Microsoft

#techsummitCH

Microsoft Tech Summit

Build your skills with the latest in cloud technologies

Microsoft

#techsummitCH

Microsoft Tech Summit

28 February – 1 March 2018, Trafo Baden

#techsummitCH

Scaling AI development with Spark on Azure

Alexandre Gattiker
Cloud Solution Architect

Alberto Arrighi
Cloud Solution Architect

Our strategy is to build best-in-class **platforms** and productivity services for an **intelligent cloud and an intelligent edge** infused with **artificial intelligence** ("AI").

Microsoft
Form 10-K 2016

Microsoft Data+AI solution

Extensible AI services

Open AI tools

Powerful Infrastructure

Rapid time to market with an
agile and productive AI platform

On-premises

Edge

Cloud

AI Built-in

Gain transformative insights with a
comprehensive platform for your data estate

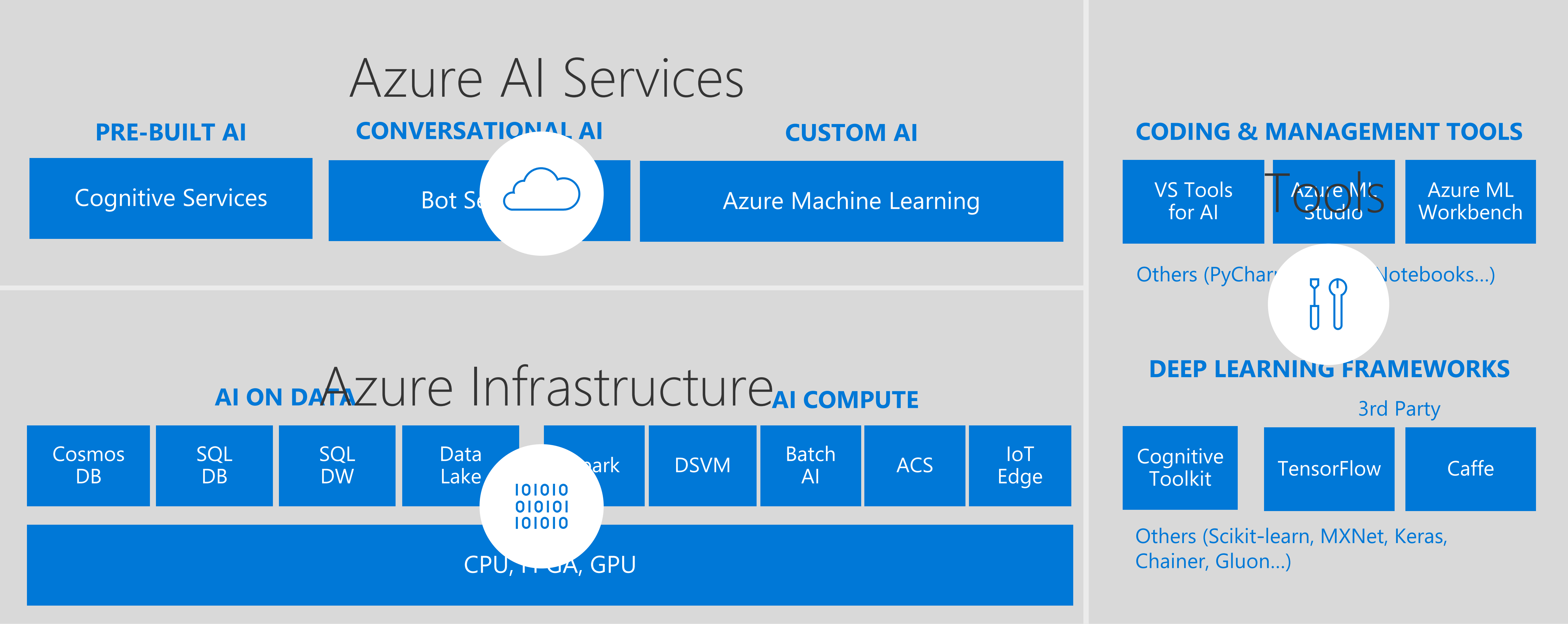
The Trusted Cloud

AI Business Solutions

Over 650,000 Partners

Innovate with confidence with
enterprise-proven solutions

Microsoft AI Platform



Azure AI Services

PRE-BUILT AI

Cognitive Services

CONVERSATIONAL AI

Bot Se

CUSTOM AI

Azure Machine Learning

Azure Infrastructure

AI ON DATA

Cosmos DB

SQL DB

SQL DW

Data Lake

AI COMPUTE

ark

DSVM

Batch AI

ACS

IoT Edge

CPU, FPGA, GPU

Tools

CODING & MANAGEMENT TOOLS

VS Tools for AI

Azure ML Studio

Azure ML Workbench

Others (PyChar Notebooks...)

DEEP LEARNING FRAMEWORKS

3rd Party

Cognitive Toolkit

TensorFlow

Caffe

Others (Scikit-learn, MXNet, Keras, Chainer, Gluon...)

Tools

Visual Studio Tools for AI

Boost productivity with code-centric AI development and Azure integration.

Azure Machine Learning Workbench

Full lifecycle support for AI and data wrangling productivity.

