Azure
Machine Learning

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.

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)

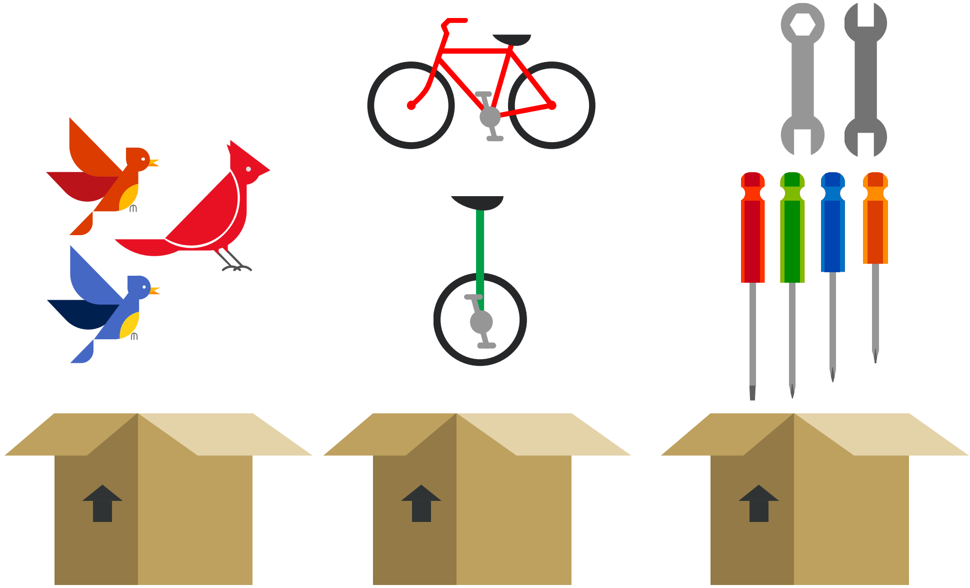
Another way of categorizing ML approaches, is to the desired output:

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

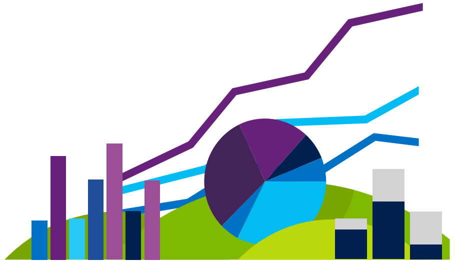
Which category

(Classification)



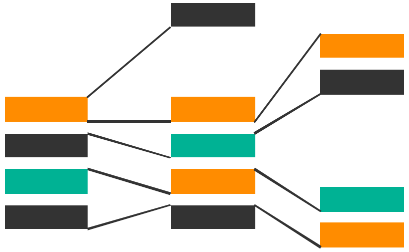
How much/many

(Regression)



Which group

(Clustering, Recommender)

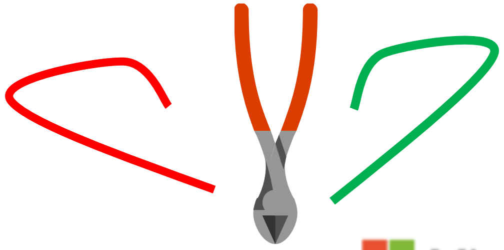


Is it odd

(Anomaly)

Which action

(Reinforcement Learning)



