

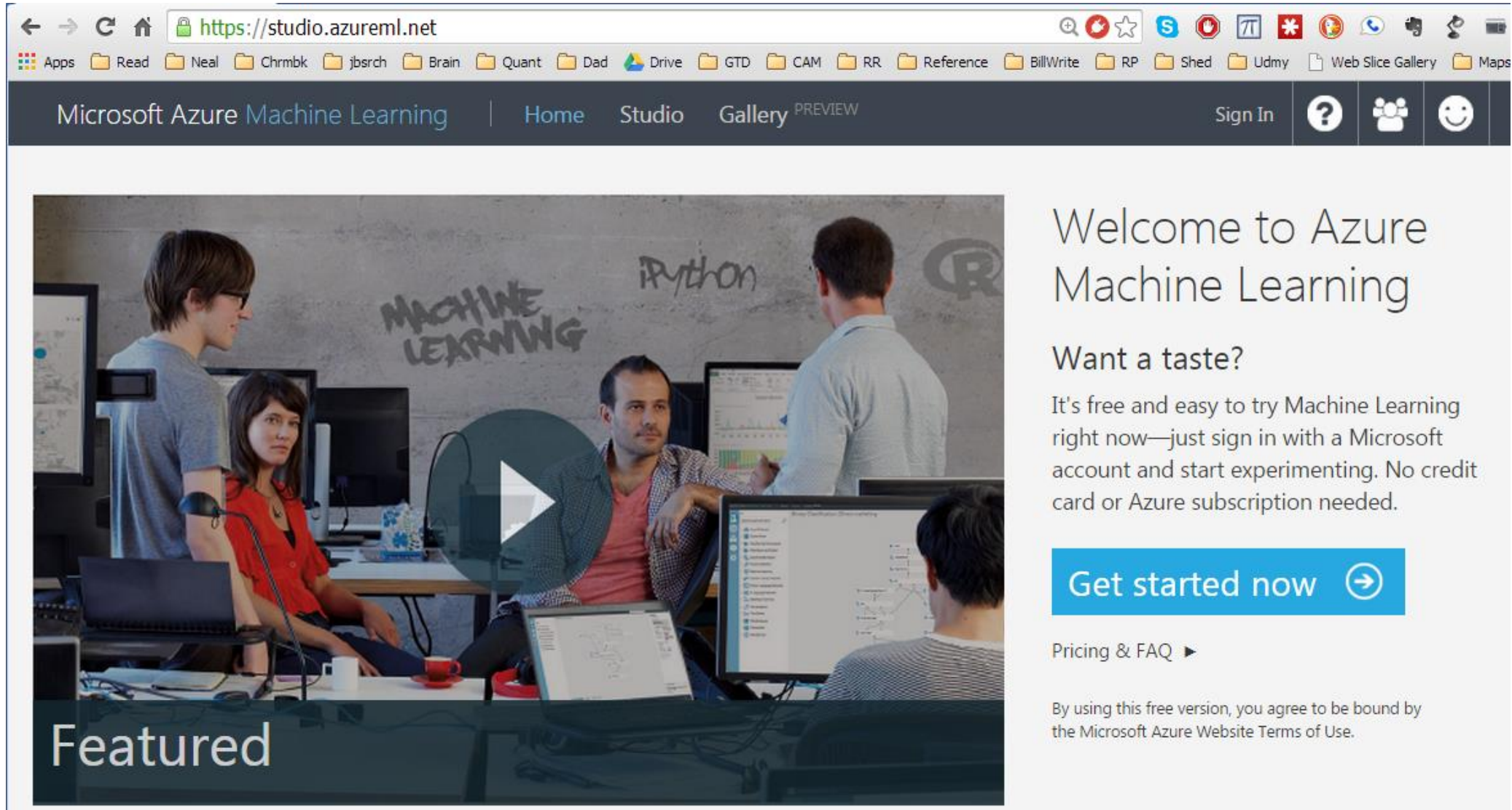


Import Data in Azure ML

Exploratory Data Analysis – Lab Session

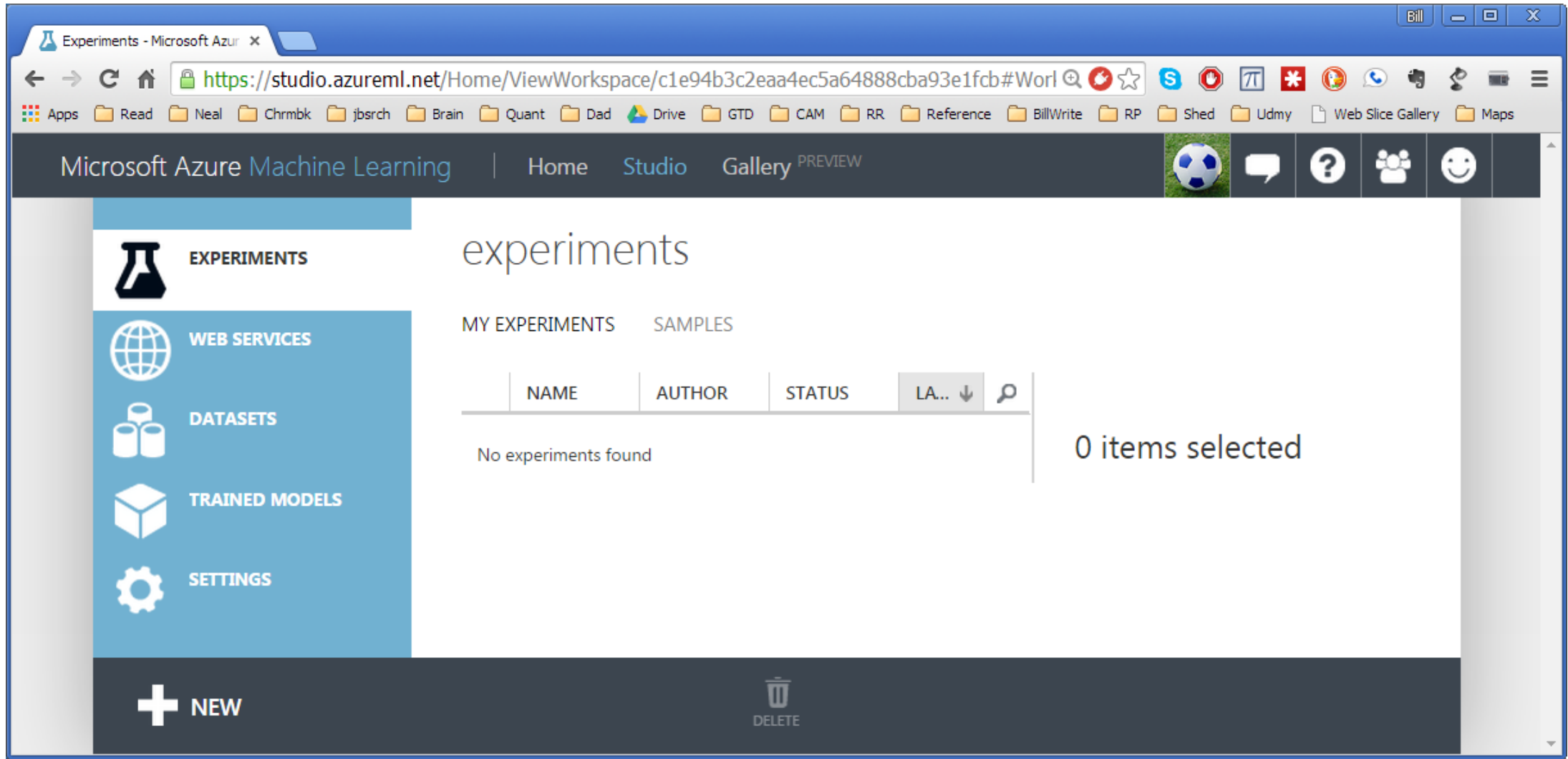
- Introduction to Notebooks
- Setting up and loading data into Notebook
- Data descriptive
- Data visualizations
- Univariate and Bi Variate plots
- Correlation
- Principal Component Analysis – dimensionality reduction
- Feature Engineering

Step 1 : Go to <https://studio.azureml.net/>

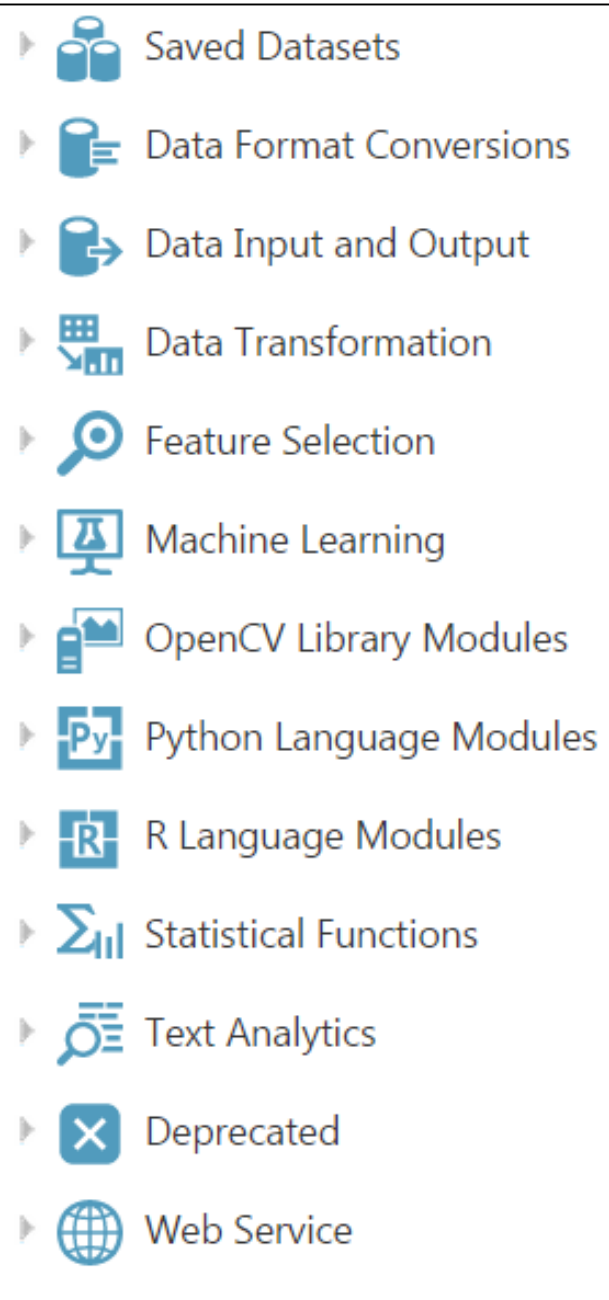


The screenshot shows the Microsoft Azure Machine Learning Studio homepage. The browser address bar displays <https://studio.azureml.net/>. The navigation bar includes links for Microsoft Azure Machine Learning, Home, Studio, and Gallery (marked as PREVIEW), along with a Sign In button and user icons. The main content area features a large video player with a play button overlay, showing a group of people working in a machine learning environment. To the right of the video, the text reads "Welcome to Azure Machine Learning" followed by "Want a taste?" and a description: "It's free and easy to try Machine Learning right now—just sign in with a Microsoft account and start experimenting. No credit card or Azure subscription needed." Below this is a prominent blue button labeled "Get started now" with a right arrow icon. Further down, there is a link for "Pricing & FAQ" and a disclaimer: "By using this free version, you agree to be bound by the Microsoft Azure Website Terms of Use." The word "Featured" is visible in the bottom left corner of the video area.

