Strata Data & AI Superstream Series

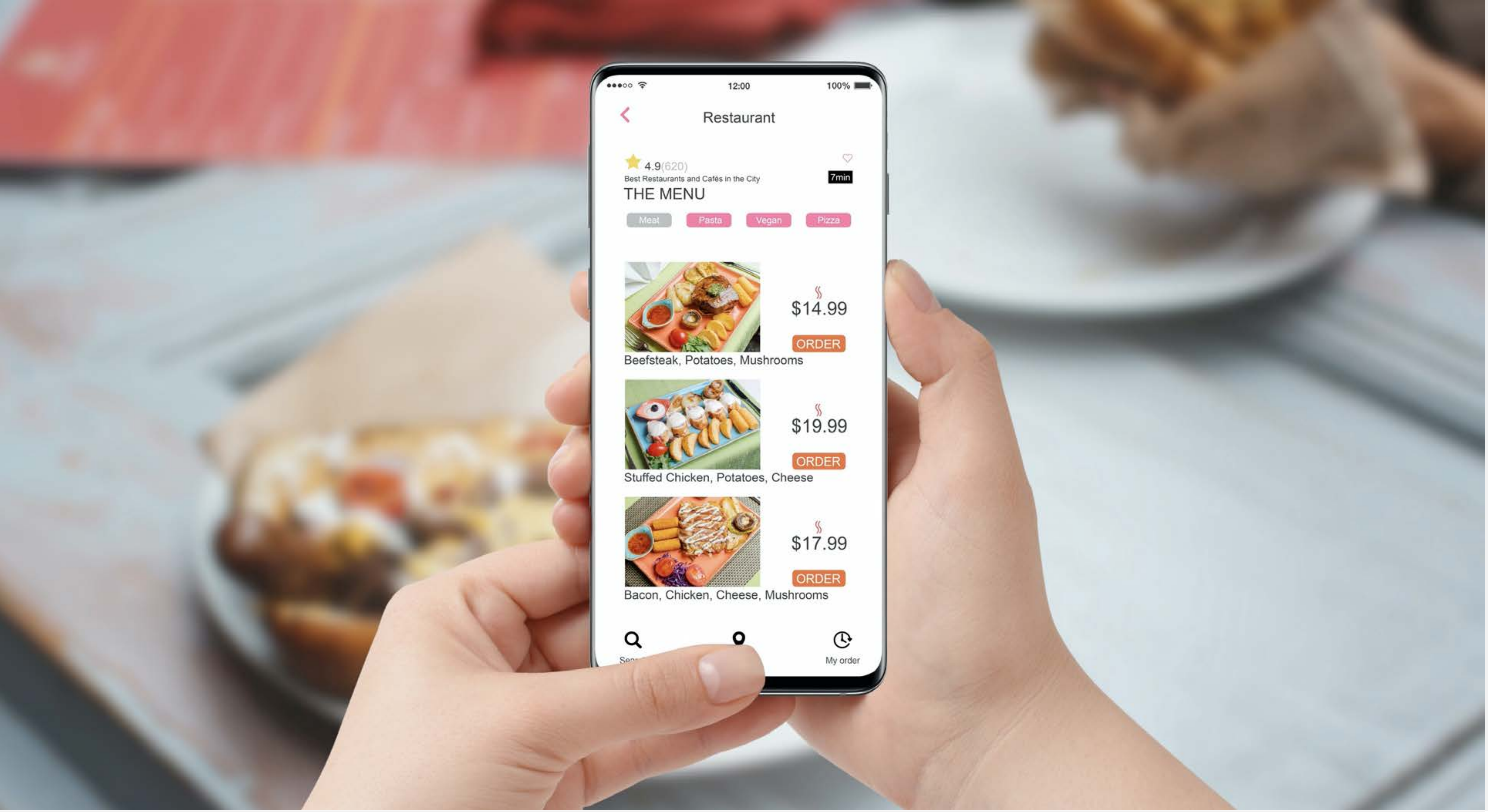
AI Inferencing with NLP at Scale with OpenVINO

Zoe Cayetano and Raymond Lo

Let’s Make “Pasta” with NLP

Voice-based Interactive Recipes App for Cooking

Where we do get our data?



12:00 100%

Restaurant

4.9(620)

Best Restaurants and Cafés in the City

7min

THE MENU

Meat Pasta Vegan Pizza

$14.99

ORDER

Beefsteak, Potatoes, Mushrooms

$19.99

ORDER

Stuffed Chicken, Potatoes, Cheese

$17.99

ORDER

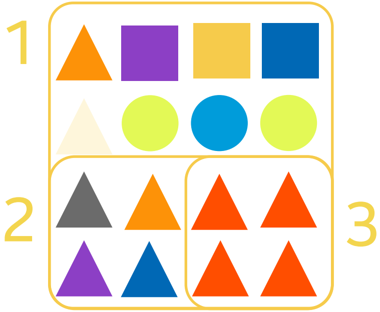
Bacon, Chicken, Cheese, Mushrooms

My order

AI Compute Considerations

How do you determine the right computing for your AI needs?

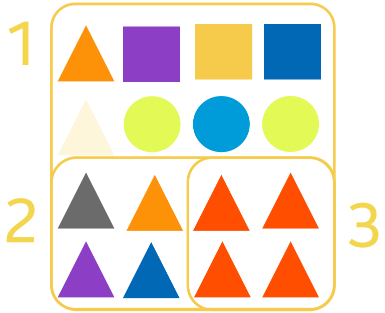
WORKLOADS



AI Compute Considerations

How do you determine the right computing for your AI needs?

WORKLOADS



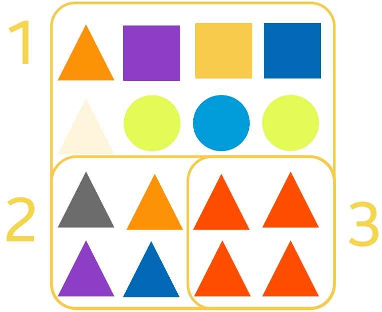
REQUIREMENTS

RESPONSE TIME
SECURITY
DATA
VERSATILITY
SCALABILITY
COST
POWER
SIZE
RELIABILITY
THROUGHPUT
EFFICIENCY
MANAGEABILITY
FLEXIBILITY
ACCURACY

AI Compute Considerations

How do you determine the right computing for your AI needs?

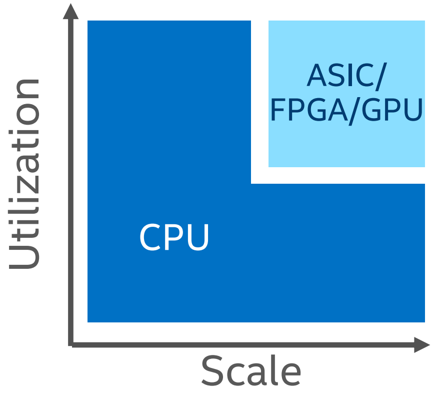
WORKLOADS



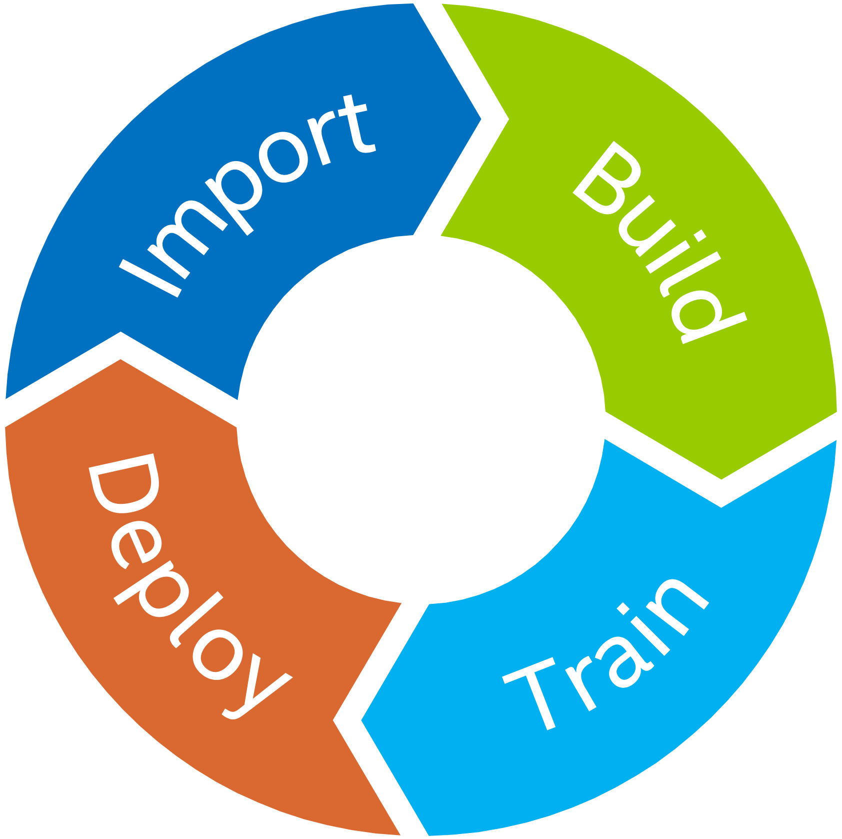
REQUIREMENTS

RESPONSE TIME
SECURITY
DATA
VERSATILITY
SCALABILITY
COST
POWER
SIZE
RELIABILITY
THROUGHPUT
EFFICIENCY
MANAGEABILITY
FLEXIBILITY
ACCURACY

