Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games

Edith Law
Machine Learning Department
Carnegie Mellon University
edith@cmu.edu

Luis von Ahn
Computer Science Department
Carnegie Mellon University
biglou@gmail.com

ABSTRACT
Since its introduction at CHI 2004, the ESP Game has inspired many similar games that share the goal of gathering data from players. This paper introduces a new mechanism for collecting labeled data using "games with a purpose." In this mechanism, players are provided with either the same or a different object, and asked to describe that object to each other. Based on each other's descriptions, players must decide whether they have the same object or not. We explain why this new mechanism is superior for input data with certain characteristics, introduce an enjoyable new game called "TagATune" that collects tags for music clips via this mechanism, and present findings on the data that is collected by this game.

Author Keywords
Human Computation, Games With A Purpose, Tagging

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

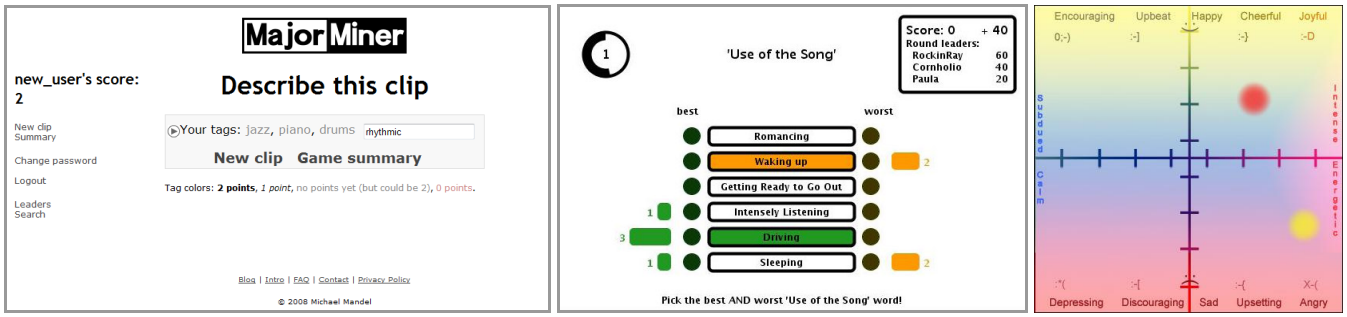
INTRODUCTION
People label, or tag, things in order to organize them and facilitate their retrieval at a later time. With the proliferation of multimedia objects on the Internet, collaborative tagging has emerged as a prevalent strategy for organizing content on the Web. A recent study shows that 28% of Internet users have tagged photos, news stories, or blog posts online [16]. Popular Web sites such as Flickr.com (photo sharing), Last.fm (music sharing) and YouTube.com (video sharing) have users contributing millions of tags each year.

However, there are two known issues with using such "social tags" as labeled data for multimedia objects. First, only the popular items are typically tagged, leaving a large proportion of the multimedia objects on the Web untagged [4]. Second, for multimedia objects with a time component, such as sound, music, and video clips, social tags found online often describe the object as a whole, making it difficult to link tags with specific content elements. This makes social tags unsuitable as data for training algorithms for music and video tagging, which rely on specific content elements being tagged (as opposed to the overall content).

Human computation is the idea of using human effort to perform tasks that computers cannot yet perform, usually in an enjoyable manner. The first human computation game, or Game With A Purpose (GWAP) [13], called the ESP Game [14], follows a specific mechanism: two players are given the same object (in this case, an image) that they are asked to describe, and the descriptions upon which the players agree become labels for that object. The ESP Game has become hugely successful: millions of image tags have been collected via the game, and a few years after its deployment, the game is still visited by a healthy number of players each day. Since then, this data collection mechanism has been adopted for other games in the domain of image tagging [1,7,10], music tagging [9,12] and knowledge extraction [11]. This mechanism is referred to as *output-agreement* [15] because players are given the same input and must agree on an appropriate output.

In this paper, we introduce TagATune, an online game developed to collect tags for music and sound clips. The initial design of TagATune [5] used the same output-agreement mechanism as the ESP Game: two players were given the same audio clip and asked to agree on a description for it. However, it was quickly apparent that this version of TagATune would not enjoy the same broad appeal as the ESP Game. This paper discusses why the output-agreement mechanism that works so well in many games failed to work for collecting data about sound clips in the TagATune prototype. Most importantly, we propose a new general mechanism for data collection in games upon which the final design of TagATune is based, and describe the conditions under which this new mechanism is applicable.

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-247-4/09/04…$5.00



(a) MajorMiner **(c) The Listen Game** **(c) MoodSwings**

Figure 1. Output-agreement human computation games for collecting music data

RELATED WORK

As explained previously, *output-agreement* games use matched outputs of the players as reliable descriptions of the input data. In Matchin [1], for example, two players are shown a pair of images and asked to vote for the one they think their partner will prefer. They are rewarded with points if their votes match. A global ranking of image preferences can then be derived from the aggregate votes. Another example is Squigl [7], a game for gathering segmentation data for images in which two players are shown the same image and an associated label, then are asked to draw an outline around the object in the image with that label. Points are awarded based on how much the two outlines of the object overlap. In PictureThis [10], players are shown a label and a list of images and asked to select the image that is the most relevant to that label. Players are again rewarded if their selections match. The output-agreement mechanism has also been extended to games for knowledge extraction, such as Ontogame and Ontotube [11], in which players are given various types of input objects (e.g., Wikipedia excerpts, YouTube videos, eBay auctions) and an ontology, then asked to annotate the input object using the given ontology. In all of these games, the reward system is the same as the one originally introduced in the ESP Game: matching on the output.

