

Which of the following AI systems were enabled by the cloud?



# MATLAB EXPO 2021

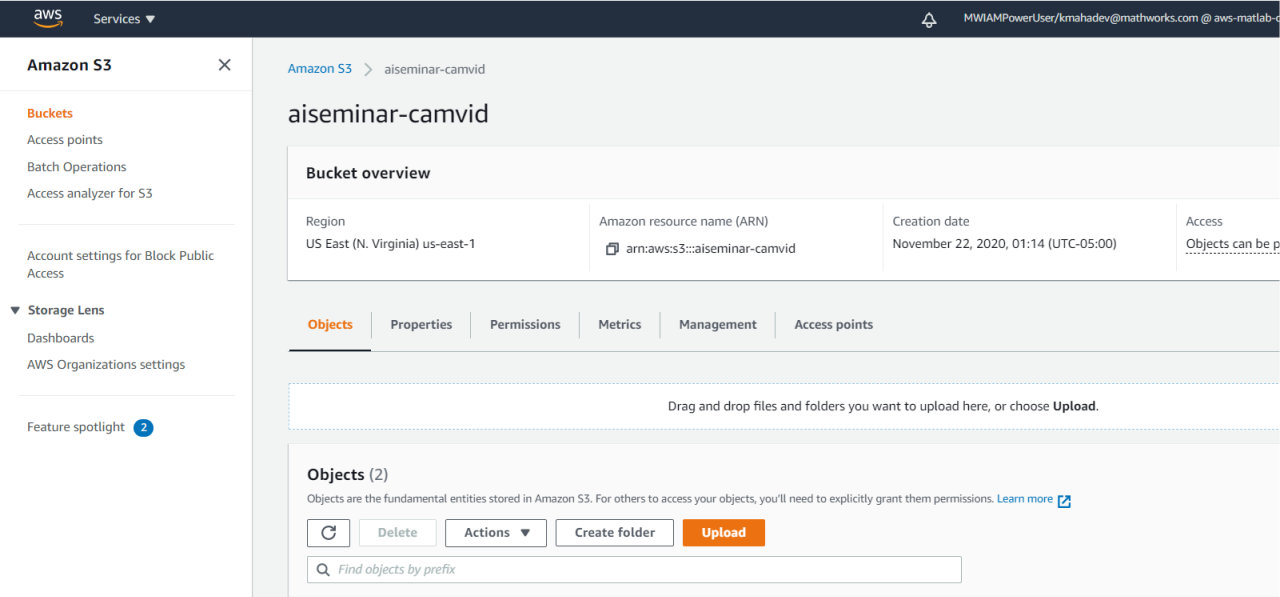
## AI Workflows in the Cloud

*David Willingham*

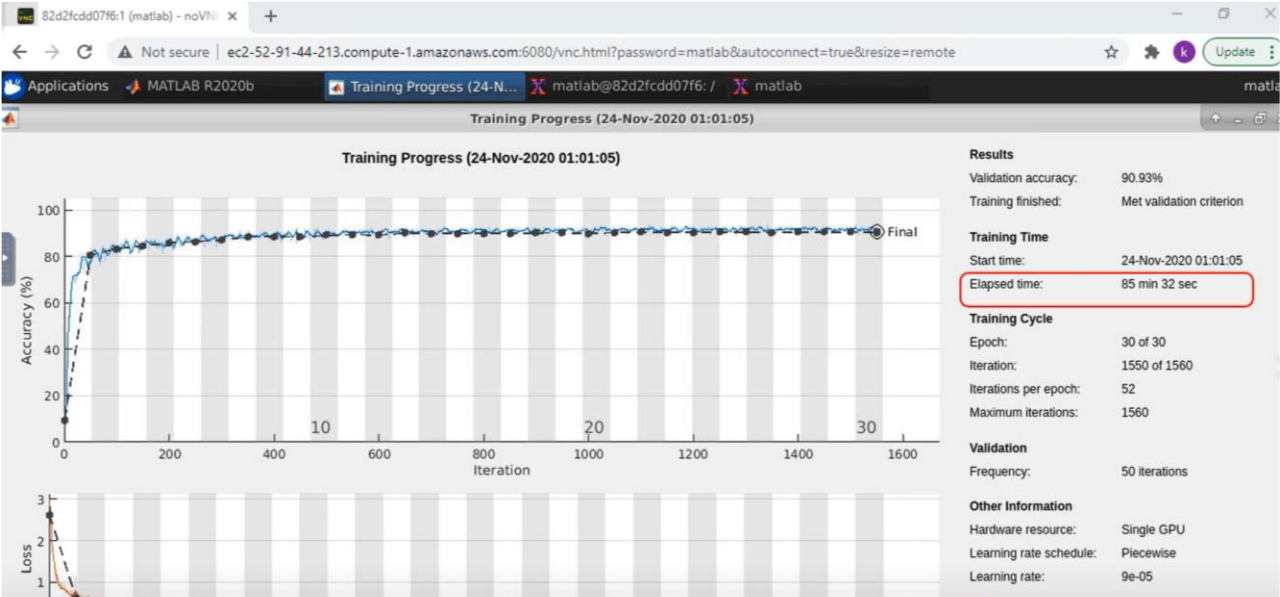


*Kishen Mahadevan*





## Accessed Data on the Cloud



## Trained a Model on the Cloud

Trial	Status	Progress	Elapsed Time	NetworkName	Training Accurac...	Training Loss	Validation Accura...	Validation L
1	Complete	100.0%	1 hr 33 min 20 sec	resnet18	90.7794	0.2229	90.9133	
2	Complete	100.0%	1 hr 5 min 27 sec	resnet50	93.3542	0.1810	92.1968	
3	Complete	100.0%	1 hr 39 min 45 sec	mobilenetv2	87.1009	0.3602	88.3391	

## Tuned a Model on the Cloud

## Deployed a Model to the Cloud

# How I got Started

I'm an experienced controls engineer, who's ramping up on Deep Learning

- Started with the free Deep Learning Onramp



## Deep Learning Onramp

Get started quickly using deep learning methods to perform image recognition.

