MAKING DATA SCIENTISTS PRODUCTIVE IN AZURE

Valdas Maksimavičius

A little exam

Microsoft Machine Learning Server

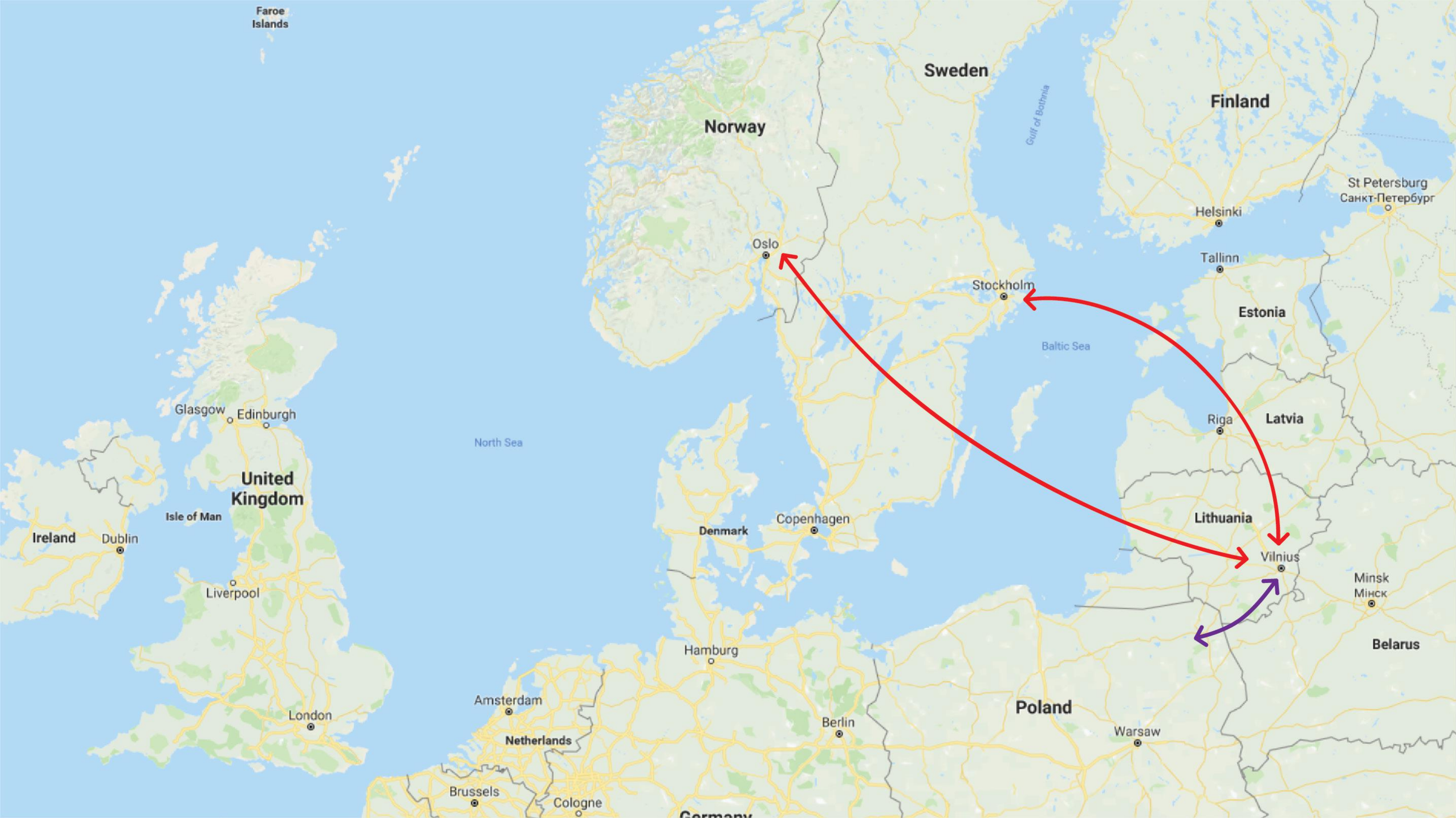
Azure Machine Learning Service

Azure Machine Learning Studio

Azure Machine Learning Workbench

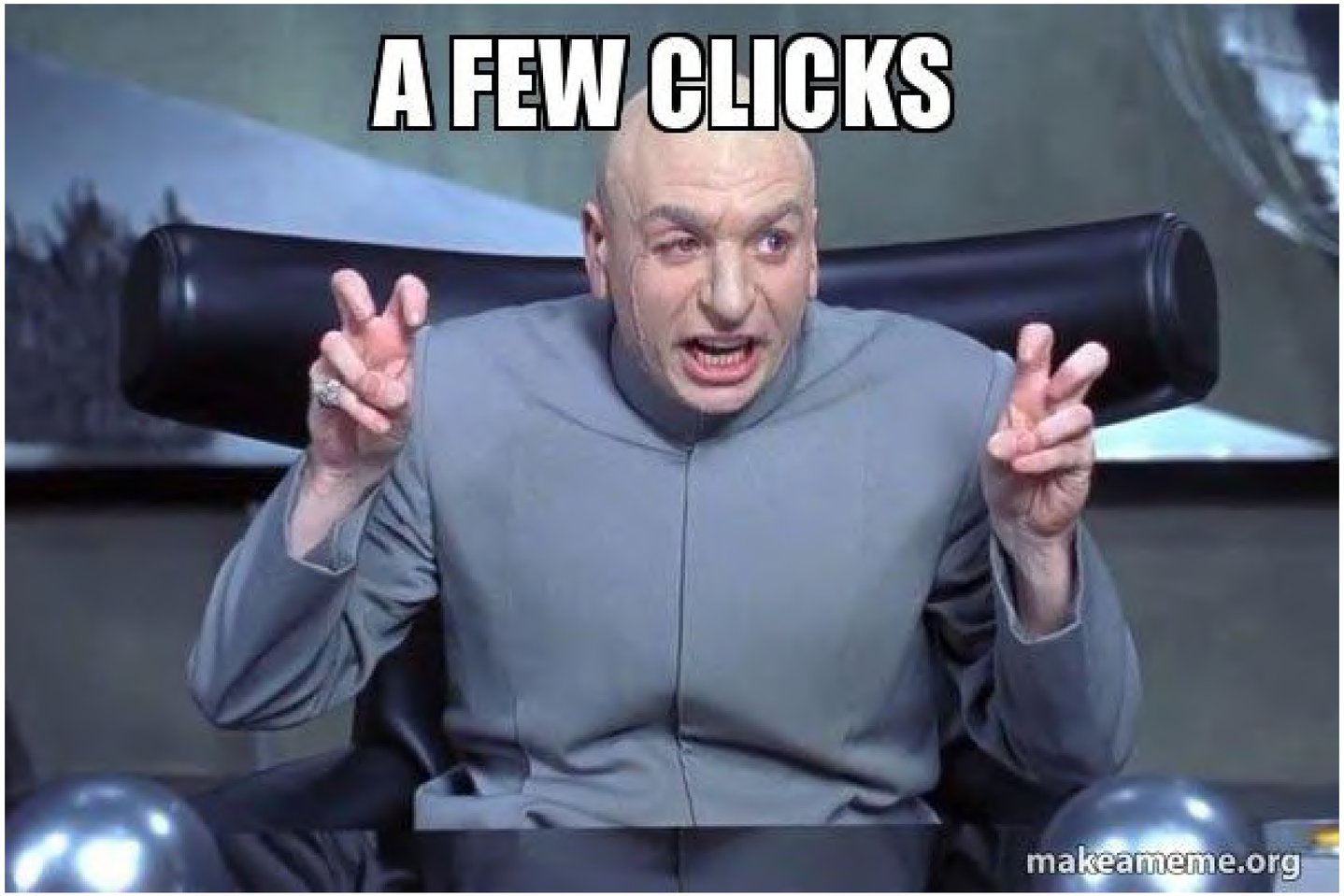
Microsoft R Server

Data Science Virtual Machine


Faroe Islands
Sweden
Finland
Norway
Gulf of Bothnia
St Petersburg
Санкт-Петербург
Helsinki
Oslo
Tallinn
Stockholm
Estonia
Baltic Sea
Glasgow
Edinburgh
Riga
Latvia
North Sea
United Kingdom
Isle of Man
Lithuania
Ireland
Dublin
Denmark
Copenhagen
Vilnius
Minsk
Мінск
Liverpool
Hamburg
Belarus
Amsterdam
Poland
London
Berlin
Netherlands
Warsaw
Brussels
Cologne


Me in 2015
“With just a few clicks, you can have a Hadoop cluster up and running”

