Data Mining (DTS 360) Presentation

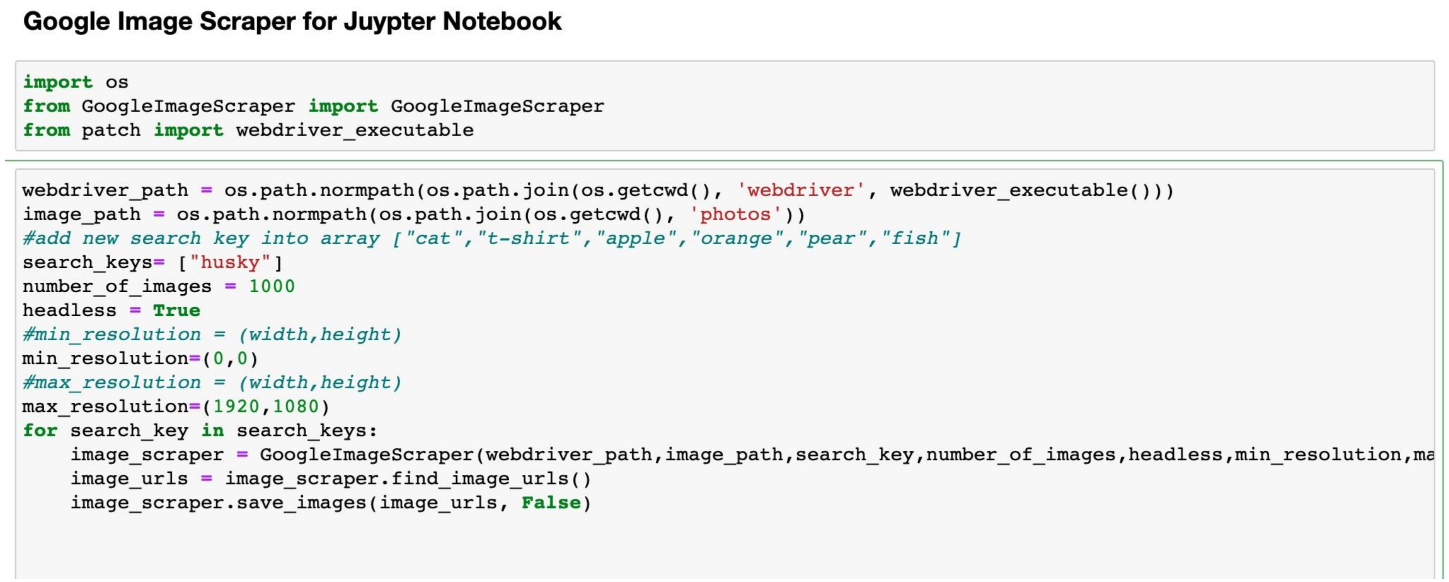
Matthew Chapman

Neural Network

- Computer Vision Neural Network
  - Jupyter Notebook
  - Google Colab
- Dogs
  - Labrador
  - Golden Retreiver
  - Great Pyrenees
  - Dachshund
  - Husky
  - Great Dane
  - German Shepard
  - Boarder Collie
  - Pitbull

Data Collection

- Jupyter Google Image Scraper
- Creates "Husky" Folder
- Different Phrases:
  - Husky
  - Huskies
  - Siberian Husky
  - Husky Alaska



Google Image Scraper for Juypter Notebook

```
import os
from GoogleImageScraper import GoogleImageScraper
from patch import webdriver_executable
```

```
webdriver_path = os.path.normpath(os.path.join(os.getcwd(), 'webdriver', webdriver_executable()))
image_path = os.path.normpath(os.path.join(os.getcwd(), 'photos'))
#add new search key into array ["cat","t-shirt","apple","orange","pear","fish"]
search_keys= ["husky"]
number_of_images = 1000
headless = True
#min_resolution = (width,height)
min_resolution=(0,0)
#max_resolution = (width,height)
max_resolution=(1920,1080)
for search_key in search_keys:
    image_scraper = GoogleImageScraper(webdriver_path,image_path,search_key,number_of_images,headless,min_resolution,ma
    image_urls = image_scraper.find_image_urls()
    image_scraper.save_images(image_urls, False)
```

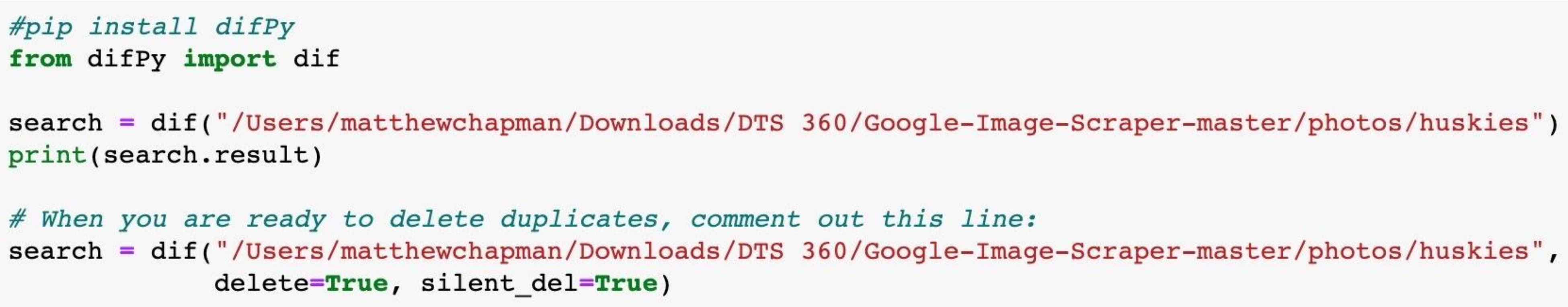
Data Cleaning

husky101.jpg

huskies45.jpg

Data Cleaning

- Delete
  - Duplicates



```
#pip install difPy
from difPy import dif

search = dif("/Users/matthewchapman/Downloads/DTS 360/Google-Image-Scraper-master/photos/huskies")
print(search.result)

# When you are ready to delete duplicates, comment out this line:
search = dif("/Users/matthewchapman/Downloads/DTS 360/Google-Image-Scraper-master/photos/huskies",
             delete=True, silent_del=True)
```

