Instructor: Mithun Prasad, PhD
Email: miprasad@Microsoft.com

Microsoft

WHAT YOU WILL BE ABLE TO DO AFTER THIS TRAINING

Build a Data Science experiment using ML studio.

Gain familiarity with Data Science components of the studio.

Customize Data Science components in the studio.

Microsoft

MACHINE LEARNING 101

The first way of thinking about ML is by the type of information or input given to a system.

1. **Supervised learning** – we get the data and the labels e.g. linear regression
2. **Unsupervised learning** – only get the data (no labels) e.g. clustering
3. **Reinforcement learning** – reward/penalty based information (feedback)

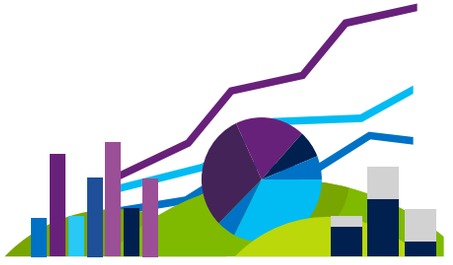
Another way of categorizing ML approaches, is to the desired output:

1. **Classification** (e.g. decision tree)
2. **Regression** (e.g. linear regression)
3. **Clustering** (e.g. k-means)
4. **Density estimation** (e.g. histograms)
5. **Dimensionality reduction** (e.g. principal component analysis

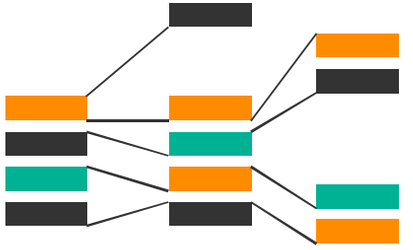
MACHINE LEARNING CAPABILITIES

Which category
(Classification)

How much/many
(Regression)

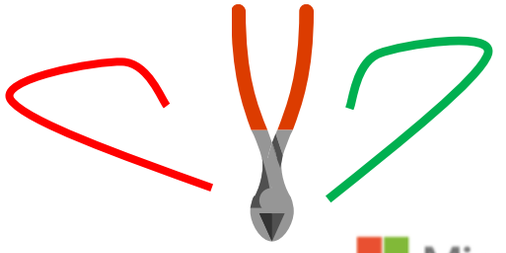


Which group
(Clustering, Recommender)



Is it odd
(Anomaly)

Which action
(Reinforcement Learning)



Microsoft

MACHINE LEARNING VISION

Make Machine Learning accessible to every enterprise, data scientist, developer, information worker, consumer, and device anywhere in the world

No need to be an expert

Microsoft

MACHINE LEARNING TERMINOLOGY

Term	Definition
Training set	set of data used to learn a model
Test set	set of data used to test a model
Feature	a variable (continuous, discrete, categorical, etc.) aka column
Target	Label (associated with dependent variable, what we predict)
Learner	Model or algorithm
Fit, Train	Learn a model with an ML algorithm using a training set
Predict	w/ supervised learning, give a label to an unknown datum(data). w/unsupervised decide if new data is weird, in which group, or what to do next with the new data
Accuracy	percentage of correct predictions ((TP + TN)/ total)
Precision	Percentage of correct positive predictions (TP/ (FP + TP))
Recall	Percentage of positive cases caught (TP/ (FN + TP))

CORTANA INTELIGENCE IN A SENTENCE

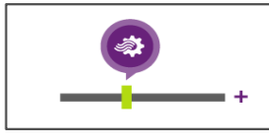
Cortana Intelligence is a Platform and a Process to perform advanced analytics from start to finish

Microsoft

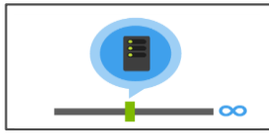
THE CORTANA INTELLIGENCE PLATFORM

Cortana, Cognitive Services, Bot Framework

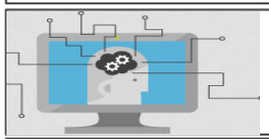
Power BI



Azure Stream Analytics



Azure HDInsight



Azure Machine Learning and MRS

Azure SQL DB, Data Warehouse, DocumentDB

Azure Data Lake

Azure Event Hubs

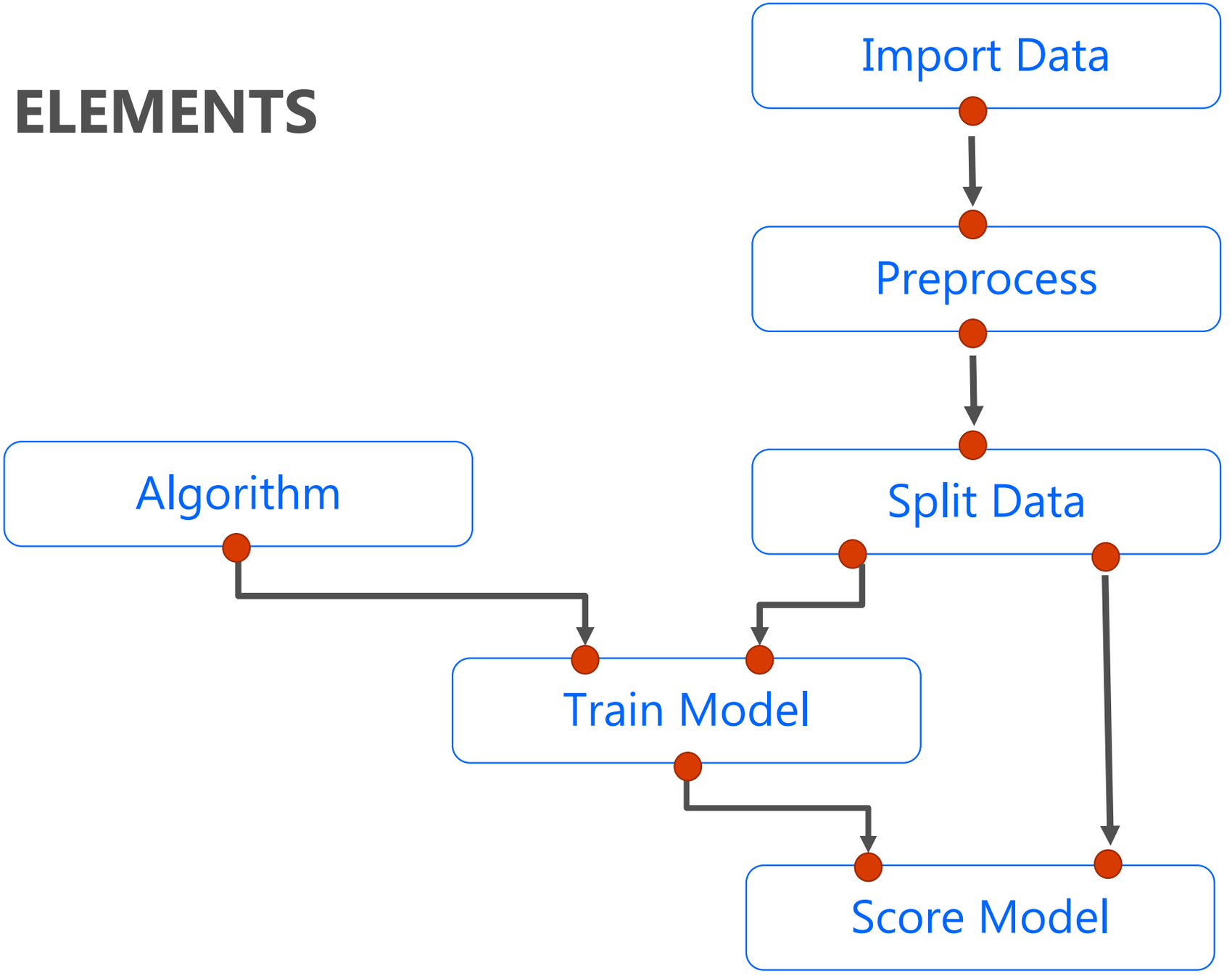
Azure Data Factory

Azure Data Catalog

Microsoft Azure

Microsoft

AZURE ML ELEMENTS



KEY AZURE ML CONCEPTS

- **Workspace**

  Roughly equivalent to an account. Cloud-based container for data, experiments, notebooks. Units for sharing.

- **Studio**

  Web-based GUI for authoring experiments in a workspace.

- **Experiment**

  A job, containing a data flow graph (DAG)

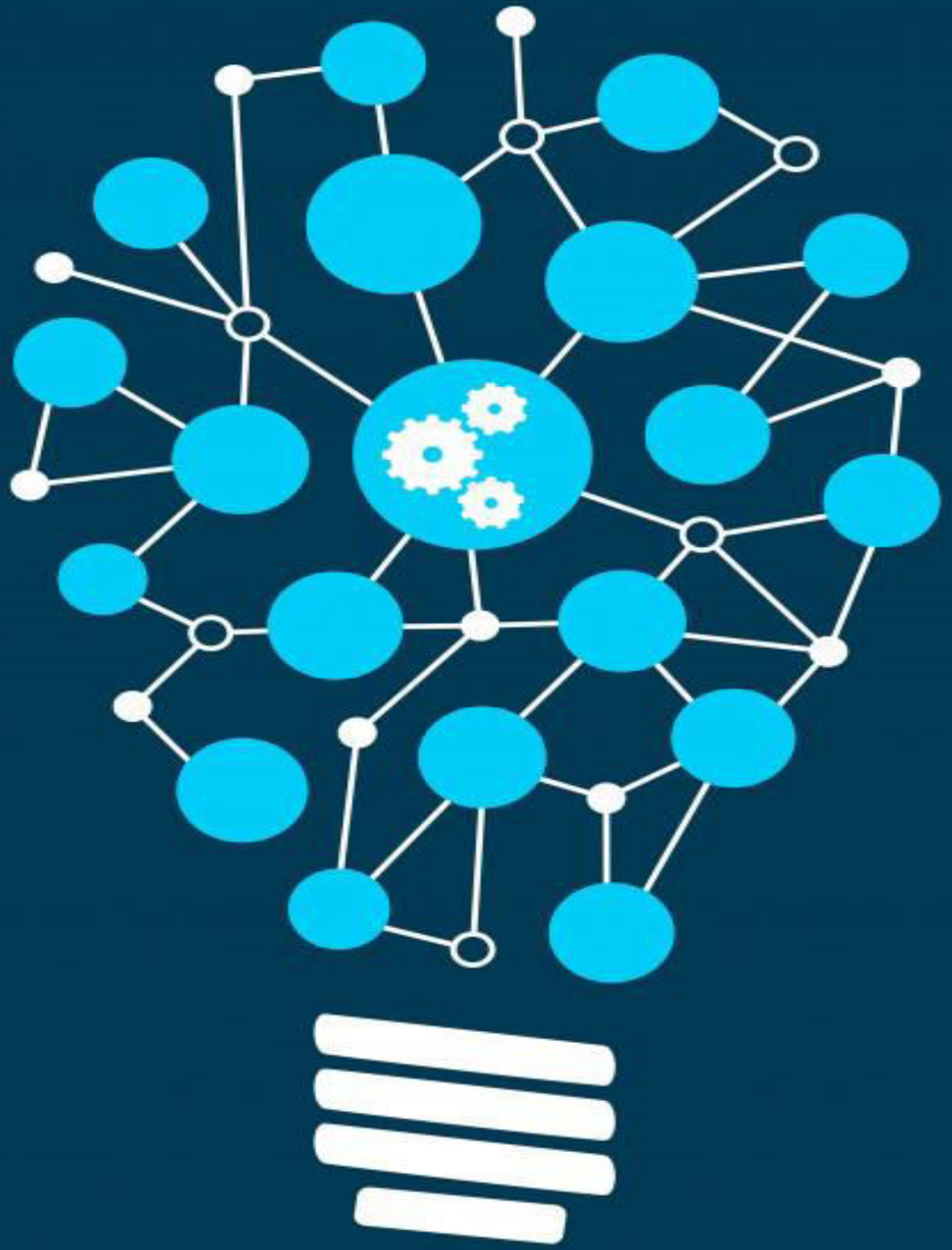
- **DataSet**

  Rectangular data structure, similar to database table.

- **Module**

  Units of work within a graph

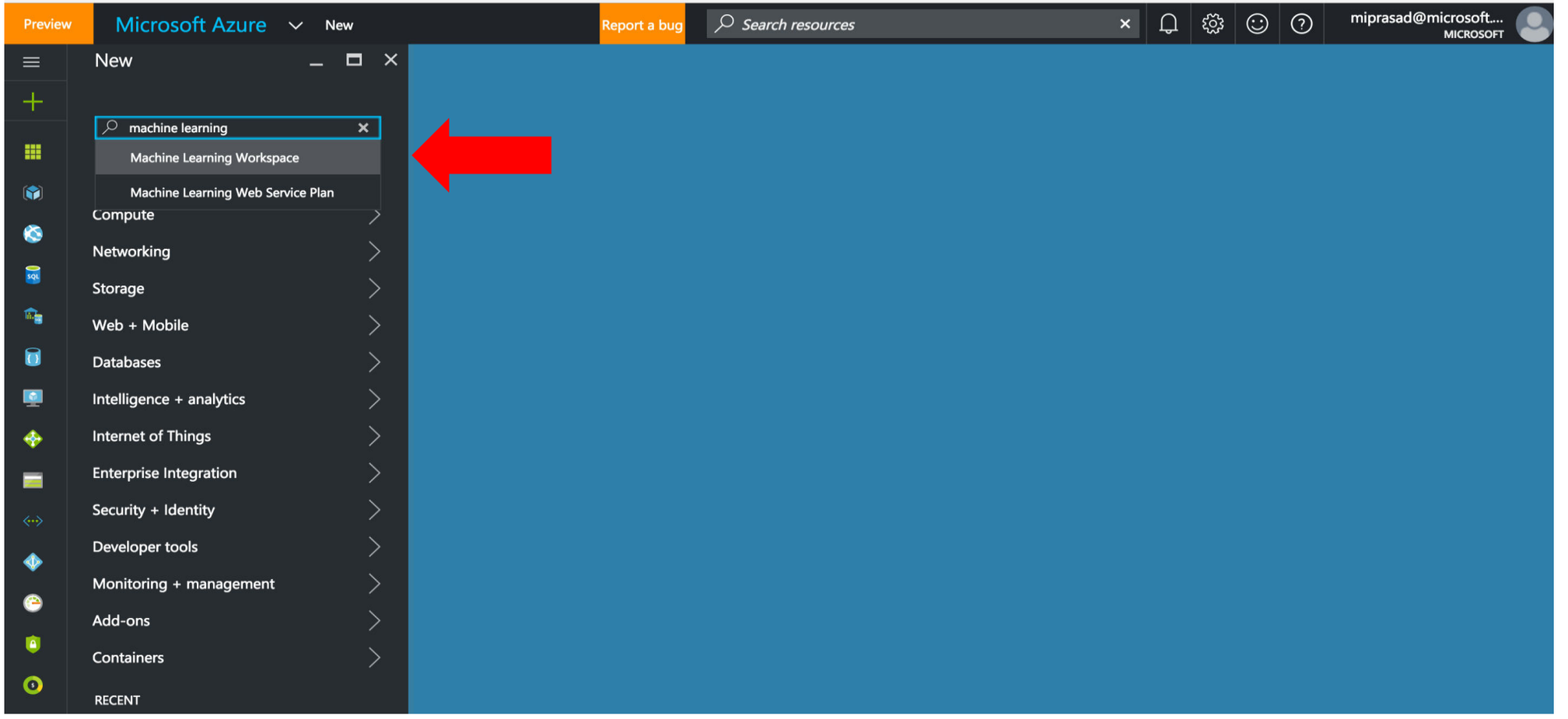
AZURE ML STUDIO AND THE TEAM DATA SCIENCE PROCESS



