Quantifying the Motown Sound:

Automatic Recording Studio and Record Label Detection in 1960's American Soul Music Through Supervised Learning

Karl Messerschmidt

Submitted in partial fulfillment of the requirements for the
Master of Music in Music Technology
in the Department of Music and Performing Arts Professions
Steinhardt School
New York University

Advisor: Dr. Brian McFee
Reader: Dr. Agnieszka Roginska

April 22, 2019

Abstract

Recordings from the mid-20th Century were made using modest equipment by today's standards. Their sonic characteristics are defined by the limitations of the technology, acoustics of the space, and musical ability and influences of the musicians, producers, and engineers behind the recordings. Certain recording studios and/or record labels, herein referred to as "Production Units," developed a recognizable "sound," or set of musical and timbral features that are in combination uniquely attributable to that unit. A set of songs recorded by five chosen Production Units between the late 1950's and early 1970's is extracted from the Million Song Dataset. Perceptual and musical features from those songs are used to build a learning representation, which is in turn used to train a classifier to predict the Production Unit. In light of particular challenges presented by the nature of the problem, all trained models generally average Accuracy and Macro Average F1 scores well above the Zero-Rule baseline score. Fine tuning of the system is performed by re-balancing the dataset and by removing outlying artists who prove to transcend studio of origin.

Acknowledgements

Thank you to my advisor Dr. Brian McFee for all his guidance through the most intense research project of my life. Thank you to all my professors at NYU, especially Dr. Juan Bello, Dr. Tae Hong Park, Dr. Schuyler Quackenbush, and DeAngela Duff for teaching me stuff I never imagined I'd get a chance to learn. Thank you to Eleanor Sparaccio for so patiently putting up with my complete inability to navigate the system at even the most basic level.

Thank you to my wife Rachel and my parents Fritz and Peggy for more things than I can ever list. Here's to the next chapter of life, when I will repay you all for everything you've given me these past four years.

Thank you to the movies "The Blues Brothers" and "Sister Act" for fostering in young Karl a life-long love and appreciation for this music. Thank you to Jeff Partridge, Arthur Hernandez, and Josh Hummel at Capital Community College for giving me the opportunity to teach my favorite subjects and to spend the past seven years researching the histories of these iconic studios and sharing them with our students. And speaking of our students, thanks to all of you for listening, learning, and inspiring me to keep at it.

Thank you to my friends who helped me survive the grad school experience: Matt Sargent for holding my hand through admissions; my brother and sister-in-law John and Helena, Allie Tedone, and Chris Przybycien for letting me crash on their couches for three summers so I wouldn't have to pay New York City rent; and Julie O'Leary for holding me accountable for finishing. Thank you to *Rick and Morty*, Pokemon Go, Arsenal Football Club, and every Smiths album for keeping me happy and amused while my life revolved around this work, and to New York City for being my favorite place in the world.

Finally, thank you to Berry Gordy, Jim Stewart and Estelle Axton, Leonard and Phil Chess, Rick Hall, Jerry Wexler, Arif Mardin, Muddy Waters, Big Willie Dixon, Smokey Robinson, Norman Whitfield, Holland-Dozier-Holland, David Porter, The Funk Brothers, The MG's, The Swampers, and all the men and women behind this music. I hope in some weird way that this work does your legacy proud.

And to everyone who's ever incorrectly included Aretha Franklin or Otis Redding in a playlist titled "Greatest Motown Hits," this is for you.

Table of Contents

List of Tables

List of Figures

1. Introduction

Soul and Rhythm & Blues music of the mid-20th century has had an immeasurable impact on American cultural output over the past century. Black American artists and composers of the era have forever shaped the course of popular music around the world despite often receiving little to no credit at the time due to oppressive segregation in the United States and to sometimes well-intended yet also damaging cultural appropriation in America in the 1950's and the UK in the 1960's. In contemporary times, the original artists have finally received a great deal of recognition from music fans and their legacy can no longer be denied. However, it can still be argued that some of their accomplishments have not been fully appreciated.

When one thinks of the most important artists of the 1960's, certain names inevitably come to mind: The Beatles, The Rolling Stones, Bob Dylan, Jimi Hendrix, Eric Clapton, etc. But for many, artists like The Supremes, Marvin Gaye, Muddy Waters, Otis Redding, and Aretha Franklin can take a second or even third thought. It's a curious situation, considering that from 1958 to 1969, only The Beatles (18) have more number 1 singles on the Billboard Hot 100 than The Supremes (12), who have as many as Elvis Presley (7) and The Rolling Stones (5) combined. In fact, only Capitol Records (20) has more number 1 singles during the same time period than Motown Records (16) - although it's a draw if Motown's Tamla (3) and Gordy (1) subsidiary labels are included (Billboard Charts Archive).

Motown Records, a small record label out of Detroit, Michigan, achieved such a high level of success in the 1960's that its uniquely distinguishable musical and sonic brand - The Motown Sound - has often been treated as musical genre unto itself. In 2011, Washington

University in conjunction with the National Endowment for the Humanities hosted a summer institute entitled "The Sock Hop and the Loft: Jazz, Motown, and the Transformation of American Culture 1959-1975." The event's Artistic Resource Guide includes literary sources referencing R&B artists Aretha Franklin, Curtis Mayfield, and James Brown, none of whom recorded for Motown Records. These inclusions, along with the event description, suggest Motown is used here in juxtaposition with the Jazz genre as a proxy for R&B and Soul music of the era in general (Brandon et al., 2011). In 2010, UK artist Craig David released the album *Signed Sealed Delivered*, a collection of covers of songs from the 1960's and 1970's by primarily Motown artists. David admitted in an interview with *The Mirror* that he "didn't actually know that Motown was a label... I thought it was an era or genre, like New Jack Swing or something. I didn't know that if you weren't on Motown Records, it wasn't Motown (Wightman, 2010, para. 3)." David isn't the only contemporary music star to have been confused. In 2009, *NME* asked a panel of famous artists to compile a list of the 50 greatest Motown tracks to celebrate the label's 50th anniversary. Caleb Followill of the band *Kings of Leon* nominated songs by both Sam Cooke and Otis Redding, neither of whom recorded for Motown (10 Incredible Motown Tracks, 2009).

The application of the Motown classification to all such music is akin to the phenomenon of Kleenex being such a ubiquitous brand of tissue that many consumers refer to all tissues as Kleenex[1] - in a sense a compliment to Motown's success, but also a discredit to the many other artists, recording studios, and labels who contributed to the American Soul lexicon. The situation

[1] Another example would be ordering a Coke at a diner that only serves Pepsi; many people will say "fine" to any cola product, but as a former soda drinker with discriminating taste, the author can attest to the fact that Pepsi is absolutely not Coke.

gives rise to the question: Is Motown truly unique in its distinctiveness, or can other labels and/or studios lay similar claim to a sonic category of their own?

It is here that a hypothesis will be made: That among a subset of American Soul records from the late 1960's until the early 1970's, there exists a unique sonic and musical quality beyond that which identifies the artist and musical genre that is common to recordings made by the same production unit, and that those qualities are discoverable via supervised machine learning using perceptual and musical features. The following describes an attempt to automatically classify by production unit a set of known recordings found in the Million Song Dataset (MSD) (Bertin-Mahieux, et al., 2011).

If this study is not measuring artist or genre, what exactly is being measured? The end goal is to identify a Production Unit, defined as the unique combination of people, technology, and environment used to make a given recording. In each of these Production Units, various factors beyond the artist and the overall song provided creative inputs to a process that led to an output. Because of the nature of recorded music, the inputs are lost to history without liner notes or other written histories to accompany the output. The goal then is to derive what the general combination of those inputs might be by examining only features of the outputs.

The creative outputs of these production units, in the form of recordings, carry with them musical, perceptual, and lyrical features that can inform us as to the nature of their creation. Although much has been written about the history of the people and places behind the creation of these recordings, that history is not conveyed to the listener when the recordings are played back. Only prior personal research into that history can allow the mind to access this historical knowledge in detail upon hearing a song. Via the Million Song Dataset, we have easy access to

the musical and sonic features of many of these songs. Although the lyrical features of the music are not presented, we can nonetheless explore what kind of knowledge, backed up by historical research, can be automatically extracted from these features using supervised learning. In a sort of reverse-creative process, this research will attempt to arrive at a song's original production unit by targeting features that describe a set of creative inputs to the recording. The following creative inputs have been identified:

- Personnel: The people behind the creation of the recording. While the artist is an obvious and heavily weighted inclusion, we should also consider the influence of the studio musicians, composers, arrangers, engineers, producers, and even possibly executives involved. The relative consistency of these auxiliary (non-artist) personnel within the selected Production Units during this time period is a strong reason for their selection (Cogan, et. al, 2003).
- Instrumentation: The instruments or combinations of instruments commonly used on recordings in particular ways by different production units. Many production units built a distinct sound around the use of specific instruments or combinations of instruments (Cogan et al., 2003). Examples include the prominent use of harmonica in many Chicago-based recordings; the very forward horn section of many Memphis productions; and the tandem guitars, multiple keyboards, and auxiliary percussion common in Detroit (Cogan et. al 2003; Licks, 1989; Gordon et al., 2007; Mayock, 2010).
- Space: The physical space in which the recording was made. The nature of recording during the selected time period predated the close-mic'ing techniques popularized in the 1970's that mitigated the sound of the room in the final recording. As such, microphones

were usually placed within a space in the room that would allow for optimal capture, making the sonic ambience often a discernible feature in the recording. It can also be hypothesized that the acoustic properties of the recording space, where a great many songs were regularly arranged and even composed outright, would influence the performance of the studio musicians and musical decisions of the arranger/composer to produce the most desirable sound (Cogan et al., 2003).

- Technology: While the live performances provided the subject of the recording, all recorded material had to pass through the lens of the technology available to each studio at the time, from microphones to console to tape machine and even reverb chambers. During this particular era it was not uncommon for important pieces of studio technology, including recording consoles, to be custom built as many commercial studio brands such as Neve and SSL were still a decade away. It can be hypothesized that each unique setup could leave a detectable sonic imprint on the final recording: a particular timbral effect of the room's acoustics, a dynamic range affected by a common use of compression or tape saturation, or a set of tempos or overall feel that may have been dictated by a favorable acoustic effect of the space. Additionally, some studios performed a historically noteworthy change in technology at various times, such as the installation of an eight track recorder at Motown or the upgrades to Atlantic and Stax studios by Tom Dowd (Cogan et al., 2003).
- Geography: In contrast to the dominance of international record charts today, popular music in the 1960's was still strongly driven by regionalism. While charting songs nationally was still the goal of each record company, regional cultures were strongly

reflected in the music of each geographic region and local charts were kept in cities such as Memphis, Nashville, Chicago, and Detroit. These regional musical cultures were reinforced when a unit from the region gained mainstream success (Cogan et al., 2003).

In the study that follows, a set of songs known to have been recorded by the selected production units during the selected time period is assembled from the Million Song Dataset based on historical research described in Chapter 2. Using the technical background described in Chapter 3, the data is extracted and prepared in Chapter 4 and the classifier is trained and tested. Chapter 5 provides an analysis of the results while Chapter 6 provides conclusions.

2. Historical Background

Historical context for this investigation is gathered in two forms: documented historical record in the form of books, articles, and documentary film; and the perceptual conclusions drawn from listening to the recordings themselves. Many years of listening experience have been used in describing the overall sounds attributed to each production unit and as such can be hard to document. As such, it is suggested that the original recordings by artists mentioned in-text be examined by interested parties. Specific details concerning the non-artist personnel, equipment, facilities, and anecdotes concerning all of the above are obtained from documented sources. Jim Cogan and William Clark's 2003 book "Temples of Sound: Inside the Great Recording Studios" is a survey of the history behind 15 prominent recording studios in the United states between approximately 1950 and 1980. The book includes detailed historical information about Motown Records, Stax Records, Chess Records, and Atlantic Records during the relevant time period. Cogan and Clark's research provides details regarding the personnel, space, and equipment featured in each studio. While the work does not provide an exhaustive list of artists with recorded for each Production Unit, album credits and information gathered from other sources mentioned herein were used to establish a time period during which each selected artist and Production Unit were connected.

2.1 Reasons for Selection of Time Period

The time period examined stretches from approximately 1957 to 1972, beginning when Chess Records moved into its new offices and studio space at 2120 South Michigan Avenue and ending when Motown Records moved its operations to Los Angeles. This time period provides for control of several conditions.

Because recording technology was limited compared to what is currently available, recording methods were simpler and more consistent across most studios. During most of the period chosen for examination, fewer than eight tracks were generally available for recording at most North American studios. In fact, until 1964 only Capitol and Atlantic Records had access to commercial eight track technology. Most of the selected studios found themselves working with three or fewer tracks prior to the mid 1960's and as such developed a recording style based around a live performance in a single room with minimal microphones and no isolation. Both the sound of the room and, crucially, the playing style of the studio musicians contributed heavily to every record. Rooms of the era were designed for acoustics that would enhance the music that would be performed. The 1970's brought on higher and higher track counts resulting in the construction of acoustically dead recording spaces that would allow for detailed processing of individual tracks. The increased use of outboard processing in the ensuing decades would hypothetically make the room and equipment a less discernible input (Cogan et al., 2003).

Another feature of the industry prior to the 1970's was that many record labels were vertically aligned business entities that carried out all levels of production in-house. While the modern recording business revolves heavily around outsourcing (different studios for each

project, freelance producers and engineers, etc), even smaller labels of the era would maintain their own recording facilities and staff them with in-house engineers, producers, A&R people, and office staff. It can reasonably be hypothesized that this insular arrangement would result in the development of a more consistent artistic style as the same people and resources would work together rather exclusively over a longer period of time (Cogan et al., 2003).

There was even some degree uniformity in the final output amongst all the selected studios, especially in the early half of the time period. When the standard for stereo LP's was adopted by the RIAA in November of 1957, it was at first marketed mostly as a high-end product for classical and highly commercial pop recordings. The selected music, marketed primarily consumers who were unlikely to invest in the latest and greatest stereo equipment, was mixed primarily in mono until stereo consumer technology became more commonplace in the mid to late 1960's (Cogan, 2003).

2.2 Selected Production Units and Criteria for Selection

In order to select the pool of Production Units included in this research, the following considerations had to be made:

- Each Production Unit must have produced a sufficiently large number of recordings representing a well-recognized contribution to the American Soul Music canon.
- Each production unit should have operated primarily within a single known recording environment during that time frame.
- The set of recordings made by each production unit represents a sufficient variety of artists.

- Finally, for the sake of practicality in the case of this study, that set of recordings would contain material selected for inclusion in the Million Song Dataset.

Based on historical research and examination of the MSD, the following production units were selected for this investigation: Motown Records, Stax Records, Chess Records, FAME Studios, and Atlantic Records. While these five production units do not by any means represent all of American soul music from their era, they each represent a significant contribution to the genre's lexicon.

