MATLAB EXPO

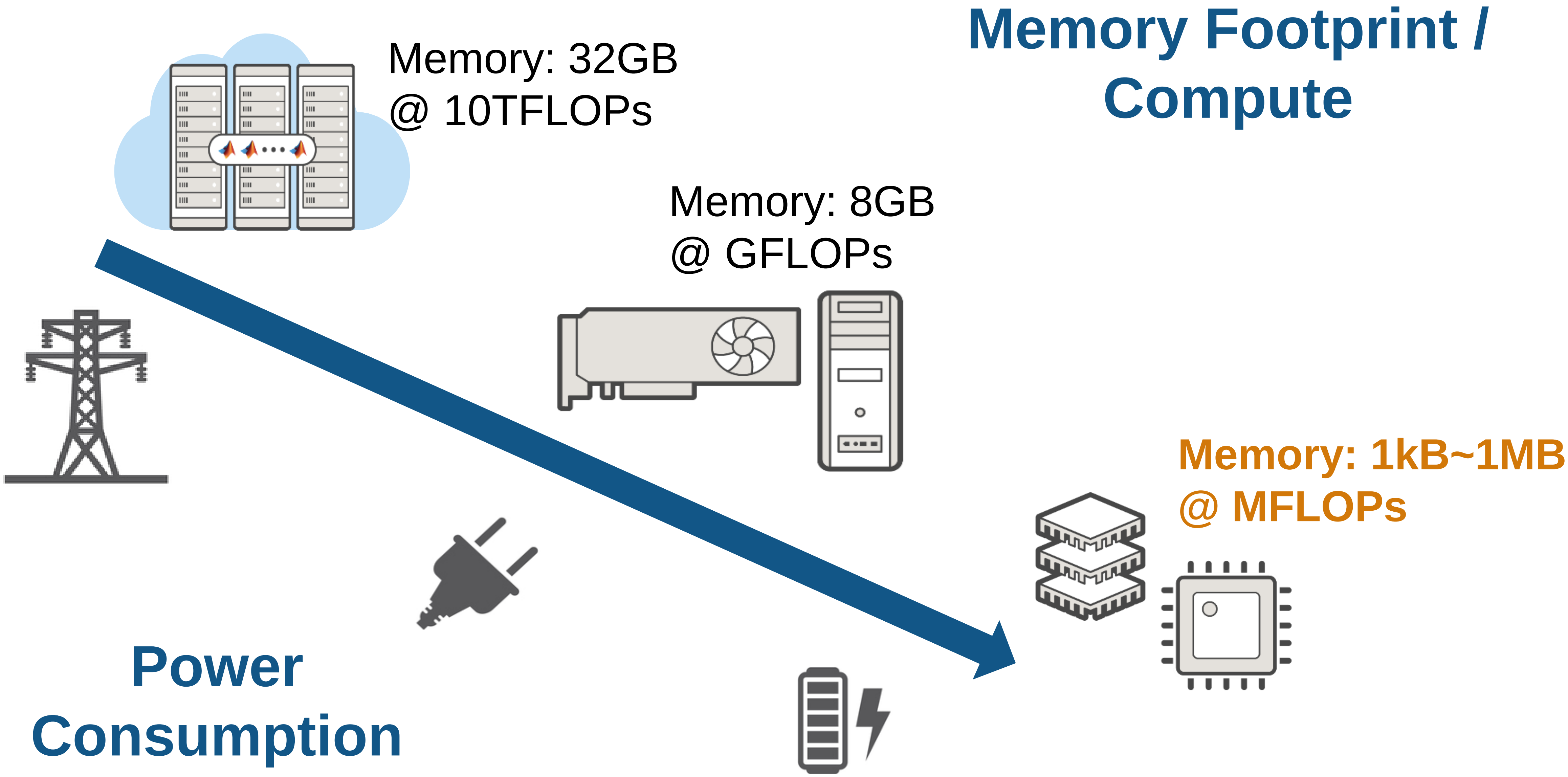
Fitting AI for Embedded Deployment

Bernhard Suhm, MathWorks

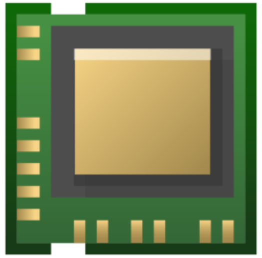
Emelie Andersson, MathWorks

Edge AI innovates many industries!

Hardware Constraints

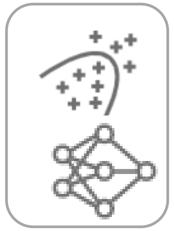


What is “Edge” (Embedded) AI?



The chip has only 500 KB memory – make that smaller

How can I make my AI smaller?



Data Scientist

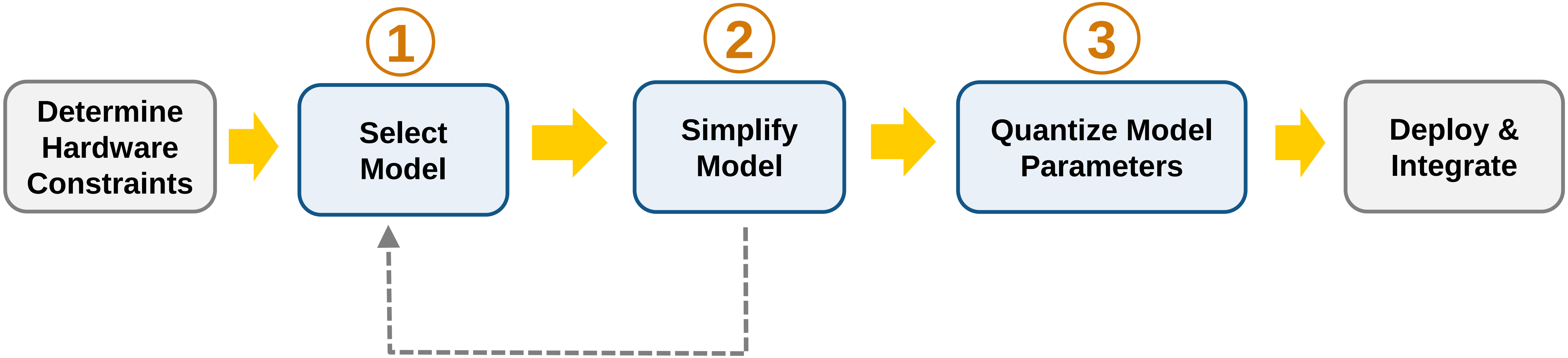
Embedded Software Engineer

Why is Edge AI (Model Compression) difficult?

