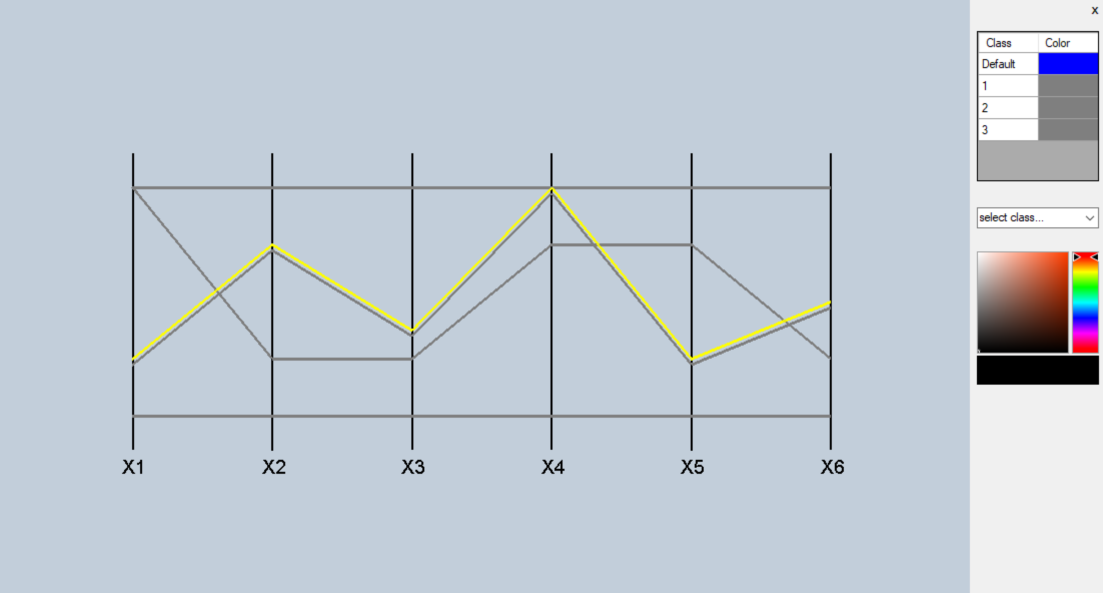
Data and Info Visualization - VisCanvas Exploration

