Catching Earworms on Twitter: Using Big Data to Study Involuntary Musical Imagery

Lassi A. Liikkanen
University of Helsinki, Helsinki, Finland

Kelly Jakubowski
Goldsmiths University of London, London, United Kingdom

Jukka M. Toivanen
University of Helsinki, Helsinki, Finland

In recent years, so-called *big data* research has become a hot topic in the social sciences. This paper explores the possibilities of big data-based research within the field of music psychology. We illustrate one methodological approach by studying involuntary musical imagery, or *earworms* in the social networking service Twitter. Our goal was to collect a large naturalistic and culturally diverse database of discussions and to classify the encountered expressions. We describe our method and present results from automatic data classification and sentiment analyses. Over six months, we collected over 80,000 tweets from 173 locations around the world to obtain the most diverse dataset collated to date related to involuntary musical imagery. Automated classifications categorized 51% of all tweets gathered, with over 90% accuracy in each category. The most prominent categories of discussion concerned reporting earworm experiences, hyperlinks to music, spreading general information about the phenomenon, and communicating thankfulness (sincerely or ironically) about receiving earworms. Sentiment analysis revealed a balance towards negative emotional expressions in comparison to reference data. This is the first study to show this negative appraisal tendency and to demonstrate the 'earworm' phenomenon on a global scale. We discuss our findings in relation to previous literature and highlight the opportunities and challenges of big data research.

Received: April 8, 2014, accepted December 8, 2014.

Key words: involuntary musical imagery, big data, methodology, music experiences, social networks

Involuntary musical imagery (INMI), or *earworms*, the experience of usually repetitive inner music in the absence of an external source, has recently received increasing attention from music psychologists. Since a few seminal studies (Bailes, 2006, 2007; Liikkanen, 2008), a number of peer-reviewed journal publications have quickly emerged around the topic. The papers have followed two main lines of investigation: describing and measuring the prevalence and phenomenology of INMI (Beaman & Williams, 2010, 2013; Beaty et al., 2013; Halpern & Bartlett, 2011; Liikkanen, 2012a, 2012b; Müllensiefen et al., 2014; Williamson & Jilka, 2013; Williamson et al., 2012; Williamson, Liikkanen, Jakubowski, & Stewart, 2014) and investigating the mechanisms that promote or impede its appearance (Byron & Fowles, 2013; Hyman et al., 2013; Liikkanen, 2012c; Williamson et al., 2012). A few papers have also discussed the distinction of INMI from clinical phenomena of similar characteristics, such as musical hallucinations and obsessions (Hemming & Altenmüller, 2012; Liikkanen, 2012d; Liikkanen & Raaska, 2013). Researchers have now accumulated some knowledge about the nature of INMI, but many questions remain unanswered regarding this elusive subject, such as whether INMI serves some functional purpose, the efficacy with which INMI can be controlled, and the effects of INMI on our everyday life.

In many disciplines of empirical inquiry, the past several years have witnessed a rise of a so-called *big data* movement. This movement is expected to influence the field of music research as well. Big data research does not refer to any one specific discipline, but represents a new methodological approach applicable to the study of human and non-human behavior alike. Currently, Internet search engines and social networking services are prominent sources of big data generation. In terms of empirical research, the big data approach can offer opportunities to gain access to data from a wider population, in comparison to the small samples often used in lab studies (Huron, 2013).

Big data is defined by the nature of the data, as well as its processing requirements. The three characteristics of big data include *big volumes*, *big velocity*, and *big variety* (Stonebaker, 2012). Volume refers to the excessive

Music Perception, volume 33, issue 2, pp. 199–216, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2015.33.2.199

quantity of data, velocity to its accumulation speed, and variety to its unstructured nature. Each characteristic sets unique requirements and the term big data can currently be associated with data possessing any (or all) of these characteristics.

Another new term associated with big data is *data mining*. Data mining refers to automated approaches to data extraction and analysis. This usually relates to unconventional statistics and methods of data analysis, for instance using machine learning for data clustering and classification. Due to automatization, the process remains somewhat opaque and its results must be validated to check how well the automated tools are performing. Although the big data approach often necessitates some degree of inherent uncertainty as the human researchers are generally incapable of exploring the entire dataset, it also presents a wealth of opportunities to learn about human behavior by using datasets that are impossible to explore with traditional methods (see Huron, 2013).

Although big data research began within the field of computer science — especially the study of networks (Barabasi & Albert, 1999; Kossinets & Watts, 2006) — a number of successful studies in the social sciences have recently emerged as well. For instance, a study using search engine queries was able to accurately predict the revenues generated by movies, video games, and music (Goel, Hofman, Lahaie, Pennock, & Watts, 2010). An experiment among 61 million American Facebook users during the 2010 Congressional election showed that social media stimulated people politically and increased voting activity (Bond et al., 2012). Psychological phenomena have also been investigated. For instance, a study of expectant mothers was able to predict postpartum changes in mood and behavior based on Twitter data (De Choudhury, Counts, & Horvitz, 2013). Another study using data from Facebook volunteers quite accurately predicted a number of personal attributions, including political and religious views, personality traits, and sexual orientation, based on the pages that each participant "liked" (Kosinski, Stillwell, & Graepel, 2013).

In the present paper, we explore the possibilities that big data has to offer for music psychology research. We document a method to study how INMI fuels discussions on a social networking service. We chose a big data approach for the present study because although the majority of people experience INMI at least every week (Liikkanen, 2012a), this topic does not seem as frequent in everyday discussions. Therefore, in order to capture naturalistic data surrounding this phenomenon, a large sample of potential input was required.

Our study aimed to collate the largest and most culturally diverse dataset related to INMI experiences to date. Our detailed goals were to 1) investigate the occurrence of INMI among Twitter users globally, 2) classify the types of INMI-related discourses on Twitter, and 3) assess the emotional valence of INMI-related tweets. Each of these questions is motivated by existing studies, but can here be assessed on a quite different scale. The Twitter data sampled for this study represents comments that were never intended for scientific research. As such, this dataset currently represents the most naturalistic set of INMI data collected, even when taking into account several previous studies that aimed to use relatively naturalistic methods such as diaries and experience sampling.

Background

In this research report, we present a brief review of the essential elements of our study — Twitter and the literature on INMI. We will begin by introducing the main features of Twitter.

Twitter is a social networking, or *microblogging*, service that was founded in 2006 in the United States. As of 2014, over 200 million registered users visit Twitter every month, 77% of them outside the US.[1] The main communication medium within Twitter is a *tweet*, which is a text message that may have at maximum 140 characters. Tweets are sent by registered users identified by an @ sign followed by a unique *username* (e.g., @drearworm). The registered username implies that the name belongs to a unique individual, however this is not necessarily the case, as one person or an organization may administer several user accounts, or artificial intelligence known as Twitter bots may generate tweets. In the US, Twitter users were most frequently 18- to 29-year-old adults (31% use Twitter vs. 12% in older cohorts) and African-American (29% use Twitter vs. 16% among white and Hispanic) in a data set collected in 2013 (Pew Research Center, 2014).

The social network within Twitter consists of users following each other. The tweets from the followed users create the stream of messages to which people normally attend. The follow relationship is unidirectional — unprotected accounts can be freely followed by all, but Twitter does not require that this person follows you, unlike in friend networks (e.g., Facebook). As a consequence, it is common for Twitter celebrities to have millions of followers, but to follow only a few others themselves. *Retweeting* is the act of forwarding someone

[1] https://about.twitter.com/company accessed February 12, 2014

else's tweet to people who follow you. This effectively circulates information and promotes viral phenomena, including Internet memes, which are actively shared new ideas that become widely recognized or mimicked (Bauckhage, 2011).

Tweets themselves embed important features. *Hashtags* — word or phrases prefixed by the number sign (e.g., #earworm) — are mechanisms used to track and group messages together. They are utilized for several purposes, for instance, to create ad hoc discussions (e.g., #grammys), to help others find tweets based on the tag (e.g., #psychology), and to comment on the previous information in the tweet (e.g., #humor). Inside a tweet, usernames can be used to post private messages or just to reference or mention the user by inserting their @username as a part of the tweet. It is also common to include hyperlinks as part of tweets. Because hyperlinks and the associated Internet addresses can be very long, Twitter automatically shortens the hyperlinks to a uniform address format for circulation and storage. For instance, a link pointing to http://www.youtube.com/watch?v=61XZA-l2PJQ becomes http://t.co/PGdodRzr in the process. It should be noted that while regular usage of Twitter can happen on a personal computer web browser or a dedicated mobile application, research and analytics access the service via an API. This acronym stands for *application programming interface*, a general term used to describe how a service allows software developers access to, and use of, the service.

