AAAI/CCC Symposium on AI for Social Good

Talk Sessions 1: Healthcare
Session Chair: Dr. Eric Horvitz

Fei Fang: It's a great honor to have Eric Horvitz from Microsoft Research to chair the session. Eric has got his PhD from Stanford, so welcome back to the campus. He's currently the technical fellow of Microsoft and is also the managing director of Microsoft Research Development. I'll give it to Eric for the opening talk.

Eric Horvitz: Being involved in research doesn't mean that you can get PowerPoint to run. There we go, great. I thought I would just share a few thoughts about healthcare today. It's a passion of mine-

Speaker 3: Could you use a mic because [inaudible 00:00:51]-

Eric Horvitz: You're recording here.

Speaker 3: Yeah.

Eric Horvitz: Okay.

Speaker 3: [inaudible 00:00:54]

Eric Horvitz: Hello, hello.

Speaker 3: It's now on the speakers.

Eric Horvitz: Good, that's good, so I can just look like a rockstar and not really be using the mic locally here. When you're in healthcare, you see lots of low hanging fruit and possibility for applying statistics, the data that's becoming available, utility theory, decision making, models, to really deliver incredible value. The way I like thinking about healthcare is we have lots of sensitive data becoming available, we tend to build predictive models, and those predictive models for specific patients, distributions over outcomes, or hidden diagnoses that we can't see without invasive tests. Most medical decision making tends to stop here in medical work, like let's see what we can do with the distributions, but the rubber meets the road with decision models, typically.

Really, how do you consider costs and benefits. Like in other fields, we have a data to prediction to decision pipeline and once you have that, you can of course think the other way and what kind of data might we collect to make our data sets more complete in the context of delivering value to patients. Of course in hospitals and in the field, and in public health, we often have to really think about, it's insight building, not just autonomous actions or recommendations. And understand the role of a system when it comes to people and decision

support. Which also brings to the fore all sorts of interesting cultural issues and ergonomics of the field.

So what are opportunities? Well, there's incredible data locked up in electronic health records. I tend to look at EHR data as goldmines where there are lots of diamonds beneath the surface and you have to just get access to it, and that's not often an easy thing when it comes to working with hospitals, as people know. There's incredible amounts of online behavioral data, Henry [Couch 00:03:14] has looked at some of this through Twitter feeds and his students. Our team has looked at quite a bit of this data when it comes to our access to anonymized logs of search activity. You can imagine new combinations of these things.

The class work that has been done in this field and we did this work ourselves as one of the contributors in this realm. As a great example, is readmissions. A 2004 data set published in 2009, New England Journal article, showed that 20% patients, Medicare reimbursed patients, were bouncing back within 30 days, 35% within 90 days, an estimated cost of $17.5 billion of avoidable costs, it was thought in that time.

We took a large database of 20 years of data, 30,000 [binarized 00:04:02] variables. This is working with [inaudible 00:04:06] and Mark [Braverman 00:04:06], and others. And created standard set of ... A predictive model that had characteristics characterized by this receiver operator characteristic curve based on a train and test.

Of course these data sets also, when you do these analysis, give you a sense for some interesting observations that are discriminatory. Some are obviously, others like a complaint sentence anywhere in the text that says fluid is a bad sign for staying out of the hospital. Very interesting to the physicians.

We built something called Readmissions Manager, which shipped worldwide as a product in those days. What doctors would see is probabilities of readmission at discharge time and little explanations coming off the logistic regression, in this case, to make it understandable.

What I'll talk about here is the fact that we discovered that hospitals systems, like Midwest unnamed hospital system in the Midwest, were making decisions based on what they called the Microsoft score. So if the score was over ... It was a probability, but they called it the Microsoft score. It was over 25, they said, "We do a special program, we have a special policy on keeping patients out of the hospital, we give them extra support, we schedule them for outpatient visits." They had a program and it said doing this is costly, but it turns out to save them $1.5 million a year and also it's very good patients.

We thought about, wait a minute? How do you really do this? You don't just pick a number out of a hat and then try to use and look at the outcome. You

want to think about the decision problem here. So we looked at congestive heart failure, I'll make a quick framing case study here to talk about because I think it frames other work [inaudible 00:05:56] might do. It's a $35 billion a year management problem. These patients are often revolving door patients, they're on medications, special diets. If they take too much salt in their meals, they tend to get fluid overloaded and come into the emergency room, what's called a tune-up, it's about seven to 10 days and it costs about $18,000 or so, and it's dangerous. So we asked the question, can we actually predict, not just predict, which patients will come back, but also, if we could intervene and lower the probability of them coming back, what's the ideal decision threshold?

