Spotify Galaxy: Visualizing High-Dimensional Music Data

Introduction

The primary analysis question driving this visualization is: *How have the musical styles and audio features evolved over time across distinct genre clusters?* To answer this, "Spotify Galaxy" presents a multi-view interface which comprises of an immersive **3D Star Field** which is representing the feature space, coordinated with **Cluster Distribution**, **Audio Radar**, and **Temporal Evolution** charts. The core view can be seen as an interactive 3D star field, songs positioned according to PCA-reduced audio features, where the color encodes the cluster membership. This is interactively linked to auxiliary analytical views including cluster distribution, a temporal duration trend line showing average duration of songs throughout the timeline, and an audio-feature radar chart.

1 Tools & Preprocessing

1.1 Tools & Technologies

I built this visualization using "ReactJs" and "Three.js" (via *@react-three/fiber*). WebGL was used for high-performance rendering of 20,000+ interactive datapoints. "Python" and it's libraries (pandas, scikit-learn) were used for data preprocessing purposes. The "D3.js" concepts and the "Recharts" were employed for coordinated 2D detail views.
The interface is a combination of Three.js WebGL scene for the galaxy view. There are standard React UI panels for legends, encodings, instructions toggles, and linked charts. "Recharts" is used to render the bar chart, duration line chart, and the radar plot. All the charts are synchronized with the React state.
This overall application was pushed on github and published publicly on the following link:
`https://ghoshnet.github.io/spotify-galaxy/`

1.2 Preprocessing & Methodology

The raw dataset consists of 169,910 songs. The following preprocessing and data analytics steps were applied:

- **1. Cleaning**: Filtered for years 1921-2020 and removed duplicates. This resulted in a dataset of approximately 19,700 songs.
- **2. Normalization**: Numerical features (like tempo, energy, etc.) were scaled using `StandardScaler` to ensure equal weight during analysis.
- **3. Dimensionality Reduction (PCA)**: I applied the Principal Component Analysis to 12 audio features and reduced them to 3 dimensions (X, Y, Z coordinates), preserving 65% of variance. Now because of this, approximately 35% of feature variability is not represented in the 3D scatter plot. Thus, instead of interpreting the distances and cluster boundaries as exact separations, they should be interpreted as approximations.
- **4. Clustering (K-Means):** Songs were grouped into 8 clusters based on their audio profiles to identify distinct genres ($k = 8$ determined via elbow method).

In order for all songs to share the same coordinate system across time, PCA was fitted once globally on the entire dataset, ensuring temporal filtering does not distort spatial layout. This allows users to observe a stable cluster structure across decades.

2 Dataset Description

The dataset consists of **19,769 items** spanning in the time range **1921–2020**.
Dataset Link is as follows:

https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-1921-2020-160k-tracks

- **Data Types**: The data includes **Ordinal** attributes (Year), **Quantitative** audio features like: Energy, Valence, Danceability etc., and **Categorical** derived attributes like "Cluster ID".
- **High Dimensionality**: The 12+ quantitative audio attributes make standard 2D scatter plots ineffective due to overplotting and inability to capture the full "shape" of the music space.
- **Derived Attributes**: I derived categorical *Cluster Labels* and numerical *PCA Coordinates* to give a structure to the visualization.

There is however, a heavy imbalance across decades, with far fewer tracks prior to the 1950s. Thus, interpretation must be done cautiously as it impacts trends in early years.

3 Visual Design & Tasks

3.1 User Tasks

- **Sonic Exploration**: User can identify songs that sound similar based on spatial proximity rather than genre tags.
- **Trend Analysis**: Observe how musical styles have evolved over decades by using the time range slider and observing the graphical visualizations. e.g.: the rise of high-energy electronic music)
- **Cluster Characterization**: Understand what defines a specific genre group. Like for example, "Cluster 3 is high energy, high valence".

3.2 Encodings & Idioms

- **Position (3D Scatter)**: X, Y, Z map to PC1, PC2, PC3. The Proximate stars share similar audio fingerprints.
- **Color (Hue)**: Encodes the K-Means Cluster, allowing distinct genre groups (e.g., "Acoustic Ballads" vs "Energetic Pop") to be visible instantly.
- **Size**: Encodes *Energy* in the 3D view. Larger stars are more energetic, adding a layer of depth and visual hierarchy.
- **Bar Chart (Cluster Distribution)**: Uses **Length** to encode the count of songs per cluster, providing a summary of the current filtered selection. (top-right)
- **Radar Chart (Audio Features)**: Uses **Radial Distance** to compare the normalized averages of 5 audio features (Danceability, Energy, Valence, etc.) for the visible data. (bottom-right)
- **Line Chart (Temporal Evolution)**: Uses **Vertical Position** to encode average song duration and **Horizontal Position** for time (Year), revealing temporal trends.(bottom-left)
- **Gloss/Bloom**: A post-processing "glow" effect is used to emphasizes dense clusters and create the "galaxy" type of aesthetic.

