ROYAL SOCIETY
OPEN SCIENCE

Accurately Predicting Hit Songs using Neurophysiology and Machine Learning

Journal:	*Royal Society Open Science*
Manuscript ID	Draft
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Merritt, Sean H.; Claremont Graduate University Gaffuri, Kevin; Claremont Graduate University Zak, Paul; Claremont Graduate University
Subject:	behaviour < BIOLOGY, neuroscience < BIOLOGY, Artificial intelligence < COMPUTER SCIENCE
Keywords:	Prediction, Immersion, Music, Neuroscience, Classification
Subject Category:	Psychology and cognitive neuroscience

Author-supplied statements

Relevant information will appear here if provided.

Ethics

Does your article include research that required ethical approval or permits?:
Yes

Statement (if applicable):
This study was approved by the Institutional Review Board of Claremont Graduate University (#3574) and all participants gave written informed consent prior to inclusion.

Data

It is a condition of publication that data, code and materials supporting your paper are made publicly available. Does your paper present new data?:
Yes

Statement (if applicable):
The data, both observed and synthetic, are available at Open ICPSR 149821.

Conflict of interest

I/We declare a competing interest

Statement (if applicable):
The senior author (PJZ) is the founder of the commercial software platform used in this study by had influence on the analysis or results.

Accurately Predicting Hit Songs using Neurophysiology and Machine Learning

By

Sean H. Merritt[1], Kevin Gaffuri[1] and Paul J. Zak[1,2]

Abstract

Identifying hit songs is notoriously difficult. Traditionally, song elements have been measured from large databases to identify the lyrical aspects of hits. We took a different methodological approach, measuring neurophysiologic responses to a set of songs provided by a streaming music service that identified hits and flops. We compared several statistical approaches to examine the predictive accuracy of each technique. A linear statistical model using two neural measures identified hits with 69% accuracy. Then, we created a synthetic set data and applied ensemble machine learning to capture inherent nonlinearities in neural data. This model classified hit songs with 97% accuracy. Applying machine learning to the neural response to first minute of songs accurately classified hits 82% of the time showing that the brain rapidly identifies hit music. Our results demonstrate that applying machine learning to neural data can substantially increase classification accuracy for difficult to predict market outcomes.

Keywords: Prediction, Immersion, Music, Neuroscience, Classification

[1] Center for Neuroeconomics Studies, Claremont Graduate University, Claremont, CA 91711-6165

[2] Immersion Neuroscience, 1887 Whitney Mesa Dr. #2358, Henderson, NV 89014

Introduction

Every day, 24,000 new songs are released worldwide (Pandora, 2018). That's 168,000 new songs every week. People are drowning in choices. The surfeit of choices makes it difficult for streaming services and radio stations to identify songs to add to playlists. Music distribution channels use both human listeners and artificial intelligence models to identify new music that is likely to become a hit. Unfortunately, the accuracy of predictions has generally been low (Prey, 2018). This has been called the "Hit Song Science" problem (McFee, Bertin-Mahieux, Ellis & Lanckriet, 2012). The inability to predict hits means that artists are often underpaid for their work and music labels misallocate production and marketing resources when seeking to build audiences for new music (Byun, 2016). The inability to curate desirable music also causes audiences move between platforms searching for music they enjoy (Prey, 2018).

People want new music, but generally prefer songs similar to those they already know (Askin & Mauskapf, 2017; Ward, Goodman & Irwin, 2014). Music streaming services have invested in technologies to identify and introduce new music customized to subscribers' existing playlists. Spotify does this with "Discover Weekly," a playlist of 30 new songs subscribers receive every Monday morning. Pandora classifies new music using 450 attributes in its Music Genome Project and introduces new music using a service called "Personalized Soundtracks" (Carbone, 2021). Tracking what people add to their playlists boosts the likelihood of songs showing up in related playlists thereby building support leading to a hit (Turk, 2021). Nevertheless, less than 4% of new songs will become hits (Interiano et al., 2018).

Predicting hits in entertainment is a long-standing problem (Litman, 1983). Predicting hit movies has been no better than a coin flip even after considering the director, the stars, budget, time of year of release, and whether or not the movie is a sequel (Lash & Zhao, 2016; Chang & 2005; Sharda & Delen, 2006). Various methods have been used to predict hit music, including the analysis of lyrics, blog postings, social media mentions, and brain activity (Singhi & Brown, 2014; Dhanaraj & Logan, 2005; Abel, Diaz-Aviles, Henze, Krause, & Siehndel, 2010; Berns & Moore, 2012; Araujo, Neto, Nakamura & Nakamura, 2017). Yet, predictive accuracy for most studies is quite low.

Ex post, experts offer rationale for why hits were "inevitable" (Rodman, 2020). Yet, the apparent inevitability that some entertainment will become popular is challenged by studies that use ex-ante self-report for prediction (Morton, 1996). Direct and indirect measures of self-reported "liking" poorly predict aggregate outcomes (Hazlett & Hazlett, 1999; Wolfers & Zitzewitz, 2004; Bar-Anan, Wilson & Hassin, 2010; John, Emrich, Gupta, & Norton, 2017). When assessing music, "liking" is often anchored to familiarity resulting in poor ratings for unfamiliar songs (Ward, Goodman & Irwin, 2014). Moreover, using Likert self-report scales to predict popularity may be asking too much of study participants. Music is meant to elicit emotional responses that arise outside of conscious awareness and are often poorly reported (Thomas & Diener, 1990; Robinson & Clore, 2002). One way to avoid the inaccuracy of self-report is to directly measure neurophysiologic responses to music.

Emotional responses emanant from multiple brain regions rather than being localized to a one or a few structures (Adolphs & Anderson, 2018). As a result, peripheral rather than central measures of neural activity may better capture the activity of neural circuits that process emotional stimuli (Mauss & Robinson, 2009) including responses to music (Koelsch, 2018; Coutinho & Cangelosi, 2011). This is consistent with the James-Lange theory of emotion in which neurophysiologic responses induce an emotional feeling (Derryberry & Tucker, 1992; McGaugh & Cahill, 2003; Barrett, 2006; Kreibig, 2010; Barrett & Westlin, 2021). While there is no one best way to measure neurophysiologic responses to emotional stimuli, peripheral measures appear to be more robust than central measures (Golland, Keissar & Levit-Binnun, 2014). For these reasons, the study here measured peripheral rather than central neurophysiologic responses.

While neural activity has been shown to add predictive power to self-reports, neural signals alone generally have poorly predictive accuracy for population outcomes (Berkman & Falk, 2013; Genevsky, Yoon & Knutson, 2017; Dmochowski et al., 2014; Falk, Berkman, Mann, Harrison, & Lieberman, 2010; Falk, Berkman, Whalen, & Lieberman, 2011; Genevsky, Yoon & Knutson, 2017). For example, a study using functional MRI to predict music popularity showed an improvement over self-report but predictive accuracy was still well-below 50% (Berns & Moore, 2012). One reason for this may be the inherent nonlinearity of neural signals used as inputs into linear predictive models. While some researchers have directly modelled the nonlinear components of neural responses (Barraza, Alexander, Beavin, Terris, & Zak, 2015), this is atypical.

A recent approach seeks to predict outcomes from neural data using machine learning. Machine learning more effectively integrates nonlinear effects into predictions (Wei et al., 2018; Guixeres et al., 2017). Nevertheless, machine learning analyses may be subject to overfitting data (Lemm, Blankertz, Dickhaus & Müller, 2011). Overfitting can be reduced by using a limited the number of neural data streams (Jabbar & Khan, 2015), an approach we take here. Our analysis compares the classification accuracy of traditional linear predictive models to machine learning models using neurophysiologic measures alone.

Methods

Participants. Thirty-three participants (47% female) were recruited from the Claremont Colleges and surrounding community. Participants ranged in age from 18 to 57 (M= 24.25, SD= 10.47). This study was approved by the Institutional Review Board of Claremont Graduate University (#3574) and all participants gave written informed consent prior to inclusion. The data were anonymized by assigning an alphanumeric code to each participant.

Procedure. After consent, participants were seated at and fitted with Rhythm+ PPG cardiac sensors (Scosche Industries, Oxnard, CA). Music was played through a speaker system to groups of 5-8 participants in a medium-sized lab. Participants were informed that they would listen to 24 recent songs and asked about their preferences for each one. They then completed a short survey on demographics. The study lasted approximately one hour and participants were paid $15 for their time. Figure 1 shows the study timeline.

