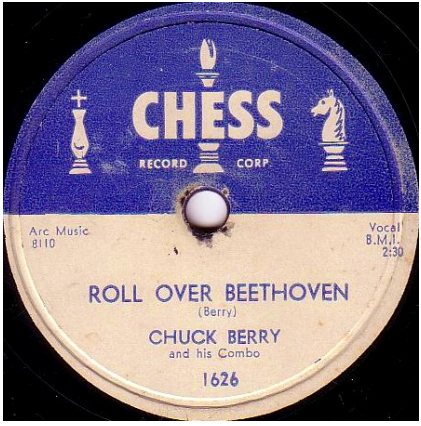
Music as Data

Tracing the Evolution of Style and Values with Modern Audio Models

UZH Economics

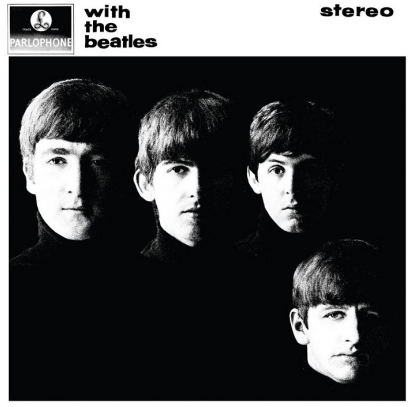
October 29, 2025

Chuck Berry (1956) vs. The Beatles (1963)



Chuck Berry (1956)
Roll Over Beethoven

▸ Play Audio

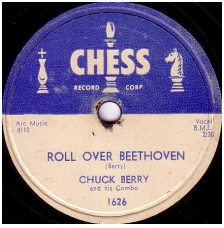


The Beatles (1963)
Roll Over Beethoven

▸ Play Audio

Introduction: The Beatles and the Transatlantic Circulation of Sound

- In the early 1960s, The Beatles emerged from Liverpool, absorbing American R&B, Chuck Berry's guitar phrasing, Little Richard's vocal timbre, and Motown's rhythm.
- They translated these Black American musical traditions into a new British pop sound that re-entered the U.S. charts during the "British Invasion"
- Can we measure this diffusion *sonically*: rhythm, tone, and production features, without relying on lyrics or cultural narratives?



Chuck Berry (1956)

Roll Over Beethoven

The Beatles (1963)

Roll Over Beethoven

Research Agenda

Guiding questions

1. How do rhythmic, harmonic, and timbral traits evolve over decades?
2. How do they diffuse across regions, scenes, and labels?
3. Which macro shocks (policy, tech, crises) shift sonic traits?
4. What do trait shifts suggest about values?

Design principles

- Use audio (not lyrics) to avoid text confounds.
- Prefer interpretable features + learned embeddings.
- Build a panel: track $\times$ year $\times$ place.

Core Tools + Methods

- **Feature extraction:** *Essentia*, *librosa*: compute interpretable descriptors such as tempo, key, spectral centroid, dynamic range, and rhythmic regularity. Useful for reproducing historical "engineered" traits and cross-validating learned embeddings.
- **Embeddings:** *MERT* or *CLAP/OpenL3* (audio encoders trained on large-scale text–audio corpora). Provide dense, high-level representations of timbre and production style, enabling similarity analysis, clustering, and cultural diffusion tracking.
- **Metadata joins:** *MusicBrainz*, *Discogs*: supply canonical identifiers for track, release, artist, label, and region. Enable linking to year, genre, and country attributes for econometric analysis.

Two Ways to Use Music as Data

Two complementary representations: features give interpretability, embeddings give depth.

1. Pre-computed / Engineered Features

- Examples: *tempo, loudness, key, energy, danceability*.
- Extracted via tools like *Essentia, librosa, Echo Nest*.
- Interpretable, standardized, reproducible.

2. Learned Embeddings

- Examples: *MERT, CLAP, OpenL3, Wav2Vec2*.
- Derived from deep audio encoders trained on text–audio pairs.
- Capture complex timbral, stylistic, and production features.
- Harder to interpret, but richer and higher-dimensional.

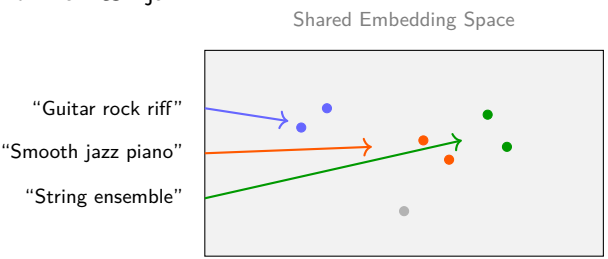
Engineered Features: The Million Song Dataset (MSD)

- Large-scale database of $\sim$1 million songs.
- Contains **pre-computed audio features** (not raw audio) generated via Echo Nest algorithms.
- Each row = one track, combining *metadata* and dozens of numerical descriptors.

