Well.. that’s ONNX-pected

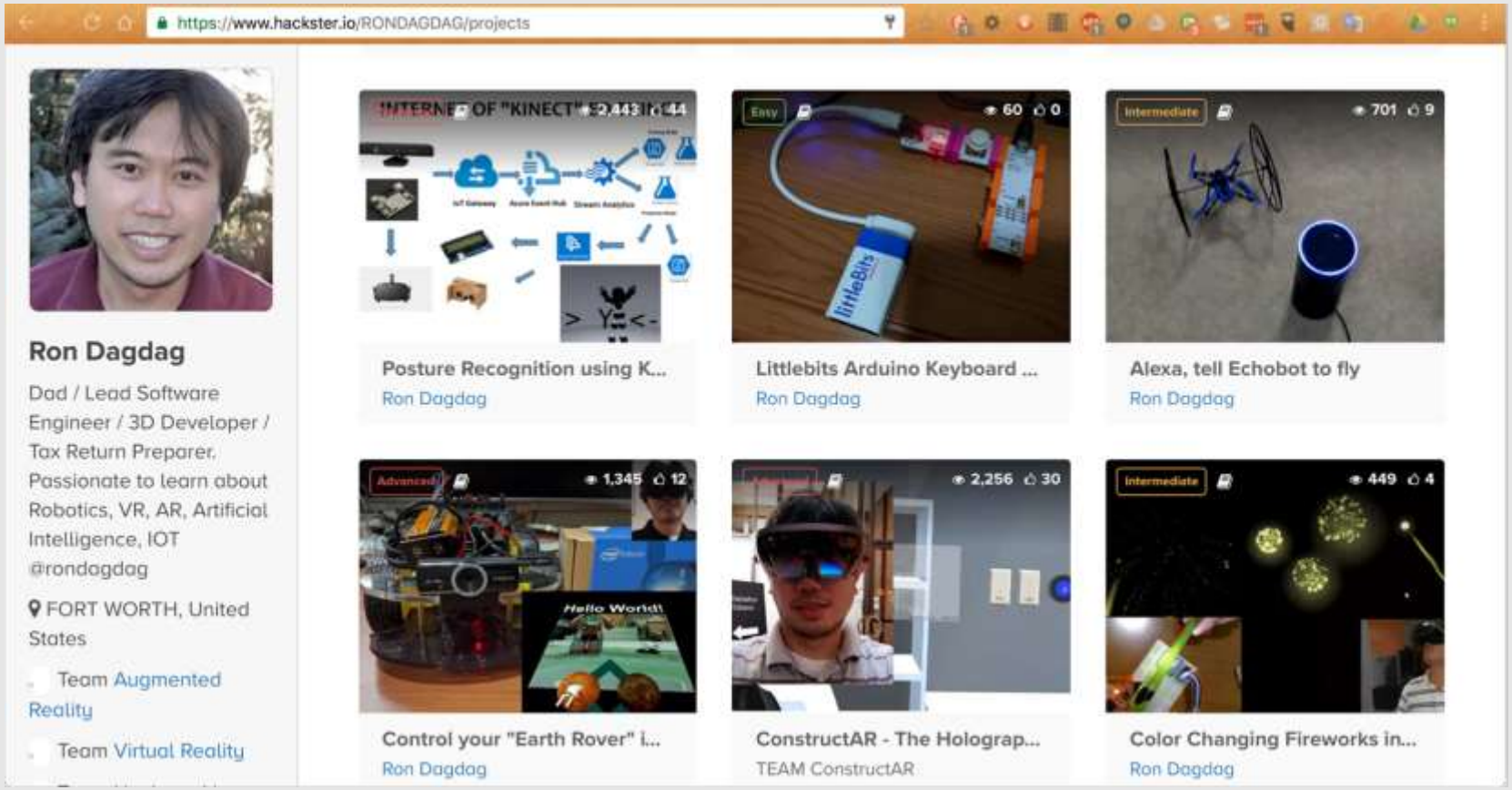
“…there are no wrong turns, only unexpected paths.”

http://bit.ly/onnxpected

Hackster Portfolio

www.dagdag.net

@rondagdag

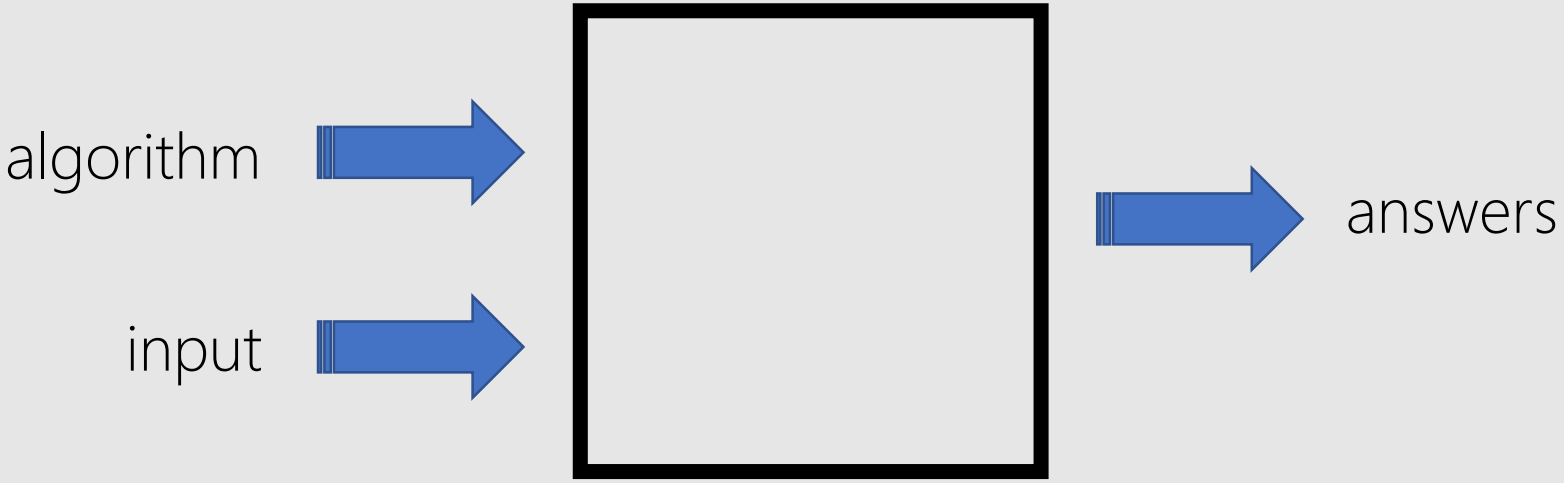


Audience Survey:

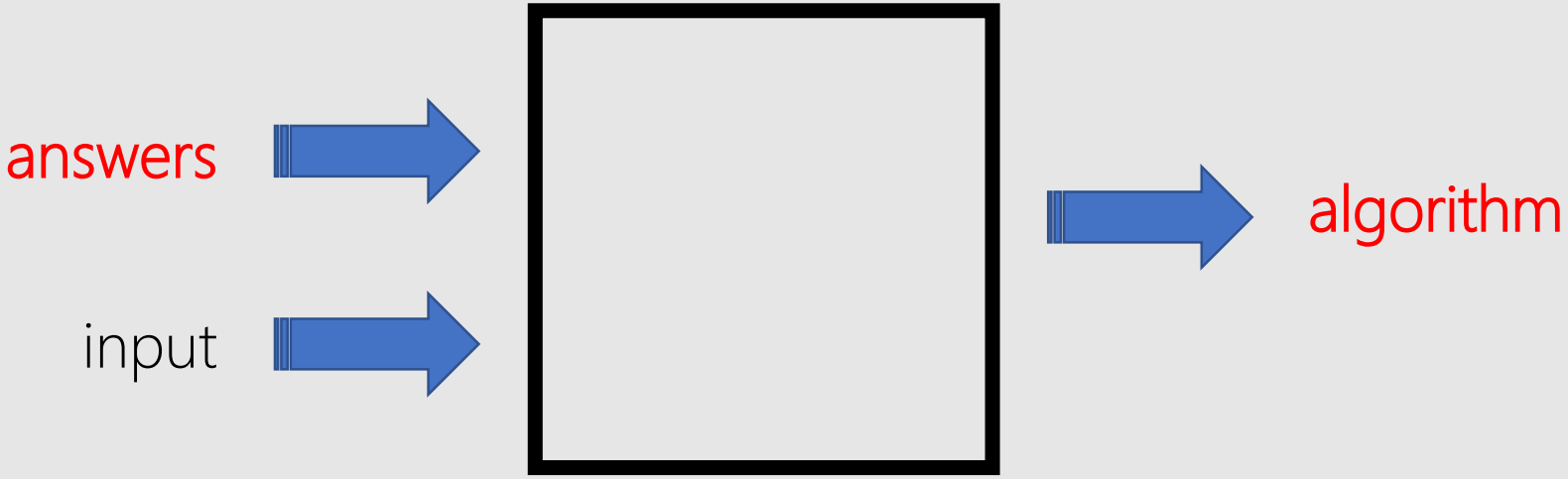
- .NET Developers
- JavaScript Developers
- Python Developers
- Data Scientists
- Data Engineers

- Pokemon Trainers?