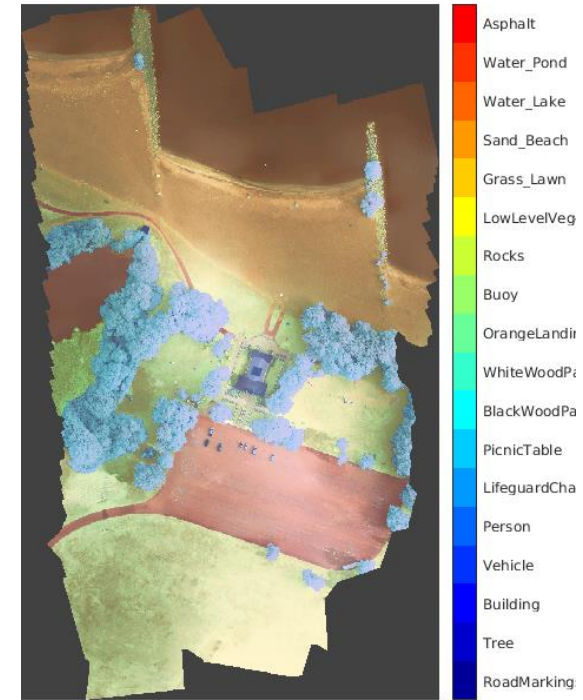
# Case Study – Semantic Segmentation

Classifying every pixel in an image/video

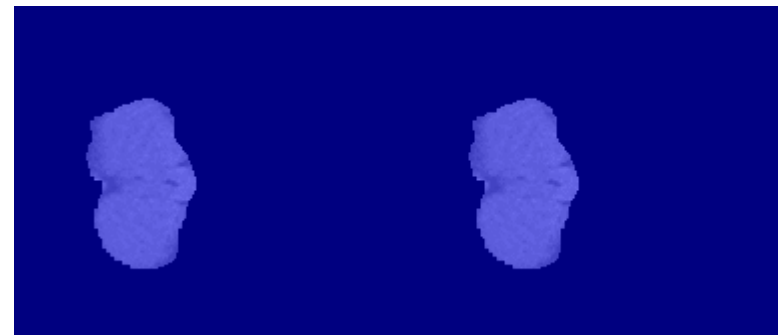
- Training Data – 557MB
- Training Time ~ 5hrs
  - NVIDIA™ Titan X with 12 GB memory



Road Segmentation for Autonomous Driving



Tracking Deforestation using Multispectral Images



3D Tumor Detection from MRI

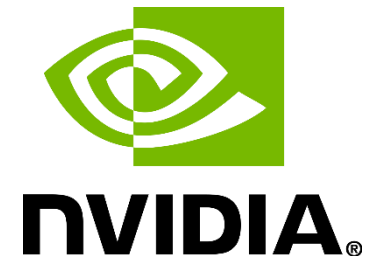
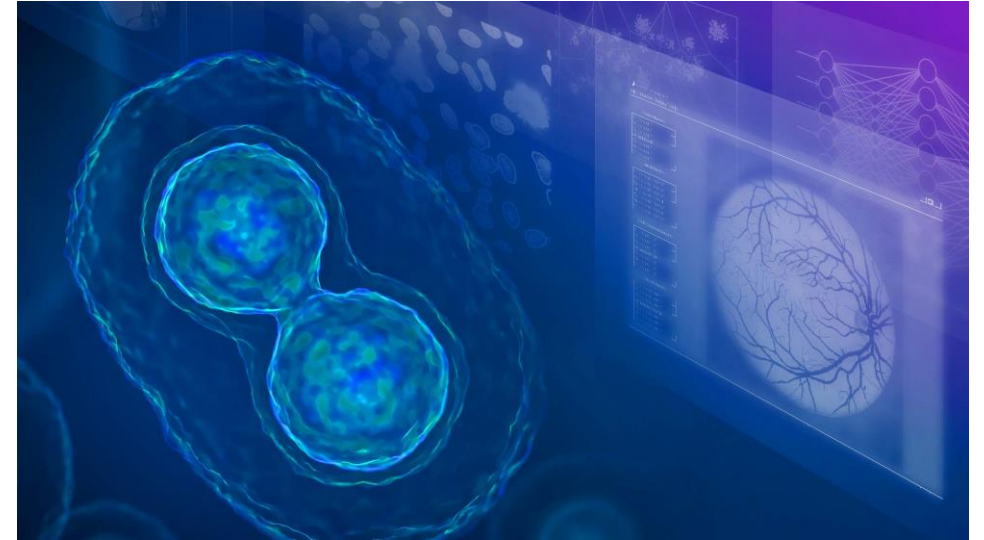
# BioMedical Devices Advances Organ Cell Growth Research

With the help of MATLAB and NVIDIA hardware on AWS

“MATLAB removes that level of friction for us so that we can just get down to the business of doing research.”

“Amazon EC2 P3 Instances provided the compute that we didn't have to go out and buy when we made the decision to scale up”

*Sam Raymond, PostDoc Stanford*

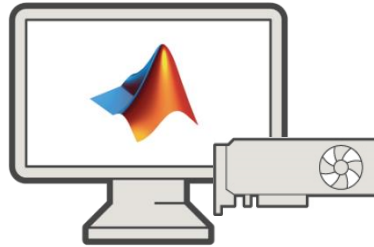


[Source: AWS-MathWorks-Case Study 2020](#)

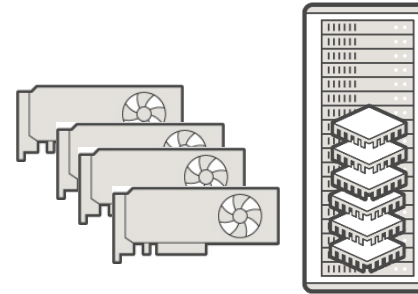
# Why perform AI on the cloud?



