# Messy Data and Machine Learning

# Final Project Proposal

Ria Pinjani

Christina Ho

Zixuan Zhou

## Proposal Summary

## In this project, we will use a number of different supervised algorithms to precisely predict individuals’ income using data collected from the 1994 U.S. Census.

## We will then choose the best candidate algorithm from preliminary results and further optimize this algorithm to best model the data. Our goal with this implementation is to build a model that accurately predicts whether an individual makes more than $50,000.

## This sort of task can arise in a non-profit setting, where organizations survive on donations. Understanding an individual’s income can help a non-profit better understand how large of a donation to request, or whether or not they should reach out to begin with. As from our previous research we have found out that the individuals who are most likely to donate money to a charity are the ones that make more than $50,000.

## Data Description

The dataset for this project originates from the UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Census+Income>

The data we investigate here consists of small changes to the original dataset, such as removing the 'fnlwgt' feature and records with missing or ill-formatted entries.

The modified census dataset consists of approximately 32,000 data points, with each datapoint having 13 features.

**Features**

* age: Age
* workclass: Working Class (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
* education\_level: Level of Education (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool)
* education-num: Number of educational years completed
* marital-status: Marital status (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)
* occupation: Work Occupation (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)
* relationship: Relationship Status (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
* race: Race (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
* sex: Sex (Female, Male)
* capital-gain: Monetary Capital Gains
* capital-loss: Monetary Capital Losses
* hours-per-week: Average Hours Per Week Worked
* native-country: Native Country (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands)

**Target Variable**

* income: Income Class (<=50K, >50K)