2.2.1 Motown Records

Motown Records was founded by Berry Gordy in Detroit, Michigan. From its inception (originally as Tamla Records) in 1959 until its move to Los Angeles in 1972, Motown achieved unprecedented popular success as a black owned label during the height of the American civil rights movement. Gordy organized Motown according to the principles of mass production that he learned while working on the assembly line at Lincoln-Mercury (Cogan et al., 2003). He employed a highly structured division of labor, each with its specific role in the process of producing a hit song. The process itself was one that he played a very active role in cultivating, especially during the early years of the label. Everyone involved in production was employed directly by Motown, including the artists, songwriters, arrangers, producers, studio musicians, and engineers.

While the writers, arrangers, and producers of Motown (among the most prolific were Smokey Robinson, Norman Whitfield, and the team of Holland-Dozier-Holland) played an important role in crafting what came to be known as "The Motown Sound," none were more

influential than the group of first-call session musicians who would play on nearly all of the 79 nation-wide top 10 singles of the Detroit era. “The Funk Brothers,” as the group dubbed itself, was comprised of a relatively steady lineup of musicians who each brought specific elements to the Motown Sound over many years, and who would often play an important role in arranging and even editing the composition of many songs. Gordy’s demand that a minimum of four songs be produced in every three-hour recording session (sometimes as many as four sessions were held in a day) often required songs to be be delivered to the studio without written parts. The Funk Brothers, under the direction of band leader Earl Van Dyke, would routinely arrange parts for the song on the spot just before recording.

As specialization was one of the hallmarks of Motown’s operational model, parts were often arranged based on the specialty of each musician, especially among those who shared duties on the same instrument. Among Motown’s three principle guitarists (Robert White, Eddie Willis, and Joe Messina), Messina’s trademark specialty was an accented strum on beats 2 and 4 that doubled with the snare drum, one of the hallmarks of the Motown Sound (visible in Fig. 2.1), while Willis was often responsible for fills and counter-rhythms. White specialized in more legato chords and strums as well as distinctive lead melodies such as the famous C major pentatonic scale at the beginning of The Temptations’ first number 1 single, “My Girl” from 1965 (Licks, 1989). Motown’s principle drummer in its early days was Benny Benjamin. Although he was increasingly replaced by fellow Funk Brothers Richard “Pistol” Allen and Uriel Jones due to recurring struggles with alcoholism prior to his death in 1969, many Motown producers would not schedule a session unless Benjamin was available. Benjamin did, however,

have one noted weakness: he struggled to play a steady shuffle pattern. For songs that required a shuffle, Allen was used a 6/8 specialist (Licks, 1989).

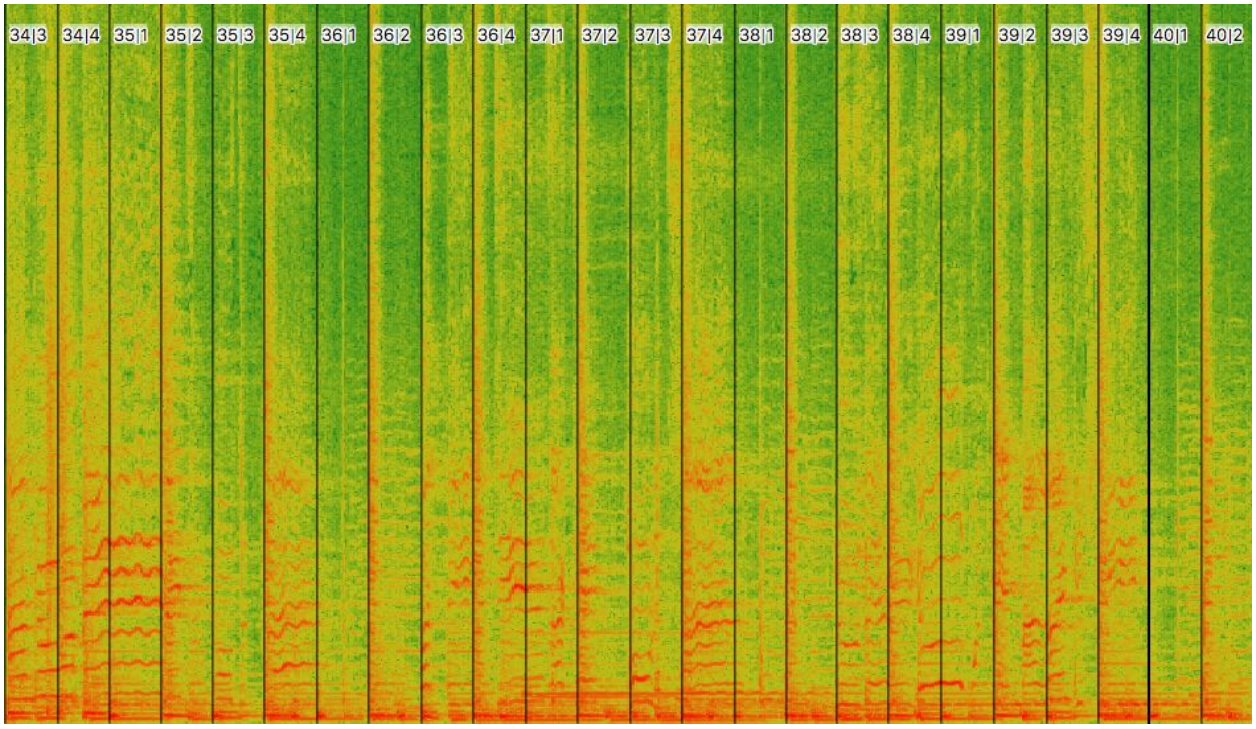

Figure 2.1. Spectrogram representation of "I Heard It Through the Grapevine," Marvin Gaye (1968). Black bands represent beat locations. The yellow vertical bands on beats 2 and 4 show snare drum hits doubled with guitar.

Perhaps no Motown musician was more critical to the signature Motown Sound than the group's principal bassist, James Jamerson. Though he was relatively unknown outside the Detroit music scene during his lifetime, much has been written about the influence of Jamerson's playing in recent decades. Although he learned to play on upright bass and continued to play and cherish the instrument his entire life, Jamerson is best known for his exclusive use of the Fender Precision bass (famously strung with LaBella flatwound strings and with the factory installed

foam mute, which many players removed, left in place) (Licks, 1989). While Jamerson was far from the only bassist to use the most popular electric bass of the era (and, arguably, of the current era as well), he took advantage of the electric instrument's faster recovery time versus its large acoustic counterpart to produce some of the most harmonically and rhythmically intricate bass lines in pop music history. While his contemporary counterparts such as Donald "Duck" Dunn (Stax) and Willie Dixon (Chess) were generally tasked with keeping a steady rhythmic foundation without taking any risks against the harmony provided by the chordal instruments and vocals, Jamerson was all too eager to explore outside the prevailing chordal and rhythmic structures common to pop. He was once famously chastised during a session by Gordy for playing off the beat. According to Gordy, after Jamerson had been given a final warning to stay on the downbeat, he waited until Gordy's back was turned during a take and played a series of syncopated notes. As soon as Gordy noticed and turned to chastise Jamerson, he had already returned to the downbeat, such was his quick thinking and prowess with the instrument (Cogan et al., 2003).

Nearly all of the recording during the Detroit era happened at Motown's headquarters, a converted house at 2648 West Grand Boulevard that Gordy dubbed "Hitsville, U.S.A." The recording studio, known officially as Studio A and unofficially as "The Snake Pit" (due to the large number of cables constantly traversing the floor), was converted from a garage space in the rear of the house that had previously been used as a photography studio. Gordy described the sound as "a thin, somewhat distorted sound with a heavy bottom (Cogan et al., 2003, p. 146)." The room was not particularly large by modern standards, resulting in very little sonic isolation. This was not an issue in Motown's early days as all recordings were done live to either two

(1959-1962) or three track tape (1962-1964). In 1964 Berry Gordy approved a massive and ambitious undertaking: the installation of an eight track tape machine in Studio A, making it only the third professional studio in the world with eight track capability (Cogan, 2003). With the installation of the eight track, Motown was capable of creating productions on par with some of the largest studios in the country. The machine, built and installed by chief engineer Mike McClain, made its debut on the session for The Supremes' 1964 hit "Baby Love." The ability to separate instruments onto individual tracks also produced a much clearer mix than previous Motown records were known for. This was perhaps most apparent in the bass. Whereas Jamerson had to share a single track with the rest of the band for most of Motown's history up until that point, the engineers were now able to reserve a track for the bass by itself. This separation allowed for individual compression and equalization of the instrument eliminating the somewhat indistinct low end of the early Motown recordings. Jamerson was now free to play more rhythmically and harmonically innovative parts that could be heard clearly in the final mix (Licks, 1989).

2.2.2 Stax Records

Stax records was originally founded as Satellite Records in 1957 by Jim Stewart in Memphis, Tennessee. Heavily influenced by Sam Phillips of Sun records (also of Memphis), Stewart initially set out to start a rockabilly and country swing-oriented label. However, Stax would go on to produce some of the most iconic Soul and R&B records of the 1960's and forged perhaps *the* definitive sound known as Memphis Soul. In a segregated city, it boldly stood out as a fully racially integrated operation (Gordon et al., 2007).

Satellite made its start in an old general store 30 miles northeast of Memphis in Brunswick, TN. The fledgling operation managed to get a couple records picked up by Mercury Records, but the space never proved adequate. In 1959 Stewart began the search for a new location in Memphis itself, specifically searching for a space that was already built for acoustics. His second in command, Chips Moman, found what would ultimately become the label's iconic new home: the Capitol Theater at 926 McLemore Avenue. Moman was a fan of Soul and Blues music and specifically sought out a location in a black neighborhood. Although initially met with suspicion, the studio made efforts to be an active member of the community and to invite local residents in. These community ties were forged thanks to two initiatives: Jim Stewart's open audition policy, modeled after Sam Phillips' similar policy at Sun Records, which allowed anyone to come and be heard; and his sister and business partner Estelle Axton's record shop in the studio lobby, which not only brought in youth from the neighborhood but also gave the Stax team first-hand insight into what sounds were popular (Cogan et al., 2003; Gordon et al. 2007).

Several of the neighborhood youth who first walked through the doors to browse the record shop would go on to become the talent that crafted the Stax Sound over the following decade. Keyboardist Booker T. Jones, guitarist and producer Steve Cropper, and bassists Lewie Steinberg (1962-1964) and Donald "Duck" Dunn (1964-1975), and songwriter David Porter were all recruited from the record shop. Along with drummer Al Jackson, Jr., who Jones recruited from a local club, they formed the rhythm section known as The MG's and would play on the majority of Stax's hit records from 1962 to 1972. The MG's were regularly joined by the Memphis Horns, led by fellow Mar-Keys alumnus Wayne Jackson. In 1964 Isaac Hayes joined the MG's while Jones was studying music at Indiana University and also became an integral part

of the Stax Sound. Jones, Cropper, Dunn, Jackson, Hayes, and Porter came to be known as "The Big Six" at Stax, such was the scope of their contributions (Cogan et al., 2003; Gordon et al. 2007).

Based on the success of Stax's (then Satellite's) first hit, "'Cause I Love You" by Rufus and Carla Thomas (1960), Jerry Wexler of Atlantic Records made a handshake deal with Stewart granting Atlantic right of first refusal to distribute all future Stax releases. Stewart was happy to agree as he was not interested in the manufacturing and distribution end of the business. From 1960 until 1968, Wexler would reciprocate by sending many Atlantic artists (such as Sam and Dave) to Stax "on loan" when he felt the specific Stax Sound would be a good fit. The relationship dissolved in 1968 when Atlantic Records was sold to Warner. Wexler had preemptively sent Stewart a written agreement to sign to preserve the relationship with the new owners, which Stewart did without reading. It was eventually discovered that the agreement included a clause that all of Stax's recordings that had been distributed by Atlantic would become property of Atlantic should Stax decide to terminate the agreement. Deeming the terms offered by the new ownership unfavorable, Stewart terminated the agreement and to this day nearly the entire Stax catalog prior to 1968 is property of Atlantic Records (Cogan et al., 2003).

The Capitol Theater was a former movie theater and as such had favorable acoustics for recording and good isolation of outside sound. It also came with odd quirks, such as the slanted concrete floor. The old projection booth was converted into the studio's control room. Playback monitoring was done through one of the cinema's giant old loudspeakers[2]. The mixes out of Stax featured more up-front instrumentation than their Motown counterparts, with a heavy emphasis

[2] A single speaker was sufficient in the early days as Stax's recordings were still being mixed in mono.

on Al Jackson's snare drum[3] and Dunn's bass. A common criticism from Wexler was that Stax mixed their vocals too low, preferring a more true to life balance of instrumentation and voice. It is perhaps here that Stax stands in its starkest contrast to the Motown Sound, which favored a loud and forward vocal sound (Cogan et al., 2003). The prolific output of the two labels and their contrasting styles make them an ideal early test case to show whether Production Units are in fact classifiable.

Just like its personnel, the equipment at Stax was also incredibly consistent over the years. The studio's centerpiece from the beginning was an Ampex 350 tape recorder. Stewart used the recorder and the rest of his equipment for the first several years of Stax's existence without any familiarity with the maintenance requirements. By 1963, the equipment had fallen into such a state of disrepair that the Stax team found itself unable to record. Wexler sent legendary Atlantic engineer Tom Dowd to Memphis to investigate. Dowd performed a complete overhaul and calibration of the equipment and Stewart and Cropper found that it sounded better than it ever had before. The same day, Rufus Thomas recorded "Walkin' the Dog," a top 10 pop and R&B hit and the first song to be recorded on Stax's now properly set up equipment (Cogan et al., 2003; Gordon et al. 2007).

2.2.3 Chess Records

Chess Records was founded in Chicago by brothers Phil and Leonard Chess. Operating under the Chess brothers' ownership from 1950 until the death of Leonard Chess in 1969. The label had existed as Aristocrat since 1947 and continued to operate under new owners General Recorded Tape (GRT) until being purchased by Sylvia Robinson's All Platinum Records in 1975

[3] Jackson's signature snare sound was achieved by placing his wallet on the rim of the drum.

and then by MCA in the 1985. However it was under Leonard and Phil that Chess produced some of the most influential records in 20th century popular music (Cogan et al., 2003).

In contrast to Motown and Stax, Chess was a label that sometimes struggled with its own identity, and conflict permeated its very existence. Its sound and style remain harder to define than some of its contemporaries. The label was known to record everything from jazz to Etta James' ballads to Fontella Bass's "post-Motown" R&B hit "Rescue Me (1965)." While it is most famously associated with its legendary studio at 2120 South Michigan Ave, Chess recorded at multiple locations throughout its existence and often accepted records from outside studios and labels including Sam Phillips' Sun Records and even the early days of Motown. The cast of studio musicians had its staples such as bassist Willie Dixon and pianist Johnnie Johnson (many of whom were Chess artists aside from their regular jobs as sidemen), but it rotated personnel more frequently (Cogan et al., 2003; Mayock, 2010).

