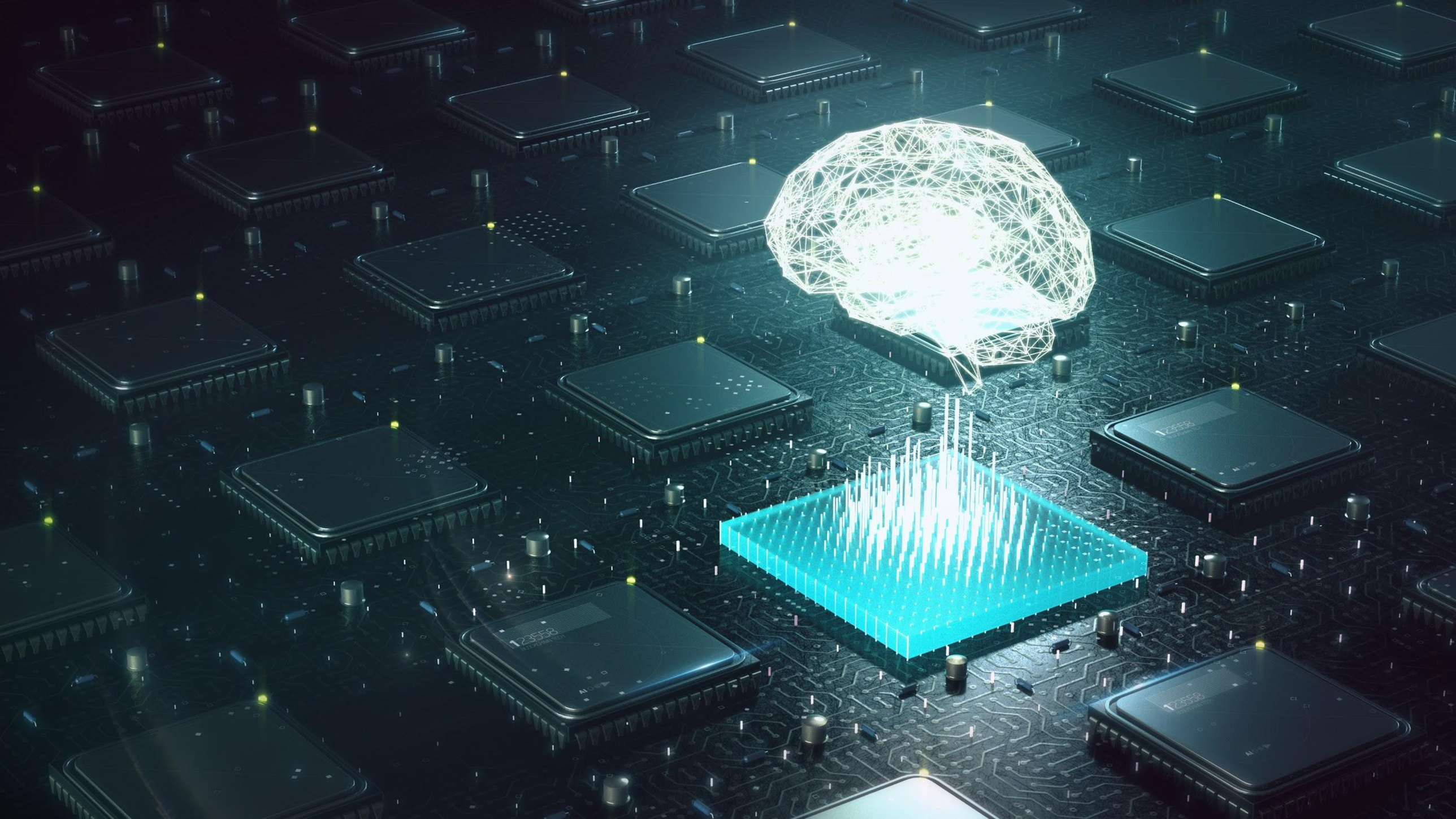
ML and AI at the edge

with Google Coral, Edge TPU and Cloud Platform

Markku Lepisto
IoT/ML Solutions Architect, JAPAC

Google Cloud

Google Cloud



Google’s AI Principles

1. Be **socially beneficial**.
2. **Avoid** creating or reinforcing **unfair bias**.
3. Be built and tested for **safety**.
4. Be **accountable** to people.
5. Incorporate **privacy** design principles.
6. Uphold high standards of **scientific** excellence.
7. Be made **available** for uses that accord with these principles.


Beam energy 4.1 to 4.4 Bev
number neg. particles, p mass. = 38 / 1,819,000 = 1 / 48000
number mesons
Momentum of neg. particle beam: 1.187 Bev/c
of neg. particles of p mass: 0.78
Energy 572 Mev


Google


Google


Machine
Learning
tab
shift

Introducing

Coral

Edge AI

Coral

Privacy

Consumers &
businesses
(privacy & compliance)

