NETS 213 Final Project Report

Basic Project Information

Name of your project

Noteworthy

Name of your teammates:

Ben Geist, Elias Kalish, Chunxi Liu, Anjali Maheshwari, Halle Wasser

Give a one sentence description of your project. Please use the name of the project in your description

Noteworthy uses machine learning algorithms and crowdsourcing to generate new music - no musicians necessary!

Logo for your project.

What problem does it solve?

The problem we are trying to solve is whether using crowdsourced ratings improves music generated by a RNN machine learning model.

What similar projects exist?

There are a number of pre-existing models and services that automatically generate music in a variety of different ways. To our knowledge, there has been no research done on whether using crowdsourced ratings leads to better music performance.

What type of project is it?

Human Computation Algorithm

What was the main focus of your team's effort

Conducting an in-depth analysis of data

How does your project work? Describe each of the steps involved in your project. What parts are done by the crowd, and what parts will be done automatically.

For our project, we will be generating songs that are tuned on a variety of different parameters and performing analysis on the results. Our project basically boils down to two main steps that are repeated many times and in different ways. We created unique songs piece by piece, where each successive piece of the song was selected by the crowd. First, we found a pretrained RNN model that could generate continuations of music based on primers we fed into it. This gave us some control over what the resulting generated continuations sounded like. The length of each primer was only 8 seconds in length because they are meant to be starting blocks from which the rest of the song is ultimately generated. We chose three different primers: a jazz primer, a classical primer and an ancient primer. The jazz primer was a simple jazz piano riff online. The classical primer we chose was the opening seconds of the iconic Dance of the Sugar Plum Fairy, by Tchaikovsky. The ancient primer is pulled from a piece called Seikilos Epitaph, the oldest full musical composition in human history, and gave surprisingly good results.

Once we had our primers picked, we ran the following process on each of them in parallel:

First, we plugged the primer into the model and generated fifteen possible continuations. We then posted the primer with each possible continuation and had multiple workers give each one a rating. We then aggregated those ratings and selected the best one, tie-breaking if necessary. We then took the primer plus continuation selected and fed that back into the model as a *new* primer, generating fifteen more continuations from that and repeating the process. We did this four times, with five second continuations, to generate pieces that were roughly 30 seconds in length.

Our final round of HITs inserted our iteratively generated piece in with a group of full length pieces generated without iteration. The goal here was to see if our iteration and crowdsourcing of ratings lead to noticeable improvements in music quality.

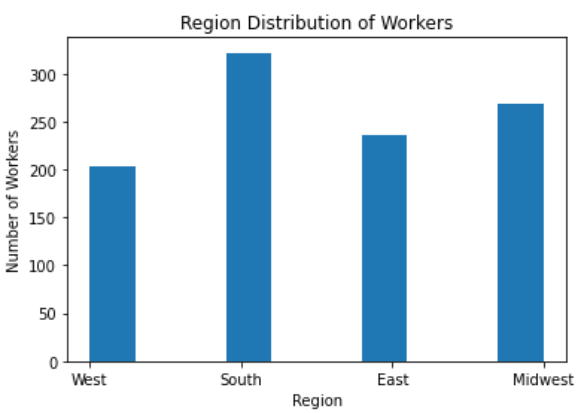
Provide a link to your final presentation video. Give the full path to your Vimeo video, and the password, if it is not public.

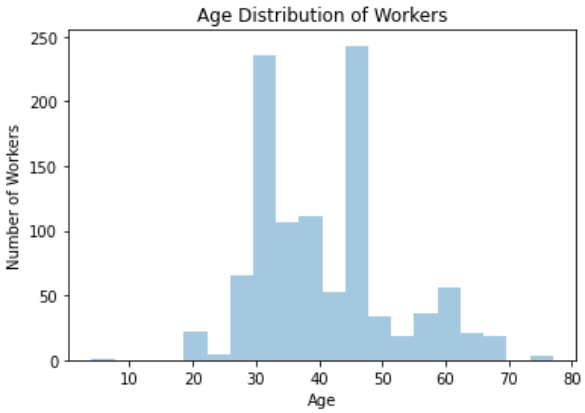
https://vimeo.com/547698214

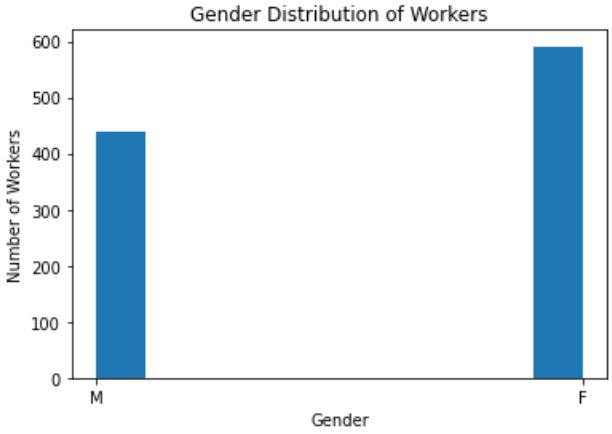
The Crowd

Who are the members of your crowd?

Our crowd is made of Amazon Mechanical Turk workers in the United States. We asked the workers their age, gender and location (state) and found the following results.







The workers are slightly more prominent in the South than in any other region and are primarily between the ages of 20 and 50 although we did have a small number of workers between 60 and 70. Furthermore, the workers are fairly evenly distributed between male and female, but are slightly more female.

For your final project, did you simulate the crowd or run a real experiment?

We ran a real experiment on Amazon Mechanical Turk.

If the crowd was real, how did you recruit participants?

We recruited participants by paying them on Amazon Mechanical Turk.

How many unique participants did you have?

