

# IBM Developer Model Asset eXchange

Nick Pentreath  
Principal Engineer

*@Mlnick*

#SAISDL6

DBG / Oct 4, 2018 / © 2018 IBM Corporation

IBM  
**CODE**

# About

@MLnick on Twitter & Github

Principal Engineer, IBM

CODAIT - Center for Open-Source Data & AI  
Technologies

Machine Learning & AI

Apache Spark committer & PMC

Author of *Machine Learning with Spark*

Various conferences & meetups



IBM

**CODE**

DBG / Oct 4, 2018 / © 2018 IBM Corporation

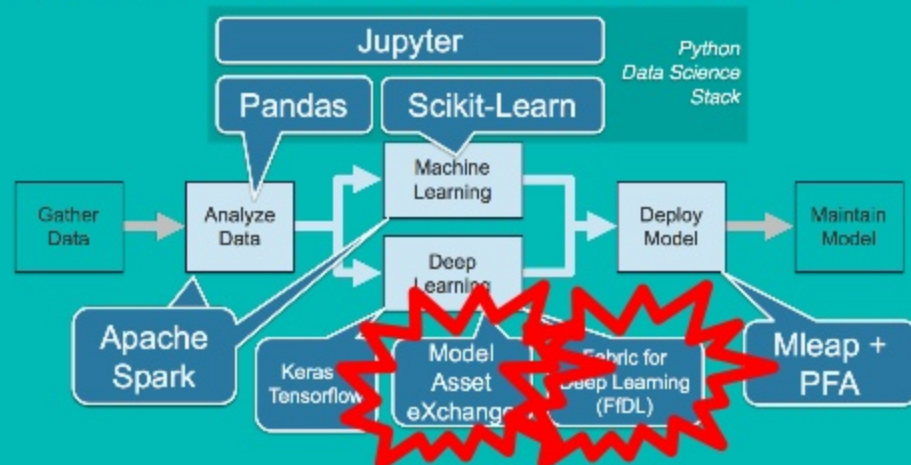
CODAIT aims to make AI solutions dramatically easier to create, deploy, and manage in the enterprise

Relaunch of the Spark Technology Center (STC) to reflect expanded mission



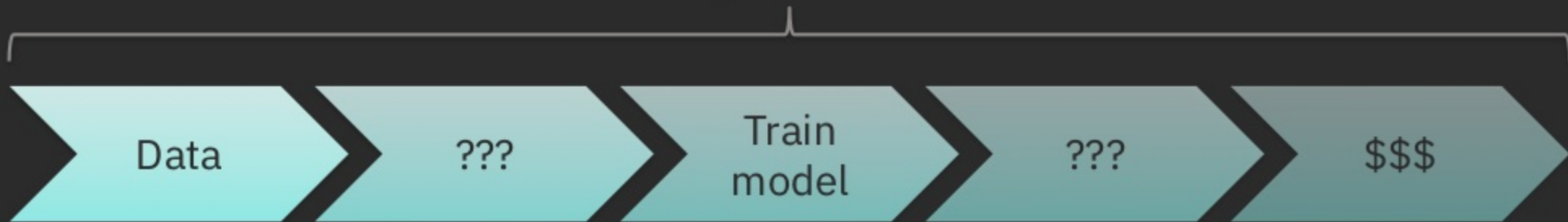
[codait.org](http://codait.org)

