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:

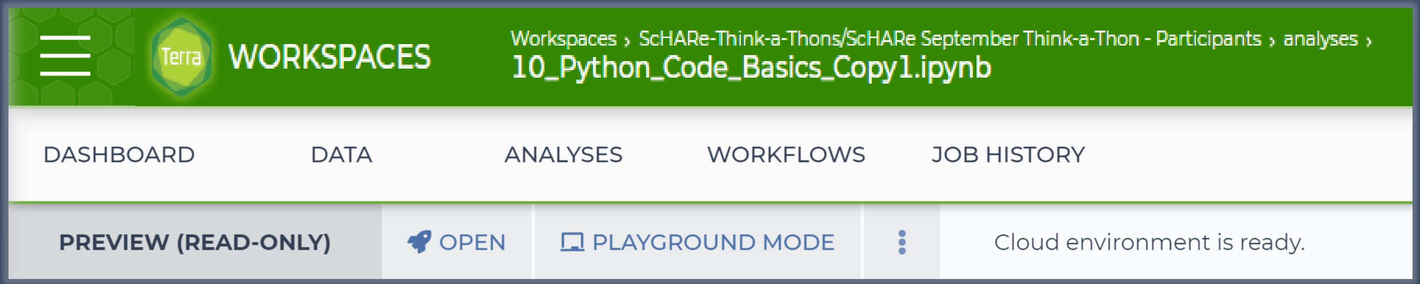
✓ 1. has provided their Gmail address and has been registered for ScHARe

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:



Workspaces › ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants › analyses ›
10_Python_Code_Basics_Copy1.ipynb

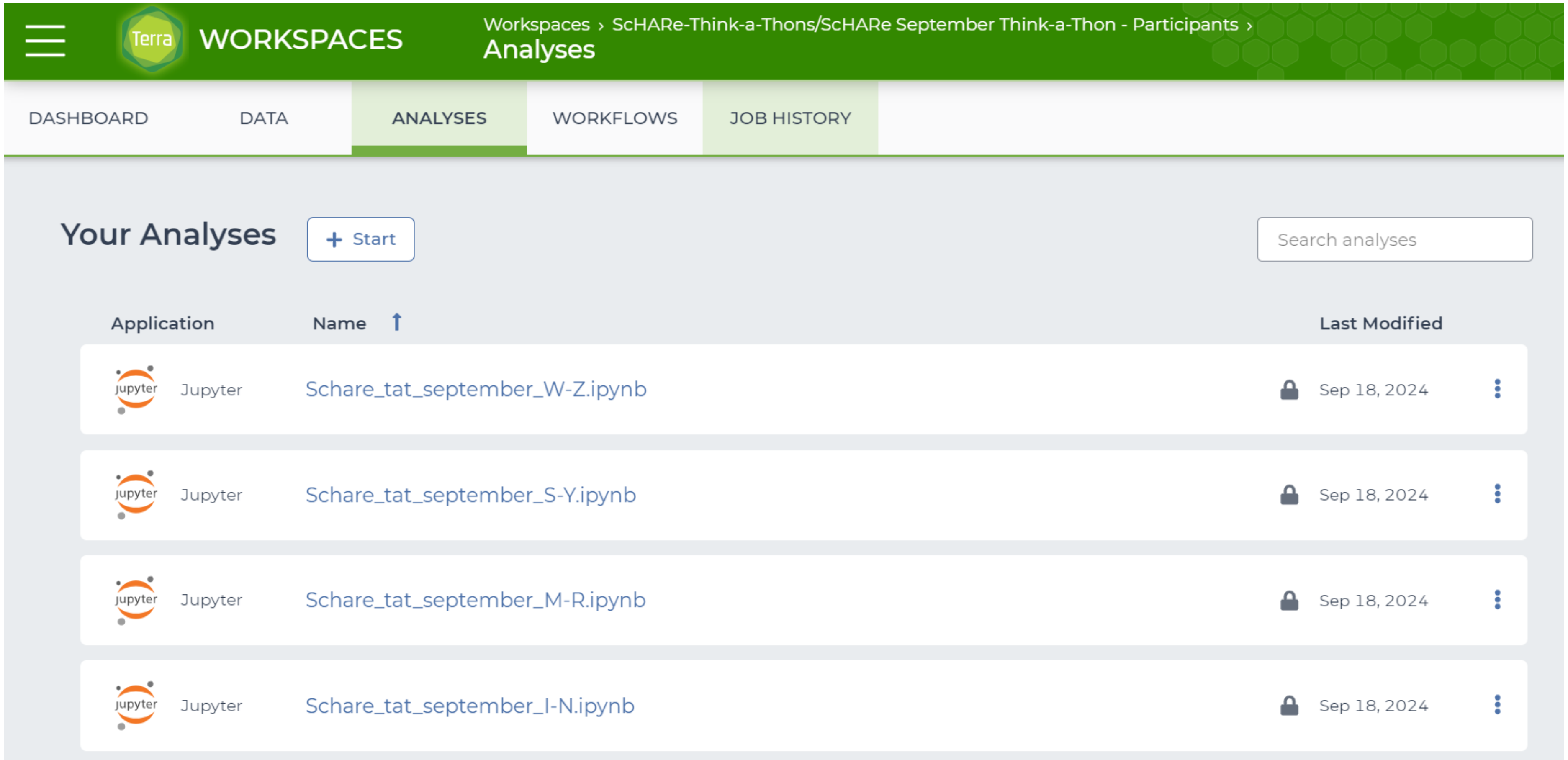
Cloud environment is ready.

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



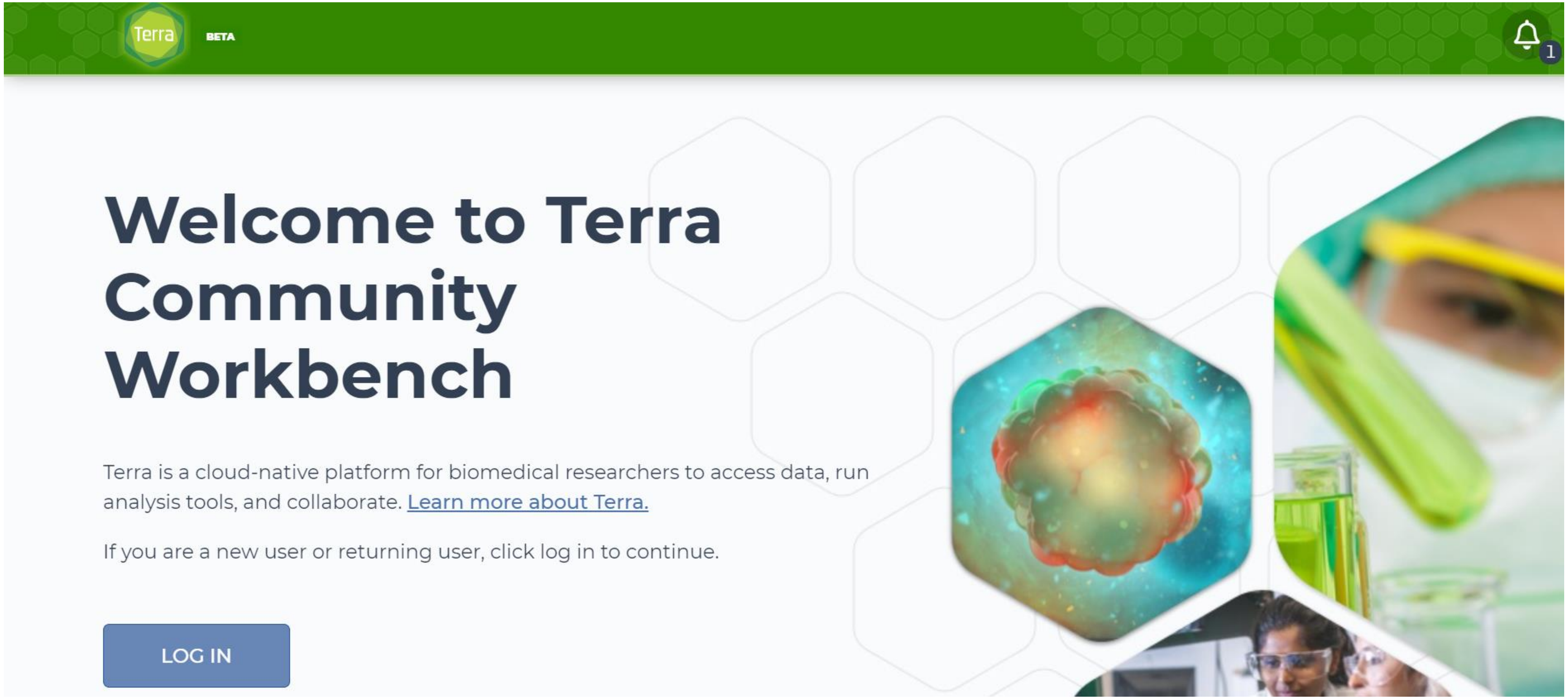
Workspaces › ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants ›
Analyses

Your Analyses

Application	Name	Last Modified
Jupyter	Schare_tat_september_W-Z.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_S-Y.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_M-R.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_I-N.ipynb	Sep 18, 2024

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

bit.ly/schare-python-notebooks-2



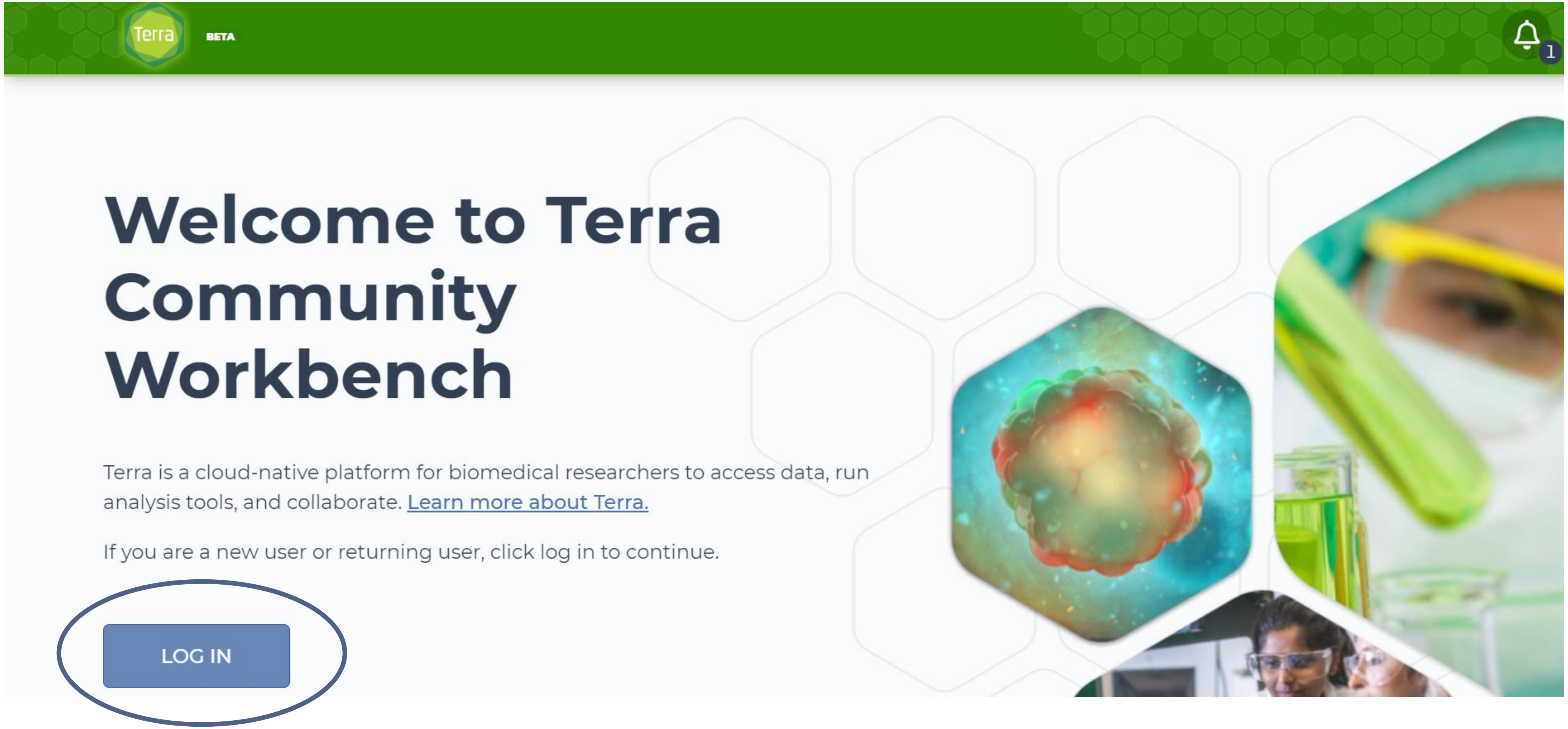
Welcome to Terra Community Workbench

Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate. Learn more about Terra.

If you are a new user or returning user, click log in to continue.

