[Duke Datathon 2025 - FAQ](https://docs.google.com/document/d/14KA0PuOynUHldiraHtaca9_6LVwmgBE2OKaFHMmJEfs/edit?tab=t.0) (this document)

## Frequently Asked Questions (FAQ)

[**Frequently Asked Questions (FAQ) 1**](#_8q32vcnvkg8h)

[Link to videos 3](#_nf0zgw1jmsa9)

[Access to data 4](#_9wpjtrw9smmf)

[PhysioNet data 5](#_9ru09veawfp4)

[Key links 5](#_c88buu19jmm6)

[How do I sign up for MIMIC-III, MIMIC-IV, eICU? 5](#_ntgw4y3gzwc3)

[I can’t figure out how to sign up for a Code of Conduct on PhysioNet. 5](#_lgff3s821sfo)

[I don’t have an MIT email address, can I still do the CITI under MIT affiliates? 6](#_p7aqwyev0frj)

[Where do I submit my completed training to PhysioNet? 7](#_qnfn2qj4luw5)

[I don’t have a PhysioNet account; is there another way? 8](#_o4kaxr1wgocm)

[CHoRUS data 9](#_e7yhb8gmbrs2)

[Common links 9](#_3v134qxvb3dj)

[What is the structure of the CHoRUS B2AI Data Enclave? 9](#_5obxo53tpec6)

[Where do I go to log in? 10](#_cwg3tzxjzeph)

[I’ve logged in, now what? 11](#_jiqq3bnc2o1r)

[I made it to the ML Workspace, how do I get my hands dirty? 12](#_izlec17njc4i)

[I’ve launched my own JupyterLab environment, where’s the data? 15](#_hbfhikslr2sr)

[Are there other ways to explore the data? 17](#_8k6xznuurihy)

[How do I train models created in Atlas in my compute instance? 18](#_9rb64ebu4hzd)

[Seoul National University (K-MIMIC data) 19](#_50vx1z1q31at)

[Key links 19](#_5ut0o4ygdz1q)

[NSTRI Platform Structure for K-MIMIC 19](#_8yagr3jsuz60)

[User Registration & Login 19](#_ntm4vw6vpczb)

[Accessing the Cloud Environment & Launching JupyterLab 20](#_tp1s04rabu6w)

[Development Environment Specifications 21](#_deruflbvevsg)

[Dataset Structure 21](#_2z2mx4k9zjx6)

[Outbound Whitelist 22](#_hq7tsyofop01)

[Important Notes 22](#_4yor3wnohpgl)

[Epic Cosmos 23](#_v3e98a8fe2sd)

[Key links 23](#_yogvw1vgzixk)

[I’m a Duke person, where do I go to register? 23](#_tm7093xf20rm)

[Can you provide examples of what work has been done in Cosmos? 23](#_k4r7awy8t8ws)

[What is Cosmos? 23](#_uco1suigb3ui)

[What data is available in Cosmos? 23](#_c3fiddv8okvu)

[How do you get access to Cosmos and Training Resources? 23](#_ls6kbv6l1c9j)

[Available Analytical Tools 23](#_jho06szayrp4)

[Cosmos Training Resources 24](#_fvit55mkrenp)

[General data questions 25](#_dtqfswxfzb07)

[Can I bring my own dataset? 25](#_51d3iq3w9b0a)

[Access to GitHub 26](#_245sco2mrqm1)

[I don’t have a GitHub ID 26](#_m3jtpt7fnkro)

[Datathon registration 27](#_x3ura2t8qsqv)

[What is the purpose of the trainee recommendation letter? What does it need to say? 27](#_53p2duh8pglm)

[Datathon operations 28](#_vu3pf85mjqw9)

[Can my team change what topic we research? 28](#_t1i8ppd70bay)

[I’m clinical, and have never done this before. Can I join? 28](#_xokscadcgbku)

[How do you define success for the datathon? 28](#_hks1h5uksn7h)

[Can I pick what team I’m in? 28](#_6exexjhuxgv)

[Can I change teams at the datathon? 28](#_a5y6ppr9hrof)

[Judging criteria 29](#_w2vh0uqzhflo)

[How will we be judged? 29](#_r3xpndlptsju)

[Running the example Colab notebook 30](#_9bcqwopbt364)

[I can’t run the cells 30](#_z2o8yytzjt5d)

[The Google Colab doesn’t have access to MIMIC / eICU 30](#_ycqms48oy9if)

[What is the presentation format? 31](#_a9lor9wjqodq)

[What are the judging criteria? 31](#_il7plbqybibn)

[What is the presentation order? 31](#_49z6kgojfzfj)

[Where can I send this abstract/paper? 31](#_uzvt7am10rb3)

[Link: conferences - healthcare data science + pulmonary/critical care 31](#_kjjxb9vnpw67)

[Unanswered questions 32](#_u4ja2841iyzt)

[**Frequently used acronyms 33**](#_s1zvbp9q70ci)

[D 33](#_7xyonqc7o2cb)

[E 33](#_pepb8obd6t9p)

[I 33](#_kmpna57t1ydu)

[S 33](#_43v6zhikm16d)

[**What resources / things would be helpful to know for next year? 34**](#_nezgk07fi7dw)

[For first-time data scientists 34](#_6ls5r6upjax6)

[For first-time clinicians 34](#_eecbozul3m5h)

[For basic scientists 34](#_v4rs231vhdah)

### 

### 

### Link to videos

* [How to set up Physionet (MIMIC, eICU) for BigQuery](https://www.youtube.com/watch?v=N6fqg2QaNN8&pp=ygUVInBoeXNpb25ldCIgYmlncXVlcnkg)
* [How to Register for GitHub (and get a GitHub ID)](https://youtu.be/uL5QCAmRbVI)
* [How to analyze PhysioNet EHR data with Jupyter/BigQuery](https://youtu.be/VDyw4ONIxMI)

### Access to data

The current data sets are:

* [MIMIC-III](https://physionet.org/content/mimiciii/1.4/)
* [MIMIC-IV](https://physionet.org/content/mimiciv/2.2/)
  + MIMIC-IV+ (extension to ) - internally offered with approval of MIT LCP
* [eICU-CRD](https://physionet.org/content/eicu-crd/2.0/)
  + eICU-CRD 2 (extension to ) - internally offered with approval of MIT LCP
* [CHoRUS](https://chorus4ai.org/)
* K-MIMIC

Affiliated data sets are acceptable, although there are no increased compute resources (e.g., for chest x-ray or waveform data) guaranteed:

* [MIMIC-CXR](https://physionet.org/content/mimic-cxr/2.0.0/)
* [MIMIC-IV-ED](https://physionet.org/content/mimic-iv-ed/2.2/)
* [MIMIC-IV Waveform](https://physionet.org/content/mimic4wdb/0.1.0/)
* [MIMIC-IV-ECHO](https://physionet.org/content/mimic-iv-echo/0.1/)

Derivative datasets are acceptable, e.g.,

* [BOLD, a blood-gas and oximetry linked dataset](https://physionet.org/content/blood-gas-oximetry/1.0/)
* Hourly MIMIC-IV data
* [Curated Data for Describing Blood Glucose Management in the Intensive Care Unit](https://physionet.org/content/glucose-management-mimic/1.0.1/)

All teams must have access to all data sets.

Duke Datathon 2025 caveat: This *may* mean that there are people on each team that does not have access to a resource (e.g., Cosmos SlicerDicer) as long as some people have access to all datasets on a team.

#### 

### PhysioNet data

#### Key links

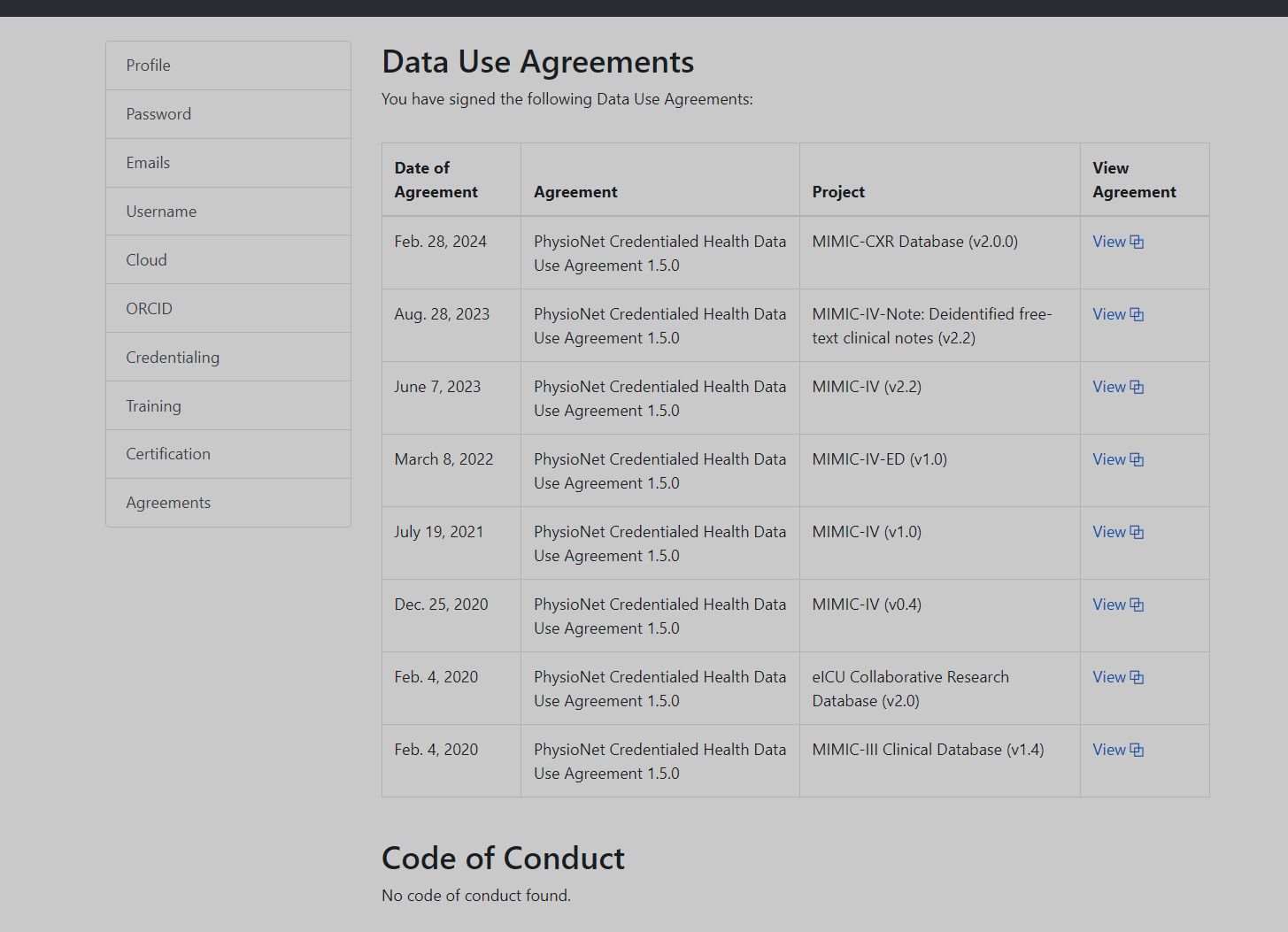
* <https://physionet.org/>

#### How do I sign up for MIMIC-III, MIMIC-IV, eICU?

Great question. We’ve created a video for you. [How to set up Physionet (MIMIC, eICU) for BigQuery](https://www.youtube.com/watch?v=N6fqg2QaNN8&pp=ygUVInBoeXNpb25ldCIgYmlncXVlcnkg)

#### I can’t figure out how to sign up for a Code of Conduct on PhysioNet.

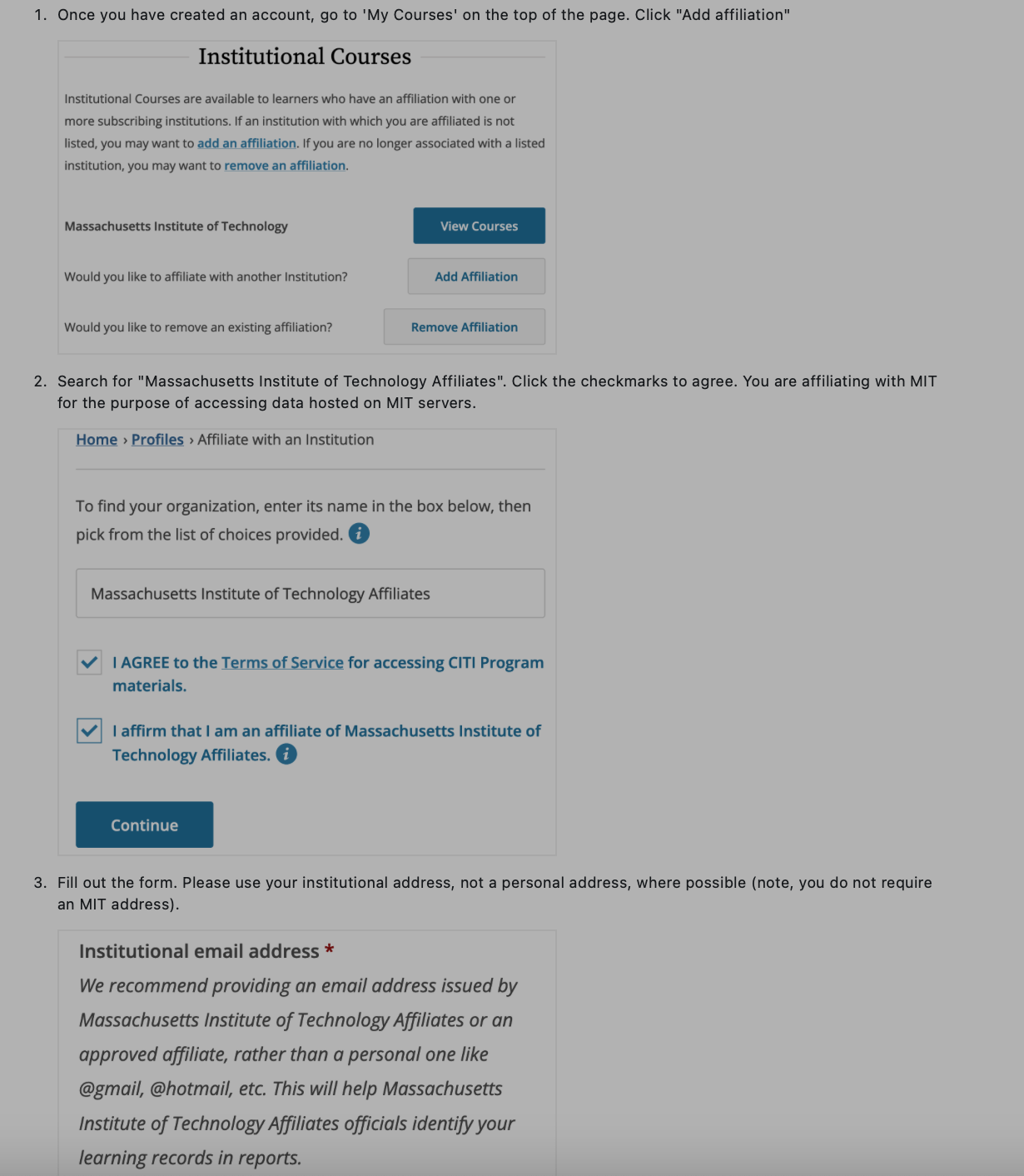
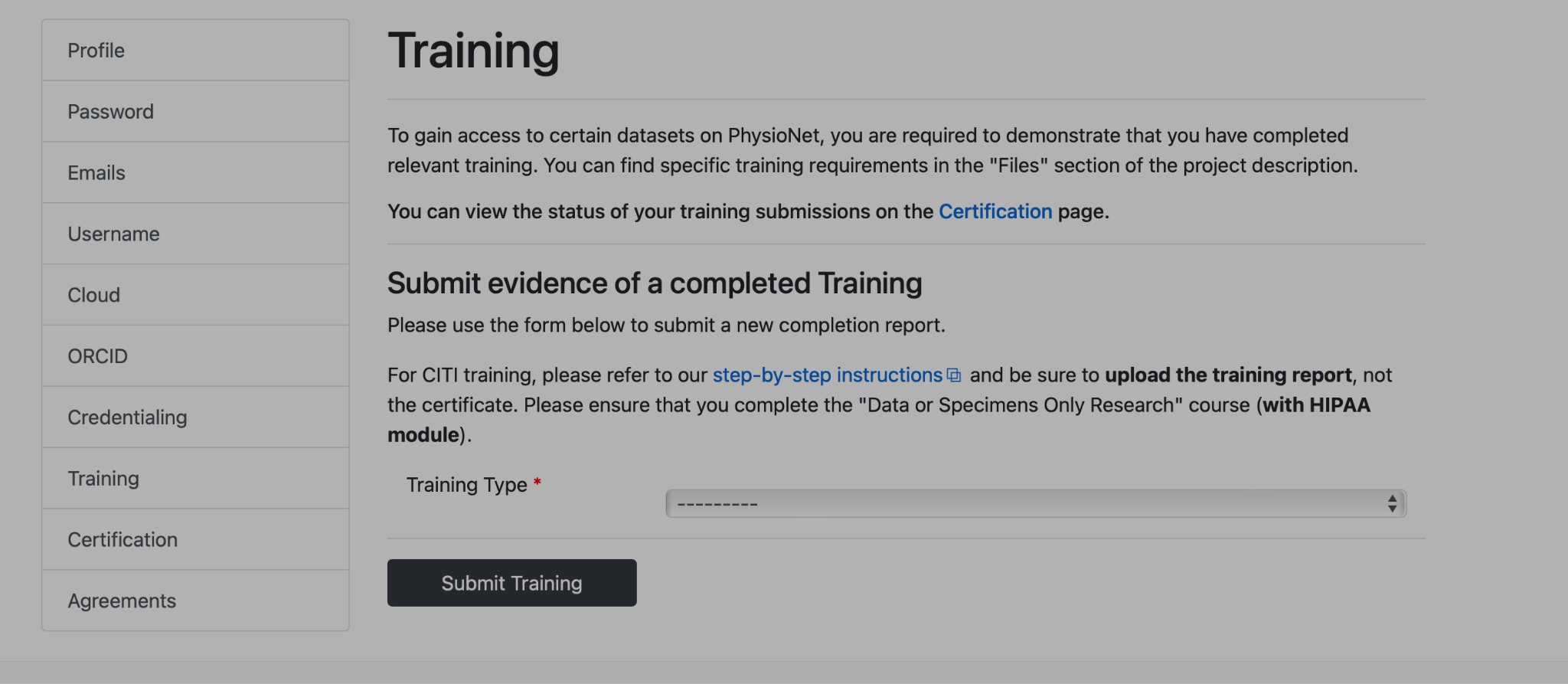
Great! You don’t need it. Here is a screenshot of Ian’s PhysioNet DUAs and Code of Conduct page as an example. Lots of DUAs. No code of conduct form seems to be needed, although Leo and Tom will probably tell Ian if he’s forgetting something.



#### 

#### I don’t have an MIT email address, can I still do the CITI under MIT affiliates?

Yes! It asks for a preferred *institutional* account (e.g., Duke.EDU, unc.EDU, etc.), but your personal account could work if you have no other options.



#### 

#### Where do I submit my completed training to PhysioNet?

Please go to your account in PhysioNet, then go to the ‘training’ tab. You’ll see it here:

#### 

#### I don’t have a PhysioNet account; is there another way?

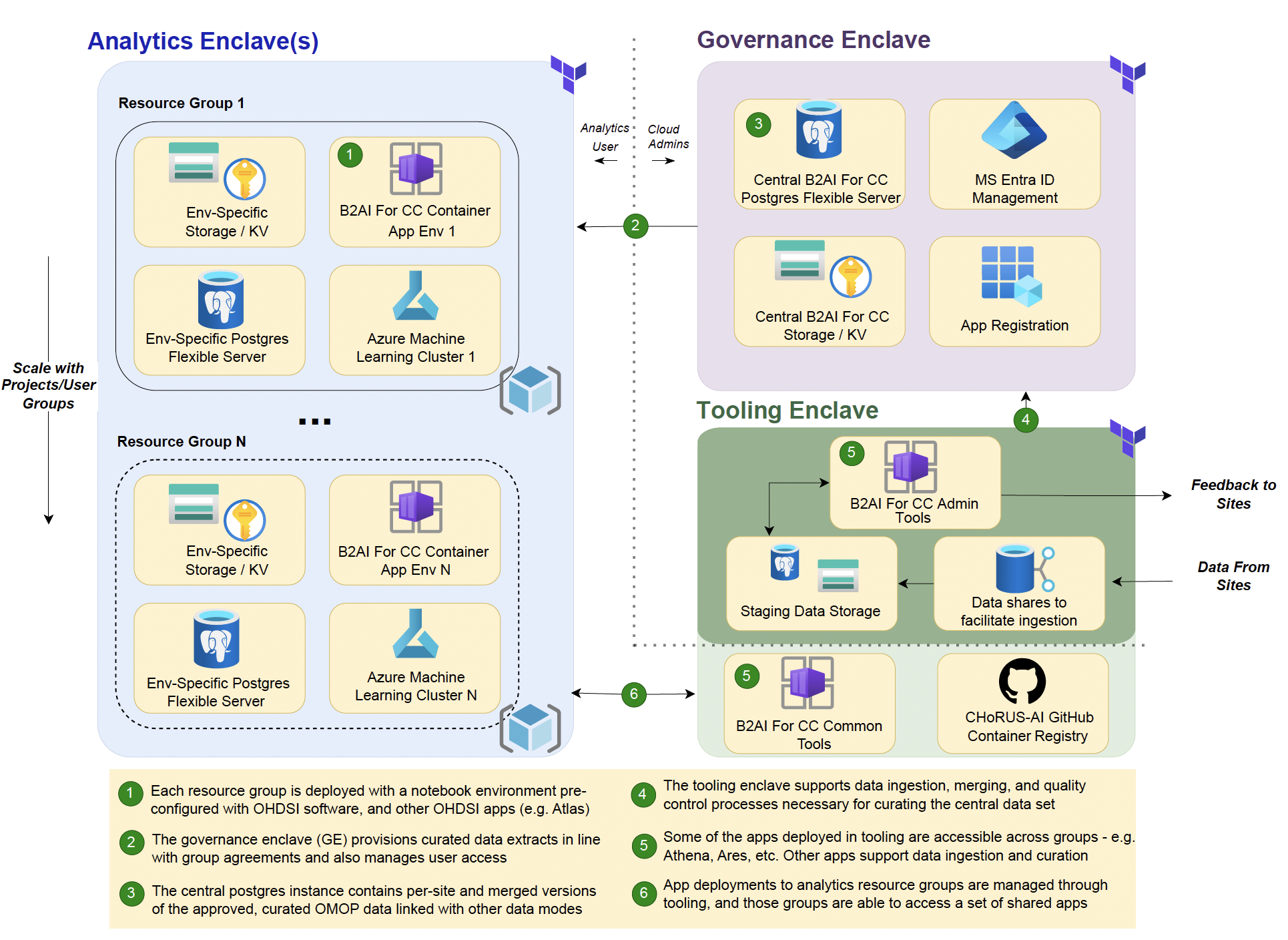
There *may* be an option (not guaranteed): <https://docs.google.com/forms/d/e/1FAIpQLSfp9otOVQdwsyxAHqEtpmsZTV72Fhjg2gEE4MuREyweugU6Fg/viewform>

### CHoRUS data

#### Common links

* Schema: <https://ohdsi.github.io/CommonDataModel/cdm54.html>
* portal: <https://ca-dukedatathon-www.blackdesert-9aaf2ae7.eastus2.azurecontainerapps.io/>

#### What is the structure of the CHoRUS B2AI Data Enclave?



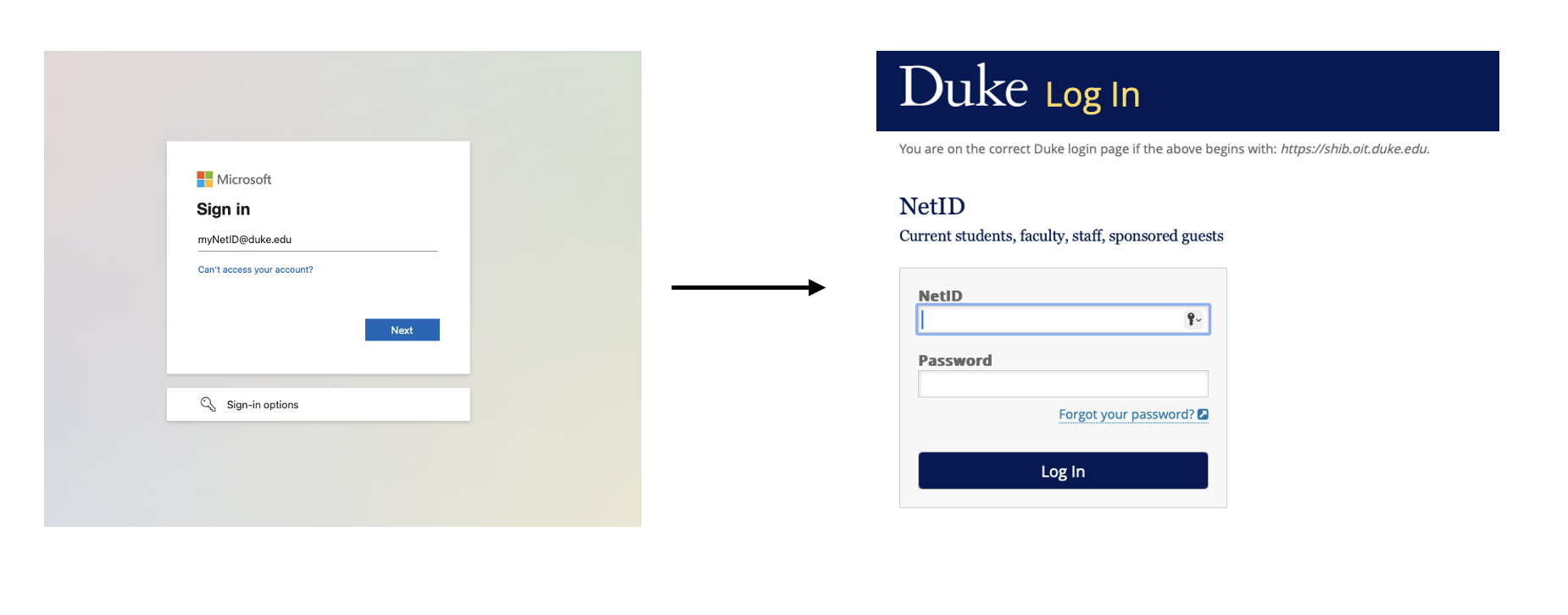
The Bridge2AI for Clinical Care (B2AI For CC, or CHoRUS) collaborative cloud consolidates rich multi-center and multimodal data in order to support complex analytic processes in machine learning (ML) and artificial intelligence (AI); such consolidation and analytic support is nontrivial and requires a flexible and scalable architecture design. We’ve devised a three-enclave approach (see above) that addresses the anticipated diversity of researcher profiles and analytic needs, and that ingests and combines large volumes of data from a heterogeneous network of fourteen healthcare centers.

#### 

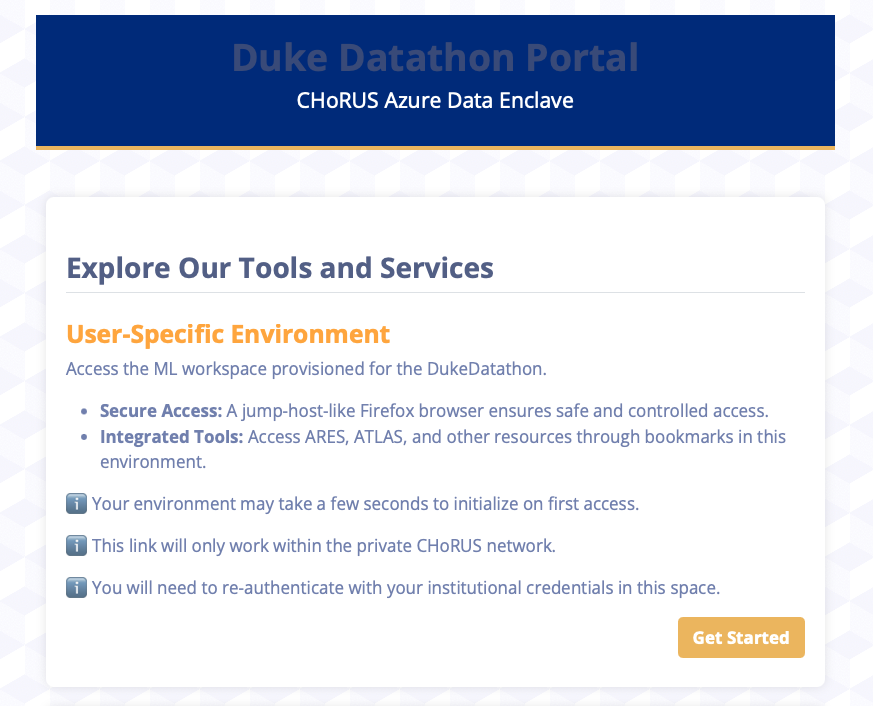
#### Where do I go to log in?

Please go here: <https://ca-dukedatathon-www.blackdesert-9aaf2ae7.eastus2.azurecontainerapps.io/>

You should arrive at a familiar Microsoft login screen. Here, you’ll enter your **DUKE** credentials (e.g. [myNetID@duke.edu](mailto:myNetID@duke.edu)) and you will be redirected to the Duke Shibboleth for a standard login.



Once successfully logged in, you’ll arrive at the CHoRUS Duke Datathon landing page:



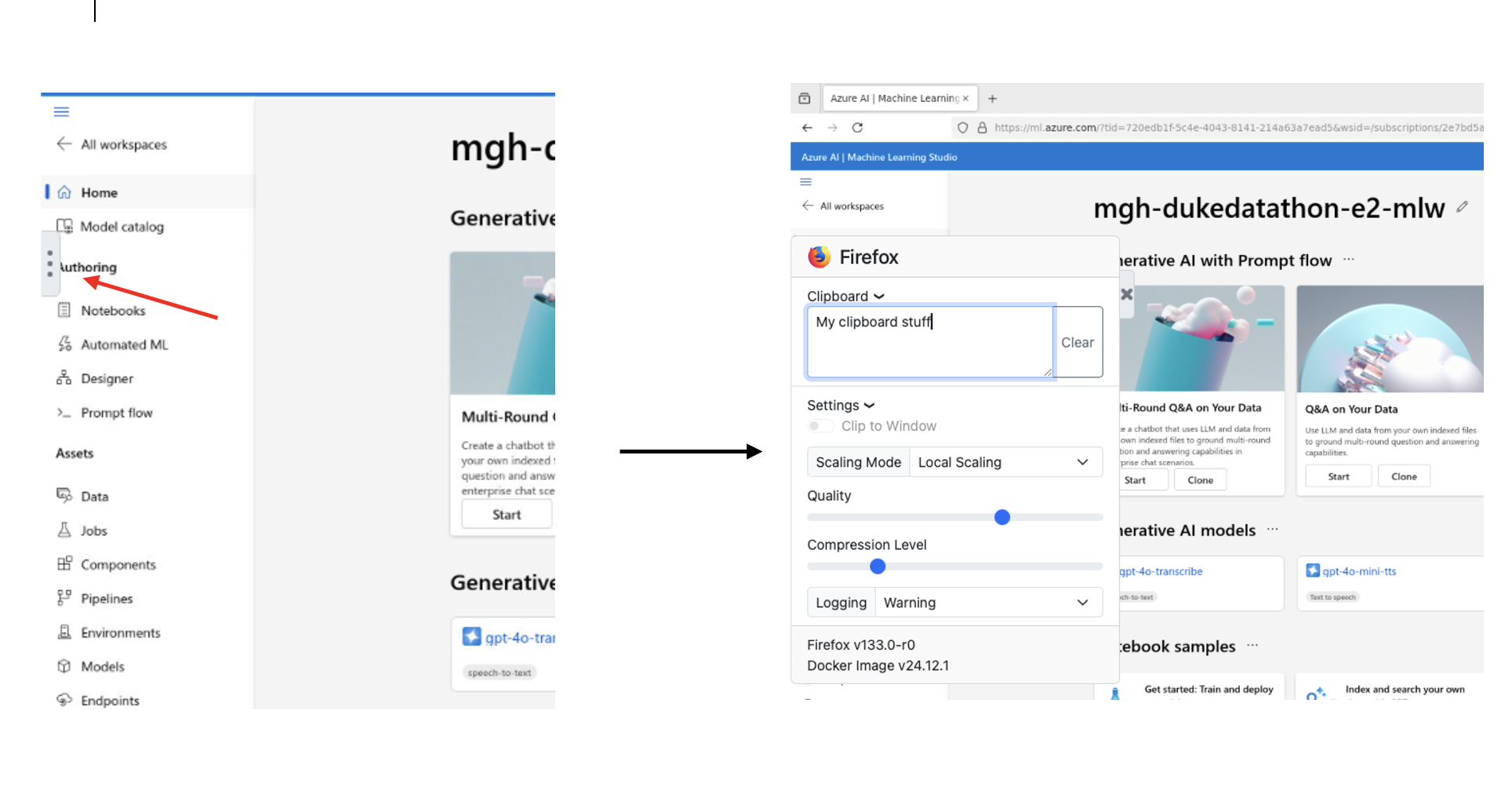
#### 

#### 

#### 

#### I’ve logged in, now what?

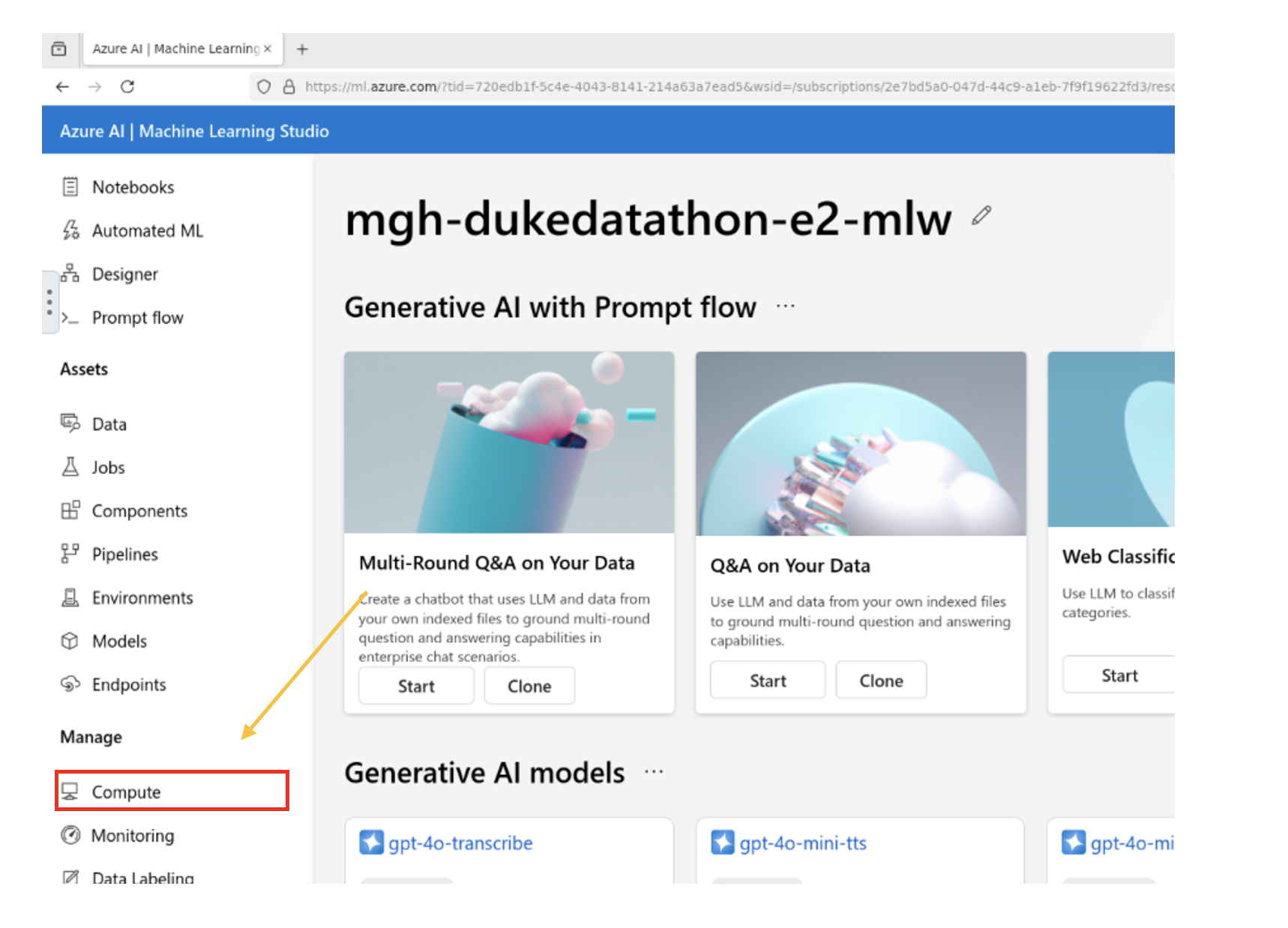
Click ‘Get Started’ on the landing page to launch your own personal virtual machine running a Firefox browser that you can interact with through VNC. The Firefox browser will launch to the URL for a Duke-Datathon-Specific Azure Machine Learning workspace. **Here, you will need to re-authenticate just like you did above, and you will be redirected to the Duke Shibboleth as before.** Note that if you are on a Mac, the control key (rather than the CMD key) will work for accessing the clipboard inside your VM, and you can pass content (e.g. passwords) into the virtual machine through the clipboard widget with shown with three dots on the left side of your window.

****

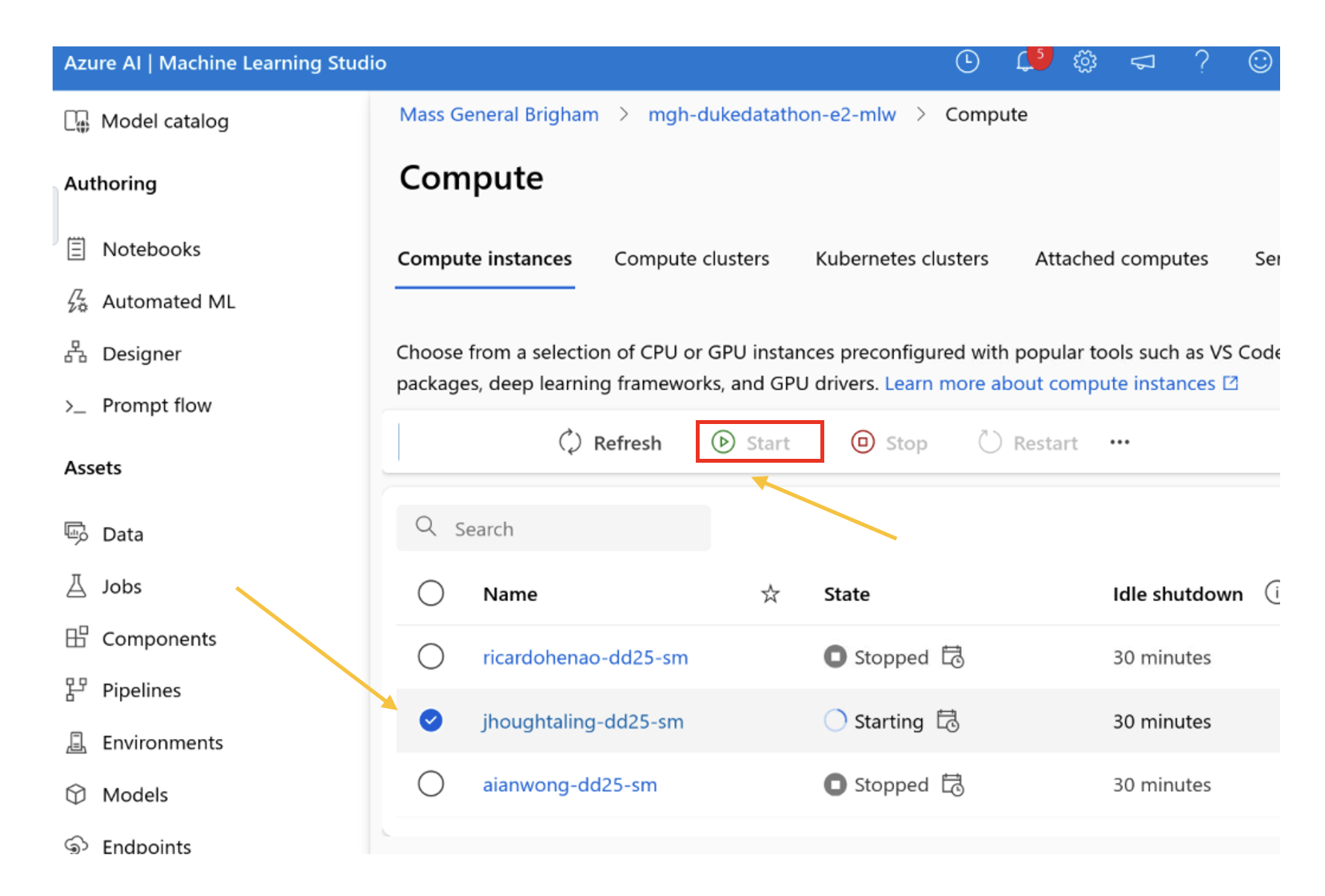
#### 

#### I made it to the ML Workspace, how do I get my hands dirty?

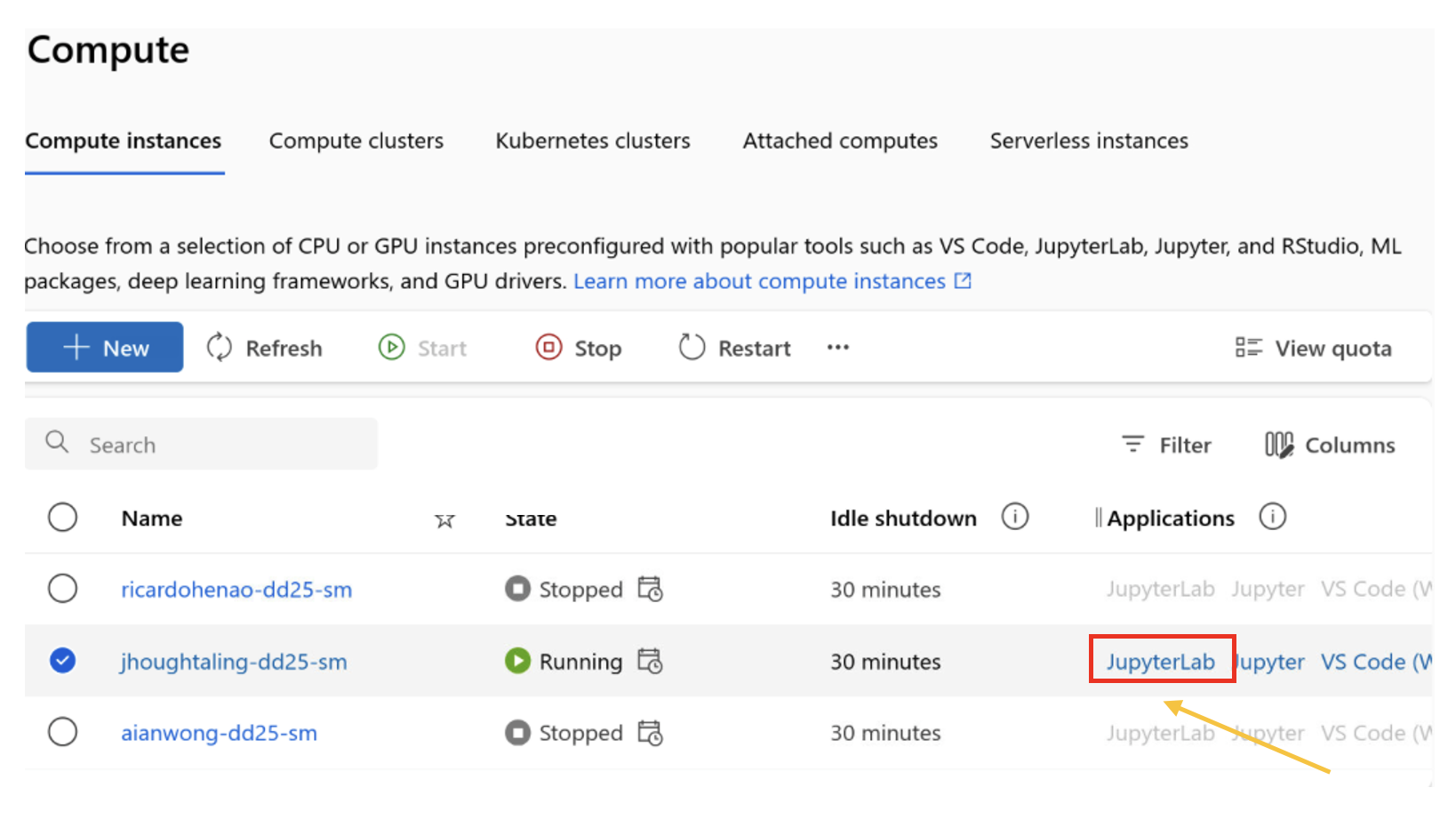
Once in the ML workspace, you’ll need to navigate to your own *personal compute instance*. You’ll first select compute on the left navigation bar:



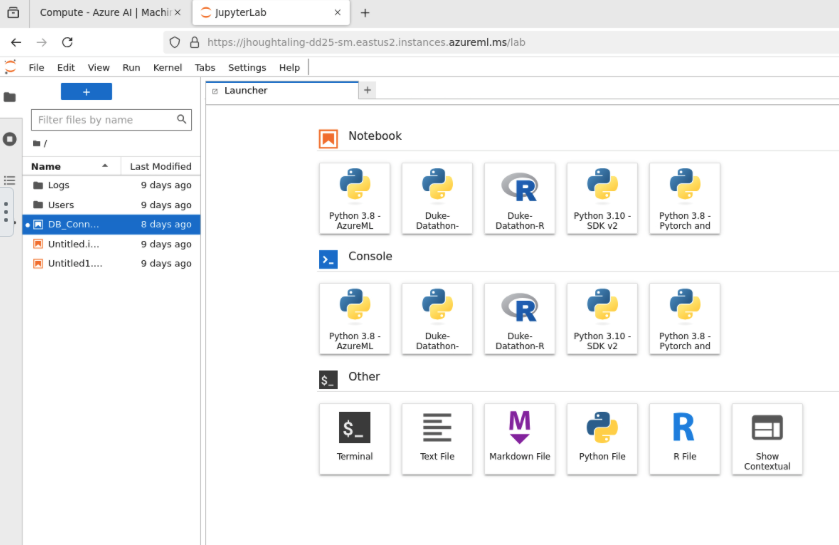
Clicking this icon will bring up a list of computes for the Datathon. Scroll to find the instance assigned to you, click it to highlight the check icon on the left, and then click the green ‘Start’ icon above:



Once your compute has started (it will shut down automatically after 30 minutes of inactivity), you’ll see a list of potential applications hosted on your compute instance; select ***JupyterLab*.**

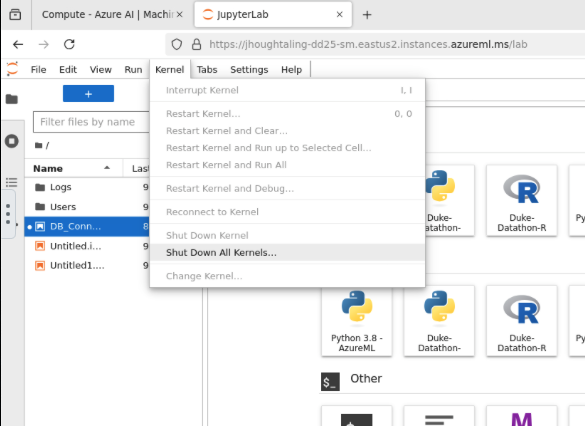
****

Clicking JupyterLab will launch a new window on your VM for JupyterLab. Note that if you’re too quick on the draw, you may get an error on startup. The JupyterLab application takes about 30 seconds to be available once the compute instance gives a running status. Inside JupyterLab you’ll see the following:



To work around some interesting constraints Microsoft has placed on the ML workspace compute deployment, we have created two Docker images to support custom kernels in this space. One image contains a number of Python packages relevant to observational health analytics, and the other contains the entire suite of OHDSI R packages for working with data in the OMOP CDM. Note that you can view the definitions of these images, along with all of the other components of the infrastructure, here: <https://github.com/chorus-ai/chorus-container-apps/tree/main/mlw-jupyter-kernel>

Also note, we recommend keeping track of, and regularly shutting down, your kernels when you are not running jobs to mitigate potentially hanging processes with Docker-based kernel management:



#### 

#### 

#### 

#### 

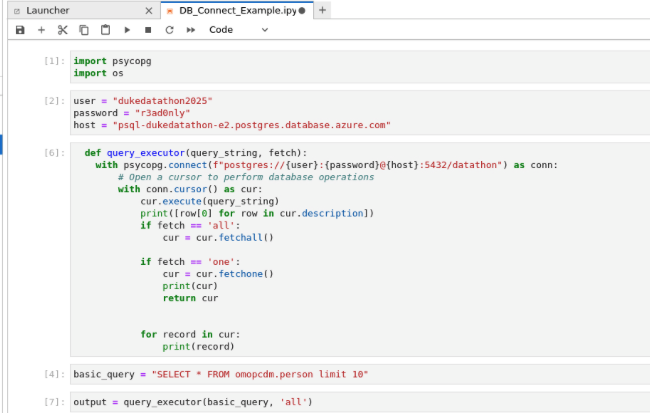
#### 

#### 

#### 

#### I’ve launched my own JupyterLab environment, where’s the data?

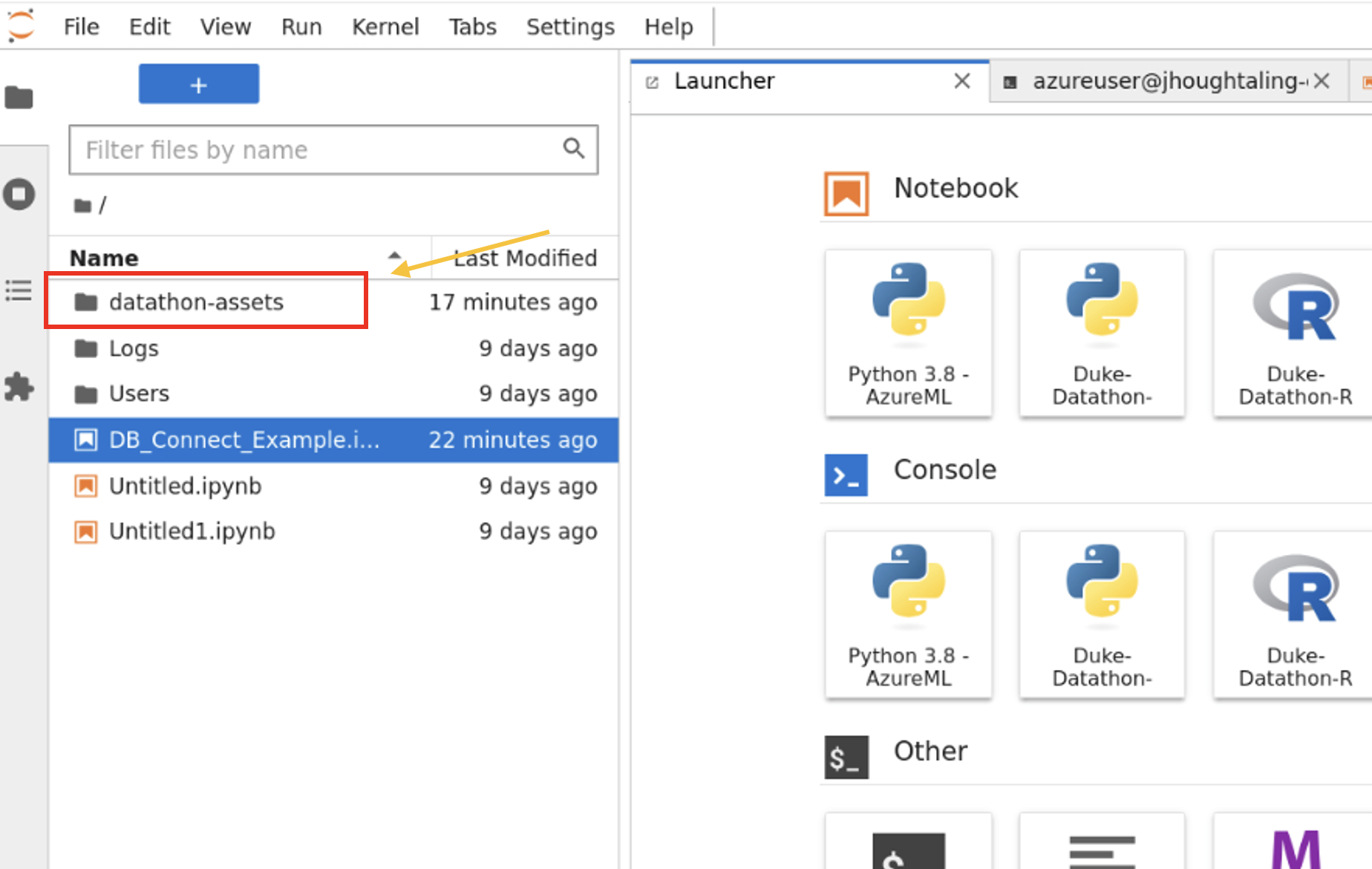
For OMOP-shaped tabular data, we’ve provisioned a common datathon Postgres instance that contains information for 18’000 individuals. You can connect to this database (datathon) using `psychopg` in Python, or `DatabaseConnector` in R, via a designated postgres user for the datathon. Note that the dukedatathon2025 user has read-only access to the omopcdm schema (where the OMOP data is located) and has read/write access to the public schema in case you need scratch tables for certain processes:



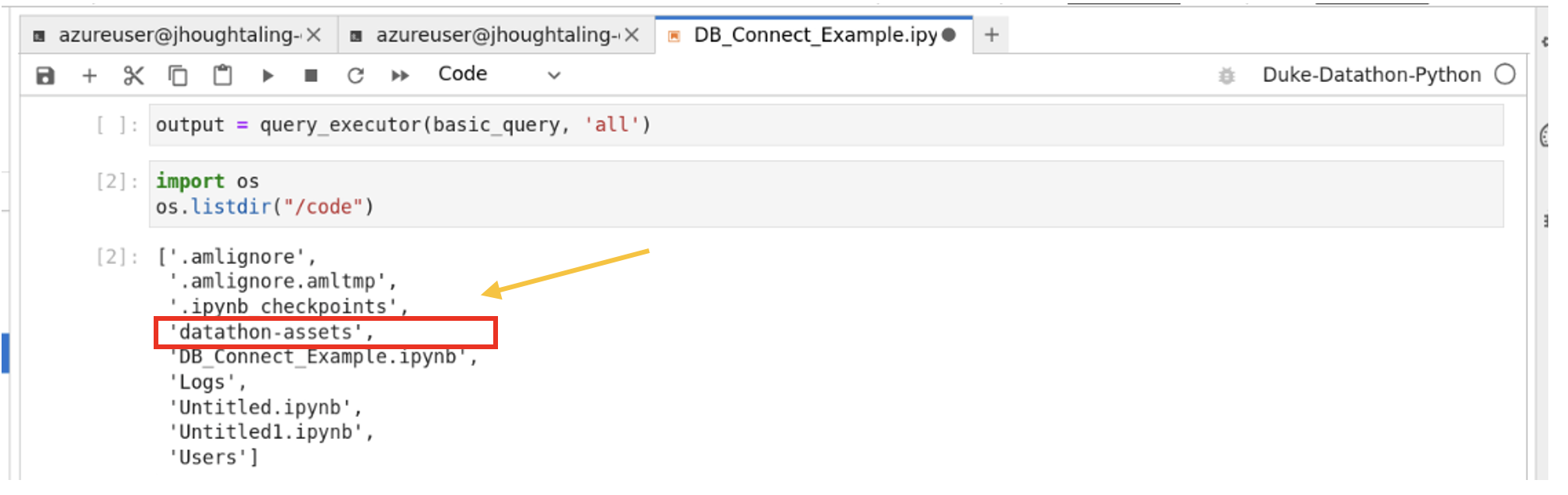
For telemetry/waveform data, there are TWO processes for access. The FIRST, and simplest, is referencing the data available in the ML workspace mount. You can find this data in several locations within the ML Workspace:



Inside the JupyterLab instance, the `datathon-assets` directory should be visible in your root folder:



Because your kernels are Docker-based, you access these files through a volume mount in the running container. This mount is located at “/code” inside the kernel, so you can access the files at “/code/datathon-assets”:



The SECOND way to access the data is through the ML Workspace datastore. This datastore can be referenced by the AzureML packages in both Python and R, and can also be referenced in formal training jobs submitted to common compute instances. MORE TO COME…

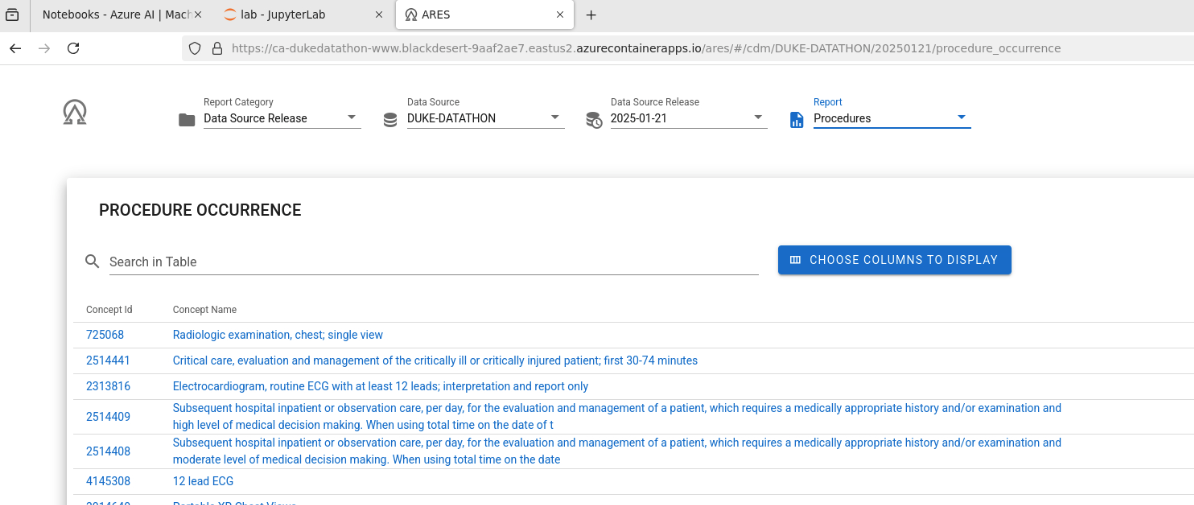
The waveform data are organized in the following structure:

datathon-assets/<SITE NUMBER>/<PERSON\_ID>/WAVEFORMS/<VISIT\_OCCURRENCE\_ID>/<WFDB FILES>

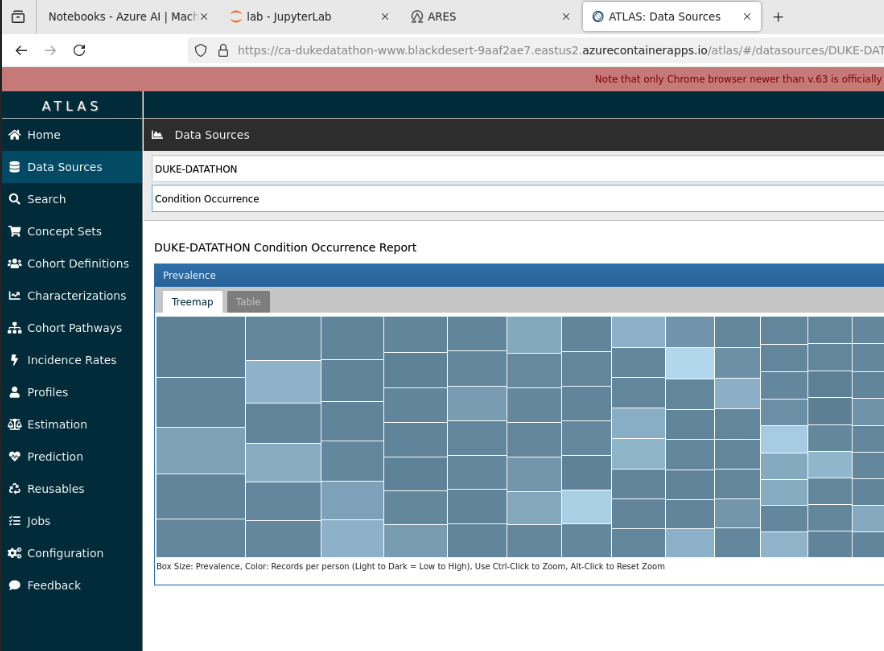
#### 

#### Are there other ways to explore the data?

Yes! You can navigate through the Firefox bookmarks to check out ARES to get a view on the OMOP data content and associated data quality:



You can also use Atlas to create concept sets, cohort definitions/phenotypes, characterizations, prediction and estimation models, etc:

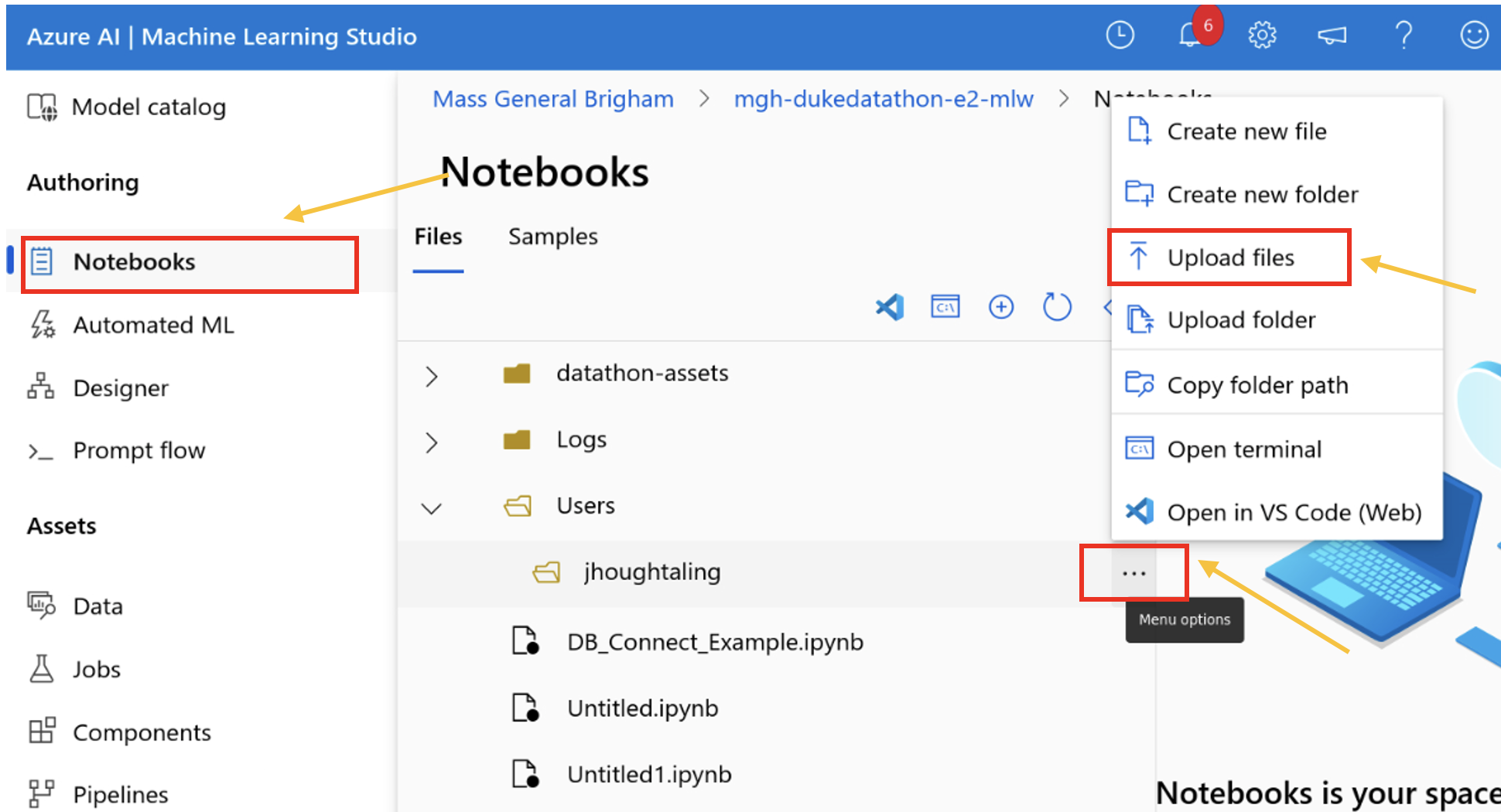


#### How do I train models created in Atlas in my compute instance?

Great question! The filesystem of the VM, while inaccessible to you, is available to both Atlas and to the ML workspace. First, create a model and download the package as a zip file:

### 

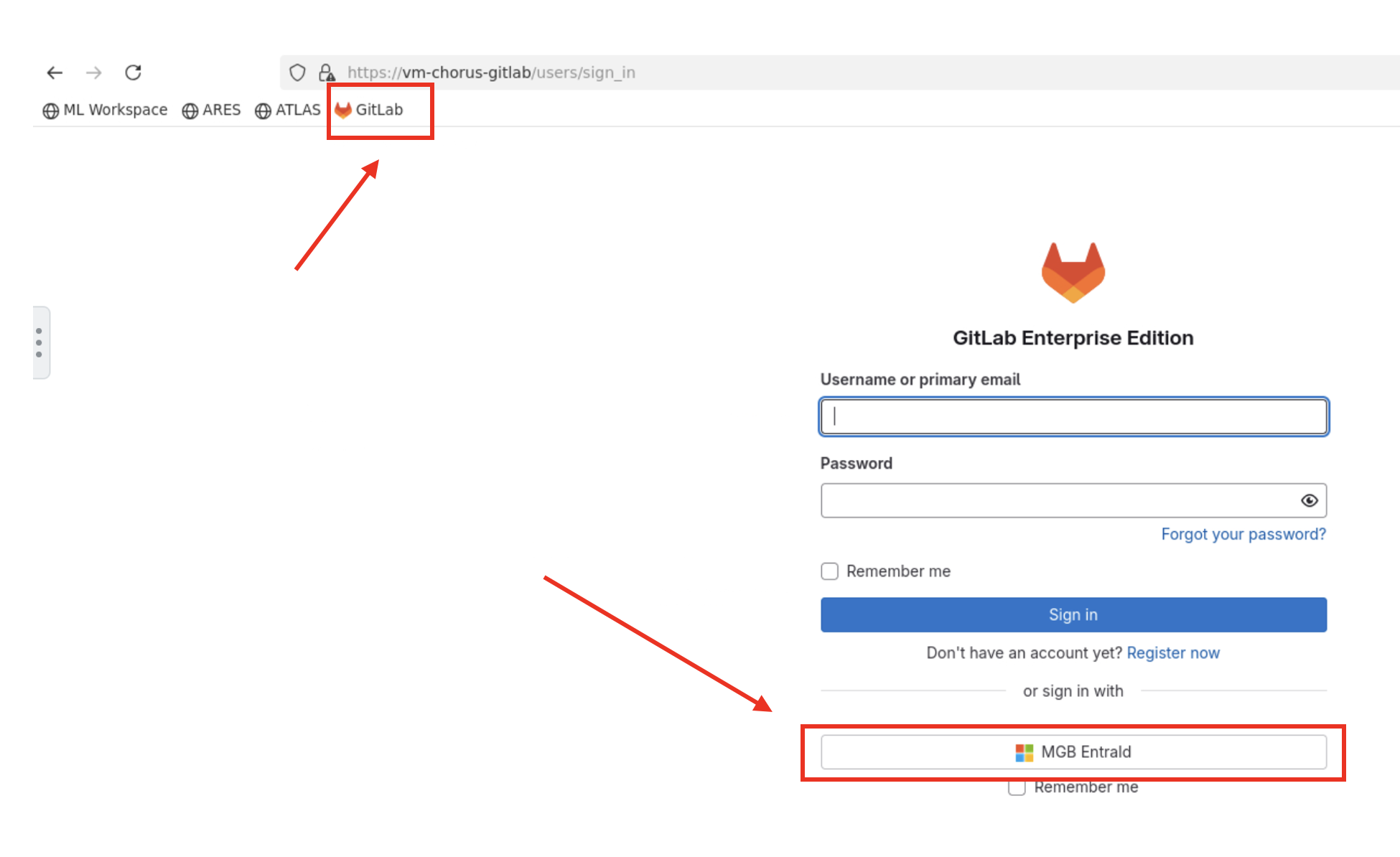
Once downloaded, head to the ML Workspace, select the ‘Notebooks’ tab, find the menu options in your upload location and select upload files:



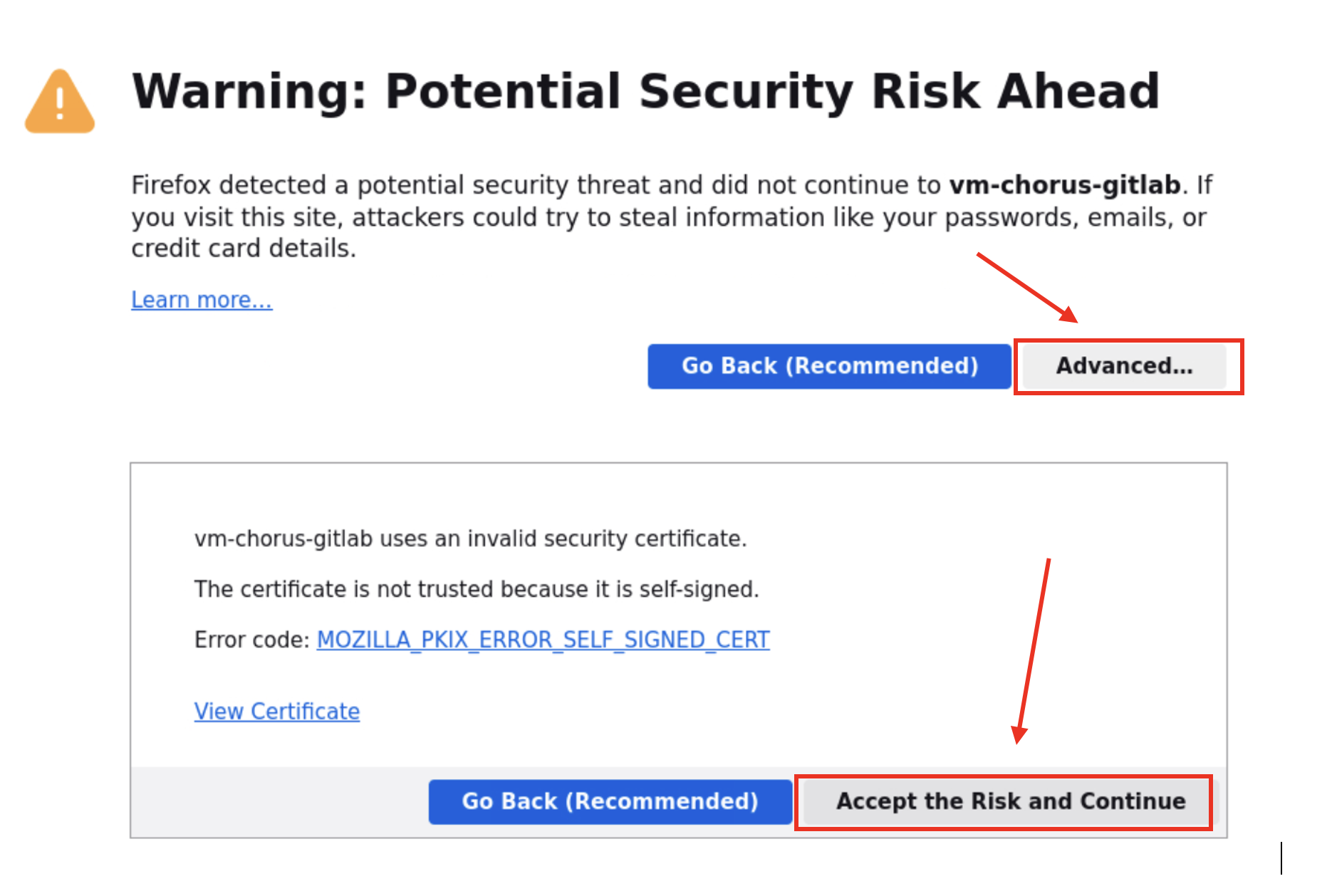
The file(s) will now be accessible in JupyterLab and inside your kernels, as described for the waveform assets above.

#### What’s the deal with GitLab?

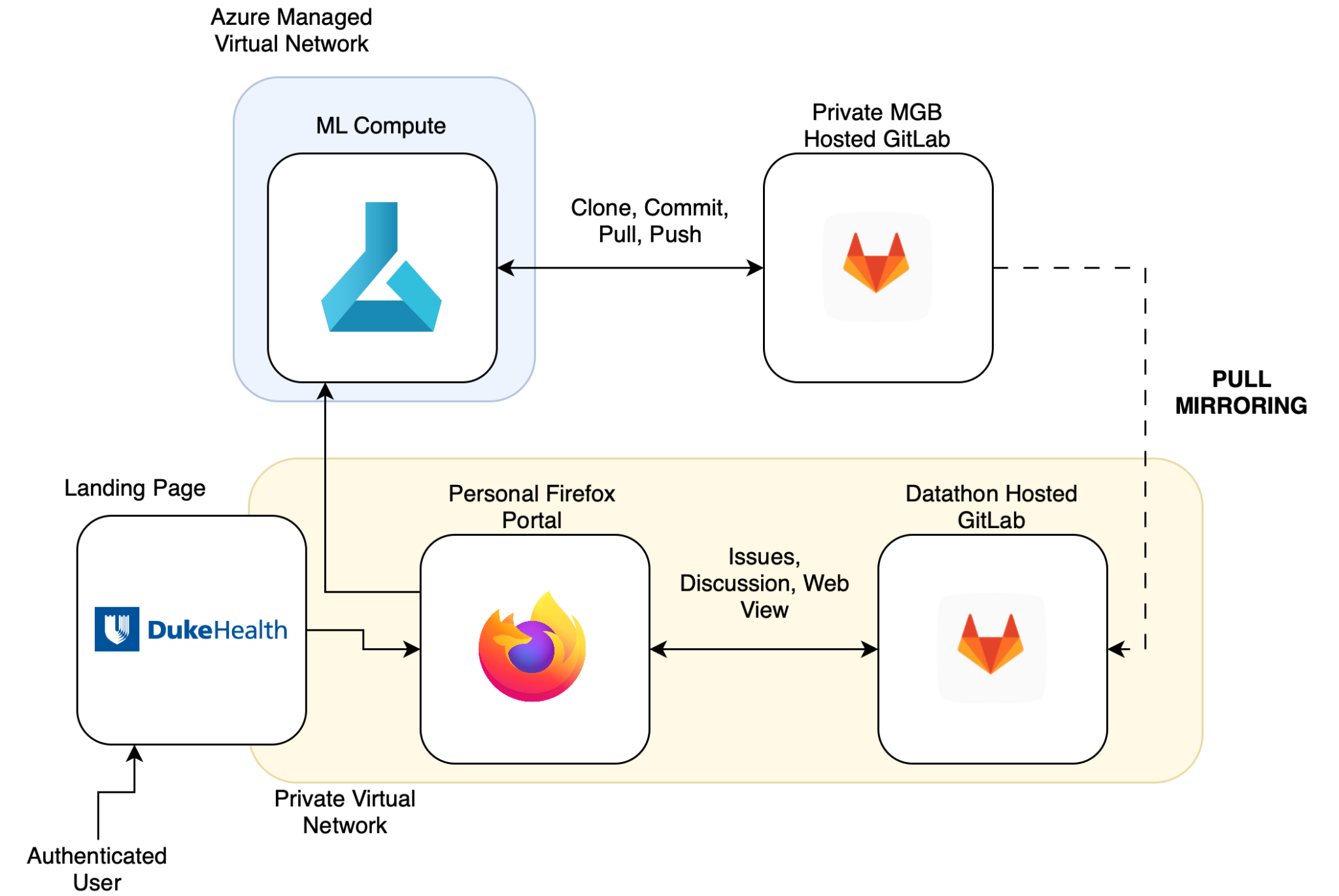
We’ve deployed a private GitLab instance inside of the Duke Datathon analytic enclave. You can get to it by clicking on the ‘GitLab’ bookmark in your bookmarks bar and then clicking on ‘MGB EntraID’ to login with your the same Microsoft credential you used above:



Note that this GitLab instance is not publicly accessible, so we have generated self-signed certificates instead of purchasing a validated certificate. Firefox will warn you about these certificates the first time you access GitLab, and you’ll need to ‘accept the risk’ to reach the login page:



Your ML workspace compute instance has heavily restricted outbound network access, and because it relies on a managed virtual network, we need to use a ***mirroring approach*** to commit and pull from your instance:



We’ve created a scoped project token that you can use to clone and make commits from your ML compute, which you can use to access the Private MGB Gitlab in your ML Compute via a terminal window:

**git clone https://oauth2:glpat-9eHw-xToNsYRHZsnQSB8@gitlab-scm.partners.org/duke-datathon/resources.git**

### 

### Seoul National University (K-MIMIC data)

#### Key links

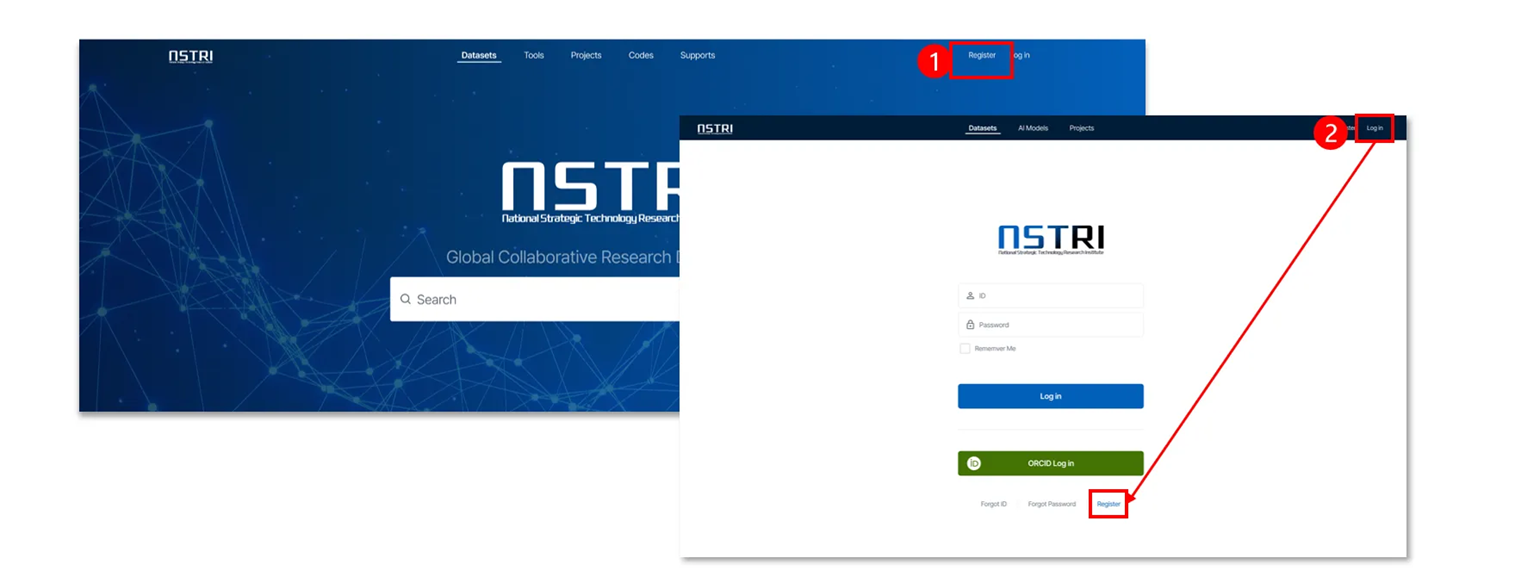
* [Statistical Overview of K-MIMIC.docx](https://docs.google.com/document/d/1uqYvuVaYYZ6Z9xhjuBw5CgsmBTRnX22R/edit?usp=drive_web&ouid=112459465598610221033&rtpof=true)
* [K-MIMIC\_Column\_Statistics.xlsx](https://docs.google.com/spreadsheets/d/1uv2Rk3dPw9PNjFwdj7Kq4sbZgSHWziaS/edit?gid=122298878#gid=122298878)
* **User Manual URL**:<https://nstridev.notion.site/NSTRI-User-Manual-11115e47dc748056add6ddf7866addd4>
* **NSTRI Platform URL**:<https://nstri.net/>

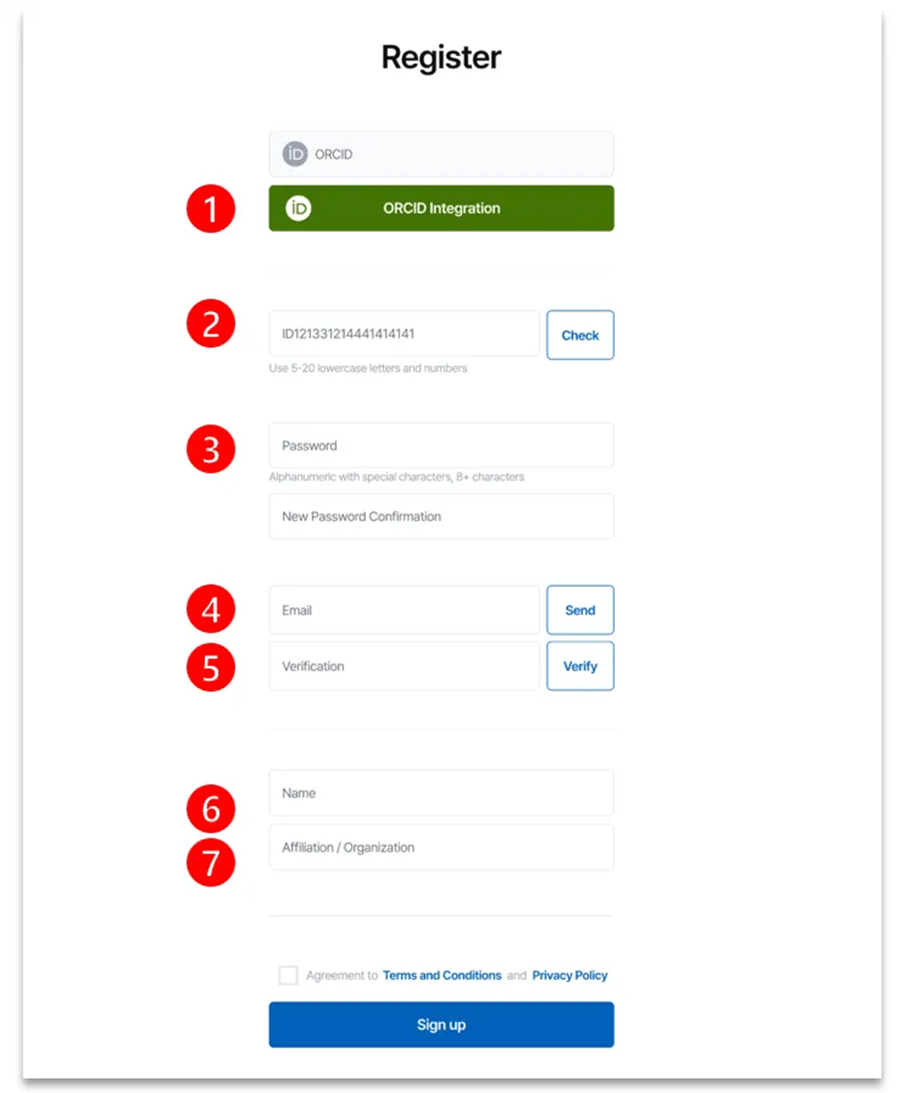
#### NSTRI Platform Structure for K-MIMIC

NSTRI is a cloud-based, closed analysis environment designed for developing AI models in healthcare.

* Each team is provided with one GPU instance.
* For security reasons, access is only available on the Duke Datathon Wi-Fi network.
* Participants will develop using JupyterLab, which runs in isolated Pods.
* The dataset includes structured data based on OMOP CDM, time-series data, and text data.

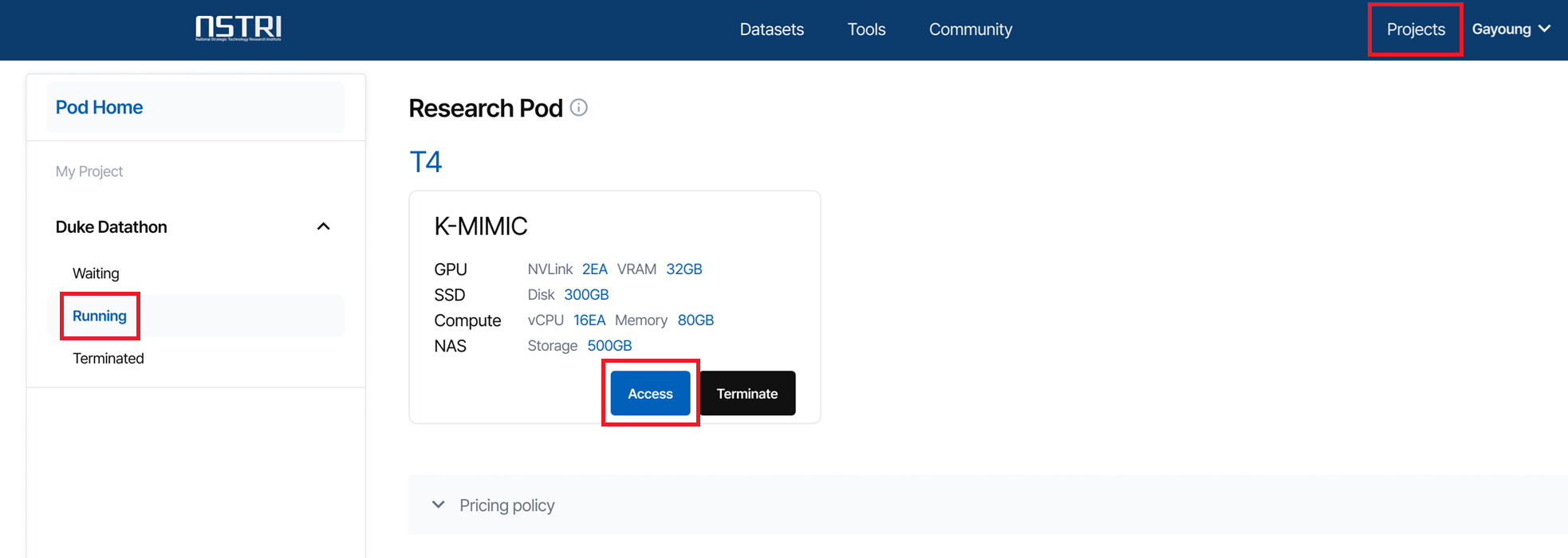
#### User Registration & Login

Click **[Register]** on the top-right of the homepage.  


1. ORCID linkage is **mandatory**.
2. **ID**: Lowercase English letters + numbers, max 20 characters.
3. **Password**: 8–16 characters, must include letters, numbers, and special characters (case-sensitive).
4. Complete email verification.

⚠️ The email used must match the one registered for the Datathon.

#### Accessing the Cloud Environment & Launching JupyterLab

1. Click [Projects] from the top menu.
2. Click [Research Pod].
3. Select [Running] based on your team name.
4. Click [Access] on the Pod card. Launch JupyterLab.

#### Development Environment Specifications

* TensorFlow and Torch GPU verification notebooks are provided (\*\_gpu\_check.ipynb)
* Administrator/root privileges are **not** provided.

| **Item** | **Specification** |
| --- | --- |
| **GPU** | NVIDIA T4 x 2 |
| **VRAM** | 32 GB |
| **vCPU** | 16 |
| **RAM** | 80 GB |
| **NAS** | 500 GB |
| **BlockStorage** | 300 GB |
| **Environment** | TensorFlow, PyTorch, scikit-learn |
| **Conda** | Miniconda |
| **GPU Driver** | 535.161.08 |
| **CUDA** | 11.8.0 |
| **cuDNN** | 9.1.0 |

#### 

#### Dataset Structure

All **bolded** folders/files are read-only.

/workspace

├── **datasets/** # Contains the complete K-MIMIC dataset  
├── **KMIMIC-EDA/** # Contains K-MIMIC EDA notebooks

├── tf\_gpu\_check.ipynb # TensorFlow GPU check notebook

└── torch\_gpu\_check.ipynb # Torch GPU check notebook

#### 

#### 

#### Outbound Whitelist

Only the following external URLs are accessible from within the environment:

* <https://pypi.org/>
* <https://anaconda.org/conda-forge/>
* <https://cran.r-project.org/>
* <https://huggingface.co/>
* <https://vitaldb.net>
* <https://physionet.org>

⚠️ Access to any other sites is blocked. Requests for additional access must be approved separately.

#### Important Notes

* **Do not terminate the Pod**: If the Pod is terminated, it cannot be recovered.
* **No reinitialization**: Pods cannot be reset or reinitialized.
* **No environment changes**: Requesting a new environment is not allowed.

### Epic Cosmos

#### Key links

#### I’m a Duke person, where do I go to register?

Please apply: <https://redcap.duke.edu/redcap/surveys/?s=3ACKTYNFATCMANDL>

#### Can you provide examples of what work has been done in Cosmos?

Please see here: [Cosmos Research Publications - January 2025.pdf](https://drive.google.com/file/d/1EatO4D_0sdzhDGFsOVGmnmOiU5Mm3kyL/view?usp=sharing)

#### What is Cosmos?

Cosmos is a massive, de-identified dataset that aggregates electronic health records (EHR) from hundreds of Epic organizations. It is designed to improve patient care by enabling new discoveries and advancing medicine. The dataset is updated in real-time, providing a living resource for research and clinical insights. Cosmos integrates inpatient and outpatient charts into a single, longitudinal patient record, including diagnoses, medications, patient-generated data, and specialty-specific data.

#### 

#### What data is available in Cosmos?

Cosmos contains data from over 300 million patients, with billions of encounters and lab results. The dataset includes a representative sample of patients, comparing well to the United States Census in terms of age, race, ethnicity, insurance coverage, and social vulnerability index. The data types available include detailed oncology visits, cancer staging, advanced lab results, hospitalizations, standard outpatient visits, and much more! Cosmos also deduplicates patient records, ensuring a single combined record for patients seen at multiple contributing organizations. For more detailed information on data domains, refer to the Galaxy guide for the [data domain encyclopedia](https://galaxy.epic.com/?#Browse/page=1!68!600!100198880&from=Galaxy-Redirect).

#### 

#### How do you get access to Cosmos and Training Resources?

To access Cosmos, you must first have both a Duke Epic account and an Epic UserWeb account. To get access to the Cosmos Portal, which includes access to our low-code analytical tool, SlicerDicer, all you need to do is submit a Cosmos Access Request to Duke’s Cosmos Provisioning Point Person via this [form](https://redcap.duke.edu/redcap/surveys/?s=3ACKTYNFATCMANDL) (<https://redcap.duke.edu/redcap/surveys/?s=3ACKTYNFATCMANDL>).

To get access to any of the training guides linked in this section, you will need an Epic UserWeb account.

#### 

#### Available Analytical Tools

Cosmos provides access to powerful tools like SlicerDicer and the Data Science Virtual Machine (DSVM). SlicerDicer, our low-code analytical tool, enables complex query and visualization capabilities, allowing users to explore aggregated data without accessing raw line-level patient data.

The DSVM offers advanced data science tools, such as Python and R, for certified users to interact with the de-identified dataset. Access to the DSVM requires additional training to ensure users are equipped to handle the data responsibly and effectively. See this [guide](https://galaxy.epic.com/?#Browse/page=1!68!642!100074365,100191866&from=Galaxy-Redirect) for more information on the types of DSVM users and how to sign up for a class!

#### 

#### 

#### Cosmos Training Resources

* Overview
  + [Welcome to Cosmos](https://galaxy.epic.com/?#Browse/page=1!68!642!100074365) – Overview Resource
  + [Cosmos Changelog](https://galaxy.epic.com/Redirect.aspx?DocumentID=100174255&PrefDocID=152082) – What new has been added
  + [Data Science White Paper Library](https://galaxy.epic.com/Redirect.aspx?DocumentID=100160836&PrefDocID=150295) – White papers for more technical users
  + [Cosmos Research Publications - January 2025.pdf](https://drive.google.com/file/d/1EatO4D_0sdzhDGFsOVGmnmOiU5Mm3kyL/view?usp=sharing)
  + [Epic Research Website](https://www.epicresearch.org/) – See all the cool research Epic has done on Cosmos
* What’s in Cosmos?
  + [Cosmos Data Domain Encyclopedia](https://galaxy.epic.com/?#Browse/page=1!68!600!100198880&from=Galaxy-Redirect) – Go to guide for overview of all data in Cosmos including known biases and limitations
  + [The Cosmos Data Dictionary](https://galaxy.epic.com/Redirect.aspx?DocumentID=100149584&PrefDocID=148191) – List all available data elements in Cosmos
    - Check the Cosmos Portal for the most up to date list!
  + [Cosmos Known Data Quality Issues](https://galaxy.epic.com/?#Browse/page=1!68!95!100208600&from=Galaxy-Redirect)
* Cosmos Training
  + [Some short refresher videos on Cosmos](https://welearning.epic.com/Browse?q=cosmos&ver=1254) including how to navigate the portal and work in SlicerDicer
  + [Sneak Peek](https://galaxy.epic.com/?#Browse/page=1!68!510!100273341&from=Galaxy-Redirect) - How does Sneak Peek work?
  + [Cosmos Super User Questions](https://galaxy.epic.com/?#Browse/page=1!68!240!100213616,100213624,100216235&from=Galaxy-Redirect) – Steps to become a Super User
  + [COS100 Tour of Cosmos](https://welearning.epic.com/Details?CSN=688028) – E-learning
  + [COS200 Querying Cosmos Data](https://welearning.epic.com/Details?CSN=548754) - E-learning
  + [COS305 Introduction to Cosmos Data Science](https://galaxy.epic.com/?#Browse/page=1!68!240!100264800&rank=4&queryid=128693859&docid=166087)
  + [COS500 Cosmos Data Model](https://galaxy.epic.com/?#Browse/page=1!68!240!100161106,100161107,100195760&from=Galaxy-Redirect)
  + [COS550 Cosmos Data Science](https://galaxy.epic.com/?#Browse/page=1!68!240!100272512&rank=1&queryid=128693922&docid=166838)
* SlicerDicer Training
  + [Cogito Bytes](https://streaming.epic.com/?category=3) – Short training videos on how several SlicerDicer features work
  + [COG170 SD Advanced Features](https://galaxy.epic.com/?#Browse/page=8400!68!240!3854936,3855098,100262377&from=Galaxy-Redirect) – Training manual on common advanced SlicerDicer methods
  + [SlicerDicer Toolkit](https://galaxy.epic.com/?#Browse/page=1!68!640!100015389&rank=1&queryid=127265888&docid=113963) – SlicerDicer documentation library
  + [Archive of Past Slice of the Month Sessions](https://galaxy.epic.com/Redirect.aspx?DocumentID=100195042&PrefDocID=156038) – Presentations of Cosmos SlicerDicer research projects from community members

### General data questions

#### Can I bring my own dataset?

The principle is that all teams can have access to all data for the event. We aim to make resources accessible to:

* all participants for at least:
  + 2 weeks before and
  + 6 months after

This should give teams sufficient time to complete analysis for a publication.

* all staff at least 2 months before and 8 months after the event
* and there’s a pathway to manage the DUAs

The datathon staff could discuss further if you reach out to med (at) aiwong (dot) com.

Otherwise, this would not be permissible, as it disrupts the playing field.

### 

### Access to GitHub

#### I don’t have a GitHub ID

Please register for one at [GitHub.com](https://github.com/). We’ve created a video for you if you’re facing any difficulties: [How to Register for GitHub (and get a GitHub ID)](https://youtu.be/uL5QCAmRbVI)

### Datathon registration

#### What is the purpose of the trainee recommendation letter? What does it need to say?

The trainee recommendation letter gives us further context on how to align you in terms of skill level. It also gives further context that you have trustworthy references who can guide you as you continue after this. It’s preferably someone who can talk about your historical research ability, even better if they can talk about your data science perspective, and finally best if they can speak to clinical data science. If not, someone who can talk about your clinical practice will suffice – hopefully, someone you think of as a mentor.

### 

### 

### Datathon operations

#### Can my team change what topic we research?

Yes! The initial topic provided is a *suggestion*. The topic is not guaranteed to be 100% novel; it may have been published before in some fashion. The team's feelings may change based upon the data accessed - perhaps the data doesn’t support that analysis. That’s all okay. Refining the question and/or pivoting is all part of the process.

#### I’m clinical, and have never done this before. Can I join?

Yes! Your interest, but not experience, is required. What you *are* required to do, however, is register for all the datasets and run the tutorial videos. You’re expected to get elbow-deep in all the data and code as best you can with everyone else on the team. While you’re not going to be an expert-level coder or data scientist overnight, this datathon is intended to reduce the activation energy - show you that EHR data science isn’t so scary after all.

Everybody codes. Nobody quits.

#### How do you define success for the datathon?

This datathon will be successful if ≥50% of the teams submit ≥1 research product (e.g., abstract, paper, poster, etc.) within 6 months of the datathon.

#### Can I pick what team I’m in?

No, teams will be assigned. The intent is to build *new* connections and ensure a diversity of experience within a team, so you are likely to be split into a team from different divisions/institutions/companies.

#### Can I change teams at the datathon?

No, teams are fixed.

### 

### Judging criteria

#### How will we be judged?

Judging has two parts:

* Peer judging: an Eurovision-style judging, where each team judges every other team.
* Formal judges: will judge your eligibility for categories

### 

### Running the example Colab notebook

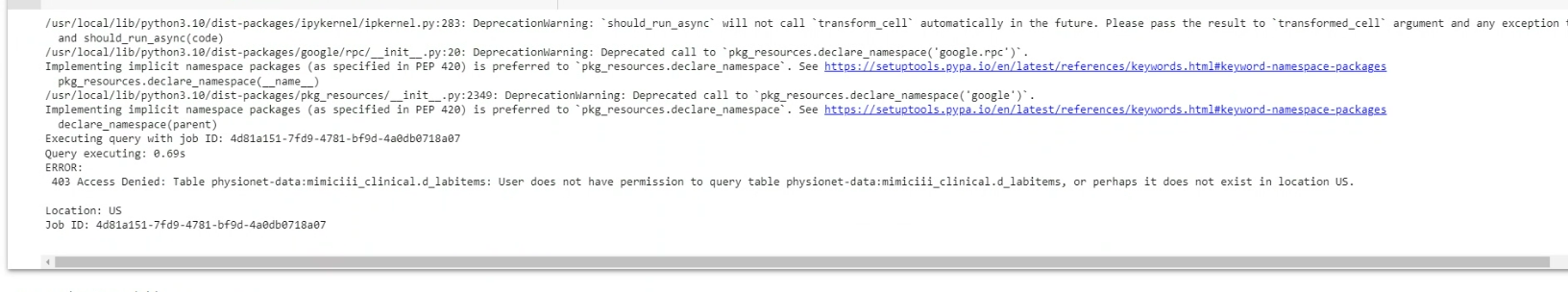
#### I can’t run the cells

You probably received an error because you’ll need to run the ‘defines’ cell first.

Consider following this instructional video. [How to analyze PhysioNet EHR data with Jupyter/BigQuery](https://youtu.be/VDyw4ONIxMI)

#### The Google Colab doesn’t have access to MIMIC / eICU

You probably received an error like this.



Be sure you:

* Logged into the right Gmail account (it’s easy to have a different one selected)
* Followed the [How to set up Physionet (MIMIC, eICU) for BigQuery](https://www.youtube.com/watch?v=N6fqg2QaNN8&pp=ygUVInBoeXNpb25ldCIgYmlncXVlcnkg)

### 

### What is the presentation format?

Please see here [mentor checkpoints and general criteria](https://docs.google.com/presentation/d/1ECLrm5w5MmyLZlA6jhYJ0PQSEoPf3wO5oqF1sbkNDC8/edit?usp=sharing)

[example judging spreadsheets](https://docs.google.com/spreadsheets/d/17ijybo4cwdiDyyNbkPuCAAevpF_hk_vCMW8-P6GS99Q/edit?usp=drive_link)

#### What are the judging criteria?

Please see here: [mentor checkpoints and general criteria](https://docs.google.com/presentation/d/1ECLrm5w5MmyLZlA6jhYJ0PQSEoPf3wO5oqF1sbkNDC8/edit?usp=sharing)

### What is the presentation order?

#### Where can I send this abstract/paper?

#### Link: [conferences - healthcare data science + pulmonary/critical care](https://docs.google.com/spreadsheets/d/12C9SGqXqEBWVbT-w2nD2TmkOlfPoEE9fe9zcVFxzwXU/edit#gid=2018761422)

### Unanswered questions

Please feel free to add any unanswered questions here. The mentors and organizers will periodically return here and answer questions

## 

## Frequently used acronyms

### D

### E

EHR: Electronic health record. Generally interchangeable with EMR.

EMR: Electronic medical record. Generally interchangeable with EHR.

### I

ICU: Intensive care unit.

### S

SCCM: Society of Critical Care Medicine. <https://sccm.org/>

## What resources / things would be helpful to know for next year?

### For first-time data scientists

(please add below)

### For first-time clinicians

(please add below)

### For basic scientists

(please add below)