

MATLAB EXPO

2021

将AI部署到生产系统及MATLAB云 workflow

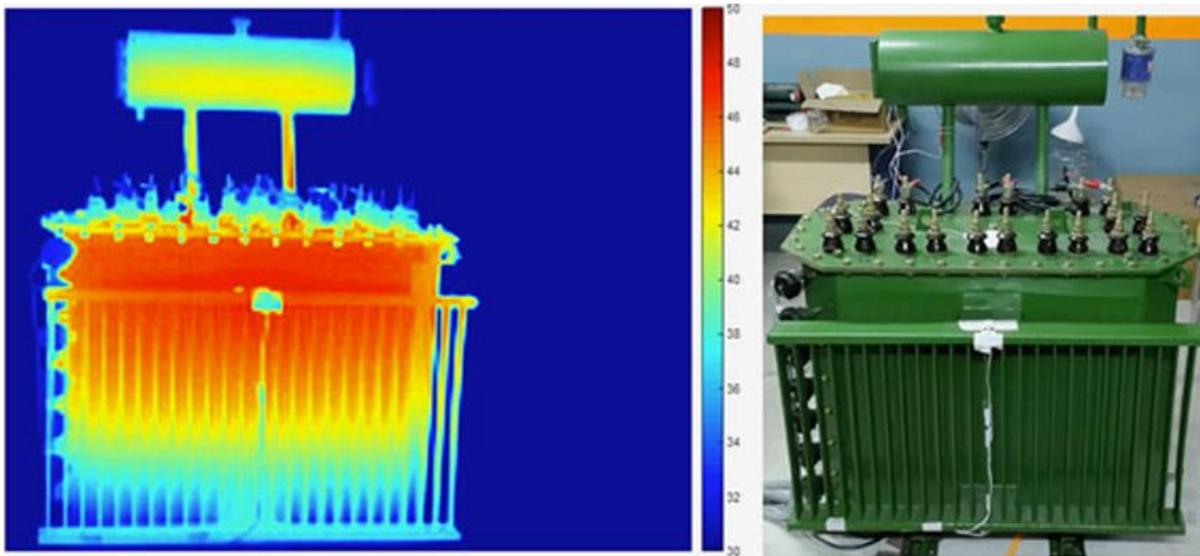
陈宜欣, MathWorks 中国





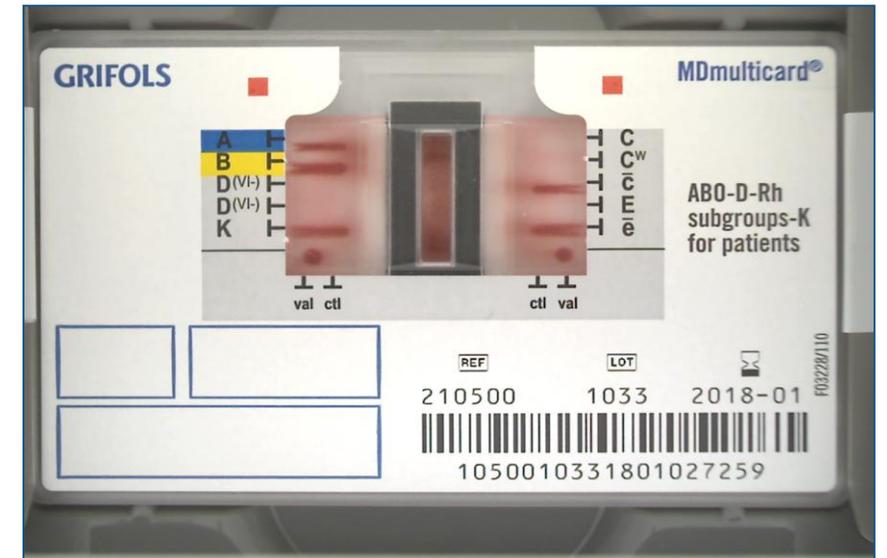
部署到嵌入式和企业系统中

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 workflows

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

Agenda

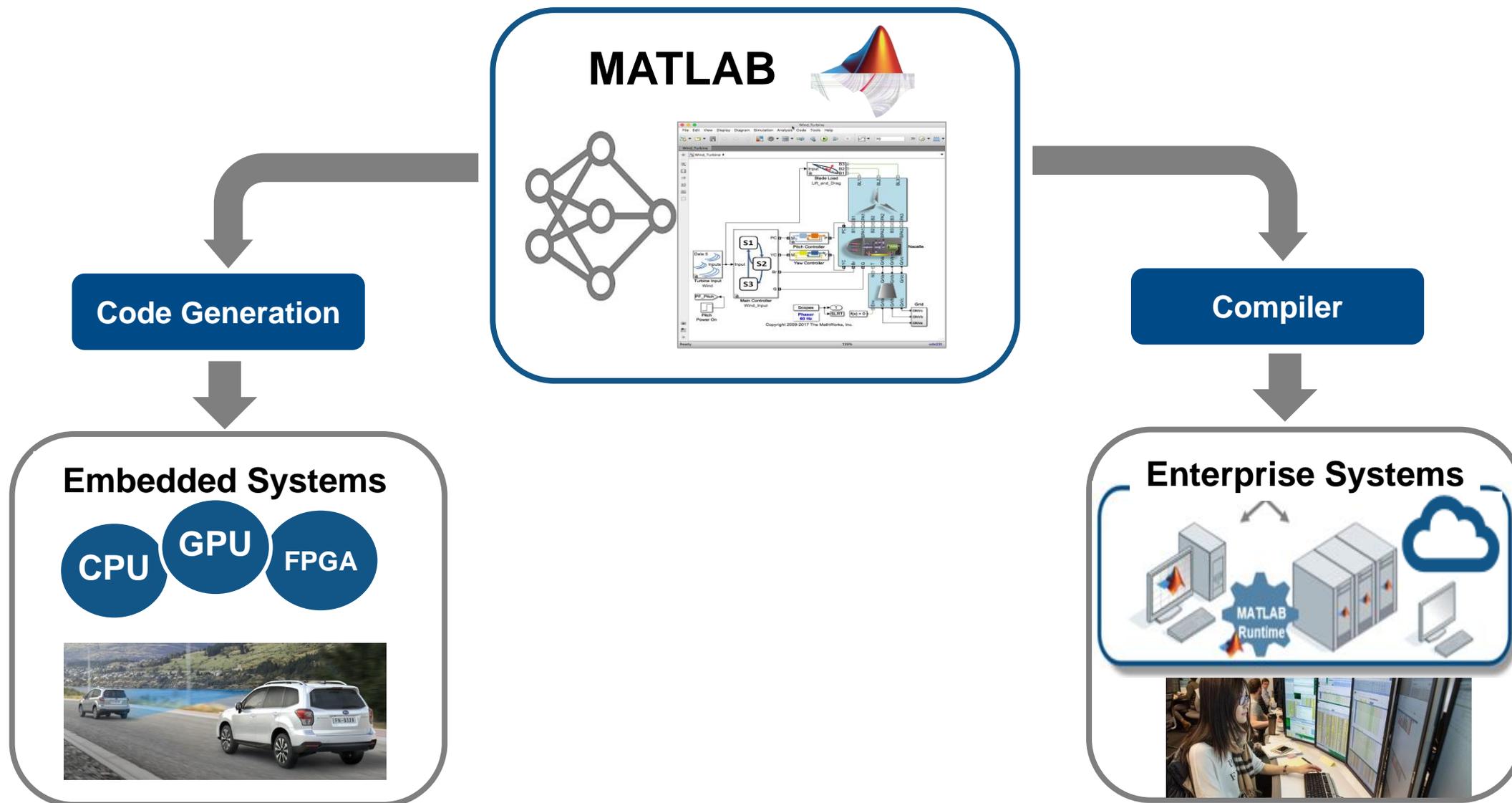
将AI部署到生产系统中

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

AI云 workflows

- 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



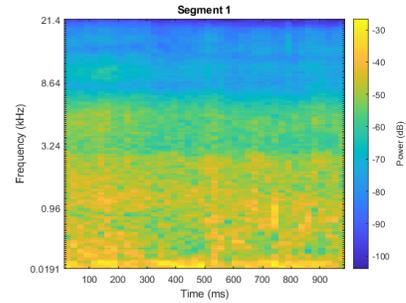
Squeezenet ~5MB
ResNet-50 ~100MB



Limited
resources

声景识别嵌入式部署

Reformat the data



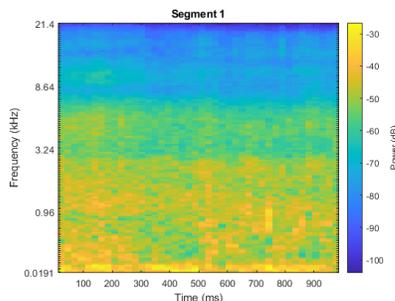
Squeezenet ~5MB
ResNet-50 ~100MB



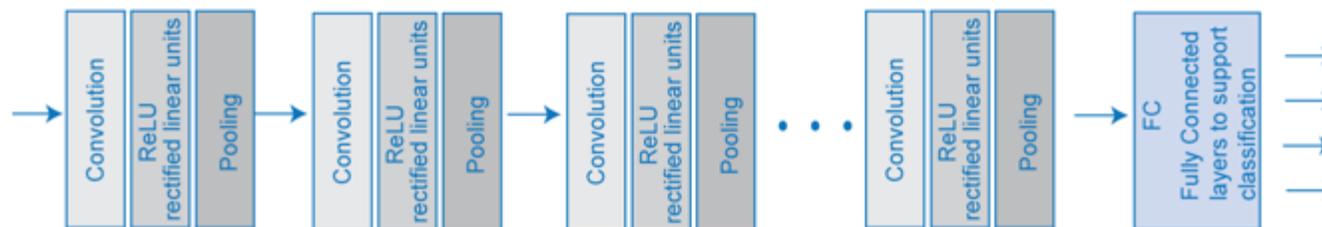
Limited
resources

声景识别嵌入式部署

Reformat the data



Convolutional Neural Networks (CNN)