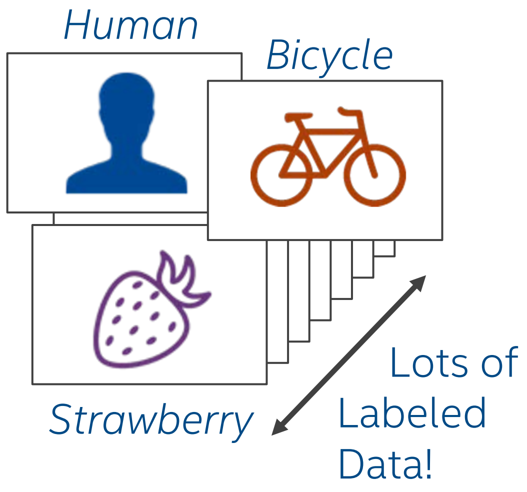
DEMAND



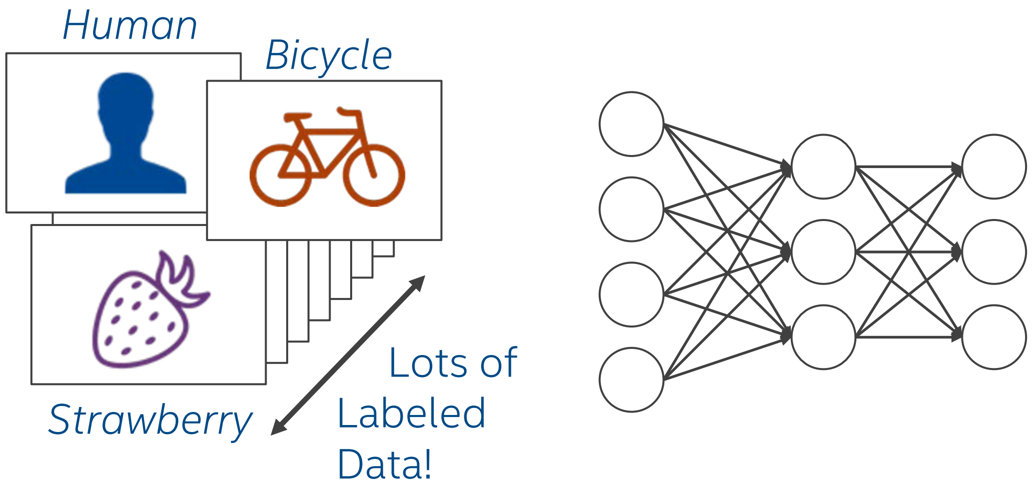
Deep Learning Development Cycle



Training vs Inference



Training vs Inference



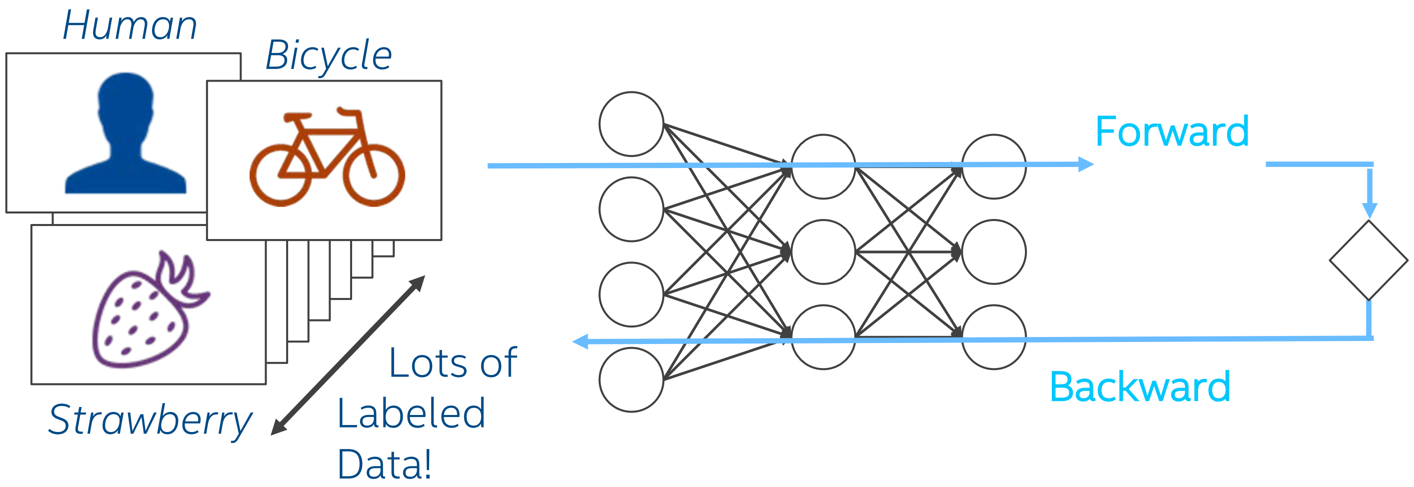
Human

Bicycle

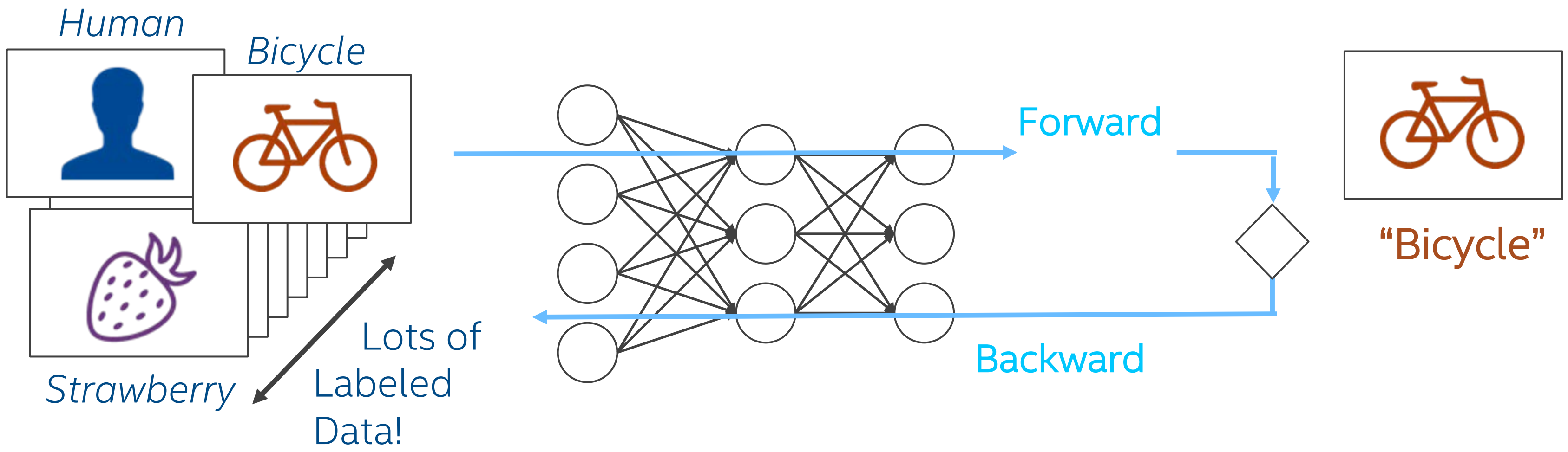
Strawberry

Lots of Labeled Data!

