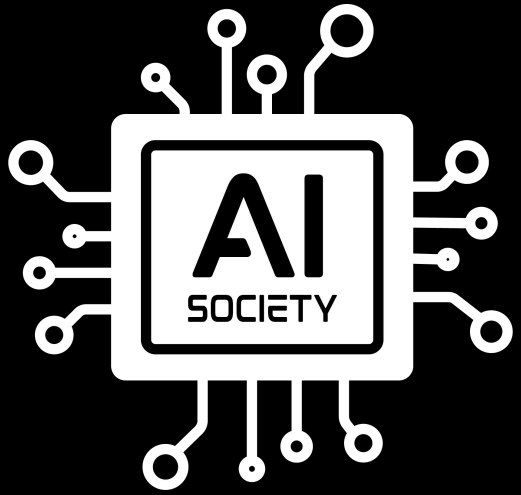


Introduction to the Machine Learning Lab

Week #1 - 4 Sept ‘24

ML Lab

What is AI Society??

Club Mission

AI Society at Arizona State University is committed to demystifying artificial intelligence and machine learning through hands-on, project-based learning. We aim to make these advanced technologies accessible to all, regardless of prior programming experience, fostering a community of innovative thinkers ready to tackle real-world challenges.

What to expect from this series??

Learn to manage and prepare data, covering everything from loading data to normalization and handling missing values.

Data Cleaning and Processing

Feature Engineering and Category Prediction

Dive into creating and selecting meaningful features from data, using techniques like tokenization and PCA. Implement and evaluate classification models such as Logistic Regression and Random Forest.

Explore clustering algorithms and dimensionality reduction to discover patterns and groupings in data.

Supervised and Unsupervised Learning Techniques

Building a Recommender System

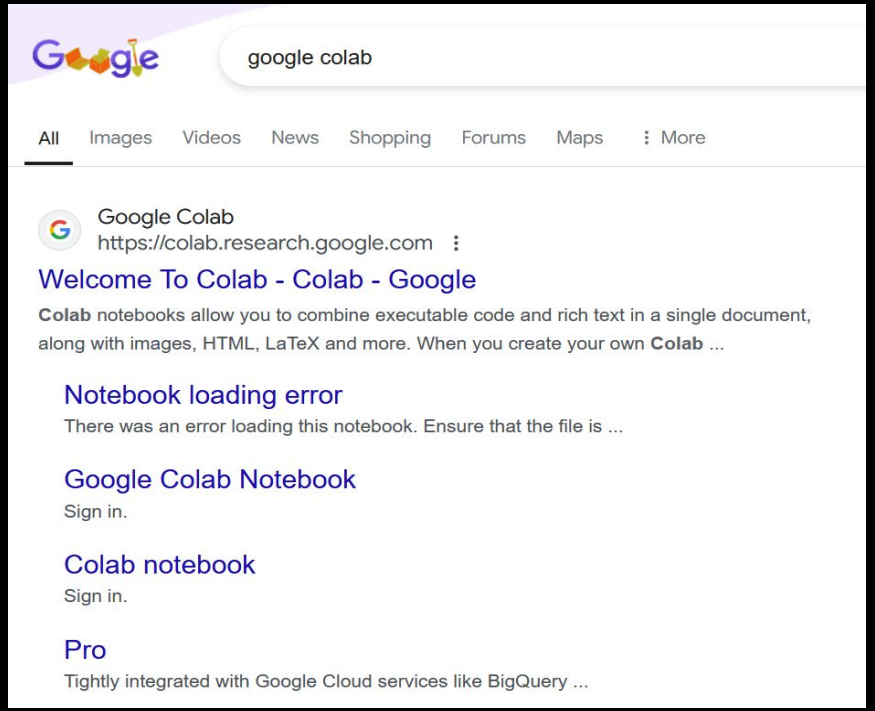
Combine all learned techniques to create a sophisticated recommender system in our final project.

