A CLOUD GURU

SageMaker Training

Scott Pletcher
INSTRUCTOR

Why GPUs?

Machine Learning = Mathematics

Computers are good with math.

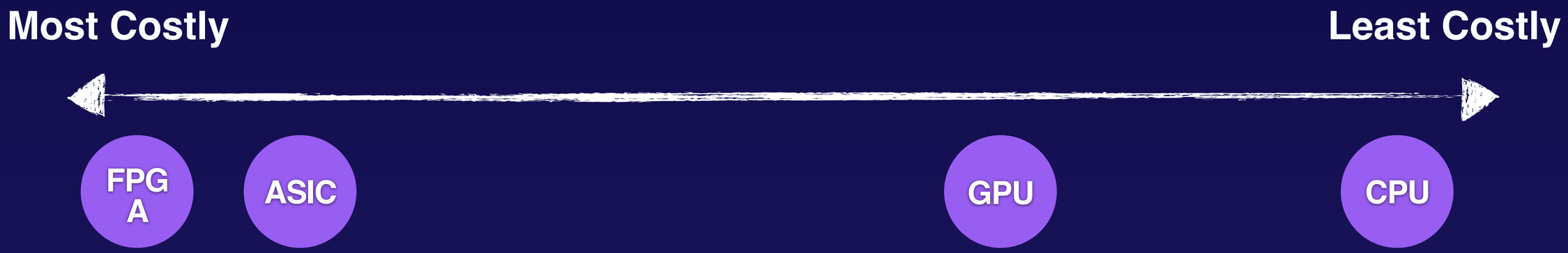
Graphical Processing Units are optimized for math.

GPUs are not the most efficient math engine.

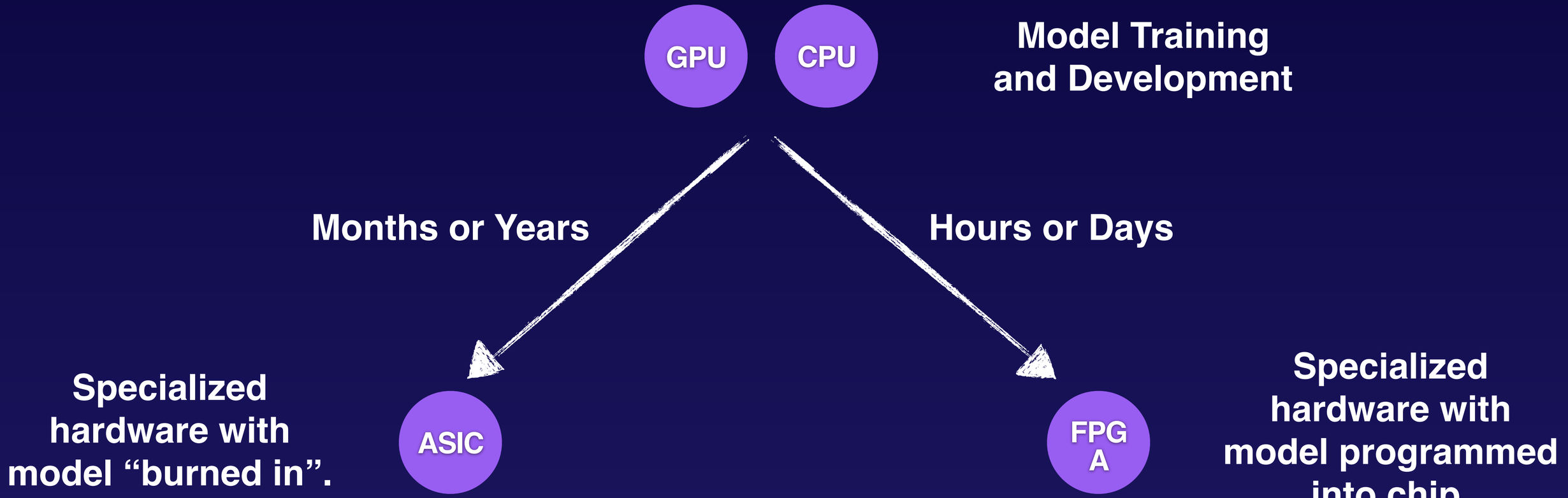
Why GPUs?

Why GPUs?

Why GPUs?



Why GPUs?


GPU
CPU
Model Training and Development
Months or Years
Hours or Days
Specialized hardware with model “burned in”.
ASIC
FPGA
Specialized hardware with model programmed into chip.

51% Attack

Group of miners controlling more than 51% of the hash rate.

