



Azure AI & ONNX

Mauro Bennici

Cedat85

 @maurobennici

SPONSOR

DIAMOND



GUCCI

splunk>



PLATINUM

improve



GOLD



CODICEPLASTICO

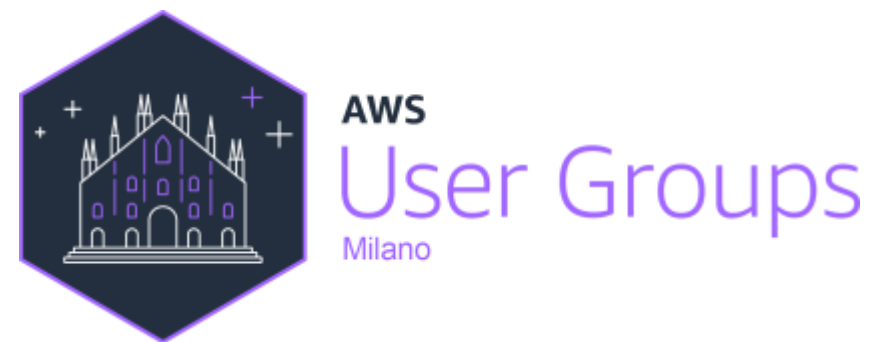


Cubbit

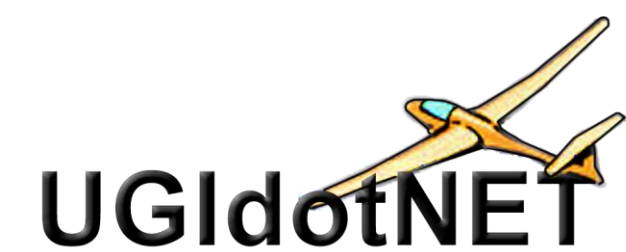
 CLOUD DAY 2022