Latency

Self-driving car
@65 miles/hr
100 ms = 9.5 ft

Bandwidth

Always-ON,
4K frames,
1 frame = 24 MB

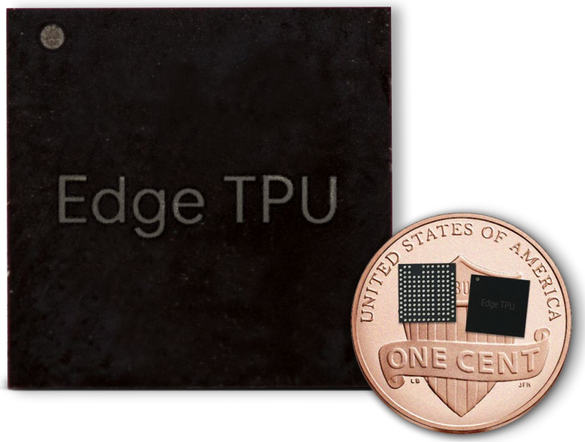
Offline

Offline, roaming,
poor connection

Cost

Streaming Cost,
perf/$$

Edge TPU



4 TOPS | 2 Watts

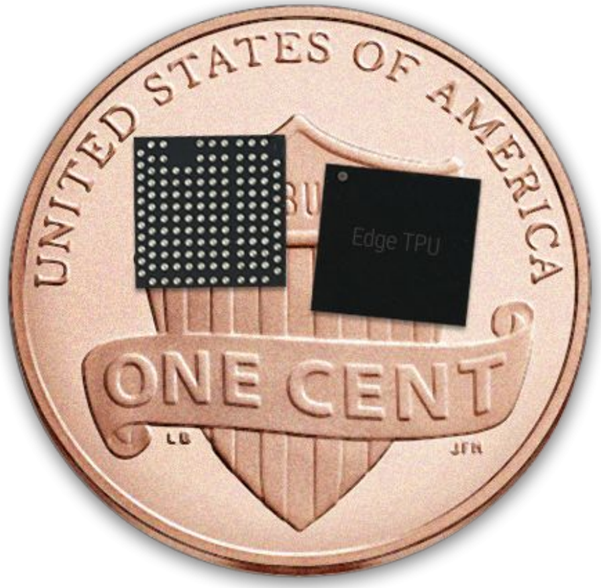
INT8 | INT16 | TFLite

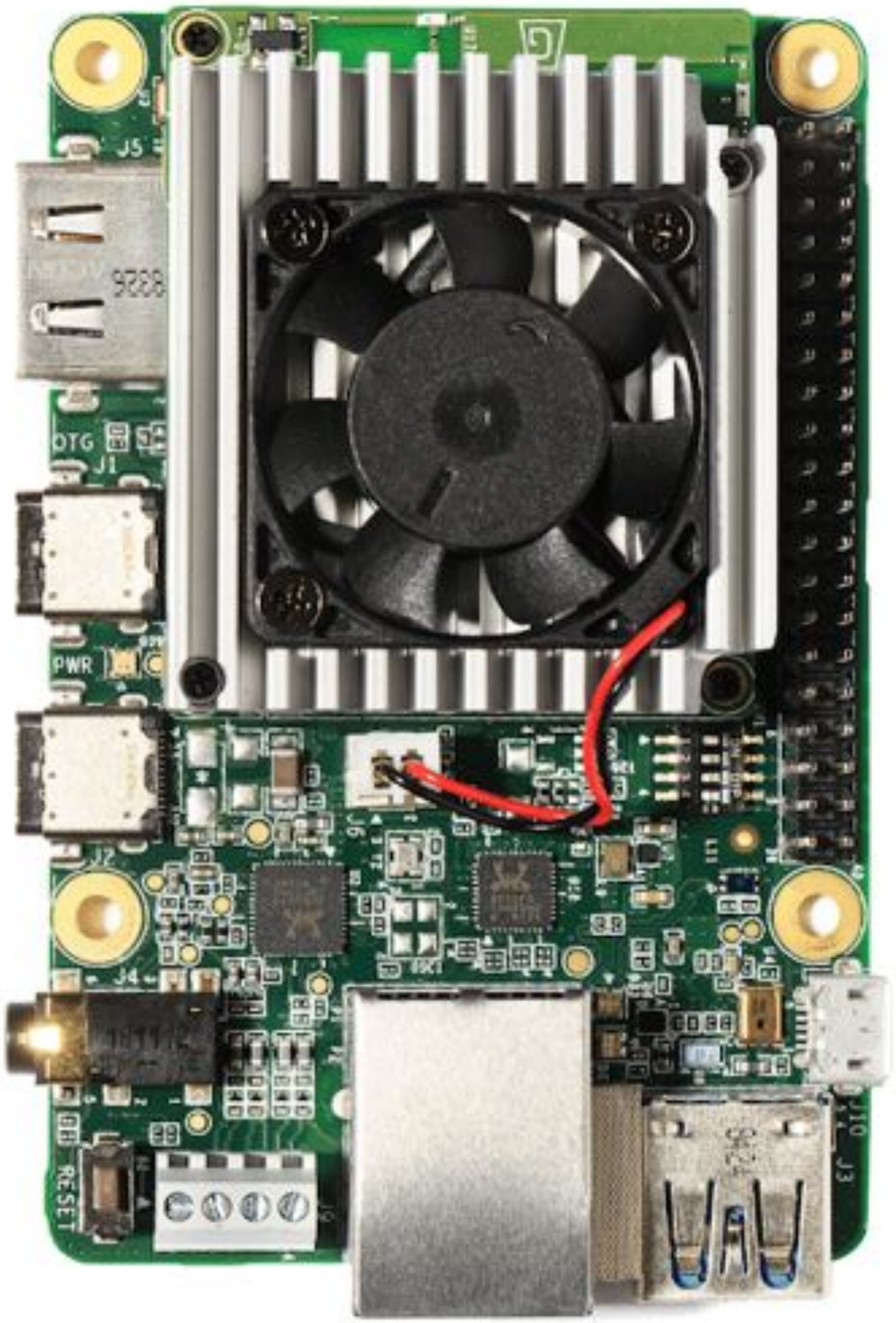
Inference accelerator

- Optimized for vision applications and convolutional neural networks
- Runs concurrent state-of -the-art models on high-resolution video, at real-time (30+ FPS)
- Full support for quantized TensorFlow Lite models

Coral boards feature the Edge TPU. A purpose-built ASIC designed to bring inference to the edge

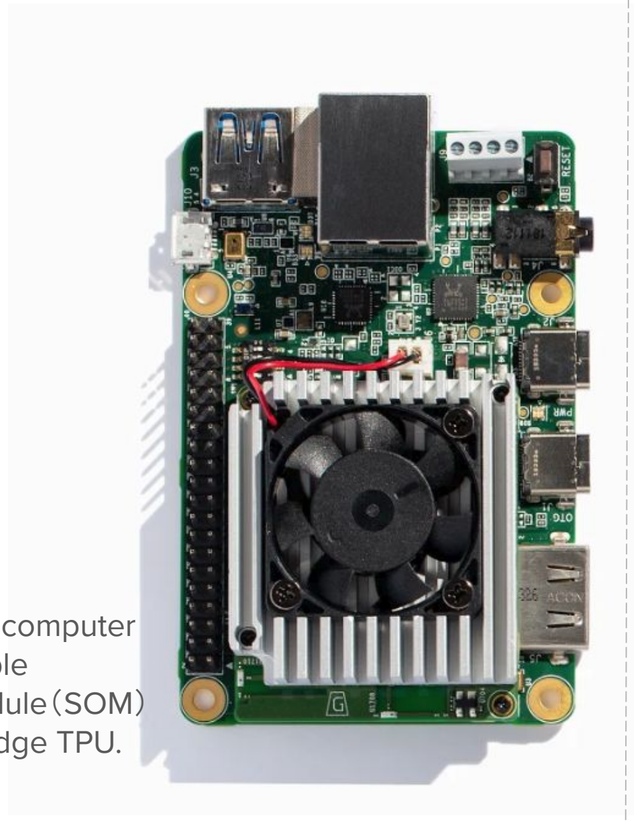
Coral Edge TPU





Coral product portfolio

Coral
www.coral.ai

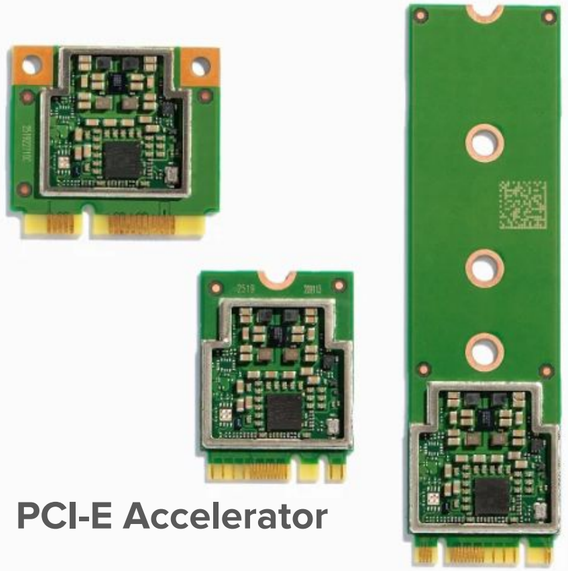


Dev Board

A single-board computer with a removable system-on-module (SOM) featuring the Edge TPU.

Available now

Price
$149.99



PCI-E Accelerator

PCI-E device for easy integration of Edge TPU into existing systems.

Available now

Price
$34.99

USB Accelerator

A USB accessory featuring the Edge TPU that brings ML inferencing to existing systems.

Available now

Price
$74.99

SOM

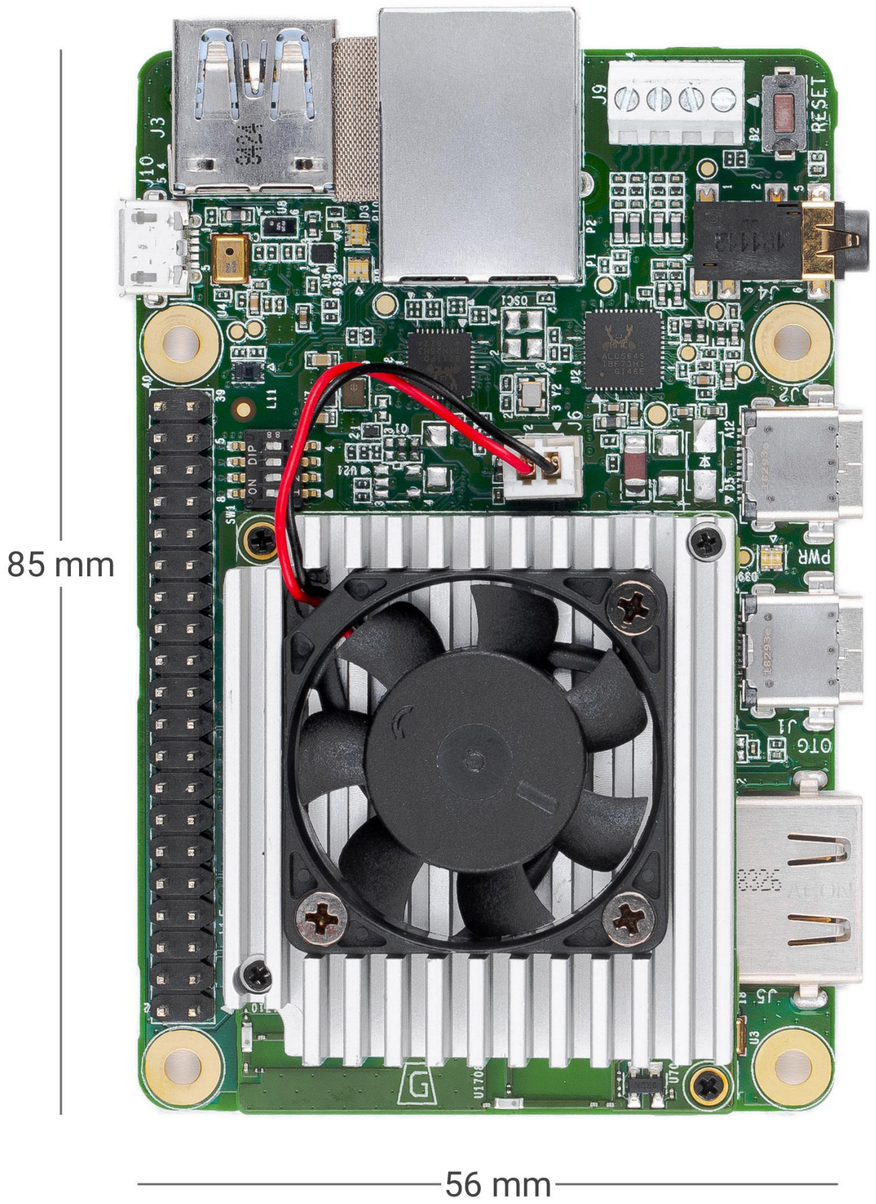
A fully integrated System on Module in a 40mm x 48mm pluggable module.

