Develop and Deploy Deep Learning Microservices

Brendan Dwyer, Software Engineer
Karthik Muthuraman, Data Scientist
IBM Center for Open Source Data and AI technologies (CODAIT)

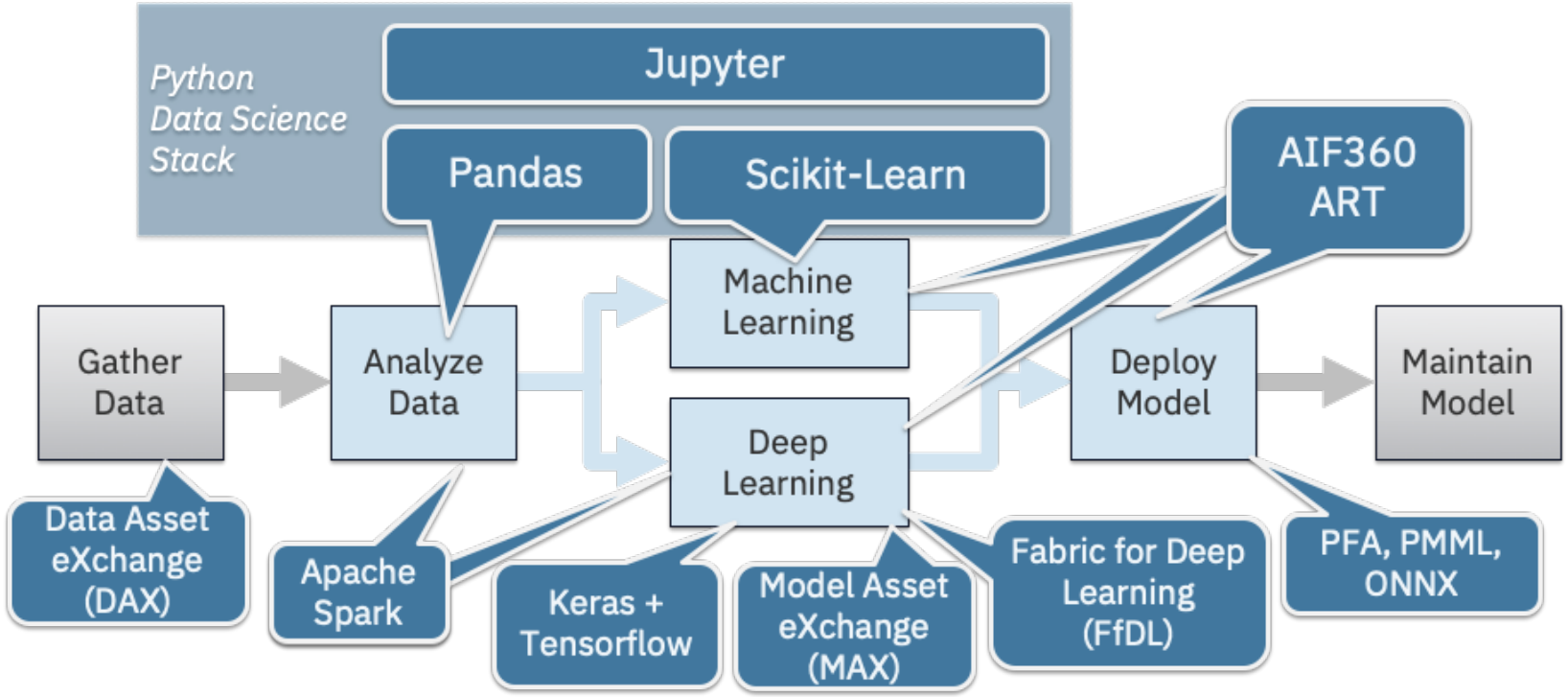
Center for Open Source Data & AI Technologies (CODAIT)

codait.org

Watson West Building
505 Howard St.
San Francisco, California

CODAIT aims to make AI solutions easier to create, deploy, and manage in the enterprise.

