Isaac Kramer, Noam Gottlieb, and Karin Osadon
67978: A Needle in a Data Haystack
Introduction to Data Science
Semester Project Proposal

Proposal 1: The Kendall Irrigation Project

In recent years, many researchers have sought to apply the tools of data science to improve resource efficiency and crop yields in agriculture. To this end, a team of researchers at the Volcani Institute (the research branch of Israel's Ministry of Agriculture) is engaged in a multiyear project aimed at increasing our understanding of how spacial heterogeneity within an agriculture field affects irrigation distribution. Some details regarding the project can be found here: `https://www.researchgate.net/project/Precision-Drip-Irrigation-in-Orchards`.

We have been in communication with the researchers on this project and they have proposed an idea related to looking for patterns in the distribution of water content within the field. Specifically, they are willing to share with us three months of sensor data from a vineyard in Israel. The sensor data includes variables such as soil temperature and volumetric water content (the water content of the soil normalized with respect to the soil's total pore volume). They would also provide us with data related to the irrigation schedule and "stem water potential," a measure of the amount of water in the plant leaf and an indicator of plant health. Our goal would be to search the data for connections between the variables and time. To look for patterns in how soil water content responds to irrigation inputs and how this may affect plant health. It may also be possible to try and connect these observations to ambient meteorological data and/or micrometeorological measurements. We believe this would be not only interesting and challenging from a data science perspective, but could also help a worthwhile project.

One of the members of our project team (Isaac) is a PhD student at the Faculty of Agriculture and has the necessary background to help the group understand the project variables. The researchers at the Volcani Center would also be happy to provide ongoing support.

Proposal 2: Using Data Mining to Identify Saline Soils

Soil salinity is a major problem affecting agricultural production worldwide, especially in arid and semi-arid regions. Measuring soil salinity, however, is costly and inefficient. To measure soil salinity, one must physically visit a given field to collect samples. These samples must then be analyzed using lab equipment. These constraints obviously pose limits on the frequency with which soil salinity can be measured in a field, as well as the extent (area) of the field that can be measured.

In recent years, remote sensing has offered an alternative means of data collection for agricultural systems. Publicly available data from satellites and private data obtained from drones offer exciting alternatives. In comparison to physical soil samples, this data is cheap to obtain, can cover much greater spacial extents, and can be collected with greater time frequency. Unfortunately, however, there is no way to directly measure soil salinity using remote sensing.

We propose a data mining project that would involve trying to discover associations between variables that can be measured using remote sensing and existing soil salinity data. This is not a novel research idea. The paper *Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates* (Vermeulen, D., and Van Niekerk, A. (2017). Geoderma, 299, 1–12) established this technique. In this project, we would attempt to imitate their research approach using different data sets.

To complete this project, we would need to find data sets related to soil salinity in a set of agricultural fields. We would also need remotely sensed data associated with these fields collected at the same time as the salinity data.

Proposal 3: The Quick Draw Dataset Quick Draw is a Google-associated A.I. project. Users of the website `https://quickdraw.withgoogle.com/` make drawings and a machine learning algorithm attempts to guess what the drawing is. Each new drawing then helps the algorithm grow in its accuracy. To date, the "game" has generated 50 million drawings, distributed over several hundred categories. The recognized drawings and associated metadata have been shared on GitHub for researchers and developers to experiment with.

We propose a project focused on using this dataset. An existing project analyzed how users draw circles. Observations included differences in starting points and path direction (counterclockwise vs. clockwise). The researchers were able to tie some of these differences to nationality, which they argue is a proxy for differences in language/writing direction and character types. We could try and continue this idea with another shape included in the Quick Draw set. For example, analyzing the data associated with drawings of squares.