Capstone 1 Project Ideas

# Agricultural Estimates

This is a Kaggle dataset for Argentina agricultural production. The set is 133k samples with 4 characteristics (Province, Department, Crop and Year) and 4 facts (Hectares sown, Hectares harvested, Tons produced, Harvest performance). This set is production yield by year starting all the way from the year 1969 to 2017, which a good amount historical information. The type of crop range is limited to grains, fruits and some not directly consumable crops like cotton.

Due to the many technological advance that has happened in the past century the crop yield across developed and developing countries has increased in recent years. But there are still some external factors that affect production, some of the biggest ones are weather, availability of labor, economy and events like draught/floods, plant diseases and pest infection. So, I am looking to answer some of the basic questions, given this data (assuming all external factors constant),

* Does the crop yield in anyway related on the area to be manages?
* Is the yield rate any different by crop or region?
* Finally, predicting the yield for next year and decade.

And if the data is enriched (if the data is available),

* Has the change in weather affected the yield?
* Does the availability of labor change the production?
* Considering any of these factors does the prediction change?

Dataset: <https://www.kaggle.com/pablolebed/agricultural-estimates-arg>

# Image Classification

Another dataset from Kaggle. This is a simple set of over 16k images of cars with class of cars. The class of car is based on make, model and year. The aim is simple, to identify a car from given image and if possible, the color of the car as well.

Dataset: <https://www.kaggle.com/jutrera/stanford-car-dataset-by-classes-folder/home>

# Movie Recommendation

This is dataset is from grouplens based on the movielens rating data sets. The current data set is of around 20 million ratings and 465,000 tag applications applied to 27,000 movies by 138,000 users. Hence there is info of movies that are being rated by user along with genres and tags on the movies.

The goal here is do a movie recommendation system that recommends a movie based on user history (previous likes and liked by similar people) and /or movie attributes (genre, actors etc.). This may sound simple enough but will take lot of iteration to reach a decent accuracy level since user behavior or choices may change over time or one odd movie can throw off the model.

Dataset: <https://grouplens.org/datasets/movielens/>