Welcome to the DS-I Africa Intro to Terra!

- All of today's slides and other materials can be found at http://broad.io/DSI-Africa-2204
- If you haven't already:
  - **Sign up for Terra** at https://elwazi.terra.bio
  - **Set up billing** by sending your Terra email to a TA in the chat
    - FYI - Billing projects/workspace deleted 1 week after workshop
- Post-workshop survey: https://www.surveymonkey.com/r/NMSWXYY

How to interact during the session

To ask a question:

- Turn on your mic and ask!
- Or send a message to the Zoom chat and a TA will respond.
- Today’s TAs from the Data Sciences Platform:
  - Anton Kovalsky
  - Liz Kiernan
  - Pamela Bretscher
  - Tiffany Miller
- Also: remember to **raise your hand** when done with a task.

Agenda Day 1

8:00 am (ET)/ 3:00 pm (GMT + 3)	**Welcome**
8:05 am ET/ 3:05 (GMT +3)	**Terra Overview and Platform Tour** An overview of Terra functionality and platform navigation.
8:35 am ET/ 3:35 pm (GMT+3)	**Billing and Secure Collaboration** An overview of billing setup and the granularity of access levels offered in Terra.
9:00 am - 10:30 am ET/ 4:00 pm - 5:30 pm (GMT+3)	**Breakout for Registration and Cloud Environment and Interactive Analysis** An introduction to the Terra Cloud Environment and applications for interactive analysis.

Agenda Day 2

8:00 am (ET)/ 3:00 pm (GMT + 3)	**Breakout for Cloud Environment and Interactive Analysis - Advanced** Finish interactive analysis module and practice finding and using additional workspace resources.
9:40 am - 10:00 am ET/ 4:40 pm - 5:00 pm (GMT+3)	**Closing Remarks, Q&A**

You will learn about:

- What Terra is and why it’s on the cloud
- Billing and secure collaboration
- Organizing data, tools, and computational resources in Terra
- Creating and running analyses on custom cloud environments
- Accessing different cloud data and moving it between resources
- Finding example resources to help get started

Icebreaker!

Let’s take a poll!

Hit Record!

