MATLAB EXPO 2021

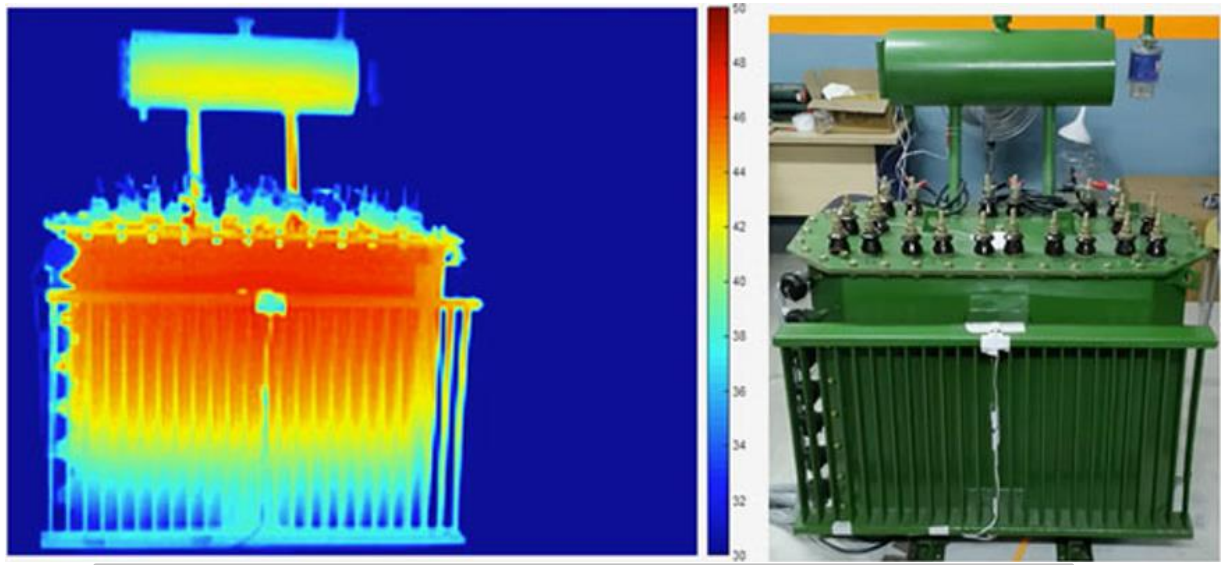
将AI部署到生产系统及MATLAB云工作流

陈宜欣， MathWorks 中国

MathWorks®

部署到嵌入式和企业系统中

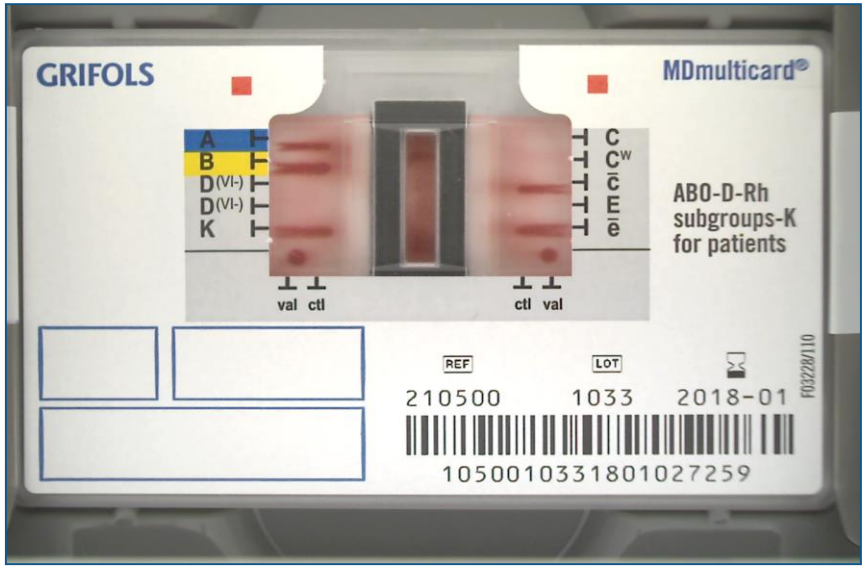
Enterprise



Health Monitoring of Distribution Transformers

SIEMENS

Embedded



Card to Classify Blood Type

IDNEO

Agenda

将AI部署到生产系统中

- Three specific challenges:
  - Limitations of Embedded hardware
  - Ongoing changes in environment or system behavior
  - Scale to production load in Enterprise systems

AI云工作流

- Data preparation, AI model design, AI model tuning, deployment

Agenda

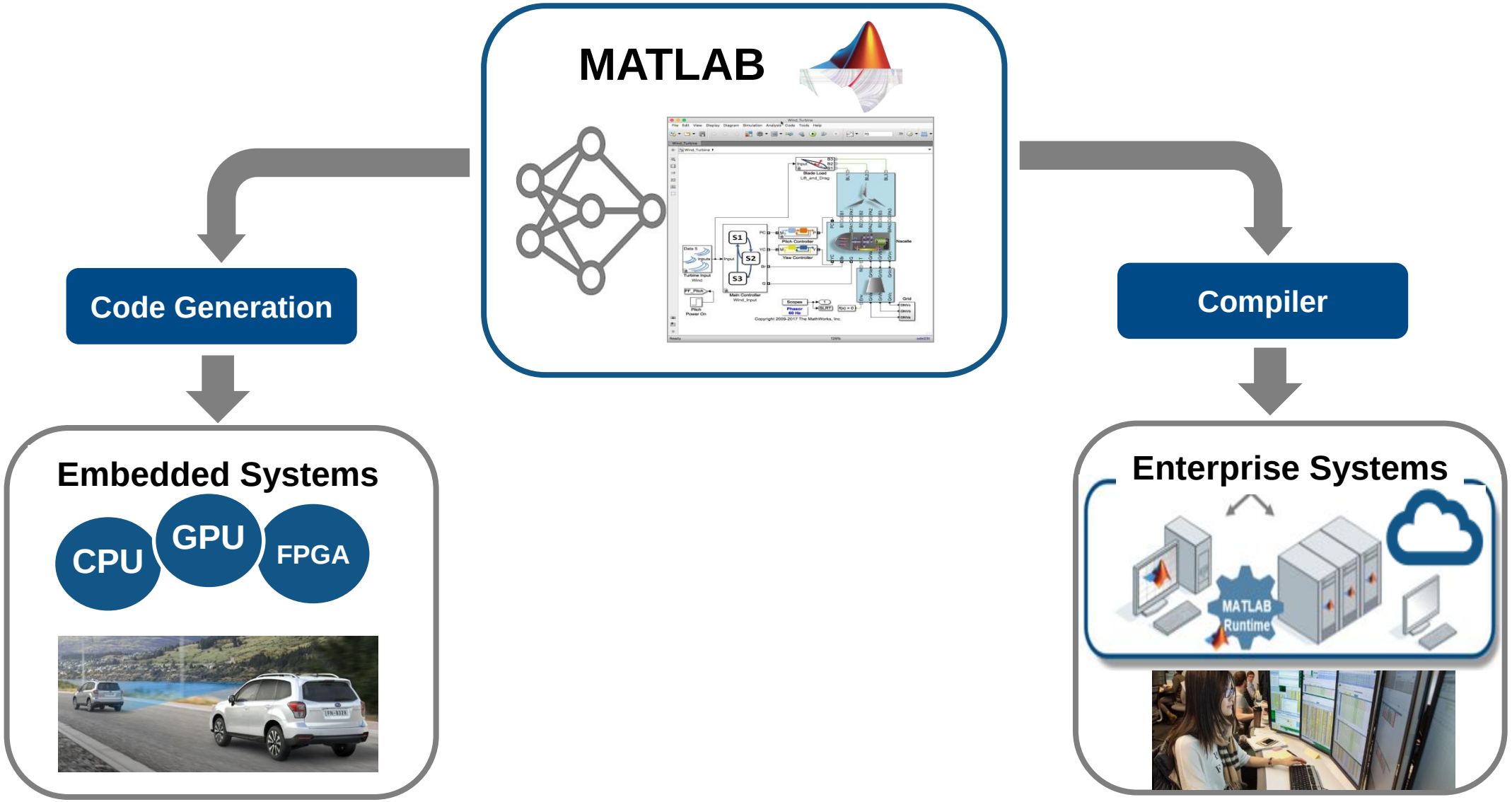
将AI部署到生产系统中

- Three specific challenges:
  - Limitations of Embedded hardware
  - Ongoing changes in environment or system behavior
  - Scale to production load in Enterprise systems

AI云工作流

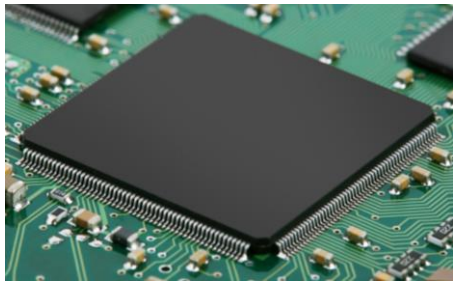
- Data preparation, AI model design, AI model tuning, deployment

将AI集成到更大系统中的两种方法



声景识别嵌入式部署

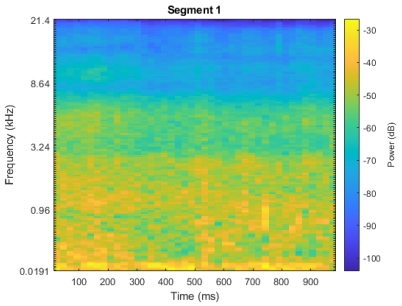
Squeezenet ~5MB
ResNet-50 ~100MB



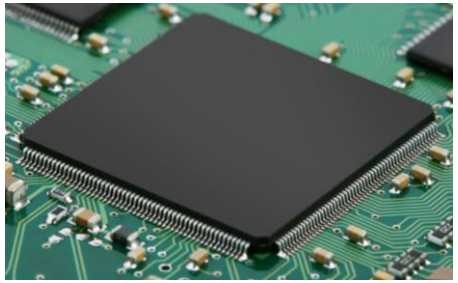
Limited resources

声景识别嵌入式部署

Reformat the data



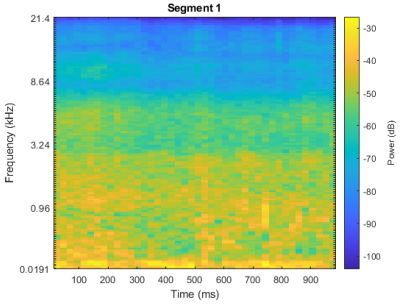
Squeezenet ~5MB
ResNet-50 ~100MB



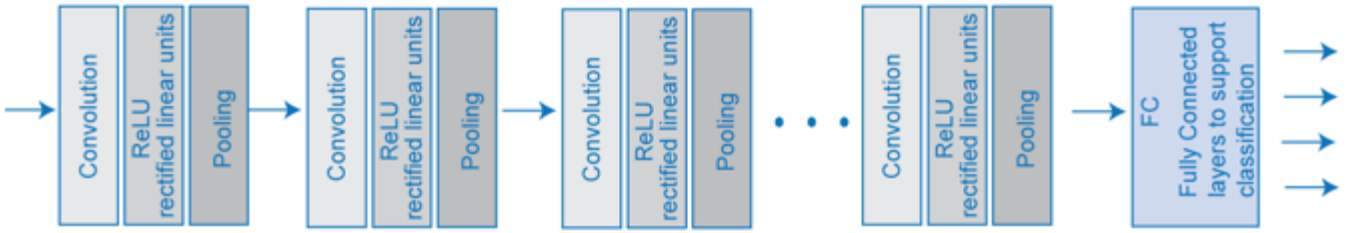
Limited resources

声景识别嵌入式部署

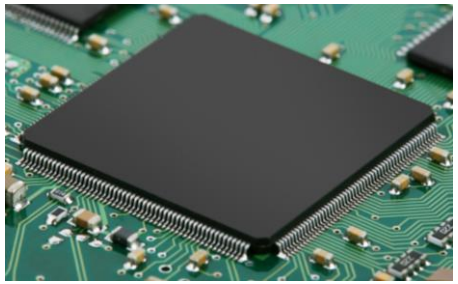
Reformat the data



Convolutional Neural Networks (**CNN**)



