AI-ML

AWS AI-ML

GOOGLE AI-ML

AZURE AI-ML

JUPYTER NOTEBOOK ENVIRONMENTS

AI-ML

Artificial Intelligence and Machine Learning (AI/ML) frameworks and architectures

AWS AI-ML

AWS AI-ML
GOOGLE AI-ML
AZURE AI-ML
JUPYTER NOTEBOOK ENVIRONMENTS

SAGEMAKER INTRODUCTION

SAGEMAKER OVERVIEW

SAGEMAKER ALGORITHMS

SAGEMAKER API

WHAT IS AMAZON SAGEMAKER?

Fully managed machine learning service

Build and training machine learning models

Deploy models into production hosted environment

WHAT DOES SAGEMAKER PROVIDE?

Development : Jupyter authoring notebook instance

- For data sources exploration and analysis

Deployment : Easy deployment at production scale

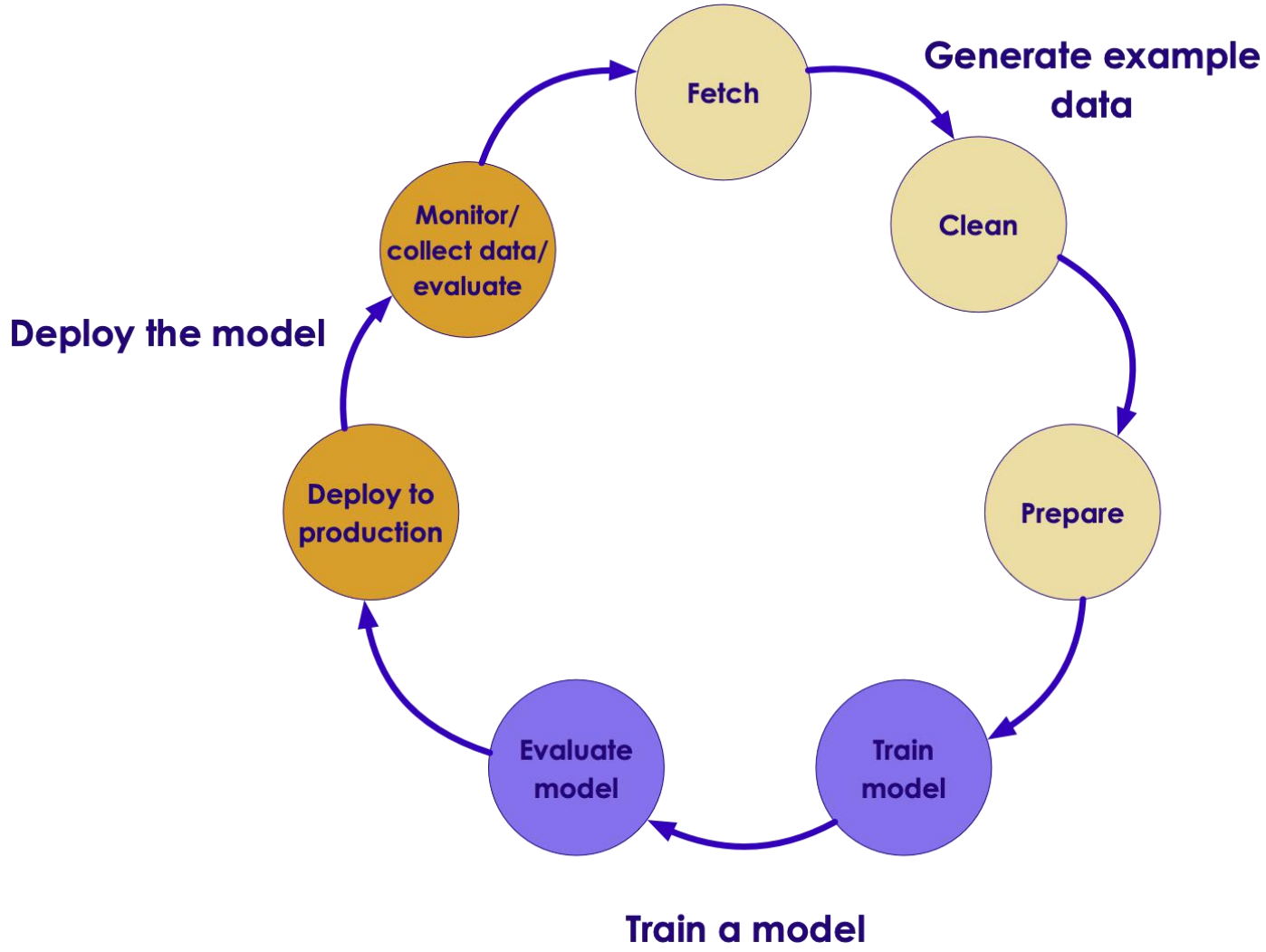
Common machine learning algorithms

- Optimized for large data sets in distributed mode

Optional bring-your-own algorithm

Billing: by the seconds

HOW DOES SAGEMAKER WORK?