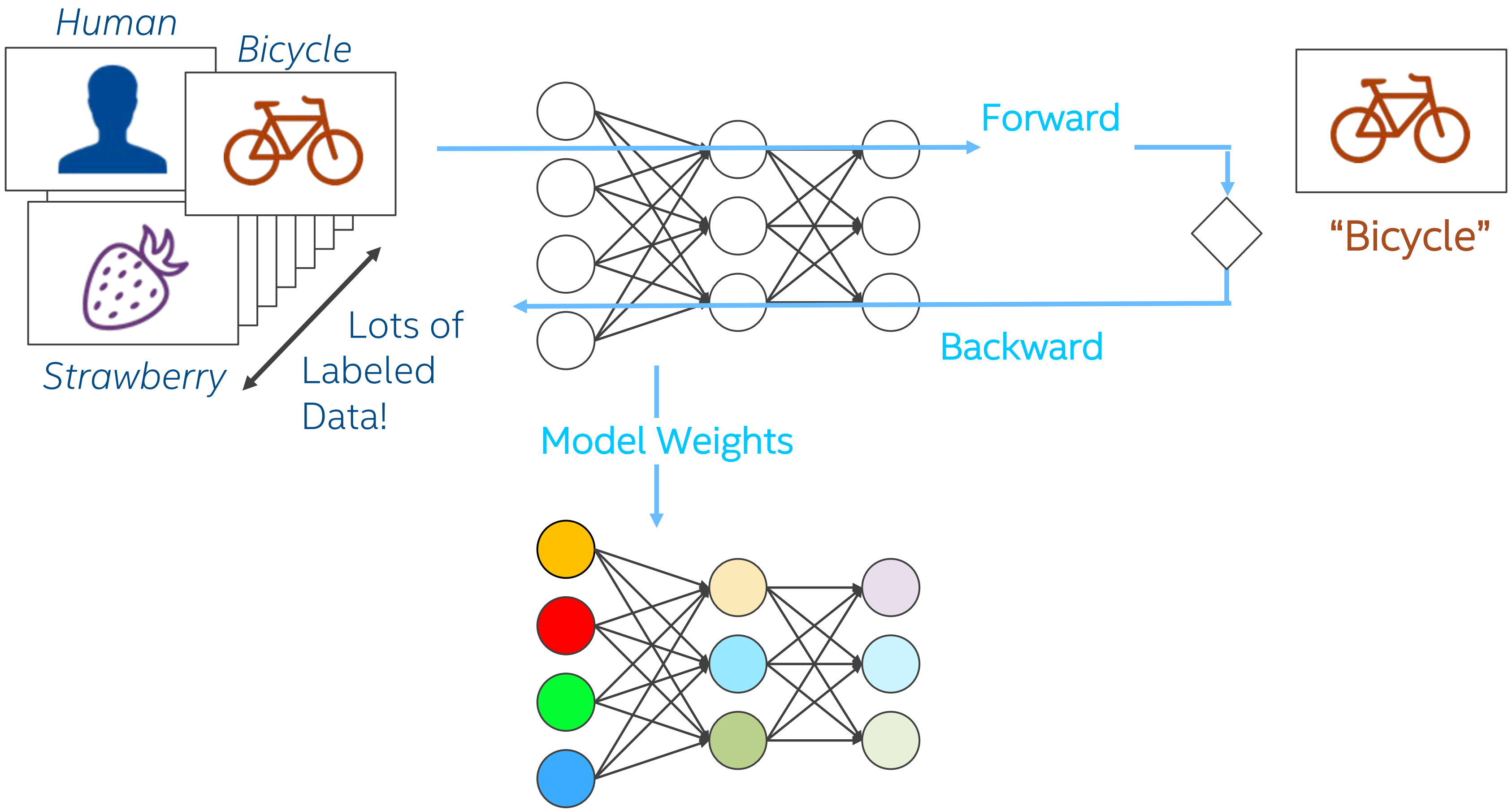
Training vs Inference



Training vs Inference

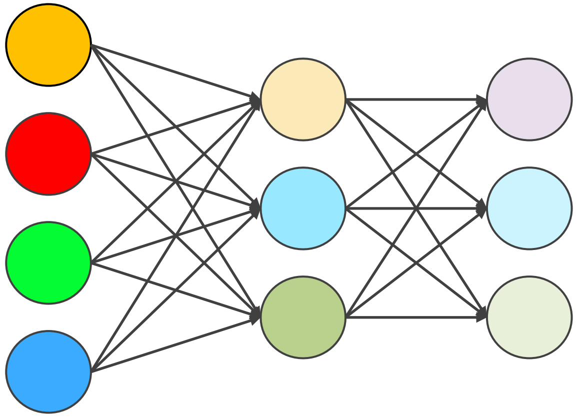


Training vs Inference

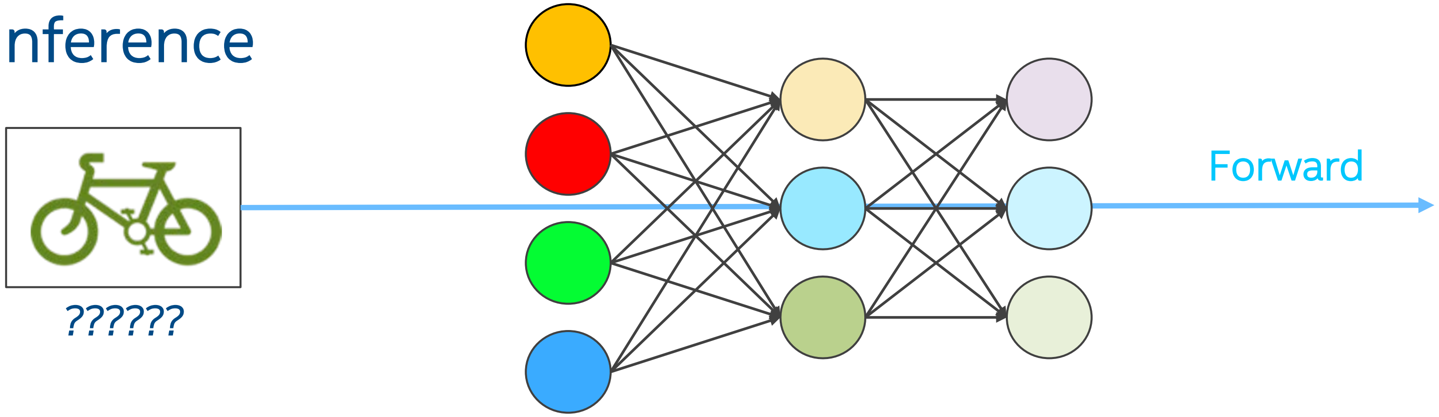


Inference

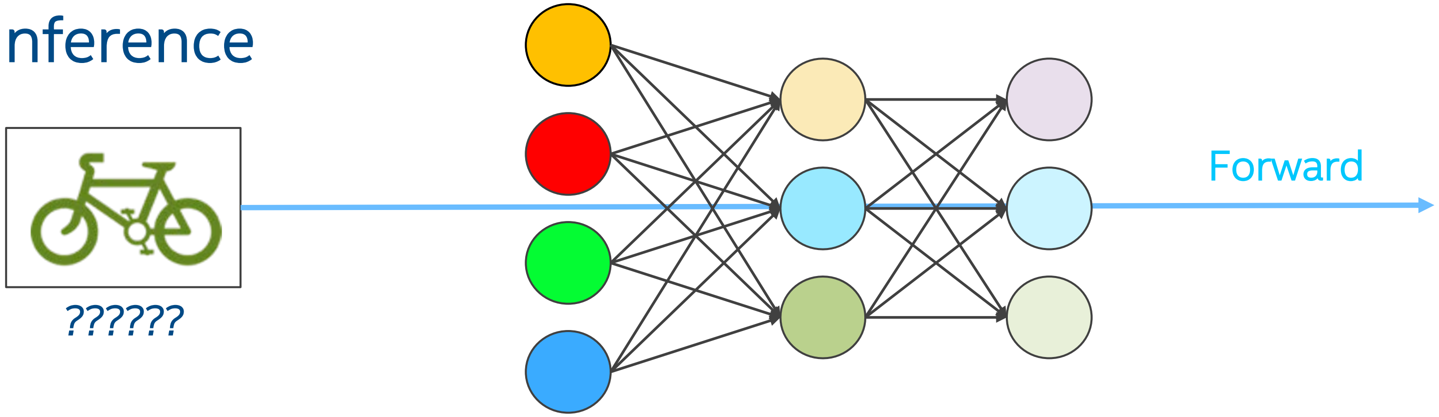
??????



Inference



Inference



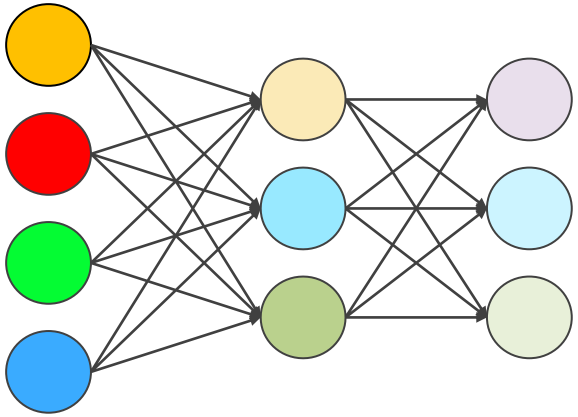
??????

Forward

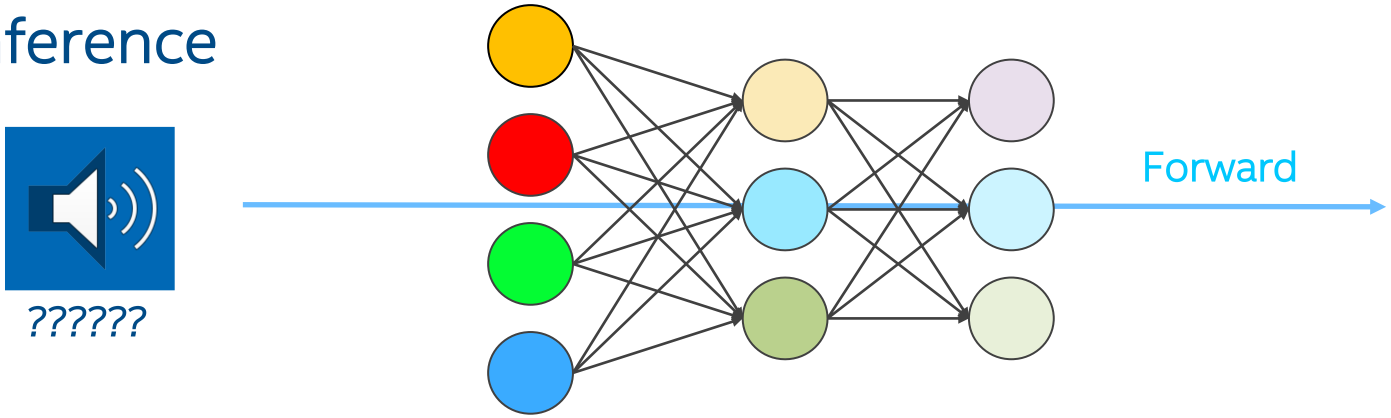
90% = “Bicycle” ?

Inference

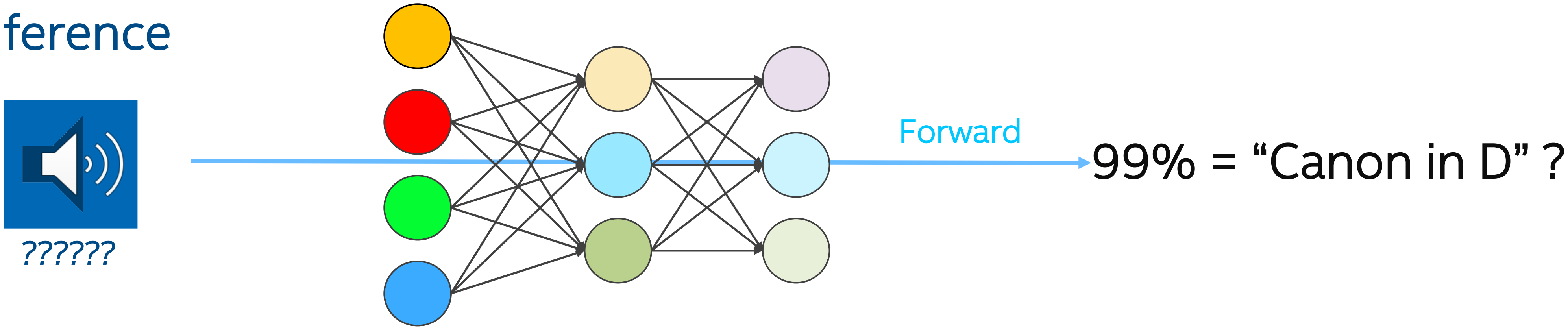
??????