Microsoft

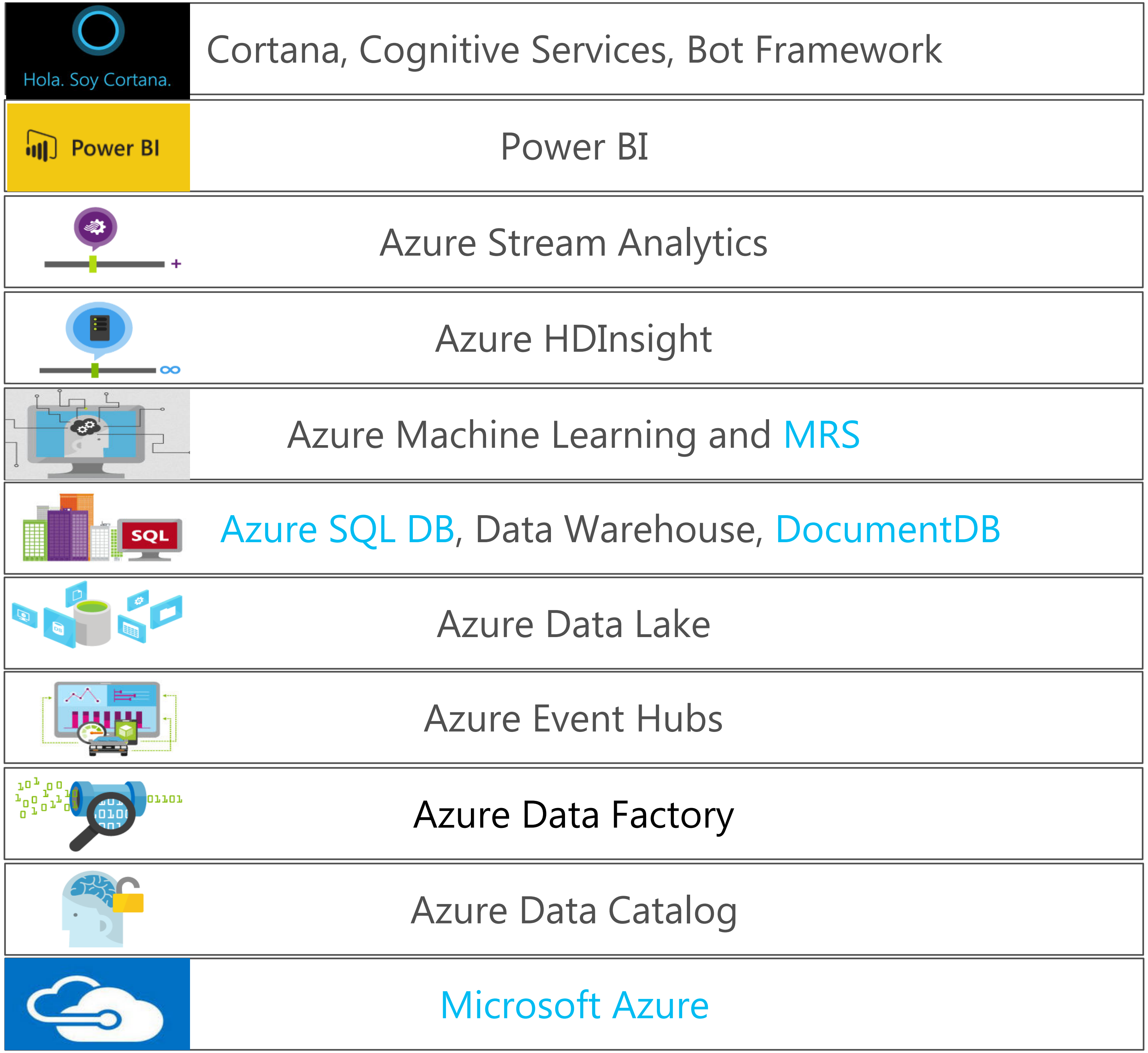
MACHINE LEARNING 101

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

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

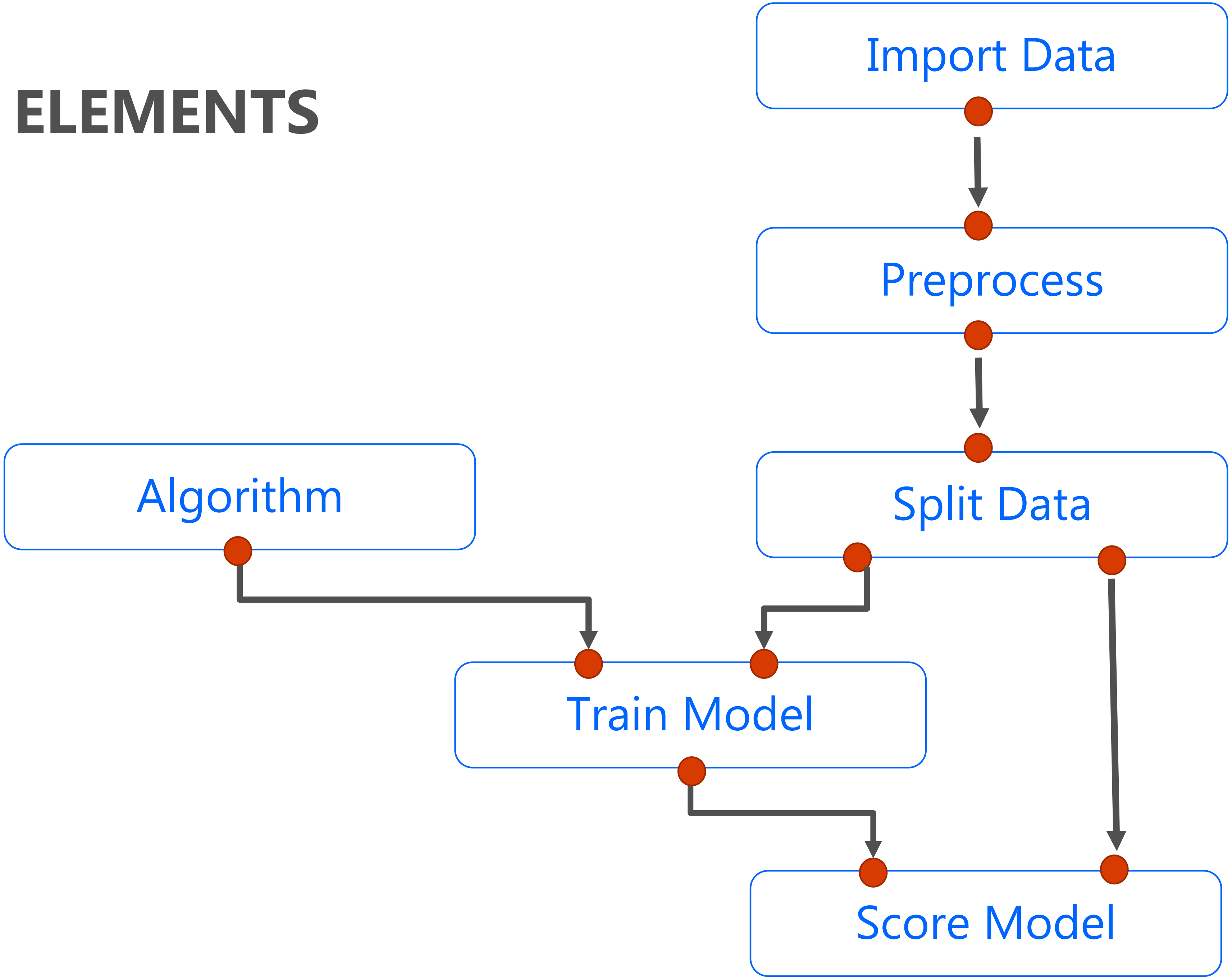
Azure Event Hubs

Azure Data Factory

Azure Data Catalog

Microsoft Azure

AZURE ML ELEMENTS



Import Data

Preprocess

Split Data

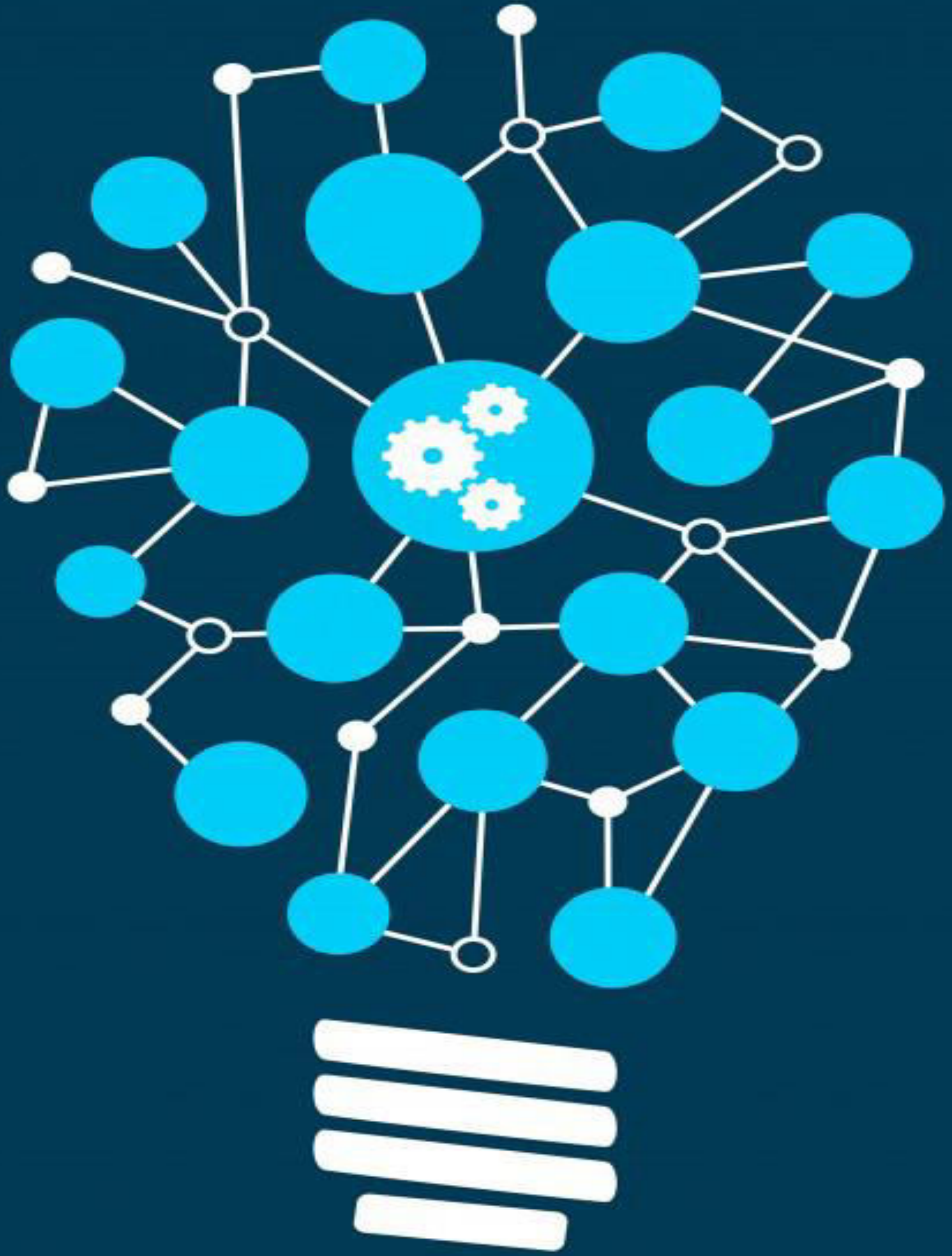
Algorithm

Train Model

Score Model

