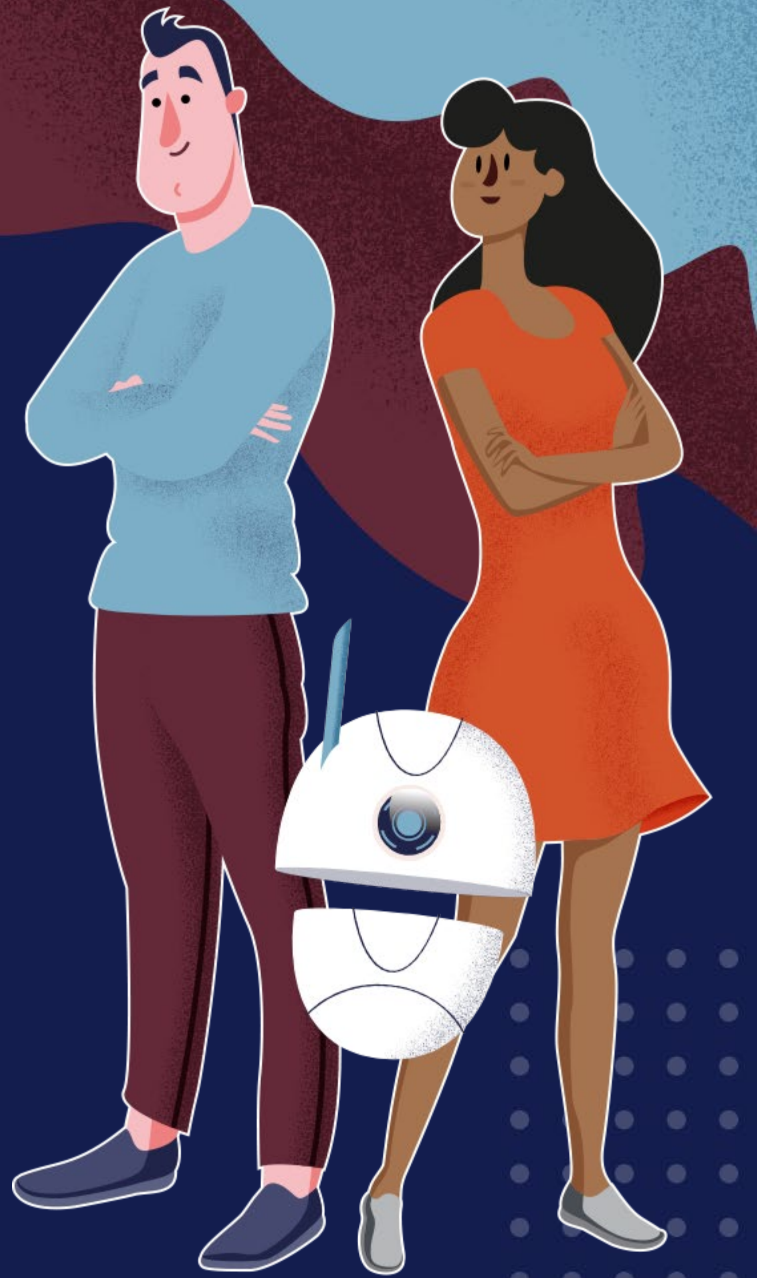
global AI bootcamp

March 4th 2023, Torino - Italy



Azure ML And ONNX

Mauro Bennici

AI Architect and AI Ethicist

global AI community

Deltatre Innovation Lab

#GlobalAIBootcamp

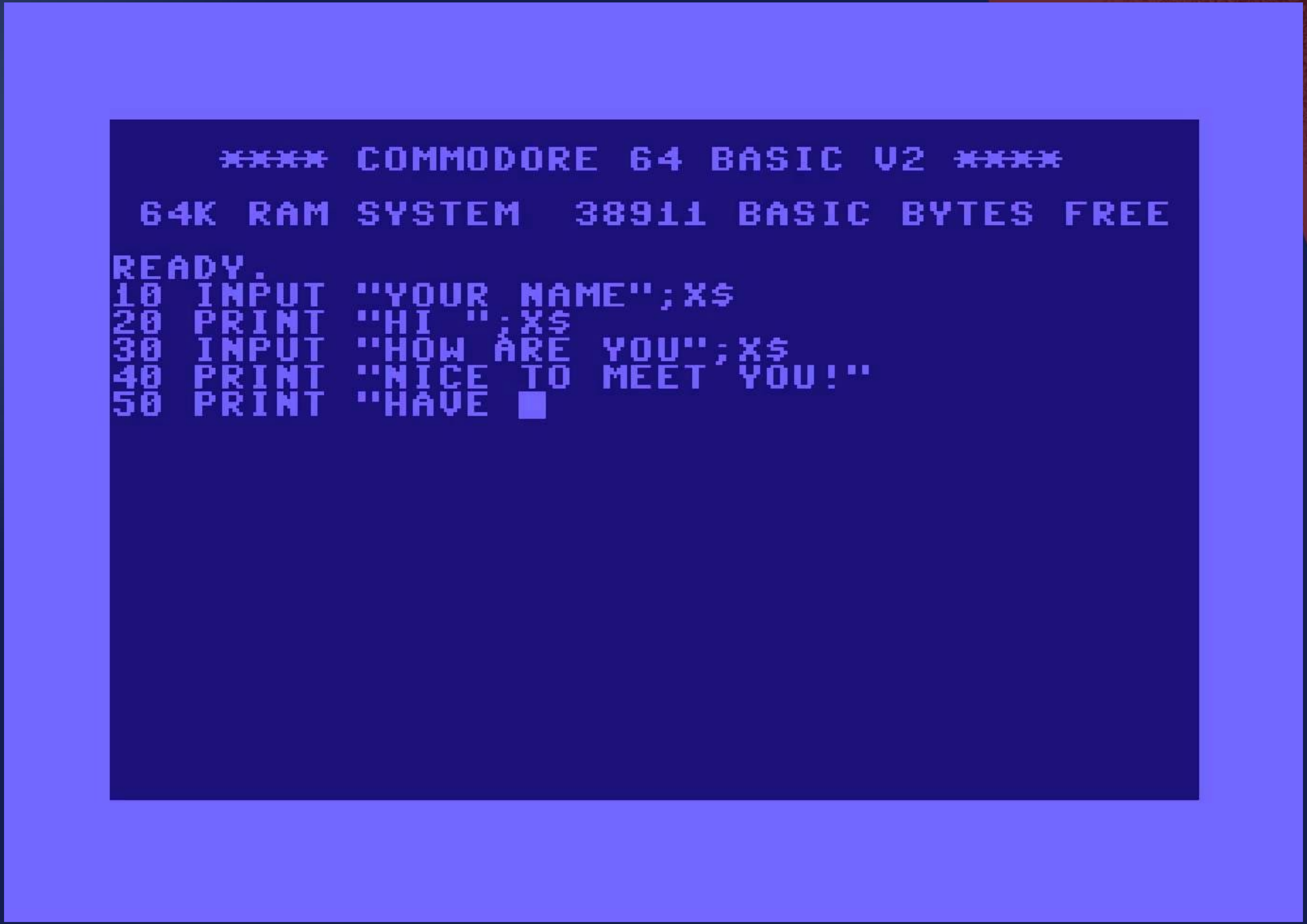
Mauro Bennici

AI Architect, Kaggle student mentor, .NET Foundation member, Data Scientist, Professional Scrum Master (PSM I), Microsoft Certified Trainer (MCT), Azure Cloud Solutions Associate (MCSA). Member of the SME Digital Alliance Working Group for Ethics and Artificial Intelligence.

