Mphasis
The Next Applied

FRB Hackathon

Machine Learning with AWS API

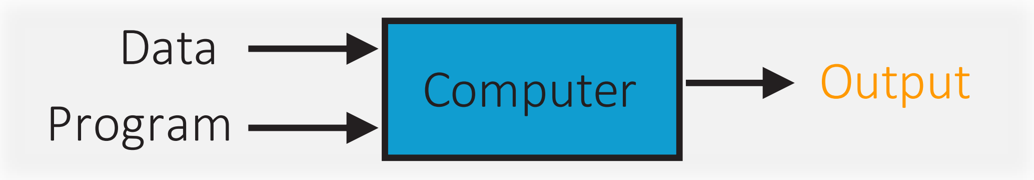
Agenda

1. Machine Learning
   A. Introduction
   B. Algorithms
   C. Cased Studies
   D. ML in Practice
2. Application Program Interface (API)
   A. Introduction
   B. ML using AWS API
      I. Comprehend: Text Analytics
      II. AWS Textract
      III. SageMaker Classification
3. Summary

Machine Learning: Introduction (1/2)

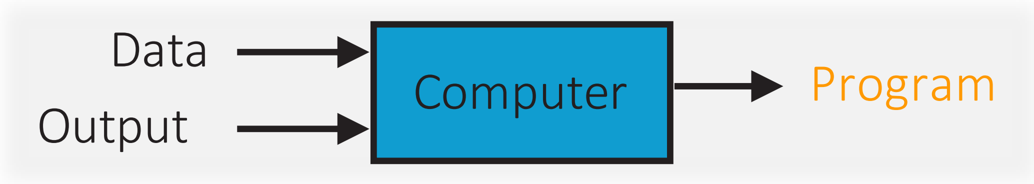
- What is Machine Learning ?

Traditional Programing



Write programs to automate the task; system will take input and execute the instructions to generate output

Machine Learning



Automating automation; Getting computers to program themselves

- Machine Learning:
  "Changes in a system that enable it to do the same task or tasks drawn from the same population more efficiently and more effectively the next time'' (Simon 1983)

- Learning: *"Learning is any process by which a system improves performance from experience." (Simon)*

Machine Learning: Introduction (2/2)

- House Price Prediction Case Study: Predict the monetary value of a house located at the Boston's area

- Traditional Approach: From our experience, we would try to understand the house features which impact the house price and then generate the equation to come up with the final Value

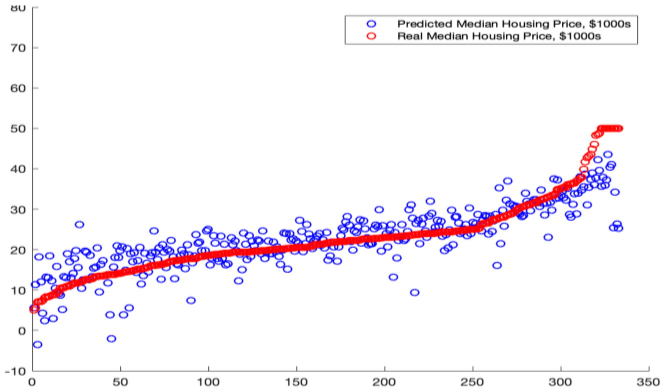
$$\text{Price} = X_0 + X_1 * \text{Floor Space} + X_2 * \text{No of Rooms} + X_3 * \text{Row House}$$

- Machine Learning:

Input Data

Price	Floor space	Rooms	Lot size	Row house	Corner house	Detached
250000	71	4	92	1	0	0
209500	98	5	123	1	0	0
349500	128	6	114	1	0	0
250000	86	4	98	1	0	0
419000	173	6	99	1	0	0
225000	83	4	67	1	0	0
549500	165	6	110	1	0	0
240000	71	4	78	1	0	0
340000	116	6	115	1	0	0

Output



Types Of Machine Learning

- Below are the two main ways in which machine can learn:
  1. Supervised Learning
  2. Unsupervised Learning

- Supervised Learning: In Supervised learning, we train the machine using data which is well "**labeled**."

  I. A supervised learning algorithm learns from labeled training data, helps you to predict outcomes for unforeseen data

  II. Formal Definition: With input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output: Y = f(X)

  III. The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data

Supervised Machine Learning: Regression

- Supervised Learning: Under Supervised learning, below are two main types of algorithms:
  I. Regression
  II. Classification

- Regression: A regression problem is when the output variable is a quantitative value value, such as house prices in Dollars, Travel Time in Minutes or Amount of rain in Centimeter

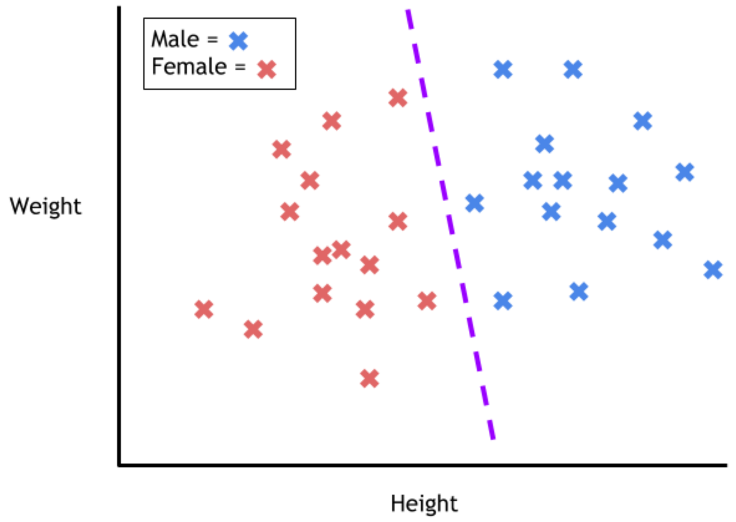
- Business use Case:
  I. Predict the number of mortgage application coming in every day to provide assistance with the personnel planning
- Solution:
  I. Build Regression model using past data with Application as dependent variables and below important features as independent variables:
     i. Interest rate
     ii. Changes in the interest rates
     iii. The amount of mortgage applications on the previous day
     iv. Holidays
     v. The day of the year

Supervised Machine Learning: Classification (1/2)

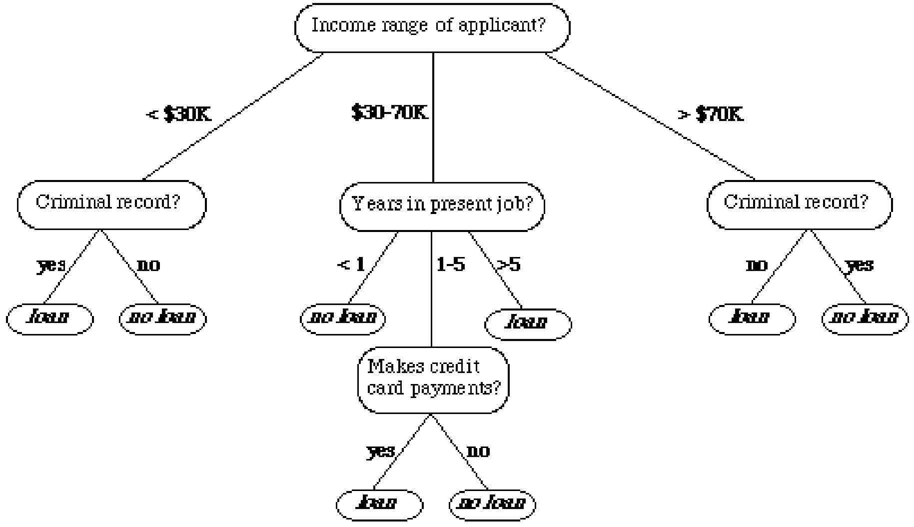
- Classification: A classification problem is when the output variable is a category;
  I. In short Classification either predicts categorical class labels or classifies data (construct a model) based on the training set and the values (class labels) and uses it in classifying new data

- For example, when filtering emails “spam” or “not spam”, when looking at transaction data, “fraudulent”, or “authorized”

Example 1: Classify Male & Female based upon their characteristics



Example 2: Loan Underwriting Use case



Supervised Machine Learning: Classification (2/2)

- Case Study:
  I. Identify if the particular customer is creditworthy or not for the applied loan(Loan Underwriting)

- Input Data:

Dependent Variables

Independent Variable

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y

- Solution:
  I. Build a decision tree classifier with dependent variables to predict the Loan status with below features:
     i. Credit History
     ii. Education
     iii. Applicant Income, Co-applicant Income
     iv. Loan Amount, Loan Amount Term

Machine Learning: In Practice

- Below is the machine learning pipeline to build and deploy the machine learning model:

