**Image Dataset Details**

**Dataset Format**

As noted in the project Description, you'll be working with a dataset of human faces.  The images have dimensionality 100x100, and are (mostly) colour.

The data is split into train\_images.npy, val\_images.npy and test\_images.npy, with a test\_private\_images.npy to come later.  Labels are also available for the training data and validation data, in train\_labels.npy and val\_labels.npy respectively.  (npy files are a standard format for files using numpy.)

The labels have the form x-y, where

* x is an identifier specific to an individual; and
* y is the part of the label that corresponds to age.

As an example, the label 12726-37 corresponds to a person who is 37 years old, with ID 12726.  (This happens to be the second person in the devset, the actor Gabrielle Anwar.)

The ages fall into 3 groups: ages 6-20, 35-40, and 55-98.  There can be multiple images of the same person: there are several images of Gabrielle Anwar, ID 12726.  The dataset is a noisy one, scraped from the web, so not all the images that should be of the same person, with the same ID, are in fact of that person.  (The ninth image in the devset, which also has ID 12726, is of another actor from the series Burn Notice that Gabrielle Anwar was a lead actor in.)  In the provided dataset, it is guaranteed that all of the images with a particular ID will be in the same age group.

In iLearn, there is also a [Jupyter notebook with a code snippet](https://ilearn.mq.edu.au/mod/folder/view.php?id=7242013), to get you started on how to read in the data.

**Kaggle Submission**

As noted in the project Description, the project is hosted at [Kaggle](https://www.kaggle.com/) InClass [competition link [here](https://www.kaggle.com/t/82679cfead68b2f27479b9e403f56718)]: that's where you'll find the dataset, and you will submit your assignment solutions there.

To submit a solution to Kaggle InClass:

* You need a Kaggle account, which is free. If you already have one, you should use that: Kaggle doesn't allow multiple accounts. See the Kaggle InClass [FAQ](https://www.kaggle.com/c/about/inclass/faqs) for details.
* You have to produce a csv file with two columns. Column A should have the string "ID" in row 1, and digits 0..N-1 in subsequent rows, with N the number of images in the dataset. Column B should have the string "Prediction" in row 1, and in subsequent rows the prediction made by your system for the image with that ID in column A. That prediction should be 0 (if the age of the person is 6-20), 1 (ages 35-40), or 2 (ages 55-98).  There is a sample submission available in the Kaggle competition (under Data), where the prediction is label 0 for every image.
* You submit this csv file at the appropriate Kaggle page for the hosted competition.

**Baselines**

At project release there is one baseline:

* The majority class baseline (baseline\_majority.csv) with accuracy 0.42207.