Supporting Information

Barrington et al. 10.1073/pnas.1014748109

SI Text

Related Work on Human Computation

Herd It and similar "games with a purpose" (1–6) offer fun and competition as incentives to motivate wide spread human participation in scientific endeavors. Other human computation approaches engage participation on a volunteer basis (7–9) or use Amazon's Mechanical Turk* to offer small monetary rewards in return for completing data labeling tasks (10–16). The majority of applications have focused on classifying text (8, 10, 13) or images (9, 11, 12, 15–17) although speech transcription (5, 14) and video labeling (7) applications also exist.

Beyond labeling of multimedia data, human computation methods have been applied to numerous fields where the so-called "wisdom of the crowds" provides insight beyond what individual experts can offer. Crowdsourcing successes include prediction markets for sports betting (18), product development scheduling (19), company stocks (20), and political races (21).

Data collected by annotation games have been used to *evaluate* the output of machine learning systems. E.g., data from the ESP-game (1) has been used as a computer vision test set (22) and both MajorMiner (2) and TagATune (4) have been used to evaluate and compare different music tagging algorithms (26, 27). To date, attempts to use human computation to *train* machine learning systems as accurately as training them from expert data have focused on using Amazon's Mechanical Turk, rather than games. Novotney, et al. (14) use Mechanical Turk to crowdsource transcriptions of phone conversations and find a small reduction in performance of the resulting speech recognition system, compared to the same system trained on expert transcriptions. Ambati, et al. (13) also use Mechanical Turk to collect 3,000 English translations of Spanish sentences and train a machine translation system that rivals a system trained on expert annotations.

The proposed game-powered machine learning moves beyond monetary incentives and collects training data for free, a potentially more sustainable and scalable approach. For example, human computation *games* for annotating multimedia data have succeeded in collecting hundreds of thousands of tags for images (1) and music (4), a significantly larger scale than most Mechanical Turk applications [e.g., thousands to tens of thousands of tags for images (11) or speech transcription (14)]. We investigate whether online annotation games, which entice players with fun and real-time social interaction with a variety of other players, can train systems competitive with those trained from experts. We focus on labeling music. In this context, the social dimension of a game offers a natural setting to gauge opinions and collect annotations. The inferior performance we observed for training from TagATune vs. expert data (see Table S3 at the end of the *SI Text*) suggested to design and investigate a new game that is adapted to and actively integrated with the machine learning system to improve overall accuracy. Table S3 confirms the resulting system compares well to training from expert data.

Herd It Minigames

In order to collect information about diverse aspects of the music as well as to enhance engagement with players, Herd It features a variety of minigames that prompt the Herd to describe the music they hear by:

- catching floating bubbles that describe emotions or instruments present in the song,
- weighing responses to yes/no questions on a scale,
- selecting the most appropriate subgenre from a grid,
- plotting emotional valence and arousal intensity on a Cartesian plane (3, 28) and
- choosing the color that best matches the music.

Each minigame requires a single mouse-click for players to indicate their chosen tag. In addition to the bubbles game depicted in Fig. 2 in the main text, screenshots in Figs. S1 and S2 illustrate the remaining Herd It minigames.

Following each minigame, the player can earn 20 bonus points by correctly naming the song or the artist they have been listening to in a multiple-choice trivia round (see Fig. S3 for an illustration). The sequence of one minigame and one trivia round is repeated for five different songs for a total of 10 rounds. At the end of 10 rounds (lasting 2–3 min), a summary screen presents the final scores, lists the songs that were played during the game and encourages players to connect with the Herd and the rest of their social network.

Growing The Herd

To ensure we would generate sufficient Herd It participation to make the game-powered machine learning system viable, we engaged in a *user-centered* design process (23) to examine the effect of a variety of design features aimed at making the game fun, popular, and possibly viral. Our primary goal in this formative design process was to create a core gameplay experience that was understood by the majority of players and discover problems with the interface that would prevent reliable data collection. We also aimed to build a social gameplay experience that would entertain players and encourage them to share with their friends. Over a 10-month period we conducted regular user-studies both in our lab and in controlled online environments. Each test included a new, previously untested group of between 5 and 50 subjects and focused on free-form *issues-based* metrics to identify and prioritize crucial issues that were detracting from the user experience (e.g., interviewer observed user mistakes or user verbally expressed frustration or confusion during or after a test session) as well as structured *self-reported* metrics to provide a quantitative evaluation of the overall experience (e.g., user filled out a questionnaire immediately after a test session) (24). To examine the effect of repeated experiences with Herd It, we also conducted follow-up evaluations with some of the previous test subjects. These follow-up tests determined that users found the game easier to play and more enjoyable during second or subsequent play sessions.

Design issues were identified using a series of in-person interviews and free-form email feedback from online test subjects. Issues-based player quotes included:

- *It took me a while to figure out the agree-o-meter.*
- *The timer was done before I even knew what the game really was.*
- *Need clearer instructions—especially the XY minigame.*
- *The interface was kinda busy—it would be nice to have some kind of demo.*
- *I could play the tutorial but not get into the game.*
- *Stuck on "connecting to the herd".*
- *The song was playing, but the screen was blank.*
- *I could view the demo and instructions but none of the genre tabs opened.*

*http://mturk.amazon.com/

- *I heard a hiphop song when it was supposed to play Jimi Hendrix's Star Spangled Banner.*
- *The two last songs didn't play.*
- *I clicked well within the time, but it said I did not click before time ran out.*

The interviewers' notes from these sessions included the following observations:

- *the timer is a little fast*
- *the game slowed down with every player*
- *the scales game needs more questions*
- *on the XY game, 2 players consistently placed their mark on the axis line*
- *a player was clicking outside the clickable area on the XY game*
- *login is frustrating and turning players off*

This iterative process identified numerous challenges, inspired design solutions and tested the effect of these redesigns on users. In particular, we detail the development timeline of some of the innovative design features that were found to be crucial for improving interface usability and gameplay efficacy:

Month 1: compute player scores from percentage agreement with the rest of the Herd, illustrated with an "Agree-O-Meter," rather than an arbitrary scoring metric,
Month 2: replace the generic results screen used for all minigames with customized feedback animations for each minigame to show players how the Herd voted,
Month 3: integrate players' existing personal data and social network via Facebook, rather than requiring to create a new identity on Herd It,
Month 4: include "name-that-tune" trivia round after each minigame, both to inform players about songs they hear and just for fun,
Month 8: precache all audio clips and game files to minimize gameplay latency.

We aimed to create an enjoyable, positive experience for Herd It players so as to maximize time spent playing and user uptake (i.e., grow the Herd), and thereby the amount of data collected. Social features inspired by this formative design process enabled players to:

Month 5: recommend Herd It, share scores and issue challenges to Facebook friends,
Month 6: chat in real-time with members of the Herd,
Month 9: share music discovered during the game with Facebook friends.

In addition to free-form reporting of issues like those above, at the end of each testing session we asked each test subject to complete an online questionnaire that collected self-reported metrics about the overall experience. More specifically, the subjects responded to the questions listed in Table S1 on a 3- or 5-point Likert scale (25). Subjects also evaluated five of the Herd It minigames after each test, on a discrete scale of five ratings ("Great," "Good," "OK," "Bad," "Awful"; see Fig. S4). This self-reported data was collected after each of six user tests, spanning 10 m of user-centered design and testing.

Table S1 and Fig. S4 show subject responses from the final user test, at which point we had confidence that players who tried the game would likely understand Herd It and contribute meaningful data. To show how each iteration of the formative design process had a direct impact on the development of Herd It, we also chart the evolution of two key user metrics in Figs. S5 and S6 (over the 10 m of user-centered design). In the second test (conducted on the first day of month 3) we see that our initial reworking of the scoring metric and the improved player feedback developed after the first test had a positive effect on the user experience. Facebook integration was added after that (throughout month 3) and introduced some technical bugs and design challenges that caused some metrics to suffer during the two subsequent user tests. For the fifth test (conducted on the first day of month 8), we experimented with hosting the audio content on a 3rd party server which caused a lot of latency and dropped audio clips, leading to numerous complaints and worse test results. The latency problem inspired the solution (implemented during month 8) of preloading all songs before the game started. Once these challenges were overcome, the final test before launching Herd It (consisting of 50 subjects) showed that 80–90% of the players evaluated these two metrics as "Good" or "Great," while the number of test subjects giving ratings of "Bad" or "Awful" was reduced to zero. Considering all results in Table S1 and Fig. S4, we see that the majority of users consistently rated their experience as "Good" or "Great" in the final prelaunch test, and said they were likely to share the game with their friends or challenge friends' high scores.

After almost a year of controlled testing was complete and the core design goals were met, Herd It was launched to the public. Once online, the most current beta version of Herd It underwent continuous, larger-scale testing. The game was exposed to over 200 online users who provided feedback actively (web surveys, emails) and passively via *live-site* metrics (24) (e.g., ratio of new visitors who registered, clickthrough rates, number of games played, time on site). During this phase of the design process, the core gameplay experience remained unchanged. We focused on enabling and testing features designed to catalyze the continuous recruitment of human music labelers, i.e., (*i*) acquire new users and (*ii*) encourage existing users to return and play more games. For example, in order to track progress and save demographic details, a new player arriving at herdit.org was required to "add" the Facebook application before playing Herd It. Although this registration step was very simple (one button click: similar to adding a Facebook friend), it proved to be a barrier as not all players understood Herd It or why they should share their personal information. Our user tests determined that we achieved almost twice as many new players ultimately registered when they were launched directly into a short demonstration game, rather than being required to register immediately. In addition to instructing players on the rules of Herd It, this scripted demo was chosen to include well-known, popular songs and fun tags that would appeal to a wide audience and entice new users to continue playing. Once registered, a new player was brought to the Herd It home page. Multiple iterations of the home page design revealed that simplicity is key: to maximize time spent playing, the home page offers just a few simple buttons that immediately launch the user into the game. Our tests found that new users were 40% more likely to play a second game, once we streamlined the home page. All ancillary features (high score tables and statistics, friend invites, more information about Herd It, music search, etc.) were removed to secondary pages accessed from a list of tabs.

Having enticed a new player to join the Herd, a number of design and gameplay features were included to offer deeper content that would encourage them to return and to invite their friends to join. Indeed, one of the motivations for the social elements in Herd It's design (i.e., multiple simultaneous players, group-based scoring, Facebook integration, sharing of songs and scores, etc.) was to aid in the viral distribution of the game. For example, players were prompted to invite their friends at the end of a game where they accomplished certain achievements (e.g., setting a high-score, advancing in rank or surpassing a friend's score). Increasing ranks were awarded as users scored more points (e.g., "beginner", "rock star", "hip hop hero") and a scoreboard page tracked each user's progress daily, weekly and monthly and compared to their friends. Users could post clips of the songs they had enjoyed while playing Herd It on their Facebook wall, sharing the

musical experience with their friends. Finally, a blog described some of the science behind Herd It and polled users about suggested improvements to the game.