Create and Train your Model

Prepare Training Data

Read input data and pre process it to prepare for model building

Choose a Model

Select the appropriate algorithm based upon the business problem

Train your Model

Model will train on input data; Calculate the performance metrics

Improve It

Iterate & Re-Train

Evaluate, customize and re-train your model until it's optimized

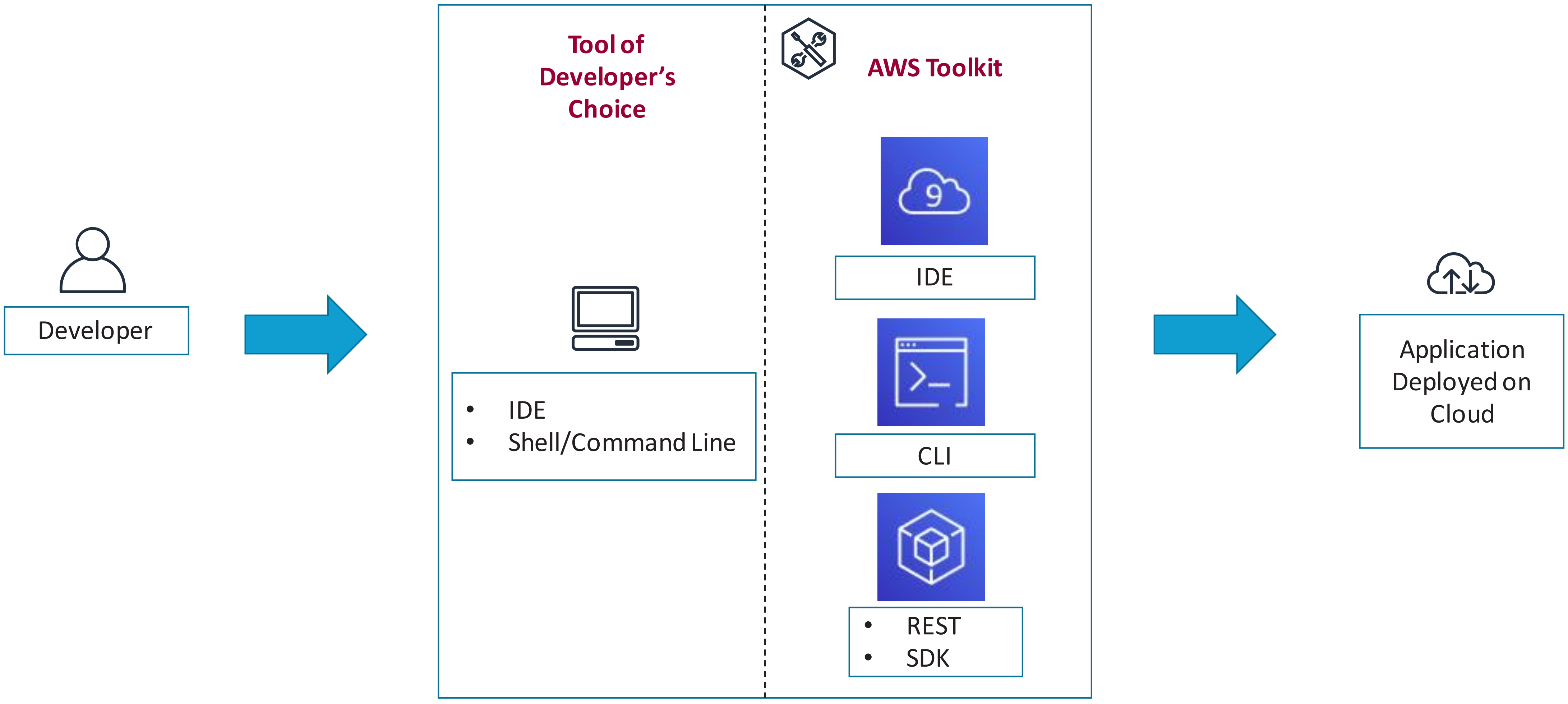
Apply It

Apply the Model

Apply your model on the future data to get the insights

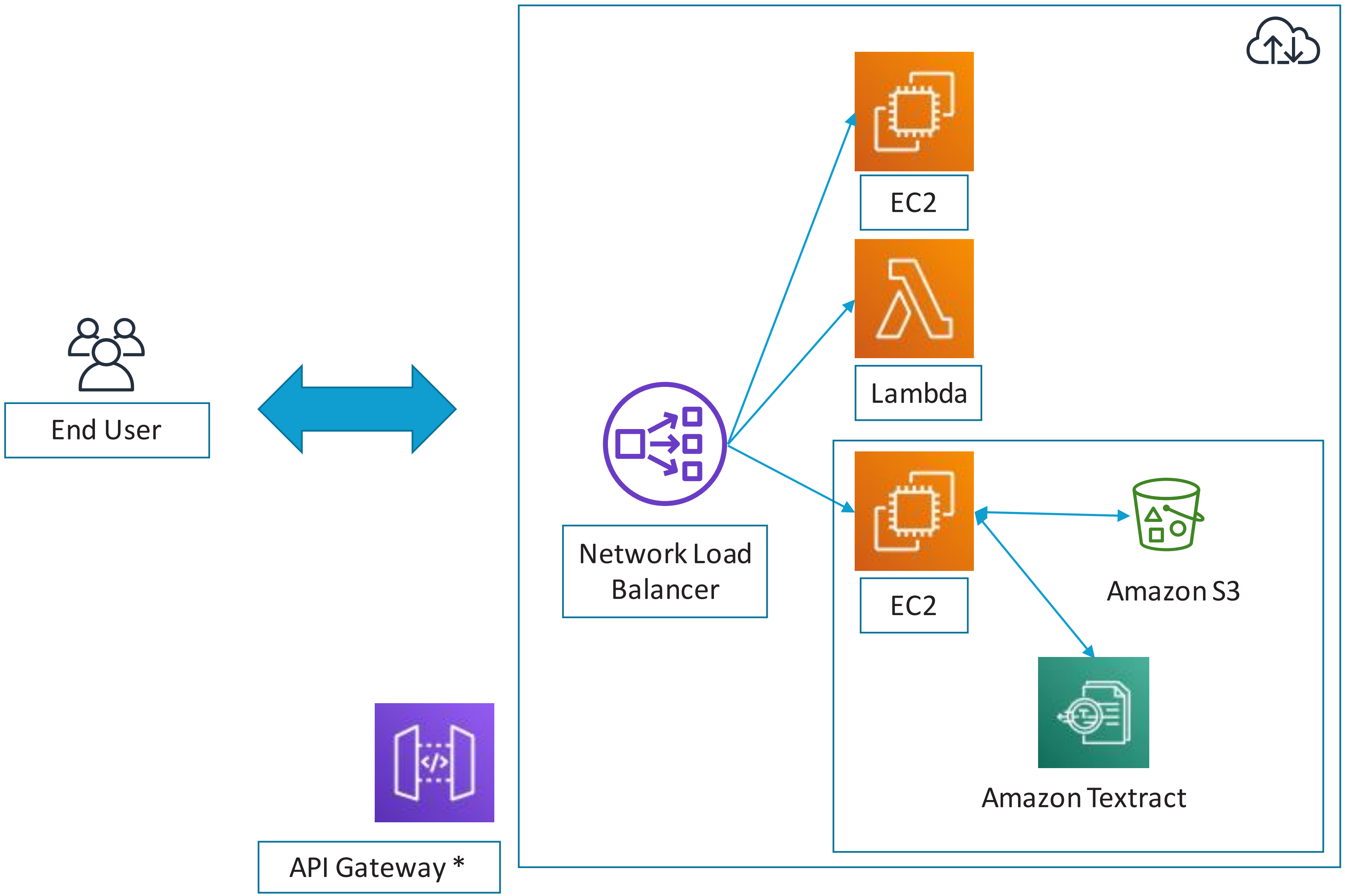
APPLICATION PROGRAM INTERFACE (API)

AWS – Building Applications



Application building tools

AWS – Application Architecture



Few AWS APIs

- Amazon Comprehend
- Amazon Forecast
- Amazon Lex
- Amazon Simple Email Service
- Amazon Personalize
- Amazon Polly
- Amazon Recognition
- Amazon S3
- Amazon Textract
- Amazon Transcribe
- Amazon Translate

AWS Python SDK

https://boto3.amazonaws.com/v1/documentation/api/latest/index.html

S3 Download file example

```
import boto3

s3 = boto3.client('s3')
s3.download_file('BUCKET_NAME', 'OBJECT_NAME', 'FILE_NAME')
```

S3 Upload file example

```
import logging
import boto3
from botocore.exceptions import ClientError


def upload_file(file_name, bucket, object_name=None):
    """Upload a file to an S3 bucket

    :param file_name: File to upload
    :param bucket: Bucket to upload to
    :param object_name: S3 object name. If not specified then file_name is used
    :return: True if file was uploaded, else False
    """

    # If S3 object_name was not specified, use file_name
    if object_name is None:
        object_name = file_name

    # Upload the file
    s3_client = boto3.client('s3')
    try:
        response = s3_client.upload_file(file_name, bucket, object_name)
    except ClientError as e:
        logging.error(e)
        return False
    return True
```

COMPREHEND API: Text Analytics

Text Analytics

Amazon Comprehend

Document Classification

- Ticket Routing
- Information Retrieval (Auditor from Annual Statements)

Sentiment Analysis

- Due Diligence – Risk Assessment
- Targeted Campaign – Voice of Customer

Entity Recognition

- Information Retrieval (Auditor from Annual Statements)

Amazon Lex

Document Classification – Annual Report (Proxy Statement) - Dataset

Amazon Comprehend

sec.gov/Archives/edgar/data/66740/000120677419001068/mmm3460801-def14a.htm#proposal_2_42

PROPOSAL 2

Ratification of the Appointment of Independent Registered Public Accounting Firm for 2019

- Ratify the appointment of PricewaterhouseCoopers LLP as 3M's independent registered public accounting firm for 2019.
- Based on its assessment of the qualifications and performance of PricewaterhouseCoopers LLP ("PwC") the Audit Committee believes that it is in the best interests of the Company and its stockholders to retain PwC.

