Erasmus School of Economics

What music are you in the mood for?

A hierarchical cluster analysis on music attributes and recommendations for mood playlists

EBB5

Group 3

6th of November 2020

Cathelijne Keser
Lucas Hagelstein
Kars-Jan Giesen
Lars Cretier

Agenda

1. Recap of hierarchical clustering theory

2. Research aim and data

3. Model diagnostics

4. Analysis and results

5. Application and conclusion

6. Discussion and future research

Theory recap: Hierarchical Clustering

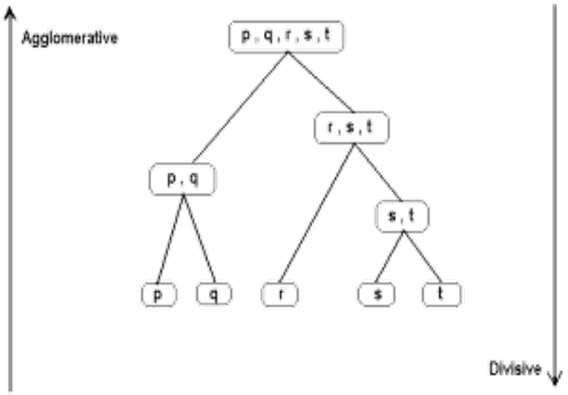
Hierarchical clustering

"Hierarchical methods proceed by stages producing a sequence of partitions, each corresponding to a different number of clusters" (Fraley & Raftery, 1998)

As opposed to K-means clustering, hierarchical clustering does **not require pre-specification** of the number of clusters K.

Agglomerative clustering

- Known as AGNES
- Bottom-up approach



Divisive clustering

- Known as DIANA
- Top-down approach

Measuring (dis)similarity between two clusters

There are various linkage methods to measure (dis)similarity:

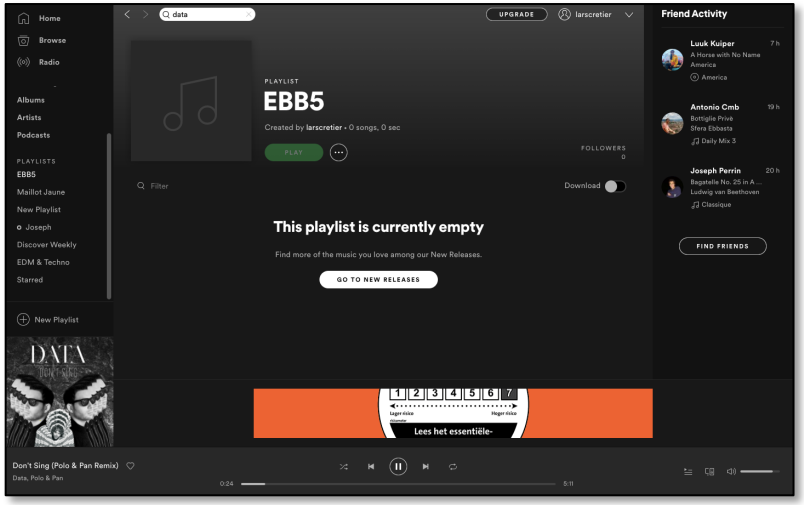
	Method	Formula	Description
1	**Single (minimum) Linkage**	$D_{12} = min_{ij} d(X_i, Y_j)$	This is the distance between the **closest** members of the two clusters
2	**Complete (maximum) Linkage**	$D_{12} = max_{ij} d(X_i, Y_j)$	This is the distance between the members that are **farthest** apart (most dissimilar)
3	**Average (mean) Linkage**	$D_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j)$	This method involves looking at the distances between **all pairs and averages all** of these distances
4	**Ward's Method**	$D_{12} = \sqrt{\frac{2.\|k\|.\|l\|}{\|k\| + \|l\|}} . \|\|\bar{x} - \bar{y}\|\|$	This method minimizes the **total within-cluster variance.** Those clusters are combined whose merger results in minimum information loss

Agenda

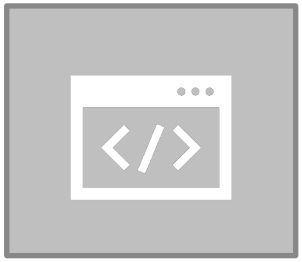
1. Recap of hierarchical clustering theory
2. **Research aim and data**
3. Model diagnostics
4. Analysis and results
5. Application and conclusion
6. Discussion and future research

Research aim

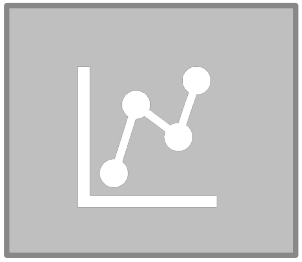
1. To provide a music recommendation approach by clustering music based on multiple attributes
2. To create music playlists for customers based on moods



Data



- Available on Kaggle
- With a developer account, you can scrape the data from Spotify



- Top 10 songs from 2010 to 2019 by year
- Worldwide

- Dataset contains 603 observations
- 14 variables in dataset

Variables

No. included of variables: 9

Variable	Explanation	Min	Max
BPM	Beats Per Minute (BPM): tempo of the song	43	206
Energy	Energy of the song by a combination of intensity and activity (100 is high energy)	4	98
Danceability	The suitability of a track for dancing based on multiple musical elements (100 is very danceable)	23	97
Loudness (dB)	The average loudness of a track in decibels (0 is loud)	-15	-2
Liveness	The presence of an audience (100 is high prob. of being live track)	2	74
Valence	The positiveness of the track (100 is very positive)	4	98
Duration	Duration of track in seconds	134	424
Acousticness	Confidence interval of the track being acoustic (100 is high prob. of being acoustic song)	0	99
Speechiness	The degree of spoken words (100 is only spoken)	3	48

Source: Spotify Developer

Data preparation

1

Cleaning the data

- **Remove NA's** and impossible values
- **Remove duplicate values** (even if it occurs multiple times in the top 50 charts)

2

Scaling the data

- Values must be **comparable** and should be measured in the **same unit**
- If variables are measured in different units, some can become **more influential** in the clustering process

Agenda

1. Recap of hierarchical clustering theory
2. Research aim and data
3. **Model diagnostics**
4. Analysis and results
5. Application and conclusion
6. Discussion and future research

Model diagnostics

1. Suitability of clustering the data
    - Hopkins Statistic
2. Choosing the appropriate linking method
    - Complete, Single, Average, Ward
    - Cophenetic correlation coefficient
3. Choosing the optimal number of K clusters
    - Within cluster sum of squares
    - Average silhouette width
    - Gap statistic

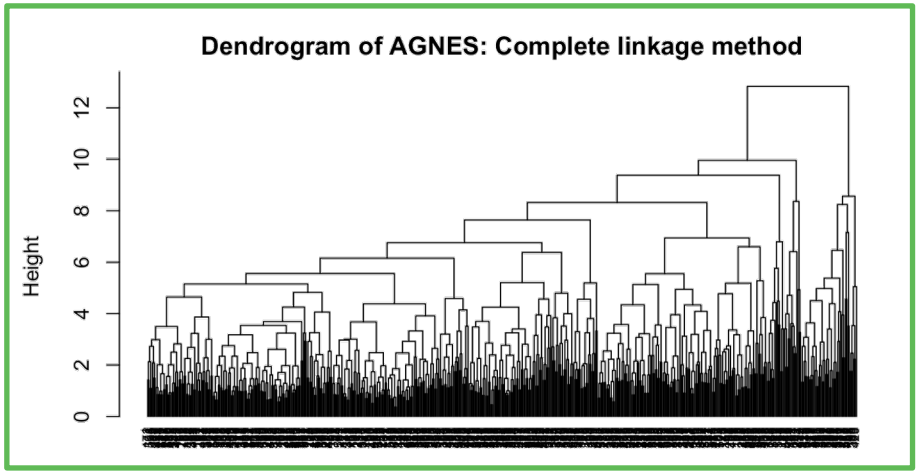
1 Suitability of clustering: Hopkins Statistic

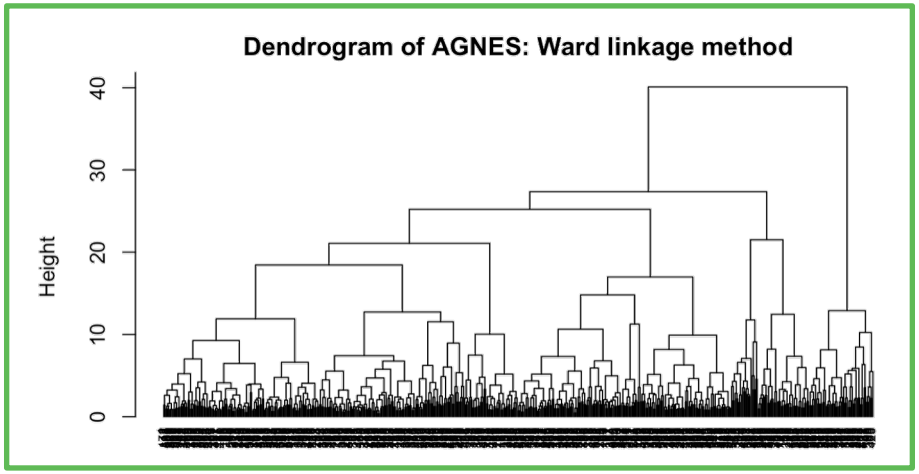
Hopkins criterion tests whether data results from a uniform random distribution.