There are three recent human computation games for collecting data about music (see Figure 1), namely MajorMiner [9], the Listen Game [12], and MoodSwings [3]. MajorMiner [9] is a single-player game in which players are asked to enter descriptions for ten-second music clips. Players receive points for entering tags that agree with tags that were previously entered for the same music clip. The scoring system encourages originality by giving a player more points to be the first to associate a particular tag with a given music clip. The Listen Game [12] is a multiplayer game in which players are asked to describe 30-second music clips by selecting the best and worst tags from six choices. In the "freestyle" rounds, players can suggest new tags for a clip. Players are rewarded based on agreement and response speed. Finally, MoodSwings [3] is a game for annotating the mood of a given piece of music in which players are asked to indicate the mood, in terms of arousal and valence, by clicking on a two-dimensional grid. Players are given points for agreeing with each other in terms of the proximity of their mouse clicks. All of these games use variants of the output-agreement mechanism. Our new game, TagATune, employs a new mechanism that we describe below.

PROBLEMS WITH OUTPUT-AGREEMENT FOR AUDIO DATA COLLECTION

The main problem with using the output-agreement mechanism to collect data for audio clips is that it can be very difficult for two players to agree on a description. Unlike images, which often contain only a few clearly identifiable objects, music can be described by abstract concepts, such as "temperature" (e.g., chilly, warm), mood (e.g., dark, angry, mysterious), or the image it evokes (e.g., busy streets, festival), as well as categorizations that have no clearly defined boundaries (e.g., acid-jazz, jazz-funk, smooth jazz). The difficulty with arbitrary sound clips is even more marked, since the content is not always readily recognizable. In designing a game that is fun to play, it is important that the task is neither too easy nor too difficult. As we will describe later in this paper, the output-agreement mechanism, when used on sound clips, can become too difficult and frustrating for the players.

The game design strategies used in MajorMiner and the Listen Game reflect this underlying problem. Because it is difficult for two players to match on a tag, MajorMiner instead uses the agreement between a player and all previous players, while the Listen game uses the agreement among a group of players for a small predefined set of tags. There are disadvantages to such design approaches: having people play by themselves eliminates the social aspects of online games and limiting players to a predefined set of tags may make the game significantly less enjoyable and useful.

Lessons Learned from the TagATune Prototype Game

The difficulty of matching tags was revealed during the testing of the first prototype of TagATune, which used the output-agreement mechanism. In the prototype game [5], two players were presented with 30-second audio clips and

asked to type descriptions for them (see Figure 2). The initial prototype served sounds only (not songs). Players were rewarded when descriptions matched. "Taboo words" [14] were also used to encourage players to enter new tags.

Although the prototype game was able to collect semantically meaningful tags, the average enjoyability rating was only 3.4 out of 5, based on a survey submitted by 54 participants in a user study [5]. Moreover, it was found that 36% of the time, players opted to pass instead of entering a description [5].

Figure 2. TagATune prototype

There were two additional opportunities for gathering informal observations on the TagATune prototype game: a game demo session at the ISMIR 2007 Conference and a human computation workshop for elementary school students (held at Creative TechNight, a weekly event run by Carnegie Mellon School of Computer Science to foster young girls' interest in technology). In both game-playing sessions, the key observation was that players were often frustrated by being unable to match on a tag. Specifically, players often entered tags that meant the same thing, but that were expressed differently (e.g., 'cars on a street' versus 'traffic'). Moreover, since players were not allowed to communicate with each other (this requirement of output-agreement games safeguards against cheating), players found no good strategies to produce tags that match, except to enter tags that were as general as possible (e.g., 'music,' 'classical') or to rely on random chance.

A NEW MECHANISM

If we would like to create a game that labels any type of audio data, including sound clips and music, the natural question that follows is what other kinds of mechanisms can be used to collect data about input objects with high description entropy? We now describe a new mechanism for such data collection using games.

In this mechanism, two players are shown either the same object or different objects and each is asked to type a description of their given object. Unlike output-agreement games, where all communication is forbidden, all of the players' descriptions are revealed to each other. Based on these descriptions, the players must decide whether they have been given the same object. The descriptions that players enter are exactly what we are interested in. In a review article [15] which appeared after the deployment of our game, the mechanism underlying TagATune was referred to as *input-agreement*, a depiction of which is shown in Figure 3. It is important to note that input-agreement is a specific case of a more general mechanism, where players are asked to compute other functions of the inputs, instead of the "same or different" function.

