**Reading: Additional Sources of Datasets**

**Estimated time: 5 mins**

In this reading, you will learn about:

* Open datasets and sources
* Proprietary datasets and sources
* Dataset license

**Open datasets and sources**

In this data-driven world, some datasets are freely available for anyone to access, use, modify, and share. These are called **open datasets**.  
Open datasets include a public license and are very useful for your journey as a Data Scientist. Some of the most informative open dataset sources are listed below.

**Government Data:**

* <https://www.data.gov/>
* <https://www.census.gov/data.html>
* <https://data.gov.uk/>
* <https://www.opendatanetwork.com/>
* <https://data.un.org/>

**Financial Data Sources:**

* <https://data.worldbank.org/>
* <https://www.globalfinancialdata.com/>
* <https://comtrade.un.org/>
* <https://www.nber.org/>
* <https://fred.stlouisfed.org/>

**Crime Data:**

* <https://www.fbi.gov/services/cjis/ucr>
* <https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html>
* <https://www.drugabuse.gov/related-topics/trends-statistics>
* <https://www.unodc.org/unodc/en/data-and-analysis/>

**Health Data:**

* <https://www.who.int/gho/database/en/>
* <https://www.fda.gov/Food/default.htm>
* <https://seer.cancer.gov/faststats/selections.php?series=cancer>
* <https://www.opensciencedatacloud.org/>
* <https://pds.nasa.gov/>
* <https://earthdata.nasa.gov/>
* <https://www.sgim.org/communities/research/dataset-compendium/public-datasets-topic-grid>

**Academic and Business Data:**

* <https://scholar.google.com/>
* <https://nces.ed.gov/>
* <https://www.glassdoor.com/research/>
* <https://www.yelp.com/dataset>

**Other General Data:**

* <https://www.kaggle.com/datasets>
* <https://www.reddit.com/r/datasets/>

**Propriety datasets and sources**

Proprietary datasets contain data primarily owned and controlled by specific individuals or organizations. This data is limited in distribution because it is sold with a licensing agreement.  
Some data from private sources cannot be easily disclosed, like public data.

National security data, geological, geophysical, and biological data are examples of propriety data. Copyright laws or patents usually bind this type of data. Proprietary datasets that mainly contain sensitive information are less widely available than open datasets.

Some standard propriety dataset sources are listed below.

**Health Care:**

<https://www.sgim.org/communities/research/dataset-compendium/proprietary-datasets>

**Financial Market data:**

<https://datarade.ai/data-categories/proprietary-market-data>

**Google Cloud based datasets:**

<https://cloud.google.com/datasets>

**Dataset licenses**

When you select a dataset, it is necessary to look into the license. A license explains whether you can use that dataset or not; or explains if you have to accept certain guidelines to use that dataset. The different license types are listed below.

1. **PUBLIC DOMAIN MARK - PUBLIC DOMAIN**  
   When a dataset has a Public Domain license, all the rights to use, access, modify and share the dataset are open to everyone. Here there is technically no license.
2. **OPEN DATA COMMONS PUBLIC DOMAIN DEDICATION AND LICENSE – PDDL**  
   Open Data Commons license has the same features as the Public Domain license, but the difference is the PDDL license uses a licensing mechanism to give the rights to the dataset.
3. **CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL CC-BY**  
   This license allows users to share and modify a dataset, but only if they give credit to the creator(s) of the dataset.
4. **COMMUNITY DATA LICENSE AGREEMENT – CDLA PERMISSIVE-2.0**  
   Like most open-source licenses, this license allows users to use, modify, adapt, and share the dataset, but only if a disclaimer of warranties and liability is also included.
5. **OPEN DATA COMMONS ATTRIBUTION LICENSE - ODC-BY**  
   This license allows users to share and adapt a dataset, but only if they give credit to the creator(s) of the dataset.
6. **CREATIVE COMMONS ATTRIBUTION-SHAREALIKE 4.0 INTERNATIONAL - CC-BY-SA**  
   This license allows users to use, share, and adapt a dataset, but only if they give credit to the dataset and show any changes or transformations, they made to the dataset. Users might not want to use this license because they have to share the work they did on the dataset.
7. **COMMUNITY DATA LICENSE AGREEMENT – CDLA-SHARING-1.0**  
   This license uses the principle of ‘copyleft’: users can use, modify, and adapt a dataset, but only if they don’t add license restrictions on the new work(s) they create with the dataset.
8. **OPEN DATA COMMONS OPEN DATABASE LICENSE - ODC-ODBL**  
   This license allows users to use, share, and adapt a dataset but only if they give credit to the dataset and show any changes or transformations they make to the dataset. Users might not want to use this license because they have to share the work they did on the dataset.
9. **CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL 4.0 INTERNATIONAL - CC BY-NC**  
   This license is a restrictive license. Users can share and adapt a dataset, provided they give credit to its creator(s) and ensure that the dataset is not used for any commercial purpose.
10. **CREATIVE COMMONS ATTRIBUTION-NO DERIVATIVES 4.0 INTERNATIONAL - CC BY-ND**  
    This license is also a restrictive license. Users can share a dataset if they give credit to its creator(s). This license does not allow additions, transformations, or changes to the dataset.
11. **CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-SHAREALIKE 4.0 INTERNATIONAL - CC BY-NC-SA**  
    This license allows users to share a dataset only if they give credit to its creator(s). Users can share additions, transformations, or changes to the dataset, but they cannot use the dataset for commercial purposes.
12. **CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 INTERNATIONAL - CC BY-NC-ND**  
    This license allows users to share a dataset only if they give credit to its creator(s). Users are not allowed to modify the dataset and are not allowed to use it for commercial purposes.