## Improving Enterprise AI Lifecycle in Open Source

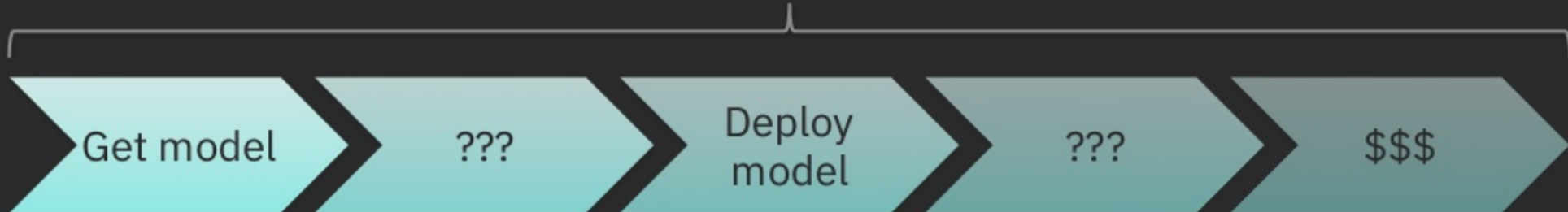


# Applying Deep Learning: Perception

Training – Data Scientist



Consumption – App Developer, Domain Expert



# Applying Deep Learning: Reality



# Step 1: Find a model

... that does what you need

... that is free to use

... that is performant enough

IBM

CODE

DBG / Oct 4, 2018 / © 2018 IBM Corporation





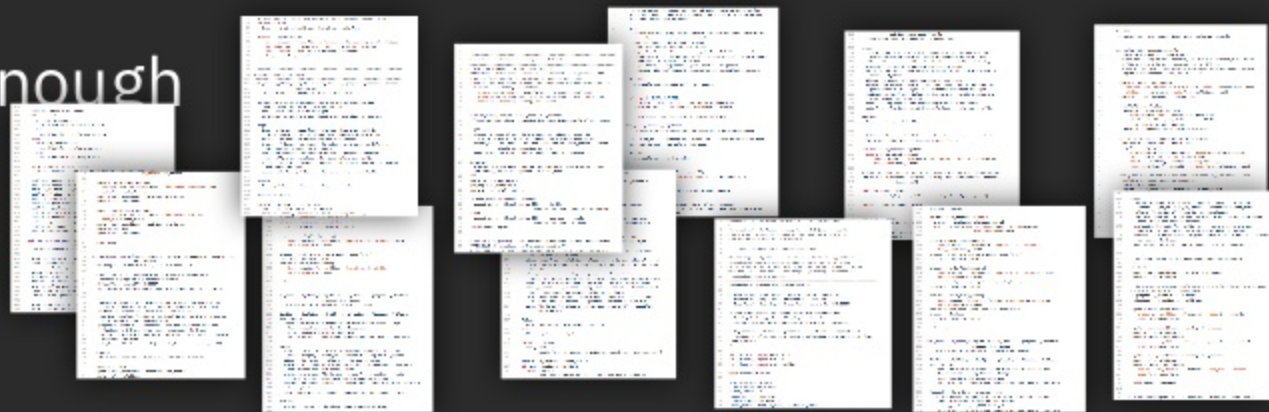
## Step 2: Get the code

Is there a good **implementation** available?

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

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

... that is **performant** enough



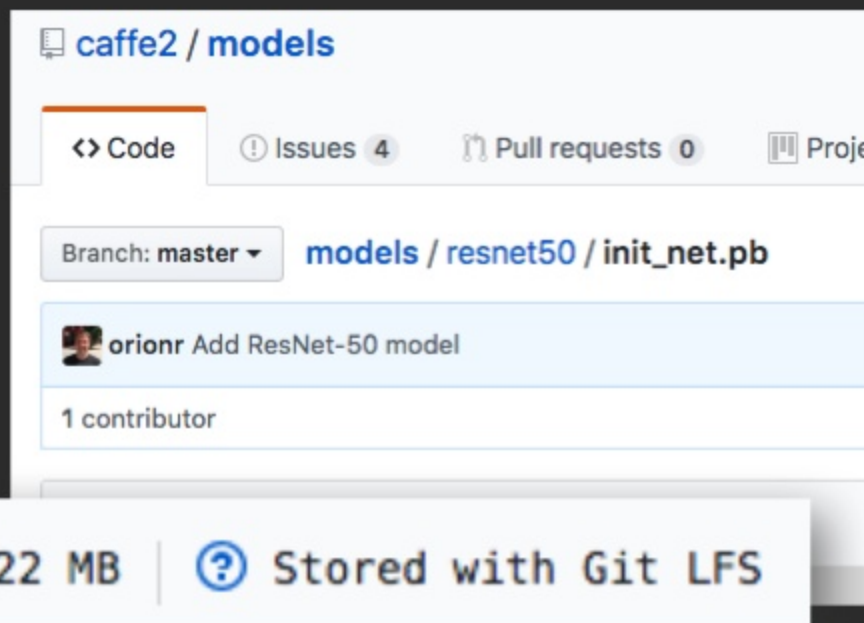
Or... Step 2: **Get** the pre-trained weights