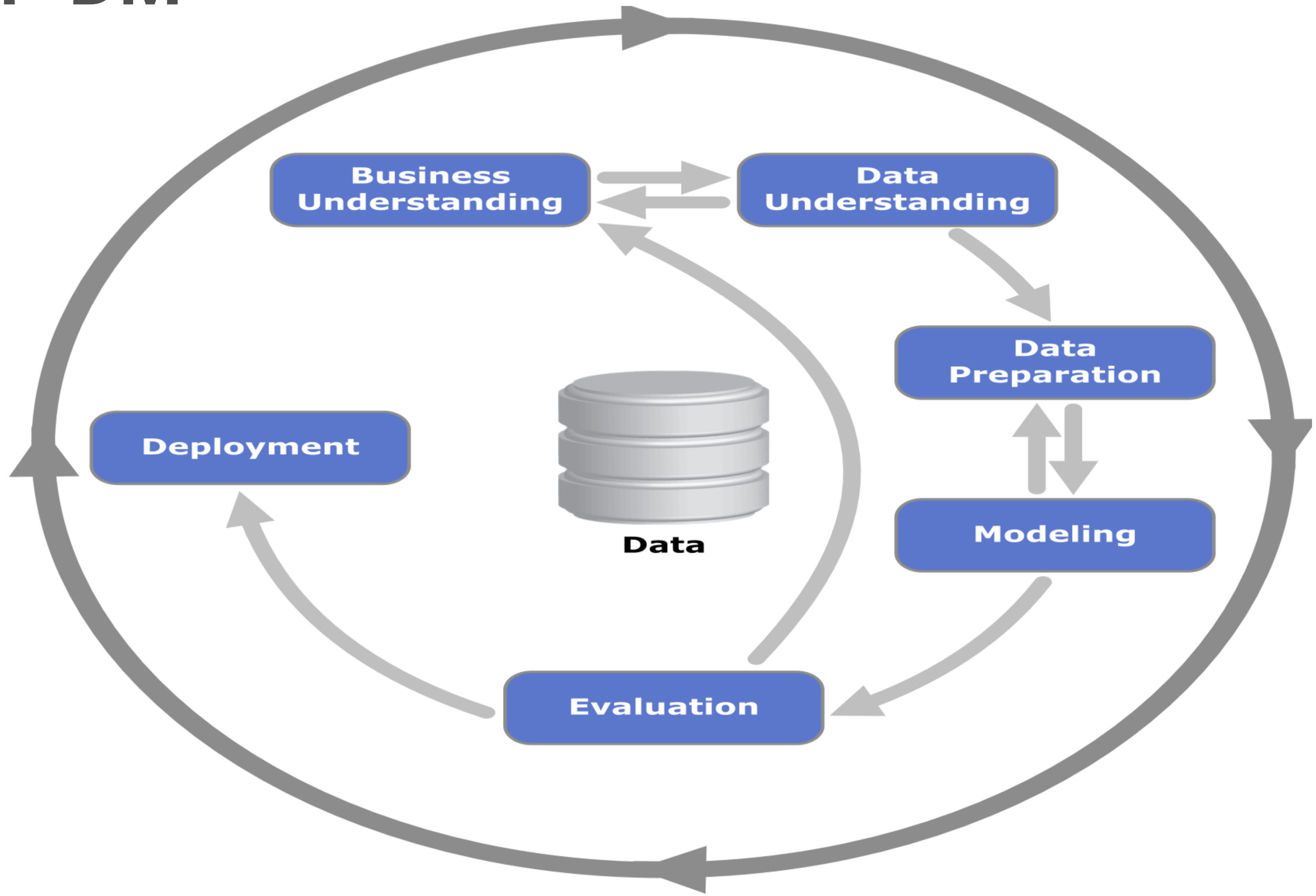
Microsoft

AZURE ML STUDIO AND THE TEAM DATA SCIENCE PROCESS

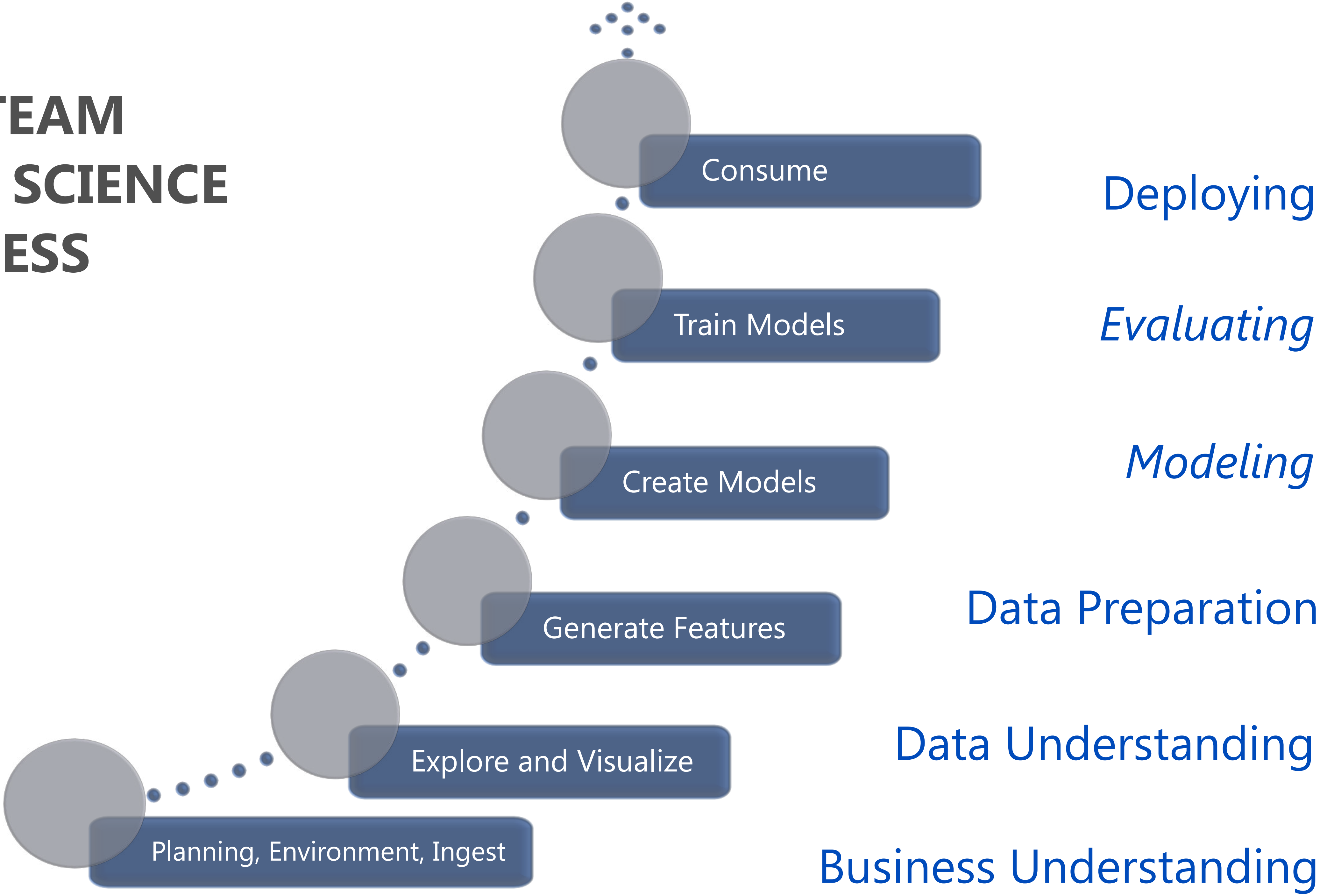


Microsoft

CRISP-DM

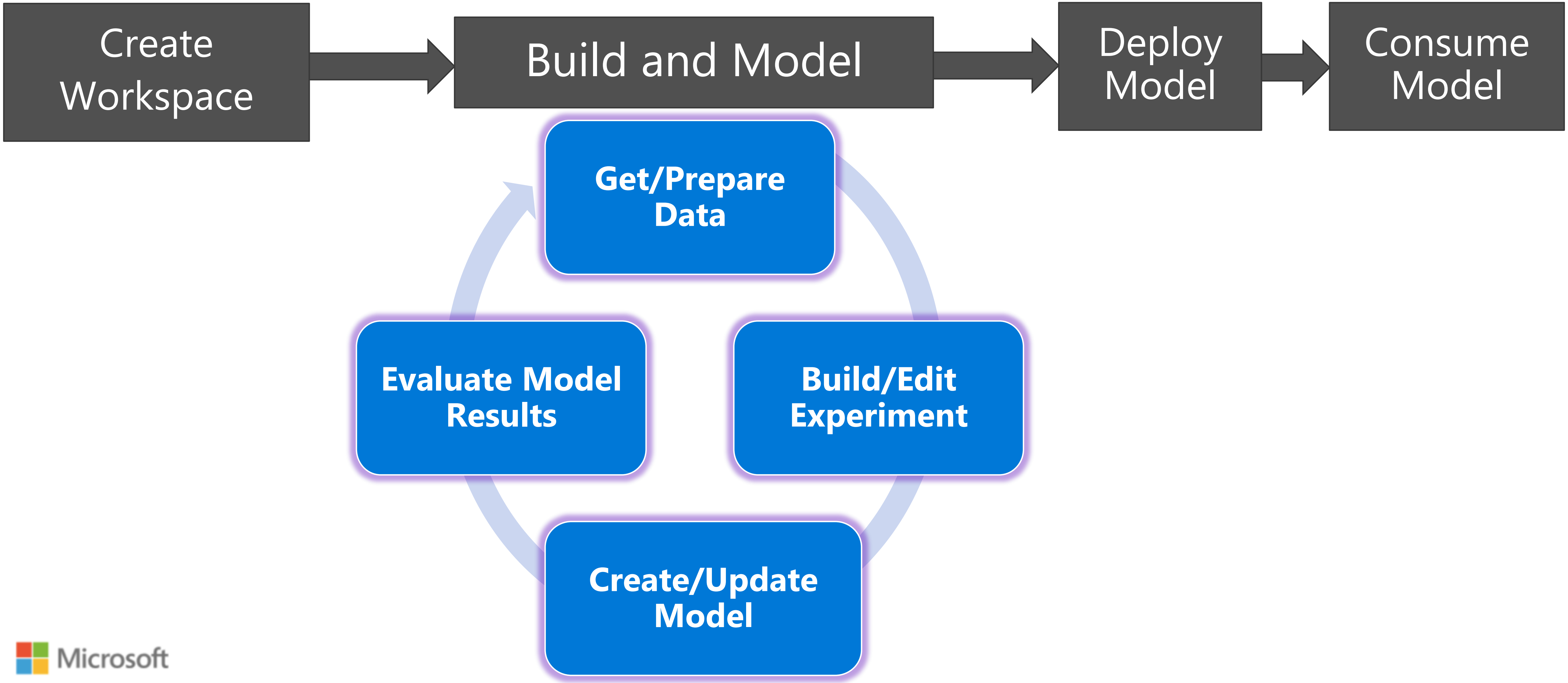

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

THE TEAM DATA SCIENCE PROCESS



Microsoft

CREATING AN EXPERIMENT



Create Workspace

Build and Model

Deploy Model

Consume Model

Get/Prepare Data

