

The logo for ScHARe, featuring the text "ScHARe" in a bold, dark blue font inside a white circle.

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	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
R	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Cloud computing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Terra	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Health disparities research	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Health outcomes research	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Algorithmic bias mitigation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Interest poll

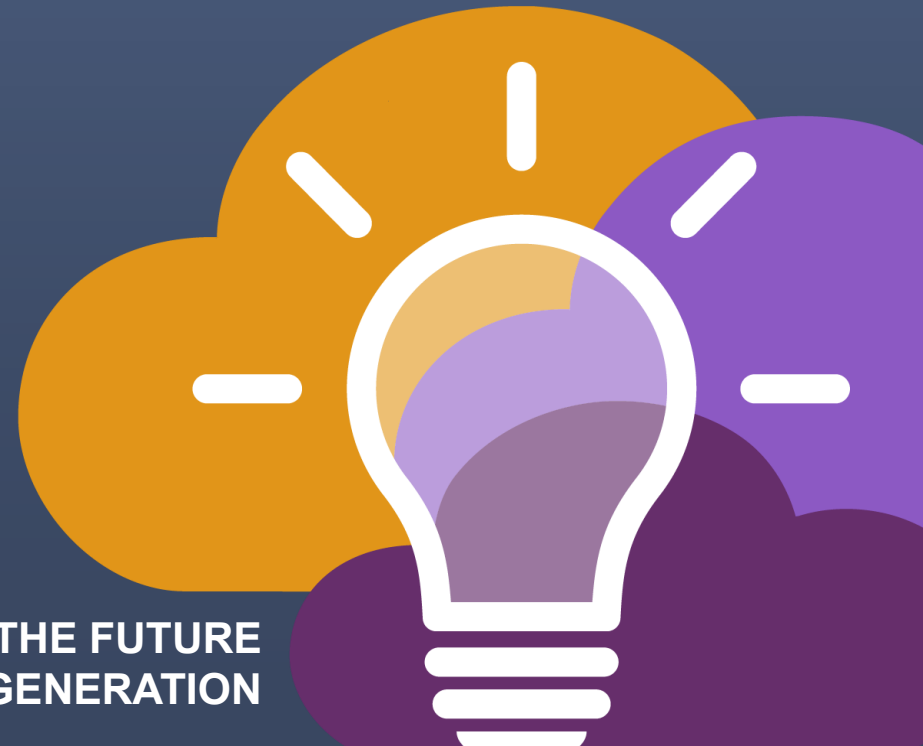
I am interested in (check all that apply):

- 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 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

ScHARe



nimhd.nih.gov/schare



ScHARe



Google Platform Terra Interface

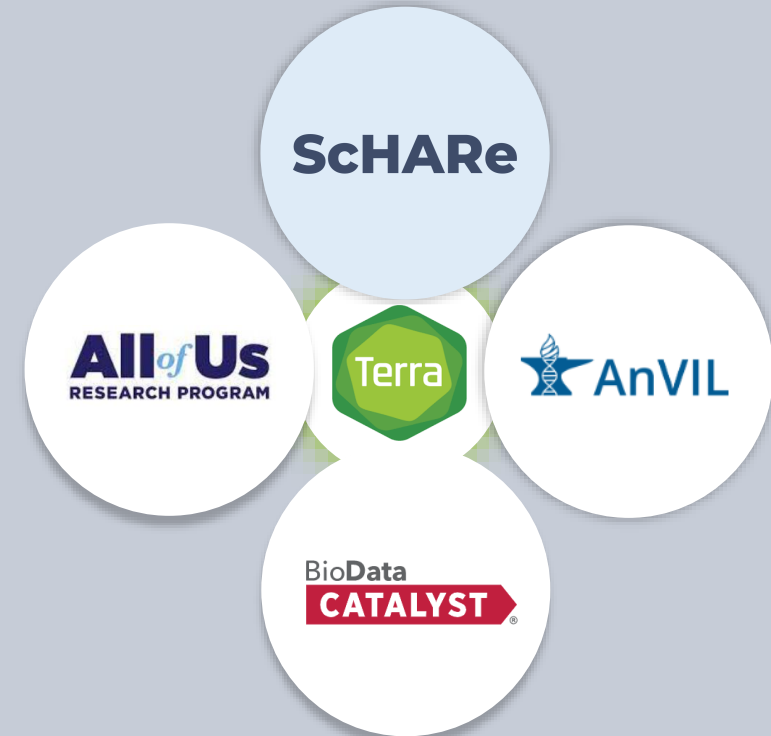
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Copy-and-paste code in Python and R
- Learning Terra on ScHARe prepares you to use other NIH platforms



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

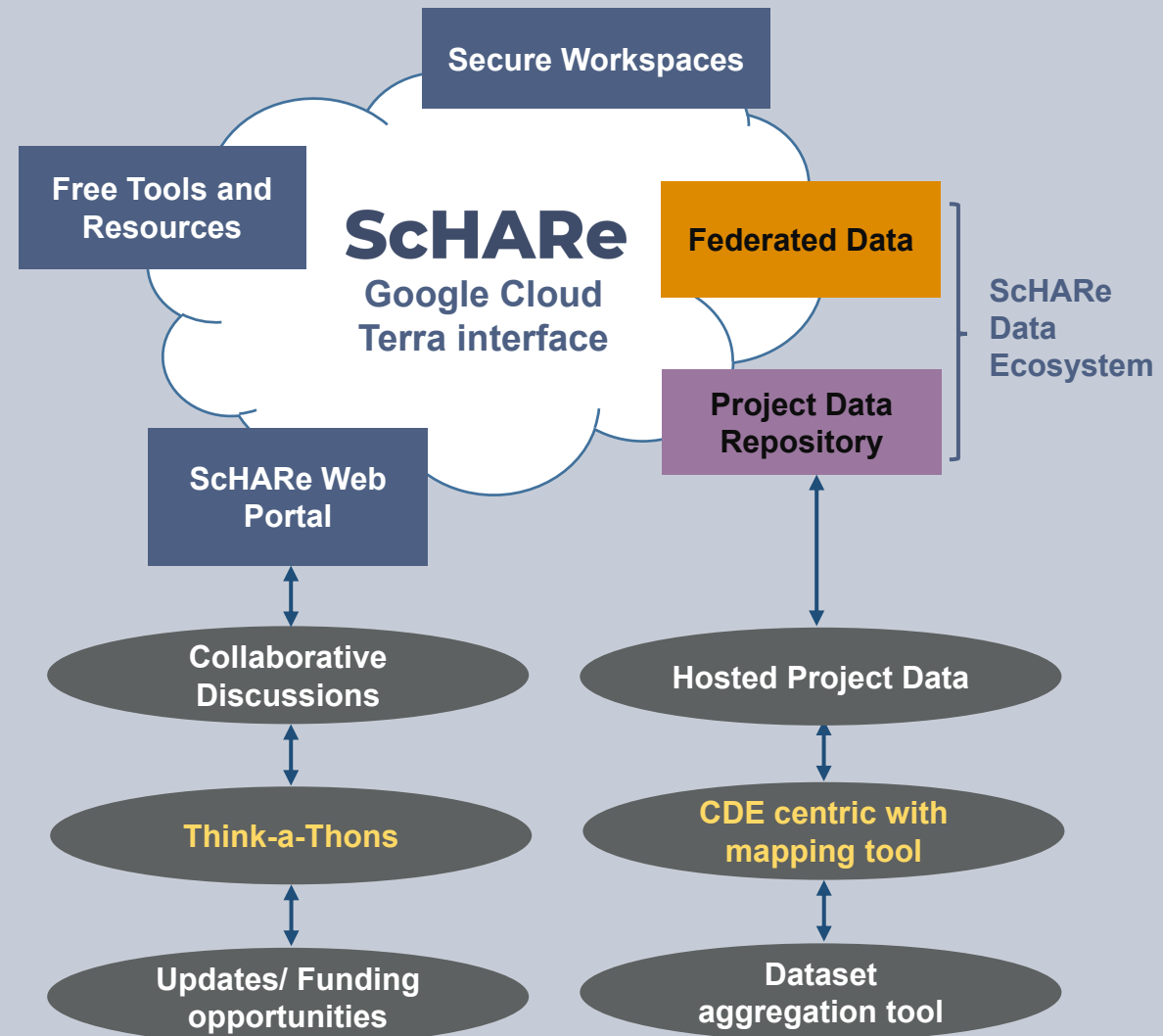


ScHARe Components

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

The screenshot displays the ScHARe Terra interface with a 'Share Workspace' modal dialog open. The dialog is titled 'Share Workspace' and contains the following elements:

- User email:** A text input field with the placeholder 'Add people or groups' and an 'ADD' button.
- Current Collaborators:** A list of collaborators with their roles and permissions:
 - calzonil2@nih.gov:** Role: Owner (dropdown), Permissions: Can share, Can compute.
 - ScHARe-Contractors@firecloud.org:** Role: Writer (dropdown), Permissions: Can share, Can compute. Includes a close button (X).
 - ScHARe-Read-Only-Access@firecloud.org:** Role: Reader (dropdown), Permissions: Can share, Can compute. Includes a close button (X).
- Share with Support:** A toggle switch currently set to 'No'.
- Buttons:** 'CANCEL' and 'SAVE' buttons at the bottom right.

The background interface shows the 'WORKSPACES' section with a search bar, a list of workspaces (e.g., 'ScHARe', 'ScHARe Think-a-Thons'), and a 'MY WORKSPACES (42)' tab selected.

- 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

WORKSPACES
Workspaces > ScHARe/ScHARe > Analyses

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

Your Analyses + START

Application	Name ↓
Jupyter	00_List of Datasets Available on ScHARe.ipynb
Jupyter	01_Introduction to Terra Cloud Environment.ipynb
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb
Jupyter	03_R Environment setup.ipynb
Jupyter	04_Python 3 Environment setup.ipynb
Jupyter	05_How to access plot and save data from public BigQuery datasets using R.ipynb
Jupyter	06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

Modular codes

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

WORKSPACES
Workspaces > ScHARe/ScHARe > ANALYSES

DASHBOARD DATA ANALYSES

WORKFLOWS

Find a Workflow

Suggested Workflows

- haplotypecaller-gvcf-gatk4
Runs HaplotypeCaller from GATK4 in GVCF mode on a single sample.
- mutect2-gatk4
Implements GATK4 Mutect 2 on a single tumor-normal pair.
- processing-for-variant-discovery-gatk4

Find Additional Workflows

Dockstore
Browse WDL workflows in Dockstore, an open platform used by the GA4GH for sharing Docker-based workflows.

- Modular codes developed for reuse
- Adding SAS

ScHARe Terra interface: access to datasets

What data?

The screenshot shows the ScHARe Terra interface in the 'Analyses' tab. The notebook '00_List of Datasets Available on ScHARe.ipynb' is open, displaying the following content:

The ScHARe Data Ecosystem

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

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

A detailed list of the datasets available in the ScHARe Data Ecosystem, including links to documentation and other helpful resources for each dataset, is available in the sections below. The datasets are categorized as follows, based on their content:

A - SOCIAL DETERMINANTS OF HEALTH

- **A1 Multiple Categories:** Datasets that include data on multiple Social Determinants of Health (SDoH) factors/indicators
- **A2 Economic Stability:** Datasets that include data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.
- **A3 Education Access and Quality** Datasets that include data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.
- **A4 Health Care Access and Quality** Datasets that include data on health literacy, use of health IT, emergency room waiting times, evidence-based preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.
- **A5 Neighborhood and Built Environment** Datasets that include data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.
- **A6 Social and Community Context** Datasets that include data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc. For incidence and prevalence of anxiety, depression, and other mental health conditions, see section 'B1 - Diseases and conditions' below
- **A7 Health Behaviors** Datasets that include data on health behaviors

B - HEALTH OUTCOMES

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

Where?