Google Colab Setup

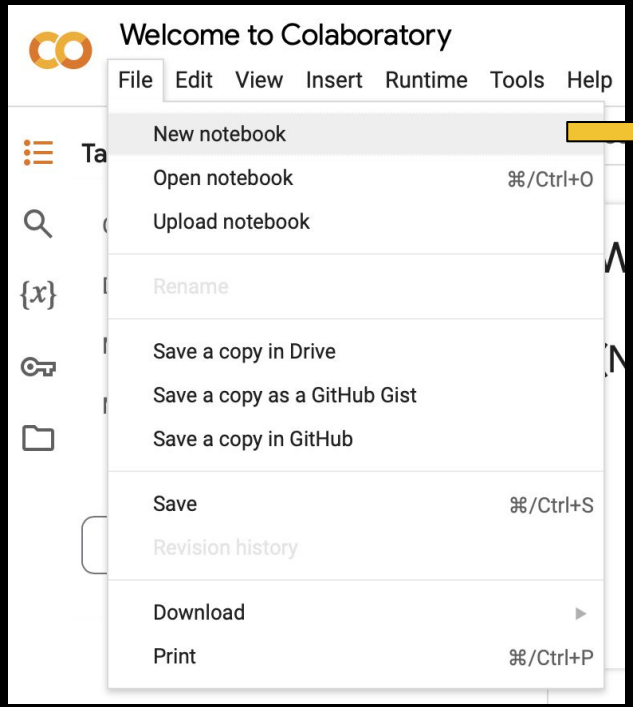
What is Colab ..... and why are we using it

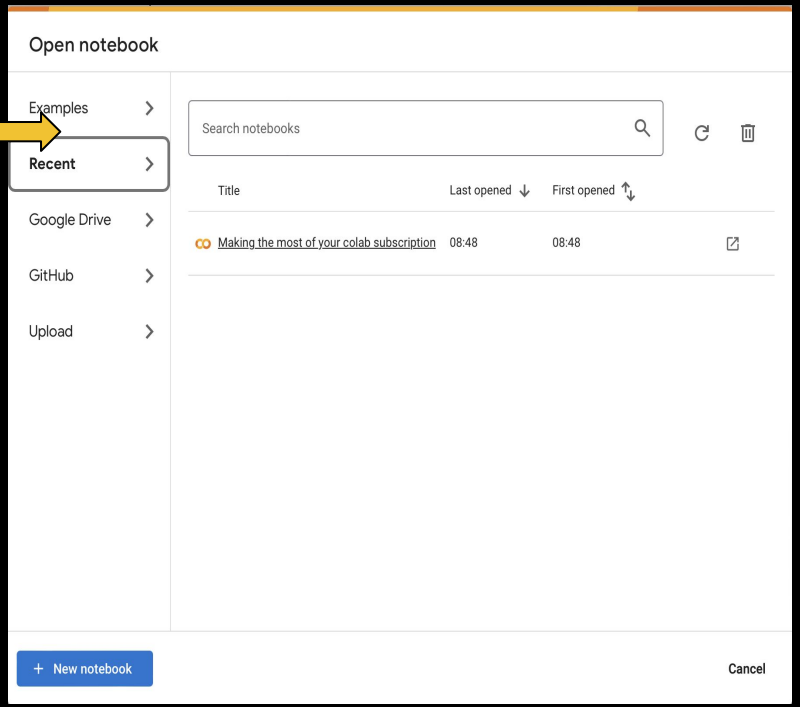
Google Colab, or Colaboratory, is a free, browser-based tool that allows users to write and run Python code. It's built around Project Jupyter code and allows users to run Jupyter notebooks without the need for local software installation.