152

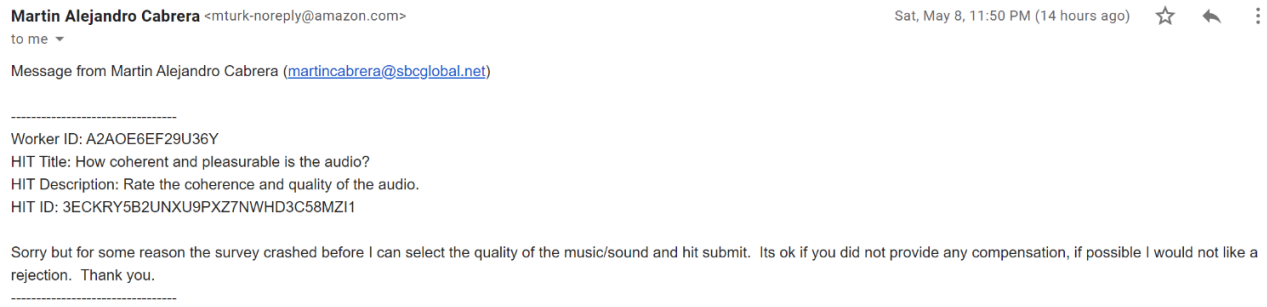
Incentives

What motivation does the crowd have for participating in your project?

The crowd's main motivations to complete our project are money, enjoyment, and reputation. The obvious incentive for our project is money because we ran our experiment on Amazon Mechanical Turk.

Another incentive we provided was enjoyment. We wanted to leverage this as much as we could as listening to music is typically seen as a fun activity. To increase the perceived enjoyment of this activity we made the task titles appealing and attached enjoyable keywords.

The final incentive we provided was reputation. Originally, we did not think that reputation was a large incentive for our HITs as we were not well known in the Turker or academic community. However, we realized reputation was an important incentive when one worker sent this email:



Martin Alejandro Cabrera <mturk-noreply@amazon.com> Sat, May 8, 11:50 PM (14 hours ago)

to me

Message from Martin Alejandro Cabrera (martincabrera@sbcglobal.net)

Worker ID: A2AOE6EF29U36Y
HIT Title: How coherent and pleasurable is the audio?
HIT Description: Rate the coherence and quality of the audio.
HIT ID: 3ECKRY5B2UNXU9PXZ7NWHD3C58MZI1

Sorry but for some reason the survey crashed before I can select the quality of the music/sound and hit submit. Its ok if you did not provide any compensation, if possible I would not like a rejection. Thank you.

Obviously the workers were very concerned with their reputation and did not want to get a bad reputation via a rejection.

How do you incentivize the crowd to participate? Please write 1-3 paragraphs giving the specifics of how you incentivize the crowd. If your crowd is simulated, then what would you need to do to incentivize a real crowd?

The original name for our HIT was “Audio Naturalness.” Quickly we found that this HIT was not garnering much attention as it had low enjoyment incentive levels. We changed the name to a more interesting name and attached more positive keywords to the HIT and saw a sharp increase in popularity. An example of the name format we decided on is “AI Music Rating-classical music” which we used for the classical music rating task. The keywords we ended up attaching were: audio, naturalness, rating, money, quick, easy, music, enjoyable, machine learning, fun, and comparison. Both of these changes helped increase the popularity of our HITs.

Another change we made to incentivize the crowd to participate was increasing the monetary payment. Originally our payment was only 1 cent, but we realized that workers were not willing to participate at this price. This resulted in our HIT getting not much attention. We increased the payment to 2 cents and immediately saw an increase in our HITs’ popularity.

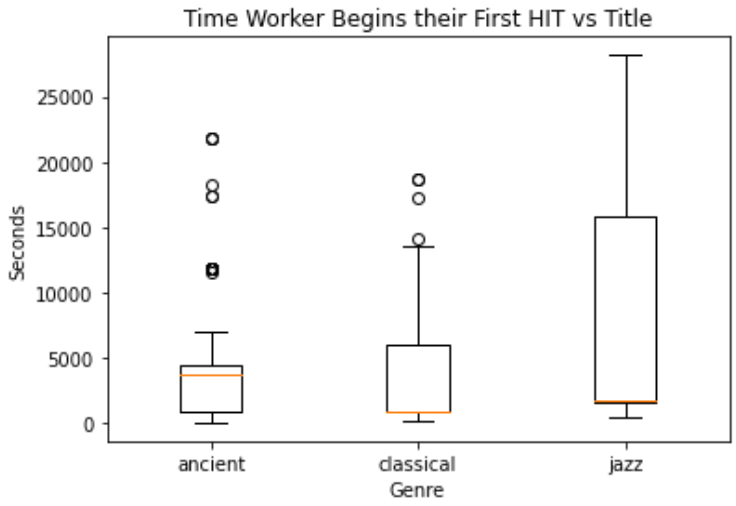
Did you perform any analysis comparing different incentives?

To measure how much of an incentive money was, we raised the payment of the fourth iteration of the ancient music to 4 cents per HIT and the payment of the fourth iteration of the classical music to 3 cents per HIT. We compared the times it took workers to do their first task in the 4 cent assignment to the times it took the workers to do the 3 cent assignments. All of these HITs were posted on Amazon Mechanical Turk around the same time and had the same name format and task format and so the only change made was the monetary incentive.

To measure the change increasing enjoyment had on HITs popularity we changed the name of the third iteration of jazz music to “Super Fun AI Music Rating- Help Create the Future of Jazz!”, the third iteration of classical music to “AI Music Rating- Classical music”, and the third iteration of ancient music to “How coherent and pleasurable is the audio?” We compared the distribution of the time it took workers to start the increased enjoyable jazz HIT, the classical music HIT as well as the ancient HIT. Because these HITs were posted at the same time, had the same format, and the same payment the only difference is the perceived enjoyableness of the HIT.

