Develop and Deploy Deep Learning Microservices

Karthik Muthuraman, Software Engineer
Saishruthi Swaminathan, Developer Advocate
IBM Center for Open Source Data and AI technologies (CODAIT)

IBM **Developer**

IBM

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

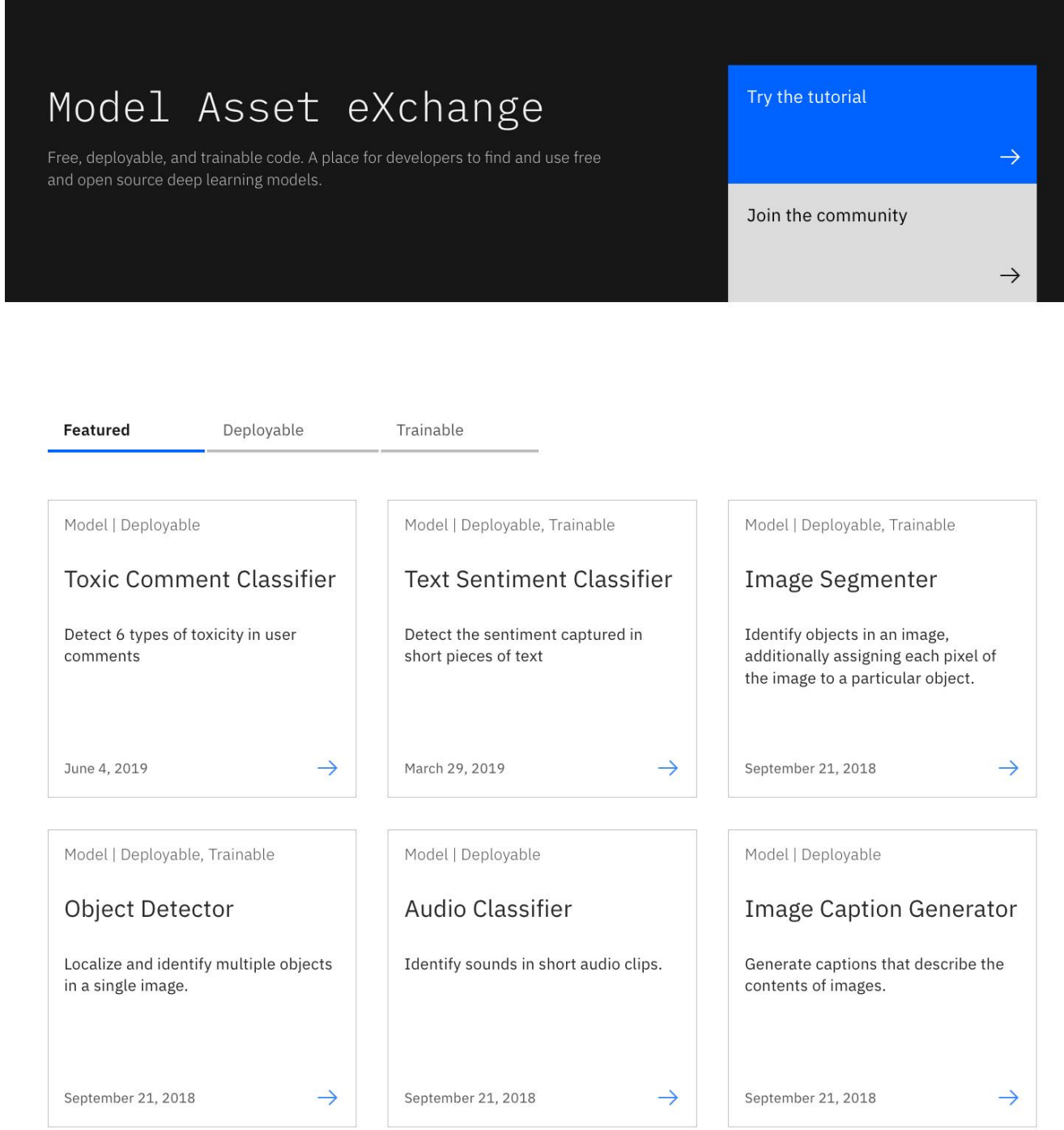
- Package inference code and model code, and pre-trained weights together