- Prevent new transactions
- Reverse completed transactions

Outsmart the Machines

ASIC — 12–18 month Time to Market

FPGA — Ultra-specialized knowledge

GPU CPU — Download new software

Training

High-Level Python library provided by Amazon SageMaker

Use the SDK for Python

CreateTrainingJob API

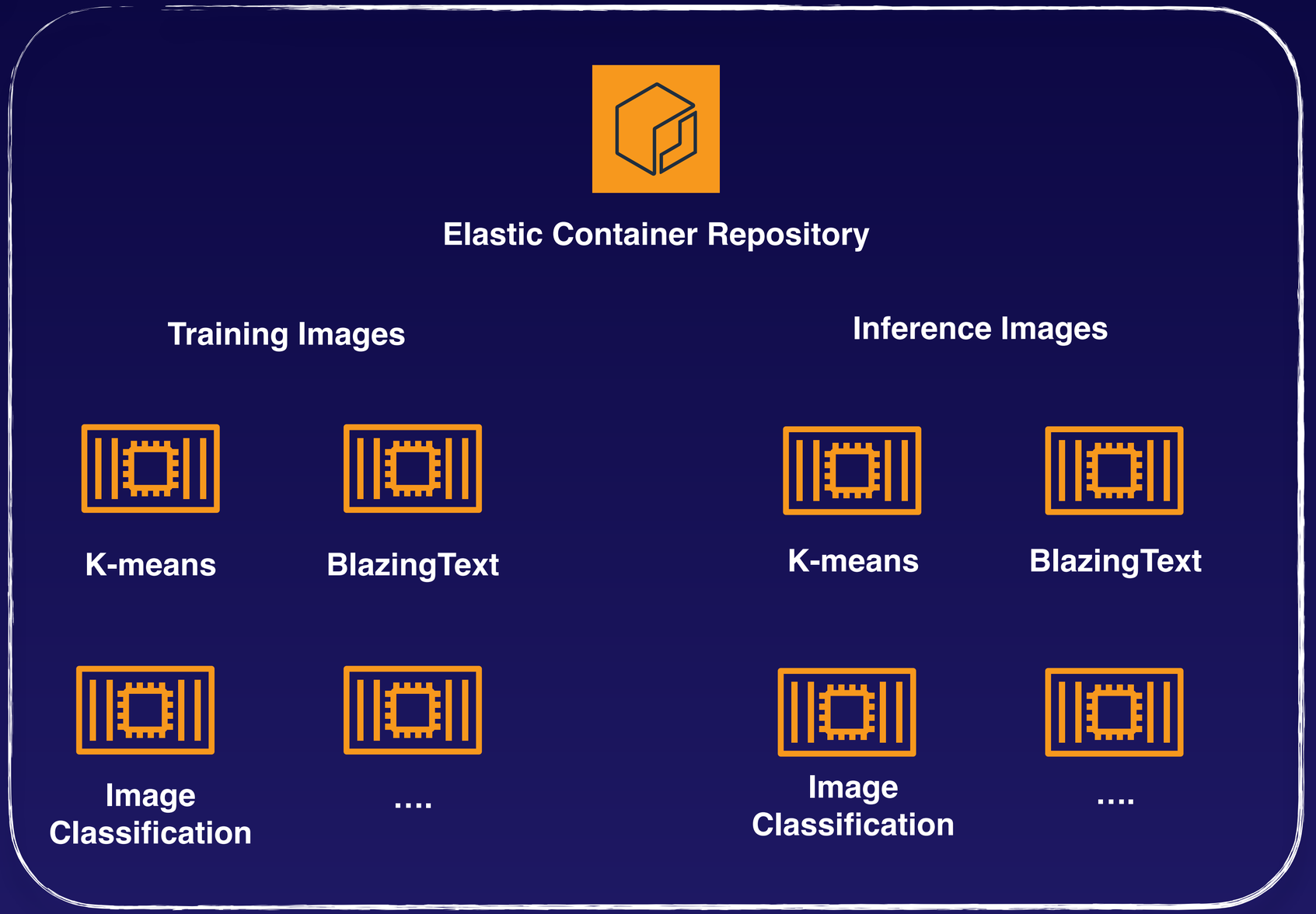
Training

High-Level Python Library

```
from sagemaker import KMeans

data_location = 's3://{}/kmeans_highlevel_example/data'.format(bucket)
output_location = 's3://{}/kmeans_highlevel_example/output'.format(bucket)


kmeans = KMeans(role = role,
                train_instance_count = 2,
                train_instance_type = 'ml.c4.8xlarge',
                output_path = output_location,
                k = 10,
                data_location = data_location)

kmeans.fit(kmeans.record_set(train_set[0]))
```

From a Jupyter Notebook, you can issue commands to launch a training job.

Training

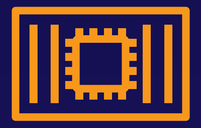

Elastic Container Repository
Training Images
Inference Images
K-means
BlazingText
K-means
BlazingText
Image Classification
....
Image Classification
....

Training

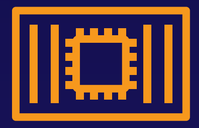
Elastic Container Repository

CreateTrainingJob API Call

Training Images



K-means



BlazingText

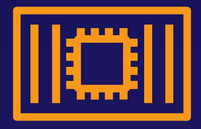
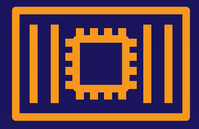


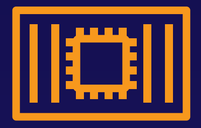
Image Classification



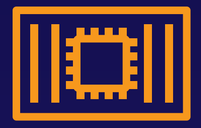
....

CreateModel API Call

Inference Images



K-means



BlazingText

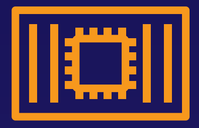
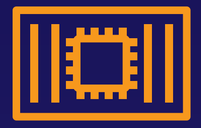


Image Classification



....

Training

Region	ECR Path
us-west-1	632365934929.dkr.ecr.us-west-1.amazonaws.com
us-west-2	174872318107.dkr.ecr.us-west-2.amazonaws.com
ap-south-1	991648021394.dkr.ecr.ap-south-1.amazonaws.com
…	…

Training

Named input source for algorithm consumption.

Algorithms Provided by Amazon SageMaker: Common Parameters

The following table lists parameters for each of the algorithms provided by Amazon SageMaker.

Algorithm Name	Channel Name	Training Image and Inference Image Registry Path	Training Input Mode	File Type	Instance Class
BlazingText	train	*<ecr_path>*/blazingtext:*<tag>*	File or Pipe	Text file (one sentence per line with space-separated tokens)	GPU (single instance only) or CPU
DeepAR Forecasting	train and (optionally) test	*<ecr_path>*/forecasting-deepar:*<tag>*	File	JSON Lines or Parquet	GPU or CPU
Factorization Machines	train and (optionally) test	*<ecr_path>*/factorization-machines:*<tag>*	File or Pipe	recordIO-protobuf	CPU (GPU for dense data)
Image Classification	train and validation, (optionally) train_lst, validation_lst, and model	*<ecr_path>*/image-classification:*<tag>*	File or Pipe	recordIO or image files (.jpg or .png)	GPU
IP Insights	train and (optionally) validation	*<ecr_path>*/ipinsights:*<tag>*	File	CSV	GPU or GPU
k-means	train and (optionally) test	*<ecr_path>*/kmeans:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPUCommon (single GPU device on one or more instances)
k-nearest-neighbor (k-NN)	train and (optionally) test	*<ecr_path>*/knn:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPU (single GPU device on one or more instances)

Training

Path to the ECR image for a given algorithm.

If using the SageMaker Python Library, it will automatically know most paths.

Algorithms Provided by Amazon SageMaker: Common Parameters

The following table lists parameters for each of the algorithms provided by Amazon SageMaker.

Algorithm Name	Channel Name	Training Image and Inference Image Registry Path	Training Input Mode	File Type	Instance Class
BlazingText	train	*<ecr_path>*/blazingtext:*<tag>*	File or Pipe	Text file (one sentence per line with space-separated tokens)	GPU (single instance only) or CPU
DeepAR Forecasting	train and (optionally) test	*<ecr_path>*/forecasting-deepar:*<tag>*	File	JSON Lines or Parquet	GPU or CPU
Factorization Machines	train and (optionally) test	*<ecr_path>*/factorization-machines:*<tag>*	File or Pipe	recordIO-protobuf	CPU (GPU for dense data)
Image Classification	train and validation, (optionally) train_lst, validation_lst, and model	*<ecr_path>*/image-classification:*<tag>*	File or Pipe	recordIO or image files (.jpg or .png)	GPU
IP Insights	train and (optionally) validation	*<ecr_path>*/ipinsights:*<tag>*	File	CSV	GPU or GPU
k-means	train and (optionally) test	*<ecr_path>*/kmeans:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPUCommon (single GPU device on one or more instances)
k-nearest-neighbor (k-NN)	train and (optionally) test	*<ecr_path>*/knn:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPU (single GPU device on one or more instances)

Training

Note the Tag attribute. It is a form of versioning for the repository.

Use :1 version tag to ensure the stable version.

Use :latest version tag for up-to-date but potentially not backward compatible.

Use :1 for production purposes.

Algorithms Provided by Amazon SageMaker: Common Parameters

The following table lists parameters for each of the algorithms provided by Amazon SageMaker.

Algorithm Name	Channel Name	Training Image and Inference Image Registry Path	Training Input Mode	File Type	Instance Class
BlazingText	train	*<ecr_path>*/blazingtext:*<tag>*	File or Pipe	Text file (one sentence per line with space-separated tokens)	GPU (single instance only) or CPU
DeepAR Forecasting	train and (optionally) test	*<ecr_path>*/forecasting-deepar:*<tag>*	File	JSON Lines or Parquet	GPU or CPU
Factorization Machines	train and (optionally) test	*<ecr_path>*/factorization-machines:*<tag>*	File or Pipe	recordIO-protobuf	CPU (GPU for dense data)
Image Classification	train and validation, (optionally) train_lst, validation_lst, and model	*<ecr_path>*/image-classification:*<tag>*	File or Pipe	recordIO or image files (.jpg or .png)	GPU
IP Insights	train and (optionally) validation	*<ecr_path>*/ipinsights:*<tag>*	File	CSV	GPU or GPU
k-means	train and (optionally) test	*<ecr_path>*/kmeans:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPUCommon (single GPU device on one or more instances)
k-nearest-neighbor (k-NN)	train and (optionally) test	*<ecr_path>*/knn:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPU (single GPU device on one or more instances)

Training

Type of input data accepted.