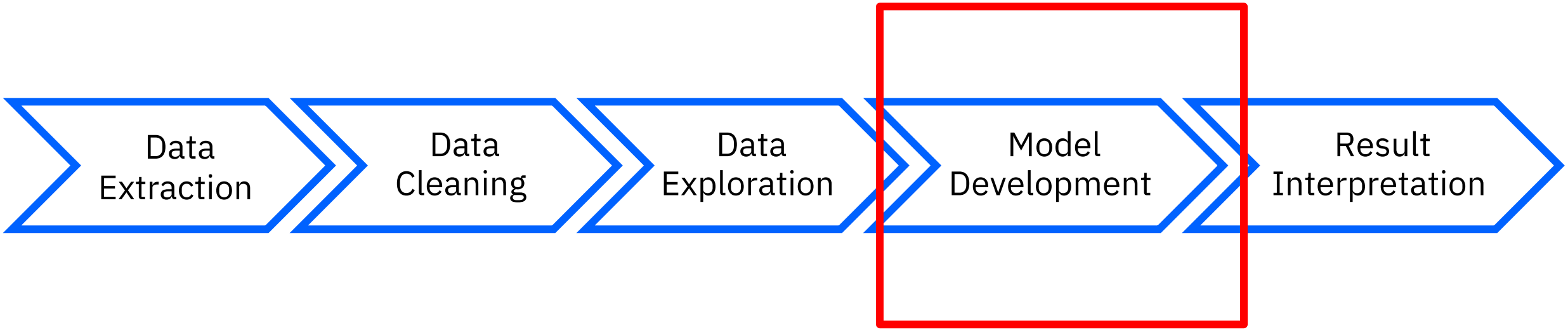
40+ open source developers!



What is Data Science?

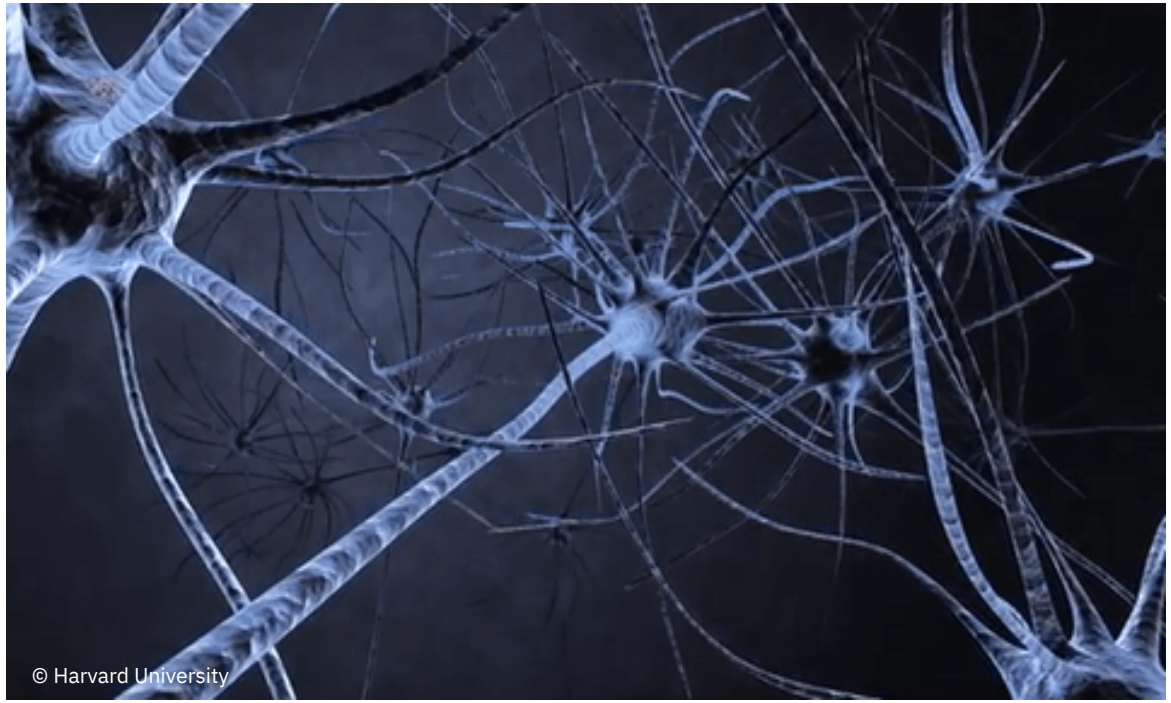
Goal	How?	Using?	Finally
Find a solution to the business problem	Transform into well posed questions	Mathematics, programming and scientific method	Communicate results and its business impact