Once the design process was complete and the game was launched, Herd It was promoted to a wide audience of wouldbe players. We leveraged a number of external promotional channels, including: personal emails to friends and coworkers; viral promotion through players' social networks (suggesting Facebook friends to invite, issuing high-score challenges to friends and sharing songs on Facebook wall); affiliate promotion by inviting musicians to include their songs in Herd It and then promote the game to their fans; media articles and interviews, both in print and online (e.g., blogs, technology news sites).

Automatic Music Tagging

Machine learning approaches to modeling the association between semantic tags and spectral patterns in a musical waveform include discriminative learning algorithms (29–35), unsupervised learning algorithms (36), and generative models (31, 37–41). Of these approaches, generative models are generally better suited to handling weakly labeled data (i.e., where songs are labeled only with the presence of some relevant tags) because they estimate audio feature distributions that naturally emerge around audio content relevant to a tag, while downweighting irrelevant outliers. Furthermore, probabilistic rankings of relevant songs for a given query tag emerge naturally from a generative model.

One of the music autotaggers used in this work, which is also the focus of our active learning approach, is implemented using the generative machine learning model of ref. 37, based on Gaussian mixture models (GMMs). This model gave rise to a top performing automatic music tagger in the 2008 MIREX evaluation (26). After collecting training data (with some data collection method), the associations between a vocabulary of tags, $\mathcal{V}$, and a training song, $\mathcal{X}$, are represented as $\mathbf{y} = (y_1, ..., y_{|\mathcal{V}|})$ where $y_i > 0$ if the tag w_i has been positively associated with the audio of $\mathcal{X}$ (e.g., if the consensus of Herd It players agrees that the tag w_i is a good description for the song) and $y_i = 0$ otherwise. Fig. S7*A* shows an example of a group of songs that are all described with the tag "romantic." Spectral feature vectors $\mathbf{x}_i$ extracted from the audio waveform at regular time intervals, represent a song as a collection of vectors, or "bag of features," $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$, where T is proportional to the length of the song (Fig. S7*B*). The system learns a GMM of the audio features for each song, using the standard expectation-maximization (EM) algorithm (42) (Fig. S7*C*). These *song-level* GMMs are then combined efficiently into a *tag-level* GMM using the hierarchical EM algorithm of (43) (Fig. S7*D*). The result is a model of $P(\mathbf{x}|w_i)$, the distribution of acoustic features $\mathbf{x}$ that are associated with tag w_i:

$$P(\mathbf{x}|w_i) = \sum_{n=1}^{N} \alpha_n^i \mathcal{G}(\mathbf{x}|\mu_n^i, \Sigma_n^i),$$

where $\mathcal{G}(\cdot|\mu, \Sigma)$ is a multivariate Gaussian distribution with mean μ and covariance Σ, and the mixture weights α_n^i are such that $\alpha_n^i \geq 0, \forall n$ and $\sum_{n=1}^{N} \alpha_n^i = 1$. In this work, we use $N = 16$ component GMMs to model each tag.

To label a new song $\mathcal{X}$ using the vocabulary of tags, modeled as above, the likelihood of the bag of features that represents the entire song is inferred under the learned tag-level models using the naïve Bayes assumption of independence between features: $P(\mathcal{X}|w_i) = \prod_{t=1}^{T} P(\mathbf{x}_t|w_i)$ (Fig. S8*C*). Posterior probabilities of each tag, for the new song $\mathcal{X}$, are found using Bayes' rule:

$$P(w_i|\mathcal{X}) = \frac{P(\mathcal{X}|w_i)P(w_i)}{P(\mathcal{X})},$$

where $P(w_i)$ is the prior probability that tag w_i will appear in an annotation and is assumed to be uniform; $P(w_i) = 1/|\mathcal{V}|$. The song prior, $P(\mathcal{X})$, is obtained by summing the song likelihoods over all $|\mathcal{V}|$ tags in the vocabulary:

$$P(\mathcal{X}) = \sum_{v=1}^{|\mathcal{V}|} P(\mathcal{X}|w_v)P(w_v).$$

The final result is a set of *semantic weights*, $P(w_i|\mathcal{X}), \forall w_i \in \mathcal{V}$, probabilities that suggest how well each tag in the vocabulary describes the song's acoustic content. The semantic weights for each tag are collected in a *semantic multinomial*, a probability distribution that provides a rich description of the acoustic content of a song (Fig. S8*D*). While the alternative autotagging algorithms examined in the main paper [i.e., (32, 33, 41)] use different models of the acoustic content associated with each tag, they each allow to compute a similar probabilistic description of the semantics of a song's content. Given a semantic query, based on a tag or set of tags, the relevant dimensions of the semantic multinomials are selected to automatically rank songs by their relevance to the query.

Automatic Tagging Examples

Table S2 presents results for seven example songs, randomly chosen from personal music collections and never presented to the system before. Each song is analyzed using the GMM autotagger trained on all 127 tags for which Herd It has collected at least 10 reliable examples and then described by inserting the most likely genre, instrument, emotion, color, and usage tags into a template sentence (as in ref. 37). The resulting "robot reviews" qualitatively illustrate that the machine learning models trained on Herd It data reliably label new music with a variety of tags. In general, musically objective tags (e.g., "hip hop" and "disco") are well modeled by machine learning while the subjective tags collected by Herd It's more whimsical minigames are harder for machine learning models to predict (e.g., the color evoked by the music or songs that are "atmospheric" or "sexy").

Music Data

In this section, we describe in more detail the data used in our experiments, including the audio features, the *MGP* data and the *CAL500* data.