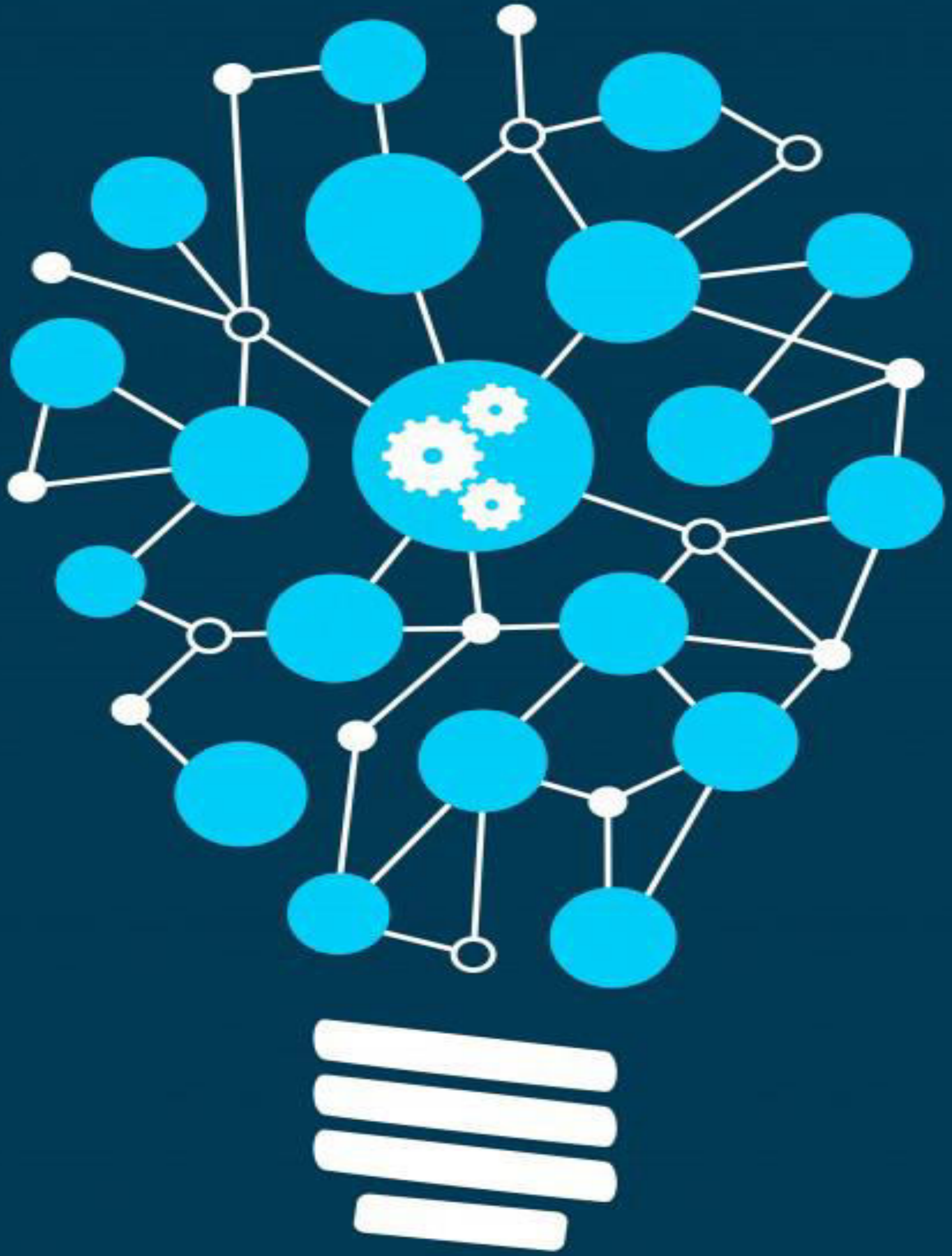
Build/Edit Experiment

Create/Update Model

Evaluate Model Results

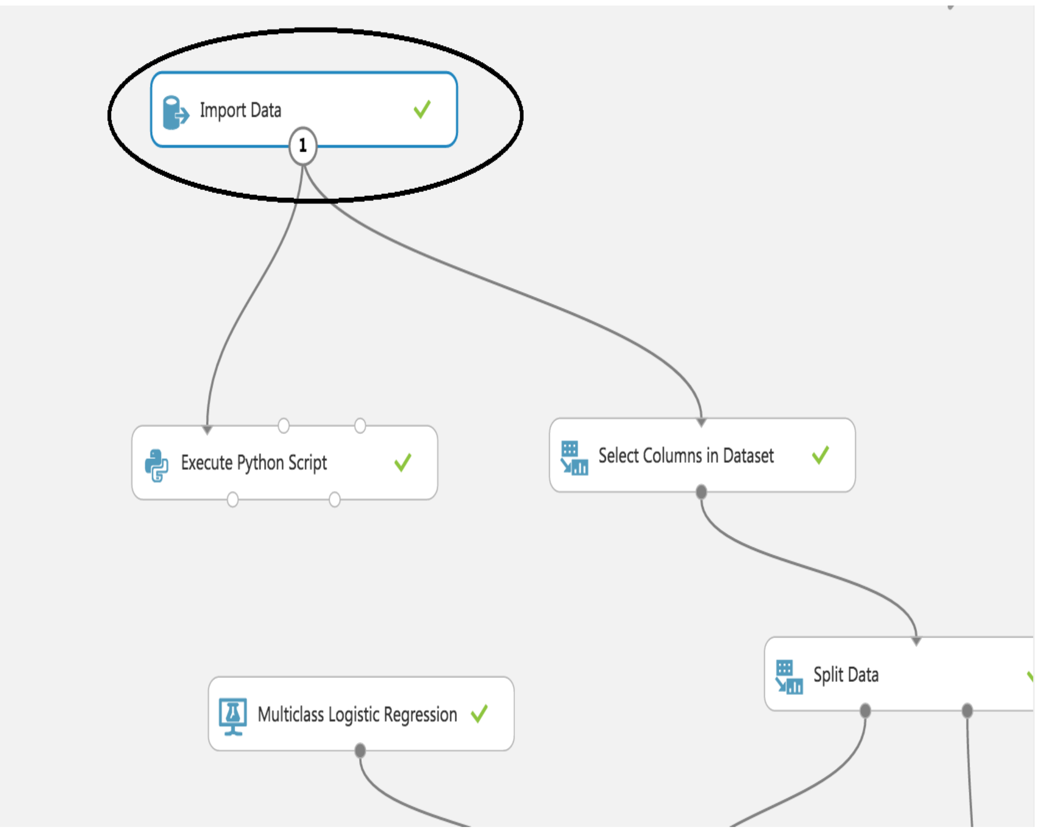
Microsoft

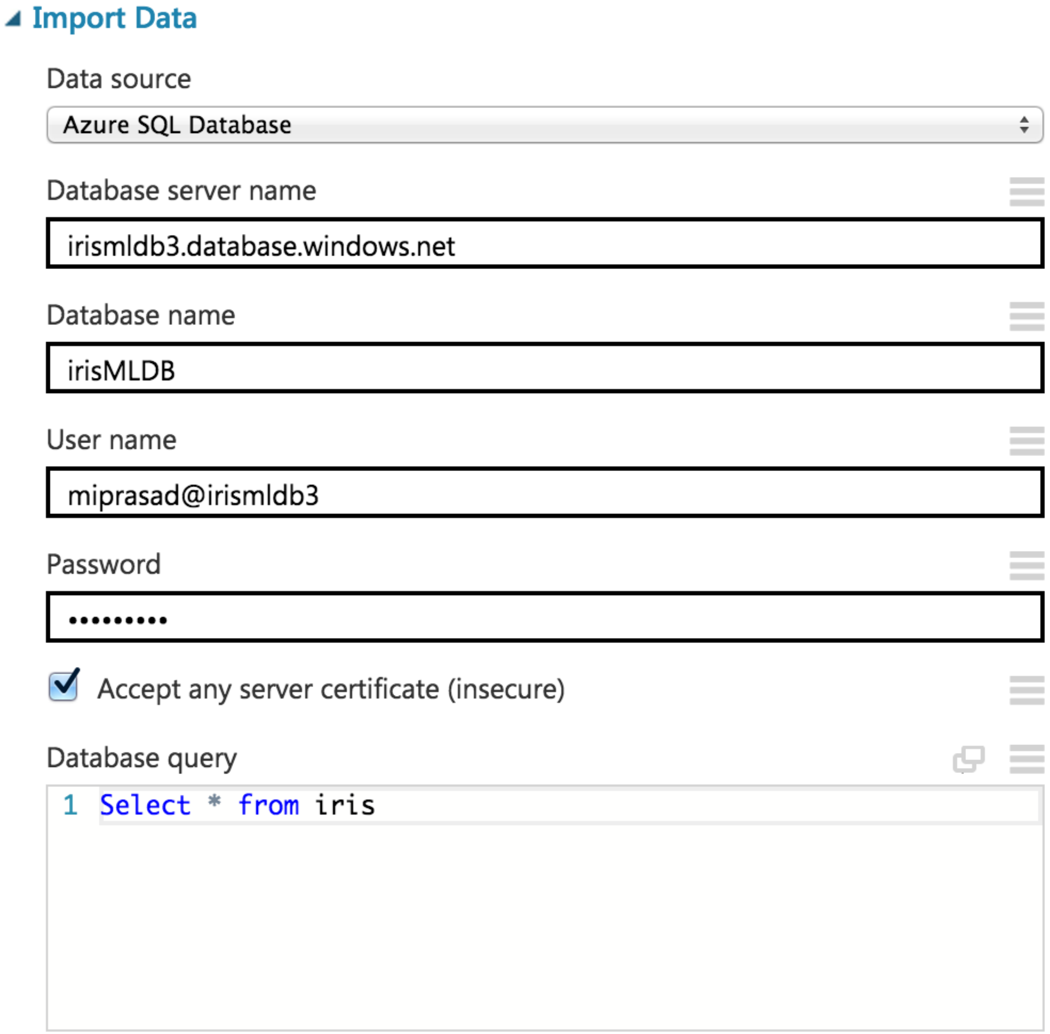
DATA INGESTION AND PREPARATION



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

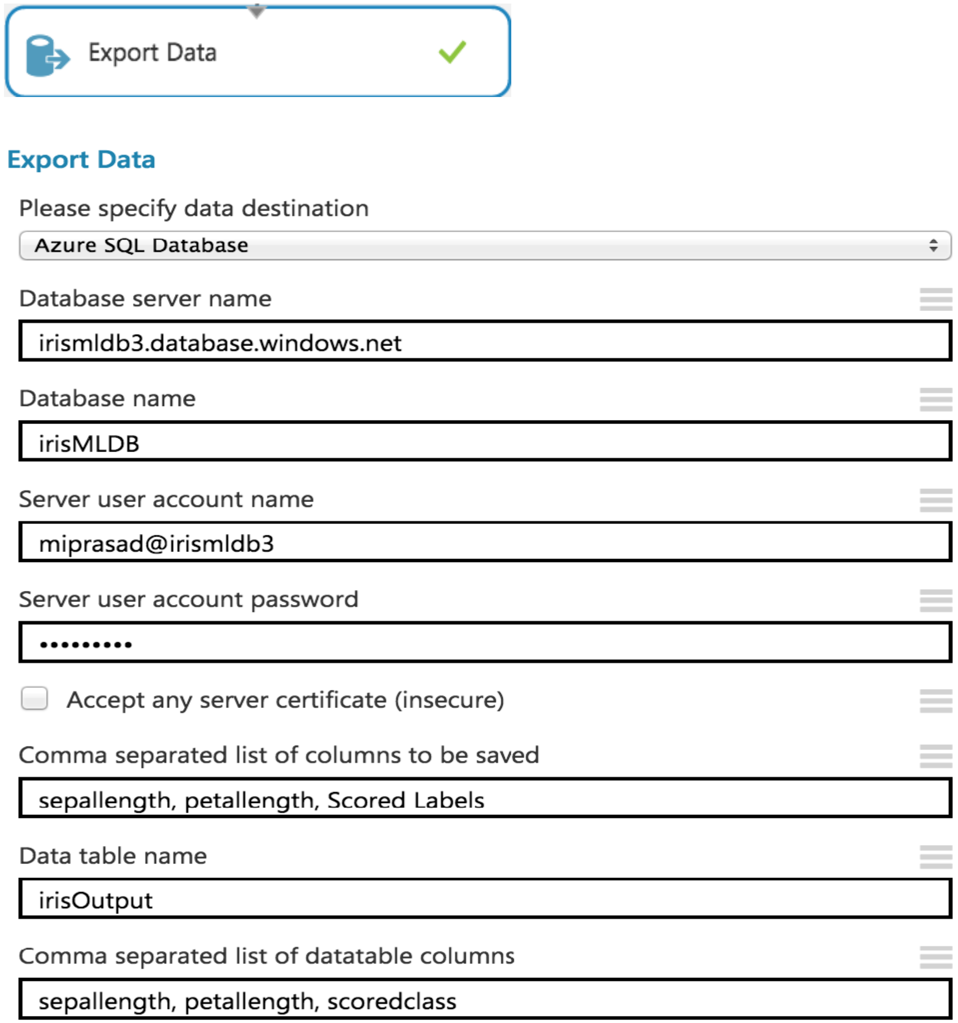
•••••••••

☑ Accept any server certificate (insecure)

Database query

```
Select * from iris
```

DATA ACCESS (EXPORT)