@ maurobennici

https://www.linkedin.com/in/maurobennici/

Everyone has their own origins

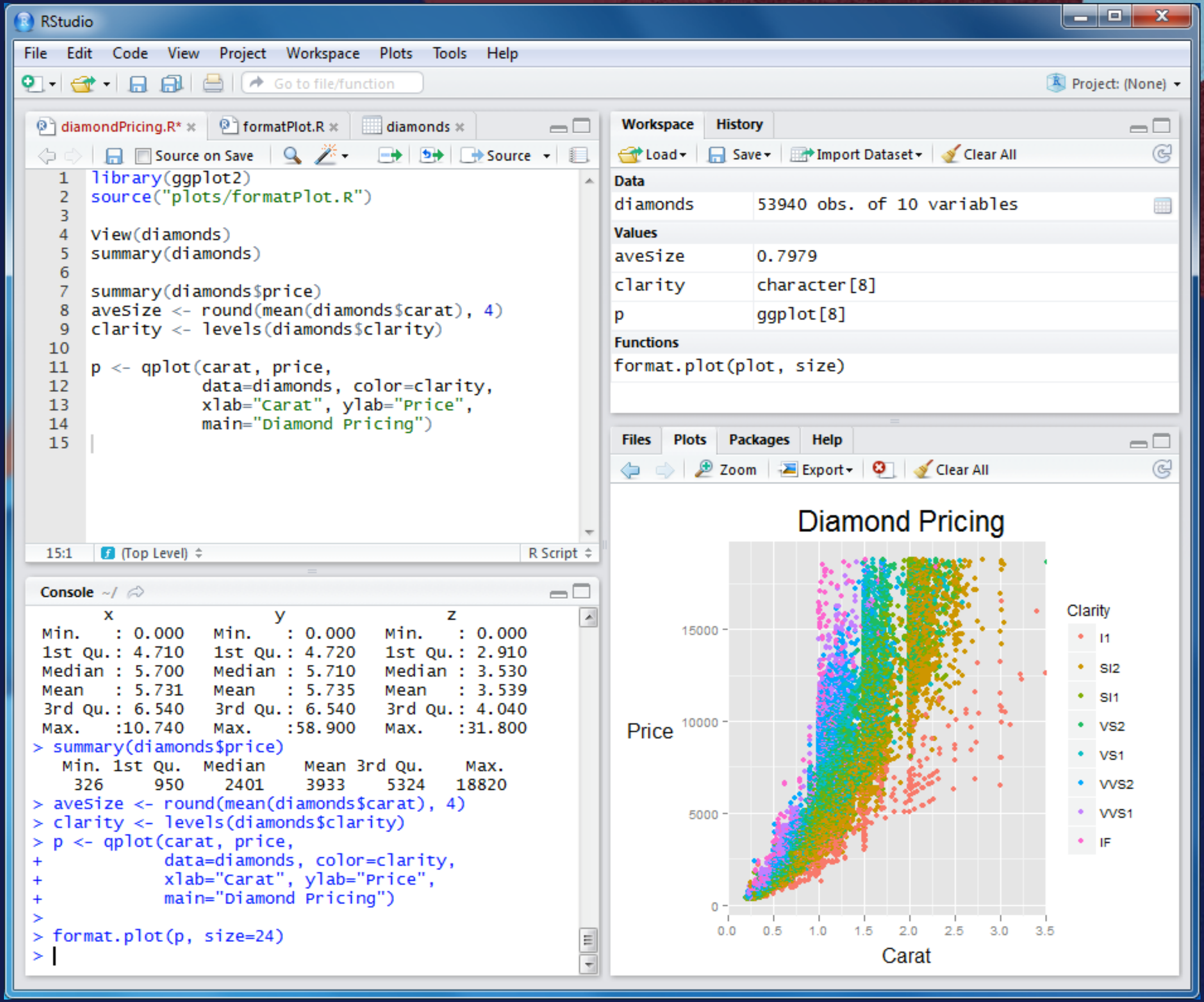


```
    **** COMMODORE 64 BASIC V2 ****

 64K RAM SYSTEM  38911 BASIC BYTES FREE

READY.
10 INPUT "YOUR NAME";X$
20 PRINT "HI ";X$
30 INPUT "HOW ARE YOU";X$
40 PRINT "NICE TO MEET YOU!"
50 PRINT "HAVE █
```

The cloud dilemma

The old way

- Data science was a laboratory science
- Data and processing took place in private data centers
- Custom software was a lot



```r
library(ggplot2)
source("plots/formatPlot.R")

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")
```

Data

diamonds	53940 obs. of 10 variables

Values

aveSize	0.7979
clarity	character[8]
p	ggplot[8]

Functions

format.plot(plot, size)

```
       x                 y                 z          
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910  
 Median : 5.700   Median : 5.710   Median : 3.530  
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539  
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :10.740   Max.   :58.900   Max.   :31.800  
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    326     950    2401    3933    5324   18820 
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
> |
```