STEP 2: TRAIN THE MODEL

First choose an algorithm

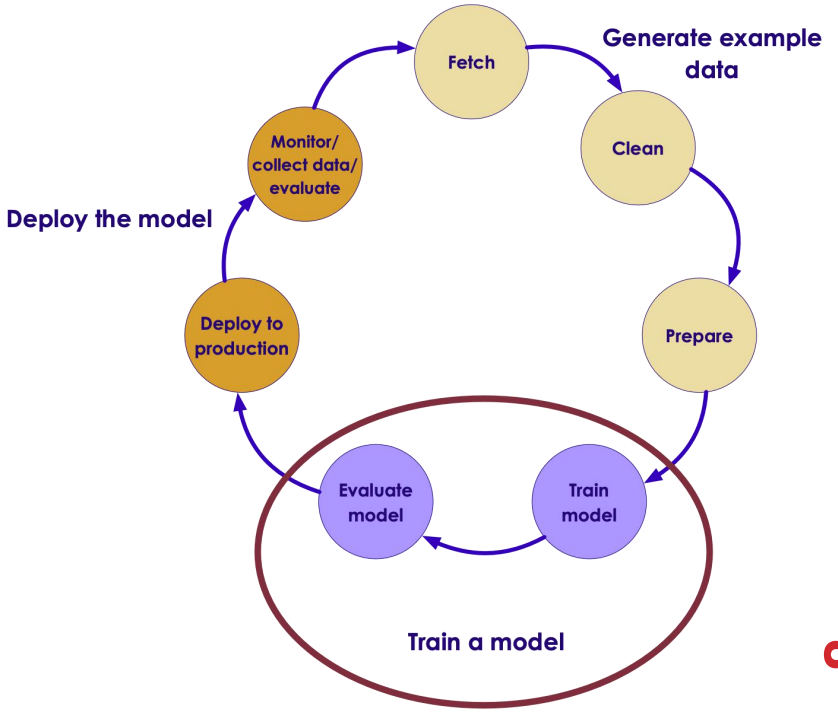
- Depends on data
- Can use algorithms provided by SageMaker
- Or bring your own

Training can consume lot of resources (CPU, GPU, TPU - Trainium when available, Memory, multiple machines)

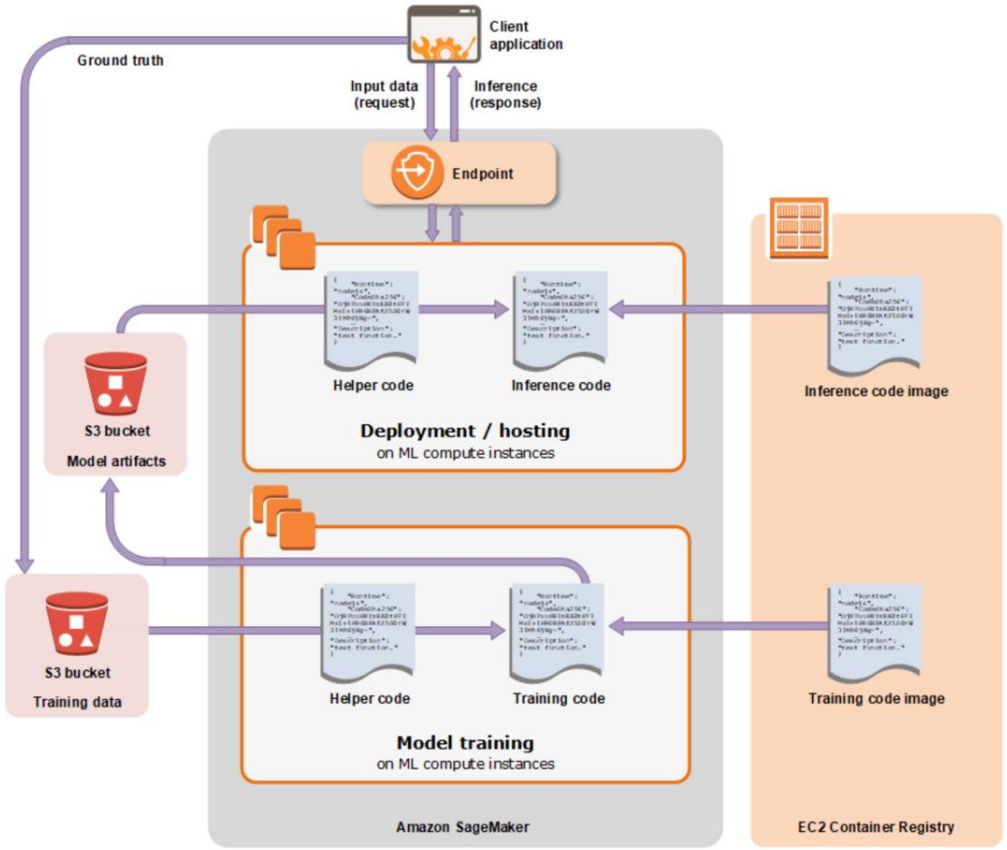
- SageMaker provides elastic scalability

Evaluating the Model:

- Can use Jupyter notebook
- Or built in SageMaker APIs



TRAINING A MODEL WITH SAGEMAKER



KERNELS SUPPORTED BY SAGEMAKER

SageMaker native algorithms

Spark

- PySpark, PySpark3, SparkR

Anaconda Python

- python2, python 3

Custom Deep Learning Python code

- TensorFlow (conda_tensorflow)
- Apache MXNet (conda_mxnet)

Your own algorithms

- Provided as Docker images

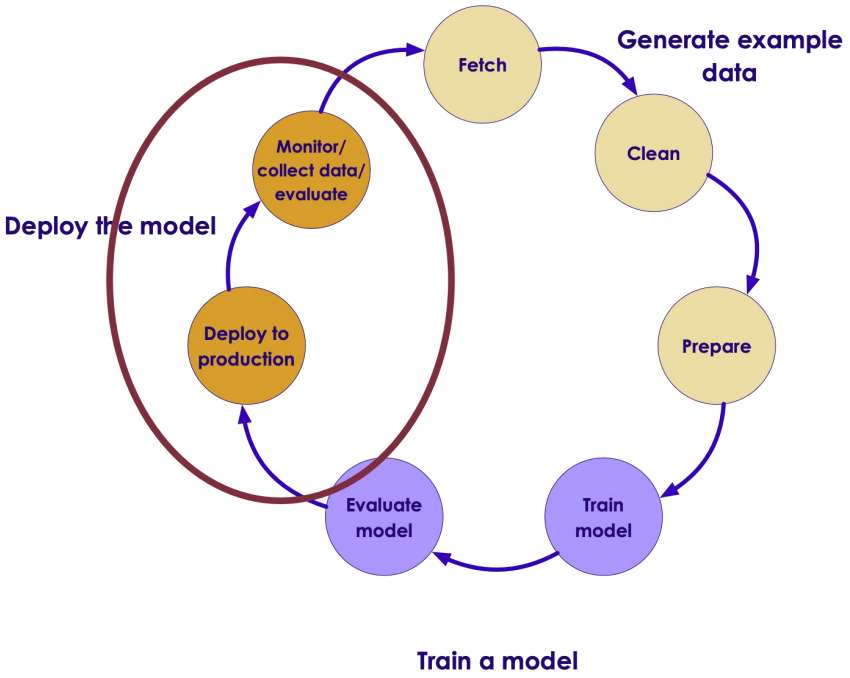
STEP 3: DEPLOY THE MODEL

Once the evaluation goes well, we can deploy the model to production:

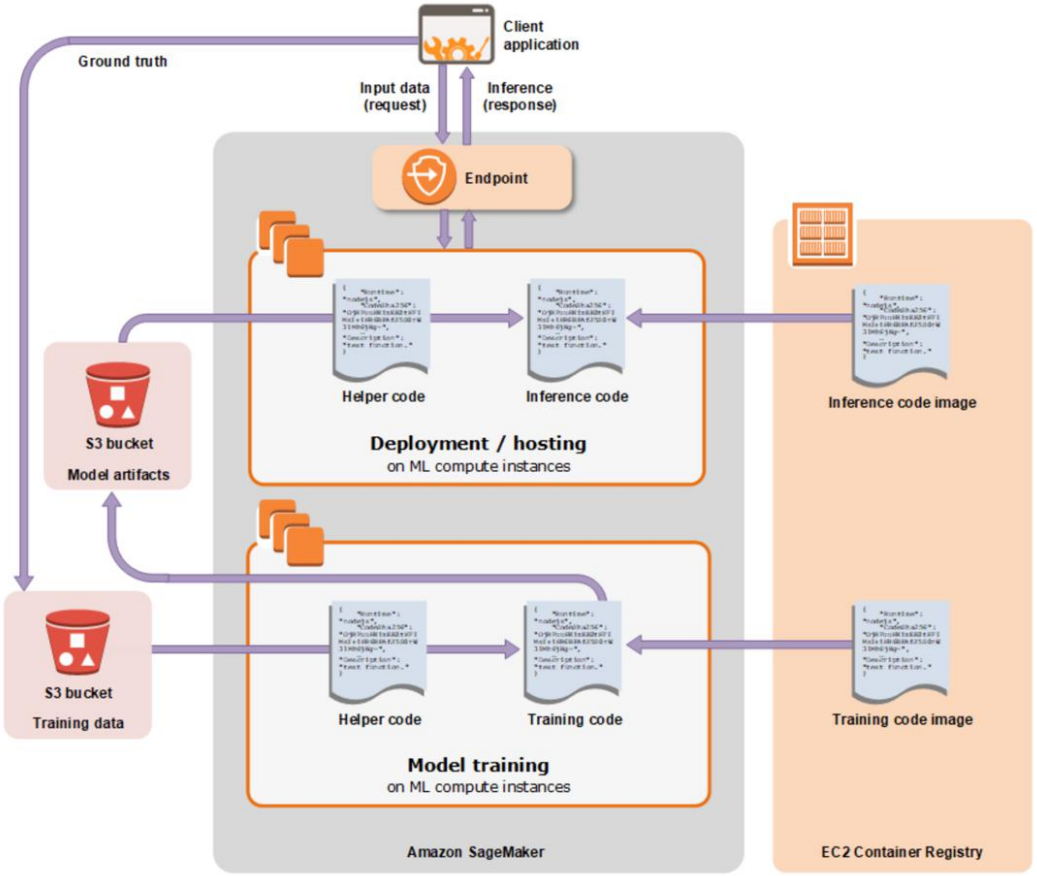
Continuous cycle:

- Monitor performance of Model
- train it with more data
- Deploy again

SageMaker makes this process easy



DEPLOYING A SAGEMAKER MODEL



GETTING STARTED WITH SAGEMAKER

Start by using the SageMaker console

Find an example Jupyter notebook and adapt it

Create new training models

Use Python or AWS SDK

Integrate SageMaker into Apache Spark pipeline

USING SPARK

Spark is a modern distributed processing engine

Features:

- Supports multiple languages Scala, Java, Python, R
- Scales to Big Data

SageMaker supports Spark

- Spark is one of the supported kernels
- Run your Spark applications in SM
- SageMakerModel extends the org.apache.spark.ml.
- SageMaker handles the data transfers

Amazon SageMaker Apache Spark Library -(https://github.com/aws/sagemaker-spark)

MONITORING AMAZON SAGEMAKER

Amazon CloudWatch monitors your AWS resources and applications

Amazon CloudWatch Logs enables you to monitor, store, and access your log files from EC2 instances, AWS CloudTrail

- That includes logs from SageMaker Docker containers

AWS CloudTrail captures API calls and related events made by or on behalf of your AWS account

- Including SageMaker calls

BUILT-IN ALGORITHMS (DIFFICULTY=EASY)

Example	What it is called	What it does
factorization_machines_mnist	factorization-machines	SVM, improved with FM
imageclassification_caltech	image-classification	Neural network full training or transfer knowledge
lda_topic_modeling	lda	LDA
linear_learner_mnist	linear-learner	Logistic regression Linear regression
ntm_synthetic	ntm	NTM - Neural Topic Modeling, cf. LDA

BUILT-IN ALGORITHMS (DIFFICULTY = EASY)

Example	What it is called	What it does
pca_mnist3	pca	PCA
seq2seq_translation	seq2seg	Machine translation Based on MXNet
xgboost_abalone (continuous values)	xgboost	Gradient Boosting Improved Random Forests Learns from previous mistakes
xgboost_mnist (classifier)	Same	Same

APPLICATIONS (DIFFICULTY = MEDIUM)

Example	What it is called	What it does
breast_cancer_prediction	linear-learner	Multi-model
ensemble_modeling	linear-learner XGBoost	Ensemble
linear_time_series_forecast	linear-learner	Adjusted for seasonality
video_games_sales_xgboost	XGBoost	Combine data from different review sources
xgboost_direct_marketing_sagemaker	XGBoost	Dealing with unbalanced classes Cleaning the data to improve model performance

SAGEMAKER ADVANCED (ADVANCED)

Example	What it is called	What it does
data_distribution_types	linear-learner	Scaling out Multiple buckets
handling_kms_encrypted_data	XGBoost	Data is encrypted KMS = Key Management Service
kmeans_bring_your_own_model	Kmeans in Python	Model conversion and Deployment
scikit_bring_your_own	A Python algorithm	Creates a Docker container Registers it with SageMaker
tensorflow_distributed_mnist	TensorFlow In Distributed Mode	Creating a job that is running a TensorFlow model

SAGEMAKER APIS / SDKS

Main (RESTful)

- https://docs.aws.amazon.com/sagemaker/latest/dg/API_Reference.html

Python - High Level Python SDK

- https://github.com/aws/sagemaker-python-sdk

SparkScala

- https://github.com/aws/sagemaker-spark

Amazon SageMaker Apache Spark Library

- https://github.com/aws/sagemaker-spark

LINEAR LEARNER

SageMaker Linear Learner is a supervised algorithm

Solves regression or classification problems

Input

- (X,Y), where
- X - multidimensional vector
- Y - numeric label

Input formats

- recordIO wrapped protobuf
- CSV

LINEAR LEARNER HYPERPARAMETERS

Parameter Name	Description
feature_dim	Number of features in input data. Required. Valid values: positive integer Default value: -
predictor_type	Whether the target variable is binary classification or regression. Required. Valid values: binary_classifier or regressor Default value: -

https://docs.aws.amazon.com/sagemaker/latest/dg/ll_hyperparameters.html

REVIEW QUESTIONS

What is SageMaker

What languages does SageMaker provide

- Scala? Java? Python? Go?

SageMaker billing is by

- Second? Minute? Hour?

What do you need to clean up?

- Instances? Redshift? Kinesis? S3?

GOOGLE AI-ML

AWS AI-ML
GOOGLE AI-ML
AZURE AI-ML
JUPYTER NOTEBOOK ENVIRONMENTS

USING GOOGLE ML

Use pre-trained machine learning (ML) models to add intelligence to your applications

Use pre-trained ML models

Vision API

Speech API

Video Intelligence API

Translation API

Natural Language API

Use your own data to train models

TensorFlow

Cloud AI Platform

LOOK, MA, NO MACHINE LEARNING KNOWLEDGE!