So we build a model. Actually, in the data set we had from the Washington Hospital Center in Washington DC, we actually had the cost of each readmission, so we can actually do a little decision analysis. Very simple decision analysis here. At the x-axis you see the probability of readmission given evidence, zero to one, and here's the full price of handling a patient who had been readmitted. If it's a .5 probability, it's the expected value of half that amount, and that's what that little decision curve looks here, or a line.

What if I actually had a magical treatment that could reduce the probability that that patient would come back and have some sort of a cost? Well, look at the efficacy here, we'll call it beta, it's going to reduce the probability of being readmitted, given evidence. What happens is, you put that in the decision model here, you say you start at zero, you're paying a price for that special treatment, and you have a curve that's going up a little bit more slowly with aggressive follow up than the standard discharge curve. These two lines coss at a decision threshold. The idea is, if you have a predictive model like we had built, you want to reason such that you choose your treatment based on whether you're above or below that decision threshold. That's the way that works. If you do that, decision theory tells us you will be ideally allocating resources to minimizing readmissions.

We have a basic, nice idea here I think that we are quite fond of is, we can actually now instead of just doing a standard [ROSC 00:08:19] curve with a test and train, we can now say, let's run a test and train on the decision model, the simulator, to see what would happen with hold out data, for example.

You can actually do studies. You can say, even if we don't know in advance the cost of some new program and how much it reduces the probability of admission, we don't know that, we can actually run a simulator. For any combination of dollars and return, reduction in readmission cost, we can compute what would happen.

Here's an example. With that hospital in DC, you see here at the bottom here, the proposed cost of an intervention, there's special kinds of programs, there's nurse visits, there's special kinds of devices you can wear, smart scales to capture fluid overload and so on. There's some assumed efficacy, how much it'll

reduce probability of readmission and we can run a simulator and it tells us, if you have an $800 intervention and it's 35% efficacy, you will reduce readmissions by 31% and you'll be reducing cost by 13%. Why is that the case? Because not all the time it worked out nicely, sometimes you'd pay money and they wouldn't have come back anyway, so they don't align perfectly.

What if someone tells you they have an $1,800 intervention and 20% efficacy, well, you lose money that way. Then I'll just end with this, because we're at 10 minutes now, end with this chart here which I think is really important for us to think about. This actually came in at our lunch conversation about simple rules versus more complicated decision analysis. What you see here is what I would call like a crystallography of intelligence. What's the value of taking an existing hospital system and with some sort of an intervention program, with an efficacy and a cost that exists, and saying, what could do I beyond a fixed policy?

Here, everything's expensive and it's not very efficacious, you would never use this program, you would never intervene. Over here, everybody you can compute in advance, would get this treatment. It's inexpensive and efficacious. But what decision theory and machine learning gives us is, it opens up this region of reflection and deliberation, patient by patient, personalizing the treatment, and it just expands, opens this area here, and that's where the value is delivered. So you have to ask the question, where do we want to use our AI systems, our decision theory offline, to just design ideal policies, boom, we have it. Proof. Proof. Where do we want a live system in the hospital that's buzzing, and aware, and working, patient by patient. Now, we have the ability to build simulators to understand all that. So I'll stop there, thanks.

So now, I get to chair the session and our first speaker today ... Oh, should I take a quick couple questions or just keep on going? No questions? Any questions at all about anything I said? Yeah, please?

Speaker 4: What are some of the dynamic approaches [inaudible 00:11:46]. You just mentioned that [inaudible 00:11:49] static case, instead of doing it offline.

Eric Horvitz: Oh, I see.

Speaker 4: [inaudible 00:11:57]

Eric Horvitz: To be honest, to get to the offline assessment, you need to sort of come at the from the [full bore 00:12:04] model approach, so you can sort of make these proofs and understand what it means besides going with a back of the envelope approach. Almost everything we've done in healthcare, and the other significant area I've worked in with [Jenna Weans 00:12:22] a former intern, with large scale electrotonic health records is predicting in advance of an adverse outcome, which patients are going to become infected in the hospital. By looking at many variables including some that are proxies for exposure and some that are proxies for susceptibility of patients.

In that work, we have actually a nice paper, I think it's nice, on getting it towards your question, which is, how much better do you do with a very, very heavy duty top of the line machine learning, opaque style machine learning algorithm, looking at 1,000 variables, versus 10 clinically known variables that are known in the literature for being associated with risk of hospital acquired infection? In this case, C. [Deficile 00:13:13], a nasty diarrheal infection. Just using that on various amounts of data and looking at the boosts you get from being deeper and smarter, and using variables that no one knew had an influence on this outcome. I'll stop there, but we can take this offline about the simple versus not simple.