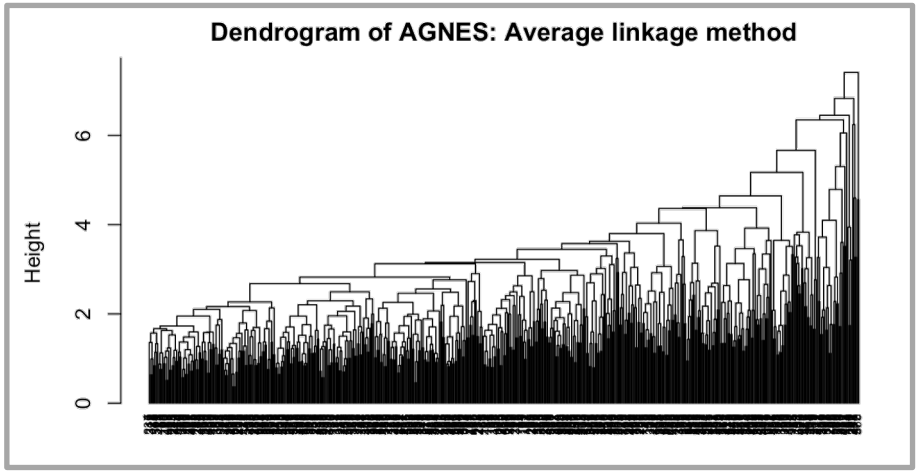
H ≈ 0.75

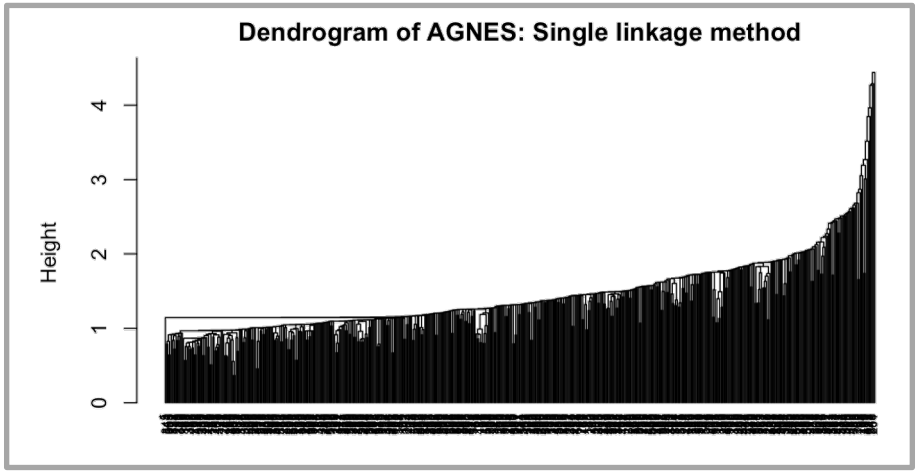
- Null hypothesis: H < 0.5, data is randomly **uniformly distributed**.
- Null hypothesis is rejected, and data contains **meaningful clusters.**

2 Choosing the appropriate linkage method









From the dendrograms, the **Complete** and the **Ward** linkage method appear to be best for interpretation purposes.

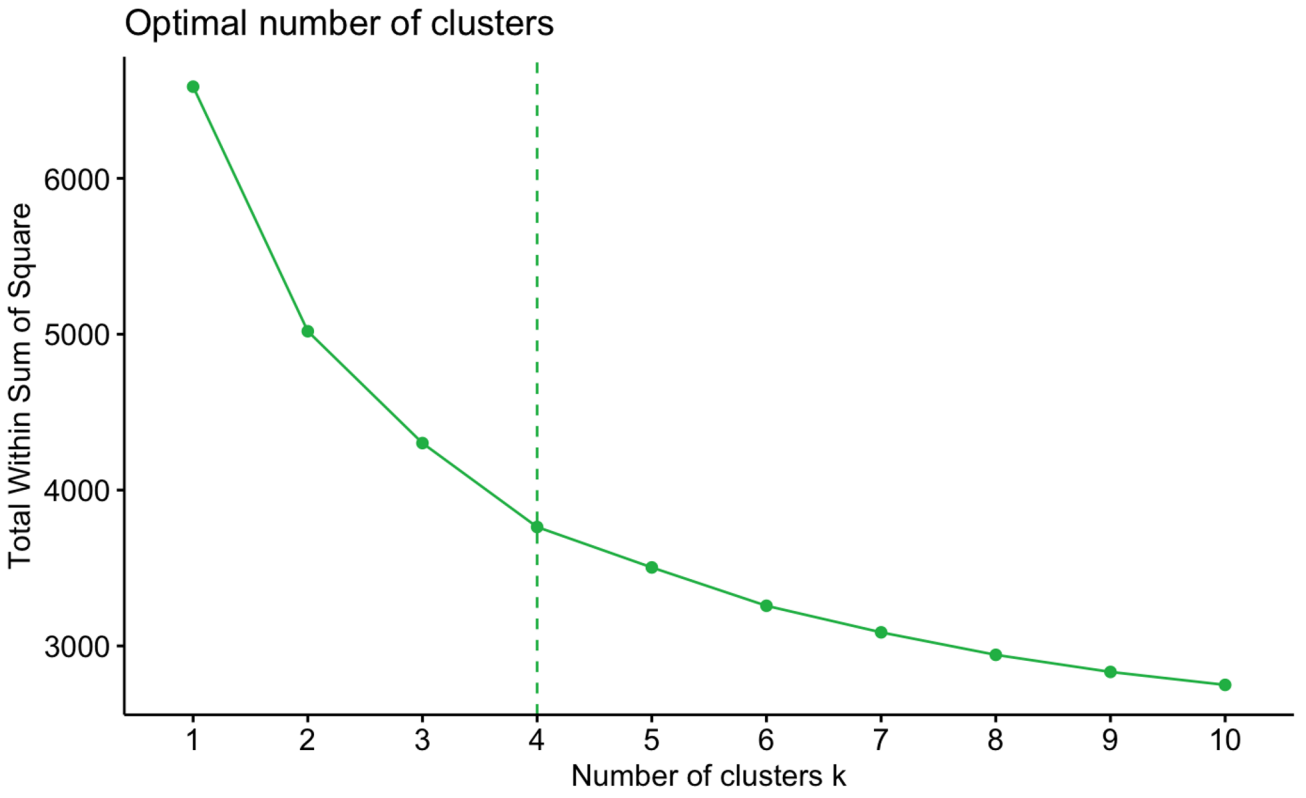
2 Choosing the appropriate linkage method

Linkage methods			
Complete	Single	Average	Ward
0.89	0.70	0.80	**0.96**

- The **cophenetic correlation coefficient** indicates how accurately the cluster tree represents dissimilarities between observations.
- From the cophenetic correlation coefficients, the **Ward linkage** method appears to be the best method.