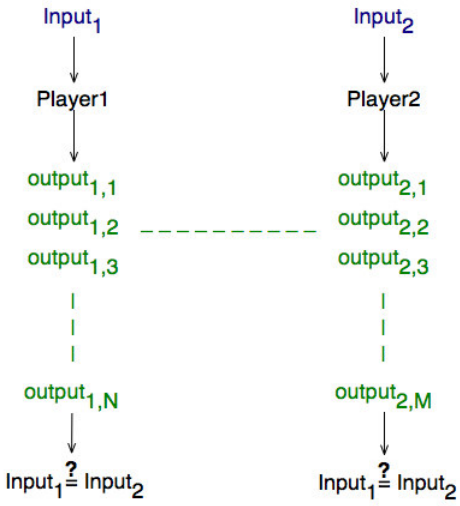


Figure 3. Input-agreement mechanism

TAGATUNE

The deployed version of TagATune is an instantiation of the input-agreement mechanism. A screenshot of the interface for a normal round of TagATune is shown in Figure 4.

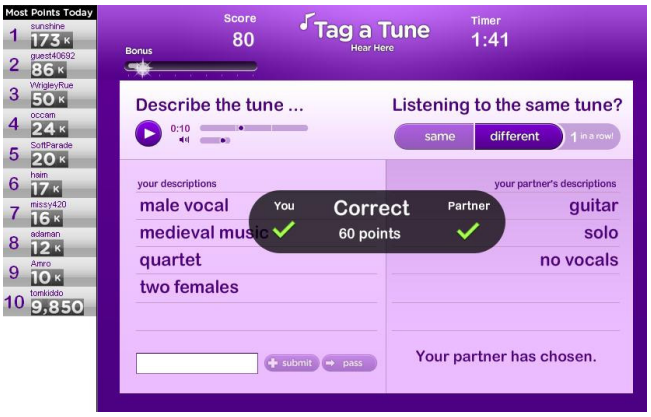


Figure 4. TagATune interface

In each round, two players are given either the same audio clip or different audio clips. They are provided with a basic music player interface to start, stop, and adjust the volume

of the audio clip to which they are listening. Each player describes the given audio clip by typing in any number of tags, which are revealed to the partner. By reviewing each other's tags, the players decide whether they are listening to the same thing by selecting either the *same* or *different* button. After both players have voted, the game reveals the result of the round to the players and presents the next round. The game lasts three minutes in total.

The inspiration for TagATune (and the input-agreement mechanism) comes from a psychology experiment [2] that studies the emergence and evolution of graphical symbol systems. The experiment involved a music drawing task, where pairs of participants were given a 30-second piece of piano music and were asked to draw on a shared virtual whiteboard. Based on the drawings, the players had to decide whether they had been given the same piece of music. Remarkably, using just drawings — whether abstract (e.g., contours, lines, or graph-like representations) or figurative (e.g., recognizable objects, figures, or scenes) — players were able to guess correctly whether their inputs were the same.

Input Data

The data currently served to the players consists of 56,670 short (~30 second) music clips from Magnatune.com and 28,715 sound clips from the FreeSound Database (http://freesound.org). Broadly speaking, the genres of music include classical, new age, electronica, rock, pop, world music, jazz, blues, heavy metal, and punk. All audio clips are provided under the Creative Commons License, allowing for much less restrictive usage than other typical music licenses. This allows audio files to be freely distributed to the public and greatly facilitates research on the data collected by the game. Moreover, the use of less well-known music minimizes the possibility that players will recognize the actual song or artist and simply describe the audio clip using tags that are already known. Finally, the shorter audio segments ensure that there is a more direct, though not guaranteed, link between content of the music and the descriptions provided.

For each round, the audio clips are selected randomly. Because the input data to the game is a pair of audio clips, the number of all possible pairs of sound and music clips is large enough that random selection suffices to ensure that players will not encounter the same pair of input data too often.

Scoring Mechanism

TagATune is a cooperative game, as can be seen from its scoring mechanism: the players score points only if they both guess correctly whether they are listening to the same audio clip. Neither gains points if one of them guesses incorrectly. This provides a natural incentive for players to be truthful to each other, which in turn, generates labeled data that accurately describes the audio clip at hand.

If TagATune were a competitive game, each player would be motivated to win against their partner, possibly by being malicious and misleading, and entering tags that did not describe the actual content of the audio clip. The consequence of this malicious behavior would be erroneously labeled data. Thus, a game that uses the input-agreement mechanism must be cooperative.

The points in TagATune compound: the more rounds the players successfully win in a row, the more points the players get for each subsequent round. This is a general scoring mechanism to motivate players that is shared by most games on the GWAP.com game portal, where TagATune is deployed.

Figure 5. Scoreboard

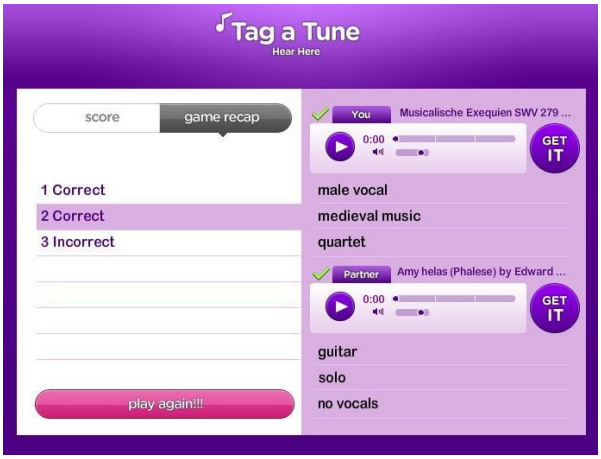


Figure 6. Game recap

A leader board is shown on the left of the main game panel throughout the game (see Figure 4). When the game is completed, a scoreboard displays the final score and the player's current GWAP level (see Figure 5).

