

Operationalizing AI - Portable ML Model Sharing across Enterprise

Adnan Masood, PhD.
masooda@stanford.edu

Microsoft Azure
+ AI Conference

CO-PRODUCED BY

Microsoft & DEVintersection



Adnan Masood

Adnan Masood, Ph.D. is an Artificial Intelligence and Machine Learning researcher, software architect, and Microsoft MVP (Most Valuable Professional) for Data Platform. As Chief Architect of AI and Machine Learning at UST Global, he collaborates with Stanford Artificial Intelligence Lab, and MIT AI Lab for building enterprise solutions.


Author of Amazon bestseller in programming languages, "**Functional Programming with F#**", Dr. Masood teaches Data Science at Park University, and has taught Windows Communication Foundation (WCF) courses at the University of California, San Diego. He is a regular speaker to various academic and technology conferences (WICT, DevIntersection, IEEE-HST, IASA, and DevConnections), local code camps, and user groups. He also volunteers as STEM (Science Technology, Engineering and Math) robotics coach for elementary and middle school students.

A strong believer in giving back to the community, Dr. Masood is a co-founder and president of the Pasadena .NET Developers group, co-organizer of Tampa Bay Data Science Group, and Irvine Programmer meetup. His recent talk at Women in Technology Conference (WICT) Denver highlighted the importance of diversity in STEM and technology areas, and was featured by variety of news outlets.



Microsoft®
Most Valuable
Professional

Wednesday, December 5, 2018

7:30am - 8:30am	Continental Breakfast					
SESSIONS	104-105	121	122	101-102	106-107	115
8:30am - 9:30am	Supercharge Your Debugging in Visual Studio 2017 Andy Sterland	Securing Web Applications and APIs with Azure Active Directory B2C Michele Leroux Bustamante & Brock Allen	Depend on Docker – Get IT done with Docker on Azure Alex Iankoulski	Machine Learning with ML.NET Ankit Asthana	 Hands-on Labs	Introduction to Azure Databricks for the Azure Developer Lino Tadros
9:30am - 9:45am	Break					
9:45am - 10:45am	Best Practices for Azure Service Fabric Applications and Clusters Chacko Daniel	Azure App Service Overview Stefan Schackow	Use the Power of the Dark-side to Control Azure (an Introduction to the Azure CLI) Dan Patrick	Microsoft Azure Machine Learning Starting Guide for Developers Lino Tadros	PRE-REGISTRATION IS REQUIRED. Each lab is limited to 25 attendees.	AI Everywhere: Open and Interoperable Platform for AI with ONNX Prasanth Pulavarthi
10:45am - 11:30am	Coffee Break - Marquee Ballroom, Expo Hall open					
11:30am - 12:30pm	Say Yes to NoSQL for the .NET SQL Developer Jeremy Likness	Surviving Event-driven Microservices – A Practical Approach on Azure Michele Leroux Bustamante	Implementing Authorization in Web Applications and APIs Brock Allen	Enable External Access to Your Custom Apps with Azure AD B2B Nick Pinheiro	Two-hour lab: End-to-end Deep Learning on Optimized Azure VMs Ben Odom & Michael Hernandez <i>Continues at 1:45</i>	Building Versatile Real-time and Batch Data Pipelines for AI Kyle Bunting
12:30pm - 1:45pm	Lunch – Marquee Ballroom, Expo Hall open					
1:45pm - 2:45pm	Python and AI in Visual Studio Code, Azure Notebooks and Azure John Lam	Chaos Engineering on Azure Paul Stack	Modernizing .NET Applications with Docker Derrick Miller	Tales from the Trenches – Building Machine Learning Models for Customer Behavior Ciprian Jichici	End-to-end Deep Learning on Optimized Azure VMs Ben Odom & Michael Hernandez	Democratization of AI with Microsoft Cognitive Services Dr. Adnan Masood

ONNX and Azure Machine Learning: Create and deploy interoperable AI models

<https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-build-deploy-onnx>

https://github.com/onnx/models/tree/master/tiny_yolov2

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

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

<https://github.com/Microsoft/onnxjs>

[https://github.com/MicrosoftDocs/azure-](https://github.com/MicrosoftDocs/azure-docs/blob/master/articles/machine-learning/service/how-to-build-deploy-onnx.md)

[docs/blob/master/articles/machine-learning/service/how-to-build-deploy-onnx.md](https://github.com/MicrosoftDocs/azure-docs/blob/master/articles/machine-learning/service/how-to-build-deploy-onnx.md)

ONNX Model Zoo: Developing a face recognition application with ONNX models

<https://medium.com/apache-mxnet/onnx-model-zoo-developing-a-face-recognition-application-with-onnx-models-64eeddb9c7a>