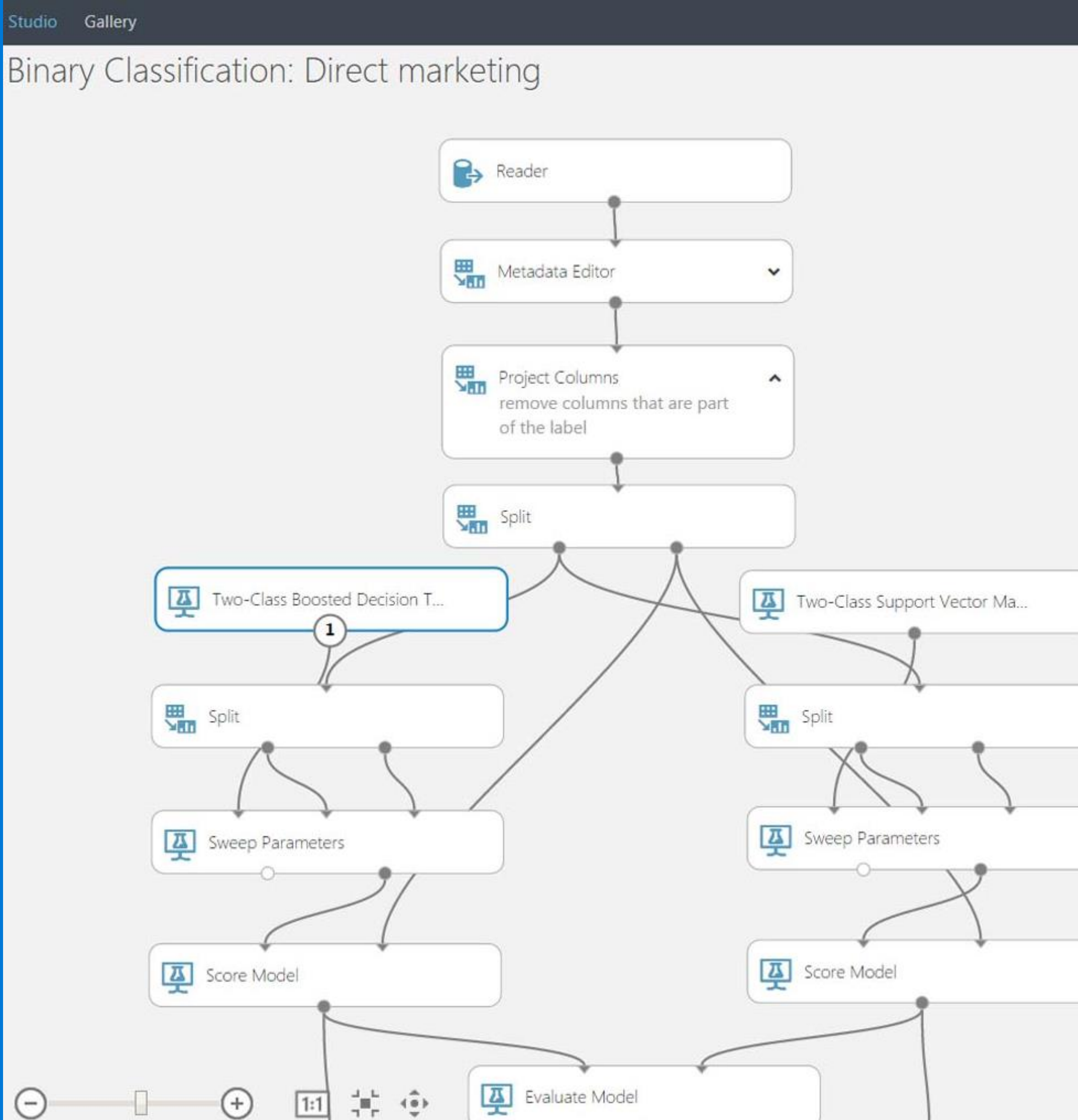
Azure Machine Learning Studio

Drag and drop machine learning development for any skillset.

Open deep learning framework support

Full support for Cognitive Toolkit, TensorFlow, Caffee and others.

Open standard for deep learning (ONNX).


Studio
Gallery
Binary Classification: Direct marketing
Reader
Metadata Editor
Project Columns
remove columns that are part of the label
Split
Two-Class Boosted Decision T...
1
Two-Class Support Vector Ma...
Split
Split
Sweep Parameters
Sweep Parameters
Score Model
Score Model
Evaluate Model
1:1