The screenshot shows the ScHARe Terra interface in the 'Data' tab. The 'Data' table is displayed, showing a list of datasets with their names and sizes. The table has columns for 'Name' and 'SizeGb'. The 'EconomicStability' category is selected, showing 62 datasets.

Name	SizeGb
EconomicStability_Id	
FoodAccessResearchAtlasData2010	0.0297
CurrentPopulationSurvey_FoodSecuritySupplement_2011	0.184
CurrentPopulationSurvey_FoodSecuritySupplement_2012	0.185
CurrentPopulationSurvey_FoodSecuritySupplement_2013	0.184
CurrentPopulationSurvey_FoodSecuritySupplement_2014	0.188
AHS_National_Household_2015	0.491
AHS_National_Mortgage_2015	0.002
AHS_National_Person_2015	0.057
AHS_National_Project_2015	0.004
CurrentPopulationSurvey_FoodSecuritySupplement_2015	0.185

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**

The screenshot shows the Terra WORKSPACES Data interface. The top navigation bar includes 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'DATA' tab is active, displaying an 'IMPORT DATA' button and a search bar for tables. A list of tables is shown on the left, with 'EconomicStability (62)' highlighted. The main table on the right lists various datasets with their names and sizes in GB.

		SizeGb
<input type="checkbox"/>	EconomicStability_id	
<input type="checkbox"/>	FoodAccessResearchAtlasData2010	0.0297
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2011	0.184
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2012	0.185
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2013	0.184
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2014	0.188
<input type="checkbox"/>	AHS_National_Household_2015	0.491
<input type="checkbox"/>	AHS_National_Mortgage_2015	0.002
<input type="checkbox"/>	AHS_National_Person_2015	0.057
<input type="checkbox"/>	AHS_National_Project_2015	0.004
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2015	0.184



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.

Health care access and quality

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

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.