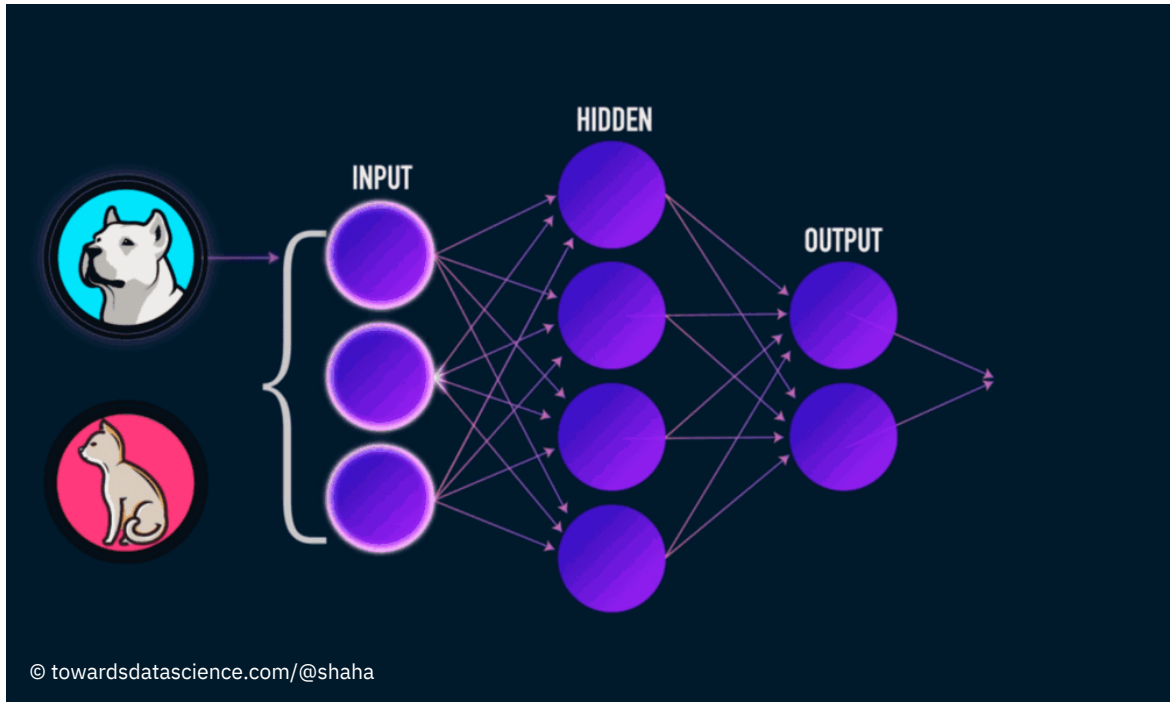
Data Science Pipeline



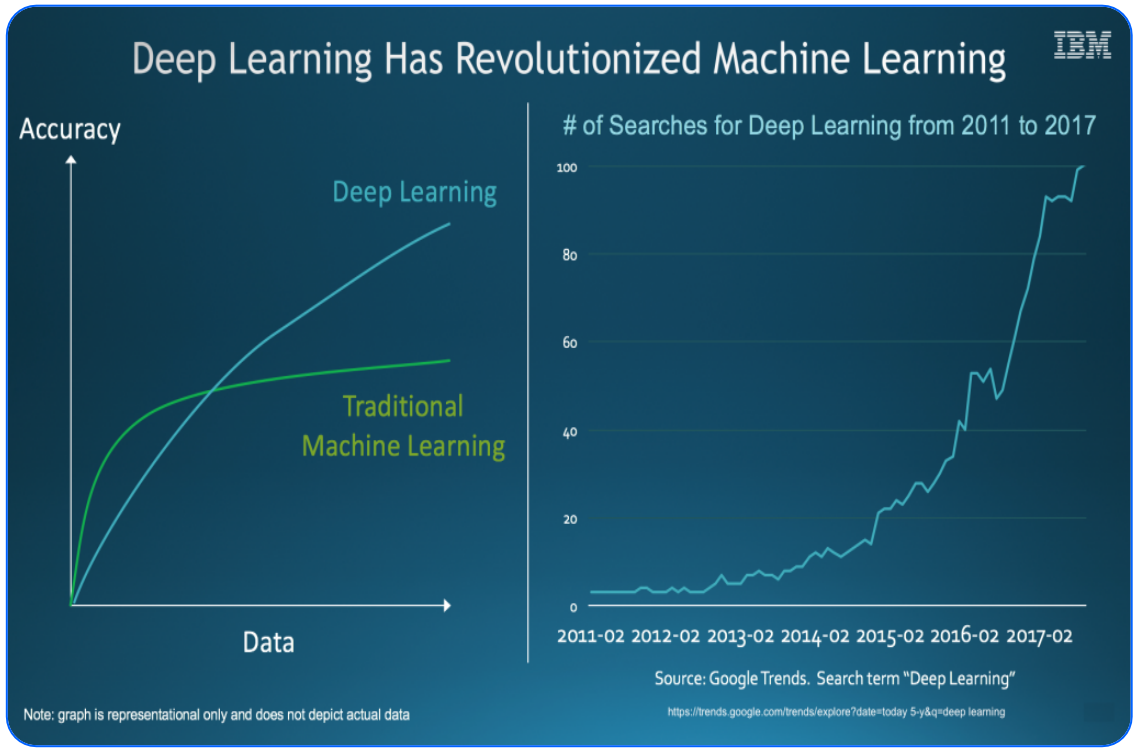
Deep Learning

Machine Learning with large deep neural networks





Why is Deep Learning so popular?