- deploy your package

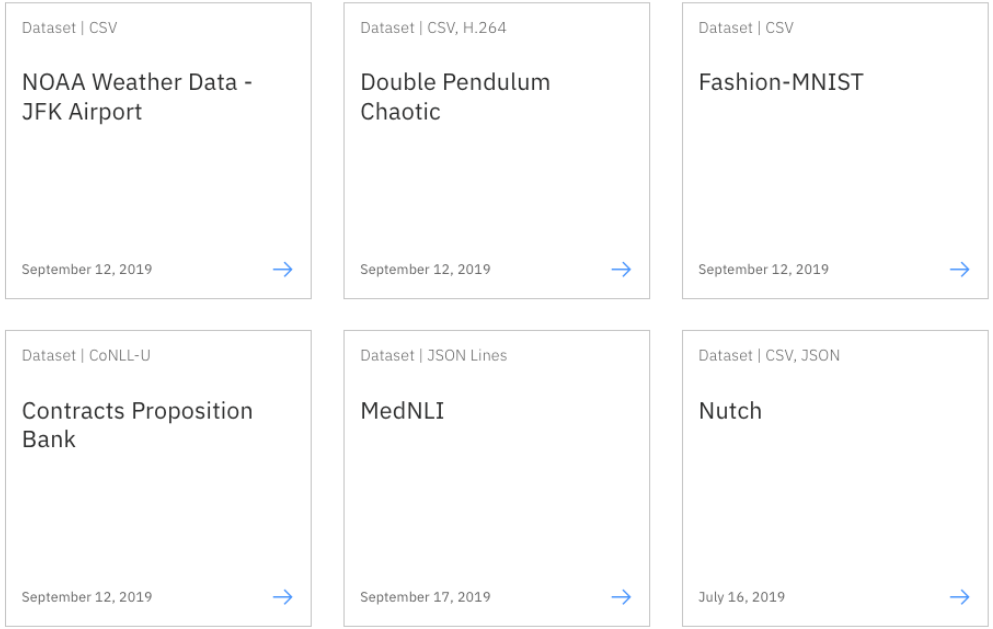
Requires time, expertise, and resources

Model Asset eXchange (MAX)