Social and community context

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

Economic stability

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

* Health behaviors

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

* Diseases and conditions

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



** 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
```

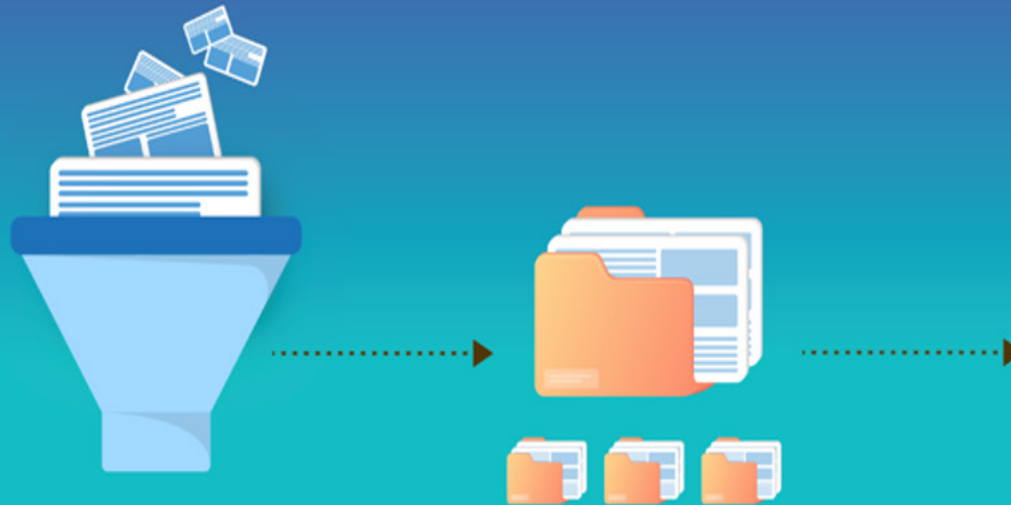

CDE benefits:

- 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

- 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

NIH Endorsed



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/

COMMON DATA ELEMENTS

NLM CDE Repository
Coded NIMHD Common Data Elements

- Labels
- Questions
- Permissible Values

A
T
O

Common Data Elements + Data

Data Access
Based On PII Levels and User Needs:

- Public
- Data Use Agreement
- Private

DATA UPLOAD

Acquired Google and ScHARe Hosted Datasets

Overview

Data Dictionaries

Data Updates

ScHARe REPOSITORY

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

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

Other Cloud Platforms
AnVil, BDC, All of Us



The screenshot shows the 'Create New Collection' form in the ScHARe Repository. The form is titled 'Create New Collection' and is located in the center of the page. It has a dark blue header with the 'Pigeon' logo and navigation links for 'About', 'Docs', 'Community', and 'Collections'. A search bar is also present in the header. The form itself is white and contains the following fields:

- NAME:** A text input field with a cursor.
- DESCRIPTION:** A large text area for entering a description.
- METADATA:** A section with a header 'METADATA' and an information icon. It contains two input fields labeled 'key' and 'value', and a plus sign button to add more metadata.
- Submit:** A button at the bottom of the form.

The left sidebar of the application is dark blue and contains navigation options: 'Recent', 'My Collections', and 'Starred'.

- 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

The screenshot shows the ScHARe Repository interface. At the top, there's a navigation bar with 'About', 'Resources', and 'Data' buttons, a search bar, and a user profile 'AB'. Below this, a sidebar on the left offers 'Create a Collection' and lists 'Most Recent' and 'Your Collections'. The main content area is titled 'pigeon@localhost / Collection Path' and features a 'CDE Configuration' section. This section includes a dropdown for 'Choose a data standard' set to 'ScHARe', 'Save', and 'Cancel' buttons. A table maps data elements from files to common data elements and column names. Below the table, a 'Status' section shows '7/22 CDEs assigned' and '0 validation errors', with a visual list of assigned and unassigned CDEs.

Home Page

← → ↻ 🏠

About Resources Data AB

+ Create a Collection

Most Recent

- Example Collection 1
- Mouseover Collection
- Example Collection 2

Your Collections

- My Collection 1
- My Collection 2
- My Collection 3

pigeon@localhost / Collection Path Admin Star 10.1k

CDE Configuration

Assign your data elements to relevant data standards like ScHARe at scale to enable more powerful analysis. Hold tab when selecting to assign multiple files or columns at once.

Choose a data standard
ScHARe

Save Cancel

File	Common Data Element	Column Name	Data Type
file2.csv	Sex	Client Age	integer
exampleTab.xlsx	Age	Smoker	
	Education Level	College	

Status 7/22 CDEs assigned 0 validation errors

✓ Address Age Education Health Insurance Orientation Sex Zipcode

✗ Annual Income Birthplace Disabilities Disease Disorders Education Employment English Proficiency Household Size Marital Status Medical Treatment Self-Reported Health Social Needs Usual Place of Care

Map project CDEs or variables to ScHARe-PhenX CDEs

ScHARe Repository

PUBLICLY AVAILABLE FALL 2024

