

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



#### **AI Workflows in the Cloud**

David Willingham

Kishen Mahadevan







aws Services ▼	ل MWIAMPowerUser/kmahadev@mathworks.com @ aws-matlab-	s 🔜 82d2fcdd07f6:1 (matlab) - noV™ 🗙 🕂	- 0 ×
Amazon S3 X		← → C 🔺 Not secure   ec2-52-91-44-213.compute-1.amazonaws.com:6080/vnc.html?password=matlab&autoconnect=true&resize=remote	🖈 🛊 🚷 Update 🚦
	Amazon S5 > aiseminar-camvid	🗳 Applications 🕠 MATLAB R2020b 🕢 Training Progress (24-N X matlab@82d2fcdd07f6: / X matlab	matla
Buckets	aiseminar-camvid	Training Progress (24-Nov-2020 01:01:05)	+ - 83
Access points Batch Operations Access analyzer for S3	Bucket overview	Training Progress (24-Nov-2020 01:01:05) Resu Valida	Its ation accuracy: 90.93%
Account settings for Block Public Access	Region     Amazon resource name (ARN)     Creation date     Access       US East (N. Virginia) us-east-1     Image: arrn:aws:s3:::aiseminar-camvid     November 22, 2020, 01:14 (UTC-05:00)     Objects can be presented on the presented on	100 Final Train 80 Final Start Elass	ing finished: Met validation criterion ing Time time: 24-Nov-2020 01:01:05 ued time: 85 min 32 sec
<ul> <li>Storage Lens</li> <li>Dashboards</li> <li>AWS Organizations settings</li> </ul>	Objects         Properties         Permissions         Metrics         Management         Access points	C 60 1 Train Epoc teration	ing Cycle h: 30 of 30 ion: 1550 of 1560
Feature spotlight <b>2</b>	Drag and drop files and folders you want to upload here, or choose <b>Upload</b> .	20 10 10 20 Maxim	num iterations: 1560
	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 P         C       Delete       Actions <ul> <li>Create folder</li> <li>Upload</li> </ul> Q       Find objects by prefix	200 400 600 800 1000 1200 1400 1600 Valid Iteration Content of the second secon	ation ency: 50 iterations r Information ware resource: Single GPU ning rate schedule: Piecewise ning rate: 9e-05
	+	Imanified a WIOCEL OIT LITE GIOUC	- □ × ☆ ★ ® ::
Applications 📣 MATLAB R2020	b 💦 Experiment Manager 🕺 [matlab] 🕺 🦹 [matlab@130e335768e		
	Experiment Manager	Semantic Segmentation	
XPERIMENT MANAGER		This example shows an application that performs semantic segmentation on images stored in cloud	
G Open →		You run this example by entering the url to the Azure Blob that contains images	
Tayout Cayout Cayout Pa	rallel Canceled Plot Matrix -	Azure Blob url	Bicyclist
XPERIMENT BROWSER	periment1 × Experiment1   Result1 ×	'wasbs://camvidblob@semar	Pedestrian
TrainNetworkProject1	xhaustive Sweep Result	Azure access keys	Car
Result1 Ex	periment1 tew Experiment Source) fform hyperparmeter tuning with different se networks such as resnet18, snet50,mobilenetv2 Start: 11/19/2020, 12:16:18 AM O Complete 3 ▲ Stopped 0 ● Error O Running 0 ☱ Queued 0 × Cancele	<ul> <li><sup>A</sup> [?sv=2019-10-10&amp;st=2020-1]</li> <li><sup>A</sup> Perform Semantic Segmentation</li> <li>- Segmentation Complete</li> </ul>	Fence SignSymbol
Trial	I Status Progress Elapsed Time NetworkName Training Accurac Training Loss Validation Accura Validation Accura Validation Accura Validation Accura		r d Verneric
2	Complete         100.01 m 5 mm 22 set resnetto         30.7734         0.2223         90.9133           Complete         100.02 l hr 5 mm 27 set resnet50         93.3542         0.1810         92.1968		Road
3	Complete 100.0% 1 hr 39 min 45 sec mobilenetv2 87.1009 0.3602 88.3391		Pole

#### **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 – <u>Semantic Segmentation</u>

Classifying every pixel in an image/video

- Training Data 557MB
- Training Time ~ 5hrs
  - NVIDIA<sup>™</sup> Titan X with 12 GB memory



