# AWS SageMaker

## CCBDA 2025

Aguirre Guaman, Pablo Fabricio    Bestak, Vojtech    Bång, Valdemar
Eberhardt, Jakob    Kosinski, Bartosz

May 12, 2025

# Agenda

- Everything-as-a-Service for Machine Learning
- Tutorial
- Conclusion

# Everything-as-a-Service for Machine Learning

- AWS SageMaker started in 2017 [1]
- Started as a platform for analytics, optimization, and machine learning
- Renamed to AWS SageMaker AI in 2024 [2]
- **Focus on the machine learning workflow**

## Data

- Fully integrated into AWS data management services, e.g. EMR [3]
- Data labeling and model validation AWS Ground Truth [4] and Mechanical Turk [5]
- Cloud services for human workforce

# Everything-as-a-Service for Machine Learning

## Development

- Jupyter notebooks for preprocessing and development
- SageMaker has a widely used Python SDK [6]
- And a IDE called SageMaker Studio [7]

# Everything-as-a-Service for Machine Learning

## Deployment

- Run pre-built or self-developed models on accelerated EC2 instances
- The more responsive, the more expensive [8]
- E.g. real-time inference with auto-scaling vs. serverless vs. batch transform overnight

# Tutorial

- Predict shared bike availability
- Load data into the cloud
- Define and train the model
- Provision it securely via an endpoint

# IAM Configuration

- ↑ Permissions for SageMaker resources are handled through IAM
- We need to grant the user access to SM-related resources
- ↓ We recommend doing it in a real (paid) AWS account to make sure all resources are available

```
"Statement": [
  {
    "Effect": "Allow",
    "Principal": {
      "Service": "sagemaker.amazonaws.com"
    },
    "Action": "sts:AssumeRole"
  }
]
```

**Figure:** Create a new IAM Role in **IAM** → **Access Management** → **Roles** called `AmazonSageMaker-TrainingExecutionRole`

# S3 Configuration

- ↑ SM can get data from all kinds of services and formats
- ↑ We settled for S3 because it is the simplest solution for testing
- ↓ Other persistent resources (e.g. RDS) can be expensive, even if we turn them off

Create new **S3 bucket**

- Set name to `ccbda-research-sagemaker`
- Leave all other settings as default

# Create Jupyter Notebook for development

- Jupyter notebooks are great for exploring and plotting data
- ↑ Run seamless on AWS
- ↑ Easy to collaborate with others
- ↓ Run on remote accelerated EC2 instances which can become expensive
- ↑ Users can download and execute notebooks step-by-step, e.g. our tutorial in `sagemaker_ml.ipynb`

# Handling the dataset

We are gonna use the **Seoul Bike Sharing Demand** dataset [9].

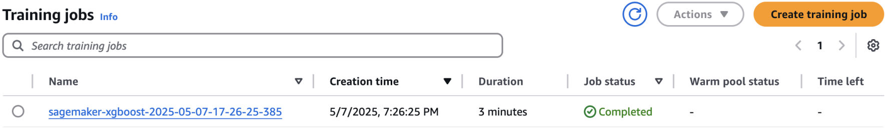
↑ Each cell in the notebook corresponds to:

- Loading the dataset
- Cleaning the data
- Splitting into train and test set
- Uploading it to our S3 bucket

# Training Job Configuration

```
xgboost_image_uri = image_uris.retrieve("xgboost",
↪  region=region, version="1.5-1")
estimator = Estimator(
    image_uri=xgboost_image_uri,
    role=role,
    instance_count=1,
    instance_type="ml.m5.large",
    volume_size=5,
    ...
)
estimator.set_hyperparameters(
    objective="reg:squarederror",
    num_round=100,
    max_depth=5,
    subsample=0.8,
    ...
)
```

# Training Job goes to the Queue



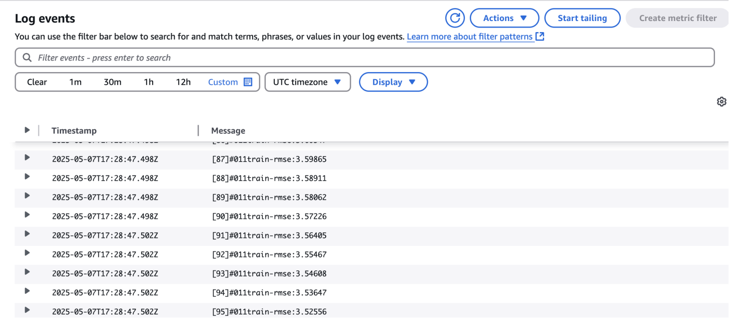
Training jobs

| | Name | Creation time | Duration | Job status | Warm pool status | Time left |
|---|---|---|---|---|---|---|
| ○ | sagemaker-xgboost-2025-05-07-17-26-25-385 | 5/7/2025, 7:26:25 PM | 3 minutes | ✓ Completed | - | - |

