On the Virtuous and the Vexatious in an Age of Big Data

David Huron
Ohio State University

The use of computers in music research has a history that in some sense precedes the invention of the computer. Back in the 1940s, Bronson (1949, 1959) pioneered the use an IBM card-sorter (a predecessor of the general purpose computer) to examine patterns in traditional British folk ballads. In the 1960s, the Princeton *Josquin Project* attempted to use computational methods to establish the network of copied manuscript sources (Hall, 1975). In the mid 1970s, Bauer-Mengelberg spearheaded the *DARMS* project—an enterprising venture to develop a comprehensive encoding language for representing and processing notated music (Erickson, 1976). No ethnomusicologist will be unaware of Lomax's ambitious *Cantometrics* (Lomax, 1968). Although these pioneering efforts might warrant the moniker "visionary," the accomplishments of these pioneers were modest and their colleagues often expressed bemused curiosity, but remained uninspired.

What was missing from most early projects was sufficient musical data to process. In the heady days of the 1960s and 70s, the optimism outpaced reality when it came to computers. Most researchers involved in computers and music thought that scanned music recognition was just around the corner. (Reliable music scanning did not appear for another three decades.) Consequently, there was considerable reluctance to engage in large-scale manual data encoding. A few scholars were eager to get on with the research, and so made early commitments to music encoding. Schaffrath's group in Germany was a notable effort (Schaffrath, 1995), as was Walter Hewlett's impressive initiative at the Center for Computer Assisted Research in the Humanities (CCARH). Additionally, there were the efforts of innumerable volunteers involved in the *Répertoire International des Source Musicales* (RISM) project (Benton, 2002). Beginning in the 1980s, I manually encoded music for one hour each evening for several years. As it turned out, reasonably reliable music scanning had to wait until the new millenium. Even now, few databases have been assembled from scanned scores. Much of this work has been carried out by Craig Sapp at CCARH.

Data Riches

The Internet has been the enabling technology for the era of Big Data. The blossoming of the Internet has made data both easier to accumulate and more widely accessible. The sources of data are varied: many people working collaboratively (e.g., Wikipedia, Human Relations Area Files; see Ember, 1997), many people working individually (e.g., YouTube, Facebook), and many people providing a large market that encourages corporate-initiated data aggregation (e.g., iTunes, Google, Amazon.com). The Human Genome Project has had an especially salutary effect—making Big Data a compelling interest among researchers and granting agencies.

In the case of music, the Internet quickly encouraged an expansion of score-based materials (e.g., International Music Score Library Project), as well as various audio and MIDI formats. Important audio resources include both commerical (e.g., LastFM, Pandora, Spotify) and non-commercial (e.g., YouTube, MIREX) initiatives. Online reference tools such as *Themefinder* provide useful sources for musical meta-data, and annotated musical data (such as Echo Nest-type music characterization) has become commonplace.

In general, researchers have made effective use of online volunteers, not just to participate in web-based surveys or experiments, but also to assist in massive data analysis projects—such as the successful Galaxy Zoo in the field of astronomy and folding@home in biology. Paid online research volunteers are readily available through Amazon's Mechnical Turk, and Vanderbilt University's ResearchMatch provides a useful web-based system for recruiting research participants that match specific demographic requirements. Finally, general-purpose search engines like Google, Yahoo, and Bing allow direct sampling of music-related opinions.

From a methodological perspective, it is helpful to distinguish two broad kinds of data: finite data (such as historical data), and unbounded data (such as prospective future data). Examples of the former include notational databases, including both score images (e.g., PDFs) and symbolic data (e.g., MusicXML, Humdrum), as well as sound archives (e.g., MP3). Although musicians continue to write and record music, much or most of the material of interest to researchers is historical (and finite) rather than prospective (and unbounded).

Music Perception, volume 31, issue 1, pp. 4–9, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2013.31.1.4

Examples of the latter include the ever-expanding volumes of behavioral data, such as through Twitter, Facebook, iTunes, Amazon, LastFM, etc. These include (performed/interpreted) MIDI data, plus innumerable sources from cookies, GPS information, IP address lists of webpage visitors, and so on.

Apart from these sources, one might also note the existence of various reference tools, innumerable lists, indexes, and other repositories, as well as the growing practice of data sharing among researchers, including the posting or archiving of experimental data (e.g., the World Data Center).