Live transcript enabled

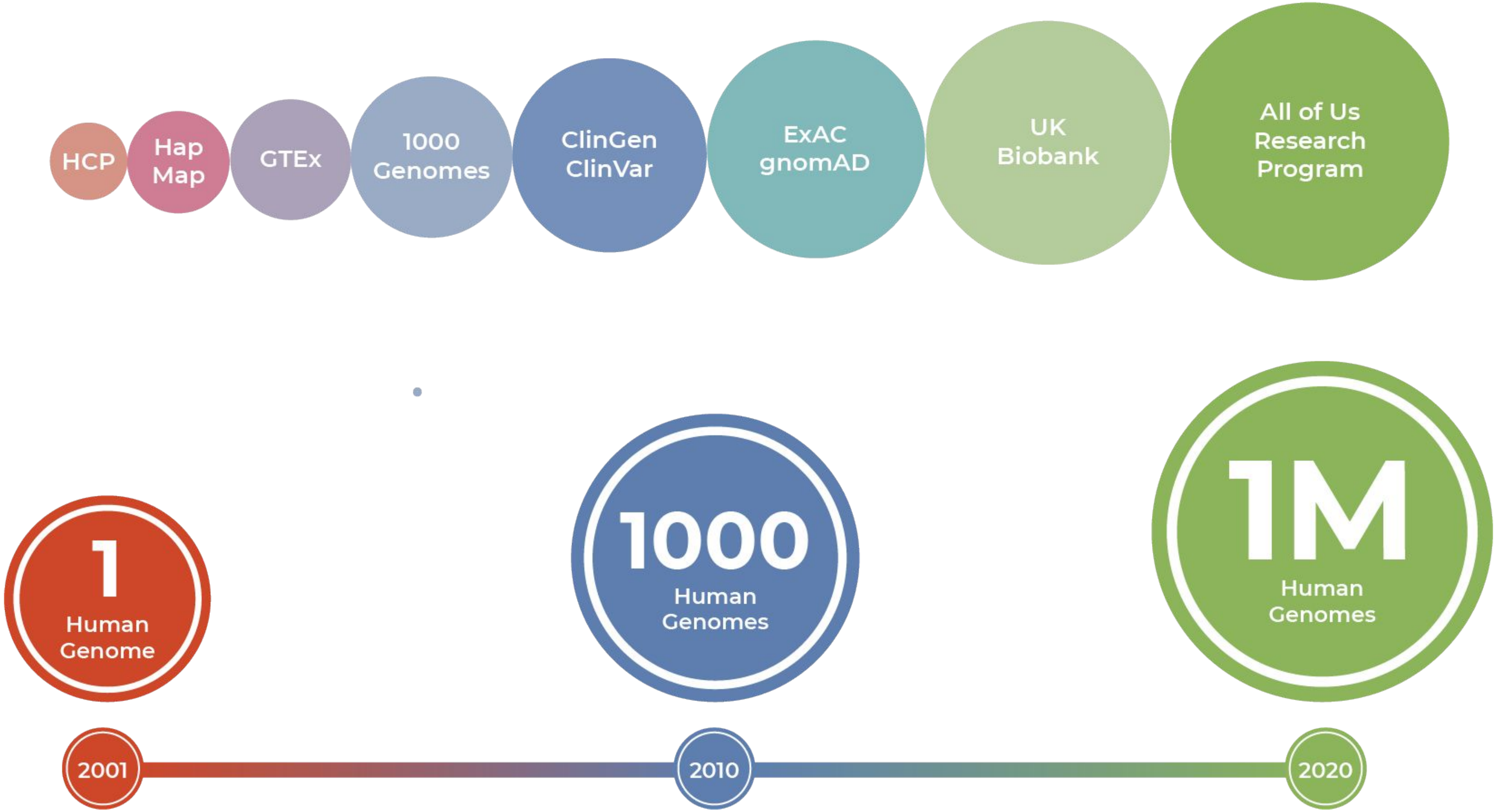
Terra

Access data, run analysis tools, and collaborate on the cloud

https://terra.bio

BROAD INSTITUTE

Data generation is on a roll



https://terra.bio

What is Terra?

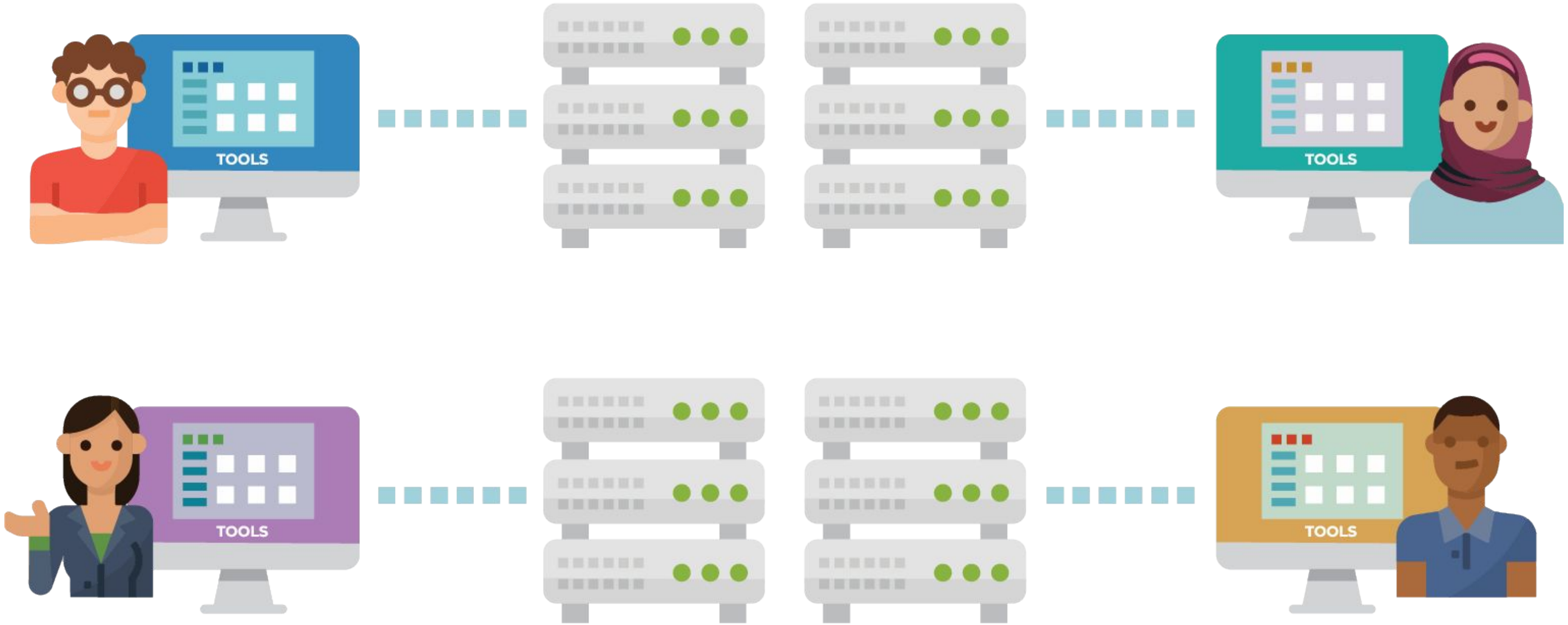
And why is it on the cloud?

