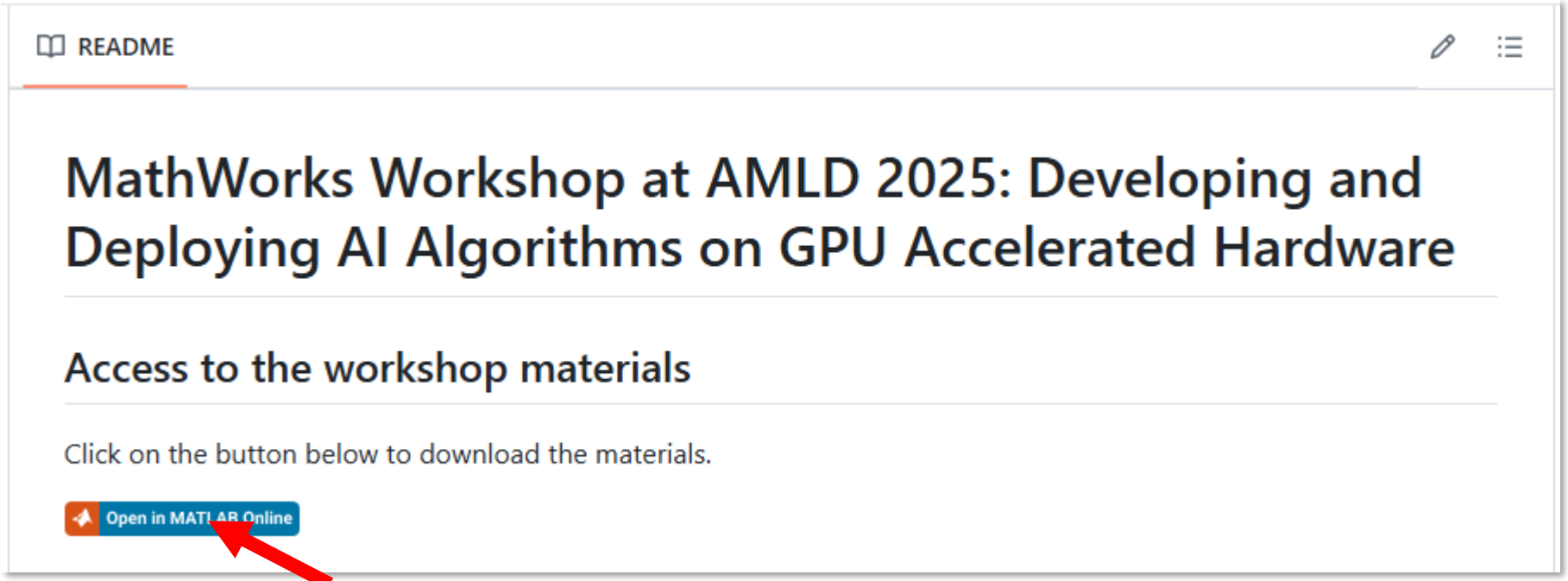
Setup for Workshop

Requirements:
- MathWorks Account
- MATLAB Online: use supported browser (Google Chrome or Microsoft Edge)

- Get the workshop materials: https://github.com/resjoehr/AMLD_2025
- Access the MATLAB Online license: https://tinyurl.com/AMLD-2025



Developing and Deploying AI Algorithms on GPU Accelerated Hardware

Workshop at Applied Machine Learning Days, 14.02.2025

MathWorks®

Speakers

Res Jöhr
Senior Application Engineer (Universities)
resjoehr@mathworks.com

Sebastien Dupertuis
Senior Application Engineer

Franziska Albers
Senior Application Engineer (Universities)

Agenda

14:00 - 14:30 **Overview of AI Deployment Workflow**


Res JÖHR, Switzerland

14:30 - 15:30 **Data Preparation and Model Development**


Franziska ALBERS, Germany

15:30 - 16:00 **Coffee break**

16:00 - 16:45 **Model Verification and Code Generation**


Sebastien DUPERTUIS, Switzerland


Res JÖHR, Switzerland

16:45 - 17:30 **Model Deployment to GPU Accelerated Hardware**


Sebastien DUPERTUIS, Switzerland


Res JÖHR, Switzerland

https://forms.office.com/r/ZdHD3p4hpS

Technical Setup

```
>>    answer = Do you have a MathWorks account?

>>    If answer==true
            Follow the short intro;
      else
            Please go to mathworks.com/login and register;
      end

>>    Access workshop materials after introduction
```

About MathWorks

We at MathWorks believe in the importance of engineers and scientists. They increase human knowledge and profoundly improve our standard of living.

We created MATLAB and Simulink to help them do their best work.

6500+ staff

34 offices worldwide

Privately held

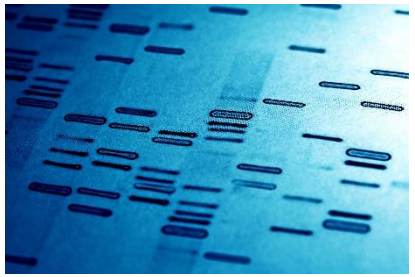
and profitable every year

MATLAB is used to design the products we rely on every day


Aerospace and Defense


Automotive


Biological Sciences


Biotech and Pharmaceutical


Communications


Electronics


Energy Production


Financial Services

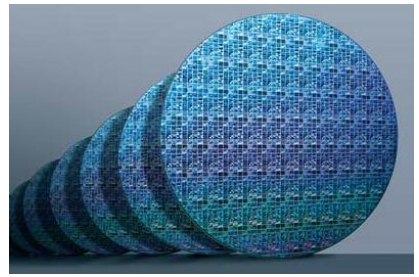

Industrial Machinery


Medical Devices


Process Industries

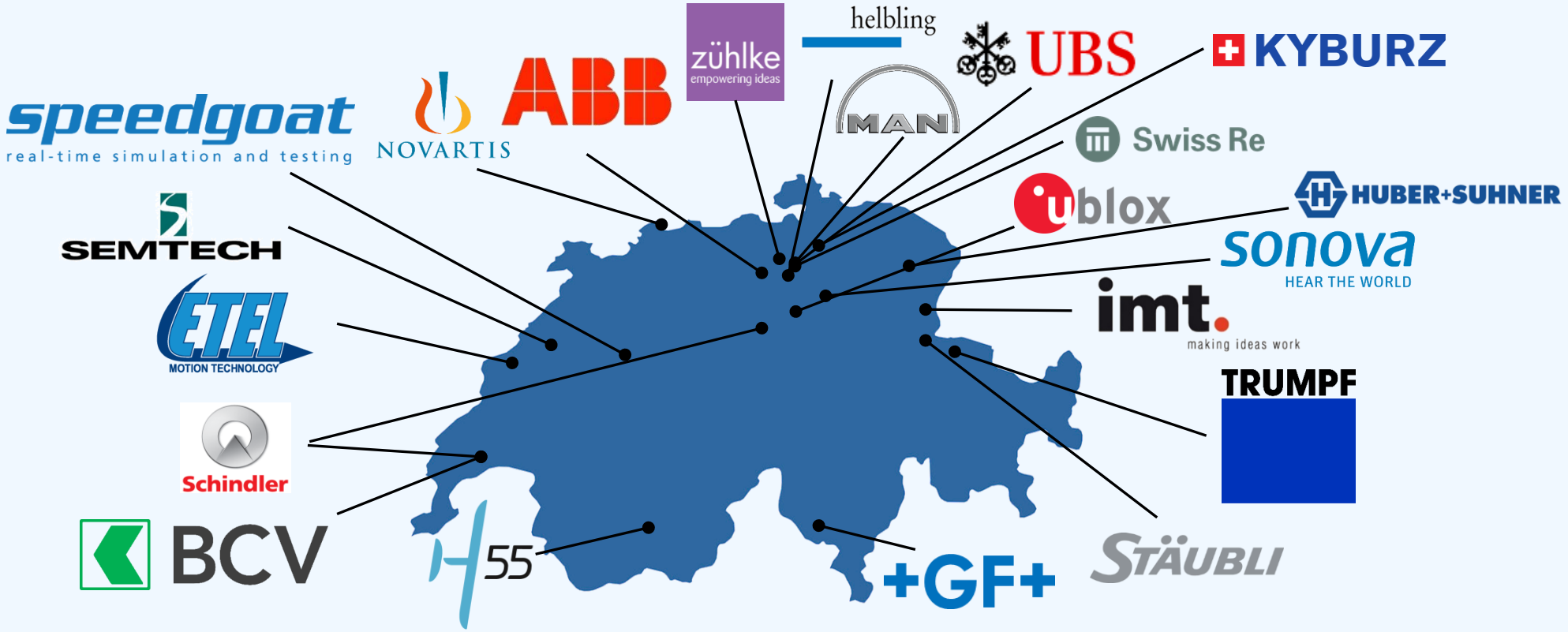

Neuroscience

Railway Systems


Semiconductors

Software and Internet

Commercial Usage in Switzerland



100,000+ business, government, and university sites

Top 10 auto manufacturers[1]

Top 10 aerospace companies[2]

Three of the top five internet companies

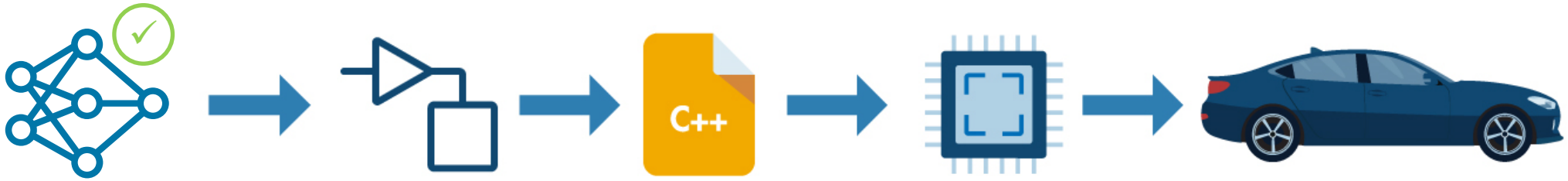
[1]OICA: *2016 World Motor Vehicle Production*

[2]PwC: *Aerospace and Defense 2017 Year in Review*

Integrating AI Algorithms into embedded Hardware

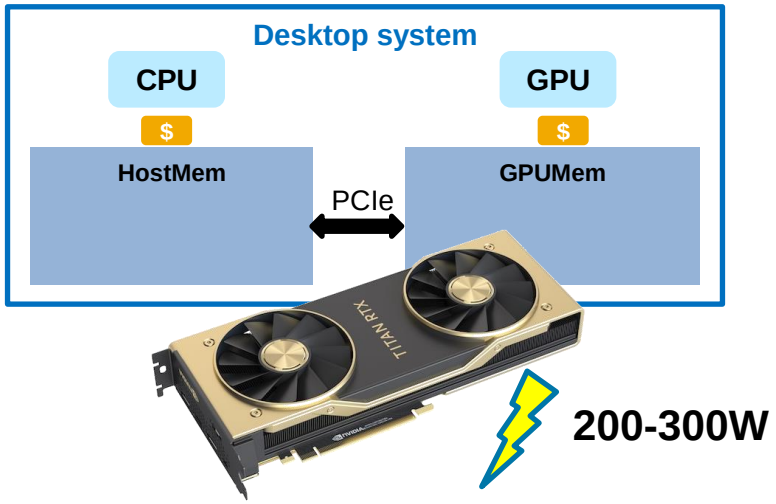
- Motivation:
  - AI algorithms can handle complex tasks that can hardly be solved using classical methods
    - Would like to use these algorithms on edge devices

- Issues:
  - How to integrate models into larger systems and deploy them to embedded hardware
    - Availability of computational resources (processor, memory)
    - Translation of code to hardware language (code generation)
  - AI models are sometimes hard to understand (black box, interpretability, robustness)
  - Models have to be verified and validated

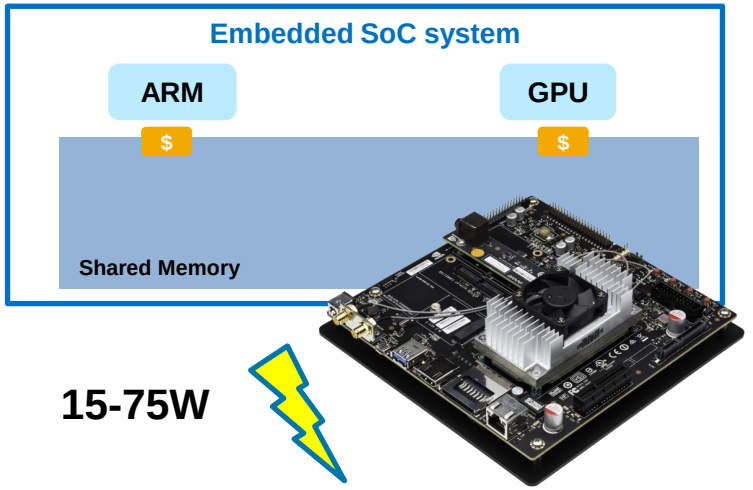


Types of GPUs

- Desktop GPUs
  - Designed for performance (more memory and cores)
  - For interactive tasks
    (e.g. simulations, video editing, 3D rendering)



- Embedded GPUs
  - Designed for integration into compact or power constrained environments
  - Optimized for specific tasks
    (e.g. real-time processing and AI inference)
  - Often requires compiled code
    (CUDA code needed to leverage benefits of NVIDIA GPUs)



Drass Develops Deep Learning System for Real-Time Object Detection in Maritime Environments

Challenge

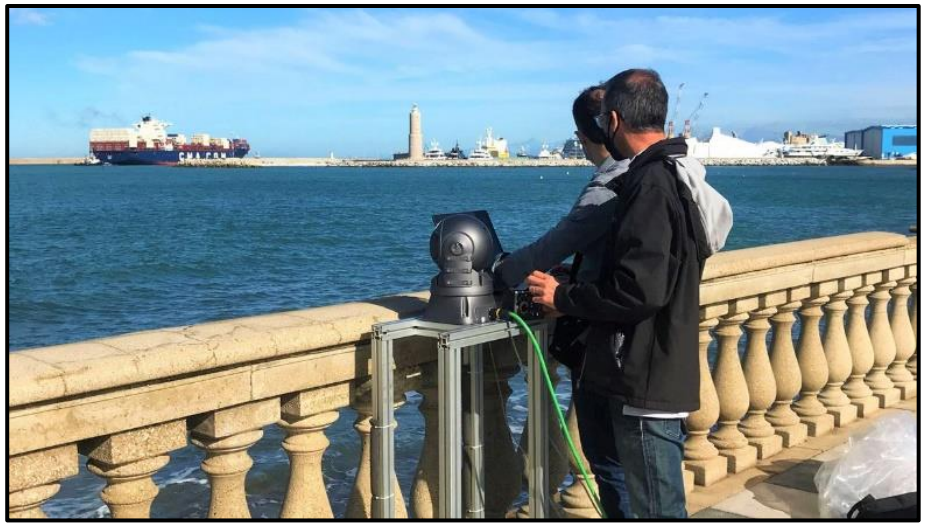
Help ship operators monitor sea environments and detect objects, obstacles, and other ships

Solution

Create an object-detection deep learning model that can be deployed on ships and run in real time

Results

- Data labeling automated
- Development time reduced
- Flexible and reproducible framework established



First day of object detection tests with optronic system prototype.

"From data annotation to choosing, training, testing, and fine-tuning our deep learning model, MATLAB had all the tools we needed—and GPU Coder enabled us to rapidly deploy to our NVIDIA GPUs even though we had limited GPU experience."

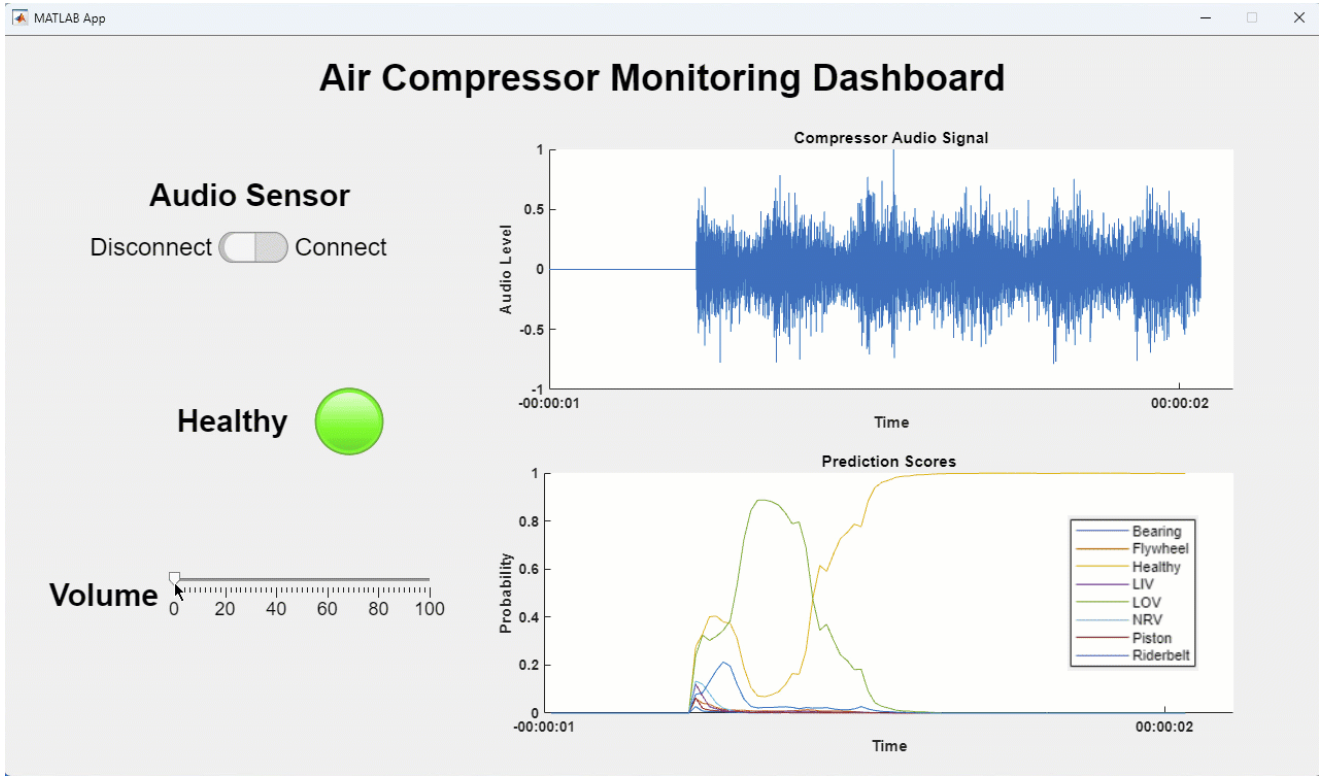
- Valerio Imbriolo, Drass Group

Link to user story

Example: Classifying Sounds

Build, test, and deploy a deep learning solution that can classify sequences in signal data.