Yet while Chess's artist roster was quite broad and diverse in comparison to Motown or Stax, the sound it primarily came to be known for was Chicago Blues, an electrified variant of the Delta Blues that spread up the Mississippi River in the early 20th century as part of the Great Migration. Chicago at that time was home to the American meat packing industry and demand for labor was high. Bereft of economic opportunity due to segregation, black workers migrated north to cities such as Chicago where labor could earn one a middle class lifestyle. With the rise of a black middle class came disposable income, and with disposable income came a booming leisure and entertainment industry, which provided opportunities for musicians. Many came from the thriving blues scene of Clarksdale, Mississippi. One of these was McKinley Morganfield, better known to the world as Muddy Waters (Mayock, 2010).

Muddy Waters was raised on Stovall Plantation, home to a community of sharecroppers and musicians that included Delta blues legends Son House and Robert Johnson. Waters grew up studying guitar under House and Johnson and in 1943 moved north to Chicago where he spent the next several years playing in Chicago's club scene. Chicago clubs and bars proved to be much louder than what Waters was accustomed to in Clarksdale, and in response he switched his acoustic guitar - the traditional instrument of the Delta blues style - for an electric guitar. Waters' electrified Delta blues would form the foundation of the Chicago blues style (Mayock, 2010).

Waters cut "I Can't Be Satisfied," his first record for Leonard Chess when he was co-owner of Aristocrat in 1948. He would serve as one of Chess's most successful and trusted artists for the rest of Leonard's life. He along with electrified harmonicist Little Walter, bassist and songwriter Willie Dixon, and fellow artist Chester Burnett - best known as Howlin' Wolf - formed a core group who would serve as a referral network for new Chess artists. In contrast to the open door policy of Sam Phillips at Sun Records (who passed along many records for distribution by Chess), Chess operated strictly on a referral system, relying on current artists identify new talent among their peers. It was through this network that the label would bring in groundbreaking artists of the era including Chuck Berry, Bo Diddley, and Buddy Guy (Cogan et al., 2003; Mayock 2010).

Although Chess recorded at several locations throughout its history (including Universal Recording), its most famous location was at 2120 South Michigan Avenue in Chicago's "Record Row" neighborhood. Abandoned at the time of purchase, Chess operated out of the recording studio on the second floor from 1957 until 1967. The room was designed and built by Jack Weiner, a protege of Universal Recording's legendary owner and chief engineer Bill Putnam.

Weiner built a false floor of concrete floated on cork to isolate vibrations from the street and installed a set of reversible panels on the south wall. One of the sides of the panels were coated in absorptive material to produce an acoustically "dead" sound, while the other side was reflective to produce a more "live" sound. As such, the room had a somewhat chameleonic nature, much like the label itself (Cogan et al., 2003).

2.2.4 FAME Studios

Florence Alabama Music Enterprises (FAME) Studios was founded in Florence, Alabama in 1959 by the team of Rick Hall and partners Billy Sherrill, and Tom Stafford. By 1960 Hall had become sole owner of FAME and the following year recorded the label's first hit and the first hit to come out of northern Alabama, "You Better Move On" by Arthur Alexander. Hall used the money earned from the record to build a new facility on East Avalon Avenue in neighboring Muscle Shoals where FAME Recording Studios continues to operate to this day (Our History, n.d.).

Located only 150 miles from Memphis, Muscle Shoals enjoyed its own rich musical history in the early 20th century. Sam Phillips, the legendary founder of Sun Records, was born in Florence and worked his first job in music as a DJ at WLAY in Muscle Shoals. WLAY had the rare distinction of being an integrated label in the deeply segregated South, playing both country and blues music (Cogan et al., 2003). The station was instrumental in the development of a strong local music scene that saw the construction of many recording studios in the area, but FAME was the first to achieve major success. Rick Hall attributed this early success to following WLAY's lead of ignoring segregationist policies and opening the studio to both white and black

musicians alike, a controversial decision at the time that would also pay huge dividends for both Sam Phillips and Jim Stewart in Memphis (Pareles, 2018; Cogan et al., 2003).

While Hall did operate a FAME Records label from 1964 until 1974, the studio primarily grew to prominence by recording for more established labels. FAME would most famously record Aretha Franklin, Wilson Pickett, Clarence Carter, and The Tams for Atlantic Records; Etta James for Chess; Joe Tex for Dial; and Tommy Roe for ABC-Paramount (Our History). Perhaps FAME's greatest asset was its house band, known officially as The Muscle Shoals Rhythm Section, but more commonly as The Swampers. The core group composed of keyboardist Barry Beckett, guitarist Jimmy Johnson, bassist David Hood, and drummer Roger Hawkins, the Swampers performed on nearly every FAME hit until 1969 as well as for other studios in the area (Cogan et al., 2003). The group also included at various times organist Spooner Oldham, bassists Tommy Cogbill and Jerry Jemmott, and guitarist Pete Carr. In the late 1960's, the group was also regularly joined by guitarist Duane Allman who had begun living in a tent in the studio parking lot until he was invited in after befriending Rick Hall and Clarence Carter. Allman's guitar work for Pickett famously caught the attention of Atlantic Records who signed him to an artist contract from which the Allman Brothers Band was born. In 1969 with help from Jerry Wexler of Atlantic Records, The Swampers left FAME to open their own studio, Muscle Shoals Sound Studios (Cogan et al., 2003 Pareles, 2018).

2.2.5 Atlantic Records

Founded in New York City in 1947 by Ahmet Ertegun and Herb Abramson, Atlantic Records was an important early proponent of Rock and Roll. In the 1950's Atlantic was home to

legendary artists including Ray Charles, Big Joe Turner, Clyde McPhatter, The Drifters, The Coasters, The Clovers, and Bobby Darin. Originally a jazz label, Atlantic artists and producers built a sound that incorporated blues, country, and gospel that would see it grow into one of the biggest labels in the United States in the ensuring decade. After Ray Charles left for ABC-Paramount in 1959, Atlantic's biggest draw for several years was its distribution deal with Stax Records. In 1964 Atlantic bought the contract of Wilson Pickett from Double L, although Pickett preferred to record in the Memphis area (notably at FAME). In 1966 it signed Aretha Franklin from Columbia (Cogan et al., 2003).

Perhaps the most game-changing acquisition Atlantic made was Jerry Wexler, a former Billboard Magazine writer who became partner at Atlantic 1953. Wexler was instrumental in establishing Atlantic as one of the premier labels for R&B and soul music. Wexler's understanding of soul music made it an attractive destination for top artists in the genre who were ready for the national stage. When Aretha Franklin left Columbia Records for Atlantic in 1966, legendary Columbia producer John Hammond opined that Wexler would do a better job than he had at getting the best out of Franklin. He was also a close collaborator with Jim Stewart at Stax and Rick Hall at FAME[4]. (Cogan et al., 2003; Pareles, 2018).

As important as Wexler and Ertegun were to shaping the musical qualities of Atlantic Records, the sonic characteristics revolved primarily around one man: Engineer Tom Dowd. Dowd was a classical musician and a physics student who was assigned to the Manhattan Project after being drafted into the army. Unable to obtain his degree in nuclear physics due to the top secret nature of is prior work, he took a job as a recording engineer and eventually began

[4] Wexler would eventually double-cross both, tricking Stewart into signing away Stax's catalog and bankrolling Hall's house band to open its own rival studio nearby after Hall signed a deal with Columbia.

freelancing for Atlantic in 1949 (Tom Dowd, n.d.). He was hired full time in 1954, although the hire was a mere formality as he had recorded nearly every session Atlantic had done as a freelancer. In those early days Dowd had very little equipment to work with and recorded most sessions using a pair of portable RCA radio broadcast mixers. At his direction in 1951 Atlantic switched from the still-standard practice of cutting records directly to acetate to the new technology of magnetic tape, resulting in highly improved audio quality. While the studio's primary recorder was a single-track, state of the art Ampex 400, Dowd would simultaneously record on his own Magnacord stereo tape recorder starting in 1952. Although stereo records would not become a commercial standard until 1958, he strongly believed that stereo was the future and his early investment proved to be a smart one (Cogan et al., 2003; Tom Dowd, n.d.).

Atlantic was forced to move several times in its early days but its primary home in the 1950's was at 234 West 56th Street in Manhattan. Prior to 1954, the offices and recording studio were one and the same, with desks and chairs being stacked against the wall during recording sessions. In 1954 when the offices were moved a block away to West 57th Street, Dowd was given free reign to redesign the studio. In 1957 he convinced Ertegun and Wexler to buy an eight track recorder, making Atlantic the first recording studio to obtain the new technology. In 1960 Atlantic opened its new state of the art studio at 11 West 60th Street (Cogan, 2003).

3. Technical Background

The technical aspects of this examination involves the identification of songs by a predetermined set of artists in the Million Song Dataset, the retrieval of perceptual and musical features from the dataset, building of a learning representation, and training a classifier to predict the Production Unit that created each recording based on that representation. Because genre and artist classification are the closest common tasks to what is attempted here, an overview of those two challenges in provided followed by a discussion of how the task at hand compares and contrasts with these more commonly covered problems. Background information on the dataset and supervised machine learning are also provided.

3.1 Music Information Retrieval

Music Information Retrieval (MIR) deals with (a) extracting descriptive features and other information from an audio recording through computational analysis, and (b) implementing systems to derive conclusions about the content or perceptual nature of the recording, usually through machine learning. Common tasks include artist (Whitman et al., 2001) and genre identification (Tzanetakis et al., 2002; Li et al., 2003), lyric transcription (Annamaria et al., 2010), similarity measurements (Whitman et al., 2001), and recommendation systems (Song, 2012). Maintaining metadata has historically been a manual task and despite being carried out by a human, errors and omissions are still common in manually maintained systems. As musical databases continue to grow in size and scope, manual maintenance becomes impractical or

impossible. Robust, reliable MIR methods provide scalable solutions for future maintenance of metadata (Foote, 1997; Whitman, et al., 2001).

3.2 Genre Classification in MIR

The problem of genre classification can provide useful insight into identifying and categorizing music based on perceptual and musical criteria. Musical genres are largely subjective labels for categorizing music. Their general function is to provide some criteria for locating music in record stores, streaming catalogs, and more. When classifying music by genre, humans have traditionally relied on descriptions of the texture, instrumentation, and rhythmic structure of the music to determine a class (Tzanetakis, 2002). Cultural and historical considerations are often made as well, but are not so easily measured. A well-versed listener to a particular genre can draw even finer distinctions between songs in a single genre to classify songs into subgenres. If a uniquely identifiable artistic or sonic style does exist for a given production unit, it would be perceptible through very similar processes. Absent historical knowledge research from sources outside the recording itself, these perceptual criteria would be the only way a listener could try to guess the origin of the recording.

Tzanetakis (2002) describes a system for genre classification using a feature set similar to that found in the MSD. A feature vector was assembled from (a) the standard deviation and mean of Spectral Centroid, Rolloff, Spectral Flux, and Zero Crossing Rate, plus the percentage of analysis windows containing less than average energy, and (b) a set of rhythmic features based on a "beat histogram," examining the period, amplitude, and ratio between the first three peaks. For classical and speech genres, the means and standard deviations of the first five Mel

Frequency Cepstral Coefficients were used. Using fifteen 30 second clips for each of fifteen musical genres, a Gaussian classifier was trained 100 times. Based on the given confusion matrix, trained classifier was 62.9% accurate overall. Li (2003) shows that both Support Vector Machines and Linear Discriminant Analysis provided significantly better performance on the same methods and dataset.

In the years since Tzanetakis and Cook's 2002 publication, often cited as the beginning of research into automatic genre recognition, much further work has been done. A thorough overview of the evaluation methods of 467 publications can be found in Sturm (2012). Pampalk et al. (2005) identified the artist and album effect, in which a classifier which is intended to detect genre will instead detect characteristics of songs by the same artist or taken from the same album. One of the principal conclusions is that the songs from the same artist and/or album should not exist in both the training and testing sets. Seyerlehner et al. (2010) is the first known study to compare the performance of several previously devised systems to a group of human subjects' performances at the task of genre classification, finding that even the most advanced systems tested performed on average about 10 percentage points worse than humans. Sturm (2014) discusses the issue of validity in the evaluation of genre recognition systems, arguing that in many systems there is no way to know that genre is truly what is being recognized as it is unknown whether the system is using relevant criteria.

3.3 Artist Classification in MIR

A musical artist is defined as the original creator and performer of a price of recorded music. Aside from distinct perceptual features such as vocal timbre, artists tend to cultivate a

particular musical style that can be identified by a listener through familiarity. In contrast to genre classification, the solution involves a one-to-one (or possibly one-to-many) mapping of a song to a specific known creator rather than applying an arbitrary descriptor. Like genre classification systems, artist classification can employ both perceptual and musical features.

Whitman, Flake, and Lawrence (2001) describe the recognition system Minnowmatch which performs artist classification using perceptually motivated features to train both Neural Networks and Support Vector Machines. The system describes the process in five stages. Starting with the original audio as standard Pulse Code Modulation (PCM) (1), a perceptual representation is extracted (2). The Perceptual Representation is generally composed of the Discrete Fourier Transform and Mel Frequency Cepstral Coefficients of the audio. The perceptual representation is then used to build a learning representation (3) that describes the perceptual representation using minimal data to aid the learning algorithm in interpreting the features. The learning process (4) trains the model and the Output/Classification stage (5) tests the model on new data.

Whitman et al. (2001) also identify the problem of choosing an "Artist Space." Representations of an artist in the dataset should be chosen such as to diversify as many variables as possible with respect to an artist's catalog. If an artist has recorded in multiple genres or with multiple producers, for example, a dataset that does not contain a representation of all of these possibilities risks the classifier inadvertently learning the producer, genre, or album rather than the artist and would be unable to properly classify the artist under different circumstances.

3.4 Contrast with Artist and Genre Classification Problems

While the investigation presented herein does rely on several of the principles and techniques used in artist and genre classification, it is distinct from a purely artist or genre classification problem in significant ways. While genre labels are decided somewhat subjectively based on rather broad differences in musical features and can be somewhat arbitrary (Tzanetakis, 2002), this study seeks to establish an objective fact - the recording studio and/or record company that produced a given recording - based on musical and textural features of songs within the same broad genre of American soul music. It also seeks to quantifiably demonstrate the musical and sonic diversity that exists within soul music of the 1960's and support the hypothesis that the artistic style of a given production unit can be detected similarly to that of a genre or artist. Although many genre classification systems have proven quite effective, the chosen task poses certain challenges that would expectedly make classification more difficult. Because all the music is selected from the same overall genre, most of the differences in musical features that are traditionally used in genre classification should be expected to have a higher degree of similarity. As such, the differences between classes are expected to be much subtler than, for example, the characteristics of the genre labels used by Tzanetakis, et al (2002).

This investigation posits the existence of both a musical and perceptual footprint unique to each selected production unit similar to that which exists for an artist. However, while in artist classification the Artist Space is selected such that the footprints belonging to the artist are detected, here the objective will be to look for a classifiable set of features surrounding multiple artists who recorded in the same place at a similar time. While these artists are distinctly

classifiable between each other, the search here is for a "common thread." If such a common thread does exist, it should not carry over in the event that the same artist records for more than one of the selected production units. Whether an adequate Learning Representation can be built so as to teach the classifier to prioritize the production unit's footprint over the artist's will be one of the subjects of the investigation (Whitman et al., 2001).