A GWAP level is a rank assigned to players who attain a specific number of points and each level carries a special title. The scoreboard also shows the player's best score for TagATune, the player's total accumulated score, the

number of points needed to achieve the next GWAP level, and the number of points needed for the player to become the top player of the day. The leader board and scoreboard are game design elements common to all games on GWAP.com, and serve to motivate the players to strive for higher scores by playing better and more frequently.

Finally, the game recap provides an opportunity for players to learn from their mistakes by reviewing their detailed performance in the game (see Figure 6). Players can also replay every audio clip that was presented to them during the game, and click the *Get It* button to download the song. This *Get It* functionality gives TagATune a dual purpose as a Web application for sampling new music.

Bonus Round

When the players reach 1,000 points, a bonus round is added along with an extra minute of game play. During the bonus round, players are asked to listen to three pieces of music or sound clips. Each must decide individually which one of the three clips is most different from the other two. If they agree, they both obtain points. Figure 7 shows the interface for the bonus round of TagATune.

Figure 7. Bonus round

The reason for including a bonus round is that it produces two types of additional data. First, similarity data for music is useful for powering and improving music recommendation systems. Second, the similarity between songs is potentially a good indication of the level of difficulty that a particular pair of songs would present during a normal round of TagATune. More specifically, two songs that are very similar will require a great number of more specific descriptions in order for the players to distinguish them. This similarity data can be used later to adjust the difficulty of the game and thus increase the enjoyment for the players.

In this paper, we focus on the efficiency of TagATune in collecting high-quality annotation data for music. Therefore, our analysis will be centered mainly on the results of normal rounds of TagATune.

Implementation

The game engine for TagATune was developed using Java and MySQL. The front-end was developed using Flash, which has the advantages of being more platform-ready and browser-compatible than Java applets and Ajax.

TAGATUNE RESULTS

In this section, we report statistics of the data collected by TagATune over the course of the first seven months since its launch on May 15, 2008.

Game Statistics

A total of 49,088 unique games were played by 14,224 unique players, equaling 439,760 normal rounds. Based on the latest statistics collected in mid-December 2008, the number of games each person played ranged from 1 to 6,286, and the total time each person spent in game play ranged from three minutes to 420 hours. The average number of games played was four. Figure 8 shows the rank-frequency curve of how many people played x number of games. The graph almost resembles a power law: there are many people who played only a few games, and a few people who played many games. We refer to this rank-frequency curve as the *player retention* curve, since the curve is a useful indicator of the proportion of players who re-visited the game and the frequency of their revisits.

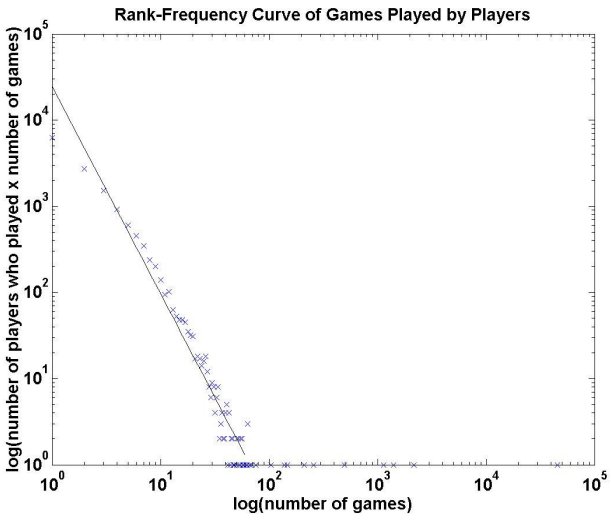


Figure 8. Number of people who played x number of games

The relative flatness of the slope of the user retention curve is a way to compare the enjoyability and popularity of different human computation games. A steep slope implies that many people played only one or a few times before abandoning the game, and not many people returned to play the game again. In contrast, a flatter slope indicates that only a few people abandoned the game after playing just a few times, while many people played a large number of games.

For example, Figure 9 shows a comparison of the player retention curves for different games on GWAP.com. The results show that the player retention curves for TagATune and Squigl are similar (in terms of slopes and intercepts), indicating that the number of players who played *x* number of games is similar between the two games, regardless of what *x* is. In comparison, there are more players for the ESP Game and Matchin, for any given number of games *x*. Finally, when compared to other games on GWAP.com, there are substantially more players who played a large number of games of Verbosity.