Recall that Pipe has better performance.

Algorithms Provided by Amazon SageMaker: Common Parameters

The following table lists parameters for each of the algorithms provided by Amazon SageMaker.

Algorithm Name	Channel Name	Training Image and Inference Image Registry Path	Training Input Mode	File Type	Instance Class
BlazingText	train	*<ecr_path>*/blazingtext:*<tag>*	File or Pipe	Text file (one sentence per line with space-separated tokens)	GPU (single instance only) or CPU
DeepAR Forecasting	train and (optionally) test	*<ecr_path>*/forecasting-deepar:*<tag>*	File	JSON Lines or Parquet	GPU or CPU
Factorization Machines	train and (optionally) test	*<ecr_path>*/factorization-machines:*<tag>*	File or Pipe	recordIO-protobuf	CPU (GPU for dense data)
Image Classification	train and validation, (optionally) train_lst, validation_lst, and model	*<ecr_path>*/image-classification:*<tag>*	File or Pipe	recordIO or image files (.jpg or .png)	GPU
IP Insights	train and (optionally) validation	*<ecr_path>*/ipinsights:*<tag>*	File	CSV	GPU or GPU
k-means	train and (optionally) test	*<ecr_path>*/kmeans:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPUCommon (single GPU device on one or more instances)
k-nearest-neighbor (k-NN)	train and (optionally) test	*<ecr_path>*/knn:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPU (single GPU device on one or more instances)

Training

Format of data accepted by algorithm.

Recall that recordIO protobuf has best performance.

Algorithms Provided by Amazon SageMaker: Common Parameters

The following table lists parameters for each of the algorithms provided by Amazon SageMaker.

Algorithm Name	Channel Name	Training Image and Inference Image Registry Path	Training Input Mode	File Type	Instance Class
BlazingText	train	*<ecr_path>*/blazingtext:*<tag>*	File or Pipe	Text file (one sentence per line with space-separated tokens)	GPU (single instance only) or CPU
DeepAR Forecasting	train and (optionally) test	*<ecr_path>*/forecasting-deepar:*<tag>*	File	JSON Lines or Parquet	GPU or CPU
Factorization Machines	train and (optionally) test	*<ecr_path>*/factorization-machines:*<tag>*	File or Pipe	recordIO-protobuf	CPU (GPU for dense data)
Image Classification	train and validation, (optionally) train_lst, validation_lst, and model	*<ecr_path>*/image-classification:*<tag>*	File or Pipe	recordIO or image files (.jpg or .png)	GPU
IP Insights	train and (optionally) validation	*<ecr_path>*/ipinsights:*<tag>*	File	CSV	GPU or GPU
k-means	train and (optionally) test	*<ecr_path>*/kmeans:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPUCommon (single GPU device on one or more instances)
k-nearest-neighbor (k-NN)	train and (optionally) test	*<ecr_path>*/knn:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPU (single GPU device on one or more instances)

Training

Type of instance required for algorithm.

Note that some are GPU only or CPU only.

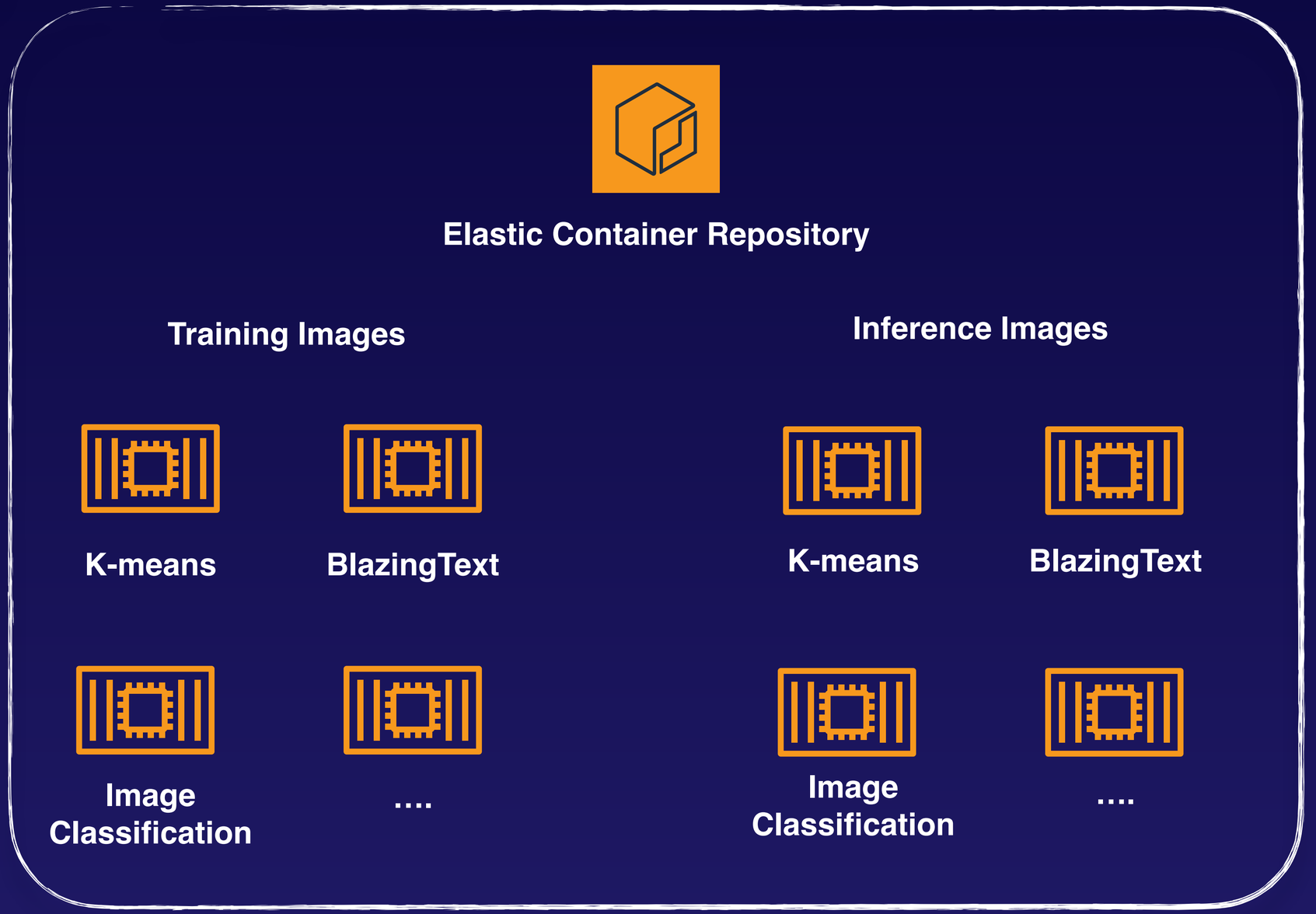
Example:
XGBoost implements an open source algorithm optimized for CPUs