***Note: Additional license types exist. Any dataset you use will include details about its license.***

# Getting started with the Model Asset Exchange and the Data Asset Exchange

In this lab, you will explore the Model Asset Exchange (MAX) and the Data Asset Exchange (DAX), which are two open source Data Science resources on IBM Developer.

## Objective of Exercise 1:

* Find open data sets on IBM Developer.
* Explore the data sets.

## Objective of Exercise 2:

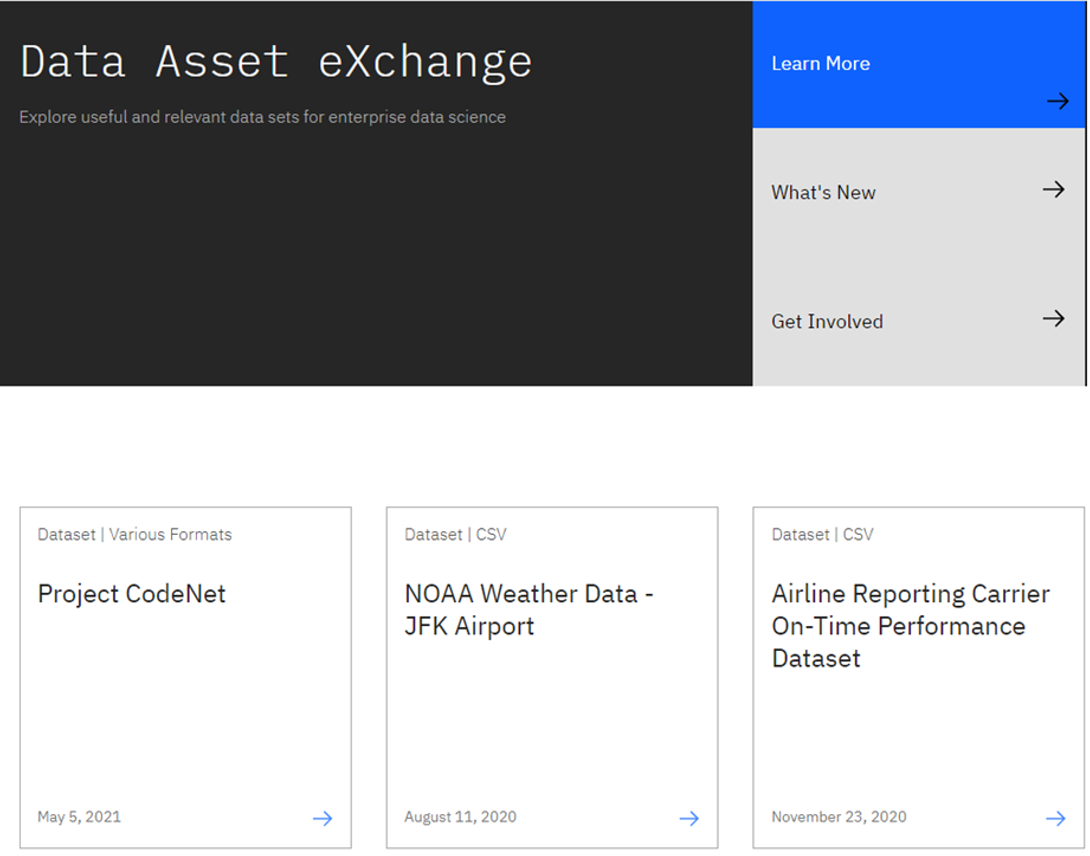
* Find ready-to-use deep learning models on the Model Asset Exchange.
* Explore the deep learning model trained to detect objects in an image.

It will take you approximately 15 minutes to complete the lab. Only a web browser is required to complete the tasks.