DATA SCIENCE & AI

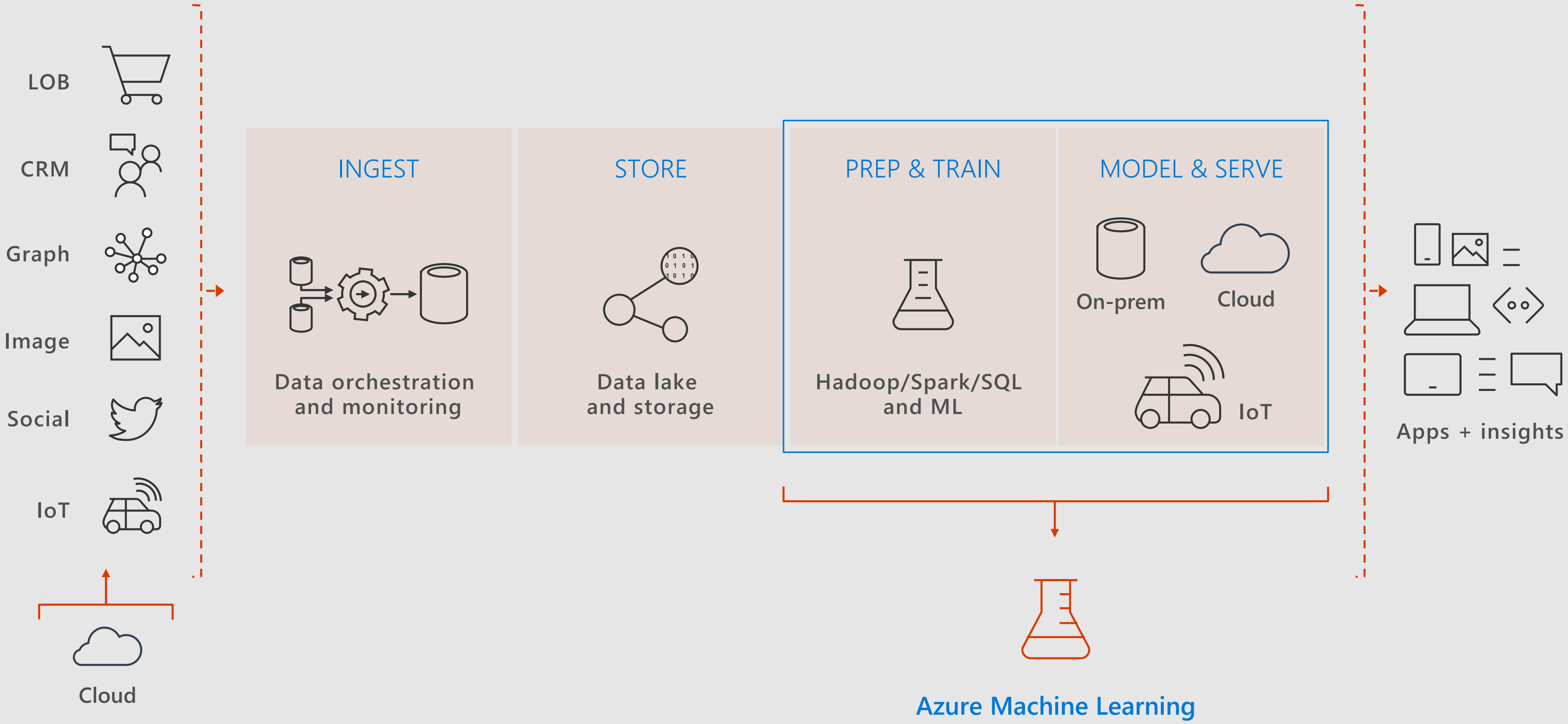
KEY TRENDS

- Accelerating adoption of AI by developers (consuming models)
- Rise of hybrid training and scoring scenarios
- Push scoring/inference to the event (edge, cloud, on-prem)
- Some developers moving into deep learning as non-traditional path to DS / AI dev
