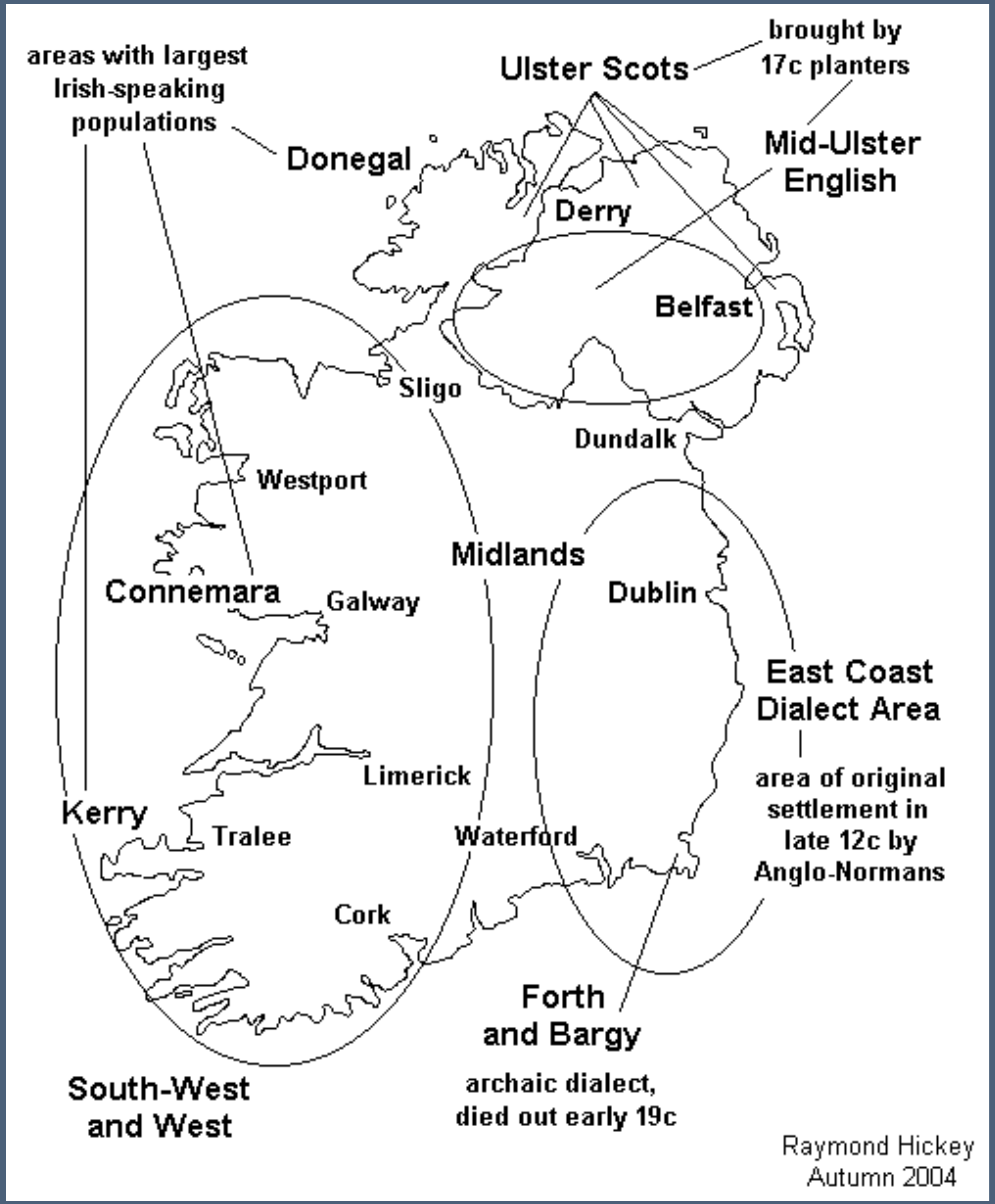
Machine Learning for Feature-Extraction and Classification of English-language Accents in Ireland

Peter Nolan
x22154116
MSCDATOP-Research in Computing

National College of Ireland



Research Question

How can we classify some selected regional accents in Ireland using data analysis techniques based on the features within the sound of their speech and on demographic characteristics of the speakers?

