

W> JavaScript
Congress 2021

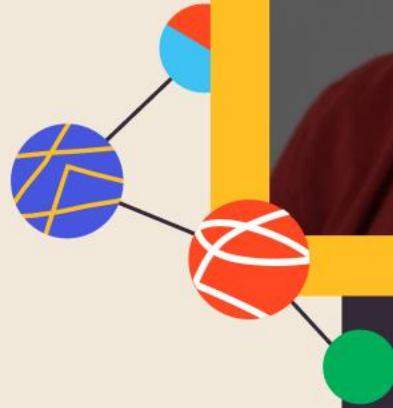
ONLINE
NOVEMBER
23-25

Making Neural Network Portable with ONNX

Ron Dagdag

Lead Software Engineer at Spacee and Microsoft MVP

jscongress.com





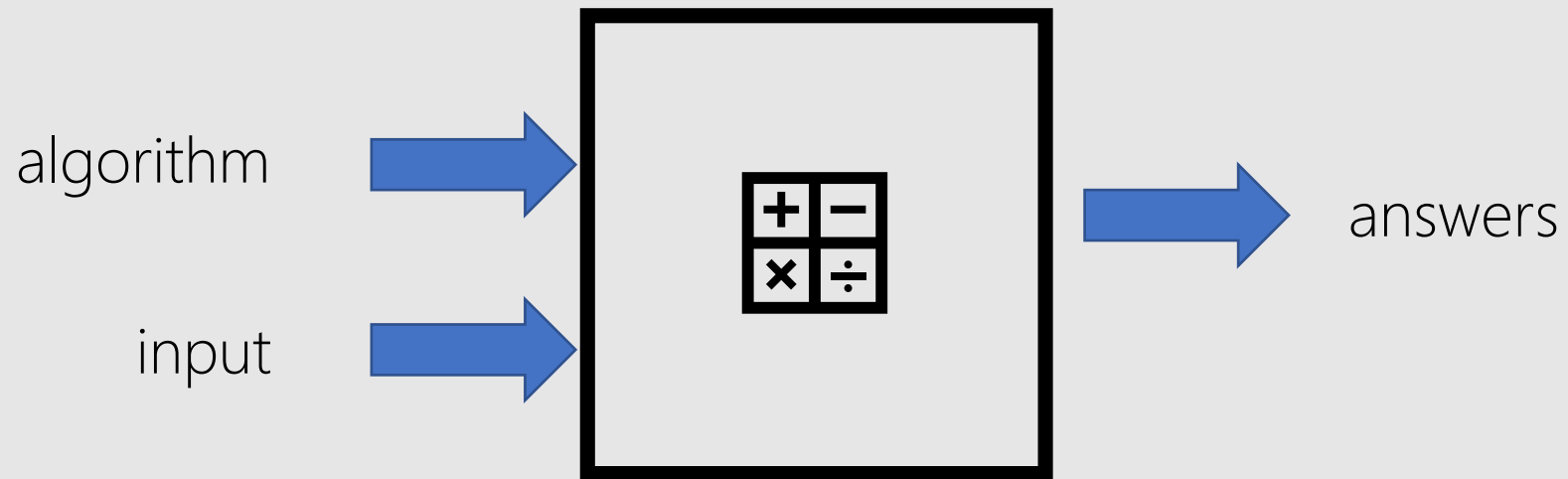
<https://bit.ly/onnxportable>

ONNX
Not ONIX
Not ONYX

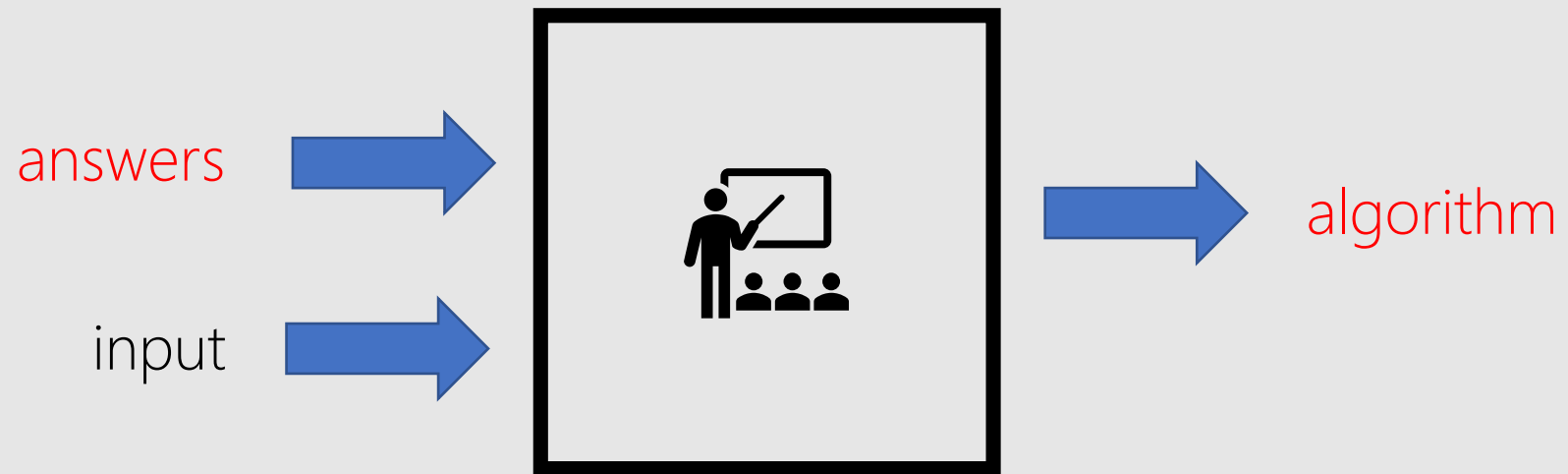




programming



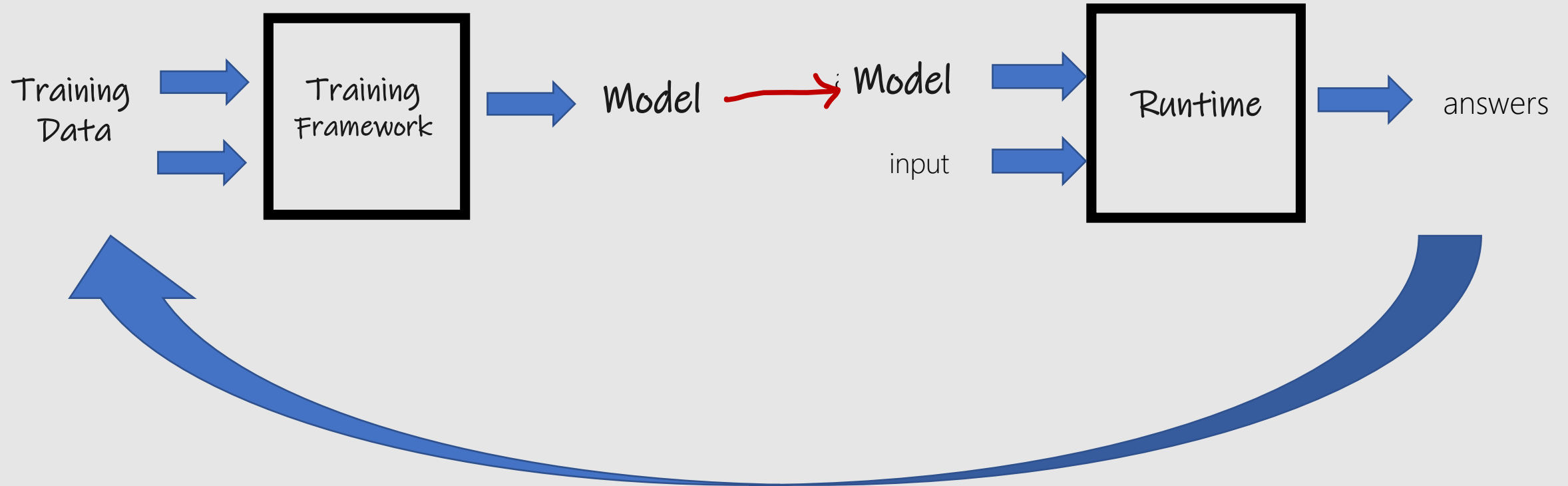
machine learning



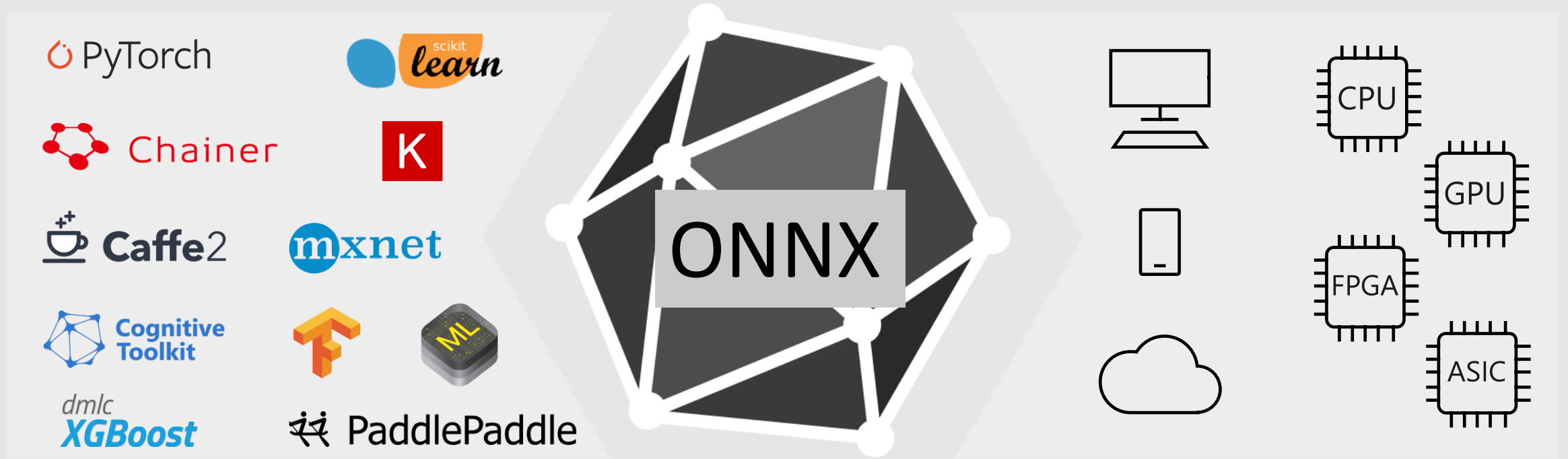
ML Primer

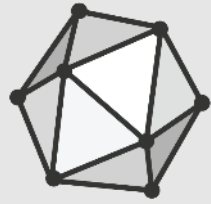
machine Learning

Inferencing



Open and Interoperable AI





ONNX

Open Neural Network Exchange

Open format for ML models

github.com/onnx

onnx.ai/



ONNX Partners

ABBYY®

Alibaba Group
阿里巴巴集团

AMD

arm

aws

Baidu 百度

BECKHOFF

BITMAIN

cādence®

CEVA®

Facebook
Open Source

GRAPHCORE

habana

HAILO

Hewlett Packard
Enterprise

HUAWEI

IBM®

Idein Inc

intel AI

MathWorks®

MAXAR

MEDIATEK

mi

Microsoft

NVIDIA

NXP

OctoML

OPEN AI LAB
开放智能

Preferred
Networks

SIEMENS

SONY

Qualcomm

sas

商汤
sensetime

skymizer

SYNOPSYS®

Tencent

unity

verizon
media

vmware®

WOLFRAM

Yandex

ZETANE



OLFAI
GRADUATE
PROJECT

Agenda

- ✓ What is ONNX, When to use ONNX
- ☐ How to create ONNX models
- ☐ How to deploy ONNX models

When to use ONNX?