https://terra.bio

How do we share large datasets?

Traditional Approach

Bring data to researchers



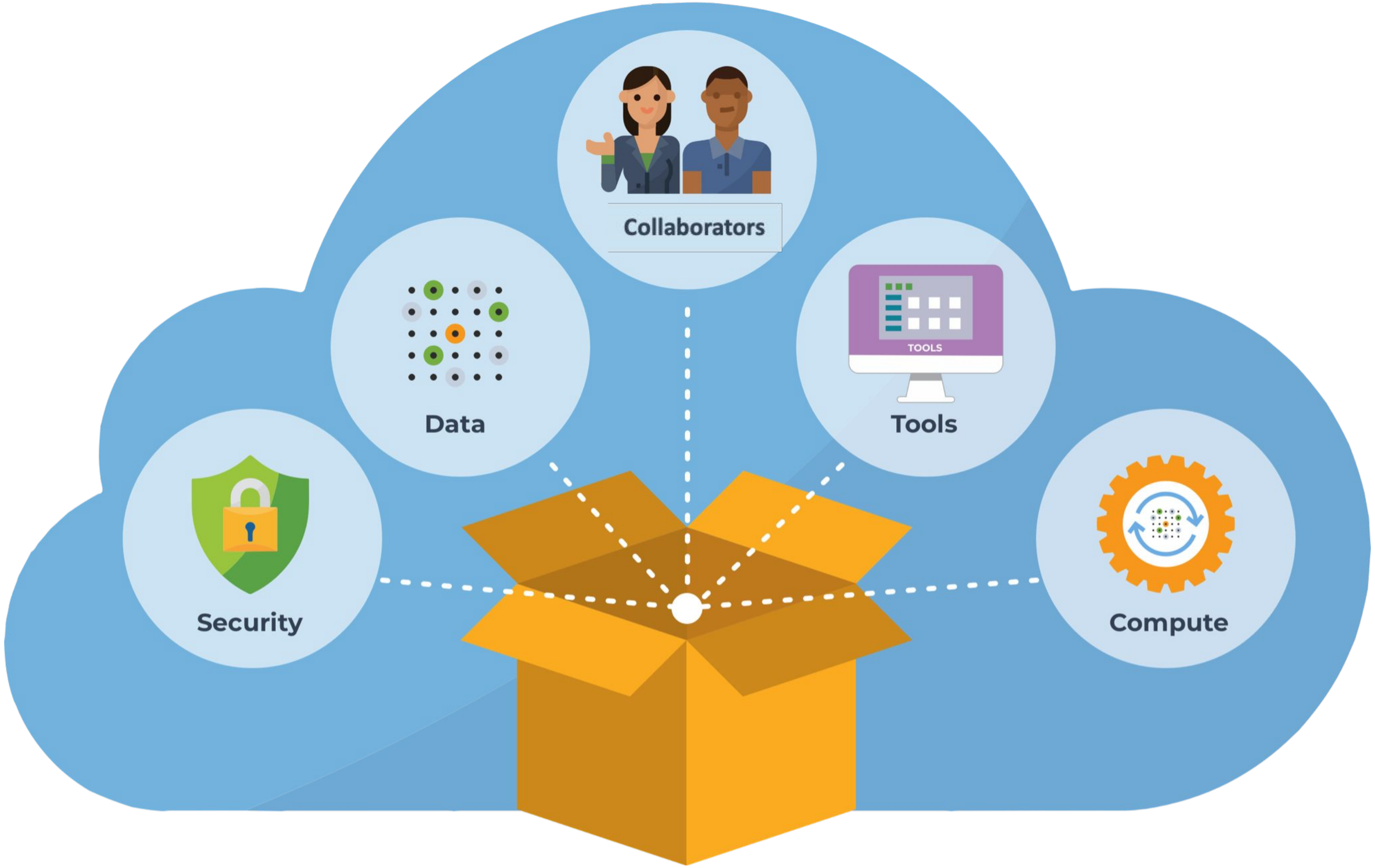
https://terra.bio

How do we share large datasets?