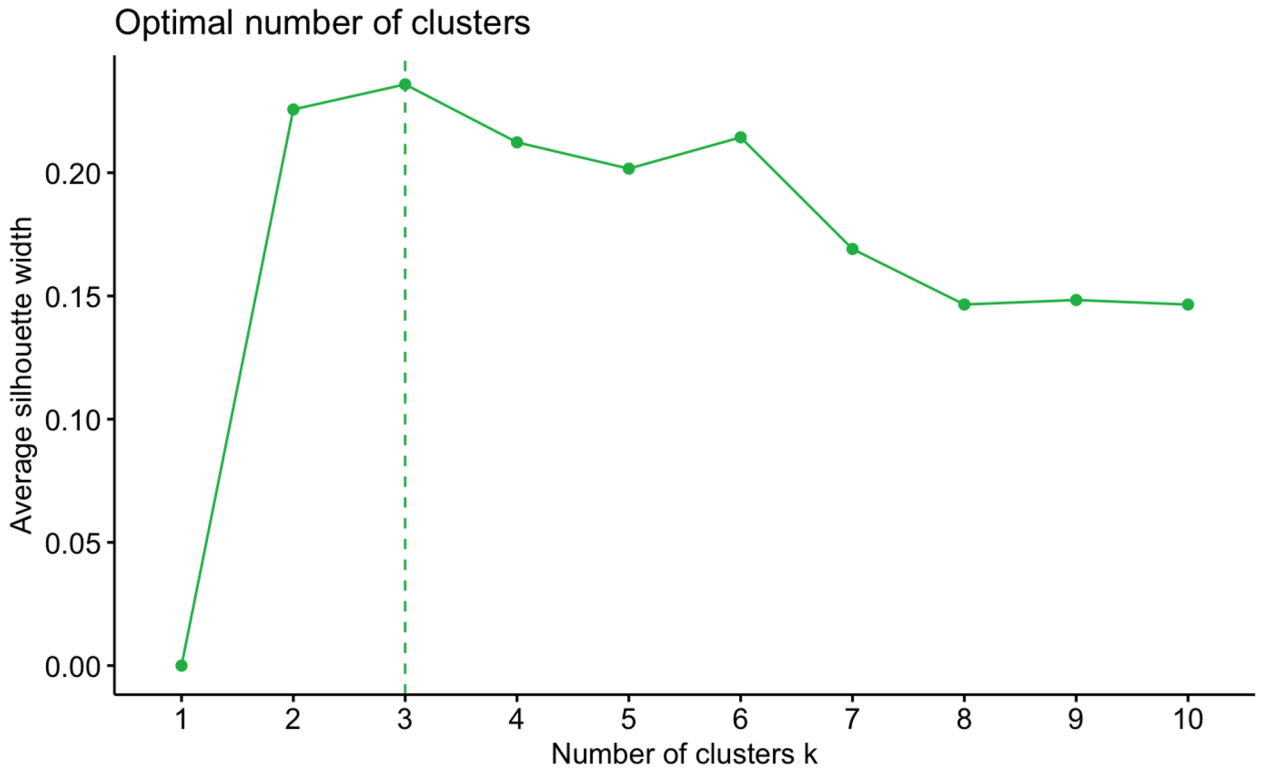
Notes: See appendix 1 for a correlation table showing the similarity in linkage method dendrogram outcomes

3 Number of K clusters: within cluster sum of squares



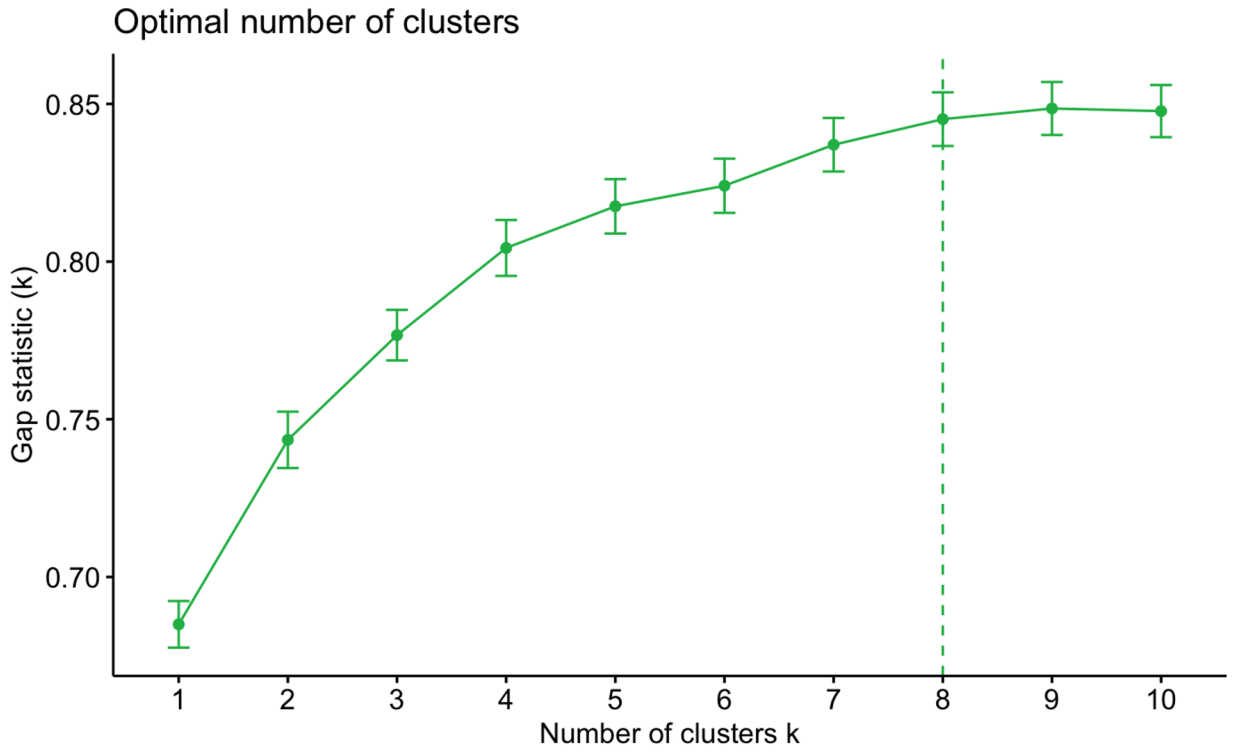
A true **'elbow'** can not be observed. However, one could argue that there is change in slope at 4 clusters, after which including 5 clusters does not add much improvement to a model that includes 4 clusters regarding the variation accounted for.

3 Number of K clusters: average silhouette



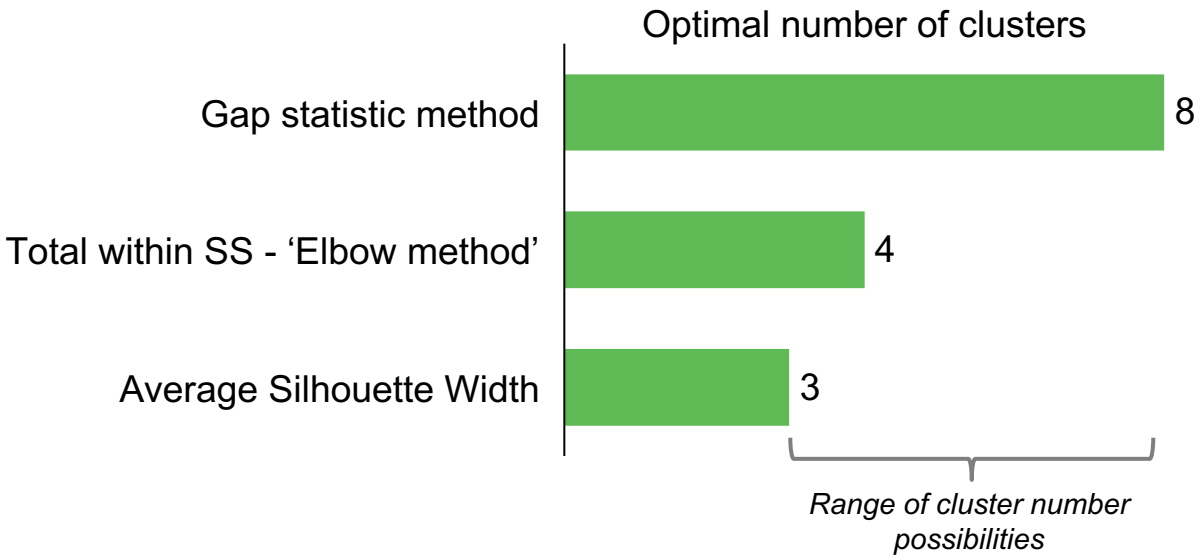
The optimal number of clusters is 3 based on the **average silhouette width**, which describes the **cohesion** of each observation within the cluster compared to the **separation** with other clusters.

3 Number of K clusters: gap statistic



The gap statistic indicates the use of **8 clusters**. The gap statistic compares the change between **within-cluster sum of squares** or variation with **expected values** under an appropriate null reference distribution.

3 Number of K clusters per diagnostic



- All methods point at a **different** optimal number of clusters.
- As the interest lies in a large number of clusters, we average the methods and **use 6 clusters**.
- Also, 6 clusters has a relative **high average silhouette width** and a **small change in the slope** of the within-cluster variation.