Click on the login button:

bit.ly/schare-python-notebooks-2

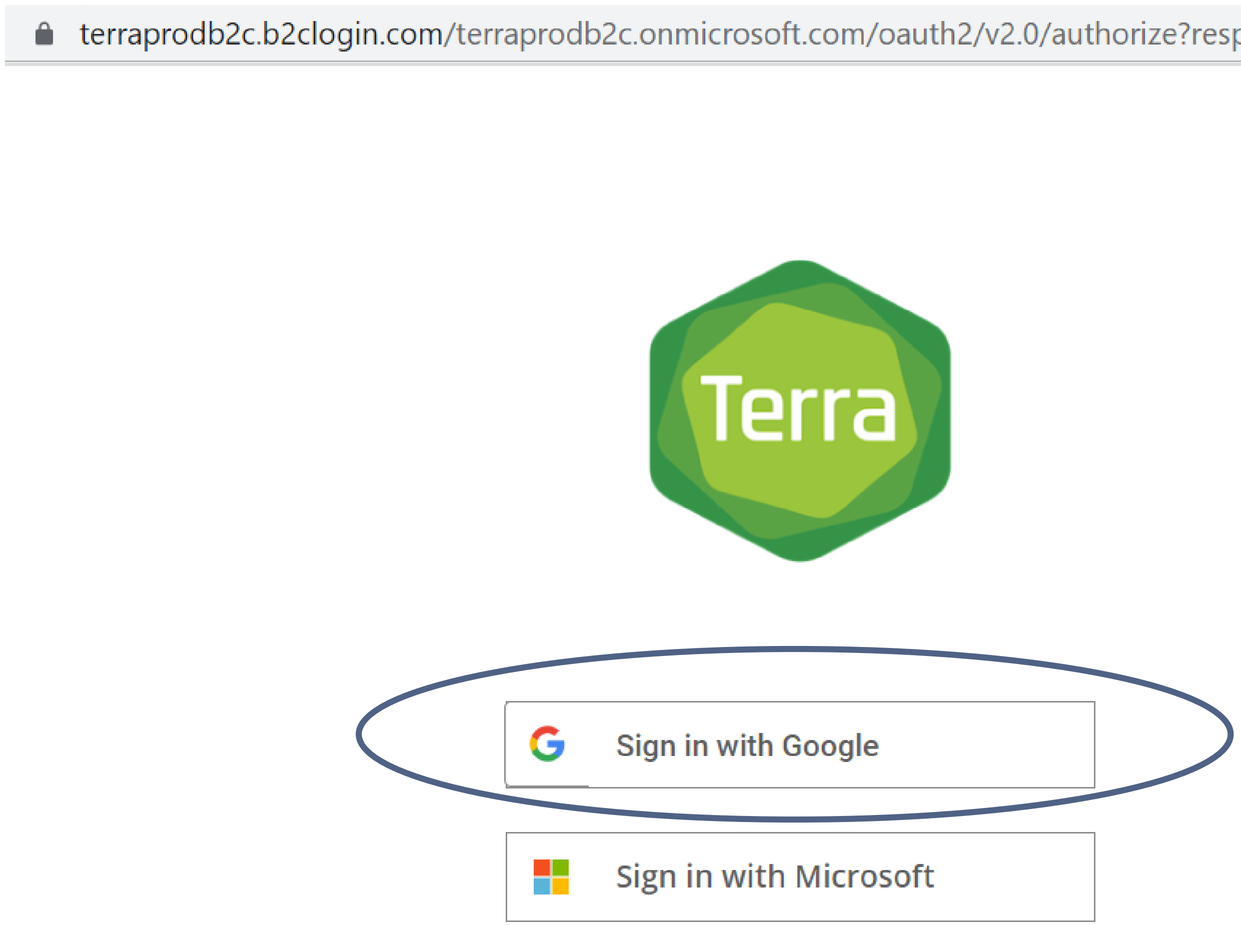


Welcome to Terra Community Workbench

Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate. Learn more about Terra.

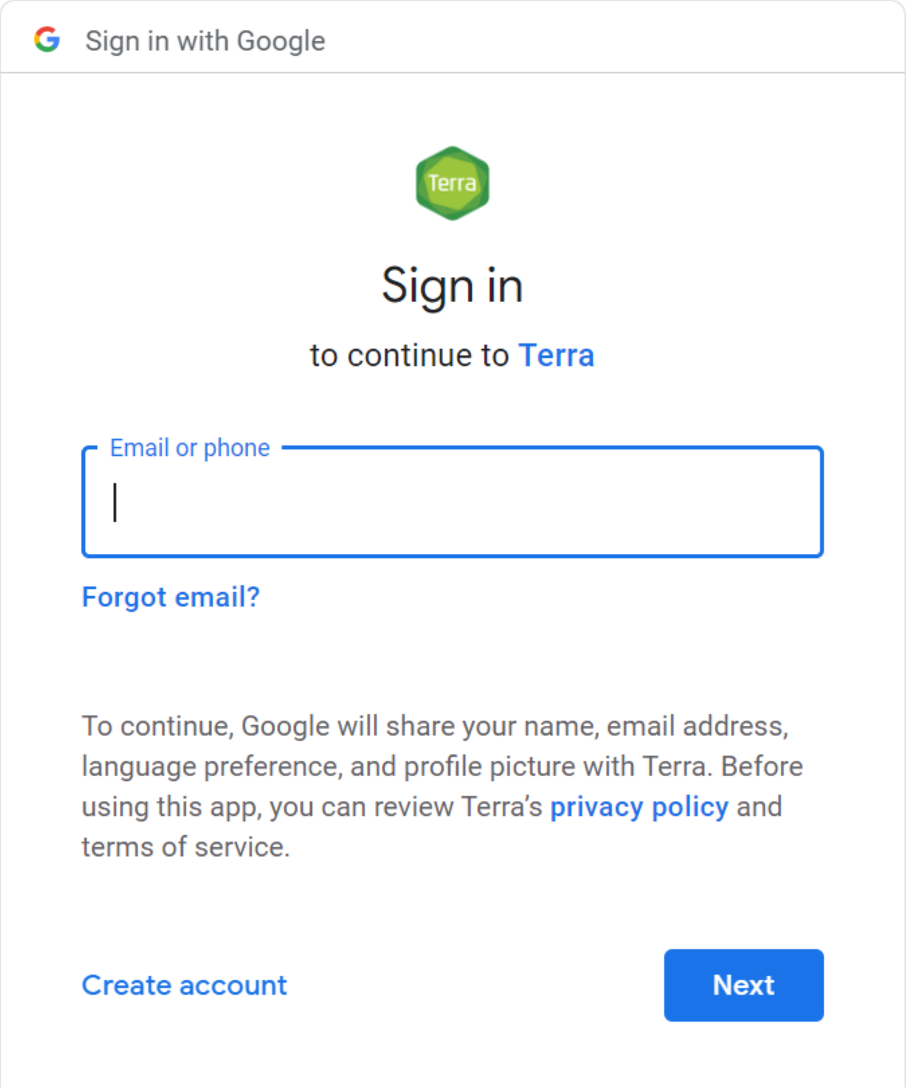
If you are a new user or returning user, click log in to continue.

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



terraprodb2c.b2clogin.com/terraprodb2c.onmicrosoft.com/oauth2/v2.0/authorize?response_mode=query&s...

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



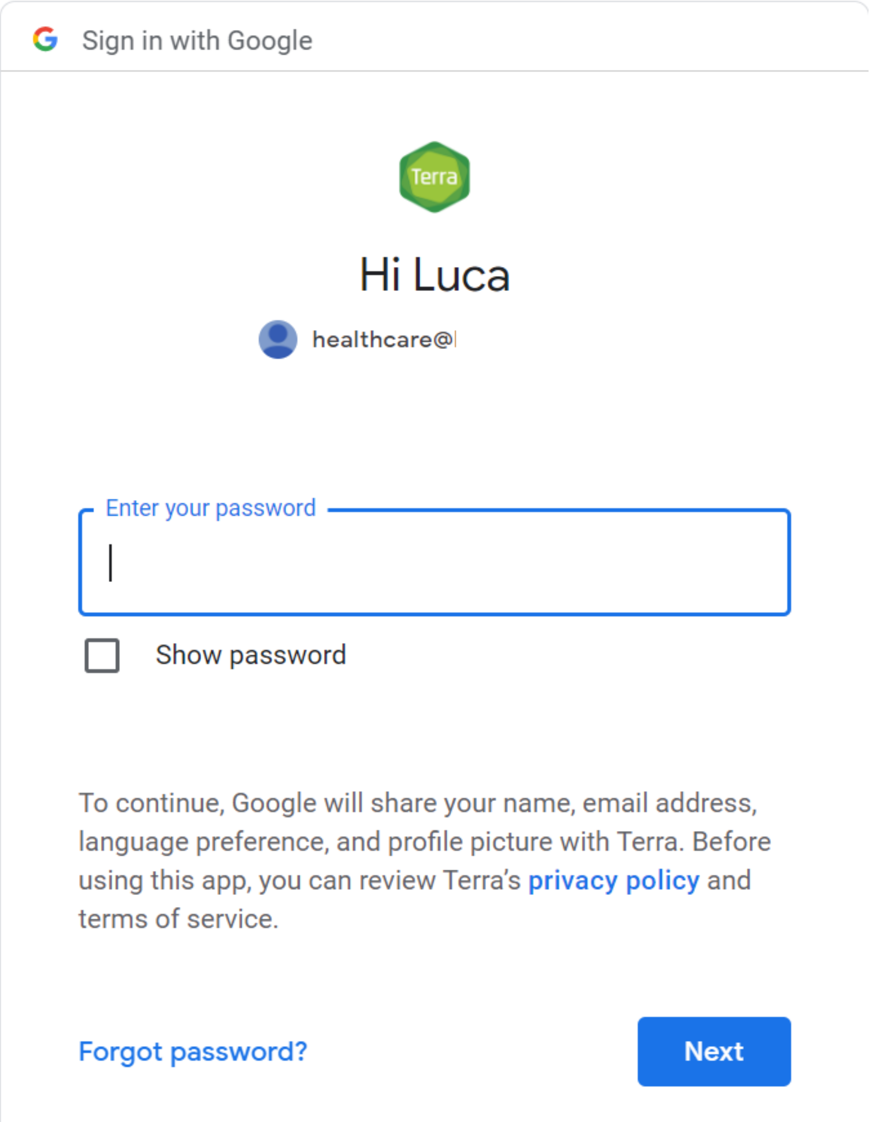
Sign in with Google

Sign in

to continue to Terra

To continue, Google will share your name, email address, language preference, and profile picture with Terra. Before using this app, you can review Terra's privacy policy and terms of service.

Input the password associated with your Gmail account:



Sign in with Google

Hi Luca

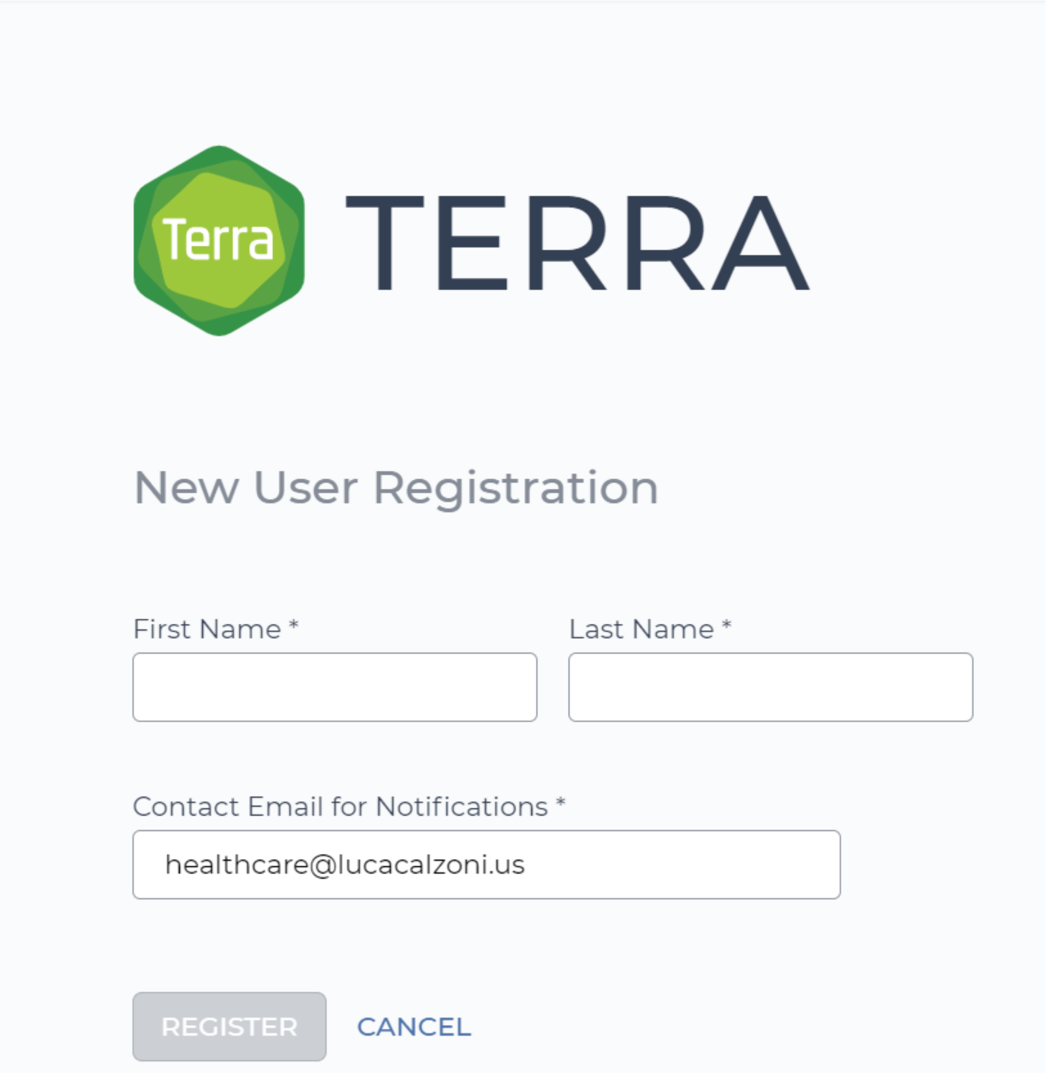
healthcare@

Enter your password

□ Show password

To continue, Google will share your name, email address, language preference, and profile picture with Terra. Before using this app, you can review Terra's **privacy policy** and terms of service.

