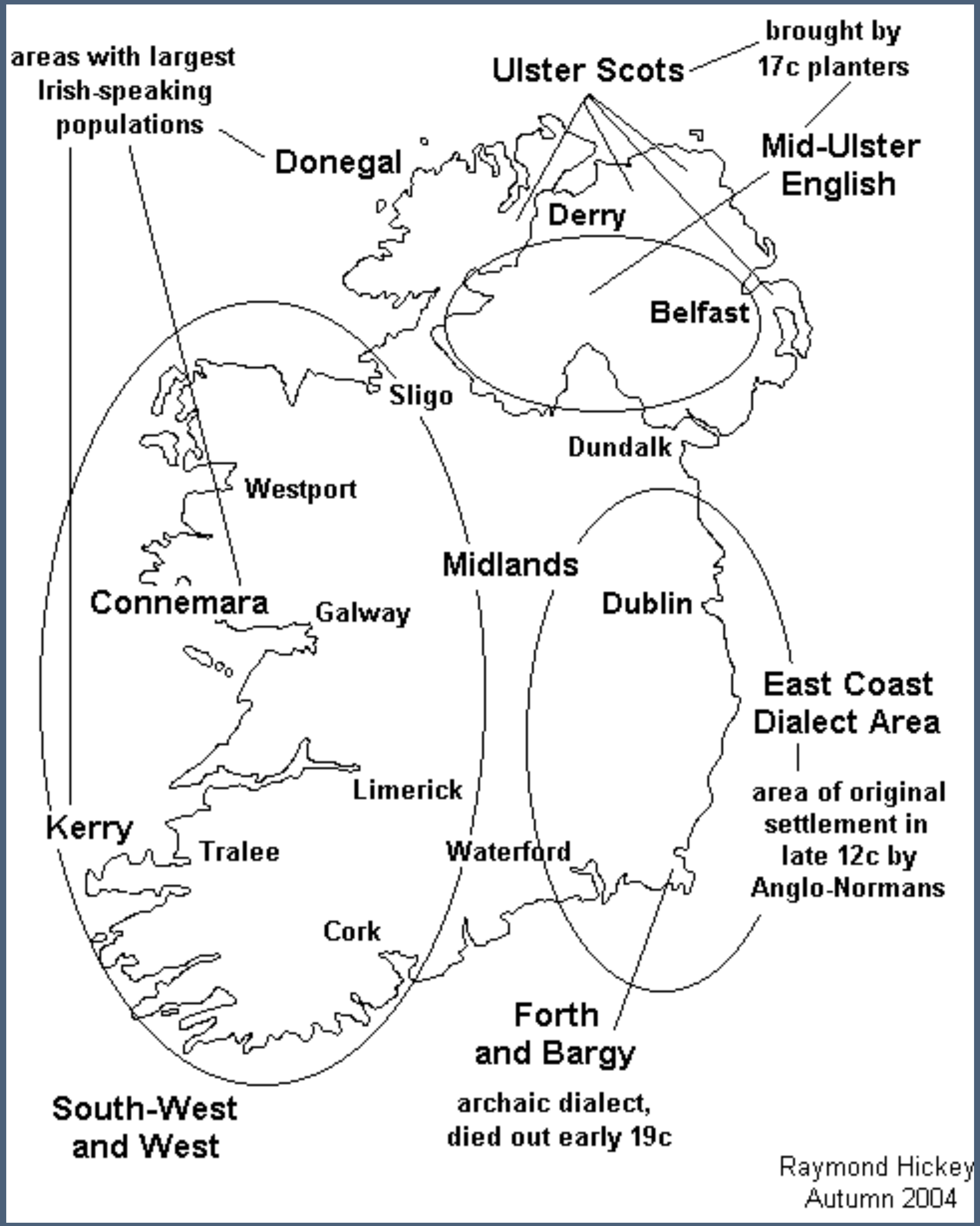
How the Irish speak English: Python Machine Learning for Classification of English-language Accents in Ireland

Peter Nolan
databeaker@gmail.com
https://github.com/dpnolan/voxpop

PyCon Ireland, Saturday 16 November 2024



About Me

- Born and raised… where?

databeaker@gmail.com

About Me

- Born and raised in south county Dublin
- Not quite a data scientist, and certainly I'm no linguist either, but the Python ecosystem allowed me to do this project
- Background: TCD B.B.S. Business
- I wrote a neural network model *by hand* in C to predict stock indices and FX rates for my thesis. Although we observed predictive power as measured in the then-current computer science literature, the models fell apart under the cost assumptions and statistical tests from finance.   One major trading house I worked at had switched off their neural net in early 1996
- London City University / Bayes Business School - Mathematical Finance M.Sc.
- National College of Ireland M.Sc. Data Analysis this year
- Python provides a free and community-supported workshop of tools, both general (NumPy, Pandas, TensorFlow) and specialised (librosa) that match those used in the research literature, enabling projects such as this

databeaker@gmail.com

How the Irish speak English: Summary

- Irish people speak English with distinctive sounds in speech: What do these sounds tell us about social identity?

- From existing academic and open-source datasets of speech samples of Irish people speaking English ("Irish English" linguists call this) from locations all over both the Republic of Ireland and Northern Ireland, I have selected one that I hope is large enough (1500+ speakers, 266 in my target classes) and with enough geographic variation for computational data analysis, namely the Sound Atlas of Irish English by Prof. Raymond Hickey (2004).

- Linguistic research in Ireland has worked with recordings since before 1900, but research patterns manually, the researcher listens to the recordings to detect patterns, which can be slow and subjective. Automation through computational data analysis is promising.

- My research project is the first attempt at machine learning using the SAIE dataset out of 179 previous citations, and the first for any machine learning for regional variation in Ireland.

- I build models to classify voice samples by location using the speaker's age, gender and speech features. Using logistic regression, random forests, MLP and CNN neural networks, I achieve significant performance (accuracy > 85%, AUC > 90%) in classification, compared to claimed state of the art of 90-95% in the research literature.

- Based on this, I propose a new Sound Atlas of Irish English, an all-island crowd-sourced effort to gather a large, open-source, up to date dataset to build public understanding, for linguistics research and for engineering better automated speech processing

What is an accent?

- In common usage, an accent is a ‘way of pronouncing a language that is distinctive to a country, area, social class, or individual’, according to the Oxford English Dictionary
- Social hierachy can enforce accent conformity, as with 'BBC English'.
- Dialect is understood differently by linguists as a distinctive grammar and vocabulary
- Accent instead refers to the identifiable differences in pronunciation of the words and sentences among a group (Hughes et al. 2012, chapter 1).
- Accent model is the classification not on matching the individual, instead capturing characteristics of the whole group through the estimation process
- Distinctive patterns remain beyond individual variations, which can be observed in the stops and starts, the tones used, letters pronounced or the words and grammar.
- Data leakage is a significant risk, the training and test datasets need to have different speakers, otherwise you are identifying each the speaker
- We detect accents manually all the time, but can automated speech processing?
- What does this tell us about the speaker's identity, individual or group characteristics e.g. personal identification, gender, age, income, geographic origin, national origin, religious?

Motivation

- Why study accent?

- Accent is DATA, and automated speech processing will perform badly or encode discrimination if done poorly or without accountability.

  Speech processing services from Amazon and Google perform worse on Northern Irish accents than any other UK or Dublin data (Markl 2022). Two Stanford students from Singapore report inferior performance for that English accent also (Sheng & Edmund 2017).

- Accent is IDENTITY

  It impacts how we understand ourselves and our changing society and how others treat us, sometimes maliciously, as shown in Northern Ireland and Yugoslavia.

Historical Background

How the Irish speak English?

- Almost all people in Ireland, north and south, have English as their first language at home, school and work since 1900, but there are wide variations in vocabulary and pronunciation
- Influence of Irish is felt, which was progressively abandoned in a pattern beginning in the east, then moving west
- Historical migration patterns of Normans, Welsh, Scots-Ulster, English and trade flows were important, in Dublin especially
- Movements after Cromwell, the Famine, then urbanisation
- Media - broadcast and now online - show different pronunciations, including English and American
- Distinctive patterns remain beyond individual variations? What does this tell us? How predictable is this?

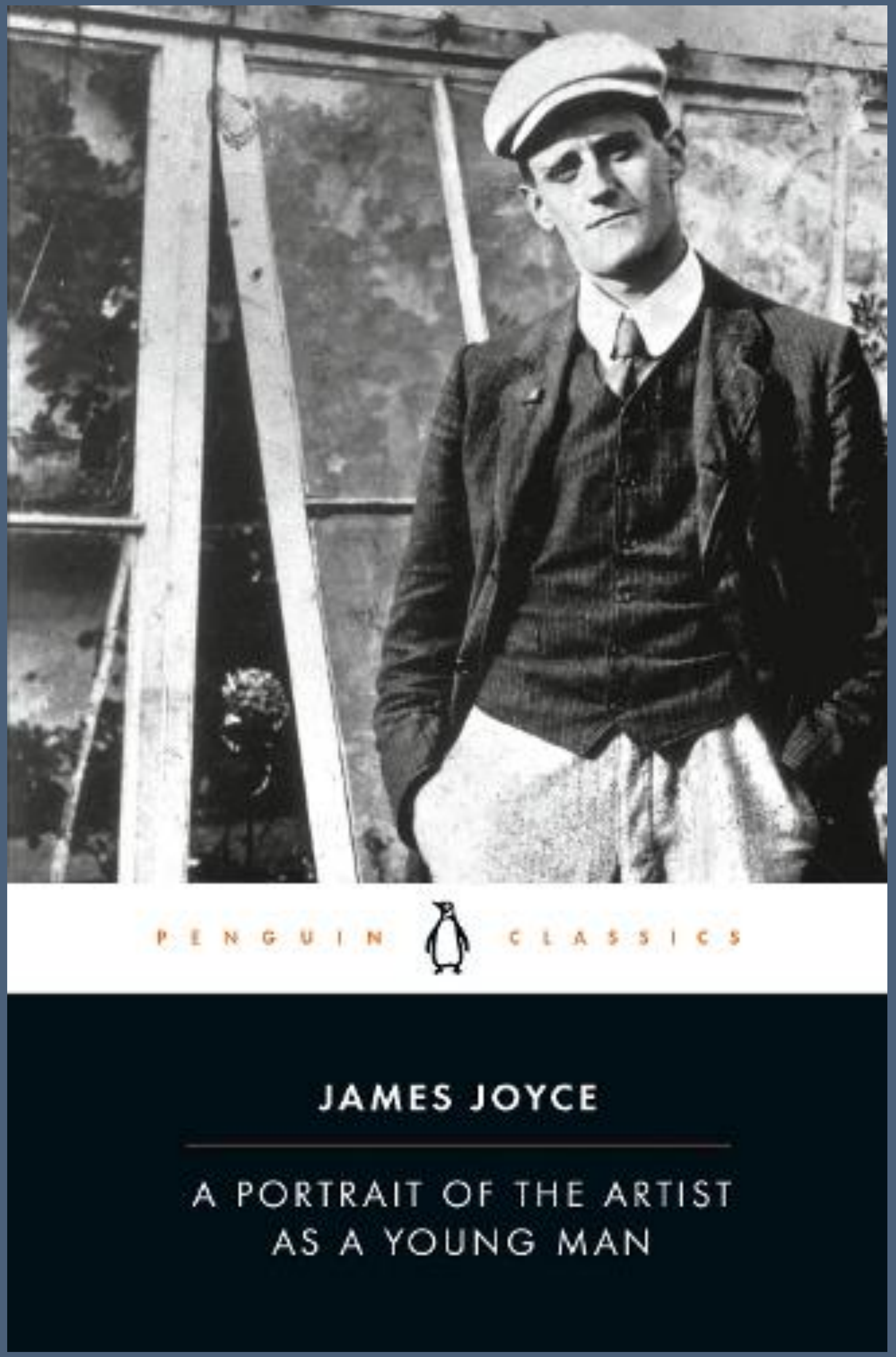
Irish Culture Focuses on Irish Speech

Literature, Comedy and Satire

- Joyce - especially in Portrait of the Artist as a Young Man, when the narrator argued with a schoolmaster, reproduced the voices of the Irish as a product of their conflict over identity

- "How different are the words home, Christ, ale, master, on his lips and on mine!"

