**Packages, API’s, Data Sets, and Models**

In this module, you will learn about the various libraries in data science. In addition, you will understand an API in relation to REST request and response. Further, in the module, you will explore open data sets on the Data Asset eXchange. Finally, you will learn how to use a machine learning model to solve a problem and navigate the Model Asset eXchange.

**Learning Objectives**

* List the tasks that a data scientist needs to perform to build a model​.
* List the various libraries: scientific, visualization, machine learning, and deep learning.
* Define REST API in relation to request and response.
* Describe data sets and sources of data.
* Explore open data sets on the Data Asset eXchange.
* Describe how to use a learning model to solve a problem.
* Navigate the Model Asset eXchange from IBM Research.

# **Packages, API’s, Data Sets, and Models**

## **Libraries for Data Science**

**libraries for data science** covers several important types of libraries in Python and other languages that help with data analysis and machine learning. Here’s a simple breakdown of the key points:

1. **Scientific Computing Libraries**:
   * These libraries help you perform complex calculations and data manipulation.
   * **Pandas**: This library is great for organizing and analyzing data in tables (called DataFrames). It makes tasks like cleaning and manipulating data easier.
   * **NumPy**: This library focuses on numerical data and allows you to work with arrays and perform mathematical operations quickly.
2. **Visualization Libraries**:
   * These libraries help you create visual representations of your data, like graphs and charts.
   * **Matplotlib**: A widely used library for making various types of plots and graphs. It’s very customizable.
   * **Seaborn**: Built on top of Matplotlib, it makes it easier to create attractive and informative visualizations, like heat maps.
3. **Machine Learning Libraries**:
   * These libraries provide tools to build models that can learn from data.
   * **Scikit-learn**: A user-friendly library for machine learning that includes tools for tasks like classification and regression.
   * **Keras**: A high-level library for building deep learning models easily. It simplifies the process of creating complex models.
4. **Deep Learning Libraries**:
   * **TensorFlow**: A powerful library for building and deploying deep learning models, often used in large-scale applications.
   * **PyTorch**: This library is popular for research and experimentation, making it easy to test new ideas.
5. **Other Languages**:
   * The video also mentions libraries used in other programming languages, like R and Scala, which have their own tools for data visualization and machine learning.

Overall, the video emphasizes that these libraries are essential tools for data scientists, allowing them to analyze data, visualize results, and build predictive models efficiently.

**libraries for data science**:

* **Scientific Computing Libraries**:
  + **Pandas**: Used for organizing and analyzing data in tables (DataFrames).
  + **NumPy**: Focuses on numerical data, allowing operations on arrays and matrices.
* **Visualization Libraries**:
  + **Matplotlib**: A versatile library for creating various types of plots and graphs.
  + **Seaborn**: Built on Matplotlib, it simplifies the creation of attractive visualizations.
* **Machine Learning Libraries**:
  + **Scikit-learn**: User-friendly library for machine learning tasks like classification and regression.
  + **Keras**: High-level library for building deep learning models easily.
* **Deep Learning Libraries**:
  + **TensorFlow**: Powerful for building and deploying deep learning models.
  + **PyTorch**: Popular for research and experimentation in deep learning.
* **Other Languages**: The video also mentions libraries in languages like R and Scala for data visualization and machine learning.

These libraries are essential tools for data scientists, enabling efficient data analysis, visualization, and model building.

## **Application Programming Interfaces(APIs)**

**Application Programming Interfaces (APIs)** and explains how they work. Here’s a simplified explanation and summary:

Simplified Explanation:

* **API**: Think of an API as a bridge that allows two different software programs to talk to each other. It helps them share information without needing to know how each other works behind the scenes.
* **Pandas Library Example**: When you use the Pandas library in Python, you can process data easily. The API lets you interact with the library's components without needing to understand all the details of how they are built.
* **REST APIs**: These are a specific type of API that work over the internet. They allow your program (the client) to request information from a web service (the resource) and get a response back. This communication follows certain rules.
* **HTTP Methods**: Data is sent and received using HTTP messages, often in a format called JSON, which is easy for both humans and machines to read.
* **Examples**: The video gives examples of REST APIs like:
  + **Watson Speech-to-Text API**: Converts spoken words into written text.
  + **Watson Language Translator API**: Translates text from one language to another.

Summary:

* An **API** allows communication between software programs.
* **REST APIs** enable internet-based communication, using HTTP methods to send requests and receive responses.
* The video highlights how APIs are used in libraries like Pandas and provides examples of specific APIs like Watson's Speech-to-Text and Language Translator.

