National COVID Cohort Collaborative

Introduction to Analyzing Real-World Data Using the National COVID Cohort Collaborative (N3C)

Spring 2024 - N3C Education and Training Domain Team

Session 1, Assignment 1

Enclave Navigation, Submitting a Data Use Request

DUE: Jan. 25, 2024

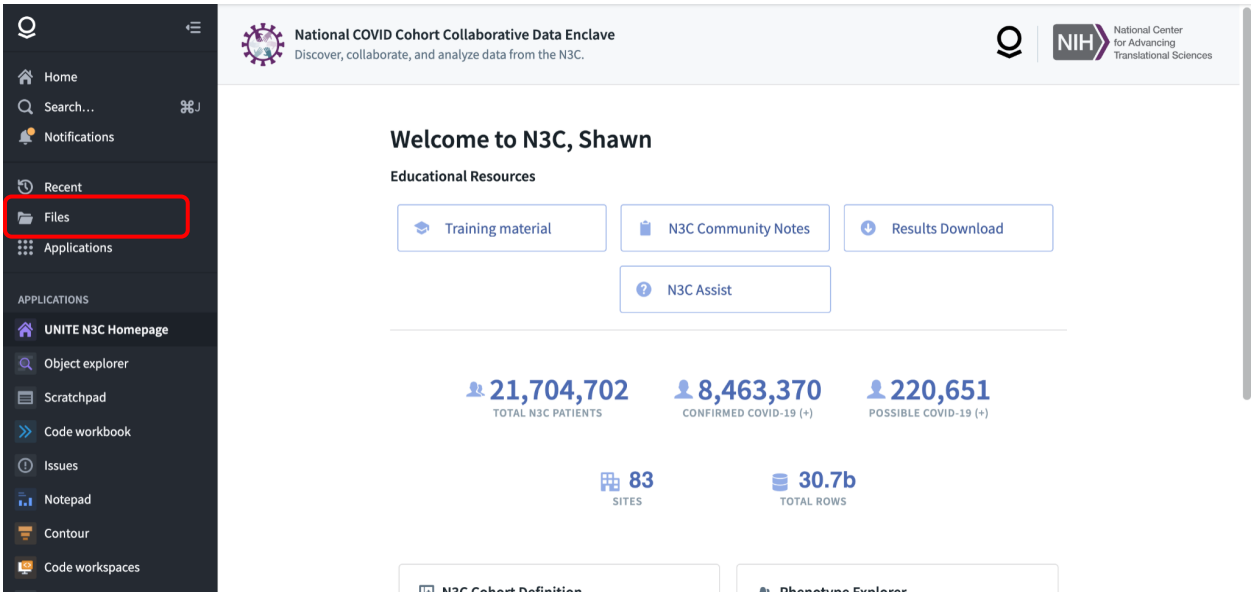
Pre-requisites

To complete this assignment, you will need to:

- Be able to log into the N3C Enclave at https://unite.nih.gov
- Have completed your Human Subjects Research Protection Training within the past 3 years, and have the date of your completion on hand. More information here.

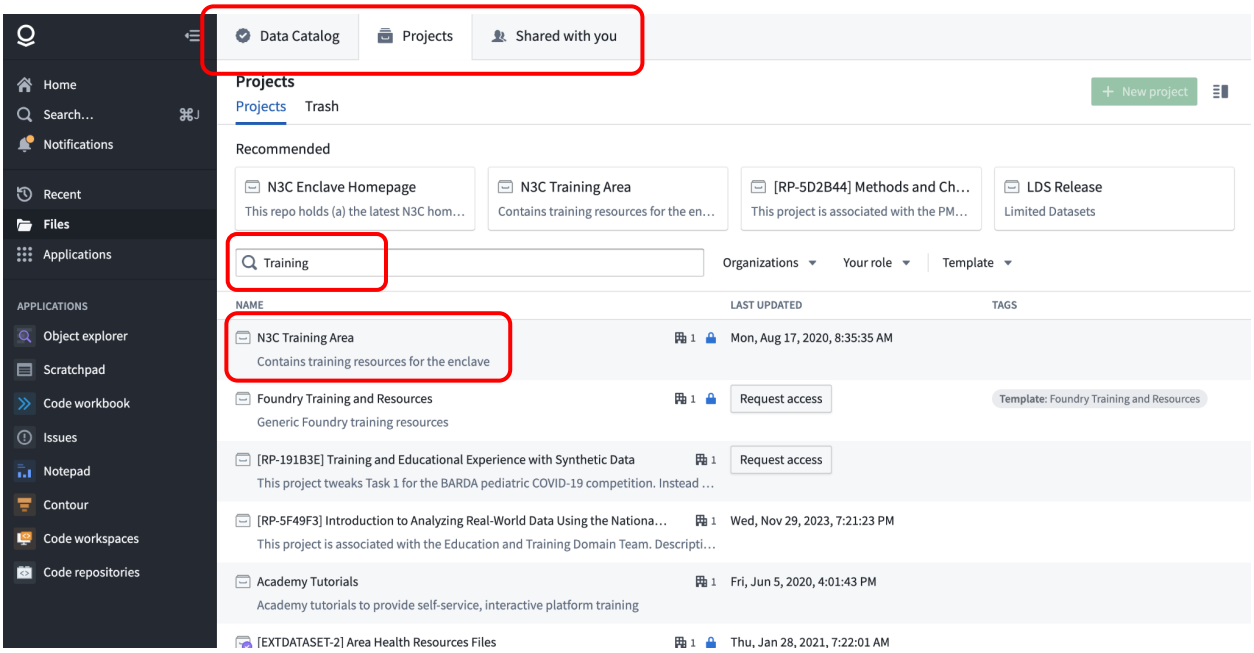
Part 1 - Workspaces, Folders, Favorites

After logging into the enclave, you'll see the enclave homepage with quick links to various resources. Start by browsing the **Files** link in the left menu bar:



