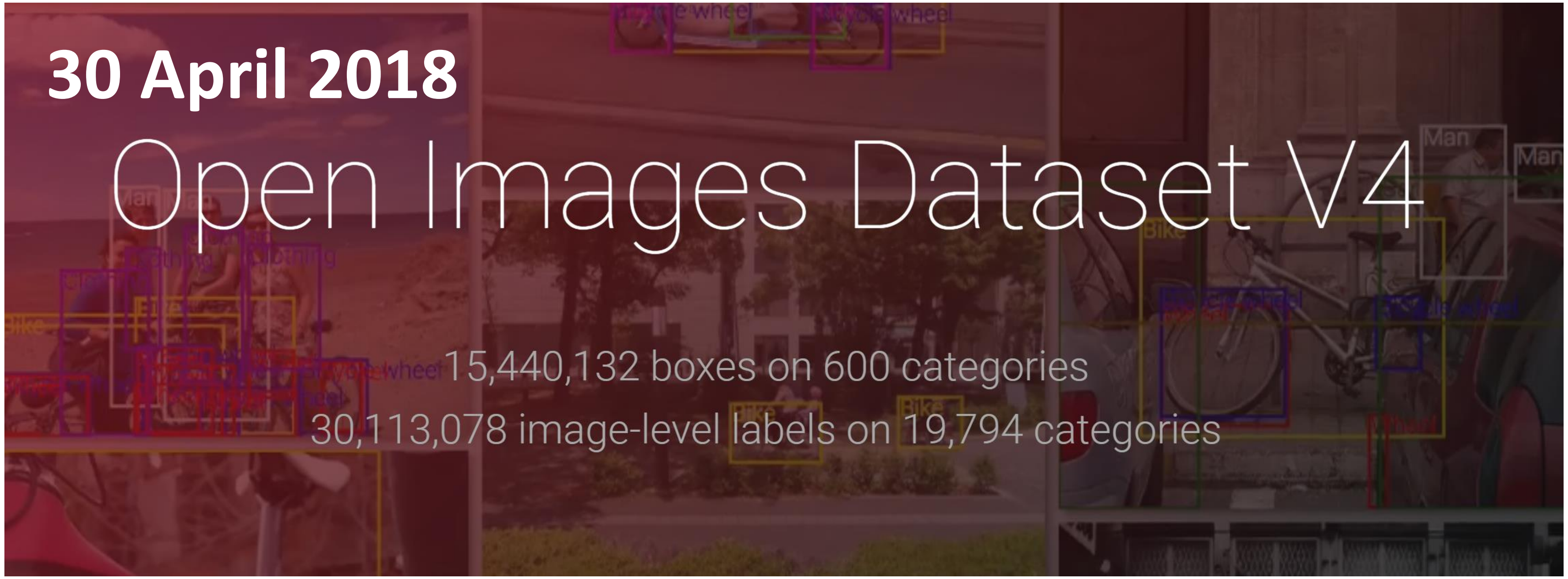
ML News

Patrycja Jakimów

07.05.2018



30 April 2018

Open Images Dataset V4

15,440,132 boxes on 600 categories

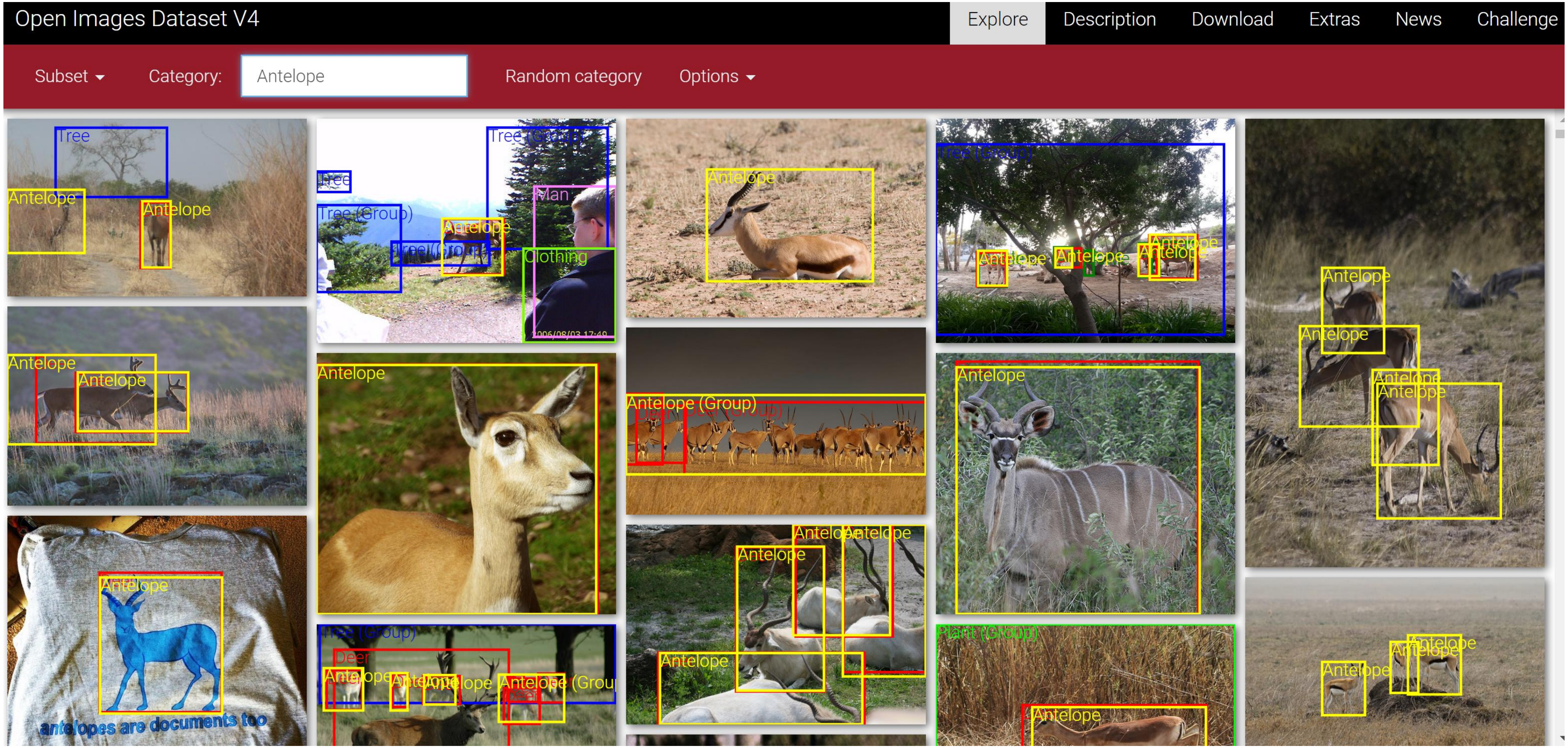
30,113,078 image-level labels on 19,794 categories