<https://onnx.ai/getting-started>

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

<https://github.com/Microsoft/Windows-Machine-Learning>

<https://github.com/Azure-Samples/cognitive-services-dotnet-sdk-samples>



Artificial Intelligence

Microsoft Practice Development Playbook

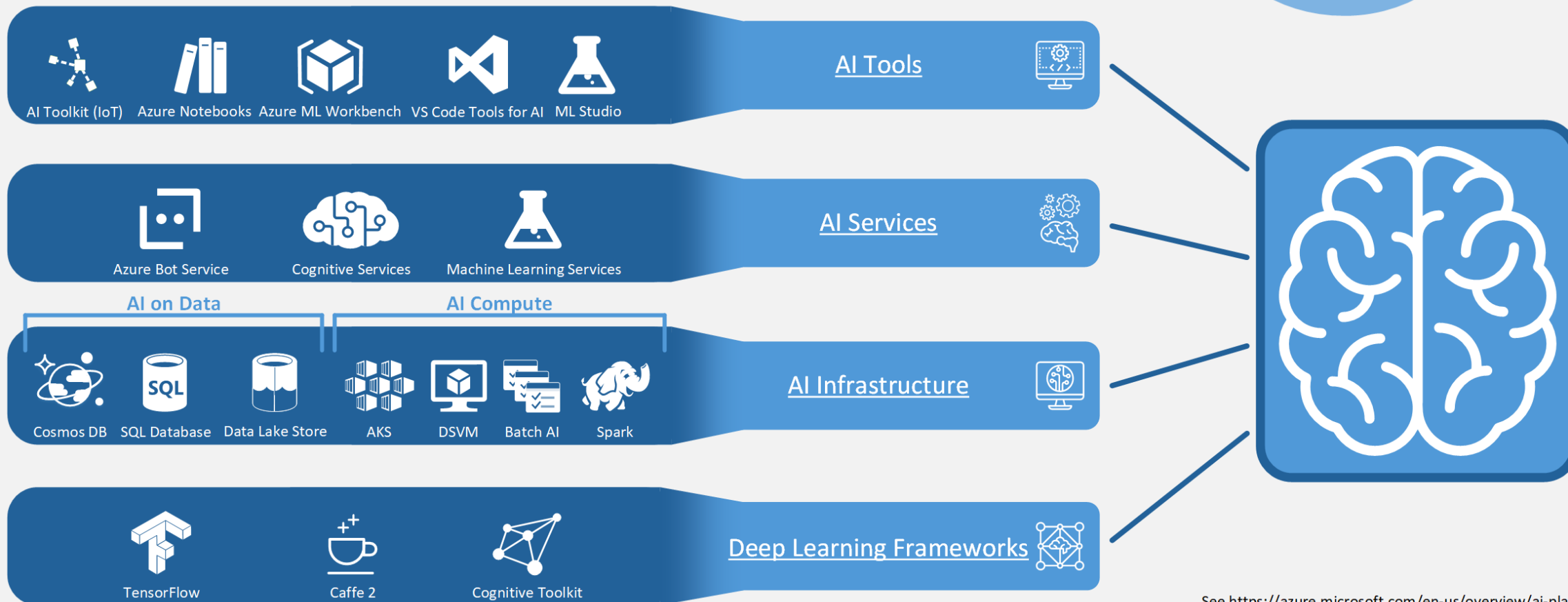


aka.ms/practiceplaybooks

Table of Contents

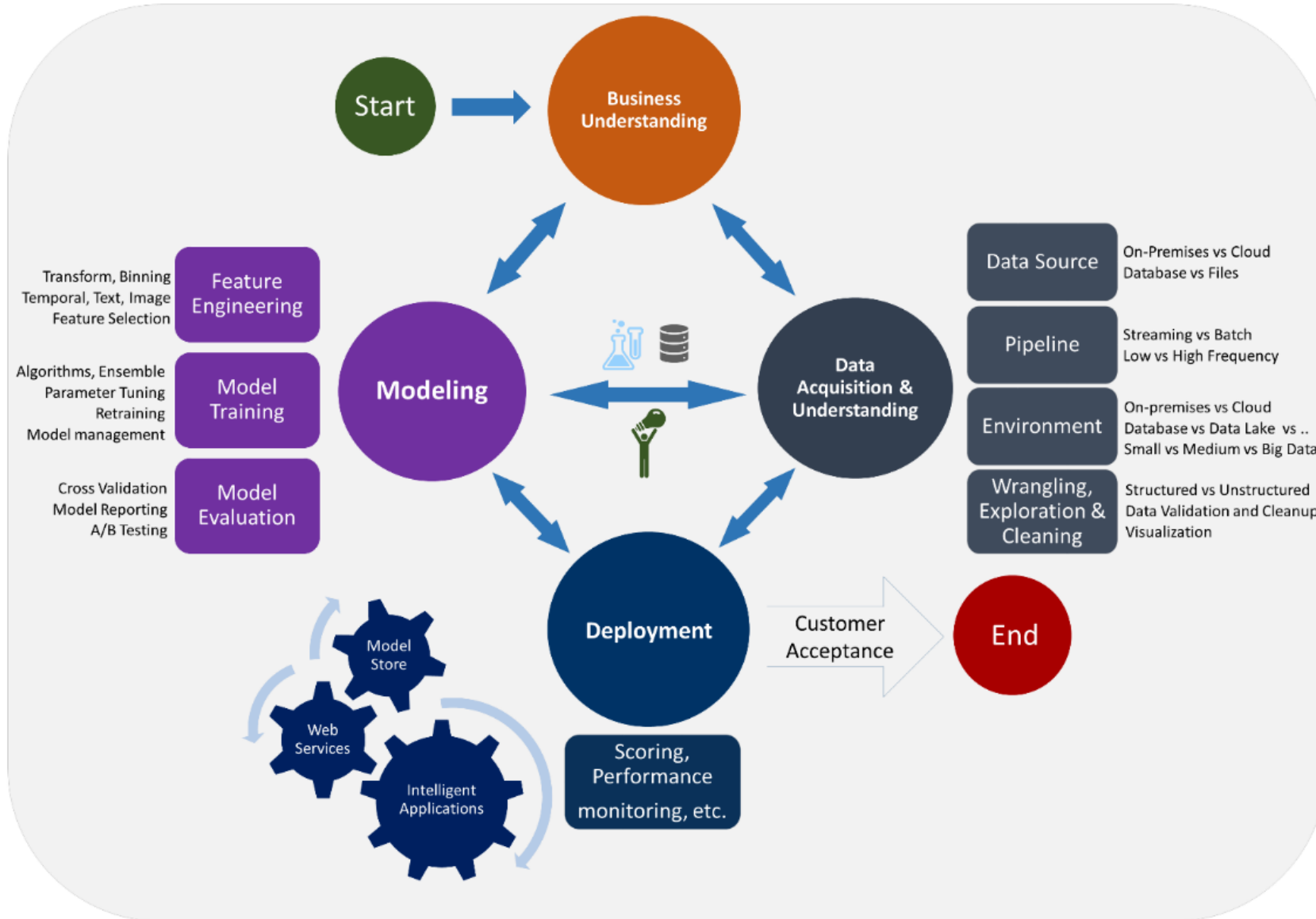
About this Playbook	2	Hire & Train	80
Partner Practice Development Framework	5	Executive Summary	81
What is Artificial Intelligence?.....	6	Hire, build, and train your team	82
AI Opportunity	8	Job Descriptions for your Technical Team	88
Industry Opportunities	9	Recruiting Resources.....	98
Define Your Strategy	17	Training & Readiness.....	99
Executive Summary.....	18	Operationalize	113
Define Your Practice Focus	19	Executive Summary	114
Understanding the AI Practice.....	20	Implement a Process.....	115
The Microsoft Approach to AI.....	23	Claim Your Internal Use Benefits	119
Pre-Built AI using Cognitive Services.....	28	Define Customer Support Program and Process	124
Building Custom AI	35	Manage and Support an AI solution in Azure.....	128
Microsoft AI Platform Summary.....	42	Support Ticket Setup and Tracking	130
Define and Design the Solution Offer.....	43	Implement Intellectual Property Offerings	131
Understanding Project Based Services.....	44	Setup Social Offerings	132
Understanding Managed Services	54	Create Engagement Checklists & Templates.....	133
Accelerate your Managed Service Model.....	60	Go to Market & Close Deals	134
Understanding Intellectual Property	61	Executive Summary	135
Define Industry Specific Offerings.....	65	Marketing to the AI Buyer.....	136
Define Your Pricing Strategy.....	66	Engage Technical Pre-Sales in Sales Conversations.....	138
Calculate Your Azure Practice Costs	69	Architecture Design Session (ADS)	140
Identify Partnership Opportunities	71	Go-to-Market and Close Deals Guide	142
Define Engagement Process.....	73	Optimize & Grow	143
Identify Potential Customers.....	74	Executive Summary	144
Join the Microsoft Partner Network	75	Understanding Customer Lifetime Value	145
Stay Informed on AI Matters	77	Guide: Optimize and Grow	147
Identify Solution Marketplaces.....	78	AI Playbook Summary	148

Microsoft AI Platform

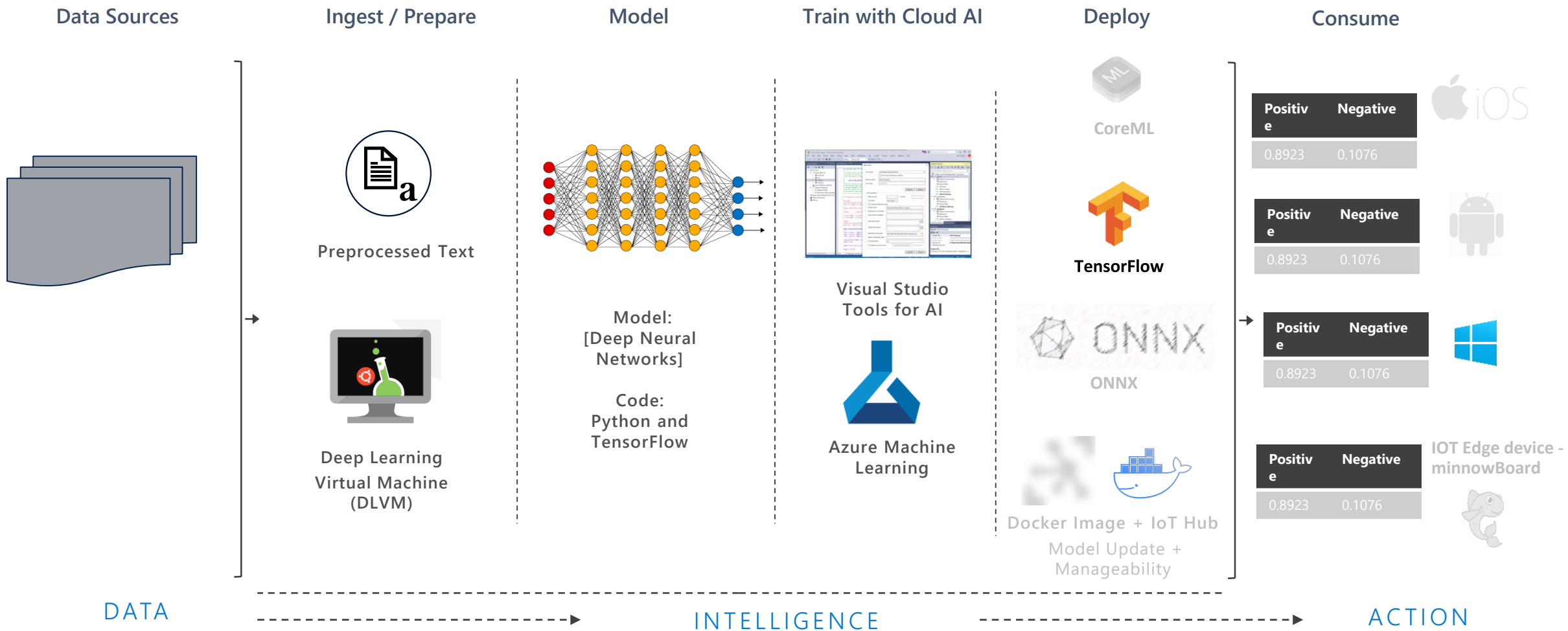


See <https://azure.microsoft.com/en-us/overview/ai-platform> for more information about the various services and features of the Microsoft AI Platform