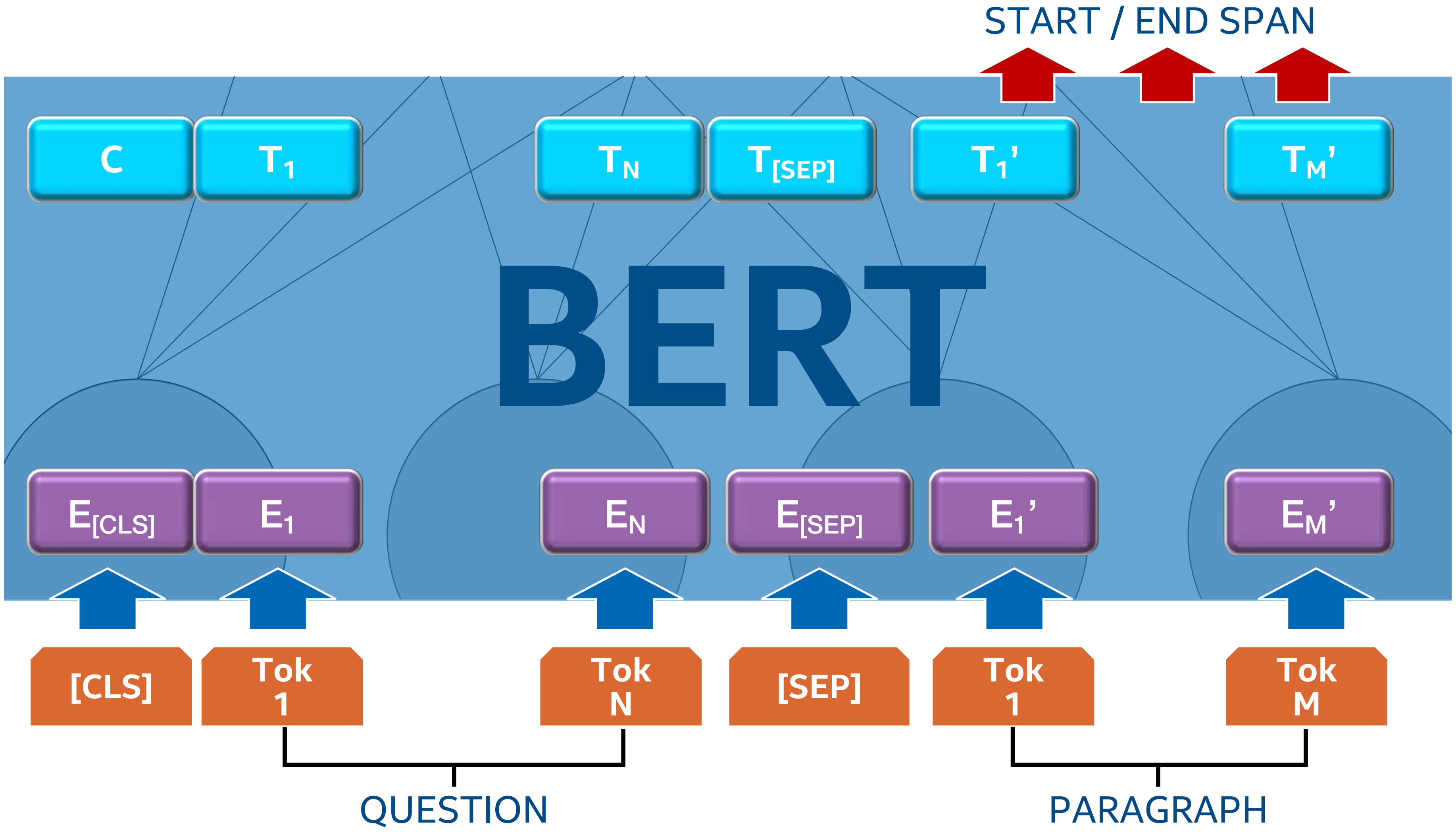
Inference

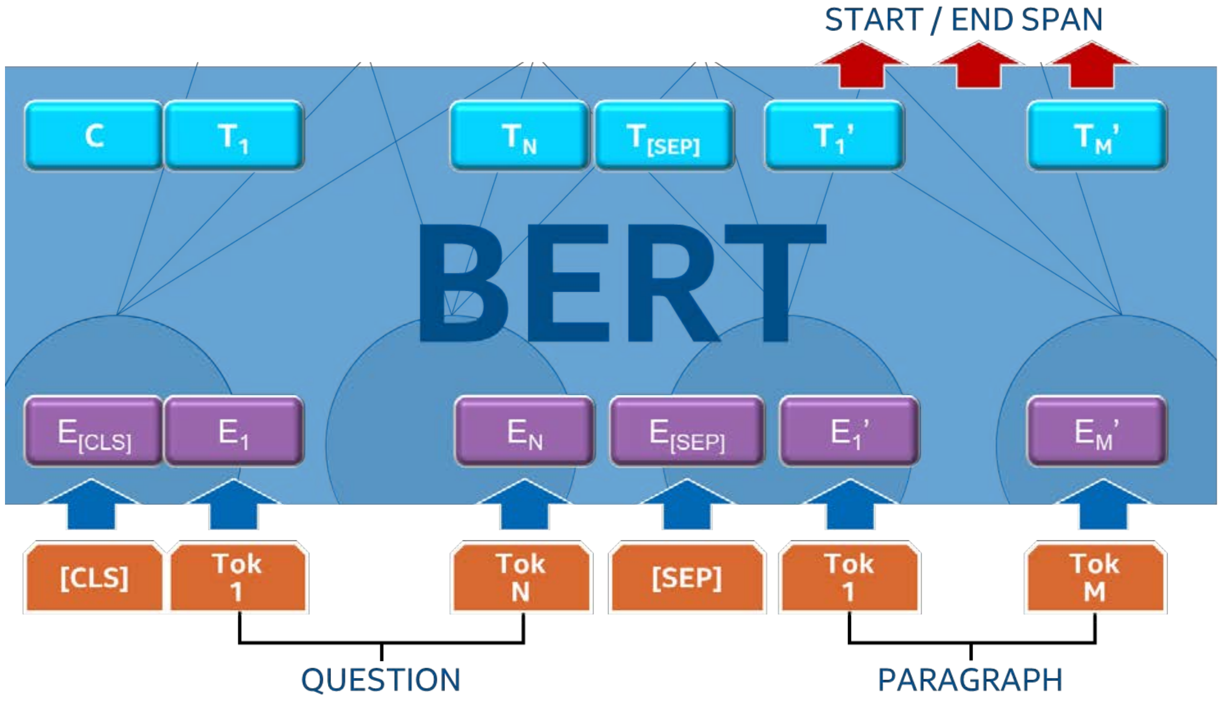


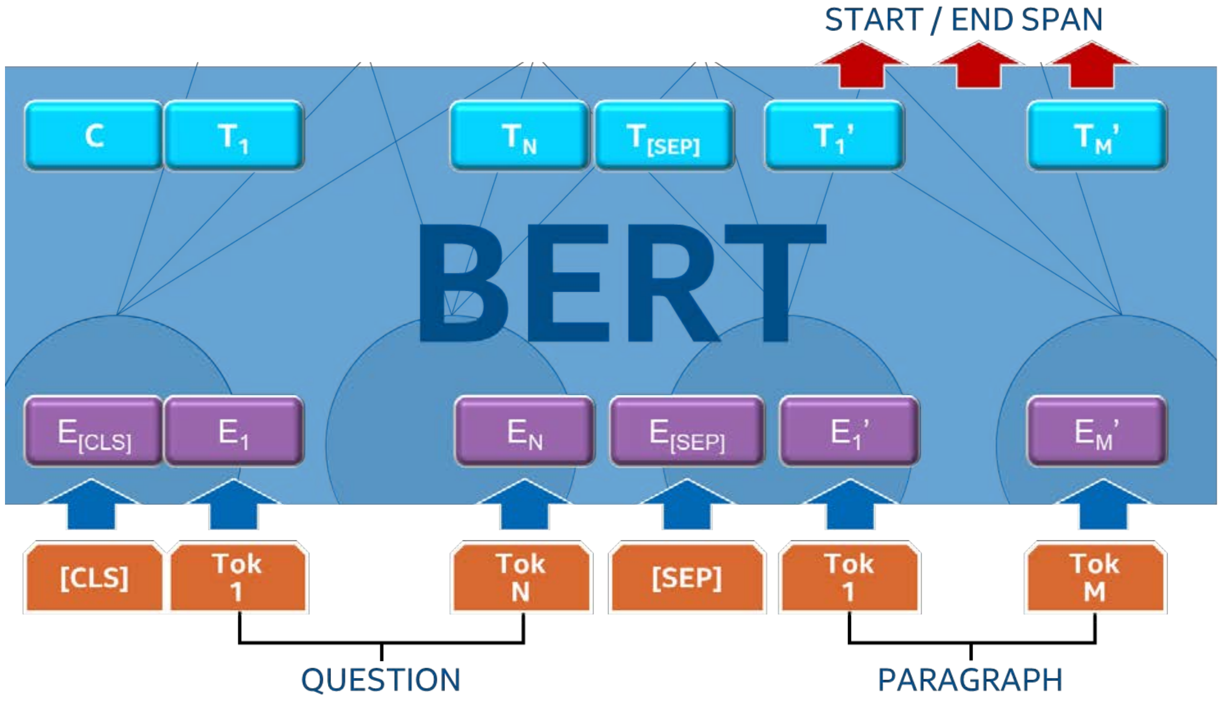
Inference



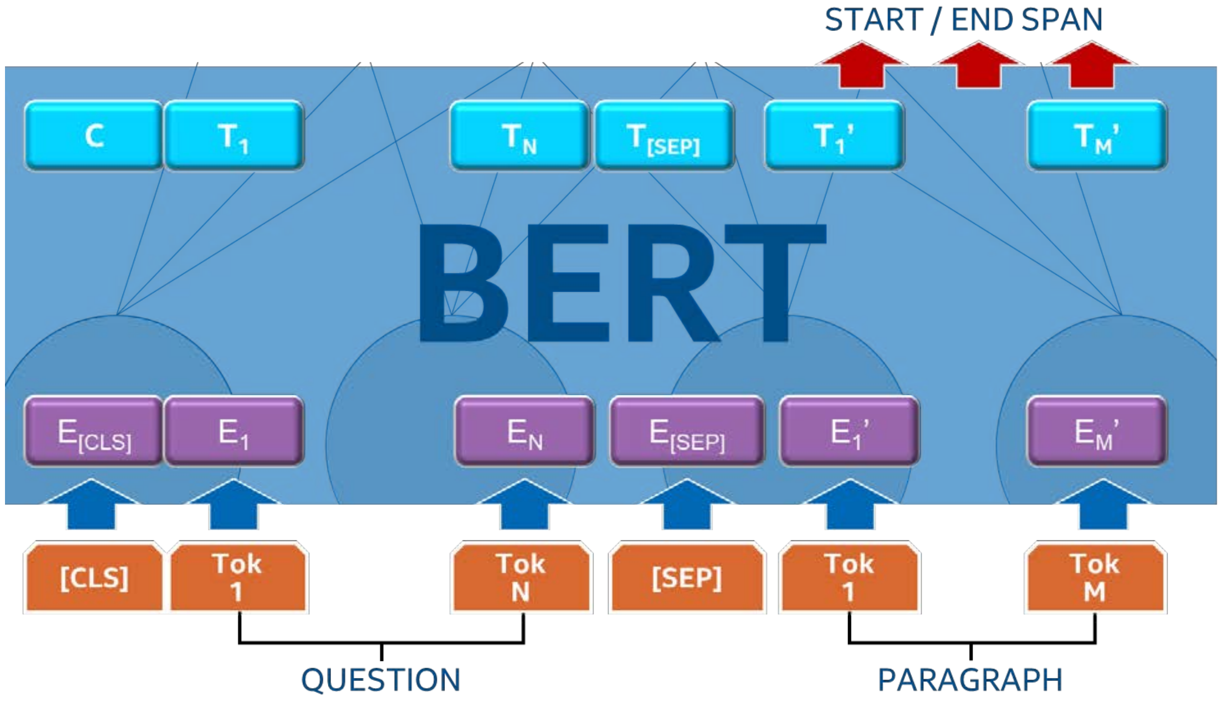
Bert Fine-Tuning (SQuAD)




START / END SPAN
C
T1
TN
T[SEP]
T1'
TM'
BERT
E[CLS]
E1
EN
E[SEP]
E1'
EM'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
Tok M
QUESTION
PARAGRAPH

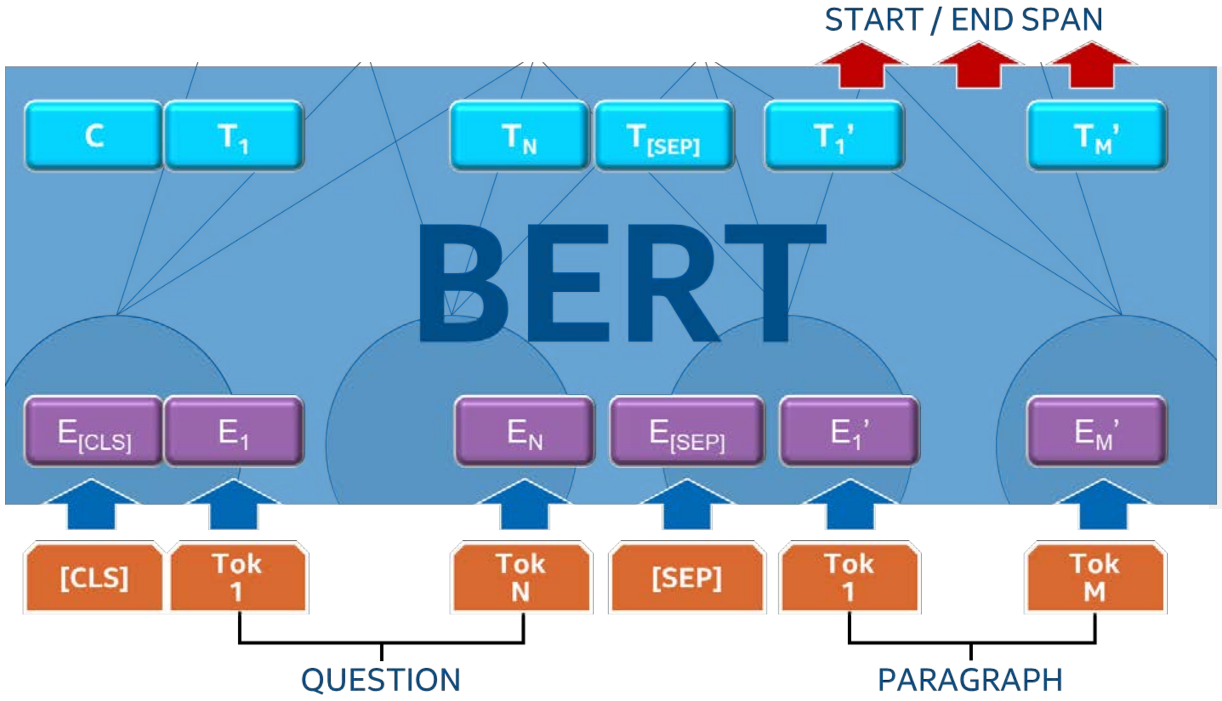


Intel was founded in Mountain View, California, in 1968 by Gordon E. Moore (known for "Moore's law"), a chemist, and Robert Noyce, a physicist and co-inventor of the integrated circuit. Arthur Rock (investor and venture capitalist) helped them find investors, while Max Palevsky was on the board from an early stage.[23] Moore and Noyce had left Fairchild Semiconductor to found Intel. Rock was not an employee, but he was an investor and was chairman of the board.[24][25] The total initial investment in Intel was $2.5 million in convertible debentures (equivalent to $18.4 million in 2019) and $10,000 from Rock. Just 2 years later, Intel became a public company via an initial public offering (IPO), raising $6.8 million ($23.50 per share).[24] Intel's third employee was Andy Grove,[26] a chemical engineer, who later ran the company through much of the 1980s and the high-growth 1990s. …



When was Intel founded?