- Trained in Python - deploy into a C#/Java/Javascript app
- High Inferencing latency for production use
- Model to run resource on IoT/edge devices
- Model to run on different OS or Hardware
- Combine running models created from different frameworks
- Training takes too long (transformer models)

Create

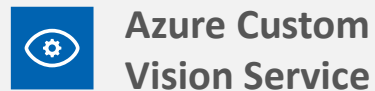
Frameworks



Native support

Converters

Services



Native support



Deploy

Cloud Services

Azure Machine Learning services

Ubuntu VM

Windows VM

Windows Devices

IoT/Edge Devices

Other Devices (iOS, Android, etc)

Native support

Converters

Frameworks



**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



Windows Devices

**Step 2:
Deploy**

Other Devices
(iOS, etc)

Native
support

Converters

A still life composition featuring several brown eggs in a cardboard carton, a pair of wire-rimmed glasses, and an open book. The text "Secret Recipe" is overlaid in white, with a vertical line separating the words. The background is dark and moody, with the objects resting on a reflective surface.

Secret Recipe

4 ways to get an ONNX model



ONNX Model Zoo



Azure 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 error from paper - ~3.6%
SqueezeNet	Iandola et al.	A lightweight CNN model with fewer parameters and less computation. Top-5 error from paper - ~4.8%
VGG	Simonyan et al.	Deep CNN model, won the ImageNet Challenge in 2014. Top-5 error from paper - ~7.4%

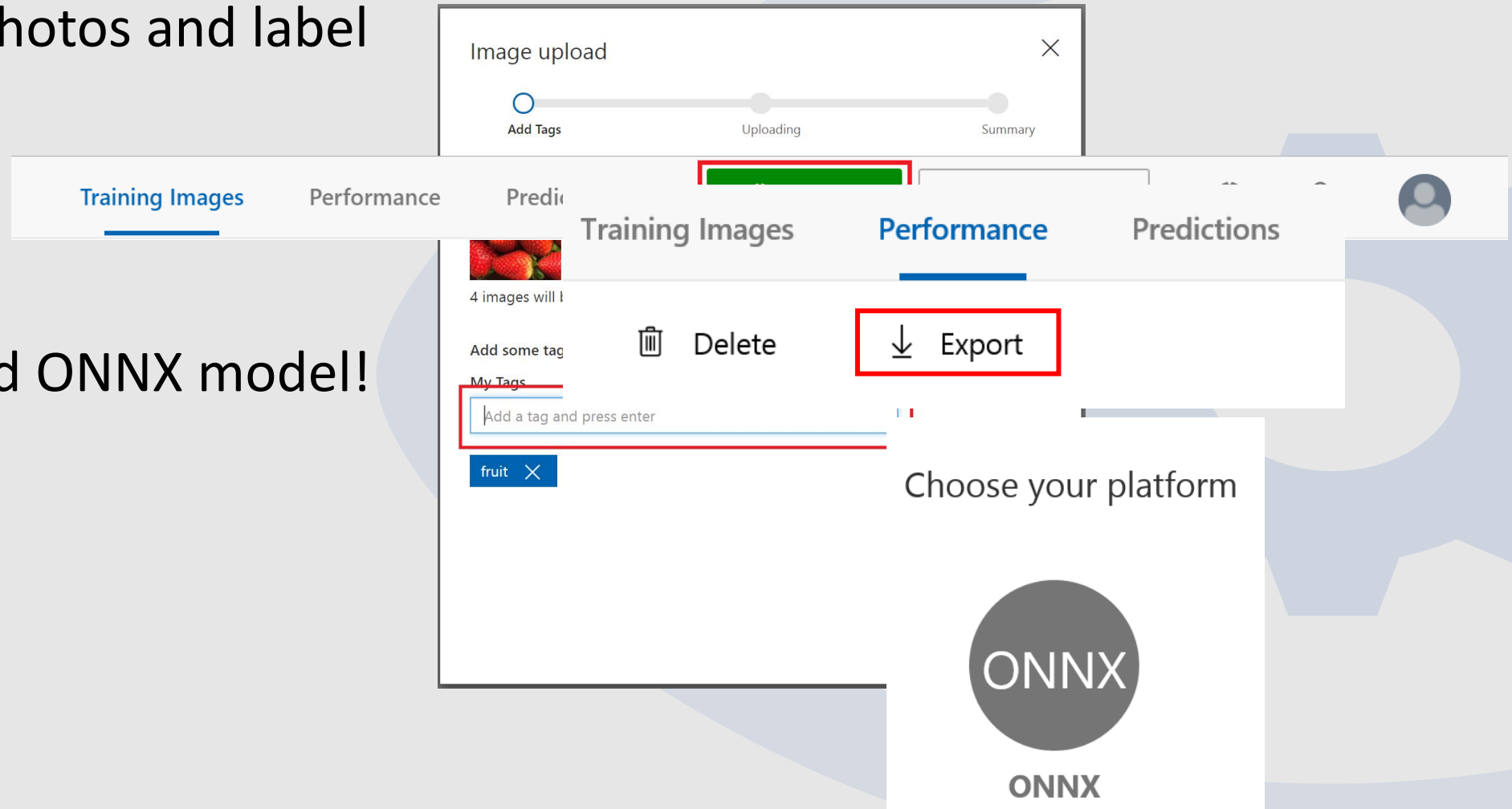
Model	Download	Checksum	Download (with sample test data)	ONNX version	Opset version	Top-1 accuracy (%)	Top-5 accuracy (%)
ResNet-18	44.6 MB	MD5	42.9 MB	1.2.1	7	69.70	89.49
ResNet-34	83.2 MB	MD5	78.6 MB	1.2.1	7	73.36	91.43
ResNet-50	97.7 MB	MD5	92.0 MB	1.2.1	7	75.81	92.82
ResNet-101	170.4 MB	MD5	159.4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