## Exercise 1: Explore deep learning datasets

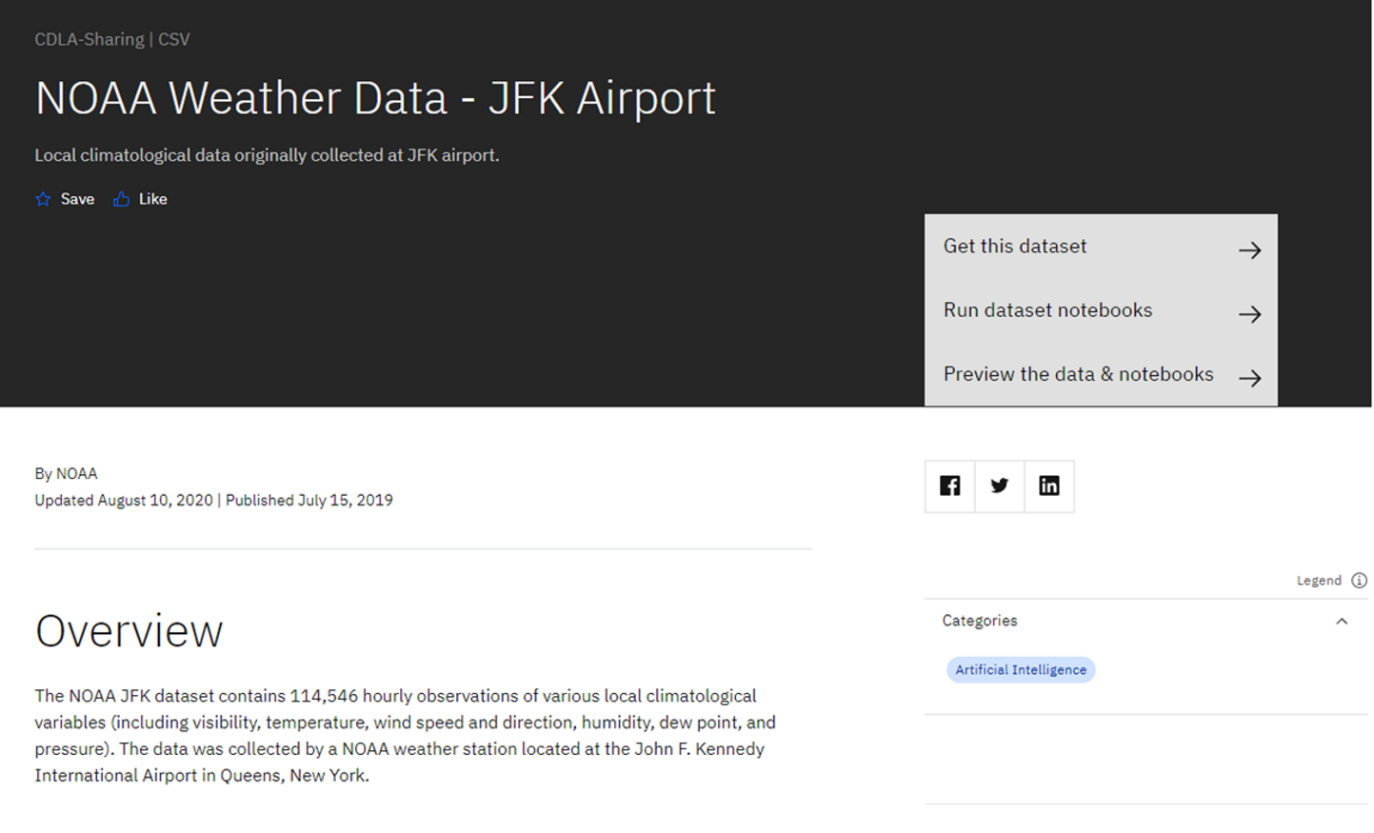
The Data Asset Exchange is a curated collection of open datasets from IBM Research and third-parties that you can use to train models.

1. Open <https://developer.ibm.com/> in your web browser.
2. From the main menu select **Open Source at IBM > Data Asset eXchange**. The DAX home page is displayed.



The collection includes datasets from the Debater project (<https://www.research.ibm.com/artificial-intelligence/project-debater/>), datasets that can be used to train models to perform document layout analysis, natural language processing, time series analysis, and more.

1. Open the NOAA Weather Data dataset (<https://developer.ibm.com/exchanges/data/all/jfk-weather-data/>), which contains data from a weather station at the John F. Kennedy Airport in New York spanning eight years. This dataset was used to train the weather forecaster model on MAX (<https://developer.ibm.com/exchanges/models/all/max-weather-forecaster/>).