The popularity and openness of Twitter has been noticed by researchers as well. There are already a great number of studies related to this service. We will mention here just a few of these studies concerning the nature of discussions on Twitter. Kwak, Lee, Park, and Moon (2010) examined some 106 million tweets to discover the content of the tweets and how they circulate. They reported that over 85% of tweets involved daily news of current events and retweets on average reached 1,000 people. They also reported that only 22% of follow relationships are reciprocal, yet any two random users are connected to one another by approximately 4.1 intermediate users. The Twitter network thus responds quickly to news and can even generate momentum of its own. Tweets have been shown to be predictive of stock market changes (Bollen, Mao, & Zeng, 2011), correlate with music record sales (Kim, Suh, & Lee, 2014), and are useful for earthquake detection and notification (Sakaki, Okazaki, & Matsuo, 2010). Finally, some studies have found that the majority of information circulated on Twitter originates from relatively few opinion leaders (Wu, Hofman, Mason, & Watts, 2011). Another study of Twitter users proposed a categorization of users based on whether they share information of general importance (e.g., news) or about themselves (Naaman, Boase, & Lai, 2009). It respectively labeled users as *informers* or *meformers* respectively, with the latter being almost thrice as numerous as the former.

Next we review INMI research to consider the previous findings that the present study can build upon and extend. It has been known for some time that INMI is a common psychological phenomenon (Liikkanen, 2012a) at least in the Western world. This phenomenon is ubiquitous, yet difficult to directly observe; therefore most previous empirical studies have aimed to capture the INMI experience using questionnaires (e.g., Baruss & Wammes, 2009; Liikkanen, 2008) and diary studies (e.g., Beaman & Williams, 2010; Halpern & Bartlett, 2011). However, these fairly naturalistic methods can be subject to recollection bias, demand characteristics, or other biases imposed by being an informed subject of a study. We aimed to observe INMI discussions and behaviors in a naturalistic setting (Twitter), in which users can discuss their feelings and experiences without specific expectations of external observation by an experimenter.

Furthermore, a recent study of two combined datasets did an elaborate qualitative analysis of INMI evaluations and behaviors (Williamson et al., 2014). The researchers found that people both enjoy and try to suppress the phenomenon through various practices of engagement with, and distraction from, the INMI tune. The study also explored the topic of how people evaluate the experience of INMI, finding that the general response to INMI was neutral or positive. This resonates with multiple past questionnaire studies in which people have reported generally positive attitudes towards INMI (Beaman & Williams, 2010; Halpern & Bartlett, 2011; Hemming, 2009; Liikkanen, 2012a; Müllensiefen et al., 2014).

As the first empirical investigation into the discourse related to INMI on Twitter, the present study was foremost descriptive, but also posed three explorative research questions related to INMI. The first goal was to find evidence of INMI experiences across geographic areas. The second aim was to explore whether INMI-related discourses on Twitter correspond to those previously reported in the literature (Williamson et al., 2012; Williamson et al., 2014), and the third aim was to investigate whether the emotional valence of INMI-related tweets, as measured by sentiment analysis, would be positively biased, as expected by previous studies that have employed small to moderate sample sizes (see previous paragraph). The methods will be

described next, but in brief, the present study consisted of a search of a small set of keywords that were used to retrieve tweets matching these terms via a programming interface.

Method

Our goal was to answer three questions around the phenomenon of INMI based on discussion taking place on Twitter. To do this, we first collected data, preprocessed it, created appropriate data mining tools and utilized them, and then validated and iteratively improved the analysis. We will describe the details of this procedure in the following subsections.

DATA COLLECTION

Twitter provides an opportunity and a challenge for research. In a filing released in early October 2013, Twitter announced that its active users post over 500 million tweets every day (Twitter Inc., 2013). The huge volume of traffic on Twitter has led the service to restrict search capabilities of the streaming API so that at maximum 1% of the total traffic can be accessed.[2] If a search produces results totaling less than this 1% of the total traffic, the API can return all related results.

To collect our data, we used the Twitter streaming API to retrieve the tweets matching a small set of keywords (*earworm, ear worm, ear-worm, stuck tune*). Because the volume of the matching results was far below the 1% limit, we believe we were able to capture all relevant public tweets during the collection period. We ran our data collector for six months starting from November 2012 until May 2013, which collected data for 24 hours every day in one-hour intervals. The technical description is provided in an Internet code repository at http://www.cs.helsinki.fi/group/discover/earworms/.

PREPROCESSING

Multiple data processing steps were required prior to the analysis of this dataset. Our procedure involved the following steps:

a) Reformatting and filtering out unnecessary data fields from the tweets
b) Converting shortened hypertext links into final destinations URLs
c) Filtering for content: filtering out irrelevant data
d) Decoding user location data

In step a) we reviewed the available data to decide which metadata fields might be informative to the subsequent data analysis. In this case, the structure of the original, native JSON data entity (see the Internet code repository) had 19 entries directly related to the tweet, 38 related to the user, and three place holders for multidimensional data entities. Out of this data, we eventually retained only seven; three fields related to the tweet (the tweet itself, date and time, and retweets count), and four related to the user (name, language, location, and time zone). The data were reformatted into a comma-separated value format for universal compatibility.

Twitter automatically converts URL links embedded in tweets to its native abbreviated format. In processing step b) we converted t.co links back to their original destinations. This allowed us to observe the Internet content to which the Twitter users were linking. Details of the scripts utilized for this purpose are included in the Internet code repository.

After the previous information enriching and compacting operations, the third step c) in the analysis concerned filtering out irrelevant data. This required some random sampling and manual review of the data. In our case, the only evident source of noise seemed to be the inclusion of "DJ Earworm" related content. This data had matched the keyword search but seemed unrelated to the topic of our inquiry, as DJ Earworm is an artist who creates music mashups and does live DJing. Thus we filtered out tweets containing clear references to DJ Earworm. This resulted in discarding approximately 1,100 tweets. Additionally, we removed all tweets that were not in English ($N = 2{,}904$) according to the language field.

Next, we intended to ascertain some descriptive information about the contributing Twitter users. Describing the subjects contributing to big data can be difficult, because information such as age or gender is not necessarily present in the data. We knew contributing users' usernames, time zones, language, the time of tweet, and their self-reported location, sometimes complemented by a *geotag*, which identified the user's location at the time of posting the tweet. However, with the exception of the user name, all other information is optional to Twitter users. For instance, geotags were missing from most of the present data and could not be relied upon.

To ascertain some idea of the users' demographics, we analyzed their self-reported, free form location information. Many people report their home country or city, but some may use GPS coordinates, for instance "6.6595699, 3.2897109" (corresponding to Lagos, Nigeria), some variation of a real name, e.g., "Only in California," or something metaphoric or witty, e.g., "On the road to riches." Given that we had over 20,000 unique entries within the location field, we utilized Google Geocoding API (part of

[2] https://dev.twitter.com/discussions/12692 accessed February 18, 2014

Maps API)[3] to resolve the locations into known countries, territories, or special geographic areas (as per ISO 3166-1 that includes 249 such names). This allowed us to decipher the locations in a consistent way. The system was able to correctly interpret most input, for example "#pantsless in Kentucky" matched "United States," but "170 yards from the Erie line" was not found (for technical details, see the Internet code repository).

DATA ANALYSIS

Data mining requires custom computational solutions. We created a series of scripts for Python, PHP, and BASH to analyze the dataset and produce results. The analysis tasks that we automated were data classification and sentiment analysis; these will be described in detail in the following section. In addition, we extracted the frequency distributions of user mentions and hyperlinks inside the tweets. Data analysis tools such as the SciPy library for Python, SPSS, and Microsoft Excel were used to derive descriptive statistics based on the results of the automated classification.

Automatic data classification. Understanding natural language, or acting upon its semantic content, has traditionally been a difficult issue for computer science and artificial intelligence. However, by working around constraints and accepting some limitations, we can extract meaning from text. For this study, we needed a solution to automatically classify tweet content with as little human supervision as possible.

We created a custom data classifier by iteratively developing the classification algorithm and then manually evaluating its effectiveness. The algorithm development, which is described in detail below, resembled the grounded theory technique (Charmaz, 2006; Corbin & Strauss, 1990) that has previously been used to analyze qualitative data in INMI research (Williamson et al., 2012; Williamson et al., 2014). The classifier development shared at least three commonalities with grounded theory: 1) the analysis is guided by the researcher, 2) the model of the data develops through time, and 3) the results are manually assessed at the end. We chose this approach over machine learning because we had no expectations of what could be discovered from the data. Machine learning methods — for instance support vector machines or Naive Bayes — require at least some manually annotated data before a classifier can be built. In the present study, we seek to provide this annotation.

