MAKING DATA SCIENTISTS PRODUCTIVE IN AZURE

(AS OF NOVEMBER 2019)

Valdas Maksimavičius

Inspiration for the talk

One thing about Microsoft - they have multiple ways to solve the same problem

Quiz

Microsoft Machine Learning Server

Machine Learning for .NET

Azure Machine Learning

Azure Machine Learning Studio

Azure Databricks

Power BI Auto ML

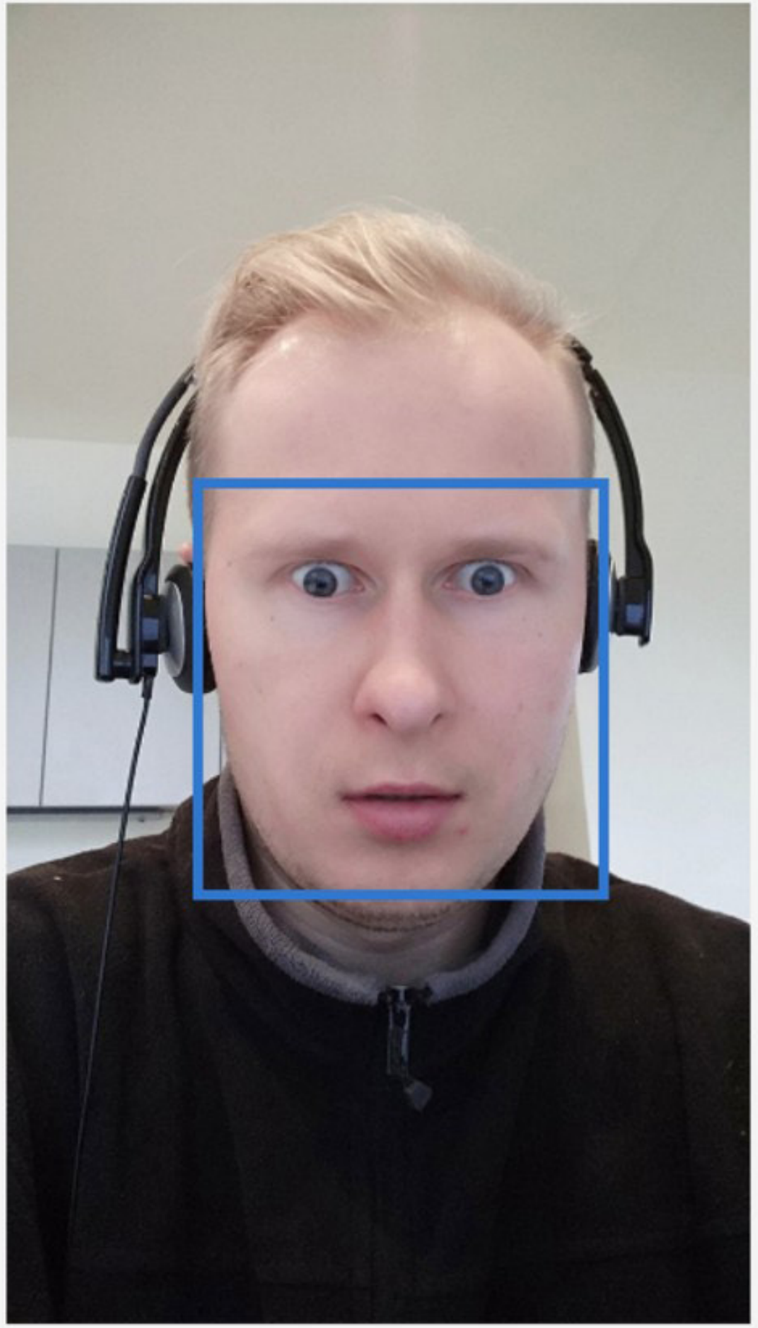
Data Science Virtual Machine

SQL Server Machine Learning Services

Azure Cognitive Services

About me

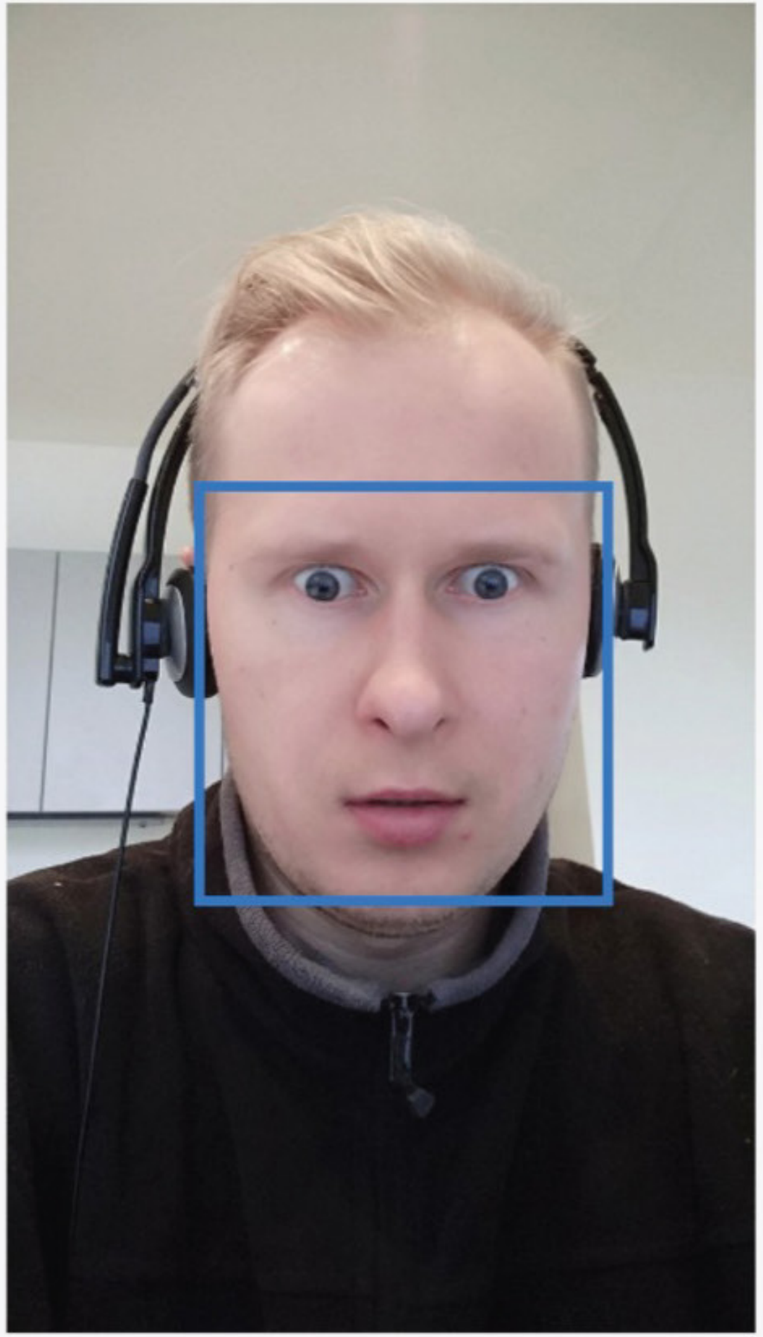
- IT Architect at Cognizant
- Financial, Manufacturing & Logistics industries
- Vilnius Microsoft Data Platform Meetup / Hack4Vilnius Hackathon
- www.valdas.blog



Detection result:

JSON:


```
[
 {
   "faceId": "f783a705-c1c9-4cf1-bb24-064f951f4e52",
   "faceRectangle": {
     "top": 415,
     "left": 163,
     "width": 366,
     "height": 366
   },
   "faceAttributes": {
     "hair": {
       "bald": 0.13,
       "invisible": false,
       "hairColor": [
         {
           "color": "brown",
           "confidence": 0.91
         },
         {
           "color": "red",
           "confidence": 0.9
         },
         {
           "color": "blond",
           "confidence": 0.58
```

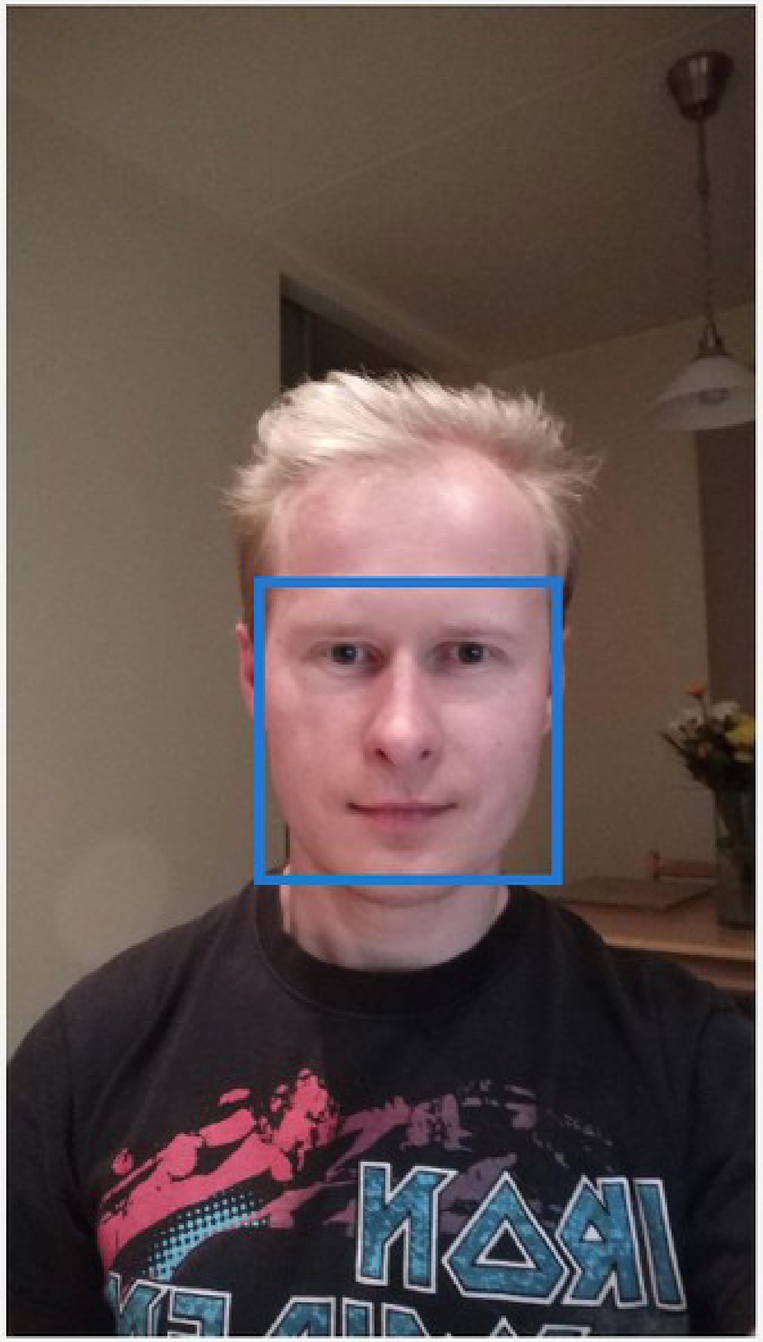


Detection result:

JSON:


```
[
  {
    "faceId": "f783a705-c1c9-4cf1-bb24-064f951f4e52",
    "faceRectangle": {
      "top": 415,
      "left": 163,
      "width": 366,
      "height": 366
    },
    "faceAttributes": {
      "hair": {
        "bald": 0.13,
        "invisible": false,
        "hairColor": [
          {
            "color": "brown",
            "confidence": 0.91
          },
          {
            "color": "red",
            "confidence": 0.9
          },
          {
            "color": "blond",
            "confidence": 0.58
```