  - Low Quality Pictures
  - Pictures that are not Huskies

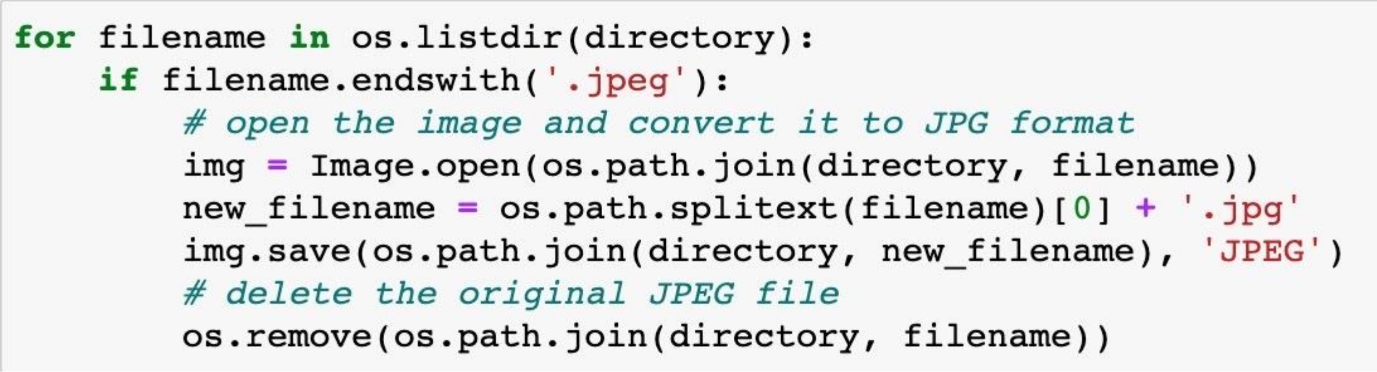
Data Labeling

- Yolo Label
- Convert to JPG or PNG
- obj.names
- Create txt file for each photo

  husky5.jpg -> husky5.txt:

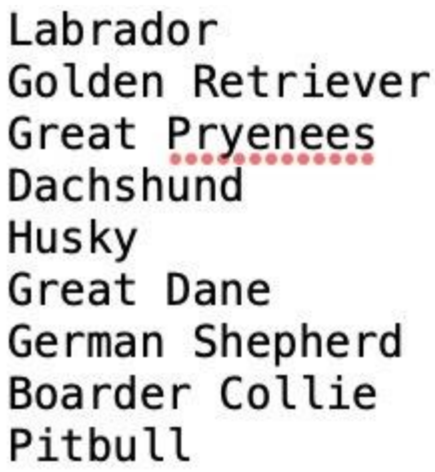
  "4 0.541806 0.488347 0.695652 0.786017"

Convert to JPG:

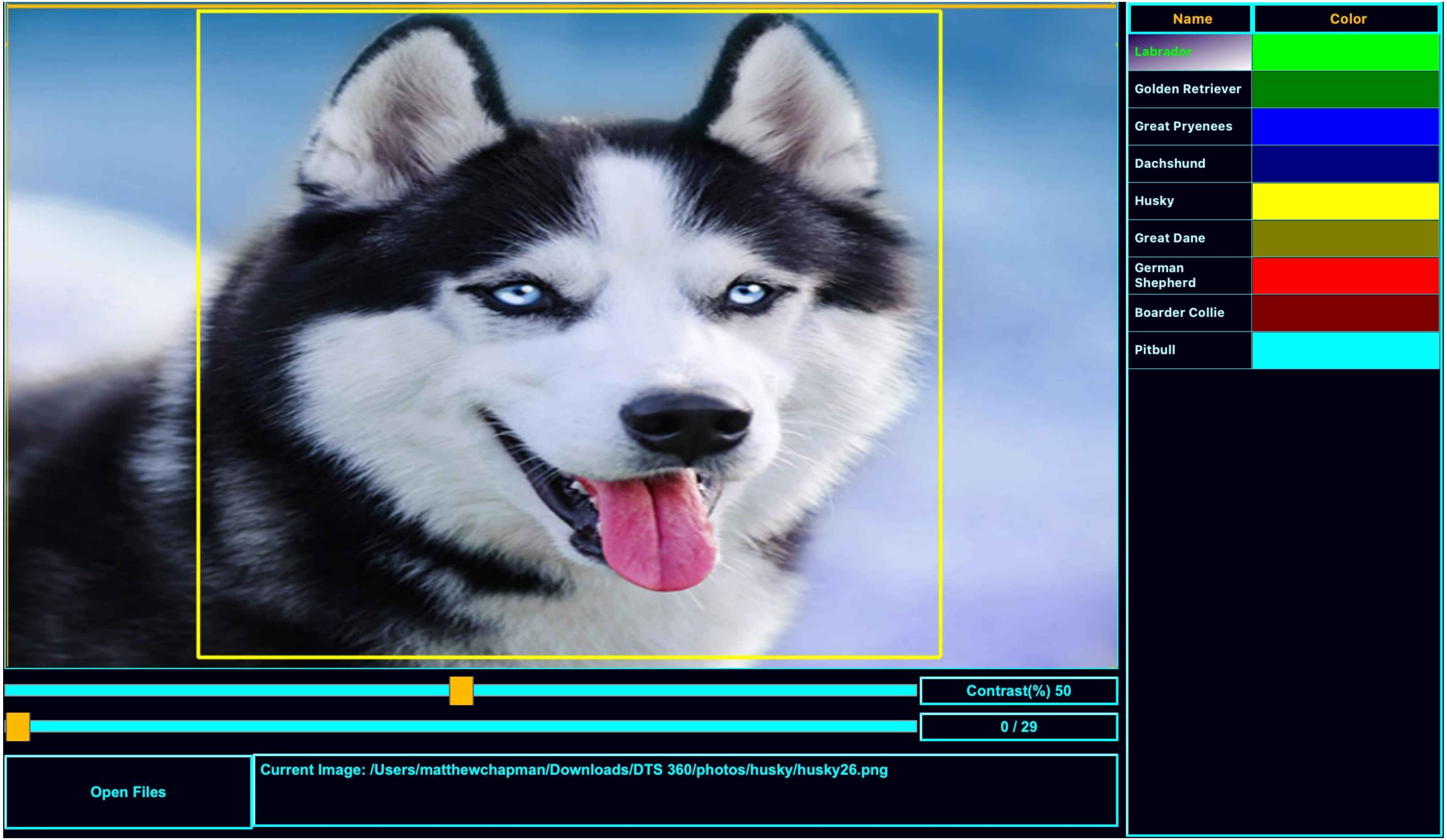


```
for filename in os.listdir(directory):
    if filename.endswith('.jpeg'):
        # open the image and convert it to JPG format
        img = Image.open(os.path.join(directory, filename))
        new_filename = os.path.splitext(filename)[0] + '.jpg'
        img.save(os.path.join(directory, new_filename), 'JPEG')
        # delete the original JPEG file
        os.remove(os.path.join(directory, filename))
```