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



New User Registration

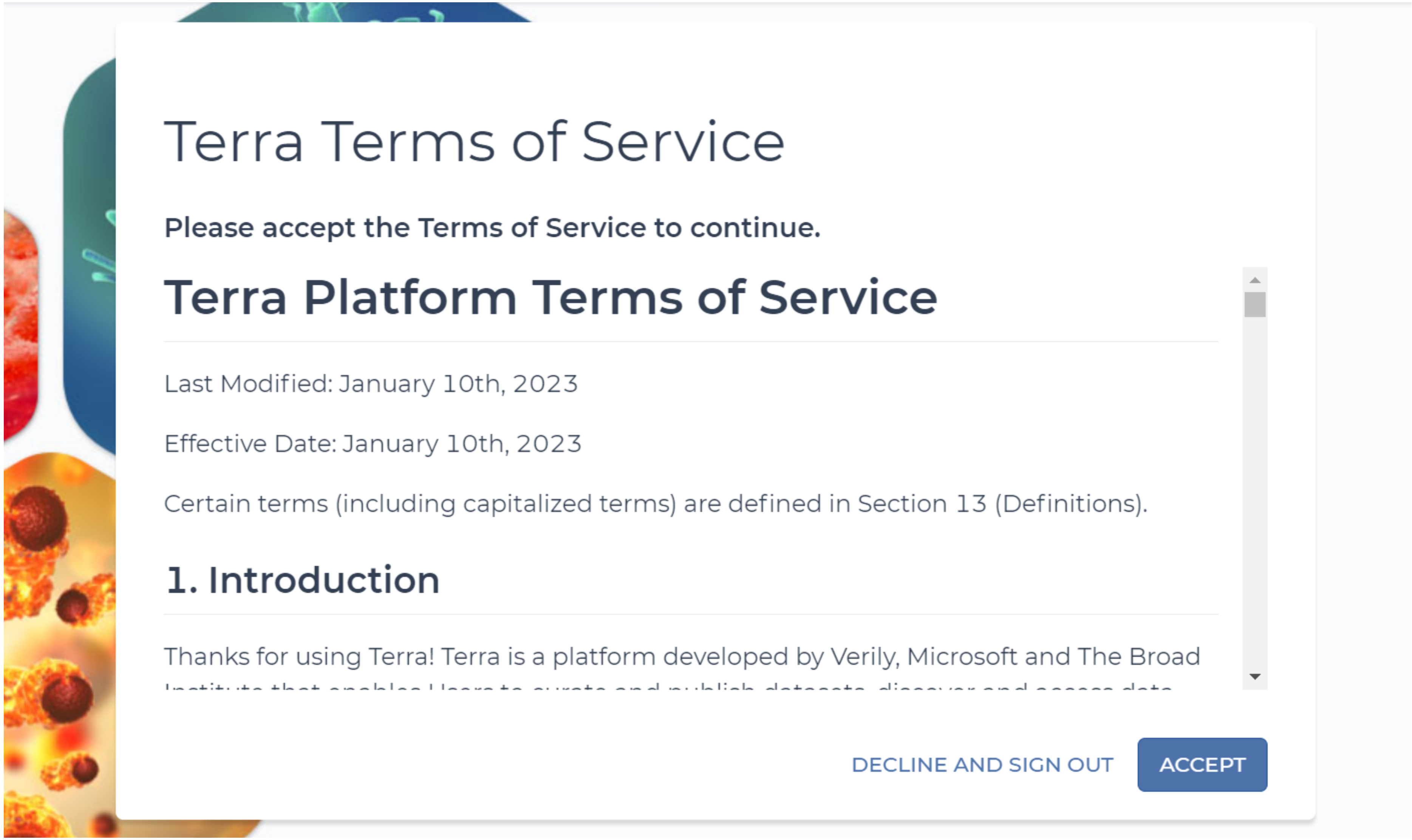
First Name *

Last Name *

Contact Email for Notifications *

healthcare@lucacalzoni.us

Accept the Terra Terms of Service:



Terra Terms of Service

Please accept the Terms of Service to continue.

Terra Platform Terms of Service

Last Modified: January 10th, 2023

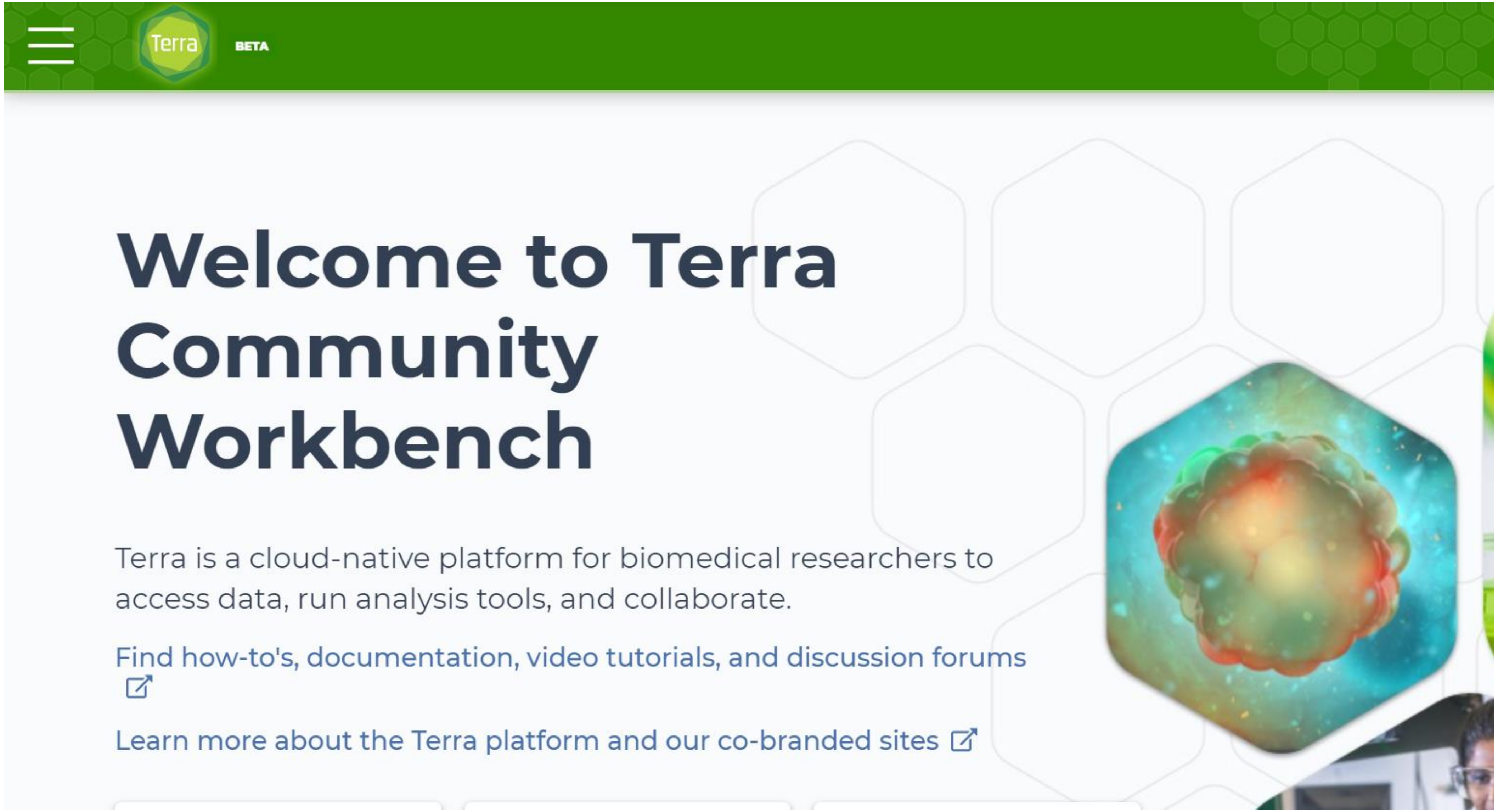
Effective Date: January 10th, 2023

Certain terms (including capitalized terms) are defined in Section 13 (Definitions).

1. Introduction

Thanks for using Terra! Terra is a platform developed by Verily, Microsoft and The Broad Institute that enables Users to curate and publish datasets, discover and access data

You will see this welcome page:



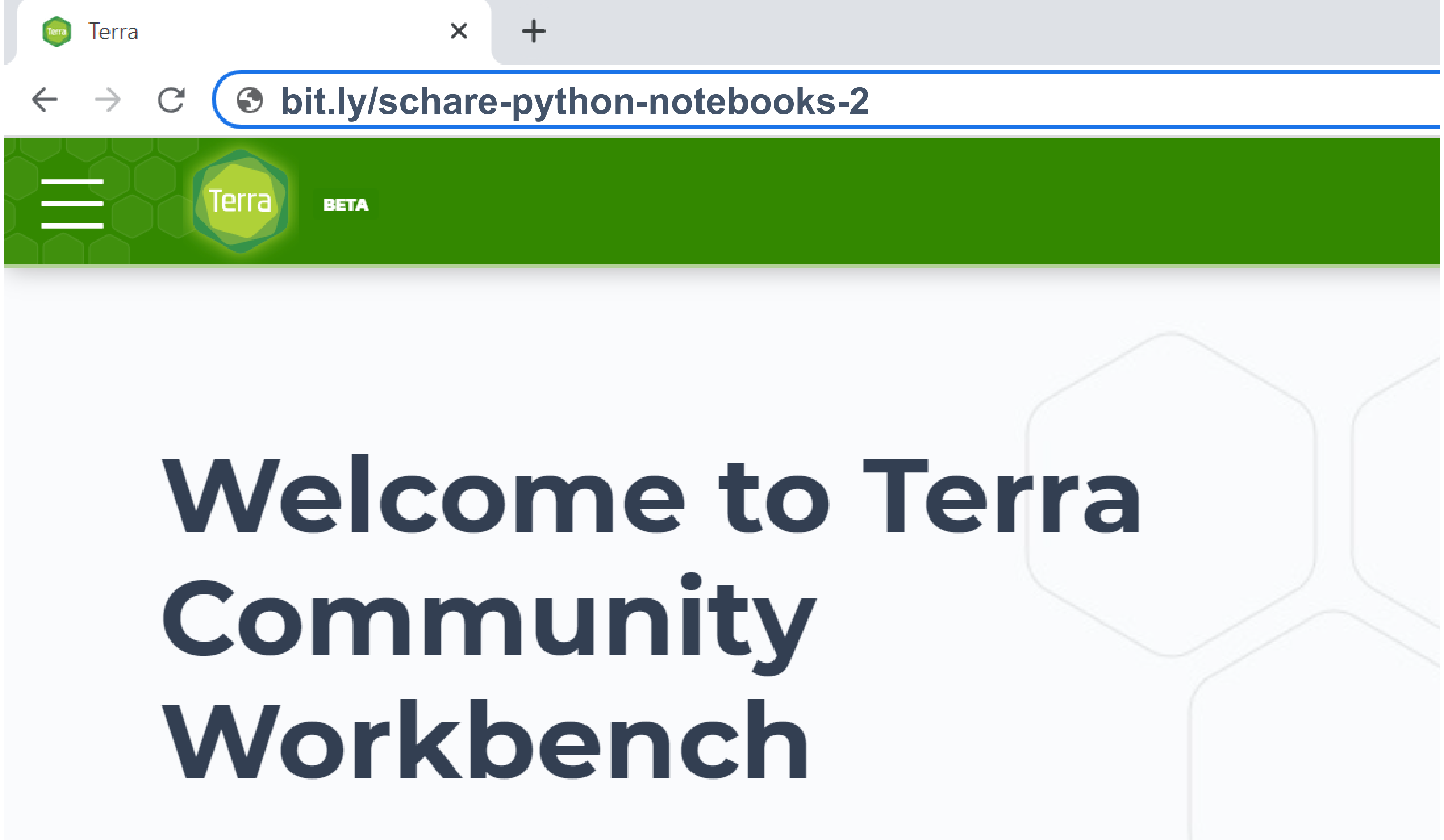
Welcome to Terra Community Workbench

Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate.