The Audit Committee is directly responsible for the appointment, compensation (including approval of all fees), retention, and oversight of the Company's independent registered public accounting firm ("Independent Accounting Firm") retained to perform the audit of our financial statements and our internal control over financial reporting.

The Audit Committee has appointed PricewaterhouseCoopers LLP ("PwC") to serve as 3M's Independent Accounting Firm for 2019. PwC has been 3M's Independent Accounting Firm since 1998. Prior to that, 3M's Independent Accounting Firm was Coopers & Lybrand from 1975 until its merger with Price Waterhouse in 1998. In accordance with SEC rules and PwC policy, audit partners are subject to rotation requirements to limit the number of consecutive years an individual partner may provide service to our Company. For lead and concurring audit partners, the maximum number of consecutive years of service in that capacity is five years. The process for selection of the Company's lead audit partner pursuant to this rotation policy involves a meeting between the Chair of the Audit Committee and the candidate for the role, as well as discussion by the full Committee and with management.

The Audit Committee annually reviews PwC's independence and performance in connection with the Audit Committee's determination of whether to retain PwC or engage another firm as our Independent Accounting Firm. In the course of these reviews, the Audit Committee considers, among other things:

Document Classification – Annual Report (Proxy Statement) - Dataset

filename	text	label
ABBOTT LABORATORIES.html	DEF 14A 1 a2222821zdef14a.htm DEF 14A.Use these links to rapidly review the do	0
ABBOTT LABORATORIES.html	).Filed by the Registrant.Filed by a Party other than the Registrant o.Check the ap	0
ABBOTT LABORATORIES.html	(Name of Registrant as Specified In Its Charter).	0
ABBOTT LABORATORIES.html	(Name of Person(s) Filing Proxy Statement, if other than the Registrant).Payment	0
ABBOTT LABORATORIES.html	(1).Title of each class of securities to which transaction applies.	0
ABBOTT LABORATORIES.html	(2).Aggregate number of securities to which transaction applies.	0
ABBOTT LABORATORIES.html	(3).Per unit price or other underlying value of transaction computed pursuant to E	0
ABBOTT LABORATORIES.html	(4).Proposed maximum aggregate value of transaction.	0
ABBOTT LABORATORIES.html	(5).Total fee paid.o.Fee paid previously with preliminary materials.o.Check box if	0
ABBOTT LABORATORIES.html	Identify the previous filing by registration statement number, or the Form or Sch	0
ABBOTT LABORATORIES.html	(1).Amount Previously Paid.	0
ABBOTT LABORATORIES.html	(2).Form, Schedule or Registration Statement No.	0
ABBOTT LABORATORIES.html	(4).Date Filed.Table of Contents.Table of Contents.Abbott Laboratories 100 Abbo	0
ABBOTT LABORATORIES.html	Glucerna.Juliana Auler, S o Paulo, Brazil.Juliana Auler is an English teacher, transl	0

filename	text	label
ABBOTT LABORATORIES.html	In October 2014, the Audit Committee appointed Ernst & Young LLP to act as audit	1
ABBOTT LABORATORIES.html	If the shareholders do not ratify the appointment of Ernst & Young LLP as auditors	1
ABBOTT LABORATORIES.html	2.Ratification of Ernst & Young LLP as auditors 3.	1
ABBOTT LABORATORIES.html	Ratification of Ernst & Young LLP as auditors 1.	1
ALCOA INC..html	Based on its evaluation, the Audit Committee has appointed PricewaterhouseCo	1
ALCOA INC..html	The Audit Committee and the Board believe that the continued retention of Price	1
ALCOA INC..html	In addition, the Audit Committee has approved, subject to shareholder ratificatio	1
BIOGEN INC..html	Ratification of PricewaterhouseCoopers LLP.	1
BIOGEN INC..html	FOR the ratification of the selection of PricewaterhouseCoopers LLP as our indep	1
BIOGEN INC..html	Ratification of PricewaterhouseCoopers LLP.	1
BIOGEN INC..html	The affirmative vote of a majority of shares present in person or represented by p	1

https://spacy.io/api/tokenizer

Document Classification – Dataset

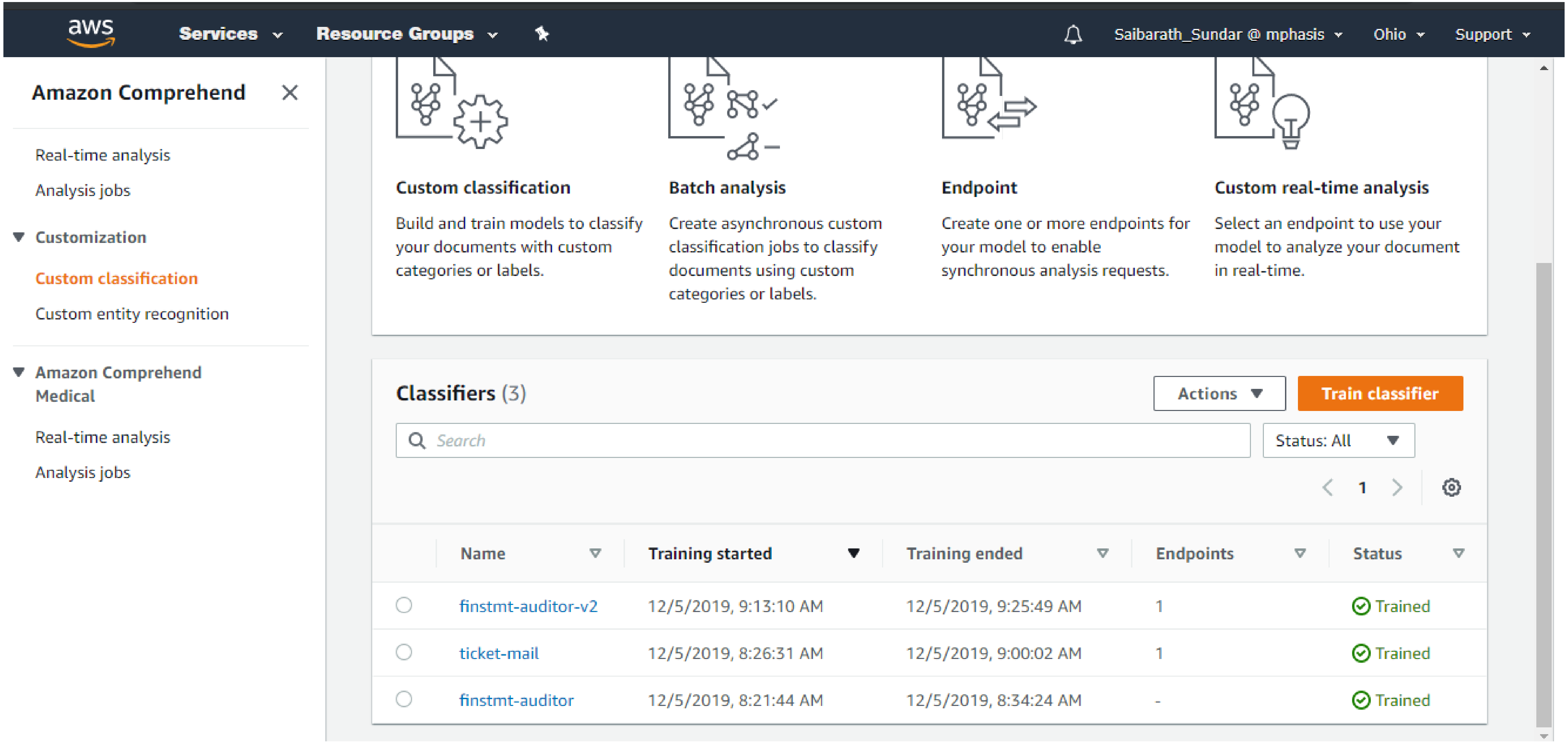
Type a prefix and press Enter to search. Press ESC to clear.

