

MATLAB EXPO

Fitting AI for Embedded Deployment

Bernhard Suhm, MathWorks



Emelie Andersson, MathWorks



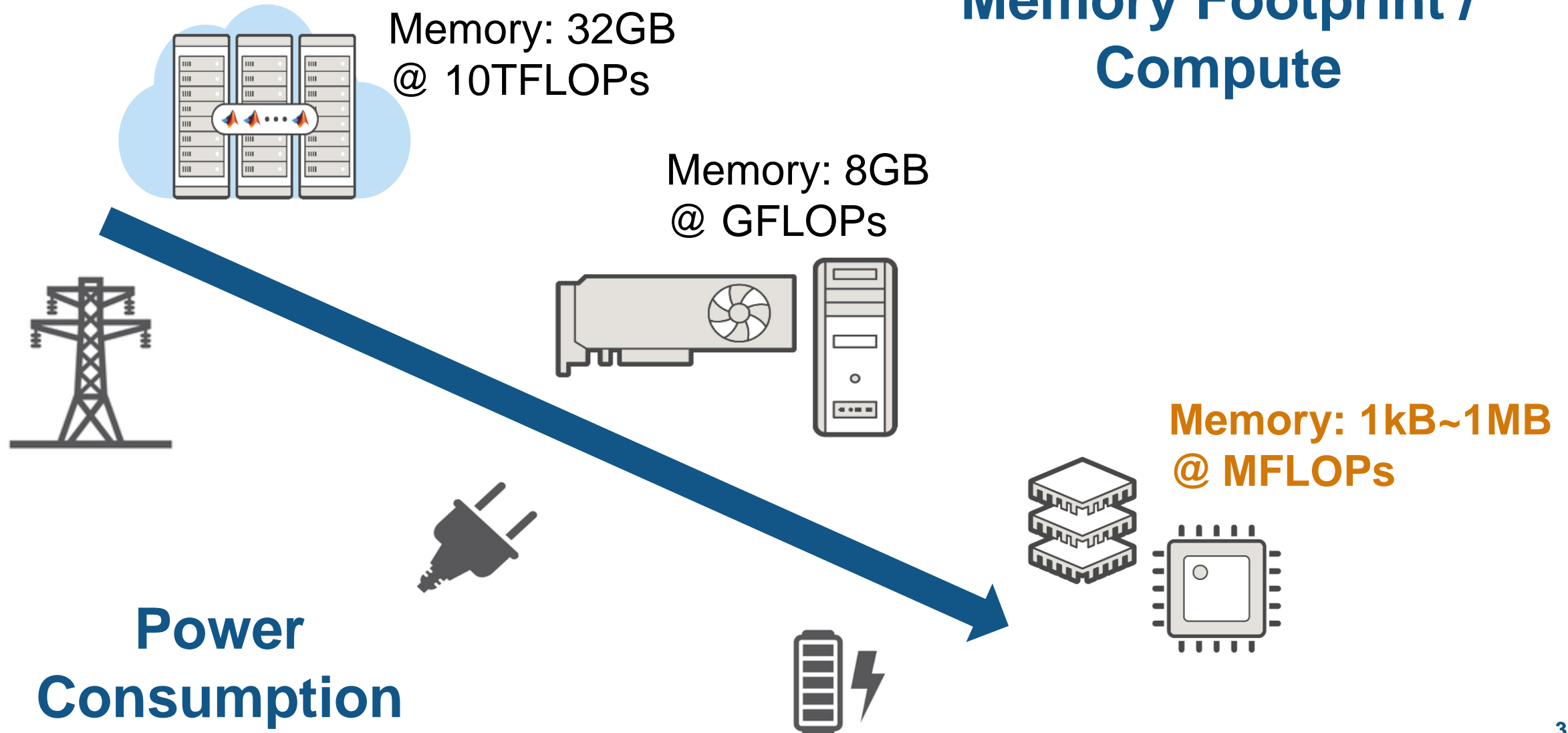


Edge AI innovates many industries!

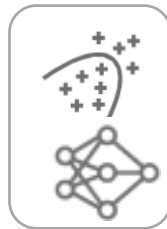
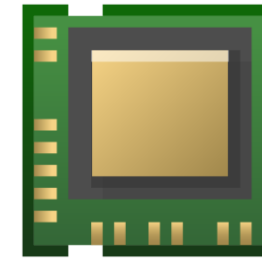


Hardware Constraints

Memory Footprint / Compute



What is “Edge” (Embedded) AI?



Data Scientist

How can I make
my AI smaller?



Embedded Software Engineer

The chip has only
500 KB memory –
make that smaller

Why is Edge AI (Model Compression) difficult?