3.5 The Million Song Dataset

The Million Song Dataset (MSD) is a database containing musical features and metadata for one million popular songs. It was compiled jointly by researchers at Columbia University and The Echo Nest and is made freely available. It contains data fields representing perceptual features such as timbre descriptors and loudness; musical features including chroma descriptors, tempo, key, beat tracking, and note onsets; and metadata including artist, album, year, geographic location, etc (Bertin-Mahieux et al., 2011). The MSD is selected because the nature of this investigation requires data from songs within a specific genre and time period. Since the MSD is currently the largest such dataset available, it should provide the highest probability of finding an adequate number of songs from each production unit that fit the necessary criteria.

3.6 Supervised Learning

Machine Learning is a branch of computer science that deals with problems for which it is impractical for a human to derive an algorithm that would work for all cases. Rather than attempt to account for all possible exceptions to a given rule for differentiating between two disjoint sets, one or more machine learning algorithms can be used to determine the best rule. In

a process called *training,* the algorithm is introduced to a dataset called the *training set* which contains a sufficiently large sample of data points which can be used to describe differentiable criteria pertaining to a single instance of the subject being investigated. The features in the dataset may be any combination of continuous, categorical, or binary types (Hastie, et al., 2008).

Supervised Learning is a branch of Machine Learning in which the algorithm is given a training set along with the desired result for each item. This allows algorithm to check its own accuracy against the correct outcome. It is then checked for accuracy using new *testing data* that (a) it has not encountered before, and (b) is not accompanied by a correct answer. Supervised Learning involves two types of learning: *Regression* and *Classification.* Regression is used to predict a continuous output, whereas Classification is used to predict a categorical response representing membership in one of a set of predetermined classes. Since this problem involves correctly predicting membership in a known class - records made by one of several selected production units - Classification will be used (Hastie, et al., 2008).

In preliminary training and testing of the collected data, all available learning algorithms were trained, including Support Vector Machines (SVM), Decision Trees, Discriminant Analysis, and k-Nearest Neighbors. SVM was selected as the learning algorithm of choice for this study based on superior training and testing results in these preliminary trials. SVM provides favorable results for music classification problems in Li (2003), Whitman (2001), and Schindler (2012). Support Vector Machines plot each data point in *n*-dimensional space where *n* is the number of features. For linearly separable data, the SVM calculates two margins, defined as the upper and lower bounds of a hyperplane that separated the two classes in space. The two margins are defined as the parallel lines in space that pass through the most extreme data points of each

class and are maximum distance from each other and from from the hyperplane. More precisely, for a weight vector *w* and displacement from the origin *b,* if the data is linearly separable then a pair (*w,b*) exists such that $y_i(wx_i + b) \geq 1$. The ideal hyperplane is found by finding the minimum magnitude of the weight vector that satisfies the above constraints by minimizing $\frac{1}{2}||w||^2$. Once the hyperplane is found, the points lying on the margins are labeled Support Vector Points. The solution to the classification problem is reduced to a linear combination of these points. Data classes that are not linearly separable may be separable by a quadratic, cubic, or other function (Cortes et al., 1995).

In practice, real world data is often not linearly separable. In such cases, a solution can be found by mapping the data to a higher-dimensional space called the *transformed feature space*. With an appropriate number of dimensions a hyperplane can be found, although the procedure becomes more computationally expensive with each dimension needed. In fact, since the transformation can be expressed as a mapping of the data to another Hilbert space H as $\Phi : Rd \rightarrow H$, the training algorithm depends on the dot product $\Phi(xi) \cdot \Phi(xj)$. The computation can be made more efficient by finding a kernel function *K* such that $K(xi,xj)=\Phi(xi) \cdot \Phi(xj)$. The kernel function allows the dot product to be calculated without having to perform the actual mapping, thus saving computational cost (Kotsiantis, 2007). A commonly used kernel function is the Gaussian, or Radial Basis Function (RBF) kernel. The Gaussian kernel is defined as:

$$K(x, y) = exp\left(-\frac{||x-y||^2}{2\sigma^2}\right)$$

The kernel is essentially a measure of the distance between the two variables *x* and *y* without concern for direction. As distance increases, similarity declines exponentially. The value of σ can be adjusted to control how far apart *x* and *y* can be for the similarity to be too small to matter. As

σ approaches zero, less and less similarity can be found until all points are completely isolated. As σ goes to infinity, all points are increasingly likely to measure as similar. A Gaussian kernel with a small σ value is considered "Fine," while larger values are considered "Coarse." (Cortes et al., 1995)

However, as increasingly high dimensional spaces are used to try to separate the data, an increasingly high number of data points must be used to define the support vectors, leaving few points outside the margins. The result is a model that is closely defined by the specific training data and may not be able to make predictions about data points that don't fall along the margin. This problem of a decision function fit too closely to a set of limited data points is called *overfitting*. An overfit model will often train to high degree of accuracy but fail to test to the same standard when new data is introduced because the rule is not sufficiently generalized (Cortes et al., 1995).

Rather than rely on increasingly high spatial dimensions, the SVM can instead use a "soft margin" which allows for a certain amount of error, or *slack*, in the model. Slack, denoted ξ, is the amount of error allowed by the model and is calculated by summing the distance ξ_i from each data point not properly separated by the hyperplane to its appropriate margin. The total slack is then multiplied by a factor *C.* The optimal hyperplane becomes the minimization over *w* and ξ of the formula:

$$min_{w,\xi}\left(\tfrac{1}{2}\|w\|^2 + C\sum_i \xi_i\right)$$

$$s.t.\ \ y_i(wx_i + b) \geq 1 - \xi_i$$

The value of *C* controls how strongly slack is taken into account in the model. As *C* becomes arbitrarily large, less error is tolerated and *w* must be forced to fit more closely to the data,

potentially at the risk of overfitting. As *C* approaches zero, more error is tolerated, potentially at the risk of the hyperplane failing to separate the data in any meaningful way (Cortes et al., 1995).

Validation is a process used during training to find the best predictive combinations of features and to prevent overfitting. K-fold cross-validation is a method that partitions the training set into *k* subsets. For each set, the model is using the other *k*-1 sets then tested against the remaining set. The process is repeated *k* times and the performance is averaged across all *k* sets. Once the optimal combination of features has been found, the model is re-trained once more using all data (Kotsiantis, 2007).

4. Methodology

4.1 Identify Artists

Before any calculations can be done, research is undertaken to compile a list of artists who are known to have recorded with a given production unit and a time period within the scope of the project that they can be confirmed to have recorded for that unit. Several sources were used in total and when possible more than one source was used to verify each artist. One of the most common resources was Discogs.com, a searchable online database of vinyl records that includes label and, in some cases, studio information. Some artists' label affiliations are well documented in the public record, such as Smokey Robinson and Motown or Muddy Waters and Chess. In cases where more rigorous validation was needed other sources were consulted.
After a list of artists with start and end years was compiled for each production unit, the MSD was queried for a list of all songs by that artist within the supplied date range.

4.2 Extract Perceptual Representation

After gathering a full list of songs, the following data fields are extracted from the MSD for each song:

- Track title, artist name, and year. These are not used by the classifier, but for examination of how different types of songs, artists, and date ranges performed during the testing phase.

- Tempo in beats per minute, as estimated by Echo Nest.
- Key as estimated by Echo Nest. The key is recorded as an integer 0 through 11, with values mapped to the chromatic scale such that 0 represents C and 11 represents B.
- Mode as estimated by Echo Nest, with 0 representing minor and 1 representing major.
- Segments Pitches. The prevailing pitch is estimated for each segment. Each estimate is comprised of a 12 element column representing each chromatic pitch with C at index 1. The value at each index is the prominence with which the pitch represented by that index is detected, normalized such that the strongest detected pitch is equal to 1. A noisy signal features all values close to 1, while a pure tone would be represented by one value at 1 and all others at 0. For a song with *n* segments, the full pitch data for each song is a matrix of 12 rows by *n* columns (Jehan, 2014).
- Segments Timbre (Echo Nest Timbral Features). The Million Song Dataset describes these as similar to Mel Frequency Cepstral Coefficients (see note in section 4.1.1). Each segment is represented by 12 floating point descriptors. For a song with *n* segments, the full pitch data for each song is a matrix of 12 rows by *n* columns.
- Segments Loudness Start, the loudness in decibels full scale of the onset of each segment.
- Segments Loudness Max, the maximum loudness in decibels full scale detected in each section.
- Segments Loudness Max Time, the time as a proportion of each section at which the maximum loudness is detected.

4.2.1 A Note on Echo Nest Timbral Descriptors Versus MFCC's

Mel Frequency Cepstral Coefficients (MFCC's) are short term audio features that describe the overall shape of the spectral envelope. They are recognized as a key feature in speech recognition systems and are popular for use in music as timbral descriptors. MFCC's are calculated via a computational model of the human ear. The audio signal is broken into frames, a Hamming window is applied and the Discrete Fourier Transform of each frame is taken. Because human hearing is logarithmic, the logarithm of the amplitude spectrum from the DFT is then calculated. The frequency bins of the log amplitude spectrum are then grouped according to the Mel scale, which approximates the widths of the critical bands of human hearing along the basilar membrane. The bands are typically calculated using 50% overlap and smoothed with a triangular window. The Discrete Cosine Transform of each frame is then taken to produce a set of coefficients. For music processing the first 13 coefficients are generally considered the most useful. (Logan, 2000)

Although the Million Song Dataset identifies the Segments Timbre features from EchoNest as "MFCC-like," no documentation is provided to explain how the features are calculated or to demonstrate their effectiveness similar to MFCC's. Additionally, Segments Timbre provide only 12 coefficients versus the 13 traditionally kept from MFCC's (Bertin-Mahieux et al., 2011). Echo Nest's "Analyzer Documentation (2014)" does give some deeper insight into the Segments Timbre feature, revealing them to be quite unlike MFCC's. Segments Timbre are described as 12 individual coefficients of 12 basis functions which are not explicitly given. The coefficients are said to be ordered by degree of importance. Some of the

represented timbral features are revealed to be (by dimension): 1. Overall loudness, 2. Brightness, 3. Spectral flatness, 4. Attack (Jehan, 2014).

In their paper "Capturing the Temporal Domain in Echo Nest Features for Improved Classification Effectiveness (2012)," Alexander Schindler and Andreas Rauber evaluate the performance of Echo Nest Segments Timbre versus MFCC's. Schindler and Rauber find that the Echo Nest features perform as a reliable alternative to MFCC's for genre classification when used as part of a complex feature set.

4.3 Build Learning Representation

After the features for each song are extracted from the MSD, the features are processed to create a set of descriptors which are assembled into a 66-element feature vector plus a response label. The dimensions of the vectors are as follows:

1	2	3	4	5	6	7	8	9-20	21-32
Label	Tempo	Pocket	Tightness	Max RMS	Dynamic Variance	Key	Mode	Mean Timbre	Std Timbre

33-44	45	46	47	48	49	50	45	51	52
Pitch Class	Down Maj 7	Down Min 7	Down Maj 6	Down Min 6	Down Perf 5	Down Tritone	Down Maj 7	Down Perf 4	Down Maj 3

53	54	55	56	57	58	59	60	61	62
Down Min 3	Down Maj 2	Down Min 2	Unison/Octave	Up Min 2	Up Maj 2	Up Min 3	Up Maj 3	Up Perf 4	Up Tritone

63	64	65	66	67
Up Perf 5	Up Min 6	Up Maj 6	Up Min 7	Up Maj 7

Table 4.1. Feature vector anatomy.

4.3.1 Response Label

The response label represents the true class of the song. The labels used are: 1 = Motown, 2 = Stax, 3 = Chess, 4 = FAME, 5 = Atlantic. Response labels are the first element of the vectors used in the training data but are kept in a separate MATLAB object for verifying testing data.

4.3.2 Tempo

The tempo of the song in beats per minute, as extracted from the MSD.

4.3.3 Pocket

Pocket is an expression of Loudness Time Max as a fraction of a duration of a beat. It attempts to describe the average time it takes for a segment to reach its maximum amplitude (musically speaking, its attack) based on tempo. It is calculated by dividing the mean of all Loudness Time Max values by the number of seconds per beat (derived from Tempo):

$$\frac{T}{60N} \sum_{n=1}^{N} L_{max}(n)$$

The feature is named for the colloquial musical term “pocket” which describes how far ahead or behind the beat rhythmic subdivisions tend to fall. It is hypothesized that different groups of musicians and different rooms may exhibit a tendency toward a certain pocket.

4.3.4 Tightness

Tightness is a similar measurement to Pocket but instead measures how far the attack time tends to vary as a proportion of one beat. It is calculated by dividing the variance of the Loudness Times Max by seconds per beat:

$$\frac{T \cdot var(L_{max})}{60}$$

It can be interpreted as a measure of consistency.

4.3.5 Max RMS

The Root Mean Squared (RMS) value of the Loudness Max of all sections is taken to closely approximate peak RMS of the recording.

4.3.6 Dynamic Variance

Dynamic Variance measures the level of consistency in the dynamic range of segments. The difference in dB FS between the Loudness Max and the Loudness Onset of each section is taken. The variance of the resulting difference vector is calculated:

$$D[n] = L_{max}[n+1] - L_{max}[n]$$

A low variance suggests a more consistent dynamic range throughout the song, which may be affected by dynamics processing or the acoustics of the room as well as instrumentation and performance.

4.3.7 Key

As stated in 4.2, the key extracted from the MSD is stored as 0 = C, 1 = C#, … 11 = B. Because key values are useful as an index value in code, the key is incremented to C = 1 … B = 12 to account for MATLAB indexing.

4.3.8 Mode

Used as extracted from the MSD (minor=0, major=1).

4.3.9 Mean and Standard Deviation of Timbral Features

While the first nine values calculated are scalar representations, the Segments Timbre features are represented sequentially as a sub-vector. The difficulty of preparing the timbre features for use is that the size of the timbral data for each song varies depending on the length of the song. It is therefore desirable to ensure a common sample size is taken from each song selected. The length in segments of the shortest song n_{min} is found. For each song, the midpoint is calculating by dividing the number of segments. of the song by 2. From the midpoint, a section of the song half the length of the shortest song is taken in each direction. The result is a sample of the song equal to the length of the shortest song in the data set. The mean of each of the 12 timbral features is then calculated across the full sample of the song. Final result is a 12-dimension vector for each song describing the average of each timbral descriptor. The values are added to the feature vector at indices 9-20. The same process is then carried out using the standard deviation of the features instead of the mean. The resulting 12 dimensions are added to the feature vector at indices 21-32.

4.3.10 Chroma Features

The remaining sub-vectors use the extracted pitch information to examine the distribution of different pitches used throughout the song and of changes in pitch (intervals). It can be hypothesized that the different musical influences of the personnel embedded with different production units would result in patterns of use of both certain scale degrees and intervallic motion. For example, a style that leans more heavily on jazz influences, such as that found at Motown, would exhibit more chromaticism and likely show a more uniform distribution of pitch than a style that draws primarily from the blues, which would likely be weighted toward something resembling a pentatonic pattern.