Find how-to's, documentation, video tutorials, and discussion forums

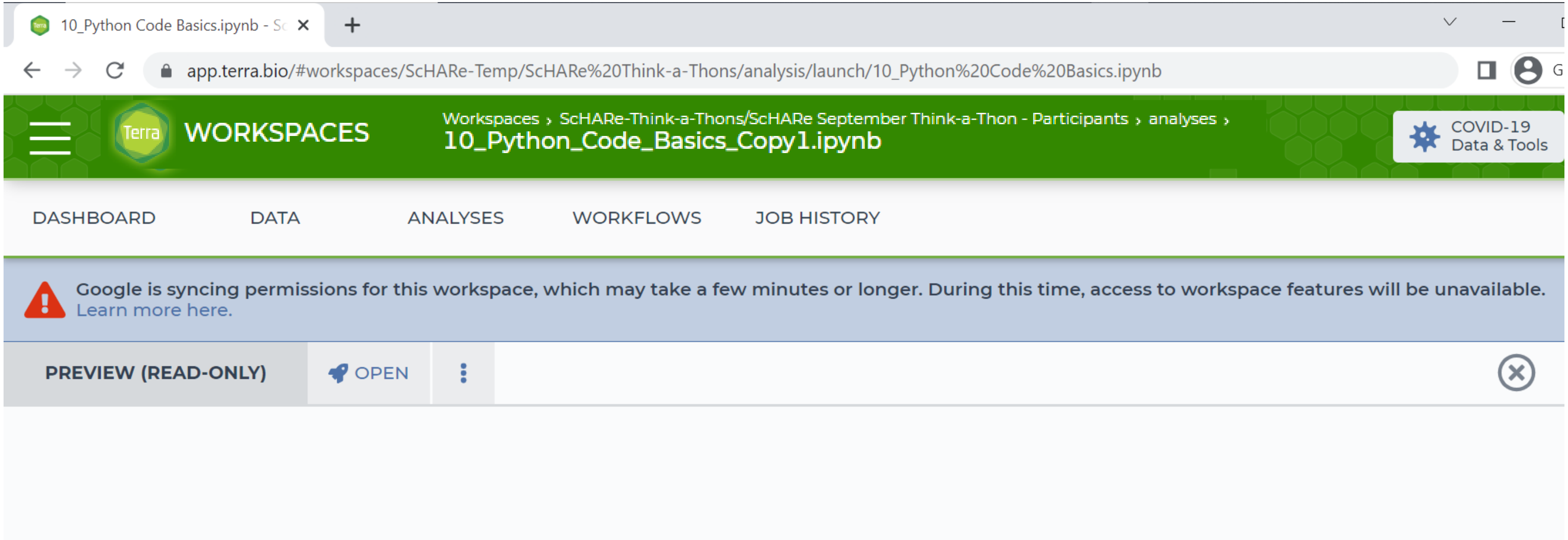
Learn more about the Terra platform and our co-branded sites

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



Welcome to Terra Community Workbench

Newly registered users might see this message:

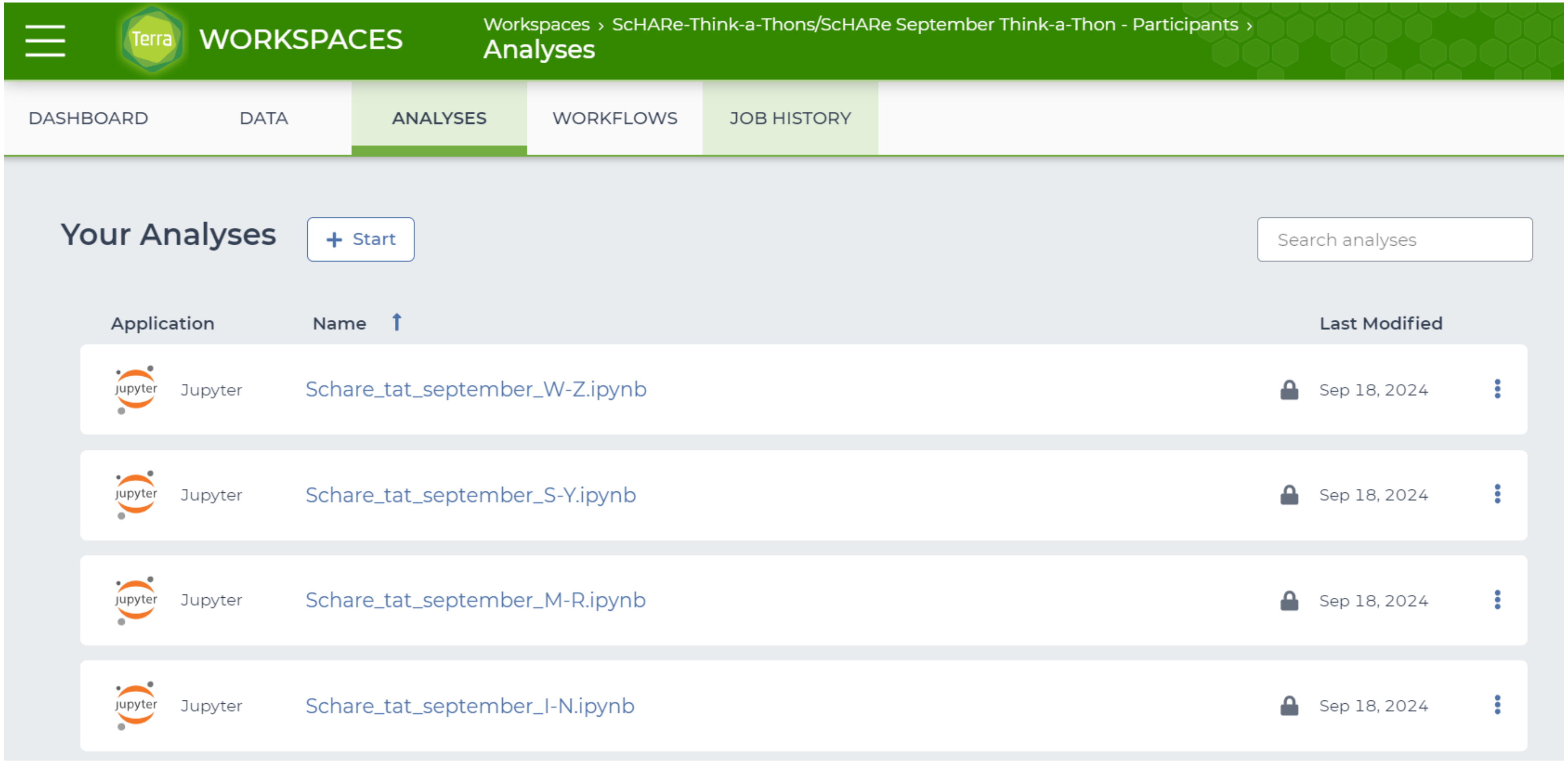


10_Python Code Basics.ipynb - Sc
app.terra.bio/#workspaces/ScHARe-Temp/ScHARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb
Terra
WORKSPACES
Workspaces › ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants › analyses ›
10_Python_Code_Basics_Copy1.ipynb
COVID-19 Data & Tools
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
PREVIEW (READ-ONLY)
OPEN

Google is syncing permissions for this workspace, which may take a few minutes or longer. During this time, access to workspace features will be unavailable. Learn more here.

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

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

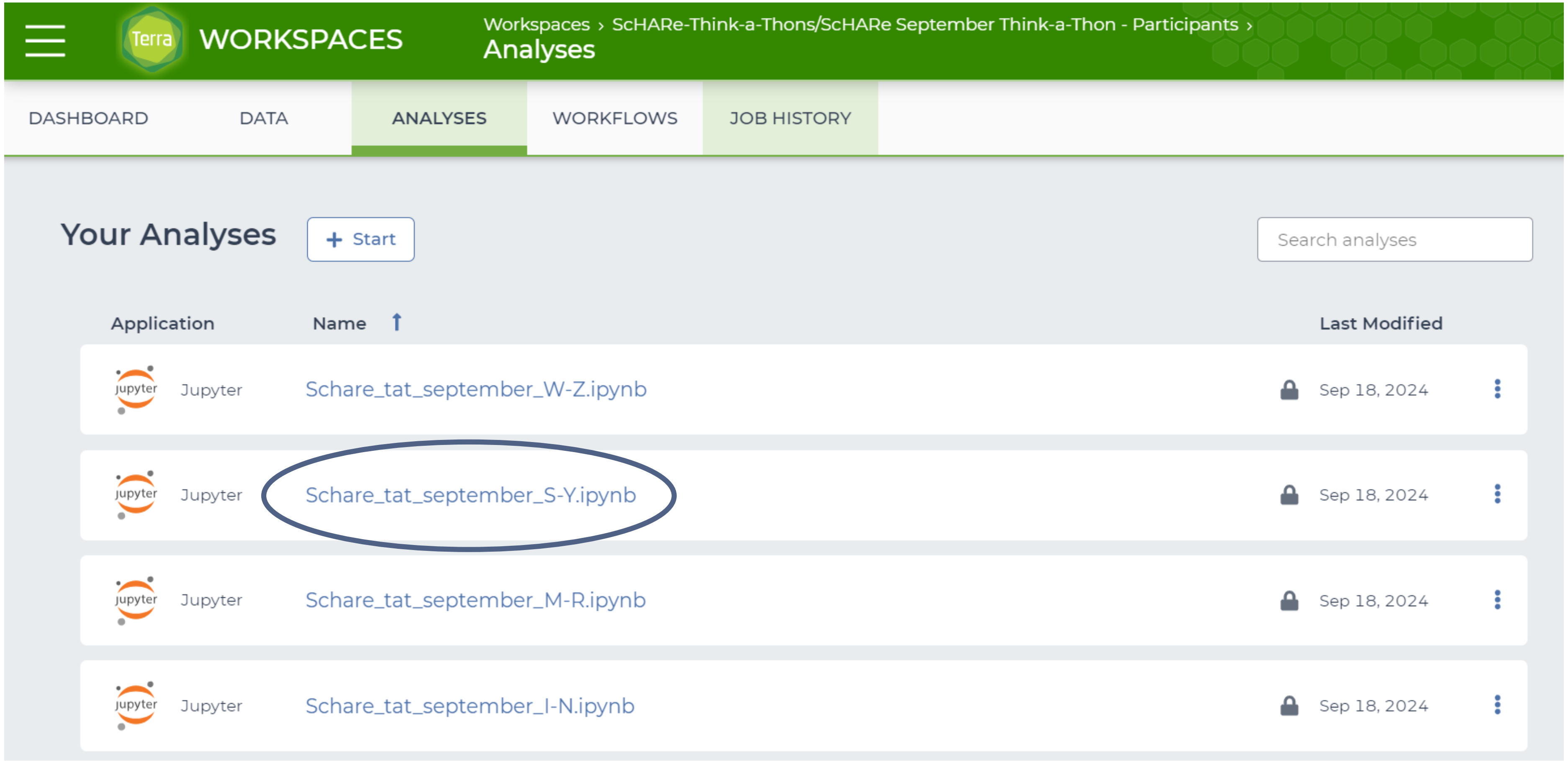


WORKSPACES
Workspaces › ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants ›
Analyses
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY

Your Analyses

Application	Name	Last Modified
Jupyter	Schare_tat_september_W-Z.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_S-Y.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_M-R.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_I-N.ipynb	Sep 18, 2024

Click on the notebook containing your last name initial:



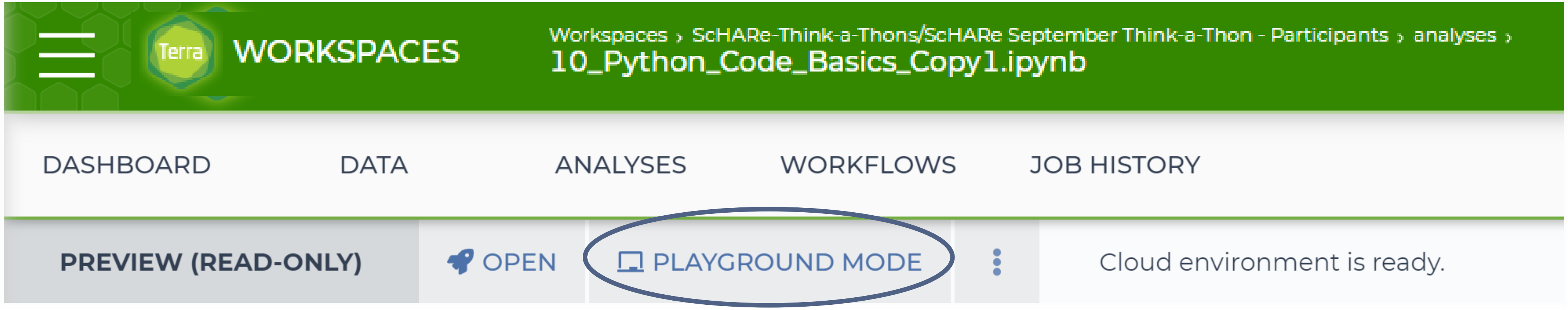
Workspaces › ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants ›
Analyses

Your Analyses

Application	Name	Last Modified
Jupyter	Schare_tat_september_W-Z.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_S-Y.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_M-R.ipynb	Sep 18, 2024
Jupyter	Schare_tat_september_I-N.ipynb	Sep 18, 2024

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

Do you see a Playground mode button?