## **Data Sets – Powering Data Science**

Explanation:

* A **data set** is a collection of information organized in a structured way, often in tables with rows and columns.
* **Types of Data**:
  + **Tabular Data**: Like a spreadsheet, where each row is an observation (e.g., weather data) and each column is a feature (e.g., temperature, humidity).
  + **Hierarchical Data**: Organized like a tree, showing relationships (e.g., family trees).
  + **Network Data**: Represented as graphs, showing connections (e.g., social media connections).
* **Open Data**: Many organizations share data publicly, allowing researchers and data scientists to access valuable information for analysis and insights.
* **Community Data License Agreement (CDLA)**: This agreement helps in sharing data legally. There are two types:
  + **CDLA-Sharing**: You can use and modify the data, but must share any changes under the same terms.
  + **CDLA-Permissive**: You can use and modify the data without needing to share changes.

Summary:

* The video explains what a data set is and its types, including tabular, hierarchical, and network data.
* It highlights the significance of open data for research and data science.
* It introduces the Community Data License Agreement, which facilitates the legal sharing of data.

## **Additional sources of data sets**

**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.

## **Sharing Enterprise Data: Data Asset eXchange**

**IBM Data Asset Exchange (DAX)**, which is a platform that provides access to a collection of high-quality open data sets. Here’s a simplified summary:

* **Purpose of DAX**: It helps users find and use open data sets that are reliable and have clear usage terms.
* **Types of Data**: DAX includes various types of data such as images, videos, text, and audio.
* **Tutorial Notebooks**: Each data set often comes with tutorial notebooks that guide users on how to clean, preprocess, and analyze the data.
* **Accessing Data**: Users can download data sets, run associated notebooks in **Watson Studio**, and explore metadata.
* **Examples**: The video mentions a specific data set, the **NOAA weather data**, which can be downloaded and analyzed using notebooks.

In summary, the video introduces DAX as a valuable resource for developers to access and work with open data sets, providing tools and tutorials to facilitate data analysis.

## **Machine Learning Model: Learning from Models to make Predictions**

Simple Explanation:

* **Machine Learning (ML)**: It's a way for computers to learn from data and make predictions without being explicitly programmed.
* **Model Training**: This is the process where a machine learning model learns patterns from data.
* **Types of Machine Learning**:
  + **Supervised Learning**: The model learns from labeled data (data with correct answers). It can be:
    - **Regression**: Predicting a number (like house prices).
    - **Classification**: Categorizing data (like identifying spam emails).
  + **Unsupervised Learning**: The model finds patterns in data without labels. For example, grouping similar customers based on their shopping habits.
  + **Reinforcement Learning**: The model learns by trying actions and receiving rewards, similar to how a mouse learns to navigate a maze for cheese.
* **Deep Learning**: A special type of machine learning that mimics how the human brain works, often used for complex tasks like image and speech recognition.

Summary:

The video introduces machine learning models, explaining how they learn from data to make predictions. It covers three main types of learning: supervised, unsupervised, and reinforcement learning, along with a brief mention of deep learning. Supervised learning involves labeled data for tasks like regression and classification, while unsupervised learning identifies patterns in unlabeled data. Reinforcement learning is about learning through rewards. Deep learning is highlighted as a powerful method for complex problem-solving.

## **The Model Asset eXchange**

**Model Asset eXchange (MAX)**, which is a free resource from IBM for deep learning models. Here’s a simplified explanation and summary:

Simplified Explanation:

* **Model Asset eXchange (MAX)** is a place where you can find ready-to-use deep learning models.
* Instead of creating a model from scratch (which takes a lot of time and data), you can use pre-trained models that are already available.
* These models can help with tasks like recognizing images, audio, and text.
* Each model comes with a service that processes input data, runs it through the model, and then gives you the output.
* The models are packaged as **Docker images**, which makes them easy to use and deploy on different platforms, including cloud services.
* You can also use **Kubernetes** to manage these models efficiently.

Summary of the Video:

* The video introduces the **Model Asset eXchange** as a free, open-source repository for deep learning microservices.
* It explains the benefits of using pre-trained models to save time and resources.
* It describes the components of a model-serving microservice, including input processing, model execution, and output processing.
* It highlights the use of Docker for packaging and deploying models and mentions Kubernetes for managing these deployments.