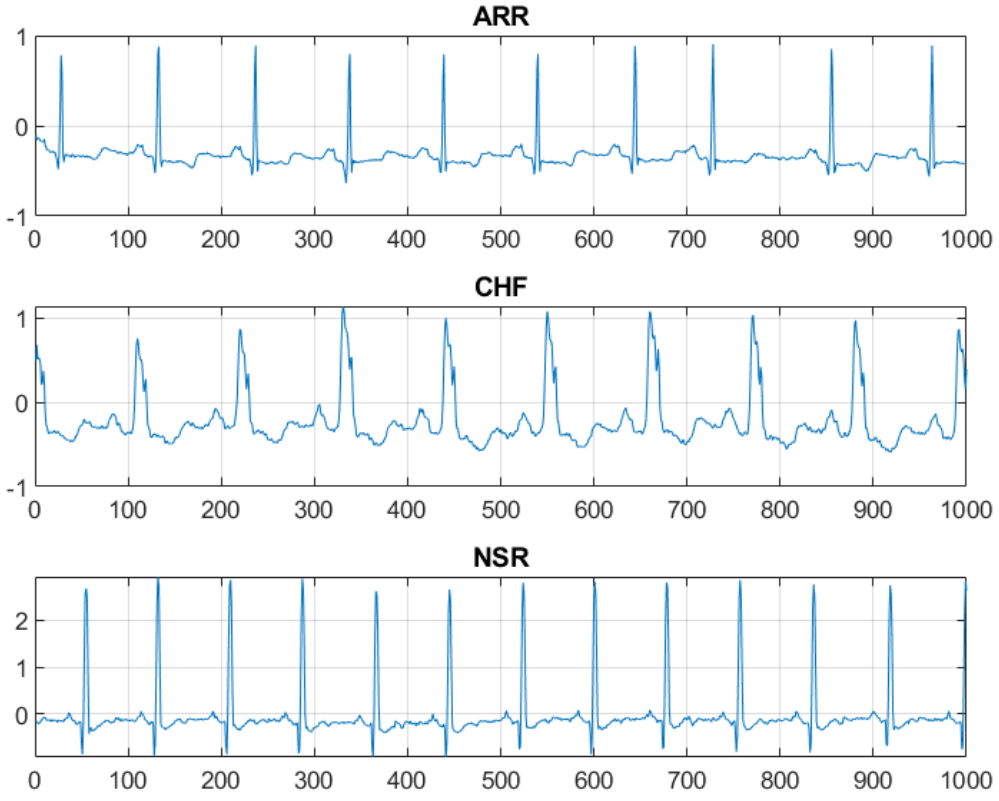
- **Goal:** Detect faults in an air compressors based on their noise



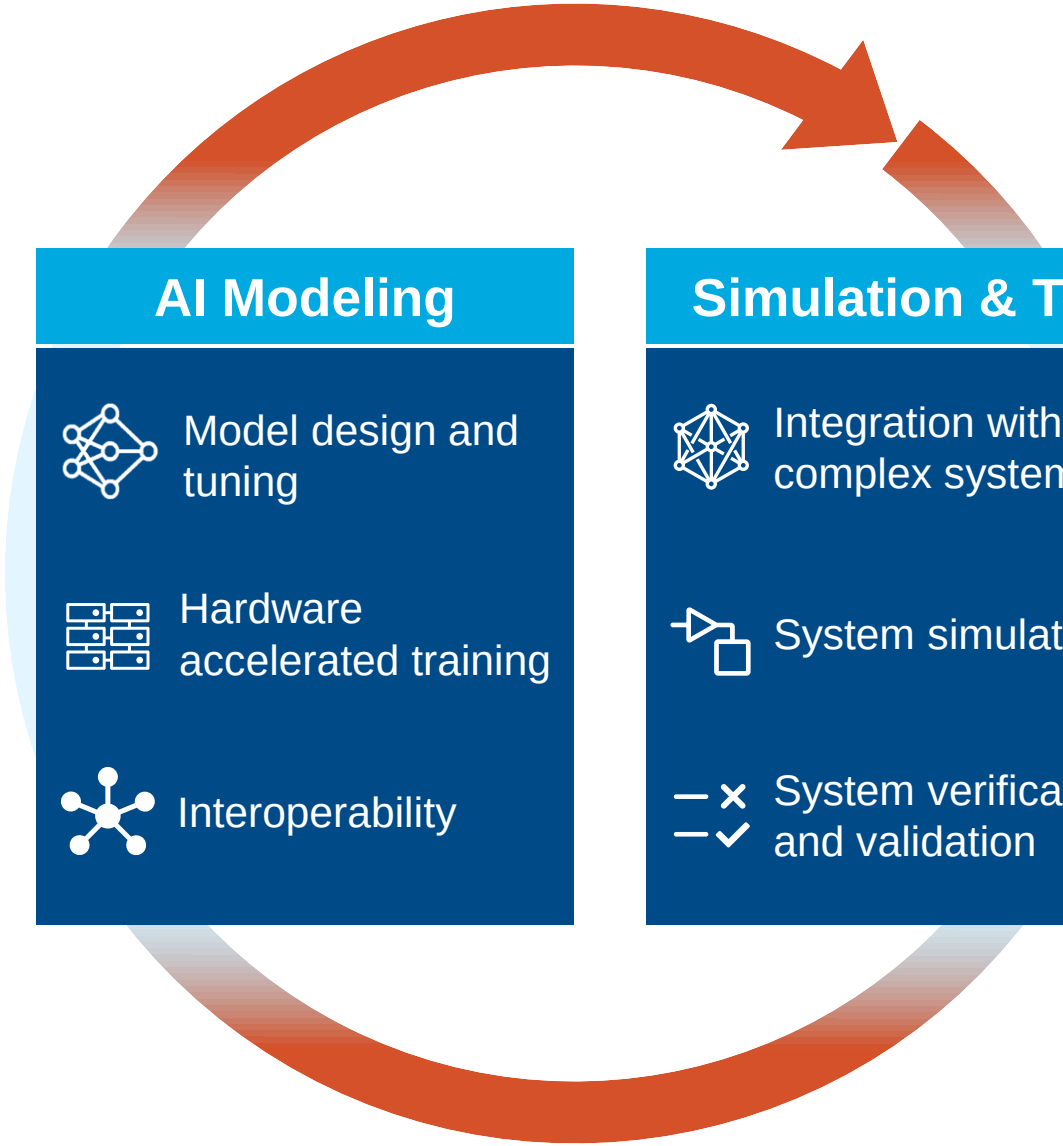
14

Today's use case: ECG Classification

- Dataset contains 3 classes of heartbeats
  - cardiac arrhythmia (ARR, N=96)
  - congestive heart failure (CHF, N=30)
  - normal sinus rhythms (NSR, N=36)
- PhysioNet databases:
  - MIT-BIH Arrhythmia Database
  - MIT-BIH Normal Sinus Rhythm Database
  - The BIDMC Congestive Heart Failure Database
- References for the dataset
- Timeseries signals with 128Hz
- 162 recordings with 65535



AI-Driven System Design



Data Preparation

- Data cleansing and preparation
- Human insight
- Simulation-generated data

AI Modeling

- Model design and tuning
- Hardware accelerated training
- Interoperability

Simulation & Test

- Integration with complex systems
- System simulation
- System verification and validation

Deployment

- Embedded devices
- Enterprise systems
- Edge, cloud, desktop

In this workshop we will …

ECG Data having 3 classes

Import/Read Data

Preprocessing Scalogram

Classification using CNNs

Feature Extraction (Wavelet Scattering)

Classification using LSTMs

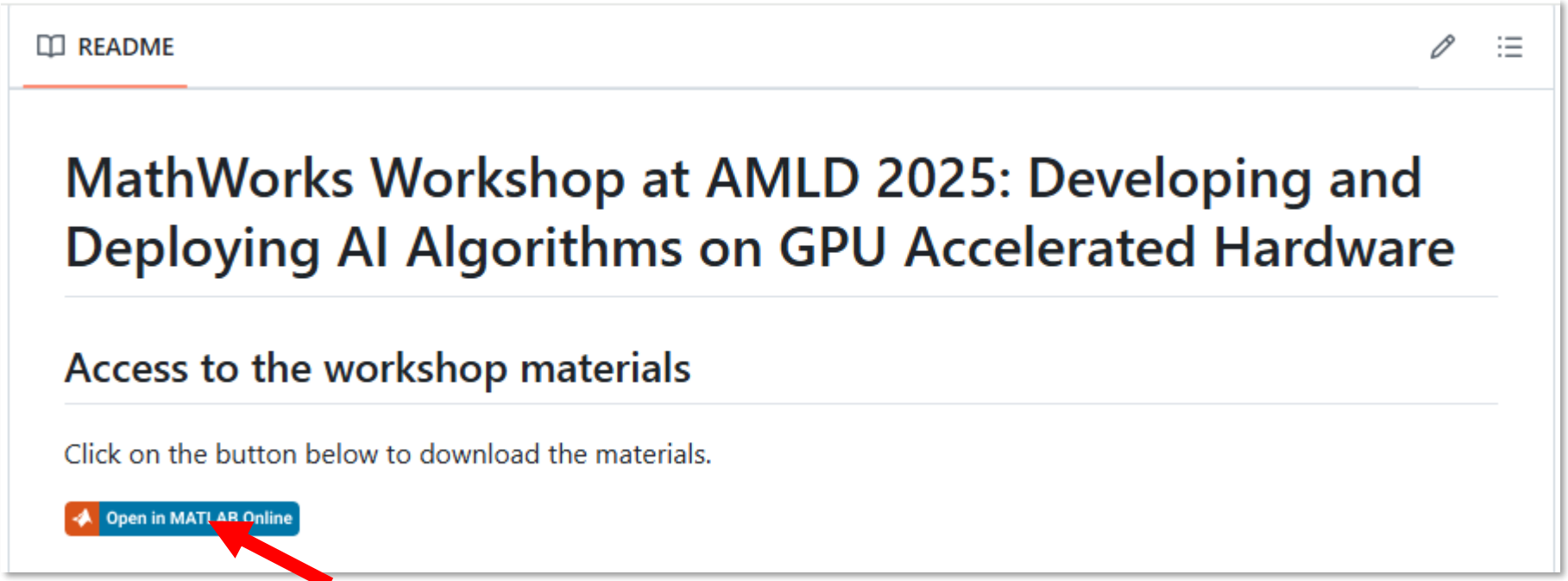
Compression Validation

Codegen and Deployment

Setup Check-In

Requirements:
- MathWorks Account
- MATLAB Online: use supported browser (Google Chrome or Microsoft Edge)

- Get the workshop materials: https://github.com/resjoehr/AMLD_2025
- Access the MATLAB Online license: https://tinyurl.com/AMLD-2025



AI-Driven System Design

Data Preparation

- Data cleansing and preparation
- Human insight
- Simulation-generated data

AI Modeling

- Model design and tuning
- Hardware accelerated training
- Interoperability

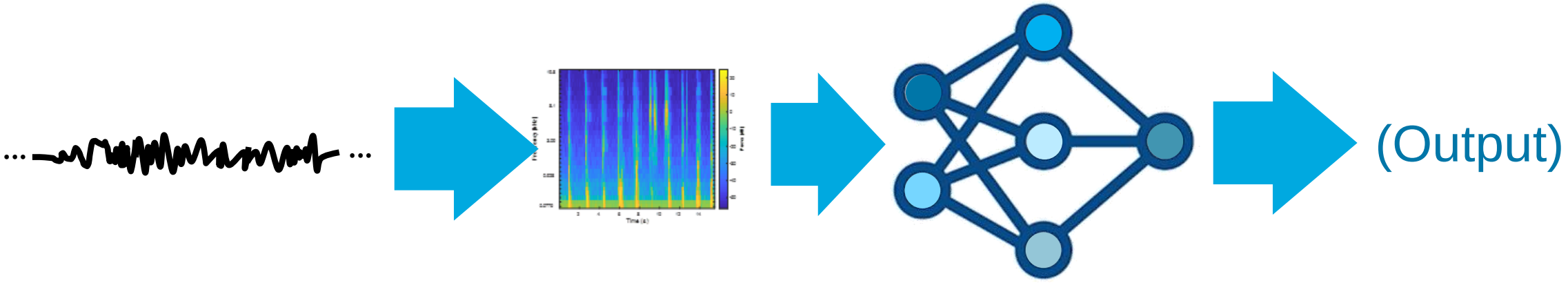
Simulation & Test

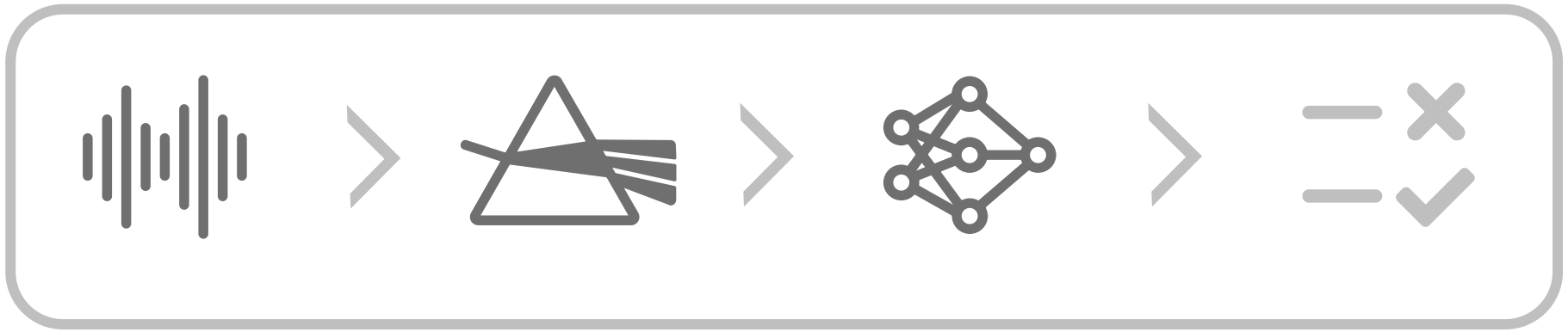
- Integration with complex systems
- System simulation
- System verification and validation

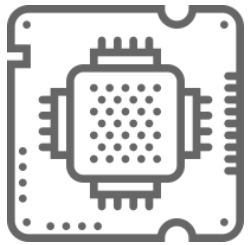
Deployment

- Embedded devices
- Enterprise systems
- Edge, cloud, desktop

Deep networks most often don’t learn directly from raw signals

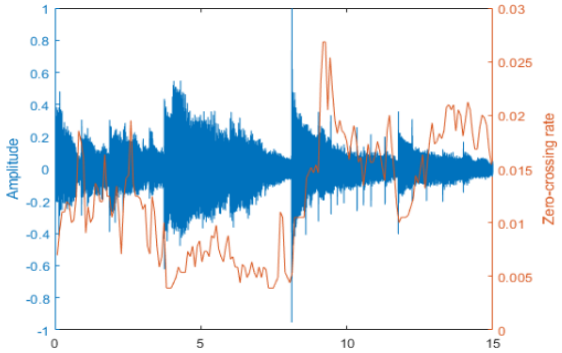




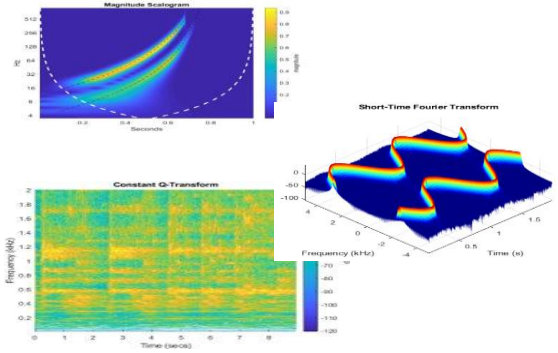


Data Preparation | AI Modeling | Simulation & Test | Deployment

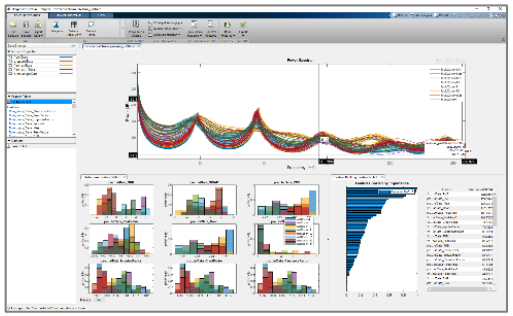
Transform raw data for useful modeling and analysis



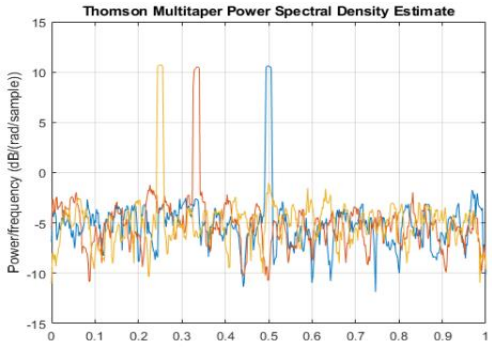
`signalTimeFeatureExtractor`



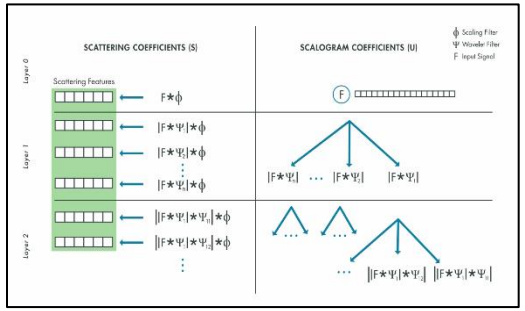
`signalFrequencyFeatureExtractor`



`diagnosticFeatureDesigner`



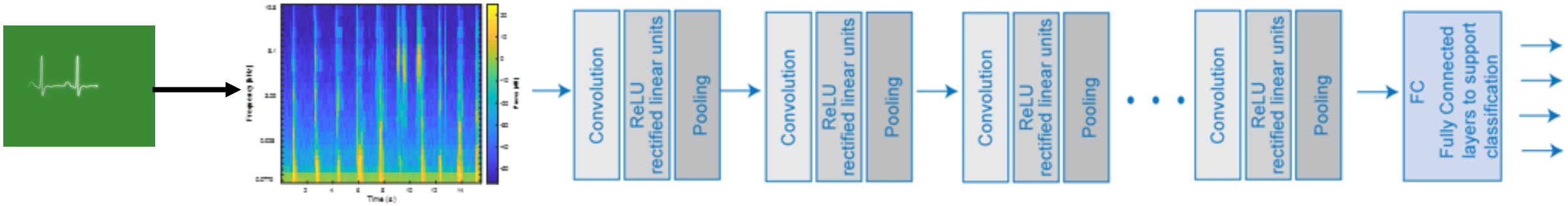
`signalTimeFrequencyFeatureExtractor`



`Wavelet based features`

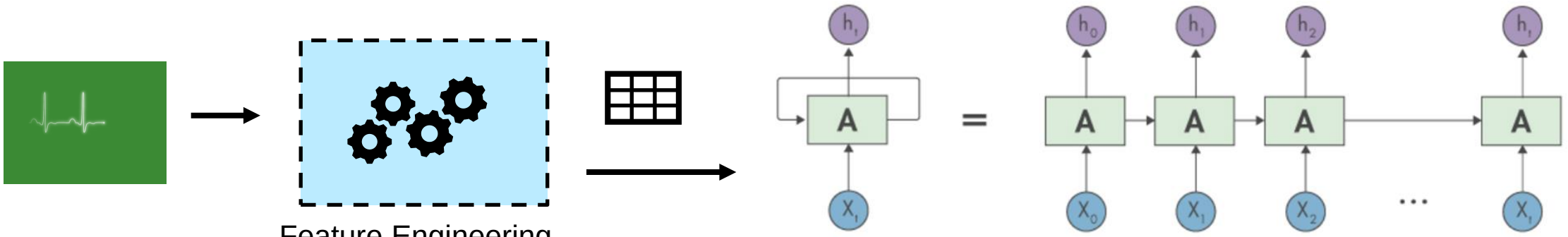
Common Network Architectures – Signal Processing

Convolutional Neural Networks (**CNN**)



Time-Frequency Transformation

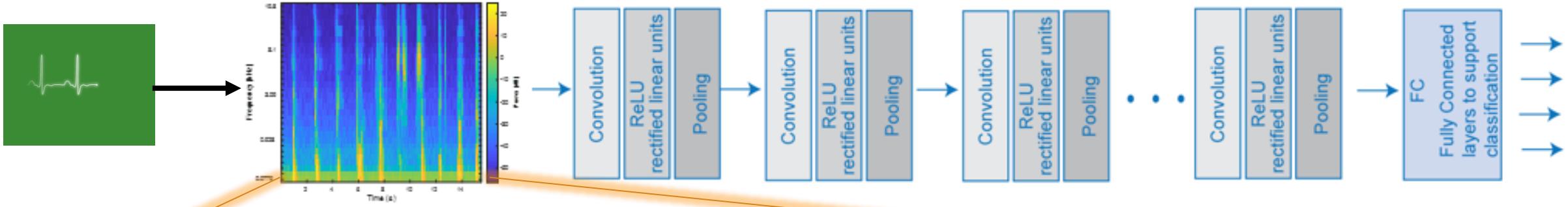
Long Short-Term Memory (**LSTM**) Networks



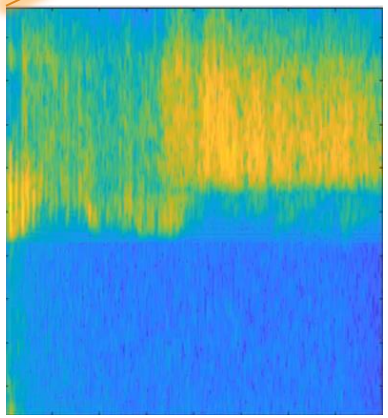
Data Preparation | **AI Modeling** | **Simulation & Test** | **Deployment**