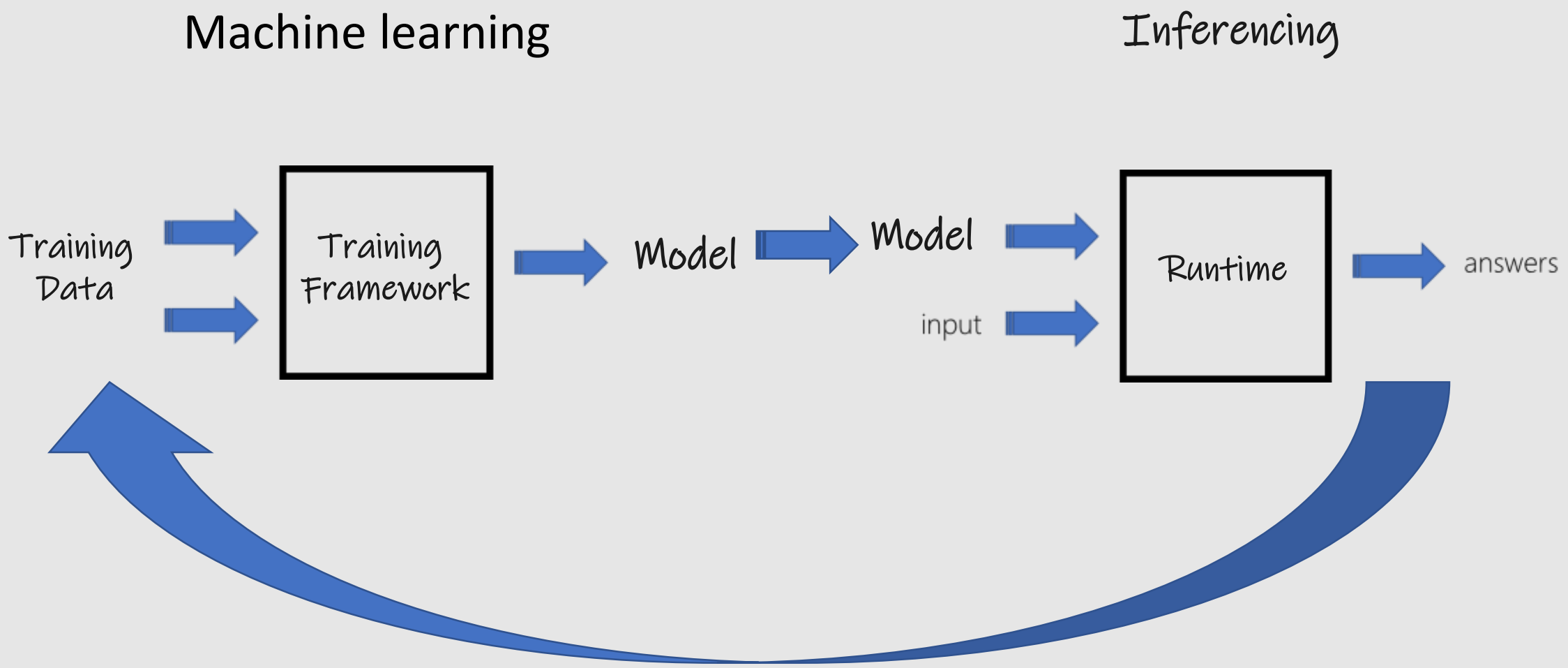
programming



machine learning

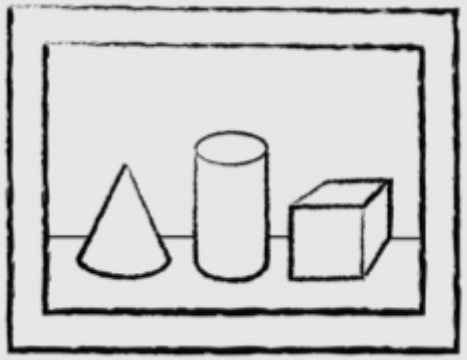


ML Primer


Machine learning
Inferencing
Training Data
Training Framework
Model
Model
input
Runtime
answers

process

identify

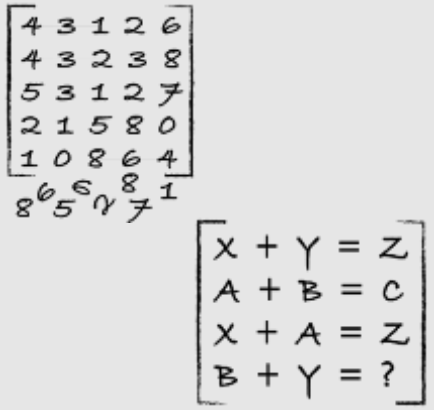


1

predict

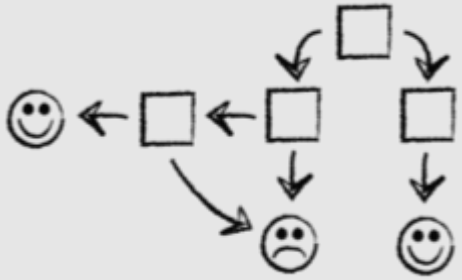
4

explore
analyze
encode



2

model



3

Open and Interoperable AI

ONNX

Open Neural Network Exchange

Open format for ML models

github.com/onnx

ONNX Partners

aws

Facebook Open Source

Microsoft

AMD

arm

Baidu 百度

BITMAIN

Hewlett Packard Enterprise

HUAWEI

IBM

intel AI

MathWorks

MEDIATEK

NVIDIA

Oath: A Verizon company

Preferred Networks

QUALCOMM

skymizer

Tencent

unity

Agenda

- ✓ What is ONNX?
- ❑ Create ONNX models
- ❑ Deploy ONNX models

Create

Frameworks

Caffe2 · Chainer · Cognitive Toolkit · mxnet · PyTorch · PaddlePaddle · ML.NET · MathWorks · dmlc XGBoost · ML · scikit learn · TF · K

Services

Azure Custom Vision Service

Native support

Converters

Native support

ONNX Model

Native support

Converters

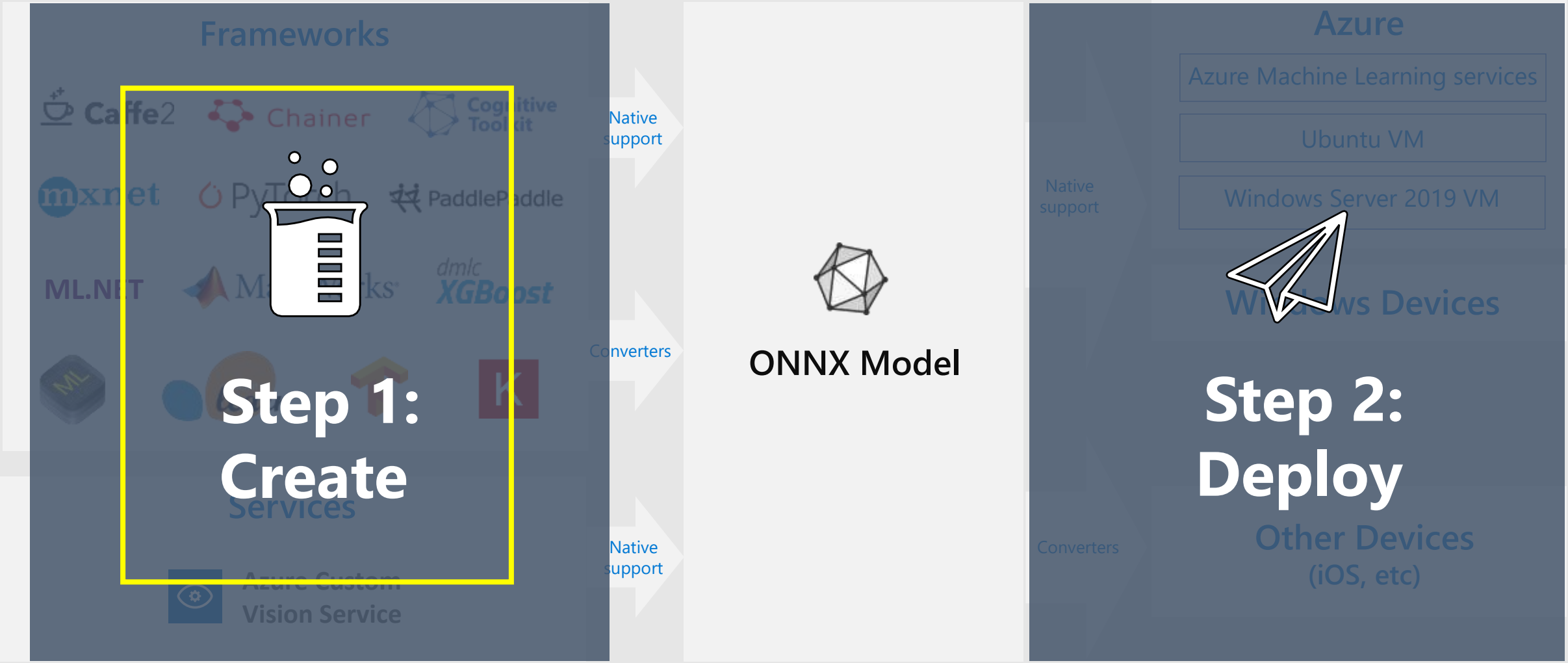
Deploy

Azure

- Azure Machine Learning services
- Ubuntu VM
- Windows Server 2019 VM

Windows Devices

Other Devices
(iOS, Android, etc)