Road Segmentation for Autonomous Driving



#### Tracking Deforestation using Multispectral Images



#### **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?



#### AI System Design Workflow

#### **Data Preparation**

Human insight

Simulation-

Data cleansing and ululu preparation

0



Model design and tuning

**AI Modeling** 



Integration with complex systems

**Simulation & Test** 



 $-\mathbf{x}$  System verification  $-\checkmark$ and validation





Embedded devices



Enterprise systems

Edge, cloud,

desktop



Hardware accelerated training

Interoperability

8

#### AI System Design Workflow

#### **Data Preparation**



0

Data cleansing and preparation

Human insight

generated data

Simulation-



Model design and tuning

accelerated training

Interoperability

Hardware

**AI Modeling** 



Integration with complex systems

**Simulation & Test** 



→ System verification→ and validation

#### Deployment



Embedded devices



Enterprise systems



Edge, cloud, desktop

#### AI System Design Workflow

#### **Data Preparation**

Image: Image:





**AI Model Design** 



**AI Model Tuning** 

Model tuning

### 

Embedded devices

**Deployment** 



Human insight

Hardware accelerated training



Simulationgenerated data



Hardware accelerated training



En

Enterprise systems



Edge, cloud, desktop

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



#### 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

#### 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 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** 



Store preprocessed data anywhere

aws Services <b>v</b>	AWIAMPowerUser/kmahadev@mathworks.com@aws-matlab-cloud-imag ▼ Global ▼ Suppopulation
Amazon S3 ×	Amazon S3 > aiseminar-camvid
Buckets Access points	aiseminar-camvid
Batch Operations Access analyzer for S3	Bucket overview
Account settings for Block Public Access	Region     Amazon resource name (ARN)     Creation date     Access       US East (N. Virginia) us-east-1     Image: arr:aws:s3:::aiseminar-camvid     November 22, 2020, 01:14 (UTC-05:00)     Objects can be public
<ul> <li>Storage Lens</li> <li>Dashboards</li> <li>AWS Organizations settings</li> </ul>	Objects         Properties         Permissions         Metrics         Management         Access points
Feature spotlight 2	Drag and drop files and folders you want to upload here, or choose <b>Upload</b> .
	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 T         C       Delete         Actions       Create folder         Upload         Q. Find objects by prefix
	Name       ▲       Type       ▼       Last modified       ▼       Size       ▼       Storage class       ▼
	E images/ Folder
	🗆 🗅 labels/ Folder

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

# Data Preparation

#### AI Model Design

Build models anywhere

#### **AI Model Tuning**



#### Compute on demand



Run models anywhere

#### Easy Access to HPC resources

It's a balance between speed and cost

#### • Which VM did you choose?



mazon EC2 Overview Featu	res Pricing Instan	ce Types 👻	FAQs Ge	tting Started	Resources +				
General Purpose Compute Optimized	Accelerated Accelerated compicalculations, graph	d Compu- uting instances nics processing,	uting use hardwar or data patt	e accelerators, c ern matching, n	or co-processo nore efficient!	irs, to perform y than is poss	n functions, s ible in softw	such as floating are running on t	point number CPUs.
Memory Optimized	P4 P3	P2 Inf1	G4	G3 F1					
Accelerated Computing Storage Optimized Instance Features Measuring Instance Performance	Amazon EC2 P 100 Gbps of no Features: Up to 8 NVI High freque High freque Supports N Provides up EFA support	5 instances delin tworking throu DIA Tesla V100 incy Intel Xeon incy 2.5 GHz (bi VLink for peer- to 100 Gbps of t on p3dn.24xla	ver high perf ghput for m GPUs, each ES-2686 v4 ase) Intel Xec to-peer GPU f aggregate i rige instance	ormance compi achine learning pairing 5,120 C Broadwell) pro- in 8175M proce communication setwork bandwi s	ute in the clou and HPC appl UDA Cores an cessors for p3 issors for p3di idth.	ud with up to lications. d 640 Tensor .2xlarge, p3.8	8 NVIDIA® V Cores xlarge, and p	100 Tensor Core p3.16xlarge. Dedicated	GPUs and up t
	Instance	GPUs	VCPU	Mem (GiB)	GPU Mem (GiB)	GPU P2P	Storage (GB)	EBS Bandwidth	Networking Performance
	p3.2xlarge	1	8	61	16	21	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.24xlarg	r 8	96	768	256	NVLink	2 x 900 NVMe	19 Gbps	100 Gigabit

