Quantifying the Motown Sound:

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

Karl Messerschmidt

FIRST DRAFT

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: Up For Grabs!

February 14, 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. Working in the space of Soul and R&B music from the late 1950's until early 1970's, songs that are known to have been recorded by various well-known and prolific production units are retrieved from the Million Song Dataset and a Perceptual Representation is built. The songs are then converted into a Learning Representation a Support Vector Machines classifier is trained to identify the production unit of origin. 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).

Because of the relative lack of public familiarity with these artists in comparison to, for example, the exhaustive documentation of seemingly every moment of the Beatles' lives from 1963 to 1970, American Soul and R&B is often thought of as something of a monolithic genre,

or at least a collection of a small handful of subgenres. This can perhaps best be exemplified by the somewhat common mislabeling of all danceable Soul music of the 1960's as "Motown." As those who are well-versed in the genre are generally aware, Motown is of course a specific record label from Detroit, Michigan that cultivated a stable of artists and a recording process that yielded a particular sound - The Motown Sound - that became something of a musical brand. Although many artists, such as Aretha Franklin of Columbia and Atlantic Records or Fontella Bass of Chess Records, sounded very similar to the artists on Motown, they never recorded for the Motown label. 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.

Motown was not the only label producing music of its kind. Many recording studios, some with a record label attached, cultivated their own unique musical styles and sonic environments at a time long before sampling and digital emulation allowed any studio anywhere to generate the exact same sounds. These studio/label combinations - referred to here as "Production Units" - are often distinguishable to a seasoned listener. Some, like the very distinct mixing style and thunderous drums and bass lines of Stax Records, are often quite easy to recognize. Others like the musically diverse, ever-moving, and often transparent Atlantic Records are harder to pin down.

It is here that an assumption 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

[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.

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, Ellis, Whitman, & Lamere, 2011).

If this study is not measuring artist or genre, what exactly is being measured? The end goal is to identify a Production Unit. 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, without liner notes or other written histories to accompany the output the inputs are lost to history. 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, we attempt to use the musical and sonic features available to extract the following historical creative inputs:

- 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, Clark, & Jones, 2003).
- Instrumentation: The instruments commonly used on recordings. Many production units built a distinct sound around the use of specific instruments or combinations of instruments (Cogan et al., 2003).
- Space: The physical space in which the recording was made. The nature of recording during the selected time period precluded 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 electrical 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. Additionally, some studios performed a historically noteworthy change in technology at various times (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 one of 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 analysis of the results while Chapter 6 provides conclusions.

2. Historical Background

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:

- The Production Unit can be argued to have an identifiable artistic style.
- The Production Unit must have produced a sufficiently large number of recordings.

- A reasonably sized continuous set of those recordings must have been made at a single location.
- That set of recordings was made within a time period that overlaps with that of the other selected production units.
- That set of recordings represents a sufficient variety of artists.
- Finally, 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. 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 er) 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, 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).

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 & Neville, 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 (which surely helped with the prevention of standing waves). 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 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).

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

[2] A single speaker was sufficient in the early days as Stax's recordings were still being mixed in mono.
[3] Jackson's signature snare sound was achieved by placing his wallet on the rim of the drum.

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 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 side of the panels was 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 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.).

[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.

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

3.1 Music Information Retrieval

Music Information Retrieval (MIR) is a subfield of Data Mining that 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 and genre identification, lyric transcription, similarity measurements, and recommendation systems. 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, Flake, & Lawrence, 2001).

3.2 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, or *features,* 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 (MathWorks, 2016).

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 (MathWorks, 2016).

Support Vector Machines (SVM's) are a popular algorithm in music, speech, and image processing and perform well with a large number of features. They are binary classifiers but can be made to classify between three or more classes by setting up permutations of binary pairs and running multiple SVM's. 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 the n a pair (*w,b*) exists such that:

$$w^T x_i + b \geq 1 \quad \forall\, x_i \in P, \text{ and}$$

$$w^T x_i + b \leq -1 \quad \forall\, x_i \in N$$

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 (Kotsiantis, 2007).

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).

One potential problem in classification models is *overfitting*, when the decision function is too closely fit to a set of limited data points so as to negatively affect performance. 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. 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 dataset 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).

3.3 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, there should be provide highest probability of finding an adequate number of songs from each production unit that fit the necessary criteria.

3.4 Genre Classification

The problem of genre classification can provide useful insight into identifying and categorizing music based on perceptual and musical criteria. 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). 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.

3.5 Artist Classification

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 in PCM format (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 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.

3.6 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).

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 used for evaluation purposes.
- 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 long 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 the 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, reveal 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 with most systems 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:

4.3.1 Response Label

The response label represents the true class of the song. The labels used are: Motown: 1, Stax: 2, Chess: 3. 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). 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. 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 ranges 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. 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.

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.

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 and the highest scoring model is saved and its accuracy scores recorded. The model is then tested using the testing data. Each classification scheme is trained in five sessions using five different training/testing splits, and the average 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 as, subjectively, as 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 the 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. All

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.