Data Science Lifecycle



Sample Real World ML Pipeline Architecture



Model Development

Training subset

Single Machine

API Wrapper

Code, Model and training subset

Model Framework

Library Dependencies

Runtime

Drivers



Windows



MacOS



Linux

Model Training

Full training data

1000s of GPUs

API Wrapper

Code, Model and training data

Model Framework

Library Dependencies

Runtime

Container Orchestration

Drivers



kubernetes



docker

Deployed Model

Production data

1000s of Nodes

Ensemble Model Routing

Outlier Detection

API Wrapper

Code + Trained Model

Model Framework

Library Dependencies

Runtime

Container Orchestration

Drivers

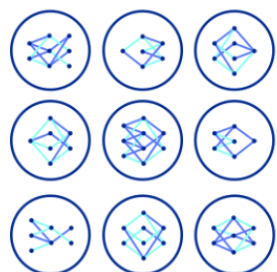


kubernetes

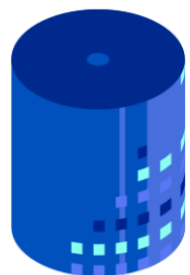


docker

AI Layer Architecture



Pre-trained ML Models

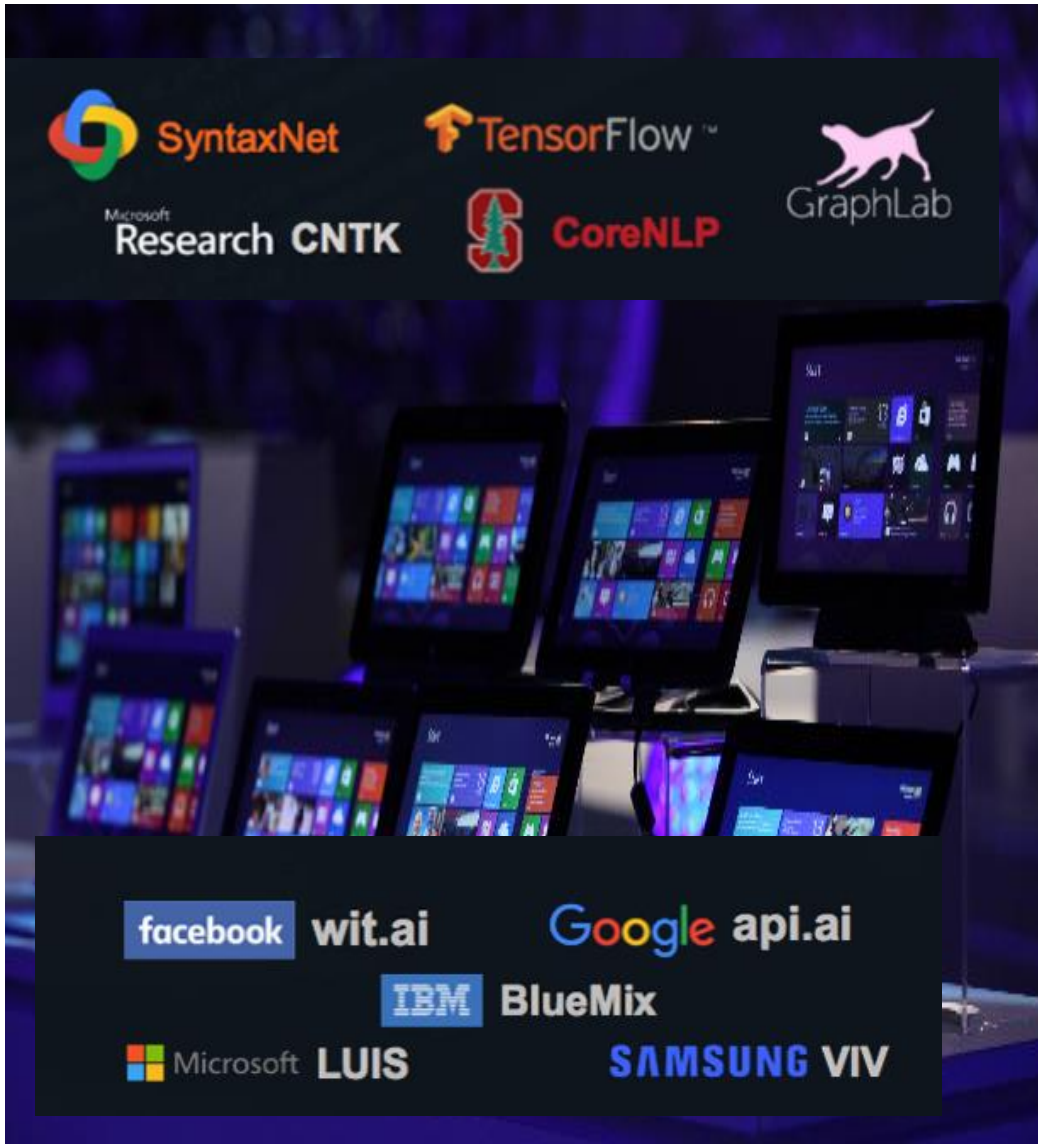


Data

Databases /
Datawarehouses /
Data lakes



End user applications



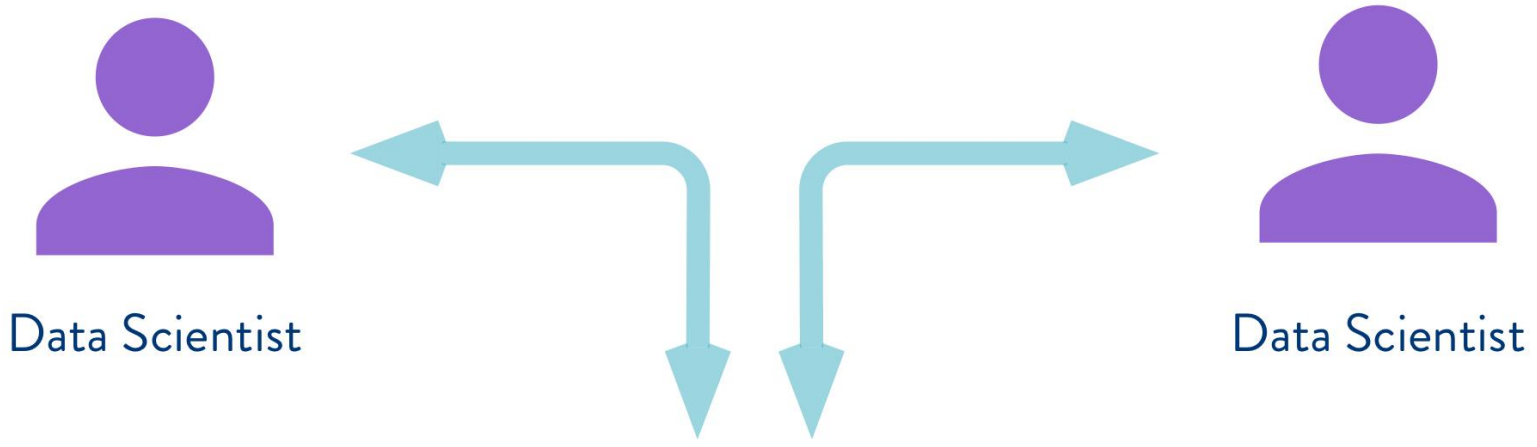
Common AI/ML Problems:

- Most libraries provide state-of-the-art algorithms but little pertinent training data
- For many conversational domains, training data may be difficult or impossible to collect
- Pre-built domains streamline development but are largely irrelevant for most apps
- Tools for building custom domains can only handle narrow models and trivial apps
- ML capabilities only scratch the surface of what is typically required for production apps

Machine Learning Development Lifecycle provides customized end to end solution from formal problem definition, domain modeling, creating training and test data, training models, evaluation of model, execution, deployment, and visualization.