Intel was founded in Mountain View, California, in 1968 by Gordon E. Moore (known for "Moore's law"), a chemist, and Robert Noyce, a physicist and co-inventor of the integrated circuit. Arthur Rock (investor and venture capitalist) helped them find investors, while Max Palevsky was on the board from an early stage.[23] Moore and Noyce had left Fairchild Semiconductor to found Intel. Rock was not an employee, but he was an investor and was chairman of the board.[24][25] The total initial investment in Intel was $2.5 million in convertible debentures (equivalent to $18.4 million in 2019) and $10,000 from Rock. Just 2 years later, Intel became a public company via an initial public offering (IPO), raising $6.8 million ($23.50 per share).[24] Intel's third employee was Andy Grove,[26] a chemical engineer, who later ran the company through much of the 1980s and the high-growth 1990s. ...



```
Intel was founded in
Mountain View,
California, in 1968 by
Gordon E. Moore (known
for "Moore's law"), …
```

When was Intel founded?

Intel was founded in Mountain View, California, in 1968 by Gordon E. Moore (known for "Moore's law"), a chemist, and Robert Noyce, a physicist and co-inventor of the integrated circuit. Arthur Rock (investor and venture capitalist) helped them find investors, while Max Palevsky was on the board from an early stage.[23] Moore and Noyce had left Fairchild Semiconductor to found Intel. Rock was not an employee, but he was an investor and was chairman of the board.[24][25] The total initial investment in Intel was $2.5 million in convertible debentures (equivalent to $18.4 million in 2019) and $10,000 from Rock. Just 2 years later, Intel became a public company via an initial public offering (IPO), raising $6.8 million ($23.50 per share).[24] Intel's third employee was Andy Grove,[26] a chemical engineer, who later ran the company through much of the 1980s and the high-growth 1990s. …

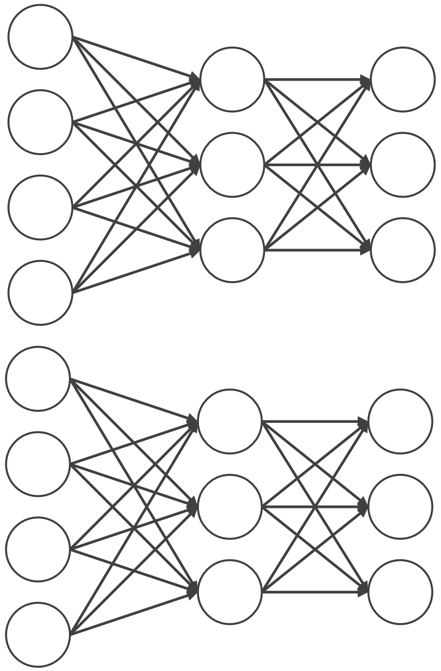
1. Build

Trained Model

TensorFlow Caffe

KALDI mxnet

ONNX

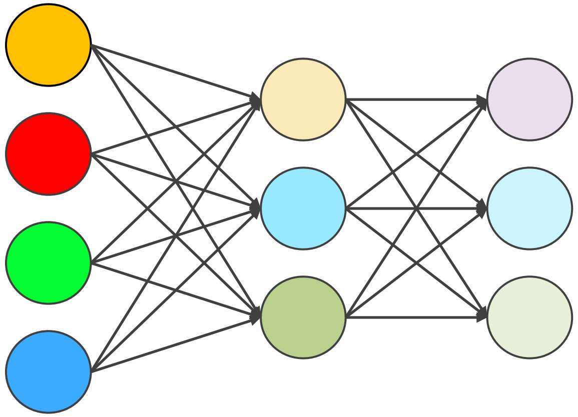


Open Model Zoo

100+ open sourced and optimized pre-trained models;

80+ supported public models

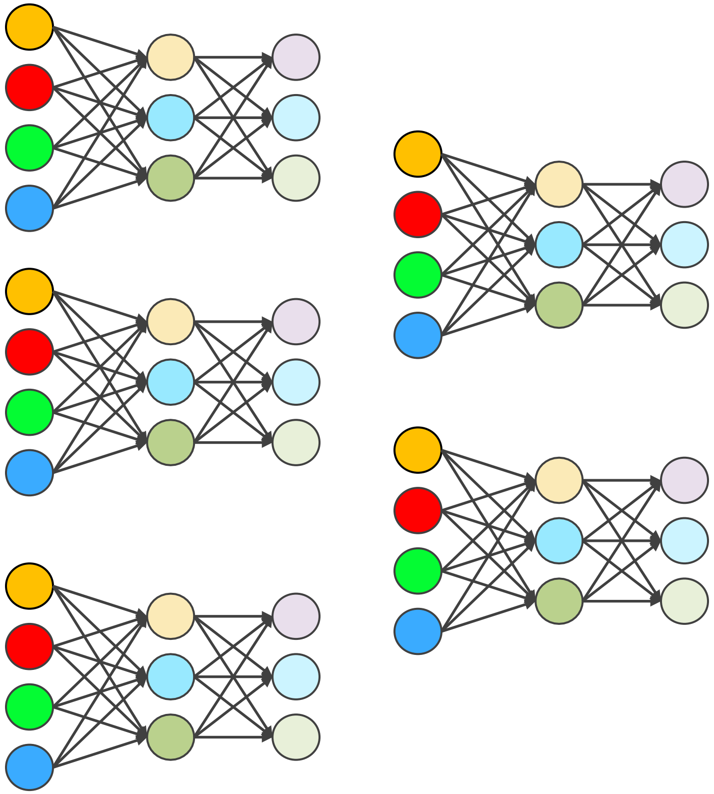
2. Optimize

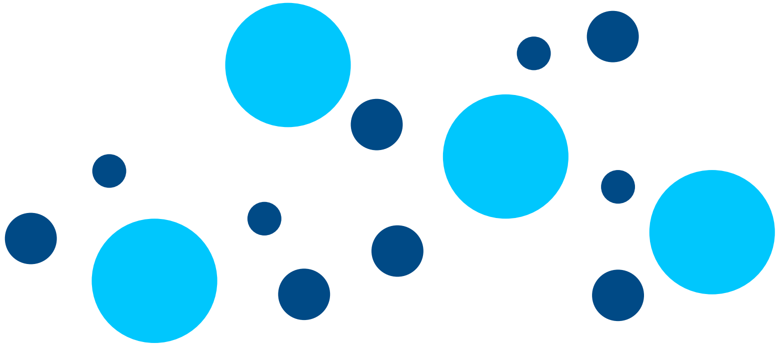


Model Optimizer

Converts and optimizes trained model using a supported framework

3. Deploy





Inference Engine

Common API that abstracts low-level programming for each hardware

OpenVINO™ toolkit + BERT

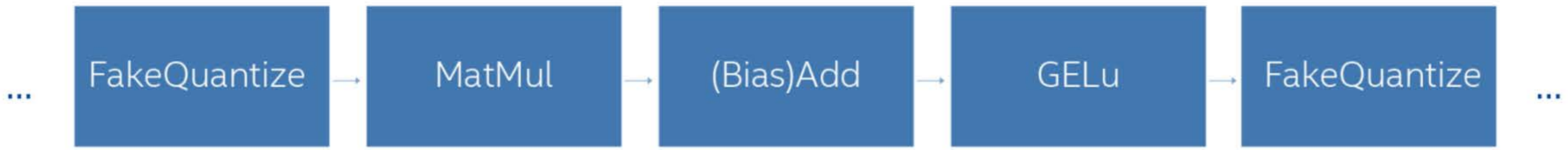
- **Model calibration** is available via the Post-training Optimization Tool for TensorFlow

