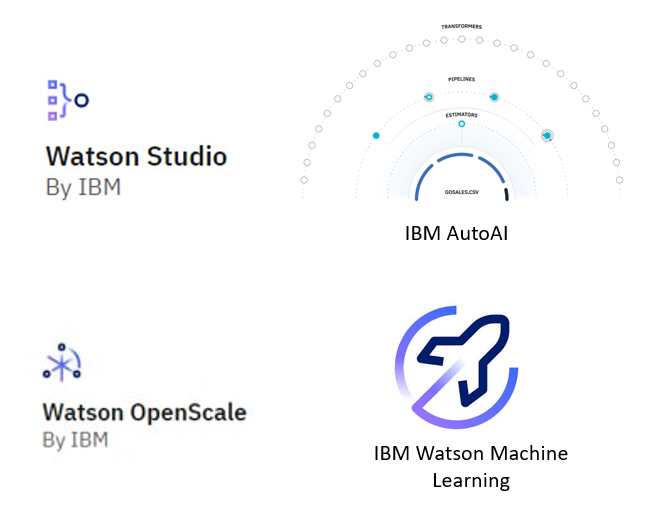
**Model Development**

Before the timestamp **2:46** in the previous video [Open Source Tools for Data Science - Part 1](https://www.coursera.org/learn/open-source-tools-for-data-science/lecture/oDcUE/open-source-tools-for-data-science-part-1)

, there is a gap in the information concerning **Model Development**. The missing information has now been included below.

IBM offers various tools and platforms tailored for model development across various domains. Here are a few examples:



* **IBM Watson Studio:** Engineered as an integrated environment, Watson Studio simplifies developing, training, and deploying models. It boasts support for multiple languages and frameworks, such as Python, R, and TensorFlow, alongside collaboration features, data preparation tools, and versatile deployment options.
* **IBM AutoAI:** A notable feature embedded within Watson Studio, IBM AutoAI streamlines the machine learning model construction process. By dynamically exploring various algorithms and hyperparameters, it aims to identify the optimal model for a given dataset.
* **IBM Watson OpenScale:** As a platform for overseeing and managing AI models in production, Watson OpenScale plays a pivotal role in ensuring model fairness, explainability, and bias mitigation. It furnishes insights into model performance and drift over time, facilitating informed decision-making.
* **IBM Watson Machine Learning:** Watson Machine Learning, available as a service on the IBM Cloud platform, enables users to scale their training and deployment of machine learning models. It seamlessly supports popular frameworks like TensorFlow, PyTorch, and scikit-learn, and offers APIs for seamless integration with other applications.

**Module 1 Summary**

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

* The Data Science Task Categories include:
  + Data Management - storage, management and retrieval of data
  + Data Integration and Transformation - streamline data pipelines and automate data processing tasks
  + Data Visualization - provide graphical representation of data and assist with communicating insights
  + Modelling - enable Building, Deployment, Monitoring and Assessment of Data and Machine Learning models
* Data Science Tasks support the following:
  + Code Asset Management - store & manage code, track changes and allow collaborative development
  + Data Asset Management - organize and manage data, provide access control, and backup assets
  + Development Environments - develop, test and deploy code
  + Execution Environments - provide computational resources and run the code

The data science ecosystem consists of many open source and commercial options, and include both traditional desktop applications and server-based tools, as well as cloud-based services that can be accessed using web-browsers and mobile interfaces.

**Data Management Tools**: include Relational Databases, NoSQL Databases, and Big Data platforms:

* MySQL, and PostgreSQL are examples of Open Source Relational Database Management Systems (RDBMS), and IBM Db2 and SQL Server are examples of commercial RDBMSes and are also available as Cloud services.
* MongoDB and Apache Cassandra are examples of NoSQL databases.
* Apache Hadoop and Apache Spark are used for Big Data analytics.

**Data Integration and Transformation Tools:** include Apache Airflow and Apache Kafka.

**Data Visualization Tools:** include commercial offerings such as Cognos Analytics, Tableau and PowerBI and can be used for building dynamic and interactive dashboards.

**Code Asset Management Tools:** Git is an essential code asset management tool. GitHub is a popular web-based platform for storing and managing source code. Its features make it an ideal tool for collaborative software development, including version control, issue tracking, and project management.

**Development Environments:** Popular development environments for Data Science include Jupyter Notebooks and RStudio.

* Jupyter Notebooks provides an interactive environment for creating and sharing code, descriptive text, data visualizations, and other computational artifacts in a web-browser based interface.
* RStudio is an integrated development environment (IDE) designed specifically for working with the R programming language, which is a popular tool for statistical computing and data analysis.