Common Network Architectures – Signal Processing

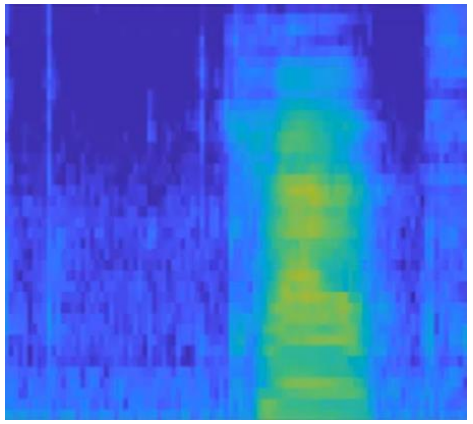
Convolutional Neural Networks (**CNN**)



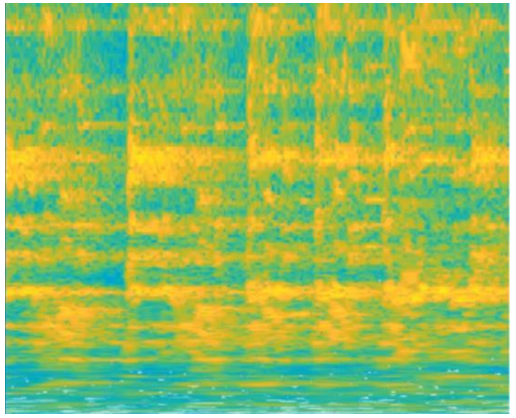
Time-Frequency Transformation



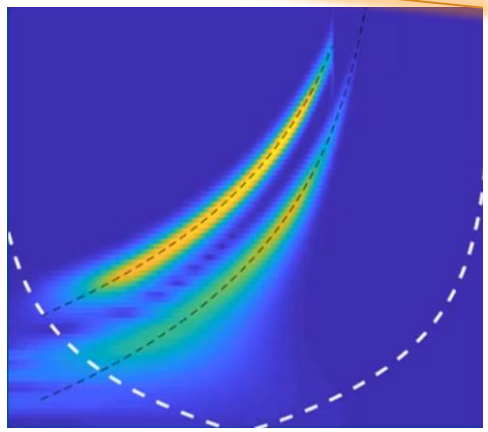
Basic spectrogram



Perceptually-spaced (e.g., Mel, Bark) Spectrogram



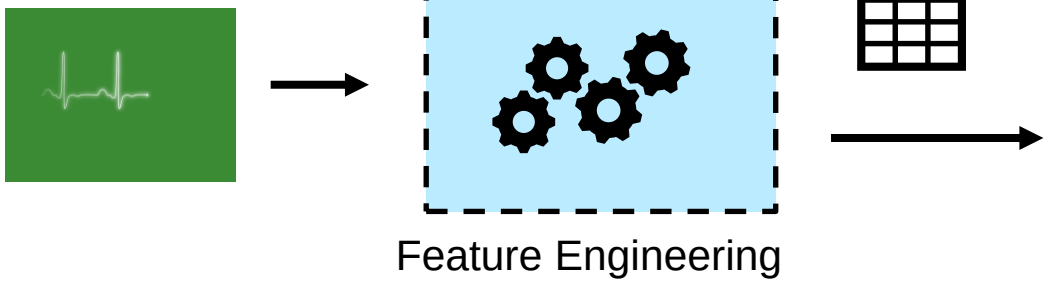
Constant Q transform



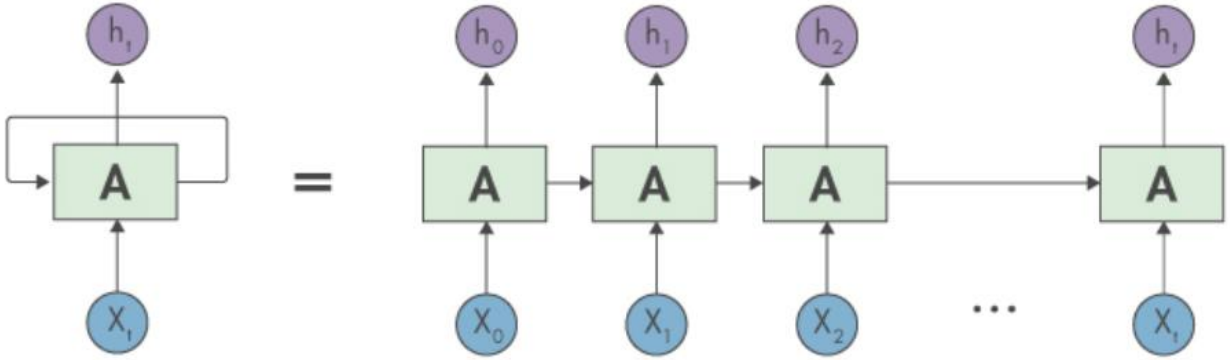
Wavelet scalogram

mathworks.com/help/signal/ug/time-frequency-gallery.html

Common Network Architectures – Signal Processing



Long Short-Term Memory (**LSTM**) Networks



mathworks.com/discovery/lstm.html

Wavelet Scattering Basic Idea

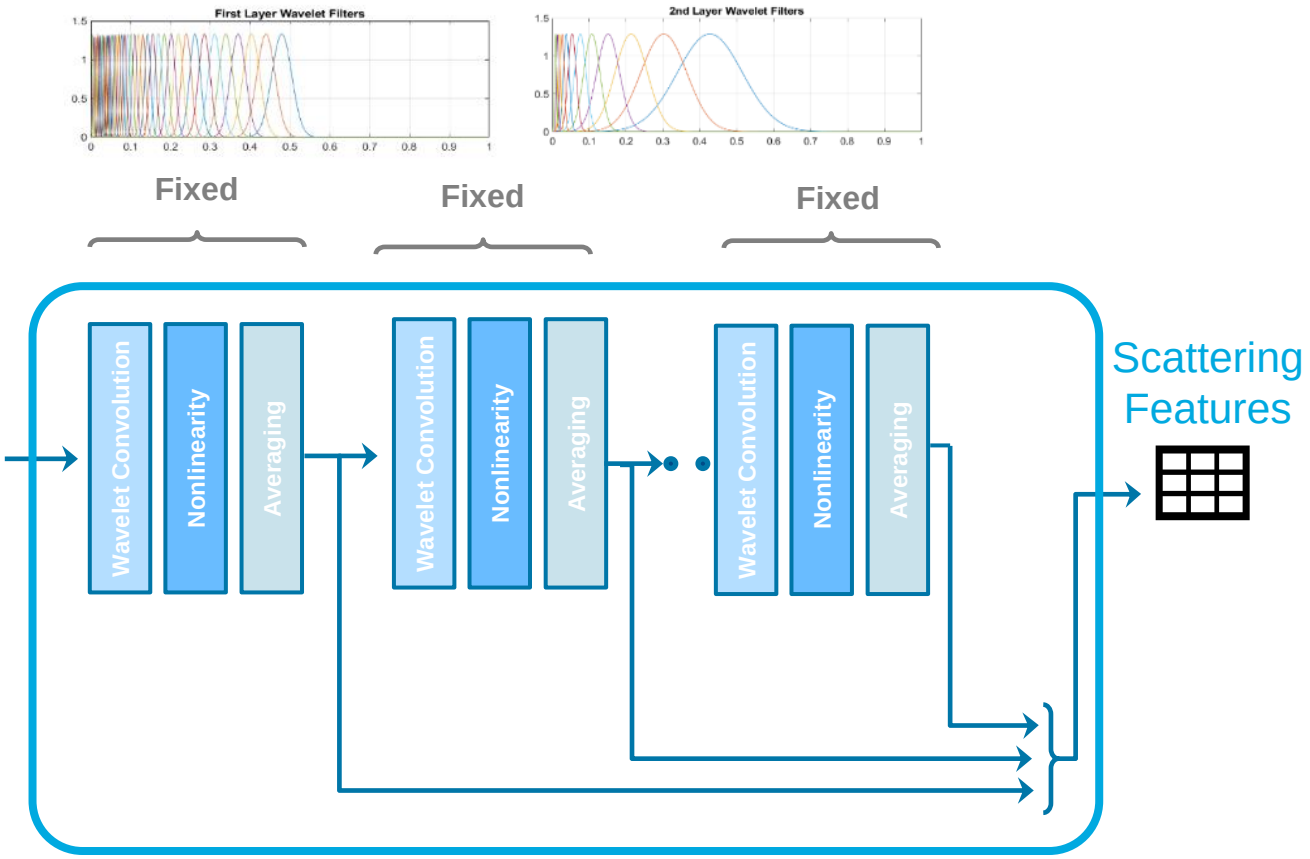
- method to generate features from 1D and 2D input data with minimal configuration
- produces data representations that
  - minimize differences *within* a class
  - while preserving discriminability *across* classes
- works with a small amount of data

Additional Resources:

Wavelet scattering Tech talk [4 min video]
Wavelet scattering for ECG [doc example]
Blog about Wavelet scattering

Deep Feature Extraction with Wavelets

- `waveletScattering`
  - Uses structure of Wavelet-based convolutional layers



  - State-of-the-art results:
    - ECG classification
    - Music genre classification
    - Spoken digit recognition
    - Acoustic scene recognition

Wavelet Scattering extracts features in a similar way to CNNs

CNN

- Step 1: Convolution
  - **Learn** convolution filters during training
- Step 2: Apply non-linearity
  - Use ReLu function
- Step 3: Apply averaging
  - Use pooling function

Wavelet Scattering

- Step 1: Convolution
  - Use wavelets as **fixed** convolution kernels
- Step 2: Apply non-linearity
  - Use **modulus function** (= calculate absolute value)
- Step 3: Apply averaging
  - Convolve with **scaling function** to apply time averaging

https://dsp.stackexchange.com/questions/78512/wavelet-scattering-explanation/78513#78513

AI-Driven System Design

Data Preparation

- Data cleansing and preparation
- Human insight
- Simulation-generated data

AI Modeling

- Model design and tuning
- Hardware accelerated training
- Interoperability

Simulation & Test

- Integration with complex systems
- System simulation
- System verification and validation

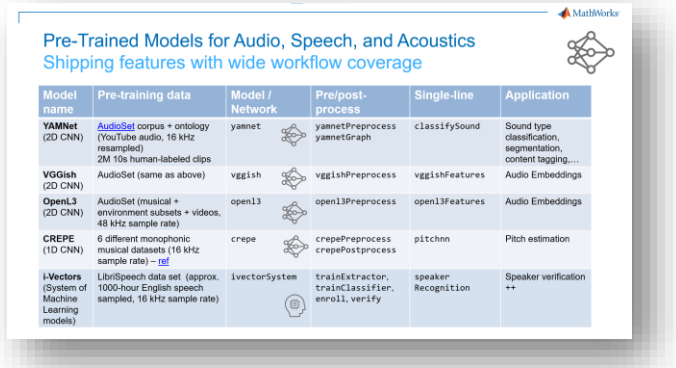
Deployment

- Embedded devices
- Enterprise systems
- Edge, cloud, desktop

Deep Learning and Machine Learning Models for Signals

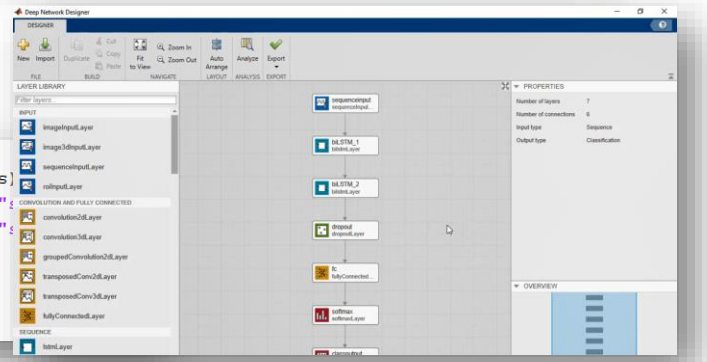
Three ways to get started in MATLAB

1. Start from a model available in MATLAB

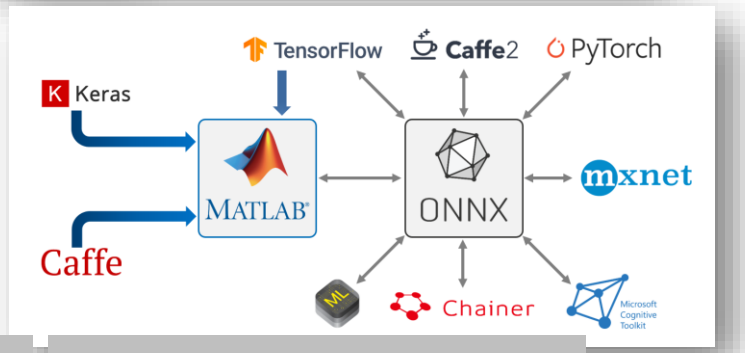


2. Define your own from scratch

```
layers = [ ...
    sequenceInputLayer(numFeatures)
    bilstmLayer(150,"OutputMode","
    bilstmLayer(150,"OutputMode","
    fullyConnectedLayer(2)
    softmaxLayer
    classificationLayer
    ];
```



3. Import one developed using another framework



Data Preparation | **AI Modeling** | Simulation & Test | Deployment

Available Models in MATLAB

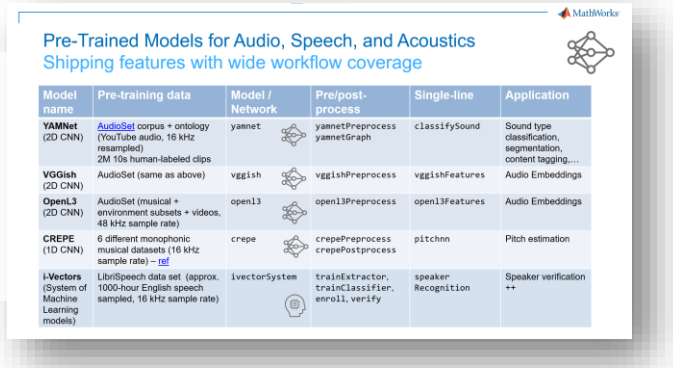
Train in MATLAB's Deep Learning Framework

Data Preparation | **AI Modeling** | Simulation & Test | Deployment

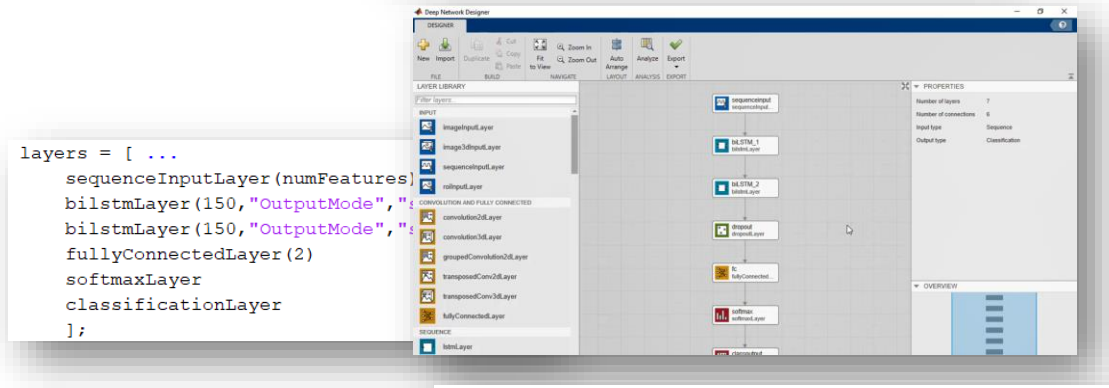
Deep Learning and Machine Learning Models for Signals

Three ways to get started in MATLAB

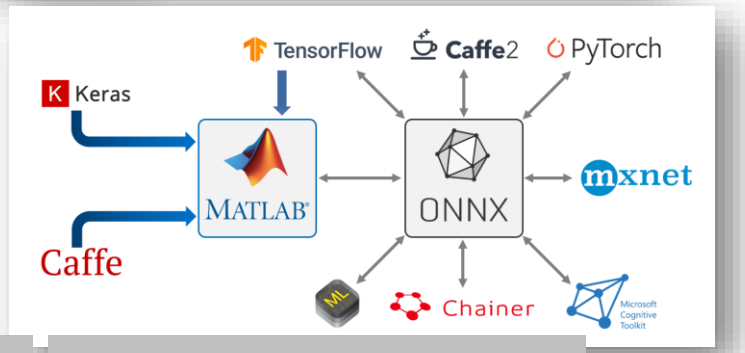
1. Start from a model available in MATLAB



2. Define your own from scratch



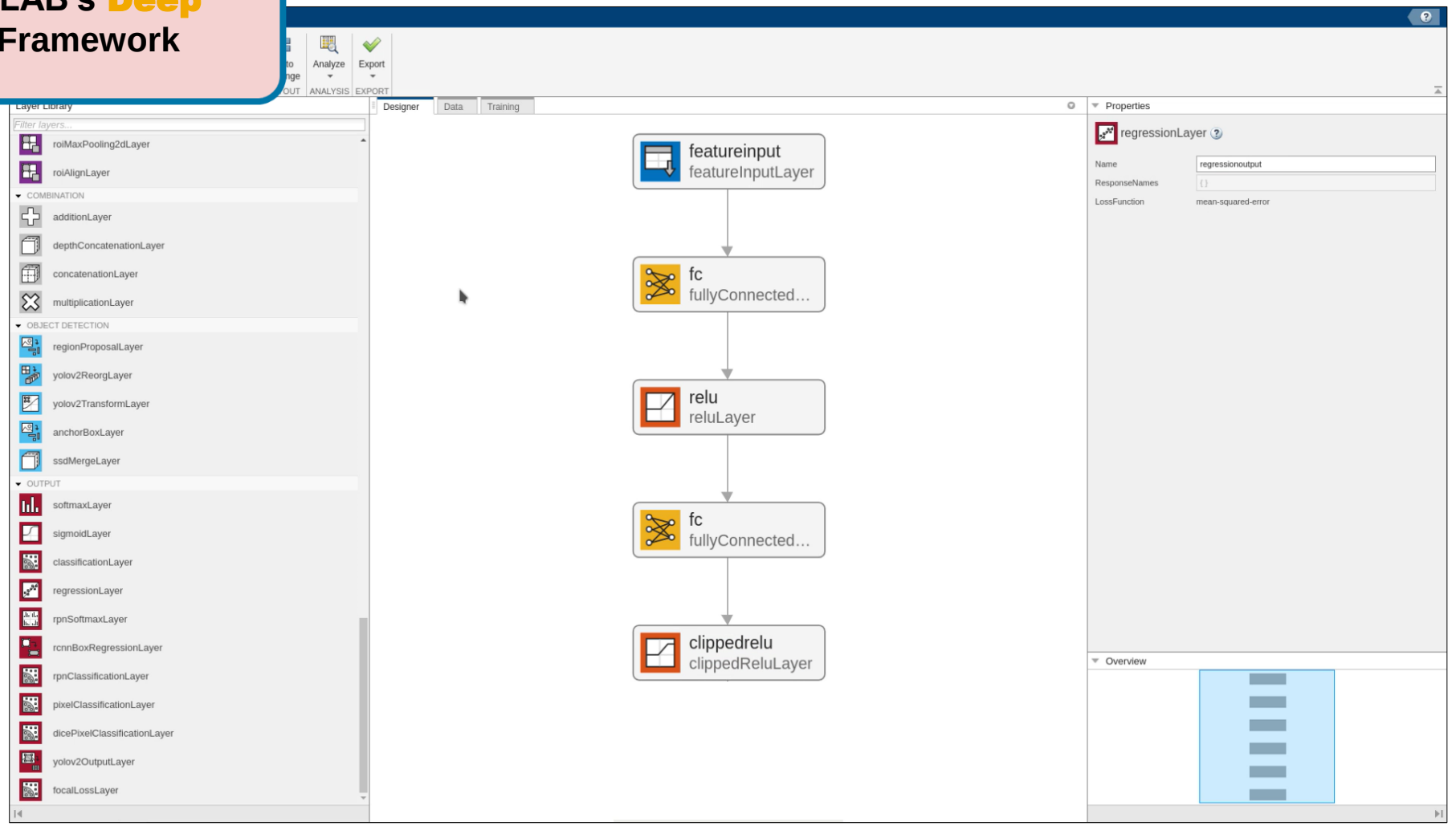
3. Import one developed using another framework



Data Preparation | **AI Modeling** | Simulation & Test | Deployment

AI Modeling

Train in MATLAB's Deep Learning Framework



Data Preparation | **AI Modeling** | Simulation & Test | Deployment

MATLAB Demo

Ch1_ECG_Classification_CNN.mlx

MATLAB Demo

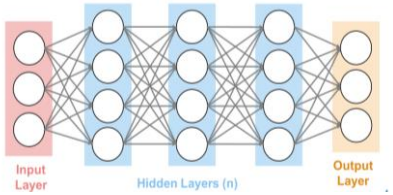
Ch2_ECG_Classification_LSTM_Wavelets.mlx