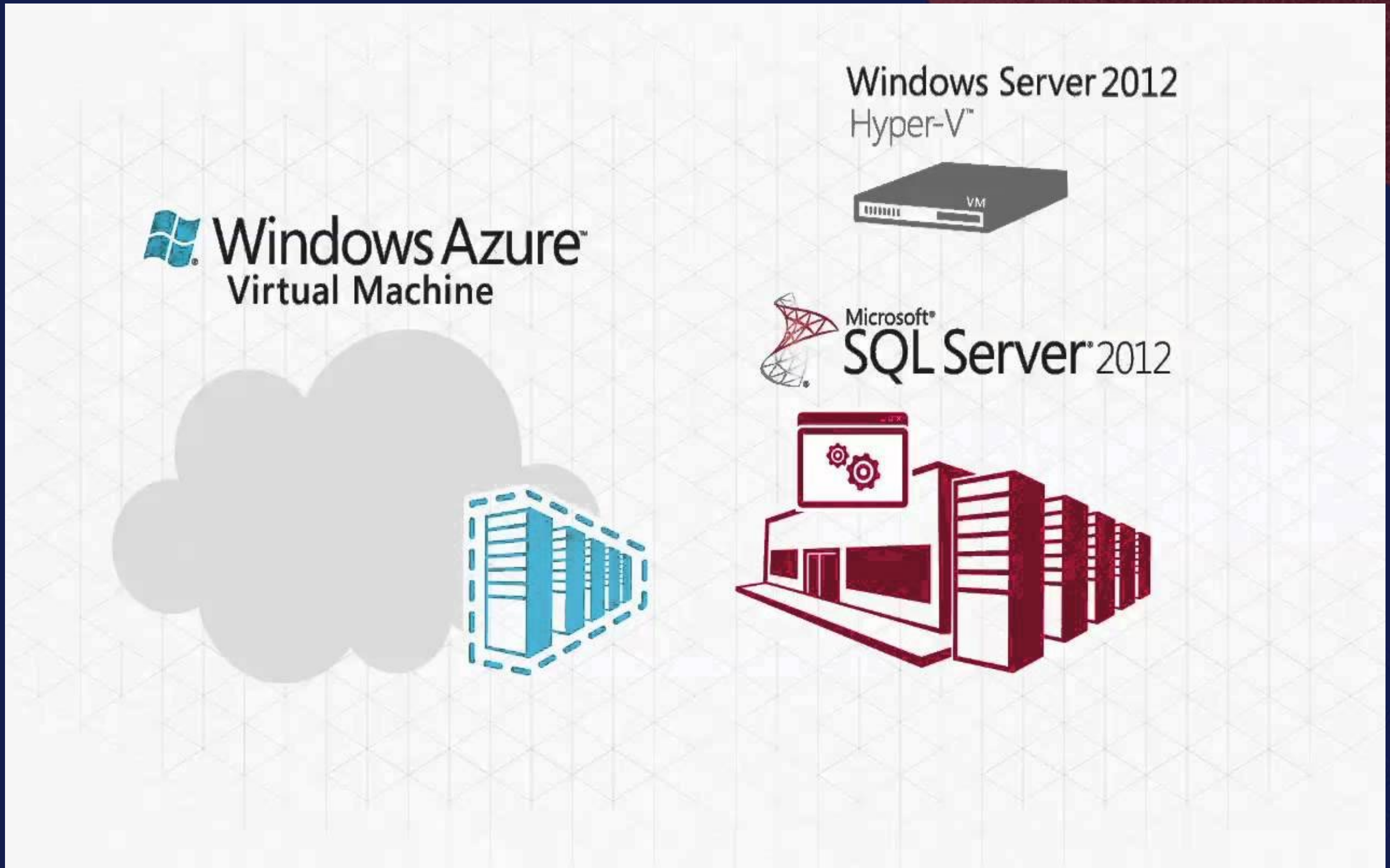
global AI community

Deltatre Innovation Lab

#GlobalAIBootcamp

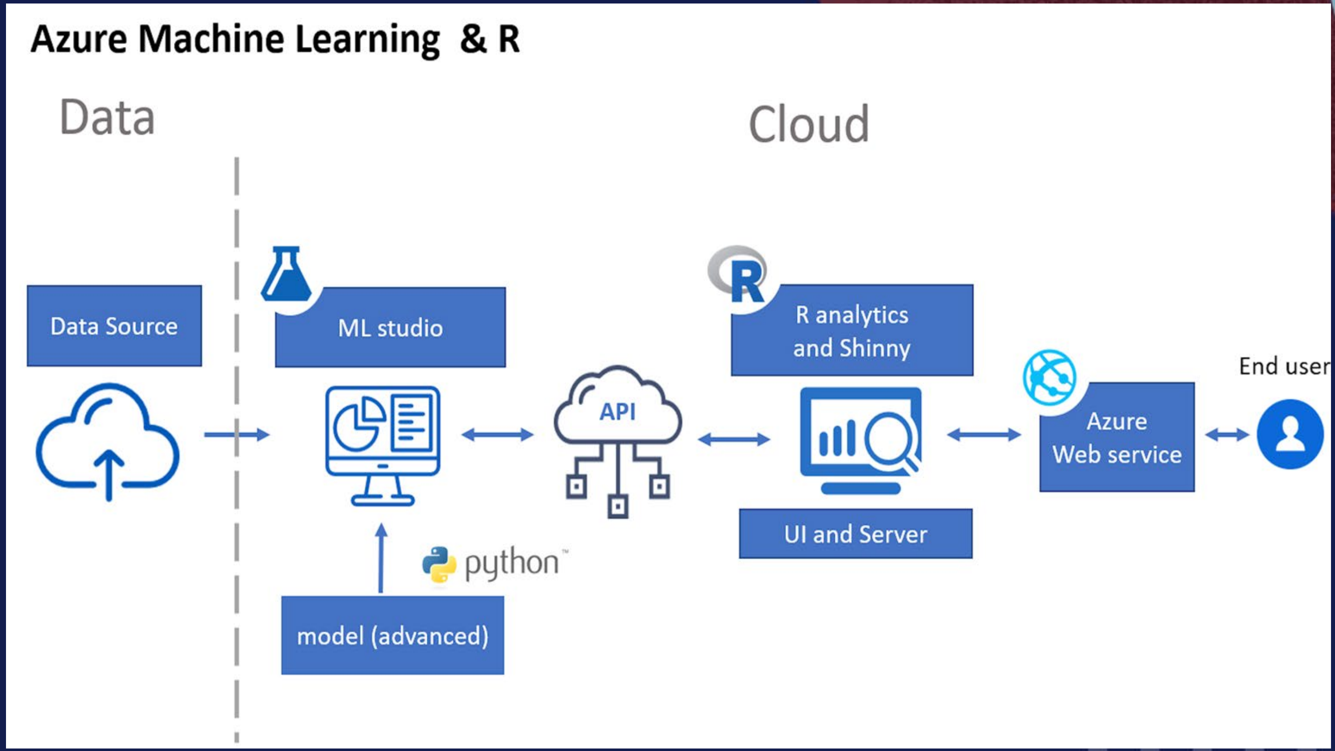
The first cloud era

- Here comes the cloud
- Online transposition of on-prem applications
- Problem shifted



The arrival of ML services

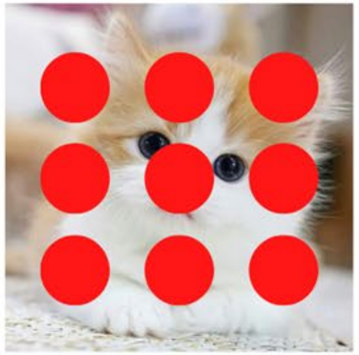
- Specialized services such as AzureML
- Everything in one place (data lake, sql, nosql, files, models, APIs)
- Vendor lock-in?
- Hardware lock-in?