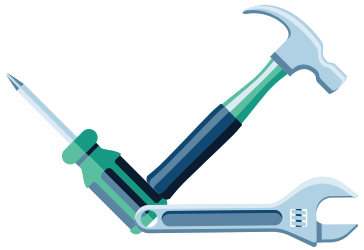
Squeezenet ~5MB
ResNet-50 ~100MB



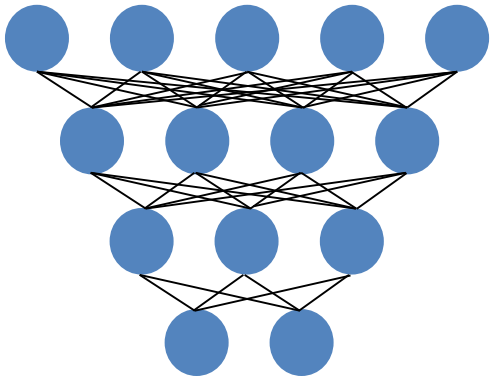
Limited resources

怎样实现嵌入式部署？

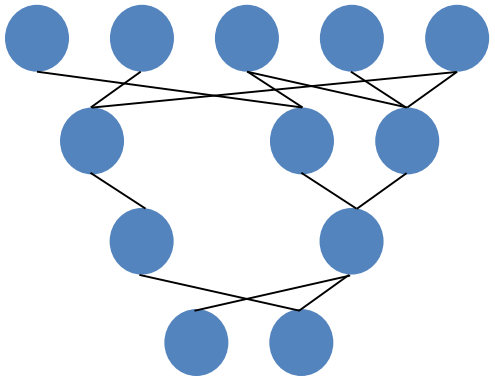
怎样实现嵌入式部署？



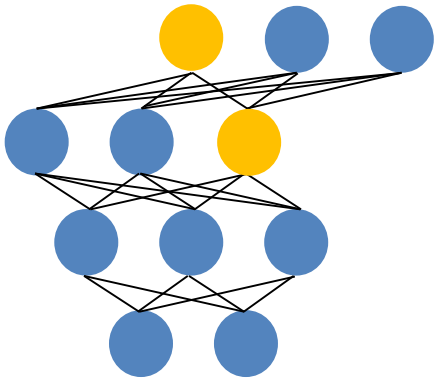
Original



Pruning



Layer Fusion

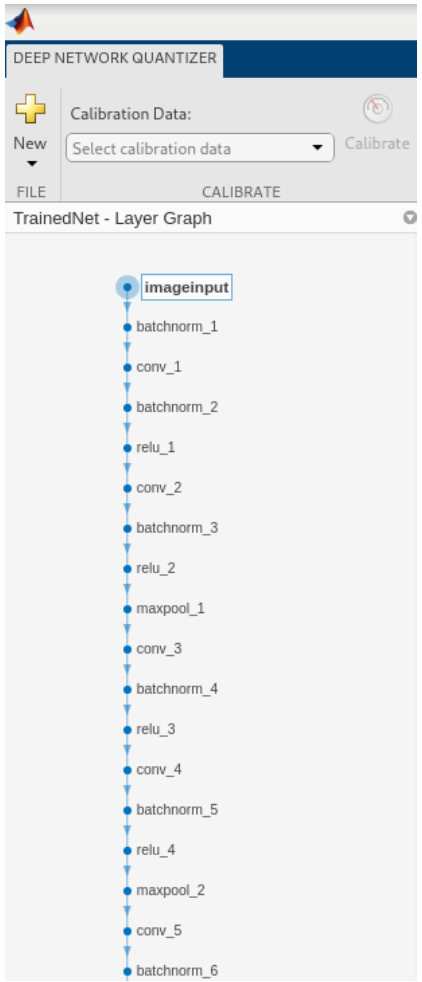


Quantizing

深度学习量化：声景分类


Calibration dataset
Trained 'Single' DNN
Calibrate
Quantize
Quantized 'INT8' DNN
Validate

深度学习量化：声景分类

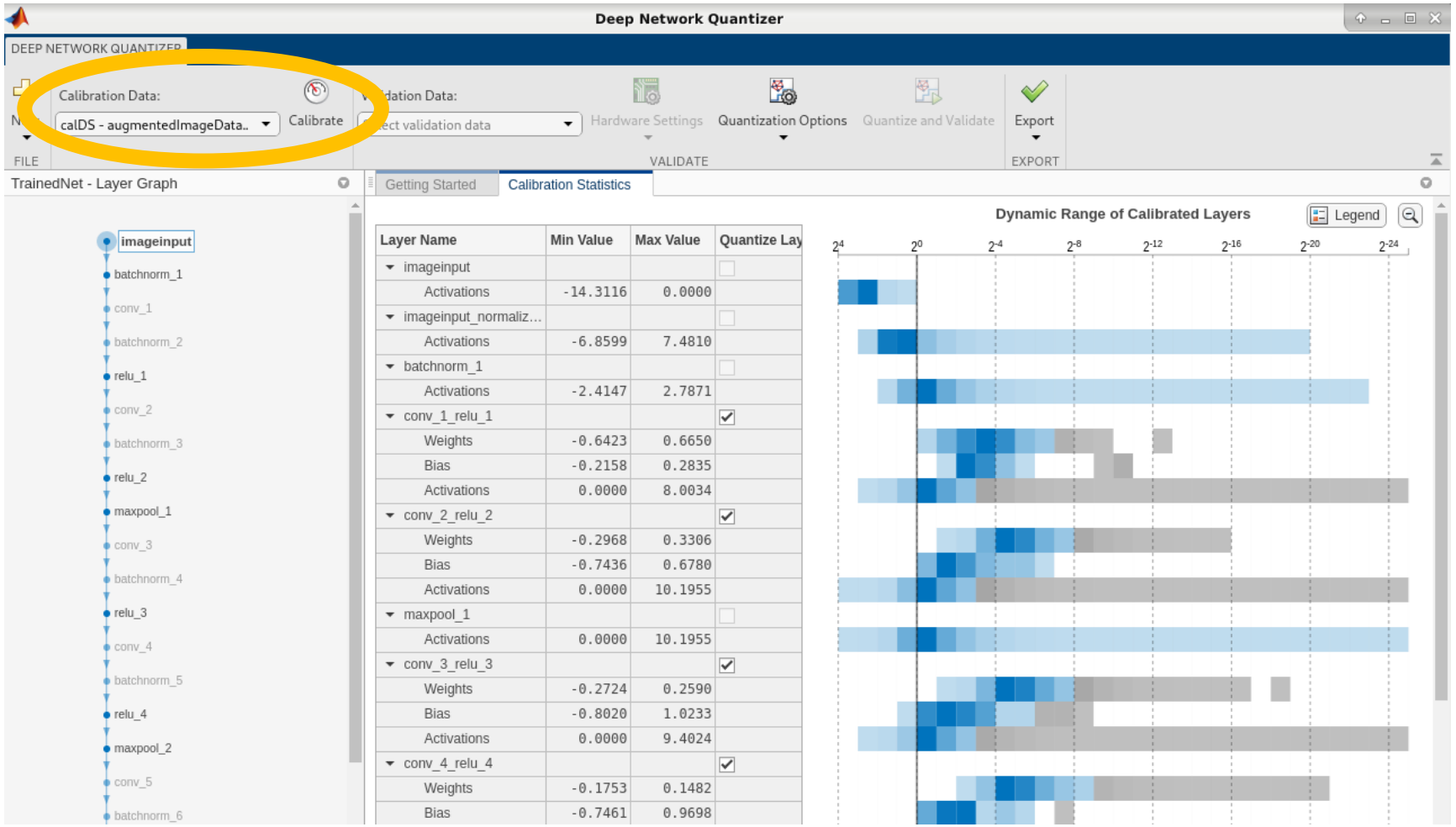


Use Deep Network Quantizer to Optimize the Inference Network

```
load('trainedNet');
analyzeNetwork(trainedNet);
numData = size(xTrain);
numData = numData(end);
augImds = augmentedImageDatastore(trainedNet.Layers(1).InputSize, xTrain, yTrain);
calDS = augImds.subset(1:floor(numData * 0.8));
valDS = augImds.subset(floor(numData * 0.8)+1:numData);
dq = dlquantizer(trainedNet, 'ExecutionEnvironment', 'GPU');
dq.calibrate(calDS)
```

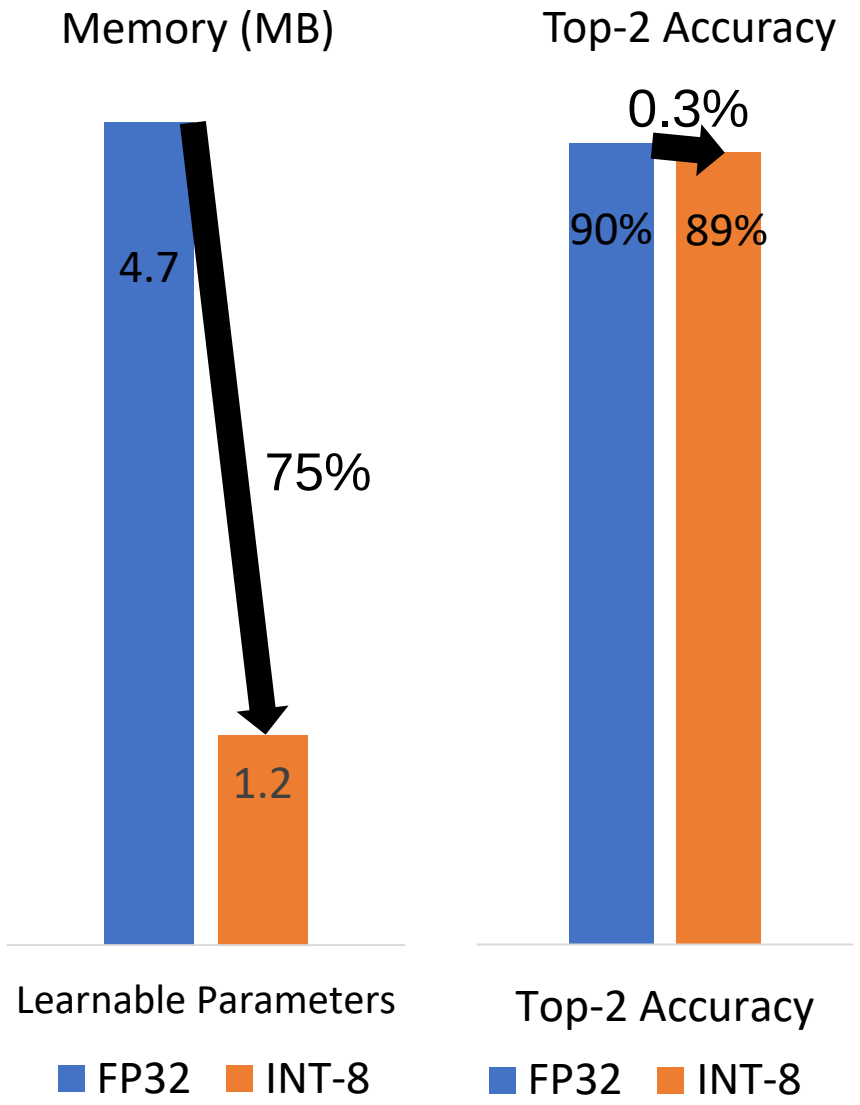
- Load trained network
- Split data: calibration – 80%, validation – 20%
- Launch Deep Network Quantizer App

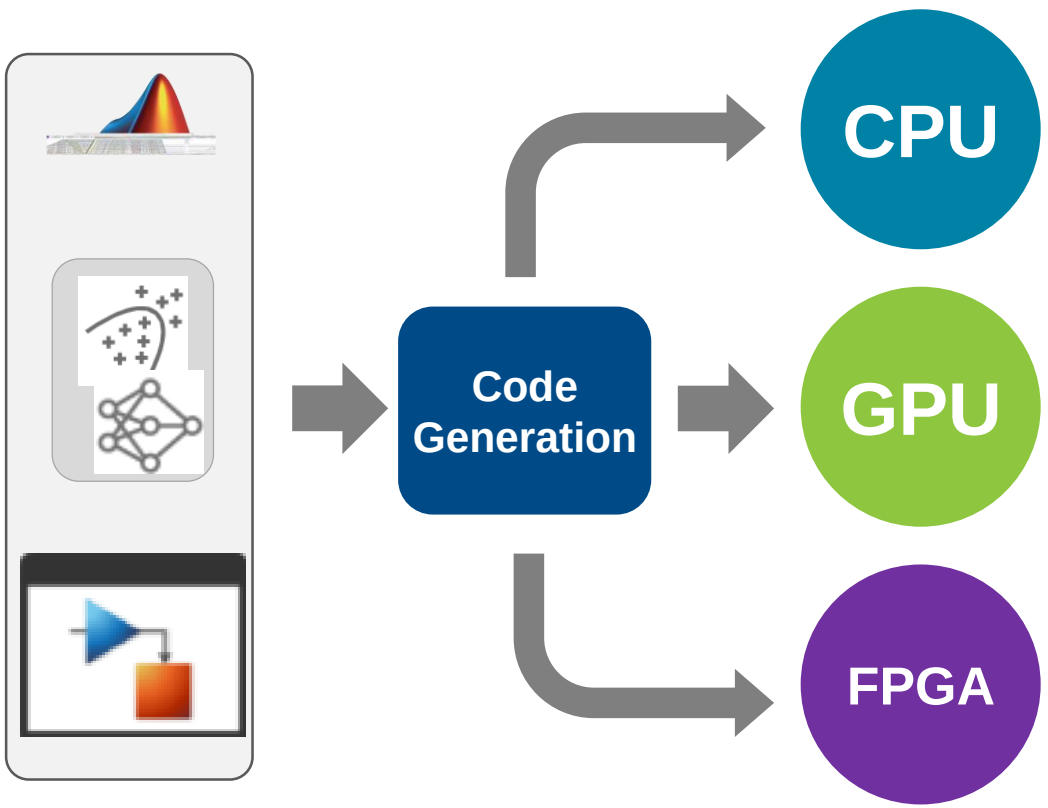
深度学习量化：声景分类


Deep Network Quantizer
DEEP NETWORK QUANTIZER
Calibration Data:
calDS - augmentedImageData..
Calibrate
Hardware Settings
Quantization Options
Quantize and Validate
Export
FILE
VALIDATE
EXPORT
TrainedNet - Layer Graph
imageinput
batchnorm_1
conv_1
batchnorm_2
relu_1
conv_2
batchnorm_3
relu_2
maxpool_1
conv_3
batchnorm_4
relu_3
conv_4
batchnorm_5
relu_4
maxpool_2
conv_5
batchnorm_6
Getting Started
Calibration Statistics
Layer Name
Min Value
Max Value
Quantize Lay
imageinput
Activations -14.3116 0.0000
imageinput_normaliz...
Activations -6.8599 7.4810
batchnorm_1
Activations -2.4147 2.7871
conv_1_relu_1
Weights -0.6423 0.6650
Bias -0.2158 0.2835
Activations 0.0000 8.0034
conv_2_relu_2
Weights -0.2968 0.3306
Bias -0.7436 0.6780
Activations 0.0000 10.1955
maxpool_1
Activations 0.0000 10.1955
conv_3_relu_3
Weights -0.2724 0.2590
Bias -0.8020 1.0233
Activations 0.0000 9.4024
conv_4_relu_4
Weights -0.1753 0.1482
Bias -0.7461 0.9698
Dynamic Range of Calibrated Layers
Legend