#### Easy Access to GPU resources

**Options chosen for setting up MATLAB** 

#### • Virtual Machine on AWS:



Easy access to a GPU

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
 C ▲ Not secure | ec2-35-153-182-187.compute-1.amazonaws.com:6080

a61b1acdf30	)7:1 (matlab) - noVN∈ 🗙	+		
$\leftrightarrow \   \ni \   {\tt G}$	A Not secure   ec2-	-35-153-182-187.co	ompute-1.amazonaws.com:60	080
🥙 Applications				
MATLAB R2020b				

-172-31-33-124:~\$ docker pull nvcr.io/partners/matlab:r2020b

2020b: Pulling from partners/matlab

#### Cloud Setup for Deep Learning in MATLAB

**Steps to using the Deep Learning Container** 



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



#### 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

• • •				Experiment Manager					
EXPERIMENT MANAGER								i i i	?
New Duplicate FILE ENVIRONMENT RUN	p Train Plo REV	ing Confusion t Matrix •	Filter Export						
EXPERIMENT BROWSER	Baseline	Tuning × Ba	seline Tuning   Result1 ×						0
DigitsClassifier     A Baseline Establishment	▼ Resul	t Details							
Sweep Initial Learning Rate     Baseline run     Baseline Tuning     Result1 (Running)     Icroer Initial Learning Rate Range	Baselir ( <u>View E</u>	e Tuning ixperiment Sour	2/7/20 DB)	20, 12:53:36 PM	Complete 7 Running 1	<b>▲</b> 5 ≌ (	Stopped 0 Queued 8	9 Erro X Can	7/16 Trials r 0 celed 0
Sweep Learning Rate Conv Size and     Add Conv-Batch-Rel µ Banks	Trial	Status	Progress	Elapsed Time	myInitialLearn	convFilterSize	Training Accu	Training Loss	Validation Ac
Wary Filter Size of First Conv2D Layer	1	Complete	100.09	0 hr 0 min 16 sec	1.0000e-6	3.0000	12.5000	2.6441	10
Train Validation Split Study	2	Complete	100.09	0 hr 0 min 15 sec	1.0000e-5	3.0000	25.7813	2.1228	20
	3	Complete	100.09	0 hr 0 min 14 sec	0.0001	3.0000	64.8438	1.0878	42
	4	Complete	100.09	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	100.09	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	C Running	30.7%	0 hr 0 min 4 sec	0.0005	4.0000			
	9	늘 Queued	0.0%		1.0000e-6	5.0000			
	10	ڬ Queued	0.0%		1.0000e-5	5.0000			
	11	🗄 Queued	0.0%		0.0001	5.0000			
	12	E Queued	0.0%		0.0005	5.0000			
	13	E Queued	0.0%		1.0000e-6	6.0000			
	14	🗎 Queued	0.0%		1.0000e-5	6.0000			
	15	ڬ Queued	0.0%		0.0001	6.0000			
	16	E 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



#### **AI Model Tuning**



On demand access to a GPU cluster

#### Parallel Multi-GPU Training in the cloud

What options are available for training at scale?

• 4 node GPU cluster chosen



On demand access to a GPU cluster



#### 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

	📣 Create Cluster - Cloud Center 🛛 🗙	+			-	
	← → C 🔒 cloudcenter.mathy	orks.com/cluster/create			☆ <b>*</b>	
Al Model Tuning	Preferences User Preferences Global Cluster Access	* Give this cluster a name MATLAB Version	GPUParallelCluster R2020b	~		
		Automatically terminate cluster	When cluster is idle	~		
		Cluster Log Level   Location & Network   Cluster Configuration Shared State Auto-Manage Cluster Access Worker Machine Type Workers per Machine Use a dedicated headnode	Low     Personal Cluster     Shareable Cluste     Set cluster firewall rules to let MATLAB comm     Double Precision GPU (p3.2xlarge, 4 core, 1     Some features are available only with a dedicated and the Type     Standard (m5.xlarge, 2 core)	a61b1acdf307.1 (matlab) - noV/ × + ← → C A Not secure ec2-35-15 Applications x matlab HOME PLOTS APPS HOME PLOTS APPS HOME PLOTS APPS File File Comp File File Comp File File Comp File Comp	Sa-182-187.compute-1.amazonaws.com:6080/vnc.html?password=matlab&autoconnect=true MATLAB R2020b MATLAB R2020b Discover Clusters Discover Clusters The following clusters were found. Select a cluster to use: Cluster Name Host Workers Type Releas CPUIParalle.uster amazon EC2 A MATLAB R2020b	
On demand access to a GPU cluster		Workers in Cluster Machines in Cluster ~ Cluster Shared Storage	Allow cluster to auto-resize Initial Count Upper Limit I I Upper Limit I I Upper Limit I I Upper Limit I I I Upper Limit I I I I I I I I I I I I I I I I I I I	☐ MATLAB.desktop  Details  Workspace Name ∠ Value  IIII- Ready		Devt Cancel