"bald": 0.13



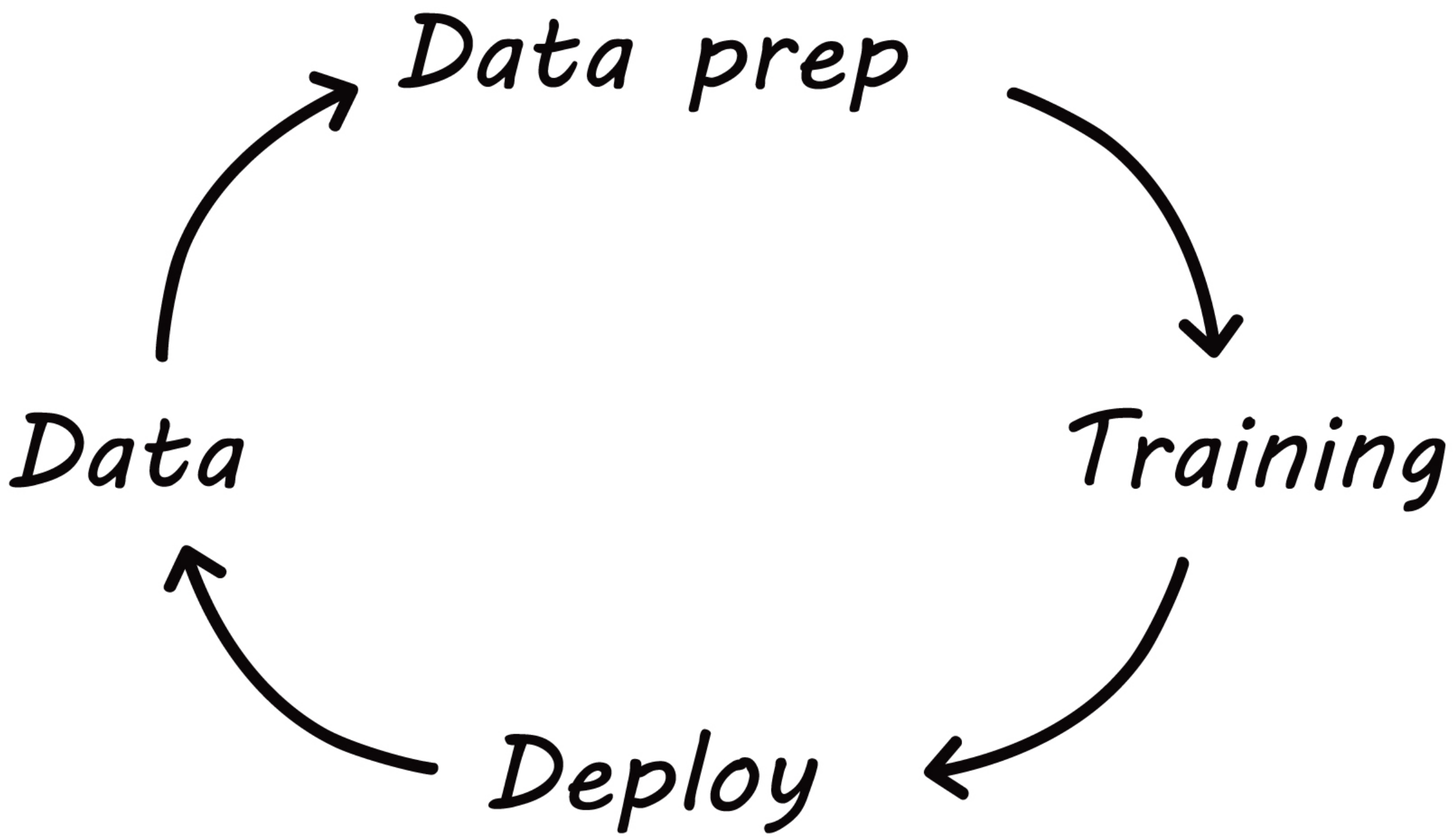
Detection result:

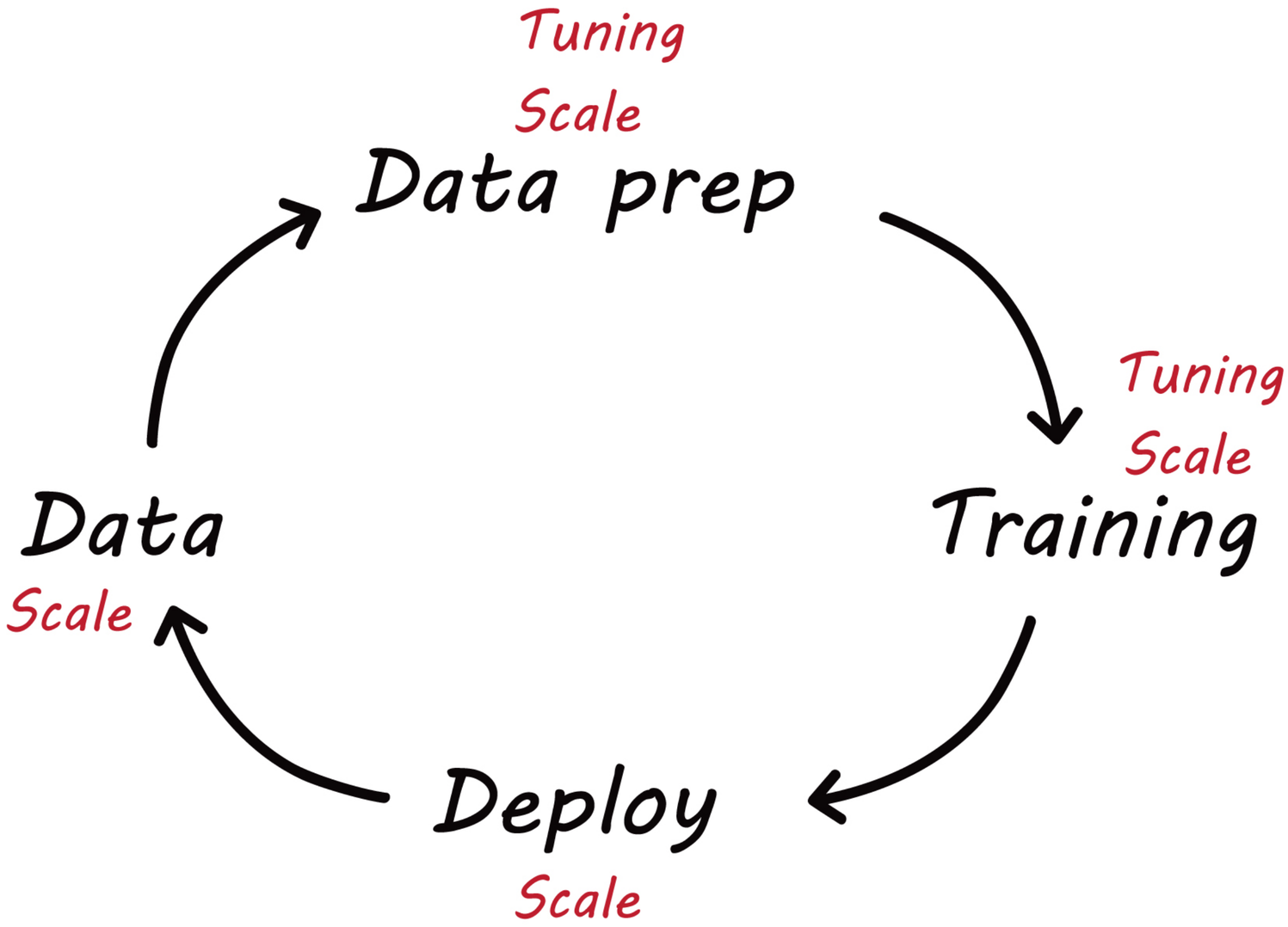
JSON:

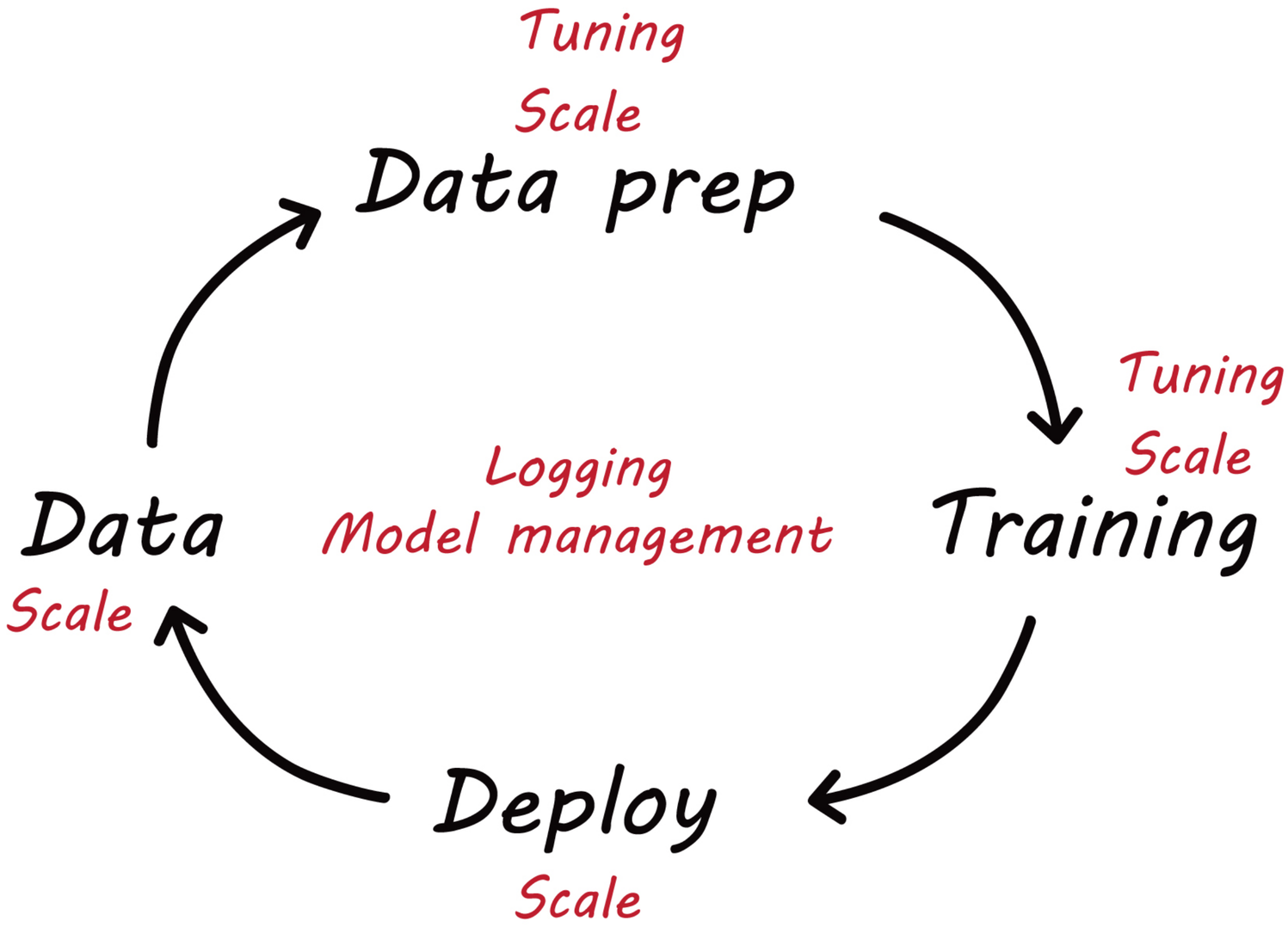

```
[
 {
   "faceId": "588778e2-bf43-48a2-9dbd-181fb26aaf41",
   "faceRectangle": {
     "top": 1644,
     "left": 704,
     "width": 883,
     "height": 883
   },
   "faceAttributes": {
     "hair": {
       "bald": 0.17,
       "invisible": false,
       "hairColor": [
         {
           "color": "blond",
           "confidence": 0.98
         },
         {
           "color": "brown",
           "confidence": 0.73
         },
         {
           "color": "gray",
           "confidence": 0.52
```


"bald": 0.17

So what do you mean by saying “Making Data Scientists Productive in Azure”?