- What datasets will support this data analysis in computational linguistics?
- What features, either the sound recording data or features calculated from it, or personal characteristics, will be useful inputs to the classification model?
- What data analysis models will perform this well?

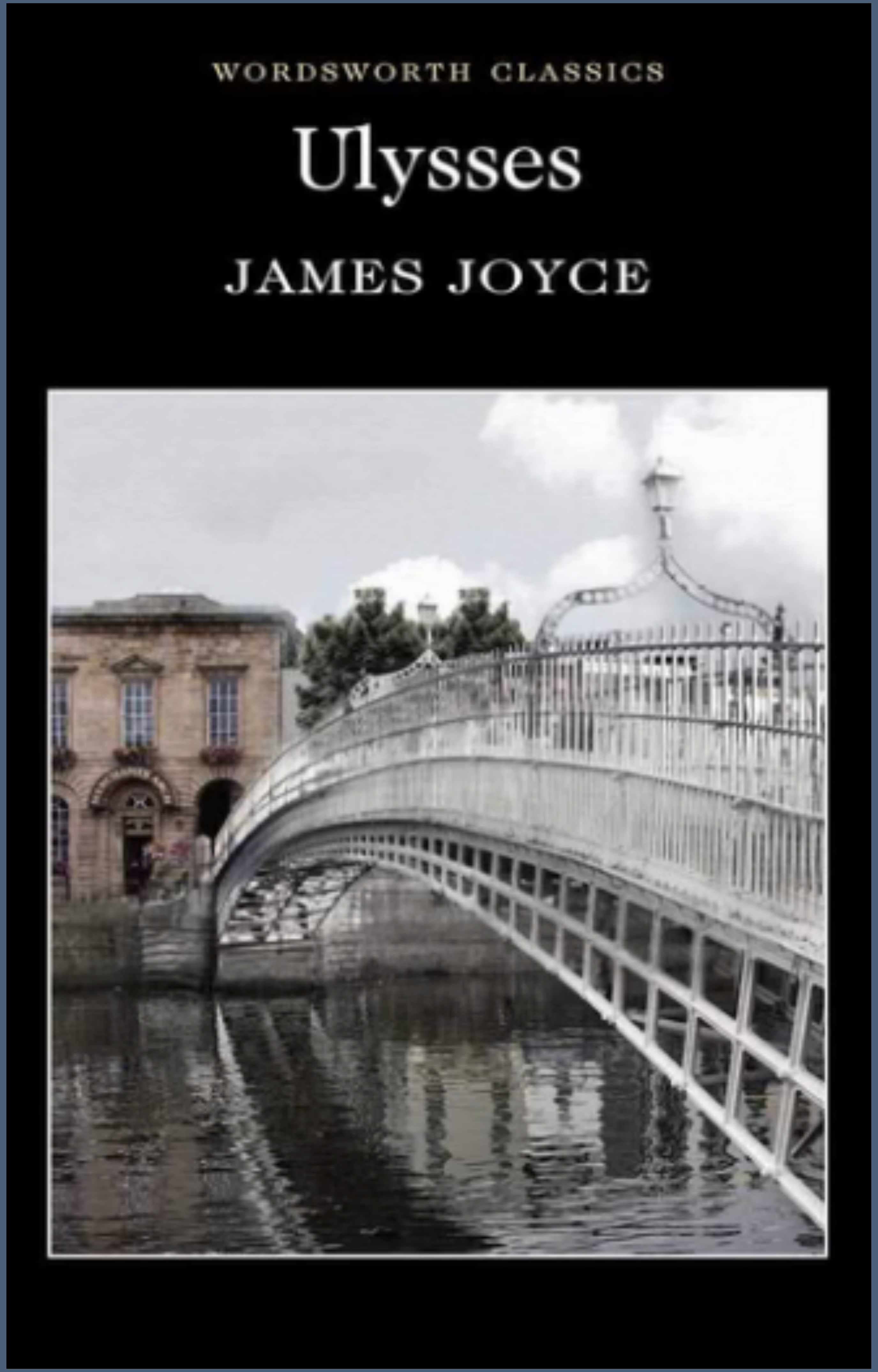
Irish Culture Focuses on Accent Diversity