obj.names:



```
Labrador
Golden Retriever
Great Pryenees
Dachshund
Husky
Great Dane
German Shepherd
Boarder Collie
Pitbull
```



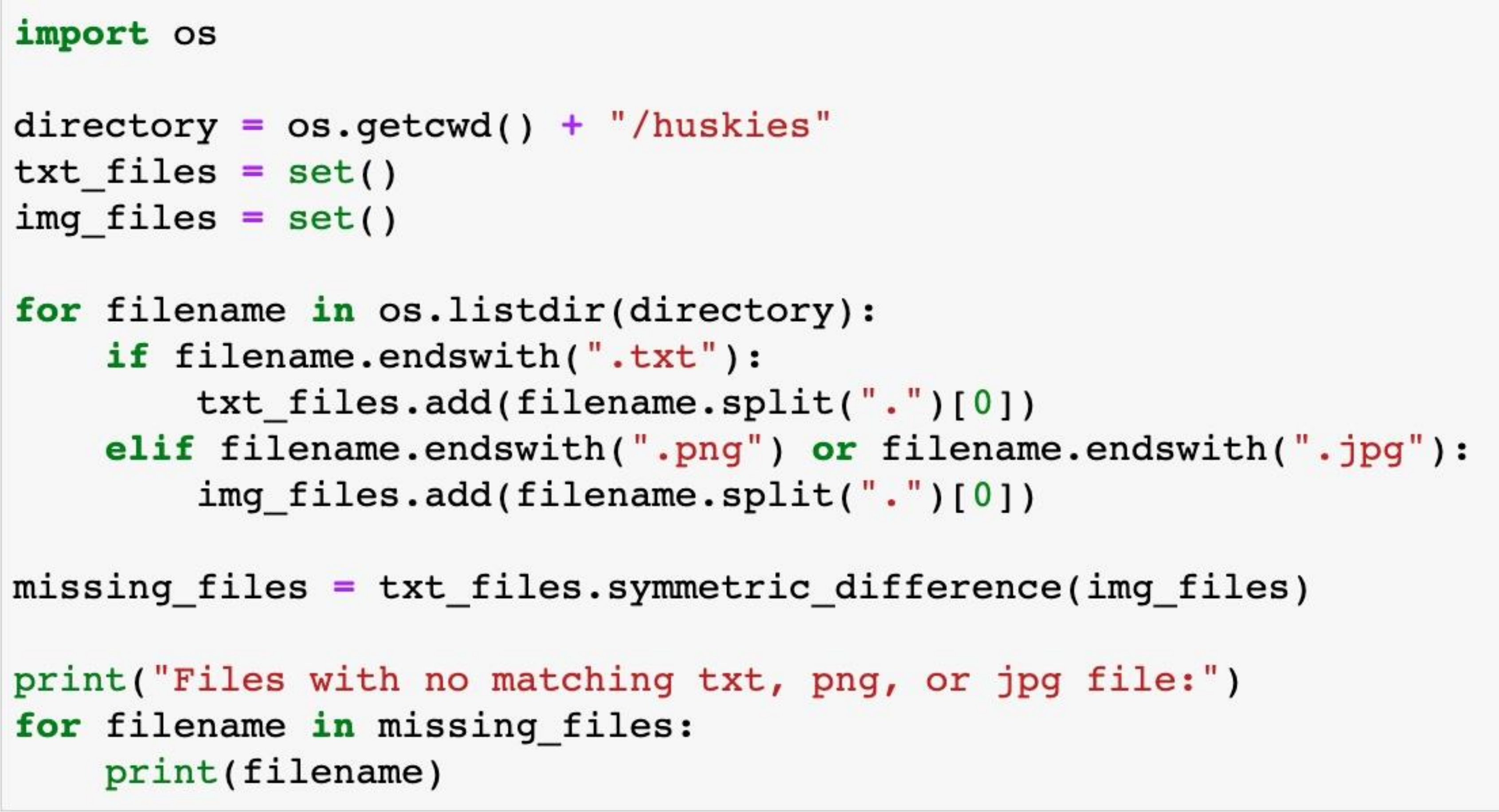
Name	Color
Labrador	🟩
Golden Retriever	🟩
Great Pryenees	🟦
Dachshund	🟦
Husky	🟨
Great Dane	🟫
German Shepherd	🟥
Boarder Collie	🟥
Pitbull	🟦