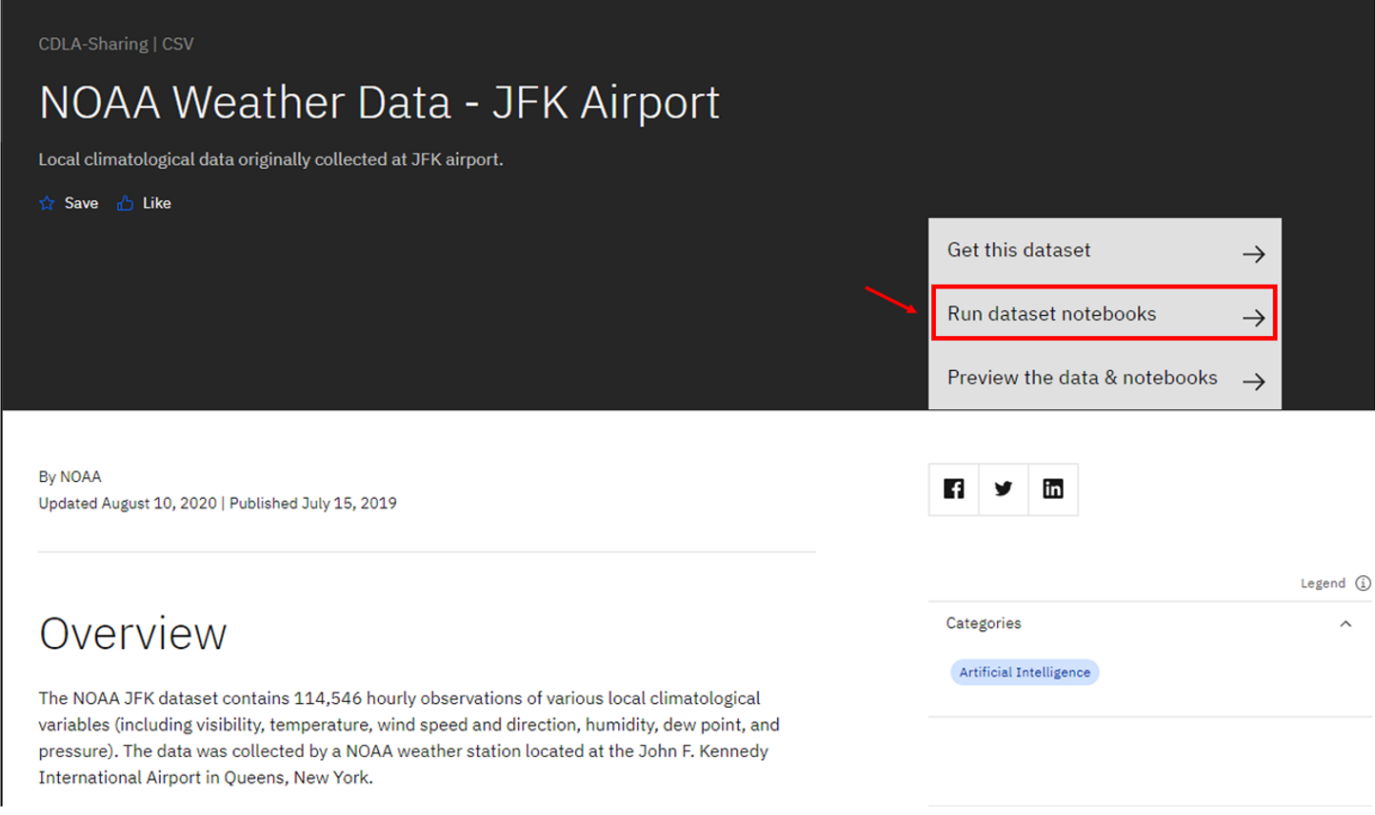


You can download the dataset using the **Get this dataset** link. Datasets are stored as compressed archives, which you can extract using any utility that supports the targz format. If you are not familiar with this file format, take a look at this short open source tutorial <https://opensource.com/article/17/7/how-unzip-targz-file>.

1. Inspect the dataset's metadata.

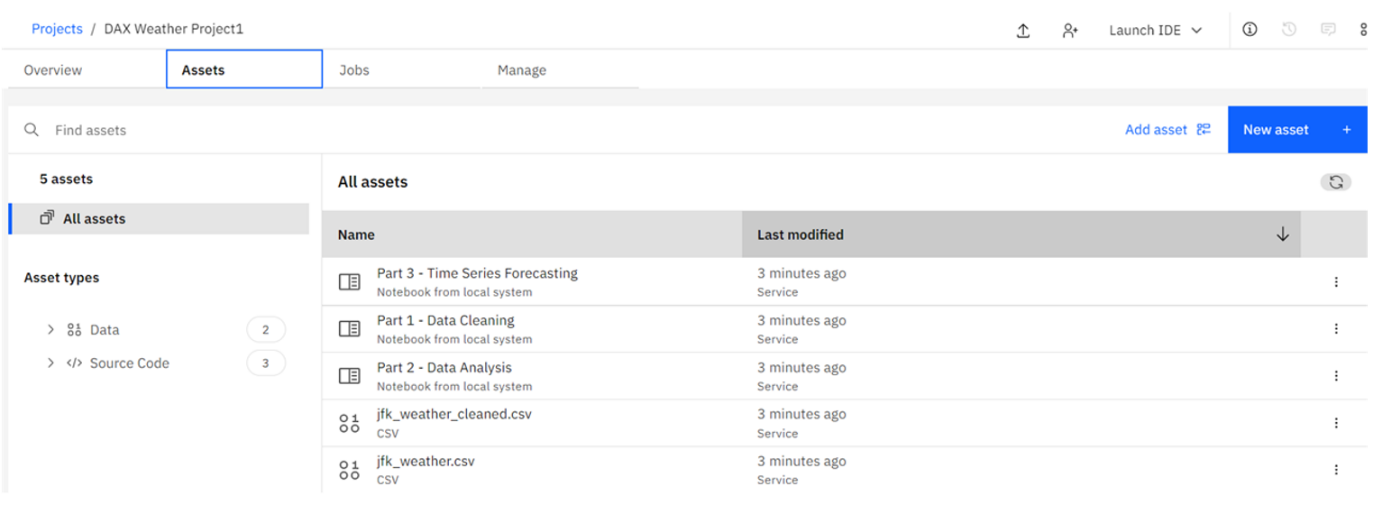
This dataset is stored as tabular data and formatted as a comma separated value (CSV) file, which is a very popular basic data exchange format.  
The dataset was published under the data science friendly CDLA-Sharing license (<https://cdla.io/>).  
The dataset contains time-series data and can be used to predict weather trends.

1. Most datasets are complemented by Python notebooks that you can use to explore, pre-process, and analyze the data. You can access the notebook (or notebooks) by clicking the **Run dataset notebooks**:



The notebooks are hosted on Watson Studio, IBM's Data Science platform. Later in this course, you'll learn more about Watson Studio notebooks and how to run them.

1. [Optional] If you are already familiar with notebooks and Watson Studio, feel free to open the link and import the project or notebook. The following example depicts the weather dataset project assets, which include the raw data file and two notebooks.