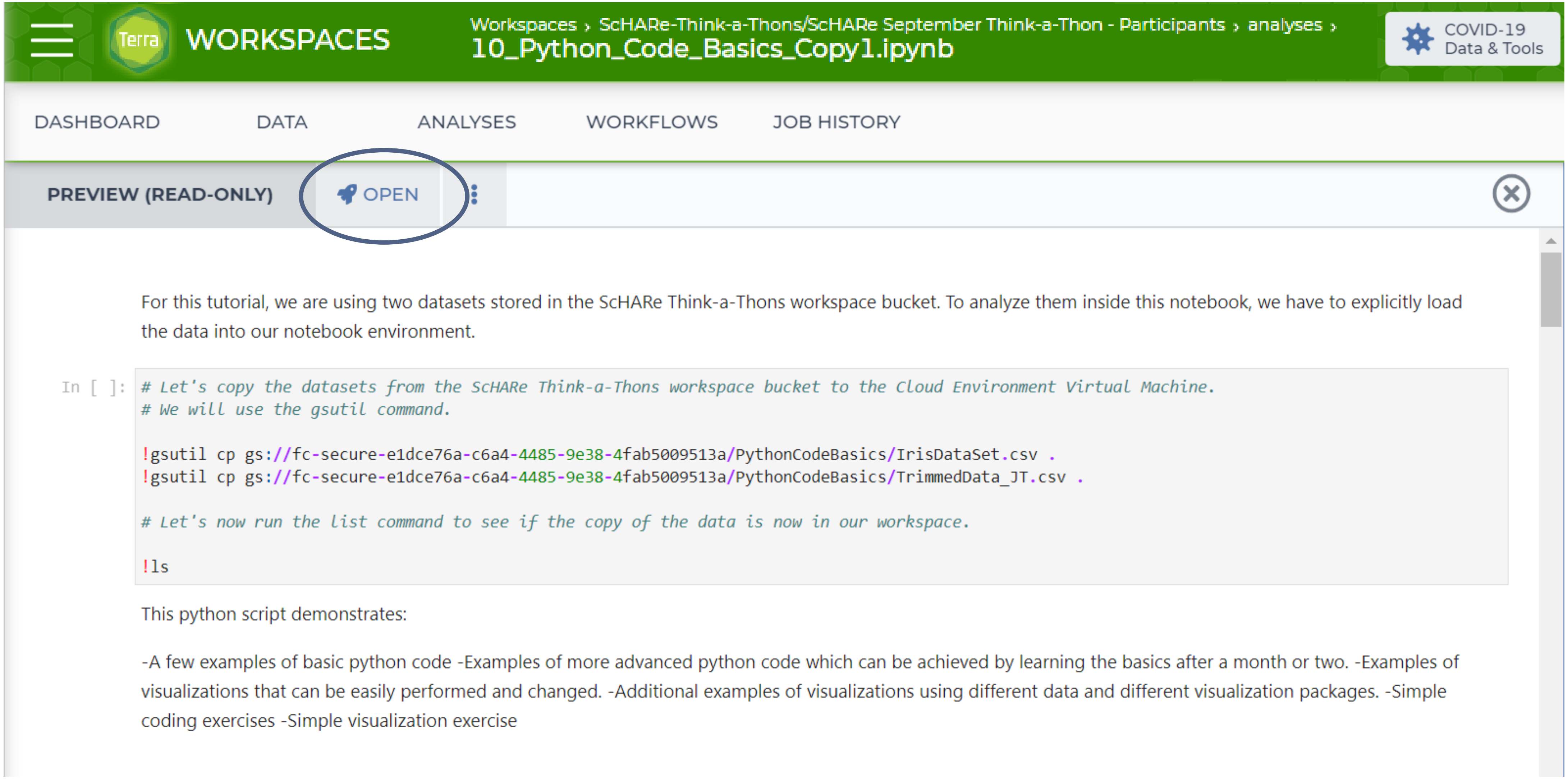


Workspaces › ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants › analyses ›
10_Python_Code_Basics_Copy1.ipynb

Cloud environment is ready.

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:



Terra
WORKSPACES
Workspaces › ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants › analyses ›
10_Python_Code_Basics_Copy1.ipynb
COVID-19 Data & Tools
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
PREVIEW (READ-ONLY)
OPEN

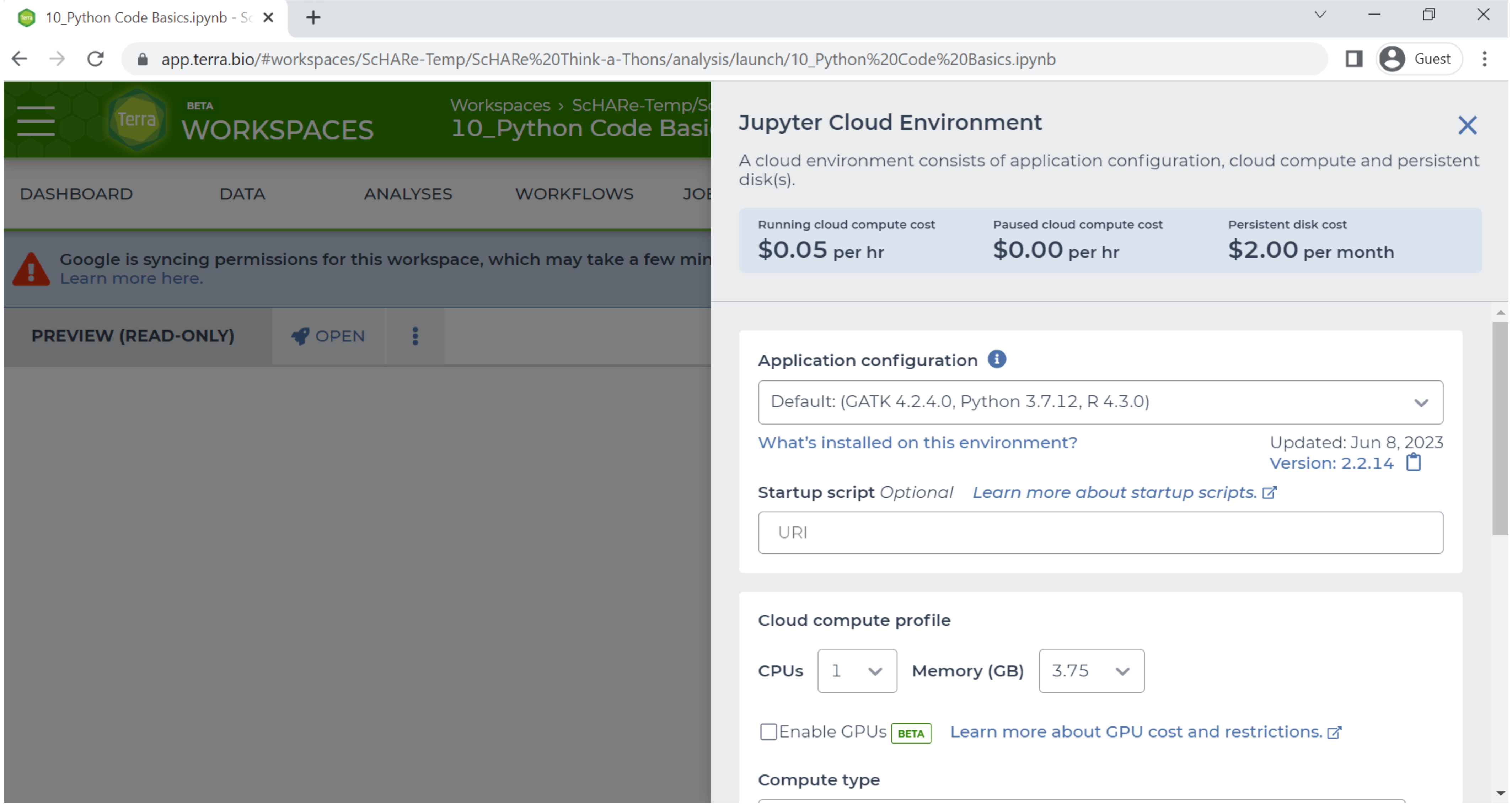
For this tutorial, we are using two datasets stored in the ScHARe Think-a-Thons workspace bucket. To analyze them inside this notebook, we have to explicitly load the data into our notebook environment.

```
In [ ]: # Let's copy the datasets from the ScHARe Think-a-Thons workspace bucket to the Cloud Environment Virtual Machine.
        # We will use the gsutil command.

        !gsutil cp gs://fc-secure-e1dce76a-c6a4-4485-9e38-4fab5009513a/PythonCodeBasics/IrisDataSet.csv .
        !gsutil cp gs://fc-secure-e1dce76a-c6a4-4485-9e38-4fab5009513a/PythonCodeBasics/TrimmedData_JT.csv .

        # Let's now run the list command to see if the copy of the data is now in our workspace.

        !ls
```

This python script demonstrates:

-A few examples of basic python code -Examples of more advanced python code which can be achieved by learning the basics after a month or two. -Examples of visualizations that can be easily performed and changed. -Additional examples of visualizations using different data and different visualization packages. -Simple coding exercises -Simple visualization exercise

Configure your virtual computer – accept the default values:



10_Python Code Basics.ipynb - Sc
app.terra.bio/#workspaces/ScHARe-Temp/ScHARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb
Guest
Terra
BETA
WORKSPACES
Workspaces › ScHARe-Temp/Sc
10_Python Code Basi
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOE
PREVIEW (READ-ONLY)
OPEN

Google is syncing permissions for this workspace, which may take a few min
Learn more here.

Jupyter Cloud Environment

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

Running cloud compute cost	Paused cloud compute cost	Persistent disk cost
$0.05 per hr	$0.00 per hr	$2.00 per month

Application configuration

Default: (GATK 4.2.4.0, Python 3.7.12, R 4.3.0)

What's installed on this environment?

Updated: Jun 8, 2023
Version: 2.2.14

Startup script *Optional* *Learn more about startup scripts.*

URI

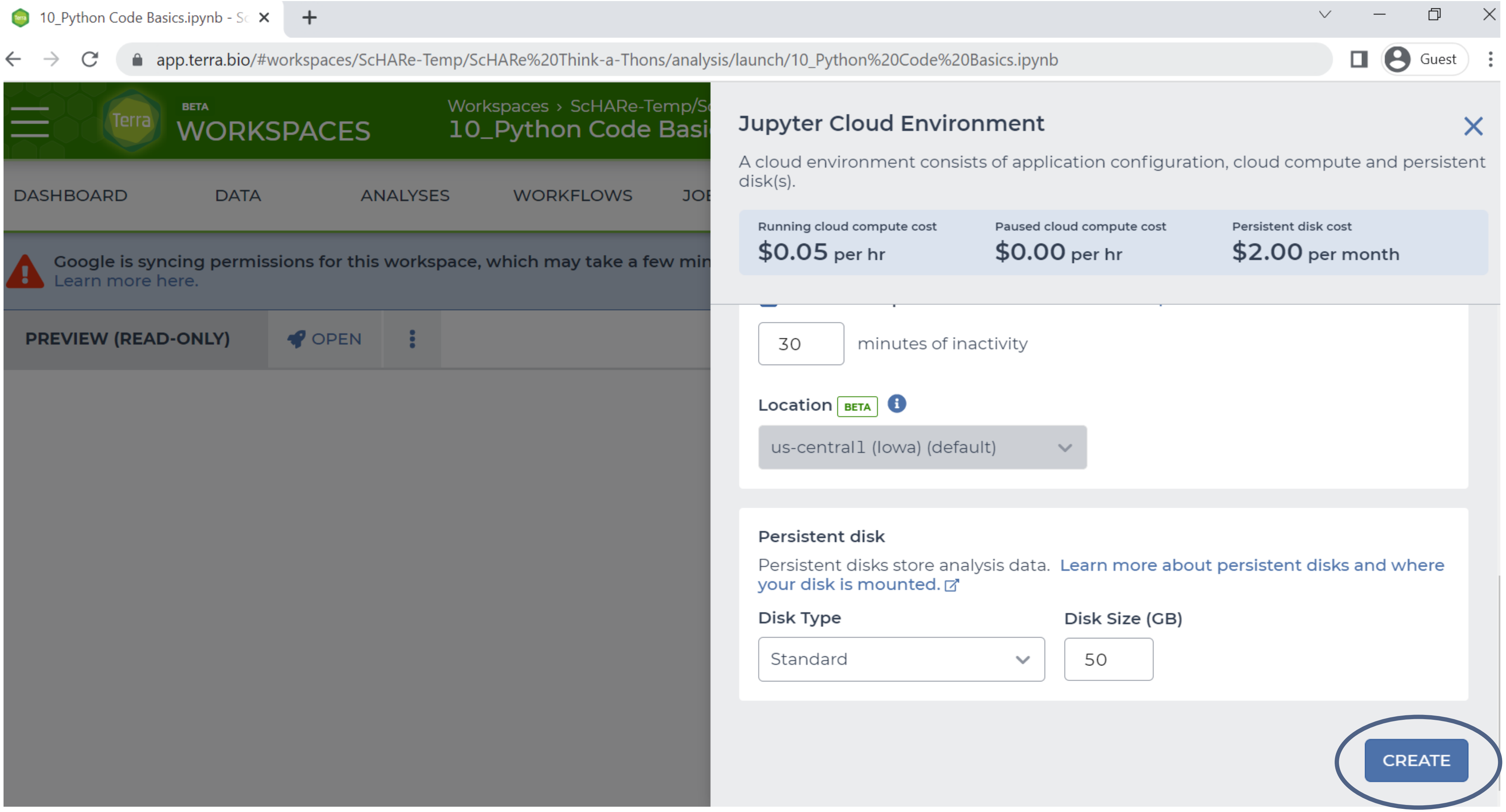
Cloud compute profile

CPUs 1 Memory (GB) 3.75

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

Compute type

Click on Create below:



10_Python Code Basics.ipynb - Sc
app.terra.bio/#workspaces/ScHARe-Temp/ScHARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb
Guest
Terra
BETA
WORKSPACES
Workspaces › ScHARe-Temp/Sc
10_Python Code Basi
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOE
Google is syncing permissions for this workspace, which may take a few min
Learn more here.
PREVIEW (READ-ONLY)
OPEN

Jupyter Cloud Environment

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

Running cloud compute cost	Paused cloud compute cost	Persistent disk cost
$0.05 per hr	$0.00 per hr	$2.00 per month

30 minutes of inactivity

Location BETA

us-central1 (Iowa) (default)