~9 million images

complex scenes with several objects

the largest existing dataset with object location annotations

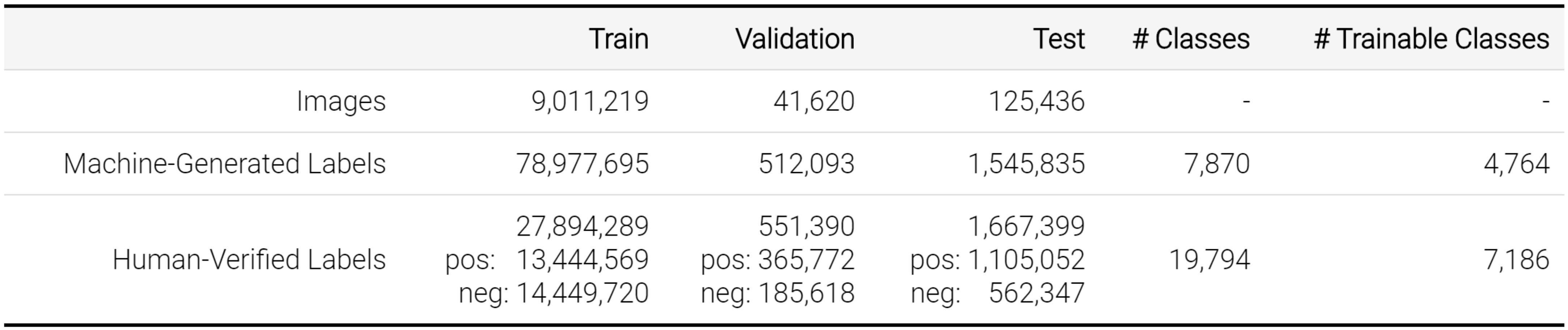
Open Image Dataset V4



Category: Antelope

Image-level labels

Table 1: Image-level labels.