Step 2 : Log in to your account



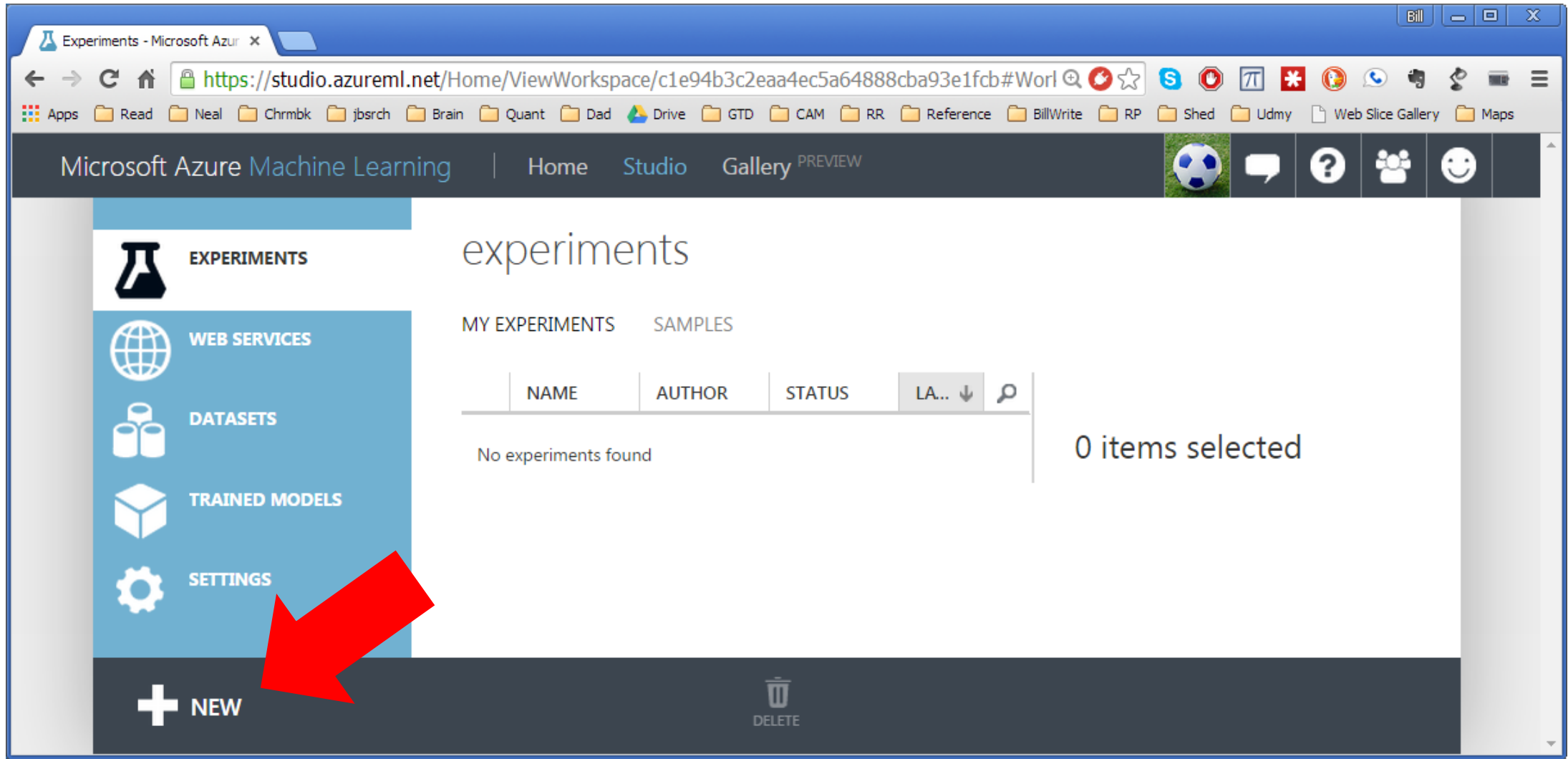
The screenshot shows the Microsoft Azure Machine Learning Studio interface. The browser address bar displays the URL: <https://studio.azureml.net/Home/ViewWorkspace/c1e94b3c2eaa4ec5a64888cba93e1fcb#World>. The top navigation bar includes 'Microsoft Azure Machine Learning', 'Home', 'Studio', and 'Gallery PREVIEW'. A sidebar on the left contains icons and labels for 'EXPERIMENTS', 'WEB SERVICES', 'DATASETS', 'TRAINED MODELS', and 'SETTINGS', along with a '+ NEW' button. The main content area is titled 'experiments' and features tabs for 'MY EXPERIMENTS' and 'SAMPLES'. Below these tabs is a table with columns: NAME, AUTHOR, STATUS, and LA... (with a dropdown arrow). The table is currently empty, displaying 'No experiments found'. To the right of the table, it says '0 items selected'. At the bottom of the main area, there is a 'DELETE' button with a trash icon.



Step 3.1 : Create New Experiment

- AML modelling ... a checklist approach
 - ☐ Create new experiment

Step 3.2 : Create a new experiment



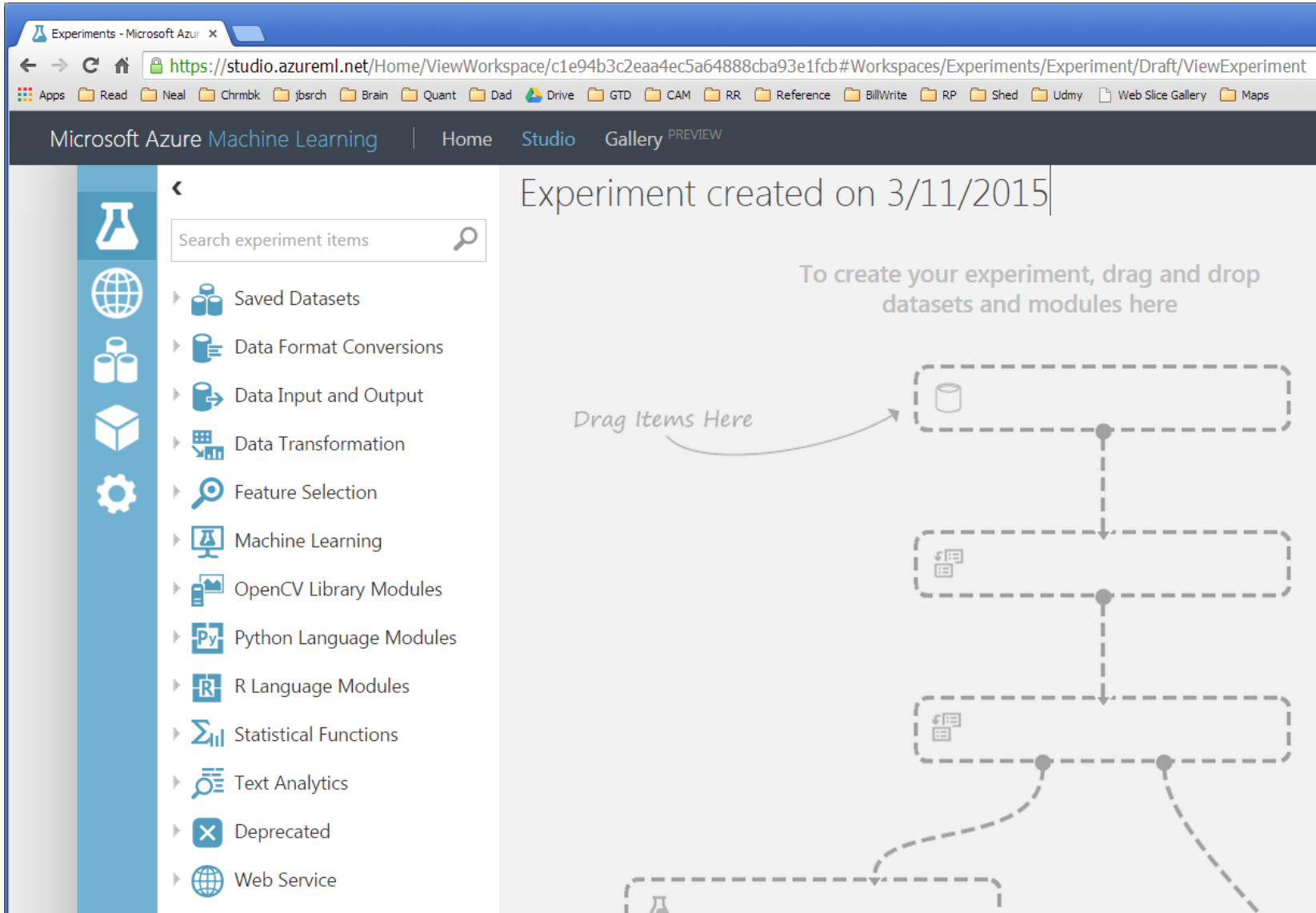
Step 3.3 : Create a new experiment

The screenshot shows the Microsoft Azure Machine Learning web interface. At the top, there's a dark header with the Microsoft Azure Machine Learning logo. Below it, a navigation bar includes 'PROJECTS' and 'EXPERIMENTS' (the latter is highlighted). The main area is titled 'experiments' and has tabs for 'MY EXPERIMENTS' and 'SAMPLES'. A table with columns 'NAME', 'AUTHOR', and 'STATUS' is visible. On the left, a 'NEW' sidebar lists options: DATASET, MODULE, PROJECT (with a 'PREVIEW' link), EXPERIMENT (highlighted with a red arrow labeled '1'), and NOTEBOOK (with a 'PREVIEW' link). The main content area features a search bar 'Search experiment templates' and a section titled 'Microsoft Samples'. It contains three cards: 'Blank Experiment' (with a plus icon and a red arrow labeled '2'), 'Experiment Tutorial' (with a green background and a right arrow icon), and 'Sample 1: Download dataset from UCI: Adult 2 class dataset' (with a database icon and a download arrow).

