

# State of the Union: Machine Learning & Amazon SageMaker

Kwunhok Chan  
Solutions Architect

Benson Kwong  
Solutions Architect



# CENTERPIECE FOR DIGITAL TRANSFORMATION

---



Customer  
experience



Business  
operations



Decision  
making



Innovation



Competitive  
advantage

40%

of digital transformation initiatives  
supported by AI in 2019

—IDC 2018

# MACHINE LEARNING IS HAPPENING IN COMPANIES OF EVERY SIZE AND INDUSTRY

Tens of thousands customers have chosen AWS for their ML workloads | More than twice as many customers using ML than any other cloud providers



# Our mission at AWS














---

Put machine learning in the  
hands of every developer


# The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities






## AI SERVICES

VISION	SPEECH		TEXT			NEW! SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	NEW! FRAUD	NEW! DEVELOPMENT	NEW! CONTACT CENTERS
												
Amazon Rekognition	Amazon Polly	Amazon Transcribe <small>+Medical NEW</small>	Amazon Comprehend <small>+Medical</small>	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens <small>For Amazon Connect</small>

## ML SERVICES

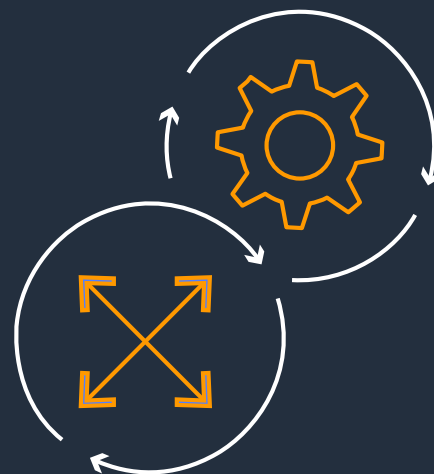
 Amazon SageMaker	Ground Truth	Augmented AI	ML Marketplace	SageMaker Studio IDE <small>NEW!</small>								Neo
				Built-in algorithms	<small>NEW!</small> Notebooks	<small>NEW!</small> Experiments	Model training & tuning	<small>NEW!</small> Debugger	<small>NEW!</small> Autopilot	Model hosting	<small>NEW!</small> Model Monitor	

## ML FRAMEWORKS & INFRASTRUCTURE

 TensorFlow			 GLUON		 Keras	Deep Learning AMLs & Containers		GPUs & CPUs	Elastic Inference	Inferentia (Inf1 instance)	FPGA
PYTORCH											



Easier to build



Easier to scale



Easier to apply

# AMAZON SAGEMAKER

Bringing machine learning to all developers

Pre-built  
notebooks for  
common problems



Collect and  
prepare  
training data

Built-in, high  
performance  
algorithms



Choose and  
optimize your  
ML algorithm

One-click  
training



Set up and manage  
environments  
for training

Optimization



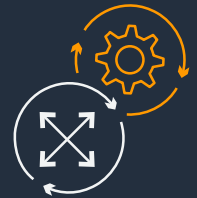
Train and  
tune model  
(trial and error)

One-click  
deployment



Deploy  
model in  
production

Fully managed with  
auto-scaling, health  
checks, automatic handling  
of node failures,  
and security checks



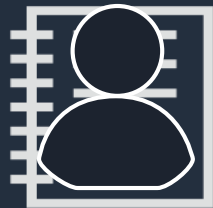
Scale and manage  
the production  
environment

# Amazon SageMaker

## Addressing challenges to machine learning



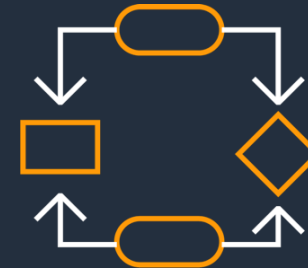
First fully integrated development environment (IDE) for machine learning  
**Amazon SageMaker Studio**



Enhanced notebook experience with quick-start & easy collaboration  
**Amazon SageMaker Notebooks (Preview)**