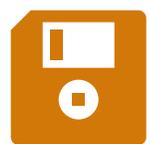


Squeezenet ~5MB
ResNet-50 ~100MB



Limited resources

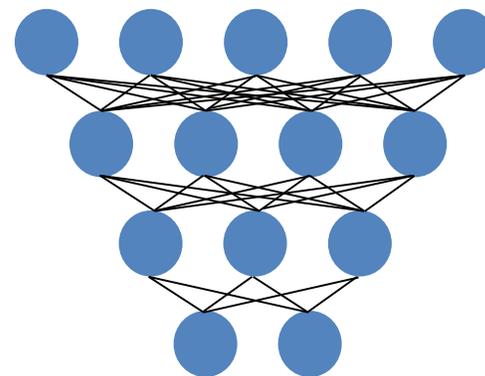
怎样实现嵌入式部署？



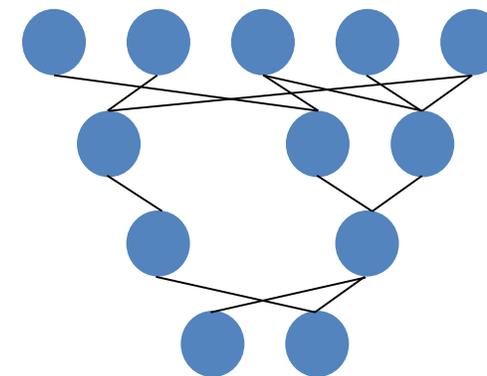
怎样实现嵌入式部署？



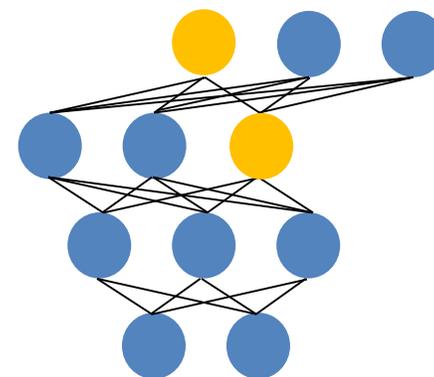
Original



Pruning



Layer Fusion



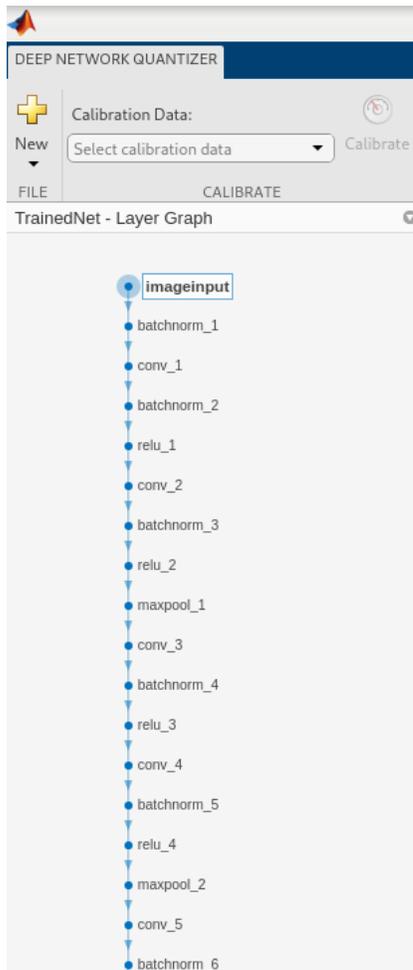
Quantizing



深度学习量化：声景分类



深度学习量化：声景分类

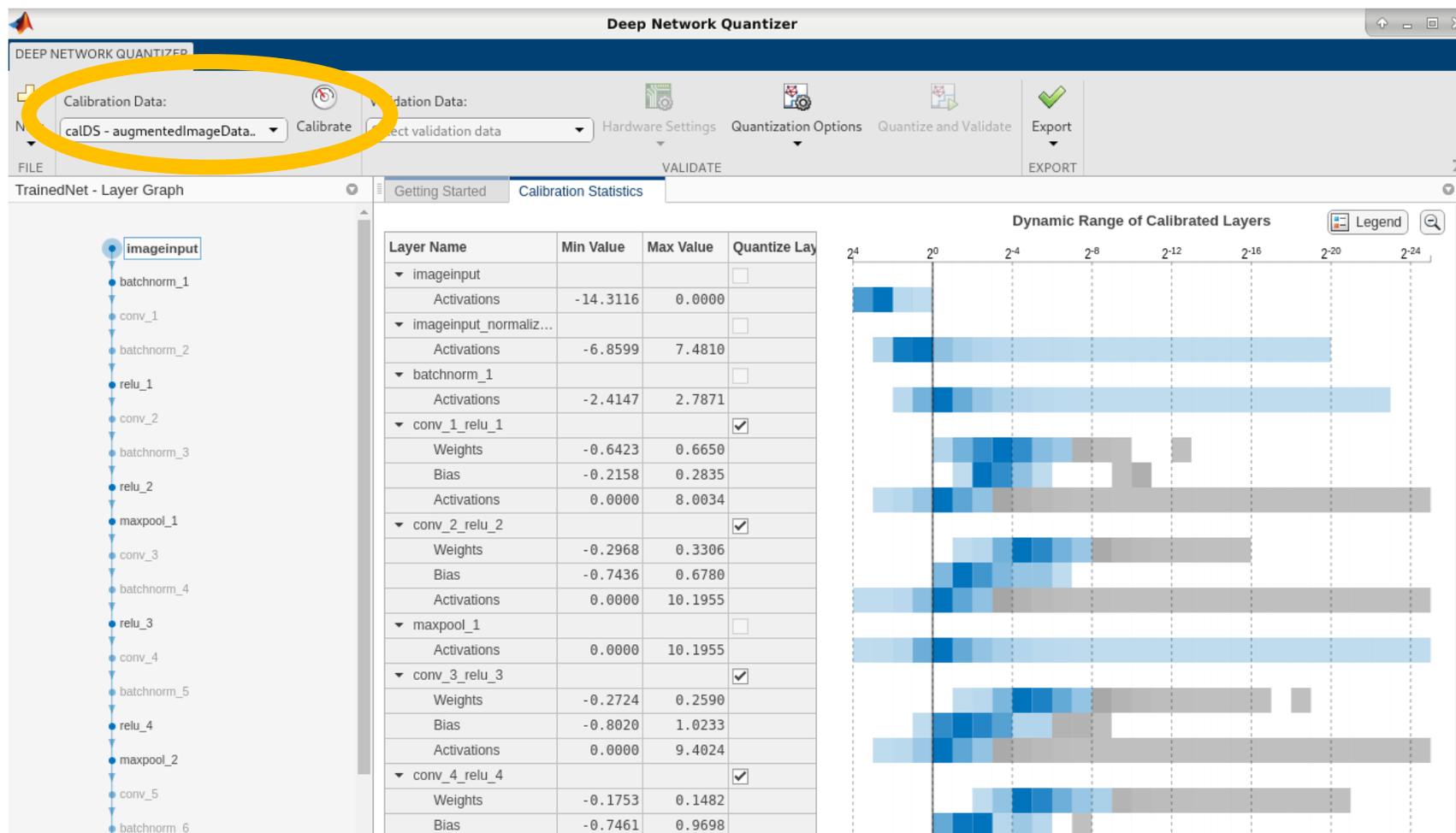


Use Deep Network Quantizer to Optimize the Inference Network

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

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

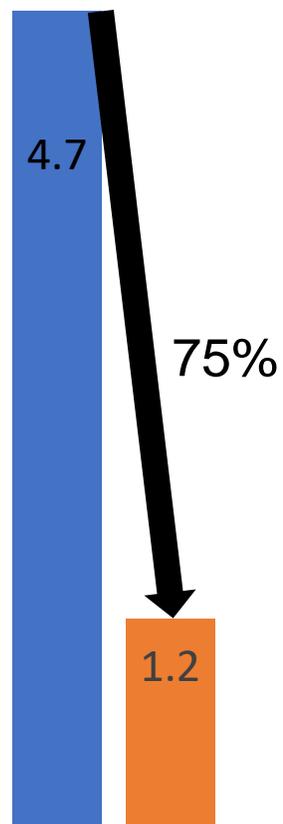
深度学习量化：声景分类



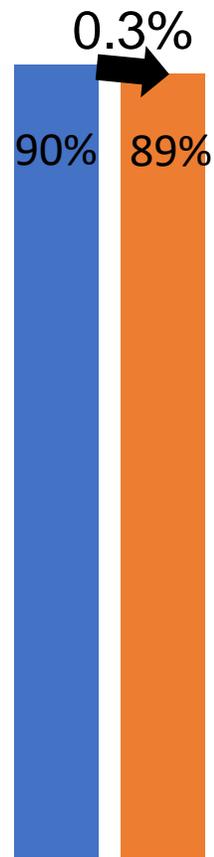
深度学习量化：声景分类

✓ Validation Results

Memory (MB)



Top-2 Accuracy

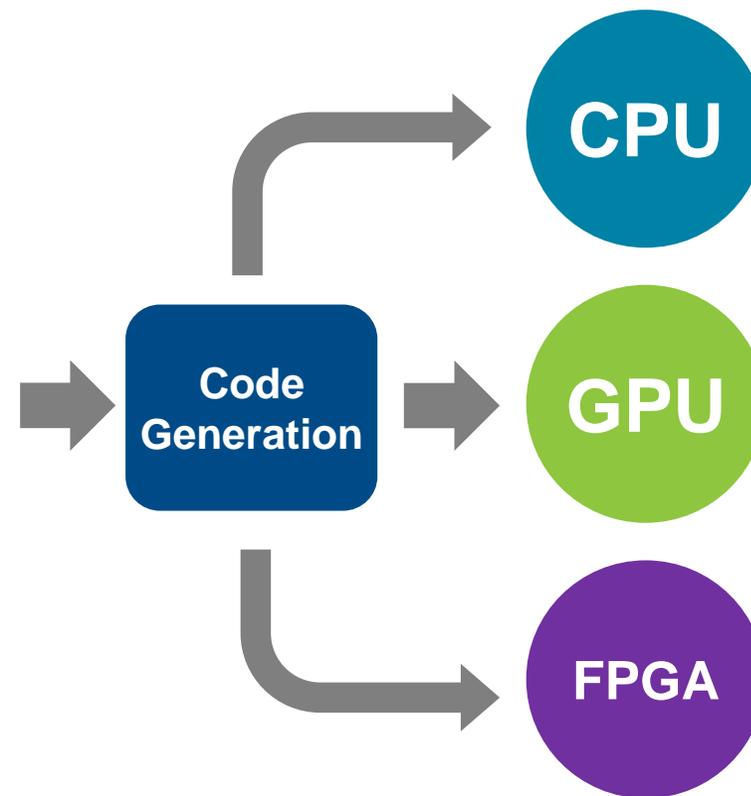
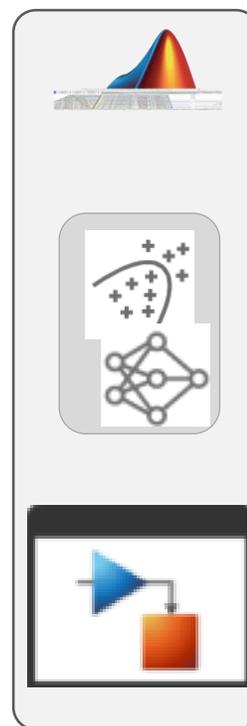