Step 3.4 : Slow down and look at the experiment

The screenshot displays the Microsoft Azure Machine Learning interface. The top navigation bar includes the title "Microsoft Azure Machine Learning", the current view "Advanced-Workshop", and user icons. The left sidebar lists various components: Saved Datasets, Trained Models, Transforms, Data Format Conversions, Data Input and Output, Data Transformation, Feature Selection, Machine Learning, OpenCV Library Modules, Python Language Modules, R Language Modules, Statistical Functions, Text Analytics, Web Service, and Deprecated. A red box highlights the "Search experiment items" input field, with a callout labeled "Filter dialog box". A red bracket groups the entire sidebar, with a callout labeled "Sources for building blocks". The central canvas, titled "Experiment created on 3/15/2016" and "In draft", shows a workflow diagram with dashed boxes and arrows. A callout labeled "Canvas on which to drag building blocks, to build models, component by component" points to the canvas area. The right sidebar contains "Properties" and "Project" tabs, with "Experiment Properties" showing "STATUS CODE" as "InDraft" and a "Summary" section for describing the experiment. A "Quick Help" section is at the bottom right. The bottom toolbar includes icons for "NEW", "RUN HISTORY", "SAVE", "SAVE AS", "DISCARD CHANGES", "RUN", "SET UP WEB SERVICE", and "PUBLISH TO GALLERY".

Step 3.5 : This is a “drag and drop” environment



Azure Machine Learning is to modelling as ...

Inserting shapes into PowerPoint is to drawing














= Modelling by dragging and dropping

Step 3.6 : Name your experiment



- Click on the title box at the top that says "Experiment Created on"
- Give it a name of your choice

Step 4.1 : Import the dataset

- ▶  Saved Datasets
- ▶  Data Format Conversions
- ▶  Data Input and Output
- ▶  Data Transformation
- ▶  Feature Selection
- ▶  Machine Learning
- ▶  OpenCV Library Modules
- ▶  Python Language Modules
- ▶  R Language Modules
- ▶  Statistical Functions
- ▶  Text Analytics
- ▶  Deprecated
- ▶  Web Service

- AML modelling ... a checklist approach
 - ☒ Create new experiment
 - ☐ Import data set

Import Data

- We have three options for importing data -
 - Get data from Hive (seamless and most optimal)
 - Get data from Azure Blob Storage
 - Get data from csv
- Final datasets for EDA --

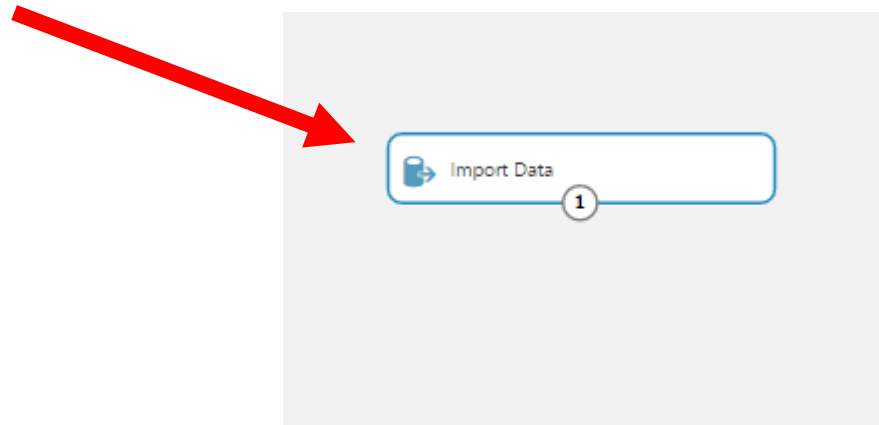
Scenario	Dataset Name
Dynamometer	Dynamometer Card Reading Data Raw Dynamometer Clusters
Predictive Maintenance of Compressors	Predictive Maintenance Raw Data
Tank Level Forecasting	Tank Level Forecasting Raw Data

Option 1 -
Import the training dataset
using a Hive Query. If option 1
did not work, skip to option 2.

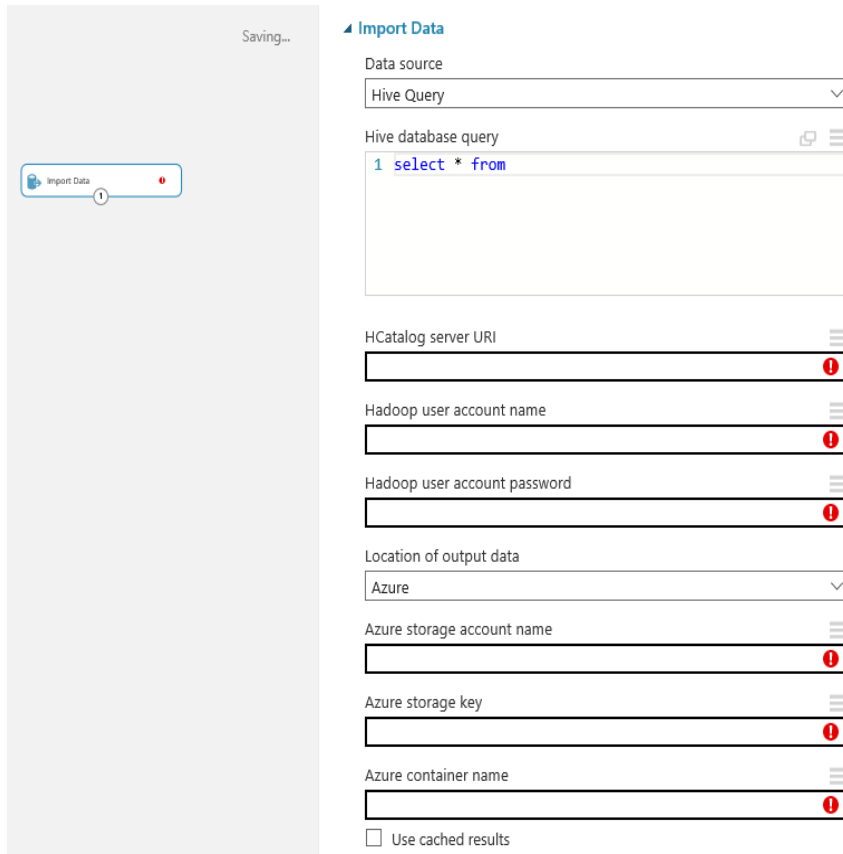
Step 4.2 : Q? How to import data using Hive Query?

- ▶ Saved Datasets
- ▶ Trained Models
- ▶ Transforms
- ▶ Data Format Conversions
- ▶ **Data Input and Output**
 - Enter Data Manually
 - Export Data
 - Import Data**
 - Unpack Zipped Datasets
- ▶ Data Transformation
- ▶ Feature Selection
- ▶ Machine Learning
- ▶ OpenCV Library Modules

- Open "Data Input and Output" from the navigation pane at the left
- Drag "Import Data" to the canvas
 - "Import Data" loads data from sources such as the Web, Azure SQL, Windows Azure Blob storage, etc.



Step 4.3 : Q? How to import data using Hive Query?



The screenshot shows the 'Import Data' configuration window. On the left, there is a sidebar with a button labeled 'Import Data' and a status indicator. The main area is titled 'Import Data' and contains several fields: 'Data source' (set to 'Hive Query'), 'Hive database query' (containing '1 select * from'), 'HCatalog server URI', 'Hadoop user account name', 'Hadoop user account password', 'Location of output data' (set to 'Azure'), 'Azure storage account name', 'Azure storage key', and 'Azure container name'. Each of these fields has a red exclamation mark icon next to it, indicating a warning or error. At the bottom, there is a checkbox labeled 'Use cached results'.

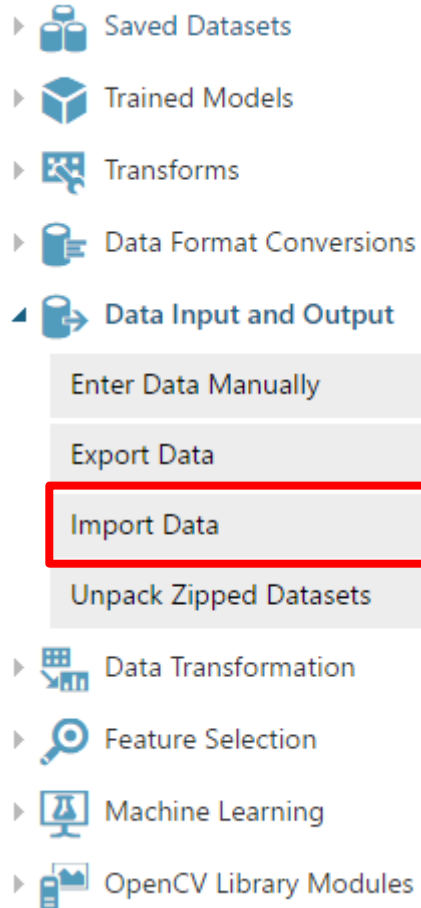
1. Machine Learning
 - Click on Import Data
 - Chose "HiveQuery" for Data source
 - Add in Hive query to pull in data
 - Update information on the cluster and storage account details
2. Pull data in as per scenario
3. Need to add select * from <tablename>
4. Add in Data Import module specific for each data set.