Microsoft

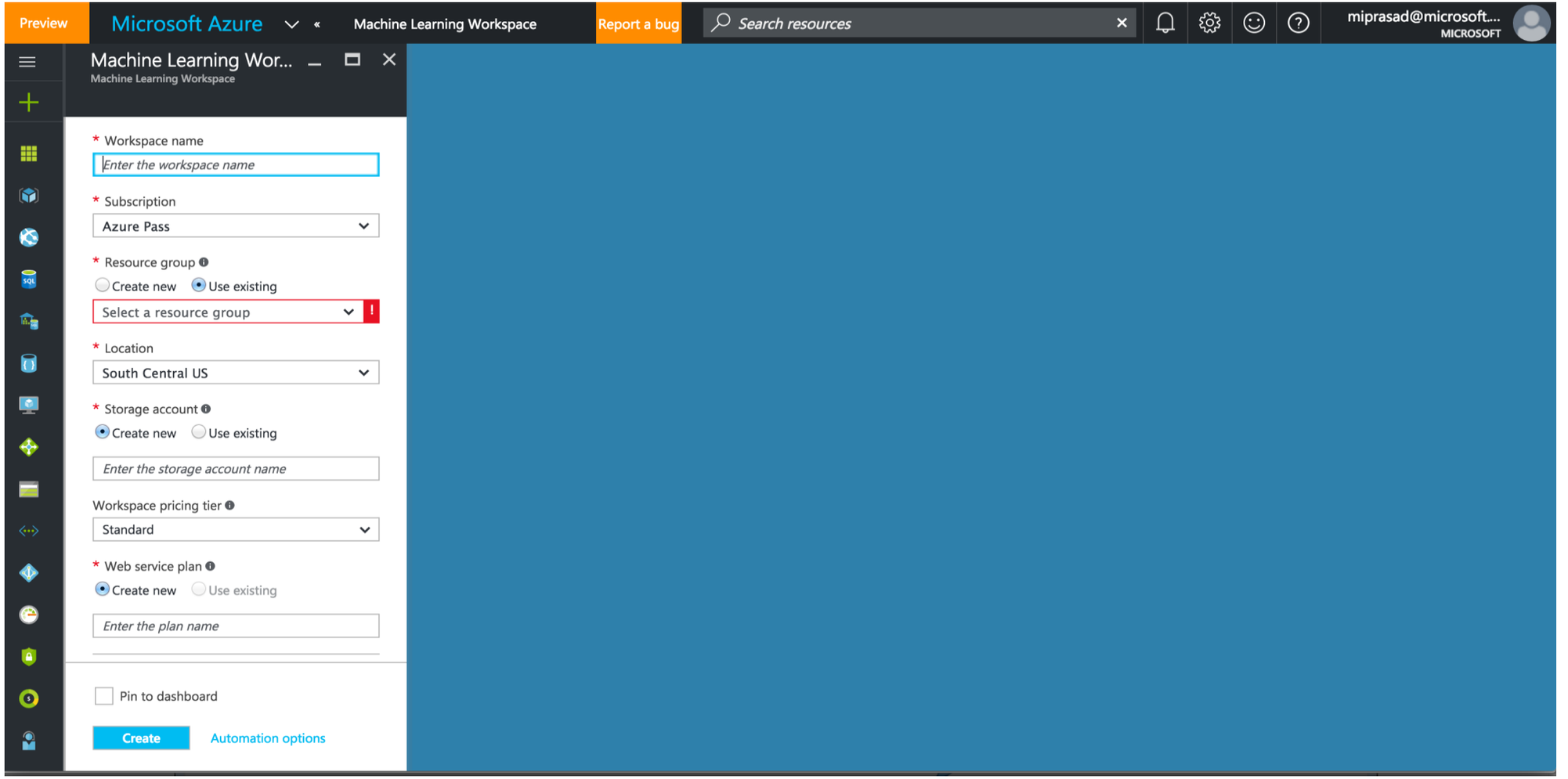
WORKSPACE

portal.azure.com



Microsoft

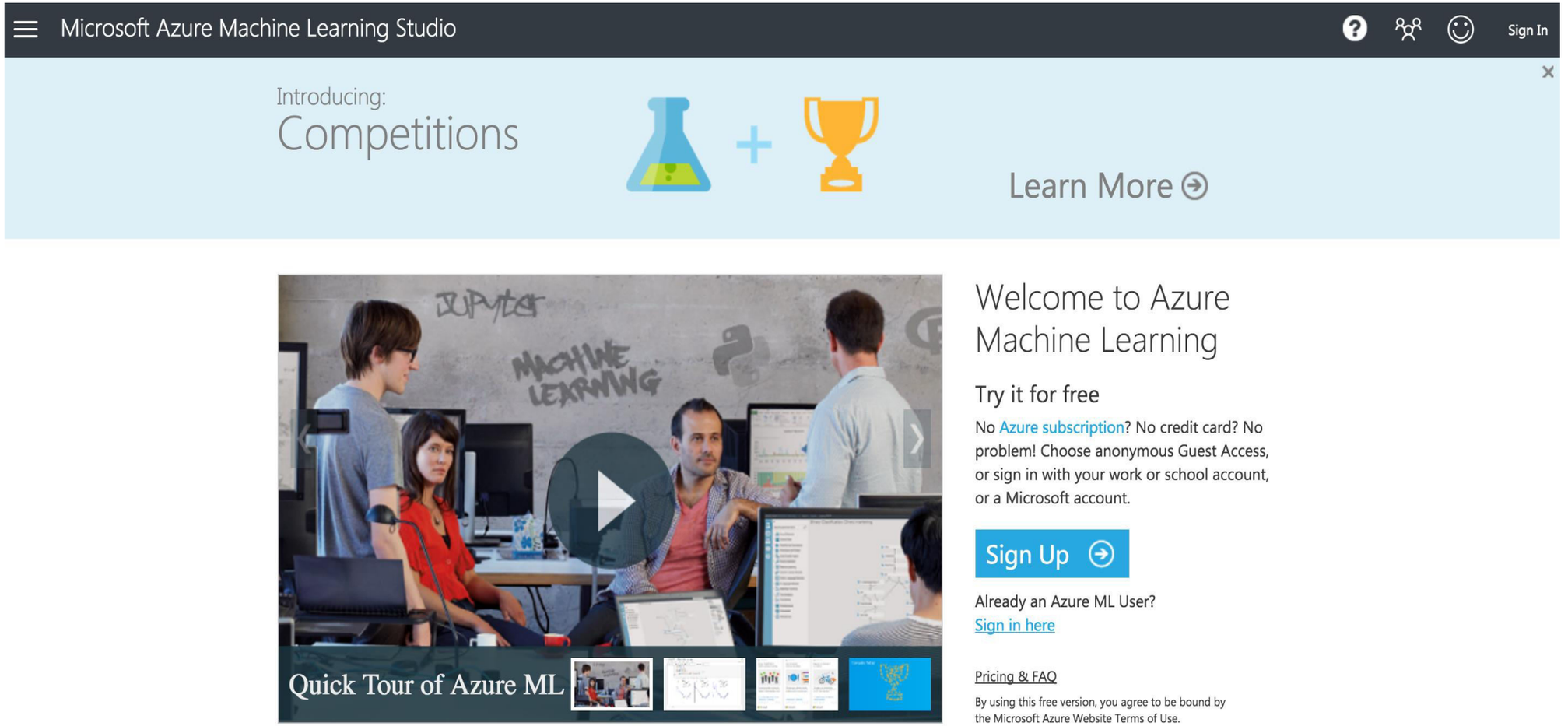
WORKSPACE



Microsoft

AZURE ML STUDIO - SIGN UP

https://studio.azureml.net/#



Microsoft

TIERS

Region: South Central US

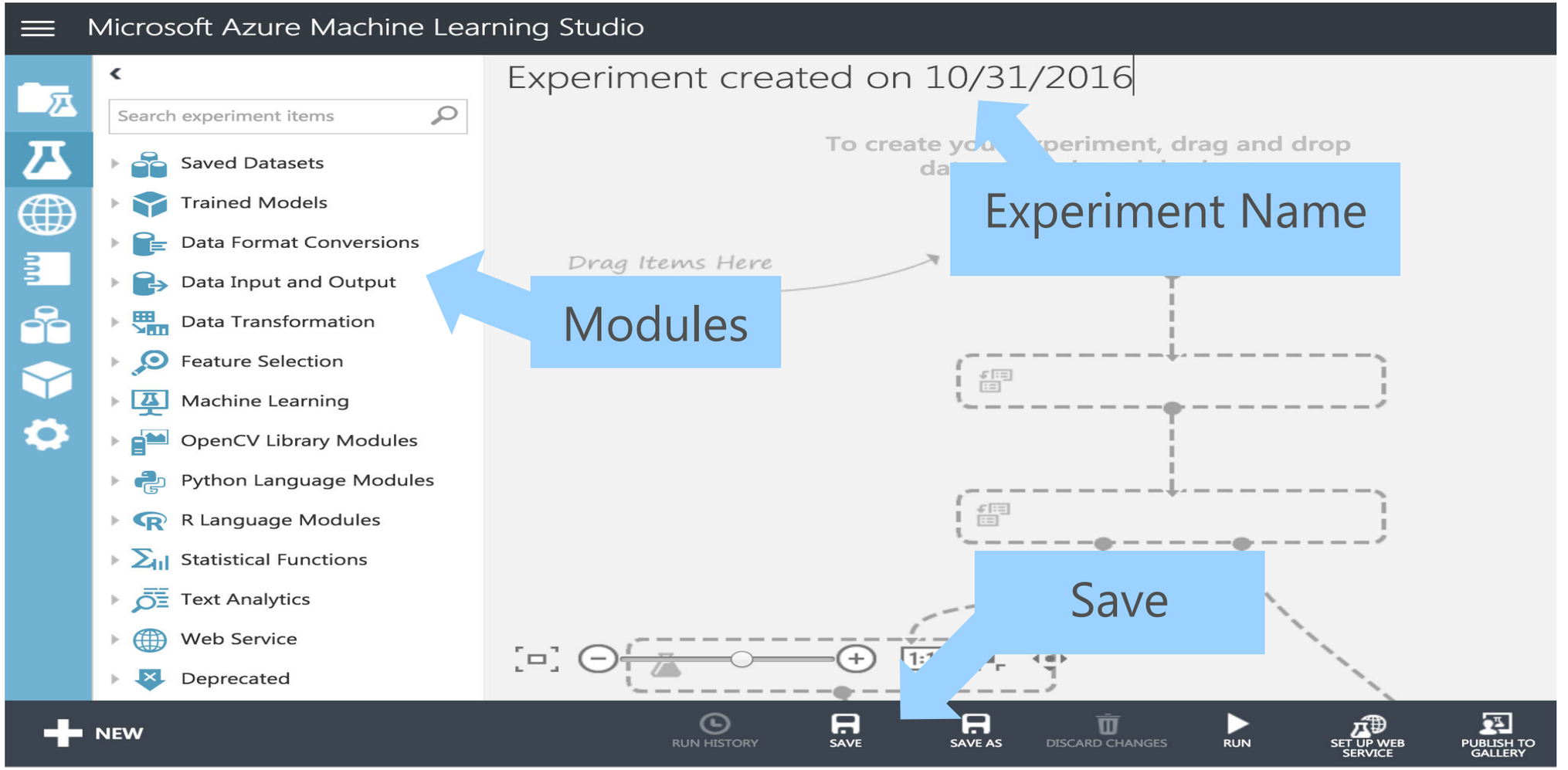
Currency: Indian Rupee (₹)

	FREE	STANDARD
Price	Free	₹600.27 per Seat per month ₹60.09 per Studio Experimentation Hour
Azure Subscription	Not Required	Required
Max Number of Modules per Experiment	100	Unlimited
Max Experiment Duration	1 hour per experiment	Up to 7 days per experiment with a maximum of 24 hours per module
Max Storage Space	10 GB	Unlimited - BYO
Read Data from On-Premises SQL Preview	No	Yes
Execution / Performance	Single Node	Multiple Nodes
Production Web API	No	Yes
SLA	No	Yes

Hourly charges only apply to active use of the service. Where multiple meters are present they are applied concurrently.