ibm.biz/model-exchange

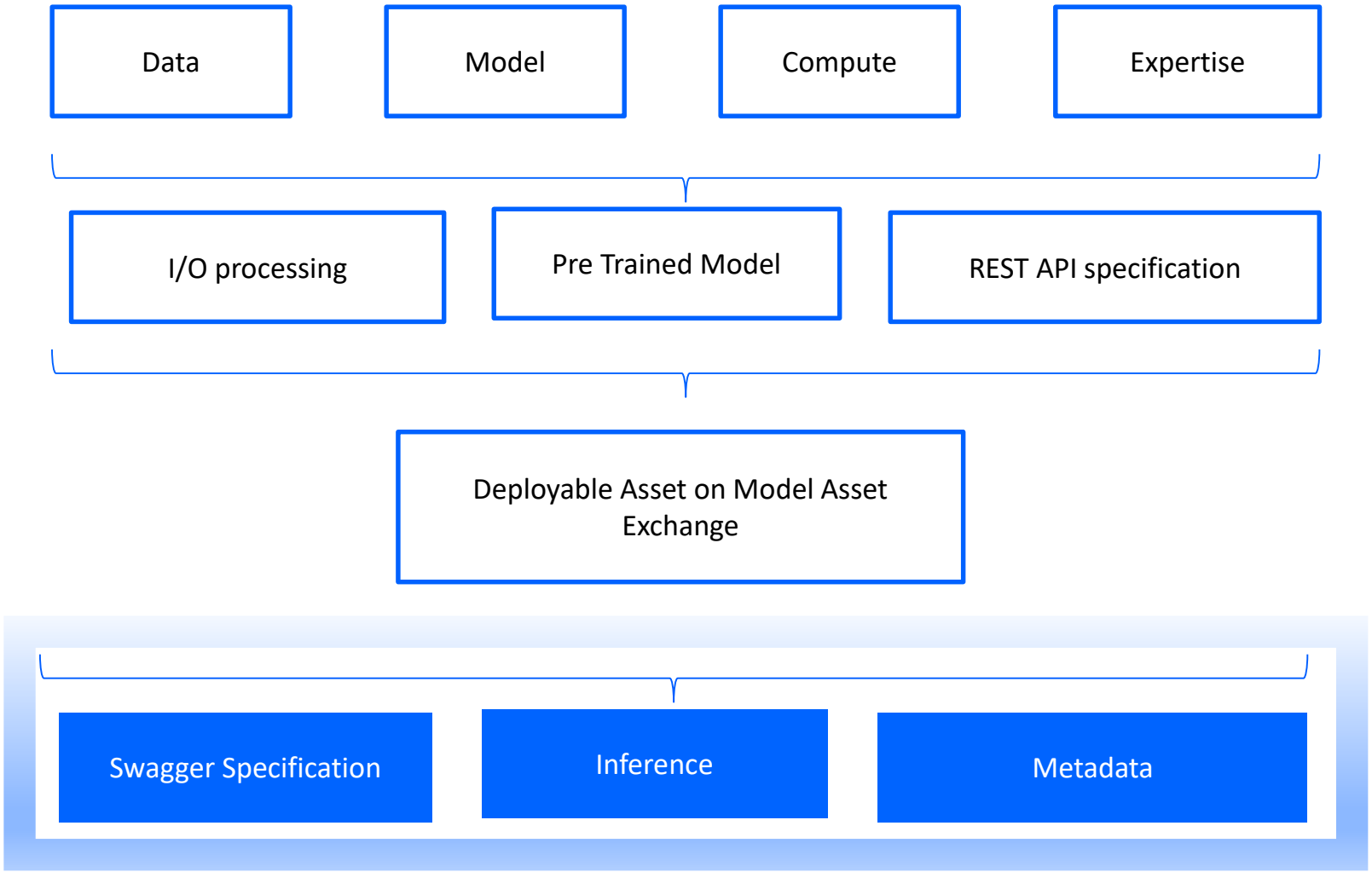


Data Asset Exchange

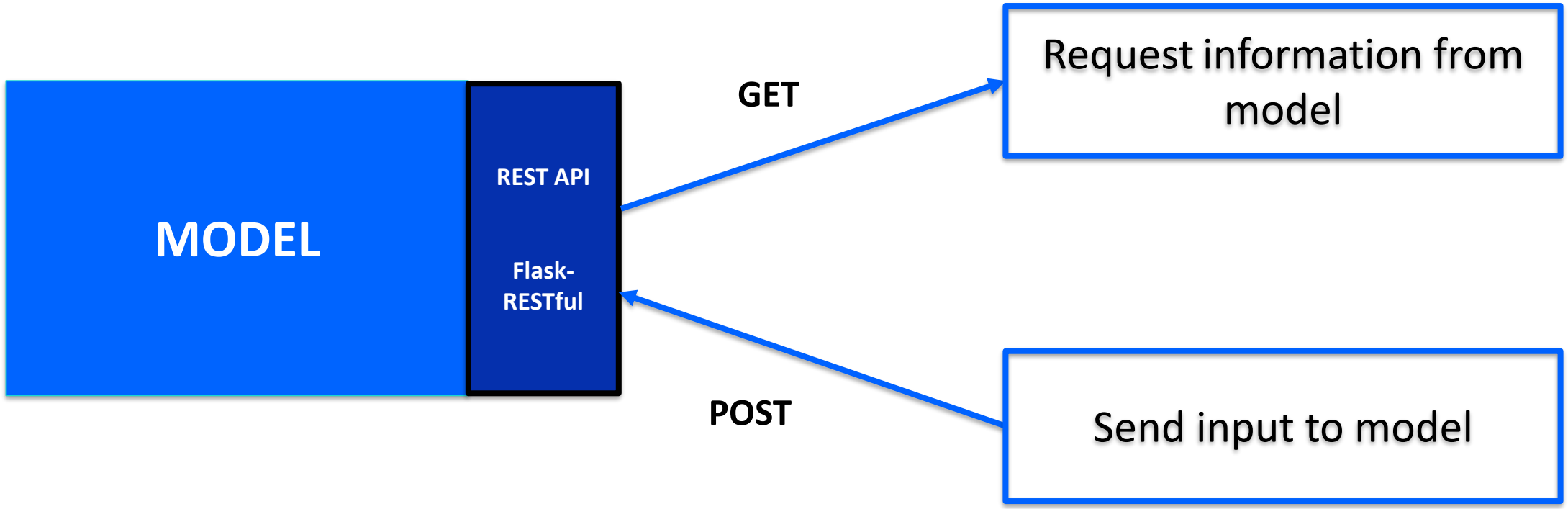


- Curated free and open datasets under open data licenses
- Standardized dataset formats and metadata
- Ready for use in enterprise AI applications
- Complement to the **M**odel **A**sset e**X**change (**MAX**)

Deployable Asset on MAX



MAX Model Consumption – REST API



Max Demo:

- Build from DockerHub
- Build locally
- Run the containe
- Perform inference through UI and Jupyter

https://ibm.biz/max-intro-tutorial

WebApp Demo:

- Build the WebApp
- Perform inference

https://ibm.biz/max-object-detector-webapp

Is there a way to run your custom trained model as a microservice?

