Intro: When you take a magnifying glass and look at a picture, you see that it's made up of thousands of dots and when you pull back, those dots become the details that create the picture. There's real power in the details, because when you have more detail, the bigger picture becomes sharper and wider, and a story emerges.

Hi, this is Amie Moreno and you're listening to "Seeing the Big Picture: Conversations on how Data and Artificial Intelligence can add the Details that Fuel Deeper Insights in the Life Sciences Industry."

Amie Moreno: Hi everybody. This is Amie Moreno here with yet another one of our data-related podcasts. And today we're going to specifically discuss natural language processing and have a bit of an NLP one-on-one and go over the basics. So, my role at Optum is Director of Data, Advanced Analytics and Tools at the Life Sciences team, so with that note, I would like to introduce my colleague, Conor Wyand, who's going to chat with us about natural language processing. Thanks for joining, Conor.

Conor Wyand:
Thanks for having me, Amie.

Amie: Can you tell me a little bit about your role here at Optum and the types of work that you do?

Conor: Sure. So, my role here within Optum is I'm a consultant on the Data and Advanced Analytic Tools team within the Life Sciences organization and I support our clients to make sure that they're maximizing and deriving the most value from our data. And so, this could be both from a technical understanding of what's in the data, what the variables are, and how to best use the variables, but also to step back and think a little bit more high-level and conceptually as to what type of information is in the data and then how to best approach your particular problems and then create an analysis that's best suited for that particular situation.

Amie: Perfect. So, let's start with the basics, Conor. We hear interchangeably the term EMR or electronic medical record and we also hear EHR. Can you tell us the difference between those two terms and why it's important that the electronic health record is so significant?

Conor: Sure. Ya, I think that's a really great question and I would say that the electronic medical record provides a single view into patient care. So, if you think about the way that data and healthcare data is stored in the clinical sense, it's stored on an electronic system and many times, there's many different electronic medical record systems within a particular facility or even within a particular hospital. And to have data from one of those medical records, gives you, like I said, a single view into that patient's

care, whereas the electronic health record is a more comprehensive, more longitudinal view of the patient and their experiences in the healthcare delivery system.

Amie: So, that gives you a more complete picture of the patient, both in an ambulatory setting as well as what's occurring while they're in hospital.

Conor: Right, it gives you insights not just into the different levels of care that they're receiving, but the type of care that they're receiving, so certain diagnostic information, other procedures that they're having performed on them, medications that they're taking, specific lab tests that they're getting performed. We have the ability to see all of these different sources of information in one centralized location. And so, an example could be someone who has heart failure, who goes to their PCP and then has to move through the network to a cardiologist. If they're on two different electronic medical record systems, you may not capture all of that information, whereas the electronic health record is a more comprehensive system and source of information such that we will then be able to see the patient as they move throughout the healthcare delivery network.

Amie: So, that's important to all the key stakeholders, right? So, you're able to understand a patient and how they're being cared for and what the outcomes in terms of their health looks like. You are able to understand a patient in such a way where you can educate physicians on when and how to treat a patient. And then, of course, for folks like brand teams and marketers, to really get that full picture of your patients in that holistic fashion helps them with their brand strategies. Great.

So, let's dive into natural language processing or we'll call it NLP for simplicity sake. Can you give us a high-level overview of what natural language processing is and some about what it includes, what types of information you can extract from it?

Conor: Sure. So, natural language processing really is a set of techniques that utilize computers to take unstructured text, which would be words, phrases, sentences, and to extract individual components of those sentences to put into structured or data tables, variables, columns that could then be used for analysis. So an example would be a patient comes in and complains and says, "Dr. So-and-So, I have severe pain in my left knee." Well, if you have that entire sentence in the context of a paragraph in a Word document, there's no way you can use that in an analysis. And so what natural language processing does is it takes that information and now we have the ability to say that that patient has pain, they have severe pain, and they have the severe pain in their knee. And so, now you can create a structured data table, including all of the different individual

components of that concept of severe knee pain that can then be used in an analysis in any capacity.

Amie: So what other kinds of information exists outside of just the symptomatology? What other kinds of information can you cull from the physician documentation?

Conor: Sure, so the physician notes are really valuable in the sense that they capture any information that otherwise wouldn't be captured in a structured data field on the back end of those EMR systems that we talked about earlier. Throughout the industry, there are many different structured codes -- diagnostic codes or procedural codes that are used for billing and used to track the care that is given to a patient -- but other particular conditions may not have the corresponding codes, so a great example would be in diabetes. We know hypoglycemia exists within a diabetic patient, but it is very infrequently documented on the patient's chart. And many times, the hypoglycemia may not go noticed in the patient's chart, but given the fact that we have the ability to mine the physician notes for this type of information, we can now extract hypoglycemia as a concept in the physician notes. Other things could be information around behavior. So you could imagine, as a patient, is taking a set of different medications, they're constantly being prescribed different medications to find the ideal drug. And there's a certain reason why they're stopping or starting or titrating medications. So, we can see as the dialogue between the patient and the physician takes place, we can extract those different reasons for either the physician behavior or the patient behavior that really gives us the best insight into the decisions that are being made for the patient.

Amie: So, you really know why the doctor is treating a patient a certain a way. Could it be a side-effect, for example?

Conor: Sure, yeah, it could be a side-effect, it could be a cost issue for the patient, it could be a formulary coverage issue.

Amie: And that type of information -- correct me if I'm wrong -- is difficult or impossible to get if you're looking at, say, a claims data asset or even electronic health records, because you really are trying to understand why and traditionally surveys are used. So, in this capacity, would you agree that it's similar information to a survey, but you're getting it in real-time and on a much greater scale? Is that fair to assume?

Conor: Yeah, I would say that's fair and I would say there's a difference between the claims data, as the claims data -- like I was talking about the codes uses standardized codes that are widely used throughout the industry -- but with the emergence of this electronic health record data, we now have a source of unstructured text where there are certain phenomena -- many

different phenomena that manifest themselves in these physician notes and the only place that we can capture this information is from the physician notes. It really provides just a much more rich clinical understanding. I mean, it's the same kind of concept as why we look at the electronic health record versus the electronic medical record and the electronic health record now gives us a more complete picture of the patient and their movement throughout the healthcare delivery system, it's the same thing on the notes side, you know, we're using the physician notes to supplement data that we're extracting from structured data fields and in ways that we couldn't, you know, we're using data that we couldn't otherwise get if we were to only rely on the medical chart.

Amie: So when we talk about physician notes, I'm envisioning an actual written or something that's got a lot of language around it. It would seem that could potentially go against HIPAA regulations in that a patient or a provider, perhaps, might be able to be identified. What's the best way to ensure that that doesn't occur?

Conor: Sure, that's a really great question and privacy and compliance is, obviously, in the healthcare system, a very big deal especially with respect to patient-level data. I think one of the things that's really important is to make sure that you're statistically de-identifying the data and that you're aware of certain minimum thresholds that exist in order to protect the anonymity of both the patient and the provider.

Amie: Thanks. So, Conor, when you talk about the healthcare delivery system, can you give us a little more detail about that. Are you talking about kind of the care continuum?

Conor: Exactly, yeah. It's the care continuum, it's all of the different ways that a patient is receiving care. The healthcare delivery system is, I would say, just the broader set of services that are provided for the patient and as a patient moves throughout the system from one physician to another, from one hospital to another, what we're doing with the EHR data is to be able to track, at the patient level, focused and centered at that patient, to see how that patient moves throughout the system and what the different types of care they're receiving at different points along their patient journey.

Amie: Another topic that's really big in healthcare right now and a lot of entities are doing this is integrating data. And we know in order to holistically understand a patient, it's important to not only look at their clinical profile, but potentially social behavior, etc. When you talk about not having the identity of a patient, how can one integrate all of these databases in a way where we're not double-counting, in a way where we could follow the patient through the care continuum and ensure it's the same patient?

Conor: So I would say the best way to integrate different data sources would be to use -- especially if it deals with a patient -- is to use patient-level and patient-identifiable information. And so while many data vendors don't have PHI, they rely on what's called a probabilistic match. So they look at other aspects of that patient profile and then match across other data points. Really, the best way to do it would be, on the other hand, is a deterministic match in which you use patient identifiable information -- name, date of birth, social security -- to be 100% confident that the patient that you're referring to is, in fact, that patient and this avoids any other consequences with probabilistic matching, which would be double-counting certain patients or even omitting certain patient records. But really, the deterministic match is the best way to integrate data sources and to be holistically confident in your practices.

Amie: So, in order to deterministically match these patients and to integrate these databases, you mentioned that the best way to do this is with patients' actual PHI. Can that be problematic and how so? And what would you do to account for that?

Conor: It can be problematic depending on how you're matching and then the scrubbing after the matching and so, I think it's very important to first use the PHI to match different data assets and different patients within a particular data asset. And then after all of the matching has been done, now you can go through and scrub for any PHI to ensure that the data set is then HIPAA-compliant for research.

Amie: Right, and all those securities and compliance components and -- I'm assuming -- people, would be in place to ensure that that happens, and it does remain compliant.

Conor: Yeah, I mean, that's 100% necessary.

Amie: So, again, thinking about integrating databases, you mentioned when you perform a probabilistic match, there can be some consequences that could negatively affect being able to track that patient across the different databases. Can you talk a little bit more about what those are?

Conor: Ya sure. So I think some of the problems with probabilistic matching is that you're not accurately representing what's taking place. So, if you were to look at a particular patient cohort, perhaps the diabetes cohort, and you're using probabilistic matching to match patients across a data set, you might double-count many of the patients. So, now your cohort of 2 million patients has now just jumped up to 2.5 or 3 million patients which doesn't really give you a very realistic patient volume for that particular cohort. Another example could be that you're omitting certain pieces of

information. So, if you're joining together and integrating patients' diagnoses records of diabetes along with their lab values of blood glucose and you're now missing some of their blood glucose labs, there's now a hole in that patient's chart or that patient's health history or that patient's health profile, where now there's a gap in your analysis. And so, there's many other different consequences of probabilistic matching, but those are just a few examples of how you can then inaccurately represent that patient's experiences in the healthcare system.

Amie: Conor, tell me about some of the challenges that researchers face when they're working with natural language processing.

Conor: I would say some of the challenges as a researcher, you don't have confidence in the output, because you don't have the ability to see every input. Traditionally, natural language processing, in the industry, has taken more of a rule-based approach, whereas utilizing and incorporating statistical techniques, specifically machine learning, gives you the ability to not only extract information on a much more complicated level, but also the ability to assess the model performance and ultimately, the accuracy of that model, so that as a researcher, you can be confident in the data that you're using.

Amie: So have you used NLP for any specific projects and if so, can you talk to me about your experience?

Conor: Sure. One of the things that I was pretty closely involved with was a project centered around prostate cancer, which was really interesting because especially in the oncology space, much of the information related to patient care is going to lie within the physician notes. You know, many of these oncology centers and particular tumor types are so specific, that it's almost impossible for them to record data in the structured EMR systems. And while there may be EMR vendors out there that have oncology-specific data sets, they might not integrate other parts of that patient care throughout the healthcare delivery system like we alluded to or talked about before. And so, while we know that there is a lot of oncology-specific information in the physician notes, the question is, how can we extract that information and then begin to assess the accuracy of that extraction process? And so, the project I was involved in really took a deep dive into prostate cancer and some of the relevant characteristics or attributes related to prostate cancer and related to the patient, drug treatment and other clinical concepts.

And what was so fascinating, I think, about the project was not only could we provide really, really rich clinical information related to the patient, but also to see that the precision and recall or really the accuracy of the output was tremendously high. And so as a researcher, it's got to be very

reassuring knowing that the data that you're using is, in fact, an accurate representation of the care that was given for the patient and then can be used confidently in an analysis.

Amie: Well Conor, thanks a lot for being here and speaking with us. I feel like I have a much better understanding of how natural language processing works.

Conor: Well it was great to be here, thanks for having me.