Exploring Representativeness in Complex and Subjective Spaces

Crystal Wang (crystw@mit.edu, 9.660)

Julia Wagner (jnwagner@mit.edu, 9.660)

Abstract

Representativeness is a construct that has been explored before in a variety of domains such as sequences of coin flips and sets of animals. This paper expands on that work and evaluates the performance of previously studied models as these domains grow increasingly complex as well as their performance in more subjective areas like music genre representation. We also further examine, in the coin flip domain, whether participants are impacted by observing other hypotheses in the same context as the current hypothesis and determine whether this impacts model performance.

Introduction

The idea of representativeness has always been an interesting concept to study, especially since humans do not always pick the most probable representatives. For example, humans tend to choose "HHTHT" as a representative over "HHHHH" even though both are equally likely. Prior research on this topic has attempted to form a rational basis for representativeness with three main models which we will call the likelihood model, the Bayesian model, and the similarity model.

Our paper will examine how these models hold up when the space becomes more complex, such as rating representatives for coin flip sequences where the coin in question always flips "heads" after two consecutive "tails." We will evaluate the three models across increasing complexities as well as varying hypotheses types (e.g. hypotheses that have only one outcome compared to hypotheses with some element of randomness). We will additionally explore if the model that best explains representativeness changes when the participants are given the sequences and hypotheses in a different order. Finally, we explore how these models perform in the subjective field of music by applying them to the representativeness of sets of songs in the Pop and Rap genres. The song similarity metrics come from Spotify audio features which we also convert into a two-dimensional multidimensional scaling (MDS) space.

Models of Representativeness

This section will briefly describe the three models used in this paper which are described in more detail in Tenenbaum & Griffiths (2001).

Likelihood

Proposed by Gigerenzer & Hoffrage (1995), the likelihood model uses the idea of probabilistic likelihood to evaluate representativeness. For a potential representative r of a larger set h, the likelihood model essentially models the level of representativeness as $P(r|h)$, giving higher representative value to more probable representatives. While the likelihood model seems intuitive in many scenarios, there exist situations where human ratings do not align with the model predictions, such as in the coin flip example comparing "HHTHT" and "HHHHH", leading to varying levels of success depending on the domain.

Additionally, the likelihood model becomes harder to quantify in subjective areas such as music where distributions and features are much harder to define. We will therefore examine in this paper how the likelihood model could be applied to musical genre representativeness.

Bayesian

The Bayesian model, introduced in Tenenbaum & Griffiths (2001), contextualizes the current hypothesis that is being considered when rating potential representatives. It incorporates the likelihood model into a larger calculation that essentially scales the likelihood by the likelihood of the potential representative for other hypotheses. In doing so, it can address some of the failures of the likelihood model by introducing a mathematical formalization of relativity consideration. In this paper, we will explore if the Bayesian model still holds up even when the entire context or set of hypotheses is not immediately presented.

Similarity

The similarity model has been researched and developed by many, such as Kahneman & Tversky (1972), and it mostly focuses on defining similarity between a representative and another member of the set using manually determined features. It then gives a higher representative value to potential representatives that have high similarity with many other members of the set.

Similarly to the Bayesian model, the similarity model also addresses the "HHTHT" and "HHHHH" discrepancy by arguing that "HHTHT" looks more similar to other sequences than "HHHHH" for several features such as number of tails, number of alternations, etc. We will evaluate in this paper how the similarity model holds up as hypotheses, and their resulting features, grow more complex and subjective.

Coin Flips

In these experiments, participants were given six coin flip sequences along with six coins that could have been used to generate the sequences and they were asked to rate the representativeness of each sequence for each coin. These participants were split into two groups: one where they rated all of the hypotheses at once for each sequence and one where they were presented the sequences and hypotheses in a random order. Additionally, the six coins were split into three groups of three where each group had a different complexity level. The ratings were made on a scale of 1 to 7 with 1 being "least representative" and 7 being "most representative." We will analyze whether the two groups of participants had differing ratings and whether the model fit was impacted by complexity.