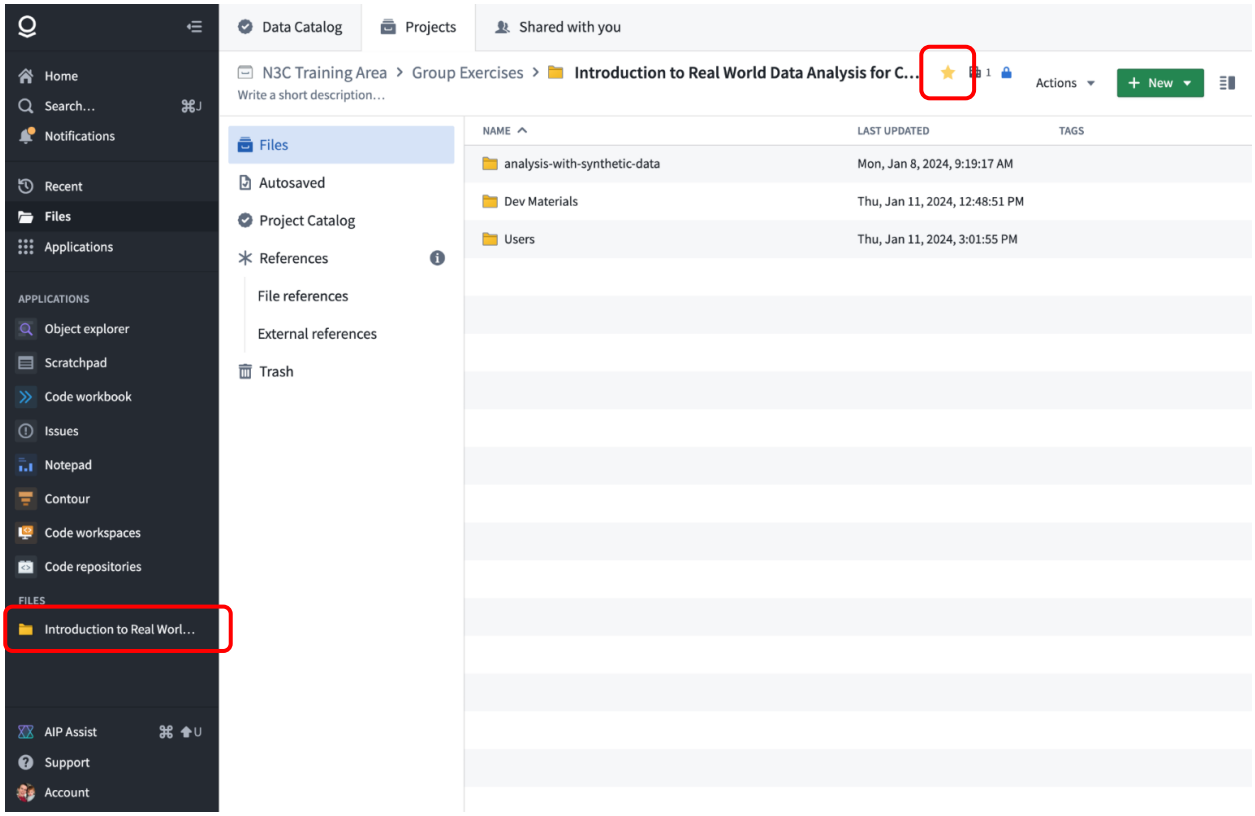
This will open a dashboard where you can browse **Projects** - also known as project workspaces. These act like folders with strict access permissions, and use a filing-drawer icon. This dashboard also provides access to the **Data Catalog**, which we'll cover later and provides collections of shortcuts to frequently-accessed files (box 1).

There are many project workspaces listed - search for "Training" to find the N3C Training Area workspace and open it (boxes 2 and 3).

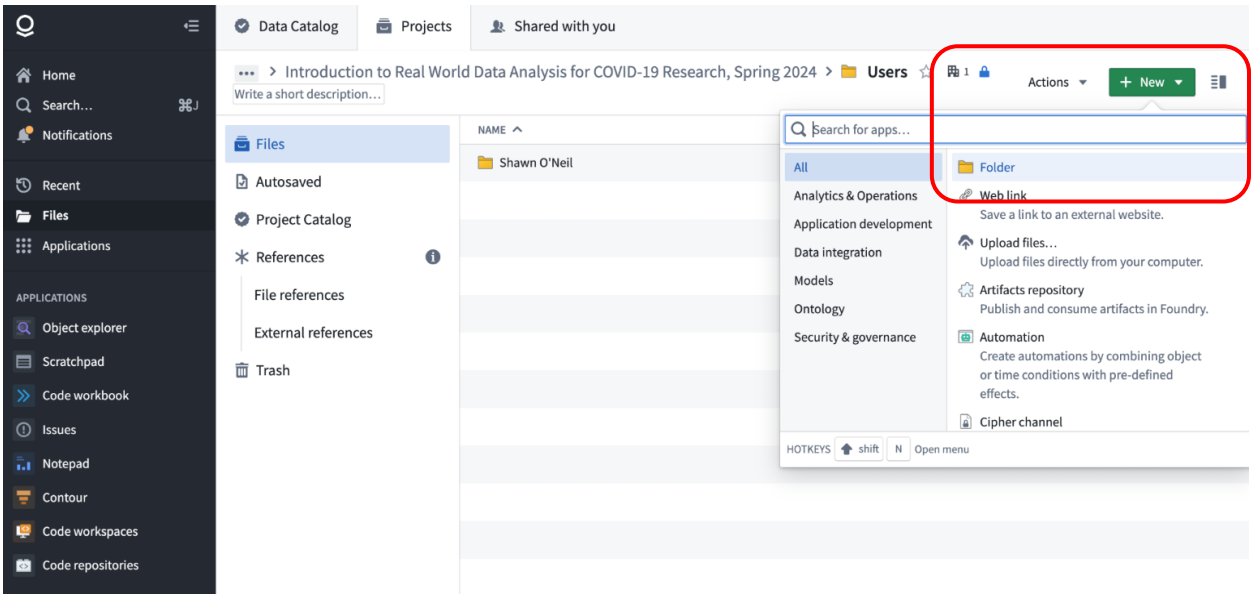


In the workspace, open the folders **Group Exercises → Introduction to Real World Data Analysis for COVID-19 Research, Spring 2024**. Once you have this folder open, click the "star" icon that is next to the "breadcrumb trail" for the folder (box 1). This will put the folder in your "favorites", which are listed in the left navigation bar under **Files** (box 2). We highly recommend using these star-favorites and the breadcrumb trails for faster navigation in the enclave!

