UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Danilo Boechat Seufitelli

Understanding Musical Success Beyond Hit Songs: Characterization and Analyses of Musical Careers

Belo Horizonte
2023

Danilo Boechat Seufitelli

Understanding Musical Success Beyond Hit Songs: Characterization and Analyses of Musical Careers

Final Version

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Mirella Moura Moro

Belo Horizonte
2023

Seufitelli, Danilo Boechat.

S496u Understanding musical success beyond hit songs: characterization and analyses of musical careers [recurso eletrônico] / Danilo Boechat Seufitelli - 2023.

1 recurso online (135 f. il., color.) : pdf.

Orientadora: Mirella Moura Moro

Tese(Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f. 123-135

1. Computação – Teses. 2. Computação – Cultura brasileira - Teses. 3. Computação – Música – Teses. 4. Sucesso profissional – Cantores brasileiros - Teses. 5. Indústria musical - Brasil – Teses. I. Moro, Mirella Moura. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III.Título.

CDU 519.6*73(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

UNDERSTANDING MUSICAL SUCCESS BEYOND HIT SONGS: CHARACTERIZATION AND ANALYSES OF MUSICAL CAREERS

DANILO BOECHAT SEUFITELLI

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores(a):

Profa. Mirella Moura Moro - Orientadora
Departamento de Ciência da Computação - UFMG

Profa. Renata de Matos Galante
Instituto de Informática - UFRGS

Prof. Flavio Vinicius Diniz de Figueiredo
Departamento de Ciência da Computação - UFMG

Prof. Flávio Luiz Schiavoni
Departamento de Ciência da Computação - UFSJ

Profa. Michele Amaral Brandão
Tecnológico em Processos Gerenciais - IFMG

Belo Horizonte, 25 de agosto de 2023.

Documento assinado eletronicamente por **Mirella Moura Moro**, **Professora do Magistério Superior**, em 12/09/2023, às 11:37, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Flavio Vinicius Diniz de Figueiredo**, **Professor do Magistério Superior**, em 13/09/2023, às 09:36, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Michele Amaral Brandao**, **Usuário Externo**, em 29/09/2023, às 16:43, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Renata de Matos Galante**, **Usuário Externo**, em 02/10/2023, às 15:24, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Flávio Luiz Schiavoni**, **Usuário Externo**, em 06/10/2023, às 16:22, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2570633** e o código CRC **A0925340**.

I dedicate this dissertation to those who faced the adversities of the Covid-19 pandemic and found a way to hold on through the power of music.

Acknowledgments

Finding the right words to express my gratitude becomes challenging as I complete this project. However, with deep appreciation and humility, I dedicate this space to express my sincere thanks to all the people who contributed directly or indirectly to the success of this long journey.

First, I would like to express my deep gratitude to my loving wife, Elaine, and my dear son, Antônio, for their infinite patience and understanding in dealing with my absence and high-stress moments. This path would have been much more difficult without your unconditional love and support. My thanks extend to a greater force that permeates the universe and illuminates my steps in all stages of this academic and personal journey. I thank God and Spirituality for guiding me, strengthening my faith, and providing wisdom to face the challenges that came my way.

To my family members, José Antônio, Cristiane, Bruno, Claudia, Micheli, Bruninha, and Mateus, thank you for always rooting for me and supporting me at every moment of this academic journey. In particular, I express my affectionate thanks to the memory of my dear grandmother Laudelina, whose love and wisdom continue to inspire me even after her departure. I'm sure she's vibrating for my conquest close to God.

I want to express my gratitude to those I met along my journey at the DCC, both in the CS-X laboratory and other spaces. Thanks to Guilherme Vezula, Luiza de Melo, Michele Brito, Michele Brandão, Natália Machado, Natércia Aguilar, Héctor Azpúrua, Jhielson Montino, and many others who contributed to an enriching learning environment and lasting friendships.

I give special thanks to Gabriel Oliveira and Mariana Oliveira, whose hands were always outstretched throughout my doctorate. Thank you for the friendship built, for all the short meetings of 2 or 4 hours (lol), and your patience in explaining simple concepts to me. Thanks also for all the fun times in our meetings (like that night that turned into old-fashioned funk), our face-to-face meetings, our online games, the daily gossip, the inside jokes, and our detective skills applied together. Our friendship was fundamental to overcoming the challenges and the moments of relaxation that made this journey even more special. Thank you for all the moments of joy, companionship and being true partners in all situations. I had promised that I would write a few pages of thanks to both of you, but I could have done better as it is tough for me to express your importance in my journey. *Ninguém solta a mão de ninguém.*

I cannot fail to mention my therapists, Lucimar, Ricardo, Luciana, and Robson,

whose support and guidance were essential for my personal and professional growth. Your contributions went beyond completing this work, allowing me to improve as a student, teacher, father, husband, and son.

Finally, my deepest gratitude goes to my esteemed advisor, Mirella Moro. In the face of her adversities, her unshakable belief in my potential, her tireless support, valuable teachings, and skillful guidance were fundamental to the success of this endeavor. Mirella is an inspiration as a professional and human being, an example to be followed in universities, always valuing and respecting the individual particularities of her advisees.

I express my deepest gratitude to all those who contributed, directly or indirectly, to this achievement. Every gesture, word of encouragement, and support was invaluable for realizing this academic dream. May this trajectory of recognition and collaboration continue to illuminate our future paths. Thank you all very much!

Finally, I would like to express my gratitude for the excellent support from UFMG and the staff of DCC (especially for Sônia) for being constantly available to solve questions. Thanks are also due to the financial support of CAPES in the form of a scholarship.

"I'm a totally different writer than I was fifteen years ago when I would be inspired by the most lame ideas. When you start off you're just connecting words, but now the song has to have an honesty to it. There are three subjects to write about: life, love, and death. The secret is writing the same old stuff and staying inspired by the simple things. That's probably the secret to life itself – finding happiness in everyday miracles."

(Graig Wiseman)

Resumo

Carreiras musicais são dinâmicas e fundamentais para a expressão artística e cultural. O seu dinamismo pode ser observado devido às diversas mudanças que ocorreram nas últimas décadas no que se refere ao consumo musical: passamos do vinil, fitas cassetes e CDs para as plataformas de streaming que, aparentemente, ficarão presentes em nossas vidas por muito tempo. O streaming trouxe consigo a alta disponibilidade de dados associados ao consumo musical e à preferência de ouvintes. Com tais dados, é possível extrair *insigths* relevantes sobre o que pode levar algumas músicas ao sucesso e outras não. Nesse cenário, surgiu uma importante área de estudo chamada *Hit Song Science*, cujo objetivo principal é desvelar a dinâmica do sucesso na indústria da música.

Colecionar músicas de sucesso pode levar artistas a experimentarem períodos de sucesso muito superior ao "comum", e tais períodos são conhecidos como *Hot Streaks*. Nesse sentido, compreender os fatores de como os diferentes perfis de artistas se destacam e alcançam seus períodos de maior sucesso pode ser crucial para a indústria da música. Dessa forma, o objetivo desta tese é identificar as características que levam os artistas a alcançarem o seus períodos de maior sucesso (*Hot Streaks*). Para isso, inicialmente, foi realizada uma profunda revisão de literatura para identificar as principais definições de sucesso, features e algoritmos utilizados, o que resultou na proposição de uma taxonomia e de um fluxo genérico para *Hit Songs Science*.

Em seguida, foi investigado como se deu a evolução do consumo musical no mercado brasileiro, analisando o período de transição da era física para a digital. Em geral, foi possível identificar que os períodos de maior sucesso dos artistas tendem a se agrupar no tempo, e detectar os períodos de *Hot Streaks*, um período contínuo de alto impacto dos artistas. Além disso, foi detectado que alguns gêneros musicais têm padrões específicos significativos para ambas as eras. Também foi realizada uma análise de perfil que revelou três clusters diferentes em ambas as épocas: Spike Hit Artists (SHA), Big Hit Artists (BHA) e Top Hit Artists (THA), que atuaram como descritores de classe de artistas de sucesso. Por fim, os estudos revelaram que os brasileiros preferem consumir músicas de artistas brasileiros, independente da era.

Por fim, foi investigado uma possível regularidade na exploração de diferentes tópicos nas carreiras artísticas antes deles experimentarem seu primeiro período de sucesso acima do normal. Para isso, foi proposto uma metodologia baseada em dados para analisar como os artistas espalham seus interesses (*Exploration*) e concentram sua atenção (*Exploitation*) em tópicos musicais distintos (ex. gêneros musicais) enquanto obtêm picos

de sucesso (*Hot Streaks*). Desta forma, foi medida a entropia das carreiras de artistas, que resulta nos graus de *Exploration* e *Exploitation*. A fase de *Exploration* indica que artistas tendem a diversificar seus tópicos de trabalho; enquanto na fase de *Exploitation* há uma certa definição do foco de trabalho, refinando suas capacidades ao longo do tempo. Os tópicos musicais são identificados por meio da detecção da comunidade em uma rede complexa de modelagem de artistas, músicas e seus gêneros. Em seguida, foi medido o grau de entropia (em termos de *Exploration* e *Exploitation*), quantificando-a em três períodos: antes, durante e depois de seus *Hot Streaks*. Os resultados mostram que os artistas exploram mais tópicos antes de atingir sua primeira onda de sucesso e também durante seu período de maior sucesso. Depois, eles tendem a reduzir a gama de tópicos, aprofundando-se em apenas alguns. Tais descobertas são relevantes para identificar e nutrir talentos criativos na indústria da música.

Palavras-chave: musical success; hot streaks; hit song science; artist careers.

Abstract

Musical careers are dynamic, expressive, and fundamental to an artistic and cultural experience. Its dynamism can be seen through the many changes in recent decades concerning music consumption: we have moved from vinyl, cassettes, and CDs to the streaming platforms that are here to stay. Streaming brought with it the high availability of data associated with music consumption and listener preference. With such data, we can extract relevant knowledge, such as what can lead some songs to success and others not. In this scenario, a critical study area called Hit Song Science emerged, whose main objective is to reveal the music industry's success dynamics. Collecting hit songs can lead artists to experience periods of success far beyond the "ordinary" periods known as Hot Streaks. In this sense, understanding how the different profiles of artists stand out and reach their most successful periods can be crucial for the music industry, which deals with the constant natural evolution of the market and needs to reinvent itself to satisfy the desires of its consumers: connect successful music and artists.

Hence, our objective in this thesis is to identify the characteristics that lead artists to reach their most successful periods. We first conducted an extensive literature review to identify the main definitions of success, characteristics, and algorithms used. As a result, we propose a taxonomy and a generic flow for Hit Songs Science. Next, we study how music consumption evolved in the Brazilian market, analyzing the transition period from the physical to the digital era. We found that artists' most successful periods tend to cluster in time, and we identified periods of Hot Streaks. Furthermore, we detected that some musical genres have significant patterns for both eras. We also performed a profile analysis that revealed three different groups in both eras: Spike Hit Artists (SHA), Big Hit Artists (BHA), and Top Hit Artists (THA), which acted as class descriptors of successful artists. Finally, we discovered that part of the Brazilian population with access to music streams prefers to consume music by Brazilian artists, regardless of the era.

Finally, we investigate a possible regularity in exploring different topics in artistic careers before they experience their first period of above-normal success. For this, we propose a data-based methodology to analyze how artists spread their interests (*Exploration*) and focus their attention (*Exploitation*) on different musical topics (e.g., musical genres) while achieving peaks of success (*Hot Streaks*). Hence, we measure the entropy of artists' careers, which results in the degrees of *Exploration* and *Exploitation*. The *Exploration* phase indicates that artists diversify their work topics. In contrast, in the *Exploitation* phase, there is a specific definition of the work focuses, refining its capabilities over time.

We identify the musical topics by detecting the community in a complex modeling network of artists, songs, and their genres. Then, we measure the entropy degree (in terms of *Exploration* and *Exploitation*), quantifying it in three periods: before, during, and after your *Hot Streaks*. Results show that artists explore more topics before hitting their first wave of success and during their most successful period. Afterward, they narrow the range of topics, delving deeper into just a few. Such findings are relevant to identifying and nurturing creative talent in the music industry.

Keywords: musical success; hot streaks; hit song science; artist careers.

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

A&R Artists and Repertories

BHA Big Hit Artists

CEO Chief Executive Officer

HSP Hit Song Prediction

HSS Hit Song Science

HS Hot Streak

ML Machine Learning

MIR Music Information Retrieval

PAA Piecewise Aggregate Approximation

PMB Pró-Música Brasil

SHA Spike Hit Artists

THA Top Hit Artists

VMA Video Music Awards

Contents

Chapter 1

Introduction

"What drives a song to become a hit?" Every year, researchers propose new formulas, algorithms, and approaches focused on finding the best solution to the hit prediction problem. There are even books on songwriting success that investigate common characteristics of successful songs in terms of melody, lyrics, and structure [17, 69, 89], and others that focus on the music business context. Jason Blume brightly informs in his book entitled *6 Steps to Songwriting Success* [17]:

> *"I've analyzed hundreds of hit songs in a variety of musical styles to try to understand why these songs became so successful - and why others had not. I've examined structures, chord progressions, melodies, lyrics, rythms, rhymes, titles, concepts, and more. After years of study I've concluded that there are no rules in songwriting. But while there are no magic formulas that guarantee hits, there are tools, techniques and principles that can help us to express ourselves to communicate what we feel in such a way that our listeners feel it, too."*

Hit song prediction aims to assess diverse perspectives over a given song and informs whether it will be a hit by considering complex attributes (e.g., emotions and internet engagement). Still, how can an algorithm-based solution return an accurate answer? This question has lead researchers to claim: Hit Song is not yet Science [91], unpredictable factors affect the market [103], and more. Despite true, such arguments have not stopped companies and startups from creating and selling their solutions (e.g., Polyphonic HMI, MixCloud, and Hit Songs Deconstructed), nor have they stopped the growing field of Music Information Retrieval (MIR) research from exploring solutions for nearly two decades (see Chapter 2). In other words, it is about choosing a good contextualized problem, defining a good set of features to feed a specialized algorithm, and testing it without bias.

With so many artists (singers) emerging each year, a fundamental question for Artists and Repertories (A&R) and music producers is: how to differentiate and identify a successful career? Such a question goes beyond music, and several works analyze successful careers based on genres, style of scientific publications, movie characteristics,

visual elements related to the artworks and even luck [48, 55, 72, 73, 106]. However, creative careers are among the most dynamic and complex, and the musical context is no different. After all, it depends on several internal and external factors to achieve success for artists and their productions.

Internally, artists need what some people call the "gift", while others already say that success comes from a lot of study and dedication. Some defend that success is a little bit of all that with a touch of luck. In other words, several internal factors may influence the artist deliver a product that has quality and appeal to the masses. Externally, artists deal with the audience of TV, radio programs, podcasts, YouTube (where their productions can be known), an association of their products to specific brands, sponsorship, events, etc. Furthermore, successful music careers demand that artistic productions be approved by the public who consumes them.

In addition to all these external factors, today's artists constantly have to reinvent themselves due to Internet and Web advances. Passman [92] expressed that for the first time in the music industry, they no longer monetize music by selling something. They monetize by the number of times listeners stream a song. Such transformation is the inflection point that completely changes the ecosystem of the phonographic business. From the advent of file-sharing technology in the late 1990s to the creation of the iPod, the music industry was on the vertex of a significant transformation - and with the newest shift to streaming music added to the social media content, such a shift has finally happened.

With content sharing on online platforms (social networks, audio/video streaming, news portals, etc.), many artists know how to use everything the internet provides to increase their fans' engagement. Synonymous with online success is gaining new followers on social networks, understanding hashtags, and discovering tools to simplify the artist's work to show their creativity in the virtual world. Artists who are very successful on social networks and accumulate millions of fans on their official profiles deal with their digital strategies: investing in trending memes, selfies, concert records, and much more. In this sense, maybe some artists are being forgotten in the mainstream media for lack of interaction with this audience.

Certainly, answering questions about success is not simple. Consider a small fraction of it: what makes a person like (or even dislike) a new song? The answer involves a complex mix of many perspectives, ranging from psychological (emotions, taste in music, personal history, etc.), social influences (e.g., everyone is listening), and intrinsic music features (tempo, energy, valence, lyrics, etc.). Music professionals (especially songwriters) will undoubtedly add their perspective to this mix. They will include creativity, uniqueness, craftsmanship, texture, quality, a genius touch, and the overall feeling, ecstasy, or introspection provided by the music.

Specifically for musical artists, the problems are the same. On the one hand, some artists have bursts of success (e.g., Hot Streaks), but they almost (or entirely) fall into

oblivion in a short time. On the other hand, there are those artists whose growth is slow, but with good management of their careers, they become great perennial artists. In this sense, understanding how to build successful careers is part of the work of producers to manage the careers of singers. As mentioned earlier, we understand that these careers reflect countless factors. However, there are those factors that we can observe certain regularity at different levels in different artist profiles. In this sense, Computer Science can help building profiles of successful artists by using computational resources.

1.1 Success Measurement

Before going any further, we must define what success is. Defining success is a difficult task not only for the music industry but also for professional careers [35], public libraries [119], and even political influence [46]. In the music scenario, we find exciting research about the musical genre Jazz linked to success, putting gender and race issues into perspective [45], and even a very relevant discussion about what it is and how to become a successful musician [100]. Discussing success can disregard essential aspects of its definition due to limited knowledge, data, or way of measuring specific attributes, etc. Nevertheless, Spotify, one of the leading music streaming platforms, has redefined success in the music industry. Traditionally, music success was measured by album sales, concert attendance, and radio appearances. Now, in the digital streaming age, success may be determined by the number of plays, followers, and engagement on platforms like Spotify. Additionally, Spotify's algorithm-driven playlists and recommendations have significantly influenced an artist's success on the platform. Being featured on popular playlists like "Today's Top Hits" or "Top Brazil" can increase substantially streams and exposure for an artist, boosting their success on the platform.

Still, it is essential to delimit what success is in data-oriented research. In this thesis, we consider a song a success when it appears on the Spotify charts. The same perspective is valid for artists. Artists who have hits on the charts (a consequence of digital music consumption) can be considered successful artists. Note that one song may be more successful than another, depending on its position. After all, a song reaching number one on the charts can be considered more successful than one that goes to number 200. However, the discussion is more complex and cannot stop there. For example, for a newly released artist, being present at *any* position on the chart can already be considered a substantial success. For example, a similar case happened with the song *Acorda Pedrinho* by the band Jovem Dionísio. A small excerpt from this song went viral on the TikTok social network. As a result, the band had its music at position 98 on Spotify's Top World

in 2022.[1] At this point, we must also define that our scope of successful songs and artists is limited to the presence on Spotify. Spotify is a reliable data source and the one we use further on. After all, countless artists are considered successful in their communities but use other platforms to promote their work (live performance, TV, radio, other streaming media, etc.). Indeed, we chose Spotify because they are the world's most popular audio streaming subscription service. According to Spotify[2], their community is of more than 515 million users, including 210 million Spotify Premium subscribers, across 184 markets.

1.2 Motivation

The entertainment industry is one of the largest and most dynamic in the world. The production of films, music, and other multimedia content follows the constant evolution of the technologies available in our society. Starting from the golden age of radio and going through the popularization of cinema and television, today we have, due to the Internet and the *streaming* era, an infinitude of titles and content for the most varied tastes available with just a few clicks. According to Forbes,[3] between 2000 and 2015, the revenue from music in the United States fell by more than half. This fact is also reflected in other contexts, such as the distribution of films and series, which had a significant drop in revenue from traditional media (cable TV channels and physical copies). Although traditional media is declining, the growth in media consumption through digital platforms (like YouTube, Spotify, and Netflix) more than made up for the drop in physical media.

The popularization of such platforms transformed the way in which society consumes multimedia content. In the music scenario, a few years ago it was essential that songs were played on the radio, and their video clips were shown on television programs to become successful; now, we easily reproduce any song and video clips at any time. Digital platforms combined with social networks have become the most prominent form of disseminating the work of artists, bringing benefits such as universal access to these contents and the elimination of geographical barriers. In this context, there are scientific studies to predict the success of a given content before its release [18, 34, 64]. This area of study became known as Hit Song Science (HSS) in the field of music. Although we found authors who previously defended the thesis that this area is not yet a science [91], the

[1] Em maio de 2022, 'ACORDA PEDRINHO,' de Jovem Dionísio, tornou-se hit viral no TikTok: https://bit.ly/jovem-dionisio-tiktok

[2] About Spotify: https://investors.spotify.com/home/default.aspx

[3] FORBES. Digital Video And Social Media Will Drive Entertainment Industry Growth In 2019: http://bit.ly/34s6my7

same authors changed their opinion [90]. Also, more and more researchers seek to study the impact of music and artist characteristics on their success [6, 66].

Moreover, one of the main devices used by artists and record labels for disseminating music and conquer new listeners is collaborating with other artists. For example, several artists in the *pop* genre collaborate with singers from *rap* or *hip-hop*, as in the song *Crazy in Love* by Beyoncé and Jay-Z, which reached the top of the American chart in 2003; as well as other mix of genres, as the song *Despacito* by Puerto Rican Latin singer Luis Fonsi featuring Puerto Rican rapper and singer Daddy Yankee, which was released in January 2017 and launched to worldwide phenomenum through a remix version with the participation of Canadian Pop sensation Justin Bieber in April 2017. One advantage of this type of collaboration is the expansion of the target audience of a song, since it will be heard among listeners of both musical genres; therefore, it will be more easily propelled on platforms and in *rankings* of songs. In this context, Silva et al. [112] analyze the importance of social factors in collaborations for musical success and show successful artists are more likely to have profiles with a high degree of interaction and diversity.

Creative careers have always instigated researchers and industry to understand how to build or improve them. Recent studies use varied machine learning techniques for pre-analyzing to propose techniques, algorithms, approaches, and studies to include or exclude features that enhance the prediction task [68, 84]. Consequently, the music industry benefits from the researchers' proactivity, as they always propose updates regarding the phonograph market nuances to computational algorithms. However, we note there is a lack of exploring the common characteristics that lead to building successful musical careers, and associating such attributes with the most relevant periods in the artists' careers. After all, a significant milestone in artistic careers is the existence of Hot Streaks, that is, those periods of high notoriety and presence in the music market. Although Hot Streaks are practically a rule in musical careers, it is not clear if there are patterns regarding their beginning. The lack of systematic explanations for Hot Streaks and the randomness of when they occur within music careers support an unpredictable view of such careers.

This behavior leads us to question whether the above-normal success links to two concepts: Exploration and Exploitation. These concepts are widely discussed in artificial intelligence and are known to be the pillars of solving search problems [28]. In summary, *Exploration* refers to the process of visiting new regions of the search space (e.g., unexplored genres and musical styles), whereas *Exploitation* is delving into a given area in search of a local optimum (e.g., focus on specific genre and musical styles). Another important concept is *entropy*, which measures the diversity or variability within a dataset and helps to evaluate the balance between Exploration and Exploitation. Entropy provides a nuanced gauge to assess the harmony between delving into the uncharted territories (e.g., unexplored genres) of creative expression and harnessing the strengths of proven artistic directions (e.g., genres already tasted).

Recently, Liu et al. [72] explore the characteristics that influence to the emergence of hot streaks in three different creative careers: scientific, cultural, and artistic. They find Hot Streaks are not associated with either exploring or exploiting behavior in isolation. Furthermore, they show real careers are complex, with heterogeneous influences operating across domains and a multitude of individual and institutional factors. In this context, we understand musical careers are very similar to such creative careers and also lack studies to characterize better what precedes the periods of Hot Streaks. Therefore, our goal is to understand how creative careers achieve success, and then discuss the success profile of artists. Such a profile complements the results of machine learning tasks that create a model for the prediction of success and generation of such content, considering the technical and social characteristics. Specifically, our goal is to build the artist profiles of creative careers in the musical careers domain.

1.3 Goals and Contributions

The music industry is a dynamic and complex environment where various intrinsic and extrinsic factors influence success. However, traditional approaches to understanding success in music (such as feature engineering and machine learning models) may not capture the nuanced nature of such an industry. Aiming to fill this research opportunity, this thesis takes a data-driven approach to understanding the mechanisms behind an artist's success. By analyzing the evolution of musical genres and the topics of an artist's career, we aim to identify regularities and patterns that lead to successful periods. This approach sheds light on the creative process in music and has implications for identifying and nurturing creative talent in the industry.

This objective reflects three specific research questions that lead to achieving the overall goal of our work as follows.

1. ***Research Question 1 (RQ1):*** *How does current research deal with Hit Song Science?* Answering such a question requires a survey, in which we describe its main research problems, frequently used data sources, musical success definitions, features, and learning methods. From this review, we define the generic workflow on Hit Song Prediction. Furthermore, we propose novel taxonomies for (*i*) success measures, (*ii*) features, and (*iii*) learning methods used to consolidate the existing knowledge in HSS. Finally, we compare the reviewed articles on such characteristics, and our analyses reveal there is not one "feature" for an ideal hit song prediction model, as its performance depends on subjective decisions made in the analysis

process. Consequently, we need to further investigate how to comprehend the success beyond machine learning techniques.

2. ***Research Question 2 (RQ2):*** *How does the evolution of music consumption in Brazil affect the occurrence of hot streaks?* The music industry has recently reinvented itself, where vinyl and compact discs have given way to streaming platforms. This digital evolution has created vast volumes of data on music consumption and artists' careers. Thus, we raise questions about market standards in such different eras. First, we perform a clustered success analysis to answer the questions: are artists' most successful periods clustered in time? How to detect the artists' most successful periods (Hot Streaks)? After witnessing the Hot Streaks, we perform a cluster analysis to understand if it is possible to distinguish the artists by their level of success. Finally, we analyze the most popular genres in the Brazilian market in different eras. We also verify if the artists udergo their first period of Hot Streak at similar times.

3. ***Research Question 3 (RQ3):*** *Do artists explore different topics in their careers before reaching their first Hot Streak?* Hot Streaks mark creative careers, bursts of high-impact work grouped in close succession. However, this prompts us to investigate characteristics that may be important for artists to reach their very successful periods, such as exploration and exploitation [72]. Specifically, in music careers, we investigate whether artists explore different topics at random before experiencing their first hot streak (exploration) and whether they specialize in specific segments from that point onwards (exploitation). Hence, we propose a way to extract relevant topics in artists' careers. Next, we apply an entropy calculation to measure the variation of topics encountered over such careers. Finally, we correlate the timing of the hot tracks with the creative trajectories of the artists to verify if changes in the characteristics of the work result in the beginning of hot streaks.

1.4 Outline

The rest of this thesis is organized as follows. To answer Research Question 1, Chapter 2 provides a comprehensive study with a complete review of the main topics of the interdisciplinary field of Hit Song Science from a computer science perspective. We also define a generic workflow for HSS, introduce taxonomies for success measures and musical features, and categorize the main current learning algorithms. To explore the changes in the music industry in the last decades, we perform cross-era comparisons between Physical

and Digital media within the music market in Brazil in Chapter 3. Specifically, to answer Research Question 2, we build artists' success time series to detect and characterize hot streak periods, defined as high-impact bursts that occur in sequence, in Physical and Digital eras. Then, we identify groups of artists with distinct success levels by applying a cluster analysis based on hot streaks' features. To understand the behavior pattern of artists before they enter their most successful periods, in Chapter 4 we apply an entropy analysis methodology. The entropy level allows to understand the degree of exploration and specialization of topics experienced by artists in their careers to draw a profile of the artists. We provide a comparative analysis of their career' periods: before, during, and after artists reach Hot Streaks to answer Research Question 3. Finally, in Chapter 5, we present this thesis's central findings, which include a summary of contributions by each research question and a discussion of this research's limitations and potential threats to validity. Further, we highlight ideas for future improvements and propose a sequence for this research. Finally, we provide a list of publications during the doctoral period.

Chapter 2

Generic Workflow and Taxonomy for Hit Song Science

Music is not only one of the most important forms of cultural expression but also entails one of the most worldwide dynamic industries. Over the last few decades, the world has witnessed a drastic change in the way people consume music, moving from physical records to streaming services. There is a wide variety of streaming formats, including premium signature, ad-supported services (such as YouTube, Vevo, and free-account Spotify), and streaming radio services (such as Pandora and SiriusXM). Combined, streaming revenues accounted for 83% of total revenues in the US music industry by the end of 2021.[1]

Indeed, music streaming services have become the main source of revenue within the global music market, reaching a total of US$ 16.9 billion by the end of 2021.[2] Such growth is mostly driven by fans' engagement on paid streaming services for the seventh consecutive year. In 2021, the total accounts for such services rose to 523 million, with an associated revenue increase of 24.3%. As a result, the landscape of the music industry has become more complex, encouraging artists to reinvent strategies to maintain their presence in the market and expand it to reach new audiences.

As an example of reinvention, "Baby Shark", a famous nursery rhyme, appeared on charts for the first time in January 2019, after a video by South Korean educational brand Pinkfong sparked a viral dance challenge. Such YouTube video clip helped the song reach No. 32 on the Billboard Hot 100 for 20 weeks, with 395 million streamings in the first half of 2019. Furthermore, the first half of 2019 saw a new milestone by over 507 billion streamings, led by singles and albums by Ariana Grande, Billie Eilish, Halsey, Khalid, BTS, Lil Nas X, and Bad Bunny that cover a massive range of genres, moods and even languages.[3] In this context, the task of predicting whether a song will become a hit or even understanding its success remains a valid concern of the music industry, which constantly seeks to increase its revenue while dealing with different audiences.

Predicting and understanding whether a song will be a hit (or not) has significant importance for different groups. For the music industry CEOs, it may assist in maximizing

[1] 2021 Year-End Music Industry Revenue Report (RIAA): http://www.riaa.com/reports
[2] IFPI Global Music Report 2022: http://gmr.ifpi.org/
[3] Nielsen Music Mid-Year Report 2019: http://www.nielsen.com/us/en/insights/report/

future success by helping to choose in whom to invest. Furthermore, by investing correctly in potential artist/song and its distribution, music providers may not only increase sales of physical and digital albums but also improve on-demand audio streaming services. Artists may also profit by identifying the most suitable songs to lead an album to stardom. For the music consumers, it may help to decide if an album is worth buying because it may potentially contain three to five hits, instead of being an *one-hit-wonder* album. In addition, consumers play an essential role in the music market as they are the musical universe "feedbackers". In other words, musical success is measured based on consumer response and updated based on their evolving preferences. Essentially, this ability is the main drive behind the research field referred to as Hit Song Science, which Pachet [90] defines as "an emerging field of science that aims at predicting the success of songs before they are released on the market".

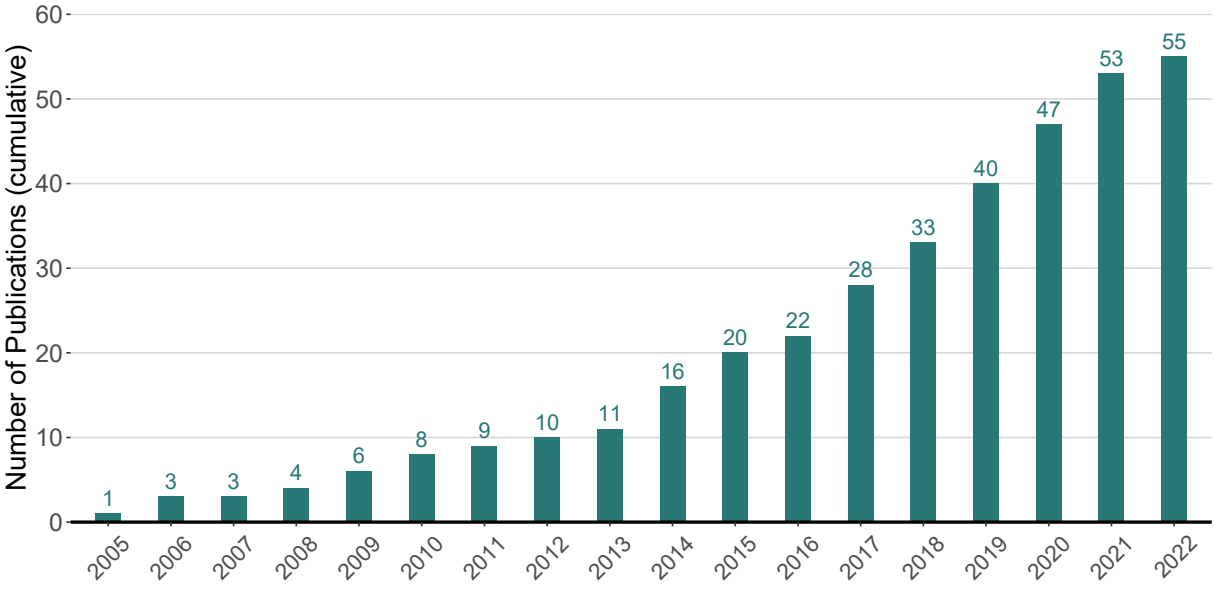
The study of HSS has two different perspectives: the prediction viewpoint and the scientific one. The former is justified by HSS involving different variable issues such as a complex combination of micro-emotions (also related to personal history), music specifics, and many other indescribable elements that escape our understanding, and how all of them may determine the success of a song. Conversely, HSS as a science is a bold view against the assumption that such variables hinder prediction, then it relies on features that would give a convenient result. The discussion of HSS being a wish or science raises many issues for the discipline. As pinpointed by Li et al. [71], HSS should also rely on "the psychology of music listening and the effects of repeated exposure, the paradoxical nature of the Western media broadcasting system, radios in particular, and the social influence human beings exert and receive from each other".

Although true, such arguments have not stopped companies and startups from creating and selling their solutions (e.g., Polyphonic HMI, MixCloud, and Hit Songs Deconstructed), nor have they stopped the research efforts from the scientific perspective. In particular, HSS is considered a specific task within MIR, a more comprehensive field that aims to extract relevant information from music content. As MIR is a well-established research field, there are survey articles that cover its main tasks, features, and applications [70, 76, 85, 134]. Regarding MIR tasks, some are more studied than others, and thus there are surveys exclusively on such topics. For example, Sturm [117] and Corrêa and Rodrigues [25] focus on music genre recognition and classification, whose goal is to determine the genre of a song from a set of musical features.

Overall, HSS is a trending research topic in academia. An evidence for such a claim is the increasing volume in publications about it, as illustrated in Figure 2.1. The publications considered in such a figure (and the remaining of the chapter) were obtained mainly from DBLP[4] until September 30, 2022, by using the following keywords: *hit song science*, *hit song prediction*, *musical success*, and *music popularity*.

[4]DBLP Digital Library: http://dblp.uni-trier.de/

Figure 2.1: Hit Song Science publications (cumulative), 2005 – 2022.

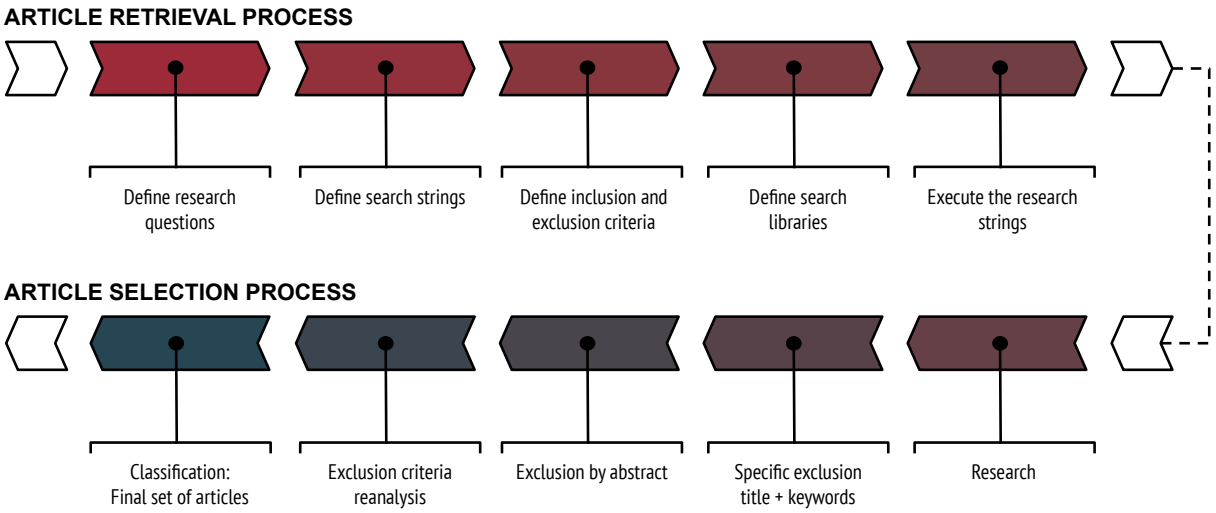


Source: The Author.