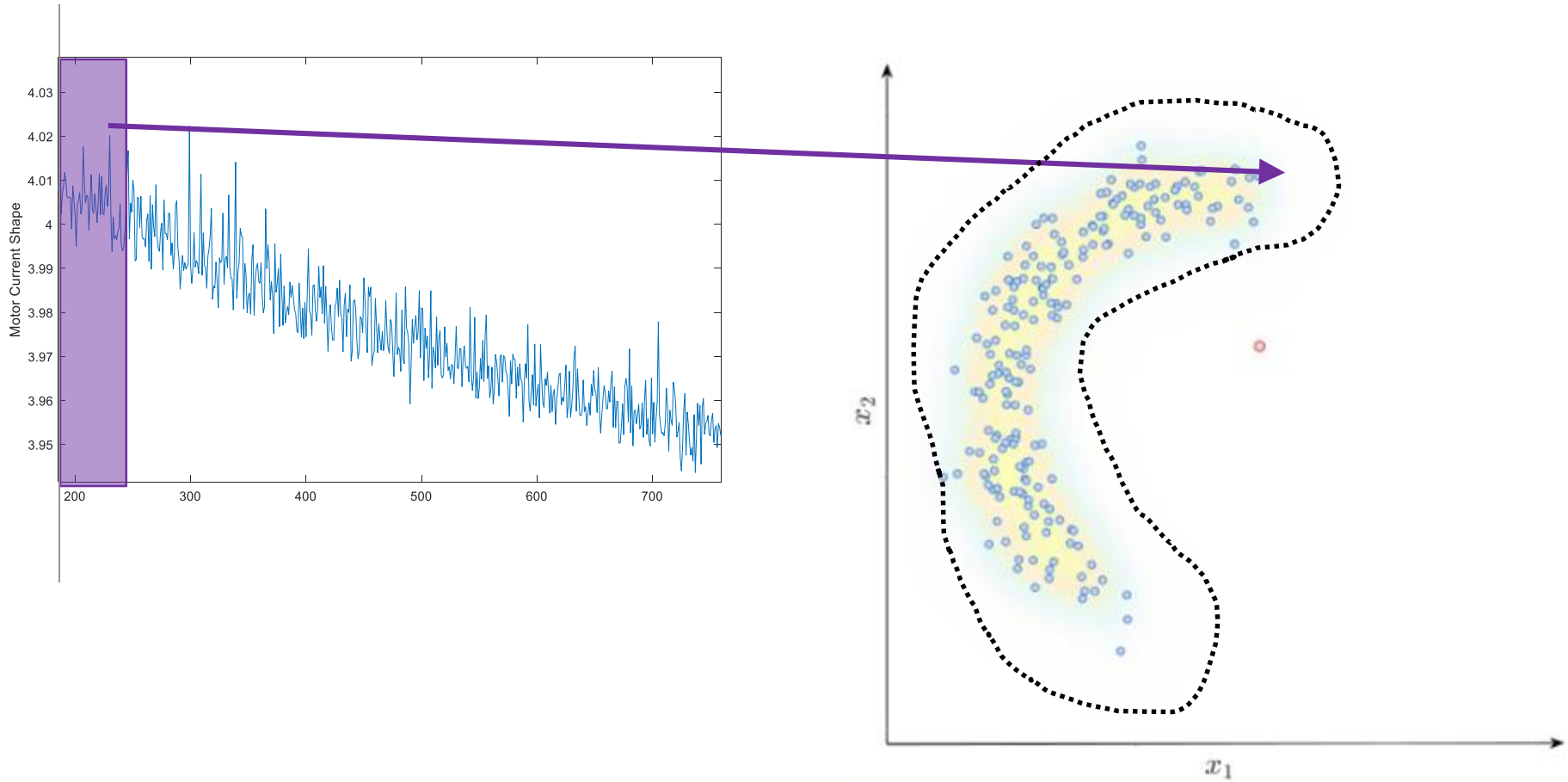
深度学习量化：声景分类

Validation Results



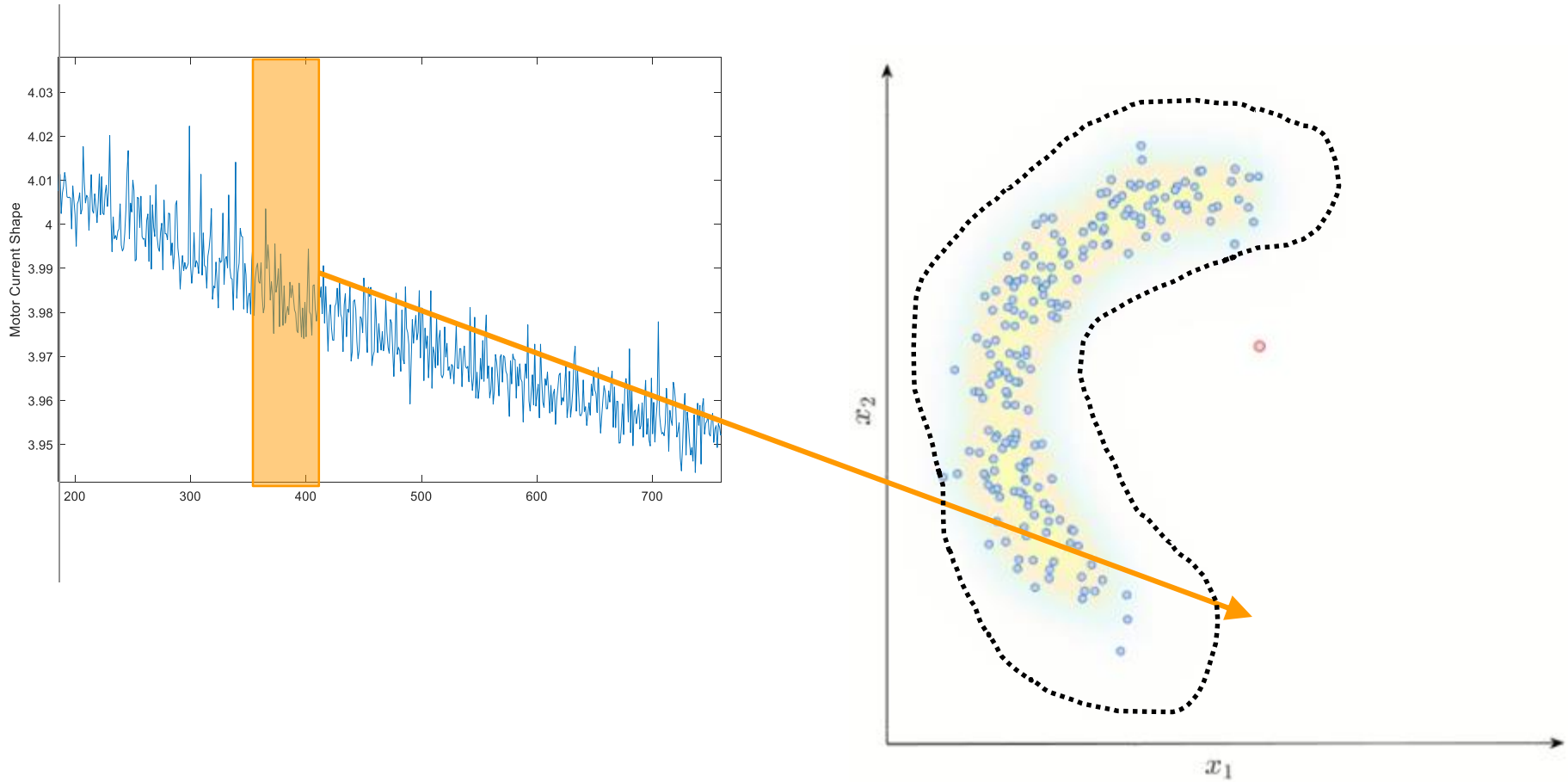


AI模型反映系统行为和环境

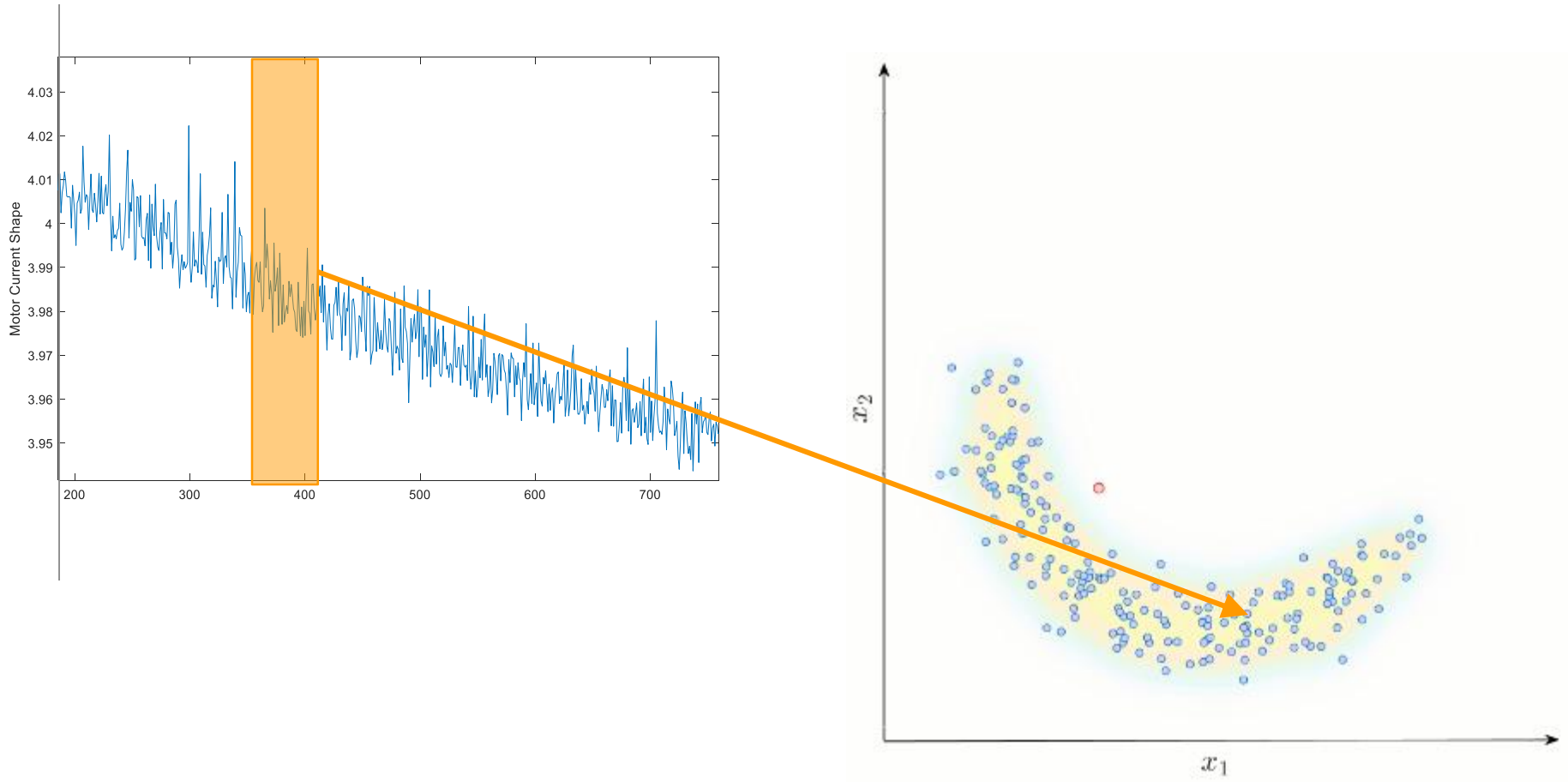


(illustration only; not based on actual data)