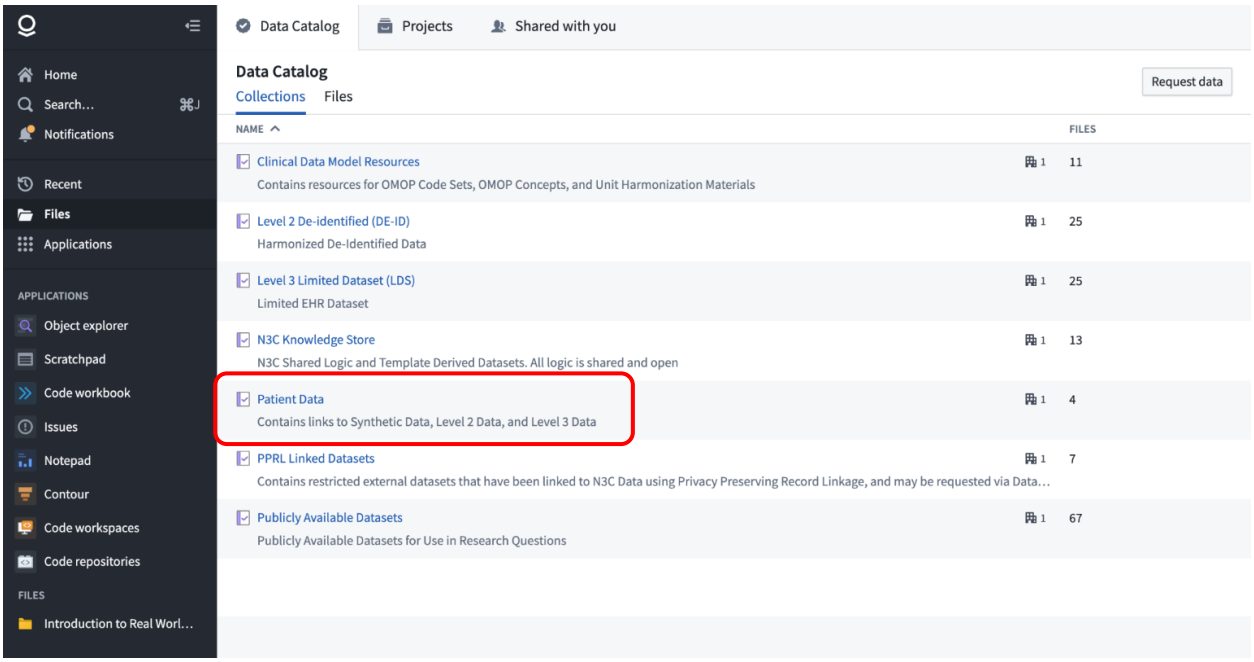
Another feature of the Enclave is the use of persistent URLs. For example, the URL displayed in the browser for the N3C Training Area workspace is https://unite.nih.gov/workspace/compass/view/ri.compass.main.folder.86a7020f-db30-4fd1-b735-bbaf53512365 - this URL will always take you directly to the N3C Training Area workspace, even if it is moved to a different location or renamed in the future. You can thus bookmark specific enclave resources in your browser, in addition to starring favorites in the interface.



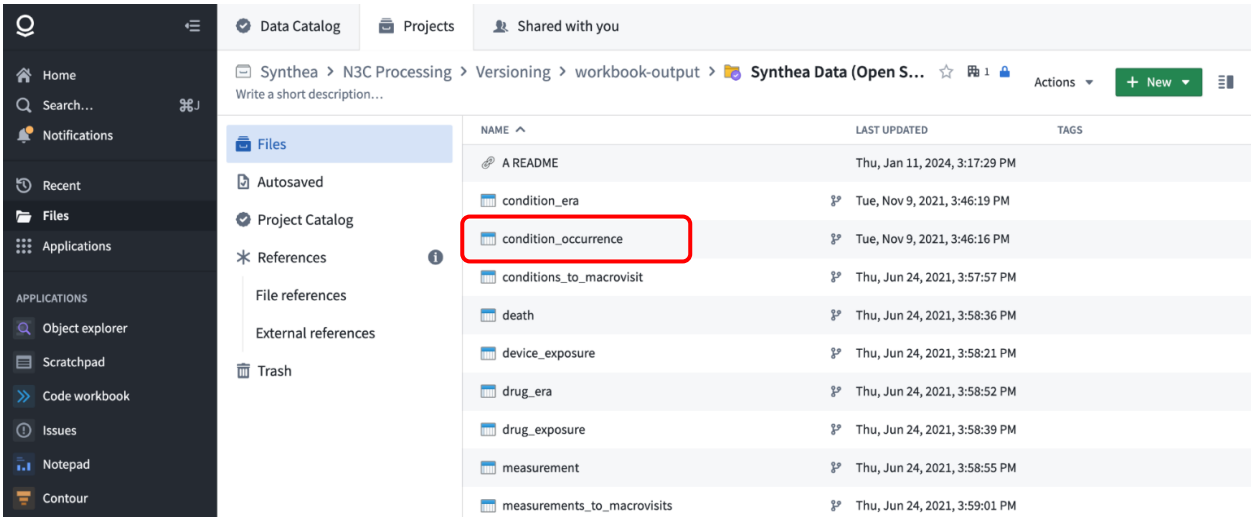
Next, open the Users folder, and create a **New Folder** using the green **New** button, naming the folder your first and last name. *If you are unable to access the workspace or create a new folder, contact Shawn O'Neil on the CD2H/N3C Slack or at shawn@tislab.org to have permissions added for you.*



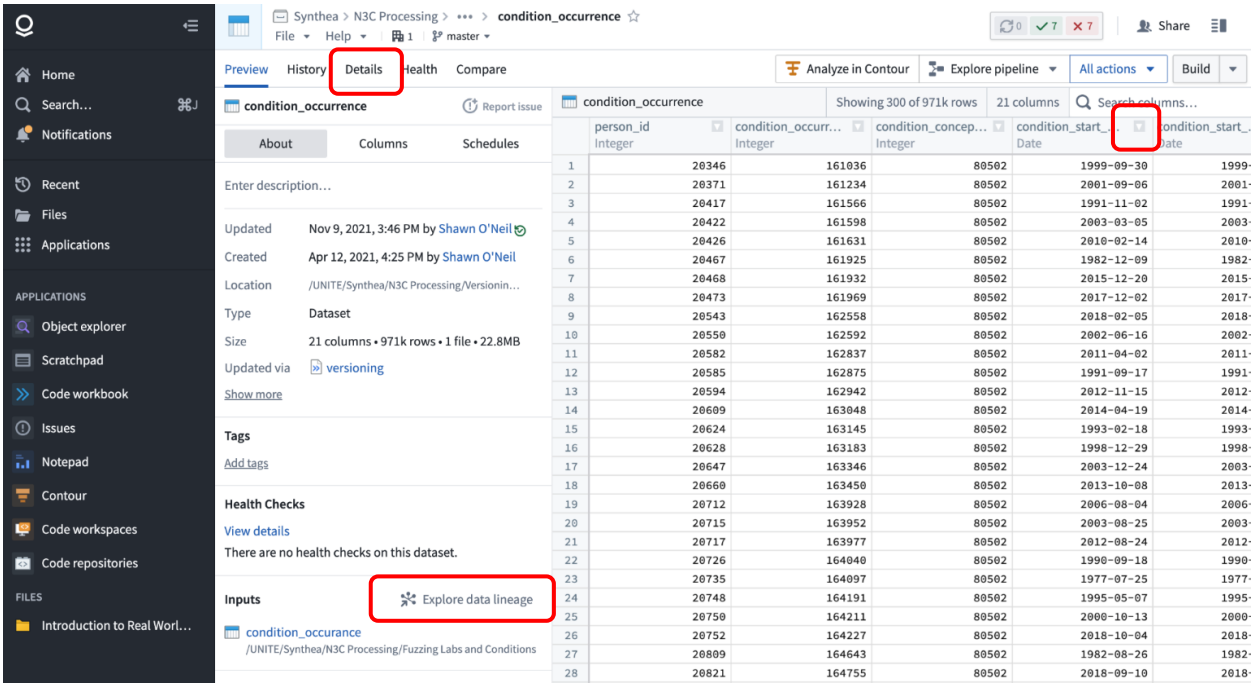
Next, let's take a look at some notional synthetic data in the enclave. Rather than find the specific location it lives in, we'll use the **Data Catalog** from the file browser to view a collection of shortcuts to commonly used resources. Open the Data Catalog tab, and then open **Patient Data → Synthea Data (Open Source Synthetic)**.