Export Data ✓

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

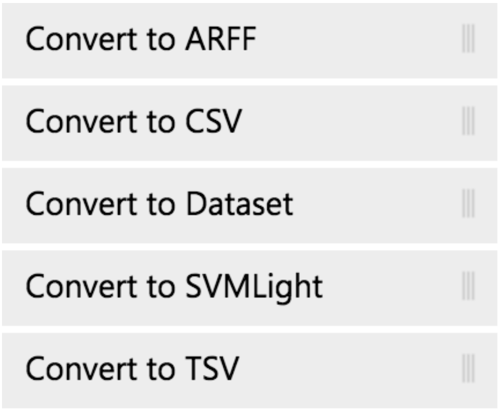
Data table name

irisOutput

Comma separated list of datatable columns

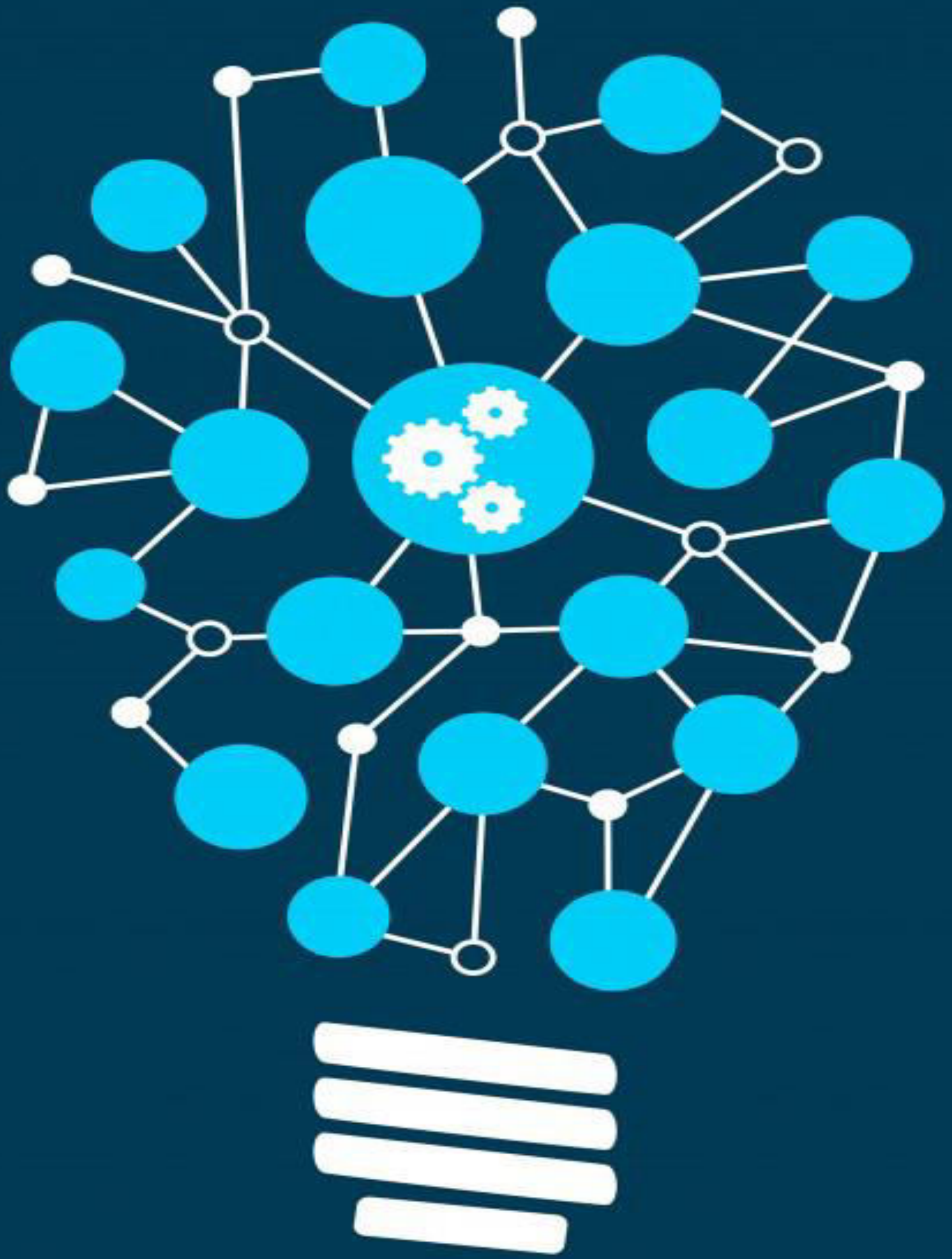
sepallength, petallength, scoredclass

Data Format Conversion



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

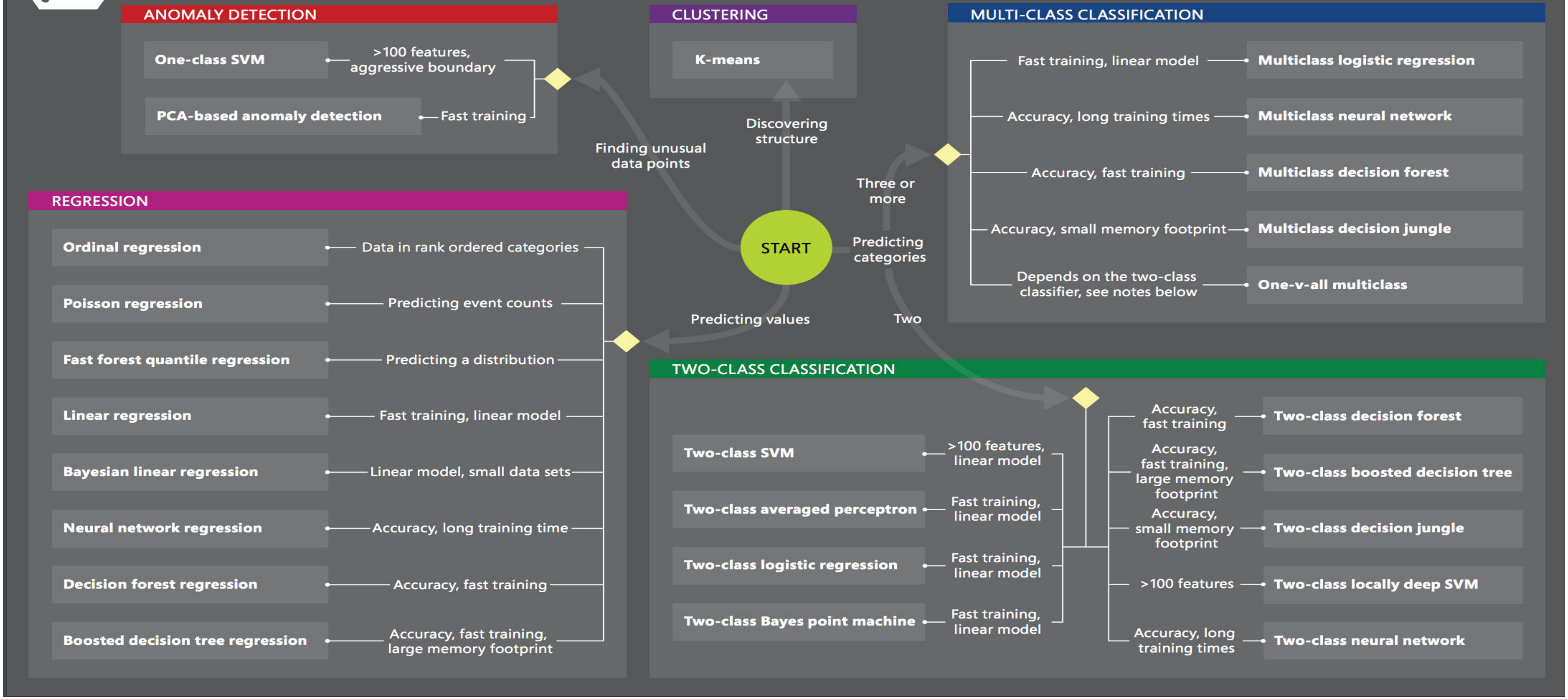
ALGORITHMS



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

Three or more

Predicting categories

Two

Predicting values

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

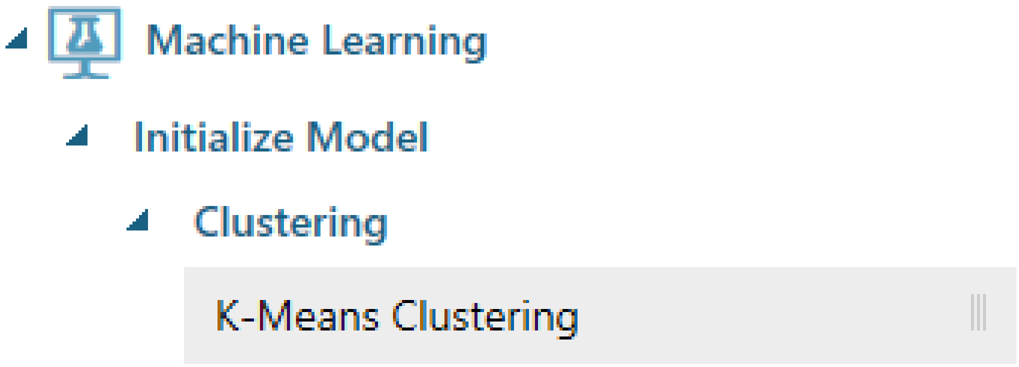
>100 features — Two-class locally deep SVM

Accuracy, long training times — Two-class neural network

Microsoft

CLUSTERING

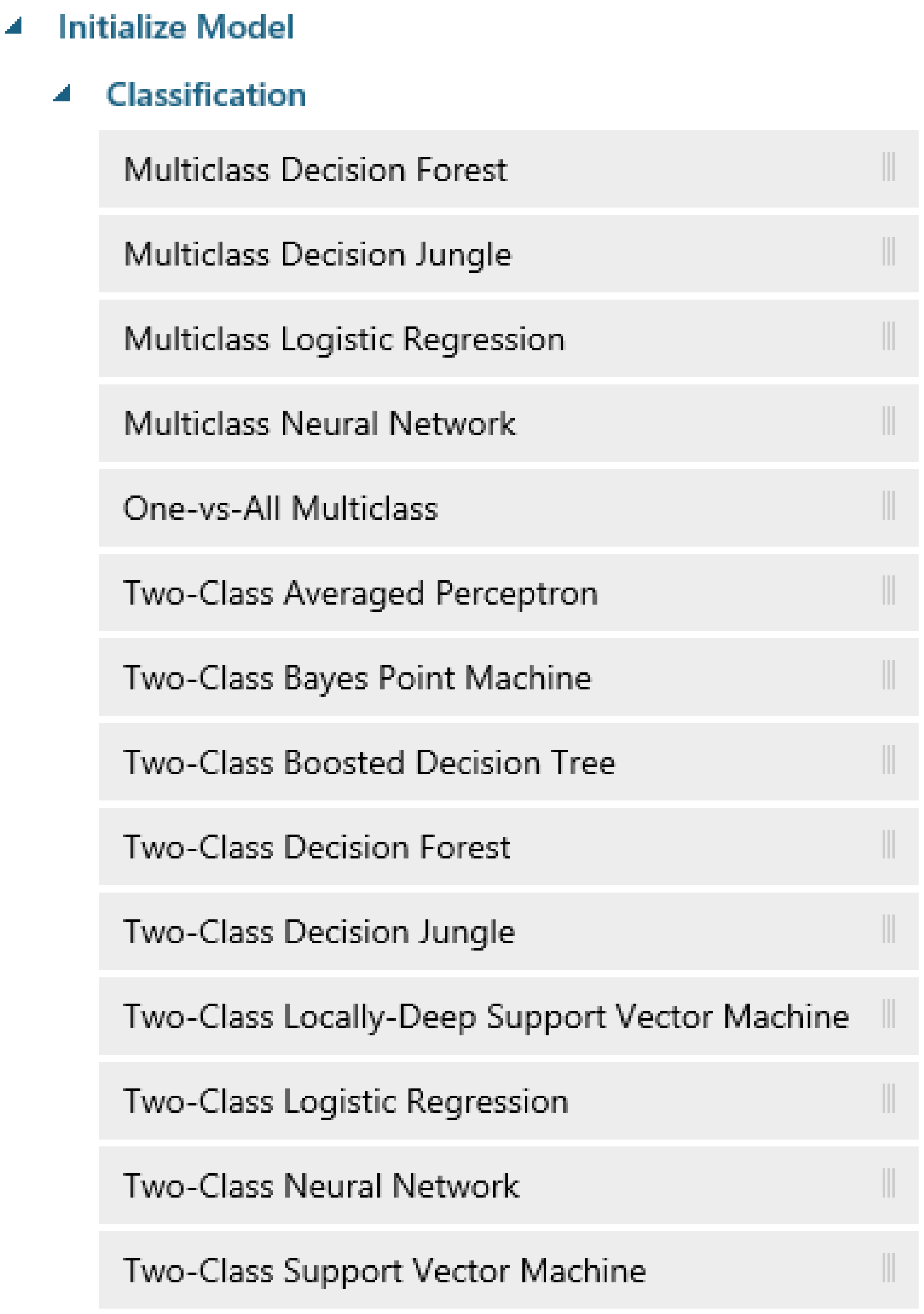
Grouping items based on defined Features



- Machine Learning
  - Initialize Model
    - Clustering
      - K-Means Clustering