The screenshot displays the ScHARe Repository web interface. At the top, there is a navigation bar with 'About', 'Resources', and 'Data' buttons, a search bar, and a user profile icon labeled 'AB'. The main content area shows a collection page for 'pigeon@localhost / Collection Path'. The collection is titled 'Big_Test Collection' and has a description: 'Description text and stuff. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, ullamco laboris nisi ut commodo consequat.' The collection has 10.1k stars and is set to 'Restricted Access' (Privacy Level) and 'Ready' (Analysis Readiness). It also shows '7/22 CDEs present in this collection' under 'ScHARe CDE Compliance'. The left sidebar contains a 'Create a Collection' button and lists 'Most Recent' and 'Your Collections'. The bottom right corner has a 'Filter by CDE' button.

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

ScHARe Repository

PUBLICLY AVAILABLE FALL 2024

The screenshot displays the ScHARe Repository interface. At the top, there is a navigation bar with the 'Pigeon' logo and links for 'About', 'Docs', 'Community', and 'Collections'. A search bar is located on the right side of the navigation bar. Below the navigation bar, the breadcrumb path 'karl / Population Data / LIVE' is shown, followed by a star icon and a row of action buttons: 'Create Readme', 'Create Folder', 'Add File', 'Add Link', 'Make Public', 'Share', 'Edit', and 'Delete'. The main content area is divided into two sections: 'ABOUT' and 'ITEMS'. The 'ABOUT' section contains the text 'Population by zip code, from an unknown source'. The 'ITEMS' section features a large dashed box for file uploads, with the text 'Drag and Drop or [Browse Files](#) to Upload' centered inside. Below this box, there is a file upload interface showing a file named 'pop...csv' with a file icon, an 'Upload Files' button, and a 'Cancel All' button. A purple line points from the text 'Aggregate datasets with drag-and-drop features' to the dashed upload area.

Aggregate datasets with drag-and-drop features

ScHARe Repository

PUBLICLY AVAILABLE FALL 2024

The screenshot displays the ScHARe Repository interface for editing a dataset. The breadcrumb path is **karl / Population Data / LIVE / population_by_zip_2010.csv**. The **Parser Type** is set to **csv**. The **Columns** section lists the following fields:

Column Name	Icon	Type	Value	Actions
minimum_age	✎	Integer		Add →
maximum_age	✎	Integer		Add →
gender	✎	String	Gender fMCdaD9I:0001	✎ ✖ ⋮
zipcode	✎	String	nlhede:7kijL9I3sx	✎ ✖ ⋮
geo_id	✎	String		Add →

The **Results** section shows: **Data available** (green check), **0 parsing errors** (green check), and **5 validation errors** (red X).

The **Table Preview** shows the following data:

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

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

- 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

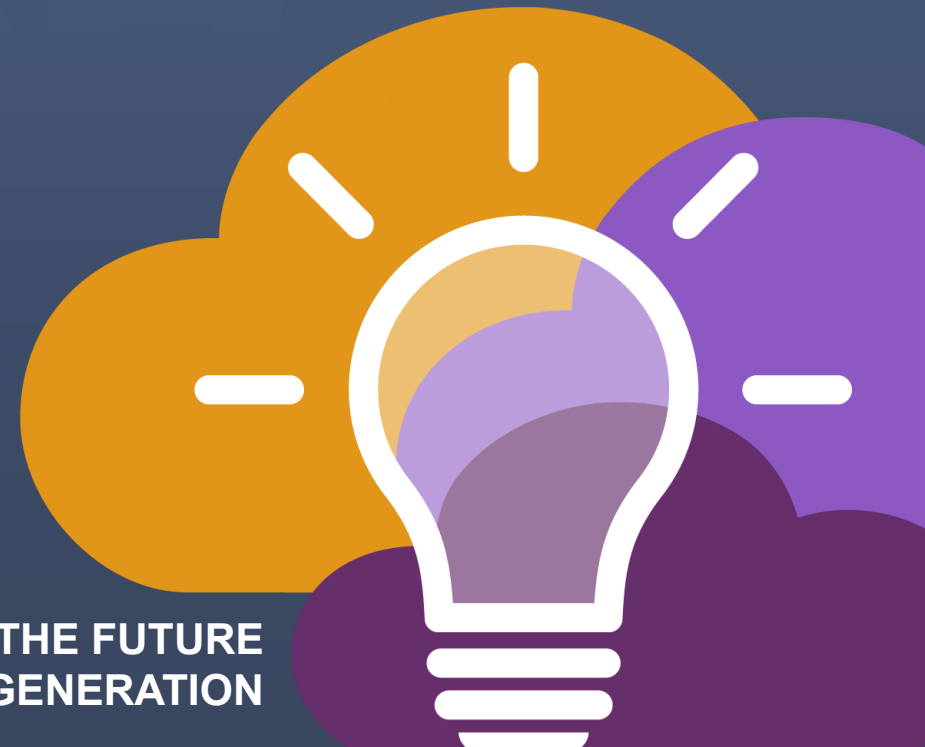
Interoperable: standardized formats and APIs enable seamless integration

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

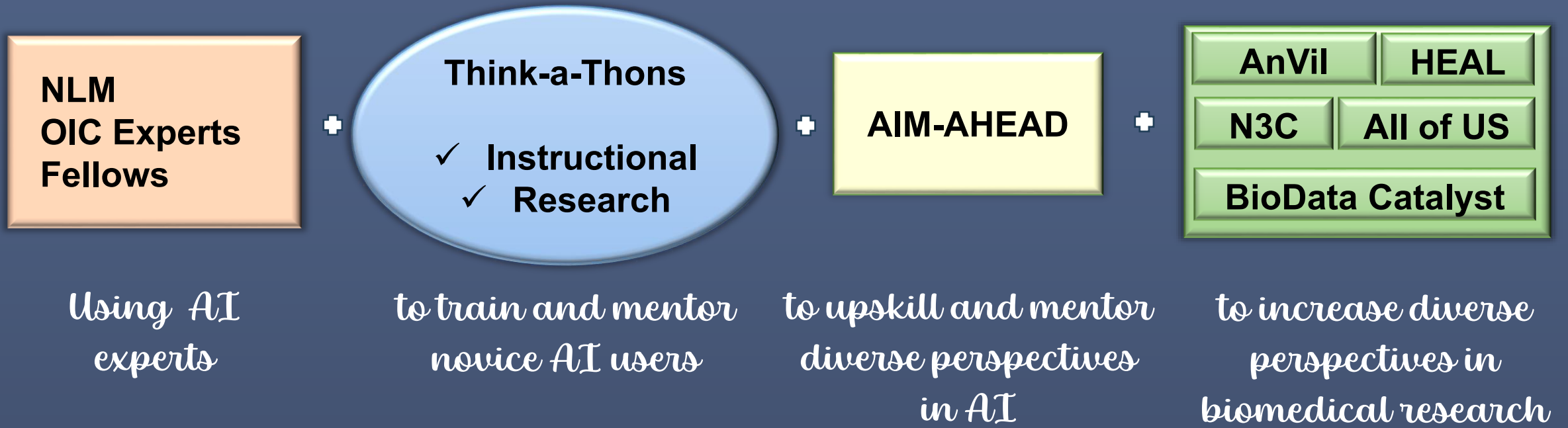
ScHARe

Training
pipeline



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Think-a-Thons training/mentoring pipeline



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

The logo for ScHARe, featuring the text "ScHARe" in a bold, dark blue font inside a white circle.

Data Management and Analysis in Python

September 18, 2024

Deborah Duran, PhD • NIMHD

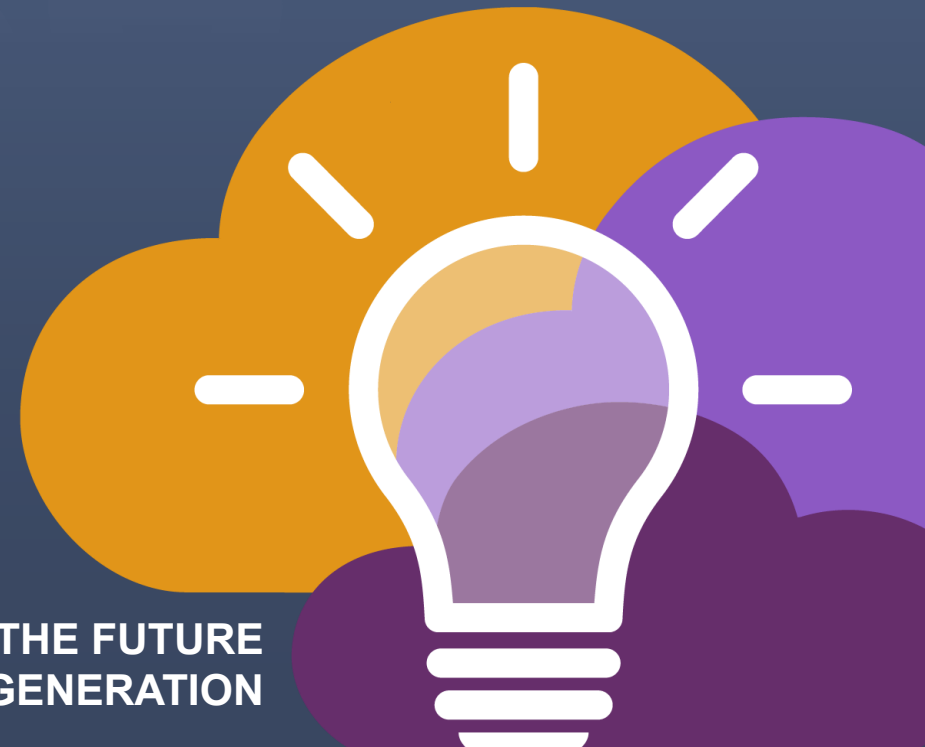
Luca Calzoni, MD MS PhD Cand. • NIMHD

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

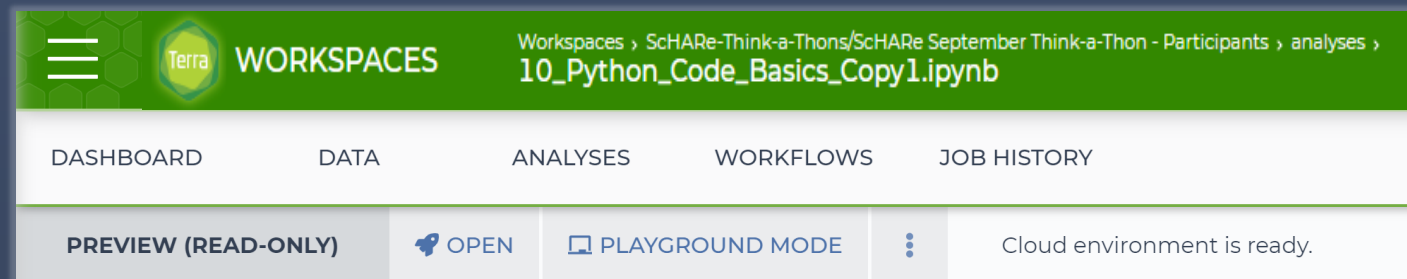
You have been:

- 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

The screenshot shows the Terra WORKSPACES interface. The top navigation bar is green and contains the Terra logo, the word 'WORKSPACES', and a breadcrumb trail: 'Workspaces > SchARE-Think-a-Thons/SchARE September Think-a-Thon - Participants > Analyses'. Below the navigation bar is a horizontal menu with tabs for 'DASHBOARD', 'DATA', 'ANALYSES' (which is selected and highlighted in green), 'WORKFLOWS', and 'JOB HISTORY'. The main content area is titled 'Your Analyses' and features a '+ Start' button and a search bar labeled 'Search analyses'. Below this is a table of 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



