# Machine Learning Project

## Subject

The goal of this project is to study the following dataset of Tweets:
https://www.kaggle.com/crowdflower/twitter-airline-sentiment

In groups of 2 or 3 (or alone), it may involve studying:

- Exploration Data Analysis (missing/null data, statistics, visualizations, ...)
- We can analyze the time series (number of tweets per day, per week, etc., depending on the location, ...)
- We can create new variables and analyze them: sentence size, do the sentences contain the words good, bad, love, etc.
- Use of SKlearn, Automl (search for the best hyper-parameters), Pycaret (try several machine learning algorithms with fixed hyper-parameters), Sktime/Facebook Prophet (time series prediction)
- Use of SK-learn TFIDF (or Universal Sentence Encoder) to create a matrix of variables from text
- Use Spacy if needed to pre-process the text
- Search for optimal variables
- Search for the best hyper-parameters (Grid-search-cv, Optuna, Hyperopt)

Each modification is an experiment to analyze.

# Deep Learning project

## Subject

On the same dataset as before, we can train several models to predict the sentiment of sentences (inspired by https://github.com/rodgzilla/teaching/blob/master/deep_nlp/lesson_05.md et https://deepnote.com/project/Efrei-M2deeplearning-a4WZZbhYS42oQQCVma74bg/%2F6.%20recurrent_networks.ipynb):

- an MLP (Dense layers)
- a CNN (1D convolution layers + dense layers)
- an LSTM (LSTM layers + dense layers)
- average the Spacy word embeddings to have one vector per sentence + train a random forest or other
- use Universal Sentence Encoding to have a vector per sentence + train a random forest or other
- add the size of the sentence as a feature
- check if the dataset contains the same number of examples per class otherwise balance it
- compare with a HuggingFace model

The Embedding layer in Keras is a matrix of shape embeddings: number of different words (vocabulary size) and number of variables to represent a word.

Each modification is an experiment to analyze.

Be inventive!

If needed: benjamin.maurice@telecomnancy.net