Is there a good **pre-trained** model available?

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

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

... that is **performant** enough





### Step 3: Verify the model you found

Check ...

... that it does what you need

... that it is free to use

... that it is performant enough



Step 4(a): Train the model



Step 4(a): Train the model





Step 4(b): Figure out how to **deploy** the model



- ... adjust **inference** code (or write from scratch)
- ... **package** your inference code, model code, and pre-trained weights together
- ... **deploy** your package

## Step 5: Consume the model

... plug in to your application

... which does not know  
(or care) about tensors



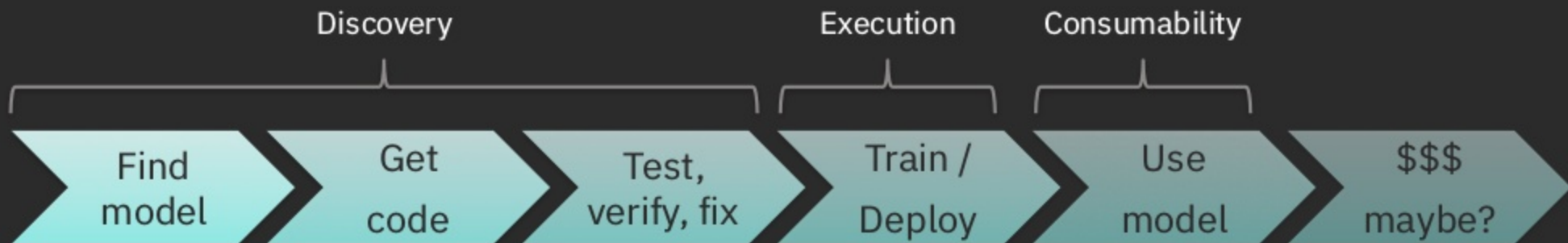
## Step 6: Profit

... hopefully





# Applying Deep Learning: Reality



Model Zoos  
(in theory)

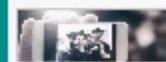




Model Zoos  
(in practice)

<http://ibm.biz/model-exchange>

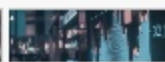
## New Asset Download



## Face it! Age Estimator

An open source machine learning model for face age estimation.

Get this model



## MobileNet Face Detector

A deep learning model for face detection using MobileNet.

Get this model



## Face it! Face to Face

A deep learning model for face-to-face video chat.

Get this model

## All models



## Detecting Faces in Images

A deep learning model for detecting faces in images.

Get this model



## Detecting Faces in Images

A deep learning model for detecting faces in images.

Get this model



## Detecting Faces in Images

A deep learning model for detecting faces in images.

Get this model



## Detecting Faces in Images

A deep learning model for detecting faces in images.

Get this model



## Detecting Faces in Images

A deep learning model for detecting faces in images.

Get this model



## Detecting Faces in Images

A deep learning model for detecting faces in images.

Get this model

## Artificial Intelligence

## CODE

Models  
Open Frameworks

## CONTENT

Announcements  
Articles  
Books  
Tutorials  
Videos

## COMMUNITY

Events  
Events

## RELATED

Deep learning  
Data science  
Deep learning  
Machine learning  
Artificial intelligence  
Speech recognition  
Visual recognition  
AI

## MODEL

## Facial Age Estimator

Recognize faces in an image and estimate the age of each face.