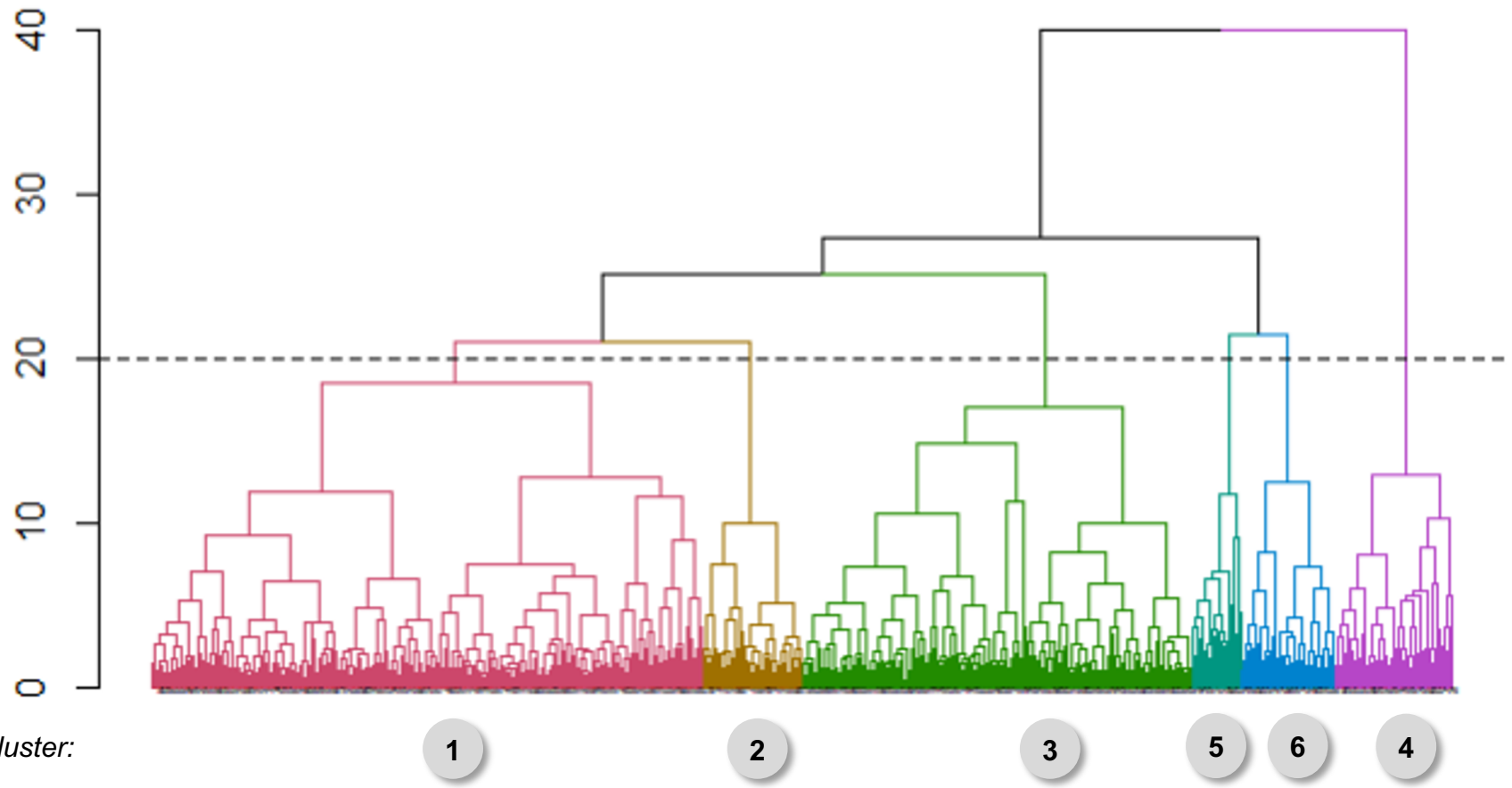
Agenda

1. Recap of hierarchical clustering theory
2. Research aim and data
3. Model diagnostics
4. **Analysis and results**
5. Application and conclusion
6. Discussion and future research

Results

Dendrogram of clusters from the analysis

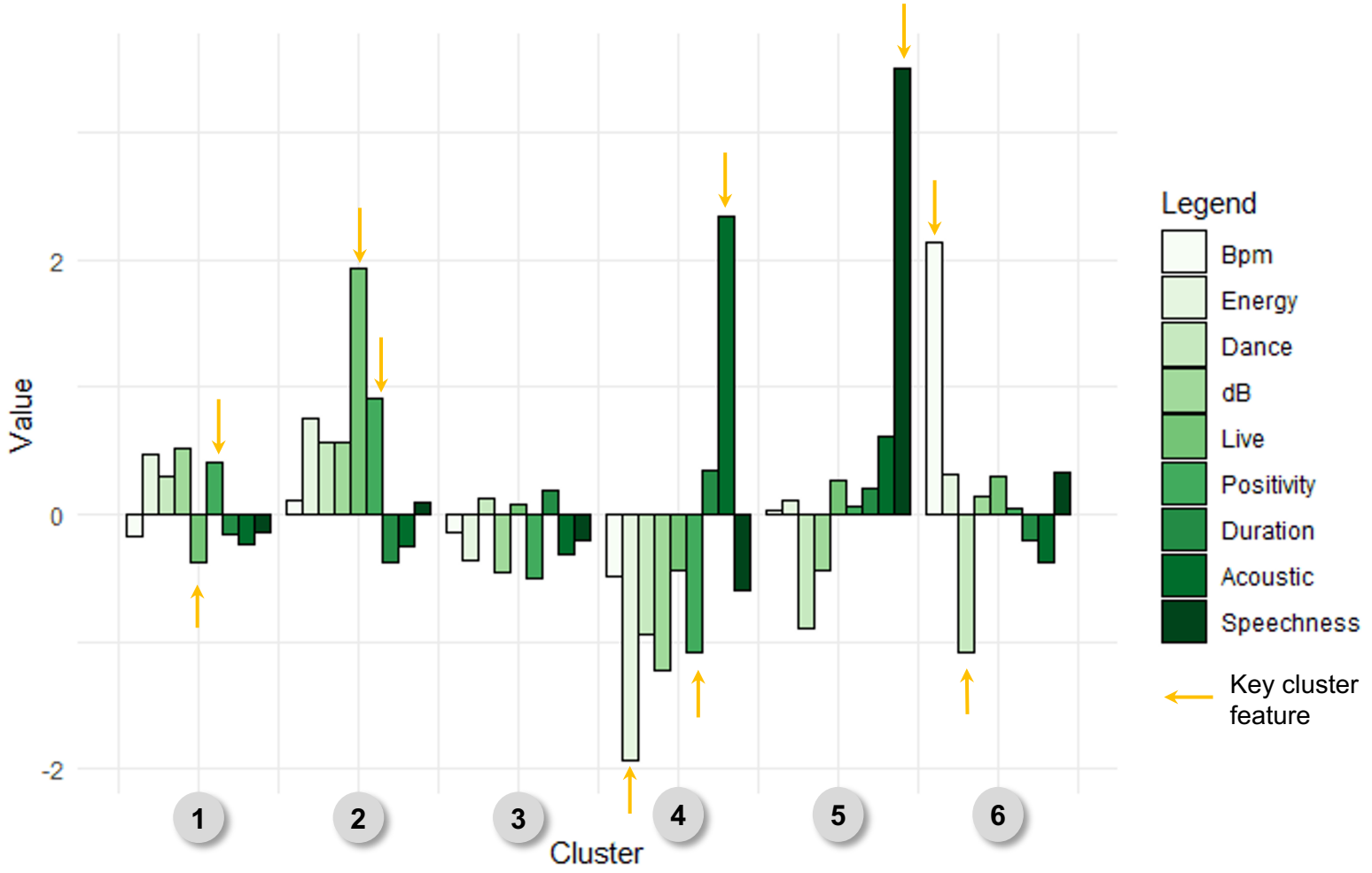
Agglomerative Clustering: Join similar clusters bottom-up



Notes: See appendix 2 and 3 for clustering outcomes with k = 4

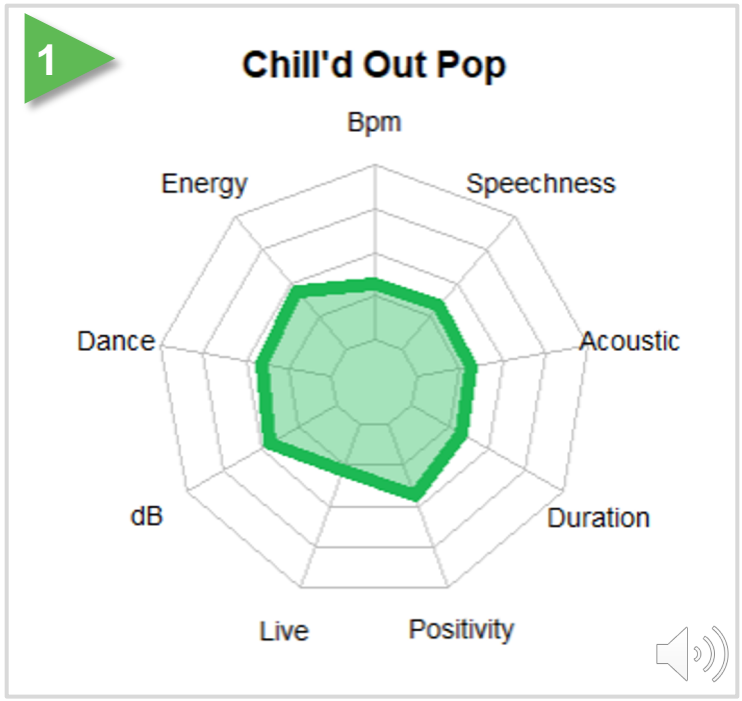
The clusters are clearly distinguishable