## **Getting Started with MAX and DAX**

**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:**

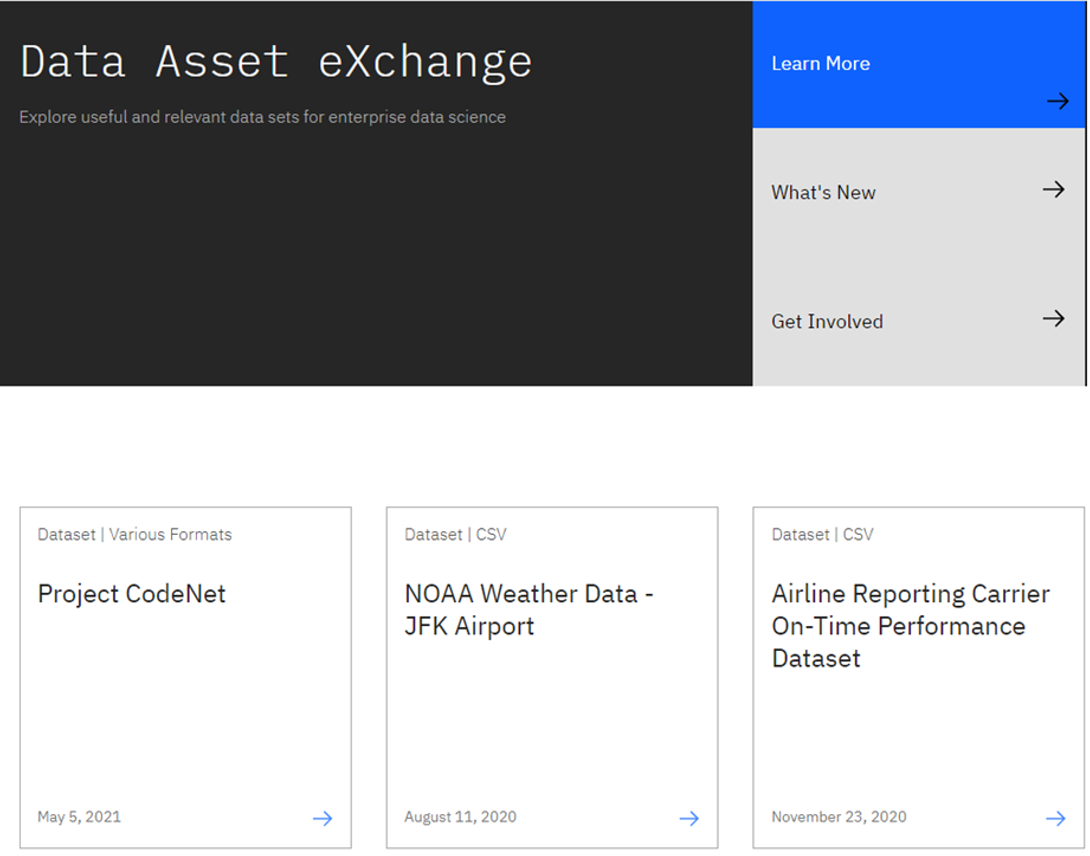
* Find ready-to-use deep learning models on the Model Asset Exchange.
* Explore the deep learning model trained to detect objects in an image.