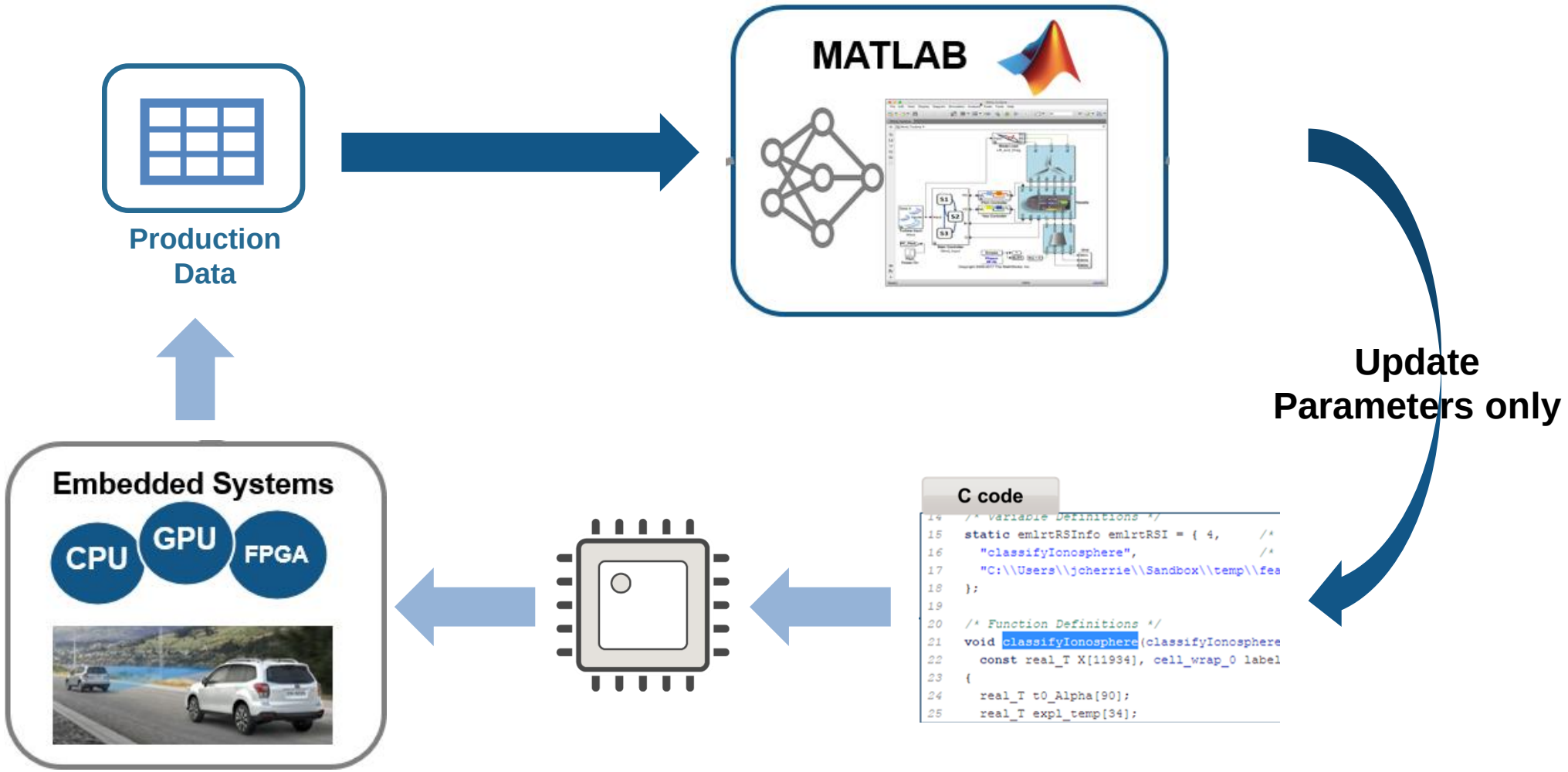
AI模型反映系统行为和环境



调整已部署的模型


Motor Current Shape
4.03
4.02
4.01
4
3.99
3.98
3.97
3.96
3.95
200
300
400
500
600
700
x_2
x_1

嵌入式部署模型更新


Production Data
MATLAB
Update Parameters only
C code
/* Variable Definitions */
static emlrtRSInfo emlrtRSI = { 4,
"classifyIonosphere",
"C:\\Users\\jcherrie\\Sandbox\\temp\\fea
};
/* Function Definitions */
void classifyIonosphere(classifyIonosphere
const real_T X[11934], cell_wrap_0 label
{
real_T t0_Alpha[90];
real_T expl_temp[34];
Embedded Systems
CPU
GPU
FPGA

Agenda

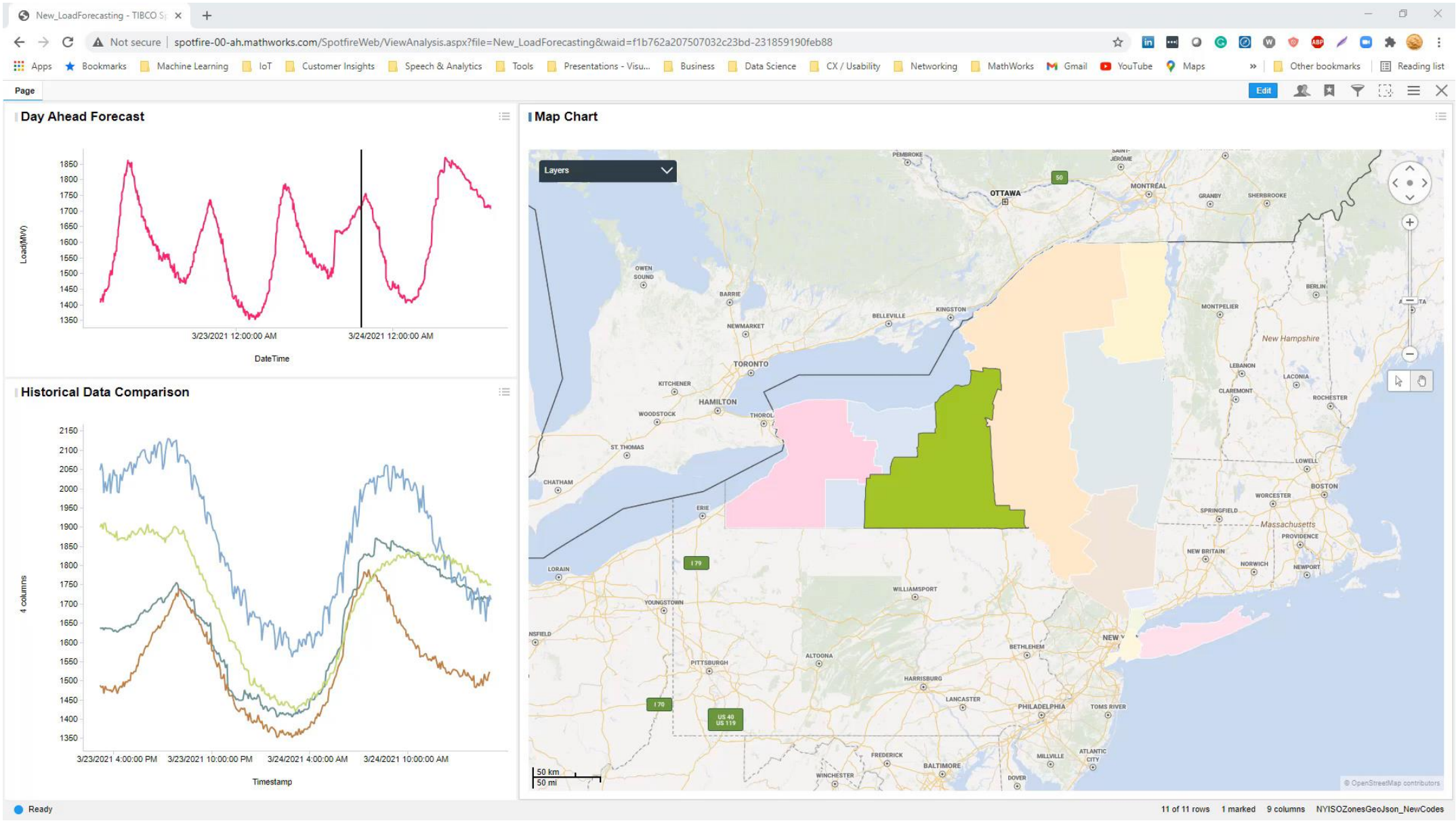
将AI部署到生产系统中

- Three specific challenges:
  - Limitations of Embedded hardware ✓
  - Ongoing changes in environment or system behavior ✓
  - Scale to production load in Enterprise systems