Algorithms Provided by Amazon SageMaker: Common Parameters

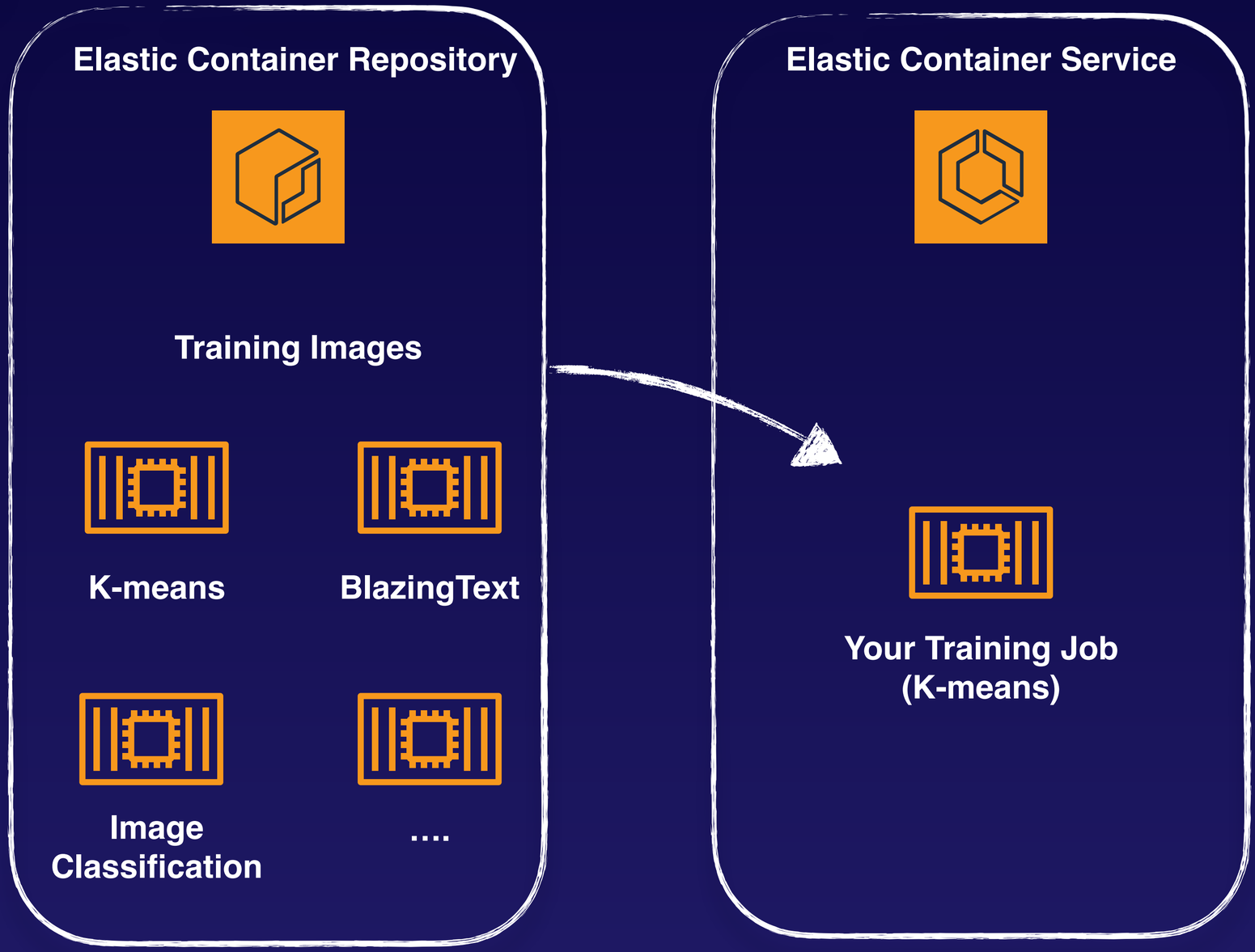
The following table lists parameters for each of the algorithms provided by Amazon SageMaker.

Algorithm Name	Channel Name	Training Image and Inference Image Registry Path	Training Input Mode	File Type	Instance Class
BlazingText	train	*<ecr_path>*/blazingtext:*<tag>*	File or Pipe	Text file (one sentence per line with space-separated tokens)	GPU (single instance only) or CPU
DeepAR Forecasting	train and (optionally) test	*<ecr_path>*/forecasting-deepar:*<tag>*	File	JSON Lines or Parquet	GPU or CPU
Factorization Machines	train and (optionally) test	*<ecr_path>*/factorization-machines:*<tag>*	File or Pipe	recordIO-protobuf	CPU (GPU for dense data)
Image Classification	train and validation, (optionally) train_lst, validation_lst, and model	*<ecr_path>*/image-classification:*<tag>*	File or Pipe	recordIO or image files (.jpg or .png)	GPU
IP Insights	train and (optionally) validation	*<ecr_path>*/ipinsights:*<tag>*	File	CSV	GPU or GPU
k-means	train and (optionally) test	*<ecr_path>*/kmeans:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPUCommon (single GPU device on one or more instances)
k-nearest-neighbor (k-NN)	train and (optionally) test	*<ecr_path>*/knn:*<tag>*	File or Pipe	recordIO-protobuf or CSV	CPU or GPU (single GPU device on one or more instances)

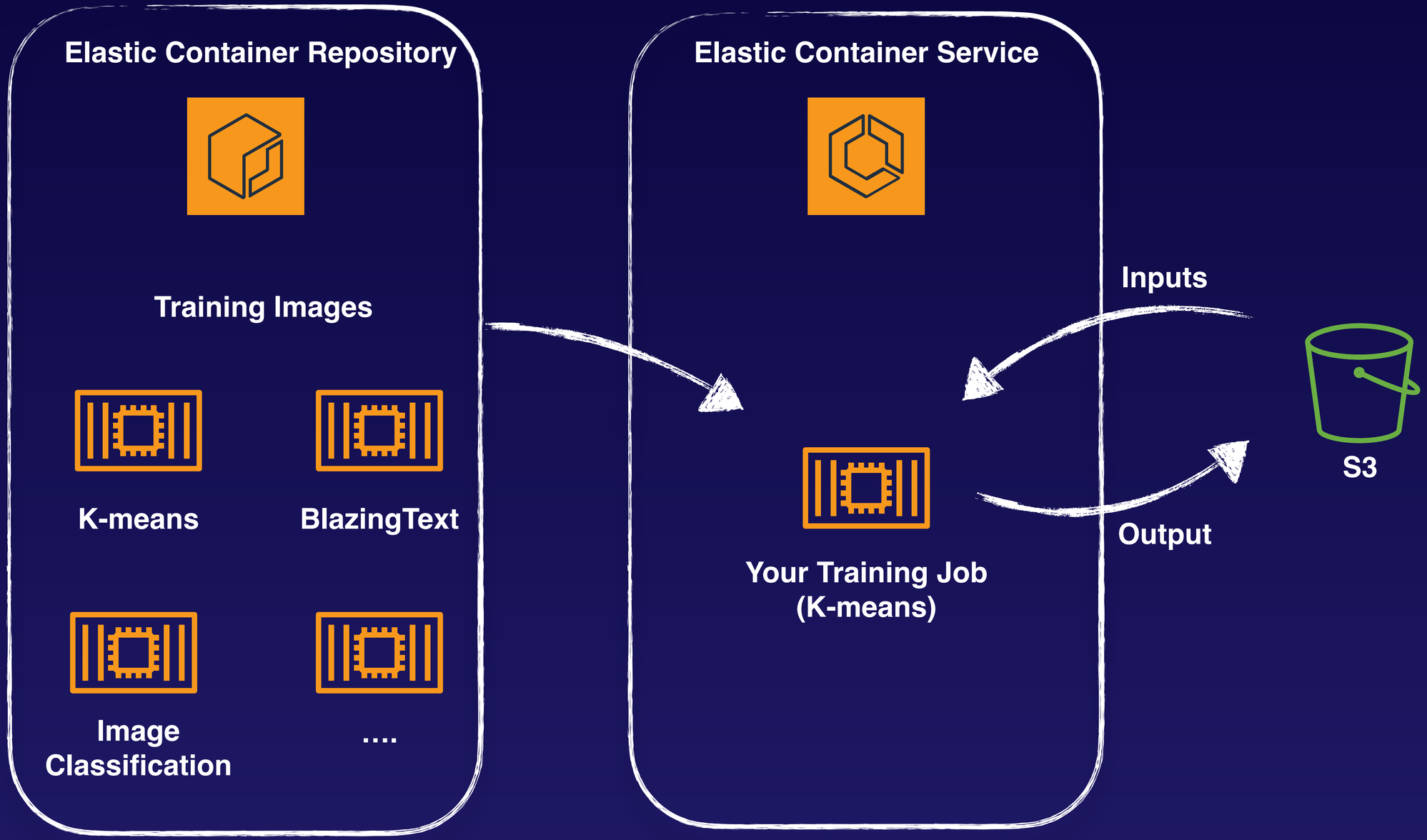
Training


Elastic Container Repository
Training Images
Inference Images
K-means
BlazingText
K-means
BlazingText
Image Classification
....
Image Classification
....