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

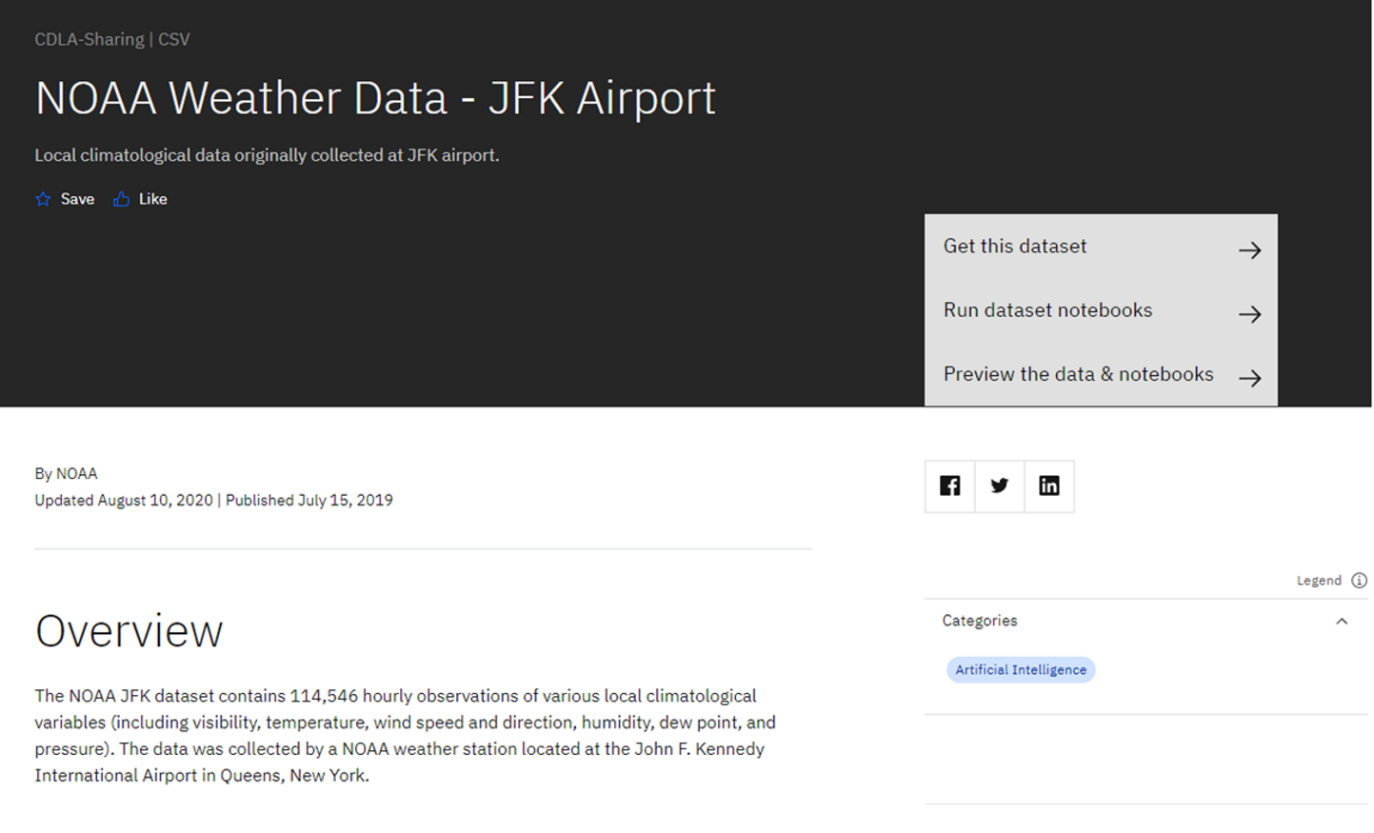
**Exercise 1: Explore deep learning datasets**

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/exchanges/data/> in your web browser.The IBM Data Asset eXchange (DAX) home page is displayed. This is an online hub for developers and data scientists to find free and open data sets under open data licenses. These datasets can be used to train models to perform document layout analysis, natural language processing (NLP), time series analysis, and more.