Image (GCS/embedded)

JSON request

Invoke Vision API

The Vision API can work off an image in Cloud Storage or embedded directly into a POST message. I'll use C

. That photograph is from http://www.publicdomainpictures.net/view-image.php?image=15842

```
# Running Vision API
import base64
IMAGE="gs://cloud-training-demos/vision/sign2.jpg"
vservice = build('vision', 'v1', developerKey=APIKEY)
request = vservice.images().annotate(body={
        'requests': [{
                'image': {
                    'source': {
                        'gcs_image_uri': IMAGE
                    }
                },
                'features': [{
                    'type': 'TEXT_DETECTION',
                    'maxResults': 3,
                }]
            }],
        })
responses = request.execute(num_retries=3)
print responses
```

```
{u'responses': [{u'textAnnotations': [{u'locale': u'zh', u'description': u'\u8bf7\u6
```

JSON response

EXAMPLES

Analyze images

Label Detection

Optical character recognition (OCR)

Landmark Detection

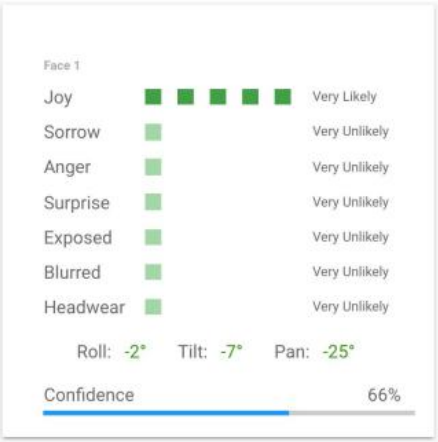
Logo Detection

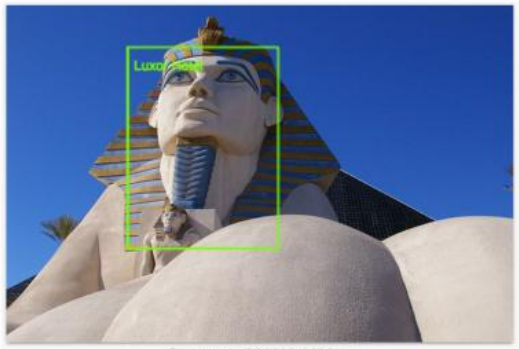
Face Detection

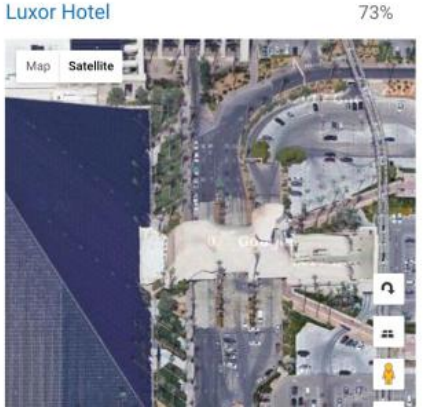
Explicit Content Detection

IMAGES

Get insight from images







QUIZ

You are developing an application that tags all surveillance video before storing the files. Which API should you use?

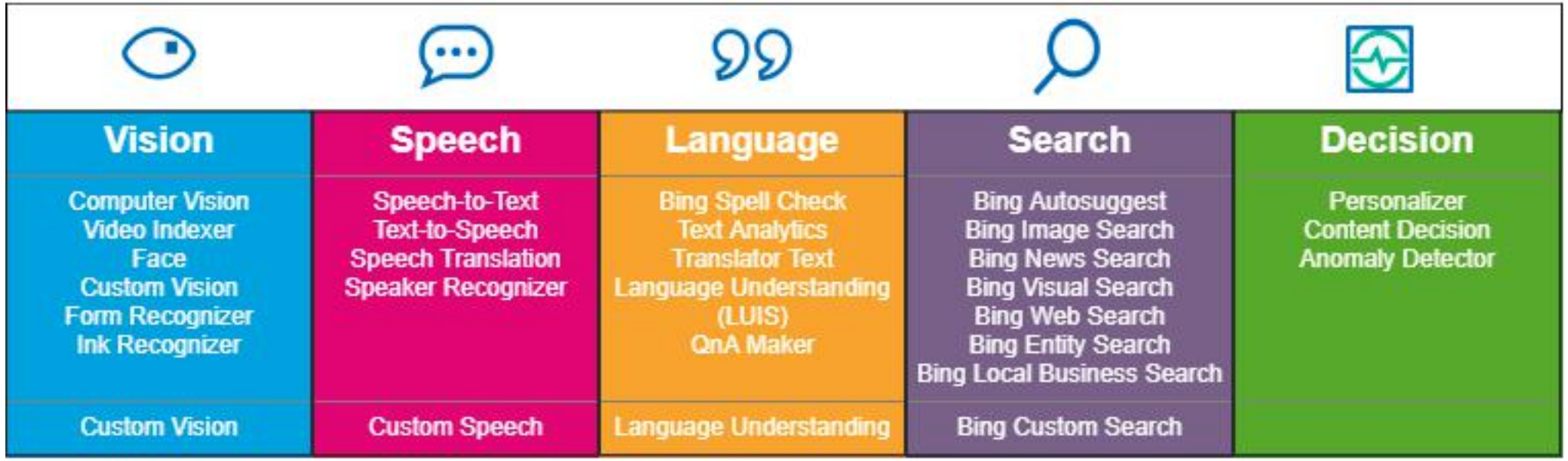
- A. Vision API
- B. Speech API
- C. Video Intelligence API
- D. Translation API