Besides such an increasing interest, to the best of our knowledge, this is the first survey chapter exclusively focused on HSS, its methods, features, and algorithms. Moreover, our goal is twofold: to answer initial questions over the field and to look for open issues within it. Our questions include: *what are the main research problems of HSS? How is musical success defined? What are the most used learning techniques for solving HSS problems? What are the challenges and research opportunities for this field?* We answer such questions by reviewing the literature on HSS to describe its main research problems, frequently used data sources, musical success definitions, features, and learning methods. From this review, we define a generic workflow on Hit Song Prediction (HSP), and novel taxonomies for (*i*) success measures, (*ii*) features, and (*iii*) learning methods used in this field to consolidate the existing knowledge in HSS and guide future research on this topic. Finally, we compare the reviewed articles on such characteristics, and our analyses reveal that there is not a "feat." for an ideal HSP model, as its performance depends on subjective decisions made in the analysis process. Hence, this survey works as a bridge between computer science and music-related fields, motivating upcoming research on Hit Song Science.

The remainder of this chapter is organized as follows. Section 2.1 presents the methodology used to retrieve this chapter's articles. Section 2.2 presents an overview of the HSS field and defines the HSP problem. The most common data sources in HSS are categorized and detailed in Section 2.3. Next, we propose and describe novel taxonomies for the most used success measures, musical features, and learning methods in Sections 2.4-2.6. In Section 2.7, we point out important research directions to guide future work on HSS. Finally, we present a general discussion and our remarks in Section 2.8.

Figure 2.2: Summary of the main steps for the literature review methodology.



Source: The Author.

2.1 Survey Methodology

In this chapter, we use a methodology comprising seven steps adapted from Kitchenham and Charters' protocol [62] to systematically search for relevant research in the Hit Song Science area. This methodology enables us to identify significant studies and better understand the current state of research in this area. The seven-step method provides a clear and structured approach to the search process, minimizing the potential for bias and ensuring the reliability of the search results. Figure 2.2 provides a visual summary of the entire process, highlighting each step of the searching process. Next, we detail each methodology step.

Step 1: Define research questions. The first step is to define research questions to guide the investigation over the state of Hit Song Science. Our questions and their goals are the following:

> ***RQ1 - What are the main research problems of HSS?***
> – Define interests and trends over time.
> ***RQ2 - How is musical success defined?***
> – Synthesize and analyze the definitions of musical success presented in the literature.
> ***RQ3 - What are the most used learning techniques for solving HSS problems?***
> – Identify proposed methods, models, and tools.
> ***RQ4 - What are the challenges and research opportunities for HSS field?***
> – Define the most addressed sub-areas, themes and trends.

Step 2: Define search strings. Solving such questions requires searching for publications that could answer them. First, we consider the most extensive Computing digital

Table 2.1: Search strings over digital libraries.

Search Strings
1."hit song" `OR` "hit song science" `OR` "hit song prediction" `OR` "hit song success"
2."hit song science" `AND` "musical success" `OR` "song success"
3. hit song (science `OR` prediction) `AND` (music *OR* musical) success
4.(music *OR* musical) success `OR` music popularity

Source: The Author.

library – DBLP,[5] and search for the term "hit song science" within title, abstract and keywords. Its results serve as input for filtering the most relevant keywords and define the final search strings, as Table 2.1 presents.

Step 3: Define inclusion criteria and general exclusion criteria. Keeping focus on our questions, we define the following criteria: *Inclusion criteria* verify if the publication is related to musical success; and *General exclusion criteria* check if it has no abstract, is only an abstract, is an old version of another study already considered, is not a primary study, and is not possible to access its full content.

Step 4: Search for publications. Searching for publications in only one digital library may compromise the research coverage. Therefore, we search for the pre-defined strings over: IEEE Xplore,[6] Scopus,[7] Science Direct,[8] and Web of Science.[9] All returned publications by the digital libraries were collected – except Scopus, as we considered only publications on Computing and Engineering.

Step 5: Define specific exclusion criteria. Based on the titles, we create specific exclusion criteria: publications outside *Computing* and *Engineering* and those dealing with other areas (e.g., biology and works such as *musical live performance*). Another exclusion criterion is papers that do not allow to predict musical success. We also exclude papers without success definitions.

Step 6: Select publications and identify common themes. By reading the abstract of the resulting publications, we discarded articles out of the inclusion criteria. Then, we reapplied the exclusion criteria to refine the results. In this stage, we promoted a group discussion to decide when an article leads to uncertainty.

Step 7: Classify publications. We elaborate a novel taxonomy by using the identified themes in the previous step, with three major classes: Success Perspective (Top-Charts, Economy, and Engagement), Musical Features (Intrinsic and Extrinsic), and Learning

[5]Digital Bibliography and Library Project: https://dblp.uni-trier.de/
[6]IEEE Xplore: https://ieeexplore.ieee.org/Xplore/home.jsp
[7]Scopus: https://www.scopus.com/
[8]Science Direct: https://www.sciencedirect.com/
[9]Web of Science: https://www.webofscience.com/wos/woscc/basic-search

Methods (Classification, Regression and Other). In this step, three volunteers manually labeled all publications (selected in step 6) by considering these three classes and their subclasses.

2.2 Hit Song Science

HSS is a multidisciplinary field where computer science meets conventional music-related topics, such as music theory, sociology of music, and culture markets. It involves the acquisition and analysis of music data with different modalities from distinct data sources. Such analysis resorts to Information Retrieval, Machine Learning, Data Mining and other advanced technologies to connect music and science. In this process, the main objective is to detect, or rather predict, whether a given song will become a chart-topping hit. Mainly, it boils down to distinguishing hits from non-hits, where a hit is usually a song featured at the top of music charts.

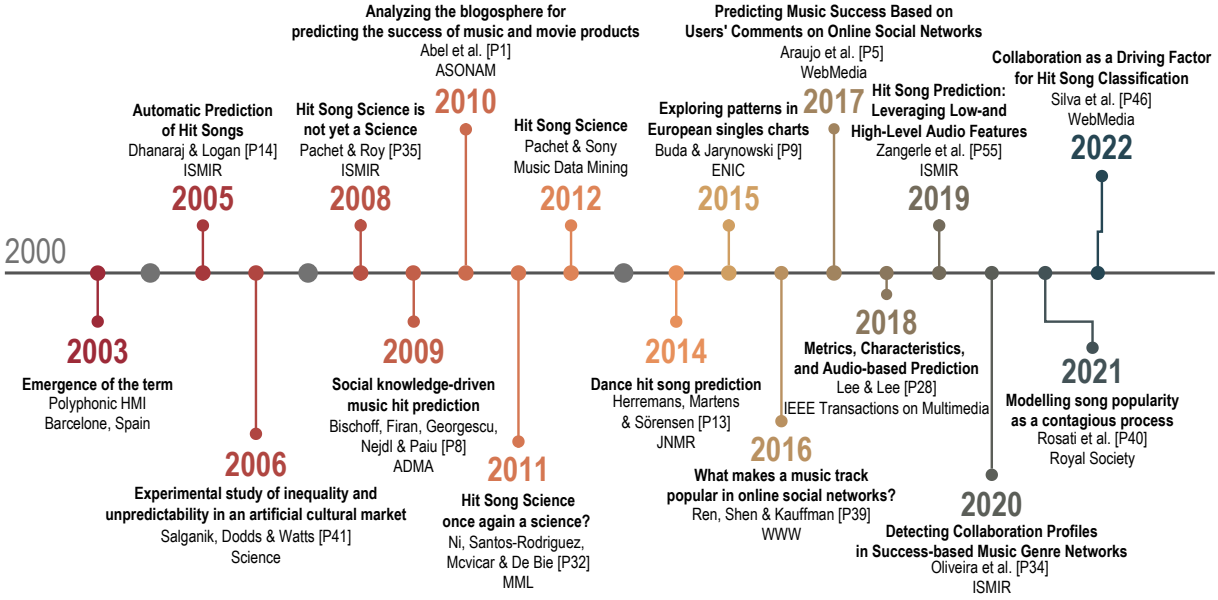
HSS emerges as a field of predictive studies to better understand the relation between the intrinsic characteristics of songs and their popularity. The premise of HSS is that popular songs are similar to the set of features that make them appealing to most people. Such attributes could then be explored through Machine Learning techniques to predict whether a song will top or appear in the popularity charts. Predicting the popularity of musical tracks provides huge benefits for all parties involved in the global music industry, as it allows improving revenues by focusing on potential hits. Moreover, predicting hits from social music media (e.g., Spotify and Deezer) can be helpful for improving revenues even further through advertising and publicity.

Despite being a recurring topic within music information retrieval (MIR), the concept of HSS was first introduced by the Polyphonic HMI,[10] in 2003. This artificial intelligence company developed a machine learning software, called Hit Song Science, which uses mathematical algorithms and statistical techniques to predict the success of a song in the current market. According to Polyphonic, its software was able to anticipate the success of artists such as Norah Jones, Jennifer Lopez, and Robbie Williams. Such a revolutionary tool allowed scientists and music enthusiasts to break down millions of past hit songs into their mathematical features. Despite such formerly unprecedented effort, it took years after the emergence of this tool for the issue to gain significant attention as a research direction, as informed in Figure 2.3.

Early works in this area focus mainly on song-related features and introduce a series of different types of approaches, including boosting classifiers [33], experimen-

[10]Polyphonic HMI, (January, 2020), http://bit.ly/polyphonic-hmi

Figure 2.3: Hit Song Science research timeline in a nutshell



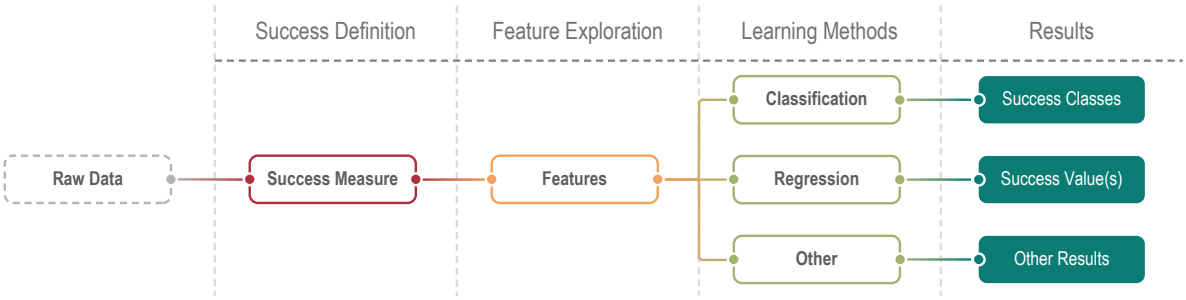
Source: The Author.

tal studies [103], adaptive algorithms [24] and Support Vector Machines [33, 91]. As more techniques for learning from data became available and widely used, recently proposed approaches are more advanced and elaborated. Specifically, more powerful methods have been applied not only to analyze bigger and more complex data but also to learn additional insights [50, 79, 129, 135]. Moreover, with the popularization of social networks, information about music consumers' tastes becomes available and easy to explore [1, 5, 14, 60, 63, 110, 131]. As a result, many studies have also relied on social information as hit song predictors [54, 99, 111, 112].

Despite the varied alternatives, most of the existing studies tackle HSS as a *Hit Song Prediction* problem, in which classification and regression are both chosen as the most common tasks. Such tasks fall under the umbrella of supervised machine learning, which is designed to learn based on a labeled dataset. However, other types of learning have also been used to predict musical success, as well as to find patterns and potential factors that influence songs' popularity. In general, such approaches include clustering algorithms [112], statistical analysis [5, 108] and prediction [32, 111], and social networks analysis [20, 112].

Overall, most of the machine learning solutions follow a predefined pipeline. Therefore, we propose a generic workflow for the *Hit Song Prediction* problem as one contribution of this survey, as shown in Figure 2.4. From one or more selected data sources, defining a proper success measure to evaluate the prediction model is required (Modeling). Moreover, such data sources guide which features will be considered as input (Feature Exploration) of the chosen Learning Methods. Such a workflow is reflected in the structure of this chapter, where each step in the prediction process is covered by a specific section.

Figure 2.4: Generic workflow for the Hit Song Prediction problem.



Source: The Author.

2.3 Music Data Acquisition

Following the premise of Hit Song Science (HSS), the first step of most approaches is to gather data regarding both song characteristics and success. However, these concepts can be seen by many facets, as the definitions are open to different visions. For instance, data about a given song can be acoustic and/or lyric-based, while its popularity may be measured considering its position within a chart or its sales revenue. Besides, information concerning consumers' behavior may be aggregated to HSS analyses to enhance the results. Therefore, using data from multiple sources is necessary and useful to build better models for analyzing and predicting musical success. In this section, we describe and classify the main and most commonly used data sources in four categories according to their purpose: popularity; acoustic characteristics; lyrics; and social behavior.

2.3.1 Common Data Sources

Regarding song popularity, research studies usually consider information such as position in charts to determine whether a song is a hit or not. The US-based magazine Billboard[11] is the most consolidated data source, providing many different types of rankings since the 1940s. The Hot 100 is the most commonly used, as it is a weekly list of the 100 most popular songs (regardless of music genre or style) in the US, considering data from radio airplay, sales, and streaming activity [9, 14, 60, 63, 67, 83, 87, 97, 111, 112, 121]. Billboard also aggregates the weekly rankings in a Year-End Hot 100 Chart, which is used in some studies within HSS [115, 116]. However, there are several studies considering other specific Billboard charts in their analyses. For example, Chon et al. [24] focus on one spe-

[11]Billboard Charts: http://www.billboard.com/charts

cific genre by using the Top Jazz Chart, based only on the albums' sales. Also, Lee and Lee [65] obtain data from The Rock Songs Chart, a weekly list of the 50 most popular rock songs. Such authors believe that this choice may produce cleaner results and better insights when focusing on specific genres.

As the world becomes more connected, and the globalization process reaches most of the countries, local engagement shapes the global music environment. In such a way, some studies consider charts from outside the US in their analyses and predictions. The United Kingdom is the second most considered market, having its charts published by the Official Charts Company[12] (OCC) [50, 54, 86]. Besides using British charts, Fan and Casey [37] also collect Chinese hit songs for comparison purposes. Moreover, there are studies considering other European countries (e.g., France, Belgium and Germany) [20, 49] and Asian markets such as South Korea [108] and Indonesia [38]. Other popularity approaches use YouTube views and likes [23] and sales data provided by platforms such as Amazon [1] and Nielsen SoundScan [9, 12, 32].

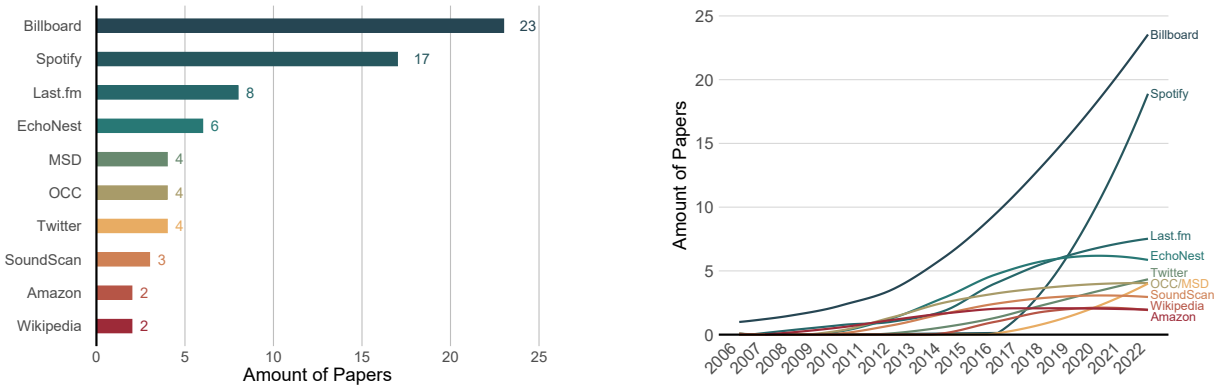
Now, changing the subject to features, acoustic characteristics of a song are important tools for describing its structure. Besides being better discussed in Section 2.5, it is important to note that they have been largely used since early HSS research studies, such as Dhanaraj and Logan [33], which use in-house databases as their data source. With the evolution of the Music Information Retrieval (MIR) field, new sources take place in HSS, as EchoNest API, with more than a trillion data points on over 34 million songs in its database [50]. Several studies use this API for extracting features such as tempo, time signature, song duration and loudness [9, 37, 49, 50, 86, 116]. Nonetheless, with the expansion of music streaming services and the acquisition of EchoNest by Spotify in 2014, its Developer API[13] is now the main source of acoustic features, thus being used by most recent studies [3, 5, 7, 8, 38, 79, 80, 83, 88, 97, 111, 112]. Nonetheless, there are still other sources used, such as the Million Song Dataset[14] (MSD) [96, 135] and AcousticBrainz[15] [54].

Data frequently used within HSS papers also include song lyrics, mainly for generating features related to rhyme and text. From the early years of HSS, there is no consensus on the best and most reliable source for song lyrics, and each study considers a different lyric source. For example, websites such as Astraweb Lyric Search [33], MetroLyrics [116], MusicSongLyrics [99] and LyricsMania [23] are used as lyrics sources by several authors. However, more recent studies, such as Martín-Gutiérrez et al. [79], use Genius, which has an exclusive API for developers to collect data in a simple and fast way, with no need to use web crawlers or HTML pages.

Finally, a different kind of data source has recently emerged: social media, which

[12]Official Charts Company: http://www.officialcharts.com/charts
[13]Spotify Developer API: http://developer.spotify.com/documentation/web-api
[14]Million Song Dataset: http://millionsongdataset.com/
[15]AcousticBrainz: http://acousticbrainz.org/

Figure 2.5: Most commonly used data sources in Hit Song Science (a) and their evolution over the years (b).



Source: The Author.

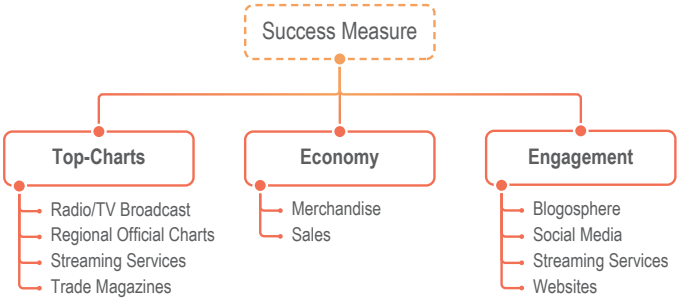
is changing the way people share their opinions and impact several areas, including the music industry. Therefore, the consumers' behavior plays a key role within musical success analysis, and online platforms such as Last.fm[16] are largely used to collect listener-based data and features [14, 32, 49, 98, 99, 110]. Moreover, blogging platforms are also an important sources of information about people's feelings on a given song, album or artist. For example, Abel et al. [1] use Spinn3r[17] to collect more than 100 million blog posts in the music domain. More recent studies collect data from social networks such as Twitter, Instagram and Facebook to analyze users' behavior related to a new musical release [5, 27, 60, 121].

2.3.2 Discussion

Songs are complex and dynamic objects that can be analyzed in different ways, and Hit Song Science (HSS) emerges as a field where studies try to use many of these facets in their models. Therefore, a convenient approach is to collect information about songs through multiple sources to complement audio-based musical success prediction. The most frequently used data sources in HSS are presented in Figure 2.5a. Note that chart organizations, such as Billboard (US) and Official Charts Company (UK), are the main sources of hit song lists, which are the basis of most studies. Only after this step authors collect song features (e.g., acoustic, lyrics, and social). Nonetheless, this choice of data source may create a regional bias, as local markets across the world behave in different ways by recognizing specific artists and music genres. As the world becomes

[16]Last.fm API: http://www.last.fm/api/
[17]Spinn3r: http://docs.spinn3r.com/

Figure 2.6: Proposed hierarchical taxonomy for music success measures from three perspectives.



Source: The Author.

more globalized, these markets constitute a driven-force within the music industry.

Furthermore, Figure 2.5b presents the evolution of data sources in HSS over the years. Over the last few decades, the world has seen a dramatic change in the way people consume music, moving from physical records to streaming services. This change is also reflected in the research within HSS, as Spotify has become the second most used data source in 2020. In addition to acoustic features, this platform offers relevant information, such as global and regional charts (due to its presence in more than 70 countries) and user behavior. Therefore, streaming platforms are becoming a powerful data source to research in HSS and MIR fields.

2.4 Measuring Success

Besides acquiring data (previous section), HSS research also defines a success metric that is fundamental for predicting success. People usually associate musical success with fame, richness and power. This might seem reasonable, but defining and measuring the success of a song can be an abstract process. Understanding music popularity remains a topic of great interest not only for music-related industries but also for researchers within the MIR community. Specifically, success measures are often summarized in music top charts, used to further understand both current and long-term rankings of a song. Moreover, the success of a song can be exploited as the target variable of prediction models.

There are different measures of musical success in the HSS literature, setting the criteria required for a song to be perceived as successful. Such definitions evolve over time and vary across cultures, making it difficult to propose a unified and objective definition. There is a clear need for a proper classification to unify such interpretations and enable a fair comparison among them. Therefore, we propose a hierarchical taxonomy in Figure 2.6 that characterizes success measures from different perspectives. The classification is

based on an overview of the existing literature, and includes three categories: Top-Charts (Section 4.1), Economy (Section 4.2) and Engagement (Section 4.3).

2.4.1 Top-Charts Perspective

Musical success has been measured by relying on top-charts provided by radio stations, trade magazines, regional markets, and streaming platforms. A top-chart is the numerical ranking of songs based mainly on retail sales (physical and digital), radio and television plays, and online streaming. Moreover, since the early 2000s, a whole range of tools have started to track the vast activity taking place online, from streaming and social network actions. As a result, new revenue sources have emerged —such as streaming platforms, digital downloads, and live shows online —updating the definition of a successful song. Still, music charts have become an increasingly reliable musical success metric, for both the industry and artists.

From a historical perspective, trade magazines have been effectively providing the main history of successful songs, such as the Billboard Magazine. Its first chart was published in 1936, *Hit Parade*, which was a term used at the time to rank popular songs based on data from manual reports filled out by radio stations and stores. A variety of song charts followed, which eventually were consolidated into the *Billboard Hot 100*[18] (best selling singles) and *Billboard 200*[19] (best selling albums). All of the Billboard charts currently combine record sales, radio airplay, digital downloads, and streaming activity. Therefore, there is an undeniable appeal for artists and record labels to be able to predict the path of their songs along with the Billboard charts —artists want to compose hit songs, and labels want to invest in more popular artists.

Even with the advent of industry-shattering changes in the world of music (namely, the introduction of digital distribution mechanisms), Billboard remains the most visible chart in the music industry. Indeed, most studies that define success from a Top-Charts perspective consider Billboard charts as a ground-truth (about 48%). Each chart summarizes popularity statistics that reflect records sales and airplay data of any given week, as well as stores song, artist, and album metadata. At the end, such charts are the most common and suitable solution adopted in the HSS literature to assess the quality of predictions.

Specifically, approaches modeling hit prediction as a classification task[20] usually

[18]Billboard Hot 100: http://www.billboard.com/charts/hot-100
[19]Billboard 200: http://www.billboard.com/charts/billboard-200
[20]Classify songs into non-hits and hits.

define a song as successful if it is featured in any weekly Billboard chart at least once [83, 115, 116]. However, in such cases, the definition of *non-hits* or *flops* is far more challenging. With no data available on less popular songs (i.e., official "flops charts"), there is no consensus on this concept. To address such a challenge, Singhi and Brown [115] and Silva et al. [113] consider *flops* as the non-top-charted songs by all singers who have hit songs on the Billboard Year-End Hot 100 chart. A similar solution is proposed by Singhi and Brown [116], who experimentally evaluate four different definitions of *flops*, ranging from a broad to a very narrow perspective. Alternatively, Middlebrook and Sheik [83] merge a Billboard dataset into a set of Spotify collected songs, thus setting the tracks that never appeared on the chart as non-hits.

Another common strategy in classification models is assuming that a track is popular if it exceeds a certain *popularity threshold* [79]. Most studies train the classifiers on several rank ranges [14, 27, 50, 60, 63, 96, 97, 121], considering hit songs as those tracks that have reached a pre-defined peak chart position. For non-hits, in general, they randomly select about the same number of hit songs from the set of music tracks with Billboard positions greater than the threshold established. For instance, when the rank 1–10 is the interval of a hit, non-hit songs are randomly selected from the 11–100 rank positions. The median value of the ranking has also been explored as a popularity boundary. Alternatively, Lee and Lee [65, 67] formulate the prediction problem as binary classification, i.e., high vs. low values of multiple popularity metrics extracted from Billboard charts, where the median value of each pattern set the boundary of the two classes (*hit* and *non-hit*).

Some researchers have also used the charts statistics as success measures, including the highest position on the chart in any week of a year (peak position) and the number of weeks the track has been or was on the chart (weeks) [9, 24]. Such an approach is more common when a regression (or ranking) task is considered, where the proposed model learns to predict numeric scores. While most previous work use the Billboard rank score as a continuous output variable [60, 63, 111, 135], Chon et al. [24] aimed to predict an album's future based on two concepts: *lifecycle* and *lifespan*. The *lifecycle* of an album is a trajectory of its weekly positions in a chart, while *lifespan* is defined to be how long the lifecycle of an album is, in terms of the number of weeks. Following a different path, Nunes and Ordanini [87] assess relative success by comparing Billboard's Hot 100 songs that reached #1 position.

Music charts based on physical sales, digital downloads and streaming activity are not exclusively present in the United States. Since 1952, the United Kingdom has released the UK Singles Chart listing the top-selling singles in the country by the New Musical Express magazine.[21] Nowadays, the chart is called Official Singles Chart,[22] which

[21] New Musical Express magazine: http://www.nme.com/
[22] Official Singles Chart: http://www.officialcharts.com/charts/singles-chart

is compiled by the Official Charts Company (OCC), listing the top-selling singles in the UK. As an alternative reliable source, the Official Singles Chart has also been explored from different perspectives, including considering different rank ranges [37, 50, 86] and simply defining songs' success as "making it" into the charts [54]. Finally, other sources of music charts featured in the literature review include Spotify Charts [8, 88], the Pandora effort [90, 91], MixRadio Charts [101], regional charts [20, 37, 49, 108, 130], radio broadcast [33], and streaming/website charts [3, 7, 98].

2.4.2 Economy Perspective

Economy indicators are also useful as conventional and quantitative measure of musical success. Such economy gauges can be quantified in terms of profits, revenues, or dividends. Fisher et al. [39] state that total revenues can be broken down into two main sources: *performance fees* and *recorded music*, whether in physical or digital format. However, along with the digital revolution, a variety of revenue earnings have emerged. The International Federation of the Phonographic Industry (IFPI),[23] a non-profit members' organization, provides an annual global recorded music report based on five segments: *physical*, *digital* (excluding streaming), *streaming*, *synchronization revenues* (revenue from the use of music in advertising, film, gaming, and TV) and *performance rights* (use of recorded music by broadcasters and public venues).

Other sources of information on recorded music sales include Amazon Sales Rank [1, 32], Nielsen SoundScan [12, 32], and Billboard units [5]. The Amazon Best Sellers Rank (BSR),[24] also known as the "Amazon Sales Rank", is a score that Amazon assigns to a specific product based on historical sales data. It has been a popular tool over the years, as Amazon is one of the largest online CD retailers. Abel et al. [1] measure the music sales performance by using such a rank, which aims at capturing the popularity of a product compared with others in its category.

Nielsen SoundScan[25] has also been an ideal source for album sales data as an information system that tracks sales of music and music video products throughout the United States and Canada. Nielsen's sales data is compiled weekly from over 39,000 retail outlets globally, and has been a trusted and vital resource for companies that want a full picture of music sales for more than two decades. Moreover, Nielsen's data serves as the primary sales source for the Billboard music charts, making it the largest source

[23] International Federation of the Phonographic Industry (IFPI): http://www.ifpi.org
[24] The Amazon Best Sellers Rank: http://www.amazon.com/Best-Sellers/zgbs
[25] Nielsen SoundScan: http://www.nielsen.com

of sales records in the music industry. In general, most studies use such data source as post factum popularity information over songs [12] and albums [32]. As an alternative, Billboard album-equivalent units can also be collected as a data source for an economy perspective. Specifically, the album-equivalent is a measurement unit that defines the consumption of music that equals the purchase of one album copy. This consumption includes streaming and song downloads plus traditional album sales. Araujo et al. [5] manually collected the weekly Billboard Top 10 record chart to gather sales units in the week of the album's release date.

2.4.3 Engagement Perspective

Music has always contained a social dimension, whether shaped through the engagement of artists or listeners. Such an aspect affects how people consume and engage with music and can impact the musical performance success. In particular, digital applications have become a powerful tool when discussing and measuring success by offering ways to share information about music —and to share the music itself. Different social media services (like Facebook, YouTube, Twitter) are designed to attract audiences and encourage them to discover new artists, share recommendations and consume music. Moreover, most music streaming platforms, such as Last.fm and Spotify, integrate music listening and social interactions into a single service. Therefore, considering social engagement metrics can be a valuable tool when measuring music success.

There are several different metrics in which social media and streaming platforms can assess success. Including views [23] and likes [23, 110] on social media; or digital downloads [103], ratings [96], number of streams [5, 38, 79, 80, 112], and play counts [32, 99, 110, 129, 131] on streaming/digital platforms. As depicted in Figure 2.5, the most commonly used digital application in HSS research is Spotify, followed by The Echo Nest platform. The former is the most popular global audio streaming subscription service today, which manages and shares over 50 million tracks.[26] The latter is a music intelligence and data platform for developers and media companies, acquired by Spotify in March 2014. As part of the Spotify platform, The Echo Nest is mainly used to give listeners the best possible personalized music listening experience. That is, it is the driving force behind the playlists professionally curated on Spotify.

In March 2016, The Echo Nest API was shut down, and developers were encouraged to move over to the Spotify API instead. The Echo Nest API provided a buzz-measuring score, called *hotttnesss*, derived from mentions on the web, mentions on music blogs,

[26]Spotify Company Info: http://newsroom.spotify.com/company-info/

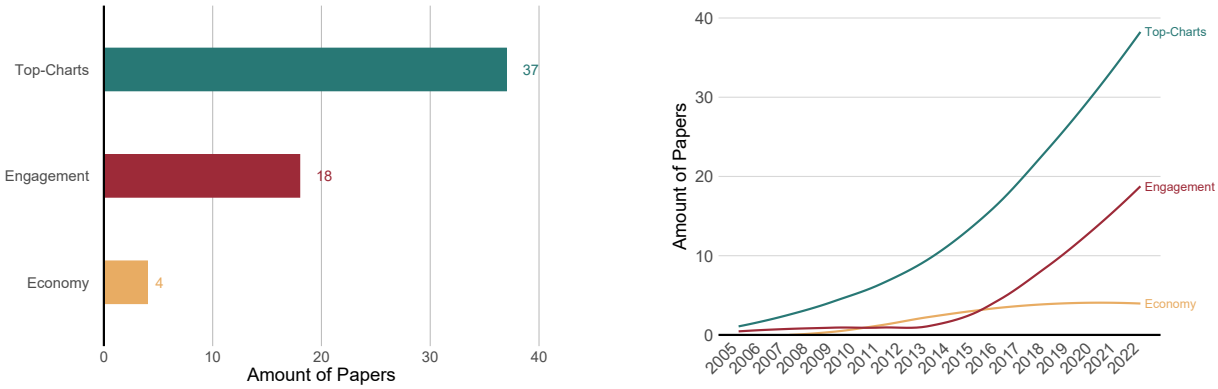
music reviews, play counts, and others. A similar score is available on Spotify Web API, called *popularity*, which is based on total number of plays compared to other tracks as well as how recent those plays are. Although Spotify uses several algorithms to determine popularity, in general, the more a song is played, the higher its score. Within the scenario of streaming services, the play count can be comparable to digital downloads. In other words, the popularity rating can translate valuable and useful information about musical success. Most studies have used such score as the label or response variable in learning models for predicting popularity [5, 38, 79]. Differently, Silva et al. [112] consider not only the popularity score but also the number of followers to define a successful artist.

Play counts have also been explored by authors using different platforms as a success measure. Examples include the music streaming service KKBOX,[27] which mainly targets the music market of Southeast Asia. It features over 50 million legal tracks and is currently available in Taiwan, Hong Kong, Japan, Singapore and Malaysia, with over 10 million users. Interested in distinguishing hits and non-hits, Yang et al. [129] consider play counts of a song from KKBOX streaming service to define song success. Likewise, Yu et al. [131] use the number of plays of one artist's songs as the target variable. However, in this case, the dataset was collected from a competition organized by the Alibaba Group[28] in China.

Besides streaming platforms, a popular music data service is the website Last.fm.[29] Founded in the United Kingdom in 2002, it has been used as a music recommender system, where a detailed profile of each user's musical taste is available. As a result, it is a useful information source about the listeners' behavior. Dewan and Ramaprasad [32] and Ren et al. [99] characterize song success based on Last.fm listeners, which are highly correlated with Amazon Sales rank. Similarly, Shulman et al. [110] define song success based on the number of people who have listened to or loved a song, almost as a sort of consumption measure. Another famous website is YouTube,[30] which offers a wide variety of user-generated and corporate media videos and encourages user engagement within the platform. With more than two billion users and one billion hours of video visualization per day, metrics like the number of views and number of like votes can be a valid measure of popularity [23].

[27] KKBOX: http://www.kkbox.com
[28] Alibaba Group: http://www.alibabagroup.com
[29] Last.fm: http://www.last.fm
[30] YouTube: http://www.youtube.com/

Figure 2.7: Most common success perspectives in Hit Song Science (a) and their evolution over the years (b).



Source: The Author.

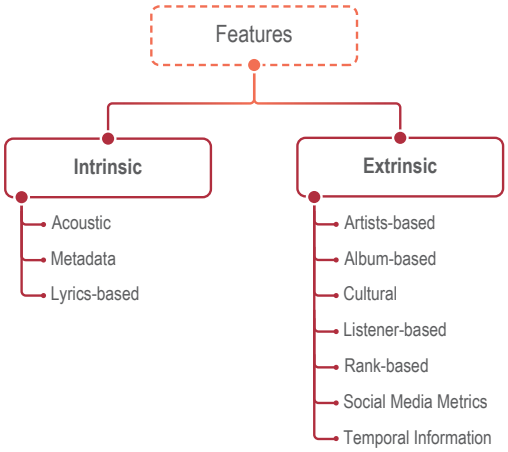
2.4.4 Discussion

In summary, the success of a song can be described from distinct perspectives, including: how many times a song has reached a top-chart, has been played on video clips, has been consumed via streaming, and so on. Essentially, existing measures of success fall into three categories: *Top-Charts*, *Economy*, and *Engagement*, representing the different points of view. Researchers traditionally seek to explain the phenomena of artistic success based on accumulated wealth, prestige, and notoriety. However, with the recent changes in the music industry, there are new markets and new possibilities to measure the attractiveness of artists and the engagement of fans.

