# Notebook 1 - Establishing the Data

**Author:** Kavan Wills

**Computing ID:** meu5cg

**Course:** DS 2023 - Communicating with Data

---

## Purpose

This notebook establishes the BiMMuDa dataset for analysis of lyrical complexity trends in Billboard's top 5 songs from 1950 to 2022.

---

## Contents



---

## How to Get the Data

This project uses the **Billboard Melodic Music Dataset (BiMMuDa)**.

To obtain the data used in this notebook:

1. Visit the dataset landing page: https://docs.google.com/document/d/17EyW-bA8oppRZ_3KloYB5Z-hJ-1L7ASldz6GH_BUd_g/edit?tab=t.0.
2. Download the file `bimmuda_per_song_full.csv`.
3. Save `bimmuda_per_song_full.csv` in the same directory as this notebook (or update the file path in the `pd.read_csv()` call below).

This notebook assumes that `bimmuda_per_song_full.csv` is available locally in that location.

---

# Describe Data Source

## Who Produced the Data

The Billboard Melodic Music Dataset (BiMMuDa) was created by Madeline Hamilton, Ana Clemente, Edward Hall, and Marcus Pearce, and published in Transactions of the International Society for Music Information Retrieval (2024).

**Citation:** Hamilton, M., Clemente, A., Hall, E., & Pearce, M. (2024). The Billboard Melodic Music Dataset (BiMMuDa). Transactions of the International Society for Music Information Retrieval, 7(1), 113-128. https://doi.org/10.5334/tismir.168

## How the Data Was Produced

The Billboard Melodic Music Dataset (BiMMuDa) was compiled through:

1. **Identifying songs** in the Billboard year-end Top 5 charts from 1950–2022.
2. **Manually transcribing** the main vocal melodies of each song into MIDI and score files (e.g., MuseScore), including segmentation into sections such as verse and chorus.
3. **Collecting lyrics** for each song from public lyrics websites and storing them as plain-text files.
4. **Extracting metadata** such as tonic, mode, and tempo from Tunebat.com, followed by manual checking and correction where needed.
5. **Computing lyrical attributes** (Number of Words, Number of Unique Words, Unique Word Ratio, Number of Syllables) directly from the lyric text files.
6. **Quality assurance** via cross-checking a subset of melodies against other corpora and manual correction.

This process produces a per-song CSV like `bimmuda_per_song_full.csv`, where each row is one song and each column is a song-level attribute, so the data are in tidy form.

---

# Import Libraries

In [42]:
```
# Data manipulation
import pandas as pd
import numpy as np

# Visualization
import matplotlib.pyplot as plt
import seaborn as sns
```

---

# Load Data

In [43]:
```
df = pd.read_csv('bimmuda_per_song_full.csv')
df.head()
```

Out[43]:

| | Title | Artist | Year | Position | Link to Audio |
|---|---|---|---|---|---|
| **0** | Goodnight Irene | Gordon Jenkins & The Weavers | 1950 | 1 | https://open.spotify.com/track/3GtfeLBXe15nyEM... |
| **1** | Mona Lisa | Nat King Cole | 1950 | 2 | https://open.spotify.com/track/5dae01pKNjRQtgO... |
| **2** | Third Man Theme | Anton Karas | 1950 | 3 | https://open.spotify.com/track/7rRGujA12UJcRUz... |
| **3** | Sam's Song | Gary & Bing Crosby | 1950 | 4 | https://open.spotify.com/track/1Wnlagmoyo7M7In... |
| **4** | Simple Melody | Gary & Bing Crosby | 1950 | 5 | https://open.spotify.com/track/75lpxrV9sLZlRvz... |

---

# COLS Table

The table below describes each feature in the dataset following the COLS (Columns) format:

- **Column Name:** Variable name in dataset
- **Type:** Data type (Categorical, Numeric, etc.)
- **Description:** What the variable represents
- **Range/Values:** Possible values or range

