# Assignment - 2 : Data Analytics with Python

## Supervised Learning

Exams are over, you don't want to do anything but take a sigh of relief or maybe Netflix and chill (read "watch a movie"). How do you decide from the mammoth directory available on DC?. You pick up one that sounds interesting, then google the IMDb rating. Just to be sure, since you don't waste your time on something like "Justice League" (no offense). How about being able to decide if you want to put your money into watching a Movie beforehand?. How about predicting the rating yourselves even before a movie is released?

**What do you need to do?**

You have been taught basic supervised learning techniques in the class. Now as a Data Analytics student, you need to figure out a way to **predict movie rating using the regression**.

**How will you do it?**

IMDb Dataset is available on the internet at https://datasets.imdbws.com/ and corresponding documentation at https://www.imdb.com/interfaces.
However, you can use any dataset you want.
If you're feeling adventurous, you're free to scrape (python is excellent at it) the data and build a custom dataset. (Not Recommended)
Important Tips:

1. Do not be overwhelmed by the verbosity of the dataset. One of the important task one has to learn is to pre-process, cleanse. You have been taught enough python that you can juggle with the dictionaries, tuples etc with ease. You can read about some of the pre-processing techniques, I shall be covering them along the course of this assignment. Also, **Pandas will simplify your life**, use it to its full potential.
2. Use internet indiscriminately, use libraries but understand how's stuff works.
3. Do not go beyond what has been taught, you can go deeper but not wider. So, if you're using some form of regression, you're allowed to learn and exploit its variants. In other words, "No deep learning stuff".

**Evaluation:**

These are the movies that you have to finally test your model on (besides your own test set):

1. Shazam
2. Pet Sematary

3. The best of Enemies
4. Hellboy

The first three movies are scheduled to release on 5th April, so you shall submit the ratings predicted along with your code by **4th April 2019, 23:59**. ( 20 days )

**Evaluation Criteria:**

**Ratings are supposed to be the range (0,10).**

**Score <= {10 - Avg ( abs(YourPrediction - ActualRating*) ) } x 100 ;**

**Avg is over 4 predictions.**

*Actual Rating = After a cool down period of 7 days on IMDb post-release.

NOTE: This won't be the only parameter for scoring. Other criteria might include, pre-processing steps, Train and Test Accuracy etc. We shall tell you soon.

**Suggestions:**

- Think something out of the box while working on the dataset ( preprocessing), you can add extra supporting attributes like "Popularity of Actors" in terms of the number of followers or "No of Hashtags on Twitter for the movies" etc. Basically, think of what can actually affect a movie's rating and how to quantify this observation.
- Use Jupyter Notebook with added notes and plots, just like you're writing a blog of your progress.
- Results may not be very promising with all the vanilla techniques that you've been taught. However, that's not our only objective. Also, regression task is a nuisance.
- DO NOT HESITATE to clarify your doubts. Since much of the stuff would be new to you, ask us anything, whenever you're stuck. You've our contacts.
- This could be a little(highly) overwhelming but know what :

"**No fine work can be done without** concentration and self-sacrifice and **toil** and doubt."

I assure you this is something that can't and shouldn't be done in the last minute