Literature, Comedy and Satire

- Joyce - especially Ulysses recorded the voices of the Irish

- Niall Toibin: A Guide to the Regional Accents of Ireland
- https://www.youtube.com/watch?v=EhLdKJnY194

- Ross O'Carroll-Kelly, the `DORT' accent in his audiobook
- https://www.youtube.com/watch?v=vlobTz-9_Fc



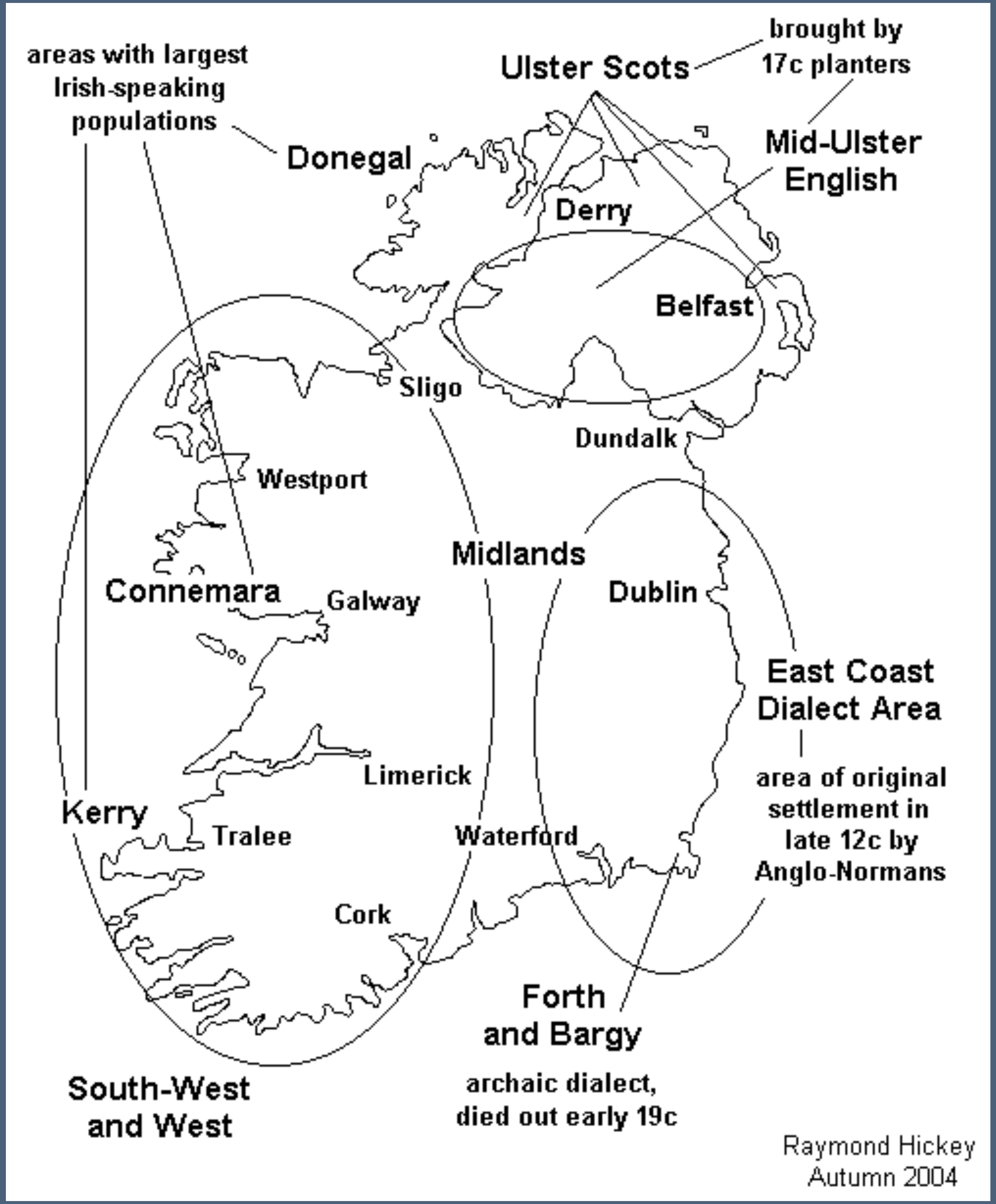
Datasets

Sound Atlas of Irish English ('SAIE')
(R.Hickey, 2004) Prof. Linguistics at UL
www.raymondhickey.com

Recordings done in public places have scripted samples from over 1500 speakers across Ireland