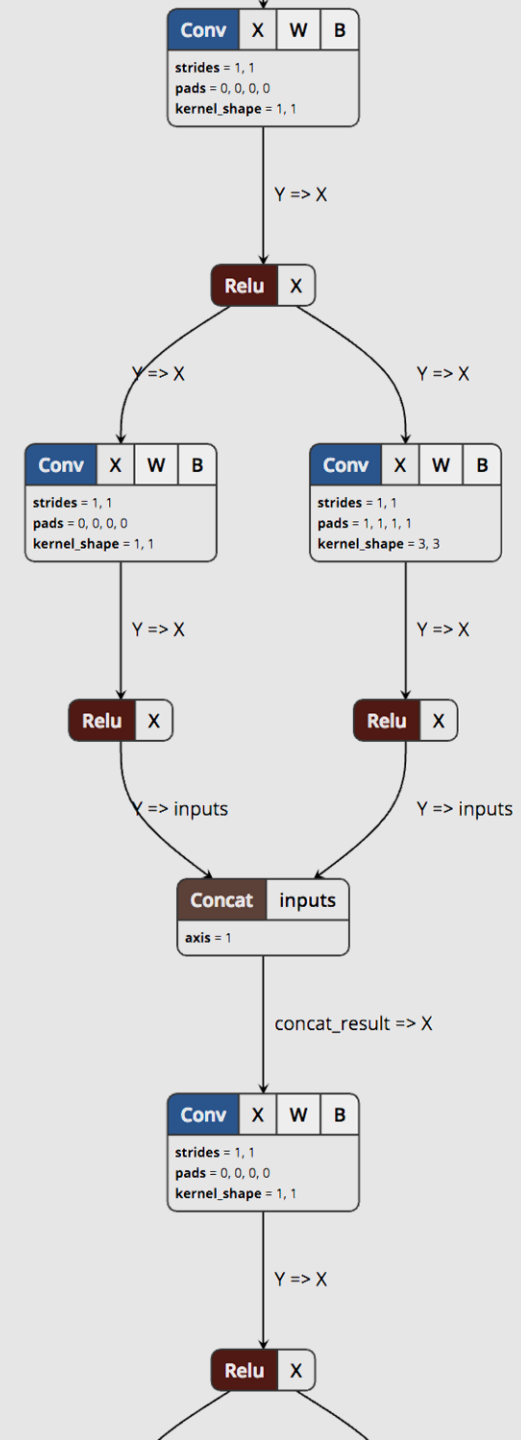
ONNX Models

Graph of operations

Netron

<https://netron.app/>

<https://lutzroeder.github.io/netron/>



Convert
models



Convert models

1. Load existing model
2. (Convert to ONNX)
3. Save ONNX model



Convert models: PyTorch

```
import torch  
import torch.onnx
```

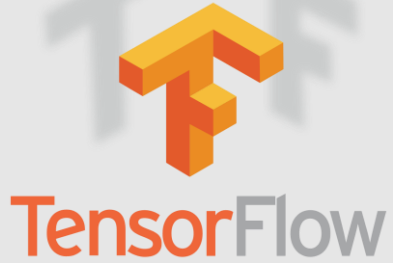
```
model = torch.load("model.pt")
```

```
sample_input = torch.randn(1, 3, 224, 224)
```

```
torch.onnx.export(model, sample_input, "model.onnx")
```



Convert models:



```
> python -m tf2onnx.convert  
    --saved-model tensorflow-model-path  
    --output model.onnx
```

<https://github.com/onnx/tensorflow-onnx>



A top-down view of a dark, textured surface, possibly a countertop, with various baking ingredients and tools. In the upper right, a metal sieve contains white powder. Below it, a red and white checkered cloth holds three white eggs. A wooden rolling pin lies horizontally across the bottom right. A metal whisk is visible in the lower left. The text "Create an ONNX Model" is centered in a white, sans-serif font, enclosed in a thin white rectangular border.