Step 1: Accessing Google Collab

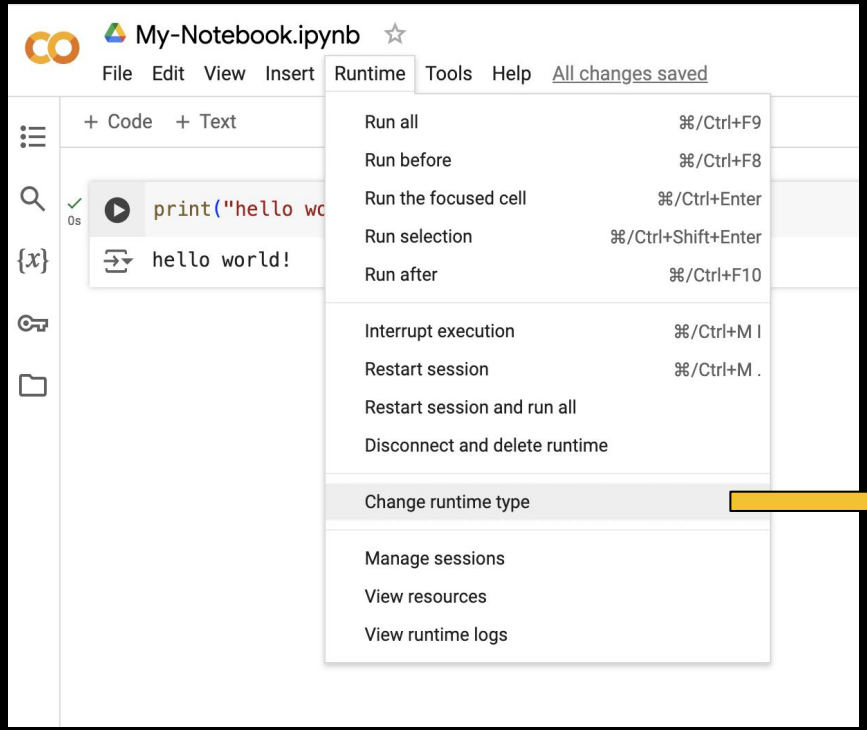


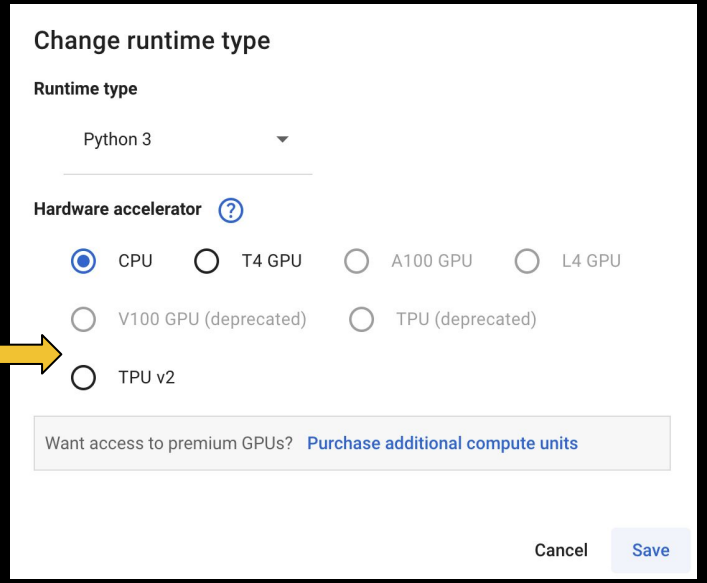
Step 2 - Creating a notebook





Step 3 - Changing the Runtime





Google Collab Link

Link to Code File

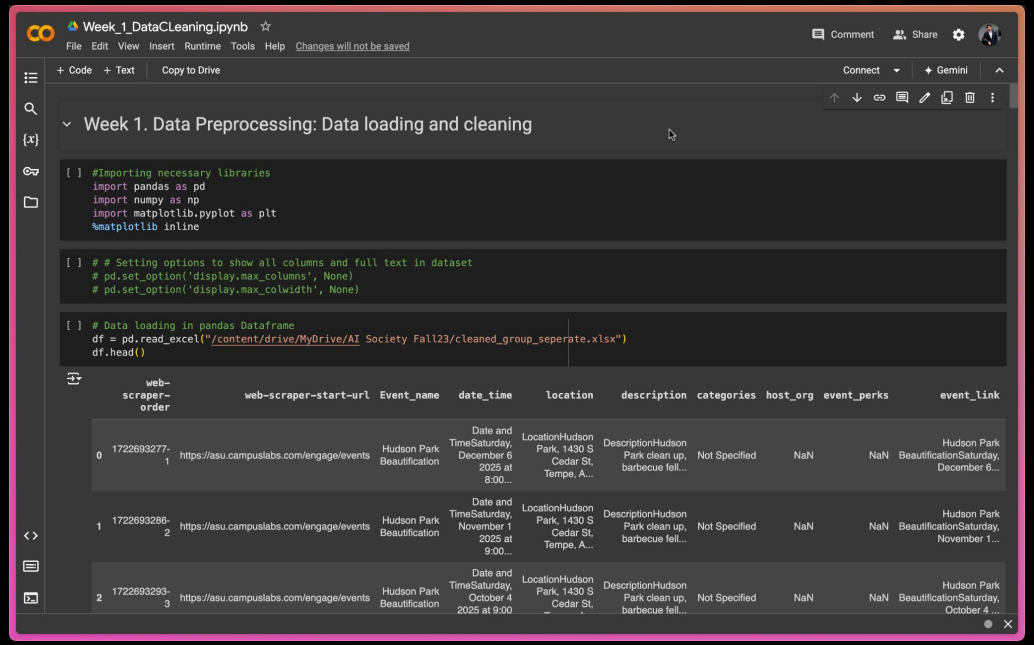
-> https://bit.ly/4ebcjD3

Link to Discord

-> https://bit.ly/AIS-Links

Link to Dataset

-> http://bit.ly/3MAVZjn



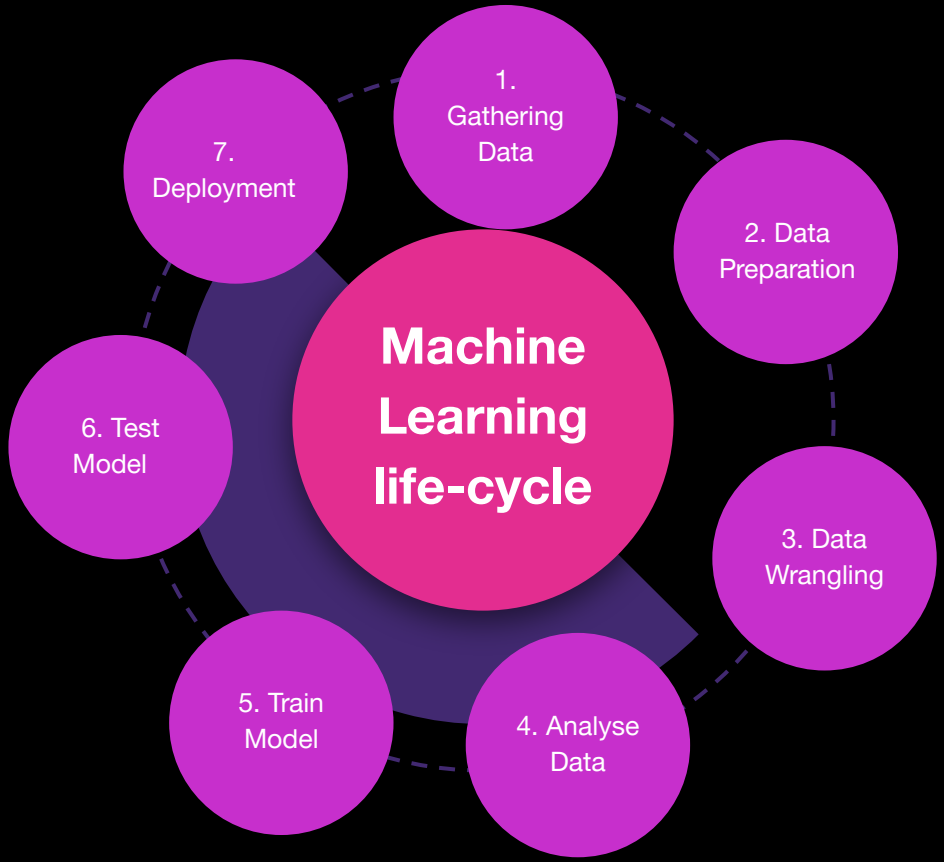
	web-scraper-order	web-scraper-start-url	Event_name	date_time	location	description	categories	host_org	event_perks	event_link
0	1722693277-1	https://asu.campuslabs.com/engage/events	Hudson Park Beautification	Date and TimeSaturday, December 6 2025 at 8:00...	LocationHudson Park, 1430 S Cedar St, Tempe, A...	DescriptionHudson Park clean up, barbecue fell...	Not Specified	NaN	NaN	Hudson Park BeautificationSaturday, December 6...
1	1722693286-2	https://asu.campuslabs.com/engage/events	Hudson Park Beautification	Date and TimeSaturday, November 1 2025 at 9:00...	LocationHudson Park, 1430 S Cedar St, Tempe, A...	DescriptionHudson Park clean up, barbecue fell...	Not Specified	NaN	NaN	Hudson Park BeautificationSaturday, November 1...
2	1722693293-3	https://asu.campuslabs.com/engage/events	Hudson Park Beautification	Date and TimeSaturday, October 4 2025 at 9:00	LocationHudson Park, 1430 S Cedar St,	DescriptionHudson Park clean up, barbecue fell...	Not Specified	NaN	NaN	Hudson Park BeautificationSaturday, October 4 ...

Python and Its Libraries

Why Python??

Python is preferred in data science due to its simplicity and readability, vast library ecosystem, and strong community support.

Core Libraries:


Machine Learning life-cycle
1. Gathering Data
2. Data Preparation
3. Data Wrangling
4. Analyse Data
5. Train Model
6. Test Model
7. Deployment

Understanding Data

What is DATA ?

Data refers to any information that can be used to train models or make predictions. This includes numbers, words, measurements or even images and sounds.

Types of data

- Structured
- Unstructured
- Semi-Structured

How important is Data??

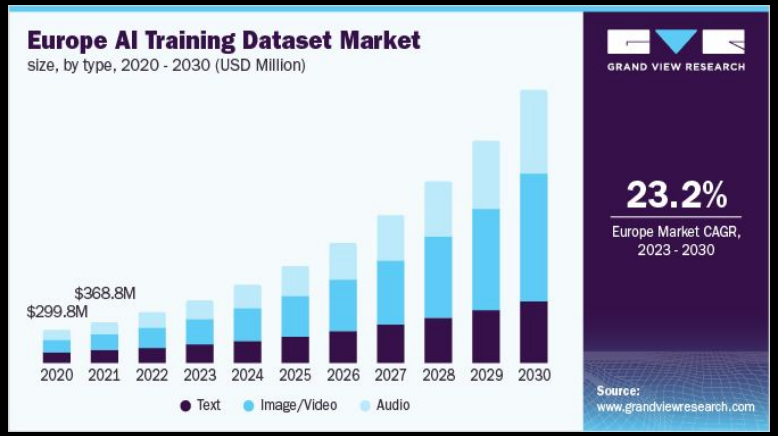
Data: The Lifeblood Of AI

At the core of every AI system lies a fundamental truth: The quality and quantity of data it ingests are paramount to its effectiveness. In essence, data is the lifeblood that fuels AI algorithms, allowing them to learn, adapt and make decisions.

NOVEMBER 1, 2023

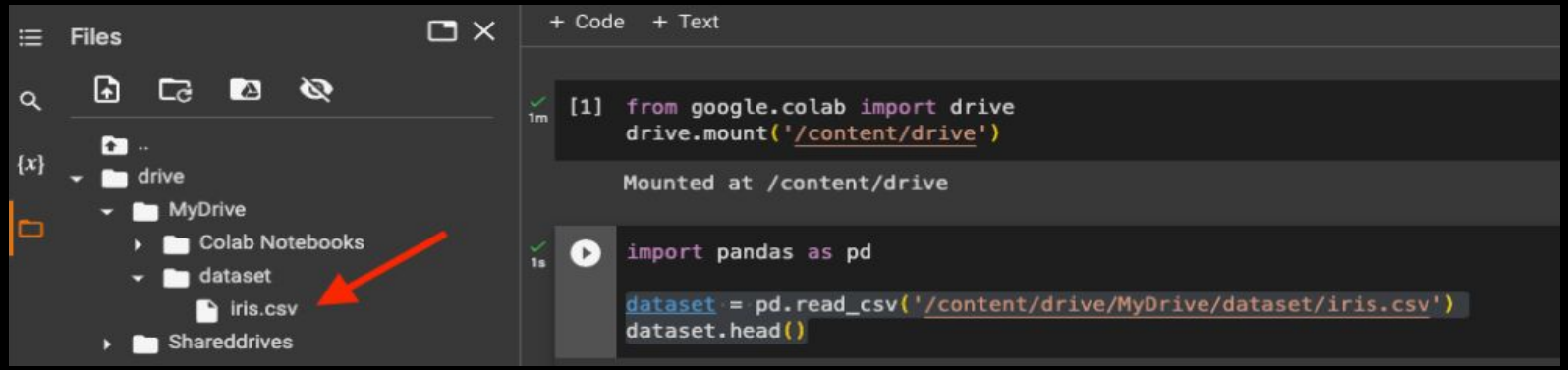
AI is only as good as the data: Q&A with Satish Jayanthi of Coalesce

AI systems obey the golden rule: garbage in, garbage out, Want good results, feed it good data.



File Handling

A CSV file is a Comma Separated Value file that stores tabular data in plain text. Each row represents a record, and each column represents a field in that record. The values in each row are separated by commas or other delimiters, such as semicolons or tabs




```
[ ]  # Let's export the data for later use.
     result.to_csv('cleaned_data.csv', index=False)
```

Data Aggregation

Data aggregation is the process of removing noise from the data by combining and organizing data from multiple sources into a single, unified body.

```
[ ]  # Group by 'Event_name' and 'datetime', then aggregate categories and location_extracted
     result = df.groupby(['Event_name', 'datetime', 'description']).agg({
         'host_org': lambda x: ', '.join(x[x != 'Unknown'].unique()),
         'event_perks': lambda x: ', '.join(x[x != 'None'].unique()),
         'categories': lambda x: ', '.join(x[x != "Not Specified"].unique()),
         'location_extracted': lambda x: ', '.join(x[x != 'Unknown'].unique())  # Exclude 'Unknown' locations
     }).reset_index()
```

Data Pre-processing

It involves transforming raw data into an understandable format. It is a critical step before any machine learning model training, ensuring the data is clean and suitable for analysis.

Feature Engineering

Definition:

Feature engineering is the process of using domain knowledge to select, modify, or create new features from raw data, enhancing the machine learning model's performance.

Importance :

- Better Model Performance
- Reduce Overfitting
- Improved accuracy
- Faster Computation

Additional Resources

Python: https://www.w3schools.com/python/

Python Video: https://www.youtube.com/watch?v=_uQrJ0TkZlc

Libraries: https://www.geeksforgeeks.org/libraries-in-python/

Pandas: https://www.w3schools.com/python/pandas/default.asp

Join our Discord for more updates and resources!

https://www.bit.ly/AIS-Links

Kahoot Time!

Group Picture!