Small is the New Big: Designing Compact AI Models for Edge Devices

GOTO Chicago
April 28th, 2020
Davis Sawyer

Brief Background: Why Deeplite?

Takeda ONCOLOGY → TANDEMLAUNCH → Deeplite AI-DRIVEN OPTIMIZER

2015-2017

2017

2018-now

How do we bring the promise of AI models to benefit daily life?

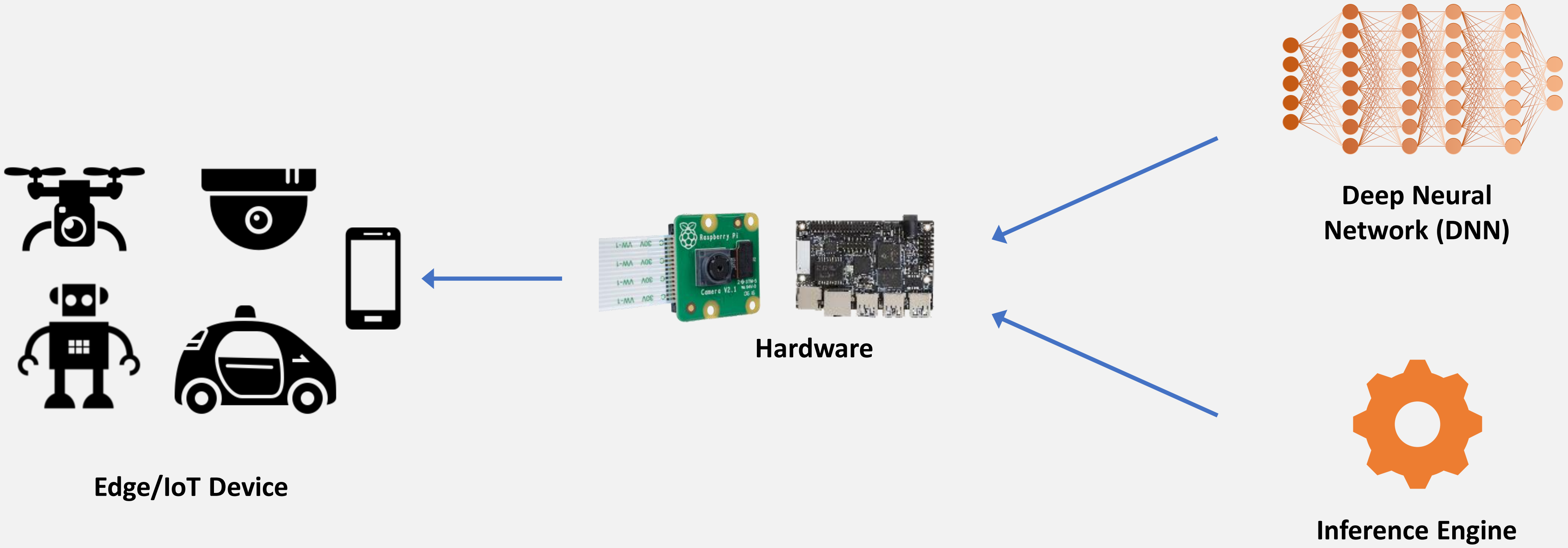
Connected & Autonomous Vehicles

Life-critical Medical Devices

Robotics & Industrial Automation

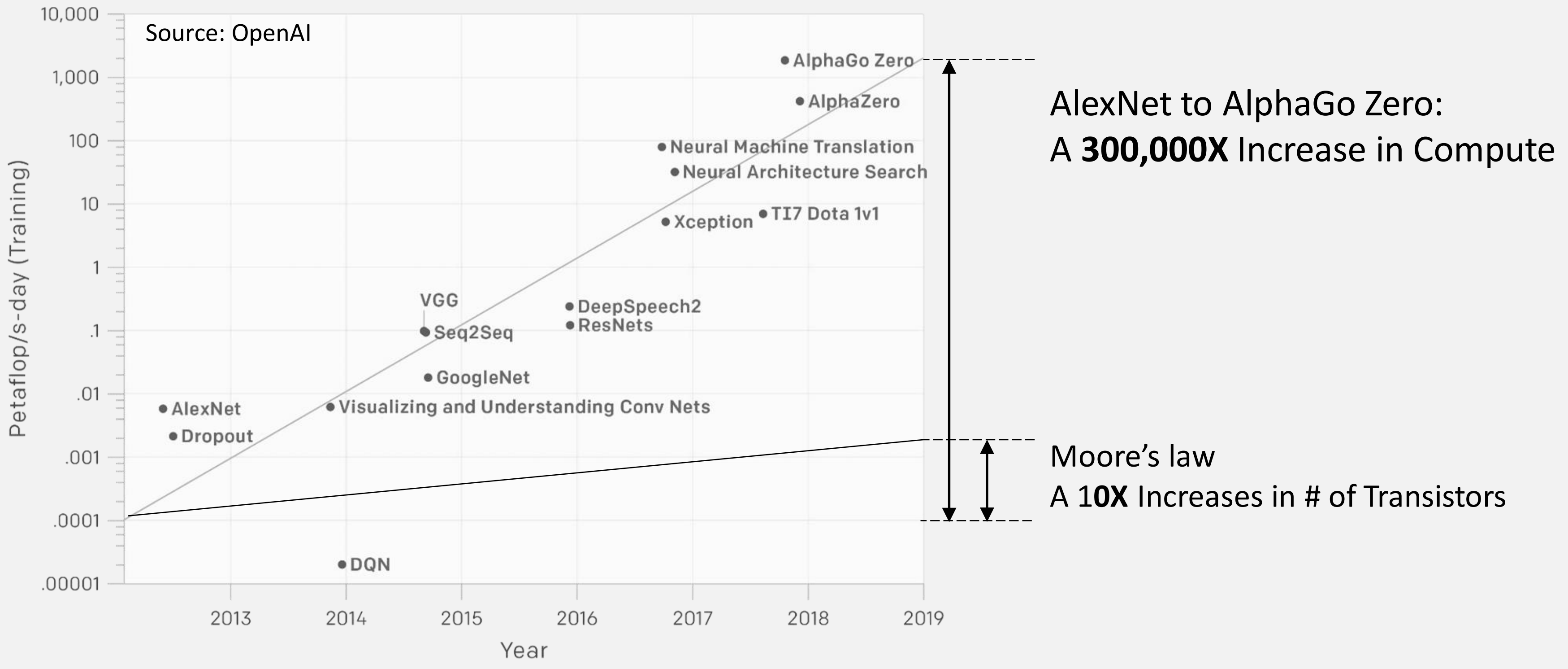
Drones, IoT & Surveillance

Embedded AI 101



Deep learning models are growing rapidly

- Deep learning outperforms humans, but comes with **huge compute cost**
- **Deeper** neural network, **better** accuracy, **more** compute required



These demands force AI to the cloud

- **Expensive hardware** required for deep learning
- **Huge power consumption** for cloud AI hardware
- **Real-time critical AI** cannot rely on the internet connection

Typical Edge AI application workflow

Memory Footprint	~>10G
Power Consumption	>~300w
Computational Complexity	> 100 TOPs
Cost (ASP)	> $5,000

Typical Cloud HW

Edge Computing Challenges

High Computational Complexity

Millions of expensive floating-point operations for each input classification are needed.