It's interesting, the paper on readmissions that we published, it's online, compared the results and power of this method to back of the ... I call them back of the envelope, but they're standard scores used in the clinics right now. There's various Apache scores for trauma care, there are scores being used, the Toronto score for readmissions, and often they're designed to be back of the envelope, the kinds of functions that ... A list of well known observations that a doctor could make and their weighted scores, basically. They're typically, from my understanding, is they're not designed with machine learning algorithms, they're are conjured by experts per intuition and then they test these scores with data. Say, "This is how well the score works."

So we actually, in our paper, looked at the score versus our approaches, and of course to even validate these methods, you have to take the score an convert it to a probability of some kind, which is interesting work in itself, and then show how well it works versus your probabilistic methods. [Melind 00:14:44], yeah.

Speaker 3: This cone of intelligence, I guess I'm wondering if that cone overlaps with the cone of interpretability or cone of understanding. That cone might be a little bit sharper in the sense that it could be there are some decisions that are easy to explain and those are outside that cone, but there are some decisions that are in the cone that are going to be harder explain.

Eric Horvitz: Well, it's a great question, and before I get to that answer, I would first say ... We just had a meeting with the Berkeley Law Institute on Friday about explainability and new European laws, it's called GDPR coming to the fore, which is requiring systems that do any kind of automated decision making or recommendations to inform the people influenced by the decisions about the logic ... Meaningfully inform them about the logic of the inference.

So all this discussion, "Well how do you explain this to thousands of Europeans when you make a decision?" It's the law now. So there's lots of discussion about, do you back off and use a simple linear model, do you use new kinds of linear models that have what are called these generalized additive functions that have shaping functions in them? Recent work has shown you can get the full power of that machine learning, almost the full power, with explainable

additive models that have mostly single terms with some pairwise and some triplet interactions that capture some of the error.

The bigger question would be, for any machine learned model, can we go from the lesser performing logistic regression style models, the classic simplified models that we think are explainable because they tend to let you see what's moving ... What single pieces of evidence would be holding everything else the same. So people consider that a good version of explainability. As you take them forward to more complex, rich models that still have that power, yet still retain some of the explainability properties.

I'm not sure if I can map that answer, which I know pretty well with the work right now, onto the cone. Certainly, you can explain the decision theoretic cost benefit arguments in the corners by just saying it's not worth it and show the cost. I will take your question as an interesting conjecture for study.

It's an honor to have Lanbo She here talking about teaching and checking of constraints for surgical tray layout.

Lanbo She: Yes.

Eric Horvitz: Lanbo, you're at Michigan State and this is with Jonathan [Connell 00:17:48] who is at IBM TJ Watson Research. This one seems hooked to a microphone, to a speaker, no? I'm just hearing echoes. I'll sit over here [inaudible 00:18:00].

Lanbo She: Good afternoon everyone, my name's Lanbo and our work is about teaching and checking of constraints for surgical tray layout. Our original goal was to build a computer agent that can on one side learn from human experts through language interaction, and on the other hand, it can acquire knowledge to teach or supervise those non-expert humans to accomplish a task. I took this idea with my mentor, Jonathan Connell, and it happened to be related with one of his projects, which goal was to build a cognitive operating room.

Basically, for operations, one very important step is to have a good preparation of these instruments. Basically, before any operations, a well trained surgical assistant needs to get out the different kinds of instruments that will be used during the operation and placed on the tray according to certain orders, or regulations. Here's a comparison between a well organized tray and a disorganized tray. Basically, as we can see with this well organized tray, the doctors can quickly find the instrument they want and quickly catch them. This will save a lot of time for the surgery, which will be very important.

Then, why do we need an automated agent to help with the preparation and supervision? Firstly, even for a well trained surgery assistant, it may require a lot of memorization about different types of instruments that will be used for different operations. Sometimes even human can make mistakes. This time, the automated agent can check the [inaudible 00:20:31] environment that already

set up by the assistant and then give suggestions if identifies any mistakes or violations. Secondly, in some situations the well trained assistants can be short handed. For example, in emergency situations or even the battlefield. In this case, the doctor may just need someone, maybe not well trained or an expert, to help with the preparation process, but they may not know how to do it or cannot do it properly. In that situation, such an automated agent can give suggestions or tell them what are the instruments that are needed and how to place them properly, and maybe even give step by step instructions. Last, these automated agents can also be used in schools to help train the students.