Figure 1: Timeline of the experiment

Neurophysiology. A commercial platform (Immersion Neuroscience, Henderson, NV) was used to measure neurophysiologic responses. Neurophysiologic immersion combines signals associated with attention and emotional resonance collected at 1Hz. The attentional response is

associated with dopamine binding to the prefrontal cortex while emotional resonance is related to oxytocin release from the brainstem (Zak, 2020; Zak & Barraza, 2018; Barraza & Zak, 2009). Together these neural signals accurately predict behaviors after a stimulus, especially those that elicit emotional responses (Barraza et al., 2015; Lin, Grewal, Morin, Johnson & Zak, 2013). The Immersion Neuroscience platform ingests device-agnostic heart rate data to infer neural states from activity of the cranial nerves using the downstream effects of dopamine and oxytocin (Ježová et al., 1985; Zak, 2012; Barraza, et al., 2015). The algorithms that measure immersion from cranial nerve activity are cloud-based and the platform provides an output file used in the analysis. We chose to measure neurologic immersion for this study because singing induces oxytocin release (Keeler et al, 2015) as does listening to music (Ooishi et al., 2017; Nilsson, 2009), though the effect is inconsistent (Harvey, 2020). The Immersion omnibus measure was expected to be more predictive than oxytocin alone or peripheral neural measures such as electrodermal activity (Ribeiro et al., 2019). Whether neurologic immersion can accurately classify hit songs is a new use of this measure.

The independent variables were average immersion for each song as well as two additional variables we derived from immersion data. The first we call peak immersion, defined as

$$\int_{t=0}^{T} (v_{it} > M_i) d_t / Im_i$$

where v_{it} is average neurophysiologic immersion across participants in song *i* at time *t* to the end of the song at time *T*, M_i is the median of the average time series of immersion for the duration of song *i* plus the standard deviation of song *i* across all participants who listened to that song divided by the sum of total immersion Im_i for song *i*. That is, peak immersion cumulates the highest immersion moments during the song relative to the song's total immersion. The second variable we created is called retreat. Neurologic retreat cumulates the lowest 20% of immersion averaged across participants for each song.

Songs. Staff from an online streaming service choose 24 songs for this study without input from the researchers. The streaming service also provided the definition of hits or flops. This resulted in a "clean" experiment as song choice could not be cherry-picked for the study and the criterion for a hit was established in advance. Thirteen songs were deemed "hits" with over 700,000 streaming listens, while the other 11 were flops. The songs had been released for no more than six months and spanned genres that included rock (Girl In Red "Bad Idea"), hip-hop (Roddy Rich "The Box"), and EDM (Tonnes and I "Dance Monkey"). Song order was counterbalanced and song start and stop times were synchronized with physiologic data. The 24 songs were used as the unit of analysis.

Surveys. After each song, participants were asked to rank how much they liked the song (1 to 10), if they would replay the song (0,1), recommend the song to their friends (0,1), if they had heard it previously to assess familiarity (0,1), and if they found the song offensive (0,1). We also showed participants lyrics from the song and lyrics created by the researchers and asked them to identify the song lyrics to measure their memory of the song (0,1).

Market Data. The streaming service provided the researchers with market data from their platform. These included the number of song streams that varied from 4,000 (NLE Choppa "Dekario") to over 32 million ("Dance Money"). Additional data included the number of streaming stations that carried the song and online likes.

Statistical Analysis. We used a sequence of statistical approaches, increasing in sophistication, to assess the predictive accuracy of neurophysiological variables. This was done so that the models can be directly compared. The analysis begins with tests of mean differences for self-report and neurophysiologic data comparing hits and flops using Student's t-tests (for readability, denoted "t-test"). Parametric relationships were examined using correlations while logistic regressions were estimated to establish predictive accuracy. Sensitivity analysis was conducted by analyzing the first minute of data and re-assessing the likelihood of a song being a hit.

In order to improve predictive accuracy, we trained a bagged machine learning (ML) model. Bagged models are a type of ensemble ML model that tests several machine learning algorithms in an attempt to improve accuracy above that of a single model (Dietterich, 2000). Bagged models do this by taking the output of each model individually and making a prediction based on the weighted average prediction of each model. The `SuperLearner` package in R was used to train and test the weighted bagged models. We included common machine learning classification algorithms in the analysis, including logistic regression, k-nearest neighbors, neural nets, and support vector machines.

Logistic regression can be considered a machine learning method since it is designed as a statistical binary classifier. Support vector machines (SVMs) trains on data by fitting a hyperplane to separate classifications. These hyperplanes can be non-linear making them well-suited for neurophysiologic data. K-nearest neighbors (KNN) uses training data to create boundaries between different classification labels. It does this by iterating through each data point and using the k-nearest observations to determine boundaries for classification. Artificial neural networks (ANN) attempt to make predictions in a way that mimics the neural patterns of the brain. It takes each variable as an input and uses a series of linear and non-linear transformations to map them into outputs. These transformations are weighted using a backpropagation algorithm seeking to improve predictive accuracy (James, Witten, Hastie & Tibshirani, 2017).

Bagged ML maximizes the predictive accuracy by comparing the predicted value of each algorithm using a training set to the actual value. It then combines algorithms by minimizing cross-validated risk (van der Laan et al. 2007; Polley & vaan der Laan, 2010) weighting each one by its contribution to accuracy. The final predicted value is calculated by the sum of the predicted value for each algorithm multiplied by the derived weights,

$$\hat{Y}_i = \sum_{j=1}^{N} \beta_j \hat{Y}_{ij},$$

where β_j is the weight of algorithm j and $\hat{Y}_{ij}$is the predicted value for song i by algorithm j. For details, see Polley & vaan der Lan (2010).

To find optimal parameter settings we used 5 fold cross-validation. The logistic regression used the optimal cost settings (1,10,100), the number of neighbors for KNN was (3,5,8,10), cost (1,10,100) and kernel (radial, polynomial, hyperbolic tangent) were used for SVM, and an activation function (linear, softmax), while layer size (1,5,10), and decay (0,1,10) were used for the ANN. Optimal settings were identified as logit C = 1, KNN k = 3, SVM C = 10, kernel = tanh, and ANN function = softmax, layer size = 5, and decay = 1.

Small data sets are not appropriate for machine learning as they lead to high bias in their results (Vabalas et al., 2019). To address this, we created a synthetic set data with 10,000 observations using the `synthpop` package in R (Nowak et al., 2016). This standard automated procedure creates observations by repeatedly randomly sampling the joint distribution of the data. This technique is used when obtaining large datasets is infeasible, including analyses of computer

vision (Mayer et al, 2018), sensitive information like hospital records (Tucker et al, 2018), and with unbalanced data (Luo et al, 2019; He et al., 2008). One-half of the synthetic data was used to train the bagged ML model and tune the hyperparameters. The other half of the synthetic data was used to test it. The Appendix compares the observed data to the synthetic dataset. Means, standard deviations and correlations are statistically identical. All participant data were used to train the models and to generate predictions.

Data Availability. The data, both observed and synthetic, are available at Open ICPSR 149821.

Results

Self-Report. Self-reported liking was statistically related to the number of streams (r = .54, N = 24, *p* = .002) when analyzing participants who were familiar with the songs. Liking was not predictive for stations (r = -.08, N = 24, p = .701) or online likes (r = .38, N = 24, *p* = .060). When analyzing songs that were unfamiliar to participants, the relationship between likes and online streams disappeared (β = -.13, N = 24, *p* = .387). This suggests an endogeneity problem: are liked songs familiar or are more familiar songs liked? In order to avoid this issue, we analyzed data only from songs with which participants were unfamiliar. These data were aggregated to the song level and were used for all subsequent analyses.

For unfamiliar songs, self-reported liking was statistically identical for hits and flops (M_{hit} = 4.49, M_{flop} = 4.48; *t*(22) = -0.05, *p* = .963, *d* = -0.02). The same held for recommend (*t*(22) = -0.21, *p* = .829, *d* = -0.09), offensive (t(22) = -0.44, *p* = .664, *d* = -0.18) and lyrics (t(20) = -0.58, *p* = .571, *d* = -0.25). None of the self-report measures correlated with steams (r = .16, p = .46), stations (-.31, p = .140), or online likes (r = .05, p = .980).