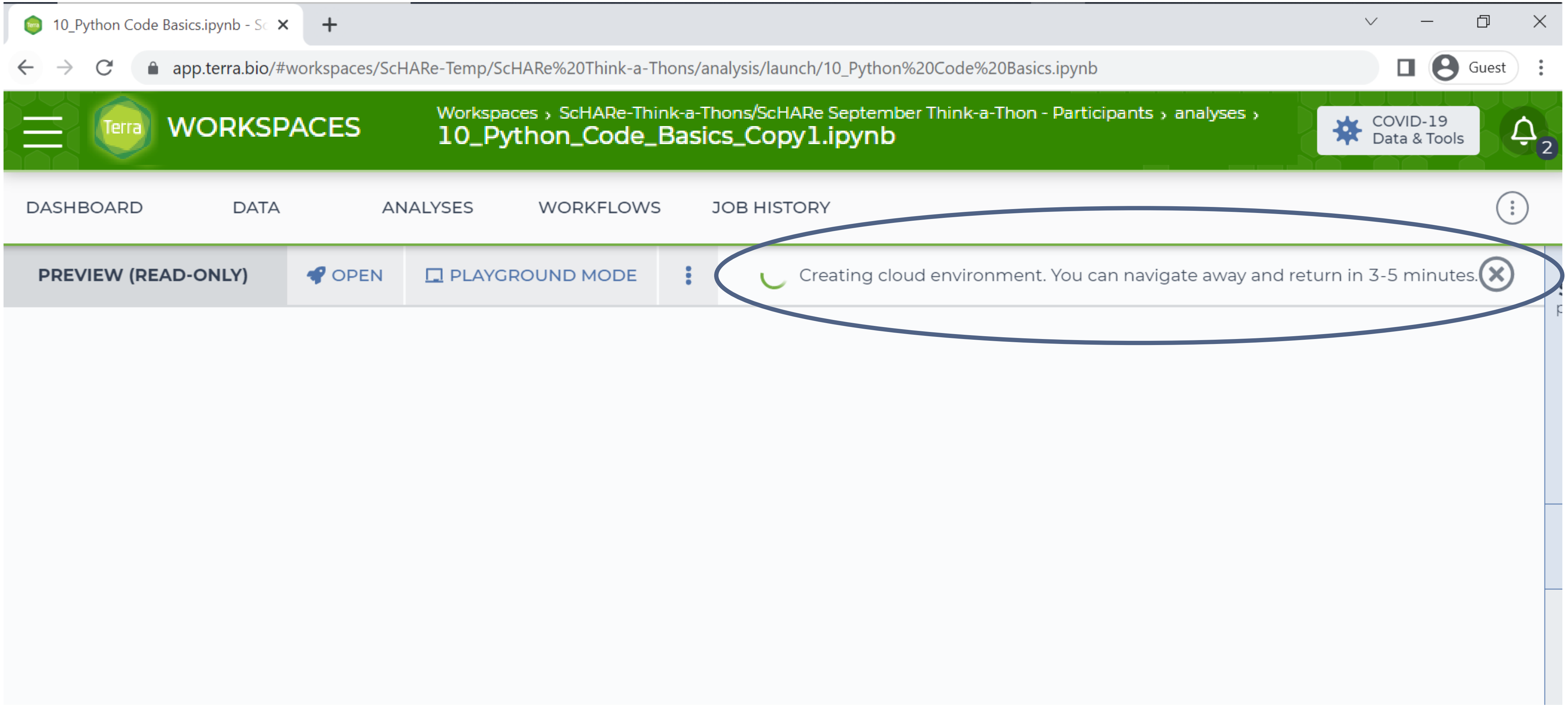
Persistent disk

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

Disk Type: Standard

Disk Size (GB): 50

It will take some time…



10_Python Code Basics.ipynb - Sc

app.terra.bio/#workspaces/ScHARe-Temp/ScHARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb

Guest

WORKSPACES

Workspaces › ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants › analyses ›

10_Python_Code_Basics_Copy1.ipynb

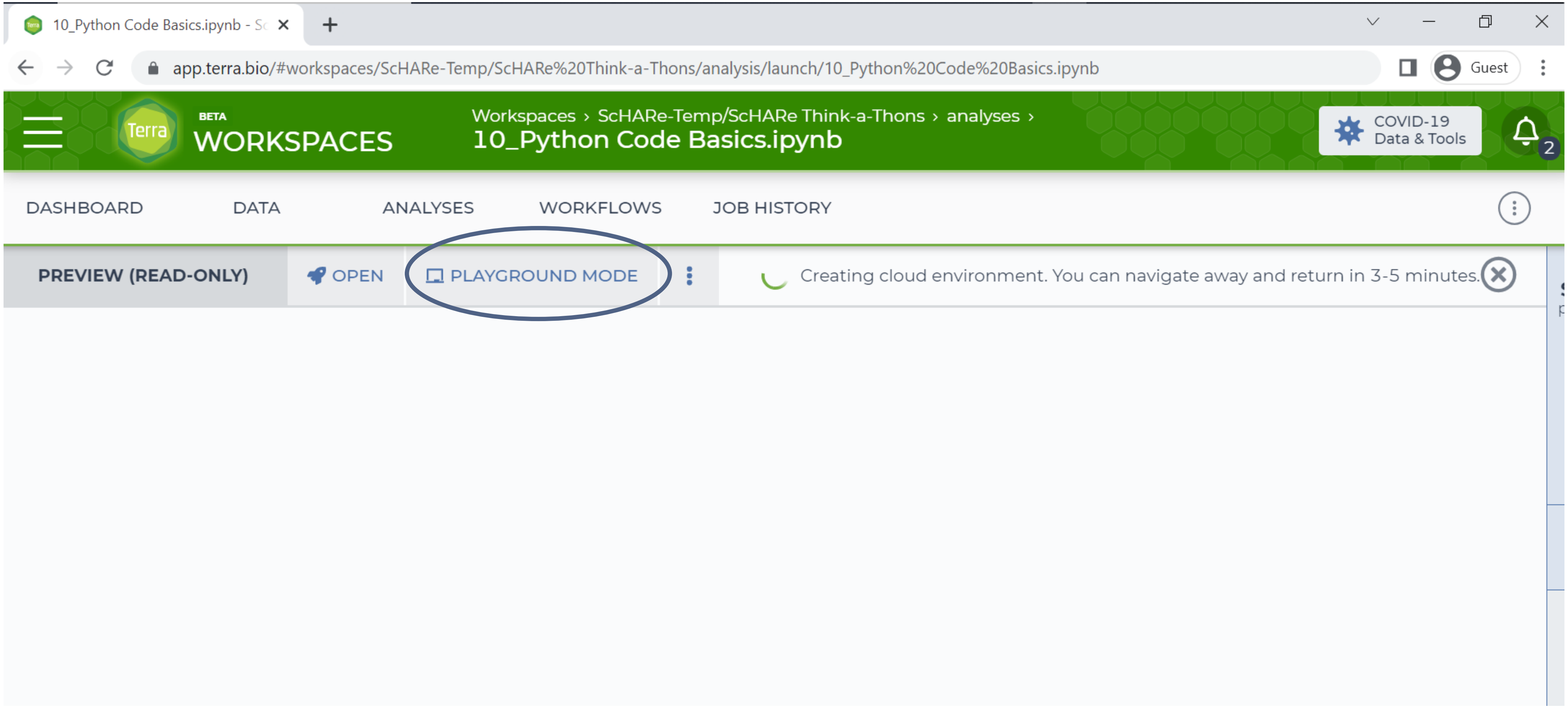
COVID-19 Data & Tools

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

PREVIEW (READ-ONLY) OPEN PLAYGROUND MODE

Creating cloud environment. You can navigate away and return in 3-5 minutes.

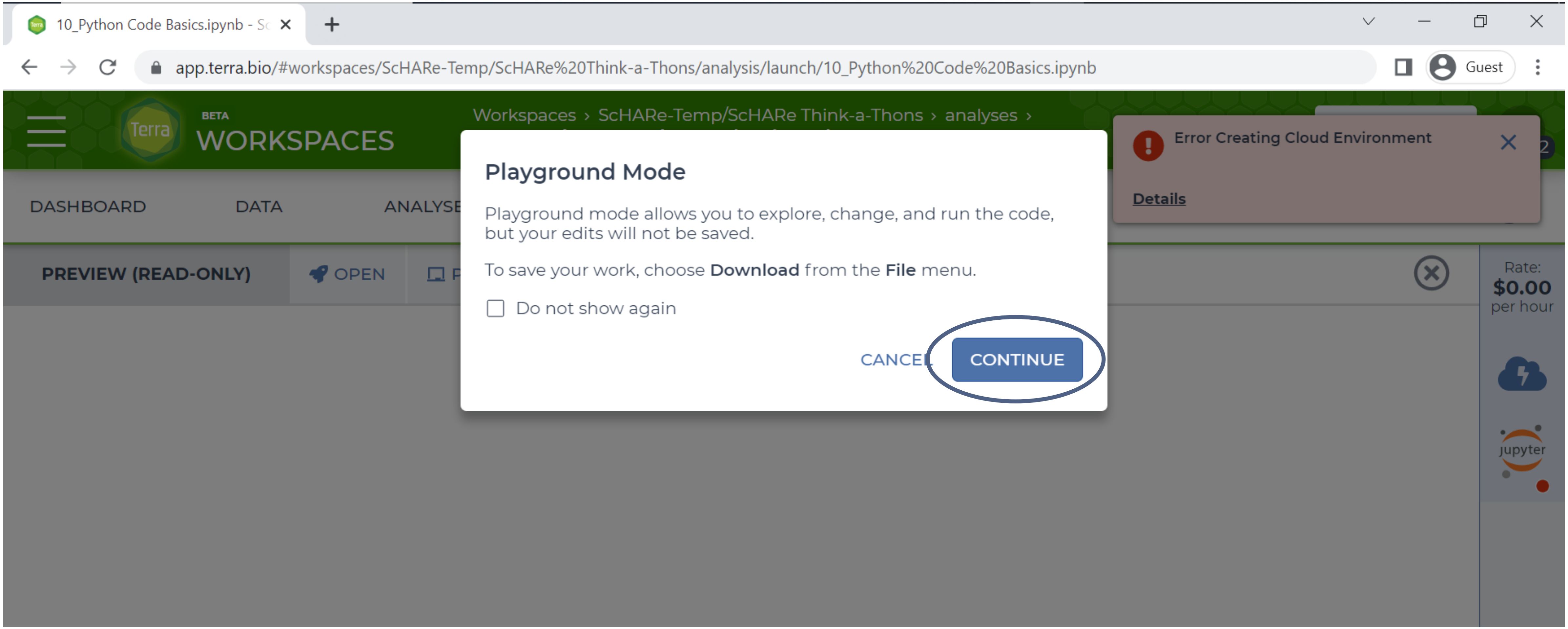
When the system is ready, click on Playground mode:



10_Python Code Basics.ipynb - Sc
app.terra.bio/#workspaces/ScHARe-Temp/ScHARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb
Guest
Terra
BETA
WORKSPACES
Workspaces › ScHARe-Temp/ScHARe Think-a-Thons › analyses ›
10_Python Code Basics.ipynb
COVID-19 Data & Tools
2
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
PREVIEW (READ-ONLY)
OPEN
PLAYGROUND MODE

Creating cloud environment. You can navigate away and return in 3-5 minutes.

Click on Continue:



10_Python Code Basics.ipynb - Sc
app.terra.bio/#workspaces/ScHARe-Temp/ScHARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb
Guest
Terra
BETA
WORKSPACES
Workspaces › ScHARe-Temp/ScHARe Think-a-Thons › analyses ›
DASHBOARD
DATA
ANALYSE
PREVIEW (READ-ONLY)
OPEN
Rate:
$0.00
per hour
jupyter

Error Creating Cloud Environment

Details

Playground Mode

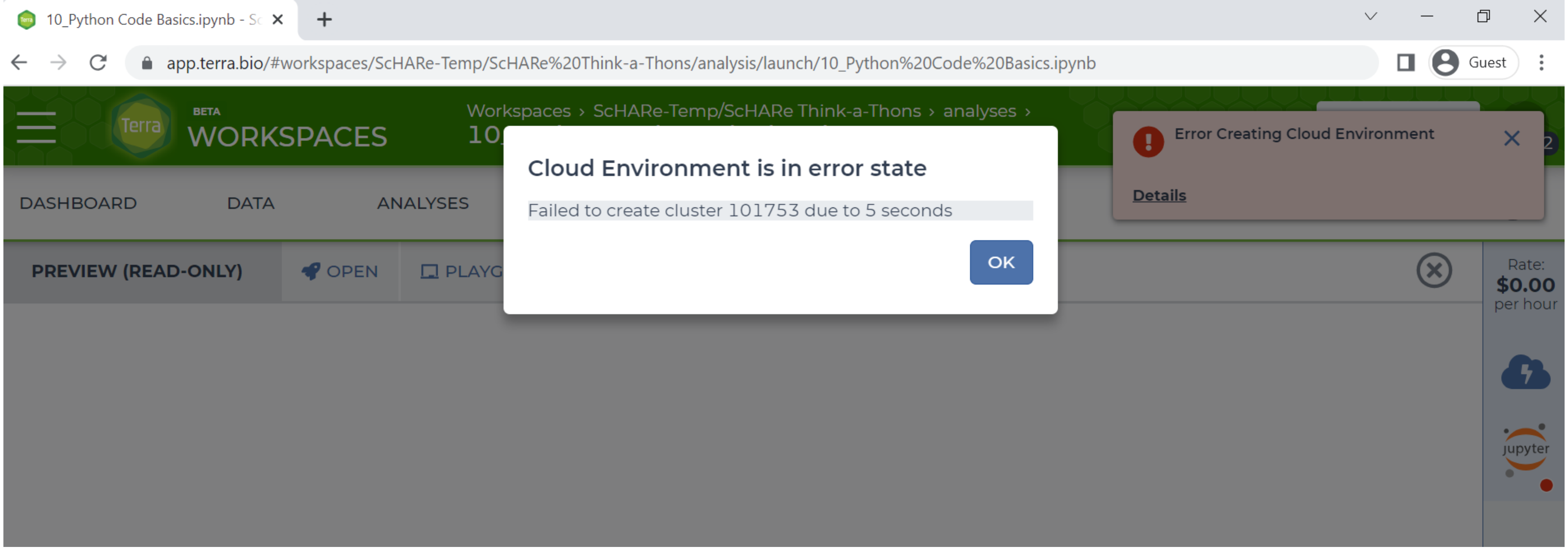
Playground mode allows you to explore, change, and run the code, but your edits will not be saved.