Get this model

Try the API

By IBM Developer Staff | Last updated: September 25, 2018

[View this model on GitHub](#)
[View this model on IBM](#)

## Overview

This repository contains code to instantiate and deploy a facial age estimation model.

## SOCIAL

[Twitter](#)
[Facebook](#)
[LinkedIn](#)
[YouTube](#)

## CONTRIBUTORS

Overviews  
Model Metadata  
References  
Licenses  
Open source license for this project



# Fabric for Deep Learning

<https://github.com/IBM/FfDL>

FfDL provides a scalable, resilient, and fault tolerant deep-learning framework

- Fabric for Deep Learning or FfDL (pronounced as 'fiddle') is an open source project which aims at making Deep Learning easily accessible to the people it matters the most i.e. Data Scientists, and AI developers.
- FfDL provides a consistent way to deploy, train and visualize Deep Learning jobs across multiple frameworks like TensorFlow, Caffe, PyTorch, Keras etc.
- FfDL is being developed in close collaboration with IBM Research and IBM Watson. It forms the core of Watson's Deep Learning service in open source.



FfDL Github Page  
<https://github.com/IBM/FfDL>

FfDL dwOpen Page  
<https://developer.ibm.com/code/open/projects/fabric-for-deep-learning-ffdl/>

FfDL Announcement Blog  
<http://developer.ibm.com/code/2018/03/20/fabric-for-deep-learning>

FfDL Technical Architecture Blog  
<http://developer.ibm.com/code/2018/03/20/democratize-ai-with-fabric-for-deep-learning>

Deep Learning as a Service within Watson Studio  
<https://www.ibm.com/cloud/deep-learning>

Research paper: "Scalable Multi-Framework Management of Deep Learning Training Jobs" [http://learningsys.org/nips17/assets/papers/paper\\_29.pdf](http://learningsys.org/nips17/assets/papers/paper_29.pdf)



## Fabric for Deep Learning (FfDL)

Deep Learning Training, Monitoring and Management



PYTORCH

Caffe

K Keras



Kubernetes – GPU/CPU/NFS Support

Cloud Hardware (GPUs and CPUs)