This will open a folder listing OMOP-formatted data tables, known as **datasets** in Enclave parlance. This dataset is notional (fake), but generated based on a statistical model of patient histories developed early in the pandemic. You can optionally learn more in the **A README** file. Next, open the **condition_occurrence** file.



Opening a dataset this way opens the Dataset Preview application, a powerful interface for quickly browsing and summarizing the table.



This table is in OMOP format, which we'll cover more in later sessions. Here are some simple data manipulation exercises you can try in the data browser interface:

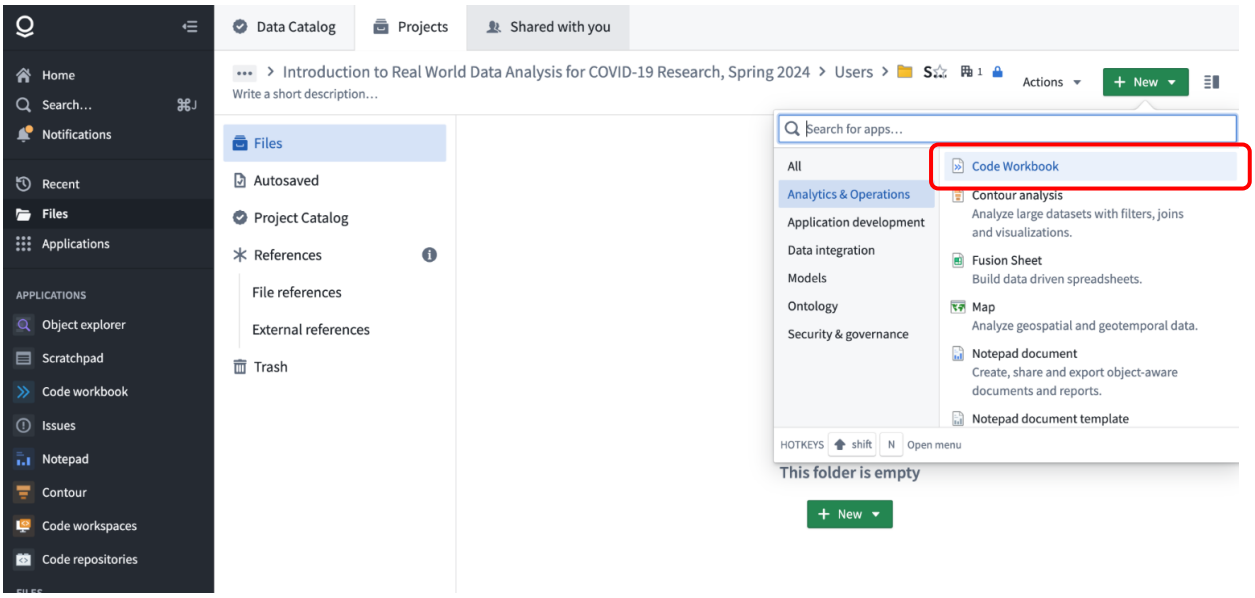
1. Use the dropdowns next to one of the columns (box 1) to **Filter** values (though note that filtering is limited to exact matches only in this basic interface), **Sort** them, or **View Stats**. The View Stats feature is particularly useful for data browsing, and can be done after filtering by one or more other columns.

2. Check out the Details tab (box 2) for more information about the dataset, noting especially the **Files** information. Datasets in the enclave are handled by Apache Spark, a high-performance distributed computing technology, which breaks large datasets over many files for efficiency (this dataset consists of just one file because it is small). For the most part we won't need to worry about this, but Spark is a powerful tool for advanced enclave usage, particularly for analyzing N3C's very large datasets!

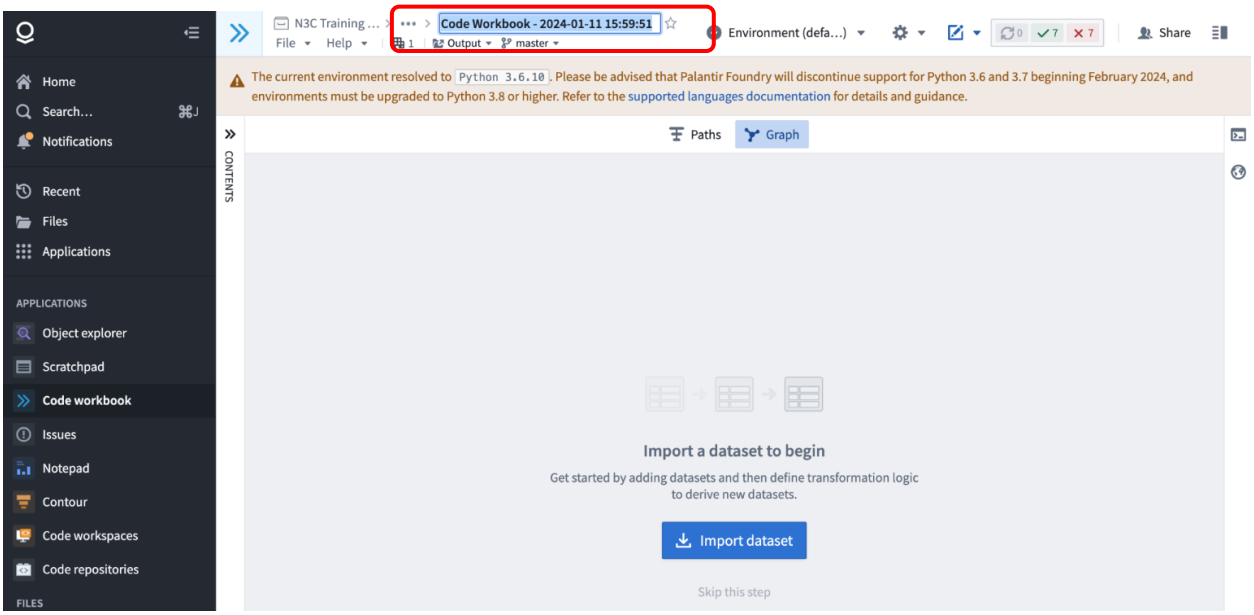
3. Visit the **Explore data lineage** tool (box 3), to see the input datasets and code that were used to create this dataset. This is another advanced tool you won't need to worry about now, but highlights that the enclave has very powerful features for data provenance.

