WSABI Summer Research Lab **Summer 2025**

Lab 19: Clustering

Instructor: Jonathan Pipping *Authors: JP, VL, CB*

19.1 Spotify Collaborative Playlists

19.1.1 Overview

In 2008, Spotify launched their *Collaborative Playlists* feature, which allows users to create and contribute to shared playlists with their friends. Since then, over a billion collaborative playlists have been created, which Spotify uses to boost engagement and recommend new music to users. This summer, our Lab created a collaborative playlist of approximately 300 songs (as of June 27th), which we will be exploring and using to make predictions on unseen data.

19.1.2 The Data

We have a dataset of approximately 300 songs, each of which includes the following features:

- **Song Name:** The name of the song
- **Album:** The album the song is featured on
- **Year:** The year the song was released
- **Artist:** The artist(s) performing the song
- **Genre:** The genre of the artist(s)
- **Contributor:** The user who added the song to the playlist
- **Song Details:** A number of features describing the song (e.g. danceability, tempo, energy, etc.)

You will begin by exploring the data to better understand the songs that make up our playlist.

19.1.3 Clustering Playlist Songs

1. Make at least 2 visualizations which help you understand the playlist's songs and features. Consider variables like genre, contributor, and other song details.
2. Filter your dataset to only include numerical features, then standardize them by subtracting their mean and dividing by their standard deviation.
3. Cluster the songs using one of the clustering methods we discussed in Lecture 19. If you select a method which requires specifying the number of clusters, do this using one of the methods we discussed in Lecture 19. If you need more help doing this, refer to the slides in `2025/supplementary`.

4. Interpret the clusters through visualizations and/or examining the songs in each cluster. Describe the types of songs in each cluster.

Next, you will fit predictive models to the data to predict who contributed a song to the playlist.

19.1.4 Prediction Competition

Using **only** the methods you've learned about in Lectures and Labs 1-19, fit a predictive model to the Spotify data to predict who added a particular song to the playlist. You are allowed to train as many models as you want, but you must submit **one** final model for the official competition. On July 30th, your final model will be tested on the songs added over the remainder of the summer, and the Lab member with the lowest prediction error (via Cross-Entropy Loss) will win a cash prize. Instructions for training and submitting your model are included below.

1. Create a new file called `yourfirstname_competition.R` in the `2025/labs/submissions` directory.
2. Set a unique 5-digit seed for your code in the `SETUP` section.
3. Filter the data to include only Lab members who contributed at least 10 songs to the playlist.
4. Fit a predictive classification model to the data, using any of the features you want. It is highly recommended that you fit multiple models and compare their performance.
5. Once you are satisfied with the performance of your model, save it to a variable in your file called `final_model`.
6. Push your code to GitHub by 11:59pm on June 29th. Any modifications after this time will not be considered.