Learnable Parameters

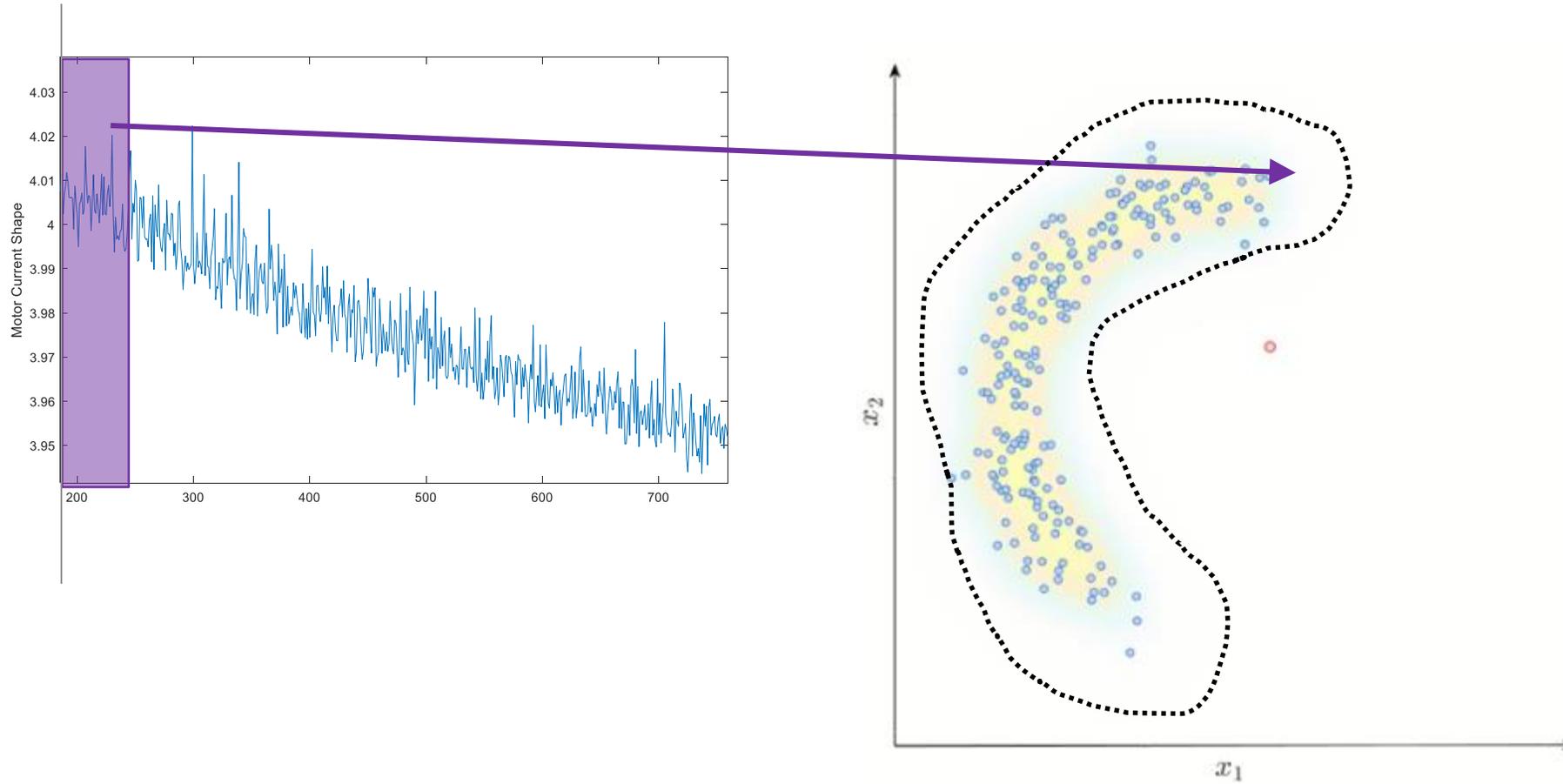
■ FP32 ■ INT-8

Top-2 Accuracy

■ FP32 ■ INT-8

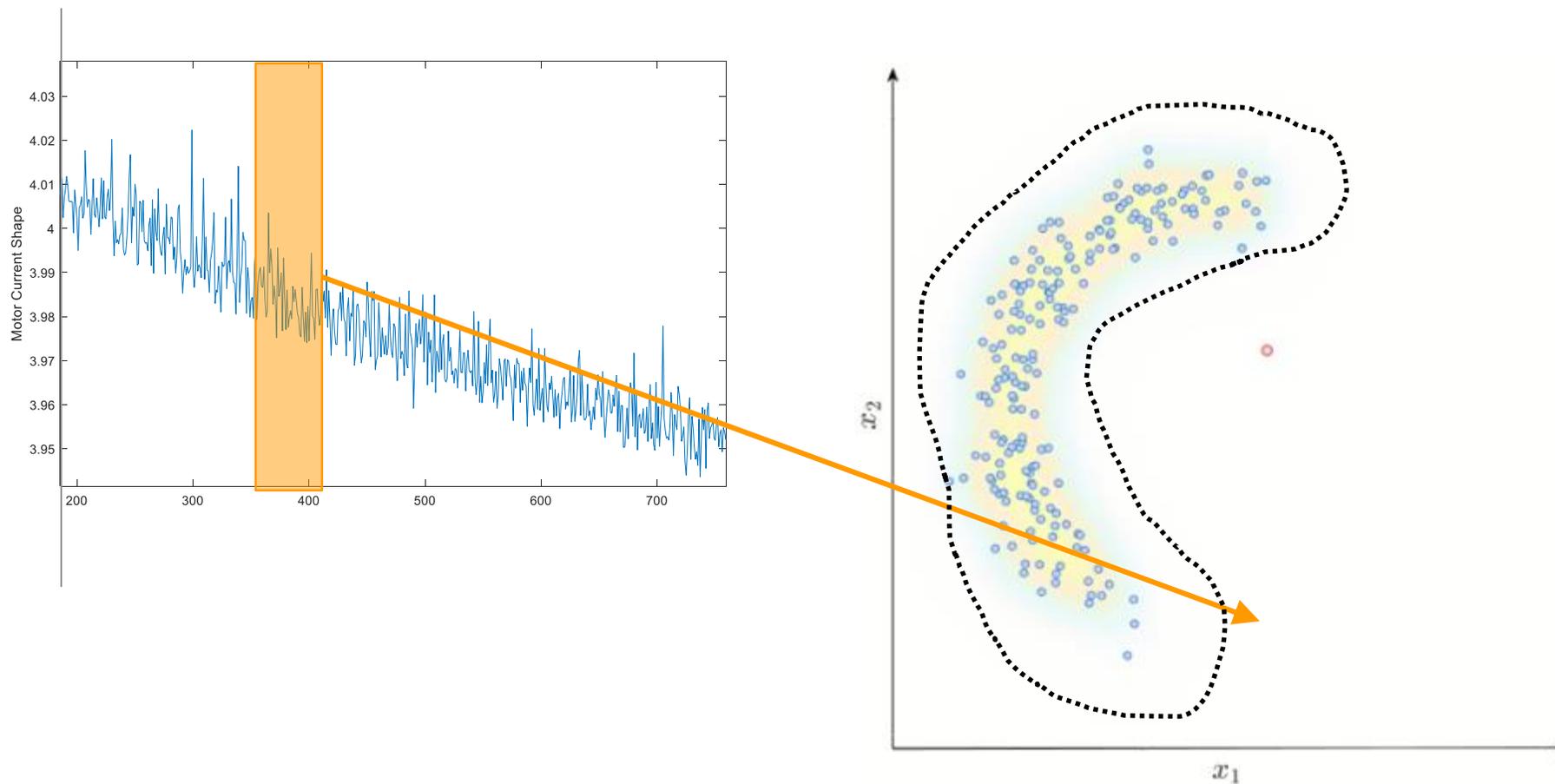


AI模型反映系统行为和环境

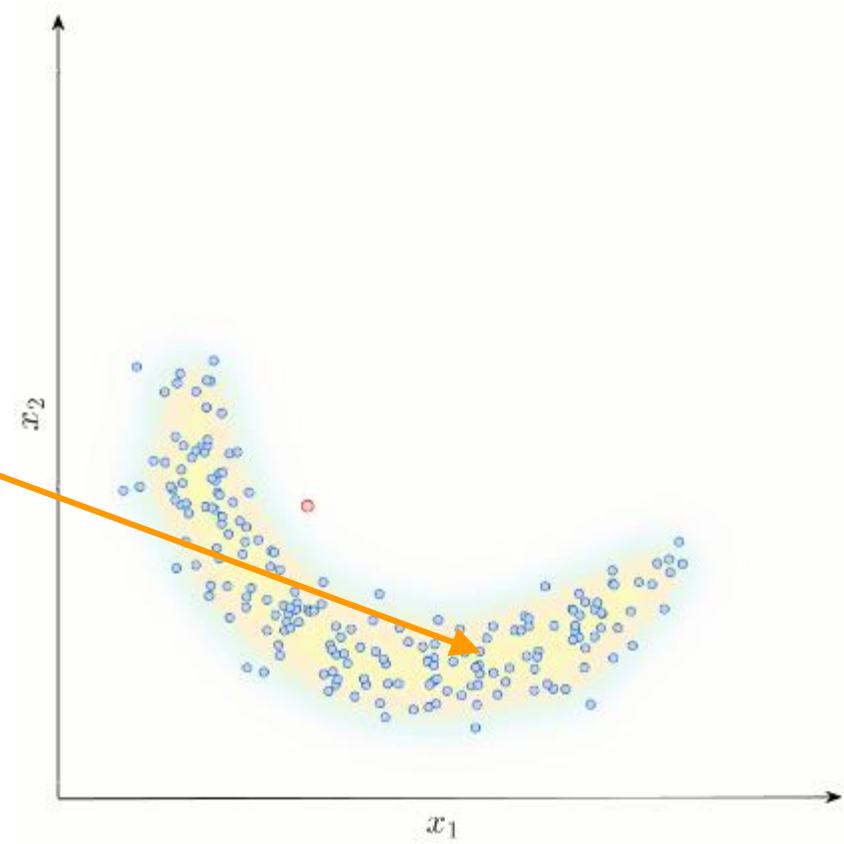
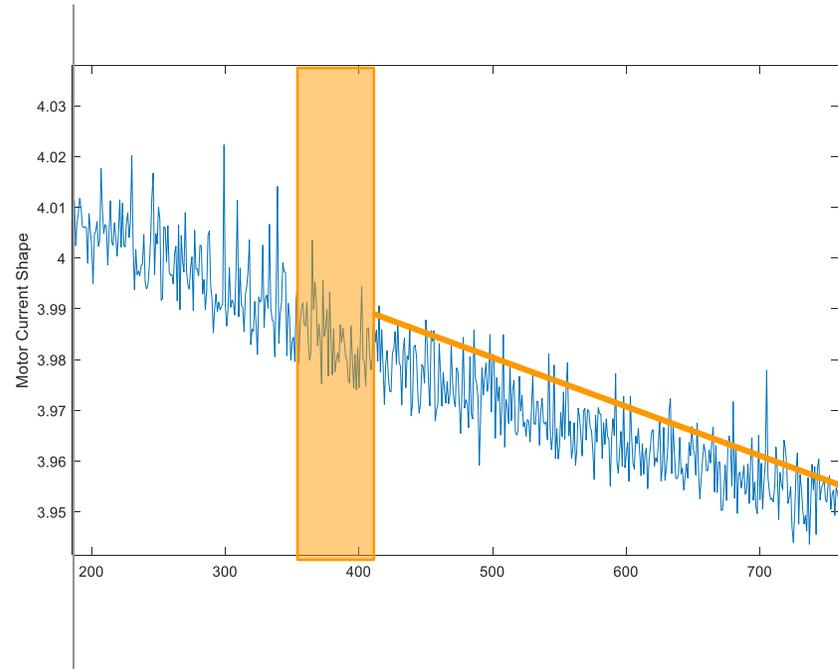


(illustration only; not based on actual data)

