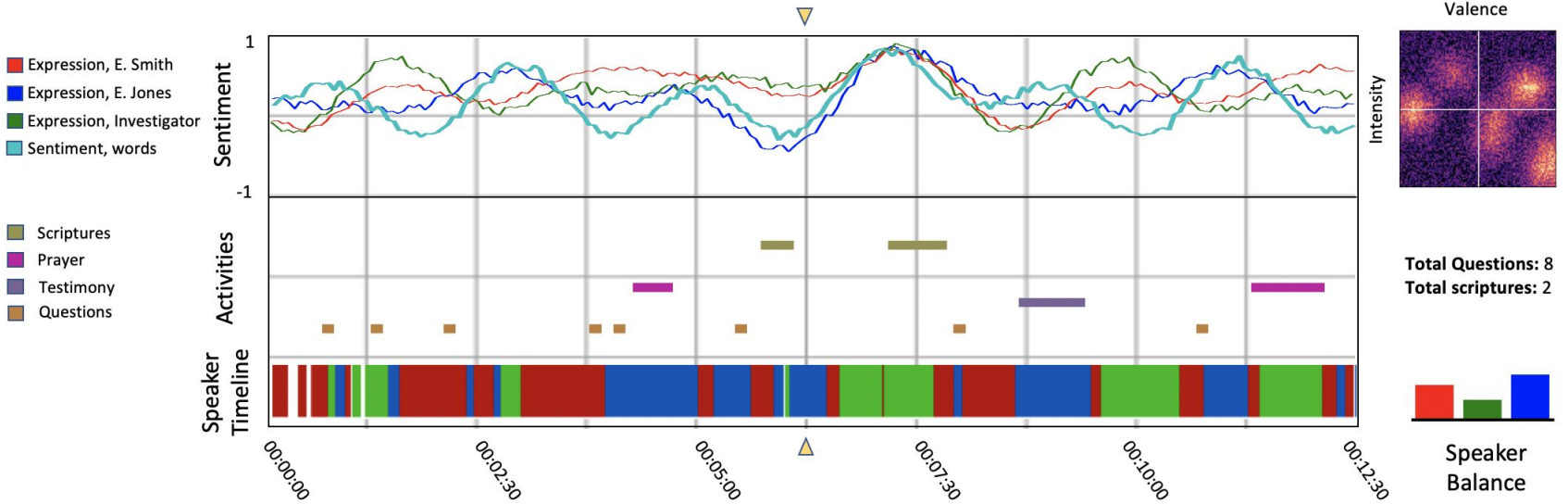
Question Detection Using Decision Tree Models

By Griffin Holt, Henry Davis, & Tysum Ruchti



Investigator | Elder Smith | Elder Jones

Sentiment

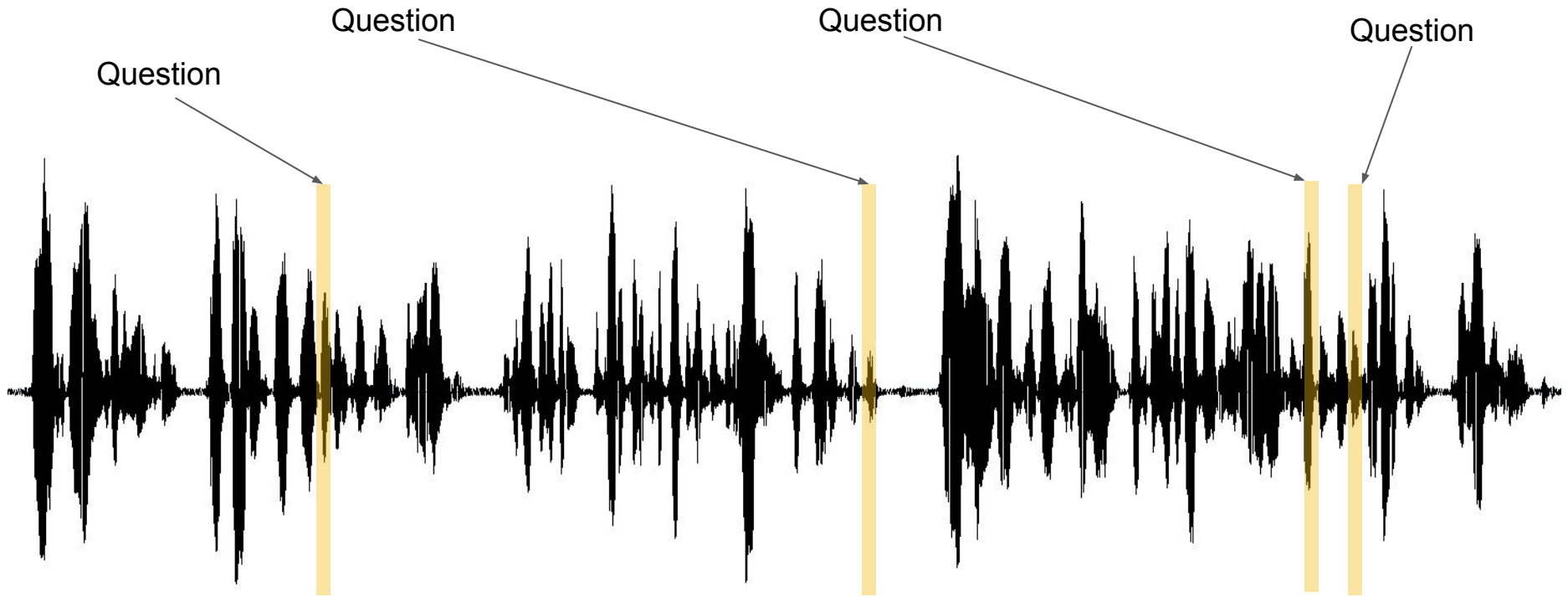
- Facial expression
- Body language
- Voice tone
- Text sentiment