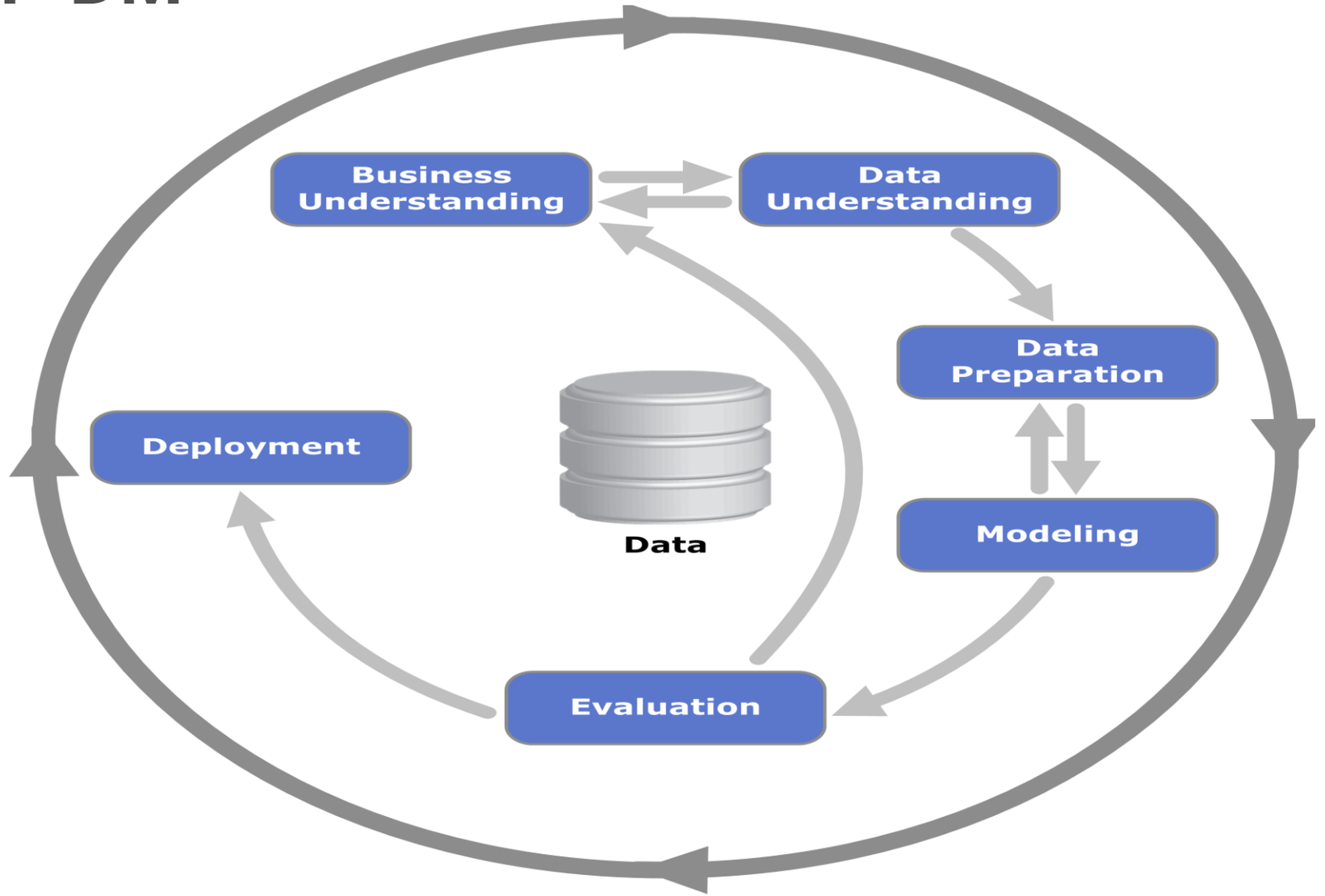
Microsoft

AZURE ML STUDIO - NAVIGATION

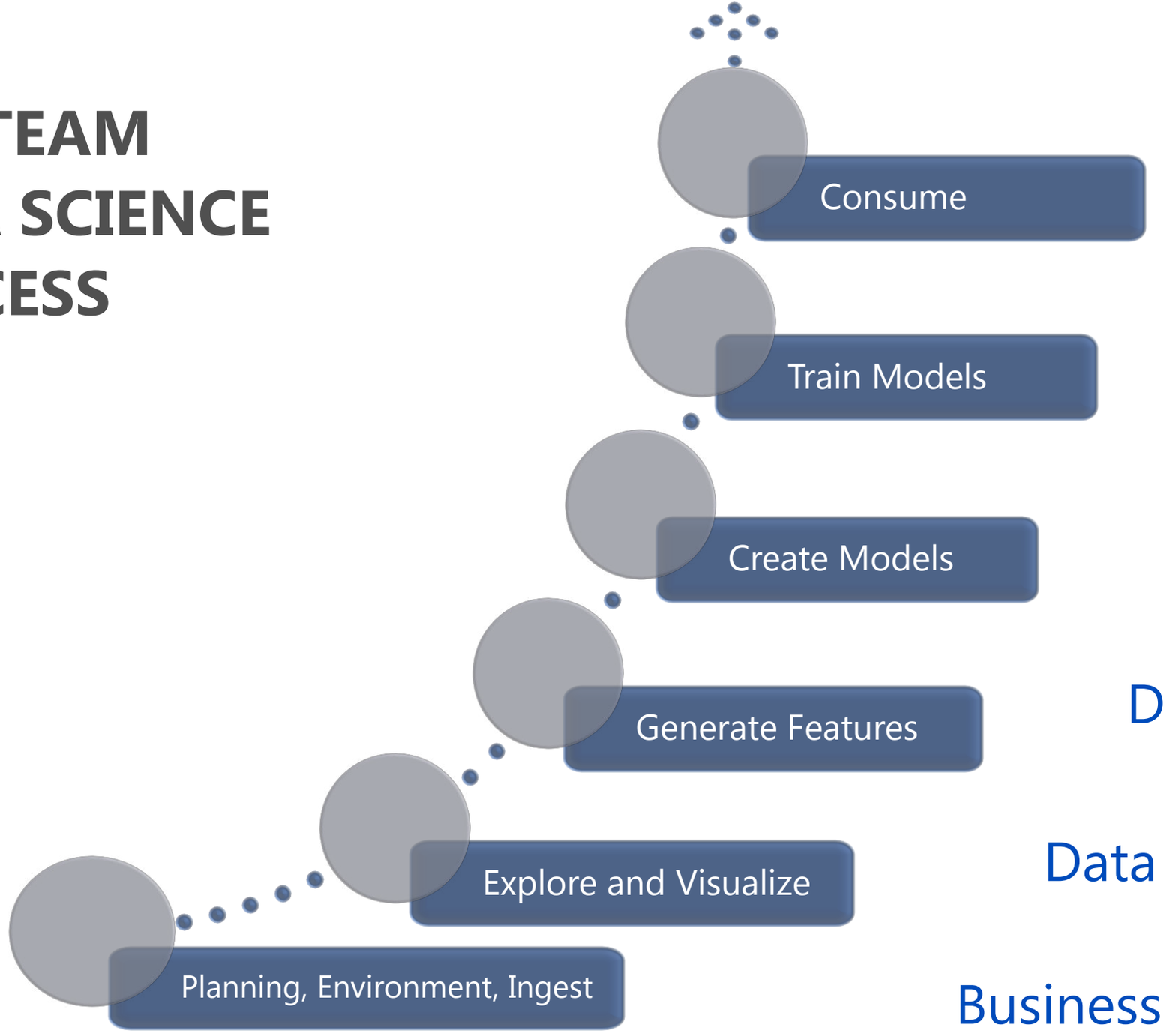


Microsoft

CRISP-DM



THE TEAM DATA SCIENCE PROCESS



Deploying

Evaluating

Modeling

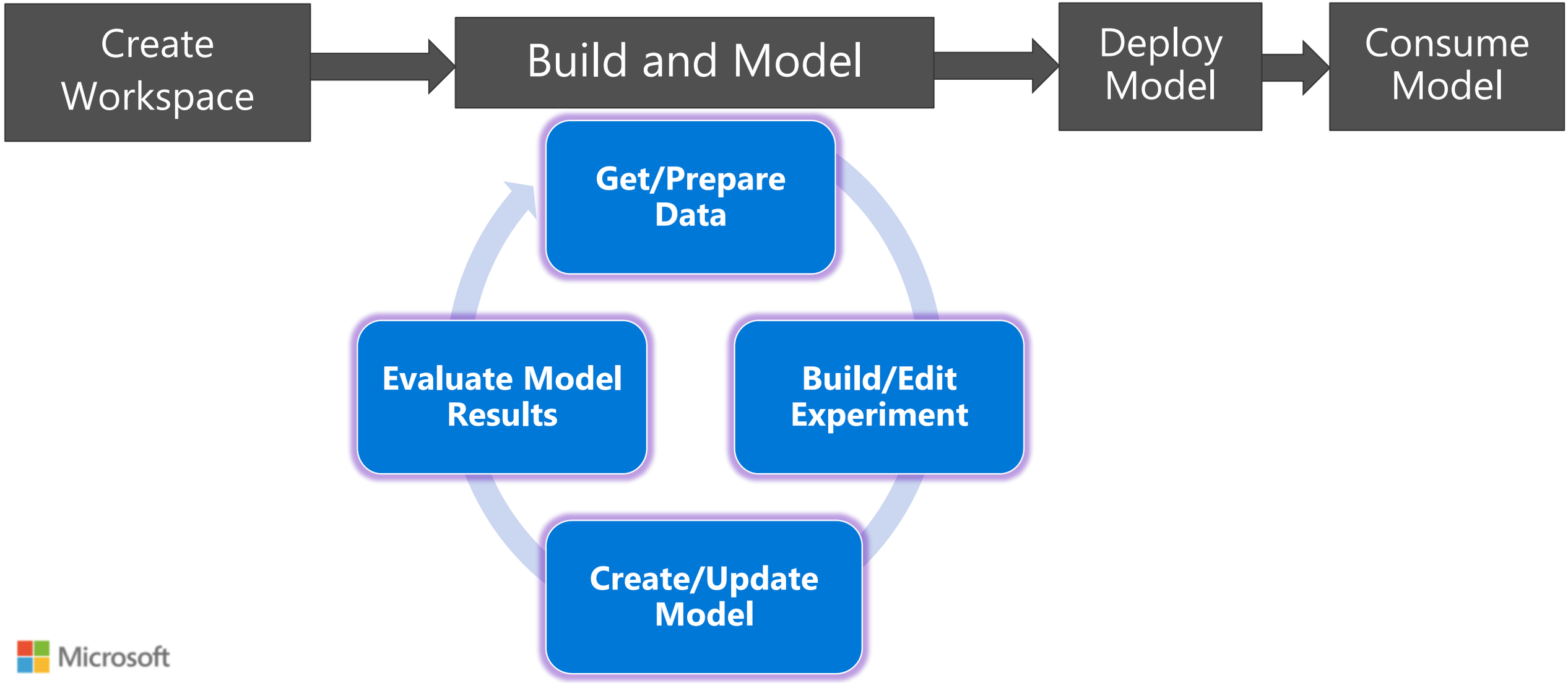
Data Preparation

Data Understanding

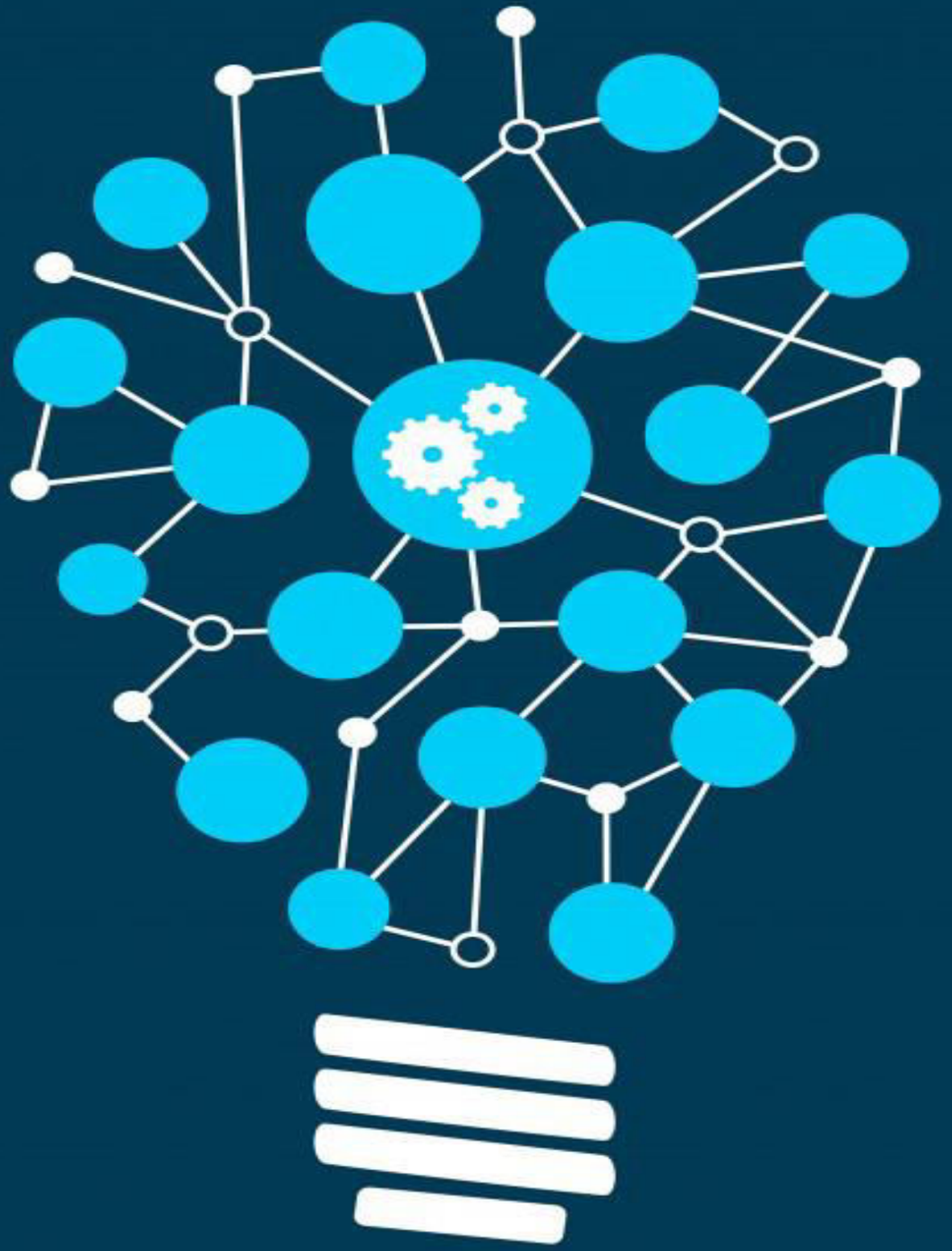
Business Understanding

Microsoft

CREATING AN EXPERIMENT

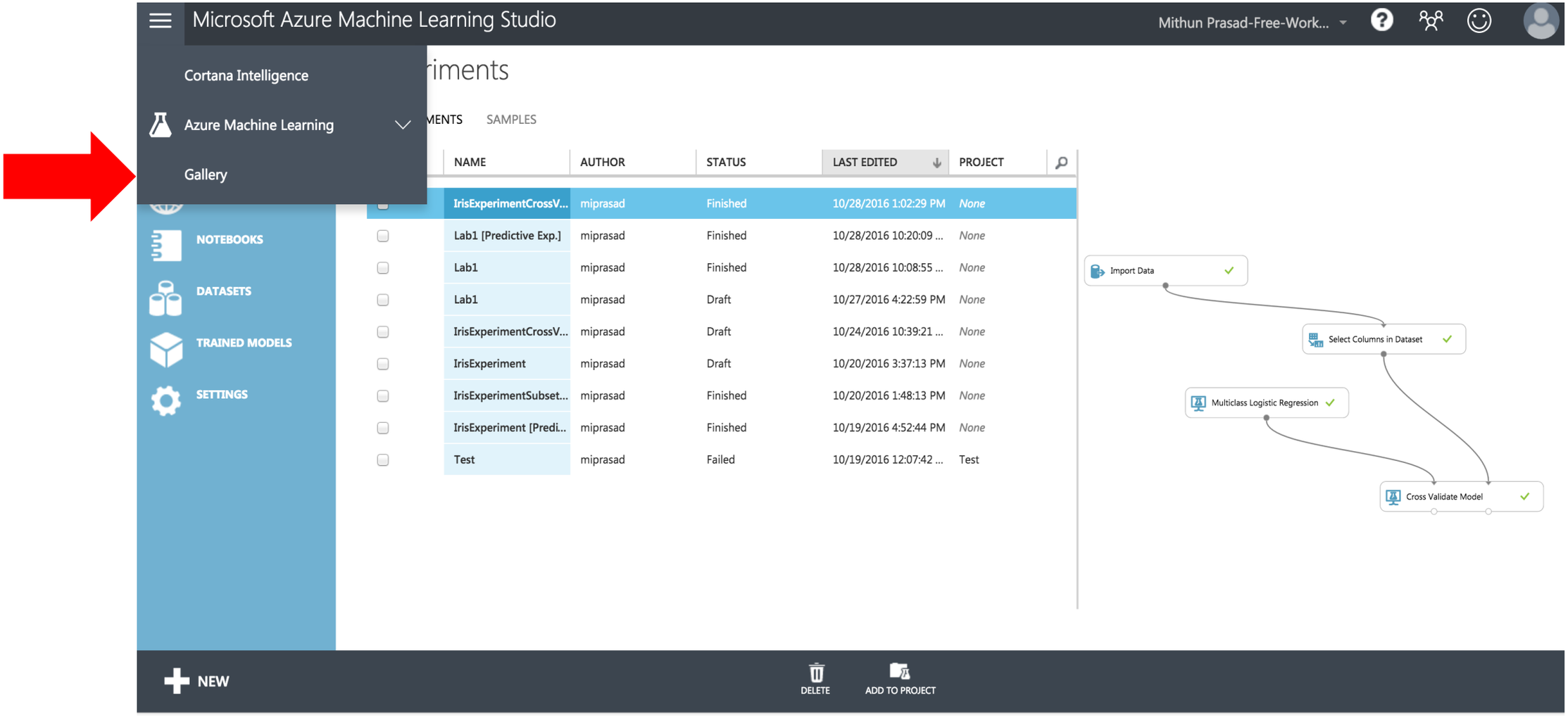


LEARNING RESOURCES



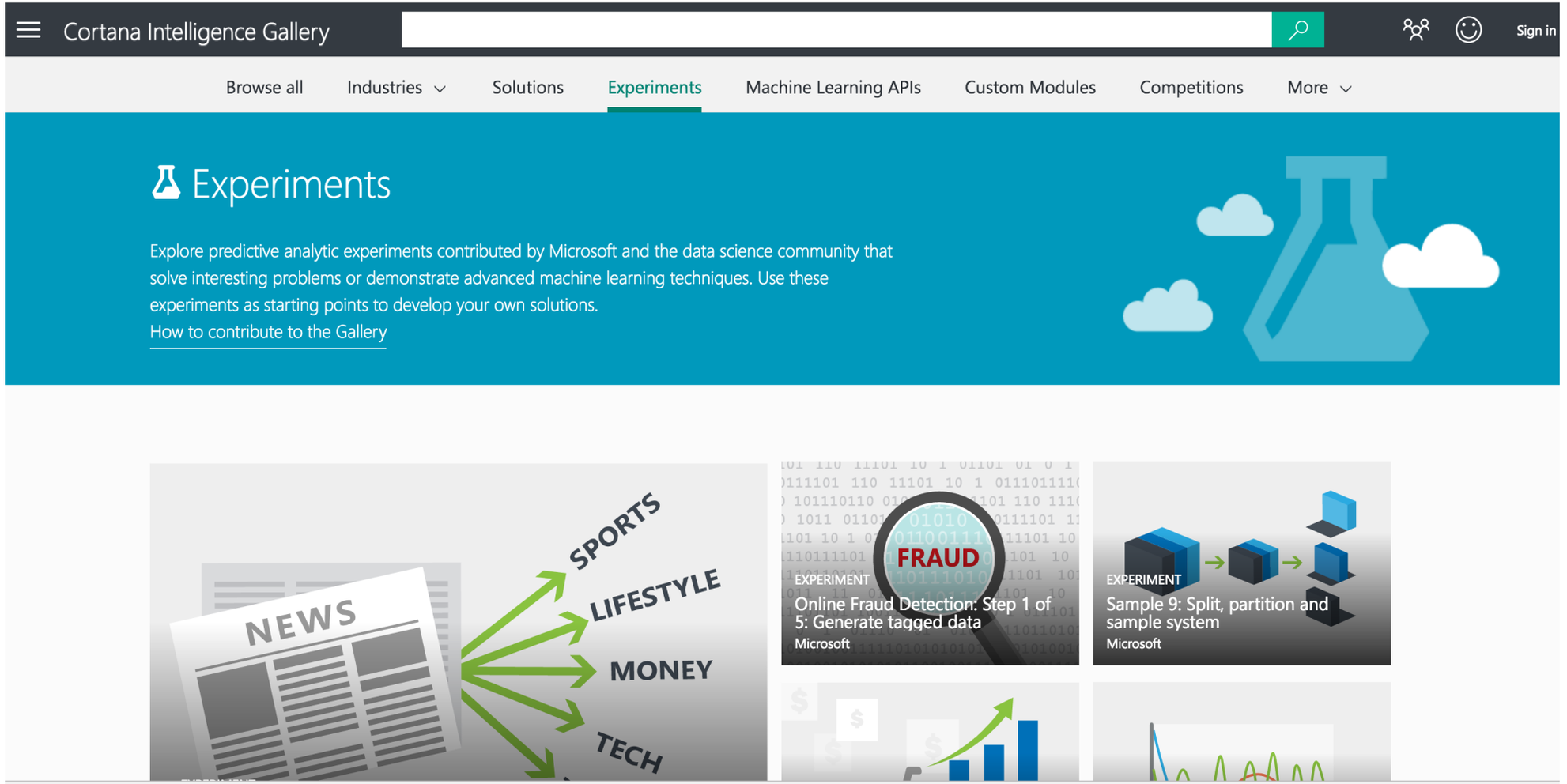
Microsoft

GALLERY



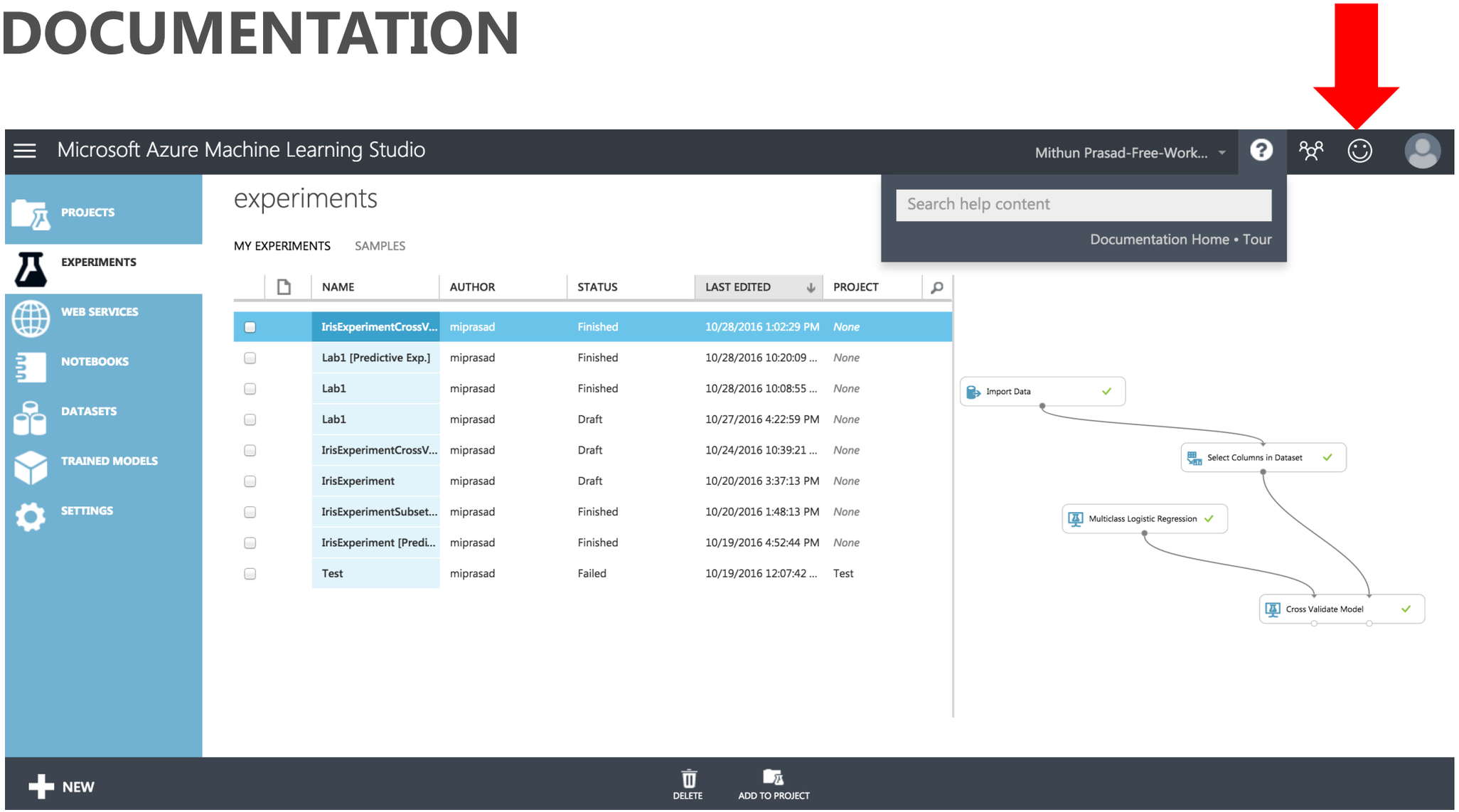
Microsoft

GALLERY



Microsoft

DOCUMENTATION



Microsoft

DOCUMENTATION

Microsoft Azure

SALES 000-800-100-3928 | MY ACCOUNT | PORTAL | Search

Why Azure | Solutions | Products | Documentation | Pricing | Partners | Blog | Resources | Support