Upload | Create folder | Download | Actions

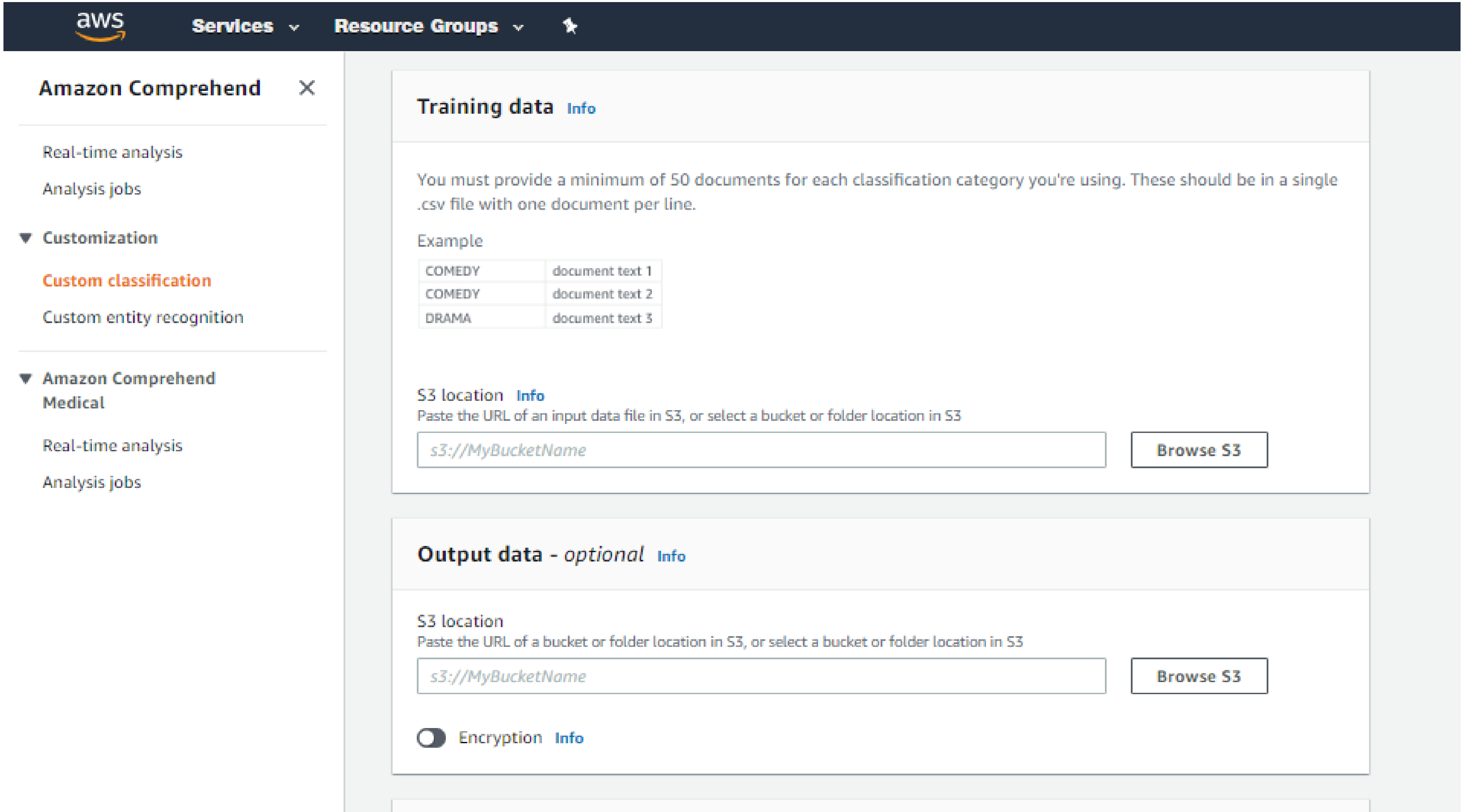
US East (Ohio)

786796469737-CLR-92bce333a8a26a778b9e30dfa84a593c	--	--	--
786796469737-CLR-b8b60933da84a5aeaae7e841e013601f	--	--	--
.write_access_check_file.temp	Dec 5, 2019 9:13:12 AM GMT+0530	0 B	Standard
finstmt.csv	Dec 5, 2019 9:11:44 AM GMT+0530	416.6 KB	Standard
ticket.csv	Dec 5, 2019 8:25:06 AM GMT+0530	12.5 MB	Standard

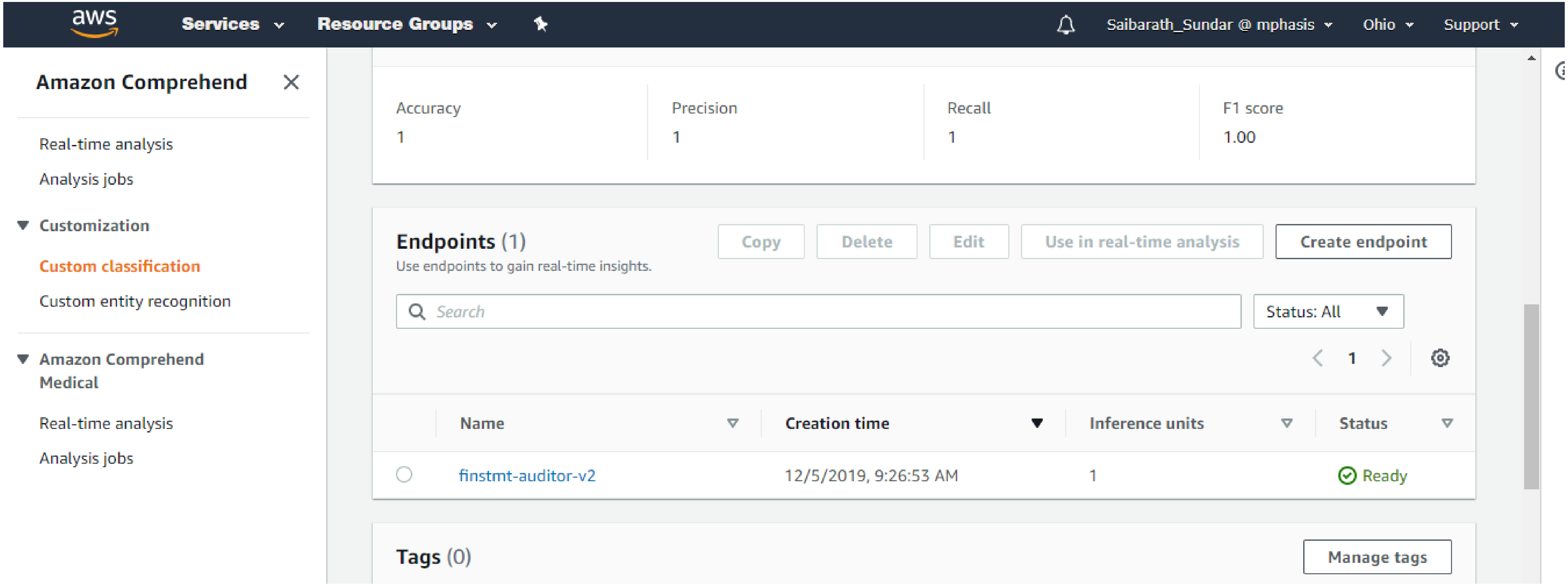
Document Classification – Training



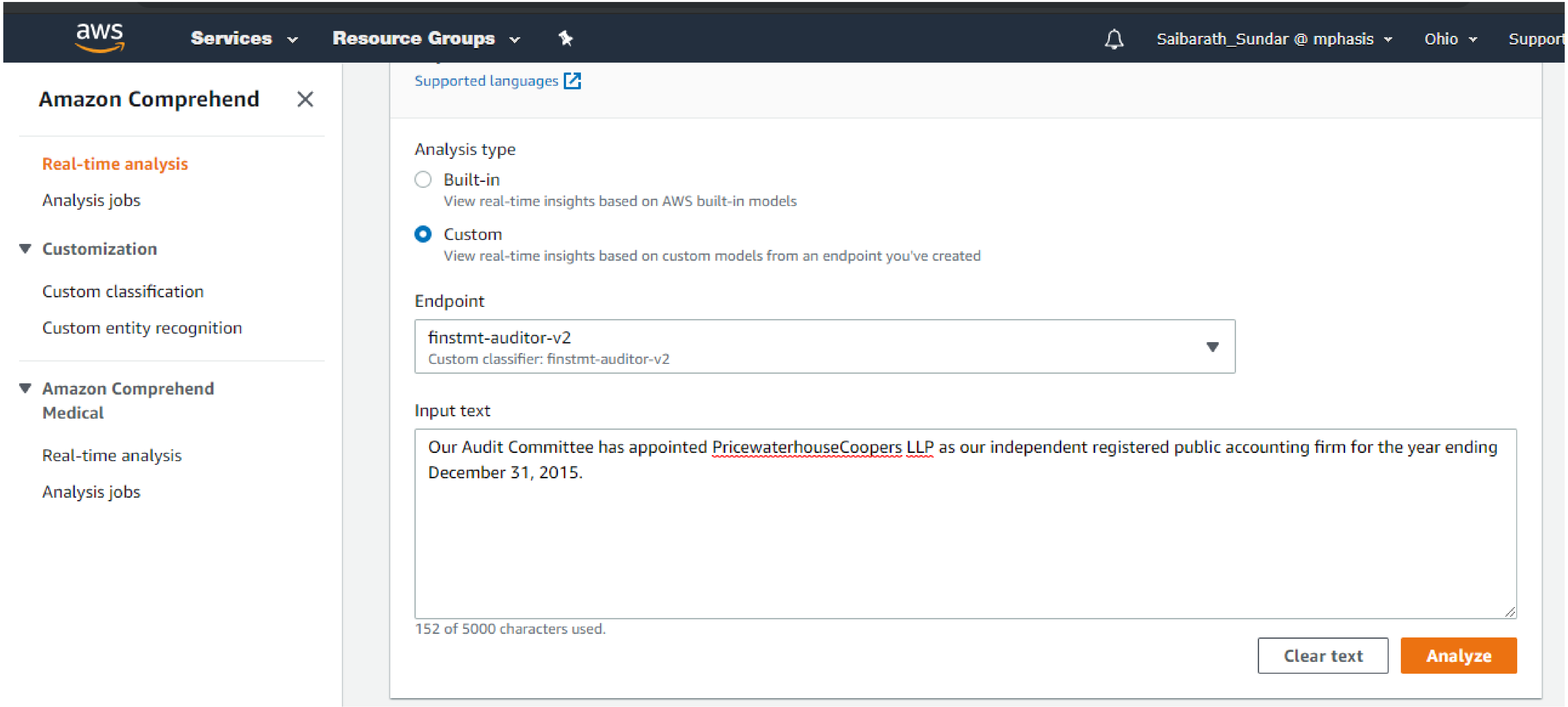
Document Classification – Training



Document Classification – Endpoint



Document Classification – Endpoint



Document Classification – Endpoint

▼ Application integration

Request and response of ClassifyDocument API. See next steps in the Documentation