Key Value Proposition:

- Not just offer an NLP library but provide expertise to work with bot framework for multiple modalities, commerce engine integration, and deployment infrastructure and expertise.



Code Data & Models

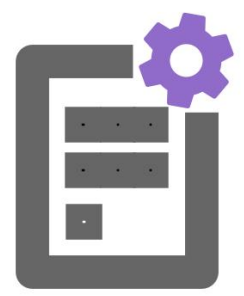


Git Server

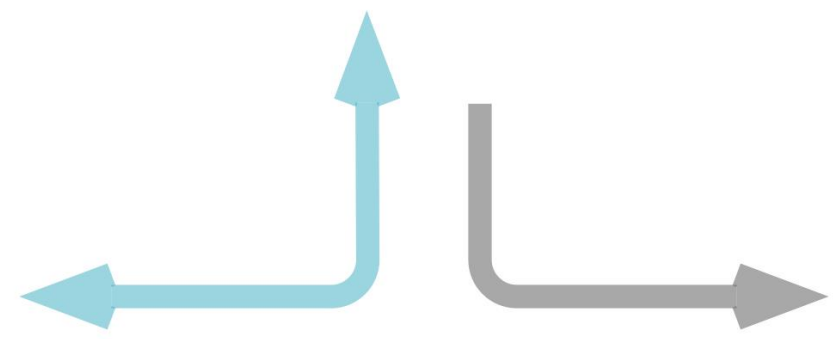
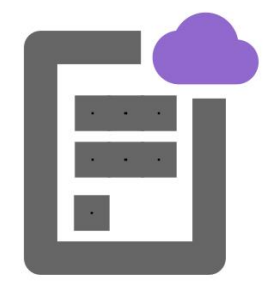