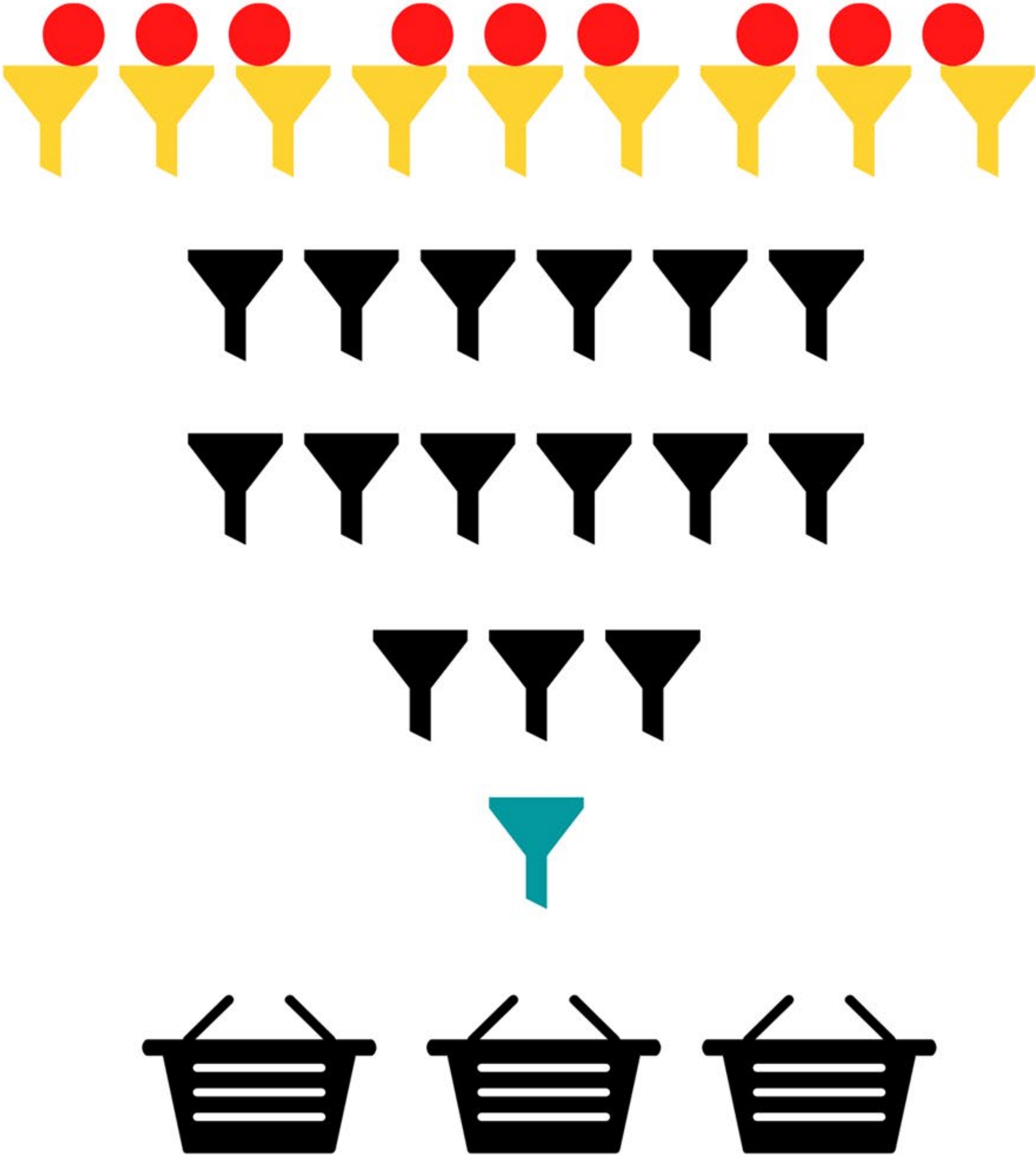


What about AI models?

- Frameworks updated frequently with breaking changes
- Custom implementations for different languages
- Incompatibility between models (frameworks)
- Increasingly high costs

One ONNX to gherm them…





Ha'DIbaH mangHom uSgheb

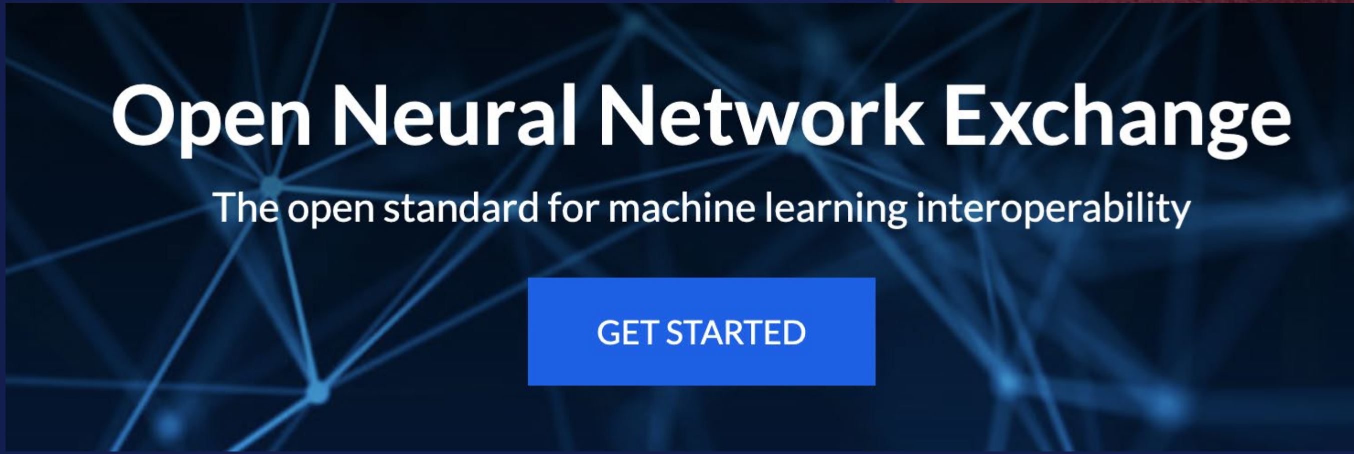
Convert

- The models eventually contain (for inference) the same thing
- Weights, structure and types
- Tools to convert models are born
- Tools to serve models are born

ABBYY
aizon
Alibaba Group
AMD
arm
aws
Bai百度
BECKHOFF
BITMAIN
cādence
CEVA
CLEARBLADE
Facebook Open Source
FURIOSA
GRAPHCORE
groq
habana
HAILO
Hewlett Packard Enterprise
HUAWEI
IBM
Idein Inc
intel
MathWorks
MAXAR
MEDIATEK
mi
Microsoft
mosaicML
NVIDIA
NXP
OctoML
ORACLE
OPEN AI LAB
Preferred Networks
SIEMENS
SONY
Qualcomm
sas
商汤 sensetime
skymizer
synopsys
Tencent
unity
verizon media
vmware
WOLFRAM
Yandex
ZAMA
ZETANE

