The Suffering Data Scientists Department

DS 5030
November 13, 2025

Erin Siedlecki, Shaveen Saadee, Anna Li, Emily Garman, Razan Habboub, & Marissa Burton


AGENDA
The Data
Markov Transition Model
Our Model
Results
Conclusion

Our goal: to model Taylor Swift's chord progressions and generate similar ones using a Markov Transition Model

The Data: Chordonomicon

- Consists of data from 666,000 chord progressions and their corresponding song's release date, decade, genre, and Spotify song and artist IDs
- Created by researchers at Artificial Intelligence and Learning Systems Laboratory at the National Technical University of Athens
- Collected the chord data using web scraping on the Ultimate Guitar platform and Spotify Web API
- Goal was to generate a large-scale dataset that can be used for advanced machine learning
- Performed experiments on chord prediction and decade and genre classification

Data Cleaning & Inspection

- We limited our search to music by Taylor Swift, ending up with 185 songs
- To clean the data, we formatted the source data and mapped chords to a 12-semitone binary representation
- The data is updated as of 12/3/24, so not all of Taylor Swift's music is captured
- We found missing values in the original dataset (filtered for Taylor Swift) for the columns release_date, decade, rock_genre, and spotify_song_id
- None of the columns containing missing values are used in our chord-specific dataset which will be used to create the chord generator, so we are not concerned with the presence of these missing values

Background

- **Chord Progressions** are sequences of chords forming the harmonic foundation of music, shaping a song's **mood** and **emotional tone**
  - **Example contrasts:**
    - ***"Blackbird"*** **–** The Beatles: soft and gentle / G - Am7 - G/B - G (I – ii^7 – I^6 – I)
    - ***"Fly Me to the Moon"*** **–** Frank Sinatra: jazz-style sequence / Am7–Dm7–G7–Cmaj7 (vi^7 – ii^7 – V^7 – I^7)
- Genres differ: **Rock/Jazz** use complex progressions for depth, while **Pop** favors simple, catchy ones
- **Taylor Swift consistently uses three main chord progressions:**
  - ***"Blank Space"*** → I–V–vi–IV (F – Dm – B ♭ – C), I-V-vi-IV (F – C – Dm – B ♭)
  - ***"You Belong With Me"*** → I–V–ii–IV (C - G - Dm - F)
  - ***"All Too Well"*** → I–V–vi–IV (C - G - Am - F)
- Taylor Swift's use of **simple, repetitive progressions** builds a **cohesive musical identity**, emphasizing lyrics and emotion over harmonic complexity
- Modeling her progressions can reveal: chord transition patterns, dominant chord types, and structure trends across songs

chord	# count
G	3833
C	3677
D	2288
F	2228
Amin	2168
Emin	1304
A	955
Dmin	514
Bmin	425
Cadd9	236

Markov Transition Model

- The variable of interest is “chord”, which lends well to a Markov Transition Model (because values are categorical)
- The Markov Chain allows us to examine a current observation (state) in a sequence and predict the next (and only the next) state based on prior data
- For this data specifically, our goal is to create a state space of chord observations from prior Taylor Swift chord progression data
- Our model will rely on the highest probabilities of one chord following another as opposed to what 4 chords together would mimic a Taylor Swift song best, which may lead to our state space sounding chaotic

