SECOND MILESTONE:

SUBJECT 4 - Big Data and Data Mining techniques applied to atmospheric tornadoes characterisation

1) Summary of events since the first milestone:

Since the last milestone, we manage to reach many of our goals. At the time, we did not have data, and we were only speculating on the kind of data we would collect. But now we have made good progress.

During several weeks of research to obtain data on Tornadoes, we contacted thirty different people and several organizations, such as NASA, NOAA, or AMS. Even though, a minority answered us, it was very rich in information. NASA and NOAA have guided us with interesting and promising data. The difficulty was that we ended up with a mass of data spread over several websites in hundreds of different projects and hundreds of files. Thus, the study of all promising cases of relevant data, has required a lot of time and investment.

On the first hand, during this bibliographic work, we were stuck on a thorny issue: the tornadoes have so small dimensions and durations of life that the sensors are unfortunately not precise enough and their frequency of sampling too high (every 6 hours at best) to study the phenomenon in meaning conditions.

But on the other hand, by searching through this mass of data, we found significant data on cyclones. As cyclones are quite bigger and much less ephemeral than tornadoes, we saw a new mean to reach our aim for this study. So, we contacted our referent teacher to ask if we may study cyclones instead of tornadoes because the former might be much more relevant than the latter. It turns out that our professor had come to the same conclusion, doing some research on his side.

During our interview with our referent teacher last week, we explored the data available on cyclone. In doing a brief search, we found that these data contained exactly what we expected to characterize the types of cyclones (shear rates, temperatures, pressures, dimensions, wind speeds, cyclone classes…).
At the end of the interview, we set several goals for the next meeting. We must therefore translate the documentation, create a clean and streamlined database and determine from the data: the viscosity, the Reynolds number and the compressibility of each cyclone.

2) The objectives in progress:

The documentation (ReadMe) on the cyclones database is long and quite complicated. Currently, we have translated about 90%. According to the documentation, there are several types of data files containing various informations. The most relevant are the "BestTrack" and "Diagnostic" files. Here is the main data:

- BestTrack: contains a dozen relevant variables and mainly ***the type of cyclone (from depression to hurricane)*** allowing the categorization of cyclones, which is our aim

- TcDiag: contains the most relevant data allowing to calculate ***the variables useful for the categorization of cyclones*** according to their compressibility, their laminar or turbulent regime... The calculated variables will be mainly the Reynolds number, shear rate, viscosity, and Mach number.

BestTrack files are very important because they list cyclones by type. TcDiag files are also very important because they contain information to determine variables important for the categorization of cyclones (shear rate, Reynolds number, viscosity, compressibility).

The idea would be to merge these two types of files into one. In theory it seems simple. But we encounter a lot of technical problems. Records are not easily crosstable. Many cyclones contained in BestTrack are not mentioned in TcDiag files. It will therefore be at the cost of a significant loss of data that we will be able to create a clean and useful database for the categorization of cyclones.

3) The last step:

Once the data is clean, we will use Data Mining Techniques to perform classifications. We will first try to categorize cyclones only with BestTrack files. This will allow us for example to know the type of cyclone according to variables such as size, speeds, pressures and others.

Lastly, crossing the two types of files will allow us to classify cyclones according to their compressibility, their viscosity, their turbulent or laminar regime. Work will therefore be more difficult, but also very instructive.