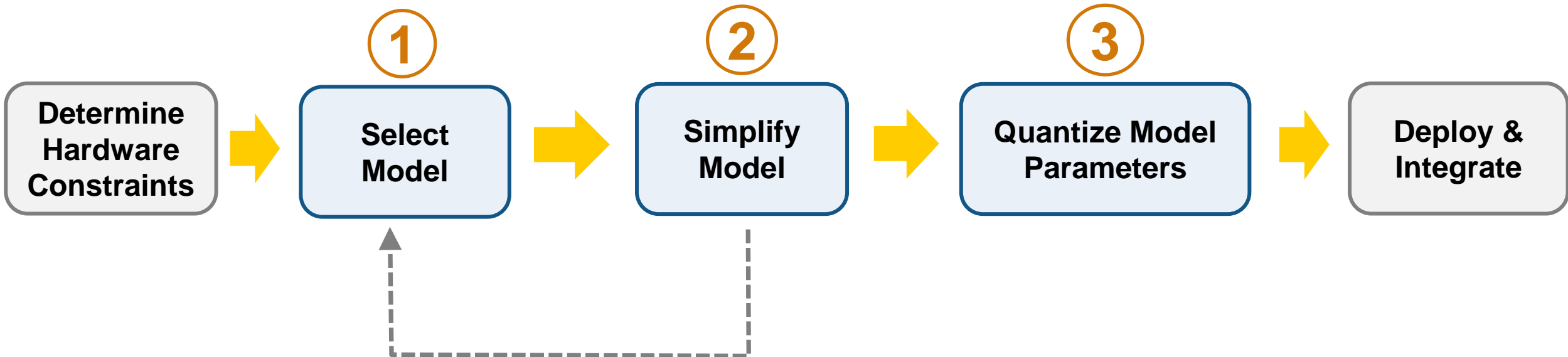


AI is often big



Knowledge Gap

Model Compression Workflow

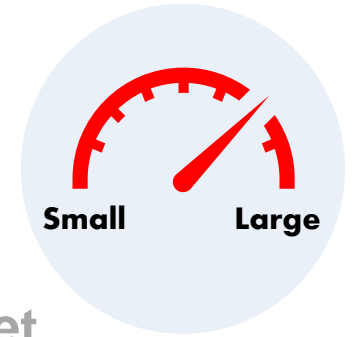


Compressing Machine Learning

Step ① Size aware model selection

Size /
Execution Time

Deep Neural Net



Gaussian Process

Kernel SVM

Ensembles

Shallow Nets

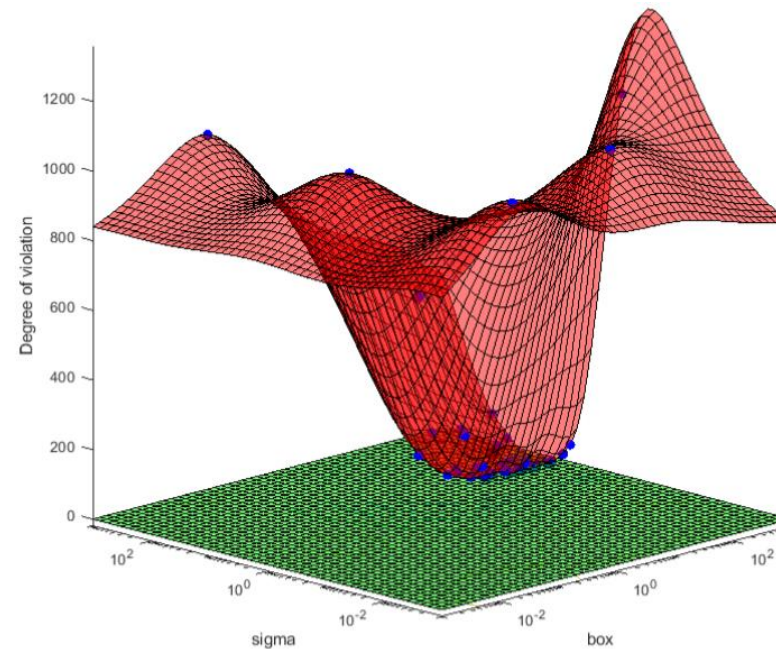
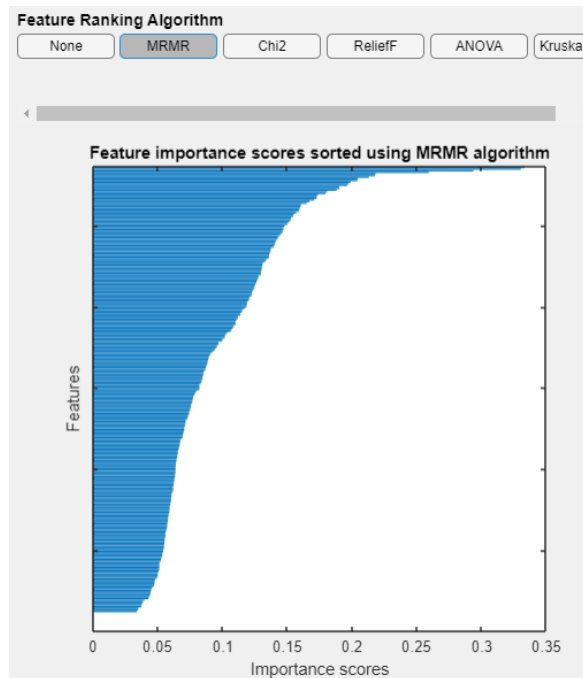
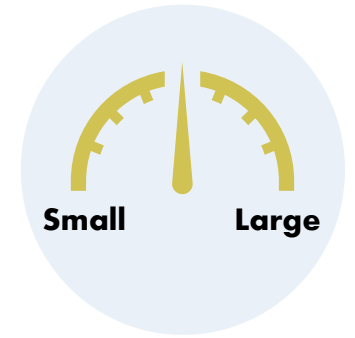
Decision
Tree

Linear
Model

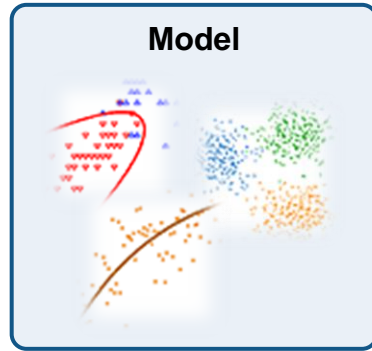
Accuracy
on Complex tasks

Step ② Simplify the structure of your model

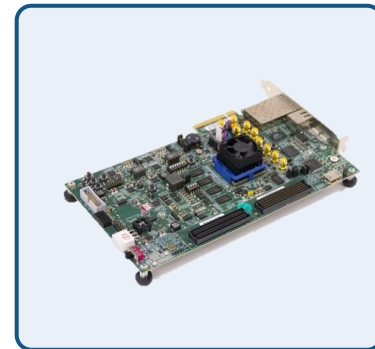
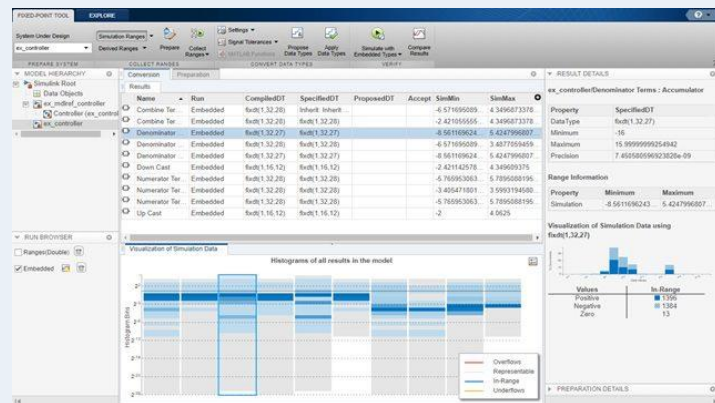
1. Fewer features
2. Tune size-relevant hyperparameters
3. Maximize accuracy given size constraint