MathWorks

#### 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?



#### AI Model Design



Build models anywhere

#### **AI Model Tuning**



#### Compute on demand



#### 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



- Each model request calls a "hot" runtime
- Scalable
  - Suitable for single calls or batch workflows



**Deployment** 

What options are available for deploying to production?

#### MATLAB Production Server from Azure Market Place



Microsoft	Azure Marketpl	ace Apps 🗸	Search Marketplace				Q			More 🗸
Products >	MATLAB Production Se	rver (PAYG)								
	Mat Mat	ATLAB Proc hWorks * * * (0) Write a review rview Plans	duction Server ( Reviews	ΡΑΥ(	G) 🗢	Save to n	ny list			
GET I	MAT NOW Serv	LAB Production Server	ver is an application tics	Overview	Applications	Sotlings	Persistence	Application Access Control	Dashboard Access Control	Loga
Categories Analytics AI + Machine L Compute Support Support Help Legal License Agreer Privacy Policy	MATL earning integr produ server MATL then or recodent enable simult	AB Production Server <sup>®</sup> is an ating MATLA8® analytics is inction enterprise application s or in Microsoft Azure. You AB, package them using M Japloy them to MATLAB Pre- ing or creating custom infra infrastructure combined wi es your application to supp- taneously.	n application server for to web, database, and is running on dedicated i can create algorithms in ATLAB Compiler SDK <sup>M</sup> , and duction Server without structure. The elasticity of th MATLAB Production Server ort many users		MATLAB F MATLAB F Ser	TLAB Executs MATLAB End Deshib Production Se NATLAB Runti Ver VM Opera Number of Lest F	on Endpoint point Status pard Version me Versions ting System t Server VMs befresh Time Refresh	https://mps14220/dtbsy.east READY 1.0.0 (R2020a (R2020a, R2016b, R2016b, R Linux 1 2.58:12 PM	an oloudapp azure.com 120160, R20164, R20170]	
	This s let yo availa provic	olution template for Azure i u quickly create, configure, ble MATLAB Production Ser des ARM templates that use	ncorporates best practices to and deploy a scalable, highly- ver environment. MathWorks preconfigured virtual		(a Be					

**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

**Steps to follow** 



#### What options are available?



#### Deployment



Run models on demand

#### Summary – What was shown?



#### 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!"



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:
  - o Assign IAM role to your instance, or
  - Use <u>AWS SAML</u> 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 <u>'Upload data to s3'</u>
- Verify data in S3 bucket.
- Refer to <u>'UploadDataToS3.txt'</u> file for an example showcasing use of AWS SAML

aws Services ▼	🐥 MMMM9SumeClam/Annshades@mathaoAs.com @ avec metable cloud imag 💌 Global 💌 Support 💌							
Amazon S3 ×	Amazon 53 > aiseminar-carnid							
Buckets Access points	aiseminar-camvid							
Batch Operations	Bucket overview							
Account settings for block Public Access	Region Amazon resource name (JAR0) Creation date Acress Ur5 East (V. Veginis) as-east-1 <b>Ö</b> anzaess3:zaitemisma-cannol November 22, 2005, 01:14 (JJTC-0500) Objects can be public							
Storage Lens Dashboards AWS Organizations settings	Objects Properties Permissions Metrics Management Access points							
Fortune contribute	Drag and drop files and folders you want to upload here, or choose Upload.							
	Objects (2)         Open set the function of a versue s3. For others to access your elaptic, you't need to explicitly your them permissions, Lawrenews (2)         O       Deletes       Accience **         O       Deletes       Accience **         O       Deletes       Accience **         O       Profestions **       Create Reduint**         Output       Output       Output         Output       Output       Output							
	Name         λ         Type         ψ         Last modified         ψ         Size         ψ         Storage class         ψ							
	Dimages/ Folder -							
	D block/ Folder							

#### Thank you



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