AI is often big

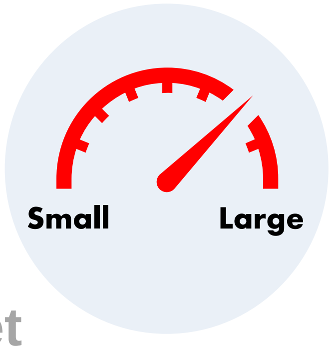
Knowledge Gap

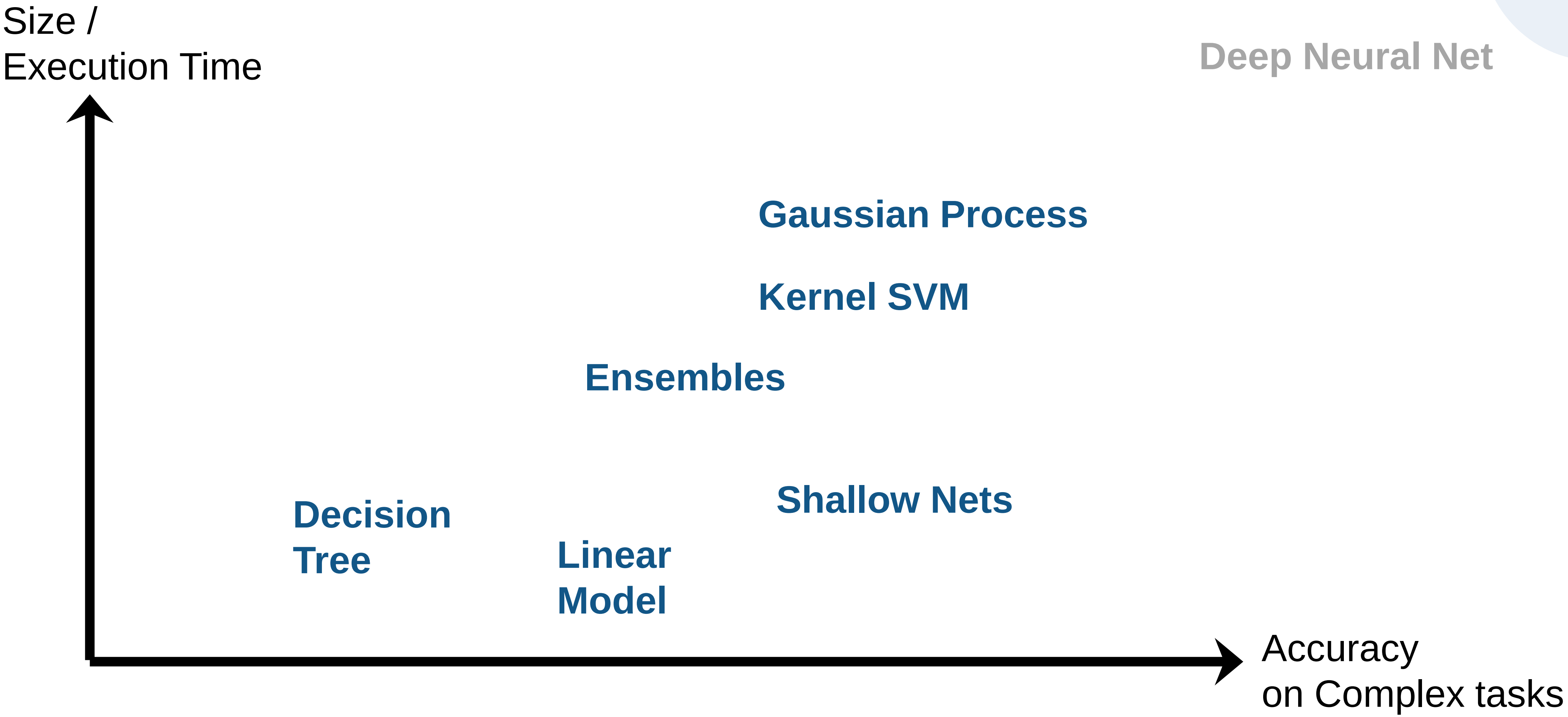
Model Compression Workflow



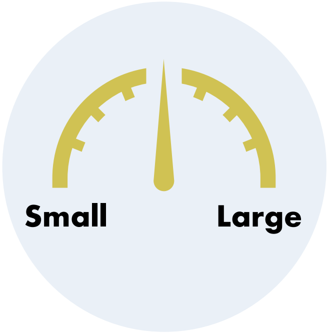
Compressing Machine Learning

Step 1 Size aware model selection

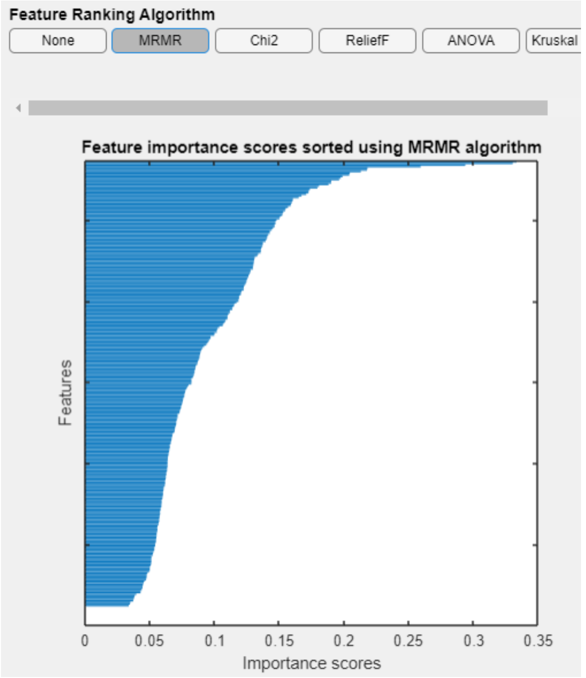


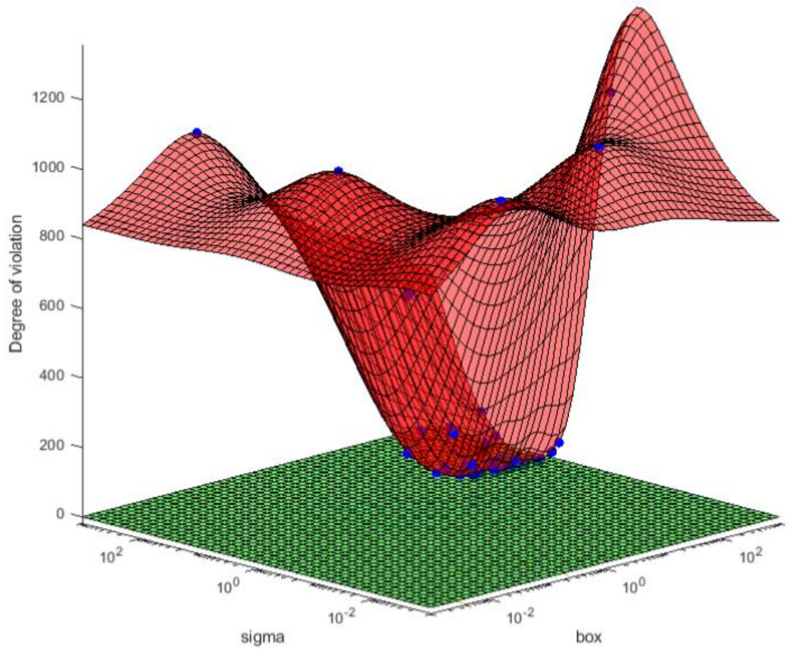


Step 2 Simplify the structure of your model

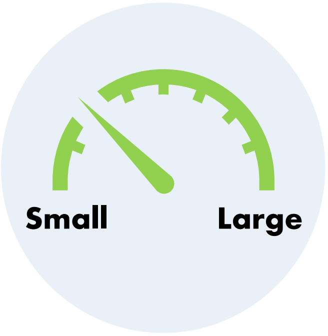


1. Fewer features
2. Tune size-relevant hyperparameters
3. Maximize accuracy given size constraint