Memory Footprint

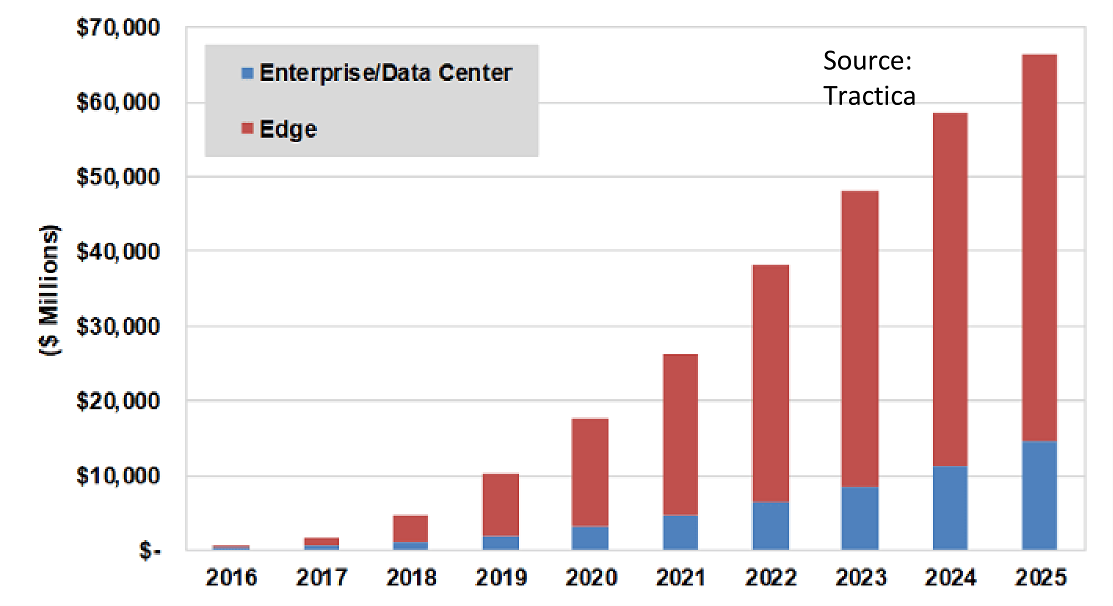
Huge amounts of weights and activations with limited on-chip memory and bandwidth.

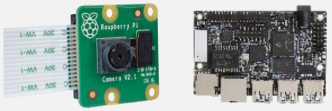
Power consumption

Deep learning requires significant power and can easily consume battery life

Time to deploy AI on edge devices

- Massive value unlocked by making AI applicable for cost-effective hardware
- **AI inference must meet strict power, speed, cost and resource constraints**