Request


```json
{
    "Text": "Our Audit Committee has appointed
        PricewaterhouseCoopers LLP as our independent
        registered public accounting firm for the year
        ending December 31, 2015.",
    "EndpointArn": "arn:aws:comprehend:us-east-2
        :786796469737:document-classifier-endpoint
        /finstmt-auditor-v2"
}
```


Response


```json
{
    "ClassifyDocumentResponse": {
        "Classes": [
            {
                "Name": "1",
                "Score": 0.9956875443458557
            },
            {
                "Name": "0",
                "Score": 0.004332461394369602
            }
        ]
    }
}
```

Entity Recognition

```
string = 'Although ratification is not required by our By-Laws or otherwise, the Board is submitting the selection of D&T to our
r = comprehend_api_entities(string, key_id,secret_key)
r
```


```
{'Entities': [{'Score': 0.9996688365936279,
   'Type': 'ORGANIZATION',
   'Text': 'D&T',
   'BeginOffset': 108,
   'EndOffset': 111}],
 'ResponseMetadata': {'RequestId': '29d6ba87-5689-4ddb-85ff-1ad8a75e1a3b',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amzn-requestid': '29d6ba87-5689-4ddb-85ff-1ad8a75e1a3b',
   'content-type': 'application/x-amz-json-1.1',
   'content-length': '112',
   'date': 'Thu, 05 Dec 2019 06:14:46 GMT'},
  'RetryAttempts': 0}}
```

Sentiment Analysis

```
string = 'All the key economic indicators have been on a downward swing for quite some time. If this slide is not checked India r
r = comprehend_api_sentiment(string, key_id,secret_key)
r
```


```
{'Sentiment': 'NEGATIVE',
 'SentimentScore': {'Positive': 0.12709932029247284,
  'Negative': 0.4827170670032501,
  'Neutral': 0.3885166645050049,
  'Mixed': 0.0016669268952682614},
 'ResponseMetadata': {'RequestId': '74a3f62f-cc53-43b5-ba77-5a50295eb36d',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amzn-requestid': '74a3f62f-cc53-43b5-ba77-5a50295eb36d',
   'content-type': 'application/x-amz-json-1.1',
   'content-length': '163',
   'date': 'Thu, 05 Dec 2019 06:14:10 GMT'},
  'RetryAttempts': 0}}
```

TEXTRACT API: OPTICAL CHARACTER RECOGNITION

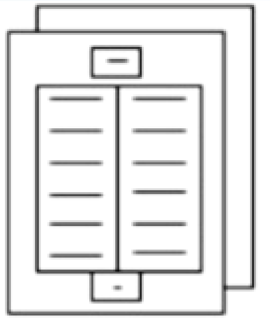
AI Services: AWS Textract

- Information in form of documents is major part of record keeping, collaborating and transacting in major domains such as Finance, Real-Estate, Medical, Legal, Insurance and Business management.
- There is a need to digitize and analyze these documents, for Search and Discovery, Compliance and Business process Automation.

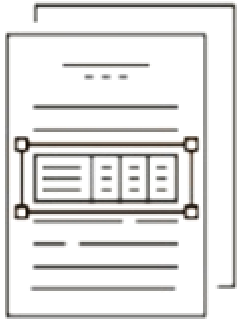
Challenges

- Manual Processing took a lot of time and effort to categorize, filter and prepare documents for analysis
- Multiple Templates created a hurdle for generic extraction algorithms
- Difficult to capture information from Complex documents – skewed, tilted
- Manual rules for Key Value pairs, Table Extraction and Form data Extraction

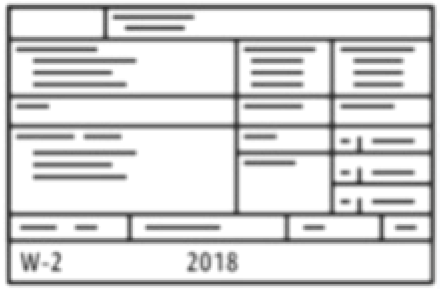
Solution



Accurate
Text Extraction
Line and paragraph Identification



Accurate
Table Extraction
Columns and Cells Identification



Accurate
Forms Extraction
Key Value pair Identification

Machine Learning Model On Sage Maker : CLASSIFICATION

AWS SageMaker: ML Services

Build Algorithms and notebooks ⟷ One Click model training, Tuning, and Testing ⟷ One Click deployment and hosting

Collect and prepare training data → Choose and optimize ML algorithm → Set up and manage environments for training → Train and tune ML models → Deploy models in production → Scale and manage the production environment

SageMaker Built-in Algorithms

- K-means Clustering
- Neural Topic Modelling
- Linear Learner
- XGBoost
- Image Classification
- DeepAR Forecasting
- BlazingText (word2vec)

Bring Your Own Algorithm

- R
- MXNet
- Tensorflow
- PyTorch
- Keras
- Apache Spark

One-click training and Hyperparameter optimization

From Command Line or Python

- Specify Container
- AWS CLI
- Algorithm
- Input Data
- Hardware

One-click Deployment

- Create an endpoint configuration for an HTTPS endpoint
- Create an HTTPS endpoint

Benefits

Cost

- Reduce ML Training time
- GPU utilization
- Support large complex models
- Lower inference costs

Data

- Automatic data labeling for training
- Faster Data cleaning and Feature Engineering

Ease of Use

- Pre defined models for training
- Automatic Optimization
- Methods for Accuracy and performance metrics

AWS SageMaker: End to End Build

Create an Amazon S3 Bucket

Stores:
Training data
Model artifacts

Create an Amazon SageMaker Notebook Instance

Fully managed EC2 instance
To create and manage
Jupyter notebooks

Create a Jupyter Notebook

Python 3 Notebook
Needs name of S3 bucket

Create Training Dataset

Download dataset
Transform and upload
to S3 bucket

Train a Model

Choose the Training
Algorithm
Set Hyperparameters and
start model training

Deploy the Model to Amazon SageMaker

Creates deployable model,
Configures the SageMaker
hosting services endpoint,
Launches endpoint