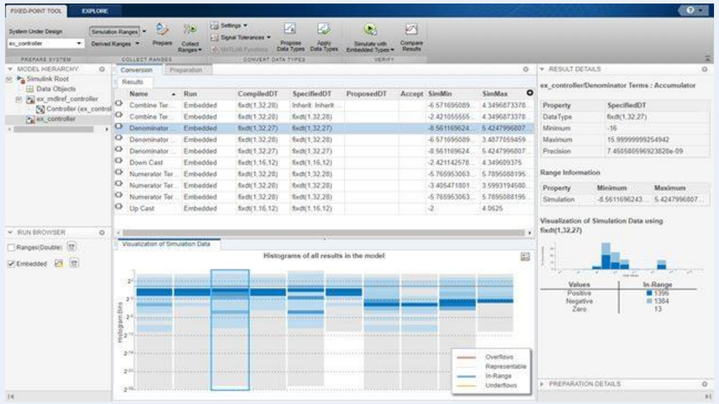


Step 3 Quantize your model



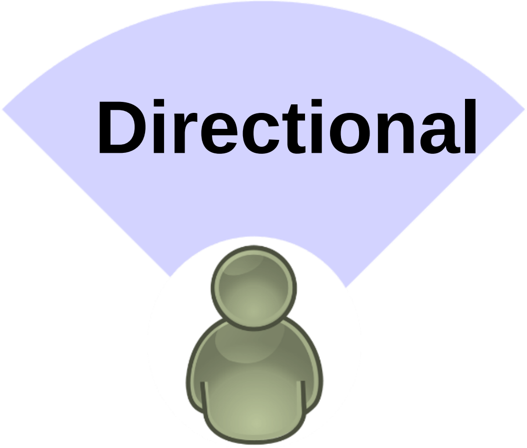
Model

Convert in Fixed-Point Designer



Demo: Embedding AI in an intelligent Hearing Aid

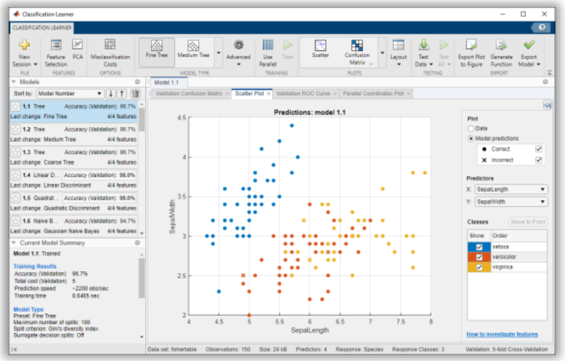
0.5 to 256 kB
on-chip memory



Functionality for Compressing Machine Learning in MATLAB

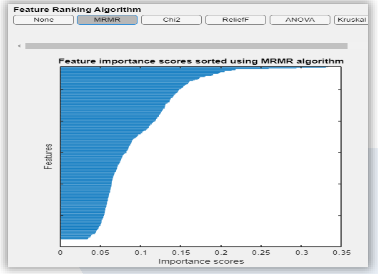
1

Classification / Regression Learner

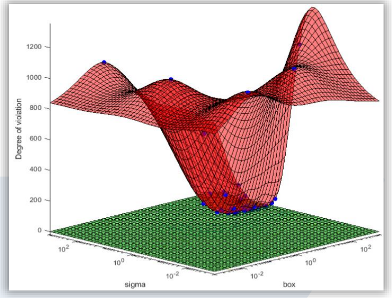


In-App Feature Selection