Andrew Struthers, Gihane Ndjeuha, Kollin Trujillo, Nathan Chapman, Nick Haviland

January 2023

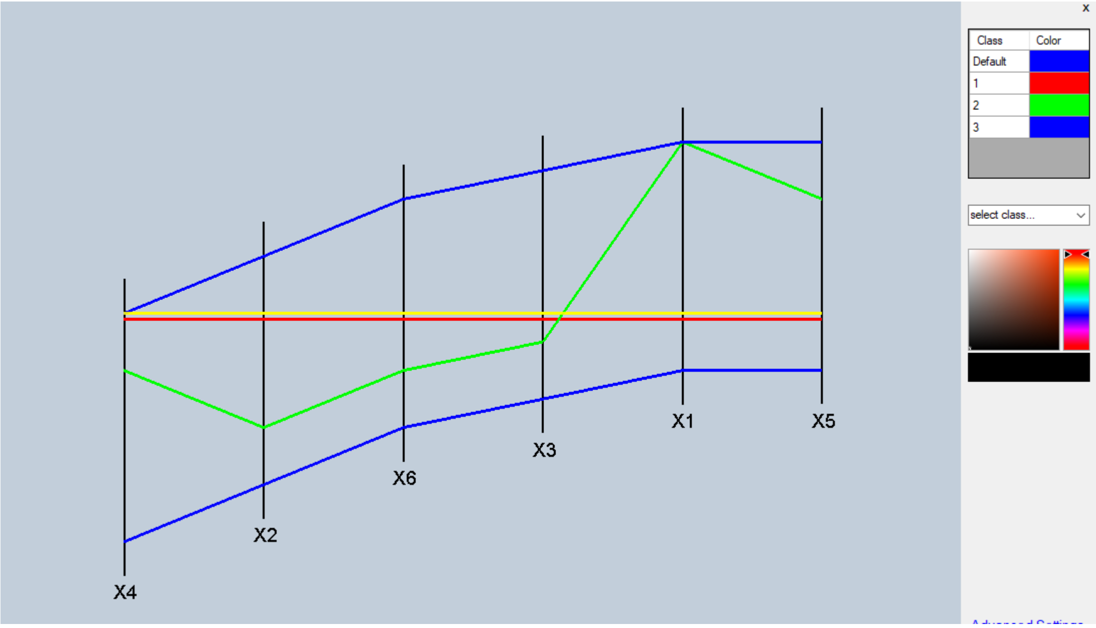
1 Part 1 - VisCanvas Introduction



Class	Color
Default	🟦
1	▨
2	▨
3	▨

Figure 1: Default dataset with zero formatting

As we can see in the dataset visualized above, the default view provided by VisCanvas is not very useful at helping us understand data. We can do a few things to help fix that issue. First off, we can provide colors to each of the classes. Here, we will assign red to class 1, green to class 2, and blue to class three. Then, to help us visualize the general trends in the data, we can select a single datapoint and then sort the dimensions such that the values of the dimensions of the selected datapoint are monotonically increasing from left to right. Doing this will sort the dimensions by changing their horizontal positions to hopefully clearly identify if there is a trend in the data, and it can help us sort and further separate the data. Then, we use the selection tools to select a single datapoint, and using the median or mean button, we can change the vertical position of each of the dimensions so that the datapoint selected is completely horizontal, and all other datapoints besides for the selected one will adjust relative to that change. This process turned the above dataset into what we can see below. This new visualization clearly shows the separation of each of the classes and is much more visually appealing than the default graph.

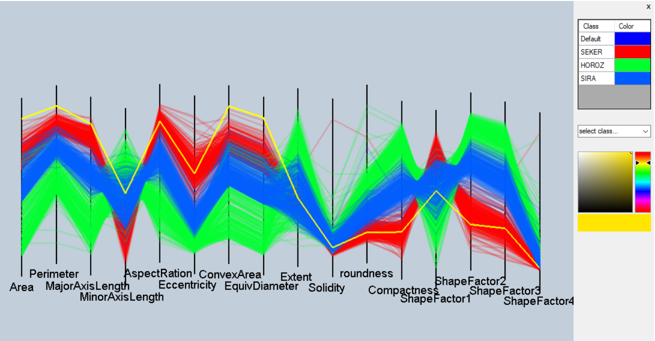


Class	Color
Default	🟦
1	🟥
2	🟩
3	🟦

Figure 2: Cleaned up, formatted sample data

2 Part 2 - Real n-D Data

2.1 Dry Bean Classification - Andrew

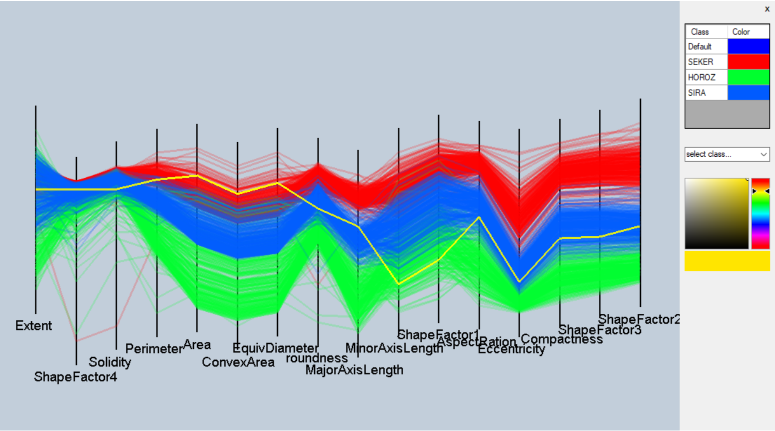


Class	Color
Default	🟦
SEKER	🟥
HOROZ	🟩
SIRA	🟦

Figure 3: Default visualization with three classes colored

The dataset I worked with consisted of 16-dimension classification data registered from high resolution images of beans. Initially, the dataset contained 13,611 images of 7 different bean classes, but I trimmed the data down to 3 classes with around 300 rows of data per class. This data came from UC Irvine Machine Learning Repository, provided by a generous donor in 2020. There are 12 dimensions and 4 shape forms of each bean calculated from the high resolution images. As we can see in the picture above, the initial data, after color coordinating each class, looks very rough. There are some patterns that we can start to see, but there is not an easily recognizable separation throughout the dimensions provided.

