# COM6018 Data Science with Python

## Week 8: Classification with scikit-learn

Jon Barker

## In this lab

Using scikit-learn to compare different classifiers.

- Preparing a face recognition task
- Performing feature pre-processing and dimensionality reduction
- Building k-NN, SVM, Random Forest and MLP classifiers
- Using GridSearchCV to perform parameter optimisation
- Using pipelines to streamline the evaluation process
- Evaluating using confusion matrices and per-class precision, recall and F1-scores

## The Task

We will use scikit-learn to recognise famous people from photographs.

- We will use scikit-learn's built-in dataset, 'Labeled Faces in the Wild'
- Details here, http://vis-www.cs.umass.edu/lfw/
- It contains 13,233 images of 5,749 famous people.
- It is designed for a face verification task, but we will use a subset of it for a classification task.

## Example Data

*Examples from the 'Labeled Faces in the Wild' dataset.*

## Example Data

*Examples from the 'Labeled Faces in the Wild' dataset.*

## The Data

```
dict_keys(['data', 'images', 'target', 'target_names', 'DESCR'])
```

images - 3D NumPy array storing original colour images (N x height x width)

data - 2D NumPy array storing pre-processed images (N x n_pixels)

target - array of N labels (stored as integers)

target_names - list of the class names

## Preprocessing

The data has been pre-processed.

- Images were cropped to the face.
- They were transformed to greyscale.
- They were resized to 125x94 pixels.

## Obtaining the Jupyter Notebook

If you have cloned and pulled the module's GitHub repository then you should see:

```
materials/labs/
├── 070_classification_with_scikit_learn.ipynb
|-- ... etc
```

The lab is `070_classification_with_scikit_learn.ipynb`. It does not require any additional data files. (The dataset is built into scikit-learn.)

Or you can download the notebook and data via links on Blackboard.

## Getting Help

- If you are stuck just raise a hand to ask for help.
- Feel free to discuss the lab with your neighbours.
- Re-read the scikit-learn tutorial notes
  - In the Git repo at `materials/tutorials/070_Classification_with_Scikit_Learn.ipynb`
  - or online at https://uos-com-6018.github.io/COM6018
- Use the scikit-learn API documentation for reference. https://scikit-learn.org/stable/modules/classes.html