Even with the advent of these revolutionary changes, the *Top-Charts* perspective remains the most accepted and, perhaps, the most reliable, as shown in Figure 2.7a. With almost 63% of the considered studies, the music charts evolved into becoming the primary source of tools measuring music success, alongside a booming music industry (Figure 2.7b). On the other hand, with the introduction of digital distribution mechanisms, social media and streaming platforms have created new ways of measuring artistic success, mainly about the engagement of listeners. This phenomenon becomes clear in Figure 2.7b, with the curve of the Engagement perspective increasing in the 21st century.

Such results may indicate that the definition of success in HSS studies depends mainly on the data available. In other words, the easy access to data on the web can be the main reason behind the vast majority of studies using the definition of success from top-charts, social media, and streaming platforms. Although the *Economy* perspective is supposed to be a strong indication of artistic success, album sales data availability is limited. Information such as profits, revenues, or dividends is generally not publicly available. The sources of data on album sales are generally in the form of rankings, without

Figure 2.8: Proposed hierarchical taxonomy for musical features. Intrinsic features are directly extracted from the song, while extrinsic ones are related to other objects and agents that influence the success of a song.



Source: The Author.

explicitly disclosing the sales figures. Therefore, to properly measure artistic success, a complete and easily accessible source of data encompassing all three perspectives is fundamental.

2.5 Musical Features

The success of a given song may be associated with a collection of factors related to the musical scenario. For example, recent studies show that characteristics including high *danceability* and low *instrumentalness* increase the popularity of songs. In other words, such songs tend to be more exciting, as a danceable music structure tends to put the audience in a good mood [3]. Going beyond explicit features, recent research also considers the strength of artist collaborations on producing hit songs, expanding research on musical success to another level [112]. Such perspectives use a large set of features (e.g., danceability, instrumentalness, etc.) in the HSS context, and these features are the basis of hit song prediction models. However, there is no unique set of features for a successful model.

In this section, we propose and describe a novel taxonomy for the most frequently used features in HSS, presented in Figure 2.8. Here, we can divide such descriptors in two main groups, according to their relation with the song itself, which is the main object of this field. First, the *Intrinsic* features (Section 2.5.1) are those directly extracted from the audio (i.e., acoustic fingerprints, lyrics) and information such as genre, duration, and the

number of artists. The second group includes all features related to *Extrinsic* agents or objects that may influence directly or indirectly the musical success, i.e., artist popularity, album sales, and the number of streams for the considered song (Section 2.5.2).

2.5.1 Intrinsic Features

Here we present the intrinsic features of music. Section 2.5.1.1 presents the Acoustic Features, which rely solely on musical data extracted from the audio (including different aspects of audio properties). Section 2.5.1.2 presents the Metadata features, which include information as title, duration, genres as well the participant singers. Finally, we discuss features based on song lyrics in Section 2.5.1.3.

2.5.1.1 Acoustic Features

Despite different definitions of music, we can certainly assume that music is an art form whose medium is sound. Moreover, it is composed of numerous core elements, including: pitch (melody and harmony), rhythm, dynamics, and the qualities of timbre and texture. However, due to the countless existing styles of music, some of these elements can be emphasized, diminished, or omitted.

There are different taxonomies for categorizing elements of music. Standard taxonomy generally divides musical descriptors into three dimensions, namely timbre, pitch, and rhythm [105]. Alternatively, hierarchical taxonomies have also been proposed, based on the dimensions of the music instead [25, 42]. However, many works use acoustic features that can be easily collected from sources such as Spotify or EchoNest. Hence, Table 2.2 summarizes the most used acoustic features followed by a brief description (mostly extracted from Spotify[31]) and the works that use such features.

In the face of such diversity of elements, several works differ on the set of features used to classify hit songs. For instance, Dhanaraj and Logan [33] extract features from each song describing its main sounds. In particular, the authors characterize sounds using Mel-frequency Cepstrum Coefficients (MFCC), which are features focusing on music timbrel aspects. Using a different source, Pachet and Roy [91] consider a set of 49 audio features taken from MPEG-7 audio standard, including spectral characteristics (Spectral

[31]Spotify API Reference: http://developer.spotify.com/

Centroid, Kurtosis and Skewness, HFC, Mel Frequency Cepstrum Coefficients), temporal (ZCR, Inter-Quartile-Range), and harmonic (Chroma) features. Such features were selected given their generality, i.e., they do not contain specific musical information or musically ad hoc algorithms. Lee and Lee [67] follow a similar path by taking 82 MPEG-7 features from each considered sound, besides the complexity features (Chroma, Rhythm, Timbre, and Arousal), and MFCC features.

Other works extract their musical features from The Echo Nest platform as song descriptor, including *tempo, time signature, song duration, loudness* [86], *mode, danceability, energy, key* [50], *liveness*, and *speechiness* [9, 37, 49]. Overall, such features match those used to represent a song globally. In addition to the set of Echo Nest features, Ni et al. [86] compute detailed summaries of the songs such as the *Coefficient of Variance of Loudness* and *Harmonic Simplicity*. Likewise, Herremans et al. [50] include two additional features to incorporate a temporal aspect: *Timbre* and *Beatdiff*. In contrast, Singhi and Brown [116] considers not only *danceability*, *loudness*, *energy, mode, tempo*, but also *mean, median*, and *standard deviation* of *timbre, pitch*, and *beat duration* vectors.

Although researchers continue to use such musical attributes, the data source has shifted. All acoustic features are now available in the Spotify API, since March 2014, when The Echo Nest was acquired by Spotify. The main features used in recent works are *acousticness, danceability, energy, instrumentalness, key, liveness, duration, mode, speechiness, tempo, time signature,* and *valence* [3, 8, 38, 80, 83, 97]. In addition to using high-level features from the Spotify API, other works include low-level features such as MFCCs [79], the *Tonnetz* [47], the *Chromagram* (vector of twelve elements indicating how much energy is released by each class tone) [107], the *Octave-based Spectral Contrast* (considers the spectral peak, the spectral valley and its difference in each sub-band) [29], the *Spectral Centroid* (frequency indicator for the energy where the spectrum is centered) [122], *Spectral Bandwidth* [79] and *Zero Crossing Rate* (ZCR) [7].

In a different perspective, Ren et al. [99] extract only *tempo* (fast, moderate, slow) and *melody* (pitch, rhythm) as acoustic features. Lee and Lee [65] extract features from audio signals in viewpoints of musical aspects, namely *chroma, rhythm*, and *timbre*. In another approach, Ren and Kauffman [98] describe the musical construct vector (MCV), with Theme, Mood, Instrumental, and Genre, which reflect how acoustic content is perceived. Such high-level semantics are extracted from lower-level musical features, such as timbre, rhythm, and tempo using machine-based methods.

Table 2.2: The most used acoustic features, their brief explanation and corresponding works.

Feature - Description	Reference Alias
acousticness - A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.	P4, P6, P16, P18, P22, P23, P25, P30, P31, P37, P46
danceability - Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is the least danceable and 1.0 is the most danceable.	P4, P6, P15, P16, P18, P20, P21, P22, P23, P25, P30, P31, P37, P46, P55
energy - Energy varies from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud and noisy; e.g., death metal has high energy, and a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.	P4, P15, P16, P18, P20, P22, P23, P25, P29, P30, P31, P37, P46
instrumentalness - This feature predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.	P4, P18, P21, P22, P23, P25, P30, P31, P37, P38, P46, P55
key - The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. $0 = C$, $1 = C\sharp/D\flat$, $2 = D$, and so on. If no key is detected, it is -1.	P6, P15, P18, P20, P22, P23, P25, P30, P31, P37, P46, P55
liveness - Liveness detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.	P4, P15, P18, P22, P23, P25, P30, P31, P37, P46
loudness - The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing the relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.	P15, P18, P20, P22, P23, P25, P30, P31, P32, P37, P46, P55
mode - Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.	P6, P15, P18, P20, P22, P23, P25, P24, P30, P31, P37, P39, P46
pitch - Pitch is a subject feature, which is determined (perceived) by what the ear judges to be the most fundamental frequency of the sound.	P17, P36, P39, P55
rhythm - As a recurring pattern of tension and release in music, it describes how certain patterns occur and recur in the music and is related to the "danceability" of the music. Beat and tempo (beat-per-minute, BPM) are two important cues that describe the rhythmic content of the music which have been utilized in music classification.	P21, P27, P28, P31, P36, P39, P55
song duration - The duration of a track in seconds as precisely computed by the audio decoder.	P3, P15, P17, P18, P20, P22, P23, P25, P30, P31, P32, P38, P46
speechiness - Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g., talk show, audiobook, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.	P4, P15, P18, P22, P23, P25, P30, P31, P37, P46
tempo - This is the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.	P15, P18, P22, P23, P25, P30, P31, P32, P37, P39, P46
timbre - Timbre is the quality of a musical note or sound that distinguishes types of musical instruments, or voices. Timbre vectors are best used in comparison with each other.	P20, P21, P27, P28, P29
time signature - It is an estimated overall time signature of a track. The time signature (meter) is a notational convention to specify the number of bits per bar (or measure).	P15, P18, P20, P22, P23, P25, P31, P32, P46
valence - With values from 0.0 to 1.0, this metric describes the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g., happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g., sad, depressed, angry).	P4, P6, P18, P22, P23, P25, P30, P31, P46

Source: The Author.

Using yet other data sources, Interiano et al. [54] gather several features, including *timbre, tonality, danceability, voice, gender* (male/female), and *mood* from AcousticBrainz. Finally, Zangerle et al. [135] extract high and low-level features using Essen-

tia's pre-compiled extractors, which provide a variety of spectral, time-domain, rhythm, and tonal descriptors. It provides around 40 basic features (e.g., MFCCs, dissonance or silence rate), 11 rhythm features (e.g., beats per minute or onset-rate), and 13 tonal features (e.g., key or harmonic pitch class profiles) that serve as low-level input for our task. The high-level features include musical genre, mood, timbre, vocals/voice, or danceability.

As an alternative, some researchers opt to use tools to extract features from audio files. A popular example is the *librosa* python package [81] for music and audio analysis that includes features as *tempo* (beats every moment) [7], MFCC (impersonates a few sections of the human discourse generation and discourse discernment), and the *Consonant Element* (the symphonic segment inside a sound flag) [96], or even the *Spectral Centroid*, the *Spectral Flatness*, and *Zero Crossings*. Another useful tool for audio feature extraction is the MeloSpySuite [40] software, which includes a set of stand-alone *commandline* tools for extracting numerical or textual characteristics of melodies. Chiru and Popescu [23] apply the Discrete Fourier Transform (DFT) signal in the WAV file to take the chart of sound intensity in function of frequency. Since the magnitude represents the sound intensity, they extract for every second of every song the frequency of highest magnitude, i.e., the largest weight of all sound frequencies that are heard in a second.

2.5.1.2 Metadata

In the music context, metadata is descriptive information about a song. It is often used for discovery and identification and is, therefore, one of the core elements in the MIR research. This type of feature usually includes basic information such as title, author, genres, and so on [1, 9, 41, 54, 83, 88, 98, 108, 135]. However, some descriptive information about the song may often not be directly related to the audio signal itself. For example, artist location, artist familiarity, artist hotness, song hotness [50], lists of song tags [14], song type (Solo, Group, or Collaboration) [108], explicity (whether a track has explicit content) [8, 83], and available markets for release [79] may summarize musical aspects as well.

Although metadata is traditionally used to provide digital identification, in the HSS context, a useful purpose is to help find relevant information and discover musical resources. Even though most meta-information is generally discarded when building a prediction model, some researchers have been seeking to assess hit song predictors through metadata. Pachet and Roy [91] use a music and metadata database provided by the HiFind Company. The HiFind metadata are grouped in 16 categories, representing specific dimensions of music including style, genre, musical setup, main instrument,

country, situation, mood, character and language. By using a different database, Nunes and Ordanini [87] also explore the instrumentation. Specifically, the authors consider the number of instrument types audible for each song as the principal independent variable.

2.5.1.3 Lyrics-based Features

Lyrics form an integral musical component and can help solve complex MIR tasks. The lyrics of a song contain specific emotional content and have more power to change mood than audio features alone Singhi and Brown [116]. However, lyrics are considerably ignored in the overall MIR research compared to acoustic features, although lyric-based features remain a useful predictor widely used in the Hit Song Prediction task. In particular, lyrics have been considered a significant component of what makes a song a hit. For example, [33] extract descriptive features based on the semantic content of songs from lyrics using a Probabilistic Latent Semantic Analysis (PLSA) method Hofmann [52], which is an effective way to assess the similarity between songs based on lyrics Logan et al. [74]. In particular, each song is converted to a vector representing the likelihood that a song is about a pre-learned topic.

Further studies explore other lyrics-based features. Singhi and Brown [115] propose a novel hit song detection model using lyric features alone. Specifically, the authors consider a complete set of 24 rhyme and syllable features of the Rhyme Analyzer Hirjee and Brown [51], including syllables per line, rhymes per line, and links per line. The CMU Pronunciation Dictionary Elovitz et al. [36] was additionally used to transcribe plain lyrics of songs to a sequence of phonemes with indicated stress, resulting in seven new metric features. The authors found the quality of hit prediction improves as lyric length increases. Moreover, Singhi and Brown [116] also show rhyme, meter and lyrics matter to hit detection, and complexity is related to being a hit.

Along with the advent of AI and machine learning, researchers have explored more advanced and powerful techniques to extract lyric-based features [97]. Ren et al. [99] apply Latent Dirichlet Allocation (LDA) [16] to learn five topic distributions from lyrics. Likewise, Ren and Kauffman [98] use LDA to build a topic model to learn the semantic themes from a dataset of 4,410 tracks. To such a task, they complement the acoustic content and give the artist's meaning behind the music. The results found that around 65% of the tracks were about "love" and "life". In a bag-of-words fashion, Chiru and Popescu [23] extract words from lyrics along with their frequencies of appearance. The authors claim that lyrics are the most useful features in identifying whether a song will be successful or not.

Finally, to assess a multimodal learning, Martín-Gutiérrez et al. [79] consider a collection of features evoked from different modalities, including text, audio and metadata. Regarding the text modality, a set of descriptors is extracted regarding the corpus of the song lyrics. By using NLP techniques, a stylometric analysis is performed resulting in the following features: the total number of sentences, the average number of words per sentence, the total number of words, the average number of syllables per word, a sentence similarity coefficient, and a vocabulary wealth coefficient. Sentiment analysis on song lyrics has also been applied to predict the nature of hit songs Raza and Nanath [97]. Recently, Kamal et al. [56] have performed a sentiment analysis to obtain a value between -1 to +1, representing the polarity of the lyrics (-1 corresponds to negative, 0 to neutral, and +1 to positive lyrics). They observed most popular songs have a neutral sentiment, and there are more popular songs with positive sentiments than negative sentiments.

2.5.2 Extrinsic Features

In this section, we present the extrinsic perspectives of the songs, which model the musical ecosystem by incorporating social media, market data and so on. In summary, Section 2.5.2.1 goes over information about artists; Section 2.5.2.2 overviews the features extracted from albums; Section 2.5.2.3 brings up discussions about cultural aspects from songs; Section 2.5.2.4 regards the listeners as important resource in HSS; Section 2.5.2.5 overviews features based on chart rankings; Section 2.5.2.6 reveals the importance of social networks to HSS; and finally, Section 2.5.2.7 discusses the temporal features of musical success.

2.5.2.1 Artist-based Features

Some features are not directly related to the music structure, such as those based on artists. Among the characteristics analyzed by several works [54, 80, 88, 98, 103], we highlight the following: (*i*) basic information (e.g., the artist's title/id) [83]; (*ii*) demographic data (e.g., age, race, gender and nationality) [87]; (*iii*) the type of artist (e.g., whether solo, group, or collaboration) [108]; and (*iv*) the awards received by the artists, associated to big record labels [99].

Other works use artist popularity to help predict hit songs. However, the set

of popularity features varies among studies. One possibility is to consider information extracted from platforms such as Last.fm, which usually provides a set of five tags from each artist in the database previously labeled by the platform users [98, 110]. Other features can be included, such as the number of tags assigned to an artist, the number of listeners, the best position ever achieved in the Billboard charts [14], and the number of followers of the artist or a popularity score [79].

Regarding artist's popularity, other studies apply network science metrics on a success-based artist network [80, 88]. For example, in [112] consider well-known metrics dependent on the node (Clustering Coefficient, Eigenvector, Degree and Weighted Degree) and related to the whole graph (Closeness, Eccentricity, and Betweenness). To deepen such analyses, Silva and Moro [111] also consider the popularity over time of the artists from the *rank_score*[32] extracted from Billboard charts.

Beyond such charts, Ren and Kauffman [98] measure artist reputation and leverage information on the news on the Grammy, American, and Billboard awards. They also consider relevant artist record label, because major labels have more resources to produce and promote high-quality tracks. A similar work by Askin and Mauskapf [9] adds a dummy variable if a song was released on a major record label. However, they also include a set of dummy variables in each model to account for the number of songs an artist had previously placed on the charts. These previous song count dummies capture artists' relative visibility or popularity at the moment of a song's release. Finally, they also construct a variable called "multiple memberships" to account for artists who have released songs under different names or band compositions.

2.5.2.2 Album-based Features

Similar to artist-based features, albums have also useful descriptive information. Indeed, additional information related not only to the social environment but also the marketing strategies may be considered. For instance, in the music industry, a music release should be an essential component of any music promotion strategy [79]. Depending on the release format, a song can reach the top of the charts more quickly. While albums and EPs are suitable to attract attention and build the fanbase, singles are more helpful to promote the album and keep fans engaged.

Listeners today have very specific behaviors regarding music consumption, in such a way that artists need to adapt their strategy to be successful. Hence, to fit such novel listening habits, labels opt to release a steady stream of singles to generate momentum

[32]The *rank_score* is the inverted rating on a success chart.

and enthusiasm. In other words, the album type can act as a significant predictor. Middlebrook and Sheik [83] consider as predictors album-based information extracted from Spotify, in addition to track, artist, and audio features. In particular, the authors explore both *album_type* (album, single, or compilation) and *album_release_date* (the date the album was first release) features. Likewise, Silva et al. [113] consider the *album_type* as predictor, as well as the albums' total number of tracks.

Also based on the marketing strategies of the artists' careers, Bischoff et al. [14] rely on a list of assumptions for their music hit prediction algorithm. One hypothesis is previous albums of the same artist have a direct influence on the future success of the songs. Thus, the popularity of an album is measured by the highest position reached on Billboard, named as *peak position*. Moreover, the authors only include the top-5 albums that reached positions in the charts, since some artists have many previously released albums.

In HSS studies, the focus is usually on determining whether or not one can predict the success of a song. However, some researchers have engaged into predicting albums' success [5, 24, 32]. Going further, Chon et al. [24] try to predict an album's future. In particular, the authors investigate not only how long an album will stay on a top-chart (i.e., the album's lifespan) but also whether an album's position can be predicted. To do so, they develop an algorithm to calculate Euclidean distances between the first weeks' sales history of a new album and the same number of weeks from the average life-cycle patterns, and determine the expected number of lifespan with the minimum distance.

2.5.2.3 Cultural Features

From a cultural perspective, we can identify a specific feature used in HSS. Specifically, Buda and Jarynowski [20] investigate distances between European countries in a Cartesian/Euclidean two dimensional way. Although the most typical representation is geographical distances, other matrices of distances are possible. For instance, the cultural map of Europe was built by political scientists based on the World Values Survey. Hence, two dominant dimensions were chosen (explaining 70% of variations between countries): traditional values versus secular-rational values on the vertical y-axis, and the survival versus self-expression values on the horizontal x-axis.

Although is no other study using cultural features in its models, recent research specialized in cuts of data for selecting different markets. For example, Fan and Casey [37] find significant differences in hit song prediction by comparing Chinese and British charts. In addition, Oliveira et al. [88] use data from Spotify to build genre networks from

eight different countries, besides the global scenario. Thus, it is possible to identify the importance of local aspects (mapped into regional genres) in defining the hits for each market.

2.5.2.4 Listener-based Features

The listeners who are music lovers can contribute a lot to an artist's success, and they usually make up the artist's fan base. In this category of features, the works seek to use the listeners' characteristics in the hit song prediction algorithms. Such characteristics can be used in several ways. For example, Pachet and Roy [91] use a set of 632 labels on the artists' popularity, and such labels were manually interpreted by listeners. On the other hand, Yang et al. [129] collect a set of Taiwanese listener records over one year, in collaboration with KKBOX Inc. The final dataset gathered the play counts of 30K users for around 125K songs.

A trend-line proposed by Shulman et al. [110] captures information about the early adopters of songs that become a hit. The features gathered are their popularity, seniority, or activity level, which might be proxies for their influence. The authors split such features into two sub-categories: roots, which are features of the first users to adopt a song; and researchers, the features averaged over other early adopters. As network density, similarity between early adopters works as: high similarity implies a niche market or one that people with similar interests are likely to adopt, while low similarity might indicate an item that could appeal to a wide variety of people.

Similarly, Herremans and Bergmans [49] work with a dataset of 854,060 records, which includes time, date, user, song, artist. Such a dataset contains a row for each song at each week (instance) and a column (feature) for each user indicating the play count to a particular song. They also enriched the dataset with a predictive feature set to 1 in case a potential song becomes a hit in the future. Furthermore, Berns and Moore [12] define a dataset using individual classifications according to the listeners' genre preferences. Each survey participant was given a list of six musical genres to rank the genres from 1 (likes the most) to 6 (likes the least). Then, the final dataset consist of a top three genre for each participant.

2.5.2.5 Rank-based Features

Dealing with hit song predictions requires reinventing the creation and use of features. Several works use features based on the musical rankings provided by entities that measure the popularity of the songs [121]. For example, Bischoff et al. [14] extract some implicit features for both artists and tracks by combining their top reached position on Billboard and their HITS scores— computed by applying the HITS algorithm on a graph using artists, tracks and tags as nodes. In a simpler approach, Koenigstein et al. [63] add the information of a song's debut rating on Billboard as an input to the proposed algorithm. Such a song's debut rating on the Billboard is a valid attribute that may have additional explanatory information for the algorithm.

In the same direction, some works consider that the ranking of each song depends on both streaming and download volumes [130]; others add the period of time that the song remained on the Billboard chart after its first entry [60, 83]. Lee and Lee [65] consider the chart's first performance as ranking scores for the first and second weeks. The initial chart performance is significant to predict popularity from two perspectives. First, it is a part of the entire popularity pattern over time and, therefore, conveys partial information about long-term popularity patterns. Second, the popularity of a song's early stage influences future popularity by increasing the song's visibility.

Ren and Kauffman [98] analyze the top-rank position before release; in other words, if a song gets in Top 50/100/150 before the new track is released. They also consider the first top-chart rank when a track first reaches a top-chart ranking on Last.fm. Shin and Park [108] include features such as the song's inaugural rank on the chart, peak rank, time to reach the peak position after appearing on the chart, and the time exit from the chart after reaching the peak.

2.5.2.6 Social Media Features

Features based on social networks can enrich the task of hit song prediction because the presence of the artist's fan bases sharing music content. Hence, a wide range of possibilities in defining features for prediction is available [121]. For instance, in the work of Salganik et al. [103], the participants made decisions about which songs to listen to; and while listening to a song, they were invited to rate it from one to five stars (the interval from "I hate It" to "I love it"). Then, to understand the social influence, the participants had the opportunity to download the song and see how many times each

song was previously downloaded. Thus, the participants received a relatively weak signal about the social influence, which they were free to use or ignore in addition to their own musical preferences.

As social interactions progressed on the internet, peer-to-peer interactions emerged. Although peer-to-peer networks are the main cause of music piracy, they are also used to sample music before purchase. In such direction, Koenigstein et al. [63] investigate the relations between music file sharing and sales using Peer-to-Peer music information. They compare file-sharing songs to their popularity on Billboard charts and show a very strong correlation (0.88-0.89). Finally, they show how this correlation can improve algorithms to predict a song's success at the Billboard by using Peer-to-Peer information.

From a distinct social perspective, Abel et al. [1] characterize the music albums on a given day by extracting the following features from the blogosphere: (i) number of posts that include the music album title; (ii) number of posts in which the artist name appears; and (iii) number of posts where the artist name and album title appear. Likewise, Kim et al. [60] collect tweets with music-related keywords and calculate the song and artist popularities as a number of tweets associated with a song and artist respectively.

Dewan and Ramaprasad [32] study the interplay between blog buzz, radio play and music sales at the album and song levels. They gather weekly data on the volume of song-level blog buzz from Google Blog search, song-level unit sales, and radio play measured by the number of "spin" from Nielsen SoundScan. The blog buzz is measured by the number of blogs that mention the artist and song names in a given week. They find the relationship between song buzz and sales is stronger for niche music comparable to mainstream music, and for less popular songs inside albums. We can relate this buzz to the commonly called early adopters. Shulman et al. [110] defend that most of the predictive power comes from looking at how quickly songs reach their first early adopters. Hence, they study the feature structures of the network around early adopters splitting them into two sub-categories: (i) ego network features to relate the early adopters to their local networks; and (ii) subgraph features that hold only connections between the early adopters.

Ren et al. [99] understand that social context positively impacted the data sample analyzed, improving music content as a predictor for a music track's popularity on the web. Thereby, for predicting such song popularity online, they consider the number of user comments in the first five weeks after the track was released as a feature. Similarly, Araujo et al. [5] collect messages from Twitter referring 30 days before the release of the examined albums. They choose such a time window to keep up with the growth in listeners' expectations as the launch date approaches. Also using data from Twitter, Tsiara and Tjortjis [121] performed sentiment analysis on collected tweets regarding a song and its artist to predict the charts in the future.

Lastly, Cosimato et al. [27] use social media features as follows: (i) the number of

reached people on social media (fans); (ii) the interests in specific albums measured by how much the album's author is a trend on the web; and (iii) the collective opinion about the album or the singer on social media. Besides, they consider that a famous singer could influence the success of a new artist by collaborating in one or more of his songs.

2.5.2.7 Temporal Information

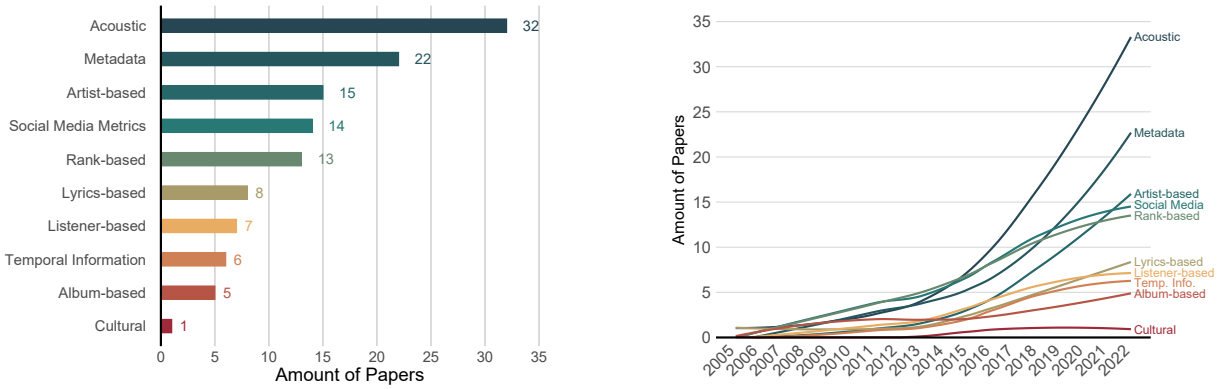
The last classification proposed for extrinsic features is temporal information. Such characteristics deal with the speed with which music reaches success (after initial single release) [20], or even how long it remains at the top. Another view is to understand the behavior of the songs before they are included in the charts, the time during their fame (already in the charts), until the moment when it stops appearing on the charts [110]. There are also works that analyze temporal distance to the release date; that is, the number of days when a music album will be or was released [1].

Although non-temporal features contribute to hit song prediction, the temporal ones are relevant because they describe the evolution of musical success and allow the identification of trends in such time-related data. Such predictors are also valuable as they make the prediction more realistic by increasing the accuracy of the models [110]. However, despite their great relevance, temporal data are difficult to be get because the main data sources do not provide them, impacting the number of studies considering such features.

2.5.3 Discussion

In this section, we discussed the main features used in works that predict a hit song. We also proposed a taxonomy to better organize the types of existing musical features divided into *Intrinsic* and *Extrinsic* features (Figure 2.8). The first group contains the intrinsic characteristics (genotypes) to the audio of the songs, as well as the information directly related, whether to describe the song or its lyrics. The second one is the group of features that describe the extrinsic characteristics associated with the songs (phenotypes). Such a group of features is composed of data obtained from artists, albums, cultural aspects (closely linked to the genre), information from listeners, data from charts, social and temporal networks.

Figure 2.9: Most commonly used musical features as predictors (a) and their evolution over the years (b).



Source: The Author.

Figure 2.9 presents the most used features and their evolution over time. Not surprisingly, the *Intrinsic* features are the most used predictors within HSS, as they are the most natural way to perceive music. In fact, this was the first set of features used by early HSS studies [33, 86, 91] (Figure 2.9b). With the popularization of the online platforms in the 2010s, social media gained notoriety in the music scene. Hence, *Extrinsic* features assessing social interactions were included in the prediction models, based on the hypothesis that the popularity of a song may be directly related to how much people talk about it online [27]. As depicted by Figure 2.9a, social media metrics are in the top three most used features and rank #1 within the *Extrinsic* ones.

Overall, music is a multimodal concept that can be described in several aspects. For instance, it can be translated into audio signals, encapsulated into the lyrics and descriptive metadata, or even into social web content. Nevertheless, most current research focuses on a unimodal or bimodal approach (e.g., acoustic + lyrics, acoustic + social, etc.). Although recent studies that consider such scenarios have good results, some aspects are not properly adopted, such as cultural, temporal, or album-based factors. Therefore, we believe that a more holistic view could better express prior knowledge of the data structure and fully exploit the valuable information that is encoded in it.

2.6 Learning Methods

One of the main steps of Hit Song Science is to discover the set of predictors that contribute to the success of a song. In general, most works use machine learning approaches for such a task. In this section, we summarize the principal techniques of machine learning used for hit song prediction. We cover the main classification algorithms

in Section 2.6.1, highlight the principal algorithms of regression in Setion 2.6.2, and go over other approaches in Section 2.6.3.

2.6.1 Classification

Several studies tackle hit song prediction as a binary classification task. Given the features of a song, the task is to classify it as a hit or a non-hit. Table 2.3 summarizes the selected works that apply Classification algorithms to Hit Song Science. Such a table includes the adopted success perspective, list of features, number of songs, classifiers utilized, and the accuracy of the best algorithm, which is in bold. There are several algorithms to address the binary classification task, the most studied in hit song prediction are as follows.

Table 2.3: Main features and machine learning methods used in classification approaches for Hit Song Science.

Year	Ref.	Success	Features	# Songs	Classifier	Acc.
2005	P14	Top-Charts	Acoustic, Lyrics-based	1,700	**SVM**, Boosting Classifiers	0.69 [A]
2008	P35	Top-Charts	Acoustic, Metadata, Listener-based	32,000	**SVM**	0.41 [B]
2009	P8	Top-Charts	Metadata, Rank-based, Artist-based, Album-based, Social Media Metrics	317,058	SVM, Naive Bayes, **Bayesian Networks**, Decision Trees	0.75
	P26	Top-Charts	Rank-based, Social Media Metrics	N/A	**Decision Trees**	0.89
2010	P1	Economy	Metadata, Temporal Info., Social Media Metrics	N/A	Naive Bayes, RBF Neural Network, SVM, **Decision Table**, One Rule, BFTree, LAD Tree, Simple Cart	0.50
2011	P32	Top-Charts	Acoustic	5,947	**Perceptron**	0.58
2014	P20	Top-Charts	Acoustic, Metadata, Artist-based	3,452	Decision Tree, **RIPPER**, Naive Bayes, Logistic Regression, **SVM**	0.85
	P24	Top-Charts	Rank-based, Artist-based, Social Media Metrics	178	**Random Forest**	0.84
	P33	Top-Charts	Metadata	2,399	**Logistic Regression**	0.65
	P47	Top-Charts	Lyrics-based	6,815	**Bayesian Network**	0.86
2015	P17	Top-Charts	Acoustic, Metadata	266	**Random Forest**	0.52
	P27	Top-Charts	Acoustic, Rank-based	867	**MLP**	0.71
	P48	Top-Charts	Acoustic, Lyrics-based	6,815	**SVM**, Bayesian Networks	0.69 [A]
2016	P39	Engagement	Acoustic, Lyrics-based, Artist-based, Social Media Metrics	1,961	Decision Tree, SVM, **Random Forest**, Bagging	> 0.7
	P43	Engagement	Temporal Info., Social Media Metrics, Listener-based	5.8M	**Logistic Regression**, Random Forest, SVM	0.81
2017	P19	Top-Charts	Acoustic, Temporal Info., Listener-based	982	RIPPER, **Logistic Regression**, SVM, Naive Bayes	0.79 [A]
2017	P38	Top-Charts	Acoustic, Metadata, Lyrics-based, Rank-based, Artist-based, Social Media Metrics	3,881	SVM, Bagging, **Random Forest**	0.80
2018	P16	Engagement	Acoustic	233	**Decision Tree**	0.73
	P21	Top-Charts	Acoustic, Metadata	500,000	**Random Forest**	0.86
	P28	Top-Charts	Acoustic	16,686	**SVM**	0.70

Continued on next page

Table 2.3: (continued from previous page)

Year	Ref.	Success	Features	# Songs	Classifier	Acc.
	P36	Engagement	Acoustic	8,000+	Logistic Regression, SVM, Naive Bayes, Random Forest, Neural Network	N/A*
2019	P3	Top-Charts	Acoustic	N/A	Ada Boost, Random Forests, Bernoulli, Gaussian Naive Bayes, **SVM**	0.89
	P12	Top-Charts	Social Media Metrics	N/A	**Random Forest**, SVM, MLP	0.97
	P31	Top-Charts	Acoustic, Metadata, Rank-based, Temporal Info., Artist-based, Album-based	1.8M	Logistic Regression, Neural Network, **Random Forest**, SVM	0.89
2020	P4	Top-Charts	Acoustic, Metadata	N/A	**SVM**, Gaussian Naive Bayes, kNN, Logistic Regression	> 0.8 [A]
	P29	Engagement	Acoustic, Metadata, Lyrics-based, Artist-based	101,939	**Deep Neural Network**	0.83
	P30	Engagement	Acoustic, Artist-based	N/A	**SVM**	0.81
	P37	Top-Charts	Acoustic, Lyrics-based	647	**Logistic Regression**, Decision Trees, Random Forest, Naive Bayes	0.52
	P49	Top-Charts	Rank-based, Social Media Metrics	N/A	Decision Table, Filtered Classifier, **Logistic Model Tree (LMT)**, Logistic Regression	0.96
2021	P25	Top-Charts	Acoustic	6,209	**Logistic Regression**	0.68 [A]
	P23	Top-Charts	Acoustic	37,236	Logistic Regression, Decision Tree, **Random Forest**, Naive Bayes, KNN, XGBoost	0.91
	P22	Engagement	Acoustic, Metadata, Lyrics-based	18,000+	**Random Forest**, SVM, Decision Tree, K-Nearest Neighbours, Logistic Regression and Naïve Bayes	0.85
	P18	Engagement	Metadata, Acoustic	130,663	Multi-variate Linear Regression, Logistic Regression, Decision Tree, Random Forest, Boosting Tree, **Neural Networks**	0.83
	P50	Engagement	Acoustic, Metadata	73,482	Logistic Regression, Random Forest, **SVM**, W&D Neural Networks, FCN	0.71
2022	P46	Top-Charts	Acoustic, Metadata, Album-based, Artists-based	911,027	Random Forest, Gradient Boosting, Support Vector Machines (**SVC** and NuSVC), MLP	0.82 [A]
	P51	Top-Charts	Acoustic, Metadata	73,482	Logistic Regression, Random Forest, **SVM**, Deep Feedforward Neural Network, KNN, CNN	0.71

Ref.: Reference alias **Acc.**: Accuracy * Final results are not presented.
[A] **AUC**: Area Under the Receiver Operating Characteristic (ROC) Curve [B] **min-f1**: minimum F-Measure

Source: The Author.