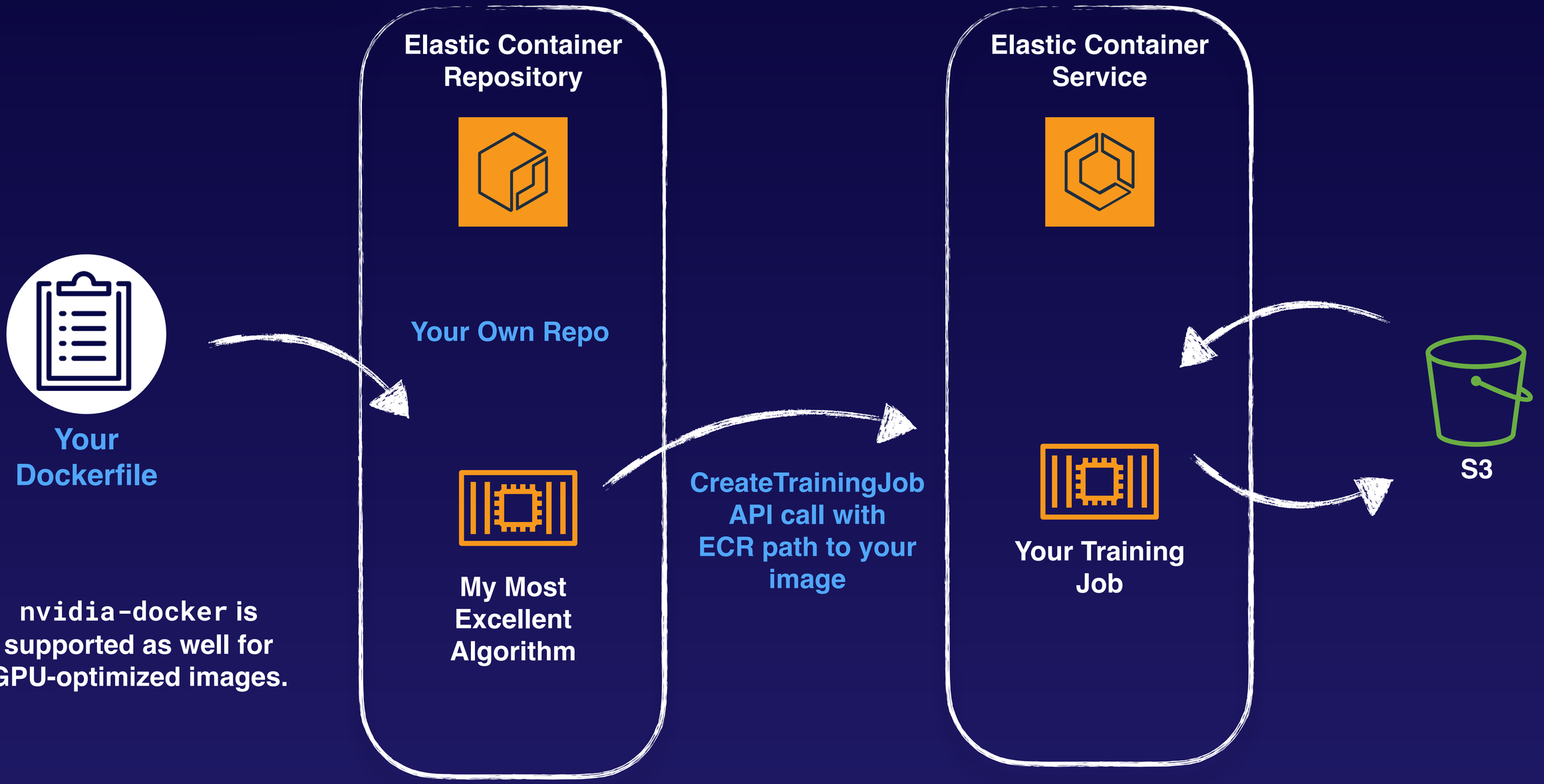
Training


Elastic Container Repository
Training Images
K-means
BlazingText
Image Classification
....
Elastic Container Service
Your Training Job (K-means)