Setup

Coin Hypotheses We used the following coin hypotheses

- Coin 1: a coin that is fair (50% chance of flipping heads, 50% chance of flipping tails)
- Coin 2: a coin that has an 80% chance of flipping heads and a 20% chance of flipping tails
- Coin 3: a coin that always flips "H T" (i.e. alternating heads and tails, starting with heads)
- Coin 4: a coin that always flips "H" after "T" but is otherwise equally random
- Coin 5: a coin that always flips "H T T"
- Coin 6: a coin that always flips "H" after "T T" but is otherwise equally random

For each coin, we used a hidden Markov model in order to calculate $P(s_i|c_j)$ for a given sequence s_i and coin c_j. The complexity of the hypothesis is the number of prior flips that are stored in the state. For example, a fair coin (Coin 1) flips independently each time so the state stores 0 prior flips and therefore its complexity is 0. Meanwhile, a coin that always flips "H" after "T" (Coin 4) requires the most recent flip to be stored in the state and therefore its complexity is 1.

Hypothesis Probabilities As described above, each coin was modeled using a hidden Markov model, where the states S_t represent the necessary amount of recent flips and the observations F_t are the actual flips at time t.

In this section, we will detail the states, the observations, and the related probabilities for each coin. We will also label the complexities of each coin as a direct result of the state definition. For the sake of concision, we will omit probabilities that can be derived from already given probabilities (i.e. the probability of tails can be assumed to be $P(T) = 1 - P(H)$ unless otherwise given).

Note that we define "always" to a probability of 0.99 in order to have a meaningful calculation for impossible sequences. This allows us to have lower probabilities for sequences that are "more impossible" than others (i.e. contain more errors/disqualifications).

Coin 1: Fair Coin

- State: since each coin flips independently, the hidden Markov model stays in one state the entire time. Call this state s.
- Initial state: $S_0 = s$ with probability 1
- Transition probability: $P(S_t = s|S_{t-1} = s) = 1$
- Observations: $P(F_t = H|S_t = s) = P(F_t = T|S_t = s) = 0.5$
- Complexity: 0

Coin 2: 80% Heads, 20% Tails

- State: since each coin flips independently, the hidden Markov model stays in one state the entire time. Call this state s.
- Initial state: $S_0 = s$ with probability 1
- Transition probability: $P(S_t = s|S_{t-1} = s) = 1$
- Observations: $P(F_t = H|S_t = s) = 0.8$
- Complexity: 0

Coin 3: Always Heads Then Tails

- State: since the next coin depends on the previous flip, we have two states, s_H and s_T representing flipping heads or tails on the most recent flip, respectively
- Initial state: $S_0 = s_T$ with probability 0.99 since the first flip should be a heads
- Transition probability:
  - $P(S_t = s_H|S_{t-1} = s_T) = 0.99$
  - $P(S_t = s_T|S_{t-1} = s_H) = 0.99$
- Observations:
  - $P(F_t = H|S_t = s_T) = 0.99$
  - $P(F_t = T|S_t = s_H) = 0.99$
- Complexity: 1 (each state stores 1 prior flip)

Coin 4: Always Heads After A Tail

- State: since the next coin depends on the previous flip, we have two states, s_H and s_T representing flipping heads or tails on the most recent flip, respectively
- Initial state: $S_0 = s_H$ with probability 0.5 and $S_0 = s_T$ with probability 0.5

- Transition probability:
  - $P(S_t = s_H | S_{t-1} = s_T) = 0.99$
  - $P(S_t = s_T | S_{t-1} = s_H) = 0.5$
- Observations:
  - $P(F_t = H | S_t = s_T) = 0.99$
  - $P(F_t = T | S_t = s_H) = 0.5$
- Complexity: 1 (each state stores 1 prior flip)

Coin 5: Always Heads, Tails, Tails

- State: since the next coin depends on the previous two flips, we have four states, s_{ij}, for $i, j \in \{H, T\}$ where i is the second most recent flip and j is the most recent flip
- Initial state: $S_0 = s_{TT}$ with probability 0.99 and $S_0 = s_{ij}$ $i \neq T$ or $j \neq T$ with probability $1/300$ each
- Transition probability:
  - $P(S_t = s_{HT} | S_{t-1} = s_{HH}) = 0.99$
  - $P(S_t = s_{TT} | S_{t-1} = s_{HT}) = 0.99$
  - $P(S_t = s_{HT} | S_{t-1} = s_{TH}) = 0.99$
  - $P(S_t = s_{TH} | S_{t-1} = s_{TT}) = 0.99$
  - $P(S_t = s_{ij} | S_{t-1} = s_{kl}) = 0$ if $j \neq l$ (impossible state transition since most recent flip j becomes second most recent flip i on the next state)
- Observations:
  - $P(F_t = T | S_t = s_{HH}) = 0.99$
  - $P(F_t = T | S_t = s_{HT}) = 0.99$
  - $P(F_t = T | S_t = s_{TH}) = 0.99$
  - $P(f_t = H | S_t = s_{TT}) = 0.99$
- Complexity: 2 (each state stores 2 prior flips)