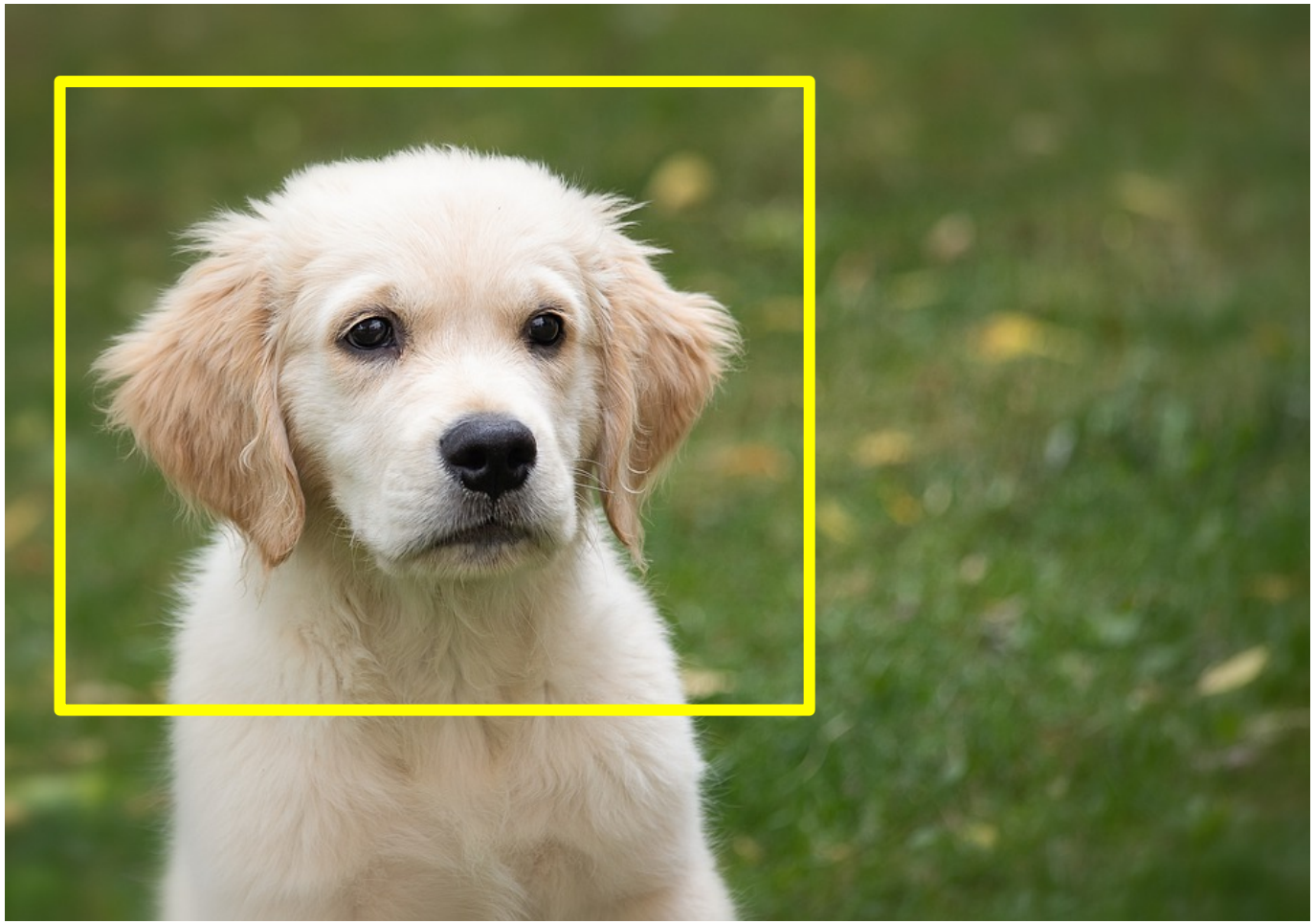
- Access to huge amount of data

- Powerful computers increases computational efficiency

- It is very good at supervised learning

- Automatically learns important features

What is in the picture?



Object Detection

Google Scholar

object detection deep learning models

Articles

About 563,000 results (**0.18** sec)

Any time
Since 2019
Since 2018
Since 2015
Custom range...

Sort by relevance
Sort by date

include patents

Deep learning

Y LeCun, Y Bengio, G Hinton - nature, 2015 - nature.com

… ConvNets were also experimented with in the early 1990s for **object detection** in natural images … applied with great success to the **detection**, segmentation and **recognition** of **objects** and regions … of biological images 54 particularly for connectomics 55 , and the **detection** of faces …

Cited by 15122 Related articles All 54 versions

Learning deep architectures for AI

Y Bengio - Foundations and trends® in Machine **Learning**, 2009 - nowpublishers.com

… For example, using knowledge of the 3D geometry of solid **objects** and lighting, we can … clear in the primate visual system [173], with its sequence of processing stages: **detection** of edges … appear

Find Model

... that does what you **need**

... that is **free** to use

... that is **performant** enough

Get Code + Cleanup

Many open source repos.