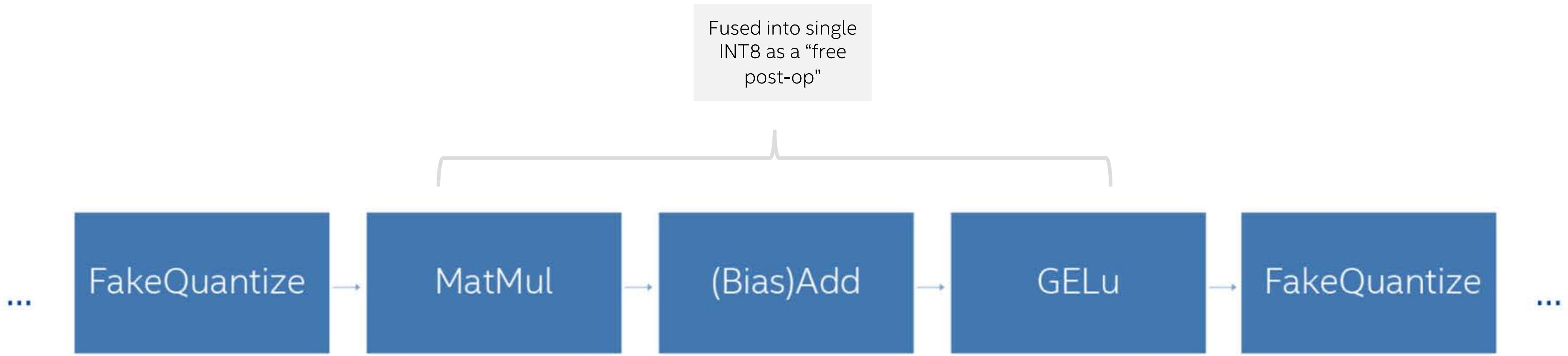
OpenVINO™ toolkit + BERT

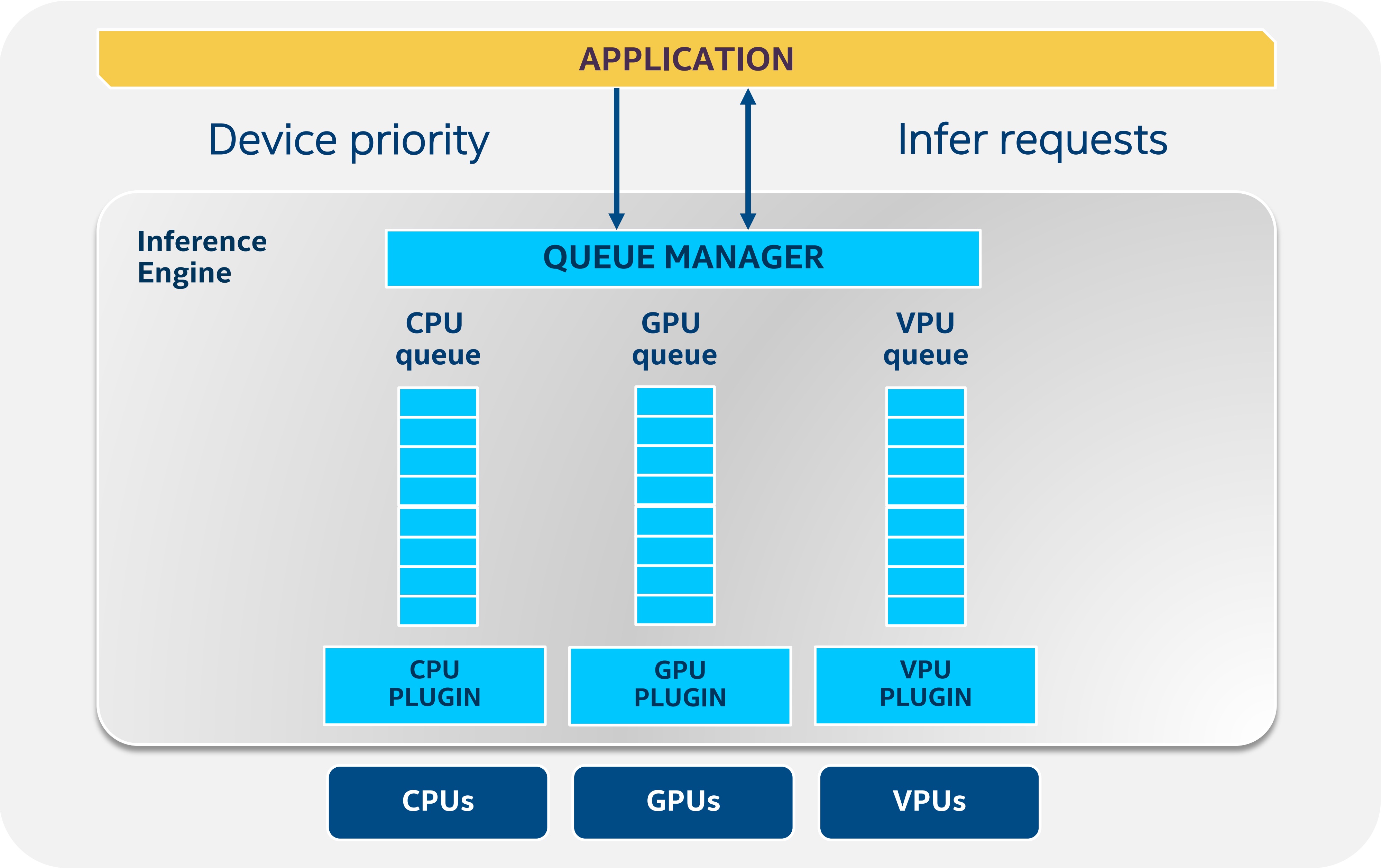
- Model calibration is available via the Post-training Optimization Tool for TensorFlow
- Model fine-tuning or re-training via PyTorch and HuggingFace recipe in **Neural Network Compression Framework**

OpenVINO™ toolkit + BERT

- Model calibration is available via the Post-training Optimization Tool for TensorFlow
- Model fine-tuning or re-training via PyTorch and HuggingFace recipe in Neural Network Compression Framework
- Open-sourced full precision (FP32) and low precision (INT8) models and a demo


...
FakeQuantize
MatMul
(Bias)Add
GELu
FakeQuantize
...


Fused into single INT8 as a “free post-op”


APPLICATION
Device priority
Infer requests
Inference Engine
QUEUE MANAGER
CPU queue
GPU queue
VPU queue
CPU PLUGIN
GPU PLUGIN
VPU PLUGIN
CPUs
GPUs
VPUs

VIDEO

Pad Thai Demo with Audio.mov

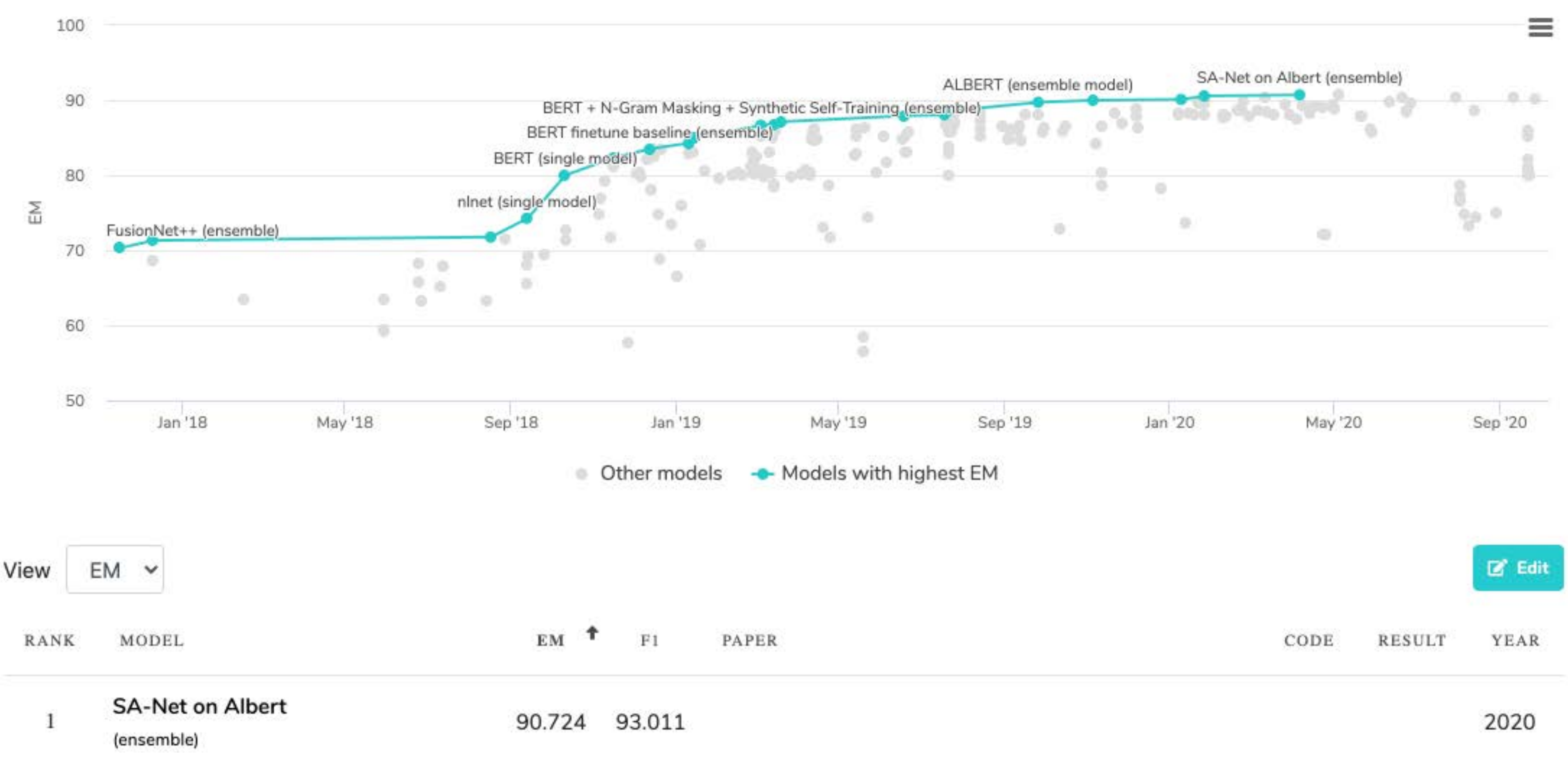
Can we do better?

VIDEO

BERT Large (Conversational).mov

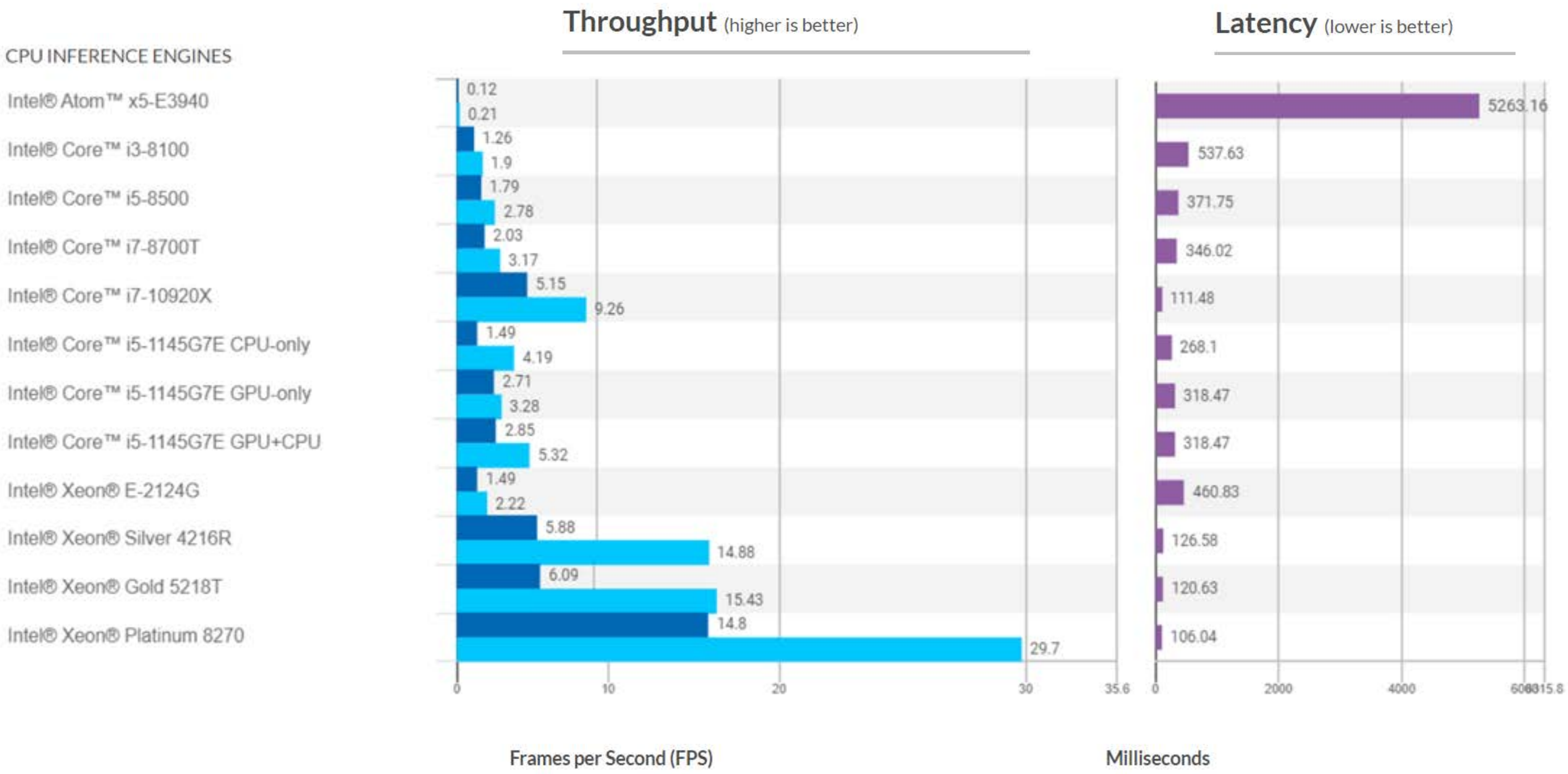
Question Answering on SQuAD2.0

Source: https://paperswithcode.com/sota/question-answering-on-squad20



RANK	MODEL	EM ↑	F1	PAPER	CODE	RESULT	YEAR
1	SA-Net on Albert (ensemble)	90.724	93.011				2020