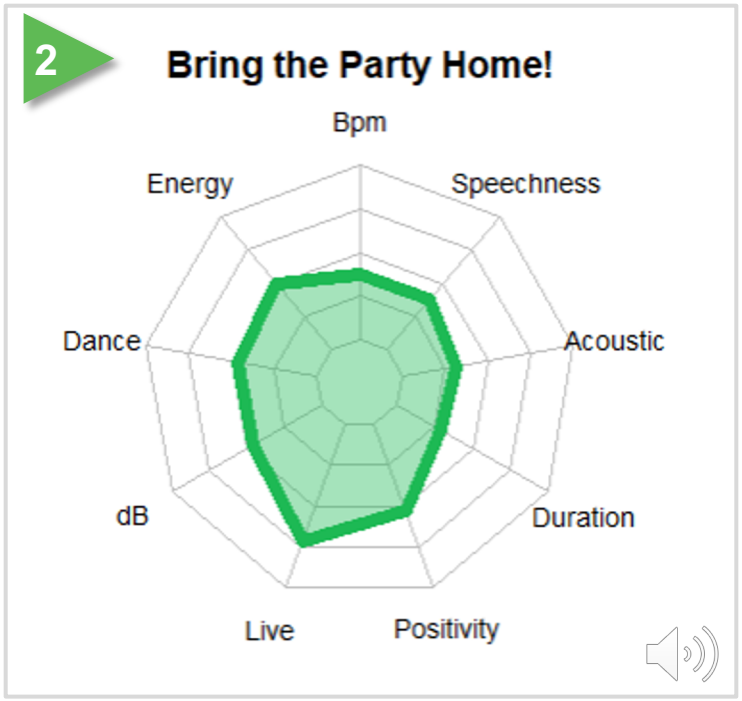
Bar plot of (scaled) cluster characteristic values



The determined clusters demonstrate **relevant and distinguishable** characteristics

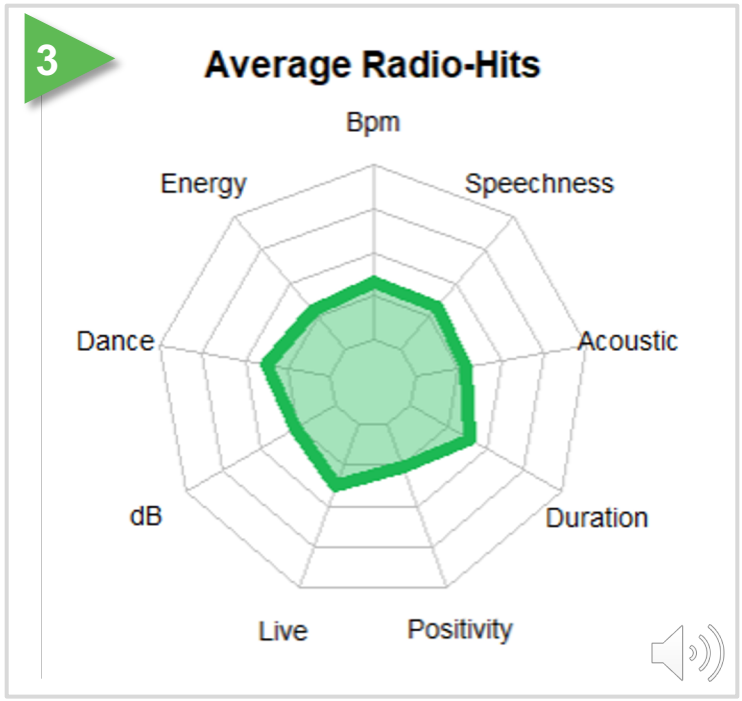
Cluster Comparison 1 and 2

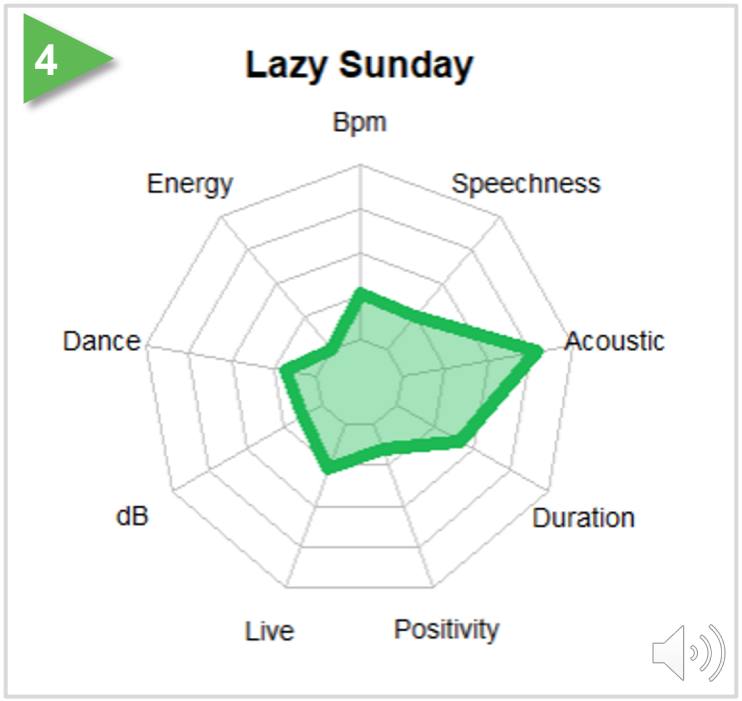




	Chill'd Out Pop	Bring the Party Home!
Size	237 obs.	44 obs.
Genre	Dance Pop (63%)	Dance Pop (64%)
Artist	Bruno Mars, Rihanna, Katy Perry	Pitbull, Jennifer Lopez, Kesha

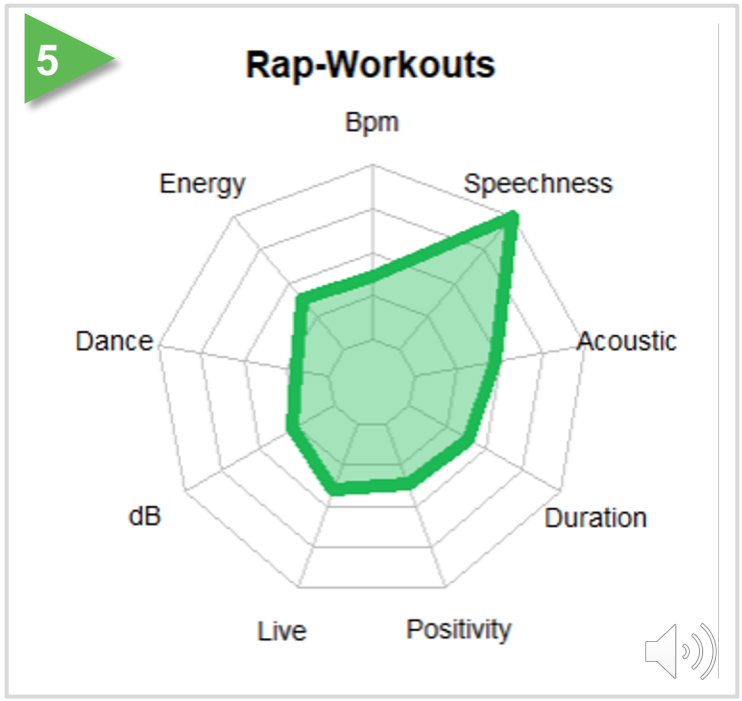
Cluster Comparison 3 and 4

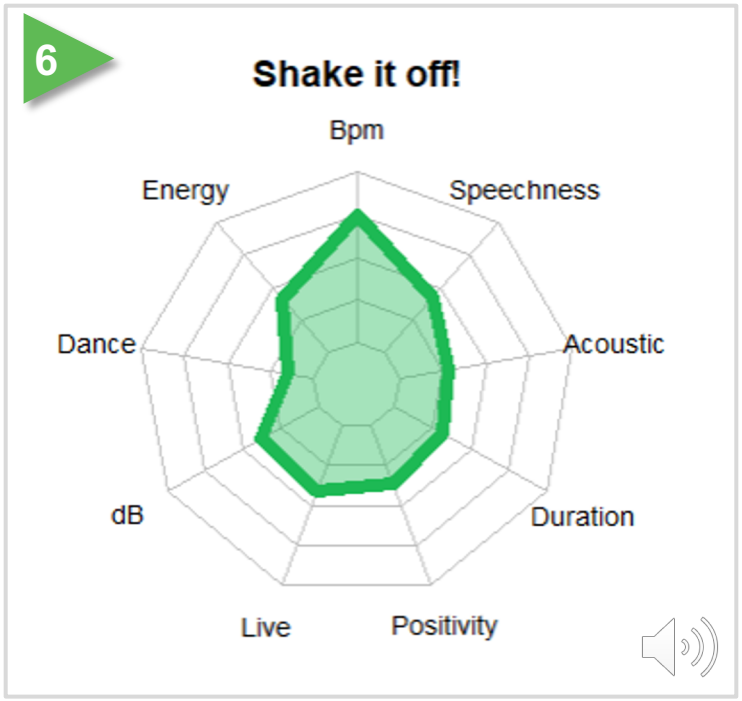




	Average Radio-Hits	Lazy Sunday
Size	175 obs.	53 obs.
Genre	Dance Pop (54%)	Dance Pop (30%)
Artist	Chainsmokers, Maroon 5	Adele, Ed Sheeran, John Legend