- The Irish speech pattern is a linguistic way of subverting a political conquest, wrote the critic Seamus Deane



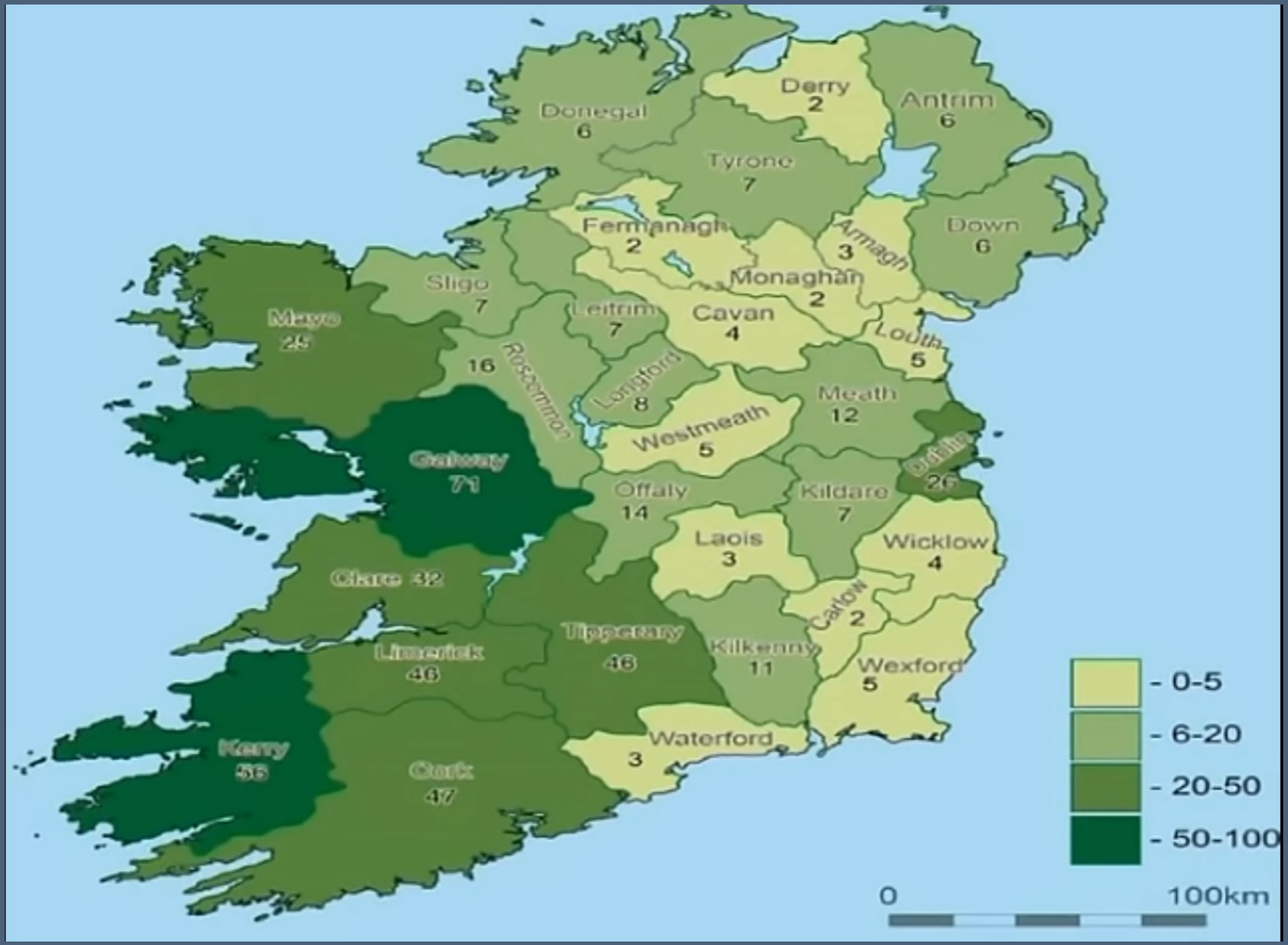
Irish Culture Focuses on Irish Speech

Literature, Comedy and Satire

Niall Tóibín: A Guide to the Regional Accents of Ireland

https://www.youtube.com/watch?v=EhLdKJnY194

Identifies and demonstrates the most common accents



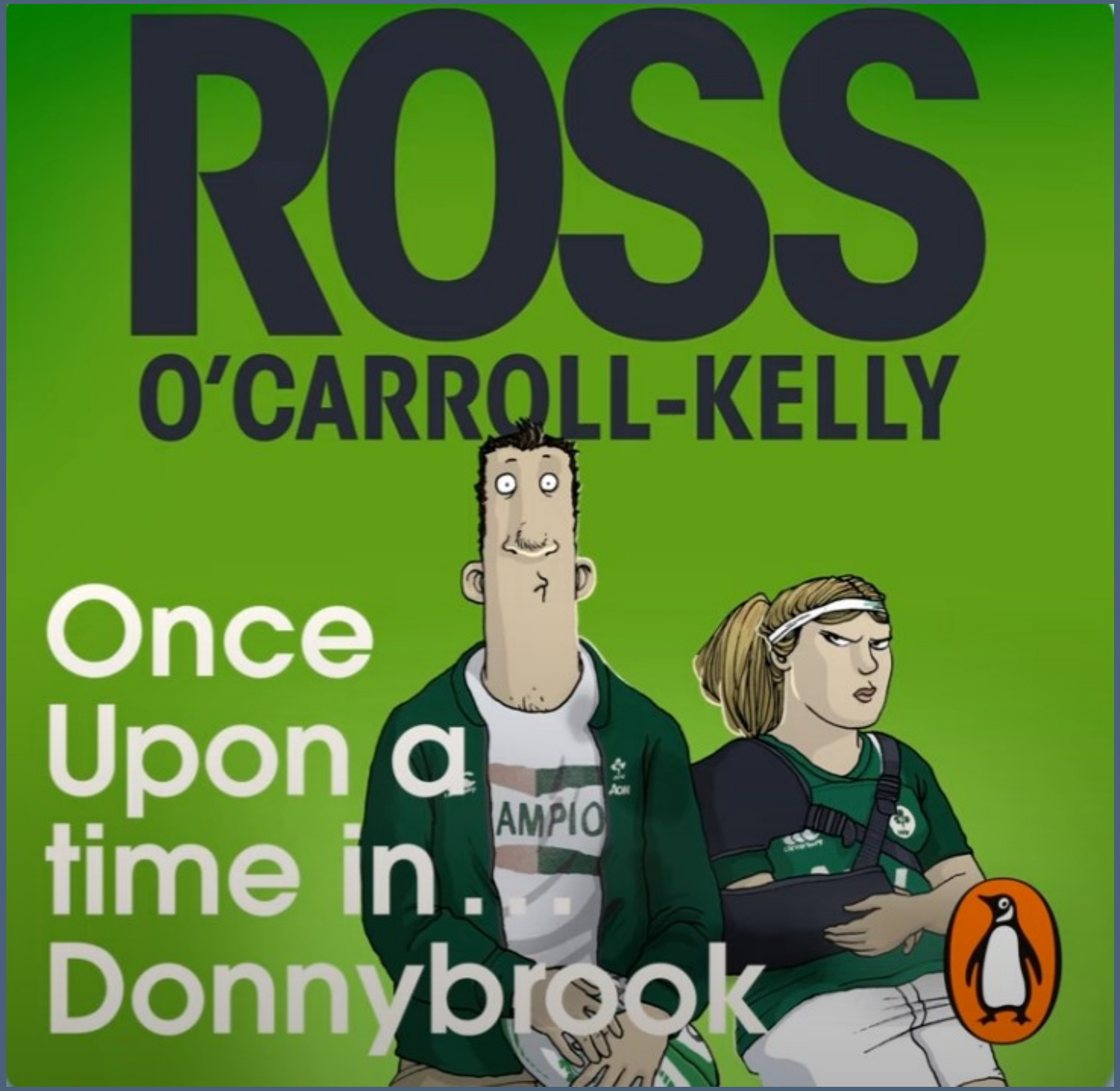
Irish Culture Focuses on Irish Speech

Literature, Comedy and Satire

Ross O'Carroll-Kelly, a stereotypical affluent south Dublin rugby-player, has given a name to the `DORT' accent, named after the 'DART' commuter rail

From just one of his audiobooks over the past 25 years as a bestseller:

https://youtu.be/vlobTz-9_Fc?t=40



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? Leads to three questions in sequence.

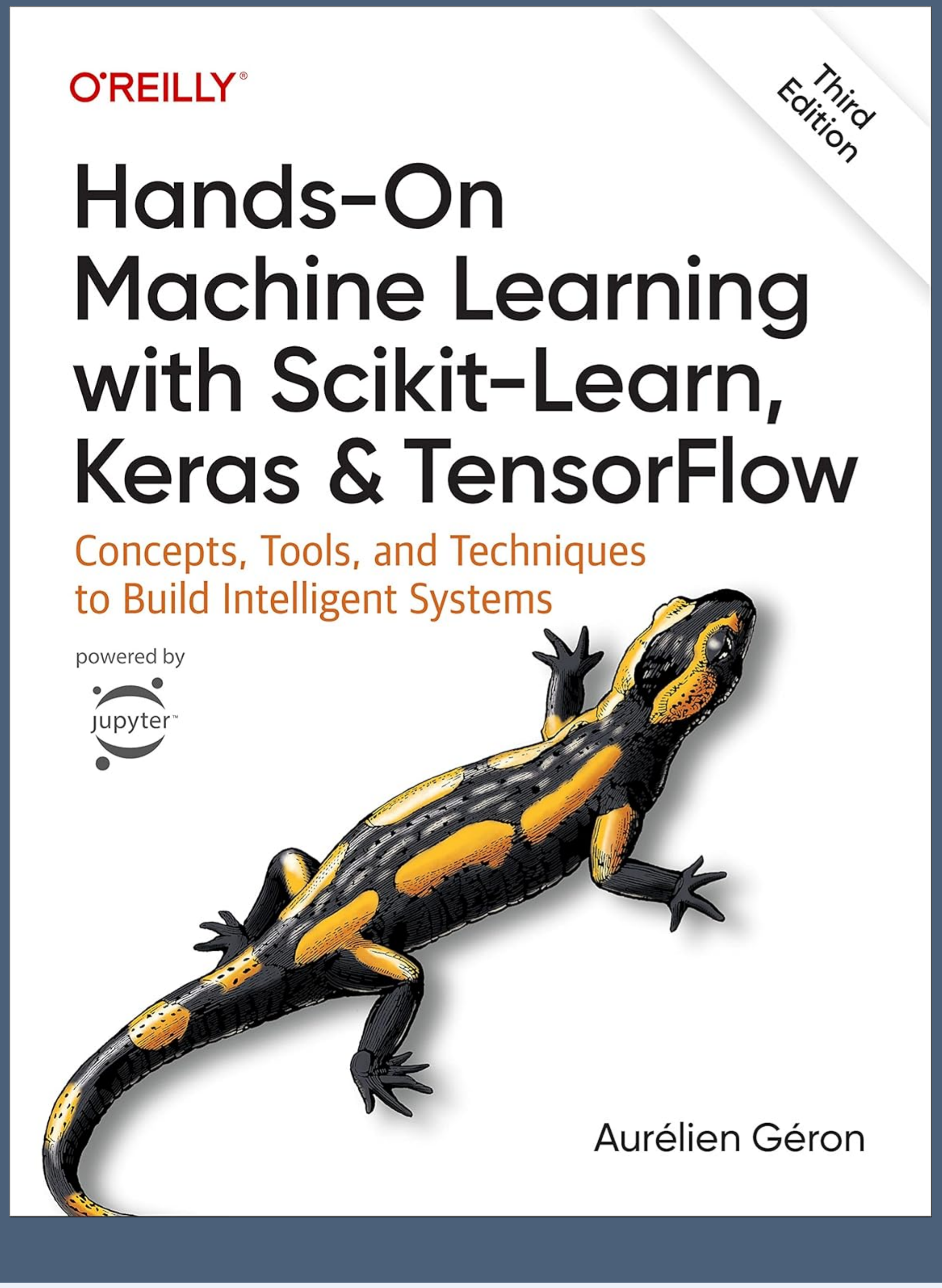
- What datasets can support this computational data analysis?

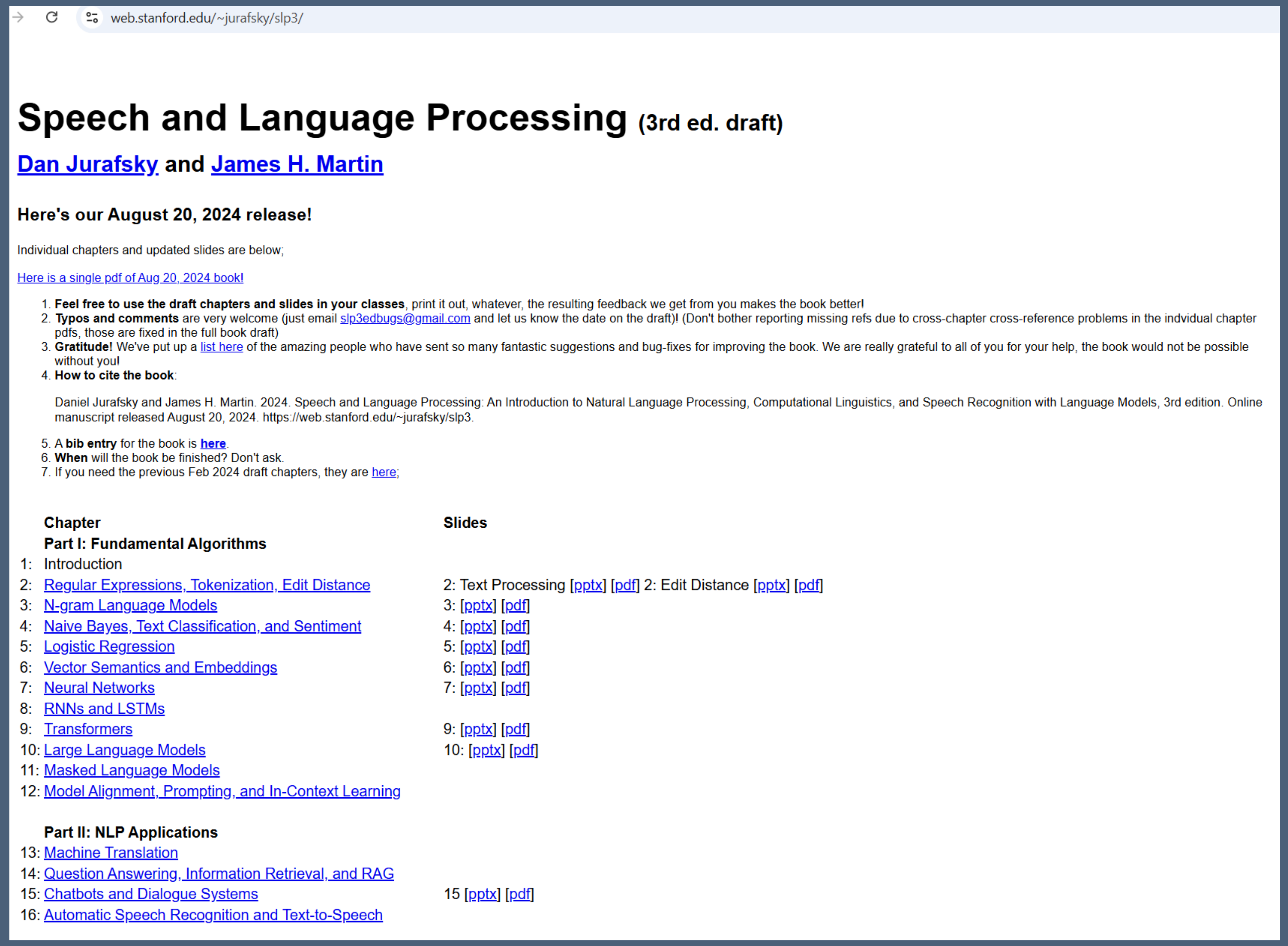
- What features, which may be either the sound recording data or features calculated from it, or personal characteristics, will be useful inputs to the classification model predicting which region the model comes from?