The screenshot shows the Terra Community Workbench landing page. At the top, there is a green header with the Terra logo and the word "BETA" on the left, and a notification bell icon with a "1" on the right. The main content area features a large heading "Welcome to Terra Community Workbench" on the left. Below the heading, there is a paragraph of text: "Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate. [Learn more about Terra.](#)" followed by another line: "If you are a new user or returning user, click log in to continue." At the bottom left, there is a blue button labeled "LOG IN". On the right side of the page, there are several hexagonal images: one showing a colorful, glowing molecular structure, and another showing a person in a lab coat and safety glasses holding a test tube.

Terra BETA

1

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.

LOG IN

Click on the login button:

bit.ly/schare-python-notebooks-2



Terra BETA

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.


LOG IN

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...




 Sign in with Google

 Sign in with Microsoft

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

 Sign in with Google



Sign in

to continue to [Terra](#)

Email or phone


[Forgot email?](#)


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.

[Create account](#)


[Next](#)

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.

[Forgot password?](#)

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



New User Registration

First Name *

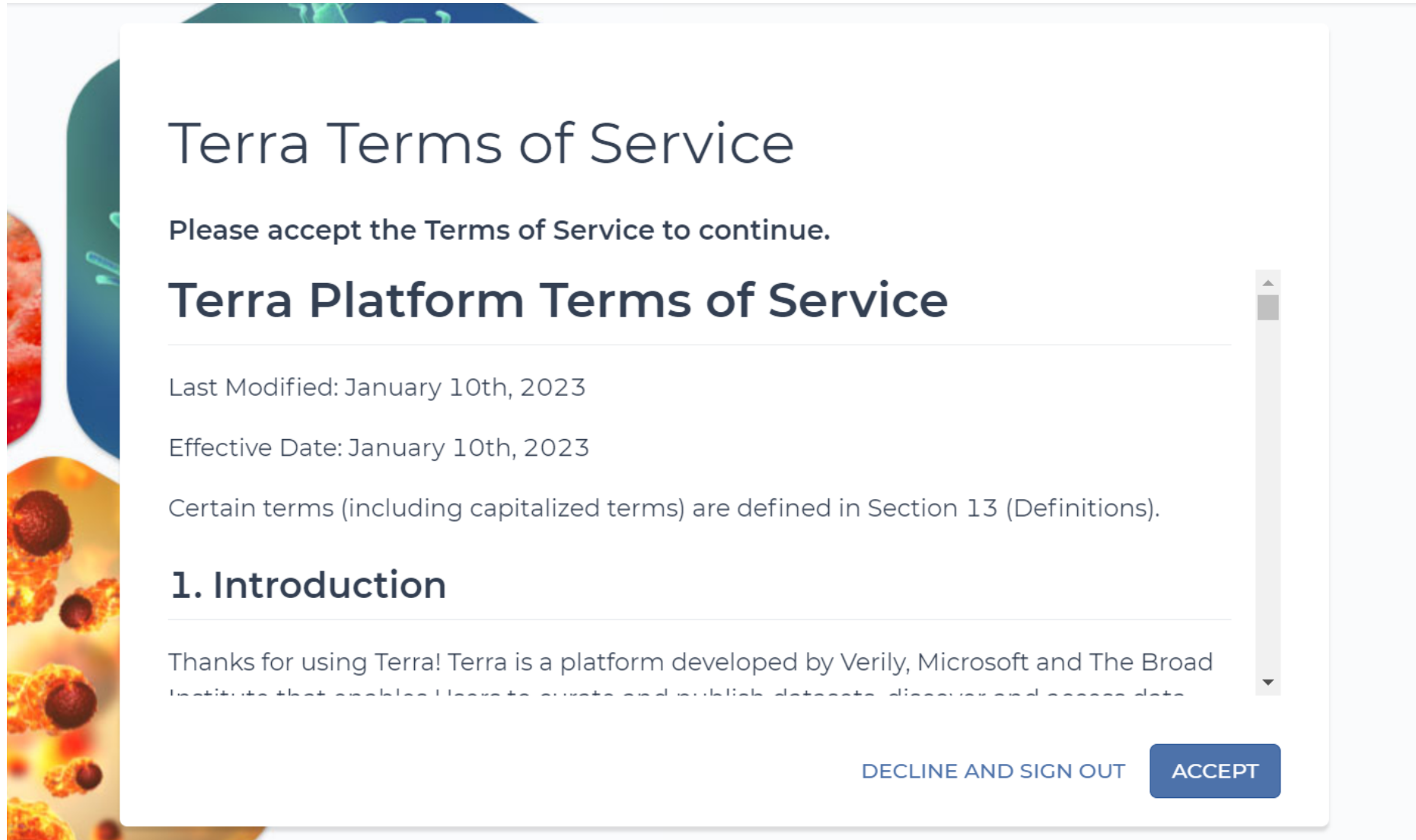
Last Name *

Contact Email for Notifications *

REGISTER

CANCEL

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

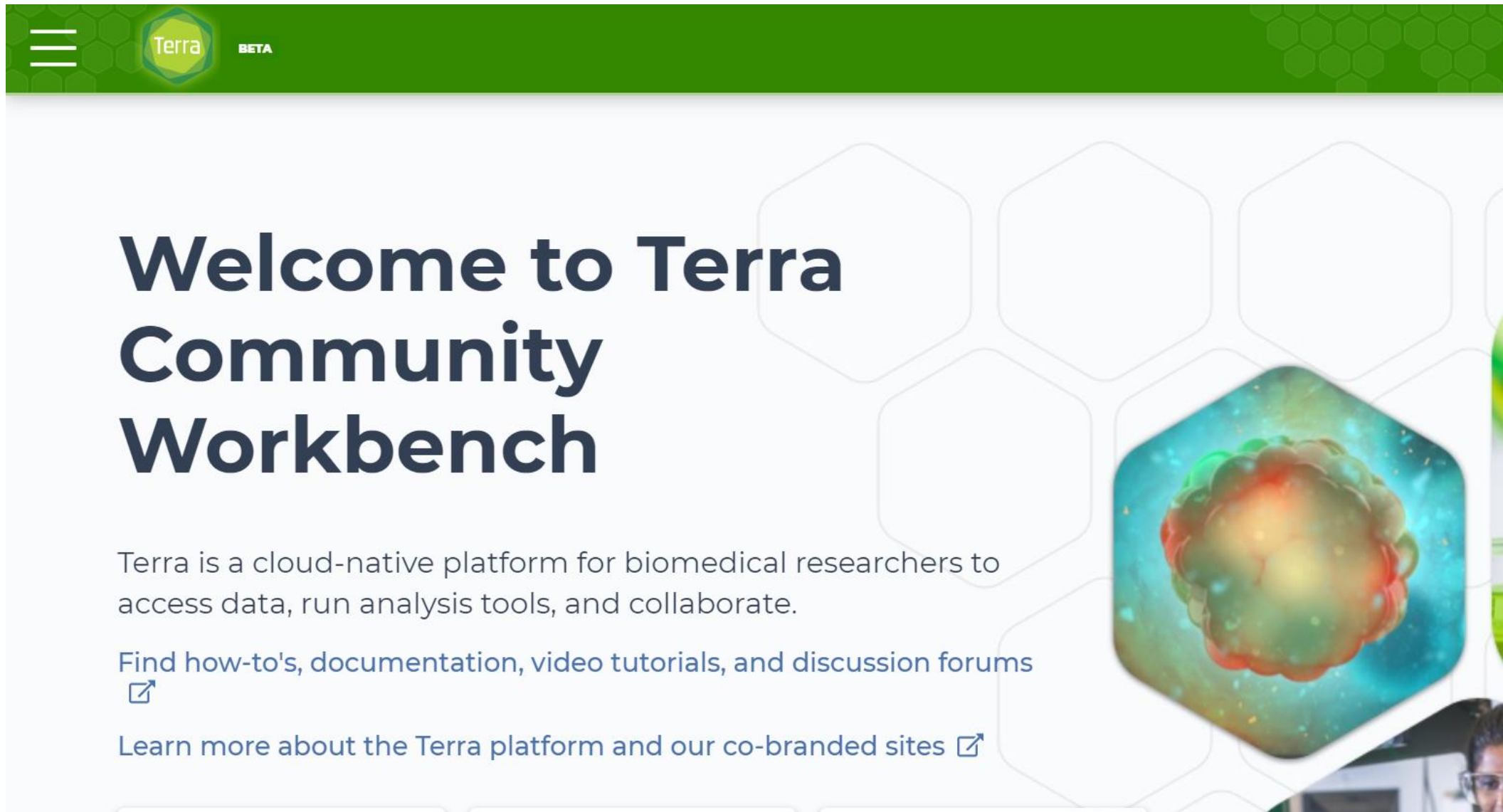
Certain terms (including capitalized terms) are defined in Section 13 (Definitions).



1. Introduction

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

[DECLINE AND SIGN OUT](#) [ACCEPT](#)

You will see this welcome page:



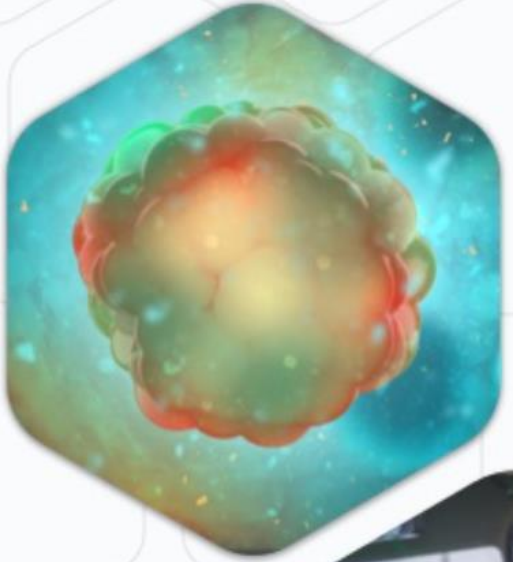
  BETA

Welcome to Terra Community Workbench

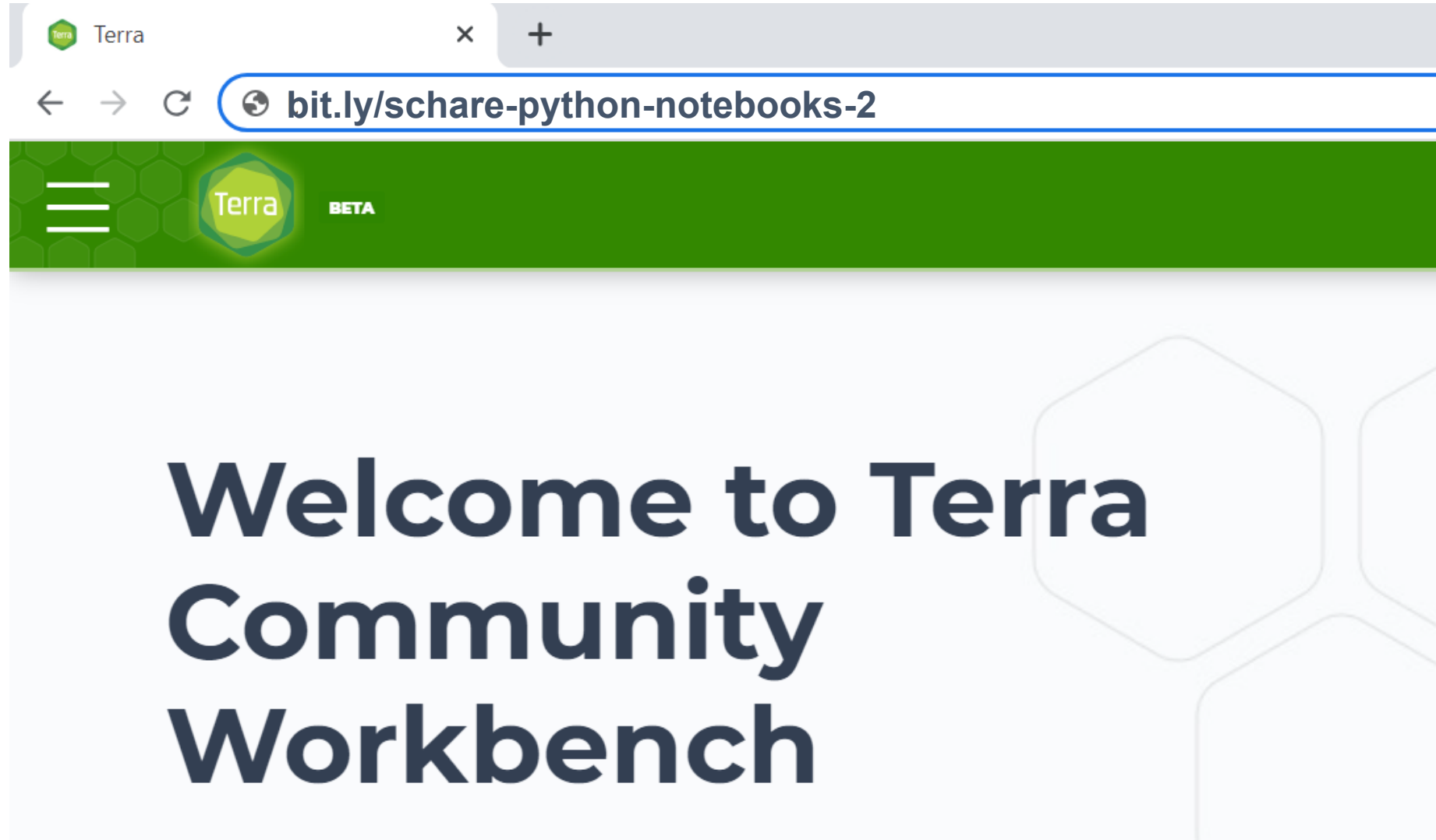
Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate.

Find how-to's, documentation, video tutorials, and discussion forums [↗](#)

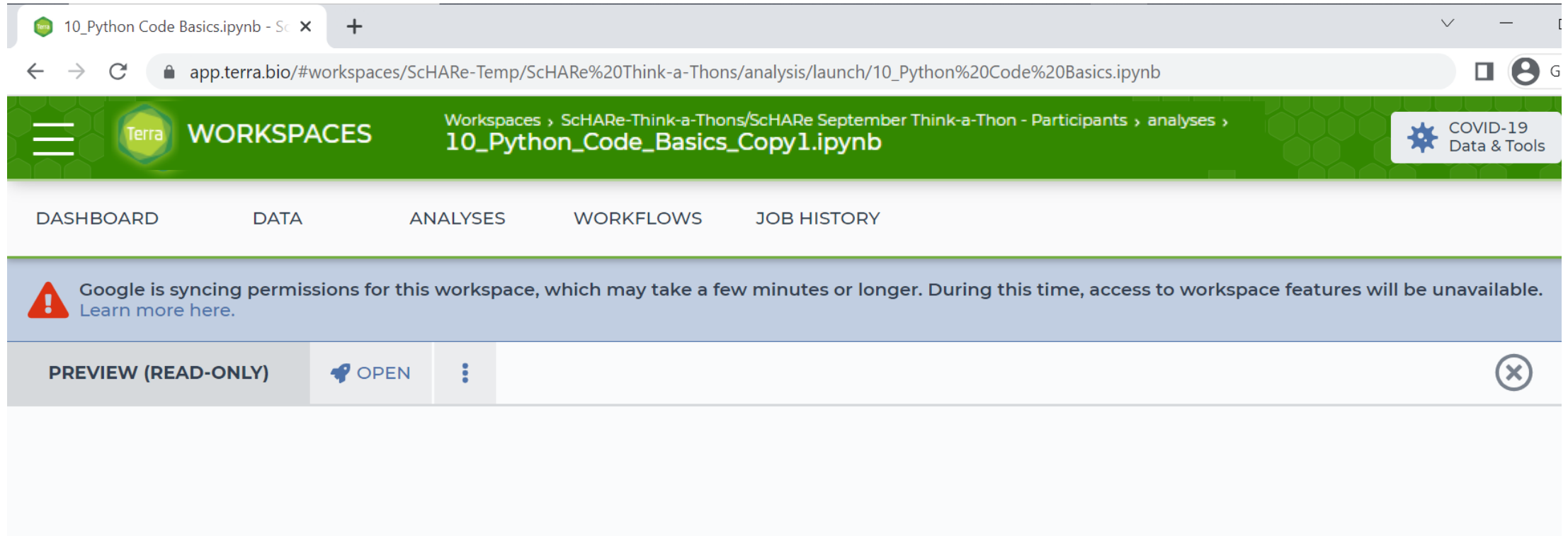
Learn more about the Terra platform and our co-branded sites [↗](#)



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



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:

The screenshot shows the Terra WORKSPACES interface. The top navigation bar is green and contains the Terra logo, the word "WORKSPACES", and a breadcrumb trail: "Workspaces > SchARE-Think-a-Thons/SchARE September Think-a-Thon - Participants > Analyses". Below the navigation bar is a horizontal menu with tabs for "DASHBOARD", "DATA", "ANALYSES" (which is selected and highlighted in green), "WORKFLOWS", and "JOB HISTORY".

The main content area is titled "Your Analyses" and includes a "+ Start" button and a search box labeled "Search analyses". Below this is a table of 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 

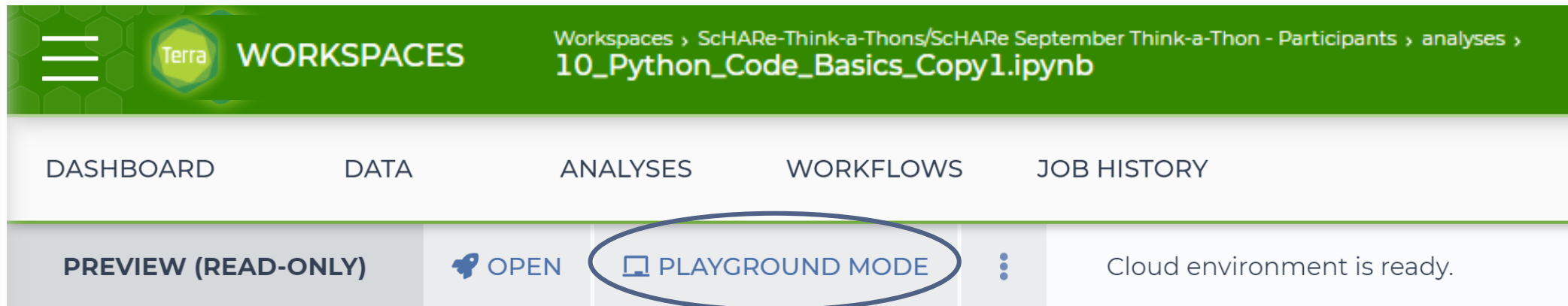
Click on the notebook containing your last name initial:

The screenshot shows the Terra WORKSPACES interface. The top navigation bar includes the Terra logo, the word 'WORKSPACES', and a breadcrumb trail: 'Workspaces > SCHaRe-Think-a-Thons/SCHaRe September Think-a-Thon - Participants > Analyses'. Below this is a secondary navigation bar with tabs for 'DASHBOARD', 'DATA', 'ANALYSES' (which is selected), 'WORKFLOWS', and 'JOB HISTORY'. The main content area is titled 'Your Analyses' and features a '+ Start' button and a search box labeled 'Search analyses'. A table lists four Jupyter notebooks, each with a 'Jupyter' application icon, a name, and a 'Last Modified' date of 'Sep 18, 2024'. The notebook 'Schare_tat_september_S-Y.ipynb' is circled in blue.

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

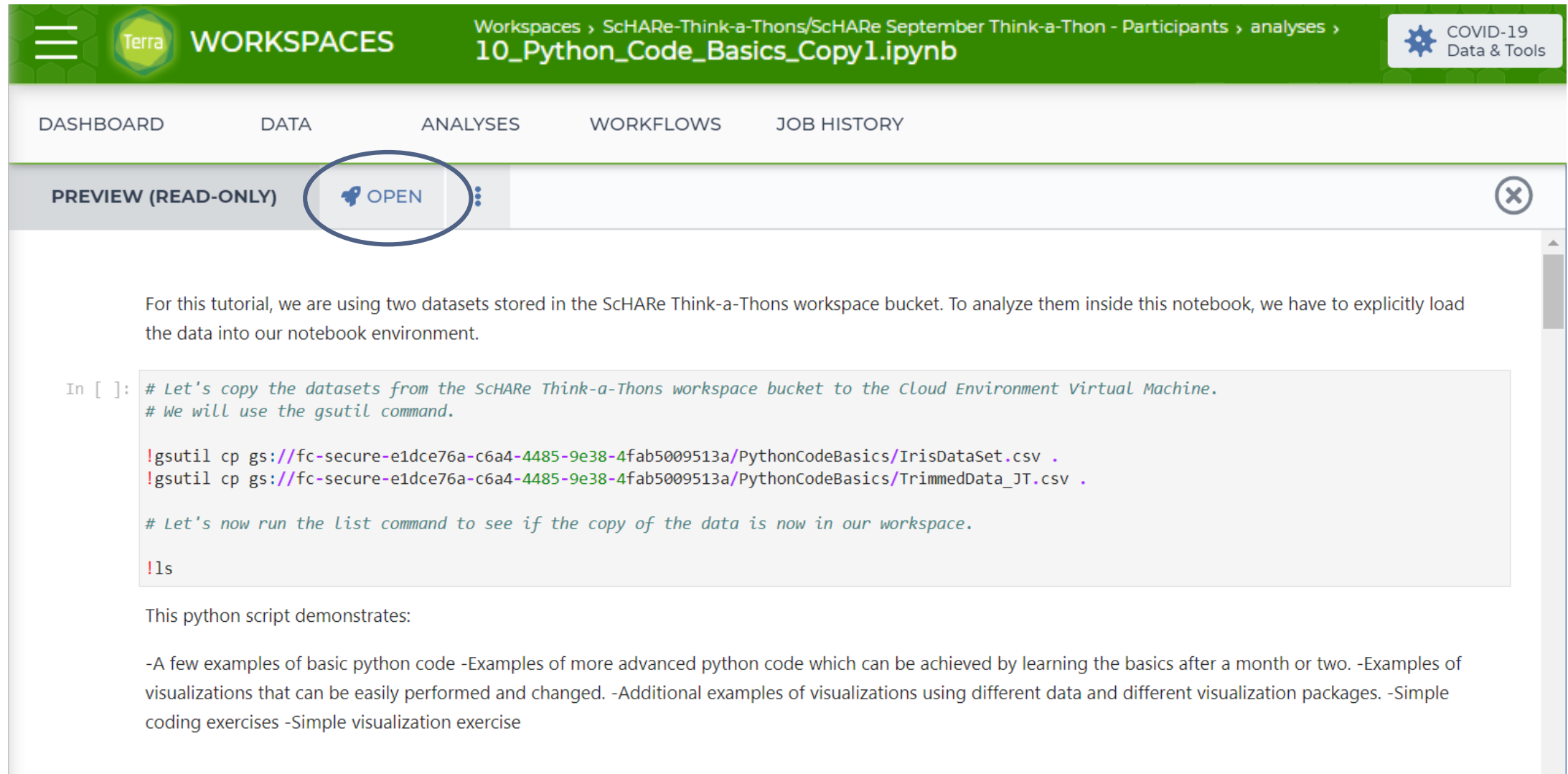
Do you see a Playground mode button?