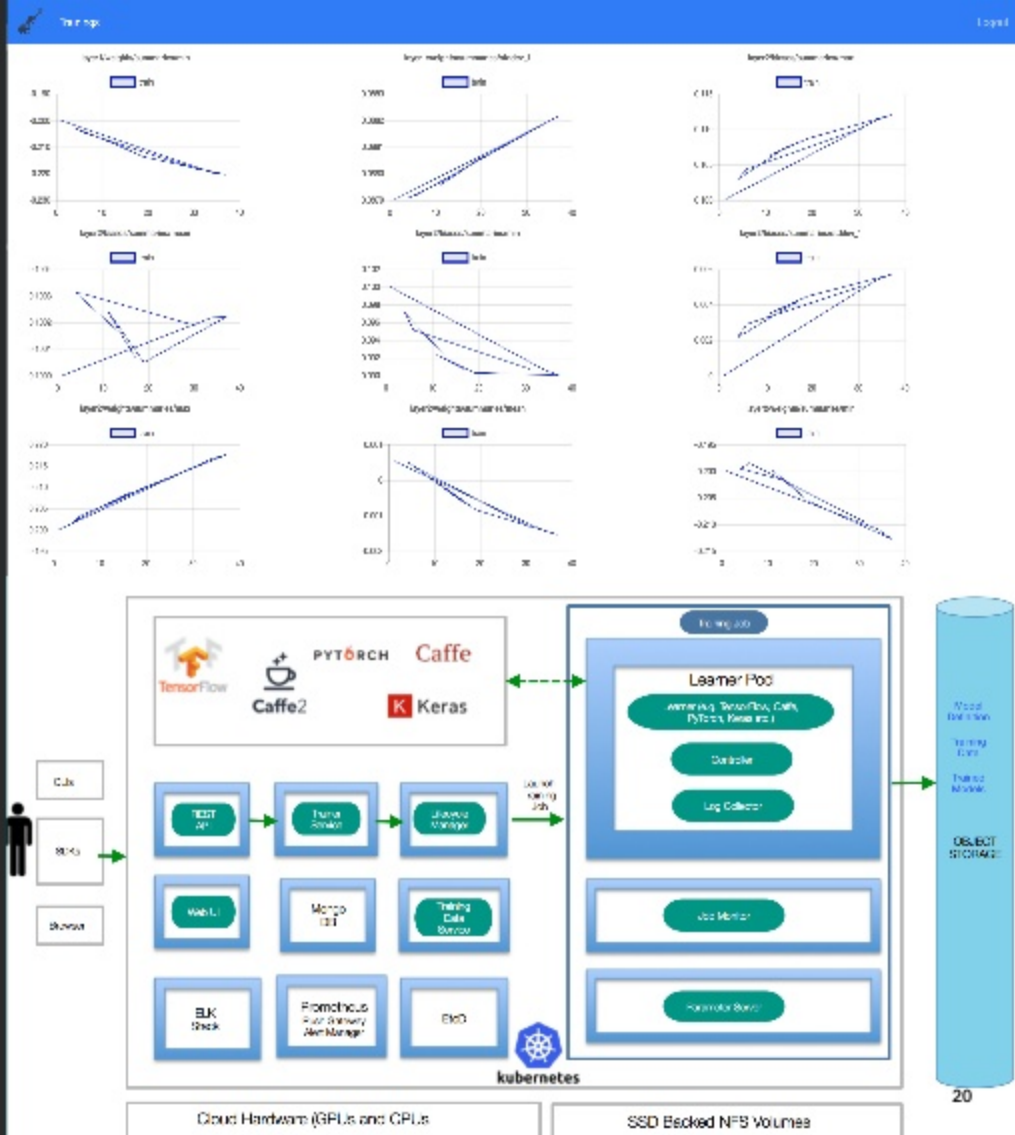
SSD Backed NFS Volumes

# Fabric for Deep Learning

<https://github.com/IBM/FfDL>

## FfDL is built using a microservices architecture on Kubernetes

- FfDL platform uses a microservices architecture to offer resilience, scalability, multi-tenancy, and security without modifying the deep learning frameworks, and with no or minimal changes to model code.
- FfDL control plane microservices are deployed as pods on Kubernetes to manage this cluster of GPU- and CPU-enabled machines effectively
- Tested Platforms: Minikube, IBM Cloud Public, IBM Cloud Private, GPUs using both Kubernetes feature gate Accelerators and NVidia device plugins





# Fabric for Deep Learning

<https://github.com/IBM/FfDL>

Just announced: Support for PyTorch 1.0  
– including distributed training and  
ONNX!

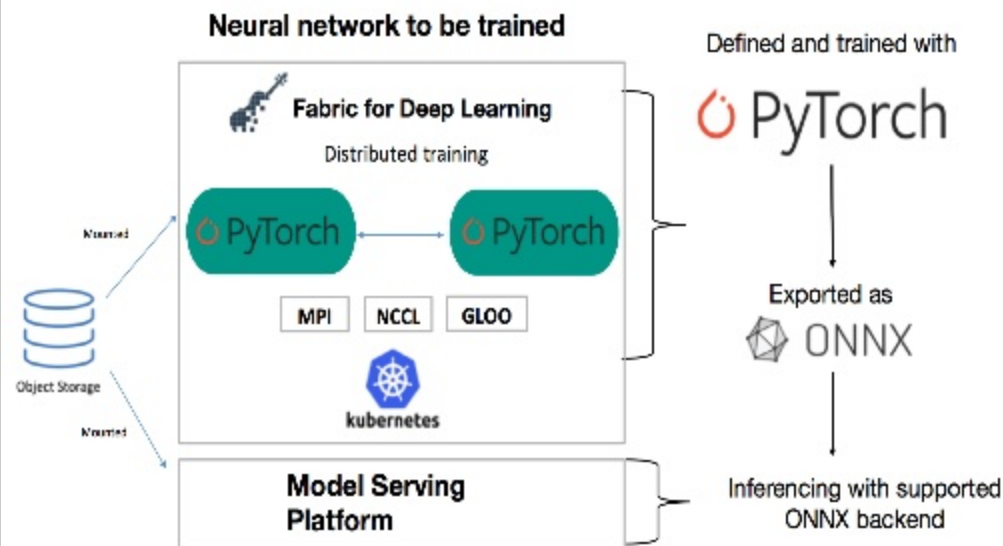
Supports distributed training via Horovod