Part 2 - Code Workbooks and Datasets

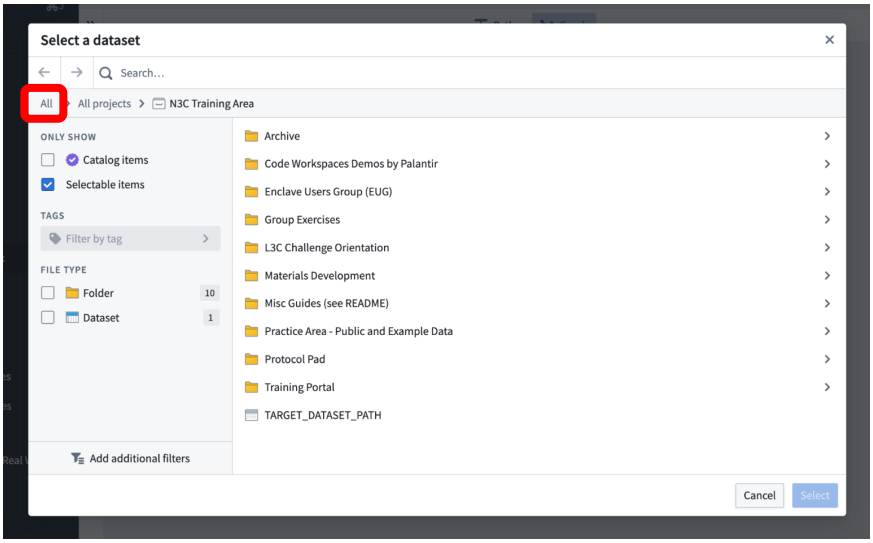
In this part we're going to create a very basic data manipulation pipeline with the Code Workbook tool, largely to practice working with workbooks, folders, and datasets together. First, navigate to your student folder from part 1 above (via your favorites list in the navigation menu!), and use the green **New +** button to create a **New Code Workbook**. It will be listed under "Analytics & Operations":



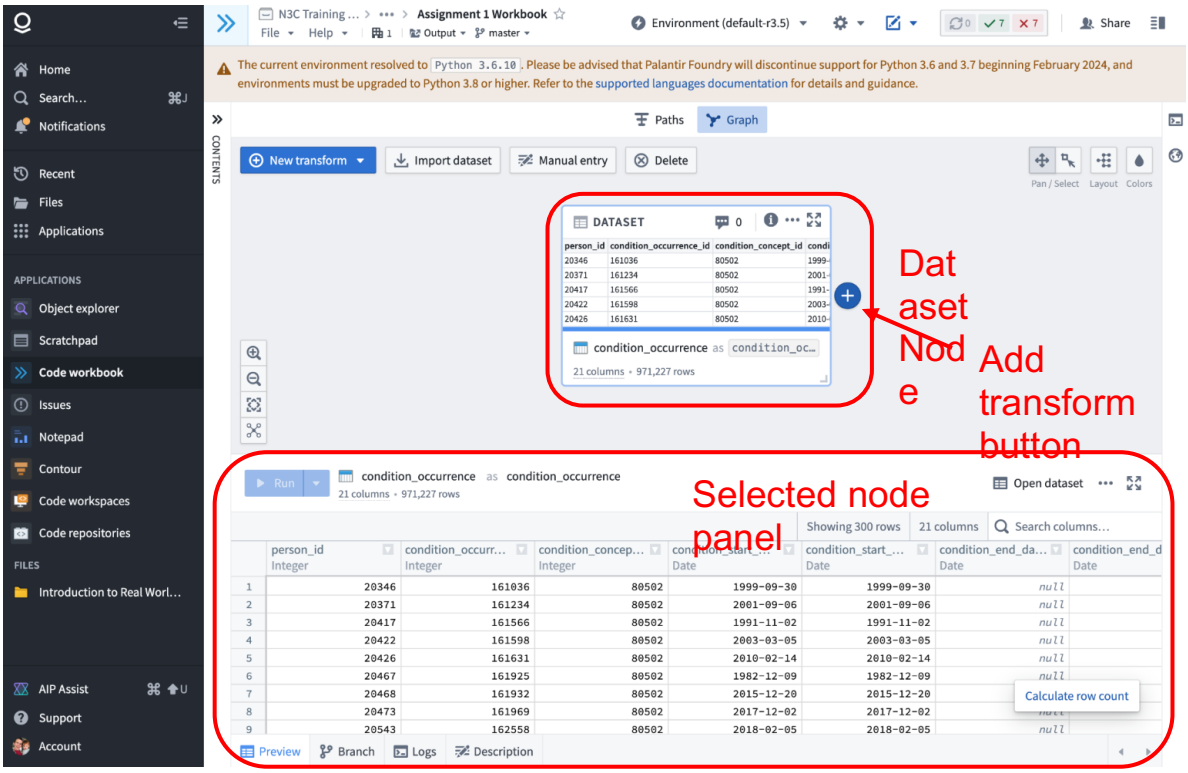
The resulting Code Workbook will be named according to current date; click the filename in the breadcrumb menu to rename it **Assignment 1 Workbook**. When prompted to rename the output folder, select **Rename**.



Now, you can click the **Import dataset** button to open the file browser and choose a dataset to import. We are going to import the **condition_occurrence** dataset we saw above, by finding it in the Data Catalog. To do that, navigate and select it using **All → Data Catalog → Patient Data → Synthea Data (Open Source Synthetic) → condition_occurrence**.



The next view is of the Code Workbook main interface, which shows the dataset we've imported as a "node" in the workspace. When a node is selected, a panel will show information about it on the bottom, and when it is highlighted a blue "Add transform" button will appear.