4.6.1 Accuracy vs. Macro Average F1

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.

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.

	Motown	*Stax*	Recall	F1
Motown	56.2	4	93.36	91.37
Stax	6.6	36.2	84.58	87.16
Precision	89.49	90.05	89.71	**89.27**
			Accuracy	Mac. Ave. F1

Table 5.1. Mean Confusion Matrix, Motown vs. Stax.

	Motown	*Stax*	Recall	F1
Motown	60.2	0	100.00	73.78
Stax	42.8	0	0.00	0.00
Precision	58.45	-	58.45	**36.89**
			Accuracy	Mac. Ave. F1

Table 5.2. Mean Baseline (Zero Rule), Motown vs. Stax.

The mean observation classes of the five Motown vs. Stax classifiers is shown in Figure 1. The baseline calculation is shown in Figure 2. 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 89.71% versus 58.45%. The overall precision is similarly high for both classes, and while there is a almost 9 percentage point disparity in recall, the two classes are clearly distinguishable in most cases. Motown and Stax averaged respective F1-scores of 91.37 and 87.16 for a Macro average F1 of 89.27, significantly higher than the baseline macro average F1 of 36.89.

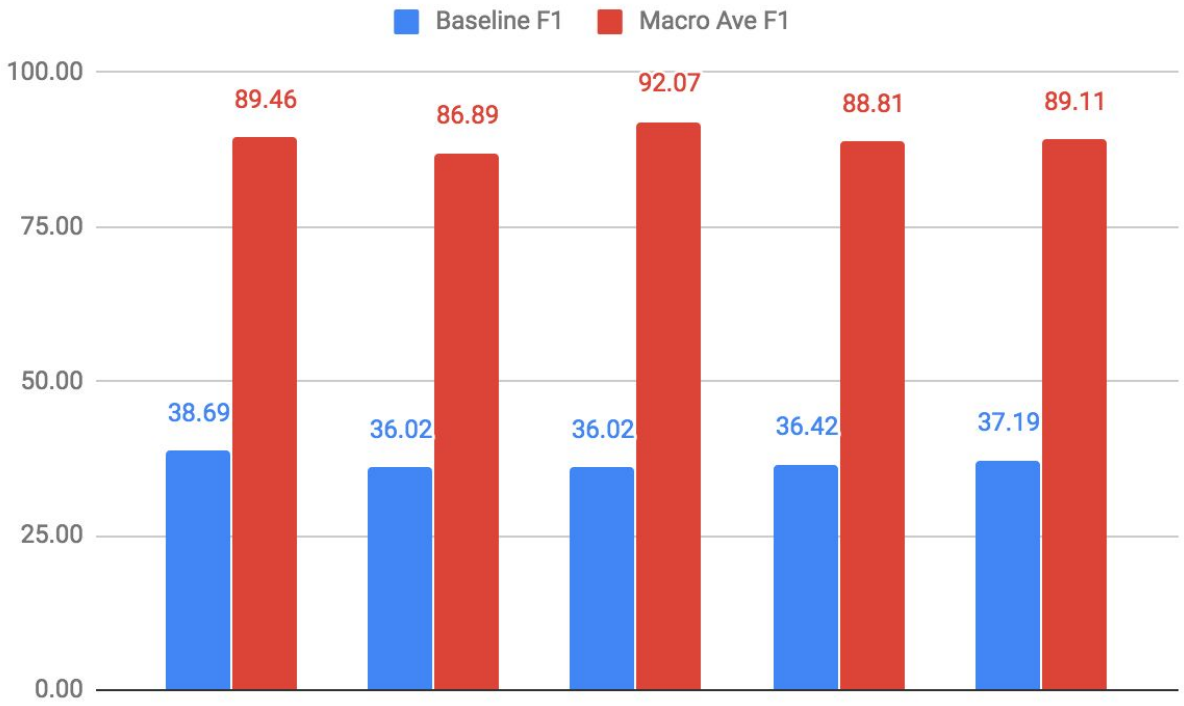

Figure 5.1. Macro Average F1 Scores of Motown/Stax Binary Classifier vs. Baseline Macro Average F1

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*	Recall	F1
Motown	51.4	4.4	1	90.49	85.65
Stax	6.8	36.6	1	82.43	81.69
Chess	5	4.4	7.4	44.05	54.35
Precision	81.33	80.62	78.72	80.85	**73.90**
				Accuracy	Mac. Ave. F1

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

	Motown	*Stax*	*Chess*	Recall	F1
Motown	58.8	0	0	100.00	66.52
Stax	43	0	0	0.00	0
Chess	16.2	0	0	0.00	0
Precision	49.83	-	-	49.83	**22.15**
				Accuracy	Mac. Ave. F1

Table 5.4. Mean Baseline (Zero Rule), 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 54.35. 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 24.14, suggesting that the trained classifiers have, at minimum, outperformed a hypothetical baseline given the balance of classes. Motown and Stax score 85.65 and 81.69 respectively for a Macro Average F1 score of 73.90, still respectable compared to the Baseline Macro Average F1 Score of 22.15. The Average Adjusted Rand Index is 0.5041.