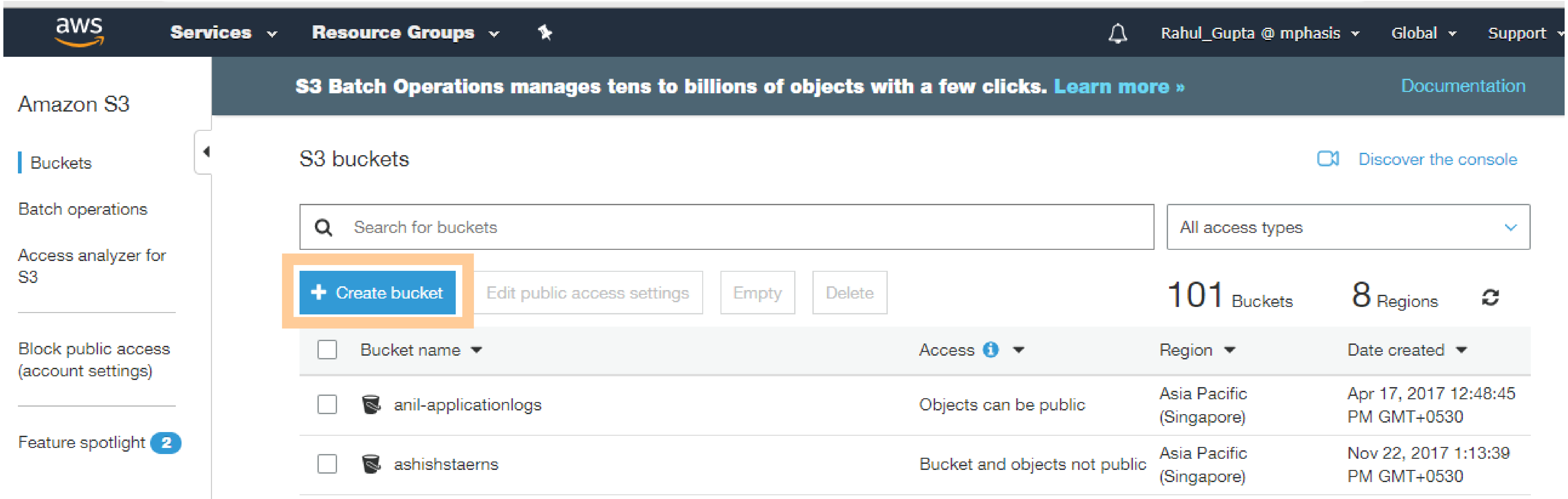
Validate the Model

validate accurate predictions
on new data

Integrating SageMaker Endpoints into Applications

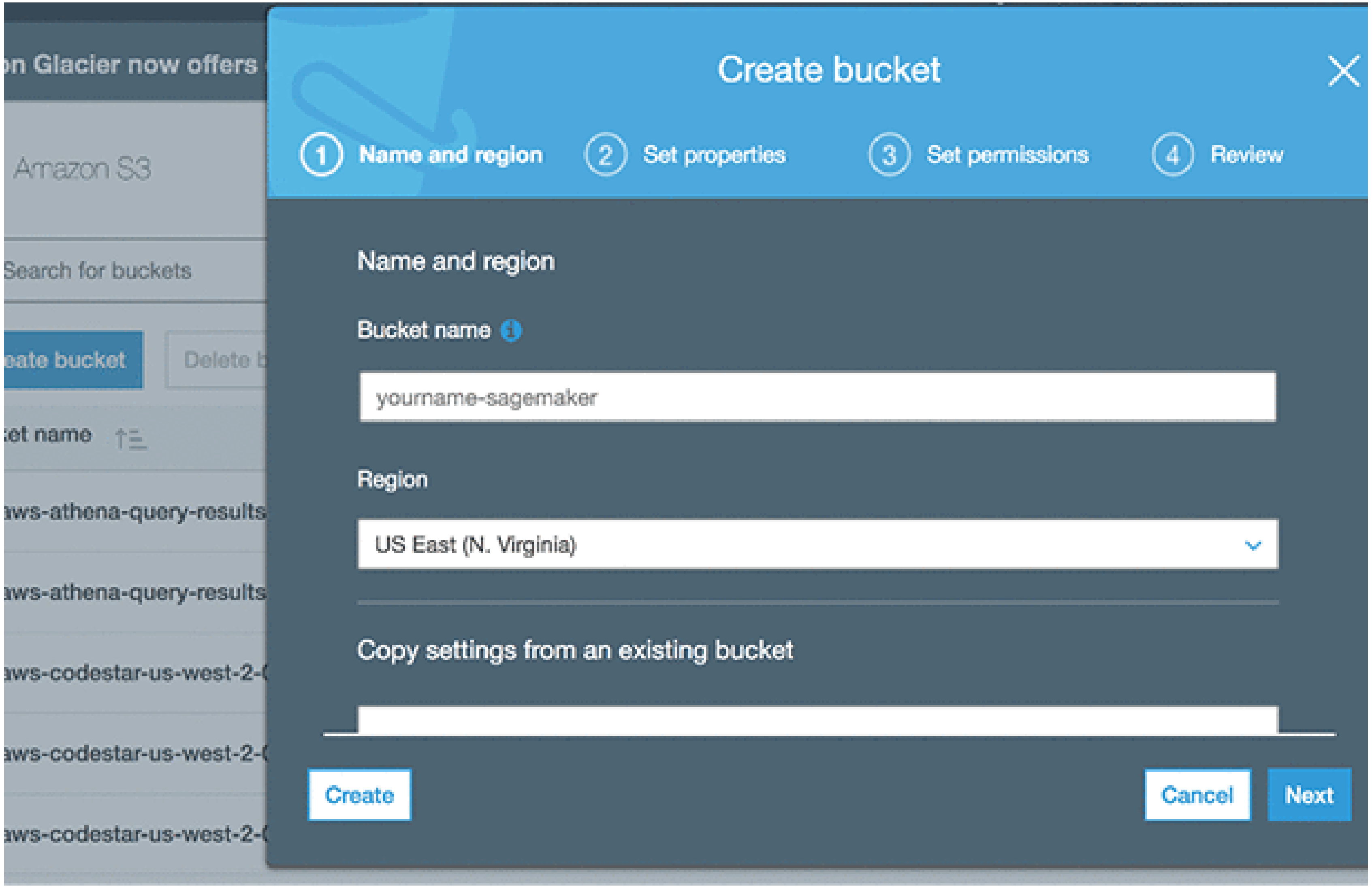
Call endpoint using set keys

Create Amazon S3 Bucket



From Services → Select Amazon S3 → Click on **Create Bucket** to open create bucket window