2.6.1.1 Random Forest

Random Forest classifiers are flexible and easy-to-use machine learning algorithms that frequently produce excellent results, even without adjusting hyperparameters. Random forest creates a set of decision trees (a forest) from a randomly selected subset of the training set. In a decision tree, the internal nodes represent a test on a feature, the

branches represent an exit from the test, and the leaves represent a vote to a class label (i.e., hit or non-hit). Next, the algorithm aggregates the votes from different decision trees to elect the final class of the test object. In other words, an object is assigned to a class that has the most votes from all trees.

In HSS, many works use the random forest algorithm to predict the success of a song [7, 41, 60, 96–99, 110]. We highlight [54] who add the 'superstar' variable (i.e., if the artist had appeared in top charts) based on music acoustic features. Such variable quantifies the contribution of purely musical characteristics in the songs' success, and suggests the time scale of fashion dynamics in popular music. Similarly, [83] test four classification models (Logistic Regression, Neural Network, Random Forest and Support Vector Machine) on a dataset with approximately 1.8 million hit and non-hit songs. Overall, the best model is the random forest, which predicts Billboard song success with 97% accuracy in [27].

2.6.1.2 Naive Bayes

Naive Bayes classifiers are supervised learning algorithms that apply Bayes' theorem with the naive assumption of conditional independence between every pair of features. In other words, such a classifier estimates the probability of a hit or non-hit holding on the assumption of musical features being conditionally independent. Due to the independence assumption, the class-conditional probability for every feature combination does not need to be calculated; only the conditional probability of each feature x given Y has to be measured. It is a practical advantage since a good estimate of the probability can be obtained without a very large training set [1, 7, 8, 14, 49, 50, 96, 97]. Although such independence assumption is a weak assumption in practice, several studies prove that naive Bayes competes fairly with more sophisticated classifiers, having similar or slightly inferior performance than other approaches. For example, [49] not only focuses on audio features to predict a hit song but also include social media listening behaviors to identify early adopters. They use a large dataset of social listening behavior from Last.FM. The first better result of AUC was the logistic regression (0.79) facing Naive Bayes to the second one (0.70), while SVM performs poorly (0.50). However, naive Bayes is especially resistant to isolated noise points, robust to irrelevant attributes.

2.6.1.3 Bayesian Networks

Bayesian networks are probabilistic graphical models that use Bayesian inference for probability computations, which are composed of random variables represented as nodes and their conditional dependencies represented as directed edges. The joint probability of the variables outlined in the directed and acyclic graph can be estimated as the product of the individual probabilities of each variable through the edges, conditioned on the node's parent variables. In other words, Bayesian networks are graphs that express how the occurrence of certain variables depends on the state of another. In hit song prediction, such a strategy is used by [14, 115, 116] with satisfactory results. For instance, Bischoff et al. [14] reach a value of 0.883 for the AUC measure, 0.788 precision, and 0.858 recall for hits, while the overall accuracy is 81.31%.

2.6.1.4 Support Vector Machine

SVM is a computer science concept for a set of supervised learning methods that analyze data and recognize patterns. The standard SVM takes a dataset as input and, for each given input, predicts which of two possible classes it is part of. SVM is a non-probabilistic binary linear classifier based on the theory of statistical learning [26]. In a summary, SVM finds a line of separation (called as hyperplane) between data from the hit and non-hit classes. Such a line seeks to maximize the distance between the closest points in relation to each of the classes.

Many HSS approaches use SVM as one of the main algorithms for hit song prediction [1, 7, 8, 14, 33, 49, 50, 56, 67, 80, 83, 91, 96, 98, 99, 110, 113, 116, 125, 126]. The first exploration within the hit song science domain is by [33]. The authors use acoustic characteristics and lyrics to build a Support Vector Machine, which is then tested over a small dataset with promising results. On the other hand, [14] trained (with Billboard as ground truth) and tested the classifier on the total set of instances (both hits and non-hits), corresponding to each of the hit class ranges, using social media data (from Last.fm) as features. Also, [116] use weighted-cost SVMs (in LIBSVM), which assign different misclassification costs to instances depending on the class they belong to. Also, using the LIBSVM package, [67] suggest three different experiments investigating the features' popularity prediction performance. They train SVMs with the extracted features to perform binary classification of each popularity metric. The boundary of the two classes is set to the median value of each popularity metric in the training data set. The

radial basis function (RBF) is used as the kernel function of the SVMs. Despite SVM overall achieving satisfactory results, the best performance of the SVM classifier achieved an accuracy of 89% in work by [7] when predicting the song's popularity two months in advance.

2.6.1.5 Decision Tree

Decision trees are non-parametric supervised machine learning methods, widely used in classification tasks. In a decision tree, a decision is made by walking from the root node to the leaf node. Although decision trees are conceptually simple, they are relevant predictors. Its complexity is logarithmic in the prediction stage. However, decision trees have some problems that can degrade their predictive power. A tree grown to its maximum depth can overfit the training set, and may degrade its predictive power to new data. Pruning the decision tree may mitigate such a problem. In addition, they are unstable models (high variance), because small variations in training data can result in completely different trees. Training several different trees and aggregating their predictions can avoid high variance.

In hit song science, decision trees are used for predicting hits as well [14, 38, 50, 63, 97, 99]. One of the most popular algorithm implementation is the C4.5 Algorithm [128]. According to [93], C4.5 Algorithm uses a divide and conquer approach to recursively build trees. Also, Decision Trees can be considered as one of the easiest models to understand classification due to the linguistic nature [77]. As mentioned by [50], the tree data structure consists of decision nodes and leaves. The class value is specified by the leaves (in this case hit or non-hit), and the nodes specify a test of one of the features. When a path from the node to a leave is followed based on the feature values of a particular song, a predictive rule can be derived [102]. As the best result for the Decision Tree classifier, [63] achieves 89% of accuracy by predicting a song's success on the Billboard in advance, using Peer-to-Peer information.

2.6.1.6 Neural Network

In computer science, neural networks are computational models inspired by the central nervous system (as the brain). They are capable of performing machine learning as

well as recognizing hidden patterns and correlations in raw data, grouping and classifying them, and - over time - continually learning and improving. Neural networks are generally presented as systems of linked neurons, which can compute input values, simulating the behavior of biological neural networks. A single neuron is a component that calculates the weighted sum of several inputs, applies a function, and forwards the results. Each neuron receives signals from input variables and passes on a weighted and treated version of that signal. In parallel, such neurons form a hidden layer of the neural network. The output of each neuron is a variable in the input of another hidden layer. Such hidden layers may be stacked, then producing a deep neural network.

Regarding research on the hit song science, few works use neural networks [1, 43, 79, 83, 96, 125, 126]. For example, Rajyashree et al. [96] define define three layers besides an input: two hidden and one output layers. The recent outbreak of Deep Learning models has changed the paradigm in pattern recognition and classification tasks in HSS. Recent research [79] propose different experiments for predicting the popularity of a song and describes a final end architecture composed of two main stages: a deep encoder and a deep neural network based on the model 1-A, which they selected as the best model since it provides good performance and a low computational cost. They achieve the best performance with an accuracy of 83.46% by using Neural Networks. Similarly, [43] shows that most songs can be found with about 83% accuracy according to audio features and artists' history profiles using Neural Networks.

2.6.1.7 Multi-layer Perceptron

MLP is an example of a neural network and can be considered as a logistic regression classifier, where the input is first transformed using a learned non-linear transformation. An MLP consists of at least three layers of nodes: input, at least one hidden layer, and output. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. However, a single hidden layer is sufficient to set MLPs as a universal approximator.

The research proposed by [65] uses MLP with one hidden layer as a classification method for predicting hit songs. They tried various numbers of hidden neurons and, as a result, the number of hidden neurons is set to 15, which produced good results. Their classifiers are trained to perform binary classification (high vs. low) of the six popularity patterns defined in the paper. MLP is also one of the classifiers used by [27] to predict the music album rank in the Billboard 200 Chart, alongside Random Forest and SVM. However, in this case, the Random Forest model is the one that performed better, with an

accuracy of 97%, while MLP achieved 86% of accuracy. Both studies are not comparable, as they consider a distinct set of features and success metrics.

2.6.1.8 Logistic Regression

Logistic regression is a supervised classification algorithm whose goal is to produce a model that allows predicting values taken by a categorical variable, often binary, from a series of continuous and/or explanatory variables. The logistic regression is distinguished from linear regression because the response variable is categorical. Logistic regression becomes a classification technique only when a threshold is defined. The setting of such threshold is a crucial aspect of Logistic regression and is dependent on the classification problem itself. As a prediction method for categorical variables, logistic regression is comparable to the supervised techniques proposed in automatic learning (decision trees, neural networks, etc.), or even the predictive discriminant analysis in exploratory statistics.

Several studies use the logistic regression algorithm applied to hit song science [8, 49, 50, 59, 83, 87, 96, 97, 110, 121, 125]. An interesting example is by Nunes and Ordanini [87], which uses logistic regression to document how the absolute number of distinct musical instrument types perceptible in a song affects its chances of being hit song. The results suggest that songs that do not follow conventional instrumentation and, instead, include an atypically low or high number of instruments have greater chances of becoming a hit. However, [110] maintains the best performance with about 81% accuracy with logistic regression. Most of the predictive power comes from looking at how quickly items reach their early adopters.

2.6.2 Regression

For classification models, the focus is toward the definition of a hit song by discrete values. On the other hand, this section highlights the models whose main objective is to predict a continuous outcome (y) variable based on the value of one or multiple predictor variables (x). This set of machine learning methods is known as Regression. Classification approaches lose relevant information about song popularity due to the binary conversion. Table 2.4 summarizes the selected works that apply Regression algorithms to Hit Song Science. It includes the adopted success perspective, list of features, number of songs,

regressors utilized, and the score of the best algorithm, which is in bold. Hence, Regression is also chosen for predicting hits, and the most applied algorithms in hit song prediction are as follows.

Table 2.4: Main features and machine learning methods used in regression approaches for Hit Song Science.

Year	Ref.	Success	Features	# Songs	Regressor	Score
2009	P26	Top-Charts	Rank-based, Social Media Metrics	N/A	**Decision Trees**	10.1 **AE**
2010	P1	Economy	Metadata, Temporal Info., Social Media Metrics	N/A	Linear Regression, **SMO Regression**, Bagging REPTree	73.28 **MAE**
2013	P15	Top-Charts	Acoustic	752	Linear Regression, **SVM**	0.39 **ER**
2014	P24	Top-Charts	Rank-based, Artist-based, Social Media Metrics	178	Linear Regression, **SVR**	0.75 $\mathbf{R^2}$
2017	P5	Top-Charts, Engagement	Album-based, Social Media Metrics	N/A	**Linear Regression**	0.96 $\mathbf{R^2}$
	P52	Engagement	Social Media Metrics, Listener-based	20,000	Linear Regression, Neural Networks	N/A*
	P38	Top-Charts	Acoustic, Metadata, Lyrics-based, Rank-based, Artist-based, Social Media Metrics	3,881	SVM, Bagging, **Random Forest**	0.73 **C**
	P6	Top-Charts	Acoustic, Metadata, Artist-based	27,000	**OLS Regression**, Binomial Regression	0.43 $\mathbf{R^2}$
	P10	Engagement	Acoustic, Lyrics-based	6,000+	**SVM**, kNN	0.16 **ER**
2019	P55	Top-Charts	Acoustic, Metadata	95,067	**Neural Networks**	43.84 **MAE**
	P54	Engagement	Metadata, Rank-based, Social Media Metrics	10,842	**SVM**, Neural Networks	90.6 **RMSE**
2020	P29	Engagement	Acoustic, Metadata, Lyrics-based, Artist-based	101,939	**Deep Neural Network**	0.09 **MAE**
	P30	Engagement	Acoustic, Artist-based	N/A	**SVM**	6.61 **MAE**
	P49	Top-Charts	Rank-based, Social Media Metrics	N/A	**SVR**, Random Forest, Bagging	4.05 **MAE**
2021	P50	Engagement	Acoustic, Metadata	73,482	**Linear Regression**, Random Forest, SVM, Convolutional Neural Network	0.47 **C**
2022	P51	Engagement	Acoustic, Metadata	73,482	**Linear Regression**, Random Forest, SVM, Deep Feedforward Neural Network, KNN , CNN	0.46 **C**

AE: Absolute Error **MAE**: Mean Absolute Error **ER**: Error Rate $\mathbf{R^2}$: Coefficient of Determination
C: Correlation **RMSE**: Root Mean Squared Error
* Although hit song prediction is addressed as a regression problem, the models are evaluated with ranking metrics.

Source: The Author.

2.6.2.1 Linear Regression

Linear regression is one of the simplest methods of Machine Learning. It is composed of an equation to estimate the expected value of a target variable y, given the values of a set of input variables X. Such regression is called linear because it considers that the relationship of response to variables is a linear function of some parameters. Linear regression models are often adjusted using least squares approach, but they can

also be adjusted by minimizing the lack of fit in some other standard (with less absolute deviations from regression), or by minimizing a penalty version of the minimum squares.

In the context of hit song science, the features (as discussed in Section 2.5) compose the set of variables X, and the expected value y is whether the song is a hit or not. Considering the task of Regression, the linear regression is the main algorithm used to predict hit songs [1, 5, 37, 60, 129]. We highlight the study of [129], which computes the audio signals to compose the feature vectors as the input of a single-layer neural network model; in other words, in a effectively linear regression.

2.6.2.2 Support Vector Regression

SVR is an extension of the SVM method (that was originally developed for class prediction) with the functionality of numerical regression. It determines the hyperplane that separates the instances of the target attribute by analyzing the distance between the instances positioned at the boundaries of the classes [123]. The family of methods derived from SVM uses functions of the kernel to produce mathematical transformations in the data, expanding the dimensionality of the representation in order to make them linearly separable. SVR is flexible as it allows to setup how much error is acceptable in the model and will find an appropriate line (or hyperplane in higher dimensions) to fit the data. The objective function of SVR is to minimize the coefficients (specifically, the loss function coefficient vector), not the squared error. SVR has an important advantage: its computational complexity does not depend on the dimensionality of the input space [10]. In addition, it has excellent generalization capabilities, with high forecasting accuracy.

There are several works using SVR on hit song science [23, 37, 60, 80, 98, 121, 131]. For example, [60] use SVR for predicting the popularity of the music based on Billboard rank. They use the radial basis function (RBF) for the kernel of the SVR model. When comparing with other regression models (linear regression and linear regression quadratic), SVR achieves a considerably high performance. It is interesting to notice that SVR allows to solve different variations of hit song prediction. For example, [98] use different combinations of Machine Learning methods to predict popularity duration, including SVR. They measure music popularity by using a Music Construct Vector with musical variables.

2.6.3 Others

HSS is a complex and dynamic domain that aims to understand how and why some songs become successful, involving analyzing multiple intrinsic and extrinsic factors related to the artist, music, and audience. Traditionally, researchers have used classification and regression models to predict a song's success, but these methods have limitations and may only consider some of the factors that influence success. As a result, there are alternative approaches to predict the success of songs.

Although there is still no consensus on the best algorithm for predicting hits, exploring different methods can provide additional insights into which factors influence the success of a song and how they interact with each other. Such alternative approaches contribute to a better understanding of the domain and can lead to more accurate predictions. The literary review methodology has pointed out 12 different approaches summarized in Table 2.5, such as experimental studies, time series prediction, rankings, neural prediction, statistical analysis, social networks, cluster analysis, and epidemiological analysis. These approaches highlight the diverse methodologies used in HSS and open directions for future research to continue exploring different perspectives to better understand musical success.

Table 2.5: Main features and methods used in other approaches for Hit Song Science.

Year	Ref.	Success	Features	# Songs	Method
2006	P41	Engagement	Social Media Metrics	48	Experimental Study
	P11	Top-Charts	Album-based	N/A	Time Series Prediction
2010	P53	Top-Charts	Rank-based	N/A	Ranking
2012	P7	Economy	Listener-based	120	Neural Prediction
2014	P13	Engagement	Rank-based, Social Media Metrics	1,000	Statistical Prediction
2015	P9	Top-Charts	Rank-based, Temporal info., Listener-based, Cultural		Social Network Analysis
2018	P42	Top-Charts	Metadata, Rank-based, Temporal info., Artist-based	7,560	Ranking and Statistical Analysis
2019	P45	Engagement	Artist-based	2,144	Statistical Analysis & Social Network Analysis
	P44	Top-Charts	Rank-based, Artist-based	7,185	Statistical Prediction
2020	P2	Top-Charts	Acoustic	100	Cluster Analysis
	P34	Top-Charts	Artist-based	13,380	Statistical Analysis & Social Network Analysis
2021	P40	Top-Charts	Metadata, Artist-based	950	Statistical & Epidemiological Analysis

Source: The Author.

- **Experimental Study.** Salganik et al. [103] measure the success of a song based on the market share of downloads of such songs.
- **Time Series Prediction.** Chon et al. [24] consider the life cycle of an album trajectory as the weekly positions from the first week to the last week on the Top

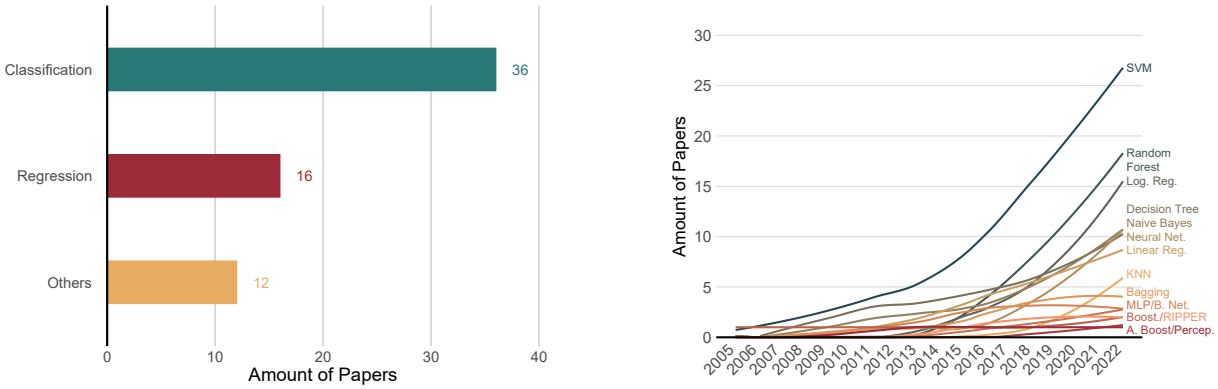
Jazz chart.

- **Rankings.** Yoo and Kim [130] measure the song's hotness by considering the high ranking songs that are given by download volumes, streaming volumes, and adjusted total volumes.
- **Neural Prediction.** With an unconventional approach, Berns and Moore [12] use functional magnetic resonance imaging (fMRI) to measure the brain responses of a small group of teenagers when listening to songs by unknown artists and, as a measure of popularity, sales (SoundScan) of these songs were totaled over three years.
- **Statistical Analysis.** Some studies use as features the number of listeners, high level of popularity of artists who have a high number of followers, and the ranking of the song/album/artist to describe the popularity over time [32, 108, 111].
- **Social Networks.** There are also studies that consider user engagement to measure the popularity of songs [20, 112].
- **Cluster Analysis.** Al-Beitawi et al. [3] perform a cluster analysis on the Top 100 Trending Spotify Song, considering ten musical features.
- **Epidemiological Analysis.** Rosati et al. [101] use the ability of the standard susceptible–infectious–recovered (SIR) epidemic model to fit the download time series of popular songs, and they conclude that such social processes underlying song popularity are similar to those that drive infectious disease transmission.

2.6.4 Discussion

The search for an ideal machine learning algorithm that predicts hit songs motivates researchers and the music industry to test various possibilities and combinations. There are algorithms with accuracy around 95%; still, in general, such works have limited scope and data, and they represent distinct facets that contribute to achieving musical success. Indeed, capturing the features necessary for prediction is hard. From another perspective, companies must deal with constant changes in the music market that generate new algorithm input features. For example, one of the first significant changes was the evolution of music consumption from physical media (CDs, DVDs) to digital media (streaming, downloads). More recently, the market is still trying to deal with trends in so-

Figure 2.10: Most common learning algorithms in Hit Song Science (a) and their evolution over the years (b).



Source: The Author.

cial networks (e.g., TikTok, Instagram, and Twitter) that produce 'successes' by making them an intersection for creative expression and playful sociality [2].

Hence, there is not *one ideal* algorithm for HSS or hit song prediction. However, people from the music industry (e.g., producers) who have access to large amounts of data (which are mostly not available to academia) can use the insights highlighted here to select or propose algorithms that work for their market. After all, in a billion-dollar industry, any percentage improvement in the bottom line can mean several million dollars in return. Still, by the universe of selected works in this survey, we note more significant favoritism towards classification algorithms, as Figure 2.10a shows. Among classification algorithms, we report a preference for Support Vector Machine and Random Forest (Figure 2.10b). Such algorithms are easy to implement and present satisfactory results. On the other hand, relatively fewer studies consider the prediction of hit songs as a regression task, and the most used algorithms are classic Linear Regression and Support Vector Regression. Other approaches use data from social networks, experimental studies, rankings, statistical analysis, clustering, time series prediction, and neural prediction.

2.7 Research Directions

Based on our overview of the techniques and tasks relevant to Hit Song Science, we now identify and discuss new directions for research that require further investigation. These topics are not the only open research problems within HSS, but they are key factors that may shed light on the science of what makes a song successful.

Dealing with multiple sources. Data integration is one of the main issues in many

Computer Science research fields. In Hit Song Science, this topic is becoming more relevant and necessary, as there is no unique data source for all necessary features and data. For instance, to the best of our knowledge, there is no data source that provides both acoustic and lyrics-based features. Furthermore, the lack of a unique and universal identifier for each music makes an integration involving several data sources very challenging. Besides, information such as the musical genre(s) of a given song is not standardized in all data sources, mainly due to the blurred line existent between music styles that are close to each other.

Regional markets' diversity. Most studies on Hit Song Science use data from the American market (e.g., Billboard Hot 100 Chart and Amazon Sales Data). This may be because the United States is the biggest music market in the world, which may facilitate the acquisition and use of such data. Research studies that consider music markets other than the USA focus mainly on European countries, such as the United Kingdom. However, there are many other relevant markets with distinct characteristics and behavior, which require an individual analysis of success. For example, South Korea, China and Brazil are among the top 10 music markets in the world[33], with a vibrant music scene and popular regional genres. Such genres have become popular in the global scenario as they connect with other well-established music genres (e.g., collaborations involving pop, k-pop, and Latin genres such as reggaeton). Therefore, as local engagement shapes the global environment, future work on HSS must consider the regional aspect, thus ensuring that music culture within such countries are accounted for.

Lack of standardized success metrics. Defining the popularity of a song is still a challenge, and each research study in HSS uses specific success metrics. In Section 2.4, we discussed and proposed a taxonomy for such metrics, but as there is no standard, researchers are unable to perform a fair comparison between their work and the existent literature on the subject. Hence, finding a way to properly generalize success would support future work on HSS to more accurately capture popularity definitions. Moreover, it would enable transposing their findings to a commonly understood metric, which then allows a complete evaluation by comparing performance with current work (as the many presented here).

Importance of social aspects. The ever-growing popularization of social networks in the last two decades has deeply changed the music industry. The propagation of songs in such platforms is fundamental in their success, as the viral phenomenon of songs in social media may lead a newly released one to stardom or even lead back a great hit from the past to the top of the charts. Since marketing has great impact on the future success of songs, it is increasingly important to consider the latest social platforms and features, which could as well give strong indications of a song's hit potential. Although

[33]IFPI Global Music Report: http://gmr.ifpi.org/

such features have been used in previous research, novel approaches on HSS need to combine both audio and social data to enhance hit prediction efficiency.

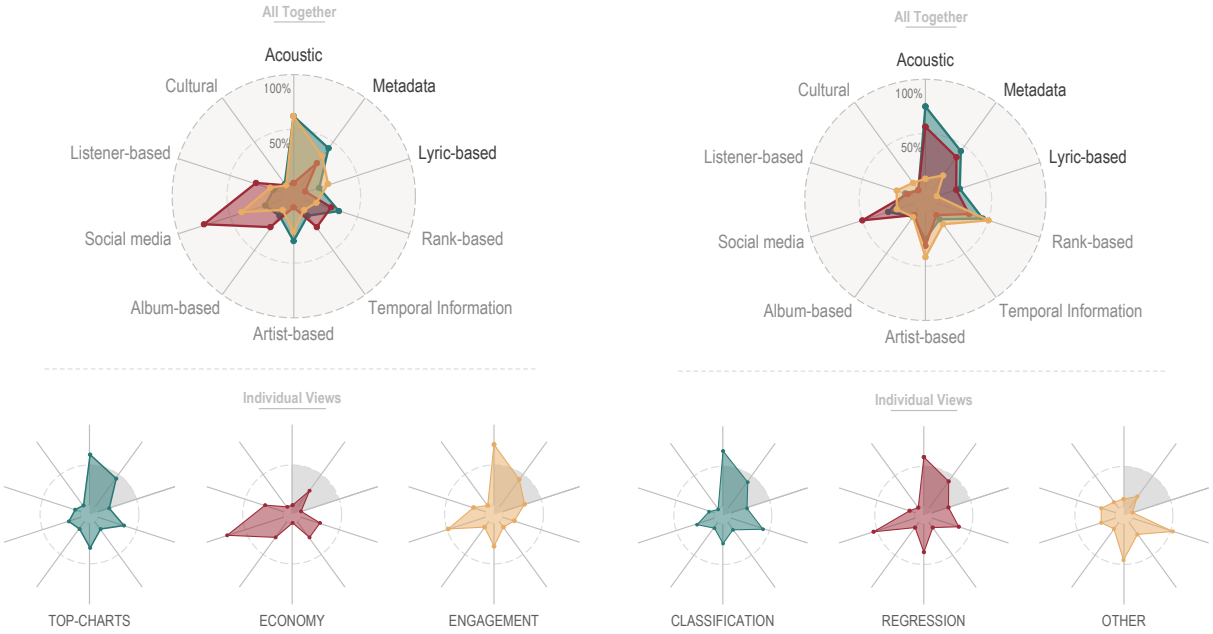
Framework for predicting and recommending hit songs. One of the main goals of Hit Song Science is to predict whether a given song will become a hit or not. Consequently, it is reasonable to extend this goal into a recommendation task in which hit songs will be returned based on the listener musical preferences. Although several studies assess such issues, there is still not a support tool for running such models and then standardizing assessment. For example, a framework on hit song prediction/recommendation should be able to receive as input a given success measure, a set of predefined features, and an algorithm to run the machine learning task (e.g., classification and/or regression). In fact, this would represent a great advance on the field of HSS, with benefits to both the academy and the music industry, as it allows scientists to better understand the success phenomenon and record label CEOs to properly invest in selected songs and artists.

2.8 Concluding Remarks

In this chapter, instead of emphasizing what worked and what did not, we aimed to present which approaches are the most used within HSS, as well as the main findings on the subject. To do so, we follow a generic workflow (see Figure 2.4) of four phases, which model this chapter's structure. First, we described the most commonly used data sources and discuss their applications. Second, we proposed a hierarchical taxonomy to classify the different success measures existent in the literature into three perspectives: *Top-Charts, Economy,* and *Engagement*. Next, we assessed the most frequently used features in hit song prediction models. Finally, we summarized the main learning methods (i.e., classification, regression and others) to predict whether a song will be a hit or not.

Despite following such a workflow, not all works use the same combinations of success measures, musical features and learning methods. Figure 2.11 compares the choice of types of features used with (a) the success perspectives and (b) the machine learning tasks. For success, most works considering the *Top-Charts* perspective use *Internal* features (mostly acoustic). However, some external ones such as *Rank-based* are also present, which is expected given the success measure choice. Both *Economy* and *Engagement* include different types of features directly related to their intrinsic nature. For instance, social media, listener-based, and rank-based features are strongly present in the *Economy* view, as they are fundamental to build up marketing strategies in order to boost sales. In contrast, works within the *Engagement* perspective rely mainly on information about

Figure 2.11: Comparison of HSS papers associating the features with the adopted success measures (a) and learning methods (b). The darker radial comprises the Internal features, and the lighter the External ones.



Source: The Author.

music consumption and social engagement (i.e., social media).

Regarding machine learning tasks, acoustic features are the most used in classification and regression approaches. Such features are usually represented in numerical variables, making easier their processing and further statistical analyses since many learning algorithms require this format as input. Furthermore, in comparison to the classification task, studies that tackle the hit song prediction as a regression problem also consider social media metrics. Finally, the external features are considerably used by other types of solutions, given the more flexible nature of such diverse techniques (e.g., clustering, statistical and social network analysis). Again, our objective is not to detail what works and what does not, as this is very particular concerning each target. However, we reaffirm our desire to exhaustively list what the research community has produced from the beginning of the HSS area to the present day.

Hit Song Science emerges as a multidisciplinary field within Music Information Retrieval aiming at predicting the success of a song before its release. Its interdisciplinary nature promotes not only benefits to the MIR community, but also the music industry as a whole. Moreover, such prediction studies may help music industry CEOs to maximize expected success by properly investing in selected songs/artists. We also identified key open research issues within HSS, revealing a broad scope for brand new improvements and advancements in such a field. Therefore, we believe that this study sheds light on the science behind musical success, serving as a base material for future research on Hit Song Science.

Still, while Machine Learning (ML) algorithms are powerful tools for predicting hit songs, it is essential to recognize that success in the music industry is based on more than measurable features. Personal creativity, inspiration, talent, and taste for musical preferences are difficult to quantify and may play a significant role in an artist's success. Therefore, a comprehensive understanding of the music market requires a more in-depth analysis beyond putting some features into ML algorithms and testing their performance.

Understanding the music market requires going through a set of mixed features, some of which are measurable and others not. In addition to social media metrics (such as followers, views, streams, likes, dislikes, and TikTok video sharing), we need to consider personal branding, musical creativity, and industry trends. We must also investigate how external factors (such as cultural shifts, economic changes, and technological advancements) influence the music ecosystem.

This thesis takes a step further to scrutinize the necessity of understanding the music market beyond machine learning techniques. Through a comprehensive literature review that includes both measurable and unmeasurable features, we aim to provide a better understanding of the factors that contribute to an artist's success. Ultimately, this will help music industry professionals make informed decisions about marketing strategies and artist development, leading to better outcomes for artists and the industry.

In conclusion, this chapter answers the first research question (RQ1 - How does current research deal with Hit Song Science?) considering the growth of the Hit Song Science area and its relevance through a comprehensive study with a complete overview of the main topics of HSS (success metrics, features, and machine learning algorithms). Furthermore, we emphasized the main research directions to evaluate open problems for future advances and improvements in such a relevant and promising field. In this sense, we identified that creative careers have always instigated researchers and the industry to understand how to build or improve them. However, we note that there is still room to explore the common characteristics that lead to successful musical careers, associating such attributes with the most relevant periods of the artists' careers. After all, a significant milestone in artistic careers is the existence of Hot Streaks, that is, those periods of great notoriety and presence in the music market. While Hot Streaks are almost standard in music careers, it could be more evident if there are any patterns to their onset. Consequently, we need to investigate further how to understand success beyond machine learning techniques, which are presented in the following chapters.

Chapter 3

Hot Streaks Modeling and Analyses

Music has experienced huge transformations in recent decades, shifting from ownership to access, from compact discs to streams, and mostly from Physical to Digital. Physical media is indeed constantly making room for the consolidation of the Digital era. Figure 3.1 shows this process, with a turning point in the early 2010s. In such a market, Physical media sales are still going on; whereas streaming services dominate music consumption, accounting for over 65% of the music industry revenue in 2021.[1] The scenario in Brazil is similar: as Latin America's largest music market, about 85.6% of music revenue comes from Digital media, against 0.6% from Physical, according to the most recent Pró-Música report.[2]

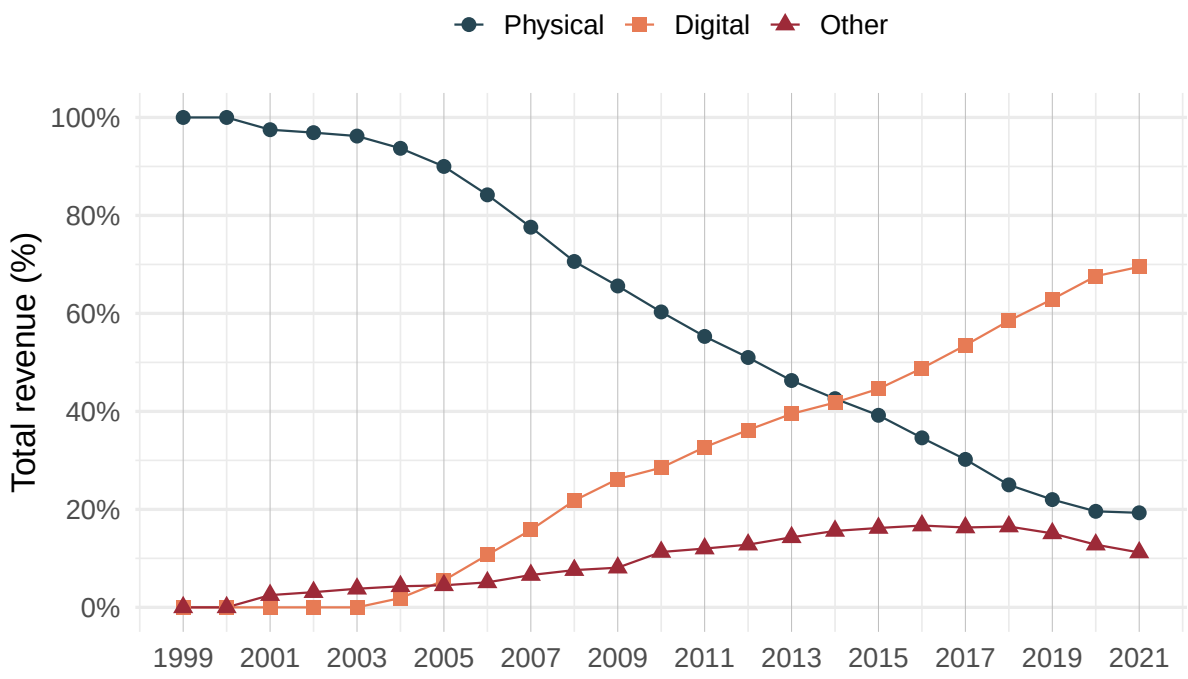
Although streaming platforms are inherently designed to not interfere with the music production process, their leading role in the music industry is unquestionable: they determine the amount paid to music content producers, and dictate the type of music accessible through their recommendation algorithms [75]. Such constant changes in the music market reinforce the investigation demand on the implications of these new media players' insertion. Specifically, the Physical to Digital era shift requires attention mainly for MIR, not only on technological-driven factors [61, 75, 127] but also on the success patterns shaping this dynamic market [79, 88, 112].

Digital media dominance has proved beneficial in several ways, such as promoting local artists and increasing listeners' engagement. Also, while musicians struggled with the COVID-19 pandemic in 2020-2021 [104], people have been reminded of the timeless healing power of music, positively impacting listeners' well-being and strengthening the connection between artists and fans. Indeed, with live music on hold and the world in lockdown, most music fans worldwide prefer to consume music through Digital ways (streaming and short video apps).[3] Still, streaming popularization has brought new challenges due to the massive volume of music-related data to process and analyze.