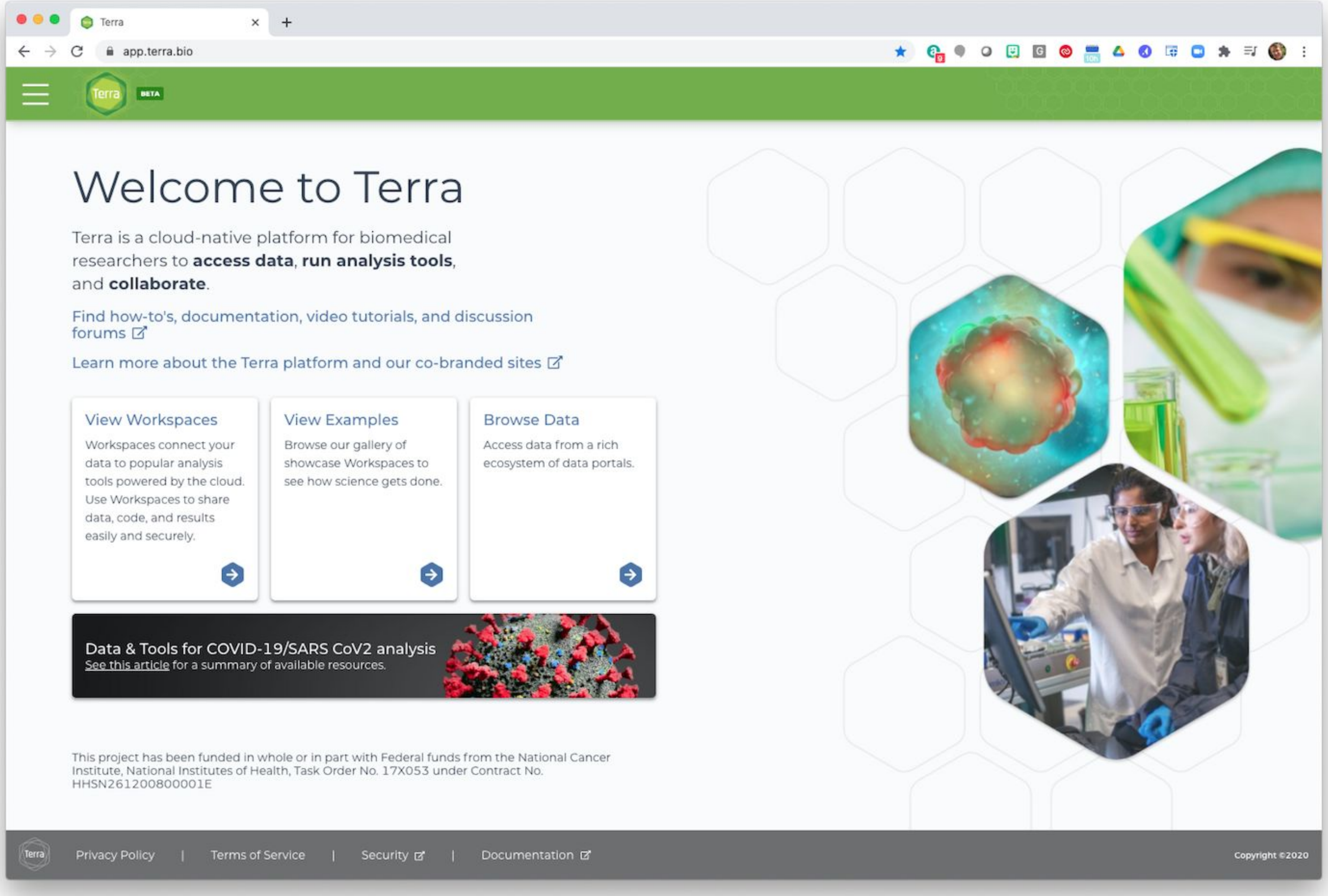
Cloud-centric Approach

Bring researchers to data

Clouds are hard to use out of the box



Terra



Welcome to Terra

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

View Workspaces
Workspaces connect your data to popular analysis tools powered by the cloud. Use Workspaces to share data, code, and results easily and securely.