Lingua franca

- One conversion format
- OpSet (versions)
- Visualizers and Analyzers



Open Neural Network Exchange

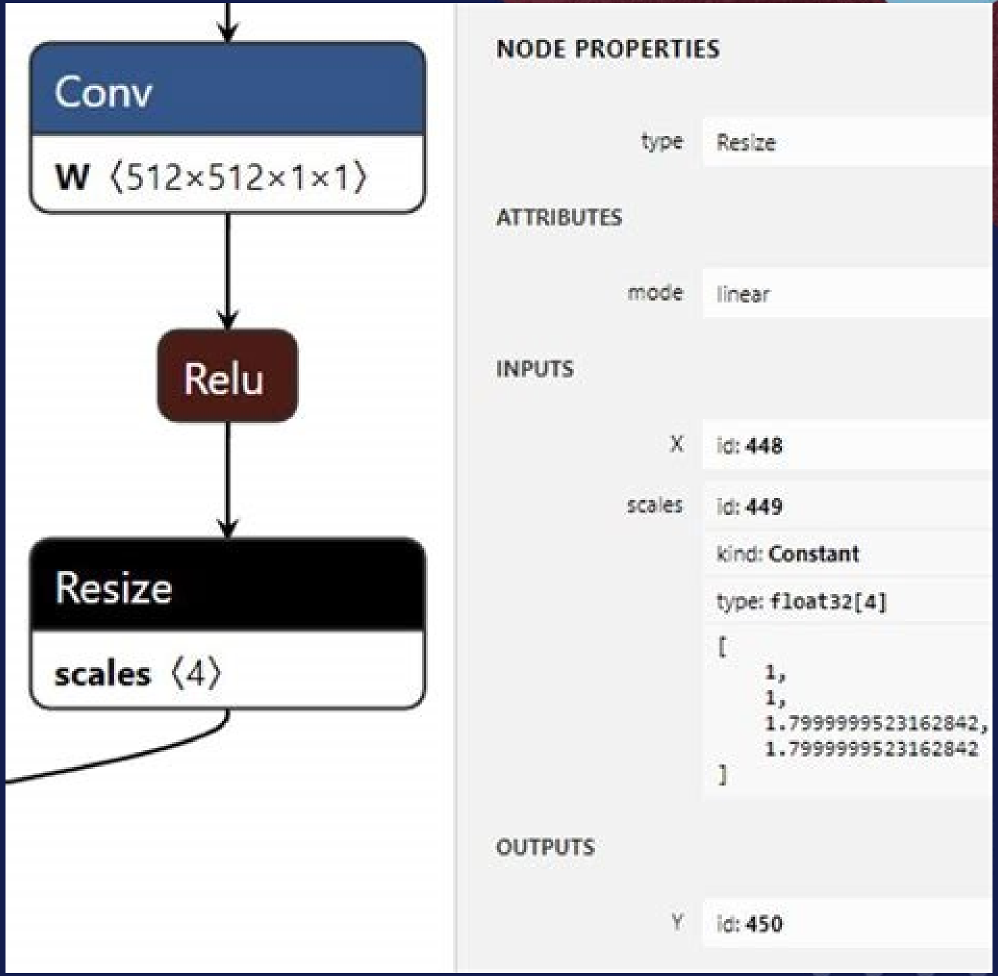
The open standard for machine learning interoperability

Load

- Change framework

- Use the preferred framework

- Do targeted testing



NODE PROPERTIES

type: Resize

ATTRIBUTES

mode: linear

INPUTS

X: id: 448

scales: id: 449

kind: Constant

type: float32[4]

```
[
    1,
    1,
    1.7999999523162842,
    1.7999999523162842
]
```