#CLOUDDAY2022

PARTNER



CLOUD DAY 2022

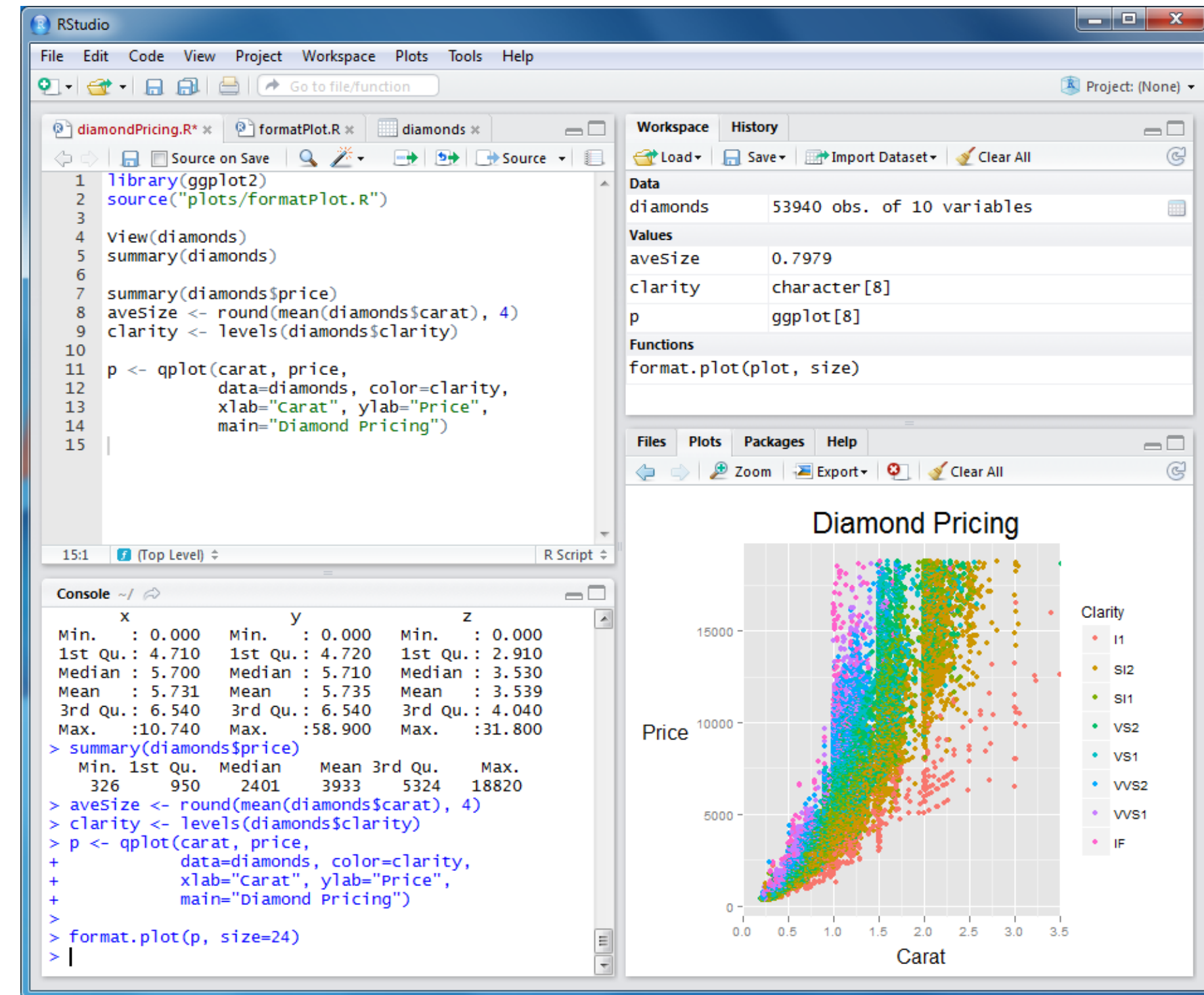


#CLOUDDAY2022

The cloud dilemma

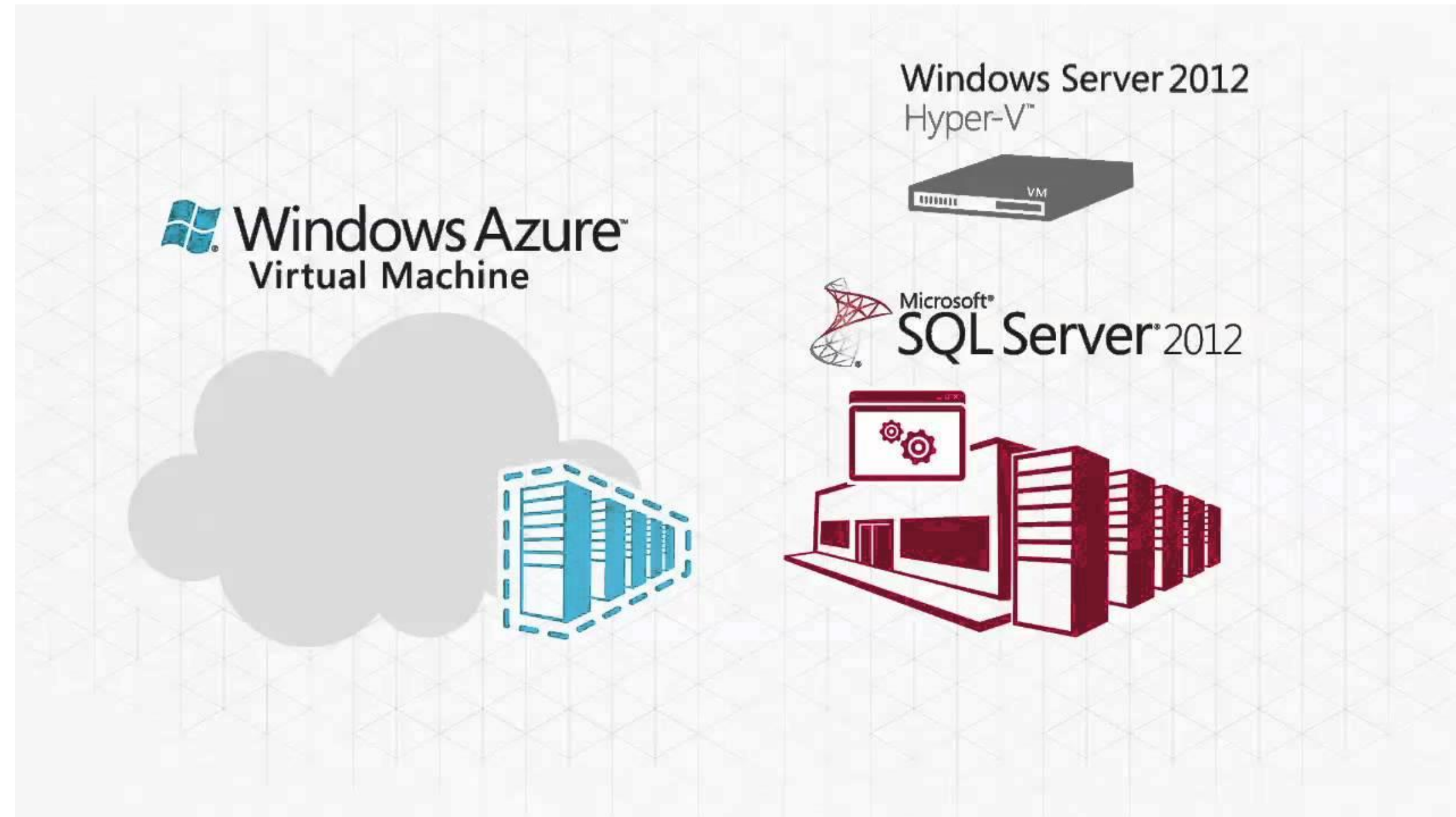
L'arrivo dei servizi ML

- La Data Science era una scienza di laboratorio