**Module 2 Summary**

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

* You should select a language to learn depending on your needs, the problems you are trying to solve, and whom you are solving them for.
* The popular languages are Python, R, SQL, Scala, Java, C++, and Julia.
* For data science, you can use Python's scientific computing libraries like Pandas, NumPy, SciPy, and Matplotlib.
* Python can also be used for Natural Language Processing (NLP) using the Natural Language Toolkit (NLTK).
* Python is open source, and R is free software.
* R language’s array-oriented syntax makes it easier to translate from math to code for learners with no or minimal programming background.
* SQL is different from other software development languages because it is a non-procedural language.
* SQL was designed for managing data in relational databases.
* If you learn SQL and use it with one database, you can apply your SQL knowledge with many other databases easily.
* Data science tools built with Java include Weka, Java-ML, Apache MLlib, and Deeplearning4.
* For data science, popular program built with Scala is Apache Spark which includes Shark, MLlib, GraphX, and Spark Streaming.
* Programs built for Data Science with JavaScript include TensorFlow.js and R-js.
* One great application of Julia for Data Science is JuliaDB.

**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.***

**Author(s)**

[Lakshmi Holla](https://www.linkedin.com/in/lakshmi-holla-b39062149/)

**Other Contributor(s)**

[Malika Singla](https://www.linkedin.com/in/malika-goyal-04798622/)

**Changelog**

| **Date** | **Version** | **Changed by** | **Change Description** |
| --- | --- | --- | --- |
| 2022-12-14 | 0.1 | Lakshmi Holla | Initial version created |

**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:**

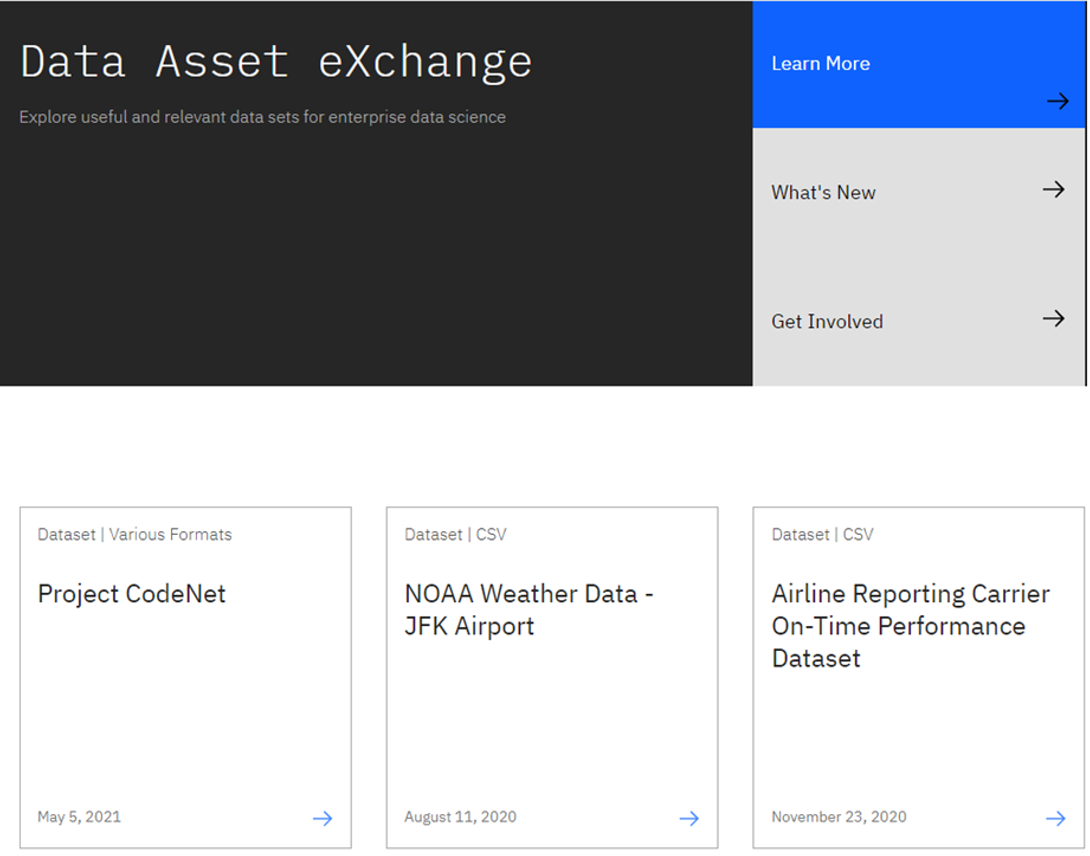
* Find ready-to-use deep learning models on the Model Asset Exchange.
* Explore the deep learning model trained to detect objects in an image.

*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.