Coin 6: Always Heads After Tails, Tails

- State: since the next coin depends on the previous two flips, we have four states, s_{ij}, for $i, j \in \{H, T\}$ where i is the second most recent flip and j is the most recent flip
- Initial state: $S_0 = s_{ij}$ with probability 0.25 for all i, j
- Transition probability:
  - $P(S_t = s_{HT} | S_{t-1} = s_{HH}) = 0.5$
  - $P(S_t = s_{TT} | S_{t-1} = s_{HT}) = 0.5$
  - $P(S_t = s_{HT} | S_{t-1} = s_{TH}) = 0.5$
  - $P(S_t = s_{TH} | S_{t-1} = s_{TT}) = 0.99$
  - $P(S_t = s_{ij} | S_{t-1} = s_{kl}) = 0$ if $j \neq l$ (impossible state transition since most recent flip j becomes second most recent flip i on the next state)
- Observations:
  - $P(F_t = T | S_t = s_{HH}) = 0.5$
  - $P(F_t = T | S_t = s_{HT}) = 0.5$
  - $P(F_t = T | S_t = s_{TH}) = 0.5$
  - $P(f_t = H | S_t = s_{TT}) = 0.99$
- Complexity: 2 (each state stores 2 prior flips)

Sequences We gave participants the following six sequences

- Sequence 1: H H H H H
- Sequence 2: H H H H H H H H H H H H H
- Sequence 3: H T H H T T H H H T T H H T
- Sequence 4: H T T H H H H H T H
- Sequence 5: H T T H T T H T T H T
- Sequence 6: H T H T H T

Groups Group 1 was given sequence 1 first, then asked to rate its representativeness for coins 1 through 6 in order, before moving onto sequence 2, and so on.

Group 2 was given each sequence-hypothesis pair in a random order and was not given a list of all of the hypotheses before beginning.

Likelihood Model For a sequence, s_i, and a coin c_j, we calculate the likelihood $P(s_i|c_j)$ by multiplying the probabilities of the state transitions necessary to generate sequence s_i using c_j. Then, the likelihoods were transformed using a power law transformation with power γ that optimized the model performance and mapped linearly from an interval of $[0,1]$ to $[1,7]$ to match the participant ratings.

Bayesian Model Priors for each hypothesis were assumed to be equal. Model predictions were calculated using Equation (4) given in Tenenbaum & Griffiths and $P(s_i|c_j)$ as calculated in the likelihood model. Then, the resulting predictions were transformed using a power law transformation with power γ as well as through the transformation $f(x) = \frac{1}{1+e^{-\alpha x+\beta}}$ where α, β, γ were chosen to optimize model fit. Finally, the transformed values were mapped linearly from an interval of $[0,1]$ to $[1,7]$ to match the participant ratings.

Similarity Model For the similarity model, we defined five features to cover our six hypotheses:

- Feature 1: number of heads
- Feature 2: number of alternating pairs (i.e. HT or TH)
- Feature 3: number of heads that are flipped right after a tail (i.e. TH)
- Feature 4: number of HTT flips (heads then tails then tails)
- Feature 5: number of heads that are flipped right after two tails (i.e. TTH)

Let f_i represent feature i and let $F_i = [f_1, f_2, .., f_5]$ be the feature vector for sequence s_i. Define a vector of parameters $W = [w_1, w_2, ..., w_5]$ that will be fit to optimize model fit. Then, we can define the similarity of two sequences s_i, s_j as

$$sim(s_i, s_j) = \exp(-W \cdot |F_i - F_j|)$$

Then, for each hypothesis, for a given sequence length n, we construct a prototype for that hypothesis by calculating the expected value of each feature across all possible sequences of length n. Denote this prototype as $p(c_i, n)$.

Finally, we can define the model prediction as

$$R(s_i, c_i) = \frac{sim(s_i, p(c_i, len(s_i)))}{\sum_{i=1}^{6} sim(s_i, p(c_i, len(s_i)))}$$

Results

Hypothesis Context Experiment First, we used a set of five sequences and a fair coin hypothesis to test for significant differences between the two groups that would influence the experiment results:

- HTHTHTH
- HTHHT
- HHHHHH
- HHHHHHHHHHHHHHHHH
- HHHTHTH

We presented every participant in both groups with these five sequences and asked them to rate the representativeness of each sequence for a fair coin hypothesis. Then, using a standard t-test for each sequence, we tested for a significant difference between the two samples and found that none were significantly different at a 5% significance level, as shown in Table 1. Therefore, we proceeded with the experiment assuming our two groups were identically distributed and the results would therefore only be influenced by the independent variable.