Contrast(%) 50

0 / 29

Current Image: /Users/matthewchapman/Downloads/DTS 360/photos/husky/husky26.png

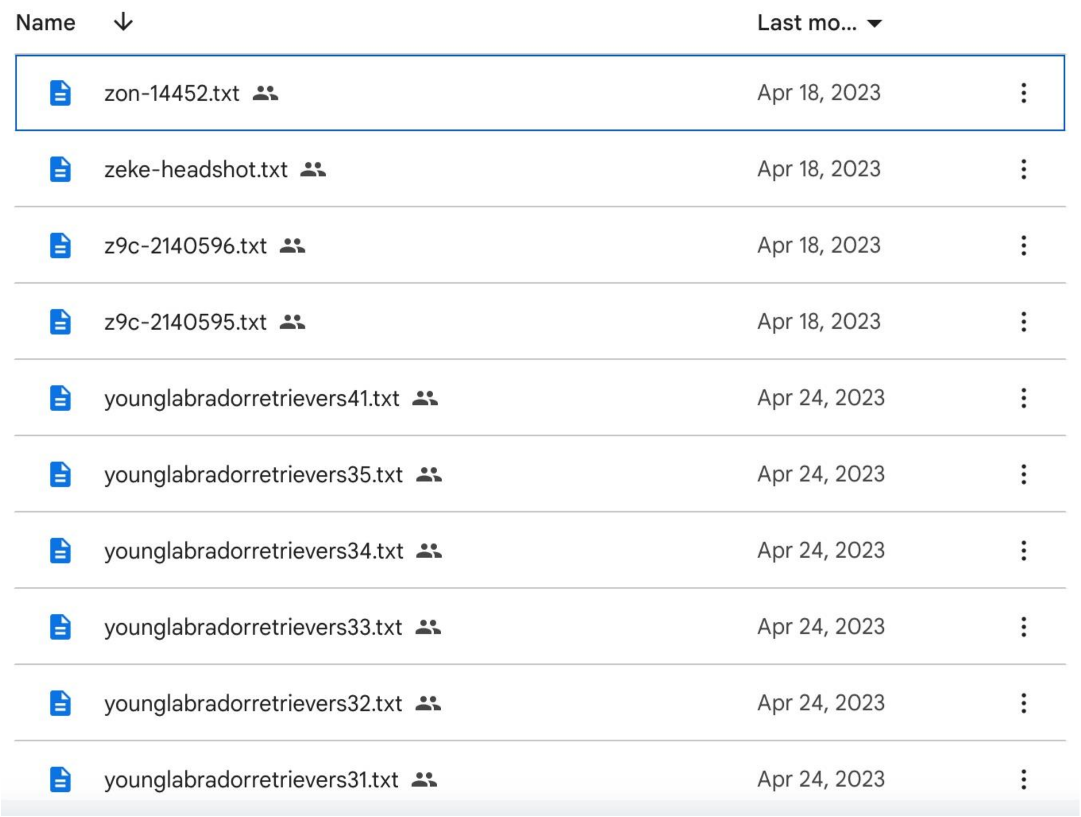
Delete Files With No Match



```python
import os

directory = os.getcwd() + "/huskies"
txt_files = set()
img_files = set()

for filename in os.listdir(directory):
    if filename.endswith(".txt"):
        txt_files.add(filename.split(".")[0])
    elif filename.endswith(".png") or filename.endswith(".jpg"):
        img_files.add(filename.split(".")[0])

missing_files = txt_files.symmetric_difference(img_files)

print("Files with no matching txt, png, or jpg file:")
for filename in missing_files:
    print(filename)
```