Finding and promoting artists with promising careers is an example of a task that has become more complex and important. In the Physical era, having a major label was essential to breakthrough; but now, artists from small or independent labels can go viral

[1] IFPI Global Music Report: https://gmr2022.ifpi.org/
[2] Pró-Música Brasil Report: https://bit.ly/ProMusica2021
[3] IFPI Engaging with Music report, 2021: https://bit.ly/IFPI-engaging-2021M

Figure 3.1: Global recorded music revenues by segment (1999–2020)



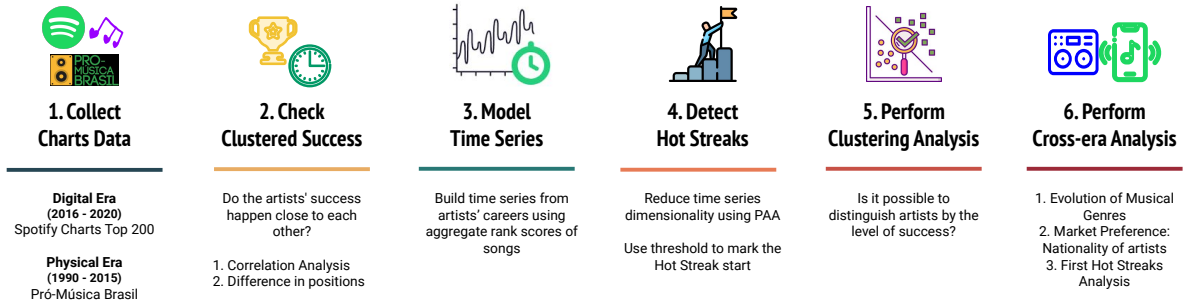
Source: The Author.

and become popular thanks to streaming services, showing just how inherently dynamic the music industry is. In such a context, combining Artificial Intelligence (AI) tools and techniques may be the key to facing the existing challenges of this task, creating significant benefits for both the artists and the A&R executives. In fact, many applications use AI-powered technology in the music industry, such as genre classification [109] and success prediction [8, 79]. Regarding the latter, identifying upcoming artists with outstanding success is crucial, as it helps planning and adjusting marketing directions for their careers.

Generally, musical careers present continuous periods of success above average, defined as hot streaks. The concept has been investigated in many domains, including science [114], social media [44], and creative careers [55, 72, 73]. In such a context, we explore over three decades in the Brazilian music market, assessing the evolution of successful careers by comparing data from the Physical (1990–2015) and Digital (2016–2020) eras. In particular, to answer RQ2 (*How does the evolution of music consumption in Brazil affect the occurrence of hot streaks?*), we build artists' success time series based on sales (Physical era) and streams (Digital era). Based on such time series, we investigate whether the most successful periods in an artist's career occur chronologically close and detect hot streak periods. Then, we perform a cluster analysis to group artists according to their success level. Finally, we characterize such hot streaks to extract insights into the temporal evolution of musical careers.

We perform a clustered success analysis to investigate if the success is grouped in

Figure 3.2: Methodology overview for identifying Hot Streaks and the success levels from musical artistic careers.



Source: The Author.

time for most artists. In addition, we analyze hot streak periods by other dimensions such as artist nationality and the position of the first hot streak. Figure 3.2 resumes the methodology process for this chapter. The remainder of this chapter is organized as follows. First, we briefly discuss related work in Section 3.1.[4] Then, we describe the data acquisition process in Section 3.2. We present a clustered success analysis in Section 3.3. We detail the methodology used to identify hot streaks in Section 3.4. We overview the identified clusters and hot streak analyses in Section 3.5. Next, we further enrich the cross-era comparison by including the evolution of genre preference, artist nationality preference in the Brazilian Market, and the first hot streak analysis in Section 3.6. Finally, we present concluding remarks in Section 3.8.

3.1 Overall Scenario for Hot Streaks in Music

Although streaming platforms are inherently designed to not interfere with the music production process, their leading role in the music industry is unquestionable: they determine the amount paid to music content producers, and dictate the type of music accessible through their recommendation algorithms [75]. Such constant changes in the music market reinforce the investigation demand on the implications of these new media players' insertion. Specifically, the Physical to Digital era shift requires attention mainly for MIR, not only on technological-driven factors [61, 75, 127] but also on the success patterns shaping this dynamic market [79, 88, 112].

After decades of intense transformations in the music market, the Digital era brought novel challenges, including a substantial volume of data. As human inspec-

[4]Although Chapter 2 goes over related work considering the whole thesis, this chapter once again overviews the specific scenario that motivates this chapter and summarizes related work on hot streaks.

tion is almost impossible for music big data scale, specialized algorithms can help with several tasks in MIR, including music recommendation [19], automatic genre classification [25, 30, 109], algorithmic composition [53] and so on. Another possible benefit is to feed machine-learning models for musical success early prediction, contributing to identify trends and new talent. Indeed, evaluating the impact of human performance is a common practice in many research fields [48, 94, 95]. The term Hot Streak emerges in such context, as the reference to a specific period within professional careers when the success is significantly higher than the average [73].

For individual and creative careers, research assessing impact is much more recent. Liu et al. [73] consider large-scale careers of artists, film directors and scientists to demonstrate that hot streaks are remarkably universal across diverse domains, yet usually unique across different careers. In this sense, Garimella and West [44] use data from Twitter, one of the most popular online social networks, and define users' impact as the reach of their content. Janosov et al. [55] also consider luck as a crucial ingredient to achieve impact in creative domains. Regarding music, they model the historical artist timelines based on the release year of songs and measure success by the total play counts obtained from Last.fm.

Nonetheless, to the best of our knowledge, no previous studies address the dynamics of music artists' success periods (i.e., hot streaks) within the Brazilian market. Also, although Brazil's high rates of music consumption, little is known about the key factors driving musical success and defining artists' promising careers. As regional markets have their own success patterns and behavior [31, 88], such individual analyses are crucial. Therefore, this work is a step forward towards understanding the specific dynamics of music artist success within the Brazilian market.

3.2 Physical and Digital Music Data Acquisition

To perform a cross-era comparative analysis between Physical and Digital media, we focus on musical success in Brazil. Our first data source is Spotify, the most popular global audio streaming service. However, its Charts only comprise data from 2016 onwards. Hence, to describe the Digital Era, we consider the range period available (2016–2020). We also use the Pró-Música Brasil platform to describe the Physical Era, with data from 1990 to 2015. Next, we detail the data acquisition processes for both Physical (Section 3.2.1) and Digital media (Section 3.2.2).

Table 3.1: Pro-Música Brasil certification levels for Brazilian and foreign artists (A&S is Albums and Singles).

	Brazilian		Foreign	
Certification	**A&S**	**DVDs**	**A&S**	**DVDs**
Gold	40,000	25,000	20,000	15,000
Platinum	80,000	50,000	40,000	30,000
Double Platinum	160,000	100,000	80,000	60,000
Triple Platinum	240,000	150,000	120,000	90,000
Diamond	300,000	250,000	160,000	125,000
Double Diamond	600,000	500,000	320,000	250,000
Triple Diamond	900,000	750,000	480,000	375,000
Quadruple Diamond	-	1,000,000	-	500,000
Quintuple Diamond	-	1,250,000	-	625,000

Source: The Author.

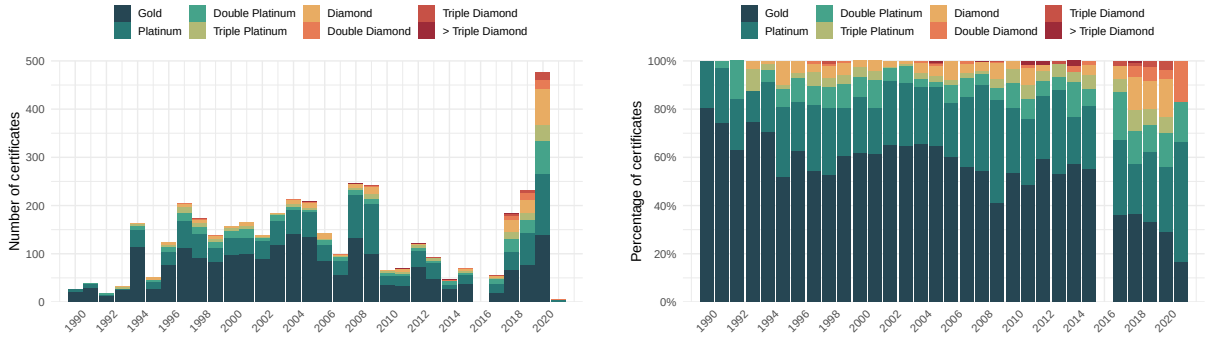
3.2.1 Physical Media

Pró-Música Brasil (PMB) is the official representative body of the record labels in the Brazilian phonographic market. It represents artists in legal and financial instances and issues certification awards, as authorized by record companies. The certification awards recognize the work of performers according to sale numbers in the form of "special discs", i.e., Gold, Platinum and Diamond discs. The data on such awards is available on its website[5] and was collected on February 5th, 2021. The final dataset comprises information on awarded artists, release year, disc category, song/album name and media type since 1990.

In PMB, the threshold sales number for each certificate depends on whether the artist is Brazilian or not, as shown in Table 3.1. However, as such information is not available in PMB, we crawled it from Wikipedia using a Python library[6]. Next, we collect the total sales for each musical work based on the certification awarded, nationality, and PMB's sales metric for the disc award. Finally, we use Spotify's API[7] to associate each artist with their respective genres specified on the streaming platform. Hence, our final dataset contains information about 4,198 musical works from 780 artists. Considering only the period between 1990 and 2015 (i.e., Physical Era), there are 3,243 musical works from 574 artists. Quantitative information on the certificates is shown in Figure 3.3.

[5]PMB Certificates: https://bit.ly/CertificatesPMB
[6]Wikipedia Python Library: https://github.com/goldsmith/Wikipedia
[7]Spotify API: https://developer.spotify.com/

Figure 3.3: Number of discs certificated (left) and its respective percentage (right) in Pró-Música Brasil between 1990–2021. In 2016, there was a metric change in the certification, hence the lack of data.



Source: The Author.

3.2.2 Digital Media

Between 2016 and 2017, there was a crucial change in PMB's metric, which moved from Physical media (i.e., DVD and CD) to Digital media (i.e., Singles and Albums), as depicted in Figure 3.4. Meanwhile, streaming was already the primary revenue source for Digital media (58.3%).[8]

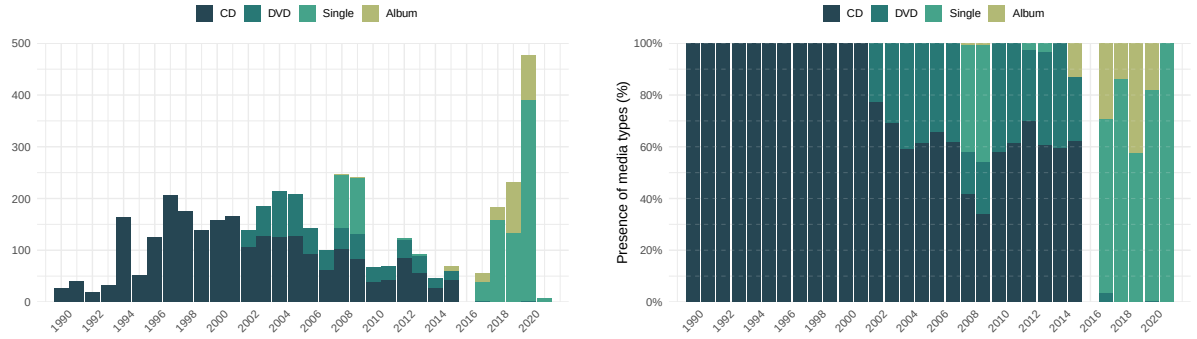
Given its relevance, we extract data referring to the Digital Era from the weekly Spotify Top 200 Chart, which corresponds to the most streamed songs in Brazil. Each chart entry contains the song's name and its artist(s), the number of streams, the song's Spotify URL and its position on the chart. We collect data from January 2017 to December 2020. We also collect artist data using the Spotify API: name, number of followers, and their list of genres. Our final dataset for Digital Era comprises 2,595 songs from 1,018 artists obtained from 108 weekly charts.[9]

3.3 Clustered Success

In this section, we present an analysis of the distribution of success over time. We follow the methodology used by [44] to investigate whether the most successful periods of each era (years or weeks) occur close to each other in artists' careers. We define artists' career as their success time series from their debut on charts to the last date of our

[8]Pró-Musica Brasil 2016: https://bit.ly/ProMusica2016
[9]MUHSIC-BR dataset available at: https://doi.org/10.5281/zenodo.5591015

Figure 3.4: Number of media type in Pró-Música Brasil (left) and its respective percentage (right) between 1990–2021. In 2016, there was a metric change in the certification, hence the lack of data.



Source: The Author.

collection (2015 for Physical Era, and December 2020 for Digital Era). To the Physical Era, we set the position $P(y_i)$ of a year y_i within a time series as its index i. The k most successful years (i.e., with the highest sales values) are denoted by Y_1, Y_2, ..., Y_k. Likewise, in Digital Era, we set the position $P(w_i)$ of a week w_i within a time series as its index i. Here, the k most successful weeks (i.e., with the highest stream values) are denoted by W_1, W_2, ..., W_k.

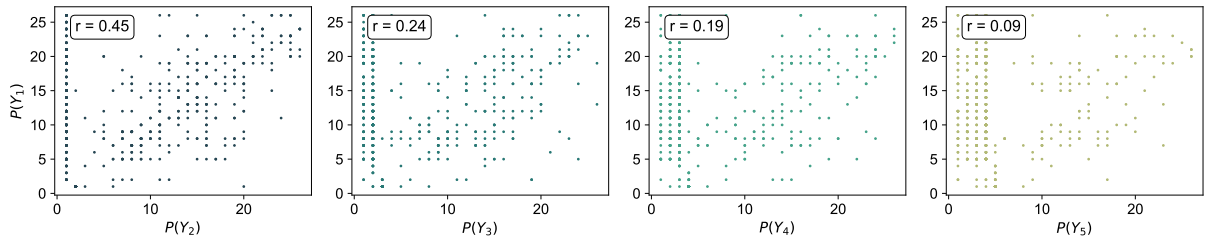
Our analyses focus on two main points. First, we investigate the timing of the most successful periods of an artist's career for both eras. Then, we look at the distribution of the difference between the positions of the two most successful periods within artists' careers. Such analyses are all made compared to shuffled careers to check the robustness of our findings, that is, if the observed effects still happen.

3.3.1 Timing of the most impactful periods

First, we analyze the positions of the five most successful years (Physical Era) and the five most successful weeks (Digital Era) within artists' careers. Figure 3.5 shows scatter plots of the first year of Physical Era $P(Y_1)$ versus the other years $P(Y_i)$ $i \in [2, 5]$, and the Pearson correlation coefficient (r) for each plot–correlation values are statistically significant ($p < 0.05$). We consider all artists from our dataset. The results show linear correlation for all artists' careers and higher Pearson coefficient when comparing the first and second most popular years. Compared with the third, fourth and fifth years, the correlation decreases. Such a finding reinforces the hypothesis that the most impacting years within an artist's career are more likely to happen close to each other.

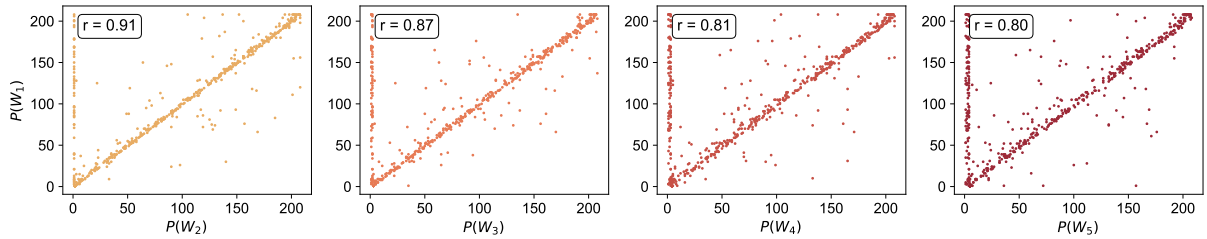
Regarding the Digital Era, Figure 3.6 presents scatter plots of the first week $P(W_1)$

Figure 3.5: Scatter plots with Pearson correlation (r) of the position of the most successful year (Physical Era) in artist careers (Y_1) with Y_2, Y_3, Y_4, and Y_5, respectively. Each point represents an artist. All correlation values are statistically significant ($p < 0.05$).



Source: The Author.

Figure 3.6: Scatter plots with Pearson correlation (r) of the position of the most successful week (Digital Era) in artist careers (W_1) with W_2, W_3, W_4, and W_5, respectively. Each point represents an artist. All correlation values are statistically significant ($p < 0.05$).

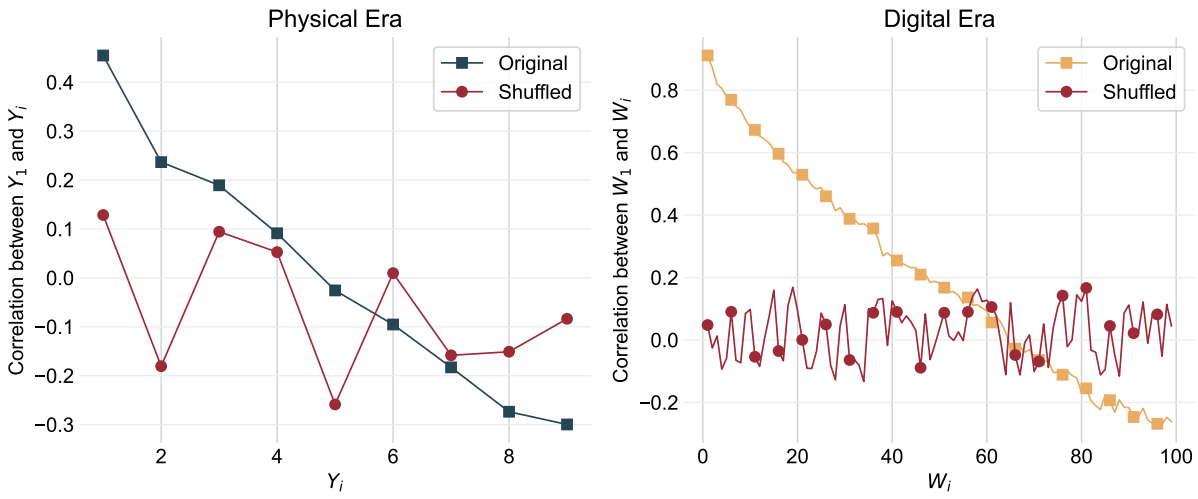


Source: The Author.

versus the other considered weeks $P(W_i)$ $i \in [2, 5]$, and the Pearson correlation coefficient (r) for each plot–correlation values are statistically significant ($p < 0.05$). Here, we also consider the totality of artists from our dataset. In contrast to the Physical Era, the results show clearer and stronger linear correlation for all artists'careers. Likewise, there is a higher Pearson coefficient concerning the first and second most popular weeks. However, although the third, fourth and fifth weeks decrease their correlation values, their value remains high. These results are different from those found for the Physical Era, and the yearly granularity of the Physical data may directly influence them. Nevertheless, it is still possible to verify that the first two weeks are more correlated than others, in both cases, reinforcing the premise that artists' most successful periods group in time.

We then expand the correlation study to compare the positions of successive periods. Figure 3.7 shows a decrease in the correlation in all artists' careers for both eras. Still, this pattern is not observed in shuffled careers, in which the correlation is always between -0.1 and 0.1 for the Physical Era, and -0.2 and 0.2 for Digital Era. Therefore, there is a general trend of clustering within the most successful periods (years or weeks) in artist careers, as they tend to happen close to each other in the success time series.

Figure 3.7: Correlation between the first and $i-th$ most successful years to the Physical Era (left), and the $i-th$ most successful weeks to the Digital era (right).



Source: The Author.

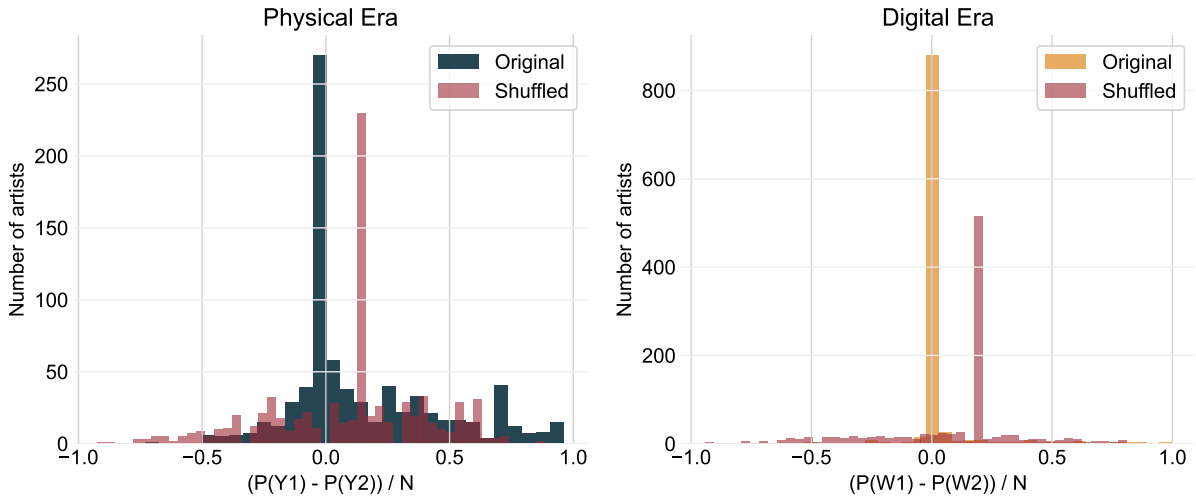
3.3.2 Difference in positions of the most successful periods

In this section, we complement the previous analyses by taking each artist's first and second most successful periods to calculate their differences in positions and verify if they happen near each other. For instance, if each artist's most successful periods are close together (years 2 and 3 for one group of artists, and years 5 and 6 for another), subtracting their positions results in 1; the final result is close to zero when normalizing such values; i.e., if such position differences have the value of 1 for most artists, their success occurs in a consecutive period, or they are grouped in time.

To calculate the difference in the positions of the top two most successful periods for the artist's careers, we consider $P(Y_1)$ and $P(Y_2)$ for the Physical Era and $P(W_1)$ and $P(W_2)$ for Digital Era. We normalize such a difference by the number N of years/weeks of the artist's time series in the corresponding era. Figure 3.8 shows the distribution has a peak around zero for all artists'careers, suggesting that these two periods (years/weeks) are close to each other on the timeline.

Note the results are similar for both eras. Such outcome agrees with the findings of the previous analyses. Further, when we shuffle artists'careers, the distribution of these differences is much distinct from the original, demonstrating that this behavior of musical careers is not random. Hence, there is strong evidence that artists may experience periods of outstanding success, or hot streaks, which we investigate in the next section.

Figure 3.8: The normalized difference between the positions of the first and second most successful periods within artists' careers. Years for Physical Era (right), and weeks for Digital Era (right).



Source: The Author.

3.4 Hot Streak Detection

To detect hot streaks in artist's time series, we rely on previous work that shows the most successful points in professional careers tend to happen close to each other [44]. Hence, we use a technique to reduce the time series dimensionality to continuous delimited periods within careers. Then, we define a hot streak as the periods in which the success (i.e., Physical sales or Digital streams) is above a certain threshold obtained from the career itself. In other words, the hot streak detection does not consider external factors (e.g., genre and time) because artists reach different levels of success, and choosing a single threshold would make the comparison unfair.

To reduce time series dimensionality, we use Piecewise Aggregate Approximation (PAA) [58]. Given a time series $X = x_1, x_2, \cdots, x_n$ of length n, PAA reduces it into a new series $\overline{X} = \overline{x_1}, \overline{x_2}, \cdots, \overline{x_N}$ with N dimensions, $1 \leq N \leq n$. The intuition is that dividing the original time series into N equal-sized segments produces N new points. The value of each segment is defined as the average of the points within such a frame (Equation 3.1). Hence, the approximation of each point on the original time series is made by simply assigning the PAA value of its corresponding segment.

$$\overline{x_i} = \frac{n}{N} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j \tag{3.1}$$

Note artists' careers may contain points with extreme values for the success metric. Therefore, PAA is a helpful tool to smooth such differences and delimit periods in the

careers. Regarding code, we use the PAA implementation of tslearn [118], a Python package for time series analysis. Its only parameter is the number of segments to split the series into (further information on values next).

Finally, we chose a specific threshold for defining the hot streak periods for each artist. Such an individualized approach is based on the percentiles of the success metric, and it allows analyzing the careers of artists with different levels of success. In other words, as success is relative for each artist, we detect HS for widely known artists with higher sales and streams, as well as independent artists who have received only a few certificates and streams.

3.5 Clustering Analysis

Here, we analyze both Physical and Digital Eras. In Section 3.5.1, we use the PMB data and Spotify's Brazil Top 200 Charts for building the time series for each artist, respectively, to such eras. Next, we characterize the hot streaks for both eras and understand their relationship to music genres in Section 3.5.2. Finally, we perform a cluster analysis to group similarly artists based on their success levels in Section 3.5.3.

3.5.1 Artists' Time Series

In the Physical Era, the evolution of an artist's success is represented by the certificates received at PMB from 1990 to 2015. We build the time series with an annual granularity because PMB provides data for the year the artist got the certificate. Hence, each point in the artist's time series corresponds to the number of sales achieved in that year. On the other hand, we use Brazil's Spotify Top 200 Chart in the Digital Era as our basis to model artists' success over time. For all artists, each point in their time series represents the accumulated success in a given week, according to the chart. In this case, the success measure considered is the total number of streams (i.e., the number of times the song was listened to on Spotify) per week. For example, Figure 3.9 presents the time series of Sandy & Junior. Their last album as a band was Acústico MTV: Sandy & Junior, released in 2007. Nevertheless, they continued to sell many records even after two years of the duo's disbandment.

From the artists' time series, we detect the hot streaks periods by first applying

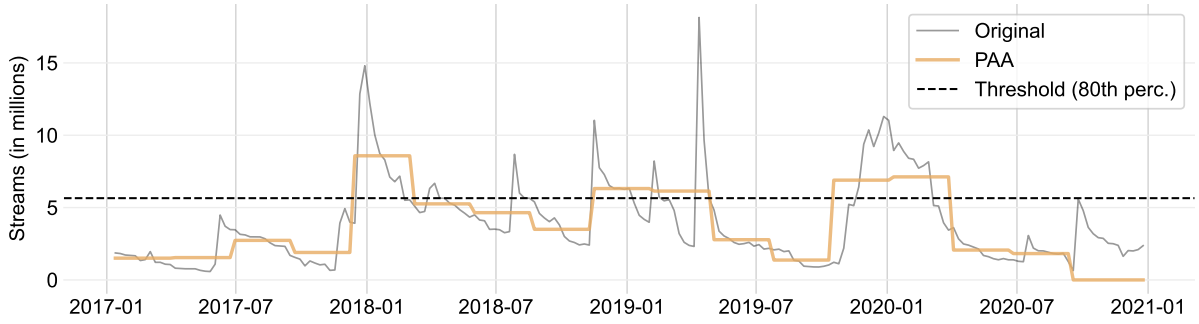
Figure 3.9: Sandy & Junior's success time series in the Physical Era (1990–2015).

Source: The Author.

PAA. To do so, we set the number of segments in which the series will be split, as this is the only parameter of the method. For Physical era, since we deal with yearly success time series, the minimum size of each segment must be two years. After extensive empirical experiments, we set the window to the minimum, which is enough to validate a hot streak. On the other hand, the digital nature of streaming platforms allows for successful weekly data to be made available almost in real-time. Hence, according to our experiments, we define 12 weeks as the size of each PAA segment for Digital era. Each segment comprises a three-month period, which is a reasonable time to analyze the continuous periods of great success in streaming platforms such as Spotify. Hence, we calculate the number of segments by dividing the time series length by the predefined size. In addition, we set the 80th percentile of the success metric in artists'time series as the threshold for defining the hot streak periods.

Figure 3.10 illustrates PAA applied to Anitta's career, currently one of the most influential Brazilian artists worldwide. On March 24, 2022, Anitta broke the record as the first Brazilian female artist to achieve the first position on Spotify Top 50 Global Chart with her single *Envolver*. Recently, Anitta has been nominated for the second consecutive year at MTV's Video Music Awards (VMA) in the Best Latin Music Video category. At the VMA 2023, the artist competed with the clip for *Funk Rave*, a song released in June as the first sample of the singer's sixth studio album. Being present in all considered weekly charts, there are three HS in her time series.The first HS is from November 2017 to April 2018, the period when she released *Vai Malandra*, which became the most streamed song on its release date.The second HS period coincides with *Veneno* and *Não Perco Meu Tempo* single releases. Finally, the third one comprises the period in which she collaborated with famous national and international artists, such as Marília Mendonça (*Some Que Ele Vem Atrás*) and Black Eyed Peas (*Explosion*).

Figure 3.10: Piecewise Aggregate Approximation (PAA) applied to Anitta's success time series in the Digital Era (2017–2020). Periods above the threshold are considered hot streaks.



Source: The Author.

Table 3.2: Main statistics on Hot Streaks grouped by Genres.

Genre	Number of Hot Streaks		Proportion of Artists (%)		Median of Hot Streaks		Max of Hot Streaks	
	Physical	Digital	Physical	Digital	Physical (year)	Digital (week)	Physical (year)	Digital (week)
Rock	164	26	23.90	4.97	2	18	8	48
Axé	162	31	28.80	5.96	2	24	8	48
Pop	131	219	24.60	41.70	2	24	8	60
Sertanejo	64	95	11.20	18.90	2	24	10	48
Gospel	57	4	9.83	1.00	2	12	8	36
Rap	4	20	0.70	3.98	4	24	4	36
Funk Carioca	3	104	0.70	22.60	2	24	4	48
Forró	2	5	0.23	1.00	2	18	2	24

Source: The Author.

3.5.2 Hot Streak Characterization

We characterize the hot streak periods identified for artists according to their musical genres. As individually considering closely related music styles may create artist overlapping and bias within the results, we define super-genres for this analysis. For example, we verify that Indie Folk is more frequently associated with Rock than any other super-genre and is then incorporated into Rock. Finally, the top eight prominent super-genres considered in Brazilian music are Rock, Axé, Pop, Sertanejo, Gospel, Rap, Funk Carioca and Forró.

Table 3.2 assesses each genre's performance regarding the number of hot streaks as well as the proportion of artists who have achieved Hot Streaks by genre, and the values

of median and maximum duration of a hot streak, respectively. All genres analyzed in the Physical Era have at least two hot streak periods. The genres of artists that concentrate higher proportions are Axé (∼29%), Pop (∼25%) and Rock (∼24%). Although this pattern happens in such genres, there is no clear correlation between genre and the number of HS. As we consider a yearly time window, achieving a high number of hot streaks is not an easy task. Regarding the number of hot streaks in the Digital Era, the genres have at least four Hot Streaks. Although all genres follow a similar trend in general, Pop presents a higher percentage of artists that have achieved more HS, around 42%. In contrast, all Gospel and Forró artists have only a 1% of presence in HS.

Next, we analyze the duration of hot streaks. We consider the median and the longest HS for each artist, as there may be more than one (HS size is always a multiple of two for Physical Era because PAA segment is set to two years). Overall, all genres have a median number of Hot Streaks of two, except for Rap, with four Hot Streaks. Nevertheless, Axé, Gospel, Pop and Rock get a remarkable maximum of eight years of Hot Streak, and Sertanejo reaches 10 years of success.

In the Digital Era, about half of the artists from all genres have 12-month long hot streaks. However, there are specific hot streak patterns when analyzing the genres individually. For instance, most Gospel artists have shorter HS periods, as the median time presents a 12-week hot streak. On the other hand, genres such as Axé, Funk Carioca, Pop, Rap and Sertanejo artists have longer hot streaks, as they have median HS periods of up to 24 weeks, i.e., six months. Still, only Pop achieves a maximum duration of Hot Streak of 60 weeks. Therefore, in contrast to the previous analysis, we note genre is relevant to describing hot streaks' longevity.

3.5.3 Cluster Analysis

We now move to the cluster analysis, which helps to better understand the characteristics of different success levels of artists achieved during the Physical and Digital Eras. We apply the K-Means algorithm in the time series and the Elbow method to find its optimal number of clusters [13]. We use the K-Means algorithm, which is the most commonly used clustering method for dividing a dataset into a set of k groups. The considered features for the algorithm include the total number of hot streaks, total sales and the time series threshold. As a result, the method outcome suggests three clusters. We name the resulting clusters according to the success metric (i.e., number of sales and streams): SHA, BHA, and THA. The main statistics of the clusters in both eras are presented in Table 3.3 and summarized as follows.

Table 3.3: Main statistics on the artist clusters in the Physical and Digital Eras.

	All		**SHA**		**BHA**		**THA**	
	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital
Number of artists	574	1,018	527	940	38	70	9	8
Average number of HS	1.3	0.6	1.3	0.5	1.8	1.6	1.9	1.6
Median sales (10^4) / Streaming (10^4)	8.5	7.6	8	3.9	152	2,190	507	10,857
Median threshold	0	0	0	0	72,500	2,052,131	300,000	7,145,116

Source: The Author.

Spike Hit Artists (SHA). This cluster contains most artists in Physical (527) and Digital (940) Eras. The median PAA threshold for such artists is 0, indicating that their sales and streams are, in general, much lower than the artists from other clusters. In addition, the average number of hot streaks for SHA is 1.3 and 0.5 for Physical and Digital Eras, respectively. To the former, the result suggests that PMB certificates happen sparsely for such artists because there are few hot streaks even with a low threshold. Artists in this cluster include Banda Eva, Claudinho & Bochecha, Coldplay, Paramore, and Latino. To the latter, we cannot say that such artists were not successful, as they are on Spotify's Top Charts. Still, they do not have many hot streaks (one HS on average). Artists in this cluster include Billie Eilish, Leo Santana, and Naiara Azevedo.

Big Hit Artists (BHA). This cluster is a bridge between the most and less successful ones, with 38 artists in the Physical era and 70 artists in the Digital era. Although BHA presents a higher number of sales when compared to SHA, its artists do not present a proportional increase in the average number of hot streaks. In fact, they have, on average, 1.8 hot streaks between 1990 and 2015 in the CD era and 1.6 hot streaks between 2016 and 2020 in Streaming era. Also, there is a substantial increase in the threshold for artists in this cluster, reaching 72,500 sales and more than 2M streams in respective eras. Madonna, Legião Urbana, Skank, and U2 are examples of Physical BHA, and Barões da Pisadinha, Dua Lipa, and Pabllo Vittar are of Digital BHA.

Top Hit Artists (THA). This cluster has most successful artists, as they have the highest median number of sales in both scenarios. The average number of hot streaks for THA is very close to the value for BHA. However, the artists in this cluster have achieved major sales success throughout the Physical Era, as they have a very high median threshold (300,000 sales). Similar, the number of Spotify streams is much higher when compared to the previous groups, which is also observed with the threshold. Therefore, Top Hit Artists in Digital Era may be considered highly successful artists, as their songs achieve a higher stream count throughout the weeks. Examples of Physical THA include Ivete Sangalo, Zeca Pagodinho, Roberto Carlos, and Sandy &Junior. All Digital THA are Brazilian, and examples include Anitta, Marília Mendonça, and Zé Neto & Cristiano.