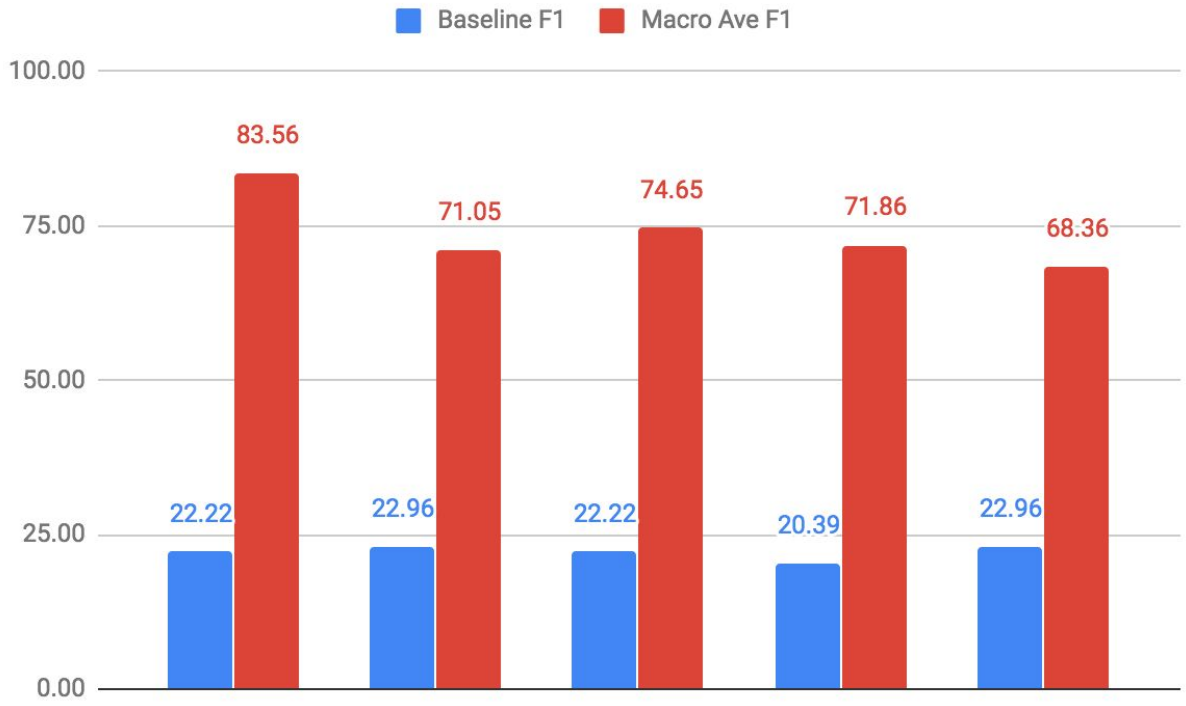

Figure 5.2. Macro Average F1 Scores of Motown/Stax/Chess Classifier vs. Baseline Macro Average F1

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*	Recall	F1
Motown	43.4	1.8	7.6	82.20	76.42
Stax	7.8	33.4	8.8	66.80	71.98
Other	9.4	7.4	27.4	62.16	61.99
Precision	71.48	78.72	62.68	70.88	**70.22**
				Accuracy	Mac. Ave. F1

Table 5.5. Mean Confusion Matrix, Motown vs. Stax vs. Other.

	Motown	*Stax*	*Other*	Recall	F1
Motown	22.8	30.4	0	42.86	40.71
Stax	18.6	32.2	0	63.39	46.33
Other	17.4	25.6	0	0.00	0.00
Precision	38.78	36.51	-	37.41	**18.15**
				Accuracy	Mac. Ave. F1

Table 5.6. Mean Baseline (Zero Rule), Motown vs. Stax vs. Chess.

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. 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 54.35 to 62.27 between the two ternary examples, the F1 scores of Motown and Stax each fell roughly 10 points to 76.42 and 71.98 respectively. As a result, the Macro Average F1 score of the model comes out to 70.22, a slight decrease versus the previous example but still an equally good comparative performance versus a baseline Macro Average F1 of 18.15.