Anonymous, but with age, gender and location data captured

Cited in Google Scholar about 175 times, but this and the literature citing it is based on human listening, not data analysis modelling



Deep Learning and Speech Processing

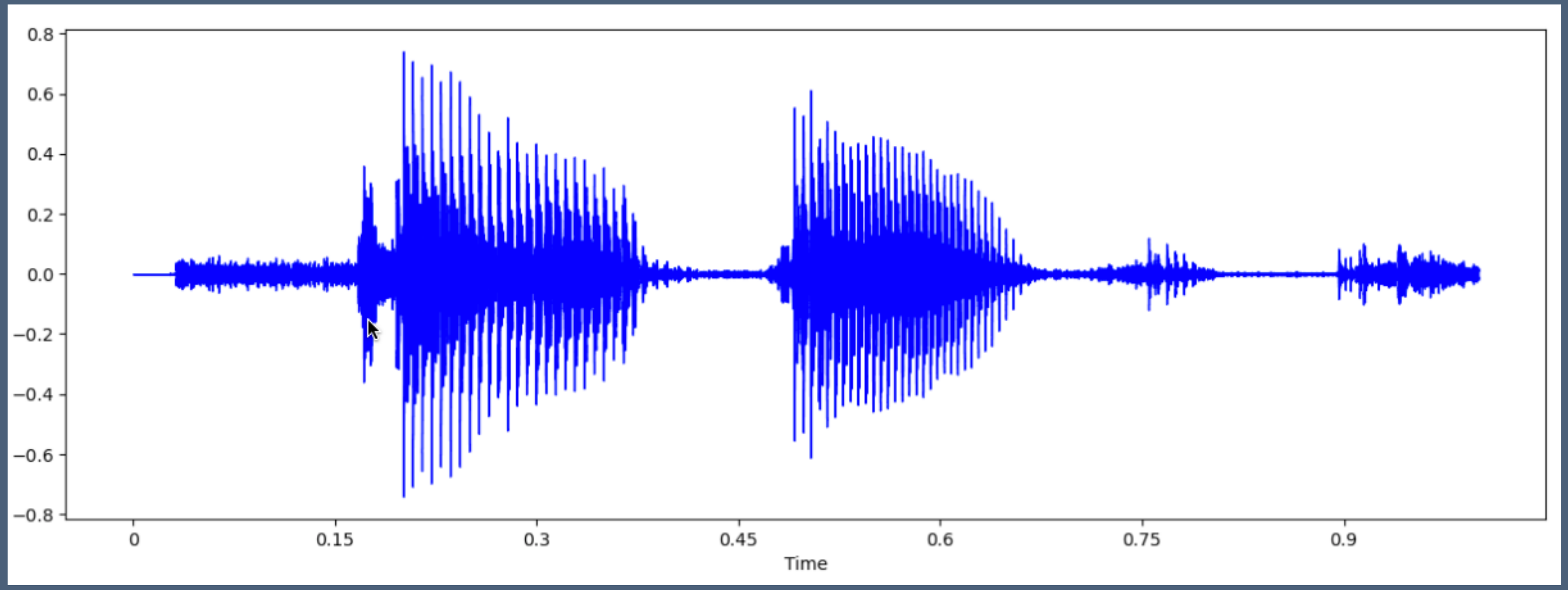
- Unsupervised learning for labelling data automatically, for automating feature detection: ‘Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups’ (2012) by Hinton, Deng, You, Abdel-Rahman et al’.