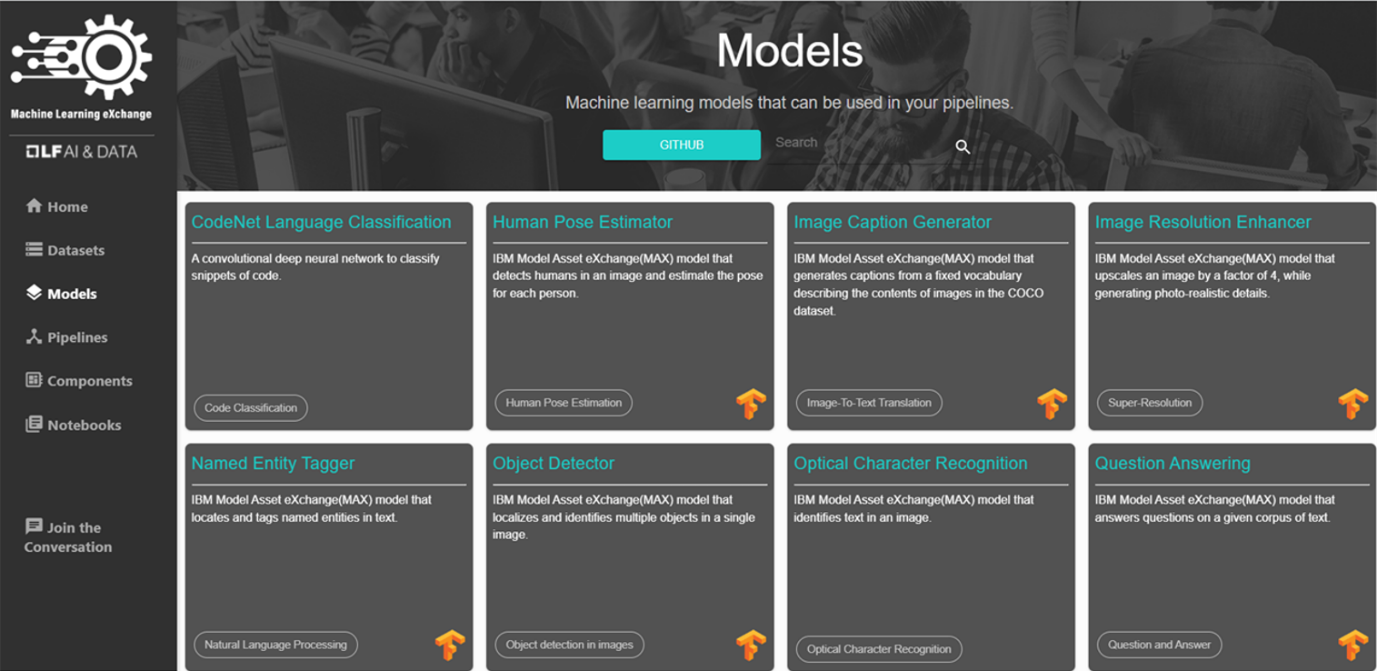


This concludes Exercise 1 of this lab, which introduced the Data Asset Exchange.

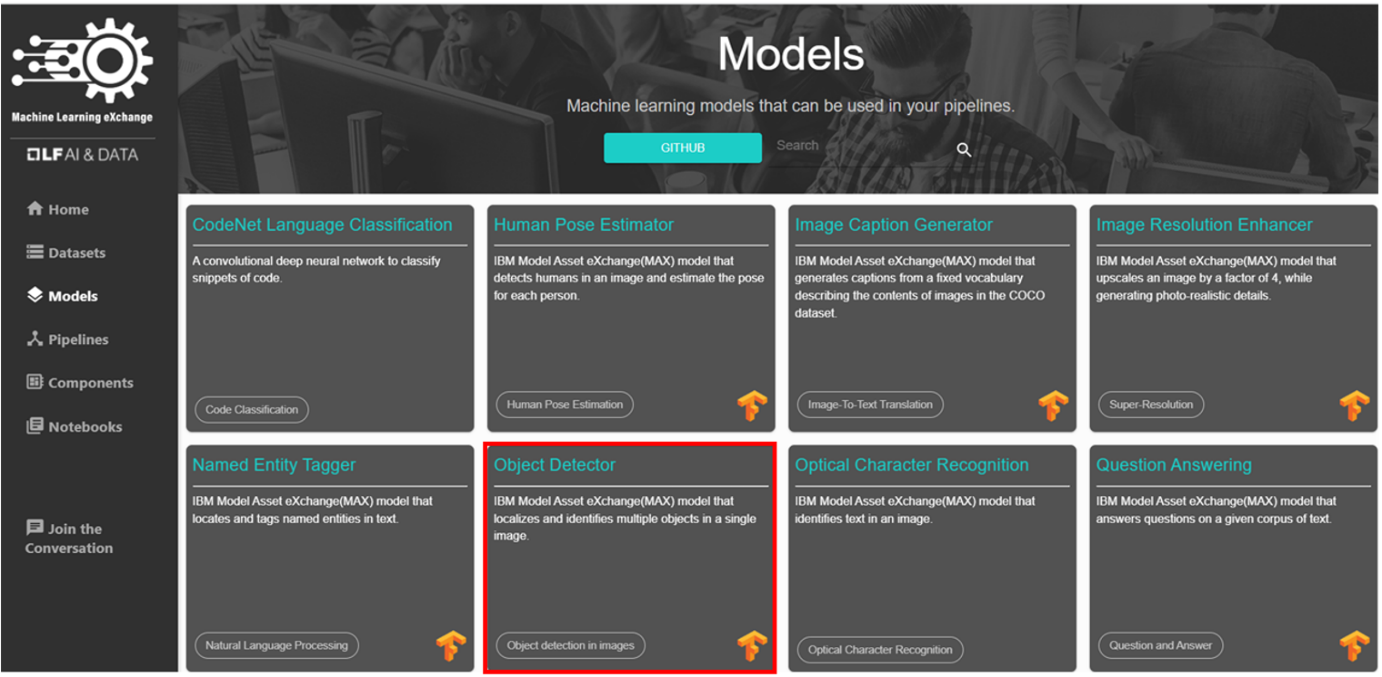
## Exercise 2 - Explore deep learning models