Data prep
Training
Deploy
Data

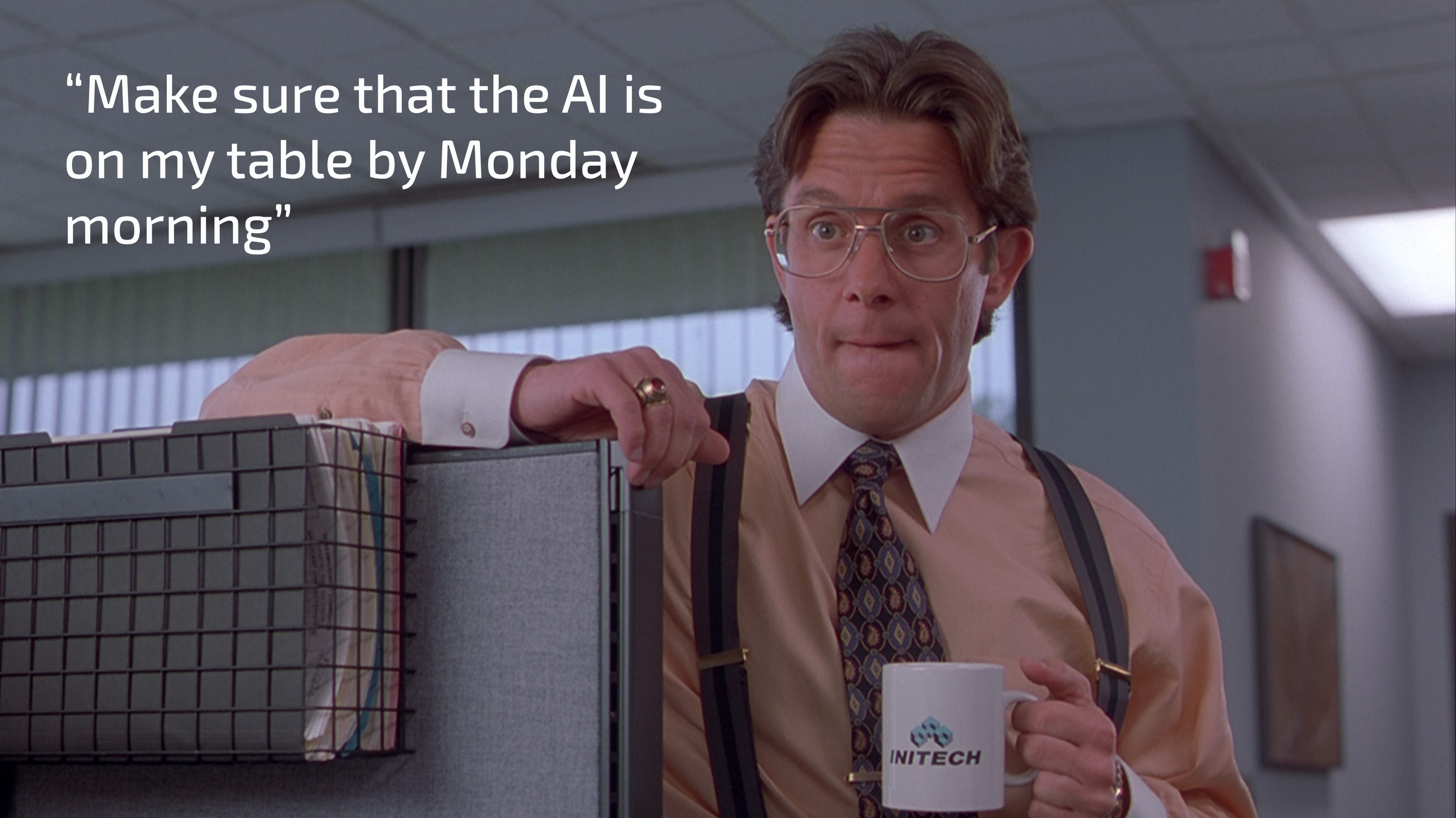

Tuning
Scale
Data prep
Tuning
Scale
Training
Deploy
Scale
Data
Scale


Tuning
Scale
Data prep
Tuning
Scale
Training
Logging
Model management
Data
Scale
Deploy
Scale

6 Data Science stories






“Make sure that the AI is on my table by Monday morning”
INITECH

Tom

- Full stack software developer
- .Net, Node.js, Vue, React

- Analyse call center's call recordings to identify unhappy customers

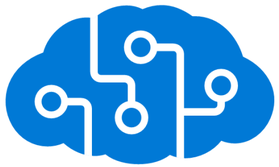
Azure Cognitive Services

What is it?

Azure services with pre-built ML models

What can you do with it?

Add intelligent features to your apps



Azure Cognitive Services - Overview

- Vision (e.g. face / scene / object recognition, video analysis)
- Speech (e.g. speaker recognition, speech-to-text)
- Language (e.g. translations, phrase extraction, QnA maker)
- Decision (e.g. content moderation, anomaly detection)
- Search (e.g. Bing search)

Azure Cognitive Services - Summary

Key benefits:

- Minimal development effort
- Easy integration via HTTP REST
- Built-in support with other Azure services
- Containers support
- Azure Virtual Network support

Azure Cognitive Services - Summary

Key benefits:

- Minimal development effort
- Easy integration via HTTP REST
- Built-in support with other Azure services
- Containers support
- Azure Virtual Network support

Considerations:

- Limited customization allowed
- Limited support for Non-English languages

ML.NET

What is it?

An open source and cross-platform ML framework

What can you do with it?

Create custom ML models using C# or F#
without leaving the .NET ecosystem

C# F#

```
//Step 1. Create a ML Context
var ctx = new MLContext();

//Step 2. Read in the input data for model training
IDataView dataReader = ctx.Data
    .LoadFromTextFile<MyInput>(dataPath, hasHeader: true);

//Step 3. Build your estimator
IEstimator<ITransformer> est = ctx.Transforms.Text
    .FeaturizeText("Features", nameof(SentimentIssue.Text))
    .Append(ctx.BinaryClassification.Trainers
        .LbfgsLogisticRegression("Label", "Features"));

//Step 4. Train your Model
ITransformer trainedModel = est.Fit(dataReader);

//Step 5. Make predictions using your model
var predictionEngine = ctx.Model
    .CreatePredictionEngine<MyInput, MyOutput>(trainedModel);

var sampleStatement = new MyInput { Text = "This is a horrible movie" };

var prediction = predictionEngine.Predict(sampleStatement);
```

ML.NET - Summary

Key benefits:

- High performance
- AutoML functionality
- Leverage TensorFlow or ONNX
- Expose a model via an ASP.NET Core Web API
- Integrate with Spark via .NET for Apache Spark (preview)
- Use ML.NET in Jupyter Notebooks (preview)

Considerations:

- Limited support for popular ML libraries (e.g. Scikit-learn, NumPy)

Mark

- Business Analyst
- Basics of statistical analysis

- Create a sales lead list

Power BI Auto ML

What is it?
Auto Machine Learning component built into Power BI

What can you do with it?
Build ML models without any code in your reporting tool

Select a field to predict — Choose a model — Select data to study — Name and train

Choose a model

Classification

Binary Prediction

Predict whether or not an outcome will be achieved.

General Classification

Distinguish between three or more outcomes.

Regression

Regression

Estimate a numeric value

Back | Next | Cancel

Select a field to predict — Choose a model — Select data to study — Name and train

Name and train your model

Model name

Purchase Intent Prediction

Description

(Optional)

Training time

The longer you train your model, the more accurate the results. Train for a short time if you just want to make sure you've selected the right data. Keep in mind, this won't result in the best model.

5 minutes — 360 minutes | 120 minutes

Training details

Model type: Binary Prediction

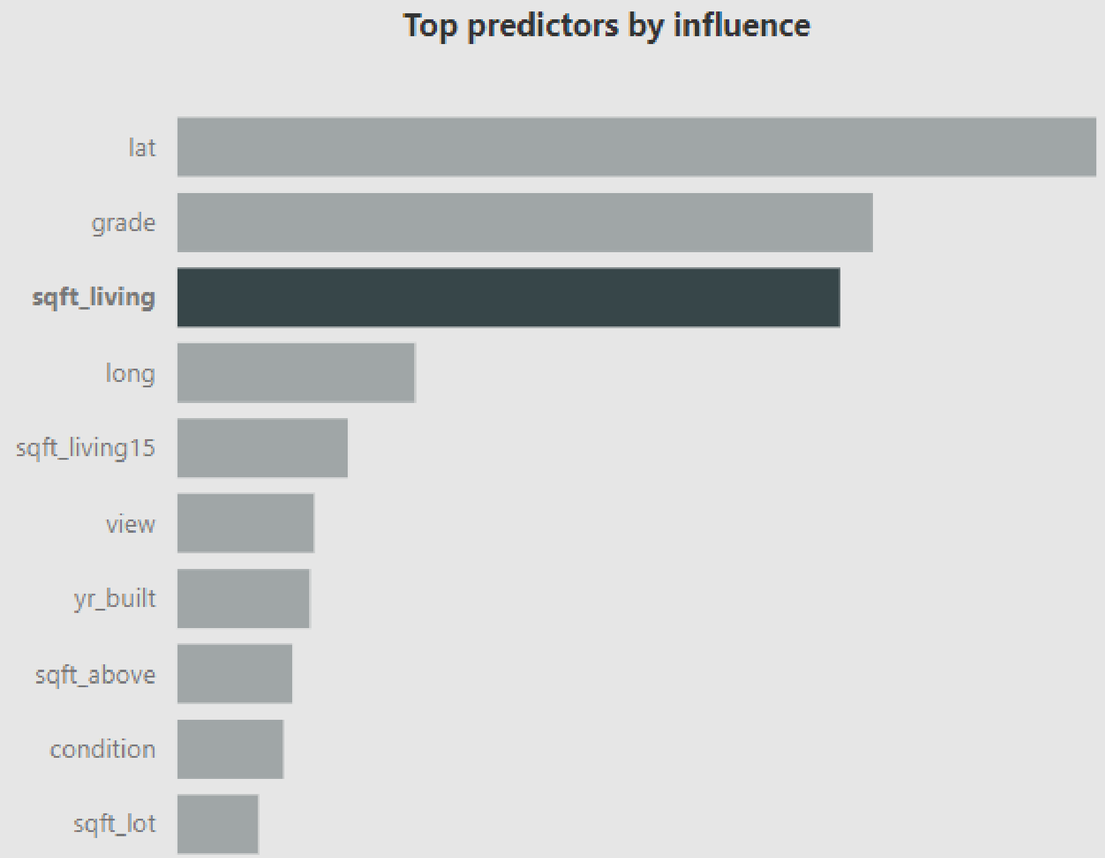
Base entity: Online visitors

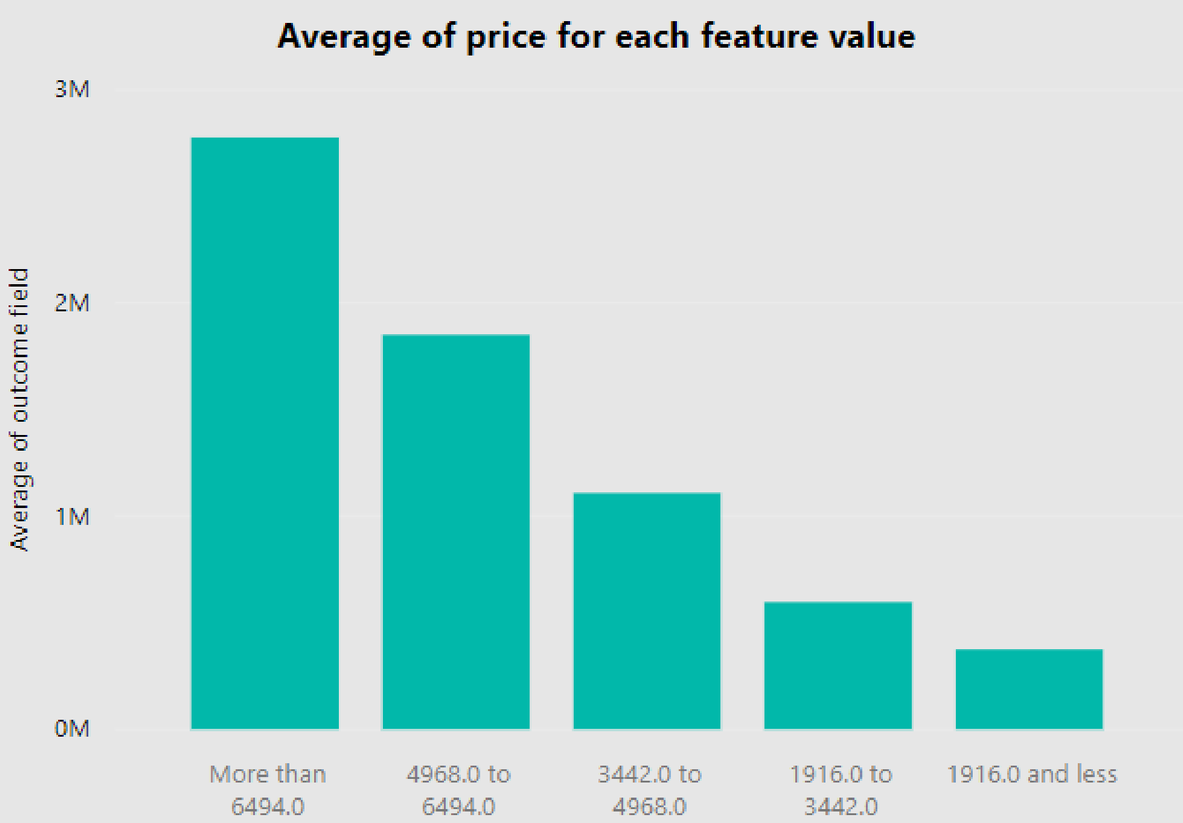
Historical outcome: Revenue

Input fields: 18

Training data: The model will take a statistically significant sample from Online visitors and train on approximately 80% of the sample data. It will then test its algorithm on the remaining sample data and report on its Prediction accuracy.

Back | Save | Save and train | Cancel


Top predictors by influence
lat
grade
sqft_living
long
sqft_living15
view
yr_built
sqft_above
condition
sqft_lot


Average of price for each feature value
Average of outcome field
3M
2M
1M
0M
More than 6494.0
4968.0 to 6494.0
3442.0 to 4968.0
1916.0 to 3442.0
1916.0 and less

Power BI Auto ML - Summary

Key benefits:

- Use Power BI dataflows to load data, transform it and build models on top of it
- Deploy models as services via Azure ML
- Get top predictors during training and explanations for each prediction

Power BI Auto ML - Summary

Key benefits:

- Use Power BI dataflows to load data, transform it and build models on top of it
- Deploy models as services via Azure ML
- Get top predictors during training and explanations for each prediction

Considerations:

- Limited selection of algorithms (binary prediction, general classification, regression)
- Paid Pro or Premium license needed

Azure Machine Learning Studio

What is it?

Drag-and-drop visual interface for ML

What can you do with it?