YES

Our tools are Open Source

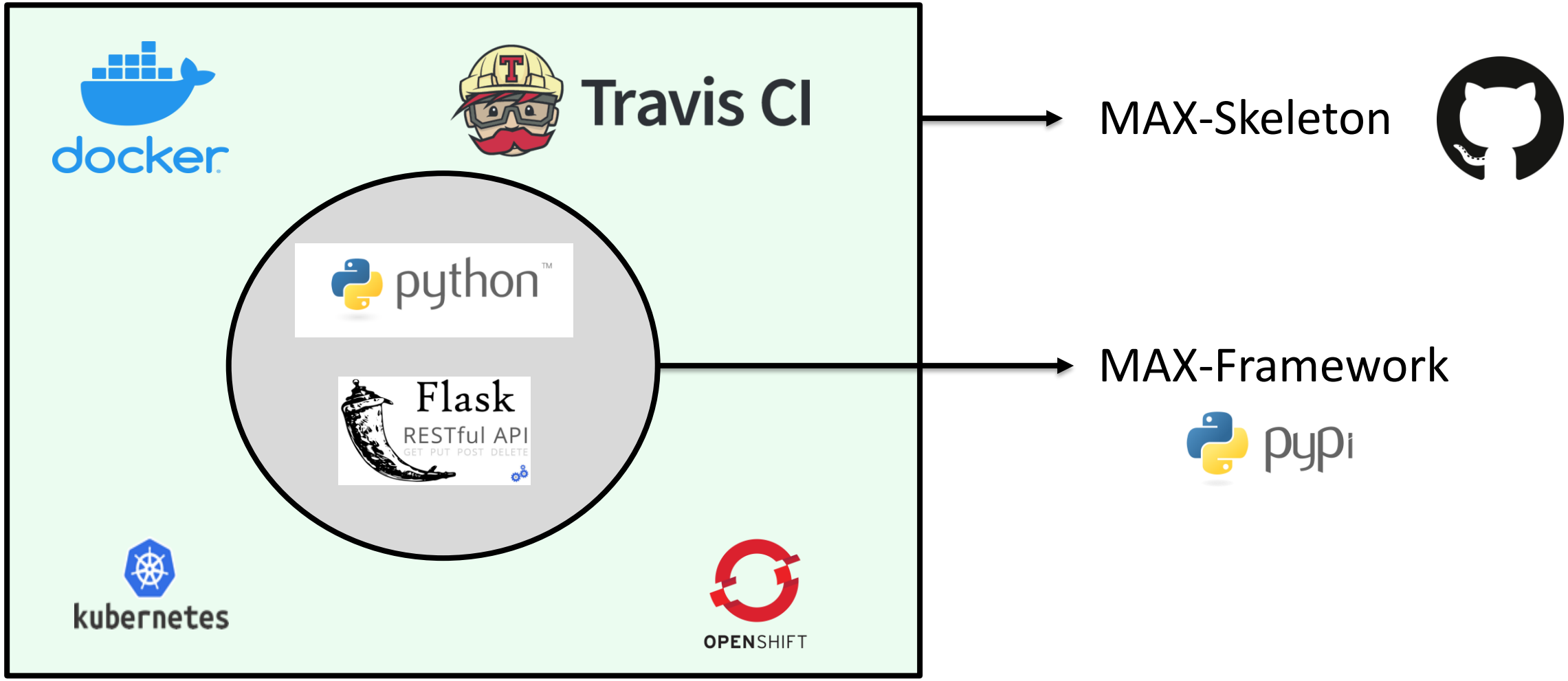
MAX-Framework

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

MAX-Skeleton

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

Components


docker
Travis CI
python
Flask
RESTful API
GET PUT POST DELETE
kubernetes
OPENSHIFT
MAX-Skeleton
MAX-Framework
pypi