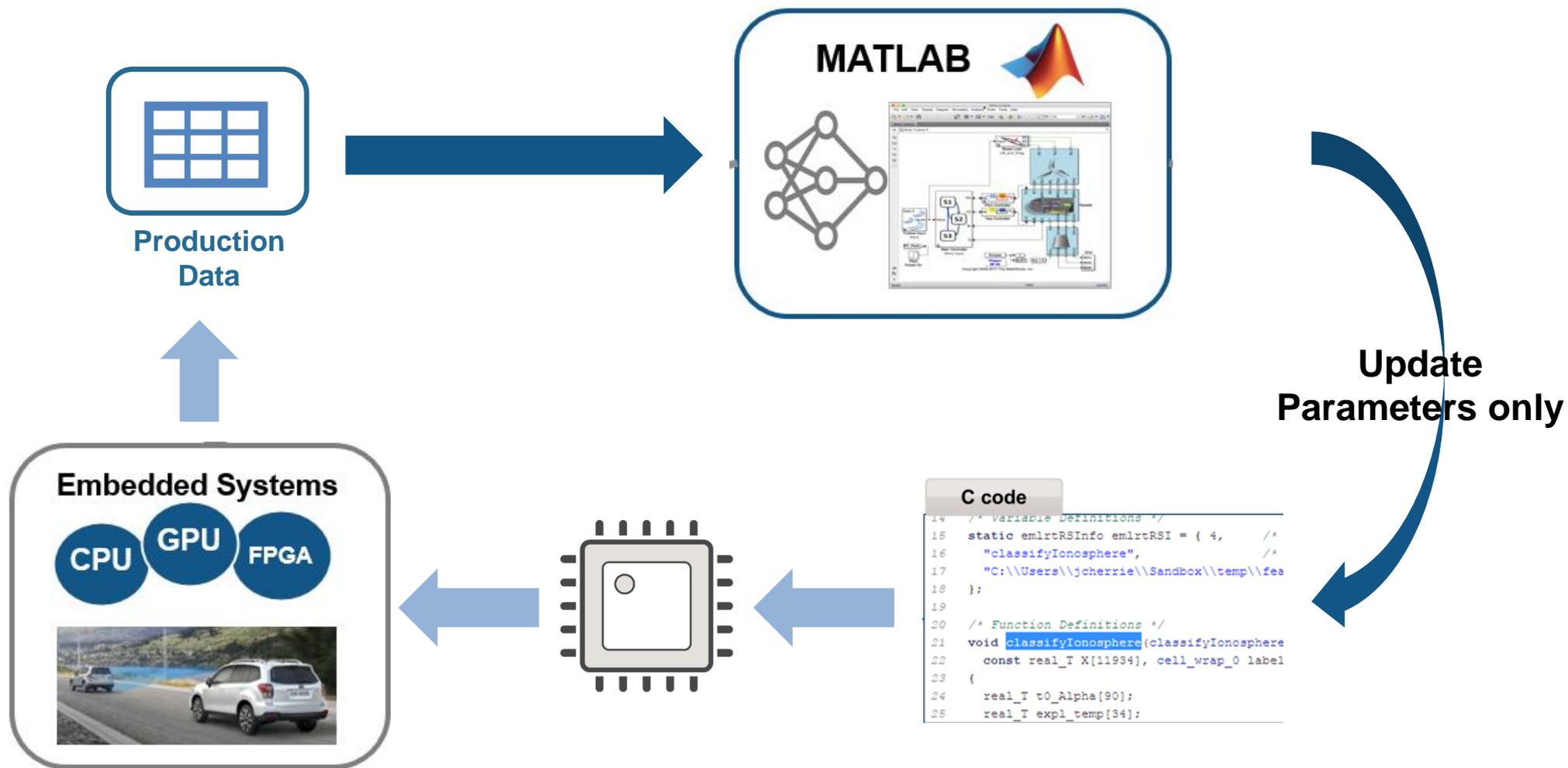
AI模型反映系统行为和环境



调整已部署的模型



嵌入式部署模型更新



Agenda

将AI部署到生产系统中

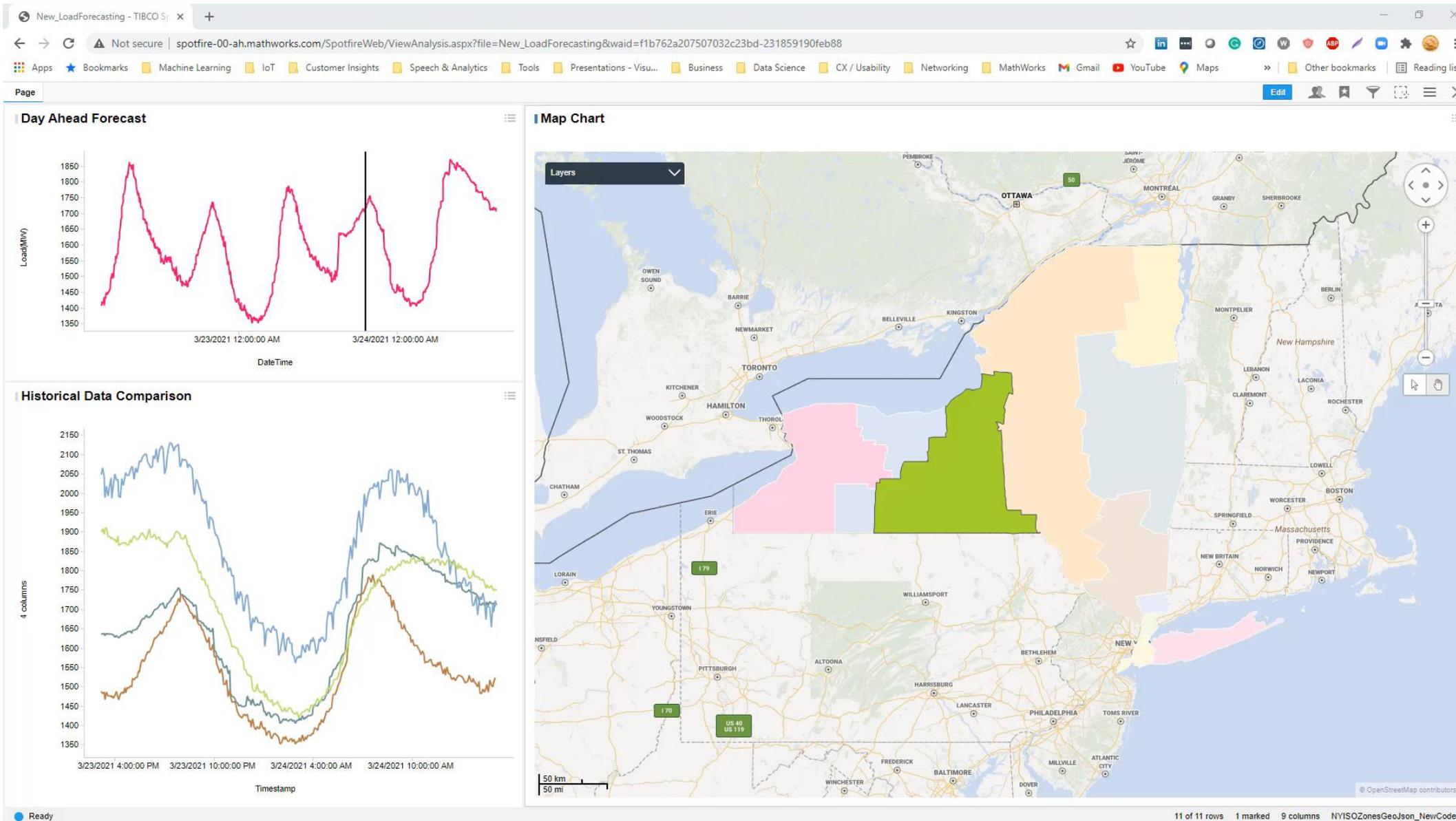
- Three specific challenges:
 - Limitations of Embedded hardware
 - Ongoing changes in environment or system behavior
 - Scale to production load in Enterprise systems



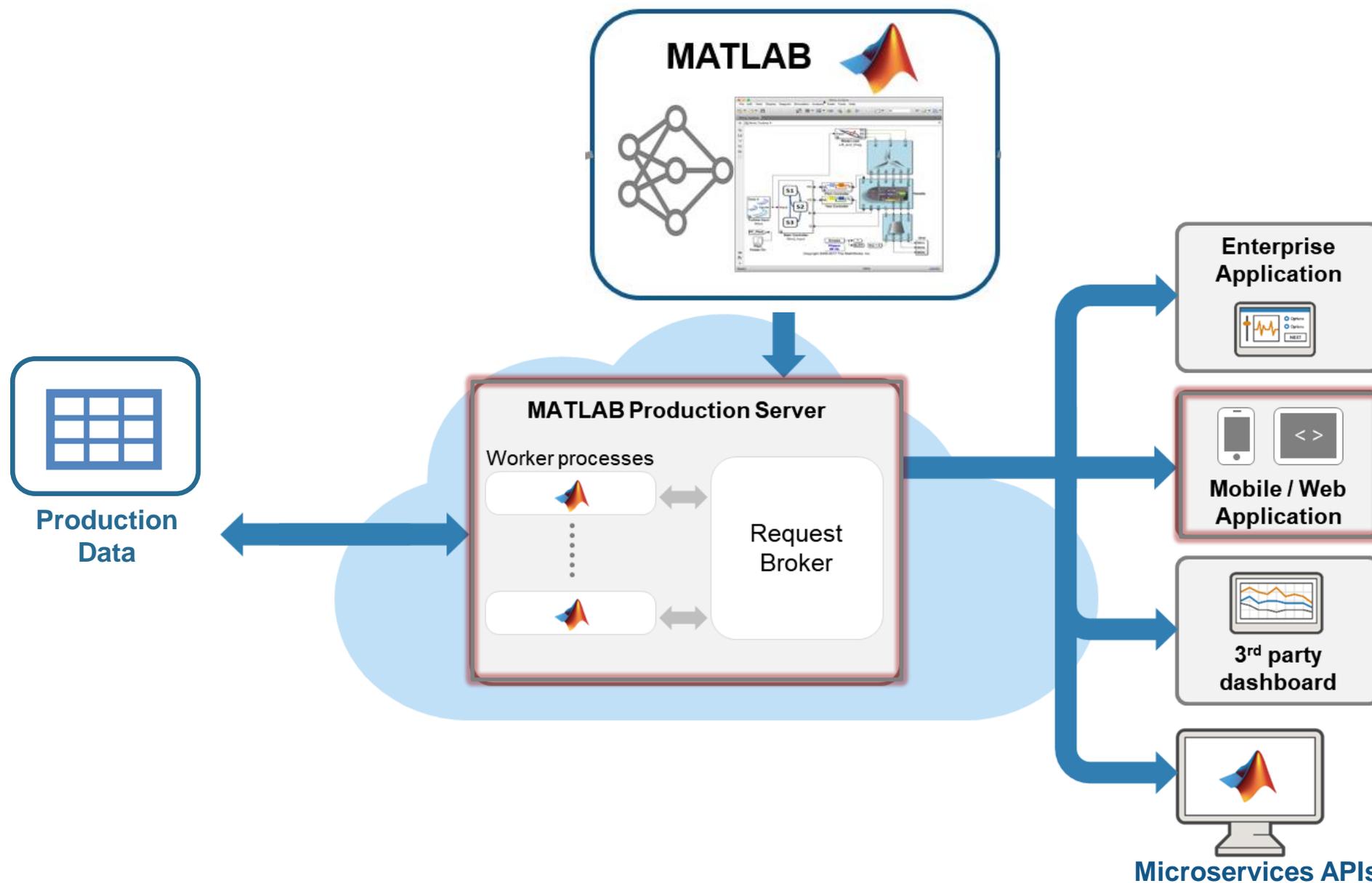
AI云 workflow

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

AI企业部署

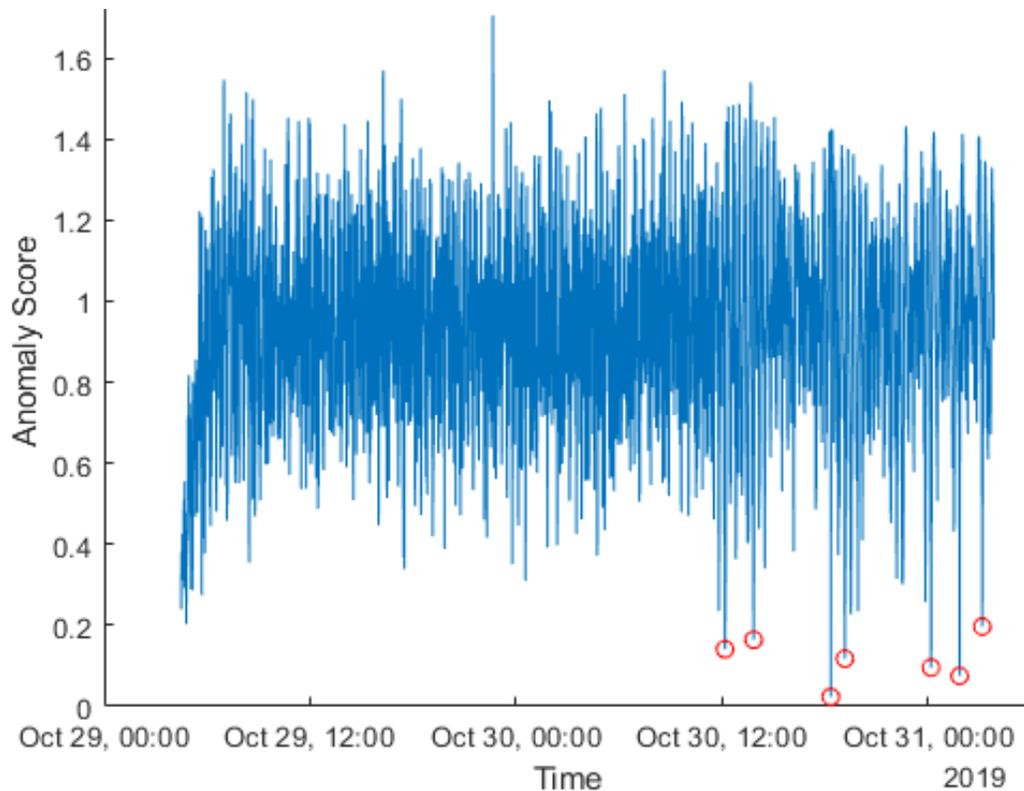


与企业系统集成并扩展到生产负载



Example: 增量健康监测

Sensor data



Anomaly Detection loop

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

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

    streamdata = sensordata(datafilter,:);

