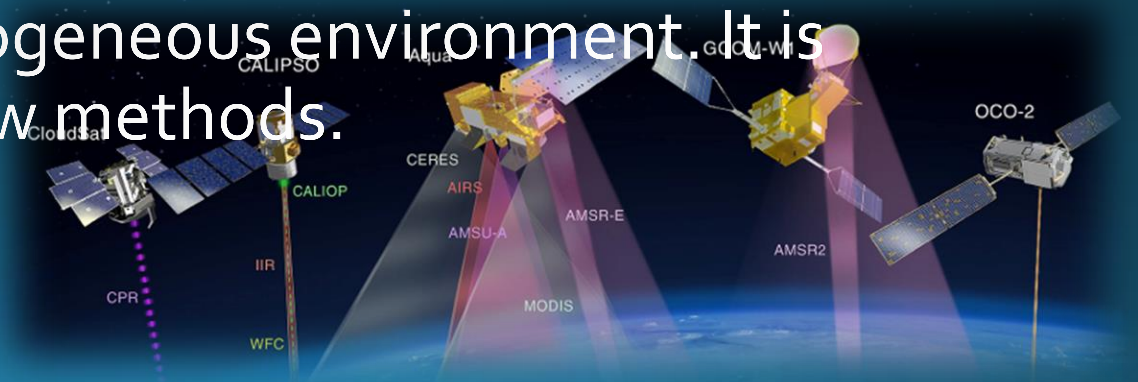
Croissant Antoine, DIA2

# Python for Data Analysis

DataSet : https://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29

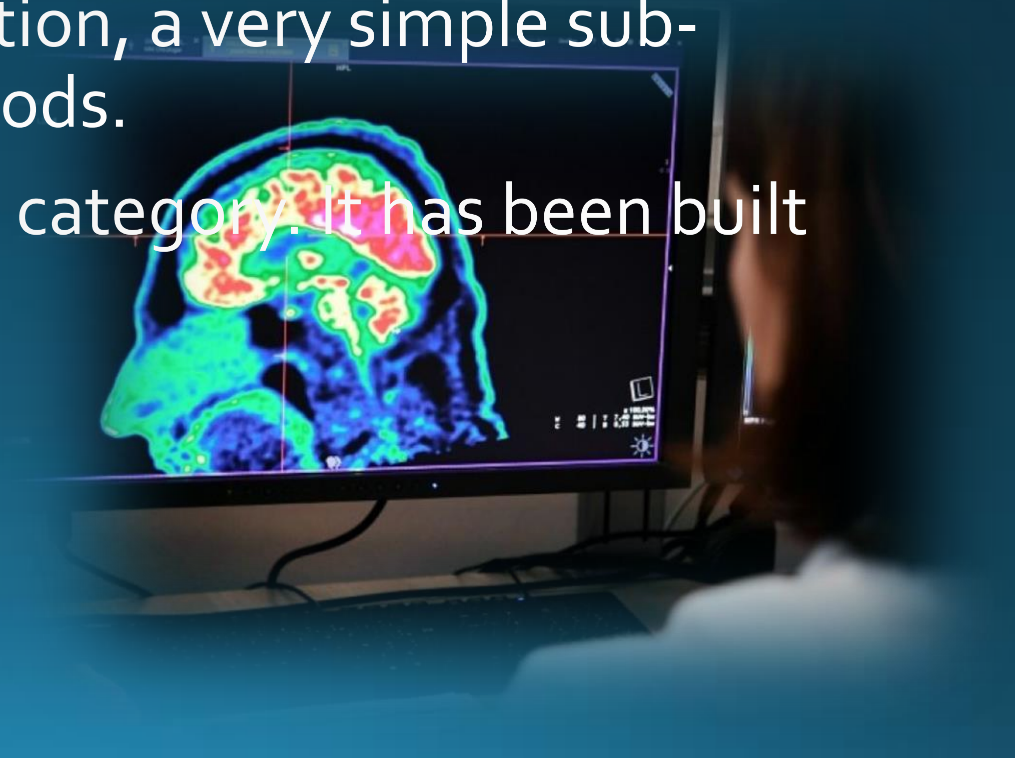
# The big picture

- It is the beginning of a new area where it is important to integrate the information which are gathered by remote sensing systems like satellite.
- In this context, remote sensing sends information with conventions, formats and information acquisition speeds.
- Classical statistical methods are based on homogeneous information and a small number of dimensions, therefore not suitable for images in a very heterogeneous environment. It is therefore important to develop new methods.