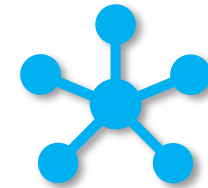
**Access Data  
anywhere**



**Build models  
anywhere**



**Compute on  
Demand**



**Run Models  
anywhere**

# AI System Design 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



# AI System Design 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

# AI System Design Workflow

## Data Preparation

 Data cleansing and preparation

 Human insight

 Simulation-generated data

## AI Model Design

 Model design

 Hardware accelerated training

 Interoperability

## AI Model Tuning

 Model tuning

 Hardware accelerated training

 Interoperability

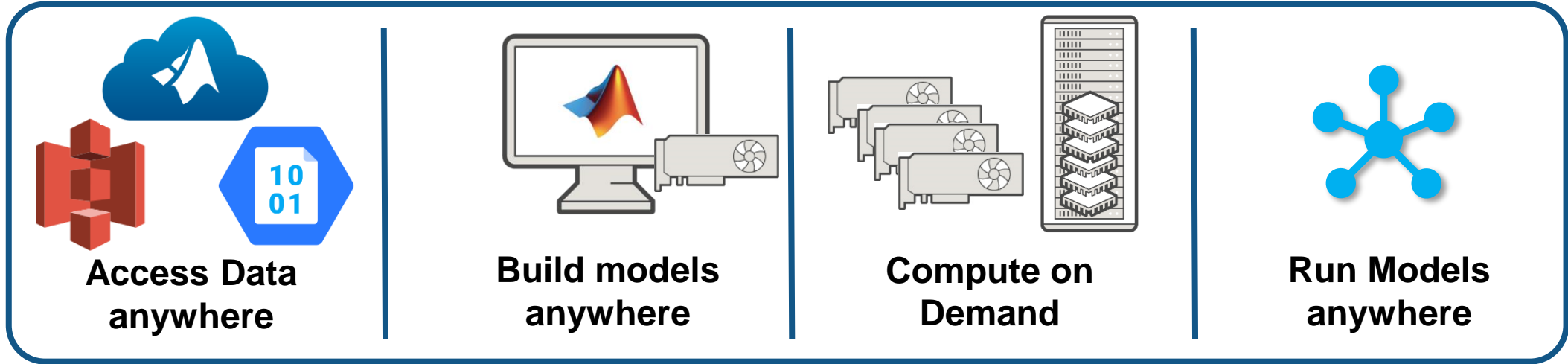
## Deployment

 Embedded devices

 Enterprise systems

 Edge, cloud, desktop

# What does AI System Design in the Cloud look like?



### Data Preparation

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

### AI Model Design

- Model design
- Hardware accelerated training
- Interoperability

### AI Model Tuning

- Model tuning
- Hardware accelerated training
- Interoperability

### Deployment

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

# What does AI System Design in the Cloud look like?

## Data Preparation



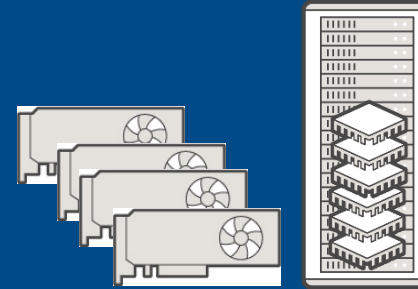
Access data  
anywhere

## AI Model Design



Build models anywhere

## AI Model Tuning



Compute on demand

## Deployment



Run models anywhere

# What does AI System Design in the Cloud look like?

## Data Preparation



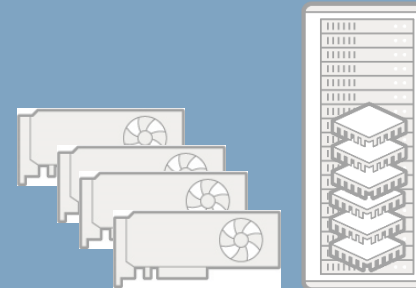
Access data  
anywhere

## AI Model Design



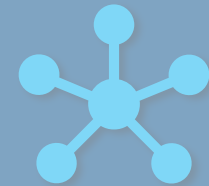
Build models anywhere

## AI Model Tuning



Compute on demand

## Deployment




Run models anywhere

# Data in the Cloud = Data accessible anywhere

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**

A graphic for 'Data Preparation' on a dark blue background. It features a white cloud icon with a white arrow pointing upwards, a cluster of red 3D cubes, and a blue hexagonal icon containing the binary code '10' over '01'.

