[SQUEAKING]

[CLICKING]

[RUSTLING]

[CLICKING]

MICHAEL SCOTT ASATO CUTHBERT: Hello, computational music analysts. I'm going to talk a little bit today about some topics that are not in themselves essentially computational although having a computer helps but are really about the intersection of mathematical and quantitative thinking with music analysis, music theory, and musicology or the study of music history. And these are some of the things that I've done in my own research that I hope will make you think about some ways that data analysis can go beyond the score and tell us something more about the history of music.

When I'm not doing computational music analysis, my other job is as a medieval music researcher, and one of the certainties I was always taught about medieval music is that we've lost most of it. We've lost so much of it, and it's just incredible how much we have.

In fact, most people always said that the music that survives is just the tip of the iceberg. It was first said by a very important medieval historian-- musicologist Nino Pirrotta, but this term was picked up from on by dozens of other people. And so when we think about what's lost, it could be a number of manuscripts that's lost or it could be things that were lost because they were never written down and we've forgotten them, but if we want to think about even the things that were once written down on songs that were once written down, the pieces that were once written down, how many of them do we no longer have any copy of at all? What is the tip of the iceberg? Is it 90%? That's what a lot of people say the visible part of the iceberg is. Is it 99% lost? 99.9% lost, or as 1%-- basically everything. So I'll call that 99.9999%.

My friend and former teacher, Sean Gallagher once asked me, hey, would there be any way-- you're a math guy-- to prove this, to show that it's true. And I immediately said, no, there's no way we could do this. Then I thought a little bit more and realized there are people who look at how things get lost and stuff. Later on I found this great article on what happens in peer-to-peer song sharing when you just start disconnecting lots of computers from the network as happened during the crackdown on Napster and things like that.

And they showed that as you go from-- if you have a 100,000 computers on a network, you might have 5 million songs. But if the network gets down to maybe 10,000 or 15,000 computers, you only lose about half the songs. So that was interesting to me. That was a little bit counterintuitive that you could really lose a lot of nodes on a network and not lose a lot of the content there. And I thought, well, computers might be something like medieval manuscripts that we lose a bunch, but maybe we still have the songs.

So I wrote two articles about this topic. First is what I'm really going to be talking about now. It's an article called "Tipping the Iceberg" in a journal called *Musica Disciplina* and another, it's a little bit harder to find. It's called Monks, Manuscripts, and Other Peer-to-Peer Song Sharing Networks of the Middle Ages. It's a great book that the University of Pennsylvania helped publish called Cantus Scriptus-- Technologies of Medieval Song. I like the idea of technology existing throughout history.

Anyhow so one of the ways that we could do this is just look at all the records of missing pieces and see how many of them we have. And that's sarcastic, but there are things that exist in the world like this-- there's a manuscript, it's actually only I think two or four pages survive now that was owned by the Duchess of Trémouille. So it's often called the Trémouille manuscript. It's in the National Library of Paris now.

And if you look at this manuscript, what-- we only have a couple pages of it, but we have the entire index, the table of contents that tells you what the names of the pieces are and what page they would have been on in the original manuscript. So we can look at-- well, we don't have this manuscript, but how many of these pieces do we-- have we lost? So we can see that I've marked here, these are all the pieces that were pretty sure we don't have copies of.

And here's all the pieces that we're pretty sure we do have copies of. And there's some pieces that we can't really be sure they're parts of medieval music traditions like the mass that have these somewhat generic names like credo or something, like if we saw, oh, we lost symphony. Well, whose symphony, things like that.

So when we overlay these two together, we see that actually those that survive really outrank those that don't. So in this particular index, we have 114 compositions that of which 109 are identifiable and 68% of them seem to survive.

Or another thing that I looked at was this collection of sonnets by cool but second tier Italian poet named Prodenzani. And one of the things that makes a lot of people think he's second tier is he throws in a lot of his songs just lists of things-- lists of all the food they eat-- ate, lists of all the dances they danced that day, and in this particular case, a list of all the songs that were sung.