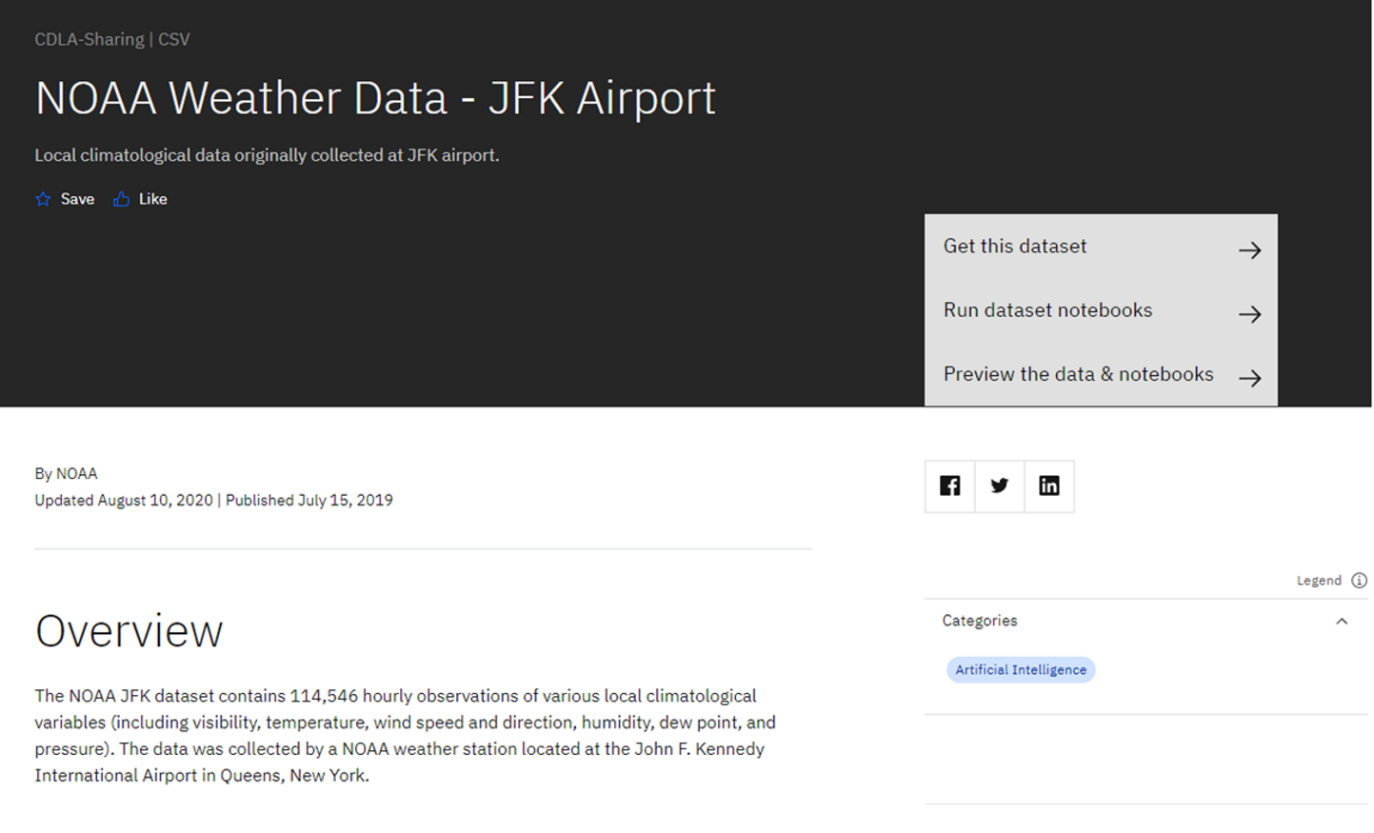
1. Open <https://developer.ibm.com/> in your web browser.
2. From the main menu select **Open Source at IBM > Data Asset eXchange**. The DAX home page is displayed.



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.  
   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/>).



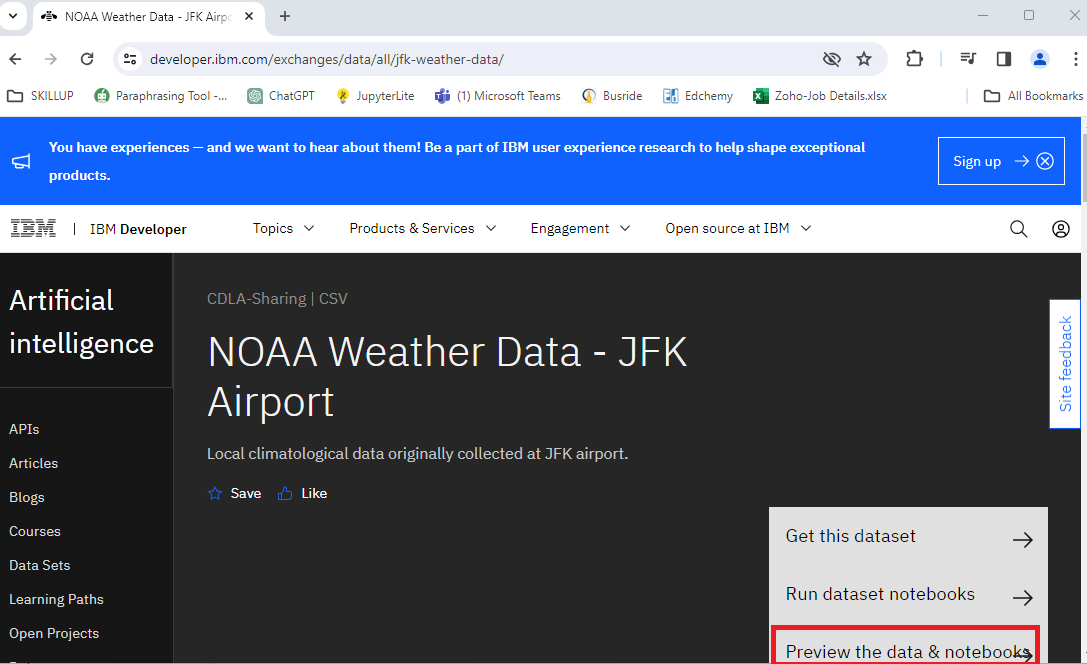
1. Inspect the dataset's metadata.