- What data analysis models will perform this classification well?

Literature Review

- Sheng & Edmund in their 2017 Stanford thesis, which is the highest ranked result from Google searches for 'machine learning' 'accent classification'. They build classification models with English language speech data alone with different accents and getting increasingly good performance as they move from using decision trees, then multi-layer perceptrons and then convolutional neural networks for accuracy over 85 to classify samples into three Asian accents in spoken English https://cs229.stanford.edu/proj2017/final-reports/5244230.pdf
- Until around 2006, automated speech processing was largely separate from neural network and deep-learning research domain, Hidden Markov Models dominated.
- Deep learning models began outperforming in speech processing once more layers, different architectures and improved optimisation algorithms became available e.g. Maas et al. (2017) and (Lesnichaia et al. 2022) showed that more hidden layers, random initialisation of network weights, newer optimisers and Dropout pruning all have improved deep learning classification performance on accent classification.
- After 2012, based on Hinton's and Bengio's work on unsupervised learning using layers of autoencoders to label samples, architectures increasingly use the sound recording directly as input.
- A recent INTERSPEECH paper Zuluaga-Gomez et al. (2023) which uses variants of the wav2vec2 models using wavenet, transformer and masking models, cross-training on non-English language data produces the highest classification accuracy of any of these examples at over 97%.
- Geron (2022) surveys the evolution of deep learning including audio processing; Jurafsky and Martin 'Speech and Language Processing' 3e (forthcoming) is free online.


O'REILLY
Third Edition
Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron



Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

Here's our August 20, 2024 release!

Individual chapters and updated slides are below;

Here is a single pdf of Aug 20, 2024 book!

1. **Feel free to use the draft chapters and slides in your classes**, print it out, whatever, the resulting feedback we get from you makes the book better!
2. **Typos and comments** are very welcome (just email slp3edbugs@gmail.com and let us know the date on the draft)**!** (Don't bother reporting missing refs due to cross-chapter cross-reference problems in the indvidual chapter pdfs, those are fixed in the full book draft)
3. **Gratitude!** We've put up a list here of the amazing people who have sent so many fantastic suggestions and bug-fixes for improving the book. We are really grateful to all of you for your help, the book would not be possible without you!
4. **How to cite the book**:

   Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. https://web.stanford.edu/~jurafsky/slp3.

5. A **bib entry** for the book is **here**.
6. **When** will the book be finished? Don't ask.
7. If you need the previous Feb 2024 draft chapters, they are here;

	Chapter	Slides
	Part I: Fundamental Algorithms	
1:	Introduction	
2:	Regular Expressions, Tokenization, Edit Distance	2: Text Processing [pptx] [pdf] 2: Edit Distance [pptx] [pdf]
3:	N-gram Language Models	3: [pptx] [pdf]
4:	Naive Bayes, Text Classification, and Sentiment	4: [pptx] [pdf]
5:	Logistic Regression	5: [pptx] [pdf]
6:	Vector Semantics and Embeddings	6: [pptx] [pdf]
7:	Neural Networks	7: [pptx] [pdf]
8:	RNNs and LSTMs	
9:	Transformers	9: [pptx] [pdf]
10:	Large Language Models	10: [pptx] [pdf]
11:	Masked Language Models	
12:	Model Alignment, Prompting, and In-Context Learning	
	Part II: NLP Applications	
13:	Machine Translation	
14:	Question Answering, Information Retrieval, and RAG	
15:	Chatbots and Dialogue Systems	15 [pptx] [pdf]
16:	Automatic Speech Recognition and Text-to-Speech	

Candidate Datasets of Irish English

- Speech Accent Archive (SAA) - Commonly-used in the literature but has only 17 Ireland or Northern Ireland samples (https://accent.gmu.edu/)

- International Corpus of English (http://ice-corpora.net/ice/) data was not available and the admins were unresponsive

- Oxford Linguistics Lab - Intonational Variation in English (IViE) has Dublin and Belfast samples, but just 6 speakers recorded in each location (http://www.phon.ox.ac.uk/files/apps/IViE/)

- Mozilla CommonVoice - crowd-sourced data including thousands of samples, labelled by gender, age and country, including from Ireland, but no more detailed geodata (https://commonvoice.mozilla.org/en)

Sound Atlas of Irish English ('SAIE')

Book and dataset by Raymond Hickey (2004)
Prof. Linguistics at Uni. Limerick and Uni. Essen - Duisberg
www.raymondhickey.com

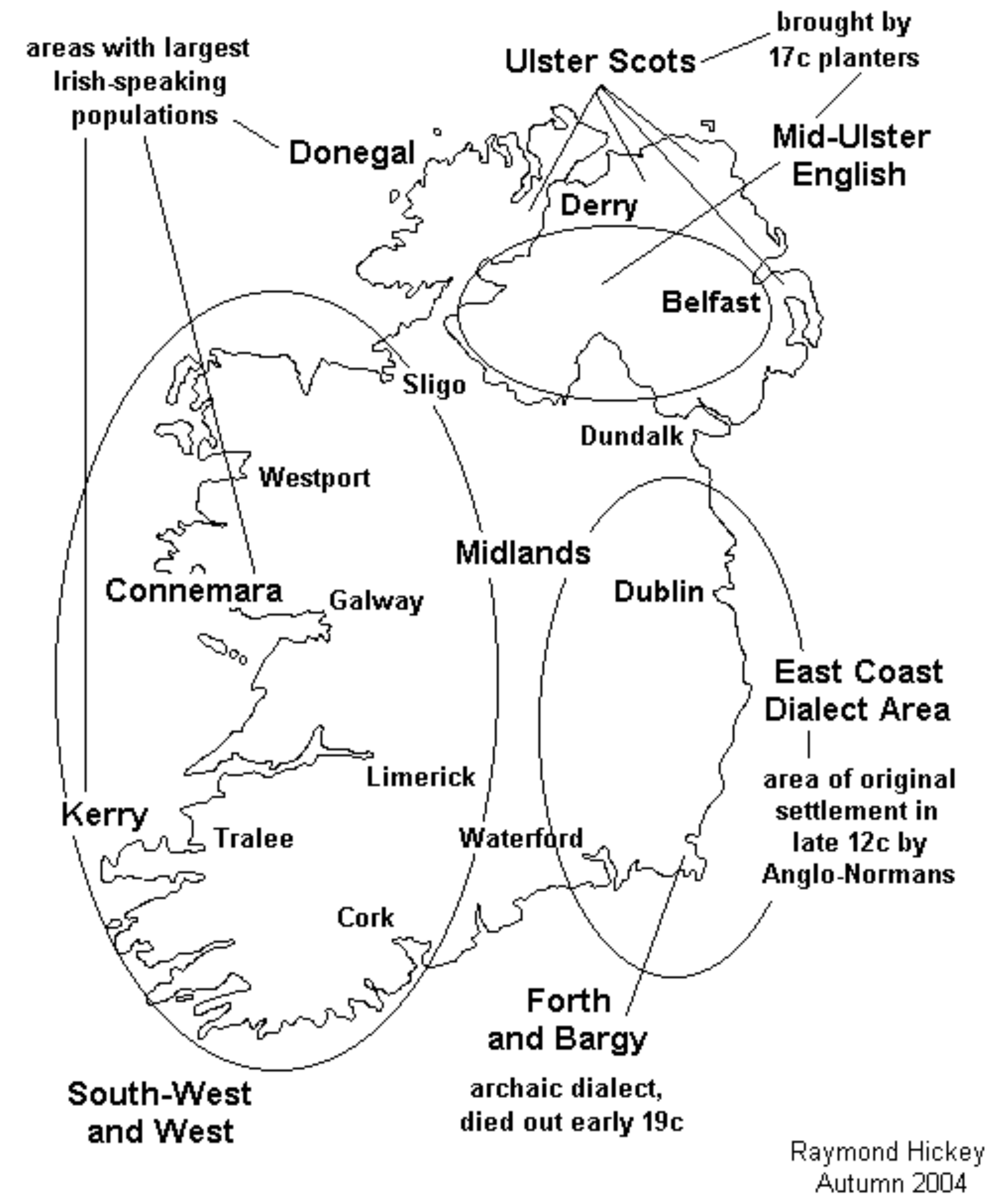
Recordings done in public places have scripted samples repeated from over 1,500 speakers

Date range from 1980 onwards

Come from all 32 counties, rural and urban areas, speakers across Ireland

Anonymously gathered, but with age, gender and location data captured

Cited in Google Scholar about 179 times as of today, but the book and the research citing it is all based on human ear to listen, detect features and classification, not any computational data analysis



A Model to Answer the Research Question

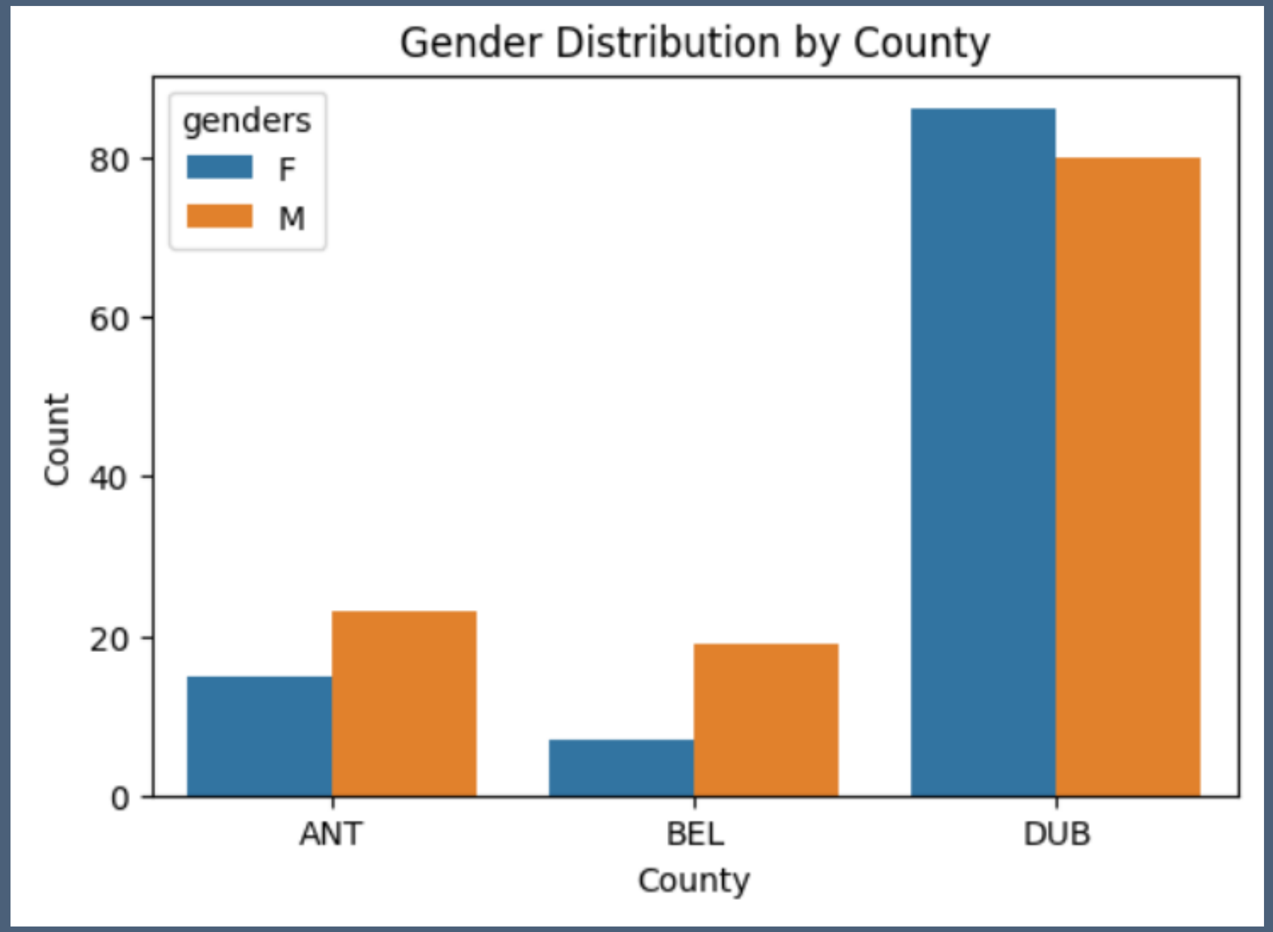
- What accent is detected on a speech sample?
- y, the target, or dependent, variable is a geographic location within Ireland
- Inputs may be speech recording data and features derived from it
- Personal characteristics - age, gender and other values may be available
- Candidate models will be classification with different models and different choices of variables,
- Largely will be a supervised learning task
- I follow much the same methodology as Sheng and Edmund 2017 here in using multiple models