Store preprocessed data anywhere



# Data in the Cloud – Bring it close to the compute

## Steps to follow

1. Generate security keys
2. Create a bucket on AWS S3 and upload
3. Verify data in S3



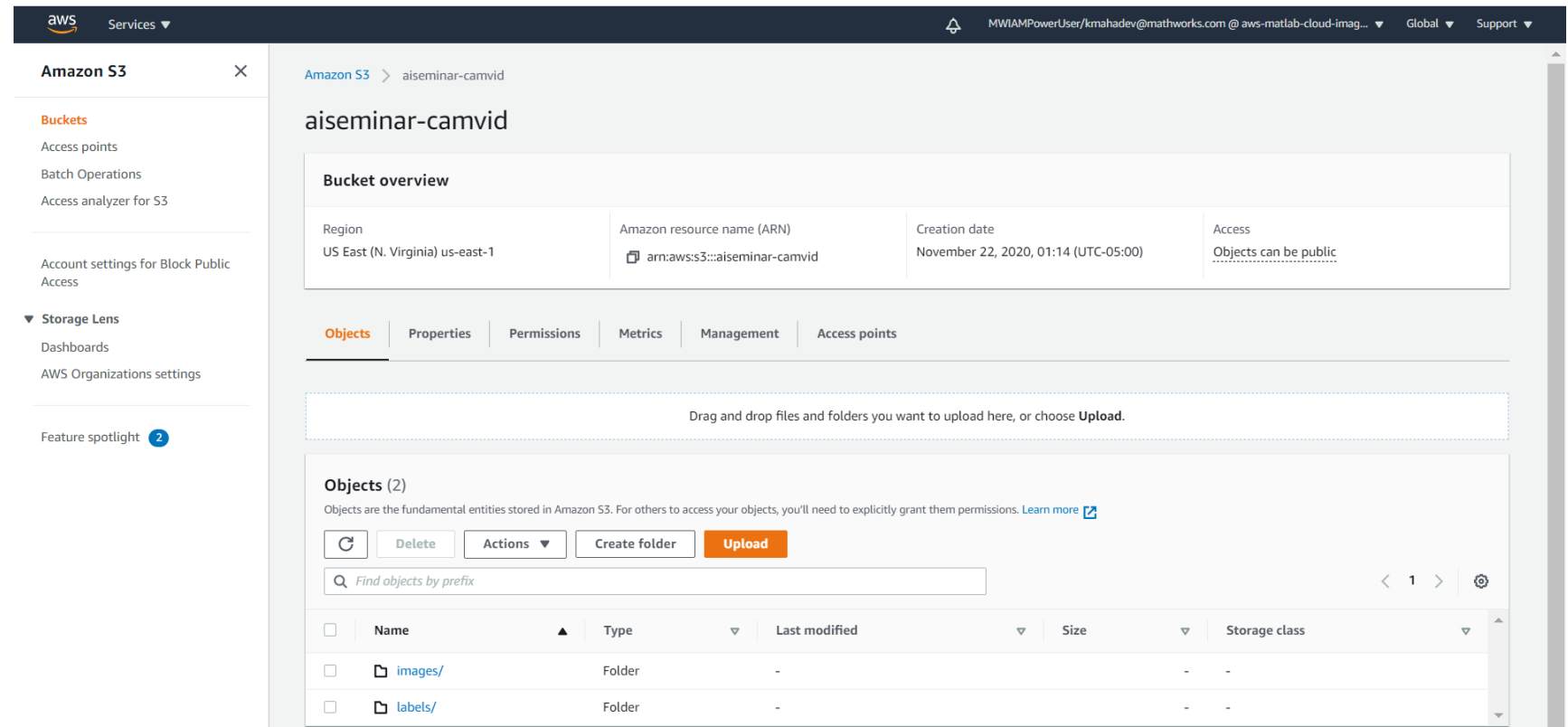
# Data in the Cloud – Bring it close to the compute

## Uploading Data to the S3

**Data Preparation**



Store preprocessed data anywhere



**Amazon S3** Services

aiseminar-camvid

**aiseminar-camvid**

**Bucket overview**

Region US East (N. Virginia) us-east-1	Amazon resource name (ARN) arn:aws:s3:::aiseminar-camvid	Creation date November 22, 2020, 01:14 (UTC-05:00)	Access Objects can be public
-------------------------------------------	-------------------------------------------------------------	-------------------------------------------------------	---------------------------------

**Objects (2)**

Objects are the fundamental entities stored in Amazon S3. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

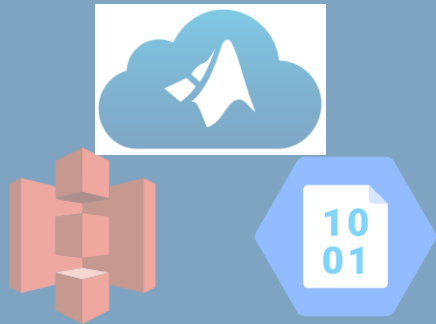
Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	images/	Folder	-	-	-
<input type="checkbox"/>	labels/	Folder	-	-	-



# What does AI System Design in the Cloud look like?

## Data Preparation



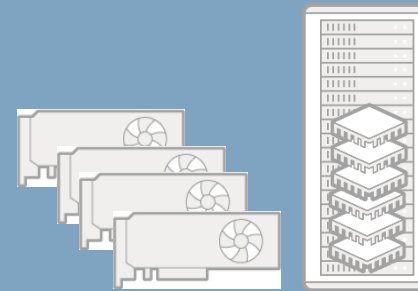
Access data  
anywhere

## AI Model Design



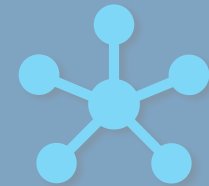
Build models anywhere

## AI Model Tuning



Compute on demand

## Deployment




Run models anywhere

# Easy Access to HPC resources

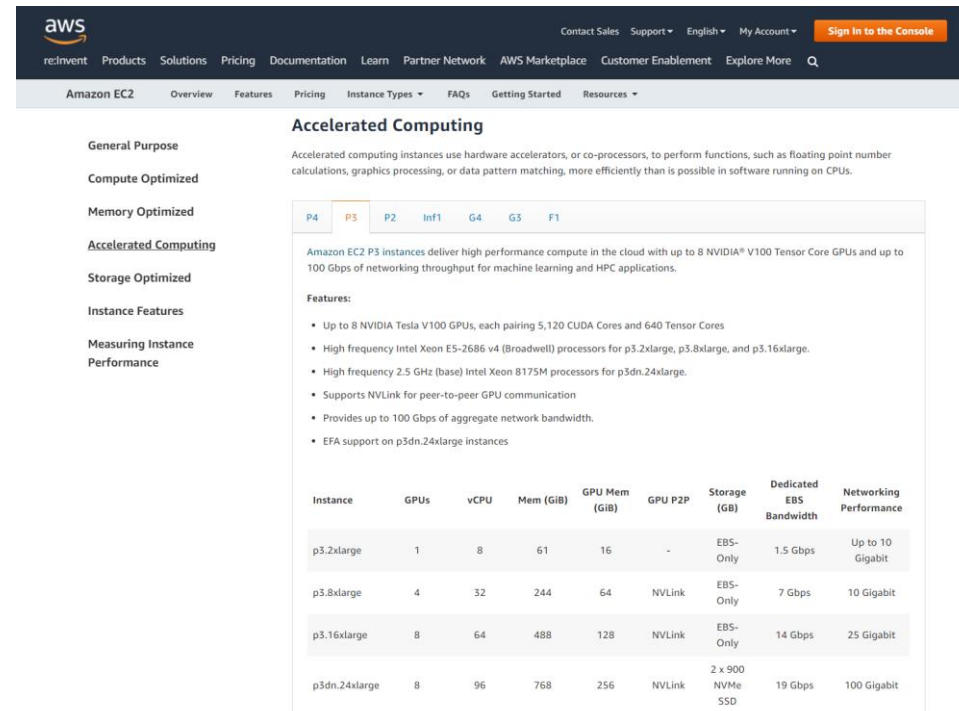
It's a balance between speed and cost

- Which VM did you choose?