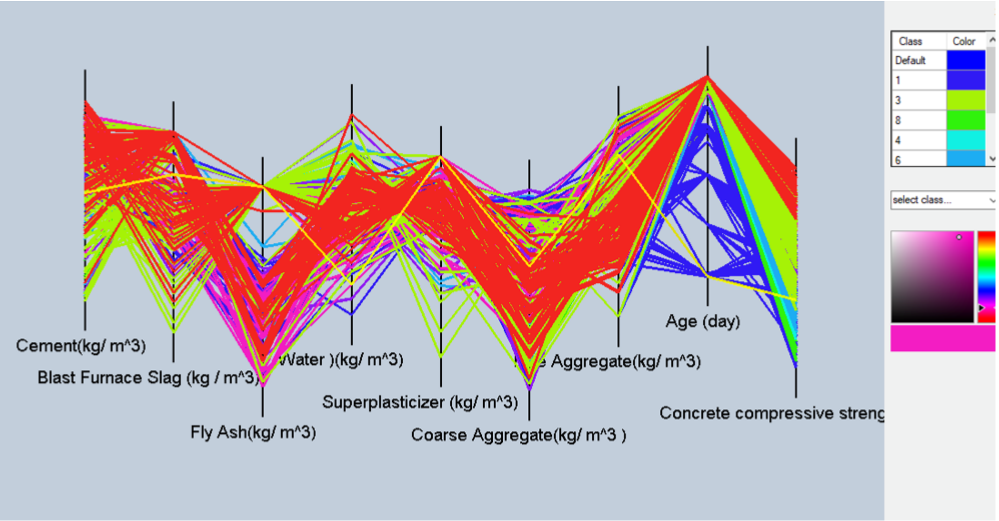
To start to paint a clearer picture, I first ordered the dimensions in a monotonically increasing pattern using an arbitrary row of the SEKER class. After flipping a few of the dimensions to further separate each of the classes, then changing some of their vertical positions so each axis label was more legible, I ended on this visualization. We can see that, in this organiztion, the SEKER, HOROZ, and SIRA classes are all fairly separated once we get past the three leftmost dimensions .



Class	Color
Default	🟦
SEKER	🟥
HOROZ	🟩
SIRA	🟦

Figure 4: Cleaned up, seperatable classes

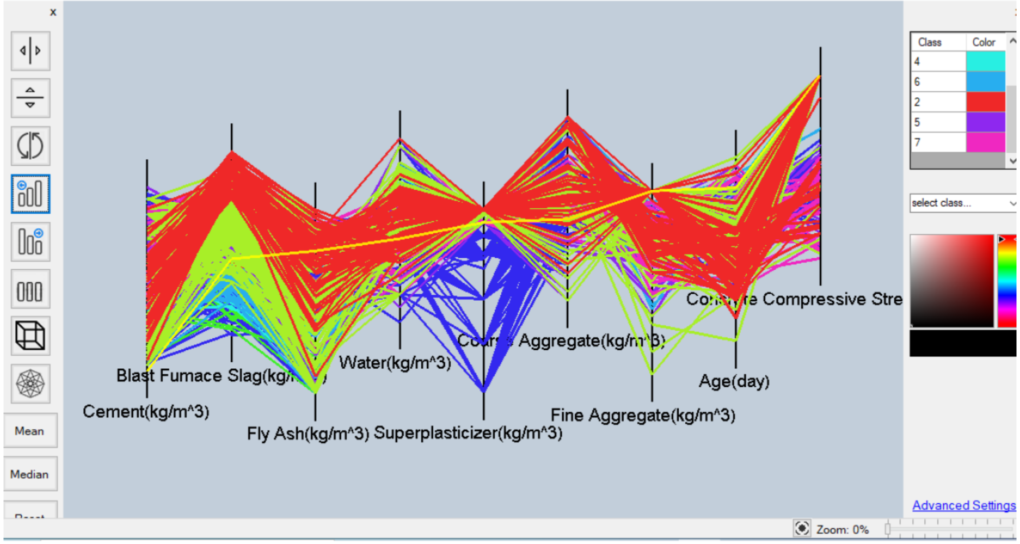
2.2 Concrete Compressive Strength - Gihane


Cement(kg/ m^3)
Blast Furnace Slag (kg / m^3)
Fly Ash(kg/ m^3)
Water)(kg/ m^3)
Superplasticizer (kg/ m^3)
Coarse Aggregate(kg/ m^3)
e Aggregate(kg/ m^3)
Age (day)
Concrete compressive streng
select class...