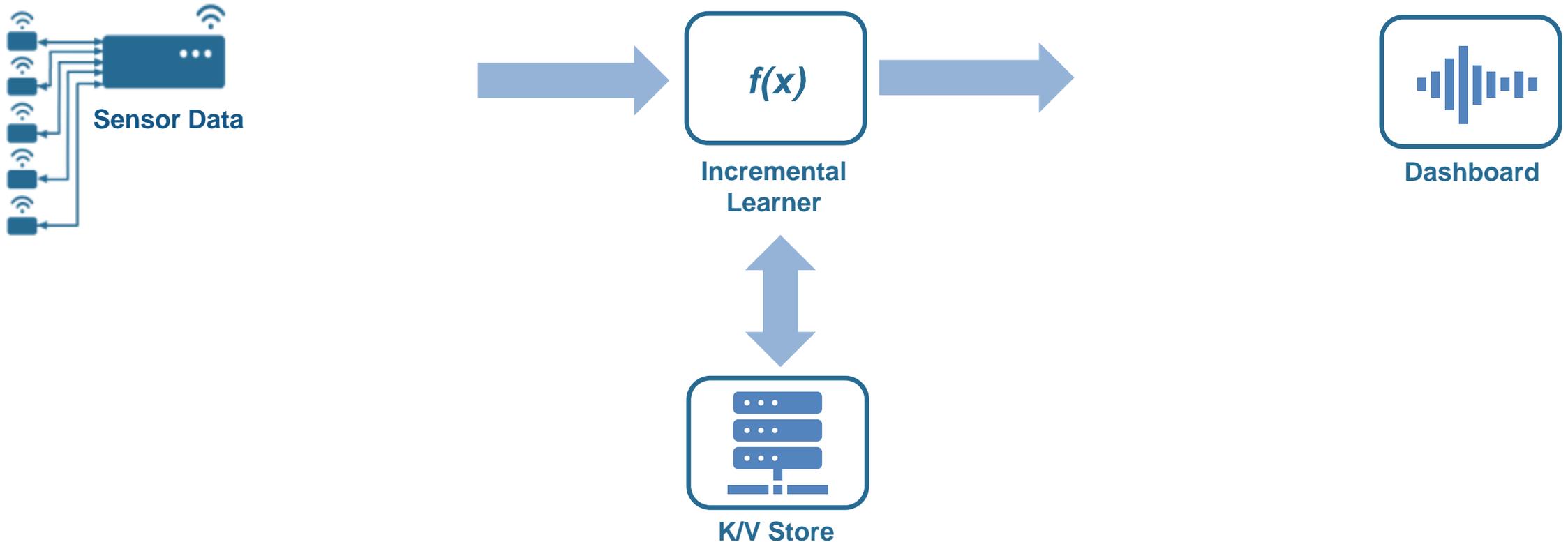
    % Detect Anomalies with incremental One-class SVM
    [nextState, results] = detectAnomalyLocal(streamdata, state);

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

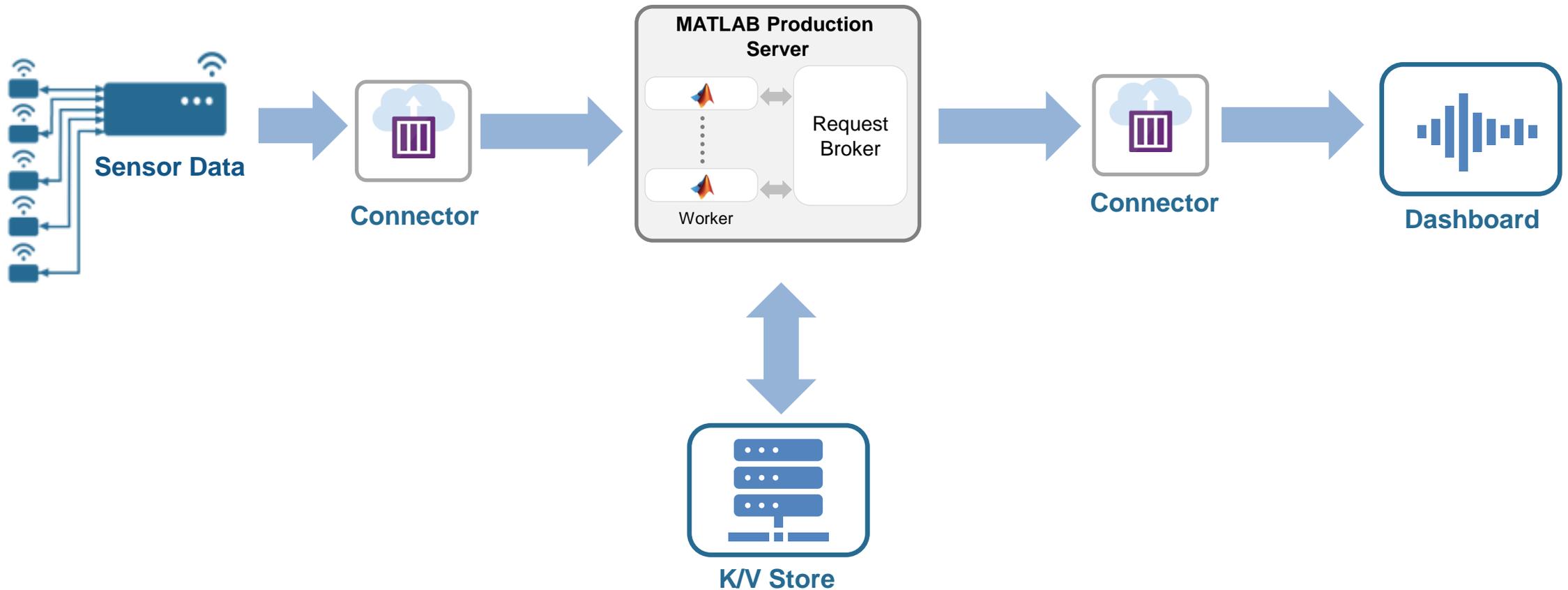
    seqn = seqn + batchsize; % |step through batch test data
end
```

流媒体架构中的增量学习

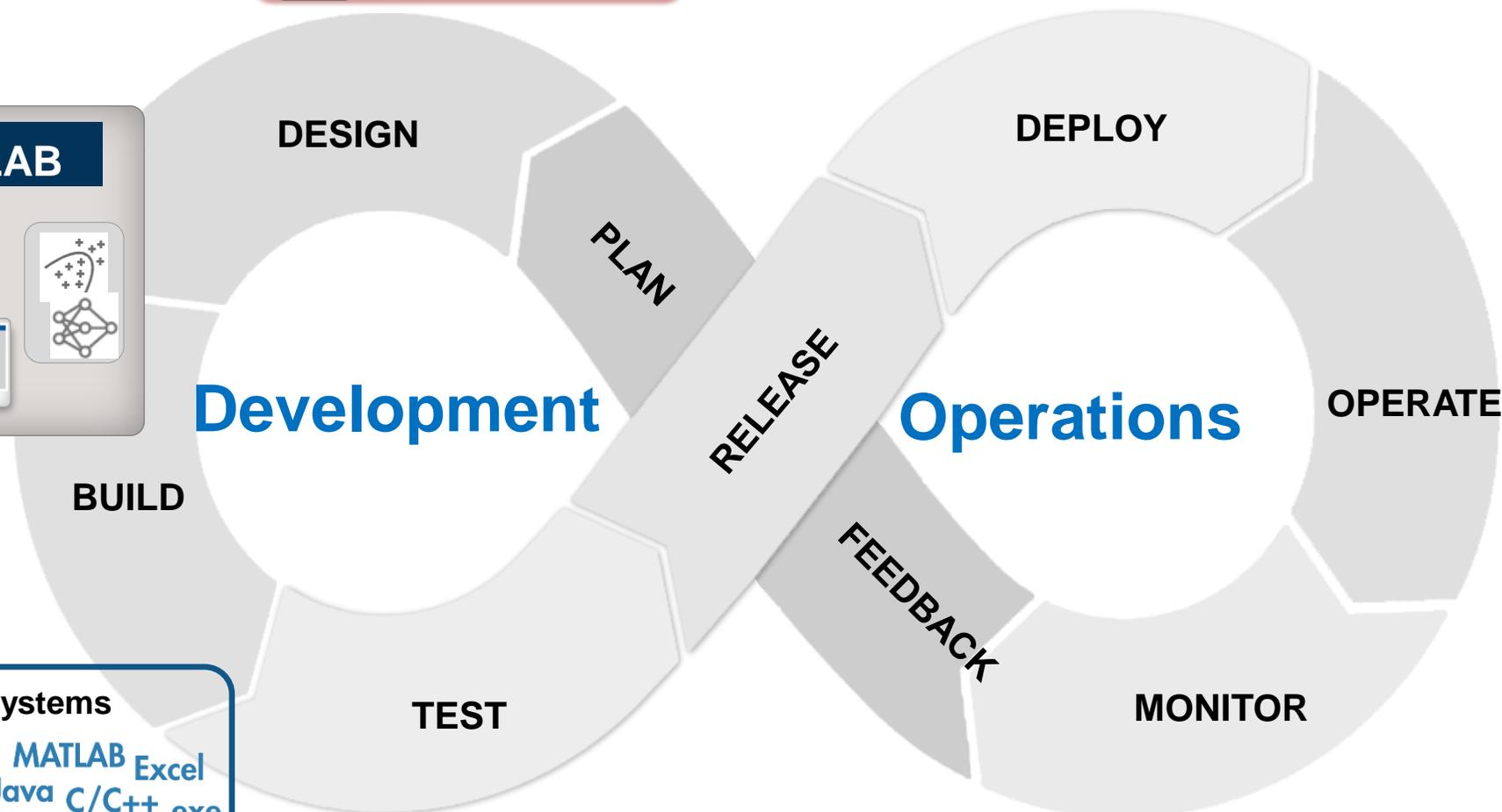
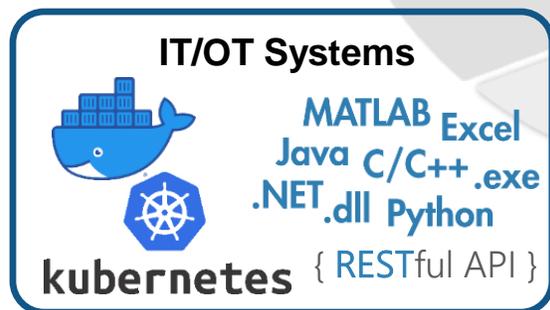
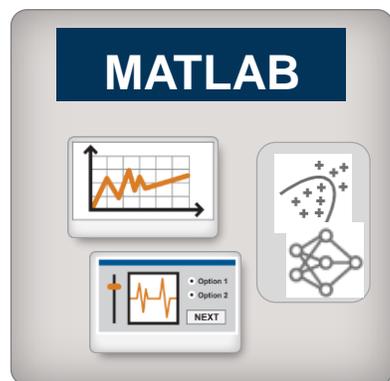
```
incMdl = incrementalLearner(mdl);  
  
