THEORY-GUIDED DATA SCIENCE: COMBINING MACHINE LEARNING WITH DOMAIN EXPERTISE TO PREDICT SPRINGFLOW

Emily Camille Pease

ABSTRACT

Traditionally, science follows a theory-based approach through which physical equations are used to model natural phenomena. In this recent era of artificial intelligence and "big data", there is a shift into a new paradigm of scientific discovery. The paradigm of theory-guided data science (TGDS) enables scientists to perform data science modeling while retaining their domain expertise to produce informed results consistent with the physical system. Predicting springflow discharge from Comal Springs using machine learning was determined to be an appropriate case study. The Edwards Aquifer in central Texas serves as the primary water supply for over 1.5 million Texans, providing water for recreational activities, businesses, and down-stream users. Additionally, these waters serve as a home to many aquatic species, eight of which are endangered or threatened. Quantifying springflow is essential in regulating groundwater resources in the Edwards Aquifer, especially during drought conditions. Here, a theory-guided predictive machine learning model for springflow estimation at Comal Springs is developed. First, feature engineering is performed to discover relations between data available in the Edwards Aquifer region, selected through theory-guided parameter initialization. Next, multiple machine learning models were explored and tested in their ability to model a complex springs system. Finally, theory-guided refinement of data science outputs was performed to make the model results consistent with what is possible in nature.

Suzanne Pierce

Dr. Suzanne A. Pierce