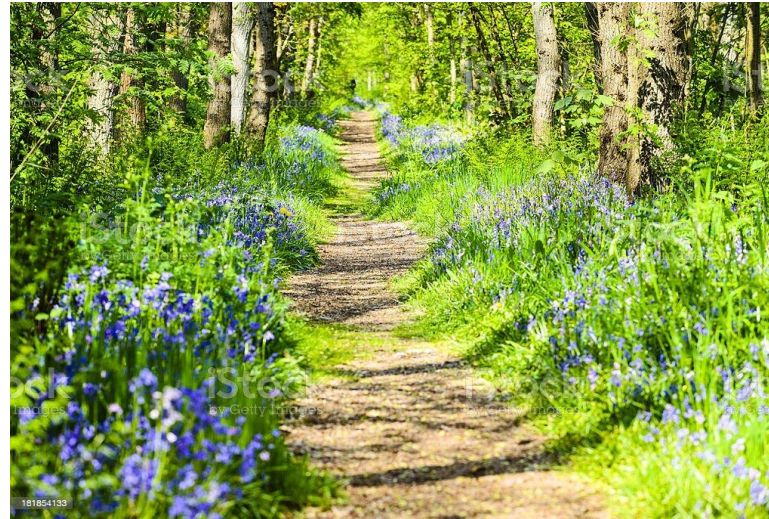
Step ③ Quantize your model



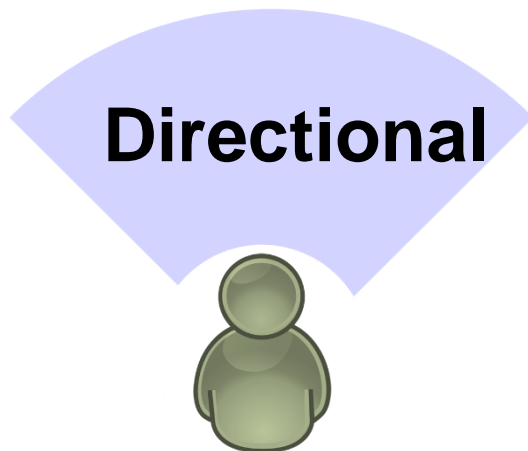
Convert in Fixed-Point Designer



Demo: Embedding AI in an intelligent Hearing Aid



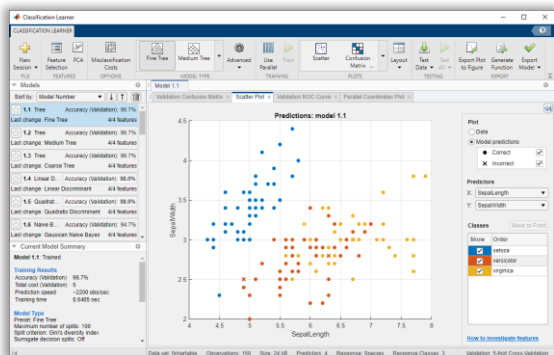
0.5 to 256 kB
on-chip memory



Functionality for Compressing Machine Learning in MATLAB

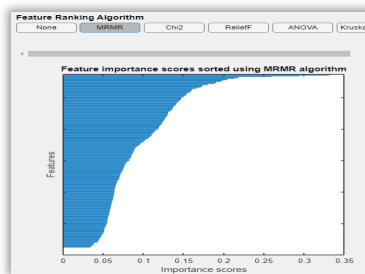
1

Classification / Regression Learner

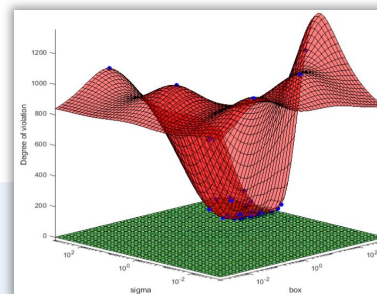


2

In-App Feature Selection

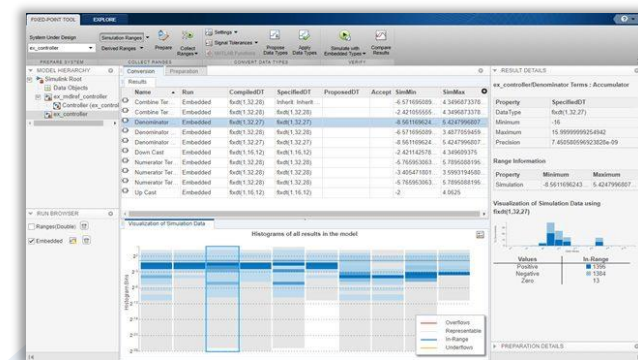


Bayesopt



3

Fixed Point Designer / Native Simulink Block



Simplify Model

Select Features

Tune Hyper-parameters

Determine Hardware Constraints

Select (Initial) Model

Quantize Model Parameters

Deploy & Integrate

Demo: Fit Machine Learning for Intelligent Hearing Aid

The screenshot displays the MATLAB R2022a interface. The top menu bar includes HOME, PLOTS, APPS, LIVE EDITOR, INSERT, and VIEW. The APPS gallery shows various toolboxes like Curve Fitter, Optimization, PID Tuner, System Identification, Wireless Waveform Gen..., Signal Analyzer, Instrument Control, SimBiology Model Builder, SimBiology Model Analyzer, MATLAB Coder, and Application Compiler. The current folder is C:\Users\bsuhm\OneDrive - MathWorks\Projects\AI with MBD\ModelCompression-HearingAid. The Live Editor window shows a script titled 'HearingAid_EXPO.mlx' with the following content:

Fitting Machine Learning onto Memory-limited Hardware

In the context of building an intelligent hearing aid, this script demonstrates the various methods available to fit machine learning onto memory-limited hardware.

Chips on hearing aids range between a few hundred down to below one kB. We'll take 50 kB as target for our example.