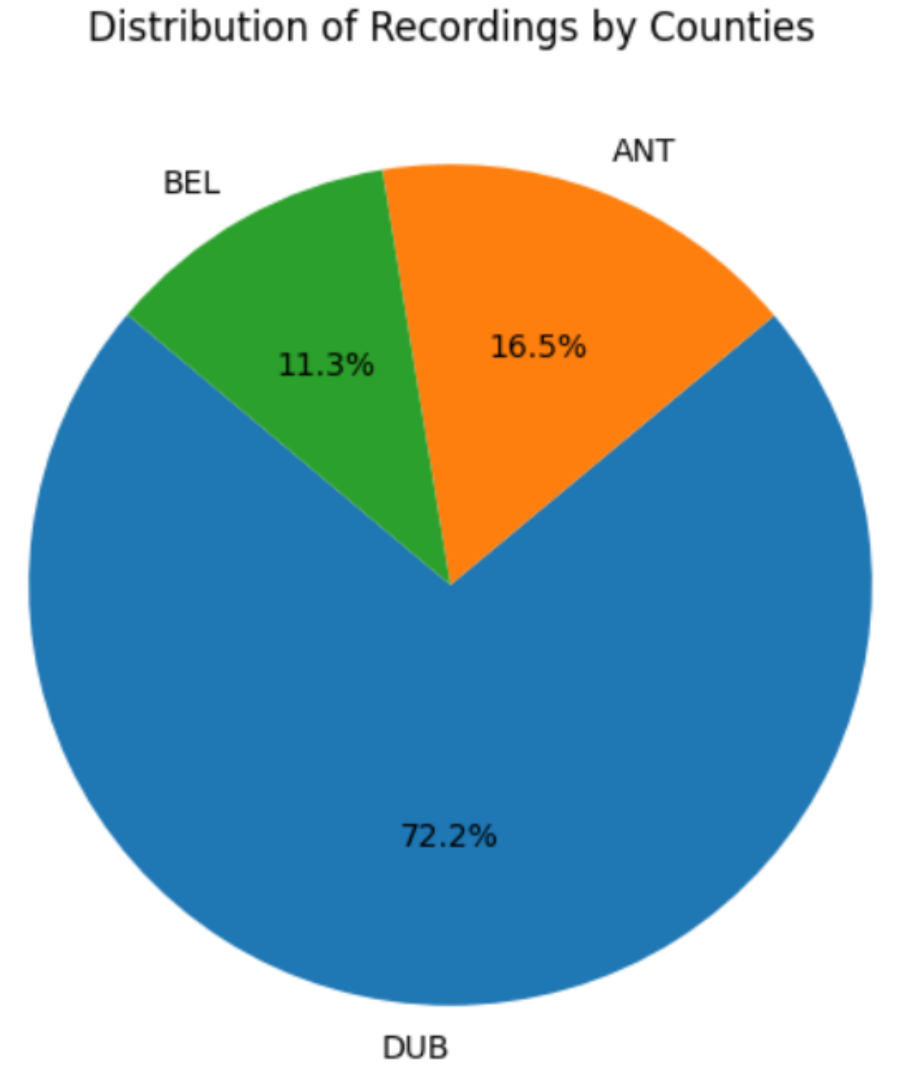
Models from Python Libraries

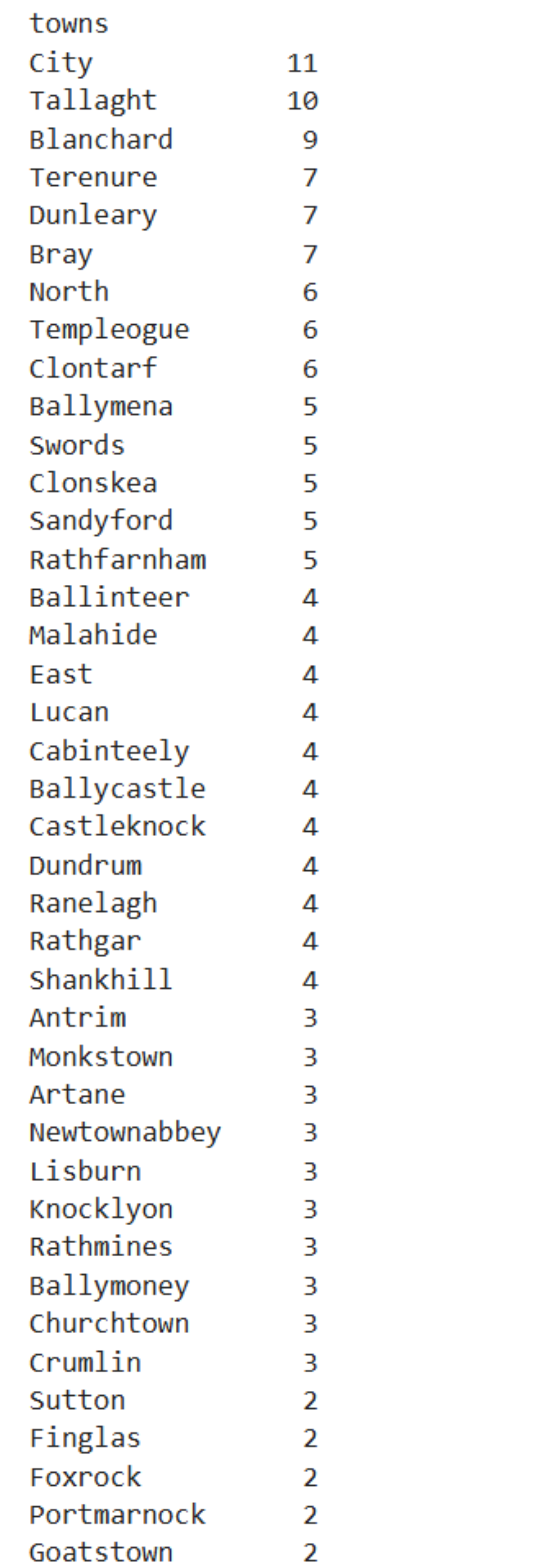
- Speech processing and feature extraction and graphing from librosa
- Logistic regression - using scikit-learn
- Random forests - using scikit-learn
- Neural network - MLP using Tensor Flow Keras
- Convolutional Neural Network - Using Tensor Flow Keras
- Classification metrics scikit-learn

Exploratory Data Analysis

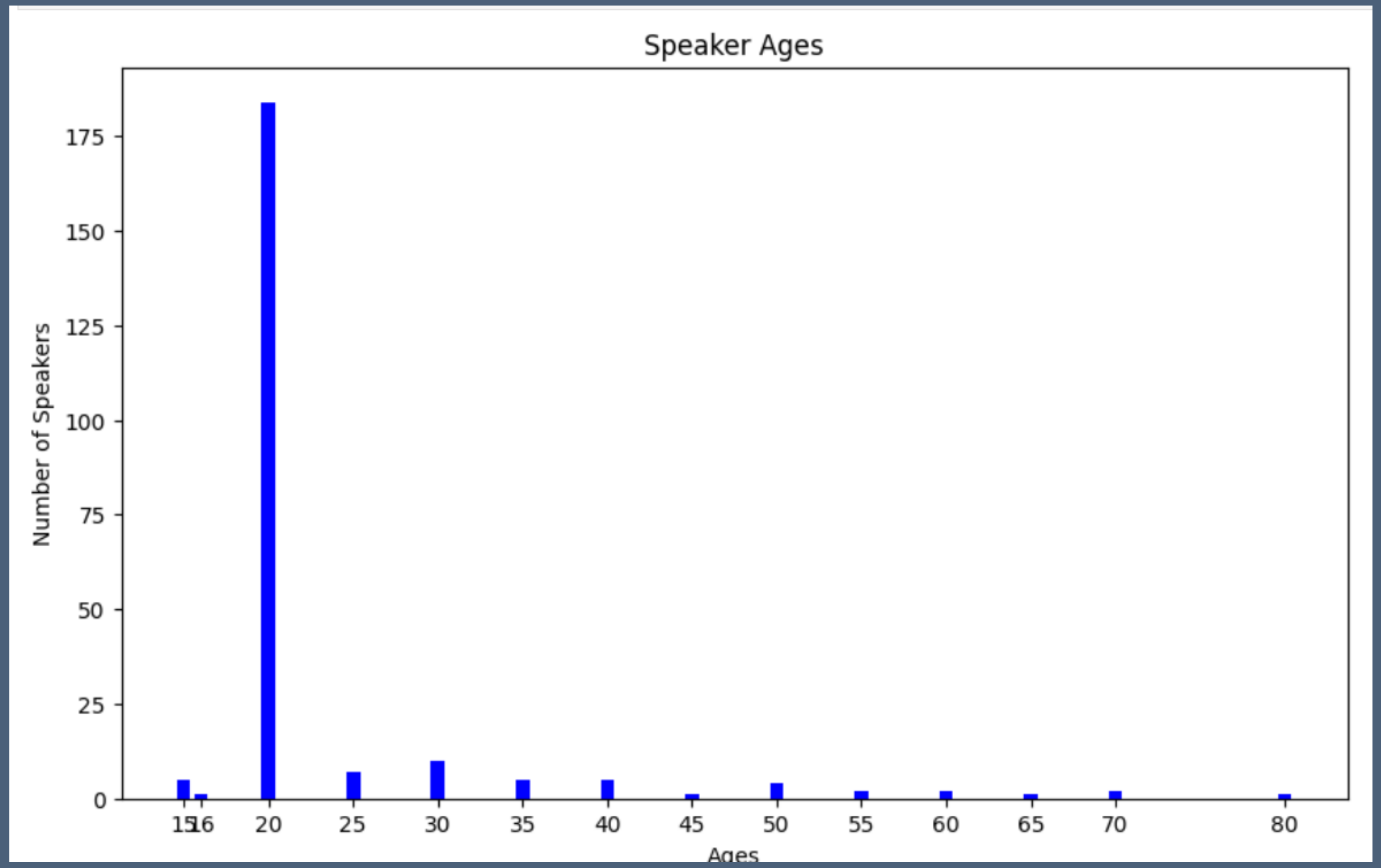
- Imbalanced dataset by County
- 122 male and 108 female speakers imbalanced in ANT+BEL
- Most speakers 'around' 20 years of age
- Speakers are not very concentrated in particular districts of Dublin
- Biggest category in Belfast dataset is 'city', while no category for the west, the traditional centre for Irish Catholic / Nationalist / Republican communities







```
towns
City            11
Tallaght        10
Blanchard        9
Terenure         7
Dunleary         7
Bray             7
North            6
Templeogue       6
Clontarf         6
Ballymena        5
Swords           5
Clonskea         5
Sandyford        5
Rathfarnham      5
Ballinteer       4
Malahide         4
East             4
Lucan            4
Cabinteely       4
Ballycastle      4
Castleknock      4
Dundrum          4
Ranelagh         4
Rathgar          4
Shankhill        4
Antrim           3
Monkstown        3
Artane           3
Newtownabbey     3
Lisburn          3
Knocklyon        3
Rathmines        3
Ballymoney       3
Churchtown       3
Crumlin          3
Sutton           2
Finglas          2
Foxrock          2
Portmarnock      2
Goatstown        2
```



Preprocessing - Extract Data and Features

- .wav sound files as the source data from Hickey, the book having a CD with sound files, visualisation files and Java code, but nothing like the open-source libraries that Python has

- Each file has one speaker repeating scripted standard sentences

- librosa library used for this, the most common in the research literature

- Jupyter notebook per county dataset, here Antrim/Belfast (ANT) and Dublin city and county (DUB)

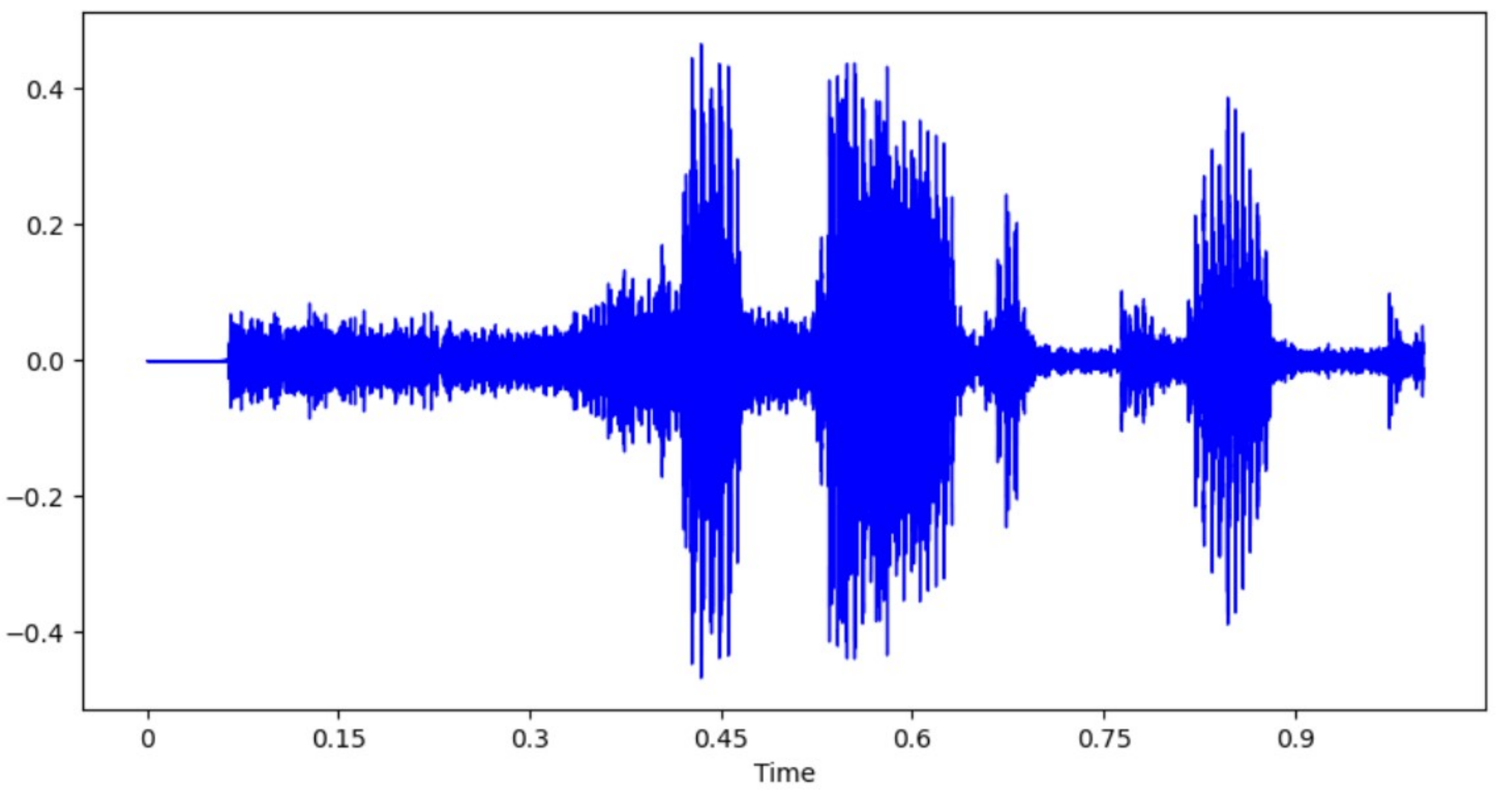
- Detect the length of the sound file in seconds