Aims

Why, one might ask, are large amounts of data important? Methodologists point to a general trade-off between two kinds of scholarly errors: claiming something to be *true, useful, or knowable* that is in fact *false, useless, or unknowable*, and claiming something to be *false, useless, or unknowable* that is in fact *true, useful, or knowable*. In attempting to minimize the first (Type I) error, we inevitably increase the likelihood of making the second (Type II) error, and vice versa. The very best way to minimize both types of error is simply to gather more evidence. This is the principal reason why collecting data is important. To the extent that minimizing knowledge-related error can contribute to health, justice, environmental well-being, and other good things, the drive toward Big Data is not merely some obsession with things numerical, or a kleptophilic compulsion to collect, but a proper moral imperative (see Huron, 1999).

Four Methodological Caveats

The virtues of Big Data notwithstanding, large data sets also raise several methodological concerns. Here I will address only four issues. First, since large amounts of data make statistical significance more likely, sampling bias is more likely to lead to spurious results. Collecting data via the Internet is fraught with many pitfalls. The Internet is not a level playing field; not all cultures and periods are equally represented. From the beginning, there has been a strong bias toward English-language materials. As much as the web feels like a global community, it remains something of a privileged club. Caution is necessary when the web is treated as a population from which the researcher samples. Where large numbers are involved, statistical significance is more likely to be an artifact of sampling bias rather than a true characteristic of the population. (With smaller samples, these sampling biases are less likely to achieve significance.) This applies not simply to the Internet, but to all large databases. Large data sets necessitate more careful consideration of issues of representative sampling.

A second, related consideration is that large volumes of data encourage us to pay more attention to effect sizes rather than statistical significance alone. Of course, this is as it should be. In studying any phenomenon, we should aim to identify the most important factors first. Usually, researchers are excited simply to find *any* statistically significant relationship. Big data encourages greater responsibility in reporting effect sizes.

A third concern arises from the relative ease of electronic query. Easy processing can quickly lead to problems of multiple tests. When a researcher accepts a 95% confidence level for a statistical test, this means that the researcher accepts a 1-in-20 chance of reporting nominally significant results that are, in fact, spurious. If a journal contains 20 articles, and each article presents a single result that is claimed to be significant at the 95% confidence level, then, on average, 1 of the 20 articles is presenting spurious results. The same logic applies to a single researcher who carries out many tests. Twenty queries of a database can similarly lead to spurious "significant" correlations. Moreover, this problem arises even if you don't perform formal statistical tests for each query. When researchers "eyeball" the results, they immediately discount relationships that don't appear to be significant. One can't avoid multiple tests by reserving formal tests for only those relationships that look promising. Each "eyeballed" relationship also counts as a test. Consequently, informal exploratory work is a perpetual temptation to be resisted.

Statistician Arthur Owen has informally suggested that researchers have a "lifetime p value" —that ought to be reduced as one continues to do further tests. The number of lifetime experiments involving human subjects carried out by a typical research psychologist is probably on the order of 100, so the "lifetime p" value is less of a concern. However, the number of tests carried out in corpora studies can balloon into the thousands. At this point, multiple tests and "lifetime p" values loom large as serious concerns.

Related to the ease of processing is a fourth concern: Large data sets require a careful delineation between a priori and post hoc theorizing. In particular, Big Data forces us to pay closer attention to the origins of hypotheses. Traditionally, methodologists make a useful distinction between the *context of discovery* and the *context of justification*; that is, between the origin of some idea, and the evidence used to test or support that idea (e.g., Duhem, 1977). Early writers in the philosophy

of science argued cogently that it does not matter where an idea comes from; the critical issue is whether a rigorous empirical test is carried out. A classic (though probably apocryphal) historical example is Kekulé's famous discovery of the structure of benzene. Having mightily struggled to decipher benzene's chemical organization, Kekulé reported having a daydream in which a snake swallowed its own tail. The dream inspired him to consider that benzene might form a ring. In conventional scientific methodology, the context of discovery (a dream) has no bearing on the context of justification (the subsequent experiments carried out by Kekulé that were consistent with a ring structure for benzene). Conventionally, we simply don't care where an idea comes from: for all we care, a useful conjecture about music might arise from the Egyptian Book of the Dead. In research, only the context of justification matters.

There is, however, at least one circumstance in which the context of discovery becomes germane. Once again, history provides an instructive example; in the theory of continental drift. Today, there is very good evidence that the continents behave as plates that slide very slowly across the earth's surface. However, the theory began in an inauspicious manner. If you look at a map of the world, there seems to be a certain similarity between the eastern coastlines of north and south America, and the western coastlines of Europe and Africa. One could easily imagine all four continents pushed together like jigsaw puzzle pieces.