Me in 2018



Vilnius Microsoft Data Platform Meetup

Looking for enthusiasts to share their stories


Join at Slido.com with #bigdata2018

Inspiration for the talk #1

Inspiration for the talk #2

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

Search AI + Machine Learning

Pricing
All

Operating System
All

Publisher
All

Machine Learning

More

Machine Learning service workspace
Microsoft

Machine Learning Model
Microsoft

Data Science Virtual Machine -
Microsoft

Machine Learning Studio Workspace
Microsoft

Machine Learning Studio Web
Microsoft

Machine Learning Studio Web
Microsoft

Search AI + Machine Learning

Pricing
All

Operating System
All

Publisher
All

Machine Learning

More

Machine Learning service workspace
Microsoft

Machine Learning Model
Microsoft

Data Science Virtual Machine -
Microsoft

Machine Learning Studio Workspace
Microsoft

Machine Learning Studio Web
Microsoft

Machine Learning Studio Web
Microsoft

Azure Databricks
Microsoft

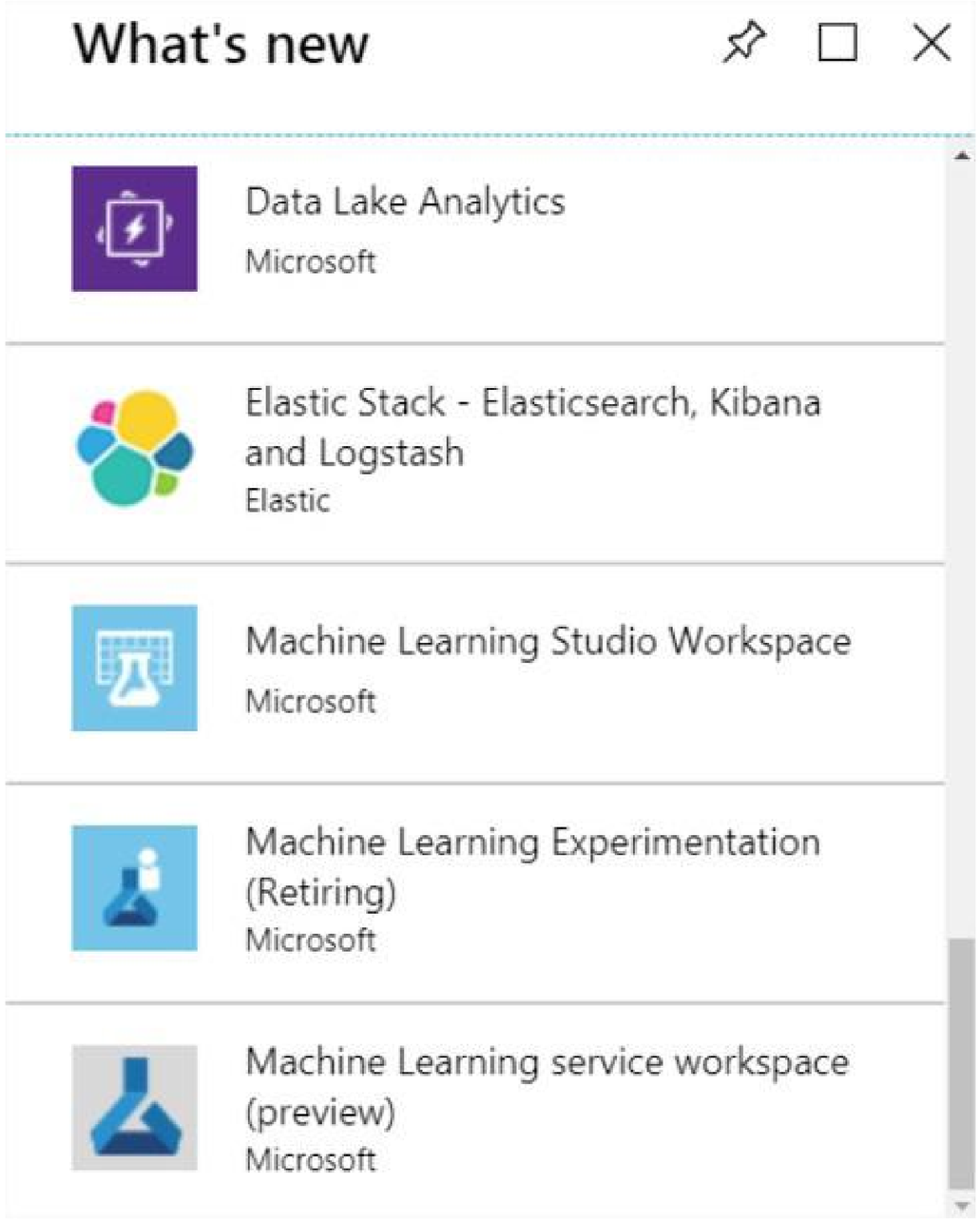
HDInsight
Microsoft

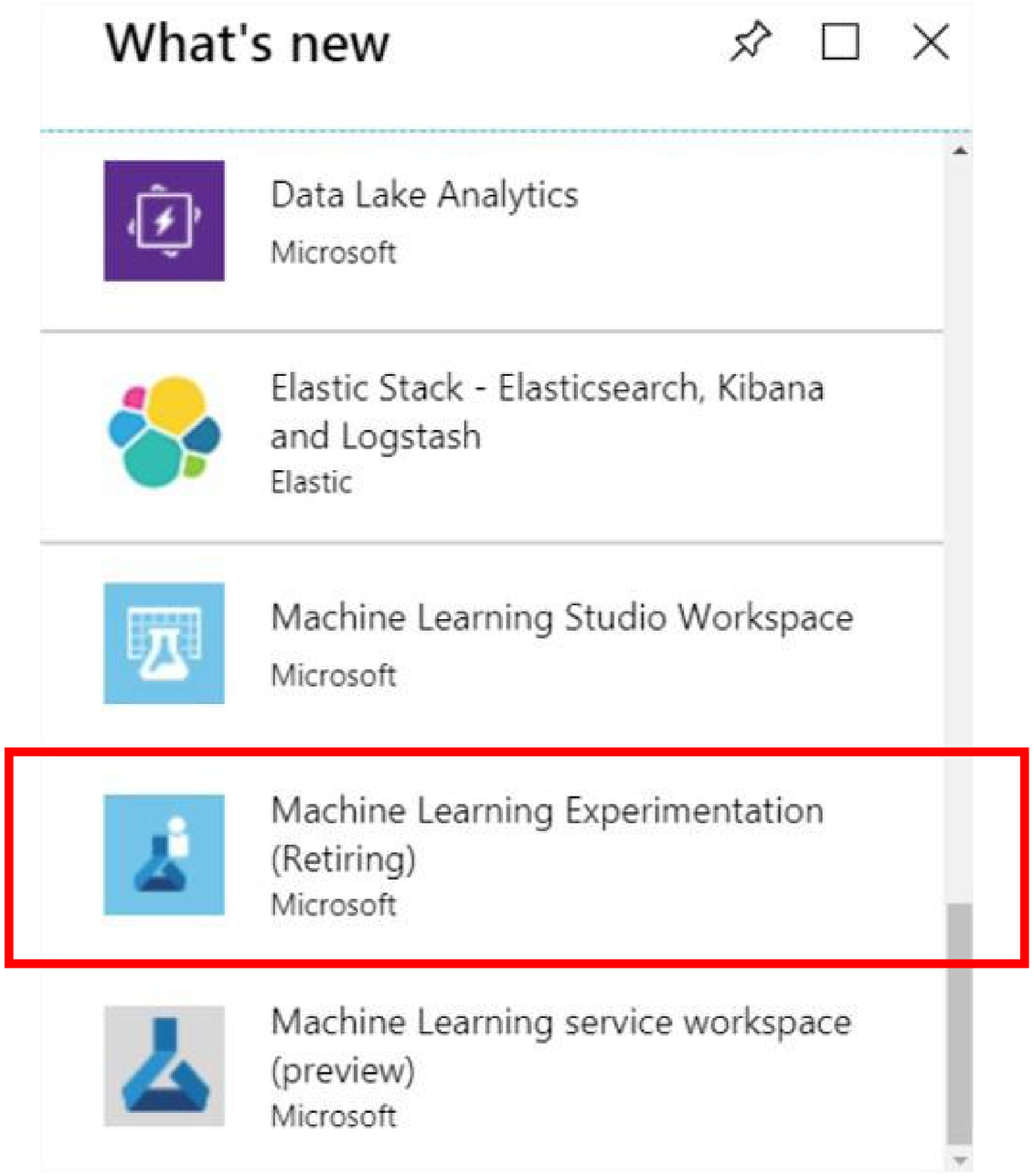
Data Factory
Microsoft

Data Lake Analytics
Microsoft

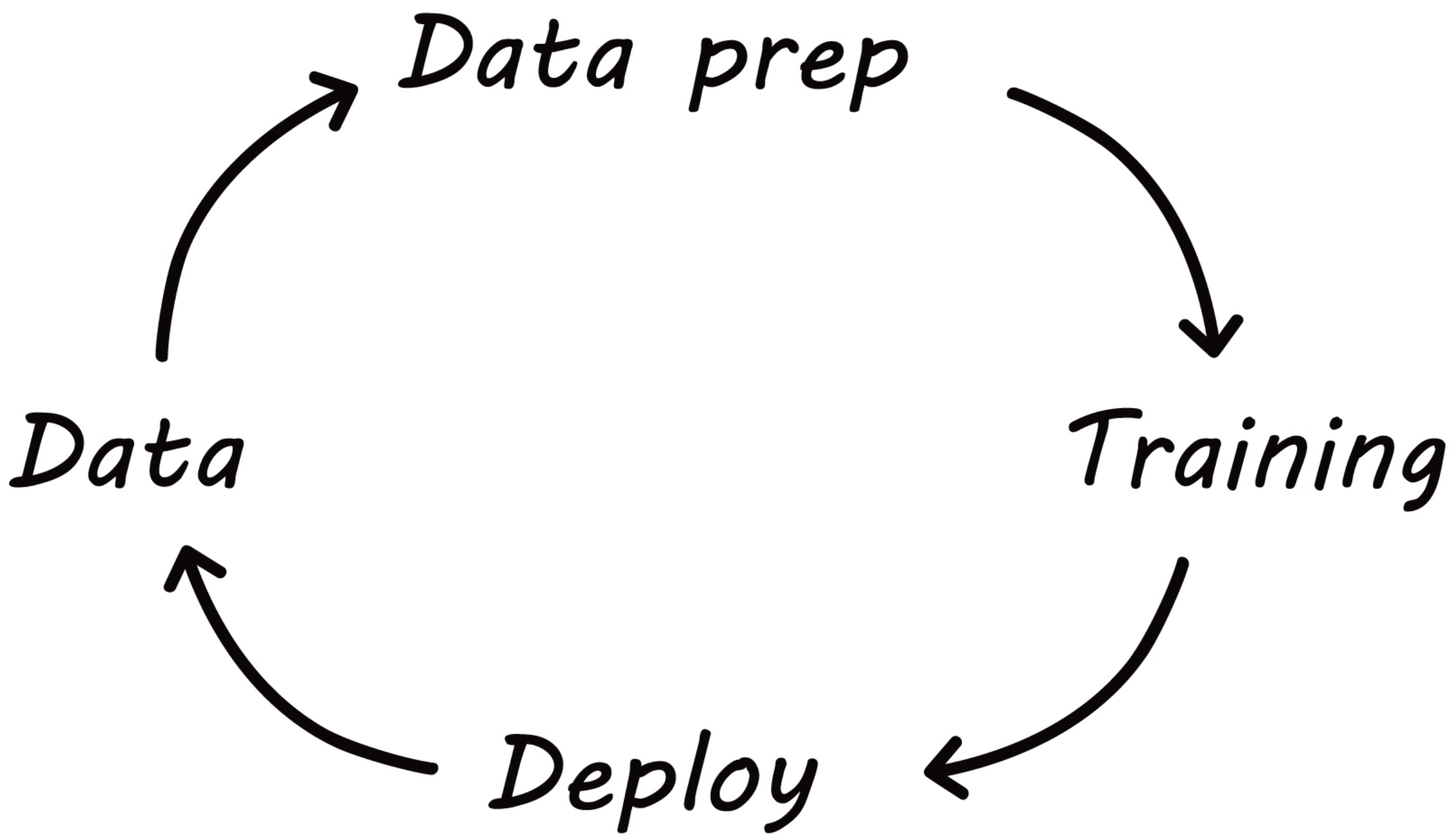
ExpressRoute
Microsoft

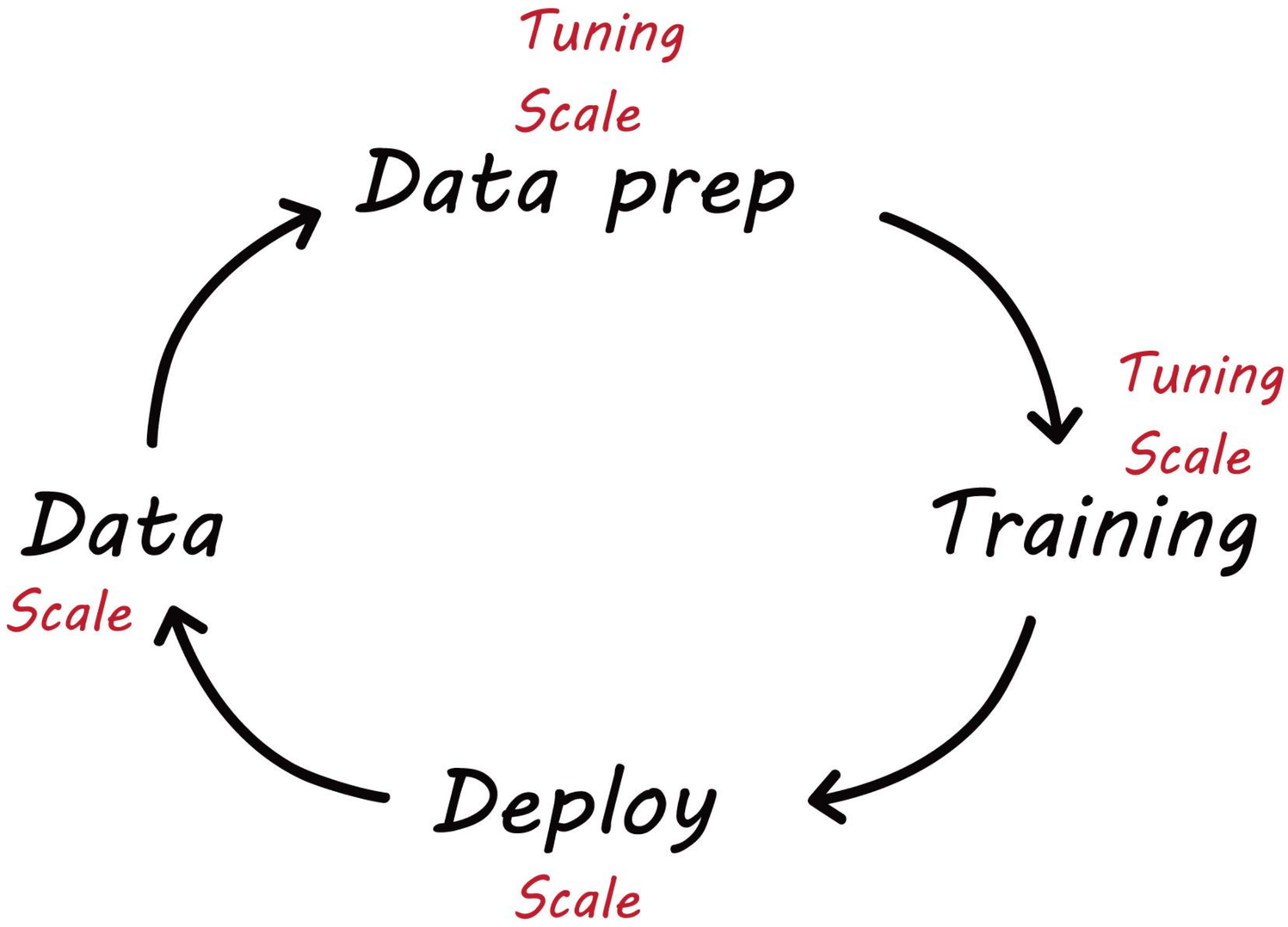
Azure Blockchain Workbench
Microsoft

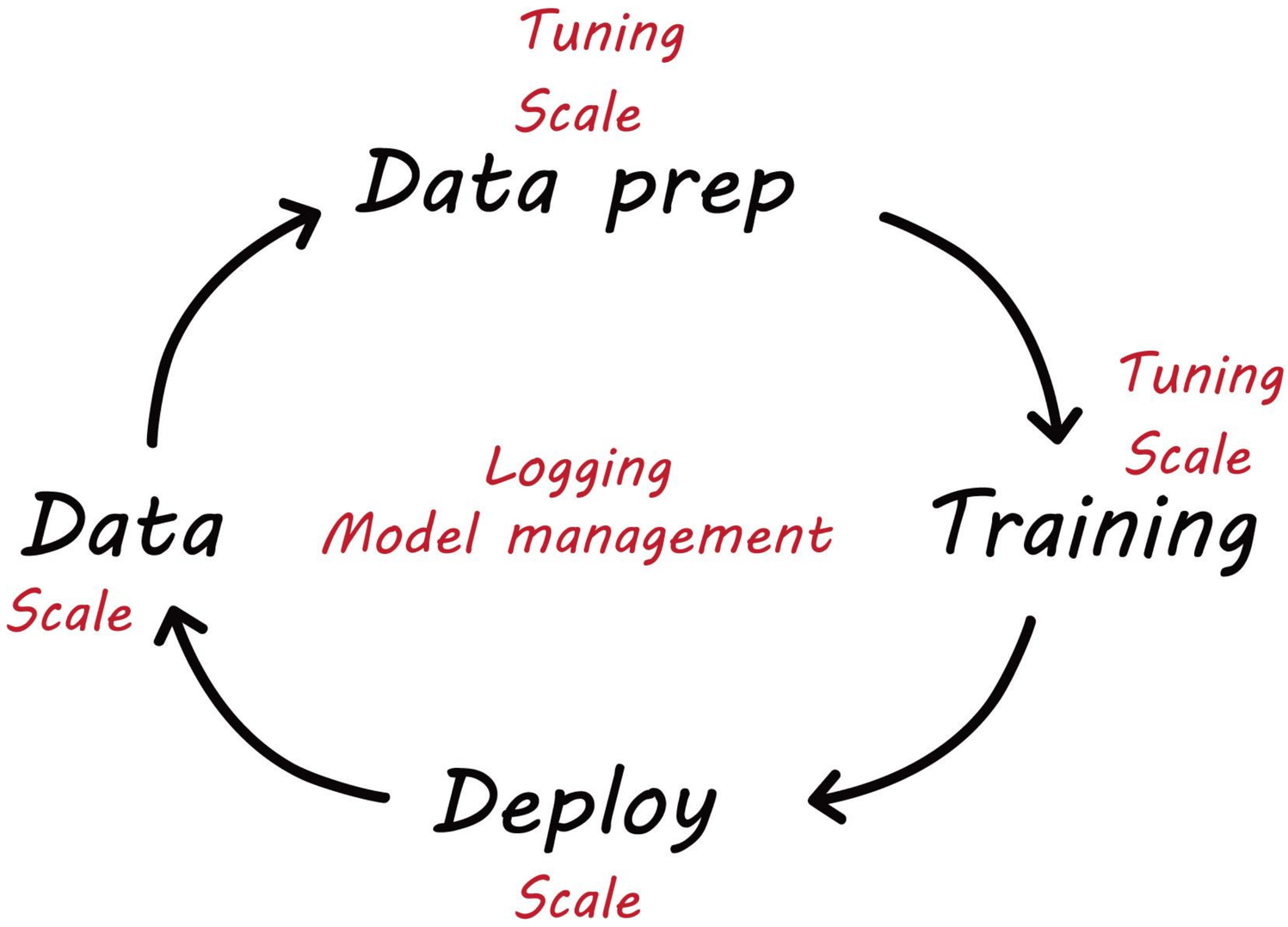

What's new
Data Lake Analytics
Microsoft
Elastic Stack - Elasticsearch, Kibana and Logstash
Elastic
Machine Learning Studio Workspace
Microsoft
Machine Learning Experimentation (Retiring)
Microsoft
Machine Learning service workspace (preview)
Microsoft