MAX-Framework

https://ibm.biz/max-framework

```python
from abc import ABC, abstractmethod


class MAXModelWrapper(ABC):
    def __init__(self, path=None):
        """Implement code to load model here"""
        pass

    def _pre_process(self, x):
        """Implement code to process raw input into format required for model inference here"""
        return x

    def _post_process(self, x):
        """Implement any code to post-process model inference response here"""
        return x

    @abstractmethod
    def _predict(self, x):
        """Implement core model inference code here"""
        pass

    def predict(self, x):
        pre_x = self._pre_process(x)
        prediction = self._predict(pre_x)
        result = self._post_process(prediction)
        return result
```

MAX-Skeleton

IBM / MAX-Skeleton Template

Watch 26 | Star 2 | Fork 5

Code | Issues 2 | Pull requests 2 | Projects 0 | Wiki | Security | Insights | Settings

A cookie-cutter / skeleton for MAX repos

Edit

Manage topics

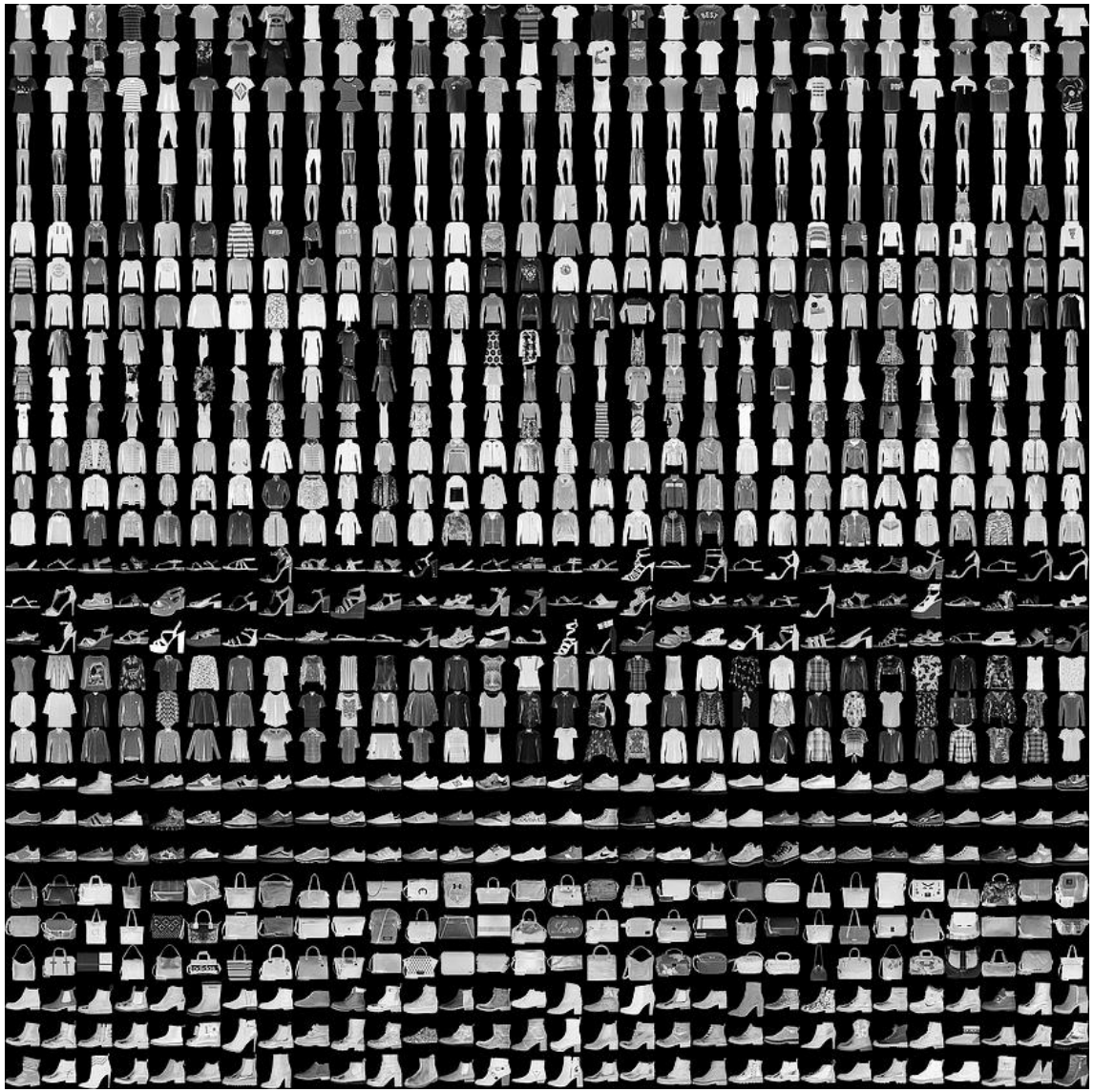
67 commits | 6 branches | 0 releases | 9 contributors | Apache-2.0

