tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"Running and Managing Fleets of Single Board Computers at Scale"

Seth Clark – Co-founder & Head of Product
Modzy

May 2, 2023

www.tinyML.org

TINY ML TALKS webcast

Thank you, **tinyML Strategic Partners**, for committing to take tinyML to the next Level, together

TINY ML TALKS webcast

Executive Strategic Partners

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Accelerate Your Edge Compute

SYNTIANT

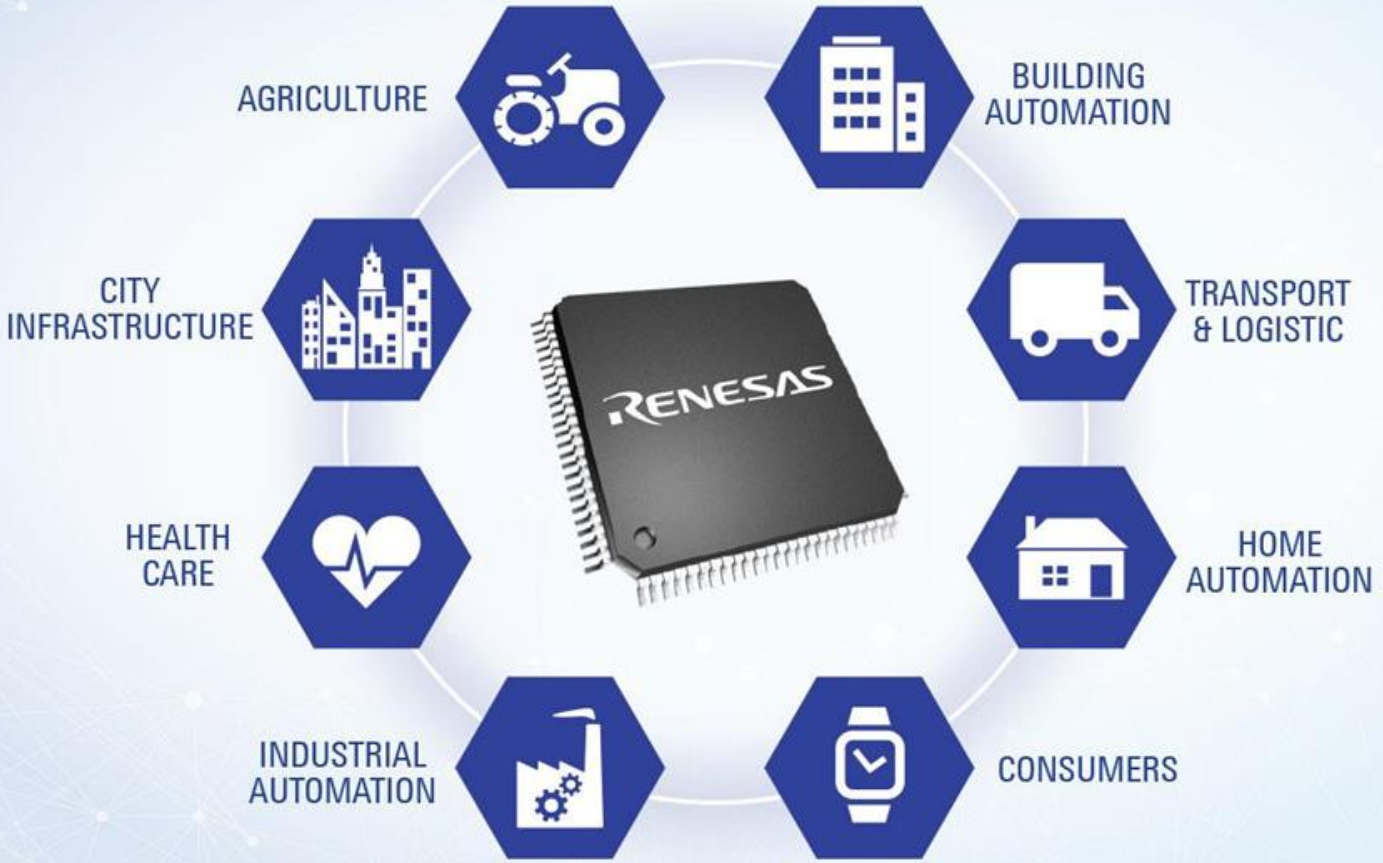
Making Edge AI A Reality

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partners

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.



