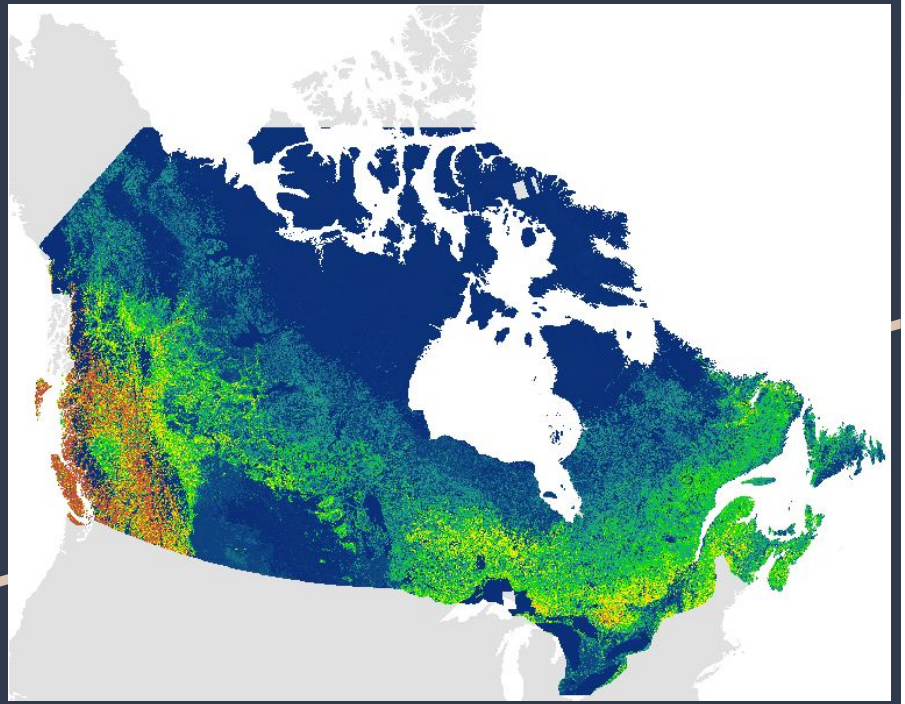
Big Data Project

Using ML to Mitigate Forest Fires in Canada

Andrew Greer & Troy Walther

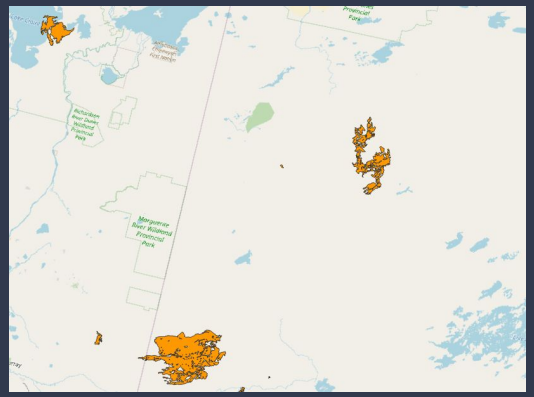
Zoom Recording Link
Passcode: *9%aEmIs

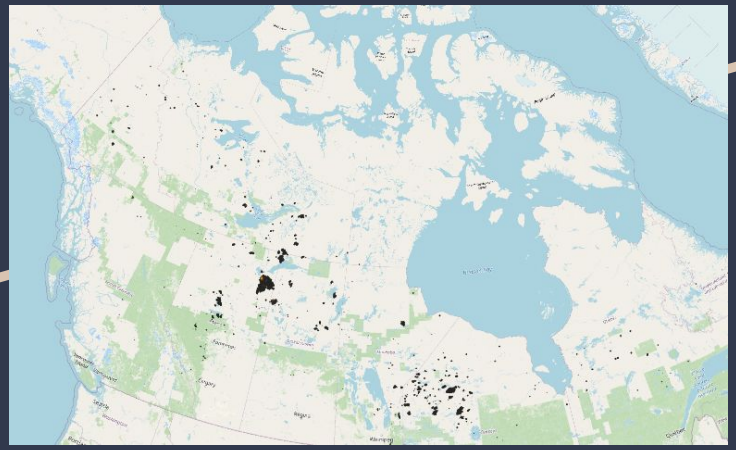
Dataset



- this dataset represents all tree genus-species percentages per 250 hectares of land across all of Canada.
- We generalised this dataset to show the majority tree type for each 250 hectare point
- Each point contains coordinates which allow us to link them with weather data and fires

Dataset cont...