3.6 Cross-era Comparison

Music is part of people's daily lives regardless of the era experienced, whether Physical or Digital. With musical consumption constantly rising, we may notice similarities between both eras. Here, we explore Hot Streak (HS) in musical careers within the Brazilian market. Such HS periods provide valuable information used in cluster analysis, in which we also notice cross-era similarities. Specifically, we discuss Brazilian listeners'main musical genre preferences in Section 3.6.1. Next, we explore the presence of Brazilian artists as the preferred consumption in the local market in Section 3.6.2. Finally, we explore when the first Hot Streaks occurs in the artists' careers in Section 3.6.3.

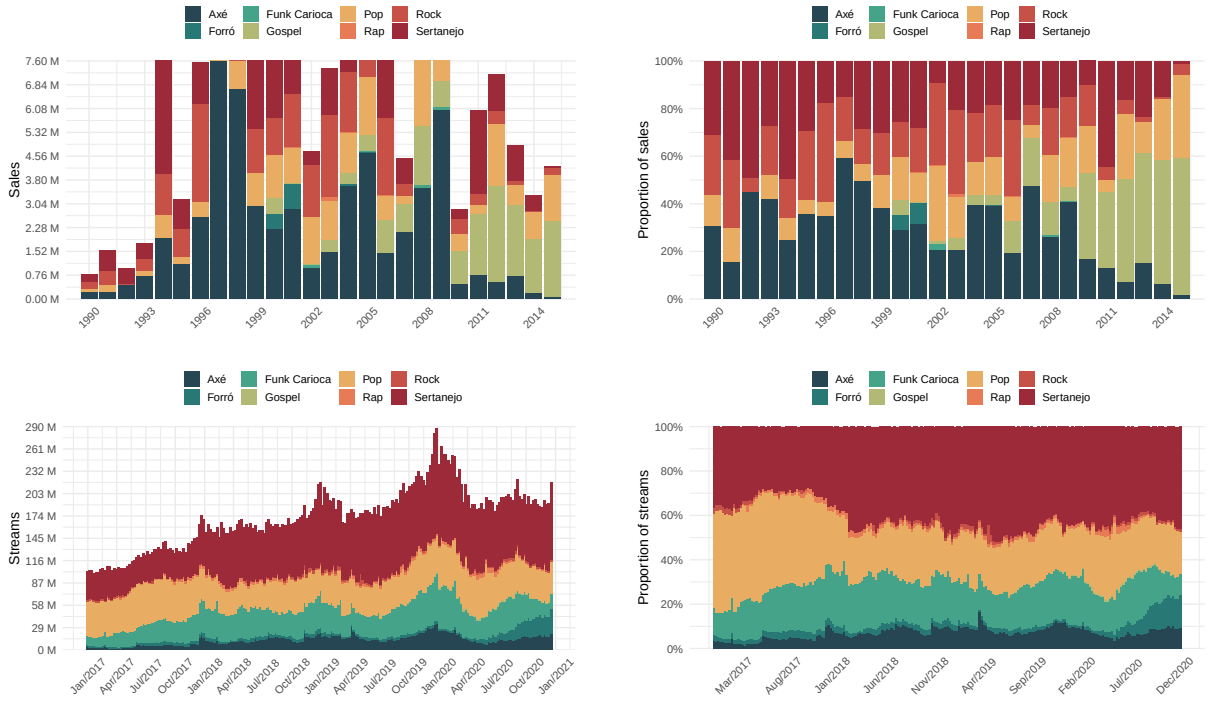
3.6.1 Genre Evolution

Musical genres express the cultural diversity existing in the country. Such diversity can be observed by the number of rhythms and musical styles as well as the specific characteristics that each of them retains. The most popular genres usually oscillate in the music market. Several factors may influence this issue, including the help provided by streaming platforms to spread cultural diversity worldwide in the form of musical genres. Hence, we analyze the temporal evolution of consumption of the main genres in the Brazilian market, both for the Physical and Digital eras.

Figure 3.11 shows the transformation in the listeners' preference regarding musical genre. For instance, there was an increase in Gospel sales in the Physical Era, but not in the Digital one. A potential reason is Gospel listeners still consume Physical media by 2015, whereas audiences from other genres had already migrated to streaming (until 2016, the PMB methodology still favored Physical sales). However, the transition of preference for musical genres over the years is notorious: in the Physical Era, the predominant rhythms were Axé (e.g., Ivete Sangalo), Sertanejo (e.g., Zezé di Camargo & Luciano) and Rock (e.g., Skank); whereas in the Digital Era, the most successful artists (THA) come from one style, Sertanejo, with more than 50% of streams in late 2020.

Overall, the Digital Era allows the appearance of new popular genres, as well as the decline of previously popular ones. For example, the prevalence of Sertanejo is remarkable over time, while Pop decreases from 2016 to 2020. Moreover, we highlight the rise of Forró in mid-2020 as a well liked genre, following the growth of popular artists who have burst the regional bubble, such as Barões da Pisadinha, Solange Almeida, and Wesley Safadão. Such a significant boost for regional artists may have been enhanced by the remarkable

Figure 3.11: Genre evolution in (top) Physical and (bottom) Digital Eras.



Source: The Author.

lives of Forró and Sertanejo artists during the COVID-19 pandemic, showing the music industry's ability to adapt. In fact, Marília Mendonça had the most-streamed YouTube live worldwide in 2020, with over 3.31 million viewers.

3.6.2 Brazilian Artists vs. Foreign Artists

Record companies have been actively working to promote local artists aiming to expand the music ecosystem. Furthermore, according to the IFPI 2022 report, fans are listening to more local artists than ever before, and their music also has the power to go global from day one. As such, the music industry needs to invest in discovering and nurturing the artists of tomorrow. There is a strong predominance of local artists' consumption in Brazil compared to other important countries, such as the USA and European countries.[10] Therefore, our purpose is to identify whether the Brazilian market follows this trend of consuming local artists in the Physical and Digital eras. Thus, we analyze the distribution of consumption of Brazilian and foreign artists in each period. Finally, we identify the representativeness of Brazilian and foreign artists in each cluster

[10]Brasileiros são os que mais ouvem a própria música entre todos os países. Link: https://bit.ly/folha-sp-musica. Access: March 26, 2022.

Figure 3.12: Comparison of the sales evolution by Brazilian artists versus foreign artists in the Physical Era in absolute values (left) and percentage values (right).

Source: The Author.

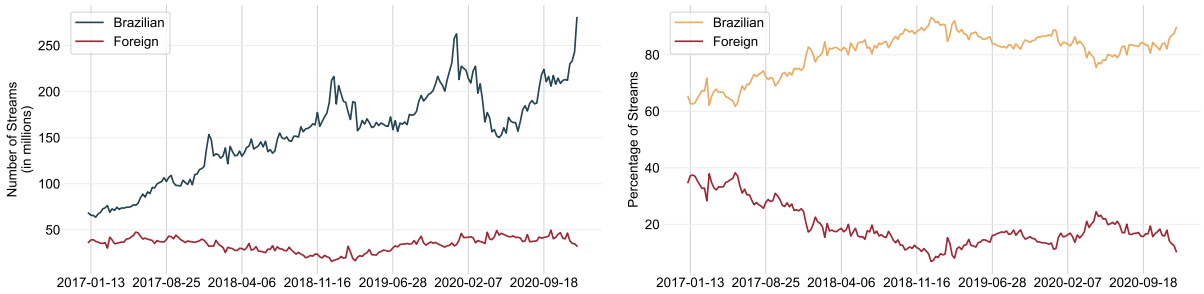
identified in the previous analyses, also separated by era.

In the Physical Era market, we notice a strong sales presence of artists whose nationality is Brazilian. Figure 3.12 shows the sales evolution in the Brazilian market, comparing sales of local artists versus foreign artists. It shows the number of sales in millions on the left, and the percentage of total sales on the right, for both Brazilians and foreign artists. There is a constant evolution of the preference for Brazilian artists, while there is a tendency for stability in foreign artists. Indeed, the sales representativeness by Brazilian artists corresponds to around 90% throughout the entire period of the Physical era. Only around 2012, there was a slight drop in consumption of local artists, but sales still represented more than 70% in such a period. Hence, we show that Brazilians do indeed favor consuming local musical artists.

Concerning the Digital Era, Figure 3.13 compares the streaming evolution by Brazilian and foreign artists by the number of streams (left) and its corresponding percentage (right). The Brazilian market in Digital Era has its beginning marked by a lower preference for foreign artists when compared to the Physical Era. In such a scenario, the sales of foreign artists in the Brazilian market reached around 40% of total streaming. One possible explanation is that different artists and genres spread more quickly worldwide with the advent of streaming platforms. However, there is a clear trend of growth in the consumption of Brazilian artists, reaching higher levels in 2020 and revealing a movement of a more significant decline in the consumption of foreign artists.

Finally, we also analyze the distribution of the presence of Brazilian artists in each of the clusters for the Physical and Digital eras identified in Section 3.5. Table 3.4 summarizes the number of Brazilian and foreign artists by era and cluster. In general, there is a strong trend in local music consumption in the Brazilian market. We highlight BHA and THA clusters, which comprise the most successful artists in both eras (i.e., higher sales and streams), including Sandy & Júnior and Anitta. In particular, the second cluster represents the paramount artists. Although the Physical era has one more foreign artist (five) than Brazilian (four), in the Digital era, all artists are Brazilian, indicating a strong preference for local artists and genres. As a result, the SHA cluster indicates

Figure 3.13: Comparison of the streams evolution by Brazilian artists versus foreign artists in the Digital Era in absolute values (left) and percentage values (right).



Source: The Author.

regular success, accounting for over 90% of the artists.

Table 3.4: Number of Brazilian and foreign artists by cluster by era. In general, there is a predominance of Brazilian artists in all groups.

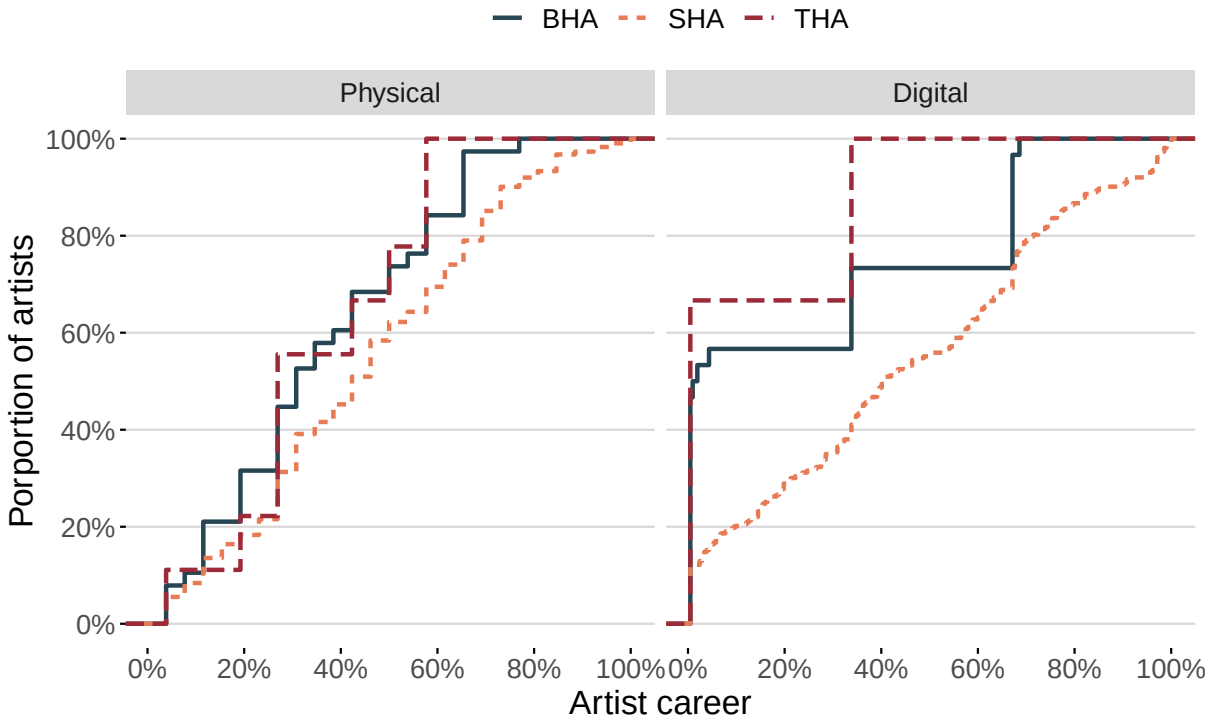
	All		**SHA**		**BHA**		**THA**	
	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital
Brazilian	366	587	342	522	20	57	4	8
Foreign	208	431	185	418	18	13	5	0
TOTAL	**574**	**1,018**	**527**	**940**	**38**	**70**	**9**	**8**

Source: The Author.

3.6.3 First Hot Streak Analysis

The last comparative aspect is the position of hot streak periods within artist careers. Previous studies on other domains show such periods are temporally localized and happen at any point in an individual's works sequence. Thus, to assess whether there is a similar behavior in the music domain, in both Physical and Digital eras, we investigate in which point artists experience their first stardom. Here, we focus only on the first point, as we detect more than one hot streak for several artists. Figure 3.14 shows the cumulative distribution of the position of the first hot streak within artist careers in both Physical (left) and Digital (right) eras, grouped by cluster. In the Physical era, in general, the cumulative distribution of the position is very similar among the three identified clusters, probably due to the data granularity (i.e., considered in years and not weeks). Nonetheless, for the BHA and THA clusters, the first explosion of success still occurs faster than the artists of the SHA cluster, considering the artists' career time.

Figure 3.14: Cumulative Distribution Function (CDF) of the position of the first hot streak within artist careers in both Physical and Digital eras, grouped by clusters. Artist timelines are described in percentages, in which 0% represent the debut week and 100% is the last year/week collected in our dataset.

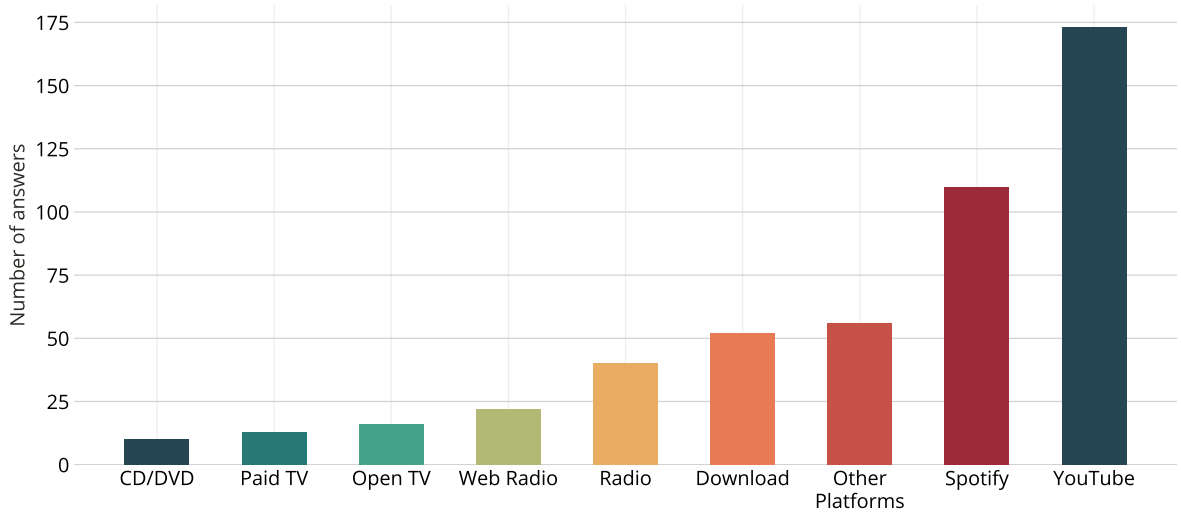


Source: The Author.

On the other hand, regarding the Digital era, the two clusters that group artists with the highest success levels (i.e., BHA and THA) stand out much more in comparison to the SHA cluster. Almost 80% of the BHA/THA artists have their first burst of success early in their careers (i.e., in the first 30% of their timelines), whereas most SHA artists reach their first hot streak much later.

Therefore, our results indicate artists who have achieved their first stardom peak earlier in their careers have a higher overall success, regardless of the musical era. It is important to note artists may have careers of different sizes depending on their debut date, as the last date in the time series is always the same (i.e., the collection date). However, the main objective of this analysis is to offer a big-picture comparison between the Physical and Digital eras, seeking to highlight the relationship between artists' levels of success and the speed of achieving the first burst of success.

Figure 3.15: Distribution of platform preference among gospel music consumers.



Source: The Author.

3.7 Cross-Era Discussion

Overall, the digital age has brought a series of changes in music consumption concerning the physical age. Physical media, such as CDs and vinyl, were the main form of access to music, and artists' income depended directly on sales of these products, in addition to other sources of income, such as ticket sales and radio audience. After all, in the physical era, as the name implies, music consumption was mainly through physical media. Then, a natural question arises from such a scenario: *Is it possible to use all the knowledge of the physical age to apply and improve models in the digital age?* Such question is intriguing, and it would be unfair to leave aside the dynamism that the internet and streaming services have added to the equation.

A practical example of this change and its respective complex analysis (as seen in Section 3.6.1) is the gospel public. Previously, in the Physical Era, with gospel music consumption predominantly based on physical media and radio, this musical genre experienced increased sales. However, when entering the digital age, there was a retention in its growth. To better contextualize this situation, we surveyed 243 active listeners of the gospel genre.[11] Figure 3.15 shows the distribution of platform preferences of this population. We found that most of these people prefer to use YouTube as their primary platform to consume gospel music. While Spotify has a significant presence, the choice of YouTube (and others) limits the digital landscape analysis when only Spotify is considered a reference platform. Therefore, understanding the preferences of the gospel public

[11] We spread a Google Forms link among listeners of the gospel genre. We ask them, *"What are the most gospel music consumption formats you use? (Select all options that apply)."*

in this digital age requires a more comprehensive approach considering the various music consumption platforms used by this segment. This paradigm shift suggests that specific strategies are required to meet the gospel audience's needs in the digital environment.

Indeed, music consumption has changed significantly with the popularization of the Internet and streaming services. Nowadays, people can access millions of songs through platforms like Spotify, Deezer, and Apple Music, video platforms like YouTube, or even independent media channels like Palco Digital and SoundCloud. From such platforms, artists earn money by streaming their music. Social media and digital marketing tools have also become essential for artists to publicize their work and build a loyal fan base. For example, today's artists need to count lives,[12] clicks (mainly likes) on their social networks, analyze the performance of their branding and sponsored ads, and even calculate the amount of streaming. A recent (and also melancholic) case that the advent of the digital age allows is to keep alive the presence of artists who made a difference in the music scene, even after their deaths. In this scenario, Marília Mendonça continues to be the streaming champion on Spotify, even after almost two years of her departure. She is the first Brazilian singer to reach 10 billion plays on such a platform.[13]

While there are similarities between the two eras, such as the need to create quality, catchy music to attract fans, the differences are significant. Therefore, it is crucial to understand that strategies and tactics that work in the physical age may be less effective in the digital age. In general, music companies and artists must adapt to this new reality and find ways to stand out in an increasingly competitive market. Artists must be more present on social media and use these platforms to build stronger connections with their fans. They must also invest in digital marketing strategies like sponsored ads and email marketing campaigns to reach a wider audience. Another critical point is that the digital age has also brought new opportunities for artists who would otherwise not have the chance to show their work to the world. Streaming platforms such as Spotify itself and SoundCloud[14] allow independent artists to share their music and build a fan base without significant financial investment.

Hence, it is not simple to say that we can take advantage of all the luggage of the physical age to improve the digital age results. However, we may learn from previous experience and routinely improve our intuition about how artists achieve success. Also, Digital Era must defend access to a quality culture so that more artists may have opportunities to publicize their works and achieve their particular success. In summary, although there are similarities between the physical and digital eras of music, it is relevant to recognize the differences and adapt strategies according to the new reality. Artists and companies must be flexible and willing to learn to continually succeed in the digital age.

[12]Marília Mendonça tem live mais vista do mundo: https://bit.ly/MariliaMendoncaRecord2020

[13]Marília Mendonça supera gigantes mundiais e atinge 10 bilhões de plays no Spotify! https://bit.ly/MariliaMendoncaSpotify2023

[14]SoundCloud: https://soundcloud.com/

3.8 Concluding Remarks

This chapter evaluated the success of musical artistic careers in the Brazilian market, comparing them in different eras: Physical Era, when listeners purchase Physical media to engage their favorite artists (e.g., LPs, CDs, and DVDs); and Digital Era, when music consumption took place mainly through streaming, which have democratized access to music, as streaming services do not necessarily require payment (just an internet connection). In this sense, comparing these eras becomes particularly relevant and valuable, as it enables to identify similar or divergent patterns that record companies can use to generate valuable insights during decision making. Thus, we performed a cross-era comparative analysis between the Physical and Digital media in the Brazilian music market, which is the largest market for the music industry in Latin America.

First, we found the artists' most successful periods tend to group in time. Motivated by such results, we built artists' success time series for both eras to identify hot streak periods, defined as continuous high-impact bursts. Next, we characterized such periods to understand the dynamics of success among artists from different musical genres. Although there are similarities among all music styles, our results showed some genres have meaningful specific patterns for both eras. Therefore, as in other studies in the MIR field, considering music genre information can be relevant for both the predictive and descriptive models. We also performed a profiling analysis that uncovered three different clusters in both eras: Spike Hit Artists (SHA), Big Hit Artists (BHA), and Top Hit Artists (THA), which acted as class descriptors of successful artists. In addition, we found artists who have achieved their first stardom peak earlier in their careers have a higher overall success, regardless of the musical era. Finally, we discover Brazilians prefer to consume music by Brazilian artists, and we highlight the proportion of Brazilian and foreign artists within each identified cluster. Such a pattern repeats in both eras.

Overall, our results shed light on meaningful insights for MIR tasks, such as prediction and recommendation. For example, the identified clusters may serve as input features for musical success prediction models. In addition, they may also help in recommending potentially successful musical partnerships and collaborations. Besides helping the scientific community, this work also benefits the music industry. Analyzing the evolution of artist careers reveals success trends in Brazil from what people consume. Indeed, our results demonstrate that although Brazilians connect with international hit songs, they still have a strong preference for local artists regardless of the era. Hence, considering individually regional markets is crucial for better comprehension of specific factors driving musical success. Finally, understanding hot streak periods and success patterns can enhance the human element in the music industry (e.g., A&R executives and record label Chief Executive Officer (CEO)) and people's relationships with music. Our findings may

help describe the listeners' behavior and musical trends, allowing the music industry to connect people to songs relevant to them.

Threats to Validity. Here, a limiting factor is that piracy had a high impact in the music consumption in Brazil, mainly in the late 2000s and early 2010s. Therefore, data collected from PMB may not precisely reflect Brazilian preference in music. In addition, although we found similar patterns between the Physical and Digital eras, each data source used considers its own success measure, which can cause biased results. Finally, we only consider artists who are recognized as successful (either through their sales or their position in stream rankings).

Chapter 4

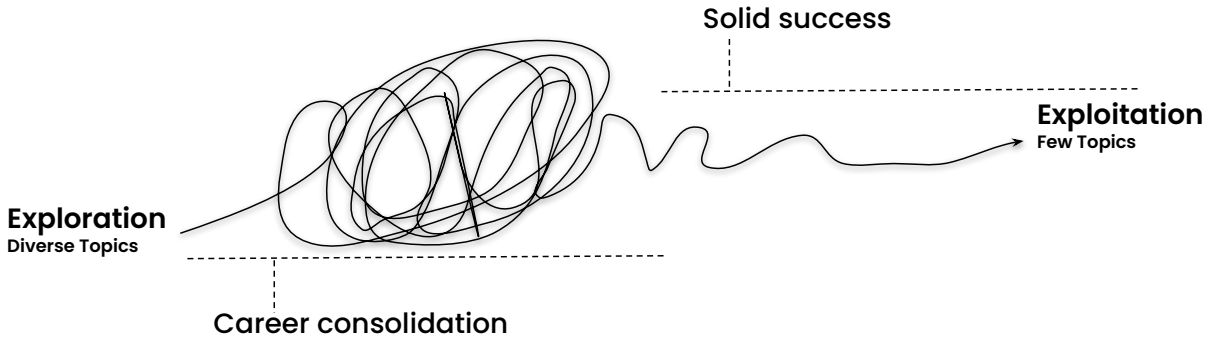
The onset of Hot Streaks in the Musical Ecosystem

Computer Science has offered solutions to a myriad of problems that are context-specific. For example, the music industry has benefited from a long and productive collaboration with computer science, and specially Databases. Indeed, many data-oriented techniques provide proper interface for accessing large datasets on music and artists, which, in turn, has boosted research in MIR. Since Byrd and Crawford's (2002) seminal position paper, various music industry problems use data-oriented solutions to automatically classifying music genres, addressing user gender bias in recommendation algorithms, and so on (e.g., [82]).

Despite such advances, the music industry still requires specialized approaches due to data-oriented challenges, such as dynamic evolution, volatile information, and diverse data sources, which instigate music managers to consider the power of data and database tools. A meaningful example of open issue in such an industry is music success: *Why do careers become successful? How to build successful careers? How to maintain success?* In order to solve such questions, researchers may use database techniques such as those based on time series. Going one step forward (or above), studying creative careers (e.g., music and literature) offers many interesting and relevant questions that can be explored over their data perspective.

Creative careers are interconnected with several factors that impact their success [133]. For example, some authors argue that the luck factor is the main ingredient to achieve success [55], whereas others claim that only hard work and resilience can build a successful career [92]. The reality is: there is no magic formula or off-the-shelf recipe for success in creative careers. Indeed, a successful career may have to undergo several phases and stages until it consolidates itself in the market. Such characteristic is also present in science, film, music, art, and even sports [132]. Aside from "regular" success, identifying the most impacting work or the most successful period within a creative career is challenging. Again, these most successful periods are known as *Hot Streaks*: a specific period during which an individual's performance is substantially better than their typical performance [73]. Further, we have already shown that it is possible to identify the

Figure 4.1: Exploration and Exploitation Phases.



Source: The Author.

presence of success above the normal in the musical context [106].

Regarding data, the music industry is a unique source of complex perspectives. First, the industry dynamics is always evolving – for example, a huge upgrade has changed it from the physical era (lead by physical LPs, CDs, and DVDs) to the current digital age (lead by streaming platforms). Second, the phonographic market is complex and formed not only by musicians but also by record companies, labels, studios, engineers, and producers. Further, another challenging dimension considers the type or genre of music, which divides the whole market into specific niches, from classic jazz to k-pop.

The latter perspective is paramount, as artists in the early stages of their careers usually explore different topics and genres. In other words, they explore features that allow them to transit among different target audiences. For example, young artists may explore new rhythms before identifying with a specific rhythm as a strong point of their careers. As their careers take off, artists tend to specialize more in a particular music segment, with less market exploration. This behavior leads us to question whether the above-normal success links to two concepts: Exploration and Exploitation.

In summary, *Exploration* refers to the process of visiting new regions of the search space; whereas *Exploitation* is the process of delving into a given area in search of a local optimum. Another important concept is *entropy*, which measures the diversity or variability within a dataset and helps to evaluate the balance between exploration and exploitation. These concepts are widely discussed in the field of artificial intelligence and are known to be the pillars of solving search problems [28]. Then, bringing such concepts to music careers, we hypothesize that artistic careers follow the exploratory phase until they reach their peak of success (hot streak). Specifically, Figure 4.1 shows a representation of the exploration and specialization phases: at the beginning of their careers, artists may explore different topics when they are still consolidating themselves in the market (different rhythms, genres, audiences, presentation formats, etc); as their careers are consolidated and known, they tend to focus on whatever works better.

Putting all together, it is still unclear if and how hot streaks relate to topics/genres explored and exploited by music artists. In other words, to the best of our knowledge, we

are the first to investigate topics *entropy* as a possible factor for Hot Streaks. Overall, our work is motivated by RQ3 (*Do artists explore different topics in their careers before reaching their first Hot Streak?*), which can be further divided into two parts:

> **a)** *What are the existing topics (genres) in a given musical career?*
> **b)** *Do musical careers reflect exploitation or exploration regarding such topics within a timeline of hot streaks?*

In this chapter, we describe a general methodology that provides a clear strategy for analyzing exploitation and exploration over hot streaks timeline (Section 4.2). Next, as the music topics are the core of entropy analysis, we propose and build a song network to extract the work topics of each artist (Section 4.3). The goal is to answer part a by extracting relevant topics that minimally explain musical careers and measuring their diversity through time (before, during, and after the careers achieve hot streaks). Next, we answer part b by analyzing entropy to understand musical career exploration and exploitation dynamics (Section 4.4). Such research may help understand how artists can manage their careers at different times, with the peak period of above-average success (i.e., their first hot streak) as the point of observation. Finally, we go over conclusions and limitations (Section 4.6).

4.1 Overall Scenario

Music has a distinct social dimension, whether shaped by the engagement of artists or listeners. Different social media services (such as Facebook, YouTube, and Twitter) are designed to engage audiences and encourage them to discover new artists, share recommendations and consume music [4]. Accordingly, the number of studies aimed at discovering the recipe for musical success has increased, defining the field of HSS, which addresses the problem of predicting the popularity of songs [90]. In this context, recent studies analyze the impact of different factors on musical success, including acoustic characteristics [78, 124], lyrics [120], collaboration [88, 112], or even engagement on social platforms [22, 27].

Creative careers have always instigated researchers and industry to understand how to build or improve them. Recent studies use machine learning techniques to propose algorithms, approaches and studies to include or exclude features that enhance the prediction task [11, 68, 84]. Consequently, the music industry benefits from the researchers' proactivity, as they always propose updates regarding the phonograph market nuances to

computational algorithms. However, we note a lack of exploring the common characteristics that lead to successful musical careers and associating such attributes with the most relevant periods in the artists' careers. After all, a significant milestone in artistic careers is the existence of Hot Streaks, that is, those periods of high notoriety and presence in the music market. Although Hot Streaks are practically a rule in musical careers, it is unclear if there are patterns regarding their beginning. The lack of systematic explanations for Hot Streaks and the randomness of when they occur within music careers support an unpredictable view of such careers.

Recently, [72] have explored the characteristics that contribute to the emergence of hot streaks in three different creative careers: scientific, cultural, and artistic. They find Hot Streaks are not associated with either exploring or exploiting behavior in isolation. Furthermore, they show real careers are complex, with heterogeneous influences operating across domains and many individual and institutional factors. In this context, we understand musical careers are similar to such creative careers and lack studies to characterize better what precedes the periods of Hot Streaks. Therefore, our goal is to understand how music careers achieve success by exploring and exploiting (entropy) topics in their careers. Such knowledge complements the results of machine learning tasks that create a model for predicting success. Further, we generate such information by considering technical and social characteristics as well.

4.2 Methodology Overview

Understanding what leads an artist to reach their most successful period (or the period in which their career gets notoriety) requires evaluating different data-driven perspectives of the music domain that evolve over time. One major perspective is the musical genre of artists and their songs. In such a context, we introduce a methodology to extract topics from artists' careers from the genres associated with them. In order words, we seek the topics (in this case, genres) associated with each song released by each artist. However, we understand that music genre classification is a complex problem in the MIR area, and is beyond our scope. Overall, the new methodology has five steps, as illustrated in Figure 4.2 and described ahead.

Figure 4.2: Five main stages to identify music genres (topics) based on hot streaks.

Source: The Author.

4.2.1 Collect Data

We built a *crawler* in Python using the Spotify API[1] to collect the data. First, we collected both Global and Brazil's Top 200 Daily chart data from Spotify Charts. Table 4.1 summarizes the columns of this dataset. For both cases, the collection date started on 2017-01-01 (Spotify's first availability date) and ended on 2022-03-13. After this period, Spotify closed access to its chart data. For both the Global and Brazilian markets, we got 379,200 chart lines. We then collected data for all artists, resulting in 286,275 artists for the Global market and 217,749 for Brazil. Finally, we collected data from all songs found on the charts, with 2,195,243 songs for the Global market and 1,447,784 for the Brazilian market.

4.2.2 Build Time Series

The evolution of an artist's success is represented by the number of daily streams received on Spotify. We use Global and Brazil's Spotify Top 200 Chart to model artists' success over time. For all artists, each point in their time series represents the total number of streams (i.e., the number of times the song was listened to on Spotify), which is our success measure. For example, Figure 4.3 presents the time series of The Weeknd. The album *After Hours*, released in 2020, continues to break records even after years. In addition to breaking all records with *Blinding Lights*, which appeared for 90 weeks on the most important chart in the United States (Billboard), the band also celebrates *Save Your Tears* with 60 weeks of charting on the Billboard Hot 100.[2]

[1]Spotify API: https://developer.spotify.com/
[2]The Weeknd Charts: https://bit.ly/theWeeknd_history

Table 4.1: Main features collected and enriched from Spotify Top 200 charts.

Spotify Top 200 Enriched Chart		
Column	**Description**	**Example**
market	Data chart from determined market	Global, Brazil
chart_type	The popularity of songs within a specific geographic area and viral charts spotlight songs going viral on a broader scale	Regional, Viral
chart_day	Reference date for the chart entry	2017-01-01
position	Song position for the chart entry	1, 2, ..., 200
song_id	Spotify song identifier	5aAx2yezTd8zXrkmtKl66Z
song_name	Song name	Starboy
artist	The participating artist (or artists)	The Weeknd, Daft Punk
streams	Number of streams for the respective chart entry	3,135,625
is_new	Verify if this is the first occurrence of the song in the chart	0, 1
is_reentry	Verify if this song has been already on the chart	0, 1
last_day_position	Previous position from a song on chart	1, 2, ..., 200
days_on_chart	Number of consecutive days from a song on the chart	1
peak_position	Chart position of a song	1
position_status	For reentry songs, this field stores the difference between the previous and actual position	-4, 2, ..., N

Source: The Author.

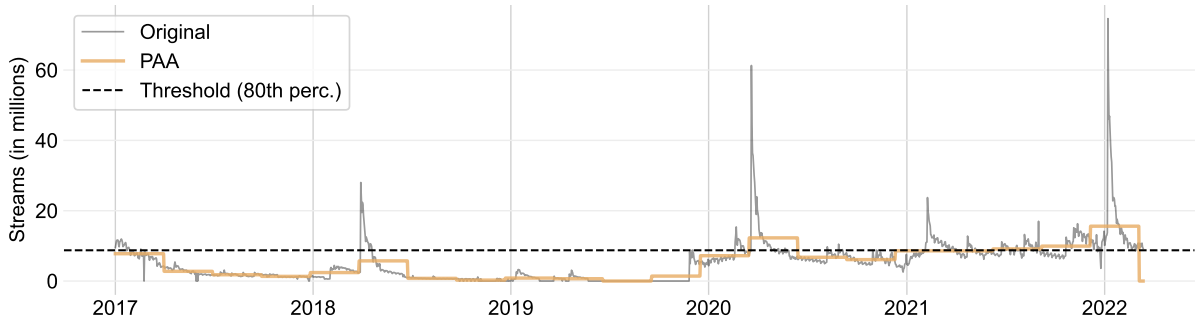
Figure 4.3: The Weeknd success time series over the Global market.

Source: The Author.

4.2.3 Detect Hot Streak

From the time series, we can now detect the artists' most successful periods, i.e., their Hot Streaks. We do so by applying the PAA method – Peacewise Aggregate Approximation, which reduces the dimensionality of the time series in N new dimensions [58]. Given a time series $X = x_1, x_2, \cdots, x_n$ of length n, PAA reduces it into a new series $\overline{X} = \overline{x_1}, \overline{x_2}, \cdots, \overline{x_N}$ with N dimensions, $1 \leq N \leq n$. The intuition is that dividing the original time series into N segments of equal size produces N new points. The value of each segment is defined as the average of the points within that frame (Equation 3.1).