Neural Network Architectures

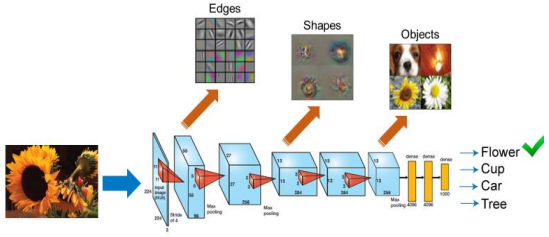
[MATLAB Deep Learning Models on Github]

Standard Framework (App workflow)

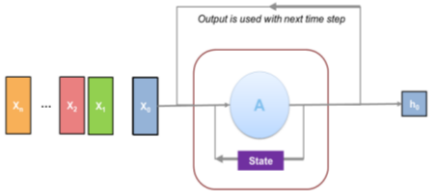
MLP
(Multi-layer Perceptrons)



CNN
(Convolutional NN)

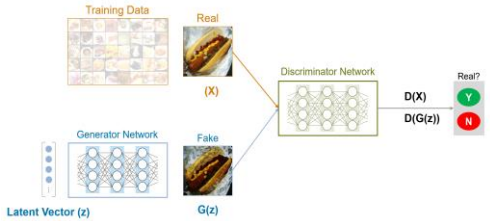


LSTM
(Recurrent NN)

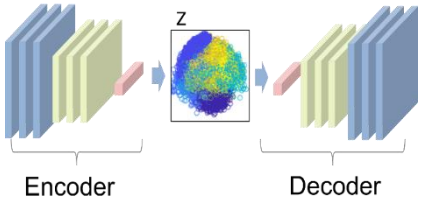


Advanced Framework (programmatic workflow)

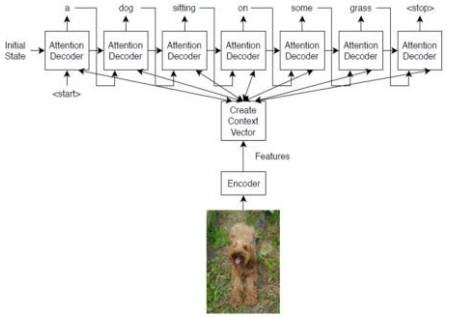
GAN
(Generative Adversarial Networks)



VAE
(Variational Autoencoder)



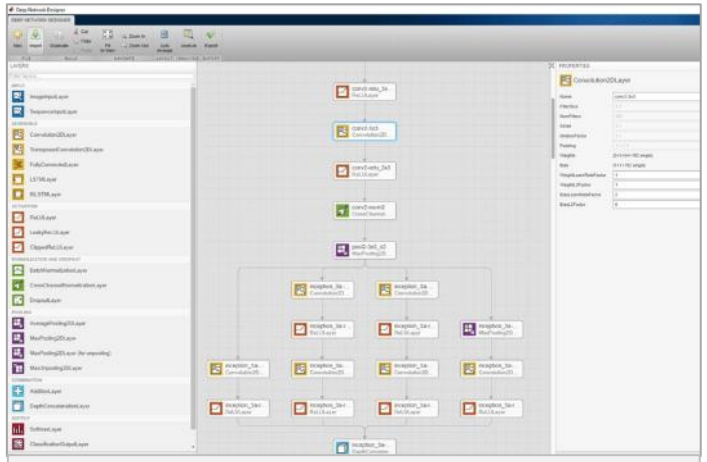
Attention Mechanisms



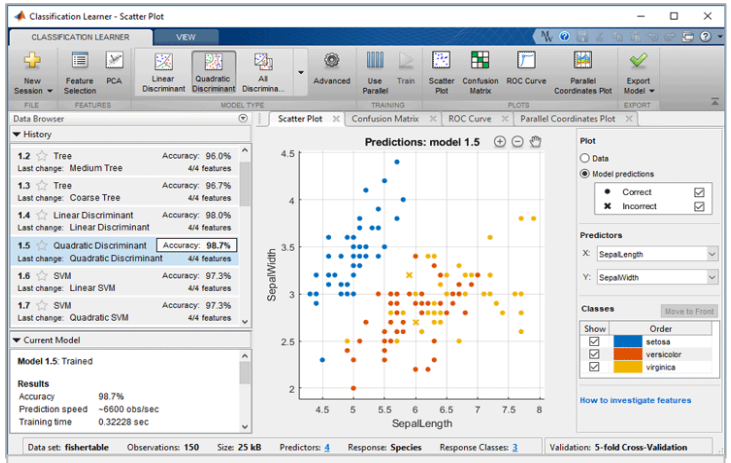
Data Preparation | **AI Modeling** | Simulation & Test | Deployment

Increase productivity using apps for design and analysis

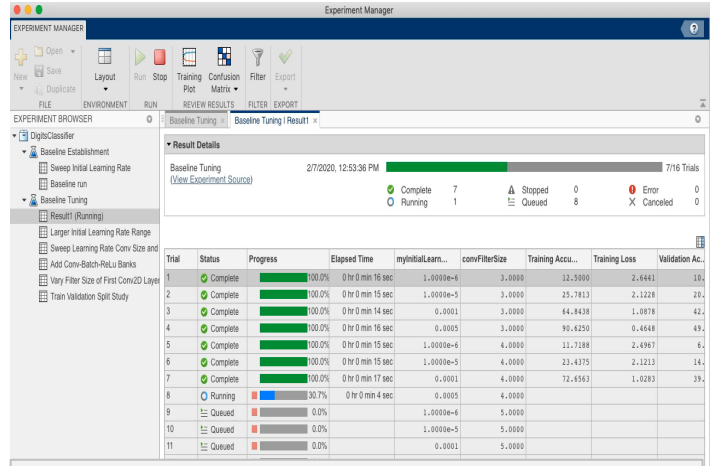
Use MATLAB Apps to design deep learning networks, explore a wide range of classifiers, train regression models, train an optical character recognition model, and more.



Deep Network Designer app to build, visualize, and edit deep learning networks.



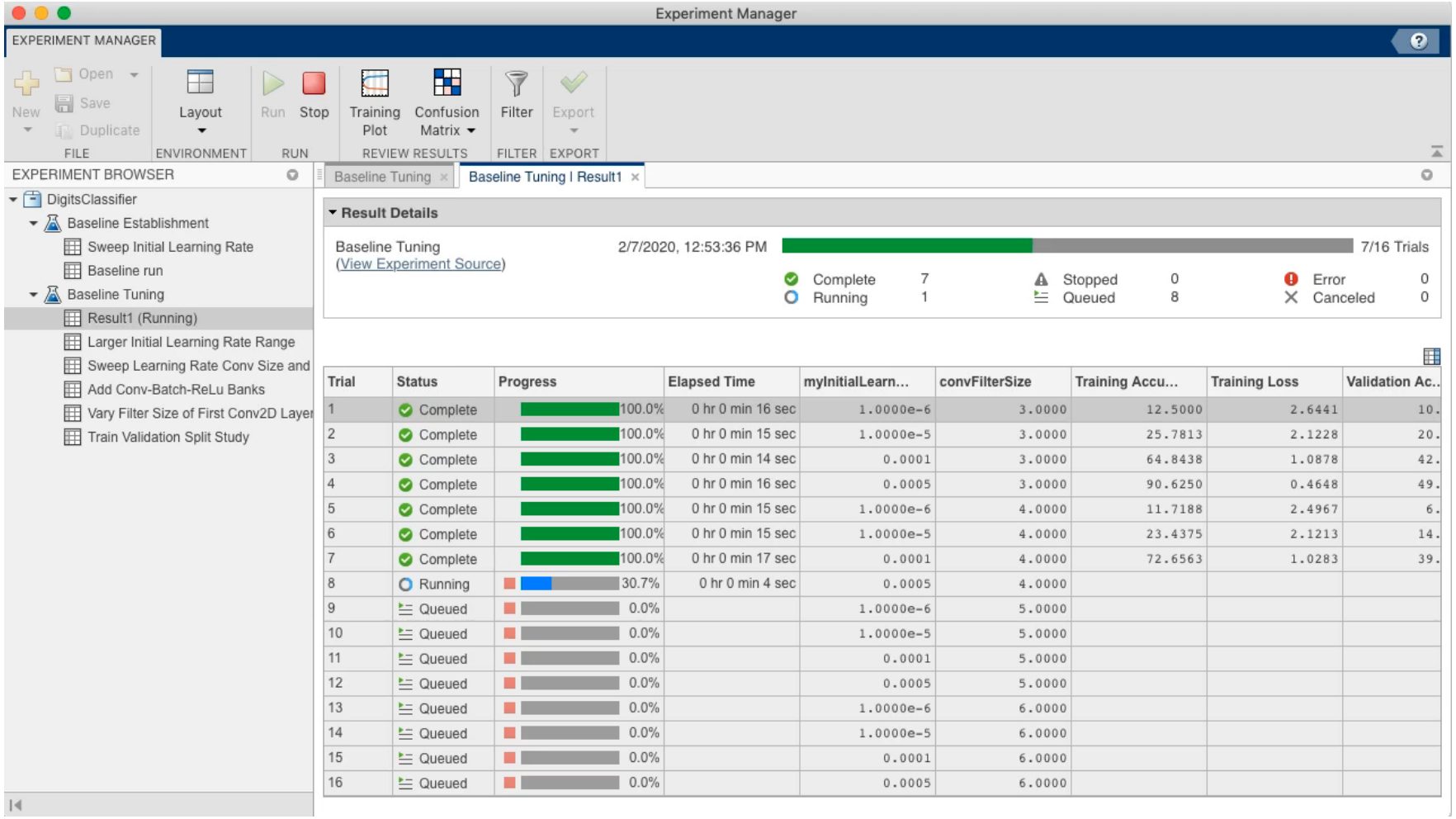
Learner apps to try different algorithms and find the best fit for data sets.



Experiment Manager app to run AI experiments to train models and compare results.

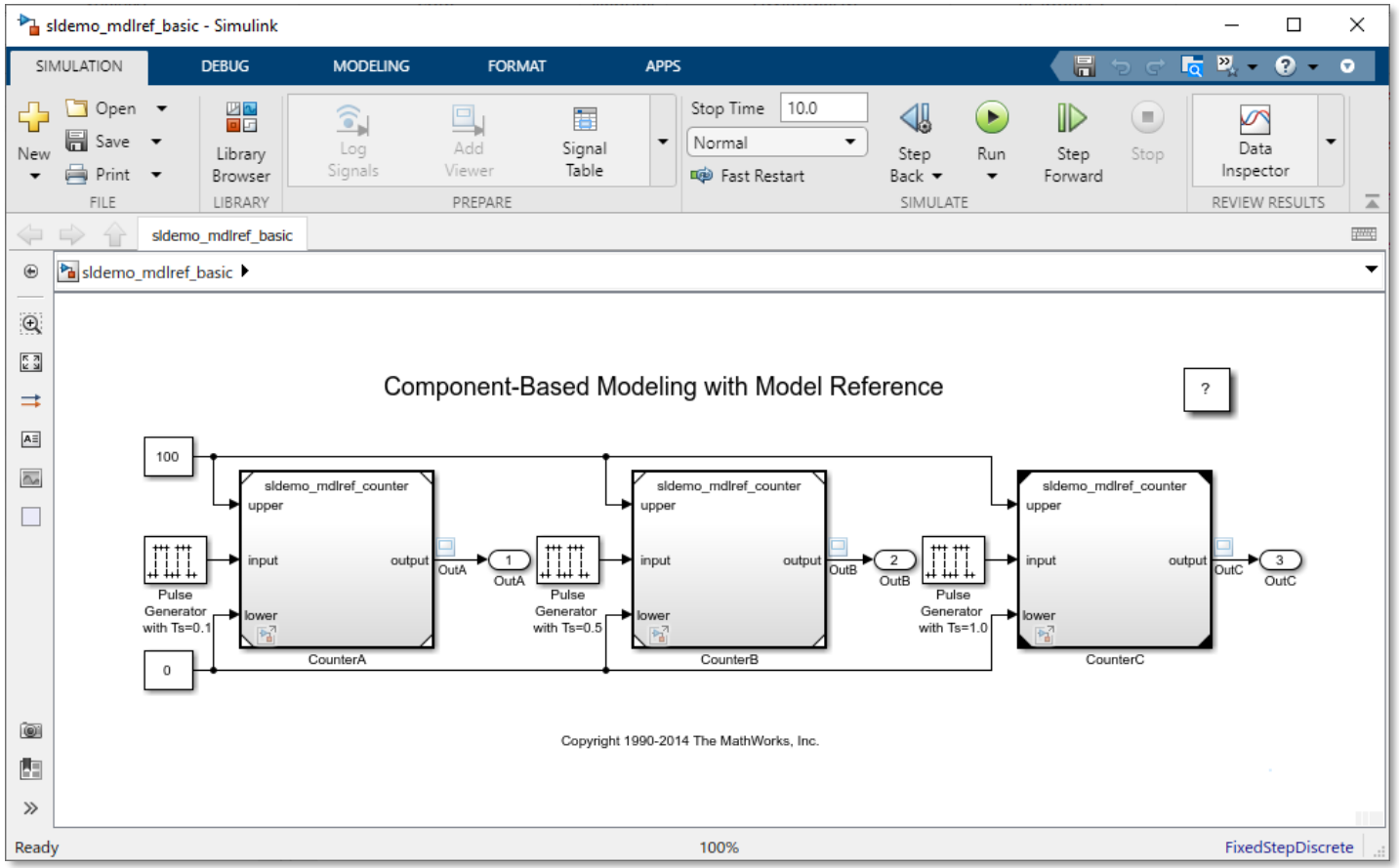
Find the Optimal Model Using Experiments

- Sweep through a range of hyperparameter values
- Test different deep network architectures and machine learning models



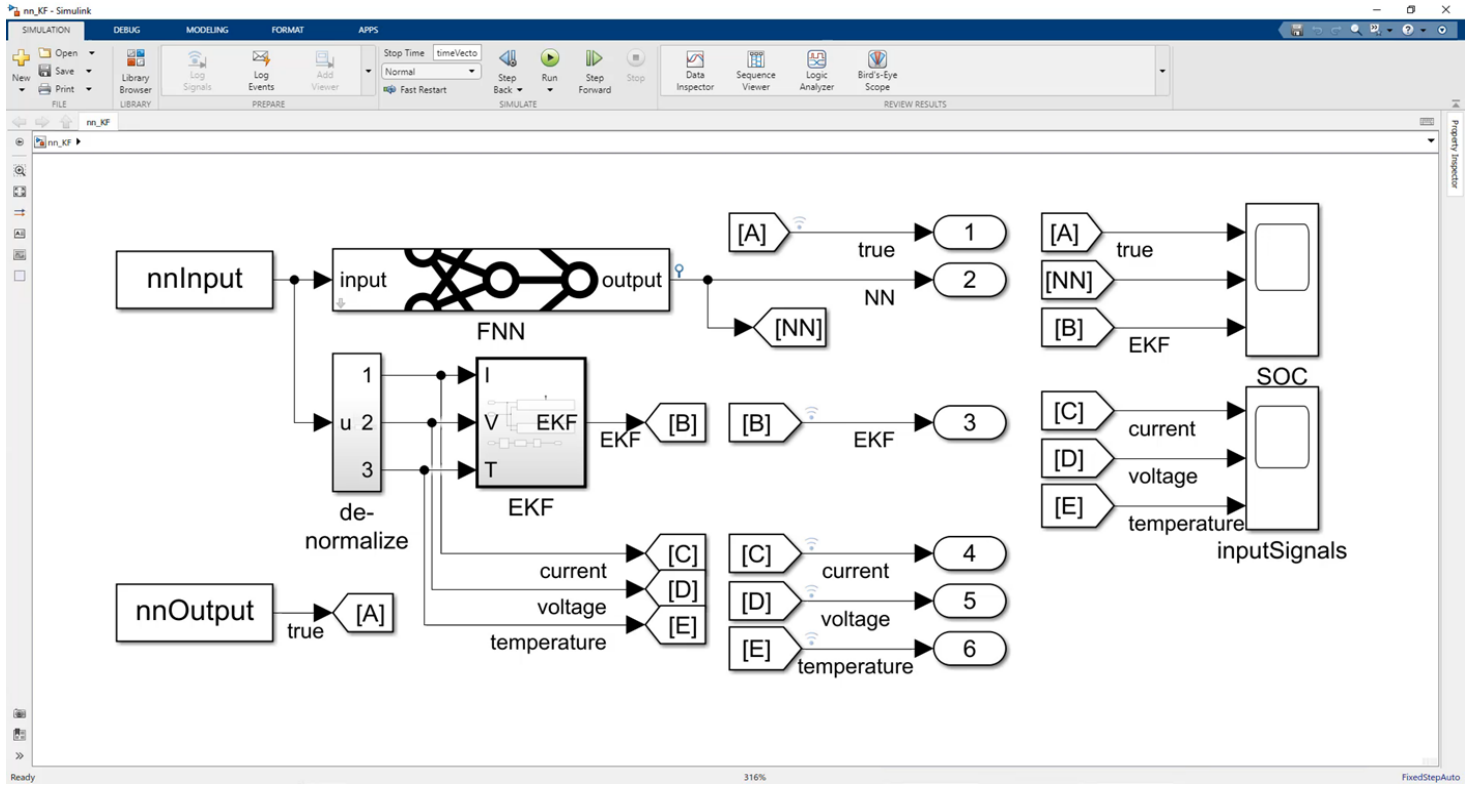
Intro to Simulink

- Design models using drag-and-drop blocks.
- Run simulations to test and validate models.
- Access to a vast library of blocks for various domains
- Generate C/C++ code for embedded systems



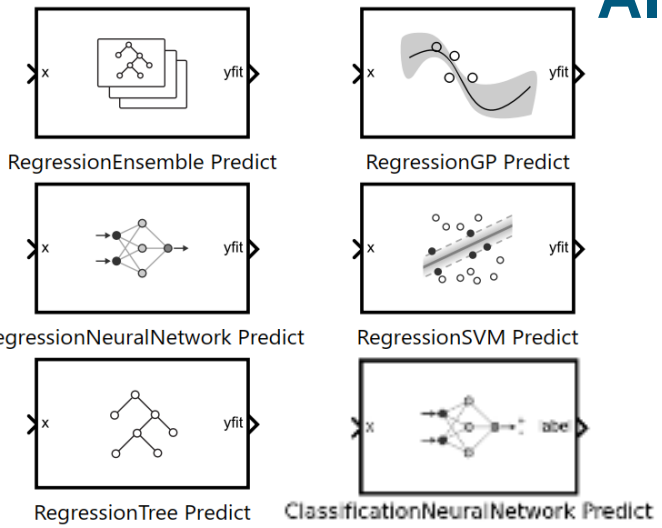
Integrate your AI model for system-level simulation and test

Integration of trained AI model into Simulink

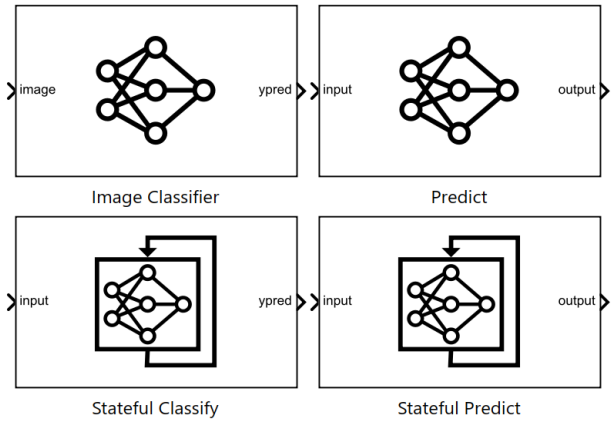


Simulink include more AI blocks for more applications

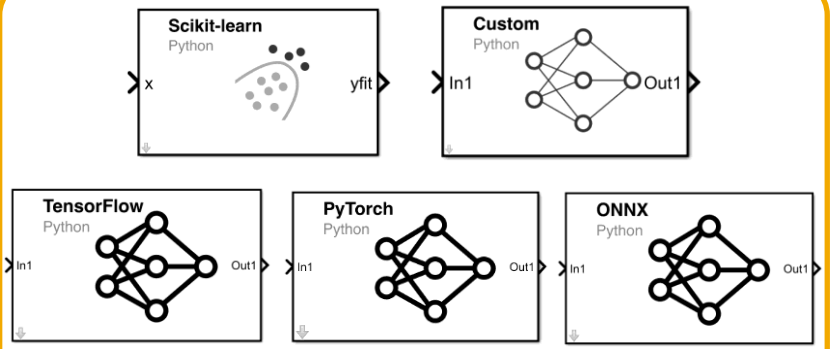
AI core



Regression and classification blocks
(Statistics and Machine Learning Toolbox)