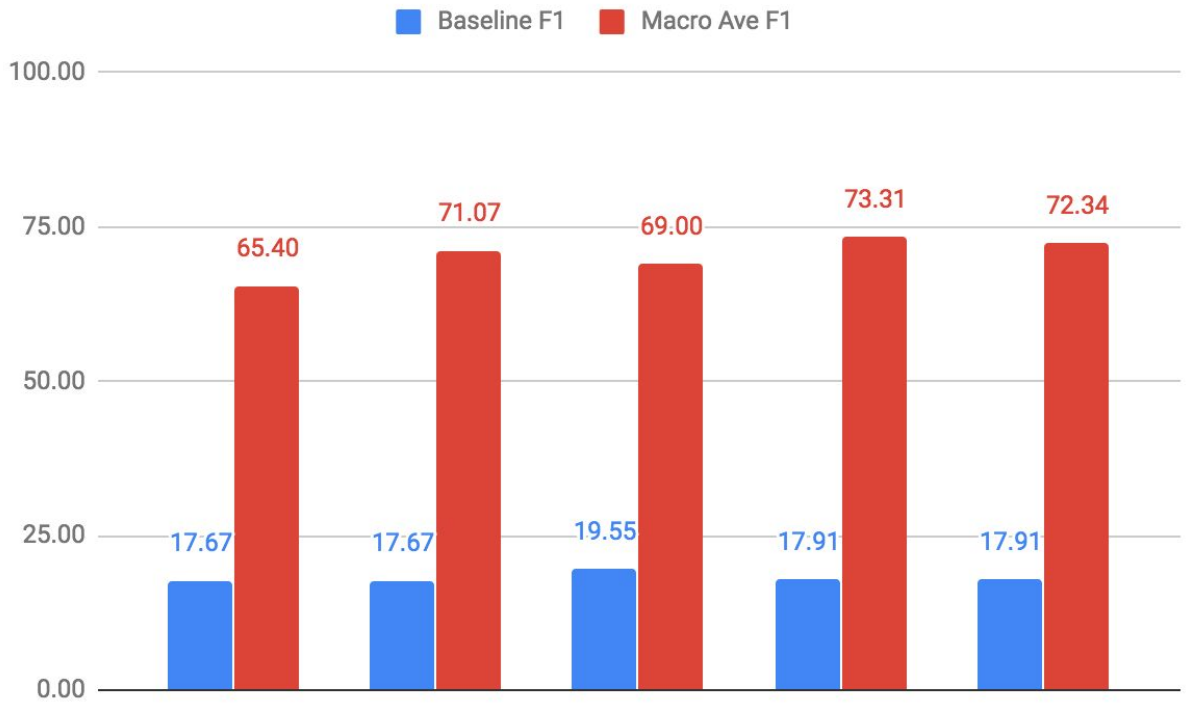

Figure 5.3. Macro Average F1 Scores of Motown/Stax/Other Classifier vs. Baseline Macro Average F1.

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*	Recall	F1
Chess	11.2	1.8	3.8	66.67	71.79
FAME	1	7.6	2.6	67.86	63.87
Atlantic	2.2	3.2	9.6	64.00	61.94
Precision	77.78	60.32	60.00	66.05	**65.87**
				Accuracy	Mac. Ave. F1

Table 5.7. Mean Confusion Matrix, Chess vs. FAME vs. Atlantic.

	Chess	*FAME*	*Atlantic*	Recall	F1
Chess	7.4	3	6.4	44.05	43.53
FAME	4.2	3.4	3.6	30.36	34.34
Atlantic	5.6	2.2	7.2	48.00	44.72
Precision	43.02	39.53	41.86	41.86	**18.64**
				Accuracy	Mac. Ave. F1

Table 5.8. Mean Baseline (Zero Rule), Chess vs. FAME vs. Atlantic.

Each class was at least once the majority class when calculating baseline scores. The Macro Average F1 reported in Figure 11 is the average of the Macro Average across all five iterations. All labels exhibited similar recall performance, 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.