- Growth of diverse hardware arms race across all form factors (CPU / GPU / FPGA / ASIC / device)

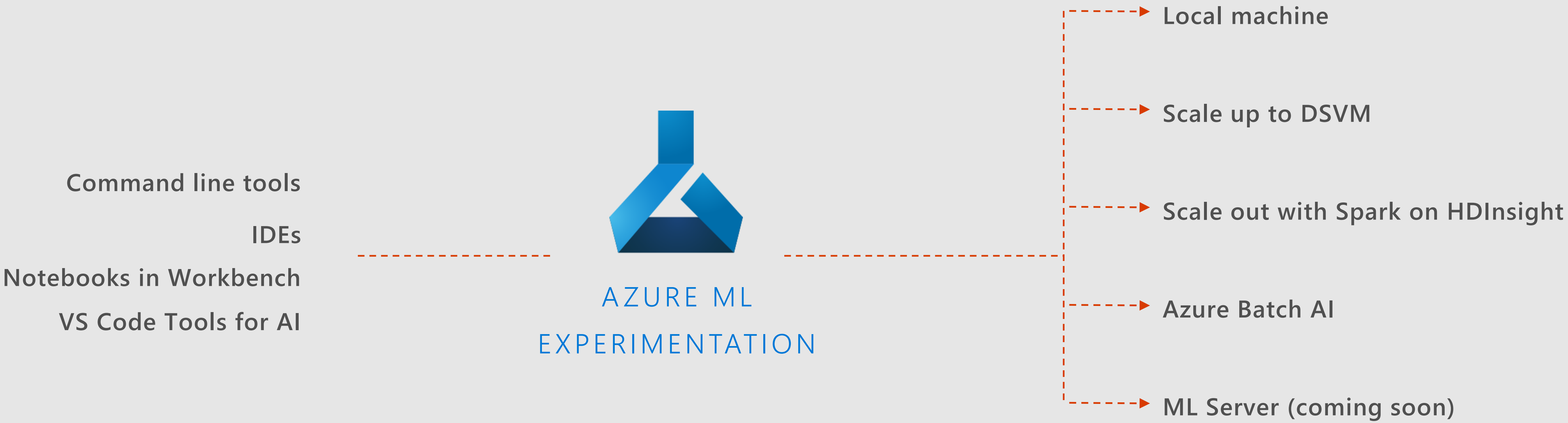
CHALLENGES

- ⚠ Data prep
- ⚠ Model deployment & management
- ⚠ Model lineage & auditing
- ⚠ Explain-ability