Frameworks
Caffe2
Chainer
Cognitive Toolkit
mxnet
PyTorch
PaddlePaddle
ML.NET
MATLAB
dmlc XGBoost
Step 1: Create
Services
Azure Custom Vision Service
Native support
Converters
Native support
ONNX Model
Azure
Azure Machine Learning services
Ubuntu VM
Windows Server 2019 VM
Native support
Windows Devices
Step 2: Deploy
Converters
Other Devices
(iOS, etc)

Secret Recipe

4 ways to get an ONNX model

ONNX Model Zoo

Custom Vision Service

Convert existing models

Train models in Azure Machine Learning

Automated Machine Learning

ONNX Model Zoo: github.com/onnx/models

Image Classification

This collection of models take images as input, then classifies the major objects in the images into a set of predefined classes.

Model Class	Reference	Description
MobileNet	Sandler et al.	Efficient CNN model for mobile and embedded vision applications. Top-5 error from paper - ~10%
ResNet	He et al., He et al.	Very deep CNN model (up to 152 layers), won the ImageNet Challenge in 2015. Top-5
SqueezeNet	Iandola et al.	A ligh fewer Top-5
VGG	Simonyan et al.	Deep Challe Top-5

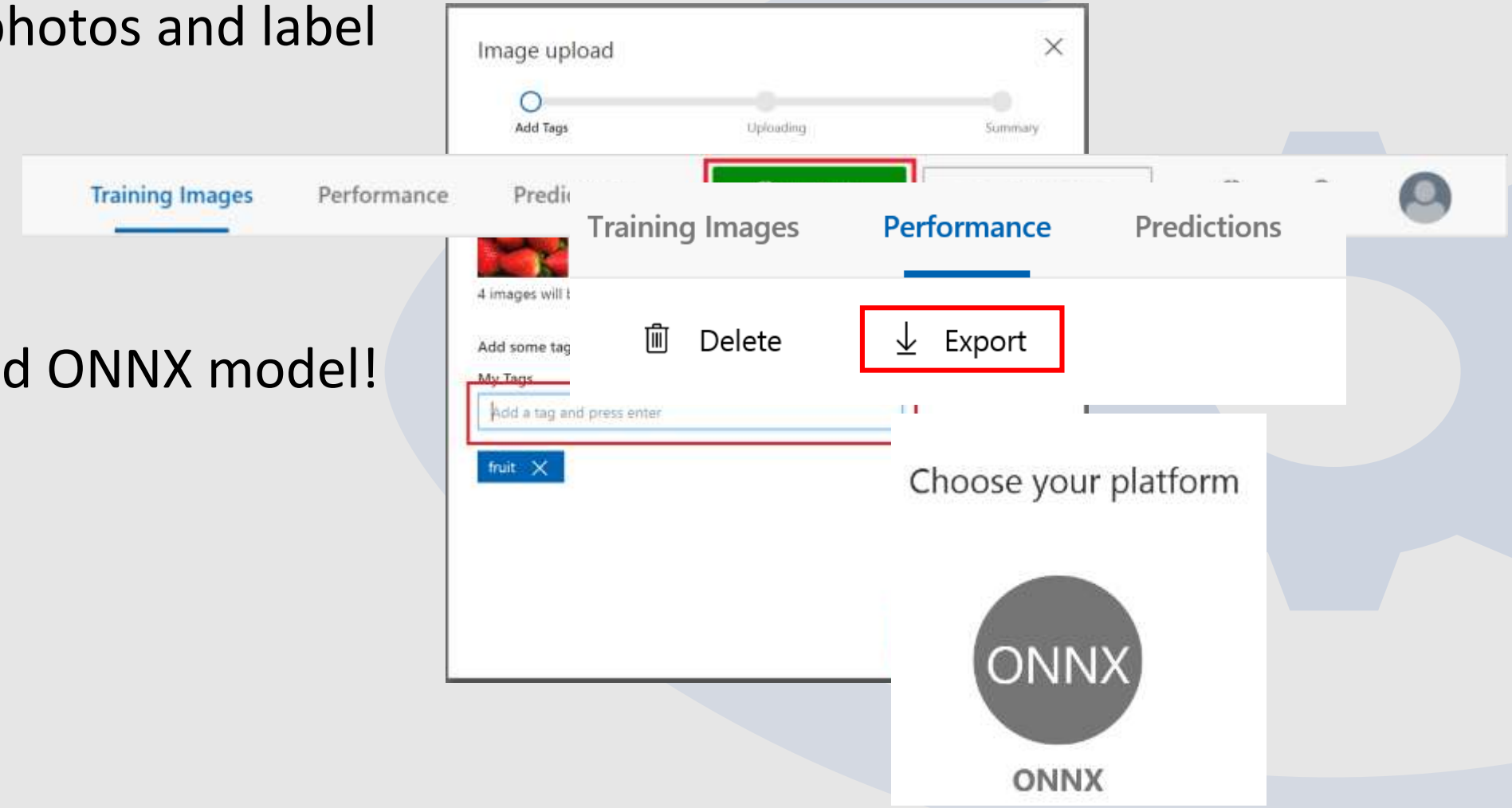
Model	Download	Checksum	Download (with sample test data)	ONNX version	Opset version	Top-1 accuracy (%)	Top-5 accuracy (%)
ResNet-18	44.6 MB	MD5	42.9 MB	1.2.1	7	69.70	89.49
ResNet-34	83.2 MB	MD5	78.6 MB	1.2.1	7	73.36	91.43
ResNet-50	97.7 MB	MD5	92.0 MB	1.2.1	7	75.81	92.82
ResNet-101	170.4 MB	MD5	159.4 MB	1.2.1	7	77.42	93.61
ResNet-152	230.3 MB	MD5	216.0 MB	1.2.1	7	78.20	94.21

Custom Vision Service: customvision.ai

1. Upload photos and label

2. Train

3. Download ONNX model!



Convert models

Convert models

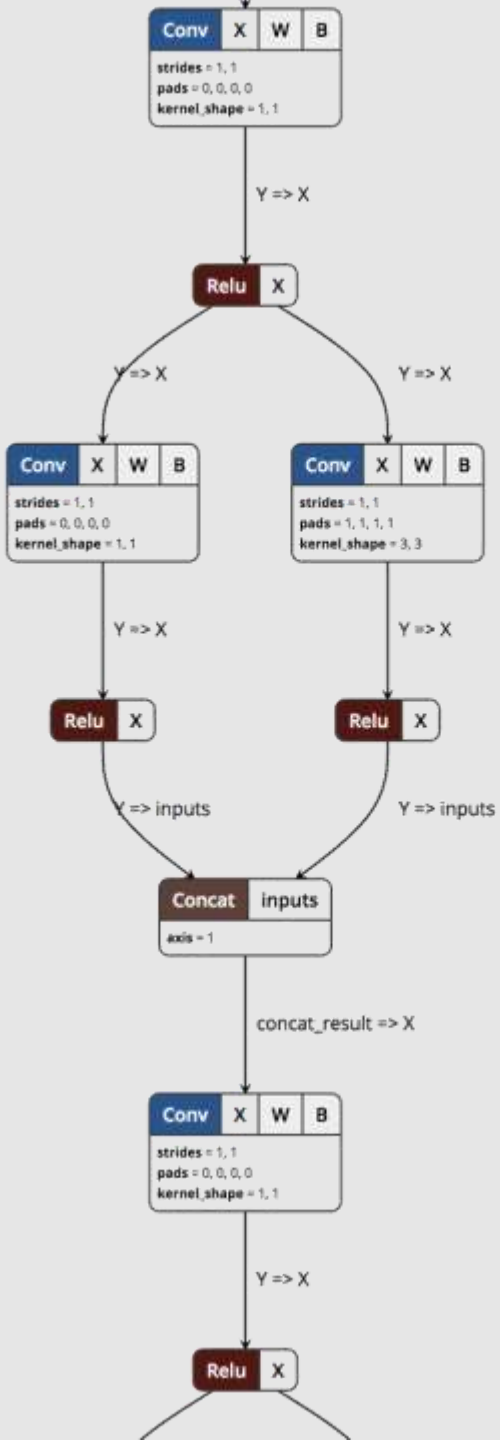
1. Load existing model

2. (Convert to ONNX)

