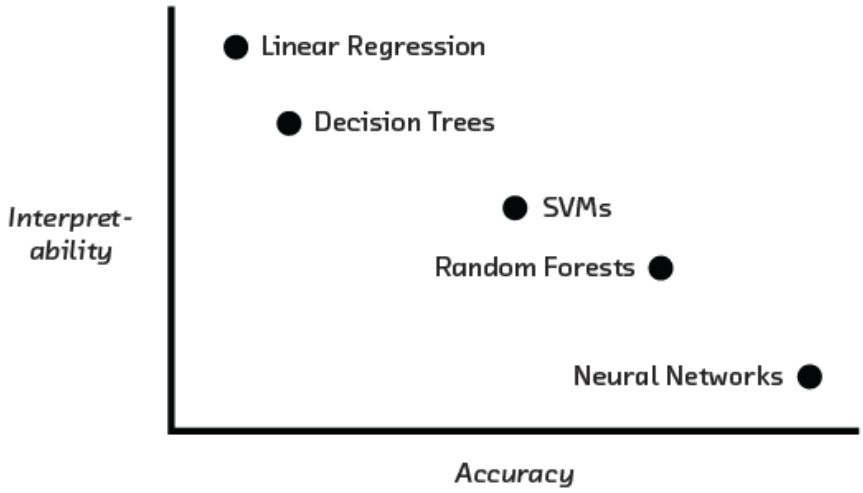
# Data Exploration Tools

## Overview

One of the greatest strengths of machine learning models is that they are able to pick up associations from data that humans are unable to parse. However, the lack of interpretability inherent in some of these models has led them to be referred to as “black boxes”. Neural networks are a prime example of this; the non-linear mapping inherent in neural networks makes it difficult to explain how the network came to a conclusion based on its data inputs.




## Interpretability Questions

One of the ways we can ameliorate this lack of interpretability is through understanding our data, specifically by asking questions like:

- Are there any outliers present? (How outliers can be handled is a complex topic, but the interested reader can delve into extreme value theory)
- Is the data homoscedastic or heteroscedastic (i.e., does variance in a variable grow with the size of the variable, or is it constant)
- Are the data normally distributed? Or do they follow other distributions (e.g., Poisson, Exponential)?
- Are missing values present? If so, will data be filled in via imputation or omitted?
- Is there any collinearity in covariates? (e.g., two separate variables with one representing temperature in Fahrenheit and the other temperature in Celsius)
- Are there any interactions between variables in our data? (e.g., relative humidity influences temperature, and vice versa)
- Is each data point independent? (i.e., is there any autocorrelation between data points)

These are very technical questions, but by better understanding our data we can usually build better models and gain additional insights that were previously hidden from view. One way of understanding our data is through the use of visualization tools and techniques. The strength of visualizations as a tool is that they replace cognition with perception. That is, instead of having to read through and interpret tables to highlight associations between data, visualizations help to uncover those associations with great ease, and often uncover new perspectives.

The cubist paintings of Pablo Picasso provide a helpful analogy. When most people look at a painting, they see a single perspective. However, Picasso specifically designed his paintings so that they had multiple interpretations, and by continuing to look at the painting you would start to see alternative perspectives within the same image. This is similar for our dataset; a single bar chart or scatter plot of our data set will only give us a low-dimensional representation of our dataset that has been filtered by our perceptive field, and only by seeing our data from multiple perspectives can we really start to understand and build higher-dimensional insights. This is where interactive visualization tools come in handy.

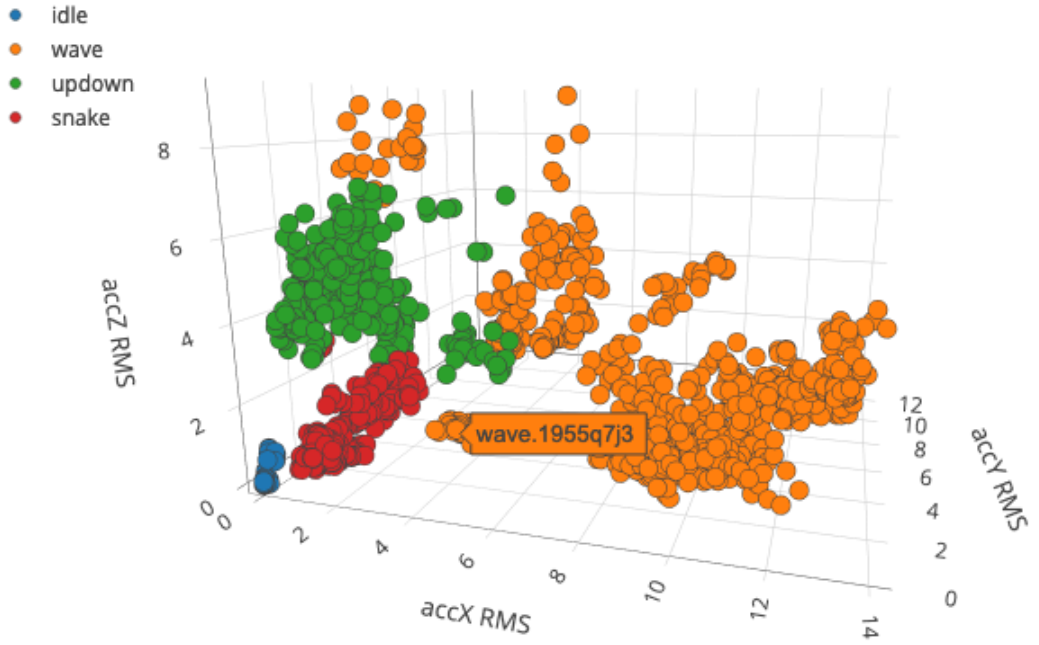
## Visualization Techniques: Dimensional Reduction

