# GCP Workshop

# *Data Driven AI Application for Healthcare and Life Sciences*

## Accelerate Vaccine Research with Google Cloud

### BigQuery, AI Platform and Healthcare APIs

## Course overview

*Duration is 2 min*

The labs introduce participants to Google BigQuery,  BigQuery Machine Learning and Notebook driven Data Analysis. In the labs, you learn how to analyze, and visualize data using Google BigQuery for peptidic epitope data sourced from NIH and available through BigQuery Public Dataset program. BigQuery has a sandbox through which you can try it out without having to sign up for Google Cloud (or having to provide a credit card).

- https://towardsdatascience.com/bigquery-without-a-credit-card-discover-learn-and-share-199e08d4a064
- To learn more about BigQuery Public Dataset and how to take advantage of free processing, you can visit: https://cloud.google.com/bigquery/public-data
- To learn more specific about Google Cloud for Healthcare and Life Sciences visit: https://cloud.google.com/solutions/healthcare-life-sciences

# Lab Setup and Requirements

## Qwiklabs setup

*Duration is 2 min*

**Before you click the Start Lab button**

Read these instructions. Labs are timed and you cannot pause them. The timer starts when you click Start Lab, shows how long Cloud resources will be made available. This Qwiklabs hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access the Google Cloud Platform for the duration of the lab.
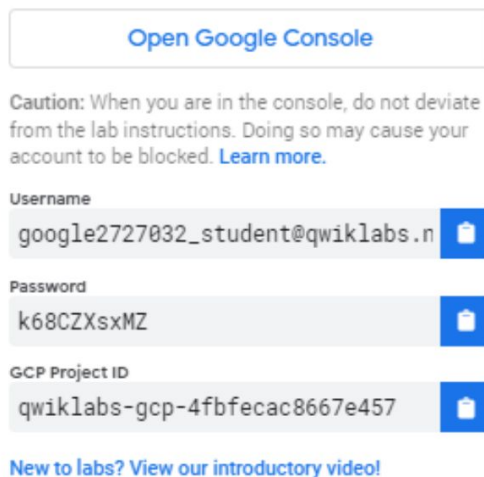
**What you need**

To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).
- *Note:* If you already have your own personal GCP account or project, do not use it for this lab.

**How to start your lab and sign in to the Console**

1. Click the **Start Lab** button. If you need to pay for the lab, a pop-up opens for you to select your payment method. On the left is a panel populated with the temporary credentials that you must use for this lab.
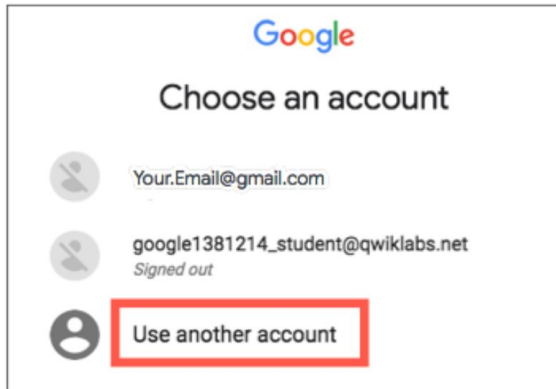
2. Copy the username, and then click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Choose an account** page.

   *Tip:* Open the tabs in separate windows, side-by-side.

3. On the Choose an account page, click **Use Another Account**.



4. The Sign in page opens. Paste the username that you copied from the Connection Details panel. Then copy and paste the password.

   *Important:* You must use the credentials from the Connection Details panel. Do not use your Qwiklabs credentials. If you have your own GCP account, do not use it for this lab (avoids incurring charges).

5. Click through the subsequent pages:
   - Accept the terms and conditions.
   - Do not add recovery options or two-factor authentication (because this is a temporary account).
   - Do not sign up for free trials.

   After a few moments, the GCP console opens in this tab.

- **Note:** You can view the menu with a list of GCP Products and Services by clicking the **Navigation menu** at the top-left, next to "Google Cloud Platform".
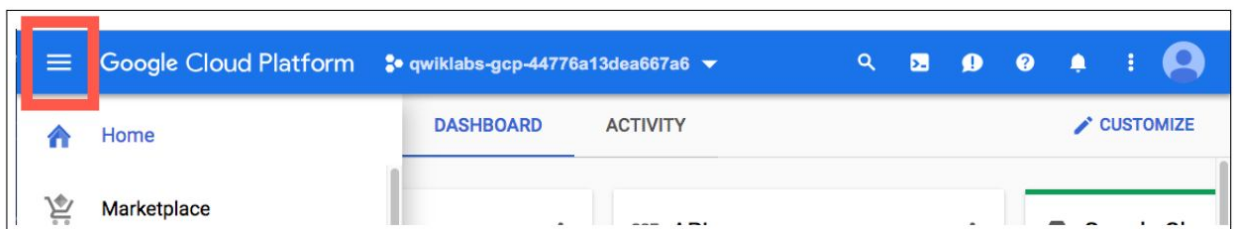
# Lab 1: Using BigQuery Interfaces