Audio Features. The method in ref. 37 used Mel-frequency cepstral coefficients (MFCCs) (44) to capture the spectral content of short-time segments (approximately 5 ms) from each song. For the GMM autotagger used in this work, we instead use the timbre coefficients computed using the feature extraction application programming interface (API) offered online by EchoNest.com, and described at http://developer.echonest.com/docs/method/get_segments/. This open API produces audio descriptors very similar in content to MFCCs but combines feature values over longer-time windows of homogenous audio (variable length but approximately 250 ms), resulting in a more concise representation of each song (i.e., 100's vs. 10,000's of feature vectors per song). Tingle, et al. recently showed that these EchoNest *timbre* feature vectors outperform MFCC feature vectors on the task of automatic music tagging when using the GMM-based system (45). As a result, we use this feature representation and similarly find a (slight) improvement in performance. For the machine learning methods that represent a song as a single feature vector [e.g.,

SVM (33) and Boosting (32)], we follow (33) and represent each song as the concatenated mean and variance of it's EchoNest timbre features. For the DTM model, which requires a constant interval between extracted feature vectors, we use the MFCCs described above.

Note: To augment *content-based* music search (based on features that describe the acoustic content only), a wealth of *metadata* [e.g., artist and song names, web search results (46), lyrics, song reviews, artist biographies, chart position, playlists, etc.] can be collected. To take advantage of this additional information to augment music search, one possible direction that has been explored, for example, is the use of kernel-based methods that can learn from multiple kernels. Such approaches combine kernels derived from the acoustic waveform with kernels derived from metadata extracted online, when available, improving the performance of using either data source in isolation [e.g., (34, 47)]. However, while metadata is readily available for popular songs, in the case of undiscovered songs, where a music search engine would be most useful, this information can not be relied upon (e.g., new songs have not yet been reviewed, unknown artists do not have lyrics or biographies available online). Thus, to satisfy the most general use case (including undiscovered music), our game-powered machine learning solution focuses solely on the only data that is guaranteed to be available: the acoustic waveform.

Training on Expert Labels: The Music Genome Project. The Music Genome Project (*MGP*), a subsidiary of the Internet radio station Pandora.com, employs musicological experts to annotate music using comprehensive surveys. Over the past ten years and at a cost of many millions of dollars, the *MGP* has labeled hundreds of thousands of songs with up to 500 tags.[†] While these labels are presumably of very high quality, the *MGP* lexicon is static: adding new styles of music or translating to another language requires laboriously retagging all songs with the new vocabulary and, as musical styles change, the tags can not be easily updated (e.g., contrast tags used to describe "classical" music with those relating to "hip hop" or even "hip hop" in 1981 with "hip hop" in 2011). Although we do not have access to the complete *MGP* data, we have collected a few *MGP* tags for 10,000 songs by retrieving publicly available information displayed on Pandora.com (45). Training machine learning models on this *MGP* subset allows to evaluate our music retrieval system when trained on data derived from a small group of expert labelers.

Evaluating on *CAL500*. At the University of California San Diego's Computer Audition Laboratory (CAL), we collected ground-truth data, similar to the Music Genome Project, by paying undergraduate music students to listen to songs and complete a survey that labeled the songs with relevant tags. The *CAL500* dataset consists of 500 songs annotated with 149 tags from categories related to musical genres, emotional content, instrumentation, vocal characteristics, and activities during which one might listen to the song (37). Each song was annotated by at least three individuals. It is important to note that, unlike most data mined from online sources, the *CAL500* dataset includes both positive and negative associations between songs and tags. Thus CAL500 is suitable for *evaluating* the output of automatic tagging systems as it is possible to judge correct—and also *incorrect*—results[‡].

The *CAL500* songs were used only for testing the machine learning models; none were included in the training sets. Each learned model is used to rank all the *CAL500* songs by their relevance to the tag. For example, the model for the tag "jazz" orders songs by how well they match patterns common to the jazz music in the training set. The ranking is evaluated in reference to the *CAL500* ground-truth labels by computing the *top-ten precision*, the proportion of relevant songs among the first ten results (e.g., the true number of "jazz" songs in the top-ten)[§]. High top-ten precision means that many songs at the top of the machine-ranked list are appropriate results for the query tag—exactly the goal of a music search engine[¶].

Dataset Availability. The song labels from all datasets collected for this work—Herd It, *MGP* and *CAL500*—are available from the CAL website: http://cosmal.ucsd.edu/cal/projects/AnnRet/.

While copyright issues preclude distribution of the audio files used in this work, all audio features are also available for download.

Comparison To Other Music Annotation Games

The first principal experiment in the main paper, comparing systems based on Herd It vs. *MGP* data, demonstrated that consensus derived from multiple nonexperts playing a human annotation game can train a multimedia retrieval system as reliably as data from expert labelers. With Herd It, we designed a unique, custom solution as opposed to using existing music annotation games such as TagATune (4) or MajorMiner (2) as a source of training data. In particular, TagATune connects pairs of players and asks them to describe the music they hear by typing tags on their computer keyboard. Based on their partner's tags, players must guess if they are listening to the same song or not. Likewise, MajorMiner's tagging is entirely user-driven. Herd It, on the other hand, is based on tag suggestion (and confirmation by users), to focus human effort on a larger, more varied vocabulary than TagATune or MajorMiner—which tend to focus on a small set of short, obvious words like, e.g., "drums," "rock," "guitar," etc. Moreover, while the aim of previous games was to have music directly labeled by humans, we designed Herd It explicitly to collect training data for labeling music with machine learning. By suggesting tags for players to confirm, rather than asking players to input tags, Herd It allows to focus data collection on tags (and songs) that are of most interest to the modeling process, in a dynamic way (enabling an active learning loop, as opposed to TagATune's or MajorMiner's more open-ended data collection).