Neurophysiologic Responses. Hit songs had higher immersion than flops (M_{hit} = 4.17, M_{flop} = 4.10; *t*(22) = -2.34, *p* = .028, *d* = -0.95) while neurologic retreat and peak immersion did not differ between hits and flops (retreat: t(22) = 2.01, *p* = .057, *d* = 0.82; peak: t(22) = -0.14 *p* = .887, *d* = -0.06). Immersion was not correlated with streams (r = .03, p = .870), nor was peak immersion (r = .26, p = .202), while neurologic retreat trended toward significance (r = -.39, *p* = .057). Immersion was negatively related to age (r = -.31, p <.001) and varied by gender (Male: 4.06, Female: 4.20; t(650) = -3.04 *p* <.001, *d* = -0.26).

Accuracy. Logistic regression models were used to assess whether neurophysiologic measures could predict hits. Model 1 only included immersion. Then Model 2 added retreat to test if it would improve accuracy. Both Model 1 and Model 2 were significantly better at predicting hits than chance ($\chi^2(1)$ = 5.18, *p* = .023; $\chi^2(2)$ = 6.55, *p* = .037). Model 1 and Model 2 correctly classified 66% and 70% of the time, respectively. Model 2 classified hits with 69% accuracy and flops with 62% accuracy using only the two neurophysiologic variables (ps<.001). While retreat and immersion are correlated with each other (r = -.51, *p* = .011), Model 2 did not suffer from multicollinearity (VIF = 1.07). The results were robust to the inclusion of an indicator for offensive lyrics (p=.68). As expected the neurophysiologic variables were not statistically related to the self-reported desire to replay the song, recommend the song to others, or the number of online likes for each song (ps>.68).

Machine Learning. A bagged ML model was trained on one-half of the synthetic data using immersion and retreat as independent variables. The ML approaches that contribute more have a higher coefficient and lower risk (Polley & van der Laan, 2010). The k-nearest neighbor model contributed the most (coef = .98, risk = .018) followed by a neural net (coef = .012, risk = .18). A logistic regression and support vector machines contributed nothing to an accurate classification.

The bagged ML model was able to accurately classify the type of song 97.2% of the time. This statistically greater than the base rate (Successes = 4,800, N = 5,000, *p* < .001) using the

exact binomial test (Holland & Wolf 1973). Examining specificity and sensitivity, hits were classified correctly 96.6% of the time and flops were classified with 97.6% accuracy (Figure 2).

Next, we assessed the bagged ML model's ability to predict hits from the original 24 song data set. The bagged ML model accurately classified songs with 95.8% which is significantly better than the baseline 54% frequency (Success = 23, N = 24, $p < .001$). Only one song, Evil Spider, was classified incorrectly. This song was a flop with nearly 54,000 streams but was classified as a hit due to its high immersion.

We conducted a bootstrap procedure with 1,000 iterations on both the bagged ML model and the logistic model to compare their accuracy for hits and flops. The logit was trained on one data set (N = 5,000) and then assessed for accuracy on another set of data (N = 5,000) for each iteration. The bagged ML model predicted hits ML: CI = [1,1]; Logistic: CI = [0.67,0.73]; $t(1998) = -115.86$, $p < .001$) and flops (ML: CI = [0.82,1.00]; Logistic: CI = [0.59, 0.63]; $t(1998) = -121.13$, $p < .001$) better than the logistic model.

The model was assessed for overfitting by running a 10-fold cross validation on the bagged ML and comparing the predictive accuracy of the training set, test set, and observed data. This analysis shows that the bagged ML does not appear to overfit the test data as the accuracy is high and consistent (James et al. 2013). As expected, the accuracy is higher on the training and test synthetic data across the k-folds (~.99) compared to the N=24 observed data (~.96). Nevertheless, across the three data sets the accuracy of the model is high, similar, and consistent (Figure 3).

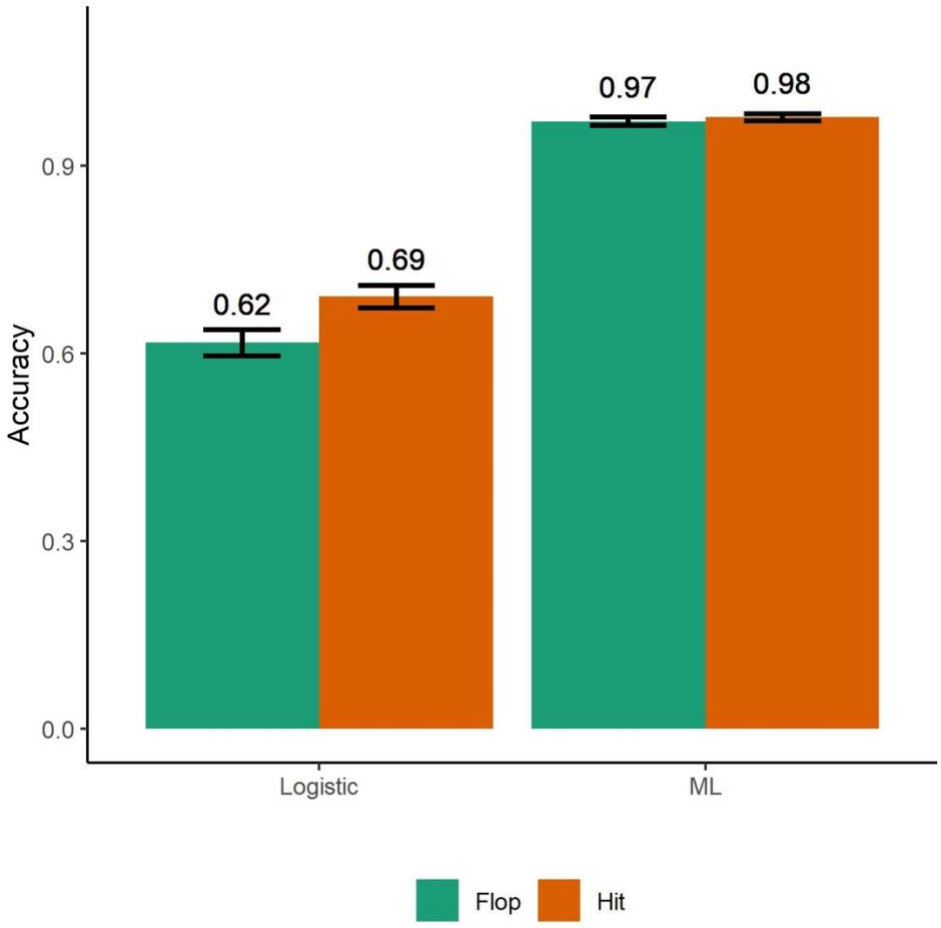


Figure 2: A traditional logistic regression model using neurophysiologic immersion and retreat as independent variables correctly classified hits with 69% accuracy and flops with 62% accuracy (N=5,000). A bagged machine learning model using the same two independent variables had accuracy of 96.6% in classifying hits while flops were classified with 97.6% accuracy. Bars are standard deviations.