View Examples
Browse our gallery of showcase Workspaces to see how science gets done.

Browse Data
Access data from a rich ecosystem of data portals.

Data & Tools for COVID-19/SARS CoV2 analysis
See this article for a summary of available resources.

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 17X053 under Contract No. HHSN261200800001E

BROAD INSTITUTE

Microsoft

verily

BROAD INSTITUTE

https://terra.bio

Core capabilities designed to support research

Data Library

Access public and access-controlled datasets

Workspaces

Bring together data and tools into secure, shareable units

Workflows

Run workflows at scale; bring your own or explore community favorites

Interactive Analysis

Analyze data with built-in applications like Jupyter Notebooks, RStudio, Galaxy

A hub in the cloud data ecosystem

https://terra.bio

Built secure to work with highly sensitive data

- ✅ Human genomes
- ✅ Clinical data
- ✅ Many other data types

FISMA Moderate
FedRAMP Moderate

FISMA
FEDERAL INFORMATION SECURITY MANAGEMENT ACT

FR
FedRAMP

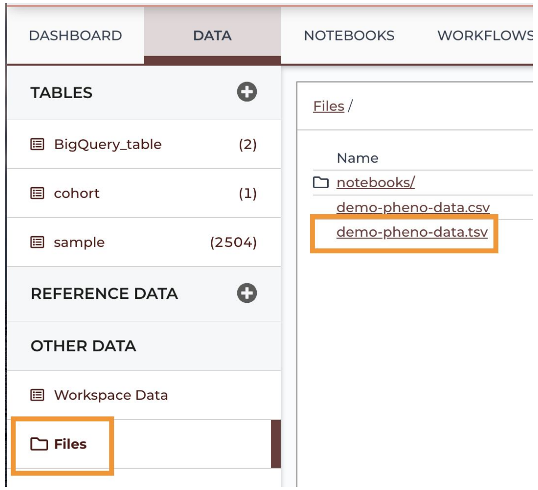
Get started today

https://terra.bio/resources/getting-started

https://terra.bio/get-started-with-free-cloud-credits

Notebook 2 Exercise 1 (5 min)

1. Go to the original Quickstart workspace (https://elwazi.terra.bio/#workspaces/fc-product-demo/Terra-Notebooks-Quickstart)
2. Select the Data page
3. Select the Files option (bottom left)



4. Download the TSV file to your computer
5. Rename the file with a unique name
6. Upload the file to your **cloned** workspace Files section using the plus icon
7. Raise your hand on zoom when done

Notebook 3 Exercise 1 (3 min)

1. Go to the Terra main menu → Library → Data
2. Under the 1000 Genomes Low Coverage option, select Browse Data
3. Pick a cohort
4. Save cohort with a simple unique name
5. Choose your clone workspace as the destination workspace
6. Raise your hand on zoom when done

Terminal Exercise 1 (3 min)

1. Select the terminal icon on your Cloud Environment widget
2. Type “pwd” and enter to see directory
    i. You should see the home directory /home/jupyter
3. Type “ls” to list the files
    a. Example result:

```
jupyter@b73f1550a0fa:~$ ls
entrypoint.out  jupyter.log  lost+found  packages  Terra-Notebooks-Quickstart-lk-20220330_test1  welder.log
```

4. Use cd to change directories
    a. Use “cd ..” to move back up a directory OR
    b. use “cd ~” to go back to the home directory
5. Take 3 min to explore:
    a. Can you find the Jupyter Notebook files?
    b. Can you find the files we copied into our VM from the Jupyter Notebooks?

Terminal Exercise 2 (5 min)

1. Return to the home/Jupyter/workspace/edit folder

2. Copy a file into the Cloud Environment using gsutil
   a. Copy this file: gs://terra-featured-workspaces/ASHG_January_2022/hg004.fq.gz
   b. What's the command to use?

