Save the Whales with your data-scientist' superpowers

Amine EL HELOU,
amine.elhelou@mathworks.fr

NOAA FISHERIES

New England Aquarium

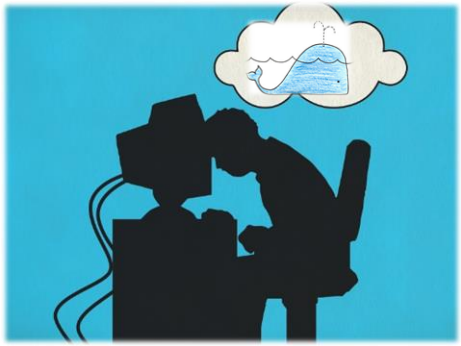
kaggle

Right Whales are in danger…

- $N \leq 500$ Right Whales left in the oceans

- Actually, whales are photographed during aerial surveys & then manually matched to a photo-identification catalog

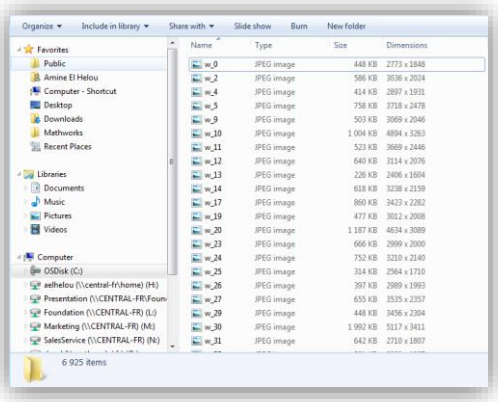
➢ **Extremely time consuming & requires special training**



✓ Real-time recognition would allow researchers to spontaneously save whales as they struggle to free themselves from fishing gears

https://www.kaggle.com/c/noaa-right-whale-recognition

The dataset



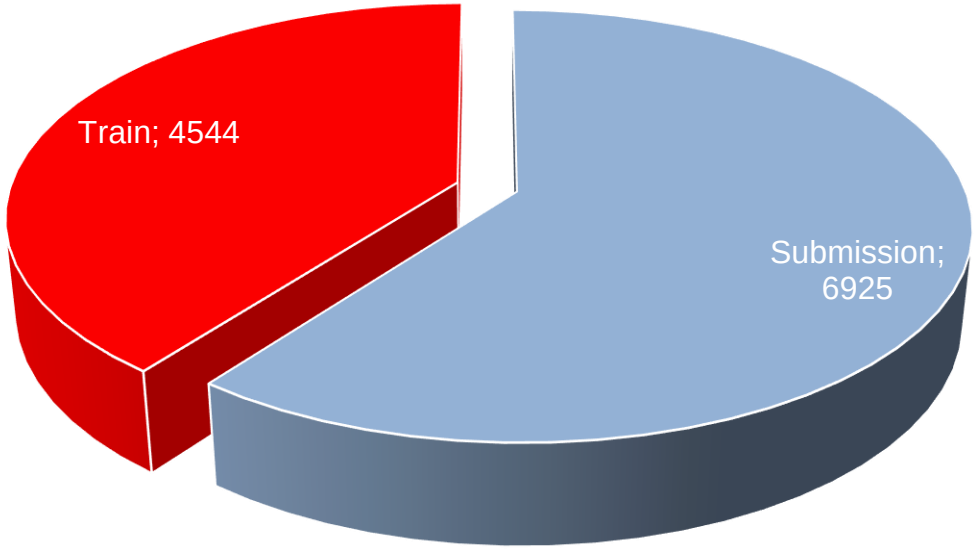
- 8.8 GB of Aerial images (11 469 .jpeg Images)

Data Files

File Name	Available Formats
sample_submission.csv	.zip (49.44 kb)
train.csv	.zip (24.12 kb)
imgs	.zip (8.73 gb)