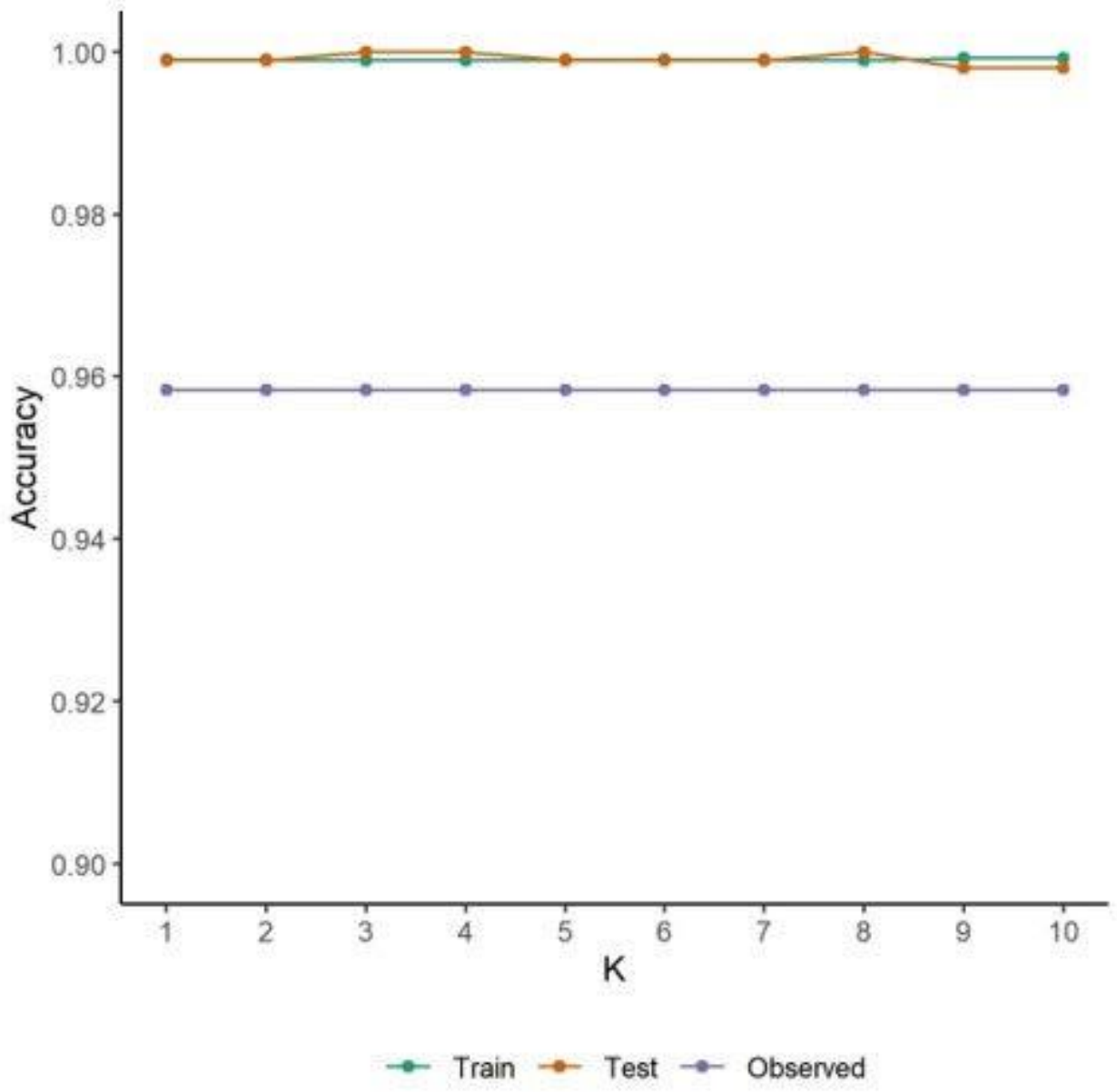


Figure 3: A 10-fold cross validation indicates the bagged ML model consistently predicts test data with ~99% accuracy. Accuracy for the observed data is consistent at ~96% indicating that the model does not overfit the data.

One Minute of Data. To establish the robustness and practical applications of our findings, we analyzed the accuracy of neurophysiology collected from the first one minute of data to identify hits. We ran the same logistic and bagged ML models described previously using only the data from the first minute of each song. We did not find a significant relationship between immersion (OR = 362.25, N = 24, p = .101) or retreat (OR = 27175.63, N = 24, p = .406) and hit songs. However, using immersion and retreat, we were able to correctly classify songs 66% of the time

using a logistic regression. Specificity and sensitivity were moderate at 77% and 56% respectively.

We created another synthetic data set to train a bagged machine learning model. Our bagged ML model had overall accuracy of 74%. It predicted hit songs with 82% accuracy and flops with 66% accuracy (Figure 4). Using the bagged ML model on the original data, we found that it was able to predict hits and flops 66% of the time. Bootstrapping the results, the bagged ML model outperformed the logistic model in classifying hit songs (ML: CI = [.80,.82], Logistic: CI = [.75,.7848]; N = 5,000, t(1998) = -41.76, $p < .001$), and flops (ML: CI = [.65,.70], Logistic: CI = [.54,.58]; t(1998) = -22.61, $p < .001$).

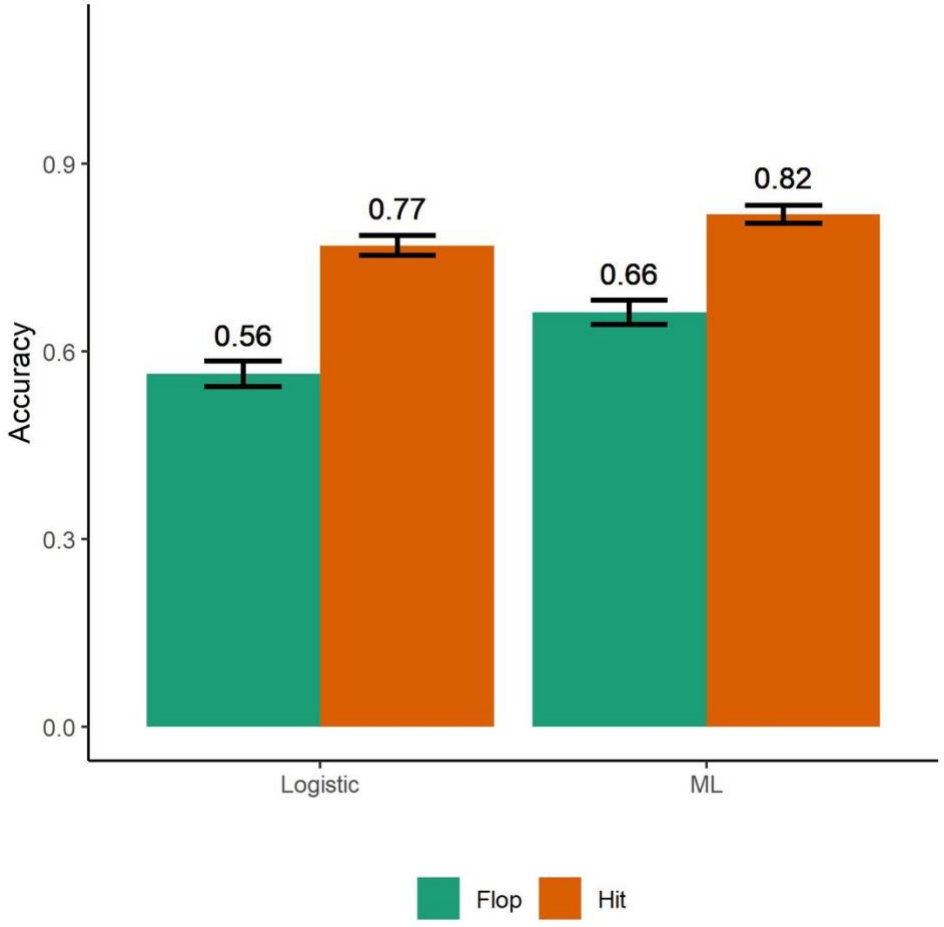


Figure 4: A logistic regression model trained on neurophysiologic *responses for the first minute of songs correctly classified hits with 77% accuracy and flops with 56% accuracy (N=10,000). The bagged machine learning model classified 74% of songs correctly. Hits were classified with 82% accuracy while flops were identified accurately 66% of the time. Model training used half of the synthetic data set (N = 5,000) using a bootstrapped evaluation of 5,000 observations per iteration for 1,000 iterations. Bars are standard deviations.*

Discussion

The key contribution of the present study is to demonstrate that neurophysiologic measures accurately identify hit songs while self-reported "liking" is unpredictive. In addition, we showed that neurophysiology, combined with machine learning, substantially improves the classification of hit songs when compared to linear statistical models. Our goal was to provide a methodology that other researchers can use to predict hit songs of different genres, in different geographic locations, and for study populations with different demographics. The approach described here can also be tested for its ability to accurately predict hits for other forms of entertainment that are known to be difficult to ascertain, including movies, TV shows, and social media posts.