Available now

Price
$114.99

Features

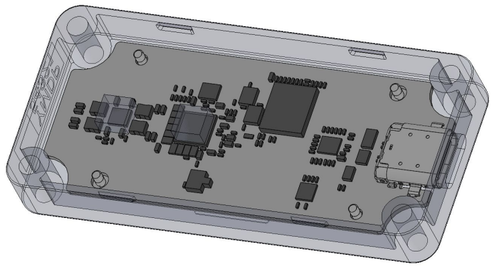
- Edge TPU Module (SOM)
  - NXP i.MX 8M SOC (Quad-core Cortex-A53, plus Cortex-M4F)
  - Google Edge TPU ML accelerator coprocessor
  - Cryptographic coprocessor
  - Wi-Fi 2x2 MIMO (802.11b/g/n/ac 2.4/5GHz), Bluetooth 4.1
  - Vivante GC7000L GPU
  - 8GB eMMC
  - 1GB LPDDR4
- USB connections
  - USB Type-C power port (5V DC)
  - USB 3.0 Type-C OTG port
  - USB 3.0 Type-A host port
  - USB 2.0 Micro-B serial console port
- Audio connections
  - 3.5mm audio jack (CTIA compliant)
  - Digital PDM microphone (x2)
  - 2.54mm 4-pin terminal for stereo speakers
- Video connections
  - HDMI 2.0a (full size)
  - 39-pin FFC connector for MIPI-DSI display (4-lane)
  - 24-pin FFC connector for MIPI-CSI2 camera (4-lane)
- MicroSD card slot
- Gigabit Ethernet port
- 40-pin GPIO expansion header (Raspberry Pi style)
- Supports Debian Linux (Mendel)

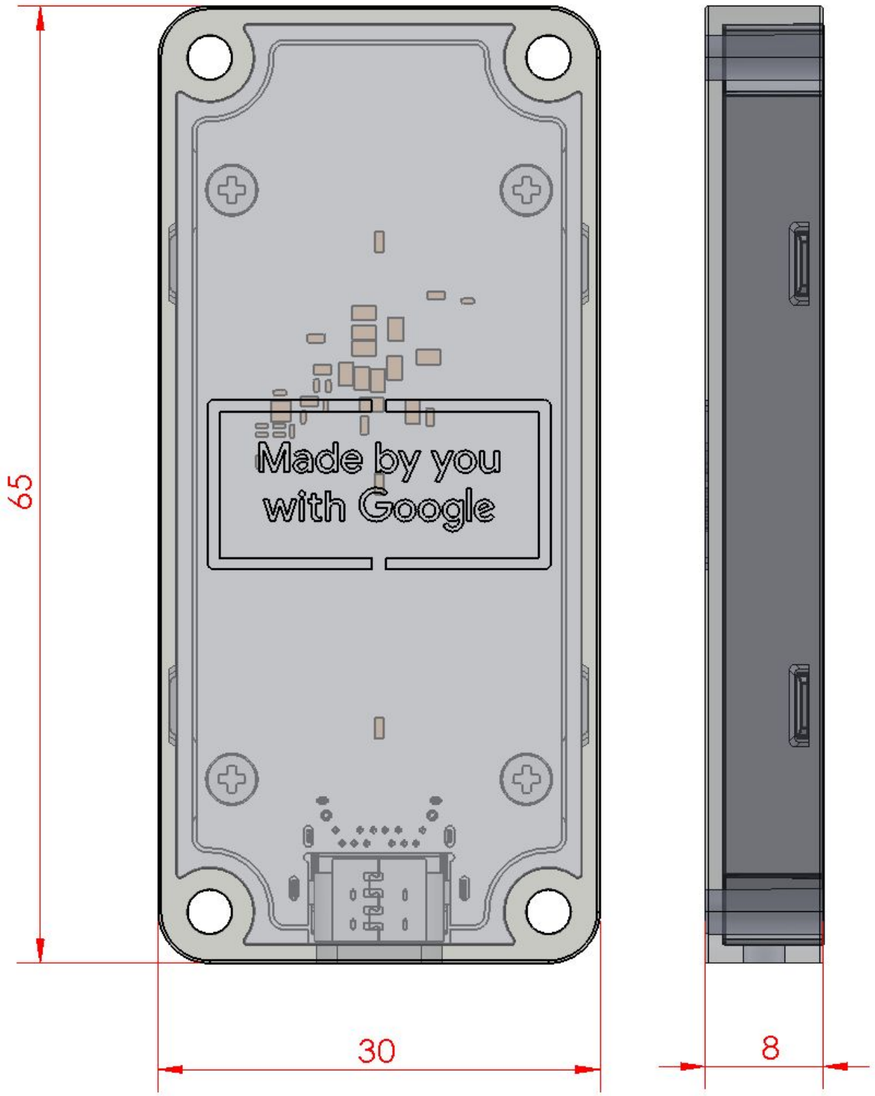


Coral

USB 3.0 accessory device featuring EdgeTPU

MCU	Cortex-M0 MCU (power, thermal monitor)
TPU	Google ASIC
RAM Memory	11 MB
Power	USB 3.0, Type-C connector
Data	USB 3.0, Type-C connector
Supported OS	Debian/Raspbian Linux
Supported ML	TensorFlow, TF Lite
Supported Models	Inception v1/2/3, Mobile Net, Daredevil





Coral

Software Toolchain

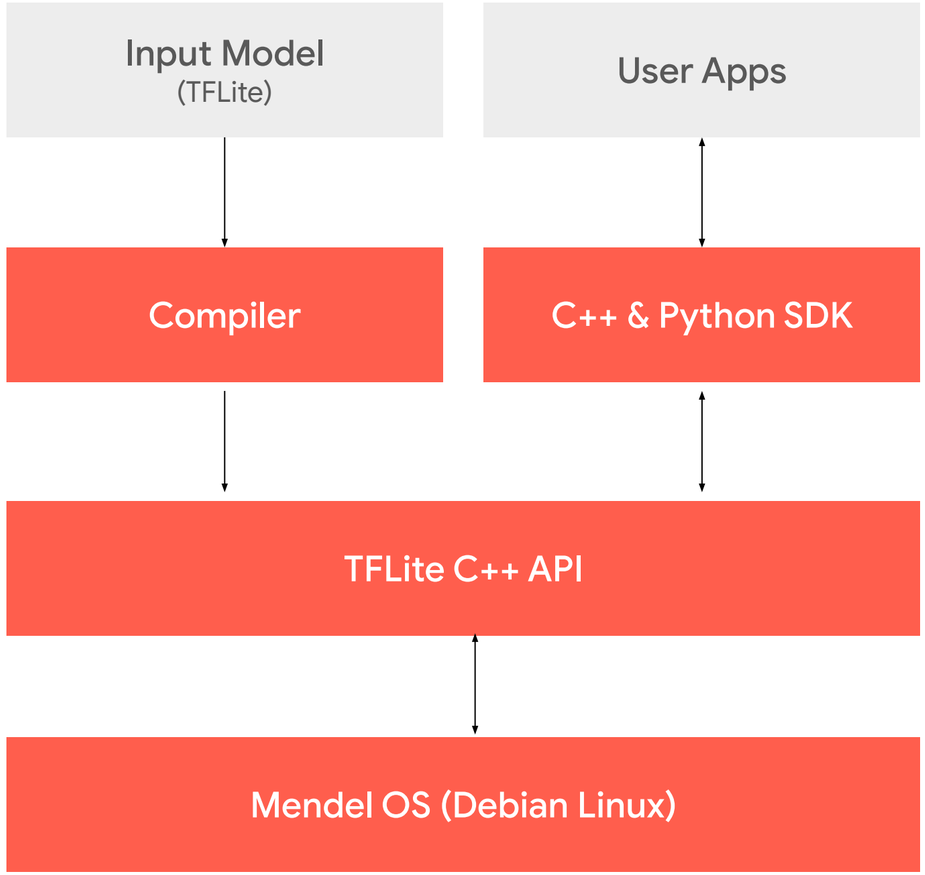
The system software and toolchain to allow developers to run, manipulate, and create with our products.

Mendel OS
A fork of the Debian OS to power our Intelligence Boards, and a

C++ & Python SDK
APIs to low level connections

Edge TPU Compiler
Converts TF graphs to run on targeted chipsets

Companion Software
Abstracts away traditional board management/coding in a high-level program



Code Example with Object Detection

```
# Initialize engine.
  engine = DetectionEngine(args.model)
  labels = ReadLabelFile(args.label) if args.label else None

# Open image.
  img = Image.open(args.input)
  draw = ImageDraw.Draw(img)

# Run inference.
  ans = engine.DetectWithImage(img, threshold=0.05,keep_aspect_ratio=True,
                               relative_coord=False, top_k=10)

  if ans:
      for obj in ans:
      box = obj.bounding_box.flatten().tolist()
      # Draw a rectangle.
      draw.rectangle(box, outline='red')
```