Cluster Comparison 5 and 6





	Rap-Workouts	Shake it off!
Size	22 obs.	42 obs.
Genre	Dance Pop (73%)	Dance Pop (38%)
Artist	Eminem, DJ Khaled, LMFAO	Katy Perry, Sia, Taylor Swift

Agenda

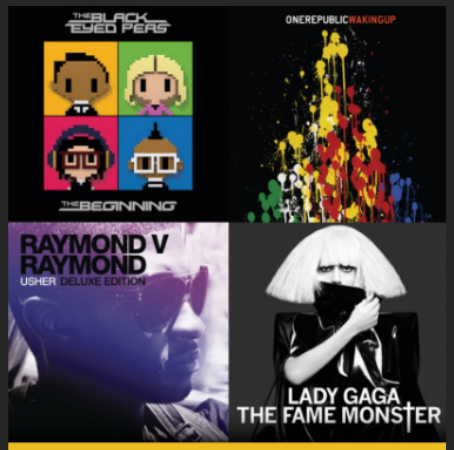
1. Recap of hierarchical clustering theory
2. Research aim and data
3. Model diagnostics
4. Analysis and results
5. **Application and conclusion**
6. Discussion and future research

The Product

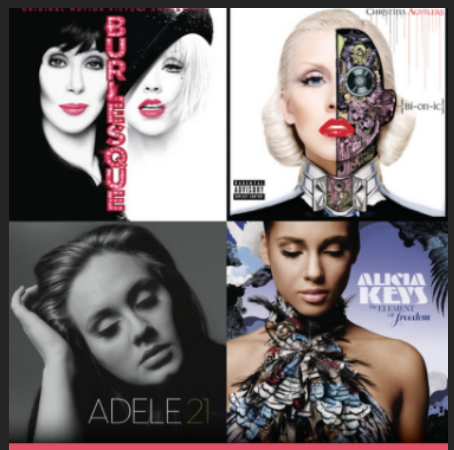
Chill'd out Pop

Bring the Party Home!

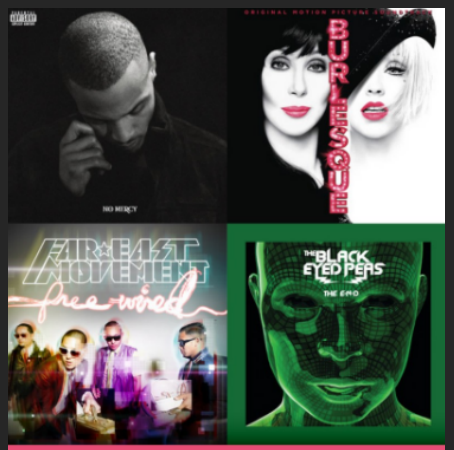
Average Radio-Hits



Lazy Sunday



Rap-Workouts



Shake it Off!

Two key outcomes with follow up steps

Conclusions

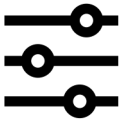
1. 6 **distinguishable clusters** based on music attributes

2. Dance music is **well represented** in the Top Charts but can be classified further

Potential next steps

1. Classification as first step to build a **recommender system**

2. **Match consumer preferences** within each cluster



3. **Detailed integratation of user-profiles** to improve matching performance

Agenda

1. Recap of hierarchical clustering theory
2. Research aim and data
3. Model diagnostics
4. Analysis and results
5. Application and conclusion
6. **Discussion and future research**

Limitations of our analysis

	Factor	Limitation	Mitigation
1	Sub optimal classification 	• Once two objects are joined, they cannot be separated • Not a globally optimising function	Use of a genetic algorithm (GA) that seeks solution wide objective[1]
2	Computational power 	• Significant computational power required • Particularly difficult for large data sets – such as Spotify's songs database	Joint dimensionality reduction clustering to increase processing speed[2]
3	Data not representative 	• Over representation of dance pop in the charts • Data comes from Top 50 playlists.	Use data more representative including more genres

1. Cowgill et al. (1999); 2. Markos, D'Enza and van de Velden (2019)

Further potential uses of hierarchical clustering in the music industry

Simplified examples

	1 Song recommender algorithm	2 Song genre classification
Explanation	Algorithms that recommend songs based on factors specific to an individual within a cluster	Algorithm to cluster personal music collections based on the genre of the music
Method	*Generate clusters then use:* - User-KNN (personalised) - Most popular item (non-personalised)	1. Computation of Inter-Recording similarities 2. Agglomerative clustering procedure 3. Genre cluster number estimation
Paper	Aguiar Neto et al. (2020)	Tsai and Bao (2010)

Thank you for listening

Any questions?

References

Aguiar Neto, F., da Costa, A., Manzato, M. and Campello, R., 2020. Pre-processing approaches for collaborative filtering based on hierarchical clustering. Information Sciences, 534, pp.172-191.

Cilibrasi, R., Vitányi, P. and Wolf, R., 2004. Algorithmic Clustering of Music Based on String Compression. *Computer Music Journal*, 28(4), pp.49-67.

Cowgill, M., Harvey, R. and Watson, L., 1999. A genetic algorithm approach to cluster analysis. *Computers & Mathematics with Applications*, 37(7), pp.99-108.

James, G., Witten, D., Hastie, T. and Tibshirani, R., n.d. *An Introduction To Statistical Learning.* Springer.

Olukanmi, P., Nelwamondo, F. & Marwala, T. Rethinking *k*-means clustering in the age of massive datasets: a constant-time approach. *Neural Comput & Applic* **32,** 15445–15467 (2020). https://doi.org/10.1007/s00521-019-04673-0

Markos, A., D'Enza, A. and van de Velden, M., 2019. Beyond Tandem Analysis: Joint Dimension Reduction and Clustering in R. Journal of Statistical Software, 91(10).

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

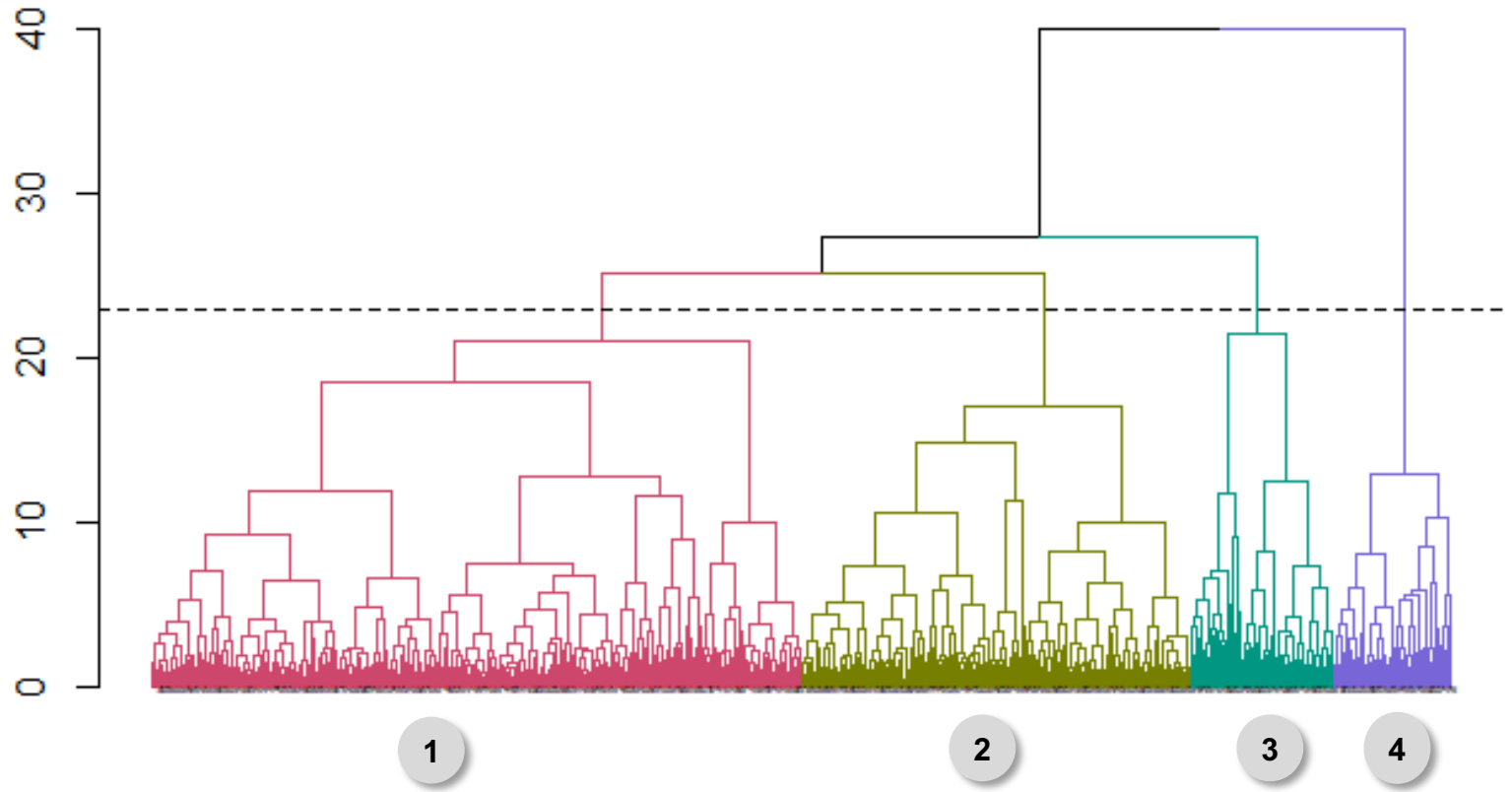
Tsai, W. and Bao, D., 2010. Clustering Music Recordings Based on Genres. *JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 26*, [online] 26, pp.2059-2074. Available at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.429.6412&rep=rep1&type=pdf> [Accessed 5 November 2020].