Build, experiment, and deploy models using pre-configured algorithms

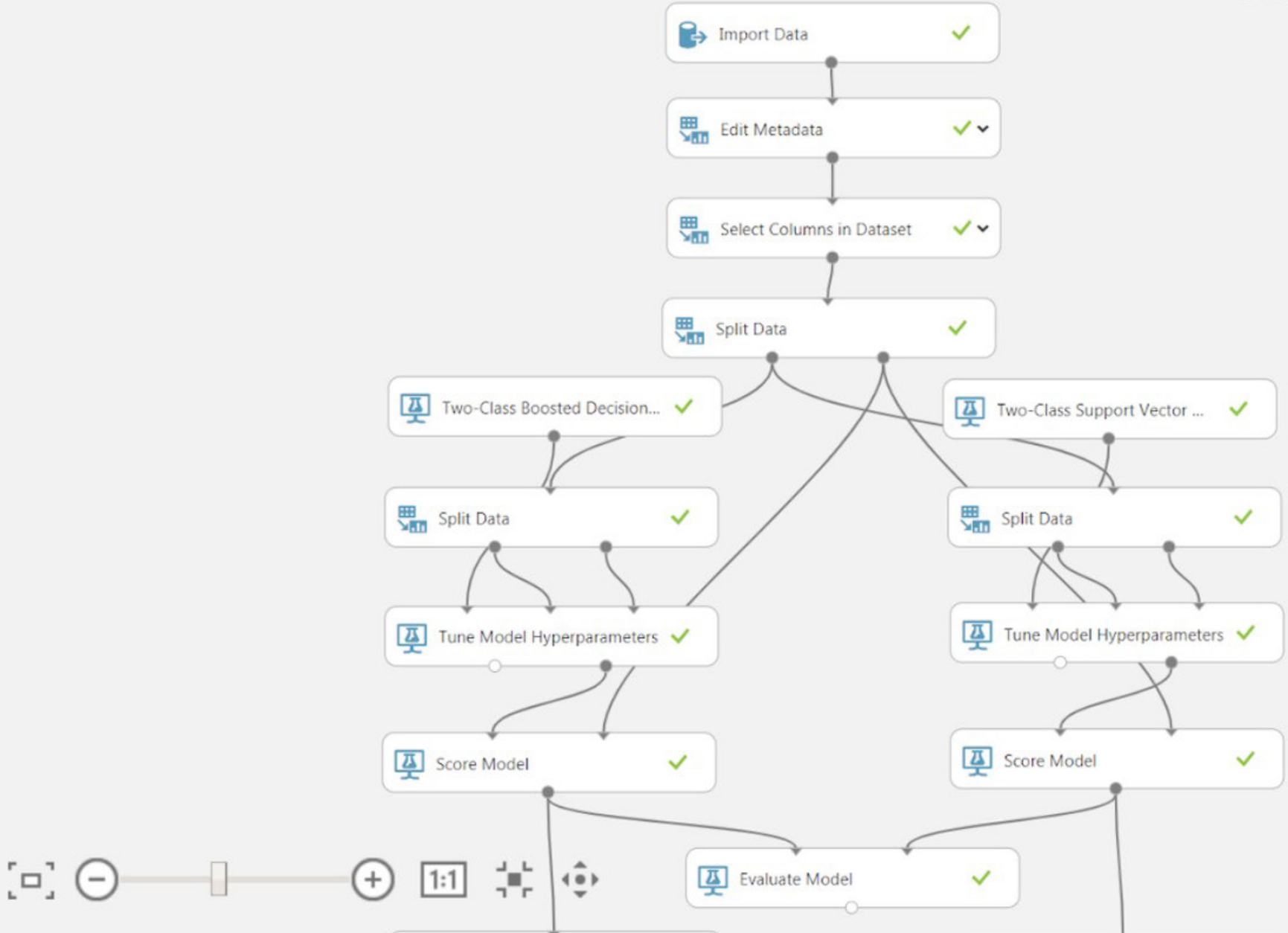
Binary Classification: Direct marketing

Finished running

Draft saved at 12:38:31

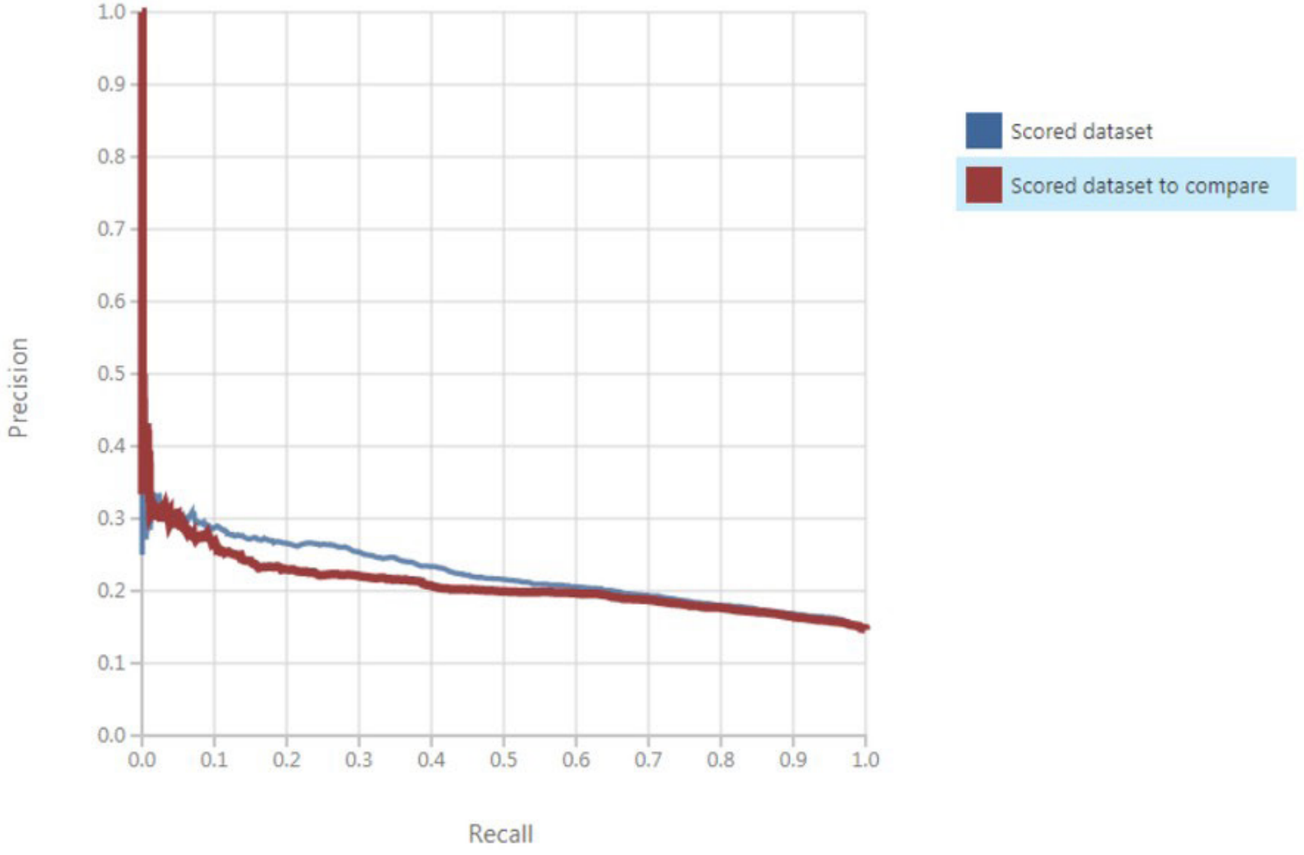
Search experiment items

- 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
- Time Series
- Web Service
- Deprecated



Binary Classification: Direct marketing › Evaluate Model › Evaluation results

ROC PRECISION/RECALL LIFT



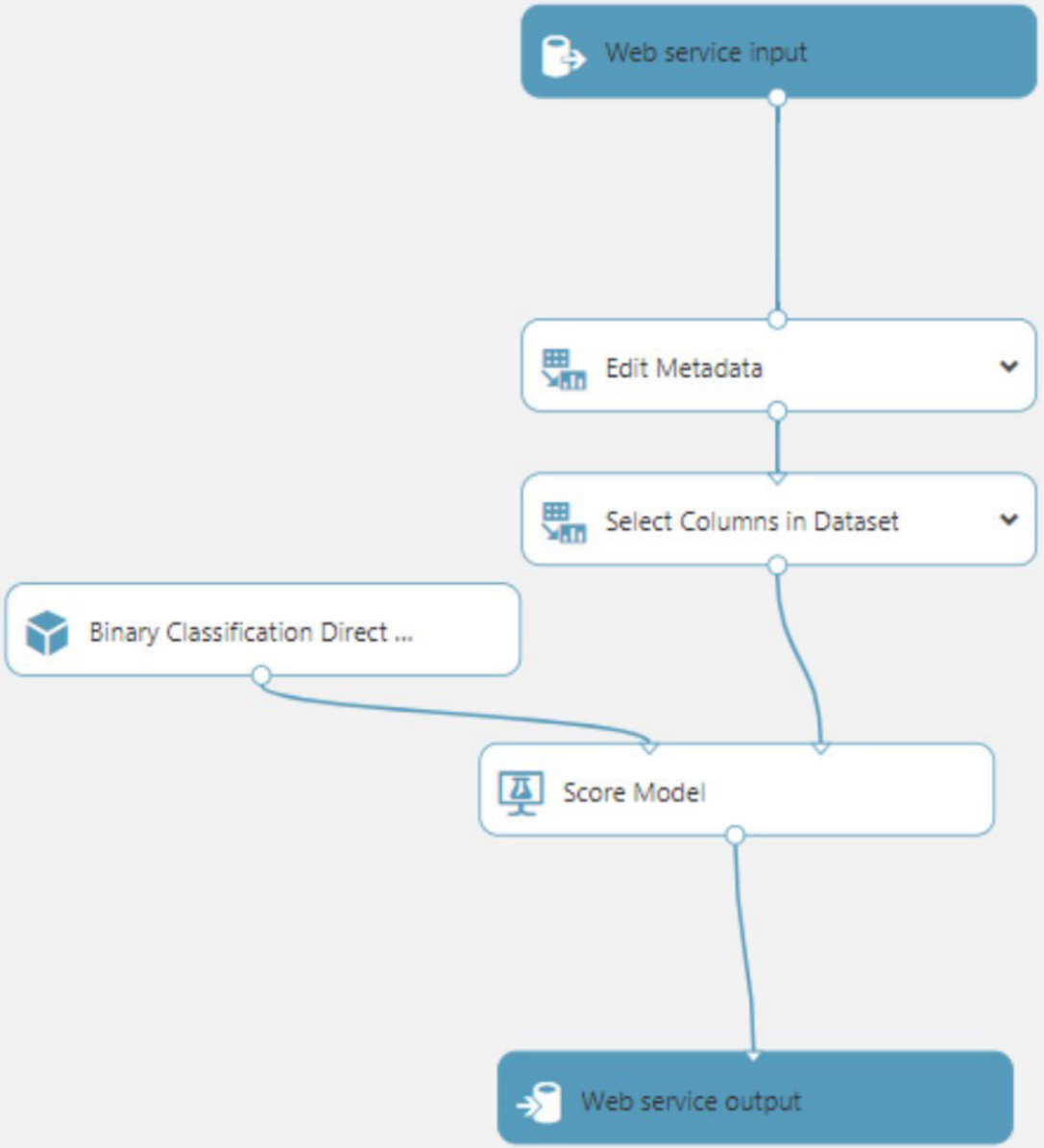
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
0	2818	0.853	1.000	0.5	0.620

False Positive	True Negative	Recall	F1 Score
0	16382	0.000	0.000

Positive Label	Negative Label
1	0

Deploy as web services

- Batch execution
- Request / Response



Azure Machine Learning Studio - Summary

Key benefits:

- Interactive visual interface
- Built-in Jupyter Notebooks for data exploration
- Direct deployment of trained models as web services
- Built-in support for other Azure services

Azure Machine Learning Studio - Summary

Key benefits:

- Interactive visual interface
- Built-in Jupyter Notebooks for data exploration
- Direct deployment of trained models as web services
- Built-in support for other Azure services

Considerations:

- Online only
- Limited scalability (the maximum size of a training dataset is 10 GB)
- Limited number of supported input and output connectors
- Limited support for custom Python/R code

Lucy

- Machine Learning Engineer
- Python, Scikit-learn, Keras, TensorFlow

- Estimate damage (repair cost) in auto insurance

Azure Machine Learning ~~Studio~~

What is it?

Managed cloud service for ML

What can you do with it?