The Model Asset Exchange is a curated repository of open source deep learning models for a variety of domains, such as text, image, audio, and video processing.

1. Open [http://ml-exchange.org/models](http://ml-exchange.org/models?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDS0105ENSkillsNetwork984-2022-01-01) in your web browser.
2. The MAX home page is displayed. In this introductory lab exercise, we are going to focus on a few MAX key features.

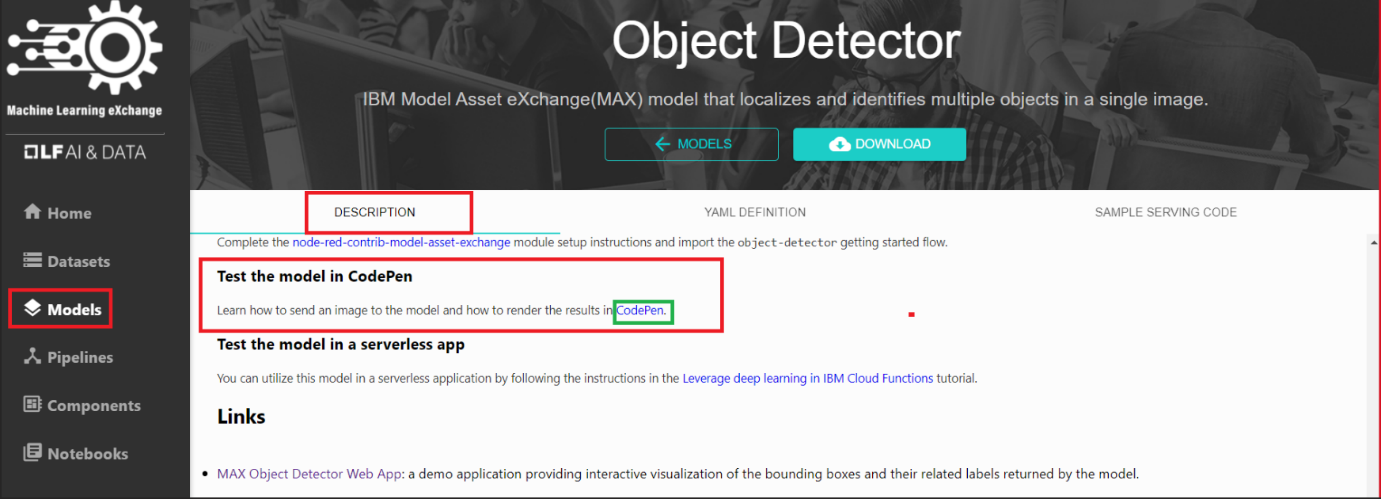


1. Select the **Object Detector** model from the list of available options.



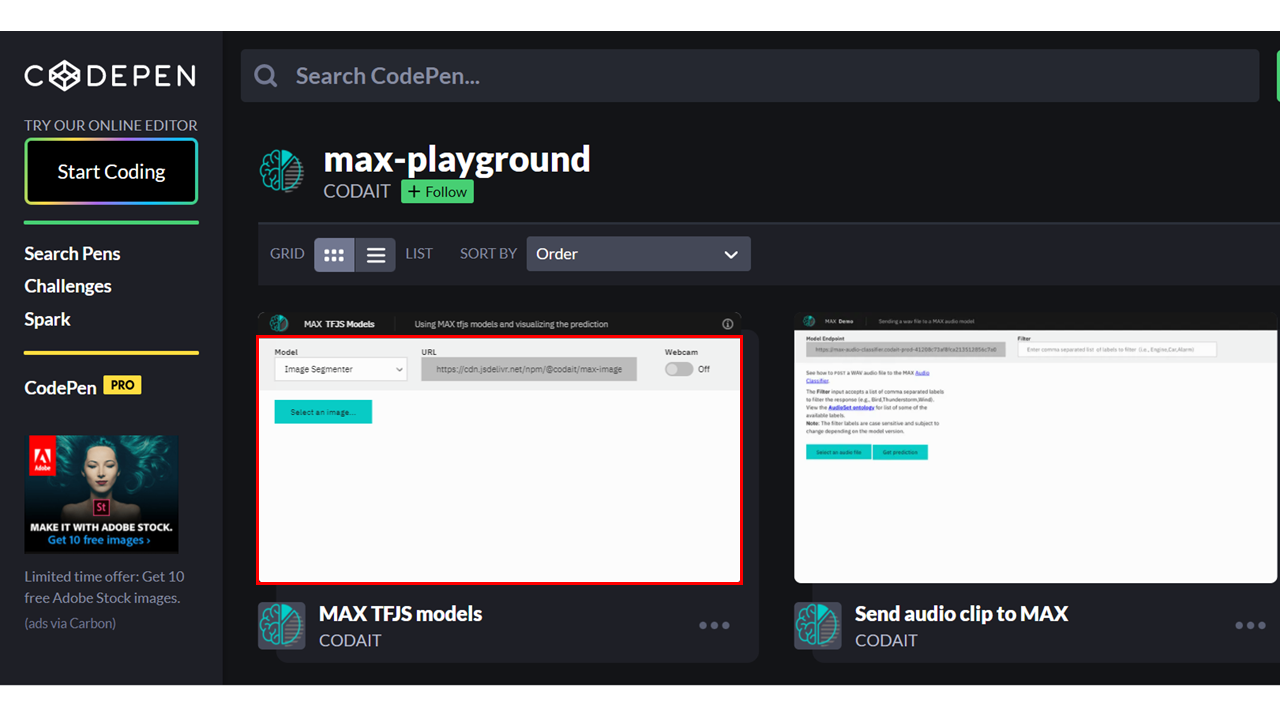
This model recognizes the objects present in an image. The model consists of a deep convolutional net base model for image feature extraction, together with additional convolutional layers specialized for the task