Here's one example of our integrated system that checked the environment. Specifically, we have a camera mounted overhead to look at this tray area and then analyze the setup, and compare his knowledge to see whether there's any violations and give feedback to the human. This is one example here. In this case, the agent identifies multiple violations. Firstly, it tells the human that this scalpel at the top left is too close to the scalpel at the top right. At the same time, the related objects will be highlighted on the screen. Based on these suggestion, the human can fix this setup. Where we can see these two objects are a little bit far away from each other. Then this time, the agent will check the environment again, and now it identified another problem where these two scissors are also too close to each other. We can see they are kind of overlapped. Then the agent will tell the human, "These two scissors are now too close to each other." After the human fix this stuff, the agent will say, "Now, the tray looks good, I don't have any suggestions for now."

Here's an example of how the agents learn new constraint setups from the human expert. In this case, the well trained assistant teaches two constraints. One is, we need three scissors in this surgery. Another is, these should be groups together. The agent will analyze these two sentences and then formulate the corresponding constraint knowledge which are represented by [inaudible 00:23:54] sentence. Then with this newly acquired knowledge, it can directly apply it to check the current environment. In this case, it identifies two scissors on the tray and it says, "There should be three scissors on the tray." The human put another scissor, which is far away from the other scissors, and then the agent will say, "The scissors at bottom right should be grouped with the other scissors." This is how the agent learns new constraints and apply it in the new situations.

The implemented system is the integration of different modules, which include language understanding, [inaudible 00:24:40], knowledge recognition, dialogue management, and language generation. Actually, each of these modules by themselves is a big research topic and here we merely focus on the integration of everything and then have a [inaudible 00:24:54] system to realize the functionality of learning and checking the layout.

At last, we introduce a cognitive agent that can give advices and assistant non-expert human to prepare the tray setup. Secondly, the agent can learn new

setup constraints through dialogue with an expert. Thirdly, the system can integrate multiple types of object properties in language like the object types and spatial descriptions. At last, this is just one example of a more general class of problems where a cognitive system can supervise non expert human to accomplish certain tasks.

So that's all for my presentation, thank you.

Eric Horvitz: A couple minutes for questions. Yes?

Speaker 6: What is the responding time of this system?

Lanbo She: Currently, we just have a [inaudible 00:26:08], so the response is very fast, even realtime.

Speaker 6: Okay.

Lanbo She: Yeah.

Eric Horvitz: I'm curious how you chose this [inaudible 00:26:28].

Lanbo She: That's a good question because what I did back in university is about human rapid interaction and that's the sort of dialogue. We also focused on learning or acquired knowledge, and when I did my summary intern at IBM, what I did back in university and then we thought, can we move it a little bit forward? Instead of just learning from human, we can also teach non-expert and then we came out with this idea.

Eric Horvitz: That's fabulous. Fabulous direction [crosstalk 00:27:03]-

Lanbo She: Yeah.

Eric Horvitz: ... lots of implications. All right, thank you very much.

Lanbo She: Yeah. [inaudible 00:27:09]

Sujoy C.: Good afternoon everyone. Today, my talk is on in search of health doubles. This is not my work, it is work by [inaudible 00:27:23]. He's an assistant professor of department of IT, [inaudible 00:27:29], I am presenting on behalf of him.

In this work, [Arturo 00:27:37] wants show how the health parameter, how the problem of finding similar persons whose health parameters predominantly matches with other persons in the world. So health doubles becomes a challenging problem because the health parameters are dynamically changing.

To better explain this in great details, let's supposed in one fine morning the doctor tells someone that he has been diagnosed with a rare disease. He also

admits that we don't have much knowledge about this disease. It cannot be cured unless the disease can be studied in greater depth. So, he wanted to find the similar persons who have same disease. In this work, outsourcing can help us to find out the similar person whose health parameter predominantly matches with similar persons so that this can be helpful to find health doubles.

These are some schematic view of this work or problem. There are three types of features, there is static features, dynamic features, and medicinal features. So static features is basically the age, sex, and blood group, et cetera. Medicinal features are basically the history of vaccinations, medicine history of persons. Dynamic features deals with some of levels of blood glucose and pressure. These are basically dynamically changing. It is very hard to find out the similar person whose health parameter are dynamically changing.

These are some quantifying formula to find out the relation between human and objects, human and features with respect to a particular time instant. Because as the health parameter are changing in time to time, so the person, similar person, are also changing for a particular person, for a specific feature or for a specific disease. This is the [inaudible 00:29:53] denotes the association, strength of association, of a particular human being, "i", and features, "j". This is the association, the this is [inaudible 00:30:03] vector for all of these features. That is, [inaudible 00:30:08], that is HK means word vector [inaudible 00:30:11] human beings. So these are the all feature vectors.