That begins at the top of-- nice for where we're about to be in the year with the violin, viola, something. They made all the may songs like "Rosetta Che Non Cambi Mai Colore," "The Rose That Never Changes Its Colors," "Je Sui Nafres Tam Fort, Dolce Sapore," et cetera, et cetera. We can basically figure out which ones of these things are titles of songs, and then we can see how many of them we know.

So "Rosetta Che Non Cambi Mai Colore" is a song that survives. So is "Je Sui Nafres Tam Fort." But some of them like toward the bottom, "Costei Sarebbe Bella in Paradiso," we don't know that song.

So we can look over, and, again, we can see how many of them survive, how many of them don't survive. And then these other four, where-- we can argue about, but there's songs with similar titles. Are they just being misquoted or are they same or is this really a song, things like that. So these are the two proportions that it could be.

So depending on what we look at, there's somewhere between 16 and 17 compositions, and three to seven of them are lost. And 10 to 12 survive. So that's about 59% to 75% of the songs here survive. If we look at all of the sonnets in this collection by Prodenzani, there's 59 identifiable compositions, and 40 of them survive. So that's 71%

So here's two pieces of evidence, both from textual sources, to try to see how much survives or how much doesn't. But can we do better?

One of the things I thought of was that there are people who spend a lot of time trying to count things that are difficult to count because you can't capture them all. There the animal population in biologists, they are people who go out, and they can't capture every deer in a forest. So what do they do? They might go out-- here's one way that they do it.

They might go out at a certain point and go and capture all the deer they can capture in a couple days. Maybe it's-- let's say it's 10, 10 deer. And they go, and they tag the deer in some way or something. Somehow they-- so they can recognize the same deer again.

Then maybe they let a month go by or something, not too long but more-- enough time that animals can move around and stuff. And then they go out, and they capture deer for a day again. And let's say they capture 10 deer again.

The point is, the second time they do a capture-- this is called capture recapture methods-- the second time they do it, they look to see how many of the deer that we captured the second time did we also capture the first time? If you think about it, if you go into a forest and you walked around pretty randomly or something and you-- all of the deer or nine out of the 10 deer that you capture the first time you see the same ones, the second time, you might think, wow, we've seen almost all the deer in the forest.

On the other hand, let's say you capture 10 deer and none of them that you captured the second time had the tags. You'd be like, wow, I know I've seen 20 deer here, but law of averages suggests there's a lot more than 20 deer. And, in fact, there's formulas that can-- you can-- that people do with this type of thing.

So I tried to figure out what would be the equivalent in music. And so I started to think, let's go and look at some equations we might make.

If we had the number of pieces-- if we knew the number of pieces that originally existed and we multiplied by the probability that any given piece would be missing, we would know the total number of missing pieces. That's fantastic. That is a very hard equation. It's very hard to argue against, but it's completely useless.

We don't know the number of pieces originally. We don't know the probability that any given piece would be missing. And therefore we don't know how many missing pieces are. So one equation, three unknowns, we don't know any of them.

But we can think through some of the things here. I'm going to go through a quicker version of this, but I have elsewhere-- well, I have in my article a little bit deeper version that talks about all the un-- the assumptions that could be in play. But let's start with how can we figure out the probability that we're missing a given piece.

Well, here's one model, and it's kind of a bad assumption maybe to start with. But here's one model. What if there's just medieval scribes had baskets of pieces, just all the pieces. They knew them all their head, there's something like that, and they randomly chose ones that they were going to write down.

Well, then the probability that a given piece would be in a particular manuscript is approximately the proportion of all of the pieces in the manuscript. That is, if R sub I is the number of pieces in the manuscript and N is the total number of pieces originally copied, then we would know the probability that a piece would be in a given manuscript or songbook. It's not exactly that number because you don't do duplication, but it's pretty close.