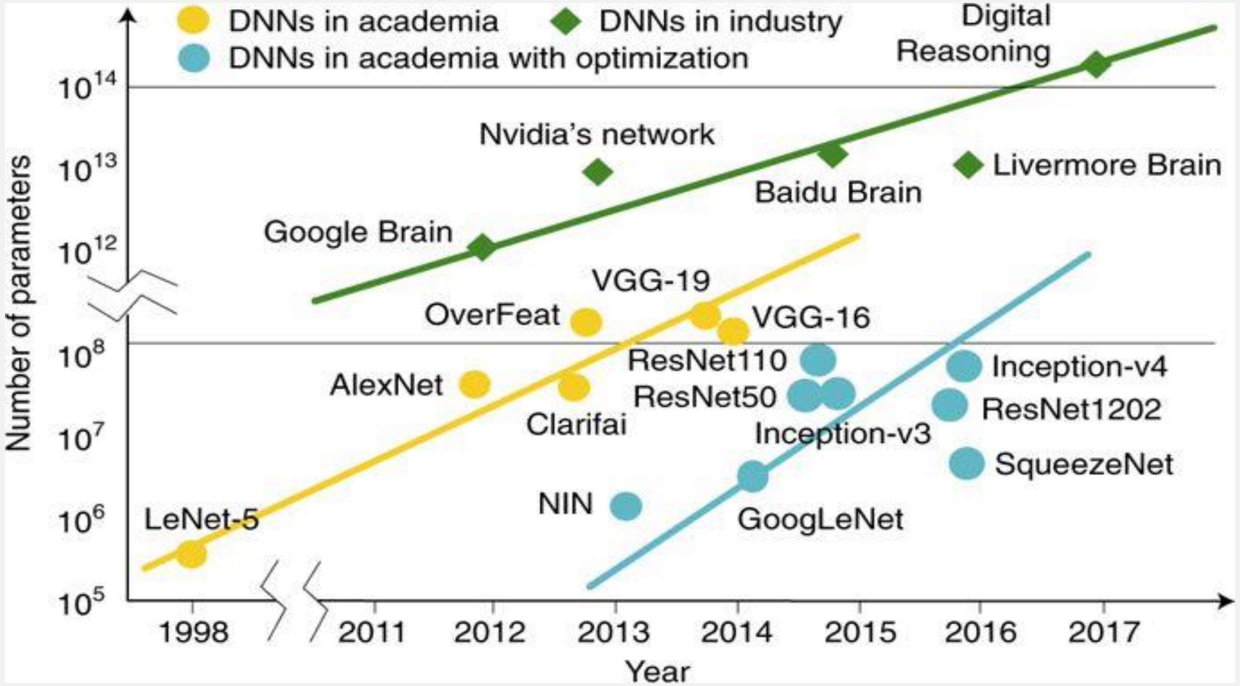


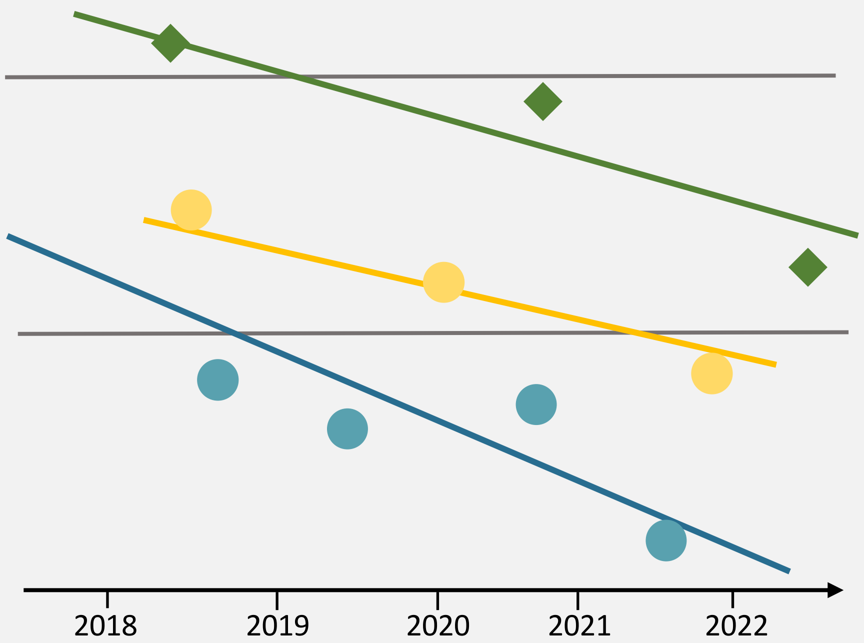


Memory Footprint	~<1M
Power Consumption	~<10w
Computational Complexity	~<10 TOPs
Cost (ASP)	~$10