Experiment management system to organize, track & compare thousands of experiments  
**Amazon SageMaker Experiments**



Automatic debugging, analysis, and alerting  
**Amazon SageMaker Debugger**



Model monitoring to detect deviation in quality & take corrective actions  
**Amazon SageMaker Model Monitor**



Automatic generation of ML models with full visibility & control  
**Amazon SageMaker Autopilot**



**Machine learning is iterative  
involving dozens of tools and  
hundreds of iterations**

Multiple tools needed for  
different phases of the  
ML workflow

+

Lack of an integrated  
experience

+

Large number of iterations

=

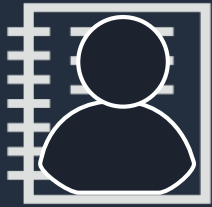
Cumbersome, lengthy processes, resulting in  
loss of productivity

NEW

# Introducing Amazon SageMaker Studio

The first fully integrated development environment (IDE) for machine learning

---



Collaboration at  
scale

Share notebooks  
without tracking code  
dependencies



Easy experiment  
management

Organize, track, and  
compare thousands of  
experiments



Automatic model  
generation

Get accurate models with  
full visibility & control  
without writing code



Higher quality ML  
models

Automatically debug  
errors, monitor models, &  
maintain high quality



Increased  
productivity

Code, build, train, deploy,  
& monitor in a unified  
visual interface



**Data science and collaboration  
needs to be easy**

Setup and manage resources

+

Collaboration across  
multiple data scientists

+

Different data science  
projects have different  
resource needs

=

Managing notebooks and  
collaborating across  
multiple data scientists is  
highly complicated

NEW

# Introducing Amazon SageMaker Notebooks

(Available in Preview)

## Fast-start shareable notebooks

---



Easy access with  
Single Sign-On (SSO)

Access your notebooks in  
seconds with your  
corporate credentials



Fully managed  
and secure

Administrators manage  
access and permissions



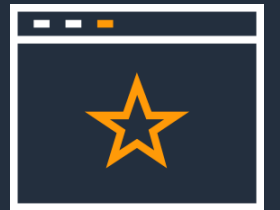
No explicit setup

Start your notebooks  
without spinning up  
compute resources



Easy collaboration

Share your notebooks  
as a URL with a single click



Flexibility

Dial up or down  
compute resources  
(Coming soon)



## **Data Processing and Model Evaluation involves a lot of operational overhead**

Building and scaling  
infrastructure for data processing  
workloads is complex

+

Use of multiple tools or services  
implies learning and  
implementing new APIs

+

All steps in the ML workflow need  
enhanced security, authentication  
and compliance

=

Need to build and manage tooling  
to run large data processing and  
model evaluation workloads

NEW

# Introducing Amazon SageMaker Processing

## Analytics jobs for data processing and model evaluation

---



Fully managed

Achieve distributed processing for clusters



Custom processing

Bring your own script for feature engineering



Container support

Use SageMaker's built-in containers or bring your own



Security and compliance

Leverage SageMaker's security & compliance features



Automatic creation & termination

Your resources are created, configured, & terminated automatically



**Managing trials and experiments is  
cumbersome**

Thousands of experiments

+

Hundreds of parameters  
per experiment

+

Compare and evaluate

=

Very cumbersome and  
error prone

NEW

# Introducing Amazon SageMaker Experiments

## Organize, track, and compare training experiments



### Tracking at scale

Track parameters & metrics across experiments & users



### Custom organization

Organize experiments by teams, goals, & hypotheses



### Visualization

Easily visualize experiments and compare



### Metrics and logging

Log custom metrics using the Python SDK & APIs



### Fast Iteration

Quickly go back & forth & maintain high-quality



## Debugging and profiling deep learning is painful

Large neural networks  
with many layers

+

Data capture with many  
connections

+

Additional tooling for analysis  
and debug

=

Extraordinarily difficult  
to inspect, debug, and profile  
the 'black box'