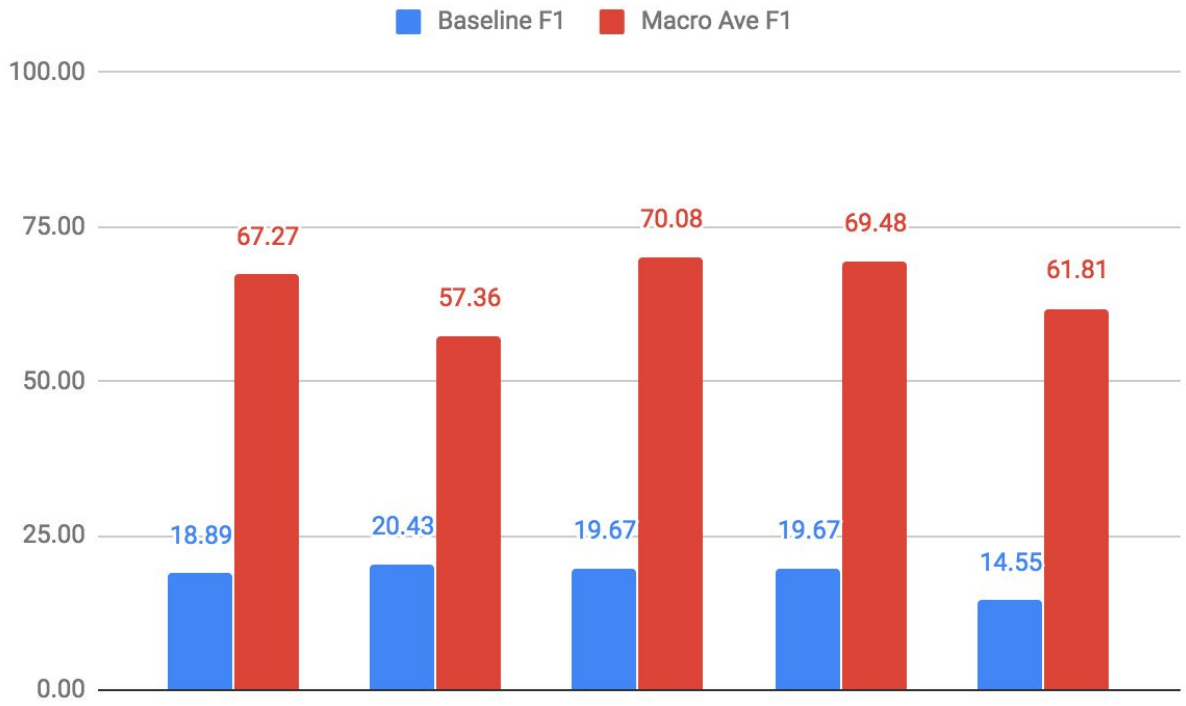

Figure 5.4. Macro Average F1 Scores of Chess/FAME/Atlantic Classifier vs. Baseline Macro Average F1.

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*	Recall	F1
Motown	49.6	3.6	1.2	1.2	2.4	85.52	77.50
Stax	8	29.8	0.8	4.6	2	65.93	67.38
Chess	4	1.6	7.4	2	0.8	46.84	56.65
FAME	2	4.6	0.2	2.6	1.2	24.53	18.93
Atlantic	6.2	3	0.8	5	2.8	15.73	21.13
Precision	71.06	69.95	71.15	16.88	30.43	62.55	**48.32**
						Accuracy	Mac. Ave. F1

Table 5.9. Mean Confusion Matrix, 5-way multiclass.

	Motown	*Stax*	*Chess*	*FAME*	*Atlantic*	Recall	F1
Motown	48	10	0	0	0	82.76	54.61
Stax	33.8	11.4	0	0	0	25.22	30.48
Chess	12	3.8	0	0	0	0.00	0.00
FAME	9.4	1.2	0	0	0	0.00	0.00
Atlantic	14.6	3.2	0	0	0	0.00	0.00
Precision	40.75	38.51	-	-	-	40.30	**11.49**
						Accuracy	Mac. Ave. F1

Table 5.10. Mean Baseline (Zero Rule), 5-way multiclass.

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 48.32 suggests serious limitations in the model's usefulness despite still performing well above the rather low baseline F1 of 11.49. The most apparent limitation is the very poor performance of FAME and Atlantic, which only serve to contribute noise to the system.