Table 1: T-Test Values for Control Questions

Sequence	T-test Value
HTHTHTH	0.0
HTHHT	0.8
HHHHHH	1.7
HHHHHHHHHHHHHHHHH	0.9
HHHTHTH	1.5

This experiment mainly attempted to address two questions: (1) would seeing the sequences and coins in a random order impact the ratings given and (2) if the ratings were impacted, would the model that best fit the ratings change? Intuition would suggest that if the other hypotheses were not presented while rating a given hypothesis, then the likelihood model would perform better than the Bayesian model since the Bayesian model relies on context and relative probabilities of the other hypotheses whereas the likelihood model considers only the current sequence and hypothesis.

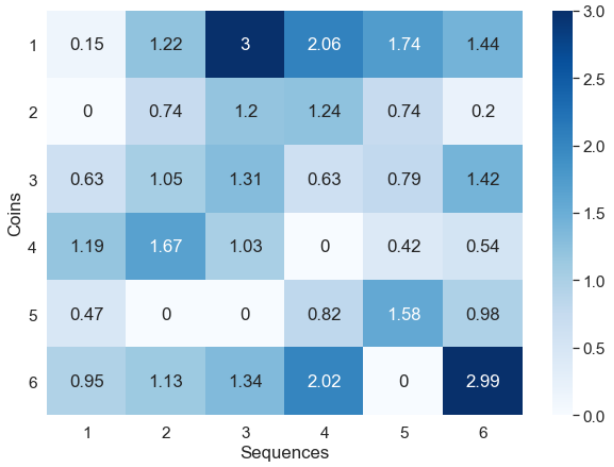


Figure 1: T-test values for each coin-sequence pair

Ultimately, using a t-test for each coin-sequence rating, we found that 4 of the ratings had a statistically significant difference at a 5% significance level. As we can see in Figure 1, the 4 (coin, sequence) pairs that saw a significant difference were $[(1,3),(1,4),(6,4),(6,6)]$.

These four outliers were relatively well-distributed in the pseudo-random order, taking positions 28, 3, 6, and 22, respectively. Therefore, we conclude that "question fatigue", or the phenomenon of a declining quality of ratings as the participant progresses through the experiment, was not the reason for this disparity.

Further analysis into the differences revealed that participants in Group 2 (no context) consistently gave ratings higher than those in Group 1. This suggests that, without the presence of other hypotheses which may appear more favorable or likely, participants give higher ratings than they would if they were made aware of these other, more likely coins.

Interestingly, the isolation of a coin-sequence pair affected the two extremes of the hypotheses: a fair coin (the simplest hypothesis) and a coin that always flips "heads" after two consecutive "tails" (the most complex hypothesis). Since we used a fair coin hypothesis for the control questions and found no significant differences there, there is strong evidence that participants are heavily influenced by context for fair coin hypotheses. Meanwhile, for the complex hypothesis (coin 6), the (6,6) pair ratings were significantly different even at the 1% level.

However, despite the difference in ratings, we still found that the best fitting model for both groups was the Bayesian model, which we can perhaps attribute to participants eventually learning more about the hypothesis space as they answer more questions. Both groups had similar r-values for

the models, and the Bayesian model remains the best fit, as seen in Table 2.

Table 2: Model R-Values for Groups 1 and 2

Model	Group 1	Group 2	Both Groups
Likelihood	0.55	0.58	0.56
Bayesian	0.73	0.76	0.74
Similarity	0.61	0.58	0.59
Best Model	Bayesian	Bayesian	Bayesian

Table 2 also shows us that the similarity model was a better fit for Group 1 than for Group 2. A possible explanation for this could be that when context is removed, as it is in Group 2, the features we defined that are irrelevant to the current hypothesis (e.g. "number of heads after two tails" is irrelevant for a fair coin hypothesis) are not considered and therefore the model's predictions are skewed.

Overall, we observe statistically significant impacts as a result of context from other hypotheses and we note that this impact was observed across the complexity spectrum, but did not affect the r-values of each of the three models. This would suggest that removing context affects the magnitude and scale of the ratings given but does not change the underlying rational basis for representativeness.

Complexity Experiment In this experiment, we sought to answer whether or not the three previously introduced models would perform worse as the coin hypotheses grew increasingly complex. We had three complexity groups, with 2 hypotheses each, across 2 groups of participants, and the resulting R-values of the best fitting model are shown in Table 3.