Train, deploy and manage models in Azure

Azure Machine Learning - Overview

- Azure ML SDK
- Data preparation
- Compute targets
- Experiment tracking
- Deployment targets

Azure Machine Learning - Compute Targets

- Your local computer

- Linux VM in Azure
- Azure ML Compute
- Azure Databricks
- Azure Container Instance
- Apache Spark for HDInsight

Azure Machine Learning - Compute Targets

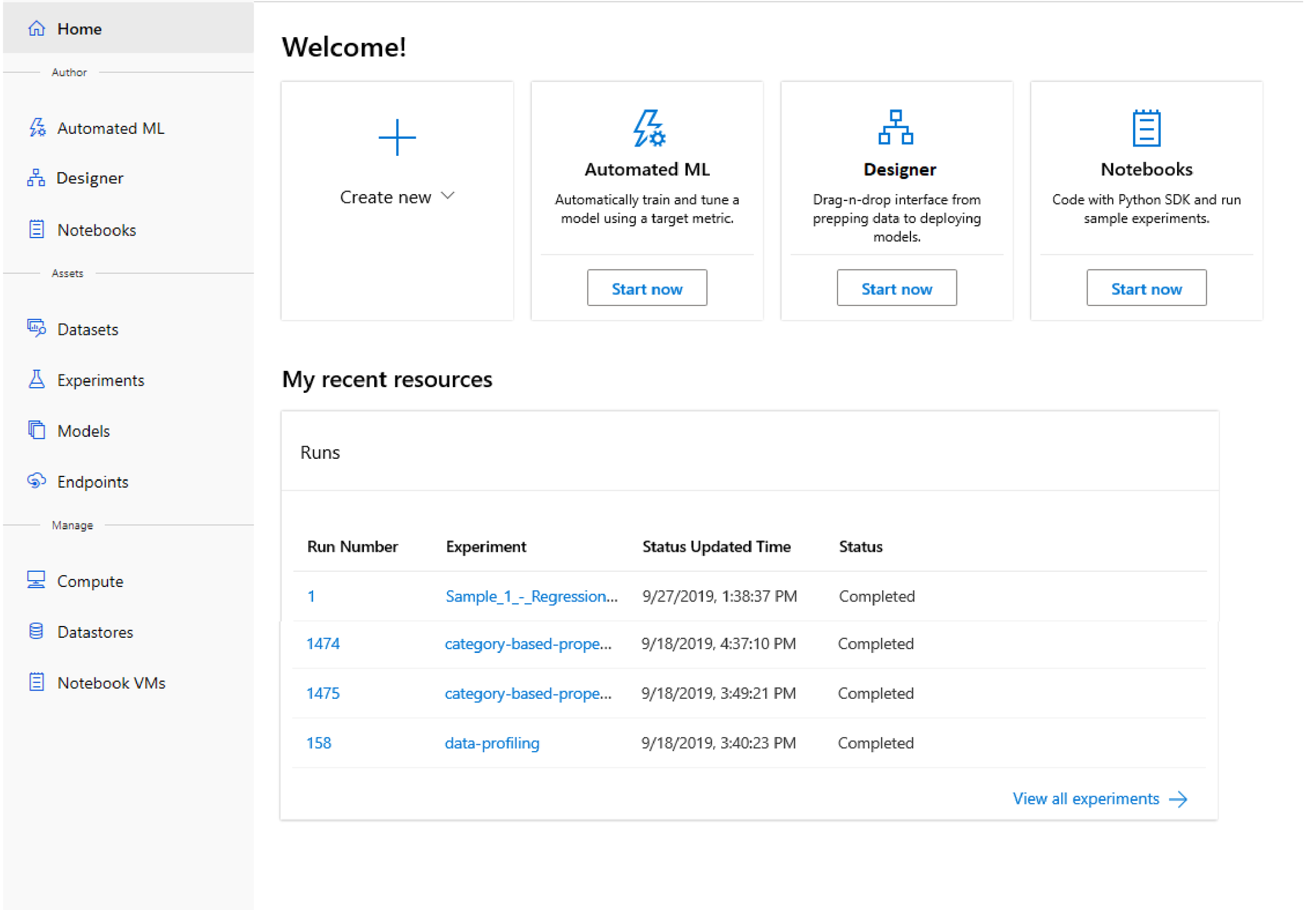
```
from azureml.core.compute import ComputeTarget, BatchAiCompute
from azureml.core import Workspace
from azureml.train.dnn import PyTorch

# Get Azure secrets
ws = Workspace.from_config()
# GPU-based Batch AI
compute_config = BatchAiCompute.provisioning_configuration(vm_size="STANDARD_NC6",
                                                           autoscale_enabled=True,
                                                           cluster_min_nodes=0,
                                                           cluster_max_nodes=4)
# Create the cluster in Azure
compute_target = ComputeTarget.create(ws, compute_config)
# Create experiment in Azure ML workspace
experiment = Experiment(ws, "My pytorch experiment")
# Set PyTorch estimator
pt_estimator = PyTorch(source_directory='./my-training-files',
                       compute_target=compute_target,
                       entry_script='train.py',
                       use_gpu=True, ...)
# Submit training scripts
experiment.submit(pt_estimator)

```

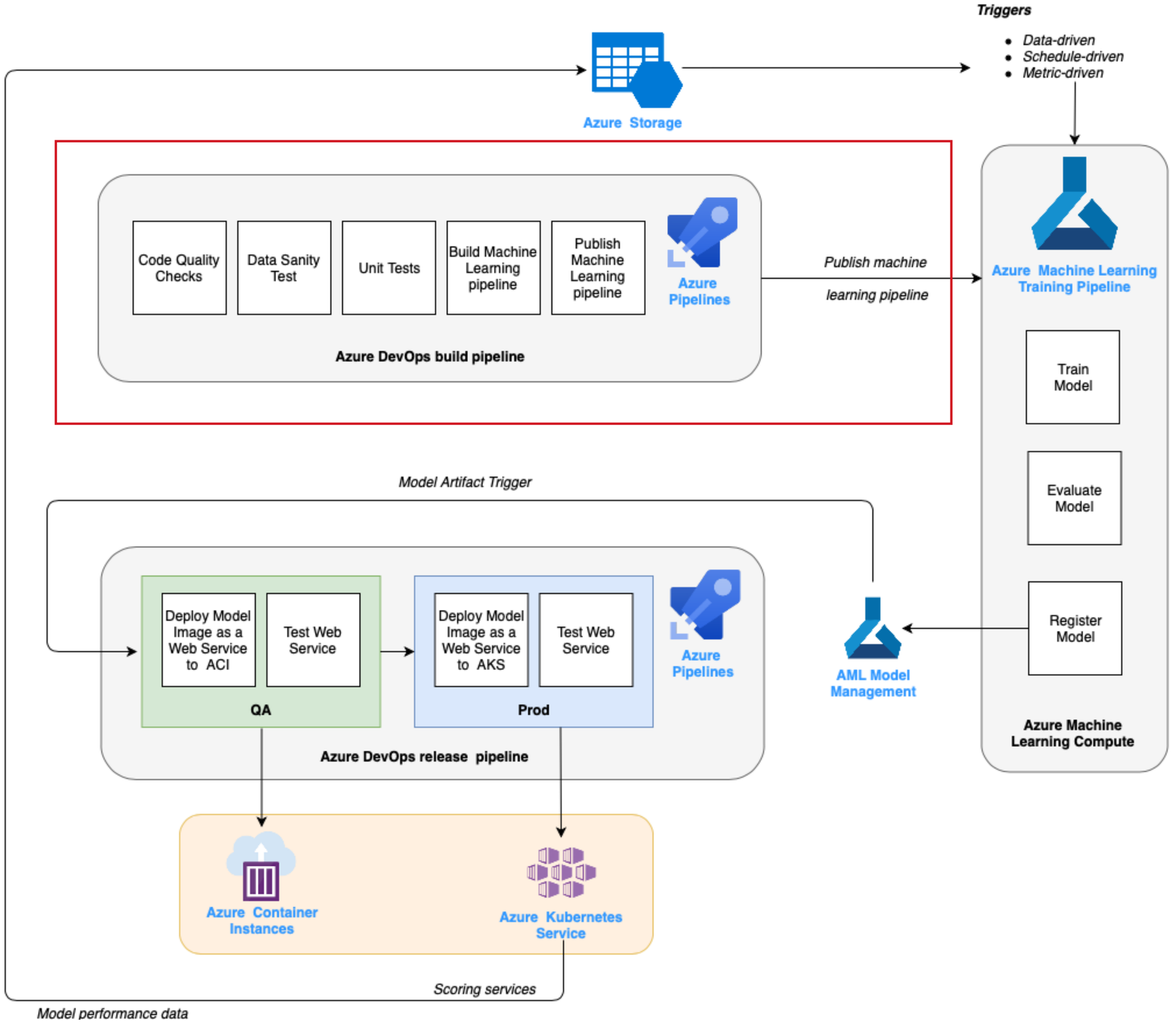
Azure Machine Learning - Experiment Tracking

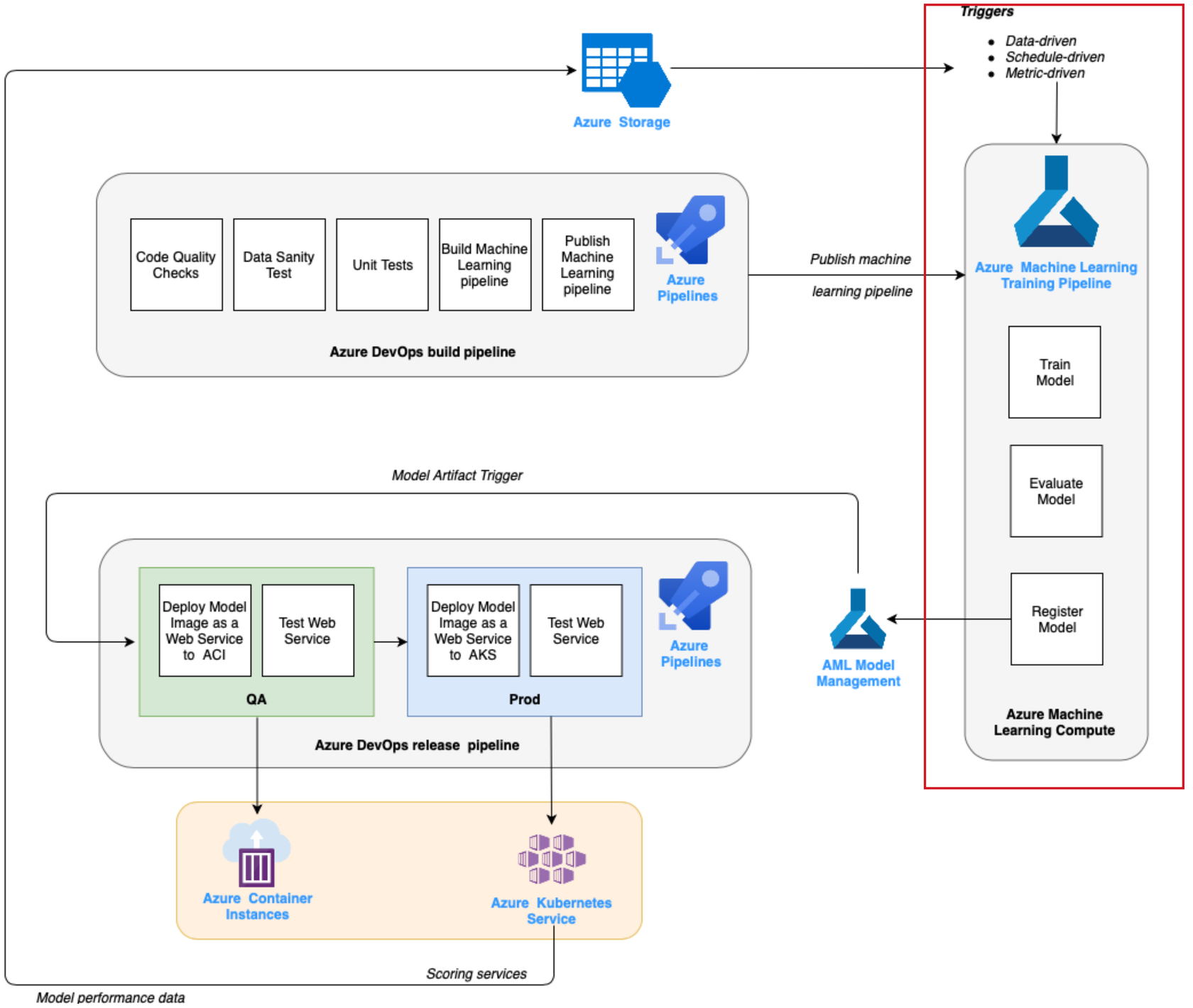
```
1
2   ws = Workspace.from_config()
3   experiment_name = 'train-in-notebook'
4   experiment = Experiment(workspace = ws, name = experiment_name)
5   # Send metrics to Azure
6   run = experiment.start_logging()
7   run.log("accuracy", 0.95)
8   run.log_list("accuracies", [0.6, 0.7, 0.87])
9   run.log_image("ROC", plt)
10  run.upload_file("best_model.pkl", "./model.pkl")
11  run.complete()
12
```