There are multiple techniques available to visualize high-dimensional data. These typically work by compressing the dimensionality of our data into a small representation that can be visualized on a two- or three-dimensional plot, as in Edge Impulse's Feature Explorer. Hence, these techniques are called **dimensionality reduction** techniques. These techniques can be linear (e.g., principal component analysis, non-negative matrix factorization) or non-linear (e.g., tSNE, UMAP).

As an example, we will look at Edge Impulse's Feature Explorer to see how misclassified examples can be spotted within a dataset. The Feature Explorer visualization shown below allows us to compress our data features into a three-dimensional image, and then highlight these images by their classification output. In this case, the classifier output is one of four different motions: "wave", "updown", "snake" and an "idle" value if there is no input. The visualization allows us to click on each of the samples and see the corresponding accelerometer waveforms of the sample. You could apply the same techniques to the compressed image or audio features.

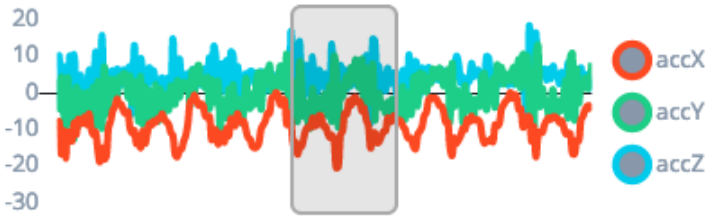
## Feature explorer (6,485 samples)