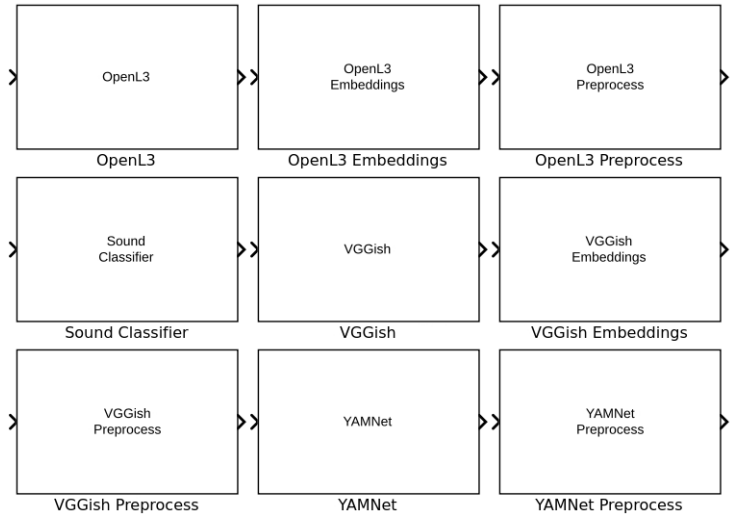


Neural network classify and predict blocks
And the list of layer blocks
(Deep Learning Toolbox)

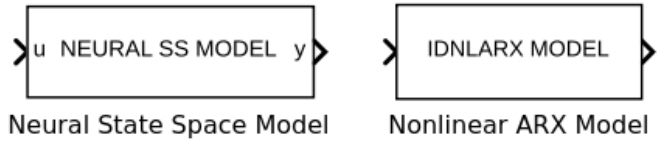


Co-execution blocks

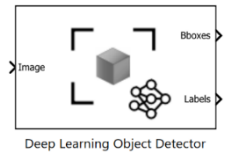
Specialized



Audio Toolbox



System Identification Toolbox



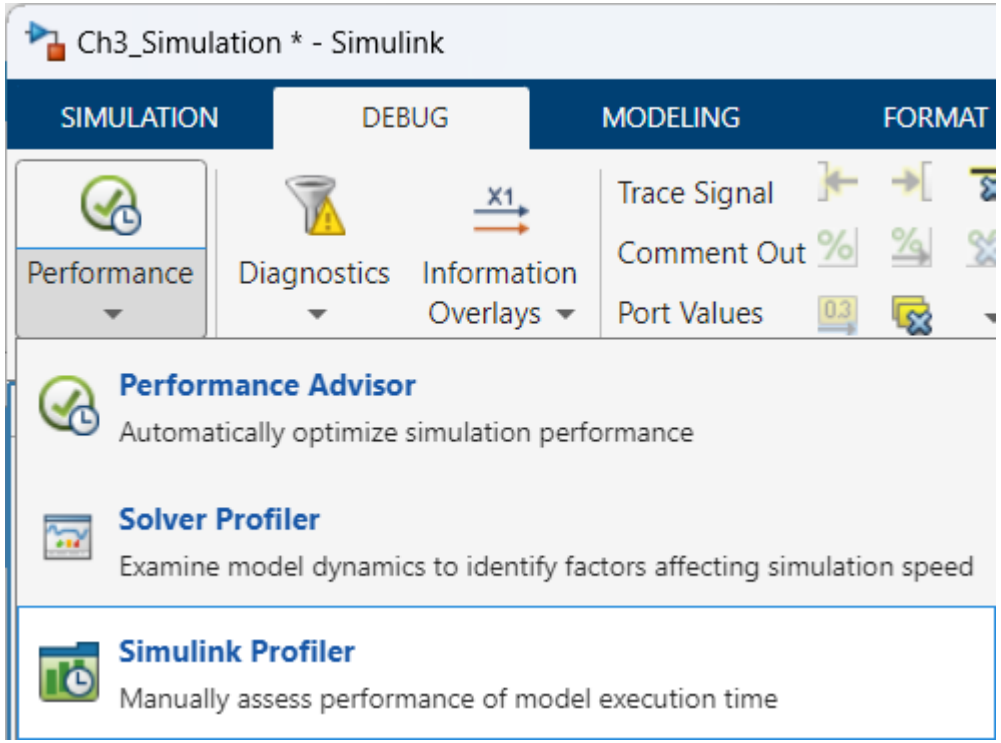
Computer Vision Toolbox

MATLAB Demo

Ch3_FunctionalVerification_Simulation.mlx

Incorporate and simulate trained models in Simulink

Profiling in Simulink



MATLAB Exercise

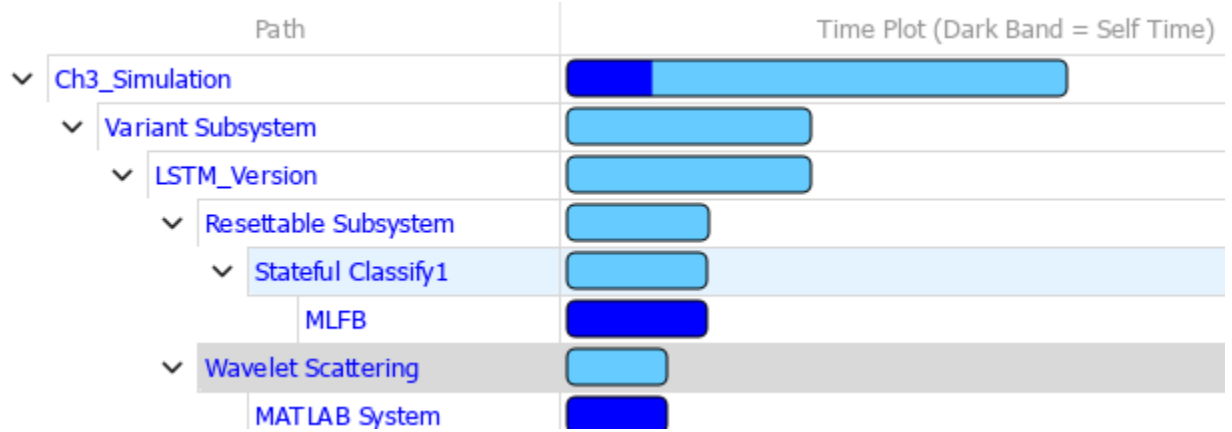
Ch3_Simulation.slx

Profile performance and compare models

MATLAB Exercise

Ch3_Simulation.slx

Profile performance and compare models



Manage AI tradeoffs for your system

	SqueezeNet	googleNet	LSTM
Preprocessing effort	🟡	🟡	🟡
Training Speed	🟡	🔴	🔵
Interpretability	🟡	🟡	🔴
Inference Speed	🟡	🔴	🔵
Model Size	🟡	🔴	🔵
Accuracy (RSME)	🟡	🔵🔵	🟡

Results are specific to this ECG classification example

Much Better 🔵🔵 Better 🔵 Okay 🟡 Worse 🔴

50

Model compression can bridge the gap between AI modelling and embedded deployment

Data Preparation

- Data cleansing and preparation
- Human insight
- Simulation-generated data

AI Modeling

- Model design and tuning
- Hardware accelerated training
- Interoperability

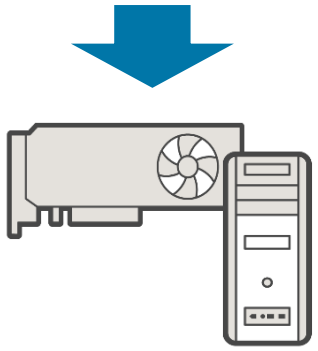
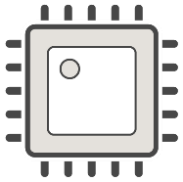
Simulation & Test

- Integration with complex systems
- System simulation
- System verification and validation

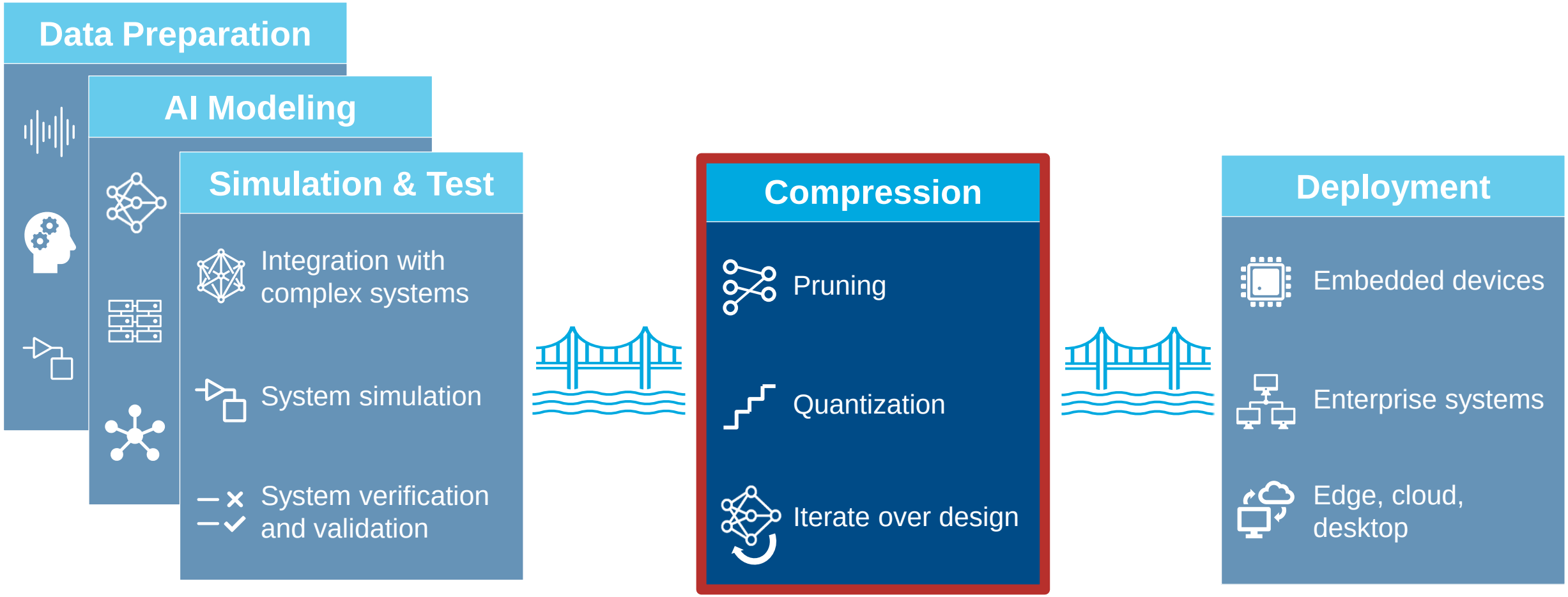
Deployment

- Embedded devices
- Enterprise systems
- Edge, cloud, desktop

Embedded Devices Have Limited Resources

	Memory	Performance	Power
	32GB	TFLOPs	
	8GB	GFLOPs	
	1kB~1MB	MFLOPs	

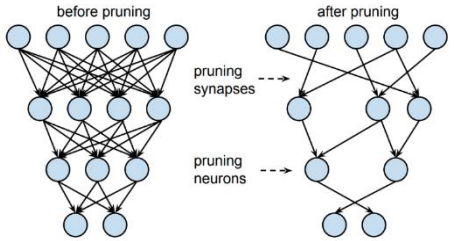
Model compression can bridge the gap between AI modelling and embedded deployment



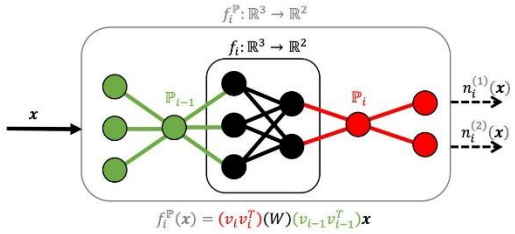
53

Model Compression

Structural Compression

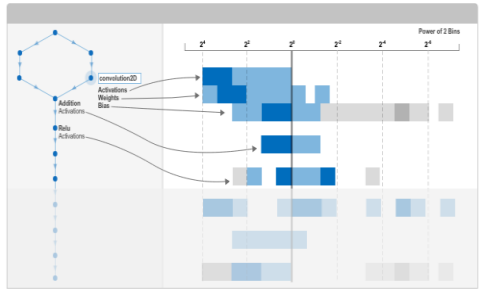


Pruning convolutional neural networks



Projection of deep neural networks

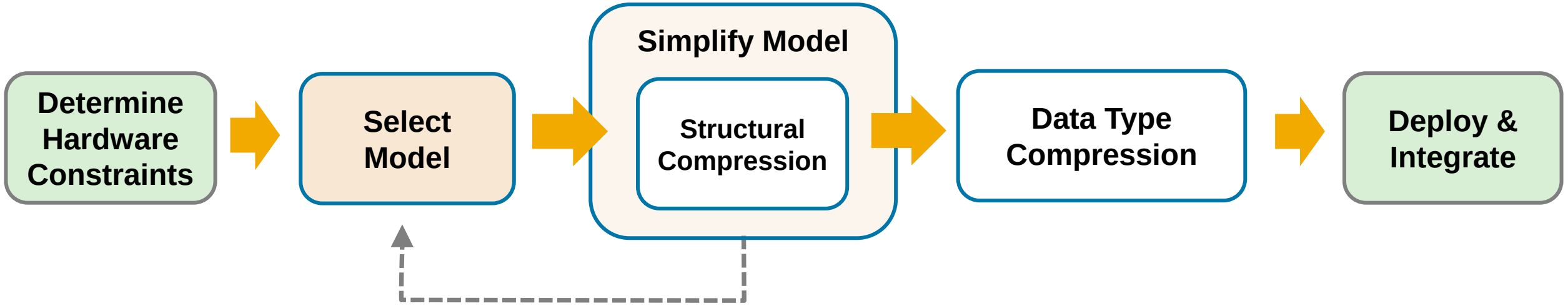
Datatype Compression



Quantization of network weights to lower precision datatypes (bfloat16, int8)

[Quantization, Projection, and Pruning]

Iterative Workflow for Model Compression



MATLAB Exercise

Ch3_Compression_ LSTM.mlx

Results: Model Compression with Projection for the CNN

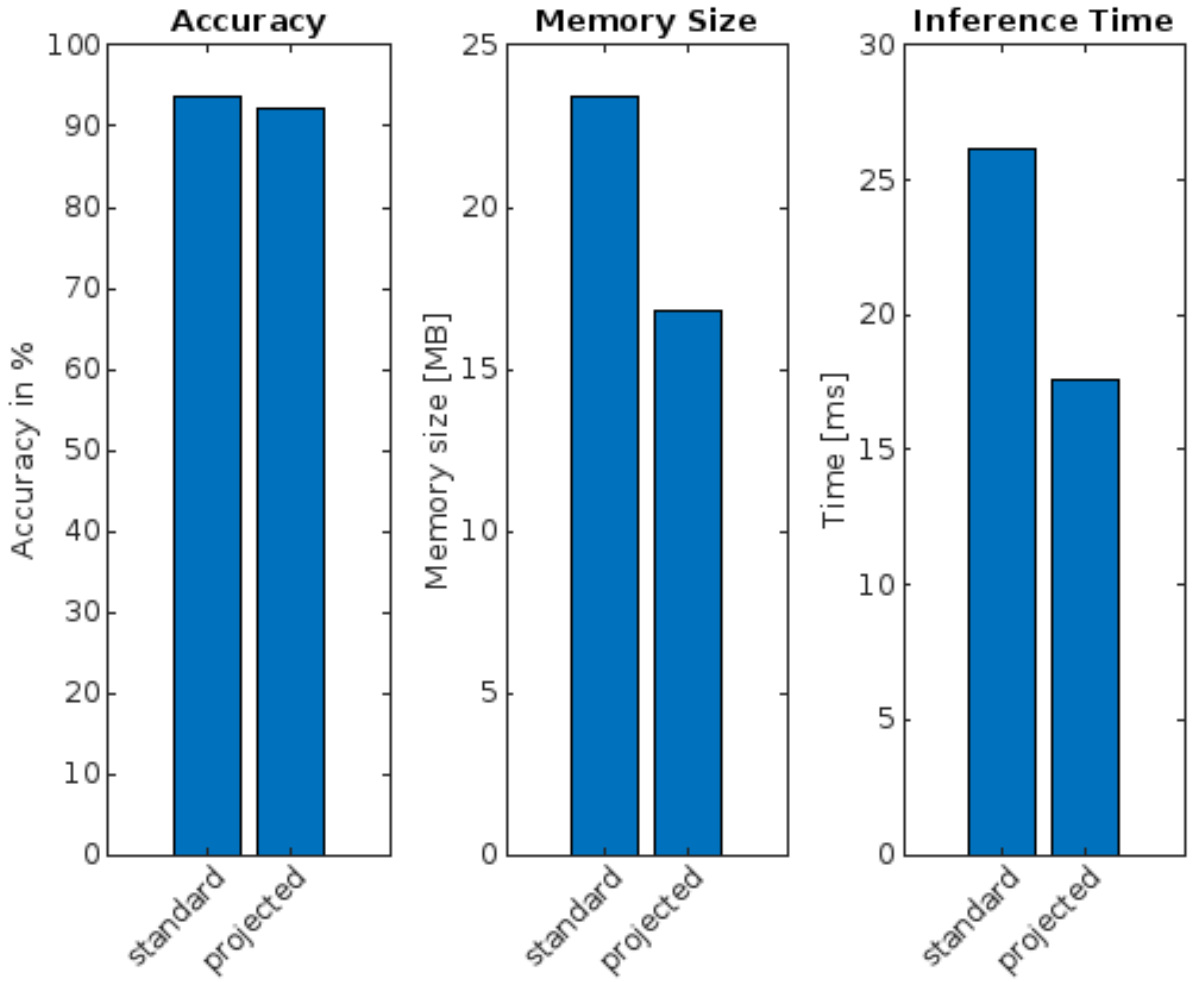
5.9M	143	0	0
total learnables	layers	warnings	errors

1.1M	143	0	0
total learnables	layers	warnings	errors

Number of total learnables decreased

Test setup

- Run on laptop CPU
- Averaged over 1000 predictions



Results: Model Compression with Projection for the LSTM

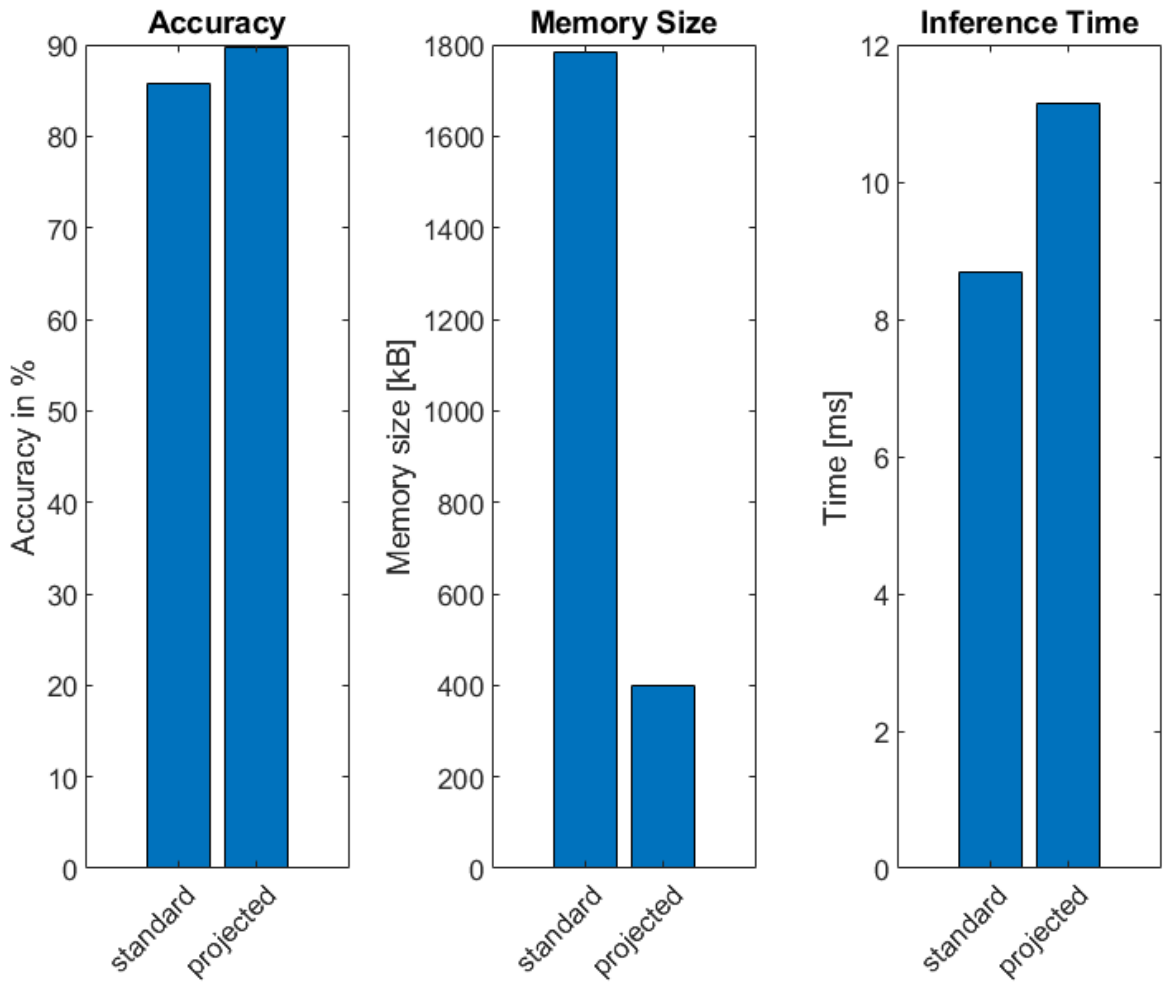
450.1k	4	0	0
total learnables	layers	warnings	errors