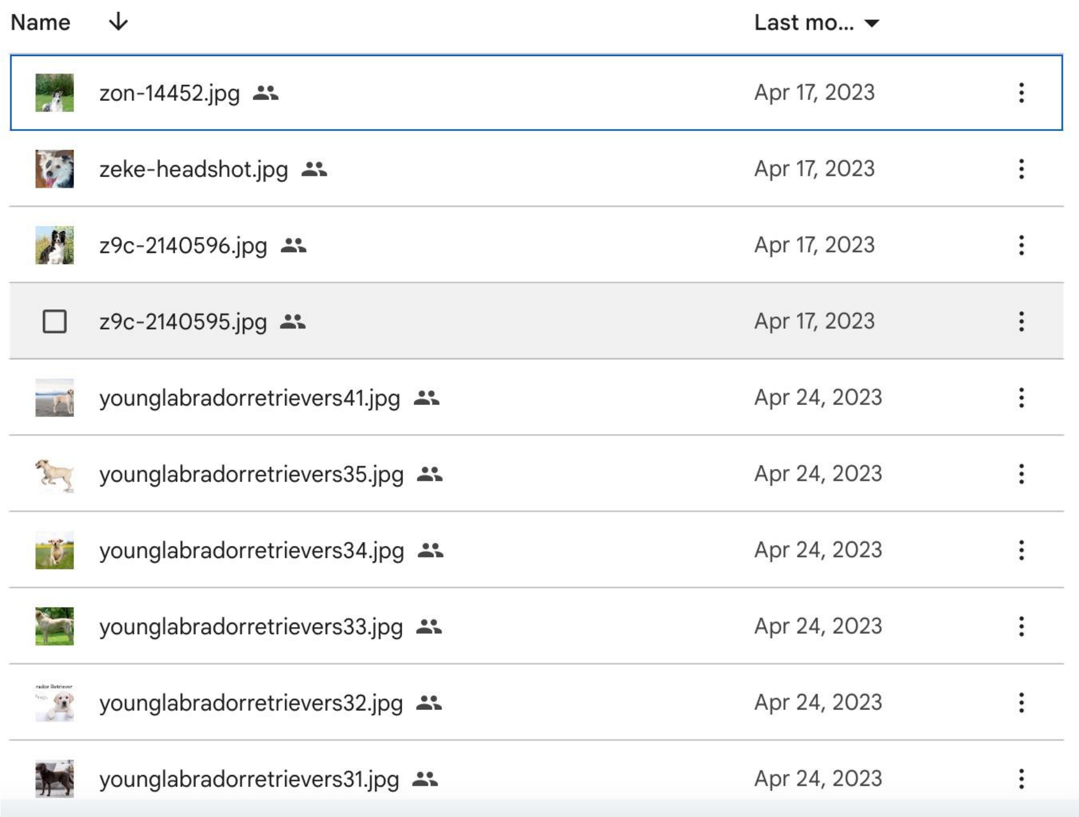
Data Merging (Google Drive)

Labels



Name ↓	Last mo... ▾
zon-14452.txt	Apr 18, 2023
zeke-headshot.txt	Apr 18, 2023
z9c-2140596.txt	Apr 18, 2023
z9c-2140595.txt	Apr 18, 2023
younglabradorretrievers41.txt	Apr 24, 2023
younglabradorretrievers35.txt	Apr 24, 2023
younglabradorretrievers34.txt	Apr 24, 2023
younglabradorretrievers33.txt	Apr 24, 2023
younglabradorretrievers32.txt	Apr 24, 2023
younglabradorretrievers31.txt	Apr 24, 2023

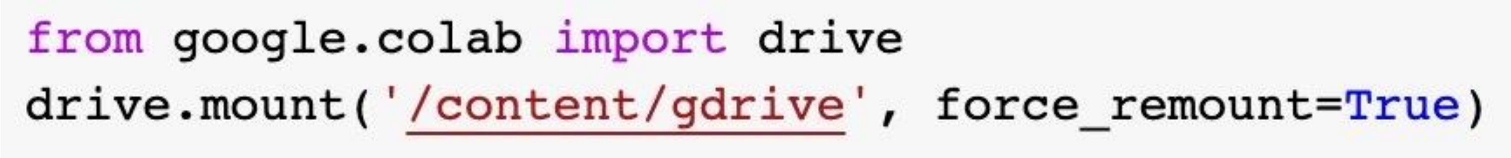
Images



Name ↓	Last mo... ▾
zon-14452.jpg	Apr 17, 2023
zeke-headshot.jpg	Apr 17, 2023
z9c-2140596.jpg	Apr 17, 2023
□ z9c-2140595.jpg	Apr 17, 2023
younglabradorretrievers41.jpg	Apr 24, 2023
younglabradorretrievers35.jpg	Apr 24, 2023
younglabradorretrievers34.jpg	Apr 24, 2023
younglabradorretrievers33.jpg	Apr 24, 2023
younglabradorretrievers32.jpg	Apr 24, 2023
younglabradorretrievers31.jpg	Apr 24, 2023

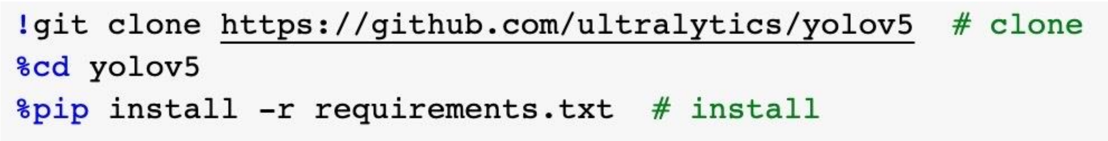
Neural Network Training (Google Colab)

1: Give Google Colab access to drive



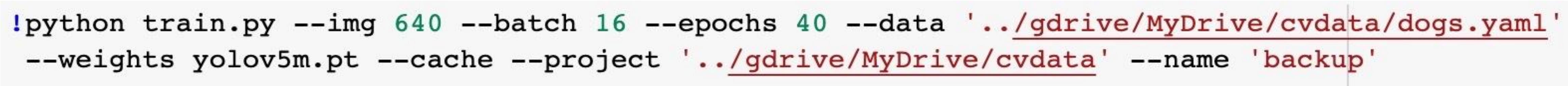
```
from google.colab import drive
drive.mount('/content/gdrive', force_remount=True)
```

2: Install



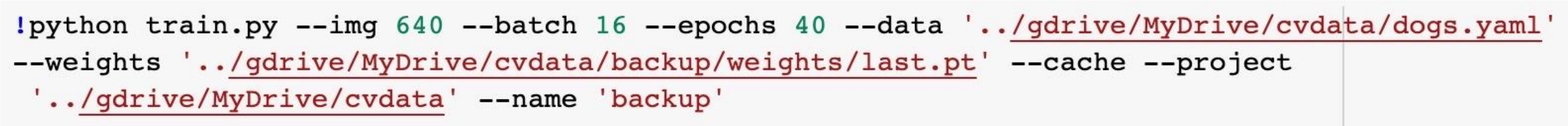
```
!git clone https://github.com/ultralytics/yolov5  # clone
%cd yolov5
%pip install -r requirements.txt  # install
```

