

Homework 5: Homebrew computer vision

1) Download the zip file below. Look at some of the images, noting that there are 50 classes in 4244 images (e.g. "goldfish", "llama", "speed-boat", ...). Caution: it's a pretty large file (~208M).

http://astro.berkeley.edu/~ishivvers/data/50_categories.tar.gz

2) Write a set of methods that takes as input one of these images, and then computes real-numbered **features** as the return.

note: you should produce **at least** 15 features. Some of them can be dumb, like the image array size and the average of the red-channel colors but some should be interesting in that they capture some of the richness of these images, like cross-correlations between the R, G, B channels, some based on edge-detection and image segmentation, etc.

3) Based on the feature set for each image, build a **random forest classifier** (`scikit-learn`). Produce metrics on your estimated error rates using cross-validation. How much better is this than the expectation with random guessing? What are the 3 most important features?

4) Make sure your final classifier can run on a directory of different images, where a call like:
`run_final_classifier("/new/directory/path/")`

on directory that contains files like:

```
validation1.jpg  
validation2.jpg  
...
```

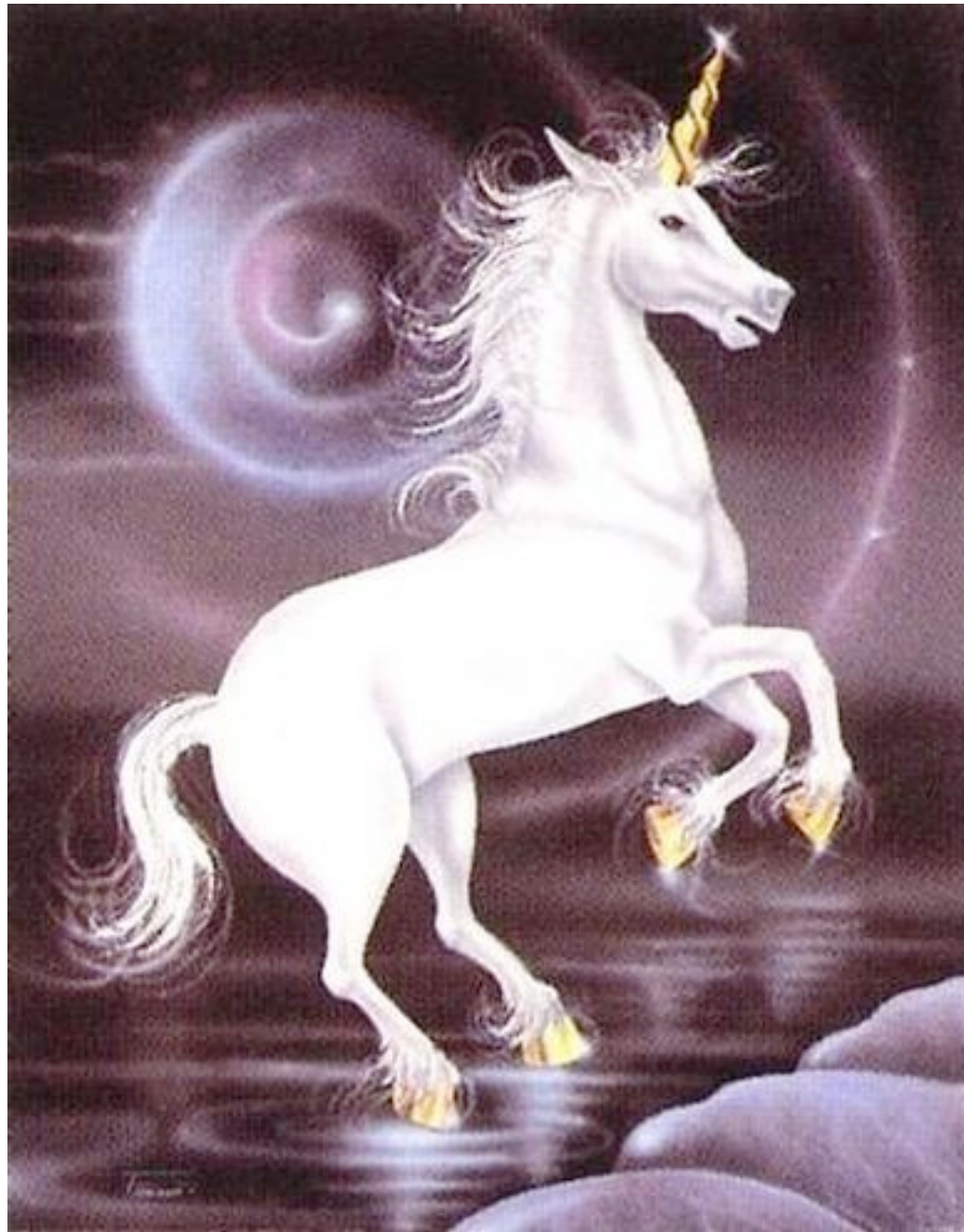
will produce an output file that looks like:

```
filename          predicted_class  
-----  
validation1.jpg   unicorn  
validation2.jpg   camel  
...
```

We will have a **validation set** to test how good your classifier is. The best classifier among those submitted will earn its writer a perfect score on this homework.

Hints: you might want to read about computer vision (CV) to get a sense of what good image features are (e.g., <http://opencv.itseez.com/>).

Good Luck!



unicorn_0012.jpg