while dataStreaming  
    featureChunk = extractFeatures(streamdata);  
    inclMdl = updateMetricsAndFit(incMdl, featureChunk, labels);  
End
```

R2020b

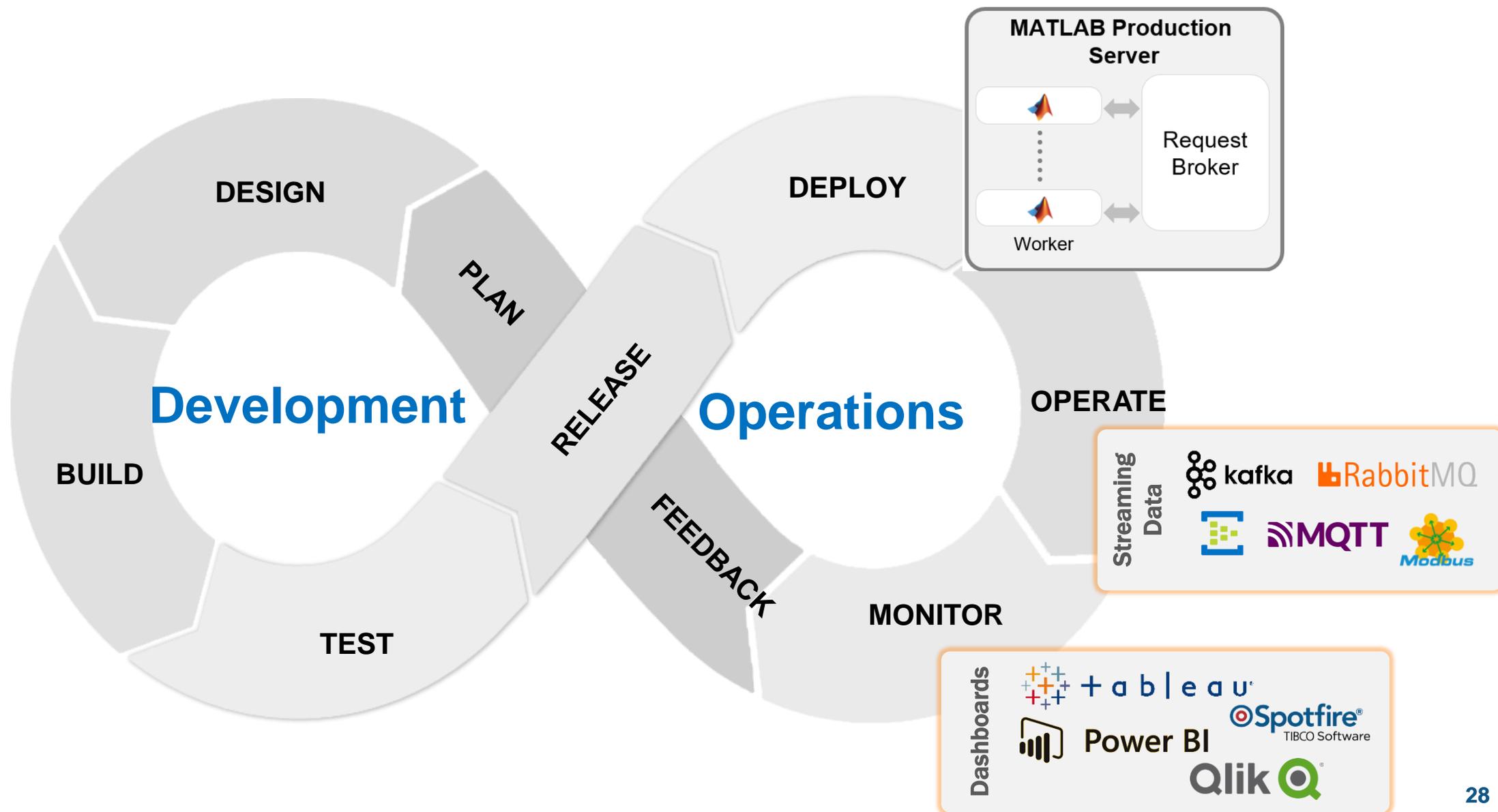
流媒体架构中的增量学习



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

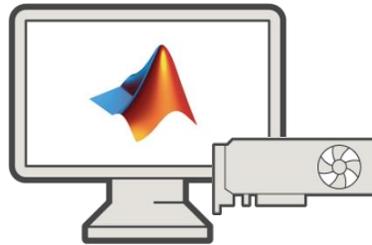
AI云 workflows

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

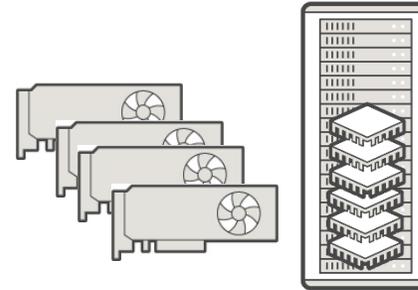
为什么要在云端实现AI?



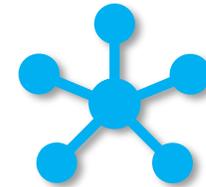
**Access Data
anywhere**



**Build models
anywhere**



**Compute on
Demand**



**Run Models
anywhere**

AI系统设计 workflow

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

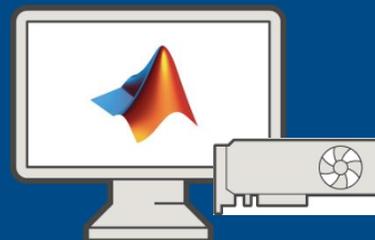
云端 AI 系统设计是什么样的？

Data Preparation



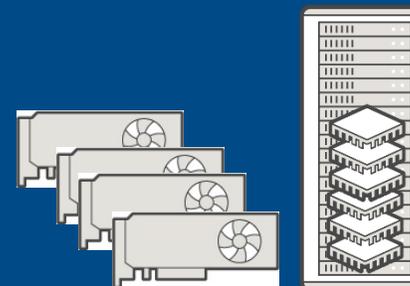
Access data
anywhere

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**

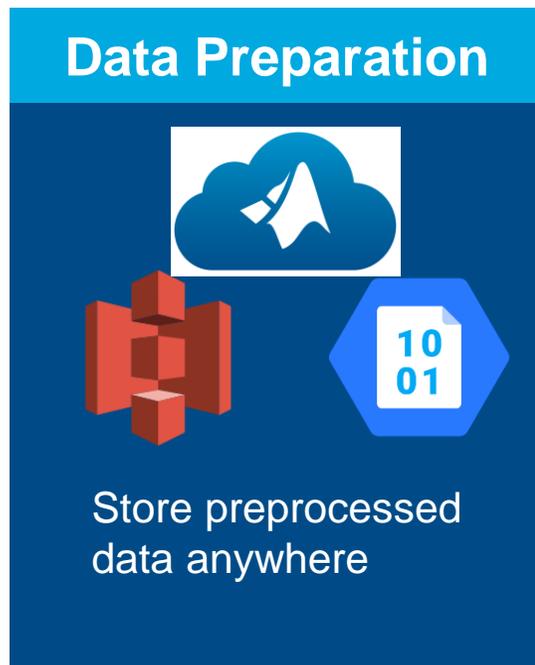
- All you need is the URL

- **Scalable**

- Deep Learning Data Sets can get BIG.
- Need more storage? No problem.

- **Secure**

- You need “Keys” to lock and unlock the data



Data Preparation

Store preprocessed data anywhere

The graphic features a blue header with the text 'Data Preparation'. Below the header, there is a white square containing a blue cloud icon with a white arrow pointing upwards. To the left of the cloud is a red 3D cube structure. To the right is a blue hexagon with a white square inside containing the binary code '10 01'. Below these icons, the text 'Store preprocessed data anywhere' is written in white.

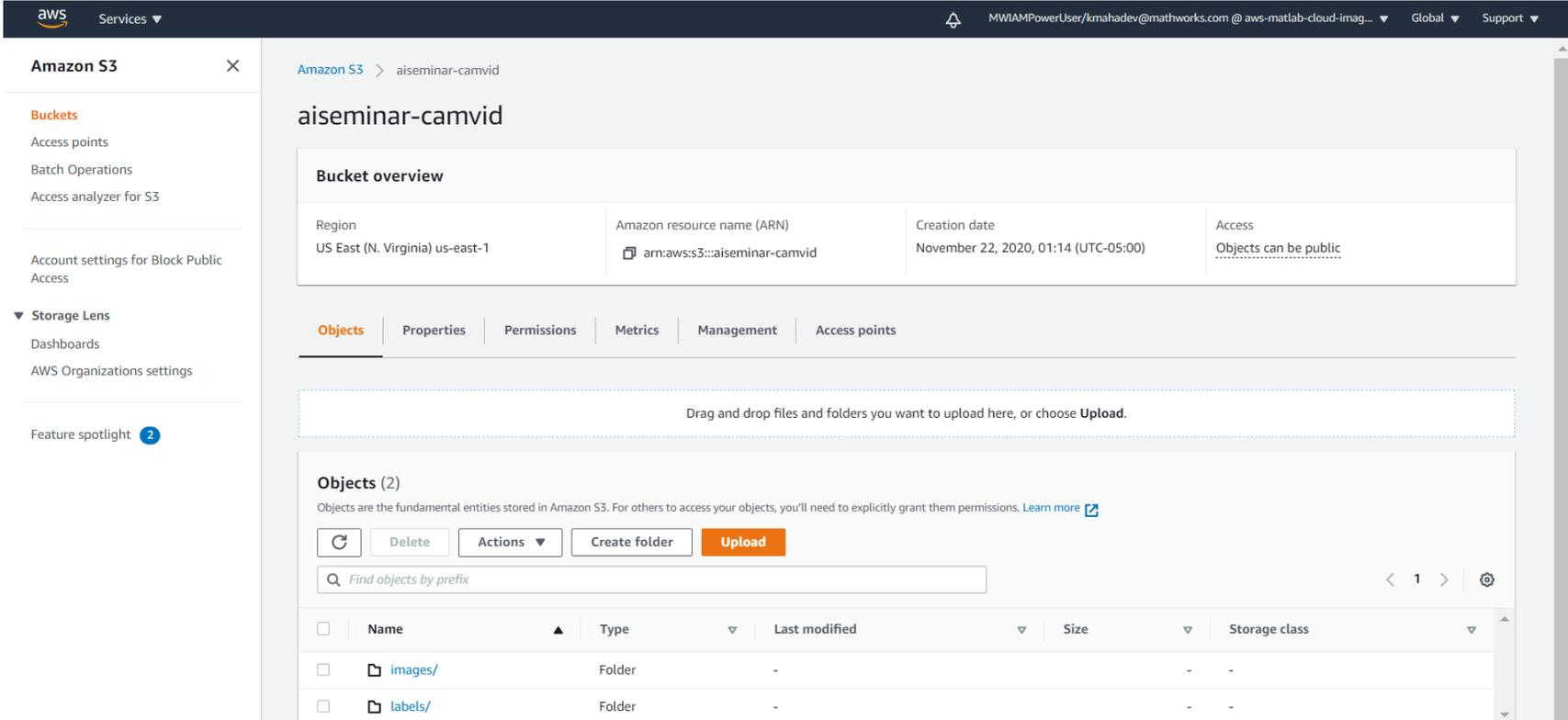
云中的数据 - 使其靠近计算

Uploading Data to the S3

Data Preparation



Store preprocessed data anywhere



The screenshot displays the Amazon S3 console interface. The left sidebar shows the 'Amazon S3' navigation menu with options like Buckets, Access points, Batch Operations, and Account settings. The main content area shows the 'aiseminar-camvid' bucket overview, including details like Region (US East (N. Virginia) us-east-1), Amazon resource name (arn:aws:s3:::aiseminar-camvid), Creation date (November 22, 2020, 01:14 (UTC-05:00)), and Access (Objects can be public). Below the overview, there are tabs for Objects, Properties, Permissions, Metrics, Management, and Access points. The 'Objects' tab is active, showing a list of objects with columns for Name, Type, Last modified, Size, and Storage class. The list contains two folders: 'images/' and 'labels/'.

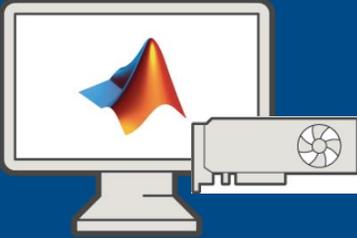
Name	Type	Last modified	Size	Storage class
images/	Folder	-	-	-
labels/	Folder	-	-	-

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

aws
Contact Sales | Support | English | My Account | [Sign In to the Console](#)

re:Invent | Products | Solutions | Pricing | Documentation | Learn | Partner Network | AWS Marketplace | Customer Enablement | Explore More |

Amazon EC2 | Overview | Features | Pricing | Instance Types | FAQs | Getting Started | Resources

General Purpose

Compute Optimized

Memory Optimized

Accelerated Computing

Storage Optimized

Instance Features

Measuring Instance Performance

Accelerated Computing

Accelerated computing instances use hardware accelerators, or co-processors, to perform functions, such as floating point number calculations, graphics processing, or data pattern matching, more efficiently than is possible in software running on CPUs.

P4 | **P3** | P2 | Inf1 | G4 | G3 | F1

Amazon EC2 P3 instances deliver high performance compute in the cloud with up to 8 NVIDIA® V100 Tensor Core GPUs and up to 100 Gbps of networking throughput for machine learning and HPC applications.

Features:

- Up to 8 NVIDIA Tesla V100 GPUs, each pairing 5,120 CUDA Cores and 640 Tensor Cores
- High frequency Intel Xeon E5-2686 v4 (Broadwell) processors for p3.2xlarge, p3.8xlarge, and p3.16xlarge.
- High frequency 2.5 GHz (base) Intel Xeon 8175M processors for p3dn.24xlarge.
- Supports NVLink for peer-to-peer GPU communication
- Provides up to 100 Gbps of aggregate network bandwidth.
- EFA support on p3dn.24xlarge instances

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
p3.8xlarge	4	32	244	64	NVLink	EBS-Only	7 Gbps	10 Gigabit
p3.16xlarge	8	64	488	128	NVLink	EBS-Only	14 Gbps	25 Gigabit
p3dn.24xlarge	8	96	768	256	NVLink	2 x 900 NVMe SSD	19 Gbps	100 Gigabit

36

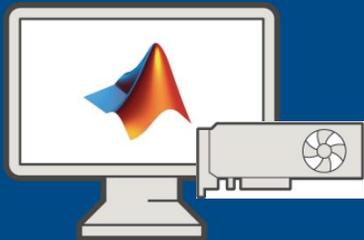
轻松访问GPU资源

Options chosen for setting up MATLAB

- 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

AI Model Design



Easy access to a GPU

- [MATLAB Deep Learning Container](#) on NVIDIA NGC store

The screenshot shows the NVIDIA NGC CATALOG interface. The top navigation bar includes 'COLLECTIONS', 'CONTAINERS', 'HELM CHARTS', and 'MODELS'. A search bar is present with the text 'Search containers'. Below the navigation, there are four container cards displayed:

- MATLAB Container:** MATLAB is a programming platform designed for engineers and scientists. The MATLAB Deep Learning Container provides algorithms, pretrained models, and apps...
- CHROMA Container:** CHROMA is a Physics application designed for solving the theory of quarks and gluons.
- Microvolution Container:** Microvolution is a high-performance 3D deconvolution application, designed to deconvolve images from widefield, confocal, two photon, light sheet, and HC...
- MILC Container:** MILC represents part of a set of codes written by the MIMD Lattice Computation (MILC) collaboration used to study quantum chromodynamics (QCD), the th...

MATLAB 中的深度学习云设置

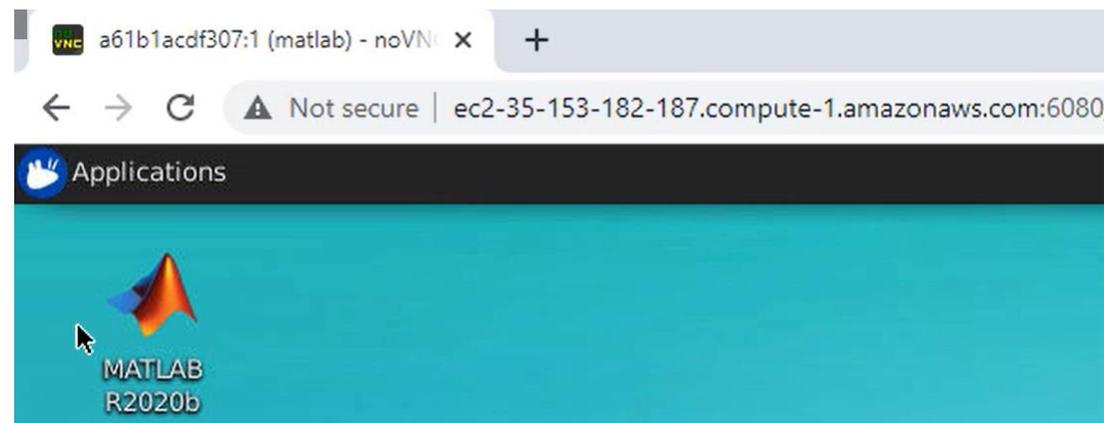
Steps to using the Deep Learning Container

1. Select and Run VM

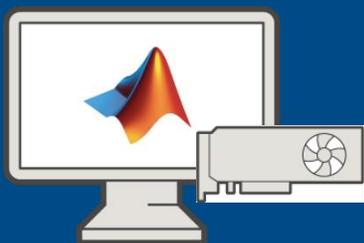
2. Run Docker

3. Remote to VM

