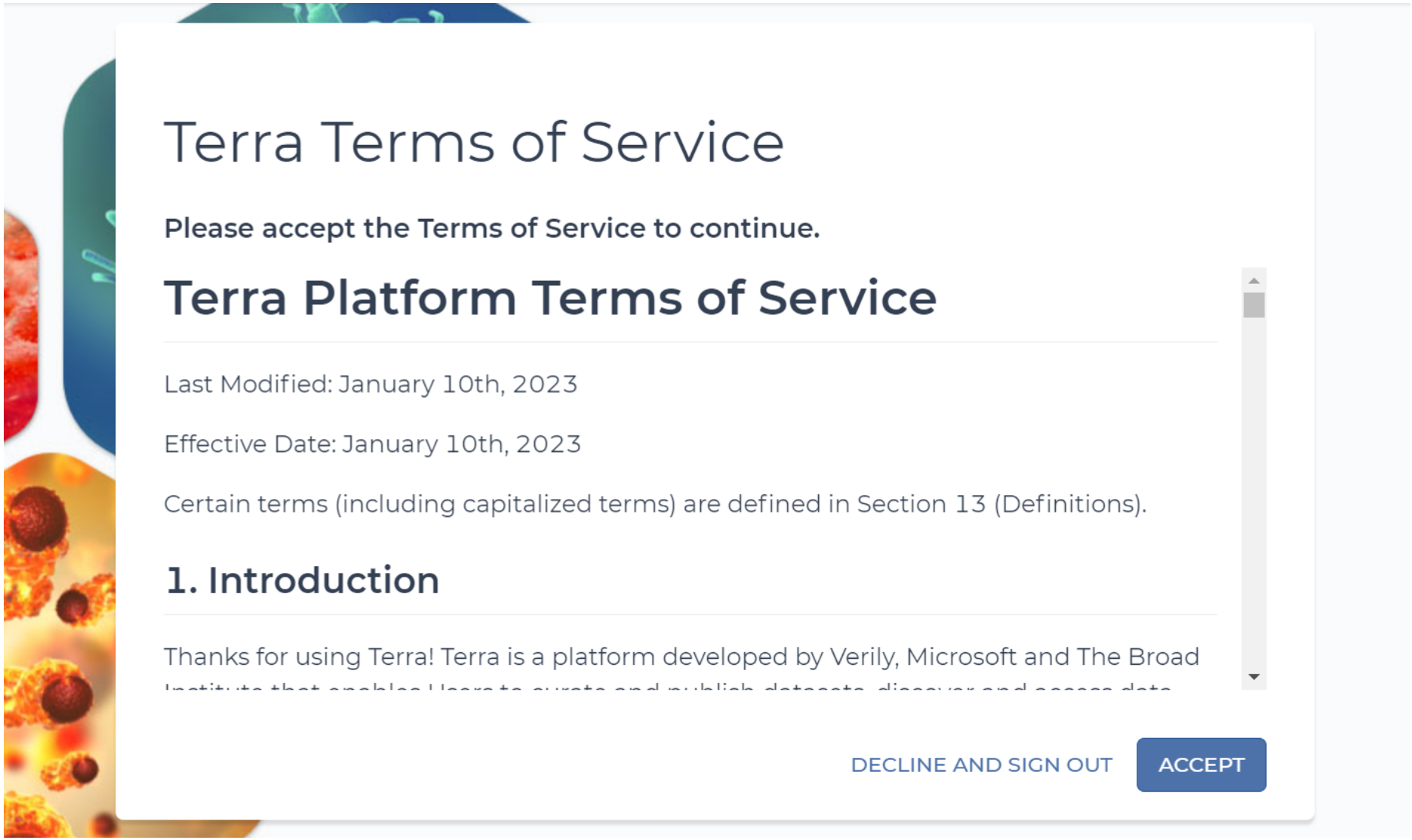
Input the password associated with your Gmail account:



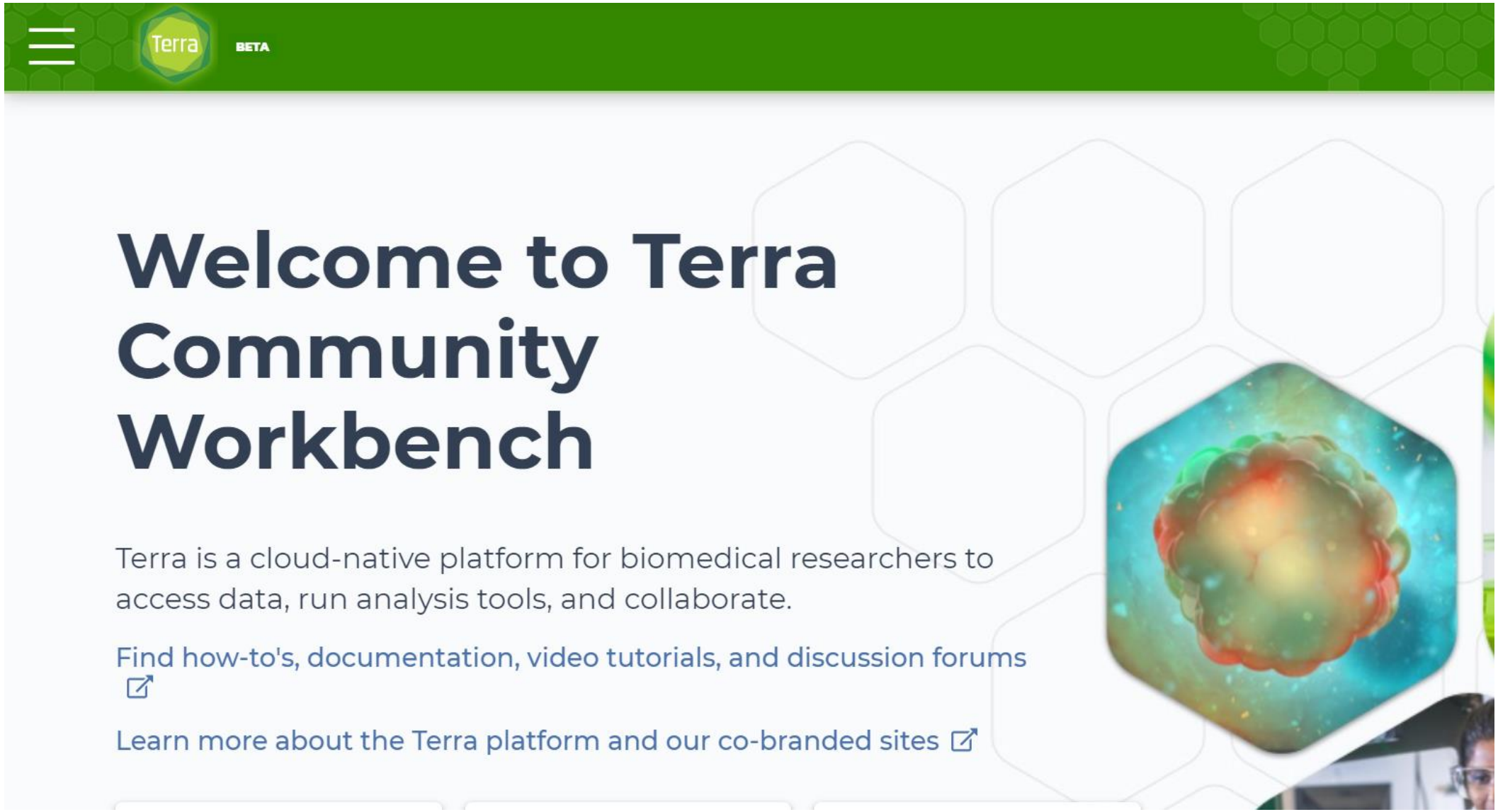
If you are new to Terra, create an account now:



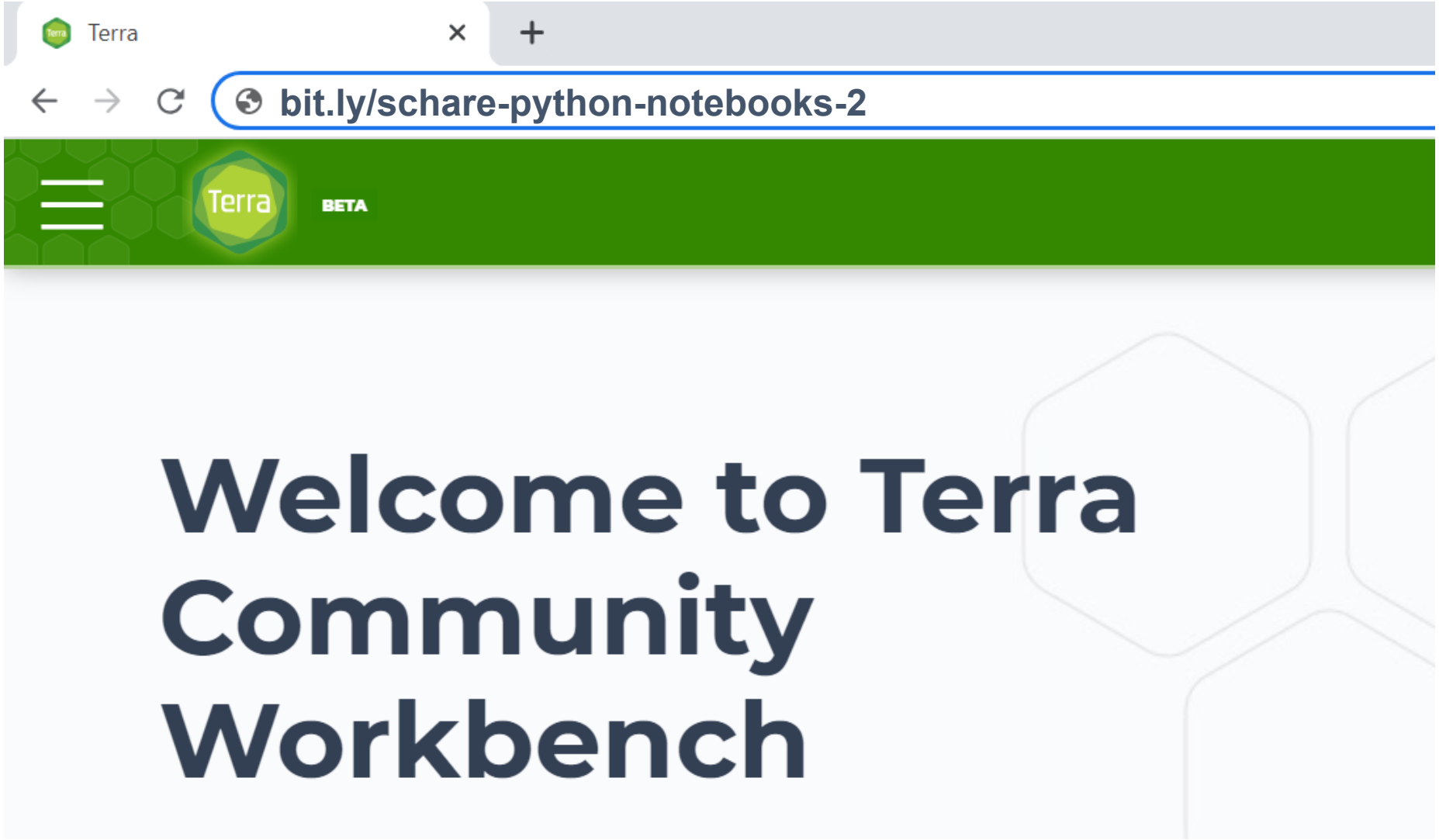
Accept the Terra Terms of Service:



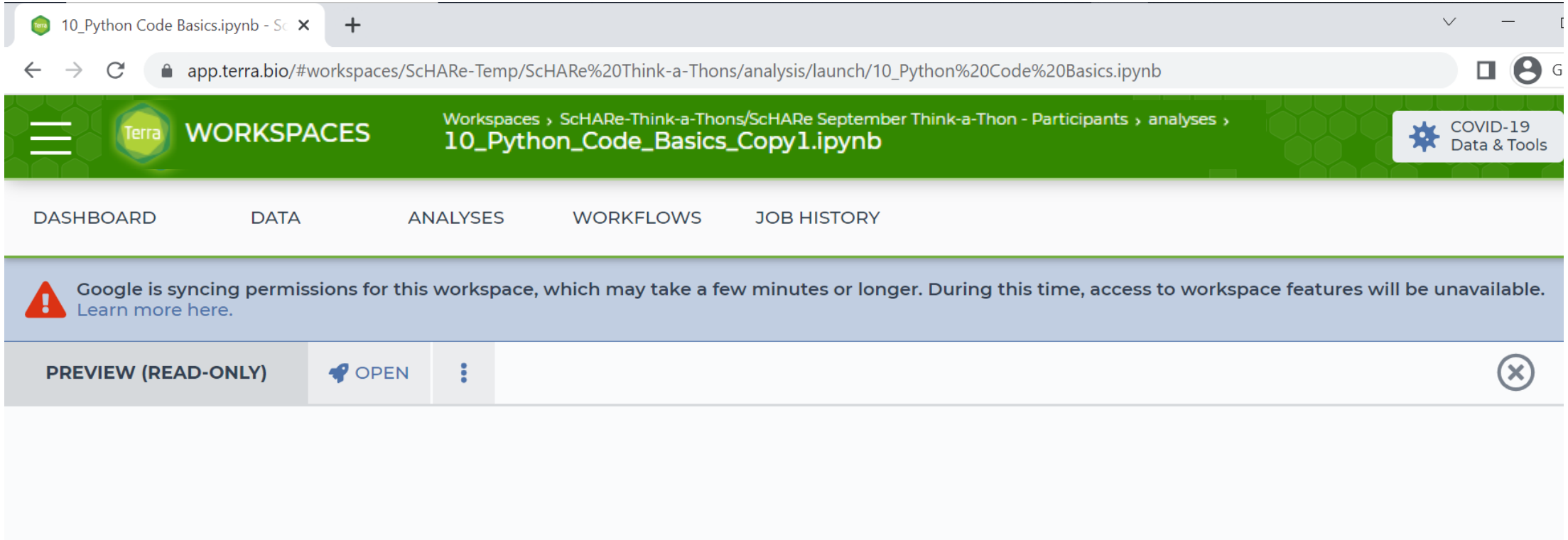
You will see this welcome page:



Paste this address in your browser: bit.ly/schare-python-notebooks-2

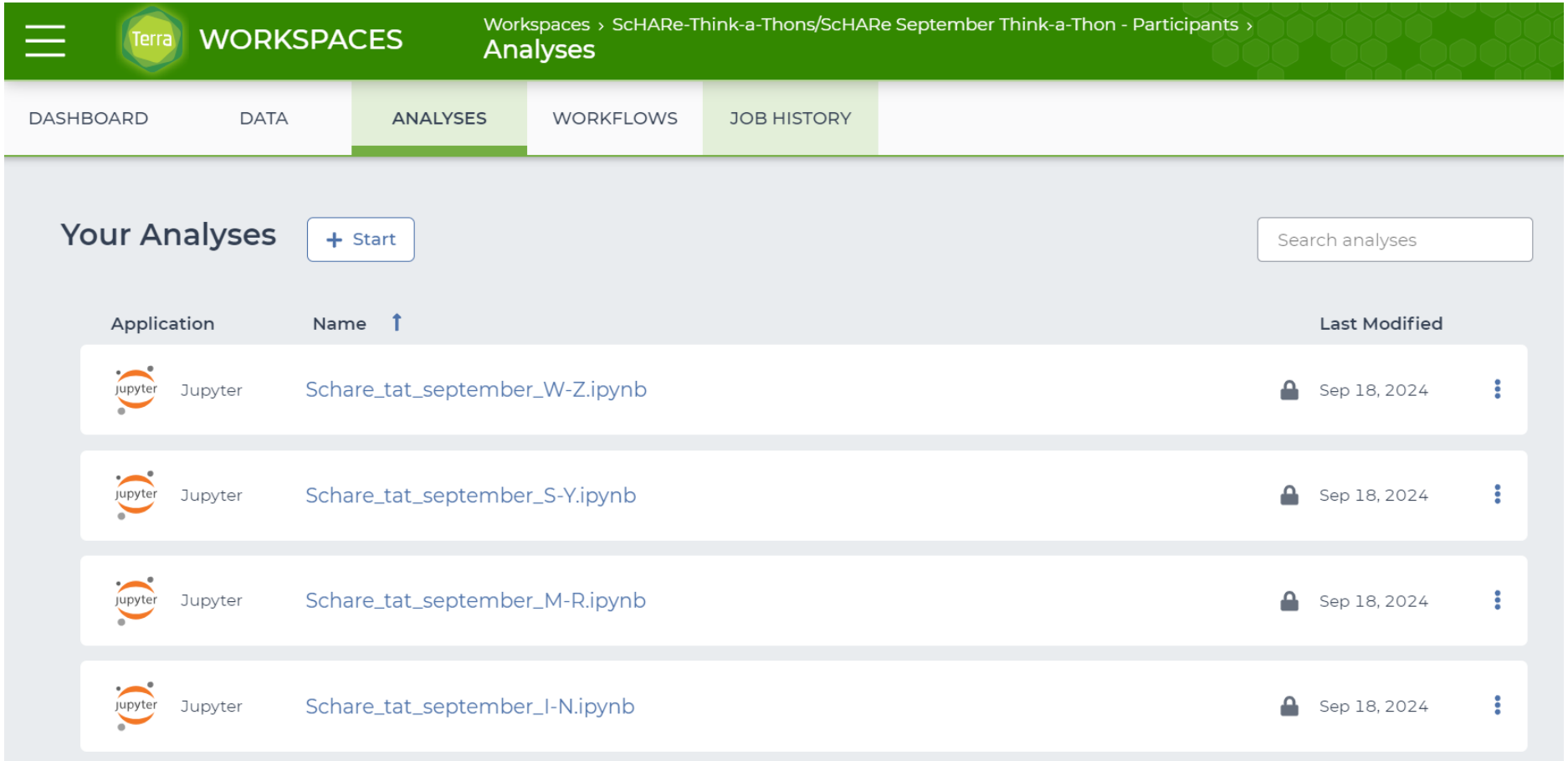


Newly registered users might see this message:

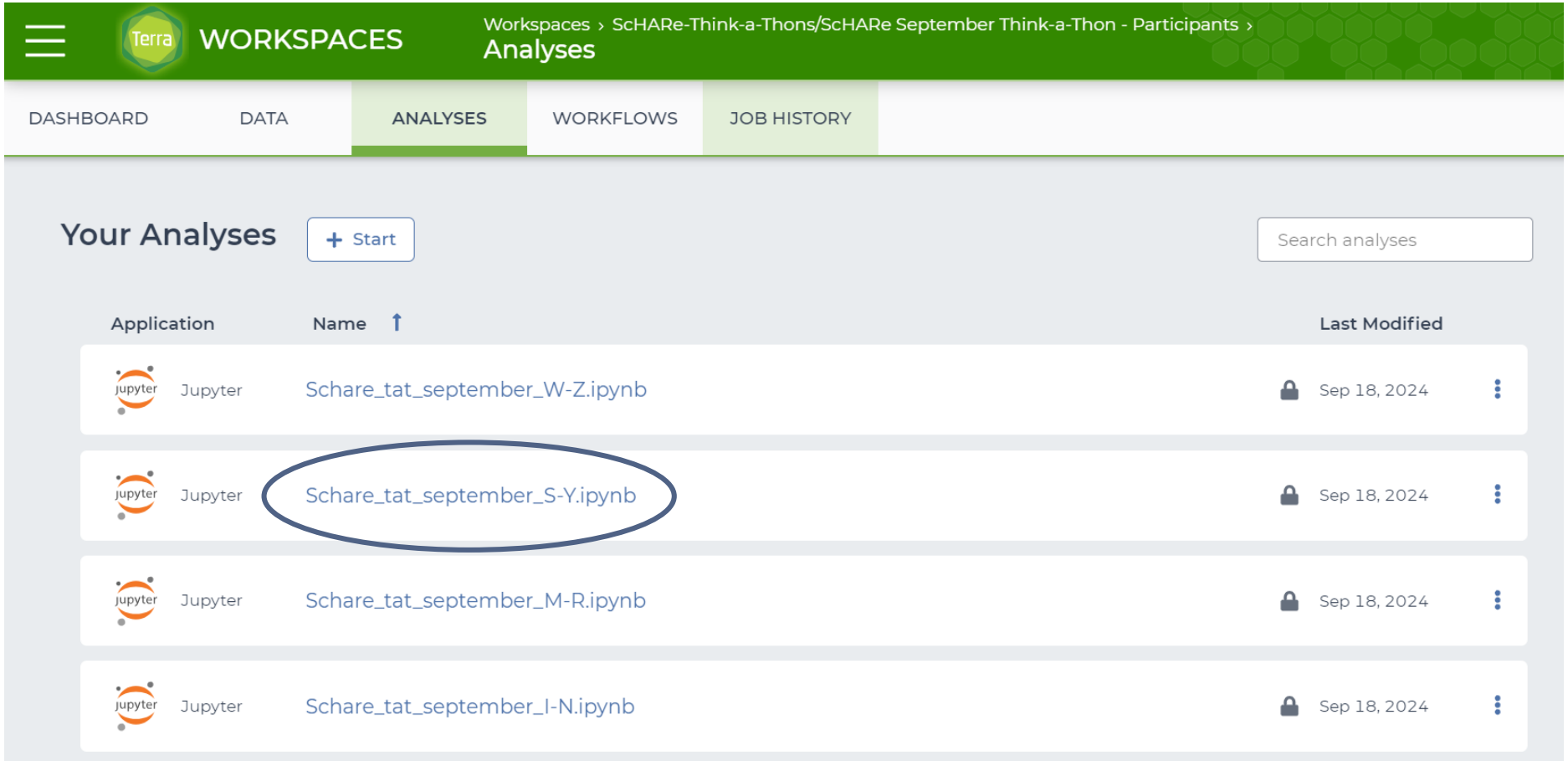


This is normal: the message should go away in a few minutes

Refreshing the page after a while, all users should see this:

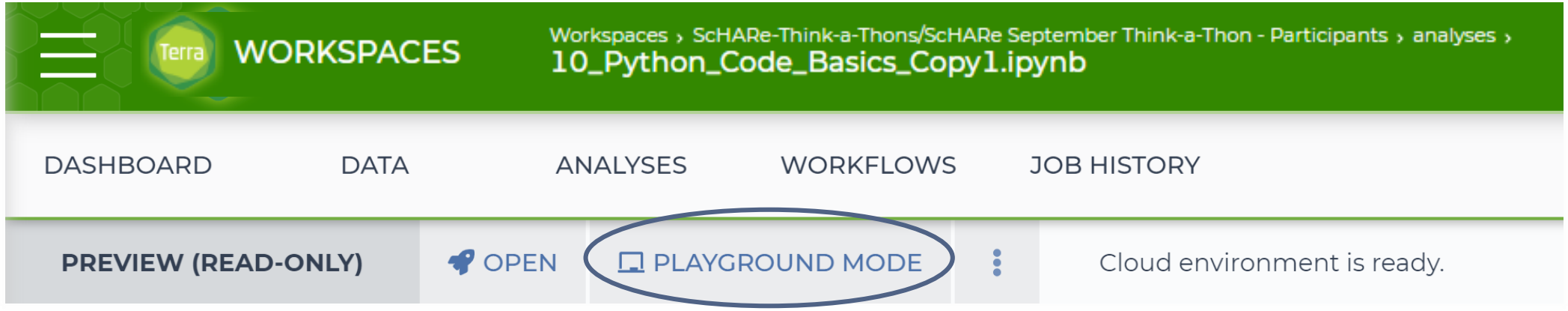


Click on the notebook containing your last name initial:



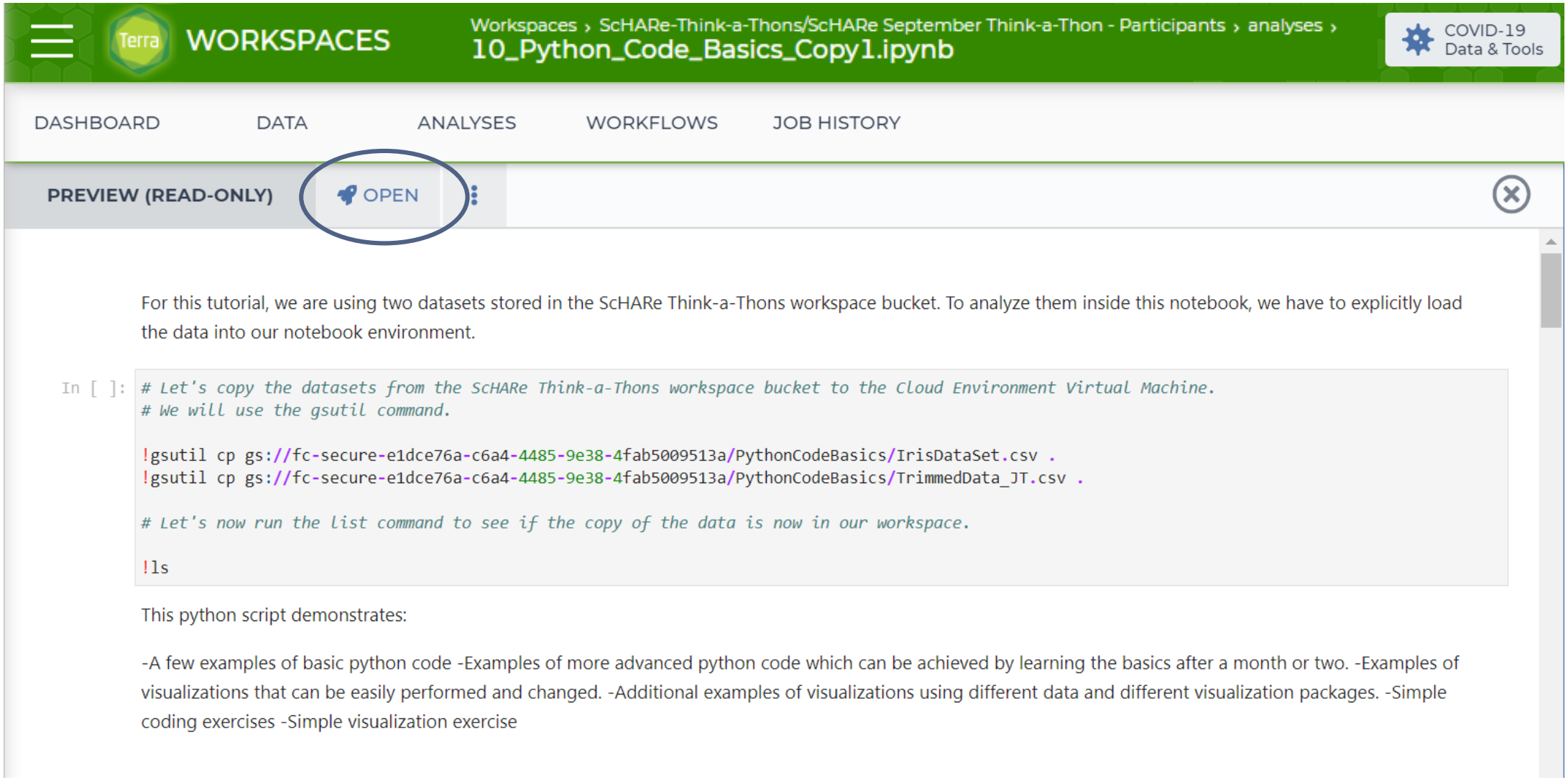
For example, if your last name starts with “S”, click on the notebook highlighted above

Do you see a Playground mode button?

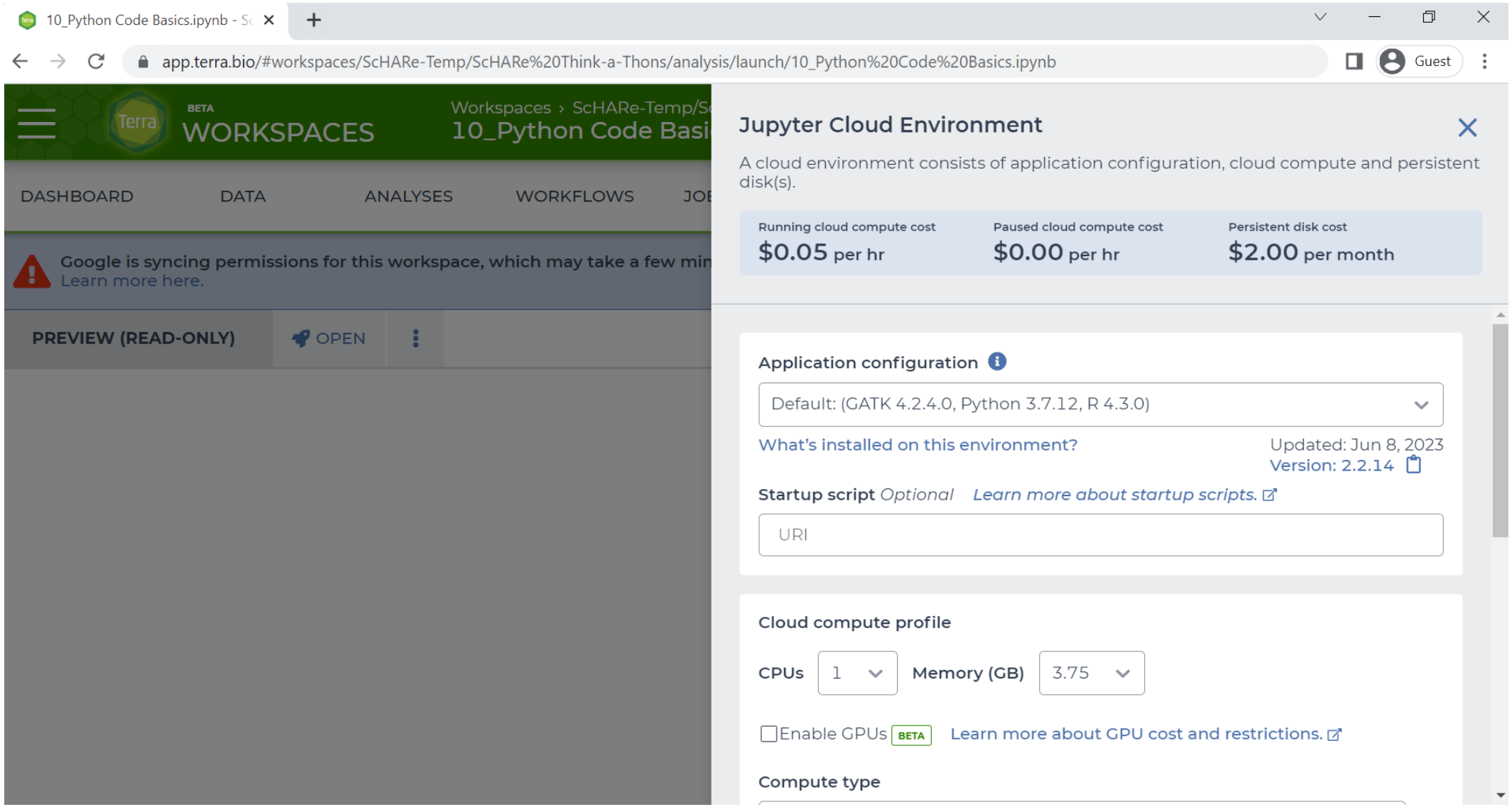


If yes, click on it to start your virtual computer. You are done!

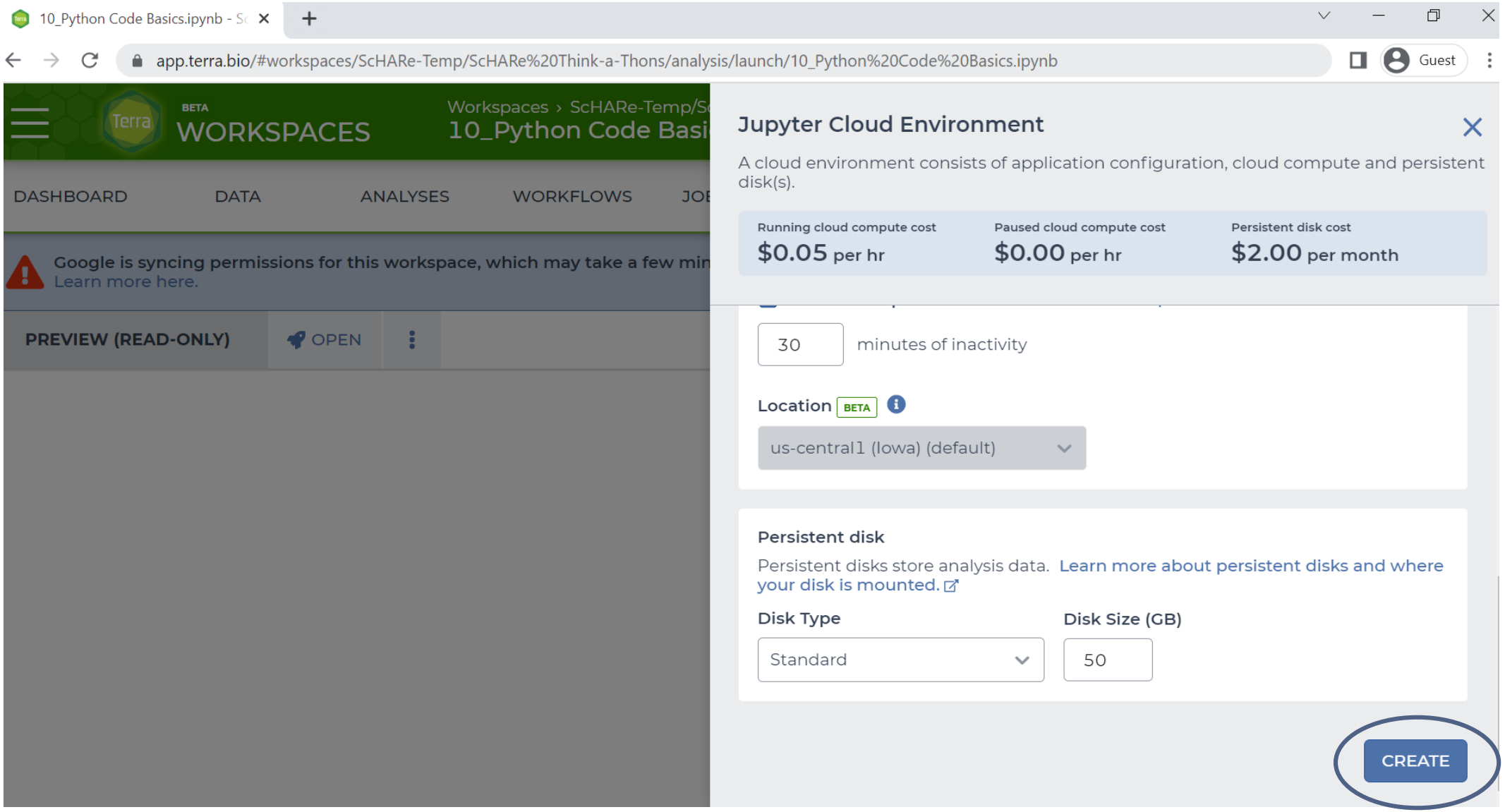
If you don’t see Playground mode, click on the Open button:



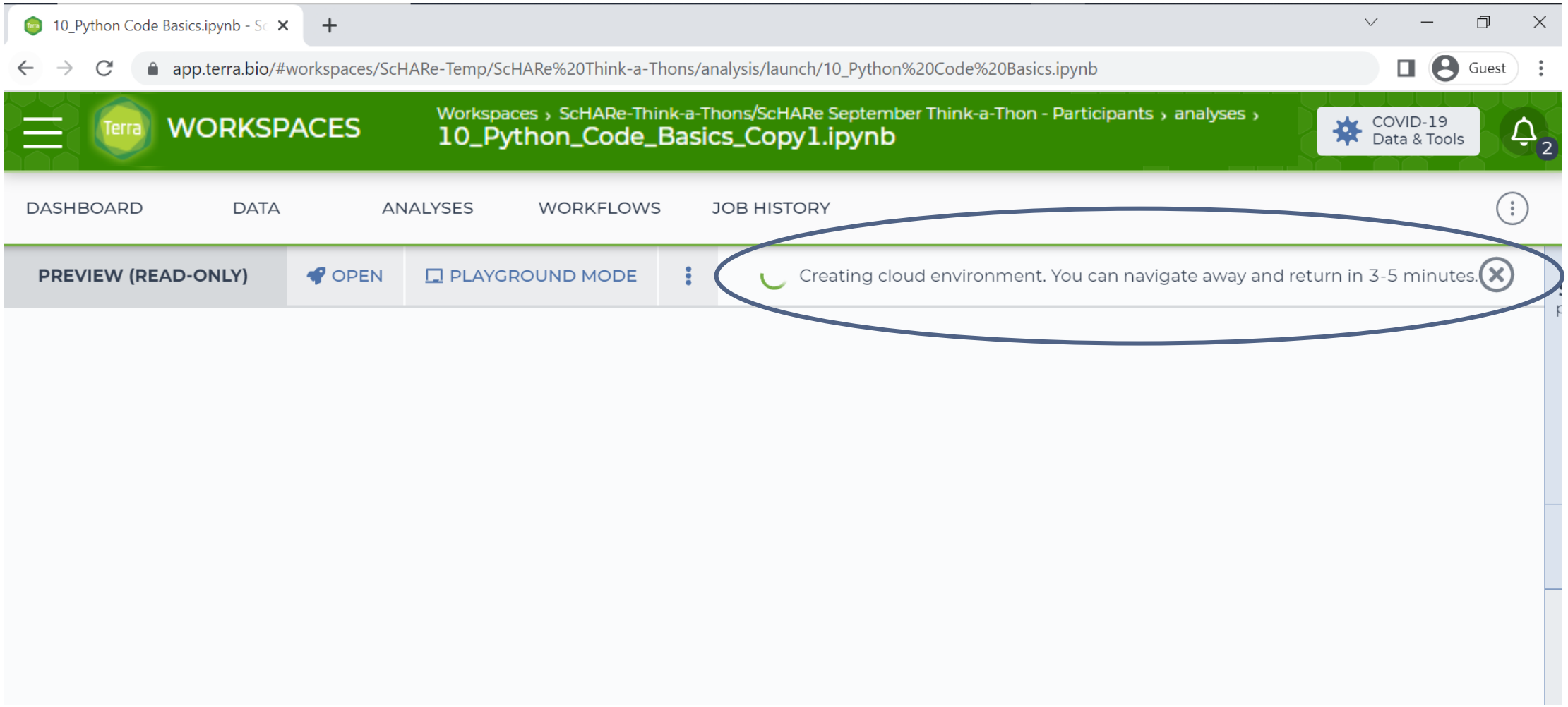
Configure your virtual computer – accept the default values:



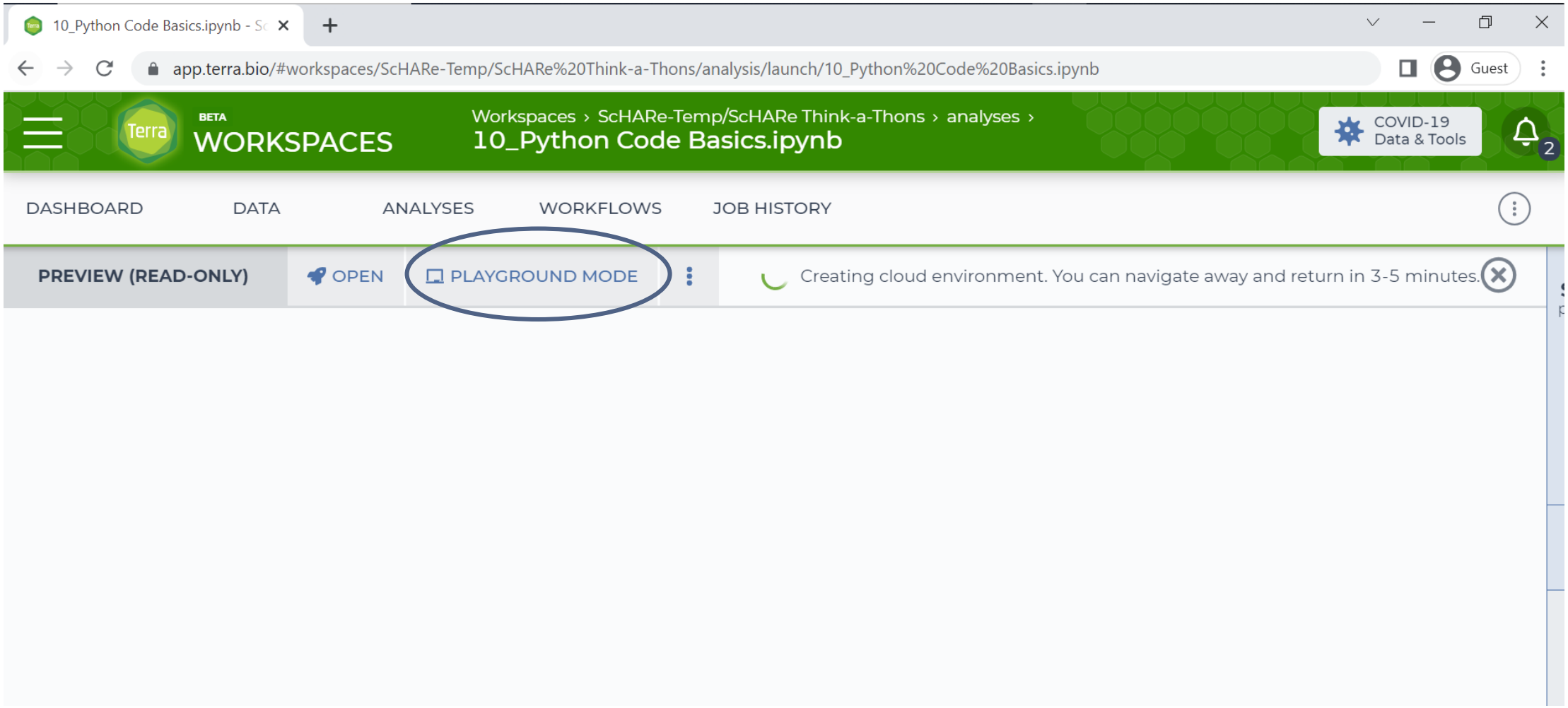
Click on Create below:



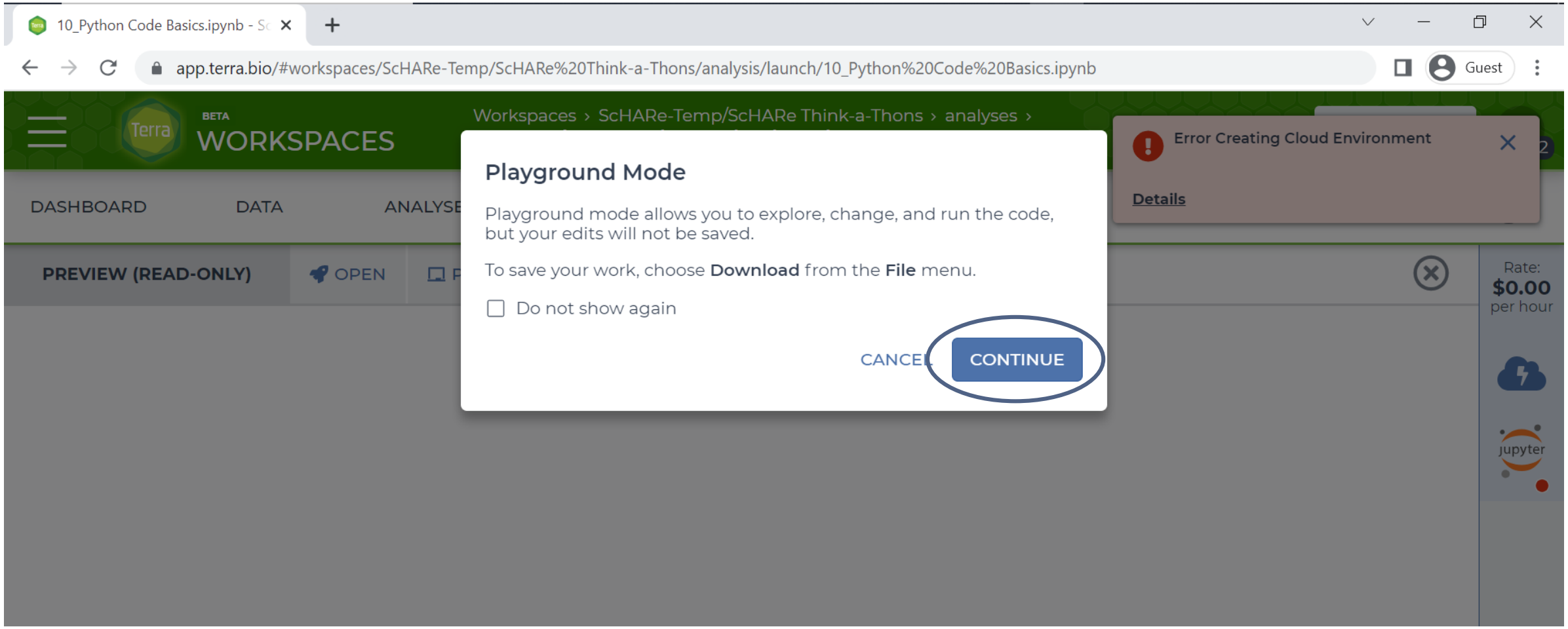
It will take some time…



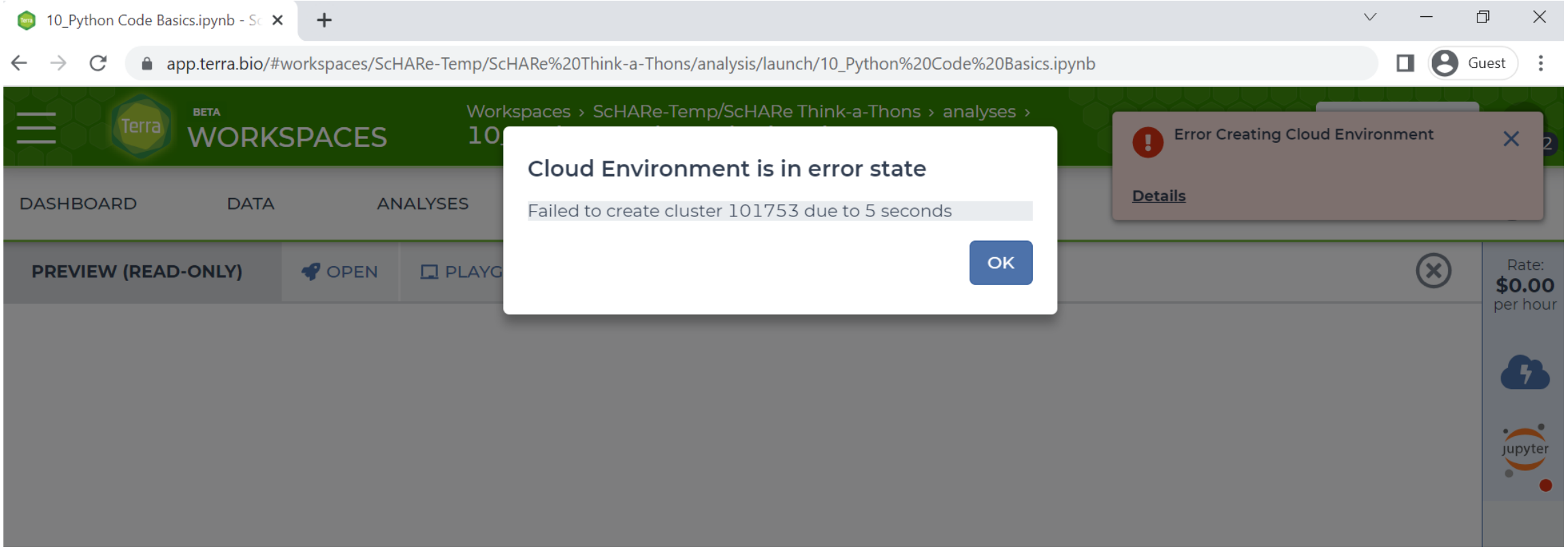
When the system is ready, click on Playground mode:



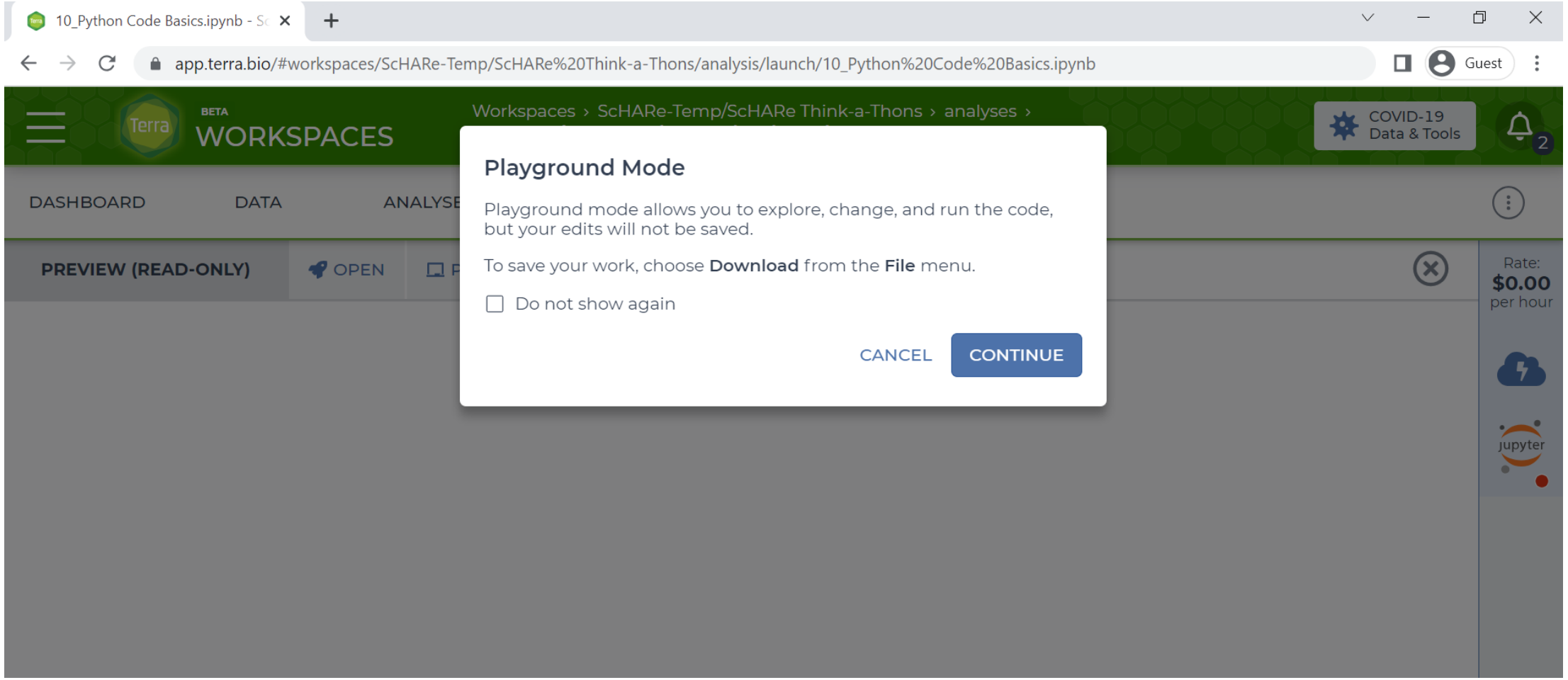
Click on Continue:



Note that you might encounter an error due to the large number of users – just try again in a few minutes:



If all goes well, you will see this:



Click on Continue. You are all set!

ScHARe

Why Python?

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

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

Terra 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

ScHARe

Data Management and Analysis in Python

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe

Guest expert

Cindy Sheffield

NIH/OD/ORS

About Cindy

Cindy is **Data Services Librarian** at the NIH Library.

She began her **library career** at the Johns Hopkins Medical Institutions with a focus on Evidenced Based Medicine. She progressed within the Welch Medical Library, leaving Hopkins as the Associate Director of Education Services.

Cindy has worked at several **federal agencies** including the Department of Homeland Security, the Department of Defense, and the Department of Health and Human Services. Within DHHS she was worked for both the National Institutes of Health and the Federal Drug Administration.

Her **focus** has always been on using key resources to identify the best evidence, and then to organize and manage that evidence in a way that makes sense for users. At the NIH she works with various user groups to support literature research and data science.

She is the Outreach Librarian for the NIH Clinical Centers, Pain and Palliative Care Team, the Eunice Kennedy Shriver, National Institute of Child and Human Development, the Administration for Children and Families, and the Office of the National Coordinator for Health Information Technology.

ScHARe

Guest expert

Sarvesh Soni

NIH/NLM

About Sarvesh

Dr. Sarvesh Soni is a Research Fellow with Dr. Dina Demner-Fushman at the National Library of Medicine.

Dr. Soni has a PhD in Biomedical Informatics from The University of Texas Health Science Center at Houston (UTHealth). He researches clinical natural language processing (NLP), focusing on question answering (QA) from both structured and unstructured data present in electronic health records (EHRs).

He implemented methods to generate paraphrases of clinical questions automatically and improve EHR QA and designed systems to automatically retrieve EHR text documents and underlying exact answer spans for given clinical information needs.

National Institutes of Health
Office of Management

ScHARe

ScHARe Think-a-Thon Series: An Introduction to Python for Data Science, Part 2

Cindy Sheffield, Biomedical Librarian, NIH Library
Sarvesh Soni, Research Fellow, National Library of Medicine

Introduction

Agenda

- Recap from August session – 10 min
- Importance of data cleaning – 10 min
- Tools for data cleaning – 10 min
- How data impacts visualizations – 10 min
- Machine Learning primer – 10 min
- Examples of Visualizations, Data Cleaning, Machine Learning – 80 min

Attendees will be able to:

- Know how to find Python libraries to help with code functionality
- Understand the importance of data cleaning
- Know what tools are available to help with data cleaning
- Visualizations and the importance of telling an accurate story
- Understand the mechanisms behind Machine Learning

Recap from Part 1:

Slido quiz

What is a Python library?

- ☐ A collection of books about Python programming
- ☐ Answer B A collection of related modules that provide specific functionality
- ☐ A place to store Python code
- ☐ A way to access Python from the command line

Python Libraries

Python Libraries – a collection of related modules that provide more extensive functionality and solve specific problems

Sample libraries:

- Numpy
- Pandas
- Matplotlib

How to find libraries:

- PyPI.org
- GitHub

Slido quiz

Which of the following are examples of Python libraries?

- ☐ Excel, OpenRefine
- ☐ Matplotlib, Pandas, Numpy
- ☐ R, SQL
- ☐ GitHub, PyPI

Data Cleaning / Data Wrangling

Data Cleaning

Ensure:

- Data accuracy
- Data consistency
- Data quality
- Efficiency

Processes:

- Parsing (First/Last Name)
- Correcting (Typos, errors)
- Standardizing (format)
- Match (id duplicates)
- Consolidating (clean presentation)

Slido quiz

Why is clean data important?

- ☐ It allows for better decision-making and saves time
- ☐ It makes data look nice without adding any practical value
- ☐ It removes all irrelevant information from public datasets
- ☐ It ensures that data can never be incorrect

Types of Dirty Data

- Corrupted
- Inaccurate
- Duplicates
- Irrelevant information

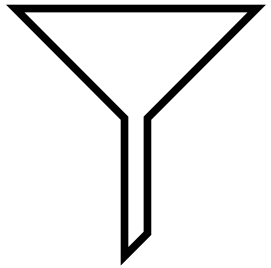
Establish quality control standards:

- Account for missing values
- De-duplication / Consolidating
- Irrelevant information
- Normalize non-standard values
- Understand outliers vs. incorrect data
- Change case if needed
- Check for bad values in fields(i.e.: alpha vs. numeric, formatting, spacing)
- Ensure overall data quality

Data Wrangling

Six step process:

- ***Explore***
- ***Transform***
- ***Clean***
- ***Enrich***
- ***Validate***
- ***Store***



Data Wrangling –

Mapping, merging, concatenating, or converting data, to transform the content, so it can be used for algorithmic processing and analysis.