*Duration is 7 min*

**Task:** In this section, you explore the epitope public dataset using the query builder in the BigQuery web UI. Scroll down to the BIG DATA services to select BigQuery.
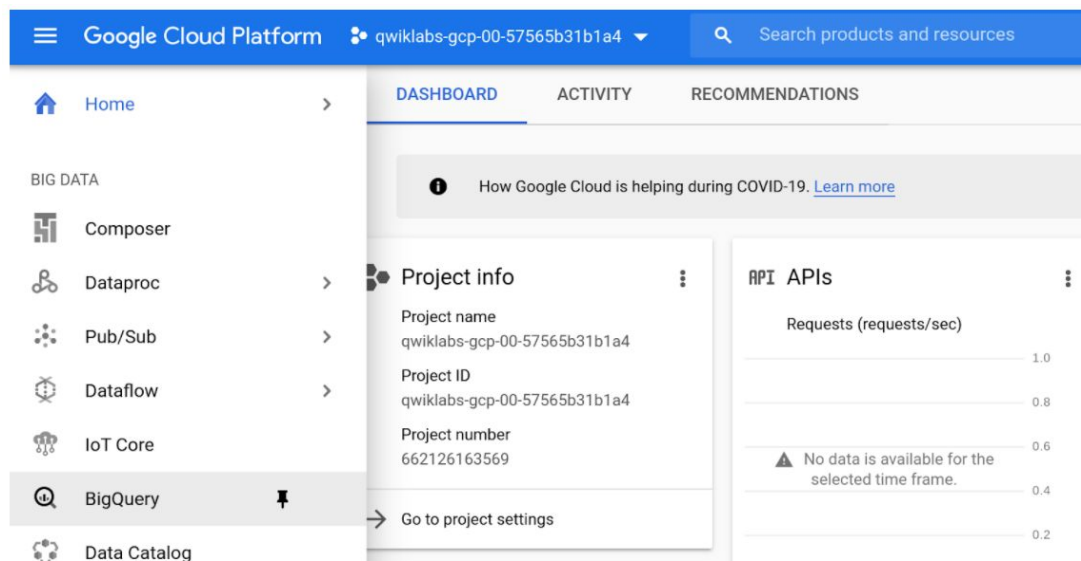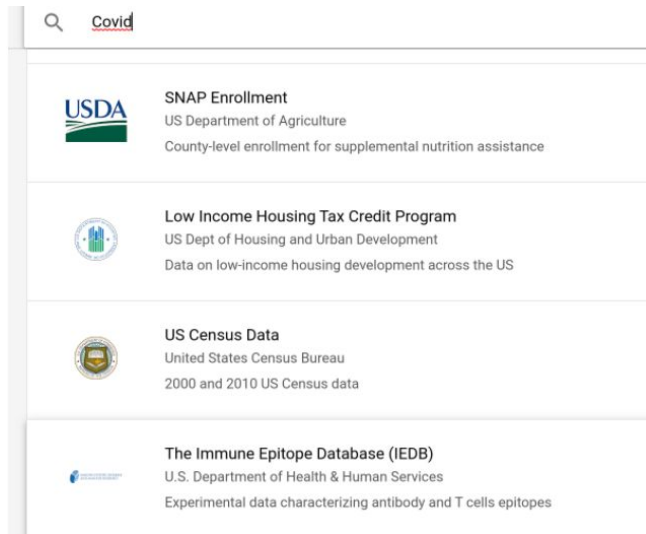


## Step 1

While on BigQuery, click on **+ ADD DATA** and select Explore **public datasets**. BigQuery provides many public datasets that you can leverage for your analysis. Explore all different datasets available.



Jignesh Mehta
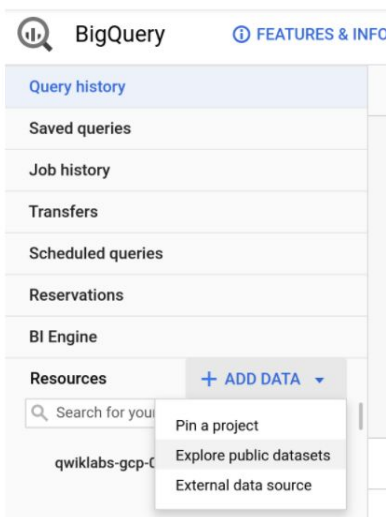
# Step 2

Search for Covid on the search box for the Public Dataset page and scroll down to select The Immune Epitope Database.