N = 447
single classes

	1 whaleID	2 ImgCount	3 FileList
1	whale_95370	47	47x1 cell
2	whale_38681	43	43x1 cell
3	whale_28892	33	33x1 cell
4	whale_90957	30	30x1 cell
5	whale_36851	30	30x1 cell
6	whale_24458	29	29x1 cell
7	whale_85464	28	28x1 cell
8	whale_51195	28	28x1 cell
9	whale_65586	27	27x1 cell
10	whale_52749	26	26x1 cell
11	whale_89615	24	
12	whale_78280		
	whale_34656		
		1	'_6656.jpg'
445	whale_08729	1	'_5252.jpg'
446	whale_05140	1	'_3576.jpg'
447	whale_04480	1	'_1067.jpg'

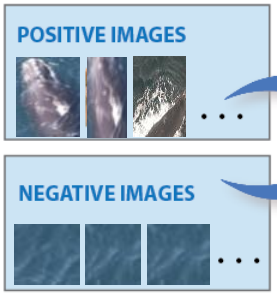


https://www.kaggle.com/c/noaa-right-whale-recognition/data

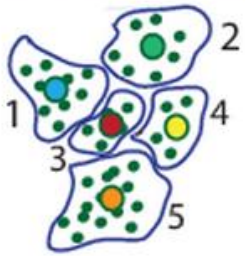
A computer vision + ML problem… ?

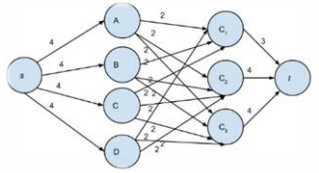
Proposed starter codes: a 2 step approach

1. Whale detection from images using a (*semi-automated*) Vision model



2. Supervised classification using features extracted from detected whale





https://www.kaggle.com/c/noaa-right-whale-recognition/details/creating-a-face-detector-for-whales

A computer vision + ML problem… ?
Obvious challenges

- Image quality variability
- Image rotations ➔ correction ?
- Manual labelling part ➔ operator dependent

- Model (& Cascade detector) implementation…?
  - One vs. All (train 1 model for each specimen vs. all rest)
  - Multi-class
    - Non-homogenous response distributions

How am I scored… ?

- Evaluation criteria == Multi-class Logarithmic Loss
  - For each image: submit a set of predicted probabilities (one for every whale)

Image	whale_00195	whale_00442	whale_02411	whale_02608	whale_02839	whale_03103	whale_03227	whale_03623	whale_03	whale_03935
w_1947.jpg	1	0	0	0	0	0	0	0	0	0
w_11096.jpg	1	0	0	0	0	0	0	0	0	0
w_10973.jpg	1	0	0	0	0	0	0	0	0	0
w_10442.jpg	1	0	0	0	0	0	0	0	0	0
w_10606.jpg	1	0	0	0	0	0	0	0	0	0
w_11167.jpg	1	0	0	0	0	0	0	0	0	0
w_1464.jpg	1	0	0	0	0	0	0	0	0	0
w_5072.jpg	1	0	0	0	0	0	0	0	0	0
w_5492.jpg	1	0	0	0	0	0	0	0	0	0
w_10124.jpg	1	0	0	0	0	0	0	0	0	0
w_9503.jpg	1	0	0	0	0	0	0	0	0	0
w_11267.jpg	1	0	0	0	0	0	0	0	0	0
w_1590.jpg	1	0	0	0	0	0	0			
w_10175.jpg	1	0	0	0	0					
	1	0	0							

sample_submission.csv

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

Submission Limits

You may submit a maximum of 5 entries per day.
You may select up to 2 final submissions for judging.

- N: number of test images
- M: number of whale labels
- y_{ij} : 0/1 if observation ***i*** belongs to whale ***j***
- p_{ij}: predicted probability

https://www.kaggle.com/c/noaa-right-whale-recognition/details/evaluation

So join the fight…

- ✓ Prizes:
  - 5000/3000/2000 $
- ✓ Deadline: 7th of January 2016

107	teams
127	players
552	entries

NOAA FISHERIES · New England Aquarium · MathWorks · kaggle

https://www.kaggle.com/c/noaa-right-whale-recognition