	Train	Validation	Test	# Classes	# Trainable Classes
Images	9,011,219	41,620	125,436	-	-
Machine-Generated Labels	78,977,695	512,093	1,545,835	7,870	4,764
Human-Verified Labels	27,894,289 pos: 13,444,569 neg: 14,449,720	551,390 pos: 365,772 neg: 185,618	1,667,399 pos: 1,105,052 neg: 562,347	19,794	7,186

pos - certain object classes are present
neg - certain object classes are absent

Boxes

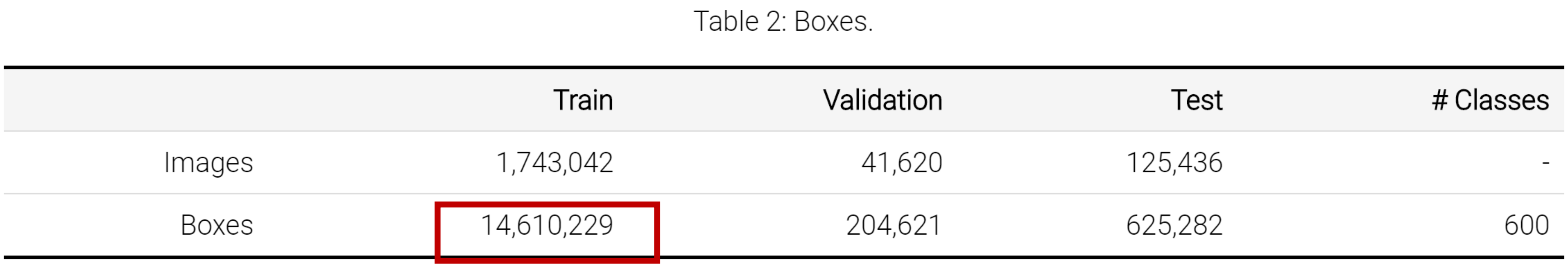


Table 2: Boxes.

	Train	Validation	Test	# Classes
Images	1,743,042	41,620	125,436	-
Boxes	14,610,229	204,621	625,282	600

90% of the boxes were manually drawn by professional annotators at Google using the efficient extreme clicking interface