What's new
Data Lake Analytics
Microsoft
Elastic Stack - Elasticsearch, Kibana and Logstash
Elastic
Machine Learning Studio Workspace
Microsoft
Machine Learning Experimentation (Retiring)
Microsoft
Machine Learning service workspace (preview)
Microsoft

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

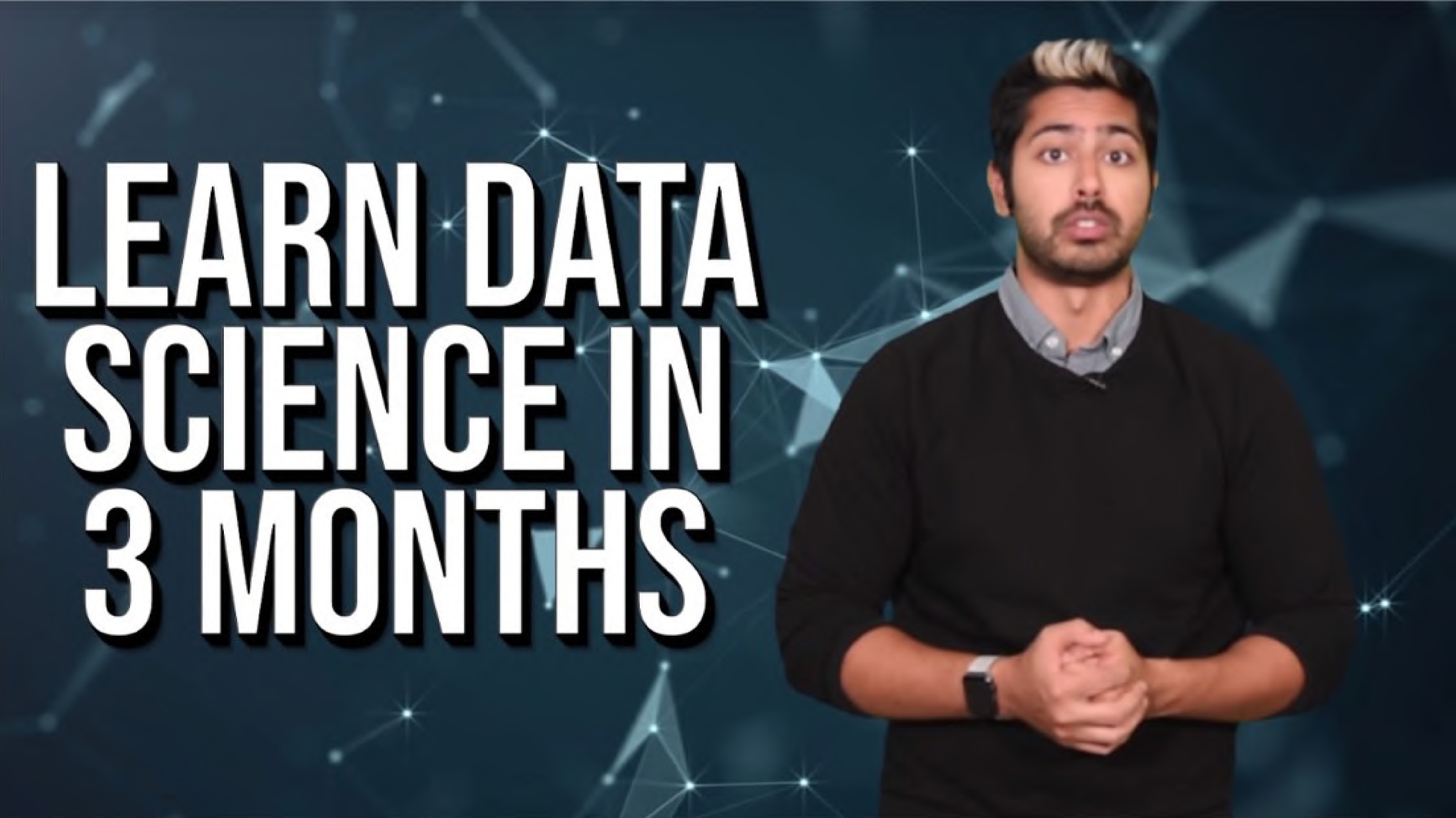




Tom

- Full stack software developer
- Android, Node.js, React

- Scan faces to decide what ad to serve


LEARN DATA
SCIENCE IN
3 MONTHS

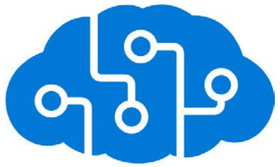
Azure Cognitive Services

What is it?

Azure services with pre-built AI and ML models

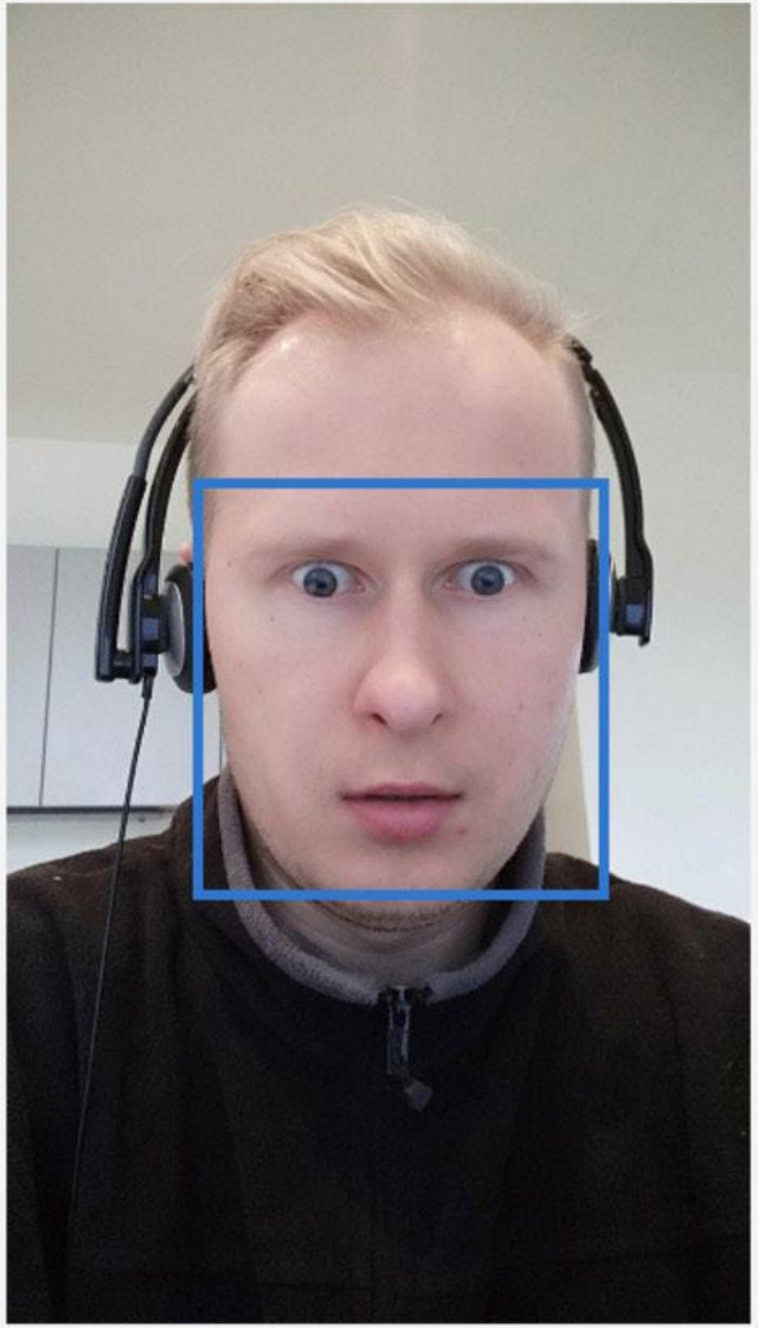
What can you do with it?

Easily add intelligent features to your apps



Azure Cognitive Services - Overview

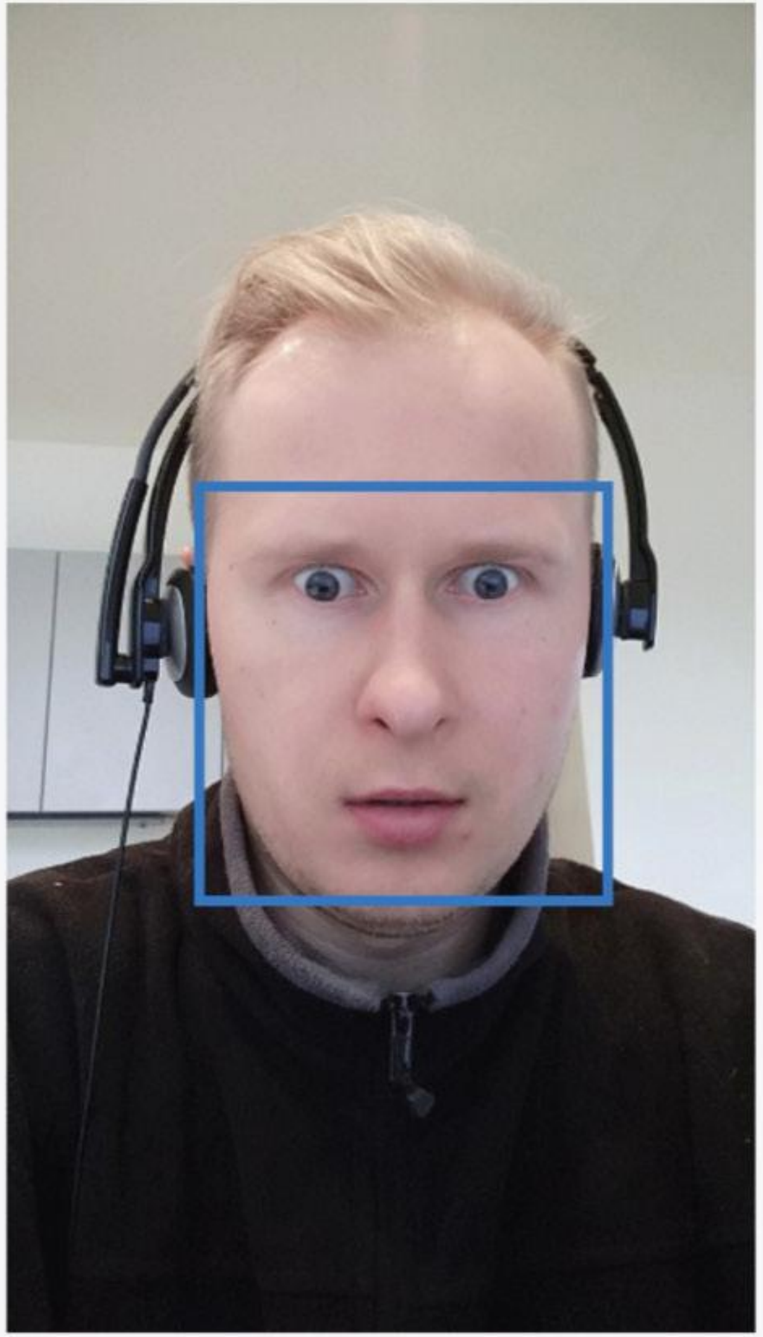
- Text analysis
- Computer vision
- Video analytics
- Speech recognition and generation
- Natural language understanding (LUIS)
- Search



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

Azure Cognitive Services - Summary

Key benefits:

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

Azure Cognitive Services - Summary

Key benefits:

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

Considerations:

- Only available over the web (an exception is the Custom Vision Service)
- Just partial customization allowed
- Limited support for Non-English languages