THE AI DEVELOPMENT LIFECYCLE



Experiment Everywhere



Command line tools

IDEs

Notebooks in Workbench

VS Code Tools for AI

AZURE ML EXPERIMENTATION

Local machine

Scale up to DSVM

Scale out with Spark on HDInsight

Azure Batch AI

ML Server (coming soon)

Experimentation service

Manage project dependencies

Manage training jobs locally, scaled-up or scaled-out

Git based checkpointing and version control

Service side capture of run metrics, output logs and models

Use your favorite IDE, and any framework

USE ANY FRAMEWORK OR LIBRARY

TensorFlow

Microsoft CNTK

APACHE Spark ML

USE ANY TOOL

CODE

Visual Studio

jupyter

PyCharm

USE THE MOST POPULAR INNOVATIONS

git

APACHE Spark

docker

python™

R

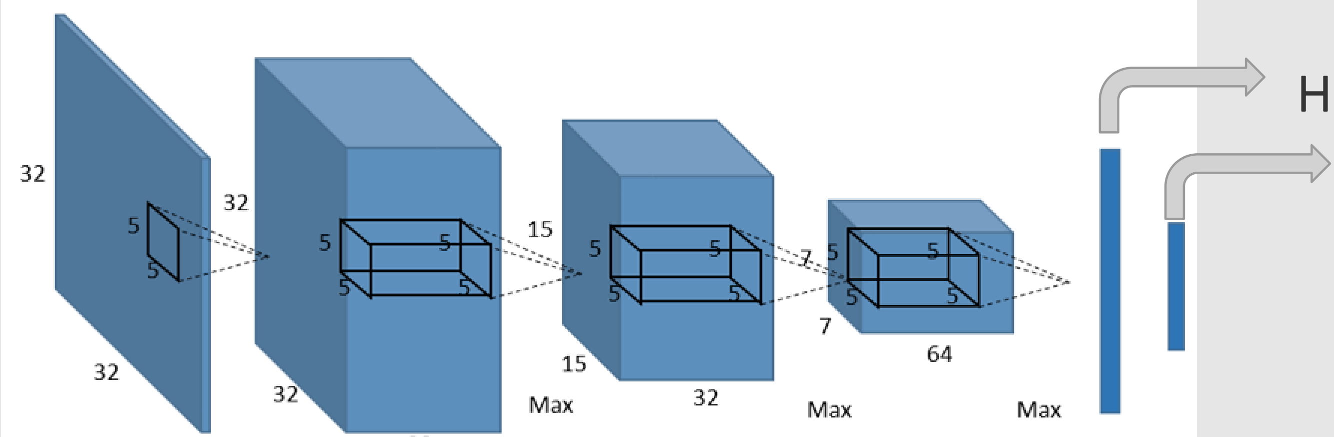
MMLSpark open-source library

- Deep learning through Microsoft Cognitive Toolkit (CNTK)
  - Scale-out DNN featurization and scoring. Take an existing DNN model or train locally on big GPU machine, and deploy it to Spark cluster to score large data.
  - Scale-up training on edge node GPUs. Preprocess large data on Spark cluster workers and feed to GPU to train the DNN.

- Scale-out algorithms for “traditional” ML through SparkML

Deep Neural Net Featurization

Basic idea: Interior layers of pre-trained DNN models have high-order information about features



High-order features

Predictions

Using “headless” pre-trained DNNs allows us to extract really good set of features from images that can in turn be used to train more “traditional” models like random forests, SVM, logistic regression, etc.

- Pre-trained DNNs are typically state-of-the-art models on datasets like ImageNet, MSCoco or CIFAR, for example ResNet (Microsoft), GoogLeNet (Google), Inception (Google), VGG, etc.