Branch: master | New pull request | Create new file | Upload files | Find File | Use this template | Clone or download

SSaishruthi and djalova Update .gitignore and test file (#29) ··· Latest commit 63bf086 24 days ago

api	Set copyright year to 2018-2019	3 months ago
core	Set copyright year to 2018-2019	3 months ago
docs	[WIP] add tutorial to README	3 months ago
samples	Clean up sample readme template	2 months ago
tests	Update .gitignore and test file (#29)	24 days ago
.dockerignore	no license for .gitignore and .dockerignore	3 months ago
.gitignore	Update .gitignore and test file (#29)	24 days ago
.travis.yml	Remove license header from travis config	2 months ago
Dockerfile	Merge pull request #11 from splovyt/complete_license	2 months ago
LICENSE	Revert LICENSE update	3 months ago
README-template.md	Update README-template.md	last month
README.md	Remove references to assets/ directory	2 months ago
app.py	Set copyright year to 2018-2019	3 months ago
config.py	Set copyright year to 2018-2019	3 months ago
max-skeleton.yaml	Remove license header from max-skeleton.yaml file (#22)	2 months ago
md5sums.txt	Import from internal repo	6 months ago
requirements-test.txt	Add list of packages required for running the tests. (#5)	4 months ago

Demo



Classify Fashion & Clothing item

Mapping

0: **T-shirt/top**
1: **Trouser**
2: **Pullover**
3: **Dress**
4: **Coat**
5: **Sandal**
6: **Shirt**
7: **Sneaker**
8: **Bag**
9: **Ankle boot**

