Milestone II Project Proposal

# A Thirsty Valley

**Aim**: Create measurable objective of sustainability

California's [Sustainable Groundwater Management Act](https://www.waterboards.ca.gov/water_issues/programs/gmp/) (SGMA) was passed in 2014 with the intention to address over pumping, halt chronic water-level declines and bring long-depleted aquifers into balance. Despite SGMA, a frenzy of well drilling has continued on large farms across the **San Joaquin Valley, the state's largest and most lucrative agricultural zone.** As a result, shallower wells supplying nearly a thousand family homes have gone dry in recent years.

Frequently, perniciously drought-inflicted California, depends on groundwater for a major portion of its annual water supply, particularly for agricultural and domestic usage. This project seeks to aid policy makers preemptively identify areas prone to overdraft and bring groundwater basins into **balanced levels of pumping and recharge.**

Focused on the San Joaquin Valley, the objectives are:

* **Supervised Learning:** Predict the depth to groundwater elevation in feet below ground surface (GSE\_GWE). This value portends shortage in a TownshipRange. Increase or decrease in GSE\_GWE will then indicate if there will be more requests for well construction. This in turn will provide a quantitative metric for whether SGMA is functioning and areas to focus on for recharge.
* **Unsupervised Learning:** cluster areas into sustainable and unsustainable areas, anomaly detection

The geographical unit of analysis chosen for this project is the Township-Range level of the [Public Land Survey System](https://en.wikipedia.org/wiki/Public_Land_Survey_System).

# The Datasets

We have collected 10 geospatial datasets from different government agencies for the 2014-2021 period, on the topics impacting water consumption:

Geo-spatial data (San Joaquin Valley PLSS), current groundwater levels and consumption through well completion reports, recharge factors and consumption factors through precipitation, water reservoir capacity, agricultural crops, population density, regional vegetation, soils survey and water shortages reports.

Basic GeoPandas ETL operations to aggregate data at TownshipRange level were followed up with pipelines, leveraged to impute missing values, normalize and scale data appropriately.

* Use Voronoi diagrams to estimate area value from point measurement (e.g precipitation reading from measurement station is used to infer TownshipRange precipitation)
* Aggregate counts or measures at the Township-Range level (e.g. the count of well shortage reports)
* Overlay Township-Range boundaries on geospatial data to estimate Township-Range values (e.g. estimate the percentage of land use for different type of agriculture crops)
* Data imputation, feature scaling and feature dimensionality reduction are performed on the aggregated data (where applicable) before doing supervised and unsupervised learning.

The post ETL dataset used for supervised and unsupervised learning is a multi-time-series (one 2014 to 2021 time series per Township-Range) multivariate (each Township-Range time series has many variables) dataset.

# Supervised Machine Learning

* Set baseline with Dummy Regressor
* Regression : Evaluation using RMSE, R2, MAE
* GridSearchCV/RandomSearchCV to tune parameters
* Check for over-fitting and regularization
* Will use PCA (unsupervised technique) to reduce features to aid the regression task, biplot visualization to enhance understanding of latent dimensions, scree plot for selection.
* Will use K Means clustering (unsupervised technique) for feature engineering.
* Will compare tree and forest regressors with linear regression algorithms. The former is not sensitive to unnormalized and unscaled data
* Will apply Lasso regression (also aids in feature selection by setting low coefficients) and Ridge regression
* Identify important features using Shapely

# Unsupervised Machine Learning

* Top-down, expectation maximization technique K-Means clustering (pre-defined- number of clusters),
* Hierarchical, Bottom-up technique : Agglomerative clustering (dendogram visualization)
* Outlier detection friendly clustering (DBSCAN) to mediate an emergency situation.
* Evaluation of inter and intra cluster scores using cost functions.

# The Team and Contribution

Both Simi Talkar and Matthieu Lienart have participated to all aspects of the project.

# The Timeline:

1. January to April 2022: Datasets collection and ETL
2. May to August 2022: Data imputation, feature dimensionality reduction, supervised and unsupervised learning
3. September, October 2022: Write report

**Appendix:**

1. Sustainable groundwater management is defined as managing water supplies in a way that can be maintained without “causing undesirable results,” such as **chronic declines in groundwater levels** or “significant and unreasonable” depletion, adverse effects on surface water, **degraded water quality** or land subsidence.

2. Drought years in California

Chart

Description automatically generated

4.Datasets