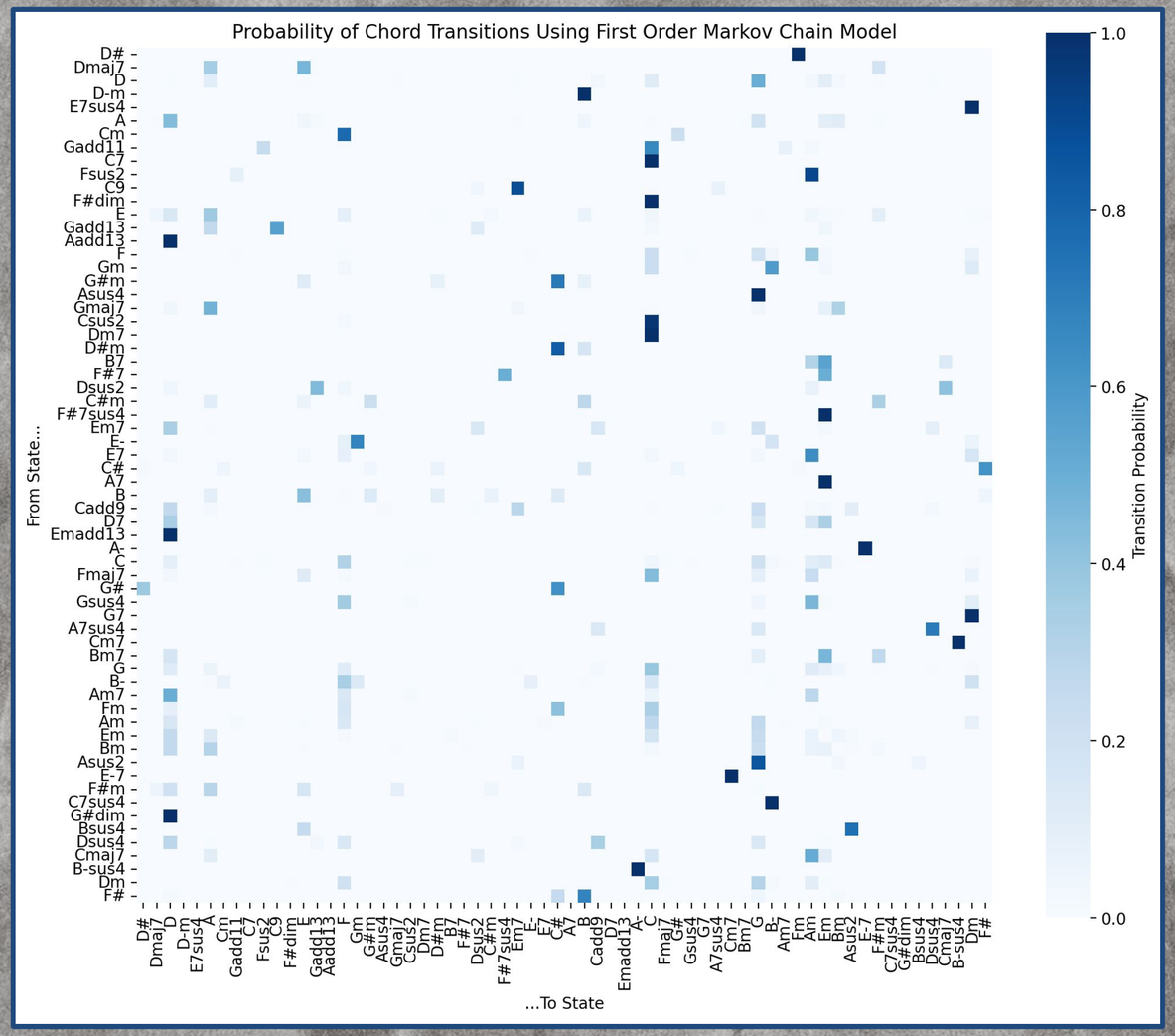
Our Model

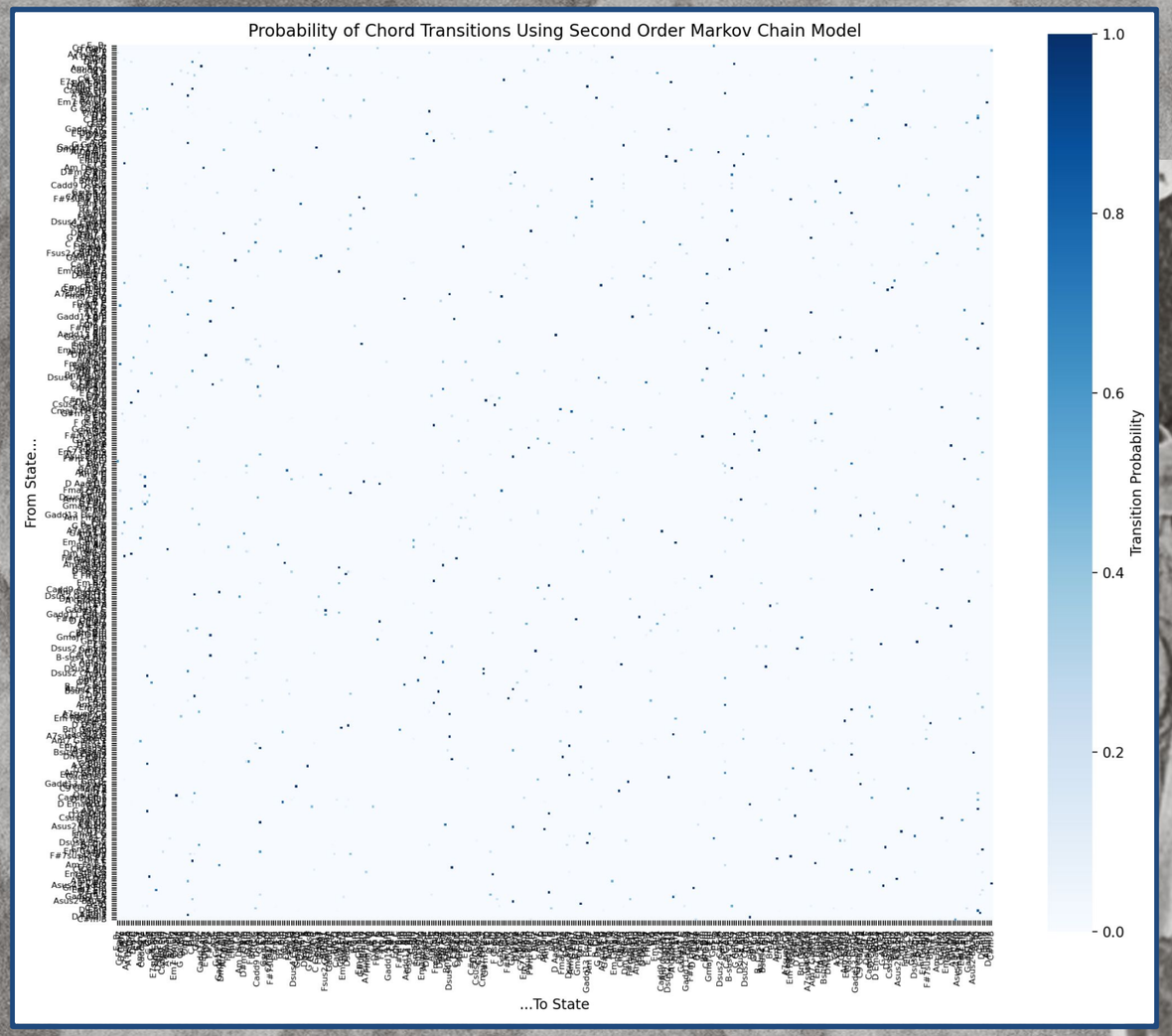
- Second-order model that looks at the previous two chords to determine the next chord
  - Initial state: C G
  - Better for capturing local musical structure

- Simulation to generate a new sequence of chords that sound similar to Taylor Swift's music transitions
  - Count how often each state is followed by another to compute transition probabilities
  - From the initial state, sample the next state based on transition probabilities
  - Keep sliding the two-chord window forward until you have enough

Our Model

```
States:
 ['G#' 'A7sus4' 'F' 'C7' 'Am' 'G' 'Asus2' 'F#7sus4' 'Cm7' 'C7sus4' 'B-sus4'
 'G#dim' 'E' 'Cmaj7' 'C9' 'Cadd9' 'Gsus4' 'G#m' 'F#' 'Bsus4' 'B' 'Fmaj7'
 'Gm' 'Asus4' 'Em7' 'Dmaj7' 'Dm' 'G7' 'Bm7' 'Am7' 'C#m' 'A-' 'Dsus4' 'Cm'
 'Gadd11' 'D#m' 'Fm' 'Gmaj7' 'Fsus2' 'Dsus2' 'D' 'F#7' 'Emadd13' 'Dm7'
 'B7' 'F#dim' 'D-m' 'E-7' 'B-' 'E7sus4' 'A' 'Gadd13' 'E7' 'C#' 'D#' 'E-'
 'F#m' 'Bm' 'D7' 'Csus2' 'Em' 'C' 'A7' 'Aadd13']
```