Class	Color
Default	🟦
1	🟦
3	🟩
8	🟩
4	🟦
6	🟦

Figure 5: Concrete strength dataset

This dataset has 1030 entries, 3 classes and 9 dimensions and 8 quantitative input variables. Each variables represent a component in the mixture of the Concrete Compressive Strength. The measurement is in kg in a m^3 mixture. We noticed that the classes 1, 2 and 3 have a higher quantity in the composition of the concrete.



Class	Color
4	🟦
6	🟦
2	🟥
5	🟪
7	🟪

Figure 6: Cleaned up and manipulated concrete compressive strength data

2.3 Raisin Dataset - Kollin

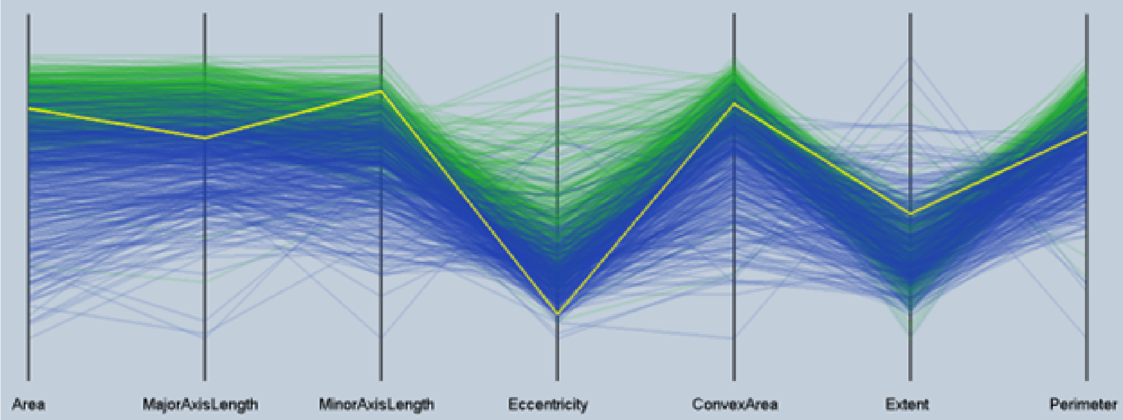


Figure 7: Raisin visualization

I chose the Raisin Dataset. This dataset consists of two classes of measurements from images of Kecimen and Besni raisin varieties. There are 900 total measurements of raisins with an equal distribution between the two varieties of raisins. The measures, or dimensions, of the data include a variety of morphological measurements. These measurements include area, major axis length, minor axis length, eccentricity, convex area, extent, perimeter, and then the classification into the two varieties. This makes the dataset 7-dimensional. The figure shown above is the default parallel line coordinate.

In an attempt to ease in the visualization of the dataset, I will shift and relocate some of the points, which showcases that the Kecimen raisins are typically larger than Besni raisins. The data is still a little difficult to decipher due to occlusion from data density, but I took conscientious effort to clearly separate the two classes of raisins.