Appendix 1: Choosing the **appropriate linking method**

Correlation table of linking methods				
	Complete	Single	Average	Ward
Complete	1.00	*N/A*	*N/A*	*N/A*
Single	0.50	1.00	*N/A*	*N/A*
Average	0.69	0.79	1.00	*N/A*
Ward	0.67	0.32	0.58	1.00

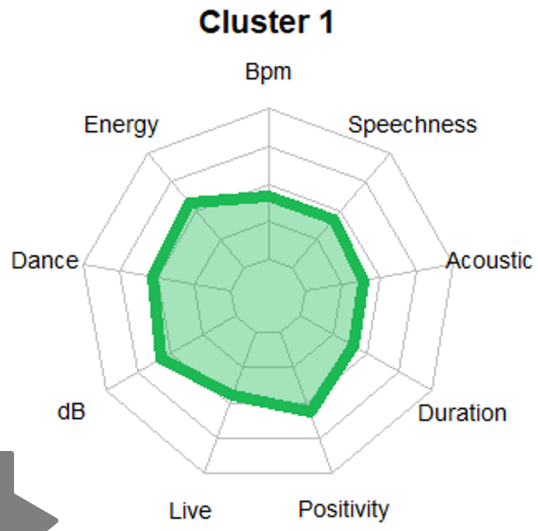
Appendix 2: Dendrogram 4-Clusters

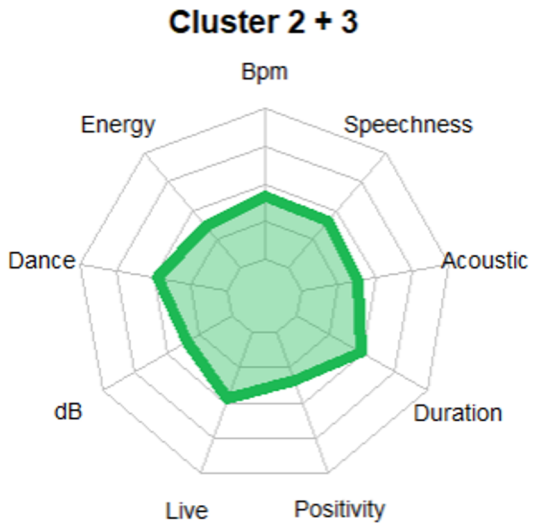
Higher dissimilarity value for clusters compared to 6 clusters