In [46]:
```
cols_table_data = [
    {
        'Column Name': 'Title',
        'Type': 'Categorical',
        'Description': 'Song title',
        'Range/Values': f'{df["Title"].nunique()} unique titles'
    },
    {
        'Column Name': 'Artist',
        'Type': 'Categorical',
        'Description': 'Artist name(s), including features',
```

```
        'Range/Values': f'{df["Artist"].nunique()} unique artists'
    },
    {
        'Column Name': 'Year',
        'Type': 'Numeric',
        'Description': 'Year song appeared in Billboard top 5',
        'Range/Values': f'{df["Year"].min()}-{df["Year"].max()}'
    },
    {
        'Column Name': 'Position',
        'Type': 'Numeric',
        'Description': 'Chart position (1=highest)',
        'Range/Values': '1-5'
    },
    {
        'Column Name': 'Link to Audio',
        'Type': 'Categorical',
        'Description': 'Spotify URL for audio track',
        'Range/Values': 'Spotify URLs'
    },
    {
        'Column Name': 'Tonic 1',
        'Type': 'Categorical',
        'Description': 'Primary key/tonic of the song',
        'Range/Values': f'{df["Tonic 1"].nunique()} unique keys'
    },
    {
        'Column Name': 'Tonic 2',
        'Type': 'Categorical',
        'Description': 'Secondary key/tonic (if modulation occurs)',
        'Range/Values': f'{df["Tonic 2"].nunique()} unique keys'
    },
    {
        'Column Name': 'Tonic 3',
        'Type': 'Categorical',
        'Description': 'Tertiary key/tonic (if multiple modulations)',
        'Range/Values': f'{df["Tonic 3"].nunique()} unique keys'
    },
    {
        'Column Name': 'Mode 1',
        'Type': 'Categorical',
        'Description': 'Primary mode (Major/Minor)',
        'Range/Values': f'{df["Mode 1"].dropna().unique()}'
    },
    {
        'Column Name': 'Mode 2',
        'Type': 'Categorical',
        'Description': 'Secondary mode (if modulation occurs)',
        'Range/Values': f'{df["Mode 2"].dropna().unique() if df["Mode 2"].nc
    },
    {
        'Column Name': 'Mode 3',
        'Type': 'Categorical',
        'Description': 'Tertiary mode (if multiple modulations)',
        'Range/Values': f'{df["Mode 3"].dropna().unique() if df["Mode 3"].nc
    },
```

```
    {
        'Column Name': 'BPM 1',
        'Type': 'Numeric',
        'Description': 'Primary tempo in beats per minute',
        'Range/Values': f'{df["BPM 1"].min():.0f}–{df["BPM 1"].max():.0f}'
    },
    {
        'Column Name': 'BPM 2',
        'Type': 'Numeric',
        'Description': 'Secondary tempo (if tempo changes)',
        'Range/Values': f'{df["BPM 2"].min():.0f}–{df["BPM 2"].max():.0f}' i
    },
    {
        'Column Name': 'BPM 3',
        'Type': 'Numeric',
        'Description': 'Tertiary tempo (if multiple tempo changes)',
        'Range/Values': f'{df["BPM 3"].min():.0f}–{df["BPM 3"].max():.0f}' i
    },
    {
        'Column Name': 'Number of Parts',
        'Type': 'Numeric',
        'Description': 'Number of distinct melodic sections (verse, chorus,
        'Range/Values': f'{df["Number of Parts"].min():.0f}–{df["Number of P
    },
    {
        'Column Name': 'Number of Words',
        'Type': 'Numeric',
        'Description': 'Total words in lyrics (with repetitions)',
        'Range/Values': f'{df["Number of Words"].min():.0f}–{df["Number of W
    },
    {
        'Column Name': 'Number of Unique Words',
        'Type': 'Numeric',
        'Description': 'Count of distinct vocabulary words',
        'Range/Values': f'{df["Number of Unique Words"].min():.0f}–{df["Numb
    },
    {
        'Column Name': 'Unique Word Ratio',
        'Type': 'Numeric',
        'Description': 'Unique Words ÷ Total Words (diversity metric)',
        'Range/Values': f'{df["Unique Word Ratio"].min():.2f}–{df["Unique Wo
    },
    {
        'Column Name': 'Number of Syllables',
        'Type': 'Numeric',
        'Description': 'Total syllables in lyrics',
        'Range/Values': f'{df["Number of Syllables"].min():.0f}–{df["Number
    }
]

cols_table = pd.DataFrame(cols_table_data)
display(cols_table)
```

| | Column Name | Type | Description | Range/Values |
|---|---|---|---|---|
| **0** | Title | Categorical | Song title | 369 unique titles |
| **1** | Artist | Categorical | Artist name(s), including features | 307 unique artists |
| **2** | Year | Numeric | Year song appeared in Billboard top 5 | 1950-2022 |
| **3** | Position | Numeric | Chart position (1=highest) | 1-5 |
| **4** | Link to Audio | Categorical | Spotify URL for audio track | Spotify URLs |
| **5** | Tonic 1 | Categorical | Primary key/tonic of the song | 15 unique keys |
| **6** | Tonic 2 | Categorical | Secondary key/tonic (if modulation occurs) | 8 unique keys |
| **7** | Tonic 3 | Categorical | Tertiary key/tonic (if multiple modulations) | 1 unique keys |
| **8** | Mode 1 | Categorical | Primary mode (Major/Minor) | ['Major' 'Minor'] |
| **9** | Mode 2 | Categorical | Secondary mode (if modulation occurs) | ['Major' 'Minor'] |
| **10** | Mode 3 | Categorical | Tertiary mode (if multiple modulations) | ['Major'] |
| **11** | BPM 1 | Numeric | Primary tempo in beats per minute | 57-174 |
| **12** | BPM 2 | Numeric | Secondary tempo (if tempo changes) | 80-167 |
| **13** | BPM 3 | Numeric | Tertiary tempo (if multiple tempo changes) | 128-128 |
| **14** | Number of Parts | Numeric | Number of distinct melodic sections (verse, ch... | 1-8 |
| **15** | Number of Words | Numeric | Total words in lyrics (with repetitions) | 12-896 |
| **16** | Number of Unique Words | Numeric | Count of distinct vocabulary words | 11-312 |
| **17** | Unique Word Ratio | Numeric | Unique Words ÷ Total Words (diversity metric) | 0.10-1.00 |
| **18** | Number of Syllables | Numeric | Total syllables in lyrics | 22-1064 |

---

# Summary

**Dataset:** BiMMuDa (Billboard Melodic Music Dataset)
**Songs:** 371 from Billboard's year-end top 5 (1950-2022)
**Focus:** Lyrical complexity analysis

---

## Key Features:

- **Number of Words** - Total word count
- **Number of Unique Words** - Vocabulary size
- **Unique Word Ratio** - Diversity metric (0-1 scale)
- **Number of Syllables** - Total syllable count
- **Year** - Temporal variable for trend analysis

---

## Next: Notebook 2 - Data Exploration