3.3 Interaction Design

- **Navigation**: Orbit controls (Rotate/Zoom/Pan) resolve 3D occlusion issues.
- **Coordinated Filtering**: A timeline slider filters the galaxy by year, updating all linked charts synchronously.
- **Details-on-Demand**: Hovering over a star reveals metadata and highlights its position in the linked charts. And clicking on the star will stop the rotation for a few seconds or unless the user clicks anywhere else.
- **Toggle and Filter**: Users can toggle cluster on and off for the Radar chart using the cluster selection panel. The filtering via the time range slider updates all the linked charts in real-time.

4 Novelty & Complexity

Novelty: The "Spotify Galaxy" visualization departs from the traditional dashboards by using an immersive 3D star-field metaphor. Unlike standard dashboards, it is using a spatial metaphor, enhanced with bloom post-processing to emulate dense audio clusters. Music is not represented as a list, here, it is represented as a space to be traversed. It integrates multiple coordinated analytical views which includes cluster distribution, duration timeline and radar profiles. This visualization bridges the gap between "artistic" and "analytical" visualization.

Complexity:

- **Technical**: Rendering of 20k points at 60fps required `InstancedMesh` optimization in Three.js, rather than standard object creation.
- **Synchronization**: State management was complex to ensure the 3D scene, timeline filter, and 2D charts remained perfectly in sync without lag.

It must be also noted that the cluster labels shown in the interface are descriptive names assigned post-hoc based on the centroid audio features. Thus, it should not be interpreted as strict genre definitions.

5 Critical Analysis

Strengths: The visualization is highly engaging and effectively unfolds the complex high-dimensional data. The clustering is visibly effective. Distinct clouds of color prove that the PCA/K-Means approach successfully grouped similar songs. Like the yellow cluster can be seen forming a well-separated island in PCA space, which suggests a unique high-energy audio signature. The overall representation looks aesthetic with interactivity. By looking at the average duration curve after the mid-2000s, we can interpret a clear structural drop reflecting modern shifts towards the shorter and optimized tracks for streaming.

Weaknesses:

- **Interpretation**: The PCA axes are abstract mathematical constructs, making it difficult for users to intuitively understand what "X" or "Y" specifically represents. Additionally, the K-Means clusters, while mathematically distinct, do not always map cleanly to familiar genre names.
- **Occlusion**: Despite rotation, dense clusters in 3D can be hard to read. Like the turquoise and pink clusters occupy overlapping PCA regions at certain regions. This indicates two distinct audio-feature groupings sharing similar sonic properties but sometimes makes it hard for users to interpret from particular points or angles.
  The Glow and bloom effects ,while aesthetic, reduce the precision of interpreting points in the dense areas.
- **Future Work**: Future iterations could implement "axis explanation" tooltips that reveal which features dominate the current view.