Over the past century, excellent evidence has been assembled that is consistent with this theory. However, this evidence was not available when the theory was first proposed. If we ask "What inspired the theory?" the answer was "Look at how the continents seem to fit together like jigsaw puzzle pieces." If we then ask "What evidence do we have that is consistent with the theory?" the answer was "Look at how the continents seem to fit together like jigsaw puzzle pieces." In other words, the same observation was used both as the inspiration for the theory, and as evidence in support of the theory. The reasoning is patently problematic. The theory of continental drift took time to be accepted precisely because of the early lack of independent evidence.

This problem of "double-use data" is an omnipresent danger in database studies. Once a researcher looks at some data, any theory formed is now post hoc. One cannot then claim that the theory was a priori and use the observations as evidence that tests the theory. Once you make an observation, you cannot pretend that you predicted that observation. With post hoc theories, one cannot legitimately use the language of prediction that is the essence of hypothesis testing.

As someone who has assembled and used large databases for some decades, I have learned to ration my exploratory activities. When introducing other scholars to musical databases, one of the hardest tasks is to dampen their enthusiasm and to impress on them the importance of not engaging in unstructured open-ended data exploration. Any fixed-length database is actually a finite resource, and each correlation calculation effectively degrades the value of the corpus.

In the case of historical or retrospective data, exploratory work should always be done with subsets of a database. Records should be kept of each exploratory test in order to correct later for multiple tests. If one is serious about doing proper empirical research, there is no such thing as "just exploring the database." Ideas should be tested using a reserved data set that did not participate in the exploratory work (e.g., Downie, 2006). Better yet, new data should be sampled or encoded for each project.

When creating a reserved sample for exploratory work, it is important to avoid using random sampling of the main data set. In the field of computer speech recognition, by way of example, recorded speech is used both to train the language model and to test the effectiveness of the resulting recognizer. Care must be taken when partitioning the training data from the reserved data. Suppose, for example, that the database of speech contains recordings of 100 speakers each speaking 100 words. If one uses half of the words spoken by all 100 speakers for the training data, this will unfairly bias any later measure of recognizer effectiveness. In order for the test to be valid, one needs to test the recognizer exclusively on new and unfamiliar speakers, not just new and unfamiliar words. That is, one needs to test using maximally independent data. Hence one should build the recognizer using a subset of the *speakers*, not merely a subset of the *utterances*.

Similarly, suppose one is interested in formulating and testing general claims about music. In exploring the corpus to form possible conjectures, one should limit oneself to music from a single composer, a single genre, a single period, or a single culture. The ensuing conjecture can then be tested against other composers, other genres, other periods, or other cultures. The test data will then exhibit greater independence, and the ensuing test better invites failure. Contrary to intuition then, one should resist assembling "representative" samples in the context of discovery. Exploratory studies, I would suggest, are best done with idiosyncratic data rather than representative data. This is done not to improve the quality of the exploratory work, but to better assure the quality of the subsequent testing. When the data are

finite, the methods employed in the context of discovery should indeed be different from the methods employed in the context of justification. As a further illustration of this point, suppose a demographer had access to a complete data set for all 1.3 billion Chinese citizens. What purpose is served by randomly dividing the data in half, carrying out exploratory research on one half and then testing the results on the reserved half? A sample size of 650 million people is very likely to be fully representative. Better to divide the data in an unrepresentative way—say, between old and young, or between urban and rural—if the aim is to invite the failure of a test of some conjecture purporting to relate to all Chinese people.

Hypotheses related to rare occurrences will quickly exhaust even a large database. For example, consider the following question: Are German, French, and Italian sixth chords more likely to be used by composers of their respective nationalities, or are these just fanciful labels? Augmented sixth chords are rare: most musical works contain none, and when they appear there is often only one. Even a sample containing several thousand works (involving potentially a million annotated harmonies) is probably insufficient to test a hypothesis like this. (In the first instance, the database is unlikely to include a sufficient volume of music by French composers.)

Even in an age of Big Data, many questions regarding music will prove to be intractable because of limitations in the amount of pertinent music, even assuming all of the pertinent sources become available online. This is already a common experience in the world of language processing.

