# SOSC 4300/5500: Assignment 1

March 6, 2024

Welcome to the first assignment! This is the link to our assignment: `https://www.kaggle.com/datasets/jtentor/blogfeedback-data-set`

Your goal is to predict the number of comments a blog will have, depending a list of 280 features constructed from texts and the metadata of the blogs. This assignment gives you an hands-on experience of using statistical models for prediction. There is no unique solution to this task. You are free to explore:

- Different models (linear/logistic regressions, LASSO, Ridge, Tree and Random Forests, or whatever models you want to try).
- Different set of predictors. Read the dataset description or use a data-driven way to find predictors that can improve out-of-sample prediction tasks.

We will use Kaggle to evaluate your prediction performance. Kaggle is a data science learning platform, on which anyone can create prediction challenges and invite other teams around the world to compete for the same prediction task. For each challenge, there is a leaderboard, from which you can easily see which team is achieving the best performance at any time. It's a great place for you to improve your machine learning skills and get some ideas for your projects.

Choose a group leader who will submit the result. You need to register a Kaggle account if you do not have one yet. Please use the name Group XXX so that we know who you are.

- Each day, you can submit **10** predictions on Kaggle to evaluate how well your algorithm performs.
- Click "Submit Predictions" button and you will be directed to a page to upload your output.
- Drag and drop the file containing predicted outcomes to Kaggle (as described in the previous section).
- Kaggle will calculate the performance metric for you. This time, it is MSE, and you should try to minimize that. What's cool about Kaggle is that you can readily see how you are performing compared with your classmates, from the **leaderboard** panel. When submitting, use your student ID as the name for you.

Please also upload your final code to Canvas (Python notebook or R markdown file). Your code should contain some documentations, like explaining key components of your codes.

## Grading policy

If you can make a successful submission on Kaggle (regardless of the performance), you will automatically get 30% of the scores. The rest 70% will be calculated based on:

1. Code notebook (Python notebook or R markdown file) on GitHub (40%): with clear documentation of methods used and attempts you have tried to improve your prediction performances. The deadline is **March 24, Sunday, 23:59**.
2. The best performance you achieved (30%). Here you are competing with your classmates on best performances.