If you compared different incentives, what analysis did you perform? If you have a graph analyzing incentives, include the graph here.

To analyze the results, we created a box-and-whisker plot comparing the response time of the three types of HITs. We found that a more enjoyable HIT improves the workers response up until a certain point, which then causes workers to become less responsive. Furthermore, we found that workers were most responsive and eager to try the classical music HIT, then the ancient music or jazz music HIT, respectively.



Between the three HITs only jazz did not have even half of its tasks completed (we waited well after both the ancient music and classical music HITs had finished). This indicates that workers want to know that their work will be enjoyable, but are more motivated by the seriousness of the work—they are treating these tasks as work and when a task presents as unprofessional workers seem to avoid it such as in the jazz HIT.

On the other hand, if a task seems slightly enjoyable, yet also professional, workers will be more inclined to complete it. The classical music HIT had a much lower average time it took for a worker to click on the HIT. Furthermore, although the classical music HIT has a larger spread than the ancient music HIT, the classical music HIT has few outliers and was completed quicker. The classical music HIT let the worker know they would listen to and rate classical music, while the ancient music HIT just let the worker know they would rate audio, and so the classical music HIT appeared more enjoyable to the worker leading to its increased completion rate and its lower average time until a worker started the HIT.

To measure how much of an incentive money was, we compared the boxplot of the first time a worker clicked on the HIT between the 4 cent HIT (ancient) and the 3 cent HIT (classical).

We clearly found that the ancient music HIT, which paid more, had a similar mean to the classical HIT, but had a much lower spread. This resulted in the higher paid HIT completing much quicker than the lower paid one. Furthermore, the higher compensation in the ancient music HIT enticed workers to complete the last few tasks which typically take the longest to complete as workers know there isn't an abundance of tasks to do in one go. Because of this and the lowered spread we clearly see that increased compensation causes incentives for the workers to do the HIT.

What the crowd gives you

What does the crowd provide for you?

The crowd provides answers to our HIT questions. They listen to either one or two pieces of machine generated music and then either rate or rank, respectively. This provides the data used to generate the following continuations that ultimately create the final generated musical pieces.

Is this something that could be automated?

No.

If it could be automated, say how. If it is difficult or impossible to automate, say why.

The main point of our project is to see if human input on music quality and coherence can help improve the quality of song outcomes using machine generated music. While the music generation is automated, the human quality of the ratings and rankings is vital to our project.

Did you train a machine learning component from what the crowd gave you?

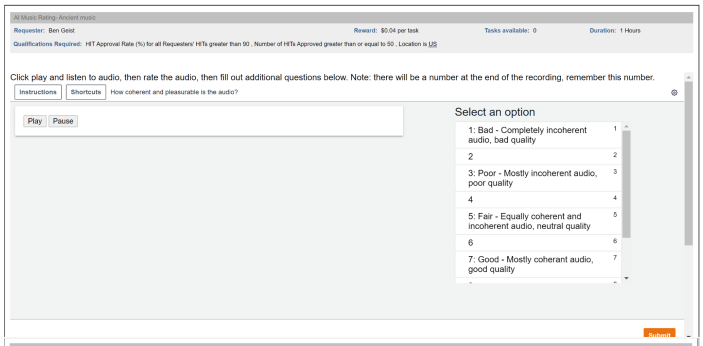
No.

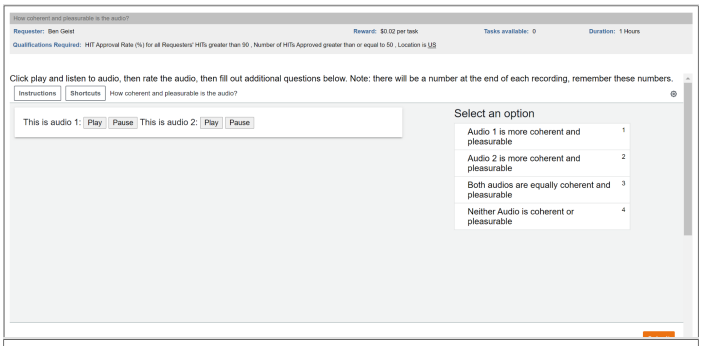
Did you create a user interface for the crowd workers? Answer yes even if it's something simple like a HTML form on CrowdFlower.

Yes.

If yes, please include a screenshot of the crowd-facing user interface in your report. You can include multiple screenshots if you want.

Rating HIT **Comparison HIT**





Describe your crowd-facing user interface. This can be a short caption for the screenshot. Alternately, if you put a lot of effort into the interface design, you can give a longer explanation of what you did.

We used two similar HITs - one which had our workers listen to a single audio clip and then give it a rating, and one in which a worker listened to two audio clips and ranked them against one another.

Both clips had a section in the upper left hand corner containing a button to play/pause the audio. We removed the common listening bar showing progress in an audio clip so that they could not use it to skip ahead and would instead be forced to listen to the full audio clip. In the upper right hand corner are the various options for rating and ranking the audio clips. Below this section are a number of text input questions for answering the quality control question(s) and providing some basic information about the worker to be used later for analysis.

Skills

Do your crowd workers need specialized skills?

No, however we did require the workers have at least 50 approved HITs and a 90% approval rating. This ensured that workers were at least somewhat skilled at MTurk.

What sort of skills do they need?

They need to be able to listen to music and have a basic understanding of what constitutes good music.

Do the skills of individual workers vary widely?

No, we relied on the basic human experience of judging music and so the skill is fairly uniform in the general population. Furthermore, we only looked at workers in the United States and so all the workers had a similar understanding of music as they are all a part of a like American culture.

