Brian Whitman (The Echo Nest Corporation)

Very large scale music understanding

Scientists and engineers around the world have been attempting something undeniably impossible-- and yet, no one could ever question their motives. Laid bare, the act of “understanding music” by a computational process feels offensive. How can something so personal, so rooted in context, culture and emotion, ever be discretized or labeled by any autonomous process? Even the ethnographical approach -- surveys, interviews, manual annotation -- undermines the raw effort by the artists, people who will never understand or even perhaps take advantage of what is being learned and created with this research. Music by its nature resists analysis. I’ve led two lives in the past ten years-- first as a “very long-tail” musician and artist, and second as a scientist turned entrepreneur that currently sells “music intelligence” data and software to almost every major music streaming service, social network and record label. How we got there is less interesting than what it might mean for the future of expression and what we believe machine perception can actually accomplish.

In 1999 I moved to New York City to begin graduate studies at Columbia working on a large “digital government” grant, parsing decades of military documents to extract the meaning of the acronyms and domain specific words. At night I would swap the laptops in my bag and head downtown to perform electronic music at various bars and clubs. As much as I tried to keep them separate, the walls came down between them quickly

when I began to ask my fellow performers and audience members how they were learning about music. “We read websites,” “I’m on this discussion board,” “A friend emailed me some songs.” Alongside the concurrent media frenzy on peer to peer networks (Napster was just ramping up) was a real movement in *music discovery*-- technology had obviously been helping us acquire and make music, but all of a sudden it was being using to communicate and learn about it as well. With the power of the communicating millions and the seemingly limitless potential of bandwidth and attention, even someone like me could get noticed. Suitably armed with an information retrieval background alongside an almost criminal naivete regarding machine learning and signal processing I quit my degree program and began to concentrate full time on the practice of what is now known as “music information retrieval.”

The fundamentals of music retrieval descend from text retrieval. You are faced with a corpus of unstructured data: time-domain samples from audio files or score data from the composition. The tasks normally involve extracting readable features from the input and then learning a model from the features. In fact, the data is so unstructured that most music retrieval tasks began as blind roulette wheels of prediction: “is this audio file rock or classical” [Tzanetakis 2002] or “does this song sound like this one” [Foote 1997]. The seductive notion that a black box of some complex nature (most with hopeful success stories baked into their names-- “neural networks,” “bayesian belief networks,” “support vector machines”) could untangle a mess of audio stimuli to approach our nervous and perceptual systems’ response is intimidating enough. But that problem is

so complex and so hard to evaluate that it distracts the research from the much more serious elephantine presence of the emotional connection underlying the data. A thought experiment: the science of music retrieval is rocked by a massive advance in signal processing or machine learning. Our previous challenges in label prediction are solved-- we can now predict the genre of a song with 100% accuracy. What does that do for the musician, what does that do for the listener? If I knew a song I hadn't heard yet was predicted "jazz" by a computer, it would perhaps save me the effort of looking up the artist's information, who spent years of their life defining their expression in terms of or despite these categories. But it doesn't *tell me anything* about the music, about what I'll feel when I hear it, about how I'll respond or how it will resonate with me individually and within the global community. We've built a black box that can neatly delineate other black boxes, at no benefit to the very human world of music.

The way out of this feedback loop is to somehow automatically understand reaction and context the same way we could with perception. The ultimate contextual understanding system would be able to gauge my personal reaction and mindset to music. It would know my history, my influences and also understand the larger culture hovering around the content. We are all familiar with the earliest approaches to contextual understanding of music -- collaborative filtering, a.k.a. "people who buy this also buy this" [Shardanand 1995] -- and we are also just as familiar with its pitfalls. Sales or activity based recommenders only know about you in relationship to others-- their meaning of your music is not what you like but what you've shared with an anonymous hive. The

weakness of the filtering approaches become vivid when you talk to engaged listeners: “I always see the same bands,” “there’s never any new stuff” or “this thing *doesn’t know me*.” As a core reaction to senselessness of the filtering approaches I ended up back at school and began applying my language processing background to music-- we started reading about music, not just trying to listen to it. The idea was that if we could somehow approximate even one percent of the data that communities generate about music on the internet-- they review it, they argue about it on forums, they post about shows on their blog, they trade songs on peer to peer networks-- we could start to model cultural reaction at a large scale. [Whitman 2005] The new band that collaborative filtering would never touch (because they don’t have enough sales data yet) and acoustic filtering would never get (because what makes them special is their background, or their fanbase, or something else impossible to calculate from the signal) could be found in world of music activity, autonomously and anonymously.

Alongside my co-founder, whose expertise is in musical approaches to signal analysis [Jehan 2005], I left the academic world to start a private enterprise, “The Echo Nest.” We are now thirty people, a few hundred computers, one and a half million artists, over ten million songs. The scale of this data has been our biggest challenge: each artist has an internet footprint of on average thousands of blog posts, reviews, forum discussions, all in different languages. Each song is comprised of thousands of indexable events and the song itself could be duplicated thousands of times in different encodings. Most of our engineering work is in dealing with this magnitude of data-- although we are not an

infrastructure company we have built many unique data storage and indexing technologies as a byproduct of our work. The set of data we collect is necessarily unique: instead of storing the relationships between musicians and listeners, or only knowing about popular music, we compute and aggregate a sort of internet-scale cache of all possible points of information about a song, artist, release, listener or event. We began the company with the stated goal to index everything there is about music. And over these past five years we have built a series of products and technologies that take the best and most practical parts from our music retrieval dissertations and package them cleanly for our customers. We sell a music similarity system that compares two songs based on their acoustic and their cultural properties. We provide tempo, key and timbre data (automatically generated) to mobile applications and streaming services. We track artists' "buzz" on the internet and sell reports to labels and managers.

The core of the Echo Nest remains true to our dogma: we strongly believe in the power of data to enable new music experiences. Since we crawl and index everything, we're able to level the playing field for all types of musicians by taking advantage of the information given to us by any community on the internet. Work in music retrieval and understanding requires a sort of wide-eyed passion combined with a large dose of reality. The computer is never going to fully understand what music is about, but we can sample from the right sources and do it often enough and at a large enough scale that the only thing in our way is a leap of faith from the listener.

References

Jonathan T. Foote. “Content based retrieval of music and audio.” In C.-C. J. Kuo et al., editor, Multimedia Storage and Archiving Systems II, Proc. of SPIE, Vol. 3229, pp. 138-147, 1997.

Tristan Jehan. “Creating Music by Listening.” Dissertation, Massachusetts Institute of Technology, 2005.

Upendra Shardanand and Pattie Maes. “Social Information Filtering: Algorithms for Automating ‘Word of Mouth.’” ACM Press, 1995, pages 210-217.

George Tzanetakis and Perry Cook. Musical Genre Classification of Audio Signals IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 10, NO. 5, JULY 2002

Brian Whitman. “Learning the Meaning of Music.” Dissertation, Massachusetts Institute of Technology, 2005.