The [Canadian Wildfire Area Burned Database] contains the Areas burned in wildfires for each year from 1986 onward. Each year can be downloaded individually, we are using 2011 (The same year as our tree database)

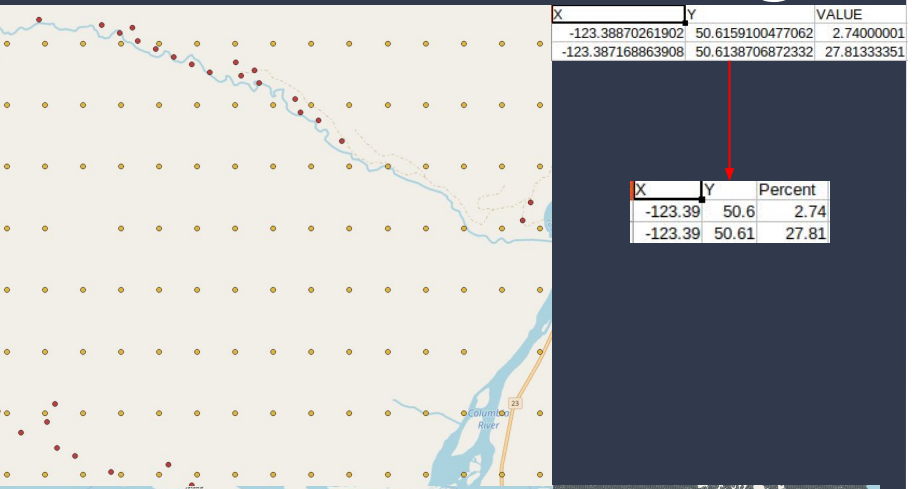
combined with weather and tree data using coordinates, fire points were grouped using their FIREID.

Dataset Cont...

The Weather Station Data warehouse contains the readings from every Weather station in Canada separated by year and month, and tracks the average temperature, and precipitation at a given latitude and longitude.

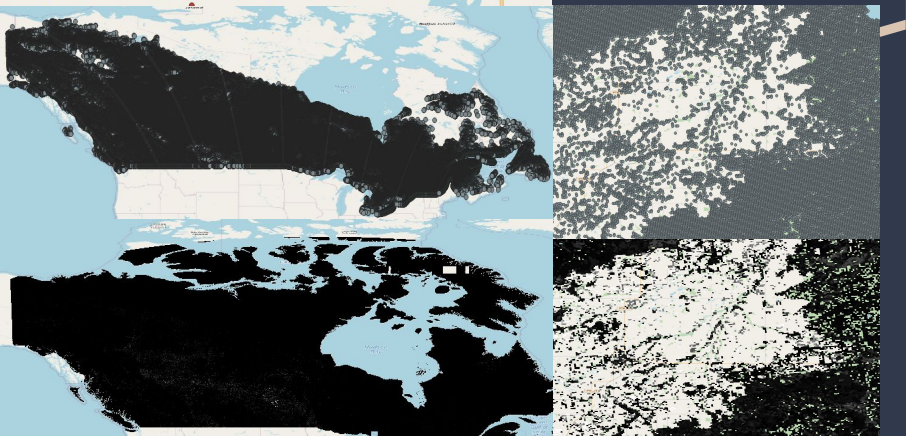
Grouped with FIREID's and coordinates in order to give algorithm an understanding of weather patterns which cause fires.

Data Preprocessing



X	Y	VALUE
-123.38870261902	50.6159100477062	2.74000001
-123.387168863908	50.6138706872332	27.81333351

X	Y	Percent
-123.39	50.6	2.74
-123.39	50.61	27.81



Tree Data

-Tree Data was converted from GeoTiff raster format to csv and simplified reducing number of trees.