# Step 3

Read about the IEDB dataset that is sourced from NIH. Scroll down to sample queries and **click on the first example of "Run this Query".** This query provides results for epitopes in a dataset for a given organism of interest, for example 'coronavirus'.



Selecting 'Run this query' will bring you back to BQ Analysis UI with the query already displayed for you. Click on the **Run** button to see the query result.

The result shows Antigen Name with total counts of epitopes assays for coronavirus available in our dataset, ordered by most number of epitopes.

For example, the resulting query shows we have 1623 epitopes for the Spike glycoprotein from an organism coronavirus.



# Step 4

On a different tab, where you had a public dataset information page for IEDB data, click on 'VIEW DATASET' to see a full list of tables for epitopes.

On BigQuery UI, you would see IEDB public dataset and all the tables within. Click on any table and select preview on the right side to see sample data, for an example, select mhc_ligand_full table and click preview.

# Step 5

Lets run few more queries to get answer for key data insights

In the Query Editor section on the right, copy or write the following query to identify

which mhc allele has most references in a binding affinity data for coronavirus?

```
SELECT
 allele_name as allele,
count(*) as counts
 FROM `bigquery-public-data.immune_epitope_db.mhc_ligand_full`
WHERE
object_type = 'Linear peptide'
AND   organism_name like '%coronavirus%'
AND   antigen_name like 'Spike%'
 GROUP BY allele_name
 ORDER By  counts desc
```

Click **Run** to review the result.

## Query editor

```
 1  SELECT
 2   allele_name as allele,
 3  count(*) as counts
 4   FROM `bigquery-public-data.immune_epitope_db.mhc_ligand_full`
 5   WHERE
 6   object_type = 'Linear peptide'
 7   AND  organism_name like '%coronavirus%'
 8   AND  antigen_name  like 'Spike%'
 9   GROUP BY allele_name
10   ORDER By  counts desc
11
```

▶ Run ▾    ⬆ Save query    ⠿ Save view    🕐 Schedule query ▾    ⚙ More ▾

## Query results    ⬆ SAVE RESULTS    📊 EXPLORE DATA ▾

Query complete (0.6 sec elapsed, 125.7 MB processed)

Job information    **Results**    JSON    Execution details

| Row | allele | counts |
|-----|--------|--------|
| 1 | HLA-A*01:01 | 95 |
| 2 | HLA-DRB1*01:01 | 87 |
| 3 | HLA-A*02:01 | 85 |
| 4 | HLA-A*26:01 | 82 |
| 5 | HLA-A*24:02 | 80 |
| 6 | HLA-A*03:01 | 75 |
| 7 | HLA-A*30:02 | 67 |
| 8 | HLA-A*23:01 | 67 |
| ⁹ | HLA-A*30:02 | 67 |

In the Query Editor section on the right, replace the query with a new query as shown below to identify what is the most effective length of linear peptide of spike protein for ligand-peptide binding reference? Click on **Run** to review the result.

```
SELECT
length(Description) as peptide_mers,
count(*) as counts
 FROM `bigquery-public-data.immune_epitope_db.mhc_ligand_full`
 WHERE
object_type = 'Linear peptide'
 AND   organism_name like '%coronavirus%'
 AND   antigen_name like 'Spike%'
 GROUP BY   1
 ORDER BY   counts desc
```

```
1  SELECT
2  length(Description) as peptide_mers,
3  count(*) as counts
4   FROM `bigquery-public-data.immune_epitope_db.mhc_ligand_full`
5   WHERE
6  object_type = 'Linear peptide'
7   AND  organism_name like '%coronavirus%'
8   AND  antigen_name  like 'Spike%'
9   GROUP BY   1
10  ORDER BY counts desc
```

**▶ Run** ▼ | ⬆ Save query | Save view | ⏲ Schedule query ▼ | ⚙ More ▼

**Query results**  ⬆ SAVE RESULTS  ⋔ EXPLORE DATA ▼

Query complete (0.7 sec elapsed, 125.3 MB processed)

Job information   **Results**   JSON   Execution details

| Row | peptide_mers | counts |
|-----|--------------|--------|
| 1   | 9            | 874    |
| 2   | 10           | 407    |
| 3   | 15           | 87     |