FREE ACCOUNT

Search for docs

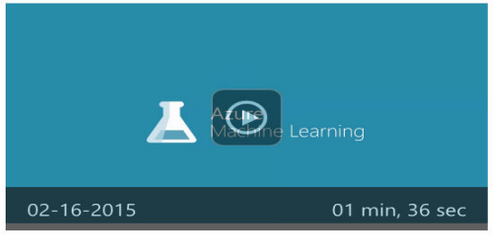
- Overview
  - What's Machine Learning?
  - Data science for beginners
  - What's the Studio?
  - Studio capabilities
  - Frequently asked questions
- Get started
- Data science process
- Algorithms
- Develop a model
- Work with data
- Extend with R & Python

Machine Learning documentation

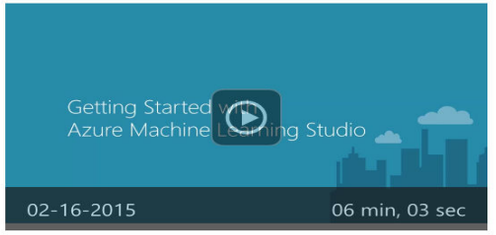
Learn how data scientists and developers can embed predictive analytics into applications

Build your first experiment in Azure Machine Learning Studio Start tutorial >

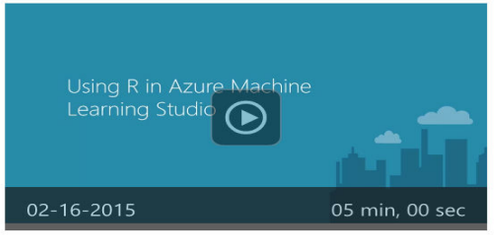
Machine Learning videos



Overview of Azure Machine Learning



Getting Started with Azure Machine Learning Studio



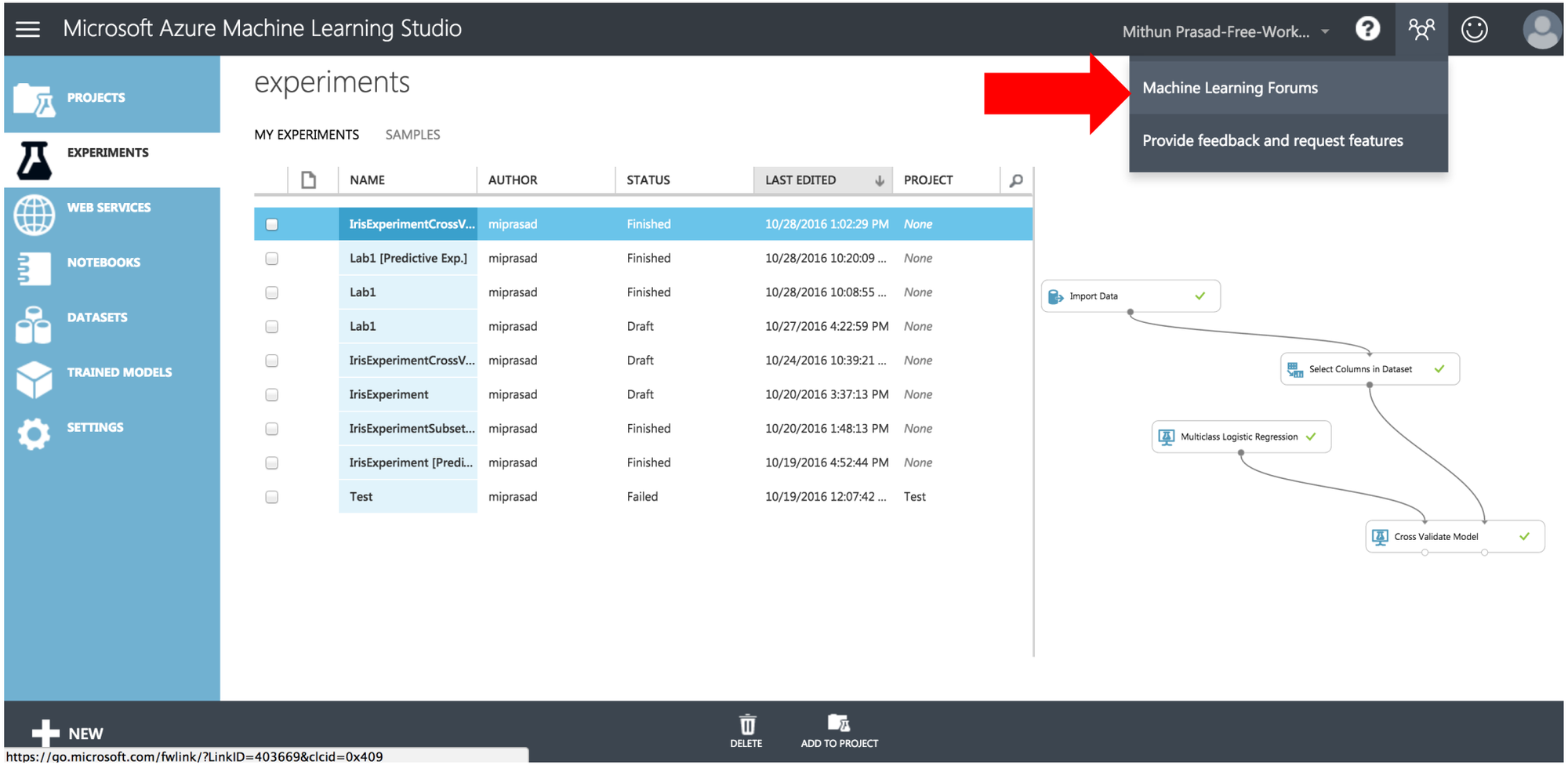
Using R in Azure Machine Learning Studio

See more videos >

Try Machine Learning Studio

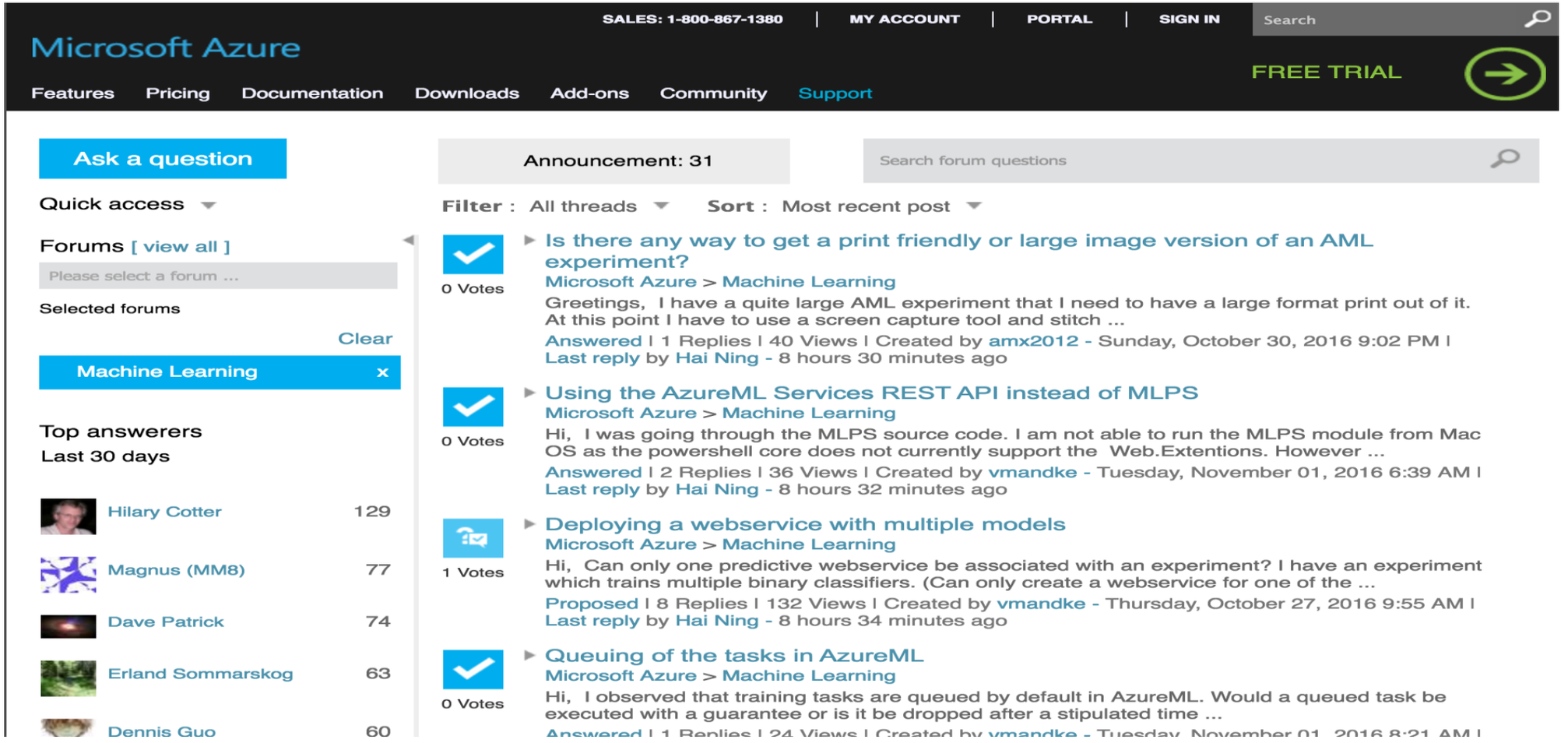
Microsoft

FORUMS



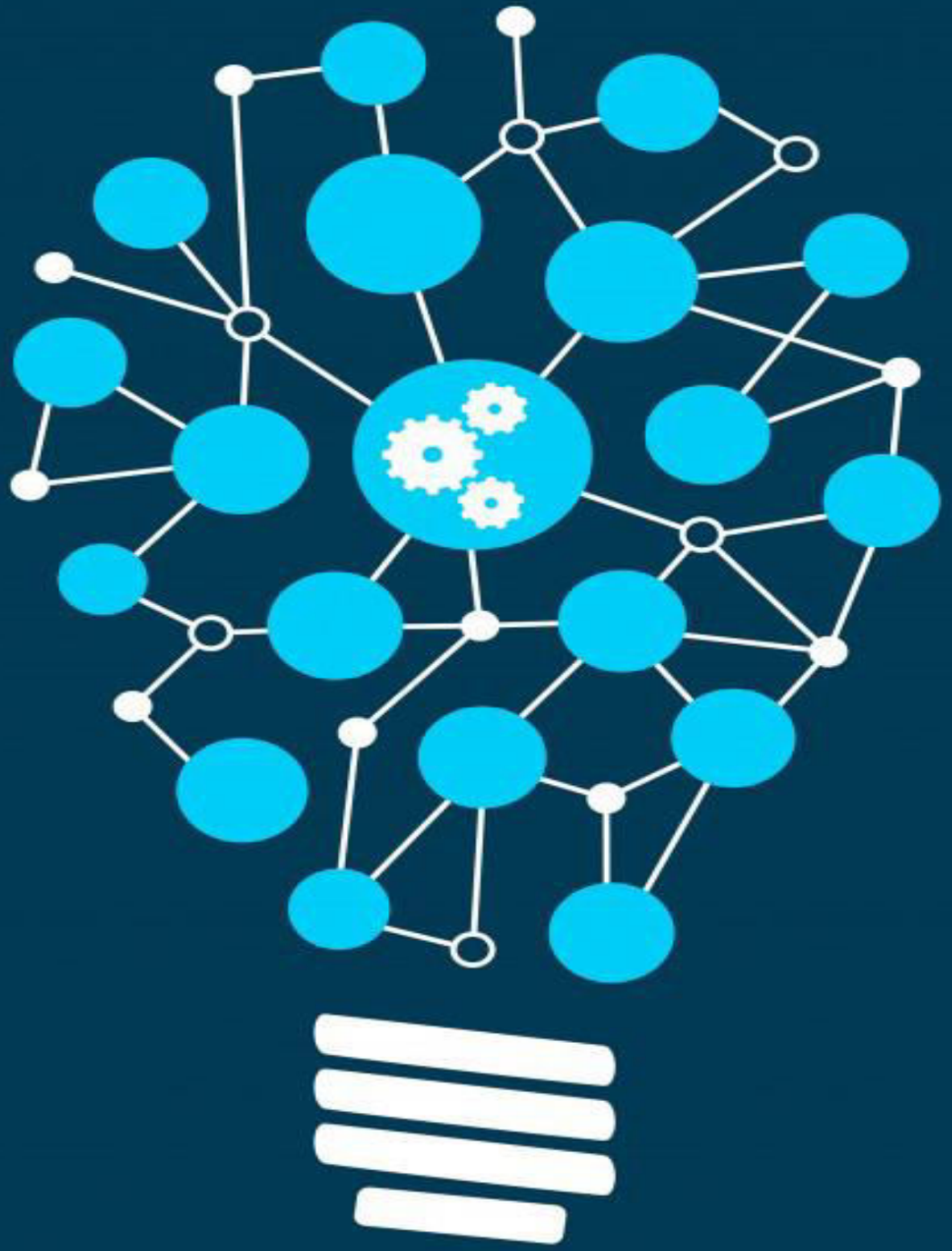
Microsoft

FORUMS



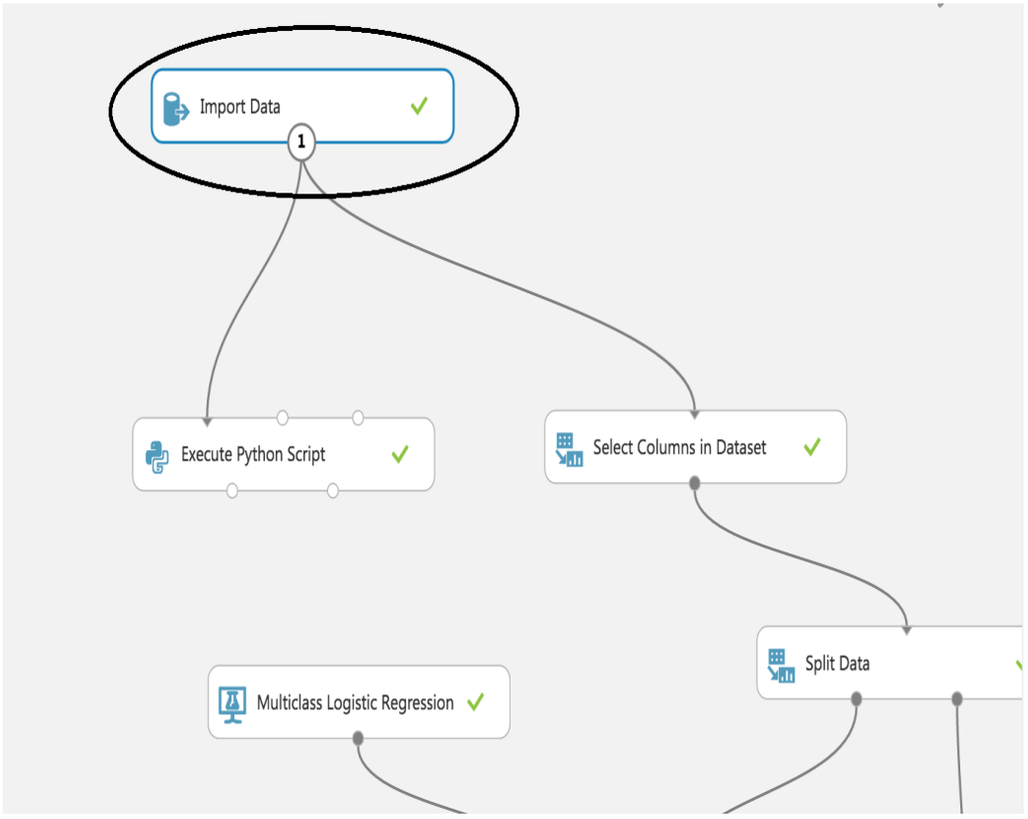
Microsoft

DATA INGESTION AND PREPARATION



Microsoft

DATA ACCESS (IMPORT)



Import Data

Data source

Azure SQL Database

Database server name

irismldb3.database.windows.net

Database name

irisMLDB

User name

miprasad@irismldb3

Password

••••••••••

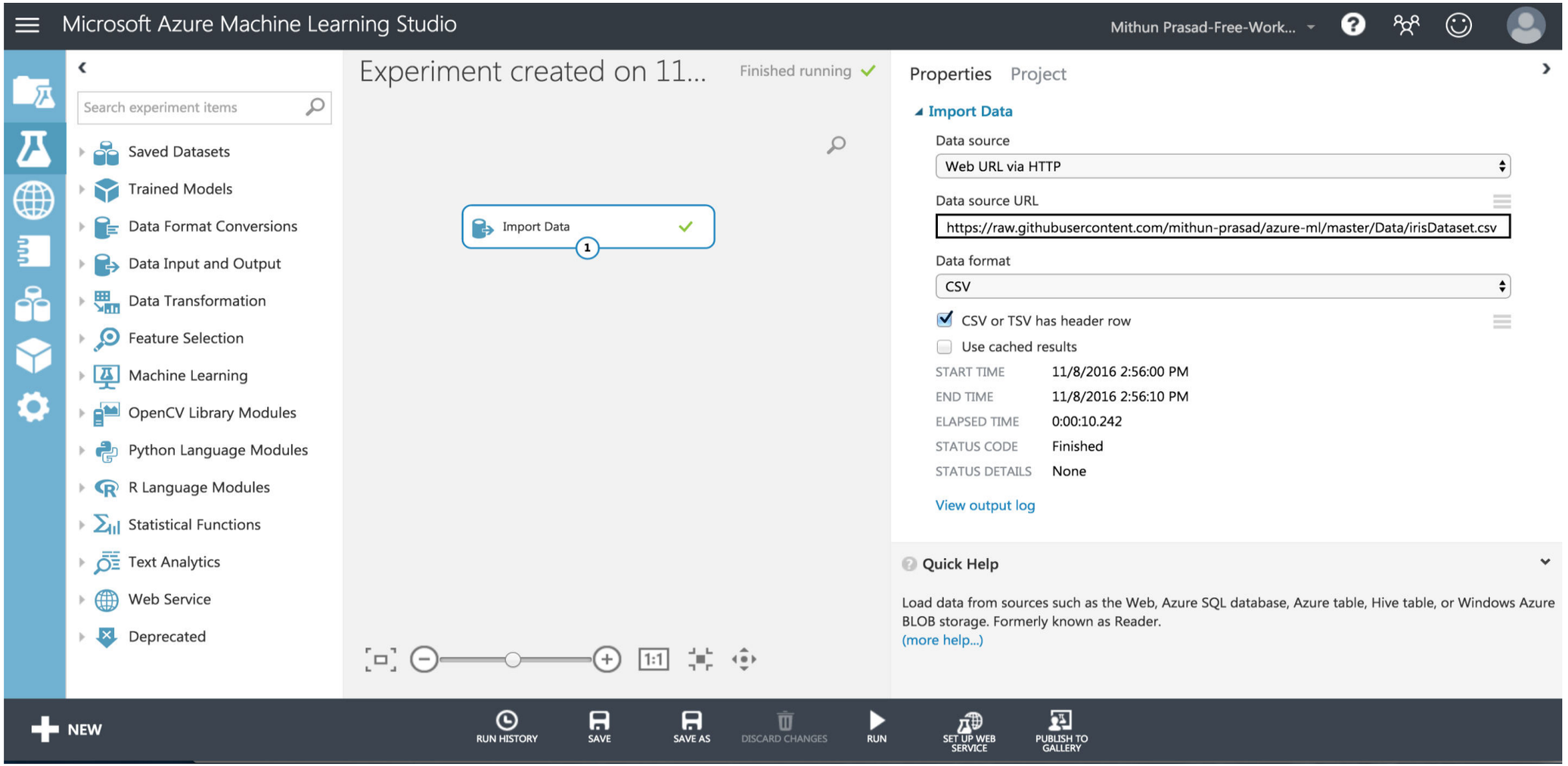
Accept any server certificate (insecure)