Scenario	Dataset Name
Dynamometer	Dynamometer Card Reading Data Raw Dynamometer Clusters
Predictive Maintenance of Compressors	Predictive Maintenance Raw Data
Tank Level Forecasting	Tank Level Forecasting Raw Data

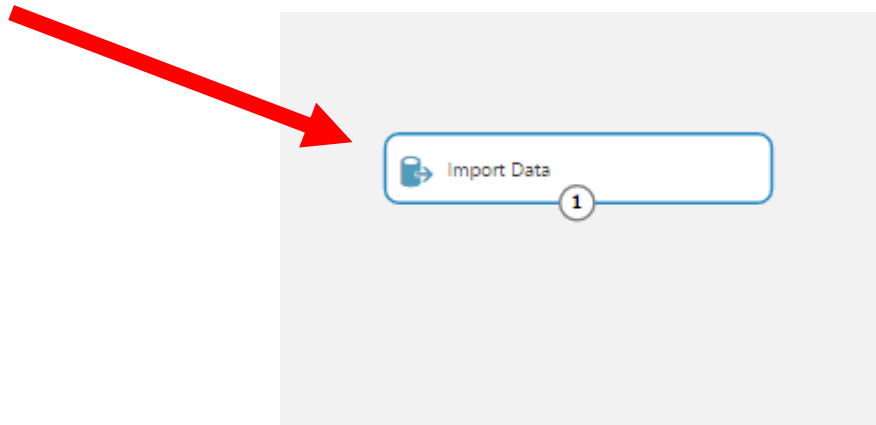
If Option 1 worked, move on to
Step 5.1

Option 2 -
Import the training dataset
from the Azure Blob Storage.

Step 4.4 : Q? How to import data from Azure Blob?



- Open “Data Input and Output” from the navigation pane at the left
- Drag “Import Data” to the canvas
 - “Import Data” loads data from sources such as the Web, Azure SQL, Windows Azure Blob storage, etc



Step 4.5 : Q? How to import data from Azure Blob?

- Machine Learning
 - Click on Import Data
 - Chose "Azure Blob Storage" for Data source
 - Chose "Storage Account" for Authentication type
 - Copy and Paste the following information without the quotes in the Import Data module
 - Account Name – **"nealworkshop"**
 - Account key –
"RER9c7kfM1e67p7p7gl+TbkE5Y6alzURg4PQc9lew+l8O+ZfU58gFjNgBW/WQm0u8N0YZQUG+wlalzFWKxyljA=="
 - Path to container – **"/oilandgas/_____.csv"**
(Look next slide for the dataset name)
 - Check the File has header row check box

Draft saved at 2:15:18 PM

Import Data

1

Import Data

Data source
Azure Blob Storage

Authentication type
Storage Account

Account name
[Redacted]

Account key
[Redacted]

Path to container, directory..
[Redacted]

Blob file format
CSV

☐ File has header row

☐ Use cached results

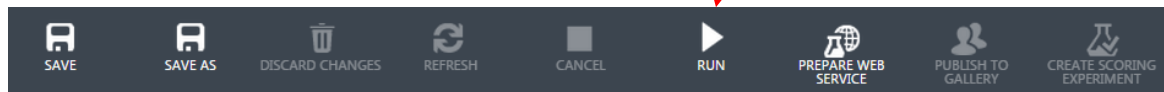
Step 4.6 : Q? How to import data from Azure Blob?

Fill the _____.CSV in the previous slide depending on which scenario you picked

Scenario	Dataset Name
Dynamometer	Dynamometer Card Reading Data Raw Dynamometer Clusters
Predictive Maintenance of Compressors	Predictive Maintenance Raw Data
Tank Level Forecasting	Tank Level Forecasting Raw Data

Step 1: Click "Save"

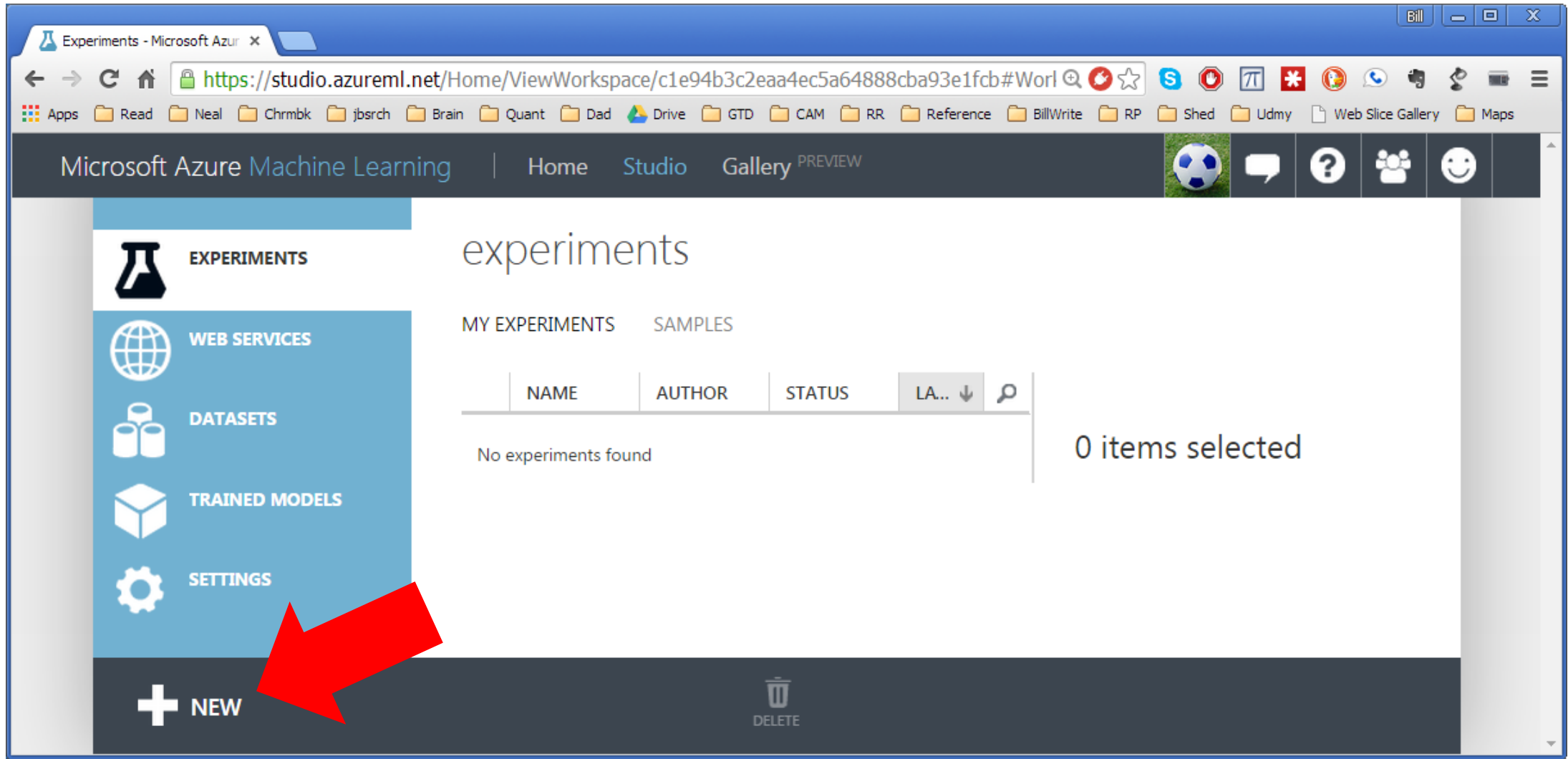
Step 2: Click "Run"



If Option 2 worked, move on to
Step 5.1

Option 3 -
Import the training dataset
from the saved datasets.

Step 4.7 : Import this tutorial's training dataset



The screenshot shows the Microsoft Azure Machine Learning Studio interface. The browser address bar displays the URL: <https://studio.azureml.net/Home/ViewWorkspace/c1e94b3c2eaa4ec5a64888cba93e1fcb#World>. The top navigation bar includes 'Microsoft Azure Machine Learning', 'Home', 'Studio', and 'Gallery PREVIEW'. The left sidebar contains a menu with icons and labels: 'EXPERIMENTS' (flask icon), 'WEB SERVICES' (globe icon), 'DATASETS' (cylinders icon), 'TRAINED MODELS' (cube icon), 'SETTINGS' (gear icon), and a '+ NEW' button at the bottom. A large red arrow points to the '+ NEW' button. The main content area is titled 'experiments' and has tabs for 'MY EXPERIMENTS' and 'SAMPLES'. Below the tabs is a table with columns: 'NAME', 'AUTHOR', 'STATUS', and 'LA...'. The table is empty, with the text 'No experiments found' below it. To the right of the table, it says '0 items selected'. At the bottom of the main area, there is a 'DELETE' button with a trash icon.

Step 4.8 : Import this tutorial's training dataset

The screenshot shows the Microsoft Azure Machine Learning web interface. At the top, there's a dark header with the Microsoft Azure Machine Learning logo. Below it, a light blue sidebar contains navigation options: 'PROJECTS' (with a folder icon) and 'EXPERIMENTS' (with a flask icon). The main area is titled 'experiments' and has tabs for 'MY EXPERIMENTS' and 'SAMPLES'. Below these tabs is a table with columns 'NAME' and 'AUTHOR'. A large dark grey panel at the bottom is titled 'NEW' and contains a list of options: 'DATASET' (with a folder icon), 'MODULE' (with a grid icon), 'PROJECT' (with a grid icon and a 'PREVIEW' label), 'EXPERIMENT' (with a flask icon), and 'NOTEBOOK' (with a document icon and a 'PREVIEW' label). A large red arrow points to the 'DATASET' option. To the right of the 'NEW' panel, there's a section titled 'FROM LOCAL FILE' with a folder icon and the text 'Upload a new dataset from a local file'.