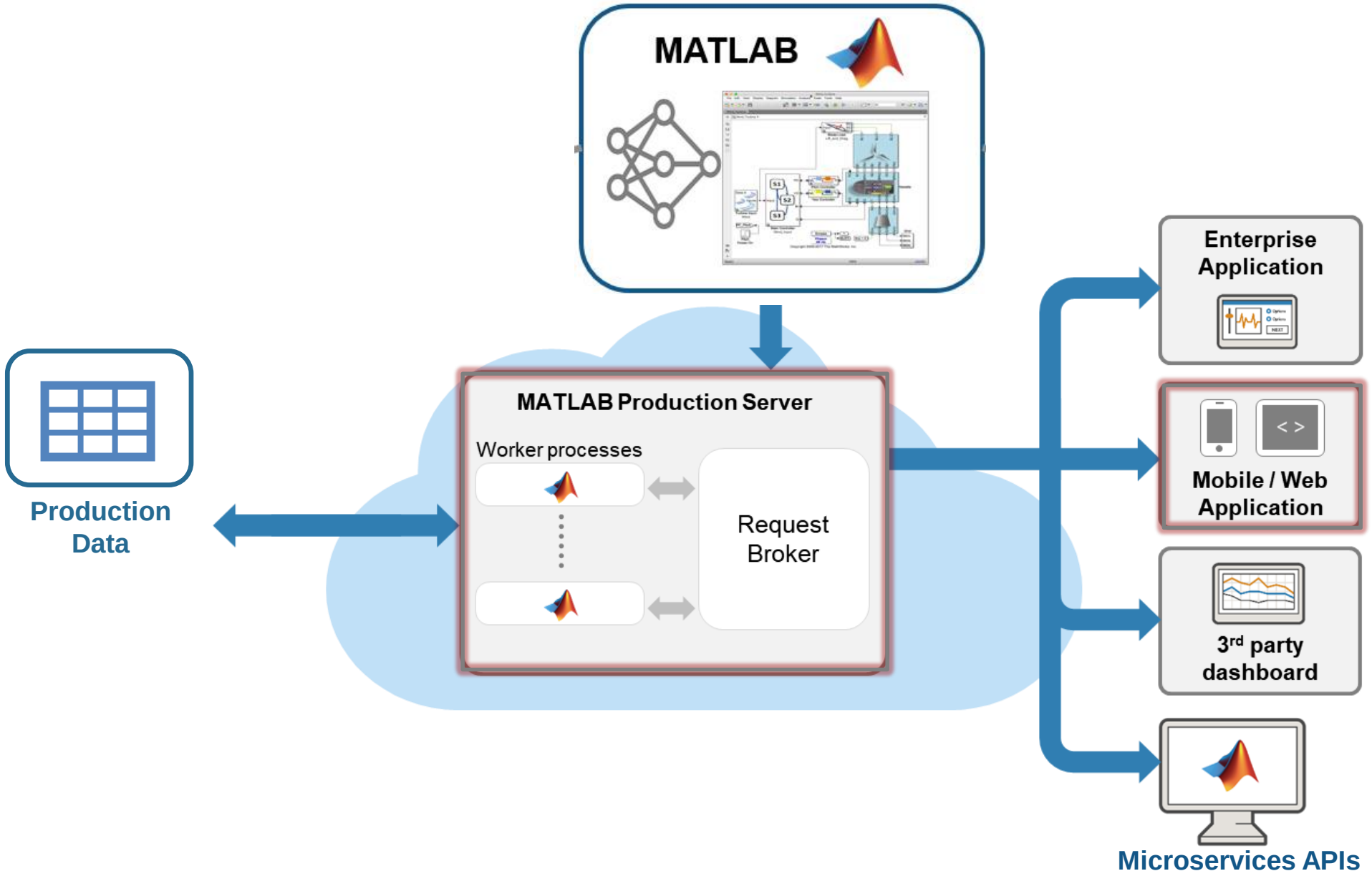
AI云工作流

- Data preparation, AI model design, AI model tuning, deployment

AI企业部署

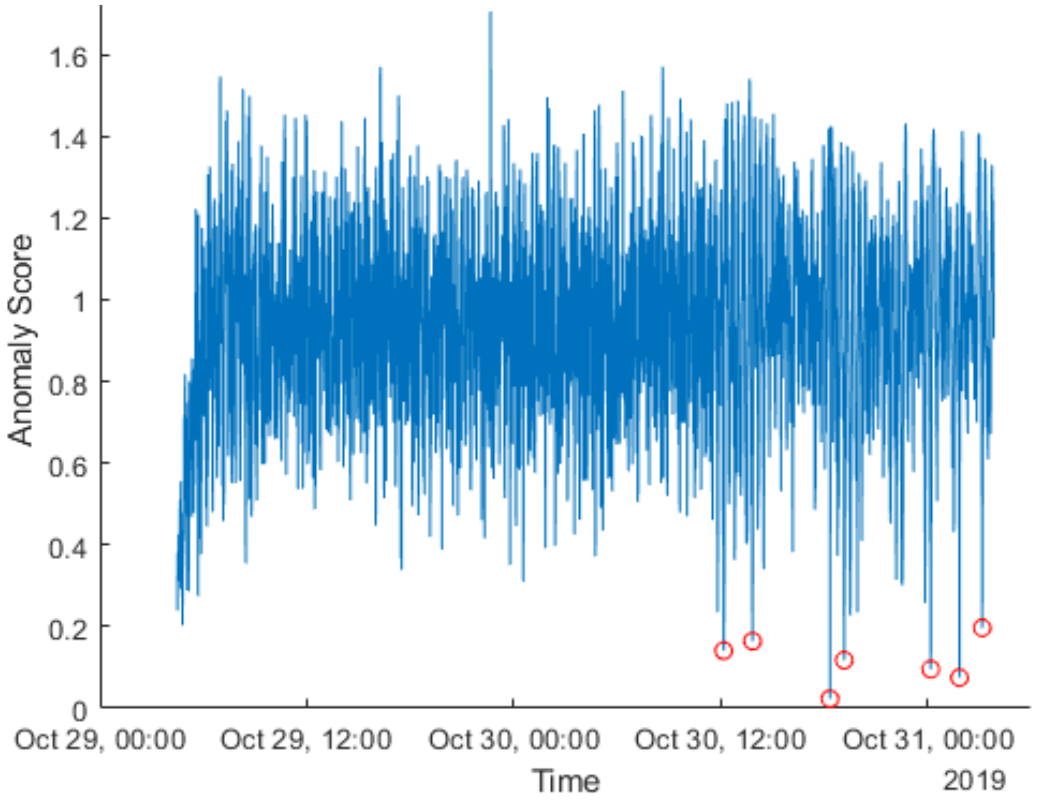

New_LoadForecasting - TIBCO S
Not secure | spotfire-00-ah.mathworks.com/SpotfireWeb/ViewAnalysis.aspx?file=New_LoadForecasting&waid=f1b762a207507032c23bd-231859190feb88
Apps
Bookmarks
Machine Learning
IoT
Customer Insights
Speech & Analytics
Tools
Presentations - Visu...
Business
Data Science
CX / Usability
Networking
MathWorks
Gmail
YouTube
Maps
Other bookmarks
Reading list
Page
Edit
Day Ahead Forecast
Load(MW)
1850
1800
1750
1700
1650
1600
1550
1500
1450
1400
1350
3/23/2021 12:00:00 AM
3/24/2021 12:00:00 AM
DateTime
Historical Data Comparison
4 columns
2150
2100
2050
2000
1950
1900
1850
1800
1750
1700
1650
1600
1550
1500
1450
1400
1350
3/23/2021 4:00:00 PM
3/23/2021 10:00:00 PM
3/24/2021 4:00:00 AM
3/24/2021 10:00:00 AM
Timestamp
Map Chart
Layers
OTTAWA
MONTRÉAL
TORONTO
HAMILTON
BELLEVILLE
KINGSTON
BARRIE
OWEN SOUND
KITCHENER
WOODSTOCK
ST. THOMAS
CHATHAM
ERIE
LORAIN
YOUNGSTOWN
PITTSBURGH
ALTOONA
WILLIAMSPORT
HARRISBURG
LANCASTER
PHILADELPHIA
BETHLEHEM
TOMS RIVER
ATLANTIC CITY
MILLVILLE
DOVER
BALTIMORE
FREDERICK
WINCHESTER
I 79
I 70
US 40
US 119
New Hampshire
Massachusetts
MONTPELIER
BERLIN
LEBANON
LACONIA
CLAREMONT
ROCHESTER
LOWELL
BOSTON
WORCESTER
SPRINGFIELD
PROVIDENCE
NEW BRITAIN
NORWICH
NEWPORT
50 km
50 mi
© OpenStreetMap contributors
Ready
11 of 11 rows
1 marked
9 columns
NYISOZonesGeoJson_NewCodes

与企业系统集成并扩展到生产负载



Example: 增量健康监控

Sensor data



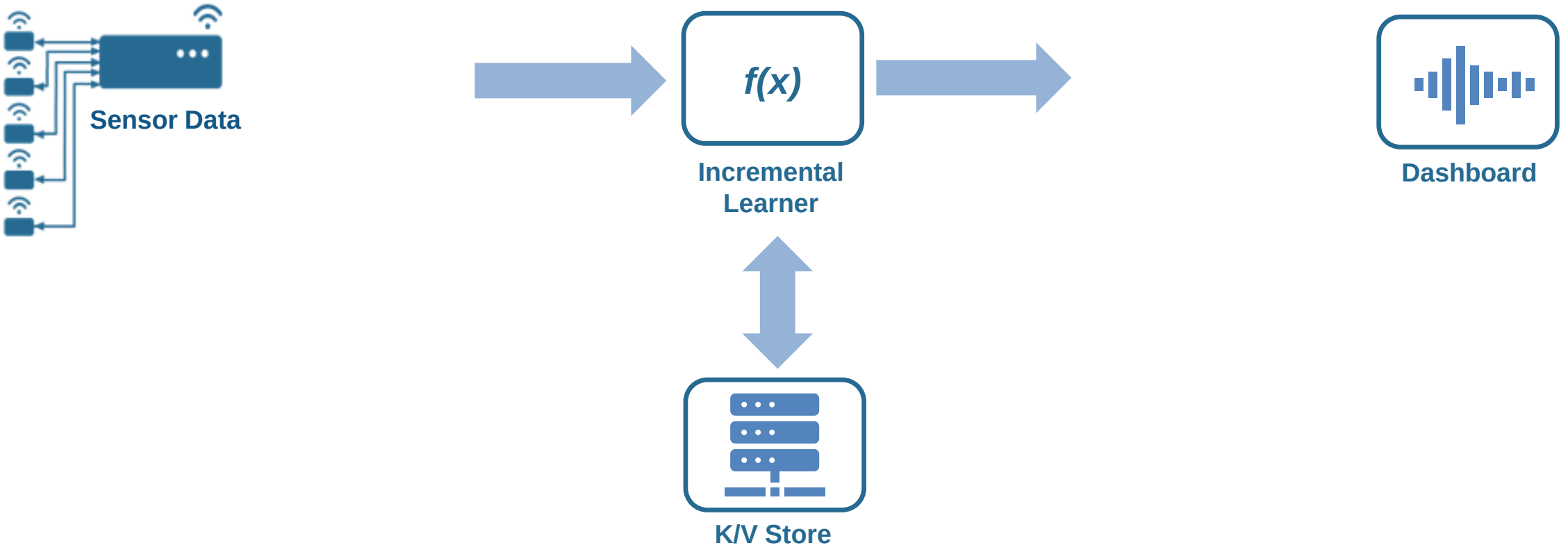
Anomaly Detection loop

```
while seqn % ... there's more data to process

    % Retrieve buffer of data
    datafilter = (sensordata.key == thisAsset) & (sensordata.Seque
        (sensordata.SequenceNumber <= seqn+batchsize);

    streamdata = sensordata(datafilter,:);

    % Detect Anomalies with incremental One-class SVM
    [nextState, results] = detectAnomalyLocal(streamdata, state);

    % Remember results and update state of incremental learner
    anomalies(datafilter) = results.anomaly;
    score(datafilter) = results.score;
    timestamps(datafilter) = results.timestamp;
    state = nextState;

    seqn = seqn + batchsize; % step through batch test data
end
```

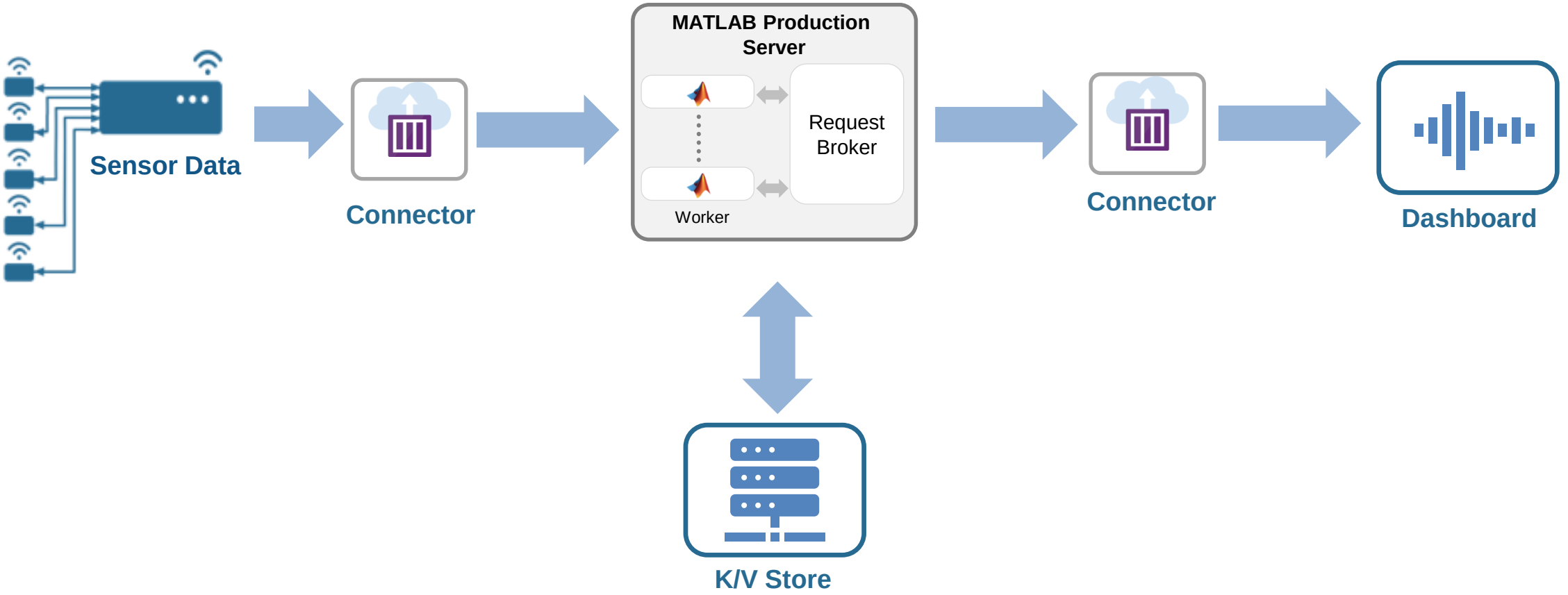
流媒体架构中的增量学习

```
incMdl = incrementalLearner(mdl);

while dataStreaming
   featureChunk = extractFeatures(streamdata);
   inclMdl = updateMetricsAndFit(incMdl,featureChunk,labels);
End
```