3. Save ONNX model

ONNX Models

Graph of operations



WinML Dashboard

https://github.com/microsoft/Windows-Machine-Learning/releases

ONNX Ecosystem Docker Image

- onnx-ecosystem: Jupyter notebook environment for getting started quickly with ONNX models, ONNX converters, and inference using ONNX Runtime.

```
> docker pull onnx/onnx-ecosystem
> docker run -p 8888:8888 onnx/onnx-ecosystem
```

http://127.0.0.1:8888/?token={tokenId}

http://127.0.0.1:8888/tree/converter_scripts

ONNX

Convert models: Keras

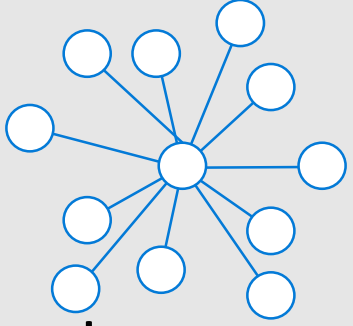
```
from keras.models import load_model
import winmltools

keras_model = load_model("model.h5")

onnx_model = winmltools.convert_keras(keras_model)

winmltools.utils.save_model(onnx_model, "model.onnx")
```

Convert models: TensorFlow

```
> pip install -U tf2onnx
> python -m tf2onnx.convert --input frozen.pb --inputs X:0
                            --outputs output:0 --output model.onnx
```

Learn more at
https://github.com/onnx/tensorflow-onnx

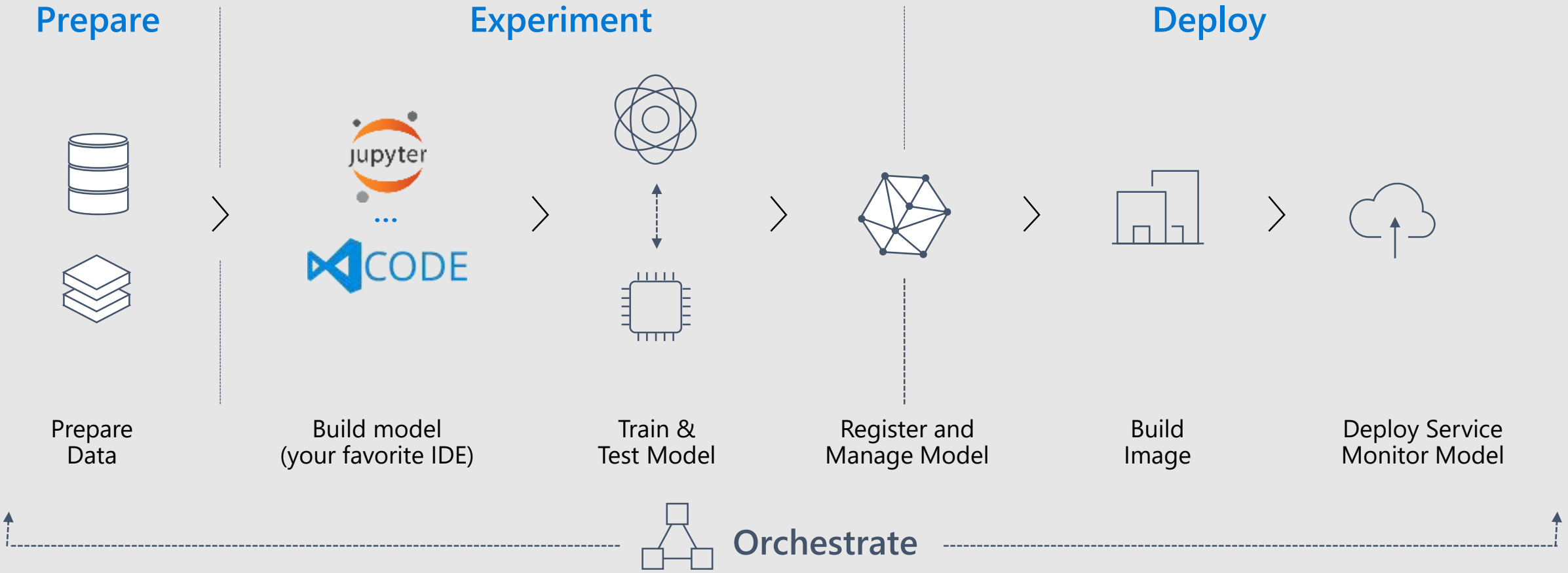
Train models in Azure Machine Learning



- Experiment locally then quickly scale with GPU clusters in the cloud
- Use automated machine learning and hyper-parameter tuning.
- Keeping Track of experiments, manage models, and easily deploy with integrated CI/CD tooling
- Clone Azure ML Gallery Samples
- https://notebooks.azure.com/import/gh/Azure/MachineLearningNotebooks/

Machine Learning

Typical E2E Process



high dimensional matrices

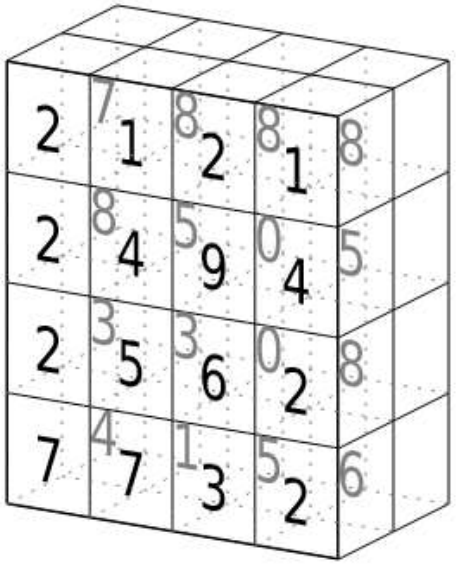
tensor

't'
'e'
'n'
's'
'o'
'r'

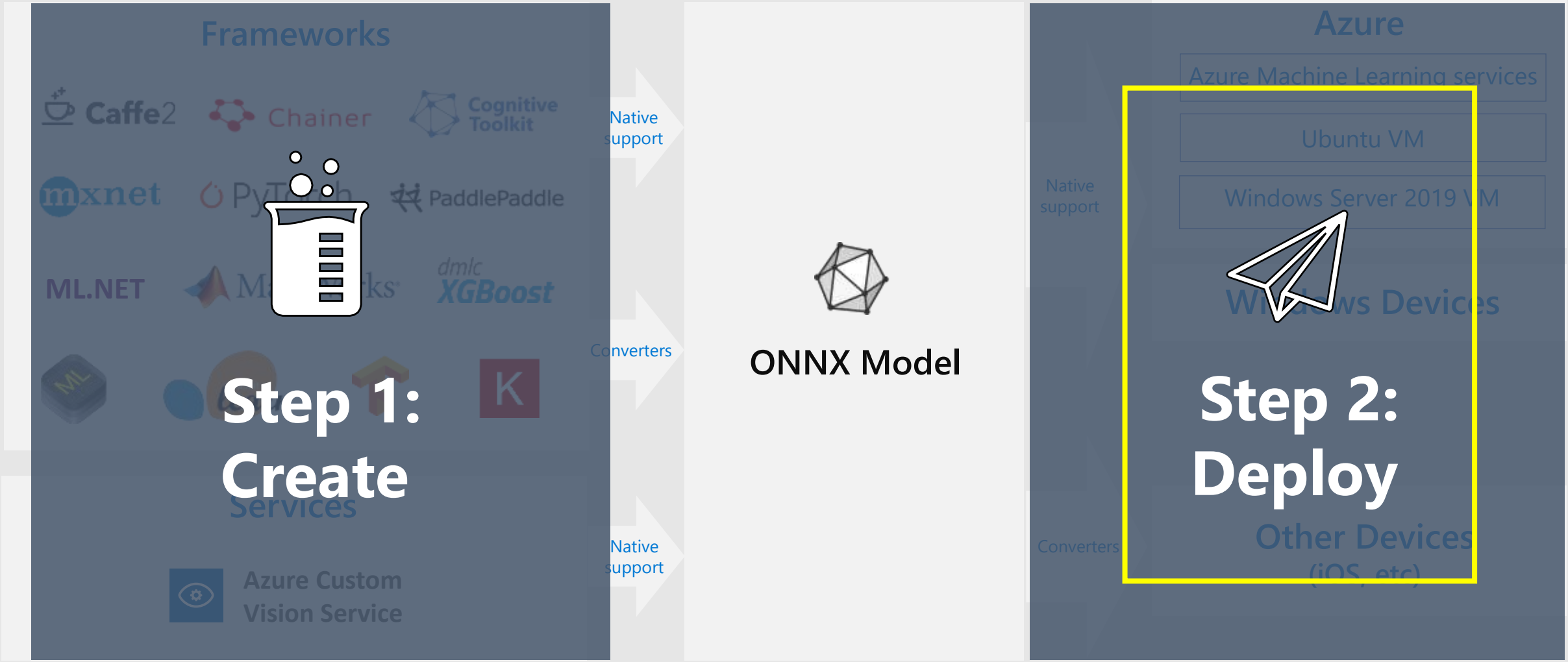
tensor of dimensions [6]
(vector of dimension 6)