X Axis: accX RMS | Y Axis: accY RMS | Z Axis: accZ RMS

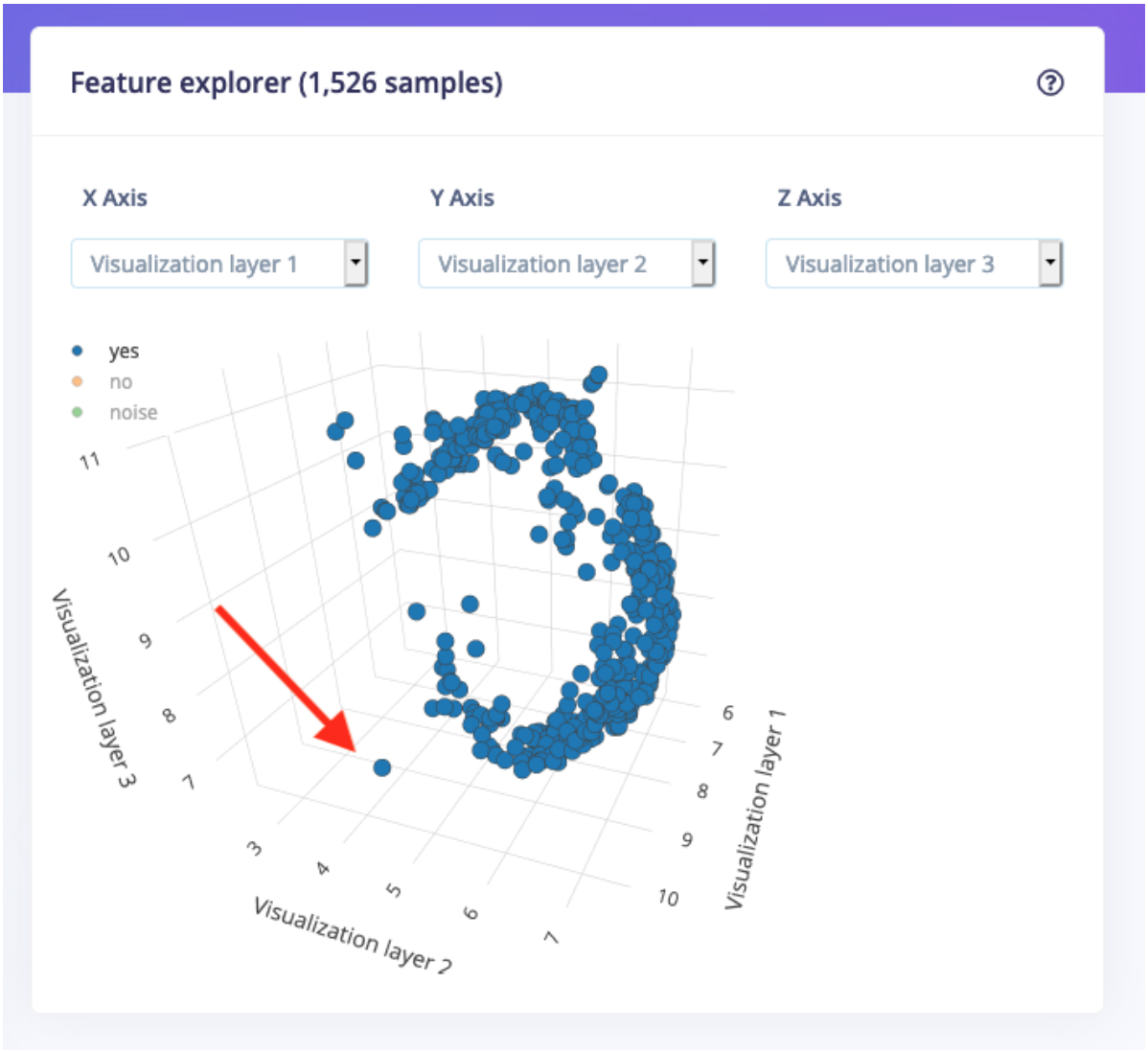



**wave.1955q7j3**
Window: 4368 - 6368 ms.
Label: wave

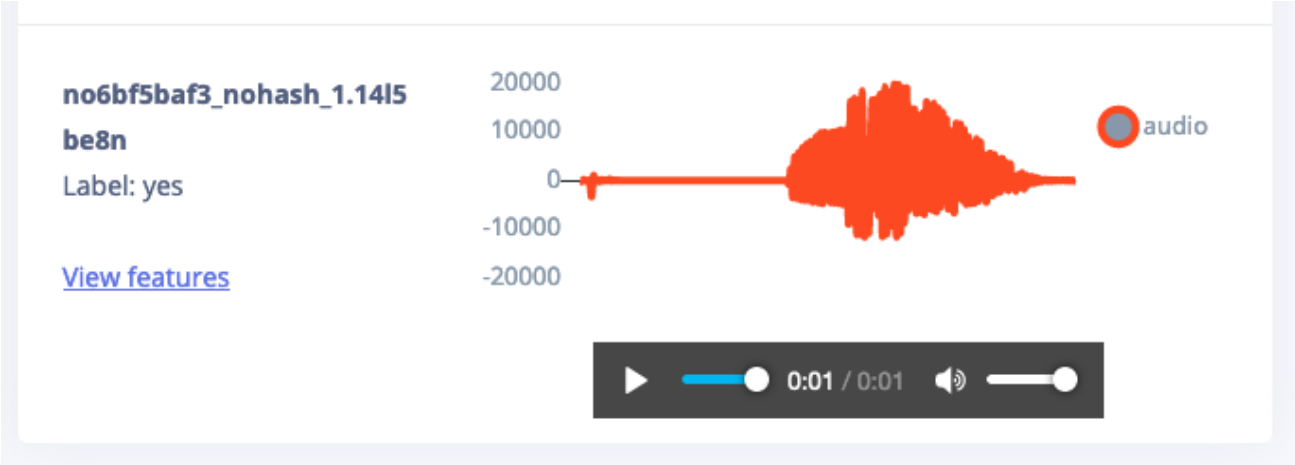
View features



To highlight how the Feature Explorer can be used, we will use an example of highlighting misclassified data samples. In the below image, we have filtered the classifier outputs to pick up only samples that correspond to the keyword “yes”. Upon clicking on the sample, we can listen to the audio sample to see what it sounds like.



Upon listening to the audio sample, it becomes clear to us that this audio sample is mislabeled, as the sample clearly sounds like “no” more than “yes”. This can be done with other samples to prevent incorrect classifications from reducing the performance of our machine learning algorithm.




## Additional Resources

Netron (model visualization)
Visualizing complex datasets in Edge Impulse
Tensorflow Embedding Projector
Tensorboard