Quality Control

Another concern relates to the handling of errors. No database is pristine or error-free. For many forms of processing, small errors do not matter. For example, a pitch error rate of 1% will have a negligible effect on key determination. However, depending on the type of processing, the error rate can balloon. A 1% pitch error rate becomes a 2% error rate when processing melodic intervals (a single wrong pitch falsifies both the preceding and ensuing melodic intervals). Moreover, a 1% pitch error rate translates into a 4% error rate for identifying four-note chords. When processing two-chord harmonic progressions, the error rate now approaches 8%. An equation for calculating the error bars for any arbitrary musical measure is available in Huron (1988).

Error rates inevitably lead to questions concerning the quality of the source material. For MIDI data, musicians are rightly interested in the quality of the performance. For encoded scores, music scholars rightly want to know about the source edition. Musicologists are inevitably disappointed to learn that a given database was not encoded using the best quality critical source. The Center for Computer Assisted Research in the Humanities takes great care to ensure that all encoded materials are in the public domain; however, this necessarily precludes making use of the latest critical editions.

The production of a critical edition is one of the most labor-intensive of all musicological activities. A group of musicologists can easily spend two or more decades assembling a critical edition for a given composer. Suppose that the youngest editor in a project is a musicologist who is 40 years of age in 2010. Further suppose that this musicologist lives to the modest age of 80. Under current copyright regimes, the resulting critical edition will not enter the public domain until the year 2150. Given recent history, there is a good chance that future legislation will extend this, so a critical edition created today may not enter the public domain before the year 2200. Musicians and scholars will pay and pay and pay for access to these materials long into the future. Even today, most music scholars simply cannot afford, say, a copy of the complete works of Shostakovich. Compared with other fields, the situation in musicology is notably grim: imagine a Shakespearean scholar who was unable to afford a personal copy of a critical edition of the complete works of Shakespeare.

In what way are we musicologists serving the scholarly and musical communities by giving our work away to prestige publishers who will hold a long-term monopoly on the best editions of all the works by a given composer? Music scholars and our scholarly organizations have been surprisingly delinquent in planning for future scholars. Widespread general electronic access to the bulk of top-quality critical editions in music is unlikely to happen in the lifetime of any scholar currently alive. We have only ourselves to blame for this regrettable situation.

Future

Within the next decade, one can reasonably expect nearly all of the history of commercially recorded music to become available online. It may take another fifty years to reach 90% availability for the non-commerical materials found in the innumerable audio archives around the world. The voluminous recorded archives in the various members of the European Radio Union, for example, may take several decades to come online. The speed of digitization will be hampered by

limited budgets in many small cash-strapped archives, repositories, and libraries. In addition, some archives may resist digitization in an understandable (but perhaps misguided) effort to preserve the monopoly value of the collection.

With regard to notated scores, the irksome task of turning pages is likely to have a facilitating effect in bringing materials online. It is not inconceivable that within ten years, the majority of music performed from "sheet music" will employ tablet displays rather than printed sources. Automatic (audio-driven) page-turning may become the norm, and so data will need to include symbolic representations of notes, not simply image files.

Publishers are likely to become less important in the future with much commercial power passing to data aggregators like EBSCO and Baker & Taylor. Only those publishers with truly large back catalogues will retain much power. Alternatively, publishers may respond to their diminishing power by forming distribution consortia, and so compete with data aggregators. One sincerely hopes that academic libraries and scholarly organizations will form appropriate alliances to create the infrastructure that ensures both accessibility and appropriate revenue streams. Alas, much of the scholarly work of recent decades (including perhaps the next decade) will be swallowed into a dark hole of distribution limbo, and consequently will play a lesser role in research for another century or so.

The reciprocal problem to making scholarly works easily available, is the problem of restricting the collecting and distribution of personal data. Personal computers began as objects over which individual users exercised nearly complete control. Computing devices are increasingly Trojan horses that permit unprecedented levels of snooping. With access to playlists, browser history, etc., the commercial world has already assembled mountains of musically pertinent behavioral data.

Imagine an academic researcher endeavoring to get human subjects approval for distributing a piece of software that spies on people's music listening habits without explicitly asking their permission. No review board would approve such a project. However, these practices are the norm in the commercial world. Even if legislative initiatives arise to defend and regain lost privacy, in the future, the greatest opportunities for conducting music research are likely to shift toward industry away from academia.

Major breakthroughs in understanding music are likely to arise from industry-related research. In the arts and entertainment industry, there has been little incentive for commercial researchers to pursue patents. Consequently, those working in the arts and entertainment industries tend to rely on trade secrecy. Shortly after I published *Sweet Anticipation* in 2006, I received a letter from a retired British advertising executive. The executive was grateful to see an idea published that his advertising agency had long ago discovered but had kept secret for decades. How many useful discoveries about music already exist as commercial secrets? Although corpus studies will prove immensely beneficial to academic researchers and greatly expand our understanding of music, the principal beneficiaries will be industry. Consequently, the ratio of public-to-proprietary knowledge about music may well decline in the future.