Training

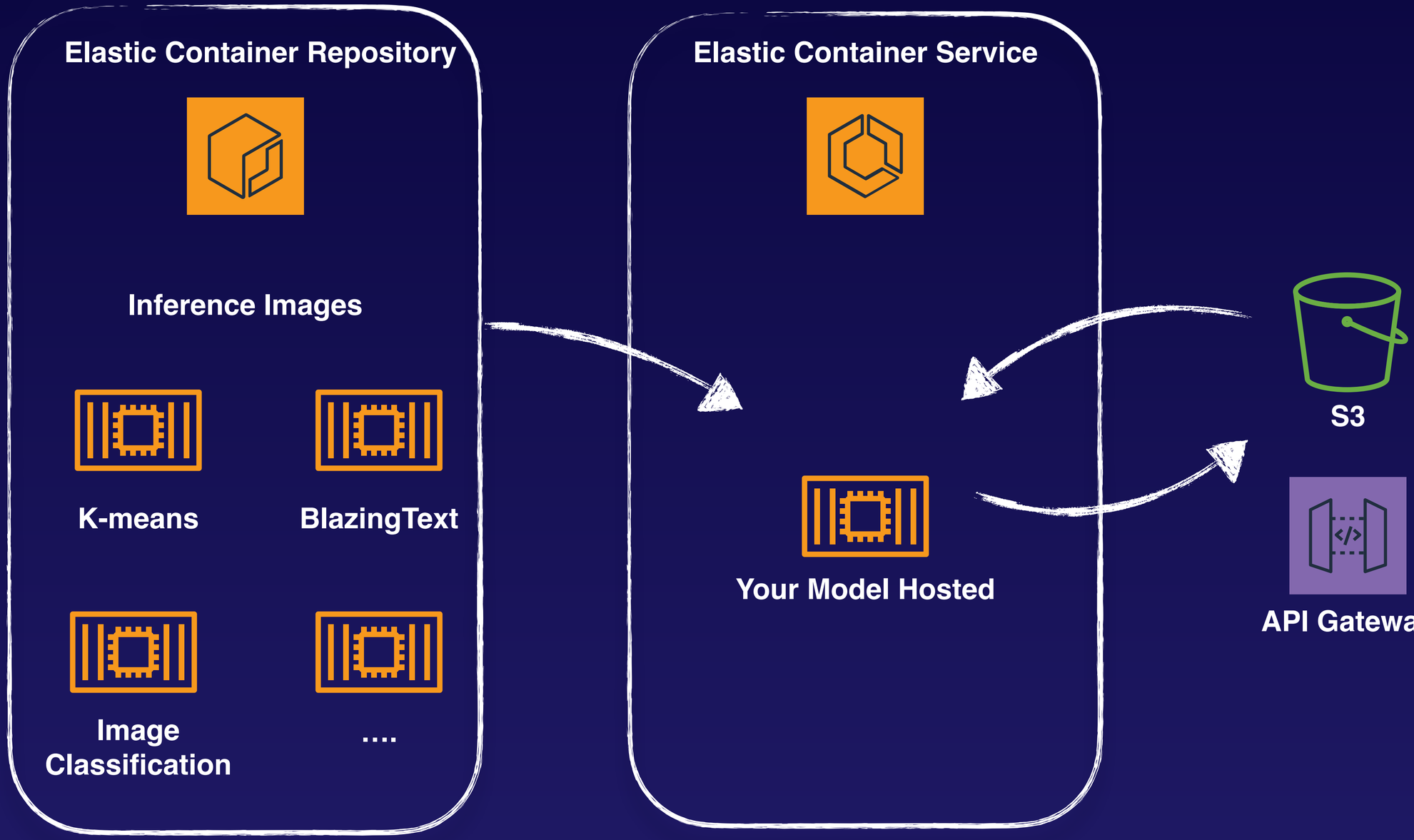

Elastic Container Repository
Training Images
K-means
BlazingText
Image Classification
....
Elastic Container Service
Your Training Job (K-means)
Inputs
Output
S3

Custom Algorithms



`nvidia-docker` is supported as well for GPU-optimized images.

Inference


Elastic Container Repository
Inference Images
K-means
BlazingText
Image Classification
....
Elastic Container Service
Your Model Hosted
S3
API Gateway

Training Process

Example of a Binary Classification Process

0	Luke	Rebel
1	JarJar	Empire
0	Han	Rebel
1	Vadar	Empire
0	Yoda	Rebel

Training Process

Example of a Binary Classification Process

0	Luke	Rebel	Does “Luke” = Evil?	N
1	JarJar	Empire	Does “JarJar” = Evil?	Y
0	Han	Rebel	Does “Han” = Evil?	N
1	Vadar	Empire	Does “Vadar” = Evil?	Y
0	Yoda	Rebel	Does “Yoda” = Evil?	N

Training Process

Example of a Binary Classification Process

0	Luke	Rebel
1	JarJar	Empire
0	Han	Rebel
1	Vadar	Empire
0	Yoda	Rebel

Does “Luke” = Evil?	N
Does “JarJar” = Evil?	Y
Does “Han” = Evil?	N
Does “Vadar” = Evil?	Y
Does “Yoda” = Evil?	N

Weak Correlation

Example of a Binary Classification Process

0	Luke	Rebel
1	JarJar	Empire
0	Han	Rebel
1	Vadar	Empire
0	Yoda	Rebel

Does “Rebel” = Evil?	N
Does “Empire” = Evil?	Y
Does “Rebel” = Evil?	N
Does “Empire” = Evil?	Y
Does “Rebel” = Evil?	N

Training Process

Example of a Binary Classification Process

0	Luke	Rebel
1	JarJar	Empire
0	Han	Rebel
1	Vadar	Empire
0	Yoda	Rebel

Does “Rebel” = Evil?	N
Does “Empire” = Evil?	Y
Does “Rebel” = Evil?	N
Does “Empire” = Evil?	Y
Does “Rebel” = Evil?	N

Very Strong Correlation

Training Process

Example of a Binary Classification Process

0	Luke	Rebel
1	JarJar	Empire
0	Han	Rebel
1	Vadar	Empire
0	Yoda	Rebel

Does “Luke” & “Rebel” = Evil?	N
Does “JarJar” & “Empire” = Evil?	Y
Does “Han” & “Rebel” = Evil?	N
Does “Vadar” & “Empire” = Evil?	Y
Does “Yoda” & “Rebel” = Evil?	N

Just as good as Rebel/Empire…

Training Process

If parameter = “Empire”, we predict “Evil (1)”.

Testing Set

1	Jabba	Empire	Given “Empire”, I predict “Evil (1)”
1	Maul	Empire	Given “Empire”, I predict “Evil (1)”
0	Leia	Rebel	Given “Rebel”, I predict “Not Evil (0)”

100% Precision! Ship it!

Training Process

CloudWatch

Information Logged

- Arguments provided
- Errors during training
- Algorithm accuracy statistics
- Timing of the algorithm

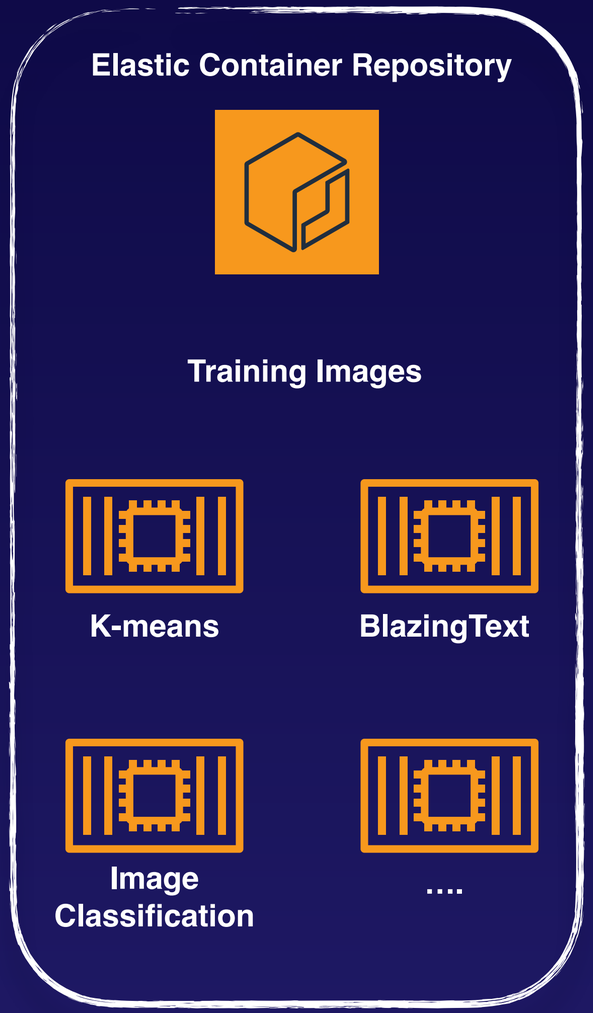
Training Process

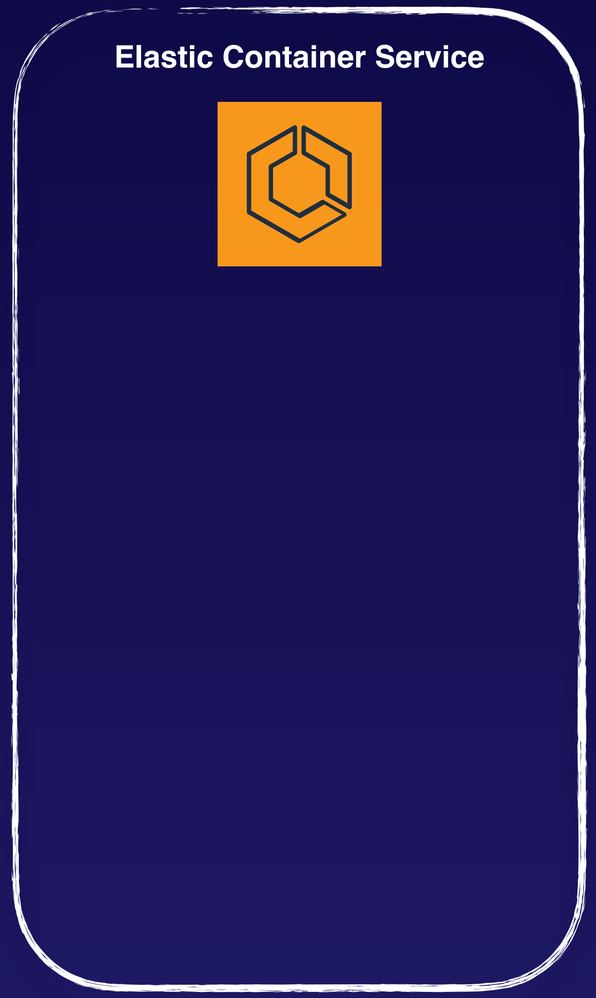
CloudWatch

Common Errors

- Error in specifying a hyperparameter such as an extra one.
- Invalid value for a hyperparameter
- Incorrect protobuf file format