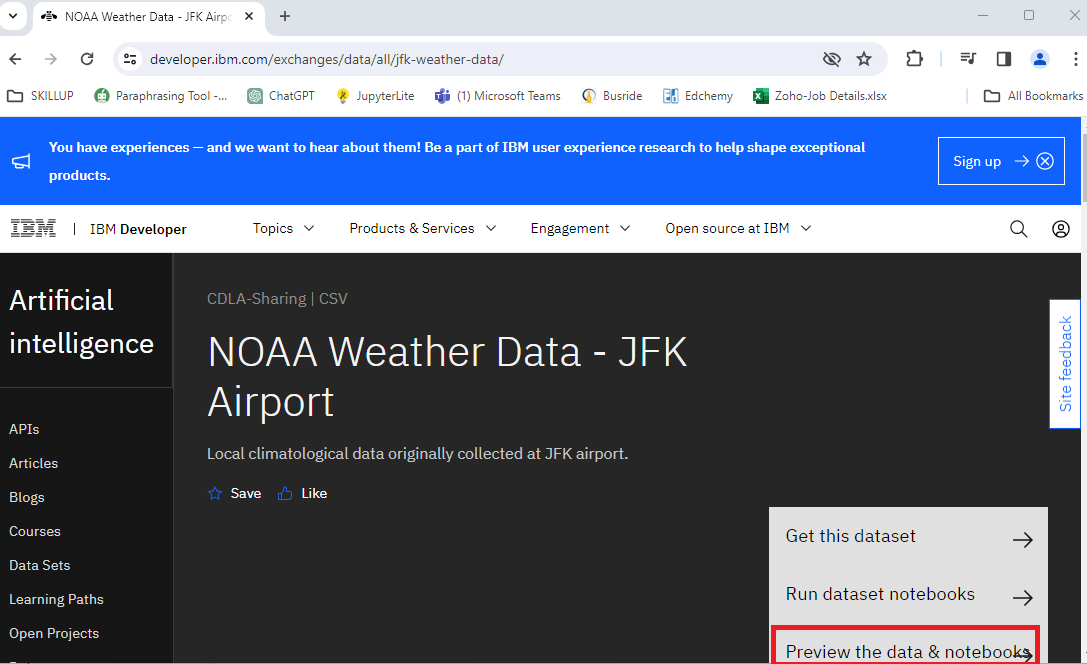
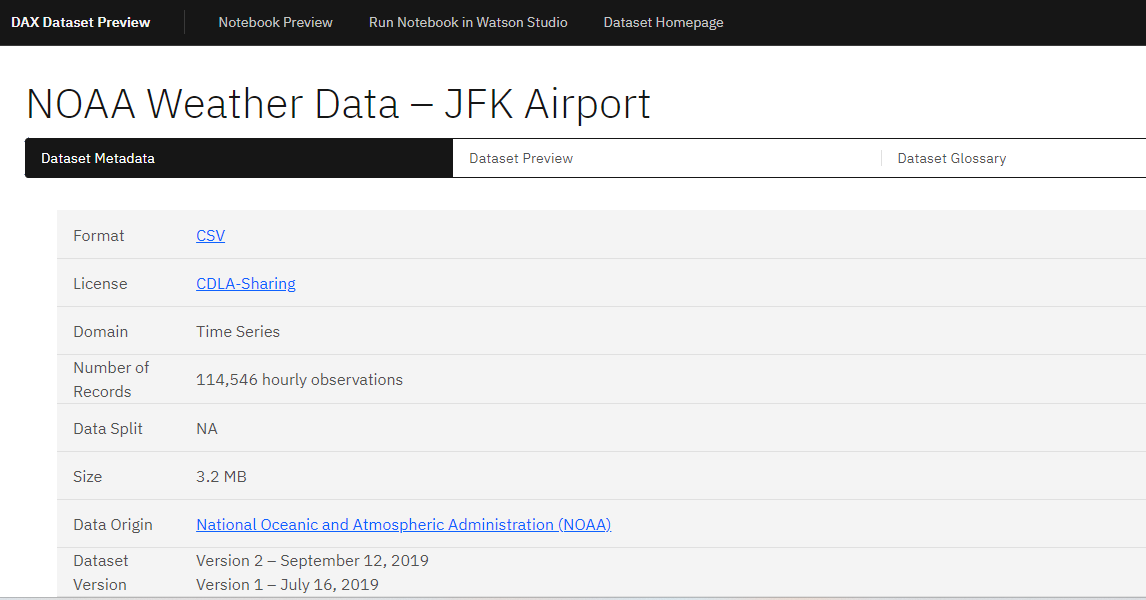
1. In this activity, you will explore **NOAA Weather Data dataset**.
2. Open the NOAA Weather Data dataset (<https://developer.ibm.com/exchanges/data/all/jfk-weather-data/>), which contains data from a weather station at the John F. Kennedy Airport in New York spanning eight years.  
   The dataset was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



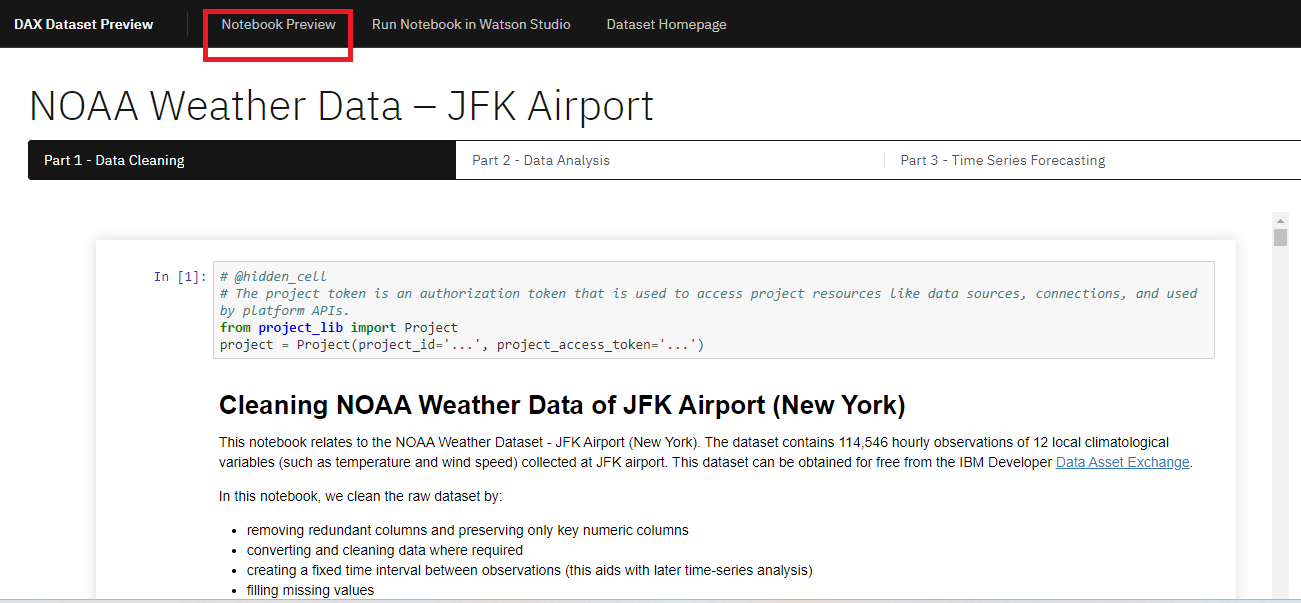
1. Inspect the dataset's metadata.  
   This dataset is stored as tabular data and formatted as a comma separated value (CSV) file, which is a very popular basic data exchange format.