## AI Model Design



Easy access to a GPU



**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.

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



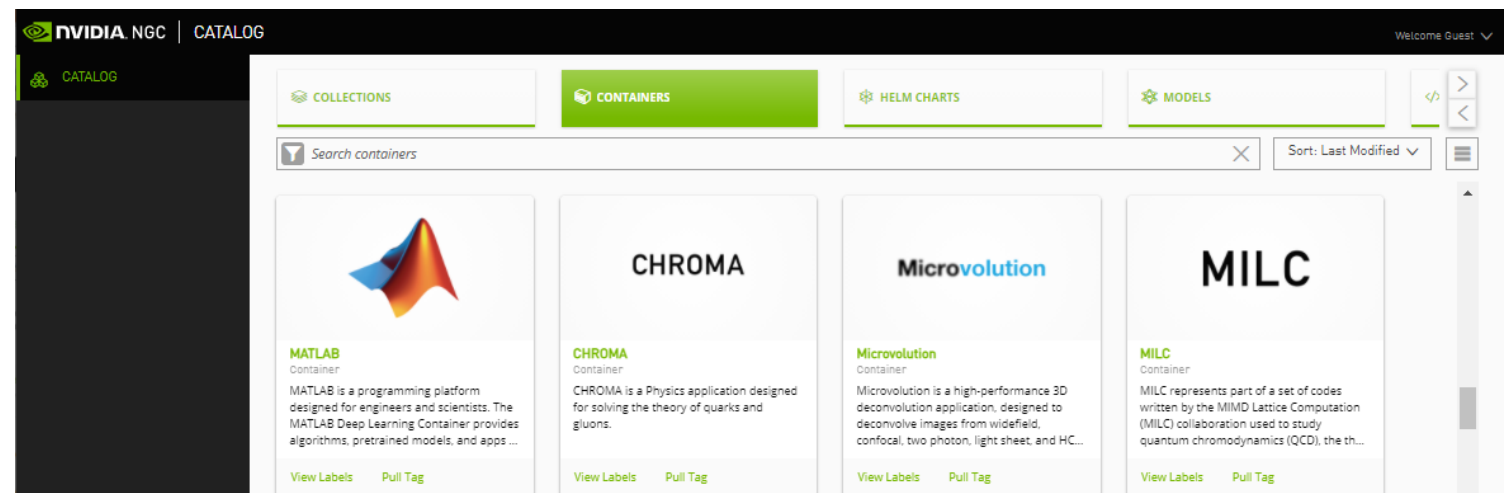
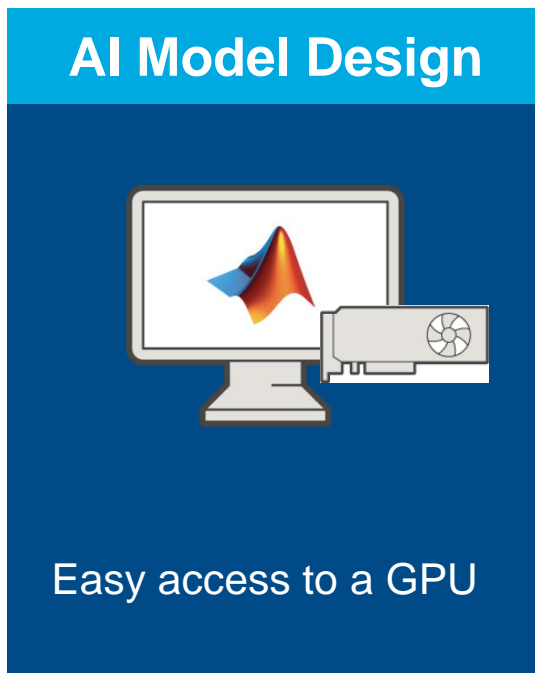
# Easy Access to GPU resources

## 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

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



# Cloud Setup for Deep Learning in MATLAB

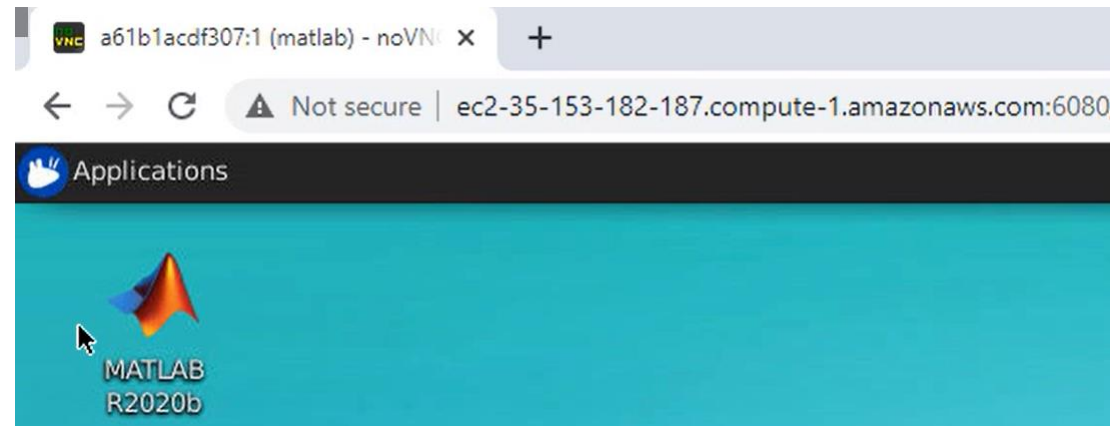
## Steps to using the Deep Learning Container

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

# Cloud Setup for Deep Learning in MATLAB

## Steps to using the Deep Learning Container

### AI Model Design



Easy access to a GPU

MathWorks

Training DL network for Semantic Segmentation

6

# What does AI System Design in the Cloud look like?

## Data Preparation



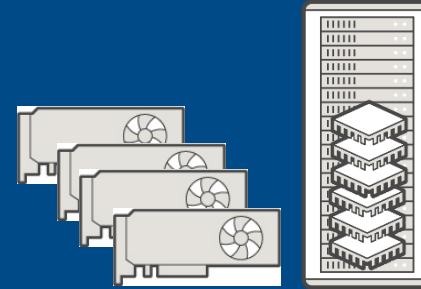
Access data  
anywhere

## AI Model Design



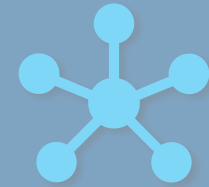
Build models anywhere

## AI Model Tuning



Compute on demand

## Deployment



Run models anywhere

# Find the Optimal Network Using Experiments

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 main window displays a table of trials with columns for Trial, Status, Progress, Elapsed Time, myInitialLearn..., convFilterSize, Training Accu..., Training Loss, and Validation Ac... The table shows 16 trials, with trials 1-7 completed, trial 8 running, and trials 9-16 queued.