S3, GCP, SSH, etc

Training

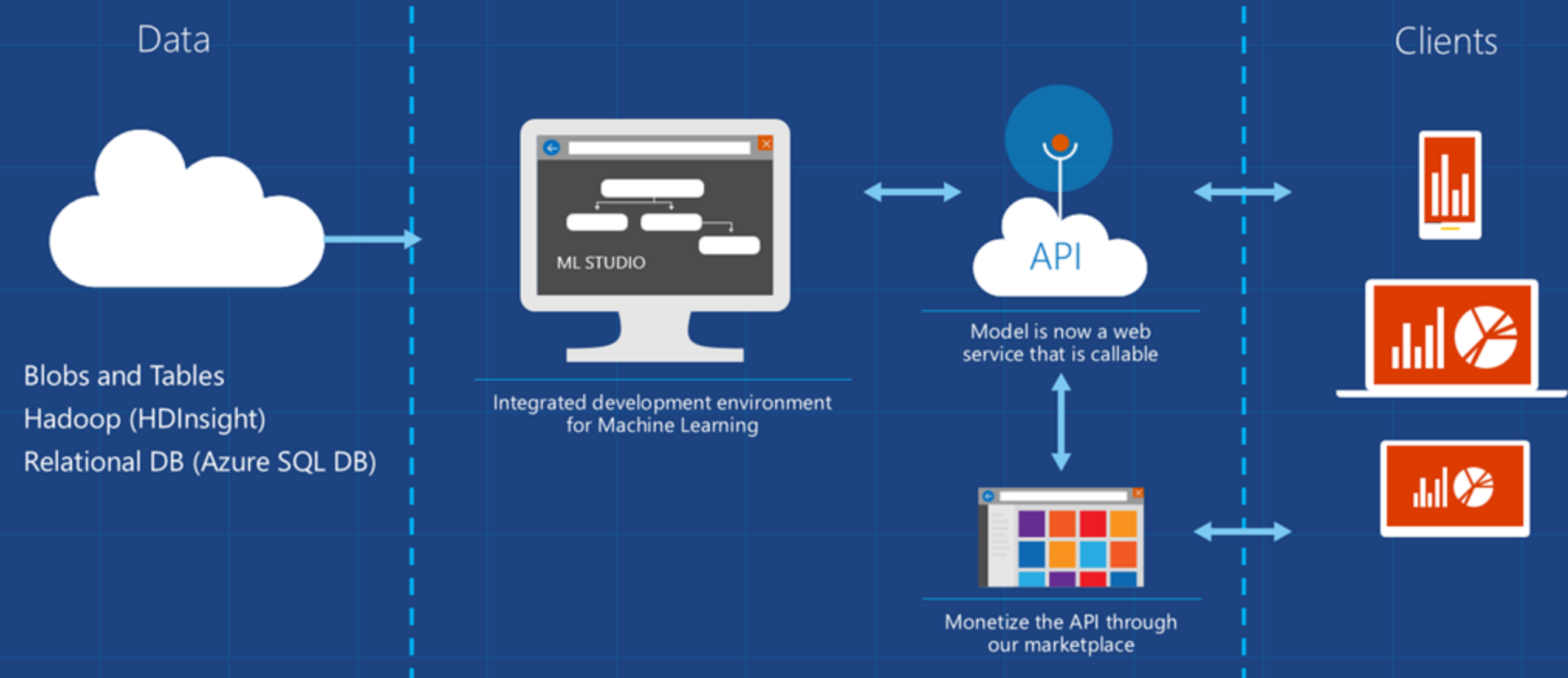


Serving



Azure Machine Learning Service

Data -> Predictive model -> Operational web API in minutes

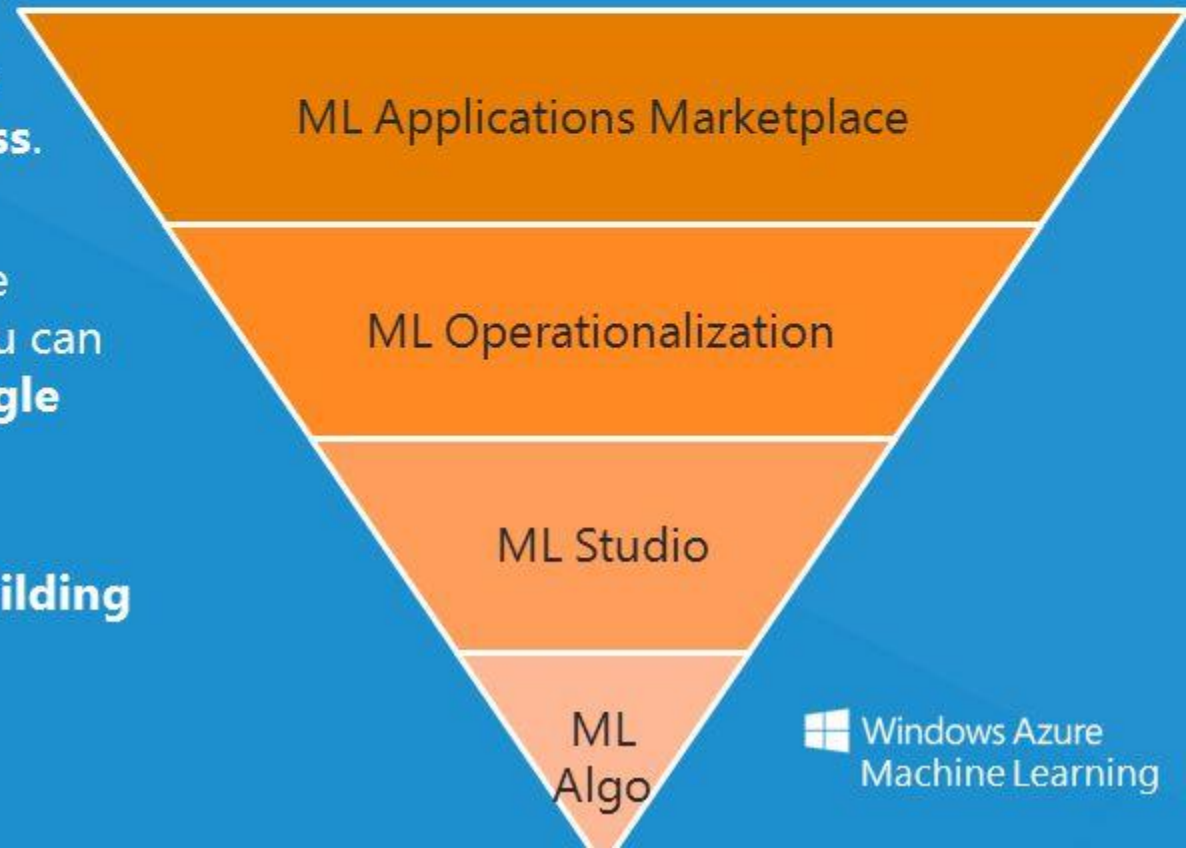


Azure Machine Learning - vision

Vision:

Make machine learning (ML) accessible to every enterprise, data scientist, developer, information worker, consumer, and device anywhere in the world.

- ML Marketplace: a marketplace/appstore for intelligent web services where an external customer can come and **consume web service applications that are relevant to their business.**
- ML operationalization: a cloud service that can host a massive selection of intelligent web services, automatically scaling. You can **put any machine learning model into production by a single click.**
- ML Studio: a easy to use browser-based solution for **rapid building and experimenting with predictive models.**
- ML Algorithms – best in class ML Algorithms and models



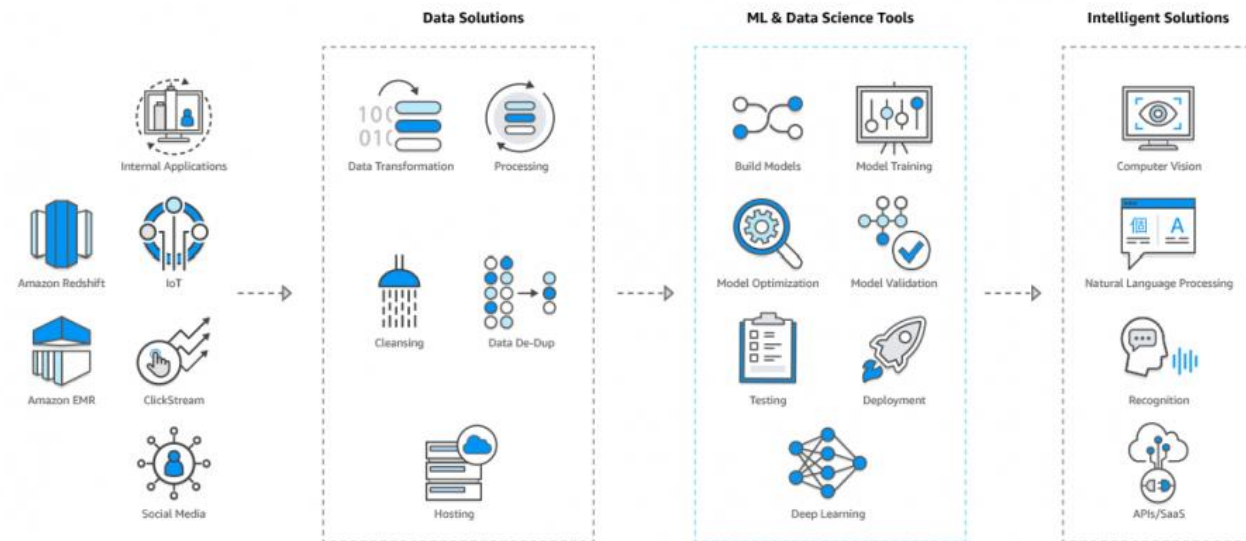
Machine Learning & Artificial Intelligence

Build intelligent applications with machine learning and data science software.

Data Solutions

ML & Data Science Tools

Intelligent Solutions



Benefits of AI in AWS Marketplace



Scalable

Solutions from software vendors in AWS Marketplace dramatically reduce your effort to deploy, scale, and maintain infrastructure, freeing up your time for focusing on data and model building.



Accessible and Fast

With AWS Marketplace, you can subscribe to and purchase solutions with one-click and use SaaS applications via your browser or via RESTful APIs.

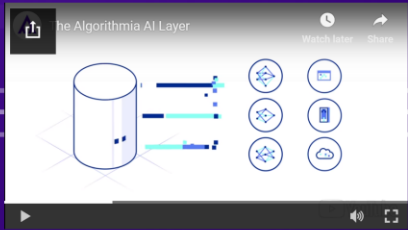


Pay as you go

AWS Marketplace has flexible payment options, like pay-as-you-go to monthly, or annual or multi-year terms.

Deploy and Manage ML Models the Smart Way

SIGN UP SEE A DEMO



Automate DevOps for ML

- ✓ Deploy any model in minutes
- ✓ Advanced Monitoring



Accelerate Your Team

- ✓ Empower your data scientists
- ✓ Collaborate Across Org



Optimize Hardware

- ✓ Serverless Microservices
- ✓ Optimized GPU usage

Welco

Face Detection

dlib/FaceDetection

This algorithm detects human faces in given images.

↓ 40k requests

Face Detection

opencv/FaceDetection

Detect faces in images

↓ 18k requests

Face Detection Box

opencv/FaceDetectionBox

Detects the faces in an image and returns an array of rectangles that contains the faces.

↓ 2.0k requests

Intuitive Machine Learning for Engineers

A platform for discovering, sharing, and discussing easy to use and pre-trained machine learning models.

Browse By Model Type

I Can't Find What I'm Looking For

Images

Tag images or extract features


Text

Understand or summarize texts

Generate

Generate novel text, images, etc.


New: Check out the models with [Live Demo](#) to test them on our servers in seconds.



Tiny YOLO in Javascript
by mikeshi [Live Demo](#) ♥ 9 ↓ 289

Detect objects in images right in your user's browser using Tensorflow.js!


[CV](#) [Mobile](#) [Food and Drink](#)



PSPNet ♥ 9 ↓ 213
by hellochick [Live Demo](#)

Detect object and partition a digital image into multiple segments at outdoor/indoor scenes.

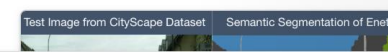
[CV](#) [NN](#) [CNN](#)



ICNet for Fast Segmentation
by oandrienko ♥ 3 ↓ 17

Perform real-time, high accuracy semantic segmentation on high resolution images

[CV](#) [Mobile](#) [NN](#)

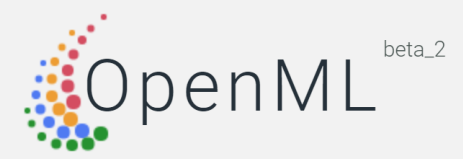


Try models and contribute feedback through joining the ModelDepot Community! [Join The Community](#)

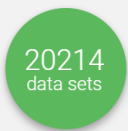
Message from Mike
Hey there, I'm Mike (not just a bot 🤖)! It would mean the world 🌍 to me if you could share what you're looking for or if you need any help! 😊

Compose your reply...

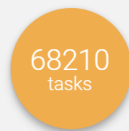




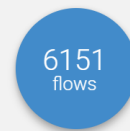
Machine learning, better, together



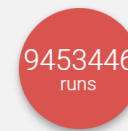
Find or add **data** to analyse



Download or create scientific **tasks**



Find or add data analysis **flows**



Upload and explore all **results** online.