# Modern approach of the problem

- Recent successes in the application of artificial intelligence to images, show us the way forward.(self-drive car, interpretation medical images…)
- People have extracted from this information, a very simple sub-image, to start studying these new methods.
- So, the data set in our study is inside this category. It has been built such that it is simple to handle .



# Purpose

- The general goal is to recognize from the images provided by several satellites the type of terrain.

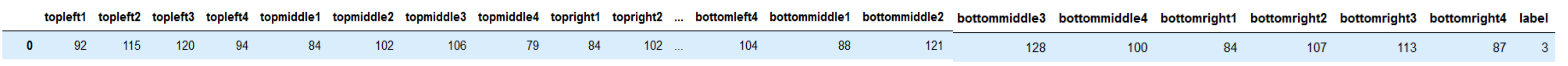
- Specifically, the aim is to predict from 4 spectral images of a terrain, the type of terrain.

- We will try to test several models and take the one that best predicts the test data set.

# Understanding the data

- So, we are given a dataset to train the data, the one on which we will apply the different models, but also a test dataset on which we will test our models, to be able to get an accuracy and compare our model
- On one line we have 37 columns, 36 of which correspond to pixels that are in 8-bit binary word (from 0 to 255) and the last column corresponds to a number indicating the classification label of the central pixel.

- To better understand, I added a name for each column to understand what that column corresponded to. Here's what it looks like on the first line



| | topleft1 | topleft2 | topleft3 | topleft4 | topmiddle1 | topmiddle2 | topmiddle3 | topmiddle4 | topright1 | topright2 | ... | bottomleft4 | bottommiddle1 | bottommiddle2 | bottommiddle3 | bottommiddle4 | bottomright1 | bottomright2 | bottomright3 | bottomright4 | label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 92 | 115 | 120 | 94 | 84 | 102 | 106 | 79 | 84 | 102 | ... | 104 | 88 | 121 | 128 | 100 | 84 | 107 | 113 | 87 | 3 |

(1)

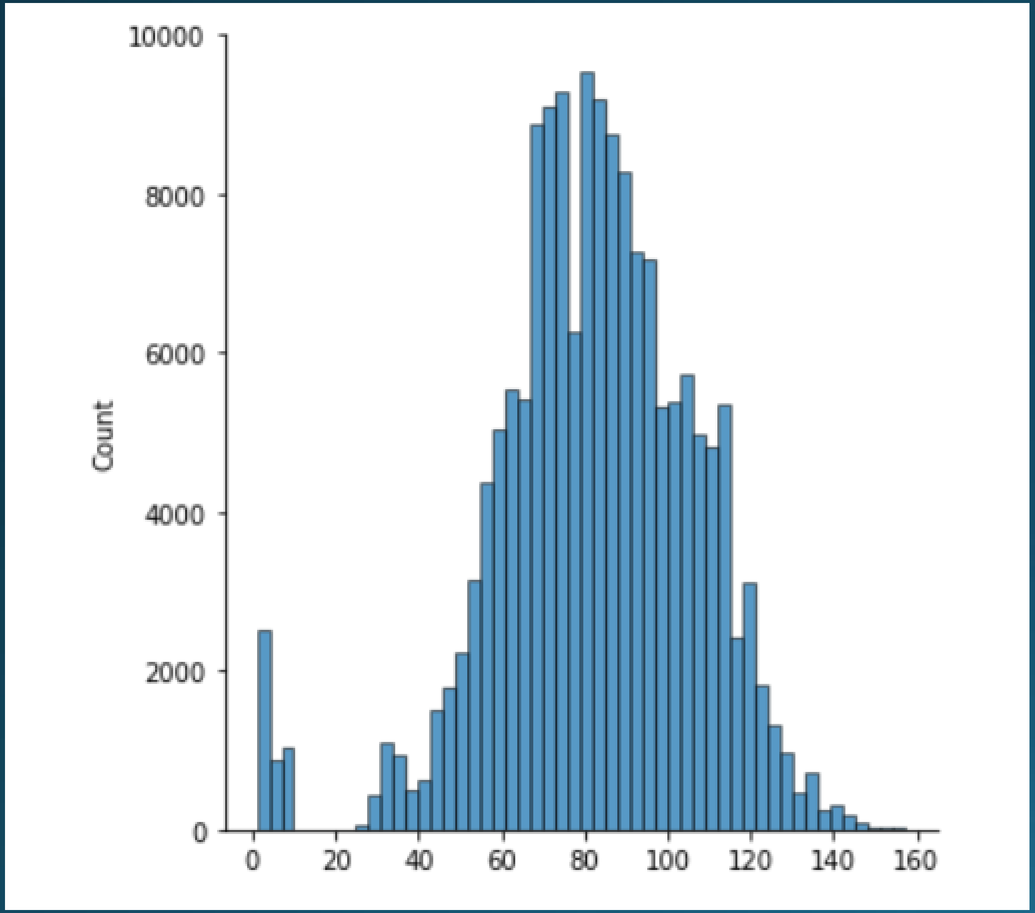
- Here is a pattern to better understand

