Currently, much of the internet includes anonymous postings, information that can be used to perform many types of analysis on the online behavior of users of social media. Few publicly available tools are able to summarize and comprehend this data. Our application aims to utilize this information and display just how much can be learned from an individual’s or community’s social media history.

Our website will provide insight into a Reddit user's activity. The main focus of our analysis is on the content of a user’s submissions. Using natural language processing (NLP) techniques, we aim to find, aggregate, and display relevant information about a user, including personally identifiable data. We chose to use Reddit as it is a medium that promotes discussion on an extremely extensive range of topics. As such, the demographic of its users is extremely diverse and there is no one “type” of user, as is the case with many other topic-specific forums. The fact that Reddit’s subreddit pool is so large, and most users posts across more than one subreddit, there are many interesting data points that can be drawn by looking at contributions across different topics. The data we provide has several uses. Most commonly, it would be a way for a user to monitor how much information they are exposing and how they are portraying themselves online, or to simply satisfy curiosity about another user’s activity.

Our application is built on the MEAN (MongoDB, Express.js, AngularJS, Node.js) stack. We are working with MVC (model-view-controller) architecture, so in our back-end, we will create a User schema with attributes corresponding to what we are displaying in the view. User data will be stored by consuming our own API which make calls to Reddit's API per request, e.g. comments, metadata. While we take some of this data raw, the majority is analyzed with NLP libraries in order to produce meaningful statistics. In the front-end, Angular's controllers handle the responses from the API endpoints and serve them to the view for end users. The information that is displayed includes metadata and NLP metrics and analysis, as well as graphs. The majority of our NLP processing will utilize Stanford CoreNLP and our data graphs will be shown using the D3.js library. Though our application uses Reddit specifically, the NLP analysis and much of the metadata analysis that we perform is generalizable to users’ data sets on any text-based social media site.