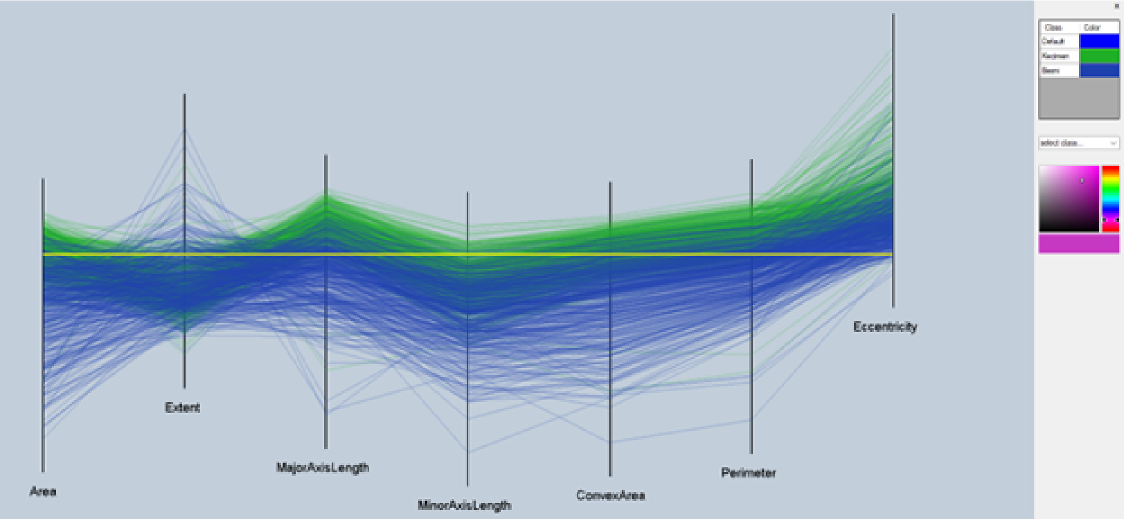
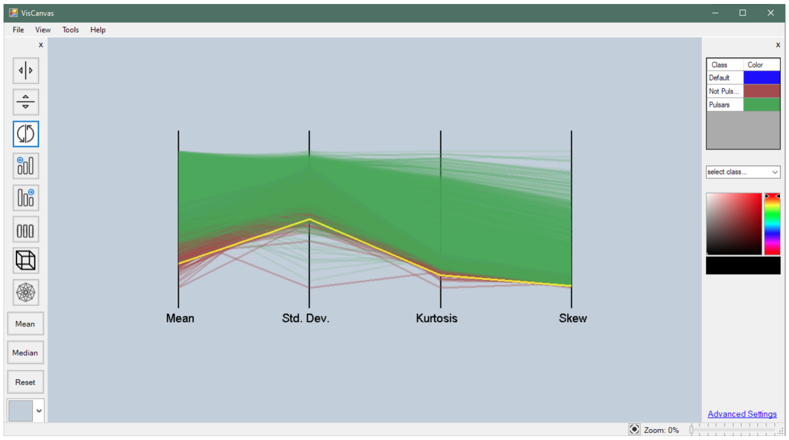


Figure 8: Cleaned up raisin data with classes showing

2.4 Pulsar Candidates in High Time Resolution Survey - Nate

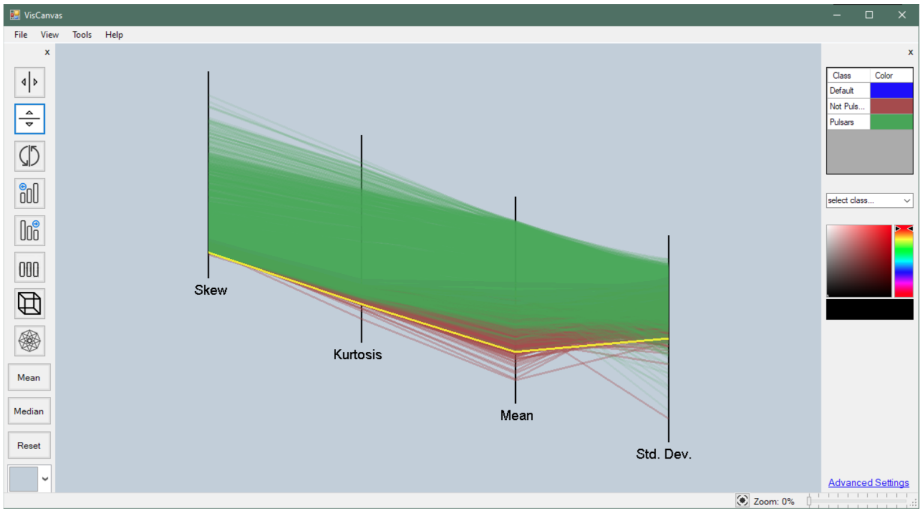


Class	Color
Default	🟦
Not Puls...	🟥
Pulsars	🟩

Figure 9: Raw pulsar candidate data

This data set is made up of 2000 (originally 17,898) entities from 2 different classes (pulsars in green and not pulsars in red), each of which has 8 dimensions. For this analysis, only a sample of the data was considered along with a further reduction in the dimensions. The discarded data was left out for the purposes of more efficient visualization. **It is important to note that the labels of the dimensions are not referencing the data itself, but rather the integrated profile of each potential pulsar.**