Figure 4.4: The Weeknd success time series over the Global market with PAA.



Source: The Author.

Thus, each point of the original time series is approximated simply by attributing the PAA value of its corresponding segment. This method helps balance days with little or no success metric values, which was quite common in the data. Its only parameter is the number of segments to divide the series into. After empirical tests, we set this parameter to 90 days (N = 90): a value that captures enough information to represent an artist's success patterns over time, without being too short that could lead to information loss and scant representation of an artist's career trajectory. Next, we define a Hot Streak as the period when success (i.e., number of streams) is above a certain threshold obtained from the career. In other words, hot streak detection does not consider external factors (e.g., genre and time) because artists reach different levels of success, and choosing a single threshold would make the comparison unfair. For example, Figure 4.4 presents the time series of The Weeknd with his Hot Streak periods identified as the yellow line above the dotted line (in black).

4.2.4 Build Songs Network

With the hot streak information from artists' careers, our next step is building a music network to extract topics from their careers before, during, and after their Hot Streaks. In summary, given two songs A and B as nodes, we link them (define an edge) if there is an interposition of genres between artists of A and B (Section 4.3 presents further details for such a process). We build such network to get more information on musical genres associated with each artist. Spotify provides a set of genres for each artist. For example, Spotify returns the following genre list for The Weeknd: ['canadian contemporary r&b', 'canadian pop', 'pop']. However, as we want to analyze the evolution of each career in terms of topics (genres) explored, we need one or more topics per song. For example, we need genres such as [Synth-pop, new wave, synthwave, and electropop]

for the music *Blinding Lights* of The Weeknd and their respective genre lists for the other songs for this band.

4.2.5 Analyze Entropy

Finally, the last step of the methodology is to apply an algorithm that calculates the degree of entropy of the distribution of topics in the artists' careers. Here, the idea is that if an artist explores different topics, they will be in the exploration phase. Otherwise, by focusing on a handful of topics, the artist must be in the exploitation phase.

To quantify the exploration and exploration behaviors reflected in the careers of musical artists, we measure the style or entropy of the topic of their music (i.e., their artistic productions). For this, we define it as the frequency an artist engages in an art style or topic, and is the number of unique styles or topics. We show such entropy calculation in Equation 4.1, where i is the set of topics (genres) explored by an artist, p_i is the frequency of "devotion" by an artist, and m is the set of unique genres of whole artist's career.

$$\tilde{H} = -\sum_{i=1}^{m} p_i \, log \, p_i \tag{4.1}$$

If an artist uses a pure exploitation strategy ($\tilde{H} = 0$), their work sits in only one style or topic (i.e., the genre tuple returned by the LDA in the next section). On the other hand, an artist uses a strategy of pure exploration ($\tilde{H} = log \, n$) when they divide attention equally in analyzing the distribution of styles or topics.

4.3 Artist's Topic Extraction

Studying exploration and exploitation of music careers requires extracting topics that artists work on or have worked on to create a timeline of the spectrum of explored topics. However, our dataset[3] has no genre associated with each song/artist's work; we only have an aggregated set of all genres already explored by artists, as Spotify only provides a list of genres for each artist. In this sense, we developed a procedure to classify musical genres using a song network.

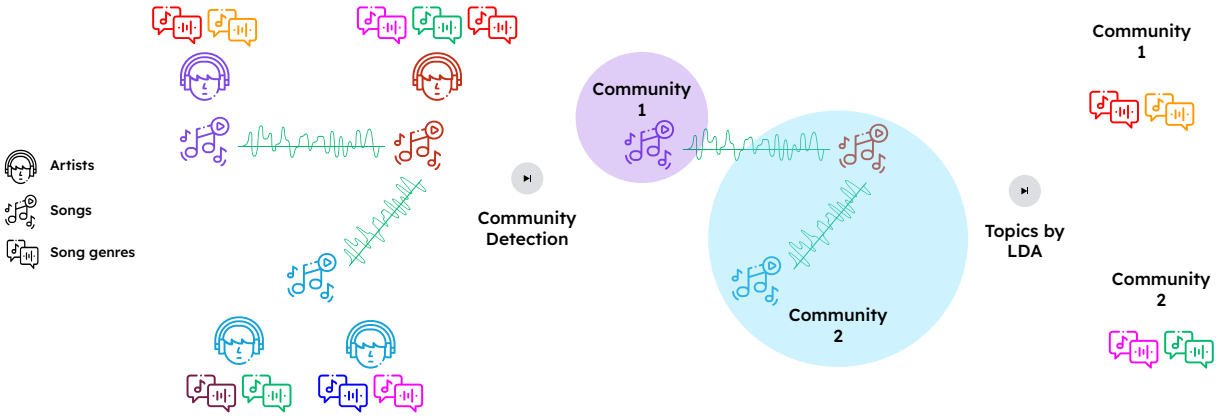
[3]MUHSIC: https://doi.org/10.5281/zenodo.5591015

Algorithm 1 Song network modeling algorithm

Input: Songs list (S) and artist genres list (G_L)
Output: Song network G

1: Create an empty edge list E
2: **for each** song s_i in S **do**
3: Add s_i in g_i ▷ Assign the respective artist's genre set $g_i \in G_L$
4: **end for**
5: **for each** pair of songs s_a and $s_b \in S$ **do**
6: **if** len$(g_a \cap g_b) > 0$: ▷ Check if the intersection between genres set g_a and g_b is not empty
7: create $s_a \overset{e_i}{\leftrightarrow} s_b$ ▷ Add and edge e_i between s_a and s_b with respective genres set g_a and g_b
8: **end for**

Figure 4.5: Topic Modeling to Song's artists.



Source: The Author.

4.3.1 Topic Modeling Procedure

Figure 4.5 summarizes the methodology to model the extraction of artists' topics. First, we create a music network based on the genres of all participating artists of the songs. Let S be a set of nodes represented by each song in our dataset, and E a set of edges connecting the songs. There is an edge e_i between two songs s_i and s_j if there is an intersection (partial or total) of the genre list of all participating artists of the respective songs s_i and s_j. Algorithm 1 summarizes the steps for network modeling.

With the music network ready, our next step is applying a community detection algorithm to identify each community's "top" genres. In other words, all songs that are part of a given community will have the most common genre of the respective community. We use Louvain's community detection algorithm,[4] a simple method to extract the community structure of a network by using Python's NetworkX library.

Finally, as the last step of identifying the artists' topics, we apply the Latent Dirichlet Allocation (LDA) algorithm to extract the most relevant genres from each recognized community. LDA is a generative probabilistic model of a corpus with the objective of modeling topics. LDA works on the premise that each topic is a set of terms, and each

[4]Networkx Louvain's Community: https://bit.ly/networkx_louvain

document is a mixture of a set of topics. The basic idea is that documents represent random mixtures over latent topics, where each topic characterizes a distribution over words [15]. We use two to model the communities' topics as the number of terms returned for each topic (i.e., a pair of song genres as a topic for each community). We also configure the algorithm to return only one topic per community. By selecting two terms per topic for each community, the result is a concise and interpretable representation of the genres associated with the community. Also, pairs allow a direct comparison between communities regarding their musical core. Next, we present the main results for both networks created: the global and Brazilian markets.

4.3.2 Topic Modeling Results

We now present the results of this phase, with the network metrics for global and Brazilian markets in Tables 4.2 and 4.3, and LDA topics for both markets in Tables 4.4 and 4.5.

Global Market. We build a network for extracting the artist's topics from the Global market. As an example of subnetwork, consider the songs:

- *Stay* by The Kid LAROI ft. Justin Bieber
- *God Is A Dancer* by Tiësto ft. Mabel
- *Look at Me!* by XXXTENTACION

The artist Justin Bieber has the genre *pop* on Spotify; The Kid Laroi has *hip hop*; Tiësto has *trance, house, dubstep, dance-pop, electro, pop, tropical house*; Mabel has *dance pop, pop, electro, tropical house*; and XXXTENTACION has *rap, hip hop*. Consequently, we now have the following pair of genres associated with the following songs:

- *Stay* assigned to [*pop, hip-hop*]
- *God Is A Dancer* to [*tropical house, dubstep, trance, electro, house, dance-pop, pop*]
- *Look at Me!* to [*hip hop, rap*].

Hence, after applying the network modeling, there is a link between the songs *Stay* and *God Is A Dancer* because the genres lists for both songs have pop in common. Likewise, there is a link between *Stay* and *Look at Me!* because of the hip-hop intersection.

Table 4.2: Music network metrics of Global Market.

Metric	Value
Number of nodes	7,725
Number of edges	15,492,869
Average degree	4011.09
Assortativity	-0.018
Average centrality degree	0.52
Density	0.52

Source: The Author.

Table 4.3: Music network metrics of Brazilian Market.

Metric	Value
Number of nodes	4,933
Number of edges	4,522,166
Average degree	1833.43
Assortativity	0.14
Average centrality degree	0.37
Density	0.37

Source: The Author.

The main statistics of the network created are summarized in Table 4.2. As the network is built by pairwise combinations, it has 15,492,869 edges and 7,725 nodes.

The next step is to identify the communities in the generated network. The Louvain algorithm has the RESOLUTION parameter, which is a critical factor in community detection algorithms as it controls the level of granularity at which communities are identified. After empiric experiments, we set such a parameter to 1.15, and Louvain returned 16 distinct communities. When the value of RESOLUTION is less than one, the algorithm favors identifying larger communities. In contrast, values greater than one tend to jeopardize the detection of smaller, more specific communities. The 16 communities identified through our analysis represent unique groupings of nodes in the network that are tightly interconnected with one another, providing valuable insights about the main genres for the songs set in each community, delimited by the LDA algorithm.

Finally, identifying the main topic of each community is crucial in establishing a distinct genre for all songs within that group. To accomplish this, we employ the LDA algorithm, which allows to isolate the most relevant themes present in each community. Overall, LDA returns a pair of genres representing each community's main topic; which then becomes the primary genre for all songs associated with that community. For instance, community zero's songs share the [k-pop, anime] genre, while community 12's songs fall under the [reggaeton, latin] genre. Table 4.4 shows the topics produced by the algorithm for each global market community. The songs previously used as examples were set as part of the following communities: *Stay* in community 13, with the topic [*pop, grime*]; *God is a Dancer* in community 1, with [*dance pop, electro*]; and *Look at Me!* in community 4, with [*hip hop, pop*].

Brazilian Market. Similarly, consider the following songs as examples of a subnetwork from the Brazilian Market:

- *Mal Feito - Ao Vivo* of Marília Mendonça ft. Hugo & Guilherme
- *Piseiro Estourou - Ao Vivo* of Os Barões Da Pisadinha
- *Avisa Que Eu Cheguei* of Naiara Azevedo ft. Ivete Sangalo

Table 4.4: LDA topics for Global Market communities.

ID	LDA Topic	ID	LDA Topic
0	[k-pop, anime]	8	[new wave, classic rock]
1	[dance pop, electro]	9	[arrocha, sertanejo]
2	[dance pop, hip hop]	10	[rap, pop]
3	[hip hop, rap]	11	[pop, trap]
4	[hip hop, pop]	12	[reggaeton, latin]
5	[r&b, rap]	13	[pop, grime]
6	[hip hop, trap]	14	[psychedelic]
7	[electro, trap]	15	[mariachi, ranchera]

Source: The Author.

Table 4.5: LDA topics for Brazilian Market communities.

ID	LDA Topic
0	[rock, mpb]
1	[k-pop, k-rap]
2	[pop, dance pop]
3	[brazilian funk, pop]
4	[sertanejo, arrocha]
5	[hip hop, rap]
6	[easy listening, lounge]

Source: The Author.

Artists Marília Mendonça and Hugo & Guilherme have the genres *sertanejo, arrocha* on Spotify; Os Barões da Pisadinha has *arrocha, forro*; Naiara Azevedo has *sertanejo, brazilian funk*; and Ivete Sangalo has *axe, samba reggae, pagode, arrocha, brazilian funk, mpb, pop*. As a result, we now have the following pair of genres associated with the following songs:

- *Mal Feito* is assigned to [*arrocha, sertanejo*]
- *Piseiro Estourou* to [*forro, arrocha*]
- *Avisa Que Eu Cheguei* to [*brazilian funk, sertanejo, samba reggae, mpb, pop, axe, pagode, arrocha*]

After applying the network modeling, there is a link between the songs *Mal Feito* and *Piseiro Estourou* because the genre lists for both songs have *arrocha* in common. Similarly, there is a link between *Piseiro Estourou* and *Avisa Que Eu Cheguei* because of the *arrocha* intersection. The main statistics of the network created are summarized in Table 4.3. Similar to the global market, as the network is built by pairwise combinations, it is expected to have many edges. Such a network has 4,522,166 edges, and the number of nodes is 4,933.

With the network created, we also run an empirical test to find the RESOLUTION parameter that controls the level of community granularity. We set this parameter to 1.3, resulting in seven distinct communities being identified. Hence, we also apply the LDA to obtain a pair of genres representing each community's main topic. We then designate this pair as the primary genre for all songs associated with that community. For instance, community zero's songs share the [rock, mpb] genre, while community 2's songs are set in the [pop, dance pop] genre. Table 4.5 shows the topics produced by the algorithm for each global market community. Here, the songs *Mal Feito, Piseiro Estourou* and *Avisa Que Eu Cheguei* previously used as an example were set as part of community 4 [sertanejo, arrocha]. Once we have assigned primary genres to each community, we can move on to

the next phase of our analysis, which involves applying entropy analysis methodology to explore the topics covered by the artists.

4.4 Exploration and Exploitation Analyses

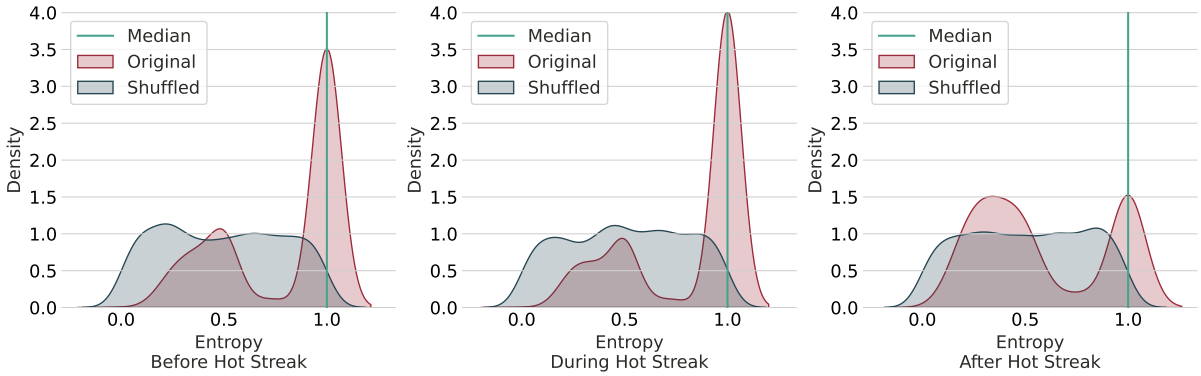
Countless factors can influence an artist's career evolution and success. The exploration and exploitation strategies have attracted interest in a wide range of expertise areas, leading us to examine their potential relationship to hot tracks in the musical ecosystem. First, we need to distinguish these two topics: exploration and exploitation. *Exploitation* allows individuals to build knowledge in a given area and refine their capabilities over time. Such a phenomenon may be relevant to understanding how to achieve hot streaks, as exploitation allows individuals to focus on a particular "area" to establish expertise in that area and gain a reputation related to such proficiency. For musical artists, we may say that an artist focuses on a small range of musical genres, for example. On the other hand, *exploration* involves individuals experimenting and searching beyond their existing or previous competence areas. Defining a parallel with music, artists explore various musical genres in their careers.

Furthermore, whereas exploration is riskier and consequently associated with more significant variation in results, it can also increase the likelihood that someone will stumble upon an innovative idea through unforeseen combinations of disparate sources. In contrast, exploitation can suppress originality and limit an individual's ability to consistently produce high-impact work over time. In this sense, the benefits and disadvantages of these contrasting approaches raise an essential question: Are the behavior of musical artists' careers a reflection of exploitation or exploration? Understanding the balance between these two strategies and how they contribute to the success of artists can help acquiring insights into the nature of creativity and innovation in the musical industry.

Indeed, the results of our study reveal an interesting pattern in the career trajectories of successful musical artists. Specifically, music artists tend to diversify their musical styles before and during their first Hot Streak period. This trend is illustrated in Figures 4.6 and 4.7, which depict the entropy distribution in three stages of an artist's career: before, during, and after their Hot Streak, for both the Global Market and the Brazilian Market. Note we also plot and compare artists' careers with random careers to check the robustness of the results. In other words, random careers are generated by the computer to compare with real careers and verify the differences.

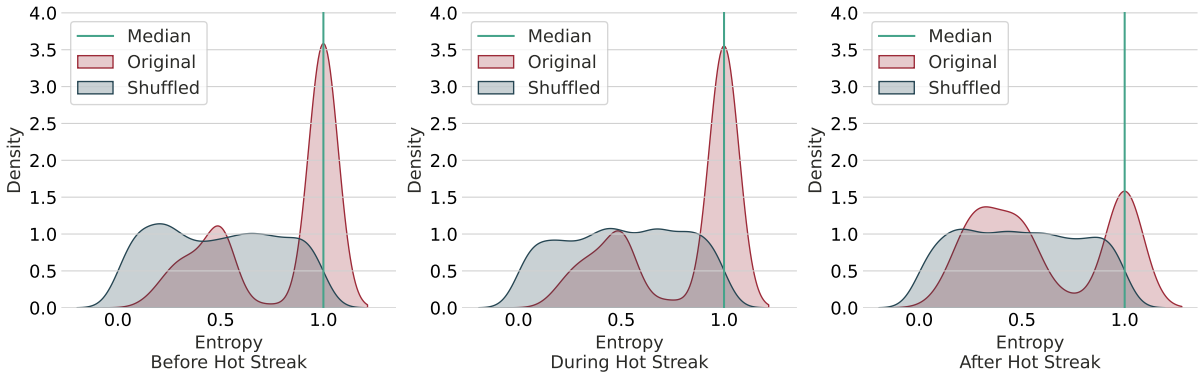
The graphics show that artists are more likely to explore fewer new rhythms and musical genres after achieving above-normal success (i.e., exploitation phase). Further-

Figure 4.6: Entropy distribution.



Source: The Author.

Figure 4.7: Entropy distribution of Brazilian Market.



Source: The Author.

more, our data revealed that some artists continue to invest in exploration even after their Hot Streak period has ended, indicating that they view exploration as a crucial aspect of their creative process. This behavior suggests that artists may feel more inclined to experiment with new sounds and styles when they have achieved a certain level of recognition and have a broader platform to showcase their work.

A possible scenario for artists who continue to explore new sounds is: they use their fame and influence to support emerging artists who work in different genres. Successful artists can continue trying and exploring new styles while supporting new talent by starring contemporary artists at the beginning of their careers. This approach also helps keeping their work fresh and relevant as they face new ideas and sounds through collaborations with emerging artists. Overall, these findings provide valuable insights into successful musical artists' strategies to sustain their success in the industry. By diversifying their musical styles and investing in exploration, these artists are better equipped to adapt to changing trends and maintain their relevance in a constantly evolving industry.

4.5 Case Study

Here, we discuss two examples of artists' careers: Jason Derulo and Anitta. They represent both markets, Global and Brazil, respectively. Table 4.6 summarizes the results for the selected artists for our discussion. At the top of such a Table are the results for Jason Derulo, while at the bottom are the results for Anitta.

Jason Derulo is an American singer, songwriter, and dancer who rose to fame in the late 2000s. He starts his musical career as a songwriter, producing tracks for prominent artists like Diddy, Lil Wayne, and Sean Kingston. According to our data, Derulo explored more topics before and during his period of Hot Streak (i.e., his highest success already registered). The entropy for both periods is 1.0, representing *Exploration* phase. The respective songs for this period include *If I'm Lucky* and *Tip Toe*. The resulting genres by our genre network (Section 4.3) for these songs are *dance pop, eletro* and *rap, pop*. We search the respective genres over the internet (Wikipedia) to better check if the resulting genres conform to the real ones. The returned genres for *If I'm Lucky* and *Tip Toe* is *pop*, agreeing with our results. After the Hot Streak period of such an artist, his associated entropy decreases to 0.36, which indicates an *Exploitation* phase. In this case, the bet is to focus on fewer topics. In fact, songs for the period, such as *Goodbye* and *1, 2, 3*, are more like *latin* rhythms and *reggaeton*. Besides, one of his latest releases is the single *Slow Low* that is classified as *pop* and *reggaeton*. Finally, our analysis suggests a dynamic shift between *Exploration* and *Exploitation* phases, characterized by a diverse exploration of genres during his Hot Streak period and a subsequent focus on specific styles in his later releases. Still, our results are based on a particular time frame and may not fully capture the entirety of his career. External factors such as marketing strategies and industry trends could also influence the observed patterns. Further research and a broader dataset would be valuable for a more comprehensive understanding of Jason Derulo's artistic evolution.

Anitta is a Brazilian singer, songwriter and entrepreneur. She gained widespread recognition in Brazil and internationally for her unique blend of pop, funk, and reggaeton music styles. According to our results in Table 4.6, Anitta is fluctuating between *Exploration* and *Exploitation* in her phases (before, during, and after hot streaks). Specifically, during her hot streak, she had the most value for entropy result (0.59) – *Exploration* tendency. Despite that, she has most of her work transitioning between *pop* and *reggaeton* genres – by our methodology and even Wikipedia. Nevertheless, we believe she explored other topics (genres) before her hot streaks. After all, her career starts mainly as a *funk* singer (or *gospel*, at her church). However, our data comprise her career only from 2017 on, capturing initial results from her investments in the international recognition.

Indeed, such a limitation does not capture when she started her career in 2010

Table 4.6: Entropy results and their respective songs and genres by Hot Streak phases for Jason Derulo and Anitta.

Global - Jason Derulo				
Hot Streak Phase	**Entropy**	**Songs**	**Genres By Wikipedia**	**Genres By Community**
Before	1.0	If I'm Lucky / Swalla	[pop] / [r&b, dancehal]	[dance pop, eletro] / [r&b, rap]
During	1.0	Tip Toe	[pop]	[rap, pop]
After	0.36	Goodbye / 1,2,3	[latin, eletro] / [latin]	[dance pop, eletro] / [reggaeton, latin]
Brazilian - Anitta				
Hot Streak Phase	**Entropy**	**Songs**	**Genres By Wikipedia**	**Genres By Community**
Before	0.48	Paradinha / Sua Cara	[dancehall, reggaeton] / [Moombahton]	[dance pop, eletro] / [dance pop, eletro]
During	0.59	Machika, Indecente	[pop latino, reggaeton] / [pop latino]	[reggaeton, latin] / [dance pop, eletro]
After	0.45	Boys Don't Cry / Envolver	[electropop, pop rock] / [reggaeton]	[dance pop, eletro] / [dance pop, eletro]

Source: The Author.

– by posting videos of herself singing on YouTube. Her first big breakthrough came in 2013 with the release of her debut single *Show das Poderosas*, which became a massive hit and established her as a rising star in the Brazilian music scene. In the subsequent years, Anitta released several successful albums and singles, earning numerous awards and accolades for her music. She quickly became one of Brazil's most prominent and influential pop stars, known for her catchy songs, vibrant performances, and captivating music videos. Anitta's popularity extended beyond Brazil, and she has collaborated with various international artists, including *Madonna*, *Cardi B*, and *J Balvin*, which further expanded her global reach.

4.6 Concluding Remarks

The music industry (like any other industry) deals with the constant natural evolution of the market, which reflects the need for adaptation to satisfy consumers. Trying to accommodate itself to potential changes, the market may explore the cycles of topics (e.g., musical genres) that are either up or down, where machine learning algorithms could be able to capture the nuances of when a cycle is about to end, and which one is predisposed to start by analyzing the available data. We know that understanding the behavior of creative careers (e.g., musical) is not a simple task due to their subjective nature (which considers factors that are not precisely measurable, such as creative capacity, personality, and other abilities). In addition, we must consider out-of-personal factors, such as the publicity force, social media power, etc. However, promoting a broader analysis of what can lead to a successful music career is still a necessary tool for the industry.

In this chapter, we analyzed how to identify regularities concerning the onset of

hot streaks in music careers through a data-driven approach. Using a song network, we proposed a methodology for extracting relevant topics associated with such careers. Then, we used a method for measuring the entropy degree on such topics to quantify the level of exploration and exploitation about the first hot streak of every musical career. Our results were promising, indicating there may be some regularity in prospecting topics to reach hot streaks. Specifically, artists explore diverse topics before hitting their first Hot Streak and during their most successful period. Then, we detected that artists tend to enter the exploitation phase as they leave their Hot Streaks periods. That is, they reduce the range of topics explored and are prone to specialize in fewer topics. Overall, our results highlight the important role of exploration and exploitation in individual careers, suggesting that a sequential view of such strategies balancing experimentation and implementation may be compelling for producing lasting careers. These findings could be relevant for identifying, training, and bringing up creative talent.

Limitations. The data-driven nature of our study leads to common limitations in this type of analysis. First, regarding the datasets used to represent large collections of music career histories: they are limited to individuals who have had sufficiently long careers to provide enough data points for statistical analysis. Second, this paper presented correlational evidence, whose main objective is to investigate empirical regularities associated with the appearance of hot streak marks.

Chapter 5

Conclusion and Final Considerations

In this chapter, we present the main conclusions of this thesis in Section 5.1 and summarize the contributions for each research question in Section 5.2. We also discuss the research's limitations and potential threats to validity in Section 5.3. Further, we highlight ideas for future improvements and propose a sequence of this research in Section 5.4. Finally, we provide a list of publications during the doctoral period in Section 5.5.

5.1 Conclusion

This thesis uncovered that Hit Song Science has been gaining ground as a legitimate scientific field since the seminal paper that challenged its validity as such [91]. Interestingly, the authors who had initially questioned its scientific nature dedicated an entire chapter to such a topic in their subsequent work. As demonstrated in Chapter 2, there has been a surge in the number of research papers devoted to this area. In this sense, our initial contribution to this field was to conduct a systematic literature review and propose a taxonomy based on the primary themes identified in the analyzed works.

Going further, we raised questions about how we could contribute to the field of Hit Song Science. From these questions, one of the main is *"What drives a song to become a hit?"* After all, this can be a guiding question for the music industry, whose precept is to find and nurture talent. In this sense, we have seen that many works discuss the best machine learning techniques, the best set of features to feed models, the best approaches between recommendation practices, and even hybrid systems that can explore the best of each of the previous methods in a single strategy.

Moving to the streaming era, Spotify is a Streaming company born in Sweden whose primary purpose was to promote access to music in digital format, without infringing copyrights, in response to The Pirate Bay, a file-sharing site of peer-to-peer digital media among its users. Another Spotify inspiration was to democratize access to culture, helping artists to be known by their future audience and also to be a facilitator in the consumption

of music by end users, providing agility in choosing and personalizing playlists to be listened to anywhere.

Spotify participated in the initial movement that kicked off the creation of other streaming companies and increasingly consolidated the digital era in the music market. In this sense, we analyzed the evolution of music consumption by comparing music consumption data from the physical and digital eras. In the physical age, artists had to fight to get ratings on Radio shows and later on TV. With these appearances, artists could sell more vinyl records and, later, CDs, and (shortly) DVDs. Generalizing physical media as records, the most successful artists sold the most. In the digital age, sales are mainly accounted for by the number of streams. The more streams an artist has, the higher their royalties. Specifically, we performed a comparison analysis between such eras for the Brazilian data, which is an up-and-coming market that has already been among the top 10 global markets in the world of music, according to the 2020 IFPI report. One of the main contributions of this analysis is that Brazilians prefer Brazilian music or music by Brazilian artists. These nuances of market preferences help to build a consumption profile. Consequently, artists and the music industry can feed on such information to better guide the direction of investments and strategies for markets locally. This knowledge goes beyond machine learning results, which typically deal with global aspects of data, making it challenging to capture specific details. In short, considering the current era of streaming, where the dissemination and consumption of foreign music have become much more possible and direct, Brazil has its particularity and globalization is not greater than its preference for its artists.

Another significant contribution of this thesis was to detect that the artists of both eras have some periods of greater prominence, the so-called Hot Streaks periods. Such periods of success led us to another question: "Is it possible to identify any pattern that precedes the existence of such periods?" From this, we performed (Chapter 4) an entropy analysis to know if there is a certain regularity in terms of Exploration and Exploitation of topics that characterize the artist's career. Hence, we proposed a data-driven approach to the artists' musical genres to associate them with their songs (since our data does not have genre information at the music level). With topics (genres) associated with each song, we can then build a timeline and detect the level of exploration and exploitation of such topics in artistic careers. Brazilian and foreign artists tend to generalize their topics before reaching a period of above-normal success (i.e., exploring diverse themes) and then focus on a specific niche after they leave their periods of success.

Next, we revisit each research question that guided this thesis. We also summarize the main contributions of each research question (RQ1, RQ2, and RQ3), respectively presented in chapters 2, 3, and 4 of this thesis.

5.2 Summary of Goals and Results

Given the open questions driving this thesis, we have summarized the main findings achieved for each of the three research questions that guided our investigation.

RQ1. How does current research deal with Hit Song? We described the main research problems, frequently used data sources, musical success definitions, features, and learning methods in HSS. As a result, we proposed a generic workflow for Hit Song Prediction, and we also suggested novel taxonomies for (*i*) success measures, (*ii*) features, and (*iii*) learning methods used to consolidate the existing knowledge in HSS. We concluded there is not one "feature" for an ideal hit song prediction model, as its performance depends on subjective decisions made in the analysis process.

RQ2. How does the evolution of music consumption in Brazil affect the occurrence of hot streaks? We performed a clustered success analysis to answer the following questions: are artists' most successful periods clustered in time? How to detect the artists' most successful periods (Hot Streaks)? First, we witnessed the Hot Streaks and then performed a cluster analysis to understand if it is possible to distinguish the artists by their level of success. Next, we analyzed the most popular genres in the Brazilian market in different eras. Finally, we also verified if the artists underwent their first period of Hot Streak at similar times. Our main conclusion is that Brazilian listeners listen to Brazilian artists as their first option, independent of the era (physical or digital).

RQ3. Do artists explore different topics in their careers before reaching their first Hot Streak? We investigated whether artists explore different topics regularly before experiencing their first hot streak (exploration) and whether they specialized in specific segments from that point onward (exploitation). Hence, we propose extracting relevant topics in artists' careers by modeling their songs as a similarity network. Next, we applied a network's community detection and LDA algorithms to extract the career topics and then calculate the topic distribution entropy to measure the variation of topics encountered over such careers. Finally, we correlated the hot tracks' timing with the artists' creative trajectories to verify changes in the work characteristics. We found that artists explored more topics before and during their hot streak periods, and then they entered to exploitation phase when they left their hot streak periods. Such findings are valid for both markets: Global and Brazil.

5.3 Limitations and Threats to Validity

There is no single formula for success in the music industry, as numerous external factors (such as market trends, culture, and social changes) may influence such an industry. Therefore, while artists in the early stages of their careers can use insights from historical success data to guide their choices and make informed decisions, it is essential to recognize the inherent limitations of such an approach.

Furthermore, it is crucial to remember that music is an art form, and many qualities beyond technical skills can influence an artist's success. Factors such as charisma, personality, and emotional connection with the audience are equally, if not more, important than musical talent itself. Therefore, finding a balance between analyzing historical data and artistic perception is necessary to build a solid and enduring musical career.

In addition, it is necessary to consider potential limitations and biases when conducting data analysis. Outliers may arise, where an artist who should have achieved success based on the analyzed data does not. Factors like luck and unpredictability can play significant roles in the music industry. The formula or methodology used for analysis may differ from the most optimal or comprehensive, and there is a risk of overlooking relevant variables. Moreover, choosing a genre can introduce complexities and challenges that may affect the outcomes of data-driven predictions. Considering these limitations, it is crucial to approach musical success cautiously and understand that it is a complex and multifaceted phenomenon that cannot be fully captured by data alone. Incorporating subjective artistic elements and individual circumstances is essential for a comprehensive understanding of an artist's journey and potential for success.

From another perspective, it is essential to recognize that other factors can influence a career and/or song's success. For example, different versions of the same song (such as remixes, live performances, or covers) can achieve more success than the original version. These variations can present acoustic differences and aspects related to the musical arrangement, such as the time of entry in the chorus or the bridge, which can impact the public's perception and preference. Therefore, understanding musical success goes beyond analyzing isolated characteristics and requires carefully considering the nuances of creating and receiving different song versions.

It is important to emphasize that although these limitations and variations exist in the musical versions, they do not invalidate our analyses. The aim is not to find a definitive musical success formula but to explore patterns and trends based on available data. We recognize that different versions of a song can have discrepant results, but that does not mean we should address all variations and characteristics in our analyses. On the contrary, the diversity and complexity of the musical world make the study of success so challenging and exciting. Therefore, even with these limitations and nuances, the analyses

carried out in this study can still provide valuable insights and contribute to a broader understanding of the phenomenon of musical success.

Lastly, and not to be overlooked, this thesis results from an individual with a strong computing background and a genuine passion for music. While the author has expertise in data analysis and computational techniques, their love for music may introduce a potential bias in the results. Despite not having a formal background or proficiency in music, this perspective allows a fresh approach to analyzing music-related phenomena from a computational standpoint. By leveraging their computational skills and combining them with a deep appreciation for music, the author aims to contribute valuable insights into the field of music analysis and shed light on the intersection of music and technology.

5.4 Future Work

In future work, we plan to identify whether ephemerality exists in other topics (e.g., acoustic features, collaboration, etc.) explored by artists associated with their Hot Streaks period. Hence, exploring other issues (features) serve as input for entropy analysis. Here, the main objective is to draw a pre-hot-streak profile of the artists. Such analysis may help the music industry understand the feature set that drives artists to achieve their hot streak periods. By proposing a successful profile of artists, the music industry will benefit from the possibility of having tools beyond machine learning techniques to understand what and how the artists' profiles stand out.

To compose this pre-hot-streaks profile, we plan to explore the entropy of collaborations between artists. We have already seen in [112] that collaboration increases the predictive power of musical success. However, it is necessary to understand the dynamics of these collaborations so that artists can leverage their gains and consequently increase the industry's power to generate valuable insights in recommending partnerships. An interesting analysis is to verify whether collaborations before the artists' most successful periods tend to be between artists of different genres, where lesser-known artists can gain space by piggybacking on already established artists. Another possibility is to verify if they make more or fewer collaborations after the artist enters his most successful period. Another analysis is to explore the acoustic features of the songs. In this case, the question to be answered is: Do artists tend to change their acoustic characteristics after achieving extraordinary success (i.e., their hot streak)?

The idea is to apply entropy analysis using our dataset's other topics (features). Exploring other issues is the input to profiling artists who still need to enter their period of hot streaks. Such a profile helps the music industry in the process of identifying potential

artists to integrate the artistic body of their company. Furthermore, this profile analysis may shed light on building a career from scratch for artists beyond their other intrinsic qualities (e.g., quality of timbre, ability to perform, mastery of musical instruments, etc.). In other words, artists who still need substantial experience but are predisposed to be talented can apply a success formula derived from historical data from previous careers. With this pre-hot streak profile added to other qualities that we could not measure (personal gift, luck, charism, etc.), further considering the relevant insights resulting from this thesis, the music industry may build a successful musical career.

5.5 Publications

We have already the following publications, all published during the Ph.D. (directly and indirectly) related to this project.

Thesis Chapters. Here, we show the publications directly related to this thesis.

1. **Seufitelli, D. B.**; Oliveira, G. P.; Silva, M. O.; Barbosa, G. R. G.; Melo, B. C.; Botelho, J. E.; Melo-Gomes, L.; Moro, M. M. From Compact Discs to Streaming: A Comparison of Eras within the Brazilian Market. REVISTA VÓRTEX, v. 10, p. 1-28, 2022. DOI.