Load Data

As starting point, we train an initial machine learning model to classify acoustic scenes, using a subset of the data used in the original example <https://www.mathworks.com/help/audio/ug/acoustic-scene-recognition-using-late-fusion.html>

We are just using the first 100 examples from the training set, with 15 scenes resulting in 1500 data points.

```

1  % load the subset of acoustic scene data we're using here
2  load("AcousticScenes-SmallTrain.mat");
3
4  c = cvpartition(trainLabels, 'HoldOut', 0.2);
5  trainSmall = xTrain(c.training, :);
6  testSmall = xTrain(c.test, :);
7

```

The Command Window shows the execution of the script:

```

>> load SelectedFeatures.mat
fx >>

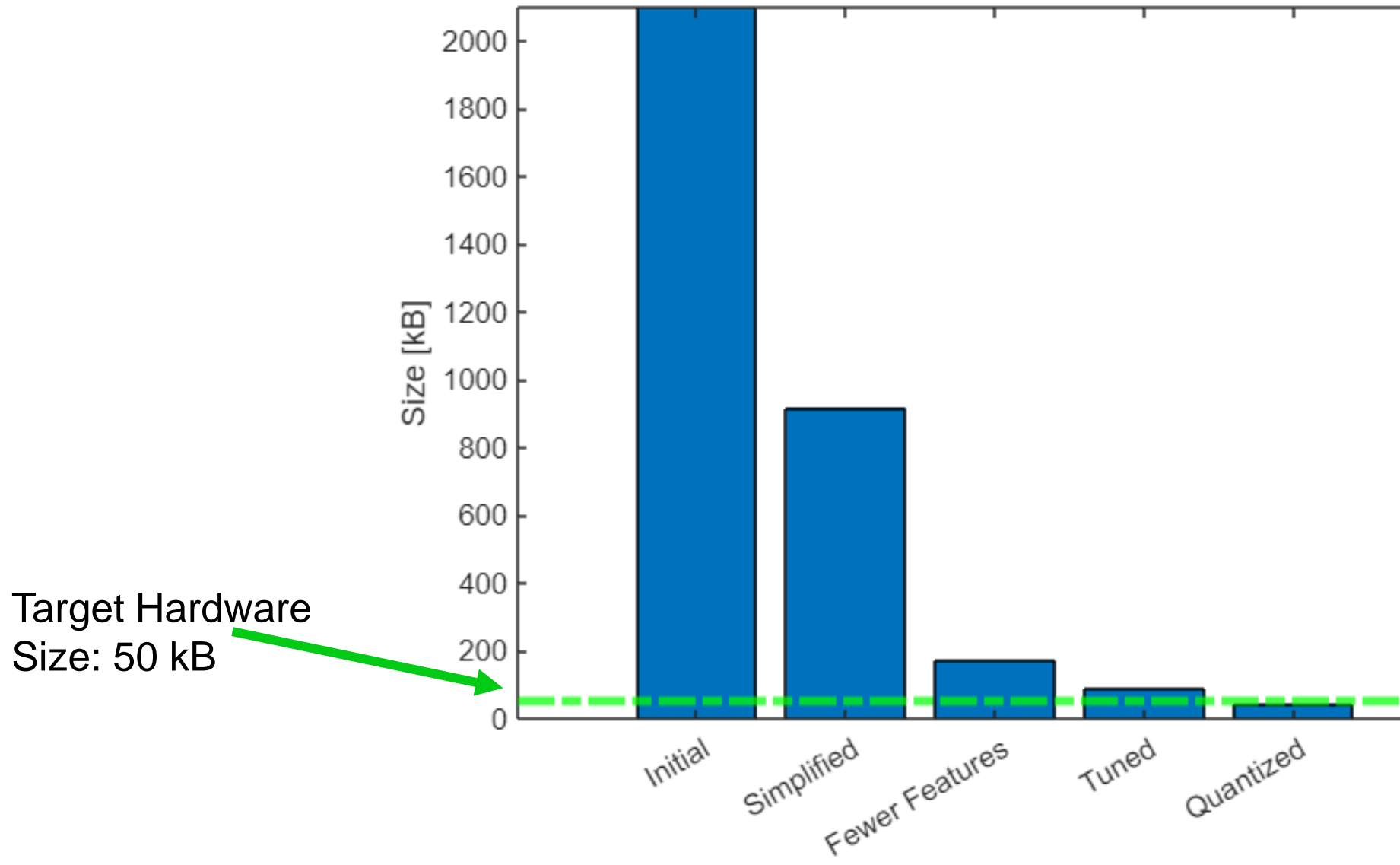
```

The Workspace window shows the following variables:

Name	Value
ans	'C:\Users\bsu...
c	1x1 cvpartition
centroids2	2x286 double
clusterIndi...	1200x1 double
testLabels	300x1 categor...
testSmall	300x286 dou...
trainLabels	1200x1 categor...
trainSmall	1200x286 do...
xTrain	1500x286 do...