Table 3: Best Model R-Values for Complexities and Groups

Complexity	Group 1	Group 2	Both Groups
1	0.55 Similarity	0.64 Similarity	0.59 Similarity
2	0.83 Likelihood	0.84 Likelihood	0.82 Likelihood
3	0.79 Bayesian	0.83 Bayesian	0.80 Bayesian

We can observe several interesting results from Table 3. First, for all complexity levels, the model that performs best is the same across Group 1, Group 2, and both groups combined. This is consistent with our findings from the hypothesis context experiment, in which the best model for the data was unaffected by the group number. Therefore, this is further proof that removing other hypotheses from the context affects ratings in terms of magnitude but not in terms of overall methodology.

Second, we can see that the model that best fits the data is different for each complexity level. For the least complex coins (i.e. a fair coin and a coin that flips heads with 80% probability), the best fitting model is the similarity model. Within the model, we can see that the most significant feature was the number of heads, which had a weight that was 4 times larger than the rest of the features for all three complexities. This suggests that, for simple hypotheses that only vary in the probability of flipping heads for each independent flip, the best model for representativeness is a slightly more advanced version of "count the number of heads".

However, when the complexity increases to 2, the best fitting model changes to the likelihood model. This is likely the complexity level at which the participants began to use more complicated probability in their judgments, and we can see that this complexity level has the highest r-values. This suggests that humans are probably most aligned with probabilistic models in moderately complex situations: situations that require more sophisticated probability theory than just counting heads, but not complicated enough that it extends beyond human intuition.

Finally, when the complexity level is equal to 3, the Bayesian model performs best and this suggests that as hypotheses grow more complex, the participants began to consider the other hypotheses when deciding a rating for the current hypothesis. This perhaps simplifies the task since we can contextualize a complicated hypothesis against a simpler hypothesis and use that context to determine the ratings.

These three different complexity levels demonstrate to us that while the Bayesian model provides the best fit for all hypotheses combined, separating apart the hypotheses exposes nuances in model performance that in reliant on model complexity.

Music Genre Representation

In the second set of experiments, we investigate applying the representativeness models to the domain of music genres in a way similar to the mammal analysis in Tenenbaum & Griffiths (2001). In particular, we gathered data separately on both pop and rap songs. For each genre, participants were given sets of 3 songs and asked to rate the representativeness of the genre as a whole on a scale from 1 (least representative) to 7 (most representative). We will explore the effectiveness of our likelihood, Bayesian, sum-similarity, and max-similarity models in aligning with human representativeness data for each of the genres.

Setup

Music Space For each genre, we chose seven songs roughly distributed with two songs before 2000, two between 2000 and 2015, and three since 2015 This was done in order to spread across the genres over time while making sure the songs were recognizable. All songs for each genre are shown in Tables 4 and 5.

Table 4: Pop Songs

Pop	Year
'Dancing Queen' - ABBA	1976
'I Want It That Way' - Backstreet Boys	1999
'Single Ladies (Put a Ring on It)' - Beyoncé	2008
'Just the Way You Are' - Bruno Mars	2010
'Congratulations' - Post Malone	2016
'bad guy' - Billie Eilish	2019
'Truth Hurts' - Lizzo	2019

Table 5: Rap Songs

Rap	Year
'Baby Got Back' - Sir Mix-A-Lot	1992
'N.Y. State of Mind' - Nas	1994
'Drop It Like It's Hot' - Snoop Dogg	2004
'Lose Yourself' - Eminem	2014
'Lucid Dreams' - Juice WRLD	2018
'WAP (feat. Megan Thee Stallion)' - Cardi B	2020
'oops!' - Yung Gravy	2020

Music Sets As in Tenenbaum & Griffiths (2001), we provide participants with sets of three songs each. We run the experiment on pop and rap each without mixing the songs. In order to avoid imposing our own bias on which sets may be important, we ask participants to rate representativeness of all 35 possible sets for each genre.

Feature Vectors All of our models necessitate some way of computing similarity between songs in order to computationally calculate the representativeness of the sets. As such, we looked into various ways to calculate pairwise song similarity.

Various online resources such as Last.fm provide similarity ratings between songs, but we find that they are quite volatile and limited only to the top 100 most similar songs at any given point due to being based on real-time user-listening-data. These restrictions mean we would not be able to survey participants on the complete set and would have a difficult time in particular finding the similarity between less similar songs.