To validate the motivation for designing Herd It as a new game, we empirically compare GMM-based models trained on TagATune and Herd It data. TagATune has released[‖] a dataset similar to Herd It's: 10,000 song clips labeled with 188 user-generated tags. MajorMiner data was not explored as it is a private evaluation set for the MIREX challenge (26). Evaluation of GMM-based autotagging models of the 14 tags that are common to TagATune, Herd It, *MGP* and *CAL500* demonstrates that, even with fewer examples of each tag, groups of Herd It players provide more reliable training data than do pairs of TagATune players (see Table S3).

[†]http://blog.pandora.com/faq/

[‡]Datasets that do not explicitly label negative associations between tags and songs are *weakly labeled*; the absence of a label may mean that the tag is truly not relevant or that no data was available for this song-tag pair.

[§]If there are only $n < 10$ relevant songs in the ground-truth, only the top n results are evaluated.

[¶]Other metrics that evaluate the complete ranking, such as mean average precision or the area under the receiver operating characteristic curve, exhibit qualitatively similar trends to the results reported here.

[‖]http://tagatune.org/Magnatagatune.html

1. von Ahn L, Dabbish L (2004) Labeling images with a computer game *In 22nd International Conference on Human Factors in Computing Systems* (ACM SIGCHI).
2. Mandel M, Ellis D (2008) A web-based game for collecting music metadata. *Journal of New Music Research* 37:151–165.
3. Kim Y, Schmidt E, Emelle L (2008) Moodswings: A collaborative game for music mood label collection *In 9th International Conference on Music Information Retrieval* (ISMIR) (PA).
4. Law E, von Ahn L (2009) Input-agreement: A new mechanism for collecting data using human computation games *In 27th International Conference on Human Factors in Computing Systems* (ACM SIGCHI).
5. McGraw I, Gruenstein A, Sutherland A (2009) A self-labeling speech corpus: Collecting spoken words with an online educational game *In INTERSPEECH International Speech Communication Association* (ISCA).
6. Cooper S, et al. (2010) Predicting protein structures with a multiplayer online game. *Nature* 466:756–760.
7. Volkmer T, Smit JR, Natsev A (2005) A web-based system for collaborative annotation of large image and video collections: an evaluation and user study *In 13th ACM International Conference on Multimedia* (ACM SIGMM).
8. Brew A, Greene D, Cunningham P (2010) Using crowdsourcing and active learning to track sentiment in online media *In 6th Conference on Prestigious Applications of Intelligent Systems (PAIS)* (IOS Press, Amsterdam, Netherlands).
9. Raykar VC, et al. (2010) Learning from crowds. *Journal of Machine Learning Research* 11:1297–1322.
10. Snow R, O'Connor B, Jurafsky D, Ng AY (2008) Cheap and fast—but is it good? Evaluating nonexpert annotations for natural language tasks *In 13th Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA).
11. Sorokin A, Forsyth D (2008) Utility data annotation with Amazon Mechanical Turk *In Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE Computer Society).
12. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. *Int J Comput Vision* 77:157–173.
13. Ambati V, Vogel S, Carbonell J (2010) Active learning and crowd-sourcing for machine translation *In 7th Conference on International Language Resources and Evaluation (LREC)* (Citeseer).
14. Novotney S, Callison-Burch C (2010) Cheap, fast and good enough: Automatic speech recognition with nonexpert transcription *In Human Language Technologies: 11th Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL HLT).
15. Nowak S, Rüger S (2010) How reliable are annotations via crowdsourcing: a study about interannotator agreement for multilabel image annotation *In 11th ACM International Conference on Multimedia Information Retrieval* (ACM MIR).
16. Vijayanarasimhan S, Grauman K (2011) Cost-sensitive active visual category learning. *Int J Comput Vision* 91:24–44.
17. Welinder P, Branson S, Belongie S, Perona P (2010) The multidimensional wisdom of crowds *In 24th Conference on Neural Information Processing Systems (NIPS)* (MIT Press, MA).
18. Dani V, Madani O, Pennock D, Sanghai S (2006) An empirical comparison of algorithms for aggregating expert predictions *In 22nd Conference on Uncertainty in Artificial Intelligence (UAI)* (AUAI Press, Arlington, VA).
19. Cherry S (2007) Bet on it. *Spectrum, IEEE* 44:48–53.
20. Surowiecki J (2004) *The Wisdom of Crowds* (Doubleday, New York, NY).
21. Berg J, Forsythe R, Nelson F, Rietza T (2008) Results from a dozen years of election futures markets research. *Handbook of Experimental Economics Results* 1:742–751.
22. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation *In 10th European Conference on Computer Vision (ECCV)* (Springer-Verlag, Berlin Heidlberg).
23. Gould J, Lewis C (1985) Designing for usability: key principles and what designers think. *Communications of ACM* 3:300–311.
24. Tullis T, Albert B (2008) *Measuring the user experience: collecting, analyzing and presenting usability metrics*, (Morgan Kaufmann, CA) .
25. Likert R (1932) A technique for the measurement of attitudes. *Archives of Psychology* 22:1–55.
26. Downie JS (2008) Audio tag classification. Music Information Retrieval Evaluation eXchange (MIREX) http://music-ir.org/mirex/wiki/2008:Audio_Tag_Classification_Results..
27. Law E, West K, Mandel M, Bay M, Downie S (2009) Evaluation of algorithms using games: the case of music tagging *In 10th International Society for Music Information Retrieval (ISMIR) Conference* (Japan).
28. Russell JA (2003) Core affect and the psychological construction of emotion. *Psychological Review* 110:145–172.
29. Slaney M (2002) Semantic-audio retrieval *27th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (IEEE, Minneapolis, MN).
30. Whitman B, Ellis D (2004) Automatic record reviews *In 10th International Society for Music Information Retrieval (ISMIR) Conference* (Spain).
31. Pampalk E, Flexer A, Widmer G (2005) Improvements of audio-based music similarity and genre classification *In 6th International Society for Music Information Retrieval (ISMIR) Conference* (London, United Kingdom).
32. Eck D, Lamere P, Bertin-Mahieux T, Green S (2007) Automatic generation of social tags for music recommendation *In 21st Conference on Neural Information Processing Systems (NIPS)* (MIT Press, MA).
33. Mandel M, Ellis D (2008) Multiple-instance learning for music information retrieval *In 9th International Society for Music Information Retrieval (ISMIR) Conference* (PA).
34. Barrington L, Yazdani M, Turnbull D, Lanckriet GRG (2008) Combining feature kernels for semantic music retrieval *In 9th International Society for Music Information Retrieval (ISMIR) Conference* (Philadelphia, PA).
35. Ness SR, Theocharis A, Tzanetakis G, Martins LG (2009) Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs *In 17th ACM International Conference on Multimedia* (ACM SIGMM).
36. Berenzweig A, Logan B, Ellis D, Whitman B (2004) A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal* 28:63–76.
37. Turnbull D, Barrington L, Torres D, Lanckriet GRG (2008) Semantic annotation and retrieval of music and sound effects. *IEEE Trans. on Acoustics, Speech and Language Processing* 16:467–476.
38. Tzanetakis G, Cook PR (2002) Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10:293–302.
39. Reed J, Lee CH (2006) A study on music genre classification based on universal acoustic models *In 7th International Society for Music Information Retrieval (ISMIR) Conference* (Victoria, Canada).
40. Hoffman M, Blei D, Cook P (2009) Easy as CBA: a simple probabilistic model for tagging music *In 10th International Society for Music Information Retrieval (ISMIR) Conference* (Kobe, Japan).
41. Coviello E, Barrington L, Lanckriet GRG, Chan AB (2010) Automatic music tagging with time series models *In 11th International Society for Music Information Retrieval (ISMIR) Conference* (Utrecth, Netherlands).
42. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39:1–38.
43. Vasconcelos N (2001) Image indexing with mixture hierarchies *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
44. Logan B (2000) Mel-frequency cepstral coefficients for music modeling *In 1st International Society for Music Information Retrieval (ISMIR) Conference* (MA).
45. Tingle D, Kim Y, Turnbull D (2010) Exploring automatic music annotation with "acoustically objective" tags *In IEEE International Conference on Multimedia Information Retrieval (MIR)* (NY).
46. Knees P, Pohle T, Schedl M, Schnitzer D, Seyerlehner K (2008) A document-centered approach to a natural language music search engine *In 30th European Conference on Information Retrieval* (ECIR).
47. McFee B, Lanckriet GRG (2009) Heterogeneous embedding for subjective artist similarity *In 10th International Society for Music Information Retrieval (ISMIR) Conference* (Kobe, Japan).