Demo

Edge TPU inference

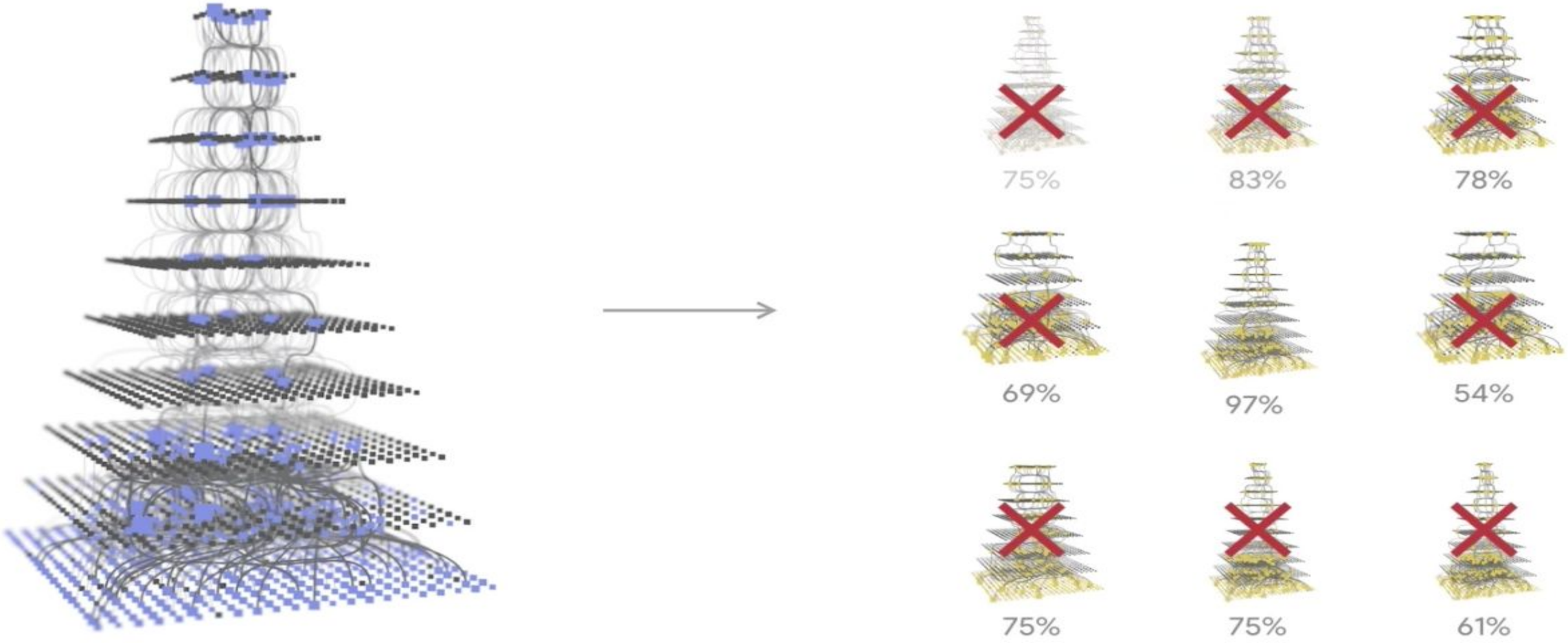
person detection and privacy protection
with pre-compiled posenet model

Google Cloud

How about custom models?

AutoML

Learning to Learn



sushi – Vision – Solution Demo

pantheon.corp.google.com/vision/datasets/ICN7492507044426984519/images?folder=&organizationId=&project=solution-demo

Google Cloud Platform | Solution Demo

Vision

- Dashboard
- Datasets
- Models

sushi | LABEL STATS | EXPORT DATA

IMPORT | IMAGES | TRAIN | EVALUATE | TEST & USE

Single-Label Classification

All images	1,937
Labeled	1,937
Unlabeled	0

Filter labels

ebi_nigiri	319
ika_nigiri	314
ikura_gunkan	325
maguro_nigiri	342
tamago_nigiri	305
uni_gunkan	332