Activities

- Scriptures
- Prayer
- Testimony
- Questions

Timeline

- Show speaker overlap

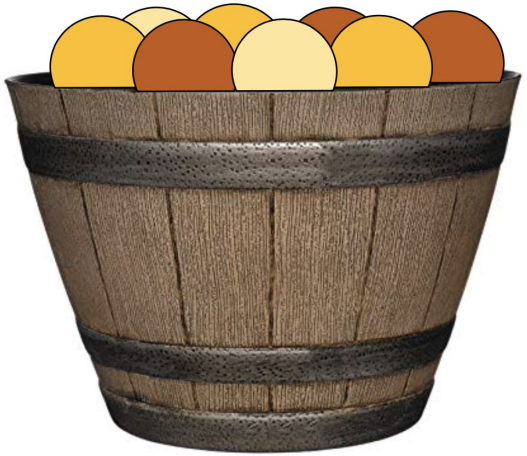

Question
Question
Question
Question

Research

- Focus: Question identification through *sound* rather than *word content*
- Most Commonly Reported Success: Adaboost Random Forests & RNNs
- Used statistics of various audio features
    - Pitch (especially towards the end of a clip) = Important
- Improved results when *sound* analysis combined with *word content* analysis

The Learning Problem

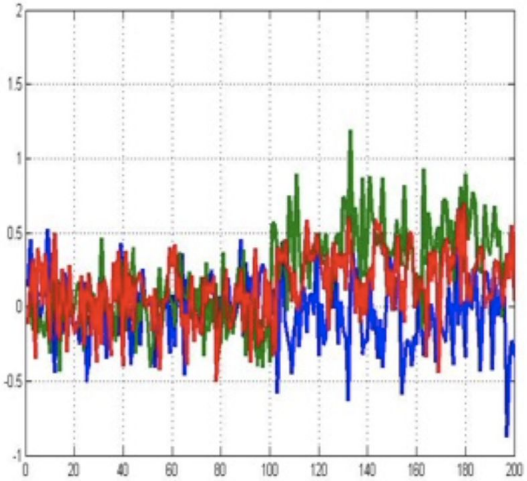
1. Bucket of Models



Smaller Bucket:
- Decision Trees

Bigger Bucket:
- Random Forests

2. Data



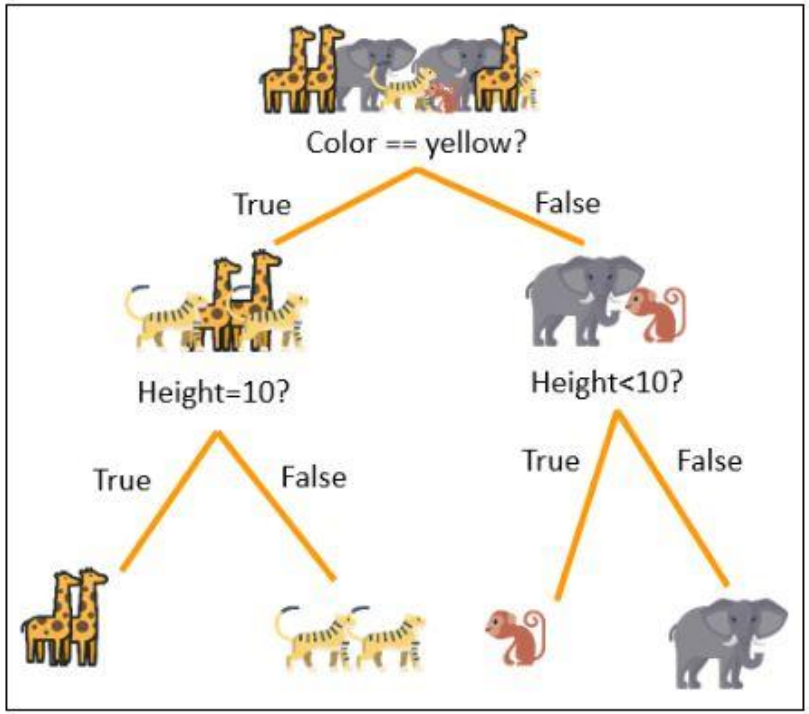
Audio files with timestamps at beginning of each question