We produced the remaining 10% semi-automatically

Data Formats

image-level labels

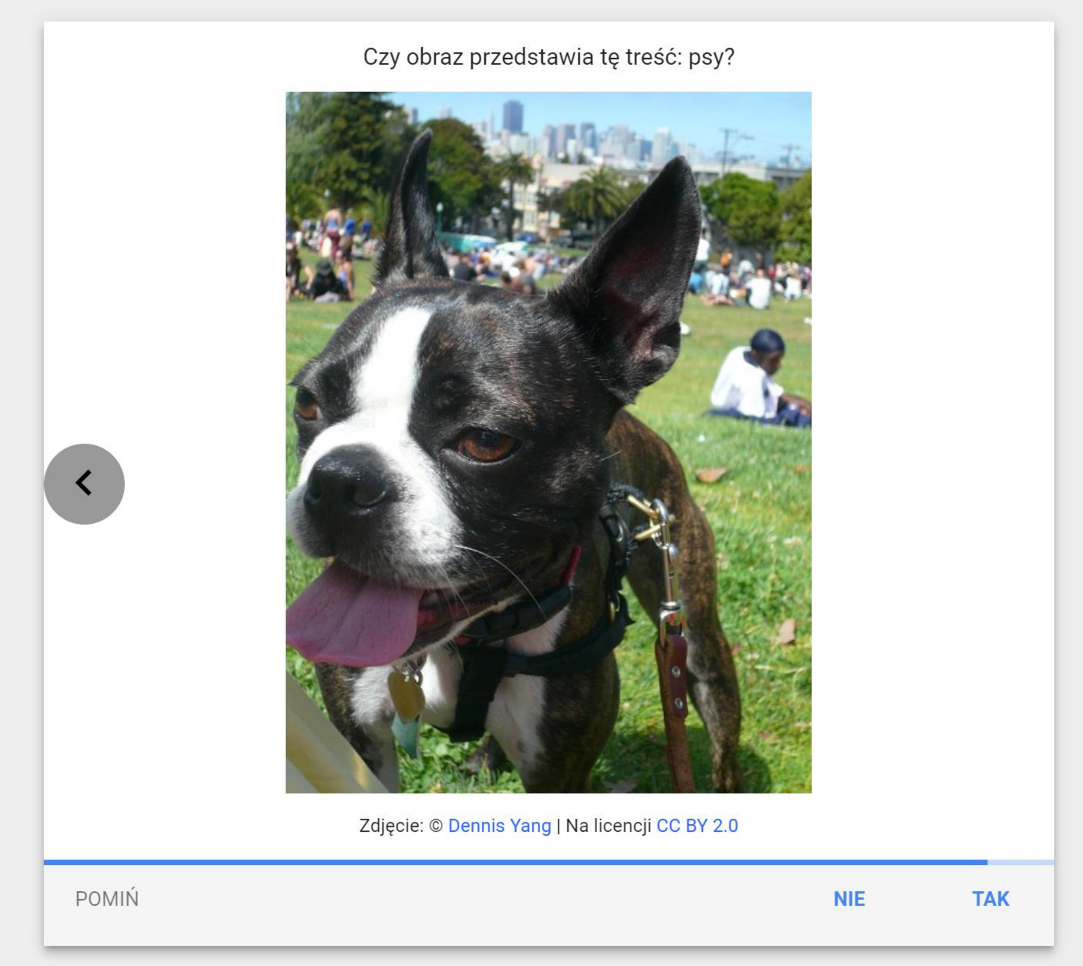
```
ImageID,Source,LabelName,Confidence
000026e7ee790996,verification,/m/04hgtk,0
000026e7ee790996,verification,/m/07j7r,1
000026e7ee790996,crowdsource-verification,/m/01bqvp,1
000026e7ee790996,crowdsource-verification,/m/0csby,1
000026e7ee790996,verification,/m/01_m7,0
000026e7ee790996,verification,/m/01cbzq,1
000026e7ee790996,verification,/m/01czv3,0
000026e7ee790996,verification,/m/01v4jb,0
000026e7ee790996,verification,/m/03d1rd,0
...
```

`Source`: indicates how the annotation was created:

- `verification` are labels verified by in-house annotators at Google.
- `crowdsource-verification` are labels verified from the Crowdsource app.
- `machine` are machine-generated labels.

`Confidence`: Labels that are human-verified to be present in an image have confidence = 1 (positive labels). Labels that are human-verified to be absent from an image have confidence = 0 (negative labels). Machine-generated labels have fractional confidences, generally >= 0.5. The higher the confidence, the smaller the chance for the label to be a false positive.

Crowdsource



Czy obraz przedstawia tę treść: psy?

