You may do this project in groups of one or of two. The answers are due by email, to daf@cs.uiuc.edu, with CS 498 in the title, by **12h00, 14 Dec (the day of the final exam).** This is a deliberately open ended activity - I'll describe the minimum, but suggest improvements. If you have a truly lousy grade on the midterm (or anticipate one on the final) this is a good method to win redemption.

Construct a dataset of either images or sound, with at least 100 training examples and at least 10 test examples each of at least two types. Possibilities include, but are not limited to: speech and music; pictures of outdoor scenes and pictures of cats; magazine covers and news pictures; male speech and female speech.

Build a feature representation for your data by vector quantization using k-means, then building a histogram. The feature representation should be trained using only training examples. Train a classifier using the training examples, then evaluate using the test examples.

Possible improvements and extensions: This is a very open ended exercise. It's a good idea to collect more than the minimum of training data. When you classify, you could try using a histogram intersection kernel (look it up; most good modern SVM codes will do this for you). It's a good idea to try multi-way classification. Here, I'd do one vs all. You should find that having a reasonable choice of unlabelled data helps build a better dictionary, and so improves classification. It's worth investigating how the choice of k in k-means affects performance. You could also look at the question of where patches should be formed --- does it help to look at every pixel or time sample, where you get lots of overlapping patches, or should you space them more widely? Notice also that the vector quantized representation loses some detail (because you represent a patch as a cluster center number, and some patches are far away from their cluster centers). One way that people deal with this, rather successfully, is to build three dictionaries with different start points, classify everything three times, then vote; this is quite successful, and worth trying.