3. Notion of “Best

1. Gini Impurity
2. Adaboost Error Function
3. ROC AUC Score

1. The Bucket of Models

The Smaller Bucket: Decision Trees

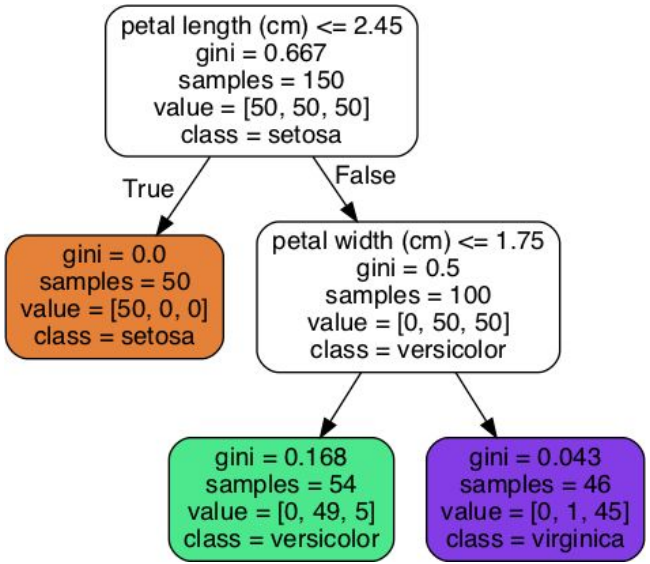


https://www.simplilearn.com/tutorials/machine-learning-tutorial/decision-tree-in-python