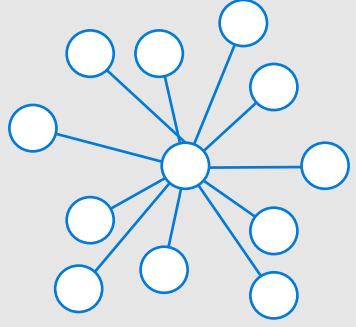
Create an ONNX Model

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/onnx-coreml>
- **Fine-tuning an ONNX model with MXNet/Gluon**
 - https://mxnet.apache.org/versions/1.3.1/tutorials/onnx/fine_tuning_gluon.html

<https://github.com/onnx/tutorials>

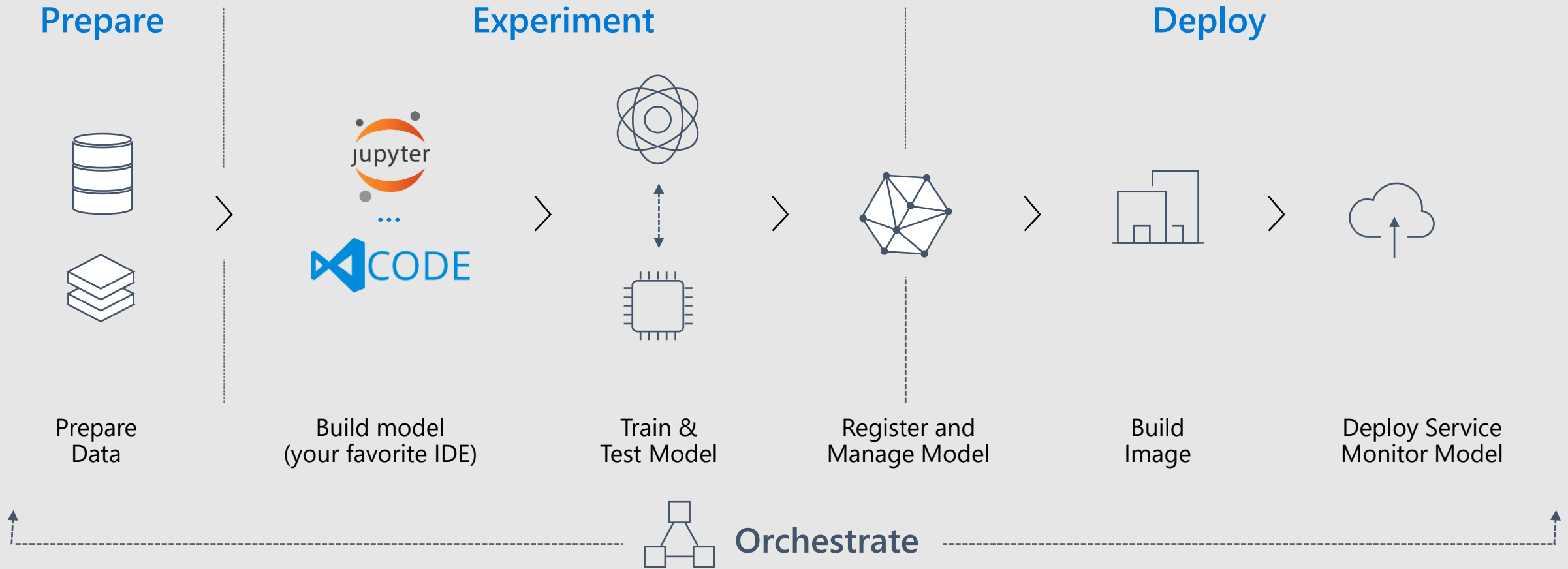
Train models in the Cloud



- 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

Machine Learning

Typical E2E Process



Frameworks



**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




**Step 2:
Deploy**

Other Devices
(iOS, etc)

Native
support

Converters

A baker in a white shirt and apron is shown from the waist down, working on a wooden table. The table is covered with a layer of white flour. A large, round loaf of bread is being shaped on the table. The baker's hands are visible, and they are wearing a green and white patterned apron. The background is a plain, light-colored wall.

Baker vs Starting a Bakery

Create

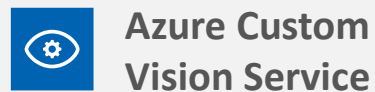
Frameworks



Native support

Converters

Services



Native support

ONNX Model

Deploy

Azure

Azure Machine Learning services

Ubuntu VM

Windows Server 2019 VM

Windows/Linux Devices

IoT Edge Devices

Other Devices
(iOS, etc)

Native support

Converters

A person's hands are visible, holding a large, round, rustic loaf of bread. The bread has a thick, golden-brown crust with some darker, caramelized spots. It is wrapped in a blue and white striped cloth. The background is a blurred wooden surface.