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. Most datasets are complemented by Python notebooks that you can use to explore, pre-process, and analyze the data. The notebooks are hosted on Watson Studio, IBM's Data Science platform. Later in this course, you'll learn more about Watson Studio notebooks, how to sign up and how to run notebooks on them.
2. For now, you can preview the dataset and the notebook (or notebooks) by clicking the **Preview the data and notebooks** as shown in the screenshot below.

  
  
This will display dataset metadata, sample records and glossary as shown in the screenshot below.  


Now here, you click on the **Notebook Preview** option on top to view the notebook hosted with this dataset. Explore the steps followed in this notebook to clean data before performing data analysis on this dataset.



**This concludes Exercise 1 of this lab, which introduced the Data Asset Exchange. You may proceed to Exercise 2.**

1. [Optional] If you have already registered and are comfortable working with notebooks on Watson Studio, you can open the link using the Run Notebook in Watson Studio option and import it into a project. If you are not acquainted with IBM Watson Studio, you can skip this step. Detailed guidance on signing up and getting started with IBM Watson Studio will be provided in Module 7 of this course.

**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.

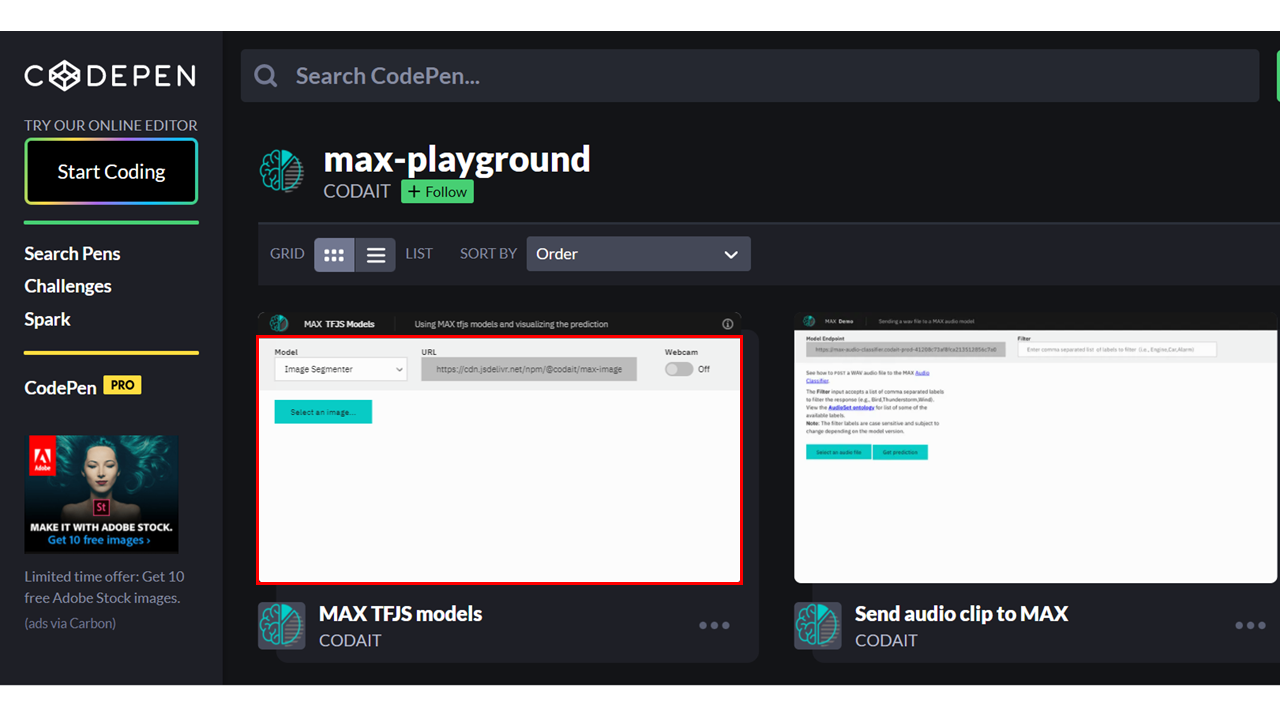
For more details, please visit - <https://github.com/CODAIT/max-central-repo> webpage.