2

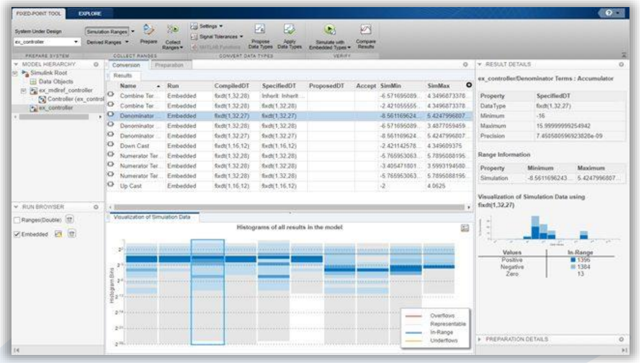


Bayesopt



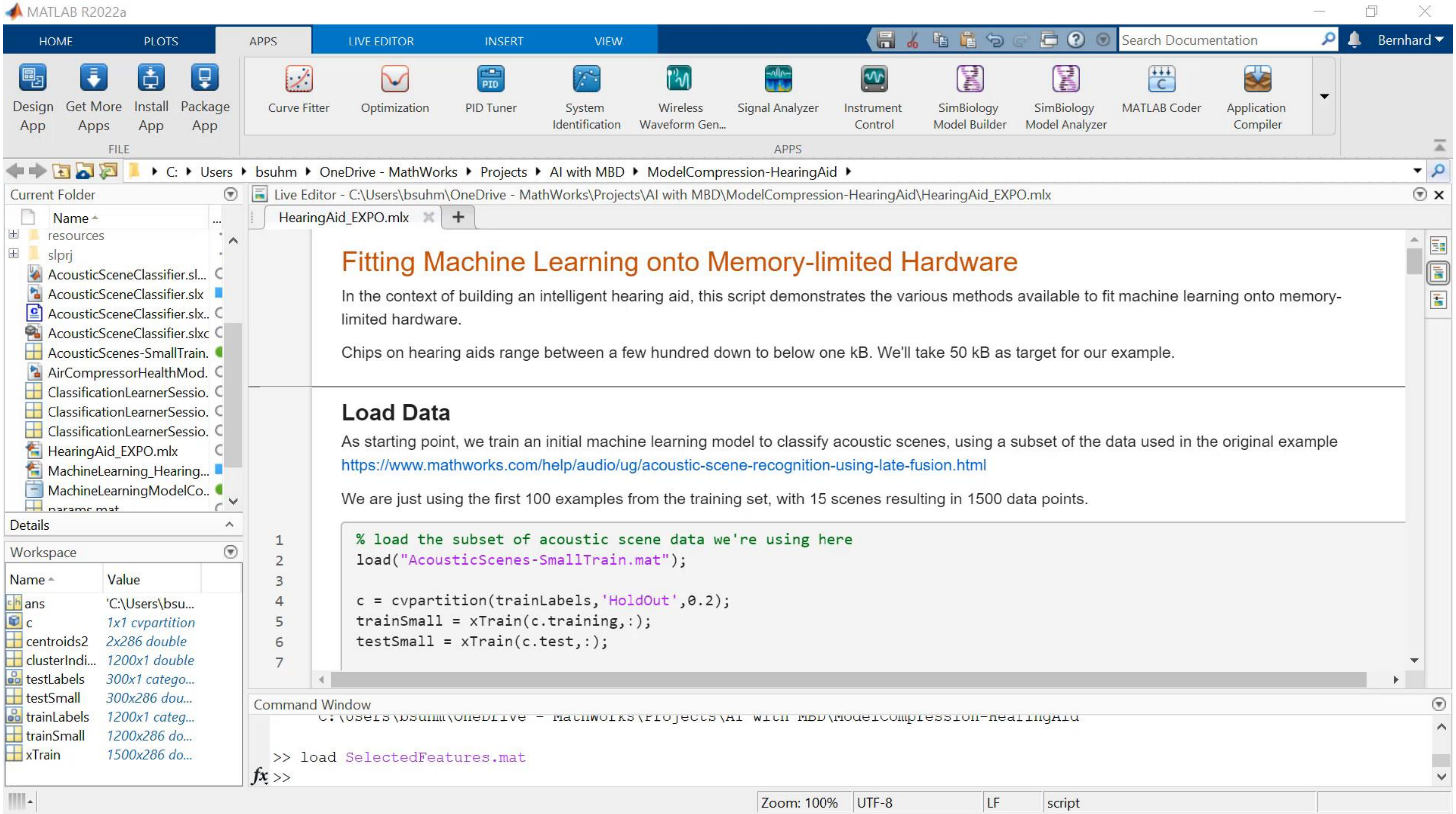
3

Fixed Point Designer / Native Simulink Block



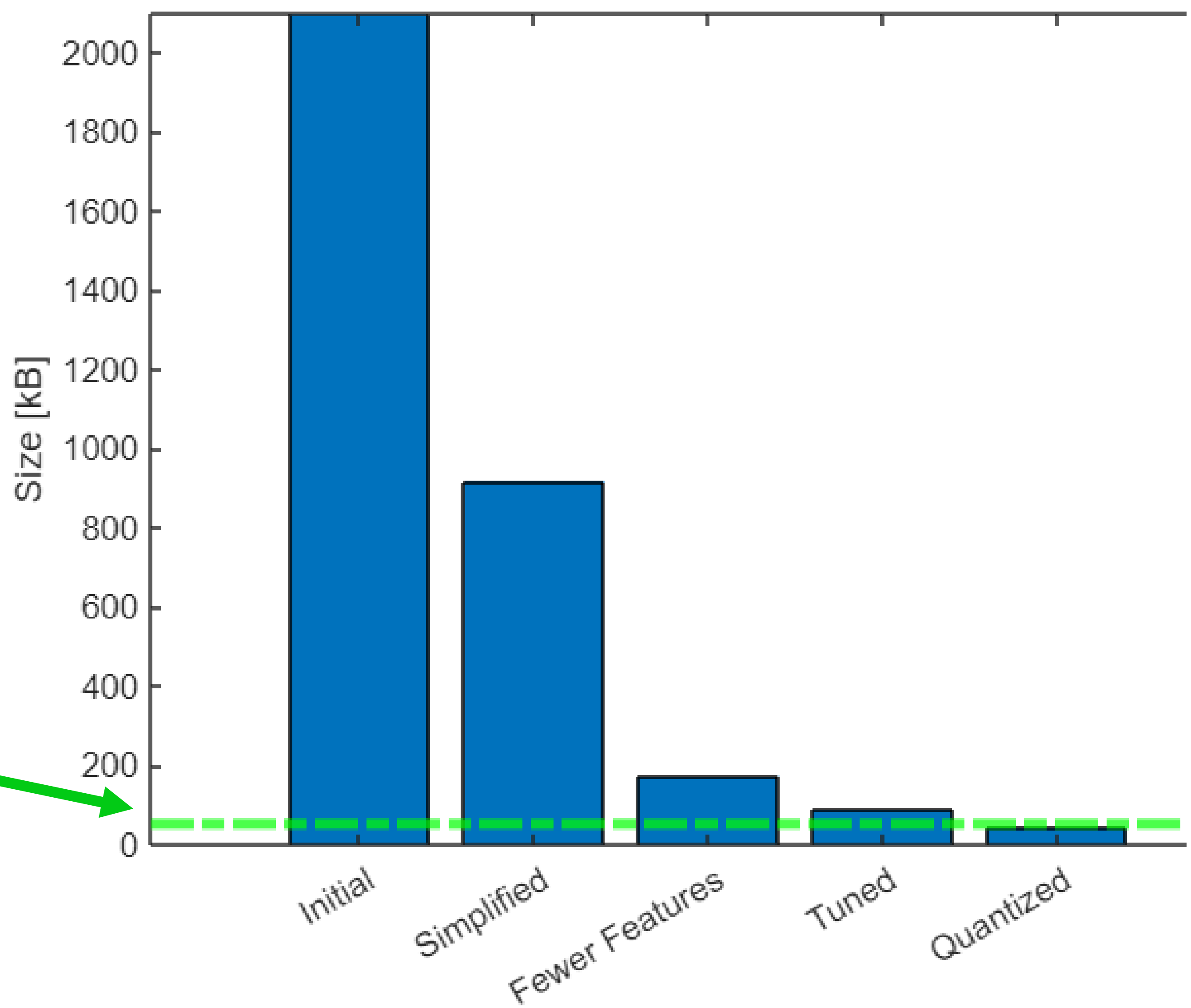
Determine Hardware Constraints → Select (Initial) Model → Simplify Model (Select Features; Tune Hyper-parameters) → Quantize Model Parameters → Deploy & Integrate

Demo: Fit Machine Learning for Intelligent Hearing Aid



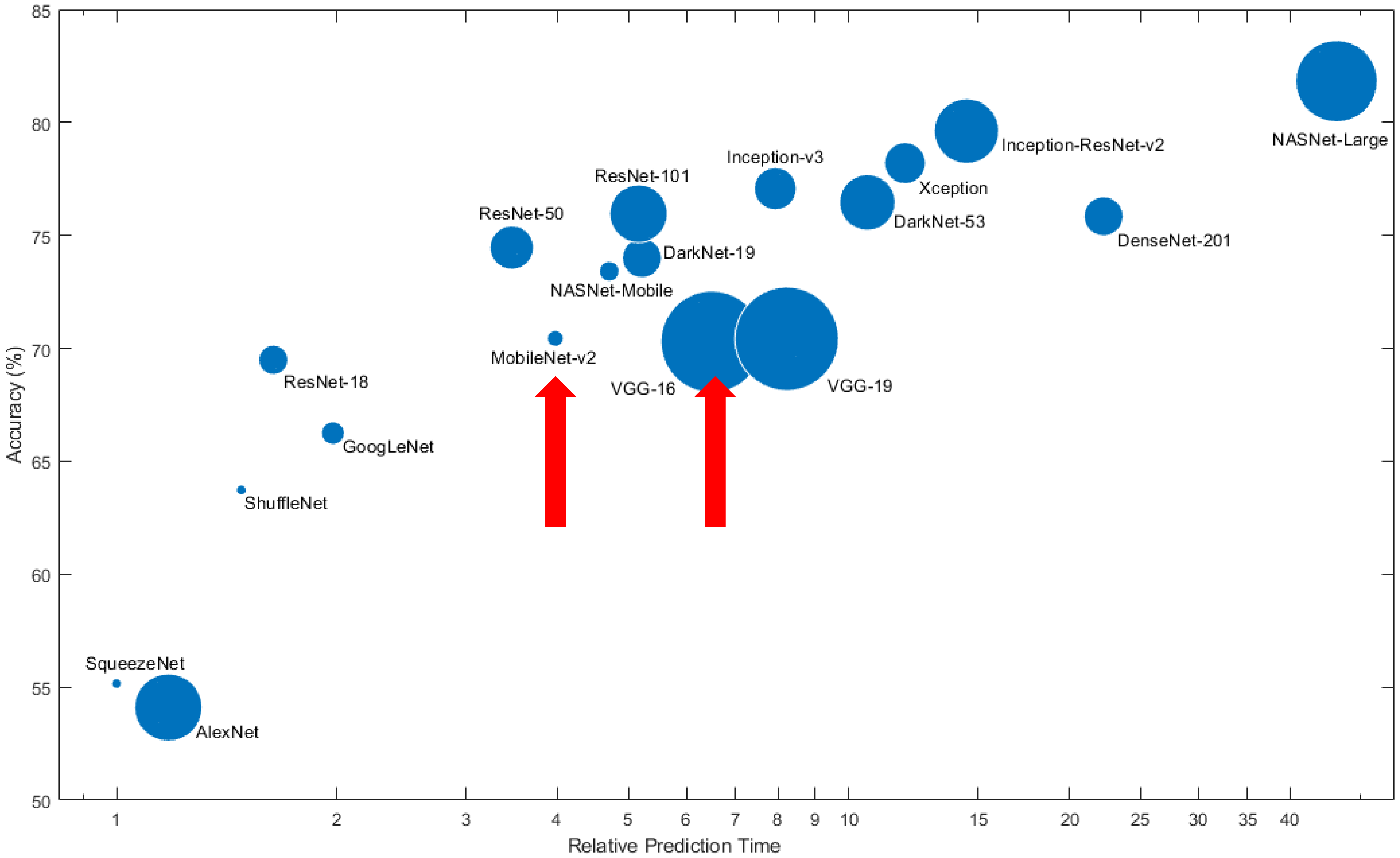
Machine Learning Demo Size Reduction by factor 20