- Dati ed elaborazioni avvenivano in data center private
- Il software custom era tantissimo



L'arrivo dei servizi ML

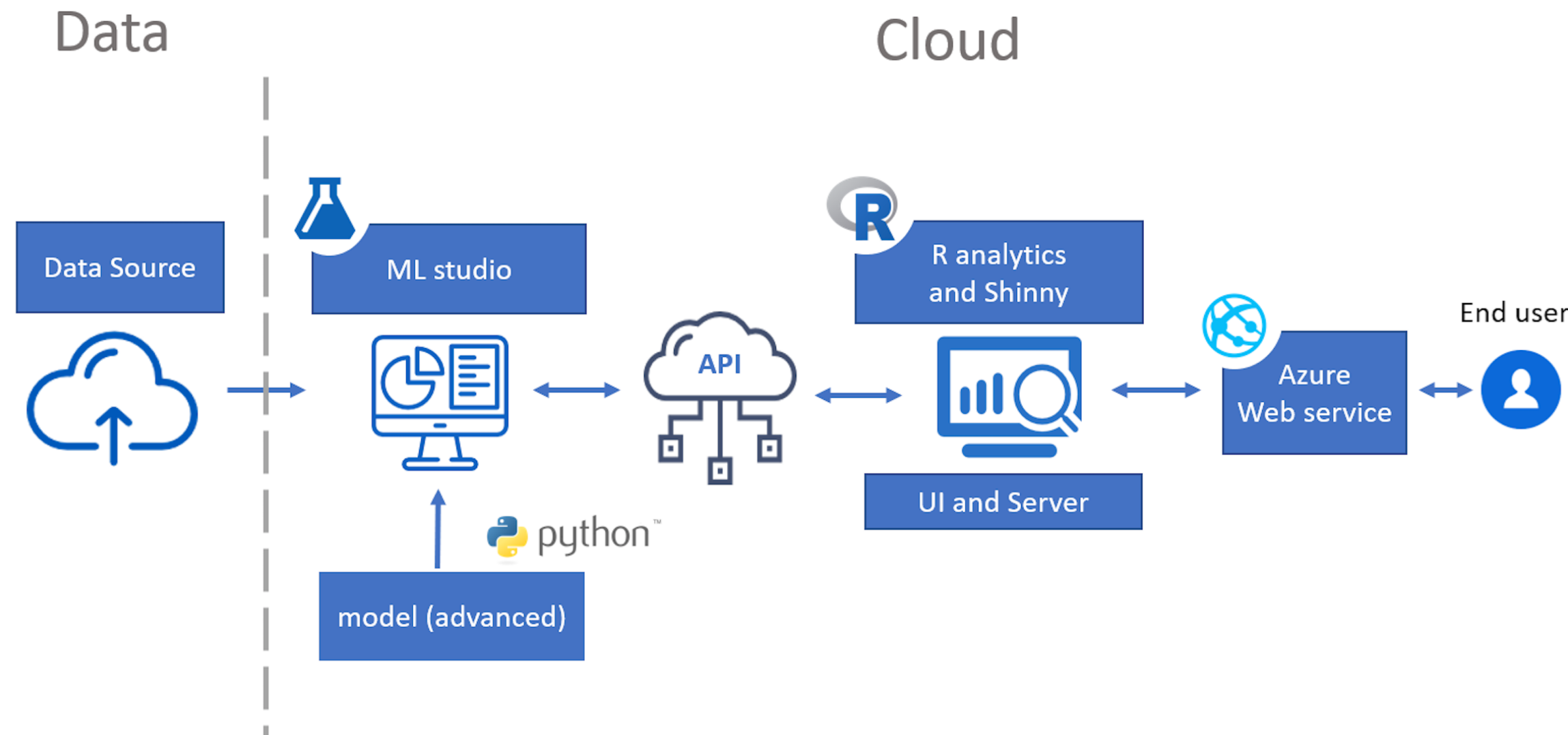
- Arriva il cloud
- Trasposizione on-line di applicativi on-prem
- Problema spostato