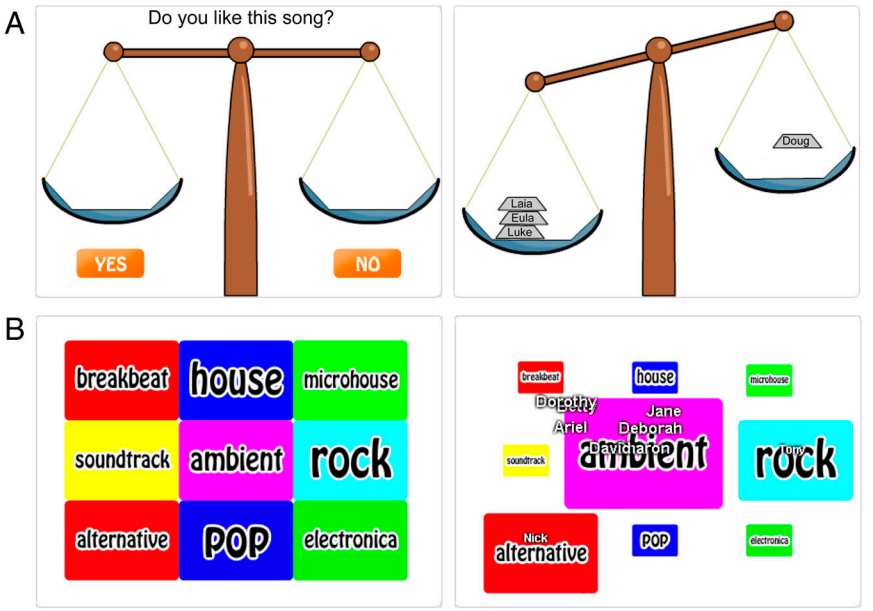

Fig. S1. Examples of Herd It's minigames that (*A*) weigh the Herd's response to a yes/no question and (*B*) determine the most appropriate subgenre for a song.

