aws SUMMIT

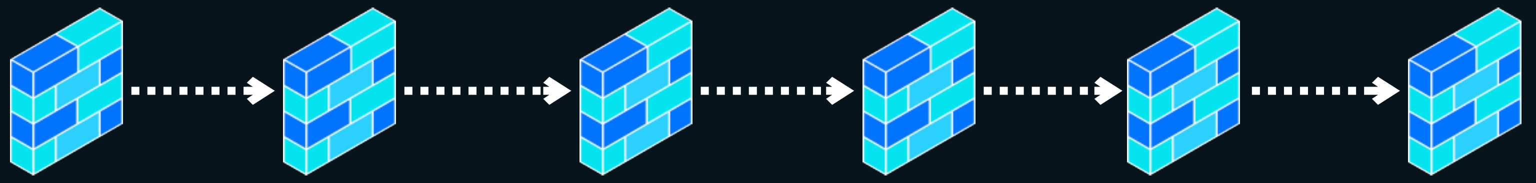
Build, train, and deploy machine learning models at scale

Julien Simon

Principal Technical Evangelist, AI and Machine Learning

@julsimon

ML is still too complicated for everyday developers



Collect and prepare training data

Choose and optimize your ML algorithm

Set up and manage environments for training

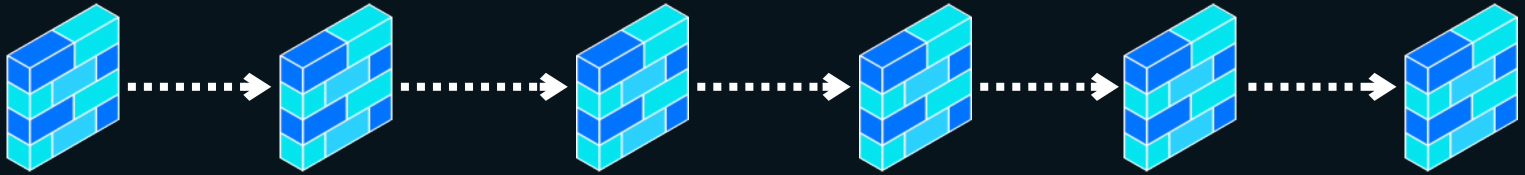
Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

Amazon SageMaker

Easily build, train, and deploy Machine Learning models



Collect and prepare training data

Choose and optimize your ML algorithm

Set up and manage environments for training

Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

Amazon SageMaker

Pre-built notebooks for common problems

Built-in, high-performance algorithms

ALGORITHMS

K-Means Clustering
Principal Component Analysis
Neural Topic Modelling
Factorization Machines
Linear Learner

XGBoost
Latent Dirichlet Allocation
Image Classification
Seq2Seq,
And more!

FRAMEWORKS

Apache MXNet
TensorFlow

Caffe2, CNTK,
PyTorch, Torch

Set up and manage environments for training

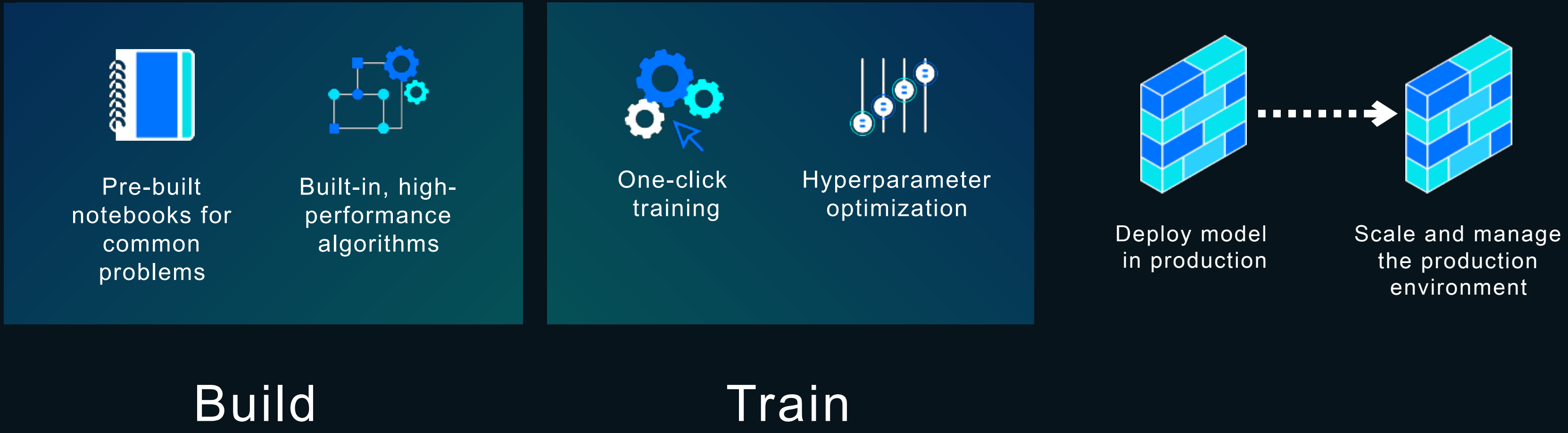
Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