HACKATHON

Bring your own data, bring your own algorithms, or build cool new features.

Next location: 17-21 September 2018, Paris, France



tensorflow / models

Watch 2,792 | Star 45,522 | Fork 27,661

Code | Issues 979 | Pull requests 326 | Projects 2 | Wiki | Insights

Models and examples built with TensorFlow

3,034 commits | 81 branches | 8 releases | 460 contributors | Apache-2.0

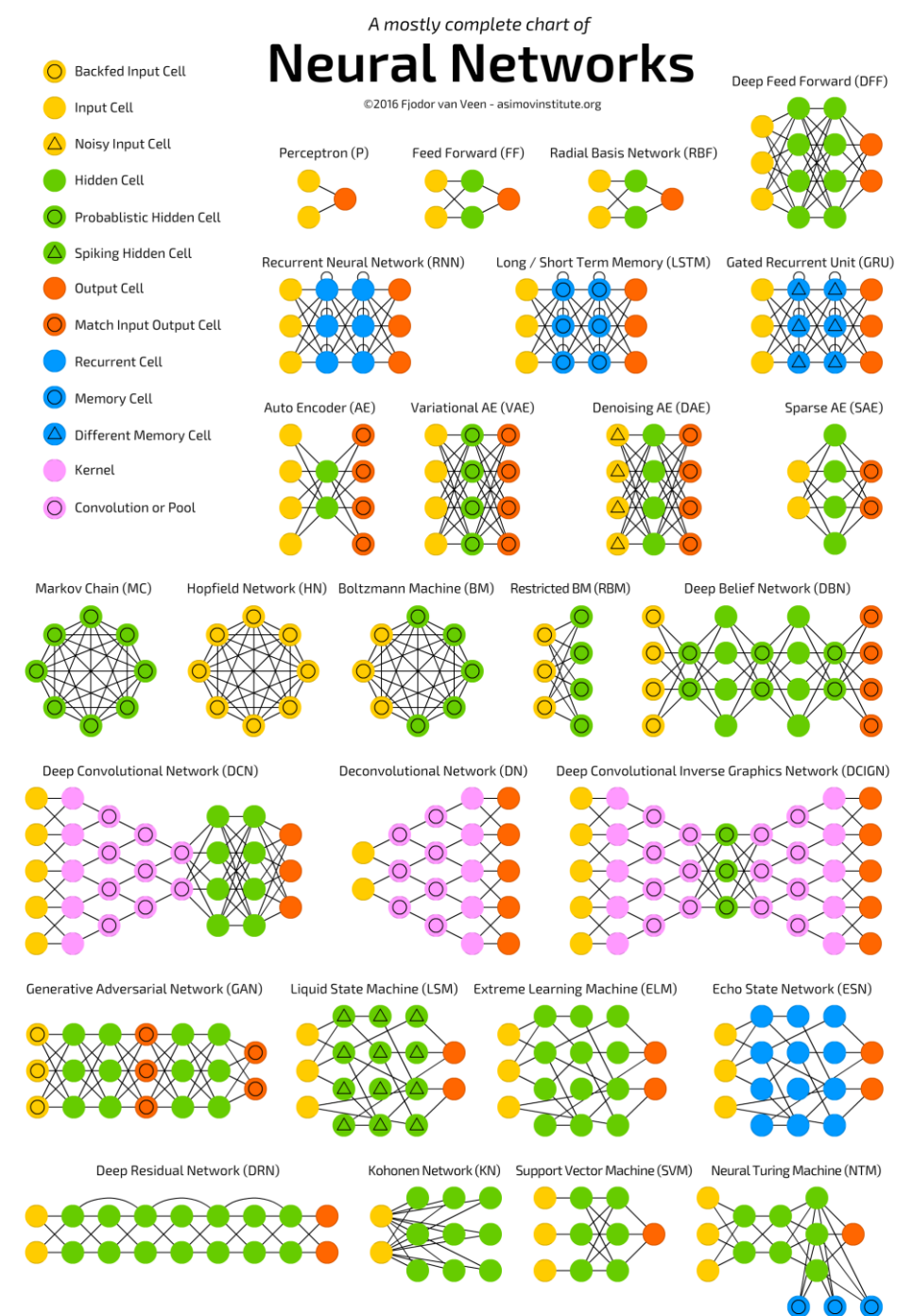
Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

cshallue Merge pull request #5862 from cshallue/master ... Latest commit 2c18130 2 days ago		
official	worker_devices moving to ds.extended (#5775)	16 days ago
research	Merge pull request #5862 from cshallue/master	2 days ago
samples	Fix #5814	6 days ago
tutorials	update the calculation of num_batches_per_epoch	a month ago
.gitignore	Fixed gitignore for mac's ds_store (#4012)	4 months ago
.gitmodules	Move the research models into a research subfolder (#2430)	a year ago
AUTHORS	Spatial Transformer model	3 years ago
CODEOWNERS	Fix dependency issues (#5815)	10 days ago
CONTRIBUTING.md	Fixing small typo	a year ago
ISSUE_TEMPLATE.md	Update ISSUE_TEMPLATE.md	10 months ago
LICENSE	Update LICENSE	3 years ago

THE ASIMOV INSTITUTE



With new neural network architectures popping up every now and then, it's hard to keep track of them all. Knowing all the abbreviations being thrown around (DCIGN, BiLSTM, DCGAN, anyone?) can be a bit overwhelming at first.



Standard?

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)



ONNX Motivation

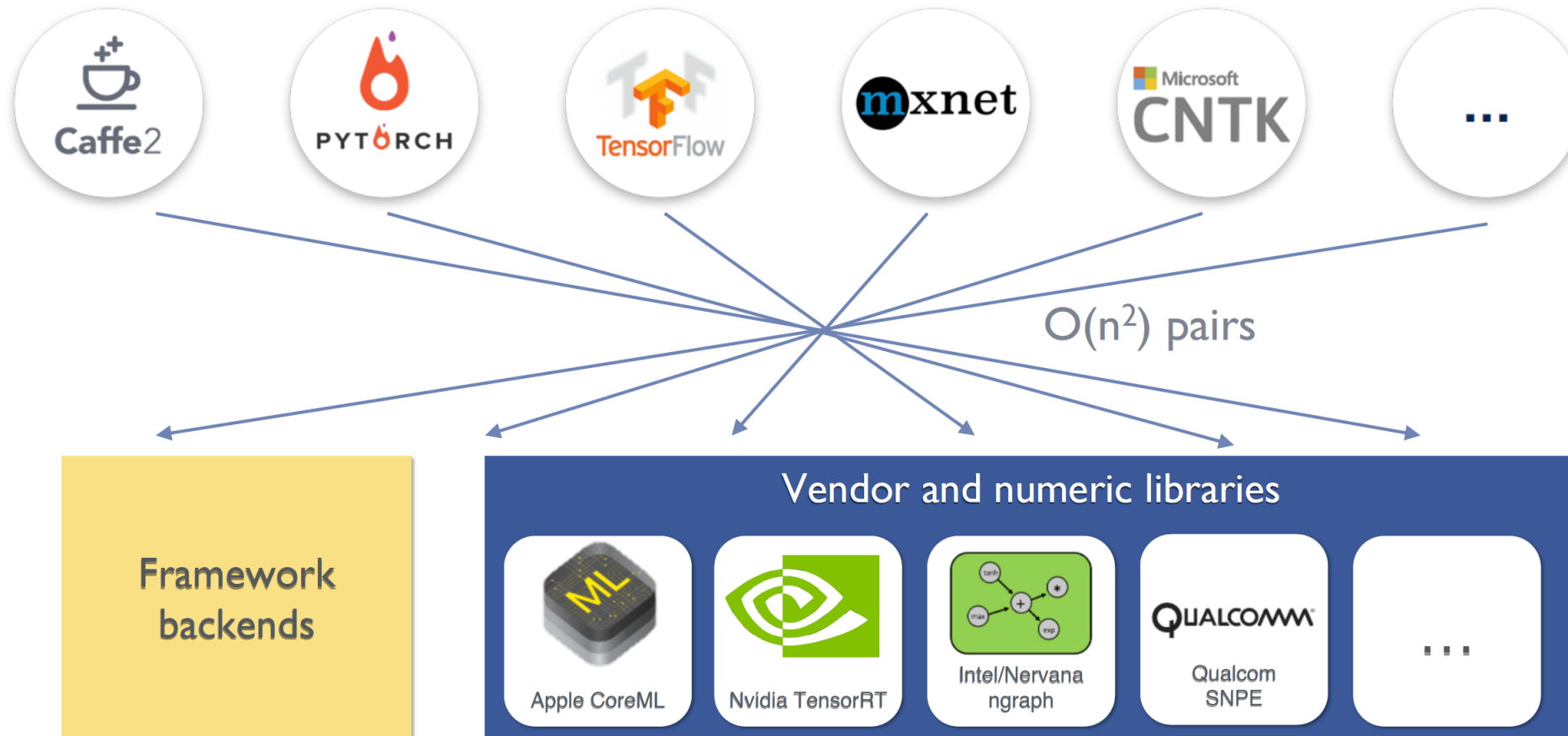
Allow interoperability between frameworks