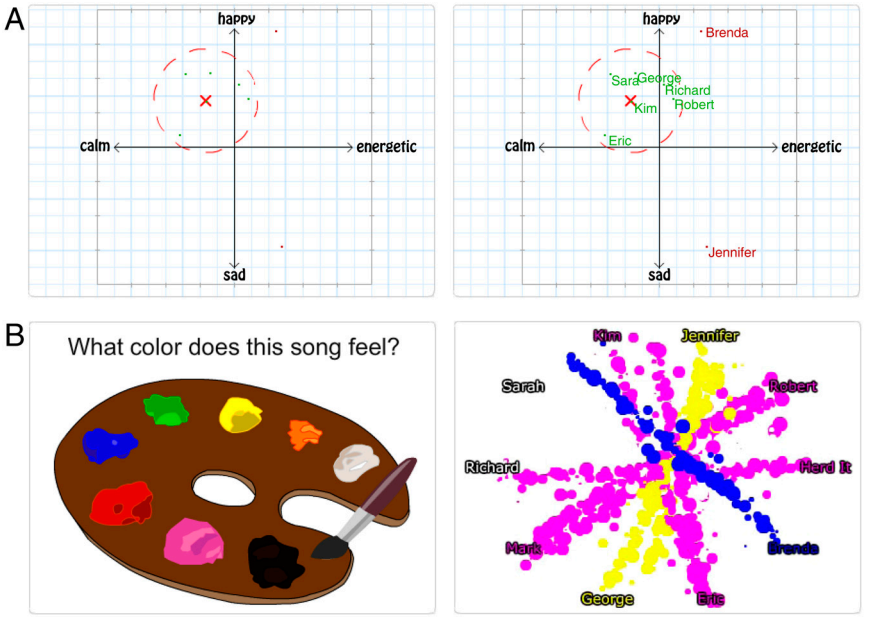

Fig. S2. More examples of Herd It's minigames that (*A*) collect two-dimensional valence and arousal on a Cartesian plane and (*B*) enquire about the color evoked by the music.

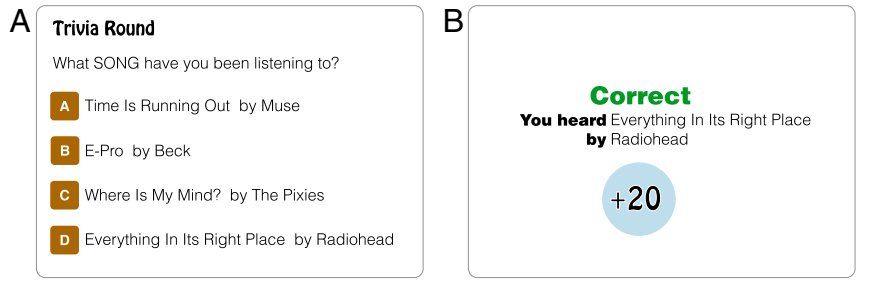

Fig. S3. Illustration of Herd It's trivia round. (*A*) Players can earn 20 bonus points for correctly naming the song that they hear. (*B*) The correct song is revealed after the player makes a choice. User testing determined that the familiar "name-that-tune" challenge of the trivia round encouraged novice players to participate in Herd It and the objective, "right/wrong" scoring provided a compelling counterpoint to the subjective, consensus-based scoring of the minigames.

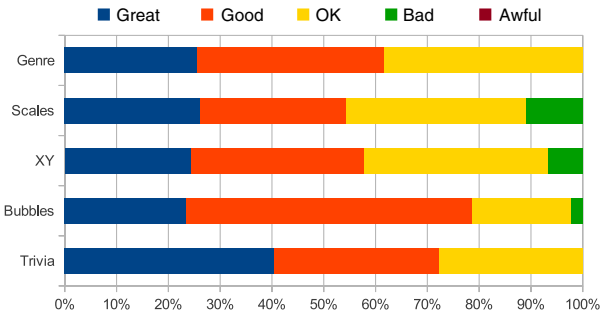

Fig. S4. Test subject evaluation of five of the Herd It minigames during the final user test.

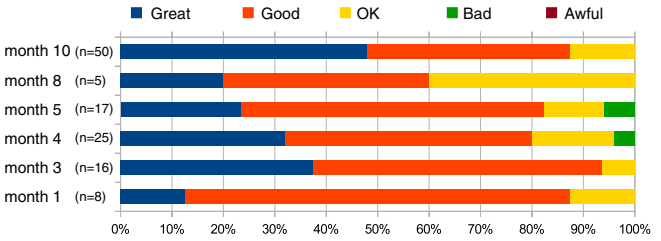

Fig. S5. Evolution of test subject responses to the question "*Overall how much did you like the game?*" during 10 m of user-centered design and development.

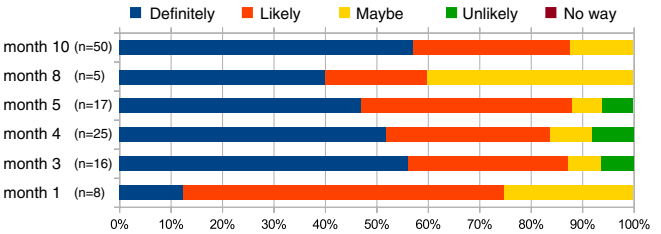

Fig. S6. Evolution of test subject responses to the question "*Would you play the game again?*" during 10 m of user-centered design and development.