As described in 4.2, the Segments Pitches data extracted from the MSD represents segment of each song with a 12-element column with each element representing one pitch class of the chromatic scale (C = 0). Each value represents the degree to which that pitch class is detected in the segment, with 1 representing the strongest detected pitch. Using the estimated key the order of the columns are rearranged so that the element representing the root is at index 1 with all other pitches ascending chromatically (such that index 2 represents the minor 2^{nd} and index 12 the major 7^{th}). The index of the max value of each column is taken, resulting in a vector containing values 1 through 12, corresponding to the most prominent pitch detected in each segment. In a separate vector, the difference between each adjoining pair of notes is then calculated to produce a vector representing the detected intervallic motion in half steps. Negative numbers imply a descending interval, positive ascending. A value of 0 represents either a unison

or octave. Compound intervals are not accounted for as octave information is not recorded in the MSD.

It should be noted here that pitch modulation is not taken into account in this system. Pitch modulation would most likely be represented as a weaker detection of the strongest pitch and a stronger than usual detection of the adjacent pitches. Detecting and incorporating modulations may be part of a potential future expansion of the work.

From these two vectors a “chroma matrix” is assembled with rows (12) representing detected pitch classes and columns (23) representing calculated intervals. The indexing of the columns and the intervals represented corresponds to: 1 = -11 (down a major 7th), 12 = 0, 23 = 11 (up a major 7th). This can serve as some source of confusion but is necessary since array indices cannot be negative or zero in MATLAB. Each cell counts the number of times each interval occurred from starting on each pitch. From this matrix the marginal distribution of pitch class is taken as a 12-element vector and added to the feature vector in dimensions 33-44. The marginal distribution of intervals is taken as a 23-element vector and added to the feature vector in elements 45-67.

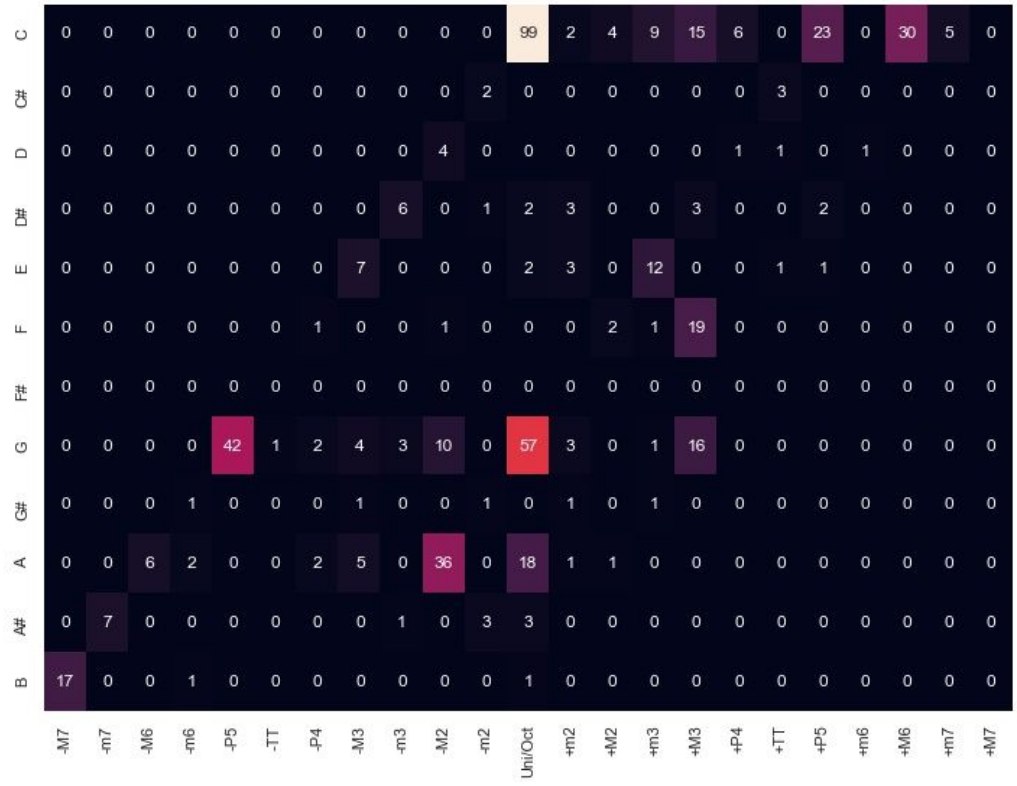

Figure 4.1. Heatmap of the Chroma Matrix for "Ain't Too Proud to Beg" by The Temptations (1966). Rows represent starting pitch classes, columns the interval to the next pitch. Cells represent the number of instances of each interval from each starting pitch.

4.4 Training

After the feature vectors are assembled for each song the full data set is randomly partitioned into training and testing sets, with 80% used for training and 20% for testing. The training set is trained in the MATLAB Classification Learner using 5-fold cross validation. Several SVM models are trained using six different kernel types: linear, quadratic, cubic, and

fine, medium, and coarse Gaussian. The kernel type with the most favorable accuracy score and confusion matrix is saved and tested using the testing data. Each classification scheme is trained in five sessions using five different training/testing splits, and the average and variance of the five best results is taken.

For each training session, a baseline score is taken using the Zero-Rule scheme. In this method, all observations are assigned to the majority class and the accuracy of such an assumption is measured. This would be the rough equivalent of a person simply labeling every classic soul or R&B song as "Motown." The baseline gives an idea of how often that assumption would be correct. For the classifier to be meaningful, it should perform better than the baseline score.

4.5 Classification Schemes

The following classification schemes are used: Motown vs Stax, Motown vs. Stax vs. Chess, Motown vs. Stax vs. Other (Chess, FAME, Atlantic combined), Chess vs. FAME vs. Atlantic, and finally all five studios against each other.

4.5.1 Motown vs. Stax

This classification scheme is undertaken first to examine the validity of the classifier. Motown and Stax are chosen as the two most represented classes in the dataset and two highly distinct styles.

4.5.2 Motown vs. Stax vs. Chess

Chess records is included as the third most represented class with a sufficiently large artist pool. While efforts were made to restrict selected Chess records to 2120 South Michigan Ave era, Chess's more diverse roster of musical genres and recording locations are expected to make classification more challenging.

4.5.3 Motown vs. Stax vs. Chess/FAME/Atlantic

In this scheme, Chess, FAME, and Atlantic are combined into a single "Other" class. The amalgamated class is closer in size to Motown and Stax providing for a more balanced dataset and gives Motown and Stax the opportunity to demonstrate distinctness from a general category of soul music.

4.5.4 Chess vs. FAME vs. Atlantic

The three smallest members of the dataset are tested against each other. Potential challenges include more noise in the results due to a smaller amount of data and a more diverse array of musical and sonic style produced at these studios, especially from the major label affiliated Atlantic Studios.

4.5.5 All vs. All

Finally, all five studios are tested against each other.

4.6 Scoring

For each trained model the following information is recorded:

- The list of all songs in the testing data, their true class and observed class.
- The baseline Accuracy and F1 score for both the testing and training sets.
- The Confusion Matrices of the training and testing results, listing number of observations of each true class vs. each observed class.

Recall and Precision are calculated from the Confusion Matrix for each class. Recall is calculated by dividing the number of True Positive observations by sum of True Positives plus True Negatives. Precision is calculated by dividing the number of True Positives by the sum of True Positives and True Negatives. An F1 Score, defined as the harmonic mean of the Recall and Precision, is then calculated for each class to produce a single class-wise accuracy metric. The mean of all F1 scores is taken to provide the Macro Average F1 score of the classifier. The Macro Average F1 of all five iterations is then averaged to obtain a final score.

The Macro Average F1 score is compared to the Macro Average F1 score of the Zero-Rule baseline algorithm. The five baselines of each random testing dataset are evaluated such that all observations are assigned the most common class. The F1 score of each class is taken, with the unrepresented classes assigned a score of 0. The Macro Average F1 score of the five baselines is taken, yielding a formula of the F1 score of the observed class divided by the number of classes. The five Baseline Macro Average F1 scores are then averaged for a final baseline score for comparison.

Depending on the nature of a classification system, Accuracy (the sum of all true positive observations divided by the total number of observations) can be misleading, as it only measures correct answers without addressing recall or precision. This can be problematic with unbalanced datasets, where a relatively high accuracy score can be obtained by simply labeling everything as the most common observation. By using the harmonic mean of recall and precision, the F1 score penalizes a recall or precision score that falls short (Shung, 2018).

5. Analysis

In general, although the overall accuracy of the classification schemes does decline as more classes and diverse options are added, the accuracy does prove to be higher than the baseline accuracy in every case. Motown Records did tend to perform the best versus other classes, followed closely by Stax. Atlantic Records generally performed poorly.

The following table represents all songs and artists used to generate the various models in the dataset, used for both training and testing:

	Motown	Stax	Chess	FAME	Atlantic	Total
Song Count	293	223	80	76	74	746
Artist Count	18	10	7	6	7	48

Table 5.1. Dataset Song and Artist Composition.

5.1 Motown vs. Stax

A total of 513 songs representing Motown and Stax were recovered from the MSD. Five different training sessions were conducted using five randomly chosen sets of 413 songs and the best scoring classifier from each session was recorded and saved for testing. The testing data for each classifier was comprised of the remaining 103 songs from the same randomized selection as the training data. The total data recovered was comprised of 293 Motown songs and 223 Stax songs. The mean confusion matrices presented here and the following analysis sections are normalized such that each row sums to 1.

	Motown	*Stax*
Motown	0.93	0.07
Stax	0.15	0.85

Table 5.2. Mean Confusion Matrix, Motown vs. Stax.

	Motown	*Stax*
Motown	1.00	0.00
Stax	1.00	0.00

Table 5.3. Mean Baseline Confusion Matrix, Motown vs. Stax.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.93	0.89	-	0.91
Stax	-	-	0.85	0.90	-	0.87
Mean	0.58	0.37	0.89	0.90	0.90	0.89

Table 5.4. Mean Scoring results, Motown vs. Stax.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.0014	0.0017	-	0.0002
Stax	-	-	0.0040	0.0021	-	0.0006
Means	0.0008	0.0001	0.0004	0.0002	0.0003	0.0003

Table 5.5. Variance Scoring results, Motown vs. Stax.

The mean observation classes of the five Motown vs. Stax classifiers is shown in Table 5.2. The baseline calculation is shown in Table 5.3. The *y*-axis represents the true class, the *x*-axis the predicted class. As expected, the results of the SVM are clearly superior to Zero-Rule,

with an overall accuracy of 0.90 versus 0.58. The overall precision is similarly high for both classes, and while there is an almost 9 percentage point disparity in recall, the two classes are clearly distinguishable in most cases. Motown and Stax averaged respective F1-scores of 0.91 and 0.87 for a Macro average F1 of 0.89, significantly higher than the baseline macro average F1 of 0.37.

Of the 513 songs extracted, 306 unique songs ended up randomly selected for the five training sets for a total of 515 total observations across five trials. Of the 306 unique songs, 34 (11.11%) were labeled incorrectly one or more times for a total of 53 incorrect predictions out of 515. Of the 34 songs to be mislabeled, 13 were Motown and 21 were Stax. Eleven songs were misclassified having only been including in the testing set once. Of those seen more than once (eleven were seen two times, six were seen three times, seven were seen four times), 9 songs never classified correctly despite multiple tries.

Two songs have the distinction of failing to classify in four attempts: “Just Be True” by David Porter (1970) and “In the Hole” by The Bar-Kays (1969), both Stax recordings. Another Stax record, “Just Keep Holding On” by Sam & Dave (1967), failed to classify correctly in three attempts. “Just Be True” and “Just Keep Holding On” are both slow ballads in 6/8 time while “In the Hole” is a slow instrumental in 4/4. Aside from tempo and relative time period, these Stax recordings also feature prominent use of horns and electric guitar. “Just Be True” and “In the Hole” both happen to make prominent use of a guitar strum in unison with the snare, similar to Joe Messina’s signature Motown rhythm.

Stax legend David Porter has the distinction of being the only artist who never classified correctly, missing five times in five attempts. Other notably high miss rates were Motown girl

group The Velvettes who classified correctly only three times in eight attempts and Stax instrumental group The Bar-Kays who had 9 misclassifications (most of any artist) in 26 attempts.

5.2 Motown vs. Stax vs. Chess

A total of 591 songs representing Motown, Stax, and Chess were recovered from the MSD. Five different training sessions were conducted using five randomly chosen sets of 473 songs and the best scoring classifier from each session was recorded and saved for testing. The testing data for each classifier was comprised of the remaining 118 songs from the same randomized selection as the training data. The total data recovered was comprised of 293 Motown songs and 223 Stax songs, but only 75 Chess songs. The imbalance in the dataset is a likely contributor to the underperformance of the Chess class as seen below:

	Motown	*Stax*	*Chess*
Motown	0.90	0.08	0.02
Stax	0.15	0.82	0.02
Chess	0.30	0.26	0.44

Table 5.6. Mean Confusion Matrix, Motown vs. Stax vs. Chess.

	Motown	*Stax*	*Chess*
Motown	1.00	0.00	0.00
Stax	1.00	0.00	0.00
Chess	1.00	0.00	0.00

Table 5.7. Mean Baseline Confusion Matrix, Motown vs. Stax vs. Chess.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.90	0.81	-	0.86
Stax	-	-	0.82	0.81	-	0.82
Chess	-	-	0.44	0.79	-	0.54
Mean	0.50	0.22	0.72	0.80	0.81	0.74

Table 5.8. Mean Scoring results, Motown vs. Stax vs. Chess.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.0007	0.0022	-	0.0010
Stax	-	-	0.0015	0.0059	-	0.0010
Chess	-	-	0.0157	0.0343	-	0.0177
Means	0.0012	0.0001	0.0030	0.0042	0.0011	0.0034

Table 5.9. Variance Scoring results, Motown vs. Stax vs. Chess.

Chess Records, outnumbered nearly 4 to 1 by Motown, unfortunately averages very poor recall among these five models, although precision remains satisfactory. In other words, the classifier does slightly worse than a coin toss when it comes to recognizing a Chess song as a Chess song, but when it does decide a song is Chess it is right slightly more than three out of four times. Motown, the most represented class, has recall on par with the binary classifier, while Motown and Stax both combine for good performance overall, though slightly less than when no third choice was presented. Chess songs were likely to be mislabeled almost equally as Motown or Stax. While only 7.5% of Motown songs were mislabeled as Stax and 1.7% as Chess, 15.8% of Stax songs were incorrectly labeled Motown and 2.9% were mislabeled as Chess.