If skills vary widely, what factors cause one person to be better than another?

The average MTurk worker should be adequate at completing this task. Individuals who are hard of hearing or deaf may not be able to complete this task. On the other hand, musicians or music lovers who understand key changes and time signatures in music will be better at deciding if the generated music is considered “good” or “bad.”

Did you analyze the skills of the crowd?

No, but we added questions at the end of the HIT that asked for the worker’s age, gender, and state that they were from. All workers are from the United States.

If you analyzed skills, what analysis did you perform? How did you analyze their skills? What questions did you investigate? Did you look at the quality of their results? Did you analyze the time it took individuals to complete the task? What conclusions did you reach?

We did not analyze the skills of the workers.

Quality Control

Is the quality of what the crowd gives you a concern?

As we utilized Amazon’s Mechanical Turk for crowdsourcing, the quality of the responses was a major concern because the platform does not require stringent qualifications for accepted workers. As each stage of the project and selected music continuations relied entirely on the worker’s responses, quality control was necessary to ensure the output was of the greatest compounded quality. However as the tasks required subjective responses, the quality control could not simply be done through standard statistical analysis and algorithms in comparison to control questions or other workers responses.

How do you ensure the quality of the crowd provides?

In short, we generated quality control measures to ensure that the worker does not randomly select a rating or ranking and that the worker cannot skip through the audio recording. In the next question, we describe these quality control measures in detail.

If quality is a concern, then what did you do for quality control? If it is not a concern, then what about the design of your system obviates the need for explicit QC?

The first part of the quality control for this project relied on the reputation system of Mechanical Turk. In order to qualify to complete HITs for this project the worker must have a HIT approval rate (%) for all requesters HITs greater than 90, have their number of approved HITs greater than 50, and reside within the United States. While this alone does not ensure that the worker produces quality responses, it does start to establish the level of their abilities. Furthermore, a step-by step instructional video is provided to each worker completing the HIT, as well as detailed and simple written instructions to ensure the worker correctly understands what is expected from them and how to successfully complete the HIT, which also helps to eliminate workers that do not speak English and careless workers.

Our tasks asked workers to rate or rank the quality of the generated portions of music on their qualities of cohesiveness with the primer and based on the level of enjoyment the piece had for the worker. As this is almost entirely a subjective measurement, quality control by comparing answers by workers to control responses was not possible, so instead two quality control features were implemented into the HIT design itself. First, at the end of every musical piece, a randomly generated two-digit number was spoken aloud that the worker must correctly report or else their HIT would not be accepted and their response would not be used. This ensures that the worker must listen to the entirety of the sound and be paying attention to the music in order to complete the task, so that the worker cannot just quickly and randomly select a rating. Second, the audio was embedded in a manner that does not allow the worker to drag the slider to the end of the piece to hear the automated number without actually listening to the musical piece.

Did you analyze the quality of what you got back? For instance, did you compare the quality of results against a gold standard? Did you compare different QC strategies? What analysis did you perform on quality?

In our analysis, we found relatively high levels of quality for each iteration based on the correct answer being recorded for the control response. The percentage of HITs that were not accepted due to the worker incorrectly inputting the control number was between 1.33% and 13.3%, with the final iteration having the lowest level of quality. As the numbers concatenated to the end of each audio were randomly generated by our quality control module, it was relatively simple to recognize and remove HITs that did not successfully follow the instructions, but as the rest of the requested answers were subjective or demographic information, using gold standard questions for them was not feasible. Instead, any response that could not successfully answer the control question was immediately removed from the mean calculation and the HITs that did respond correctly were accepted as the user's opinion.

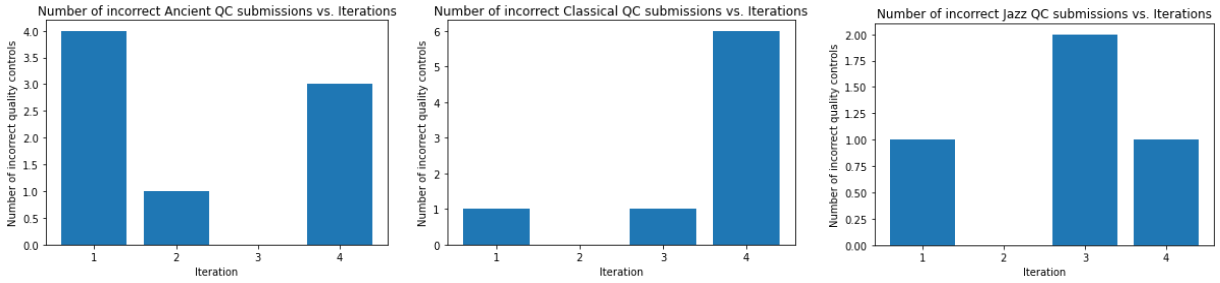
What questions did you investigate? What conclusions did you reach?

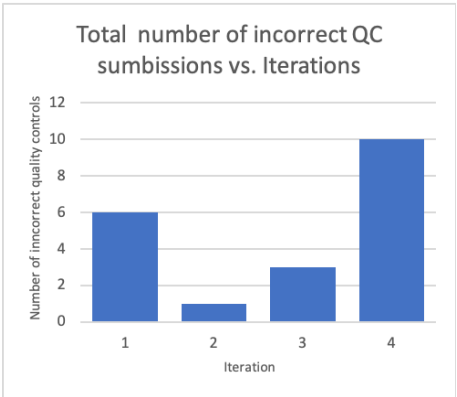
In our analysis of quality control, we investigated two questions to assess the best strategies for scaling up the project in the future:

- Will the quality of work produced for each HIT decrease as the length of the audio, and time it took for the worker to complete the task, increasem?
- Is there a difference in quality between the rating and ranking HIT designs?