renesas.com

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

TINY ML TALKS webcast

Gold Strategic Partners

ANALOG DEVICES

AHEAD OF WHAT'S POSSIBLE™

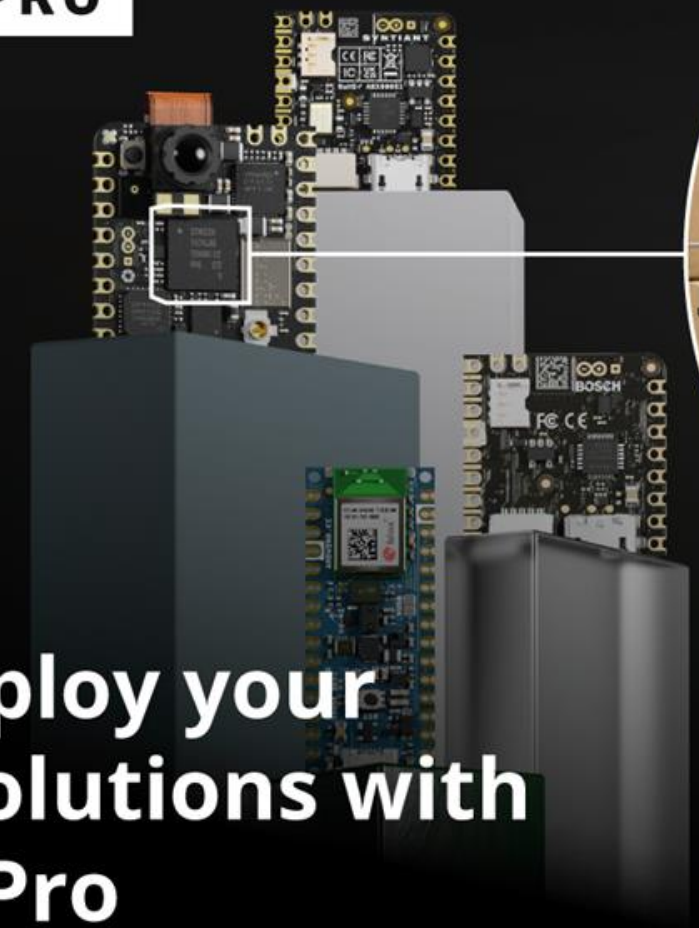
ANALOG DEVICES

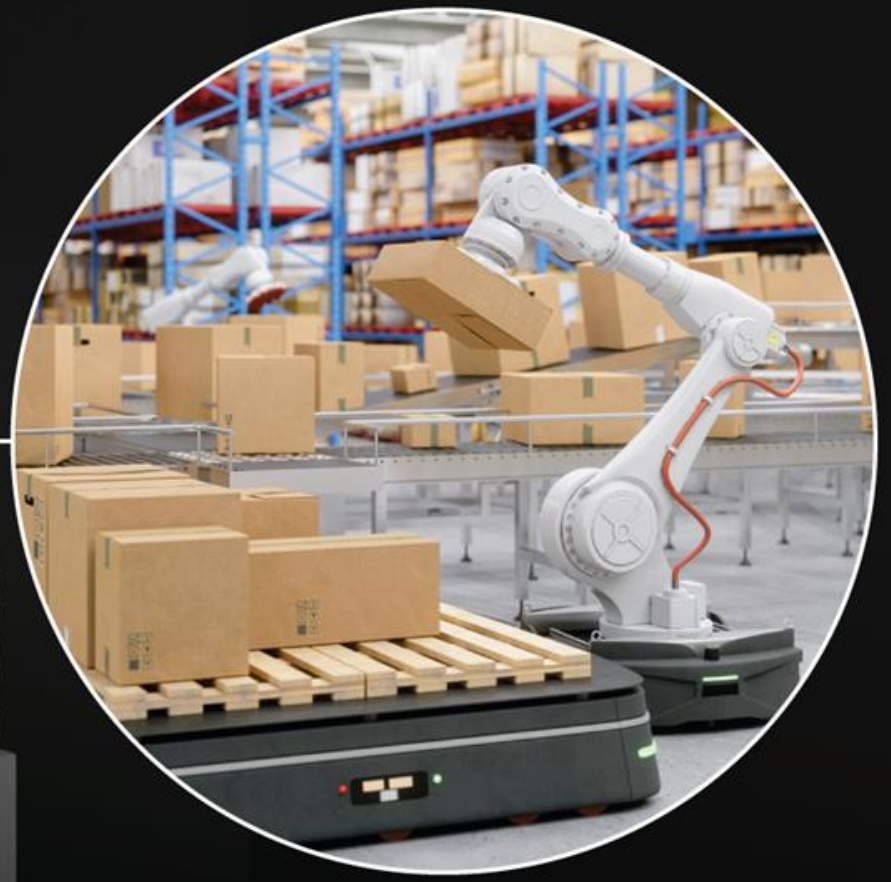
AHEAD OF WHAT'S POSSIBLE™

Where what if becomes what is.

Witness potential made possible at analog.com.

ARDUINO PRO™





Easily deploy your tinyML solutions with Arduino Pro

arduino.cc/pro

Made In Italy

arm AI

Powering tinyML Innovation

Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more:
www.arm.com/techtalks

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com


Microsoft

The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product

SensiML

Analytics Toolkit Suite

sensiml.com/tinyML

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

www.st.com/ai

synaptics

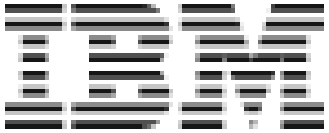
ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

TINY ML TALKS webcast

Silver Strategic Partners



Join Growing tinyML Communities:

14.4k members in
47 Groups in 39 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

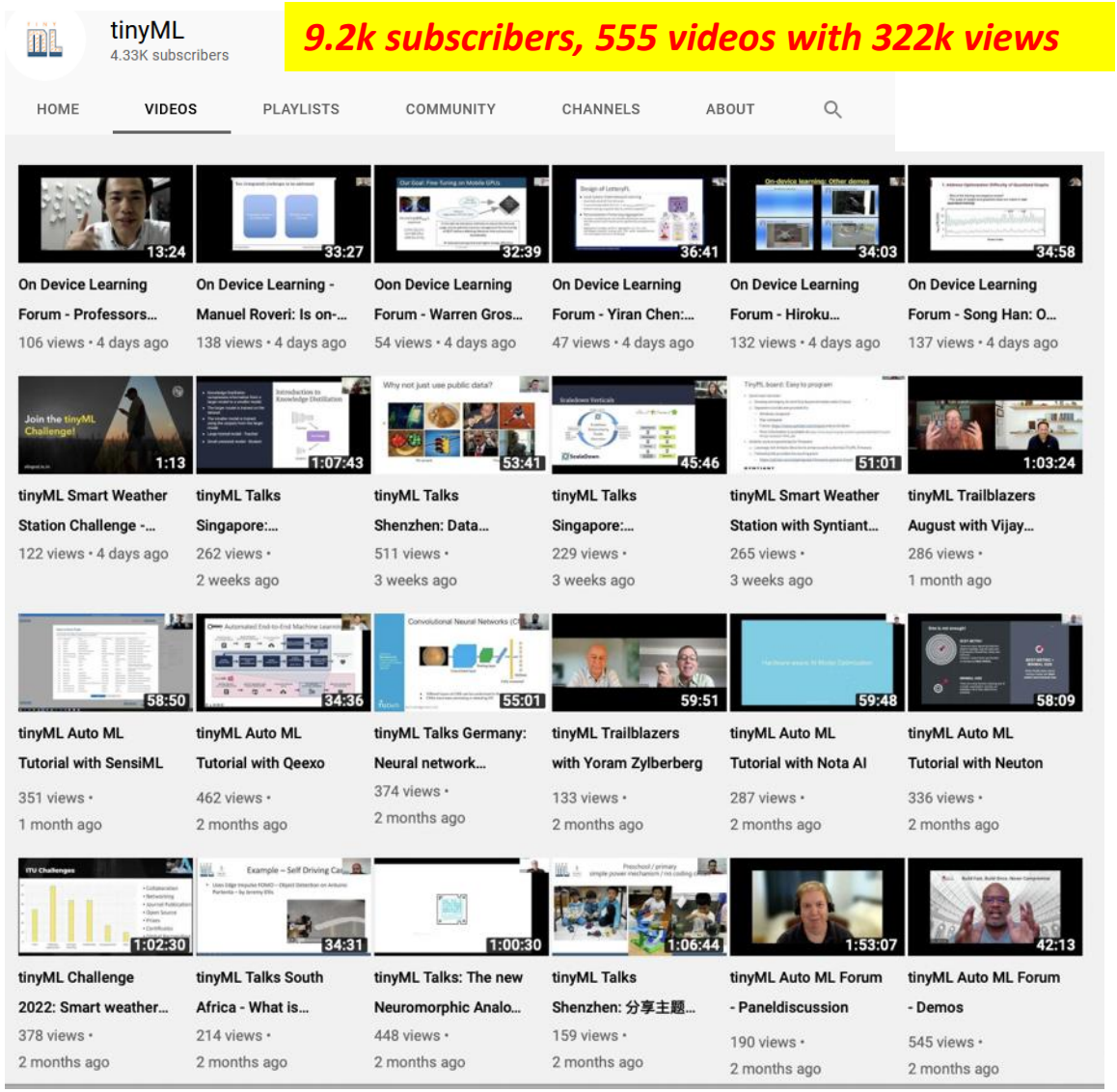
4k members
&
11.6k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