25.8k	4	0	0
total learnables	layers	warnings	errors

Number of total learnables decreased

Test setup

- 1000 predictions
- Averaged over 100 times



Data Preparation | AI Modeling | Simulatio | Compression | yment

Manage AI tradeoffs for your system

	googleNet	compressed	LSTM	compressed
Preprocessing effort	●	●	●	●
Training Speed	●	●	●	●
Interpretability	●	●	●	●
Inference Speed	●	●	●	●
Model Size	●	●	●	●●
Accuracy (RSME)	●	●	●	●

Results are specific to this ECG classification example

Much Better ●● Better ● Okay ● Worse ●

62

MATLAB & Simulink support hardware workflows

Data Preparation

Data cleansing and preparation

Human insight

Simulation-generated data

AI Modeling

Model design and tuning

Hardware accelerated training

Interoperability

Simulation & Test

Integration with complex systems

System simulation

System verification and validation

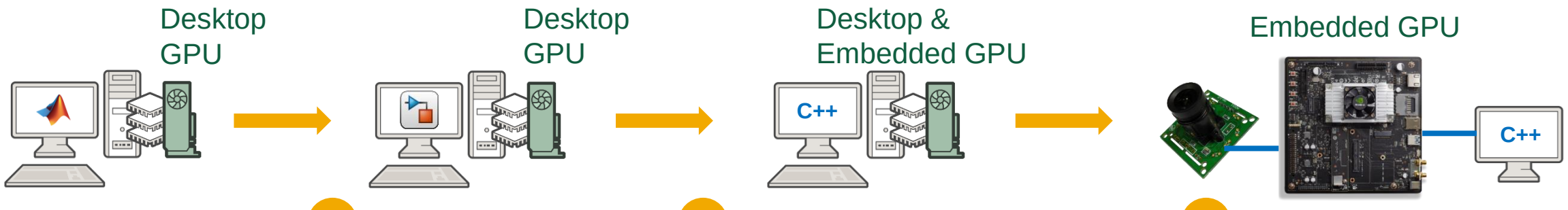
Deployment

Embedded devices

Enterprise systems

Edge, cloud, desktop

Algorithm Design to Embedded Deployment Workflow



1 Model Development

2 Simulation & Verification

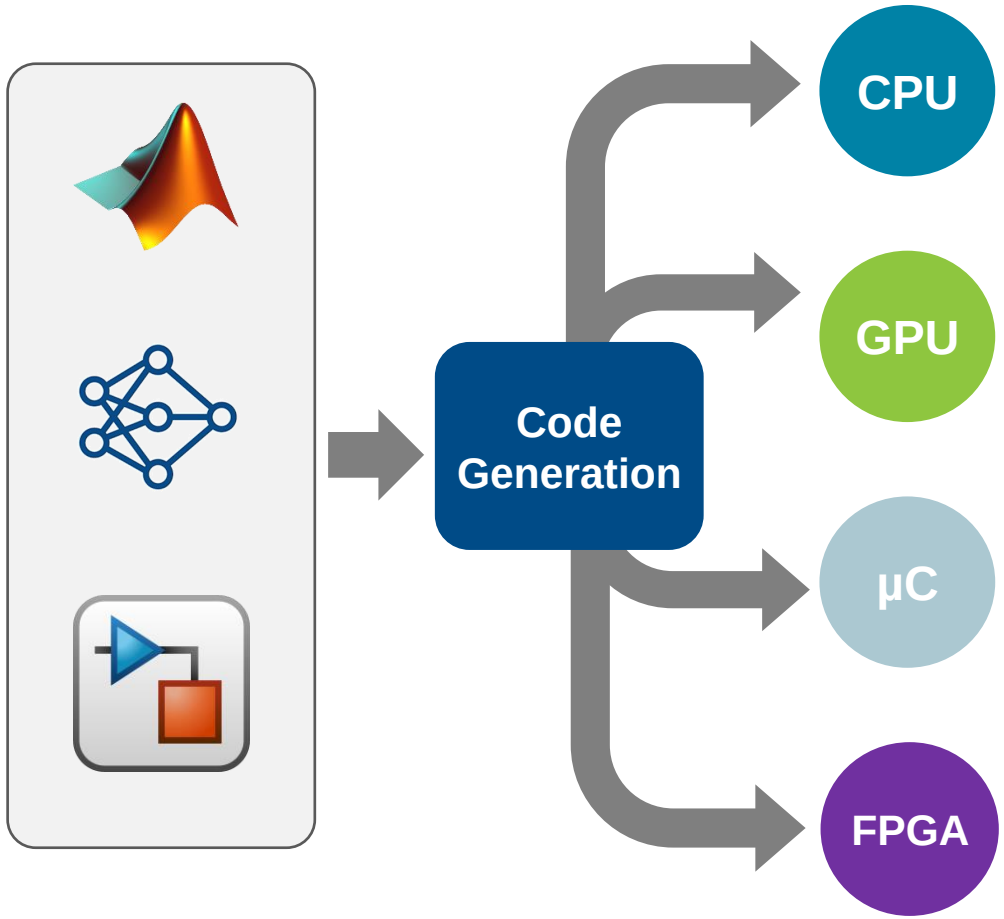
3 Deployment integration-test

4 Real-time test

High-level language
Deep learning framework
Large, complex software stack

C/C++
Target-optimized libraries
Optimize for memory & speed

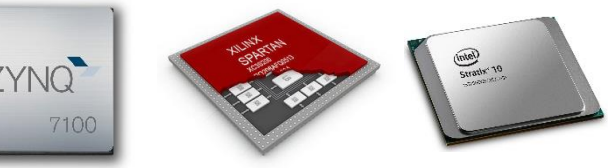
Deploy to target with zero coding errors



Any CPU
No Library needed

oneDNN
Library

ARM Compute
Library



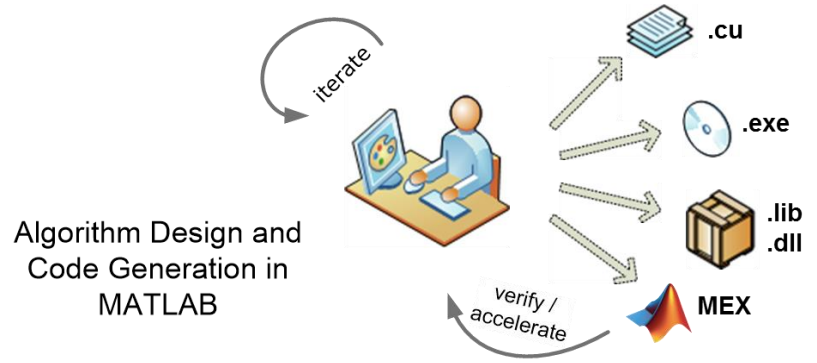
Why Use CUDA code generation ?

Pains: Hand code

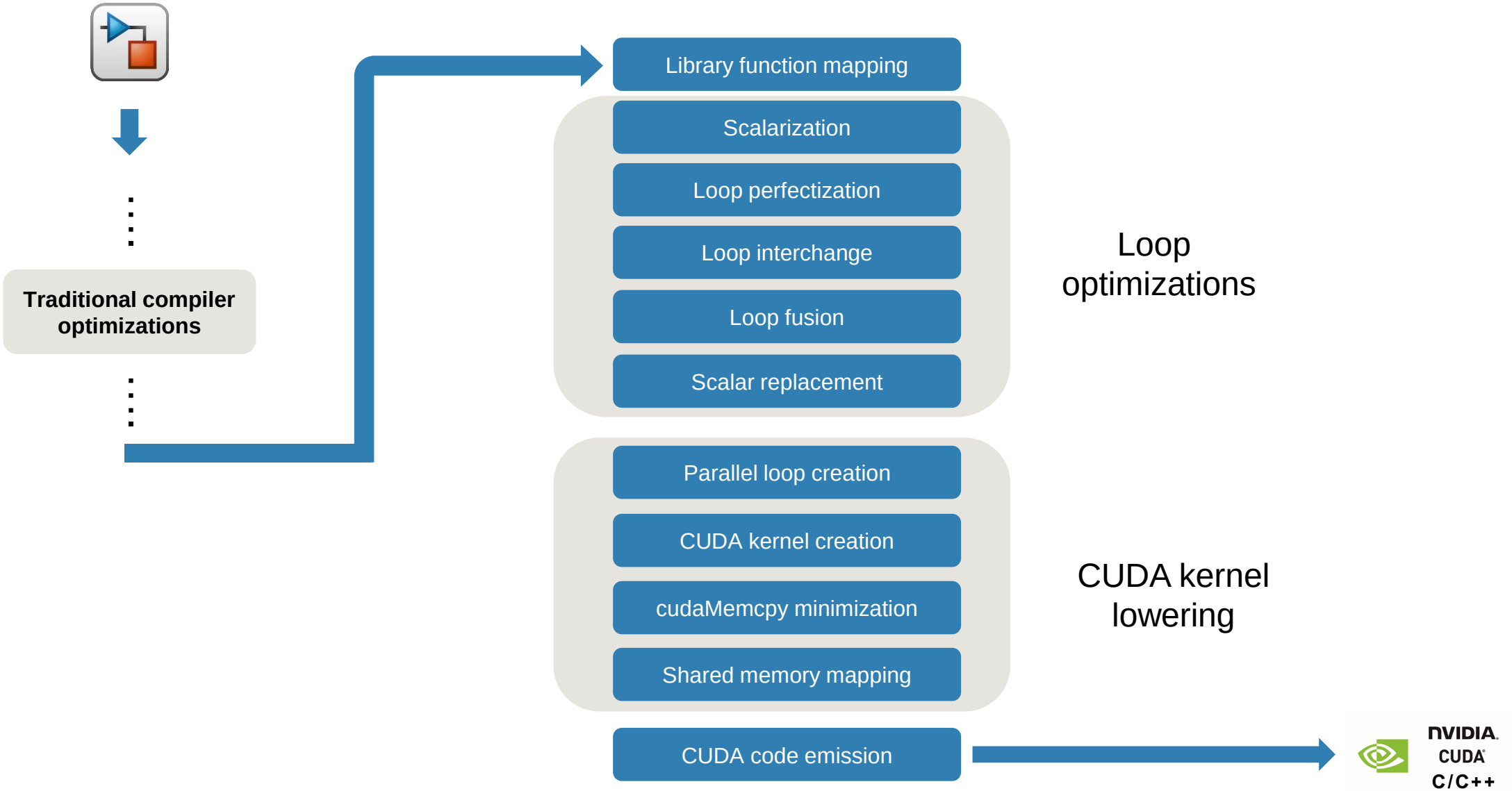
- **Cannot code in CUDA**
- Time consuming
- Manual Coding Errors
- Multiple implementations
- Expensive

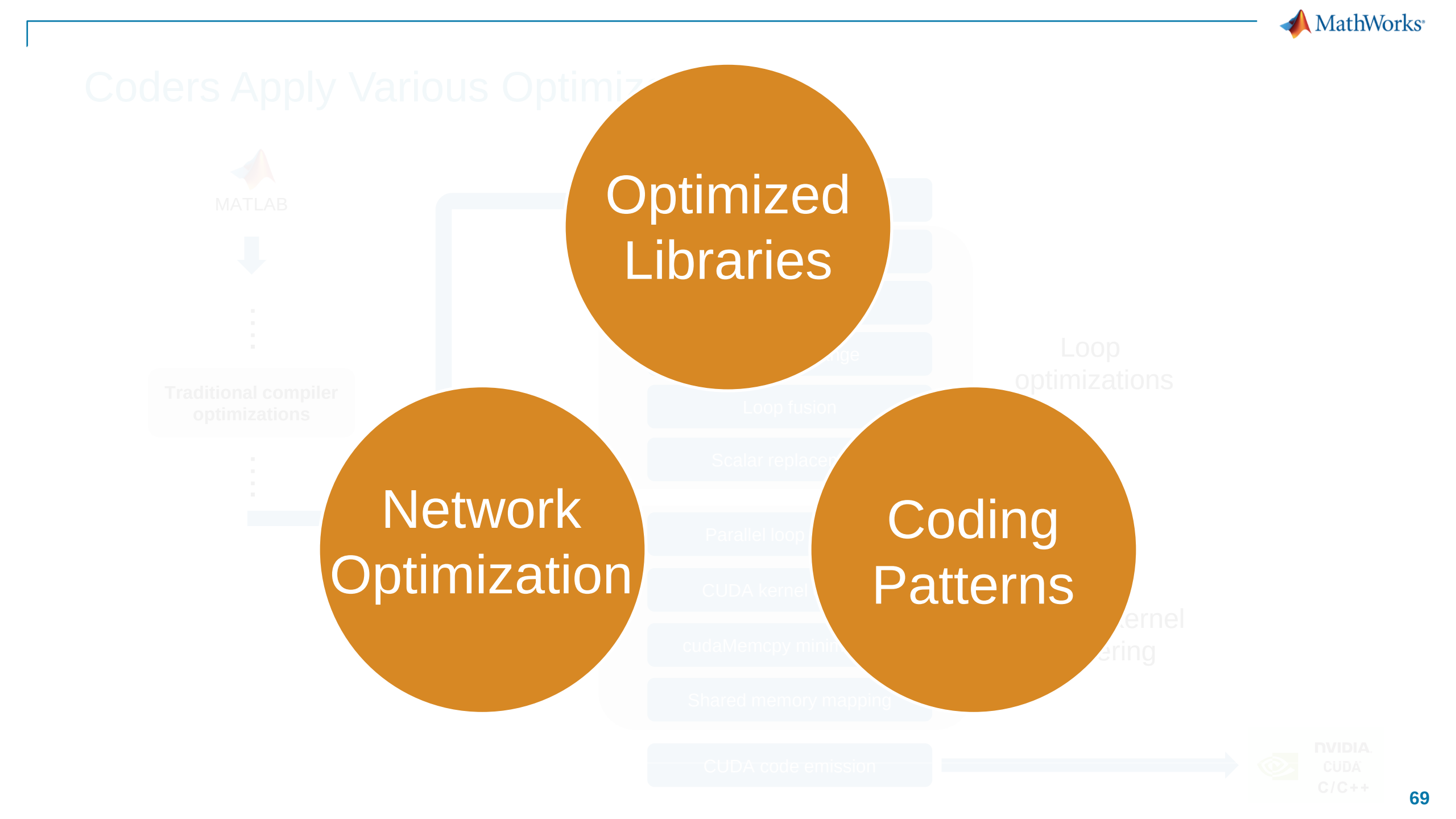
Solution: GPU Coder

- Automatically convert to CUDA
- Get to optimized CUDA faster
- Eliminate manual coding errors
- Maintain Single “Truth”
- Stay within MATLAB/Simulink at a higher level