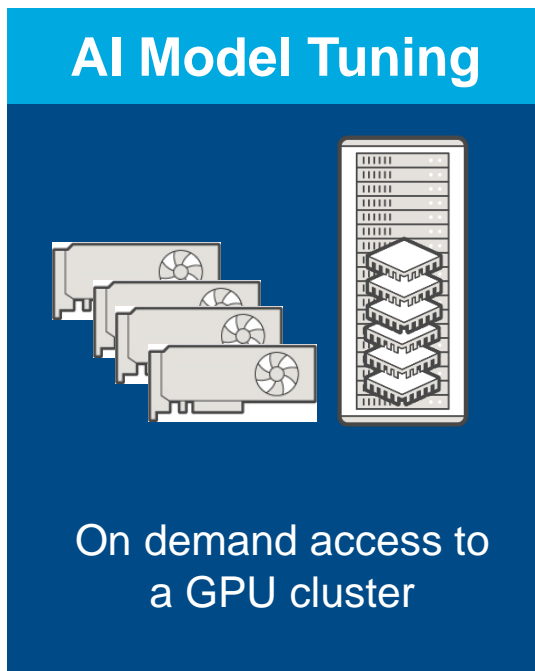
Trial	Status	Progress	Elapsed Time	myInitialLearn...	convFilterSize	Training Accu...	Training Loss	Validation Ac...
1	Complete	100.0%	0 hr 0 min 16 sec	1.0000e-6	3.0000	12.5000	2.6441	10.
2	Complete	100.0%	0 hr 0 min 15 sec	1.0000e-5	3.0000	25.7813	2.1228	20.
3	Complete	100.0%	0 hr 0 min 14 sec	0.0001	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

# Tune and Compare Networks with Scalable Compute

## Using High Performance Cloud Instances

- **Iterate**
  - Running Experiments requires many trial iterations
- **Wait**
  - If 1 trial can take hrs to days, Experiments can take days to weeks
    - Our example takes 1.5hrs to train
    - 10trials = 15hrs
    - 100trials = 150hrs or 6.25days
- **Scale**



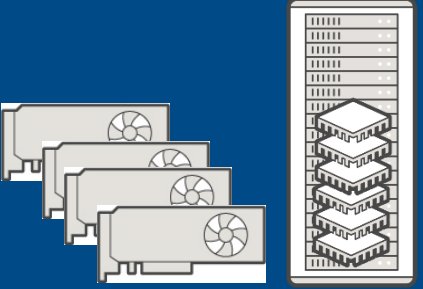


# Parallel Multi-GPU Training in the cloud

What options are available for training at scale?

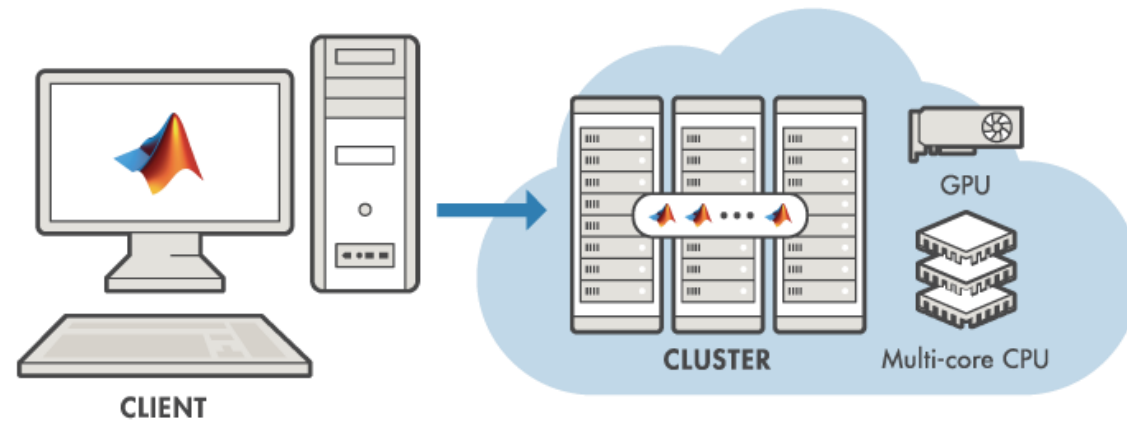
- 4 node GPU cluster chosen

**AI Model Tuning**



On demand access to a GPU cluster

The illustration shows four server racks with fans on the left and a single server rack with multiple GPU cards on the right, all set against a dark blue background.

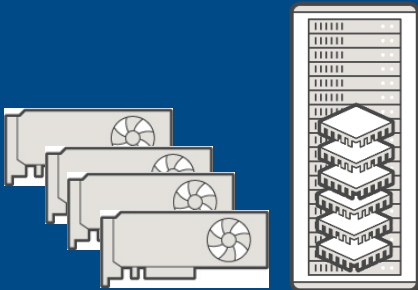


# Parallel Multi-GPU Training in the cloud

## Steps to add a cluster to a MATLAB session

1. Setup Parallel Server in MathWorks Cloud Center
2. Change Default Cluster

### AI Model Tuning



On demand access to  
a GPU cluster

The screenshot shows the 'Create Cluster' interface in the MathWorks Cloud Center. The cluster name is 'GPUParallelCluster', the MATLAB version is 'R2020b', and the cluster log level is 'Low'. The worker machine type is 'Double Precision GPU (p3.2xlarge, 4 core, 1 GPU)'. The headnode machine type is 'Standard (m5.xlarge, 2 core)'. The number of workers per machine is set to 1, and the number of machines in the cluster is 19 (including the headnode).