[3] https://developers.google.com/maps/documentation/geocoding/ accessed February 18, 2014

Grounded theory is a qualitative analysis technique that is appropriate for synthesizing and categorizing large text-based data sets in which responses are widely variable. Participant responses (in this case tweets) are categorized and collated into summary themes that emerge and develop as the analysis progresses. This technique allows for large datasets to be summarized in terms of a small number of categories to determine overall patterns and similarities within the initially large number of responses. Here the procedure was adapted to the situation, to enable processing of a large dataset through automatic classification.

The automated classifier was built using regular expressions, a standard computer science tool for matching text against search patterns. A set of regular expressions (classification rules), was generated for each category to be discovered (these are documented in the Internet code repository). In some cases, category rules were supplemented with an inverse matching set. Each classifier was independent and binary (yes/no), meaning that a tweet might be classified for inclusion in multiple categories. The categories were defined manually as the model was developed. For the first iteration, we randomly sampled 100 tweets from the full dataset and used grounded theory techniques to establish the first candidate categories and corresponding rules, which were refined later on and throughout the classifier development. The development of rules for the classifier progressed in an iterative loop consisting of four steps:

1) Random sampling of categorized and uncategorized data
2) Manual analysis of the samples
3) Rule creation, modification, and removal
4) Application of the new rule set

In the *sampling* phase, we randomly selected 20 tweets from each classified category and 50 tweets from the unclassified tweets from the dataset. Next, we manually analyzed this subset of data to discover opportunities to create, modify, or remove categories and observe any modifications needed to be made to the classifier. After this, the updated classification rules were applied to the original dataset by re-running the analysis script to generate a new classification. Having performed this, the loop would restart and the newly classified data was randomly sampled again.

The classification script informed us about the number of tweets classified after each run so we could observe the changes in distribution of classified categories and the residual, unclassified tweets. Overall, twelve major revisions of the rules were required to develop the classifier used to obtain the results we report.

There are currently no established procedures for validating text classifiers for data such as tweets. Approaches in previous research vary (e.g., Genc, Sakamoto, & Nickerson, 2011; Nishida, Banno, Fujimura, & Hoshide, 2011), thus our approach was to use two independent coders to check random samples of classified and unclassified tweets for false positive (classified), false negatives (unclassified), and undetected categories (unclassified). Using a script, we randomly picked 50 samples from each classified category ($N = 8$) and 200 unclassified tweets were randomly selected for assessment after the classifier was considered finalized. As such, we manually checked $8*50 + 200 = 600$ tweets. This provided us 2% unit accuracy estimates within classified and tweets and 0.5% among unclassified. The inter-rater reliability in coding was high; the initial assessed kappa was .97 for the categorized items (yes or no coding) and .78 for the unclassified items (nine options for coding). After discussing all false instances and possible discrepancies in coding, the differences were resolved into a unanimous consent. This procedure was also an adaptation of the procedure previously used with a grounded theory approach (Williamson et al., 2014).

Sentiment analysis. Sentiment analysis, which is also known as opinion mining, refers to a variety of natural language processing methods used to determine the attitude of a speaker or a writer (Pang & Lee, 2008). Sentiment analysis (SA) methods have become widely applied as they can provide valuable information about given products or services by automatically analyzing large bodies of text (Pang & Lee, 2008). The methods used to perform SA utilize various machine learning and data mining approaches to evaluate the affective state of the text. Often, SA attempts to classify a given text according to its valence, i.e., to tell whether the text is positive, negative, or neutral. Advanced SA tools (see, e.g., Strapparava & Mihalcea, 2008) attempt to extract more information about the emotional content of the text, for instance, to classify texts as angry, sad, and happy.

Traditionally, SA has proven much more robust for longer texts, and classifying polarity of short texts, such as tweets, has proven challenging. Previous attempts to analyze sentiment of short texts (e.g., Pak & Paroubek, 2010; Thelwall, Buckley, & Paltoglou, 2011), have made use of a wide variety of different features including n-grams, part-of-speech tags, and valence values of sentiment lexicons. The sentiment lexicons are special purpose dictionaries that contain positivity and negativity values for words, and they are usually manually crafted. We decided to use a simple but relatively robust bag-of-words approach to perform basic SA on our data. This approach means that every tweet is treated as a collection of words and the sentence structure is not taken into account. The idea is to use a hand-crafted sentiment lexicon that contains positivity and negativity values for a large number of English words to analyze the tweets word-by-word for total positivity and negativity by summing up affective values of the individual words. We refer to these sums of positivity and negativity values as the positive and negative sentiment score, respectively. Because of the independent assessment, the analyzed text can potentially have both high negativity and high positivity. To simplify reporting, calculated *sentiment balance* values by subtracting the negativity value from the respective positivity value.

Rudimentary SA was performed on the data following the preprocessing phase. The analysis was performed with a custom built script (the Internet code repository) that used SentiWordNet 3.0 (Esuli & Sebastiani, 2006) to assess positivity and negativity of individual tweets. SentiWordNet is a widely used lexical resource that provides positivity and negativity values for a relatively large set of English words (currently around 117 000 words). In order to determine whether the SA results reflected emotions specifically associated with earworm discussion or whether these results simply mirrored the emotional content of Twitter discourse in general, we collected two reference data sets by sampling tweets in English 1) without any search criteria and 2) using "song" as search term. Thus, in the second case the reference data was related to music but was not specifically related to earworms. Both reference datasets were sampled twice in February 2014 with a week in between. Each time the sampling took place in one day, collating at least 20,000 tweets at a time. Identical SA was performed on this data to establish norms for emotionality among these tweets that were not sampled by INMI related keywords. The reference data samples were acquired because Twitter currently prohibits third parties from publishing tweet corpora, such as the past Edinburgh Twitter corpus (Petrovic, Osborne, & Lavrenko, 2010). We statistically tested for differences in positivity and negativity values between the earworm data and the reference data by using the nonparametric Wilcoxon rank sum test.

Results

We present our findings in two parts: first describing the data on an aggregate level using descriptive statistics, and then providing insights from the automatic classification and sentiment analysis.

TABLE 1. *Descriptive statistics about contributing users and the data set*

	N	%
Users		
Number of contributing users	56 626	
Number of locations reported	23 580	
Decoded locations	17 789	75.6%
Number of unique countries	173	
Tweets		
Number of tweets	80 620	
Number of retweets	3 074	
Number of hashtags	39 752	
Number of URL hyperlinks	16 422	
Number of unique hashtags	11 933	30.0%
Number of unique URLs	11 669	71.1%

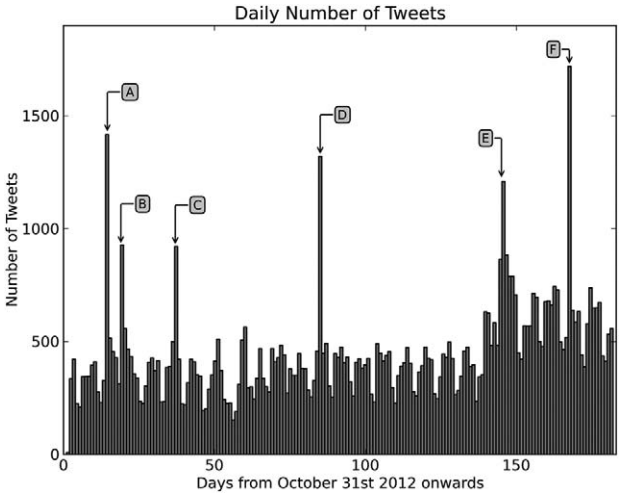

FIGURE 1. The distribution of tweets across the data collection period of 183 days.

USERS

Our final dataset included 80,620 tweets from 56,626 unique usernames (see Table 1 for more details). In relation to the total volume of Twitter traffic, earworm discussions represent a miniscule fraction; less than one in every 100,000 tweets was included in our data, assuming average 58 million Tweets a day in 2013.[4]

We used the location data to infer the geographic origin of the tweets. 66,477 of the tweets in our dataset contained information about the user's location. Out of 23,530 unique locations, we resolved 17,789 (75.6%), representing 173 countries and other geographic locations. The countries with the most contributions were the US (32.6%), UK (26.7%), Canada (7.0%), Australia (4.7%), and India (2.3%). These countries contributed over 1,500 tweets each and cumulatively accounted for 86.6% of all the tweets with geographic information. There were single contributions from many ($N = 27$) small countries such as Vatican City, Laos, Burma, and Aruba (see Appendix 1 for a complete list).