Mark

- Business Analyst
- Basics of statistical analysis

▹ Create a sales lead list

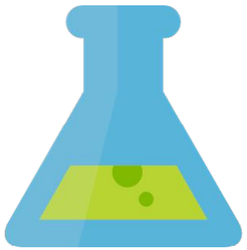
Azure Machine Learning Studio

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



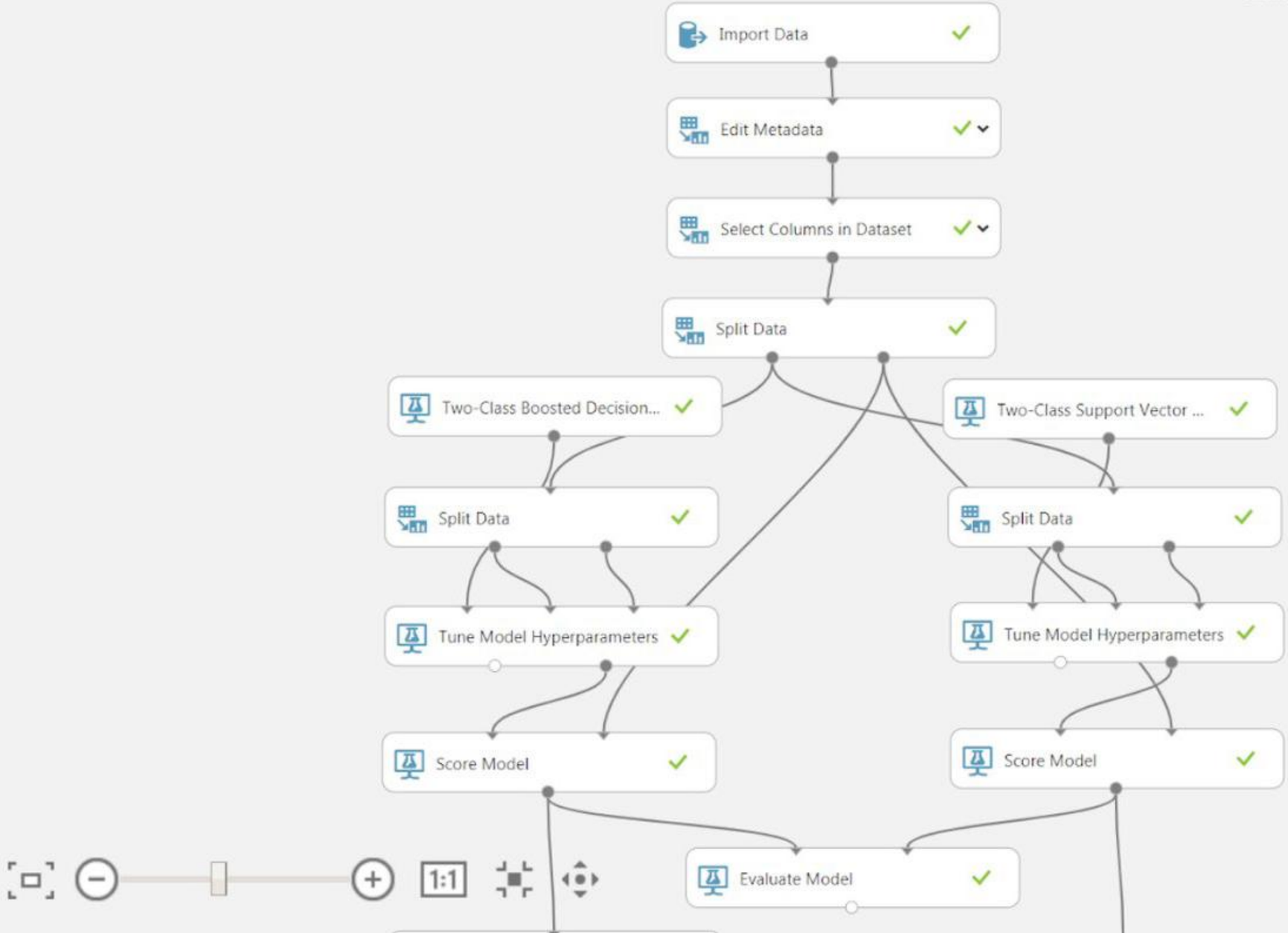
Binary Classification: Direct marketing

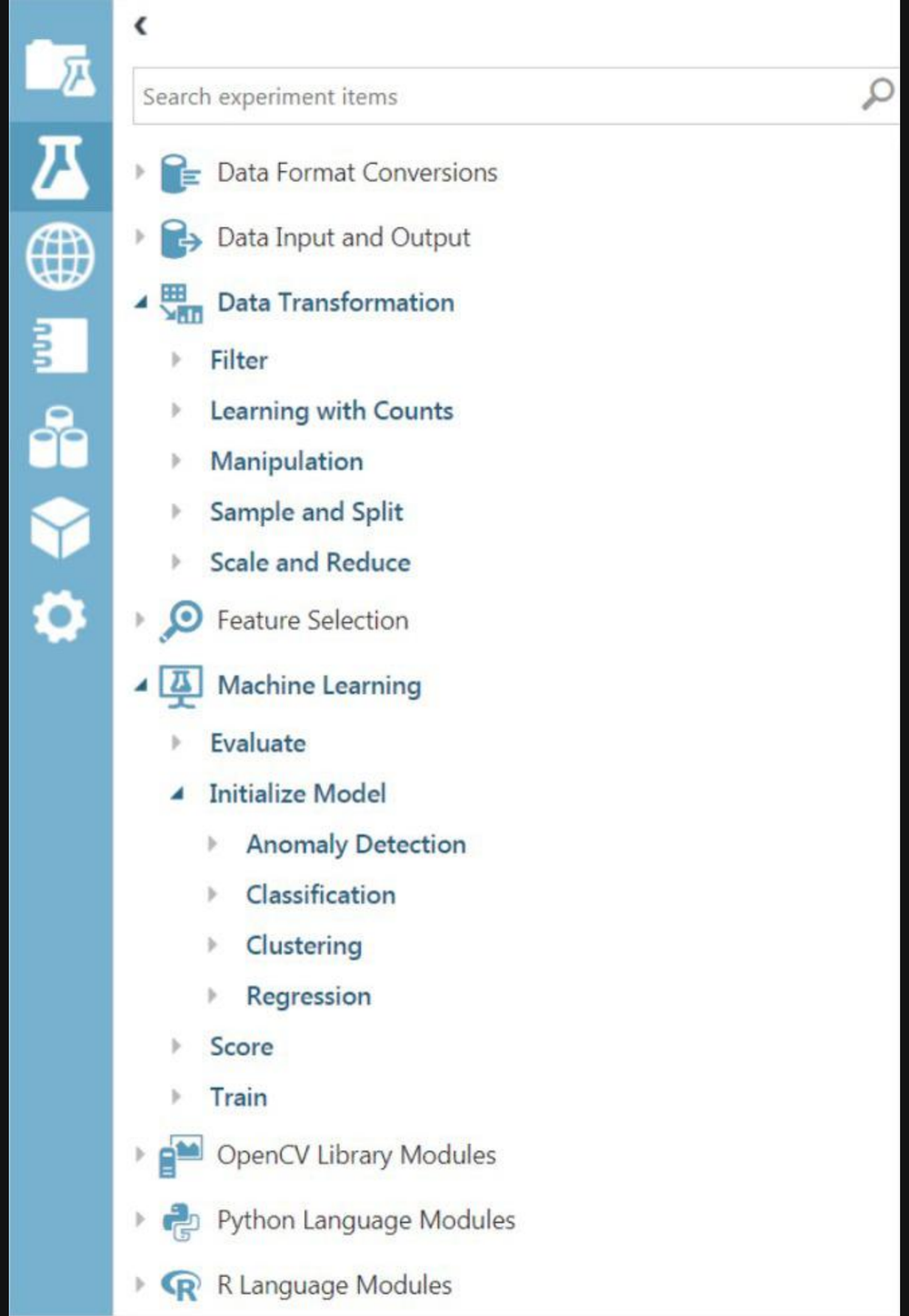
Finished running

Draft saved at 12:38:31

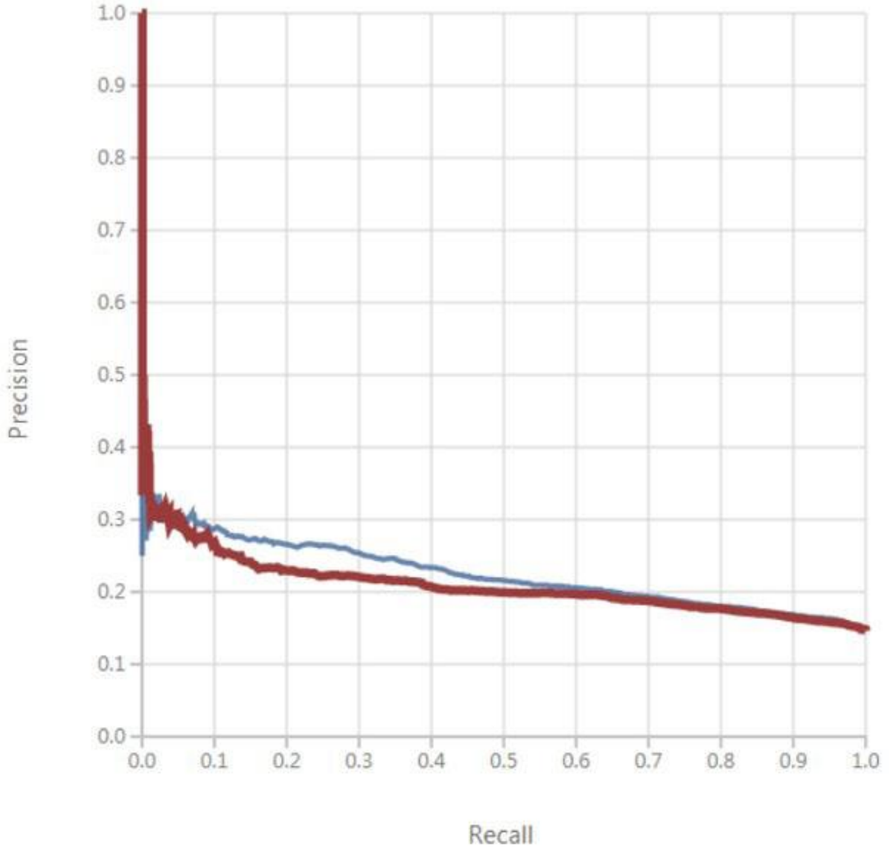
Search experiment items

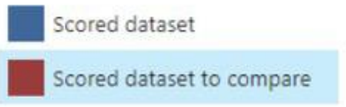
- Saved Datasets
- Trained Models
- Transforms
- Data Format Conversions
- Data Input and Output
- Data Transformation
- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Time Series
- Web Service
-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

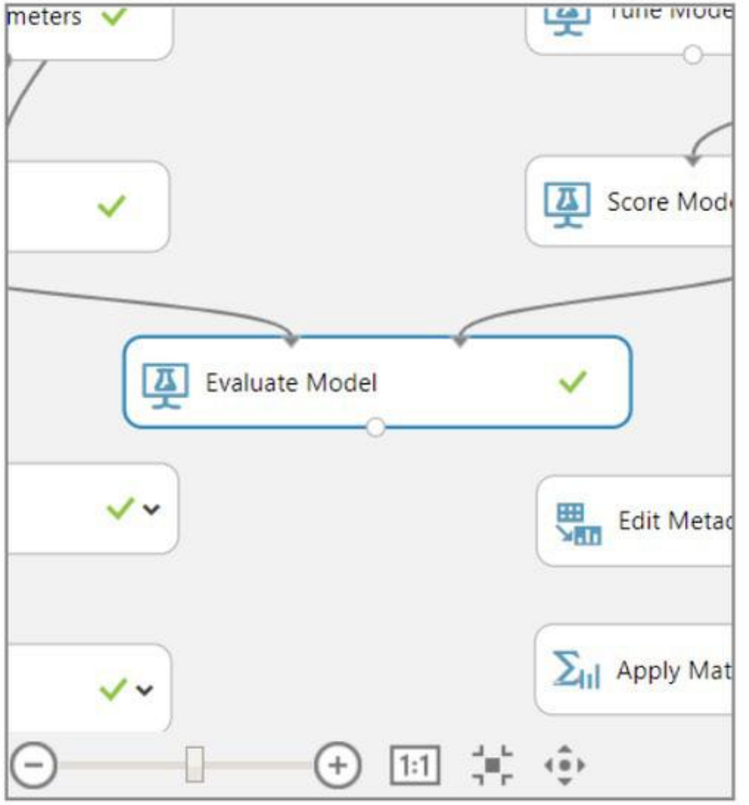



Search experiment items
Data Format Conversions
Data Input and Output
Data Transformation
Filter
Learning with Counts
Manipulation
Sample and Split
Scale and Reduce
Feature Selection
Machine Learning
Evaluate
Initialize Model
Anomaly Detection
Classification
Clustering
Regression
Score
Train
OpenCV Library Modules
Python Language Modules
R Language Modules

ROC PRECISION/RECALL LIFT







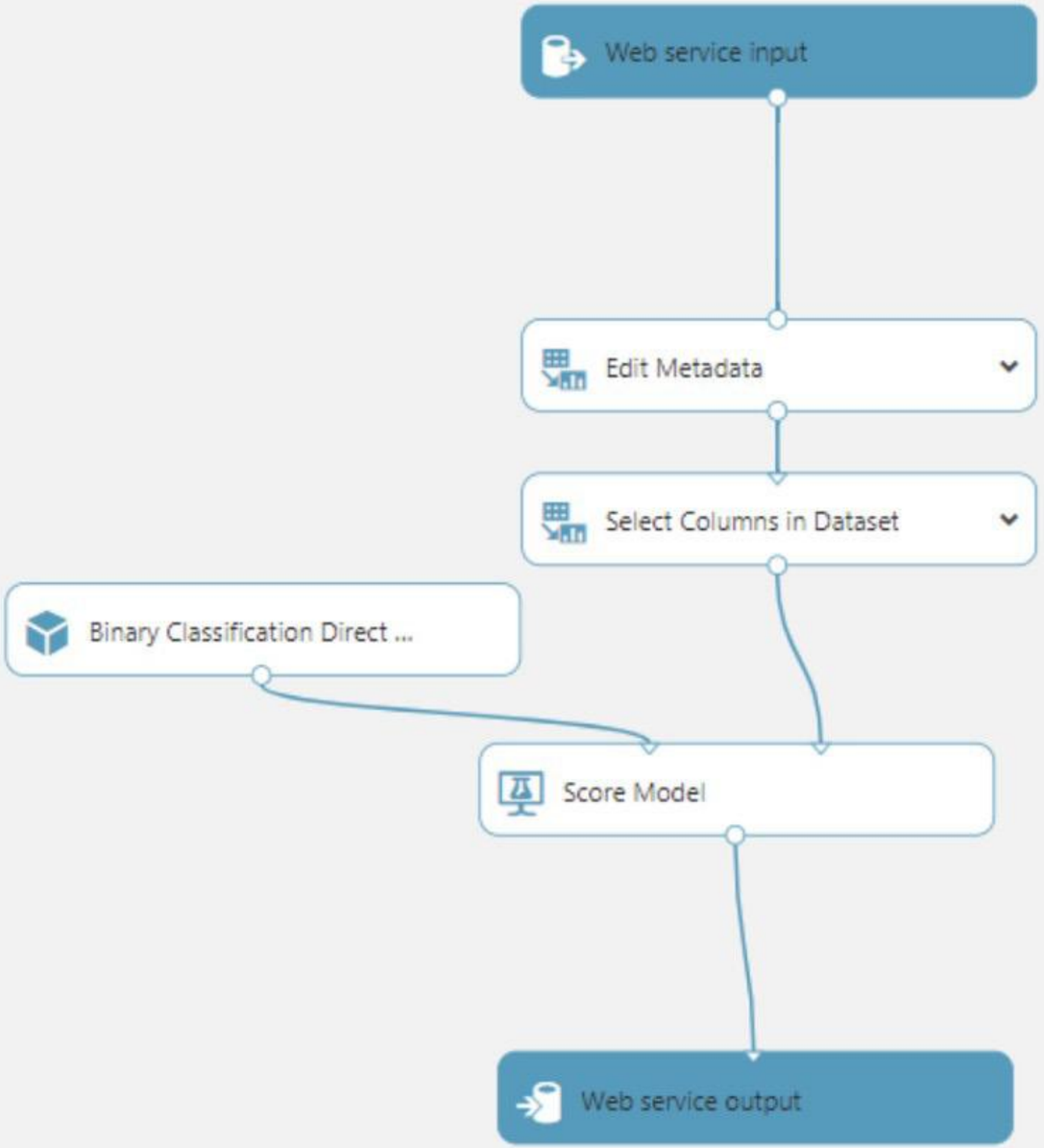
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
0	2818	0.853	1.000	0.5	0.620