AZURE AI-ML

AWS AI-ML
GOOGLE AI-ML
AZURE AI-ML
JUPYTER NOTEBOOK ENVIRONMENTS

MICROSOFT AI APPROACH

Prebuilt AI: Cognitive Services

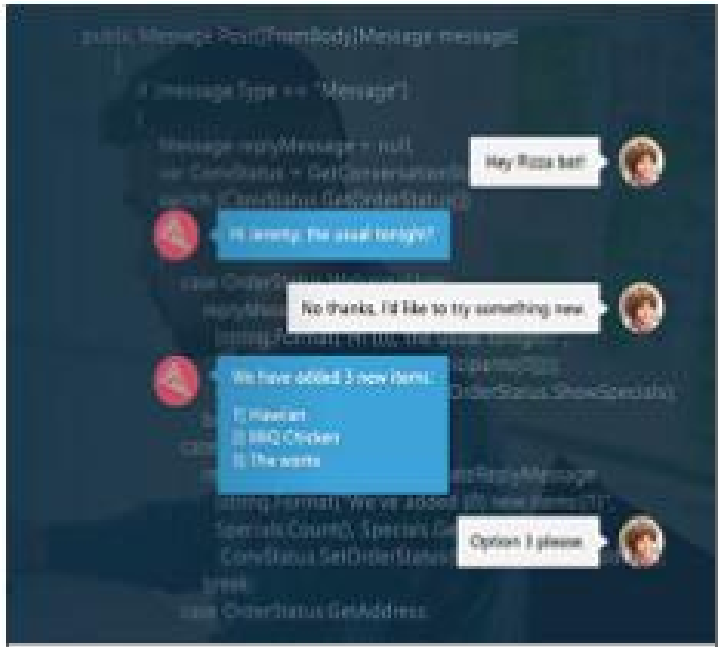


Vision	Speech	Language	Search	Decision
Computer Vision Video Indexer Face Custom Vision Form Recognizer Ink Recognizer	Speech-to-Text Text-to-Speech Speech Translation Speaker Recognizer	Bing Spell Check Text Analytics Translator Text Language Understanding (LUIS) QnA Maker	Bing Autosuggest Bing Image Search Bing News Search Bing Visual Search Bing Web Search Bing Entity Search Bing Local Business Search	Personalizer Content Decision Anomaly Detector
Custom Vision	Custom Speech	Language Understanding	Bing Custom Search	

AZURE COGNITIVE SERVICE REAL WORLD

FACE

PERSONALIZER

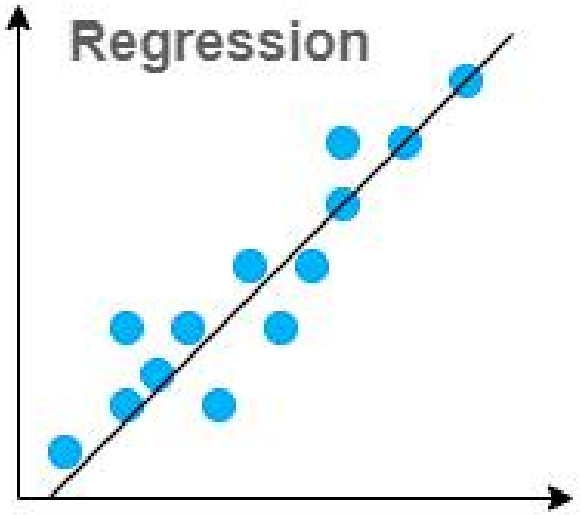


LANGUAGE UNDERSTANDING

SPECTRUM OF AI IMPLEMENTATIONS

ML CATEGORIES

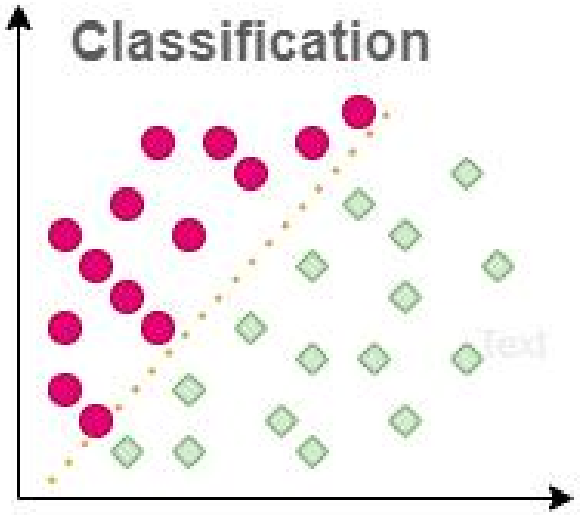
How many?