References

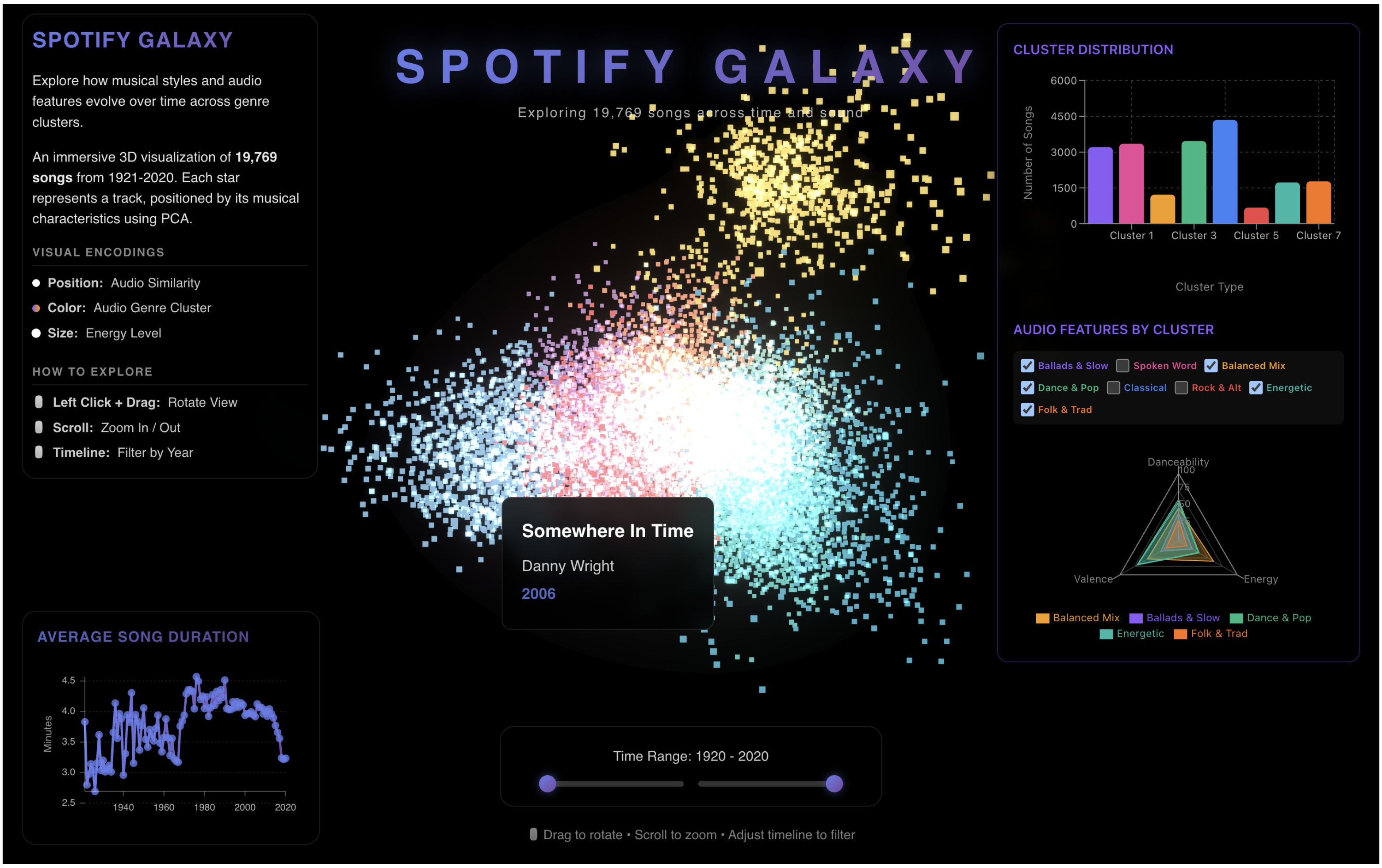
[1] Yamac Eren Ay. *Spotify Dataset 1921–2020, 160k+ Tracks*. Kaggle, April 2021. `https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-1921-2020-160k-tracks`

[2] Three.js. *Three.js Documentation*. `https://threejs.org/docs/`

[3] Tamara Munzner. *Visualization Analysis and Design*. A K Peters/CRC Press, 2014. `https://www.crcpress.com/9781466508910`

Annexure

I acknowledge the use of Co-pilot AI for summarising initial notes and proofreading the final draft.



SPOTIFY GALAXY

Explore how musical styles and audio features evolve over time across genre clusters.

An immersive 3D visualization of **19,769 songs** from 1921-2020. Each star represents a track, positioned by its musical characteristics using PCA.

VISUAL ENCODINGS

- **Position:** Audio Similarity
- **Color:** Audio Genre Cluster
- **Size:** Energy Level

HOW TO EXPLORE

- **Left Click + Drag:** Rotate View
- **Scroll:** Zoom In / Out
- **Timeline:** Filter by Year

SPOTIFY GALAXY

Exploring 19,769 songs across time and sound

Somewhere In Time

Danny Wright

2006

CLUSTER DISTRIBUTION

AUDIO FEATURES BY CLUSTER

☑ Ballads & Slow □ Spoken Word ☑ Balanced Mix
☑ Dance & Pop □ Classical □ Rock & Alt ☑ Energetic
☑ Folk & Trad

AVERAGE SONG DURATION

Time Range: 1920 - 2020

Drag to rotate • Scroll to zoom • Adjust timeline to filter