Typical Edge HW

Edge Computing Solution: Small is the New Big



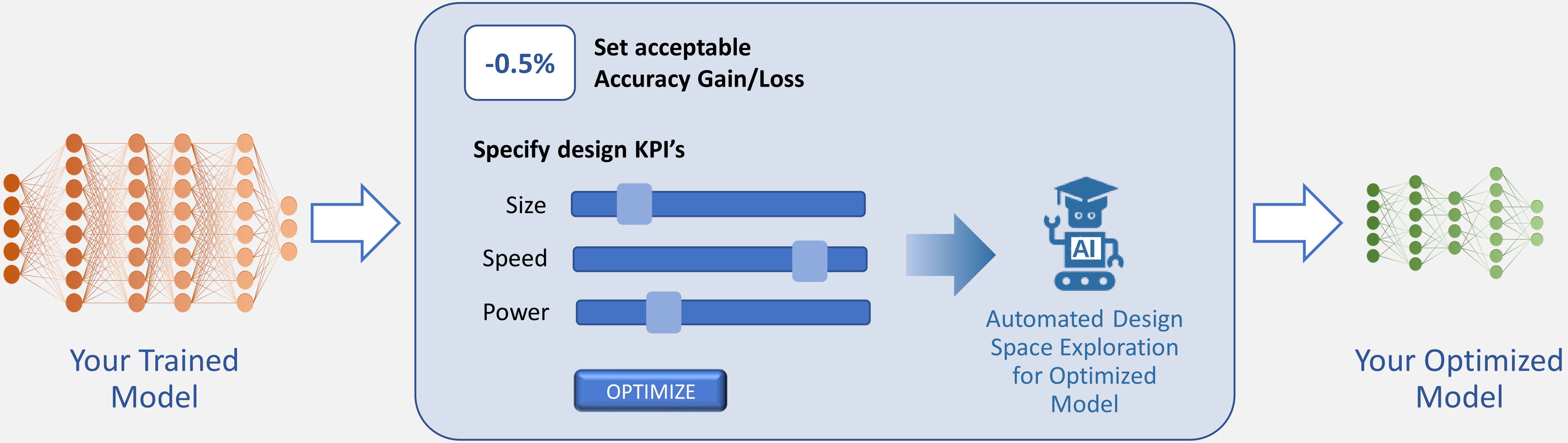


The Past

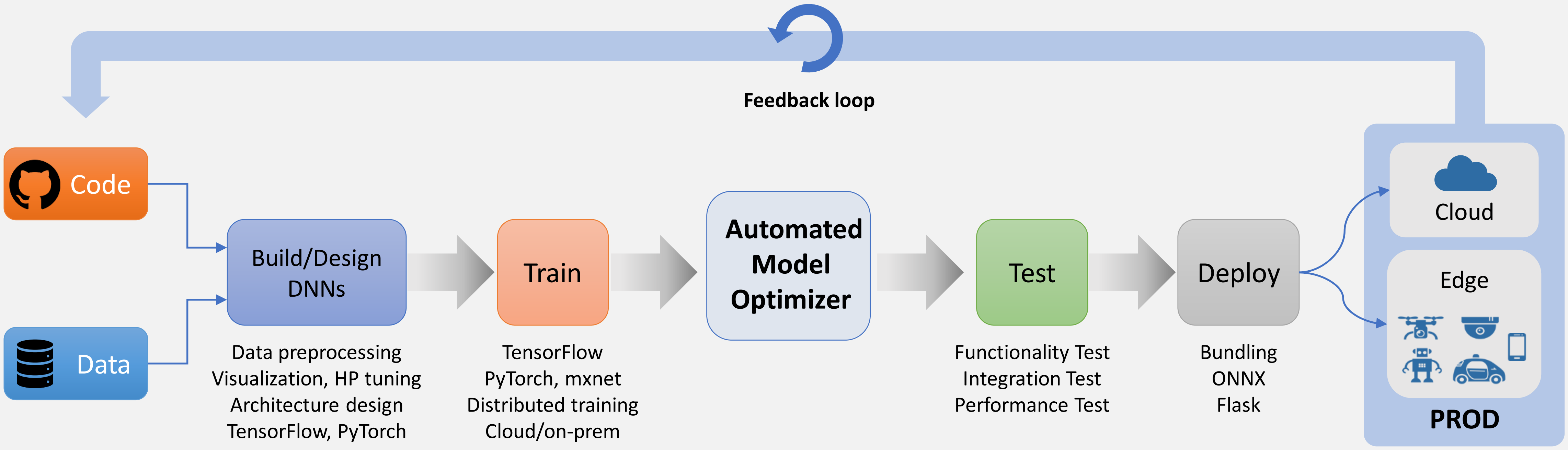
The Future

Designing compact deep learning models

Automated, intelligent optimization methods help AI engineers to automatically create faster, smaller & more efficient model architectures for production edge devices.