What will revenue be in Q3 in Latin America for product X?

How many salespeople will we have at the end of Q4?

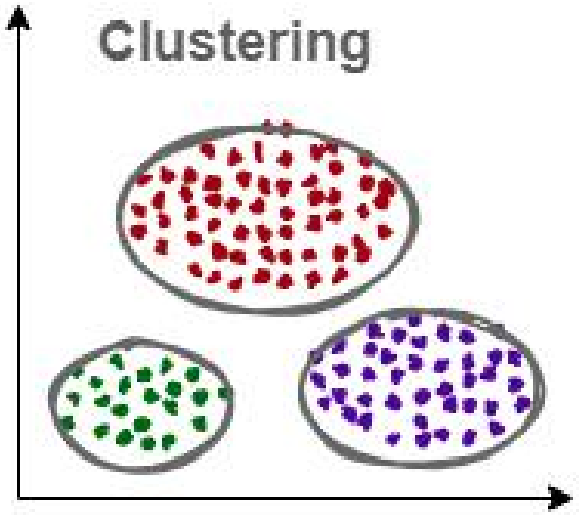
What Category?



What is a probability of a customer to purchase this insurance package?

What is a probability the customer will churn to a competitive product?

In what group?



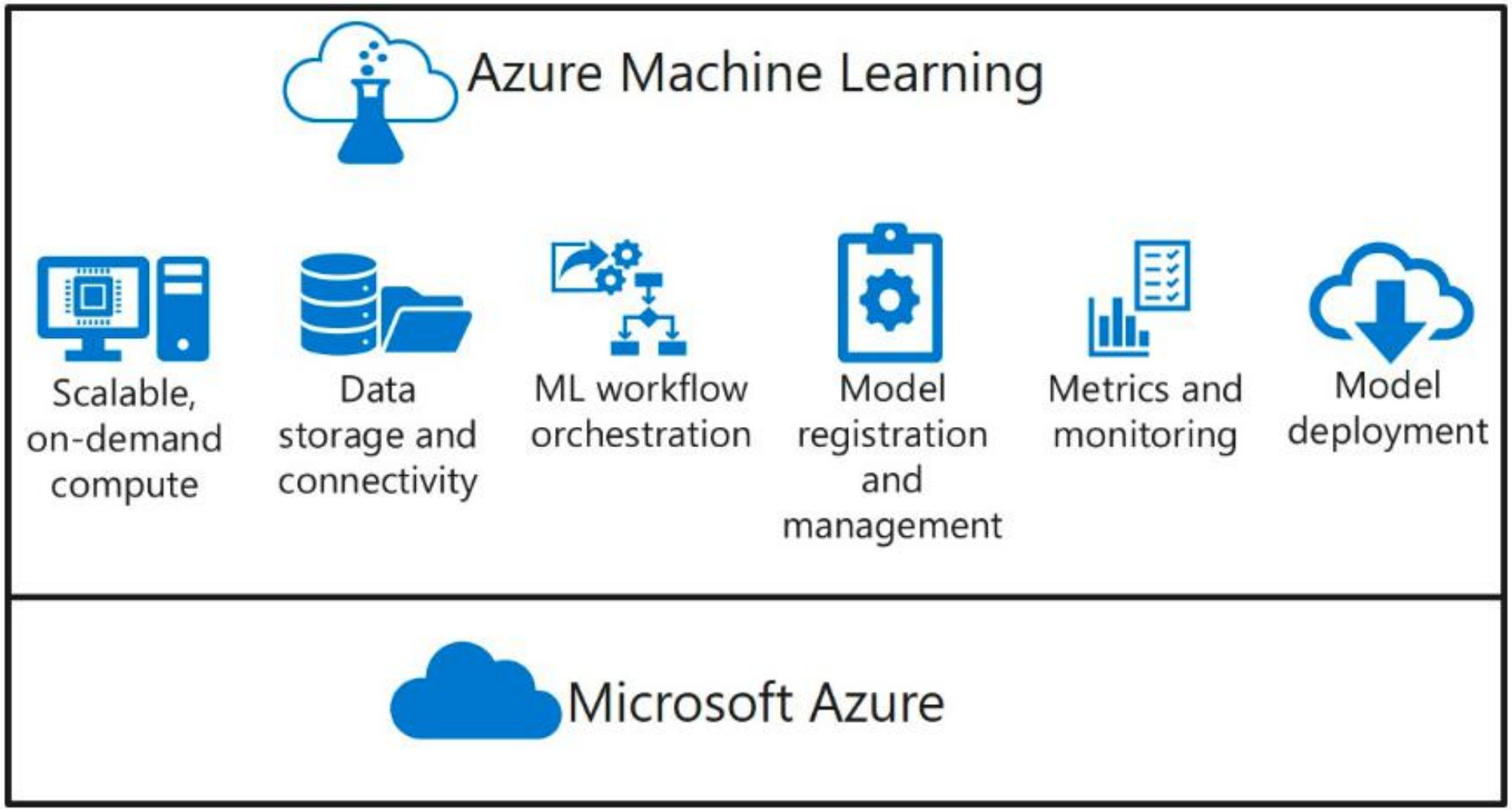
What group of patients have symptoms of this virus?