Step 4.9 : Import this tutorial's training dataset

The screenshot shows the Microsoft Azure Machine Learning Studio interface. A modal dialog titled "Upload a new dataset" is open. The dialog contains the following fields and options:

- SELECT THE DATA TO UPLOAD:** A button labeled "Choose File" and a text field showing "No file chosen".
- ☐ This is the new version of an existing dataset
- ENTER A NAME FOR THE NEW DATASET:** A text input field.
- SELECT A TYPE FOR THE NEW DATASET:** A dropdown menu with the text "Select a dataset type..." and a downward arrow.
- PROVIDE AN OPTIONAL DESCRIPTION:** A text area.

A large red arrow points from the left sidebar towards the "Choose File" button. The sidebar on the left has a "DATASETS" section highlighted. The background of the studio shows a workspace with various icons and a "selected" label.

Step 4.10 : Import this tutorial's training dataset

Microsoft Azure Machine Learning Studio

Advanced-Workshop

PROJECTS

EXPERIMENTS

WEB SERVICES

NOTEBOOKS

DATASETS

TRAINED MODELS

SETTINGS

experiments

MY EXPERIMENTS SAMPLES

	NAME	AUTHOR	STATUS	LAST EDITED	PROJECT
	Oil & Gas - Tank Level For...	achal_mallaya	Draft	9/28/2016 5:39:58 PM	None
	Oil & Gas - Tank Level For...	achal_mallaya	Finished	9/28/2016 5:25:19 PM	None
	OK Training - Tank Level F...	sailaja.karthik	Draft	9/28/2016 5:19:53 PM	None
	Oil & Gas - Brine Analysis ...	achal_mallaya	Draft	9/28/2016 3:53:47 PM	Oklahoma Training Worksh...
	Oil & Gas - Dynamometer ...	achal_mallaya	Draft	9/28/2016 1:58:14 PM	Oklahoma Training Worksh...
	Oil & Gas - Dynamometer ...	achal_mallaya	Draft	9/28/2016 1:57:51 PM	Oklahoma Training Worksh...
	Oil & Gas - Dynamometer ...	achal_mallaya	Draft	9/28/2016 1:57:30 PM	Oklahoma Training Worksh...
	Oil & Gas - Binary Classifi...	achal_mallaya	Draft	9/28/2016 5:33:52 AM	Oklahoma Training Worksh...
	Oil & Gas - Linear Regressi...	achal_mallaya	Finished	9/28/2016 5:26:28 AM	Oklahoma Training Worksh...
	Oil & Gas - Linear Regressi...	achal_mallaya	Draft	9/28/2016 5:25:59 AM	Oklahoma Training Worksh...
	Oil & Gas - Dynamometer ...	achal_mallaya	Finished	9/27/2016 6:33:37 PM	Oklahoma Training Worksh...
	Oil & Gas - Linear Regressi...	Ryan	Draft	9/27/2016 4:33:20 PM	None
	OK Cluster, Classify, Regre...	eric.hullander	Finished	9/26/2016 12:32:05 PM	None
	Brine Analysis K Means Cl...	sailaja.karthik	Draft	9/26/2016 4:00:55 AM	None
	OK - K Means Clustering ...	sailaja.karthik	Finished	9/26/2016 3:54:17 AM	None
	OK Predictive Maintenan...	eric.hullander	Draft	9/22/2016 2:02:02 PM	OK Oil and Gas Workshop
	Oil & Gas - Binary Classifi...	sailaja.karthik	Failed	9/20/2016 12:30:00 PM	None
	Binary Classification: Brea...	Microsoft	Draft	9/14/2016 4:14:28 PM	None
	OK Predictive Maintenan...	Microsoft	Draft	9/14/2016 1:14:28 PM	OK Oil and Gas Workshop
	Aerospace W... Regres...	zperkel	Finished	9/14/2016 1:14:28 PM	Aerospace W...

1

2

→

Select Columns in Dataset

Normalize Data

Edit Metadata

Edit Metadata

Split Data

Select Columns in Dataset

Filter Based Feature Selection

Two-Class Decision Forest

SMOT

Two-Class Boosted Decision T...

Train Model

Train Model

Score Model

Score Model

Evaluate Model

Upload of the dataset 'Tank Level Forecasting Training Data.csv' has completed.

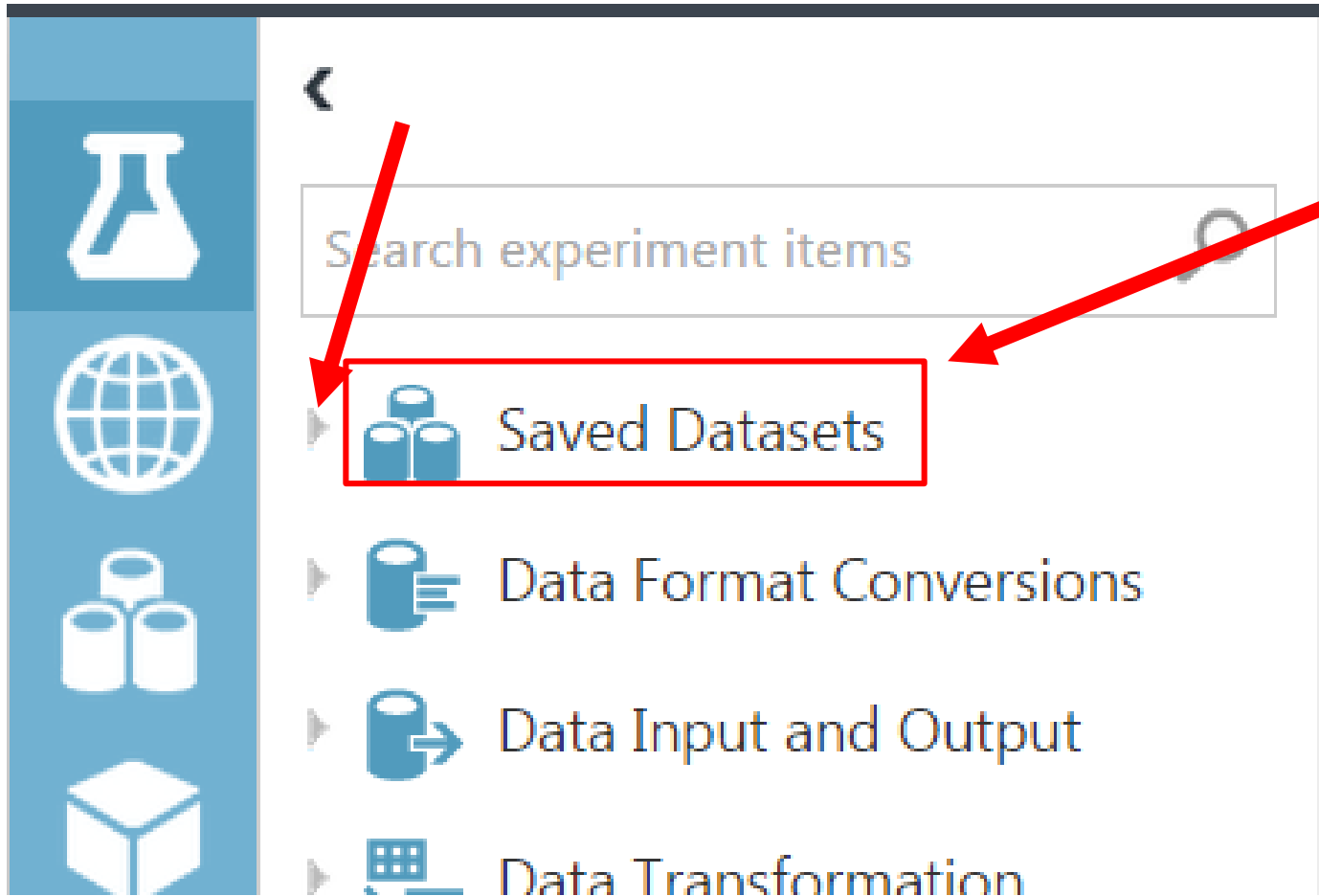
NEW

DELETE

ADD TO PROJECT

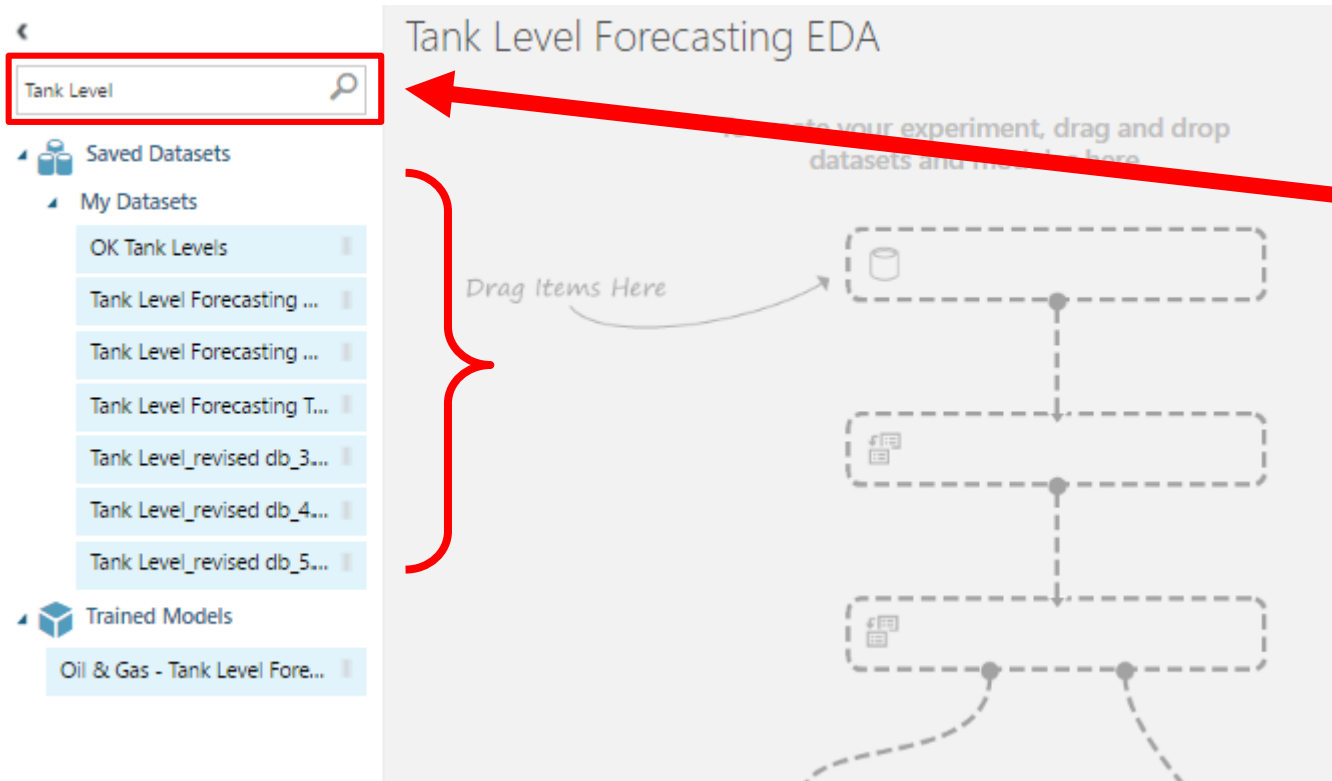
COPY TO

Step 4.11 : Open "Saved Datasets"



