**INFO 659 SPRING 2023**

**Project Proposal: Predicting Data Science Job Salaries**

***Anh Nguyen, Amira Bendjama, Hong Doan***

## Business Question:

The goal of this project is to create a predictive model that estimates the salary range for data science jobs. The lack of transparency in salary often creates difficulties for job seekers such as recent graduates in data science who are trying to understand the job market and negotiate salaries. This model will provide them with a better understanding of the range of salaries they can expect and increase their earning potential. Overall, we expect this project to be a valuable contribution to the field of data science and job market transparency.

## Models to be used:

We will use three different models to predict data science job salaries: Linear Regression, Random Forest, and Decision Tree. Since we are not sure about the type of relationship between salaries and other features, our reasoning was: In case of a linear relationship, Linear regression might have better prediction, since it is a simple yet powerful model that can estimate the relationship between the dependent variable (salary) and one or more independent variables (job title, location, experience, etc.). However, in case of a more complex relationship, Random Forest and Decision Tree models might perform better, since they are both tree-based models that can handle categorical variables and nonlinear relationships between features. Consequently, the contrast between these models will help us better understand our problem in a deeper way.

## Dataset:

The [dataset](https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries) used for this project is available on Kaggle and contains information about data science jobs, such as job title, company location, salary, experience level, and more. The dataset has over 600 rows and 12 columns, providing enough data for modeling and analysis. The dataset is aggregated from the website ‘ai-jobs.net.’