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.

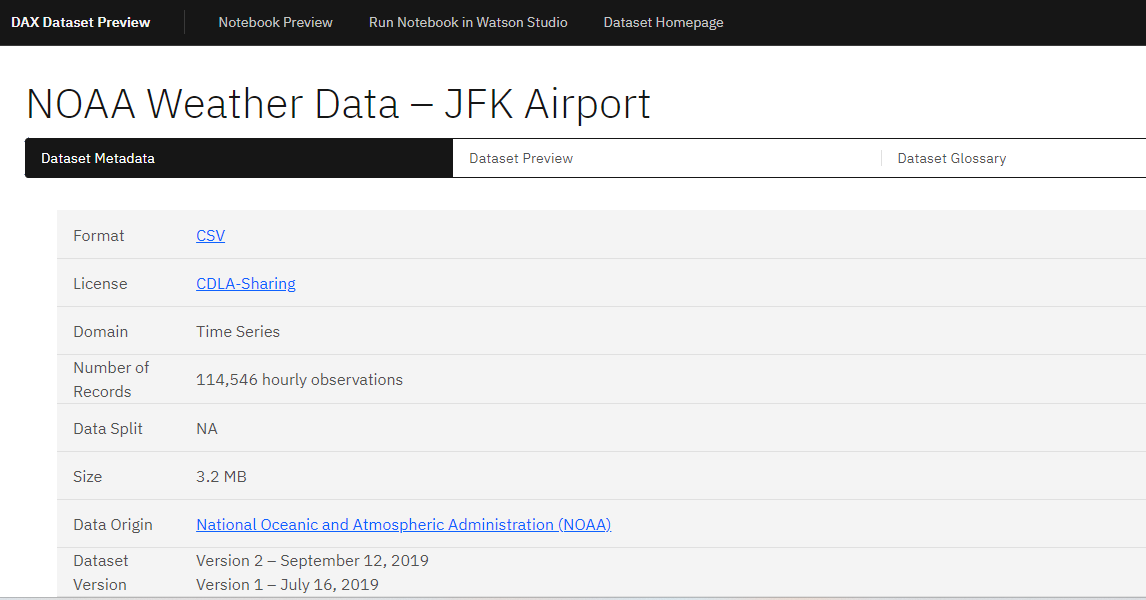
You can download the dataset using the **Get this dataset** link. Datasets are stored as compressed archives, which you can extract using any utility that supports the targz format. If you are not familiar with this file format, take a look at this short open source tutorial <https://opensource.com/article/17/7/how-unzip-targz-file>.

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.

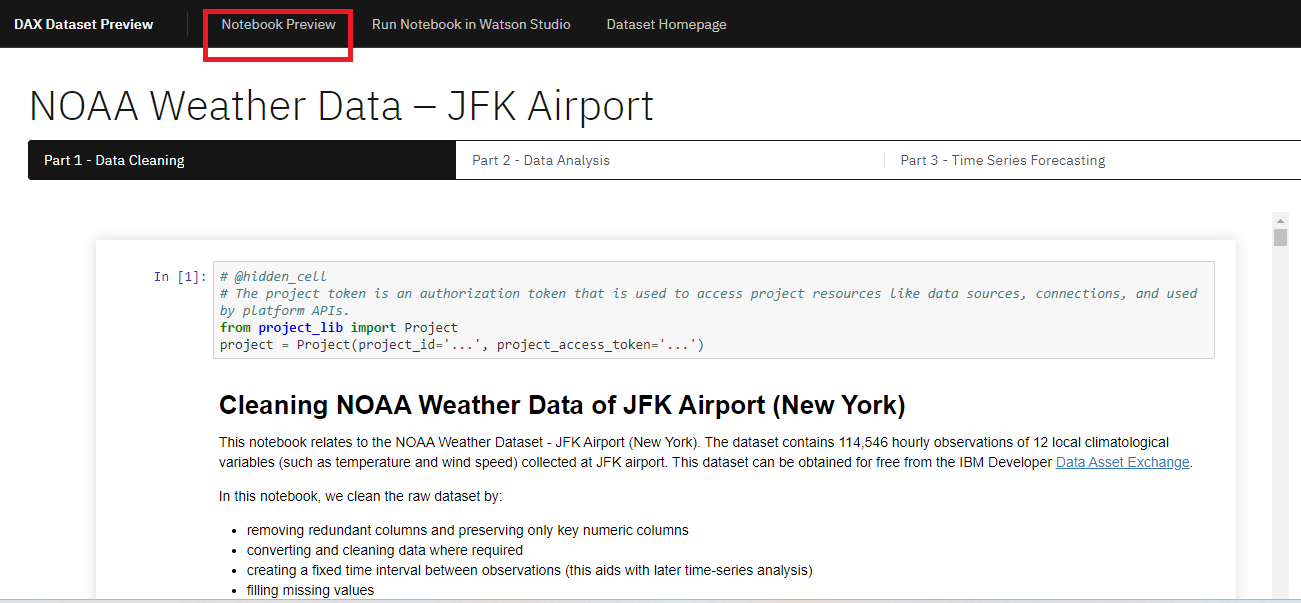
For now, you can preview the dataset and the notebook (or notebooks) by clicking the **Preview the data and notebooks** as shown in the screenshot below.