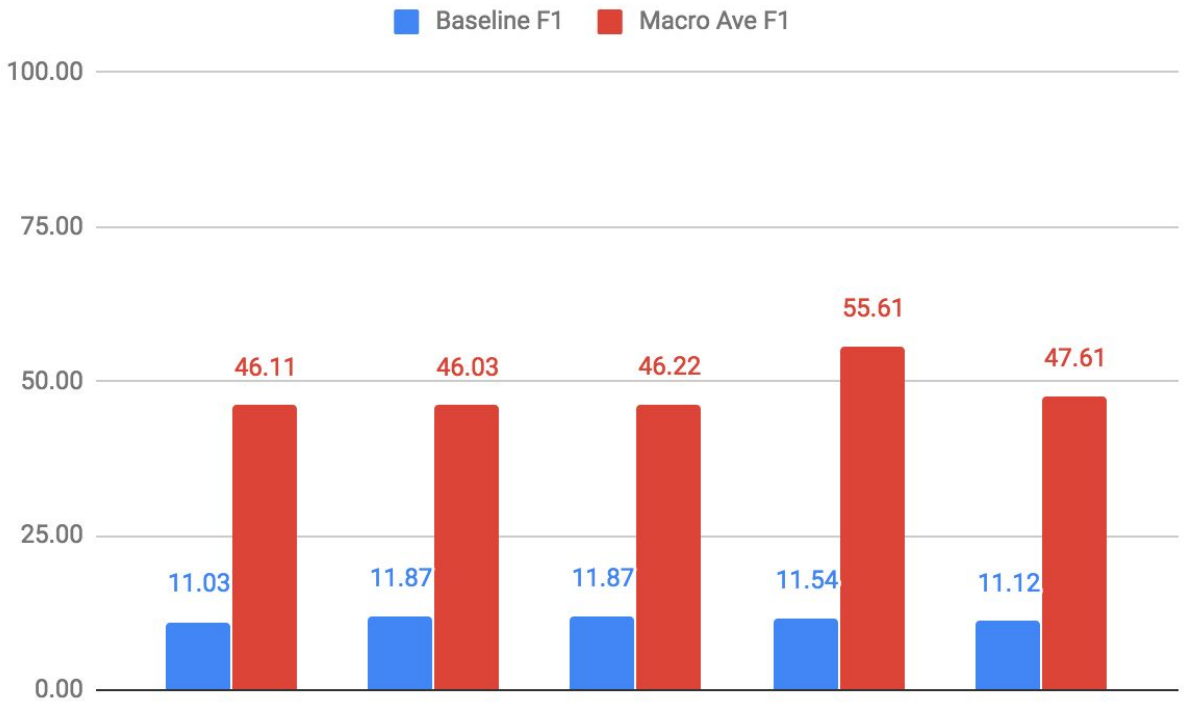

Figure 5.5. Macro Average F1 Scores of All vs. All Classifier vs. Baseline Macro Average F1.

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, 28% of Atlantic songs were identified as FAME. However, only 11% of FAME songs were identified as Atlantic.
- Rather, 43% of FAME songs 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 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.

5.5.1 The Aretha Franklin Effect

The high degrees 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.

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	0	0	0	0
Stax	0	0	0	0	0
Chess	0	0	0	0	0
FAME	3	1	0	2	0
Atlantic	10	6	8	6	0

Table 5.11. 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. The model is trained on the rest of the dataset and then shown Stevie Wonder:

	Motown	*Stax*	*Chess*	*Fame*	*Atlantic*
Motown	28	1	1	2	9
Stax	0	0	0	0	0
Chess	0	0	0	0	0
FAME	0	0	0	0	0
Atlantic	0	0	0	0	0

Table 5.12. 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%. This result suggests that a definitively Motown sonic and musical characteristic *does* exist, is classifiable, and Mr. Wonder generally exhibits it.

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
Motown	33	1	1	2	4
Stax	0	0	0	0	0
Chess	0	0	0	0	0
FAME	0	0	0	0	0
Atlantic	0	0	0	0	0

Table 5.13. Confusion Matrix of songs by Stevie Wonder only with Aretha Franklin omitted from training.

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*	Recall	F1
Motown	8.6	0.6	2.8	1	1	61.43	60.99
Stax	1.2	7.8	1.4	2	0.6	60.00	61.42
Chess	2.8	1.4	10.2	0.2	0.4	68.00	63.35
FAME	0.6	2.2	2	8.4	0.8	60.00	62.22
Atlantic	1	0.4	0.8	1.4	4.4	55.00	57.89
Precision	60.56	62.90	59.30	64.62	61.11	61.56	**61.18**
						Accuracy	Mac. Ave. F1

Table 5.14. Confusion Matrix, All vs. All with Motown and Stax limited to 70 random songs each, Aretha Franklin omitted.

	Motown	*Stax*	*Chess*	*FAME*	*Atlantic*	Recall	F1
Motown	3.2	0	7.8	3	0	22.86	23.88
Stax	2.8	0	8.4	1.8	0	0.00	0.00
Chess	2.6	0	9.8	2.6	0	65.33	36.70
FAME	2.6	0	7.8	3.6	0	25.71	26.87
Atlantic	1.6	0	4.6	1.8	0	0.00	0.00
Precision	25.00	-	25.52	28.13	-	25.94	**8.23**
						Accuracy	Mac. Ave. F1

Table 5.15. Baseline Confusion Matrix, All vs. All with Motown and Stax limited to 70 random songs each, Aretha Franklin omitted.

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 61.18 performs well above the

baseline of 8.23. The accuracy does drop slightly to 61.56 from 62.55, but the baseline score of the balanced dataset is much lower versus the unbalanced (40.30 vs. 25.94).