NEW

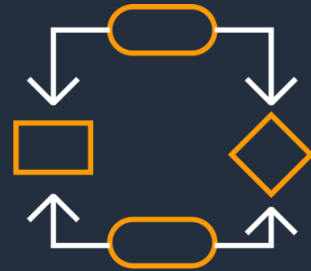
# Introducing Amazon SageMaker Debugger

## Analysis & debugging, explainability, and alert generation



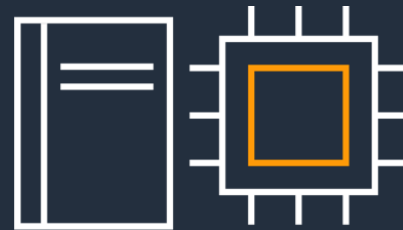
Relevant data  
capture

Data is automatically  
captured for analysis



Data analysis &  
debugging

Analyze & debug data with  
no code changes



Automatic error  
detection

Errors are automatically  
detected based on rules



Improved productivity  
with alerts

Take corrective action  
based on alerts



Visual analysis  
and debugging

Visually analyze & debug  
from SageMaker Studio



**Deploying a model is not the end.  
You need to continuously monitor  
models in production and iterate**

Concept drift due to  
divergence of data

+

Model performance can  
change due to unknown  
factors

+

Continuous monitoring involves a  
lot of tooling and expense

=

Model monitoring is  
cumbersome but critical

NEW

# Introducing Amazon SageMaker Model Monitor

## Continuous monitoring of models in production



Automatic data  
collection

Data is automatically  
collected from your  
endpoints



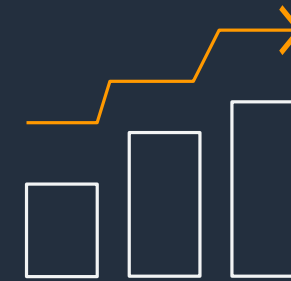
Continuous  
Monitoring

Define a monitoring  
schedule and detect  
changes in quality against  
a pre-defined baseline



Flexibility  
with rules

Use built-in rules to  
detect data drift or write  
your own rules for  
custom analysis



Visual  
Data analysis

See monitoring results,  
data statistics, and  
violation reports in  
SageMaker Studio



CloudWatch  
Integration

Automate corrective  
actions based on Amazon  
CloudWatch alerts



**Successful ML requires  
complex, hard to discover  
combinations**

of algorithms, data, parameters

Largely explorative &  
iterative

+

Requires broad and  
complete  
knowledge of ML domain

+

Lack of visibility

=

Time consuming,  
error prone process,  
even for ML experts

NEW

# Introducing Amazon SageMaker Autopilot

## Automatic model creation with full visibility & control

---



Quick to start

Provide your data in a tabular form & specify target prediction



Automatic model creation

Get ML models with feature engineering & automatic model tuning automatically done



Visibility & control

Get notebooks for your models with source code



Recommendations & Optimization

Get a leaderboard & continue to improve your model

# Demo

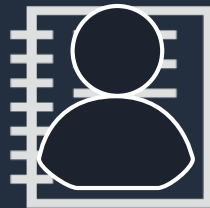


# Amazon SageMaker

## Addressing challenges to machine learning



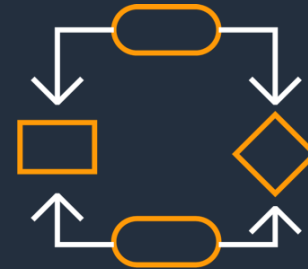
First fully integrated development environment (IDE) for machine learning  
**Amazon SageMaker Studio**



Enhanced notebook experience with quick-start & easy collaboration  
**Amazon SageMaker Notebooks (Preview)**



Experiment management system to organize, track & compare thousands of experiments  
**Amazon SageMaker Experiments**



Automatic debugging, analysis, and alerting  
**Amazon SageMaker Debugger**



Model monitoring to detect deviation in quality & take corrective actions  
**Amazon SageMaker Model Monitor**



Automatic generation of ML models with full visibility & control  
**Amazon SageMaker Autopilot**