Training with Apache Spark





Training with Apache Spark

Elastic Container Repository

Training Images

K-means

BlazingText

Image Classification

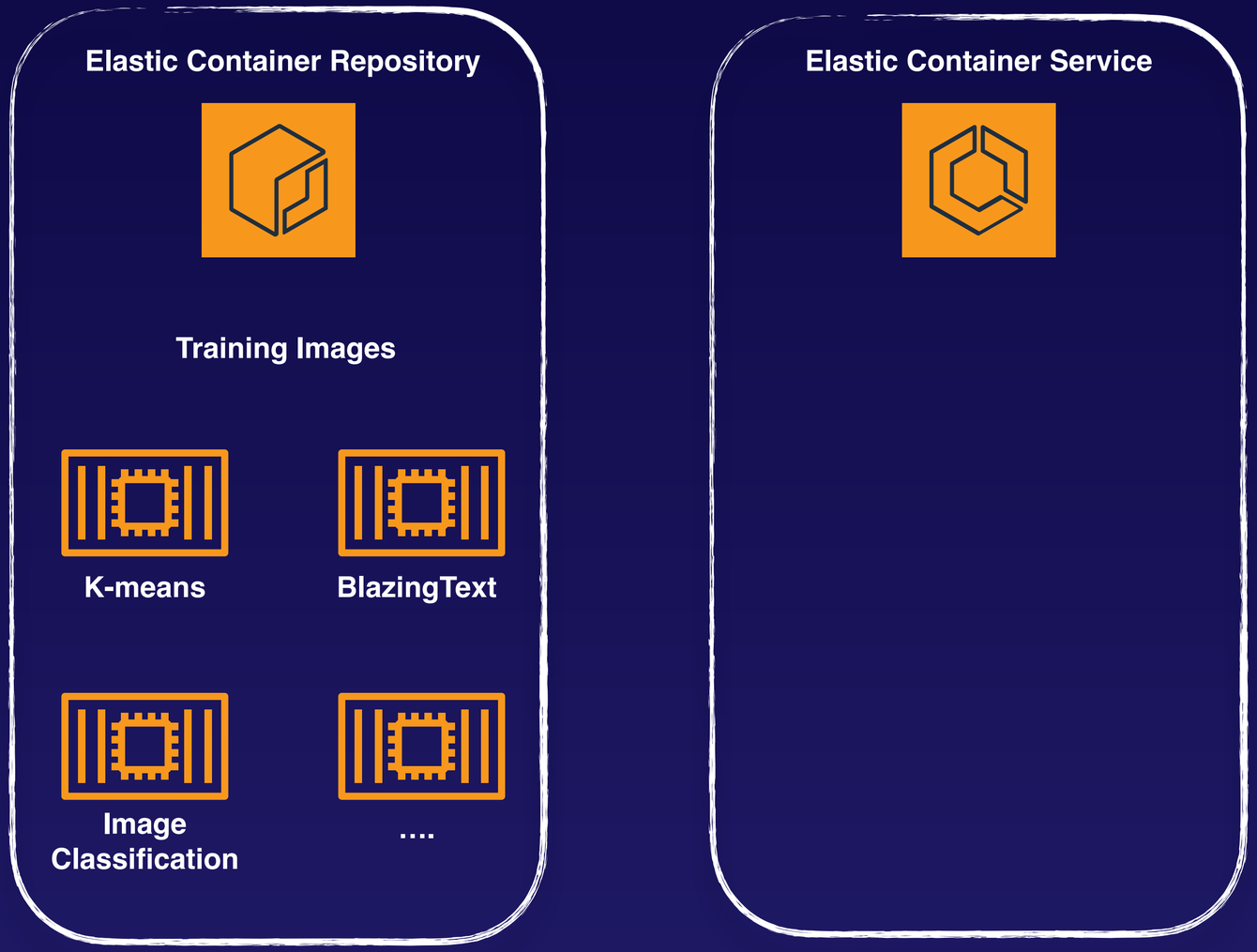
....