3: Train Model



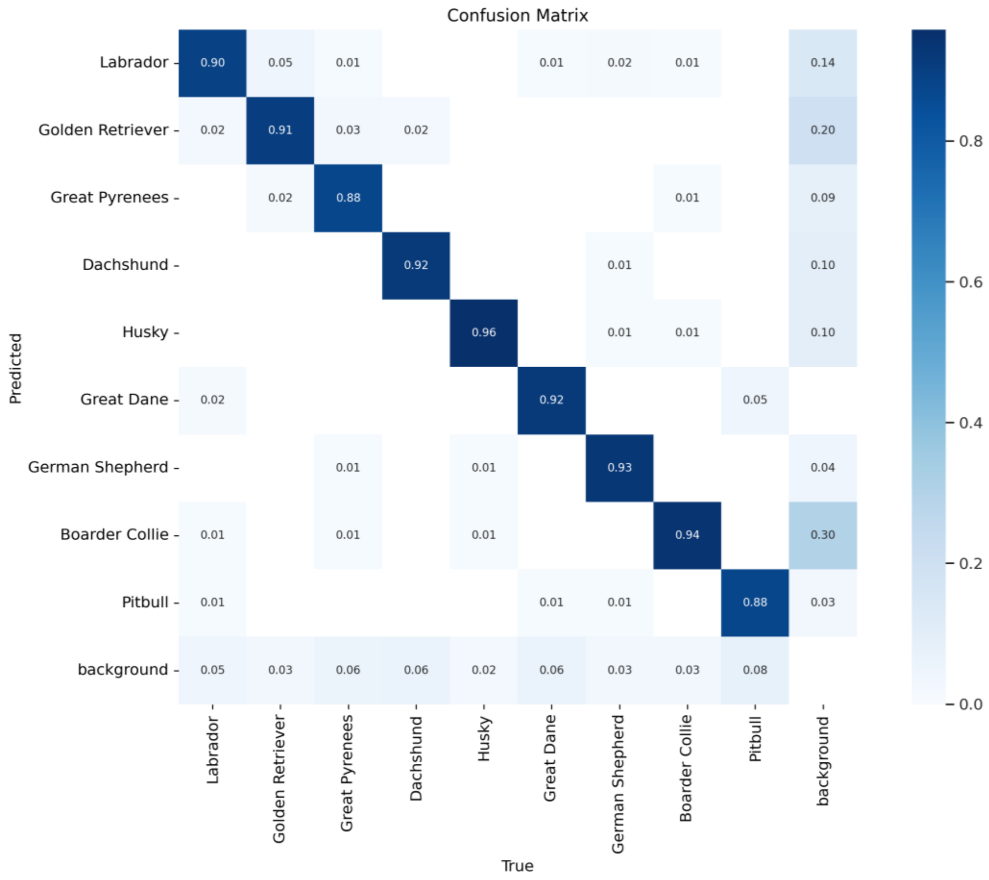
```
!python train.py --img 640 --batch 16 --epochs 40 --data '../gdrive/MyDrive/cvdata/dogs.yaml'
 --weights yolov5m.pt --cache --project '../gdrive/MyDrive/cvdata' --name 'backup'
```

4: Train with new weights



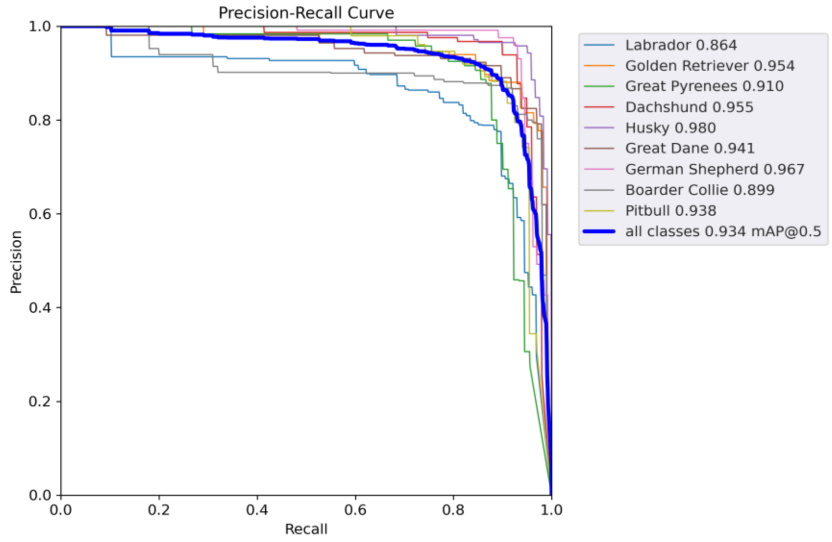
```
!python train.py --img 640 --batch 16 --epochs 40 --data '../gdrive/MyDrive/cvdata/dogs.yaml'
--weights '../gdrive/MyDrive/cvdata/backup/weights/last.pt' --cache --project
 '../gdrive/MyDrive/cvdata' --name 'backup'
```

