# Machine Learning Engineer Nanodegree

## Capstone Proposal

Zeyad Obaia
January 28th, 2019

### Domain Background

After centuries of intense whaling, recovering whale populations still have a hard time adapting to warming oceans and struggle to compete every day with the industrial fishing industry for food.

To aid whale conservation efforts, scientists use photo surveillance systems to monitor ocean activity. They use the shape of whales' tails and unique markings found in footage to identify what species of whale they're analyzing and meticulously log whale pod dynamics and movements. For the past 40 years, most of this work has been done manually by individual scientists, leaving a huge trove of data untapped and underutilized.

### Problem Statement

The problem is a straight forward classification problem the goal is to to build an algorithm to identify individual whales in images.

Image classification is a wide area of research that has always fascinated me, and the work that has been done using **Convolutional Neural Networks** has shown a lot of success.

### Datasets and Inputs

The dataset can be found on Kaggle, This training data contains thousands of images of humpback whale flukes. Individual whales have been identified by researchers and given an Id. The challenge is to predict the whale Id of images in the test set.

What makes this such a challenge is that there are only a few examples for each of 3,000+ whale Ids, So my kernel should try to find patterns from small number of datapoints.

The dataset is around 6 GB and is over 25,000 images not of the same size. I will be using the whole dataset, I will have to resize the images before feeding them to the network, this will

happen in the data preprocessing step.
About 4.2GB will be used as training data and the rest will be used for testing/validation.

## Solution Statement

We start with a clean dataset, fortunately the amount of data cleaning will be minimum.

Once the data is ready, We will train a Multi-layer Perceptron to set a benchmark.

Then we are going to train at least one Convolutional Neural Network to see if we can beat the benchmark we set earlier.

Finally we will use Image Augmentation and fine-tune our model/ try different optimizer to achieve better accuracy.

## Benchmark Model

I intend to use a simple Multi-layer Perceptron (MLP) with fully connected layers as a benchmark model to define as baseline before converting to using Convolutional Neural Networks.

## Evaluation Metrics

Using **categorical crossentropy** as a loss function and **categorical accuracy** / **f1_score** as evaluation metrics seems like a sound choice since we are dealing with multiple types of whales that we are trying to classify and the dataset is quite unbalanced.

## Project Design

The following steps describe the workflow for this project:

1) Data preprocessing

The data provided by Kaggle needs a bit of prepossessing:

- I will use one hot encoding to categorize the labeled data.
- Resizing the images might become useful.

2) Constructing Benchmark model and test it

- Build an MLP with fully connected layers as a benchmark model.
- Evaluate the model against the test dataset.

3) Constructing Final Model using CNNs and beat the benchmark result.

- Build Convolutional Neural Network / fine-tune a pre-trained network (transfer learning)
- Use dropout layers to overcome overfitting.
- Use Max Pooling & Average Pooling layers to extract important features like edges.

4) Add Image Augmentation and Fine tune the model.

- Image Augmentation to produce images with different angles
- Try different architectures, loss functions and optimizers to produce a better result.