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.



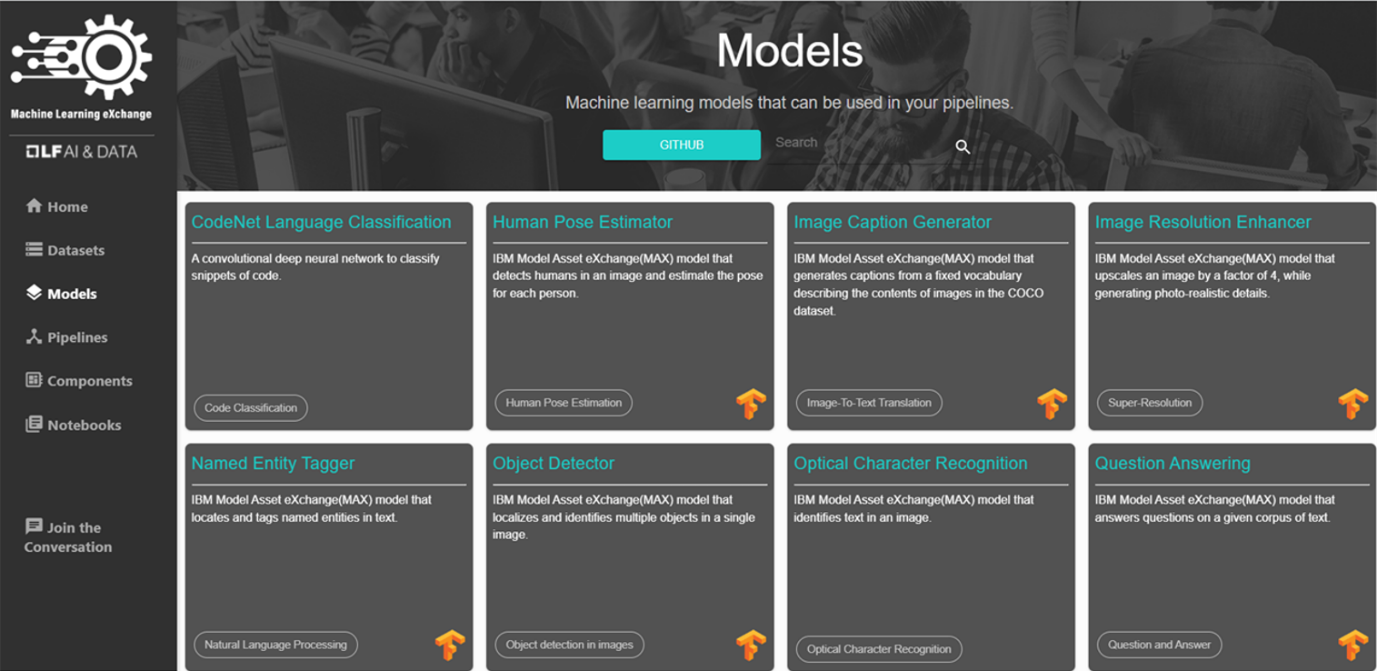
**This concludes Exercise 1 of this lab, which introduced the Data Asset Exchange. You may proceed to Exercise 2.**

1. [Optional] If you have already registered and are comfortable working with notebooks and 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 Week 7 of this course.