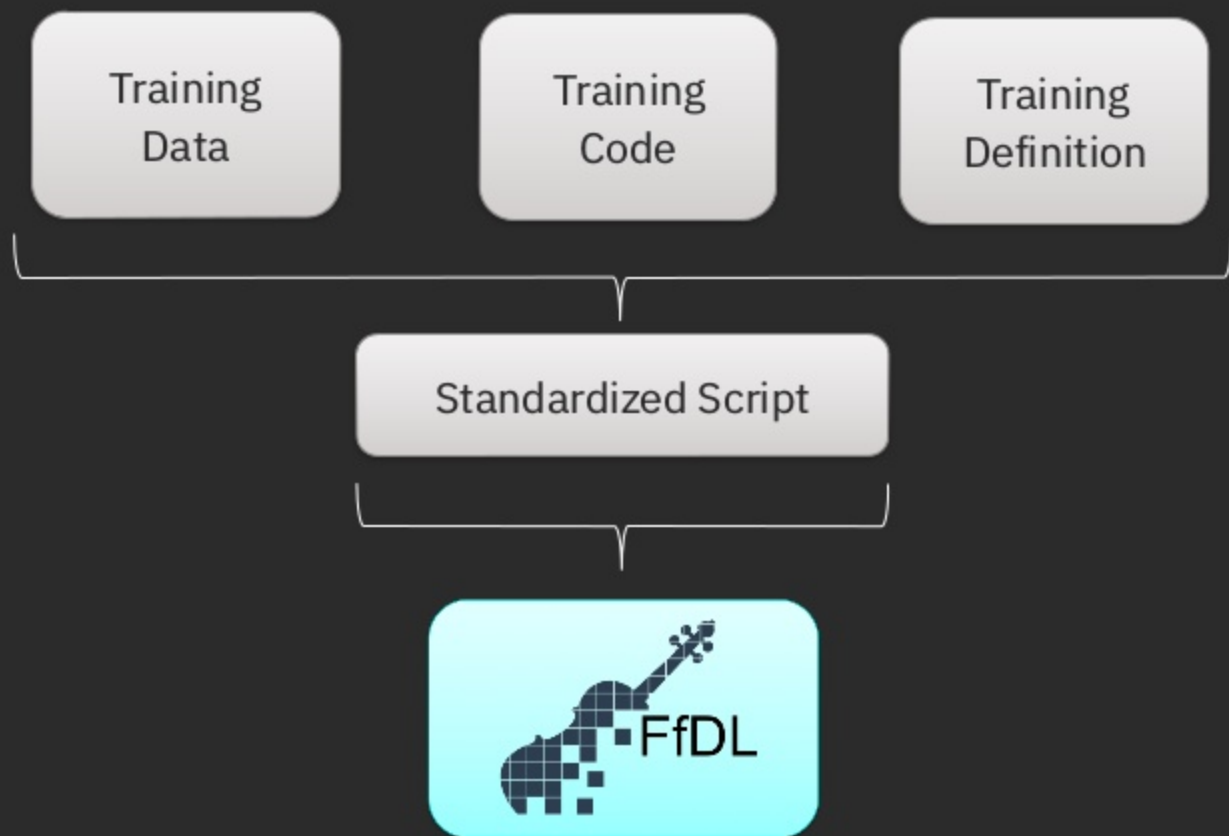
FfDL Github Page  
<https://github.com/IBM/FfDL>