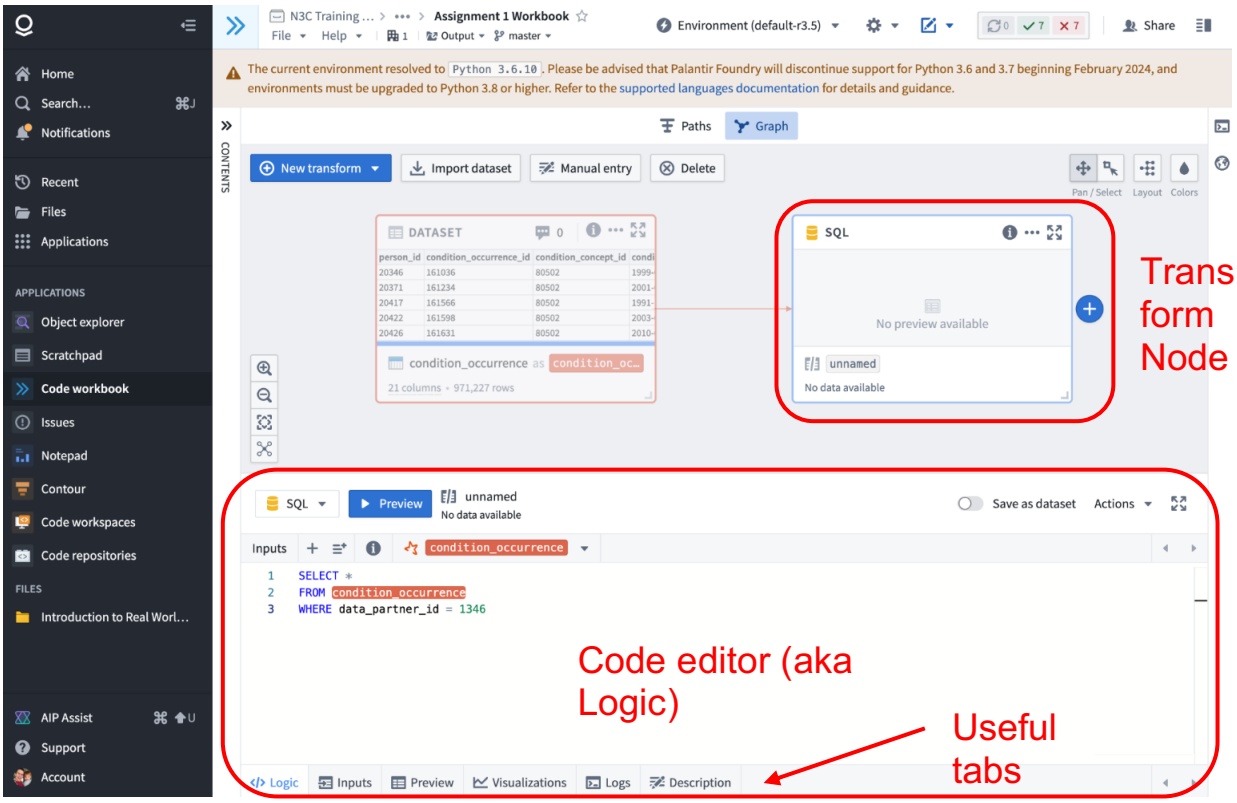


NOTE: There is generally no need to "save" files as they are edited in the enclave - they are continually saved as you work.

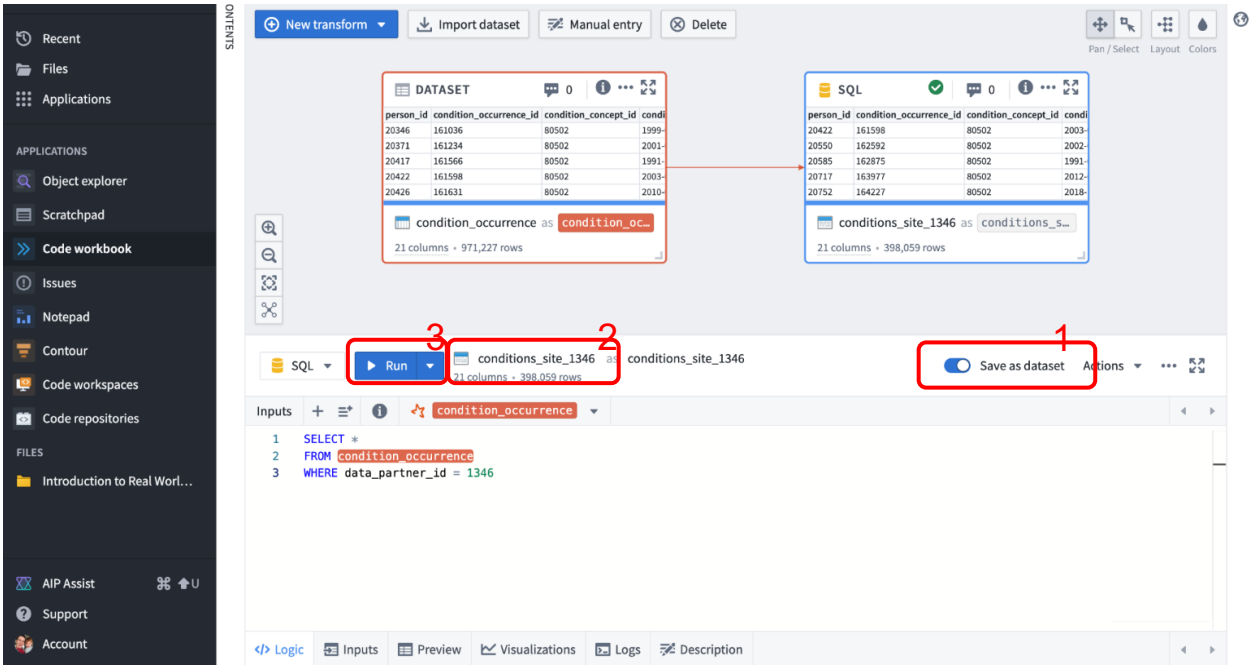
Next, add a transform to the dataset by clicking the blue + "Add Transform" button, and selecting **SQL code**. This will create a new SQL "transform" node, with an arrow connecting the dataset and the transform. The lower panel will show a code editor, and some tabs you may want to browse.

Edit the code to read `SELECT * FROM condition_occurrence WHERE data_partner_id = 1346` as shown, and click **Preview** to run the code. This will result in the code being run, and the result being shown in the Preview tab.

(Side note: `data_partner_id` is not a standard column of OMOP data, but is an N3C addition indicating which data partner each row of data comes from. Although these data are notional, we've added `data_partner_id` columns for realism.)