What are groups of customers with similar purchasing?

PREPARE - DEPLOY

All the logic that's behind deep learning can be accelerated on specialized hardware, like GPUs and FPGAs

AZURE MACHINE LEARNING



QUIZ

Which of the following descriptions accurately describes Azure Machine Learning?

- A. A Python library that you can use as an alternative to common machine learning frameworks like Scikit-Learn, PyTorch, and Tensorflow.
- B. A cloud-based platform for operating machine learning solutions at scale.
- C. An application for Microsoft Windows that enables you to create machine learning models by using a drag and drop interface.

QUIZ

You are using the Azure Machine Learning Python SDK to write code for an experiment. You must log metrics from each run of the experiment, and be able to retrieve them easily from each run. What should you do?

- A. Add print statements to the experiment code to print the metrics.
- B. Save the experiment data in the outputs folder.
- C. Use the log* methods of the Run class to record named metrics.

FIELD-PROGRAMMABLE GATE ARRAYS ON AZURE

What are FPGAs?

- FPGAs contain an array of programmable logic blocks, and a hierarchy of reconfigurable interconnects.
- The interconnects allow these blocks to be configured in various ways after manufacturing.
- Compared to other chips, FPGAs provide a combination of programmability and performance.

FPGAs make it possible

- achieve low latency for real-time inference (or model scoring) requests.
- Asynchronous requests (batching) aren't needed.
- Batching can cause latency, because more data needs to be processed. Implementations of neural processing units don't require batching;
- The latency can be many times lower, compared to CPU and GPU processors.

DOUG BURGER AZURE ARCHITECT ON FPGA

YouTube

https://www.youtube.com/watch?v=iJo_sSzioxM&t=2659s

JUPYTER NOTEBOOK ENVIRONMENTS

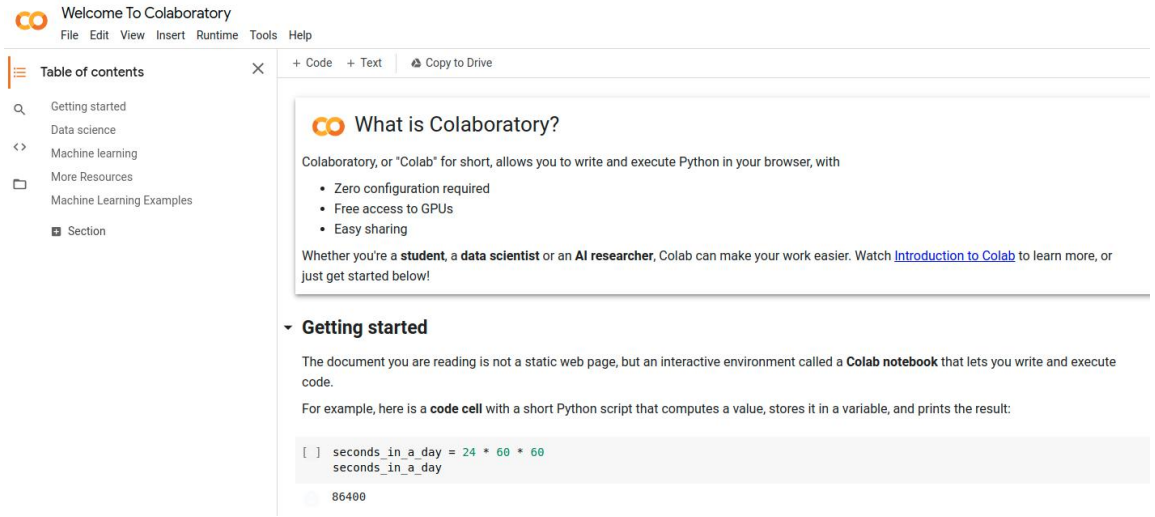
AWS AI-ML
GOOGLE AI-ML
AZURE AI-ML
JUPYTER NOTEBOOK ENVIRONMENTS

WHAT IS COLABORATORY?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing
- Whether you're a student, a data scientist or an AI researcher, Colab can make your work easier. Watch Introduction to Colab to learn more, or just get started below!

COLAB AT A GLANCE



COLAB CONTENTS

Getting started

Data science

Machine learning

More Resources

Machine Learning Examples

Let's do a demo!

KAGGLE AT A GLANCE

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment.

Access free GPUs and a huge repository of community published data & code.

Inside Kaggle you'll find all the code & data you need to do your data science work.

Use over 50,000 public datasets and 400,000 public notebooks to conquer any analysis in no time.

https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data

Host your book or research paper on Kaggle!

MORE REFERENCES

PyTorch, TF, and other stacks - in every cloud

Technical considerations to use ML at scale

https://d1.awsstatic.com/whitepapers/aws-power-ml-at-scale.pdf

CONGRATS ON COMPLETION