# 3110 Final Project

Nick Carney, Max Ivry

# Problem Statement

The primary goal of this project is to demonstrate the susceptibility of Los Angeles crime Data to linking attacks, matching the names and information of criminals to their crime using the auxiliary dataset of the Condemned Inmate List of the California Department of Corrections and Rehabilitation. Additionally, a differentially private re-analysis of the dataset using the Laplace Distribution on several features shows a potential need for more deidentification measures. It is important to note that, while some re-identification was achieved, all information used in the project is publicly available, and this project only serves as an educational practice with no mal intent.

# Implementation

We scraped the California Condemned Inmate List’s website using beautifulsoup to gather information about condemned inmates in the state of California, such as their name, age, date of offense, and date sentenced… We compared this dataset to a Kaggle dataset consisting of over 2 million entries of Los Angeles crime data, with features consisting of offense data and others but not name, and attempted to find links between the two datasets.  We did this between the two original datasets, then applied some data privacy principles to add differential privacy. We first tried to generalize the date values, and then tried adding noise to the date values to compare their results.

# Results - No DP

We were able to obtain 9 exact-matches and 10 half-matches between the two datasets.  There were 131 potential matches to be found (had committed a crime between 2010 and 2019) and we were able to have a 50% guess about what crime they had committed  which is just under 15%. We are skeptical of the validity of these  results because many of the matched crimes that are supposed to linked with condemned felons are petty charges or misdemeanors.

# Results - DP

When adding generalization to the dates data, we were not able to find any exact or half-matches because the data was too general to have utility to an adversary. We generalized the day to the nearest 10th day, and the year to the nearest decade. With this generalization, we were able to prevent adversaries from having a good chance to guess the crime and associated assailant.

We also added laplace noise to the day and year values which caused us to find very many matches (exact+half matches of ~⅓ of the data) we think due to some luck. These matches are for the most part different from the results with no DP, allowing us to conclude that this does not allow an adversary to confidently guess inmates with their crimes accurately.

# References

Condemned Inmate List: https://www.cdcr.ca.gov/capital-punishment/condemned-inmate-list-secure-request/

Los Angeles Crime Data 2010- 2020: https://www.kaggle.com/datasets/sumaiaparveenshupti/los-angeles-crime-data-20102020/data.

Near, Joseph P., and Chiké Abuah. Programming Differential Privacy. vol. 1, 2021. programming-dp.com.