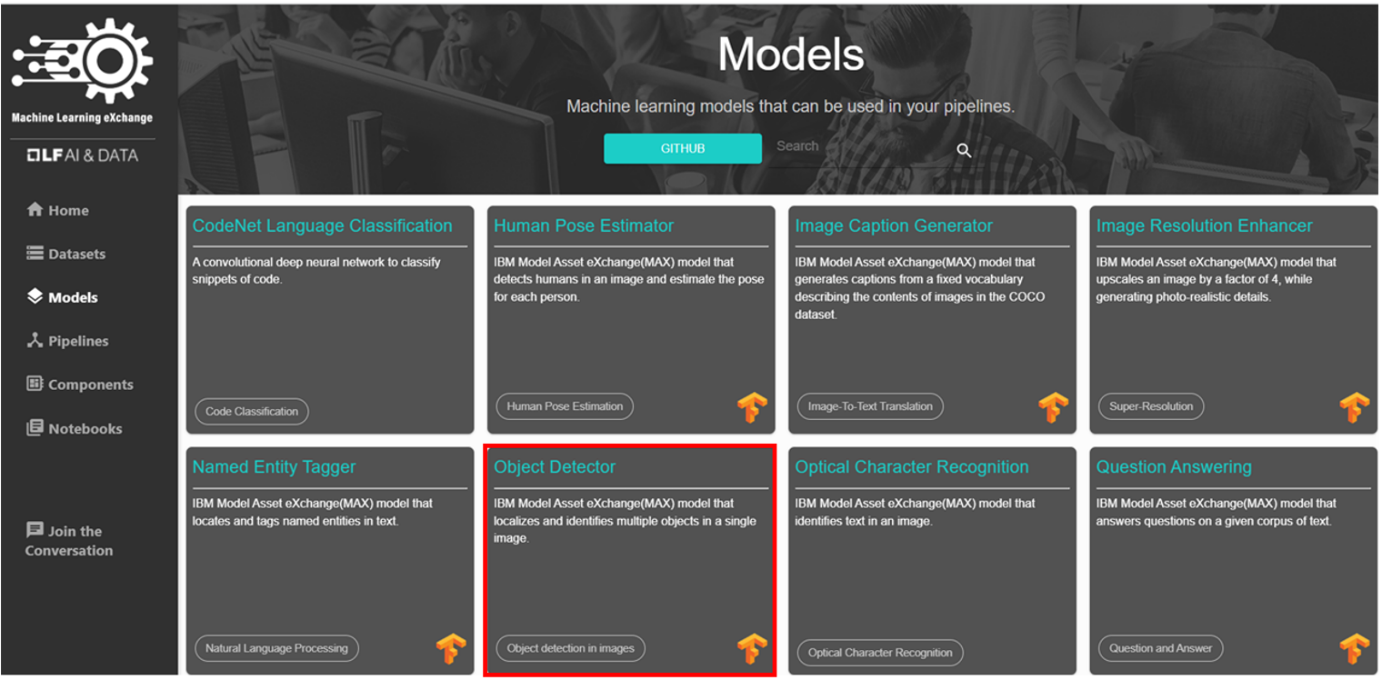
**Exercise 2 - Explore deep learning models**

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.