Build

Amazon SageMaker



Pre-built notebooks for common problems

Built-in, high-performance algorithms

One-click training

Hyperparameter optimization

Deploy model in production

Scale and manage the production environment

Build

Train

Amazon SageMaker

Pre-built notebooks for common problems

Built-in, high-performance algorithms

Build

One-click training

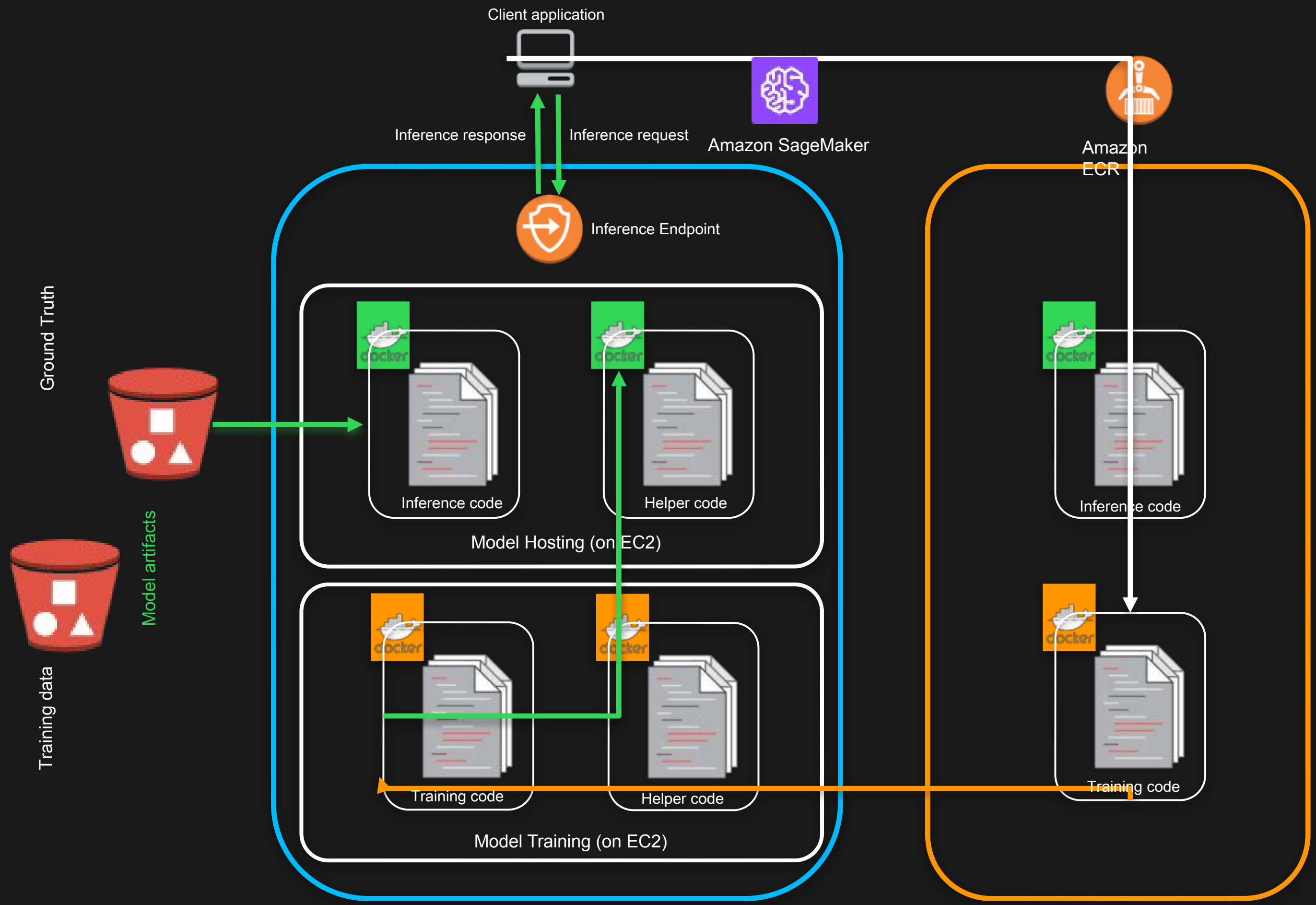
Hyperparameter optimization

Train

One-click deployment

Fully managed hosting with auto-scaling

Deploy



Client application
Inference response
Inference request
Amazon SageMaker
Amazon ECR
Inference Endpoint
Ground Truth
Model artifacts
Training data
Inference code
Helper code
Model Hosting (on EC2)
Training code
Helper code
Model Training (on EC2)
Inference code
Training code

Open Source Containers for TF and MXNet

https://github.com/aws/sagemaker-tensorflow-containers

https://github.com/aws/sagemaker-mxnet-containers

- Customize them
- Run them locally for development and testing
- Run them on SageMaker for training and prediction at scale

Bring your own container