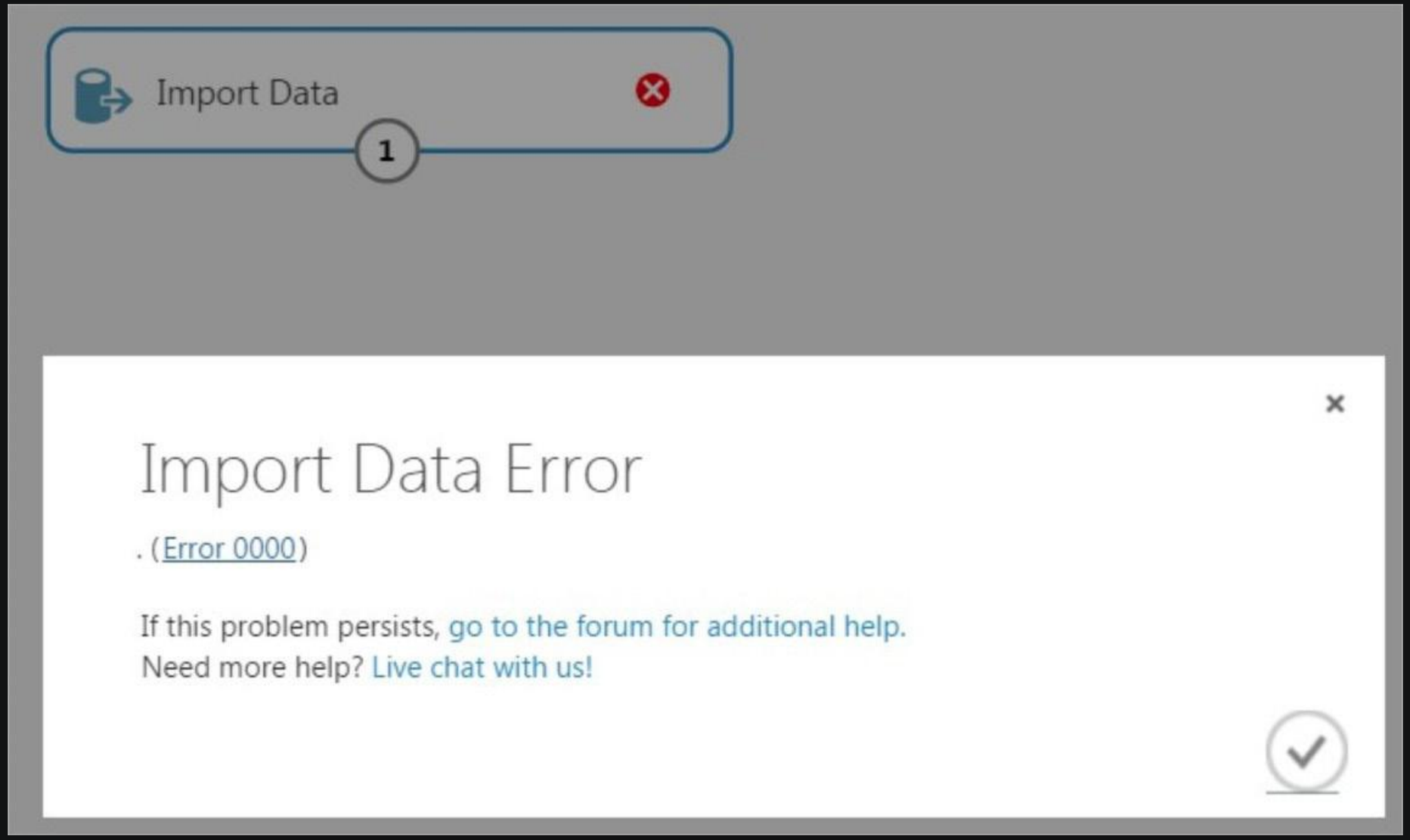
False Positive	True Negative	Recall	F1 Score
0	16382	0.000	0.000

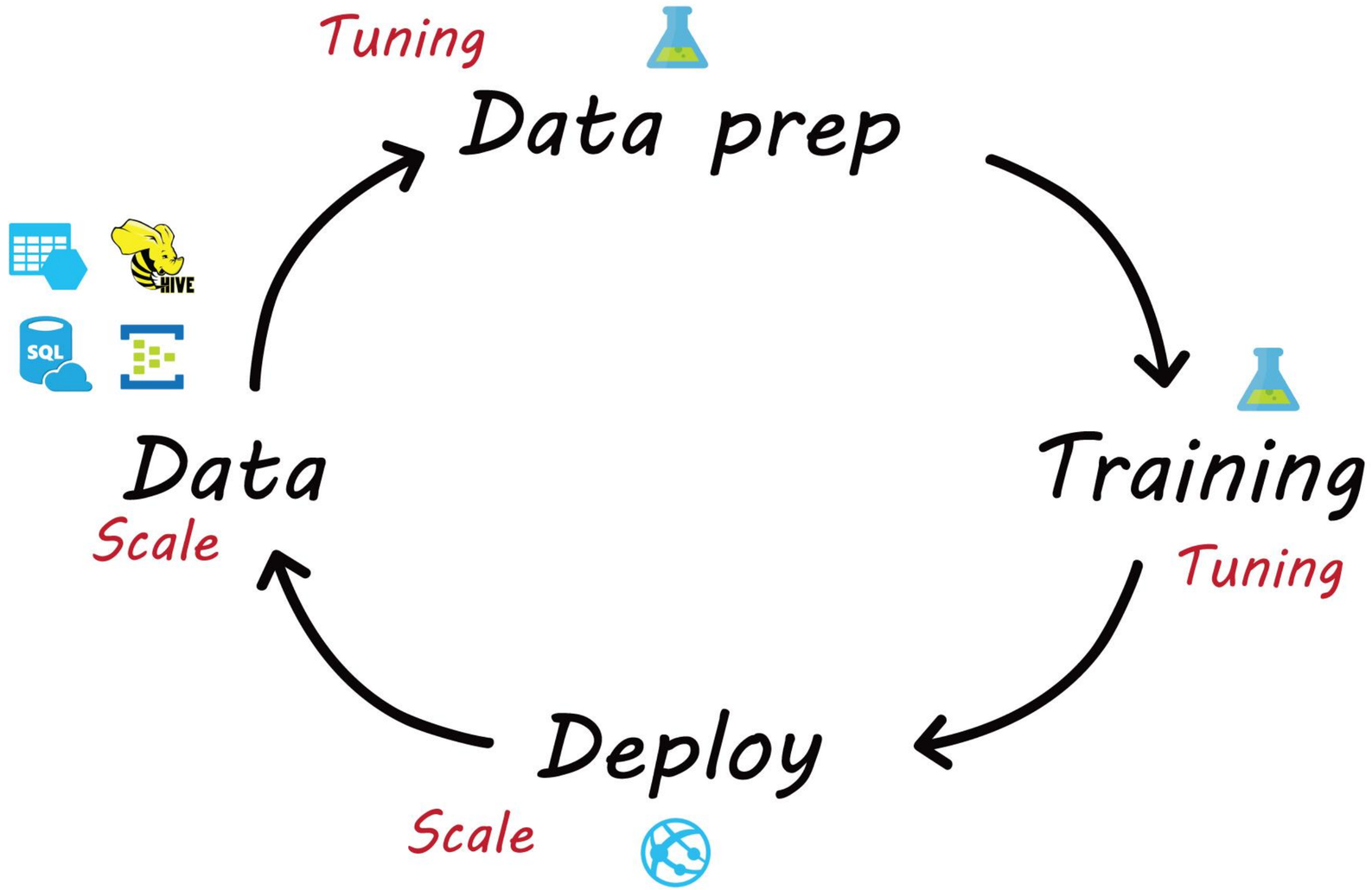
Positive Label	Negative Label
1	0

Deploy as web services

- Batch execution
- Request / Response




Import Data
1
Import Data Error
. (Error 0000)
If this problem persists, go to the forum for additional help.
Need more help? Live chat with us!


Tuning
Data prep
HIVE
SQL
Data
Scale
Training
Tuning
Deploy
Scale

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:

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

Lucy

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

- Estimate damage (repair cost) in auto insurance

Missing headlight: 300 Eur

Azure Machine Learning ~~Studio~~ Service

What is it?

Managed cloud service for ML

What can you do with it?

Train, deploy and manage models in Azure using Python and CLI

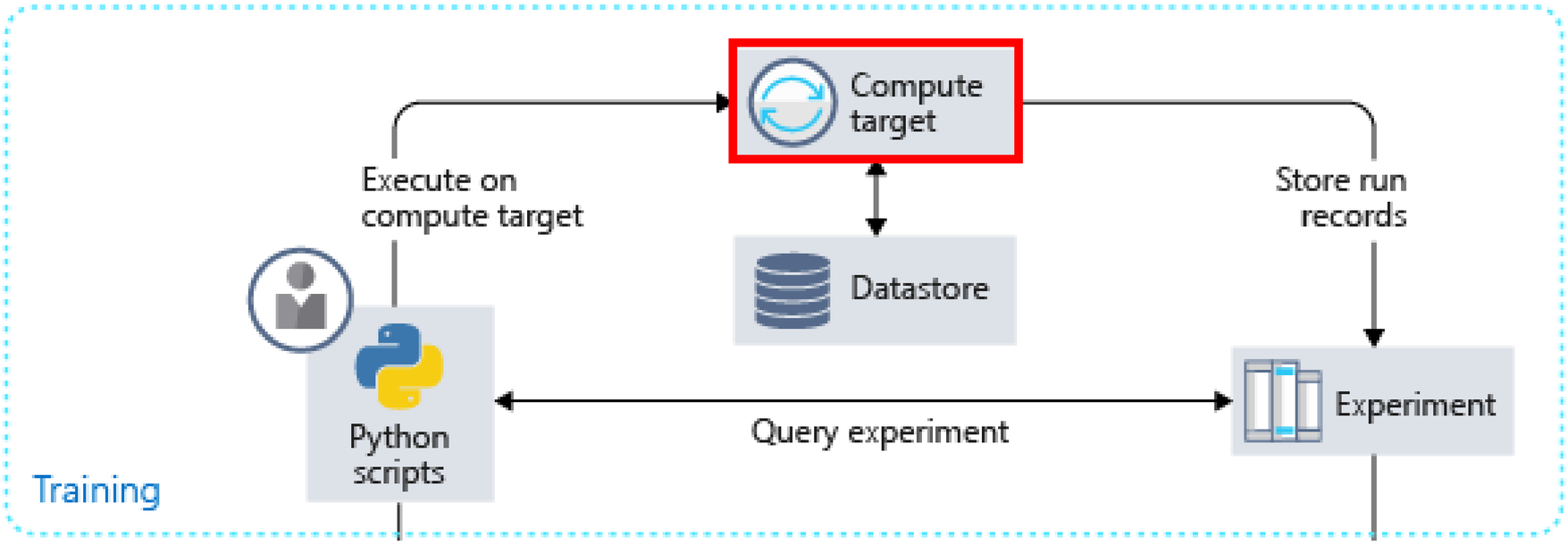
Azure Machine Learning Service - Overview

- Python SDK
- Data preparation
- Compute targets
- Experiment tracking
- Deployment targets

Azure Machine Learning Service - Overview

- Python SDK
- Data preparation
- Compute targets
- Experiment tracking
- Deployment targets

Azure Machine Learning Service - Compute Targets



Azure Machine Learning Service - Compute Targets

- Your local computer

Azure Machine Learning Service - Compute Targets

- Your local computer

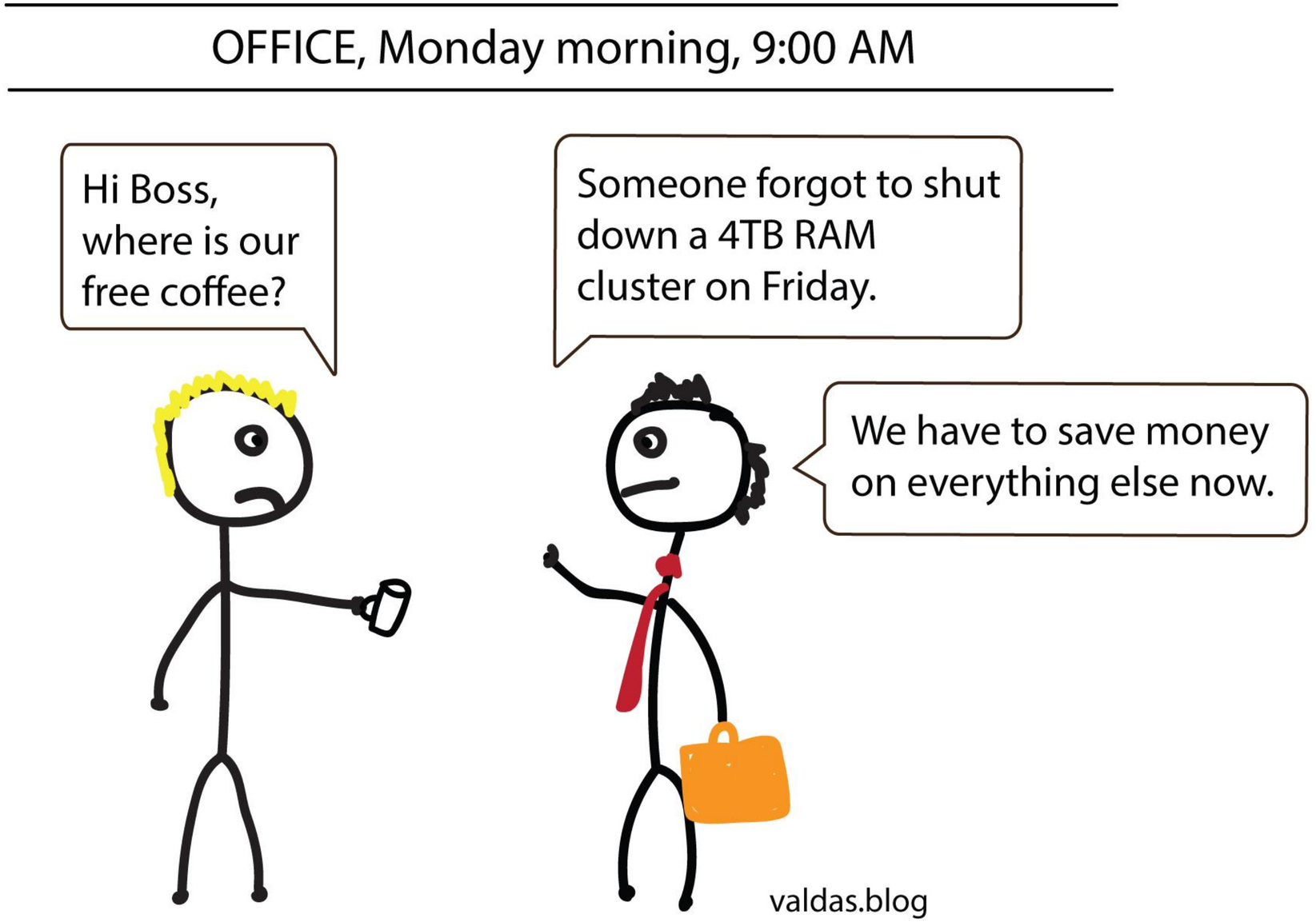
- Linux VM in Azure
- Azure Batch AI Cluster
- Azure Databricks
- Azure Container Instance
- Apache Spark for HDInsight