The poor recall of Chess drags its F1 score down to only 0.54. The variance of Chess is also much greater than the more well-represented classes, suggesting that its distribution is heavily distorted. It can be noted that if a baseline were to be taken which labeled all observations Chess, the resulting F1 score for Chess alone would be 0.24 suggesting that the trained classifiers have, at minimum, outperformed a hypothetical baseline given the balance of classes. Motown and Stax score 0.86 and 0.82 respectively for a Macro Average F1 score of 0.74, still respectable compared to the Baseline Macro Average F1 Score of 0.22. The success rate of the model is also affected by a single outlying case which significantly outperformed the other four, visible in Figures 5.2 and 5.3 (p. 77-78).

Of the 591 songs extracted, 397 unique songs by 33 unique artists were used in the five testing sets for a total of 590 observations across five trials. Of the 397 unique songs, 76 (19.14%) were labeled incorrectly one or more times for a total of 111 incorrect predictions out of 590. Of the 76 songs to be mislabeled, 15 were Motown, 29 were Stax, and 32 were Chess. In contrast to the model in 5.1, only 9 songs (12%) that were missed at all were classified correctly in another attempt, 23 (30%) were never scored correctly in more than one attempt, and the remaining 44 that were missed (58%) only appeared once. The decline in repeated songs is likely due to the expansion of the size of the data pool with the number of trials remaining constant.

Twenty of the 33 artists experienced at least one misclassification. The two artists to fail to register a single correct classification were once again David Porter (6 of 6 instances all classified as Motown) and Chess legend Etta James, famous for soulful, jazz-influenced ballads that were distinctly different from Chess's more typical blues and early Rock and Roll artists. As such, it is not a surprise that James failed to classify correctly. All four instances were predicted

as Stax. Three other Chess legends - Chuck Berry (0.70), Muddy Waters (0.56), and Howlin' Wolf (0.53) - all had an error rate higher than 50%. Motown's poorest performer was The Velvettes, who misclassified 5 out of 7 times after also failing five times in 5.1.

A particularly noteworthy feature is that of the 76 songs that misclassified, only three misclassified as more than one incorrect class. Marvin Gaye's "What's Going On" (Motown) classified twice as Stax and once as Chess. The Velvette's "Needle in a Haystack" (Motown) classified three times as Stax and once as Chess. Both songs failed to classify on multiple attempts in 5.1. Additionally, Rufus Thomas's "Sixty Minute Man" (Stax) classified once each as Motown and Chess.

5.3 Motown vs. Stax vs. Chess/FAME/Atlantic

A total of 734 songs representing all studios were recovered from the MSD. Five different training sessions were conducted using five randomly chosen sets of 587 songs and the best scoring classifier from each session was recorded and saved for testing. The testing data for each classifier was comprised of the remaining 147 songs from the same randomized selection as the training data. In this example, Chess, FAME, and Atlantic were grouped together as a single "Other" class. The total data recovered was comprised of 293 Motown songs, 223 Stax songs, and 218 Other songs. While the criteria of the Other class is much less defined, the dataset is much more balanced than in the previous ternary example.

	Motown	*Stax*	*Other*
Motown	0.82	0.03	0.14
Stax	0.16	0.67	0.18
Other	0.21	0.17	0.62

Table 5.10. Mean Confusion Matrix, Motown vs. Stax vs. Other.

	Motown	*Stax*	*Other*
Motown	0.43	0.57	0.00
Stax	0.37	0.63	0.00
Other	0.40	0.60	0.00

Table 5.11. Mean Baseline Confusion Matrix, Motown vs. Stax vs. Other.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.82	0.71	-	0.76
Stax	-	-	0.67	0.79	-	0.72
Other	-	-	0.62	0.63	-	0.62
Mean	0.37	0.18	0.70	0.71	0.71	0.70

Table 5.12. Mean Scoring results, Motown vs. Stax vs. Other.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.0020	0.0007	-	0.0009
Stax	-	-	0.0027	0.0087	-	0.0029
Other	-	-	0.0033	0.0009	-	0.0014
Means	0.0005	0.0001	0.0010	0.0012	0.0009	0.0010

Table 5.13. Variance Scoring results, Motown vs. Stax vs. Other.

As a result of the data being more closely balanced in this examination, Motown was not always the most represented class in each iteration of testing data. In fact, Stax was the most represented class three out of five times. The Mean Confusion Matrix for the baseline shown Figure 8 reflects this, as both Motown and Stax were selected as the majority class in different iterations.

With the introduction of a more balanced dataset at the cost of a more poorly defined third class, the recall and precision of Motown and Stax, the previously highly performing classes, fell. Motown maintains the best recall, with Stax and Other in a similar range. The high variance of the third class is also brought within a more normal range with classes 1 and 2. With the expansion of the third class to include FAME and Atlantic songs and to be more in balance with classes 1 and 2, Other has much improved recall performance versus Chess in the previous exercise. Precision performance falls for all classes, although Stax experience only a slight decrease versus a rather significant drop-off of more than ten percentage points each for Motown and Other.

While the F1 score of class 3 improved from 0.54 to 0.62 between the two ternary examples, the F1 scores of Motown and Stax each fell roughly 10 points to 0.76 and 0.72 respectively. As a result, the Macro Average F1 score of the model comes out to 0.70, a slight decrease versus the previous example but still an equally good comparative performance versus a baseline Macro Average F1 of 0.18.

Of the 734 songs extracted, 496 unique songs by 45 unique artists were used in the five testing sets for a total of 736 observations across five trials. Of the 496 unique songs, 149 (30%) were labeled incorrectly one or more times for a total of 211 incorrect predictions out of 736. Of

the 149 songs to be mislabeled, 34 were Motown, 54 were Stax, and 61 were one of Chess, FAME, or Atlantic. In contrast to 5.2, 29 songs (19%) that were missed at all were classified correctly in another attempt, 44 (29.53%) were never scored correctly in more than one attempt, and the remaining 76 that were missed (51%) only appeared once. The near doubling in number of songs that misclassified at least once vs. 5.2 is at least partially explained by the increase in the size of the data pool. The increase in percentage of of songs that missed but eventually classified correctly (19% vs. 12%) suggests slightly more confusion between trials, but also a greater likelihood of a song eventually being classified correctly. The decrease in percentage of songs that never classified correctly despite multiple repetitions (51% vs. 58%) could hint at an increase in usability.

Thirty eight of forty five artists had at least one song classify incorrectly. Seventeen artists were misclassified in 50% or more of all observations. Five artists failed to classify correctly at all: David Porter of Stax (11 instances, 5 songs), Etta James of Chess (6 observations, 4 songs), Isaac Hayes of Stax (4 observations, 3 songs), Bobby Moore and the Rhythm Aces of Chess (2 observations, 1 song), and Big Joe Turner of Atlantic (1 observation, 1 song). Porter once again consistently misclassified as Motown in all observations.

Before moving onward to model 5.4, which will not look at the Motown and Stax classes, it is worth acknowledging six artists who, in all three models thus far, have not had a single misclassification: Barrett Strong (10 observations), David Ruffin (17), Diana Ross (38), The Jackson 5 (21), Marvin Gaye and Kim Weston (24), and The Supremes (66). All six are Motown artists. If any conclusions can be drawn from this research thus far, it may be that if someone

were to ask to have the Motown Sound explained, one would be well advised to hand them an album by one of these six artists.

5.4 Chess vs. FAME vs. Atlantic

A total of 215 songs representing Chess, FAME, and Atlantic were recovered from the MSD. Five different training sessions were conducted using five randomly chosen sets of 172 songs and the best scoring classifier from each session was recorded and saved for testing. The testing data for each classifier was comprised of the remaining 43 songs from the same randomized selection as the training data. The total data recovered was comprised of 80 Chess songs, 61 FAME songs, and 74 Atlantic songs.

	Chess	*FAME*	*Atlantic*
Chess	0.67	0.11	0.23
FAME	0.09	0.68	0.23
Atlantic	0.15	0.21	0.64

Table 5.14. Mean Confusion Matrix, Chess vs. FAME vs. Atlantic.

	Chess	*FAME*	*Atlantic*
Chess	0.44	0.18	0.38
FAME	0.38	0.30	0.32
Atlantic	0.37	0.15	0.48

Table 5.15. Mean Baseline Confusion Matrix, Chess vs. FAME vs. Atlantic.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Chess	-	-	0.61	0.78	-	0.72
FAME	-	-	0.46	0.60	-	0.64
Atlantic	-	-	0.83	0.60	-	0.62
Mean	0.42	0.19	0.64	0.66	0.66	0.66

Table 5.16. Mean Scoring results, Chess vs. FAME vs. Atlantic.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Chess	-	-	0.0232	0.0021	-	0.0065
FAME	-	-	0.0240	0.0230	-	0.0116
Atlantic	-	-	0.0247	0.0183	-	0.0091
Means	0.0003	0.0006	0.0027	0.0038	0.0031	0.0030

Table 5.17. Variance Scoring results, Chess vs. FAME vs. Atlantic.

Each class was at least once the majority class when calculating baseline scores. The Macro Average F1 reported in Table 5.16 is the average of the Macro Average across all five iterations. All labels exhibited similar overall recall performance, albeit with very high variance between repetitions, but Chess does outperform FAME and Atlantic's precision scores. As a result, Chess retains the highest F1 score with FAME and Atlantic scoring closely with each other. The relationships between FAME and Atlantic are discussed in the following sections.

Of the 215 songs extracted, 140 unique songs by 18 unique artists were used in the five testing sets for a total of 215 observations across five trials. Of the 140 unique songs, 54

(38.57%) were labeled incorrectly one or more times for a total of 73 incorrect predictions out of 215. Of the 54 songs to be mislabeled, 20 were Chess, 15 were FAME, and 19 were Atlantic. Eighteen songs (33.33%) that were missed at all were classified correctly in another attempt, 11 (20.37%) were never scored correctly in more than one attempt, and the remaining 25 that were missed (46.30%) only appeared once.

Despite the removal of Motown and Stax, certain underperformers continued to misclassify under the newly introduced labels. The Etta James Chess classic "At Last," which consistently classified as Stax in 5.2 and Motown in 5.3, registered twice as FAME and once as Atlantic. James's position as a musical outlier in the Chess family was discussed earlier, but the song's failure to classify correctly in any model (true in fact of any of James's songs) further underscores her unique position with the label.

5.5 All vs. All

A total of 734 songs representing all studios were recovered from the MSD. Five different training sessions were conducted using five randomly chosen sets of 587 songs and the best scoring classifier from each session was recorded and saved for testing. The testing data for each classifier was comprised of the remaining 147 songs from the same randomized selection as the training data. The total data recovered was comprised of 293 Motown songs, 223 Stax songs, 75 Chess songs, 76 FAME songs, and 67 Atlantic songs.

	Motown	*Stax*	*Chess*	*FAME*	*Atlantic*
Motown	0.86	0.06	0.02	0.02	0.04
Stax	0.18	0.66	0.02	0.10	0.04
Chess	0.25	0.10	0.47	0.13	0.05
FAME	0.19	0.43	0.02	0.25	0.11
Atlantic	0.35	0.17	0.04	0.28	0.16

Table 5.18. Mean Confusion Matrix, All vs. All.

	Motown	*Stax*	*Chess*	*FAME*	*Atlantic*
Motown	0.83	0.17	0.00	0.00	0.00
Stax	0.75	0.25	0.00	0.00	0.00
Chess	0.76	0.24	0.00	0.00	0.00
FAME	0.89	0.11	0.00	0.00	0.00
Atlantic	0.82	0.18	0.00	0.00	0.00

Table 5.19. Mean Baseline Confusion Matrix, All vs. All.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.86	0.71	-	0.78
Stax	-	-	0.66	0.70	-	0.67
Chess	-	-	0.47	0.71	-	0.57
FAME	-	-	0.25	0.17	-	0.19
Atlantic	-	-	0.16	0.30	-	0.21
Mean	0.40	0.11	0.48	0.52	0.63	0.48

Table 5.20. Mean Scoring results, All vs. All.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.0003	0.0035	-	0.0015
Stax	-	-	0.0028	0.0119	-	0.0045
Chess	-	-	0.0029	0.0135	-	0.0046
FAME	-	-	0.0094	0.0070	-	0.0077
Atlantic	-	-	0.0036	0.0128	-	0.0057
Means	0.0004	0.0000	0.0012	0.0032	0.0010	0.0017

Table 5.21. Variance Scoring results, All vs. All.

With all five classes treated as distinct, the classifier runs into trouble. While the performance remains better than baseline, the Macro Average F1 score of 0.48 suggests serious limitations in the model's usefulness despite still performing well above the rather low baseline F1 of 0.11. The most apparent limitation is the very poor performance of FAME and Atlantic, which only serve to contribute noise to the system.

Of the 734 songs extracted, 435 unique songs by 49 unique artists were used in the five testing sets for a total of 735 observations across five trials. Of the 435 unique songs, 163 (37.47%) were labeled incorrectly one or more times for a total of 267 incorrect predictions out of 735. Of the 163 songs to be mislabeled, 25 (15%) were Motown, 49 (30%) were Stax, 25 (15%) were Chess, 26 (16%) were FAME, and 38 (23%) were Atlantic. Twenty songs (12%) that were missed at all were classified correctly in another attempt, 68 (42%) were never scored correctly in more than one attempt, and the remaining 75 that were missed (46%) only appeared once.

While not very useful for accurately identifying production units, some insightful observations can still be made:

- Atlantic Records did commonly contract with Rick Hall's FAME Studios to record Atlantic artists and so a reasonable assumption might be that FAME and Atlantic recordings would share a degree of similarity. Sure enough, on average 5 out of 17.8 Atlantic songs (28%) were identified as FAME. However, only 1.2 of 10.6 FAME songs (11%) were identified as Atlantic.
- Rather, an average of 4.6 out of 10.6 FAME songs (43%) were incorrectly labeled as Stax. This may reasonably be explained by the fact that Memphis, TN and Muscle Shoals, AL are only 150 miles apart and that local geography and similarity of local culture may have played a key role in forging a high similarity between the two production units. However, only 4.6 of 45.2 (10%) of Stax songs were misclassified as FAME.
- Atlantic Records was known to distribute for and to sub-contract with Stax Records as well. A combined 45% of Atlantic songs were classified as either Stax or FAME, with 35% classified as Motown. However, songs from those three labels were not highly likely to classify as Atlantic. Atlantic was the oldest and largest of the five labels at this time. The fact that its catalog is often confused with those of smaller production units whose own oeuvre is seldom confused with Atlantic may be an indicator of a major label looking to smaller labels enjoying great local success as a blueprint for its own recordings.

- The 45% of Atlantic recordings labeled as Stax or FAME may also serve as an indicator that particular record might have been made elsewhere. As noted in the motivation for this research, studios of origin are often not passed along with relevant historical information about a song. While the origins of well known hits are usually well documented, the information may have fallen by the historical wayside regarding other songs. While the classifier is hardly sufficient to serve as definitive proof of a song's true origin, it may be sufficient to raise questions.
- While Atlantic's Jerry Wexler was known to do regular business with Stax and FAME, he had no such known arrangement with Berry Gordy at Motown. An unreciprocated 35% confusion rate may be suggestive of a conscious attempt by Atlantic to imitate the Motown Sound.