Where does this fit in an ML/AI Workflow



Levels of Optimization

Data labeling | DNN architecture design

AI Frameworks: mxnet, PYTORCH, TensorFlow, Keras

Hyper parameter tuning: SIGOPT, Auptimizer

Training Infrastructure: NVIDIA, aws, Microsoft Azure

Design and Train models (1)

Replaces Manual/Traditional Optimization
(Pruning, INT8 Quantization, etc.)

Content aware optimizations (2)

Computational Graph Optimization

TensorRT, TensorFlowLite, XILINX, ARM, OpenCL, OpenVINO, NVIDIA CUDA, LLVM, intel, tvm

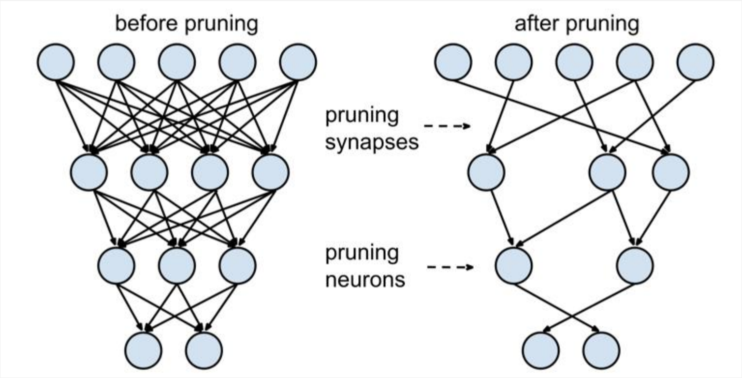
Compilers and Low-Level Platform aware Optimization (3)

Target Hardware: Intel CPU GPU FPGA ARM GRAPHCORE MYTHIC ETC.

Hardware Accelerators (4)

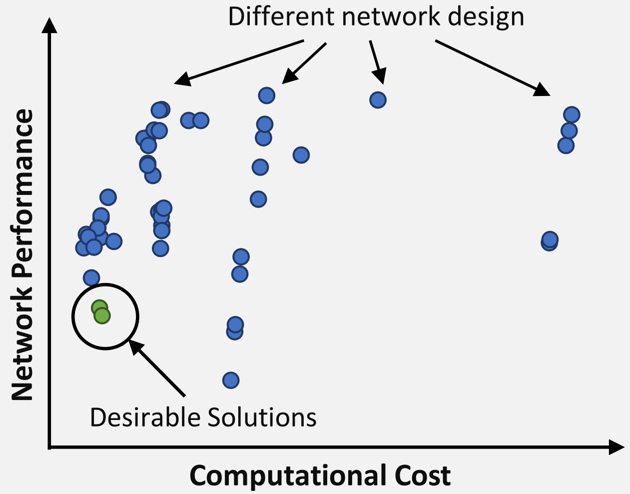
Types of Optimization

Pruning and Network Approximation



AutoML and Neural Architecture Search (NAS)

Automated Design Space Exploration



Our Focus

Optimization Benchmarks – Computer Vision

Application	Model	Compression[3]			Complexity Reduction (FLOPs) [3]	Accuracy Drop (%)	Dataset
		Original Size	Optimized Size	Improvement			
Image classification	VGG19	80MB	2.16MB	**x37**	x5	**<1%**	CIFAR100
	Resnet50	98MB	6.71MB	**x14.6**	x6	**<1%**	CIFAR100
	Resnet18	45MB	3.16MB	**x14.2**	x6	**<1%**	CIFAR100
	Mobilenet-v1.0	12.8MB	530KB	**x22**	x5	**~1.5%**	Visual Wake Words
	Industry use case[1]	45MB	1.8MB	**x25**	x4	**<1%**	Subset of Imagenet
Activity Recognition	Industry use case[2]	1.9MB	59KB	**x32**	x100	**~0%**	Custom dataset
Object Detection	ResNet50-SSD300	54MB	18MB	**x3**	x3	**~0%**	Subset of COCO2017