The MATLAB R2020b session shows the 'Discover Clusters' dialog box with the following table of discovered clusters:

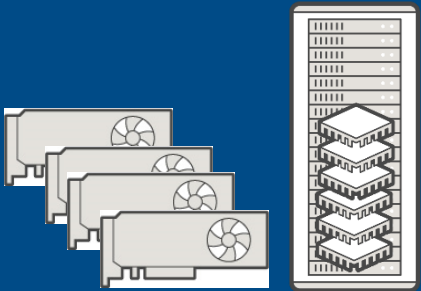
Cluster Name	Host	Workers	Type	Release	Profile Name
GPUParallelClusters	Amazon EC2	1	MATLAB Job Scheduler	R2020b	

# Parallel Multi-GPU Training in the cloud

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

# What does AI System Design in the Cloud look like?

## Data Preparation



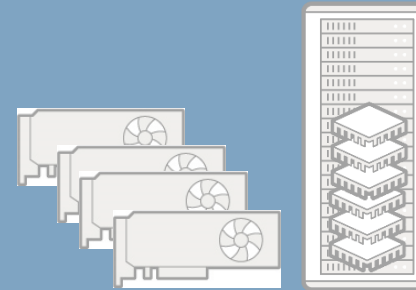
Access data  
anywhere

## AI Model Design



Build models anywhere

## AI Model Tuning



Compute on demand

## Deployment

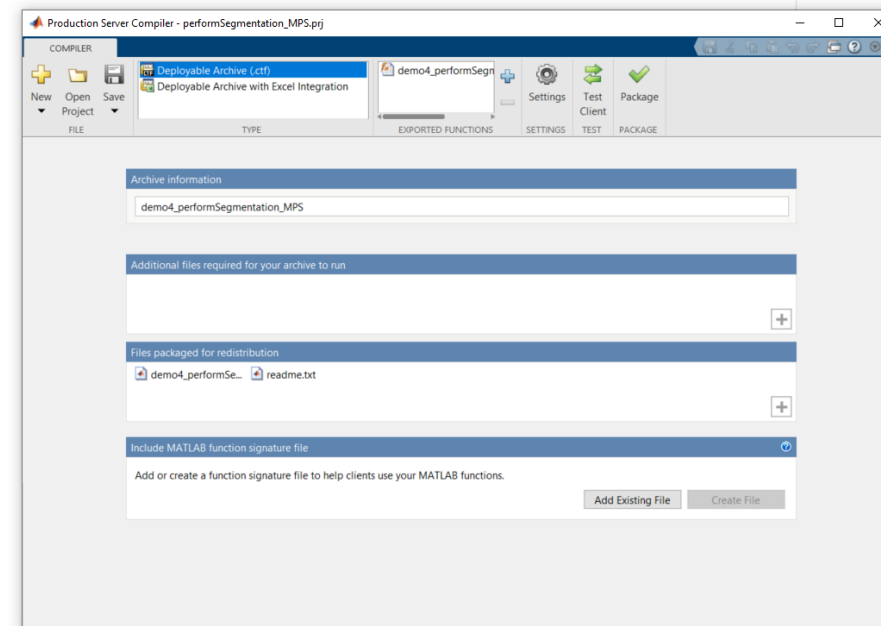


Run models anywhere

# Preparing the trained model for deployment

Steps to follow before deploying to the cloud

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



# Models Deployed in the Cloud = Models Available Anywhere

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



# Deploying MATLAB Models in the Cloud

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. On the right, there is a configuration table for the deployment.

Property	Value
MATLAB Execution Endpoint	https://mat4ezco0d0dayeastus.cloudapp.azure.com
MATLAB Endpoint Status	READY
Dashboard Version	1.0.0
MATLAB Production Server Version	R2020a
MATLAB Runtime Versions	[R2020a, R2019b, R2019a, R2018b, R2018a, R2017b]
Server VM Operating System	Linux
Number of Server VMs	1
Last Refresh Time	2:58:12 PM

# Deploying MATLAB Models in the Cloud

## Steps to follow

1. Start MATLAB Production Server VM
2. Upload compiled function
3. Enable model caching
4. Link with a front end of your choice