YouTube



Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

www.youtube.com/tinyML

TINY ML TALKS webcast

TINY ML FOUNDATION

EMEA 2023

https://www.tinyml.org/event/emea-2023

More sponsorships are available: sponsorships@tinyML.org

TINY ML TALKS webcast

Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions

TINY ML TALKS webcast

Seth Clark

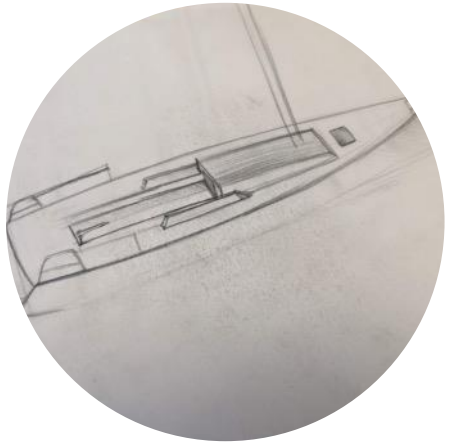
Seth Clark is Head of Product & Co-founder at Modzy. Prior to founding Modzy, Seth served as product manager for a number of successful analytics products. He also served as Principal at Booz Allen Hamilton, where he led complex projects at the intersection of Data Science and Software Development. He has degrees in engineering from the University of Southampton and the Massachusetts Institute of Technology.

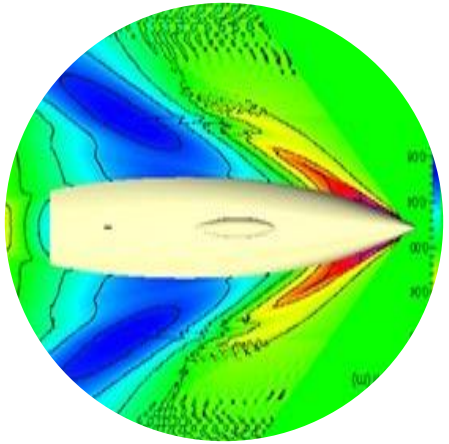
modzy®

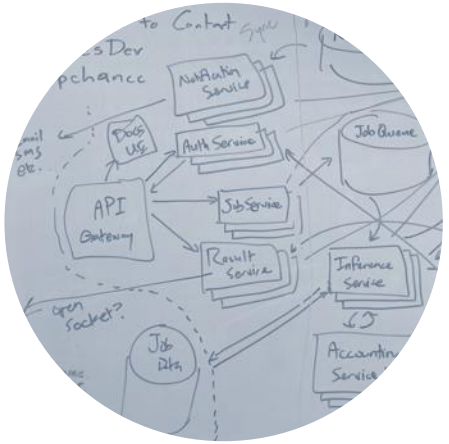
Running and Managing Fleets of Single Board Computers (SBCs) at Scale