[1] Based on ResNet18 architecture

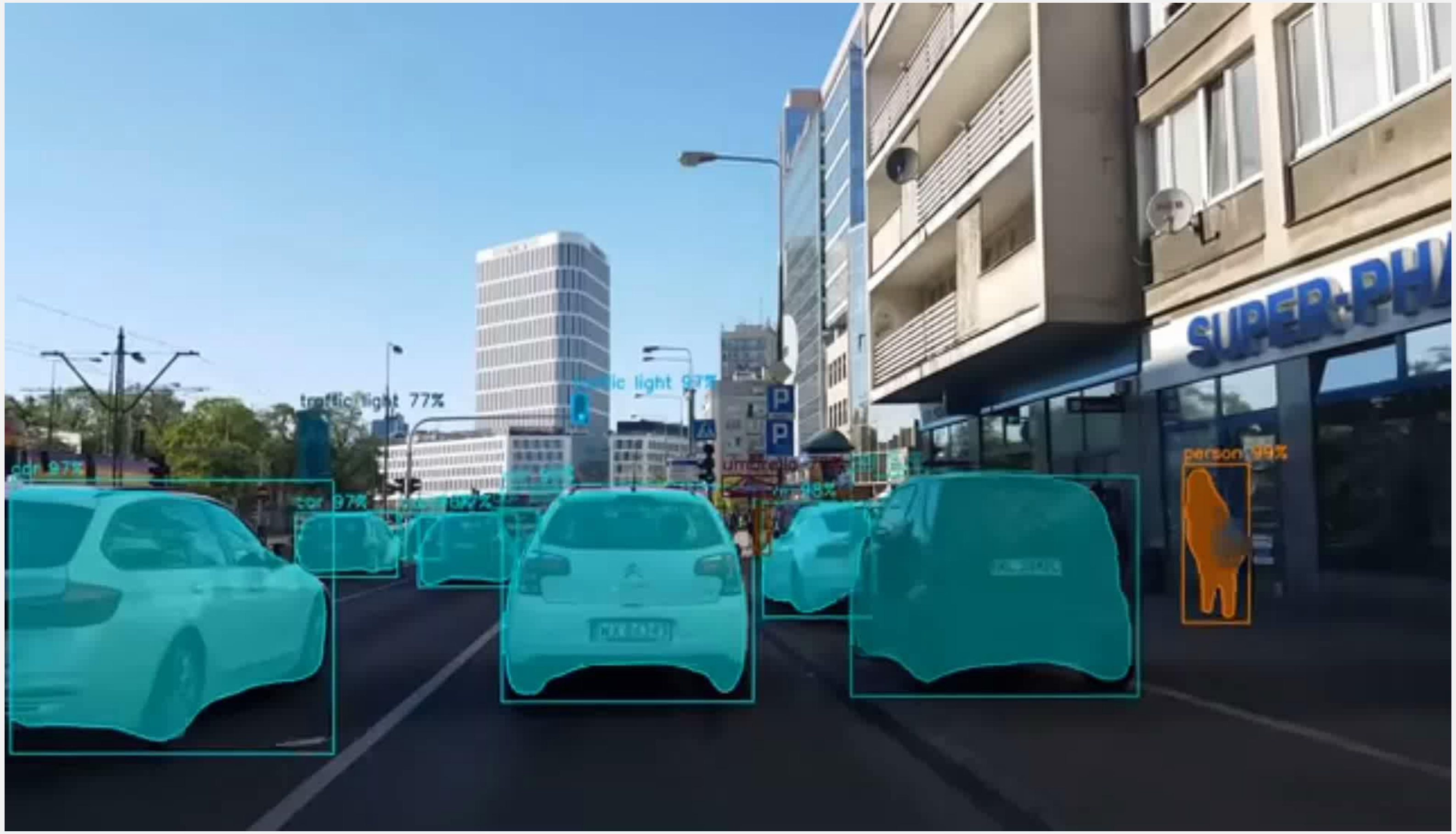
[2] Based on custom NN architecture

[3]**Results obtained purely using content-aware optimization (models in FP32). Further memory, speedup and energy savings available using platform-aware optimizations (INT8, mixed precision, binary weights etc.) and inference engine**