CLASSIFICATION

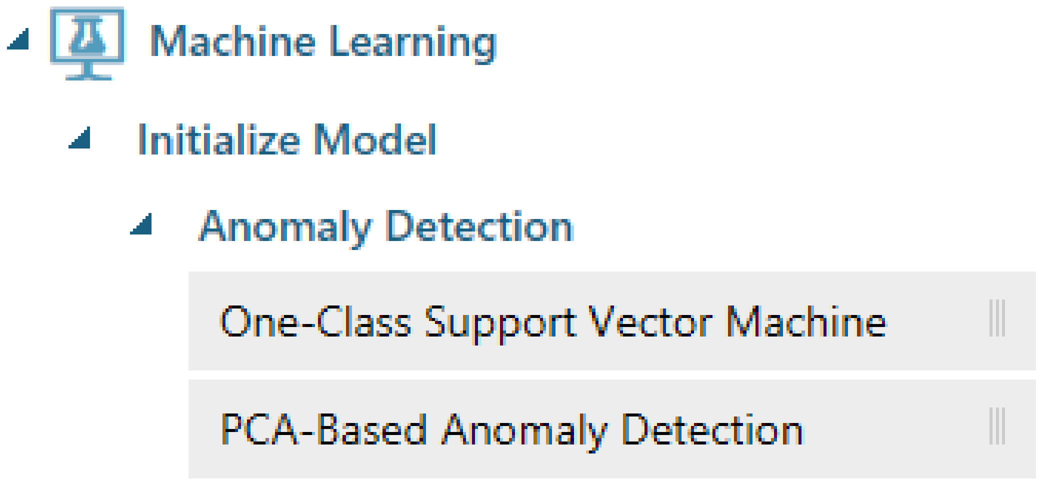
Predicting the class or category for a single instance of data



- Initialize Model
  - Classification
    - Multiclass Decision Forest
    - Multiclass Decision Jungle
    - Multiclass Logistic Regression
    - Multiclass Neural Network
    - One-vs-All Multiclass
    - Two-Class Averaged Perceptron
    - Two-Class Bayes Point Machine
    - Two-Class Boosted Decision Tree
    - Two-Class Decision Forest
    - Two-Class Decision Jungle
    - Two-Class Locally-Deep Support Vector Machine
    - Two-Class Logistic Regression
    - Two-Class Neural Network
    - Two-Class Support Vector Machine

ANOMAY DETECTION

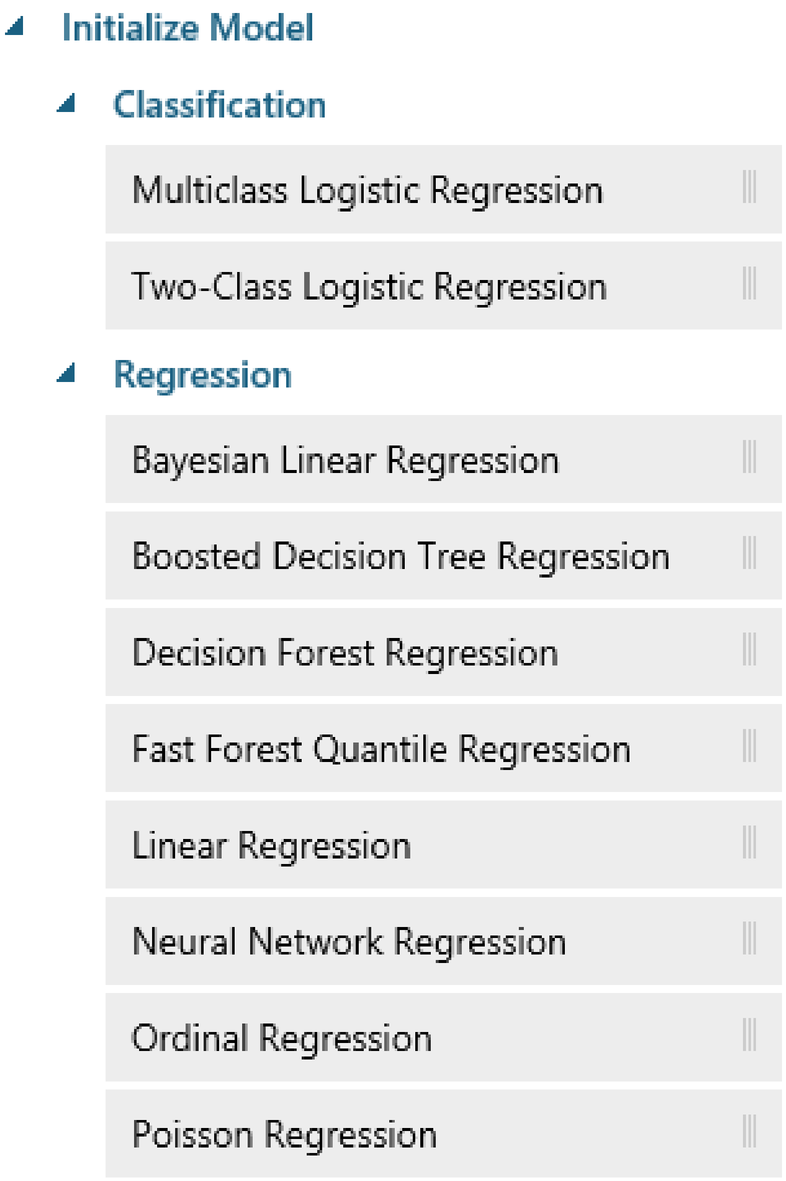
Selecting items based on unusual or suspicious patterns



- Machine Learning
  - Initialize Model
    - Anomaly Detection
      - One-Class Support Vector Machine
      - PCA-Based Anomaly Detection

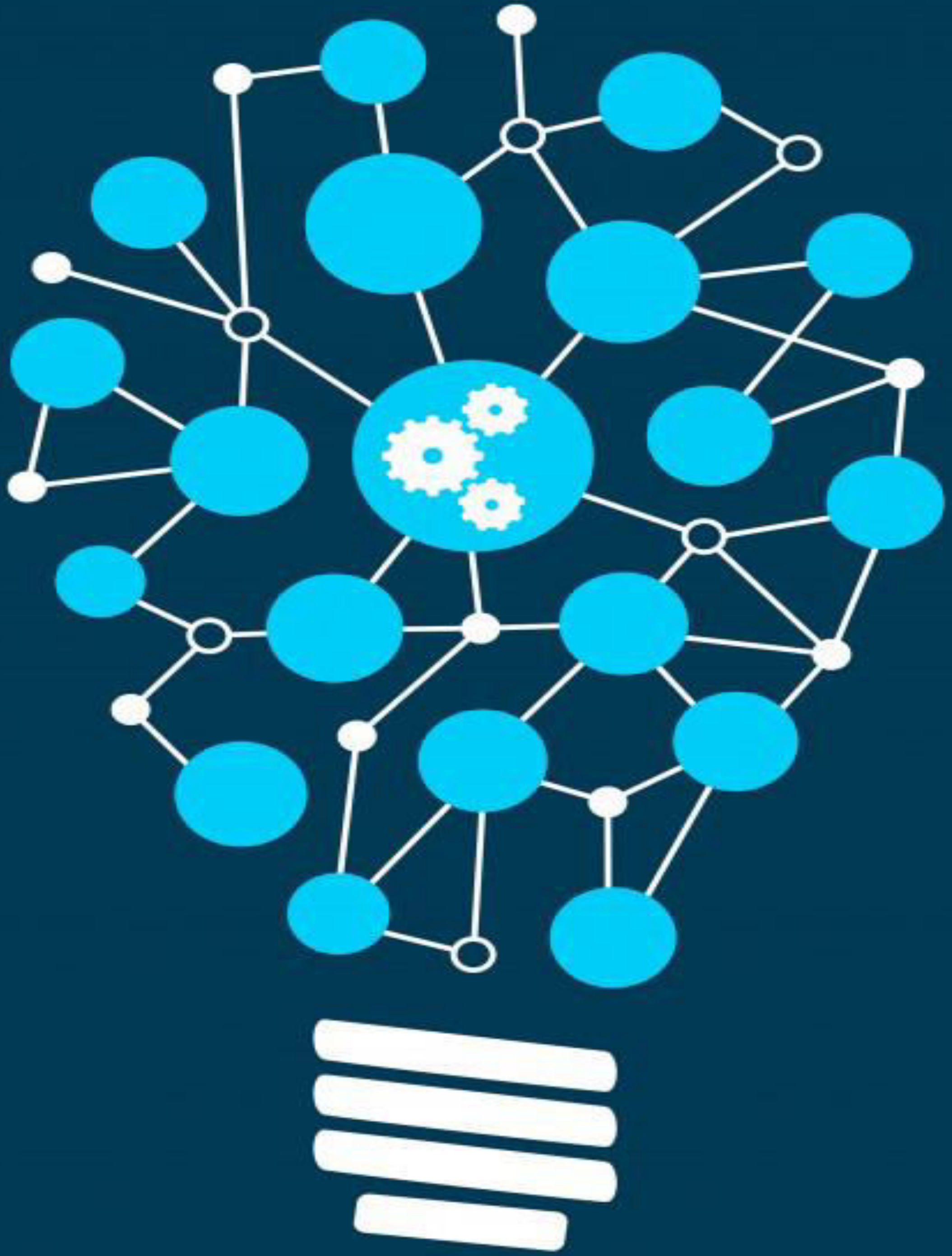
REGRESSION

Predicting the value of a datum given its history



- Initialize Model
  - Classification
    - Multiclass Logistic Regression
    - Two-Class Logistic Regression
  - Regression
    - Bayesian Linear Regression
    - Boosted Decision Tree Regression
    - Decision Forest Regression
    - Fast Forest Quantile Regression
    - Linear Regression
    - Neural Network Regression
    - Ordinal Regression
    - Poisson Regression