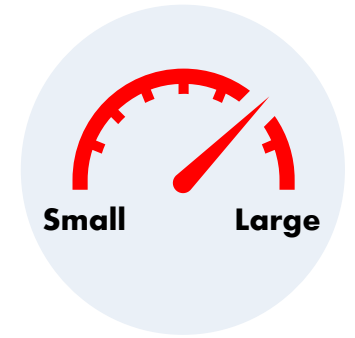
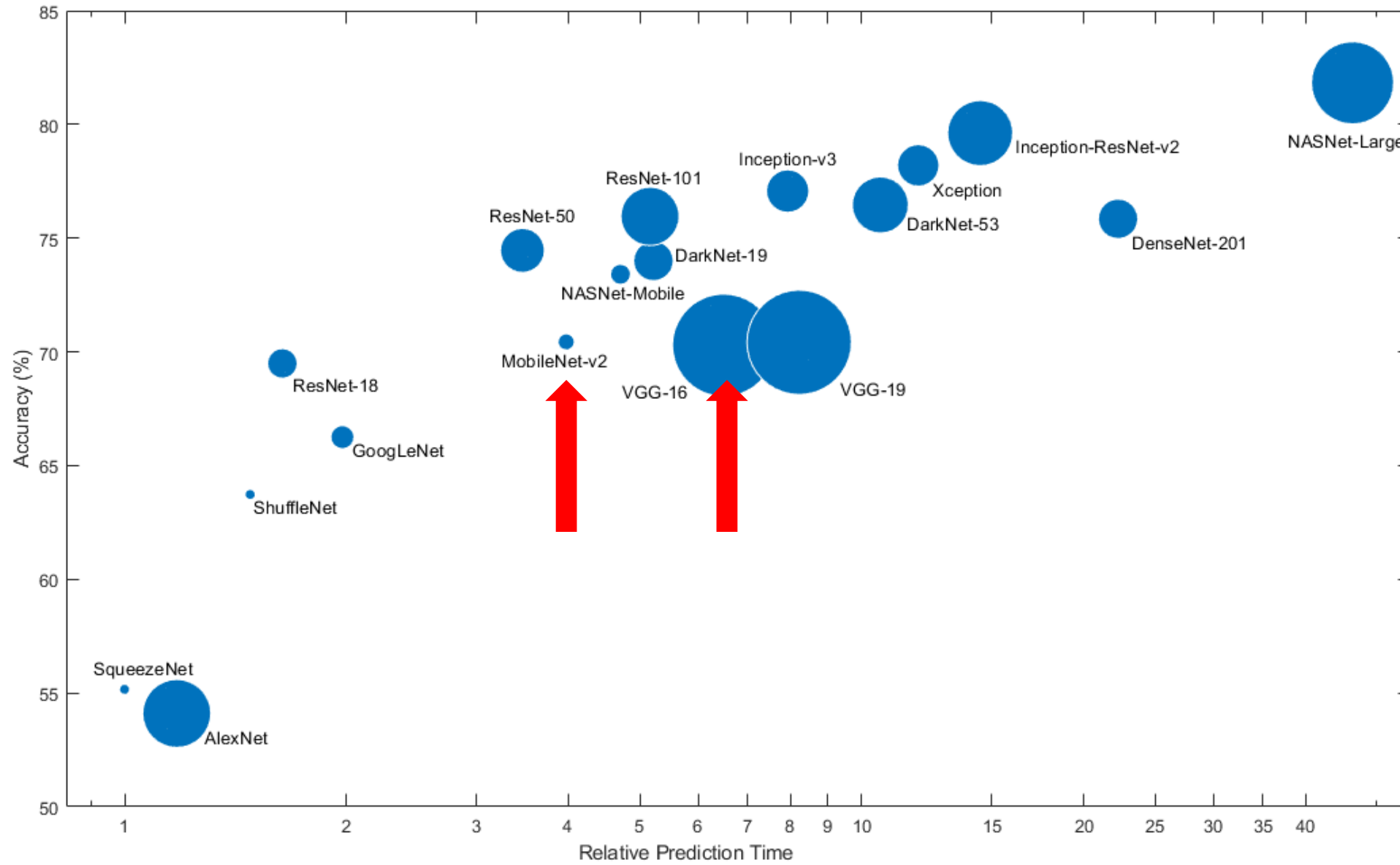
The status bar at the bottom shows Zoom: 100%, UTF-8, LF, and script.

Machine Learning Demo Size Reduction by factor 20



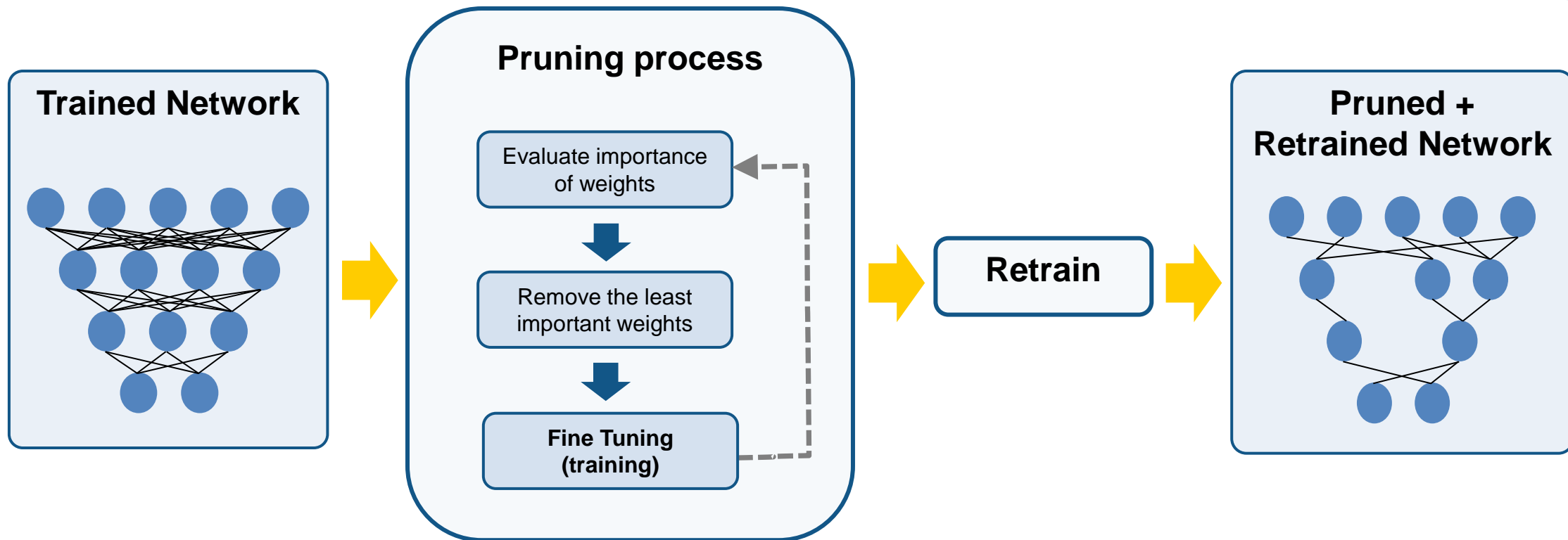
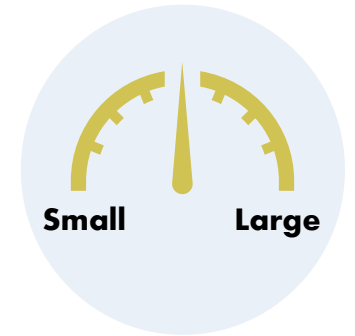
Compressing Deep Learning