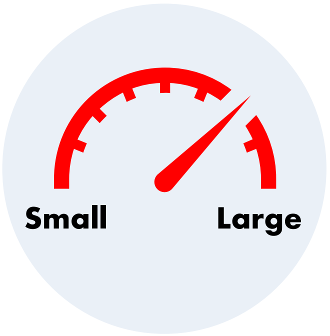
Target Hardware
Size: 50 kB



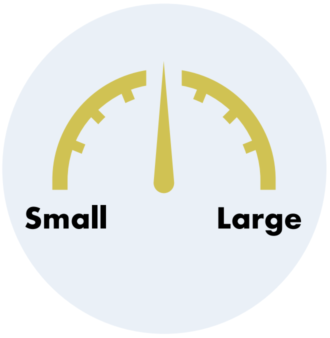
Compressing Deep Learning

Step 1 Size aware model selection

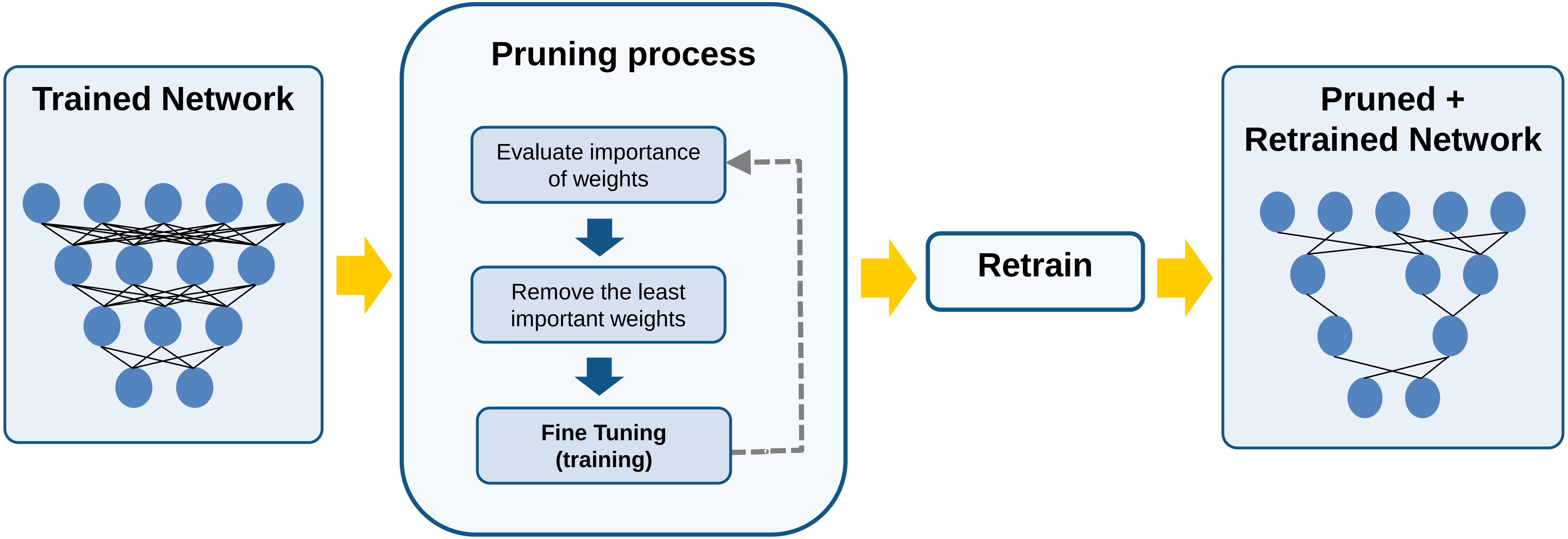




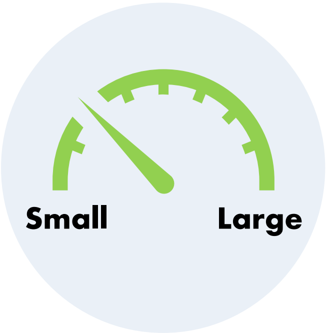
Step 2 Smart pruning

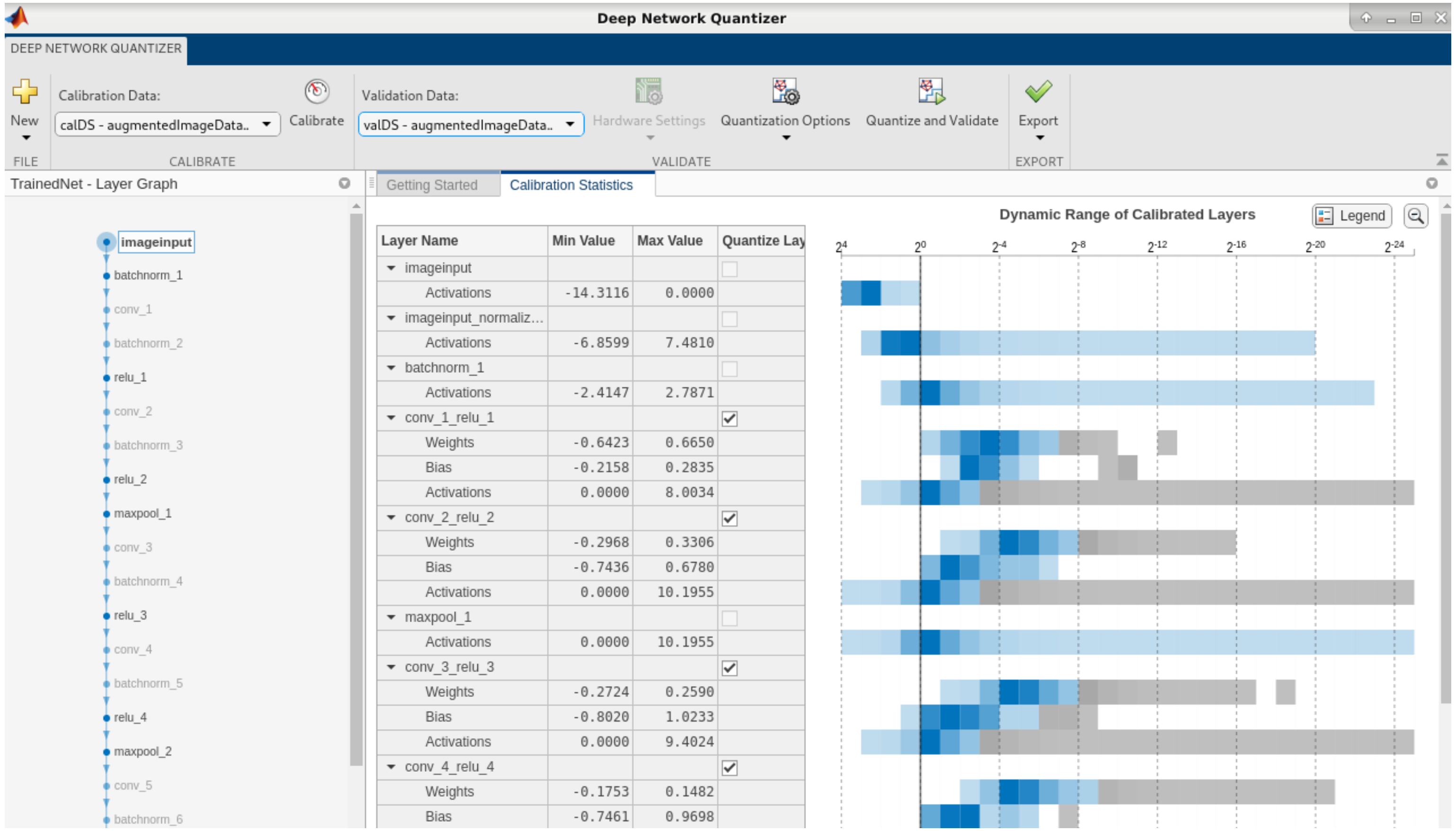


Remove **unimportant** parts of the network