of object detection, trained on the COCO data set. The input to the model is an image, and the output are extracted objects from the image, appropriately labeled.

1. Scroll down and in **Test the model in CodePen** click **CodePen** hyperlink as highlighted below:

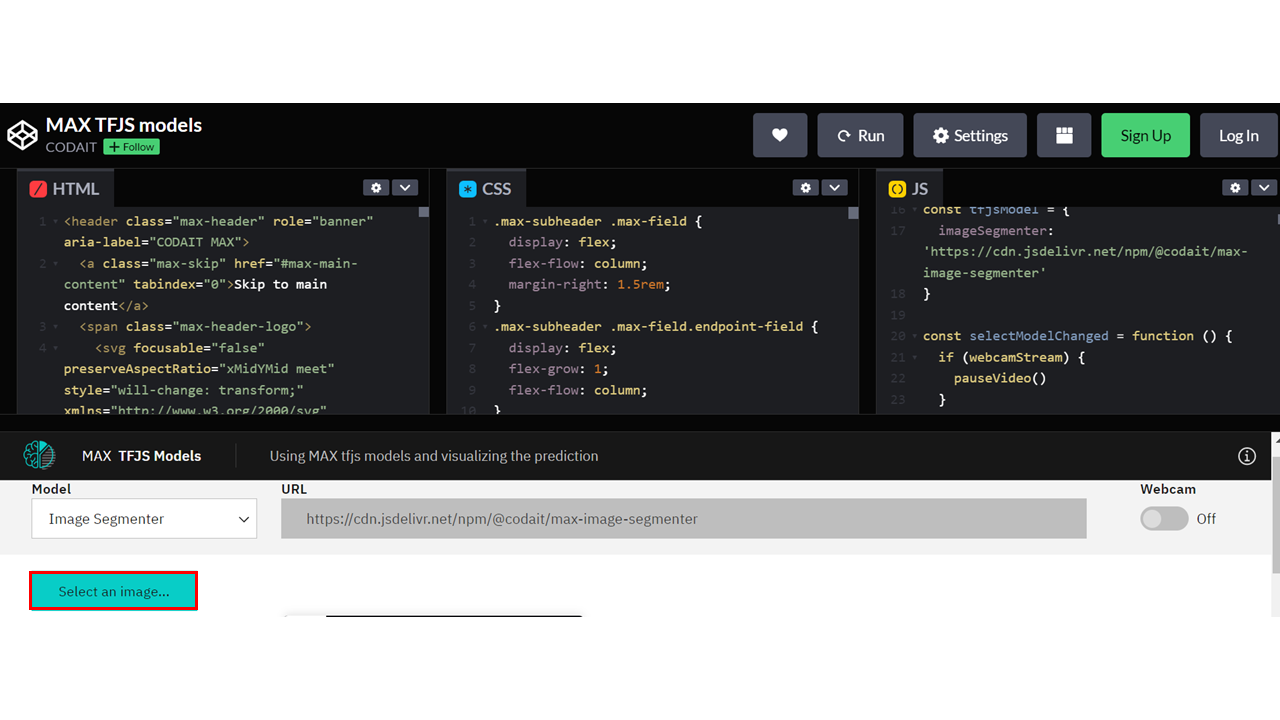


CodePen is a social development environment. At its heart, it allows you to write code in the browser and see the results of it as you build. It is a useful and liberating online code editor for developers of any skill and is particularly empowering for people learning to code.

Some of the models are already built for you to test. Let's test one of the models. Click **MAX TFJS models**.



1. Upload an image. You may choose images with a person, dog, cat, truck, car, and so on, which are labels the model has been trained on.