The curated list includes deployable models, which you can run as a microservice locally or in the cloud on Docker or Kubernetes, and trainable models where you can use your own data to train the models. Some of the models are already built for you to test. Let's test one of the models.

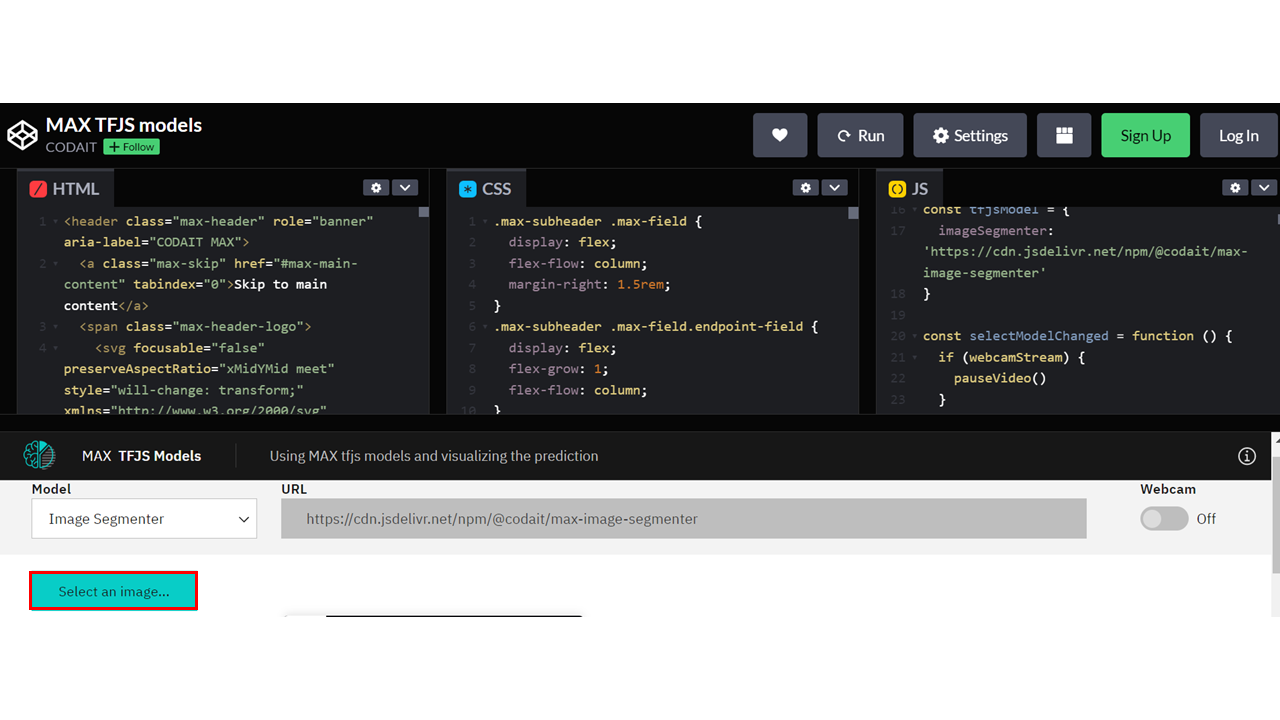
In this exercise, explore the **Object Detector** model hosted on CodePen platform. This model recognizes the objects present in an image. The model consists of a deep convolutional net base model for image feature extraction, together with additional convolutional layers specialized for the task of 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.

**CodePen** is a social development environment. At its heart, it allows 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.

1. Navigate to **[CodePen](https://codepen.io/collection/DzdpJM/" \t "_blank)** webpage.
2. Select **MAX TFJS models** as shown in the screenshot below. Here the **Image Segmenter**, 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.