L'arrivo dei servizi ML

- Servizi specializzati come AzureML
- Tutto in un unico posto (data lake, sql, nosql, file, modelli, API)
- Vendor lock-in?
- Hardware lock-in?

Azure Machine Learning & R



E i modelli di AI?

- Framework aggiornati spesso con breaking changes
- Implementazioni custom per diversi linguaggi
- Incompatibilità tra i modelli (framework)
- Costi sempre più alti



Un ONNX per ghermirli...

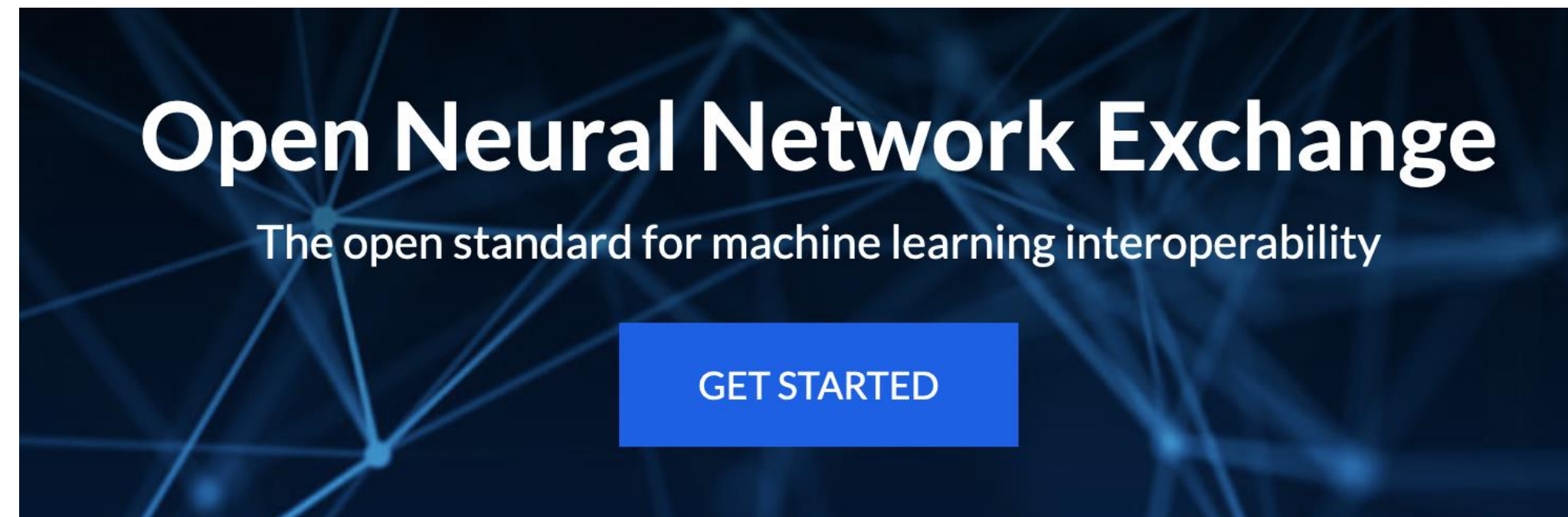
Convertire

- I modelli alla fine contengono (per l'inferenza) la stessa cosa
- Pesi, struttura e tipi
- Nascono i tool per convertire i modelli
- Nascono i tool per servire i modelli



Lingua franca

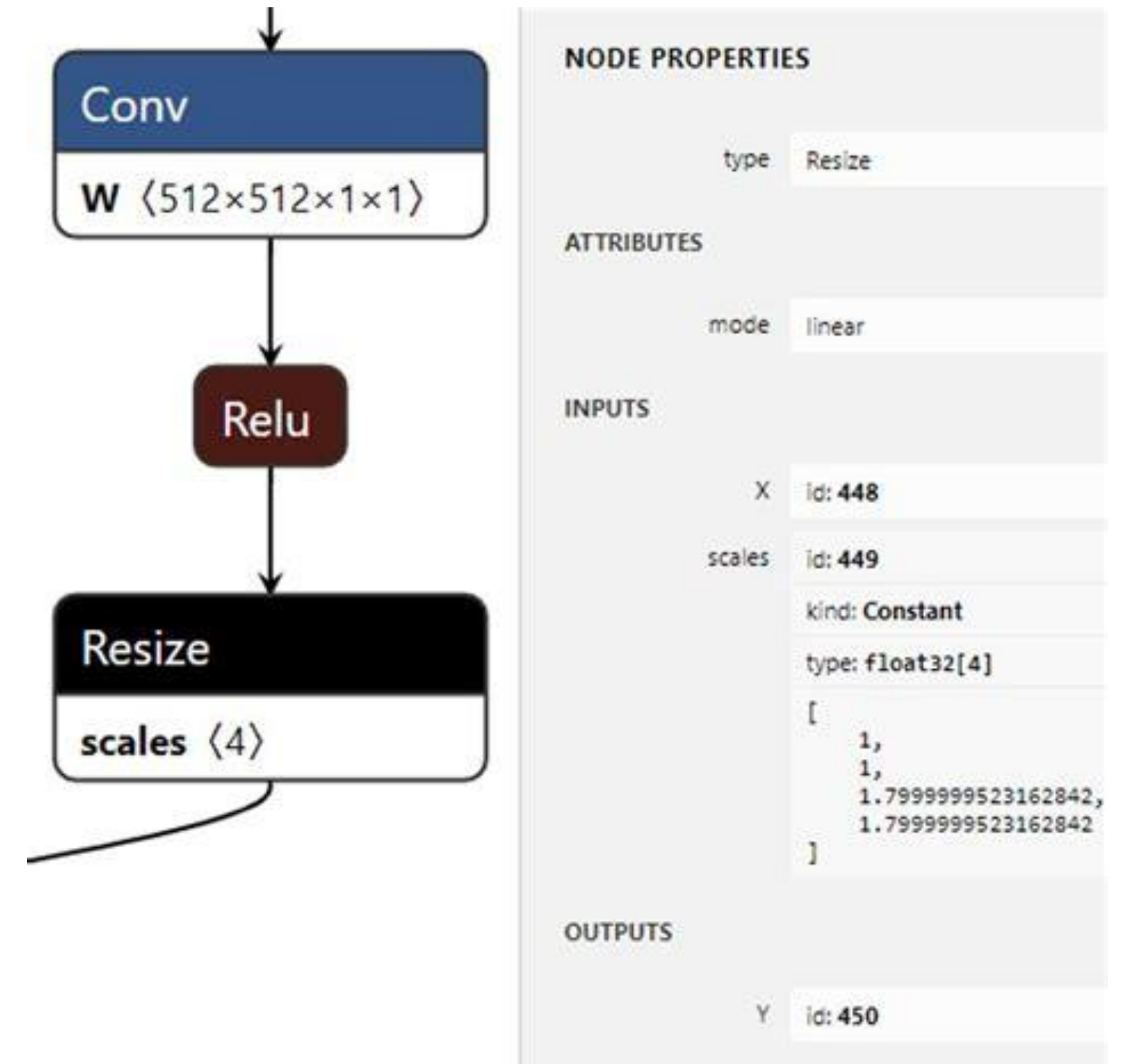
- Un solo formato di conversione
- OpSet
- Visualizzatori e Analizzatori