| | | |
|---|---|---|
| TOP<br>LEFT | TOP<br>MIDDLE | TOP<br>RIGHT |
| MIDDLE<br>LEFT | MIDDLE<br>MIDDLE | MIDDLE<br>RIGHT |
| BOTTOM<br>LEFT | BOTTOM<br>MIDDLE | BOTTOM<br>RIGHT |

(2)

- Each line has four spectral bands represented as above (2), so with 9 pixel, which gives 4x9=36 columns, 37 with the label. But the data are classified as follows: Look at the first image (1).

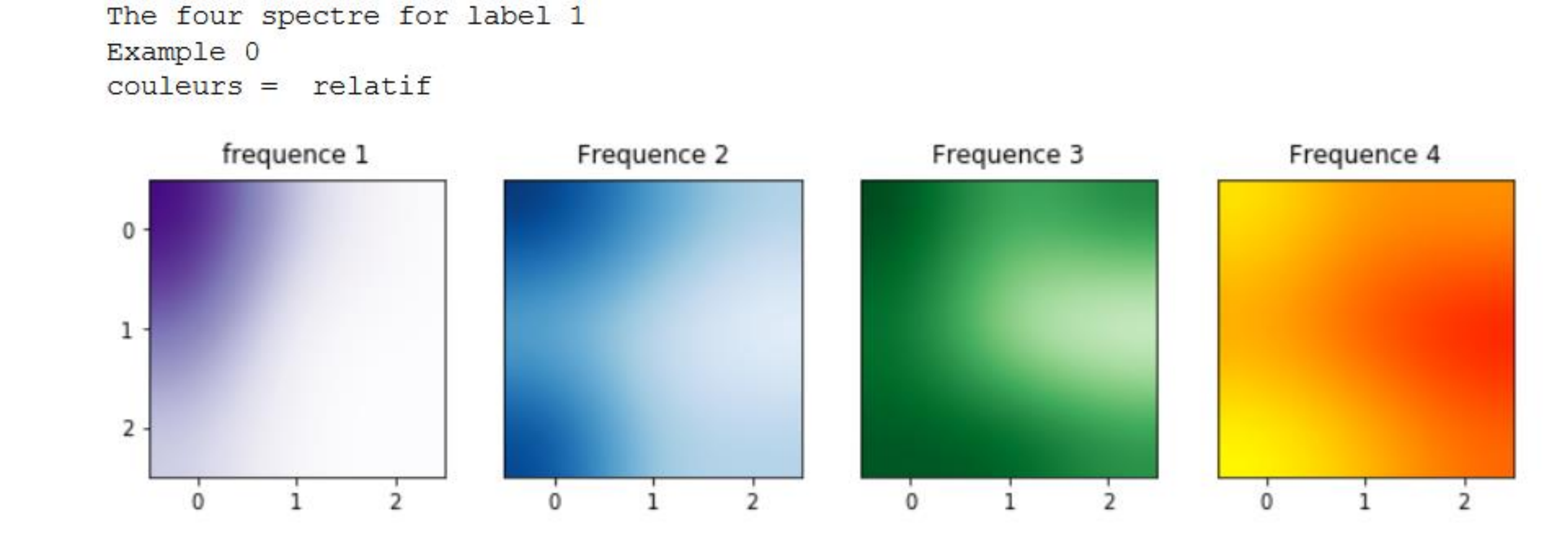
# Distribution of data



Intensity of color is given by a number between 0 and 255

This graph shows that the images have a Good contrast.  Not the maximum because all intensities are below 147