Contrary to our prediction, the number of HITs that did not successfully answer the control questions demonstrated upward trends as the length of the audio increased, excluding the first iteration. There was significant variation in the number of incorrect responses in each iteration based on the audio styles of ‘jazz’, ‘ancient’, and ‘classical’, with the highest levels of failure for the first and last iterations when looking at the total sum. This is demonstrated in the graphs below. Additionally, there was no significant difference between the number of incorrect quality control questions for the ranking and rating HITs demonstrating that the level of worker engagement was comparable between each design, however as our results noted, the ranking HIT design was ineffective for our objective, since the probability of one being ranked above the other is always close to 0.5.

Do you have a graph analyzing quality control? If you have a graph analyzing quality control, include the graph here.





Aggregation

How do you aggregate the results from the crowd?

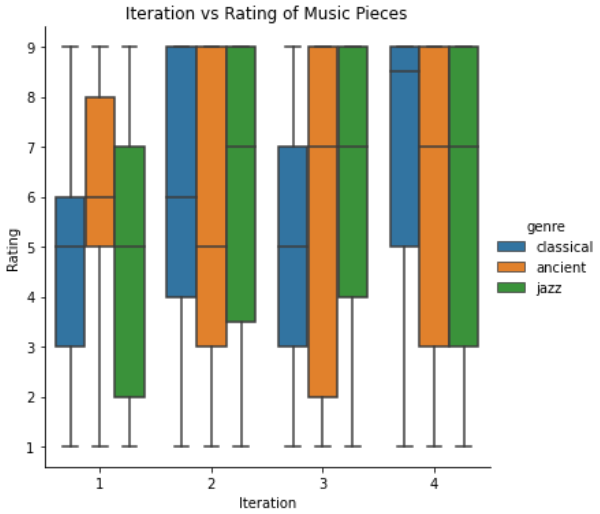
We aggregated our results by finding the mean and standard deviation of each of the 15 musical pieces from the 3 genres at each iteration.

Did you analyze the aggregated results?

We chose the top song by selecting the audio pieces with the highest mean. In the case that two or more pieces had similar means and were less than 1 of the minimum standard deviations (between the pieces) away from each other, we listened to the songs collectively and decided which one sounded best.

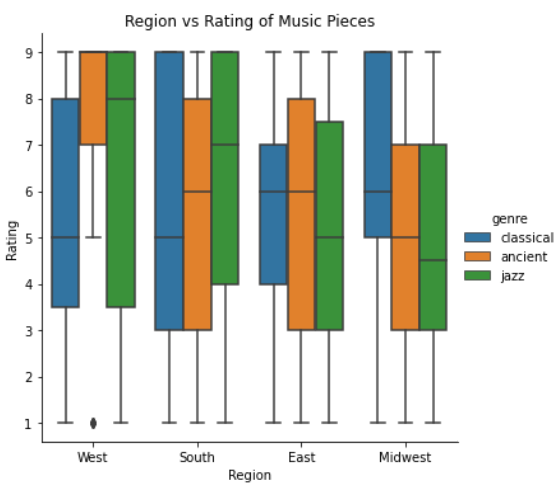
What analysis did you perform on the aggregated results? What questions did you investigate? Did you compare aggregated responses against individual responses? What conclusions did you reach?

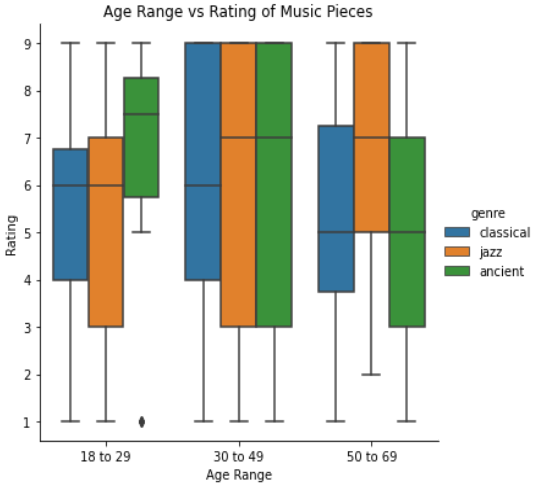
One question we aimed to investigate was how the standard deviations and means of the musical pieces differed between the genres and over iterations. We wanted to see if our iterations actually improved the quality of music as iterations progressed. We also analyzed how different factors about our workers, such as age, location in the US and gender, affected their musical preferences and ratings.

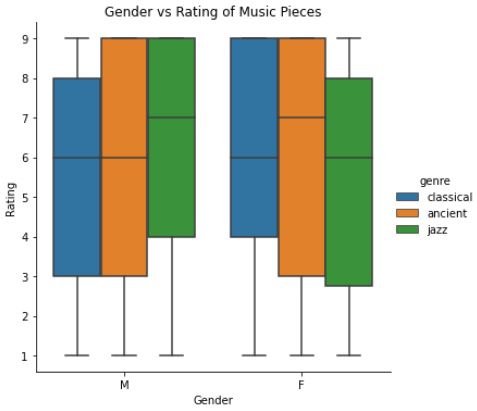


The above box plot maps ratings of the genres across iterations. There is a fairly large amount of rating variance, but this is to be expected due to the wide range of continuations generated at each iteration. Some will be very good, the best of which will be used in the next iteration, but there also was usually a large number of bad continuations. The main takeaway from the graph is that ratings of the music increased as iterations went on. There was a little bit of up and down for some genres during the middle iterations but by the last iteration the average rating for each

genre was higher than the first iteration, proving that this method of crowdsourced ratings does have some merit.