We examined the temporal accumulation of data over the collection period. The accumulation of tweets was quite stable throughout the sampling period, as illustrated by Figure 1. The accumulation averaged to 433 tweets a day, with a slight increase towards the end of the data collection period (12,458 tweets in November 2012 vs. 16,978 tweets in April 2013). On average, more tweets were posted on weekdays than during the weekend. Tuesday was the busiest day ($M = 498$ tweets) in contrast to weekdays ($M = 478$ tweets Monday - Friday). The weekends saw almost 1/3 fewer tweets, with Saturday and Sunday averaging to 319 tweets per day.

[4] http://www.statisticbrain.com/twitter-statistics/ accessed February 18, 2014

These patterns match previous reports of temporal patterns for tweeting in general (e.g., TrackMaven, 2013).

From the daily frequency distribution seen in Figure 1 we can observe some peaks. These occurred when a tweet was actively retweeted (tweets A-D, F) or a piece of information captured the attention of multiple users at same time (tweet E). Looking at the four days, in which over 900 tweets were captured, we have two instances of accumulating retweets (A and D), and one instance of a news item (E) apparently catching the attention of several Twitter users independently without retweets. The content of these popular tweets are listed below:

A) 11th of November 2012, B) 18th of November 2012, and D) 23rd of January 2013

From **@WhatTheFFacts**, follower count at the time: over 4 million

```
Tweet 1.
Sometimes you might think you're hearing
a piece of music in your head even though it's
not being played. This is known as the
"earworm".
```

C) 6th of December 2012

From **@VEVO**, follower count: over 498,000

```
Tweet 2.
Are we sure that category wasn't for Biggest
Earworms? #RecordOfTheYear #GRAMMYNoms
```

E) 25th of March 2013

From multiple users shared via **Metro.co.uk** website

Tweet 3.
Lady Gaga tune stuck in your head? A tricky anagram will get rid of those…: But try anything too difficult…http://metro.co.uk/2013/03/25/lady-gaga-tune-stuck-in-your-head-a-tricky-anagram-will-get-rid-of-those-annoying-earworms-3557974/

F) 16th of April 2013, From **@manuel_c** (follower count: over 300 000)

Tweet 4.
Stumbled across @KatDahlia - reminds me of a mix of @NellyFurtado & @MIAuniverse. Am hoping to hear a lot more of her! #music #earworm #jam

TWEETS

Although our keyword-based search for earworms did produce a wealth of results, it is evident that the word "earworm" is not yet fully established in the vocabulary of Internet users. About one in ten (11.7%) wrote earworm as "ear worm" or "ear-worm." Similar variation is visible among the hashtags found in the data. Looking into the 11,933 unique hashtags present in the data (total $N = 18{,}804$ hashtags), only 154 hashtags were used at least ten times. Among these, we observed 26 that were variations of the word "earworm" (e.g., #earworms!, #EarWorm, #earworms) and these tended to be the most popular. We can thus estimate that almost half of *all hashtags* embedded in the data were directly referring to earworms. The five next most used hashtags (after the earworm-related tags) were #RecordOfTheYear, #GRAMMYNoms, #tune, #nowplaying, and #music (from 546 to 250 instances each, respectively). The most frequently used artist-related hashtag was that of #stompintom ($N = 58$), ahead of #Bowie ($N = 18$).

The hyperlinks embedded in the tweets were unique more often than the hashtags. 87% of all hyperlinks were posted only once and 75 URLs were posted ten times or more (see Table 1). The most frequently shared references were informative articles in online magazines or Wikipedia about earworms. The most popular music link was the song *Dumb Ways To Die*, a public service announcement created by Metro Trains in Melbourne and published on YouTube. Of all posted links, there were 7,123 unique links to YouTube, making it the most frequently linked source (total $N = 8532$, 73.1% of all links). A list of the most frequently linked music content on YouTube is included in Appendix 2 and as a YouTube playlist[5] (https://www.youtube.com/playlist?list=PLFF185AyoHpojr7fZB8d0GwsOIyZAXjvG).

Retweets were fairly uncommon in the dataset. Only two tweets obtained over 100 retweets. The tweet from @WhatTheFFacst quoted above (tweet A) had the highest retweet count at 1,241, whereas the second most retweeted message from @VEVO (tweet C) had 368 retweets.

AUTOMATIC TWEET CATEGORIZATION

We developed a classification system comprised of eight categories. The categories are labeled, described, and illustrated below in alphabetical order.

A. Causal attribution. User comments about receiving or transmitting earworms to somebody, intentionally or unintentionally. This expresses beliefs about causality of INMI, in terms of its transmission either to or from the user.

Tweet 5.
Nate gave me this earworm, simply by saying the title. Have to listen to kill it. (oh yeah, you are welcome: D) http://open.spotify.com/album/6awK1ZiNXG9NOwPbAGWE5w
By **@adenpenn** on Wed Mar 06 01:57:20 2013

B. Cure and riddance. User expresses the need or attempts to eradicate earworms, or offers advice on coping with them.

Tweet 6.
RT @ThatEricAlper: Hey, I just met you, and this is crazy, but researchers figure out how to get rid of earworms, so read it maybe? http…
By **@noblemusicnyc** on Thu Jan 03 21:08:10 2013

C. Earworm nominations. User expresses the belief that a defined song has the potential to cause an earworm experience. This nomination does not indicate of whether the user actually experiences the earworm.

Tweet 7.
Song I tweeted last night, it's a beautiful earworm with heartfelt lyrics. No Man - All that you are http://www.youtube.com/watch?-v=i16CkdJ7lO8&feature=youtu.be
By **@davesleney** on Fri Nov 02 09:24:06 2012

D. Earworms information. User provides general information about earworms, their study, or asks generally for information about earworms.

Tweet 8.
An ''Earworm'' describes a song, hook and or chorus that you can't get out of your head.
By **@always_known** on Mon Apr 01 20:32:26 2013

[5] https://www.youtube.com/playlist?list=PLFF185AyoHpojr7fZB8d0GwsOIyZAXjvG

E. Grateful attribution. A variety of causal attribution in a particularly positive tone; the user expresses thankfulness (sincere or ironic) towards an identified party for receiving an earworm.

> Tweet 9.
> Thanks @lolaredgal for giving me the earworm "Call Me Maybe." I'll need to listen to two hours of "Last Christmas" to get it out.
> By **@jonathanmaze** on Wed Dec 19 20:26:21 2012

F. Music links. User posts a link to an internet music or video service in association with an earworm

> Tweet 10.
> Off to Greece today - resulting earworm: "She came from Greece she had a thirst for knowledge": http://www.youtube.com/watch?v=yuTMWgOduFM
> By **@sunnydeveloper** on Thu Feb 28 20:25:38 2013

G. My Earworms. User reports experiencing a current, habitual, or recent personal earworm. The song or the artist may or may not be identified, essential is the actual earworm experience.

> Tweet 11.
> I have Some Nights stuck in my head but I only know the tune to the chorus so I've been singing it as a weird voodoo African chant instead
> By **@s0dalite** on Sat Mar 30 17:30:29 2013

H. Viral worms. This is purely quantitative category referring to popular tweets or retweets about earworms that appear multiple times in an identical or nearly identical form.

> Tweet 12.
> Get that tune out of your head - scientists find how to get rid of earworms via @Telegraph http://www.telegraph.co.uk/science/science-news/9950143/Get-that-tune-out-of-your-head-scientists-find-how-to-get-rid-of-earworms.html
> By **@Live4ever_Ezine** on Sun Mar 24 21:40:22 2013

The tweets that could not be automatically classified were called Unclassified. We provide two examples.

> Tweet 13.
> Keep me stuck inside your head like your favorite tune
> By **@_bezzyy** on Tue Mar 19 00:12:51 2013

> Tweet 14.
> When people ask my Halloween costume I sing the "by mennen" jingle. (my costume is an earworm.)
> By **@SoBrianRowe** on Thu Nov 01 02:13:09 2012

CATEGORIZATION AND VALIDATION RESULTS

The automatic classifier was able to classify 51% of the tweets ($N = 80{,}620$). The results of the classification are displayed in Table 2. It should be kept in mind that categories were not discreet, so a tweet could be assigned multiple category codes. As a result, the total of number of assigned category codes was 51,934 in contrast to the 44,214 unique tweets coded. Correspondingly, 84% of all categorized tweets had only a single code.

The dominant category of tweets was "Category G. My earworms" (28% of all tweets), which consisted of people reporting their INMI experiences. This was sometimes (19% of instances) complemented with a hyperlinked music resource that further identified the song in question. As we can see from the high proportion of user references, these tweets were often parts of a discussion or directed to a particular user as discussion starters or replies. The music link tweets (Category F) were the next most popular category (11% of all). These tweets always linked to an Internet music source, but mentioned other Twitter users infrequently (36.4% included user mentions).