Step 3 Quantize your model





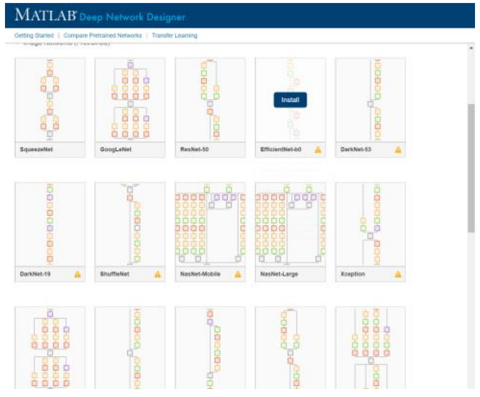
Deep Learning Demo: Scene classification

Classify 10 classes

More difficult problem → more complex model

Functionality for Compressing Deep Neural Nets

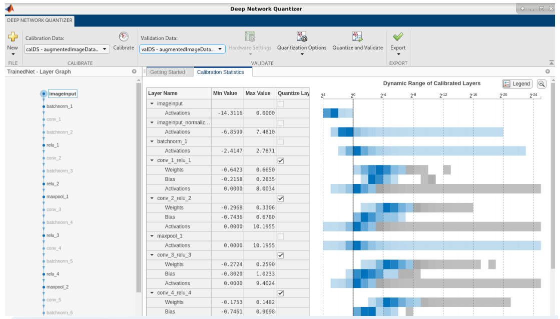
1 Deep Network Designer



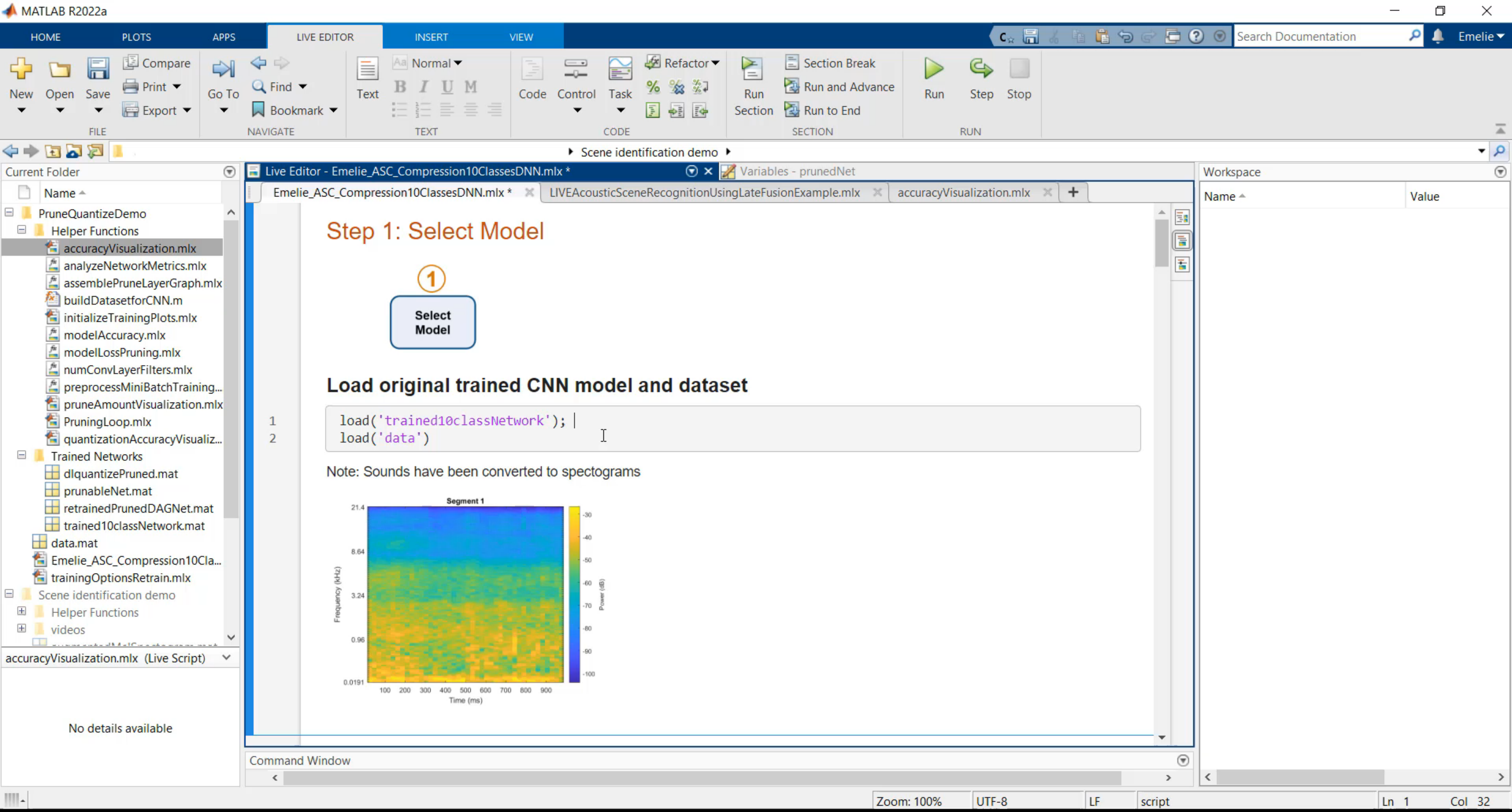
2 Taylor Pruning

```
taylorPrunableNetwork(net)
```