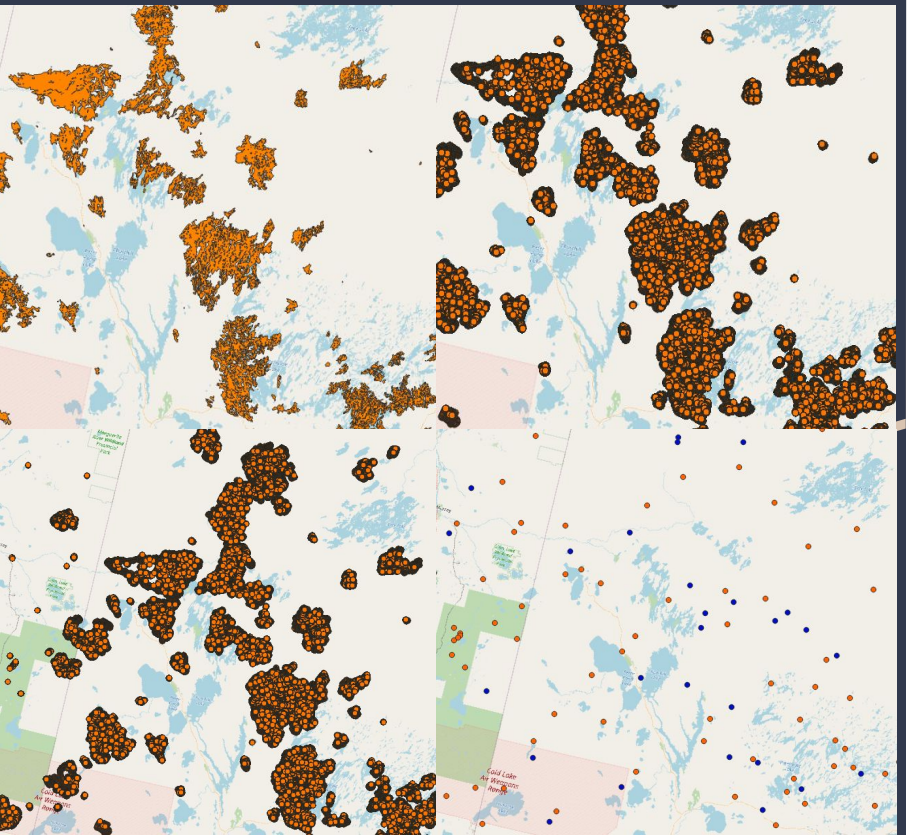
-Coordinates were rounded to 2 decimal places from 12.

-Species were merged into one Genus, IE Pice_Mar Pice_Neg, Pice_SPP all get averaged and merged into Pice.csv.

-processing time with multiprocessing was 84 hours. The total compute hours based on a 6 core CPU (12 logical processors) is 505.17 core hours, or 21 core days.

-Final features are X, Y, Percent

Data Preprocessing



Fire Data

-Fire Data was converted from shape file to CSV by extracting Vector points

-Fire were grouped by their fireID

-Grouped fires without valid dates (requires one of SDATE, AFSDATE, or EDATE to match weather) are dropped

-Negative values are added using random coordinates which are excluded from the fires.

-Fire data contains X, Y, and BURNCLAS as features.

Data Preprocessing

Weather Data

-Weather data came in CSV files in the form of 12 monthly summary files

- Any rows which lack a value for Tm or P (Temperature or precipitation) will have that value imputed using the nearest neighbour based on min (distance X + distance Y + (date_difference * 1000)) the date difference * 1000 ensures that the closest measurement is in the same month