In the Query Editor section on the right, replace the query with a new query as shown below to identify what is the strong binding affinity between 9 or 10 mer peptide of spike protein with HLA allele? List results in order of high positive binding to negative based on quantitative measure. Click on **Run** to review the result.

```
SELECT
antigen_name,
Description as peptide,
Parent_protein,
assay_group as binding_type,
allele_name as allele,
        qualitative_measure,
quantitative_measurement as result_score
FROM `bigquery-public-data.immune_epitope_db.mhc_ligand_full`
WHERE
object_type = 'Linear peptide'
AND     organism_name like '%coronavirus%'
AND     assay_group   != 'qualitative binding'
AND     allele_name like 'HLA-%'
AND     antigen_name like 'Spike%'
AND     length(Description) IN (9,10)
ORDER BY     result_score
```

# Step 6

**Important**: This step is necessary to build and run machine learning models within BigQuery.

You will create a new dataset named 'corona' for your BigQuery project. This dataset will hold your own tables and models as you need them. On the BigQuery UI left panel, highlight the name of your project and then click on **+ CREATE DATASET** from right panel middle toolbar section.



Give the dataset a name 'corona', leave the rest of the parameters to defaults and click on the **Create Dataset** button.

Newly created dataset 'corona'  is visible on the left panel under your project. Think of a dataset as equivalent to 'schema' of training databases.



# Step 7

To create a table within a dataset, you have multiple options such as creating a table manually, create a data transfer job or even just point to a file on Google storage or from your own local machine. You will download the following file, click on the url below to get a csv file that describes the amino acids basic properties.

https://storage.cloud.google.com/bq_epitope_workshop/files/amino_acids.csv

Once you have downloaded the file, highlight the 'corona' dataset from the left panel and click on **+CREATE TABLE** button from the middle toolbar on the right panel.

On the create table wizard, select as following:
Create table from: **Upload**
Select file: browse and select the **csv** you just downloaded
File format: **CSV**
Table name: **amino_acid**
Scheme - Auto detect: **check** to auto create table from CSV
Expand Advanced Options to set
 - Header rows to skip: 1

Leave the rest of the parameters to default values and click on the **Create table** button.

Create table

Source

| Create table from: | Select file: ⓘ | | File format: |
|---|---|---|---|
| Upload ▼ | amino_acids.csv | Browse | CSV ▼ |

Destination

⦿ Search for a project      ◯ Enter a project name

| Project name | Dataset name | Table type ⓘ |
|---|---|---|
| qwiklabs-gcp-00-4bf2aa883446 ▼ | corona ▼ | Native table ▼ |

Table name

amino_acid

Schema

Auto detect
☑ Schema and input parameters

> ⓘ Schema will be automatically generated.

Advanced options ⌃

Write preference:

Write if empty ▼

Number of errors allowed: ⓘ          Unknown values: ⓘ

0                                       ☐ Ignore unknown values

Field delimiter: ⓘ

Comma ▼

Header rows to skip: ⓘ      Quoted newlines: ⓘ          Jagged rows: ⓘ

1                           ☐ Allow quoted newlines      ☐ Allow jagged rows
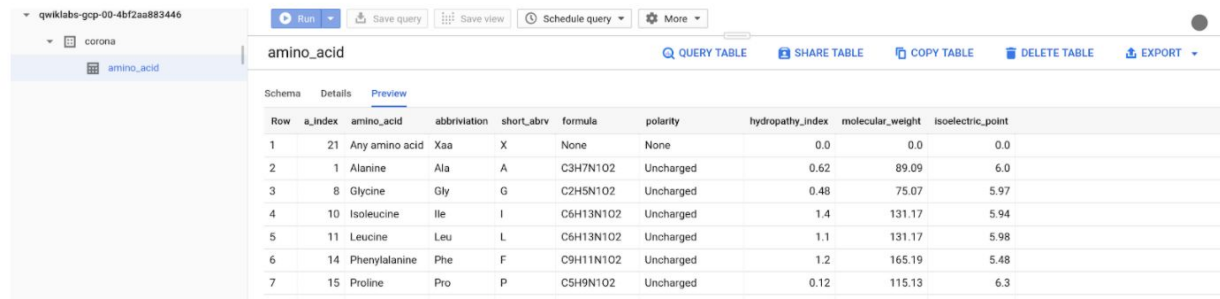
Encryption
Data is encrypted automatically. Select an encryption key management solution.
⦿ Google-managed key
    No configuration required
◯ Customer-managed key
    Manage via Google Cloud Key Management Service

Create table      Cancel

If you do not see the table created, refresh the page to make sure and review the table by selecting it.