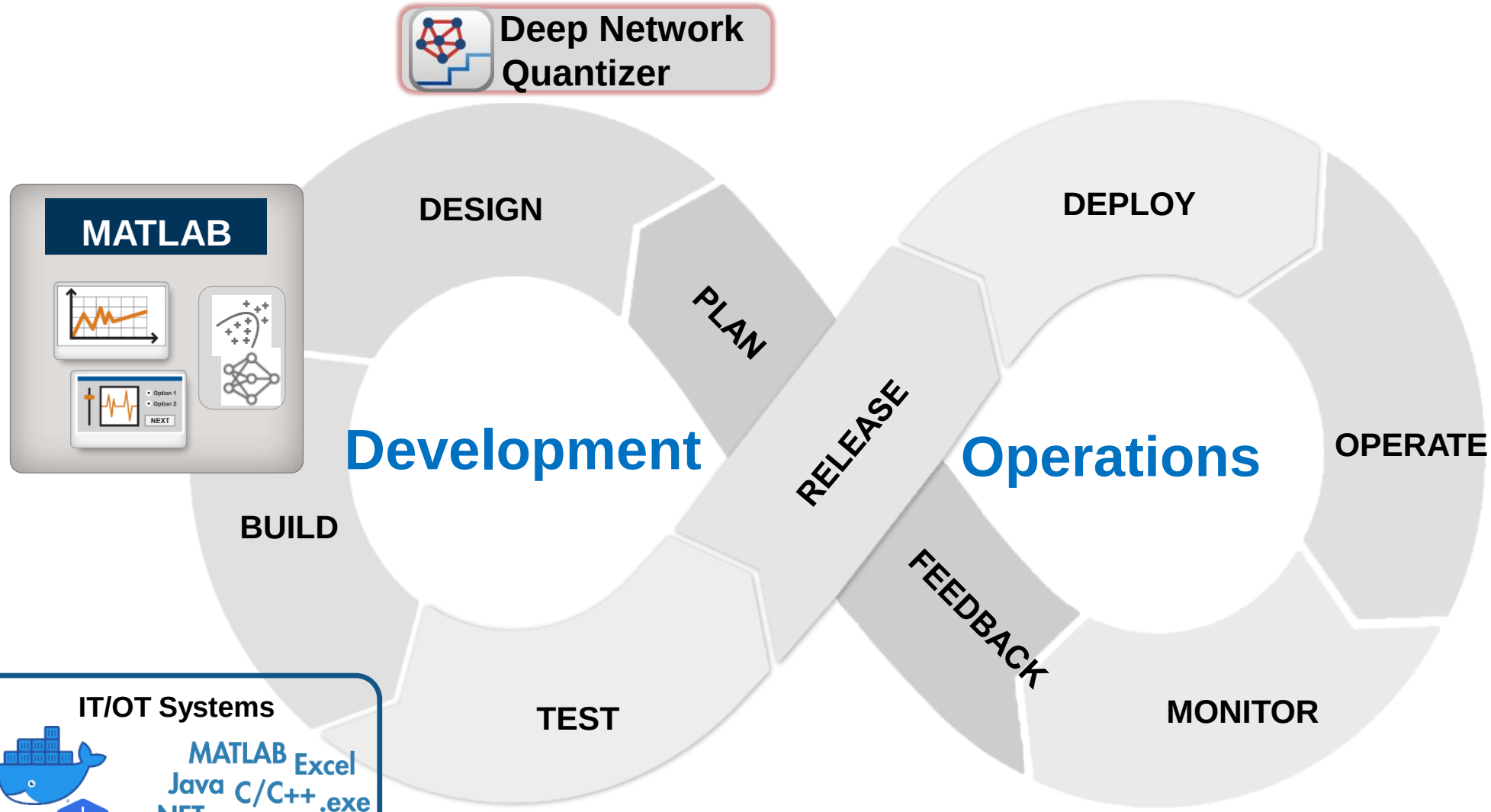
R2020b



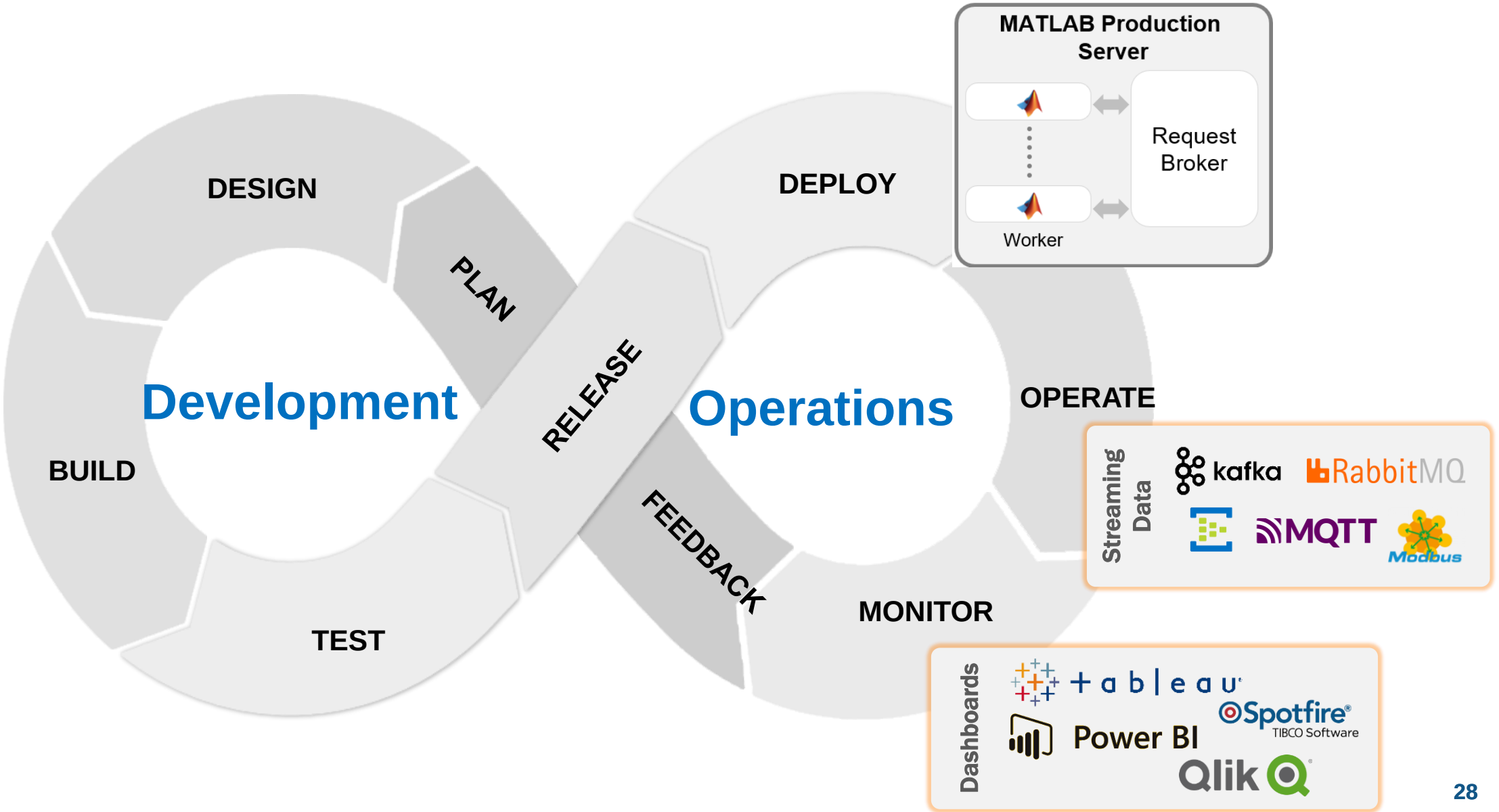
流媒体架构中的增量学习



无需重新编码即可操作 AI



无需重新编码即可操作 AI- Model DevOps



Agenda

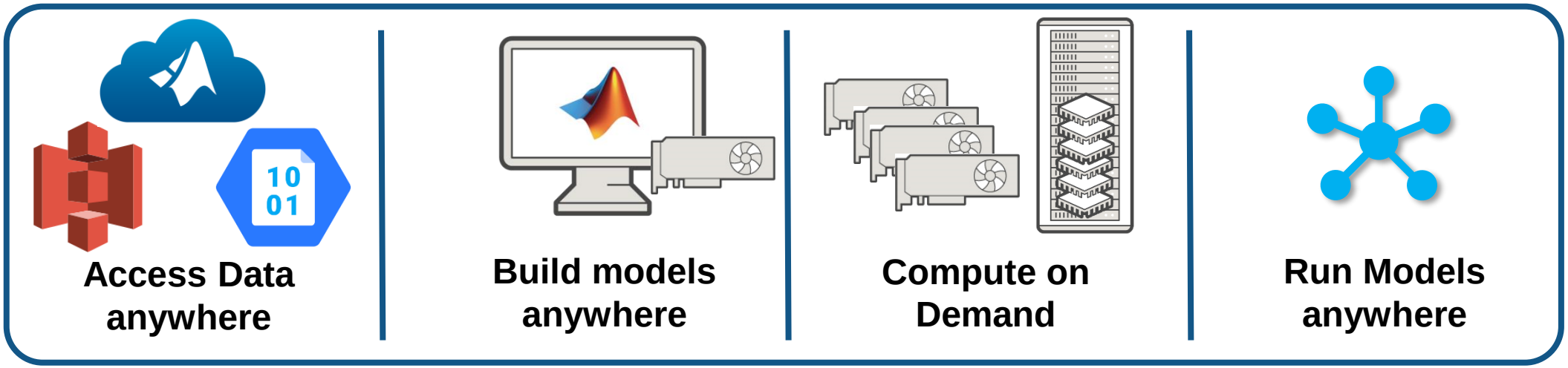
将AI部署到生产系统

- Three specific challenges:
  - Limitations of Embedded hardware
  - Ongoing changes in environment or system behavior
  - Scale to production load in Enterprise systems

AI云工作流

- Data preparation, AI model design, AI model tuning, deployment

为什么要在云端实现AI?



AI系统设计工作流

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

31

云端 AI 系统设计是什么样的？

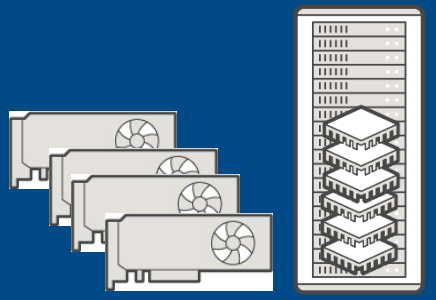
Data Preparation

Access data anywhere

AI Model Design

Build models anywhere

AI Model Tuning



Compute on demand

Deployment

Run models anywhere

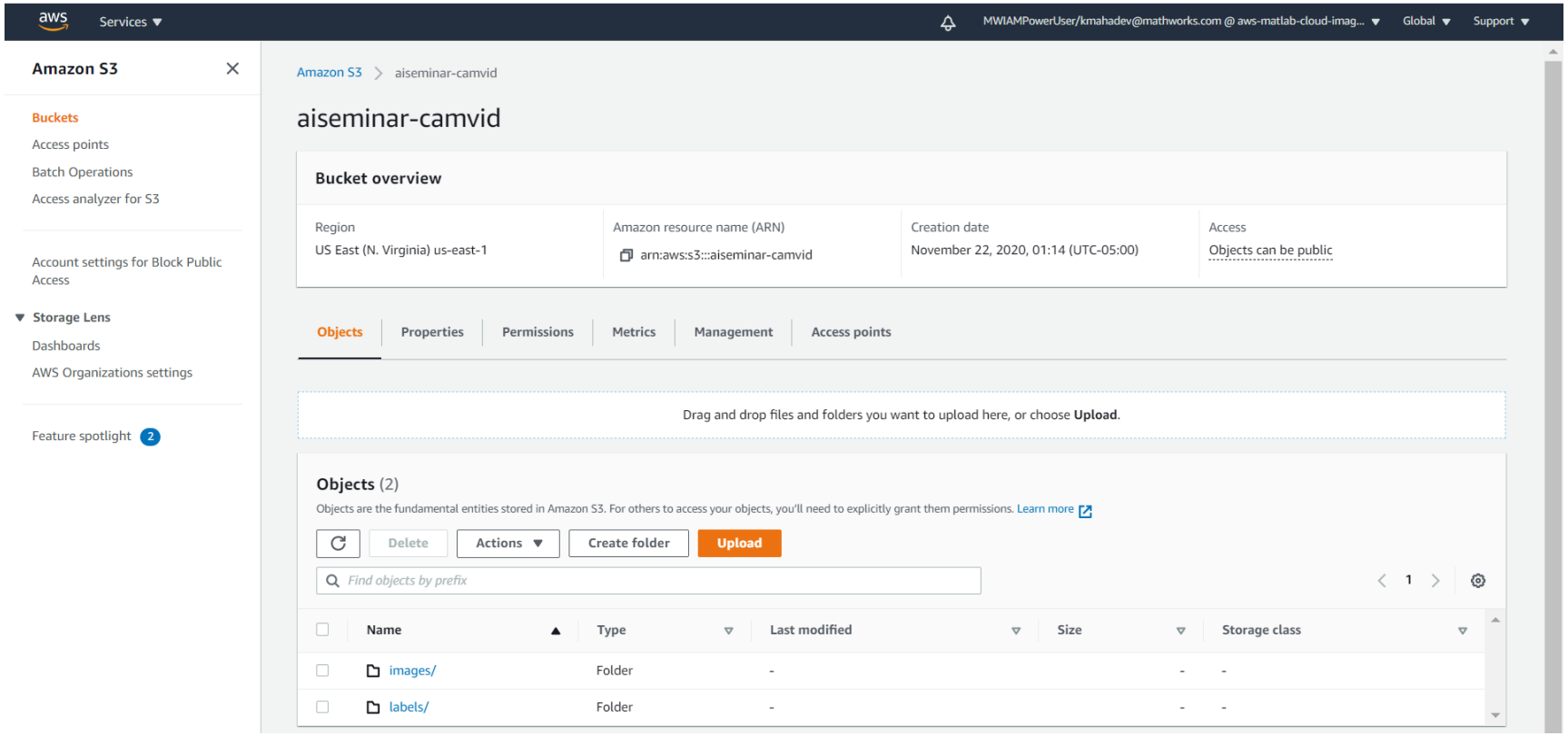
云端数据 = 数据随处可访问

Enabling Shareable, Scalable and Secure storage

- **Shareable**
  - All you need is the URL
- **Scalable**
  - Deep Learning Data Sets can get BIG.
  - Need more storage? No problem.
- **Secure**
  - You need “Keys” to lock and unlock the data

云中的数据 - 使其靠近计算

Uploading Data to the S3



35

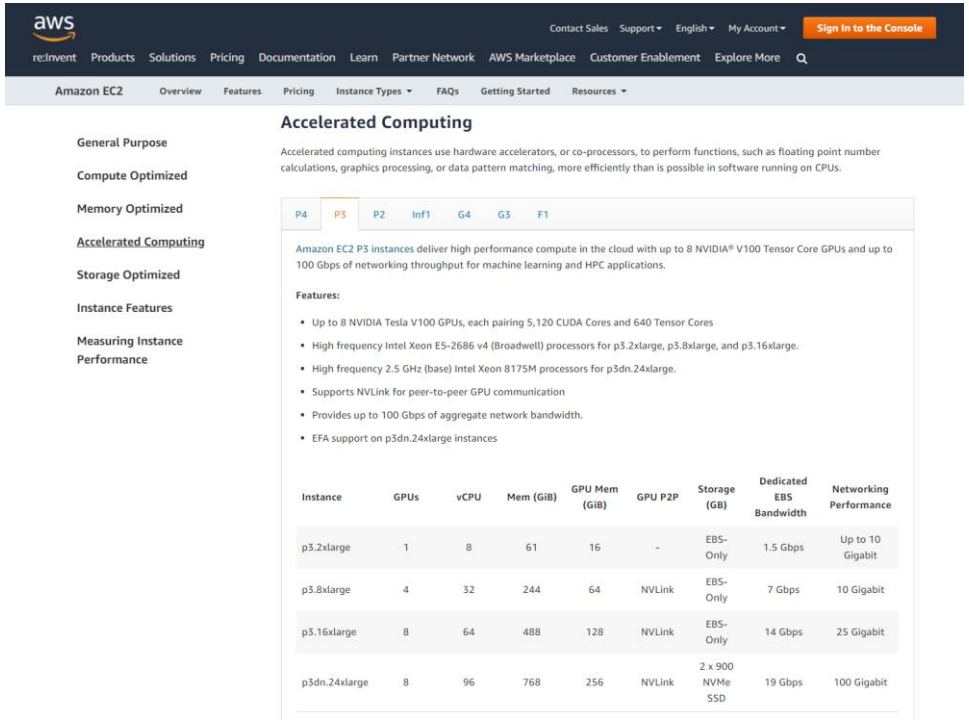
轻松访问HPC资源

It’s a balance between speed and cost

- Which VM did you choose?