Cloud or Edge

Deploy in Cloud

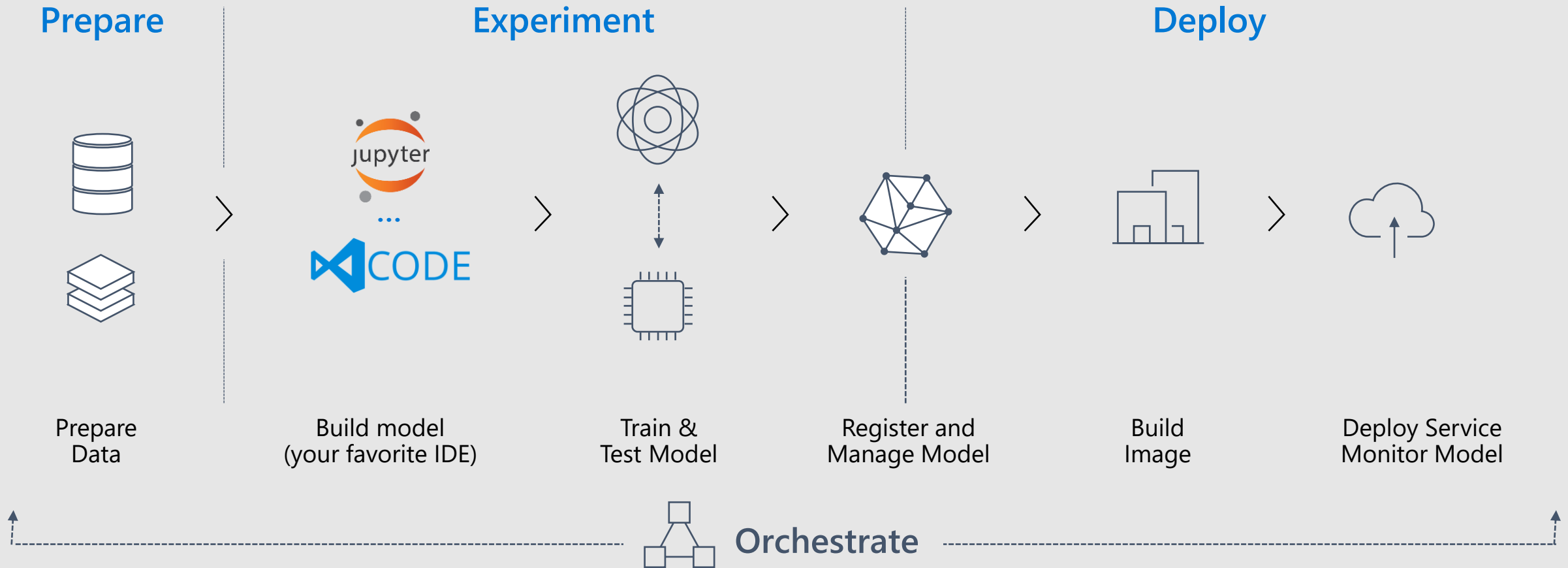
- Model management services
- Deploy as web service
- Capture model telemetry



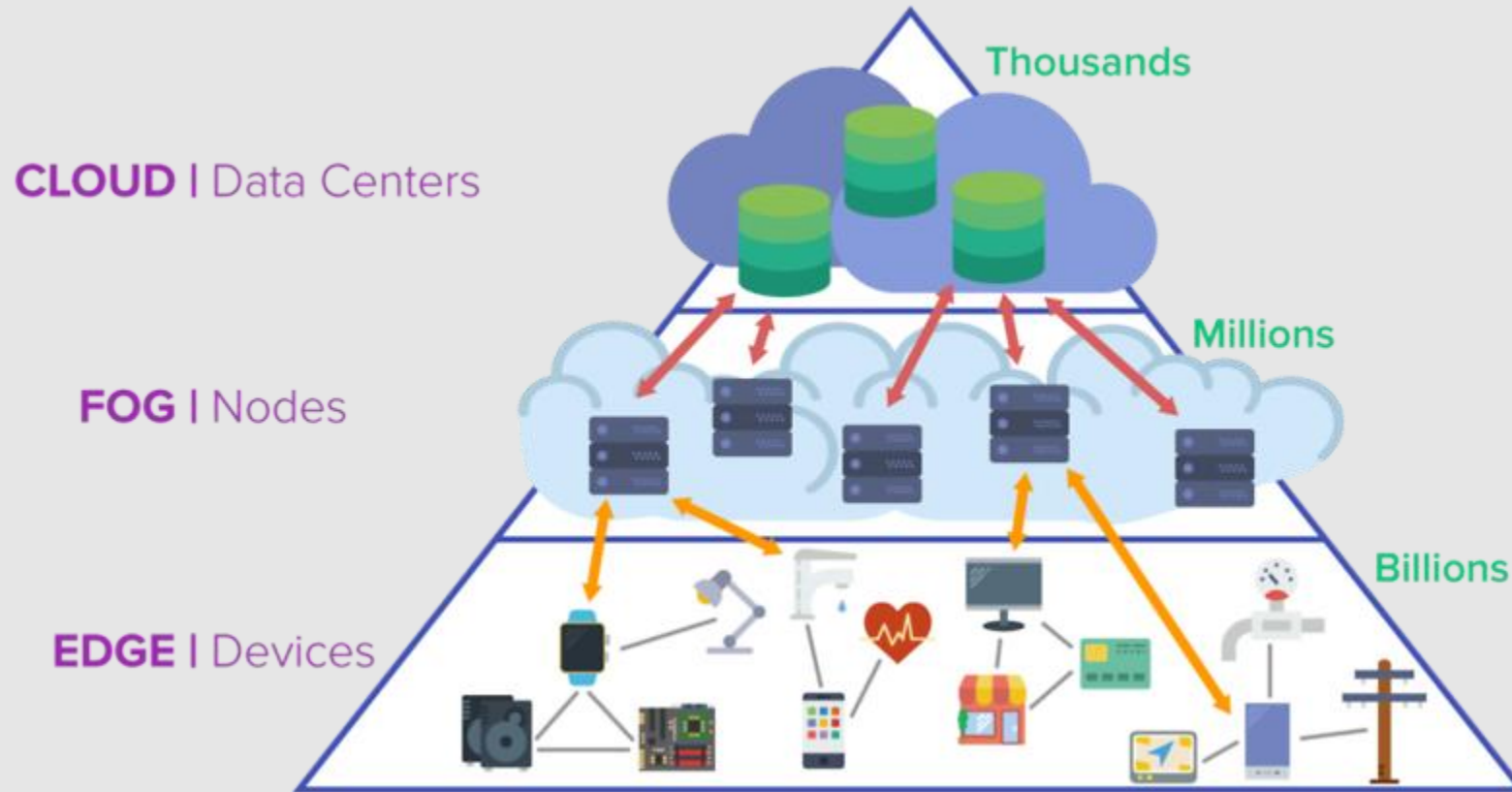
Azure
Machine Learning

Machine Learning

Typical E2E Process



What is the Edge?