Azure Machine Learning Service - Compute Targets

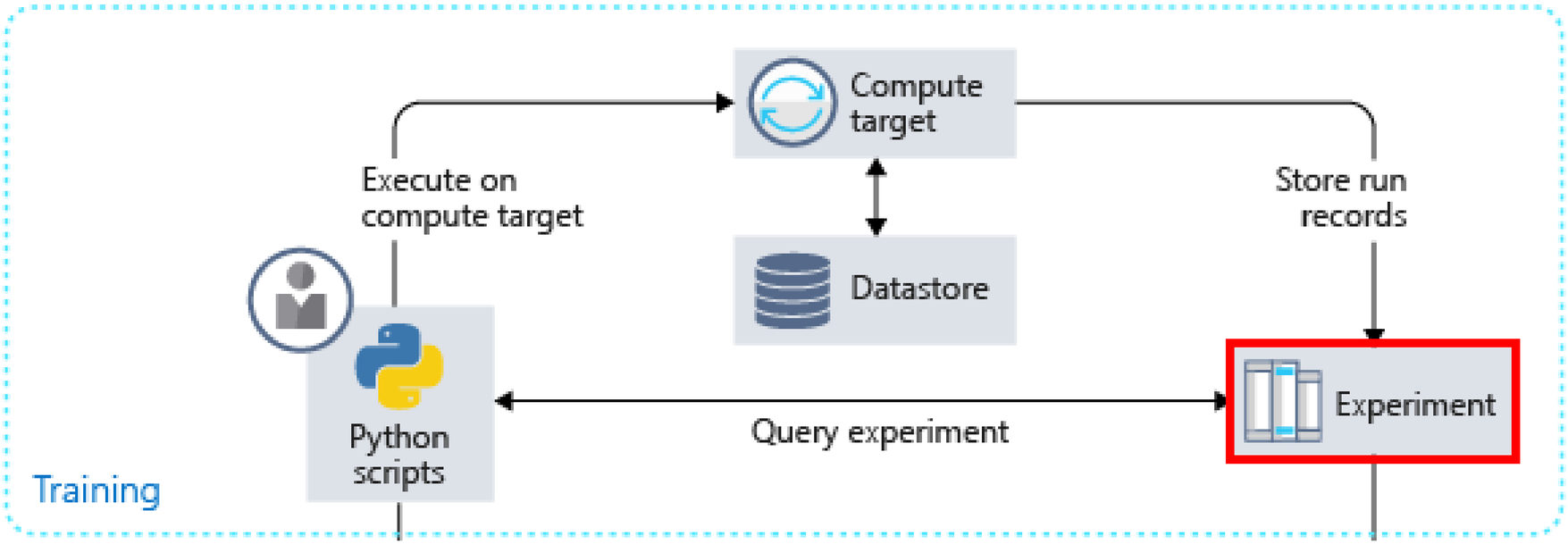
```python
from azureml.core.compute import ComputeTarget, BatchAiCompute
from azureml.core import Workspace
from azureml.train.dnn import PyTorch

# 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)

```


OFFICE, Monday morning, 9:00 AM
Hi Boss, where is our free coffee?
Someone forgot to shut down a 4TB RAM cluster on Friday.
We have to save money on everything else now.
valdas.blog

Azure Machine Learning Service - Experiment Tracking



Azure Machine Learning Service - 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
```

Edit Metadata

```
1 from azureml.train.widgets import RunDetails
2 RunDetails(run).show()
```

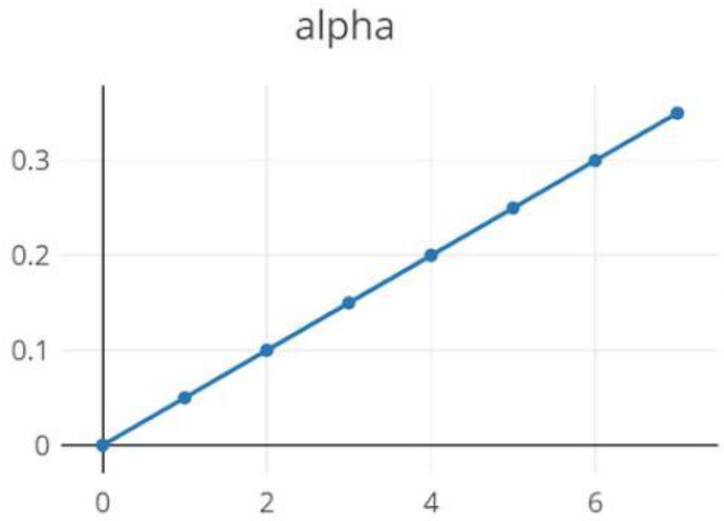
Run Properties

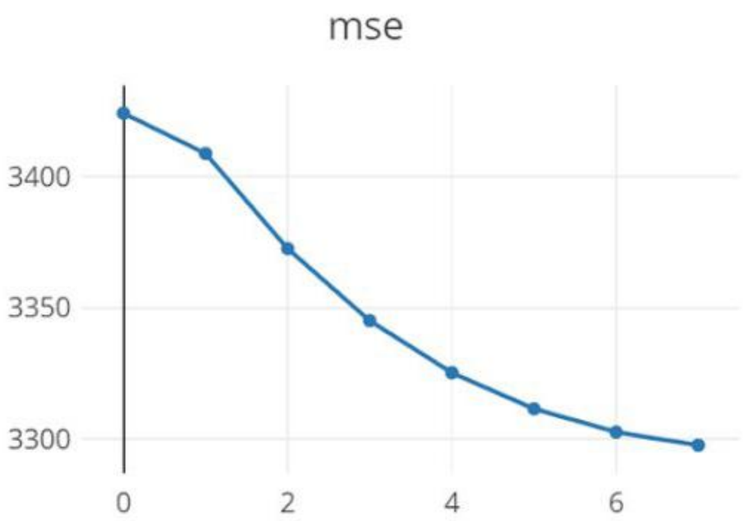
Status	Running
Start Time	9/15/2018 7:15:37 PM
Duration	0:00:20
Run Id	train-on-local_1537053337_839d0780
Arguments	N/A

Output Logs

```
Uploading experiment status to history service.
Adding run profile attachment azureml-logs/80_driver_log.txt

alpha is 0.00, and mse is 3424.32
alpha is 0.05, and mse is 3408.92
alpha is 0.10, and mse is 3372.65
alpha is 0.15, and mse is 3345.15
alpha is 0.20, and mse is 3325.29
alpha is 0.25, and mse is 3311.56
alpha is 0.30, and mse is 3302.67
```





Click here to see the run in Azure portal

Iteration	Pipeline	Iteration metric	Best metric	Status	Duration	Started
0	Scale MaxAbs 1, extra trees	0.93564621	0.93564621	Completed	0:02:45	Sep 5, 2018 9:58 PM
1	Normalizer, kNN	0.98376915	0.98376915	Completed	0:02:44	Sep 5, 2018 9:58 PM
2	Standardize, SGD classifier	0.95410701	0.98376915	Completed	0:02:36	Sep 5, 2018 9:58 PM
3	Standardize, gradient boosting	0.97145517	0.98376915	Completed	0:03:03	Sep 5, 2018 9:58 PM
4	Robust Scaler, SGD classifier	0.86735521	0.98376915	Completed	0:02:24	Sep 5, 2018 9:58 PM

Pages: **1** 2 3 4 Next Last 1 of 4

Azure Machine Learning Service - Deployment Targets

Native support:

- Azure Container Instance
- Azure Kubernetes Service
- Azure IoT Edge

Azure Machine Learning Service - Deployment Targets

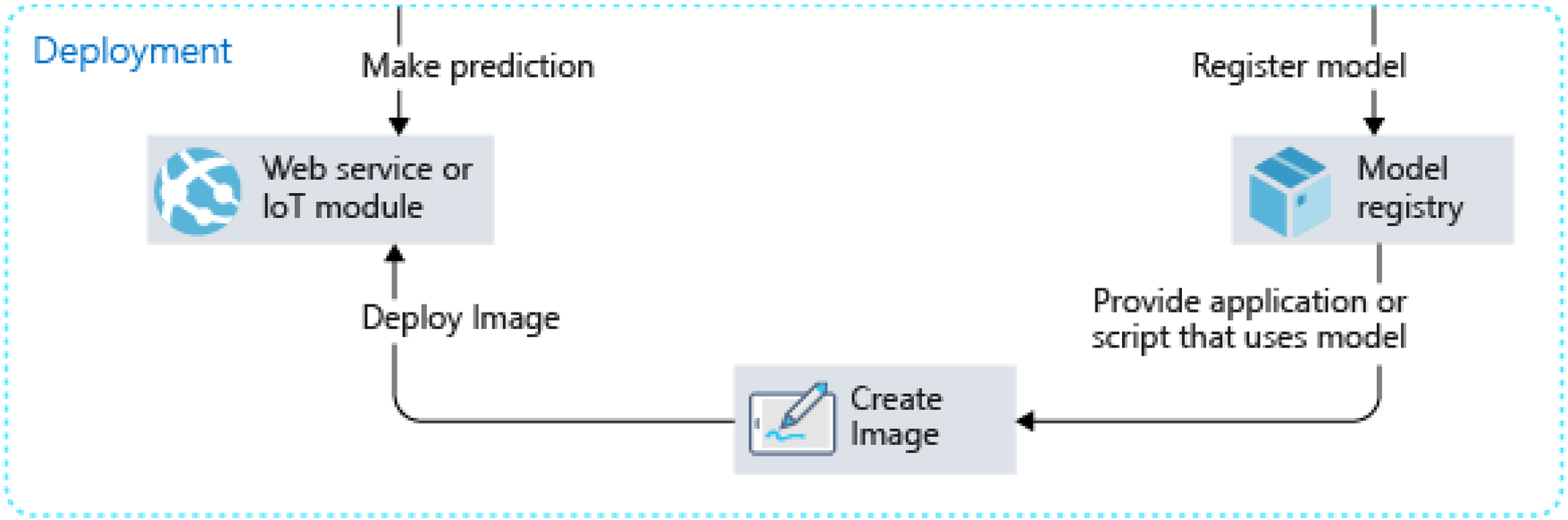
Native support:

- Azure Container Instance
- Azure Kubernetes Service
- Azure IoT Edge

Requires rework:

- Linux VMs
- Other cloud providers

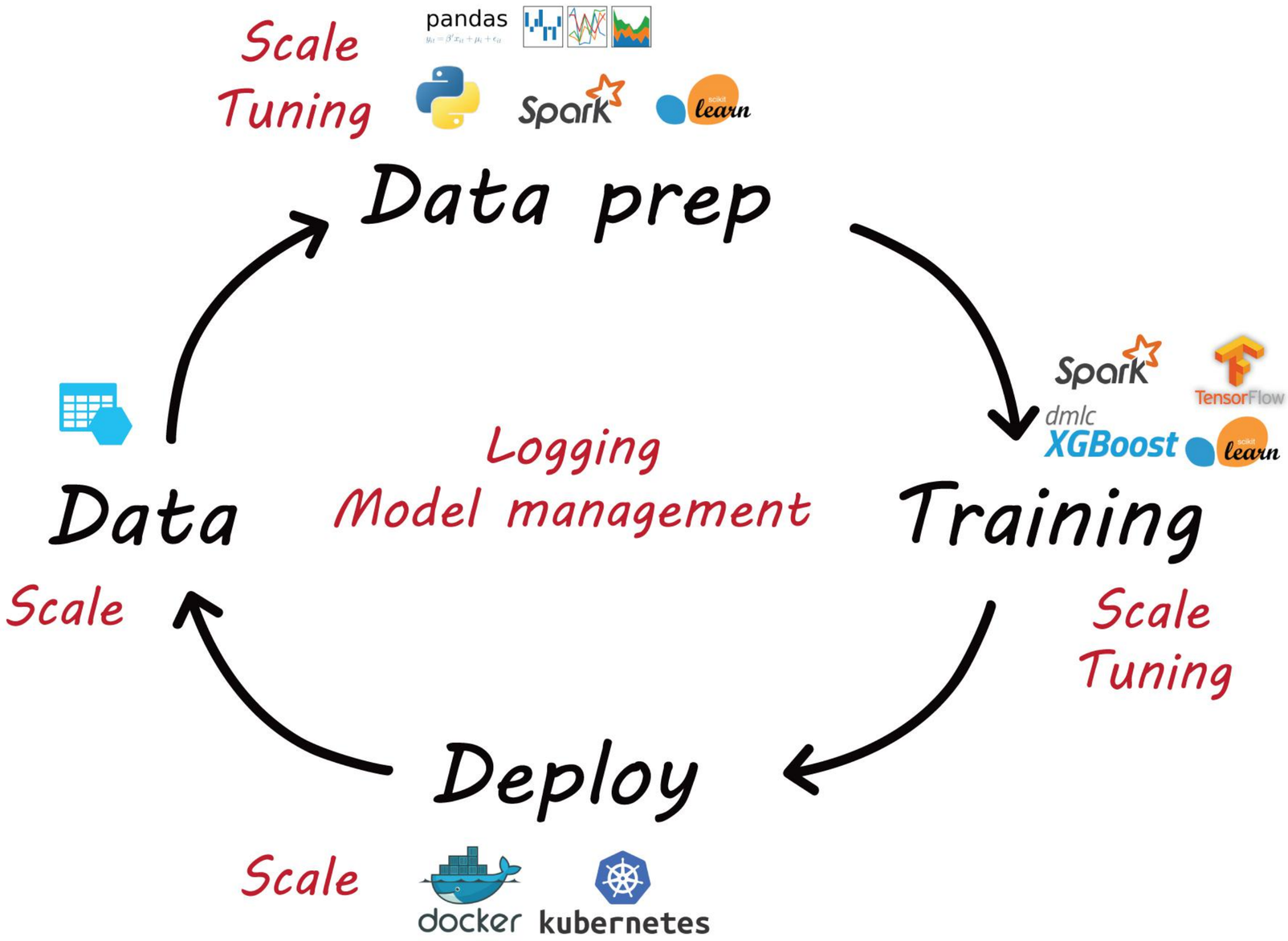
Azure Machine Learning Service - Deployment Targets



Azure Machine Learning Service - Deployment Targets