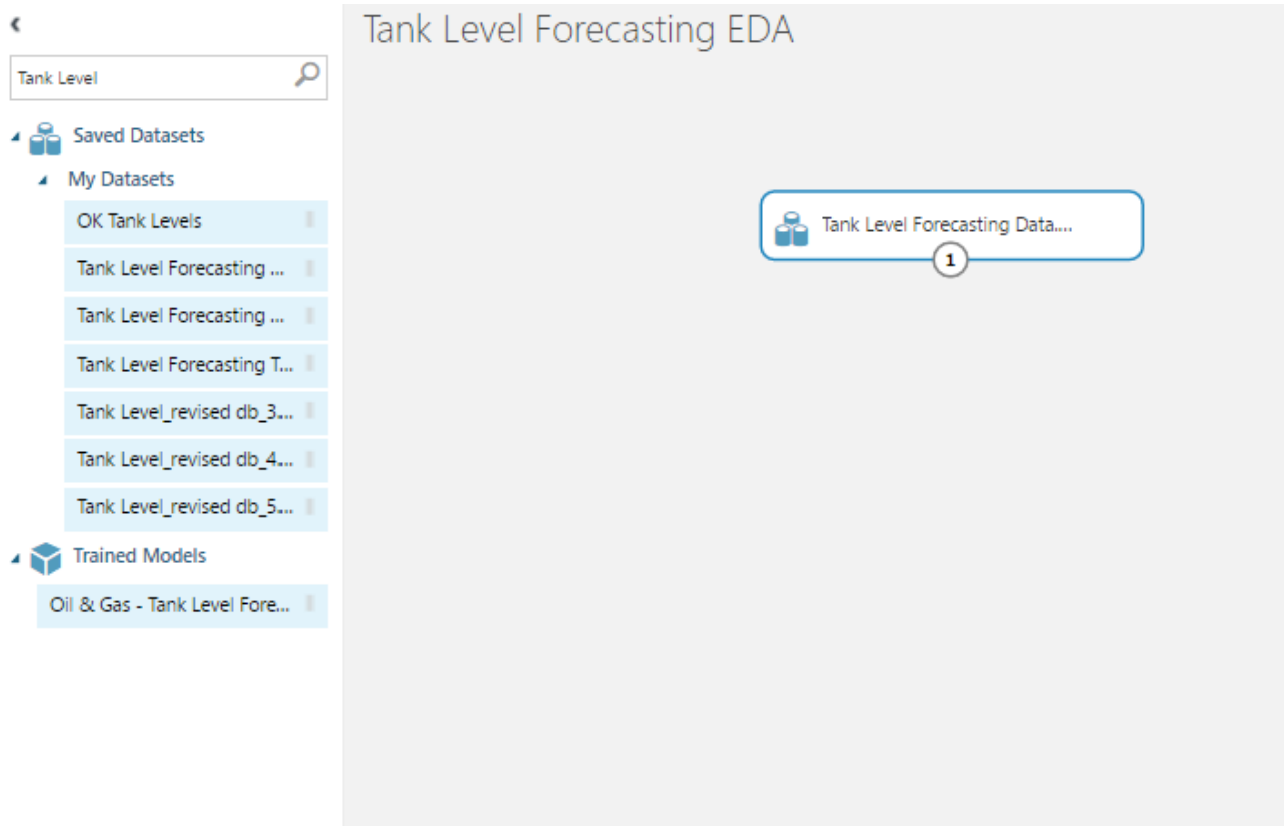
- By clicking on the triangle at the left of "Saved Datasets"

Step 4.12 : Take a second to notice the MANY datasets



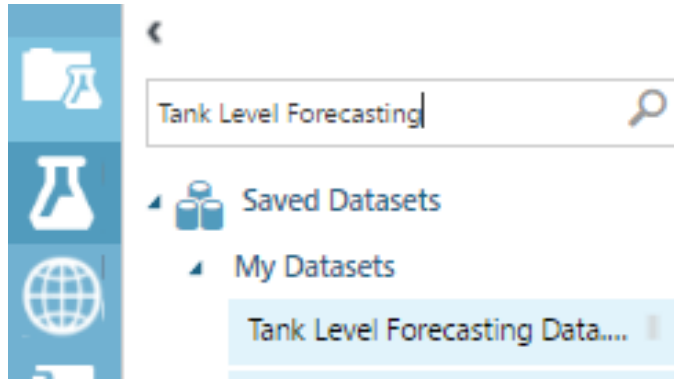
- To filter to the data set for this tutorial ...
- Type the name of your dataset in "Search experiment items" dialog box
- The data set list will reduce to our data set for this tutorial

Step 4.13 : Drag the data set to the experiment

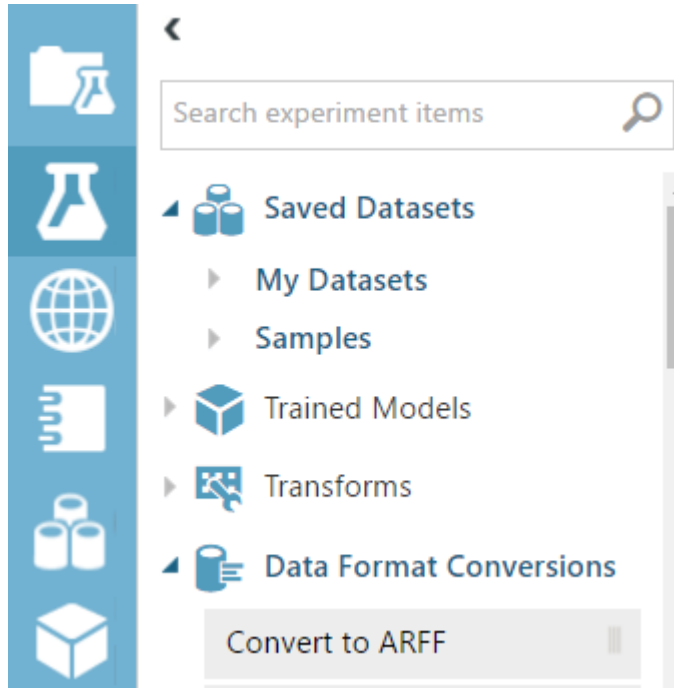


- *Note* when you drag the first element of your model to the canvas ... all the guides disappear
- Now, where are all the tools that were at the left Azure Machine Learning?
- They are still there, ... but we need to un-filter to see them

Step 4.14 : Backspace over "data"
















- Backspace to remove the dataset name from the "Search experiment items" dialog



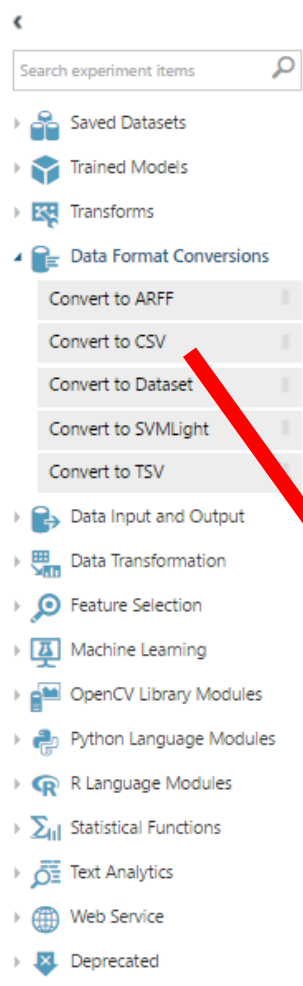
- Then click the triangle at left of "Saved Datasets" to close the dataset list

Step 5.1 : Convert to CSV

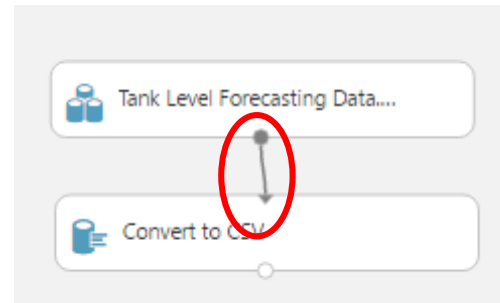
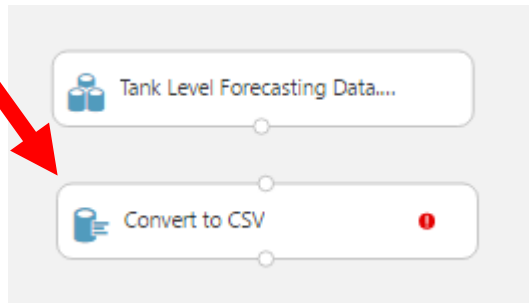
- ▶  Saved Datasets
- ▶  Data Format Conversions
- ▶  Data Input and Output
- ▶  Data Transformation
- ▶  Feature Selection
- ▶  Machine Learning
- ▶  OpenCV Library Modules
- ▶  Python Language Modules
- ▶  R Language Modules
- ▶  Statistical Functions
- ▶  Text Analytics
- ▶  Deprecated
- ▶  Web Service