3	1	4	1
5	9	2	6
5	3	5	8
9	7	9	3
2	3	8	4
6	2	6	4

tensor of dimensions [6,4]
(matrix 6 by 4)



tensor of dimensions [4,4,2]


Frameworks
Caffe2
Chainer
Cognitive Toolkit
mxnet
PyTorch
PaddlePaddle
ML.NET
dmlc XGBoost
Step 1: Create
Services
Azure Custom Vision Service
Native support
Converters
Native support
ONNX Model
Azure
Azure Machine Learning services
Ubuntu VM
Windows Server 2019 VM
Native support
Windows Devices
Step 2: Deploy
Converters
Other Devices (iOS, etc)

Baker vs Starting a Bakery

Create

Frameworks

Caffe2 | Chainer | Cognitive Toolkit

mxnet | PyTorch | PaddlePaddle

ML.NET | MathWorks | dmlc XGBoost

ML | scikit learn | TF | K

Services

Azure Custom Vision Service

Native support

Converters

Native support

ONNX Model

Native support

Converters

Deploy

Azure

- Azure Machine Learning services
- Ubuntu VM
- Windows Server 2019 VM

Windows Devices

Other Devices
(iOS, etc)

Cloud or Edge

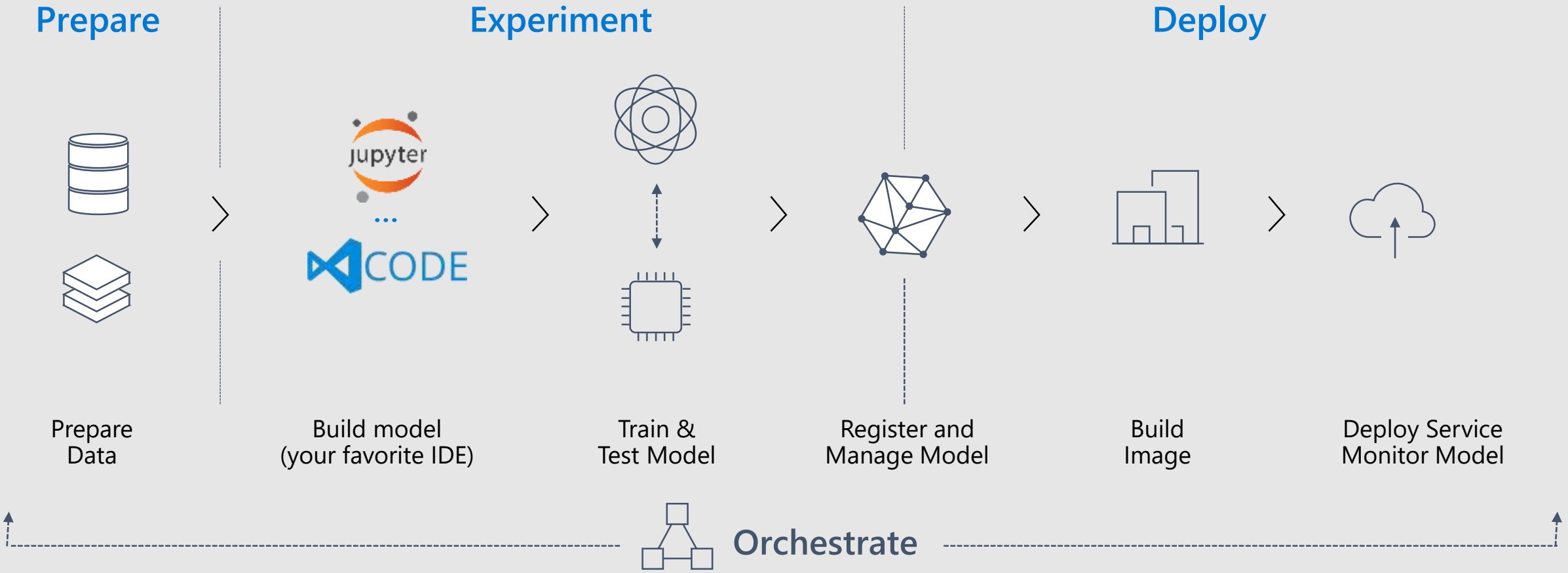
Deploy with Azure Machine Learning

- Model management services
- Deploy as web service to ACI or AKS
- Capture model telemetry

Azure
Machine Learning

Machine Learning

Typical E2E Process



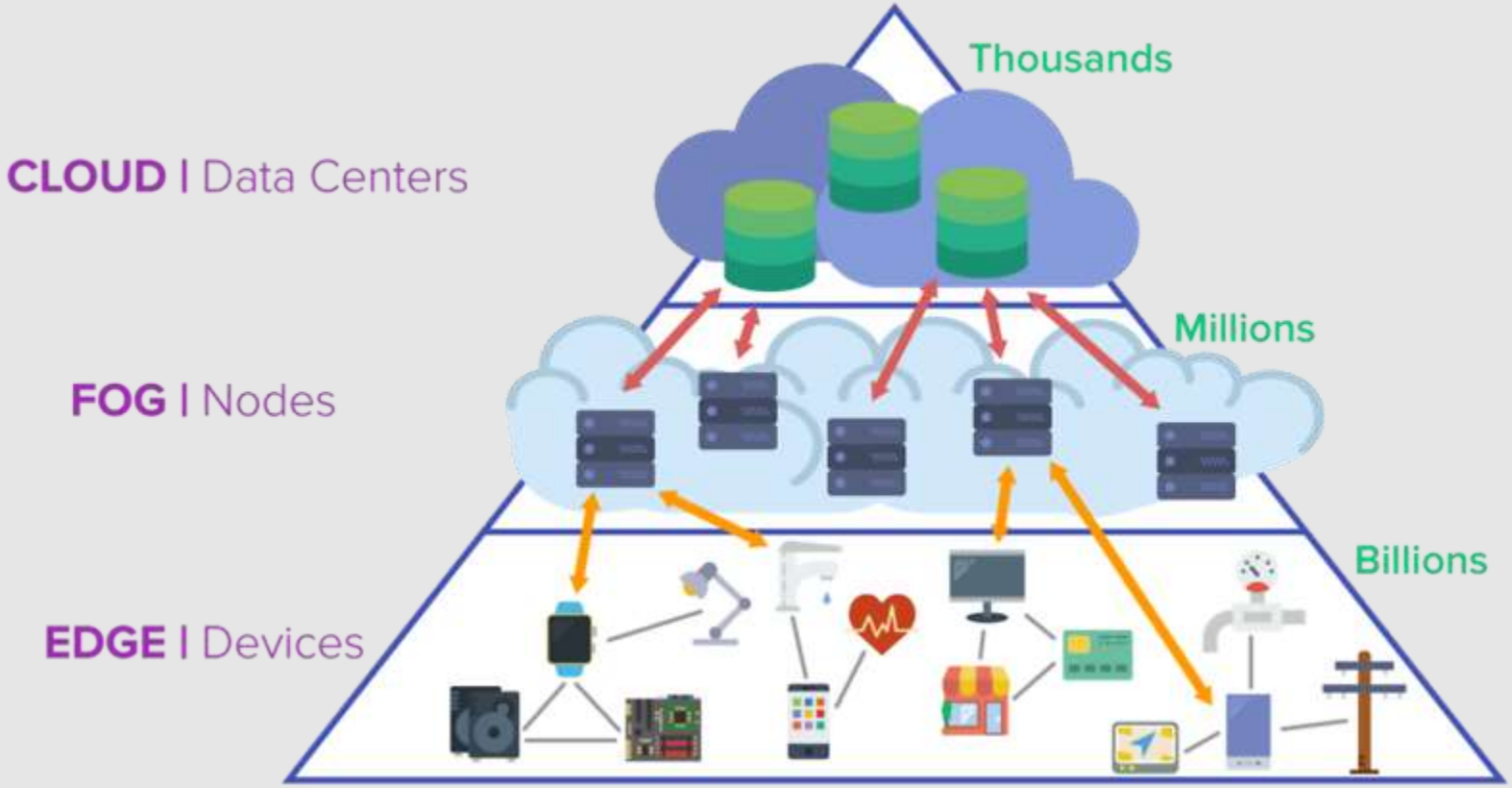
Deploy a VM with your ONNX model in the Cloud