Model Evaluation

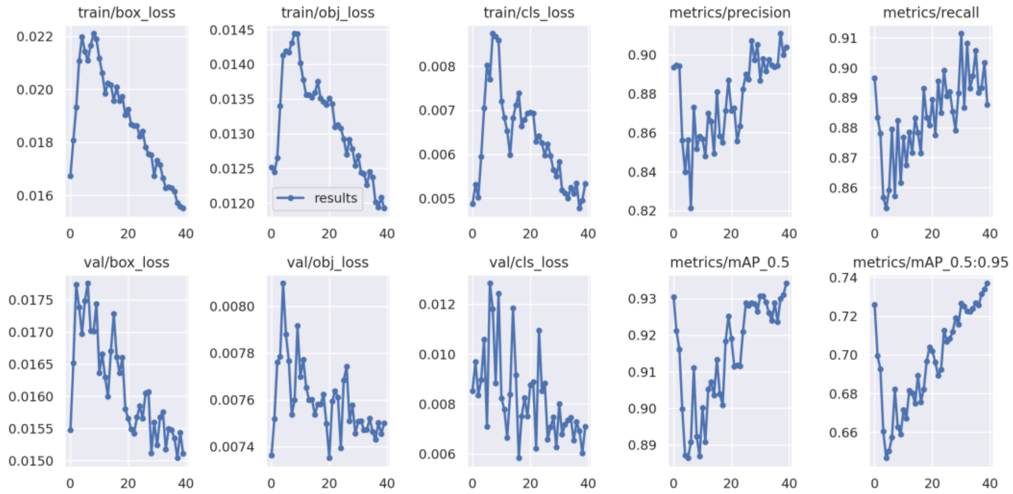


Confusion Matrix

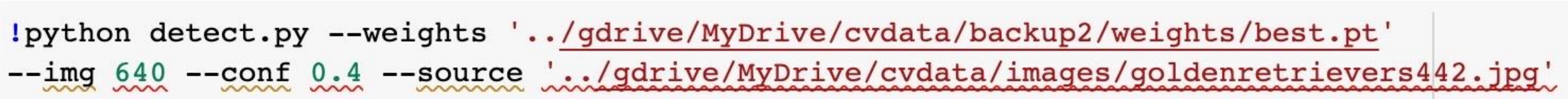
	Labrador	Golden Retriever	Great Pyrenees	Dachshund	Husky	Great Dane	German Shepherd	Boarder Collie	Pitbull	background
Labrador	0.90	0.05	0.01			0.01	0.02	0.01		0.14
Golden Retriever	0.02	0.91	0.03	0.02						0.20
Great Pyrenees		0.02	0.88					0.01		0.09
Dachshund				0.92			0.01			0.10
Husky					0.96		0.01	0.01		0.10
Great Dane	0.02					0.92			0.05	
German Shepherd			0.01		0.01		0.93			0.04
Boarder Collie	0.01		0.01		0.01			0.94		0.30
Pitbull	0.01					0.01	0.01		0.88	0.03
background	0.05	0.03	0.06	0.06	0.02	0.06	0.03	0.03	0.08	