Forecasting one's own behavior based on reflection is fraught and using self-report to predict market outcomes of entertainment is nearly always a fool's errand (Brenner & DeLamater, 2016; Sheeran, 2002; Woodside & Wilson, 2002). Even experts cannot identify high quality goods and services from their imitators (Ashenfelter & Jones, 2013; Almenberg, Dreber & Goldstein, 2014). While people want to hear new music, they prefer music that is similar to familiar songs generating a bias in self-reports. (Ward, Goodman & Irwin, 2014). The "Hit Song Science" problem is typically addressed by mining very large datasets (McFee, Bertin-Mahieux, Ellis & Lanckriet, 2012). We took a different approach, collecting neurophysiologic data from a small set of songs chosen by a streaming service who identified hits using their own criteria. We showed that self-reported liking only identifies hit songs if one was already familiar with the song. This is most likely due to endogeneity: participants are more likely to report they like a song if they have heard it often. Once we removed participants' familiar songs, self-reported liking ceased to predict hits.

The use of neurophysiologic data to predict aggregate outcomes is an approach that has been labelled "brain as predictor" or "neuroforecasting." This approach captures neural activity from a small group of participants to predict population outcomes (Berkman & Falk, 2013; Genevsky, Yoon & Knutson, 2017; Dmochowski et al., 2014). Neurologic data have been shown to predict outcomes more accurately than self-reports for sunscreen use, smoking reduction strategies, watching TV, and crowdfunding requests (Falk, Berkman, Mann, Harrison, & Lieberman, 2010; Falk, Berkman, Whalen, & Lieberman, 2011; Genevsky, Yoon & Knutson, 2017). While estimating predictive models using neural data is an improvement over poorly-predicting self-reported measures, the accuracy of neural forecasts have generally been no better than 50%. The closest published study to the report here used fMRI data from 28 people to predict the popularity of 20 songs. One brain region, the nucleus accumbens, was correlated with aggregate outcomes but was only able to correctly classify hits with 30% accuracy (Berns & Moore, 2012).

Our analysis showed that two measures of neurophysiologic immersion in music identified hits and flops with 69% accuracy using a traditional linear logistic regression model. A logistic regression using only the first minute of the song was nearly as accurate at 66% and was 77% accurate at classifying which songs were hits. This is a substantial improvement over the existing literature. Most of the models cited above using neural data to predict aggregate outcomes have focused on attentional responses. The neurophysiologic data we used convolves attentional and emotional responses and this may account for our improved predictive accuracy (Zak & Barraza, 2018; Zak, 2020; Lench, Flores & Bench, 2011). Emotional responses are a key component of persuasive communication because emotions capture the subjective value of an experience (Cacioppo, Cacioppo & Petty, 2018; Falk & Scholz, 2018; Doré et al., 2019; Barraza, Alexander, Beavin, Terris & Zak, 2015). The analysis here indicates that emotional responses also appear to determine which songs become hits.

Applying a bagged machine learning model to neural data improved its predictive accuracy from 69% to 97%. We also demonstrated the robustness and practical use of our approach by correctly classifying hits using the bagged ML model with 82% accuracy using the first minute of songs. It is worth noting that no demographic or self-report data were used in these

models. Further, our findings are unlikely due to chance. Machine learning using neural data has been used to identify mental illness (Khodayari-Rostamabad, 2013; Amorim et al., 2019; Stahl, 2012), epilepsy (Shoeb & Guttag, 2010; Buettner, 2019), stress (Subhani et al., 2017), and to recognize emotions (Zhang et al., 2020). Marketing researchers have applied machine learning to neural data to predict views of Superbowl ads and behavioral responses to advertising (Wei et al., 2018; Guixeres et al., 2017). As of this writing, machine learning models of music have used lyrical content rather than neural data to classify hits with only moderate accuracy (Singhi & Brown, 2014; Dhanaraj & Logan, 2005). We extended these approaches by using neural responses to music from a modest number of people to identify hits. Future work could connect neural responses to lyric classifications for additional insights.

Rather than choose a single machine learning algorithm, our use of an ensemble model eliminated a manual search for the best approach. The analysis showed that a K-nearest neighbors' (KNN) algorithm was responsible for the majority of the explanatory power. While machine learning models have been called "black boxes," our analysis showed that hit songs have higher immersion than flops and do so with a large effect size (*Cohen's d = .95*). Hits also produce less neurologic retreat than flops with a similar effect size (*Cohen's d = .82*). Another reason to use machine learning to classify hits is that neurophysiologic data are inherently nonlinear. Unlike logistic regressions, KNN's incorporate non-linear relationships making it ideally suited to neural data.

While the accuracy of the present study was quite high, there are several limitations that should be addressed in future research. First, our sample was relatively small so we are unable to assess if our findings generalize to larger song databases. The large effect sizes indicate the results are likely to be similarly accurate if other songs were tested. We created a synthetic data set to train the machine learning model. These data, while generated from human neural responses, may have overweighted subtle relationships not evident in the original data. Nevertheless, this approach has become standard when access to large samples is not available (Hoffmann et al., 2019). The use of synthetic data allows researchers to gather less direct participant data with a small or no loss in accuracy. While we found high accuracy using the observed data, we did not have access to an outside sample of songs to validate the model further. This means our model might have overfitted the data.

Measuring emotional responses using neuroscience technologies provides a new way for artists, record producers, and streaming services to delight listeners with new music. Our contribution is to show that neuroscience measurements quite accurately classify hits and flops. As neuroscience technologies enter into general use, the ability to curate music and other forms of entertainment to give people just what they want will improve existing recommendation engines benefiting artists, distributors, and consumers.

References

Abel, F., Diaz-Aviles, E., Henze, N., Krause, D., & Siehndel, P. (2010, August). Analyzing the blogosphere for predicting the success of music and movie products. In 2010 International Conference on Advances in Social Networks Analysis and Mining (pp. 276-280). IEEE.

Adolphs, R., & Anderson, D. J. (2018). *The Neuroscience of Emotion: A New Synthesis*. Princeton University Press.

Almenberg, J., Dreber, A., & Goldstein, R. (2014). Hide the label, hide the difference?. *American Association of Wine Economists*.

Amorim, E., Van der Stoel, M., Nagaraj, S. B., Ghassemi, M. M., Jing, J., O'Reilly, U. M., ... & Westover, M. B. (2019). Quantitative EEG reactivity and machine learning for prognostication in hypoxic-ischemic brain injury. *Clinical Neurophysiology*, *130*(10), 1908-1916.

Araujo, C. V., Neto, R. M., Nakamura, F. G., & Nakamura, E. F. (2017, October). Predicting music success based on users' comments on online social networks. In *Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web* (pp. 149-156).

Ashenfelter, O., & Jones, G. V. (2013). The demand for expert opinion: Bordeaux wine. *Journal of Wine Economics*, 8(3), 285-293.

Askin, N., & Mauskapf, M. (2017). What makes popular culture popular? Product features and optimal differentiation in music. *American Sociological Review*, 82(5), 910-944.

Bar-Anan, Y., Wilson, T. D., & Hassin, R. R. (2010). Inaccurate self-knowledge formation as a result of automatic behavior. *Journal of Experimental Social Psychology*, 46(6), 884-894.

Barraza, J., & Zak, P. (2009). Empathy toward strangers triggers oxytocin release and subsequent generosity. *Annals of the New York Academy of Sciences*, 1167(1), 182-189.

Barraza, J. A., Alexander, V., Beavin, L. E., Terris, E. T., & Zak, P. J. (2015). The heart of the story: Peripheral physiology during narrative exposure predicts charitable giving. *Biological Psychology*, 105, 138-143.

Barrett, L. F. (2006). Are emotions natural kinds?. *Perspectives on Psychological Science*, 1(1), 28-58.

Barrett, L. F., & Westlin, C. (2021). Navigating the science of emotion. In *Emotion Measurement* (pp. 39-84). Woodhead Publishing.

Berkman, E. T., & Falk, E. B. (2013). Beyond brain mapping: Using neural measures to predict real-world outcomes. *Current Directions in Psychological Science*, 22(1), 45-50.

Berns, G. S., & Moore, S. E. (2012). A neural predictor of cultural popularity. *Journal of Consumer Psychology*, 22(1), 154-160.

Brenner, P. S., & DeLamater, J. (2016). Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias. *Social Psychology Quarterly*, 79(4), 333-354.

Buettner, R., Frick, J., & Rieg, T. (2019, December). High-performance detection of epilepsy in seizure-free EEG recordings: A novel machine learning approach using very specific epileptic EEG sub-bands. In *ICIS*.

Byun, C. (2016). *The economics of the popular music industry: Modelling from microeconomic theory and industrial organization*. Springer.