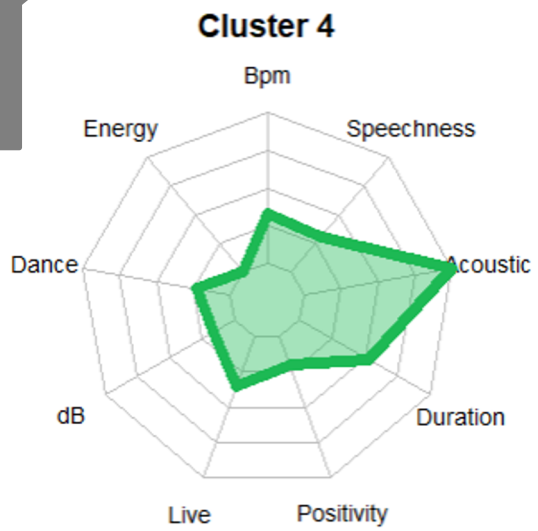


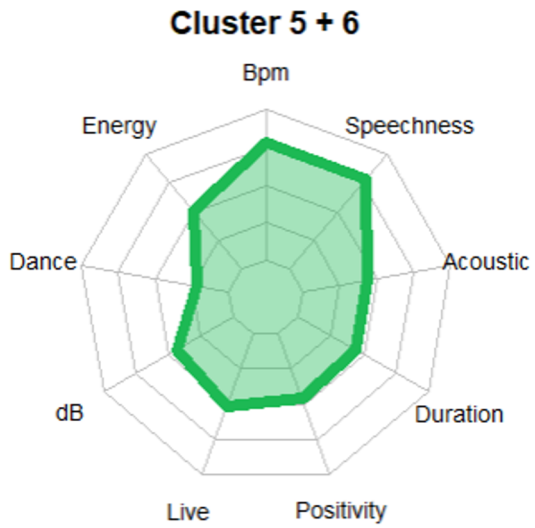
Appendix 3: 4 Cluster Alternatives

Clusters maintain distinguishable and interpretable characteristics

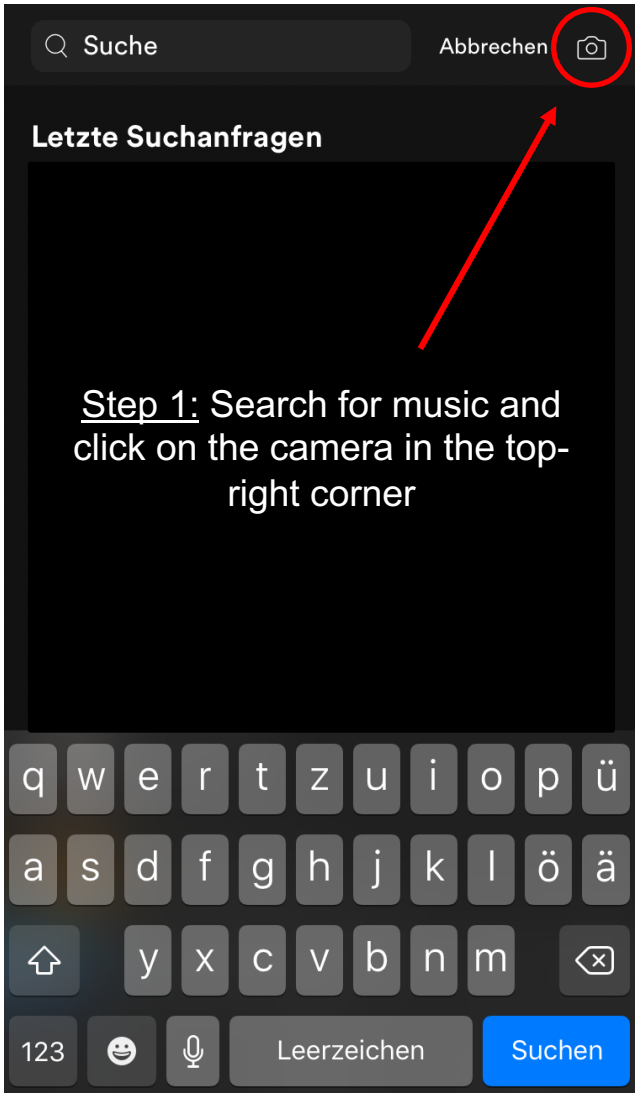


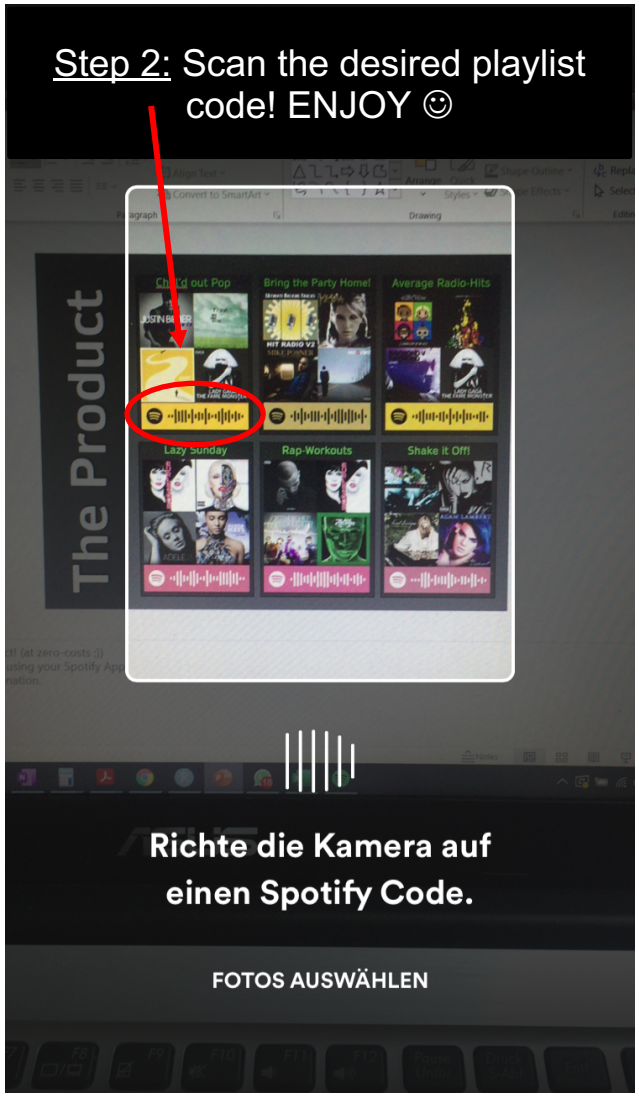




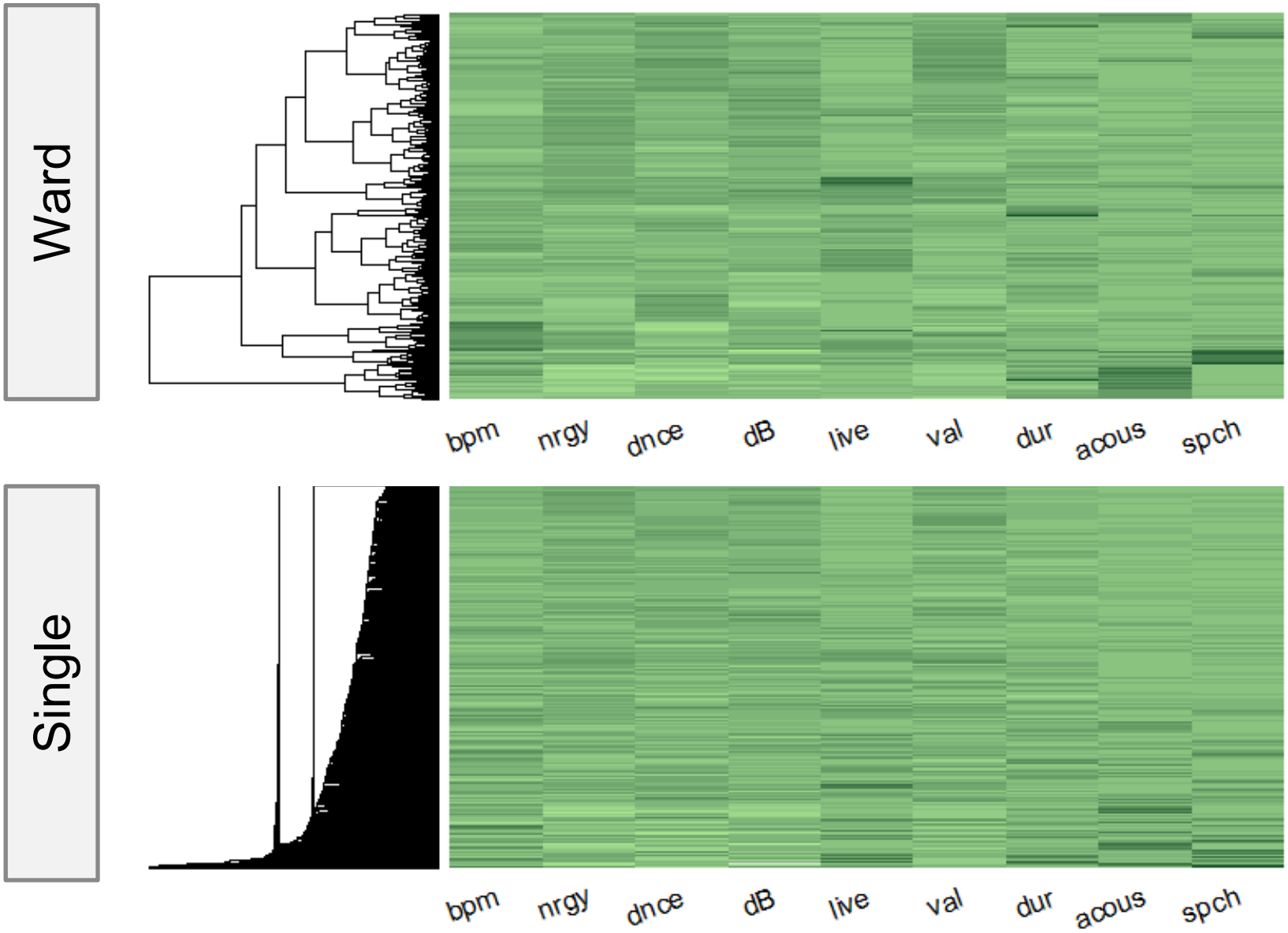


Appendix 4: Scan the Playlists

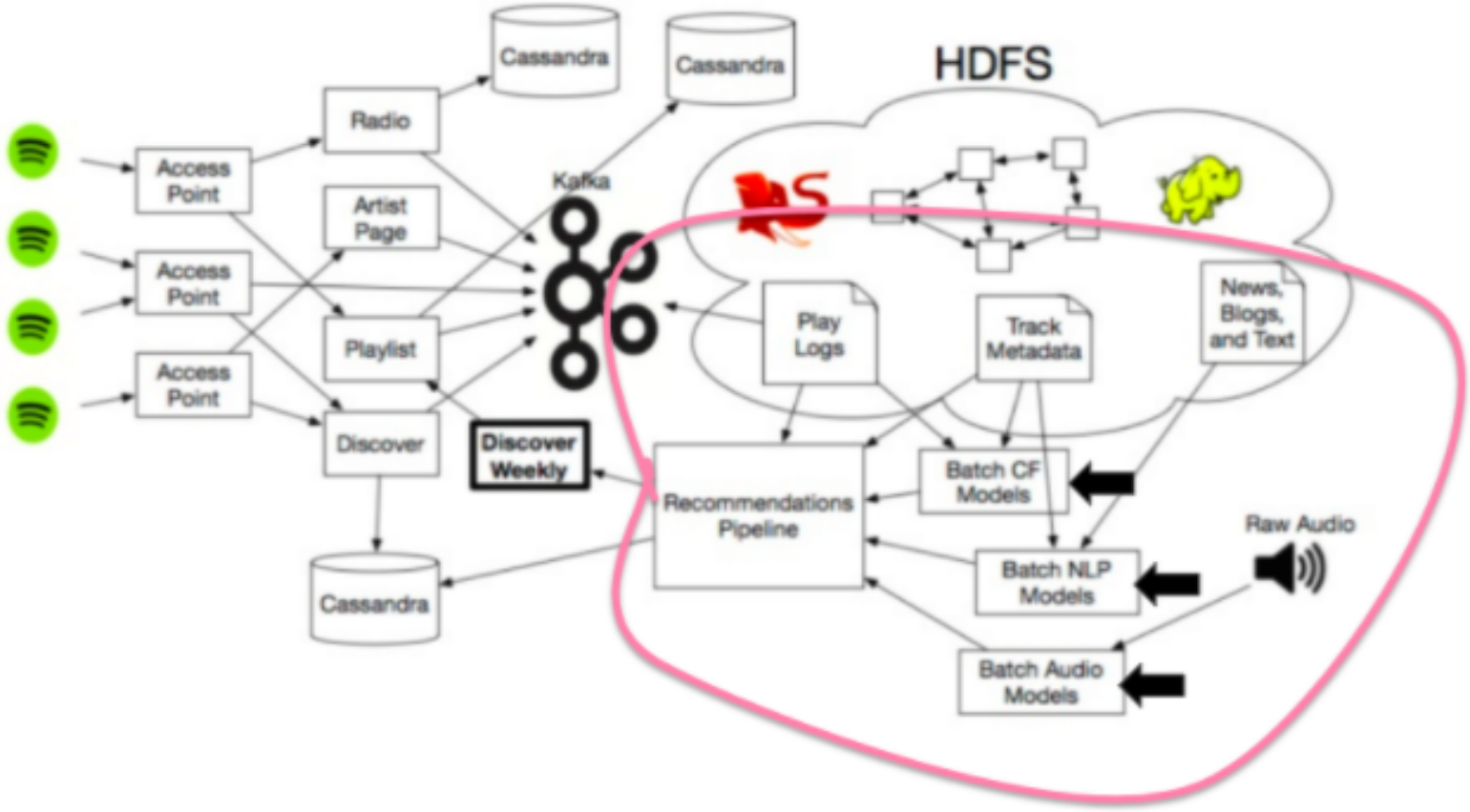




Appendix 5: Dissimilarity Matrices



Appendix 6: How Spotify actually creates its music playlists



Source: The Next Big What