AI Model Design

Easy access to a GPU



轻松访问GPU资源

Options chosen for setting up MATLAB

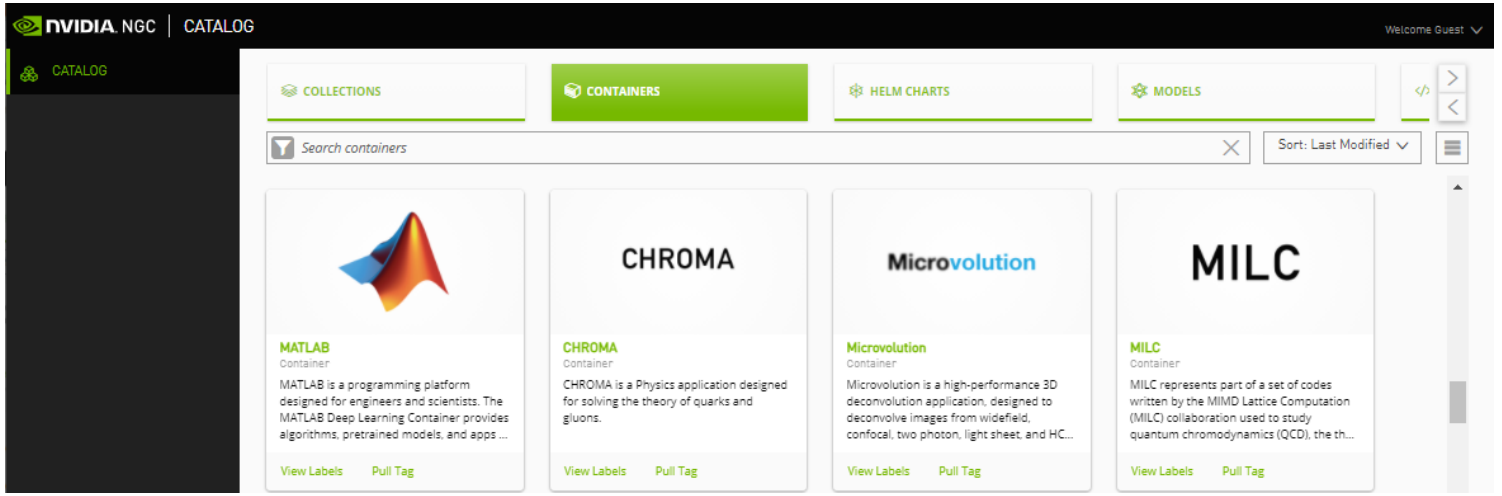
AI Model Design

Easy access to a GPU

- Virtual Machine on AWS:

Instance	GPUs	vCPU	Mem (GiB)	GPU Mem (GiB)	GPU P2P	Storage (GB)	Dedicated EBS Bandwidth	Networking Performance
p3.2xlarge	1	8	61	16	-	EBS-Only	1.5 Gbps	Up to 10 Gigabit

- MATLAB Deep Learning Container on NVIDIA NGC store



MATLAB 中的深度学习云设置

Steps to using the Deep Learning Container

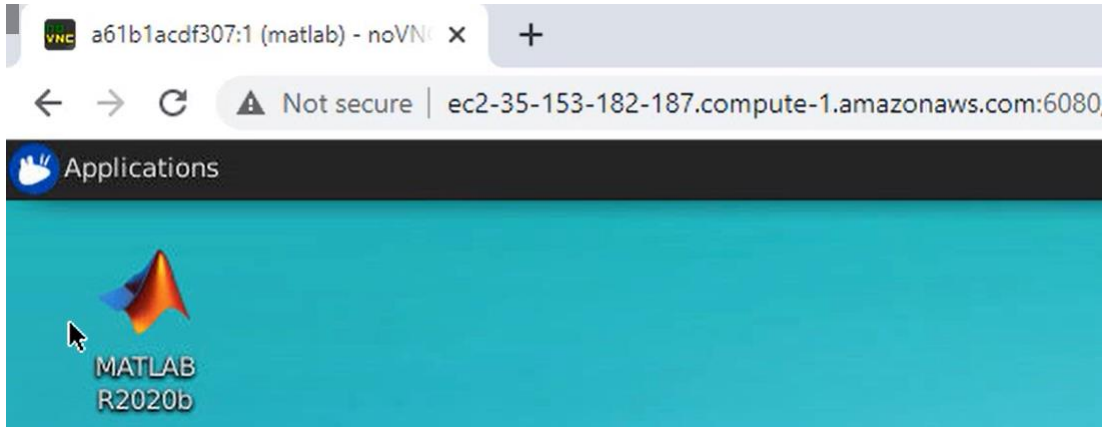
AI Model Design

Easy access to a GPU

1. Select and Run VM
2. Run Docker

```
ubuntu@ip-172-31-33-124:~$ docker pull nvcr.io/partners/matlab:r2020b
r2020b: Pulling from partners/matlab
Digest: sha256:fc07fle83badc807ef5e2341afa2e23cc5c297d54e5f144e81fe4d6075d74486
Status: Image is up to date for nvcr.io/partners/matlab:r2020b
nvcr.io/partners/matlab:r2020b
ubuntu@ip-172-31-33-124:~$ docker run -it --rm -p 5901:5901 -p 6080:6080 --gpus all --s
hm-size=512M nvcr.io/partners/matlab:r2020b
```

3. Remote to VM



38

MATLAB 中的深度学习云设置

Steps to using the Deep Learning Container

AI Model Design

Easy access to a GPU

MathWorks

Training DL network for Semantic Segmentation

6

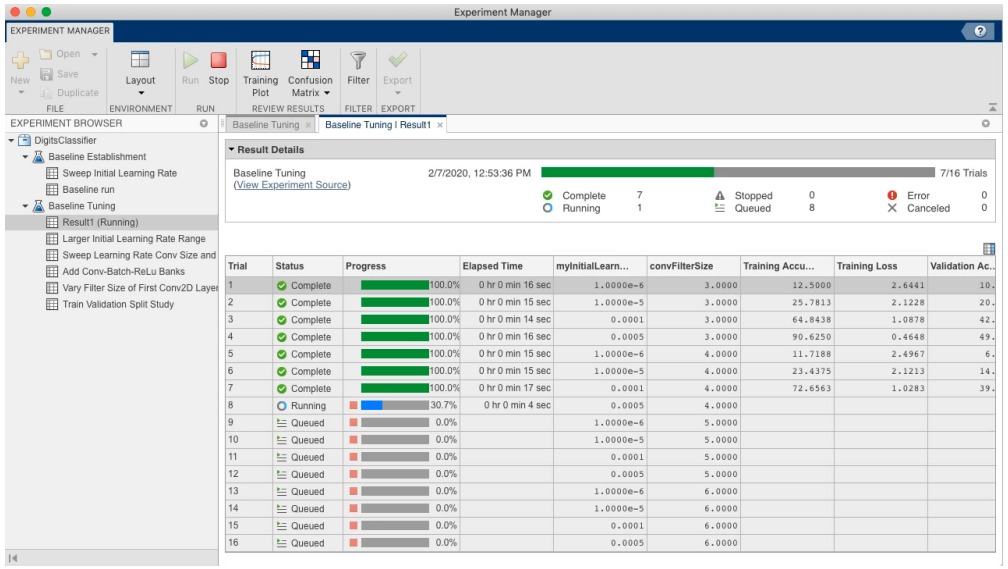
使用实验找到最佳网络

Run experiments to train networks and compare the results.

- Sweep through a range of hyperparameter values
- Compare the results of using different data sets
- Test different deep network architectures

Experiment Manager App

- Reduces the need to code & manually manage experiments



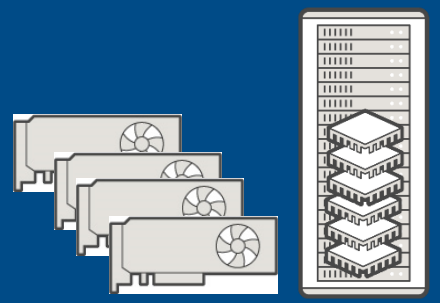
Experiment Manager app to manage multiple deep learning experiments, analyze and compare results and code

40

云端并行多 GPU 训练

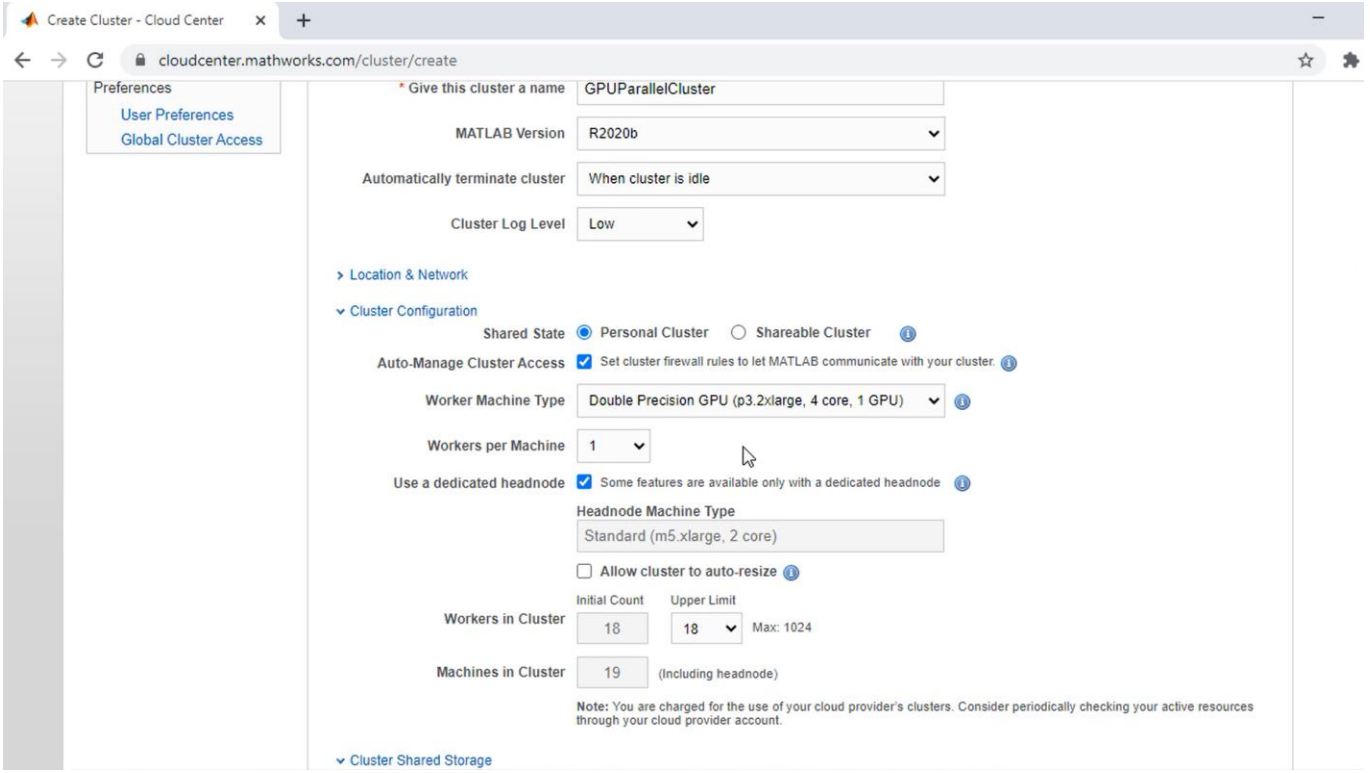
Steps to add a cluster to a MATLAB session

AI Model Tuning



On demand access to a GPU cluster

1. Setup Parallel Server in MathWorks Cloud Center

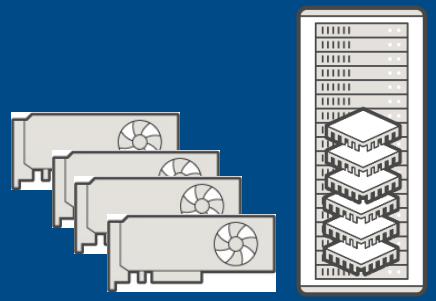


云端并行多 GPU 训练

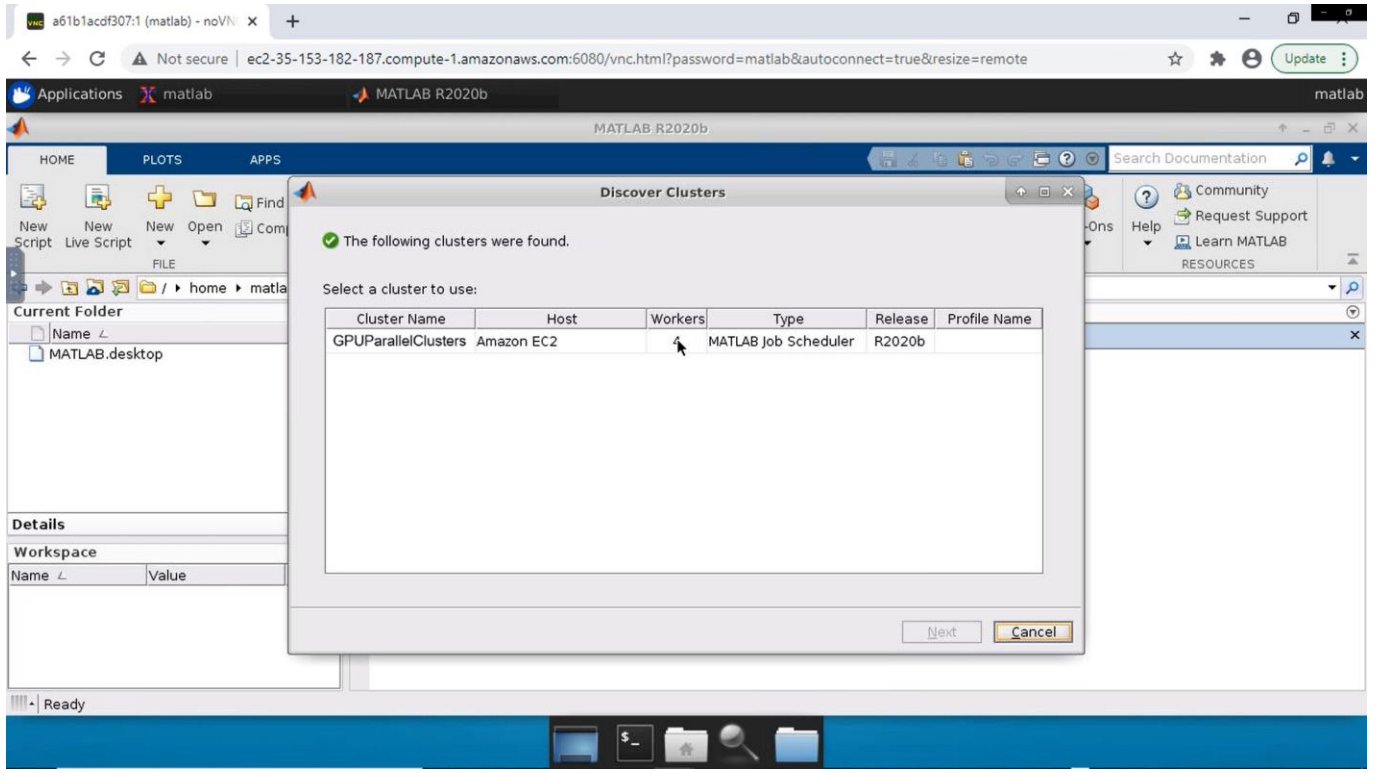
Steps to add a cluster to a MATLAB session

2. Change Default Cluster

AI Model Tuning



On demand access to a GPU cluster

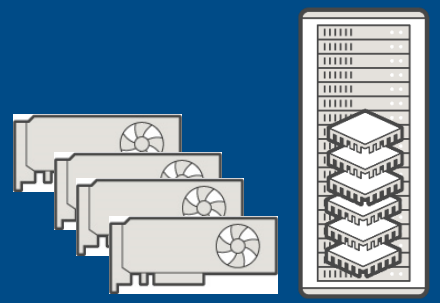


云端并行多 GPU 训练

What options are available for training at scale?