Cacioppo, J. T., Cacioppo, S., & Petty, R. E. (2018). The neuroscience of persuasion: A review with an emphasis on issues and opportunities. *Social Neuroscience*, 13(2), 129-172.

Carbone, S. (2021). Spotify vs Pandora. What's in the box? SoundGuys, Retrieved April 15, 2021. https://www.soundguys.com/spotify-vs-pandora-36915/.

Chang, B. H., & Ki, E. J. (2005). Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. *Journal of Media Economics*, 18(4), 247-269.

Coutinho, E., & Cangelosi, A. (2011). Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, 11(4), 921.

Cyders, M. A., & Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity?. *Clinical Psychology Review*, 31(6), 965-982.

Derryberry, D., & Tucker, D. M. (1992). Neural mechanisms of emotion. *Journal of Consulting and Clinical Psychology*, 60(3), 329.

Dhanaraj, R., & Logan, B. (2005, September). Automatic Prediction of Hit Songs. In *ISMIR* (pp. 488-491).

Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., & Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications*, 5(1), 1-9.

Doré, B. P., Tompson, S. H., O'Donnell, M. B., An, L. C., Strecher, V., & Falk, E. B. (2019). Neural mechanisms of emotion regulation moderate the predictive value of affective and value-related brain responses to persuasive messages. *Journal of Neuroscience*, 39(7), 1293-1300.

Falk, E. B., Berkman, E. T., Mann, T., Harrison, B., & Lieberman, M. D. (2010). Predicting persuasion-induced behavior change from the brain. *Journal of Neuroscience*, 30(25), 8421-8424.

Falk, E. B., Berkman, E. T., Whalen, D., & Lieberman, M. D. (2011). Neural activity during health messaging predicts reductions in smoking above and beyond self-report. *Health Psychology*, 30(2), 177.

Falk, E., & Scholz, C. (2018). Persuasion, influence, and value: Perspectives from communication and social neuroscience. *Annual Review of Psychology*, 69, 329-356.

Genevsky, A., Yoon, C., & Knutson, B. (2017). When brain beats behavior: Neuroforecasting crowdfunding outcomes. *Journal of Neuroscience*, 37(36), 8625-8634.

Golland, Y., Keissar, K., & Levit-Binnun, N. (2014). Studying the dynamics of autonomic activity during emotional experience. *Psychophysiology*, 51(11), 1101-1111.

Guixeres, J., Bigné, E., Ausin Azofra, J. M., Alcañiz Raya, M., Colomer Granero, A., Fuentes Hurtado, F., & Naranjo Ornedo, V. (2017). Consumer neuroscience-based metrics predict recall, liking and viewing rates in online advertising. *Frontiers in psychology*, *8*, 1808.

Harvey, A. R. (2020). Links between the neurobiology of oxytocin and human musicality. *Frontiers in Human Neuroscience*, *14*, 350.

Hazlett, R. L., & Hazlett, S. Y. (1999). Emotional response to television commercials: Facial EMG vs. self-report. *Journal of Advertising Research*, 39(2), 7-7.

Hoffmann, J., Bar-Sinai, Y., Lee, L. M., Andrejevic, J., Mishra, S., Rubinstein, S. M., & Rycroft, C. H. (2019). Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. *Science Advances*, 5(4), eaau6792.

Hollander, M. & Wolfe, D. A. (1973), Nonparametric Statistical Methods. New York: John Wiley & Sons. Pages 15–22.

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science*, 5(5), 171274.

Jabbar, H., & Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, 163-172.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112, p. 18). New York: Springer.

Ježová, D., Jurčovičová, J., Vigaš, M., Murgaš, K., & Labrie, F. (1985). Increase in plasma ACTH after dopaminergic stimulation in rats. *Psychopharmacology*, 85(2), 201-203.

John, L. K., Emrich, O., Gupta, S., & Norton, M. I. (2017). Does "liking" lead to loving? The impact of joining a brand's social network on marketing outcomes. *Journal of Marketing Research*, 54(1), 144-155.

Keeler, J. R., Roth, E. A., Neuser, B. L., Spitsbergen, J. M., Waters, D. J. M., & Vianney, J. M. (2015). The neurochemistry and social flow of singing: bonding and oxytocin. *Frontiers in Human Neuroscience*, 518.

Khodayari-Rostamabad, A., Reilly, J. P., Hasey, G. M., de Bruin, H., & MacCrimmon, D. J. (2013). A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. *Clinical Neurophysiology*, *124*(10), 1975-1985.

Koelsch, S. (2018). Investigating the neural encoding of emotion with music. *Neuron*, 98(6), 1075-1079.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84(3), 394-421.

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems*, 33(3), 874-903.

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K. R. (2011). Introduction to machine learning for brain imaging. *Neuroimage*, 56(2), 387-399.

Lin, P-Y., Grewal, N.S., Morin, C., Johnson, W.D., & Zak, P.J. (2013). Oxytocin increases the influence of public service advertisements. *PLoS ONE*, 8(2).

Litman, B. R. (1983). Predicting success of theatrical movies: An empirical study. *Journal of Popular Culture*, 16(4), 159.

Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: a meta-analysis of experimental emotion elicitations. *Psychological Bulletin*, 137(5), 834.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, 23(2), 209-237.

McFee, B., Bertin-Mahieux, T., Ellis, D. P., & Lanckriet, G. R. (2012, April). The million song dataset challenge. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 909-916).

McGaugh, J. L., & Cahill, L. (2003). Emotion and memory: Central and peripheral contributions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Series in Affective Science. Handbook of Affective Sciences* (p. 93–116). Oxford University Press.

Moraffah, R., Karami, M., Guo, R., Raglin, A., & Liu, H. (2020). Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, *22*(1), 18-33.

Morton, A. (1996). Folk psychology is not a predictive device. *Mind*, 105(417), 119-137.Nilsson, U. (2009). Soothing music can increase oxytocin levels during bed rest after open-heart surgery: a randomised control trial. *Journal of Clinical Nursing*, *18*(15), 2153-2161.

Ooishi, Y., Mukai, H., Watanabe, K., Kawato, S., & Kashino, M. (2017). Increase in salivary oxytocin and decrease in salivary cortisol after listening to relaxing slow-tempo and exciting fast-tempo music. *PloS ONE*, *12*(12), e0189075.

Pandora. (2018). Personal communication to PJZ, New York City, Dec. 12.

Polley, E. C. & van der Laan. 2010. M. J. Super learner In prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 266. https://biostats.bepress.com/ucbbiostat/paper266

Prey, R. (2018). Nothing personal: Algorithmic individuation on music streaming platforms. *Media, Culture & Society*, 40(7), 1086-1100.

Ribeiro, F. S., Santos, F. H., Albuquerque, P. B., & Oliveira-Silva, P. (2019). Emotional induction through music: Measuring cardiac and electrodermal responses of emotional states and their persistence. *Frontiers in Psychology*, *10*, 451.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6), 934.

Rodman, S. 2020. David Foster on his hit songs and working with Whitney Houston, Celine Dion, and more in his new Netflix documentary. yahoo! news, retrived 5/6/2021 https://news.yahoo.com/david-foster-hit-songs-working-134058286.html.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), 243-254.

Sheeran, P. (2002). Intention—behavior relations: a conceptual and empirical review. *European Review of Social Psychology*, 12(1), 1-36.

Shoeb, A. H., & Guttag, J. V. (2010). Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 975-982).

Singhi, A., & Brown, D. G. (2014). Hit song detection using lyric features alone. *Proceedings of International Society for Music Information Retrieval*.

Stahl, D., Pickles, A., Elsabbagh, M., Johnson, M. H., & BASIS Team. (2012). Novel machine learning methods for ERP analysis: a validation from research on infants at risk for autism. *Developmental neuropsychology*, *37*(3), 274-298.

Subhani, A. R., Mumtaz, W., Saad, M. N. B. M., Kamel, N., & Malik, A. S. (2017). Machine learning framework for the detection of mental stress at multiple levels. *IEEE Access*, *5*, 13545-13556.

Thomas, D. L., & Diener, E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology*, 59(2), 291.

Turk, V. (2021). How to bust your Spotify feedback loop and find new music. *Wired*, Retrieved April 15, 2021. https://www.wired.com/story/spotify-feedback-loop- find-new-music/

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).