Database query

```
1 Select * from iris
```

Microsoft

IMPORT VIA URL



Microsoft

DATA ACCESS (EXPORT)

Data Format Conversion

- Convert to ARFF
- Convert to CSV
- Convert to Dataset
- Convert to SVMLight
- Convert to TSV

Export Data

Please specify data destination

Azure SQL Database

Database server name

irismldb3.database.windows.net

Database name

irisMLDB

Server user account name

miprasad@irismldb3

Server user account password

••••••••

Accept any server certificate (insecure)

Comma separated list of columns to be saved

sepallength, petallength, Scored Labels

Data table name

irisOutput

Comma separated list of datatable columns

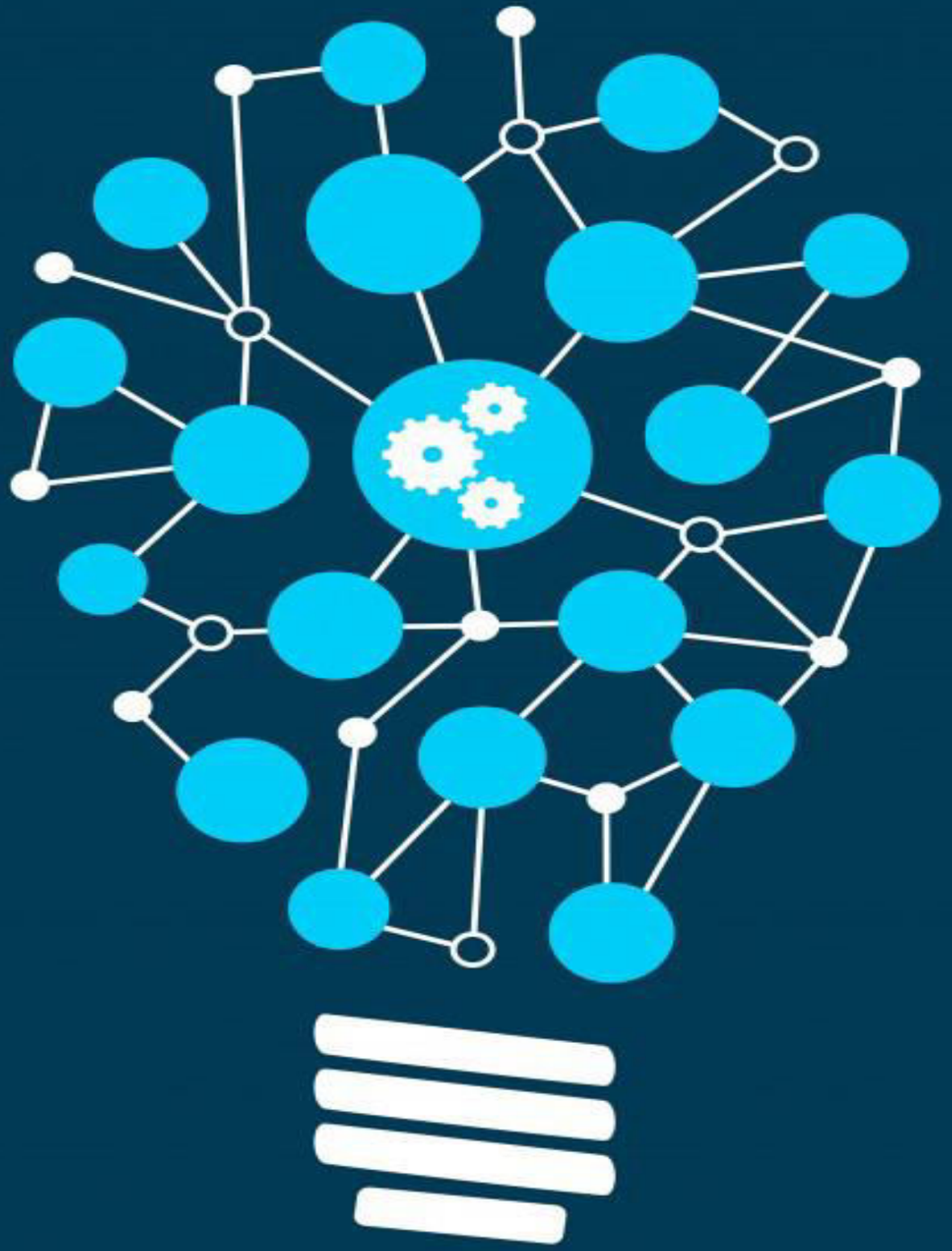
sepallength, petallength, scoredclass

Microsoft

ARFF FILE

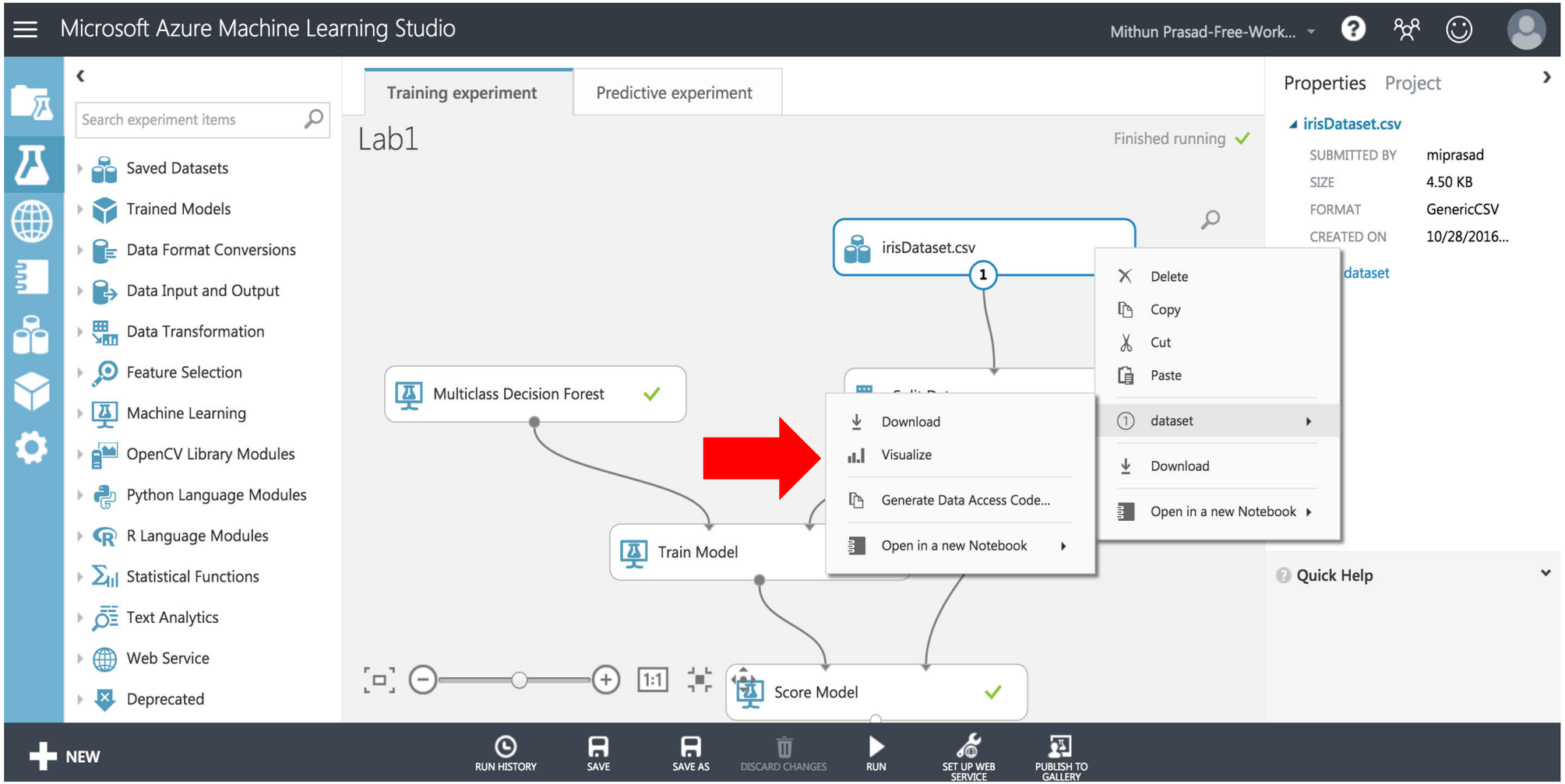
```
% 1. Title: Iris Plants Database
%
% 2. Attribute Information:
%     1. sepal length in cm
%     2. sepal width in cm
%     3. petal length in cm
%     4. petal width in cm
%     5. class:
%         -- Iris Setosa
%         -- Iris Versicolour
%         -- Iris Virginica
|
@RELATION iris

@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class     {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
```

Exploratory Statistics



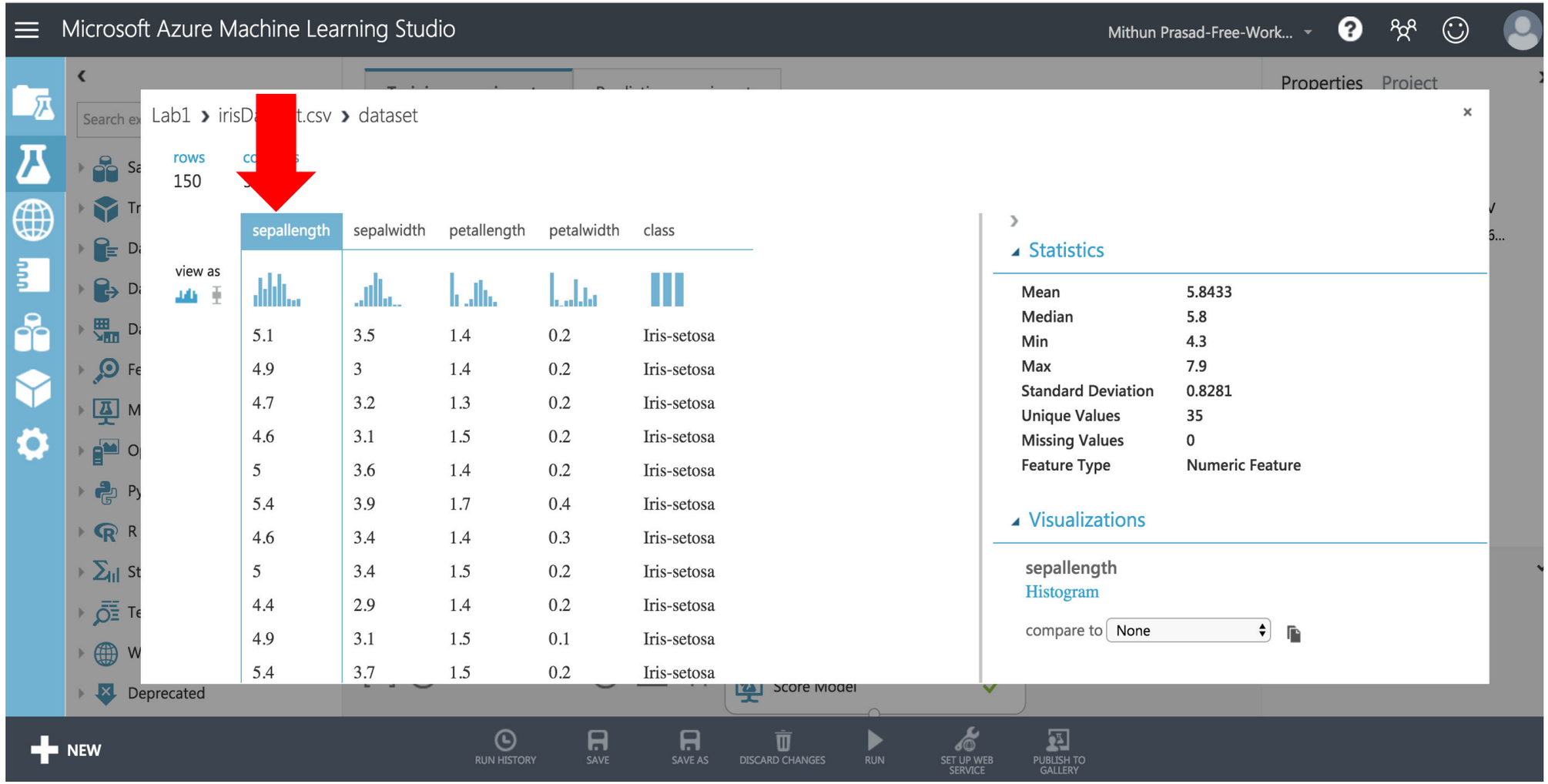
Microsoft

STATISTICS



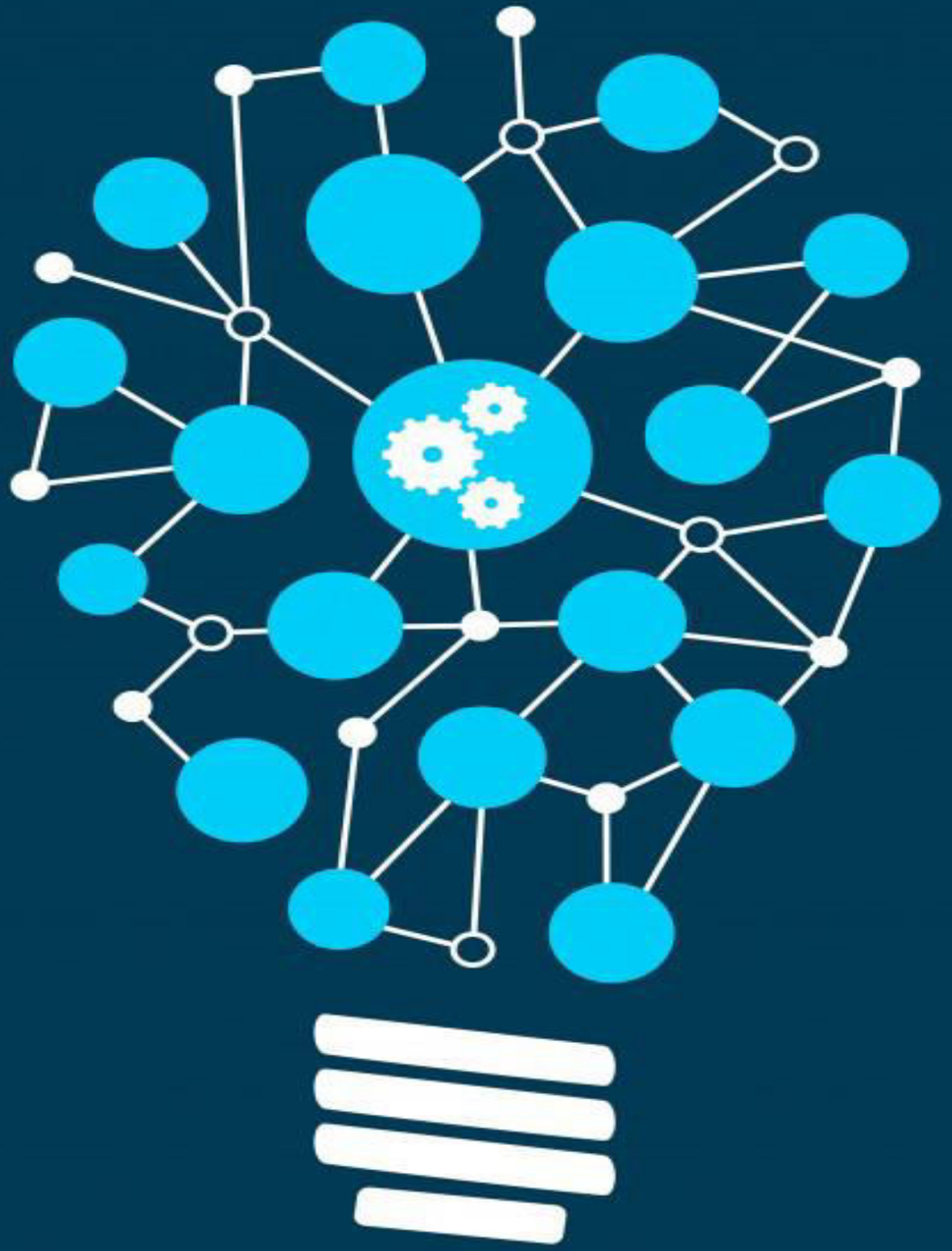
Microsoft

STATISTICS



Microsoft

ALGORITHMS



Microsoft

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

Finding unusual data points

Discovering structure

Predicting categories

Predicting values

Three or more

Two

ANOMALY DETECTION

- **One-class SVM** — >100 features, aggressive boundary
- **PCA-based anomaly detection** — Fast training