tinyML Talks

2 May 2023





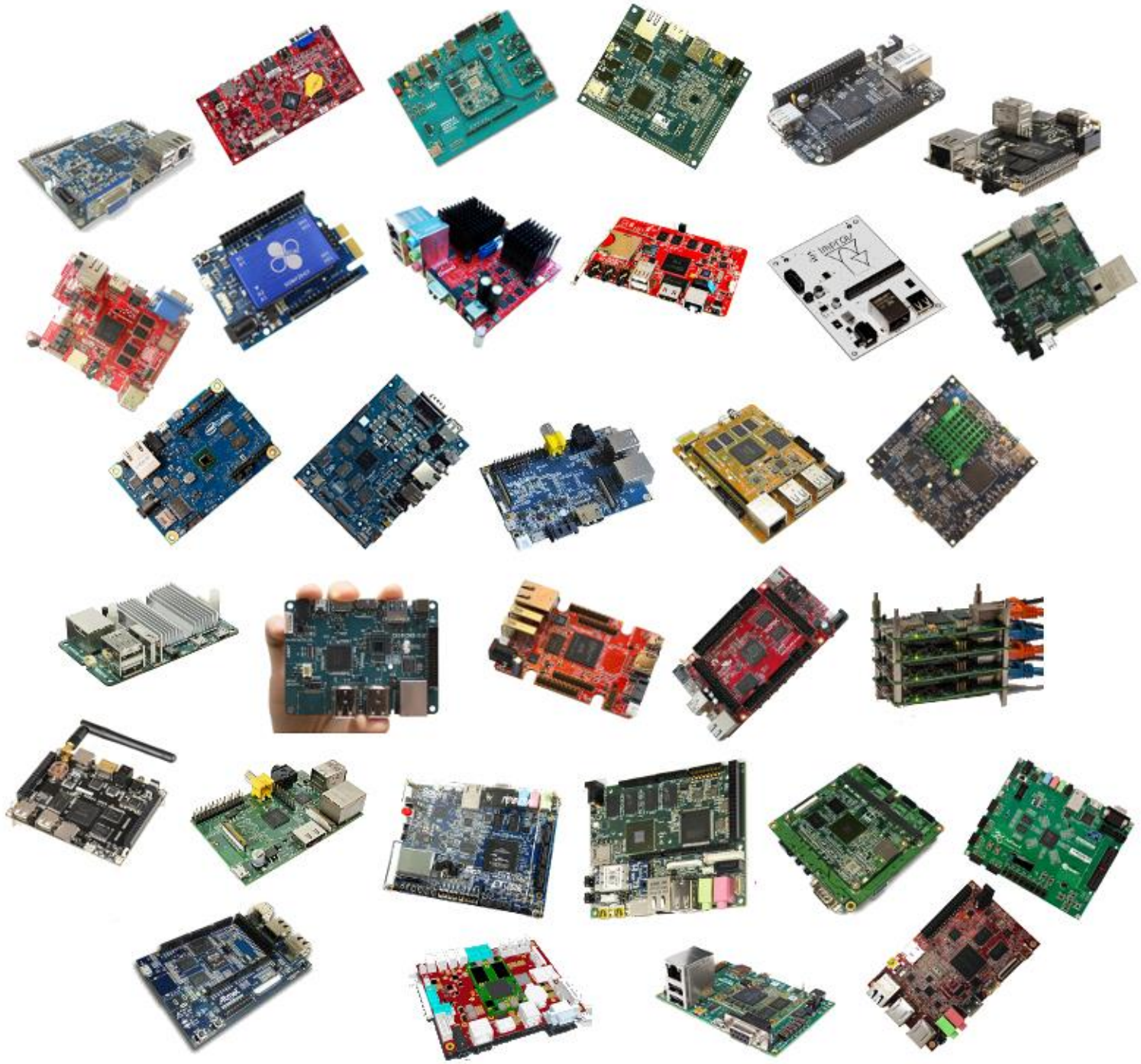


How I got here

modzy®

modzy

2020s: The ~~Wild West~~ Golden Age of SBCs



modzy

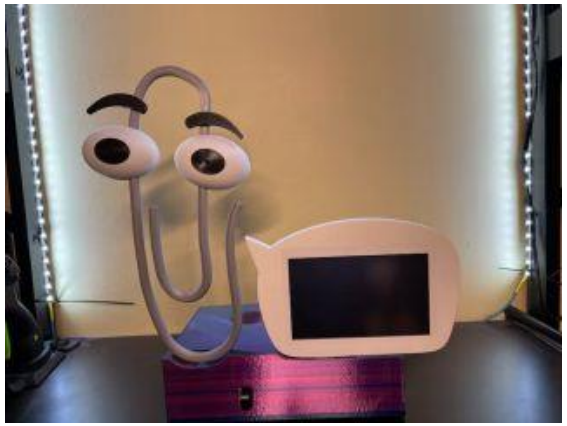
SBCs are doing amazing things

ZeroPhone

Pi Zero-based open-source mobile phone (that you can assemble for 50$ in parts)

Jetson Clean Water AI

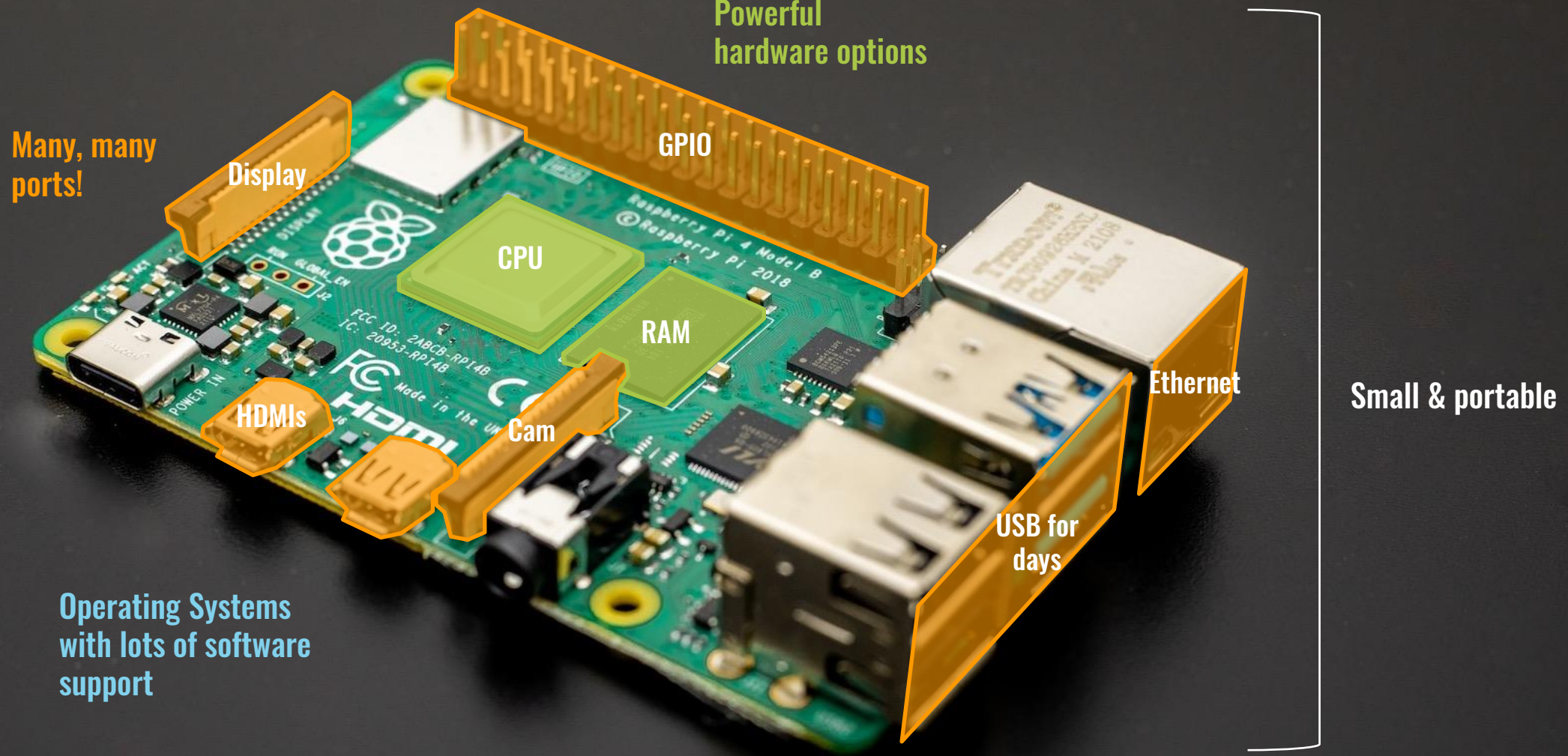
Using AI object detection to detect water contamination



ClippyGPT

Raspberry Pi 3B powered Clippy bot with ChatGPT

Why we love SBCs



SBCs are a convenient alternative to microcontrollers

SBCs	Microcontrollers
Operating System with an optional user interface	No Operating System
Built in networking	No networking
Lots of software options	Limited to software supported by microcontroller’s IDE

Businesses are increasingly adopting SBCs for serious applications

- 3D cameras
- 5G modules
- Robotics
- Smart cities
- Smart manufacturing
- IoT
- Fleet management
- …