It's actually the people have argued against me that, well, you have to correct for that error, but you only have to correct for the probability of duplication if the number of pieces in that manuscript is close to some major percentage of the total number of pieces originally copied. So that's already getting that.

Anyhow so this is the probability that a given piece would be in a given manuscript. From here, we can work out what's the probability the piece is not in a given manuscript. And we just subtract that from one. Great. So that's pretty good.

Now let's say you have two independent manuscripts. They're being copied in different places, different people, whatever, something like that. What would be the probability that the piece would not be in any of these two manuscripts.

And so you just multiply the probability that it wouldn't be in one manuscript and not in the other, and so R sub 1 and R sub 2, they're going to depend on how many pieces are in the manuscript. But this would be the probability that some piece existing or not would be in-- would not be in any of two manuscripts.

So there's about 85 surviving medieval manuscripts in the period I'm most interested in, Italy from 1370-1420 just after the Black Death and during the time of the Great Papal Schism. So there's about 85 of these manuscripts, so what's the probability that any given piece would not be in any of the 85 surviving manuscripts? And it looks something like this.

So this is something that it's a big long number, but you can see that right at the very end, this-- whatever this works out to is the equal to the probability that a piece would be missing. Because if a piece is not in any of the surviving manuscripts, then it's missing to us today.

So this is great, but we know all these r numbers because they're simply the number of pieces in a manuscript today, not in originally. So we can take this, and we can go back to our original equation with three unknowns and substitute it into for that P of m place. So we can see that now we have one equation and two unknowns. It's still a pretty honking equation, but, well, that's good. One equation, two unknowns, still too many unknowns. We need two equations at least if we're going to do two unknowns.

But we can look at this last answer, this m. What are the total number of missing pieces. Well, the total number of missing pieces is just the number of pieces that originally existed, which we don't know, minus the number of pieces that survive, which we do know. If we have a pretty good catalog of everything that survives today, then we can figure that out.

So now we go back, and we see that we have actually just one equation and one unknown. It's a pretty big equation. In fact, a lot of the literature that I was reading from the '60s and '70s and '80s was about you can't possibly solve an equation of this dimensionality exactly. So here's all the sophisticated math that we can do to try to approximate it. But, in fact, nowadays it's pretty simple.

What we can do is we can try different numbers for N, the number of original pieces, starting from one more than the number of surviving pieces or something. Try all these numbers into here until the left side and the right side approximately balance. And we can do that with computers.

So when I did this, it's-- this is an article I think like 13, 14 years old, it's before Python was big. So I used Perl, and I just did find N, find the hypothetical total number of manuscripts given all the things that we know, and so on and so on. And we're able for each of four different genres-- these are different types of music-- we're able to figure out what percentage of the pieces are probably missing.

These are-- I'm-- there's error bars on this. It's rough guesses, but we can see that the percentage of missing pieces may be in only one-- in only one of the genres is it predicted to be more than the number of total number of pieces than the number of surviving pieces. And, in fact, for some of these genres, we predict that, hey, we have almost everything there.

So one of the things we can look at, instead of saying how many pieces were once copied but no longer exist, we can take that total number and say, wow, maybe we have all but 25% of them.

Now there were a lot of assumptions that went into that. Scribes are collecting things randomly and so on and so forth and that there's not pieces that are more popular than each other and so on. And so in the article that I wrote, there's ways that you can compensate for this and ways that we can test.

For instance, we can hold out 10% of the manuscripts at random and then say not just how many pieces would we expect to-- would we expect were lost. But if we gained 10 more manuscripts, how many more pieces would we expect to have that we didn't know already and then we can look at? Now put those 10 manuscripts back and voila. I think

I can't remember off the top of my head right now, one of those things I publish it so I don't have to remember, but I think that the percentage of pieces that were lost I think it was within 10% of something or-- so these models were pretty, pretty good.

So we can use this mathematical thinking and just a little bit of programming to answer questions that were thought to be completely unanswerable in history. Thanks.