Field	Example Value	Description
track_id	TRMMMYQ128F932D901	Unique ID for the track
artist_name	The Beatles	Artist metadata
title	A Hard Day's Night	Song title
year	1964	Release year
tempo	140.05	Global tempo (BPM)
key	9	MIDI key (0=C, 11=B)
mode	1	1 = major, 0 = minor
loudness	-5.42	Average loudness (dBFS)
energy	0.812	Normalized energy (0–1)
danceability	0.645	From Echo Nest model

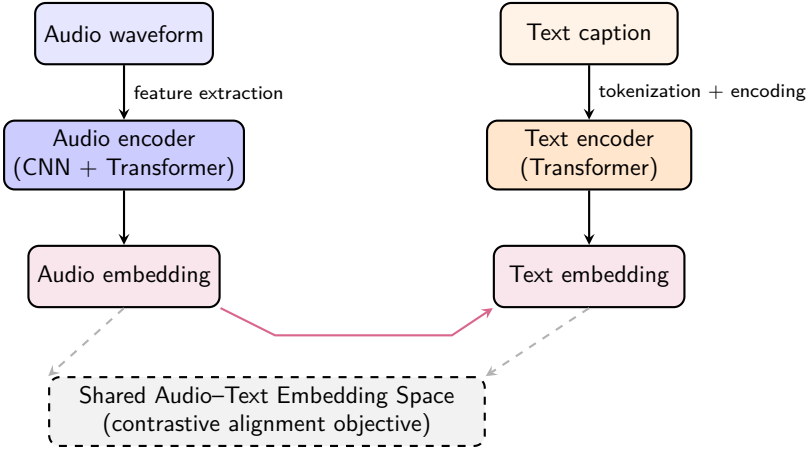
Learned Embeddings (CLAP / MERT)

- **Goal:** Represent each audio clip as a dense vector capturing musical style.
- **Method:** Deep encoders (e.g., *CLAP*, *MERT*) learn from millions of audio examples, some with text captions, others self-supervised.
- **Result:** Songs that sound alike or share stylistic traits end up close together in embedding space.
- **Applications:**
  - Clustering songs by sound similarity
  - Tracing stylistic drift over time
  - Zero-shot search: "find tracks that sound like R&B jazz"



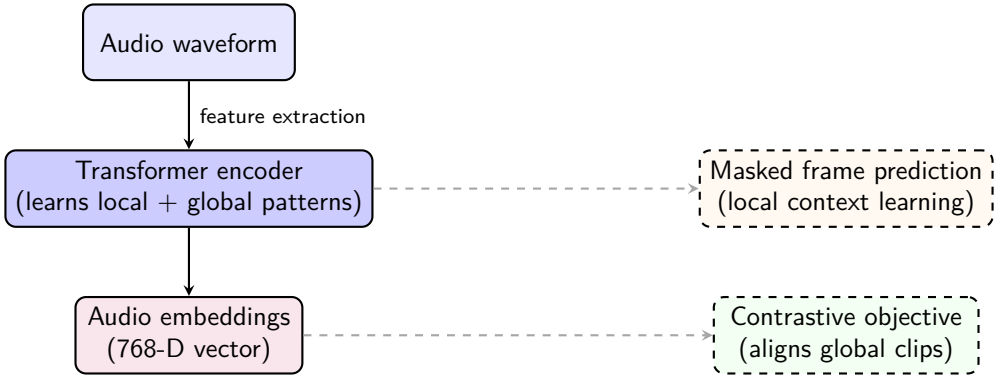
How CLAP Learns Meaning from Audio and Text

CLAP (Contrastive Language–Audio Pretraining): A multimodal model that learns a shared representation for **audio and text** by aligning them through a contrastive learning objective (similar to CLIP for images).