ADD NEW LABEL

Filter images

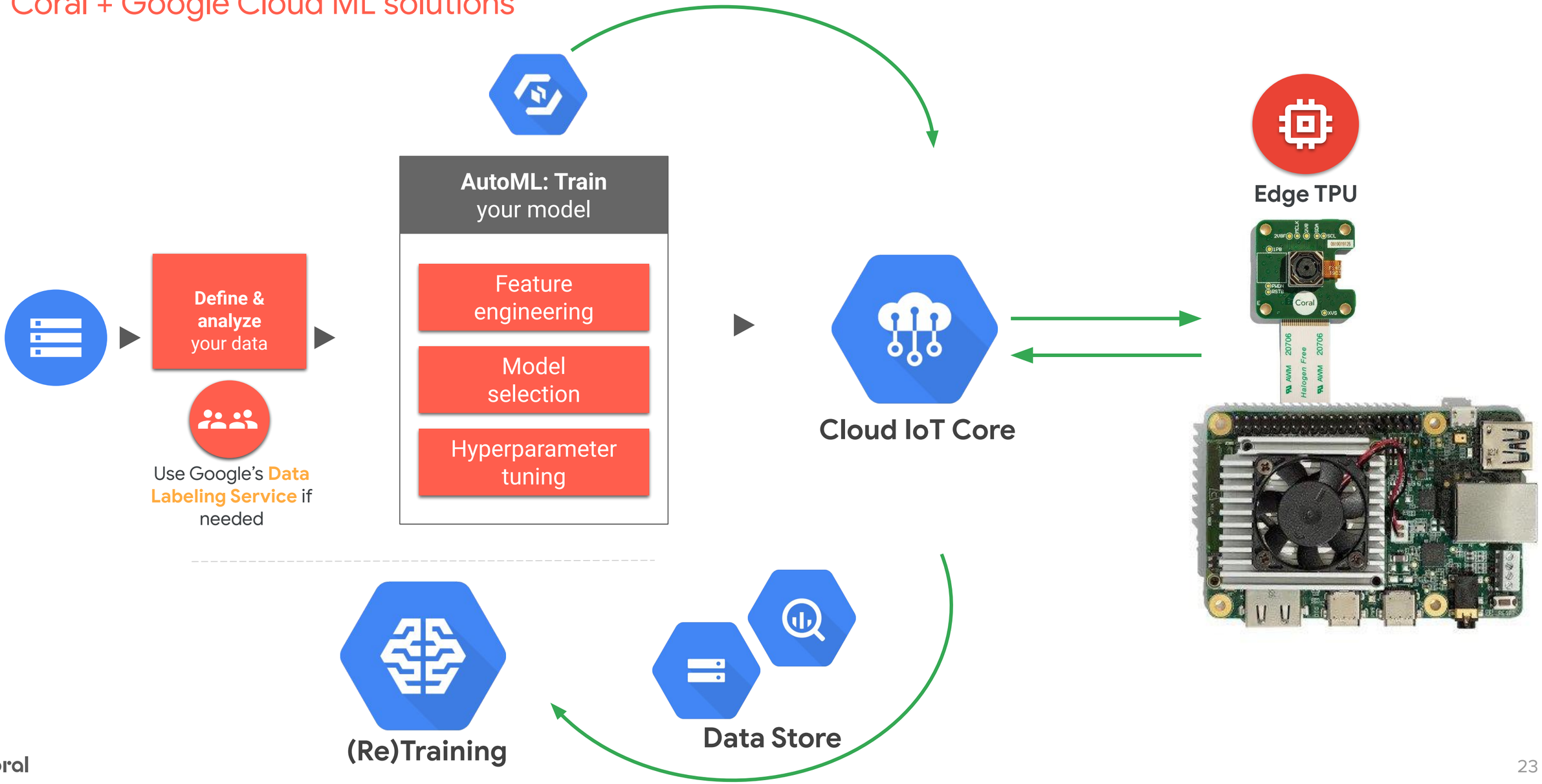
Images per page: 50 | 1 – 50 of many

Show debug panel

How it works with Cloud

Coral + Google Cloud ML solutions

Coral
www.coral.ai



Demo

Edge TPU inference

with AutoML custom sushi model

Google Cloud

Demos are cool..
but we live in the Real World™

Customer Case

Manufacturing Visual Inspection

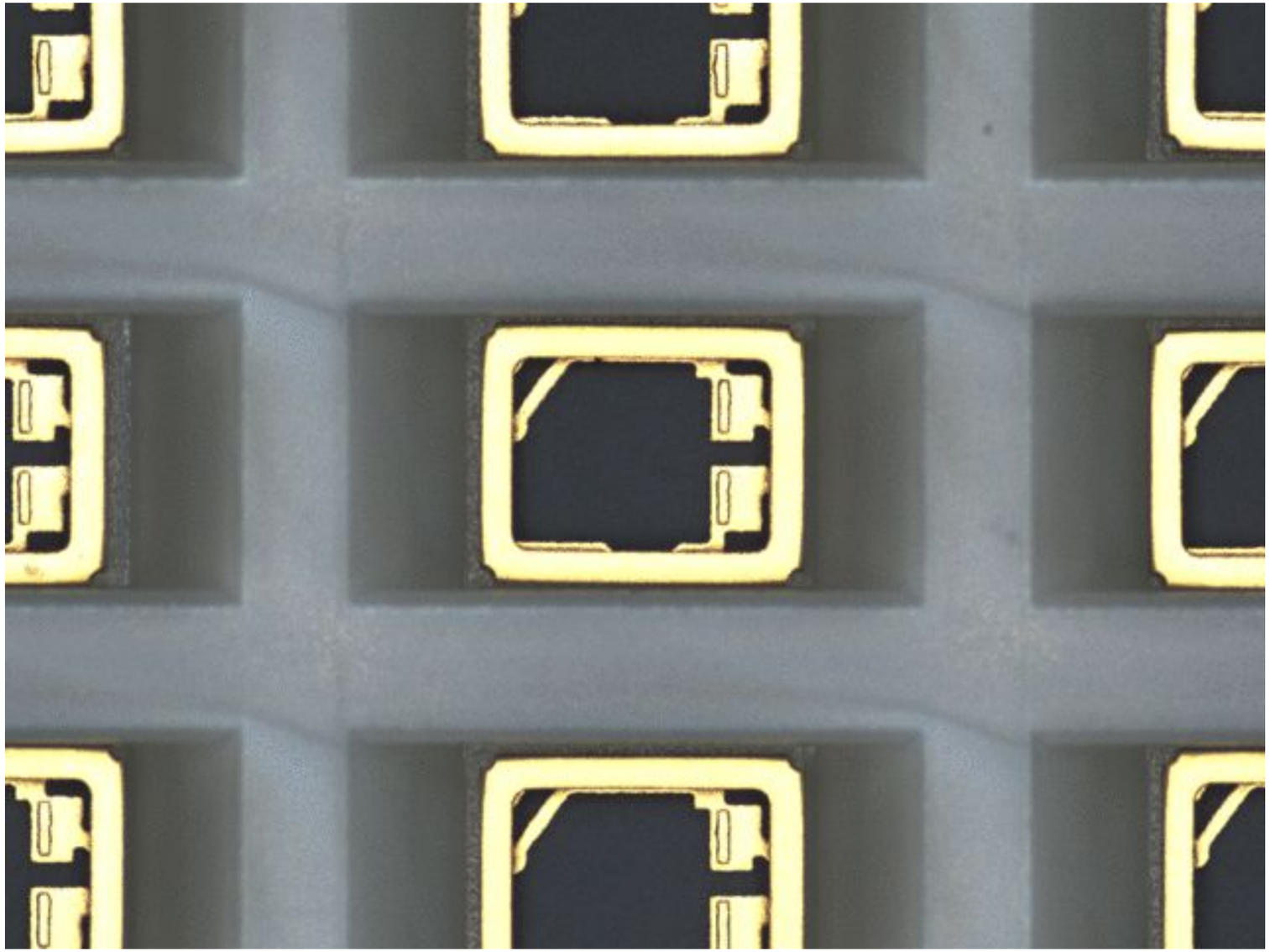
Google Cloud

The Problem

Multiple objects partially visible

Target object moves around, rotated at random angles

Defects are small blemishes on the object



Google Cloud

Customer Initial Testing

Standard AutoML Usage

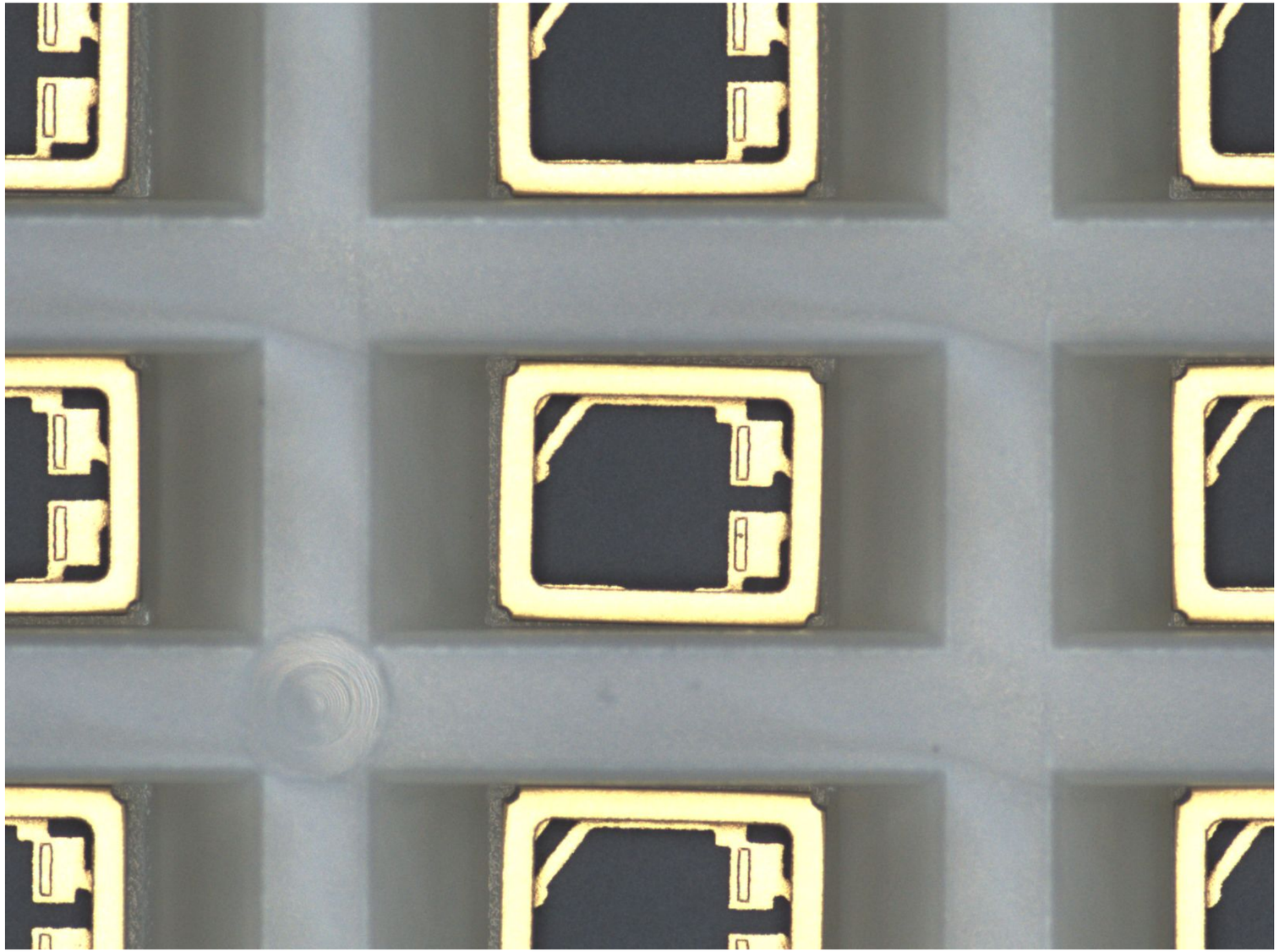
True label \ Predicted label	ng_xbp	ng_dp	ng_sp	ok
ng_xbp	14.3%	71.4%	-	14.3%
ng_dp	-	71.4%	14.3%	14.3%
ng_sp	-	-	71.4%	28.6%
ok	-	-	-	100.0%