Thus, we create feature vectors for each song using the Spotify API through its ID. Specifically, we use a large number of each track's audio features as well as characteristics from its track and album. The full list of features used includes 'danceability', 'energy', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', 'duration_ms', 'popularity', and 'release_date'. Every feature is normalized to be a rating between 0 and 1 inclusive based on the minimum and maximum of the range for all songs in both genres.

In order to test subsets of the features, we run trials in which we remove a random subset of the features from the vector and evaluate the resulting models' r values when comparing with human data on a variety of seeds. The features that perform the best overall for each genre are used in the next section.

Multidimensional Scaling We apply 2D multidimensional scaling to our feature vectors from the previous section, similarly to Tenenbaum & Griffiths (2001), for use in our models. Given that the sklearn algorithm is not deterministic, we run 500 trials with a randomized seed to see the best output per model which will be discussed in later sections. The MDS output affects both the likelihood and Bayesian model success.

Likelihood Model In order to calculate the likelihood of each song, we create a multivariate normal distribution from all songs in the genre. This uses the mean from the MDS coordinates as well as the covariance matrix. Total likelihood of a set is simply calculated by plugging in individual song coordinates to the respective distribution and multiplying the three probabilities together.

Bayesian Model As was done for animals, we assume people's hypothesis space includes the category of all pop/rap songs and infinite alternate hypotheses. We thus use Equation 7 from Tenenbaum & Griffiths (2001) to model all hypotheses as Gaussian distributions in a two-dimensional feature space from our MDS analysis of our Spotify feature vectors. As in the original paper, we also plot the one-standard-deviation contours for the full song space as well as three samples that differ in representativeness.

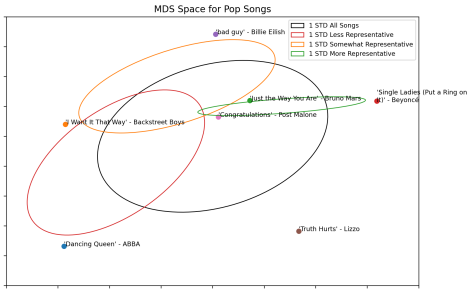
Sum-Similarity and Max-Similarity Models We explore two similarity models as ways to convert our pairwise similarity metrics to measures of setwise similarity. Sum-similarity takes the sum of the similarity of each song in the set with every other song in the song space. Max-similarity takes the maximal similarity between each song in the song space and every song in the set and sums them together.

Results

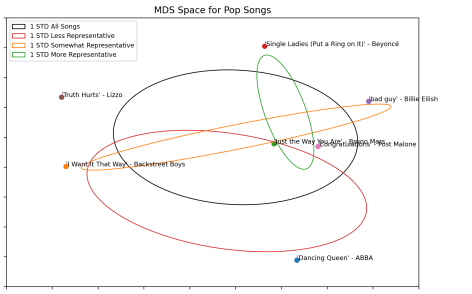
Pop Music

Feature Vectors In testing the random feature subsets for pop music, we find that models perform the best when all features are used: 'danceability', 'energy', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', 'duration_ms', 'popularity', and 'release_date'. This suggests that many factors may go into humans' concepts of representativeness with regards to this genre. In particular, some of the features may have only subtle differences such as 'danceability' and 'energy' that is important in a genre such as pop music where the songs might not have stark differences from each other.

MDS Space Our MDS algorithm includes randomness, meaning that different MDS outputs can result in the models



(a) Optimizing Bayesian model



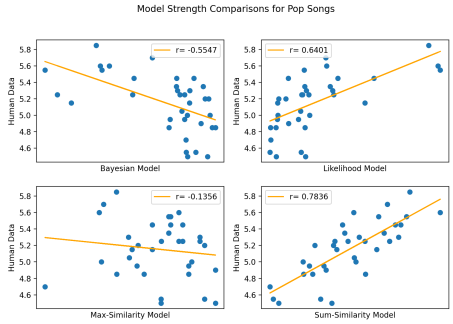
(b) Optimizing likelihood model

Figure 2: Representativeness for Pop Songs. The black ellipses show the 1 standard deviation contours for all pop songs. The colored ellipses show a representative sample (green: 'Single Ladies (Put a Ring on It)' - Beyoncé, 'Congratulations' - Post Malone, 'Just the Way You Are' - Bruno Mars), a somewhat representative sample (orange: 'bad guy' - Billie Eilish, 'Just the Way You Are' - Bruno Mars, 'I Want It That Way' - Backstreet Boys), and a less representative sample (red: 'Dancing Queen' - ABBA, 'I Want It That Way' - Backstreet Boys, 'Congratulations' - Post Malone) as rated by participants.