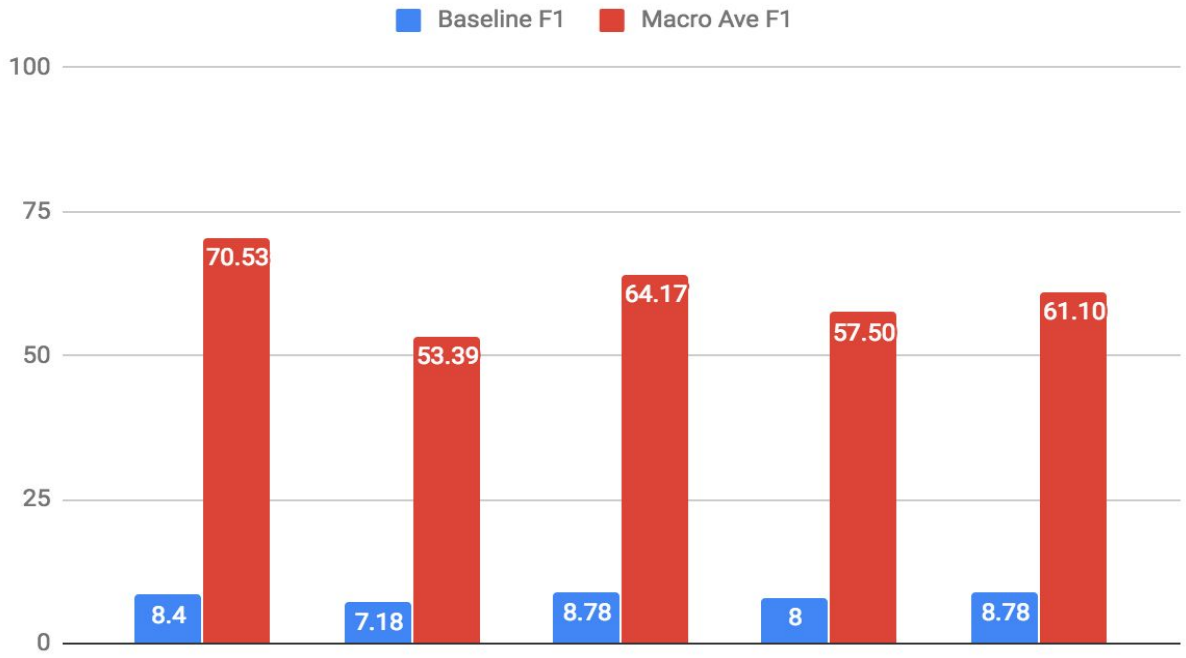

Figure 5.6. Macro Average F1 Scores of All vs. All (Balanced) Classifier vs. Baseline Macro Average F1.

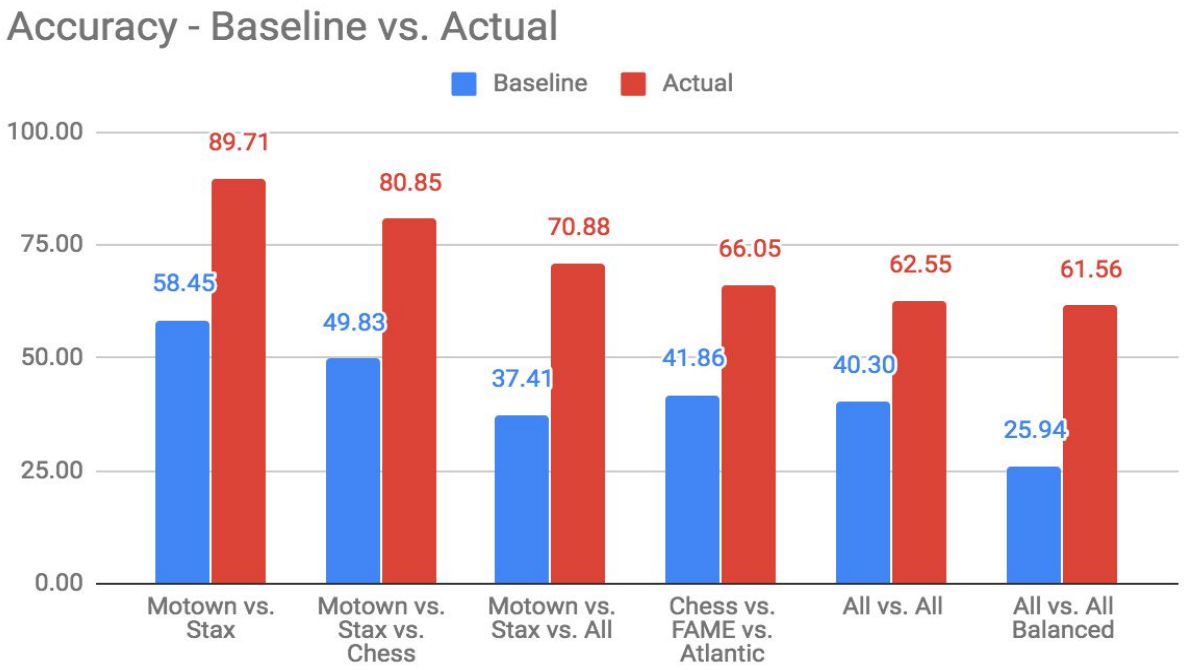

Figure 5.7. Accuracy of All Classifiers vs. Baseline Accuracy.