Create Amazon S3 Bucket



Specify a bucket name

→

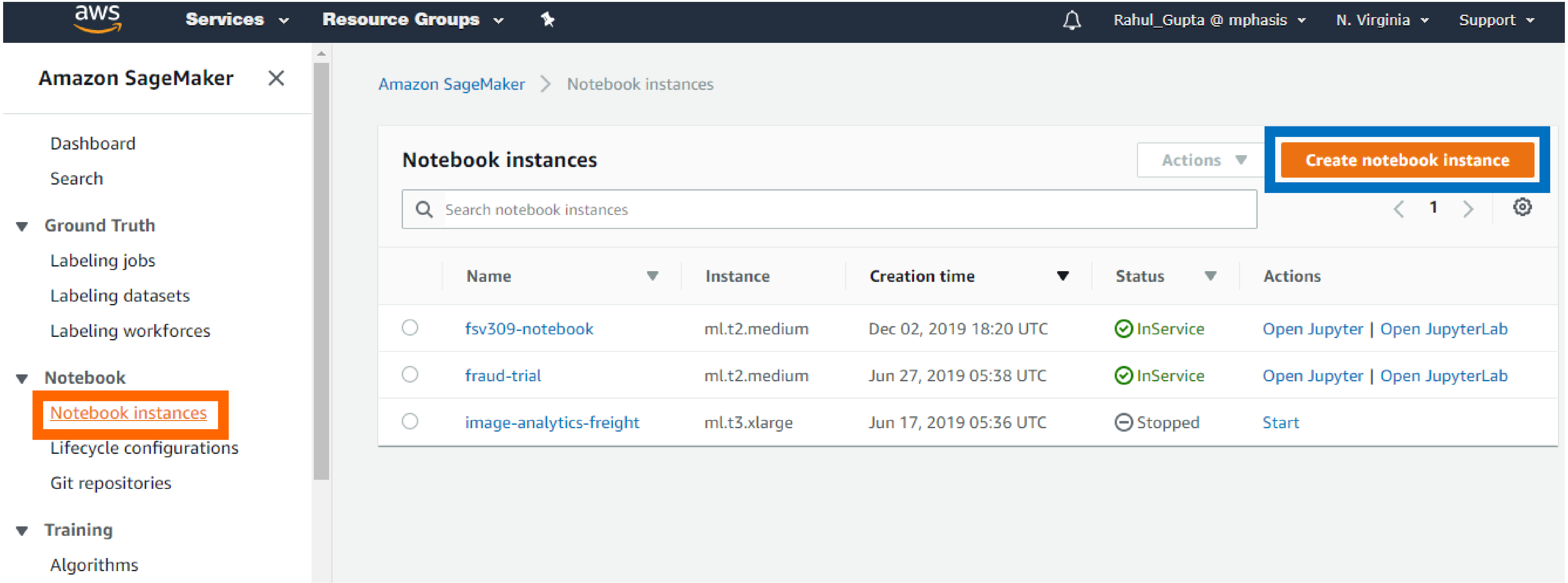
Select Region

→

Click on **Create**

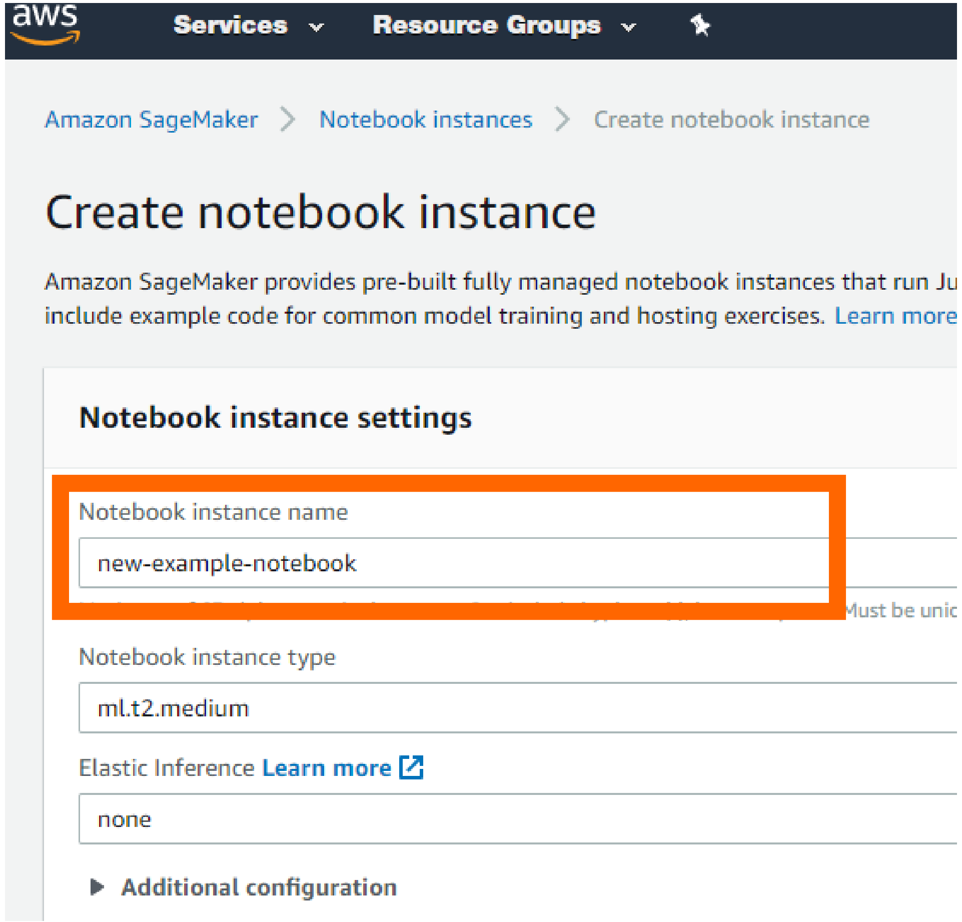
An S3 bucket with default properties and permissions will be created. They can be changed later per requirement.

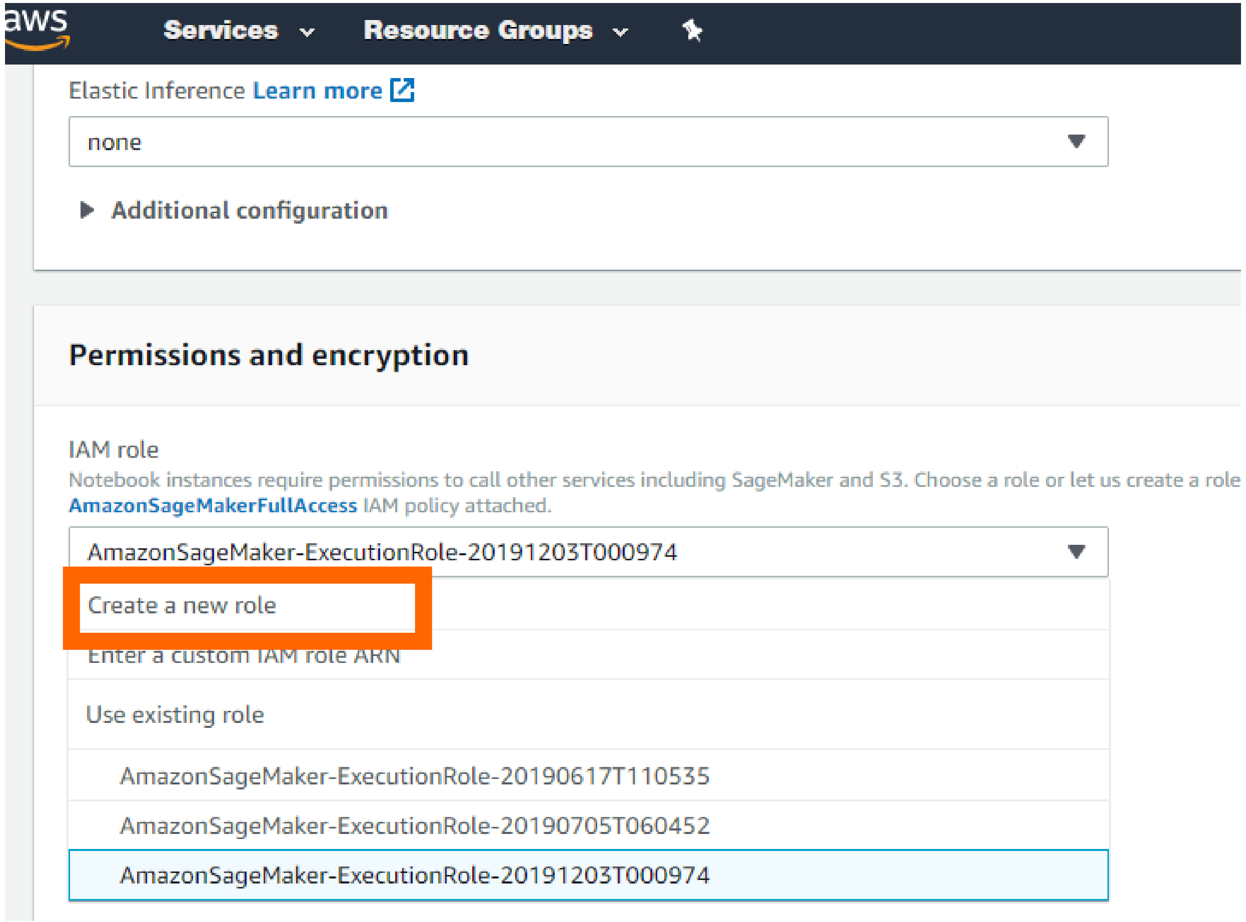
Create notebook instance



From Services → Select Amazon SageMaker → Click on **Notebook instances** → Click on **Create notebook instance**