- AML modelling ... a checklist approach
 - ☒ Create new experiment
 - ☒ Import data set
 - ☐ Convert to CSV

Step 5.2 : Add convert to CSV module to your experiment

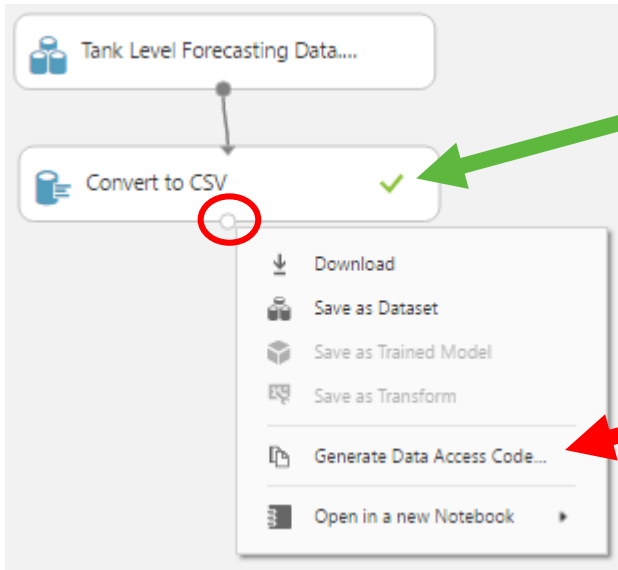
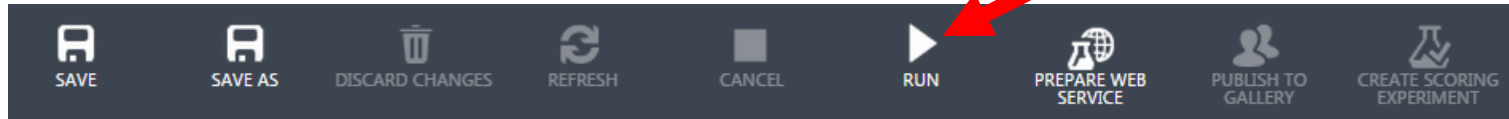


- Open "Data Format Conversion" from the navigation pane at the left
- Drag "Convert to CSV" to the canvas
 - "Convert to CSV" Converts data input to a comma-separated values format
- Next, click and hold on the bottom middle circle of your "Dataset" module
- While holding down the mouse button, drag the line to the top middle circle of "Convert to CSV" module



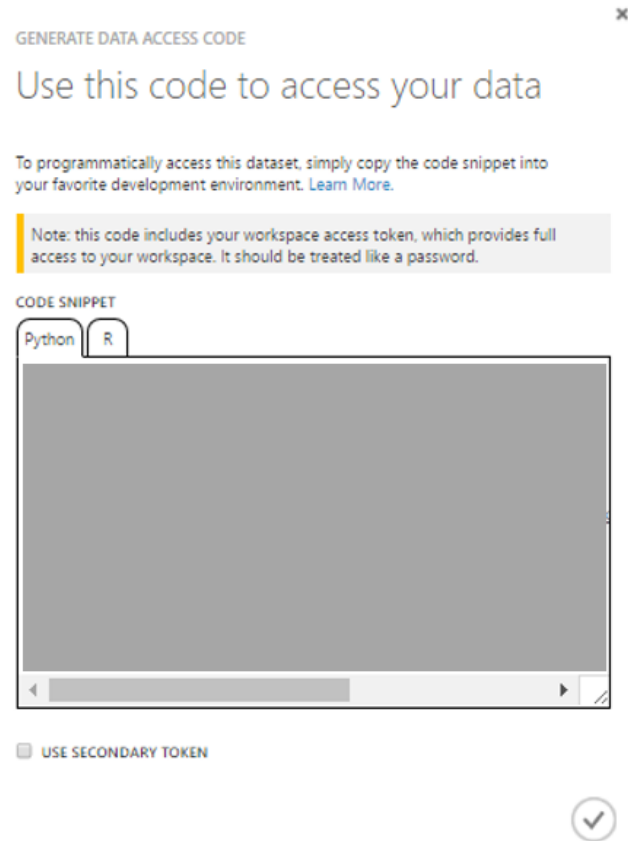
Step 5.3 : Run the experiment

Click "Run"

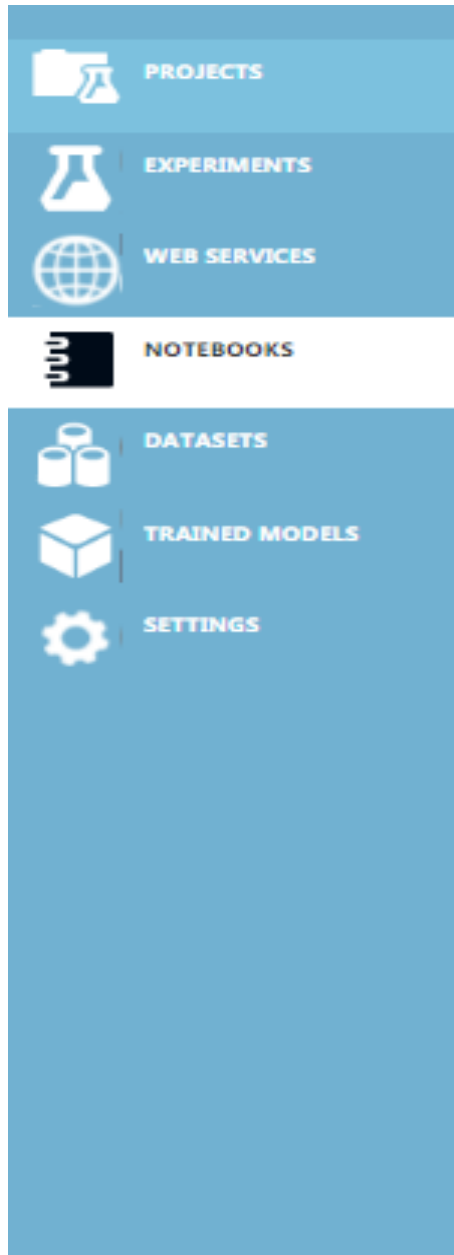


- After Azure Machine Learning "runs" all the steps in your model, you will see a green check mark in the "Convert to CSV" box
- Next right click on the bottom-middle circle in "Convert to CSV" and select "Generate Data Access Code"

Step 5.4 : Generate access code



- You will find data access code one each for Python and R which you will need to import data into your Python Notebooks
- Copy it to your clipboard (Ctrl-C)



Step 5.5 : Open Python or R Notebook in Azure ML

- AML modelling ... a checklist approach
 - ☒ Create new experiment
 - ☒ Import data set
 - ☒ Convert to CSV
 - ☐ Open notebook in AML

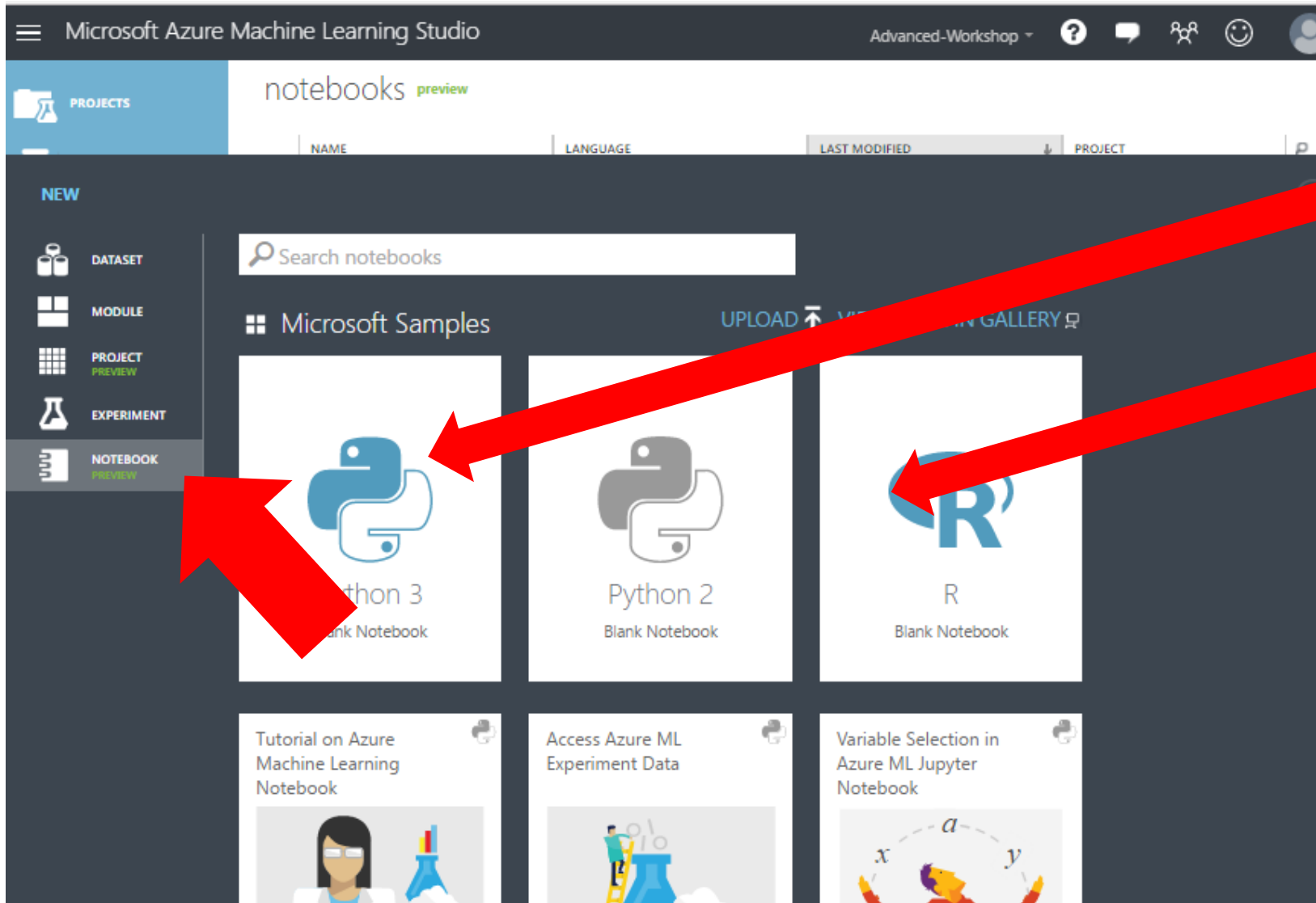
Step 5.6 : Create a new Python or R Notebook

The screenshot shows the Microsoft Azure Machine Learning Studio interface. The top navigation bar includes the title 'Microsoft Azure Machine Learning Studio', a dropdown menu 'Advanced-Workshop', and several icons for help, chat, collaboration, feedback, and user profile. The left sidebar contains navigation links for PROJECTS, EXPERIMENTS, WEB SERVICES, NOTEBOOKS (highlighted), DATASETS, TRAINED MODELS, and SETTINGS. The main content area is titled 'notebooks preview' and displays a table of existing notebooks.