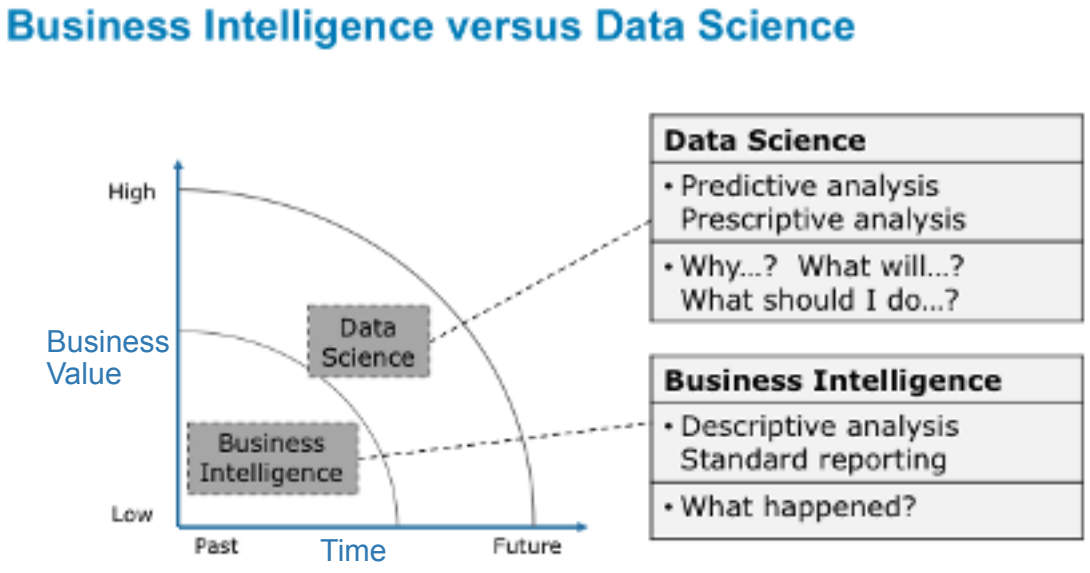
Slido quiz

Which of the following is part of the data wrangling process?

- ☐ Transforming data to prepare it for analysis
- ☐ Writing code in a programming language
- ☐ Saving data as images
- ☐ Downloading data from the internet

Benefits of Clean Data

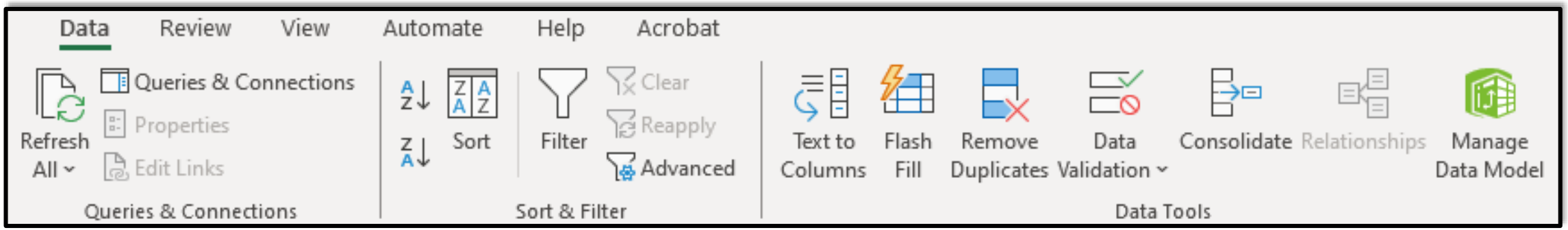
- Allows for informed decision making, and it is the precursor to artificial intelligence.
- Enhances efficiencies by saving time, effort, and resources.
- Improves satisfaction for consumers and producers
- In Public Health and Regulatory environments, it helps to maintain trust and avoid legal actions.



Tools

- Excel: Functions within Excel
- R: dplyr, tidyr, rrefine
- Python: Pandas, NumPy
- OpenRefine

Excel



Open Refine [openrefine.org]

- Open Source
- Desktop application
- Data cleanup and transformation
- Faceting
- Clustering
- Reconciling

OpenRefine:

- is ‘a tool for working with messy data’
- works best with data in tabular format
- can help split data into more granular parts
- can help match local data to other data sets
- can help enhance a data set with data from other sources

Tutorial: Library Carpentry: OpenRefine: https://librarycarpentry.org/lc-open-refine/instructor/aio.html

Slido quiz

What is OpenRefine used for?

- ☐ Word processing
- ☐ Creating spreadsheets
- ☐ Data cleanup and transformation
- ☐ Developing websites

Books/Articles on Data Cleaning

- 1. Buttrey S, Whitaker LR. *A data scientist's guide to acquiring, cleaning and managing data in R*. 1st edition ed. THEi Wiley ebooks. Wiley; 2017.
- 2. Gueta T, Carmel Y. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological informatics*. 2016;34:139-145. doi:10.1016/j.ecoinf.2016.06.001
- 3. Martin N, Martinez-Millana A, Valdivieso B, Fernández-Llatas C. Interactive Data Cleaning for Process Mining: A Case Study of an Outpatient Clinic's Appointment System. Springer International Publishing; 2019:532-544. *Lecture Notes in Business Information Processing*.
- 4. Mertz D. *Cleaning data for effective data science : doing the other 80% of the work with Python, R, and command-line tools*. Packt Publishing; 2021.
- 5. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med*. Oct 2005;2(10):e267. doi:10.1371/journal.pmed.0020267
- 6. Walker M. *Python Data Cleaning Cookbook : Prepare Your Data for Analysis with Pandas, NumPy, Matplotlib, Scikit-Learn, and OpenAI*. Packt Publishing, Limited; 2024.
- 7. Wang X, Wang C. Time Series Data Cleaning: A Survey. *IEEE access*. 2020;8:1866-1881. doi:10.1109/ACCESS.2019.2962152

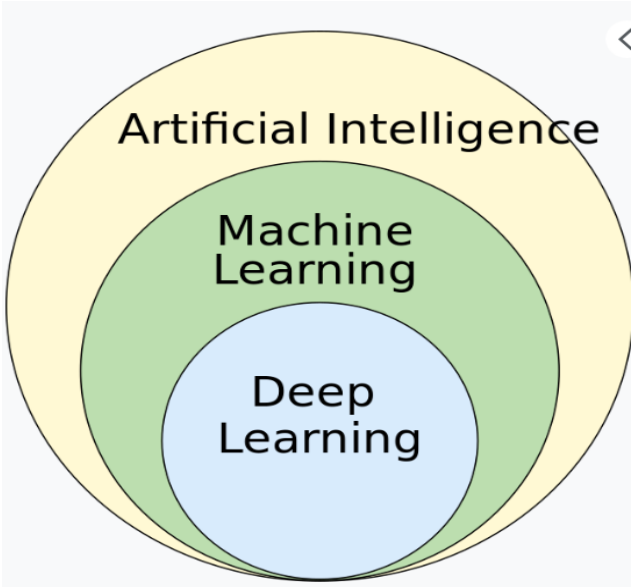
Visualizations

- Use to better understand data
- Prepare the data so it tells an accurate story
- Understand the data and any potential bias

Machine Learning

Machine Learning and Artificial Intelligence (AI)

- Machine Learning - type of AI and CS
- Improves how software systems process and categorize data
- Focuses on the use of data and algorithms
- Imitate human learning
- Gradually improving its accuracy
- ML algorithms imitate human learning
- ML algorithms improve over time as they take large data sets



https://bootcamp.berkeley.edu/blog/how-does-machine-learning-work/

Types of Machine Learning

Types of Machine Learning

Supervised Learning

Classification
Uses an algorithm to accurately assign test data into specific categories, such as separating apples from oranges.

Regression
Uses an algorithm to understand the relationship between dependent and independent variables.

Unsupervised Learning

Clustering
Data mining technique for grouping unlabeled data based on their similarities or differences.

Association
Uses different rules to find relationships between variables in a given data set.

Dimensionality
Used when the number of features (or dimensions) in a given data set is too high.

Semi-Supervised Learning

Combines supervised and unsupervised machine learning techniques.

Uses smaller labeled data to guide classification and feature extraction from a larger unlabeled data set.

Reinforcement Learning

Learns as it goes through trial and error, rather than being trained using sample data, like in supervised learning.

https://bootcamp.berkeley.edu/blog/how-does-machine-learning-work/

Types of Machine Learning - Supervised

- Taught by example
- Training data is fed into an algorithm and teaches to categorize based on pre-set characteristics
- Algorithm can similarly sort raw data
  - Good at classifying data into pre-set categories
    Example: identify spam emails or telling images apart

https://bootcamp.berkeley.edu/blog/how-does-machine-learning-work/

Types of Machine Learning - Unsupervised

- Uses algorithms to sort unlabeled and unstructured data
- Algorithms discover data patterns without human intervention
- Good situations without clear delineations between different data categories
- Example:
  - Recommend similar types of research projects or publications

https://bootcamp.berkeley.edu/blog/how-does-machine-learning-work/

Types of Machine Learning – Semi-supervised

- Combines supervised and unsupervised machine learning to sort or identify data
- Involves labeling some data
- Involves rules and structure for the algorithm to use to start sorting and identifying data
- A small amount of tagged data improves an algorithm's accuracy
- Example: classify content in scanned documents: typed and handwritten

https://bootcamp.berkeley.edu/blog/how-does-machine-learning-work/

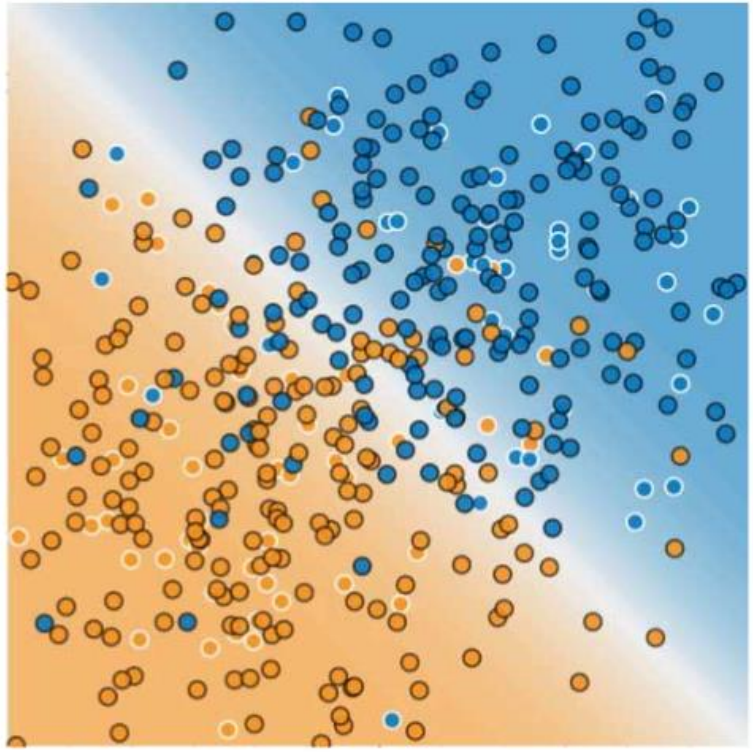
Types of Machine Learning – Reinforcement