- Deep learning for speech processing: ’Building DNN acoustic models for large vocabulary speech recognition’ (2017) by Maas, Ng et al.

- Large Acoustic Models - CommonAccent, claims 93 to 97\% accuracy: (’Common Accent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on CommonVoice’ (2023) by Zuluaga-Gomez et al.

Input Variables

Sound recording is a simple x, y data series



Mel Frequency Cepstrum Coefficients (MFCC) can be calculated: The cepstrum, the inverse of the power of the wave in the sub-periods

Also, the age and gender are available in the SAIE

A Model to Answer the Research Question?

- What accent is detected on a speech sample?
- Target, or dependent, variable is a geographic location within Ireland
- Inputs may be - speech recordings data and features derived from it
- Personal characteristics - age, gender and other values may be available
- Candidate models will likely be classification models, trained by supervised learning
- Short and simple example from the literature: Sheng and Edmund use SVM, tree, neural networks and convolutional NN to distinguish among three Asian accents (https://cs229.stanford.edu/proj2017/final-reports/5244230.pdf)

Logistic Regression

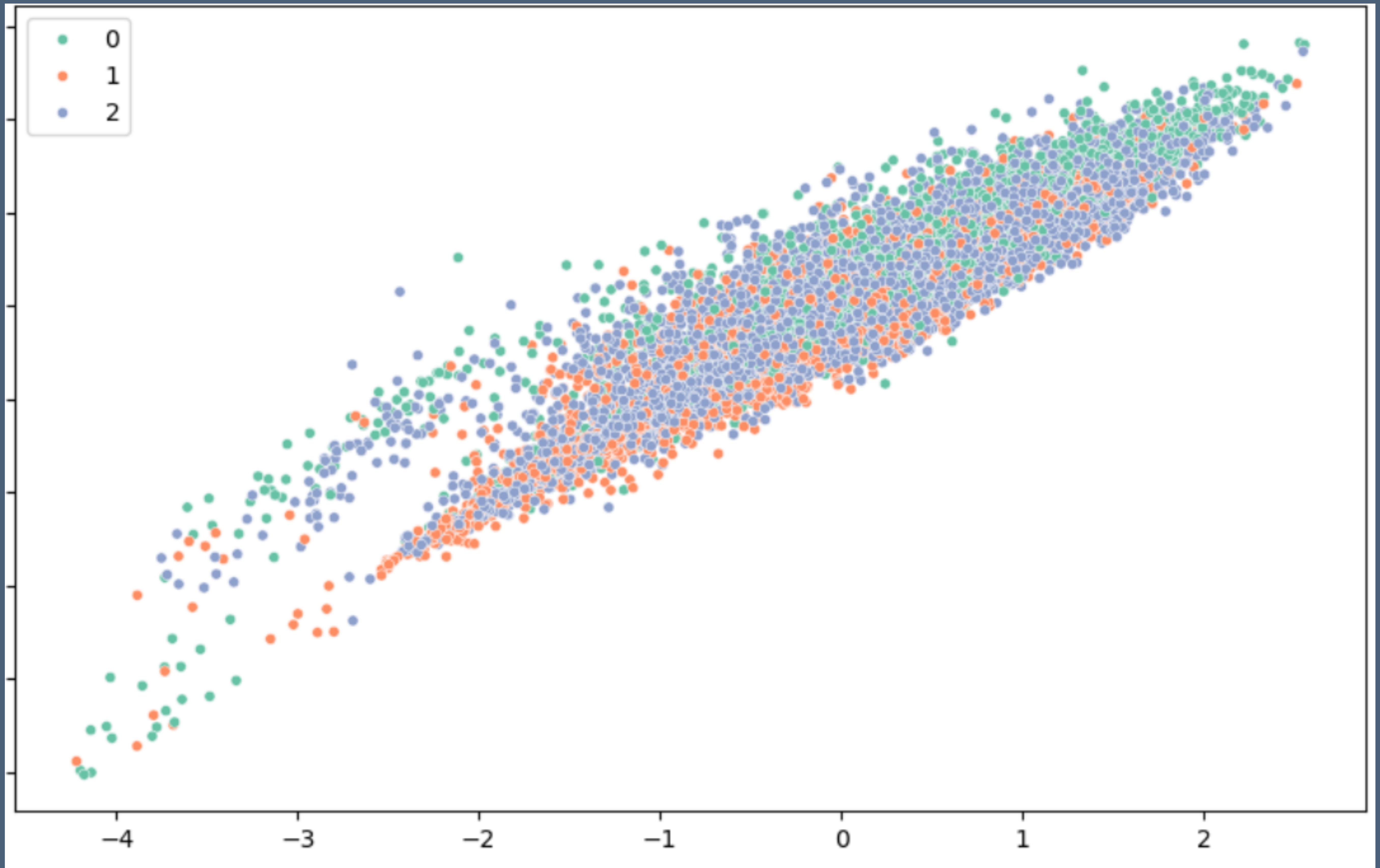
- Jupiter notebook clustering12.ipynb
- Accent classification Belfast or Dublin run as Y variables
- MFCC, age and gender as X variables
- Run with combinations of variables, add more as long as classification metrics improve
- MFCC with age and gender was best combination of variables ROC-AUC of 91.6%