# Using Kubernetes for ML is hard to manage and scale

Build and manage services within Kubernetes cluster for ML

+

Make disparate open-source libraries and frameworks work together in a secure and scalable way

+

Requires time and expertise from infrastructure, data science, and development teams

=

Need an easier way to use Kubernetes for ML

NEW

# Introducing Amazon SageMaker Operators for Kubernetes

Kubernetes customers can now train, tune, & deploy models in Amazon SageMaker

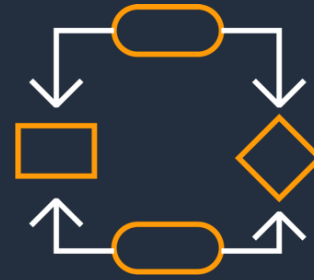
---



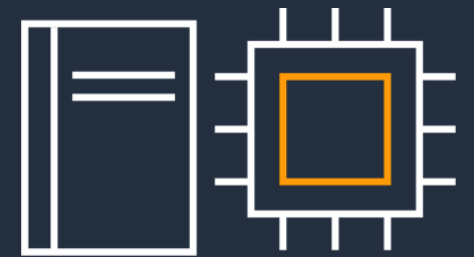
Train, tune, and deploy models in SageMaker



Orchestrate ML workloads from your Kubernetes environments



Create pipelines and workflows in Kubernetes



Fully managed infrastructure in SageMaker

**Deploying models at scale is hard  
to manage and not cost-effective**

Large number of per-user  
models or similar models

+

Different access patterns for  
all models – some highly  
accessed, others infrequently  
accessed

+

Need to have all models in  
production and available to serve  
inferences at low latency

=

High deployment costs and  
challenges in managing scale

# Introducing Amazon SageMaker Multi-model Endpoints

## Deploy and manage thousands of models

---



Easy to deploy & manage models

Store trained models in Amazon S3



Deploy multiple models on an endpoint

Serve all models from a single endpoint



Invoke target model

Concurrently invoke multiple models on the same endpoint



Automatic memory handling

Your memory is managed based on traffic



Significant cost savings

Get improved endpoint & instance utilization

**Training ML models can get  
expensive**

Training ML models can last  
anywhere between few  
minutes to weeks

+

Want to use EC2 Spot  
instances, but they can get  
interrupted

+

ML model training needs to be  
unaffected by interruptions

=

Need to build complex tooling  
to use Spot instances for  
Training ML Models

# Introducing Amazon SageMaker Managed Spot Training

## Save up to 90% in training costs

---



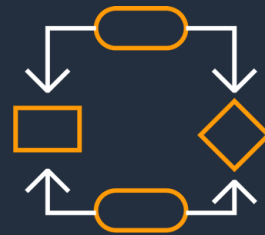
Up to 90% savings

Save training costs compared to Amazon EC2 On-Demand instances



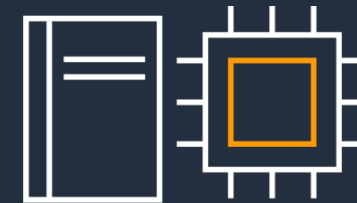
All SageMaker training capabilities

Take advantage of Automatic Model Tuning & Reinforcement



No more interruptions

Spot capacity is managed & interruptions are automatically handled



Support for algorithms & frameworks

Get support for built-in and your own algorithms & frameworks



Full visibility

Visualize your cost savings for each training job

# Resources

- Amazon SageMaker developer resources - <https://aws.amazon.com/sagemaker/developer-resources/>
- Amazon SageMaker sample notebooks - <https://github.com/awslabs/amazon-sagemaker-examples>

# AWS Innovate Online Conference – AI/ML Edition

February 19, 2020 12:30 – 4:30pm



Register Today

Dive deep into any of the **20+ breakout sessions** across **six tracks** delivered by AWS experts; explore key concepts, use cases, best practices, live demos, and live Q&A to learn how other organizations are using **AI and Machine Learning**.

# Thank you!

