# Accomplishments

## Download and decompress the seizure prediction dataset. (At least 2 hours!)

The dataset comprises seven large gzipped tarfiles files ranging from about 3 to 15GB. Downloading takes several hours. Decompress them to seven subdirectories using 7-zip and the following procedure:

* Open the file in 7-zip
* Double click on it. This takes 5-10 minutes and seems to mount the tarfile.
* The generic icon will now turn to a folder icon. Select it and press the extract button.
* Select the extraction destination as the directory in which you’re assembling the dataset and decompress. This actually extracts the folder and files to the target directory and takes another 5-10 minutes.

The dataset comprises a single directory with seven subdirectories Dog\_1…5 and Patient\_1..2, each of which has hundreds of .mat formatted files representing 10 minutes of captured EEG data for a specific epileptic individual.

## Write a program to tabulate basic data set properties. (3 hours)

Wrote a program to examine the dataset files and tabulate some statistics in order to gain some insight into its properties.

* Interictal files are those collected at a time distant from a seizure event.
* Preictal files are those collected at a time ranging from five minutes to 1 hours and five minutes before a seizure event.
* Test files are those which require classification as preictal or interictal.

Summary statistics for the data are as follows:

|  |  |
| --- | --- |
| Total Interictal Files | 3766 |
| Total Preictal Files | 301 |
| Total Test Files | 3935 |
| Total Files | 8002 |
| Data Bytes | 114 (GB) |

The following table summarizes details for the individual subdirectories.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dir | Interictal Files | Preictal Files | Test Files | Electrodes | Samples | Length (S) | Frequency (Hz) | File Bytes (MB) | Dir Bytes (GB) |
| Dog\_1 | 480 | 24 | 502 | 16 | 239766 | 600 | 399.61 | 7.67 | 7.72 |
| Dog\_2 | 500 | 42 | 1000 | 16 | 239766 | 600 | 399.61 | 7.67 | 11.8 |
| Dog\_3 | 1440 | 72 | 907 | 16 | 239766 | 600 | 399.61 | 7.67 | 18.6 |
| Dog\_4 | 804 | 97 | 990 | 16 | 239766 | 600 | 399.61 | 7.67 | 14.5 |
| Dog\_5 | 450 | 30 | 191 | 15 | 239766 | 600 | 399.61 | 7.2 | 4.83 |
| Patient\_1 | 50 | 18 | 195 | 15 | 3000000 | 600 | 5000.00 | 95.8 | 25.2 |
| Patient\_2 | 42 | 18 | 150 | 24 | 3000000 | 600 | 5000.00 | 147 | 31 |

Some observations:

* The individual .mat files can vary in size by a factor of roughly two for the same individual. It seems that Matlab stored the integer valued sample data as 16 bit ints in most cases, but if any single sample magnitude exceeded 32767 it promoted the storage to 32 bit ints. The file bytes column represents average file size, and the dir bytes column represents the sum of all file sizes in the directory.
* The number of electrodes varies between individuals, even those of the same species. The learning framework must accommodate this.
* The ratio of preictal to interictal samples is quite low. This reflects the fact that seizures are rare events that are difficult to capture while the vast majority of time is spent during interictal periods. Even this low ratio is much higher than the true ratio of preictal to interictal time for an individual.
* I confirmed that the length, sampling frequency, number of samples, and number of electrodes are consistent for every file in a subdirectory. (This takes quite a while to run, because it opens all 8002 mat files and so has to read 114GB of data.)

## Develop a class to act as an interface to the dataset

The idea here is to hide the application specific complexity and enormity of the dataset by providing a class with a simple interface to act as a proxy to the data.

This class should provide:

* A constructor that accepts a path to the dataset.
* A method to select the individual and return the electrode count for that patient.
* A method to partition training and validation subsets.
* A method to set the number of samples per pattern.
* A method to retrieve a specified number of training, validation, or test patterns.