Report 7

Reflection about our results and the significance of our findings

- I am worried about the quality and significance of our data.

- In our original dataset, I queried spotify's database by looking for 32 different musical genres (API call - python). For each one of these genres, I got up to 50 artists. Then I got the complete discography of each one of these artists (with a limitation of 50 albums per artist). This yielded 1.3k artists and 7k albums after removing duplicates and albums with an unusual number of tracks.

- This was the most efficient way that I found for querying albums from Spotify (API methods - python). An alternative approach would be to query directly for artists and their discographies, but then I would need a large list of artists to begin with. WIth the query by genre, I was able to get up to 50 artists for each genre, but then I had to have many different genres.

- The problems: there were albums with very low popularity, from "not so common" musical genres, such as "neurofunk", "celtic" and "sludge". I included these genres hoping that they would increase the representativity of our data, and also increase our sample size.

- New query and data collection: one way to possibly control for the quality of our dataset is to **query albums by labels**. In this approach I can get, for instance, all ECM and Sony records that have been recorded throughout history.

- My intuition is that this will offer not only more albums, but also a set of albums with higher quality than the albums that I previously got from Spotify.

- This is not possible through Spotify's API, so I've started working on a **parallel query from Discog's API**. So far, I've been able to get around 200k albums, which is considerably more than our previous sample of 7k. Next step would be to find the corresponding data from each album on Spotify. (API methods and calls here - python).

- With this new query, we will be sure that there was a label behind the production and distribution of our albums, and this might entail a larger control over the way that the albums are produced (if compared to independent productions, for instance).

- Just out of curiosity, I ran our analysis on ECM albums, and pretty much all of our results were slightly improved. This doesn't mean anything on its own, but it makes me curious as to where we can go with a new and better dataset.

- **Next steps**: My plan is to finnish this new query, and re-run every analysis that we have made so far. I believe that I can do it by the end of the next month or so.

- **Final thoughts:** Even though I am worried about the quality of our data, I believe we have some interesting results reported in the paper. We might need to think if this is worth publishing or not. If so, where would be a good place.