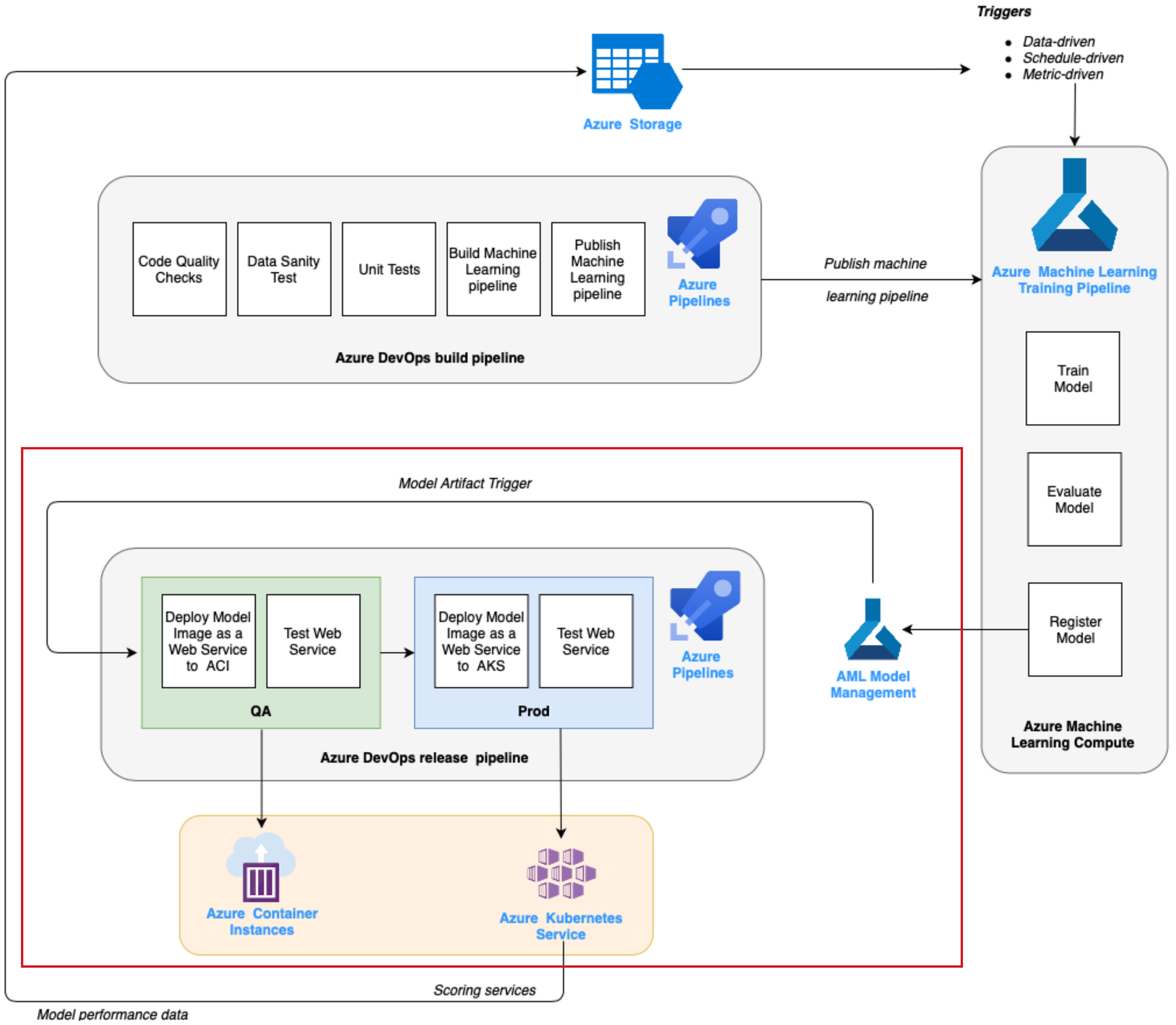
Azure Machine Learning - Deployment Targets

- Azure ML Compute
- Azure Kubernetes Service / Container Instances
- Docker image
- App Service
- Azure Functions
- IoT Edge


Home
Author
Automated ML
Designer
Notebooks
Assets
Datasets
Experiments
Models
Endpoints
Manage
Compute
Datastores
Notebook VMs
Welcome!
Create new
Automated ML
Automatically train and tune a model using a target metric.
Start now
Designer
Drag-n-drop interface from prepping data to deploying models.
Start now
Notebooks
Code with Python SDK and run sample experiments.
Start now
My recent resources
Runs
Run Number
Experiment
Status Updated Time
Status
1
Sample_1_-_Regression...
9/27/2019, 1:38:37 PM
Completed
1474
category-based-prope...
9/18/2019, 4:37:10 PM
Completed
1475
category-based-prope...
9/18/2019, 3:49:21 PM
Completed
158
data-profiling
9/18/2019, 3:40:23 PM
Completed
View all experiments


Triggers
Data-driven
Schedule-driven
Metric-driven
Azure Storage
Code Quality Checks
Data Sanity Test
Unit Tests
Build Machine Learning pipeline
Publish Machine Learning pipeline
Azure Pipelines
Azure DevOps build pipeline
Publish machine learning pipeline
Azure Machine Learning Training Pipeline
Train Model
Evaluate Model
Register Model
Azure Machine Learning Compute
Model Artifact Trigger
Deploy Model Image as a Web Service to ACI
Test Web Service
QA
Deploy Model Image as a Web Service to AKS
Test Web Service
Prod
Azure Pipelines
Azure DevOps release pipeline
AML Model Management
Azure Container Instances
Azure Kubernetes Service
Scoring services
Model performance data


Triggers
Data-driven
Schedule-driven
Metric-driven
Azure Storage
Code Quality Checks
Data Sanity Test
Unit Tests
Build Machine Learning pipeline
Publish Machine Learning pipeline
Azure Pipelines
Azure DevOps build pipeline
Publish machine learning pipeline
Azure Machine Learning Training Pipeline
Train Model
Evaluate Model
Register Model
Azure Machine Learning Compute
Model Artifact Trigger
Deploy Model Image as a Web Service to ACI
Test Web Service
QA
Deploy Model Image as a Web Service to AKS
Test Web Service
Prod
Azure Pipelines
Azure DevOps release pipeline
AML Model Management
Azure Container Instances
Azure Kubernetes Service
Scoring services
Model performance data


Triggers
Data-driven
Schedule-driven
Metric-driven
Azure Storage
Code Quality Checks
Data Sanity Test
Unit Tests
Build Machine Learning pipeline
Publish Machine Learning pipeline
Azure Pipelines
Azure DevOps build pipeline
Publish machine learning pipeline
Azure Machine Learning Training Pipeline
Train Model
Evaluate Model
Register Model
Azure Machine Learning Compute
Model Artifact Trigger
Deploy Model Image as a Web Service to ACI
Test Web Service
QA
Deploy Model Image as a Web Service to AKS
Test Web Service
Prod
Azure Pipelines
Azure DevOps release pipeline
AML Model Management
Azure Container Instances
Azure Kubernetes Service
Scoring services
Model performance data

Azure Machine Learning - Summary

Key benefits:

- Central management of scripts and run history
- Run model training scripts locally (offline), and then scale out to the cloud
- Management and deployment of models to the cloud or edge devices
- Integration with Azure Dev Ops
- Added support for R (preview)

Azure Machine Learning - Summary

Key benefits:

- Central management of scripts and run history
- Run model training scripts locally (offline), and then scale out to the cloud
- Management and deployment of models to the cloud or edge devices
- Integration with Azure Dev Ops
- Added support for R (preview)

Considerations:

- Investigate MLflow to track metrics and manage models

Bradley

- Data Scientist / Engineer
- Apache Spark / SQL / Python / Scala
- Wants to spend more time outdoors than exploring beta version tools

- Build an enterprise data lake and data science environment

Azure Databricks

What is it?

Spark-based analytics platform

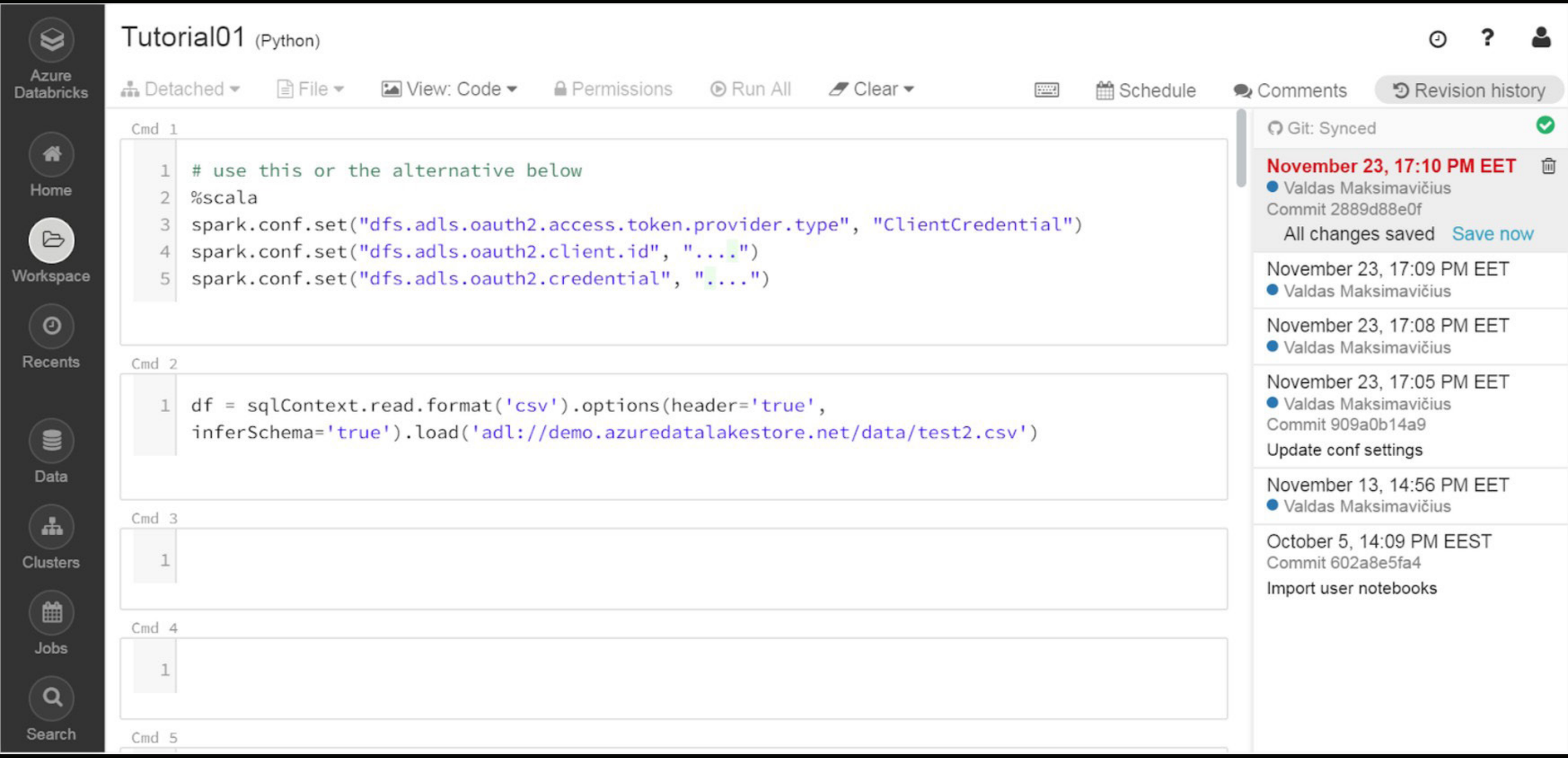
What can you do with it?

Build and deploy models and data workflows

Azure Databricks - Overview

Collaborative Workspace

- Notebooks
- User access
- Git integration


Azure Databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Tutorial01 (Python)
Detached
File
View: Code
Permissions
Run All
Clear
Schedule
Comments
Revision history
Cmd 1
use this or the alternative below
%scala
spark.conf.set("dfs.adls.oauth2.access.token.provider.type", "ClientCredential")
spark.conf.set("dfs.adls.oauth2.client.id", "...")
spark.conf.set("dfs.adls.oauth2.credential", "...")
Cmd 2
df = sqlContext.read.format('csv').options(header='true',
inferSchema='true').load('adl://demo.azuredatalakestore.net/data/test2.csv')
Cmd 3
Cmd 4
Cmd 5
Git: Synced
November 23, 17:10 PM EET
Valdas Maksimavičius
Commit 2889d88e0f
All changes saved
Save now
November 23, 17:09 PM EET
Valdas Maksimavičius
November 23, 17:08 PM EET
Valdas Maksimavičius
November 23, 17:05 PM EET
Valdas Maksimavičius
Commit 909a0b14a9
Update conf settings
November 13, 14:56 PM EET
Valdas Maksimavičius
October 5, 14:09 PM EEST
Commit 602a8e5fa4
Import user notebooks

Azure Databricks - Overview

Collaborative Workspace

- Notebooks
- User access
- Git integration

Databricks Runtime

- Apache Spark
- Rest APIs
- Libraries

Azure Databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Create Cluster

New Cluster

Cancel | Create Cluster

2-100 Workers: 864.0-43200.0 GB Memory, 128-6400 Cores, 32-1600 DBU
1 Driver: 432.0 GB Memory, 64 Cores, 16 DBU

Cluster Name

big_data_conf_2019

Cluster Mode

Standard

Pool

None

Databricks Runtime Version Learn more

Runtime: 6.1 (Scala 2.11, Spark 2.4.4)

New This Runtime version supports only Python 3.