- Used for decision-making in a complex, uncertain environment
- Game-like rules system designed to maximize algorithm's score
- Programmers define rules; computer starts without guidance
- Computer learns through trial and error for optimal solutions
  - Example: used for language processing, self-driving vehicles and game-playing AIs

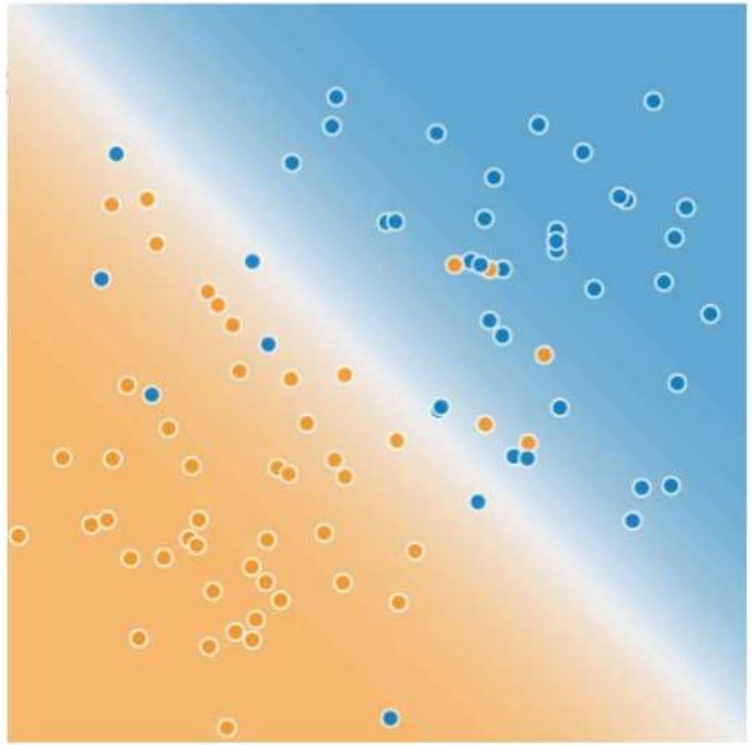
https://bootcamp.berkeley.edu/blog/how-does-machine-learning-work/

Training Set and Test Set

- **Training Data Set**—a subset to train a model.
- **Test data set**—a subset to test the trained model.



Training Data



Test Data

https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data

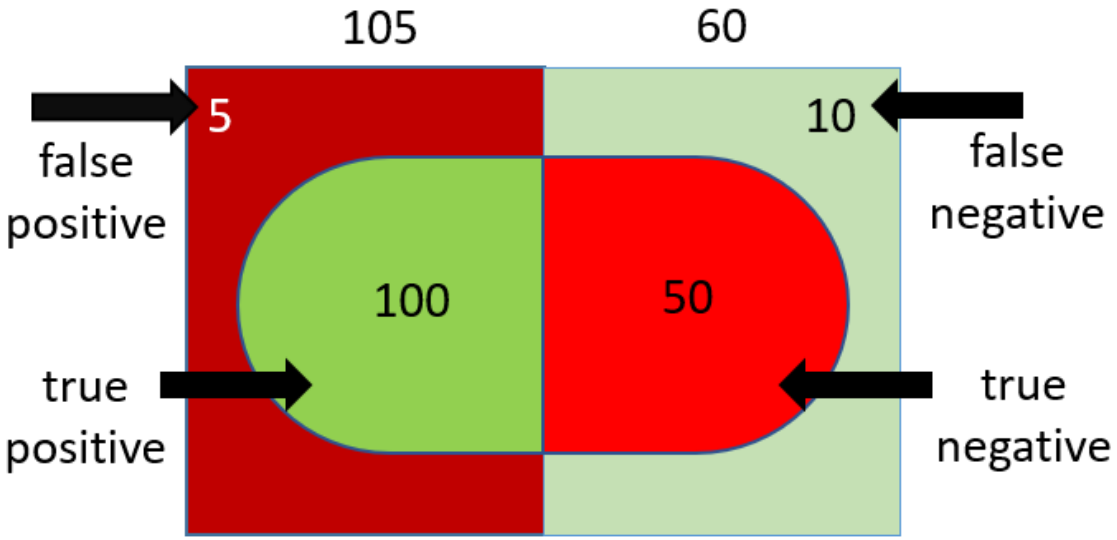
Evaluating Machine Learning Performance

		Actual (ex. Manual coding)		
		Positive	Negative	
ML model/ Algorithm Predictions	Positive	True Positive (TP)	False Positive (FP)	Positive Predictive Value
	Negative	False Negative (FN)	True Negative (TN)	Negative Predictive Value
		Sensitivity	Specificity	

- **Accuracy:** how much did the model get right; % of predictions the model or algorithm gets correct; = (TP + TN)/(TP+FN +FP+TN)
- **Precision:** also called positive predictive value (PPV); the quality of the positive predictions; % of positive predictions that were correct; =TP/TP+FP
- **Sensitivity:** also referred to as recall; measures how well a model can detect positive instances; =TP/TP+FN
- **Specificity:** measures how well the model identifies negatives instances; =TN/TN+FP
- **F1 score:** also used to assess accuracy of the model and it accounts for both precision and recall; =TP/TP + ½(FP+FN)

Machine Learning Performance Example

tp = AI Model found true positive = 100
fp = AI Model marked as positive; but negative = 5
tn = AI Model found true negative = 50
fn = AI Model marked as negative; but positive = 10



- **Accuracy**: how much did model get right;

  = (tp + tn)/(tp + fn +fp + tn) = 150 / 165 = **.9091**

- **Precision**: positive predictive value (PPV);

  = tp / tp + fp = 100 / 105 = **.9523**

- **Sensitivity**: recall; true positive instances;

  =tp/(tp+fn) = 100/100 + 10 = 100/110 = **.9091**

- **Specificity**: negatives;

  =tn/tn+fp = 50/50 + 5 = 50/55 = **.9091**

- **F1 Score**: assesses accuracy; precision and recall;

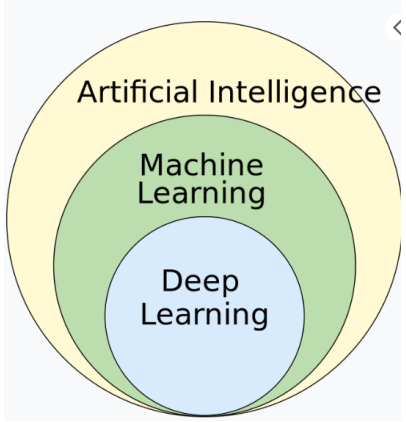
  *= 2 (precision * recall / precision + recall)*

  Or = TP/TP + ½(FP+FN)

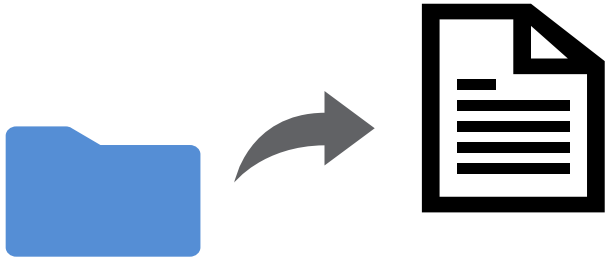
  = 100/100 + .5(10 + 5)

  = 100/107.5 = **.9302**

Deep Learning



- Subset of Machine Learning
- Teaches computers to process data similar to human brain
- Recognize picture patterns, text, sounds and other data
- Produce insights and predictions based on data
- Use to automate tasks typically done by humans:
  - describe images
  - transcribe files into text



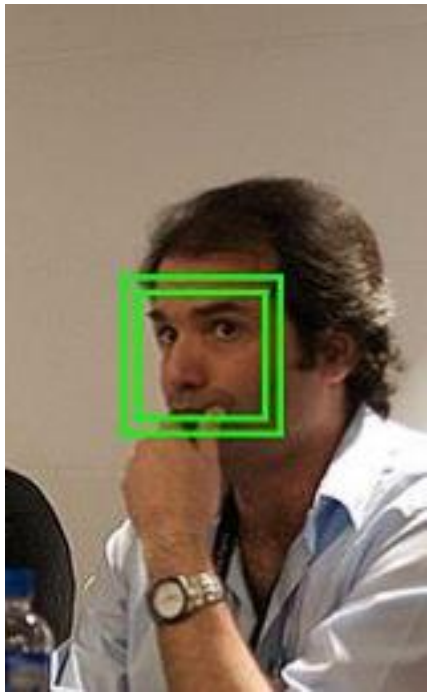
Deep Learning (Cont'd)

Used in everyday products:

- Digital assistants
- Voice-activated television remotes
- Fraud detection
- Automatic facial recognition

Uses of Deep learning:

- Self-driving cars
- Defense systems
- Medical image analysis
- Factories



Natural Language Processing (NLP)

- NLP is an artificial intelligence technique
- Subset of machine learning
- Allows machines to process and understand language like humans
- Uses computational linguistics combined with machine learning, deep learning and statistical modeling
- Understands intent and sentiment
- Stores information and context to strengthen future responses

https://monkeylearn.com/blog/natural-language-processing-examples/

Natural Language Processing (NLP) Benefits

- **Text analysis and data mining**
  - helps scientists extract valuable information from vast amounts of unstructured text data
- **Automated Literature Review**
  - allows for automated literature; speeds up gathering and summarizing research
- **Semantic Search and Information Retrieval**
  - enhances search engines, enabling more relevant results
- **Language Translation**
  - enables translation between different languages

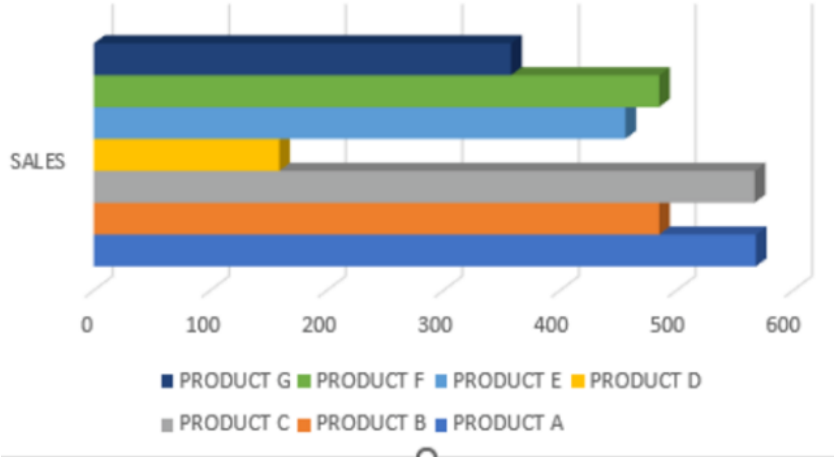
Natural Language Processing (NLP) Benefits (Cont'd)