Tweets conveying general information about earworms or their study were also common (Category D. Earworms information). They nearly always referenced a user (92%), but rarely included links to Internet content (9%). The fourth largest category was "Category E. Grateful attribution." Interestingly, users hardly ever thanked another Twitter user for an earworm by mentioning them (5.8%) and the use of hyperlinks was similarly infrequent (10.3%), making these statements exceptionally isolated in terms of social media.

Less than 4% of tweets were classified as "Category H. Viral worms." These tweets were also often coded in the categories "Category B. Cure and riddance" (23% of instances) and "Category D. Earworms information" (11% of instances). Because their definition included the retweet criterion, the proportion of included user references was very high (83%). "Cure and riddance" tweets were almost equally as frequent as "Viral worms" (3.2%). They involved a typical rate of user references, but more hyperlinks than the average (51.3% vs. 46.8%), primarily due to users sharing article links related to dispelling earworms (see Tweet 3).

"Category C. Earworm nominations" revealed a novel way of using the word earworm to describe a song (see Tweet 7). These types of tweets were uncommon (1.1% of the data) and they had a high user mention rate (60.6%), making them similar to the "Category G. My earworms" category. In some of the tweets reviewed during the manual analysis, users' language implied that

TABLE 2. *The results of automatic classification in terms of discovered category instances, user mentions, included hyperlinks, classification accuracy within the category (false positive) and among unclassified tweets (false negatives).*

Category	Instances	%	User mentions	Included hyperlinks	False positives	False negatives
Unclassified	39 248	48.7%	53.9%	11.6%	–	–
G. My Earworms	22 531	28.0%	44.7%	18.8%	12%	11.5%
F. Music links	9 099	11.3%	36.4%	99.9%	0%	0.0%
D. Earworms information	5 187	6.4%	91.6%	8.8%	2%	3.5%
E. Grateful attribution	4 711	5.8%	5.8%	10.3%	6%	0.0%
H. Viral worms	2 700	3.4%	82.6%	34.7%	2%	0.0%
B. Cure and riddance	2 610	3.2%	51.3%	46.5%	2%	1.5%
C. Earworm nominations	874	1.1%	60.6%	10.1%	20%	2.0%
A. Causal attributions	752	0.9%	1.3%	20.7%	4%	3.5%
Classified M			46.8%	31.2%	6%	3%

they might have also been experiencing an earworm, but this could not be ascertained.

The smallest category was "Category A. Causal attribution." It also included a novel use of the term earworm as a verb, e.g., "be sure to earworm somebody today." Although our category definition demanded that the tweet must acknowledge the causal relationship (incoming or outgoing) clearly, users hardly ever attributed another Twitter users by mentioning them (1.3%). Instead many pointed out songs through hyperlinks as sources of their (or someone else's future) earworms.

Finally, following previous research on Twitter (Naaman et al., 2009), we regrouped our categories to reflect the proportion of tweets containing general (informing tweets: categories B, C, & D) and personal (meforming tweets: categories A, E, & G) information, while omitting ambiguous categories that could fall into either classification (categories F and H). We found that the present data were comprised of 77% meforming tweets ($N = 25{,}875$) and 23% ($N = 7{,}752$) informing tweets.

FINDINGS FROM MANUAL VALIDATION

Through manual validation we evaluated how well the classifier performed by hand coding random samples of the data. The accuracy across all categories in this inspection was 94% (6% of codes were false positives; Table 2). The weighted accuracy of the classifier, which takes into consideration the differences in category size, was 93%. The proportion of false negatives among the unclassified tweets ($N = 39{,}248$) was higher. Overall we discovered that 21% of the unclassified tweets should have been included in some of the categories (A-H). From Table 2 we can see that the distribution of false negatives was similar to the overall distribution, however "Music links," "Viral worms," and "Grateful attribution" had a zero false negative rate. Furthermore, some 8.5% of the unclassified tweets (4.1% of all) were interpreted as invalid for the data; that is, it was clear or strongly suspected that they were unrelated to the INMI phenomenon.

Another discovery made in the manual validation was the emergence of a new category called *Song references*. Many people had the habit of tweeting some lyrics or a song title preceded or followed by the word earworm, often as a hashtag (e.g., "Mamma mia, here I go again! #earworm"). These tweets were quite frequent, representing approximately 18.5% of the unclassified tweets (9.0% of total sample). These tweets were distinct from the discussions in "My Earworms" or "Earworm nominations" categories because they lacked the necessary information to understand whether an actual INMI experience or dispositional properties of the music were being discussed.

In the final stage, we looked for any missed emergent themes by coding a random sample of 200 unclassified tweets following ground theory practice. We calculated that 70.5% of these unclassified tweets had various nonrepetitive themes, but these did not display evident linguistic regularities and were not numerous enough (only 1 to 13 tweets each, under 10% of the sample) to warrant their own category or theme. These residual categories included such content as appraising earworms as good or bad, making an exclamation with little other information provided (e.g., "earworm!!"), or talking about music more generally.

SENTIMENT ANALYSIS

We performed SA on both the earworm tweets and the reference data. This was done in order to compare the emotionality of earworm discussions to the emotional tone of Twitter discussions in general (unfiltered reference 1 and 2 datasets), as well as those specifically based

TABLE 3. *Results of the sentiment analysis for earworms data and four reference data samples (Unfiltered reference 1 & 2, Music reference 1 & 2).*

Dataset	N	Positivity M	Positivity SD	Negativity M	Negativity SD	Balance M	Balance SD
Earworms	80 642	0.0595	0.180	0.0681	0.187	−.00428	0.1940
Unfiltered 1	30 861	0.0414	0.155	0.0383	0.148	.00154	0.1570
Unfiltered 2	24 773	0.0315	0.137	0.0291	0.130	.00118	0.1370
Music 1	28 093	0.0393	0.151	0.0399	0.150	−.00029	0.1559
Music 2	22 327	0.0324	0.137	0.0284	0.128	.00202	0.1361
Reference data combined	106 054	0.0366	0.147	0.0344	0.141	.00107	0.1479

around music but not related directly to earworms (music reference 1 and 2 datasets). The results showed that both the positivity and negativity values of the earworm tweets were higher for earworm tweets than both sets of comparison data. Normalized SA for the tweets showed that the earworm tweets were generally higher in emotionality than the comparison data, and this effect was confirmed by a Wilcoxon rank sum test ($p < .01$). The results are displayed in Table 3.

The results show that the average emotional balance (based on the sum of positivity and negativity of each tweet) of the earworm data was negative whereas in the reference data, the balance was positive. Of all four reference data samples, only one of the reference datasets (Music reference 1) also displayed a negative balance, however its balance score was less than one tenth of the value of the earworm data's balance (see the balance illustration in Figure 2).

Within the earworms dataset, we compared the sentiment scores of uncategorized and categorized tweets. This revealed that uncategorized tweets had on average a 0.018 higher positivity and 0.016 higher negativity scores (respective Cohen's $d = 0.101$ and 0.088). We manually reviewed the results of the SA, discovering that many of the tweets receiving high scores were in fact from the emergent category "Song reference" that had not been automatically categorized. Examples of highly emotional tweets, positive (15) and negative (16), below:

Tweet 15.
@Nicholosophy best Twitter earworm so far. Well done Sir, well done.
By **@mindyhoyden** on Fri Nov 02 01:01:34 2012

Tweet 16.
I had thought that "Bela Legosi's Dead" had managed to kill the "When the Levee Breaks" earworm I've been stuck with --- but I was wrong.
By **@cthulhim** on Thu Apr 11 19:41:40 2013

A final observation concerns the distribution of sentiment scores across tweets conveying personal versus informational content. Meforming tweets (Naaman et al., 2009) were higher in both positivity and negativity in contrast to informing tweets ($p < .001$). The difference was especially strong for positivity (Cohen's $d = 0.159$) but also considerable for negativity (Cohen's $d = 0.083$). However, even though the personal tweets had higher emotionality scores, the difference between positivity and negativity (emotionality balance) was more negative among the informational tweets (balance $M = -0.0213$ and $M = -0.0086$; informers and meformers respectively; $p < .001$). This emphasizes the polarization of sentiment in meforming tweets.

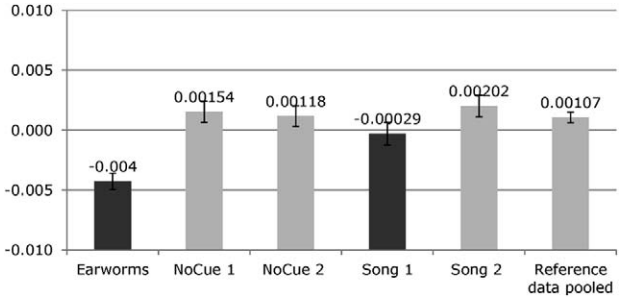

FIGURE 2. Average emotional balance of earworm tweets and two reference tweet sets (Music and Unfiltered tweets). The error bars indicate the standard error of the mean.

