Harmonic progression is one of the cornerstones of tonal music composition and is thereby essential to many musical styles and traditions. Previous studies have shown that musical genres and composers could be discriminated based on chord progressions modeled as chord n-grams. These studies were however conducted on small-scale datasets and using symbolic music transcriptions.

In this work, we apply pattern mining techniques to over 200,000 chord progression sequences out of 1,000,000 extracted from the I Like Music (ILM) commercial music audio collection. The ILM collection spans 37 musical genres and includes pieces released between 1907 and 2013. We developed a single program multiple data parallel computing approach whereby audio feature extraction tasks are split up and run simultaneously on multiple cores. An audio-based chord recognition model (Vamp plugin Chordino) was used to extract the chord progressions from the ILM set. To keep low-weight feature sets, the chord data were stored using a compact binary format. We used the CM-SPADE algorithm, which performs a vertical mining of sequential patterns using co-occurence information, and which is fast and efficient enough to be applied to big data collections like the ILM set. In order to derive key-independent frequent patterns, the transition between chords are modeled by changes of qualities (e.g. major, minor, etc.) and root keys (e.g. fourth, fifth, etc.). The resulting key-independent chord progression patterns vary in length (from 2 to 16) and frequency (from 2 to 19,820) across genres. As illustrated by graphs generated to represent frequent 4-chord progressions, some patterns like circle-of-fifths movements are well represented in most genres but in varying degrees.

These large-scale results offer the opportunity to uncover similarities and discrepancies between sets of musical pieces and therefore to build classifiers for search and recommendation. They also support the empirical testing of music theory. It is however more difficult to derive new hypotheses from such dataset due to its size. This can be addressed by using pattern detection algorithms or suitable visualisation which we present in a companion study.