MODEL SCORING AND EVALUATION

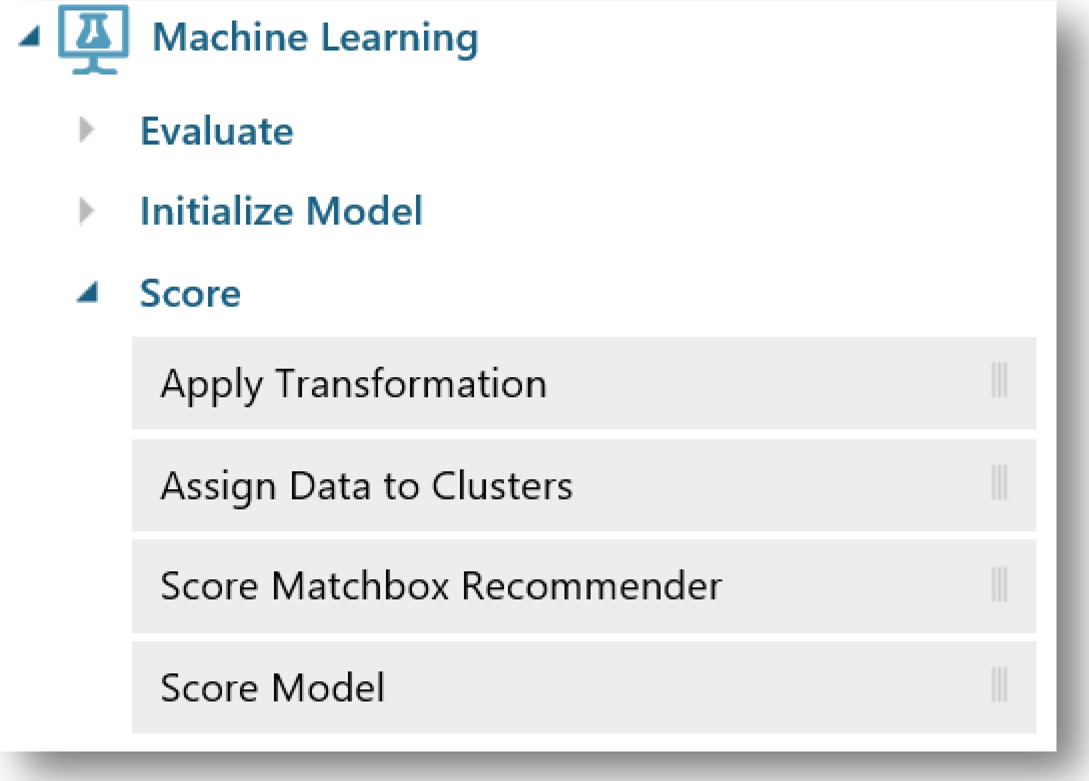


Microsoft

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

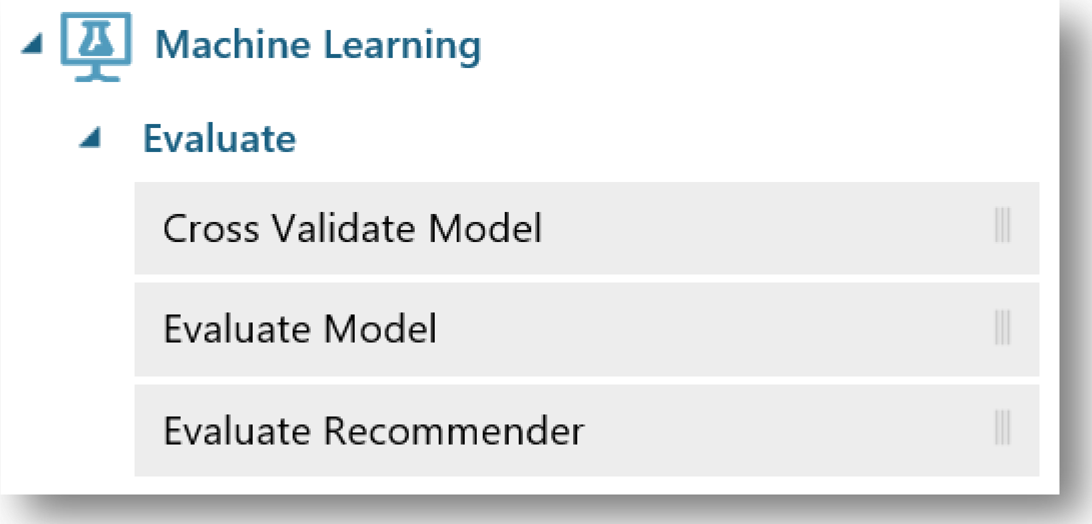
EVALUATING A MODEL

Metrics for Classification Models

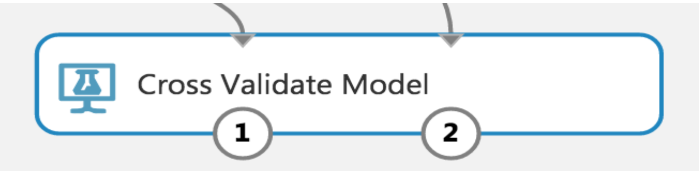
- Accuracy, Recall, Precision, F1-Score
- AUC
- Average Log Loss
- Training Log Loss

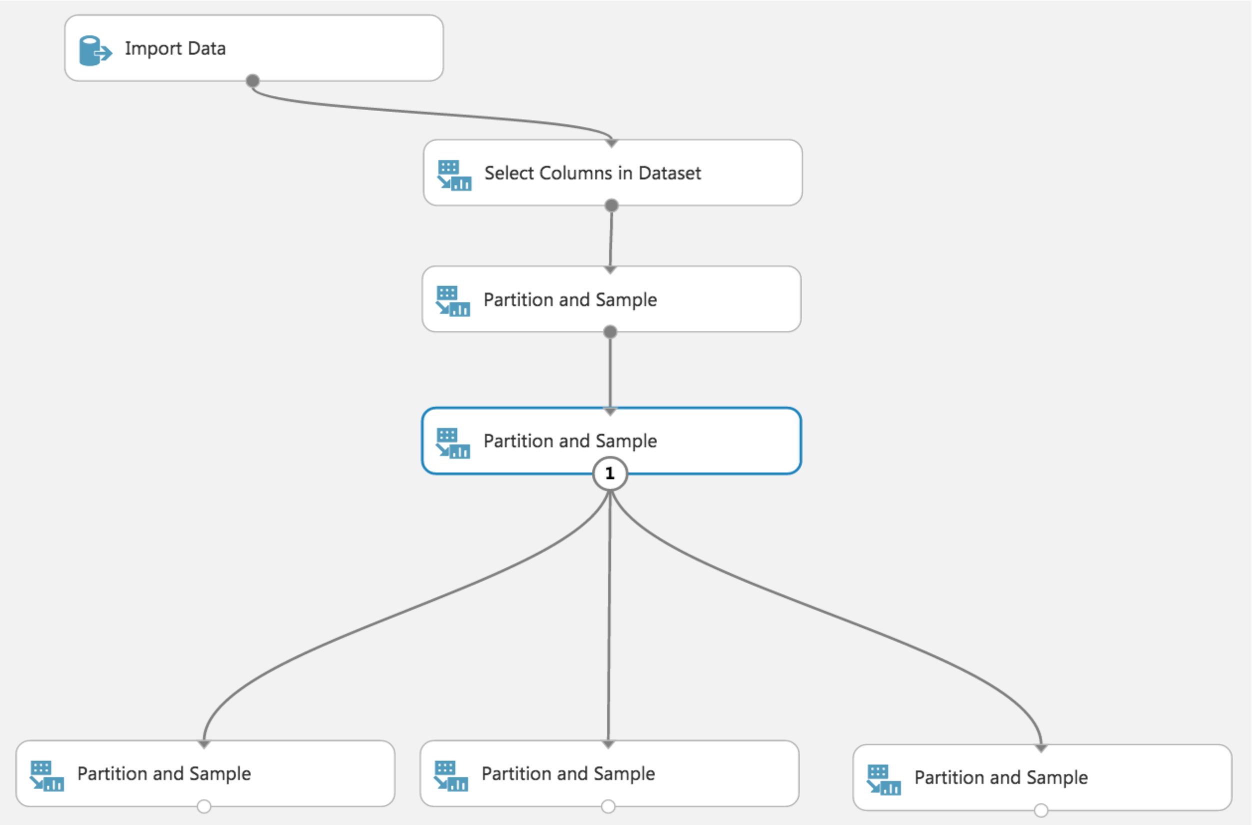
Metrics for Regression Models

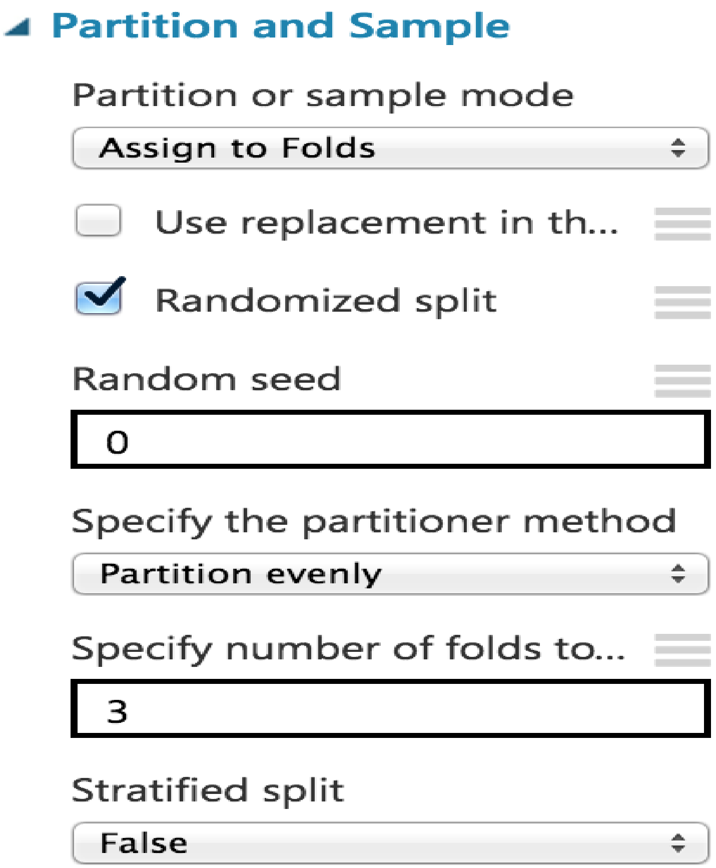
- Mean absolute error (MAE)
- Root mean squared error (RMSE)
- Relative absolute error (RAE)
- Relative squared error (RSE)
- Coefficient of determination



CROSS VALIDATION







◢ **Partition and Sample**

Partition or sample mode

Assign to Folds

☐ Use replacement in th...

