# Motivation

> ISOLET (Isolated Letter Speech Recognition) data set

> 150 Subjects spoke [A-Z] aloud, twice each

> 617 Features from audio signal processing, and pitch tracking normalized between -1 and 1 where possible

| f2 | f3 | f4 | f5 | f6 | f7 | f8 | f9 | f10 | ... | f609 | f610 | f611 | f612 | f613 | f614 | f615 | f616 | f617 | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -0.0930 | 0.1718 | 0.4620 | 0.6226 | 0.4704 | 0.3578 | 0.0478 | -0.1184 | -0.2310 | ... | 0.4102 | 0.2052 | 0.3846 | 0.3590 | 0.5898 | 0.3334 | 0.6410 | 0.5898 | -0.4872 | 1 |
| -0.1198 | 0.2474 | 0.4036 | 0.5026 | 0.6328 | 0.4948 | 0.0338 | -0.0520 | -0.1302 | ... | 0.0000 | 0.2954 | 0.2046 | 0.4772 | 0.0454 | 0.2046 | 0.4318 | 0.4546 | -0.0910 | 1 |
| 0.2124 | 0.5014 | 0.5222 | -0.3422 | -0.5840 | -0.7168 | -0.6342 | -0.8614 | -0.8318 | ... | -0.1112 | -0.0476 | -0.1746 | 0.0318 | -0.0476 | 0.1112 | 0.2540 | 0.1588 | -0.4762 | 2 |
| -0.0096 | 0.2602 | 0.2554 | -0.4290 | -0.6746 | -0.6868 | -0.6650 | -0.8410 | -0.9614 | ... | -0.0504 | -0.0360 | -0.1224 | 0.1366 | 0.2950 | 0.0792 | -0.0072 | 0.0936 | -0.1510 | 2 |
| 0.0946 | 0.6082 | 0.6216 | -0.1622 | -0.3784 | -0.4324 | -0.4358 | -0.4966 | -0.5406 | ... | 0.1562 | 0.3124 | 0.2500 | -0.0938 | 0.1562 | 0.3124 | 0.3124 | 0.2188 | -0.2500 | 3 |
| 0.0306 | 0.3546 | 0.4448 | -0.1022 | -0.4184 | -0.6388 | -0.4370 | -0.4396 | -0.6654 | ... | 0.6626 | 0.7350 | 0.3734 | 0.6626 | 0.3012 | 0.1808 | 0.2290 | 0.6144 | 0.3254 | 3 |
| -0.0102 | 0.2132 | 0.2018 | -0.6146 | -0.8380 | -0.8130 | -0.7240 | -0.8062 | -0.8996 | ... | 0.0526 | -0.0702 | -0.0350 | 0.0702 | 0.1578 | 0.1930 | 0.4562 | 0.4562 | -0.3860 | 4 |
| -0.1580 | 0.1764 | 0.1820 | -0.6378 | -0.8400 | -0.7280 | -0.6654 | -0.7978 | -0.7904 | ... | 0.2912 | -0.1646 | 0.1140 | 0.0126 | -0.0380 | 0.0886 | 0.2912 | 0.3670 | 0.1646 | 4 |
| 0.0424 | 0.2166 | 0.2124 | -0.4564 | -0.6200 | -0.7112 | -0.6602 | -0.6942 | -0.7920 | ... | 0.8868 | 0.8868 | 0.6792 | 0.6038 | 0.2264 | 0.7924 | 1.0000 | 0.9246 | 0.5284 | 5 |
| -0.0940 | 0.2868 | 0.2964 | -0.5326 | -0.7204 | -0.7518 | -0.7398 | -0.8482 | -0.8386 | ... | 0.6130 | 0.6130 | 0.6130 | 0.3226 | 0.6130 | 0.2904 | 0.5484 | 0.5162 | 0.3548 | 5 |

> Fanty & Cole (1991)

# Code Snippet

```
1  from pyspark.ml.feature import RFormula
2
3  formula = RFormula(
4      formula="class ~ .",
5      featuresCol="features",
6      labelCol="label")
7
8  output = formula.fit(trainingDF).transform(trainingDF)
9  output.select("features", "label").show(10)
```

> Spark's Special Sauce 🍔 is doable with R formatting using RFormula

> Supports ~ : + - .

