UNIVERSITY OF TWENTE.

EMBEDDED AI

DSI SEMINAR BY SEBASTIAN BUNDA MSc

WHO AM I?

- PhD Student @ DMB and CAES
  - Subject: Embedded AI
- Background in Electrical Engineering: 2015-2022
  - Specialisation Computer Vision & Biometrics @Twente

UNIVERSITY OF TWENTE.

EMBEDDED AI

ARTIFICIAL INTELLIGENCE
DEFINITION

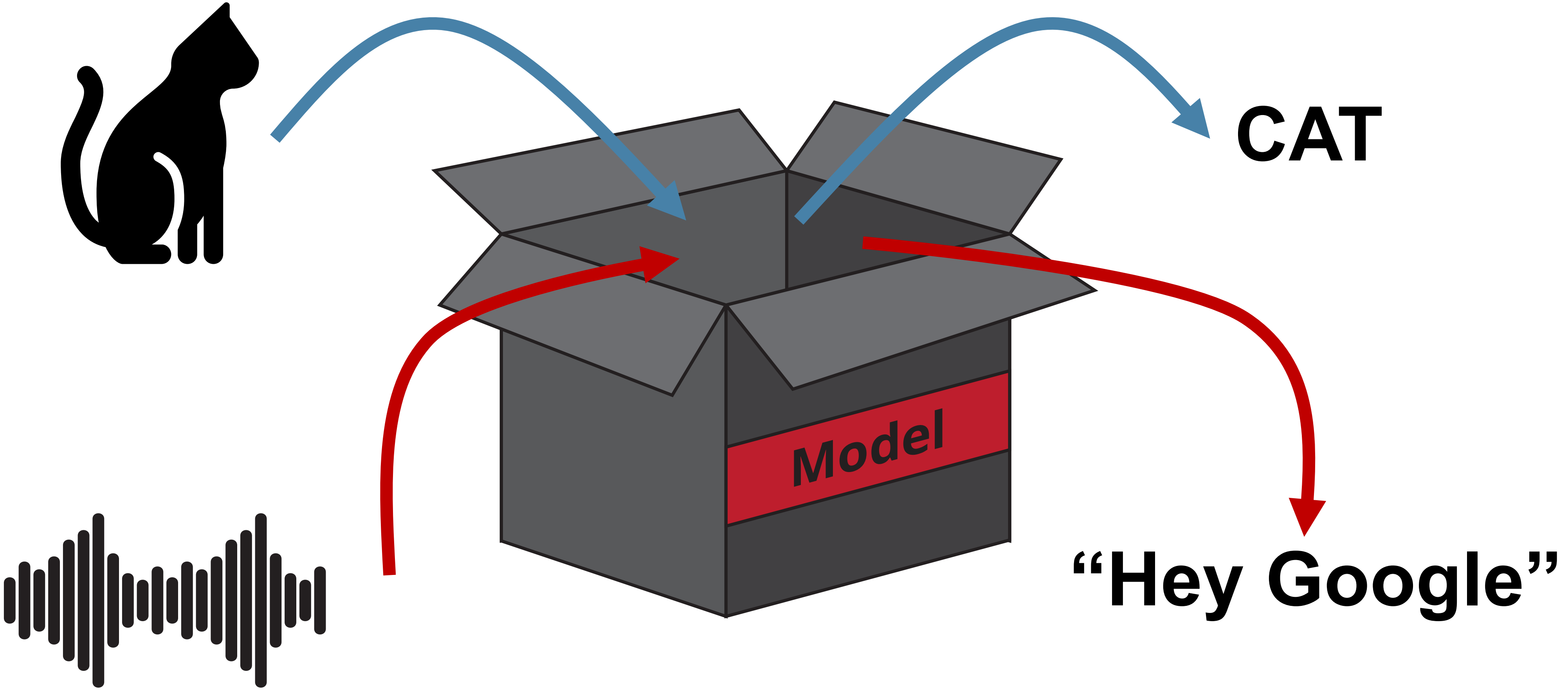
Artificial Intelligence is an umbrella term for various computational strategies able to display **human-like capabilities** such as reasoning and learning

*Examples: **machine learning**, **robotics** and **natural language processing***

Key Concepts

MACHINE LEARNING

DEEP LEARNING MODEL



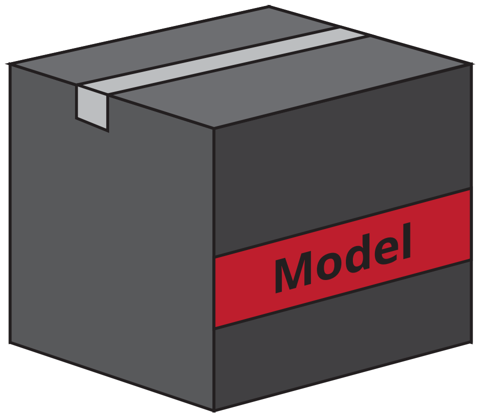
TYPES DEEP NEURAL NETWORKS

Convolutional Neural Network

e.g. Image Classification

Transformer

e.g. Natural Language Processing



Recurrent Neural Network

e.g. Speech Processing

Generative Adversarial Networks

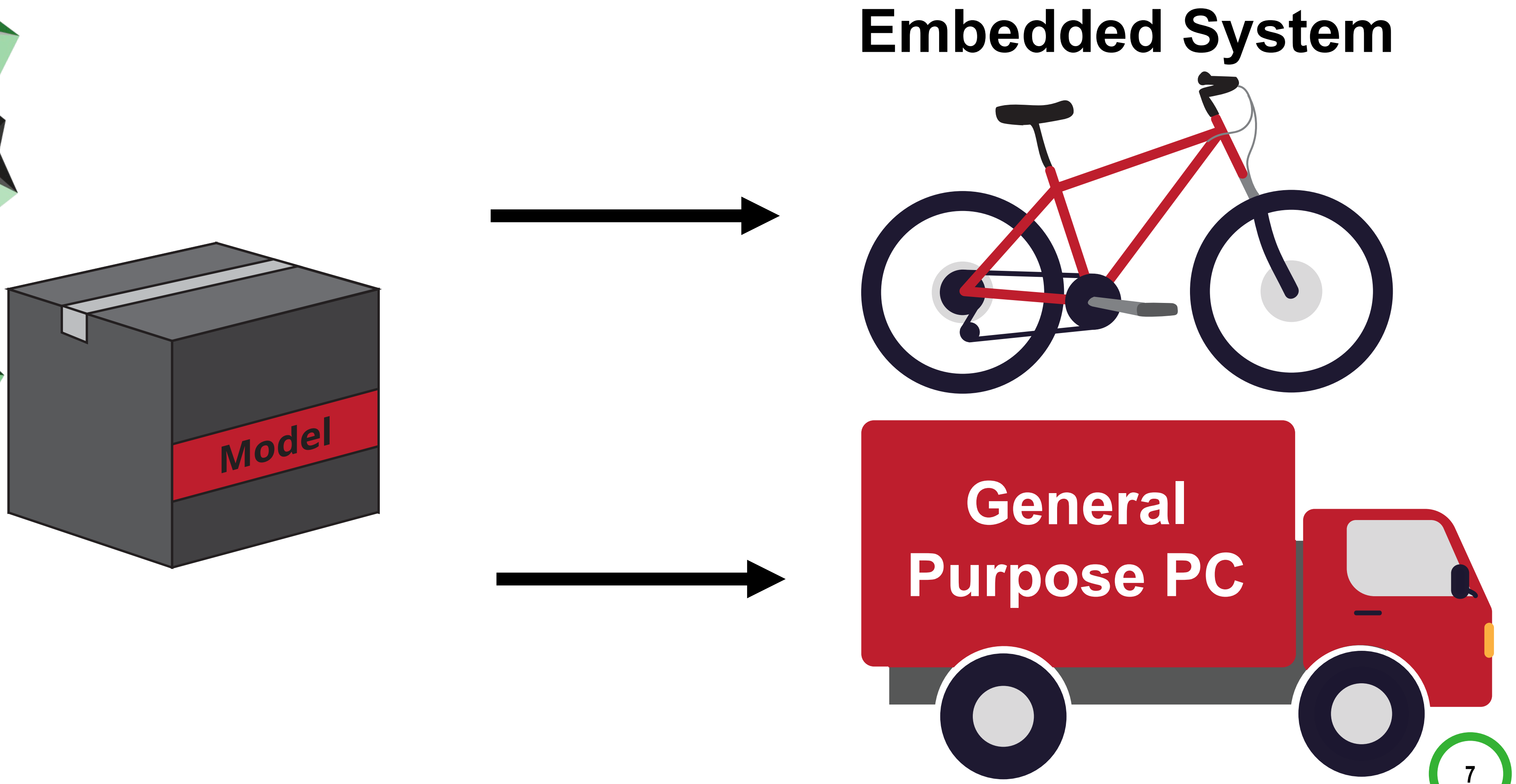
e.g. Synthetic Face Generation

Key Concepts

UNIVERSITY OF TWENTE.

IMPLEMENTING AI ON EMBEDDED SYSTEMS

CHALLENGES

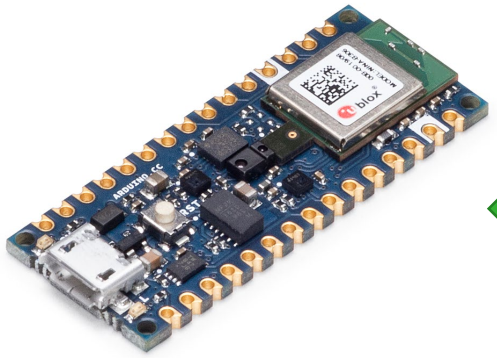


Key Concepts

EMBEDDED SYSTEMS

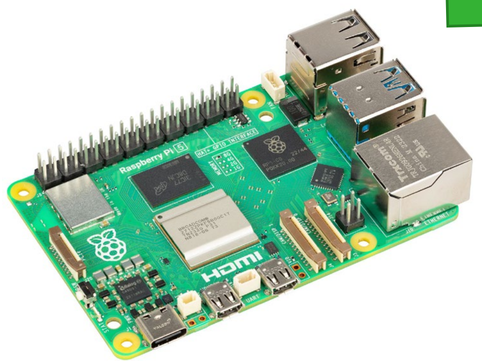
A DEVICE DESIGNED FOR ONE SPECIFIC TASK WITHIN A SYSTEM

Embedded System

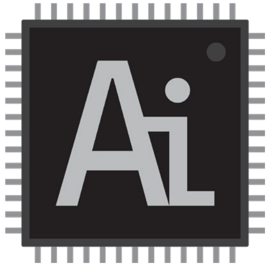


MICROCONTROLLERS

FPGA'S



MICROPROCESSORS



ASIC'S

IMPLEMENTING AI ON EMBEDDED SYSTEMS

COMPARE EMBEDDED SYSTEM WITH PC

Embedded System

- Special Purpose Hardware
- Pre-programmed firmware
- Real-time response and efficiency is key
- Little power consumption
- Cheap
- Local data processing

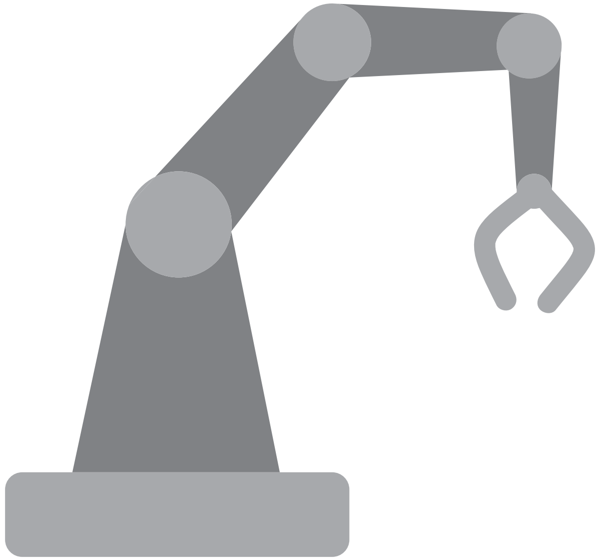
- Generic Hardware
- Programmable by user
- Performance and capacity is key
- High power consumption
- Expensive
- Cloud-based data processing

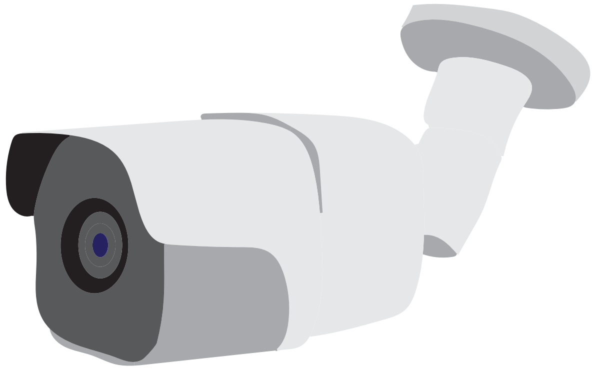
Key Concepts

APPLICATIONS

Personal Healthcare

Predictive Maintenance



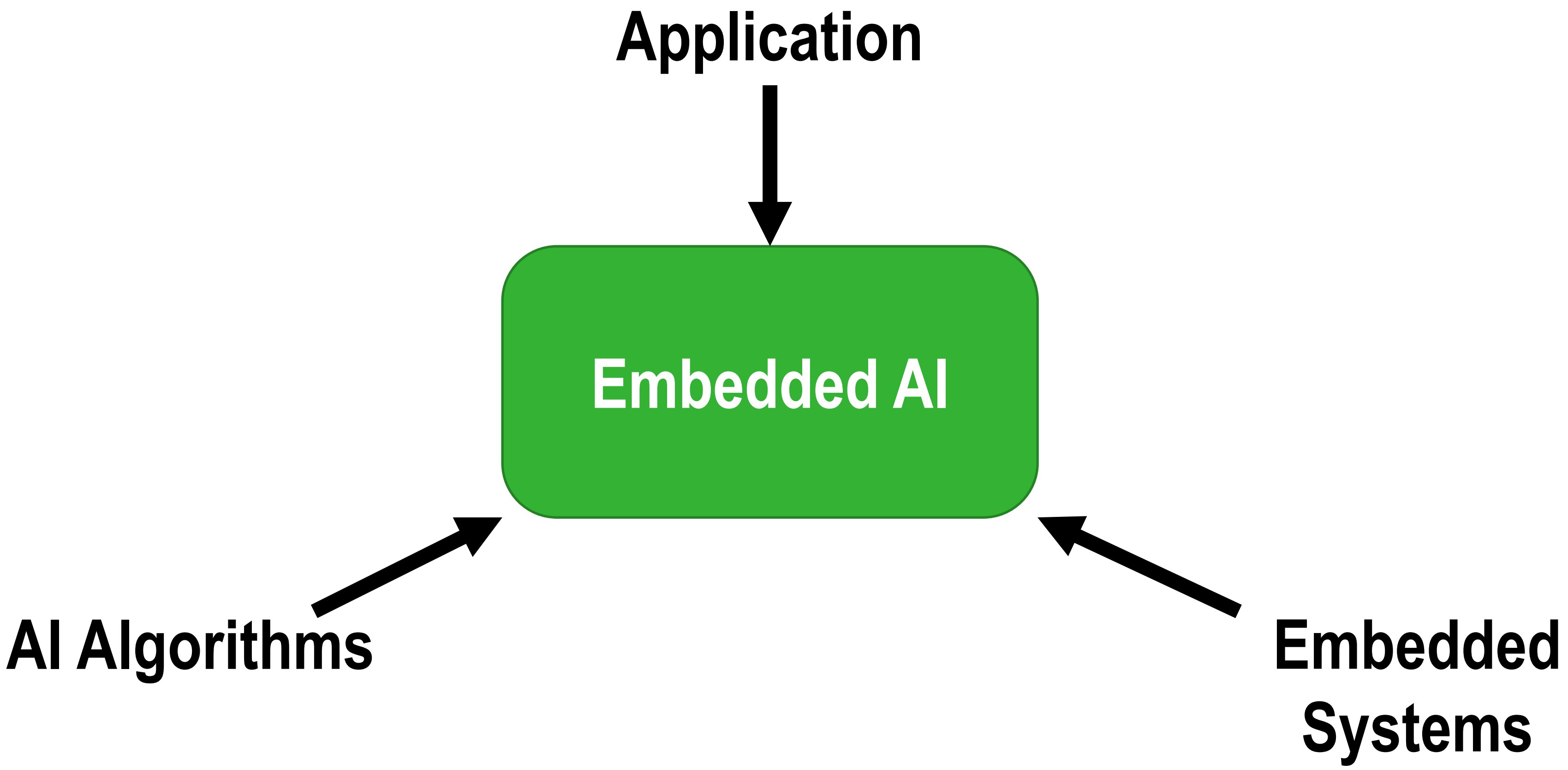


Smart Camera's

Key Concepts

UNIVERSITY OF TWENTE.

EMBEDDED AI



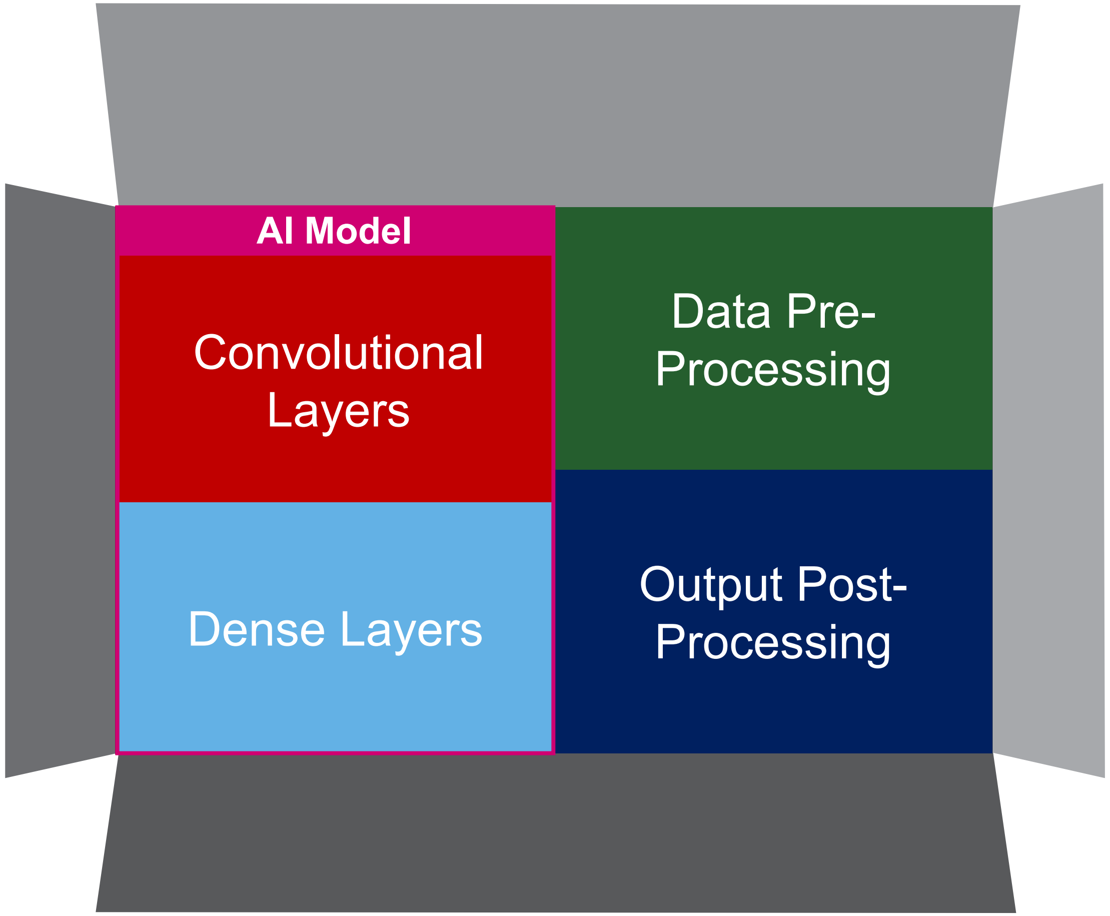
Key Concepts

UNIVERSITY OF TWENTE.

MY RESEARCH

MODEL DESIGN & PERFORMANCE

- AI model contains several different layers

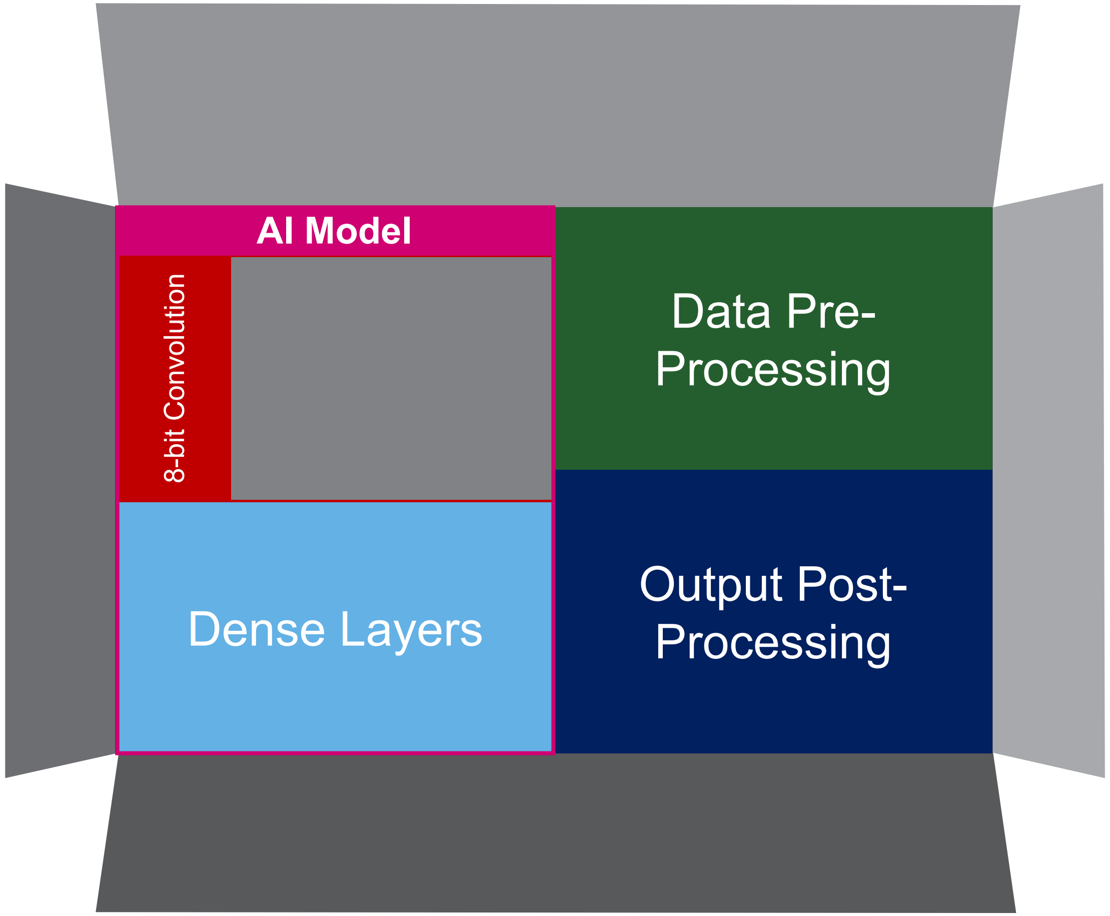


Techniques
Efficient Computation

UNIVERSITY OF TWENTE.

MODEL DESIGN & PERFORMANCE

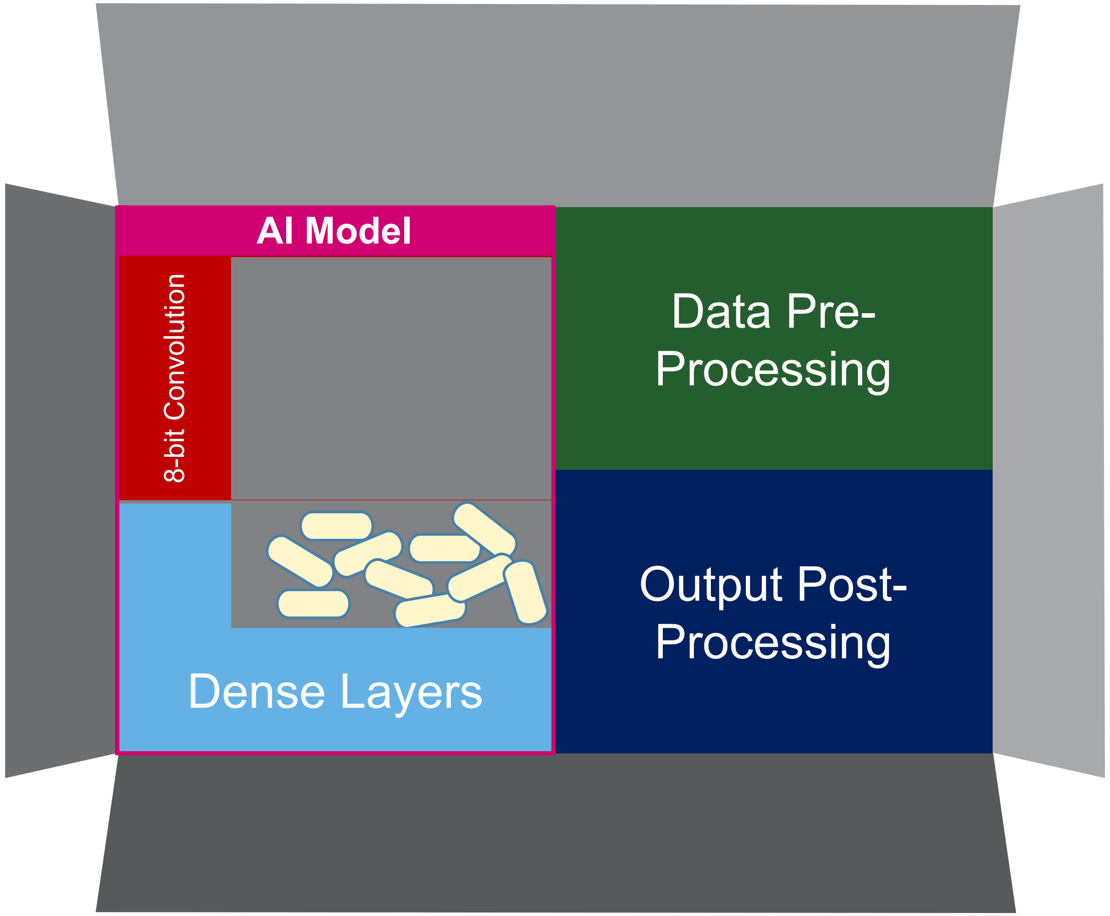
- Quantization
  - Approximating a 32-bit number with an 8-bit number
  - Can reduce footprint by a factor 4



Techniques
Efficient Computation

MODEL DESIGN & PERFORMANCE

- Pruning
  - Removing connections in neural network that do not contribute
  - Like packaging peanuts that fill up space



Techniques

Efficient Computation

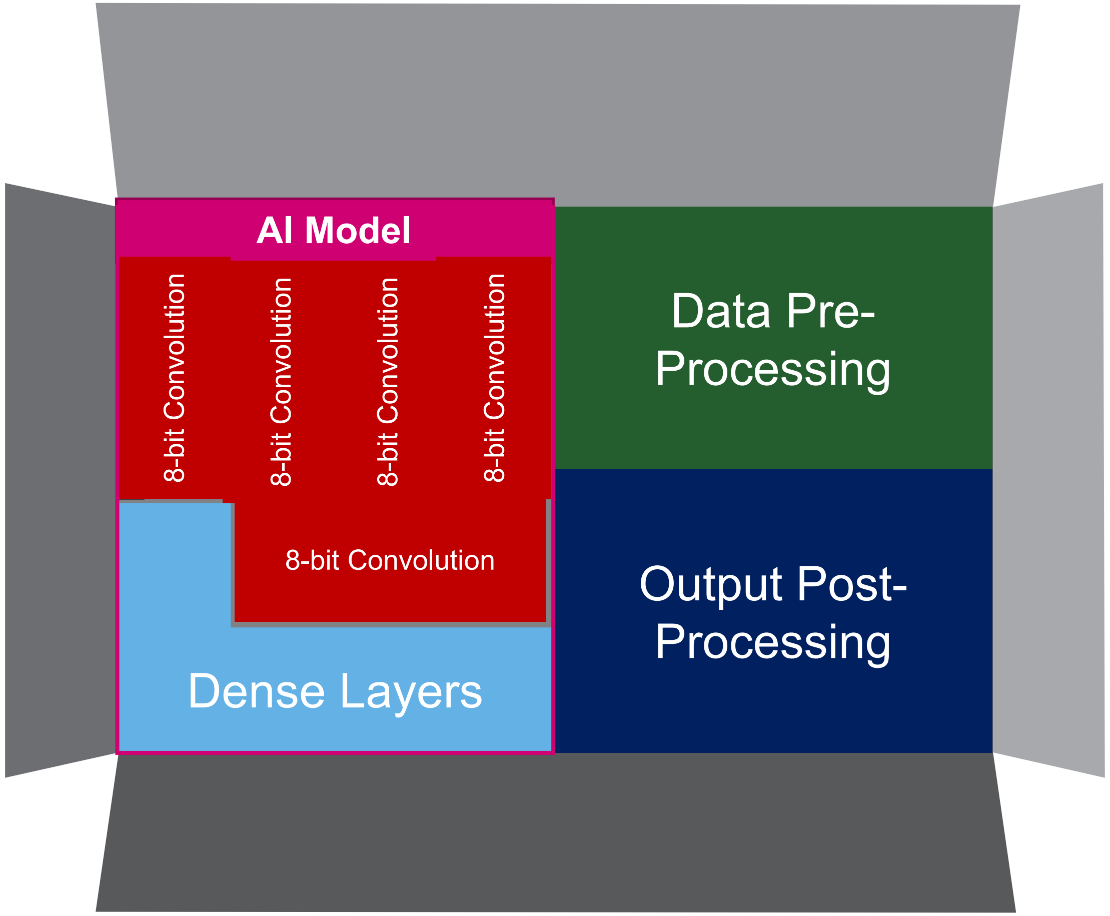
UNIVERSITY OF TWENTE.

EFFECT OF QUANTIZATION ON FACE RECOGNITION

Face Recognition Dataset	32-bit	8-bit	4-bit	2-bit
Post Training Quantization	98.85%	94.65%	63.15%	51.55%
Quantize Aware Training	98.85%	98.68 ± 1.15%	98.63 ± 0.18%	93.45 ± 0.66%

S. Bunda, L. Spreeuwers and C. Zeinstra, "Sub-byte quantization of Mobile Face Recognition Convolutional Neural Networks," 2022 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2022, pp. 1-5, doi: 10.1109/BIOSIG55365.2022.9897025.

OPTIMIZING HARDWARE USAGE



Beware power **consumption** and **latency!**

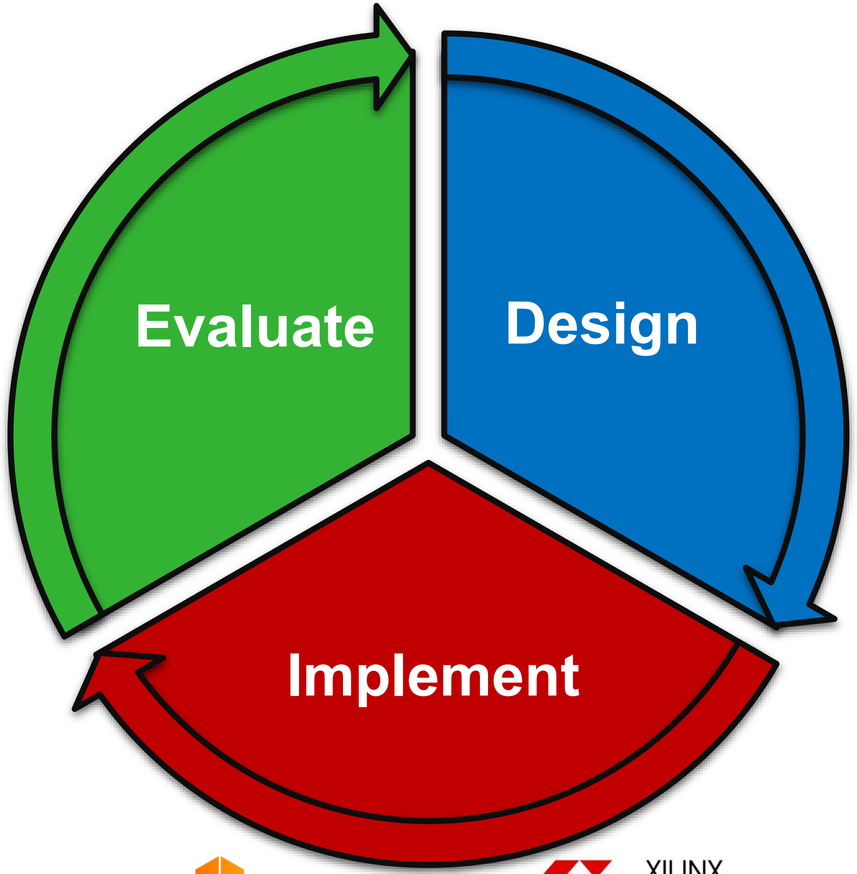
Techniques
Efficient Computation

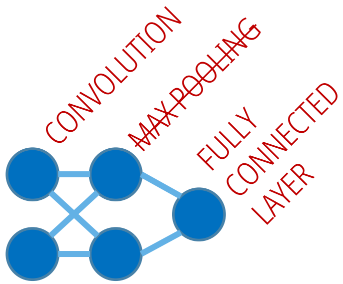
UNIVERSITY OF TWENTE.

MY RESEARCH

	Required	Measured
Accuracy	> 99 %	**98%**
Latency	< 100ms	**89ms**

Our current research and goals





Create/Adapt Model Architecture

TensorFlow Lite

XILINX VITIS AI

UNIVERSITY OF TWENTE.

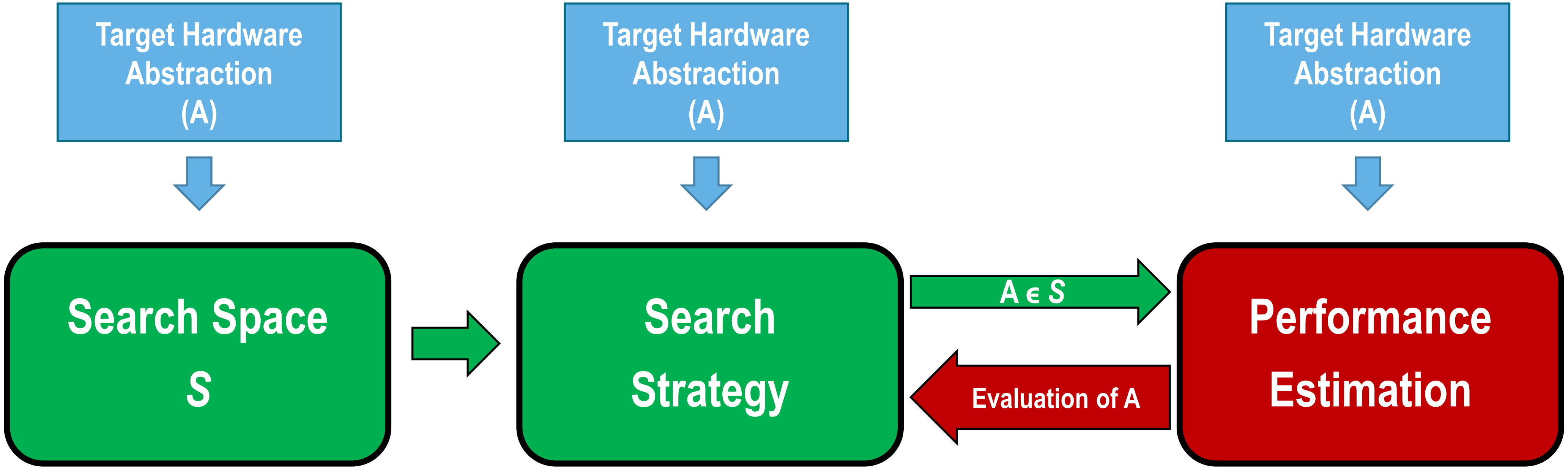
HARDWARE CO-DESIGN NEURAL ARCHITECTURE SEARCH

Optimize design by:

- Proposing models that fit within the **target hardware memory**
- Search for optimal **implementation**
- Evaluate based on e.g. **accuracy**, **latency** and **energy**

Our current research and goals

HARDWARE CO-DESIGN NEURAL ARCHITECTURE SEARCH



Our current research and goals

UNIVERSITY OF TWENTE.

ACTIVE PROJECTS

SOME ACTIVE PROJECTS

- Image classification optimization using FINN-aware neural architecture search
- Vision-based object distribution detection using YOLO and Raspberry Pi
- Efficient Transformer Networks by researching 8-bit Arithmetics for Transformers
- Hardware acceleration for Genetic Selective Sweep Detection using modern technologies

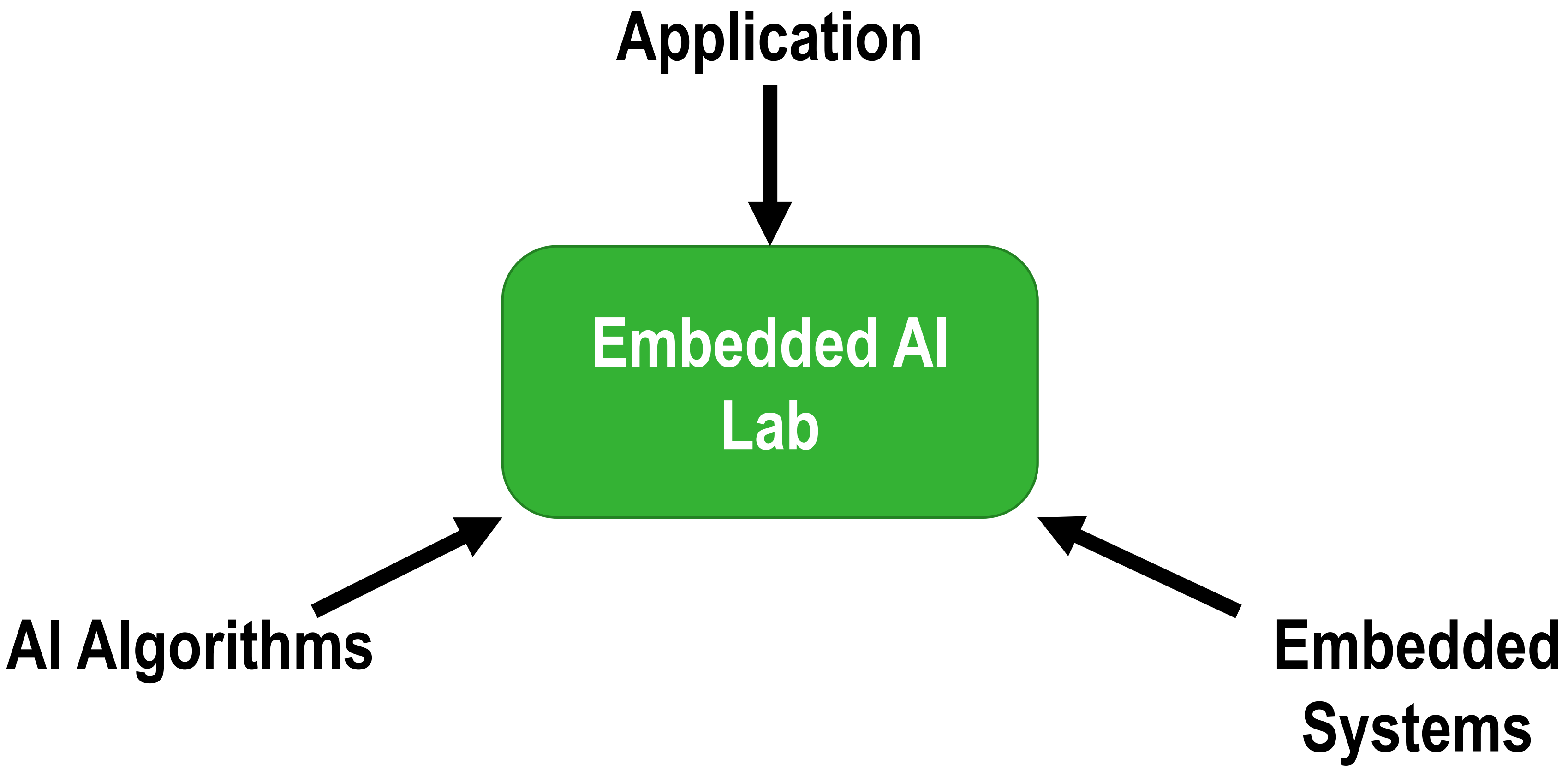
Our current research and goals

OUTLOOK OF EMBEDDED AI

- Expect new efficient AI-acceleration hardware
  - AMD Ryzen AI NPU
  - Neuromorphic Processors
- Leverage the knowledge of the domain of AI and Embedded Systems for Specific Applications

Our current research and goals

EMBEDDED AI LAB



OBJECTIVES EMBEDDED AI LAB

- **Objective 1**: Create an Embedded AI Community at UT
  - Creating new contacts and collaborations within the University of Twente
- **Objective 2**: Inspire using DSI Embedded AI Seminars
  - Everyone is welcome to present their work and start discussions
- **Objective 3**: Stimulate student projects through collaboration and teaching
  - Co-supervise students on Embedded AI topics
  - Teaching the basics through Master course

PLANNING

9th of October

Mario Ruiz from AMD on their new Neural Processing Units (NPUs)

November

Qing Wang on his research on Embedded AI in TU Delft

December

Talk on Neuromorphic Computing

EMBEDDED AI LAB

- Simulate **collaboration** on the topic of **Embedded AI** within University of Twente and with industry
- Develop **AI** methods to achieve State-of-the-Art performance with **limited resources**
- Build **Demonstrators** to show applications

SIGN UP FOR NEXT SEMINAR 9TH OF OCTOBER

QUESTIONS?

EXTRA SLIDES

ARTIFICIAL INTELLIGENCE

ARTIFICIAL INTELLIGENCE

A system that is able to display human-like capabilities such as reasoning and learning

Natural Language Processing

Robotics

Automatic Programming

MACHINE LEARNING

Algorithms whose performance improves by being exposed to more data

Random Forest

Support Vector Machines

Principle Component Analysis

Decision Trees

Logistic Regression

k-means Clustering

DEEP NEURAL NETWORKS

ML algorithm based on multi-layered neural networks

Convolutional Neural Networks

Transformer Networks

Recurrent Neural Network

TENSORFLOW LITE & TINYML

COMPILATION TOOLING

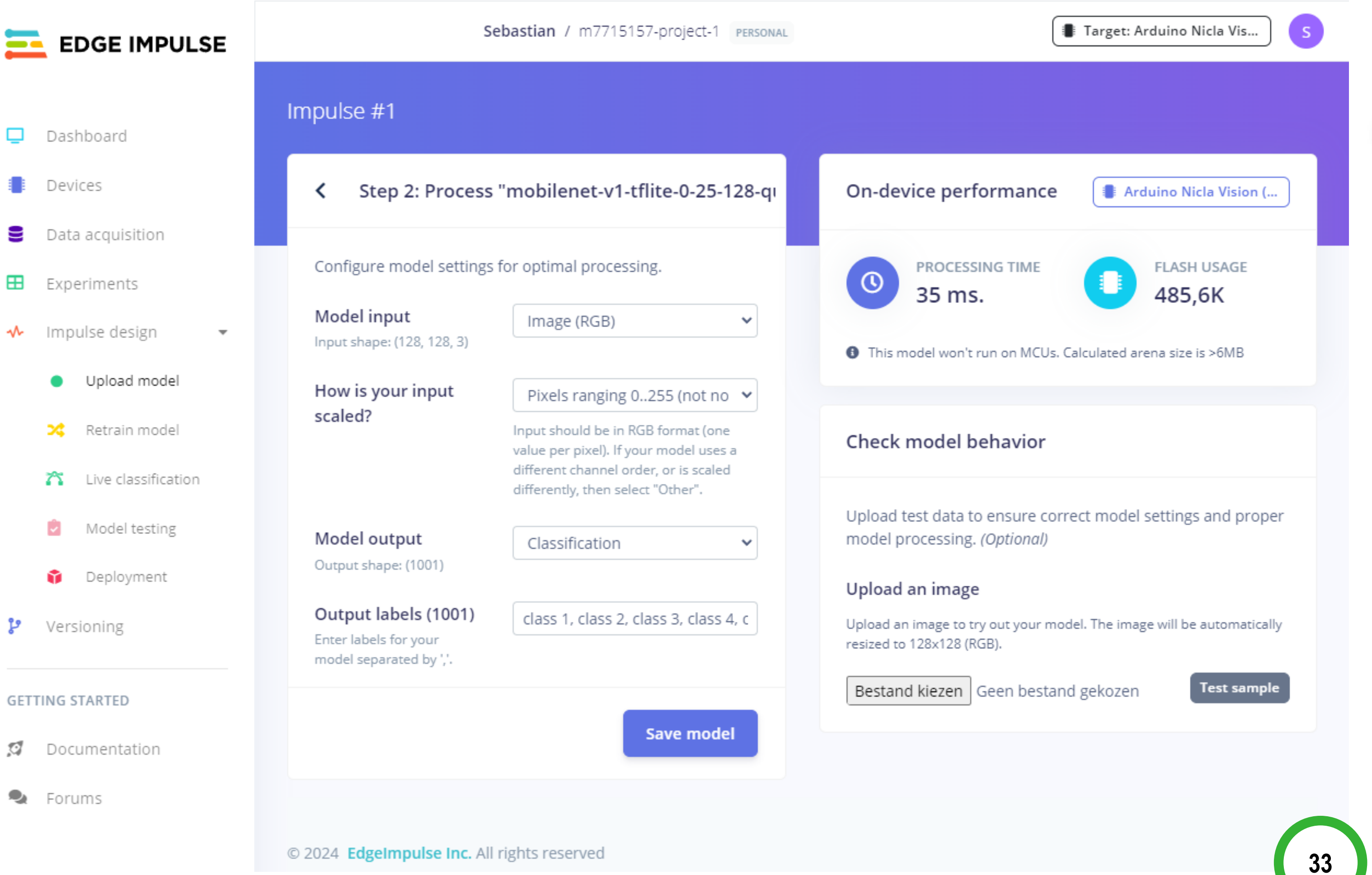
- tensorflow.org/lite/microcontrollers
- tinyurl.com/tinyml-book

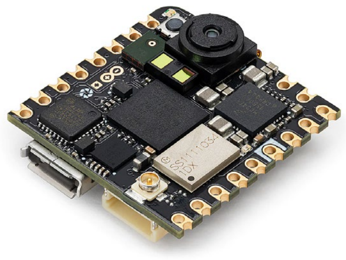
Overview of Available Tools

UNIVERSITY OF TWENTE.

EDGE IMPULSE

DEPLOYMENT TOOLING

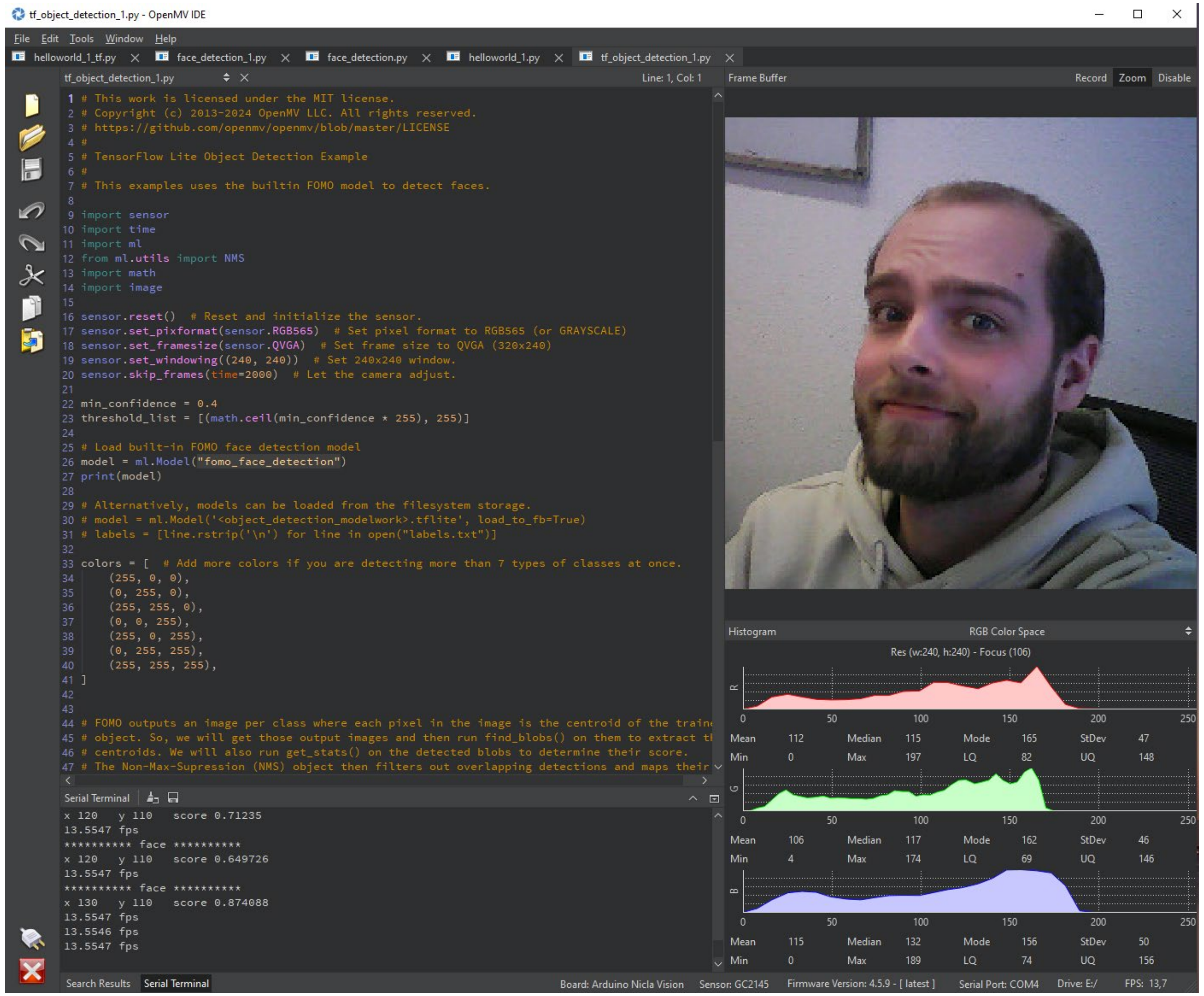


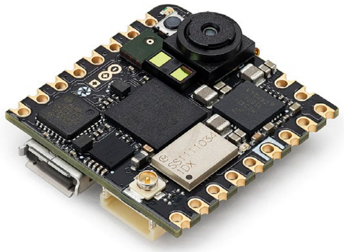


Overview of Available Tools

MICROPYTHON IN OPENMV

DEPLOYMENT TOOLING



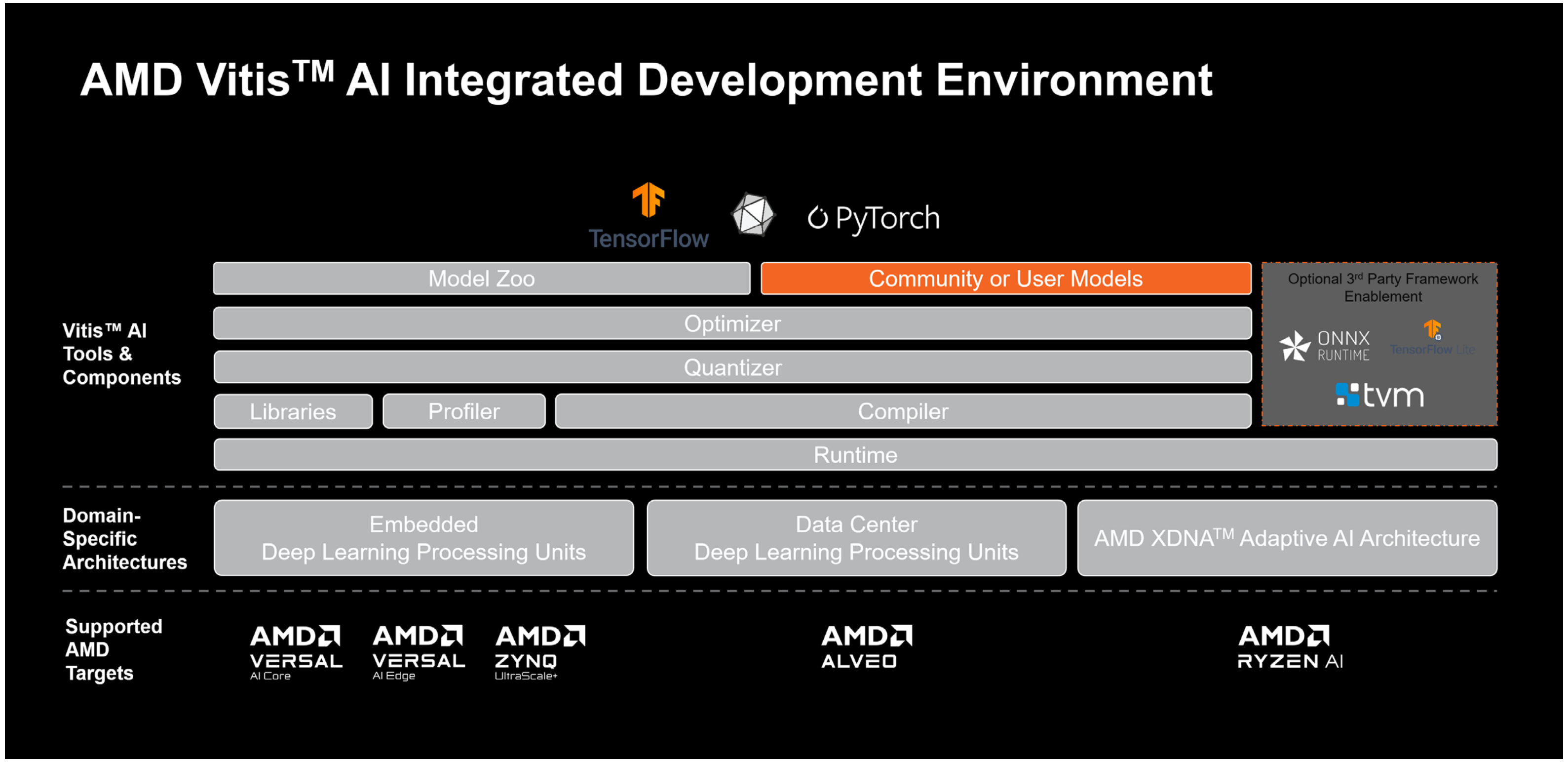


Overview of Available Tools

UNIVERSITY OF TWENTE.

VITIS AI & FINN

DEPLOYMENT TOOLING



https://xilinx.github.io/Vitis-AI/3.5/html/index.html
https://xilinx.github.io/finn/

Overview of Available Tools

STM32CUBE.AI

- Optimize and deploy Deep Neural Networks on STM32 microcontrollers:
  - https://stm32ai.st.com/stm32-cube-ai/