Discussion

We have reported a study of involuntary musical imagery using a big data approach, which has previously been uncommon in music psychology. We used this method to study discussions of INMI via the social media service Twitter. We sought answers to three research questions pertinent to INMI and established some new findings in return. We uncovered evidence that the earworm experience is a widespread psychological phenomenon reported in locations throughout the globe. We found that users openly discuss the types of music that they experience as earworms and potential causes and cures for these via their Twitter network. Finally, we discovered that people discuss INMI in more

negative emotional terms on Twitter than other topics, including music in general. The present work has expanded the study of INMI using a novel methodological approach, and demonstrated one way in which empirical musicology can appropriate freely accessible big data sources for research purposes.

In response to our research questions, we found evidence of INMI's ubiquitous character. In our data, INMI tweets were contributed from 173 geographic locations around the world. 32% of the tweets included in our sample originated from the US. However, this figure is close to the 23% share of account by US users reported by Twitter. The bias in our sample towards countries such as the US and UK is likely explained by the fact that we only included English tweets in our data. Despite the limitations imposed by linguistic constraints, we were still able to include a large amount of geographically diverse data, thus present the first large-scale multi-cultural study of INMI. Our findings thus support the idea that INMI is an integral part of music cognition among humans.

The results of our classification procedure show that Twitter users frequently share their earworm experiences and related music. The "My earworm" reports were the most frequent type of categorized message. Among uncategorized tweets, the song references were numerous. This indicates, firstly, that many Twitter users have taken up the use of earworm as a common word to pinpoint a certain conscious experience and are willing to share their experiences with the public. These data give credence to the previous studies of the commonality of INMI (Bailes, 2007; Liikkanen, 2012a). We also observed that the temporal pattern of earworm tweets follows the general rhythm of tweeting (TrackMaven, 2013), such that tweeting is more frequent over the weekdays than over the weekends.

Second, the current results also partially replicate those of Williamson et al. (2014), who found that people employ both active and passive strategies as a means of coping with the INMI experience. The present data showed that patterns such as active coping by engaging with an earworm tune, and passive coping by allowing a tune to fade out, are also reported by Twitter users. We confirmed that people engage in various INMI coping strategies, express their desires to expel their INMI, and seek out INMI cures via their Twitter network. Although our exploration did not provide as detailed a picture of INMI coping behavior as the aforementioned paper, it reflects the existence of these behavioral patterns in a more heterogeneous and geographically distributed sample than demonstrated by the previous work. There also emerged a new type of INMI behavior: a public display of annoyance over INMI in social networks.

The third notable finding from the classification of tweets concerns the intuitive acknowledgement of the dynamic and even viral nature of INMI. This was expressed in the belief that INMI can be passed on from one person to another. This has been only touched upon in previous studies (Liikkanen, 2012b; Williamson et al., 2014), but was demonstrated in the present study in over 5,000 tweets classified under Causal or Grateful attribution. These show that many people intuitively understand the role of triggers (Williamson et al., 2012) in the onset of INMI and can thus appreciate the causal relationship in receiving or sending earworms to others, for instance, via social networks.

We also found that the personal tweets (meforming) were more emotional and on average more negative than the tweets containing general information (informing). The research on INMI has thus far concluded (e.g., Williamson et al., 2014) that the majority of INMI experiences are not bothersome, but the current results indicate that there is a bias towards reporting highly emotional experiences on Twitter — especially in a negative tone. The finding of the negative tone of earworms tweets goes somewhat against the previous literature (Beaman & Williams, 2010; Halpern & Bartlett, 2011; Hemming, 2009; Liikkanen, 2012a), which has repeatedly demonstrated that people generally evaluate their INMI experiences as neutral or positive.

This apparent discrepancy is likely due to the method of data collection. Because our data consists of voluntary or spontaneous reports of INMI, rather than input from participants in a scientific experiment, we were able to observe discussions without any experimental bias. This suggests that people are more likely to complain about the annoying INMI experiences rather than celebrate the good ones in public, even though the latter might be more representative of what goes on in their heads. Based on the high negative and positive sentiment scores among the personal tweets, both instances occur. Additional research is needed to look into this issue and distinguish whether the earworm songs (as cited by Twitter users) or the descriptions of INMI experiences have a negative wording.

The nature of INMI discussion on Twitter conforms very much to the overall pattern of Twitter discussions. First, more earworm tweets were sent on weekdays than during weekends, which is also true of tweeting in general (Gao, Abel, Houben, & Yu, 2012). The previously observed distinction of prominently talking about oneself and infrequently spreading impersonal information (or meforming and informing; Naaman et al., 2009)

were present in the earworm tweets as well. Although the balance was similar, we observed more informational tweets (our data was comprised of 35% informers vs. Naaman's 20%). Our finding that the earworm tweets rarely reach wider audiences is also in line with the findings about opinion leadership (Wu et al., 2011); out of 50,000 contributing users, only a handful of users received retweets by numerous others.

Different kinds of networks are often discussed in relation to studies of social networking services (e.g., Kossinets & Watts, 2006). In this regard, earworm discussions reveal a quite sparse, not very tightly connected network. Many of the users contributing to the data were not engaging in conversations, but rather just providing updates about their lives or the daily news. Without an elaborate network analysis, we can tell by the relatively small number of retweets and user references that earworm discussions are somewhat lonesome cries in the micro blogosphere (i.e., Twitter), with infrequent echoes from other users.

LIMITATIONS

In terms of big data, one can question whether the dataset of 80,620 tweets from 56,626 unique users used in the present study warrants the label. This is a justified question if we only observe sample sizes, as some meta-analyses with traditionally collected data have involved data from over 50,000 people (for instance, Jokela et al., 2013 analyzed 76,150 cases). However, there are important methodological differences. The first is that the presented methodology can be easily scaled up to work on datasets up to hundreds of times larger without running into computational limitations. The data we have gathered — largely independent, spontaneous verbal reports of INMI — would have been impossible to come by without a social network service such as Twitter. As we estimated how rare the INMI-relevant tweets were (1/1000 of 1% of the total traffic on Twitter), it seems difficult to come up with such a dataset through any other means than deriving it from a big data source. Second, we have drawn our observations from publicly available data, which was not generated for research purposes. While the data processing required much effort to control various sources of noise, the results are all the more interesting for being naturalistic observations even if sampled from a distinct, public social media.

Another methodological concern is that the reference data for the sentiment analysis was collected only over a period of a couple of days whereas the earworm data was collected over a period of six months. This may have possibly affected the results. An alternative approach for validating the results of sentiment analysis would be to use a known reference dataset, if available. Future research should look into collaborating with partners who hold such data.

Finally, the selection of search terms may have biased our data. The term earworm, which we used as a main search keyword, has entered English vocabulary and dictionaries quite recently. Its inclusion may mean that more people who have been "early adopters" of that term have been included in the data, with unknown consequences. Careful screening of other potential keywords, such as "stuck song," should be carried out in future studies to optimize the amount of relevant data collected. This is important planning as Twitter data cannot be re-sampled in retrospect.

FUTURE WORK

The present study has evoked ideas about big data methods, Twitter data, and INMI. For instance, there were many indications that earworms spread via Twitter, from shared music references and links. This could be further explored. For instance, one might investigate where people derive their musical influence and the triggers for INMI. This could be analyzed starting from tweets categorized for "My earworms" or the two attribution categories and followed up to user mentions to include Twitter discussion beyond the sample we have explored (e.g., earworms reported after attending to a music-related Tweet not including any earworm reference). One might also further explore the large sample of songs named in the "Earworm nominations" and "My earworm" categories in an attempt to discover any underlying musical (or extramusical) features that might cause these songs to be regarded as particularly "earwormy" by the general public.

The collected data are considerable and have not been fully exploited. We are committed to making it available for interested researchers for further study. We believe it contains useful information about many features of INMI. For instance, following recent ideas of "cure tunes"— specific tunes used to expel earworms without becoming stuck themselves (Williamson et al., 2014) — the plethora of musical references and tweets categorized as "Cure and riddance" suggests there is much potential to look further into the subject of specific earworm-displacing tunes. We were also quite drastic in our decisions to discard some data, particularly thousands of non-English tweets. Taking a closer look into discussions by users in different languages might provide new insights and corroborate the evidence on the universality of INMI phenomenon. Naturally, this study should sample data with appropriate keywords, not only English ones.