Precision-Recall Curve

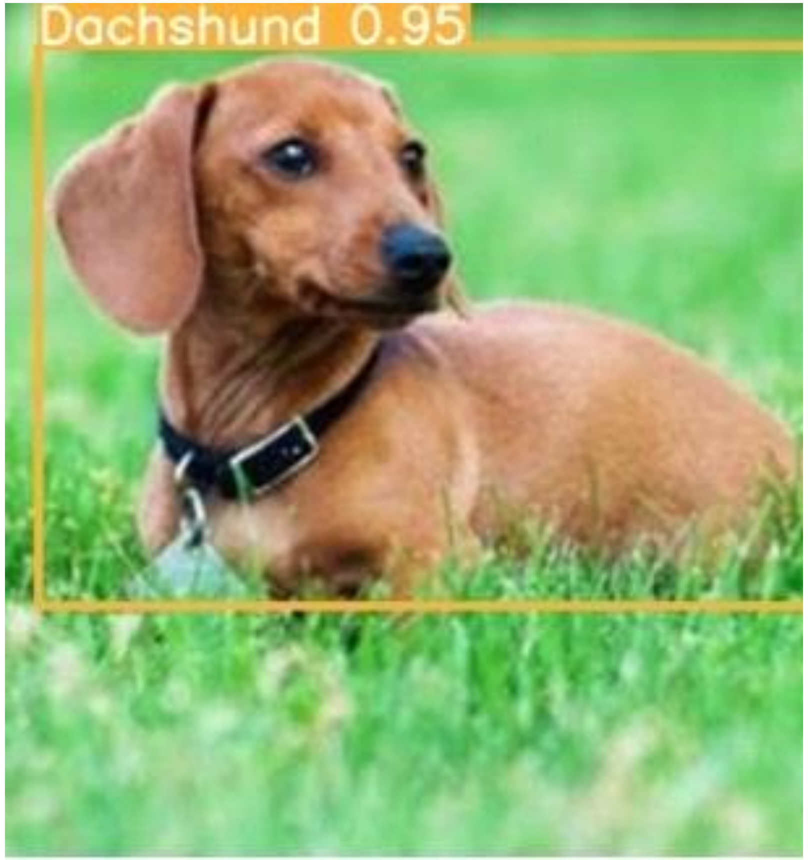


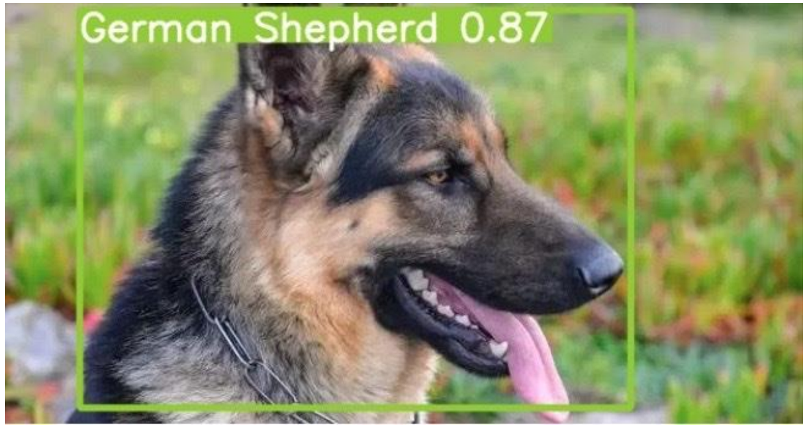
Using the Model

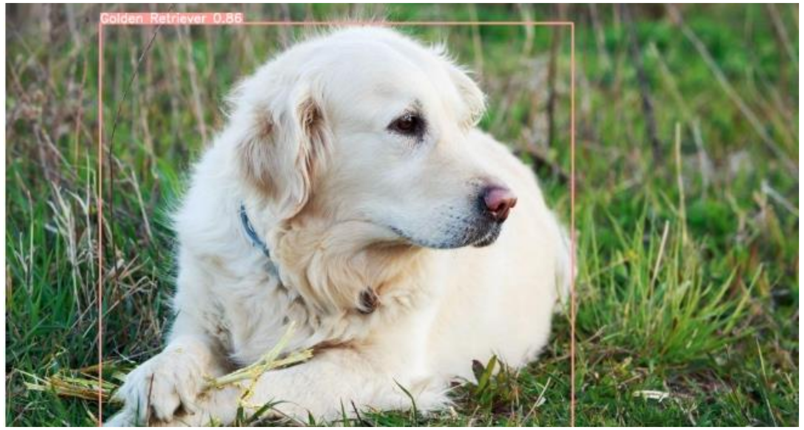


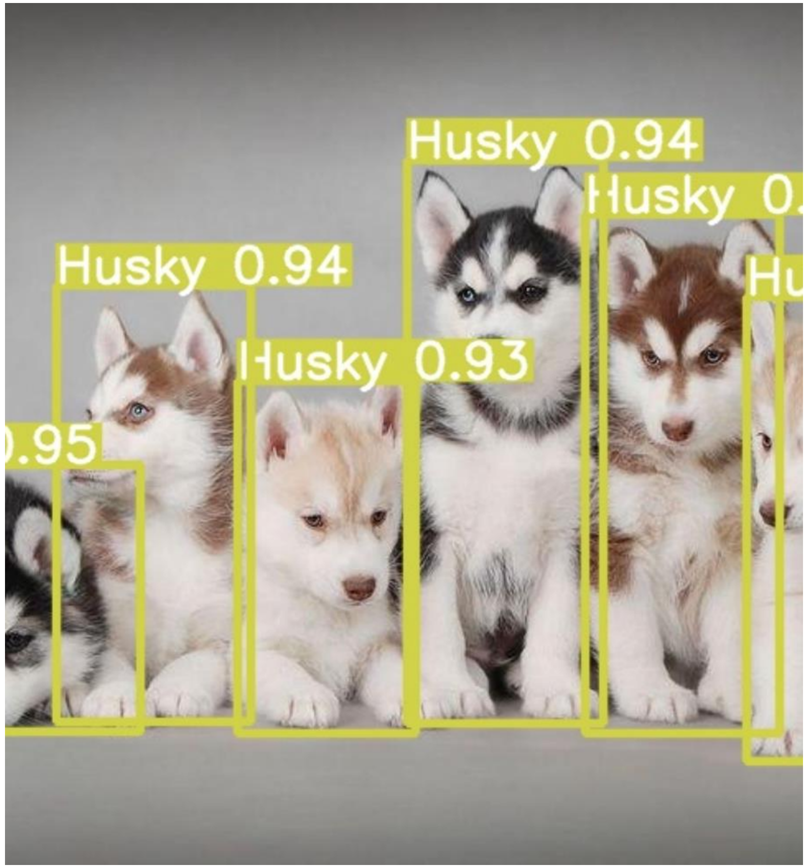
```
!python detect.py --weights '../gdrive/MyDrive/cvdata/backup2/weights/best.pt'
--img 640 --conf 0.4 --source '../gdrive/MyDrive/cvdata/images/goldenretrievers442.jpg'
```

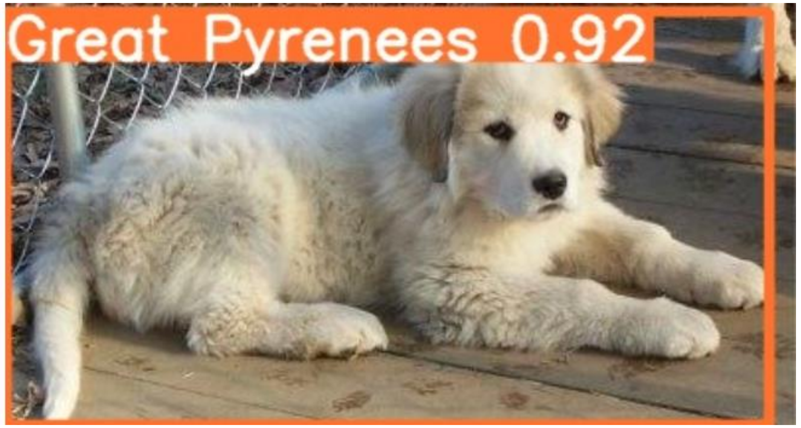
Using the Model











Thoughts

- Most Challenging:
  - Collecting, Cleaning, and Preparing Data (Didn't always work)
  - Google Colab
  - Understanding the model
- Longest:
  - Labeling
- Most Interesting:
  - Training Model
  - Using Model