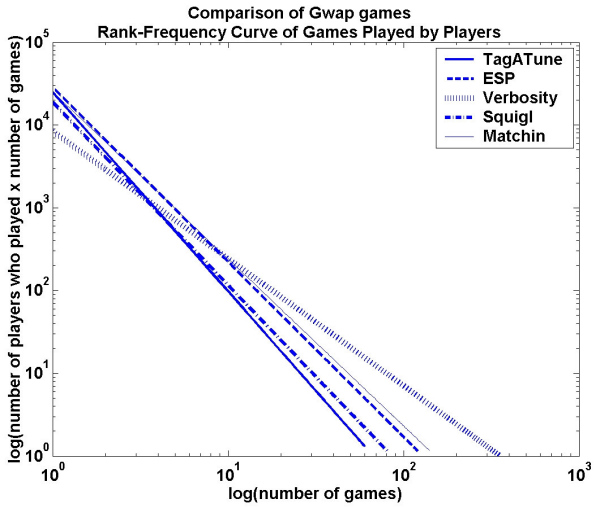


Figure 9. Player retention for games on GWAP.com

Of the 439,760 rounds, players only passed on 2,203 rounds, or 0.50% of the total number of rounds. In contrast to the 36% pass rate of the prototype version of TagATune, this indicates that players are less likely to give up on a round when the new mechanism is used. In 97.36% of the rounds, both players voted *same* or *different* before the end of the round. We refer to these rounds as *completed* rounds. The remaining 2.64% are called *missed* rounds, where one or both players did not submit their vote, most likely due to a timeout of the game.

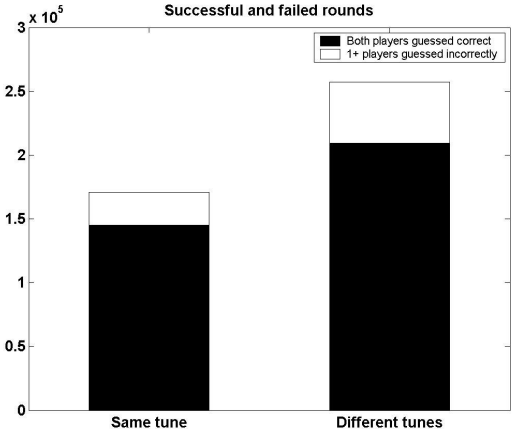


Figure 10. Successful versus failed rounds

Of the completed rounds, 80% were *successful,* meaning that both players guessed correctly whether they were listening to the same tune or different tunes; while 20% of the rounds were *failures,* where one or both players guessed incorrectly. Figure 10 provides a summary of these statistics. The success rate for rounds in which the tunes were the same was 85%, whereas the success rate for rounds in which the tunes were different was 81%, suggesting that it may be slightly harder to distinguish between tunes that are different.

Tag-Based Statistics

Prior to compiling statistics on the tags, a basic level of preprocessing was performed to convert all tags into lowercase, delete leading and trailing spaces, and remove punctuation marks (such as ?, !, ., *, - and quotation marks). After preprocessing, there were a total of 512,770 tags collected, of which 108,558 were verified by at least two players and 70,908 were unique. Based on this, the average number of tags generated per minute of play is approximately four.

Tag	Count	Tag	Count
classical	37,781	no vocals	6,126
guitar	30,093	soft	5,642
piano	27,718	sitar	5,413
violin	19,525	no vocal	5,285
slow	18,485	classic	5,228
strings	17,484	male	5,216
rock	17,413	singing	5,059
techno	15,627	solo	5,047
opera	14,512	vocals	5,014
drums	13,667	cello	4,966
same	12,610	loud	4,957
flute	12,149	woman	4,321
fast	11,435	pop	4,213
diff	11,046	male vocal	3,951
electronic	10,333	choir	3,576
ambient	8,733	violins	3,454
beat	7,683	new age	3,390
yes	7,352	beats	3,387
harpsichord	7,261	no voice	3,252
indian	7,255	harp	3,172
female	7,071	voice	3,080
vocal	6,964	weird	3,056
no	6,659	instrumental	2,946
synth	6,530	dance	2,896
quiet	6,167	female vocal	2,873

Table 1. Head List: top 50 most frequently used tags

The 50 most frequently used tags (the "head list") are shown in Table 1. There are a few observations. First, as also confirmed in other studies [9,12], the most common tags used to describe music fall into the categories of genre (e.g., classical, rock, techno), instrumentation (e.g., guitar,

piano, violin, drums, singing), or aspects of the music itself (e.g., fast, soft).

Second, there are some tags in the "head list"—specifically 'same,' 'diff,' 'yes,' 'no'—that have nothing to do with the content of the music, but instead are communication vehicles between partners in a game. Players use the words 'same' or 'diff' to signal their decision for that round to their partner. Although these *communication tags* are problematic, they are relatively easy to filter out since they often occur in the same formats.

A third observation is that this game generates *negation tags*, which are tags that describe what is *not* in the audio file, e.g., 'no vocals.' This is also a consequence of communication between the partners. For example, if one player types 'singing,' their partner might type 'no vocals' to indicate the difference between his or her tune and that of the partner. Other examples of *negation tags* include 'no piano,' 'no guitar,' 'no drums,' 'not classical,' 'not English,' 'not rock,' 'no lyrics,' etc. Negation tags are a unique product of TagATune and its underlying input-agreement mechanism, and are not often found in output-agreement games where communication is forbidden.