Step ① Size aware model selection

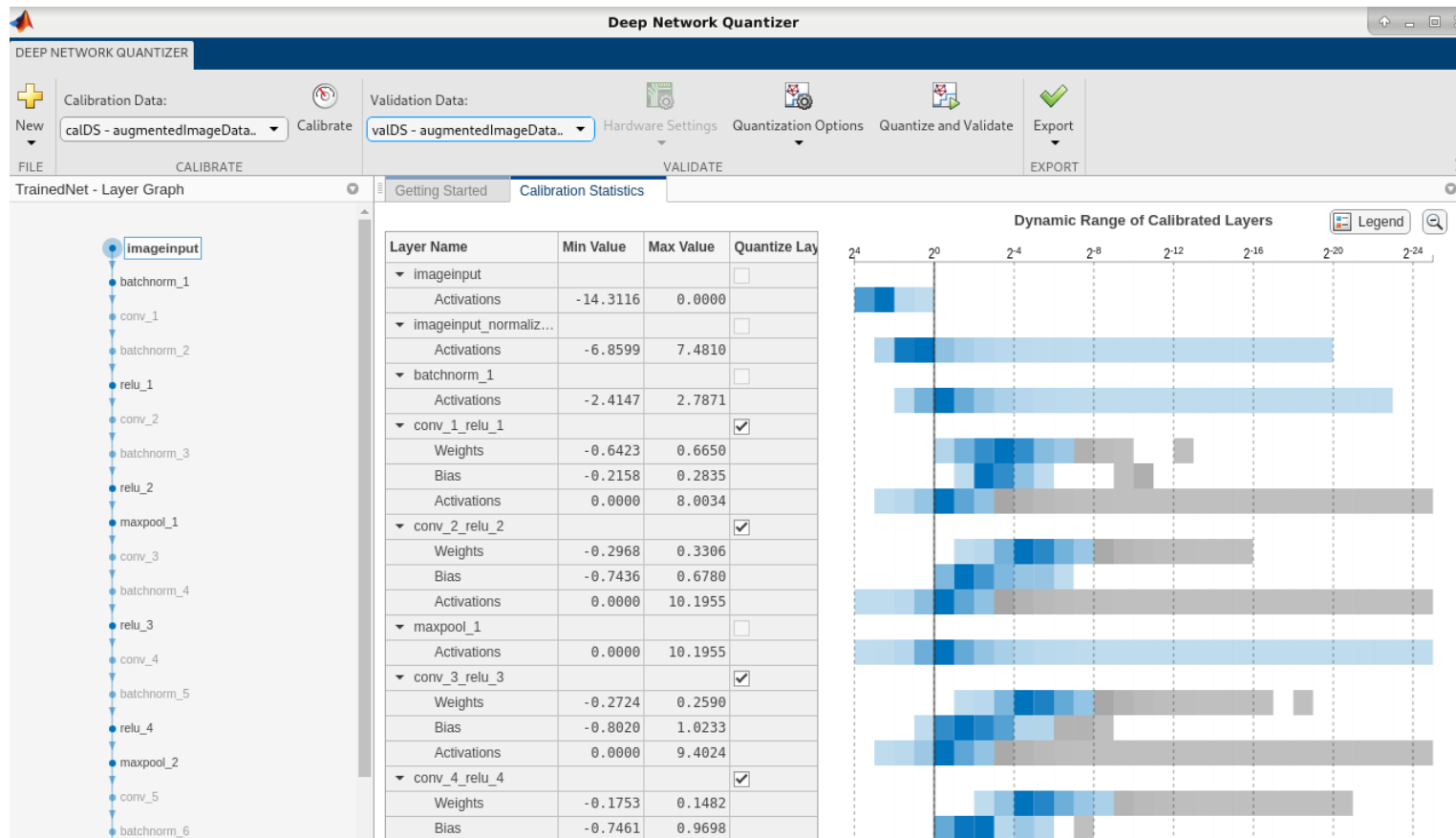


Step ② Smart pruning

Remove **unimportant** parts of the network



Step ③ Quantize your model



Deep Learning Demo: Scene classification

Classify 10 classes

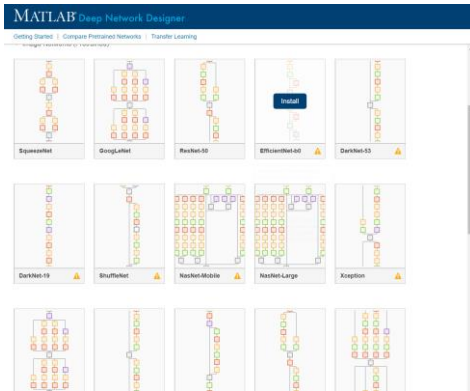
More difficult problem → more complex model



Functionality for Compressing Deep Neural Nets

1

Deep Network Designer



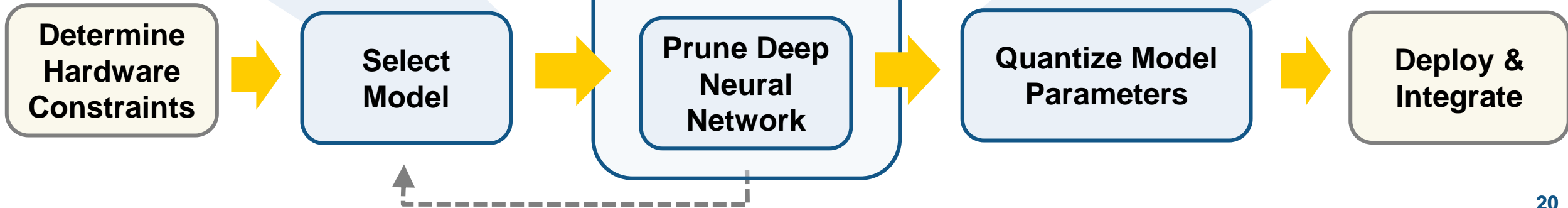
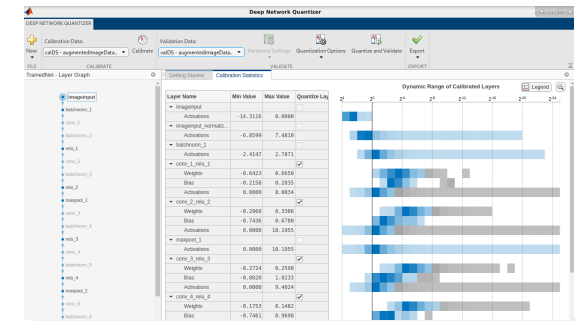
2