Ward, M. K., Goodman, J. K., & Irwin, J. R. (2014). The same old song: The power of familiarity in music choice. *Marketing Letters*, 25(1), 1-11.

Wei, Z., Wu, C., Wang, X., Supratak, A., Wang, P., & Guo, Y. (2018). Using support vector machine on EEG for advertisement impact assessment. *Frontiers in Neuroscience*, *12*, 76.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107-126.

Woodside, A. G., & Wilson, E. J. (2002). Respondent inaccuracy. *Journal of Advertising Research*, 42(5), 7-18.

Yang, T., Lee, D. Y., Kwak, Y., Choi, J., Kim, C., & Kim, S. P. (2015). Evaluation of TV commercials using neurophysiological responses. *Journal of Physiological Anthropology*, *34*(1), 1-11.

Zak, P. J. (2012). *The Moral Molecule: The Source of Love and Prosperity*. Random House.

Zak, P. J., & Barraza, J. A. (2018). Measuring immersion in Experiences with biosensors. Proceedings of the 11th *International Joint Conference on Biomedical Engineering Systems and Technologies*. doi:10.5220/0006758203030307

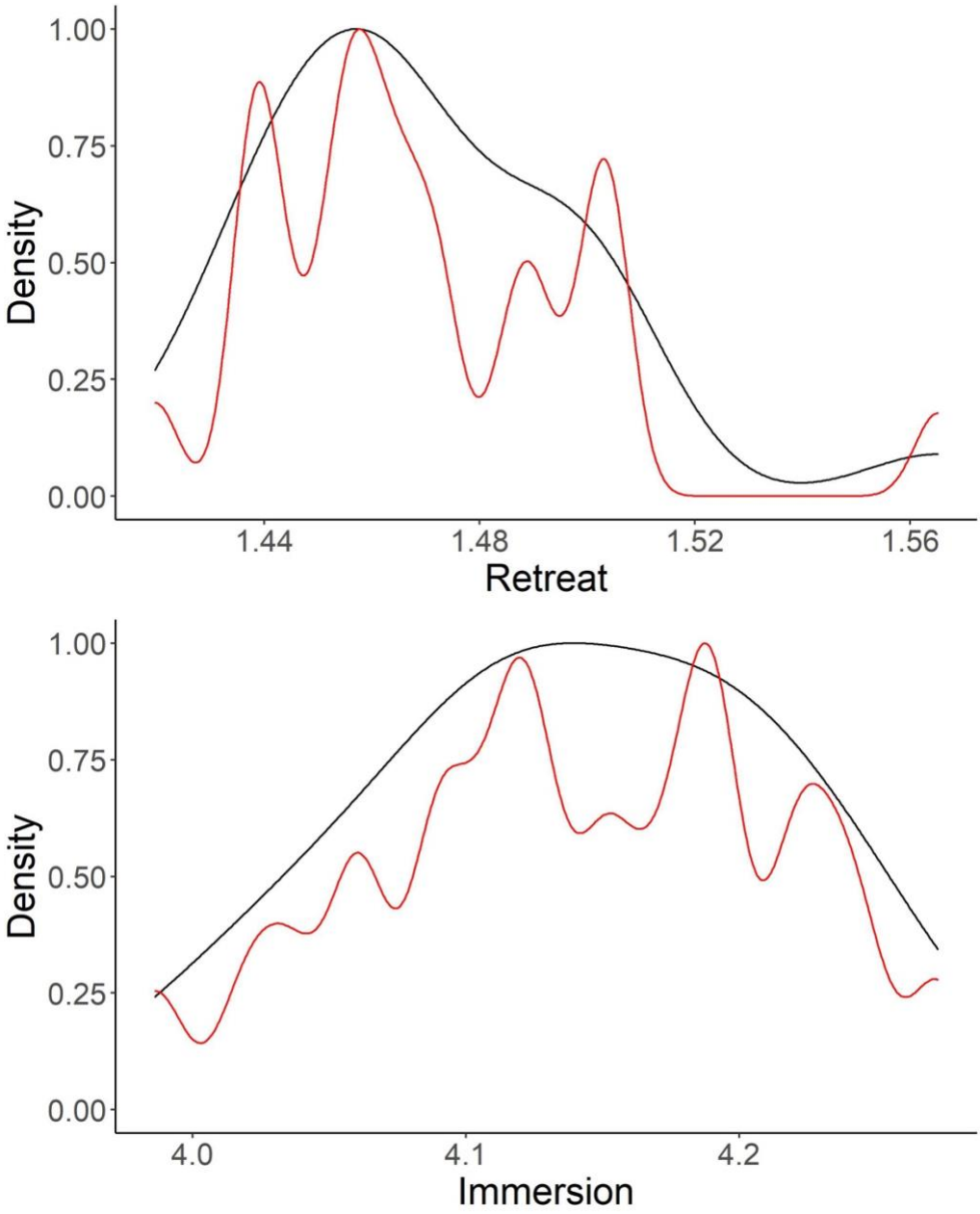
Zak, P.J. (2020). Neurological correlates allow us to predict human behavior. *The Scientist*. Oct. 1

Zak, P.J. (2022). *Immersion: The Science of the Extraordinary and the Source of Happiness*. NY: Lioncrest.

Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103-126.

Appendix

Figures A1-A3 show the distribution of the data with the observed data in black and the synthetic data in red.


1.00
0.75
0.50
0.25
0.00
Density
1.44
1.48
1.52
1.56
Retreat
1.00
0.75
0.50
0.25
0.00
Density
4.0
4.1
4.2
Immersion

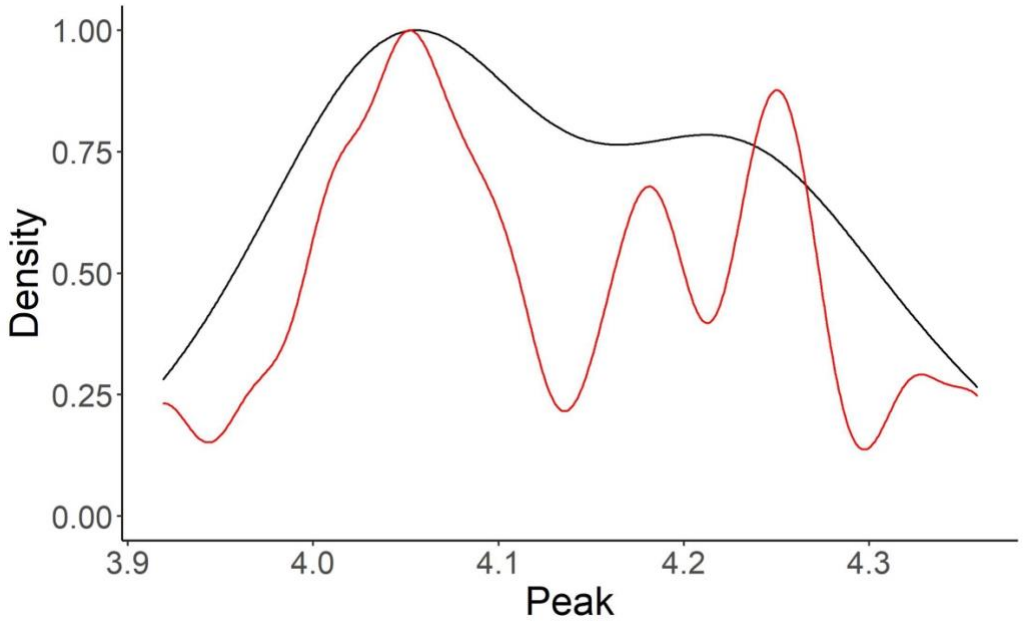


Table A1: Comparison of summary statistics for the observed and synthetic datasets.