1. Open <http://ml-exchange.org/models> 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.*

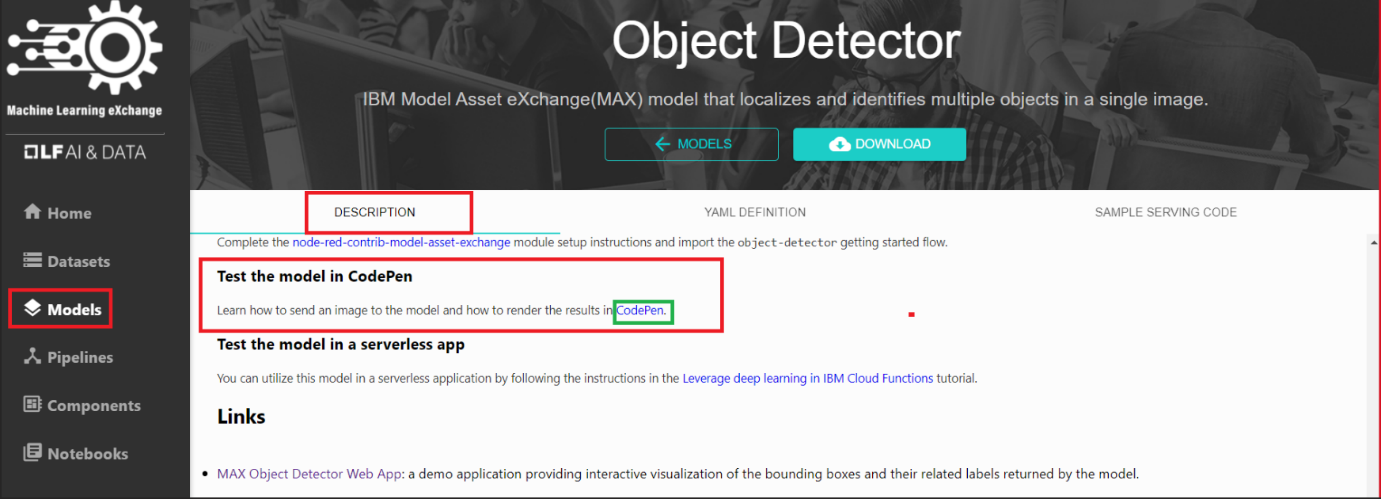


1. Select the **Object Detector** model from the list of available options.



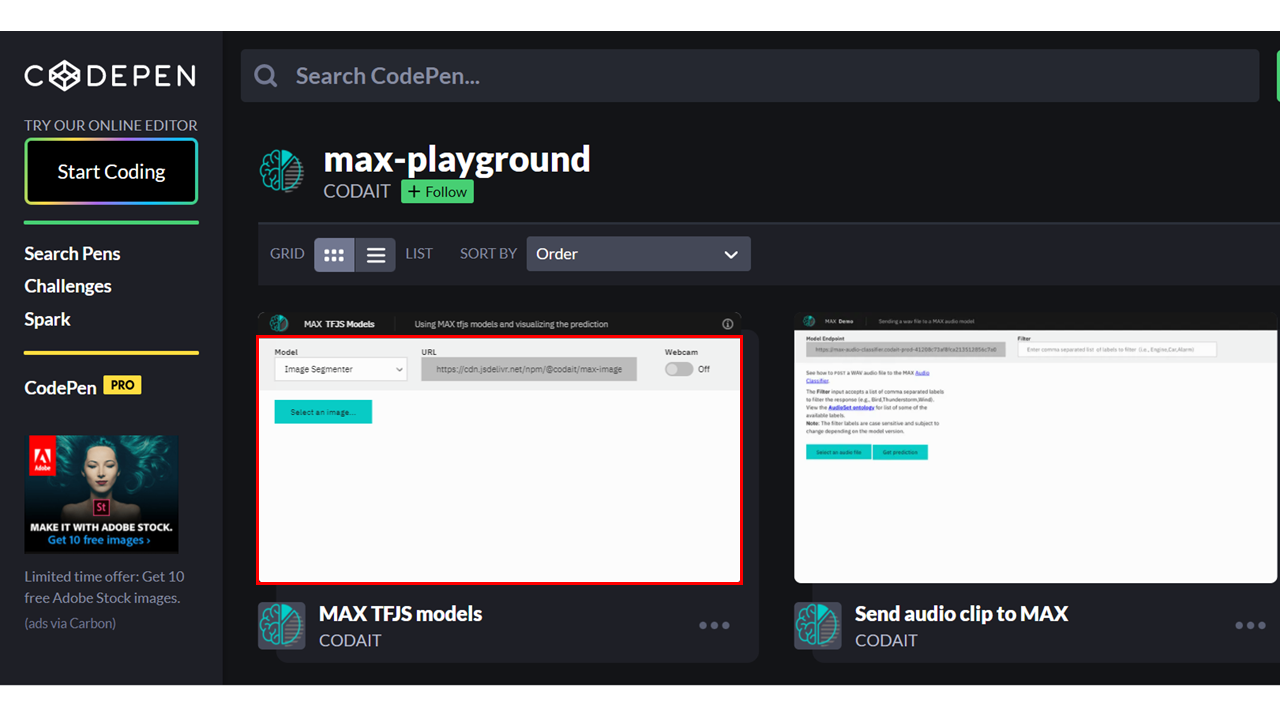
This model recognizes the objects present in an image. The model consists of a deep convolutional net base model for image feature extraction, together with additional convolutional layers specialized for the task of object detection, trained on the COCO data set. The input to the model is an image,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

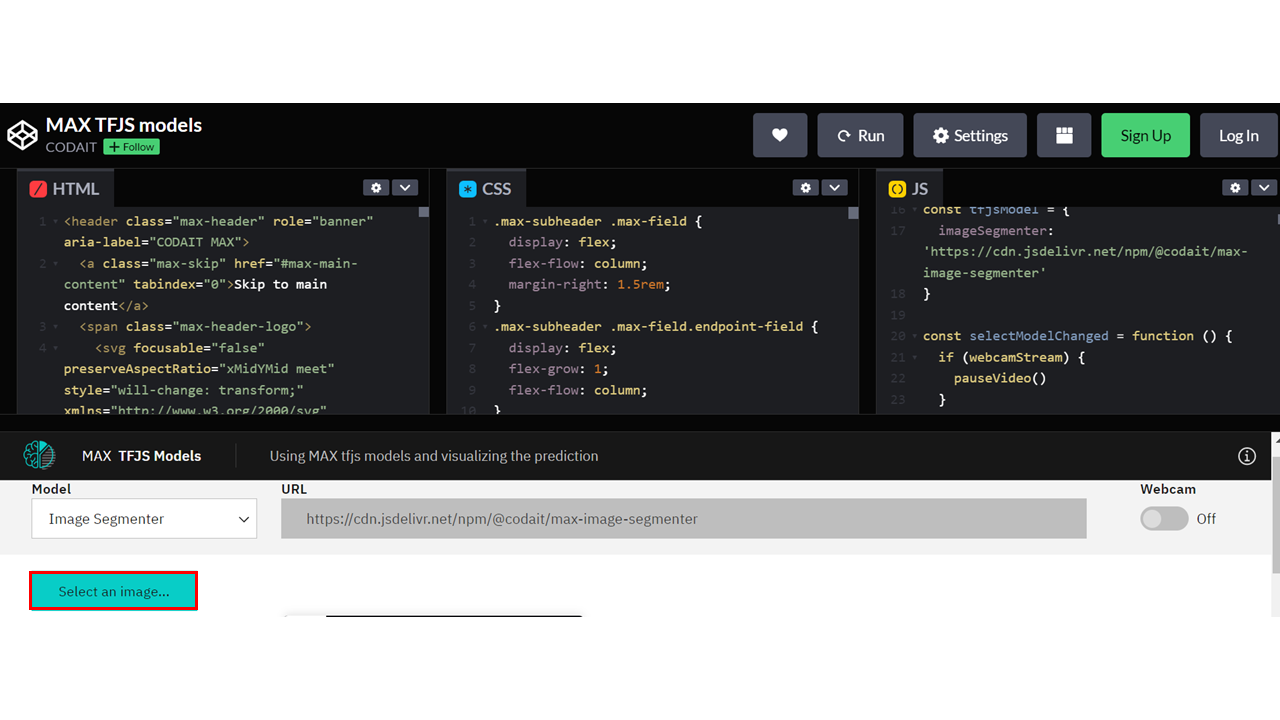


CodePen is a social development environment. At its heart, it allows you to write code in the browser and see the results of it as you build. It is a useful and liberating online code editor for developers of any skill and is particularly empowering for people learning to code.