ONNX e i runtime

Load

- Cambiare framework
- Usare il framework preferito
- Fare test mirati



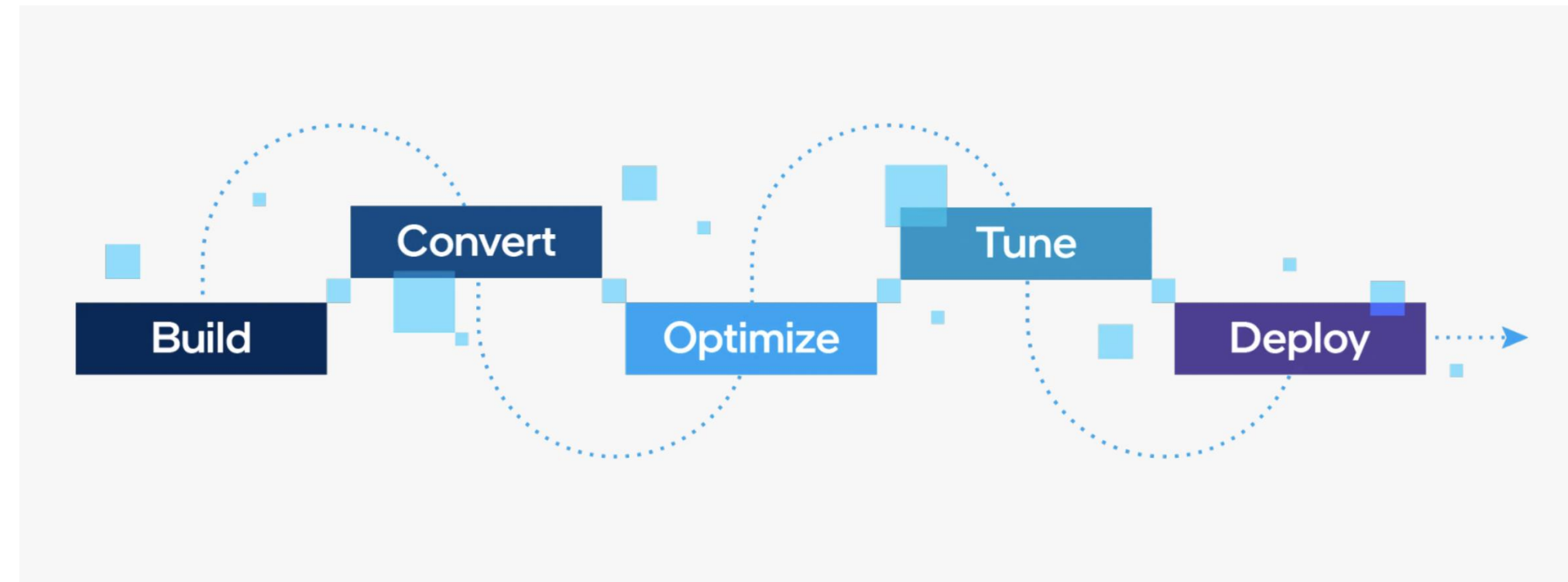
Run?

- ONNX Runtime
- Inferenza con il modello di scambio
- Ottimizzazione

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

- Libreria e runtime per linguaggio e hardware
- Librerie di ottimizzazione sul modello ONNX (Intel OpenVino)
- (beta per Pytorch) Training direttamente in ONNX



DEMO



Grazie!