The screenshot shows the Terra WORKSPACES interface. The top navigation bar is green and contains the Terra logo, the word "WORKSPACES", and a breadcrumb trail: "Workspaces > ScHARe-Think-a-Thons/ScHARe September Think-a-Thon - Participants > analyses > 10_Python_Code_Basics_Copy1.ipynb". Below this is a white navigation bar with tabs for "DASHBOARD", "DATA", "ANALYSES", "WORKFLOWS", and "JOB HISTORY". The main content area has a grey header "PREVIEW (READ-ONLY)", followed by an "OPEN" button with a key icon, a "PLAYGROUND MODE" button with a laptop icon (circled in blue), a vertical ellipsis menu icon, and a status message "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:



The screenshot shows the Terra WORKSPACES interface. At the top, there is a green header with the Terra logo and the text 'WORKSPACES'. To the right of the header, there is a breadcrumb trail: 'Workspaces > SchARE-Think-a-Thons/SchARE September Think-a-Thon - Participants > analyses > 10_Python_Code_Basics_Copy1.ipynb'. In the top right corner, there is a 'COVID-19 Data & Tools' button. Below the header, there is a navigation bar with tabs: 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'ANALYSES' tab is selected. Below the navigation bar, there is a 'PREVIEW (READ-ONLY)' section. In this section, there is a button labeled 'OPEN' with a blue icon of a person, which is circled in blue. To the right of the 'OPEN' button is a close button (an 'X' in a circle). Below the 'PREVIEW (READ-ONLY)' section, there is a text area containing the following text:

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:

The screenshot shows the Terra Jupyter Cloud Environment configuration page. The browser address bar shows the URL: `app.terra.bio/#workspaces/SchARe-Temp/SchARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb`. The page title is "10_Python Code Basics.ipynb - Sc".

The main content area is titled "Jupyter Cloud Environment" and includes a description: "A cloud environment consists of application configuration, cloud compute and persistent disk(s)." Below this, a table shows the costs:

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

The "Application configuration" section includes a dropdown menu for the environment configuration, currently set to "Default: (GATK 4.2.4.0, Python 3.7.12, R 4.3.0)". Below this, it shows "What's installed on this environment?" with a version of 2.2.14, updated on Jun 8, 2023. There is also a field for a "Startup script" (URI) and a checkbox for "Enable GPUs" (BETA).

The "Cloud compute profile" section shows the configuration for the virtual computer: CPUs set to 1 and Memory (GB) set to 3.75. There is also a checkbox for "Enable GPUs" (BETA) and a link to "Learn more about GPU cost and restrictions."

The "Compute type" section is partially visible at the bottom of the configuration panel.

Click on Create below:

10_Python Code Basics.ipynb - Sc x +

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

Terra BETA WORKSPACES

Workspaces > SCHaRe-Temp/SCHaRe%20Think-a-Thons/analysis/launch/10_Python Code Basics

DASHBOARD DATA ANALYSES WORKFLOWS JOBS

Google is syncing permissions for this workspace, which may take a few minutes. [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

CREATE

It will take some time...

The screenshot shows a web browser window with the URL `app.terra.bio/#workspaces/SchARe-Temp/SchARe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb`. The page header includes the Terra logo, the word "WORKSPACES", and a breadcrumb trail: "Workspaces > SchARe-Think-a-Thons/SchARe September Think-a-Thon - Participants > analyses > 10_Python_Code_Basics_Copy1.ipynb". There are also buttons for "COVID-19 Data & Tools" and a notification bell with a "2" badge. The main navigation bar contains "DASHBOARD", "DATA", "ANALYSES", "WORKFLOWS", and "JOB HISTORY". Below this, there are buttons for "PREVIEW (READ-ONLY)", "OPEN", and "PLAYGROUND MODE". A blue oval highlights a notification message: "Creating cloud environment. You can navigate away and return in 3-5 minutes." with a close button (X) on the right.

When the system is ready, click on Playground mode:

The screenshot shows a web browser window with the URL `app.terra.bio/#workspaces/SchARE-Temp/SchARE%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb`. The page header is green and contains the Terra logo, the word "BETA", and "WORKSPACES". The breadcrumb trail reads "Workspaces > SchARE-Temp/SchARE Think-a-Thons > analyses > 10_Python Code Basics.ipynb". There is a "COVID-19 Data & Tools" button and a notification bell with a "2" badge. Below the header is a navigation bar with "DASHBOARD", "DATA", "ANALYSES", "WORKFLOWS", and "JOB HISTORY". The main content area has a control bar with "PREVIEW (READ-ONLY)", "OPEN", and "PLAYGROUND MODE" buttons. The "PLAYGROUND MODE" button is circled in blue. To the right of the buttons is a status message: "Creating cloud environment. You can navigate away and return in 3-5 minutes." with a close button.

Click on Continue:

10_Python Code Basics.ipynb - Sc x +

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 P

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

Error Creating Cloud Environment

Details

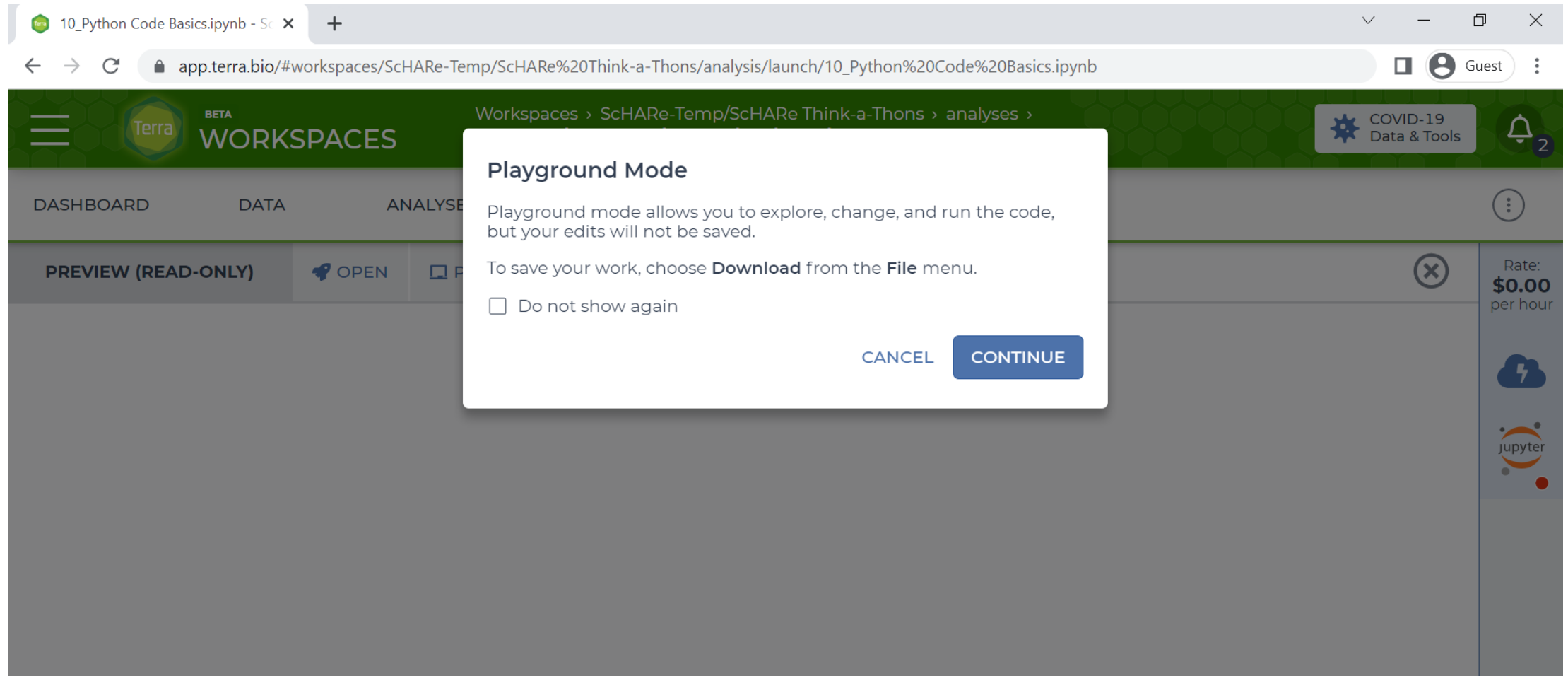
Rate: \$0.00 per hour

jupyter

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

The screenshot shows a web browser window with the URL `app.terra.bio/#workspaces/SCHaRe-Temp/SCHaRe%20Think-a-Thons/analysis/launch/10_Python%20Code%20Basics.ipynb`. The page displays the Terra workspace interface. A central modal dialog box is open, titled "Cloud Environment is in error state", with the message "Failed to create cluster 101753 due to 5 seconds" and an "OK" button. In the background, a notification banner reads "Error Creating Cloud Environment" with a "Details" link. The interface includes a navigation menu with "DASHBOARD", "DATA", and "ANALYSES", and a sidebar with "PREVIEW (READ-ONLY)", "OPEN", and "PLAY" buttons. The top right shows the user is logged in as "Guest".

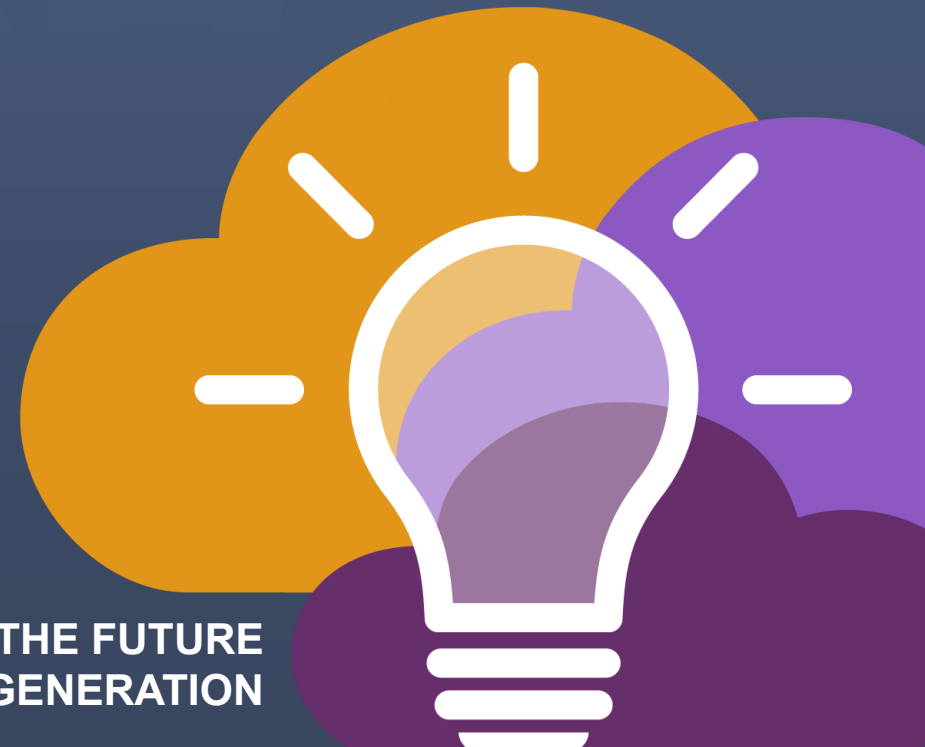
If all goes well, you will see this:



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.quanhub.com/python-for-data-science/
[coursera.org](https://www.coursera.org)

What is R?

R is a **programming language** for statistical computing and graphics

It is used by data miners, bioinformaticians and statisticians for data analysis

Users have created **packages** to augment its functions

Third-party **graphical user interfaces** are also available, such as Rstudio



supports **both Python and R**

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.

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

- 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 – 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

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

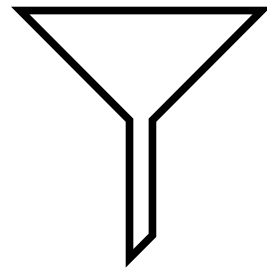
- Corrupted
- Inaccurate
- Duplicates
- Irrelevant information

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

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.

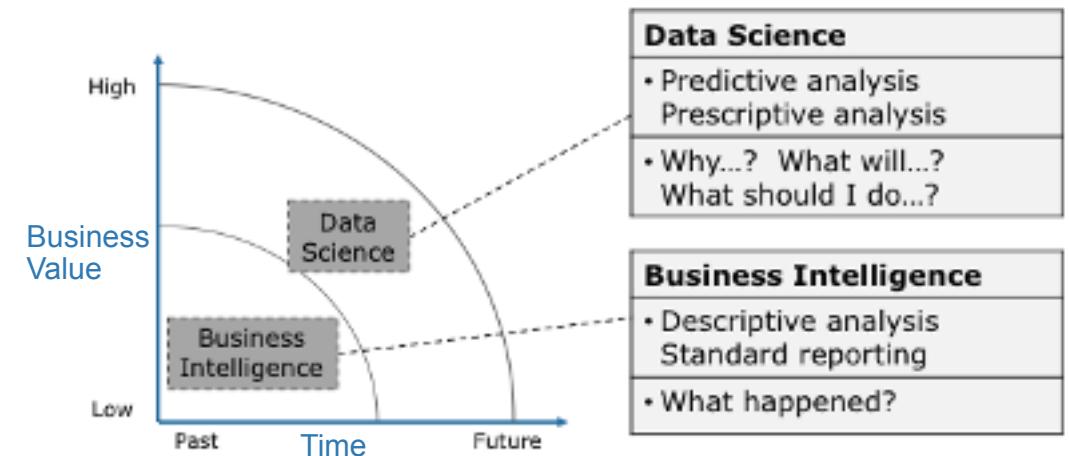
Slido quiz

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

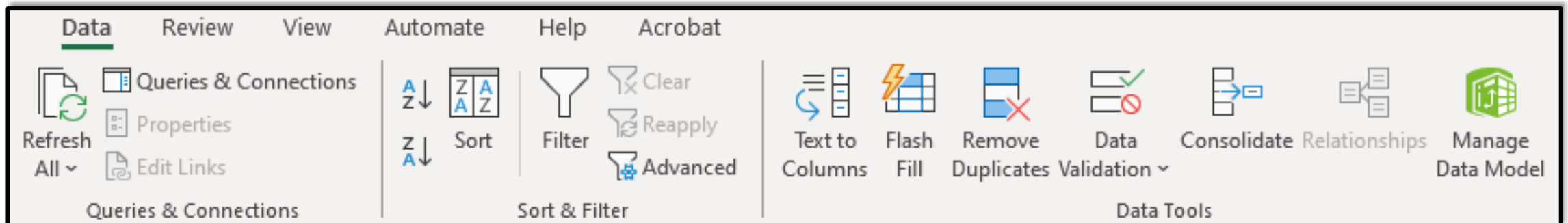
- Allows for informed decision making, and it is the precursor to artificial intelligence.
- Enhances efficiencies by saving time, effort, and resources.
- Improves satisfaction for consumers and producers
- In Public Health and Regulatory environments, it helps to maintain trust and avoid legal actions.

Business Intelligence versus Data Science



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

Excel



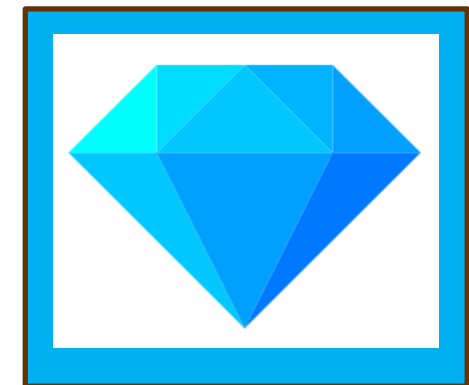
- Open Source
- Desktop application
- Data cleanup and transformation
- Faceting
- Clustering
- Reconciling

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?

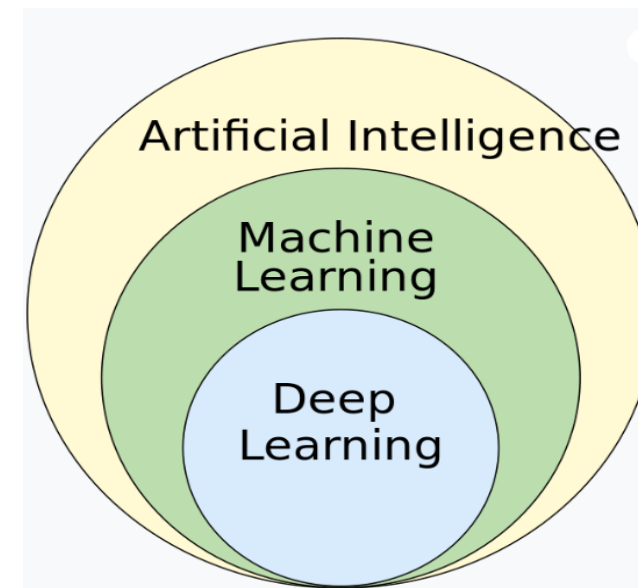
- Word processing
- Creating spreadsheets
- Data cleanup and transformation
- Developing websites

- 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

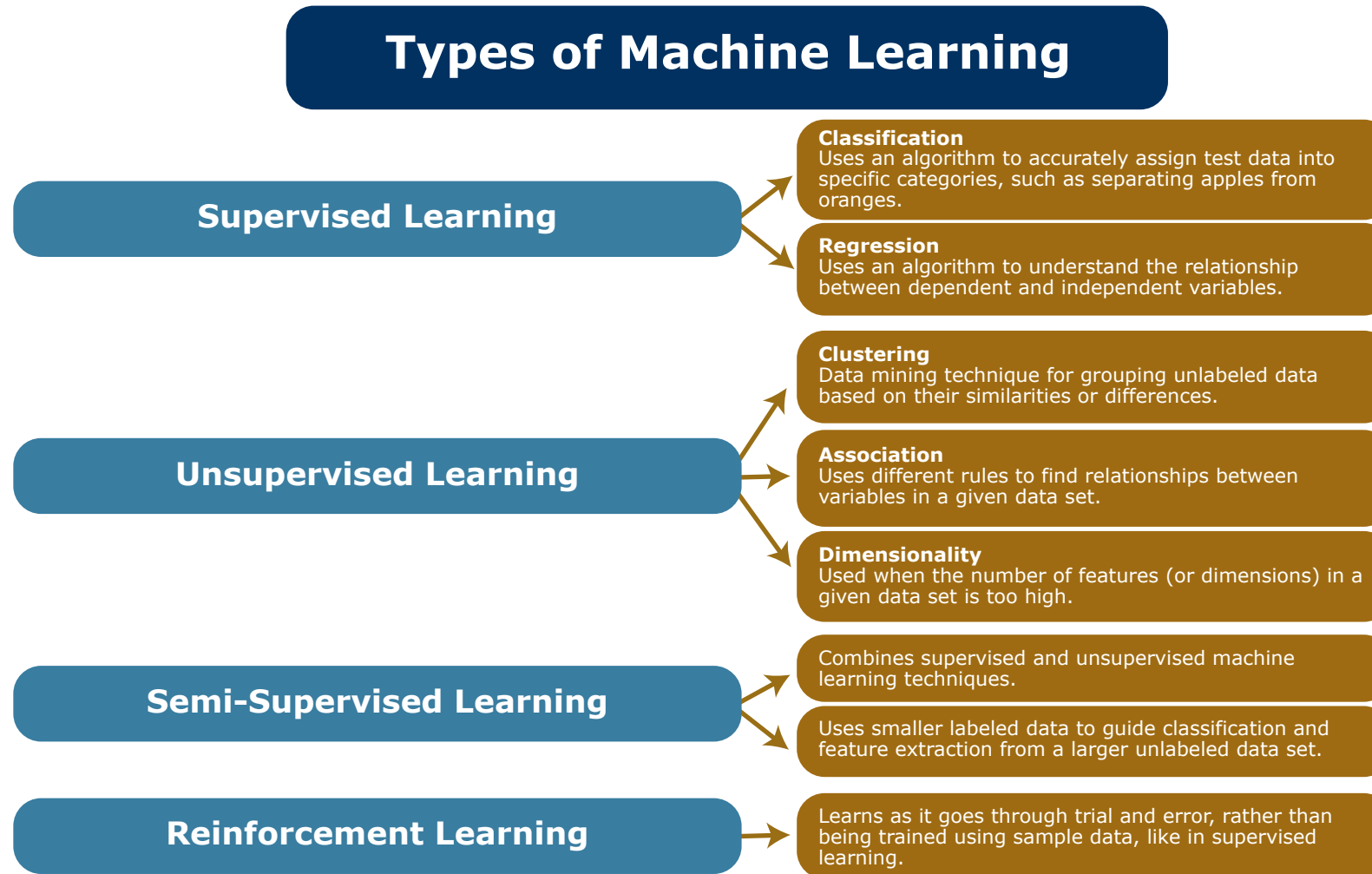
- Use to better understand data
- Prepare the data so it tells an accurate story
- Understand the data and any potential bias

Machine Learning

- Machine Learning - 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/>



- 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

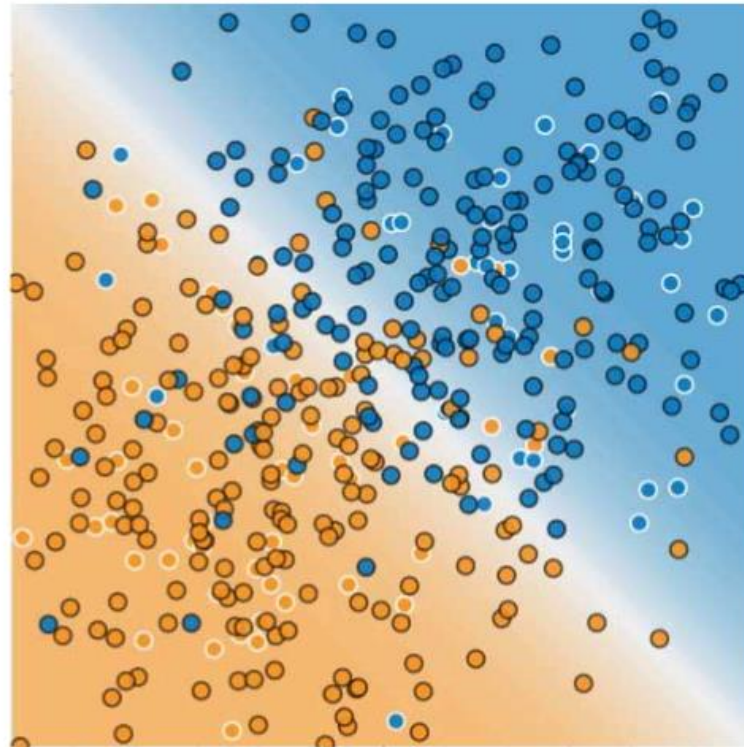
- Uses algorithms to sort unlabeled and unstructured data
- Algorithms discover data patterns without human intervention