To find the association of different feature vector, this a S HS, [HUA 00:30:23], this is the cosine similarity. Here, you can see there are two types of features vector, there is positive features and negative features. Here you can see that some vaccinations is a positive features and negative features can be a disease. Human disease can be a negative features. So before comparing this similarity, we need to care about what are the positive and negative features to combine this similarity.

Now, this is quantifying the similarity score of, final score, by which we can find the association of two human beings of a particular feature at a particular time instant.

To view consider this problem in a global view, [Arthur 00:31:13] shows that this problem can be treated as an [inaudible 00:31:17]. In this [inaudible 00:31:19], the notes are basically the crowd worker who are processing the information from one person to another persons. So the challenge is that as the parameters, dynamic parameters, are dynamic, so that's why it's very challenging to find out what will be the ... How the information will flow from one layer to another layer.

So there are three steps. There is a knowledge extraction, learning, and search. This is the search for health doubles here. You can see that if we consider this node as a crowd worker and if each of these crowd worker ... The [inaudible

00:32:02] basically to flow the information in first. So everyone should communication with other. In that case, if there are "M" number of crowd workers, it will be MC2 number of combinations, so it will be order of M squared.

But in this work, Arthur wants to show that if the hierarchical nature of graph can be considered and crowd workers can be layers as a hierarchical way, and then the information can be flowed through hierarchical way, and this can reduce the complexity in order of M [inaudible 00:32:38]. That is height of this tree.

The major challenges is that because of the diversity of parameter and health condition of different people, it is very hard to find the health doubles. We realize that the success of this proposed method entirely depends on the involvement of crowd workers, so we need more involvement, more input from this crowd worker, to flow this information very fast.

These are the conclusion that it's very, really challenging to find out the health doubles because it can solve, say for example, for a particular area, if we need the blood group of a persons to know the information of this blood group. So it will be very helpful if the health doubles can be decided in advance. The challenge is that there are still some problems that we need to address because the information flow requires more number of input. These are the [inaudible 00:33:44] references. Thank you.

Eric Horvitz: Any comments? I know there are some health double data sets, right? Because-

Sujoy C.: Yeah. For this time being, as the work is in ongoing stage, so I think that other is working on it, so it will be better to [inaudible 00:34:13] data set, but my knowledge is that the data set is not publicly available until now and he's [inaudible 00:34:21].

Eric Horvitz: Right, but I meant like Washington State has a credible twin data set that they put together over the years.

Sujoy C.: Yeah.

Eric Horvitz: There are some rising sets of data sets of twins for studies of influence of environment on people, by understanding people who have at least the same genomic foundation. It's interesting to think about even maybe the relevance of that kind of gold standard in training up your models, for example. Any other comments or questions? Okay, great. Thank you very much.

Sujoy C.: Thank you.

Eric Horvitz: Sorry for the wires here, be careful. Next we have Arnaud Delaunay working with Jean Guérin, wandering detection within an embedded system for Alzheimer's suffering patients.

Arnaud Delaunay: Hi everyone, my name is Arnaud, I'm really glad to be here to present our work on wandering detection within an embedded system for Alzheimer's suffering patients. I've worked on this topic with Jean Guérin who is the CTO of Co-Assist. Co-Assist is the startup developing connected watches for Alzheimer's suffering patients.

For my part, I'm student at Ecole Polytechnic and [inaudible 00:36:01] in a company called [inaudible 00:36:03] in Paris. Today, I'm going to present you what is wandering why is it a danger for Alzheimer's suffering patients. The state of the art on this topic of detecting wandering behaviors, present Co-Assist, the watch, and then the solutions, our resolutes.

The wandering problem for Alzheimer's suffering patients, what is wandering? Well, it's just when your lose your sense of direction and it's due to dementia pathologies and Alzheimer's is one of dementia pathologies. It's really, really frightening for the patients who is losing itself outside, outdoor, and for the relatives, as well. It's mainly due to the inability to recognize your surroundings and when you have memory loss. So sometimes it can lead the patients to really serious risks, to be lost outdoor.

What solution exists today to counter this kind of wandering behaviors? Well, I would just speak about the most famous one, which is geo-fencing. The idea of geo-fencing is just to trigger and alert each time you go beyond a boarder of a predetermined zone. So the pros is, really easy to implement, it works well, it's precise, it's secure, but one of the limit of this solution is that it affects the freedom of the people because it's like a virtual leash or a virtual jail for the patients suffering from Alzheimer's.

Many people has worked on the topic of detecting wandering behaviors and I just want to stress two main groups of people who have worked on it. [inaudible 00:37:49] in the early 90's and [inaudible 00:37:51] in 2007, have worked on the classification of the different patterns of wandering. So here are the different patterns, it's like pacing between A and B, doing returns. Having a random trip between A and B with a lot of change of directions. Or lapping around the points.

These are the key notions they figure out and another group of researchers in 2012 built an algorithm to detect wandering behavior based on GPS traces. So they used that asset from Microsoft Asia called Geo Life, with a frequency of one point every second. What's the definition of sharp points? It's when you change your directions and counting the number of sharp points within a distance range. They were able to detect a wandering behaviors like you see on these kind of trips.

For Co-Assist, the solution we wanted to build was ... The aim of the system was to be embedded in a watch, in a connected watch. You see the watch they are building for Alzheimer's suffering patients, [inaudible 00:38:58] and with also some full detection system in it.

The question is now how we've implemented a detection solution that runs in the embedded system. Well, just let me recall the goals, it's really important to understand that we wanted to have [inaudible 00:39:15] for the device because we cannot imagine selling a device to people suffering from memory loss, a device that you have to remember to recharge every night. Well, it's not possible to sell this kind of product to Alzheimer's suffering patients, so the autonomy was the key points of our goals. So the approach is to answer that was to lower the GPS core frequency. We used the acceleration to set different modes, whether we are in the house or outdoor. When we are outdoor, we use the maximum frequency of one [inaudible 00:39:49] every 30 seconds to keep a longer autonomy for the watch.

Another goal was to have the computation embedded, so using a light computation pipes, we [inaudible 00:40:02] or use really [inaudible 00:40:04] model inside the watch and still have a precise and sensible detection solution.

Our work, it's like an approximation of a [inaudible 00:40:15] model. Meaning that we use the state of previous positions to compute the state of the new one and we check the states. If the state is beyond the thresholds, we trigger the alert. Basically, you can seen an example on this trip where, for example here, there is a change of directions, but these people has been working for a long time, so it's not trigger the states to increasing the stats. Here, for example, you have a lot of change of directions in a small distance range, so it will increase the states, and when the states pass beyond a threshold, the alert is triggered. With this kind of algorithm, you have to define four main parameters, and that's when we use a training set and a testing set to define these parameters in order to have an optimized model, precise, and sensible. So optimized for [AUCT 00:41:12] score.

The results, we have to understand that for the process, for the test process, we have in [parallela 00:41:19]. The test for the scores, the metrics, of our model, M batter life test to be able to have a longer life expectancy for the battery in the watch. So we build a [inaudible 00:41:36] using the algorithm of [inaudible 00:41:39] researcher group. We split it in training and testing tests. We lower the frequency to have a frequency for our model embedded in the watch and [inaudible 00:41:49] implementation of our solution. We had good metrics compared to the work done by the researchers in 2012 and in [parallela 00:42:01] we can see that the estimated battery life [inaudible 00:42:03] 21 days for a connected watch for people suffering from Alzheimer's. It's well enough for them to wear this kind of watch. Here, we can see an example of how we tried to optimize the thresholds, the parameters you saw before.

So, what's next? Well, we thought about today in the algorithm, we just increased by one the states, but we can use a continuous increment of the states, meaning more like predicting the level of the states could be like an increased improvement in our model. We want to try to ... It's always the problem to have [inaudible 00:42:53] data sets, so using a custom [inaudible 00:42:56] training sets with really data coming from Alzheimer's suffering patients wearing the watch we built. At least at the end, release the beta, including the detection system within the watch. So stay tuned and thank you for your attention.

Eric Horvitz: [inaudible 00:43:18] audience.

Speaker 9: How will you be using the signal from the system? What intervention will it drive?

Arnaud Delaunay: From the system, we have GPS antenna, which is giving us the GPS points-

Speaker 9: I meant the decision, if it tells you the patient's wandering [inaudible 00:43:40].

Arnaud Delaunay: Yeah. What happens when we trigger the alert?

Speaker 9: Yeah.

Arnaud Delaunay: Yeah, okay. There is a chain of relatives ... Well, the chain can be decide by the patient and his relative and it's a chain of people who are alerted by the alert. So maybe, for example, the chain be at first you have to call my son, and then if the son doesn't answer, you can call the, I don't know, the closest house helping [inaudible 00:44:14] patients, and if he doesn't answer, we can call an emergency solution. So there is a chain of people triggered by the alert.

Speaker 10: A follow up, a key who is false positives.

Arnaud Delaunay: Exactly. We cannot have an unprecise model for that.

Speaker 10: [crosstalk 00:44:36]

Arnaud Delaunay: Yeah, but we still want a sensible model because we cannot sell a product saying, "Okay, don't worry, we'll detect when there is wandering," so the model still has to be sensible to not let pass any real true positive wandering case.

Eric Horvitz: Way in the back, first, and then behind you, then you, and then the third person here. Thanks.

Speaker 11: That was a nice talk. In the approaches for detection of wandering, one of the drawbacks of the first approach you mentioned is it limits the freedom of the person. So I was wondering in the second approach, the one that you detect the turns, the sharp turns, aren't you still limiting the person's freedom?

Arnaud Delaunay: That's a good question. What is the freedom of movement for the people and where are the limits of freedom of this movement? Because we are targeting people suffering from a disease and the symptoms of this disease can hurt these people, so at the level or at another, you have to watch these people to help them if there is a problem with his current position. Some patients suffering from Alzheimer will tell you, "Okay, I don't want people to watch me at every second," so the algorithm where just the relatives have access to the GPS points is not a good solution for them. Other people just will say, "Oh, I will be more assured if I know that someone is watching on me every time and I want this kind of solutions." You always have limits on freedoms, but to answer your question, compared to geo-fencing solutions, enable the people to move wherever they want around the world, and it will be still detecting wandering situations.

Speaker 12: So, there seems to be a bit of dichotomy between global solutions to this problem, which is the geo-fencing, versus the local solution which is the personal device that's on the wrist. Have you put any thought into actually trying to hybridize this where it's not just one or the other, but some combination of such? Where if a person is wandering outside of their typical area, it is an indicator of a more extreme example. [inaudible 00:47:15] in the last question, but there was a lot of limiting factors, either one, you're making trade-offs, is there a way where marrying both systems gives you more flexibility and more freedom?

Arnaud Delaunay: Yeah, I think it's a great remark and we haven't thought of this kind of combination. I think, yeah, there is some place to work on that. Yeah, for sure. Thank you.

Eric Horvitz: Wait, we have another question. Don't get too excited just yet.

Speaker 13: What about privacy, in danger of hacking by criminals who then know that the old people is not at home and all that kind of stuff?

Arnaud Delaunay: Actually, there is no communication between the watch and the international, all the computation are done locally, and it only trigger the alerts whenever the states is higher than the thresholds, so there is no communication between the watch and the cloud, if it's not a wandering behavior. Yes?

Eric Horvitz: I have a quick question, a follow up on two questions ago. I'd be curious to understand if you've done some failure analysis. So when you have a false positive, to understand what was going on, maybe it's something about landmarks that would help you to reduce that false positive rate? There actually is something interesting that you considered to be wandering, but actually was goal directed, have you looked at that?

Arnaud Delaunay: Yeah, it's a quite difficult question because the data sets we used for our training set and testing sets, is not only wandering people, and it has been handle labeled by doctors and researchers. So we-

Eric Horvitz: Can't answer it, yeah.

Arnaud Delaunay: ... wouldn't be sure of being really a false positive about wandering or not, but it was just like we were using these kind of patterns-

Eric Horvitz: Right.

Arnaud Delaunay: ... to say, "Okay, this is the wandering patterns."

Eric Horvitz: It'd be interesting to-

Arnaud Delaunay: Yeah.

Eric Horvitz: ... if you actually did have actual ground truth-

Arnaud Delaunay: Yeah.

Eric Horvitz: ... and to look at failures there to see what they actually are. It'd be interesting to know-

Arnaud Delaunay: Yeah.

Eric Horvitz: ... to help you out with that.

Arnaud Delaunay: Yeah, for sure, but this is a [inaudible 00:49:22] problem, like having the ground truths on the sets.

Eric Horvitz: Right.

Arnaud Delaunay: Thank you very much.

Eric Horvitz: Thank you. Now we have a paper by [Bolee 00:49:34], Yevgeniy [inaudible 00:49:34], [inaudible 00:49:36] Lee, and Bradley [Malin 00:49:37]. Eugene's going to present on sanitizing large scale medical records before publishing, which addresses a really interesting problem with HIPPA and so on. Thanks.

Yevgeniy V.: Okay, thank you for having me here. Yes, this is about AI for social good. One of the arguments is that the increasing amounts of data promotes a tremendous amount of social good, the challenge is the privacy challenge. A lot of this data, especially in medical domains, has sensitive information. Things like people's name, their address, social security numbers, and so on. This talk really is broader than that, but the focus is on medical and clinical domains. Specifically, things like clinical notes, which are inherently hard to share because the

information in those clinical notes is not structure so it's not directly labeled. In structure data, you can just remove the name column and whatnot. In clinical notes or in any text data, that in itself identifying what are names or what is sensitive information itself, is a complicated task. That's what this talk is about.