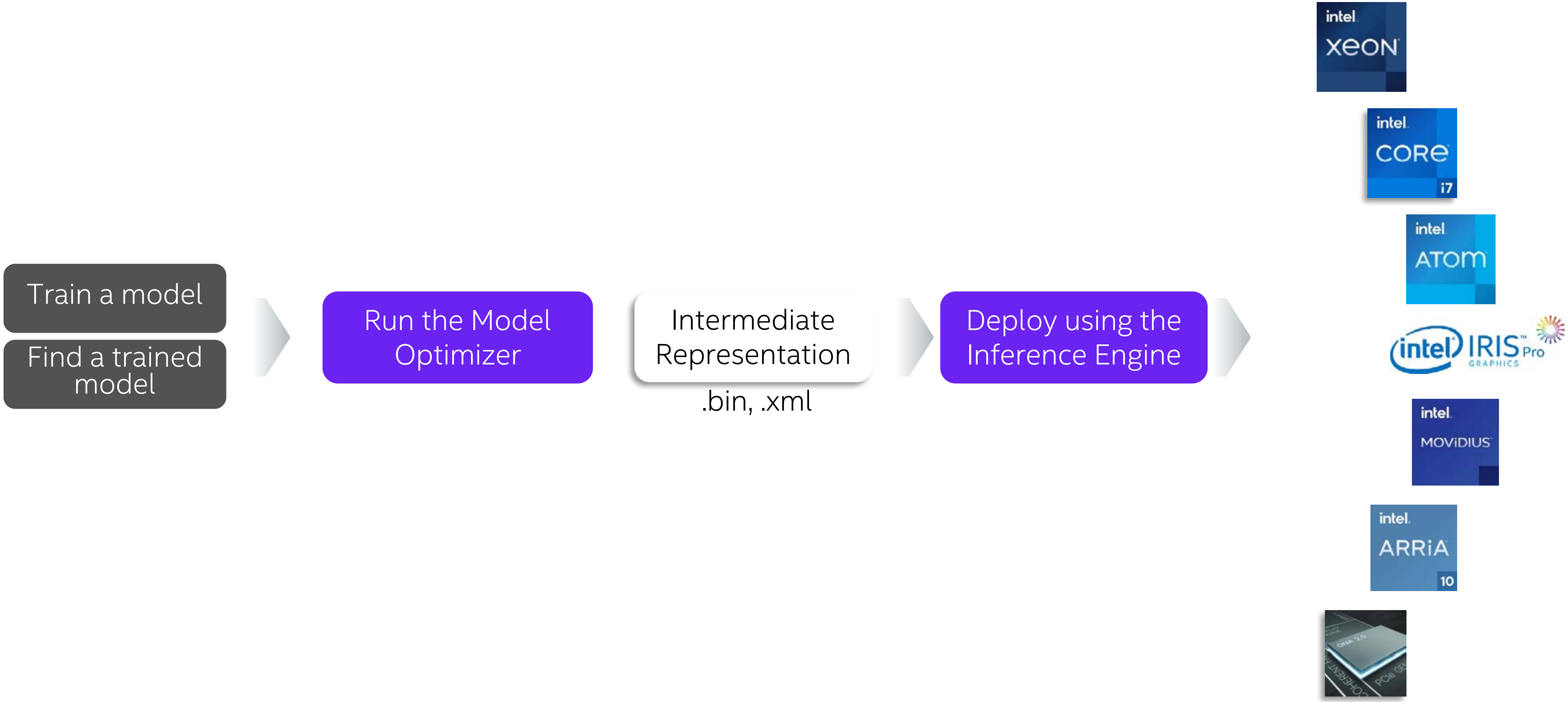
bert-large-uncased-whole-word-masking-squad-int8-0001



Source: https://docs.openvinotoolkit.org/latest/openvino_docs_performance_benchmarks.html

Get Started

Typical workflow from development to deployment



Train a model

Find a trained model

Run the Model Optimizer

Intermediate Representation

.bin, .xml

Deploy using the Inference Engine

What's New in the 2021.1 Release

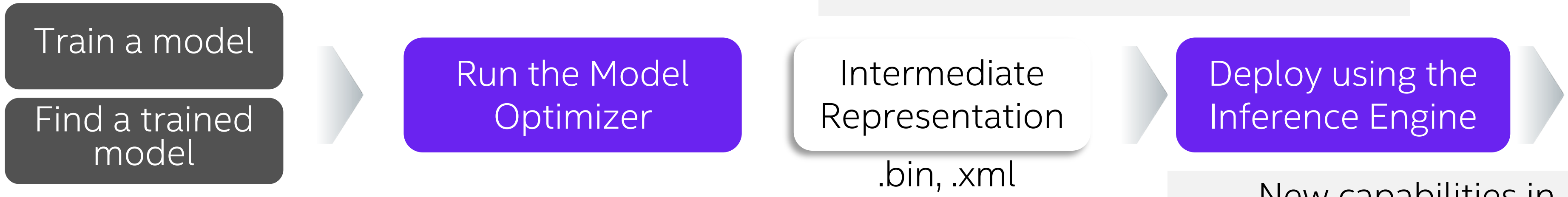
Typical workflow from development to deployment

TensorFlow

Support for TensorFlow 2.x

Support for Tiger Lake (10th Generation Intel® Core® Processors)

Integration of DL Workbench and Intel® DevCloud for the Edge



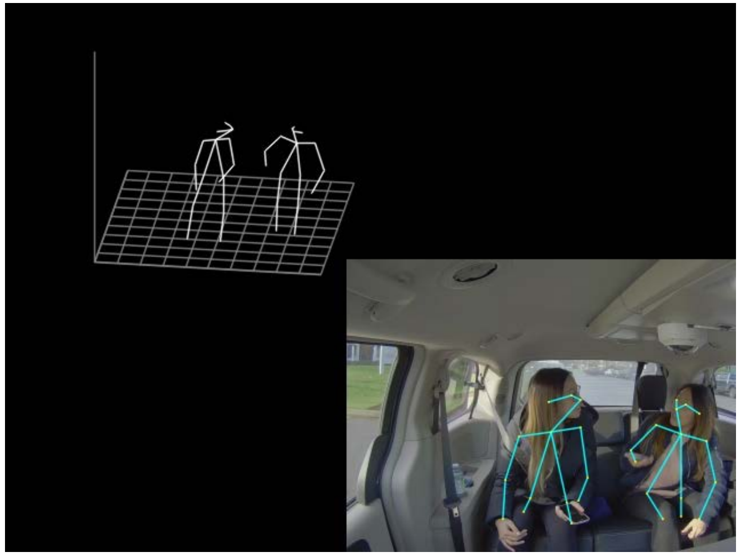
New capabilities in OpenVINO™ Model Server

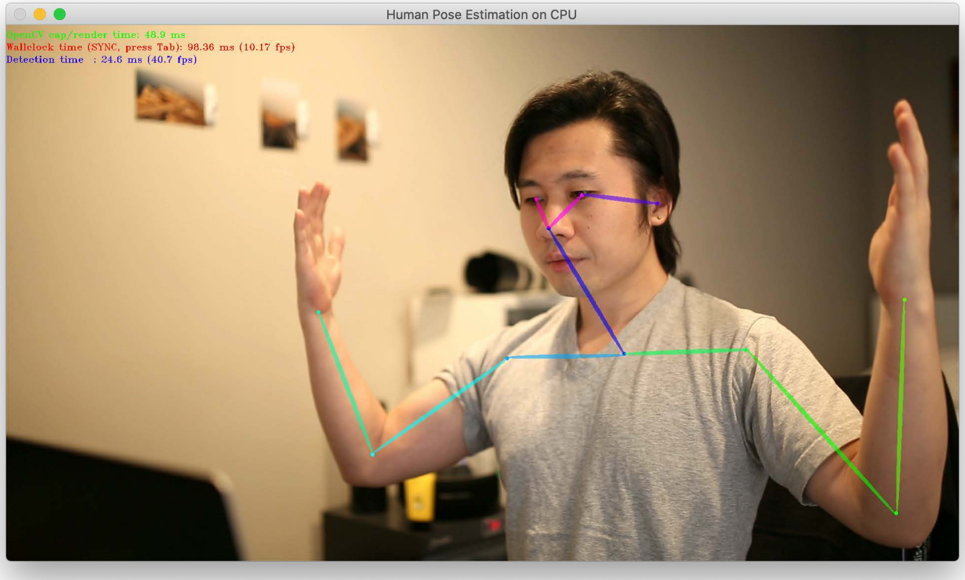
Introducing non-computer vision workloads

Support for GNA 2.0

35+

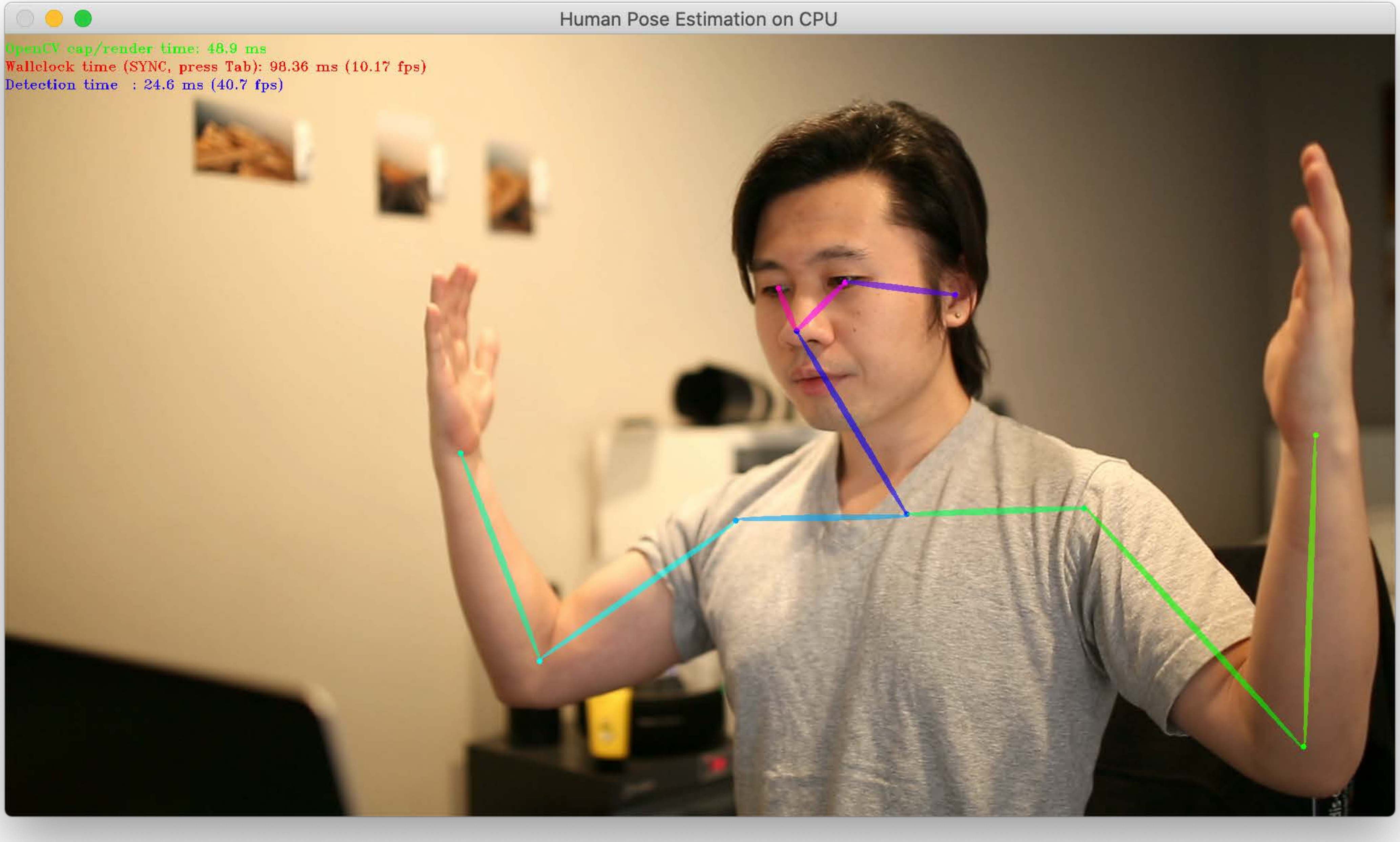
Open Source Deep Learning Demos





Human Pose Estimation on CPU

```
OpenCV cap/render time: 48.9 ms
Wallclock time (SYNC, press Tab): 98.36 ms (10.17 fps)
Detection time  : 24.6 ms (40.7 fps)
```