Steps

Find a suitable **dataset** for the task

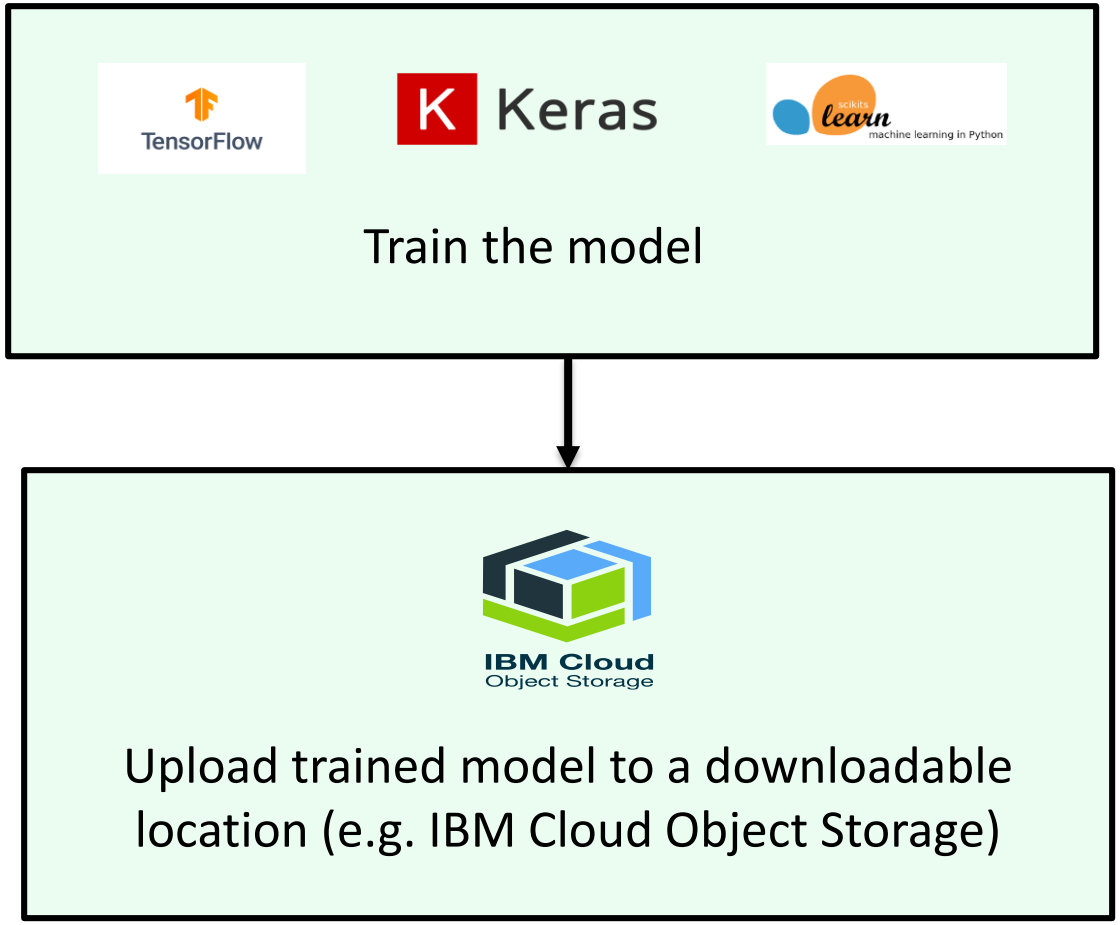
↓

Fashion-MNIST

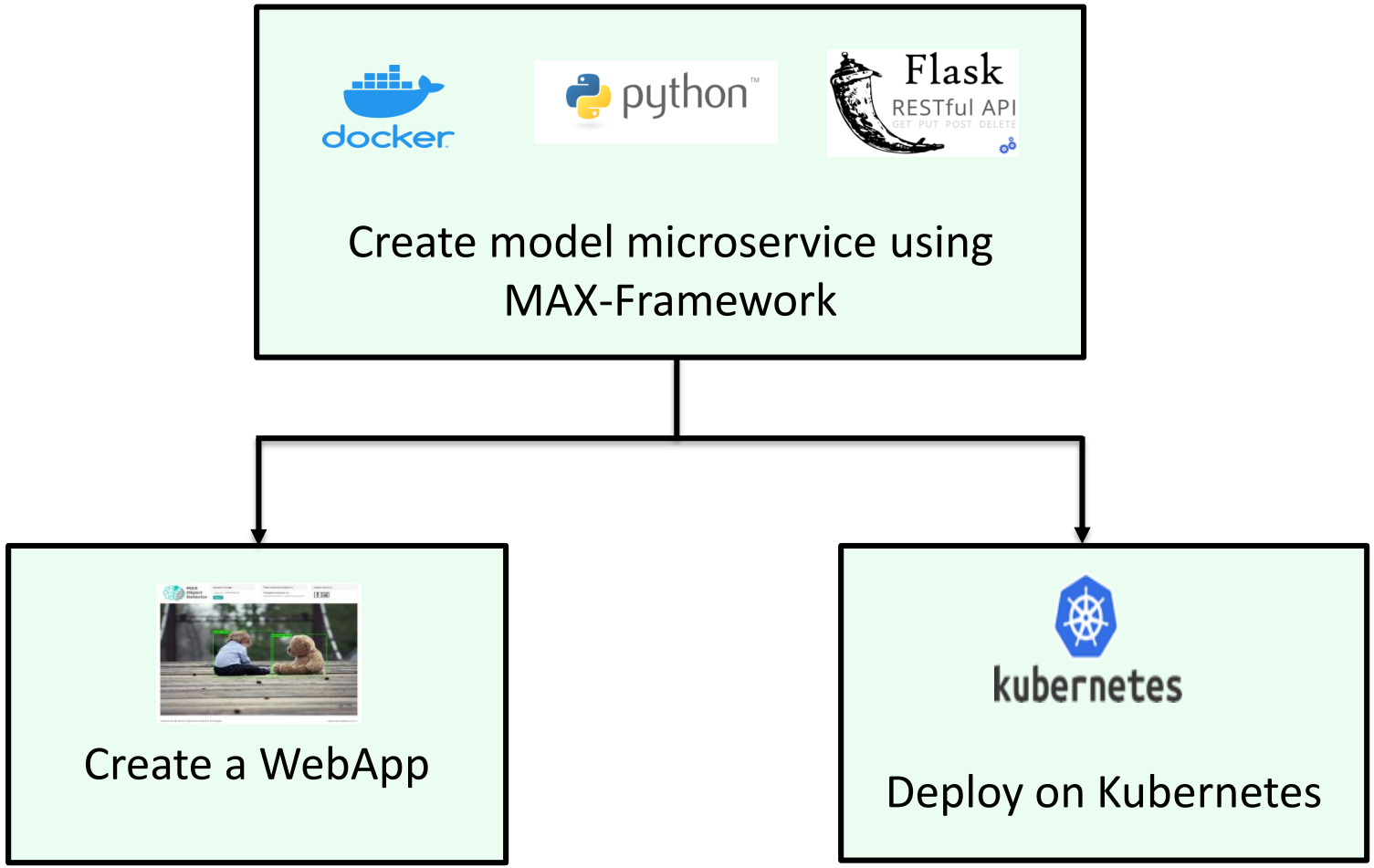
↓

Load data from Data Asset eXchange (**DAX**)

Steps


TensorFlow
Keras
scikit learn
machine learning in Python
Train the model
IBM Cloud
Object Storage
Upload trained model to a downloadable location (e.g. IBM Cloud Object Storage)

Steps


docker
python
Flask
RESTful API
Create model microservice using MAX-Framework
Create a WebApp
kubernetes
Deploy on Kubernetes

Requirements for Wrapping a Model

- Docker
- Python IDE or code editors
- Calculate sha512sum value for the model files.
- Pre-trained model weights stored in a downloadable location
- List of required python packages
- Input pre-processing code
- Prediction/Inference code
- Output post-processing code

Get prepared - 1

Install Docker, Python IDE or Code editors

Get location of pre-trained model

&

Calculate sha512sum value

Get list of required python package to run the scripts

numpy==1.14.1
Pillow==5.4.1
h5py==2.9.0
tensorflow==1.15

Get prepared - 2