Population Hermeneutics

Any set of observations is open to more than one interpretation. In empirical research, competing theories are evaluated by formulating a critical experiment where diverging predictions are tested against newly collected data. So what happens when researchers have access to the entire population? A scholar interested in the characteristic "Motown sound" need not be content with a musical sample, when she/he can have access to the entire catalogue of recordings produced by Motown Records. When researchers have access to the population, then one can no longer use prediction as a tool for discriminating between competing theories. Once all of the data are in hand, all conjectures are tainted by the possibility of post hoc origins.

In these circumstances, researchers can no longer rely on the rhetoric of hypothesis testing. Instead, scholars will simply endeavor to show that, on balance, one interpretation is more parsimonious, or more coherent, or better dovetails with other theories, or is parallel in structure to similar phenomena in other domains, or better fits with commentary fragments made by major figures, and so on. In short, the scholarly rhetoric changes from prediction to hermeneutics.

One definition of a "statistic" (singular) is that it is an estimate of a population parameter. The field of statistics is primarily (though not exclusively) a discipline whose aim is to estimate properties of some population on the basis of a subset. However, once the population parameters are known, much (though not all) of statistics becomes irrelevant. Once the population is accessible, empirical science is replaced by empirical hermeneutics.

On the one hand, Big Data will inevitably encourage music scholars to become more conversant with statistical methods. On the other hand, for finite data sets, as competing theories proliferate, Big Data will ultimately move research away from hypothesis testing, back to the

hermeneutic rhetoric familiar to all data-impoverished fields.

Having said this, however, we have a long way to go before reaching this "ironic" state. In the meantime, we live in an age of rapidly expanding research resources, and therefore, rapidly expanding opportunities for addressing questions about music. It is appropriate to pause, pinch ourselves, and celebrate the extraordinary fact that so much music, from so many different sources, is so readily available in devices that can fit in the palm of your hand.

Ultimately, disciplines are defined not by their methods but by the questions they ask. The development of new methods, however, can often make it easier to pursue certain questions. Conscientious scholars focus on the questions, and then acquire whatever tools best allow them to address those questions. Never before have researchers had access to such powerful tools for posing and addressing music-related questions, both the trivial and the consequential. These resources provide unprecedented opportunities for the current generation of music researchers to better serve the public good.

References

Benton, R. (2002). Répertoire international des source musicales. In S. Sadie (Ed.), *The new Grove dictionary of music and musicians* (2nd ed., Vol. 21, p. 194). Oxford, UK: Oxford University Press.

Bronson, B. H. (1949). Mechanical help in the study of folk song. *Journal of American Folklore*, *62*, 81-86.

Bronson, B. H. (1959). Toward the comparative analysis of British-American folk tunes. *Journal of American Folklore*, *72*, 165-191.

Downie, S. (2006). The music information retrieval evaluation eXchange (MIREX). *D-Lib Magazine*, *12*. Retrieved from http://www.dlib.org/dlib/december06/downie/12downie.html

Duhem, P. (1977). *La théorie physique: Son objet, sa structure* [The aim and structure of physical theory] (P. P. Wiener, Trans.). New York: Atheneum.

Ember, M. (1997). Evolution of the human relations area files. *Cross-Cultural Research*, *31*, 3-15.

Erickson, R. F. (1976). *DARMS: A reference manual* [Duplicated typescript]. Binghamton, NY.

Hall, T. (1975). Some computer aids for the preparation of critical editions of Renaissance music. *Tijdschrift van de Vereniging voor Nederlandse Musiek Geschiedenis*, *25*, 38-53.

Huron, D. (1988). Error categories, detection and reduction in a musical database. *Computers and the Humanities*, *22*, 253-264.

Huron, D. (1999). The new empiricism: Systematic musicology in a postmodern age. *Ernest Bloch Lectures* [Public lectures]. University of California, Berkeley. Retrieved from http://www.music-cog.ohio-state.edu/Music220/Bloch.lectures/3.Methodology.html

Lomax, A. (1968). *Folk song style and culture*. New Brunswick, NJ: Transaction Publishers.

Schaffrath, H. (1995). *The Essen folksong collection*. Stanford, CA: Center for Computer Assisted Research in the Humanities.