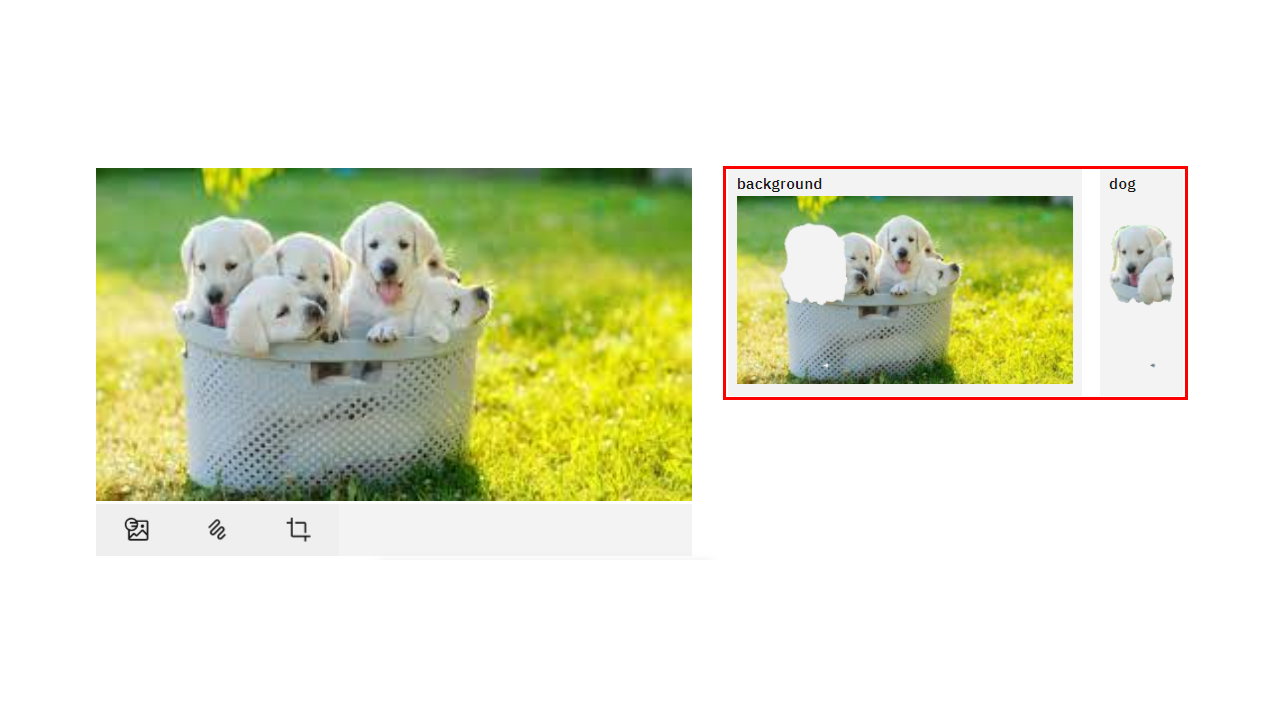
1. Click on **Select Image** and upload an image. You may choose images with a person, dog, cat, truck, car, and so on, which are labels the model has been trained on.

Click here for all the labels the model is trained on

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



You can see the output of the prediction on the basis of the uploaded image.



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 (MAX).

*You can also watch a demo of the object detector model*[*here*](https://video.ibm.com/recorded/128825527?utm_source=skills_network&utm_content=in_lab_content_link&utm_id=Lab-IBMDeveloperSkillsNetwork-DS0105EN-SkillsNetwork)*.*

## **Summary**

**Module 3 Summary**

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

* Python offers a diverse library ecosystem for data science, covering scientific computing (Pandas, NumPy), visualization (Matplotlib, Seaborn), and high-level machine learning (Scikit-learn). These libraries offer tools for data manipulation, mathematical operations, and simplified machine learning model development.
* Application Programming Interfaces (APIs) facilitate communication between software components. REST APIs, specifically, facilitate internet communication and access resources like storage. Key API terms include client (user or code accessing it), resource (service or data), and endpoint (API's URL).
* Machine learning models analyze data and identify patterns to make predictions and automate complex tasks—the three fundamental types of machine learning are supervised, unsupervised, and reinforcement learning. Supervised learning includes regression and classification models for predictive modeling and pattern recognition. Deep learning, an advanced subset of machine learning, mimics the brain's processing, enabling intricate problem-solving in various domains.
* The Community Data License Agreement (CDLA) facilitates open data sharing by providing clear licensing terms for distribution and use, and the IBM Data Asset eXchange (DAX) site contains high-quality open data sets.
* The Model Asset eXchange (MAX) provides a wealth of pre-trained deep learning models, empowering developers with readily deployable solutions for various business challenges.