# Machine Learning MC886

University of Campinas (UNICAMP), Institute of Computing (IC)
Assignment #1, 2019s2, Prof. Sandra Avila

## Objective

Explore **linear regression** alternatives and come up with the best possible model to the problems, avoiding overfitting. In particular, predict the traffic volume of the Metro Interstate from their attributes (e.g., temperature, clouds, weather) using the Metro Interstate Traffic Volume dataset (`https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume#`).

## Activities (10 pts)

1. (0.25 pts) Split the data for providing your results and avoid overfitting.
   Keep in mind that friends don't let friends use testing data for training :-).
2. (3 pts) Perform Linear Regression. **You should implement your solution** and compare it with `sklearn.linear_ model.SGDRegressor` ("linear model fitted by minimizing a regularized empirical loss with SGD"[1]). What are the conclusions?
3. (0.75 pts) Plot the cost function vs. number of iterations in the training set and analyze the model complexity. What are the conclusions? What are the actions after such analyses?
4. (1 pts) Use different Gradient Descent (GD) learning rates when optimizing. Compare the GD-based solutions with Normal Equation. **You should implement your solutions**. What are the conclusions?
5. (4 pts) Prepare a 4-page (max.) report with all your findings. It is UP TO YOU to convince the reader that you are proficient on linear regression and the choices it entails.
6. (1 pts) You should provide a single Jupyter notebook with your solution (in Python 3 code).

## Dataset

The Metro Interstate Traffic Volume dataset contains the traffic volume and 8 attributes of 48,204 examples.

### Dataset Information:

- There are 8 attributes as follows:

  1: **holiday**: US National holidays plus regional holiday, Minnesota State Fair
  2: **temp**: average temp in kelvin
  3: **rain_1h**: amount in mm of rain that occurred in the hour
  4: **snow_1h**: amount in mm of snow that occurred in the hour
  5: **clouds_all**: percentage of cloud cover
  6: **weather_main**: short textual description of the current weather
  7: **weather_description**: longer textual description of the current weather
  8: **date_time**: hour of the data collected in local CST time
  target **traffic_volume**: traffic volume

- The data is available at `https://archive.ics.uci.edu/ml/machine-learning-databases/00492/Metro_Interstate_Traffic_Volume.csv.gz`

[1] `http://scikit-learn.org`

## Deadline

Monday, **September 2** in the beginning of the class, 7 pm.
Penalty policy for late submission: You are not encouraged to submit your assignment after due date. However, in case you did, your grade will be penalized as follows:

- September 3 7pm : grade * 0.75
- September 4 7pm : grade * 0.5
- September 5 7pm : grade * 0.25

## Submission

On the deadline day, bring your 4-page printed report. The template for report is available at `https://www.dropbox.com/s/nc6d89otr8ekvjd/report-model.zip`. Please, print on both sides of the page.
The report should be written in Portuguese or English.
**Submit a zip file, with the code and the report (PDF file), via Moodle**.
This activity is NOT individual, it must be done in pairs (two-person group).