Starting with CNTK, Caffe2 and PyTorch

Allow hardware vendor to focus on one IR in their backend optimization

Allow train in one toolkit and deploy in another

Deep Learning Frameworks Zoo



Open Neural Network Exchange



Caffe2



PYTORCH



TensorFlow



ONNX

Shared model and operator representation

From $O(n^2)$ to $O(n)$ pairs

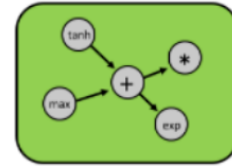
Framework
backends



Apple CoreML



Nvidia TensorRT



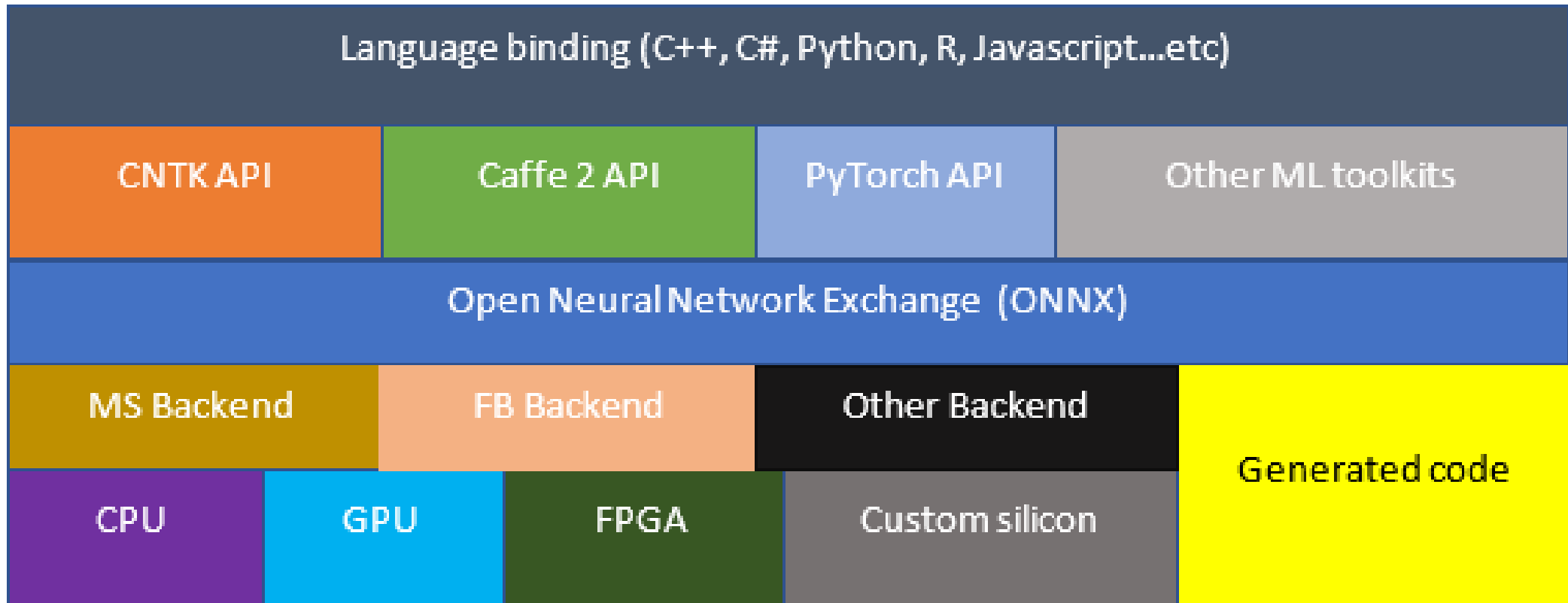
Intel/Nervana
ngraph



Qualcom
SNPE



ONNX Vision





ONNX

Linux	Windows
build passing	build passing

[Open Neural Network Exchange \(ONNX\)](#) is the first step toward an open ecosystem that empowers AI developers to choose the right tools as their project evolves. ONNX provides an open source format for AI models. It defines an extensible computation graph model, as well as definitions of built-in operators and standard data types. Initially we focus on the capabilities needed for inferencing (evaluation).

Caffe2, PyTorch, Microsoft Cognitive Toolkit, Apache MXNet and other tools are developing ONNX support. Enabling interoperability between different frameworks and streamlining the path from research to production will increase the speed of innovation in the AI community. We are an early stage and we invite the community to submit feedback and help us further evolve ONNX.



PyTorch

PyTorch is the framework for *AI research* at Facebook which enables rapid experimentation

Flexibility

Debugging

Dynamic neural networks

Not optimized for production and mobile deployments (Python)

When research projects produce valuable results, *the models need to be transferred to production.*

Traditionally, rewriting the training pipeline in a product environment with other frameworks.

ONNX Runtime for inferencing machine learning models now in preview

Posted on October 16, 2018



 **Faith Xu**, Senior Program Manager, Machine Learning Platform

We are excited to release the preview of ONNX Runtime, a high-performance inference engine for machine learning models in the [Open Neural Network Exchange \(ONNX\)](#) format. ONNX Runtime is compatible with ONNX version 1.2 and comes in Python packages that support both [CPU](#) and [GPU](#) to enable inferencing using [Azure Machine Learning service](#) and on any Linux machine running Ubuntu 16.

ONNX is an open source model format for deep learning and traditional machine learning. Since we launched ONNX in December 2017 it has gained support from more than 20 leading companies in the industry. ONNX gives data scientists and developers the freedom to choose the right framework for their task, as well as the confidence to run their models efficiently on a variety of platforms with the hardware of their choice.