Another way to investigate the reasoning behind the relative performance of the five labels is to measure the effectiveness of the individual artists in each Production Unit's space. An artist who represents a significant portion of a given Production Unit's total instances will have a strong impact on that label's performance. An artist with only a small number of songs but who performs consistently will also have an impact, even more so for labels of a small population size.

To visualize the impact of various artists in the dataset, an artist's total number of representations across all five iterations is plotted against that artist's individual accuracy across all five iterations. In the resulting scatterplot (Figure 5.1), the closer an artist near the top of the *y* axis tests consistently accurately, while an artist to the right on the *x* axis represents a large

number of observations. The closer an artist is to the top right corner, the more significantly that artist contributes to the label's results.

The plot suggests that Motown and Stax are aided by several highly significant artists, in particular Stevie Wonder and Marvin Gaye for Motown, and Booker T. & The MG's, Carla Thomas, and Otis Redding for Stax. Chuck Berry and Howlin' Wolf are the most significant Chess artists with accuracy scores above 50%, but their smaller sample sizes compared to other highly significant artists under other labels limits how well they can help Chess's scores. The majority of FAME and Atlantic's artists are present in the lower left quadrant, showing that the most common artists present under those two labels lacked the critical mass and/or accuracy to swing the classifier in that label's favor.

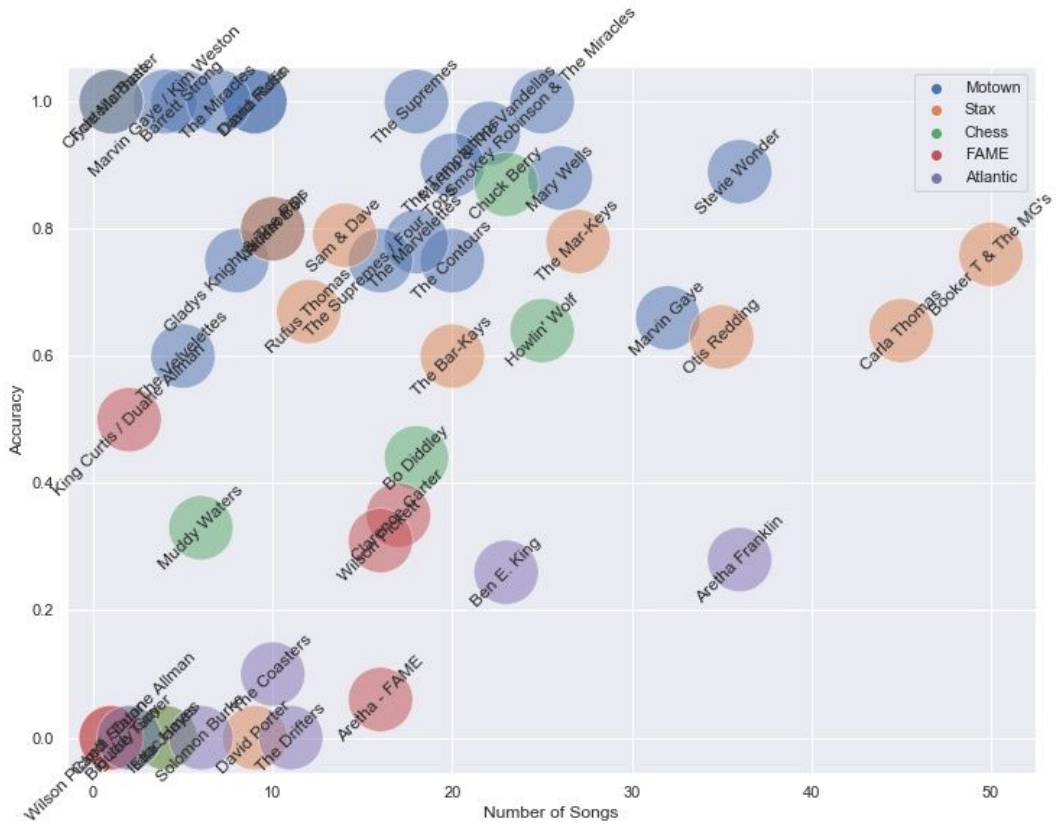

Figure 5.1. Artist Effectiveness.

5.5.1 The Aretha Franklin Effect

The high degree of confusion may also be caused by the Queen of Soul, Aretha Franklin. A native of Detroit, MI, Franklin was courted to the Tamla label by Gordy in 1960 but declined believing the label was not yet developed enough (Aretha Franklin, 2001). After a stint at Columbia Records from 1960 until 1966, Franklin signed with Atlantic Records. Her first album released on Atlantic, "I Never Loved a Man the Way I Love You," was recorded at FAME in 1967. In 1968 she recorded "Lady Soul" and "Aretha Now" at Atlantic Studios in New York

(Cogan et al., 2003). This situation sees Aretha Franklin as the only artist known to be included under two labels in the dataset - a problem considering she represents a significantly large portion of both classes. Further compounding the situation is Franklin's Detroit upbringing and her close friendship with Berry Gordy, all contributing to a signature sound that is highly similar to and commonly confused with Motown itself (Aretha Franklin, 2011).

The potential impact of Aretha Franklin's association with two production units highlights the need to consider "Studio Space," as per Whitman (2001) and the Artist Space problem. Franklin recorded one album at FAME, "I Never Loved a Man (The Way I Love You) (1967)." It is worth noting that some of the tracks were not finished in Muscle Shoals due to a fight between Rick Hall and Franklin's then-husband and manager Ted White. Jerry Wexler brought the Swampers who had been playing on the record to New York City to finish the album at Atlantic's West 60th Street facility (Cogan, et al., 2003). With the number and name of tracks finished in New York relatively unknown, the change of facility does inevitably compromise the Studio Space, although the consistency of personnel may potentially offset. Franklin's next two albums in 1968 were recorded at Atlantic. The inclusion of all three albums in the dataset creates an Artist Space for Franklin that contains members of two Studio Spaces, but with relatively clearly delineated subsets.

Aretha Franklin is represented twice in the scatterplot in Figure 5.1. Her Atlantic recordings appear in the right-hand portion of the plot, showing her to be the largest member of the Atlantic class but with an accuracy hovering around only 30%. Her FAME recordings appear in the lower left quadrant very close to the *x* axis, suggesting both a small sample size and a poor accuracy score.

An obvious test for how significant the Aretha Franklin Effect is in the current model is to withhold all songs by Aretha Franklin as the testing set and to train the model on everything that is left (restricting Motown and Stax to 70 songs apiece to create a more balanced data pool). Unfortunately, the final result did not bode well. Out of six Aretha Franklin songs recorded at FAME and 30 recorded at Atlantic, only two FAME and zero Atlantic recordings classified correctly.

	Motown	*Stax*	*Chess*	*Fame*	*Atlantic*
Motown	0.00	0.00	0.00	0.00	0.00
Stax	0.00	0.00	0.00	0.00	0.00
Chess	0.00	0.00	0.00	0.00	0.00
FAME	0.50	0.17	0.00	0.33	0.00
Atlantic	0.33	0.20	0.27	0.20	0.00

Table 5.22. Confusion Matrix of songs by Aretha Franklin only.

One possible interpretation is that Aretha Franklin is simply more similar to Motown than any other label. Stylistically, this is not an unreasonable explanation as 1.) Franklin was raised in Detroit and would likely have absorbed the same influences as other Detroit based musicians, 2.) the fact that Berry Gordy courted Franklin to sign with Motown suggests he saw her as a good fit for the Motown Sound, 3.) Motown was established as a very successful label and a rival to Atlantic by 1967 and it would make sense for Franklin and Jerry Wexler to try to emulate the Motown Sound.

Another less generous yet highly possible interpretation is that the classification model cannot handle classifying an artist it has not encountered in training. To test this hypothesis, a similar training/testing situation is set up that holds out Stevie Wonder for testing. Wonder's

Artist Space is comprised of 41 songs, all under the Motown label, spanning every year from 1963 to 1970 with the sole exception of 1965. In model 5.5, Wonder misclassified in only 4 out of 36 observations for an accuracy score of 89%. The model is trained on the rest of the dataset and then shown Stevie Wonder:

	Motown	*Stax*	*Chess*	*Fame*	*Atlantic*
Observations	28	1	1	2	9
Percentage	0.68	0.02	0.02	0.05	0.22

Table 5.23. Confusion Matrix of songs by Stevie Wonder only.

In contrast to Ms. Franklin, Stevie Wonder tests with 68% accuracy. Curiously enough, the most significant source of confusion is with Atlantic at 22%.

One curious side note: Of the 13 Stevie Wonder songs that classified incorrectly, only three were recorded prior to 1966. One can only speculate as to the significance if any, but 1966 could mark a point at which other labels, especially Atlantic began imitating the Motown Sound. As of the end of the previous year, Motown and its sub-label Tamla had landed 11 number 1 singles on the Billboard Hot 100, dwarfing Atlantic and its sub-label Atco's four (in fact, the Atlantic imprint itself hadn't had a number 1 single since The Drifters' "Save the Last Dance For Me" in 1960) (Billboard Charts Archive). Again, the classifier does not provide any conclusive proof of such an influence.

To attempt to control for confusion that Aretha Franklin might be providing, the previous Stevie Wonder holdout experiment was performed one more time, this time omitting all Aretha Franklin songs from the training or testing data. The result tests 80.49% of Wonder's music correctly as Motown, while the confusion rate with Atlantic falls to 9.75%.

	Motown	*Stax*	*Chess*	*Fame*	*Atlantic*
Observations	33	1	1	2	4
Percentage	0.80	0.02	0.02	0.05	0.10

Table 5.24. Testing results of songs by Stevie Wonder only with Aretha Franklin omitted from training.

For good measure, the classifier was also tested using a newly selected artist who fit none of the labels. David Bowie was selected as the random testing artist due to his large and diverse catalog of music. The 83 David Bowie songs retrieved from the MSD resulted in the following results:

	Motown	*Stax*	*Chess*	*FAME*	*Atlantic*
Observations	39	9	3	11	21
Percentage	0.47	0.11	0.04	0.13	0.25

Table 5.25. Results of model tested on David Bowie, Aretha Franklin omitted.

While Bowie's music does test more randomly than Stevie Wonder's, the distribution is much closer to Aretha Franklin's, suggesting a possible bias toward Motown. To test the degree of bias, the classifier is re-trained and tested on a significant Stax performer, Otis Redding:

	Motown	*Stax*	*Chess*	*FAME*	*Atlantic*
Observations	8	24	0	7	1
Percentage	0.20	0.60	0.00	0.18	0.03

Table 5.26. Results of model tested on Otis Redding, Aretha Franklin omitted.

While not quite as strong as Stevie Wonder's revised solo performance of 80%, if the purpose of the model were to guess which Production Unit Redding belonged to in his career, Stax would be the most likely (and correct) result.

5.6 All vs. All Revisited

The results of test cases against the Aretha Effect beg the question as to what effect could be obtained by a.) omitting Franklin's music as a stylistic outlier, and b.) rebalancing the data set to bring Motown and Stax to within a similar range of other classes. In a final test case, the classifier is trained once more with a regular 80/20 train/test split using all classes, but with all Aretha Franklin songs omitted. The classifier is trained five times as in 5.5 and the average results taken.

	Motown	*Stax*	*Chess*	*FAME*	*Atlantic*
Motown	0.61	0.04	0.20	0.07	0.07
Stax	0.09	0.60	0.11	0.15	0.05
Chess	0.19	0.09	0.68	0.01	0.03
FAME	0.04	0.16	0.14	0.60	0.06
Atlantic	0.13	0.05	0.10	0.18	0.55

Table 5.27. Mean Confusion Matrix, All vs. All with Balanced Dataset.

	Motown	*Stax*	*Chess*	*FAME*	*Atlantic*
Motown	0.23	0.00	0.56	0.21	0.00
Stax	0.22	0.00	0.65	0.14	0.00
Chess	0.17	0.00	0.65	0.17	0.00
FAME	0.19	0.00	0.56	0.26	0.00
Atlantic	0.20	0.00	0.58	0.23	0.00

Table 5.28. Mean Baseline Confusion Matrix, All vs. All with Balanced Dataset.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.61	0.61	-	0.61
Stax	-	-	0.60	0.63	-	0.61
Chess	-	-	0.68	0.59	-	0.63
FAME	-	-	0.60	0.65	-	0.62
Atlantic	-	-	0.55	0.61	-	0.58
Mean	0.26	0.08	0.61	0.62	0.62	0.61

Table 5.29. Mean Scoring results, All vs. All with Balanced Dataset.

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	0.0096	0.0141	-	0.0091
Stax	-	-	0.0308	0.0291	-	0.0167
Chess	-	-	0.0252	0.0315	-	0.0163
FAME	-	-	0.0154	0.0260	-	0.0070
Atlantic	-	-	0.0160	0.0646	-	0.0145
Mean	0.0007	0.0000	0.0052	0.0032	0.0042	0.0043

Table 5.30. Variance Scoring results, All vs. All with Balanced Dataset.

The final result sees a nearly 13 point increase in the Macro Average F1 score versus the underperforming model in Section 5.5. The Macro Average F1 of 0.61 performs well above the baseline of 8.23. The accuracy does drop slightly to 0.62 from 0.63, but the baseline score of the balanced dataset is much lower versus the unbalanced (0.40 vs. 0.26).

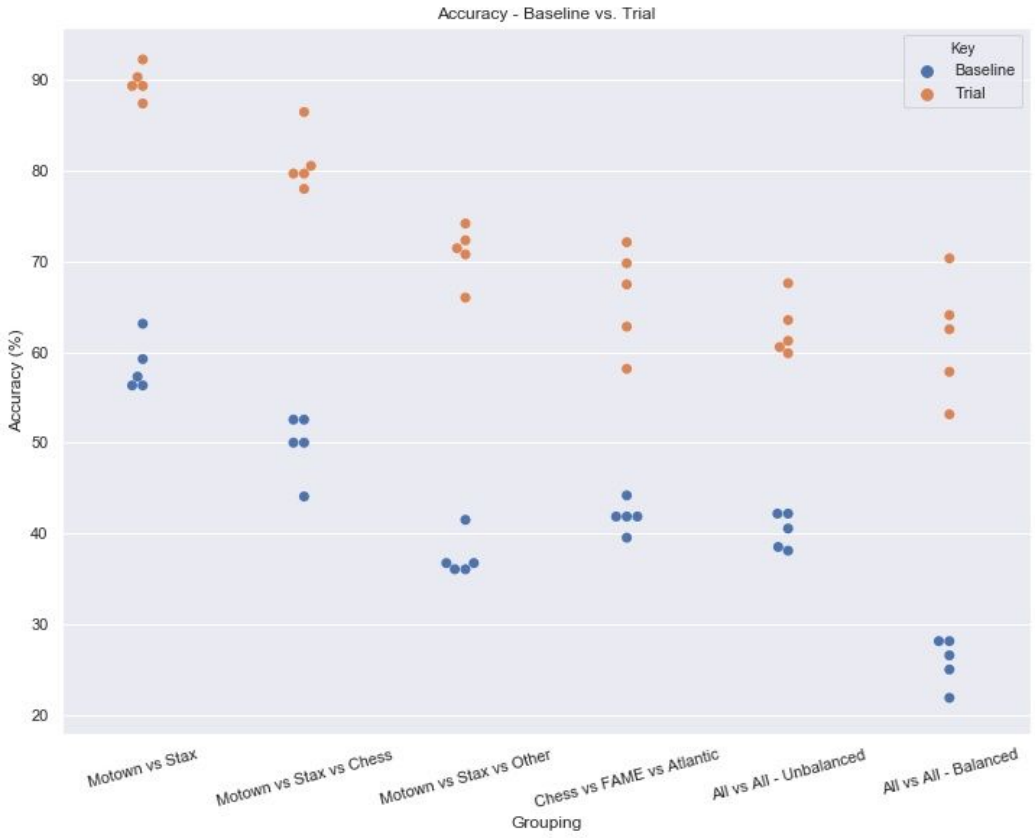

Figure 5.2. Trial vs. Baseline Accuracy, Point Plot.