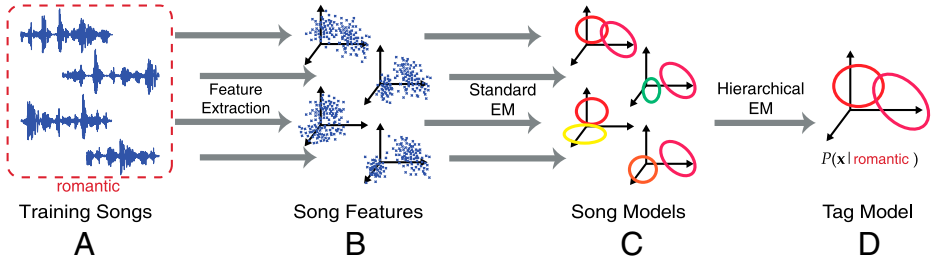

Fig. S7. Training an automatic music tagger. (*A*) Human labelers provide a training set of songs that have been reliably labeled with a given tag (e.g., "romantic"). (*B*) A "bag of features" represents the waveform of each training song. (*C*) A GMM of the acoustic features of each song is learned using the EM algorithm (42). (*D*) The GMMs for each song are combined efficiently into a single GMM that models the acoustic features predictive of the tag using a hierarchical EM algorithm (43).

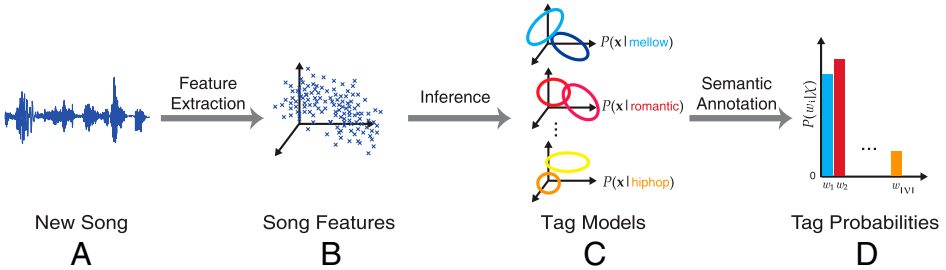

Fig. S8. Automatically tagging a new song. (*A*) A new, unlabeled song is to be analyzed by the automatic tagging system. (*B*) A "bag of features" represents the waveform of the song. (*C*) The features are compared to previously learned models of each tag. (*D*) Tag probabilities are obtained, providing a semantic profile that describes the song's acoustic content.

Table S1. Self-reported metrics collected from questionnaires after each of 6 user tests over 10 m

Question	Subject response				
	Hard core	Always	Occasionally	Rarely	Never
How much do you play Internet games?	4.0%	8.0%	46.7%	25.3%	16.0%
	great	good	Ok	bad	awful
How do you like each minigame? (5 questions)			*see Fig. S4*		
	very		somewhat		not at all
Are you aware of other people playing with you?	49.0%		44.9%		6.1%
	completely	mostly	sometimes	rarely	not at all
Do you understand how the scores are calculated?	28.0%	50.0%	16.0%	4.0%	2.0%
	absolutely	somewhat	not really	no	distracting
Is music necessary in the game?	100.0%	0.0%	0.0%	0.0%	0.0%
	great	good	ok	bad	awful
Overall, how much did you like the game?	47.9%	39.6%	12.5%	0.0%	0.0%
	definitely	likely	maybe	unlikely	no way
Would you play the game again?	57.2%	30.6%	12.2%	0.0%	0.0%
	definitely	likely	maybe	unlikely	no way
Would you recommend the game to your friends?	47.0%	34.7%	12.2%	6.1%	0.0%
	definitely	likely	maybe	unlikely	no way
Would you try to beat a friend's high-score if you saw it on your Facebook homepage?	43.8%	20.8%	29.2%	6.2%	0.0%

Results are shown from the final test with 50 subjects.

Table S2. Automatic music summaries produced by GMM-based machine learning models trained on Herd It data

Song	Automatic annotation
Wham! "Careless whisper"	This *soft rock* song features *male lead vocals*, feels *mellow*, evokes the color *white* and would be good to listen to *on a rainy day*.
Neil Young "Heart of gold"	This *folk* song features *bass guitar*, feels *slow*, evokes the color *orange* and would be good to listen to *late at night*.
Michael Jackson "The way you make me feel"	This *disco* song features *drum set*, feels *catchy*, evokes the color *yellow* and would be good to listen to *at dusk*.
Metallica "One"	This *rock* song features *male lead vocals*, feels *atmospheric*, evokes the color *orange* and would be good to listen to *in the morning*.
Lady Gaga "Poker face"	This *hip hop* song features *drum set*, feels *happy*, evokes the color *red* and would be good to listen to *at a party*.
The Flying Burrito Brothers "White line fever"	This f*olk-rock* song features *piano*, feels *acoustic*, evokes the color *orange* and would be good to listen to *in the morning*.
Eminem "Kill you"	This *hip hop* song features *drum machine*, feels *sexy*, evokes the color *black* and would be good to listen to *late at night*.

For each song, the tags in italics are automatically determined by the automatic tagging system to be the most appropriate genre, instrument, emotion, color and time.

Table S3. Comparison of the average number of training examples available from various data sources and the resulting music tagging performance

Dataset	Songs per tag	Top-ten precision
TagATune(4)	88	0.368
MGP(45)	851	0.422
Herd It	46	0.424

Top-ten precision, averaged over the 14 tags in common between all three data sources and *CAL500*, is evaluated in reference to the *CAL500* ground-truth, after training GMM-based models for each tag. While collecting the fewest number of songs for each tag, Herd It clearly provides more reliable examples for training machine learning models than TagATune. Models trained on examples collected by Herd It perform significantly better than those learned from TagATune data (paired t-test, 95% significance level).