True label \ Predicted label	ng	ok
ng	56.3%	43.8%
ok	-	100.0%

Google Cloud

The Solution

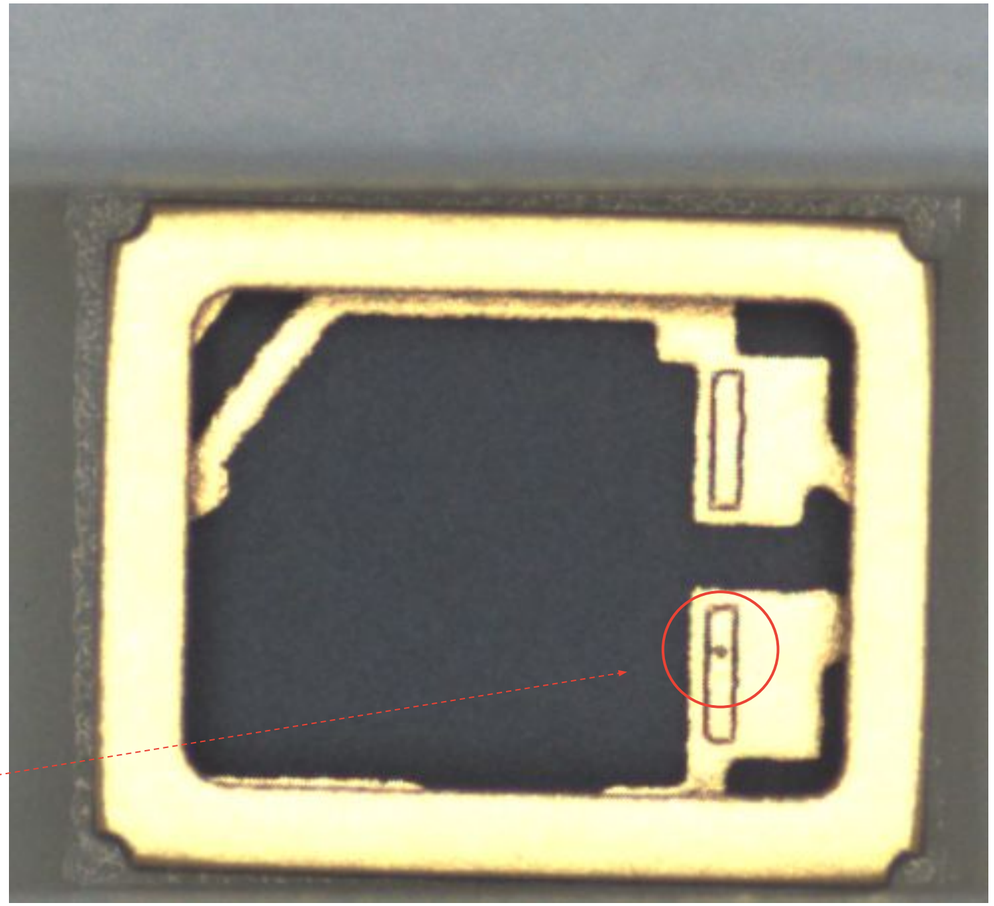
Pre-processing



Google Cloud

Pre-processing Steps

Rough crop of target object

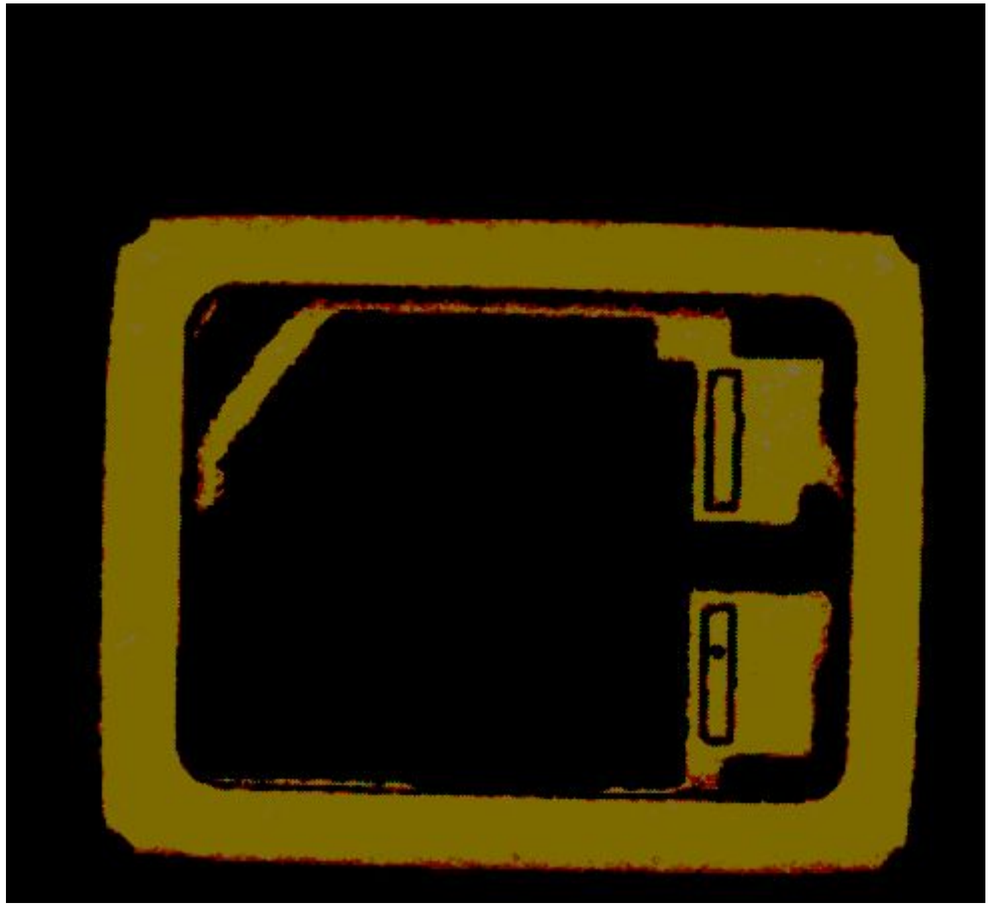


A defect

Pre-processing Steps

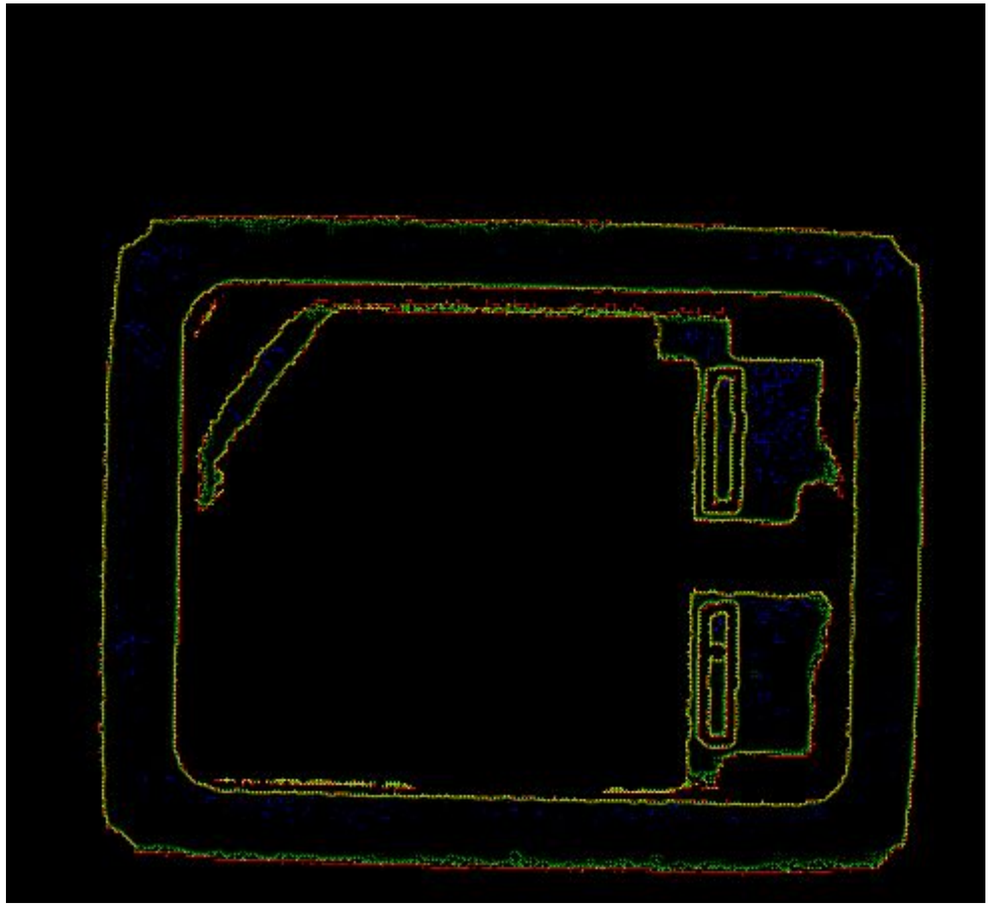
Remove background

- Filter background pixels to zero
- Remaining non-zero pixels are the object



Pre-processing Steps

Detect edges