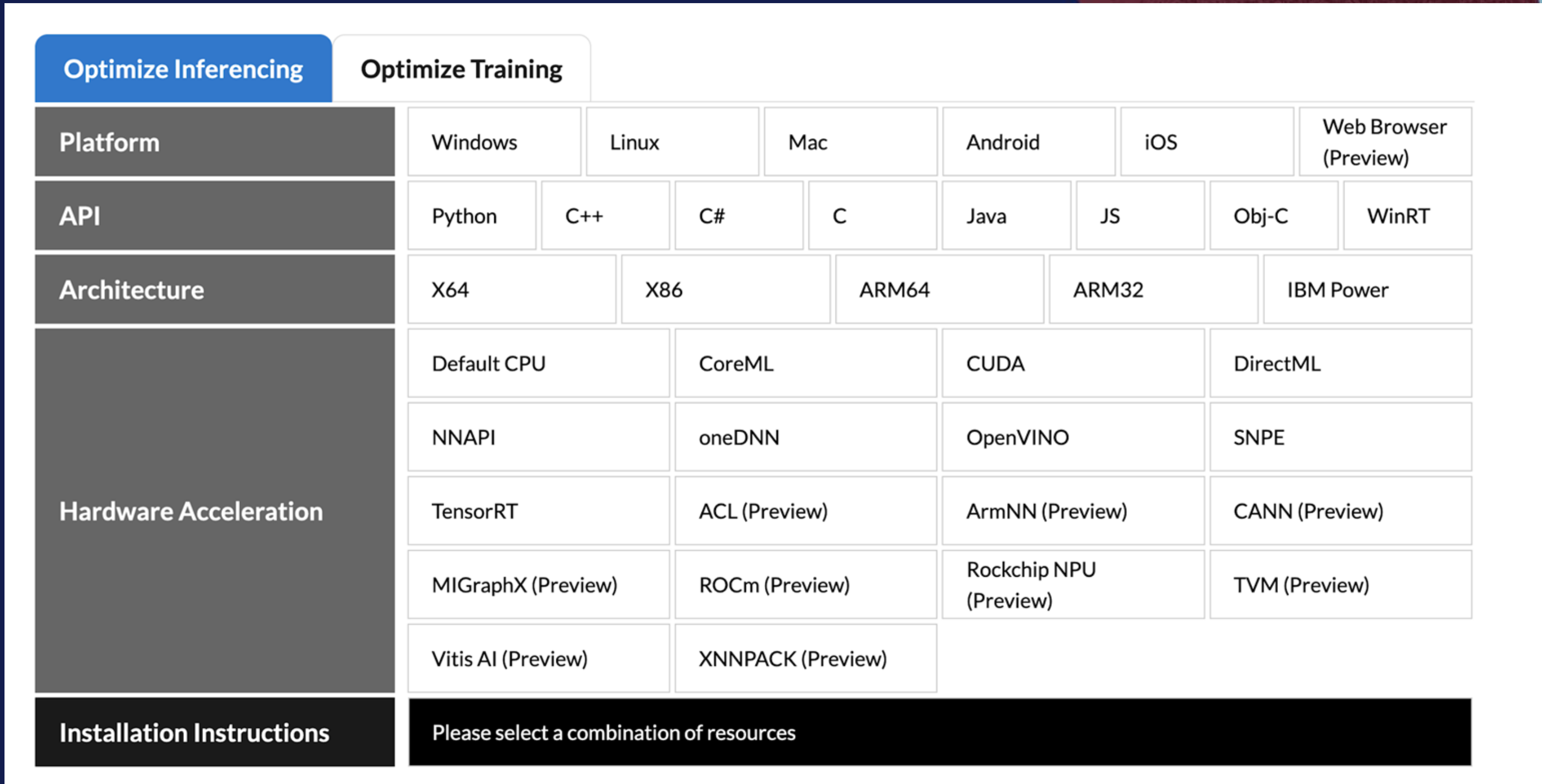
OUTPUTS

Y: id: 450

ONNX and runtimes

Run

- ONNX Runtime
- Inference with the exchange model
- Optimization



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	CoreML	CUDA	DirectML				
	NNAPI	oneDNN	OpenVINO	SNPE				
	TensorRT	ACL (Preview)	ArmNN (Preview)	CANN (Preview)				
	MIGraphX (Preview)	ROCm (Preview)	Rockchip NPU (Preview)	TVM (Preview)				
	Vitis AI (Preview)	XNNPACK (Preview)						
Installation Instructions	Please select a combination of resources							

All-in

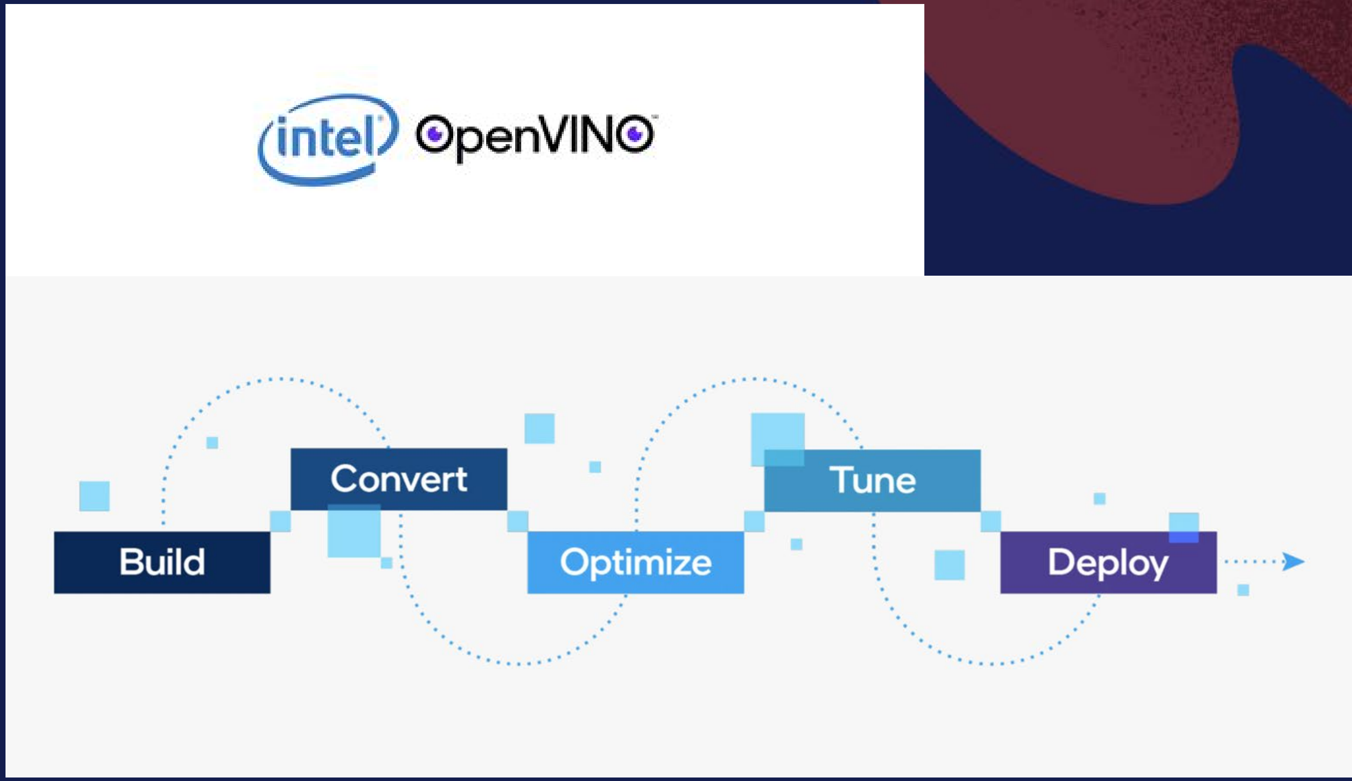
- Library and runtime for language and hardware

- Optimization libraries on the ONNX model

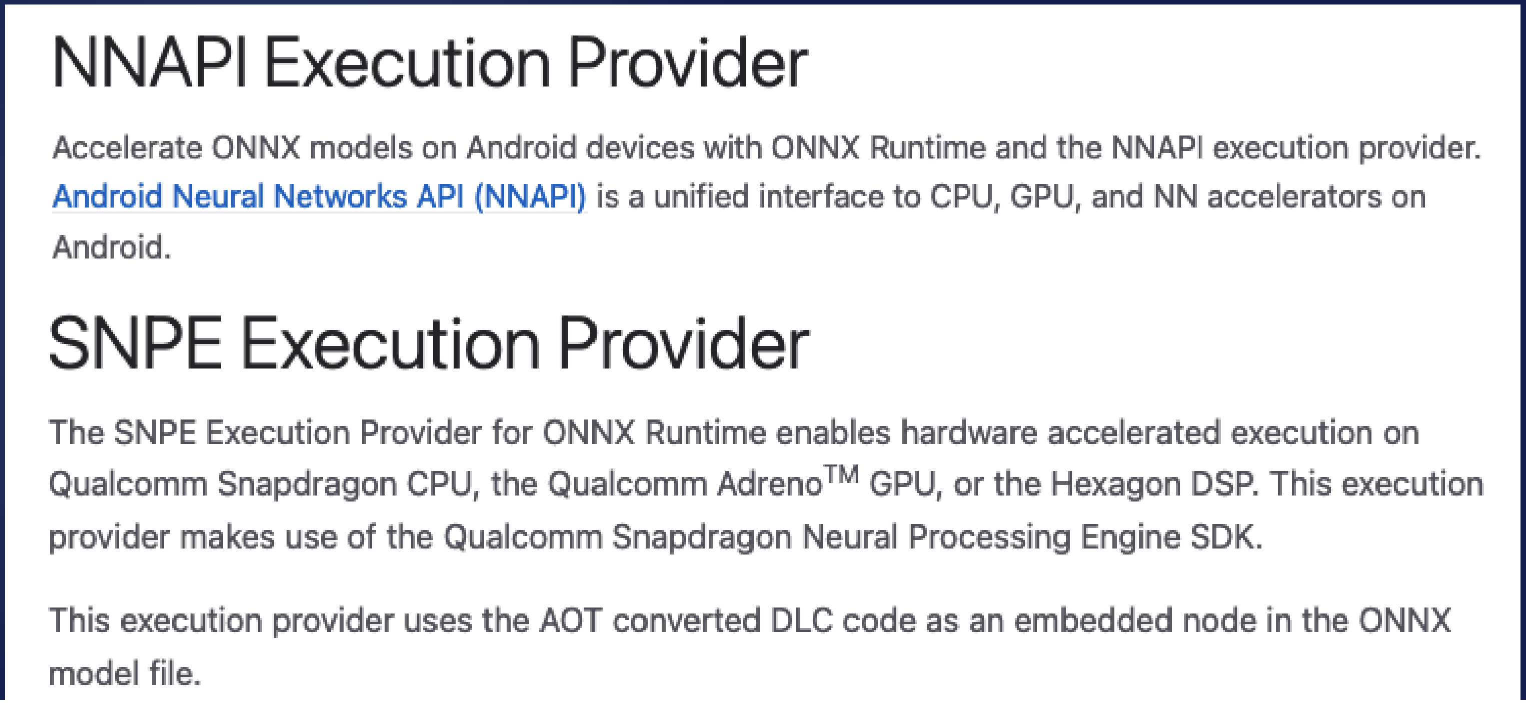
- Training directly in ONNX (ORModule for PyTorch)