CLUSTERING

- **K-means**

MULTI-CLASS CLASSIFICATION

- Fast training, linear model — **Multiclass logistic regression**
- Accuracy, long training times — **Multiclass neural network**
- Accuracy, fast training — **Multiclass decision forest**
- Accuracy, small memory footprint — **Multiclass decision jungle**
- Depends on the two-class classifier, see notes below — **One-v-all multiclass**

REGRESSION

- **Ordinal regression** — Data in rank ordered categories
- **Poisson regression** — Predicting event counts
- **Fast forest quantile regression** — Predicting a distribution
- **Linear regression** — Fast training, linear model
- **Bayesian linear regression** — Linear model, small data sets
- **Neural network regression** — Accuracy, long training time
- **Decision forest regression** — Accuracy, fast training
- **Boosted decision tree regression** — Accuracy, fast training, large memory footprint

TWO-CLASS CLASSIFICATION

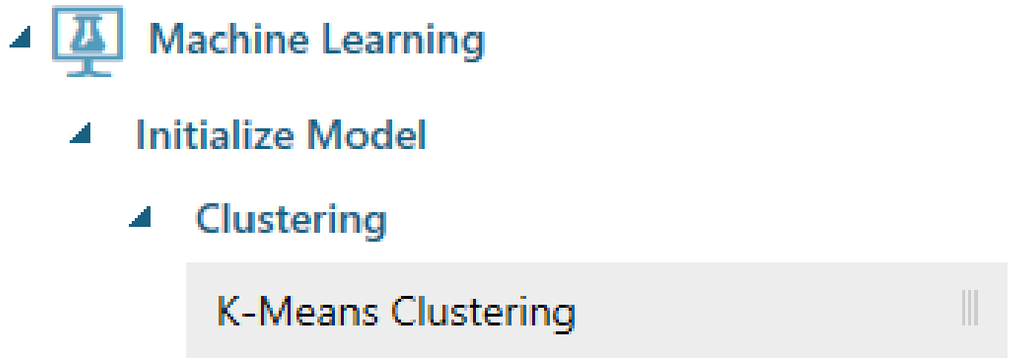
- **Two-class SVM** — >100 features, linear model
- **Two-class averaged perceptron** — Fast training, linear model
- **Two-class logistic regression** — Fast training, linear model
- **Two-class Bayes point machine** — Fast training, linear model

- Accuracy, fast training — **Two-class decision forest**
- Accuracy, fast training, large memory footprint — **Two-class boosted decision tree**
- Accuracy, small memory footprint — **Two-class decision jungle**
- >100 features — **Two-class locally deep SVM**
- Accuracy, long training times — **Two-class neural network**

Microsoft

CLUSTERING

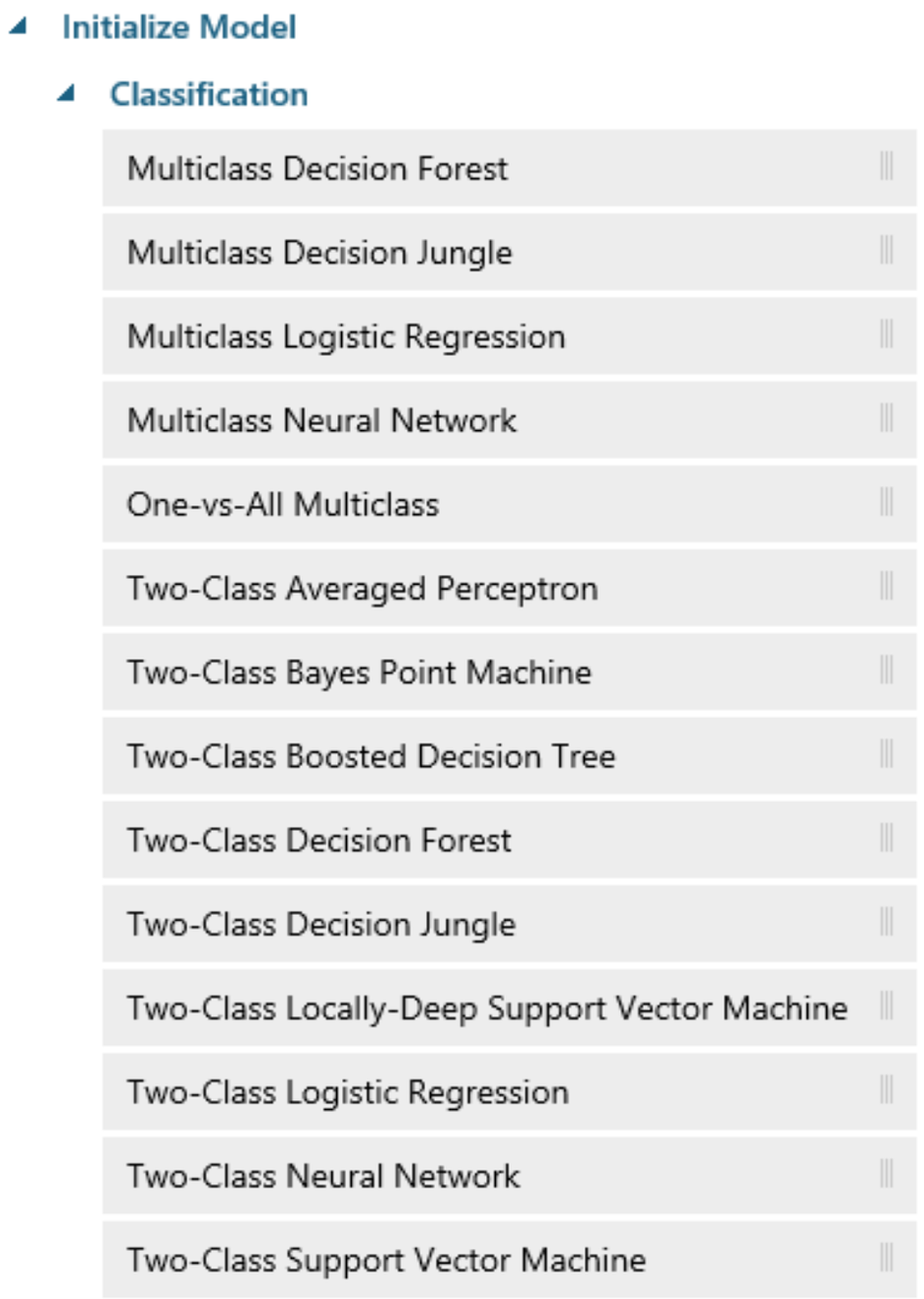
Grouping items based on defined Features



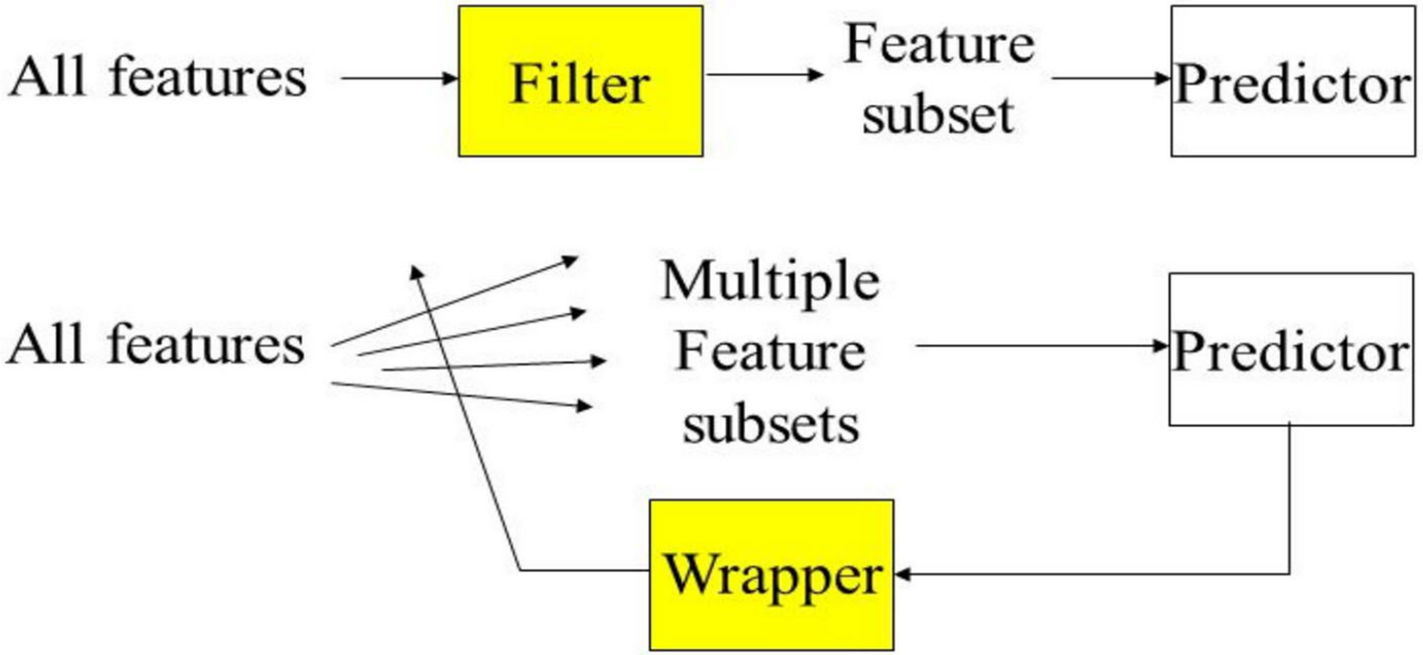
Microsoft

CLASSIFICATION

Predicting the class or category for a single instance of data



FEATURE SELECTION (FILTER)



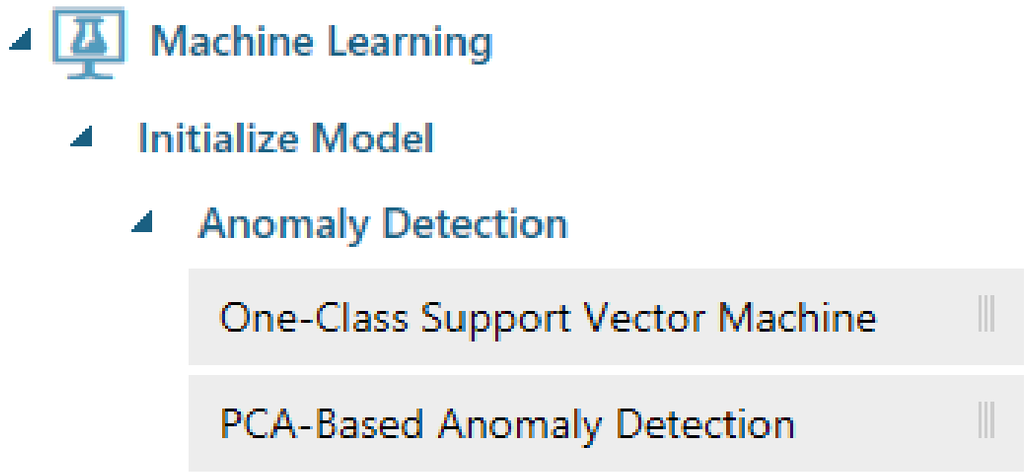
Filter Algorithms

- ✓ Pearson Correlation
- Mutual Information
- Kendall Correlation
- Spearman Correlation
- Chi Squared
- Fisher Score
- Count Based

Microsoft

ANOMAY DETECTION

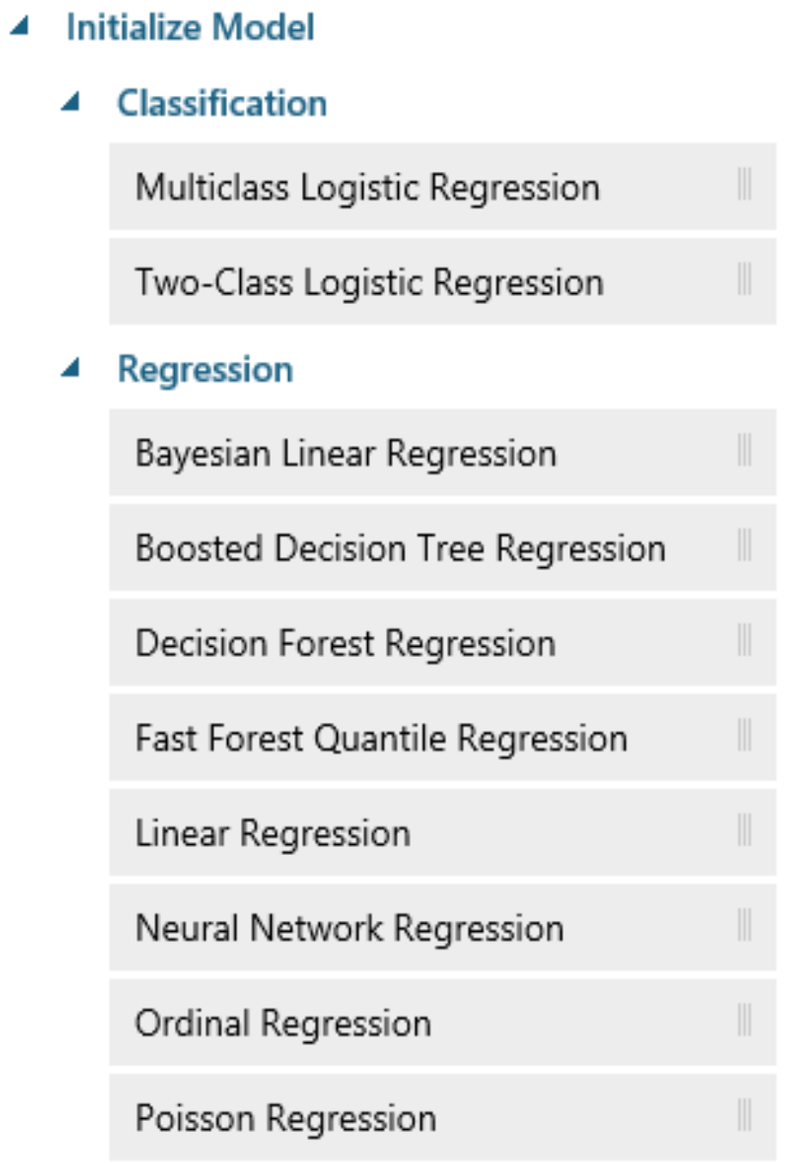
Selecting items based on unusual or suspicious patterns