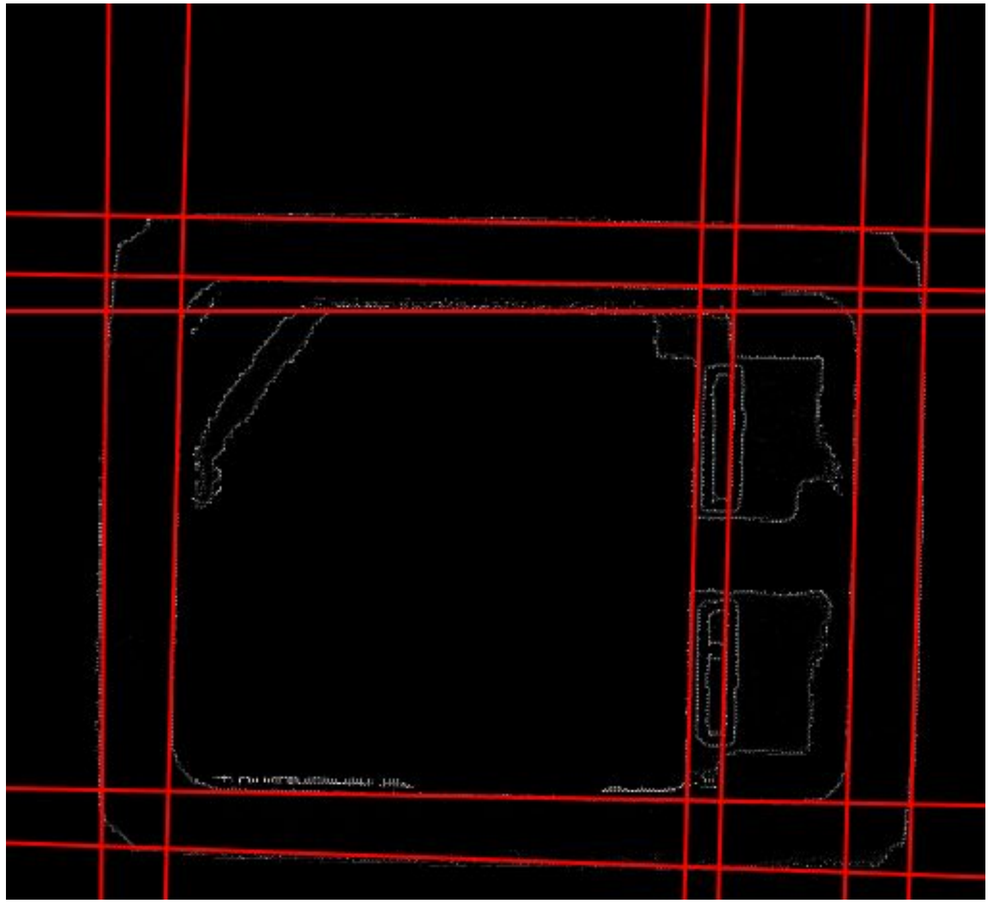


Google Cloud

Pre-processing Steps

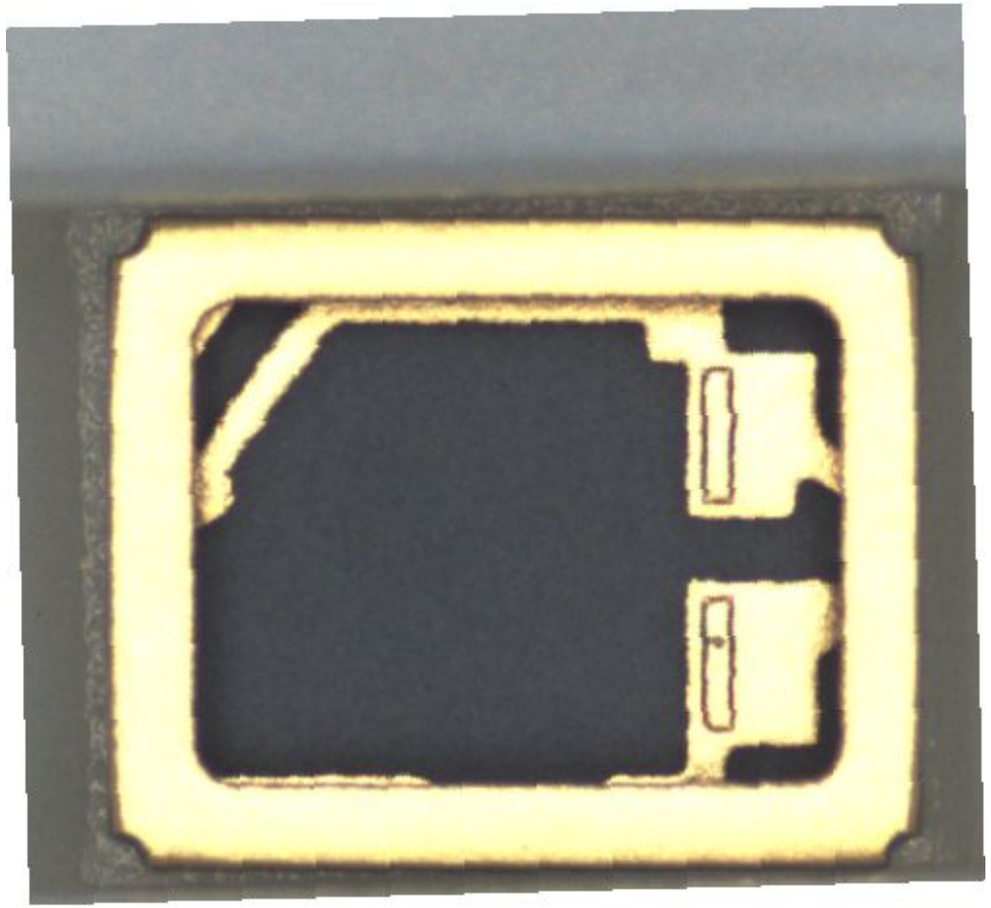
Detect lines with Hough transform

- Calculate rotation angle
- Remove outliers



Pre-processing Steps

(Counter) rotate image to zero degrees

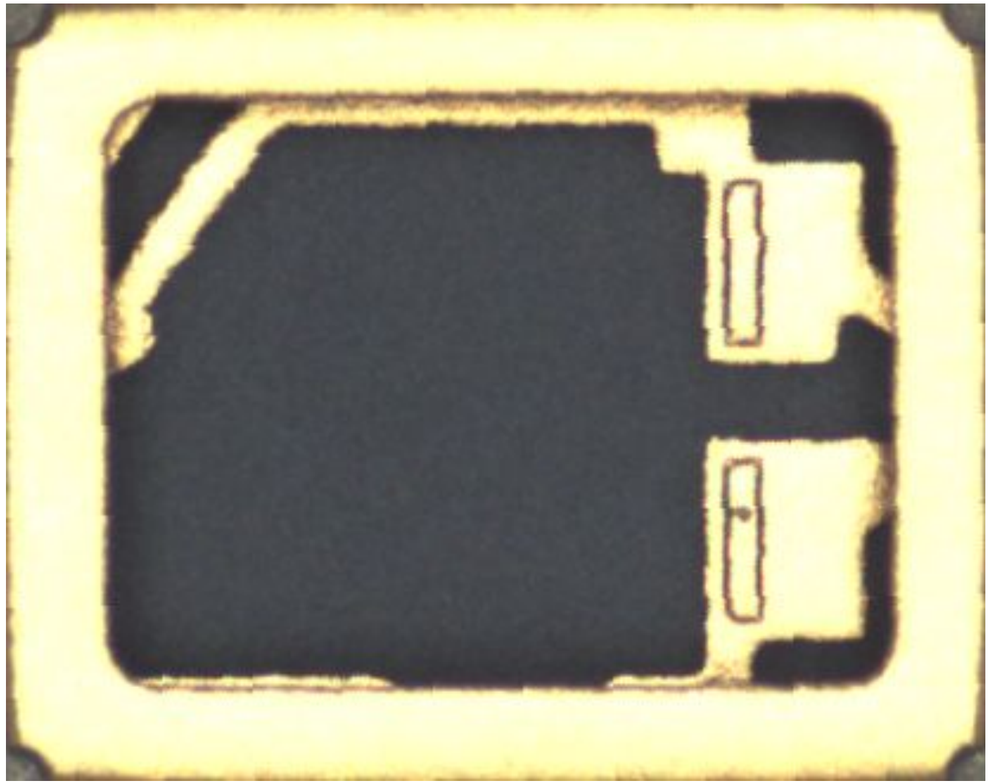


Google Cloud

Pre-processing Steps

Detect bounding box and crop tightly

→ **Training image for AutoML**



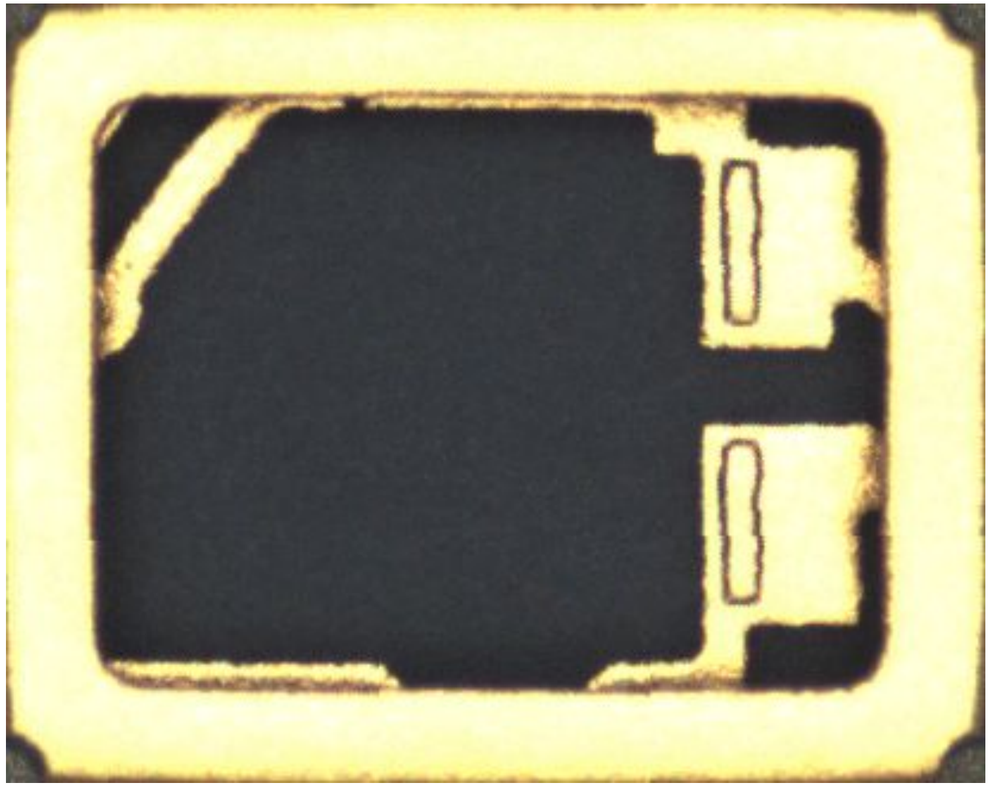
EdgeTPU Test Results
using Coral DevBoard