Clustering

Same notebook as logistic regression

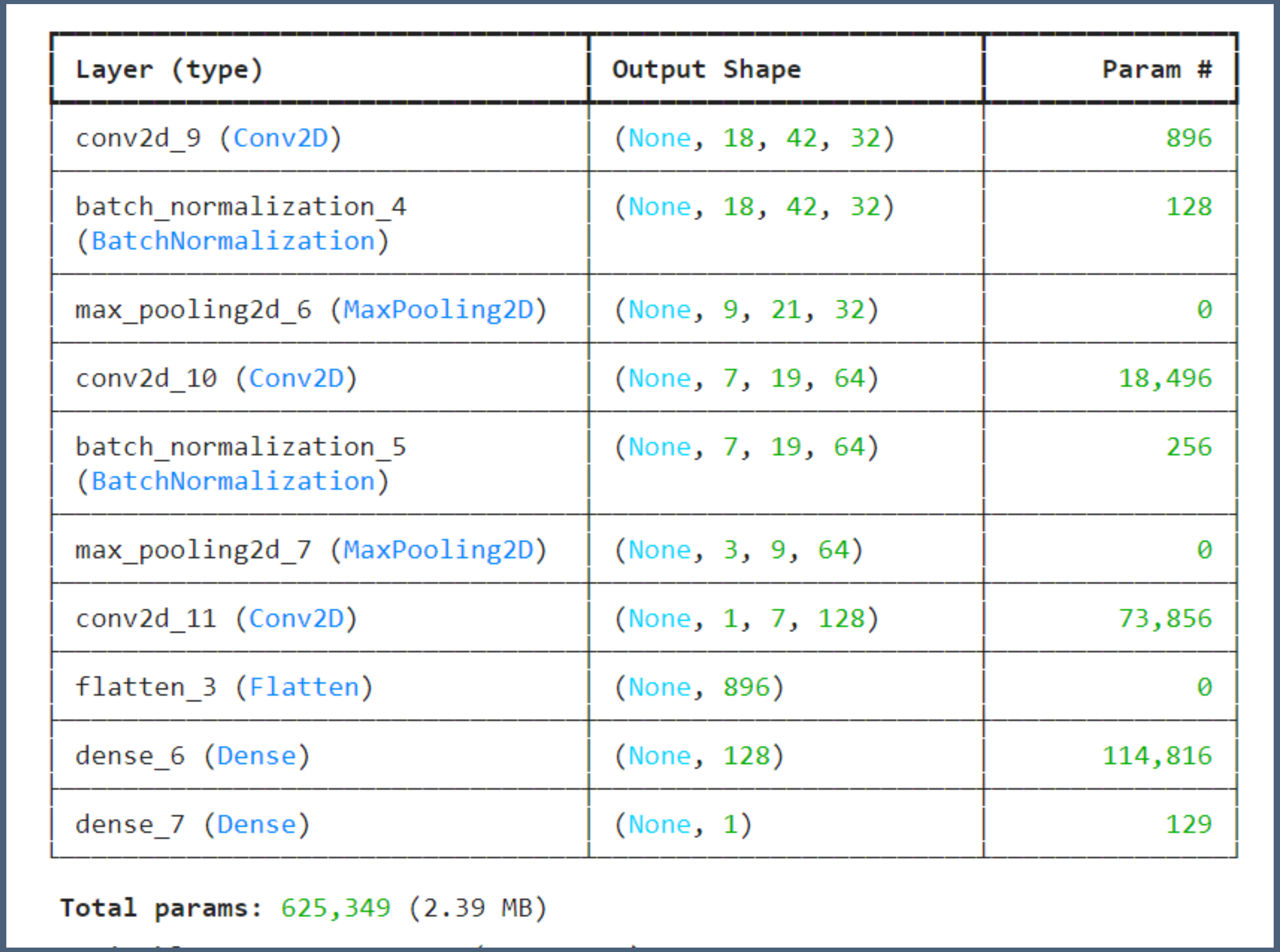
Clustering MFCC, gender and age with n=2 and n=3 k-means algorithms