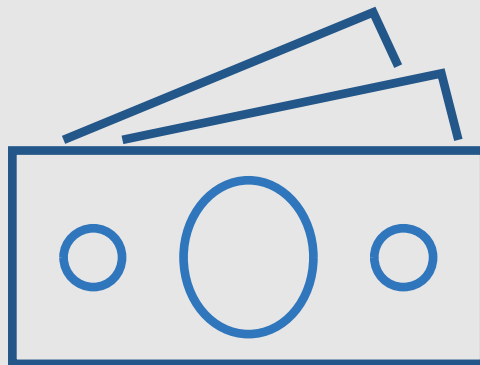


[Imagimob AB](#)

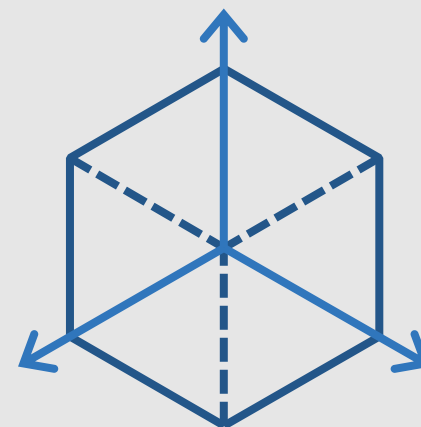
AI on the edge



Low latency



Scalability



Flexibility

ONNX Runtime

- High performance inference engine for ONNX models
- Founded and Open Sourced by Microsoft under MIT License
- Supports full ONNX-ML spec
- Extensible architecture to plug-in hardware accelerators
- Ships with Windows 10 as WinML
- onnxruntime.ai



ONNX

ONNX Runtime

Get Started Easily

Optimize Inferencing	Optimize Training							
Platform	Windows	Linux	Mac	Android	iOS	Web Browser (Preview)		
API	Python	C++	C#	C	Java	JS	Obj-C	WinRT
Architecture	X64	X86	ARM64	ARM32	IBM Power			
Hardware Acceleration	Default CPU	CUDA	DirectML	oneDNN	OpenVINO			
	TensorRT	NNAPI	ACL (Preview)	ArmNN (Preview)	CoreML (Preview)			
	MIGraphX (Preview)	NUPHAR (Preview)	Rockchip NPU (Preview)	Vitis AI (Preview)				
Installation Instructions	Install Nuget package Microsoft.ML.OnnxRuntime.Gpu Refer to docs for requirements.							

ONNX
Runtime
JavaScript

Node.js binding

Web

React Native

ONNX Runtime Node.js

- Node.js binding
- ONNX model inferencing
- Electron
- Uses web assembly

Install

```
# install latest release version  
npm install onnxruntime-node
```

Import

```
// use ES6 style import syntax (recommended)  
import * as ort from 'onnxruntime-node';
```

```
// or use CommonJS style import syntax  
const ort = require('onnxruntime-node');
```




Node.js Demo

ONNX Runtime Web (ORT-Web)

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

Install

```
# install latest release version  
npm install onnxruntime-web  
  
# install nightly build dev version  
npm install onnxruntime-web@dev
```

Import

```
// use ES6 style import syntax (recommended)  
import * as ort from 'onnxruntime-web';
```

```
// or use CommonJS style import syntax  
const ort = require('onnxruntime-web');
```


Why inference in the browser



It's faster



It's safer and helps with
privacy



It works offline



It's cheaper

Why not in the browser?



**THE MODEL IS TOO LARGE AND
REQUIRES HIGHER HARDWARE SPECS.**



DOWNLOADED ONTO THE DEVICE



Web Browser Demo

React Native

- score pre-trained ONNX models
- ONNX Runtime Mobile
- light-weight inference solution
- Android and iOS

Install

```
# install latest release version  
npm install onnxruntime-react-native
```