```
ubuntu@ip-172-31-33-124:~$ docker pull nvcr.io/partners/matlab:r2020b
r2020b: Pulling from partners/matlab
Digest: sha256:fc07f1e83badc807ef5e2341afa2e23cc5c297d54e5f144e81fe4d6075d74486
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 --shm-size=512M nvcr.io/partners/matlab:r2020b
```



AI Model Design

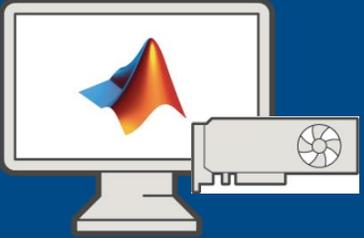


Easy access to a GPU

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

The screenshot shows the Experiment Manager app interface. The top menu bar includes options like New, Save, Duplicate, Layout, Run, Stop, Training Plot, Confusion Matrix, Filter, and Export. The main area displays a list of trials for a 'Baseline Tuning' experiment. The table below shows the details of these trials.

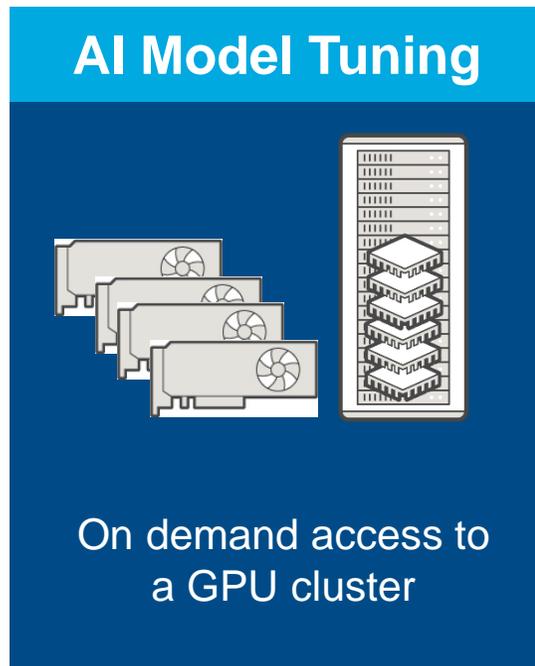
Trial	Status	Progress	Elapsed Time	myInitialLearn...	convFilterSize	Training Accu...	Training Loss	Validation Ac..
1	Complete	100.0%	0 hr 0 min 16 sec	1.0000e-6	3.0000	12.5000	2.6441	10.
2	Complete	100.0%	0 hr 0 min 15 sec	1.0000e-5	3.0000	25.7813	2.1228	20.
3	Complete	100.0%	0 hr 0 min 14 sec	0.0001	3.0000	64.8438	1.0878	42.
4	Complete	100.0%	0 hr 0 min 16 sec	0.0005	3.0000	90.6250	0.4648	49.
5	Complete	100.0%	0 hr 0 min 15 sec	1.0000e-6	4.0000	11.7188	2.4967	6.
6	Complete	100.0%	0 hr 0 min 15 sec	1.0000e-5	4.0000	23.4375	2.1213	14.
7	Complete	100.0%	0 hr 0 min 17 sec	0.0001	4.0000	72.6563	1.0283	39.
8	Running	30.7%	0 hr 0 min 4 sec	0.0005	4.0000			
9	Queued	0.0%		1.0000e-6	5.0000			
10	Queued	0.0%		1.0000e-5	5.0000			
11	Queued	0.0%		0.0001	5.0000			
12	Queued	0.0%		0.0005	5.0000			
13	Queued	0.0%		1.0000e-6	6.0000			
14	Queued	0.0%		1.0000e-5	6.0000			
15	Queued	0.0%		0.0001	6.0000			
16	Queued	0.0%		0.0005	6.0000			

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

云端并行多 GPU 训练

Steps to add a cluster to a MATLAB session

1. Setup Parallel Server in MathWorks Cloud Center



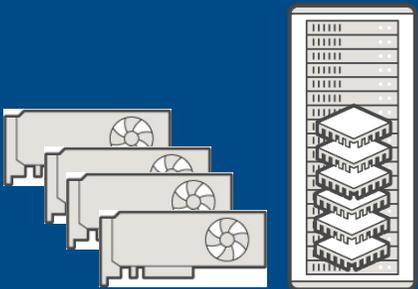
The screenshot shows the 'Create Cluster' page in the MathWorks Cloud Center. The browser address bar is 'cloudcenter.mathworks.com/cluster/create'. The page is titled 'Create Cluster - Cloud Center'. The main content area is divided into sections: 'Preferences', 'Location & Network', and 'Cluster Configuration'. The 'Cluster Configuration' section is expanded, showing options for 'Shared State' (Personal Cluster selected), 'Auto-Manage Cluster Access' (checked), 'Worker Machine Type' (Double Precision GPU (p3.2xlarge, 4 core, 1 GPU)), 'Workers per Machine' (1), 'Use a dedicated headnode' (checked), 'Headnode Machine Type' (Standard (m5.xlarge, 2 core)), and 'Allow cluster to auto-resize' (unchecked). The 'Workers in Cluster' section shows 'Initial Count' and 'Upper Limit' both set to 18, with a 'Max: 1024' label. The 'Machines in Cluster' section shows 19 machines, including the headnode. A note at the bottom states: 'Note: You are charged for the use of your cloud provider's clusters. Consider periodically checking your active resources through your cloud provider account.'