Finally, even among the most frequently used tags, there are still many equivalent tags that were considered distinct due to differences in spelling, wording, and pluralization. This property of the data is useful for search, since keywords entered by users can be just as varied. However, as a dataset for training machine learning algorithms, this indicates the need for more post-processing.

In contrast to the head list, the "tail list" consists of tags that have been used very infrequently. Some of the tags are simply uncommon, e.g., 'helicopter sound,' 'Halloween,' 'cookie monster vocals,' 'wedding reception music,' etc.

The tail list tags can be divided into four categories: misspelled tags, longer communication tags, compound tags (tags that contain multiple descriptors), and transcription tags (tags that transcribe lyrics). Examples of each kind are shown in Table 2.

Compound Tags	Transcription Tags
eastern female voice	fill me up...
long slow tones	rain on my parade
trombone and guitar	from shore to shore
light violin	the highest of sunny days
piano male voice	he'll never love you the way
Misspelled Tags	**Communication Tags**
churhc music	pick sooner
coubtry	you have to give me info
otiental sound	you're good too
instrumental	hello :)
ipano	yes agree

Table 2. Tail List examples

While the first two types of tail list tags are not of interest to us, compound tags can be converted into individual keywords for search, and tags which transcribe lyrics are invaluable since they support the prevalent strategy of searching for music by lyrics.

Tune-Based Statistics

On average, each game serves about nine songs. After seven months of game play, there were a total of 30,237 audio clips annotated and 108,558 *verified* (confirmed by at least two players) tags collected. Throughout this paper, the term "verified" is used to refer to tags that have high confidence (because they have been independently generated by multiple players) and "unverified" to refer to tags that have low confidence.

Figure 11 shows the number of the audio clips that have been tagged by x number of players. The data indicates that 92% of the audio clips have been annotated by two or more players, 61% have been annotated by ten or more players, and 26% have been annotated by 20 or more players. In order to attain a high level of confidence about the tags, an important criterion is that most songs are evaluated by multiple players. These results show that even using a simple random selection strategy for picking songs to present to players, this criterion is satisfied.

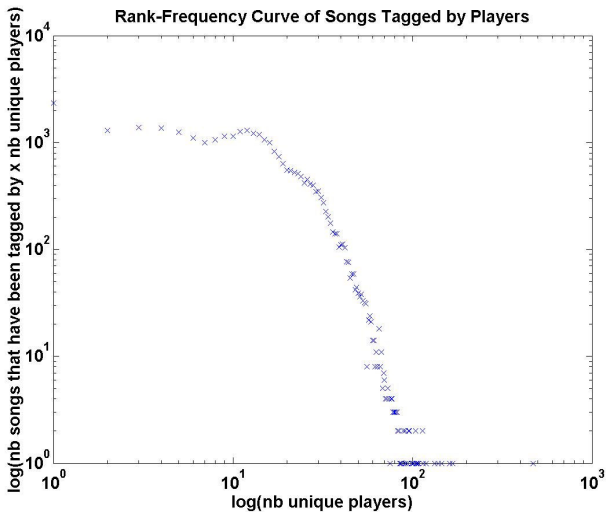


Figure 11. Number of songs that are tagged by *x* number of unique players

One question is whether allowing free-form text entry and open communication between partners results in tags that are accurate descriptions of the audio clips. In order to evaluate the quality of the tags, we conducted an experiment that evaluated how well the collected tags described the audio clips based on a small sample of the data.

Methodology

Twenty music clips with at least five verified tags were chosen at random, then 100 participants were solicited via Mechanical Turk (http://www.mturk.com) to answer a set

of 20 questions. For each question, the participant was given a music clip and was asked to answer four sub-questions. The first two sub-questions pertained to the quality of the verified tags, i.e., tags that were confirmed by at least two players. The second two sub-questions pertained to the quality of unverified tags, i.e., tags that were entered once only for that particular audio clip. Note that the number of verified and unverified tags varies among different music clips. On average, each music clip had around 7 verified tags and 17 unverified tags. The two sub-questions for the verified tags were as follows:

1. Which of the following tags would you use to describe the piece of music to someone who could not hear it?
2. Which of the following tags have *nothing* to do with the piece of music (i.e., you don't understand why they are listed with this piece of music)?

The same two sub-questions were asked for the unverified tags; we will refer to them in order as questions 3 and 4. For each question, participants were asked to count the number of tags that would be appropriate answers and to respond by a picking a number from a combo box.

Results

We retained results from 80 of the participants who spent at least 1,000 seconds on the task, which is the time needed to listen to the entire audio clip for each question plus at least five seconds to answer each of the four sub-questions. Note that for this experiment, we did not perform any post-processing to remove the easily filterable junk words—such as 'same,' 'diff,' 'yes,' 'no'—before presenting the tags to the participants. This is because we were also interested in finding out whether there were fewer junk words among the *verified* tags than *unverified* tags.