End of Lab 1. Congratulations!

# Lab 2:

# Data Analysis with Data Science Notebook

*Duration is 8 min*

**Task:** In this section, you will wear a hat of data scientist and will perform analysis on epitope dataset to identify and narrow down the possible attributes of peptides to build machine learning models.  You will perform analysis using Jupyter notebook as a datalab from Google AI Platform.

## Step 1

First create a notebook environment within Google Cloud. GCP provides quick access to many frameworks and tools as well as choice of CPUs and GPUs for a notebook runtime environment. You will create a notebook with PyTorch 1.4 without GPU.
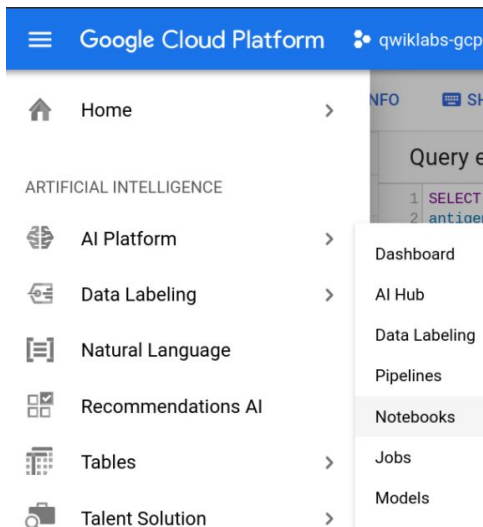
From the Home menu on Google Cloud panel, select Notebook within Artificial

Intelligence services.



Click on **+ NEW INSTANCE** and select PyTorch 1.4 Without GPU.



You can leave default settings or select a region close to your data, usually for these labs, you

can keep region settings to us-central1. Click on the **CREATE** button. In about 30 seconds your

notebook environment will be ready to use.

Once ready, open the notebook instance by clicking on **OPEN JUPYTERLAB.**

**Click on Python 3** Notebook to launch the notebook editor panel.
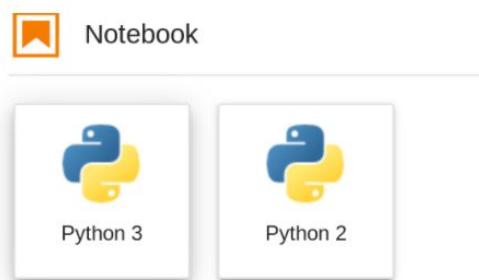


Make sure, you have completed Lab 1 to create a dataset named 'corona', which is required for hosting the machine learning model you will create in the lab 3.
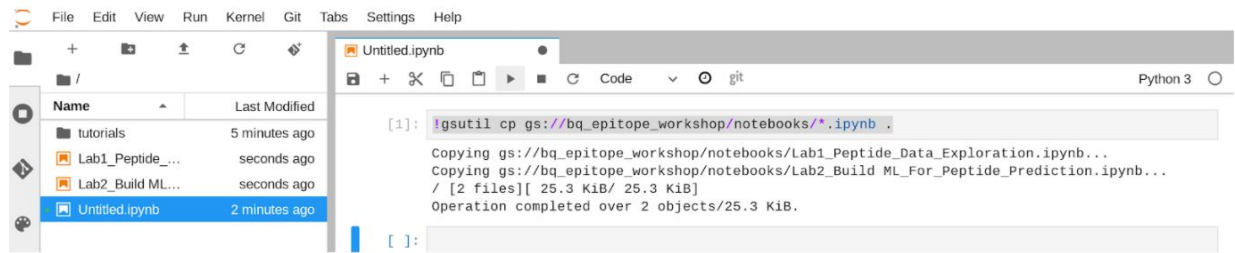
## Step 2

From file menu, create New → Notebook
Enter the following command in a section as shown below and click on the 'Run' / Play button to execute. Alternatively, you can execute any selected cell in a notebook by pressing shift+enter.

```
!gsutil cp gs://bq_epitope_workshop/notebooks/*.ipynb .
```

Command will copy two lab notebook files to your environment as shown below.