	NAME	LANGUAGE	LAST MODIFIED	PROJECT
<input type="checkbox"/>	Tank Level Forecasting EDA R v2	Python 2	9/28/2016 9:54:44 AM	None
<input type="checkbox"/>	Tank Level Forecasting EDA R	R	9/28/2016 9:27:05 AM	None
<input type="checkbox"/>	Dynamo PCA Py	Python 2	9/22/2016 2:26:39 PM	None
<input type="checkbox"/>	Dynamo PCA R-Copy1	Python 2	9/13/2016 2:03:31 PM	None
<input type="checkbox"/>	Dynamo PCA R	Python 2	9/9/2016 11:05:02 AM	None

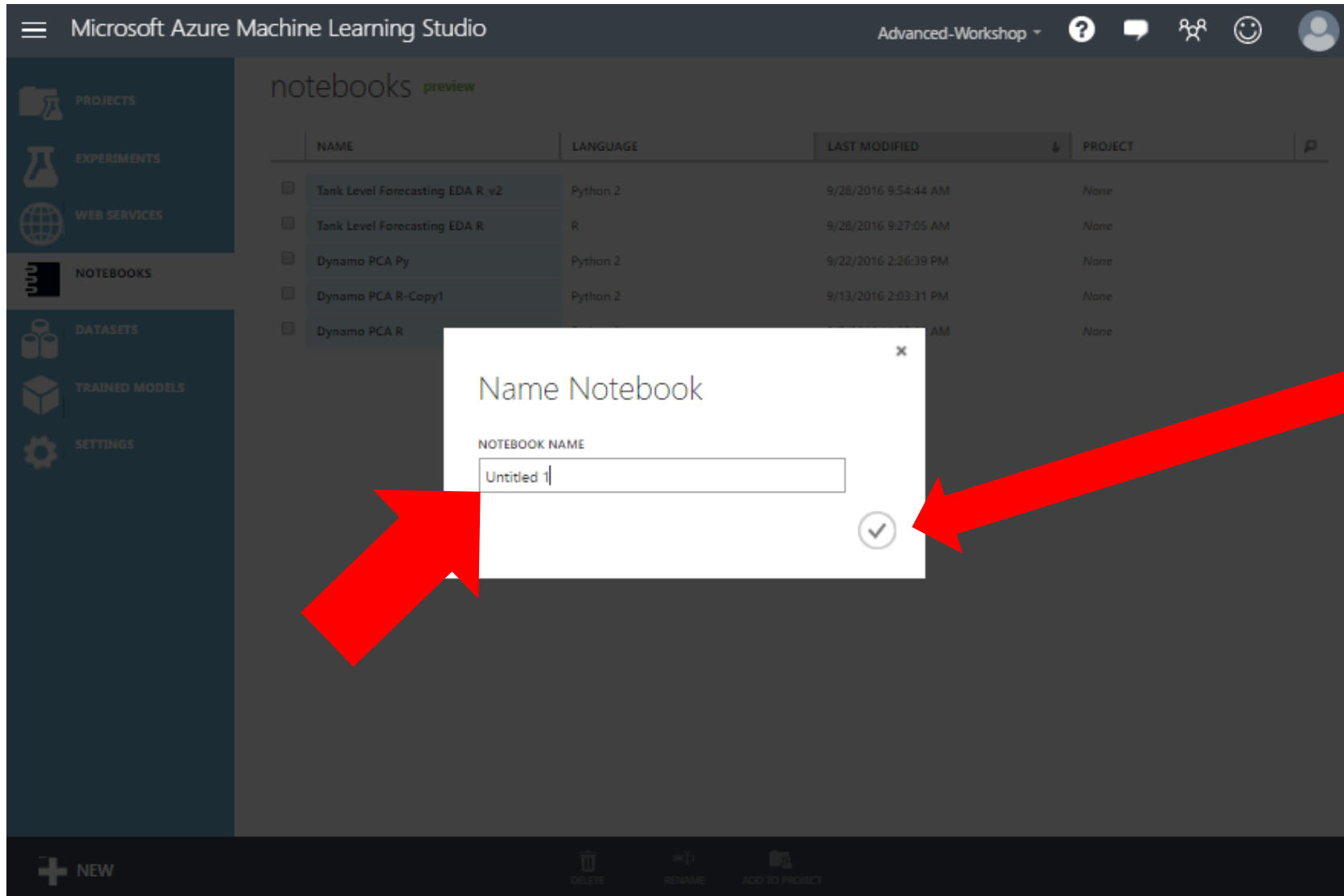
At the bottom of the interface, there is a dark grey bar with several buttons: '+ NEW' (highlighted with a red arrow), 'DELETE', 'RENAME', and 'ADD TO PROJECT'.

Step 5.6 : Create a new Python or R Notebook



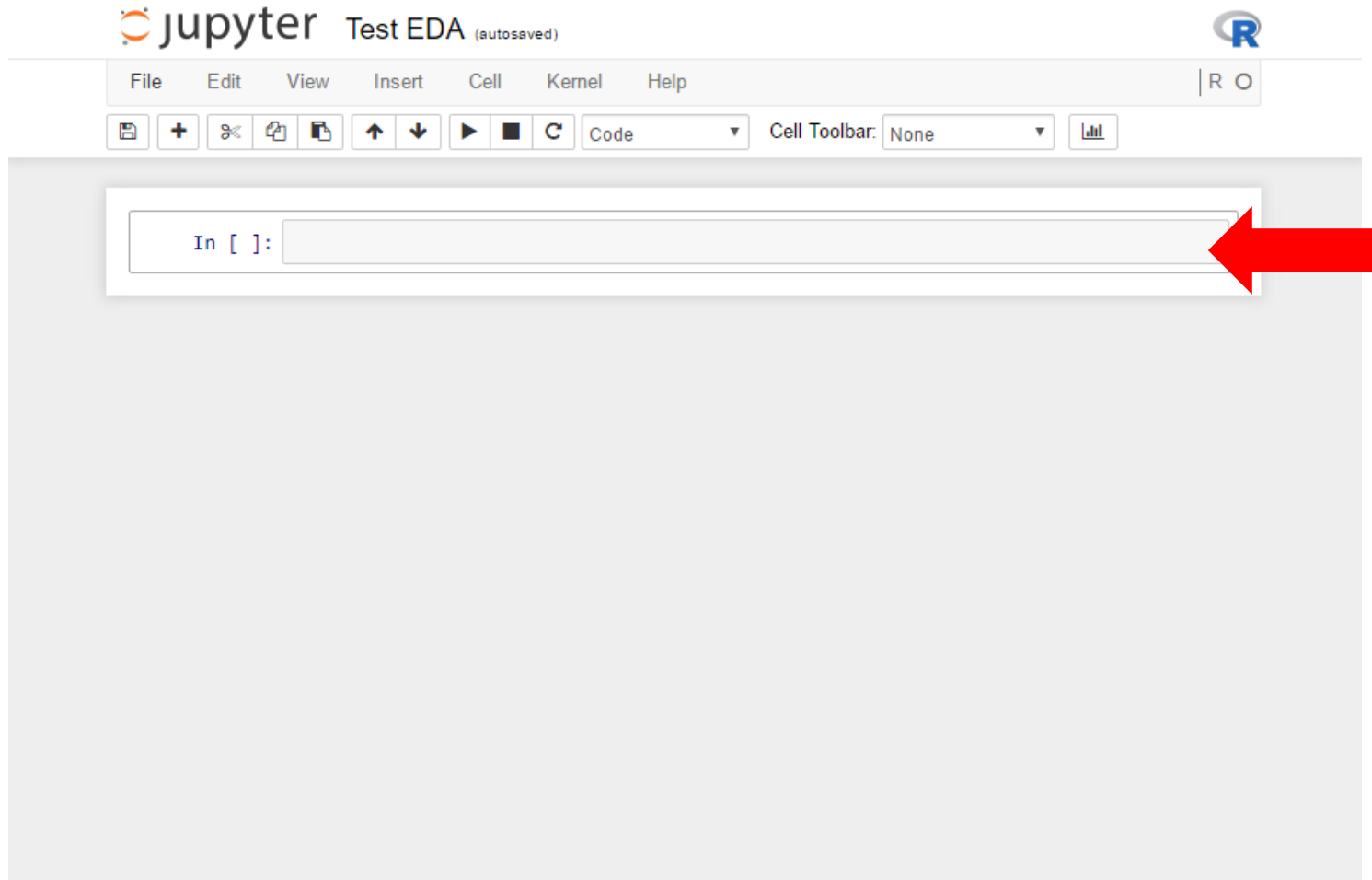
- Chose an R or Python Notebook based on your choice

Step 5.7 : Import this tutorial's training dataset



- Give a name to your notebook
- Click on check mark

Step 5.8 : Import this tutorial's training dataset



- Paste in the data access code you had copied earlier onto the clipboard depending on whether it is a Python or R Notebook