- Good situations without clear delineations between different data categories
- Example:
 - Recommend similar types of research projects or publications

- 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

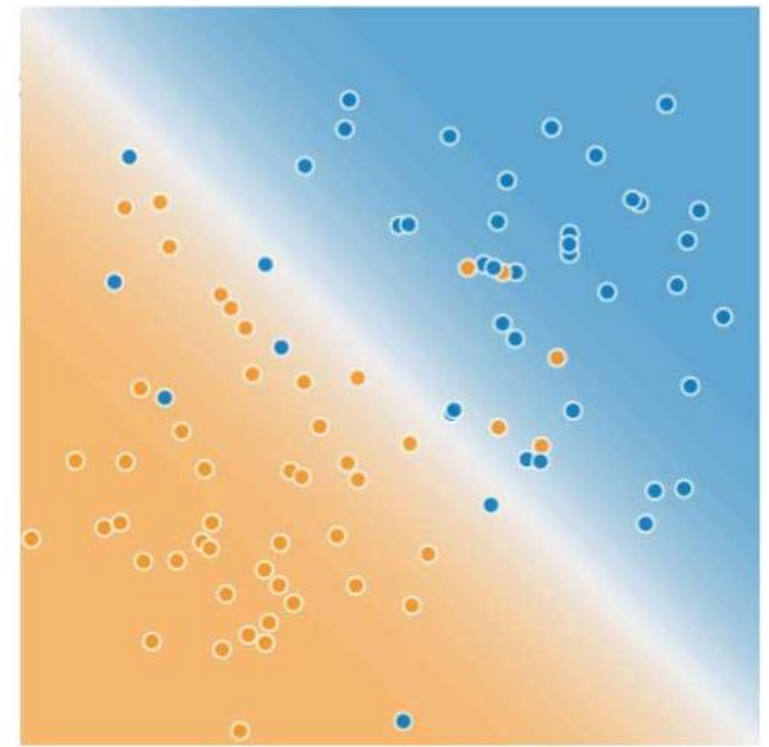
- Used for decision-making in a complex, uncertain environment
- Game-like rules system designed to maximize algorithm's score
- Programmers define rules; computer starts without guidance
- Computer learns through trial and error for optimal solutions
 - Example: used for language processing, self-driving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

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

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;
= $TP/TP + \frac{1}{2}(FP+FN)$

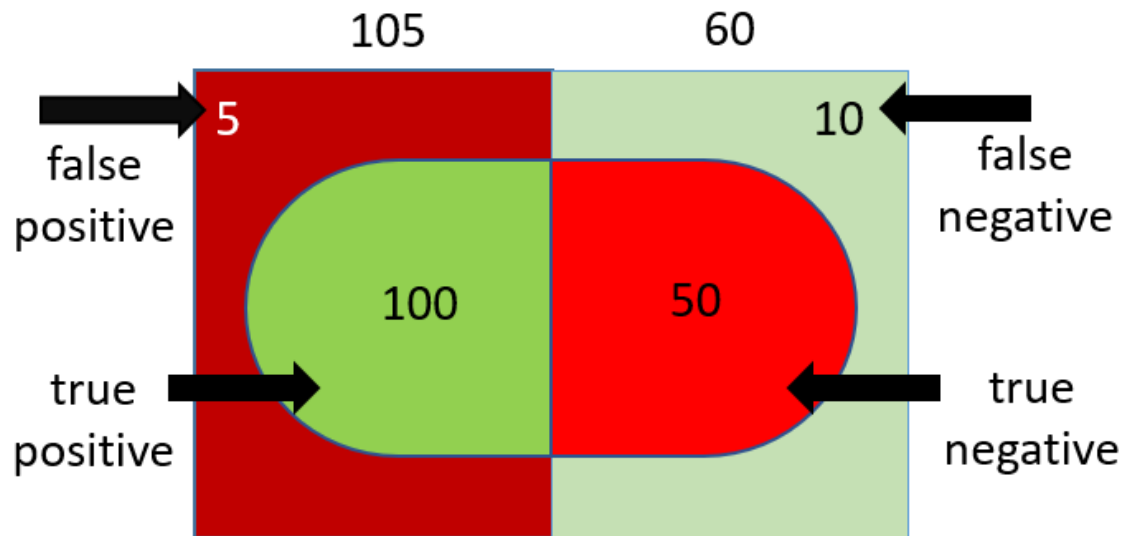
Machine Learning Performance Example

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;

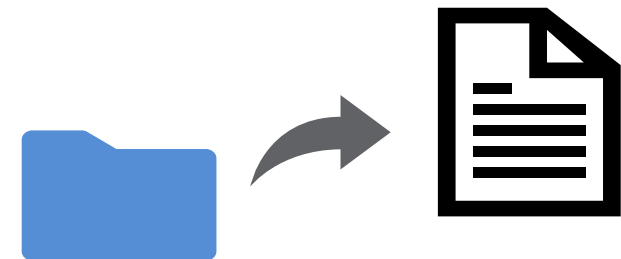
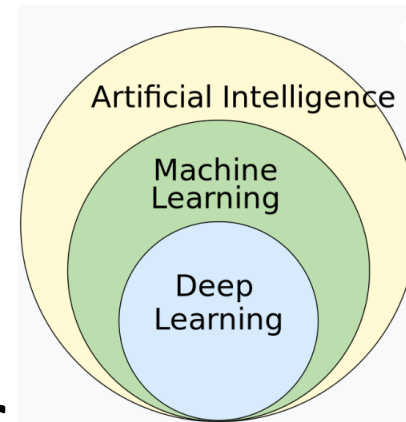
$$= 2 (precision * recall / precision + recall)$$

$$\text{Or} = TP / TP + \frac{1}{2}(FP + FN)$$

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

$$= 100 / 107.5 = .9302$$

- 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

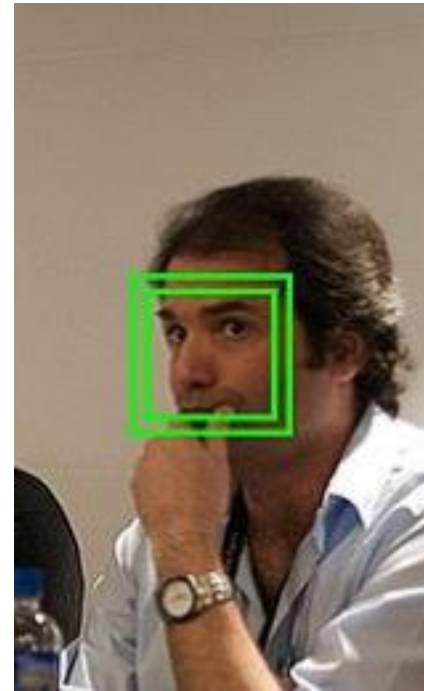


Used in everyday products:

- 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



- **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

- **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

The logo for ScHARe, featuring the text "ScHARe" in a bold, dark blue font inside a white circle.

Data Management and Analysis in Python

September 18, 2024

Deborah Duran, PhD • NIMHD

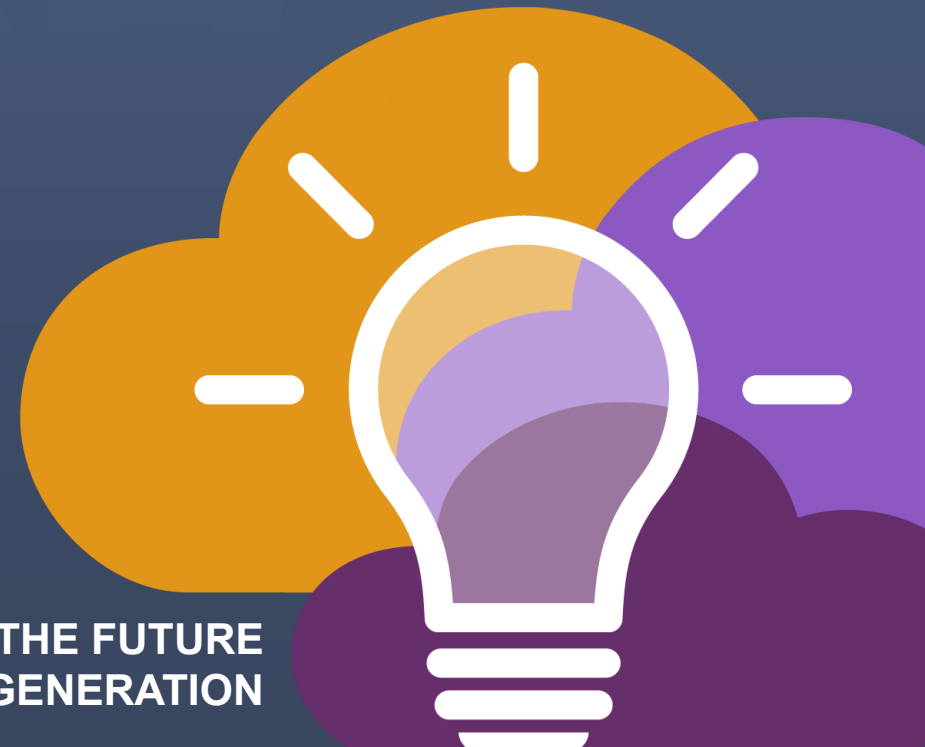
Luca Calzoni, MD MS PhD Cand. • NIMHD

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

A screenshot of a webpage from FreeCodeCamp. The page has a dark blue header with the FreeCodeCamp logo and a tagline. Below the header, there are two main sections of text, each with a bold title and a paragraph of description. The first section is titled 'Python Tutorial for Beginners (Learn Python in 5 Hours)' and describes a course by TechWorld with Nana. The second section is titled 'Scientific Computing with Python' and describes a certification course.

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

Catalog > Computer Science Courses > HarvardX's Computer Science for Web Programming



Harvard University: CS50's Introduction to Computer Science

An introduction to the intellectual enterprises of computer science and the art of programming.



12 weeks

6–18 hours per week



Self-paced

Progress at your own speed

There is one session available:

4,974,616 already enrolled! After a course session ends, it will be [archived](#) .

Starts Jul 19

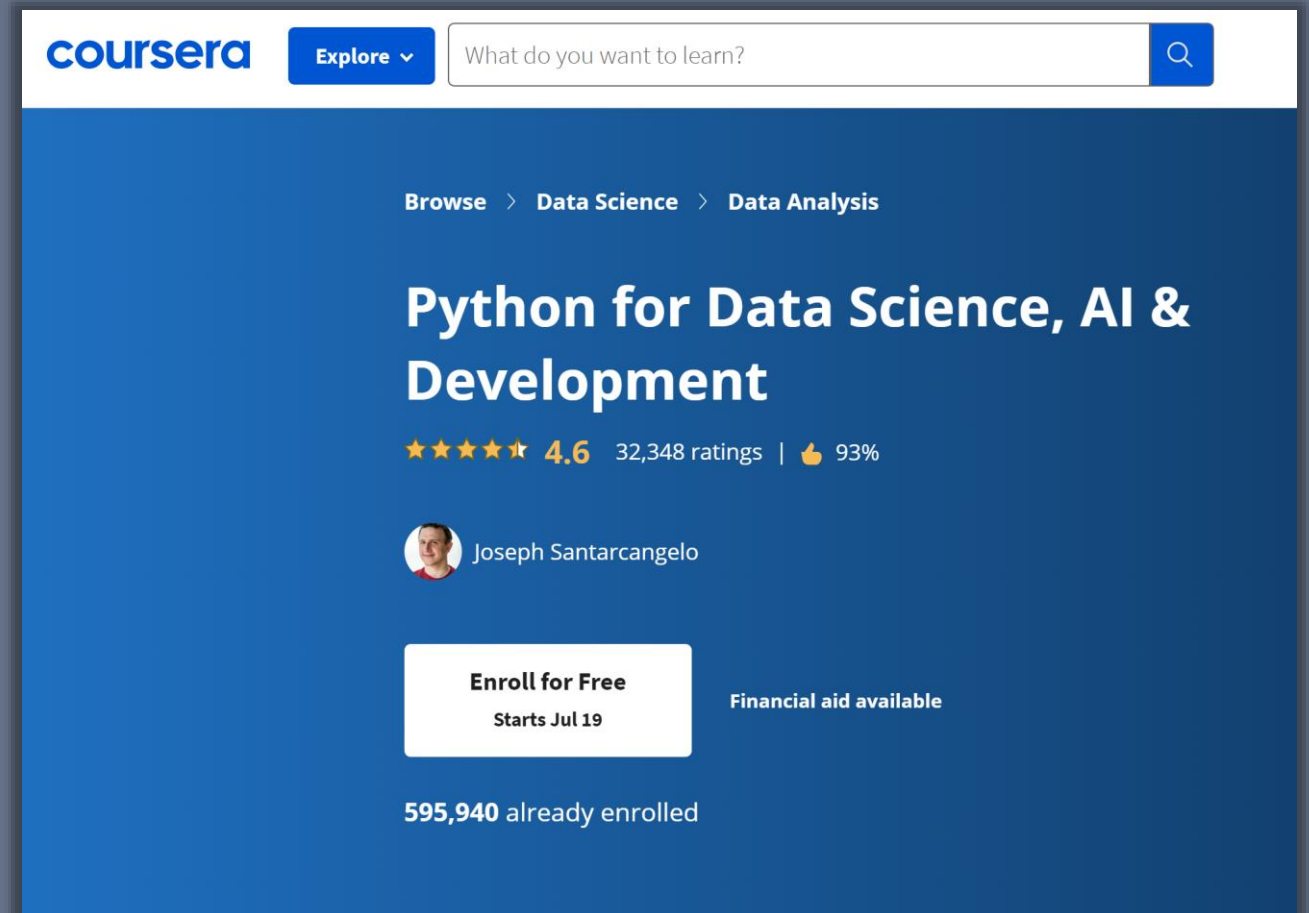
Ends Dec 31

Enroll

Python resources

Coursera

Free and Paid Python Courses




The screenshot shows the Coursera website interface. At the top, there is a navigation bar with the Coursera logo, an 'Explore' dropdown menu, and a search bar containing the text 'What do you want to learn?'. Below the navigation bar, the breadcrumb trail reads 'Browse > Data Science > Data Analysis'. The main heading for the course is 'Python for Data Science, AI & Development'. Below the heading, there is a rating of 4.6 stars based on 32,348 ratings, and a thumbs-up icon indicating a 93% approval rate. The instructor's name, Joseph Santarcangelo, is displayed next to his profile picture. A prominent white button with black text says 'Enroll for Free' and 'Starts Jul 19'. To the right of this button, it says 'Financial aid available'. At the bottom of the course card, it states '595,940 already enrolled'.

coursera Explore ▾ What do you want to learn? 🔍

Browse > Data Science > Data Analysis

Python for Data Science, AI & Development

★★★★☆ 4.6 32,348 ratings | 👍 93%

 Joseph Santarcangelo

Enroll for Free
Starts Jul 19

Financial aid available

595,940 already enrolled

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

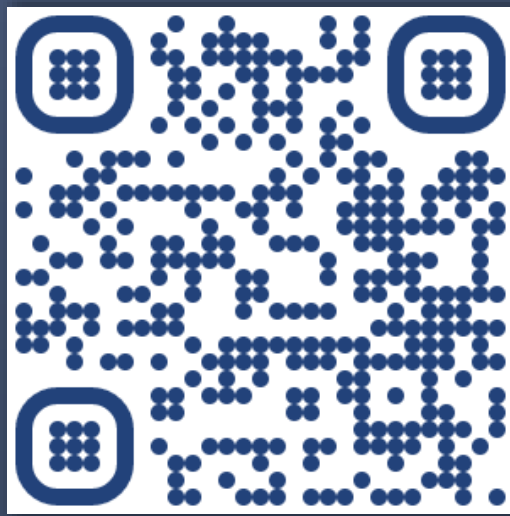
SchARE

Next Think-a-Thons:



bit.ly/think-a-thons

Register for SchARE:



bit.ly/join-schare



schare@mail.nih.gov