To save your work, choose **Download** from the the **File** menu.

□ Do not show again

CANCEL CONTINUE

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

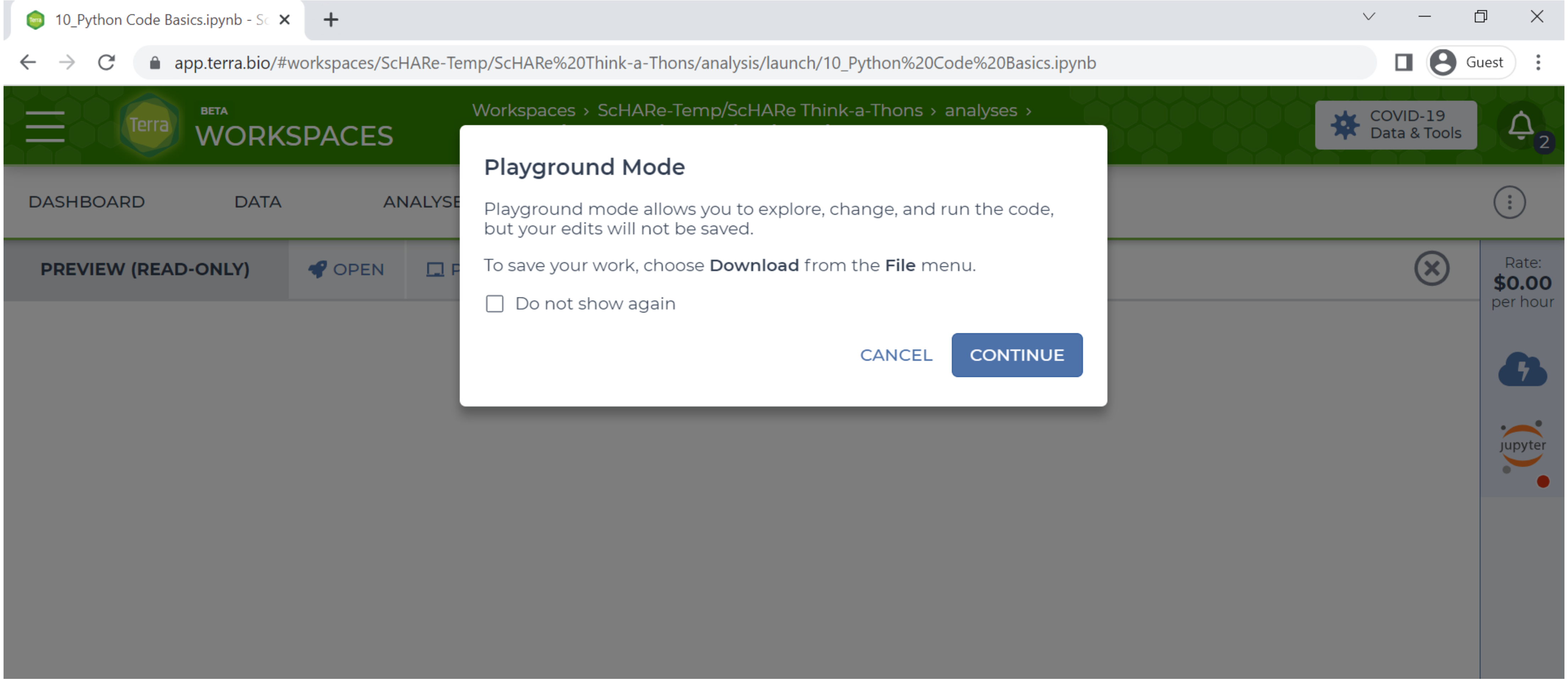

10_Python Code Basics.ipynb - Sc
app.terra.bio/#workspaces/ScHARe-Temp/ScHARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb
Guest
Terra
BETA
WORKSPACES
Workspaces › ScHARe-Temp/ScHARe Think-a-Thons › analyses ›
10
DASHBOARD
DATA
ANALYSES
PREVIEW (READ-ONLY)
OPEN
PLAYG
Rate:
$0.00
per hour
jupyter
Details
OK

Error Creating Cloud Environment

Cloud Environment is in error state

Failed to create cluster 101753 due to 5 seconds

If all goes well, you will see this:


10_Python Code Basics.ipynb - Sc
app.terra.bio/#workspaces/ScHARe-Temp/ScHARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb
Guest
Terra
BETA
WORKSPACES
Workspaces › ScHARe-Temp/ScHARe Think-a-Thons › analyses ›
COVID-19
Data & Tools
2
DASHBOARD
DATA
ANALYSE
PREVIEW (READ-ONLY)
OPEN
Rate:
$0.00
per hour
jupyter

Playground Mode

Playground mode allows you to explore, change, and run the code, but your edits will not be saved.

To save your work, choose **Download** from the **File** menu.

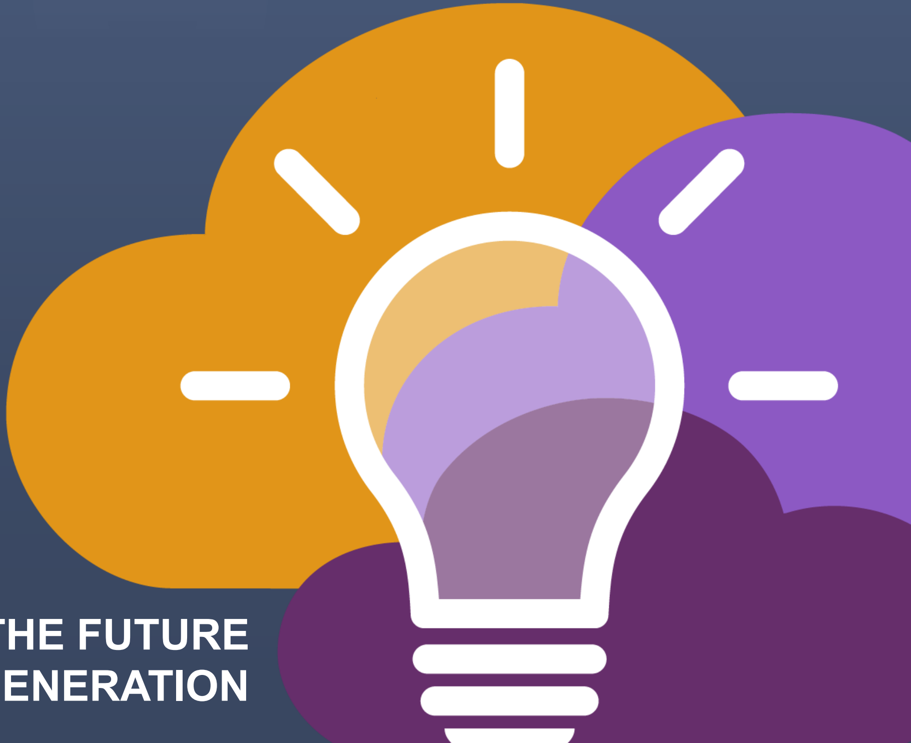
☐ Do not show again

CANCEL   CONTINUE

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?

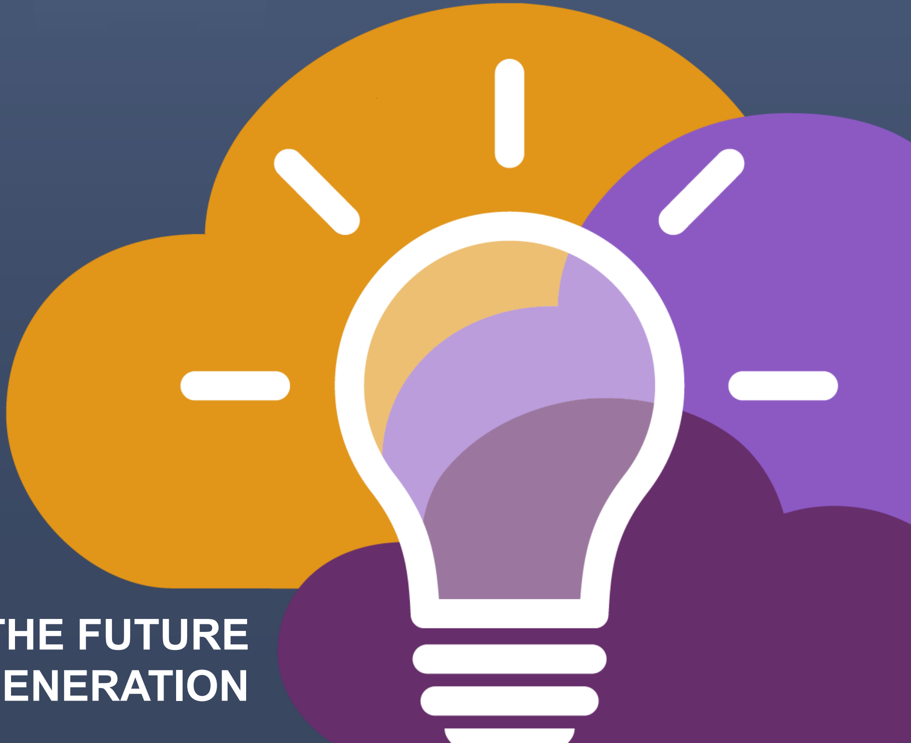
Python is the **perfect** programming language **for people without any coding experience**, as it has a simple syntax and is very accessible to beginners

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.

NIH 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

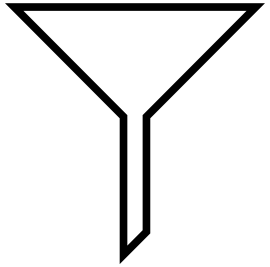
Establish quality control standards:

- 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

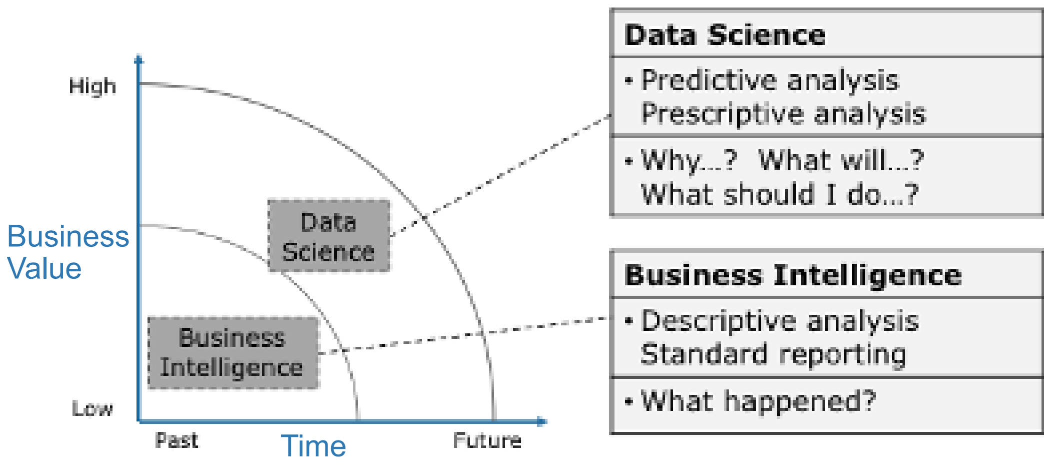
☐ Saving data as images

☐ 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.

Business Intelligence versus Data Science



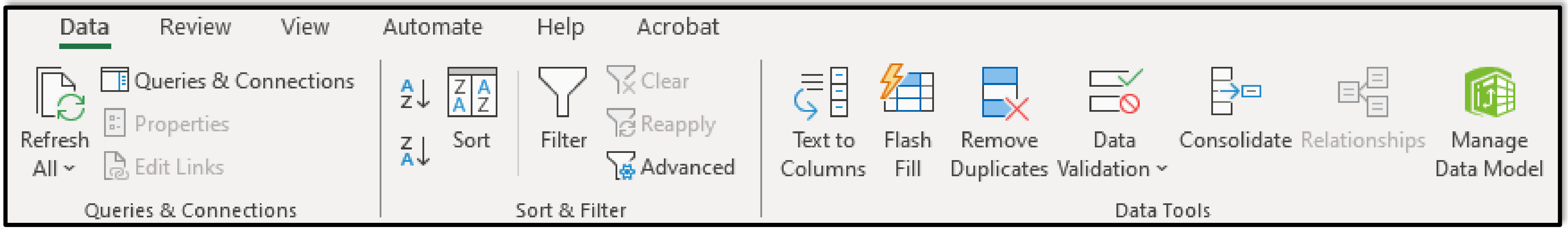
Data Science
- Predictive analysis
  Prescriptive analysis
- Why...? What will...?
  What should I do...?

Business Intelligence
- Descriptive analysis
  Standard reporting
- What happened?