**Figure:** Created training job

# CloudWatch Integration

- ↑ Training and deployment logs are integrated in CloudWatch



You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns

```
Timestamp                   Message
2025-05-07T17:28:47.498Z    [87]#011train-rmse:3.59865
2025-05-07T17:28:47.498Z    [88]#011train-rmse:3.58911
2025-05-07T17:28:47.498Z    [89]#011train-rmse:3.58062
2025-05-07T17:28:47.498Z    [90]#011train-rmse:3.57226
2025-05-07T17:28:47.502Z    [91]#011train-rmse:3.56405
2025-05-07T17:28:47.502Z    [92]#011train-rmse:3.55467
2025-05-07T17:28:47.502Z    [93]#011train-rmse:3.54608
2025-05-07T17:28:47.502Z    [94]#011train-rmse:3.53647
2025-05-07T17:28:47.502Z    [95]#011train-rmse:3.52556
```

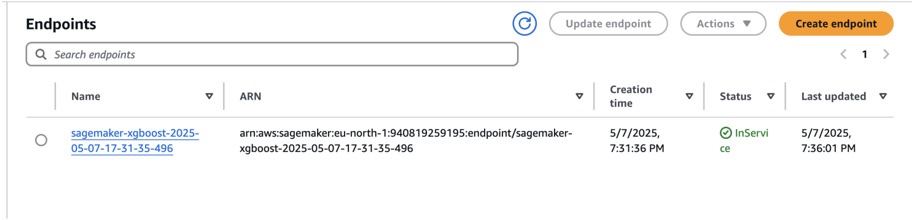
**Figure:** Training Job Logs

# Deploying Endpoint from the Notebook

- ↑ With a single line, we can deploy the trained model
- ↑ We have full control from here, e.g. about instance size, scaling policy

```
predictor = estimator.deploy(initial_instance_count=1,
↪  instance_type="ml.m5.large")
```

# The Endpoint is deployed & scaled automatically



| | Name | ARN | Creation time | Status | Last updated |
|---|---|---|---|---|---|
| ○ | sagemaker-xgboost-2025-05-07-17-31-35-496 | arn:aws:sagemaker:eu-north-1:940819259195:endpoint/sagemaker-xgboost-2025-05-07-17-31-35-496 | 5/7/2025, 7:31:36 PM | ✓ InService | 5/7/2025, 7:36:01 PM |

**Figure:** Published endpoint

# Conclusion & Opinion

- ↑ AWS SageMaker AI is feature complete for all aspects of ML
- ↑ We got started very fast thanks to the notebooks
- ↑ We have fine grain control over expected performance and QoS
- ↑↓ AWS provides adequate tooling, e.g. Python SDK
- ↑ Features industry-standard libraries, e.g. TensorFlow

# References & Questions I

[1] Ron Miller. *AWS releases SageMaker to make it easier to build and deploy machine learning models*. 8 May 2025. 2017. URL: https://techcrunch.com/2017/11/29/aws-releases-sagemaker-to-make-it-easier-to-build-and-deploy-machine-learning-models/.

[2] AWS. *Amazon SageMaker AI Rename*. https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html. Accessed: 2025-05-08. 2025.

[3] Amazon Web Services. *Amazon EMR*. https://aws.amazon.com/emr/. Accessed: 2025-05-08.

[4] Amazon Web Services. *SageMaker Ground Truth*. https://aws.amazon.com/sagemaker-ai/groundtruth/. Accessed: 2025-05-08.

# References & Questions II

[5] Amazon Web Services. *Using the Amazon Mechanical Turk Workforce*. https://docs.aws.amazon.com/sagemaker/latest/dg/sms-workforce-management-public.html. Accessed: 2025-05-08.

[6] Amazon Web Services. *Amazon SageMaker Notebook Instances and SDK*. https://docs.aws.amazon.com/sagemaker/latest/dg/nbi.html. Accessed: 2025-05-08.

[7] Amazon Web Services. *Amazon SageMaker Studio*. https://aws.amazon.com/sagemaker-ai/studio/?nc1=h_ls. Accessed: 2025-05-08.

[8] Amazon Web Services. *Model deployment options in Amazon SageMaker AI*. https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-deployment.html. Accessed: 2025-05-08.

# References & Questions III

[9] UCI Machine Learning Repository. *Seoul Bike Sharing Demand*. https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand. Accessed: 2025-05-05. 2020. DOI: 10.24432/C5F62R.