		N	Mean	S.D.	r_hit	r_immersion	r_peak	r_retreat
Hit	Observed	24	0.54	0.51	-			
	Synthetic	10000	0.54	0.49	-			
Immersion								
	Observed	24	4.13	0.08	0.44	-		
	Synthetic	10000	4.14	0.07	0.44	-		
Peak								
	Observed	24	4.13	0.12	0.03	-0.11	-	
	Synthetic	10000	4.13	0.12	-0.02	-0.01	-	
Retreat								
	Observed	24	1.47	0.03	-0.39	-0.51	-0.18	-
	Synthetic	10000	1.47	0.03	-0.11	-0.37	-0.23	-

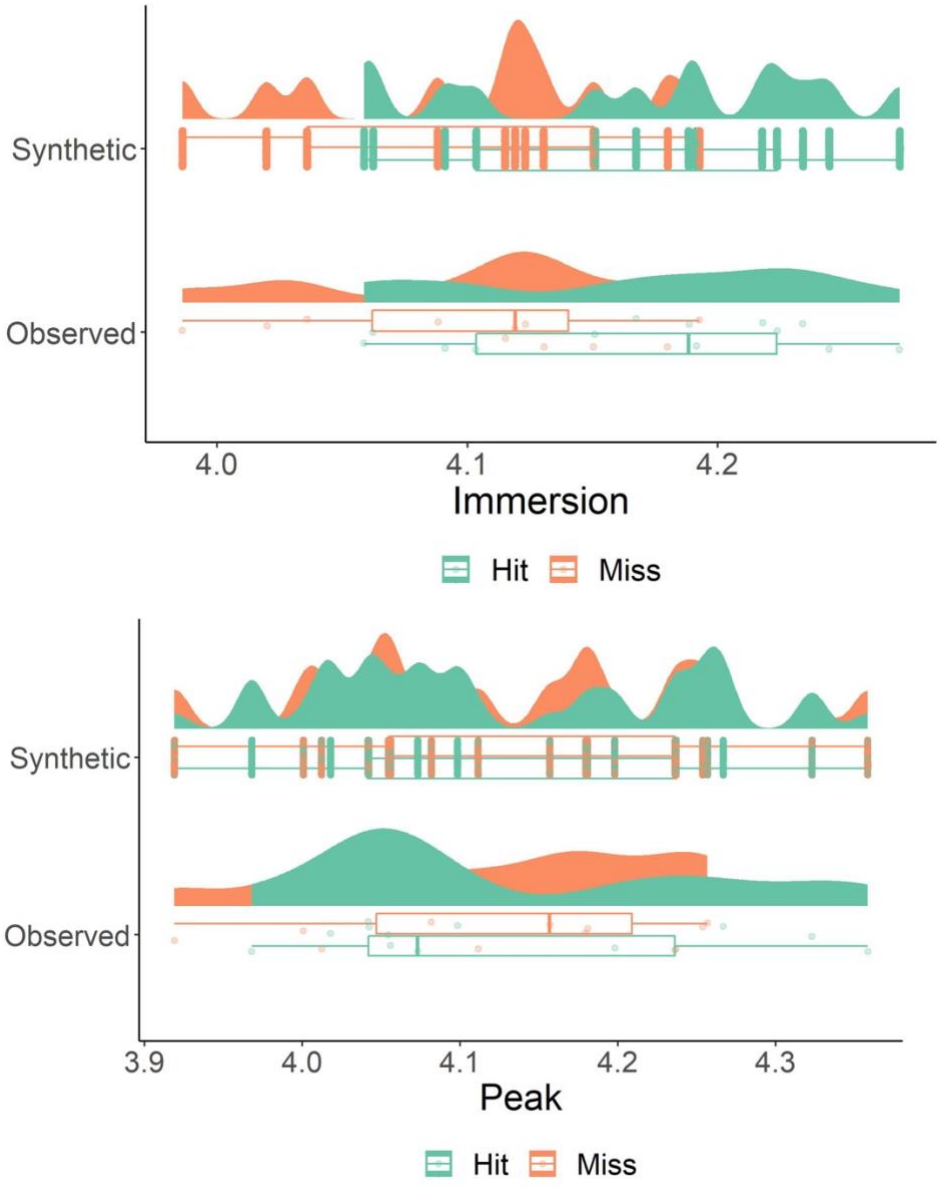
No significant differences were found between the observed and synthetic data for means (t-tests), standard deviations (F-tests), or correlations (t-tests) as shown in Table A1. P-values for each test are shown in Table A2.

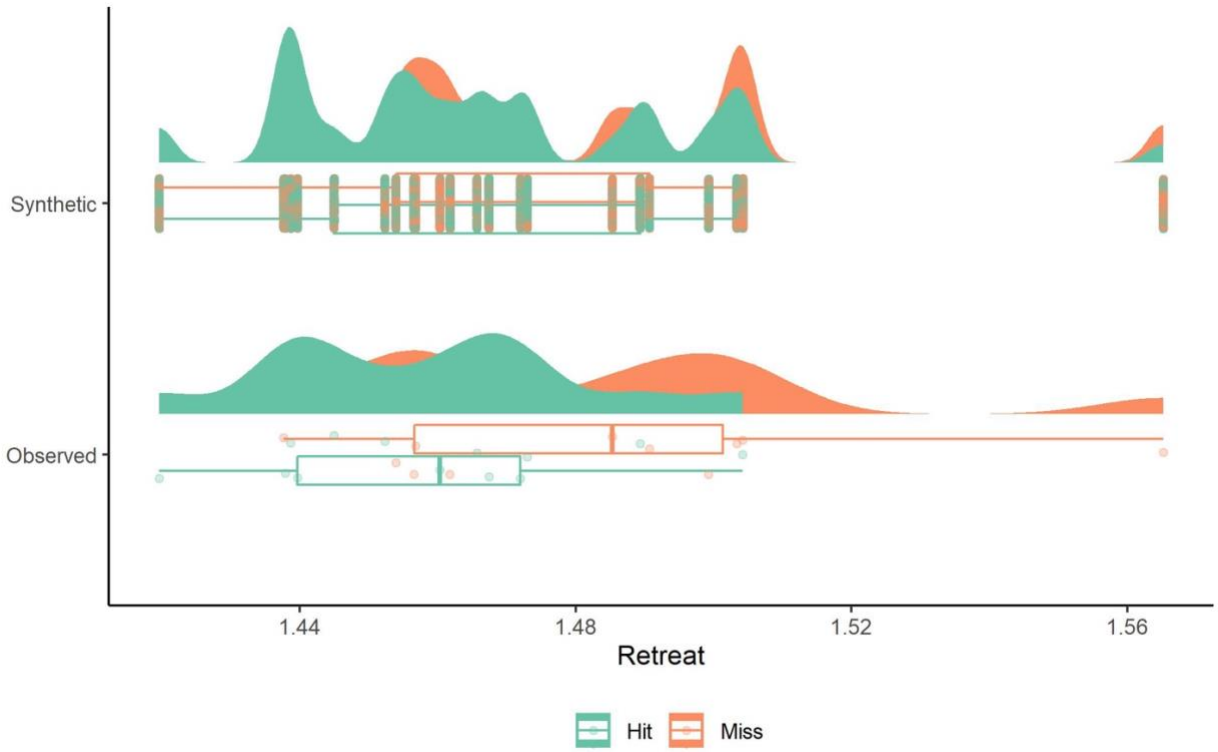
		Mean	SD	r_hit	r_immersion	r_peak	r_retreat
Hit							
	Statistic	-.04	1.05	-			
	df	23.11	23,9999	-			
	p	.969	.804	-			
Immersion							
	Statistic	-0.04	1.04	0.00	-		
	df	23.11	23,9999	10022	-		

	p	.971	.816	1.00	-		
Peak							
	Statistic	-0.04	1.04	1.06	-0.46	-	
	df	23.11	23,9999	10022	10022	-	
	p	.970	.816	.291	.649	-	
Retreat							
	Statistic	0.13	1.08	-1.37	-0.79	0.24	-
	df	23.10	23,9999	10022	10022	10022	-
	p	.896	.707	.172	.429	.813	-

We established two baselines for comparison to the bagged ML model. The first is the base rate probability of selecting a hit song. In both the synthetic and observed data this is 54%. The second baseline was established by taking the dominant observed outcome in the data (hits) and predicting how many hits in the test set. This baseline was 68%. The figures below compare the hit and miss rates for all three neurophysiologic variables for the original data and synthetic data.

Figures A4-A6: A comparison of the hit and miss rate for each neurophysiologic variable for the observed and synthetic datasets. Means, standard deviations, and ranges (25th and 75th percentiles) are shown.


Synthetic
Observed
4.0
4.1
4.2
Immersion
Hit
Miss
Synthetic
Observed
3.9
4.0
4.1
4.2
4.3
Peak
Hit
Miss


Synthetic
Observed
1.44
1.48
1.52
1.56
Retreat
Hit
Miss