Code to load the trained model

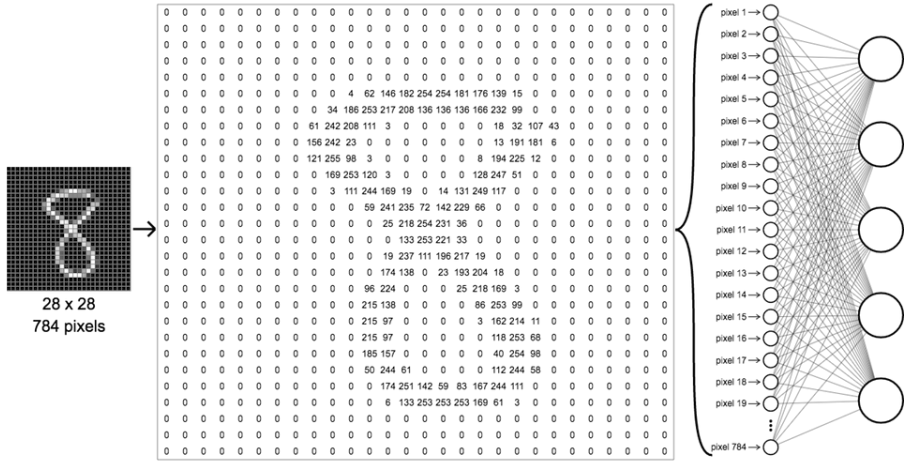
tf.keras.models.

load_model

(model_path)

Code to process the input image

Image -> array



Prediction code

model.predict

(image_array)

Get prepared - 3

Decide Output Variables

Should output contain only predicted item name or it should also have probability of the prediction?

Result = Bag

(Or)

Result = Bag

Probability = 0.95

Code to extract the desired response variables from the prediction

For 3 items in dataset, result will be

[0.98, 0.5, 0.2]

Get the item with maximum probability

Model Wrapping

https://github.com/CODAIT/presentations/tree/master/workshops/MAX-Model-Wrapping

WebApp

https://github.com/CODAIT/max-fashion-mnist-tutorial-app

How to contribute ?

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

Resources

- MAX on IBM Developer

https://ibm.biz/model-exchange

- Learning path

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

- DAX on IBM Developer

http://ibm.biz/dax-tutorial-get-started

@ibmcodait

@codait

ibm.biz/max-slack

Thank You!

IBM