The results of this survey are summarized in Figure 12. As desired, for question 1 the mean was 78.26% (s.d.=9.45), equal to roughly 5-6 out of 7 tags. This indicates that the verified tags are useful for describing the audio clip. The mean of 16.67% (s.d.=8.59) for question 2, or roughly 1 out of 7 tags, indicates that there are very few of the verified tags that do not describe the audio clip at all. This small error can be attributed mostly to the easily filterable junk words that we decided to present to the participants during this experiment (such as 'same,' 'diff,' etc.).

The results for questions 3 and 4 indicate the quality of the unverified tags. One would expect the mean percentage for question 3 to be lower than for question 1, and the mean percentage for question 4 to be higher than for question 2. This is exactly what is observed in the results. For question 3, the mean is 51.84% (s.d.=7.33), which is equivalent to 8-9 out of 17 tags, indicating that in general, a smaller proportion of the unverified tags are useful for describing an audio clip. For question 4, the mean is 36.61% (s.d.=6.8) or 6 out of 17 tags, suggesting that a greater proportion of the unverified tags have nothing to do with the content of the music than the verified tags. The difference between the percentage of good quality tags in question 1 and 3 is statistically significant ($F(1,38)=92.74$, $p << 0.001$), and likewise for the difference between question 2 and 4 ($F(1,38)=62.96$, $p << 0.001$).

However, it is worth noting that there are usually many more unverified tags than verified tags. In some ways, the result is surprising in that a non-trivial proportion of the unverified tags actually describe the content of the music. This implies that the tail list of the collected tags is still potentially useful as data for search.

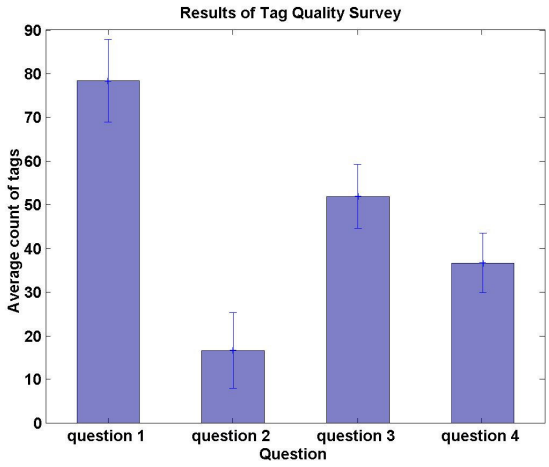


Figure 12. Results of questions 1-4 in tag quality survey

THE NEW MECHANISM REVISITED

One of the key ideas of the output-agreement mechanism first utilized in the ESP Game is that labeled data can be assigned high confidence if it is verified by multiple players, which motivated the use of *agreement*. In this paper, we have shown that agreement is neither the only nor always the best mechanism for data extraction. In this section, we outline the major characteristics of the input-agreement mechanism and the conditions under which it is most applicable for data collection.

Multiple Levels of Verification

The input-agreement mechanism allows multiple opportunities to verify that a tag is in fact a good description for an audio clip. First, each player's descriptions are *implicitly* verified by their partner during the game; that is, players will only choose 'same' if they believe that their partner's descriptions are appropriate for the audio clip they themselves are listening to. Likewise, players will only choose 'different' if they believe that their partner's descriptions do not adequately describe the audio clip. In other words, the task of guessing whether the players are listening to the same or different audio clips is a good indicator of whether the tags are appropriate for the audio clips.

A second level of verification takes place offline after the data has been collected, where descriptions become official

	TagATune	MajorMiner	The Listen Game	MoodSwings
Users	14,224	490	440	100
Clips Labeled	30,237	2,300	250	1,000
Data collected	108,558 verified tags	12,000 verified tags	26,000 choices	50,000 Valence-arousal labels
Unique Tags	70,908	6,400	120	Not applicable

Table 3. Comparison of human computation games for music (some of these statistics are taken from [3,9])

tags for the audio clip only if they are verified by greater than x players. The higher x is, the more confidence we have about the appropriateness of the descriptions for the audio clip. This utilizes the idea of *agreement* that is prevalent in the output-agreement games. However, the main difference here is that agreements between tags are not captured *during* the game, but *afterwards*. This is essential for collecting descriptions for data which has high description entropy—such as sounds, music, and videos—where agreement of descriptions between two partners is difficult to attain during the game, and which, in turn, may cause user dissatisfaction.

Lack of Cheating Strategies

The prevention of cheating is one of the major issues in the design of human computation games. In the ESP Game, for example, a pair of players can cheat if they settle on a strategy of typing in the same tag in order to match with each other, regardless of the content of the image. This problem is usually addressed by two countermeasures: (1) adding a delay in the player matching process so it is not guaranteed that two people who click 'play' simultaneously will be matched, and (2) giving players inputs for which the correct answers are already known.

An important property of the input-agreement mechanism introduced in this paper is that there is no obvious strategy for cheating. While our goal is to collect tags for audio clips, the objective of the game is not to tag, but to judge from the tags entered whether the players are listening to the same audio clip. There are three basic features of the input-agreement mechanism that result in a lack of cheating strategies, as well as a lack of need for cheating: (1) neither player holds the ground truth, (2) each player must derive this ground truth from the other's descriptions, and (3) players are rewarded only if both of them obtain the ground truth. In short, by being truthful to each other, players increase their probability of obtaining the ground truth and scoring points, which as a result, generates valid descriptions for the audio clips served in the game. The pre-agreed cheating strategies that are potentially detrimental to an output-agreement game are not a problem here, because the players are allowed to communicate anyway.