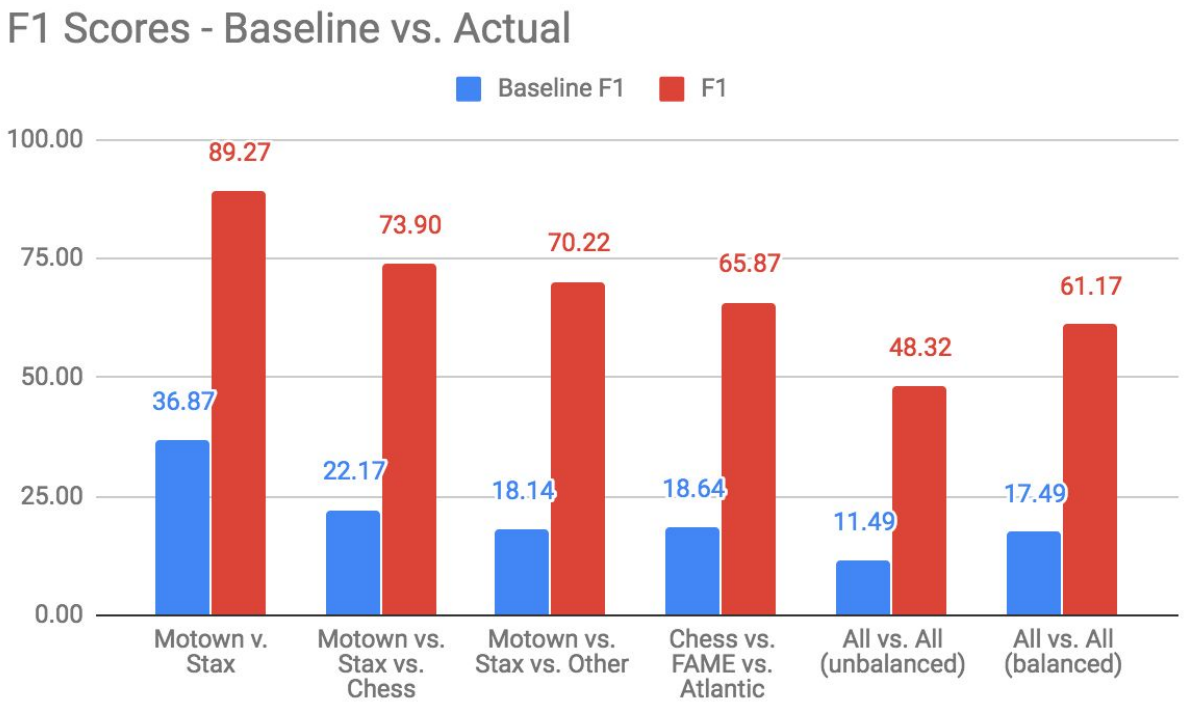

Figure 5.8. F1 Scores of All Classifiers vs. Baseline F1.

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	58.45	36.87	88.97	89.77	89.71	89.27
Motown vs. Stax vs. Chess	49.83	22.17	72.32	80.22	80.85	73.90
Motown vs. Stax vs. Other	37.41	18.14	70.33	70.96	70.88	70.22
Chess vs. FAME vs. Atlantic	41.86	18.64	63.53	66.03	66.05	65.87
All vs. All (unbalanced)	40.30	11.49	47.71	51.90	62.55	48.32
All vs. All (balanced)	25.94	17.49	60.89	61.70	61.56	61.17

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	82.60	74.78	78.39
Stax	71.95	76.45	73.93
Chess	55.00	71.74	61.54
FAME	43.56	47.27	48.34
Atlantic	51.35	50.51	46.98

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	81.87	75.72	78.61
Stax	73.45	78.07	75.56
Chess	55.53	72.35	62.38
FAME	53.08	62.47	63.04
Atlantic	69.17	60.56	59.91

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 compliment, 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.

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.

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.

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

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.

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).

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.

Appendix A: Results in Detail

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	93.36	89.49	-	91.37
Stax	-	-	84.58	90.05	-	87.16
Mean	58.45	36.87	88.97	89.77	89.71	89.27

5.1 Motown vs. Stax

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	90.49	81.33	-	85.65
Stax	-	-	82.43	80.62	-	81.69
Chess	-	-	44.05	78.72	-	54.35
Mean	49.83	22.17	72.32	80.22	80.85	73.90

5.2 Motown vs. Stax vs. Chess

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	82.20	71.48	-	76.42
Stax	-	-	66.80	78.72	-	71.98
Other	-	-	61.99	62.68	-	62.27
Mean	37.41	18.14	70.33	70.96	70.88	70.22

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

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Chess	-	-	61.11	77.78	-	71.79
FAME	-	-	46.15	60.32	-	63.87
Atlantic	-	-	83.33	60.00	-	61.94
Mean	41.86	18.64	63.53	66.03	66.05	65.87

5.4 Chess vs. FAME vs. Atlantic

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	85.52	71.06	-	77.50
Stax	-	-	65.93	69.95	-	67.38
Chess	-	-	46.84	71.15	-	56.65
FAME	-	-	24.53	16.88	-	18.93
Atlantic	-	-	15.73	30.43	-	21.13
Mean	40.30	11.49	47.71	51.90	62.55	48.32

5.5 All vs. All, Unbalanced Data

	Baseline Accuracy	Baseline F1	Recall	Precision	Accuracy	F1
Motown	-	-	61.43	60.56	-	60.99
Stax	-	-	60.00	62.90	-	61.42
Chess	-	-	68.00	59.30	-	63.35
FAME	-	-	60.00	64.62	-	62.22
Atlantic	-	-	55.00	61.11	-	57.89
Mean	25.94	17.49	60.89	61.70	61.56	61.17

5.6 All vs. All, Balanced Data