3. After you copy the file, use "ls" to make sure it's in your environment

Exercise: Copying Notebooks into Workspace (5 min)

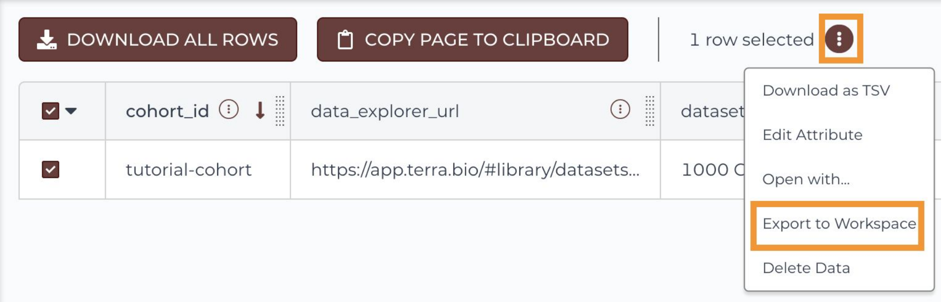
1. Navigate to the ASHG workspace Notebook: https://elwazi.terra.bio/#workspaces/ASHG2021/2021-ASHG-Galaxy-SV-Discovery/notebooks
2. Select the vertical three-dot icon next to the Notebook
3. Select Copy to Another Notebook
4. Choose your cloned workspace and export to the workspace
    a. This moves the .ipynb file to your workspace Google Bucket
5. Open the Notebook in edit mode
    a. This moves the .pynb file to your virtual computer
6. Restart kernel and run all cells
7. If time remains, search another Notebook that interests you
8. Raise hand when done

Exercise: Copying data into a Workspace (5 min)

1. Navigate to Terra showcase section

2. Find a workspace of interest and explore it

3. Copy data from the data table into your cloned workspace
   a. Select the checkbox next to the data of interest
   b. Use the three-dot icon to copy to another workspace
   c. Choose your clone of the Notebooks quickstart



☑	cohort_id	data_explorer_url	dataset
☑	tutorial-cohort	https://app.terra.bio/#library/datasets...	1000 C

4. Note the link to the workspace that interested you - we will share it when we regroup

Additional Resources

- Leanpub Courses
  - Intro to Terra
  - Billing and Secure Collaboration
  - Data Tables
  - Pipelining with Workflows
  - Introduction to the Cloud Environment
- Terra Support
  - Getting started
  - Terra Youtube videos

Survey

Please take our survey (5 min):

https://www.surveymonkey.com/r/NMSWXYY

Q&A

Questions?