1. **Load Azure ML workspace – (config.json - subscription, resources)**
2. **Registering your model with Azure ML**
3. **Write Score File - python**
4. **Write Environment File (conda dependencies) - yml**
5. **Setup inference configuration – entry script, conda file, Dockerfile**
6. **Deploy the model to ACI – type of cpu, mem**

Demo

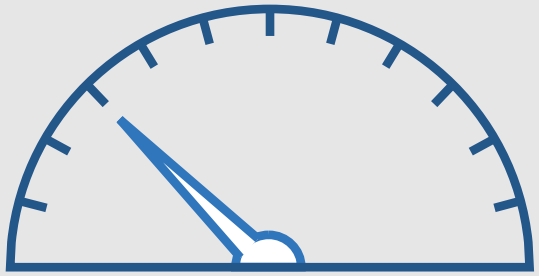
https://github.com/rondagdag/onnx-pected/tree/master/ONNX-AML

What is the Edge?

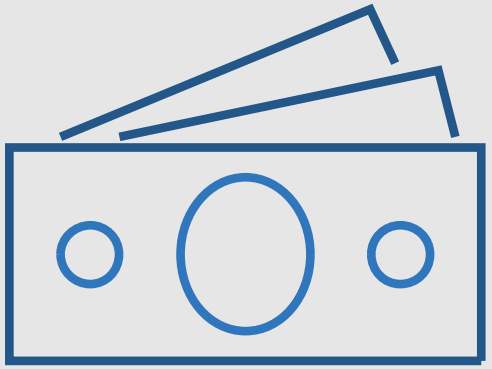


Imagimob AB

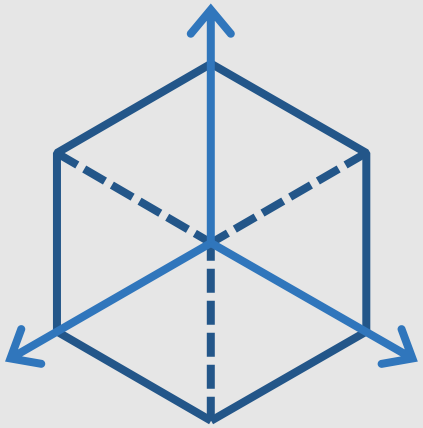
AI on the edge



Low latency



Scalability



Flexibility

ONNX as an intermediary format

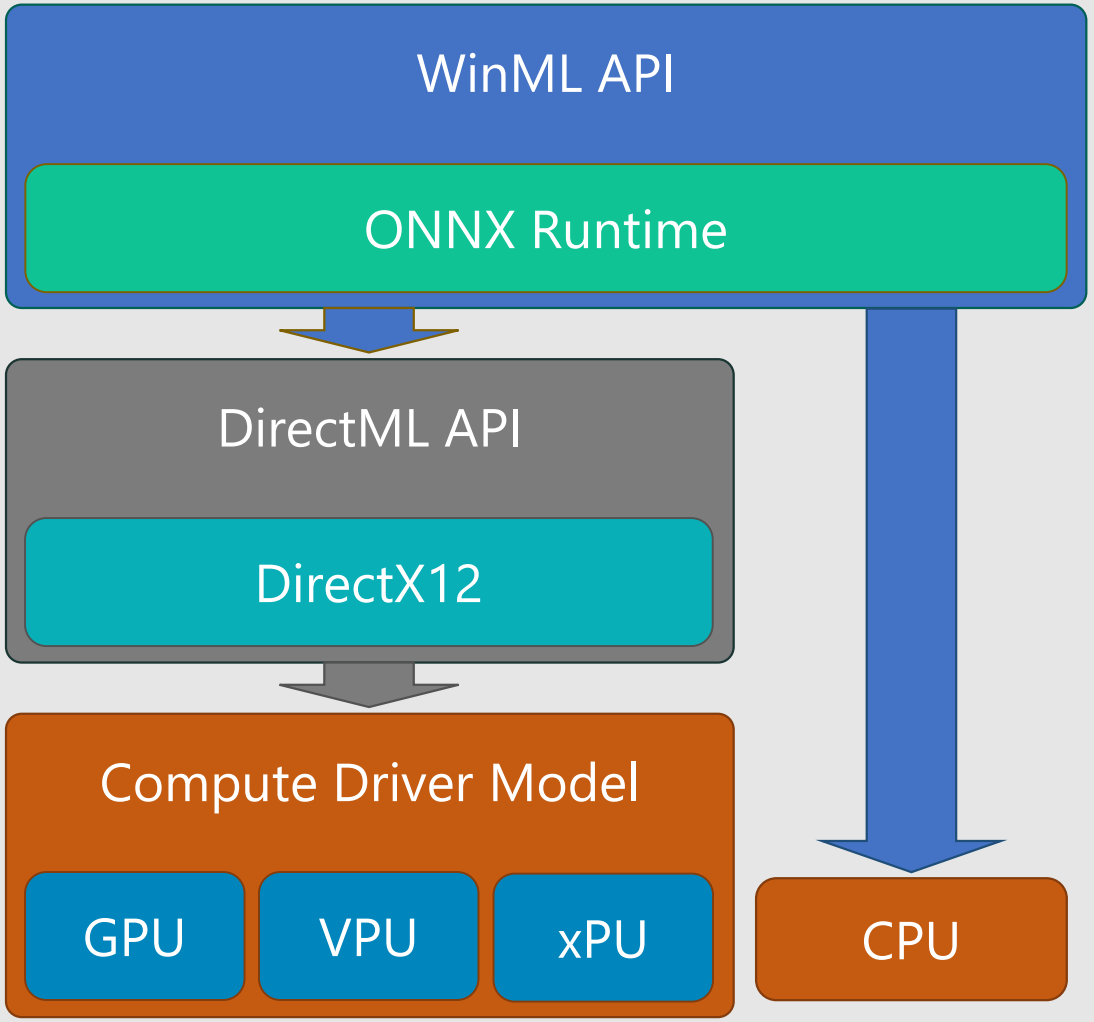
- **Convert to Tensorflow for Android**
  - Convert a PyTorch model to Tensorflow using ONNX
- **Convert to CoreML for iOS**
  - https://github.com/onnx/tutorials/blob/master/examples/CoreML/ONNXLive/README.md
- **Fine-tuning an ONNX model with MXNet/Gluon**
  - https://mxnet.apache.org/api/python/docs/tutorials/packages/onnx/fine_tuning_gluon.html

Deploy to Windows Devices

Windows ML

- Available across Windows family of devices
- Hardware abstraction via DirectML
- Unified API for Win32 and WinRT
- Optimized for performance
- Virtualization ready

Windows AI platform



- WinML
  - **Practical**, simple model-based API for ML inferencing on Windows
- DirectML
  - **Realtime, high control** ML operator API; part of DirectX family
- Compute Driver Model
  - Robust **hardware reach**/abstraction layer for compute and graphics silicon

Demo

https://github.com/rondagdag/onnx-pected/tree/master/GenerateONNX-AutoML

ONNX Runtime

- High performance runtime for ONNX models
- Supports full ONNX-ML spec (currently v1.2+)
- Extensible architecture to plug-in hardware accelerators
- Simple Python API