Research vs Production code

Code license?

Train

Multiple frameworks

- TensorFlow

- PyTorch

- Keras

Data License?

Deploy + Consume

- Adjust inference code

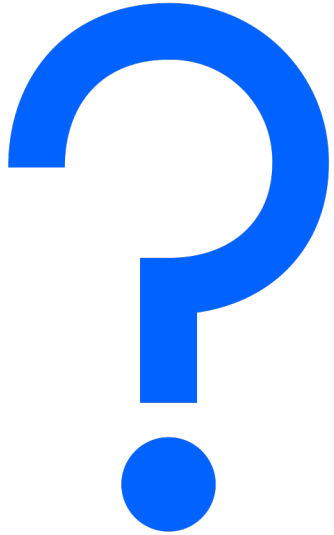
- Package inference code and model code, and pre-trained weights together

- deploy your package

Requires time, expertise, and resources

Developer

Learn deep learning to create an AI powered application?

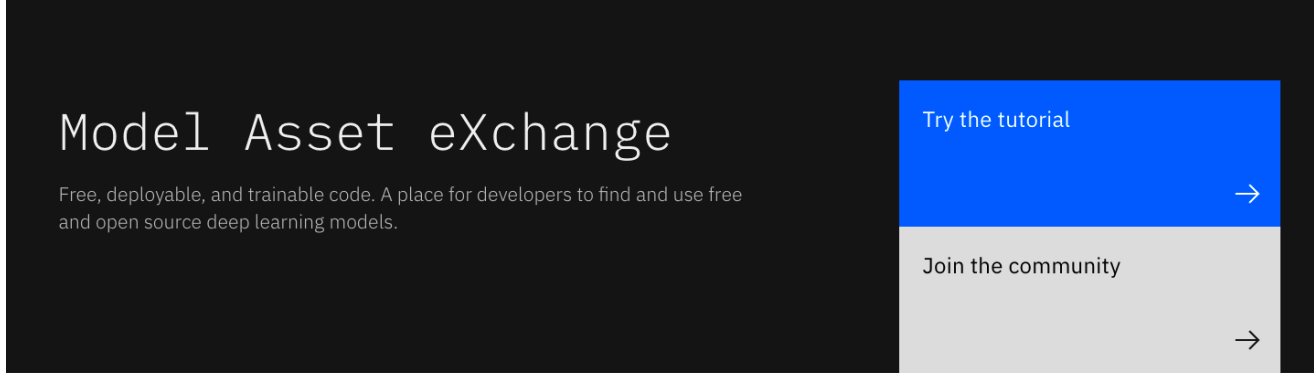


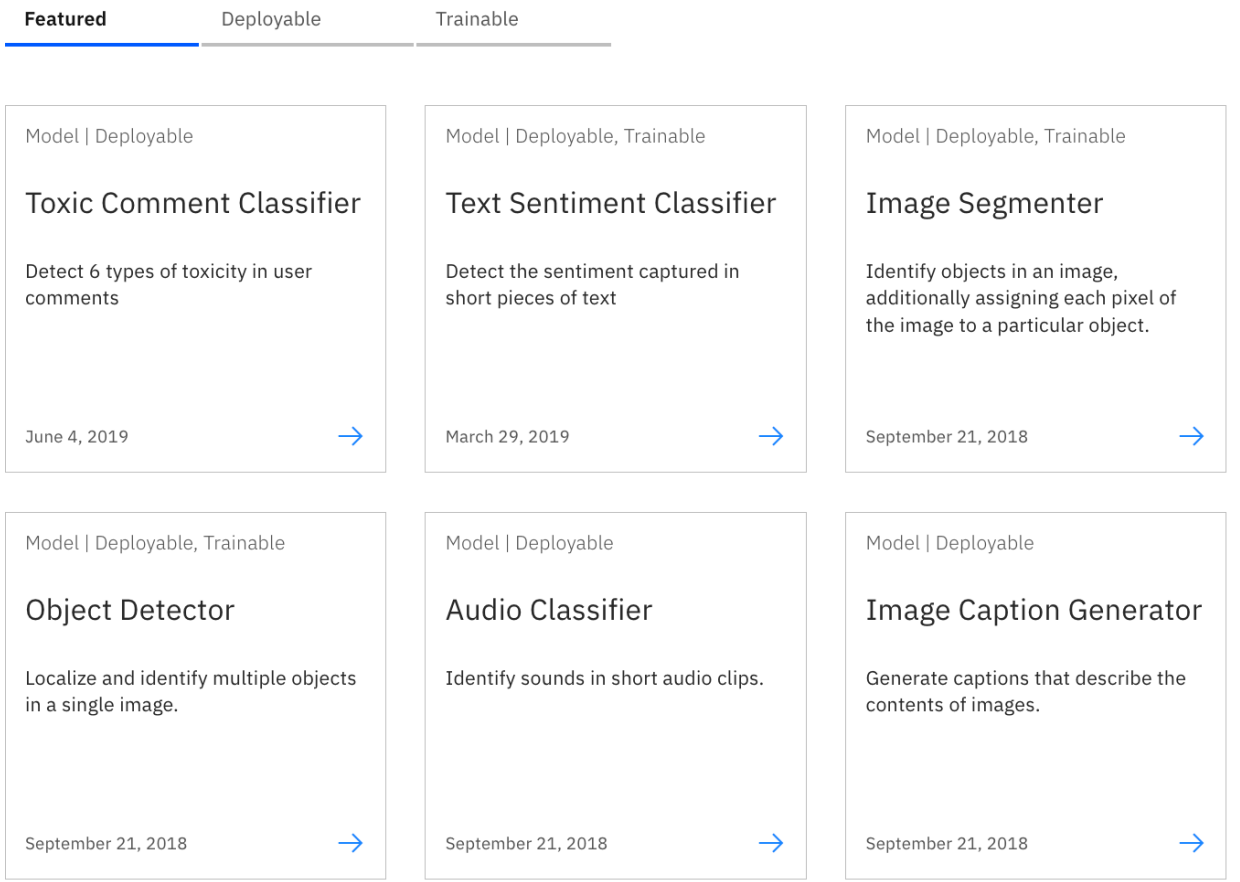
Data Scientist

Will the models be ready to use or train? How to deploy after training?

Model Asset eXchange (MAX)