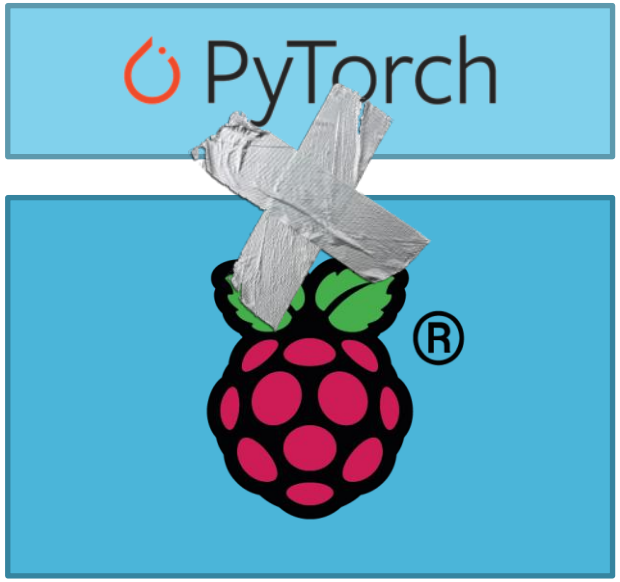
$2.94 billion market in 2022

modzy

SBCs are also great for machine learning

- Write, compile, and run custom software in any language you like
- Download ML frameworks, libraries, and even fully trained models
- Plug in peripherals like cameras and sensors for computer vision and anomaly detection
- Use built in hardware accelerators (e.g. GPUs) or plug one in via USB
- Add an HDMI cable and a monitor and you've got a fully-functional app

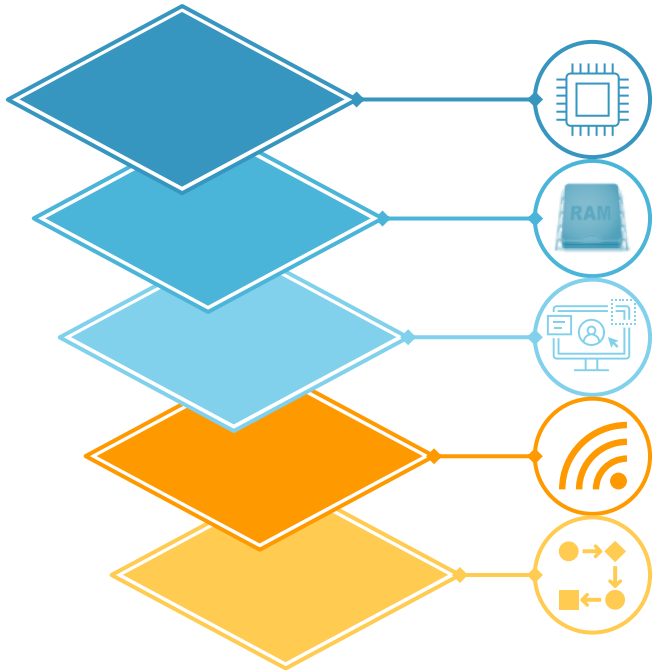
SBCs are great, and so is machine learning, so...



?

modzy

Setting up an SBC takes time and effort



Device Setup
OS installation, system-level dependency configuration, GPU access

Resource Constraints
RAM allocation, number of CPU cores, shared across all app processes

External Software
Custom data connections, monitoring or logging tools

Security
Device and application security, application accessibility, networking

Model Requirements
Programming language(s), model framework dependencies

modzy

Running ML apps brings distinctive challenges

- Minimizing resource needs of models
- Running on multiple chipsets (arm, x86/AMD, 32-bit & 64-bit)
- Managing code for pre-processing data & post-processing predictions
- Working with hardware accelerators (GPUs, TPUs, and FPGAs)
- Connecting to sensors and local data sources
- Gracefully handling intermittent network availability
- Monitoring model performance
- Improving and updating models overtime

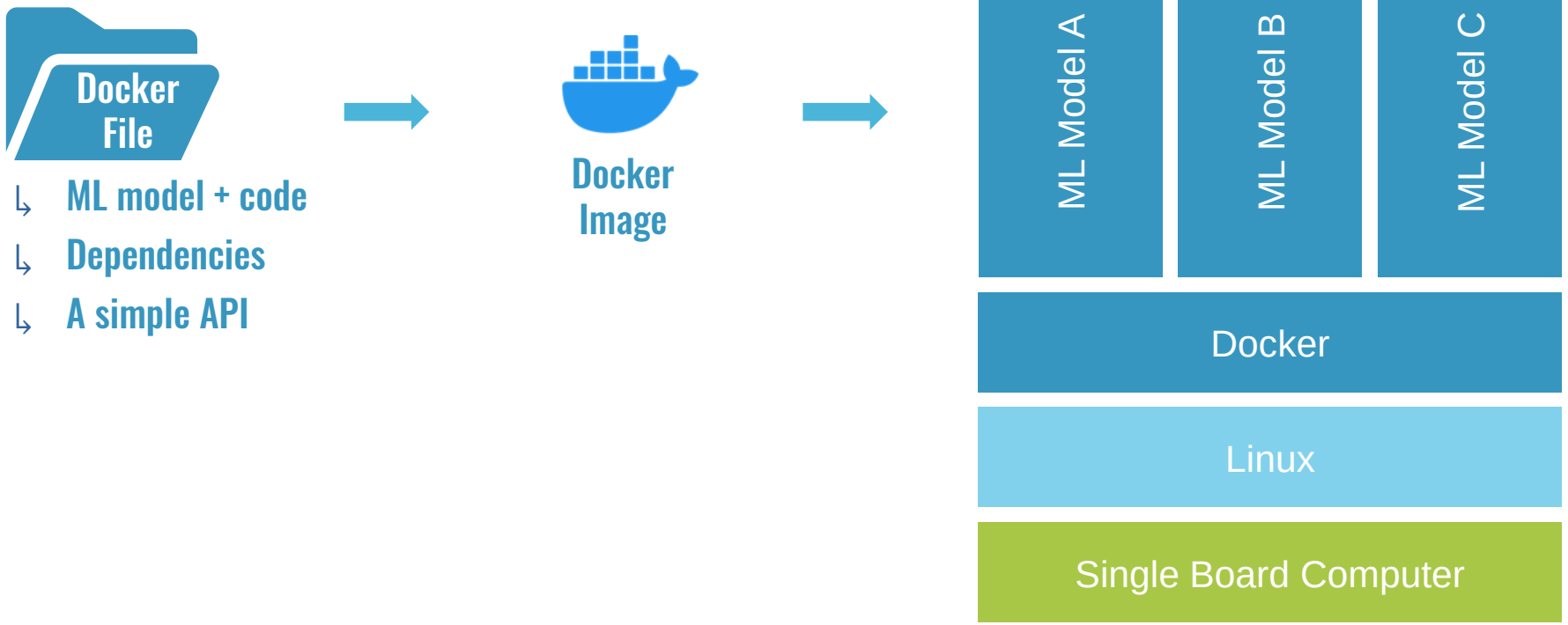
modzy

Running models by hand on 1000s of SBCs would be madness, so what if you could…

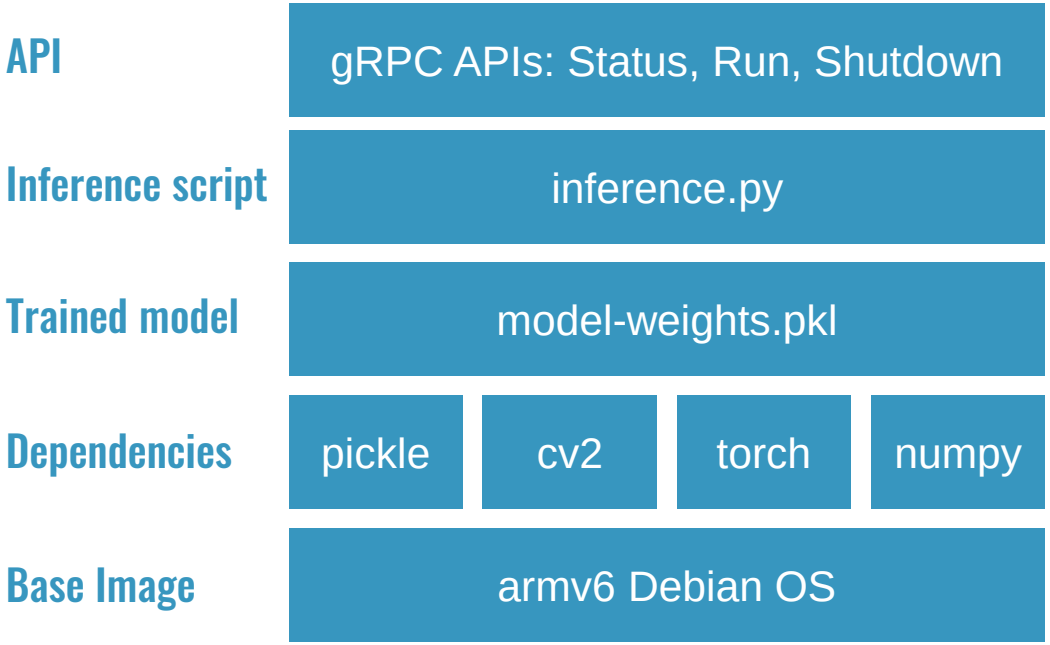
- Send model code and dependencies to an SBC dependably, repeatably, and scalably
- Connect to local data sources and other applications
- Run quickly and use as few resources as possible
- Monitor models while they're running
- Will keep working even when the wifi is down


Enter the container

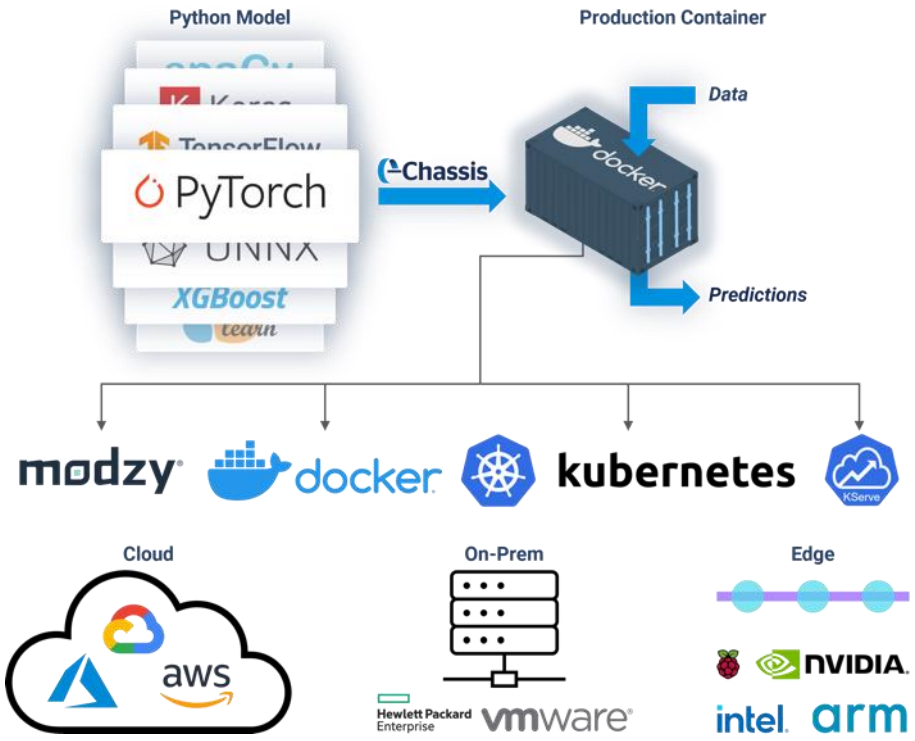
The basic idea



A hypothetical ML container image



Chassis.ml: An easy way to build model containers



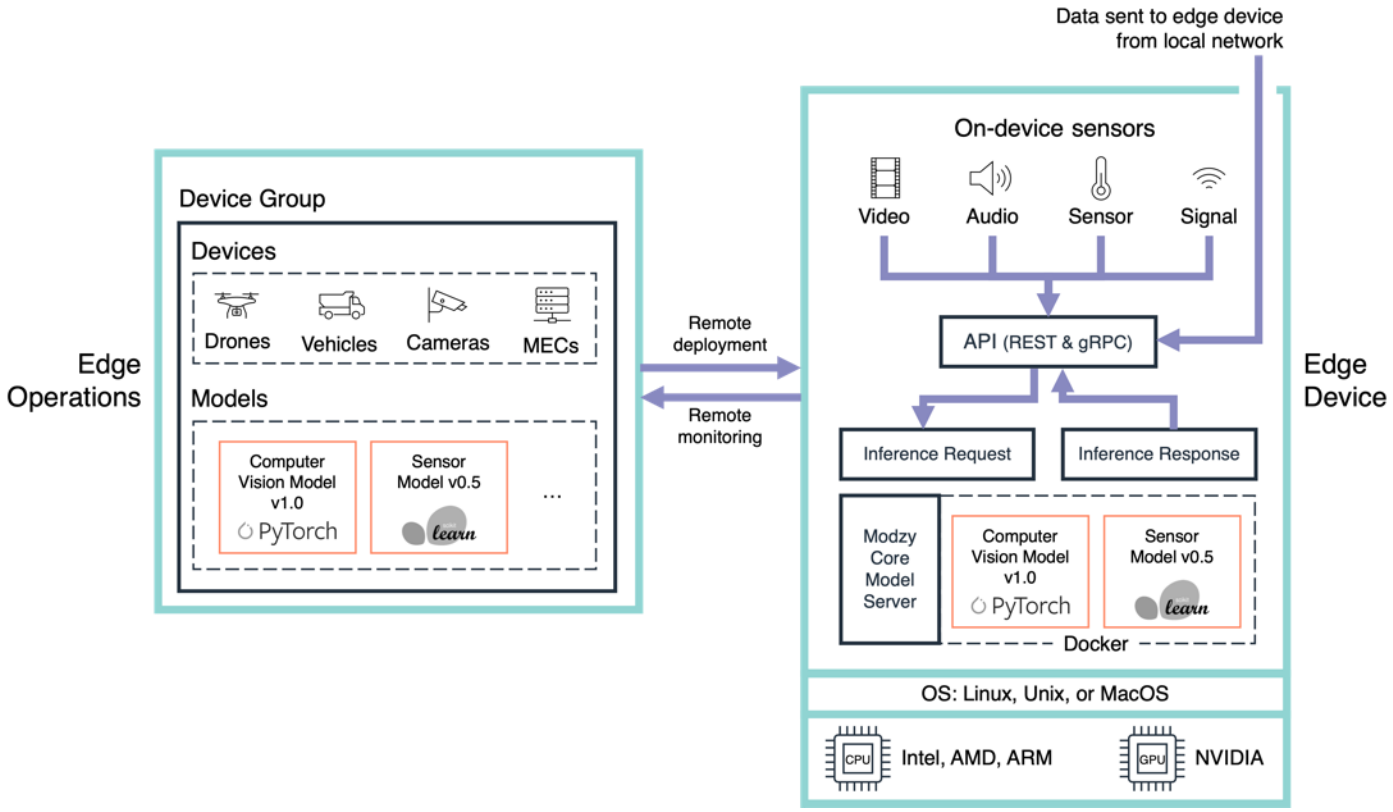
Website: https://chassis.ml
Github: modzy/chassis

modzy

A few more ingredients for managing containers on lots of SBCs

- Chassis.ml containers
- Dockerhub
- Docker
- Raspberry Pi, Jetson Nano, Intel Upboard
- Modzy Core
- Nats.io + Jetstream

How it works


Data sent to edge device from local network
Edge Operations
Device Group
Devices
Drones
Vehicles
Cameras
MECs
Models
Computer Vision Model v1.0
PyTorch
Sensor Model v0.5
learn
...
Remote deployment
Remote monitoring
Edge Device
On-device sensors
Video
Audio
Sensor
Signal
API (REST & gRPC)
Inference Request
Inference Response
Modzy Core Model Server
Computer Vision Model v1.0
PyTorch
Sensor Model v0.5
learn
Docker
OS: Linux, Unix, or MacOS
CPU
Intel, AMD, ARM
GPU
NVIDIA

modzy

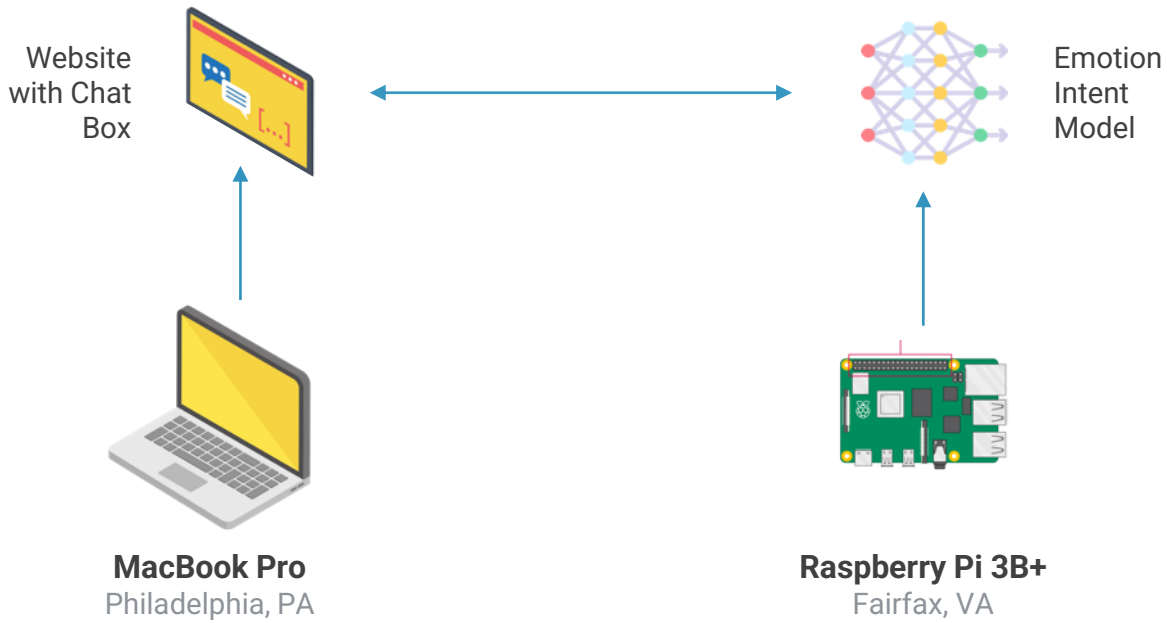
Benefits of container-based ML on SBCs

- Supports both ARM and x86/AMD chipsets, 32 or 64-bit
- Built-in support for GPUs
- Models never break due to missing dependencies
- Network connectivity is only needed initially to deploy models
- Models can be monitored at any time
- New models and versions can be deployed centrally to many, many devices at once

modzy®

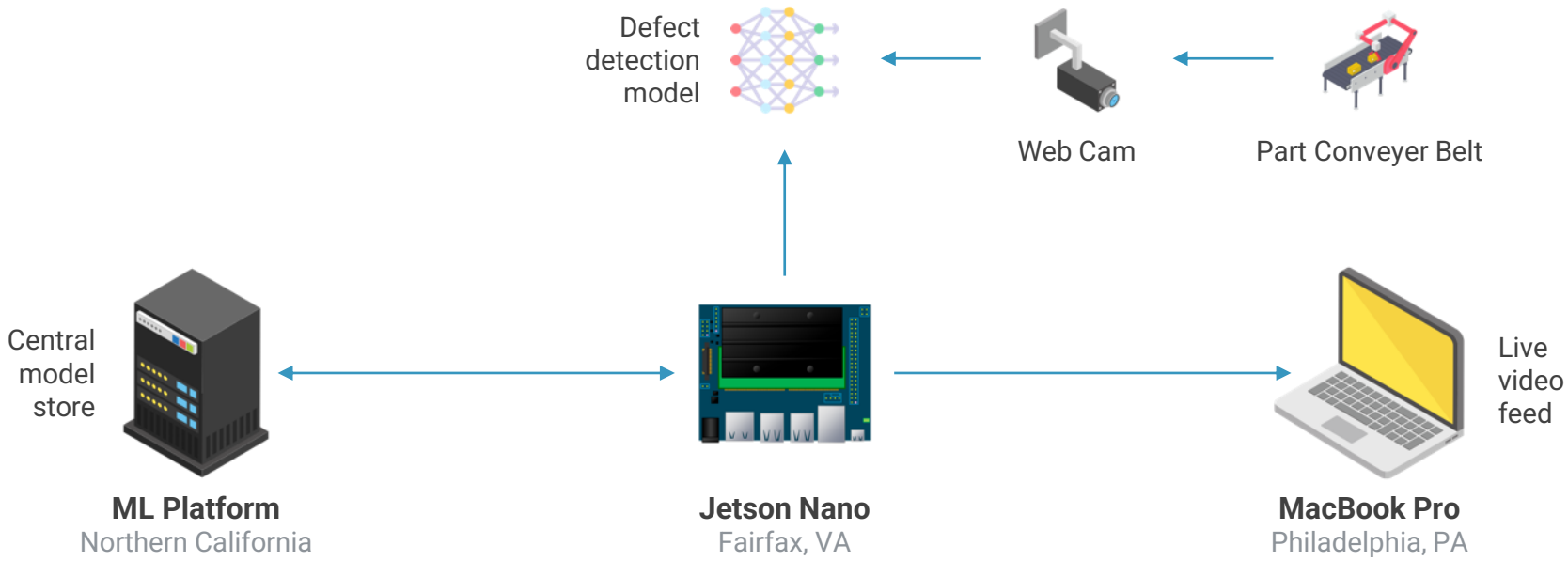
Examples

Example: Edge-centric NLP App



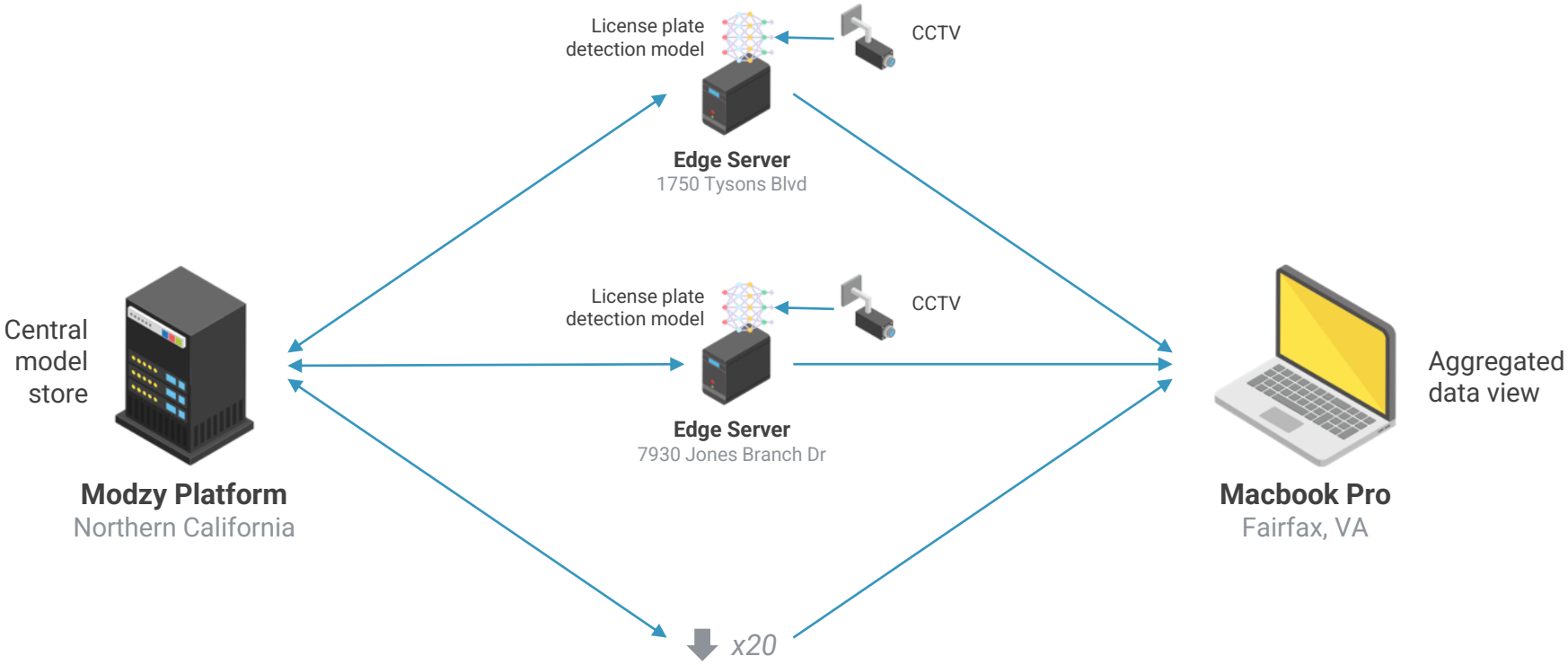
modzy®

Example: Industrial Defect Detection



modzy®

Example: Computer Vision Parking Lots



modzy®

Final thoughts

- Use an edge-centric architecture anytime you need low-latency, non-stop performance on or offline, or models that need to run in more than one place
- Start small, but plan for massive scale

modzy®

info@modzy.com

@getModzy

getModzy

discord.gg

Seth Clark

Thanks for joining our webinar!

Scan the QR code to join our Discord server for more resources.

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org