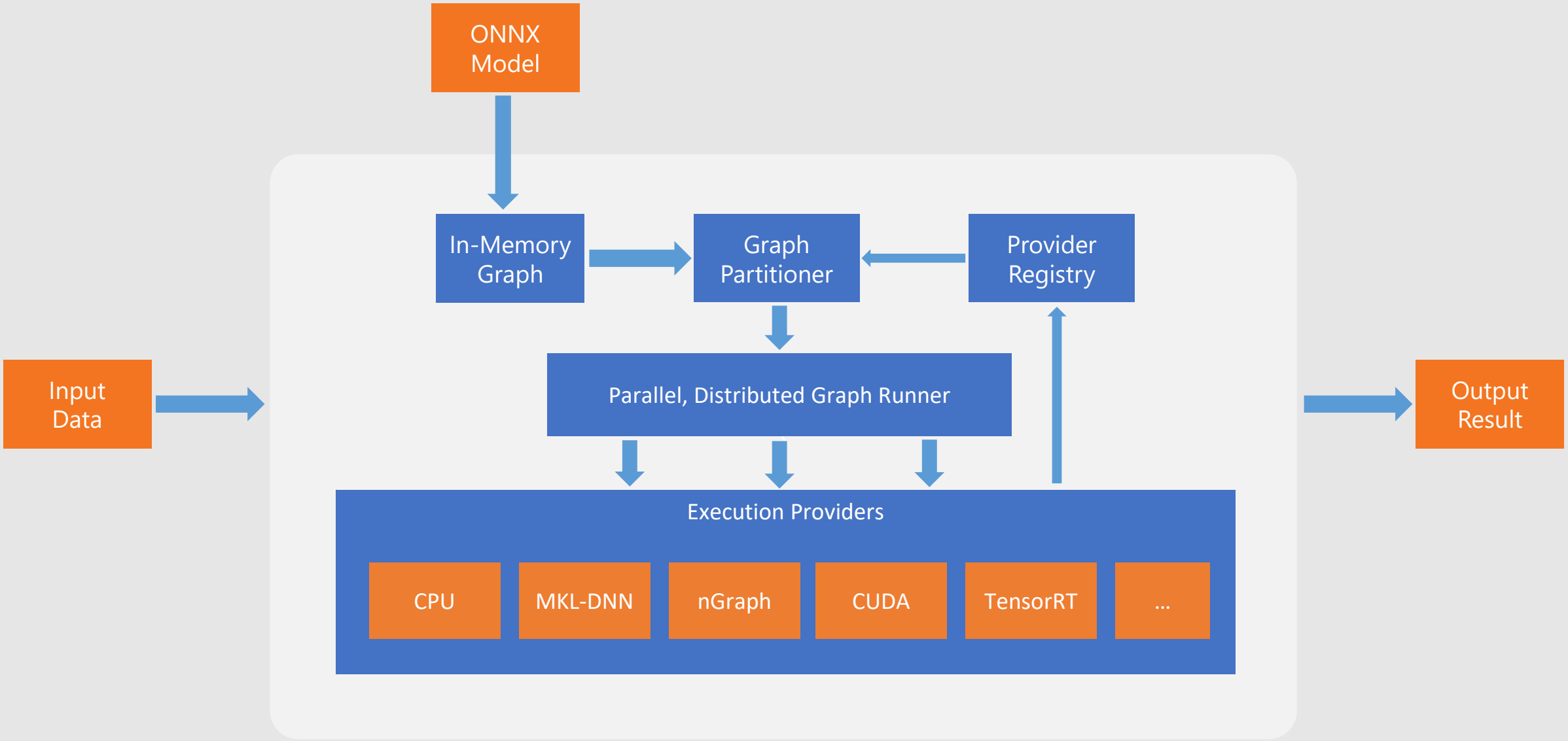
ONNX Runtime

Get Started Easily

Select your requirements and use the resources provided to get started quickly

OS	Windows	Linux	Mac		
Language	Python (3.5-3.7)	C++	C#	C	Java
Architecture	X64	X86	ARM64	ARM32	
Hardware Acceleration	Default CPU	CUDA	TensorRT	DirectML	MKL-DNN
	MKL-ML	nGraph	NUPHAR	OpenVINO	
Installation Instructions	Install Nuget package Microsoft.ML.OnnxRuntime				

ONNX


ONNX Model
In-Memory Graph
Graph Partitioner
Provider Registry
Input Data
Parallel, Distributed Graph Runner
Output Result
Execution Providers
CPU
MKL-DNN
nGraph
CUDA
TensorRT
...

ONNX Runtime

- Office team saw a **14.6x reduction in latency** for a grammar checking model (thousands of queries per minute)
- Azure Cognitive Services saw a **3.5x reduction in latency** for an optical character recognition (OCR) model
- Bing QnA saw a **2.8x reduction in latency** for a model that generates answers to questions
- Bing Visual Search saw a **2x reduction in latency** for a model that helps identify similar images

https://cloudblogs.microsoft.com/opensource/2019/05/22/onnx-runtime-machine-learning-inferencing-0-4-release/

ONNX Runtime - Python API

```
import onnxruntime

session = onnxruntime.InferenceSession("mymodel.onnx")

results = session.run([], {"input": input_data})
```

ONNX Docker Image

- onnx-docker-cpu: Image with ONNX, PyTorch, Tensorflow support
- onnx-docker-gpu: Image with ONNX, PyTorch (CUDA), Caffe2 support
- onnx-ecosystem: Jupyter notebook environment for getting started quickly with ONNX models, ONNX converters, and inference using ONNX Runtime.

ONNX Runtime Server

- **ONNX Runtime Server**
  - https://github.com/onnx/tutorials/blob/master/tutorials/OnnxRuntimeServerSSDModel.ipynb

ONNX

Reference implementation to use ONNX Runtime with Azure IoT Edge

- https://github.com/Azure-Samples/onnxruntime-iot-edge

ONNX.js

- ONNX.js is a JavaScript library for running ONNX models on browsers and on Node.js.
- ONNX.js has adopted Web Assembly and WebGL technologies
- optimized ONNX model inference runtime for both CPUs and GPUs.

https://github.com/microsoft/onnxjs

ONNX

ONNX.js

Compatibility

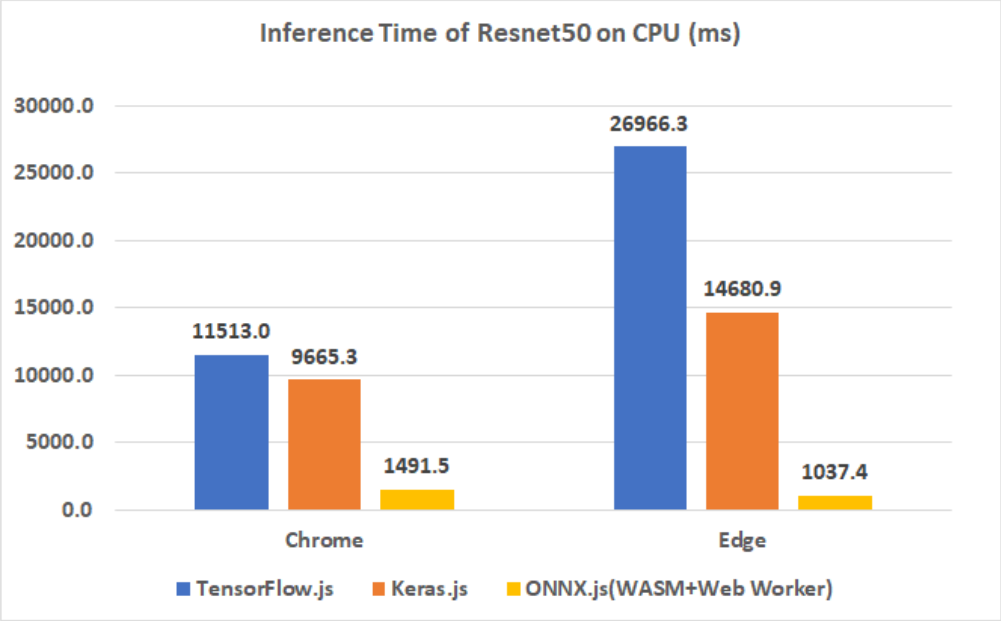
Desktop Platforms

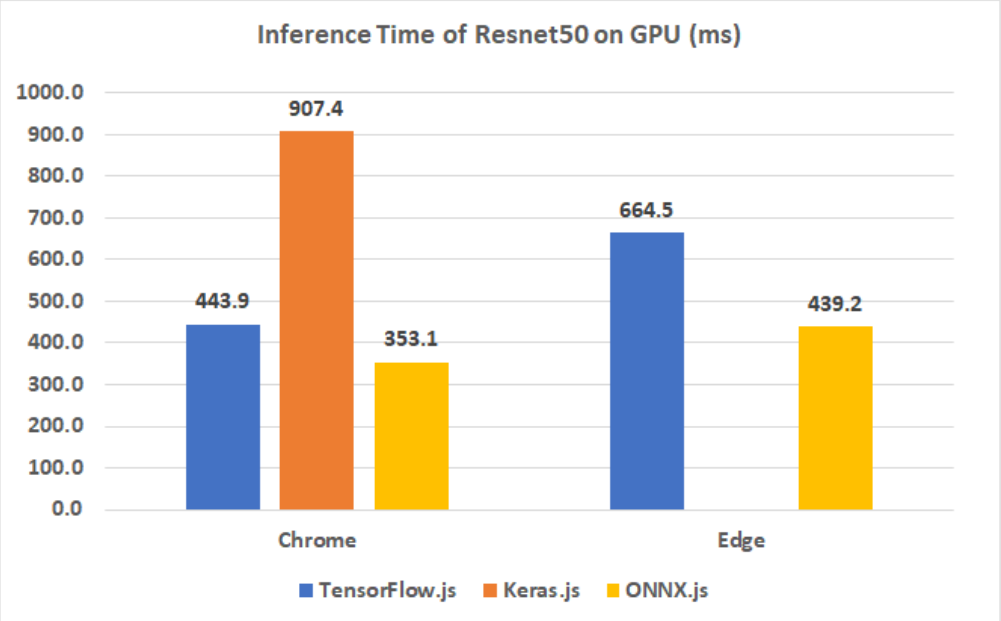
OS/Browser	Chrome	Edge	FireFox	Safari	Opera	Electron	Node.js
Windows 10	✔	✔	✔	-	✔	✔	✔
macOS	✔	-	✔	✔	✔	✔	✔
Ubuntu LTS 18.04	✔	-	✔	-	✔	✔	✔