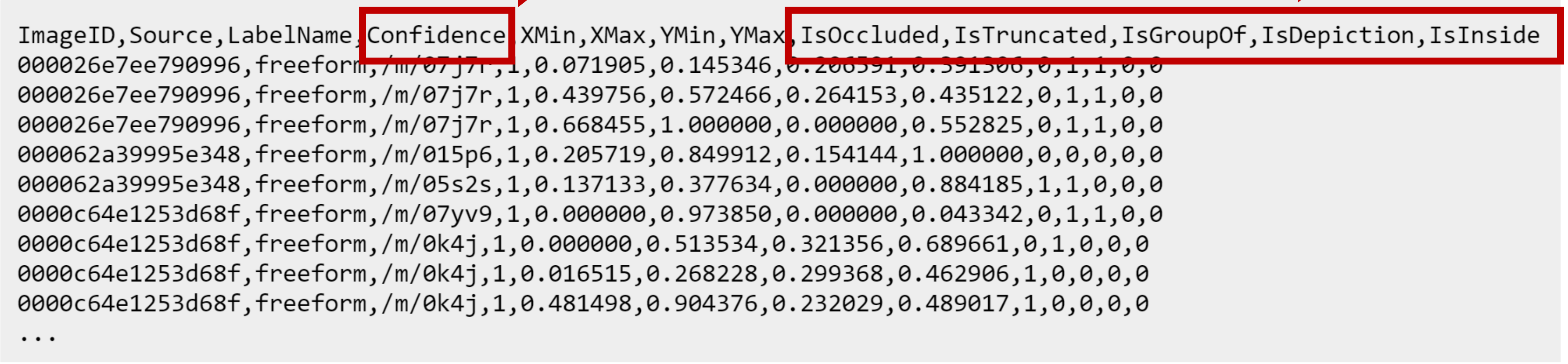
Zdjęcie: © Dennis Yang | Na licencji CC BY 2.0

Data Formats

bounding box

always 1

additional attributes
for the validation and test sets

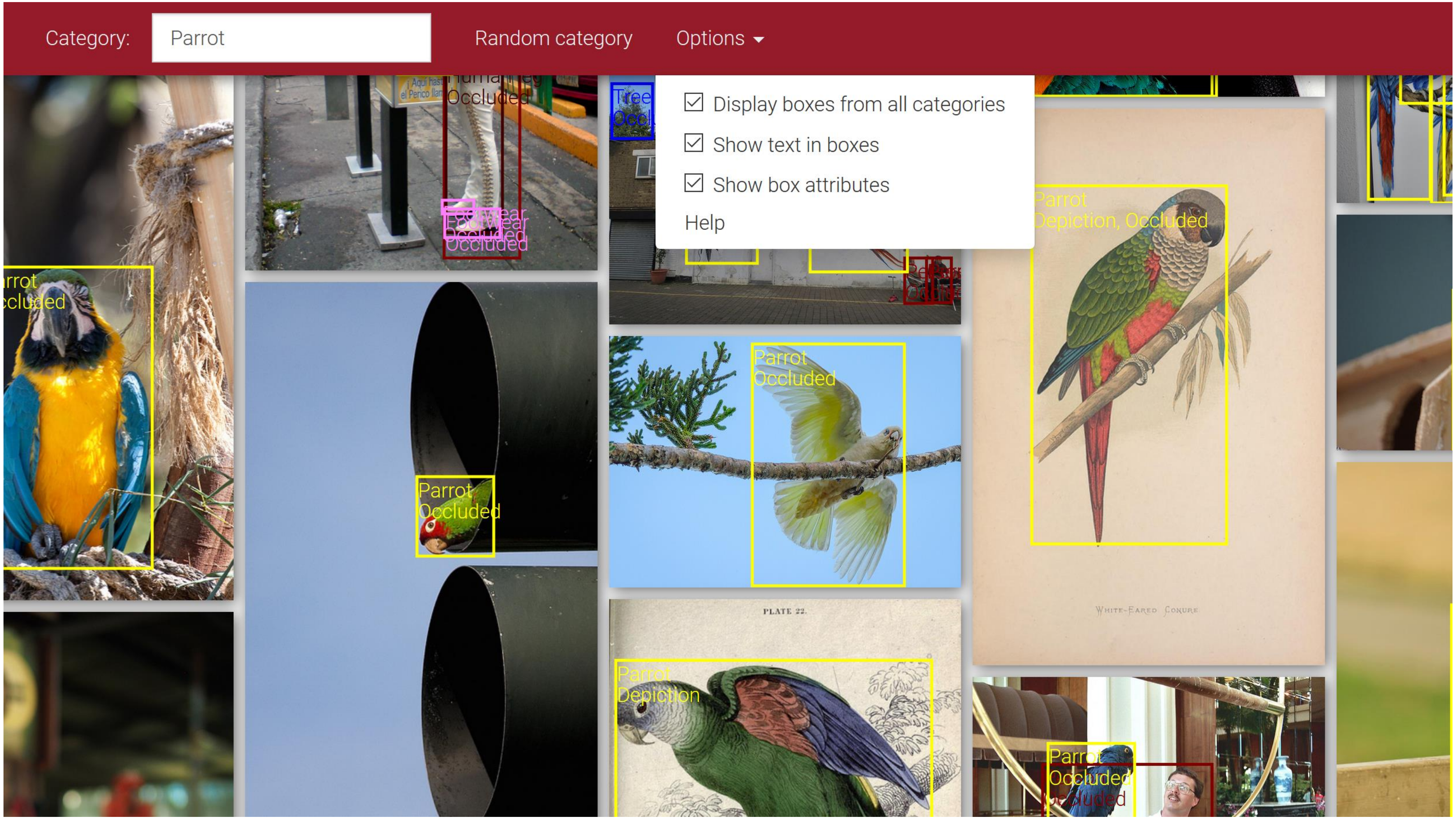


```
ImageID,Source,LabelName,Confidence,XMin,XMax,YMin,YMax,IsOccluded,IsTruncated,IsGroupOf,IsDepiction,IsInside
000026e7ee790996,freeform,/m/07j7r,1,0.071905,0.145346,0.206591,0.391306,0,1,1,0,0
000026e7ee790996,freeform,/m/07j7r,1,0.439756,0.572466,0.264153,0.435122,0,1,1,0,0
000026e7ee790996,freeform,/m/07j7r,1,0.668455,1.000000,0.000000,0.552825,0,1,1,0,0
000062a39995e348,freeform,/m/015p6,1,0.205719,0.849912,0.154144,1.000000,0,0,0,0,0
000062a39995e348,freeform,/m/05s2s,1,0.137133,0.377634,0.000000,0.884185,1,1,0,0,0
0000c64e1253d68f,freeform,/m/07yv9,1,0.000000,0.973850,0.000000,0.043342,0,1,1,0,0
0000c64e1253d68f,freeform,/m/0k4j,1,0.000000,0.513534,0.321356,0.689661,0,1,0,0,0
0000c64e1253d68f,freeform,/m/0k4j,1,0.016515,0.268228,0.299368,0.462906,1,0,0,0,0
0000c64e1253d68f,freeform,/m/0k4j,1,0.481498,0.904376,0.232029,0.489017,1,0,0,0,0
...
```