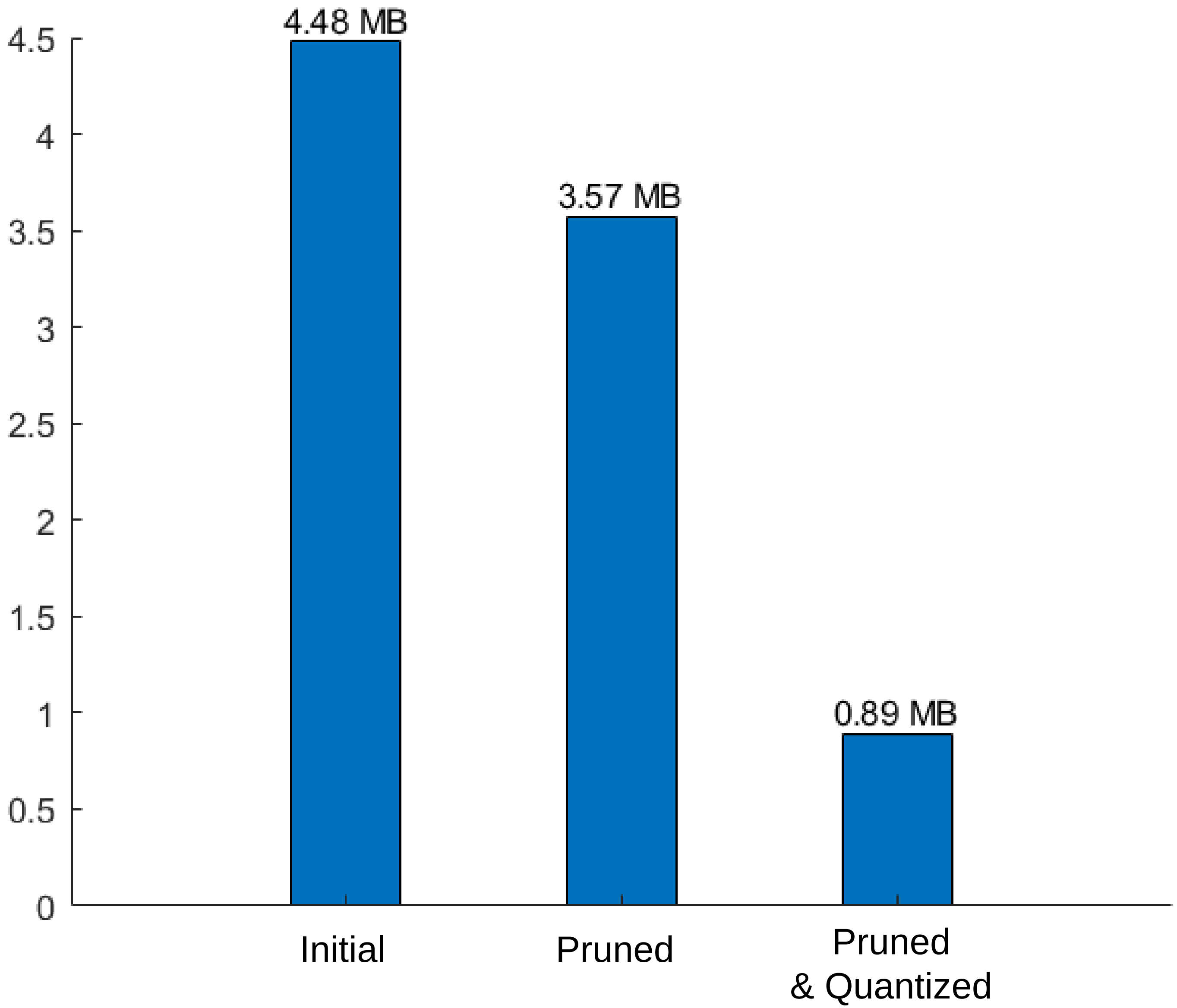
3 Deep Network Quantizer



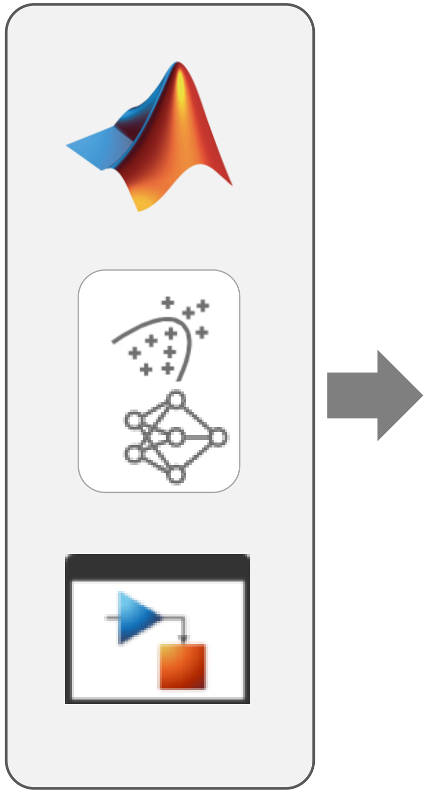
Determine Hardware Constraints → Select Model → Simplify Model (Prune Deep Neural Network) → Quantize Model Parameters → Deploy & Integrate

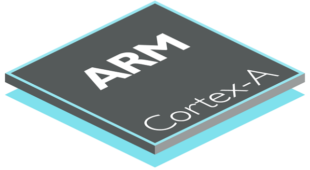

MATLAB R2022a
HOME
PLOTS
APPS
LIVE EDITOR
INSERT
VIEW
Search Documentation
Emelie
New
Open
Save
Compare
Print
Export
FILE
Go To
Find
Bookmark
NAVIGATE
Text
Normal
TEXT
Code
Control
Task
Refactor
CODE
Run Section
Section Break
Run and Advance
Run to End
SECTION
Run
Step
Stop
RUN
Scene identification demo
Current Folder
Name
PruneQuantizeDemo
Helper Functions
accuracyVisualization.mlx
analyzeNetworkMetrics.mlx
assemblePruneLayerGraph.mlx
buildDatasetforCNN.m
initializeTrainingPlots.mlx
modelAccuracy.mlx
modelLossPruning.mlx
numConvLayerFilters.mlx
preprocessMiniBatchTraining...
pruneAmountVisualization.mlx
PruningLoop.mlx
quantizationAccuracyVisualiz...
Trained Networks
dlquantizePruned.mat
prunableNet.mat
retrainedPrunedDAGNet.mat
trained10classNetwork.mat
data.mat
Emelie_ASC_Compression10Cla...
trainingOptionsRetrain.mlx
Scene identification demo
Helper Functions
videos
accuracyVisualization.mlx (Live Script)
No details available
Live Editor - Emelie_ASC_Compression10ClassesDNN.mlx *
Variables - prunedNet
Emelie_ASC_Compression10ClassesDNN.mlx *
LIVEAcousticSceneRecognitionUsingLateFusionExample.mlx
accuracyVisualization.mlx
Step 1: Select Model
1
Select Model
Load original trained CNN model and dataset
load('trained10classNetwork');
load('data')
Note: Sounds have been converted to spectograms
Segment 1
Frequency (kHz)
21.4
8.64
3.24
0.96
0.0191
100 200 300 400 500 600 700 800 900
Time (ms)
Power (dB)
-30
-40
-50
-60
-70
-80
-90
-100
Workspace
Name
Value
Command Window
Zoom: 100%
UTF-8
LF
script
Ln 1
Col 32