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.

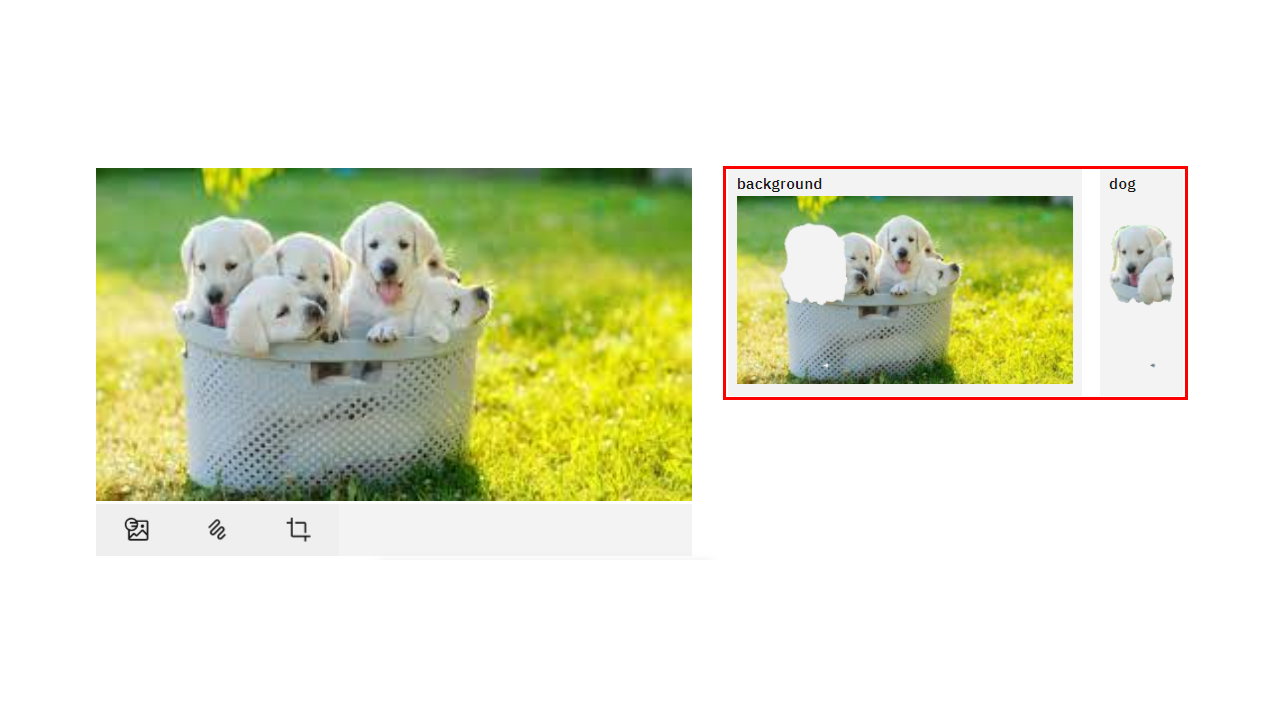
Click here for all the labels the model is trained on 

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_source=skills_network&utm_content=in_lab_content_link&utm_id=Lab-IBMDeveloperSkillsNetwork-DS0105EN-SkillsNetwork).

**Author(s)**

**Joseph Santarcangelo**

**Other Contributor(s)**

Lavanya

**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.

**Module 4 Summary**

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

* Jupyter Notebooks are used in Data Science for recording experiments and projects.
* Jupyter Lab is compatible with many files and Data Science languages.
* There are different ways to install and use Jupyter Notebooks.
* How to run, delete, and insert a code cell in Jupyter Notebooks.
* How to run multiple notebooks at the same time.
* How to present a notebook using a combination of Markdown and code cells.
* How to shut down your notebook sessions after you have completed your work on them.
* Jupyter implements a two-process model with a kernel and a client.
* The notebook server is responsible for saving and loading the notebooks.
* The kernel executes the cells of code contained in the Notebook.
* The Jupyter architecture uses the NB convert tool to convert files to other formats.
* Jupyter implements a two-process model with a kernel and a client.
* The Notebook server is responsible for saving and loading the notebooks.
* The Jupyter architecture uses the NB convert tool to convert files to other formats.
* The Anaconda Navigator GUI can launch multiple applications on a local device.
* Jupyter environments in the Anaconda Navigator include JupyterLab and VS Code.
* You can download Jupyter environments separately from the Anaconda Navigator, but they may not be configured properly.
* The Anaconda Navigator GUI can launch multiple applications.
* Additional open-source Jupyter environments include JupyterLab, JupyterLite, VS Code, and Google Colaboratory.
* JupyterLite is a browser-based tool.