How MERT Learns Musical Meaning

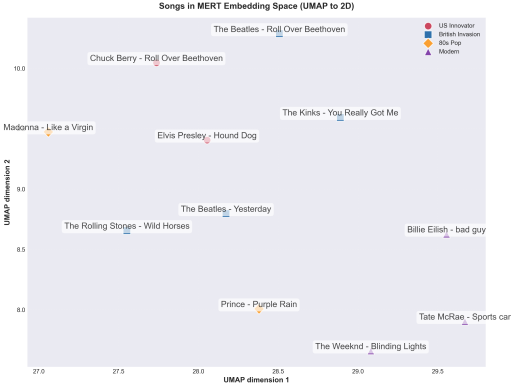
MERT (Music Understanding Transformer): A deep **audio-only** model trained in a self-supervised way to learn music representations from large unlabeled datasets.



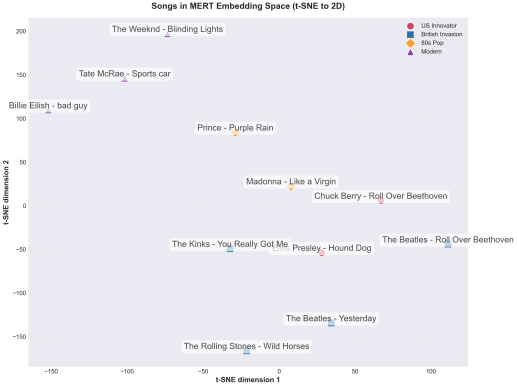
Demo: Mapping Songs in MERT Embedding Space

We take 11 representative songs spanning four major eras and genres (early rock, British Invasion, 80s pop, and modern pop) and project their high-dimensional **MERT audio embeddings** into 2D using UMAP and t-SNE to visualize stylistic proximity `GitHub Repo`.

UMAP projection



t-SNE projection



Identification Ideas

1. **Event Studies:** Exploit discrete shocks (streaming rollout, algorithm updates, or new production technologies) to trace changes in audio embeddings and stylistic diversity before vs. after.
2. **Diffusion and Convergence:** Track cross-country or cross-genre similarity in embedding space. Use timing of platform entry or cultural exposure as quasi-random variation.
3. **Within-Artist and Label Variation:** Compare an artist's sound before and after producer or label switches. Fixed-effects framework isolates production-side influences.
4. **Etc.**

Data Access Challenges

Challenge: *Finding copyright-safe, historical audio data for large-scale analysis.*

- Most commercial recordings are under copyright, so direct downloads (e.g., YouTube-to-MP3) are **not legally feasible**.
- For this **demo**, I converted each song from YouTube to MP3 manually, but this method **does not scale** or meet research standards.

Existing open datasets:

- **Million Song Dataset (MSD)** – metadata + 7digital 30 s previews.
- **MTG-Jamendo** – Creative Commons songs (mostly 2000s–2010s).
- **Free Music Archive (FMA)** – open-licensed, genre-balanced tracks.
- **AcousticBrainz** – large-scale feature dataset linked to MSD.

Open datasets offer valuable proxies but limited coverage of earlier decades.

Building a Feasible 1960–1980 Dataset

One approach to construct a legal panel of popular songs across decades.

1. Song lists:

- Billboard Hot 100 (US) and Official Charts UK Top 100, 1960–1980.
- Keep: `title`, `artist`, `chart_date`, `rank`, `country`.

2. Audio linkage:

- Fuzzy-match `artist/title/year` to the Million Song Dataset metadata.
- Retrieve `7digital track ID` and use the 7digital API to obtain 30 s previews.
- Follows the approach of *Mauch et al. (2015)* for legal large-scale analysis.

Then embed songs with MERT/CLAP...

Key Papers

- **Serrà et al. (2012), Measuring the Evolution of Contemporary Western Popular Music** Large-scale analysis of pitch, timbre, and loudness (1950–2010) shows reduced timbral variety but increasing loudness and harmonic uniformity.
- **Mauch et al. (2015), The Evolution of Popular Music: USA 1960–2010** Uses audio features and topic modeling to identify stylistic "revolutions" (e.g., 1983, 1991) and long-run trends in harmony and timbre.
- **Interiano et al. (2018), Musical Trends and 50 Years of Data** Analyzes 500,000 songs to show declining happiness and rising relaxation in popular music lyrics and sound.
- **Tzanetakis & Cook (2002), Musical Genre Classification of Audio Signals** Foundational work introducing timbral, rhythmic, and pitch-based features for automatic genre recognition.
- **Hendricks & Sorensen (JPE, 2009), Information and the Skewness of Music Sales** Empirical analysis of how information diffusion affects the concentration of sales in the music industry, showing that better information leads to greater inequality in market outcomes.

Thank you!