Transfer learning enables us to train effective models where we don’t have enough data, computational power or domain expertise to train a new DNN *from scratch*

Performance *scales with executors*

DNN Featurization using MML-Spark

```
cntkModel = CNTKModel().setInputCol("images").
       setOutputCol("features").setModelLocation(resnetModel).
       setOutputNode("z.x")

featurizedImages = cntkModel.transform(imagesWithLabels).
       select(['labels','features'])

model = TrainClassifier(model=LogisticRegression(),labelCol="labels").
       fit(featurizedImages)
```

The DNN featurization is incorporated as SparkML pipeline stage. The evaluation happens directly on JVM from Scala: no Python UDF overhead!

Image Processing Transforms

DNNs are often picky about their input data shape and normalization.

We provide bindings to OpenCV image processing operations, exposed as SparkML PipelineStages:

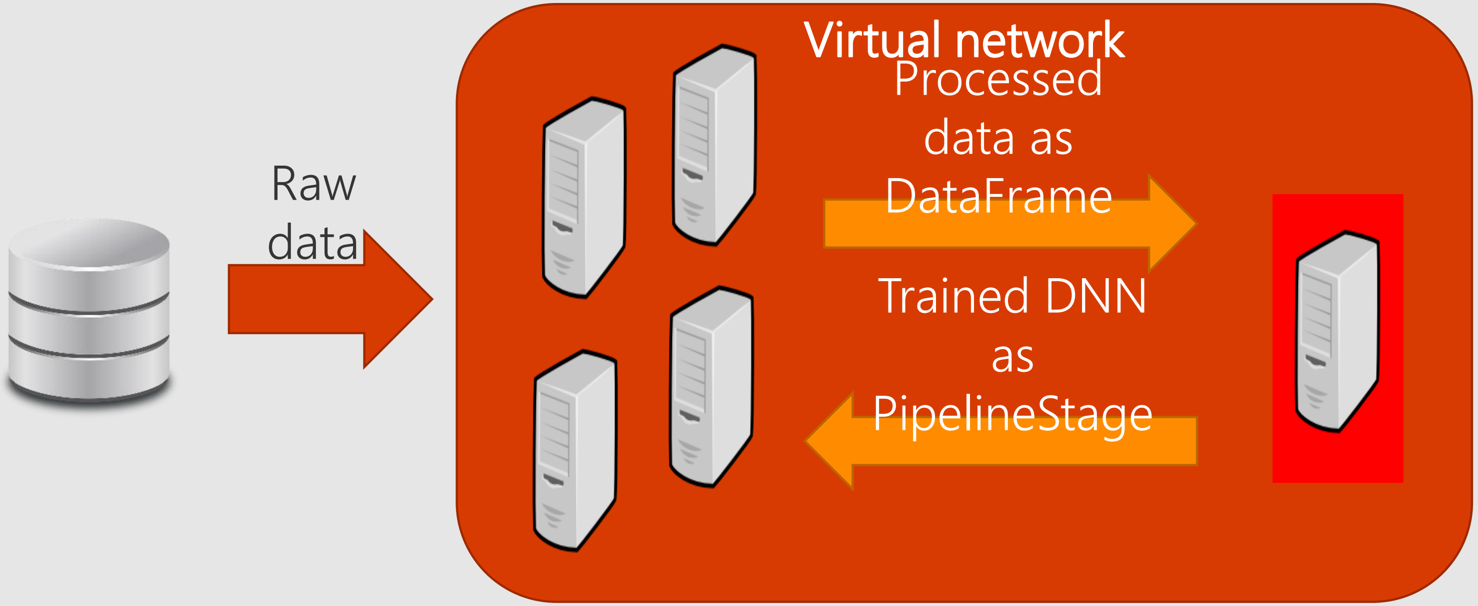
```
tr = ImageTransform().setOutputCol("transformed")
    .resize(height = 200, width = 200)
    .crop(0, 0, height = 180, width = 180)

smallImages = tr.transform(images).select("transformed")
```

