# MEMO

**To:** Geeks Without Bounds Manager

**From:** Employees (Team Snow: Alexis Rouge Carrassat, Fan Liu, Anthony Niznik)

**Date:** Sept 22 2016

**Subject:** Pump it Up – Data Mining the Water Table (Hosted by DrivenData)

---

**Background**

Access to a clean and reliable supply of water is a human right for all countries. Tanzania, like many poor countries, suffers from serious issues of water quality. Tanzania uses ground water for their water supply which often times contains pollutants, such as human waste and water-borne illnesses such as malaria. Geeks Without Bounds, being a nonprofit humanitarian organization, is compelled to address Tanzania's water dilemma. Since we conduct hackathons for humanitarian projects, this provides a great opportunity to help a whole country.

There is a huge need to determine which water wells are functioning and which wells need repair. It is imperative that we help Tanzanian's Ministry of Water and Irrigation with this current water issue since it will give our nonprofit global exposure and opportunities for more volunteers to give their services to our organization.

In the dataset that we received from Taarifa, a group of technologists who are trying to solve water delivery problems, we have 39 features including region code, population, water quality, ground quality and etc., for around 60,000 different wells in Tanzania.

**Objectives**

- Understand the features that determine whether a well will be functional or not. We will see which features directly influence the quality of a well
- Predict, given a set of data and features, if a well is functional
- Find patterns in the features that make a well functional in order to give advice for future wells' constructions.

**Proposed Methods**

We are mainly dealing with a classification problem. Given the amount of features and data we that have, we believe there are plenty of ways to approach this problem.

For example, we can apply models such as random forest, logistic regression, k-means clustering or neural network to solve this problem. One challenge we may face could be how to perform efficient feature engineering to improve the performance of our model. Another algorithm we want to work with is the Patient Rule Induction Method and it will be our starting point to deal with the data. Indeed, this method gives rules that the features should follow in order to optimize the percentage of functional wells.