For my final capstone, I'd like to know which features make a song popular: specifically, can we accurately predict whether or not a song will break into the United States Billboard Top 100 for a given timeframe? For this project I will use two separate data sources: the Billboard Top 100 for the past 12 years, as well as Spotify streaming data starting in 2008. We will need to use a web scraper for the first data collection, and can simply use the Spotify API (developer.spotify.com) for the other data set. For the sake of simplicity, we will focus exclusively on whether a song made it into the top 10 of the top 100.

The questions of particular relevance to both music labels and distribution platforms; in particular, if the top 3 labels in the world (which control ~70% of the world's music supply) are able to more accurately predict which songs are going to be hits, they will know where to invest marketing and promotion dollars. Similarly, platforms such as Spotify and Pandora will be more inclined to surface these songs on their Discovery/recommendation lists, thereby increasing user satisfaction relative to competitive alternatives such as Apple or Amazon.

The idea will be to focus in which features are most correlated with breaking into the top 100 -- again, these are separate data sets, so we will need to know to what extent features like length of song, tempo, volume, "energy," more correlate to a song's ability to break into the Billboard top 100. Why focus on streaming data?

For years, the music industry was in decline -- the prevalence of online pirating and companies such as NASPERS made it so that the music industry had gone ex growth since 1999, with the industry bottoming in 2014 at $6.7Bn domestically. Today, the music industry is back to double digit growth, with recorded music growing 13% to $11.1Bn in 2019 thanks in large part to the advent of streaming, as the likes of Spotify, Pandora, Apple, Amazon, and others have brought in hundreds of millions of users to the on-demand and ad supported ecosystem. These streaming services are offsetting declining revenues from physical distribution (discs) as well as radio -- as such, while physical sales and radio share time are also used in determining the Billboard Top 100, they are becoming increasingly less relevant as time goes on.

The goal of the project will be to see if we can predict the Top 10 of the top 100 for the last 3 years, using the last 12 years of data. If successful, the engine can be used to power both recommendation engines from a distributor's standpoint, and it can inform investment decisions at the label level. It could potentially fuel creativity at the artist level, but we'd be hard pressed to find any artists who take a pure data science approach to their expression.