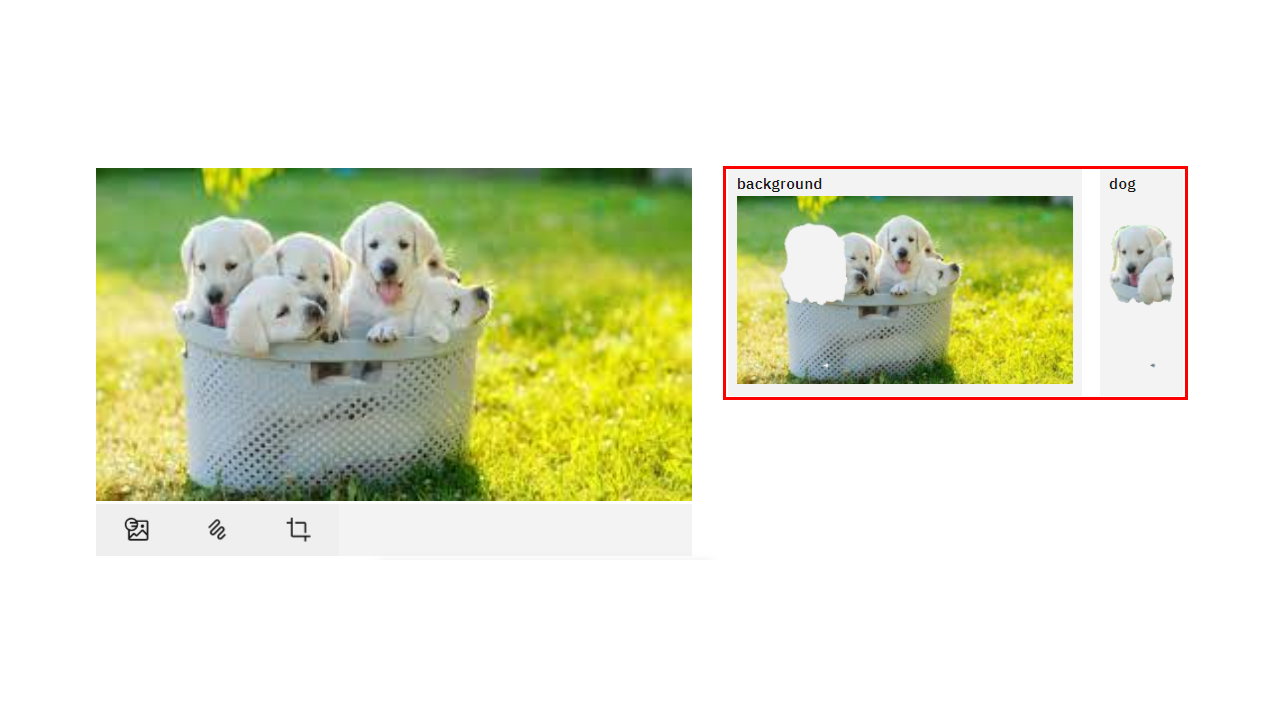


Here we are using **Image Segmenter**, which divides an image into regions or categories that correspond to different objects or parts of objects. Every pixel in an image is allocated to one of a number of these categories.

1. Click the icon **Extract prediction** as shown below:



You will now be able to see the output of the prediction on the basis of the image you upload.



Here the background and the dog image are separated, showing two different parts of the image. **You can also try the webcam option, which will show the real-time prediction by the toggle-on webcam option.**

This concludes Exercise 2 of this lab, which introduced the Model Asset Exchange.

*Optionally you can watch a demo of the Object detector model*[*here*](https://video.ibm.com/recorded/128825527?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDS0105ENSkillsNetwork984-2022-01-01)*.*

# **Module 3 Summary**

Congratulations! You have completed this module. At this point in the course, you know:

* Libraries usually contain built-in modules that provide different functionalities.
* You can use data visualization methods to communicate with others and display meaningful results of an analysis.
* For machine learning, the Scikit-learn library contains tools for statistical modeling, including regression, classification, clustering, and so on.
* Large-scale production of deep-learning models use TensorFlow, a low-level framework.
* Apache Spark is a general-purpose cluster-computing framework that allows you to process data using compute clusters.
* An application programming interface (API) allows communication between two pieces of software.
* API is the part of the library you see while the library contains all the components of the program.
* REST APIs allow you to communicate through the internet and take advantage of resources like storage, data, artificially intelligent algorithms, and much more.
* Open data is fundamental to Data Science.
* Community Data License Agreement makes it easier to share open data.
* The IBM Data Asset eXchange (DAX) site contains high-quality open data sets.
* DAX open data sets include tutorial notebooks that provide basic and advanced walk-throughs for developers.
* DAX notebooks open in Watson Studio.
* Machine learning (ML) uses algorithms – also known as “models” – to identify patterns in the data.
* Types of ML are Supervised, Unsupervised, and Reinforcement.
* Supervised learning comprises two types of models, regression and classification.
* Deep learning refers to a general set of models and techniques that loosely emulate the way the human brain solves a wide range of problems.
* The Model Asset eXchange is a free, open-source repository for ready-to-use and customizable deep-learning microservices.
* MAX model-serving microservices are built and distributed on GitHub as open-source Docker images.
* You can use Red Hat OpenShift, a Kubernetes platform, to automate deployment, scaling, and management of microservices.
* Ml-exchange.org has multiple predefined models.