```
from azureml.core import Workspace
from azureml.core.image import ContainerImage
from azureml.core.webservice import AciWebservice
# Get Azure secrets
ws = Workspace.from_config()
# Save a model in Azure ML workspace
model = Model.register(model_path="local_model.pkl", workspace=ws, ...)
# Specify scoring and Conda dependencies files
image_config = ContainerImage.image_configuration(execution_script="score.py",
                                                  runtime="python",
                                                  conda_file="myenv.yml", ...)
# Create an image in Azure ML workspace
image = ContainerImage.create(name="myimage1",
                              models=[model],
                              image_config=image_config,
                              workspace=ws, ...)
# Create Azure Container Instances
aci_config = AciWebservice.deploy_configuration(cpu_cores=1, memory_gb=1, ...)
# Create a web service based on
service = Webservice.deploy_from_image(deployment_config=aci_config,
                                       image=image,
                                       workspace=ws, ...)

```


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

Azure Machine Learning Service - Summary

Key benefits:

- Central management of scripts and run history
- Run model training scripts locally, and then scale out to the cloud
- Deployment and management of models to the cloud or edge devices
- Start development locally (offline)

Azure Machine Learning Service - Summary

Key benefits:

- Central management of scripts and run history
- Run model training scripts locally, and then scale out to the cloud
- Deployment and management of models to the cloud or edge devices
- Start development locally (offline)

Considerations:

- Still in preview
- Python only

Rick

- Specializes in R
- Not allowed to push data to Azure

- Create personalized treatment based on individual health data

Microsoft Machine Learning ~~Service~~ Server

What is it?

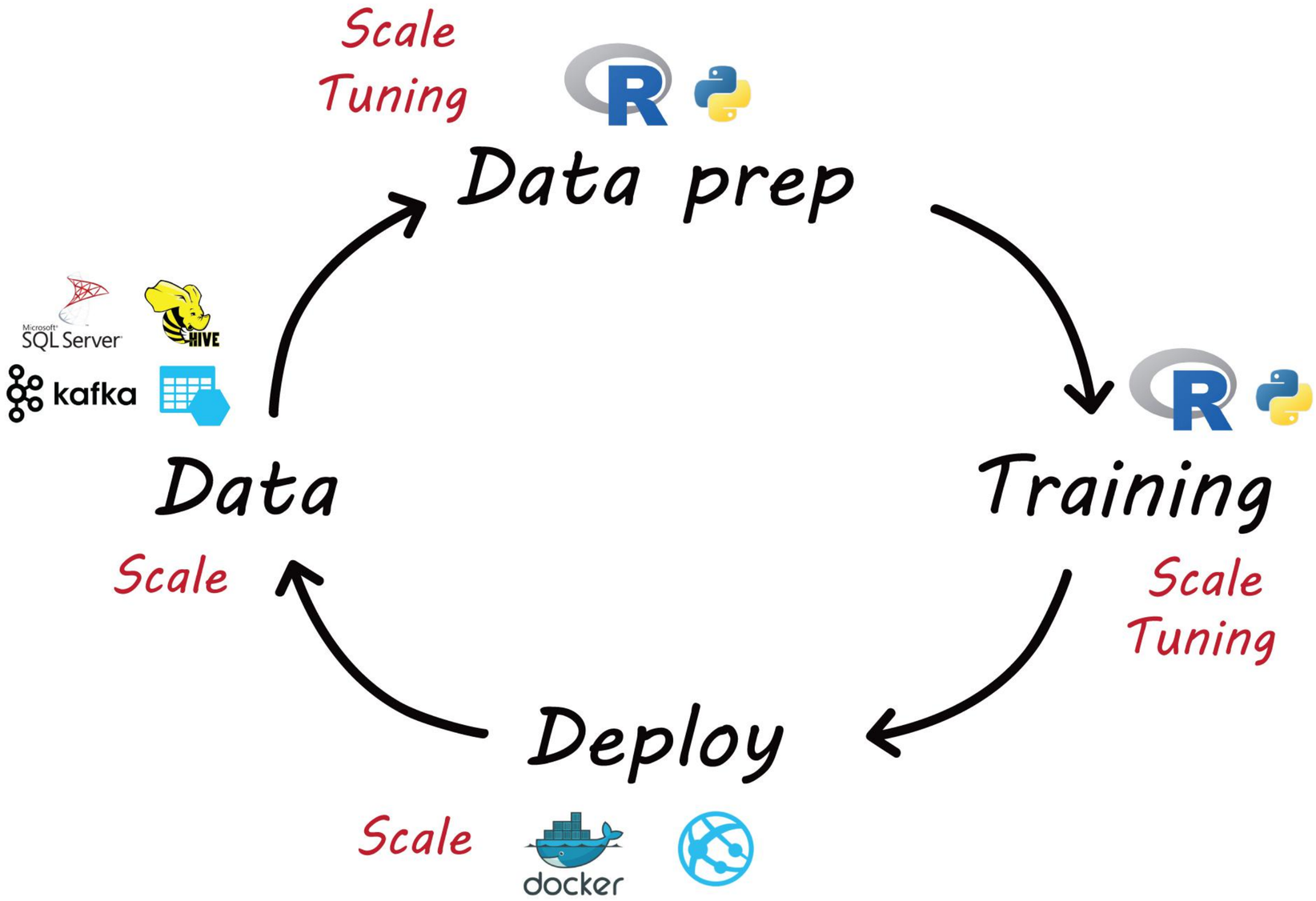
Cross-platform standalone server for predictive analysis

What can you do with it?

Build and deploy models with R and Python

Microsoft Machine Learning Server - 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


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

Microsoft Machine Learning Server - Summary

Key benefits:

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

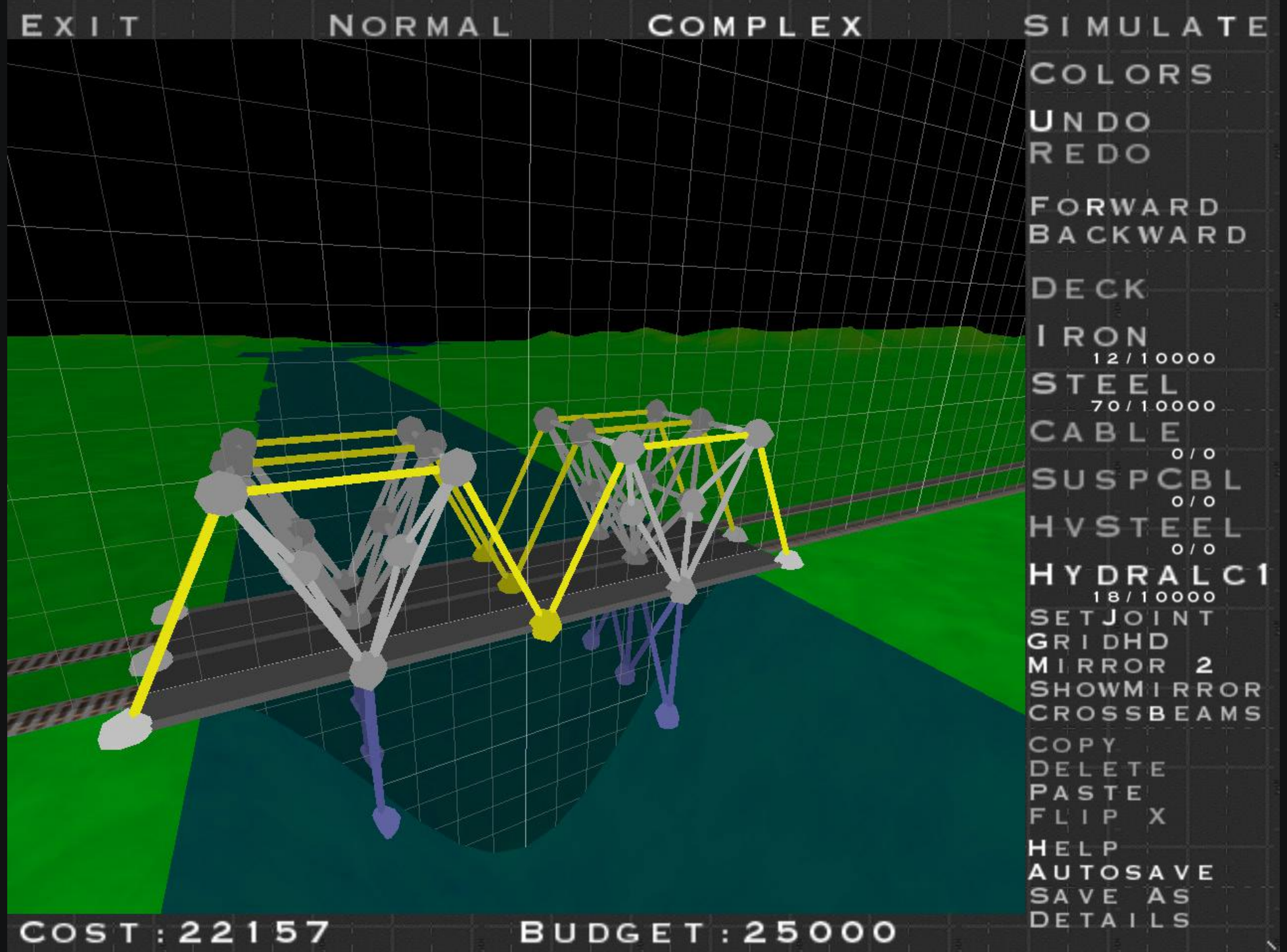
Considerations:

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

Bradley

- Data Scientist
- Apache Spark, SQL
- Wants to spend more time outdoors than exploring new toys

- Create a solution to help structural engineers better monitor the health of bridges


EXIT
NORMAL
COMPLEX
SIMULATE
COLORS
UNDO
REDO
FORWARD
BACKWARD
DECK
IRON
12/10000
STEEL
70/10000
CABLE
0/0
SUSPCBL
0/0
HVSTEEL
0/0
HYDRALC1
18/10000
SETJOINT
GRIDHD
MIRROR 2
SHOWMIRROR
CROSSBEAMS
COPY
DELETE
PASTE
FLIP X
HELP