Deep Learning Demo Size Reduction by factor 5



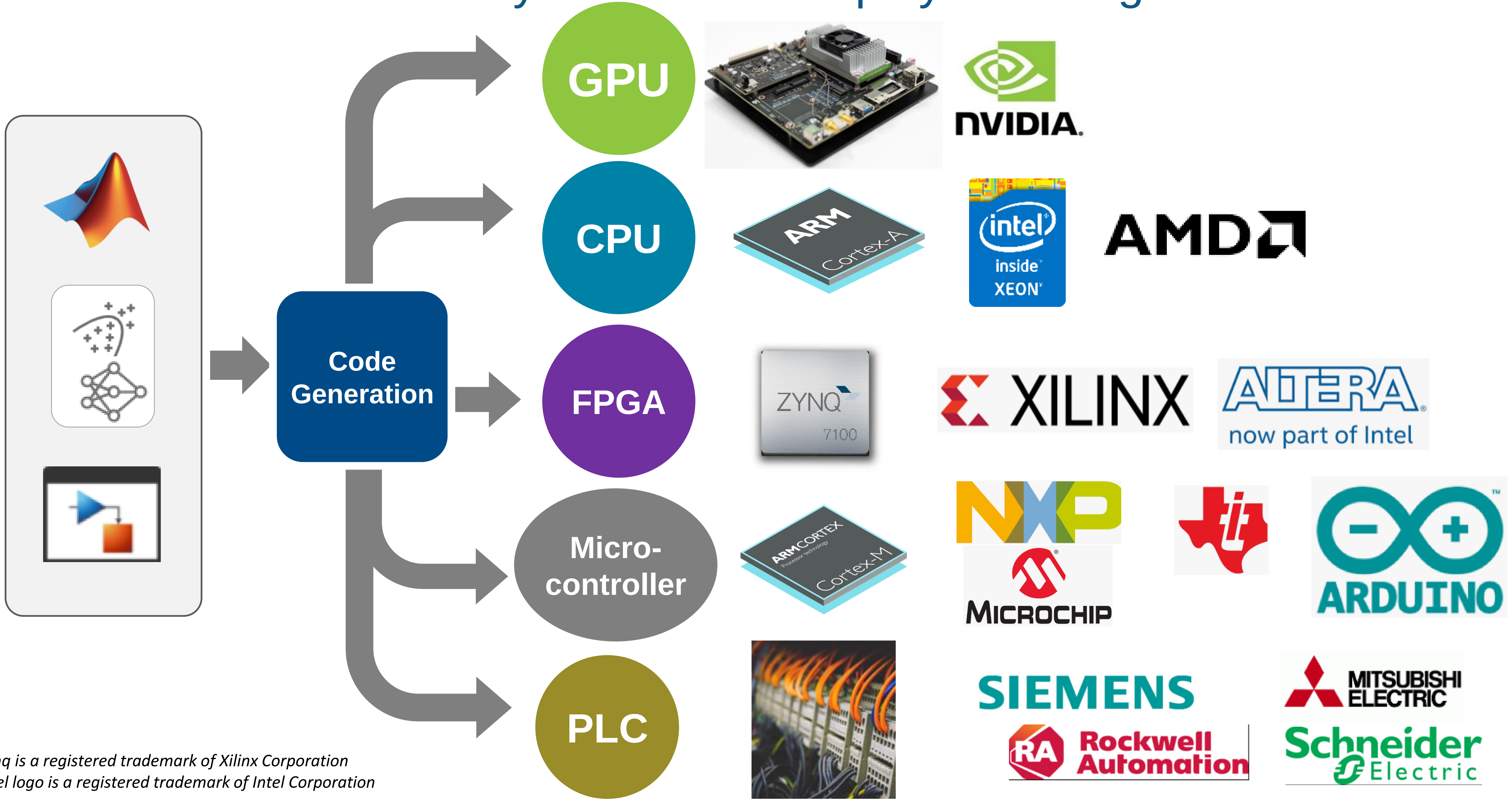
One Codebase – Many Embedded Deployment targets





Zynq is a registered trademark of Xilinx Corporation
Intel logo is a registered trademark of Intel Corporation

One Codebase – Many Embedded Deployment targets



Zynq is a registered trademark of Xilinx Corporation
Intel logo is a registered trademark of Intel Corporation

Conclusions

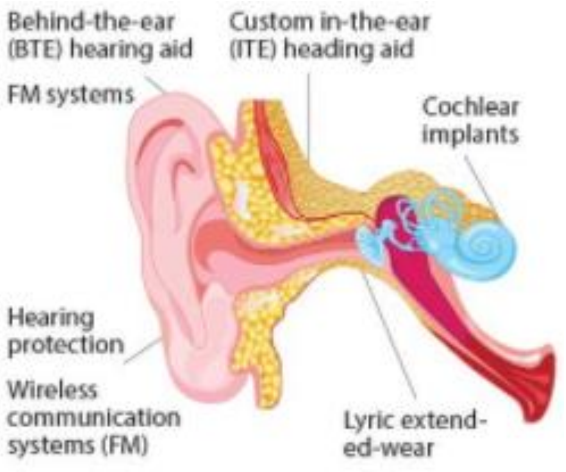
You can fit AI for many applications onto limited hardware

MathWorks tools make fitting AI models on constrained hardware a lot easier

Same high-level Workflow for any type of AI

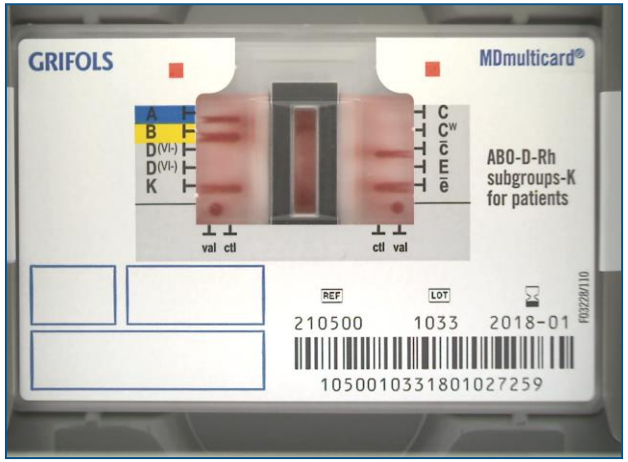
Which constraints are most challenging for your application?

Learn More



[Hearing Implant using MBD](#) sonova HEAR THE WORLD

[Autonomous Tractor](#) MONARCH



[Card to Classify Blood Type](#) IDNEO

To get your started:

[Learn about Embedded Deployment](#)

[Quantization of classification SVM](#) (Doc)

[Deploy Hand-Gesture Classifier onto Arduino](#) (Doc)

[Generate C/C++ Code from Simulink](#) (Video)

[Quantizing a Deep Neural Network](#) (Video)

(Doc)

Smaller Models are often Better!

www.linkedin.com/in/emelie-andersson-ai
www.linkedin.com/in/bernhard-suhm-ai

MathWorks®

MATLAB EXPO