Intel OpenVINO

- Libraries and runtimes for Intel hardware
- Multi Device Clusters
- Training directly in ONNX (TF and Pytorch in two lines)



DEMO



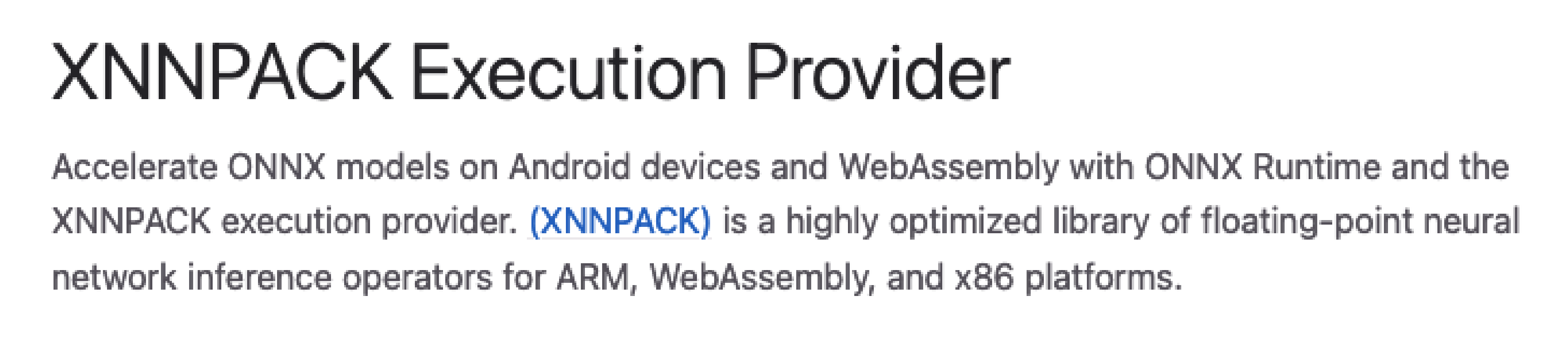
NNAPI Execution Provider

Accelerate ONNX models on Android devices with ONNX Runtime and the NNAPI execution provider. Android Neural Networks API (NNAPI) is a unified interface to CPU, GPU, and NN accelerators on Android.

SNPE Execution Provider

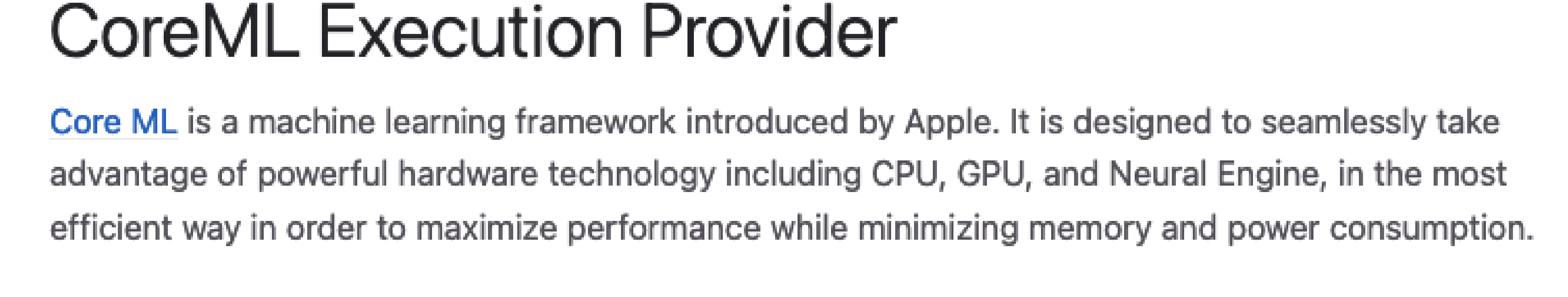
The SNPE Execution Provider for ONNX Runtime enables hardware accelerated execution on Qualcomm Snapdragon CPU, the Qualcomm Adreno$^{\text{TM}}$ GPU, or the Hexagon DSP. This execution provider makes use of the Qualcomm Snapdragon Neural Processing Engine SDK.

This execution provider uses the AOT converted DLC code as an embedded node in the ONNX model file.