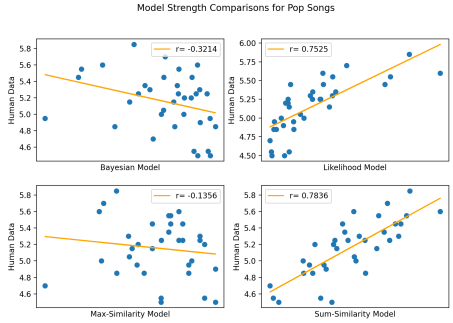
being more and less effective at predicting representativeness. Since it affects both the Bayesian and likelihood model, we provide and analyze two MDS spaces in Figures 2a and 2b. The first of these is optimized for the Bayesian model and the other the likelihood model.

There are many similarities between the spaces when ignoring rotation, such as the songs around our green ellipse. However, 'bad guy' and 'I Want It That Way' appear in different places. We see the differences these MDS outputs have on our r-values in the next section. Interestingly, these spaces do not seem to follow the pattern that the more representative set has more overlap with the contour representing the entire set. This could provide some intuition as to the inefficacy of the Bayesian model as compared to its application on animals.

Model Strengths We look at the human representativeness predictions versus the model predictions when using the



(a) Optimizing Bayesian model



(b) Optimizing likelihood model

Figure 3: Scatter plots comparing representative judgements for 35 sets of pop songs with four model predictions

MDS coordinates from Figure 2. This is seen in Figure 3.

In optimizing for the Bayesian model, we see an r value at an average performance of -0.55. In optimizing for the likelihood model, we see an r value perform decently at 0.75. In either case, the max-similarity r value is quite low at -0.14 and the sum-similarity r value is the highest at 0.78.

Interestingly, this seems to differ from the initial application to the animals domain in a few ways. We see a negative correlation for both the Bayesian and max-similarity models with a positive correlation to the likelihood and sum-similarity models. The Bayesian model performs noticeably worse in this domain than in previous domains which could be explained by the lack of correspondence in the MDS space to representativeness from a human perspective. In fact we do see that the more representative samples are smaller ellipses that could lead to a negative correlation. This suggests that humans, particularly in pop music, might sometimes place importance on songs that aren't necessarily representative or similar because they are standout songs in the genre. The likelihood model does well but also shows a higher concentration around lower likelihoods - in particular it does not distinguish between the lower end of our range of representativeness. The sum-similarity model appears to be successful on the data. This suggests that all songs in the set are vital to representativeness, but perhaps the MDS space is not accurate in predicting the additional subjectivity in this domain

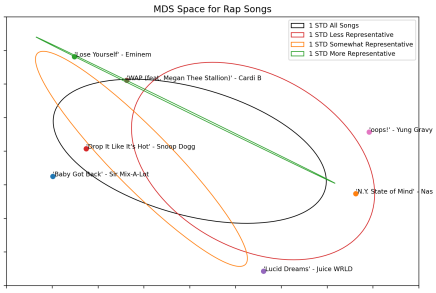


Figure 4: Representativeness for Rap Songs. The black ellipses show the 1 standard deviation contours for all pop songs. The colored ellipses show a representative sample (green: 'N.Y. State of Mind' - Nas, 'WAP (feat. Megan Thee Stallion)' - Cardi B, 'Lose Yourself' - Eminem), a somewhat representative sample (orange: 'Lose Yourself' - Eminem, 'Drop It Like It's Hot' - Snoop Dogg, 'Lucid Dreams' - Juice WRLD), and a less representative sample (red: 'WAP (feat. Megan Thee Stallion)' - Cardi B, 'oops!' - Yung Gravy, 'Lucid Dreams' - Juice WRLD) as rated by participants.

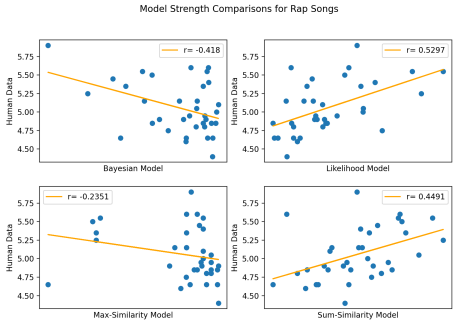


Figure 5: Scatter plots comparing representative judgements for 35 sets of rap songs with four model predictions

due to losing nuances in the similarity features.