Training of DNNs on GPU node

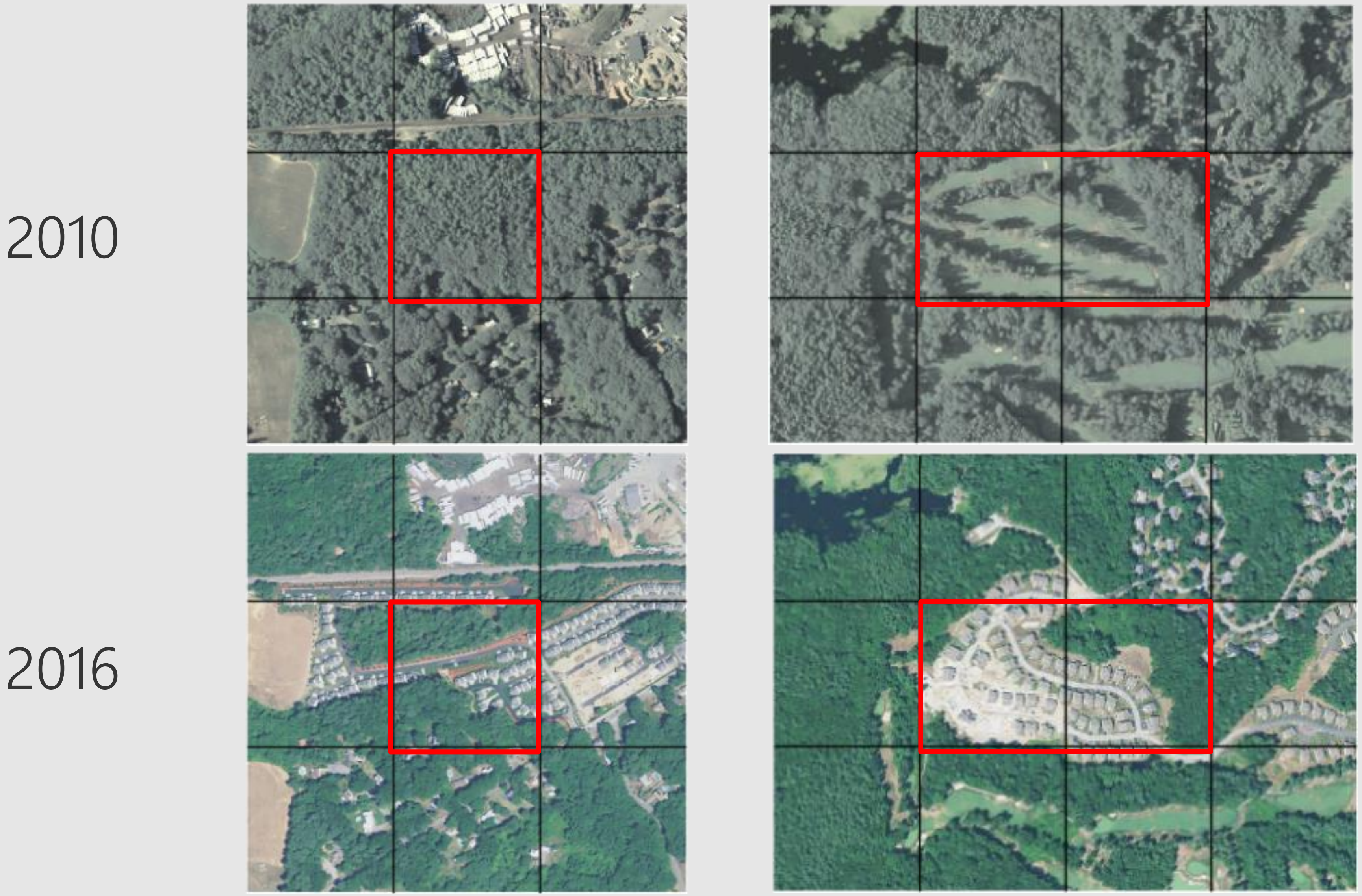
GPUs are very powerful for training DNNs. However, running an entire cluster of GPUs is often too expensive and unnecessary.

Instead, load and prep large data on CPU Spark cluster, then feed the prepped data to GPU node on virtual network for training. Once DNN is trained, broadcast the model to CPU nodes for evaluation.