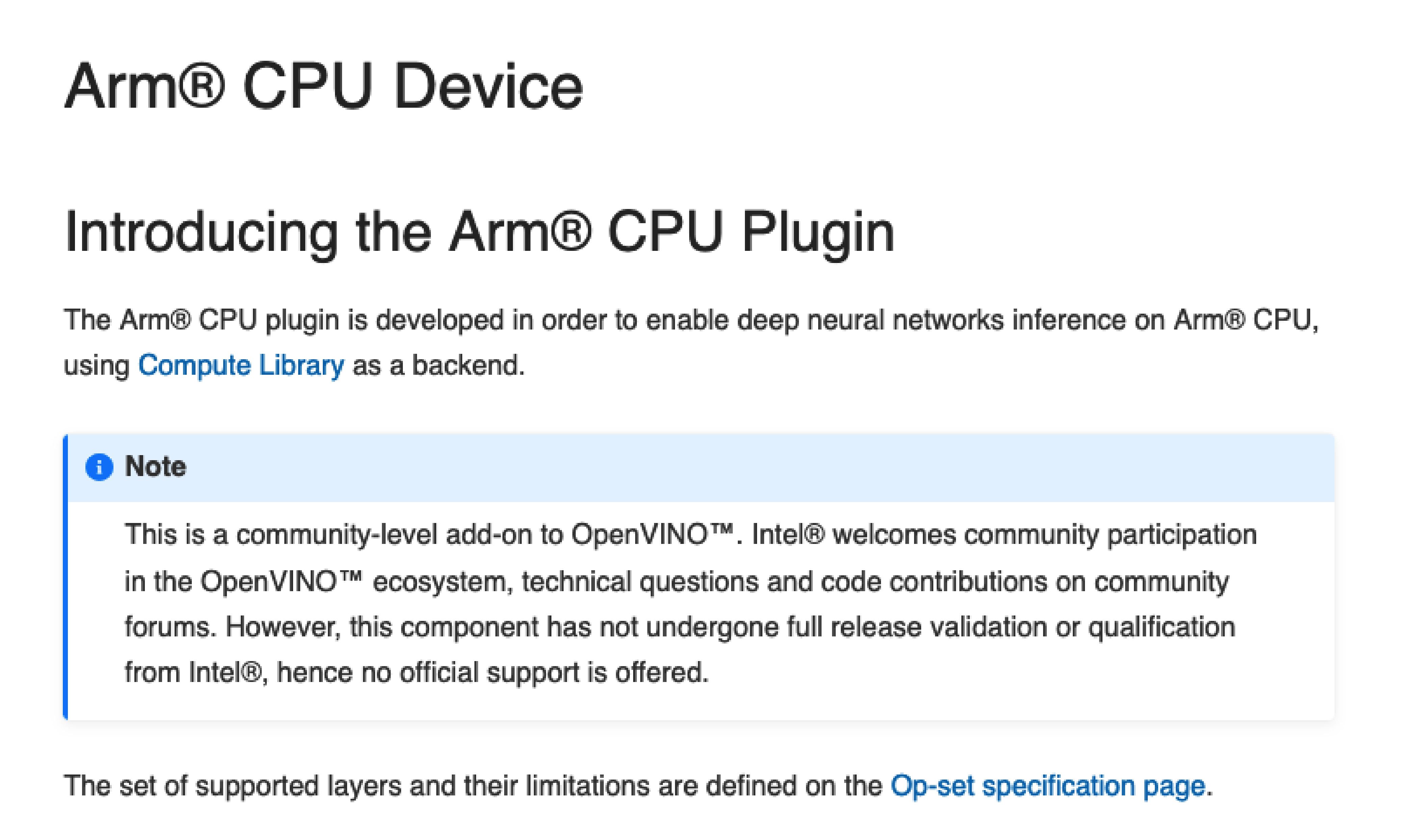
XNNPACK Execution Provider

Accelerate ONNX models on Android devices and WebAssembly with ONNX Runtime and the XNNPACK execution provider. (XNNPACK) is a highly optimized library of floating-point neural network inference operators for ARM, WebAssembly, and x86 platforms.



CoreML Execution Provider

Core ML is a machine learning framework introduced by Apple. It is designed to seamlessly take advantage of powerful hardware technology including CPU, GPU, and Neural Engine, in the most efficient way in order to maximize performance while minimizing memory and power consumption.



Arm® CPU Device

Introducing the Arm® CPU Plugin

The Arm® CPU plugin is developed in order to enable deep neural networks inference on Arm® CPU, using Compute Library as a backend.

> **Note**
>
> This is a community-level add-on to OpenVINO™. Intel® welcomes community participation in the OpenVINO™ ecosystem, technical questions and code contributions on community forums. However, this component has not undergone full release validation or qualification from Intel®, hence no official support is offered.

The set of supported layers and their limitations are defined on the Op-set specification page.

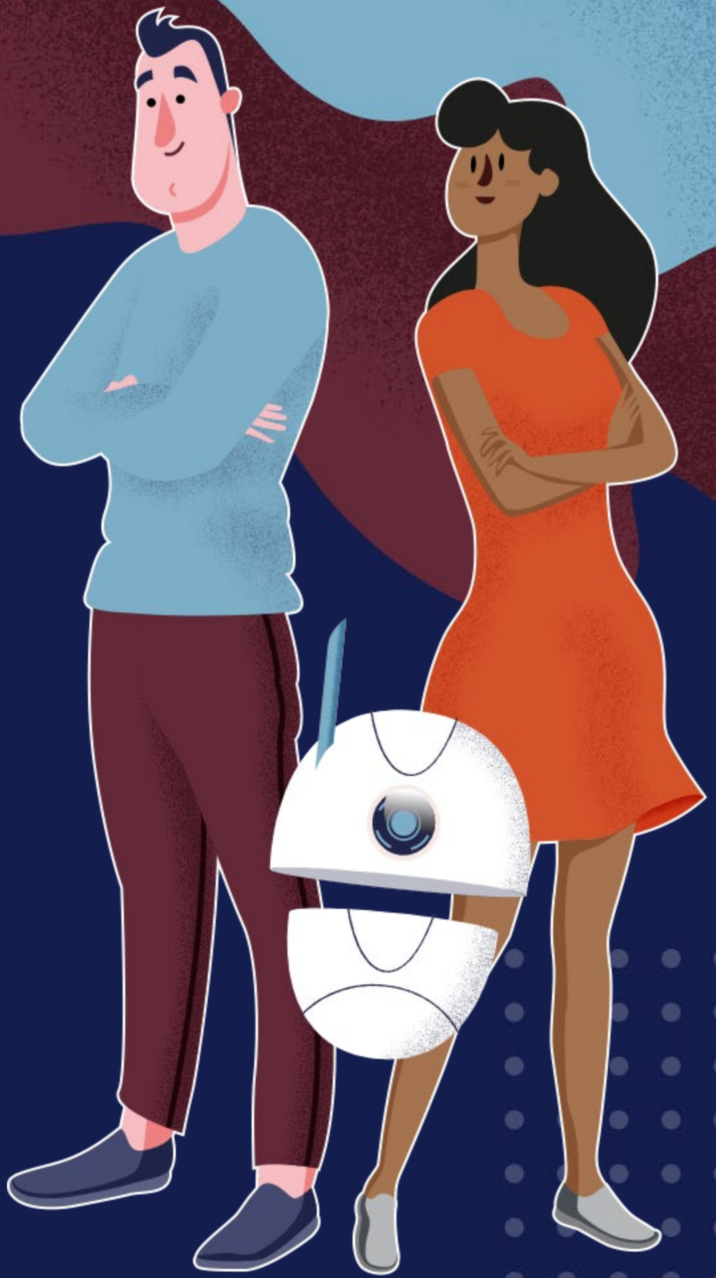
global AI bootcamp

March 4th 2023, Torino - Italy

Thank you! Q & A Time ☺

Mauro Bennici

AI Architect and AI Ethicist



global AI community

Deltatre Innovation Lab

#GlobalAIBootcamp