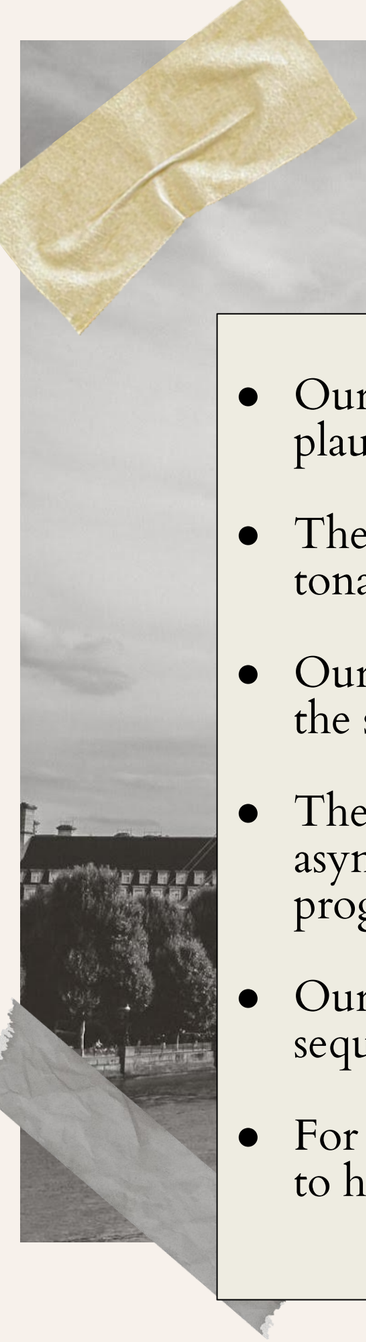


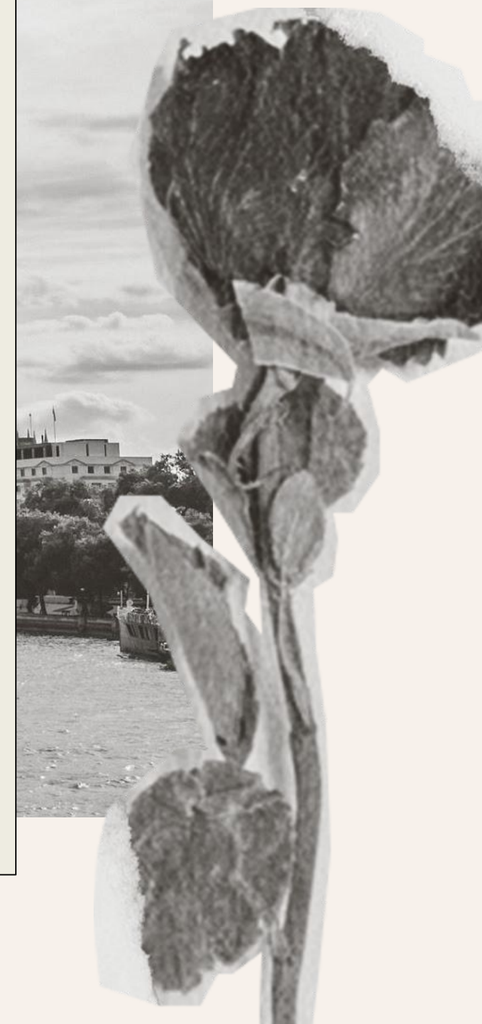


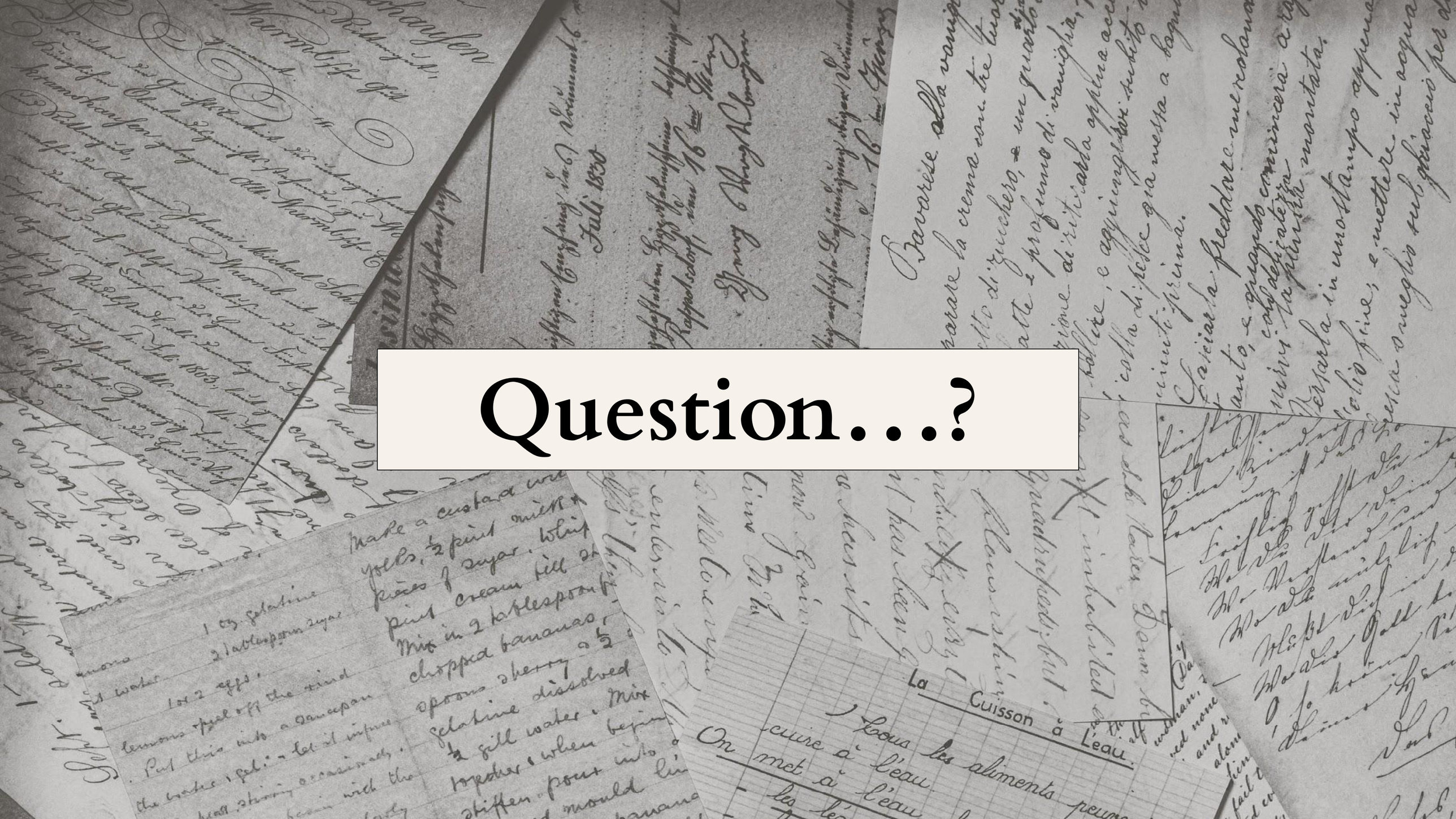
A New Taylor Swift Collab…

So Long, Markov

- Our model successfully reproduced the same set of chords as the training data and encoded plausible local transitions, but the simulated sequences differed perceptibly from the real songs
- The generated MIDI sounded less coherent and lacked the larger-scale structure, repetition, and tonal stability present in the original compositions
- Our approach effectively captured short-term dependencies between chords, but was limited by the simplicity of the Markov assumption and the sparsity of the training data
- The probability that we accurately predict a chord progression longer than 2 chords is asymptotically close to 0, so we should not rely on a Markov chain to mimic a typical 4-chord progression in Taylor Swift's discography
- Our model ignored rhythm, melody, and long-range harmonic structure, which led to generated sequences that often sounded musically inconsistent compared to the original compositions
- For future work we could use larger and more diverse datasets, apply smoothing or regularization to handle rare transitions, and/or explore higher-order or hybrid probabilistic models






Question…?