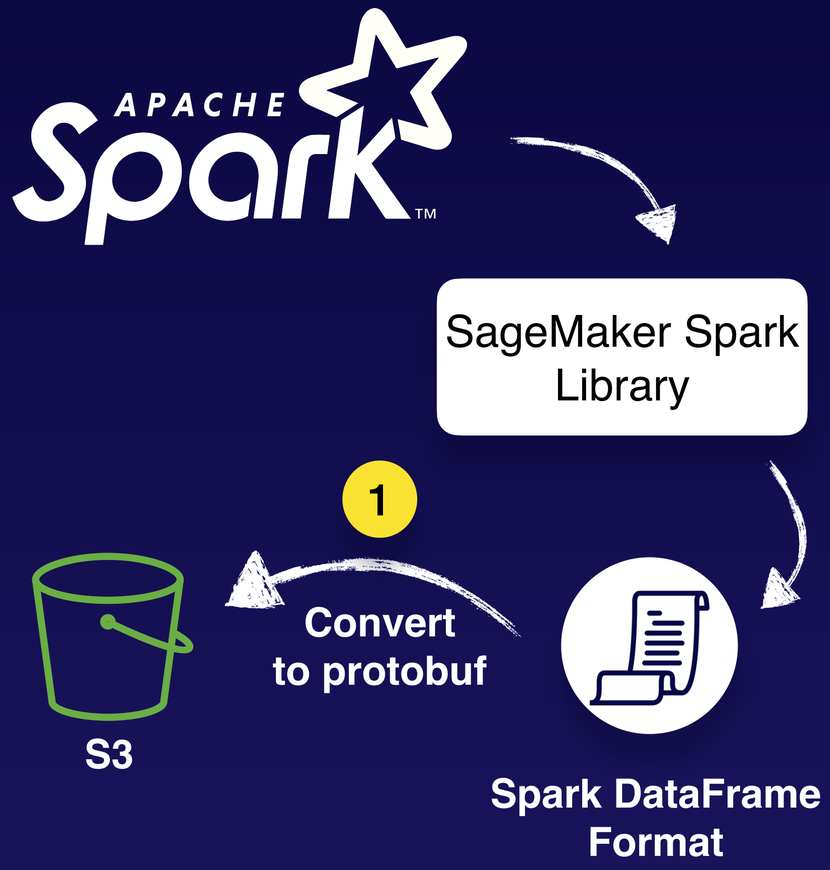
Elastic Container Service

APACHE Spark™

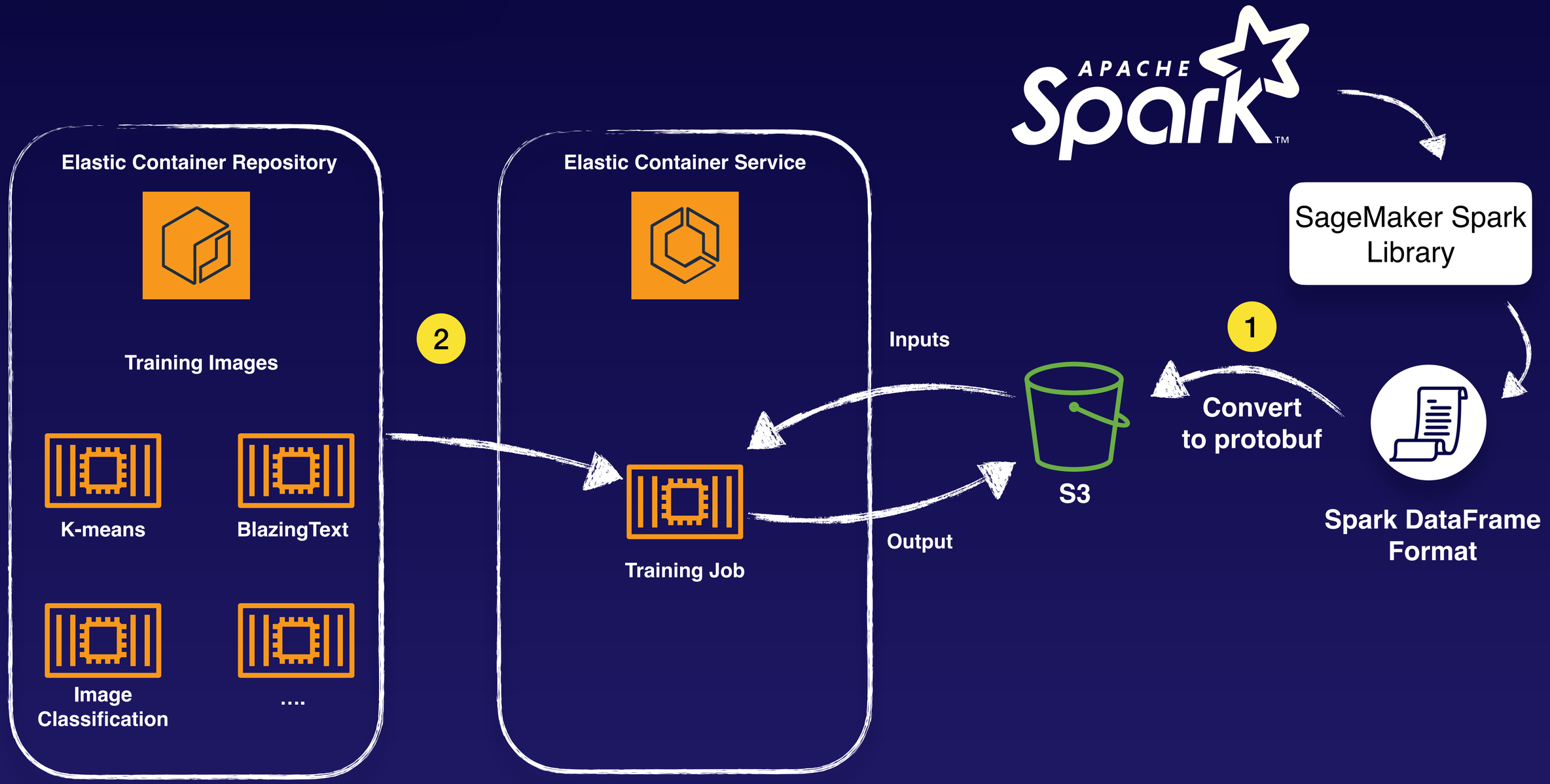
SageMaker Spark Library

Training with Apache Spark

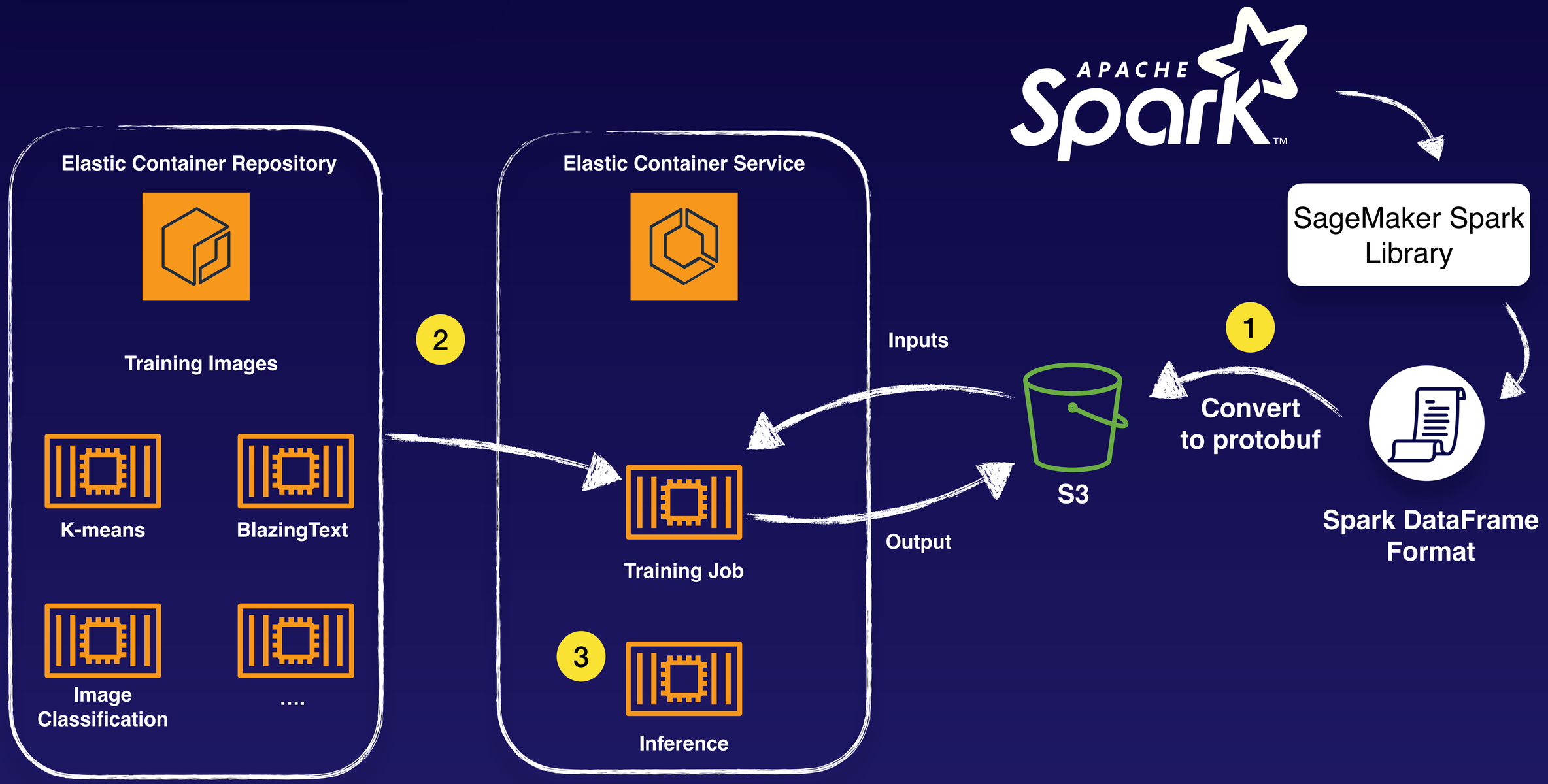




Training with Apache Spark



Training with Apache Spark



SAGEMAKER TRAINING

Training with Apache Spark

A CLOUD GURU