# Human Representation of the data



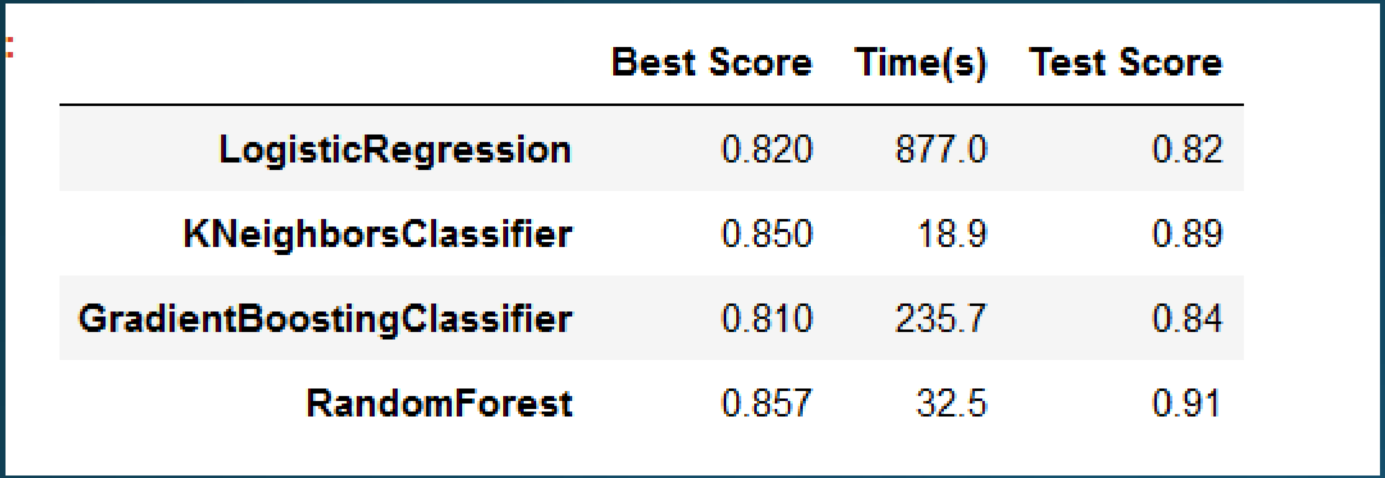
```
The four spectre for label 1
Example 0
couleurs =  relatif
```

The colors have been chosen for a better intuition,  the frequencies are not the one of the colors

Given the python color scheme, the data are represented with maximum contrast of the image

We use bicubic spline to create an illusion a smothness of the data to facilitate their intuitiveness

# 4 types of classifiers tried



| | Best Score | Time(s) | Test Score |
|---|---|---|---|
| **LogisticRegression** | 0.820 | 877.0 | 0.82 |
| **KNeighborsClassifier** | 0.850 | 18.9 | 0.89 |
| **GradientBoostingClassifier** | 0.810 | 235.7 | 0.84 |
| **RandomForest** | 0.857 | 32.5 | 0.91 |

Random Forest give us the best score. It is not unexpected. This method is difficult to overrun.
Only deep learning can do better

# Conclusion

- As said in the documentation accompagning the data, the 6 categories of terrain can be easly discriminated with these 4 times 9 pixels. This is very few pixels compared to the type of images that we can get from today sensors (60 mega pixels cameras and up).
- The fact that we can identify the type of terrain from only 9 pixels in 4 colors, means that we will be able to interpret  large images with a lot of details.
- As usual , random forest is a very efficient method, but of course modern deep neural networks should be even more powerful, but only needed for more complex images