AUTOSAVE
SAVE AS
DETAILS
COST:22157
BUDGET:25000

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 - Overview

Collaborative Workspace

- Notebooks
- User access
- Git integration

Databricks Runtime

- Apache Spark
- Rest APIs
- Libraries

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

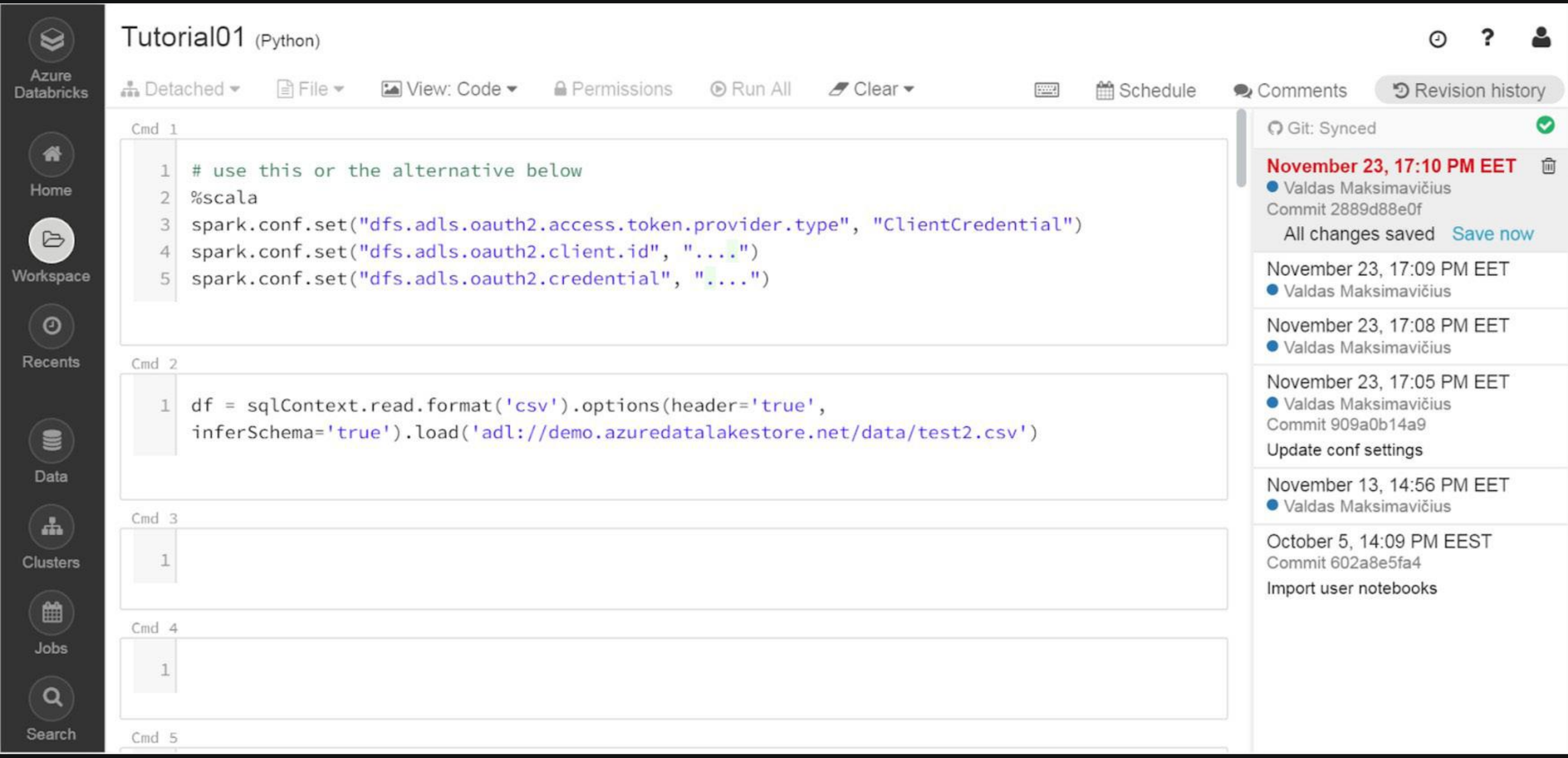
Azure Databricks - Overview

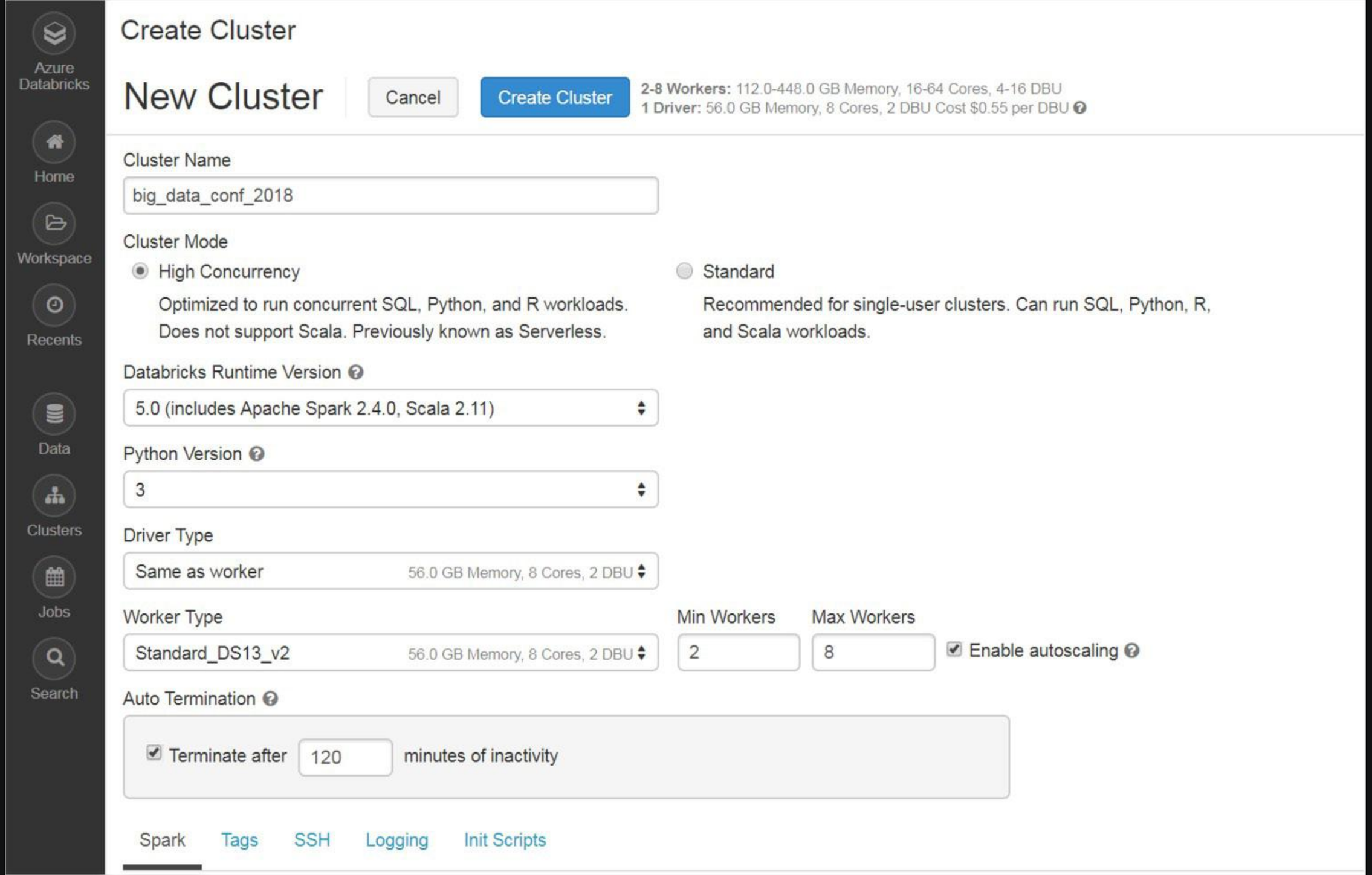
Collaborative Workspace
- Notebooks
- User access
- Git integration

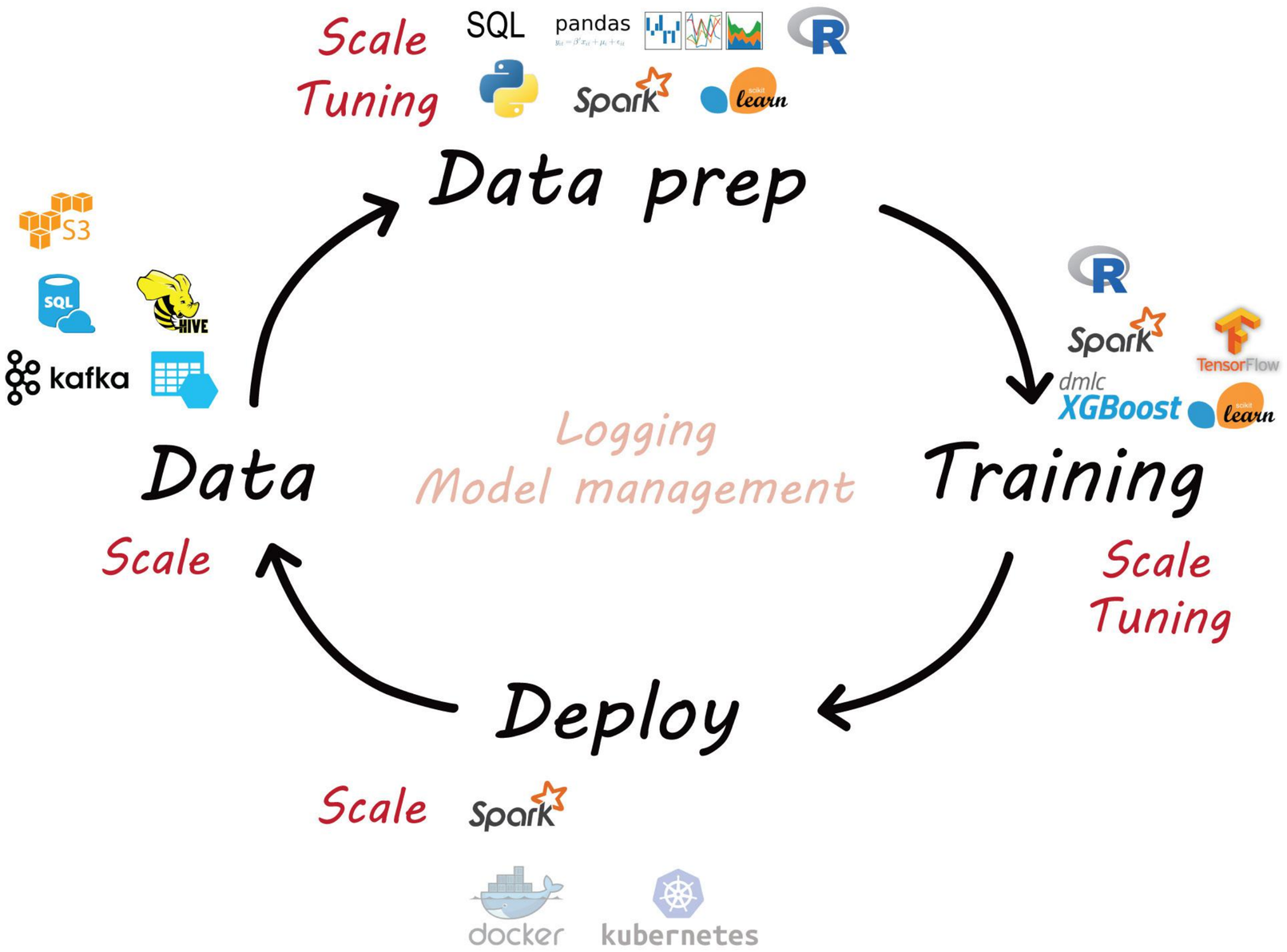
Deploy Jobs & Workflows
- Job scheduler
- Notifications & logs
- Multi-stage pipelines

Databricks Runtime
- Apache Spark
- Rest APIs
- Libraries

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


Azure Databricks
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
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Create Cluster
New Cluster
Cancel
Create Cluster
2-8 Workers: 112.0-448.0 GB Memory, 16-64 Cores, 4-16 DBU
1 Driver: 56.0 GB Memory, 8 Cores, 2 DBU Cost $0.55 per DBU
Cluster Name
big_data_conf_2018
Cluster Mode
High Concurrency
Optimized to run concurrent SQL, Python, and R workloads.
Does not support Scala. Previously known as Serverless.
Standard
Recommended for single-user clusters. Can run SQL, Python, R, and Scala workloads.
Databricks Runtime Version
5.0 (includes Apache Spark 2.4.0, Scala 2.11)
Python Version
3
Driver Type
Same as worker
56.0 GB Memory, 8 Cores, 2 DBU
Worker Type
Standard_DS13_v2
56.0 GB Memory, 8 Cores, 2 DBU
Min Workers
2
Max Workers
8
Enable autoscaling
Auto Termination
Terminate after
120
minutes of inactivity
Spark
Tags
SSH
Logging
Init Scripts


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

Azure Databricks - Summary

Key benefits:

- Probably the most mature development environment for ML on the Azure platform
- Nicely integrated with other Azure services

Azure Databricks - Summary

Key benefits:

- Probably the most mature development environment for ML on the Azure platform
- Nicely integrated with other Azure services

Considerations:

- Online only

Joshua

- Data Scientist
- Research and development

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

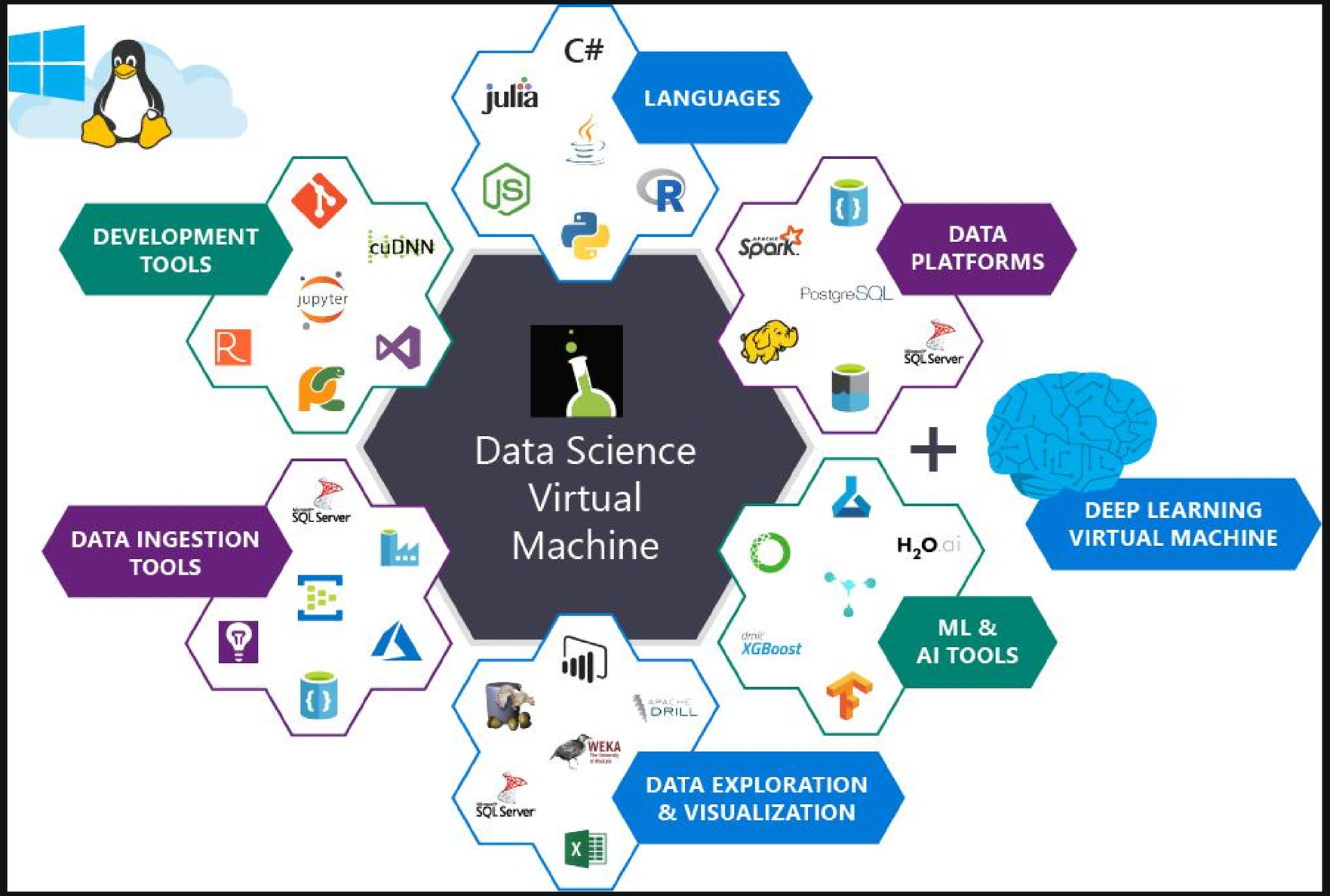
Data Science Virtual Machine

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
DATA EXPLORATION & VISUALIZATION
APACHE DRILL
WEKA
SQL Server
DEEP LEARNING VIRTUAL MACHINE

Data Science Virtual Machine???
It’s not the cloud! It’s just another VM in the datacenter!

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
- Included the latest versions of all commonly used tools and frameworks
- Virtual machine options include highly scalable GPU images

Considerations:

- Online only

A little exam

Azure Machine Learning Studio

Azure Machine Learning Service

(Azure Machine Learning Workbench)

Microsoft Machine Learning Server

(Microsoft R Server)

Data Science Virtual Machine

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