## Step 3

Double click on Lab1_Peptide_Data_exploration notebook to open it in an editor panel. The notebook cells have informational sections as well as executional sections. Notebooks are designed to be self-explanatory, hence read through it carefully and execute a cell in the order it is presented. To execute a cell, click on a play icon on the tool-bar of the notebook or press shift+enter. Observe the result of the cell you execute before proceeding to the next cell.
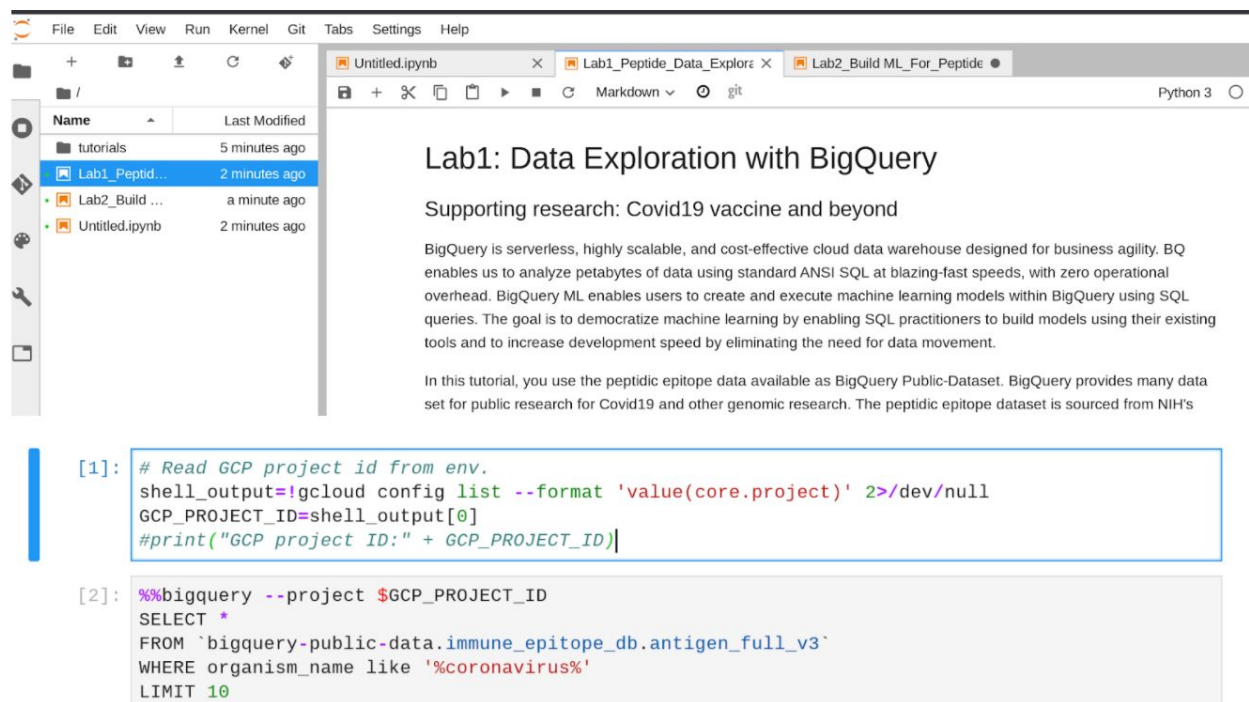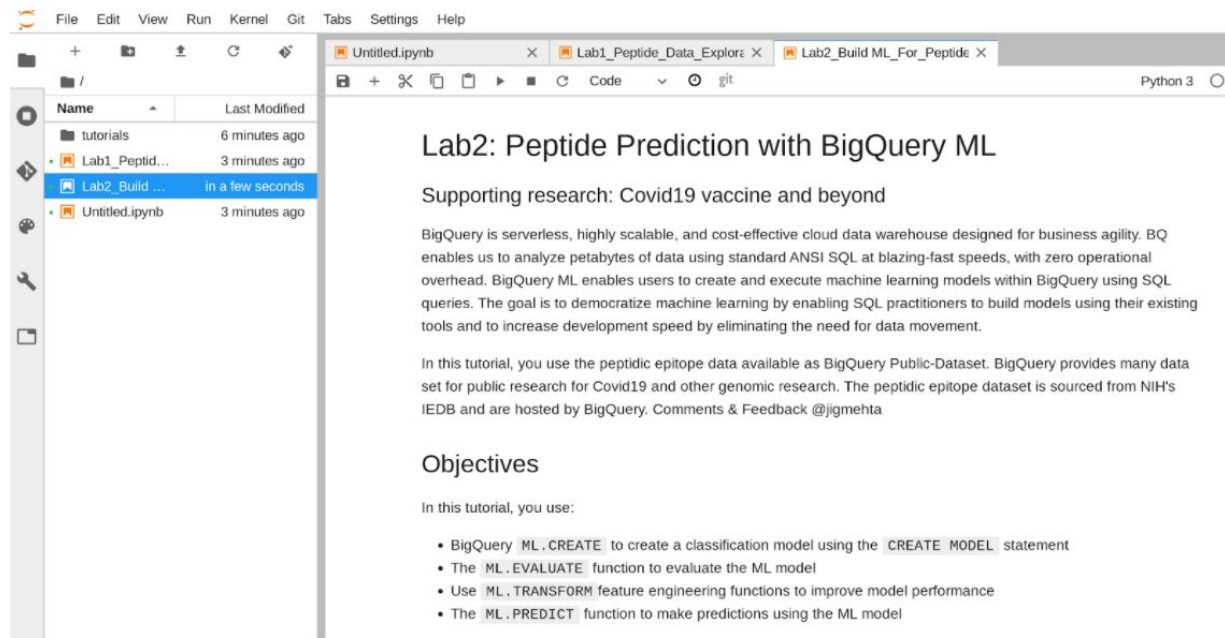