10x speedup on ARM mobile CPU

Optimized vs. Unoptimized model on Android phone

Accelerating Autonomous Perception

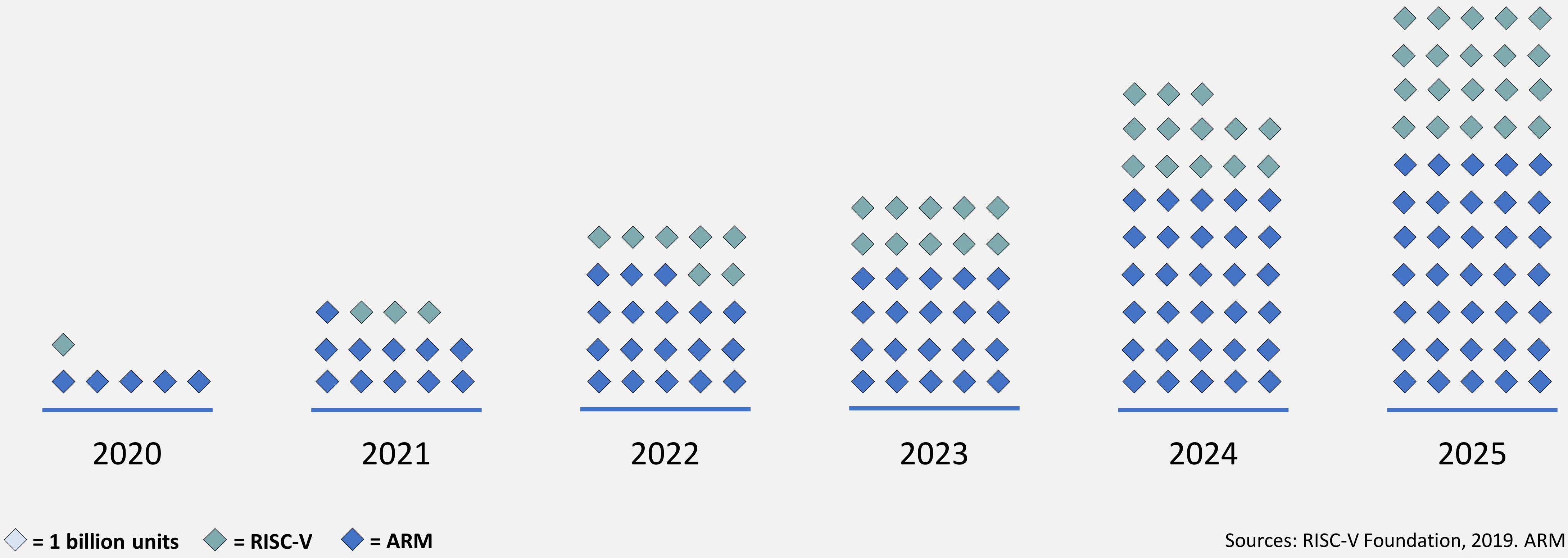


Expensive hardware required ~$10,000/GPU

Deep learning consumes ~20% of battery

AI on Low cost, low power chips

+100 billion IoT devices with ARM and RISC-V shipped over next 5 years



Sources: RISC-V Foundation, 2019. ARM IoT, 2020.

Bringing AI to daily life

- **Enable scalable** data centers and cloud services
- **Unlock new opportunities** by making DNNs applicable for edge computing
- **Reduce time to market** and engineering effort drastically

Thank you!

For more information and questions please contact:

Davis Sawyer, Co-founder and CPO, Deeplite Inc.
davis@deeplite.ai
info@deeplite.ai