2. **Seufitelli, D. B.**; Oliveira, G. P.; Silva, M. O.; Scofield, C.; Moro, M. M. Hit Song Science: A Comprehensive Survey and Research Directions. Journal of New Music Research. **(Under second review)**

3. **Seufitelli, D. B.**; M. M. From Exploration to Exploitation: Understanding the Evolution of Music Careers through a Data-driven Approach. *In:* Simpósio Brasileiro de Banco de Dados (SBBD), 2023, Belo Horizonte.

4. **Seufitelli, D. B.**; Oliveira, G. P.; Silva, M. O.; M. M. MGD+: An Enhanced Music Genre Dataset with Success-based Network. *In:* Dataset Showcase Workshop (DSW) - Simpósio Brasileiro de Banco de Dados (SBBD), 2023, Belo Horizonte.

Cultural-related Publications. Next, we show the publications chronologically sorted (Ascending).

1. Oliveira, G. P.; Silva, M. O.; **Seufitelli, D. B.**; Lacerda, A. M.; Moro, M. M. Detecting Collaboration Profiles in Success-based Music Genre Networks. In: 21st

International Society for Music Information Retrieval Conference, 2020, Montreal. LINK.

2. Barbosa, Gabriel R. G.; Melo, Bruna C.; Oliveira, Gabriel P.; Silva, Mariana O.; **Seufitelli, Danilo B.**; Moro, Mirella M. Hot Streaks in the Brazilian Music Market: A Comparison Between Physical and Digital Eras. In: Simpósio Brasileiro de Computação Musical, 2021, Brasil. p. 152-732. DOI.

3. Oliveira, Gabriel P.; Barbosa, Gabriel R. G.; Melo, Bruna C.; Silva, Mariana O.; **Seufitelli, Danilo B.** ; Moro, Mirella M. MUHSIC: An Open Dataset with Temporal Musical Success Information. In: Dataset Showcase Workshop, 2021, Brasil. Anais do III Dataset Showcase Workshop (DSW 2021), 2021. p. 65. DOI.

4. Silva, Mariana O.; Scofield, Clarisse; Oliveira, Gabriel P.; **Seufitelli, Danilo B.**; Moro, Mirella M.. Exploring Brazilian Cultural Identity Through Reading Preferences. In: Brazilian Workshop on Social Network Analysis and Mining, 2021, Brasil. p. 115-126. DOI.

5. Silva, Mariana O.; Scofield, Clarisse; Melo-Gomes, L.; Botelho, J. E.; Oliveira, Gabriel P.; **Seufitelli, Danilo B.**; Moro, Mirella M. Brazilian Reading Preferences in Goodreads: Cross-state and Cross-region Analyses. Brazilian Journal of Information Systems (iSys), 2022. LINK.

6. Oliveira, Gabriel P.; Barbosa, Gabriel R. G.; Melo, Bruna C.; Botelho, Juliana E.; Silva, Mariana O.; **Seufitelli, Danilo B.**; Moro, Mirella M. Musical Success in the United States and Brazil: Novel Datasets and Temporal Analysis. Journal of Information and Data Management - JIDM, 2022. DOI.

7. Silva, Mariana O.; Oliveira, G. P.; **Seufitelli, D. B.**; Moro, M. M. Collaboration as a Driving Factor for Hit Song Classification. In: 28th Brazillian Symposium on Multimedia and the Web (WebMedia), 2022, Curitiba - PR. DOI.

8. Melo-Gomes, L.; **Seufitelli, D. B.**; Oliveira, G. P.; Silva, Mariana O.; Moro, M. M. Análise do Sucesso Musical no Brasil Utilizando Dados do Twitter. In: SBBD WTAG - Workshop de Trabalhos de Alunos de Graduação, 2022, Búzios. Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados. Porto Alegre: Sociedade Brasileira de Computação - SBC, 2022. DOI.

9. Silva, M. O.; Oliveira, G. P.; **Seufitelli, D. B.**; Moro, M. M. Collaboration-Aware Hit Song Prediction. Journal on Interactive Systems. DOI

10. Silva, Mariana O.; Oliveira, G. P.; **Seufitelli, D. B.**; Moro, M. M. Temporal Success Analyses in Music Collaboration Networks: Brazilian and Global Scenarios. REVISTA VÓRTEX, 11(2), 1–27, 2023. DOI

11. Oliveira, Gabriel P.; Silva, Mariana O.; **Seufitelli, Danilo B.**; Barbosa, Gabriel R. G.; Melo, Bruna C.; Moro, Mirella M. Hot streaks in the music industry: Identifying and characterizing above-average success periods in artists' careers. Scientometrics. **(Under Submission)**

Forensic-related Publications. As the previous focus of the thesis was Digital Forensics, we published some works on such a topic as follows.

1. Mata, W. R. R.; **Seufitelli, D. B.**; Michele A. Brandão. JusBD: Um Banco de Dados para Obtenção de Informações do Poder Judiciário. In: 34th Simpósio Brasileiro de Bancos de Dados - Dataset Showcase, 2019, Fortaleza. LINK.

2. **Seufitelli, D. B.**; MATA, W. R. R.; SOUZA, R.; Michele A. Brandão; Moro, M. M. Characterization and Analysis of Open Brazilian Judiciary Data. In: 26th Brazillian Symposium on Multimedia and the Web (WebMedia), 2020, São Luís. DOI.

3. **Seufitelli, D. B.**; MOURA, A. F. C.; Fernandes, A.; Siqueira, K.; Michele A. Brandão; Moro, M. M.. Forense Digital e Bancos de Dados: um Survey. In: Simpósio Brasileiro de Banco de Dados (SBBD), 2021, Rio de Janeiro. DOI.

4. **Seufitelli, D. B.**; Michele A. Brandão; Moro, Mirella M. Exploring the Intersection between Databases and Digital Forensics. Journal of Information and Data Management - JIDM, 2022. DOI.

5. **Seufitelli, D. B.**; Michele A. Brandão; Fernandes, A.; Siqueira, K.; Moro, Mirella M. Where do Databases and Digital Forensics meet? A Comprehensive Survey and Taxonomy. SIGMOD Record. **(Accepted for publishing)**

Courses and Didactic Material.

1. Pimentel, João; Oliveira, Gabriel; Silva, Mariana; **Seufitelli, Danilo**; Moro, Mirella. Ciência de Dados com Reprodutibilidade usando Jupyter. Jornada de Atualização em Informática 2021. 1ed.: SBC, 2021, v. , p. 11-59. DOI.

Bibliography

1. Abel, F., Diaz-Aviles, E., Henze, N., Krause, D., and Siehndel, P. Analyzing the blogosphere for predicting the success of music and movie products. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM*, pages 276–280, Odense, Denmark, 2010. IEEE Computer Society. doi: 10.1109/ASONAM.2010.50. **P1**.

2. Abidin, C. Mapping internet celebrity on tiktok: Exploring attention economies and visibility labours. *Cultural Science Journal*, 12(1):77–103, 2020. doi: doi:10.5334/csci.140.

3. Al-Beitawi, Z., Salehan, M., and Zhang, S. Cluster analysis of musical attributes for top trending songs. In *Hawaii International Conference on System Sciences*, pages 1–7, Maui, Hawaii, 2020. ScholarSpace. doi: 10.24251/HICSS.2020.017. **P2**.

4. Amorim, A., Murrugarra-Llerena, N., Silva, V., de Oliveira, D., and Paes, A. Modelagem de tópicos em textos curtos: uma avaliação experimental. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 254–266, Porto Alegre, Brasil, 2022. SBC. doi: 10.5753/sbbd.2022.224314.

5. Araujo, C. V. S., Neto, R. M., Nakamura, F. G., and Nakamura, E. F. Predicting music success based on users' comments on online social networks. In *Brazilian Symposium on Multimedia and the Web*, pages 149–156, Gramado, Brazil, 2017. ACM. doi: 10.1145/3126858.3126885. **P5**.

6. Araújo, C. V. S., Neto, R. M., Nakamura, F. G., and Nakamura, E. F. Predicting music success based on users' comments on online social networks. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, pages 149–156, 2017. doi: 10.1145/3126858.3126885.

7. Araujo, C. V. S., de Cristo, M. A. P., and Giusti, R. Predicting music popularity using music charts. In *IEEE International Conference On Machine Learning And Applications*, pages 859–864, Boca Raton, USA, 2019. IEEE. doi: 10.1109/ICMLA.2019.00149. **P3**.

8. Araujo, C. V. S., de Cristo, M. A. P., and Giusti, R. A model for predicting music popularity on streaming platforms. *RITA*, 27(4):108–117, 2020. doi: 10.22456/2175-2745.107021. **P4**.

9. Askin, N. and Mauskapf, M. What makes popular culture popular? product features and optimal differentiation in music. *American Sociological Review*, 82(5):910–944, sep 2017. doi: 10.1177/0003122417728662. **P6**.

10. Awad, M. and Khanna, R. *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Apress Media, New York, USA, 2015. ISBN 978-1-4302-5989-3.

11. Baldo, F., Grando, J., Weege, K., and Bonassa, G. Adaptive fast xgboost for binary classification. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 13–25, Porto Alegre, Brasil, 2022. SBC. doi: 10.5753/sbbd.2022.224291.

12. Berns, G. S. and Moore, S. E. A neural predictor of cultural popularity. *Journal of Consumer Psychology*, 22(1):154–160, 2012. doi: 10.1016/j.jcps.2011.05.001. **P7**.

13. Bholowalia, P. and Kumar, A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *Int'l J. of Computer Applications*, 105(9), 2014.

14. Bischoff, K., Firan, C. S., Georgescu, M., Nejdl, W., and Paiu, R. Social knowledge-driven music hit prediction. In *International Conference Advanced Data Mining and Applications*, pages 43–54, Beijing, China, 2009. Springer. doi: 10.1007/978-3-642-03348-3_8. **P8**.

15. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

16. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

17. Blume, J. *Six Steps to Songwriting Success, Revised Edition: The Comprehensive Guide to Writing and Marketing Hit Songs*. Billboard Books, New York, USA, 2008.

18. Borges, H., Hora, A. C., and Valente, M. T. Predicting the popularity of github repositories. In *Proceedings of the The 12th International Conference on Predictive Models and Data Analytics in Software Engineering*, pages 9:1–9:10, 2016. doi: 10.1145/2972958.2972966.

19. Borges, R. and Queiroz, M. A probabilistic model for recommending music based on acoustic features and social data. In *Simpósio Brasileiro de Computação Musical*, pages 7–12, 2017.

20. Buda, A. and Jarynowski, A. Exploring patterns in european singles charts. In *European Network Intelligence Conference*, pages 135–139, Karlskrona, Swede, 2015. IEEE Computer Society. doi: 10.1109/ENIC.2015.27. **P9**.

21. Byrd, D. and Crawford, T. Problems of music information retrieval in the real world. *Information Processing & Management*, 38(2):249–272, 2002. doi: 10.1016/S0306-4573(01)00033-4.

22. Calefato, F., Iaffaldano, G., and Lanubile, F. Collaboration success factors in an online music community. In *Association for Computing Machinery GROUP*, Sanibel Island, USA, 2018. doi: 10.1145/3148330.3148346.

23. Chiru, C. and Popescu, O. Automatically determining the popularity of a song. In *Rough Sets - International Joint Conference*, pages 392–406, Olsztyn, Poland, 2017. Springer. doi: 10.1007/978-3-319-60837-2_33. **P10**.

24. Chon, S. H., Slaney, M., and Berger, J. Predicting success from music sales data: a statistical and adaptive approach. In *ACM Workshop on Audio and Music Computing Multimedia, AMCMM*, pages 83–88, Santa Barbara, CA, USA, 2006. ACM. ISBN 1595935010. doi: 10.1145/1178723.1178736. **P11**.

25. Corrêa, D. C. and Rodrigues, F. A. A survey on symbolic data-based music genre classification. *Expert Systems with Applications*, 60:190–210, apr 2016. doi: 10.1016/j.eswa.2016.04.008.

26. Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, sep 1995. doi: 10.1023/A:1022627411411.

27. Cosimato, A., Prisco, R. D., Guarino, A., Malandrino, D., Lettieri, N., Sorrentino, G., and Zaccagnino, R. The conundrum of success in music: Playing it or talking about it? *IEEE Access*, 7:123289–123298, 2019. doi: 10.1109/ACCESS.2019.2937743. **P12**.

28. Črepinšek, M., Liu, S.-H., and Mernik, M. Exploration and exploitation in evolutionary algorithms: A survey. *ACM computing surveys (CSUR)*, 45(3):1–33, 2013.

29. Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, pages 113–116, Washington, USA, 2002. IEEE Computer Society. doi: 10.1109/ICME.2002.1035731.

30. de Araújo Lima, R., de Sousa, R. C. C., Lopes, H., and Barbosa, S. D. J. Brazilian lyrics-based music genre classification using a BLSTM network. In Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., and Zurada, J. M., editors, *Artificial Intelligence and Soft Computing - 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part I*, volume 12415 of *Lecture Notes in Computer Science*, pages 525–534. Springer, 2020. doi: 10.1007/978-3-030-61401-0_49.

31. de Melo, G. B. V., Machado, A. F., and de Carvalho, L. R. Music consumption in Brazil: an analysis of streaming reproductions. *PragMATIZES - Revista Latino-Americana de Estudos em Cultura*, 10(19):141, 2020. doi: 10.22409/pragmatizes.v10i19.40565.

32. Dewan, S. and Ramaprasad, J. Social media, traditional media, and music sales. *MIS Quarterly*, 38(1):101–121, mar 2014. doi: 10.25300/MISQ/2014/38.1.05. **P13**.

33. Dhanaraj, R. and Logan, B. Automatic prediction of hit songs. In *International Conference on Music Information Retrieval*, pages 488–491, London, UK, 2005. ISMIR. **P14**.

34. Dhanaraj, R. and Logan, B. Automatic prediction of hit songs. In *ISMIR 2005, 6th International Conference on Music Information Retrieval*, pages 488–491, 2005.

35. Dyke, L. S. and Murphy, S. A. How We Define Success: A Qualitative Study of What Matters Most to Women and Men. *Sex Roles*, 55(5):357–371, September 2006. doi: 10.1007/s11199-006-9091-2.

36. Elovitz, H. S., Johnson, R. W., McHugh, A., and Shore, J. E. Automatic translation of english text to phonetics by means of letter-to-sound rules. Technical report, Naval Research Laboratory, Washington D.C., 1976.

37. Fan, J. and Casey, M. Study of chinese and uk hit songs prediction. In *Proceedings of International Symposium on Computer Music Multidisciplinary Research*, pages 640–652, Marseille, France, 2013. The Laboratory of Mechanics and Acoustics. **P15**.

38. Febirautami, L. R., Surjandari, I., and Laoh, E. Determining characteristics of popular local songs in indonesia's music market. In *International Conference on Information Science and Control Engineering*, pages 197–201, Zhengzhou, China, 2018. IEEE. doi: 10.1109/ICISCE.2018.00050. **P16**.

39. Fisher, C., Pearson, M. M., Goolsby, J. R., and Onken, M. H. Developing measurements of success for performing musical groups. *Journal of Services Marketing*, 24 (4):325–334, jul 2010. doi: 10.1108/08876041011053033.

40. Frieler, K., Abeßer, J., Zaddach, W.-G., and Pfleiderer, M. Introducing the jazzomat project and the melo (s) py library. In *International Workshop on Folk Music Analysis*, pages 76–78, Amsterdam, Utrecht, 2013.

41. Frieler, K., Jakubowski, K., and Müllensiefen, D. Is it the song and not the singer? hit song prediction using structural features of melodies. *Jahrbuch Musikpsychologie*, 25:41–54, dec 2015. **P17**.

42. Fu, Z., Lu, G., Ting, K. M., and Zhang, D. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319, dec 2011. doi: 10.1109/TMM.2010.2098858.

43. Gao, Andrea. Catching the earworm: Understanding streaming music popularity using machine learning models. *E3S Web Conf.*, 253:03024, 2021. doi: 10.1051/e3sconf/202125303024. **P18**.

44. Garimella, K. and West, R. Hot streaks on social media. In *International Conference on Web and Social Media*, pages 170–180, 2019.

45. Griffin, G. Defining success: African American women in the jazz industry, 1935-1965. Master's thesis, Texas Christian University, 2021.

46. Grills, S. et al. Generic social process and the problem of success-claiming: Defining success on the margins of Canadian federal politics. *Qualitative Sociology Review*, 18(3):54–69, 2022.

47. Harte, C., Sandler, M., and Gasser, M. Detecting harmonic change in musical audio. In *ACM Workshop on Audio and Music Computing Multimedia*, pages 21–26, Santa Barbara, USA, 2006. ACM. doi: 10.1145/1178723.1178727.

48. Hendricks, D., Patel, J., and Zeckhauser, R. Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988. *The Journal of finance*, 48(1): 93–130, 1993.

49. Herremans, D. and Bergmans, T. Hit song prediction based on early adopter data and audio features. In *International Society for Music Information Retrieval Conference, ISMIR - Late Breaking Demo*, Suzhou, China, 2017. ISMIR. **P18**.

50. Herremans, D., Martens, D., and Sörensen, K. Dance hit song prediction. *Journal of New Music Research*, 43(3):291–302, 2014. doi: 10.1080/09298215.2014.881888. **P19**.

51. Hirjee, H. and Brown, D. G. Rhyme analyzer: An analysis tool for rap lyrics. In *International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, 2010. ISMIR.

52. Hofmann, T. Probabilistic latent semantic indexing. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, USA, 1999. ACM. ISBN 1581130961. doi: 10.1145/312624.312649.

53. Holopainen, R. Making complex music with simple algorithms, is it even possible? *Revista Vórtex*, 9(2), 2021. doi: 10.33871/23179937.2021.9.2.4516.

54. Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., and Komarova, N. L. Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society open science*, 5(5):171274, may 2018. doi: 10.1098/rsos.171274. **P20**.

55. Janosov, M., Battiston, F., and Sinatra, R. Success and luck in creative careers. *The European Physical Journal Data Science*, 9(1):9, 2020. doi: 10.1140/epjds/s13688-020-00227-w.

56. Kamal, J., Priya, P., Anala, M. R., and Smitha, G. R. A classification based approach to the prediction of song popularity. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–5, Chennai, India, 2021. IEEE. doi: 10.1109/ICSES52305.2021.9633884. **P22**.

57. Kaneria, A. V., Rao, A. B., Aithal, S. G., and Pai, S. N. Prediction of song popularity using machine learning concepts. In K V, S. and Rao, K., editors, *Smart Sensors Measurements and Instrumentation*, pages 35–48, Singapore, 2021. Springer Singapore. ISBN 978-981-16-0336-5. **P23**.

58. Keogh, E. J. and Pazzani, M. J. Scaling up dynamic time warping for datamining applications. In *Special Interest Group on Knowledge Discovery and Data Mining*, pages 285–289. ACM, 2000. doi: 10.1145/347090.347153.

59. Kim, S. T. and Oh, J. H. Music intelligence: Granular data and prediction of top ten hit songs. *Decision Support Systems*, 145:113535, 2021. doi: https://doi.org/10.1016/j.dss.2021.113535. **P25**.

60. Kim, Y., Suh, B., and Lee, K. # nowplaying the future billboard: mining music listening behaviors of twitter users for hit song prediction. In *International Workshop on Social Media Retrieval and Analysis*, pages 51–56, Gold Coast, Australia, 2014. ACM. doi: 10.1145/2632188.2632206. **P21**.

61. Kischinhevsky, M., Vicente, E., and De Marchi, L. Em busca da música infinita: os serviços de streaming e os conflitos de interesse no mercado de conteúdos digitais. *Fronteiras-estudos midiáticos*, 17(3):302–311, 2015.

62. Kitchenham, B. and Charters, S. Guidelines for performing systematic literature reviews in software engineering. Technical report, Un of Durham, 2007.

63. Koenigstein, N., Shavitt, Y., and Zilberman, N. Predicting billboard success using data-mining in p2p networks. In *IEEE International Symposium on Multimedia*, pages 465–470, San Diego, USA, 2009. IEEE. doi: 10.1109/ISM.2009.73. **P22**.

64. Kong, Q., Rizoiu, M., Wu, S., and Xie, L. Will this video go viral: Explaining and predicting the popularity of youtube videos. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 175–178, 2018. doi: 10.1145/3184558.3186972.

65. Lee, J. and Lee, J.-S. Predicting music popularity patterns based on musical complexity and early stage popularity. In *Workshop on Speech, Language & Audio in Multimedia*, pages 3–6, Brisbane, Australia, 2015. ACM. doi: 10.1145/2802558.2814645. **P23**.

66. Lee, J. and Lee, J. Predicting music popularity patterns based on musical complexity and early stage popularity. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia, SLAM 2015, Brisbane, Australia, October 30, 2015*, pages 3–6, 2015. doi: 10.1145/2802558.2814645.

67. Lee, J. and Lee, J. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, 20(11):3173–3182, mar 2018. doi: 10.1109/TMM.2018.2820903. **P24**.

68. Lee, K., Park, J., Kim, I., and Choi, Y. Predicting movie success with machine learning techniques: ways to improve accuracy. *Information Systems Frontiers*, 20 (3):577–588, 2018. doi: 10.1007/s10796-016-9689-z.

69. Leikin, M.-A. *How to Write a Hit Song.* Hal Leonard, Milwaukee, USA, 5 edition, 2008.

70. Li, H., Tang, Z., Fei, X., Chao, K., Yang, M., and He, C. A survey of audio MIR systems, symbolic MIR systems and a music definition language demo-system. In *IEEE International Conference on e-Business Engineering*, pages 275–281, Shanghai, China, 2017. IEEE Computer Society. doi: 10.1109/ICEBE.2017.51.

71. Li, T., Ogihara, M., and Tzanetakis, G. *Music Data Mining.* CRC Press, Inc., USA, 1st edition, 2011. ISBN 1439835527.

72. Liu, L., Dehmamy, N., Chown, J., Giles, C. L., and Wang, D. Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nature Communications*, 12(1):5392, September 2021. doi: 10.1038/s41467-021-25477-8.

73. Liu, L. et al. Hot streaks in artistic, cultural, and scientific careers. *Nature*, 559 (7714):396–399, 2018. doi: 10.1038/s41586-018-0315-8.

74. Logan, B., Kositsky, A., and Moreno, P. J. Semantic analysis of song lyrics. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 827–830, Washington, USA, 2004. IEEE Computer Society. doi: 10.1109/ICME.2004.1394328.

75. Marchi, L. D. and Ladeira, J. M. Digitization of music and audio-visual industries in brazil: new actors and the challenges to cultural diversity. *Cahiers d'Outre-Mer*, 71(277):67–86, January 2018. doi: 10.4000/com.8716.

76. Maršık, L., Pokornỳ, J., and Ilcık, M. A survey on music retrieval systems using microphone input. In *Annual International Workshop on DAtabases, TExts, Specifications and Objects*, CEUR Workshop Proceedings, pages 131–140, Aachen, 2015. CEUR-WS.org.

77. Martens, D., Baesens, B., and Van Gestel, T. Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(2):178–191, July 2008. doi: 10.1109/TKDE.2008.131.

78. Martín-Gutiérrez, D., Hernández-Peñaloza, G., Belmonte-Hernández, A., and Álvarez, F. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8:39361–39374, 2020. doi: 10.1109/ACCESS.2020.2976033.

79. Martín-Gutiérrez, D., Hernández Peñaloza, G., Belmonte-Hernández, A., and Álvarez García, F. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8:39361–39374, feb 2020. doi: 10.1109/ACCESS.2020.2976033. **P25**.

80. Matsumoto, Y., Harakawa, R., Ogawa, T., and Haseyama, M. Context-aware network analysis of music streaming services for popularity estimation of artists. *IEEE Access*, 8:48673–48685, mar 2020. doi: 10.1109/ACCESS.2020.2978281. **P26**.

81. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. librosa: Audio and music signal analysis in python. In *Python in Science Conference*, pages 18–25, Austin, Texas, 2015. SciPy Organizers. doi: 10.25080/Majora-7b98e3ed-003.

82. Melchiorre, A. B. et al. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, 58(5):102666, 2021. doi: 10.1016/j.ipm.2021.102666.

83. Middlebrook, K. and Sheik, K. Song hit prediction: Predicting billboard hits using spotify data. *CoRR*, abs/1908.08609, 2019. **P27**.

84. Moghaddam, F. B., Elahi, M., Hosseini, R., Trattner, C., and Tkalcic, M. Predicting movie popularity and ratings with visual features. In *14th International Workshop on Semantic and Social Media Adaptation and Personalization*, pages 1–6, Larnaca, Cyprus, 2019. doi: 10.1109/SMAP.2019.8864912.

85. Murthy, Y. V. S. and Koolagudi, S. G. Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review. *ACM Computing Surveys*, 51(3):45:1–45:46, June 2018. doi: 10.1145/3177849.

86. Ni, Y., Santos-Rodriguez, R., Mcvicar, M., and De Bie, T. Hit song science once again a science? In *International Workshop on Machine Learning and Music*, pages 355–360, Sierra Nevada, Spain, 2011. ACM. **P28**.

87. Nunes, J. C. and Ordanini, A. I like the way it sounds: The influence of instrumentation on a pop song's place in the charts. *Musicae Scientiae*, 18:392–409, sep 2014. doi: 10.1177/1029864914548528. **P29**.

88. Oliveira, G. P., Silva, M. O., Seufitelli, D. B., Lacerda, A., and Moro, M. M. Detecting collaboration profiles in success-based music genre networks. In *International Society for Music Information Retrieval Conference*, pages 726–732, Montreal, Canada, 2020. ISMIR. **P30**.

89. Oliver, B. *How [Not] To Write A Hit Song!: 101 Common Mistakes to Avoid If You Want Songwriting Success*. CreateSpace Independent Publishing Platform, Scotts Valley, USA, 2013.

90. Pachet, F. Hit song science. In Tao Li, G. T., Mitsunori Ogihara, editor, *Music Data Mining*, chapter 10, pages 305–326. CRC Press, New York, USA, 2011.

91. Pachet, F. and Roy, P. Hit song science is not yet a science. In *International Conference on Music Information Retrieval, ISMIR*, pages 355–360, Philadelphia, USA, 2008. ISMIR. **P31**.

92. Passman, D. *All You Need to Know About the Music Business: 10th Edition*. Simon & Schuster, 2019. ISBN 978-1-5011-2218-7.

93. Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, Elsevier, San Francisco, USA, 1993. ISBN 1-55860-238-0.

94. Raab, M., Gula, B., and Gigerenzer, G. The hot hand exists in volleyball and is used for allocation decisions. *Journal of Experimental Psychology: Applied*, 18(1): 81, 2012.

95. Rabin, M. and Vayanos, D. The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, 77(2):730–778, 2010.

96. Rajyashree, R., Anand, A., Soni, Y., and Mahajan, H. Predicting hit music using midi features and machine learning. In *International Conference on Communication and Electronics Systems*, pages 94–98, Coimbatore, India, 2018. IEEE. doi: 10.1109/CESYS.2018.8724001. **P32**.

97. Raza, A. H. and Nanath, K. Predicting a hit song with machine learning: Is there an apriori secret formula? In *International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, pages 111–116, online, 2020. IEEE. doi: 10.1109/DATABIA50434.2020.9190613. **P33**.

98. Ren, J. and Kauffman, R. J. Understanding music track popularity in a social network. In *European Conference on Information Systems*, pages 374–388, Guimarães, Portugal, 2017. AIS. **P34**.

99. Ren, J., Shen, J., and Kauffman, R. J. What makes a music track popular in online social networks? In *International Conference Companion on World Wide Web*, pages 95–96, Montreal, Canada, 2016. doi: 10.1145/2872518.2889402. **P35**.

100. Ricker, C. Defining success. *American Music Teacher*, 66(1):12–14, 2016.

101. Rosati, D. P., Woolhouse, M. H., Bolker, B. M., and Earn, D. J. D. Modelling song popularity as a contagious process. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2253):20210457, 2021. doi: 10.1098/rspa.2021.0457. **P40**.

102. Ruggieri, S. Efficient c4. 5 [classification algorithm]. *IEEE Transactions on Knowledge and Data Engineering*, 14(2):438–444, mar 2002. doi: 10.1109/69.991727.

103. Salganik, M. J., Dodds, P. S., and Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, feb 2006. doi: 10.1126/science.1121066. **P36**.

104. Sandroni, C., Ferreira, D. M., Requião, L. P. d. S., Sandroni, C., and Lima, M. G. A covid-19 e seus efeitos na renda dos músicos brasileiros. *Revista Vórtex*, 9(1), 2021. doi: 10.33871/23179937.2021.9.1.4175.

105. Scaringella, N., Zoia, G., and Mlynek, D. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, apr 2006. doi: 10.1109/MSP.2006.1598089.

106. Seufitelli, D. B., Oliveira, G. P., Silva, M. O., Barbosa, G. R. G., Melo, B. C., Botelho, J. E., Melo-Gomes, L. d., and Moro, M. M. From compact discs to streaming: A comparison of eras within the brazilian market. *Revista Vórtex*, 10(1), 2022.

107. Shepard, R. N. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America*, 36(12):2346–2353, 1964. doi: 10.1121/1.1919362.

108. Shin, S. and Park, J. On-chart success dynamics of popular songs. *Advances in Complex Systems*, 21(3-4):1850008, 2018. doi: 10.1142/S021952591850008X. **P37**.

109. Shinohara, V., Foleiss, J., and Tavares, T. Comparing meta-classifiers for automatic music genre classification. In *Simpósio Brasileiro de Computação Musical*, pages 131–135, 2019.

110. Shulman, B., Sharma, A., and Cosley, D. Predictability of popularity: Gaps between prediction and understanding. In *International Conference on Web and Social Media*, pages 348–357, Cologne, Germany, 2016. AAAI Press. **P38**.

111. Silva, M. O. and Moro, M. M. Causality analysis between collaboration profiles and musical success. In *Brazilian Symposium on Multimedia and the Web*, page 369–376, Rio de Janeiro, Brazil, 2019. ACM. doi: 10.1145/3323503.3349549. **P39**.

112. Silva, M. O., de Alencar Rocha, L. M., and Moro, M. M. Collaboration profiles and their impact on musical success. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 2070–2077, Limassol, Cyprus, 2019. doi: 10.1145/3297280.3297483. **P45**.

113. Silva, M. O., Oliveira, G. P., Seufitelli, D. B., Lacerda, A., and Moro, M. M. Collaboration as a Driving Factor for Hit Song Classification. In *Proceedings of the 28th Brazillian Symposium on Multimedia and the Web*, page 66–74, Curitiba, Brazil, 2022. doi: 10.1145/3539637.3556993. **P46**.

114. Sinatra, R., Wang, D., Deville, P., Song, C., and Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science*, 354(6312), 2016.

115. Singhi, A. and Brown, D. G. Hit song detection using lyric features alone. In *International Society for Music Information Retrieval Conference (ISMIR): Late-Breaking Demo*, Taipei, Taiwan, 2014. ISMIR. **P41**.

116. Singhi, A. and Brown, D. G. Can song lyrics predict hits. In *International Symposium on Computer Music Multidisciplinary Research*, pages 457–471, Plymouth, UK, 2015. The Laboratory of Mechanics and Acoustics. **P42**.

117. Sturm, B. L. A survey of evaluation in music genre recognition. In *Adaptive Multimedia Retrieval*, pages 29–66, New York, USA, 2012. Springer. doi: 10.1007/978-3-319-12093-5_2.

118. Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K., and Woods, E. Tslearn a machine learning toolkit for time series data. *J.Mach.Learn.Res*, 21:118:1–118:6, 2020.

119. Teasdale, R. M. Defining success for a public library makerspace: Implications of participant-defined, individualized evaluative criteria. *Library & Information Science Research*, 42(4):101053, 2020. doi: 10.1016/j.lisr.2020.101053.

120. Trindade, I., Resendo, L., Andrade, J., and Komati, K. Análise das letras das músicas brasileiras mais tocadas nas rádios das Últimas seis décadas. In *Simpósio Brasileiro de Bancos de Dados Workshop de Tabalhos de Alunos de Graduação*, pages 1–7. SBC, 2021. doi: 10.5753/sbbd_estendido.2021.18155.

121. Tsiara, E. and Tjortjis, C. Using twitter to predict chart position for songs. In *IFIP Artificial Intelligence Applications and Innovations*, pages 62–72, Neos Marmaras, Greece, 2020. Springer. doi: 10.1007/978-3-030-49161-1_6. **P43**.

122. Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, nov 2002. doi: 10.1109/TSA.2002.800560.

123. Vapnik, V. The support vector method of function estimation. In Suykens, J. and Vandewalle, J., editors, *Nonlinear Modeling*, pages 55–85. Springer, New York, USA, 1998. doi: 10.1007/978-1-4615-5703-6_3.

124. Vötter, M. et al. Novel datasets for evaluating song popularity prediction tasks. In *IEEE International Symposium on Multimedia (ISM)*, pages 166–173. IEEE, 2021. doi: 10.1109/ISM52913.2021.00034.

125. Vötter, M., Mayerl, M., Specht, G., and Zangerle, E. Novel datasets for evaluating song popularity prediction tasks. In *IEEE International Symposium on Multimedia, ISM 2021, Virtual Event, November 29 – December 1, 2021*, pages 166–173. IEEE, 2021. doi: 10.1109/ISM52913.2021.00034. **P50**.

126. Vötter, M., Mayerl, M., Specht, G., and Zangerle, E. Hsp datasets: Insights on song popularity prediction. *International Journal of Semantic Computing*, pages 1–23, May 2022. doi: 10.1142/S1793351X22400104. **P51**.

127. Waldfogel, J. How digitization has created a golden age of music, movies, books, and television. *Journal of economic perspectives*, 31(3):195–214, 2017.

128. Witten, I. H., Frank, E., and Hall, M. A. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, Elsevier, San Francisco, USA, 3rd edition, 2011. ISBN 9780123748560.

129. Yang, L., Chou, S., Liu, J., Yang, Y., and Chen, Y. Revisiting the problem of audio-based hit song prediction using convolutional neural networks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 621–625, New Orleans, USA, 2017. IEEE. doi: 10.1109/ICASSP.2017.7952230. **P44**.

130. Yoo, B. and Kim, K. Online music ranking service: Ranking mechanism based on popularity and slot effect. In *Pacific Asia Conference on Information Systems*, pages 615–626, Taipei, Taiwan, 2010. AISeL. **P45**.

131. Yu, H., Li, Y., Zhang, S., and Liang, C. Popularity prediction for artists based on user songs dataset. In *Proceedings of International Conference on Computing and Artificial Intelligence*, pages 17–24, Bali, Indonesia, 2019. ACM. doi: 10.1145/3330482.3330493. **P46**.

132. Yucesoy, B. and Barabási, A.-L. Untangling performance from success. *The European Physical Journal Data Science*, 5(1):17, April 2016. doi: 10.1140/epjds/s13688-016-0079-z.

133. Yucesoy, B., Wang, X., Huang, J., and Barabási, A.-L. Success in books: a big data approach to bestsellers. *The European Physical Journal Data Science*, 7(1):7, April 2018. doi: 10.1140/epjds/s13688-018-0135-y.

134. Zampoglou, M. and Malamos, A. G. Music information retrieval in compressed audio files: A survey. *New Rev. Hypermedia Multimedia*, 20(3):189—-206, July 2014. doi: 10.1080/13614568.2014.889223.

135. Zangerle, E., Vötter, M., Huber, R., and Yang, Y. Hit song prediction: Leveraging low- and high-level audio features. In *International Society for Music Information Retrieval Conference*, pages 319–326, Delft, The Netherlands, 2019. ISMIR. **P47**.