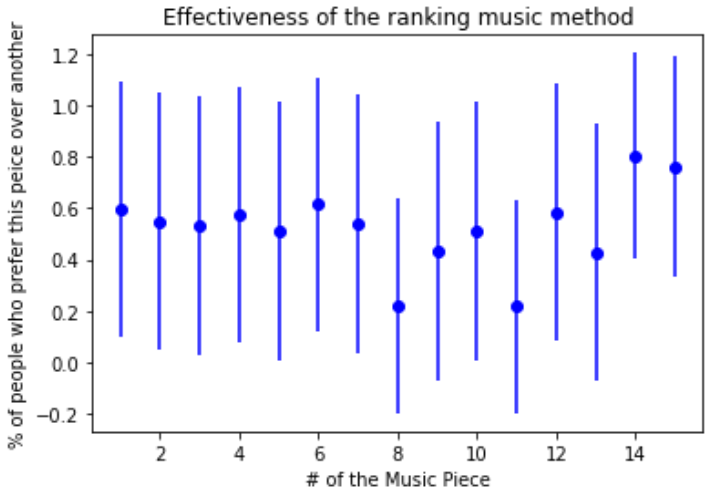




The above graphs show the variance of genre preference across age, gender and location. Younger people tended to prefer the ancient genre over the other too, while the oldest demographic preferred jazz. The middle age group didn't show significant preference for a genre. Men showed a greater preference for jazz while women showed a greater preference for ancient music. The differences across different regions in the US was significant: the West and South showed a much larger preference for jazz while the East and Midwest showed a greater preference for classical music. All these differences show that there are biases in the population relating to age, location and gender. Since our respondents are not uniform, these biases likely are also present in our data.

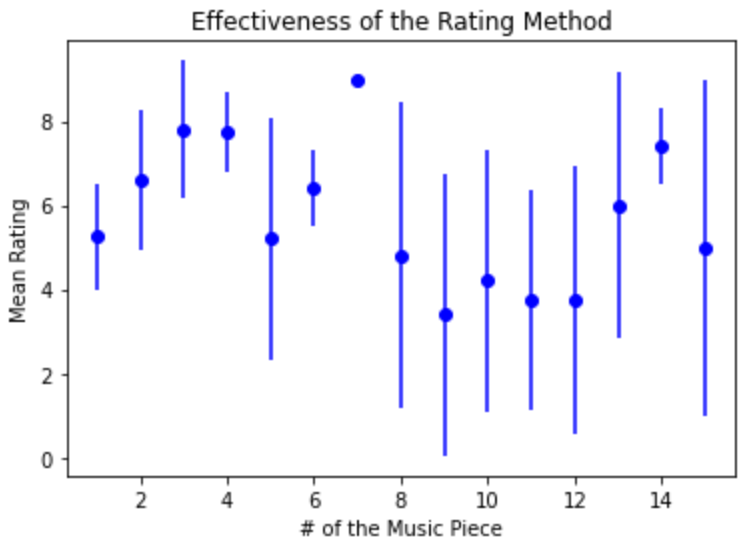
Another question we wanted to investigate was if our method of obtaining and aggregating data was worthwhile. To test if rating musical pieces was an effective method to get information on the quality of a piece we also asked workers to select the higher quality piece between two pieces. They had the option to choose one of 4 choices: "Audio 1 is more coherent and pleasurable," "Audio 2 is more coherent and pleasurable," "Both Audios are equally coherent and pleasurable," or "Neither audio is coherent and pleasurable." We then calculated the mean number of times audio 1 and audio 2 were preferred and the standard deviation of their

preference. For instance, if half the number of times audio x was pitted against another audio and audio X was preferred or both audios were equally pleasurable, then it would have a mean preference of .5 and a standard deviation of $\left(\frac{\Sigma\,(x_i-.5)^2}{n}\right)^{1/2} = \sqrt{\frac{n\,(.5^2)}{n}} = \ .5$. We performed this analysis on the first iteration of the “ancient” genre music.

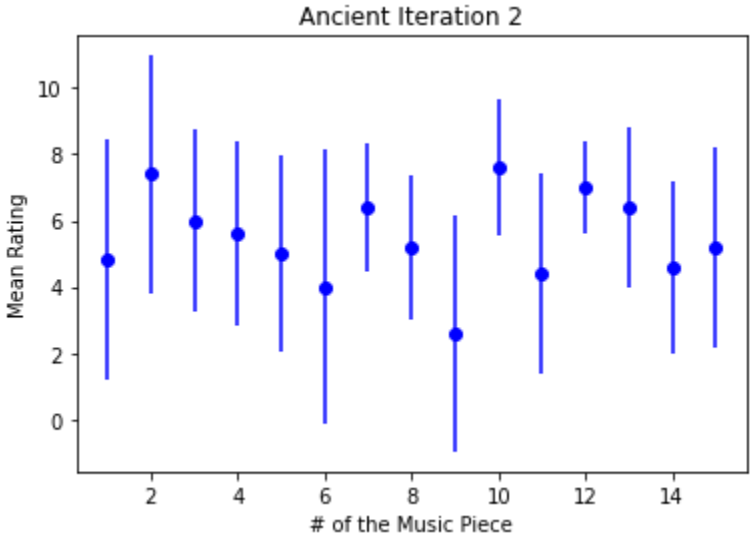


As you can see in the graph below most means are near 0.5 and most of our standard deviations are near 0.5 corresponding to just half of the results preferring the first audio over the second when given the option between two audios. We ran this by comparing every music piece with every other music piece (that wasn't itself). For each pair, we showed 3 workers with piece A first and piece B second and 3 workers with piece B first and piece A second. This showcases how workers were mainly choosing music pieces randomly in this method and so we concluded it was not viable.