The attributes have the following definitions:

- `IsOccluded`: Indicates that the object is occluded by another object in the image.
- `IsTruncated`: Indicates that the object extends beyond the boundary of the image.
- `IsGroupOf`: Indicates that the box spans a group of objects (e.g., a bed of flowers or a crowd of people). We asked annotators to use this tag for cases with more than 5 instances which are heavily occluding each other and are physically touching.
- `IsDepiction`: Indicates that the object is a depiction (e.g., a cartoon or drawing of the object, not a real physical instance).
- `IsInside`: Indicates a picture taken from the inside of the object (e.g., a car interior or inside of a building).

`XMin`, `XMax`, `YMin`, `YMax`: coordinates of the box, in normalized image coordinates.



Category: Parrot

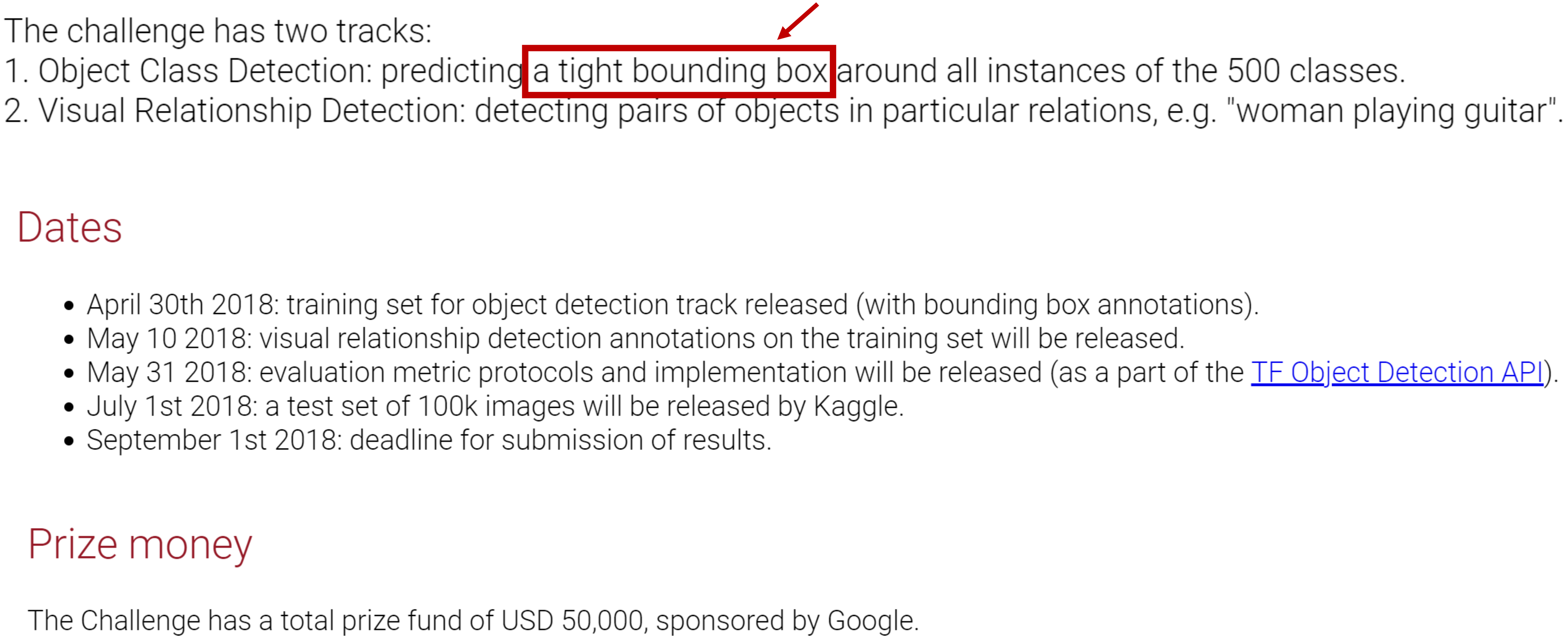
☑ Display boxes from all categories

☑ Show text in boxes

☑ Show box attributes

Open Images Challenge 2018

broad case: clothing



The challenge has two tracks:
1. Object Class Detection: predicting a tight bounding box around all instances of the 500 classes.
2. Visual Relationship Detection: detecting pairs of objects in particular relations, e.g. "woman playing guitar".

Dates

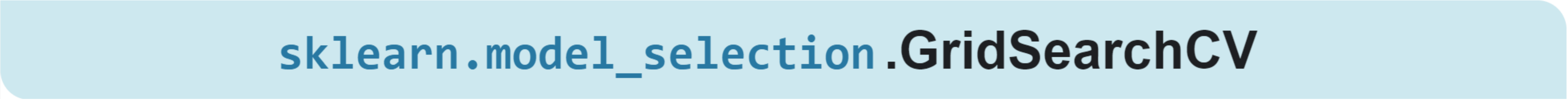
- April 30th 2018: training set for object detection track released (with bounding box annotations).
- May 10 2018: visual relationship detection annotations on the training set will be released.
- May 31 2018: evaluation metric protocols and implementation will be released (as a part of the TF Object Detection API).
- July 1st 2018: a test set of 100k images will be released by Kaggle.
- September 1st 2018: deadline for submission of results.

Prize money

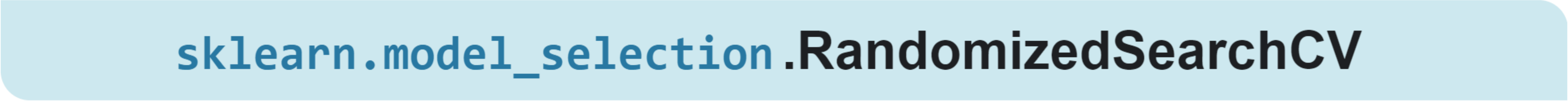
The Challenge has a total prize fund of USD 50,000, sponsored by Google.