FfDL / PyTorch 1.0 Blog Post  
<https://developer.ibm.com/blogs/2018/10/01/announcing-pytorch-1-support-in-fabric-for-deep-learning/>

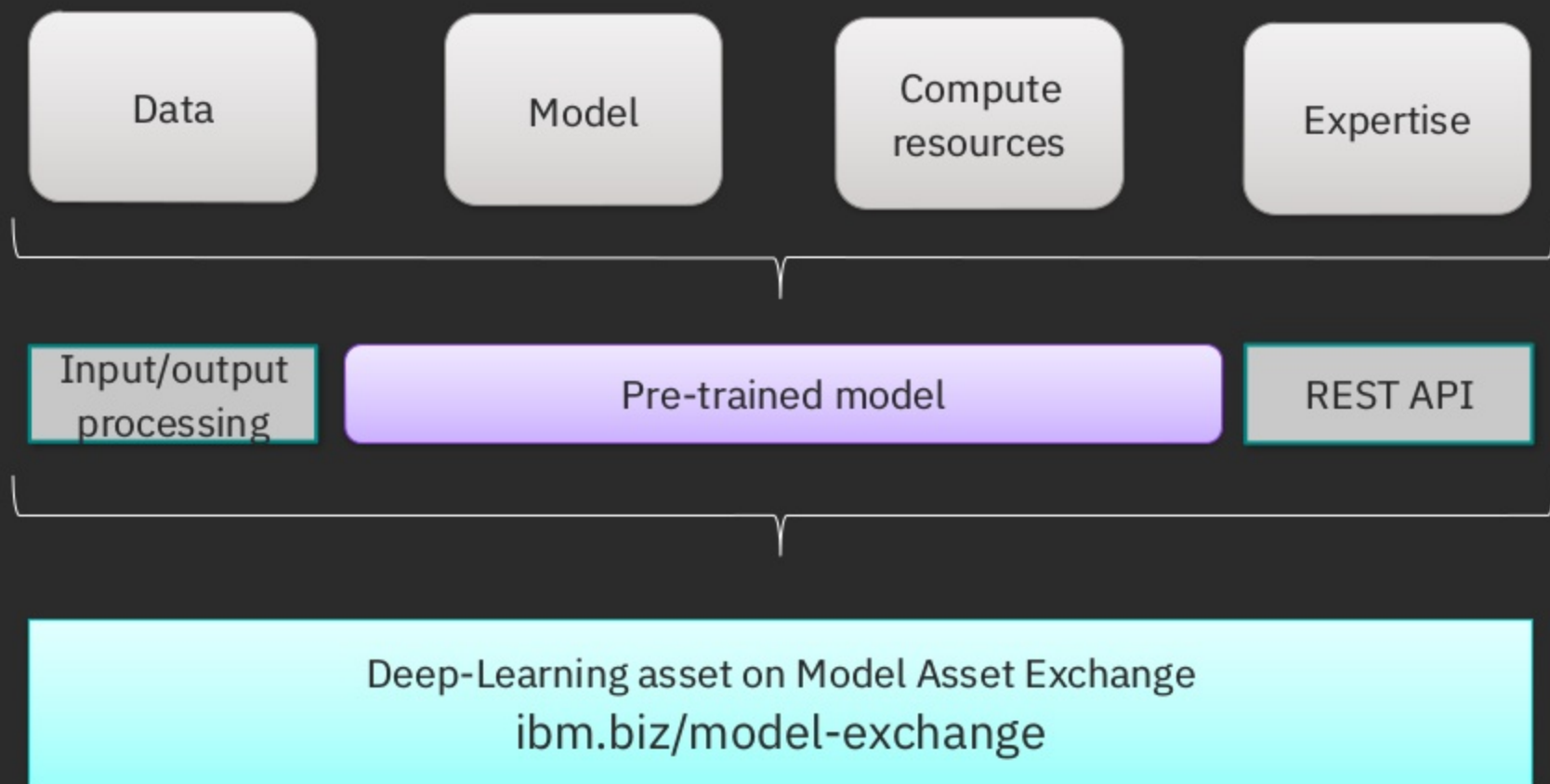
FfDL / Horovod Blog Post  
<https://developer.ibm.com/code/2018/07/18/scalable-distributed-training-using-horovod-in-ffdl/>