In contrast, the rating method paired with our selection process for the best of the pieces worked better. Below you can see the equivalent graph to the one above, showing the means and standard deviation of different musical pieces ratings in iteration 1 of the “ancient” genre:



The means and standard deviations are clearly more varied and thus less often caused by chance. We see that piece 7 is the clear winner in this situation as it has the highest mean and a standard deviation of 0. In some cases the answer is less clear, such as the second iteration of the ancient genre.



Here we have multiple high means and no clear front runner. However, if we sort our results by mean (descending) and standard deviation (ascending), we can find the most consistently, highest rated pieces, demonstrating the worker's confidence in the audio's quality and cohesiveness. From here, we found the best piece by collectively listening and deciding on the answer, crowdsourcing within.

Scaling Up

What is the scale of the problem that you are trying to solve?

The scale of the problem/question is dependent on how large we choose to scale a number of different factors. The main places to scale would be:

1) Length of the song
2) Number of iterations per song
3) Total number of songs
4) Number of different continuations for each iteration
5) Number of people reviewing each option per iteration

Our project was more along the lines of scientific research to see if this approach led to significant improvements in song quality. Therefore, our scale was smaller, using 3 songs of length around 30 seconds, 4 iterations per song, 15 different possible continuations for each iteration and 5 workers looking at each continuation.

Would your project benefit if you could get contributions from thousands of people?

Yes

If it would benefit from a huge crowd, how would it benefit?

As shown above, there are a number of different facets of the project that could be scaled up given a large number of people. Some would benefit greatly from a larger sample size, while others would provide potential issues if scaled up substantially. Options two through five in the list above would all benefit in different ways if scaled up to a significantly larger crowd.

Number of iterations means that number of different continuations we do per song until we get to the desired length. For example, if our primer was 10 seconds, and we wanted a 50 second song, we could do 10 continuations of 4 seconds each, 8 continuations of 5 seconds each, etc. Having a larger crowd would allow us to more efficiently parse through many iterations more quickly. Furthermore, having a large crowd could help eliminate some of the population specific biases we illustrated earlier.

Similarly, a large crowd would mean we could post more songs and get results back faster for each song.

Options 4 and 5 are the main places where a large crowd would be a benefit for our project. One of the major bottlenecks for getting data for our project was tweaking the number of continuations per iteration and the number of people reviewing each continuation, the challenges of which will be discussed in the next section. A larger crowd would enable us to efficiently evaluate more continuations for each iteration; more continuations per iteration means a greater chance of finding that perfect continuation for the primer, and thus would lead to better outcomes. Similarly, a larger crowd would allow us to have more people review each continuation, leading to a higher quality in the ratings of each continuation.

What challenges would scaling to a large crowd introduce?

As mentioned above, the main bottleneck for our project was how time consuming each iteration was - having multiple people review 15 different continuations per iteration, with multiple

iterations, means the number of assignments/tasks per song increases drastically as each of these factors is scaled up. While a larger crowd might potentially solve this if it were sufficiently large, there would still be the factor of cost. Increasing the number of tasks/iterations would increase quality but also sharply increase cost.

One of the main challenges we considered with scaling up was song length. Scaling up song length means we have to either:

1) Increase the number of iterations, which would lead to an increase in time/resources required per song
2) Increase the length of each continuation, would would lead to a decrease in quality/consistency in the song

Furthermore, the total time required for a worker to complete each HIT increases as the song length is increased since the worker has to listen to the entirety of the song in order to complete the HIT. This problem is exacerbated in our comparison HIT design, where a worker has to listen to more than one song and rank them in order to complete the HIT. This total time to complete rises drastically as song length increases.

Did you perform an analysis about how to scale up your project? For instance, a cost analysis?

No.

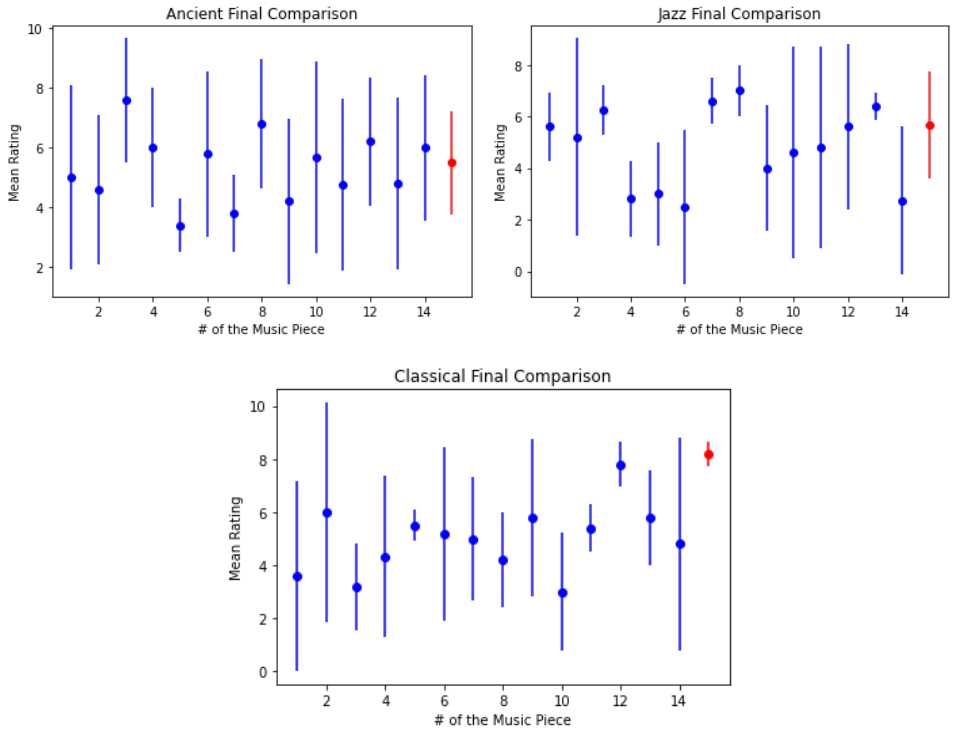
Project Analysis

Did your project work? How do you know? Analyze some results, discuss some positive outcomes of your project.