Mobile Platforms

OS/Browser	Chrome	Edge	FireFox	Safari	Opera
iOS	✔	✔	✔	✔	✔
Android	✔	✔	Coming soon	-	✔

ONNX.js


Inference Time of Resnet50 on CPU (ms)
30000.0
25000.0
20000.0
15000.0
10000.0
5000.0
0.0
11513.0
9665.3
1491.5
26966.3
14680.9
1037.4
Chrome
Edge
TensorFlow.js
Keras.js
ONNX.js(WASM+Web Worker)

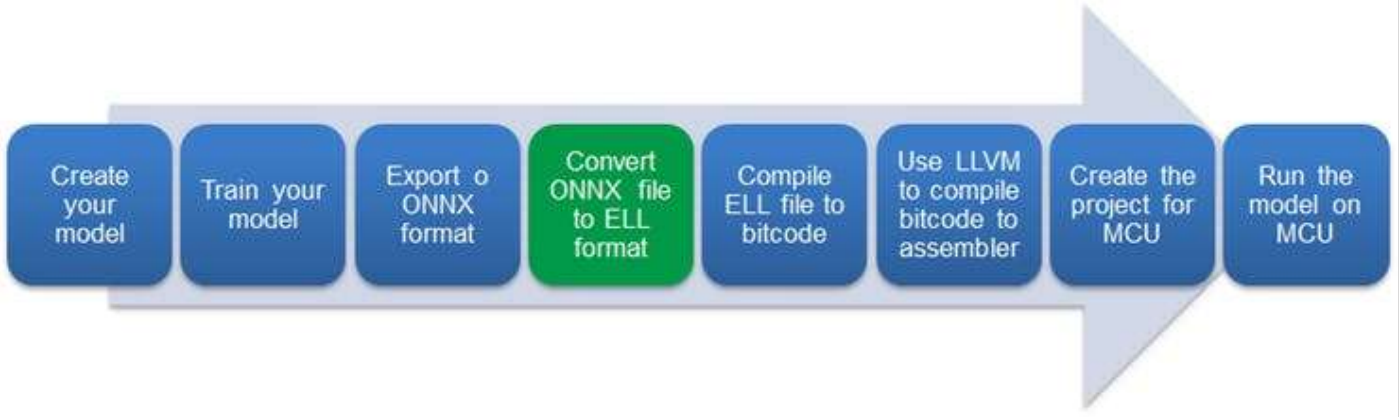

Inference Time of Resnet50 on GPU (ms)
1000.0
900.0
800.0
700.0
600.0
500.0
400.0
300.0
200.0
100.0
0.0
443.9
907.4
353.1
664.5
439.2
Chrome
Edge
TensorFlow.js
Keras.js
ONNX.js

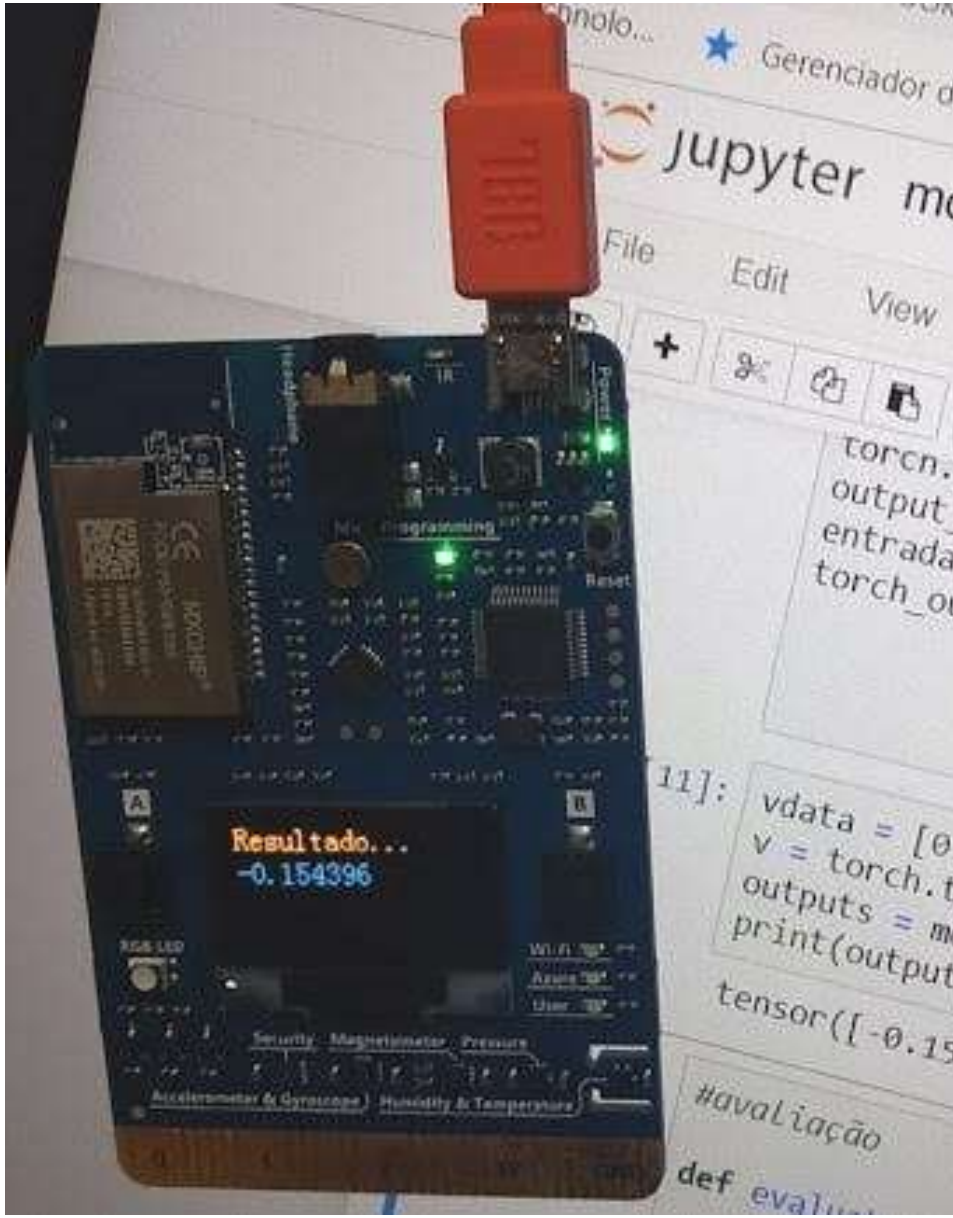
Demo

https://github.com/rondagdag/onnx-pected/tree/master/webmnist-master

Wait… there’s more

- Embedded Learning Library
  - https://github.com/microsoft/ELL
- Machine Learning Model Running on Azure IoT Starter Kit
  - https://www.hackster.io/waltercoan/machine-learning-model-running-on-azure-iot-starter-kit-f9608b





Recap

- ✓ What is ONNX

  ONNX is an open standard so you can use the right tools for the job and be confident your models will run efficiently on your target platforms

- ✓ How to create ONNX models

  ONNX models can be created from many frameworks

- ✓ How to deploy ONNX models

  ONNX models can be deployed with Windows ML, .NET/Javascript/Python and to the cloud with Azure ML and the high performance ONNX Runtime

Try it for yourself!

ONNX Runtime is available now!

```
pip install onnxruntime
pip install onnxruntime-gpu
```

Documentation and samples at aka.ms/onnxruntime

Source for Demo:

https://github.com/rondagdag/onnx-pected

http://bit.ly/onnxpected

Ron Dagdag

- Sr. Software Engineer/Voice AI Assistant Specialist at Crestron Electronics
- Microsoft MVP award
- Hackster DFW Ambassador meetup.com/Hackster-DFW
- Dallas Littlebits Chapter Leader meetup.com/amRobotics
- Dallas AR/VR Development meetup meetup.com/Dallas-Virtual-Reality
- *"Opinions expressed are solely my own and do not express the views or opinions of my employer."*