☑ Randomized split

Random seed

0

Specify the partitioner method

Partition evenly

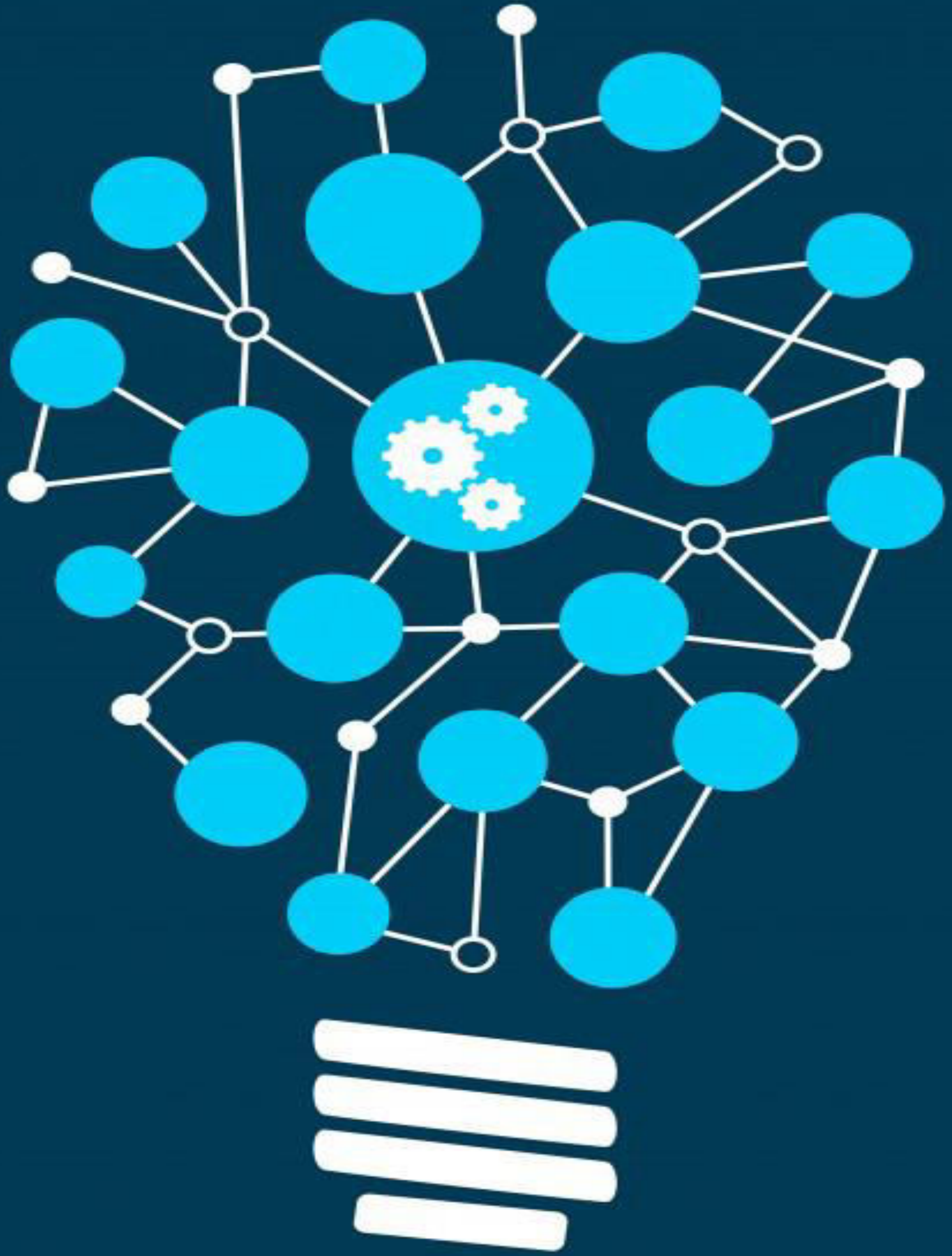
Specify number of folds to...

3

Stratified split

False

CUSTOMIZATION

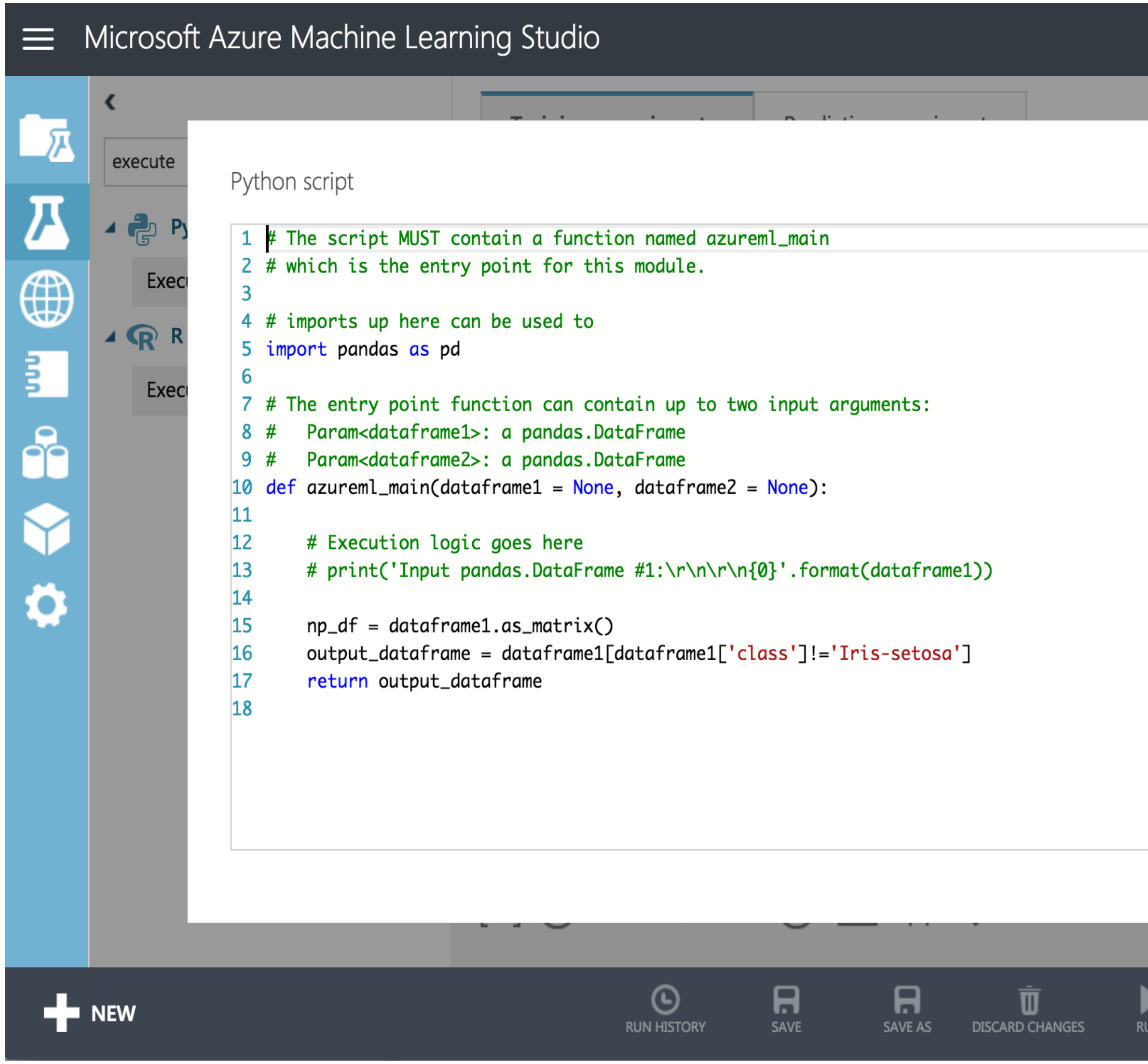


Microsoft

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



```
# The script MUST contain a function named azureml_main
# which is the entry point for this module.

# imports up here can be used to
import pandas as pd

# The entry point function can contain up to two input arguments:
#   Param<dataframe1>: a pandas.DataFrame
#   Param<dataframe2>: a pandas.DataFrame
def azureml_main(dataframe1 = None, dataframe2 = None):

    # Execution logic goes here
    # print('Input pandas.DataFrame #1:\r\n\r\n{0}'.format(dataframe1))

    np_df = dataframe1.as_matrix()
    output_dataframe = dataframe1[dataframe1['class']!='Iris-setosa']
    return output_dataframe

```

HOW TO USE EXECUTE PYTHON SCRIPT

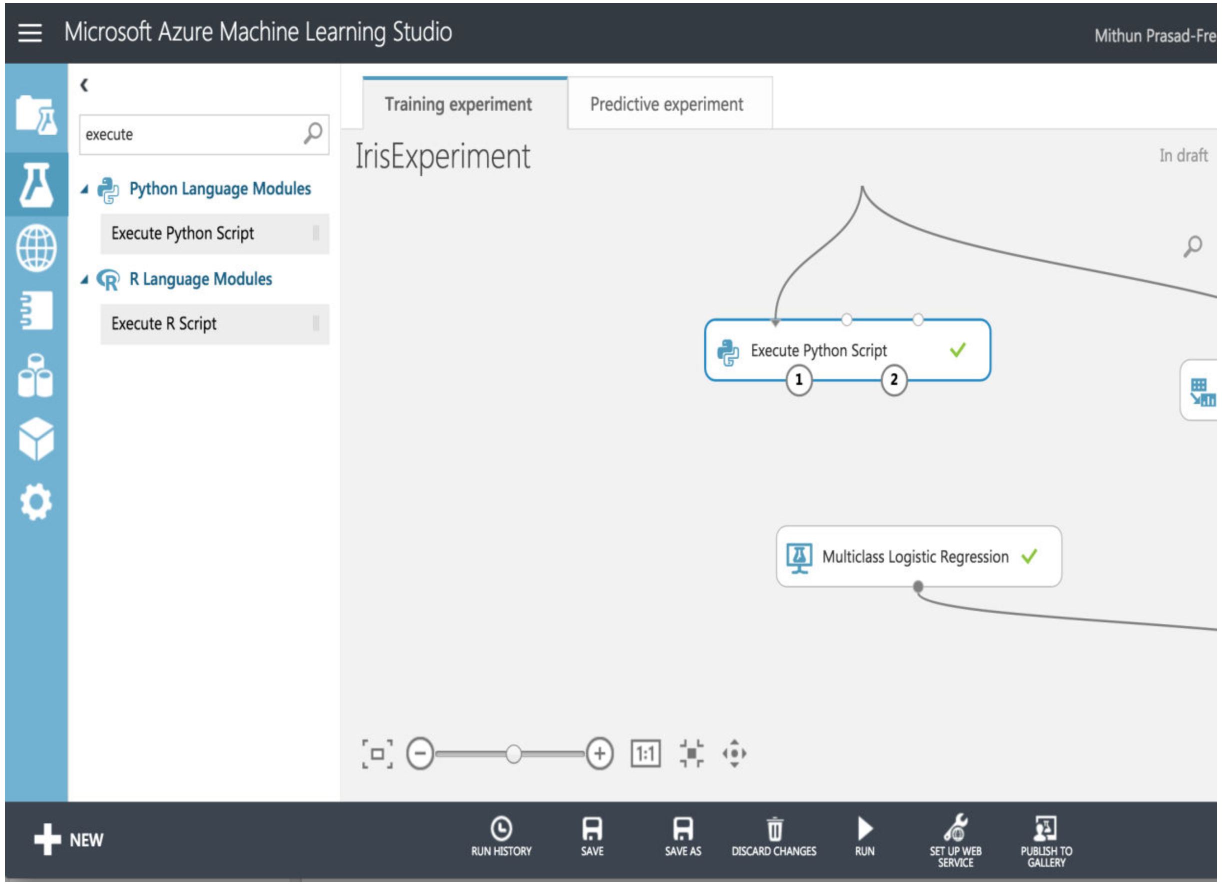
1. Add the **Execute Python Script** module to your experiment.

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 Azure Machine Learning Studio
Mithun Prasad-Fre
execute
Python Language Modules
Execute Python Script
R Language Modules
Execute R Script
Training experiment
Predictive experiment
IrisExperiment
In draft
Execute Python Script
1
2
Multiclass Logistic Regression
1:1
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY