CS167: Machine Learning

Introduction to Machine Learning

Thursday, February 1st, 2024

Alimoor Reza | Assistant Professor | CS Department | Drake University

Recap

- Course introduction and syllabus

- Briefly talked about ML topics

- Other logistics

Today's Agenda

- Topics:
  - **Introduction to Machine Learning**

  - Introduction to Google Colab

  - Python Lab

  - Introduction to Pandas (a library in Python)

An Example: Classifying my dog

- Imagine we want to classify which image depicts a specific dog we want to identify
- Our training data might look something like this:

An Example: Classifying my dog

- Then, when we have some new pictures of my dogs, the idea is that we can make a **prediction** based on previous data as to whether it is **Zoey** or **Ace** in the photo.

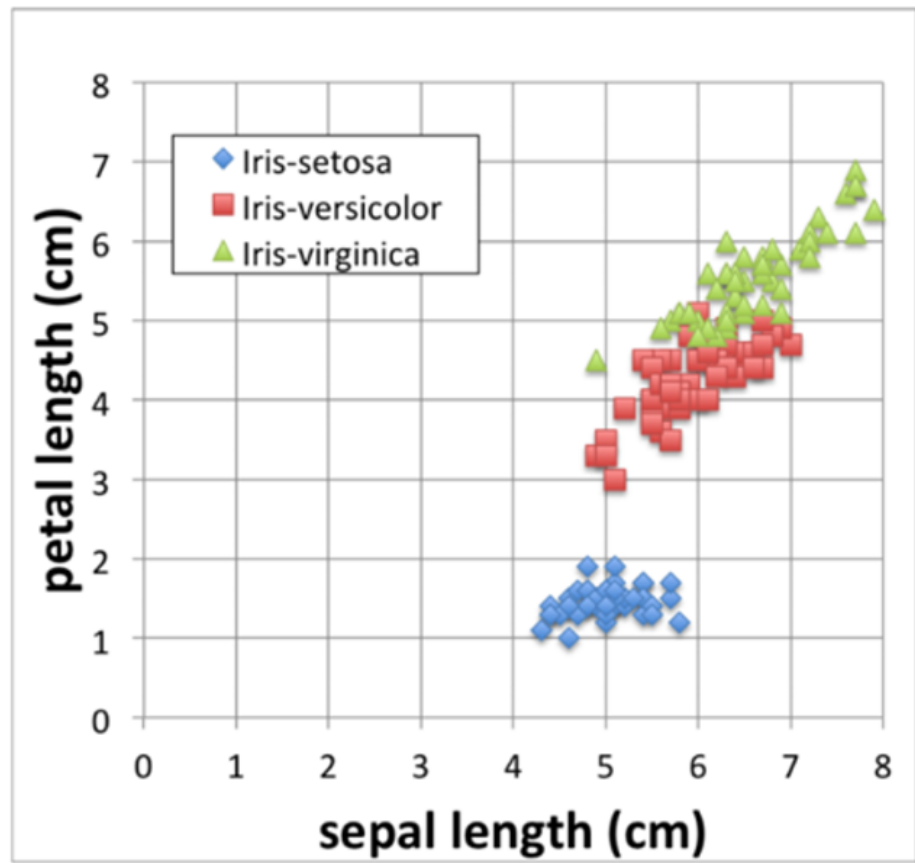# Another Example: What species of Iris is this?

- Use the [Poll link](#) to answer this question

Imagine you found this beautiful flower while on a walk and took the following measurements:

5.1 cm petal length
7.2 cm sepal length

What species do you think it is?

Machine Learning Variations

- We are going to learn about a lot of different types of machine learning in CS167. Here are a few categories to look out for:

  - **classification:** identify which category it goes in, eg, *'Spam or ham?'*, *'Eric or Tim?'*, *'Fish, amphibian, reptile, bird, or mammal'*
  - **regression**: real-valued labels eg, price of Bitcoin, tomorrow's temperature, etc.
  - **supervised learning:** data has labels, goal is to predict the labels of new instance
  - **unsupervised learning:** data does not have a label, the goal is to analyze/cluster the examples
  - **other issues:** missing data, sequential data, outlier anomaly detection, and many more
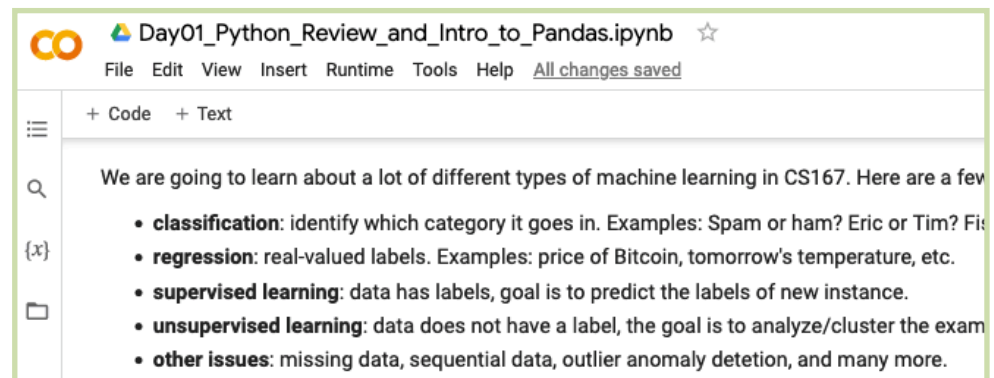
Today's Agenda

- Topics:
  - Introduction to Machine Learning

  - **Introduction to Google Colab**

  - Python Lab

  - Introduction to Pandas (a library in Python)

Introduction to Google Colab

- What is Google colab?

  - Colab is a great browser-based tool for developing iPython Notebooks. Colab allows anybody to write and execute Python code through the browser, and is especially well suited to machine learning, data analysis and education.

  - It makes it really easy to open up someone else's notebook and run it yourself. It also makes it easy to annotate and explain what is happening in the code using Markdown cells.

Google Colab Demonstration

- Switch to Google colab demo …

Today's Agenda

- Topics:
  - Introduction to Machine Learning

  - Introduction to Google Colab

  - Python Lab

  - Introduction to Pandas (a library in Python)

Let's do some Python activities on Colab

- On Blackboard, you should find a file called `Day01_Notes.txt` This is the outline for today's Python Refresher course.

- For the rest of the class period, here's what I'm expecting:
  - Please don't just copy, paste, and run the code...
    - take a second before running it and discuss with your group what you think is going to happen. This really helps us to understand where our underlying assumptions are incorrect.

# Let's do some Python activities on Colab

- For the rest of the class period, here's what I'm expecting:

  - When the outcome is different than you expected, dig in, and try to figure out why. If you're not sure why, raise your hand, and I'll come help explain
    - what does this difference say about how Python works?

  - Answer the questions from the slides in text cells in your notebook.
    - if something was of interest, add a text cell and describe what was of note.

Let's get started!

- Step 1: Create a new notebook

- There are two ways to do this:

  - From Google Drive: https://drive.google.com/
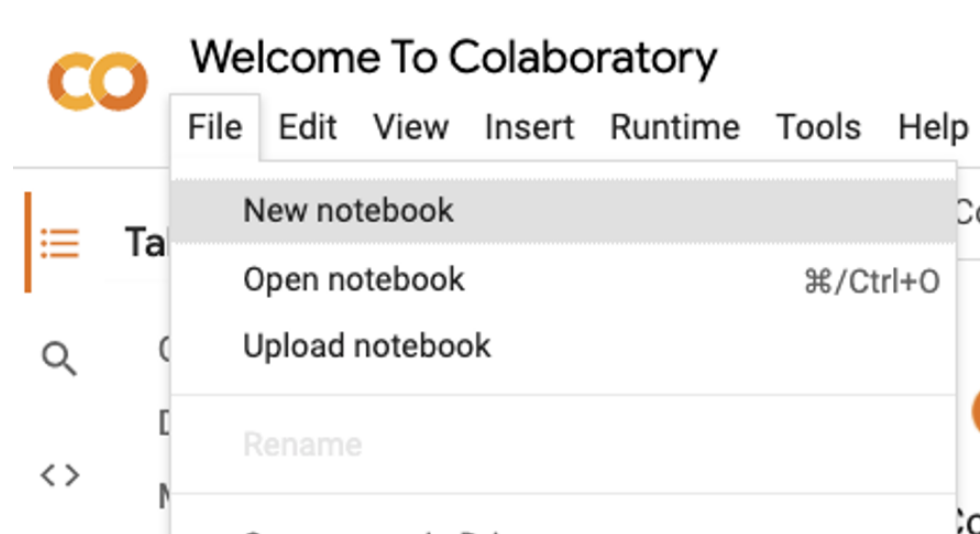
  - From Colab: https://colab.research.google.com/

Step 1: Create a new notebook

- From Google Drive: https://drive.google.com/
- You may need to 'connect more apps', and select Google Colaboratory.

Step 1: Create a new notebook

- From Colab: https://colab.research.google.com/
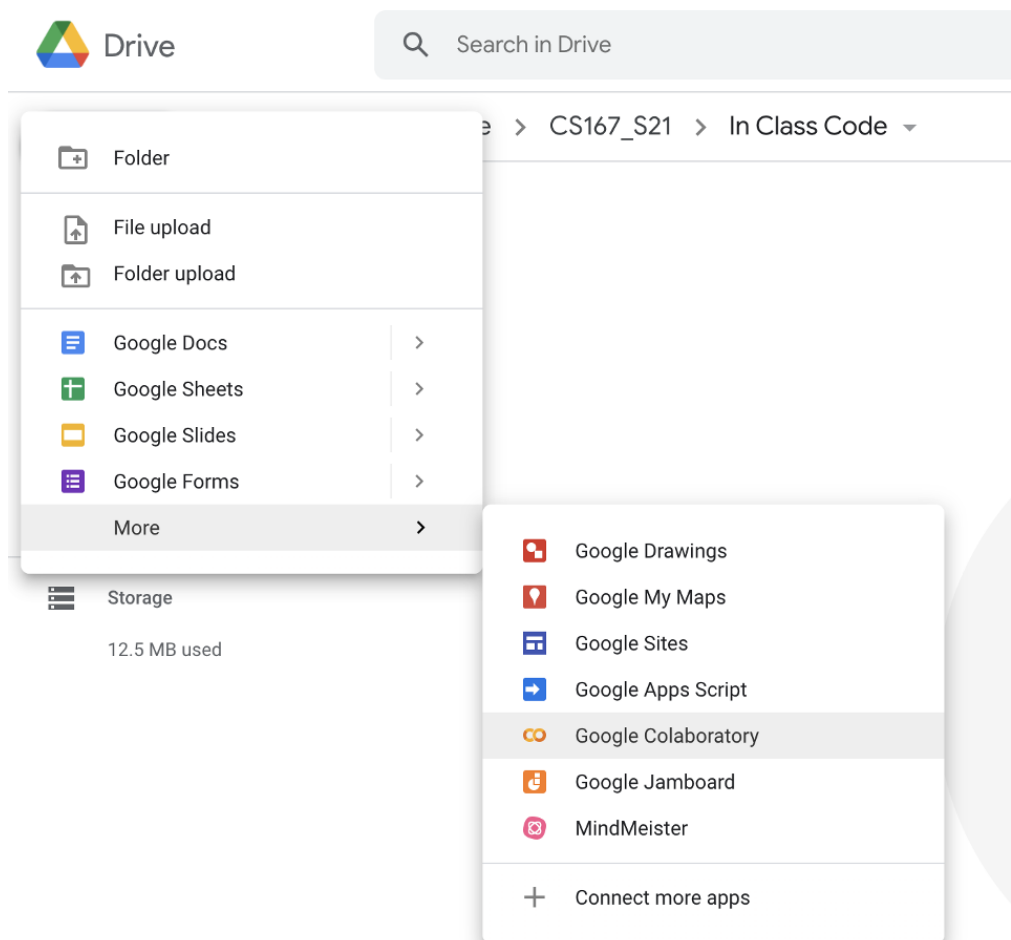
Wait, what am I supposed to be doing?

- Make sure you give your notebook a name (maybe `Day01_notes.ipynb`), and save it to your CS167-Notes Github repository. Your workflow for the rest of class should look something like this:

  - you should have the `Day01_Notes.txt` file open, as well as your Colab Notebook.
  - Copy a section of text from the .txt file and paste it into a new cell in your Colab Notebook.
  - Take a minute and look over the code and predict what will happen. Some cells have specific instructions as to what you should be trying to predict.
  - Run the cell, and see if your prediction was correct.
    - If so, great! Move on.
    - If not, even better--you get to dig into why your expectations were different than how it actually worked, which is a great opportunity to learn something new :)

  - Move on to the next cell and repeat!

Wait, what am I supposed to be doing?

- For the Python review, I'm going to let you all go through it at your own pace, but I will interrupt and go over some of the `pandas` material when I notice people ready to get into it. I'll be walking around to provide moral support, nudges in the (hopefully) correct direction

  - Feel free to flag me down if you get stuck, or don't understand how/why something is working.

Today's Agenda

- Topics:
  - Introduction to Machine Learning

  - Introduction to Google Colab

  - Python Lab

 - **Introduction to Pandas (a library in Python)**

Accessing Data

- Google Colab is a cloud-based tool, which means that we need to store our data in the cloud as well. We cannot simply reference our local data and expect it to work.

- Go ahead and download the `restaurant.csv` file from Blackboard. It is in the Datasets folder.
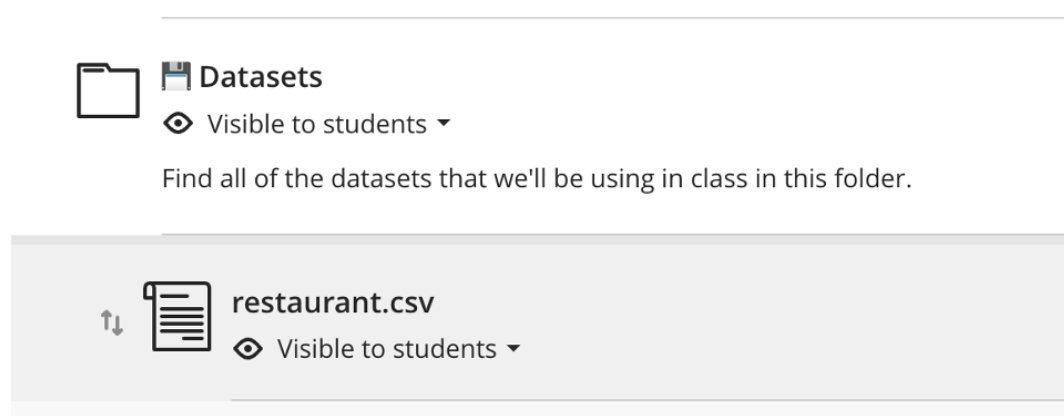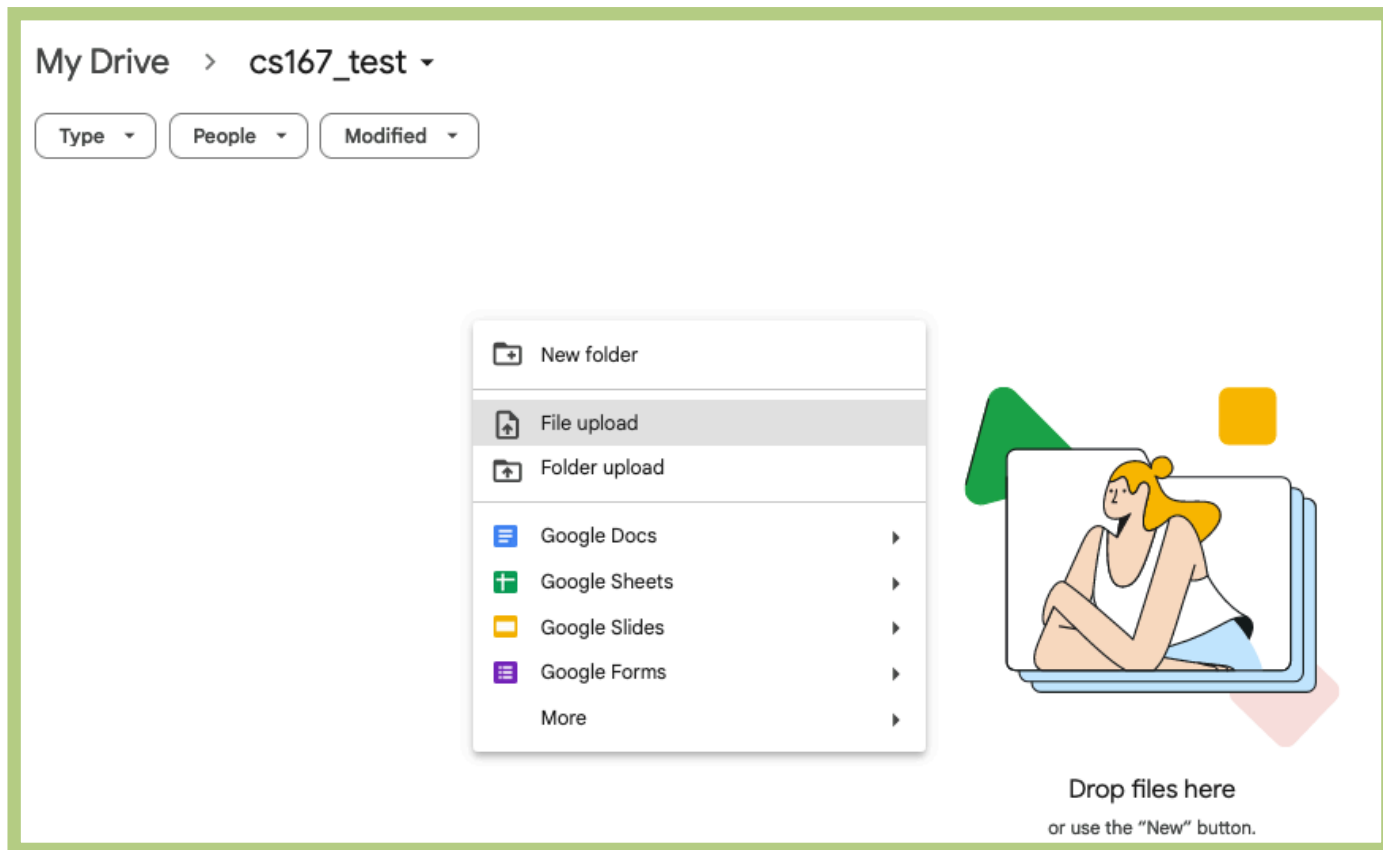
Uploading File to Google Drive

- Upload the `restaurant.csv` to your Google Drive.
  - First go to: drive.google.com
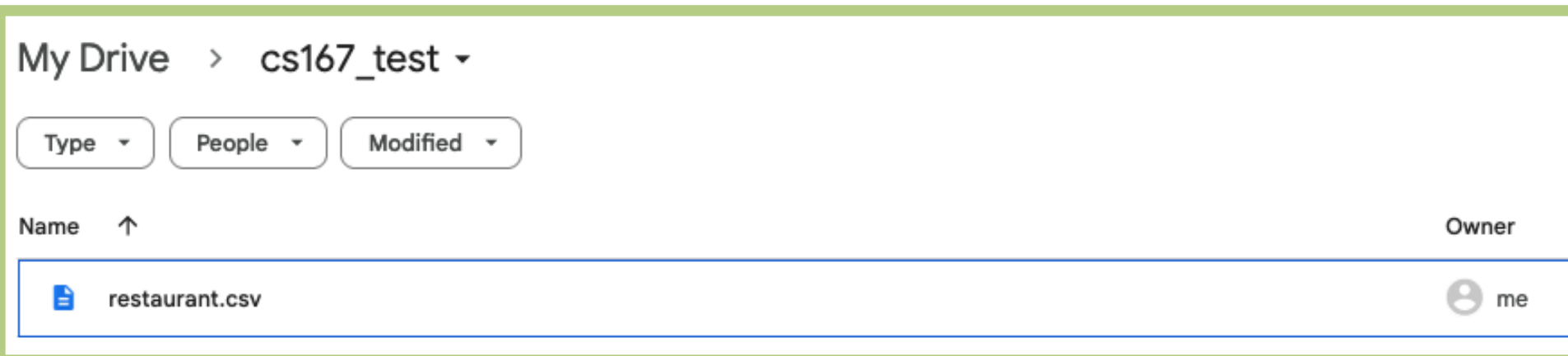 - Then, create a directory/folder (by right-clicking your mouse) as shown below:
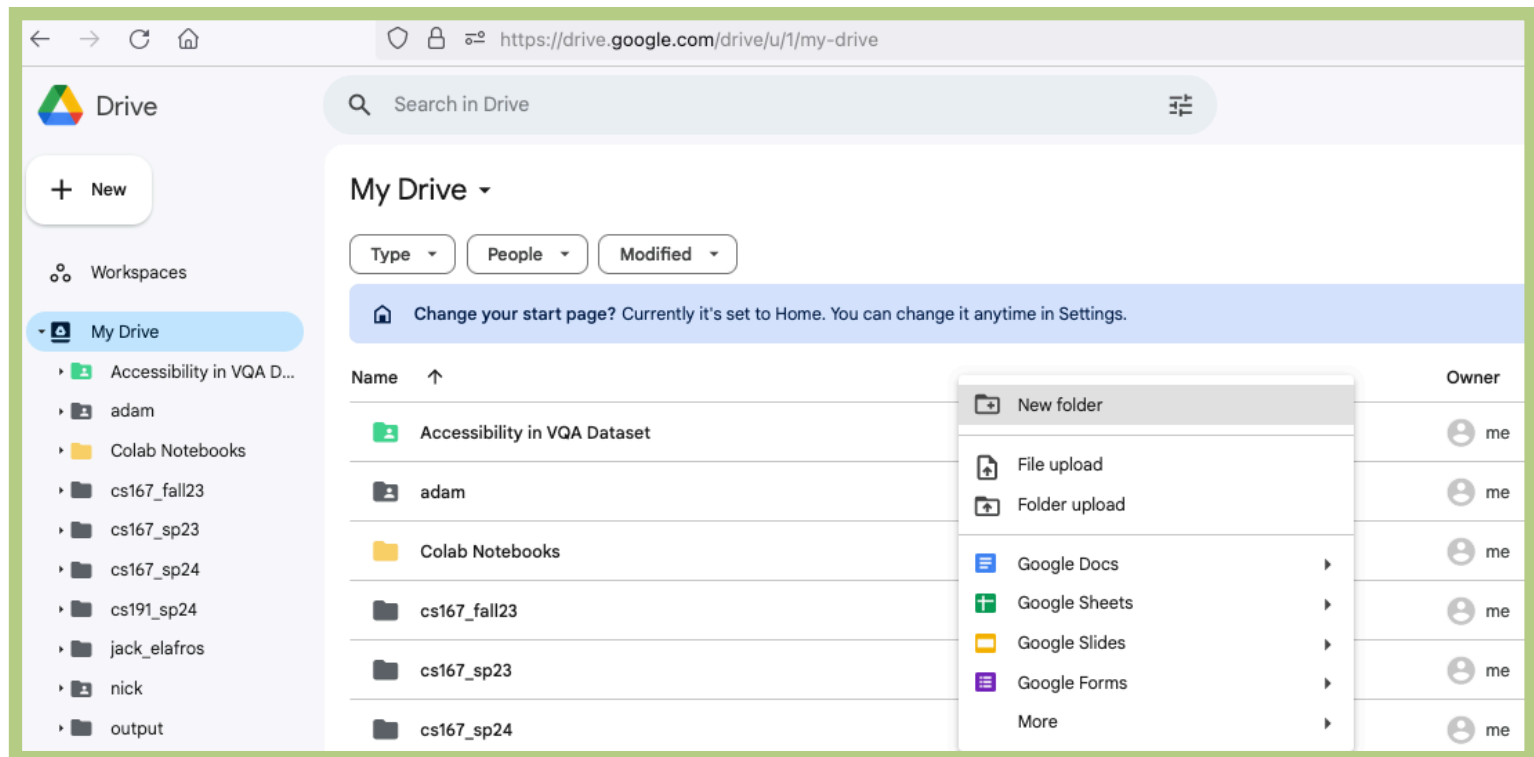
Uploading File to Google Drive

- Upload the `restaurant.csv` to your Google Drive.
  - Go to the directory (eg, 'cs167_test') and then click the 'File upload' option as follows:

Uploading File to Google Drive

- Upload the `restaurant.csv` to your Google Drive.
  - Result of your upload looks as follows:

Accessing Data

- To access this file in Google Colab, you'll need a little bit of code.

```
[ ]  # The first step is to mount your Google Drive to your Colab account.
     #You will be asked to authorize Colab to access your Google Drive. Follow the steps they lead you through.

     #this will only work in Google Colab.

     from google.colab import drive
     drive.mount('/content/drive')
```

- Do a demonstration …

Accessing Data

- **Pandas** is a super powerful Python data analysis library.
  - it's built on top of another super powerful library called `numpy`

- Using Google Colab, `pandas` should already be installed. If you see `In [*]` next to a cell, it means your computer is working on the task

Creating a DataFrame

- In `pandas`, there are two main datatypes, the first of which is a `DataFrame`.

- [Pandas Documentation](#) defines `DataFrames` as:
  - *'Two-dimensional, size-mutable, potentially heterogeneous tabular data'*
  - basically, think of `DataFrames` as our excel sheets--two dimensional, tabular data
  - Each column has a name, and you can use these names to filter and create subsets of data
  - often, you'll see `DataFrames` abbreviated to `df`

DataFrame Demonstration

```
[ ]  #in colab, your path should look something like this:
     path = '/content/drive/MyDrive/CS167/datasets/restaurant.csv' #'dataset/restaurant.csv'
     data = pd.read_csv(path)
     print("data is a ", type(data))
```

Helpful Method Alert: `df.head()`

- The `.head()` method can be called on any DataFrame, and by default will display the first 5 lines rows of the data, as well as the names of the columns.
  - if you want it to display more than 5 rows, you can provide a number as an argument to the method.

- In iPython notebooks, whatever you leave at the end of a cell will automatically output.

- So, when you put those two facts together, you get this nifty functionality:

```
[ ] data.head()
```

	alt	bar	fri	hun	pat	price	rain	res	type	est	target
0	Yes	No	No	Yes	Some	$$$	No	Yes	French	0-10	Yes
1	Yes	No	No	Yes	Full	$	No	No	Thai	30-60	No
2	No	Yes	No	No	Some	$	No	No	Burger	0-10	Yes
3	Yes	No	Yes	Yes	Full	$	No	No	Thai	10-30	Yes
4	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	No

DataFrame Shape

- Want to know the dimensions of your DataFrame? Use `df.shape`

```
data.shape

(12, 11)
```