- Check if the file is wav, as some are mislabelled and are really .mp3 and if so, convert to .wav

- Filenames encode the metadata e.g.DUB_Artane_M_20_(2).wav here, shows county Dublin ('DUB' or 'ANT' or 'BEL'), unstructured town or suburb name, gender code ('M' or 'F'), age range code and a count

- The metadata for each file is extracted to a Pandas DataFrame along with the raw sound data

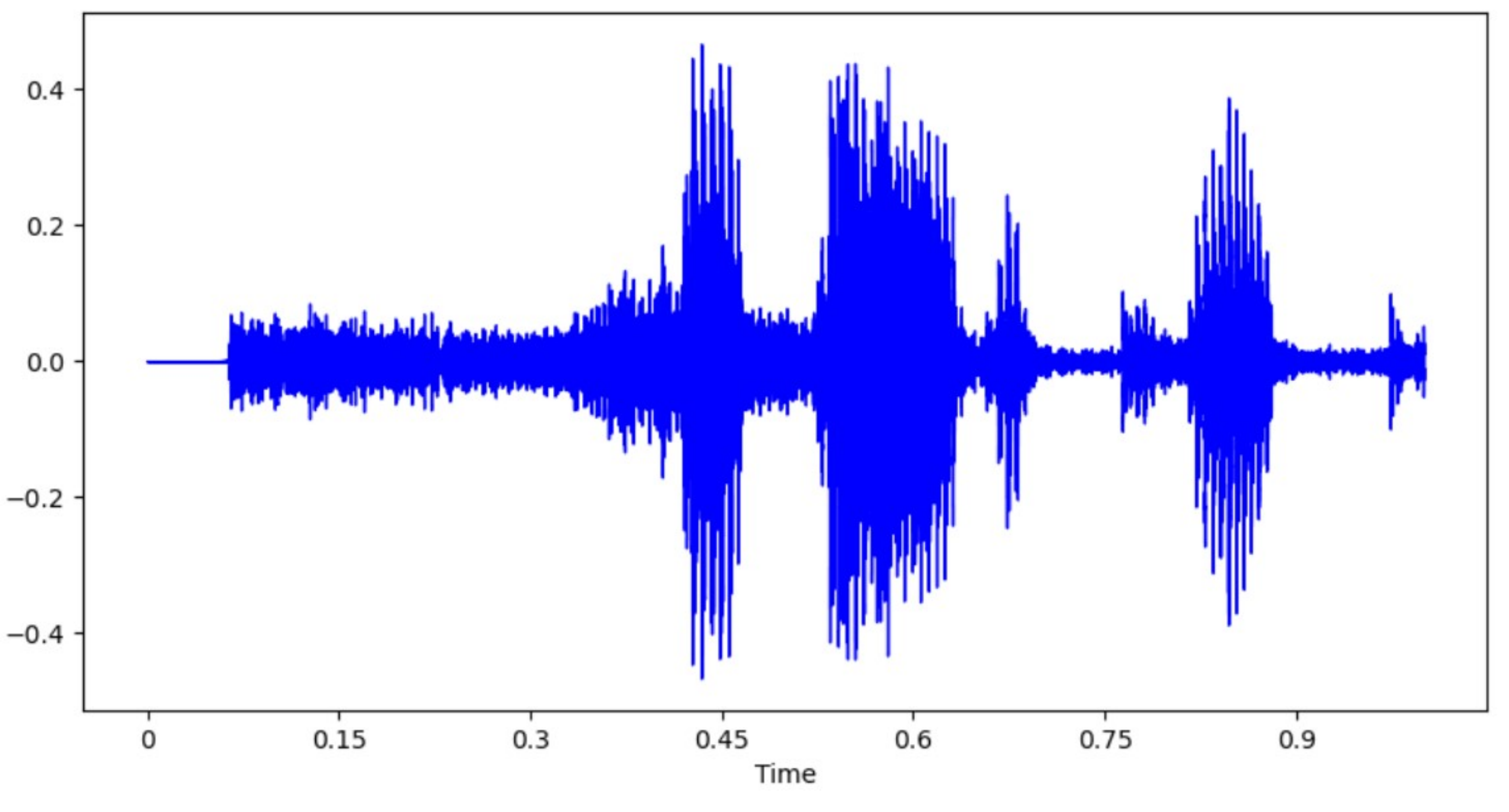
- Sample data is sound wave value against time, graphed here for 1 second



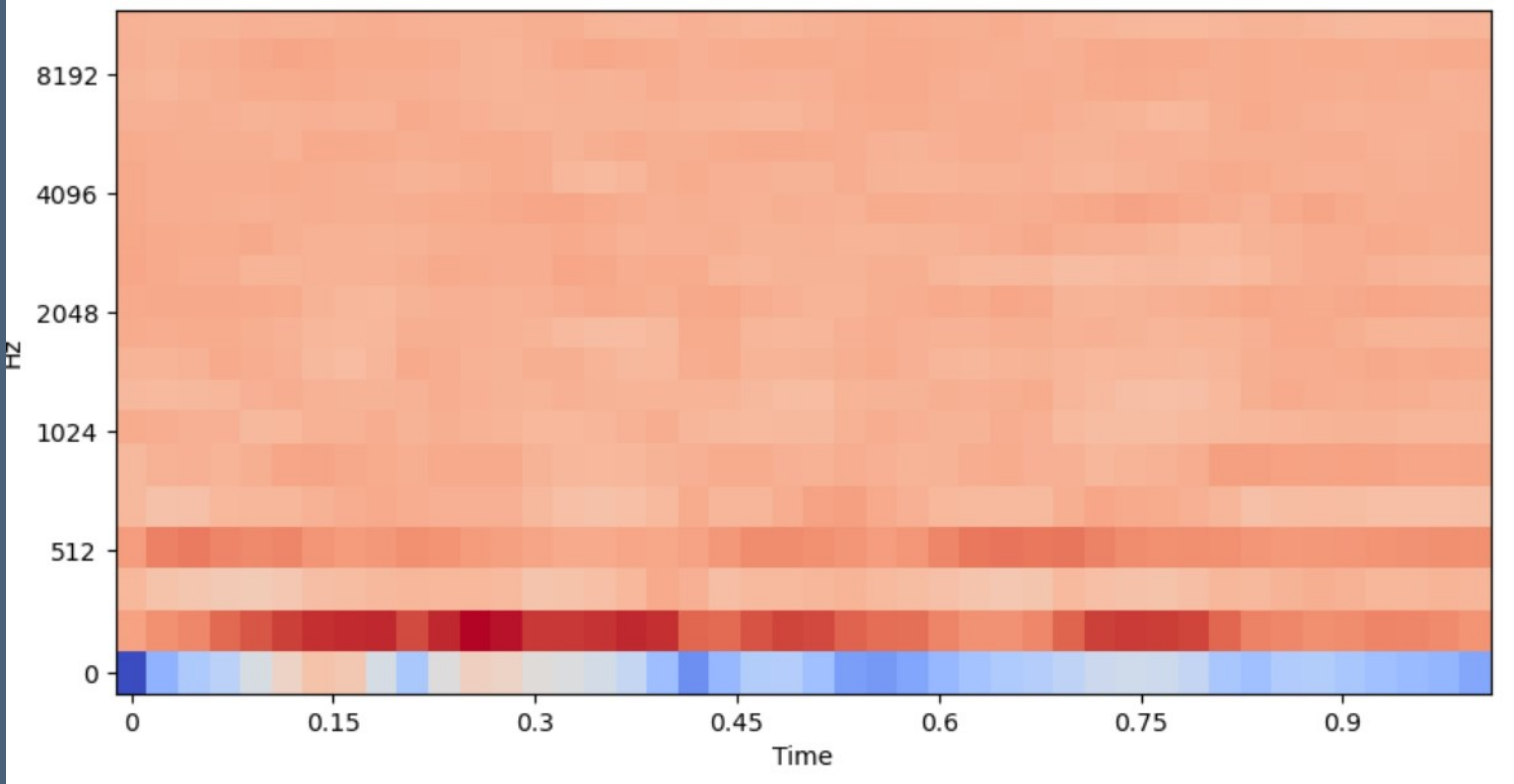
Preprocessing - Calculate Sound Features

From the sound recording, librosa is used to calculate a Mel Frequency Cepstrum Coefficients (MFCC), commonly used in telecoms, seismic and speech data analysis

MFCC is calculated using a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.



x-axis = time
y-axis = the frequency in Hz on a mel scale, a log scale that matches human sound perception
hue = power, the original x value squared



Estimates - Data

- Each speaker recording is split into samples of 1 second, literature shows ranges from 0.5 to 4 seconds
- MFCC for 20 sub-periods are calculated, the librosa default value, so around 50 milliseconds each
- Accent classification for the Belfast or Dublin datasets, binary variable (0/1) as Y variable
- Age and gender as categorical X variables
- From the speech sample, the MFCC, MFCC delta (lagged values) and MFCC delta 2 (lagged two periods) are numerical X variables
- The risk exists that the difference may be from recording equipment used generating distinctive sounds
- Forward-stepwise estimation - start with simple models, then add more variables and see if the performance improves

Classification Metrics

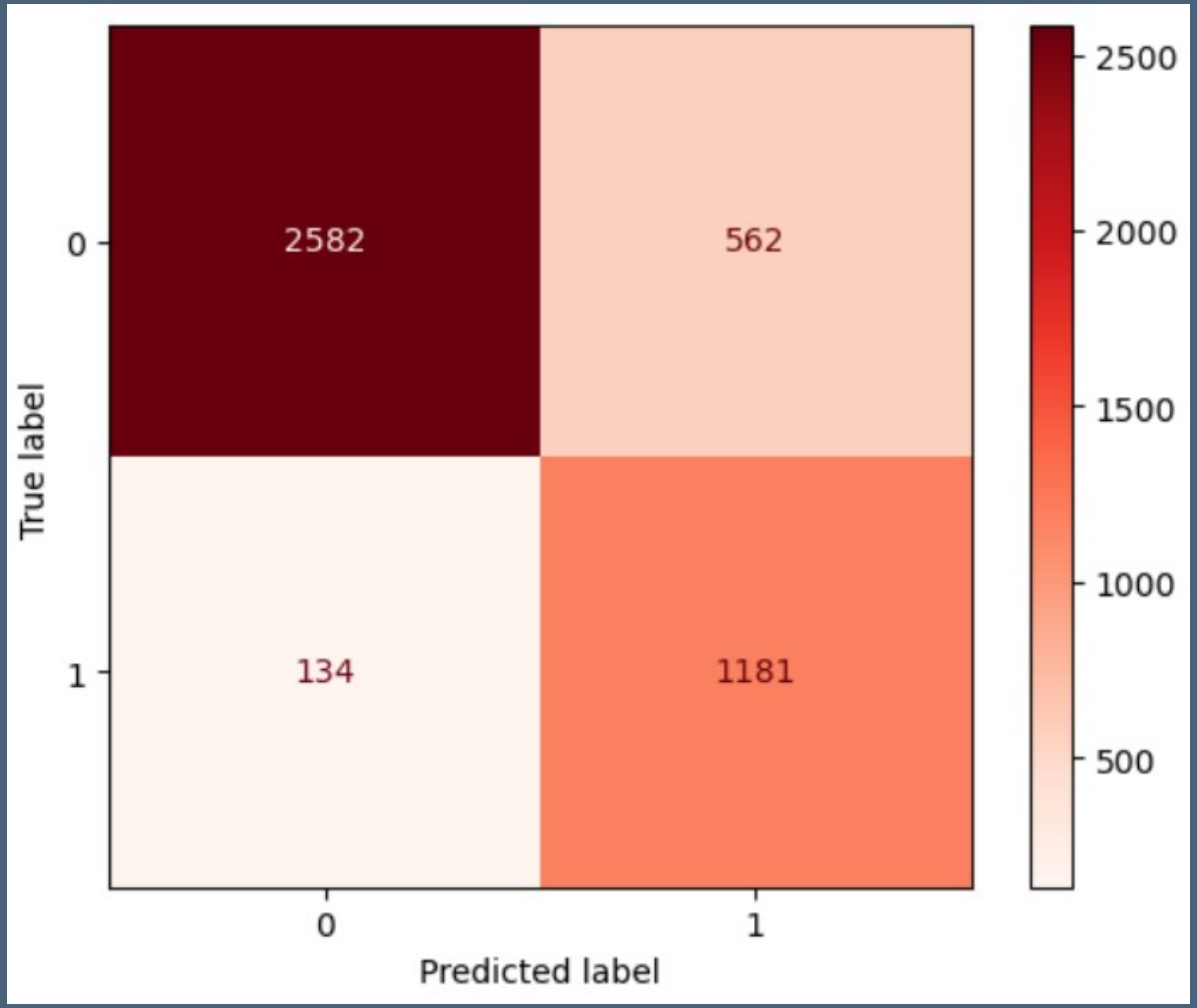
- Classification metrics calculations come from scikit-learn library also