# Trainable Models



# Deployable Models



# Deployable Models

Deep-Learning asset on Model Asset Exchange

Deploy

Microservice

Swagger specification

Inference endpoint

Metadata endpoint

# Deployable Models

## Highlights

- Image, audio, text, healthcare, time-series and more
- Pre- / post-processing & inference wrapped up in Docker container
- Generic API framework code - Flask RESTPlus
- Swagger specification for API
- One-line deployment locally and on a Kubernetes cluster
- **Code Patterns** demonstrating how to easily consume MAX models

This model can be deployed using the following mechanisms:

- Deploy from Dockerhub:

```
docker run -it -p 5000:5000 codett/max-facial-age-estimator
```

- Deploy on Kubernetes:

```
kubectl apply -f https://raw.githubusercontent.com/IDR/NAX-facial-Age-Estimator/master/max-fa
```

- Locally: follow the instructions in the [model README](#) on GitHub

POST /max/facial/analyzeVideoInputData

Parameters Cancel

name UNIQUE ID

image 4 MB MAX

file (REQUIRED)

Response Response content type: application/json

URL

URL: curl -X POST "http://localhost:5000/max/facial/analyzeVideoInputData" -H "accept: application/json" -H "content-type: multipart/form-data" -F "image=@./sampleImage.jpg"

Request URL

http://localhost:5000/max/facial/analyzeVideoInputData

Server response

Code 200

Response body

```
{
  "status": "ok",
  "message": " ",
  "prediction": 40,
  "max": 100,
  "min": 100,
  "age": 40
}
```

# Summary and Possible Future Directions

## Current status

- 22 models (4 trainable)
- Image, audio, text, healthcare, time-series and more
- 3 [Code Patterns](#) demonstrating how to consume MAX models in a web app
- [Code Pattern](#) on training an audio classifier using [Watson Machine Learning](#)
- One-line deployment via Docker and on a Kubernetes cluster

## Potential Future

- More deployable models – breadth and depth
- More trainable models - transfer learning in particular
- New MAX web portal launching soon
- More MAX-related content:
  - [Code Patterns](#)
  - Conference talks, meetups
  - Workshops
- Enhance production-readiness of MAX models
- Improve MAX API framework



# IBM Developer Model Asset eXchange

- Free, open-source deep learning models.
- Wide variety of domains.
- Multiple deep learning frameworks.
- Vetted and tested code and IP.



<http://ibm.biz/model-exchange>

## IBM Code Model Asset Exchange

Accelerate business value with vetted open source deep learning models

Try it on IBM Cloud



Machine



IBM Developer

Topics

Community

More open source at IBM

Search



Artificial Intelligence

CODE

Models  
Open Frameworks

CONTENT

Announcements  
Articles  
Books  
Tutorials  
Videos

COMMUNITY

Events  
Blogs

RELATED

Deep learning  
Data science  
Deep learning  
Machine learning  
Artificial intelligence  
Speech recognition  
Vision

MODEL

## Facial Age Estimator

Recognize faces in an image and estimate the age of each face.

Get this model

Try the API

By IBM Developer Staff | Last updated September 25, 2018

On GitHub | On IBM Cloud | On AWS

## Overview

This repository contains code to instantiate and deploy a facial age estimation model.

SOCIAL



CONTRIBUT

Overview  
Model Metadata  
References  
Licenses  
Open source license for this project

Thank you!



[codait.org](https://codait.org)



[twitter.com/MLnick](https://twitter.com/MLnick)



[github.com/MLnick](https://github.com/MLnick)



[developer.ibm.com](https://developer.ibm.com)



FfDL

MAX



Sign up for IBM Cloud and try Watson Studio!

<https://ibm.biz/BdYbTY>

<https://datascience.ibm.com/>