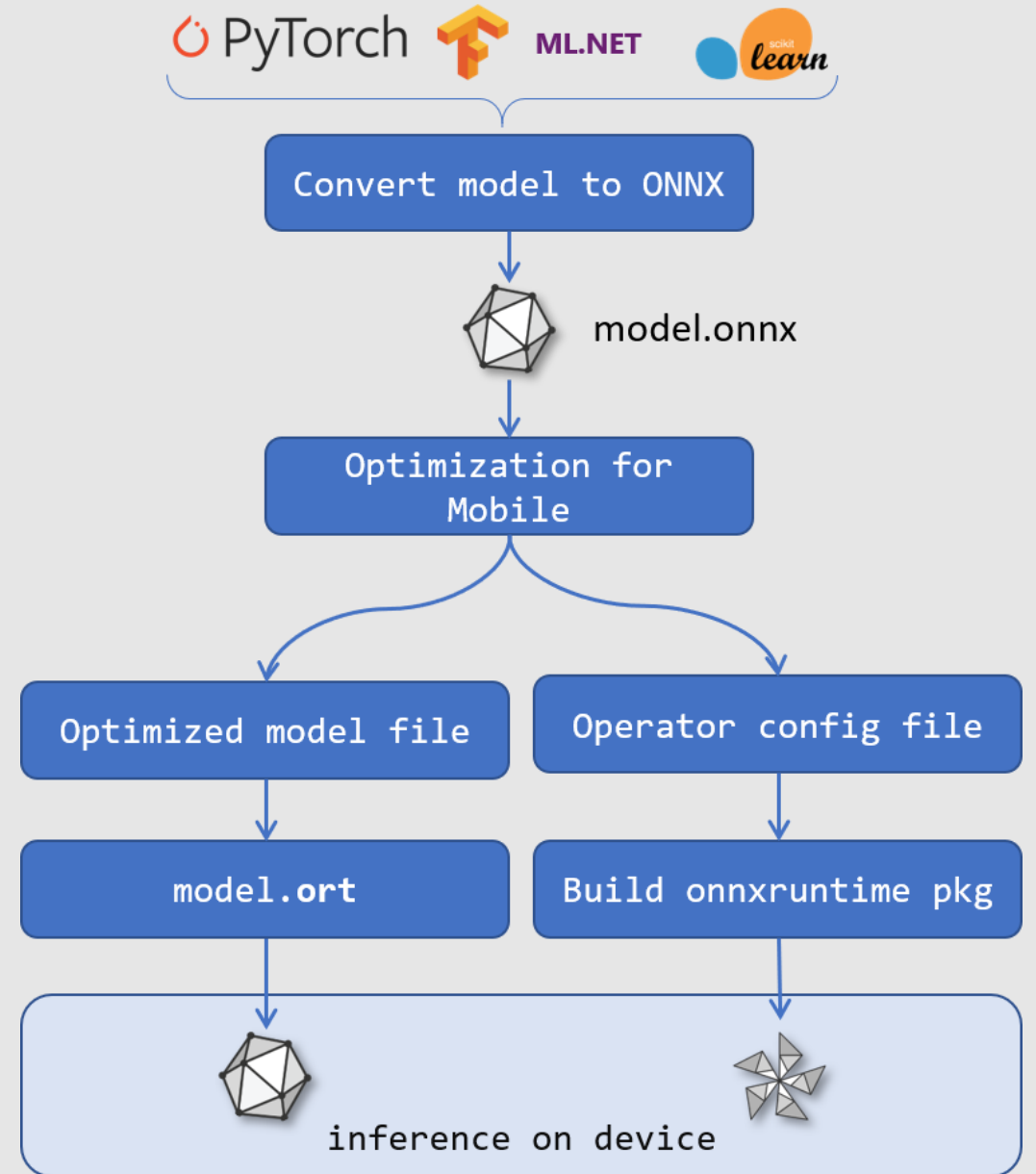
Import

```
// use ES6 style import syntax (recommended)  
import * as ort from 'onnxruntime-react-native';
```

```
// or use CommonJS style import syntax  
const ort = require('onnxruntime-react-native');
```

ONNX Runtime Mobile

- minimizes the binary size
- pre-optimized ONNX model to an internal format ('ORT format model')



Compatibility Chart

Compatibility

OS/Browser	Chrome	Edge	Safari	Electron	Node.js
Windows 10	wasm, webgl	wasm, webgl	-	wasm, webgl	wasm
macOS	wasm, webgl	wasm, webgl	wasm, webgl	wasm, webgl	wasm
Ubuntu LTS 18.04	wasm, webgl	wasm, webgl	-	wasm, webgl	wasm
iOS	wasm, webgl	wasm, webgl	wasm, webgl	-	-
Android	wasm, webgl	wasm, webgl	-	-	-



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 Node, Web Browser, React Mobile using high performance ONNX Runtime

About Me

Ron Dagdag

A waiter in a black tuxedo and white shirt with a black bow tie is holding a silver tray. On the tray is a single cupcake with white frosting, colorful sprinkles, and a single lit candle. The background is a soft, out-of-focus light color.

Lead Software Engineer at Spacee

5th year Microsoft MVP awardee

Personal Projects
www.dagdag.net

<https://linktr.ee/rondagdag>

Email: ron@dagdag.net
Twitter [@rondagdag](https://twitter.com/rondagdag)

Connect me via Linked In
www.linkedin.com/in/rondagdag/

Thanks for geeking out with me about ONNX




<https://bit.ly/onnxportable>

Hackster Portfolio

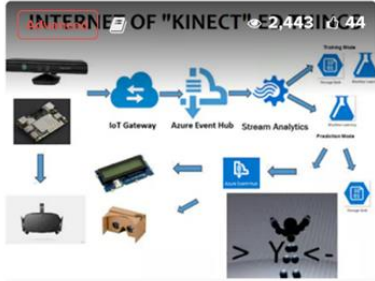
www.dagdag.net

@rondagdag




Ron Dagdag
Dad / Lead Software Engineer / 3D Developer / Tax Return Preparer.
Passionate to learn about Robotics, VR, AR, Artificial Intelligence, IOT
@rondagdag
FORT WORTH, United States
Team [Augmented Reality](#)
Team [Virtual Reality](#)

INTERNET OF "KINECT" 2,443 16 44



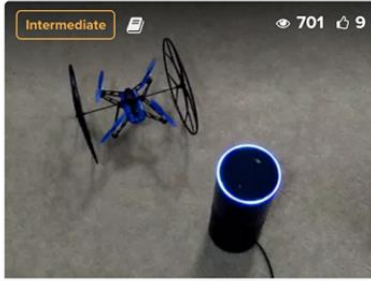
Posture Recognition using K...
Ron Dagdag

Easy 60 0




Littlebits Arduino Keyboard ...
Ron Dagdag

Intermediate 701 9



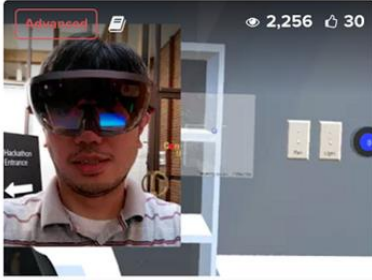
Alexa, tell Echobot to fly
Ron Dagdag

Advanced 1,345 12




Control your "Earth Rover" i...
Ron Dagdag

Advanced 2,256 30



ConstructAR - The Holograp...
TEAM ConstructAR

Intermediate 449 4



Color Changing Fireworks in...
Ron Dagdag