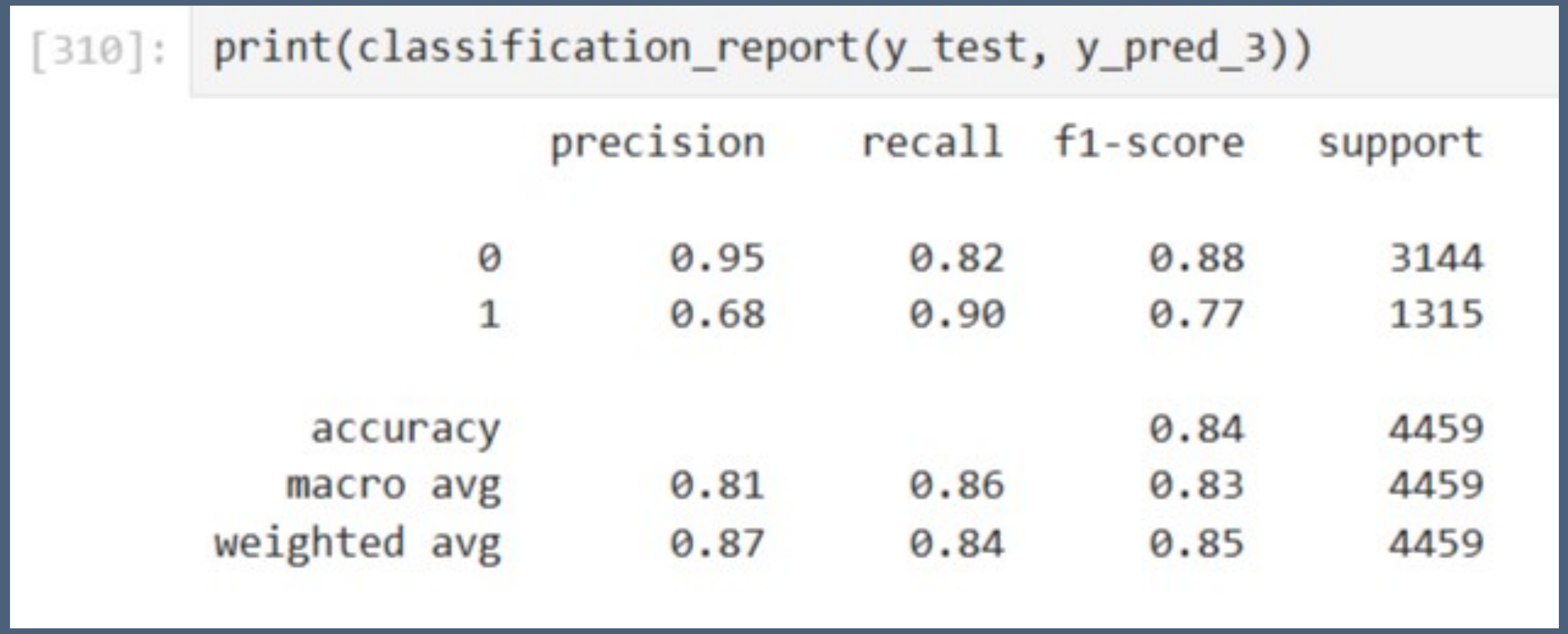
- Note that the two classes are *imbalanced*, significantly different from equal numbers, so the adjustment is needed in the model estimation e.g.
  lr3 = LogisticRegression(...class_weight='balanced')

- Confusion Matrix - a table that categorises predictions as True or False by output, shown for log reg model 2, X = MFCC
  True Negative =
  True Positive =
  False Negative =
  False Positive =

- classification_report gives summary statistics e.g.
  Accuracy = TN + TP / (TN+TP+FN+FP) =

-





```
[310]: print(classification_report(y_test, y_pred_3))

              precision    recall  f1-score   support

           0       0.95      0.82      0.88      3144
           1       0.68      0.90      0.77      1315

    accuracy                           0.84      4459
   macro avg       0.81      0.86      0.83      4459
weighted avg       0.87      0.84      0.85      4459
```

Classification Metrics

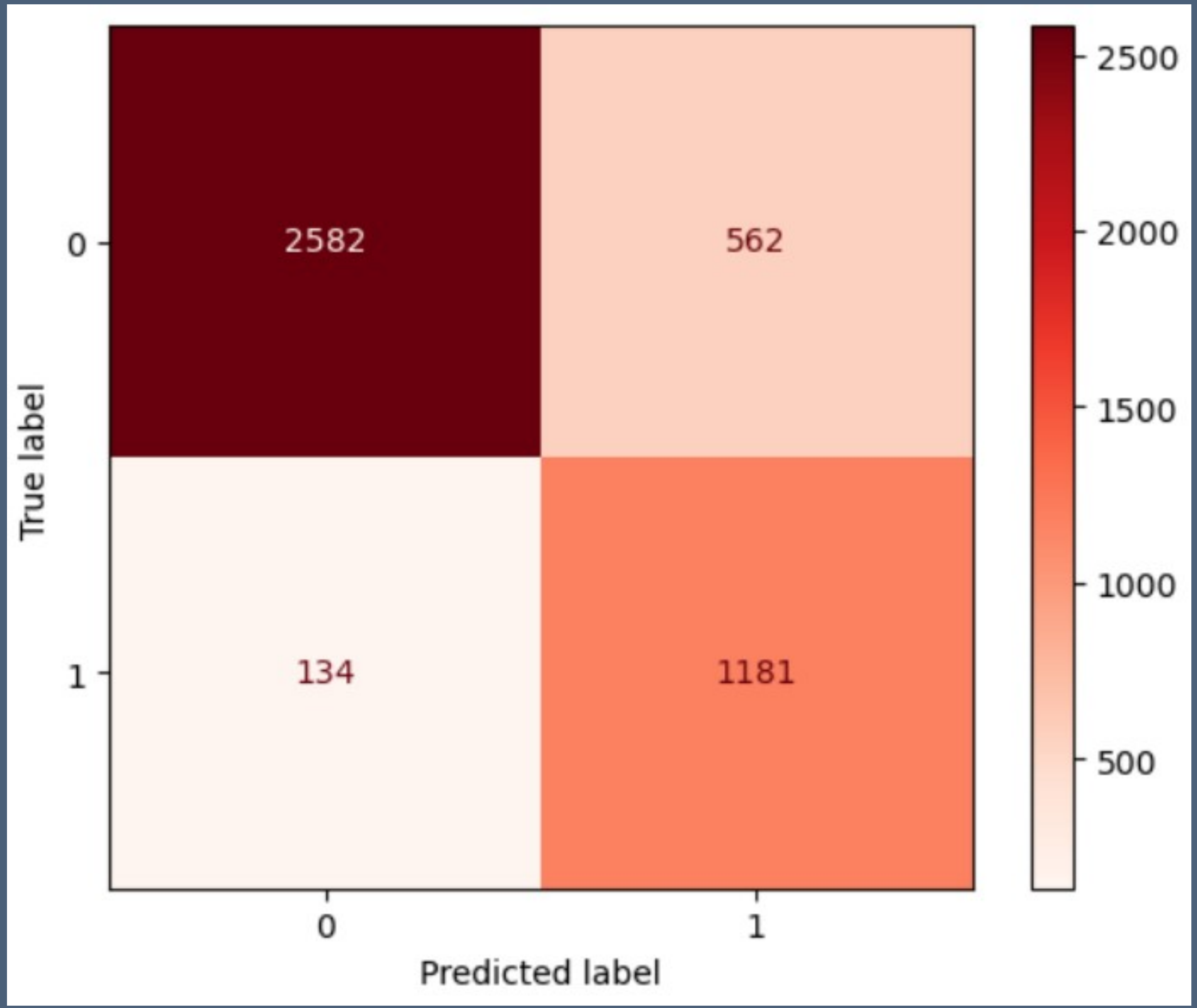
- Classification metrics calculations come from scikit-learn library also

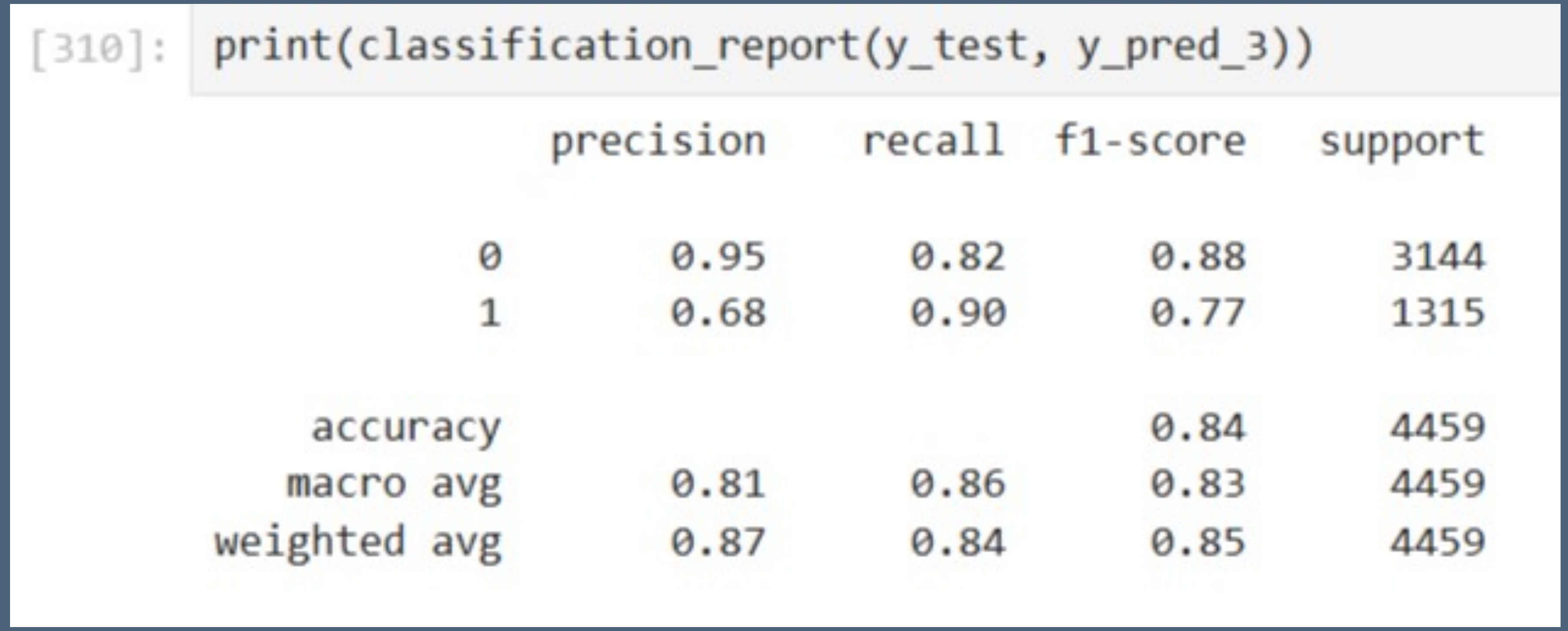
- Note that the two classes are *imbalanced*, significantly different from equal numbers, so the adjustment is needed in the model estimation e.g.
  lr3 = LogisticRegression(…class_weight='balanced')

- Confusion Matrix - a table that categorises predictions as True or False by output, shown for log reg model 2, X = MFCC
  True Negative = 2582
  True Positive = 1181
  False Negative = 134
  False Positive = 562

- classification_report gives summary statistics e.g.
  Accuracy = TN + TP / (TN+TP+FN+FP) = 84.4%

- Precision, Recall, F1 are calculated also - whether they are useful depends on the cost or effect of an error





```
[310]: print(classification_report(y_test, y_pred_3))

              precision    recall  f1-score   support

           0       0.95      0.82      0.88      3144
           1       0.68      0.90      0.77      1315

    accuracy                           0.84      4459
   macro avg       0.81      0.86      0.83      4459
weighted avg       0.87      0.84      0.85      4459
```

Classification Metrics

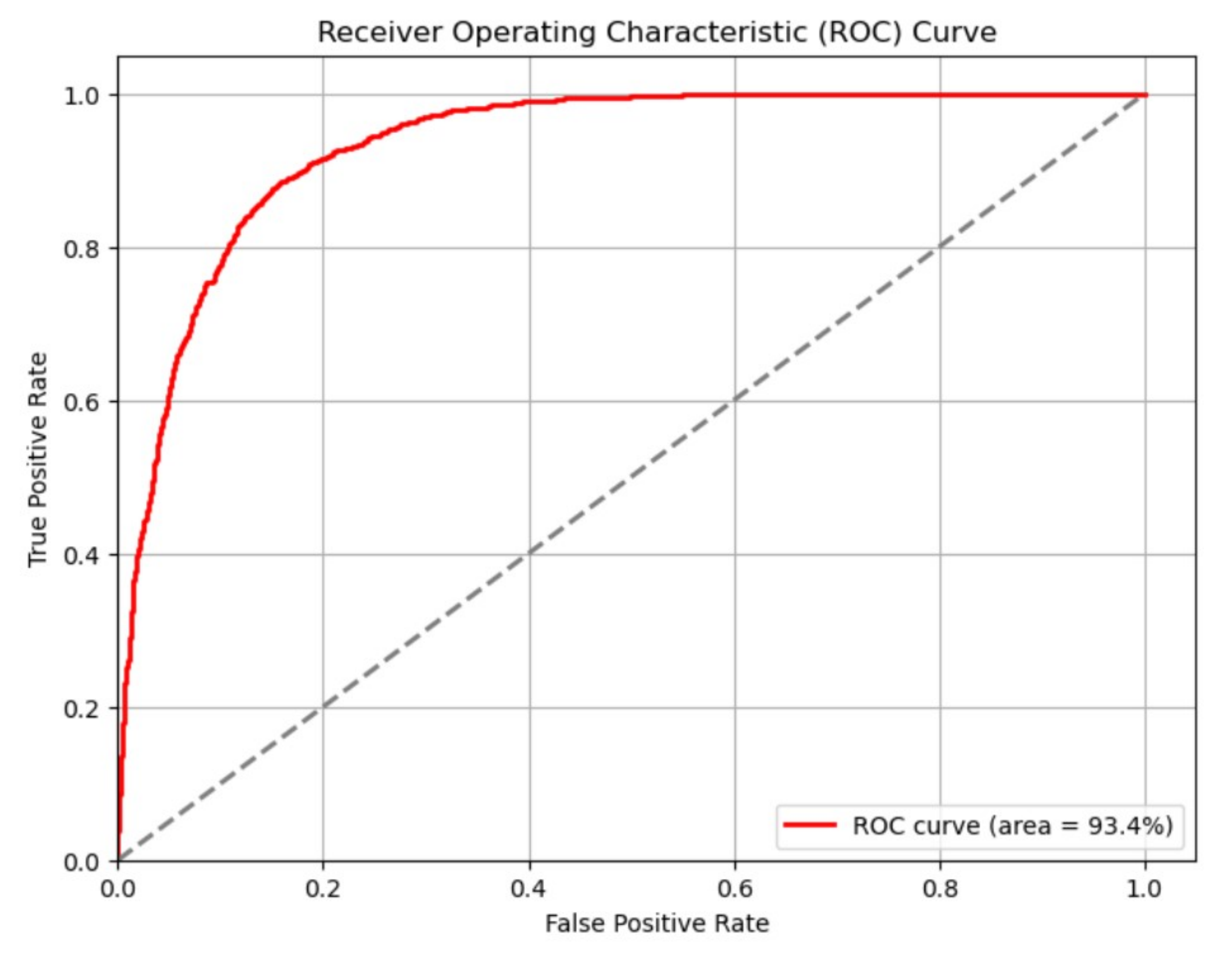
- The ROC evaluates classification model peformance

- Predictions are sorted by the model's estimated probability of the positive class, with the largest values first.