ibm.biz/model-exchange



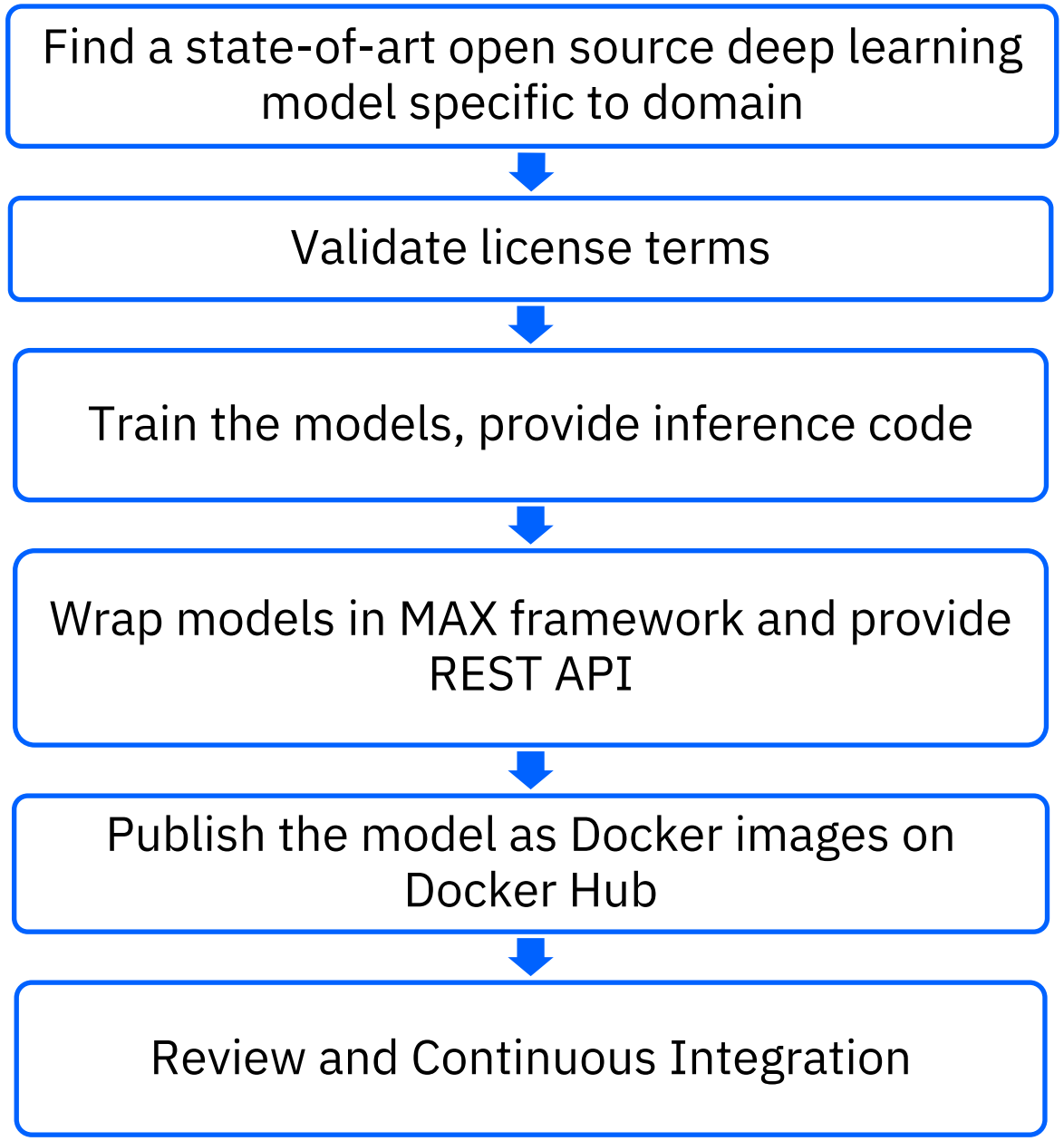


What is MAX?

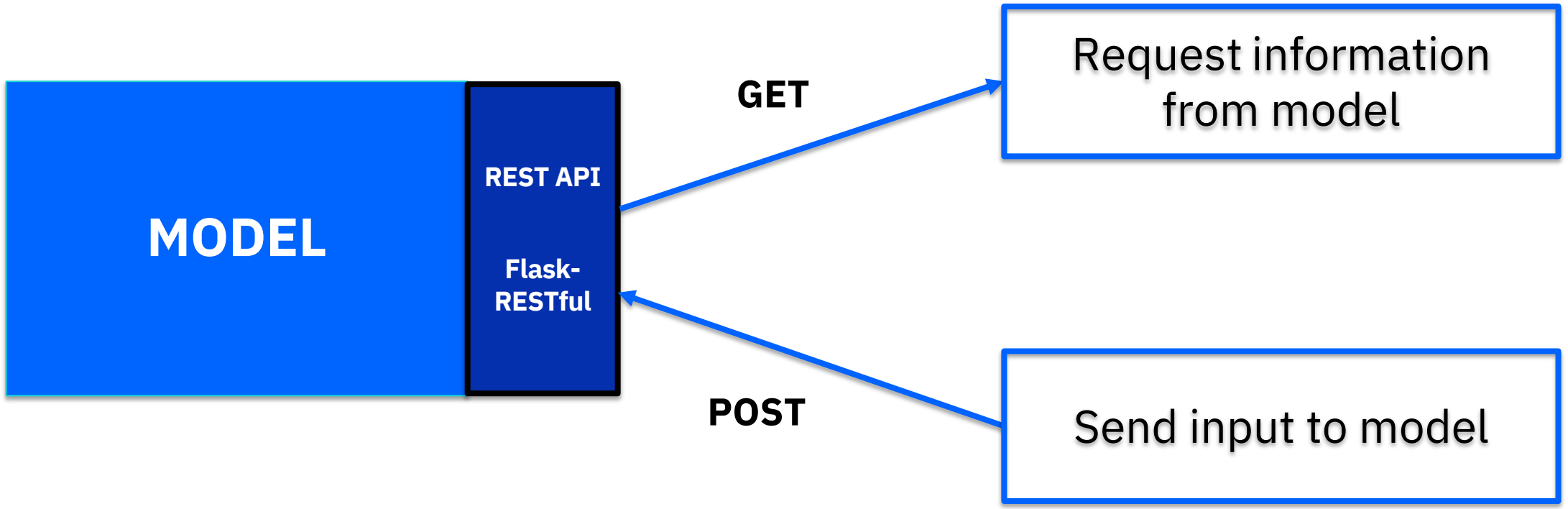
- One place for **state-of-art** open source deep learning models
- Wide variety of domains
- Tested code and IP
- **Free and open** source
- Both trainable and deployable versions
- Get started with 1 command:

docker run -it -p 5000:5000 codait/max-object-detector

Behind the Scenes



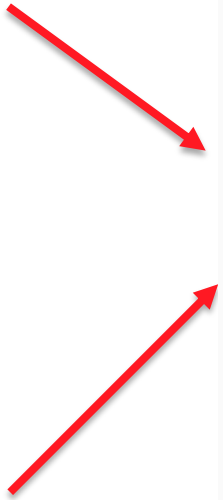
MAX Model Consumption – REST API



MAX Object Detector 1.3.0

[Base URL: /]
http://max-object-detector.max.us-south.containers.appdomain.cloud/swagger.json

Localize and identify multiple objects in a single image.