https://github.com/aws/sagemaker-container-support

- Integration with SageMaker Python SDK Estimators, including:
  - Downloading user-provided Python code
  - Deserializing hyperparameters (preserving their Python types)
- bin/entry.py, the Docker entrypoint required by SageMaker
- Reading in the metadata files provided to the container during training
- nginx + Gunicorn HTTP server for serving inference requests

https://github.com/awslabs/amazon-sagemaker-examples/tree/master/advanced_functionality/scikit_bring_your_own
https://github.com/awslabs/amazon-sagemaker-examples/tree/master/advanced_functionality/r_bring_your_own

Amazon EC2 C5 instances

C5: Next Generation Compute-Optimized Instances with Intel® Xeon® Scalable Processor

AWS Compute optimized instances support the new Intel® AVX-512 advanced instruction set, enabling you to more efficiently run vector processing workloads with single and double floating point precision, such as AI/machine learning or video processing.

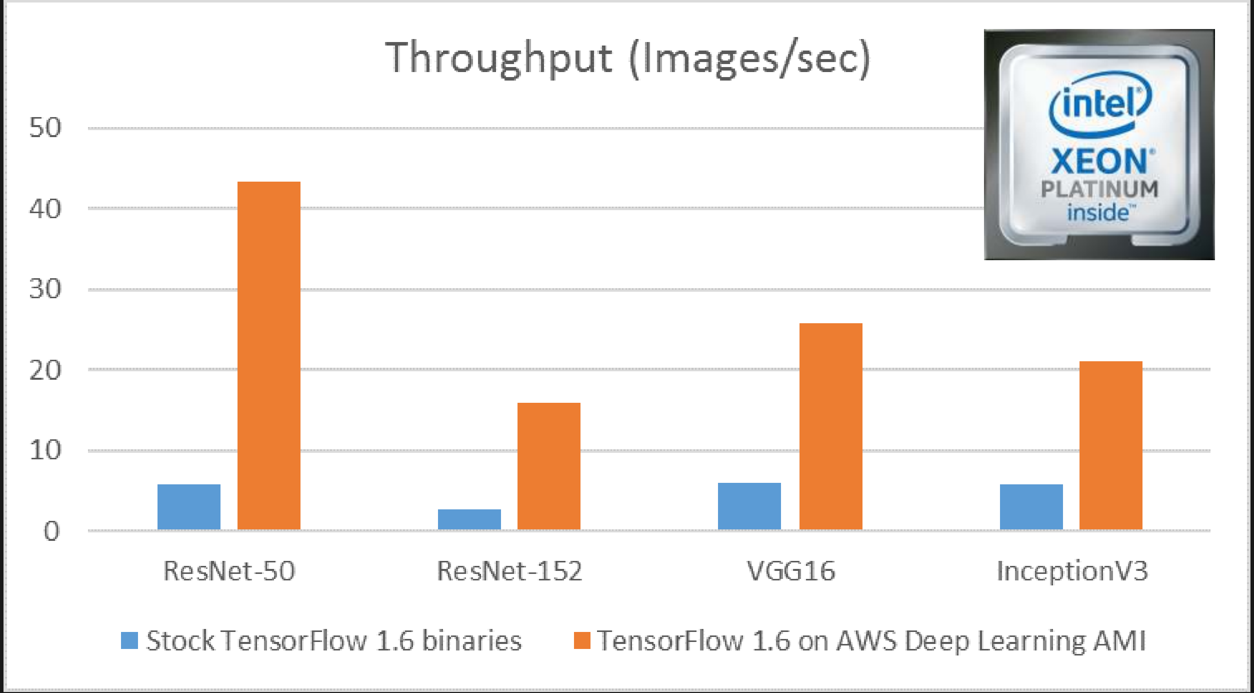
25% improvement in price/performance over C4