The methods we used can certainly be improved upon. Although we achieved very good results when classifying the appropriate tweets, the number of tweets incorrectly left unclassified could be further minimized. The algorithmic analysis could be improved by attempting to enable it to recognize fragments of lyrics and song titles, which make up a great part of the unclassified tweets, but might belong to the "My earworms" category – or at least to the "Song reference" category. However, this task is nontrivial because song titles and lyrics have no regularity on a lexical or grammatical level, and the number of instances to match against (all English song titles and artist names) is considerable, especially when Twitter users do not necessarily spell the names or recall the lyrics correctly.

Different types of machine learning approaches usually used in analyzing big data should be considered in the future now that we have annotated data available. Their challenge is the lack of interpretative aid that human assisted heuristic data analysis such as that applied here can offer. Machine learning typically functions as a black box and provides results that can be hard to interpret, particularly from a qualitative point of view. We believe that our tentative categorization of what is out there and annotated data can help others to utilize machine learning approaches to study similar datasets and make meaningful discoveries using even greater sample sizes.

CONCLUSIONS

The earworm data we collected over a period of six months shows that many little streams make up a big river. Considering that earworms represent a private phenomenon not prevalent in everyday discussions – even if prevalent in everyday consciousness – it was not at all guaranteed that this approach would yield meaningful data in adequate quantities. However, a substantial quantity data was successfully sampled and categorized so that insights about INMI could be derived.

Explorative work holds many surprises. We were surprised by both the negativity of discussions surrounding earworms, as well as the fact that we discovered as much relevant data about INMI as we did. There is an interesting trend in modern day society of bringing out mental experiences from the private consciousness to be shared with an ad hoc social circle of fellow Twitter users. This seems to be inherent to the nature of social networking, which as such opens up opportunities for a new kind of research in music psychology as well.

Author note

We thank Helsinki Institute for Information Technology HIIT at Aalto University and University of Helsinki for providing infrastructural support for the research as well as Taneli Vähäkangas for technical support in data acquisition. We also thank Professor Mari Tervaniemi, our action editors, and the anonymous reviewers for the insights provided during the publication process.

Correspondence concerning this article should be addressed to Lassi A. Liikkanen, Institute of Behavioral Sciences, University of Helsinki, Finland, P.O. Box 9. FI-00014 University of Helsinki. E-mail lassi.liikkanen @helsinki.fi

References

Bailes, F. (2006). The use of experience-sampling methods to monitor musical imagery in everyday life. *Musicae Scientiae, 10,* 173-190.

Bailes, F. (2007). The prevalence and nature of imagined music in the everyday lives of musical students. *Psychology of Music, 35*, 1-16.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509-512. doi: 10.1126/science.286.5439.509

Baruss, I., & Wammes, M. (2009). Characteristics of spontaneous musical imagery. *Journal Of Consciousness Studies, 16*(1), 37-61.

Bauckhage, C. (2011, July). *Insights into internet memes.* Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media ICWSM-11, Barcelona, Spain.

Beaman, C. P., & Williams, T. I. (2010). Earworms ('stuck song syndrome'): Towards a natural history of intrusive thoughts. *British Journal of Psychology, 101*, 637–653. doi: 10.1348/000712609X479636

Beaman, C. P., & Williams, T. I. (2013). Individual differences in mental control predict involuntary musical imagery. *Musicae Scientiae 17,* 398-409. doi: 10.1177/1029864913492530

Beaty, R. E., Burgin, C. J., Nusbaum, E. C., Kwapil, T. R., Hodges, D. A., & Silvia, P. J. (2013). Music to the inner ears: Exploring individual differences in musical imagery. *Consciousness and Cognition, 22*(4), 1163-1173. doi: http://dx.doi.org/10.1016/j.concog.2013.07.006

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*, 295-298.

Byron, T. P., & Fowles, L. C. (2015). Repetition and recency increases involuntary musical imagery of previously unfamiliar songs. *Psychology of Music 43,* 375-389. doi: 10.1177/0305735613511506

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis.* London: Sage.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13*(1), 3-21.

De Choudhury, M., Counts, S., & Horvitz, E. (2013, May). *Predicting postpartum changes in emotion and behavior via social media.* Paper presented at the the 2013 ACM Annual Conference on Human Factors in Computing Systems (CHI'13), Paris, France.

Esuli, A., & Sebastiani, F. (2006, May). *Sentiwordnet: A publicly available lexical resource for opinion mining.* Paper presented at the5th Conference on Language Resources and Evaluation.

Gao, Q., Abel, F., Houben, G.-J., & Yu, Y. (2012). A comparative study of users' microblogging behavior on Sina Weibo and Twitter. *Lecture Notes Computer Science, 7379*, 88-101. doi: 10.1007/978-3-642-31454-4_8.

Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). Discovering context: Classifying tweets through a semantic transform based on Wikipedia. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of augmented cognition. Directing the future of adaptive systems* (pp. 484-492). Berlin: Springer.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences, 107*, 17486-17490.

Halpern, A. R., & Bartlett, J. C. (2011). The persistence of musical memories: A descriptive study of earworms. *Music Perception, 28*, 425-432.

Hemming, J. (2009). Zur phänomenologie des "ohrwurms" [the phenomenology of "earworms"]. In W. Auhagen, C. Bullerjahn, & H. Höge (Eds.), *Musikpsychologie - musikalisches gedächtnis und musikalisches lernen* (Vol. 20, pp. 184-207). Göttingen: Hogrefe.

Hemming, J., & Altenmüller, E. (2012, July). *When an everyday-phenomenon becomes clinical: The case of long-term earworms.* Paper presented at the 12th International Conference on Music Perception and Cognition (ICMPC) and 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, Thessaloniki, Greece.

Huron, D. (2013). On the virtuous and the vexatious in an age of big data. *Music Perception, 31*, 4-9.

Hyman, I. E., Burland, N. K., Duskin, H. M., Cook, M. C., Roy, C. M., McGrath, J. C., & Roundhill, R. F. (2013). Going gaga: Investigating, creating, and manipulating the song stuck in my head. *Applied Cognitive Psychology, 27,* 204-215. doi: 10.1002/acp.2897

Jokela, M., Batty, G. D., Nyberg, S. T., Virtanen, M., Nabi, H., Singh-Manoux, A., & Kivimäki, M. (2013). Personality and all-cause mortality: Individual-participant meta-analysis of 3,947 deaths in 76,150 adults. *American Journal of Epidemiology, 178*, 667-675.

Kim, Y., Suh, B., & Lee, K. (2014, July). *#Nowplaying the future billboard: Mining music listening behaviors of Twitter users for hit song prediction*. Paper presented at the Proceedings of the First International Workshop on Social Media Retrieval and Analysis, Gold Coast, Queensland, Australia.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*, 5802-5805. doi: 10.1073/pnas.1218772110

Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science, 311*, 88-90.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). *What is Twitter, a social network or a news media?* Paper presented at the the 19th International Conference on World Wide Web (WWW'10), Raleigh, NC.

Liikkanen, L. A. (2008, August). *Music in everymind: Commonality of involuntary musical imagery.* Paper presented at the 10th International Conference on Music Perception and Cognition, Sapporo, Japan.

Liikkanen, L. A. (2012a). Musical activities predispose to involuntary musical imagery. *Psychology of Music, 40*, 236-256. doi: http://dx.doi.org/10.1177/0305735611406578

Liikkanen, L. A. (2012b, July). *New directions for understanding involuntary musical imagery.* Paper presented at the12th International Conference on Music Perception and Cognition (ICMPC) and 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, Thessaloniki, Greece.

Liikkanen, L. A. (2012c). Inducing involuntary musical imagery: An experimental study. *Musicae Scientiae, 16*, 217-234. doi: 10.1177/1029864912440770

Liikkanen, L. A. (2012d). Involuntary music among normal population and clinical cases. *Advances in Clinical Neuroscience and Rehabilitation, 12*(4), 12-13.

Liikkanen, L. A., & Raaska, K. (2013). Treatment of anxiety from musical obsessions with a cognitive behaviour therapy tool. *BMJ Case Reports, 8,* 1-4. doi: 10.1136/bcr-2013-201064

Müllensiefen, D., Fry, J., Jones, R., Jilka, S., Stewart, L., & Williamson, V. J. (2014). Individual differences predict patterns in spontaneous involuntary musical imagery. *Music Perception, 31*, 323-338.

Naaman, M., Boase, J., & Lai, C.-H. (2009, February). *Is it really about me?: Message content in social awareness streams.* Paper presented at the the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW'10), Savannah, GA.

Nishida, K., Banno, R., Fujimura, K., & Hoshide, T. (2011, October). *Tweet classification by data compression.* Paper presented at the the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web (DETECT'11), Glasgow, UK.