Taylor Pruning

```
taylorPrunableNetwork(net)
```

3

Deep Network Quantizer



HOME PLOTS APPS LIVE EDITOR INSERT VIEW

New Open Save Print Export Go To Find Bookmark

Text Normal B I U M Code Control Task Refactor Run Section Break Run and Advance Run Step Stop

FILE NAVIGATE TEXT CODE SECTION RUN

Search Documentation

Current Folder ▶ Scene identification demo ▶

Current Folder

- 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

```

1 load('trained10classNetwork'); |
2 load('data')
    
```

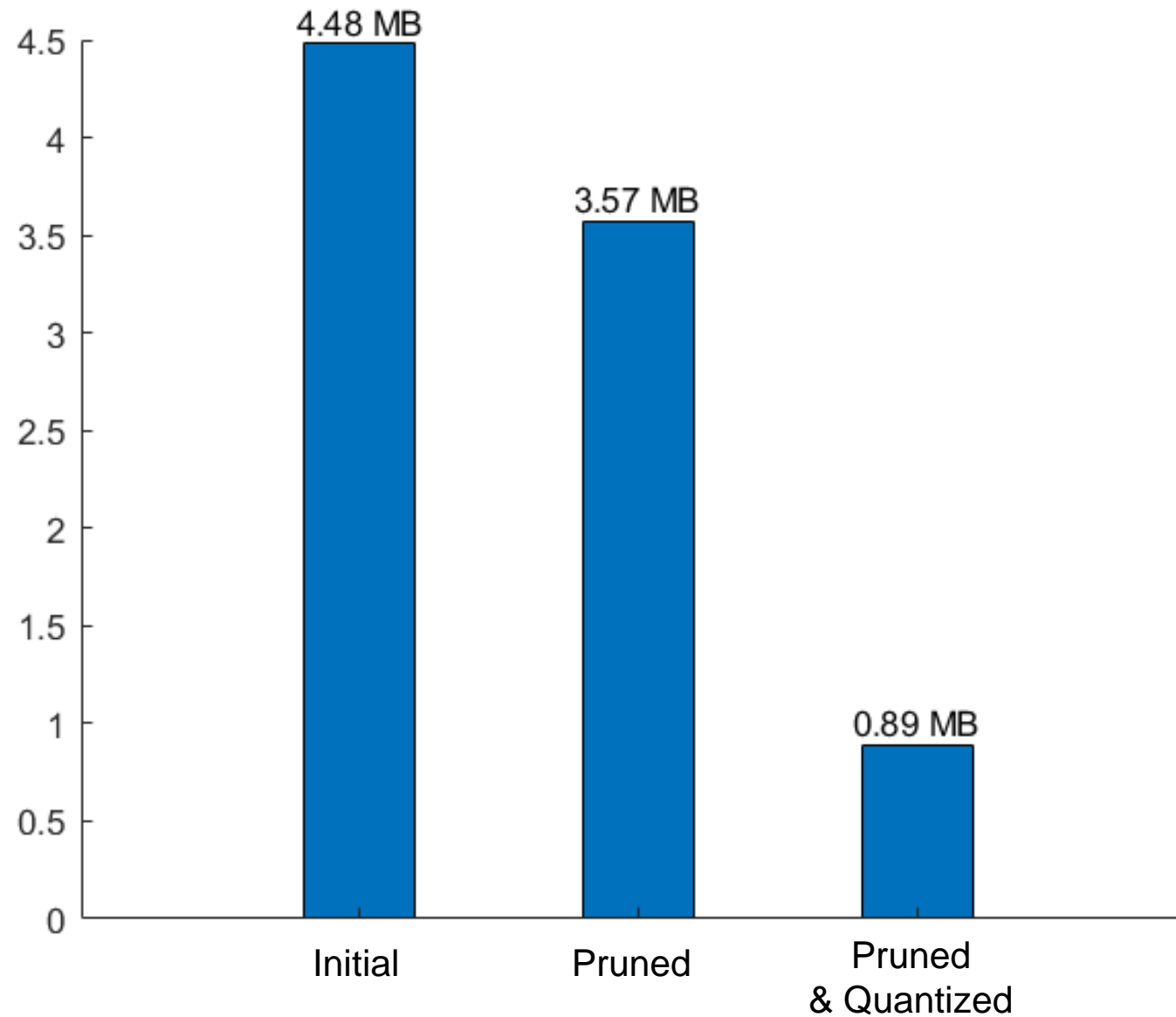
Note: Sounds have been converted to spectrograms

Workspace

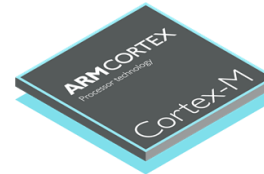
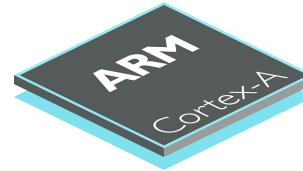
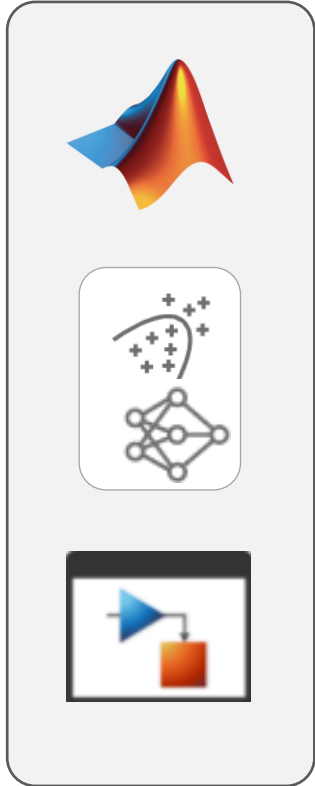
Name	Value