This completes Lab 2! Congratulations!

# Lab 3:

# Machine Learning Model with BigQuery

*Duration is 7 min*

**Task:** In this section, you will build a machine learning model with data already available in bigquery, by leveraging simple SQL statements. Then, you will evaluate the model and create another model to get better results from machine learning. At last, you will use the machine learning model you created to predict binding affinity of a peptide to a mhc ligand. Carefully execute cells in the order it is presented. To execute a cell, click on a play icon on the tool-bar of the notebook or press shift+enter. Observe the result of the cell you execute before proceeding to the next cell.

## Step 2: Build First Classification Model

Building ML models with BigQuery is as simple as writing SQL statements; makes ML modeling accessible to even SQL developers and analysts. We will create a simple classification model to predict for a given peptide if there is strong binding affinity with certain HLA Allele.

Following statement creates a classification model using logistic regression. We are selecting feature columns of Allele and peptide of specific mers to classify if a peptide is a good candidate for vaccine testing.

- Note: we are filtering data for peptides with length of 9 or 10 mers only. Also, since we run multiple samples, we are randomizing samples by 80% of data for learning.

```
[3]: %%bigquery --project $GCP_PROJECT_ID
CREATE OR REPLACE MODEL `corona.Classification_model_P1`
OPTIONS
(
model_type='logistic_reg',
input_label_cols=['Qualitative_Measure']
)
AS
SELECT
 Qualitative_Measure, Description, Allele_Name, Quantitative_measurement
 FROM
   `bigquery-public-data.immune_epitope_db.mhc_ligand_full`
 WHERE length(Description) IN (9,10)
 AND organism_name like '%coronavirus%'
 AND Assay_Group = 'qualitative binding'
```

Once you complete the notebook, you can monitor model artifacts and additional details such as evaluation and confusion matrix for a model within a BQ UI. Simply select a model from your project dataset and click on the right panel for Training and Evaluation tabs to review details of model characteristics.
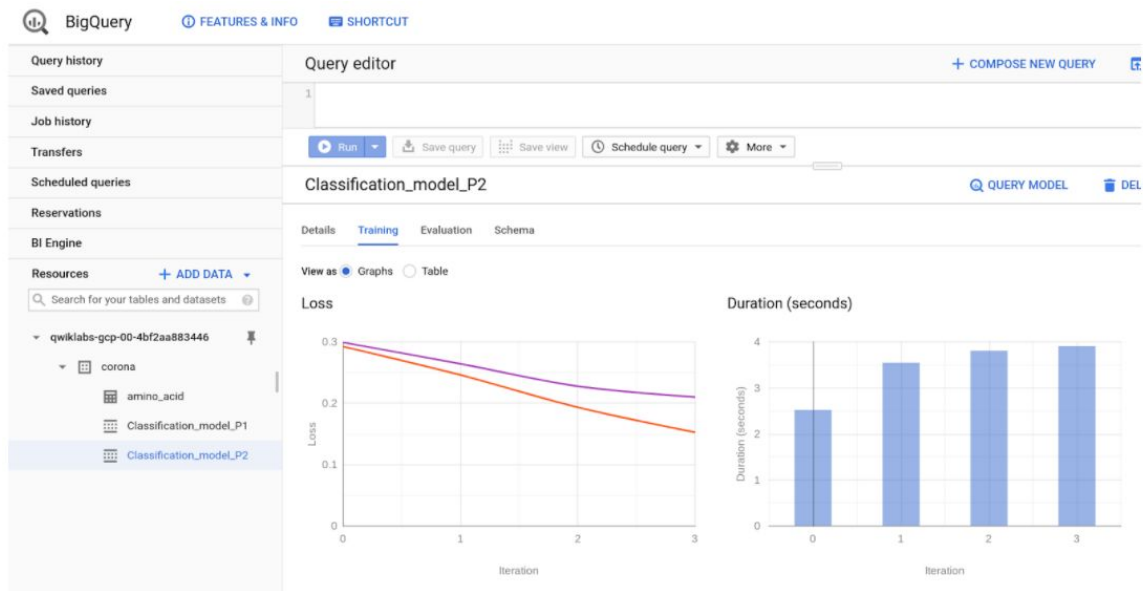
## Confusion matrix