model Model information and inference operations

GET **/model/labels** Return the list of labels that can be predicted by the model

GET **/model/metadata** Return the metadata associated with the model

POST **/model/predict** Make a prediction given input data

Parameters Cancel

Name	Description
image * required **file** *(formData)*	An image file (encoded as PNG or JPG/JPEG) Browse... No file selected.
threshold **number** *(query)*	Probability threshold for including a detected object in the response in the range [0, 1] (default: 0.7). Lowering the threshold includes objects the model is less certain about. 0.7

Execute

Responses Response content type application/json

Code	Description
200	*Success* **Example Value** Model

```
{
  "status": "string",
  "predictions": [
    {
      "label_id": "string",
      "label": "string",
      "probability": 0,
      "detection_box": [
        0
      ]
    }
  ]
}
```

ibm.biz/serverlessNA2020

Model | Deployable, Trainable

Object Detector

Localize and identify multiple objects in a single image.

Get this model

Try the API

Try the web app

Try in a Node-RED flow

By IBM Developer Staff

Updated September 21, 2018 | Published March 20, 2018

Overview

This model recognizes the objects present in an image from the 80 different high-level classes of objects in the COCO Dataset. The model consists of a deep convolutional net base model for image feature extraction, together with additional convolutional layers specialized for the task of object detection, that was trained on the COCO data set. The input to the model is an image, and the output is a list of estimated class probabilities for the objects detected in the image. The model is based on the SSD Mobilenet V1 object detection model for TensorFlow.

Technologies (3)

Artificial intelligence | Deep learning | Visual recognition

Products & Services (3)

Model Asset Technologies (2)

MAX-Skeleton

- Template to create a deployable MAX model.
- Contains all the code scaffolding and imports MAX Framework.
- ibm.biz/max-skeleton

MAX-Framework

- A pip installable python library.
- Wrapper around flask
- Abstracts out all basic functionality of the MAX model into MAXApp and MAXApi abstract classes.

MAX and Serverless

ibm.biz/max-serverless

ibm.biz/serverlessNA2020

Tutorial

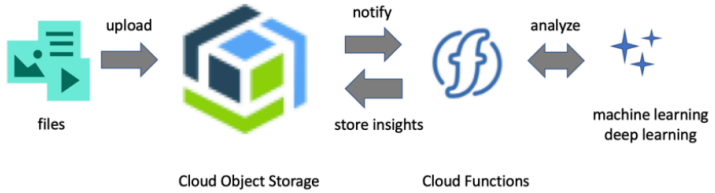
Leverage deep learning in IBM Cloud Functions

Automatically analyze the content of Cloud Object Storage buckets

By Patrick Titzler
Published October 14, 2019

Based on Apache OpenWhisk, IBM Cloud Functions is a Functions as a Service (FaaS) platform that makes it easy to build and deploy serverless applications.

In this tutorial, you'll build a serverless application using IBM Cloud Functions that monitors the content of a Cloud Object Storage bucket and analyzes the content of images that are uploaded to the bucket by a human or an automated process. For illustrative purposes, analysis is performed by a deep learning microservice from the Model Asset eXchange and analysis results are stored as JSON files in the same bucket.



You can easily adapt the outlined approach to take advantage of hosted cognitive services, such as those provided by IBM Watson®, and to store results in a NoSQL datastore like Cloudant® or a relational database.

Learning objectives

By completing this introductory tutorial, you learn how to monitor a Cloud Object Storage bucket for changes (new objects, updated objects, or deleted objects) using Cloud Functions and how to use deep learning microservices from the Model Asset eXchange to automatically analyze those objects in near real time.

Upon completion of the tutorial, you know how to use the IBM Cloud CLI to set up change monitoring for a Cloud Object Storage bucket and how to derive information from uploaded objects, such as images, audio, video or text files in near real-time using deep learning microservices.

Technologies (6)

Artificial intelligence | Deep learning | Machine learning | Microservices | Object Storage | Serverless

Products & Services (3)

Table of Contents

Learning objectives
Prerequisites
Download the tutorial artifacts from GitHub
Estimated time
Steps
Create a regional bucket
Configure your IBM Cloud CLI
Create a Cloud Object Storage trigger
Create an action
Create a rule
Test the serverless application
Extending the serverless application
Perform a different kind of analysis
Automatically remove analysis files
Support additional object types
Summary

Resources

- MAX on IBM Developer
  https://ibm.biz/model-exchange

- GitHub
  https://github.com/CODAIT/max-central-repo

- Learning path
  https://developer.ibm.com/series/create-model-asset-exchange/

- MAX ServerLess tutorial (IBM Cloud Functions)
  https://ibm.biz/model-serverless

@ibmcodait

@codait

ibm.biz/max-slack

Thank You!

ibm.biz/serverlessNA2020