- **Knowledge Representation**
  - convert textual information into structured data
- **Sentiment Analysis**
  - understand public opinion and reactions to scientific breakthroughs or research findings.
- **Question-Answering Systems**
  - enables specific questions and receive relevant answers from large databases or scientific literature
- **Automated Report Generation**
  - generate summaries, abstracts, or reports automatically, reducing manual effort

Natural Language Processing (NLP) Benefits (Cont'd)

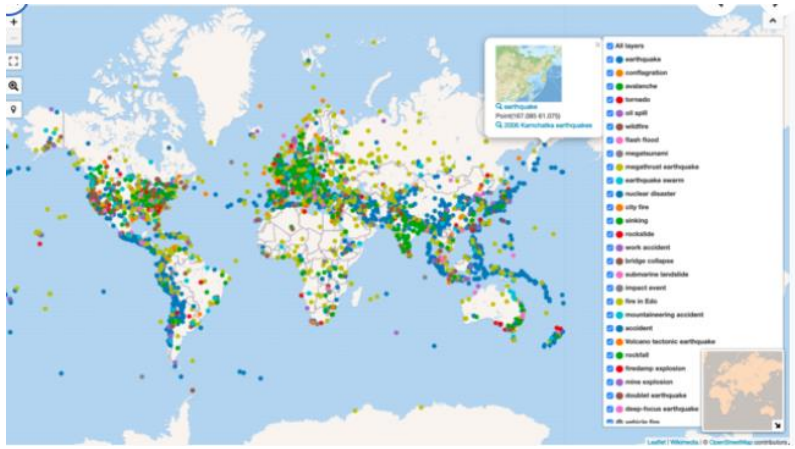
Clinical Applications

- analyze electronic health records
- extract important medical information
- diagnosing patients
- identifying patterns in medical data



Data Interpretation and Visualization

- interpret and understand complex scientific data
- generate visualizations



ScHARe

Data Management and Analysis in Python

September 18, 2024

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

ScHARe

Python tutorials and resources

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

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

freeCodeCamp(🔥)

Learn to code — free 3,000-hour curriculum

Python Tutorial for Beginners (Learn Python in 5 Hours)

In this TechWorld with Nana YouTube course, you will learn about strings, variables, OOP, functional programming and more. You will also build a couple of projects including a countdown app and a project focused on API requests to Gitlab.

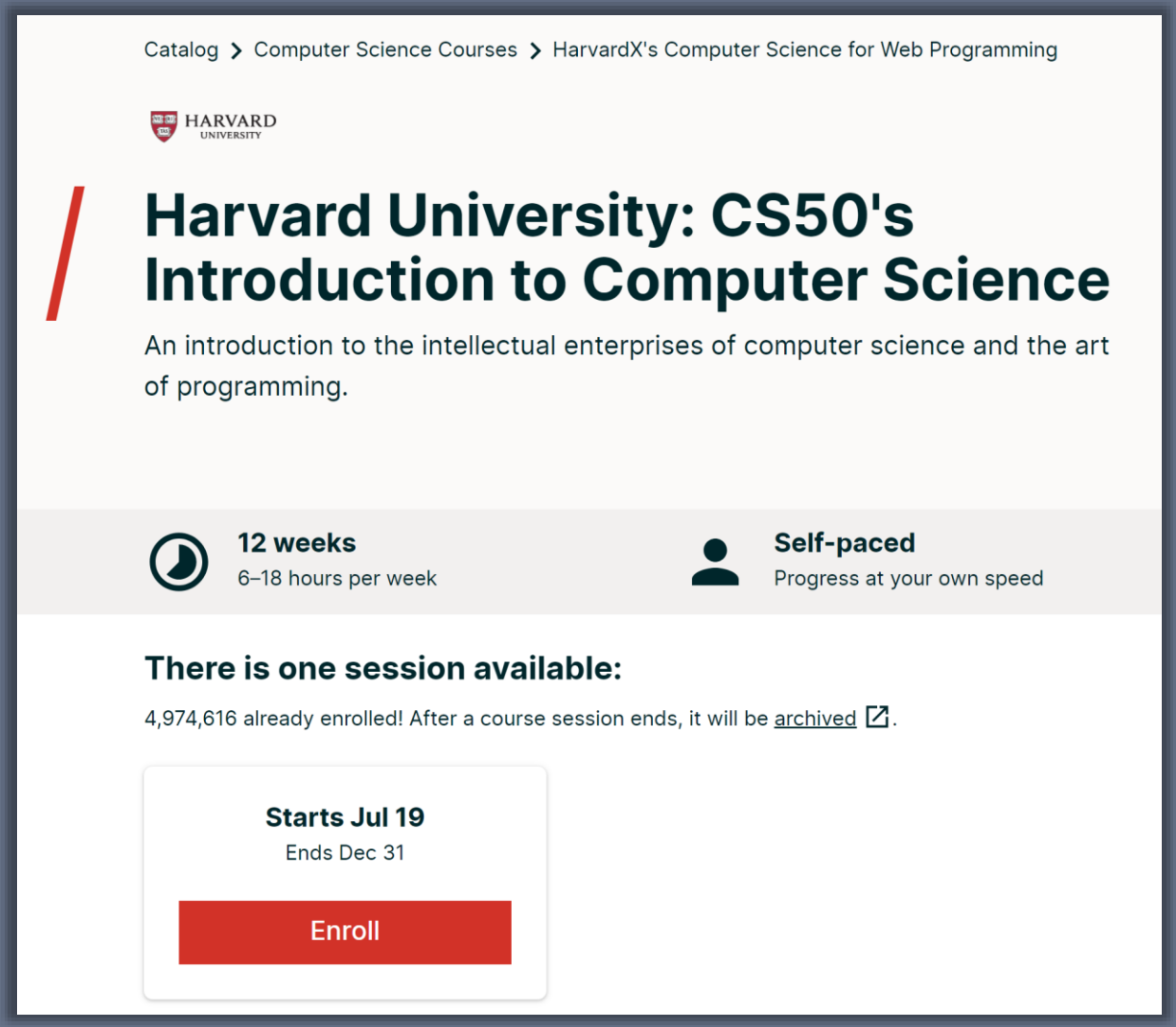
Scientific Computing with Python

In this freeCodeCamp certification course, you will learn about loops, lists, dictionaries, networking, web services and more.

Python resources

Harvard

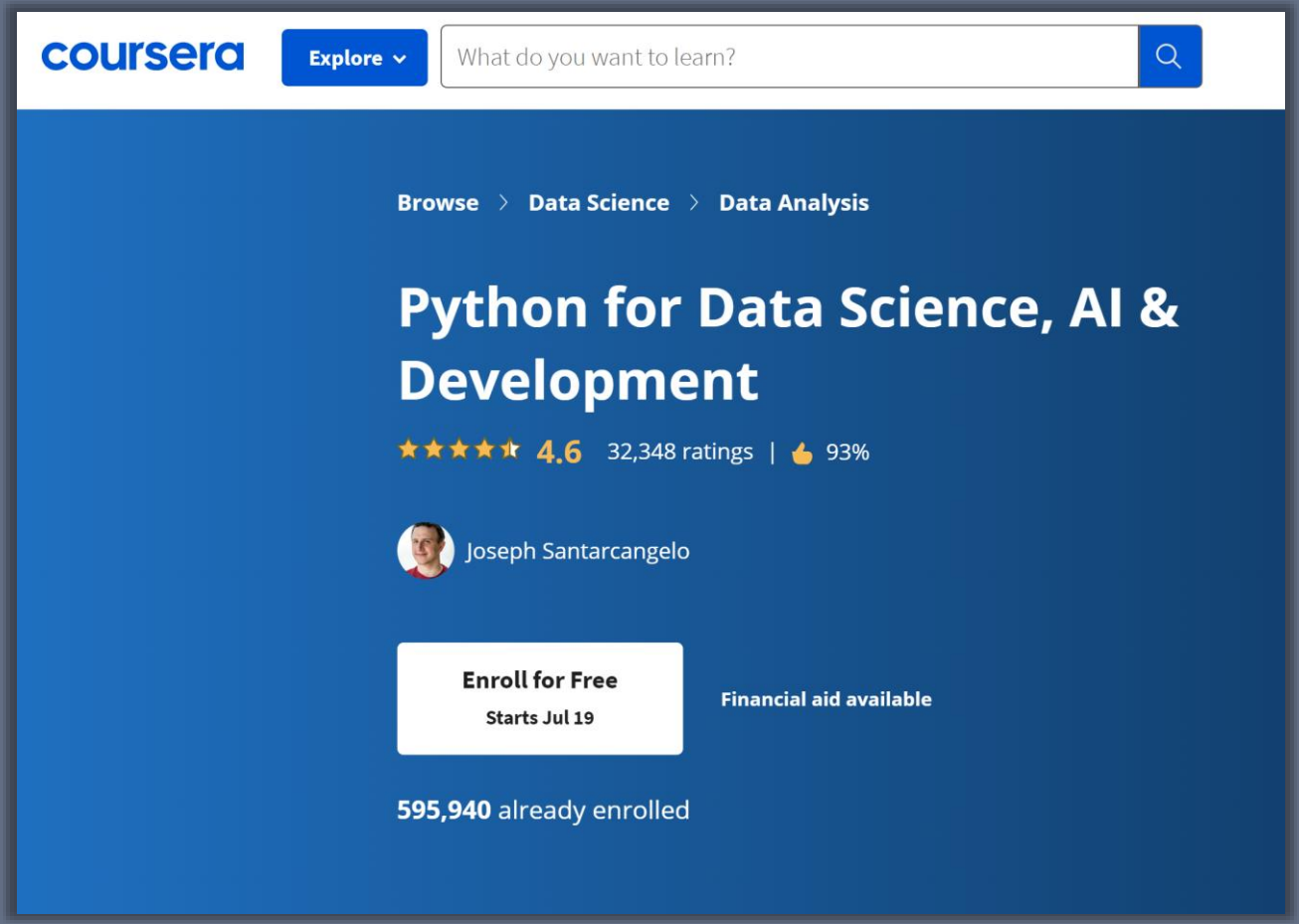
Free Python Course



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

Bootcamp Types Reviews Resources About BestColleges.com

Top 10 Free Python Courses

Google's Python Class

Students with some programming language experience can learn Python with Google's intensive two-day course. While there are no official prerequisites, students need a basic understanding of programming language concepts, such as if statements.

Learners initially explore strings and lists using lecture videos and written materials. A coding exercise follows each section, and the exercises become increasingly complex.

This Python course gives students hands-on practice with complete programs, working with text files, processes, and HTTP connections.

Microsoft's Introduction to Python Course

Students can learn Python online and build a simple input/output program with Microsoft's introductory Python course. There are no prerequisites for this short, eight-unit, 16-minute class.

This online Python course is part of Microsoft's Python learning paths. It prepares students with the concepts and basic skills to pursue more advanced learning.

Students explore Python code, where to run Python apps, learn how to declare variables, and use the Python interpreter. They also learn how to access free resources.

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

ScHARe

Thank you

Evaluation poll

1. **Rate how useful this session was:**

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

Evaluation poll

2. **Rate the pace of the instruction for yourself:**

- ☐ Too fast
- ☐ Adequate for me
- ☐ Too slow

Evaluation 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

ScHARe

Next Think-a-Thons:

bit.ly/think-a-thons

Register for ScHARe:

bit.ly/join-schare

schare@mail.nih.gov