Pak, A., & Paroubek, P. (2010, May). *Twitter as a corpus for sentiment analysis and opinion mining.* Paper presented at the Conference on Language Resources and Evaluation LREC'10, Malta.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Petrovic, S., Osborne, M., & Lavrenko, V. (2010, June). *The Edinburgh Twitter corpus.* Paper presented at the the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, Los Angeles, California.

Pew Research Center. (2014). *Social media update 2013.* http://www.pewinternet.org/2014/01/08/social-media-update-2013/twitter-users/

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). *Earthquake shakes Twitter users: Real-time event detection by social sensors.* Paper presented at the the 19th International Conference on World Wide Web (WWW'10), Raleigh, NC.

Stonebaker, M. (2012). *What does big data mean?* Web blog entry at blog@CACM. http://cacm.acm.org/blogs/blog-cacm/155468-what-does-big-data-mean/fulltext

Strapparava, C., & Mihalcea, R. (2008, March). *Learning to identify emotions in text.* Paper presented at the the 2008 ACM Symposium on Applied Computing (SAC'08), Ceara, Brazil.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology, 62*(2), 406-418. doi: 10.1002/asi.21462

TrackMaven. (2013, 23 DEC 2013). *The best time to get retweets.* Retrieved from http://trackmaven.com/blog/2013/12/the-best-time-to-get-retweets/

Twitter, Inc. (2013). *Form s-1 registration statement. Registration no. 333-*. San Francisco, CA.

Williamson, V. J., & Jilka, S. R. (2014). Experiencing earworms: An interview study of involuntary musical imagery. *Psychology of Music, 42,* 653-670. doi: 10.1177/0305735613483848

Williamson, V. J., Jilka, S. R., Fry, J., Finkel, S., Müllensiefen, D., & Stewart, L. (2012). How do "earworms" start? Classifying the everyday circumstances of involuntary musical imagery. *Psychology of Music, 40*, 259-284. doi: 10.1177/0305735611418553

Williamson, V. J., Liikkanen, L. A., Jakubowski, K., & Stewart, L. (2014). Sticky tunes: How do people react to involuntary musical imagery? *PLoS ONE, 9*(1), e86170. doi: 10.1371/journal.pone.0086170

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, March-April). *Who says what to whom on Twitter.* Paper presented at the 20th International Conference on World Wide Web (WWW'11), Hyderabad, India.

Appendix 1: List of contributing geographic regions (based on ISO 3166-1 classification)

Location	No. of contribut.
United States	21656
United Kingdom	17717
Unidentified location	8889
Canada	4677
Australia	3140
India	1517
Ireland	762
Philippines	645
South Africa	639
New Zealand	605
Mexico	579
Singapore	564
Indonesia	538
Germany	423
Malaysia	281
The Netherlands	222
Japan	194
United Arab Emirates	192
Brazil	187
France	179
Finland	169
Kenya	108
Spain	107
Nigeria	98
Italy	96
Egypt	94
Sweden	91
Poland	86
Austria	85
Belgium	83
Pakistan	76
Russia	75
Hong Kong	53
Switzerland	53
China	52
Saudi Arabia	51
Iran	46
Norway	45
Thailand	43
Colombia	42
Argentina	41
The Bahamas	40
Cyprus	39
Chile	38
Turkey	38
Taiwan	36
Israel	35
Greece	34
Jamaica	34
Zimbabwe	34
Jersey	33
Peru	33
Uganda	31
Czech Republic	29
Lebanon	29
Denmark	28
Ghana	28
South Korea	26
Venezuela	26
Hungary	24
Nepal	24
Portugal	24
Bangladesh	20
Kuwait	20
Ecuador	19
Puerto Rico	19
Qatar	18
Romania	18
Greenland	16
Serbia	15
Barbados	14
Iceland	14
Uruguay	14
Isle of Man	13
Costa Rica	11
Estonia	11
Namibia	11
Trinidad and Tobago	11
Botswana	10
Iraq	10
Jordan	10
Kazakhstan	10
Panama	10
Slovenia	10
Tanzania	10
Vietnam	10
Honduras	9
Latvia	9
The Gambia	9
Ukraine	9
Bahrain	8
Belize	8
Maldives	8
Dominican Republic	7
El Salvador	7
Morocco	7
Paraguay	7
Solomon Islands	7
Syria	7
Zambia	7
Guernsey	6
Haiti	6
Mauritius	6
Bulgaria	5
Cambodia	5
Cameroon	5
Croatia	5
Mozambique	5
Brunei	4
Cuba	4
Guatemala	4
Kosovo	4
Monaco	4
New Caledonia	4
Sri Lanka	4
Tunisia	4
Turkmenistan	4
Algeria	3
Bermuda	3
Bosnia and Herzegovina	3
Curacao	3
Guam	3
Guyana	3
Libya	3
Madagascar	3
Micronesia	3
Nicaragua	3
Papua New Guinea	3
Slovakia	3
Tuvalu	3
Albania	2
Antigua and Barbuda	2
Azerbaijan	2
Belarus	2
Bhutan	2
Chad	2
Fiji	2
Georgia	2
Grenada	2
Lithuania	2
Luxembourg	2
Malawi	2
North Korea	2
Northern Mariana Islands	2
Saint Lucia	2
Somalia	2
Suriname	2
Tajikistan	2
Afghanistan	1
Andorra	1
Angola	1
Anguilla	1
Aruba	1
Central African Republic	1
Congo	1
Democratic Republic of the Congo	1
Faroe Islands	1
Laos	1
Lesotho	1
Liberia	1
Mali	1

Location	No. of contribut.
Mongolia	1
Myanmar (Burma)	1
Netherlands	1
Niger	1
Oman	1

Location	No. of contribut.
Reunion	1
Rwanda	1
Seychelles	1
Sierra Leone	1
Svalbard and Jan Mayen	1

Location	No. of contribut.
Tokelau	1
Tonga	1
Uzbekistan	1
Vatican City	1

Appendix 2: Most popular music links

Each link was tweeted at least five times

YOUTUBE VIDEO ID	YOUTUBE VIDEO TITLE
IJNR2EpS0jw	Dumb Ways to Die
6q0dsG8fTHY	DJ Earworm Mashup - United State of Pop 2012 (Shine Brighter)
QK8mJJJvaes	MACKLEMORE & RYAN LEWIS - THRIFT SHOP FEAT. WANZ (OFFICIAL VIDEO)
NarbqFTrXbE	For the Win - Team Unicorn (Featuring Weird Al Yankovic, Aisha Tyler, Grant Imahara)
8N_tupPBtWQ	Muppet Show - Mahna Mahna . . . m HD 720p bacco . . . Original!
5_sfnQDr1-o	Baby Monkey (Going Backwards On A Pig) - Parry Gripp
Z0GFRcFm-aY	R.E.M. - It's The End Of The World
jnlRdu4Wvuk	Lord Sij – Shuup
zvCBSSwgtg4	The Lumineers - Ho Hey (Official Video)
A1VTaAYH3Hc	Badtameez Dil - Full Song - Yeh Jawaani Hai Deewani | Ranbir Kapoor, Deepika Padukone
If5MF4wm1T8	Pop Danthology 2012 - Mashup of 50+ Pop Songs
kfVsfOSbJY0	Friday - Rebecca Black - Official Music Video
OpQFFLBMEPI	P!nk - Just Give Me A Reason ft. Nate Ruess
tydgBBrmRfg	The 1975 - Chocolate at Radio 1's Future Festival
8UVNT4wvIGY	Gotye - Somebody That I Used To Know (feat. Kimbra) - official video
ASO_zypdnsQ	PSY - GENTLEMAN M/V
e-fA-gBCkj0	Bruno Mars - Locked Out Of Heaven [OFFICIAL VIDEO]
Ek0SgwWmF9w	Muse – Madness
0gEVaniPOmU	Of Monsters and Men - Mountain Sound
9e9NSMY8QiQ	Tegan and Sara - Closer [OFFICIAL HD MUSIC VIDEO]
bii-PIGprv8	The Pink Panther Show Original Opening HQ
ghb6eDopW8I	Of Monsters and Men - Little Talks
lcOxhH8N3Bo	Bonnie Tyler - Total Eclipse of the Heart
Leg3uiu0g1c	Doves – Pounding
Qsy7kJyizoc	Sam and the Womp | Bom Bom (Official Video)
VvcohzJvviQ	Peaches

Note. A YouTube playlist containing a unique mix of the titles mentioned in this table can be found at https://www.youtube.com/playlist?list=PLFF185AyoHpojr7fZB8d0GwsOIyZAXjvG