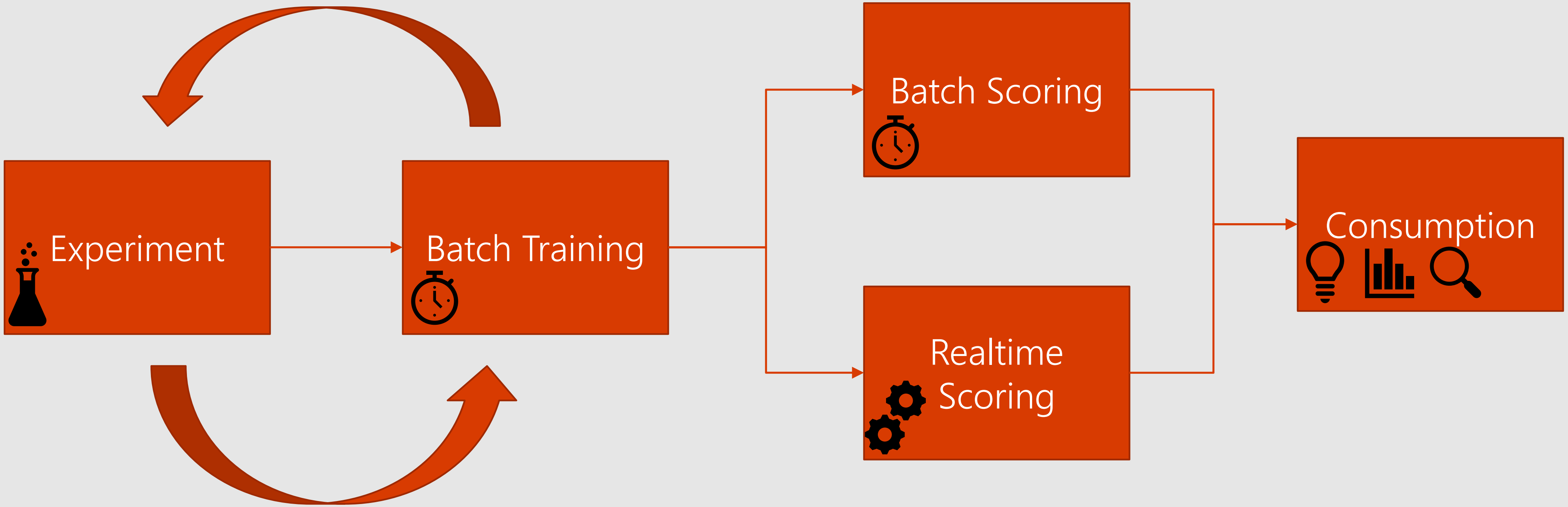
```
learner = CNTKLearner(brainScript=brainscriptText, dataTransfer='hdfs-mount',
    gpuMachines=‘my-gpu-vm’, workingDir=‘file:/tmp/’).fit(trainData)
predictions = learner.setOutputNode(‘z’).transform(testData)
```



Application: Finding newly-developed regions



High level workflow



Execution environments

Model Training

Azure Batch

Azure Batch AI
Spin up a cluster with hundreds of GPUs to train quickly, then tear down when finished

For fastest distributed learning: copy data to cluster-adjacent NFS

Model Training / Batch Scoring

APACHE Spark™

HDInsight Spark
Apply the trained model to large, static datasets in your Azure Data Lake Store

Real-Time Predictions

Create a web service
Incorporate real-time predictions from the model into your applications

Incoming data are uploaded to web server for scoring

DATA → INTELLIGENCE → ACTION

Demo

https://github.com/Azure/MachineLearningSamples-AerialImageClassification

What’s next

- Go and get started!
- Learn more!
- Tell us what you think!

http://aka.ms/aml_deep_dive

Microsoft

#techsummitCH

Please Complete your Session Evaluations

Get your cool IoT Dev Kit!

Fill out your feedback form and turn it in before you leave.

Microsoft