Appendix

https://terra.bio

Foundational design principles

Open-source

Standards-based

Modular

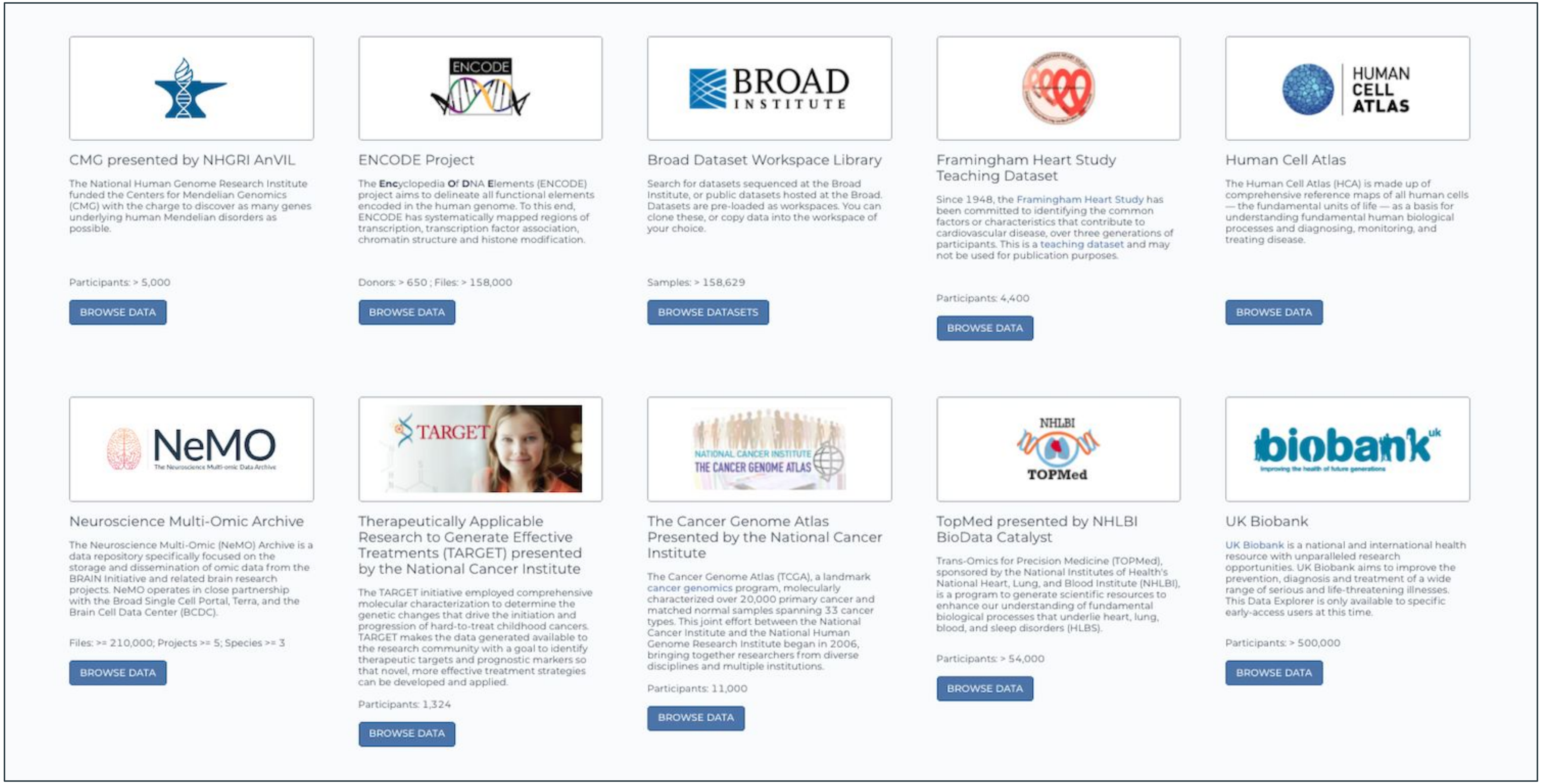
Community-driven

DATA BIOSPHERE

Global Alliance for Genomics & Health

NCPI

Access to a rich catalog of data hosted by various organizations



CMG presented by NHGRI AnVIL

The National Human Genome Research Institute funded the Centers for Mendelian Genomics (CMG) with the charge to discover as many genes underlying human Mendelian disorders as possible.

Participants: > 5,000

ENCODE Project

The **Enc**yclopedia **O**f **D**NA **E**lements (ENCODE) project aims to delineate all functional elements encoded in the human genome. To this end, ENCODE has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification.

Donors: > 650 ; Files: > 158,000

Broad Dataset Workspace Library

Search for datasets sequenced at the Broad Institute, or public datasets hosted at the Broad. Datasets are pre-loaded as workspaces. You can clone these, or copy data into the workspace of your choice.

Samples: > 158,629

Framingham Heart Study Teaching Dataset

Since 1948, the Framingham Heart Study has been committed to identifying the common factors or characteristics that contribute to cardiovascular disease, over three generations of participants. This is a teaching dataset and may not be used for publication purposes.

Participants: 4,400

Human Cell Atlas

The Human Cell Atlas (HCA) is made up of comprehensive reference maps of all human cells — the fundamental units of life — as a basis for understanding fundamental human biological processes and diagnosing, monitoring, and treating disease.

Neuroscience Multi-Omic Archive

The Neuroscience Multi-Omic (NeMO) Archive is a data repository specifically focused on the storage and dissemination of omic data from the BRAIN Initiative and related brain research projects. NeMO operates in close partnership with the Broad Single Cell Portal, Terra, and the Brain Cell Data Center (BCDC).

Files: >= 210,000; Projects >= 5; Species >= 3

Therapeutically Applicable Research to Generate Effective Treatments (TARGET) presented by the National Cancer Institute

The TARGET initiative employed comprehensive molecular characterization to determine the genetic changes that drive the initiation and progression of hard-to-treat childhood cancers. TARGET makes the data generated available to the research community with a goal to identify therapeutic targets and prognostic markers so that novel, more effective treatment strategies can be developed and applied.

Participants: 1,324

The Cancer Genome Atlas Presented by the National Cancer Institute

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between the National Cancer Institute and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

Participants: 11,000

TopMed presented by NHLBI BioData Catalyst

Trans-Omics for Precision Medicine (TOPMed), sponsored by the National Institutes of Health's National Heart, Lung, and Blood Institute (NHLBI), is a program to generate scientific resources to enhance our understanding of fundamental biological processes that underlie heart, lung, blood, and sleep disorders (HLBS).

Participants: > 54,000

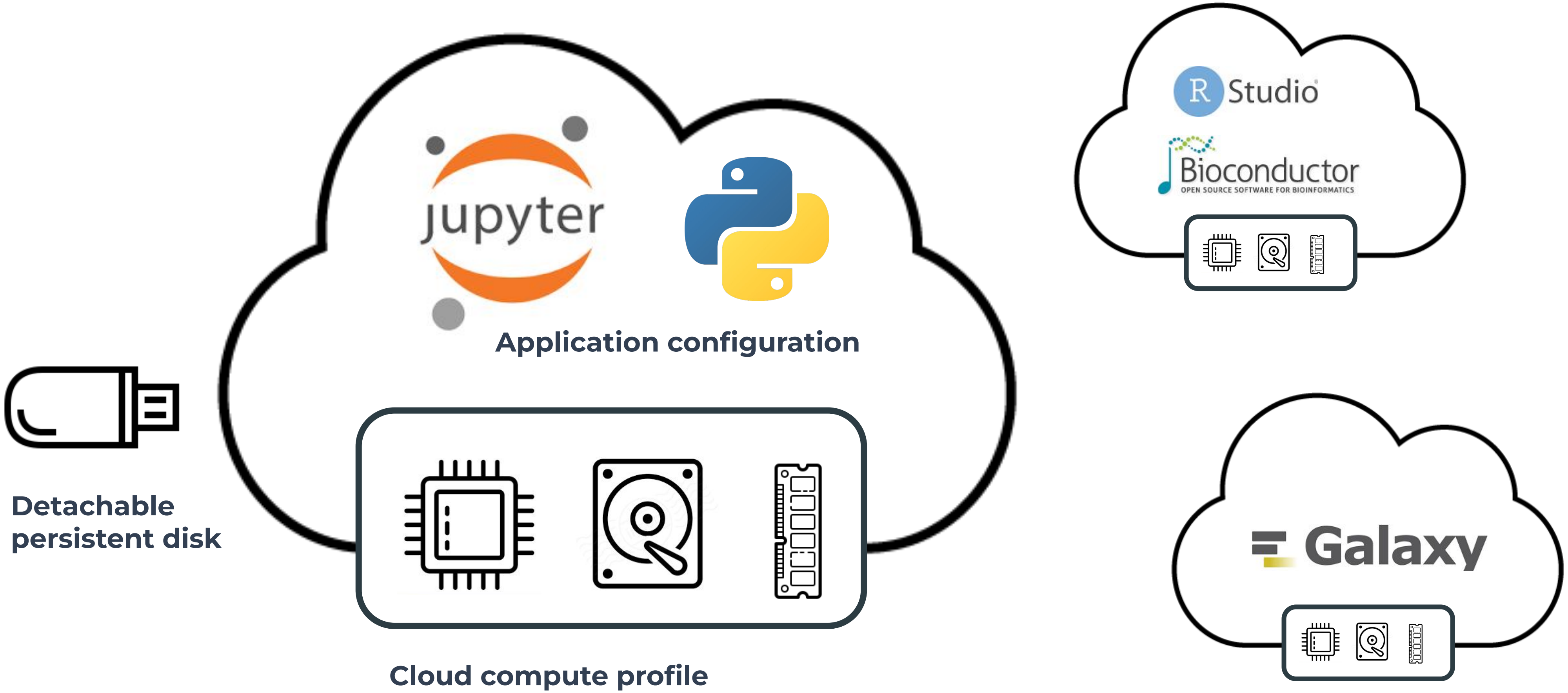
UK Biobank

UK Biobank is a national and international health resource with unparalleled research opportunities. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses. This Data Explorer is only available to specific early-access users at this time.

Participants: > 500,000

https://terra.bio

A highly flexible system of customizable cloud environments for interactive analysis



Application configuration

Detachable persistent disk

Cloud compute profile
Processor / Local storage / Memory