Increased Complexity of Collected Tags

One of the common problems in output-agreement games is that in their efforts to match with each other, players choose to enter short, obvious, and general descriptions. This problem is alleviated, but not completely solved, by the introduction of "taboo" words [14].

In input-agreement games, the goal is not to match on the tags, but to provide descriptions of the input data that are as detailed and accurate as possible so that the partners can guess the ground truth successfully. This allows tags to be longer and more varied. This is evident in the results obtained from the experiment presented in the previous section, showing that a non-trivial number of the longer, more complex tags in the tail list are valid descriptions of the audio clips.

Conditions of Applicability

As mentioned previously, the input-agreement mechanism can be applied to collect data about input objects with high description entropy. In fact, the TagATune game can be readily transformed to handle images, videos and text.

The applicability of the input-agreement mechanism is not limited to multimedia objects. Indeed, since the launch of TagATune, two games [6,8] have already been developed in the domain of Web search using a modified version of the input-agreement mechanism.

CONCLUSION

Until recently, efforts of research in human computation largely have centered around the development of games, with less focus on the invention of new mechanisms for data collection. The main contribution of this paper is the introduction of the input-agreement mechanism, a new method for collecting data in human computation games.

We developed a game called TagATune that uses this new mechanism to collect tags for music and sound clips, and presented statistics on the data collected during the seven-month period after the game was launched. The results (see Table 3) show that the popularity and throughput of TagATune are superior to other human computation games for collecting music metadata. Moreover, this new mechanism is readily extensible to images and videos, and is already being adopted for collecting data in other domains where the input data is text-based.

Future work includes using the data gathered from the bonus rounds of the game to adjust the difficulty of the

game, and to study the potential impact of this adjustment on the quality of the labeled data.

ACKNOWLEDGEMENTS

Many thanks to Jean-Julien Aucouturier for the insightful discussion, to Mike Crawford and Edison Tan for their help with the successful deployment of the game, and to Susan Hrishenko and the CHI 2009 reviewers for their feedback on this paper. This work was partially supported by generous gifts from the Heinz Endowment and the Fine Foundation, and by an equipment grant from Intel Corporation. Luis von Ahn was partially supported by a Microsoft Research New Faculty Fellowship and a MacArthur Fellowship.

REFERENCES

1. Hacker, S. and von Ahn, L. Matchin: Eliciting User Preferences with an Online Game. To appear in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems* (Boston, April 4-9). ACM, New York, 2009.
2. Healey, P., Swoboda, N., Umata, I., and King, J. Graphical language games: interactional constraints on representational form. *Cognitive Science*, 31:285-309, 2007.
3. Kim, Y.E., Schmidt, E., and Emelle, L. MoodSwings: A collaborative game for music mood label collection. In *Proc. 9th Intl. Conf. on Music Information Retrieval* (Philadelphia, September 14-18). 231-236, 2008.
4. Lamere, P. Social tagging and music information retrieval. *Journal of New Music Research.* 37(2):101-104, 2008.
5. Law, E., von Ahn L., Dannenberg, R., and Crawford M. TagATune: a game for music and sound annotation. In *Proc. 8th Intl. Conf. on Music Information Retrieval* (Vienna, September 23-27). 361-364, 2007.
6. Law, E., Mityagin, A., and Chickering, M. Intentions: A game for classifying search query intent. In submission.
7. Lee, B. and von Ahn, L. Squigl: A Web game to generate datasets for object detection algorithms. In submission.
8. Ma, H., Gupta, A., and Chandrasekar, R. Page hunt, page race and page match: using competitive and collaborative games to improve Web search and understand user behavior. In submission.
9. Mandel, M. and Ellis, D. A Web-based game for collecting music metadata. *Journal of New Music Research.* 37(2):151-165, 2009.
10. Mityagin, A. and Chickering, M. PictureThis. http://club.live.com/Pages/Games/GameList.aspx?game=Picture_This
11. Siorpaes, K. and Hepp, M. Games with a purpose for the semantic Web. *IEEE Intelligent Systems,* 23(3):50-60, 2008.
12. Turnbull, D., Liu, R., Barrington, L., and Lanckriet, G. A game-based approach for collecting semantic annotations of music. In *Proc. 8th Intl. Conf. on Music Information Retrieval* (Vienna, September 23-27). 535-538, 2007.
13. von Ahn, L. Games with a purpose. *IEEE Computer Magazine*, 39(6):96-98, 2006.
14. von Ahn, L. and Dabbish, L. Labeling images with a computer game. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems* (Vienna, April 24-29). ACM, New York, 319-326, 2004.
15. von Ahn, L. and Dabbish, L. Designing games with a purpose. *Communications of the ACM,* 51(8):58-67, 2008.
16. Weinberger, D. How tagging changes peoples relationship to information and each other. Pew Internet & American Life Project. http://www.pewinternet.org/pdfs/PIP_Tagging.pdf