No visible clustering visible on either case or with the same model with MFCC only



Convolutional Neural Network

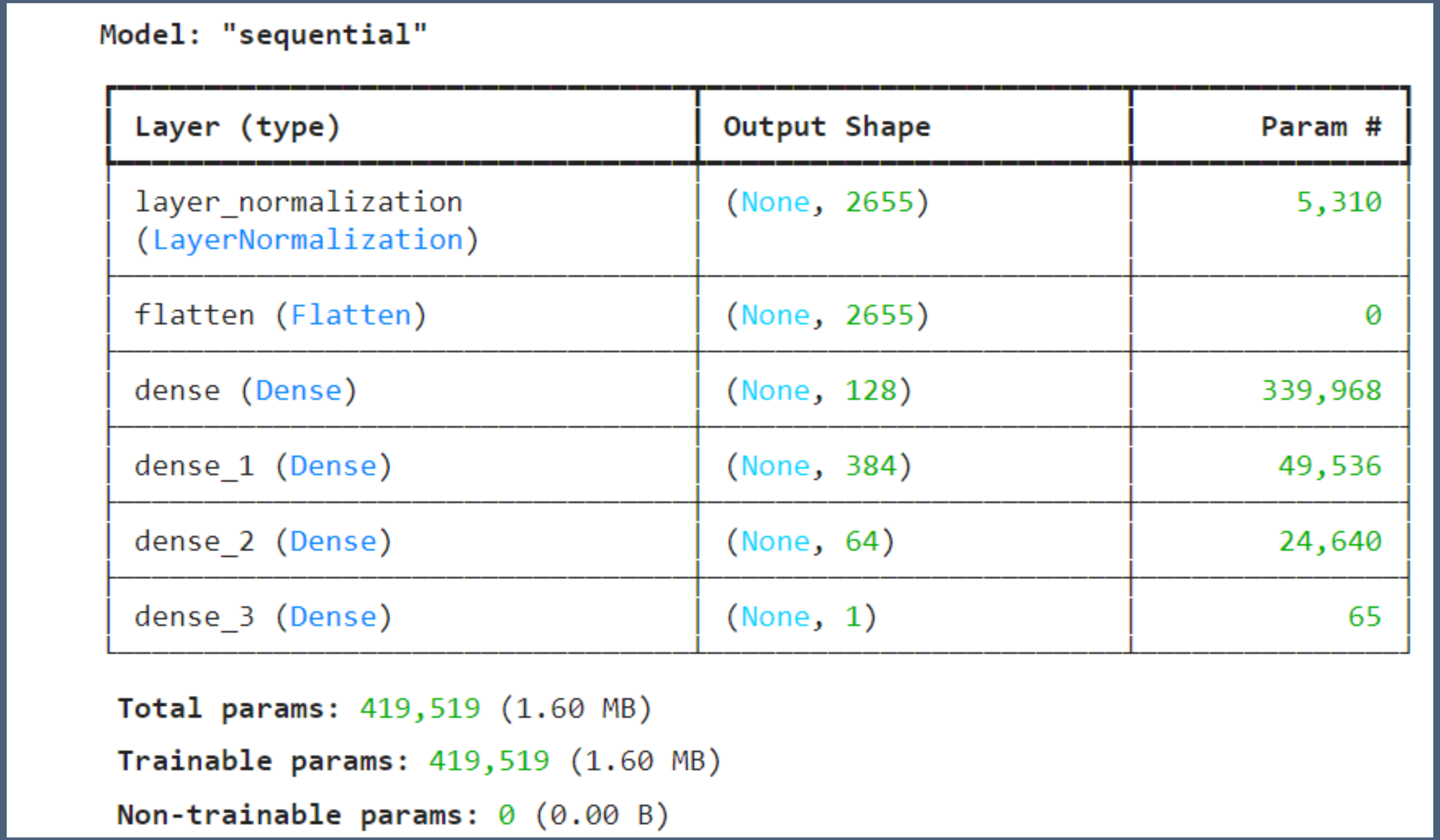
- Y is the region of the accent
- Keras Tuner for random selection of the architecture hyper parameters
- Target, or dependent, variable is a geographic accent location within Ireland
- MFCC and MFCC delta 1 and delta 2 first and second time differences were fed in as X variables
- MFCC speech data alone was achieving test accuracy in the low ninety percentages and ROC AUC of 96%



```
┃ Layer (type)                    ┃ Output Shape           ┃       Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩
│ conv2d_9 (Conv2D)               │ (None, 18, 42, 32)     │           896 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ batch_normalization_4           │ (None, 18, 42, 32)     │           128 │
│ (BatchNormalization)            │                        │               │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ max_pooling2d_6 (MaxPooling2D)  │ (None, 9, 21, 32)      │             0 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ conv2d_10 (Conv2D)              │ (None, 7, 19, 64)      │        18,496 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ batch_normalization_5           │ (None, 7, 19, 64)      │           256 │
│ (BatchNormalization)            │                        │               │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ max_pooling2d_7 (MaxPooling2D)  │ (None, 3, 9, 64)       │             0 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ conv2d_11 (Conv2D)              │ (None, 1, 7, 128)      │        73,856 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ flatten_3 (Flatten)             │ (None, 896)            │             0 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_6 (Dense)                 │ (None, 128)            │       114,816 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_7 (Dense)                 │ (None, 1)              │           129 │
└─────────────────────────────────┴────────────────────────┴───────────────┘
 Total params: 625,349 (2.39 MB)
```