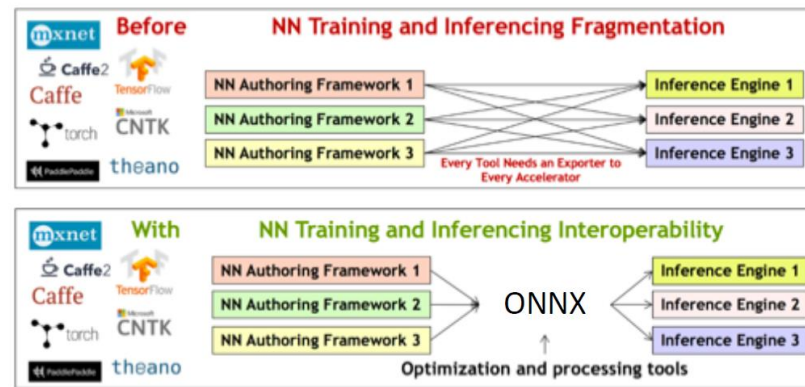
Importing and Exporting from frameworks

Framework / tool	Installation	Exporting to ONNX (frontend)	Importing ONNX models (backend)
Caffe2	onnx/onnx-caffe2	Exporting	Importing
PyTorch	part of pytorch package	Exporting, Extending support	coming soon
Cognitive Toolkit (CNTK)	built-in	Exporting	Importing
Apache MXNet	onnx/onnx-mxnet	coming soon	Importing [experimental]
Chainer	chainer/onnx-chainer	Exporting	coming soon
TensorFlow	onnx/onnx-tensorflow	coming soon	Importing [experimental]
Apple CoreML	onnx/onnx-coreml	coming soon	Importing



Interoperability

- Having at disposal several libraries how we can interoperate between them for reusing training for inference, or transfer learning?
- Fight against fragmentation



- For a while Caffe models have been used for exchange, ONNX or NNEF are proposed as interoperable solutions
 - **Open Neural Network Exchange Format or Neural Network Exchange Format**
- Tools around ONNX
 - Direct or indirect support for specific libraries
 - Runtime support by Nvidia TensorRT

ONNX

- Which kind of format is ONNX?
 - Based on Google Protobuf serialization
 - Describes network layers eventually with trained parameters
 - Node, Graph, Attribute, Operator, Value, Shape
 - All operators here:
<https://github.com/onnx/onnx/blob/master/docs/Operators.md>
- Example with TF
 - <https://github.com/onnx/tutorials/blob/master/tutorials/OnnxTensorflowImport.ipynb>
- Repository of Pre-trained Networks
 - <https://github.com/onnx/models>
 - E.g. ResNet-50 is 92MB

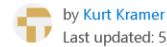
Browser address bar: <https://azure.microsoft.com/en-us/resources/samples/cognitive-services-onnx-customvision-sample/>

Microsoft Azure | Contact Sales: 1-800-867-1389 | Search | My account | Portal | Sign in

Overview | Solutions | Products | Documentation | Pricing | Training | Marketplace | Partners | Support | Blog | More | [Free account >](#)

[Samples](#) / [Cognitive Services](#) / Sample application for ONNX models exported from Custom Vision Service

Sample application for ONNX models exported from Custom Vision Service



by Kurt Kramer

Last updated: 5/8/2018 [Edit on GitHub](#)

[Browse on GitHub](#)

[Download as .zip](#)

This sample application demonstrates how to take a model exported from the [Custom Vision Service](#) in the ONNX format and add it to an application for real-time image classification.

Getting Started

Prerequisites

- [Windows SDK - Build 17110+](https://www.microsoft.com/en-us/software-download/windowsinsiderpreviewSDK)(<https://www.microsoft.com/en-us/software-download/windowsinsiderpreviewSDK>)
- [Visual Studio 17](#)
- [Windows 10 Insider Preview](#)
- An account at [Custom Vision Service](#)

Quickstart

- clone the repository and open the project in Visual Studio
- Build and run the sample Application

Open community

- Framework agnostic
- GitHub from the beginning
- Close partnerships and OSS contributions



Facebook
Open Source

Microsoft



NVIDIA

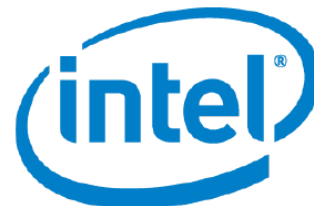


arm

QUALCOMM[®]



HUAWEI





ONNX

Get Involved!

ONNX is a community project.

<https://onnx.ai>

<https://github.com/onnx>



Facebook
Open Source

Microsoft



Microsoft
Cognitive
Toolkit



CNTK Latest Features (v2.2, v2.3)

New tutorials/examples/manuals

NCCL2 support

MKL-DNN integration

ONNX support

C#/.NET API

R-binding for CNTK

Model simplification/compression support

New ops and perf-improvements

Tensorboard support

Open Neural Network Exchange (ONNX)

ONNX is an open format to represent deep learning models

Supported by:

CNTK

PyTorch

Caffe 2

MxNet

Enabled interop-ability between frameworks

For more information: <https://onnx.ai/>



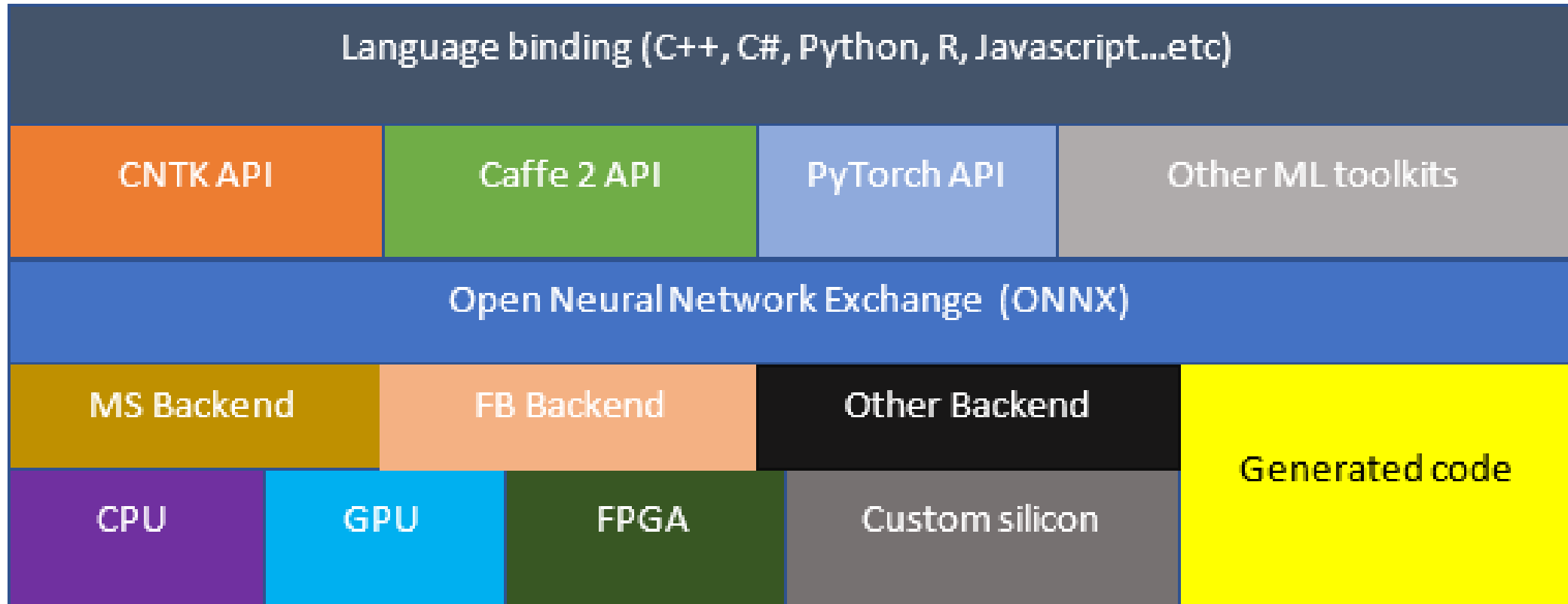
ONNX Motivation

Allow interoperability between frameworks

Allow hardware vendor to focus on one IR in their backend optimization

Allow train in one toolkit and deploy in another

ONNX Vision



ONNX Status in CNTK

V1 release in Github, focus on the basics

Support only inference, no loop, no condition and no gradient

Supported by CNTK, Caffe2, PyTorch and MxNet

Upcoming work:

Refined RNN support

Loop and control

Converter for other toolkits are coming soon

Open Neural Network Exchange (ONNX)

An open source intermediate representation (IR) of computation graph (<https://github.com/onnx/onnx>)

With defined common OPs and their semantics

Released on Sep. 7, 2017

Collaboration between Microsoft and Facebook

A share library with a Caffe2 example as reference

Permissive MIT license and no patents

In Conclusion

Think Operationalization

You have choices

Sharing is Caring

Microsoft Azure
+ AI Conference

CO-PRODUCED BY

Microsoft & DEVintersection

Thank You!

<https://ONNX.AI>

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