- Beginning at the origin, each prediction moves the curve vertically for a correct classification or horizontally for an incorrect one.

- The area under the ROC curve ('AUC') usually scales from a minimum 0.5 up to 1 for perfect prediction for all examples.

-



Logistic Regression

- Logistic regression estimates the probability that that one observation belongs to one of the two classes.

- Accent classification has Belfast or Dublin binary categorical variable (0/1) as Y variable

- MFCC features are the numerical X-variables, age and gender as categorical X variables

- Forward-stepwise evolution of the datasets - I run the logistic regression models, with different combinations of X variables i.e.
  (2) gender and age alone
  (3) MFCC alone
  (4) MFCC plus gender and age
  (5) then MFCC_delta with gender and age
  (6) MFCC_delta_2 with age and gender
  (7) All MFCC features, age and gender

- MFCC with age and gender was best combination of variables ROC-AUC of 95.3% and accuracy of 88.1%

Model	X	Accuracy	AUC
2	Gender, Age	70.8%	70.7%
3	MFCC	84.4%	93.4%
4	MFCC, Gender, Age	88.1%	95.3%
5	MFCC_delta, Gender, Age	71.5%	75.1%
6	MFCC_delta_2, Gender, Age	67.3%	76.6%
7	All	86.6%	94.7%

Random Forest

- Random Forest uses an ensemble of decision tree classifier models, trained on different random subsets of the training dataset. The most popular classification becomes the ensemble prediction.

- Accent classification has Belfast or Dublin binary categorical variable (0/1) as Y variable

- MFCC features are the numerical X-variables, age and gender as categorical X variables

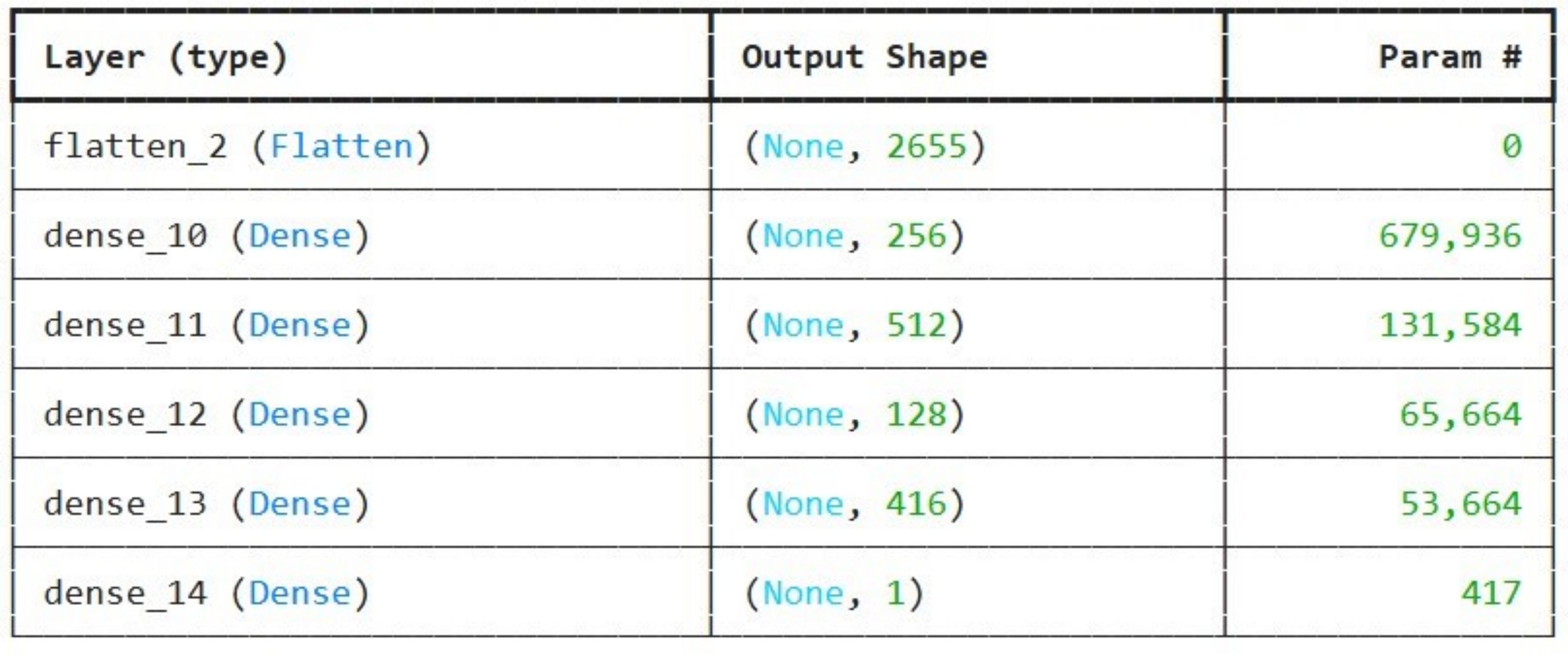
- Forward-stepwise evolution of the datasets - I run the random forest, with combinations of variables as before i.e.
  (2) gender and age alone
  (3) MFCC alone
  (4) MFCC plus gender and age
  (5) then MFCC_delta with gender and age
  (6) MFCC_delta_2 with age and gender
  (7) All MFCC features, age and gender

- MFCC with age and gender was best combination of variables ROC-AUC of 95.3% and accuracy of 88.1%

Model	X	Accuracy	AUC
2	Gender, Age	66%	69.7%
3	MFCC	84.4%	93.3%
4	MFCC, Gender, Age	86.3%	94.9%
5	MFCC_delta, Gender, Age	74.1%	76%
6	MFCC_delta_2, Gender, Age	71.2%	73.3%
7	All	84.4%	94.7%

Neural Network MLP

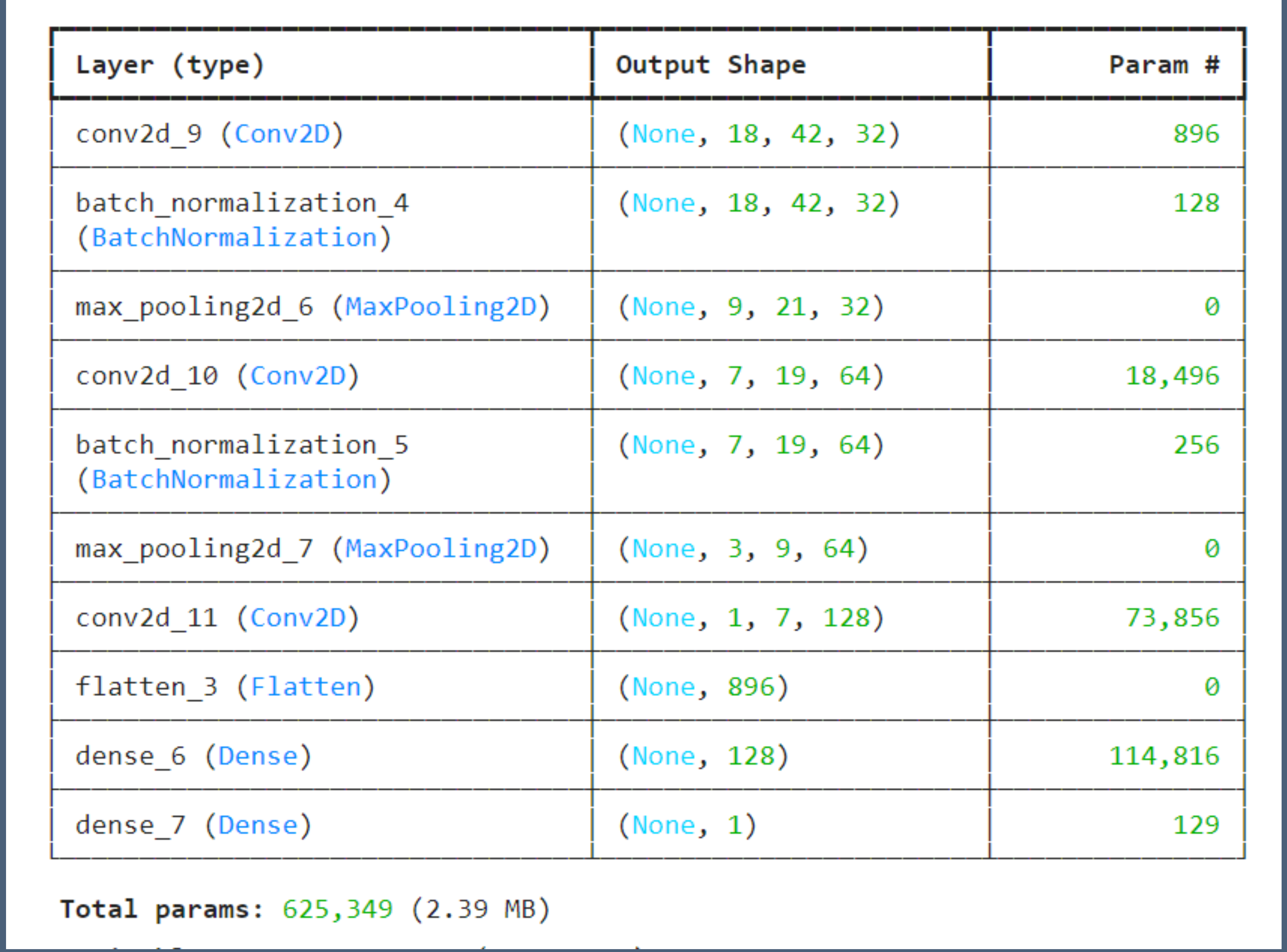
- Multi-layer Perceptron architecture
- Hyperparameters were selected using the Keras Tuner to choose among values at random from the neuron numbers, learning rate and choice of an extra (fourth) layer
- Target, or dependent, variable is a geographic accent location within Ireland
- MFCC was fed in as X variable
- Age gender were added as one-hot encoded inputs also to X
- Test dataset accuracy of 84.3%, AUC 88.4%
- Test accuracy 80-88% are typical and ROC AUC of 92%



Layer (type)	Output Shape	Param #
flatten_2 (Flatten)	(None, 2655)	0
dense_10 (Dense)	(None, 256)	679,936
dense_11 (Dense)	(None, 512)	131,584
dense_12 (Dense)	(None, 128)	65,664
dense_13 (Dense)	(None, 416)	53,664
dense_14 (Dense)	(None, 1)	417

Convolutional Neural Network

- Y is the region of the accent
- MFCC were fed in as X variable
- Test dataset accuracy was 91.84%, AUC of 95.1%
- MFCC speech data alone was achieving test accuracy in the 89--92% range.