This table shows the percentage of actual labels that were classified correctly (in blue) and incorrectly (in grey). The last column shows the percentage of total samples for the corresponding actual label.

| Actual labels | Predicted labels Negative | Positive | Positive-High | Positive-Intermediate | Positive-Low | % samples |
|---|---|---|---|---|---|---|
| Negative | 88.71% | 0.24% | 2.06% | 3.03% | 5.95% | 40.31% |
| Positive | 58.97% | 15.38% | 23.08% | - | 2.56% | 1.91% |
| Positive-High | 15.9% | - | 62.96% | 12.42% | 8.71% | 22.46% |
| Positive-Intermediate | 24.78% | - | 30.32% | 26.24% | 18.66% | 16.78% |
| Positive-Low | 34.04% | 0.26% | 14.78% | 19.53% | 31.4% | 18.54% |

This completes Lab 3! Congratulations!

# End your lab environment
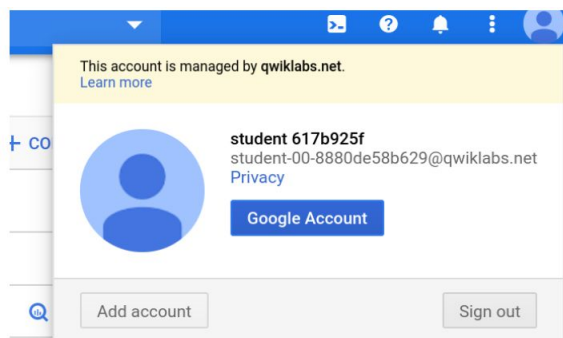
*Duration is 1 min*

## Step 1

From Google Cloud Console Home menu, select Notebooks from AI Platform services.
Select the notebook instance you have created, click on **DELETE** to close and remove
it.
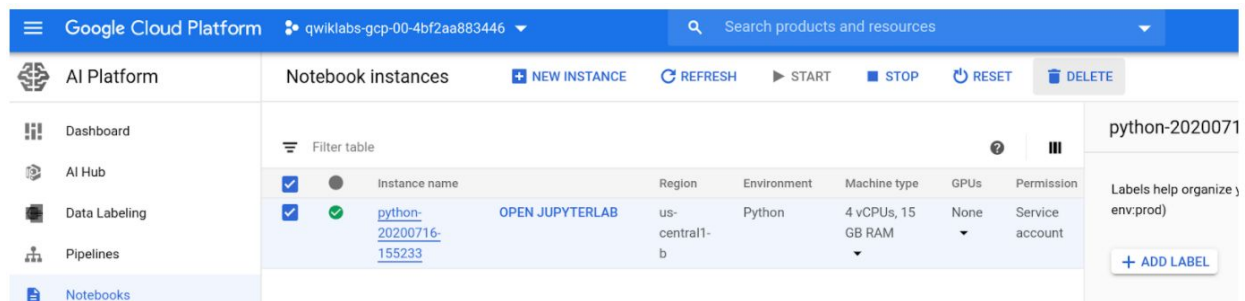


## Step 2

Sign out from your Google Cloud lab console.

When you've completed your lab, click **End**. Qwiklabs removes the resources you've
used and cleans the account for you.



## Step 3

You'll be asked to rate the lab experience. Select the applicable number of stars, enter a
comment, and then click **Submit**.

 Jignesh Mehta

**Note:** When you have completed your lab, click **End**. Qwiklabs removes the resources you've used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click **Submit**.

**Note:** The number of stars indicates the following:

- 1 star = Very dissatisfied
- 2 stars = Dissatisfied
- 3 stars = Neutral
- 4 stars = Satisfied
- 5 stars = Very satisfied

You may close the dialog if you don't want to provide feedback.

**Additional Resources**

- For more information about Google Cloud Training and Certification, see
  https://cloud.google.com/training/
- For more Google Cloud Platform Self-Paced Labs, see http://run.qwiklabs.com

For feedback, suggestions, or corrections, please use the **Support** tab.