Microsoft

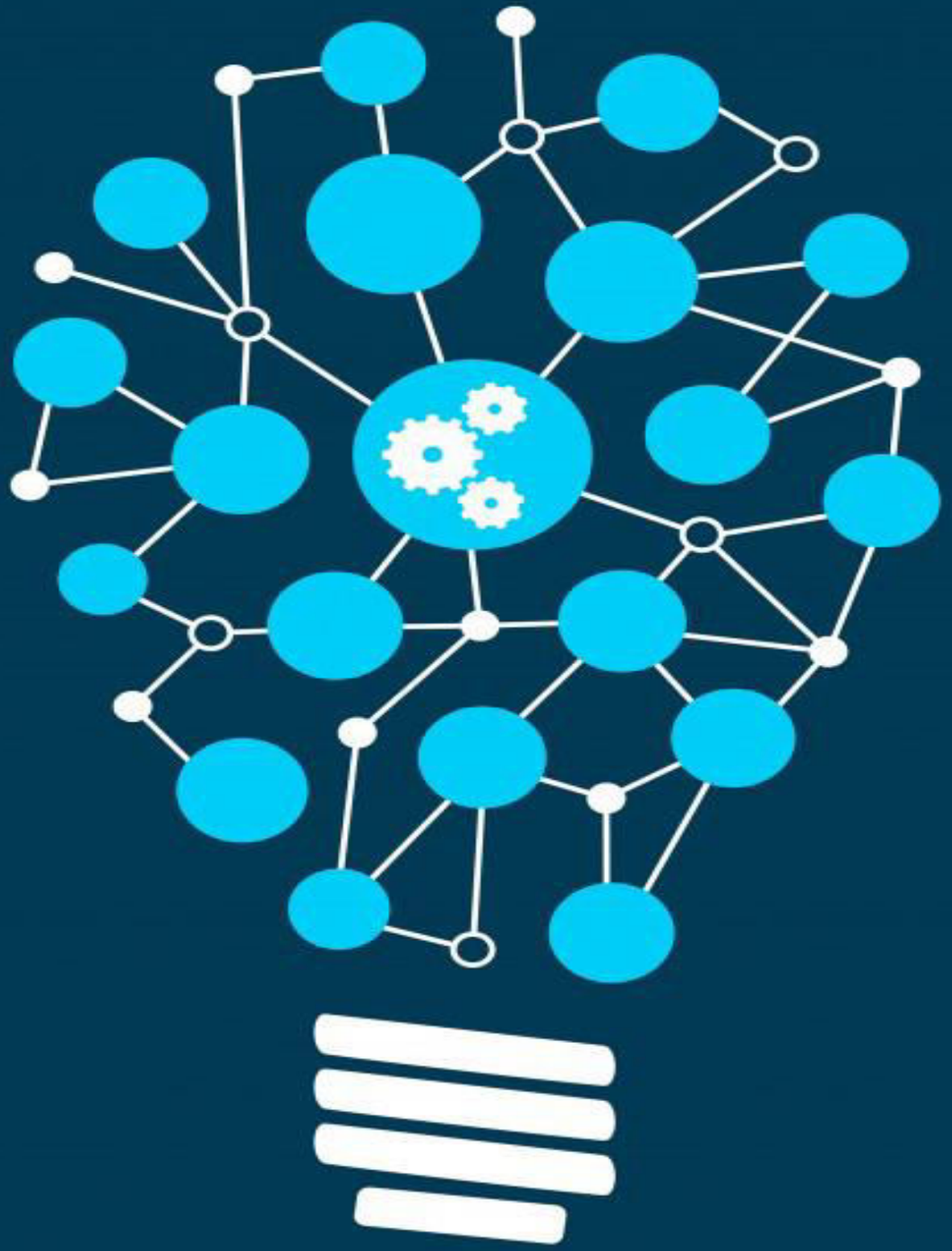
REGRESSION

Predicting the value of a datum given its history



Microsoft

MODEL SCORING AND EVALUATION

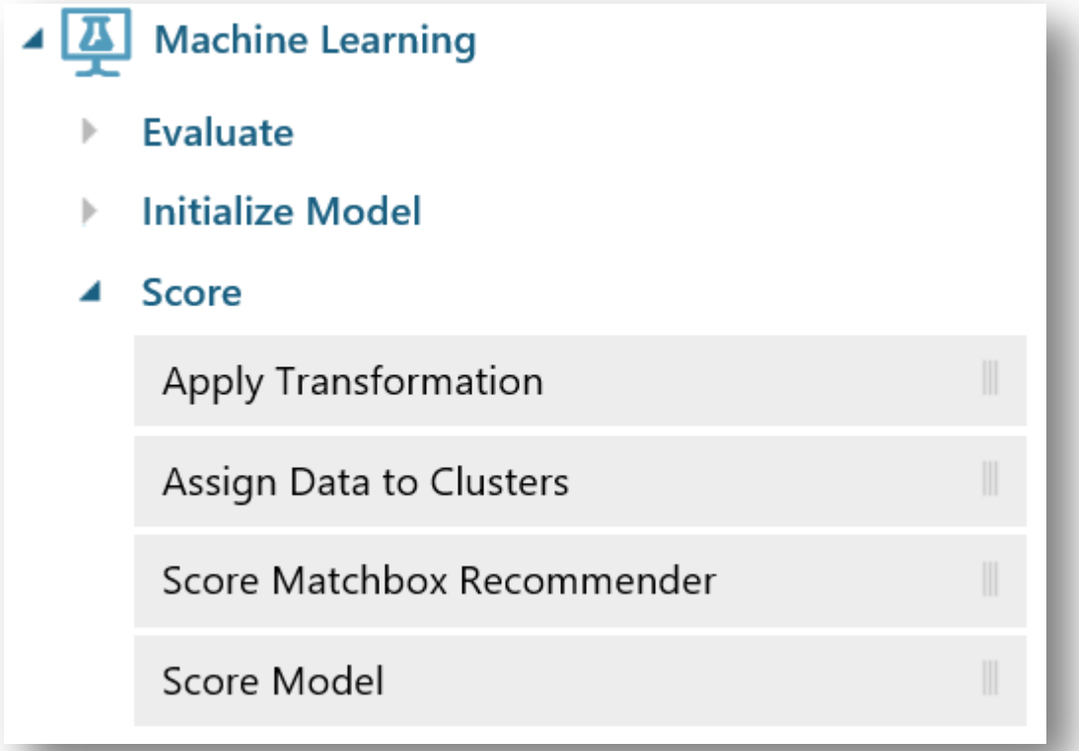


Microsoft

SCORING A MODEL

Apply a trained model to:

- A list of recommended items
- Forecasts for time series models
- Estimates of projected demand, volume, or other numeric quantity, for regression models
- Cluster assignments
- A predicted class or outcome, for classification models
- Probability scores associated with these outputs



Microsoft

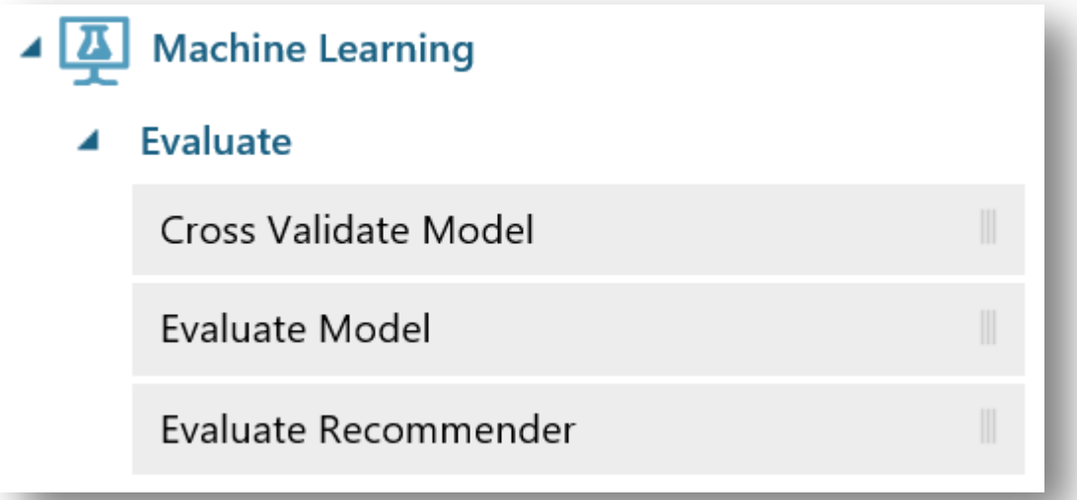
EVALUATING A MODEL

Metrics for Classification Models

- Accuracy, Recall, Precision, F1-Score
- AUC
- Average Log Loss
- Training Log Loss

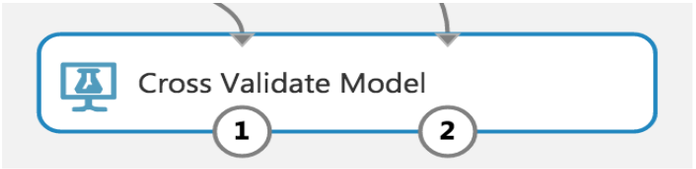
Metrics for Regression Models

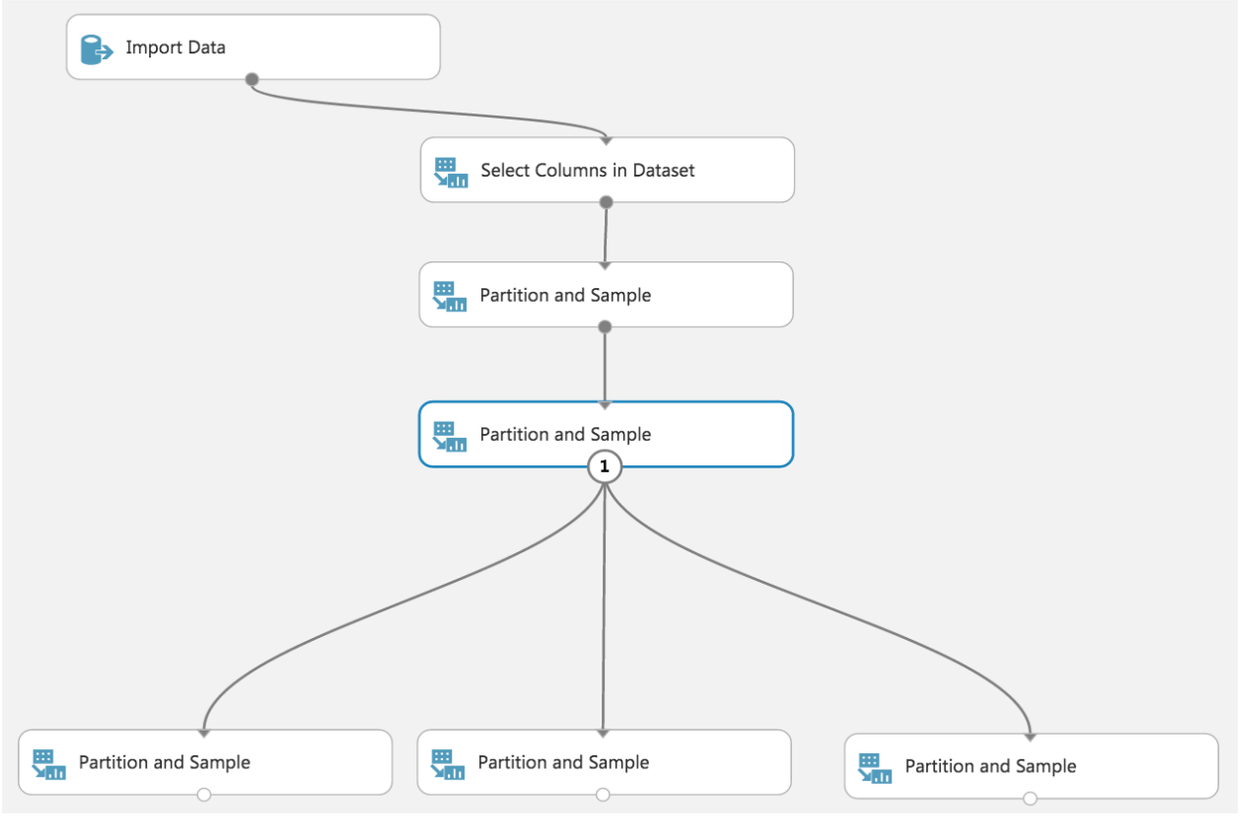
- Mean absolute error (MAE)
- Root mean squared error (RMSE)
- Relative absolute error (RAE)
- Relative squared error (RSE)
- Coefficient of determination

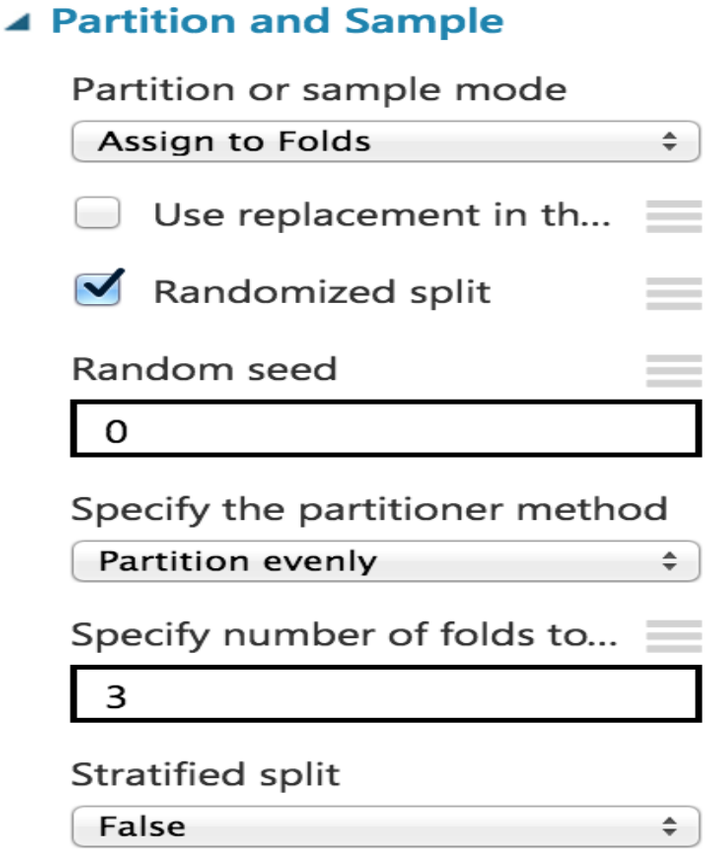


Microsoft

CROSS VALIDATION

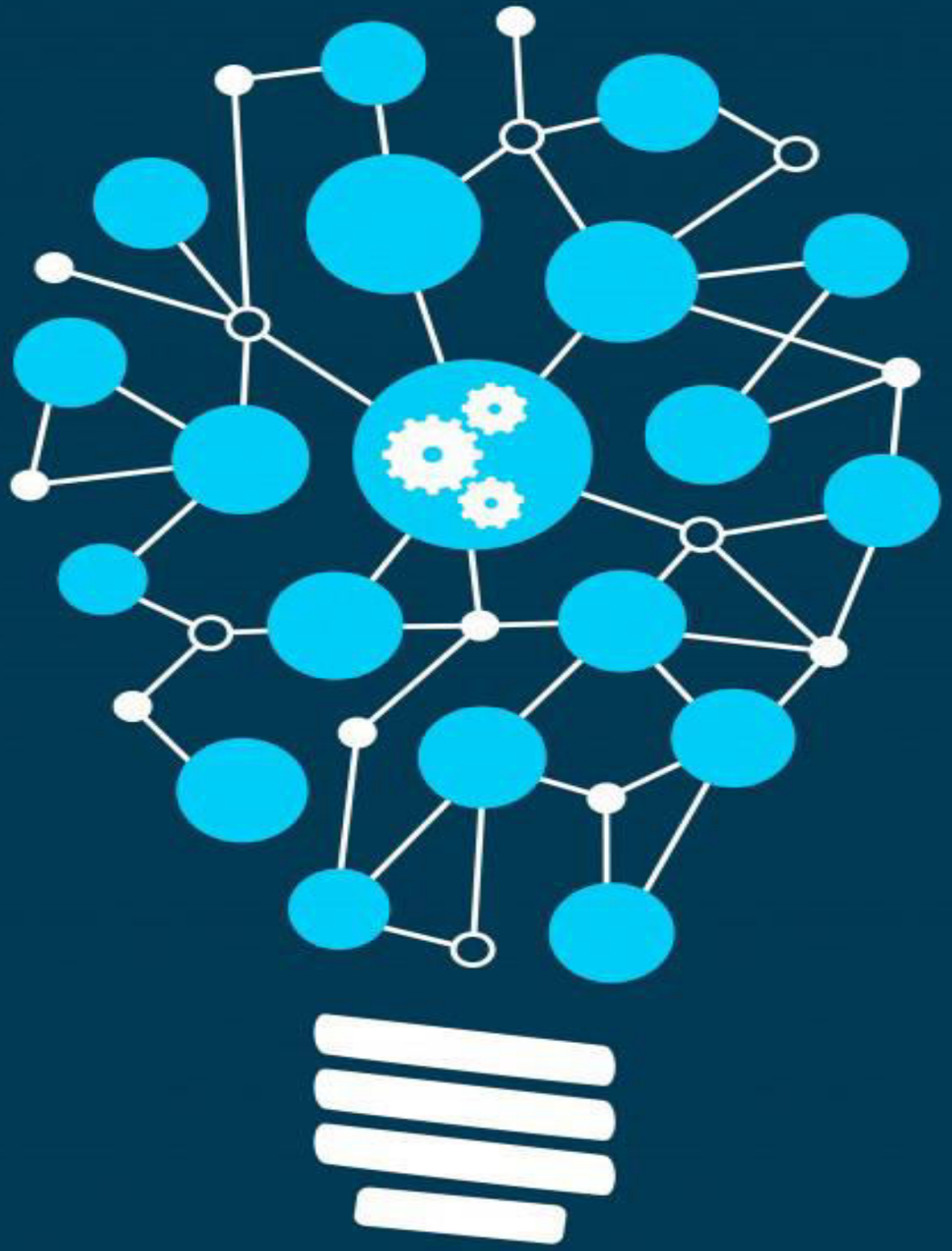






Microsoft

CUSTOMIZATION



Microsoft

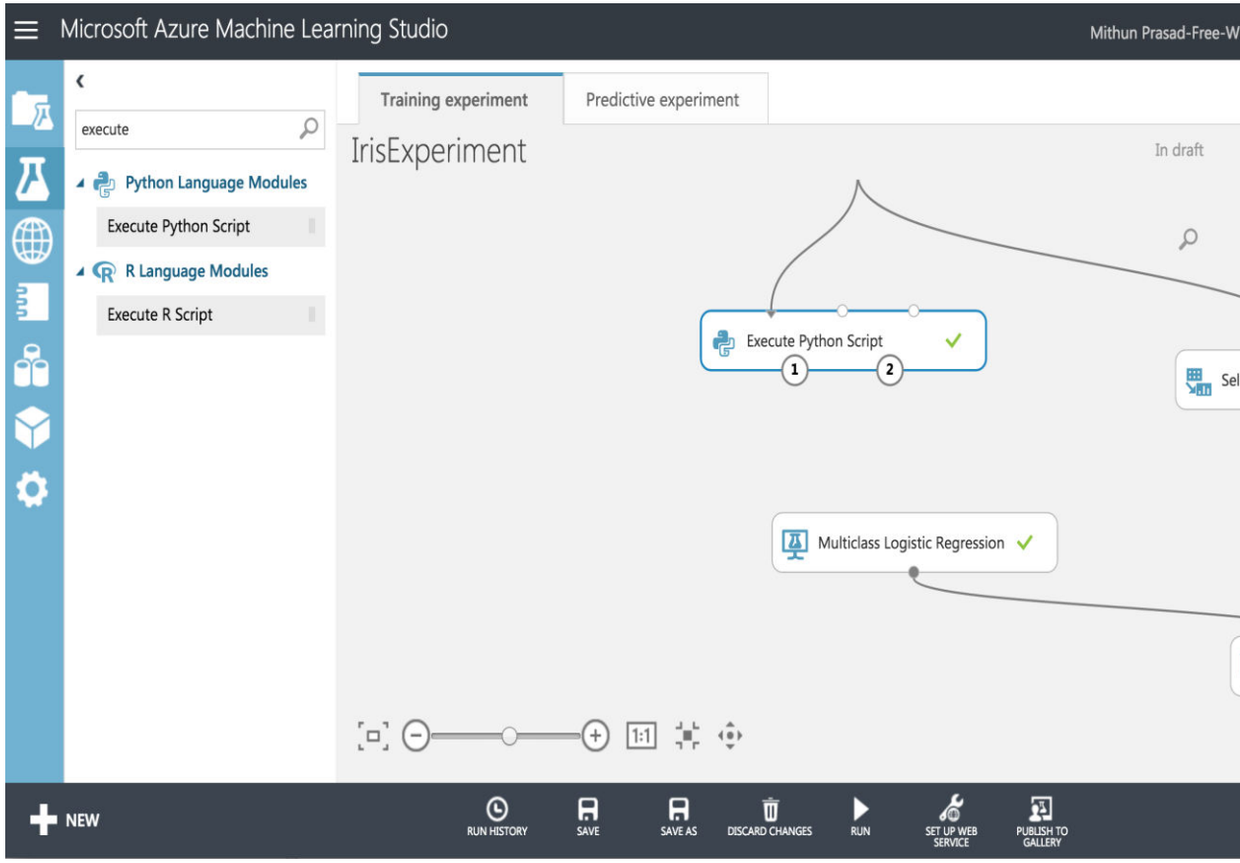
HOW TO USE EXECUTE PYTHON SCRIPT

1. Add the **Execute Python Script** module to your experiment.
2. Connect any datasets that you want to use for input. You can also provide a zipped file containing custom resources.