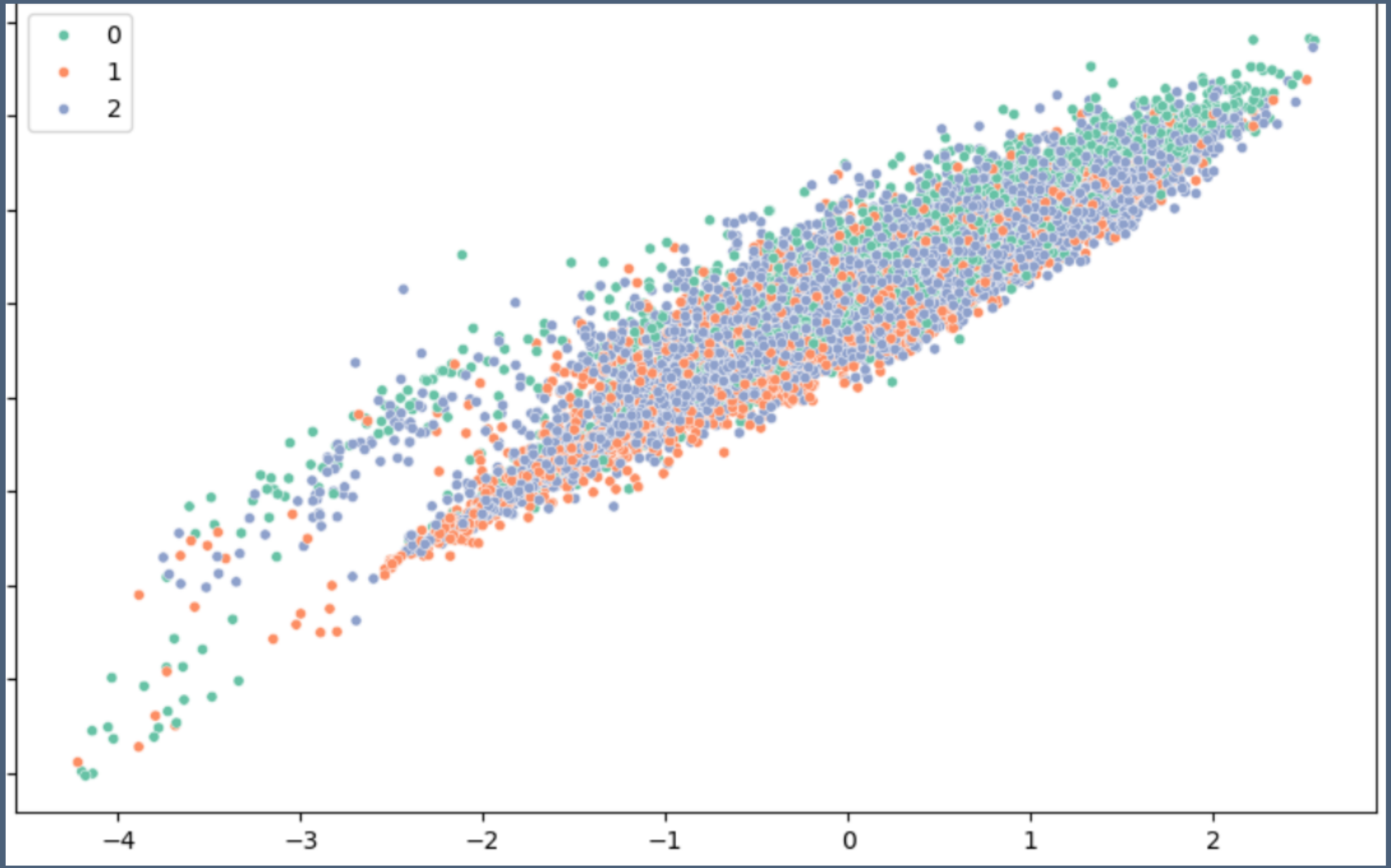
Layer (type)	Output Shape	Param #
conv2d_9 (Conv2D)	(None, 18, 42, 32)	896
batch_normalization_4 (BatchNormalization)	(None, 18, 42, 32)	128
max_pooling2d_6 (MaxPooling2D)	(None, 9, 21, 32)	0
conv2d_10 (Conv2D)	(None, 7, 19, 64)	18,496
batch_normalization_5 (BatchNormalization)	(None, 7, 19, 64)	256
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)

Clustering Analysis

Clustering MFCC, gender and age with n=2 and n=3 k-means algorithms on the country was tried too

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



Conclusions & Future Research Questions

- With logistic regression, speech data from MFCC, on its own and added to personal characteristics give significantly positive classification performance, accuracy of 80-88%.
- Random forest models almost matched that.
- MLP improved on those results and then CNN more so, up to 92%.
- Wide range of more evolved deep-learning models to apply in future research: 'Large Audio Models' claim 90-94% accurate classification, but my LAM examples (not shown here) came nowhere close: investigation continues.
- More data, assured data quality, and the ability to create links to other variables are key to more robust estimation and more advanced machine learning models.
- What future questions?

Future Research Questions?

- More data and better assurance around data quality will are key to more robust estimation from future work
- More advanced machine learning models should be adapted and tried
- Combining data from other languages, Irish and Scots in particular
- Mutli-modal data - add video, social media metrics and consumer behaviour to speech data
- Data can be gathered at an institution level, especially secondary schools, which have a catchment area, income level and a religious identity
- The ability to create links to other personal, demographic and social variables will need design and consultation of the personal and geographic data to match government and statistical boundaries while preserving privacy
- How do accents vary across Dublin city and county? Suburbs, neighbourhoods, streets? What mobility or economic status affects this?
- How much variation is there by geography e.g. is a Belfast accent closer to modern Scots English or other Irish English?
- How do regional accents in Irish English match with Irish language pronunciation?
- Do Dublin and Liverpool match?
- Are effects of inward migration seen in the common speech?

Next Steps - More and Better Data

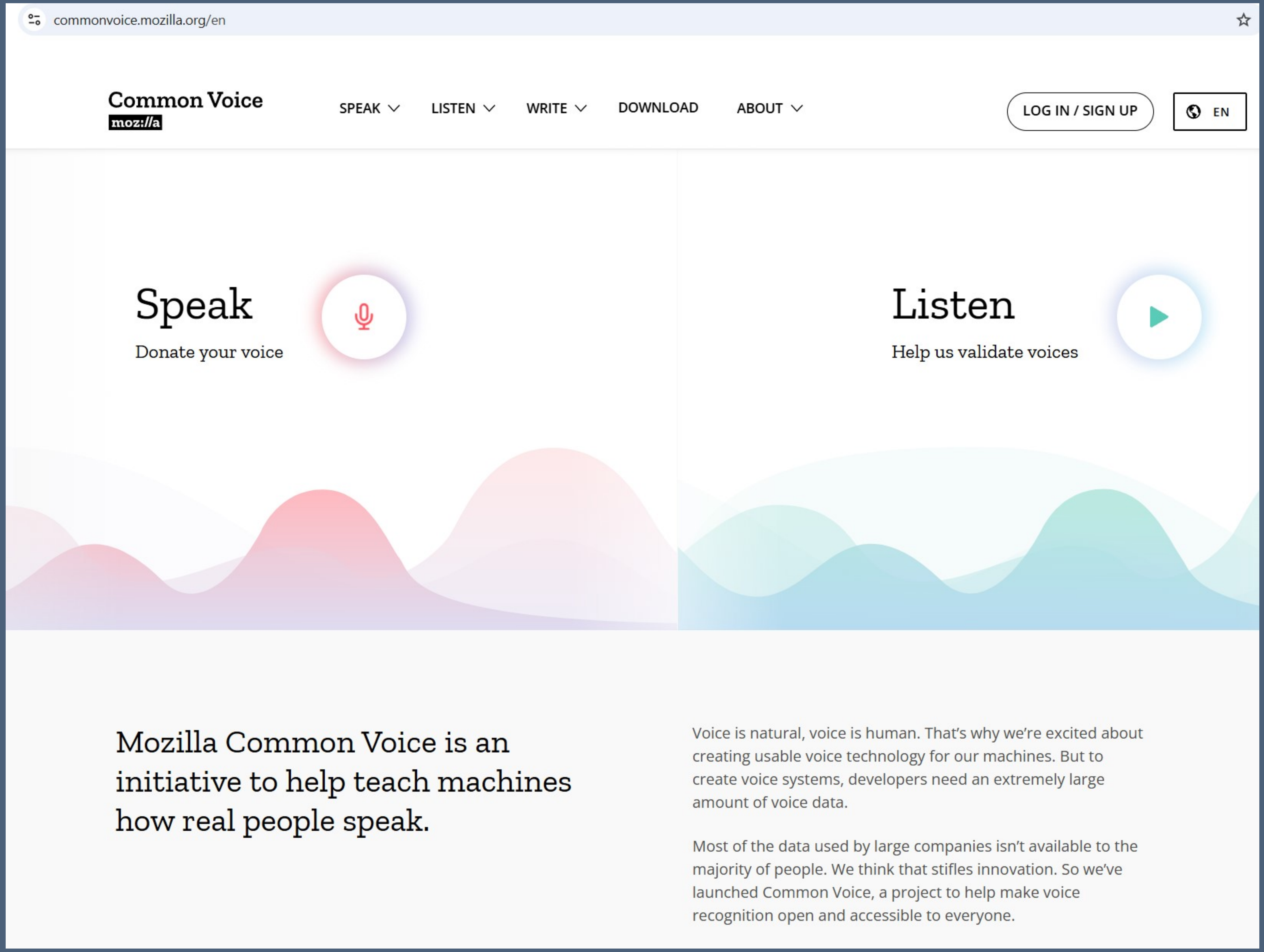
- Crowd-sourcing data online is cheaper and easier than manual collection e.g. http://commonvoice.mozilla.org/en

- I am in discussions with Mozilla and Irish universities about creating an academic partnership to do this

- CommonVoice Data is labelled by country of origin as Irish, by age and gender, but does not give more detailed geographic or social data

- We can add data fields to existing recordings and seek new samples

- Email me at databeaker@gmail.com if you would like to participate in future research or data collection



Mozilla Common Voice is an initiative to help teach machines how real people speak.

Voice is natural, voice is human. That's why we're excited about creating usable voice technology for our machines. But to create voice systems, developers need an extremely large amount of voice data.

Most of the data used by large companies isn't available to the majority of people. We think that stifles innovation. So we've launched Common Voice, a project to help make voice recognition open and accessible to everyone.

Privacy Implications

- Speech samples cover more of the population, as speech processing apps are used more, almost all this is held within the large tech companies, where it may be accessed to your detriment e.g. Saudis and Twitter

- Individual identification is no longer secure against impersonation by voice-cloning e.g. https://elevenlabs.io, so spear-phishing fraudsters will impersonate you to make transfers, or tell your relatives that you need money

- Network owners or governments can gather and identify speech samples also without users knowing, split the cables in a locked room or hack the data centres and cables between them (Snowden); GCHQ gathers mobile phone signals using planes over British cities (R.Aldrich, 'GCHQ',2019).

- Individual identification possible by crossing with other datasets - customer records, social media

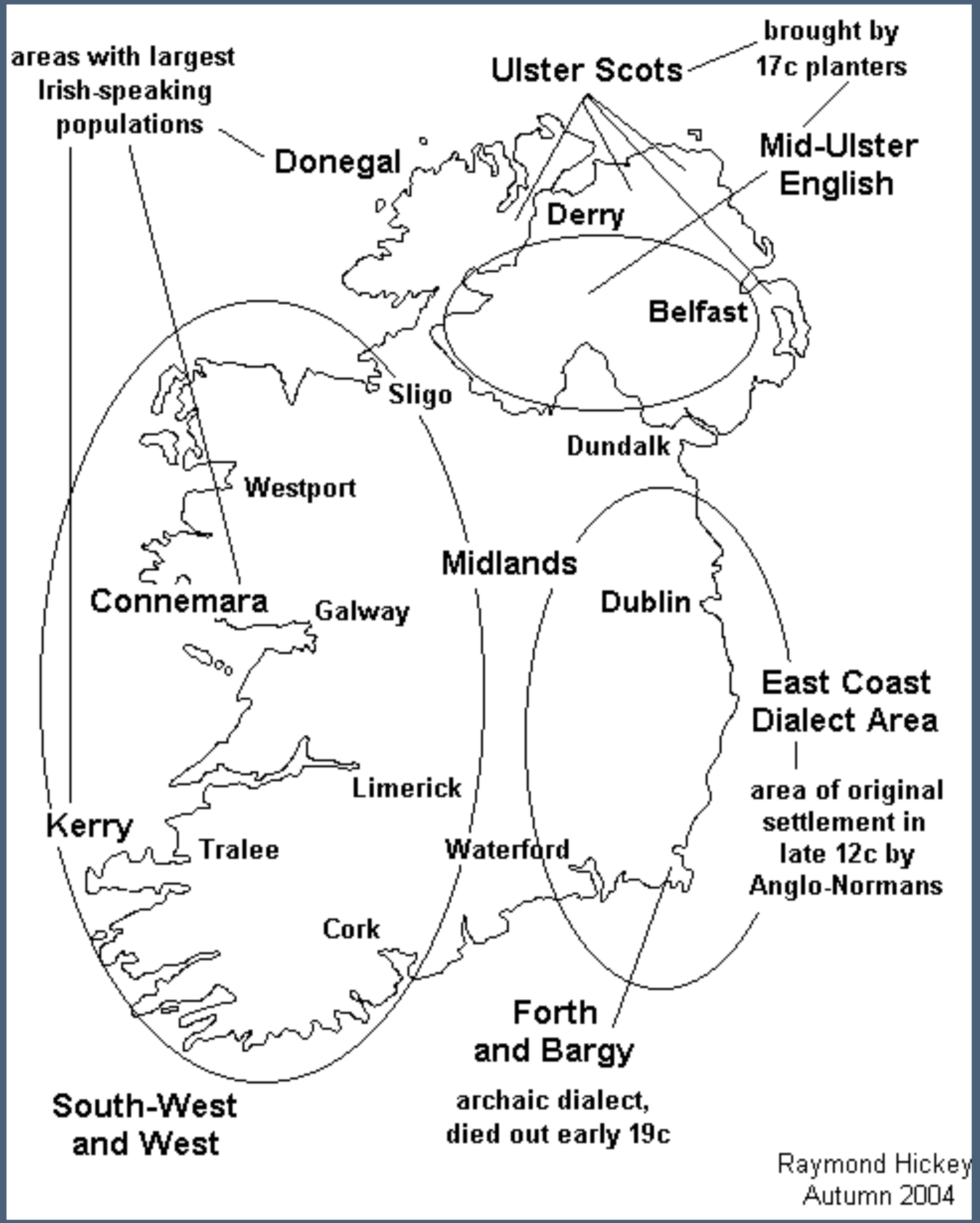
Privacy Compliance

- Signoff for compliance with Irish data protection legislation
- Anonymisation by aggregation, printing only age range, gender county, town and district is mandatory
- No personally identifying data should be gathered
- No email addresses or IP or other network data published
- Subjects should be 18+ years old
- Sample contributor conditions specify no sharing or sale of data and legal ownership by a trustee
- Users have the contracted right to withdraw consent for use at any time
- Record with no name, address, only an email for an individual or institution

How the Irish speak English: Python Machine Learning for Classification of English-language Accents in Ireland

Peter Nolan
databeaker@gmail.com
https://github.com/dpnolan/voxpop

PyCon Ireland, Saturday 16 November 2024


areas with largest Irish-speaking populations
Donegal
Ulster Scots
brought by 17c planters
Mid-Ulster English
Derry
Belfast
Sligo
Dundalk
Westport
Midlands
Connemara
Galway
Dublin
East Coast Dialect Area
area of original settlement in late 12c by Anglo-Normans
Limerick
Kerry
Tralee
Waterford
Cork
Forth and Bargy
archaic dialect, died out early 19c
South-West and West
Raymond Hickey
Autumn 2004