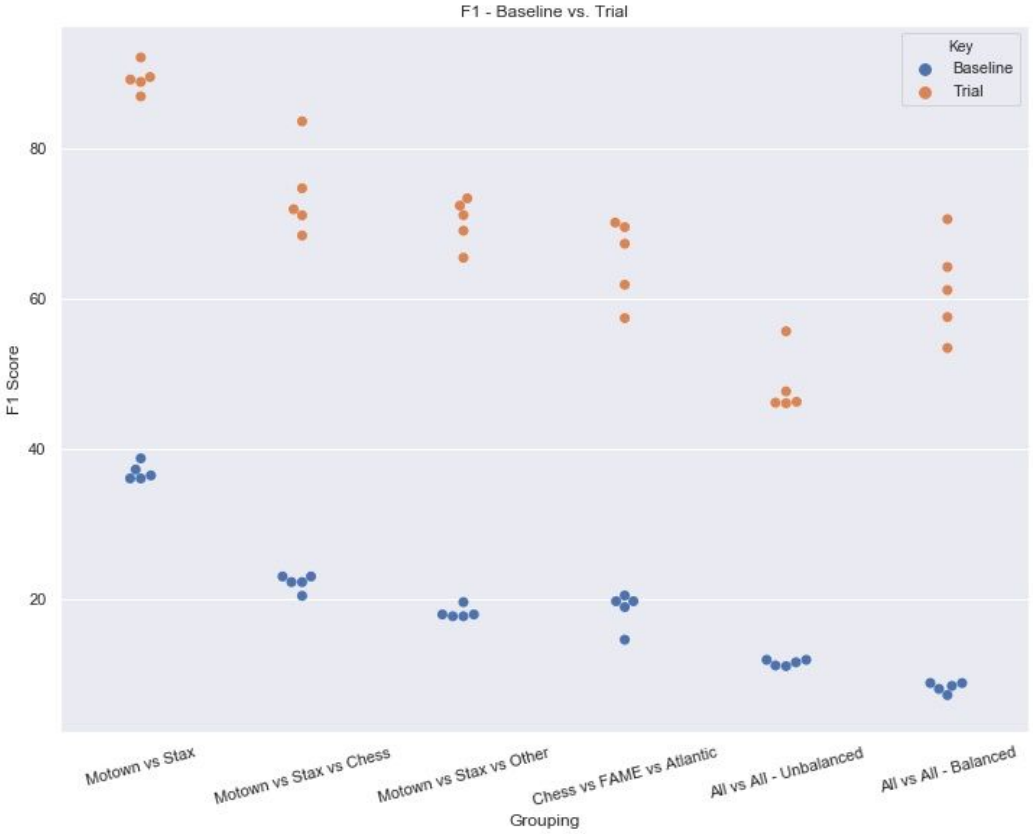

Figure 5.3. Trial vs. Baseline Macro F1, Point Plot.

6. Conclusions

6.1 Summary of Key Results

In every case tested, the classifier outperformed the baseline score in both Accuracy and Macro Average F1 score. A summary of the final results of each classifier allows for general conclusions to be drawn:

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown vs. Stax	0.58	0.37	0.89	0.90	0.90	0.89
Motown vs. Stax vs. Chess	0.50	0.22	0.72	0.80	0.81	0.74
Motown vs. Stax vs. Other	0.37	0.18	0.70	0.71	0.71	0.70
Chess vs. FAME vs. Atlantic	0.42	0.19	0.64	0.66	0.66	0.66
All vs. All (unbalanced)	0.40	0.11	0.48	0.52	0.63	0.48
All vs. All (balanced)	0.26	0.17	0.61	0.62	0.62	0.61

Table 6.1. Summary of mean performance metrics of all classifiers.

Although all classes did not perform equally well, enough of a distinction exists to conclude that a differentiable quality between Production Units of the era does exist and can be measured in most cases. The overall performance of the classifier does decline as more classes are added, with a 34% decline in Accuracy between a binary choice and the five class model.

Examining the two cases in which the dataset was largely unbalance - Motown vs. Stax vs. Chess (5.2) and All vs. All Unbalanced (5.5) - there is a notable decline in Recall and/or. Precision, leading to a noticeable disparity between the F1 Score and Accuracy of those classifiers. In the case of 5.2, there is an approximately 8 percentage point difference in Recall vs. Precision due to the relatively poor Recall score of the significantly smaller Chess class (44.05%). In 5.5, Chess produced similarly poor recall (46.84%) while the similarly small FAME and Atlantic classes produced Recall and Precisions scores well below 50%, resulting in sub-par average performance in both metrics as well as a drop in F1 performance vs. Accuracy.

Indeed, the unbalanced All vs. All model in which both Motown and Stax are heavily represented versus other labels, the results suggest that although the classifier does outperform the baseline method of just labeling all Soul music as "Motown," it can only be counted on to be right about half the time. When all classes are more equally represented - and the sub-genre defying Aretha Franklin is removed as an outlier - it improves to about 3 out of 5 while the likelihood of blindly guessing correctly falls to just 1 in 4.

	Recall	Precision	F1
Motown	0.83	0.75	0.78
Stax	0.72	0.76	0.74
Chess	0.55	0.72	0.62
FAME	0.44	0.47	0.48
Atlantic	0.51	0.51	0.47

Table 6.2. Average Recall, Precision, and F1 by Production Unit.

Two labels, Motown and Stax Records, exhibited very good classifiability in nearly all circumstances. Their disproportionate representation in the MSD compared to other Soul

production units of the era perhaps speaks to their success and popularity. The prolific Chess Records underperformed in Recall in heavily unbalanced datasets, but generally did well in Precision. FAME Studios who primarily recorded artists on behalf of other labels, and Atlantic Records who routinely farmed out recording and dealt in a very diverse artist pool, were harder for the classifier to distinguish. However, when the rather ineffective unbalanced All vs. All classifier from section 5.5 is omitted in deference to the much better performing balanced All vs. All of 5.6, an improvement is evident:

	Recall	Precision	F1
Motown	0.82	0.76	0.79
Stax	0.73	0.78	0.76
Chess	0.56	0.72	0.62
FAME	0.53	0.62	0.63
Atlantic	0.69	0.61	0.60

Table 6.3. Average Recall, Precision, and F1 by Production Unit, minus underperformer.

If 5.6 is treated as an improved replacement for 5.5 rather than a complement, all Production Units score greater than 50% in all metrics, including much improved F1 scores for FAME and Atlantic.

The generally superior performance of Motown and Stax and the increase in effectiveness when Detroit native Aretha Franklin is omitted may not prove anything definitively, but it does seem to support a common narrative regarding the popularity of The Motown Sound and Memphis Soul. As time went on, it would make sense that older labels such as Atlantic and Chess would look to younger and "hotter" ones for inspiration. It is important to note that none of these labels or data points were created in isolation from each other. Berry Gordy and Jim

Stewart absolutely listened to records by Chess and Atlantic, Rick Hall was geographically close enough to Memphis to share its musical culture. Jerry Wexler and Leonard Chess were most certainly paying attention to the hits coming out of Memphis and Detroit and taking notes. Considering the challenge of overcoming this network of influences, the achieved performance of the trained classifiers appears quite understandable and satisfactory.

6.2 Future Work

One shortcoming of the current study that should be addressed is that it does not compare the accuracy of the classifier to that of an average human listener - in other words, could a person classify these songs as well or better, and how well listened would such a person need to be? A Soul Music connoisseur may likely outperform the machine, but would someone who is not familiar with the music at all struggle to do better? And how would such a listener perform after a perfunctory lesson on the history and sound of the various labels?

It also may be valuable to see how such a system would perform using Mel Frequency Cepstral Coefficients in place of Echo Nest Segments Timbre. MFCC's may potentially improve the performance of the classifier and would open the code to work with more available datasets, as MFCC"s are a more standard feature. However, doing so would require the building of a dataset for this express purpose - a Soul Music Dataset. The construction of a dataset of such a scale is beyond the scope of this project.

One aspect of performance not currently taken into strong account is rhythmic features. Although the Pocket and Tightness features do examine on a beat level the rhythmic

interpretation of within the work, a more "macro" feature may be conceptualized to explore rhythmic variation on a large scale.

Another important variable that cannot be adequately accounted for is mastering, or rather modern digital re-mastering. Mastering has a profound impact on the perceptual features of a recording, and at present there is no reliable way to determine source material in the MSD that has been re-mastered. The only cursory filter that can be implemented at this time is to eliminate songs known to be from the era whose date field in the MSD is far removed from the actual release date.

Finally, a sufficiently large round of tests similar to the David Bowie experiment may be used to generate a weight vector to correct for bias in the system.

With the currently achieved accuracy, the system may have uses in artist recommendation systems. While artists from the same label can be found using meta tags, such tags can be errantly applied. Further, no such tag exists at the recording studio level. If a listener tends to prefer music from a particular production unit, consciously or not, the system may be able to assist with finding similar songs.

With improved performance, the system may potentially be useful in forensic analysis. With many details of these recordings lost to history, the ability to identify the likely source of a recording can be an important step in helping the men and women behind its creation finally get due credit for their contributions to our musical history.

References

Billboard Charts Archive. (n.d.). Retrieved from https://www.billboard.com/archive/charts.

Brandon, C., Fraction, R., Gracey, B., Giannapoulos, A., Helseth, K., Lucas, E., … Severin, D. (2011). The Sock Hop and the Loft: Jazz, Motown, and the Transformation of American Culture, 1959-1975. National Endowment for the Humanities. Retrieved March 6, 2019 from https://humanities.wustl.edu/files/cenhum/imce/Artistic%20Resource%20Guide.pdf.

Wightman, C. (2010). Craig David 'thought Motown was a genre.' *Digital Spy, March 20, 2010.* Retrieved March 6, 2019 from https://www.digitalspy.com/music/a209774/craig-david-thought-motown-was-a-genre/.

10 Incredible Motown Tracks You Haven't Heard. (2009). *NME, January 9, 2009.* Retrieved from https://www.nme.com/blogs/nme-blogs/10-incredible-motown-tracks-you-havent-heard-1188563.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

Cogan, J., Clark, W., & Jones, Q. (2003). *Temples of Sound: Inside the Great Recording Studios*. Chronicle Books.

Licks, D., Jamerson, J., & Gordy, B. (1989). *Standing in the Shadows of Motown.* Hal Leonard.

Gordon, R. & Neville M. (Directors). (2007, August 1). Respect Yourself: The Stax Records Story. [Television series episode]. Mark Crosby, Robert Gordon, Morgan Neville, Rupert Smith, & John Walker (Producers), *Great Performances.* London, UK: BBC Four.

Mayock, J. (Director). (2010, November 12). Roll Over Beethoven: The Chess Records Story. [Television series episode]. James Mayock (Producer), *Legends.* London, UK: BBC Four.

Our History. (n.d.). Retrieved February 11, 2019 from https://famestudios.com/our-history/.

Ward, E. (2016). How Clarence Carter Put Fame Records On The Map. Retrieved February 11, 2019 from https://www.npr.org/2016/05/10/477490697/how-soul-great-clarence-carter-put-fame-records-on-the-map.

Pareles, J. (2018). Rick Hall, Architect of the Muscle Shoals Sound, Dies at 85. *The New York Times, January 3, 2018.* Retrieved February 11, 2019 from https://www.nytimes.com/2018/01/03/obituaries/rick-hall-muscle-shoals-dies.html.

Tom Dowd. (n.d.). Retrieved February 12, 2019 from https://www.rockhall.com/inductees/tom-dowd.

Foote, J. (1997). An Overview of Audio Information Retrieval. National University of Singapore.

Whitman, B., Flake, G., & Lawrence, S. (2001). Artist Detection in Music with Minnowmatch. Neural Networks for Signal Processing XI, 2001. Proceedings of the 2001 IEEE Signal Processing Society Workshop, pages 559–568. IEEE, 2002.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Ed.). New York, NY: Springer.

MathWorks (2016). Machine Learning With Matlab. Retrieved from https://www.mathworks.com/campaigns/offers/machine-learning-with-matlab.html.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. Machine Learning 20: 273. https://doi.org/10.1007/BF00994018

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica (03505596) 31 (3).

Tzanetakis, G., Essl, G., & Cook, P. (2002). Automatic Musical Genre Classification Of Audio Signals. IEEE Transactions on speech and audio processing 10 (5), 293-302.

Li, T., Tzanetakis, G. (2003). Factors in Musical Genre Classification of Audio Signals. Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.

Sturm, B. (2014). A Survey of Evaluation in Music Genre Recognition. Adaptive Multimedia Retrieval.

Pampalk, E., Flexer, A., & Widmer, G. (2005). Improvements of Audio-Based Music Similarity and Genre Classification. In ISMIR (Vol. 5, pp. 634-637).

Seyerlehner, K., Widmer, G., & Knees, P. (2010). A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In International Workshop on Adaptive Multimedia Retrieval (pp. 118-131). Springer, Berlin, Heidelberg.

Sturm, B. L. (2014). The state of the art ten years after a state of the art: Future research in music information retrieval. Journal of New Music Research, 43(2), 147-172.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. Cambridge ISMIR 270, 1-11.

Jehan, T., DesRoches, D. (2014). Analyzer Documentation. The Echo Nest Corporation.

Schindler, A., Rauber, A. (2012). Capturing the Temporal Domain in Echonest Features for Improved Classification Effectiveness. Proc. Adaptive Multimedia Retrieval. (Oct. 2012).

Annamaria, M. & Virtanen, T. (2010). Automatic Recognition of Lyrics in Singing. EURASIP Journal on Audio, Speech, and Music Processing. 2010. 10.1155/2010/546047.

Song, Y., Dixon, S., & Pearce, M. (2012). A survey of music recommendation systems and future perspectives. In 9th International Symposium on Computer Music Modeling and Retrieval (Vol. 4).

Shung, K. P. (2018, Mar 15). *Accuracy, Precision, Recall or F1?* Retrieved from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Aretha Franklin. (2001). Retrieved February 12, 2019 from https://www.michiganrockandrolllegends.com/mrrl-hall-of-fame/70-aretha-franklin.