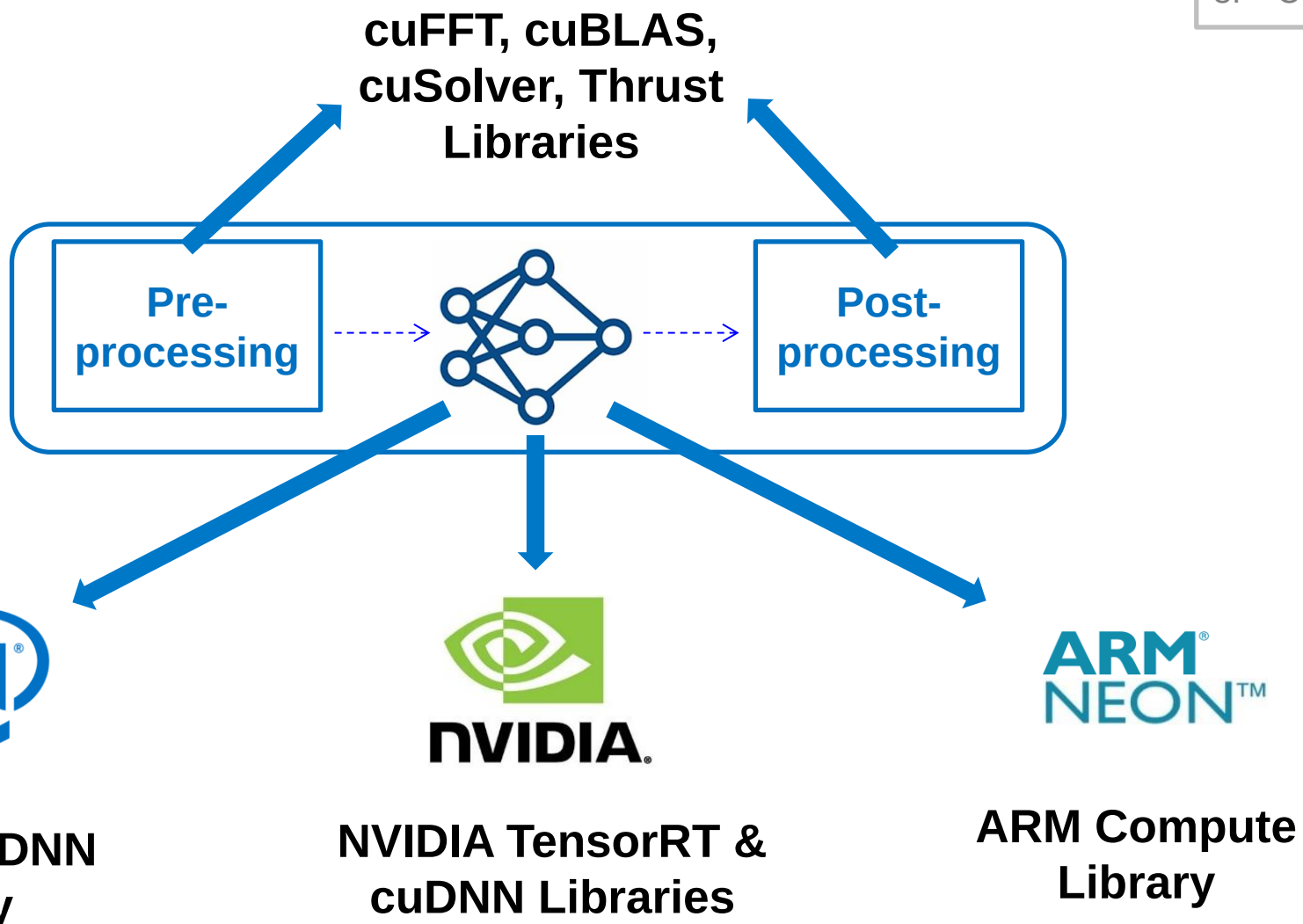


GPU Coder Apply Various Optimizations

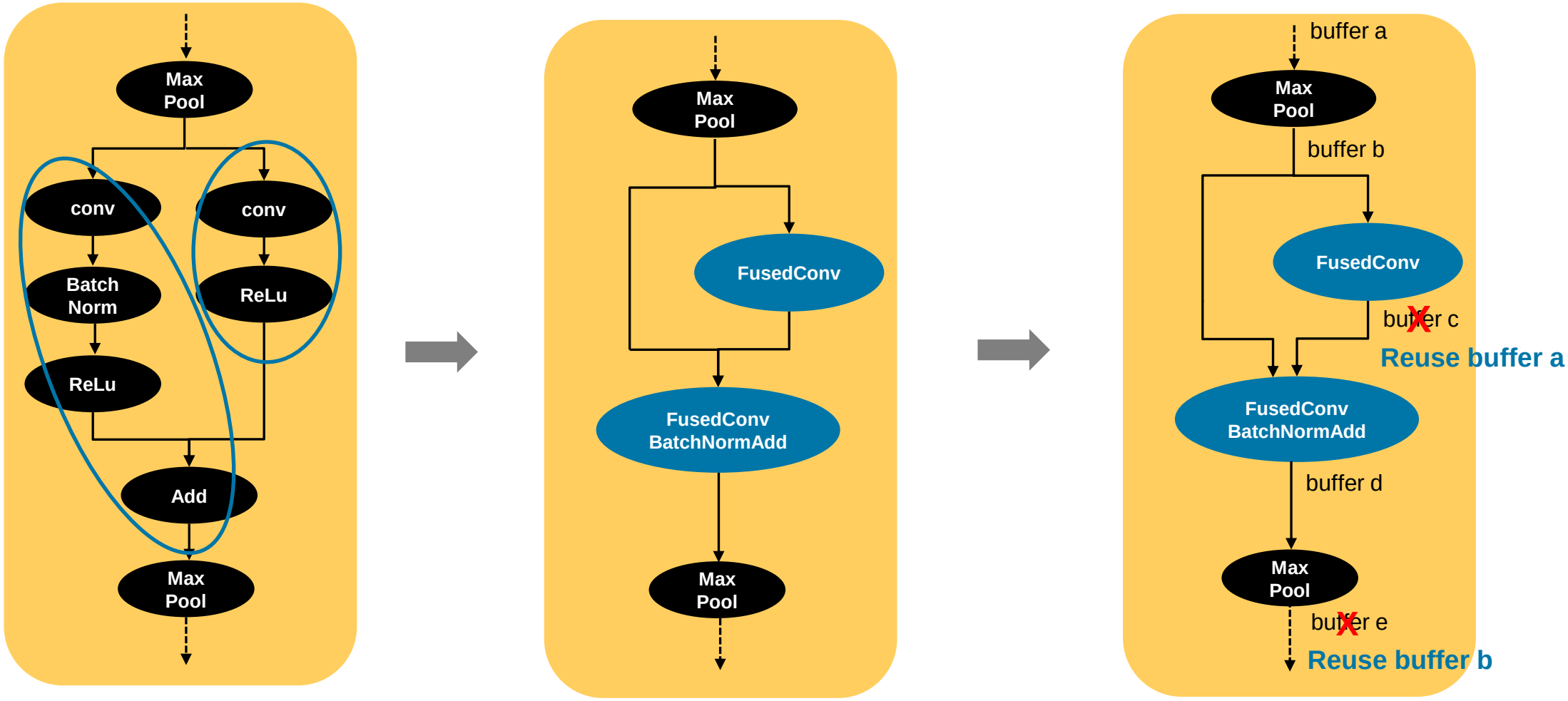



Optimized Libraries
Network Optimization
Coding Patterns

Generated Code Calls Optimized Libraries



Deep Learning Network Optimization



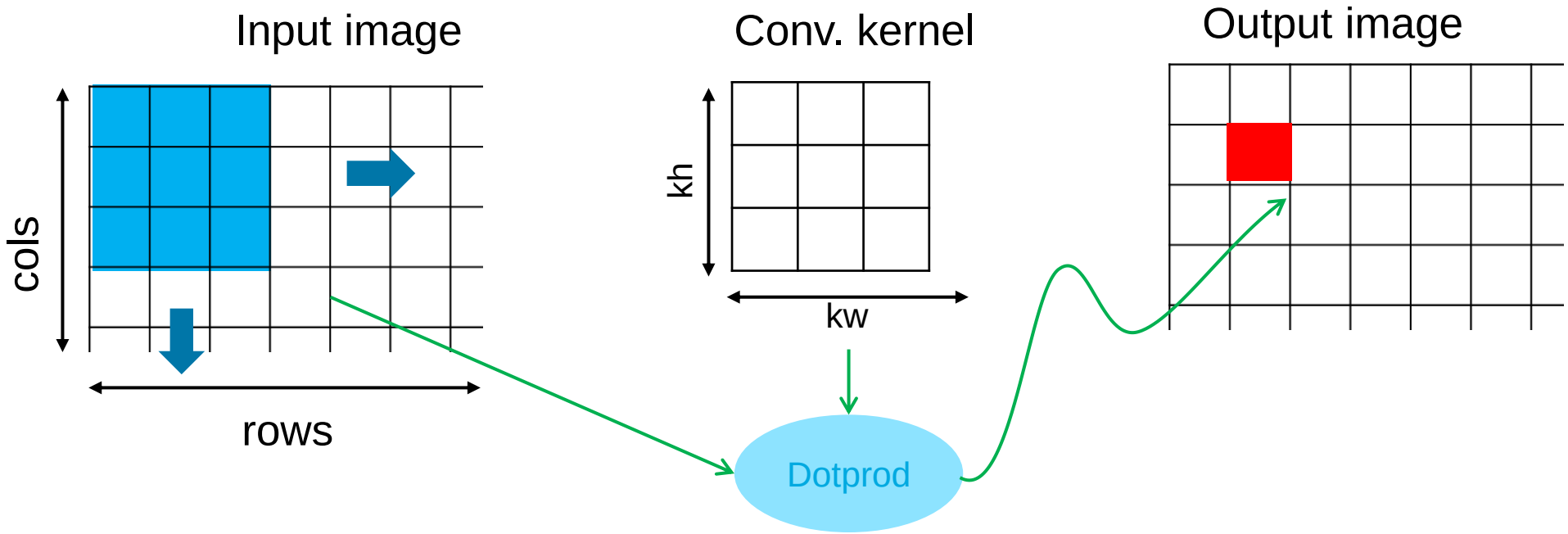
Network

Layer fusion
Optimized computation

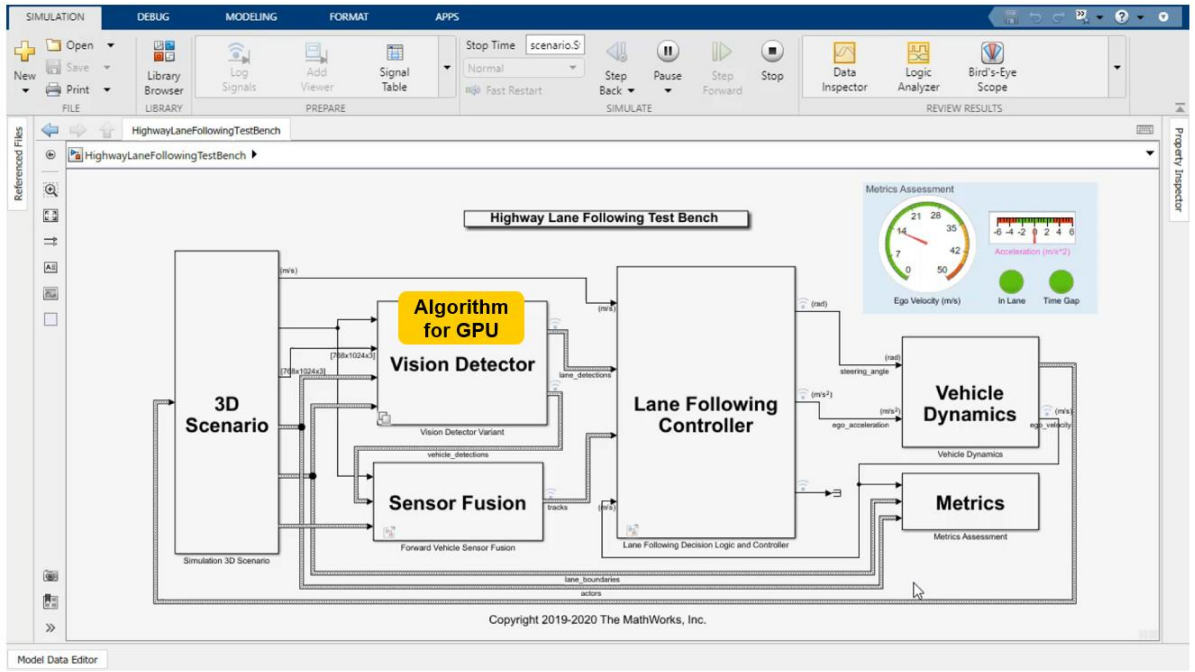
Buffer minimization
Optimized memory

Coding Patterns: Stencil Kernels

- Automatically applied for image processing functions (e.g. imfilter, imerode, imdilate, conv2, …)
- Manually apply using *gpucoder.stencilKernel()*

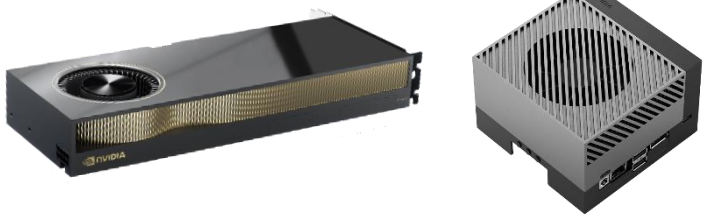


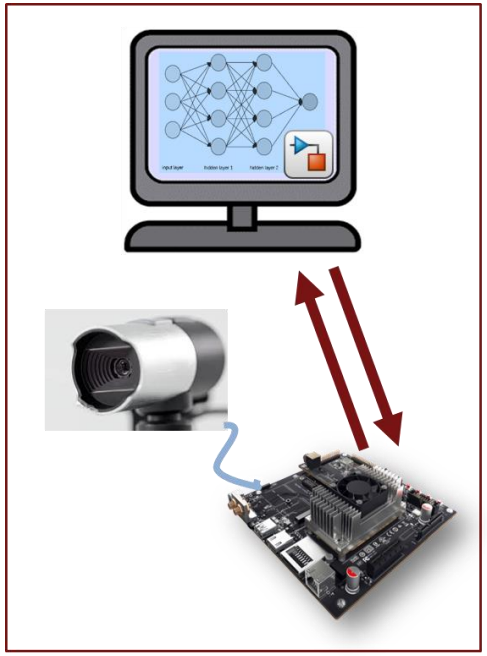
GPU Coder for Simulink Workflows



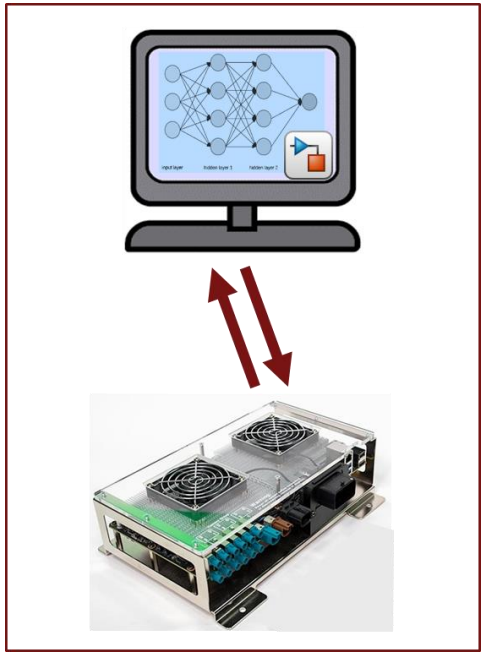
Deploy Standalone Application

GPU Coder



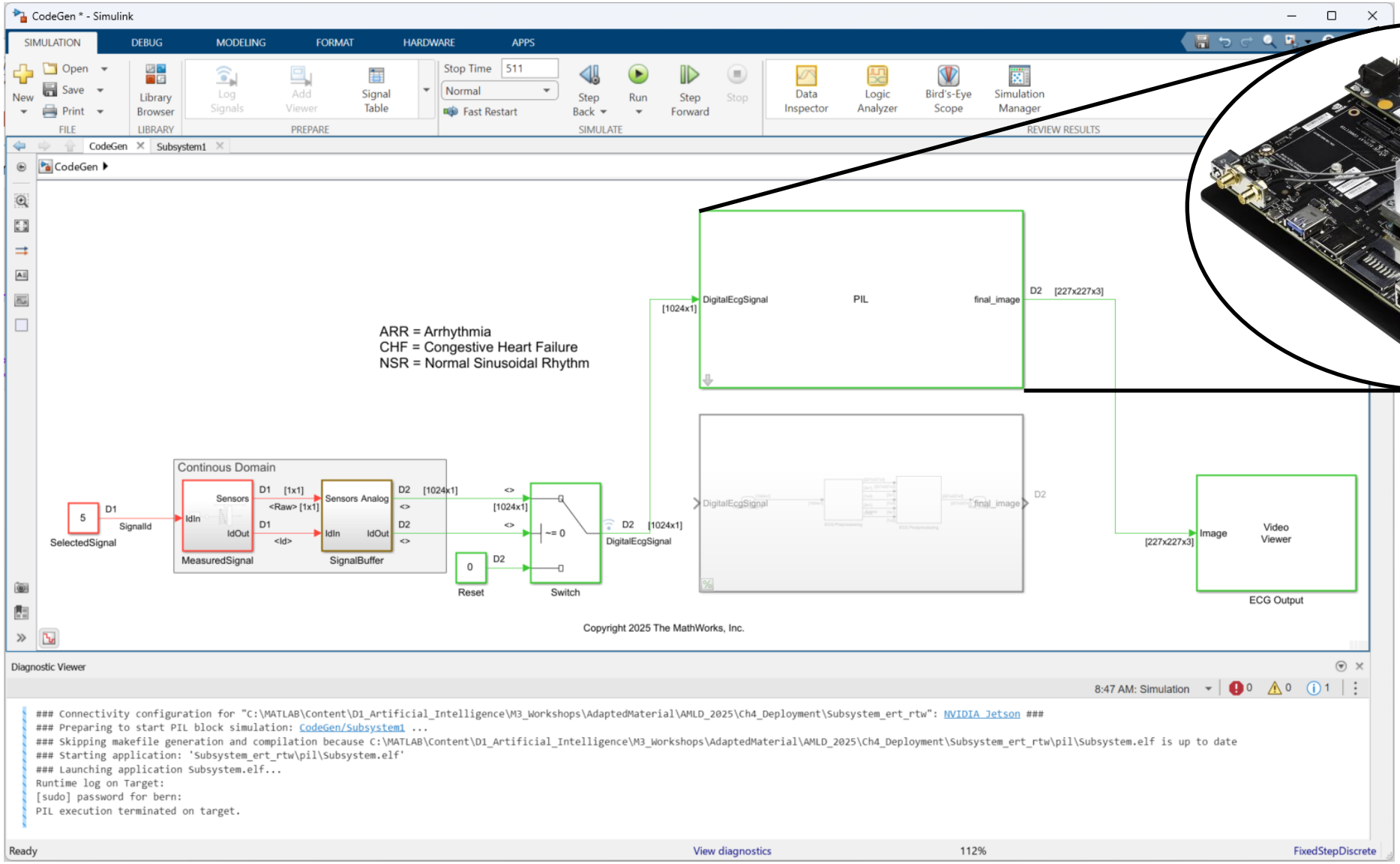


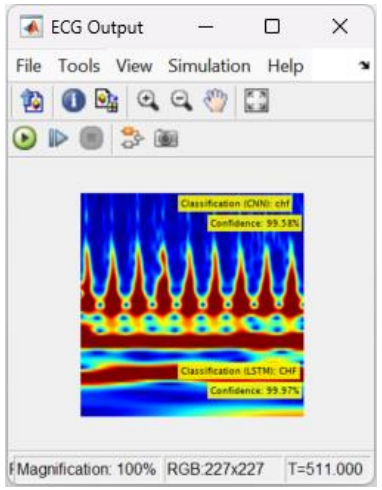
Access Peripheral from Simulink



Software-in-Loop, Processor-in-Loop, External Mode

Demo: Test algorithm on the Jetson



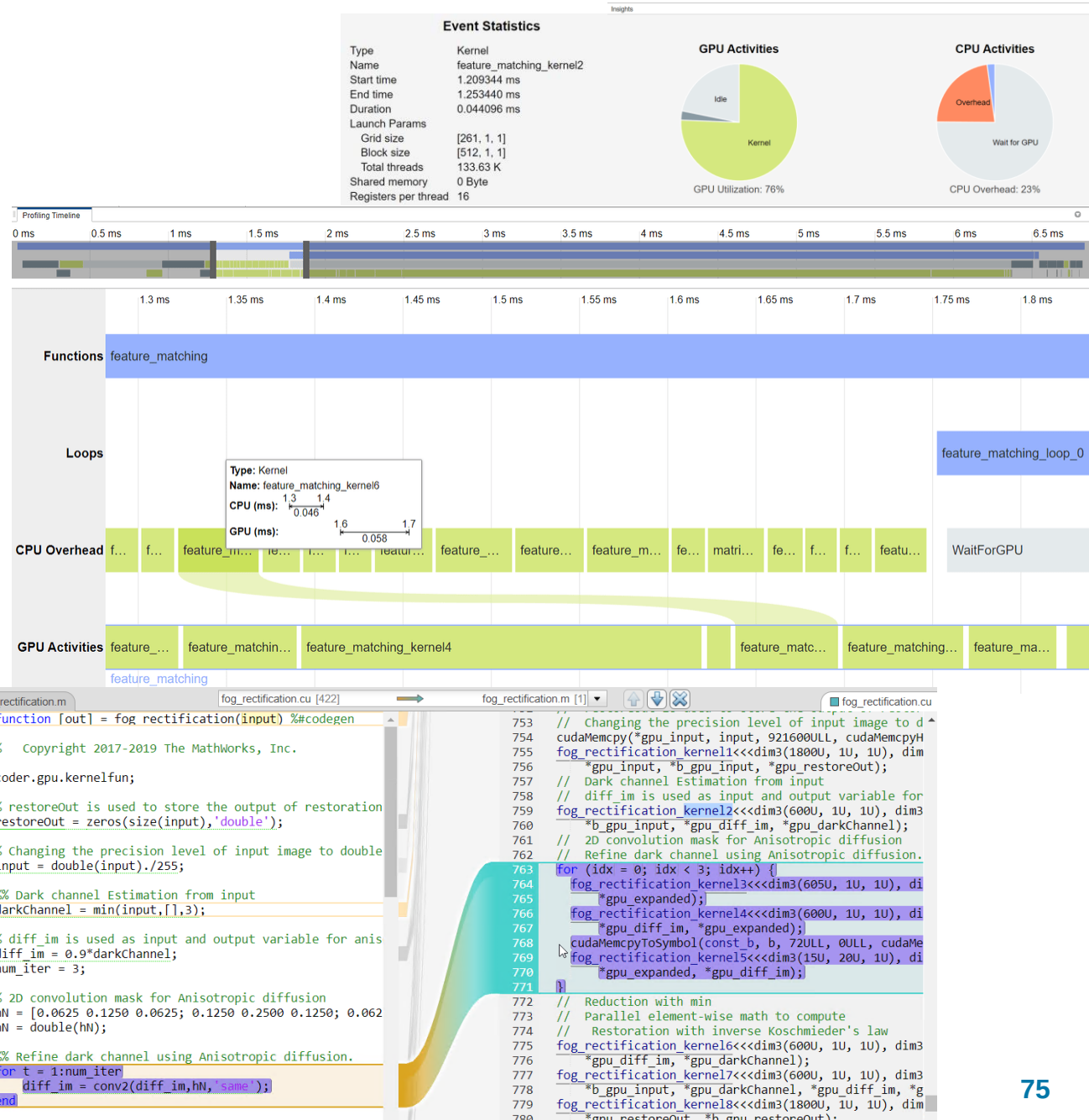


How to Optimize Further?