### Deployment



Run models on  
demand



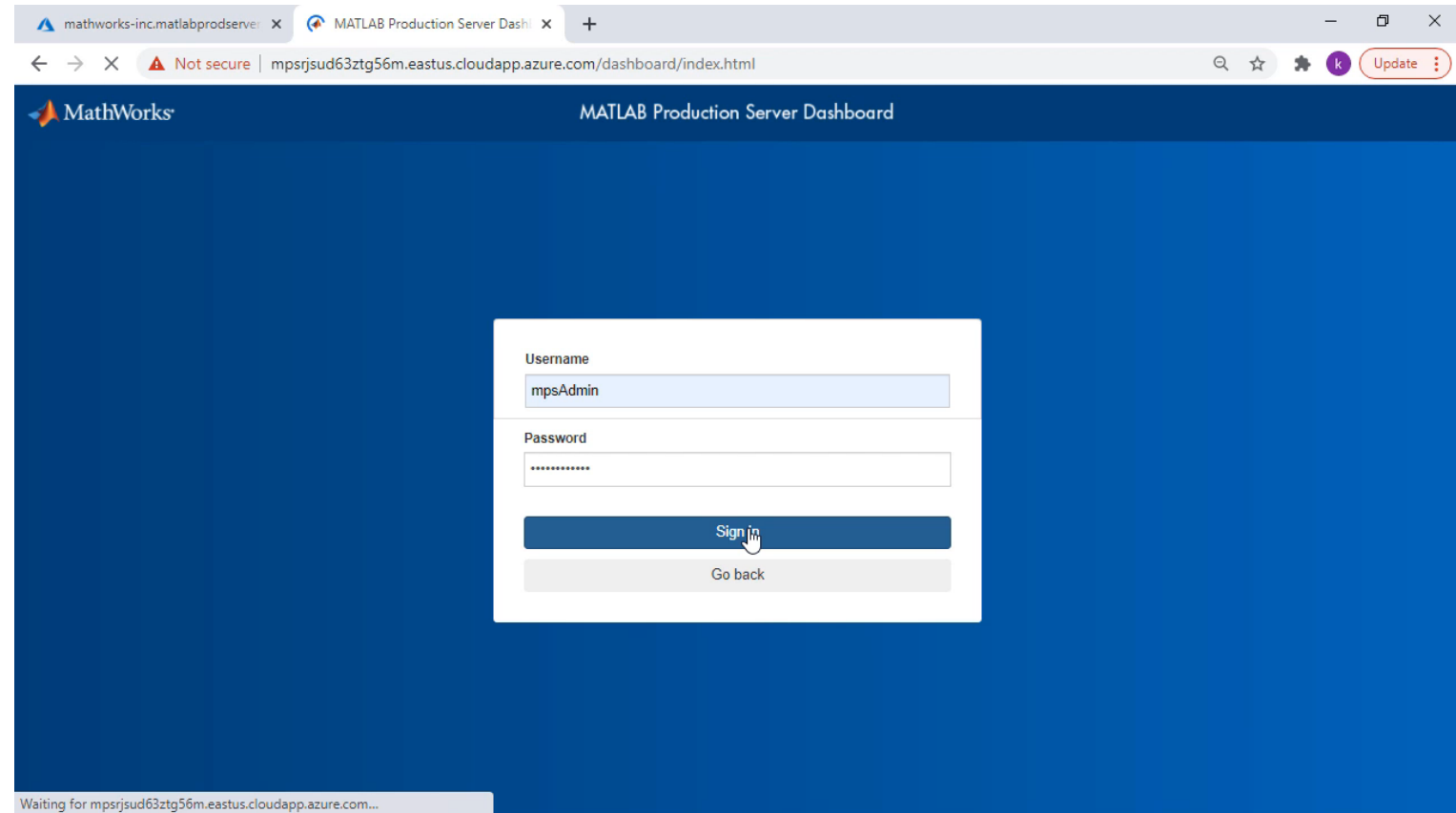
# Deploying MATLAB Models in the Cloud

## Steps to follow

### Deployment



Run models on demand



# Deploying MATLAB Models in the Cloud

## What options are available?

### Deployment



Run models on demand



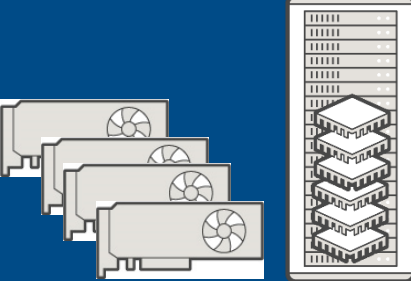

The screenshot shows a web browser window titled "Semantic Segmentation". The address bar contains the URL: `C:/Filetransfer/File%20transfer/Working/Working/TMDG/DL%20Matlab%20Cloud/AISeminar/demo4/deploy/demo4Client.html`. The page content includes:

- Semantic Segmentation** (Section Header)
- 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** (Section Header)
- 
- Azure access keys** (Section Header)
- 
- 
- Segmentation Complete

On the right side of the page, there is a street scene image with a legend for semantic segmentation. The legend includes the following categories and colors:

- Bicyclist (Green)
- Pedestrian (Yellow)
- Car (Red)
- Fence (Blue)
- SignSymbol (Purple)
- Tree (Light Green)
- Pavement (Dark Blue)
- Road (Grey)
- Pole (Light Blue)
- Building (Orange)
- Sky (White)

# Summary – What was shown?

<p><b>Data Preparation</b></p>  <p>Access data anywhere</p>	<p><b>AI Model Design</b></p>  <p>Build models anywhere</p>	<p><b>AI Model Tuning</b></p>  <p>Compute on Demand</p>	<p><b>Deployment</b></p>  <p>Run models anywhere</p>
----------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------

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

## What's next?

“My manager heard about the results of my project. I’ve been asked to develop and run a training course to over 100 colleagues!”

**MATLAB EXPO**

Integrating AI into Model-Based Design

# Resources from today's talk in **Handouts Tab**

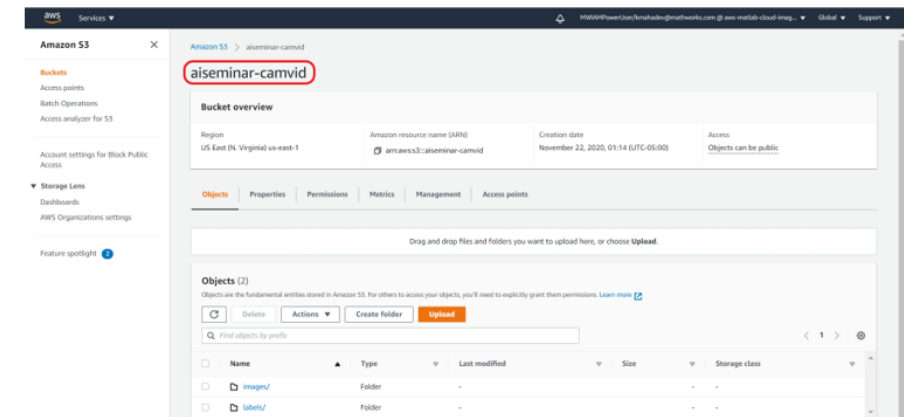
- **Recipes for each section**
- **Examples for each section**

## AI Workflows in the Cloud – Handout Recipes

Example: [Semantic Segmentation](#)

Data preparation:

- **Generate secure access keys:**
  - Assign [IAM role to your instance](#), or
  - Use [AWS SAML](#) to create temporary access keys and set them as env variables.
- **Create a bucket** on AWS S3 and upload data using steps mentioned in the '[Upload data to s3](#)'
- **Verify data** in S3 bucket.
- Refer to '[UploadDataToS3.txt](#)' file for an example showcasing use of AWS SAML



# MATLAB EXPO 2021

Thank you