Tuning the hyper-parameters of an estimator

Article on Data Science PL



sklearn.model_selection.GridSearchCV

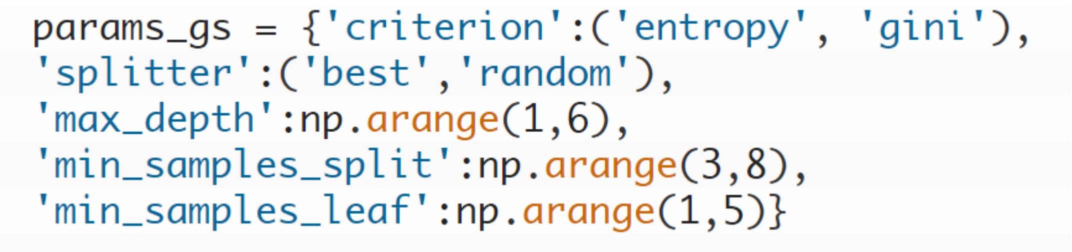


sklearn.model_selection.RandomizedSearchCV

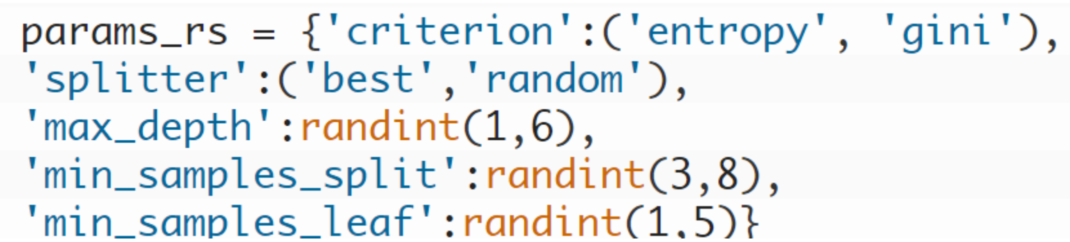
best_params_ : dict

best_score_ : float

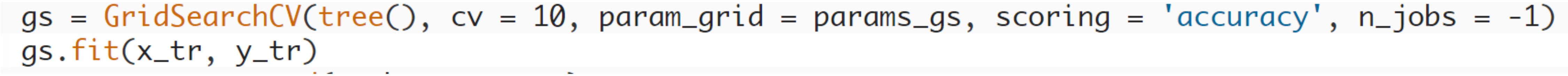
Example



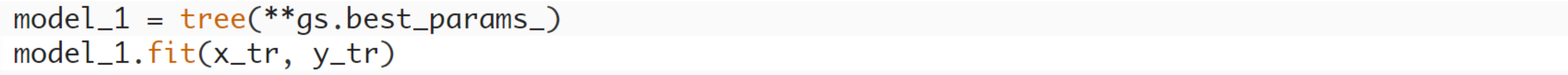
```
params_gs = {'criterion':('entropy', 'gini'),
'splitter':('best','random'),
'max_depth':np.arange(1,6),
'min_samples_split':np.arange(3,8),
'min_samples_leaf':np.arange(1,5)}
```



```
params_rs = {'criterion':('entropy', 'gini'),
'splitter':('best','random'),
'max_depth':randint(1,6),
'min_samples_split':randint(3,8),
'min_samples_leaf':randint(1,5)}
```



```
gs = GridSearchCV(tree(), cv = 10, param_grid = params_gs, scoring = 'accuracy', n_jobs = -1)
gs.fit(x_tr, y_tr)
```



```
model_1 = tree(**gs.best_params_)
model_1.fit(x_tr, y_tr)
```

Resources

- Google Research Blog

https://research.googleblog.com/

- Open Images Dataset v4

https://storage.googleapis.com/openimages/web/index.html

- Crowdsource

https://crowdsource.google.com/

- Data Science PL Group on Facebook

https://www.facebook.com/groups/datasciencepl/

- 2 proste i skuteczne metody optymalizacji parametrów modelu

https://mateuszgrzyb.pl/2-proste-i-skuteczne-metody-optymalizacji-parametrow-modelu/

- RandomizedSearchCV

http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

- GridSearchCV

http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Thank you