Neural Network MLP

- Y is the region of the accent
- Manual selection of the architecture
- Target, or dependent, variable is a geographic accent location within Ireland
- MFCC and MFCC delta 1 and delta 2 first and second time differences were fed in as X variables
- Age and gender were added as one-hot encoded inputs also
- Test accuracy 82-89% are typical and ROC AUC of 92%



```
Model: "sequential"
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━┓
┃ Layer (type)                    ┃ Output Shape           ┃       Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩
│ layer_normalization             │ (None, 2655)           │         5,310 │
│ (LayerNormalization)            │                        │               │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ flatten (Flatten)               │ (None, 2655)           │             0 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense (Dense)                   │ (None, 128)            │       339,968 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_1 (Dense)                 │ (None, 384)            │        49,536 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_2 (Dense)                 │ (None, 64)             │        24,640 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_3 (Dense)                 │ (None, 1)              │            65 │
└─────────────────────────────────┴────────────────────────┴───────────────┘
 Total params: 419,519 (1.60 MB)
 Trainable params: 419,519 (1.60 MB)
 Non-trainable params: 0 (0.00 B)
```

Conclusions and Future Research

- Speech data and personal characteristics from SAIE with simple logistic regression give very good classification performance
- Relatively simple MLP and CNN in turn improved on those results
- Wide range of more evolved deep-learning models to apply in future research
- More data can now be gathered cheaply and widely using online crowd-sourced efforts