MathWorks

AI Model Tuning



On demand access to a GPU cluster

Run hyperparameter tuning in parallel using Experiment Manager

准备用于部署的训练模型

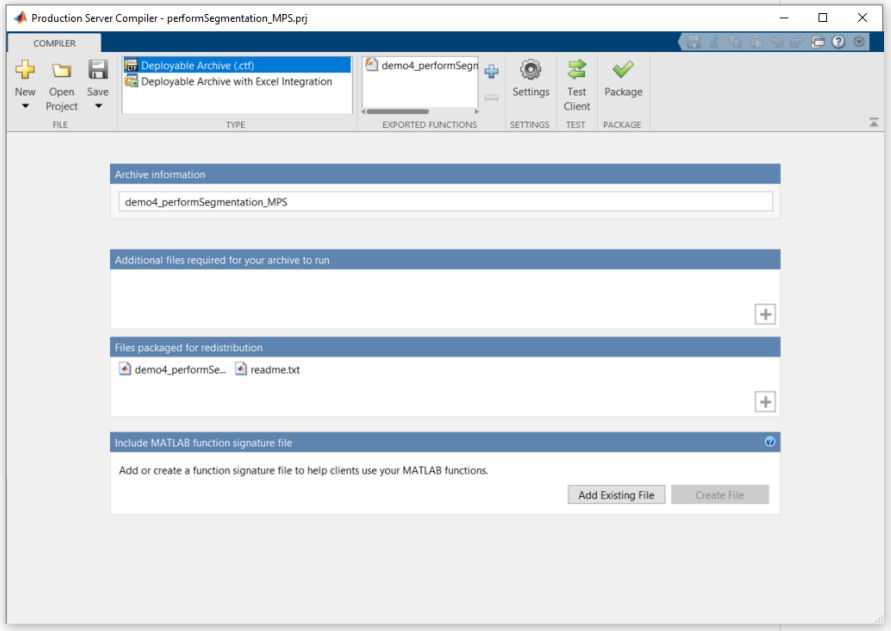
Steps to follow before deploying to the cloud

Deployment

Run models on demand

1. **Create a Function that runs the trained model**

2. **Package (Production Server Compiler App)**



部署在云端的模型 = 模型随处可用

Providing external users access on demand

- **Accessible**
  - Get access to the latest model
- **Available**
  - Each model request calls a “hot” runtime
- **Scalable**
  - Suitable for single calls or batch workflows

将MATLAB模型部署到云端

What options are available for deploying to production?

MATLAB Production Server from Azure Market Place

Deployment

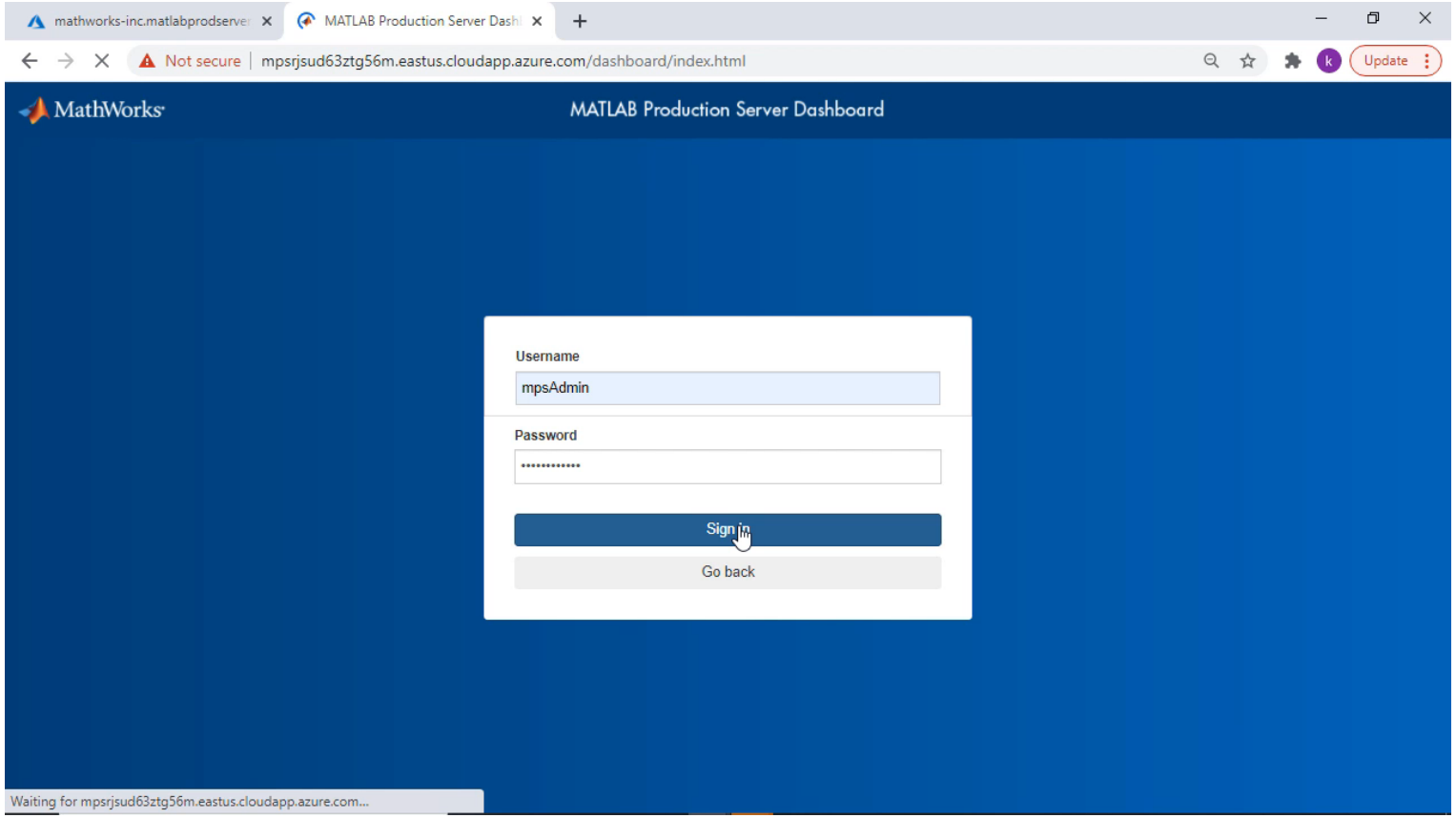
Run models on demand

将MATLAB模型部署到云端

Steps to follow

Deployment

Run models on demand

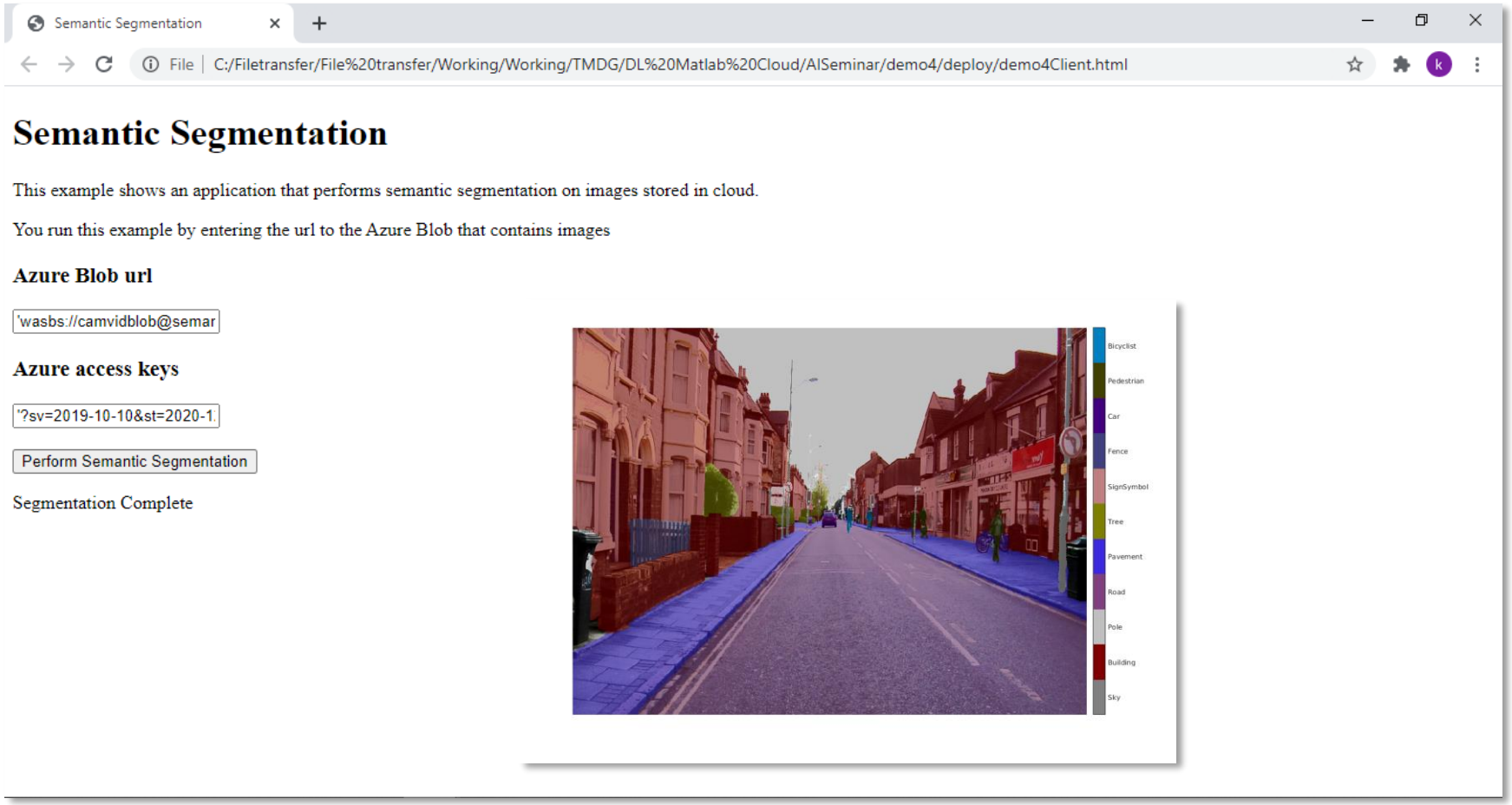


将MATLAB模型部署到云端

What options are available?

Deployment

Run models on demand



总结

Deploy to Embedded and Enterprise systems from one codebase

Tools for handling deployment-specific challenges:

- Fit models to embedded hardware with Quantization / Fixed-Point conversion
- Scale to data and users with MATLAB Production Server
- Incrementally adapt deployed models to maintain performance

Design, Deploy and Maintain AI-powered systems in one framework

总结

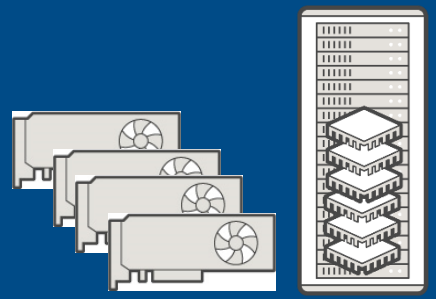
Data Preparation

Access data anywhere

AI Model Design

Build models anywhere

AI Model Tuning



Compute on Demand

Deployment

Run models anywhere

Labeled data	Prototype	Run Experiments	Run Model anywhere
Stored in s3/blob	MATLAB running via Deep Learning Container from NGC (NVIDIA)	MATLAB Parallel Server from Cloud Center	MATLAB Production Server – PAYG

50

MATLAB EXPO 2021

Thank you