> Allowed a fast vectorization of the 617 features in the ISOLET dataset

# Visualization

> RandomForestClassifier()

> F1: **83.3%**

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

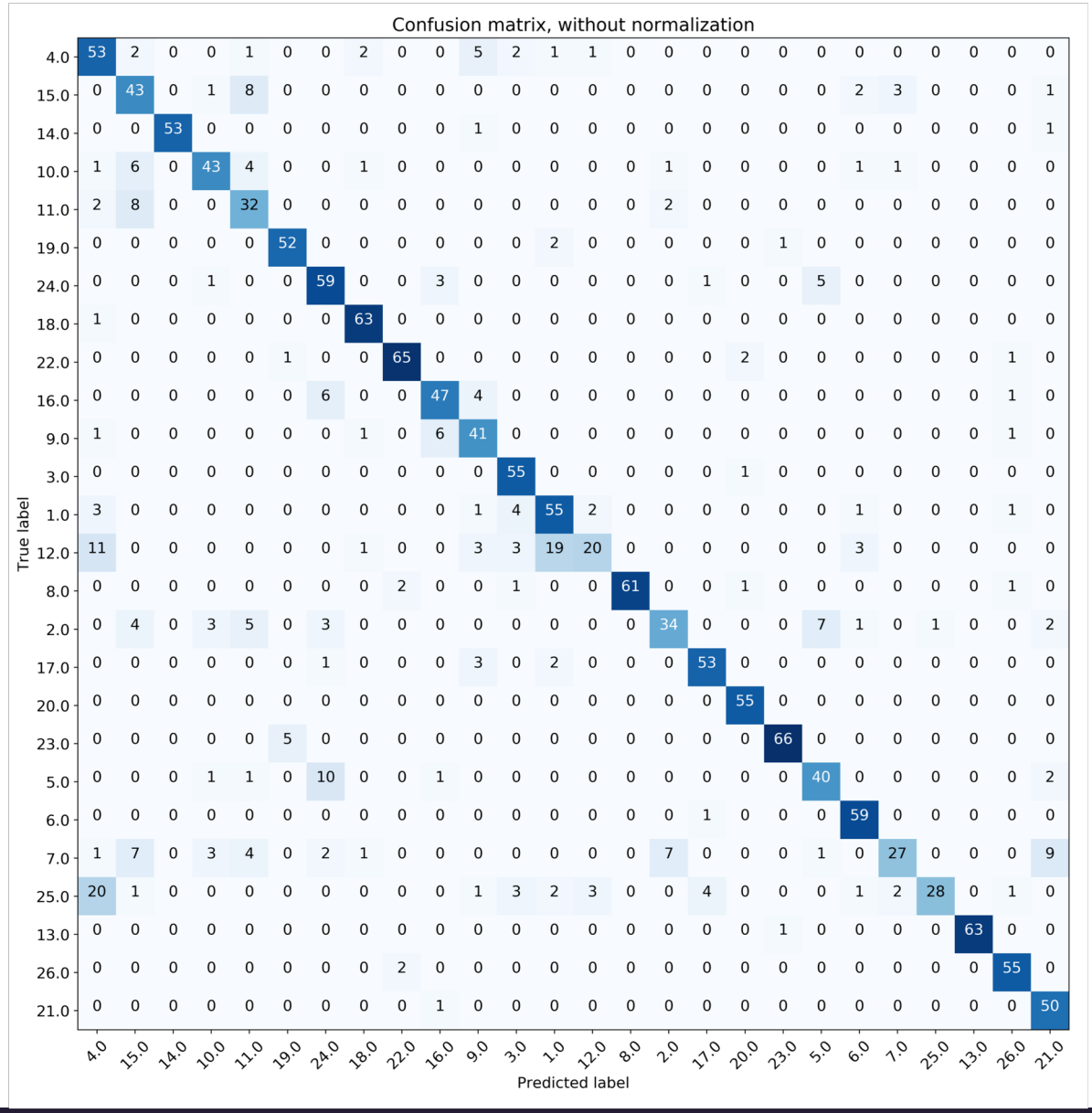
> Accuracy: **83.7%**

$$Accuracy = \frac{truepositives + truenegatives}{totalexamples}$$

> DecisionTreeClassifier() was worse

> F1: 50.5%

> Accuracy: 56.4%



# Visualization

> RandomForestClassifier()

> F1: **83.3%**

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

> Accuracy: **83.7%**

$$Accuracy = \frac{truepositives + truenegatives}{totalexamples}$$

> DecisionTreeClassifier() was worse

> F1: 50.5%

> Accuracy: 56.4%