Rap Music

Feature Vectors For rap music, we find the models are most successful with a subset of features used. These are 'acousticness', 'instrumentalness', 'key', 'liveness', 'loudness', 'mode', 'valence', 'popularity', and 'release_date'. This excludes 'danceability', 'energy', 'tempo', 'speechiness', and 'duration_ms. The first three of these excluded features seem that they would be more suited toward the pop category when you need a fine grained way to distinguish songs, and the last two features do not particularly seem like features that humans would internalize when thinking of overall song to genre similarity. It should be noted that 'speechiness' is used in Spotify to separate podcasts and music more so than songs. In the features we keep, we do not need quite as many minor details as before, indicating that perhaps rap sees a lesser variety of sound across its genre as compared to pop. This is additionally supported by a smaller range of representativeness ratings in our human data. It should also be noted that this chosen set of features is optimal for the likelihood model but not the similarity models, which does better once more with all features as in pop. We analyze this set as the best likelihood r value is higher than the best similarity r value with their respective optimal features chosen.

MDS Space Although our MDS algorithm still includes randomness, the most successful seed for the Bayesian and likelihood model on rap music happened to be the same. Thus, we analyze only one MDS space in this section seen in Figure 4.

Once more, much more severely with our representative set, we see a smaller area covered with the more representative contours than the less representative contours.

Model Strengths As before, we look at the four models in Figure 5 to analyze their success against the human representativeness data.

We see an r value at -0.4180 for the Bayesian model, 0.5297 for the likelihood model, -0.2351 for the max-similarity model, and 0.4491 for the sum-similarity model. Noticeably, this is less successful for all models except for max-similarity (which only has a loose negative correlation) as compared to the pop music experiments.

In rap music, we see an even smaller range of representativeness which could lead to poorer results. The likelihood and sum-similarity models still both are comparable in success and have more similar distributions than in the previous genre. The other two models, however, show a lot of bunching toward the higher values which suggests the overall Gaussian space could be predicting too high of values or the individual sample Gaussians could be calculating too low of values based on the MDS coordinates.

Again, it appears that there is a level of subjectivity in the representativeness space that is not captured either by the initial feature vectors or the conversion to the MDS space. This could be introduced by human biases or general cultural intuition regarding songs, whereas our features capture strictly objective data. The models find some correlation but are clearly missing a piece to predict representativeness more accurately.

Overall Music Discussion

In general, a disconnect between Spotify features and humans' concept of similarity in addition to the close range of representativeness both likely contribute to the lower success of the models. The data shows that people did not differentiate largely in representativeness between many of the sets, and there often was not strong agreement in sets' representative values. This implies that a model would also have a difficult task replicating the predictions, especially without a crowd-sourced similarity feature. Having something that draws upon listening usage might be able to contain underlying distributions of societal biases and incorporate qualitative

similarities into the models that would do a better job with understanding representativeness.

Conclusion

In this paper, we have shown several results. First, we found strong evidence that hypothesis context can affect participant ratings but ultimately does not affect which model performs best. Second, the fit of a model depends on the complexity of the hypothesis and our data suggests that the rational basis for representativeness changes as these complexities change.

Within the domain of music, we have also found that objective features have some success on predicting song set representativeness but likely require data including subjective similarity information to be more successful as has been seen in other domains. In particular, the lack of cultural and societal bias in Spotify features seems to leave a hole in the MDS representation and results in an inability to distinguish between some song sets. Additionally, this issue is more prominent in the rap than pop genre. We also see that the likelihood and sum-similarity models seem to work best in the highly subjective domains which could be influenced by the missing crowd-sourced features.

Acknowledgements

We would like to thank our project TA, Sihan Chen, as well as Professor Tenenbaum for their continued support and advice throughout the project and the semester.

Contribution Statement

We designed the experiments together, which included selecting the hypotheses, sequences, and song sets. We also consulted with each other for the model designs, across both the coins and music domains. After creating the Google Form and collecting data together, we divided the remaining work between the two of us. Crystal wrote the code and ran the analysis for the coins experiments, while Julia did the same for the music representation. Although we split up for the respective experiments, we still consulted with each other throughout on the data and interpreting the results. Finally, we each wrote about our respective experiments in the paper and Crystal handled the introductory sections.

All code and data used in this paper can be found at https://github.com/cwcrystal8/9.660-final-project

References

Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*, 684–704.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cog. Psych.*, *3*, 430–3.

Tenenbaum, J. B., & Griffiths, T. (2001). The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 1036–1041).