**Capstone-3 project ideas**

**Project 1: Employing ML algorithms to better understand and analyze future trend of violations (Time series analysis)**

Background:

Crime data can be useful to determine the efficacy of crime prevention steps and the safety of cities and societies. However, it is a difficult task to predict the crime accurately because the number of crimes is increasing day by day. The objective of this study is to apply time series to predict the crime rate to facilitate practical crime prevention solutions. Machine learning can play an important role to better understand and analyze the future trend of violations. Different time-series forecasting models will be used to predict the crime. These forecasting models are trained to predict future violent crimes. The FBI crime data is fascinating and one of the most interesting data sets for analyzing time series data to chart changes in crime rates at the national level over a [20-year period](https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/tables/table-1).

Datasets:

1. Crime in the United States by volume and rate per 100,000 Inhabitants, 1997–2016 <https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-1>
2. Estimated number of arrests in United States, 2016

<https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-18>

1. How many crimes came to the attention of law enforcement in my city in 2016?

Offenses Known to Law Enforcement by State by City, 2016 <https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-6/table-6.xls/view>

**2) Project 2:**

Employing machine learning algorithm to group or cluster soybean genotypes based on genetic similarity (Genetic diversity Analysis; is an example of Unsupervised Learning Clustering)

Soybean Germplasm includes 18,480 domesticated and 1168 wild soybean accessions introduced from 84 countries or developed in the United States, consists of wild, landrace and North American cultivars. This collection was genotyped with the SoySNP50K Bead Chip consists of >50K single-nucleotide polymorphisms (SNPs).

**Data source:**

1. The SoySNP50K iSelect BeadChip has been used to genotype the USDA soybean germplasm (Song et al., 2015). Complete data set for 20,087 G. max and G. soja accessions genotyped with 42,509 SNPs is available for Wm82.a2 in vcf format <https://soybase.org/snps/soysnp50k_wm82.a2_41317.vcf.gz>.
2. Minor Allele Frequencies (MAF) of SoySNP50K SNPs. Contains 40,841 of the 47,337 SNPs (86%) had minor allele frequencies >10% among the landraces, elite cultivars and the wild soybean accessions <https://soybase.org/snps/snp50k_maf.txt>
3. Table S1, from Song et al., 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes| Genomes| Genetics 50(10):1999-2006.)

Table S1. Description of 19,648 Glycine max and G. soja accessions genotyped with the SoySNP50K BeadChip. Data include the PI (Plant Introduction) number, Species (G. max vs. G. soja), Country of origin, Cultivar name if applicable, Maturity Group, 99.9% similar to another accession (Y/N), Landrace or North American Elite cultivars used for analysis and seed weight of accessions used in the association analysis of seed weight.

[Soy\_USDA\_Lines\_descritpion.xlsx](file:///C:\Users\gellima\Desktop\Springboard_DScience\CAPSTONE_3\'Soy_USDA_Lines_descritpion.xlsx')