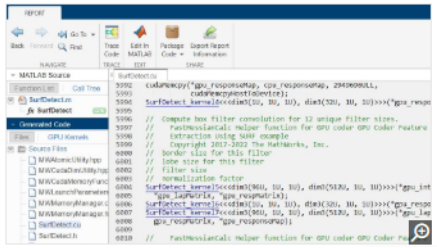
Profiling and Bidirectional Traceability **Tools**

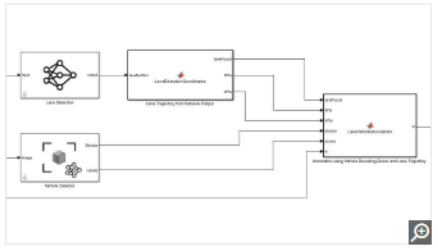
- Use the GPU Coder Performance Analyzer to profile the generated CUDA code
  - Identify bottlenecks & opportunities to optimize performance

- Use bidirectional traceability to map to/from CUDA code back to MATLAB code
  - Helps you to understand how CUDA kernels are created from your MATLAB algorithms



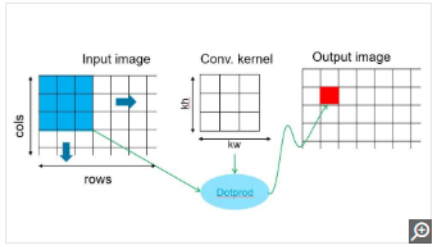
Outlook

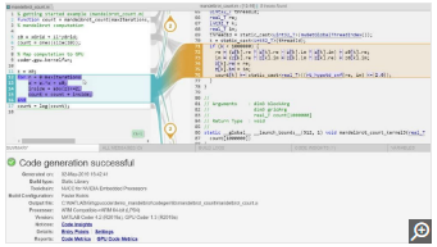


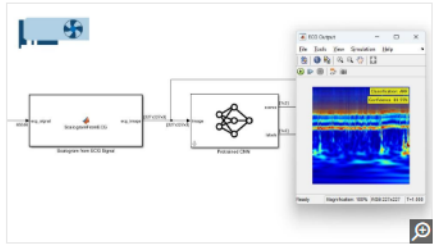


Generate CUDA Code from MA

Compile and run CUDA code generated fror
MATLAB algorithms on popular NVIDIA GP
desktop RTX cards to data centers to embe
Jetson and DRIVE platforms. Deploy the ge
code royalty-free to your customers at no c





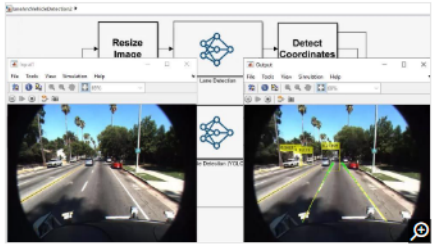


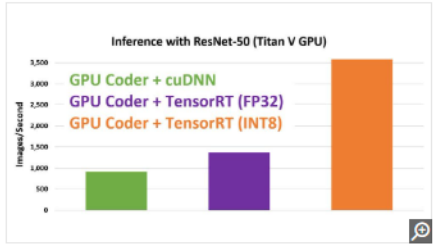
Use Design Patterns to Boost Performance

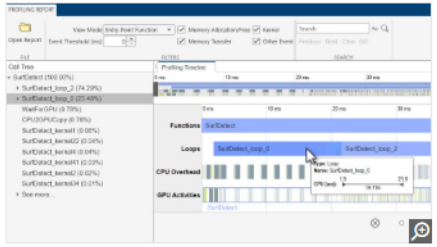
Design patterns, including stencil processin
reductions, are applied automatically when
to increase the performance of generated c
can also manually invoke them using specif
pragmas.

Log Signals, Tune Parameters, and

Accelerate MATLAB and Simulink







Generate Code for Deep Learning

Deploy a variety of predefined or customized deep learning networks to NVIDIA GPUs. Generate code for preprocessing and postprocessing along with your trained deep learning networks to deploy complete algorithms.

Optimize Generated Code

GPU Coder automatically applies optimizations including memory management, kernel fusion, and auto-tuning. Reduce memory footprint by generating INT8 or bfloat16 code. Further boost performance by integrating with TensorRT.

Profile and Analyze Generated Code

Use the GPU Coder Performance Analyzer to profile generated CUDA code and identify opportunities to further improve execution speed and memory footprint.

https://www.mathworks.com/products/gpu-coder.html

Deep Learning Workflow in MATLAB and Simulink

Deep Neural Network Design + Training

Reference model

Transfer learning

Train in MATLAB

Trained DNN

Keras
TensorFlow
ONNX
PyTorch

Model importer

Application Design

Application logic

Standalone Deployment

Coders

intel

MKL-DNN Library

NVIDIA

TensorRT and cuDNN Libraries

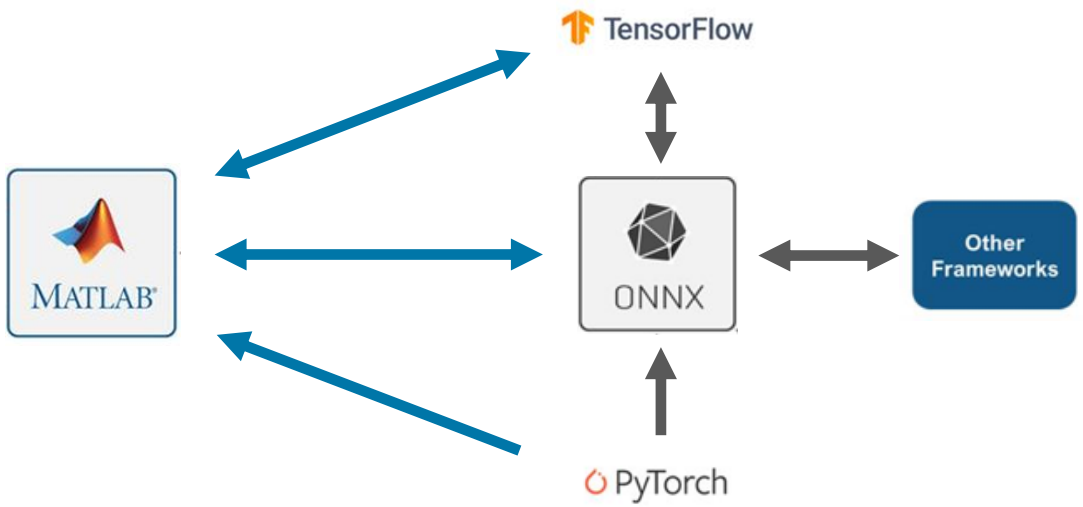
ARM NEON™

ARM Compute Library

Access pretrained models

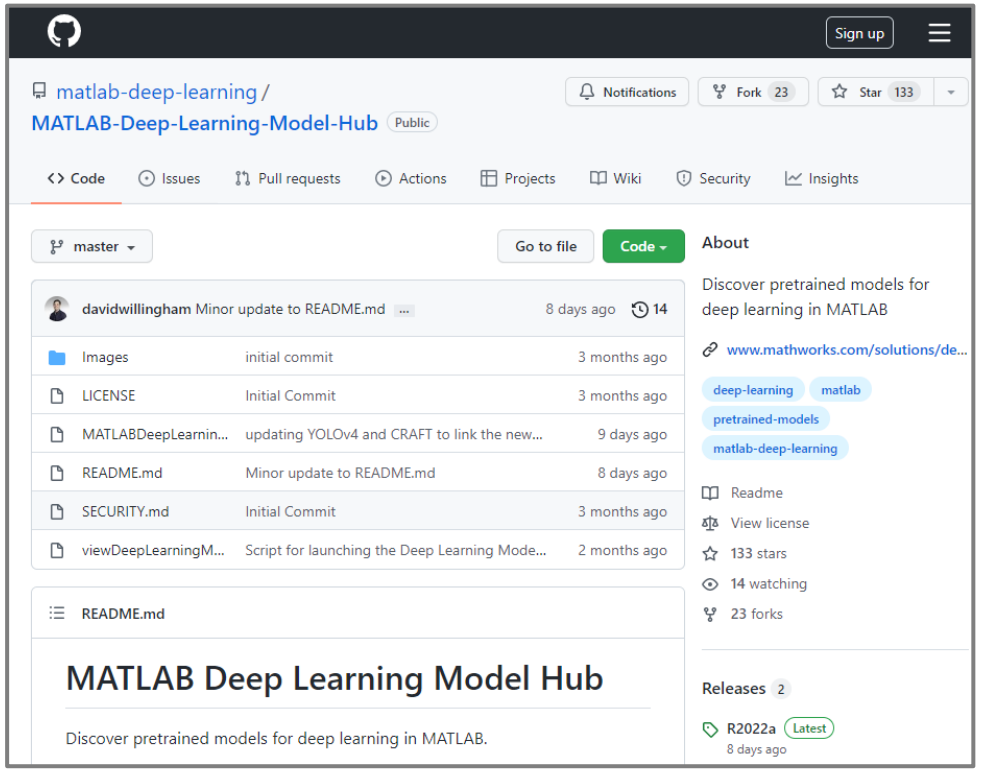
Take advantage of the knowledge provided by pretrained networks to learn new patterns in new data

Import it from other platforms



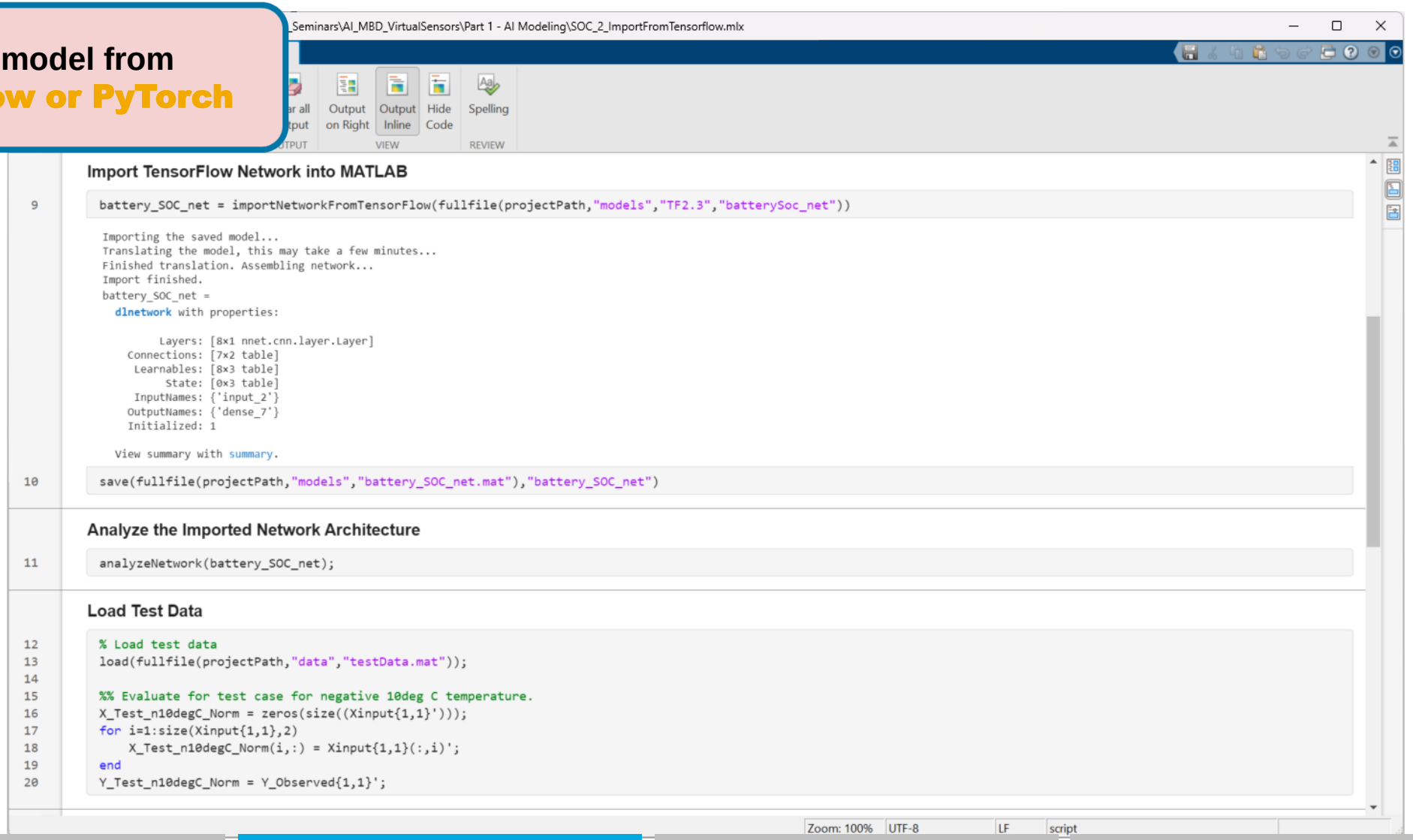
Find one directly in MATLAB

https://github.com/matlab-deep-learning/MATLAB-Deep-Learning-Model-Hub

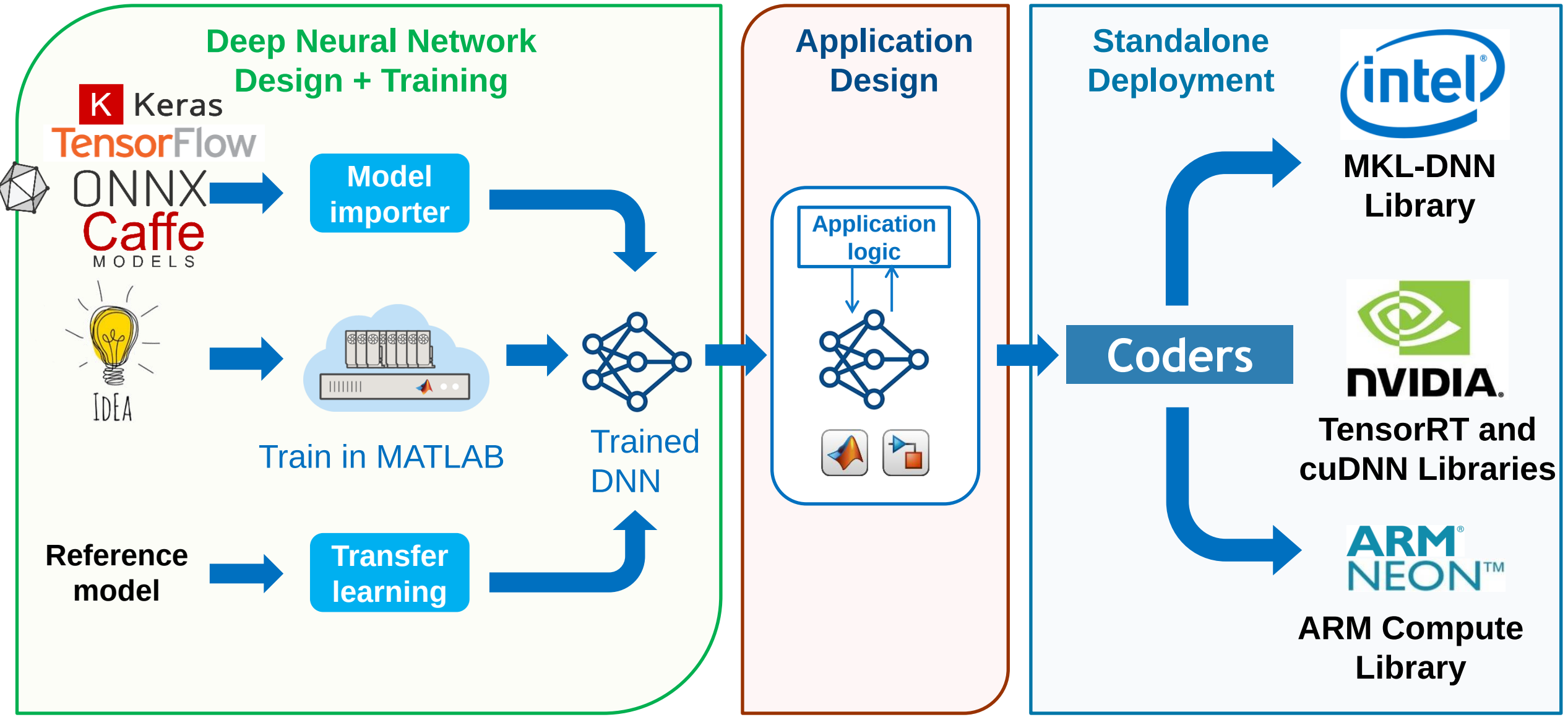


AI Modeling

Import model from
TensorFlow or PyTorch



Deep Learning Workflow in MATLAB and Simulink

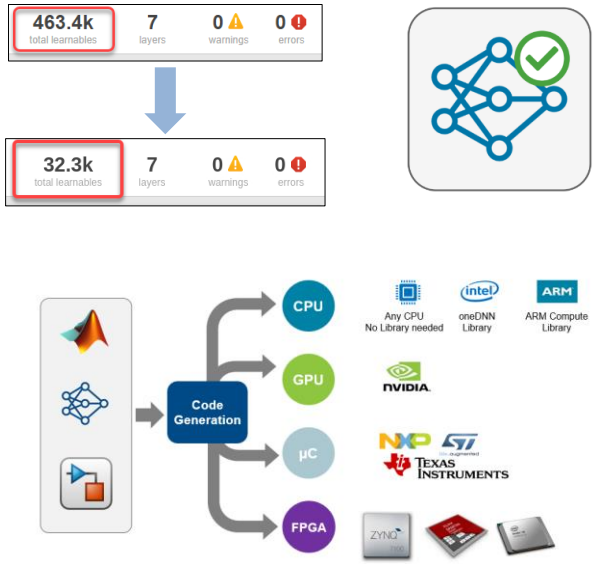


Key takeaways

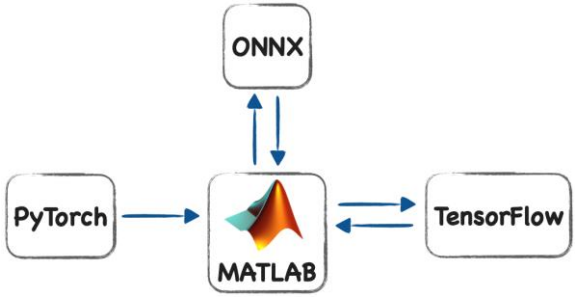
- Easily design and develop AI models and integrate them in system level simulations
- Optimize neural network to fit hardware requirements
- Deployment AI applications on embedded systems using code generation
- Interoperability with open-source AI frameworks

AI + Model-Based Design

Compression, Verification, Code Generation

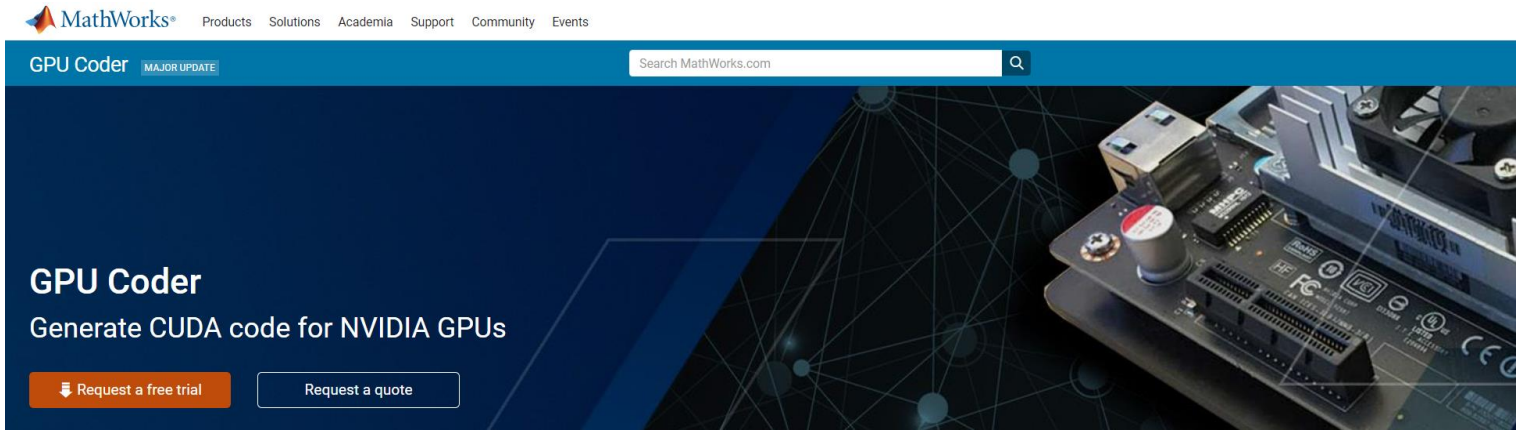


Interoperability



Next Steps

- To learn more, including playing with a trial license, visit:
  - www.mathworks.com/products/gpu-coder
- Read white paper on generating CUDA code from MATLAB
- Attend 4-hour hands-on workshop
  - Arrange with your account manager



Q&A