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

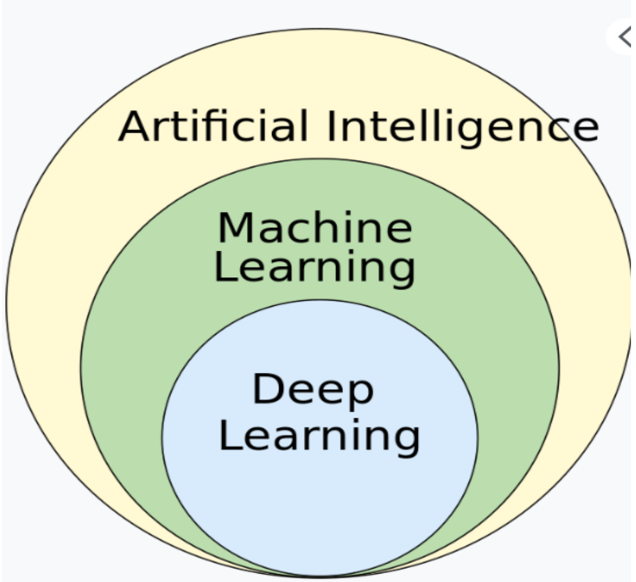
Visualizations

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

Machine Learning

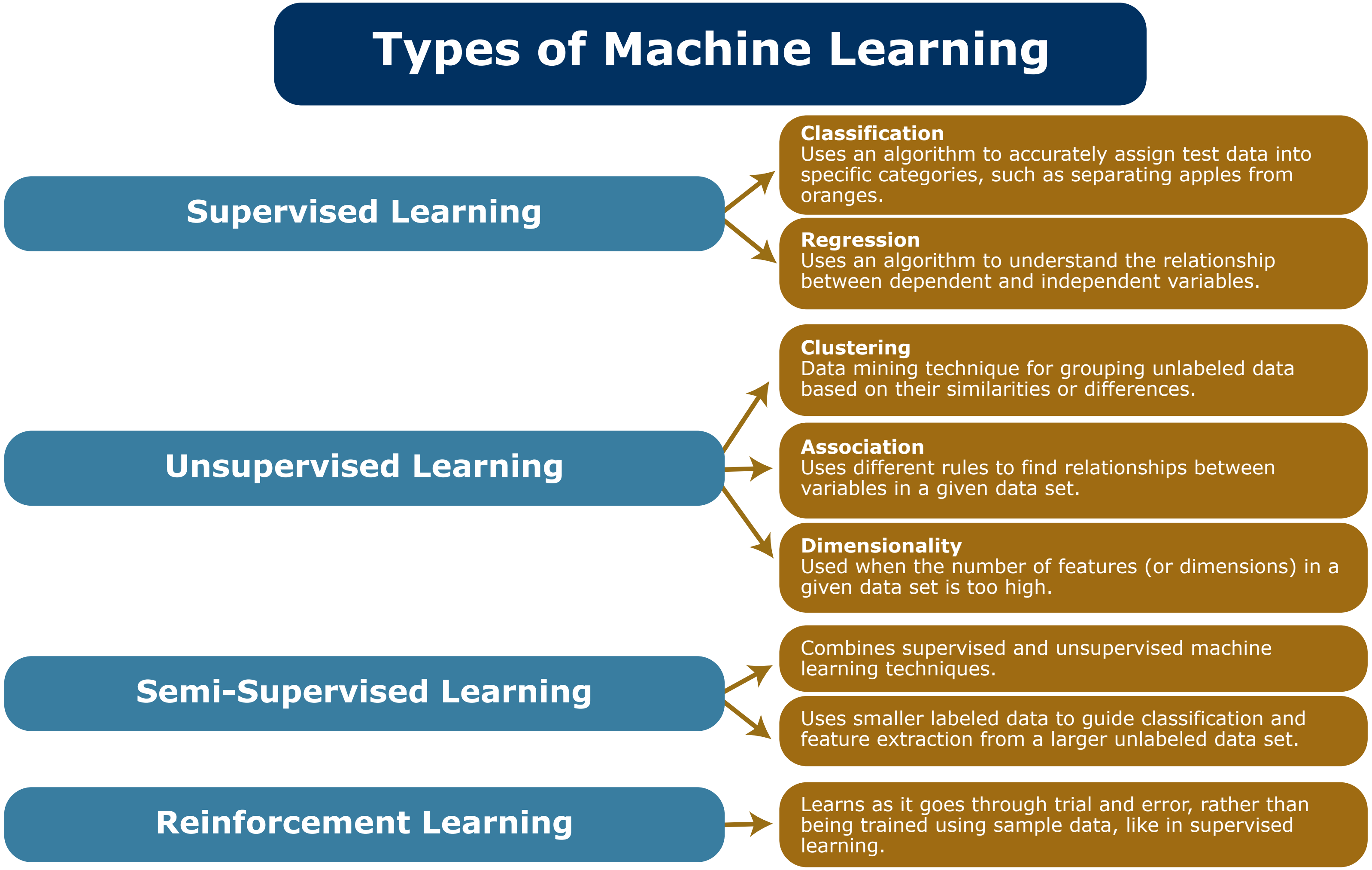
Machine Learning and Artificial Intelligence (AI)

- 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/

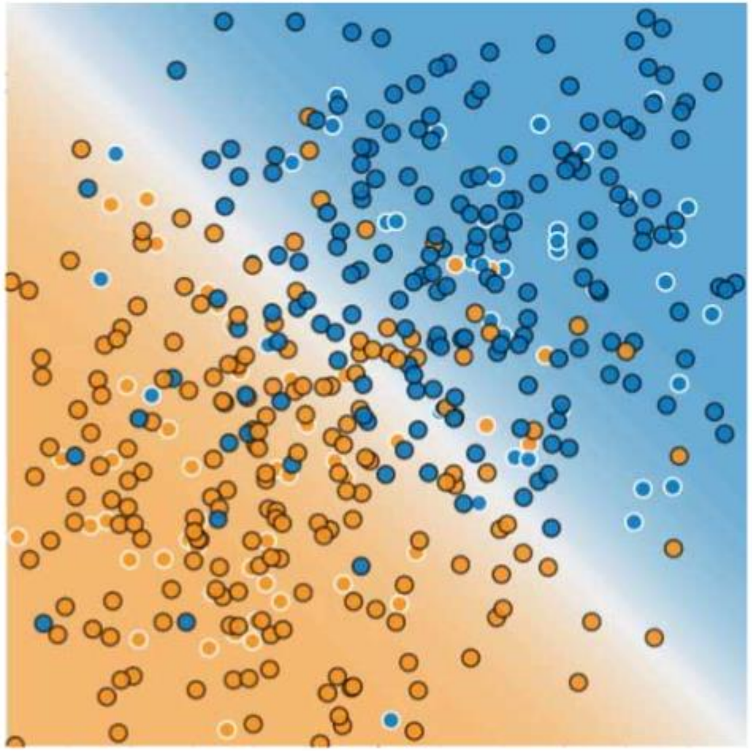
Types of Machine Learning – Reinforcement

- 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

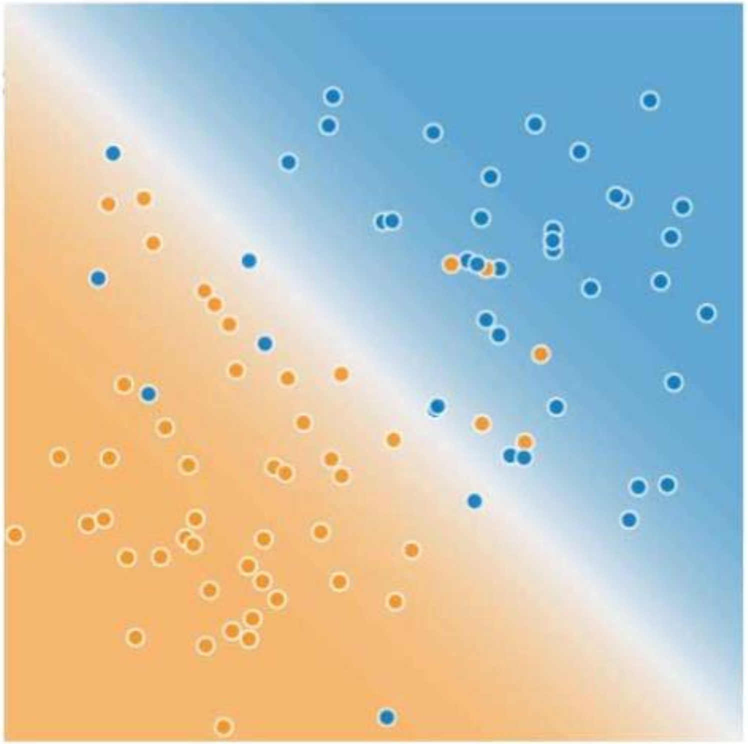
https://bootcamp.berkeley.edu/blog/how-does-machine-learning-work/

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

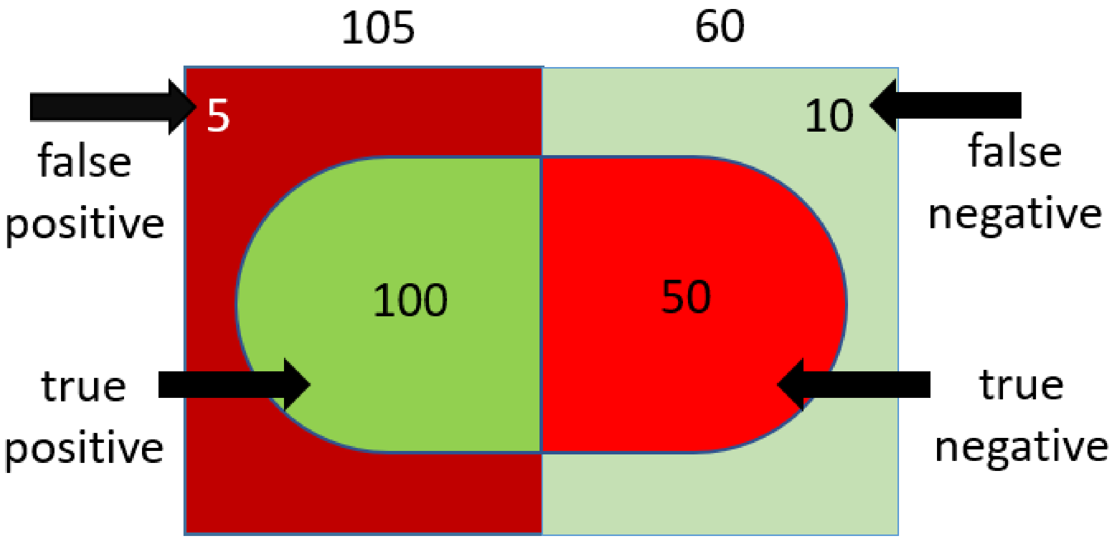
Evaluating Machine Learning Performance

		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)*

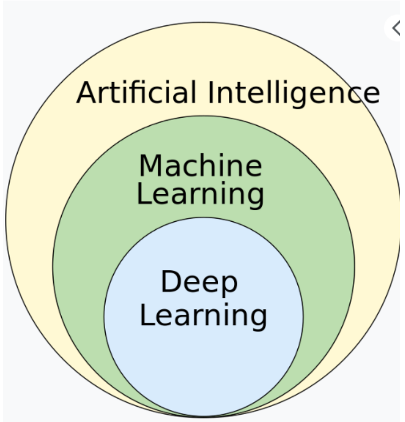
  Or = TP/TP + ½(FP+FN)

  = 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



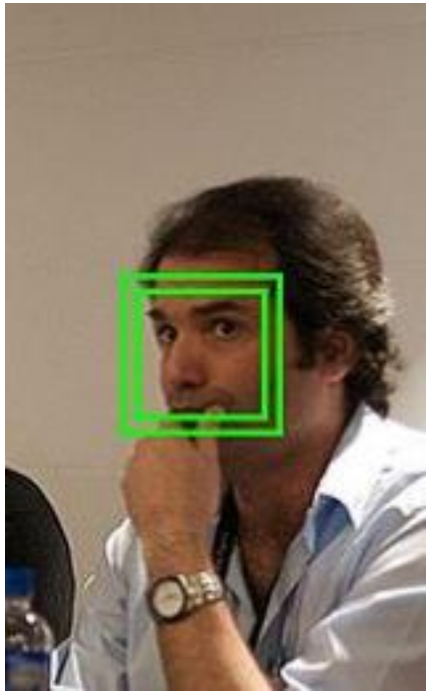
Deep Learning (Cont’d)

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

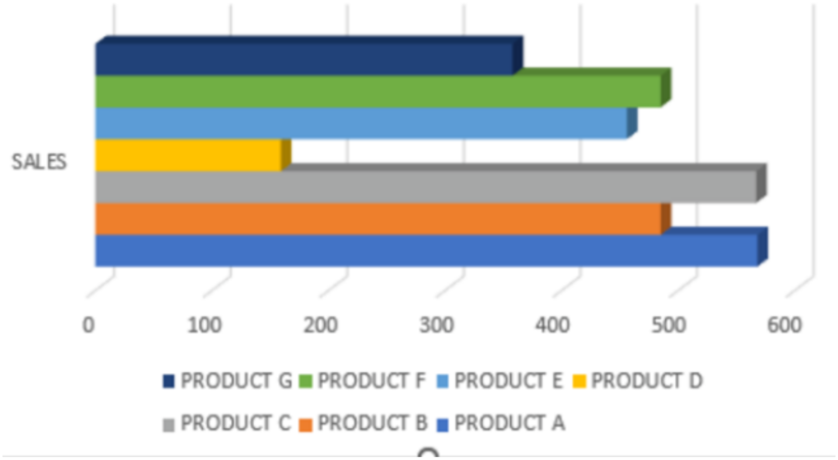
Natural Language Processing (NLP) Benefits (Cont’d)

- **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