Autopilot Options

Enable autoscaling

Terminate after 120 minutes of inactivity

Worker Type	Min Workers	Max Workers
Standard_E64s_v3 432.0 GB Memory, 64 Cores, 16 DBU	2	100

Driver Type

Same as worker 432.0 GB Memory, 64 Cores, 16 DBU

Azure Databricks - Overview

Collaborative Workspace
- Notebooks
- User access
- Git integration

Databricks Runtime
- Apache Spark
- Rest APIs
- Libraries

Deploy Jobs & Workflows
- Job scheduler
- Notifications & logs
- Multi-stage pipelines

Azure Databricks - Overview

Collaborative Workspace
- Notebooks
- User access
- Git integration

Deploy Jobs & Workflows
- Job scheduler
- Notifications & logs
- Multi-stage pipelines

Databricks Runtime
- Apache Spark
- Rest APIs
- Libraries

Security
- Single sign-on (SSO)
- Access control list (ACL)
- Secrets via Key Vault

Azure Databricks - Summary

Key benefits:

- The most mature development environment for ML on the Azure platform
- Seamless integration with MLflow & Azure ML
- Integrated with other Azure services (e.g., Azure Data Factory, Azure Key Vault)
- Delta Lake support

Azure Databricks - Summary

Key benefits:

- The most mature development environment for ML on the Azure platform
- Seamless integration with MLflow & Azure ML
- Integrated with other Azure services (e.g., Azure Data Factory, Azure Key Vault)
- Delta Lake support

Considerations:

- Online only
- Cost includes the price of virtual machines and Databricks fee

Joshua

- Data Scientist
- Research and development

- “I need a sandbox to learn and evaluate new tools”

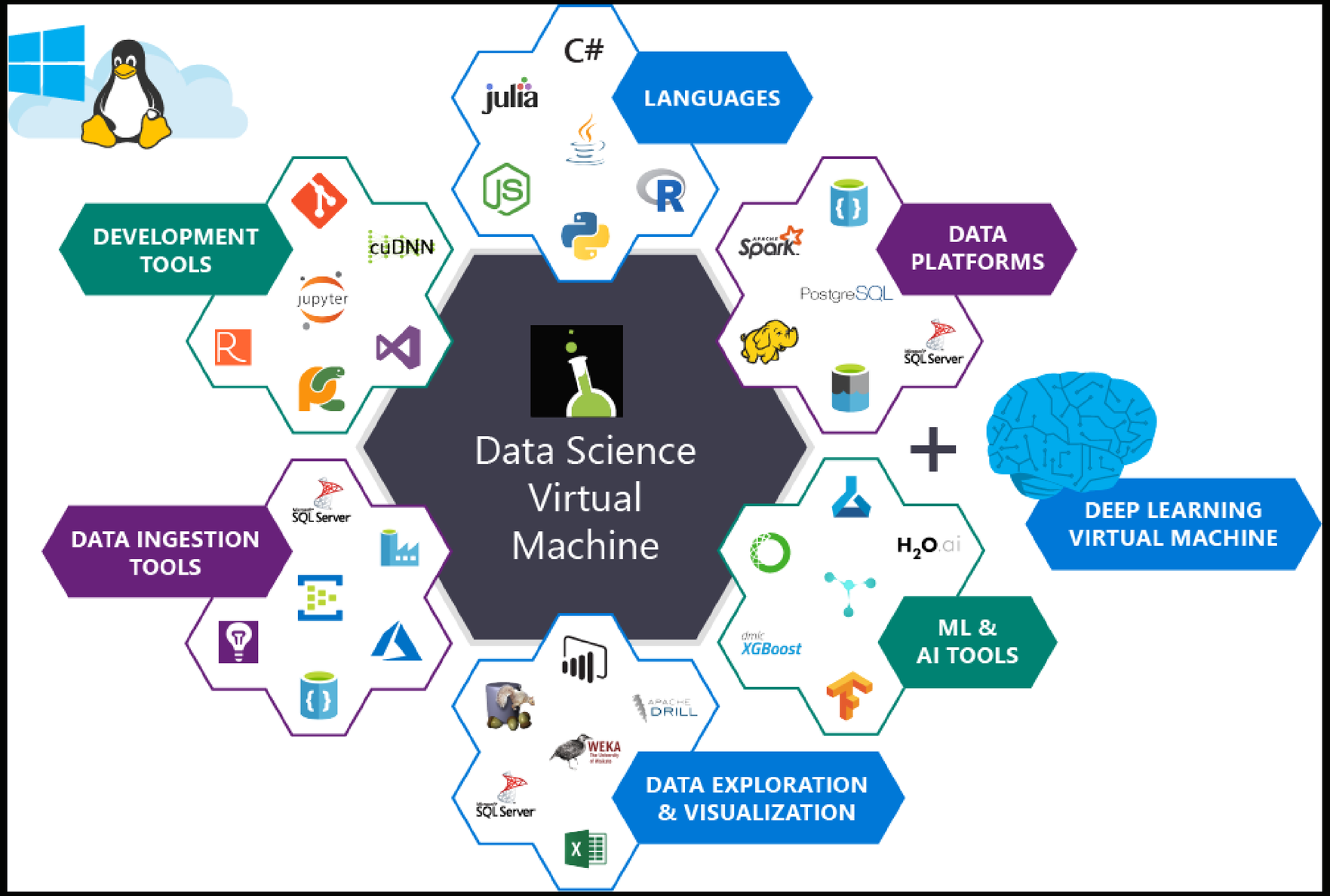
Data Science Virtual Machine

What is it?

A virtual machine with pre-installed data science tools

What can you do with it?

Develop ML solutions in a pre-configured environment


LANGUAGES
C#
julia
DEVELOPMENT TOOLS
cuDNN
jupyter
DATA PLATFORMS
Spark
PostgreSQL
SQL Server
Data Science Virtual Machine
DATA INGESTION TOOLS
SQL Server
ML & AI TOOLS
H2O.ai
XGBoost
DEEP LEARNING VIRTUAL MACHINE
DATA EXPLORATION & VISUALIZATION
APACHE DRILL
WEKA
SQL Server

Azure Data Science Virtual Machine - Summary

Key benefits:

- Probably the most complete development environment for ML on the Azure platform
- Reduced time to install, manage, and troubleshoot data science tools and frameworks
- Virtual machine options include highly scalable GPU images
- A dedicated geospatial with ArcGIS distribution

Considerations:

- Online only
- You need to take care of VM management

I asked for Cloud and they said NO

Rick

- Data Scientist
- Python, R, ScikitLearn, etc.

- Create personalized treatment based on individual health data

Microsoft Machine Learning ~~Service~~ Server

What is it?

Cross-platform standalone server for predictive analysis

What can you do with it?

Build and deploy models with R and Python

Microsoft Machine Learning Server - Summary

Key benefits:

- Built on a legacy of Microsoft R Server and Revolution R Enterprise
- Install on Windows / Linux
- Advanced security options
- Deploy R and Python models as web services

Microsoft Machine Learning Server - Summary

Key benefits:

- Built on a legacy of Microsoft R Server and Revolution R Enterprise
- Install on Windows / Linux
- Advanced security options
- Deploy R and Python models as web services

Considerations:

- You need to deploy and manage Machine Learning Server in your enterprise

SQL Server Machine Learning Services

What is it?

A built-in SQL Server feature to support machine learning

What can you do with it?

Execute Python and R scripts with relational data

SQL Server Machine Learning Services - Example

```
CREATE PROCEDURE predict_species (@model VARCHAR(100))
AS
BEGIN
    DECLARE @nb_model VARBINARY(max) = (
            SELECT model
            FROM iris_models
            WHERE model_name = @model
            );

    EXECUTE sp_execute_external_script @language = N'Python'
        , @script = N'
import pickle
irismodel = pickle.loads(nb_model)
species_pred = irismodel.predict(iris_data[["Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width"]])
iris_data["PredictedSpecies"] = species_pred
OutputDataSet = iris_data[["id","SpeciesId","PredictedSpecies"]]
print(OutputDataSet)
'
        , @input_data_1 = N'select id, "Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width", "SpeciesId" from iris_data'
        , @input_data_1_name = N'iris_data'
        , @params = N'@nb_model varbinary(max)'
        , @nb_model = @nb_model
    WITH RESULT SETS((
                "id" INT
              , "SpeciesId" INT
              , "SpeciesId.Predicted" INT
                ));
END;
GO

```

SQL Server Machine Learning Services - Summary

Key benefits:

- Run your scripts where the data resides and eliminate transfer of the data across the network to another server
- Encapsulate predictive logic in a database function

SQL Server Machine Learning Services - Summary

Key benefits:

- Run your scripts where the data resides and eliminate transfer of the data across the network to another server
- Encapsulate predictive logic in a database function

Considerations:

- Assumes a SQL Server database as the data tier for your application
- Limited scalability
- Long list of known issues

