**Data Mining Project Proposal** (MScA 31008, Professor U. Pamuksuz)

**Suggested dataset:** The Office of Foreign Labor Certification (OFLC) generates program data that is useful information about the immigration programs including the H1-B visa. The disclosure data updated annually is available at (<https://www.foreignlaborcert.doleta.gov/performancedata.cfm>)

The actual dataset on kaggle: **H-1B Visa Petitions 2011-2016**

**Dataset overview**: The dataset contains the H-1B petition data from 2011 to 2016. The H1-B visa is a category of employment-based, non-immigrant visas for temporary foreign workers in the United States. In this dataset, the data is more related with the employment and case details instead of the information of H1-B candidates. The columns include case status, employer name, worksite coordinates, job title, prevailing wage, occupation code and year filed. The details are presented below:

* CASE\_STATUS: Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.
* EMPLOYER\_NAME: Name of employer submitting labor condition application.
* SOC\_NAME: Occupational name associated with the SOC\_CODE. SOC\_CODE is the occupational code associated with the job being requested for temporary labor conditions, as classified by the Standard Occupational Classification (SOC) System.
* JOB\_TITLE: Title of the job
* FULL\_TIME\_POSITION: Y = Full Time Position; N = Part Time Position
* PREVAILING\_WAGE: Prevailing Wage for the job being requested for temporary labor condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer’s minimum requirements for the position.
* YEAR: Year in which the H-1B visa petition was filed
* WORKSITE: City and State information of the foreign worker’s intended area of employment
* lon: longitude of the Worksite
* lat: latitude of the Worksite

**Dataset structure:** 1 csv files, 3 million records, 11 columns, some missing data

**Business problem:** Utilize the dataset to review the general trends of international recruiting and categorize the job titles into different domains to compare hiring interests of different industries.

**Examples of business questions:**

* Is the number of petitions with the Data Science job title increasing over time?
* Which part of the US has the most Software Engineer jobs?
* Which industry has the most number of Data Scientist positions?
* Which Employers submit the most number of H-1B visa applications?
* What are the most common Job Titles applied for by the high applicant employers?
* What occupations are more likely to qualify for an H-1 B visa?

Key points:

* And Feature Mining
* feature transformation
* Don’t run the code, understand and explain what you used to class
* This is a supervised learning model, we need to come up with the outcomes for the h1b status
* But also use unsupervised learning methods such as t-SNE, KDE, feature mining etc.
* geospatial implementation:
  + <https://towardsdatascience.com/visualizing-geospatial-data-in-python-e070374fe621>
  + Which regions have predictable / unpredictable accuracy

Timeline : (3 weeks)

* clean the data -- 1 week
  + feature mining
  + decide how many columns should input in the model
* apply unsupervised learning -- 1 week
  + t-SNE
  + KDE
* apply supervised learning -- 2 week
  + logistic regression
  + random forest
  + naive bayes
  + decision trees
  + bagging boosting
  + gradient boosting model
  + stochastic gradient decient model
  + perceptron classifier
  + xgboost (failed)
* geospatial + presentation -- 3 week

Slides:

* Executive summary:
  + business problems, our goal
  + Explain data points + variables
* EDA & clean process & analysis:
  + cleaning and processing
  + visualizations
  + geospatial
* Modeling
  + Model results: try 8 models, important features
  + split and balance data(SMOTE and why )
  + used different version of cleaned data:
  + evaluate and choose the best performing one
  + Apply our best model on holdout dataset
  + why we choose certain model, why we use certain evaluation metrics
* Finding and strategy & future blueprint: results of our analysis:
* Recommendation engine