Human Pose Estimation on CPU

```
OpenCV cap/render time: 48.9 ms
Wallclock time (SYNC, press Tab): 98.36 ms (10.17 fps)
Detection time  : 24.6 ms (40.7 fps)
```

VIDEO

Body Pose.mov

VIDEO

SF Inpainting.mov

Resources and Community Support

Vibrant community of developers, enterprises and skills builders

QUALIFY

INSTALLATION

PREPARE

HANDS ON

SUPPORT

Resources and Community Support

Vibrant community of developers, enterprises and skills builders

QUALIFY

- Use a trained model and check if framework is supported

- or –

- Take advantage of a pre-trained model from the Open Model Zoo

INSTALLATION

- Download the Intel® OpenVINO™ toolkit package from Intel® Developer Zone, or by YUM or APT repositories
- Utilize the Getting Started Guide

PREPARE

- Understand sample demos and tools included
- Understand performance
- Choose hardware option with Performance Benchmarks
- Build, test and remotely run workloads on the Intel® DevCloud for the Edge before buying hardware

HANDS ON

- Visualize metrics with the Deep Learning Workbench
- Utilize prebuilt, Reference Implementations to become familiar with capabilities
- Optimize workloads with these performance best practices
- Use the Deployment Manager to minimize deployment package

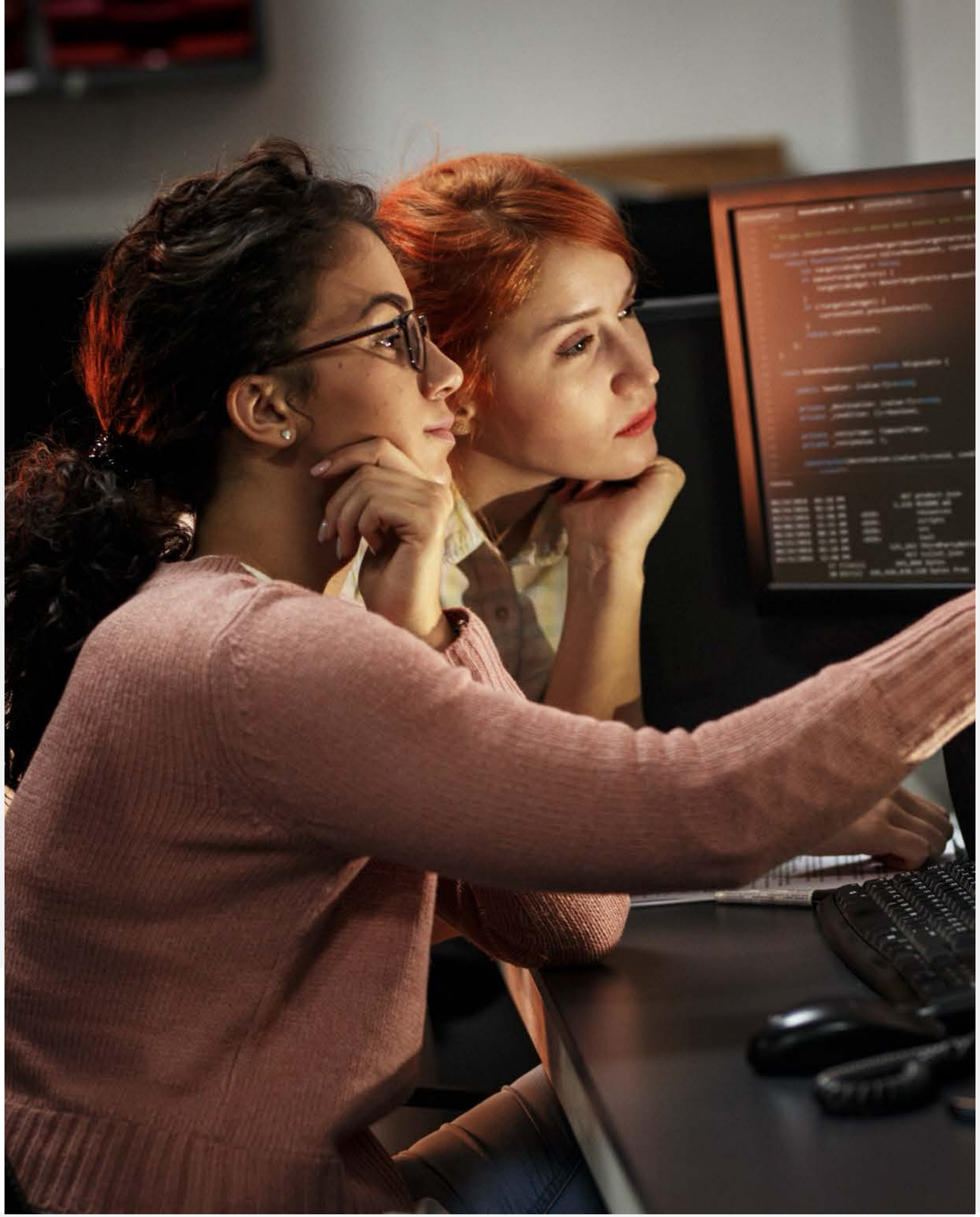
SUPPORT

- Ask questions and share information with others through the Community Forum
- Engage using #OpenVINO on Stack Overflow
- Visit documentation site for guides, how to's, and resources
- Attend training and get certified
- Ready to go to market? Tell us how we can help

Ready to get started?

Download directly from Intel for free:

https://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit.html



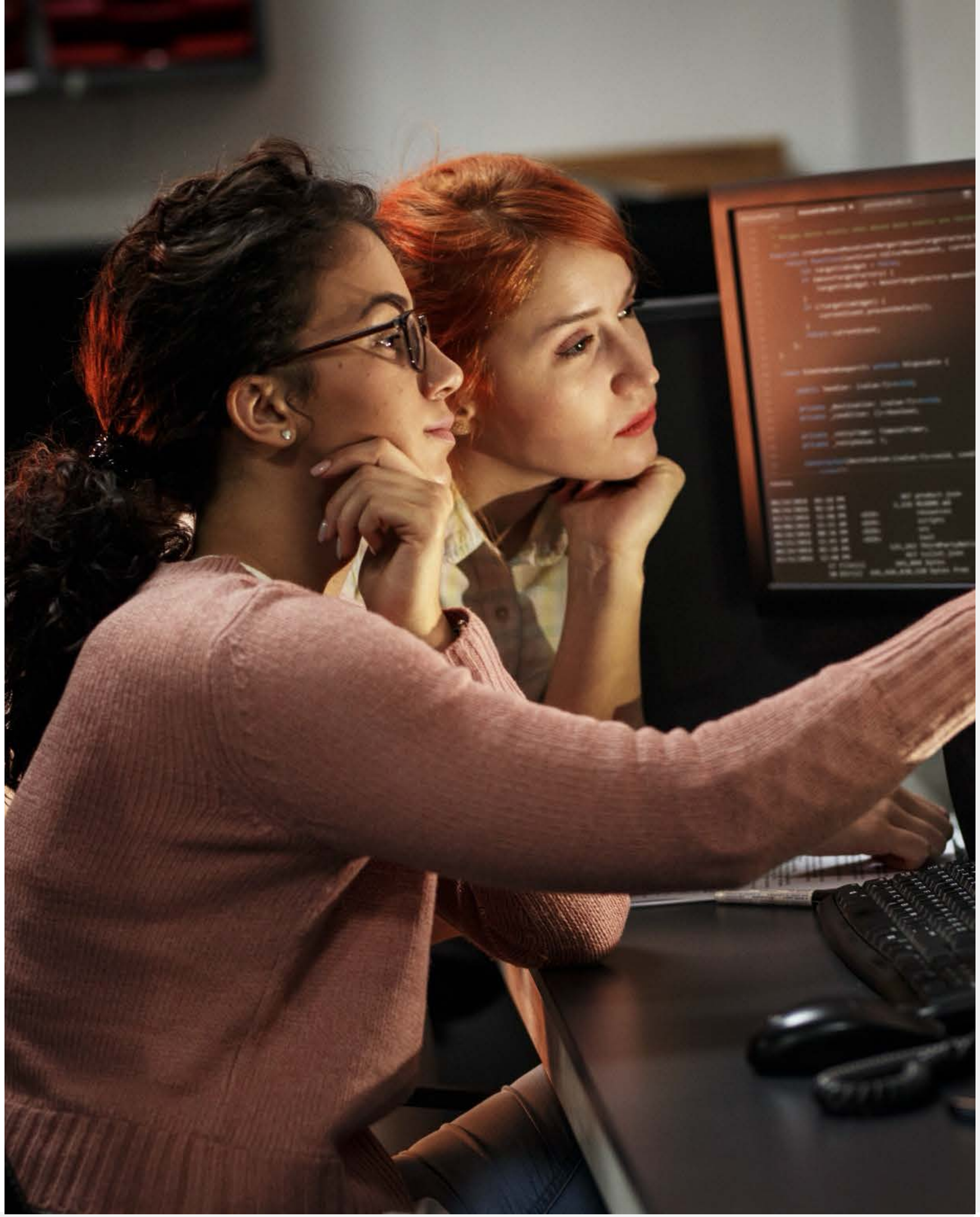
Ready to get started?

Also available from Intel's Edge Software Hub | Intel® DevCloud for the Edge | PIP | DockerHub | Dockerfile | Anaconda Cloud | YUM | APT

Build from source:

https://github.com/openvinotoolkit/openvino

https://gitee.com/OpenVINO-Toolkit



intel®

Notices and Disclaimers

- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.
- Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.
- Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
- Your costs and results may vary.
- Intel technologies may require enabled hardware, software or service activation.
-

Optimization Notice

[1] Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804.

[2] Software and workloads used in performance tests may have been optimized for performance only on microprocessors from Intel. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. Consult other information and performance tests while evaluating potential purchases, including performance when combined with other products. For more information, see Performance Benchmark Test Disclosure. Source: Intel measurements, as of June 2017.