-Once imputation is done weather data is exported

-Weather data contains Long, Lat, P, and Tm

Data Preprocessing

X	Y	Tm	P	tree_genus	BURNCLAS
-140.259159520593	66.8571674805877	12.9	31.6	Pice	4
-136.612391393738	67.540101380588	7.3	16	Pice	4
-137.052146471854	67.278038161089	7.3	16	Pice	4
-139.334882642017	66.410795582083	12.9	31.6	Pice	4
-135.468622017598	67.5681625286617	10.1	45.8	Pice	4
-129.857489961437	68.5748888035075	13.6	48.2	Pice	4
-134.753473732719	67.2095840358642	10.1	45.8	Pice	4
-136.790327191531	66.4170035169357	7.3	16	Pice	4
-138.032518745239	65.8686864581537	12.1	29	Pice	4
-136.932927363097	65.7587201631901	7.3	16	Pice	4
-136.32085119695	65.6561921464804	10.1	45.8	Pice	4
-140.257298876319	64.1278260969475	12.1	29	Abie	4
-134.170620550791	66.0948375391246	11.7	45.9	Pice	4
-140.156898025913	63.79087137812	12.1	29	Pice	4
-139.939473953817	63.2237081130601	13.7	44.2	Pice	4
-140.272725523617	62.8830217430921	13.7	44.2	Popu	4
-139.36043465228	63.2343705797827	13.7	44.2	Popu	4
-138.020332158668	63.722096078822	13.7	44.2	Pice	4
-138.152270971263	63.586716862288	13.7	44.2	Pice	4
-138.706226451541	63.1388539481365	12.1	29	Pinu	4
-134.756339646669	64.1310775751824	13.2	19.6	Pice	4
-138.81072747654	62.7112477638504	8.5	26.1	Pice	4

Final Fire Data

-All 3 processed datasets are combined

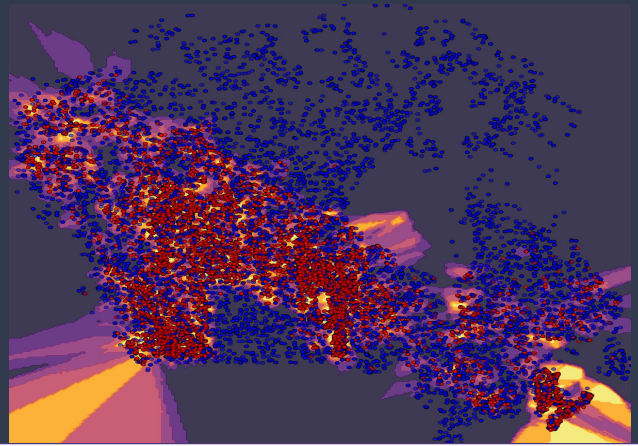
-Each fire is assigned a tree Genus

-For each coordinate in the fire dataset, search for the nearest weather station data based on date and X, Y values

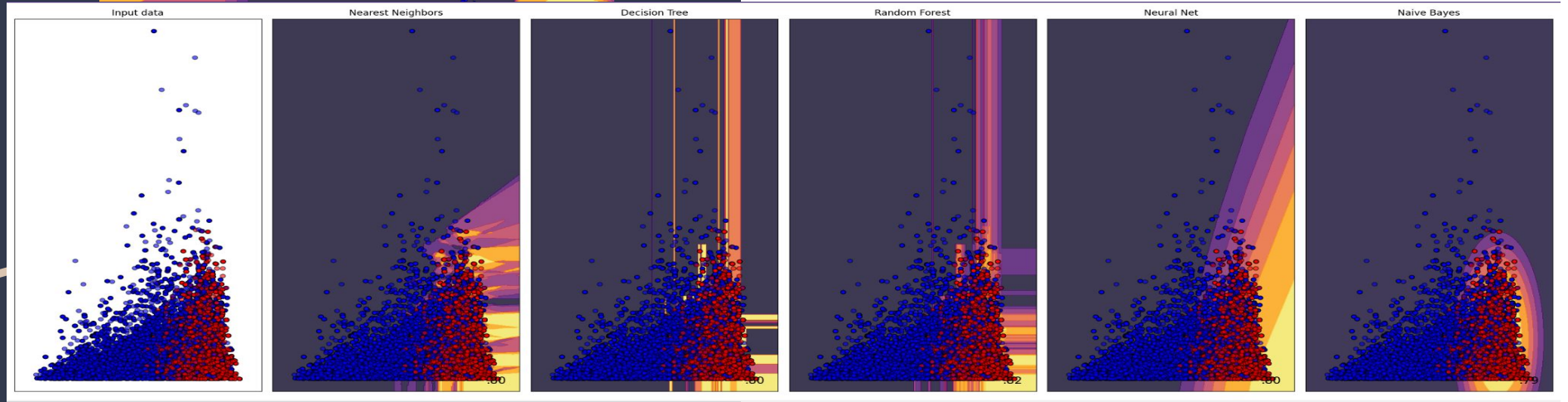
- Final CSV with good data is exported with features:

X, Y, Tm, P, Tree_Genus and BURNCLAS

Results & Methodology



- We evaluated multiple classification algorithms such as K-nn, Decision Tree, Random Forest, Neural Net, and Naïve Bayes (see below, *x* axis is temperature, *y* axis is precipitation)
- Left is K-nn with Longitude vs Latitude as *x* and *y* variables
- We chose Random Forest since in gave us the highest accuracy in our testing



Results & Methodology Cont'd

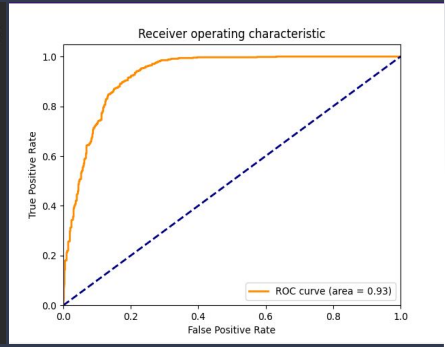
2015 Test-train Split

ORIGINAL DATA:

	X	Y	Tm	P	tree_genus	BURNCLAS
0	-140.259160	66.857167	12.9	31.6	0	1.0
1	-136.612391	67.540101	7.3	16.0	0	1.0
2	-137.052146	67.278038	7.3	16.0	0	1.0
3	-139.334883	66.410796	12.9	31.6	0	1.0
4	-135.468622	67.568163	10.1	45.8	0	1.0
...	...	...	...	...	...	...
3955	-95.740000	68.460000	-33.6	35.8	0	0.0
3956	-83.890000	68.970000	-28.2	6.9	0	0.0
3957	-107.600000	56.560000	9.5	14.0	1	0.0
3958	-107.740000	59.090000	-13.1	13.4	1	0.0
3959	-97.270000	62.550000	-18.2	33.2	0	0.0

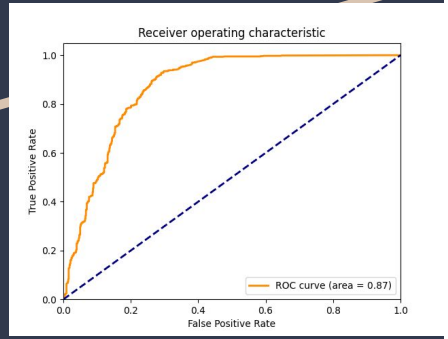
[3960 rows x 6 columns]

2015 Test-Train Split Accuracy: 86.61055853098699 %



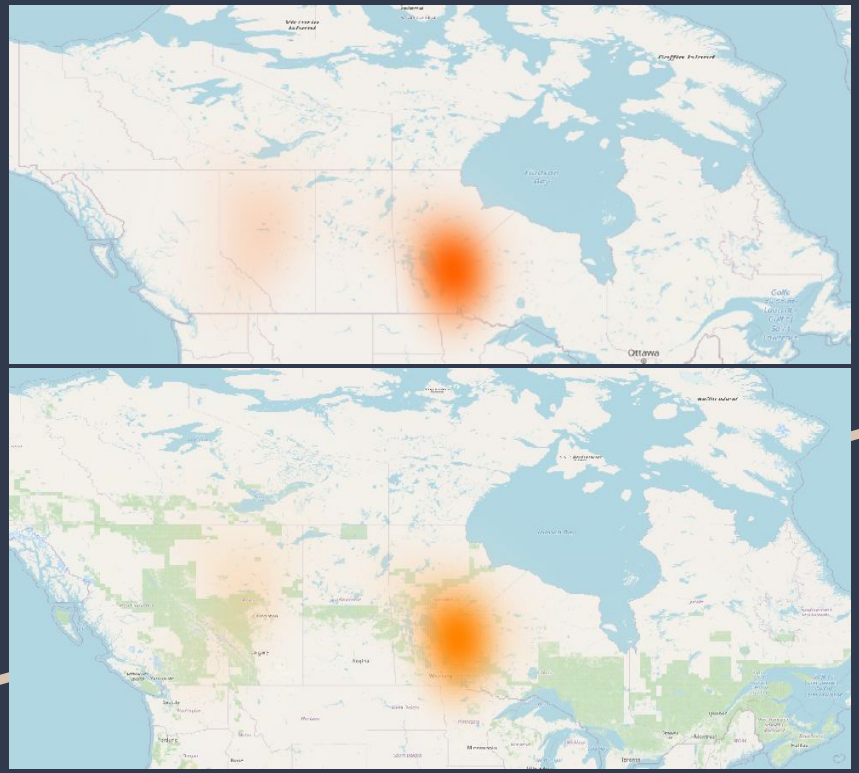
2011 data using model trained on 2015

```
5-fold CV results:
Accuracy:          0.8123835202761003
Std. Deviation: 0.06563437792776153

True positives: 642
True negatives: 605
False positives: 157
False negatives: 120

Total cases: 1524
```