Command Window

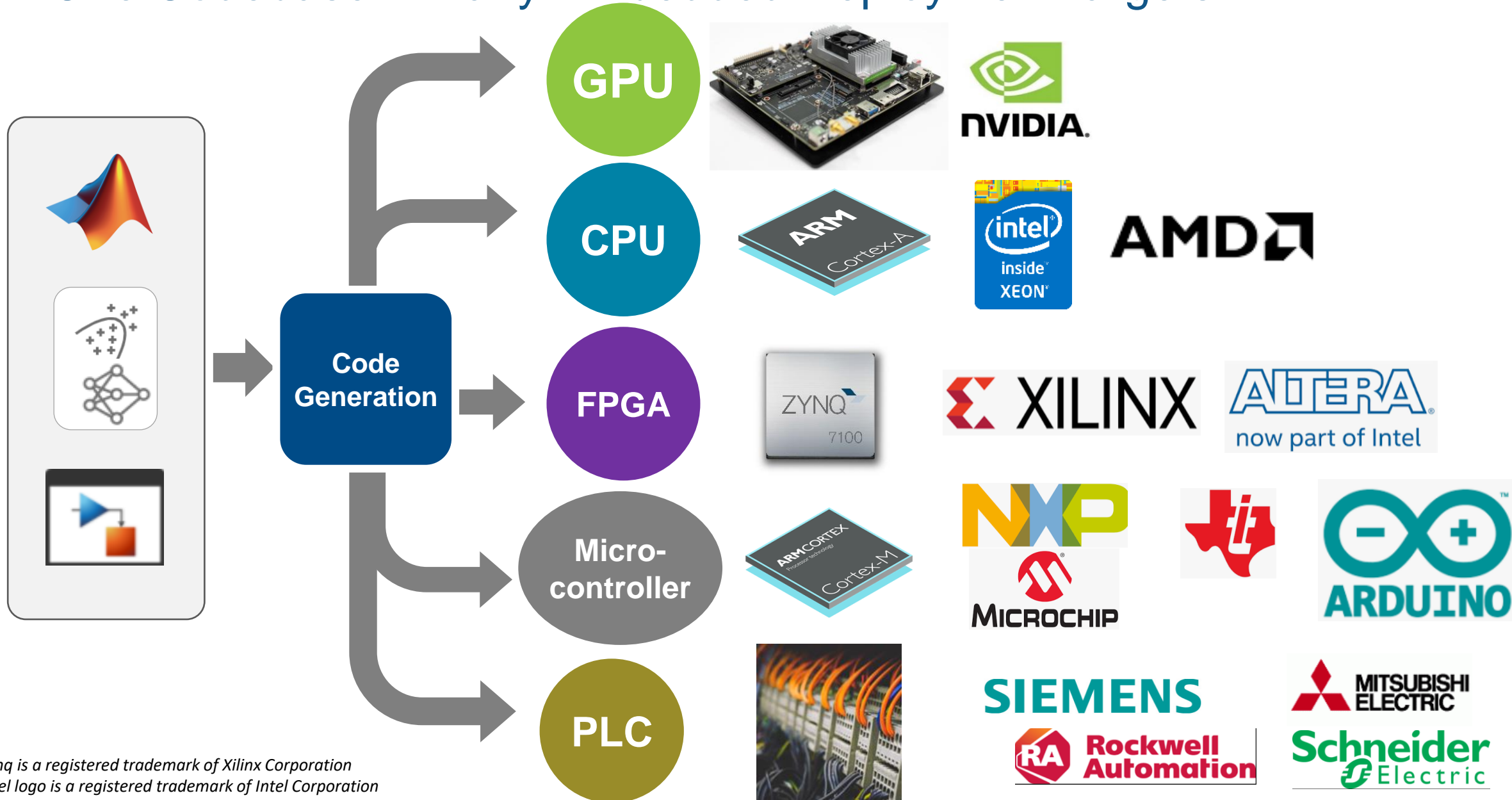
Deep Learning Demo Size Reduction by factor 5



One Codebase – Many Embedded Deployment targets



One Codebase – Many Embedded Deployment targets



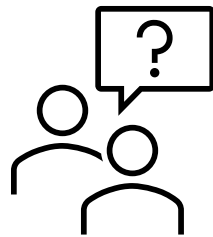
Zynq is a registered trademark of Xilinx Corporation
 Intel logo is a registered trademark of Intel Corporation

Conclusions

You can fit AI for many applications onto limited hardware

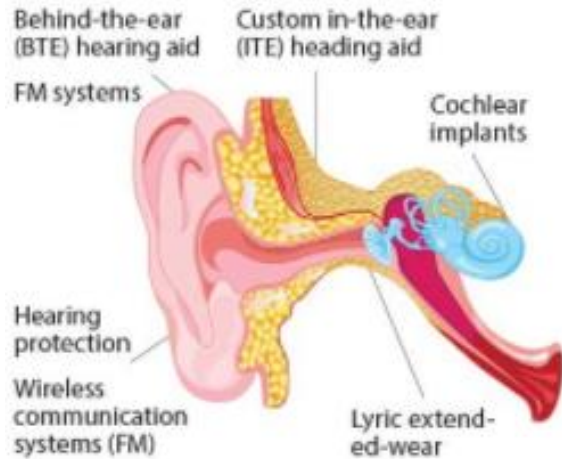
MathWorks tools make fitting AI models on constrained hardware a lot easier

Same high-level Workflow for any type of AI



Which constraints are most challenging for your application?

Learn More



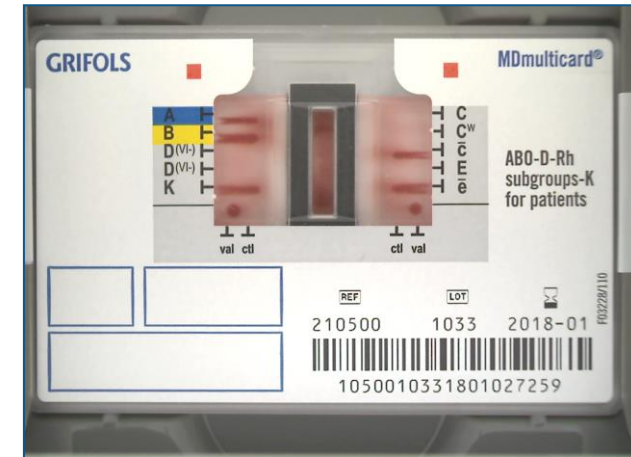
[Hearing Implant using MBD](#)

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