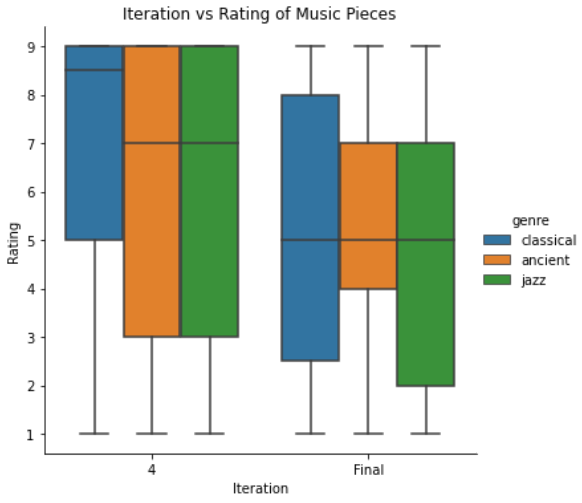
Our project was somewhat successful. We saw improvements in the mean scores from iteration 1 to iteration 4 across all genres. Furthermore, our analysis of the crowd's biases for certain genres was certainly successful and illuminating. To measure the final success of our project, for each genre we generated 14 other songs of the same length that our iterative approach produced. We then asked workers to rate these 15 in total songs using the same format as each iteration.

Do you have a graph analyzing your project? If you have a graph analyzing your project, include the graph here.

We found that while our iterative approach music piece was always in the top 50% of ratings (for the classical genre, it was the top piece) our pieces did not always achieve the top result as expected. Thus we constructed a way to create decent, but maybe not optimal music pieces. In the graphs below, the read point represents the music piece created from the iterative process.



However, when we analyzed the overall distribution of iteration 4 and the final comparison rounds we saw that our mean ratings for iteration 4 were much higher than the mean ratings of the final round.



This demonstrates that our iterative approach is a good way to produce consistently good music, while producing long songs in one go using the AI is a good way to produce consistently average, but sometimes one off good music.

What were the biggest challenges that you had to deal with?

One of our biggest challenges was finding primers that actually created good music when input into our model. We ran into this problem during the first posting of our HIT on MTurk Sandbox for our classmates. Our primer was too long and complex, and our continuations from the primer were also too long. The first wave of ratings was almost universally the lowest possible score as a result. Fine tuning our primers and parameters was critically important to generating better music and useful data.

Another big issue was with the timing of the HITs. We had a very large number of HITs required for this project, many of which could not be parallelized. These HITs would often take a longer amount of time as they went on - the first 40 assignments would go by very quickly, then the next dozen slower and the final dozen slower still. This is likely due to workers wanting to do batches of HITs, and thus being less inclined to select our tasks when they could not do so due to low remaining amounts.

Were there major changes between what you originally proposed and your final product?

Yes.

If so, what changed between your original plan and your final product?

Our original design did not involve using the crowd to help us generate individual songs. The first iteration of our project was to generate music from multiple models, and then have the crowd select the best model. From there, we would use crowd ratings to help us tune hyperparameters which led to the best music from that model.

We ended up veering away from that version since it did not use the crowd in interesting enough ways and overall was not as exciting. With help from our advisors we eventually settled on the current version of our project.

What are some limitations of your product? If yours is an engineering-heavy project, what would you need to overcome in order to scale (cost/incentives/QC…)? If yours was a scientific study, what are some sources of error that may have been introduced by your method.

There are definitely some limitations in our version of this project. One big limiting factor is the model we selected - while we could improve the primers and generate more continuations, the music will only be as good as this model can produce.

Cost and time were two big factors that limited the scale of the data we could collect. As mentioned in the scaling up section, our music would have better outcomes if we had more continuations for the crowd to choose from. However, adding more continuations drives up the cost and time required for the HITs to complete.

Some sources of error were the differing times of the day at which the HITs were posted for each iteration. This potentially introduced population biases and so skewed our results. Furthermore, as we showed different segments of the population had biases and so this could introduce error as well. A final source of error that could be present are network effects. Since the same workers could, and often did, work on all of our HIts at the same time, our results

could be showcasing the opinions of workers who heavily worked with us rather than a non-subjective, consensus opinion.

Did your results deviate from what you would expect from previous work or what you learned in the class?

Our project followed a similar idea as *Exploring Iterative and Parallel Computation Processes* by Little et al. In their paper they found that for image descriptions iterative processes were superior to parallel processes. In our project we reached a similar conclusion, but also found out that parallel processes were good at creating one off, but not consistently good results.

If your results deviated, why might that be?

We believe our results deviated slightly because the metric for what makes good music is much more broad and subjective than the metric for what makes good image descriptions. These additional factors in quality music created more opportunities and pathways for the AI to create one off good music.

Technical Challenges

Did your project require a substantial technical component? Did it require substantial software engineering? Did you need to learn a new language or API?

One technical challenge of this project is to generate music from a primer. Because we decided to use a pretrained model, this limits us from using the more complicated models, which do not provide this functionality. The final model we decide to use is an RNN-based model that is not intended for this task, and thus quality and coherency of the generated music was largely dependent on the primer.

One of our team members was already familiar with music generating machine learning models, which reduced the amount of new material we would have to learn as a group.

If your project required a substantial technical component, describe the largest technical challenge you faced.

There was not a large technical challenge since our focus was on the effectiveness of using the crowd iteratively in improving the quality of model generated music, rather than the model itself.