# Proposal: Predicting GitHub Repository Popularity

**Grace Austen**
Clemson University: School of Computing
gausten@clemson.edu

**Jacob Schechter**
Clemson University: School of Computing
jcschec@clemson.edu

## 1 Introduction

Our project will be centrally focused on GitHub, the widely used platform that supports software development and version control using Git. In addition to software development and version control, users can search for projects and interact by starring, forking, watching, raising issues, starting discussions, and more. Developers can also create GitHub Pages, websites that contain additional helpful information for the associated repository. It is the aim of our project to determine if the attributes of GitHub repositories are influential to their popularity and whether or not certain features can be used to predict a repository's popularity.

## 2 Problem Statement

The purpose of this project is to determine whether or not the features tracked within GitHub repositories can be used to predict a repository's popularity. It is our aim to understand which features possess the greatest influence in repository's associated popularity.

## 3 Dataset

We will use the dataset, "Most Popular GitHub Repositories (Projects)" provided by Canard on Kaggle. The dataset consists of over 215,000 GitHub repositories with their associated features. The tracked repository features include: name, description, URL, date created, date last updated, URL for the GitHub homepage (if one exists), size of repository in bytes, number of stars, number of forks, number of open issues, number of GitHub watchers, primary language, license used, list of topics or tags, whether or not issue tracker is enabled, if the repository uses GitHub projects, whether or not the repository has downloadable files, if the repository has an associated wiki, if it has GitHub pages enabled, discussions enabled, is a fork, is archived, is a template, and the name of the default branch.

## 4 Methodology

We will utilize MATLAB for the purpose of our data analysis and modeling. The machine learning techniques we will employ are PCA (Principle Component Analysis), LASSO Regression, and raw features for the purpose of performance comparison. PCA will be applied for dimensionality reduction by identifying key components while minimizing information loss. LASSO regression will be used to simplify the model and aid in feature selection to retain only the most important features. Raw features will allow for feature impact on popularity prediction to be directly interpreted.
Our process first involves data preprocessing: the handling of outliers, normalization of features, special encoding for names, and PCA for dimensionality reduction. We then focus on feature selection, determining column inclusion, exploring tag encoding options, and decide on the inclusion of descriptions in our analysis. Lastly, our regression techniques include: principal component regression using PCA for feature reduction, Lasso regression for effective feature selection to prevent possible overfitting, and raw features for direct interpretability and insight into popularity prediction.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).