Confusion Matrix

	Predicted NG	Predicted OK
True NG	95.10%	4.90%
True OK	1.47%	98.53%

Performance

Average inference latency	8.98 ms

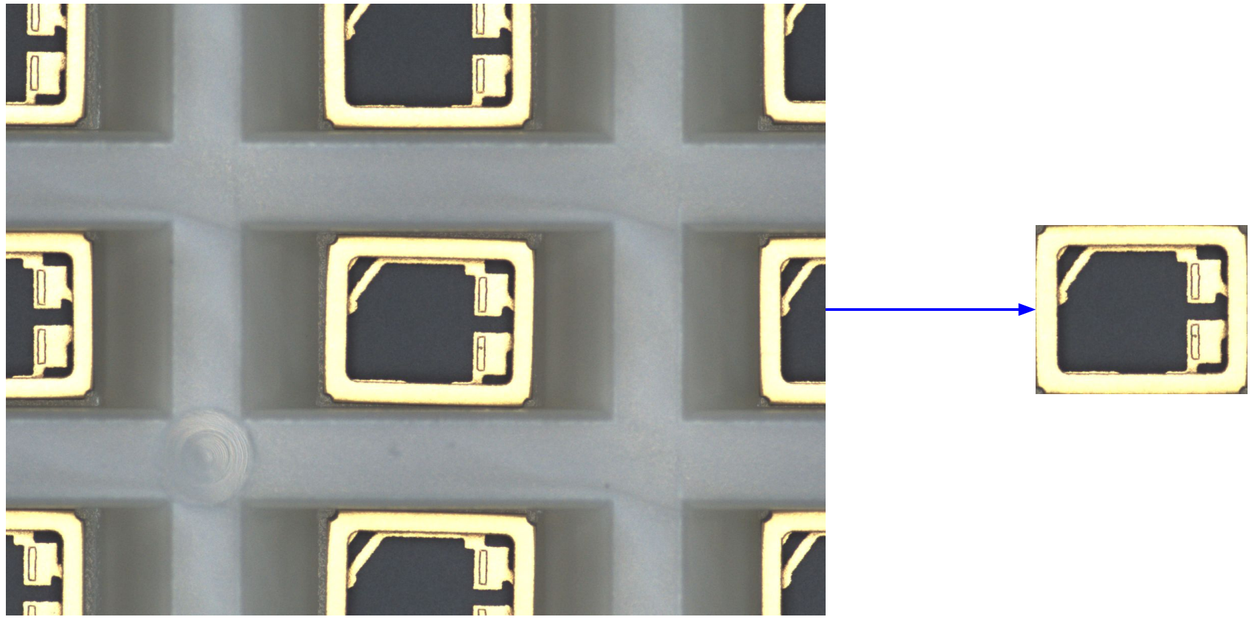


Google Cloud

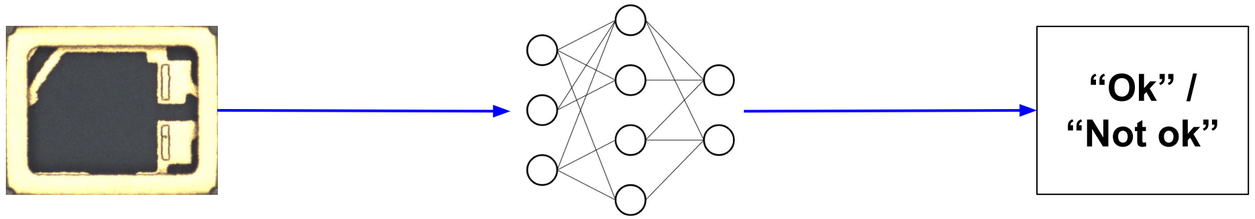
Demo

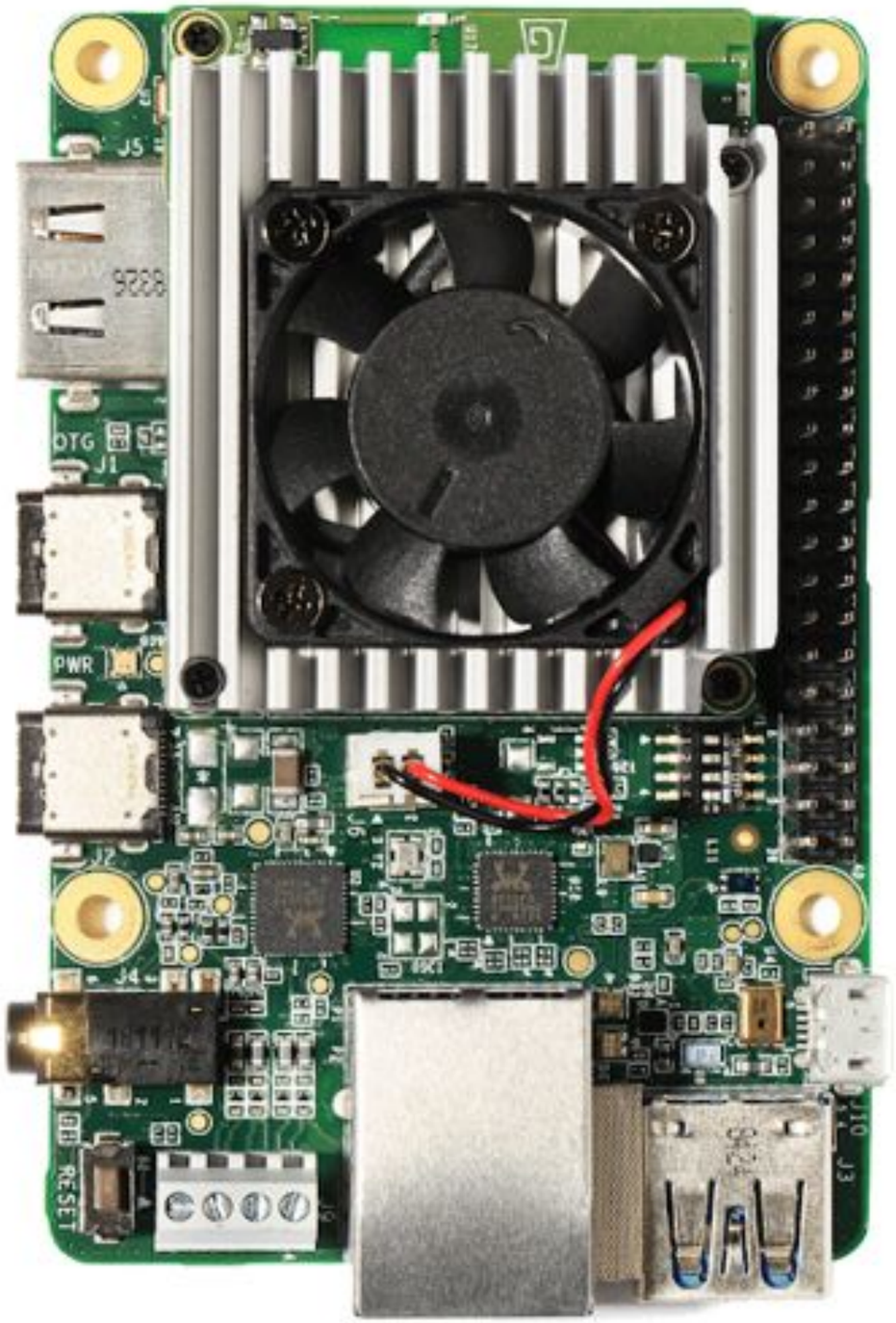
Data Pre-processing & Coral Edge TPU inference

CPU



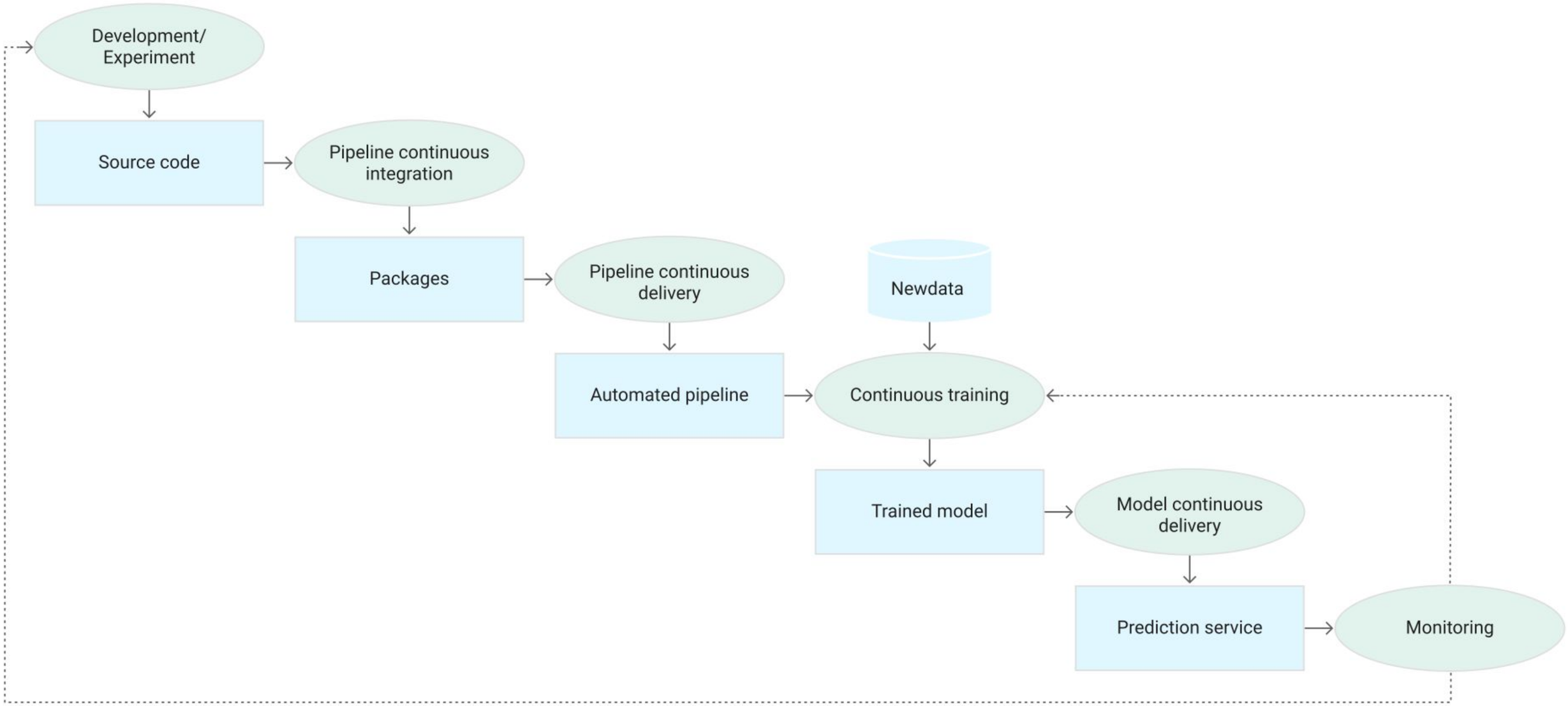
Edge TPU





The model is not static tho..
We need lifecycle management

ML Ops


Development/
Experiment
Source code
Pipeline continuous
integration
Packages
Pipeline continuous
delivery
Newdata
Automated pipeline
Continuous training
Trained model
Model continuous
delivery
Prediction service
Monitoring

Thank You!

Google Cloud

@markkulepisto