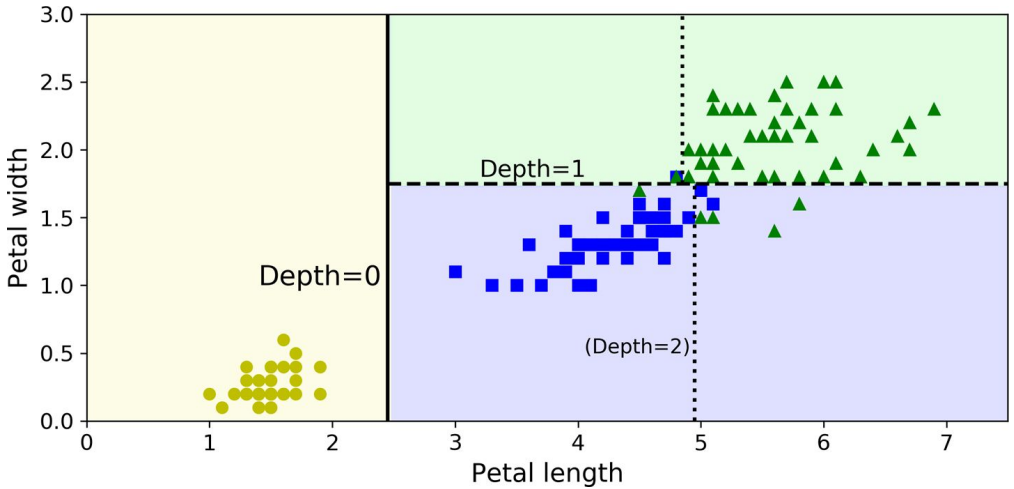
1. The Bucket of Models

The Smaller Bucket: Decision Trees

Tree Visualization



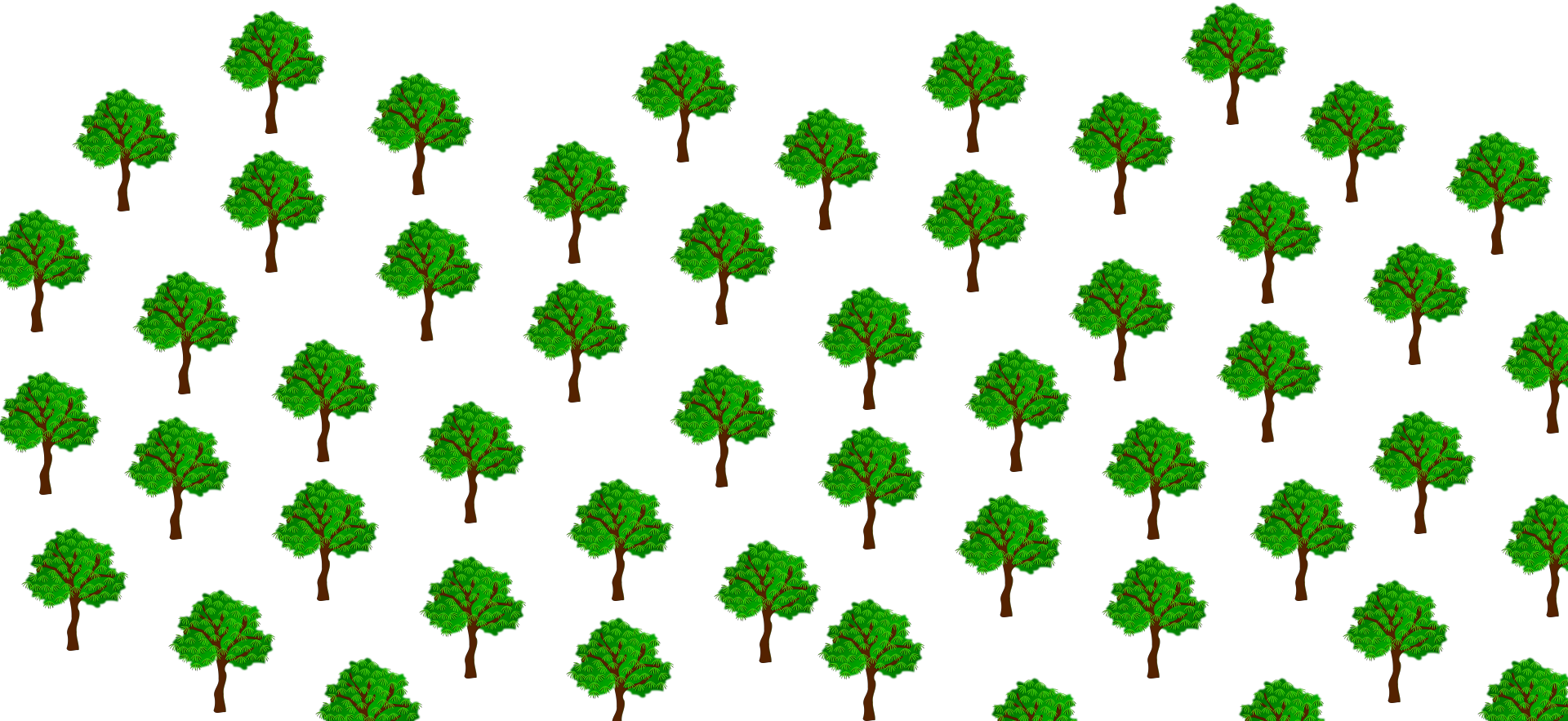
Division of Feature Space Visualization



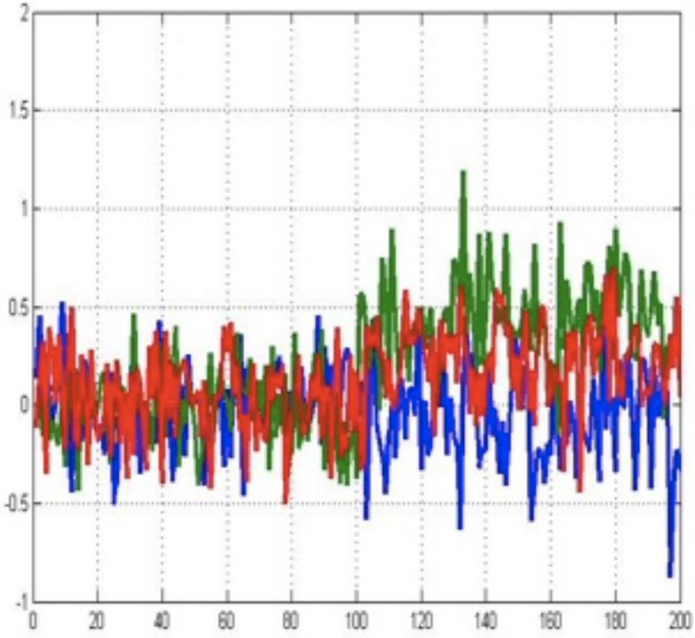
Chapter 6: Decision Trees. (2019). In A. Géron (Ed.), *Hands-On Machine Learning with Scinkit-Learn, Keras & TensorFlow* (2nd ed., pp. 175-188). Sebastopol, CA: O'Reilly. Retrieved December 15, 2020, from https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ch06.html#trees_chapter

1. The Bucket of Models

The Bigger Bucket: Random Forests of 200 Decision Trees

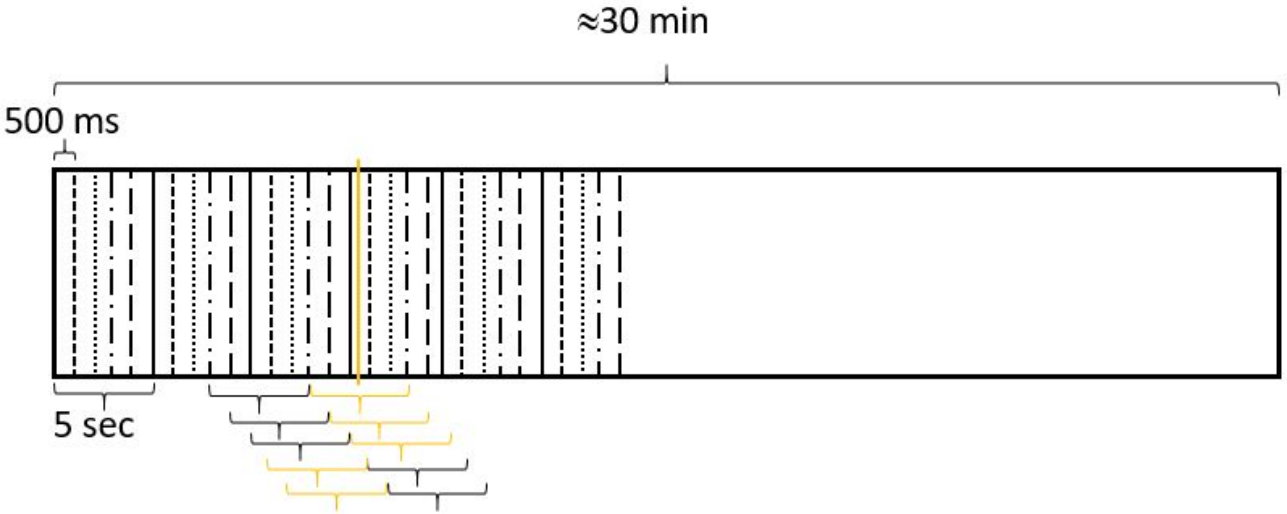


2. Data



28 Videos, each w/ XML files containing annotations

2. Data



- Rolling Windows of 5-second clips every 500 ms
- ~105,000 total clips created from 28 videos
- Randomly selected 20,000 clips
  - Trained on 16,000 (80%)
  - Tested on 4,000 (20%)

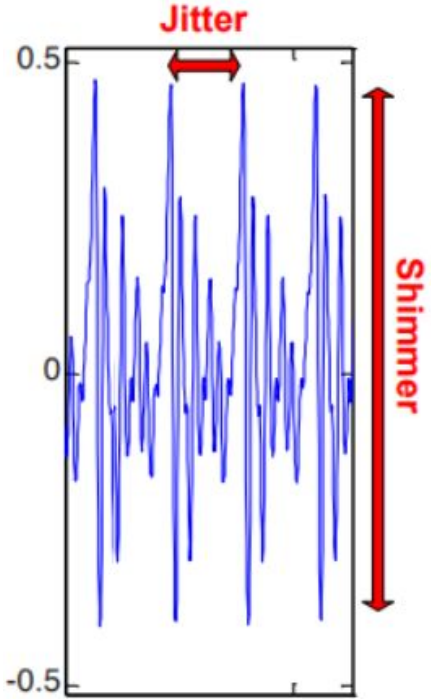
2. Data

Jitter

$$\frac{N}{N-1}\frac{\sum_{i=1}^{N-1}|T_{i+1}-T_i|}{\sum_{i=1}^{N}T_i}$$

Shimmer

$$\frac{N}{N-1}\frac{\sum_{i=1}^{N-1}|A_{i+1}-A_i}{\sum_{i=1}^{N}A_i}$$

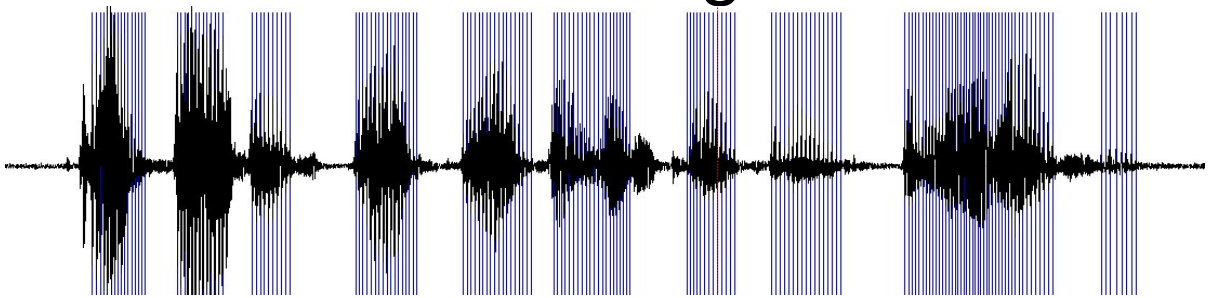


RMS- Frame Energy

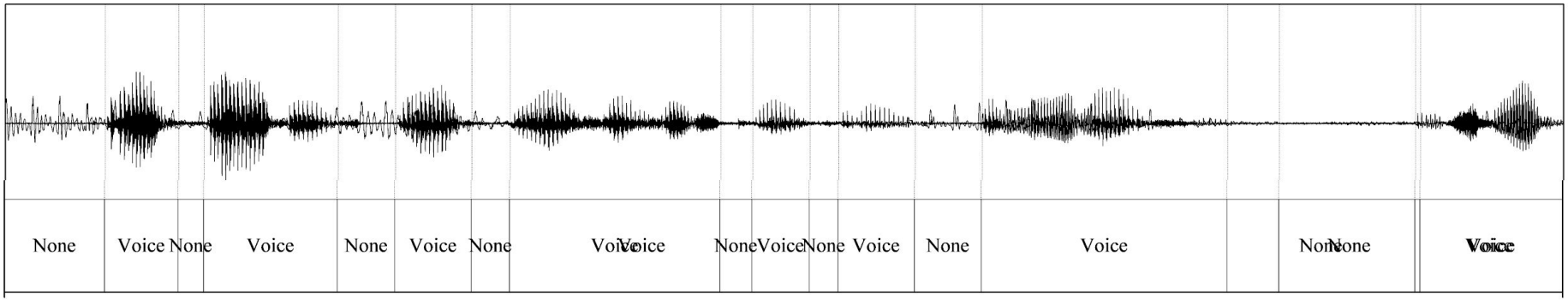
$$\sqrt{\frac{1}{N}\sum_{n}|x(n)|^2}$$

2. Data

Zero-Crossing Rate



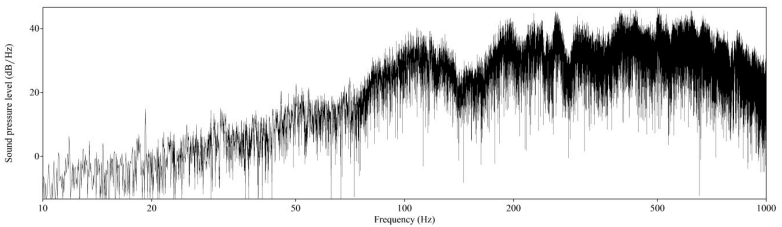
Number of Pauses



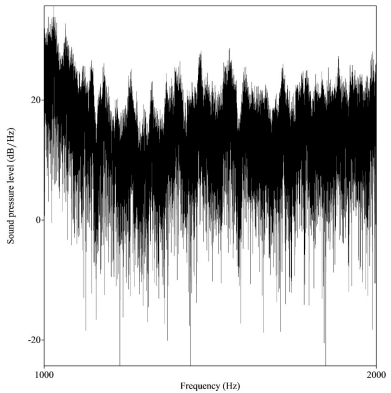
2. Data

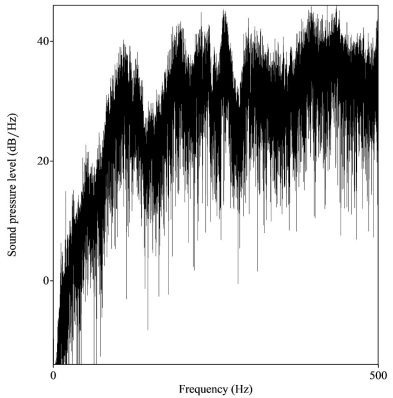
Power Spectral Density (PSD)

$$\int_{t_1}^{t_2} |e^{(-i2\pi ft)}x(t)|^2\,dt$$



Spectral Balance
(1 to 2kHz) vs (0 to 0.5kHz)



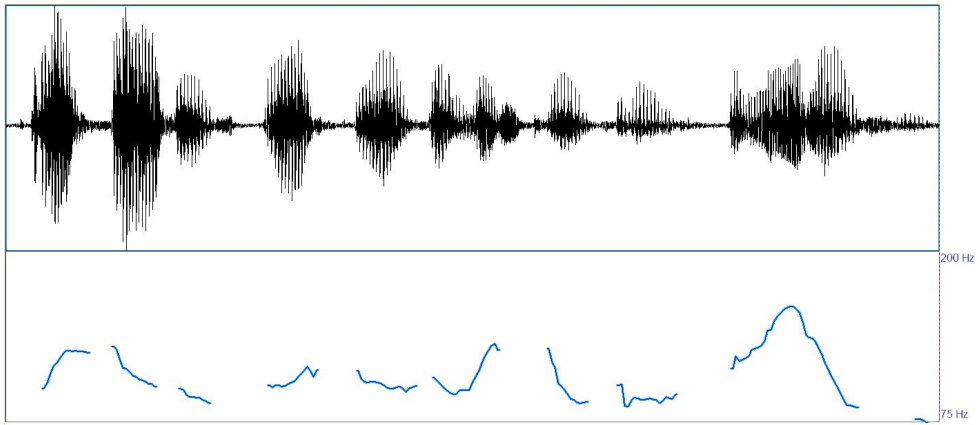


2. Data

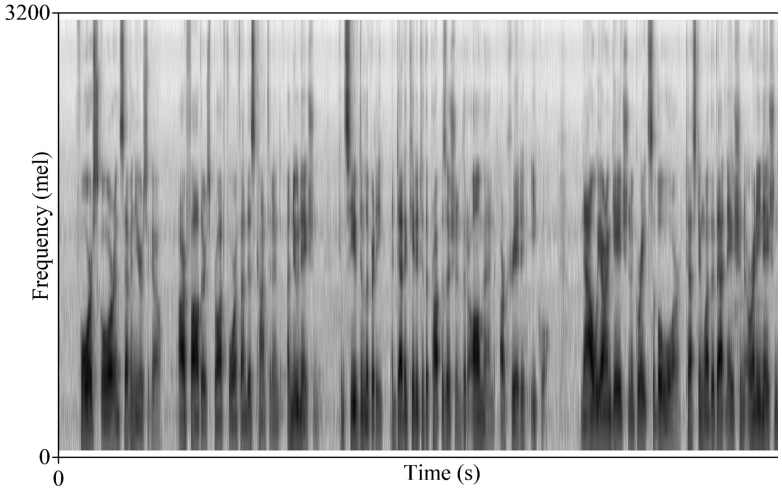
Signal Intensity

$$10 \log_{10} \left(\frac{1}{4 \times 10^{-10} (t_2 - t_1)} \int_{t_1}^{t_2} x^2(t)\, dt \right)$$

Fundamental Frequency



Mel Frequency Cepstral Coefficients (MFCC)



Input Variables

Jitter
Shimmer
Root Mean Square (RMS) Frame Energy
Zero-Crossing Rate
of Pauses
Pause Length
Power Spectral Density
Spectral Balance
Signal Intensity
Fundamental Frequency
MFCC Coefficients (Default = 13)

X

Mean
Standard Deviation
Median
Maximum
Minimum
Range
Skew
Kurtosis
Least-Squares: Slope, Offset, MSE
First-Last Pts Line: Slope, Offset, MSE
% of Rising Slopes

X

Full 5-Second Clip
Last 500 Milliseconds
Last 200 Milliseconds

=

363 Possible Variables for each 5-Second Clip

We chose a subset of these:

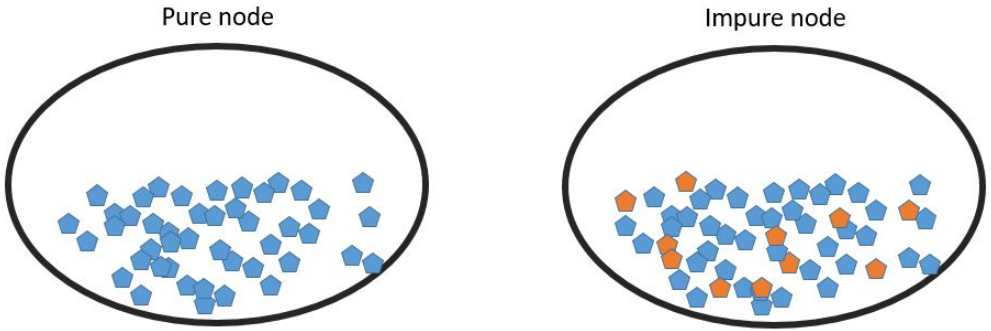
194 Selected Variables for each 5-Second Clip

3. Notion of “Best”

1. Gini Impurity
2. Adaboost Error Function
3. ROC AUC Score

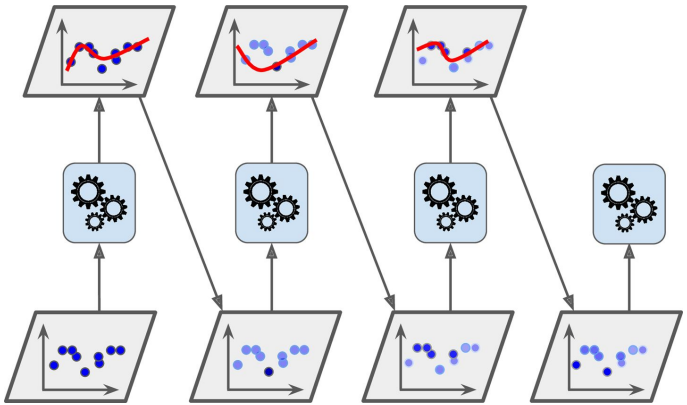
3. Notion of “Best”: Gini Impurity

$$G(k) = \sum_{i=1}^{J} P(i)(1 - P(i))$$



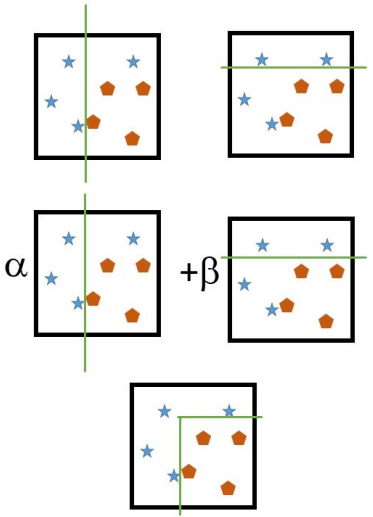
3. Notion of “Best”: Adaboost Error Function

Adaboost Algorithm for Finding a Random Forest Model



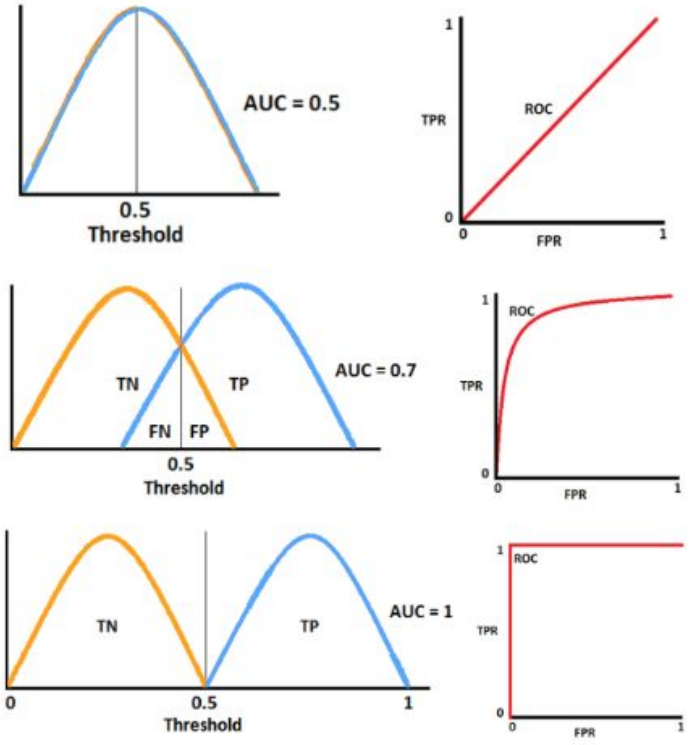
AdaBoost Error Function

$$\epsilon_m = \frac{\sum_{i=0}^{n} w_i^{(m)} (1 - \delta_{x,y})}{\sum_{i=0}^{n} w_i^{(m)}}$$



Chapter 7: Ensemble Learning and Random Forests. (2019). In A. Géron (Ed.), *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow* (2nd ed., pp. 189-212). Sebastopol, CA: O'Reilly.

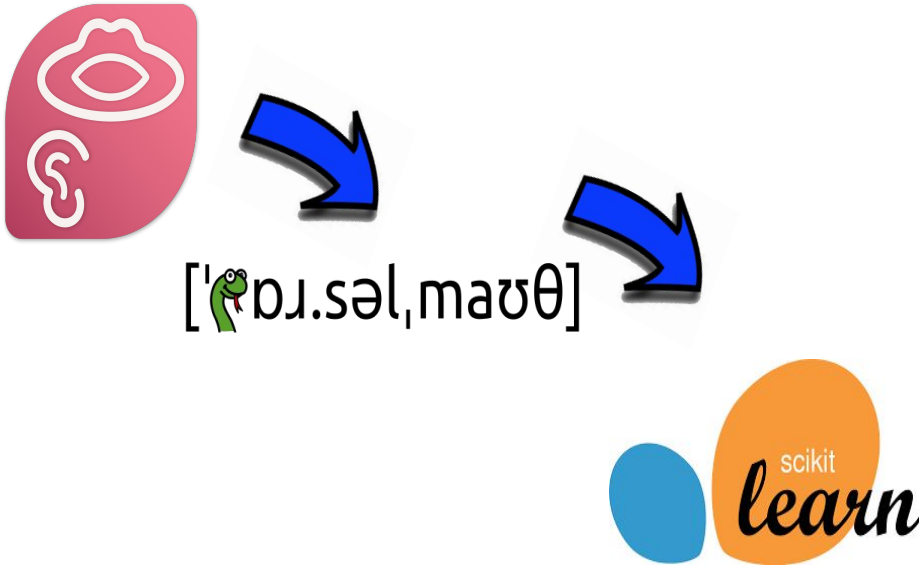
3. Notion of “Best”: ROC-AUC



towardsdatascience.com

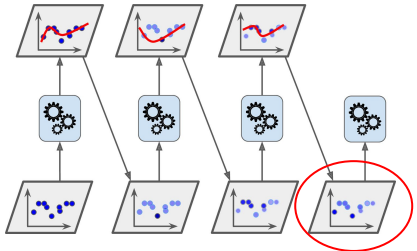
Methods

1. Grid Search on hyperparameters
    a. Gini Impurity vs. Entropy
    b. Max. Depth of the Trees
    c. Learning Rate (0.25, 0.5, 0.75)
2. 5-Fold Cross Validation
3. Tools Used
    a. Praat
    b. Parselmouth
    c. Scikit-Learn



Results

Visualization of Last Decision Tree



Most Important Features

Importance	Audio Feature	Statistic	Clip Portion	Decrease in Gini Impurity
1	**Fundamental Frequency f0**	Median	**Last 500 Milliseconds**	**0.384219106**
2	**Fundamental Frequency f0**	Minimum	Last 200 Milliseconds	**0.38409663**
3	Signal Intensity	Standard Deviation	Full Clip	**0.288251239**
4	**Fundamental Frequency f0**	Standard Deviation	**Last 500 Milliseconds**	**0.195265416**
5	**Fundamental Frequency f0**	% of Rising Slopes	Full Clip	**0.059541905**
6	6th MFCC	Mean	**Last 500 Milliseconds**	**0.054253965**
7	Signal Intensity	% of Rising Slopes	Last 200 Milliseconds	**0.051263641**
8	3rd MFCC	Standard Deviation	Full Clip	**0.050303281**
9	2nd MFCC	Mean	Full Clip	**0.040921364**
10	11th MFCC	Mean	**Last 500 Milliseconds**	**0.0406494**
11	Signal Intensity	Median	**Last 500 Milliseconds**	**0.040223961**
12	**Fundamental Frequency f0**	Range	**Last 500 Milliseconds**	**0.040138624**
13	9th MFCC	Mean	Full Clip	**0.027013746**
14	**Fundamental Frequency f0**	Slope of the First-Last Line	Full Clip	**0.026454863**
15	**Fundamental Frequency f0**	Mean	**Last 500 Milliseconds**	**0.020519628**
16	12th MFCC	Mean	**Last 500 Milliseconds**	**0.016024721**
17	**Fundamental Frequency f0**	Slope of the Least-Squares Line	Last 200 Milliseconds	**0.015291915**

Performance acc. to Notion of “Best”

ROC AUC Score: 0.5134

Overall Accuracy Score: 88.6%

“Question” Acc. Score: 2.8%

“Not Question” Acc. Score: 99.9%

On Test Sample of 4000 clips:

	Actual: “Question”	Actual: “Not a Question”
Predicted: “Question”	13 *(True Positive)*	4 *(False Positive)*
Predicted: “Not a Question”	452 *(False Negative)*	3531 *(True Negative)*

Future Work to Improve the Model

- Train on more even proportion of questions vs. non-questions
- Use all ~105,000 clips that we had (requires a lot of time)
- Use only most important features to make the model simpler
- Experiment with better data orientation techniques
- Run a larger combination of Grid Search hyperparameters
    - Ex: Clip Sizes, Clip Periods, Learning Rates, Max. Depth, Max. # of Features, etc.
- Combine with word content analysis already created by the MTC group

Thank You