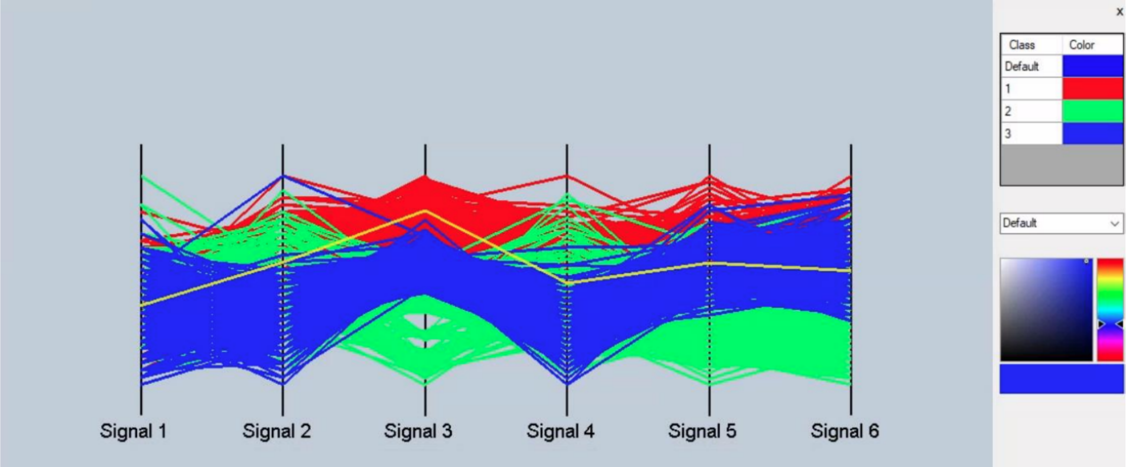
The strongest correlation between these data and the positive classification of a pulsar lies within the standard deviation of the integrated profile. The correlation can be seen in that most of the standard deviation data for positive classifications lies within a compact region. Similarly, and to a much more significant degree, a correlation can be drawn between the false classification of a pulsar and the skew of its integrated profile.



Class	Color
Default	🟦
Not Puls...	🟥
Pulsars	🟩

Figure 10: Manipulated correlation data

2.5 Wireless Indoor Localization - Nick



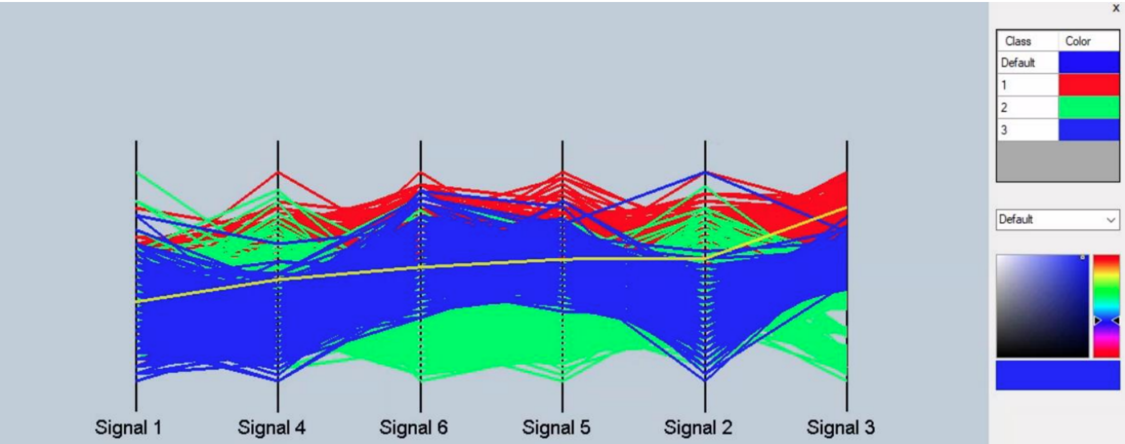
Class	Color
Default	🟦
1	🟥
2	🟩
3	🟦

Default

Figure 11: Raw localization dataset

I chose to use the Wireless Indoor Localization. Originally, this data set contained 2,000 observations from 4 different classes. I trimmed this data set down to include only 3 classes, and 1,500 observations. This data set measures Wi-Fi signal strength from 6 different smartphones, measured in an unmentioned unit. These 6 different smartphones and 3 rooms make this a 6-dimensional data set with three classes. The figure above shows the data set as it was originally imported into VisCanvas, with the only change being the color change for the three classes.

The figure below shows that room one, shown in red, tends to have a better overall signal strength than either of the other two rooms. It also shows that the highest overall signal strength tends toward the third smartphone.



Class	Color
Default	🟦
1	🟥
2	🟩
3	🟦

Default

Figure 12: Manipulated signal data