- First results are using test-train split on data from 2015. We consistently predicted with ~86% accuracy.
- Second set of results were predicting on 2011 using the Random Forest model trained on the 2015 data. This gave us an accuracy of ~80% with a somewhat worse but still acceptable false positivity rate. We also performed 5-fold CV on the 2011 data predictions
- For hyperparameter tuning, we found that *m* = 2 features gave the best results, using a max tree depth of 11 and *n*-esitmators value of 20 trees.
- We found that feature scaling hurt our results by 5% so we opted not to scale the training data.

Results & Methodology Cont’d



Heatmap

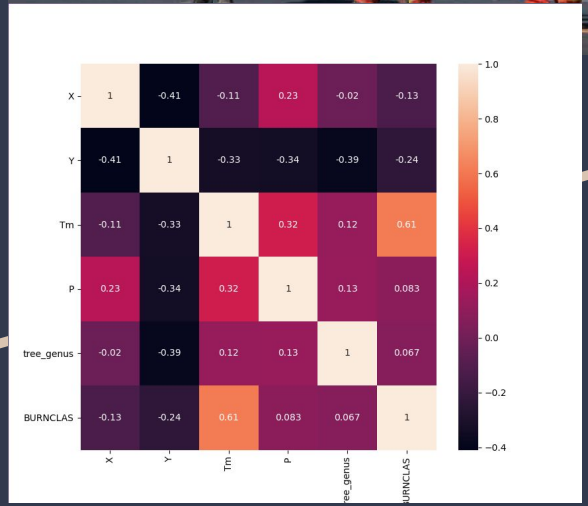
- Heat map of predicted vs actual 2011

- Predicted Top, Actual Bottom; year 2011

Tools/Libraries

- Python 3.8.6
- IntelliJ PyCharm
- Github
- Sklearn
- Pandas
- Matplotlib.pyplot
- Numpy
- Seaborn
- NJIT

Conclusions and Future Work



In our testing, one of the strongest predictors of fire risk was mean temperature, with tree genus and precipitation being weaker predictors (see correlation matrix in the bottom left). As global temperatures rise as a result of climate change, we can say with near certainty that instances of wildfires will increase in the future. For example, from the data we have from years 2011-2015, 2015 had the most fires by far (3960). By comparison, 2011 had 1524 fires and 2013 had 2400

References / Resources

https://cwfis.cfs.nrcan.gc.ca/datamart

https://ftp.maps.canada.ca/pub/nrcan_rncan/Forests_Foret/canada-forests-attributes_attributs-forests-canada/2011-attributes_attributs-2011/

https://climate.weather.gc.ca/historical_data/search_historic_data_stations_e.html?searchType=stnProv&timeframe=1&lstProvince=ON&optLimit=yearRange&StartYear=1840&EndYear=2021&Year=2021&Month=3&Day=31&selRowPerPage=25&txtCentralLatMin=0&txtCentralLatSec=0&txtCentralLongMin=0&txtCentralLongSec=0&startRow=26

https://open.canada.ca/data/en/dataset/ec9e2659-1c29-4ddb-87a2-6aced147a990