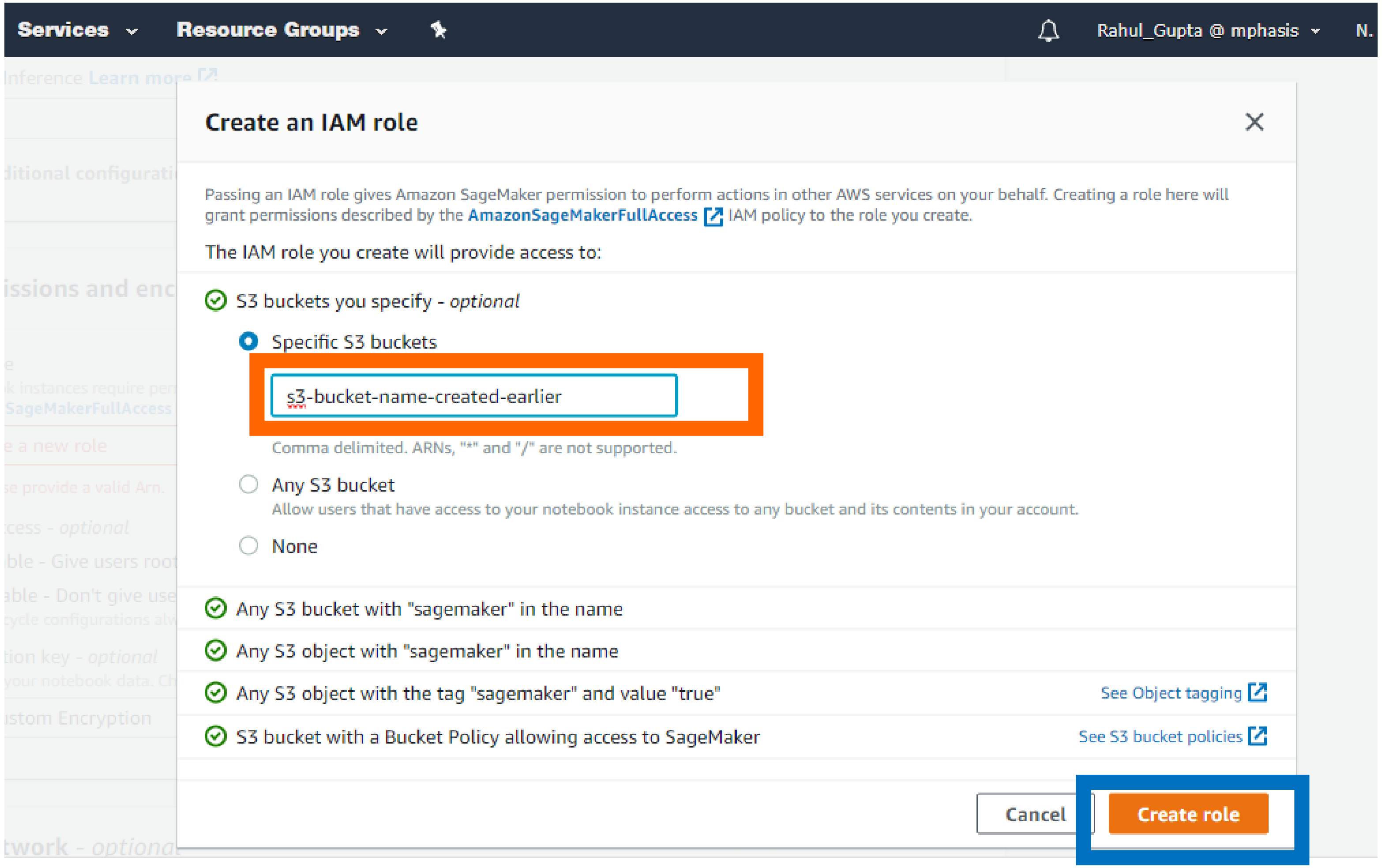
Create notebook instance





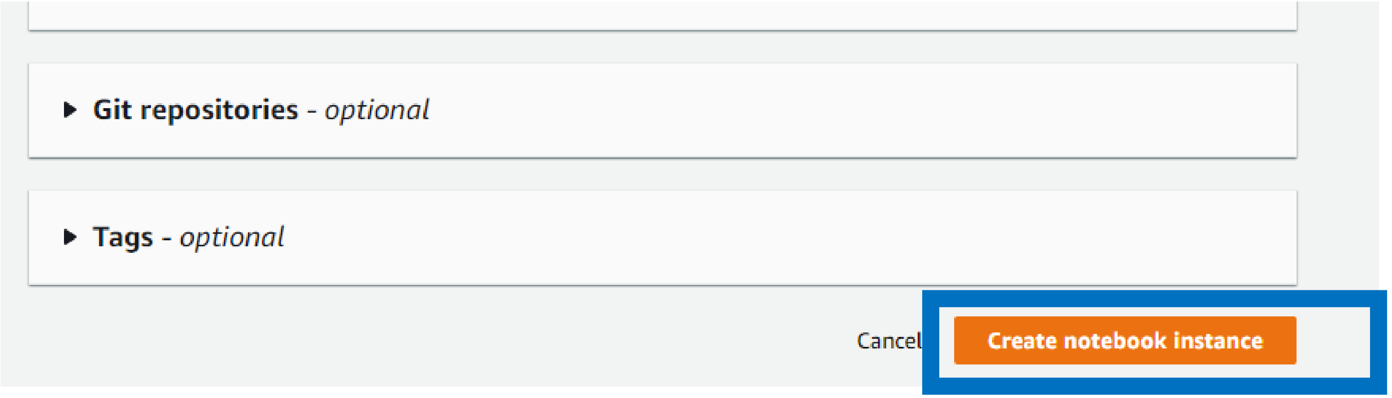
Specify a name for Notebook instance →
Select **Create a new role** from **Permissions and encryption** dropdown

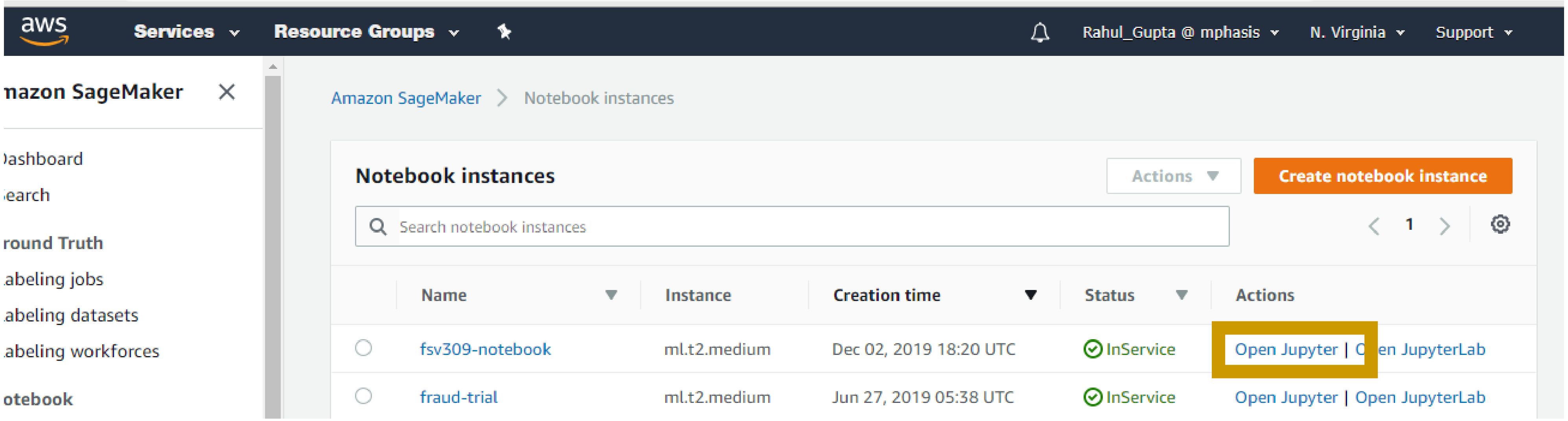
Create notebook instance



Specify name of S3 bucket created earlier→
Click on **Create role**

Create notebook instance

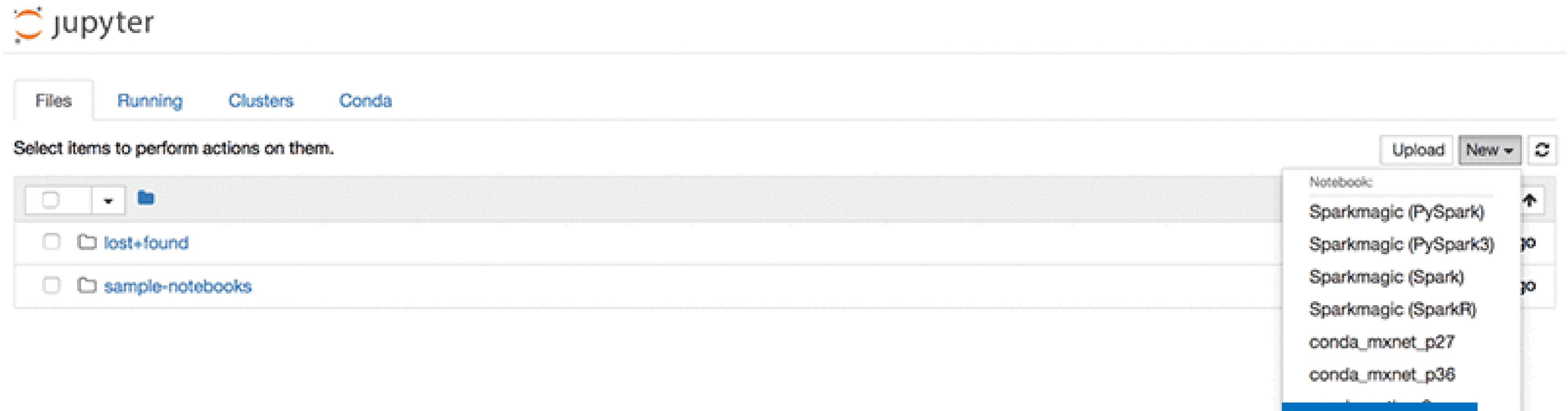




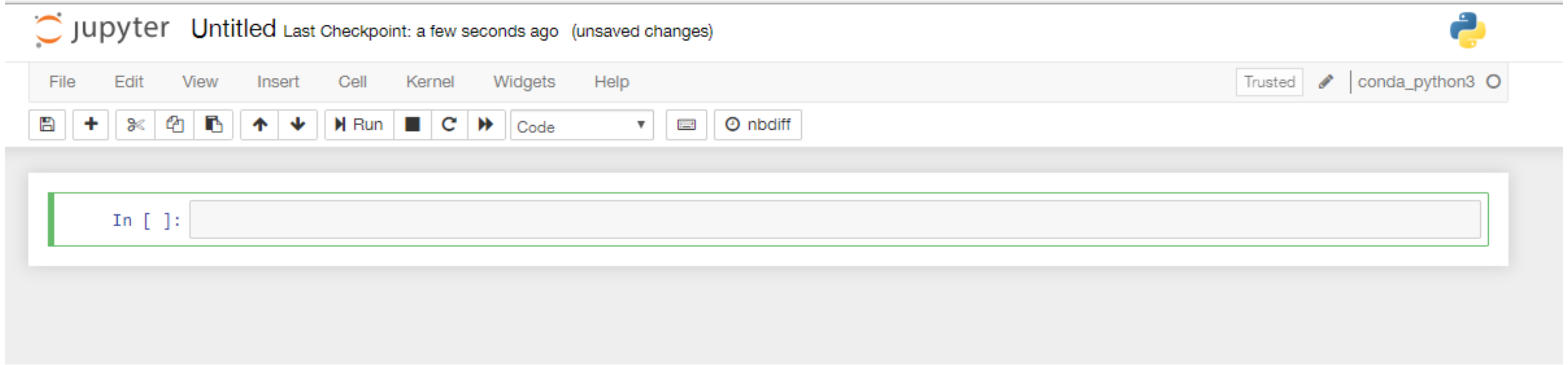
Click on **Create notebook instance**

→ Once instance's Status is **InService**, Click on **Open Jupyter**

Create notebook instance



On the new page, Select New dropdown and Click on **conda_python3**



Thank You

- In case of any doubt/queries, please feel free to reach out to us at below contact details:

Saurabh Singh: saurabh.singh07@mphasis.com
Sai Barath Sunder: sai.sundar@mphasis.com
Rahul Gupta: rahul.gupta@mphasis.com

THANK YOU

About Mphasis

Mphasis (BSE: 526299; NSE: MPHASIS) applies next-generation technology to help enterprises transform businesses globally. Customer centricity is foundational to Mphasis and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized (C=X2C$^{2}_{TM}$=1) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization are key to building strong relationships with marquee clients. Click here to know

Important Confidentiality Notice