云端并行多 GPU 训练

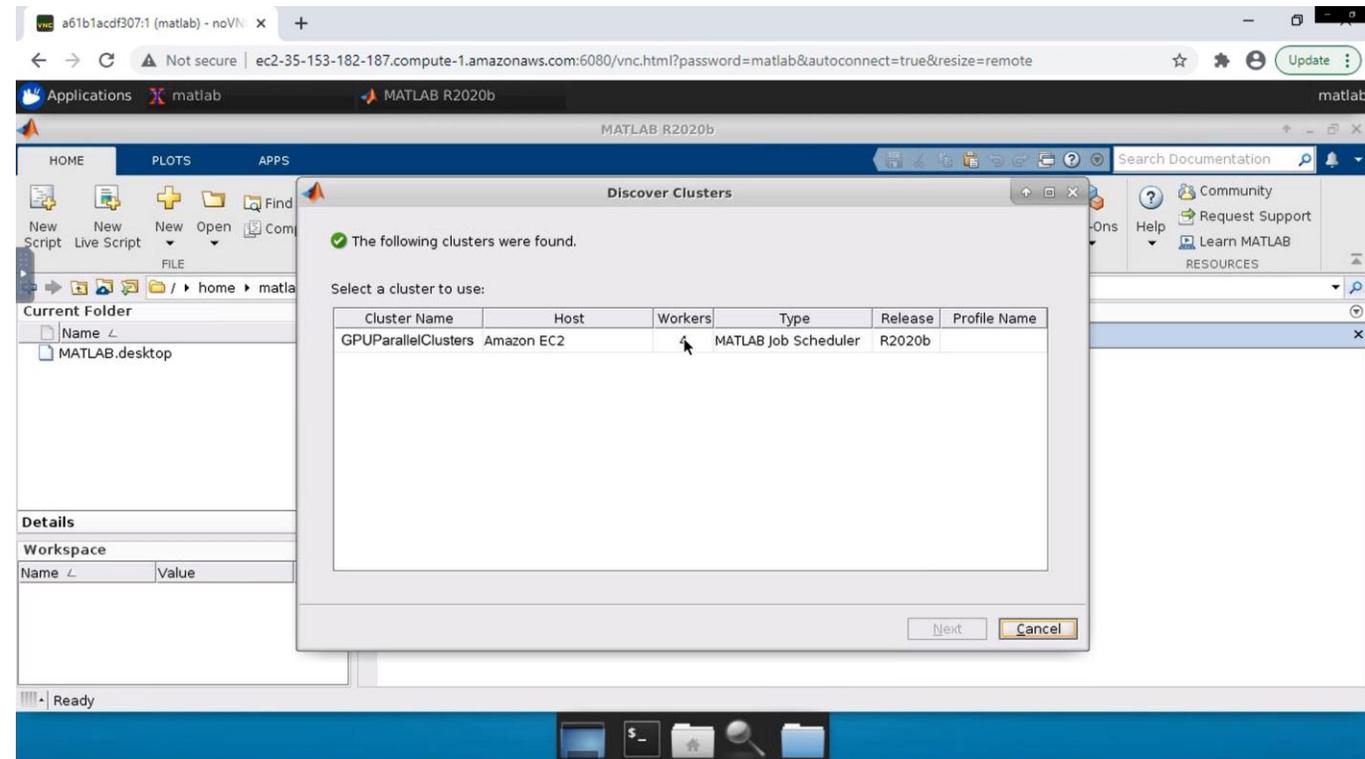
Steps to add a cluster to a MATLAB session

2. Change Default Cluster

AI Model Tuning



On demand access to
a GPU cluster

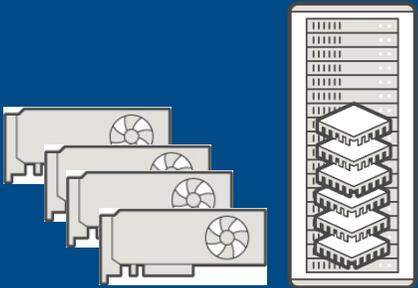


云端并行多 GPU 训练

What options are available for training at scale?



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

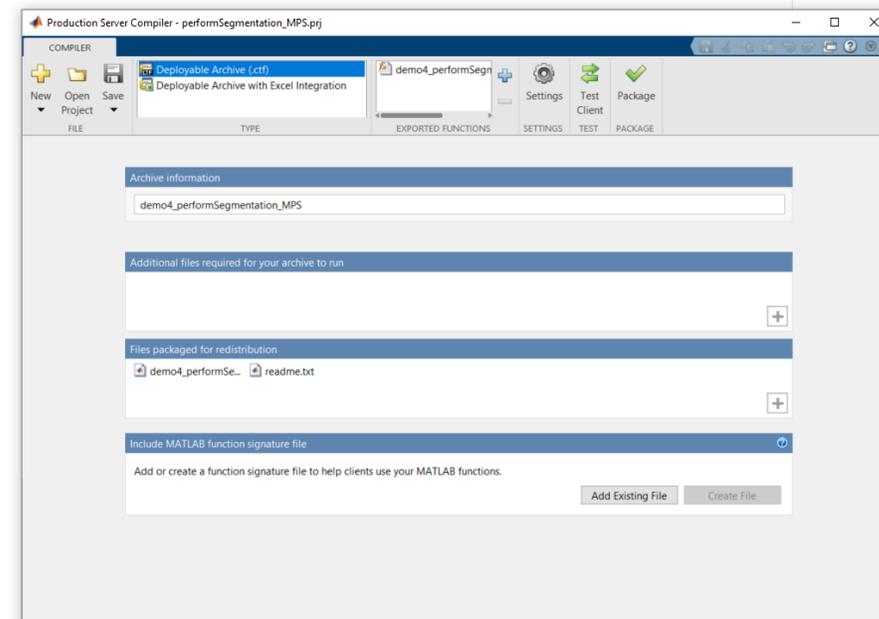
Steps to follow before deploying to the cloud

1. Create a Function that runs the trained model
2. Package (Production Server Compiler App)

Deployment



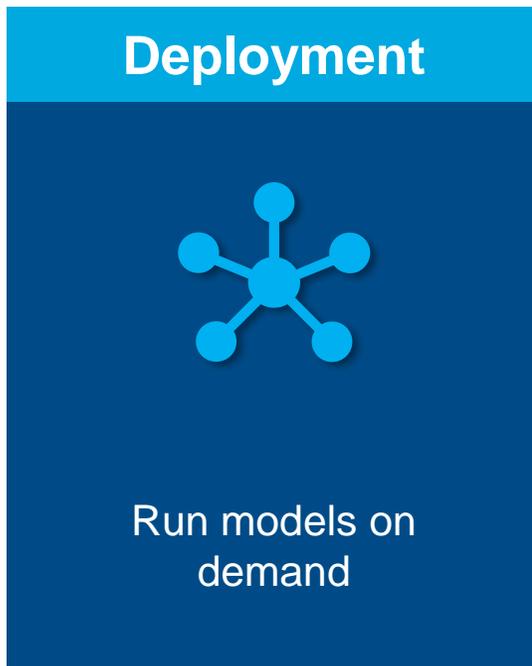
Run models on demand



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

The screenshot shows the Azure Marketplace page for MATLAB Production Server (PAYG) by MathWorks. The page includes a search bar, navigation tabs for Overview, Plans, and Reviews, and a detailed description of the product. A 'GET IT NOW' button is visible. The right side of the page displays a configuration dashboard with various settings and status indicators.

Overview	Applications	Settings	Persistence	Application Access Control	Dashboard Access Control	Logs
MATLAB Execution Endpoint	https://mat4ez00d0dayeastus.cloudapp.azure.com					
MATLAB Endpoint Status	READY					
Dashboard Version	1.0.0					
MATLAB Production Server Version	R2020a					
MATLAB Runtime Versions	[R2020a, R2019b, R2018a, R2017b, R2017a, R2017c]					
Server VM Operating System	Linux					
Number of Server VMs	1					
Last Refresh Time	2:58:12 PM					

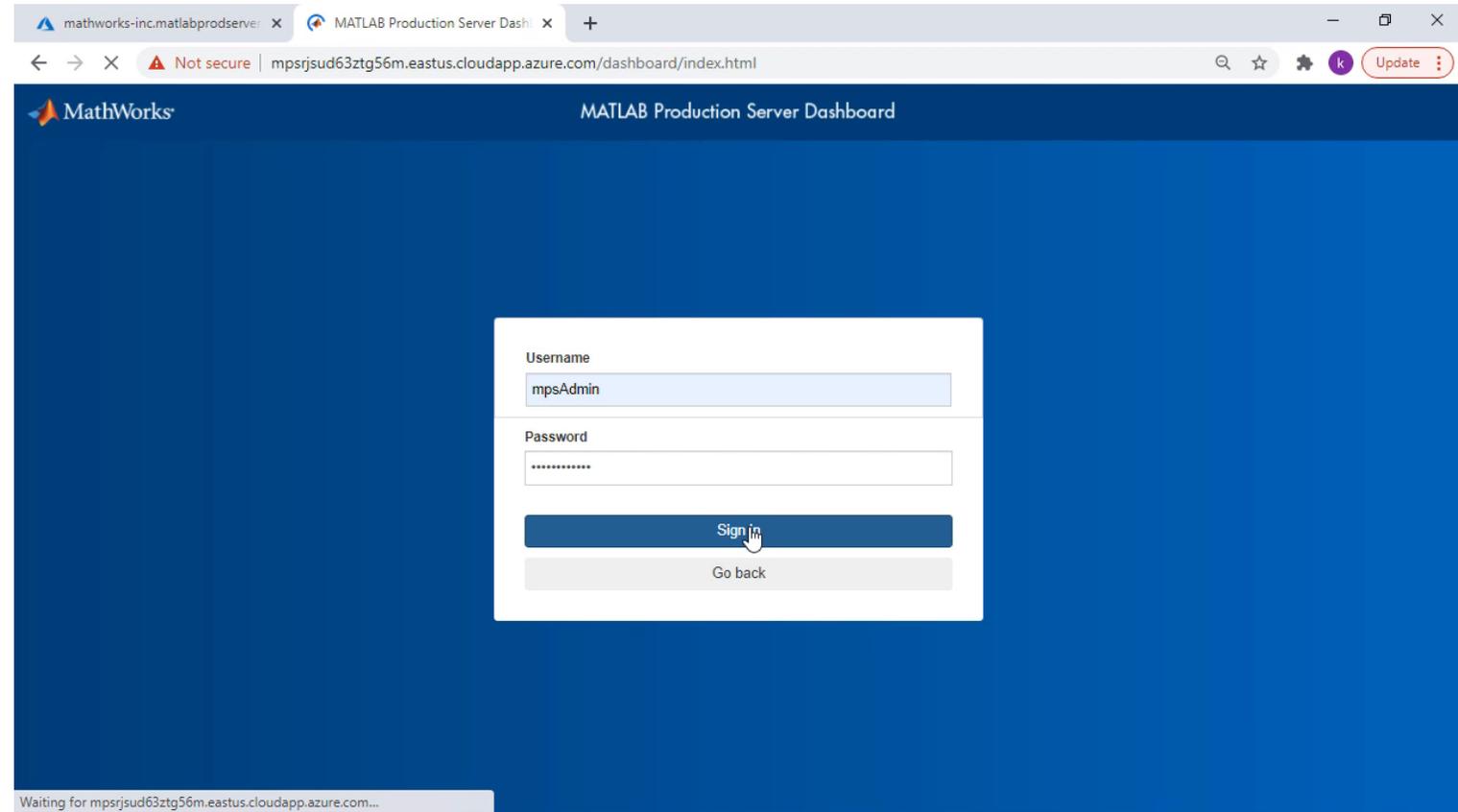
将MATLAB模型部署到云端

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

Semantic Segmentation

This example shows an application that performs semantic segmentation on images stored in cloud.

You run this example by entering the url to the Azure Blob that contains images

Azure Blob url

`'wasbs://camvidblob@semar'`

Azure access keys

`'?sv=2019-10-10&st=2020-1'`

Segmentation Complete

Category	Color
Bicyclist	Blue
Pedestrian	Green
Car	Purple
Fence	Light Blue
SignSymbol	Light Green
Tree	Light Blue
Pavement	Light Blue
Road	Light Blue
Pole	Light Blue
Building	Light Blue
Sky	Light Blue

总结

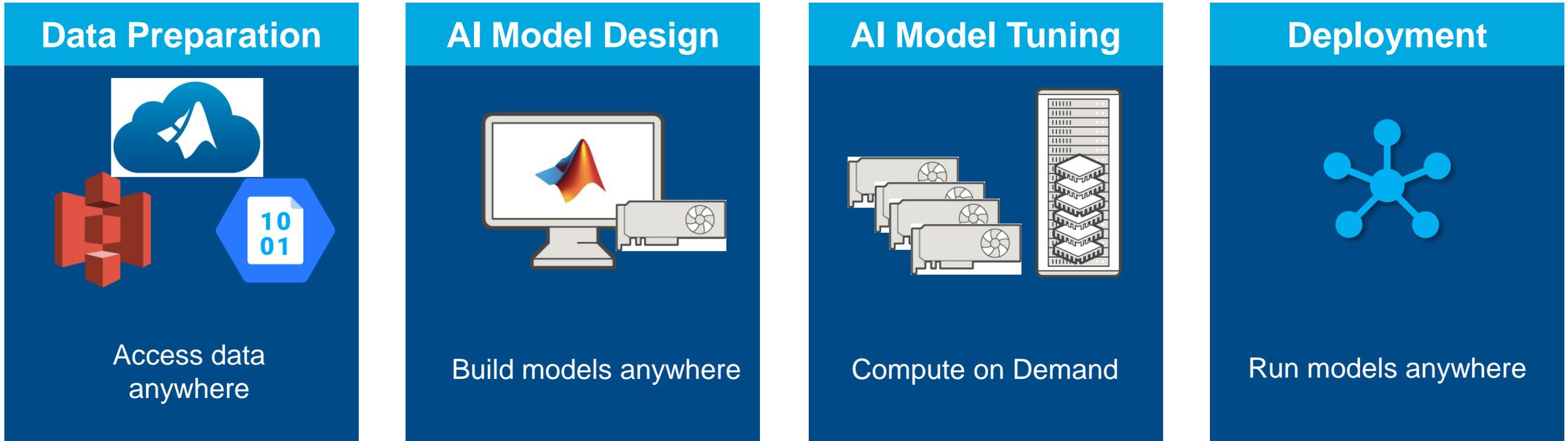
Deploy to Embedded and Enterprise systems from one codebase

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



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

MATLAB EXPO

2021

Thank you



© 2021 The MathWorks, Inc. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See [mathworks.com/trademarks](https://www.mathworks.com/trademarks) for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.