C4	C5
36 vCPUs	72 vCPUs **2X vCPUs**
“Haswell”	“Skylake”
4 Gbps to EBS	12 Gbps to EBS **3X throughput**
60 GiB memory	144 GiB memory **2.4X memory**

AVX 512

2X performance

Faster TensorFlow training on C5



https://aws.amazon.com/blogs/machine-learning/faster-training-with-optimized-tensorflow-1-6-on-amazon-ec2-c5-and-p3-instances/

Amazon EC2 P3 Instances

The fastest, most powerful GPU instances in the cloud

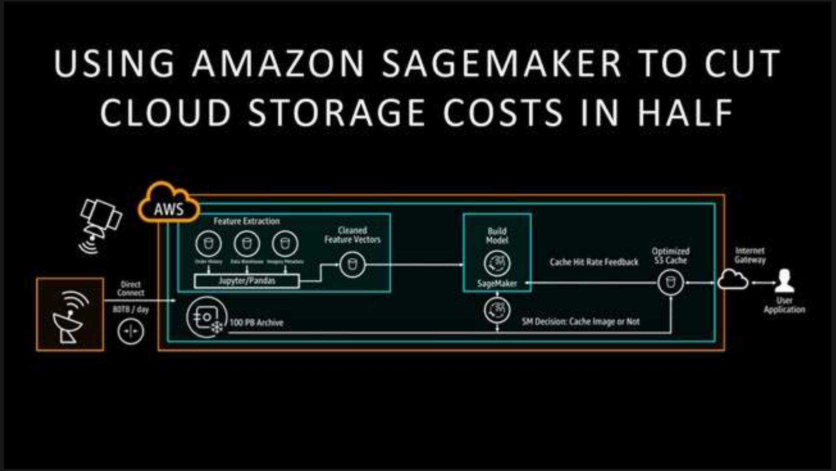
- P3.2xlarge, P3.8xlarge, P3.16xlarge
- Up to eight NVIDIA Tesla V100 GPUs in a single instance
  - 40,960 CUDA cores, 5120 Tensor cores
  - 128GB of GPU memory
- 1 PetaFLOPs of computational performance *– 14x better than P2*
- 300 GB/s GPU-to-GPU communication (NVLink) *– 9x better than P2*

Digital Globe

https://aws.amazon.com/solutions/case-studies/digitalglobe-machine-learning/

DigitalGlobe™

- Operating Earth imaging satellites and providing image analysis services.
- Over 100 PB of imagery.
- Extensive use of Machine Learning on SageMaker to extract information from images.
- Working with the AWS ML Lab, built a predictive model reducing cloud storage costs by 50%.



USING AMAZON SAGEMAKER TO CUT CLOUD STORAGE COSTS IN HALF

DEMOS

https://github.com/juliensimon/dlnotebooks

Thank you!

https://aws.amazon.com/sagemaker
https://github.com/awslabs/amazon-sagemaker-examples
https://github.com/aws/sagemaker-python-sdk
https://github.com/aws/sagemaker-spark

https://medium.com/@julsimon
https://youtube.com/juliensimonfr

Julien Simon

Principal Technical Evangelist, AI and Machine Learning

@ julsimon