Quiz

Azure Cognitive Services

Machine Learning for .NET

Power BI Auto ML

Azure Machine Learning Studio

Azure Machine Learning

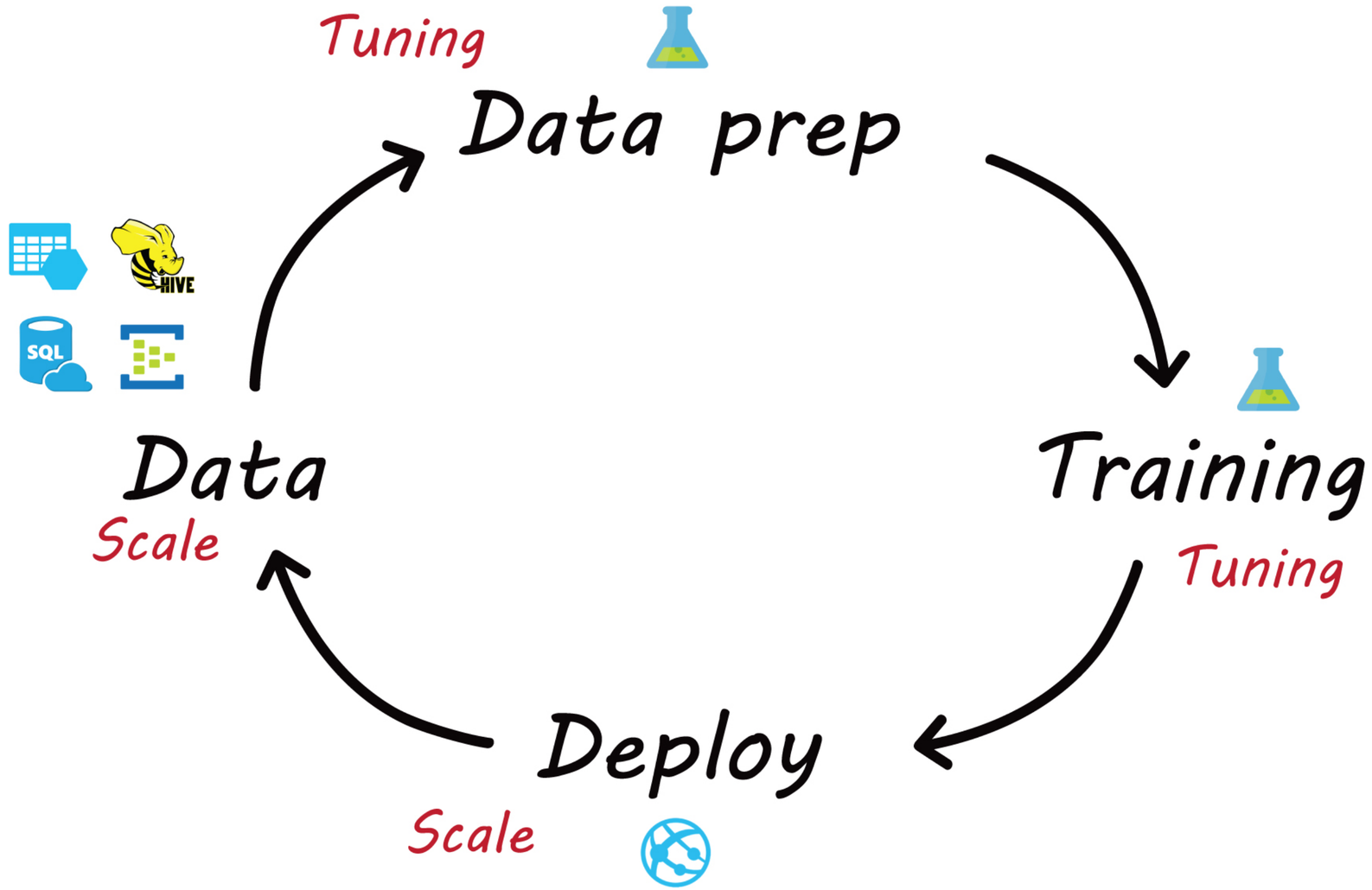
Azure Databricks

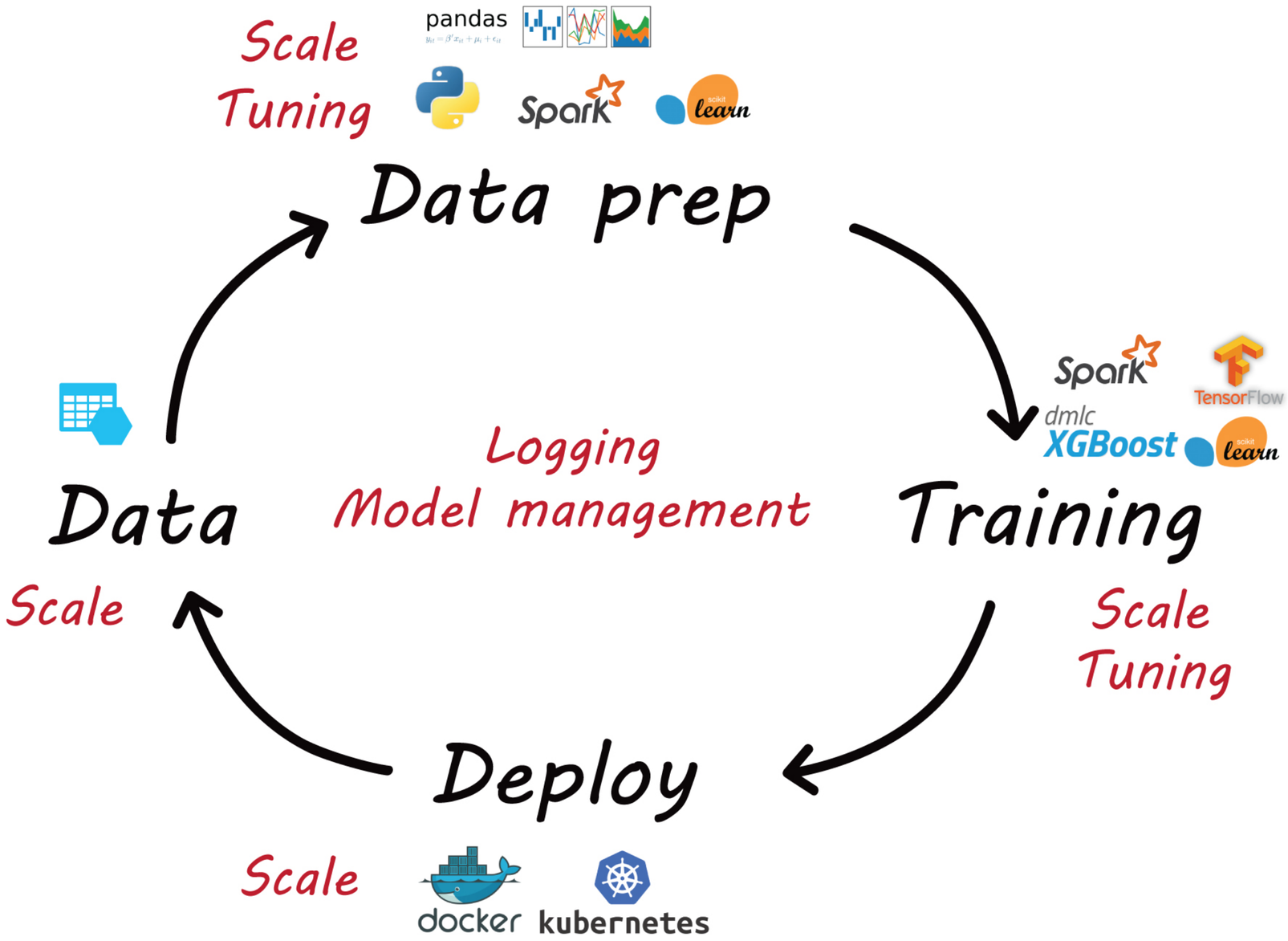
Data Science Virtual Machine

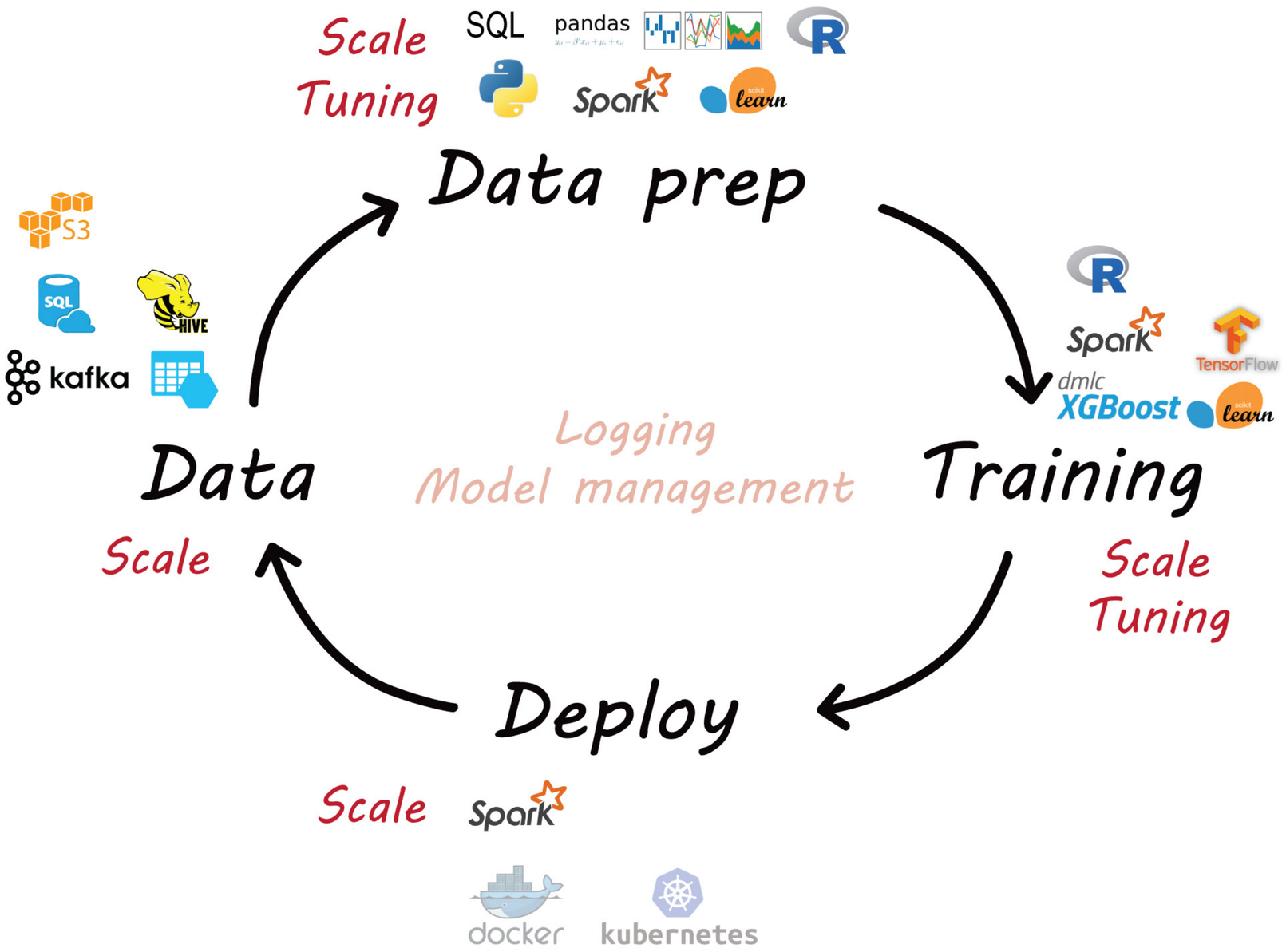
Microsoft Machine Learning Server

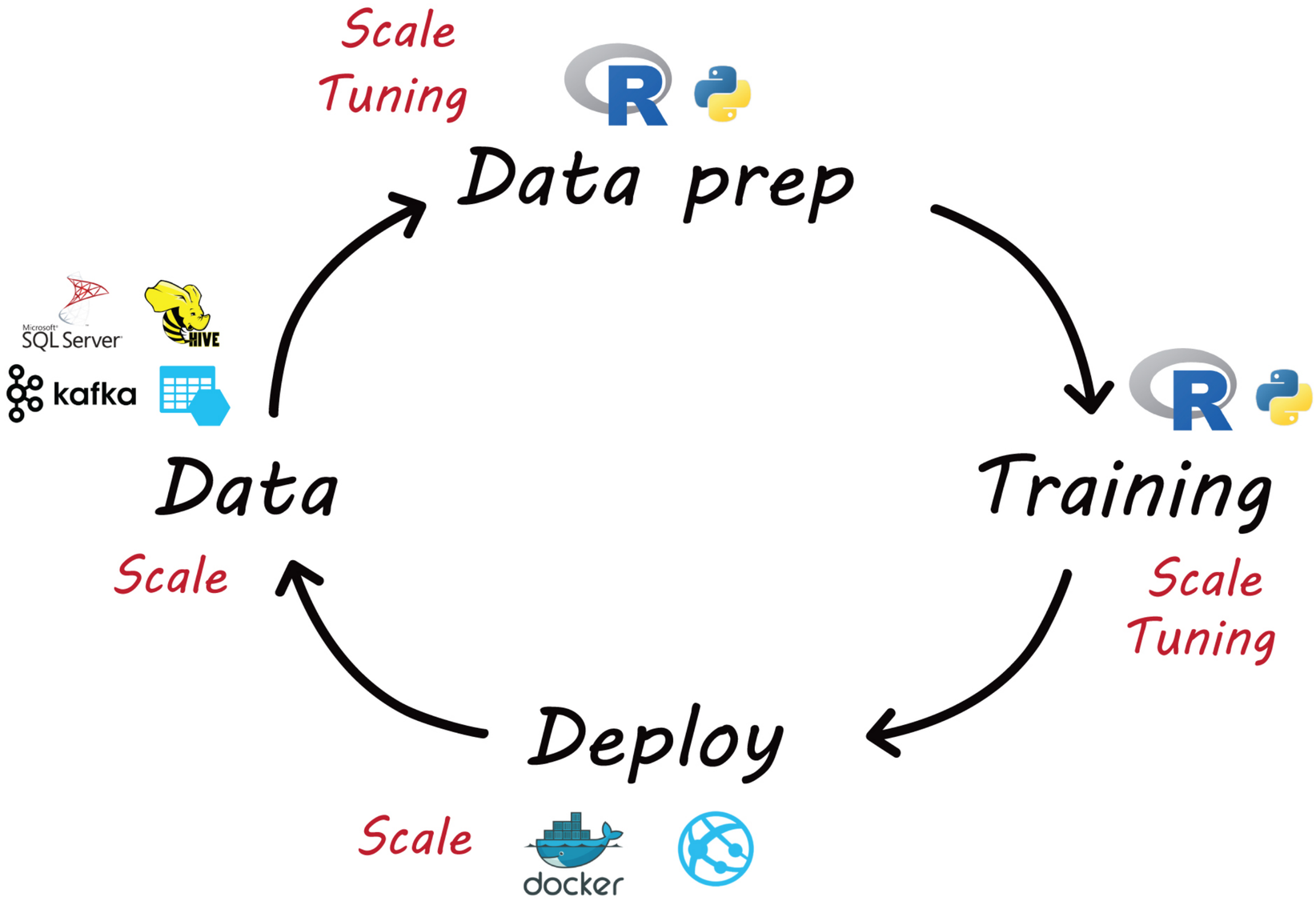
SQL Server Machine Learning Services

valdas@maksimavicius.eu
linkedin.com/in/valdasm
valdas.blog


Tuning
Data prep
HIVE
SQL
Data
Scale
Training
Tuning
Deploy
Scale


Scale
Tuning
pandas
Spark
scikit learn
Data prep
Spark
TensorFlow
dmlc
XGBoost
scikit learn
Logging
Model management
Data
Scale
Training
Scale
Tuning
Deploy
Scale
docker
kubernetes


Scale
Tuning
SQL
pandas
Spark
scikit learn
Data prep
S3
SQL
HIVE
kafka
R
Spark
TensorFlow
dmlc
XGBoost
scikit learn
Logging
Model management
Data
Scale
Training
Scale
Tuning
Deploy
Scale
Spark
docker
kubernetes


Scale
Tuning
Data prep
Training
Scale
Tuning
Deploy
Scale
docker
Data
Scale
Microsoft SQL Server
HIVE
kafka

Microsoft Machine Learning Server - Overview

- A new name for Microsoft R Server
- Install on Windows / Linux / Hadoop cluster
- Deploy models as web services packaged as container images
- Satisfy security and compliance needs of any enterprise