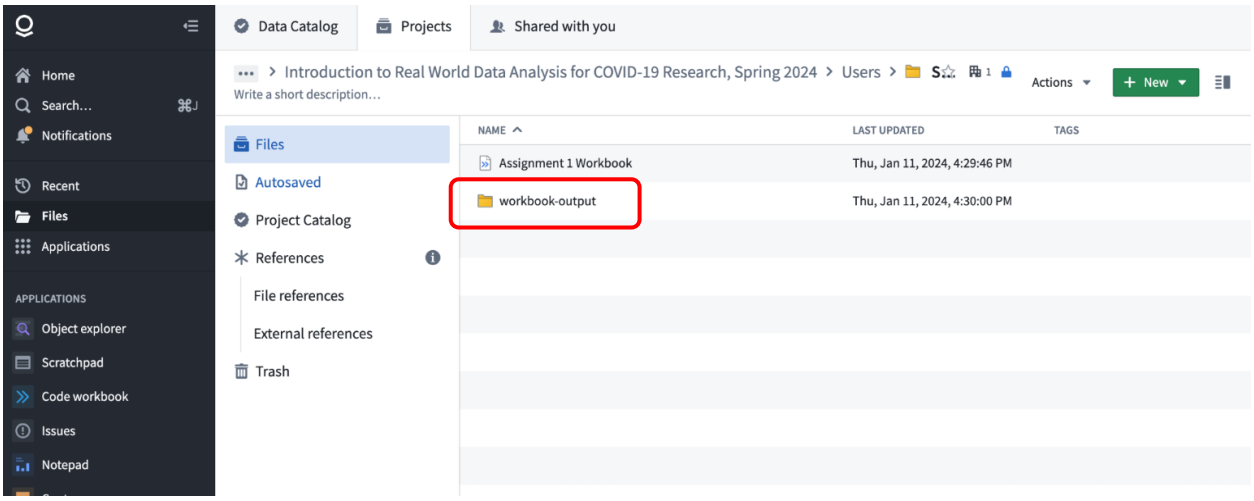


Now, we've only previewed this result - it is not saved out as a dataset, and if we were to re-open this workbook we would lose the preview. It is much more common to use the **Save as dataset feature** (box 1), giving the output reasonable filename a name like conditions_site_1346 (box 2; you'll also be asked if you want to rename the node in the workbook, say yes), and clicking Run (box 3).



Running the node executes the code, and this time the result is saved to a dataset in the enclave. **NOTE: if the code or upstream input datasets change, the result dataset will *not* be updated unless it is explicitly re-run.**

Finally, you should be able to navigate back to your personal folder (where you created the Assignment 1 Workbook), and see a new **workbook-output** folder. In here you can navigate to the **Assignment 1 Workbook** folder (named after the code workbook by default), and find your conditions_site_1346 dataset for review!

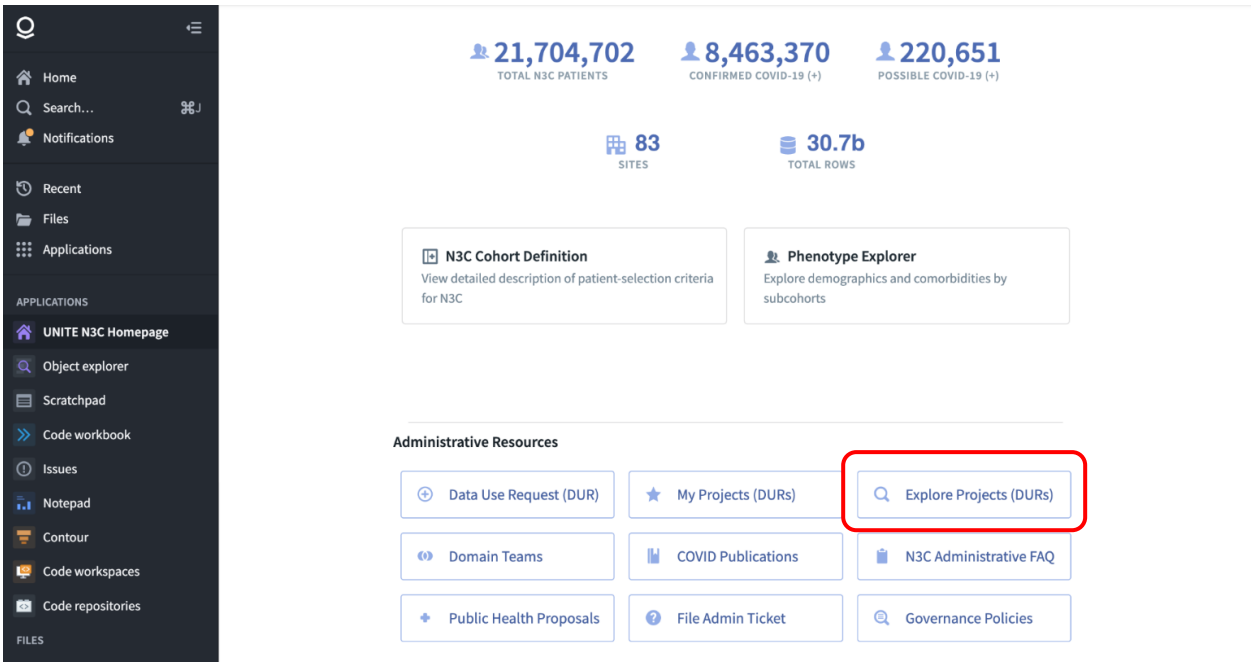


Part 3 - Joining a DUR

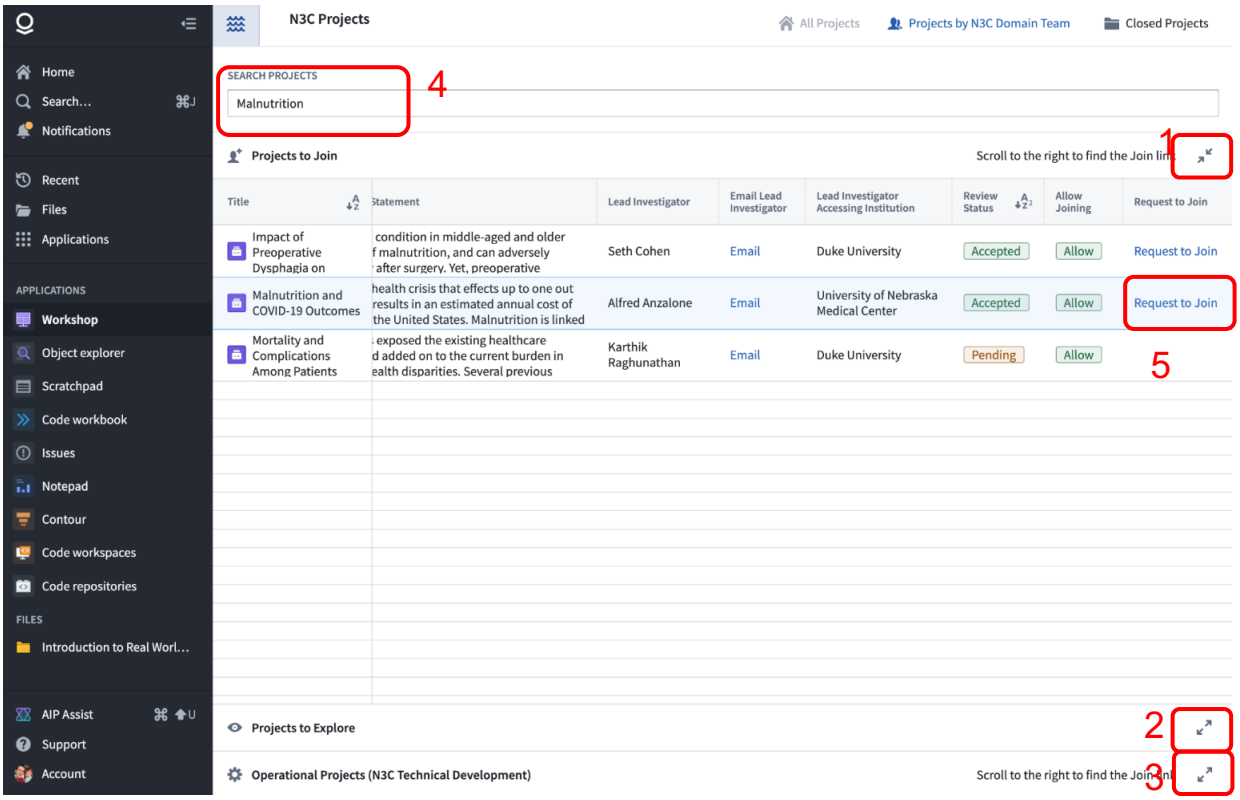
In a later part of this course we will be working with real Level 2 data in a research project workspace. To access the workspace, we need to join the DUR and get approved. We're going to find the DUR via the N3C dashboard, but here's a direct link just in case you have trouble: Malnutrition and COVID-19 Outcomes.

If you have any questions or issues, please reach out to your course instructors! For more information on N3C DURs in general, check out this chapter in the Guide to N3C, and this tutorial video.

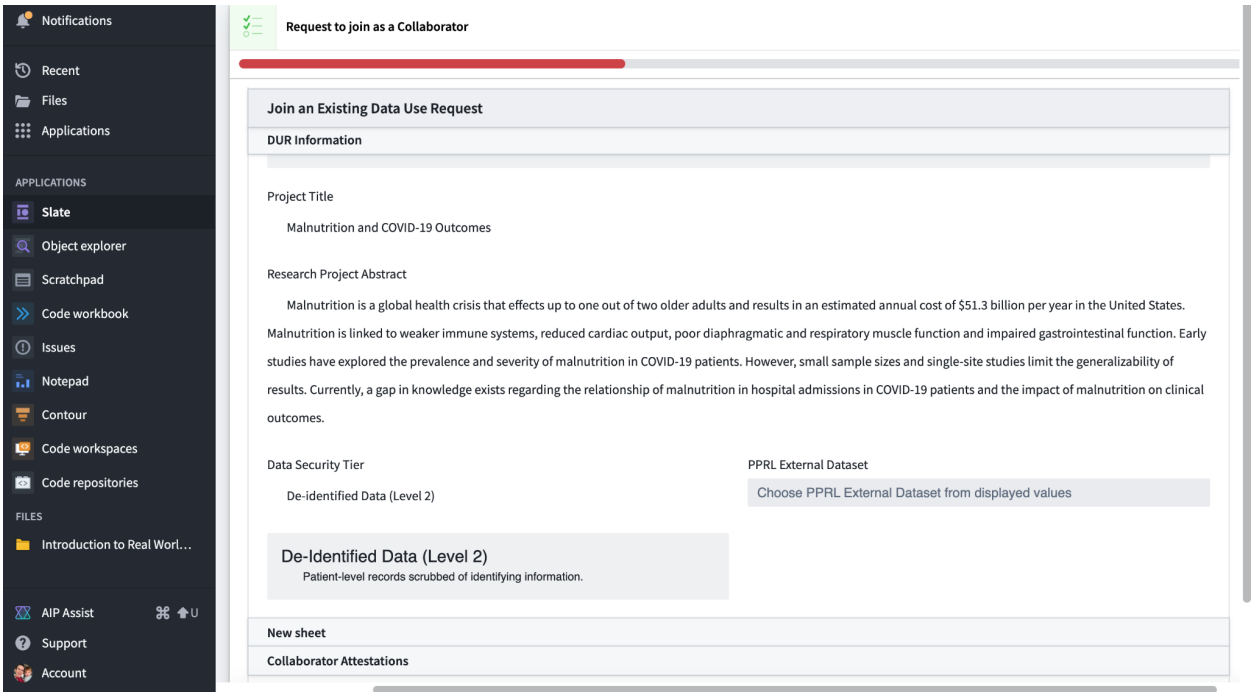
First, from the Enclave homepage, open the **Explore Projects** dashboard:



This will open a dashboard showing all N3C DURs, which come in three kinds: those configured as joinable by the project lead (box 1), those not listed as joinable (box 2), and operational DURs used by N3C staff (box 3). Use the search box to look for "Malnutrition", and in the entry for **Malnutrition and COVID-19 Outcomes** (box 4), scroll to the right to find the corresponding **Request to Join** link (box 5), and open it.



This will open the DUR form for the project, in which the project lead has specified a Title, Abstract, and Data Level for the DUR.



To join the DUR and eventually get access to the project workspace, scroll down and fill out the required information:

- Read and attest to the institutional DUA.

- Verify that you have completed the required NIH IT Security training in the past year (the (i) button provides a link - the 2023 refresher is the needed one).

- Verify that you have completed appropriate Human Subjects Research Protection Training in the past 3 years, and provide the date of completion. (More information here.)

- Read and attest to the User Code of Conduct.

- Answer the question “Does your Institution policy require IRB review for use of Level 2 data?”
    - Note that most institutions do not require IRB review for Level 2 (De-Identified) data, which is commonly not considered human subjects research. However, we cannot guarantee this for your institution, and you should check with your IRB office if unsure.

- Read and attest to the N3C Download policy.
    - Importantly, you are not allowed to screenshot or record video of row-level patient data (notional data such as we have worked with in this assignment is ok, but be careful!) All results for publication, such as figures and tables, must be approved by a submission-and-review process before they can be exported.

- Acknowledge that if you plan to use publicly available datasets alongside patient data you will need to ensure it is consistent with your institution’s policies.
    - Some institutions regard linking De-Identified patient data with public data (e.g. US Census or other data) an increased risk of identification, potentially triggering the need for IRB review otherwise not required. Again, this is not typical, but we cannot guarantee this for your institution, and you should check with your IRB office if unsure.

- Finally, click Submit!

What happens next: Your request to join the DUR will be sent to the project lead (Dr. Anzalone) for approval; once approved by the lead, it will head to the N3C Data Access Committee (DAC) for approval; once approved, your account will be given permissions to access the corresponding workspace and you will receive an email as well.

To see the status of your DUR request, you can navigate to the **My Projects** dashboard of the enclave homepage.