Dataset1. An optional dataset from your Machine Learning Studio workspace, containing input data or values.

Dataset2. A second dataset, also optional.

Script bundle. A zipped file containing custom resources.

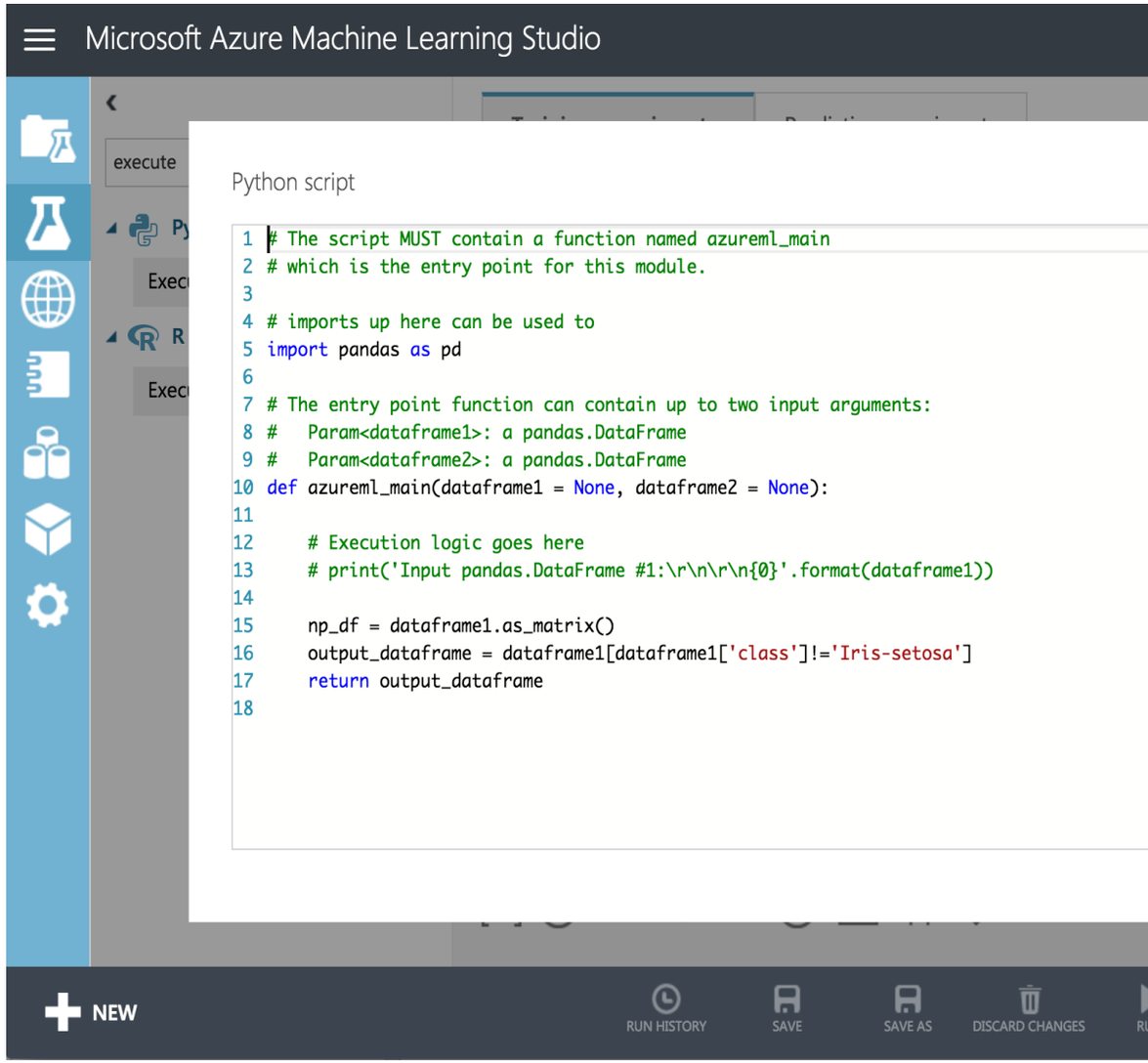


Microsoft

CUSTOM TASKS

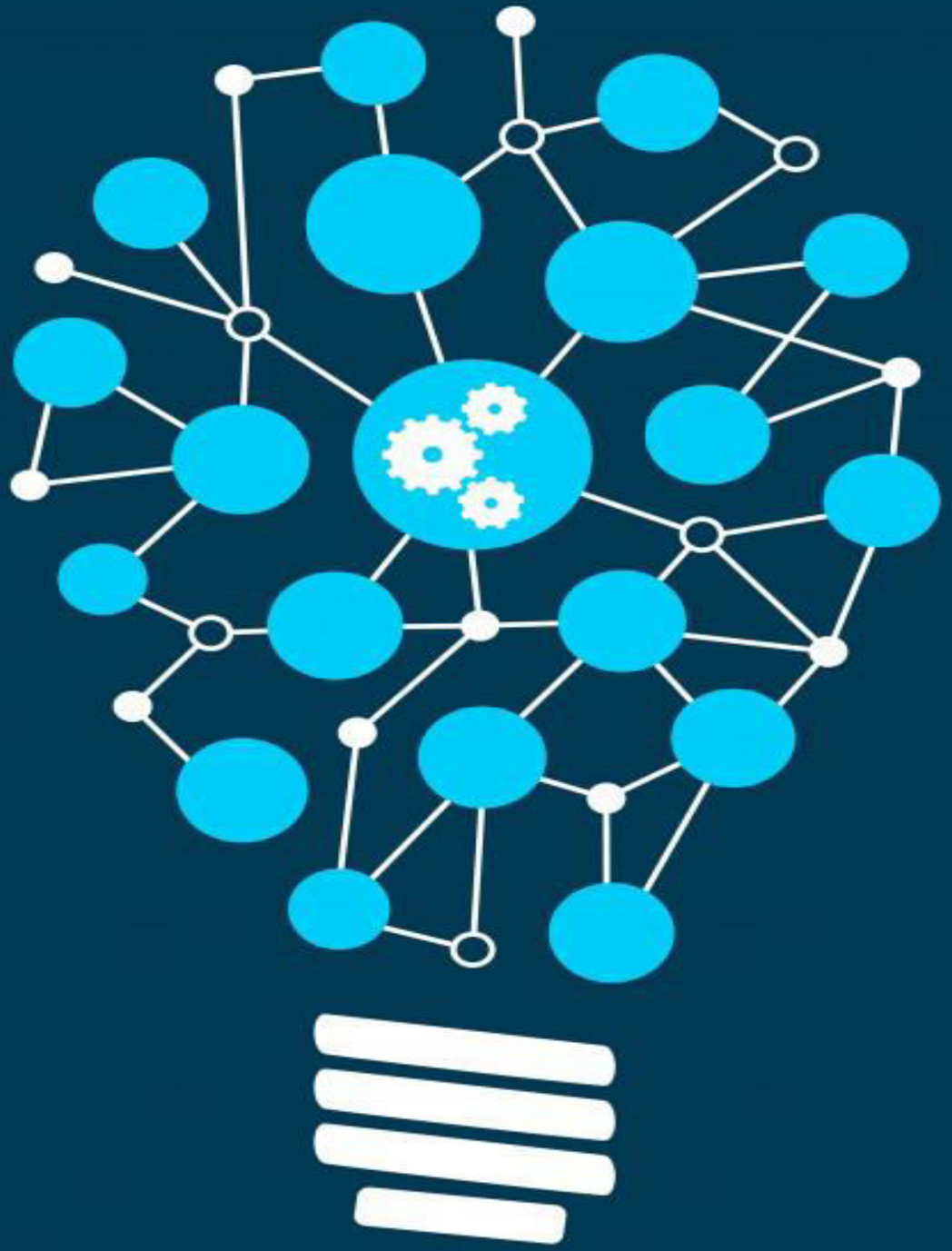
Integrate Ipython notebooks with Azure Machine Learning to perform custom tasks:

- Visualization
- Use Python client libraries to enumerate datasets and models in your workspace
- Read, load, and manipulate data



Microsoft

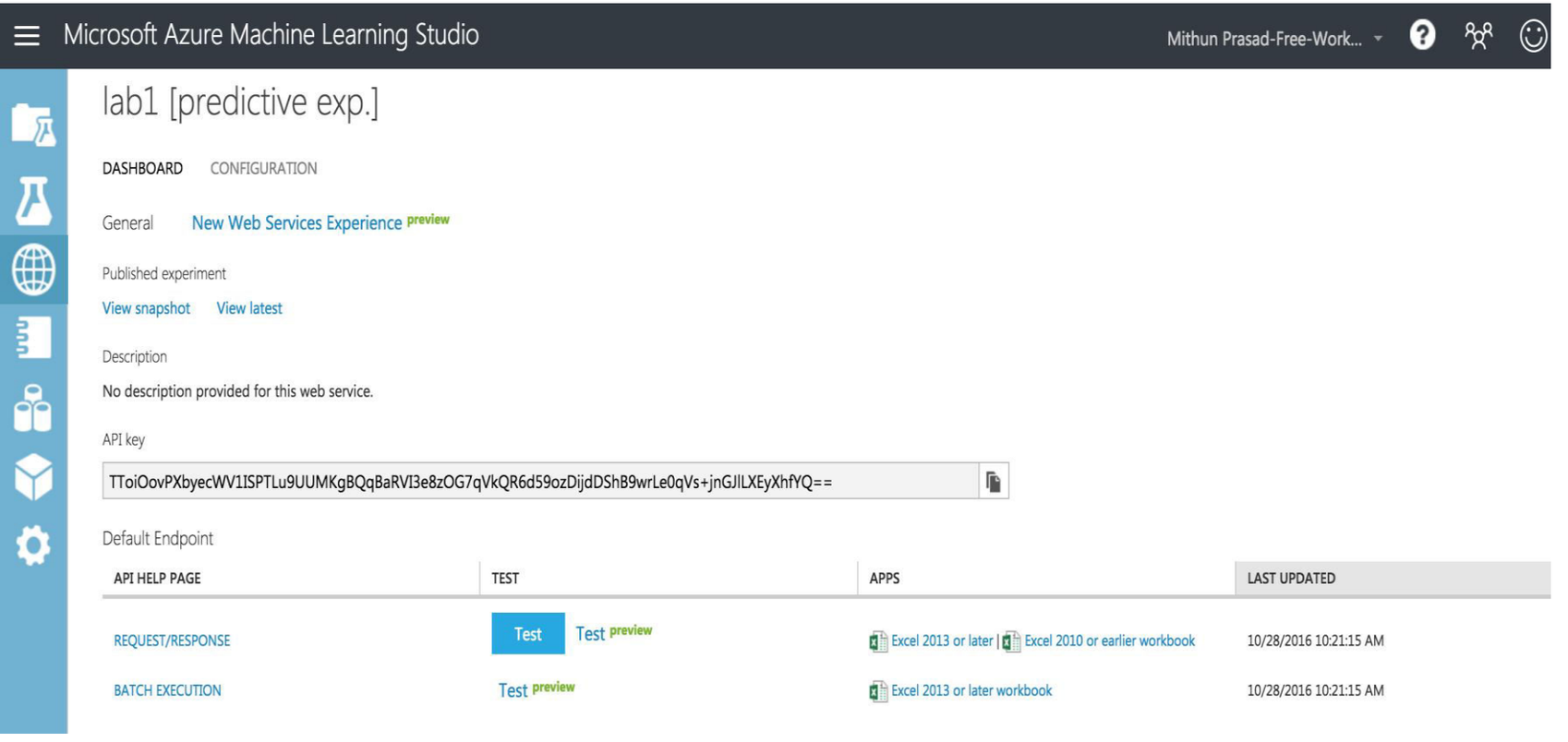
DEPLOY AS A WEB SERVICE

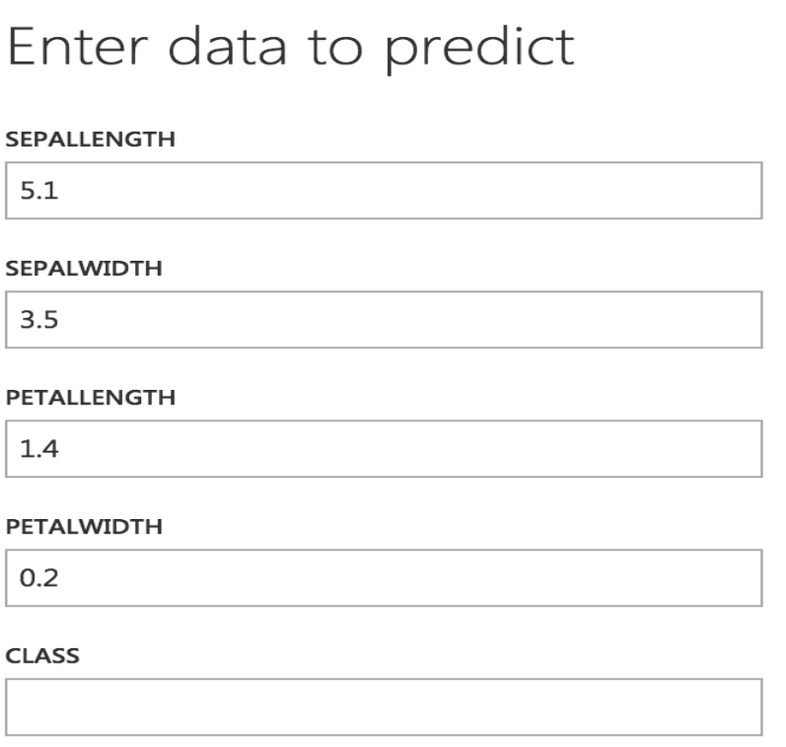


Microsoft

WEB SERVICE

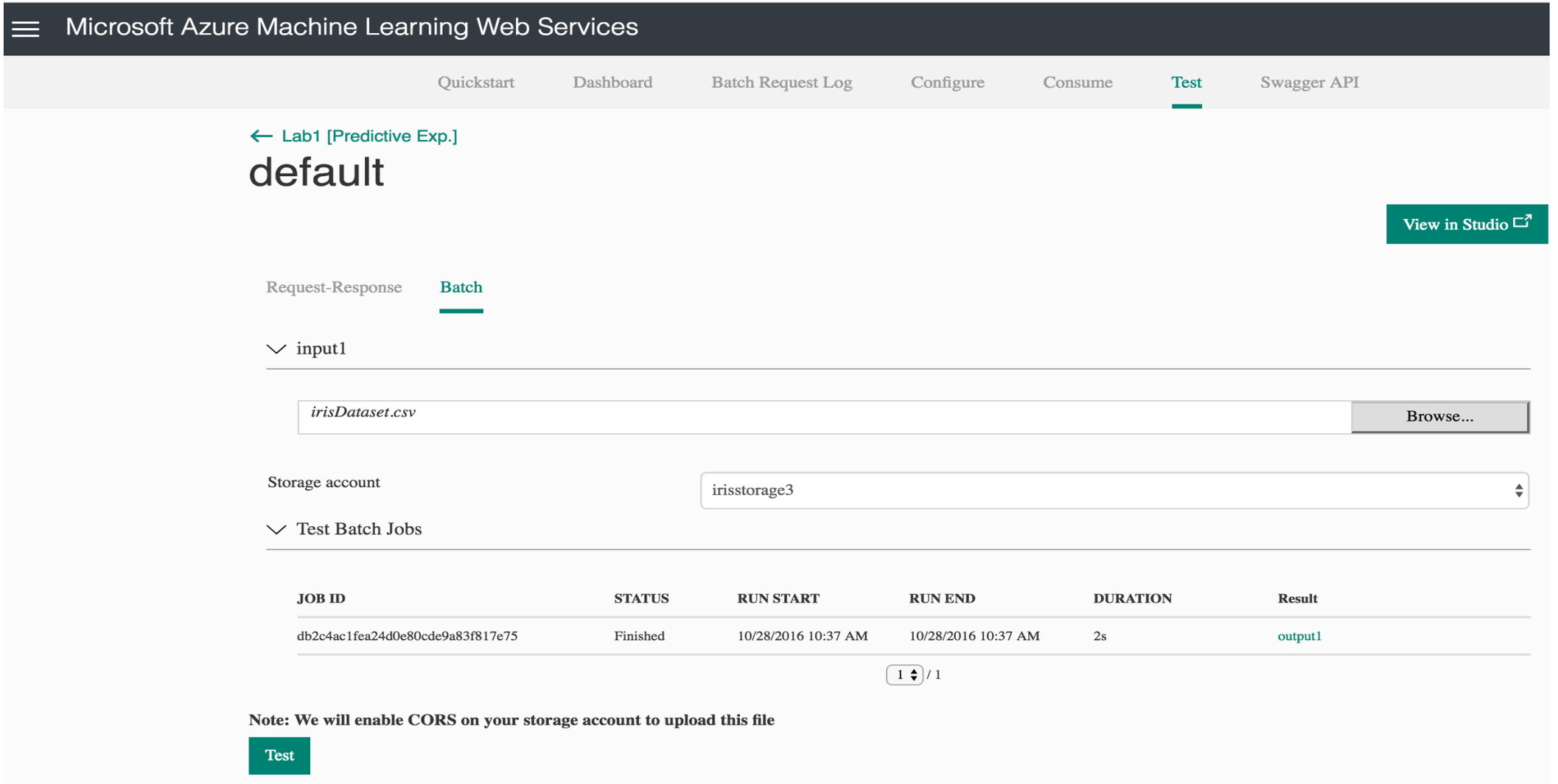
1. Deploy the experiment as a web service





Microsoft

BATCH CLASSIFICATION



Microsoft