There are a host of approaches for doing this at scale using machine learning. The setup is fairly straightforward. So you have some human labelers that identify, here's a name, let's call this plus one, here's not a name, let's call this minus one, and you create these dichotomies of labels. You have this labeled data and you use your favorite machine learning tool, commonly it's something like a CRF or NLP domains, and try to predict what is sensitive information that you can subsequently extract from this data.

Learning is great, but just like any tools we develop, it makes mistakes. Two kinds of mistakes, to be precise. False negatives, or under reduction, is the things that are actual names or sensitive information that you don't catch, and subsequently when you share that data, you leak it to the public or to whoever you share it with. False positives are the over reduction, things that you suppress, even though it has no sensitive content. The reason that's bad is because you're suppressing information that could potentially be useful in analytics. So you want to suppress as little as possible, but of course you want to suppress the sensitive stuff.

So that's really the challenge. The privacy part is the first one. You want to be sure to suppress to an acceptable level of risk, the stuff that's actually sensitive and not leak it to the public. So that's the challenge.

In prior work, implicitly the threat [inaudible 00:52:06] prior work or these machine learning approaches, implicitly the threat model has been that there is some human adversary who is trying to actually reading these documents and identifying what is the sensitive information. Now, in reality, human adversaries can also use machine learning and that's what this talk is about, how can we quantify this kind of adversarial behavior in a precise way and build a threat model out of this?

Here's a threat model for an attacker that would also use machine learning to facilitate discovery of sensitive entities. So let's just say that you've released some amount of data after you've applied machine learning to it. We're going to assume that the adversary has some budget for manual inspection. What does that mean? They take, they do some pre-processing using machine learning, and then they manually verify whether some entities that are highly prioritized are actually sensitive or not.

The approach is as following, that was posit the attacker would use. So they run their machine learning model, let's call it [laronic 00:53:01] classifier H, to predict sensitive entities now in a data that's actually been published, that they have access to. They rank the predicted positives prior to predicted negatives,

which is presumably why you would use machine learning to begin with. Then they go up until their budget, and grab and inspect manually those instances that are ranked in the first B spots.

Now, we endow adversary with two superpowers. The first superpower is perfect verification. If they see something, they know if it's an identifier or not, if it's a sensitive entity or not, that's one. The second one is optimality in the machine learning sense. We assume here that, at least for purposes of analysis, that the adversary can actually learn the best accuracy learning model. In practice, you can kind of simulate this very easily by just, again, having the adversary label the data manually or a portion of the data that they receive, and sort of learn the model this way. But for analysis purposes, assume that they are optimal in accuracy sense.

So let's now revisit the typical framework that people use to apply machine learning to sanitized data, in this context. So step one, to publish [inaudible 00:54:06] as a classifier, let's call this H. They remove all predicted positives and then release the data. Now comes the adversary, they act according to the threat model I just described, which I'm not going to describe again. Now, here's the challenge, and also the opportunity. The challenge is that now you potentially have given up some sensitive entities adversary can identify. The opportunity is the same way, you can actually, as a publisher, simulate what the adversary would do, and then use that to judge whether you should do something else before you release the data, or just let the data go.

That's the key insight. In effect, we can do this in an iterative fashion which gives rise to a fairly straightforward, [inaudible 00:54:47] algorithm, or we call it [greedy 00:54:49] sanitize. The idea is as follows. You start with some data set, you learn a classifier H, you remove the predicted positives, the predicted sensitive entities, you learn again a classifier and its restricted data. You have some new predicted positives, you remove those, you keep going. Now, obviously if you keep going like this forever, you're going to remove everything, so you've got to stop at some point and the key question is, when do you stop?

In order to answer this question, we need to more precisely define what it is that you are trying to actually accomplish, which is to save publisher utility, as well as the losses. So we assume that whenever you publish the data, anytime you release something, the publisher gains C for every non-sensitive entity that's released, and loses L, which is some quantity we specify a priori, for sensitive entity that's actually being leaked. That's identified by the adversary.

So publisher's goal is going to be now trading of these two things. You want to share more data, that's the C part, but at the same time, you want to limit identifiable entities that you leak, that's the L part. What do we do? We [inaudible 00:55:57] this [greedy 00:55:58] iterative algorithm, essentially one of the marginal benefits, no longer outweigh the marginal costs. Which I will call

the local optimum because we're doing this greedily and we stop at the first chance we get. So it's local optimum.

Now, it turns out that we can characterize in terms of false positives and false negatives of the simulated adversary model, hence the subscript A, when do we stop? It's basically, false positives are too many relative to the true positive here, and the relationship here is L over C, which is sort of intuitive, the relative value relative to cost over data. Now, something you might be wondering, will Greedy iterate, will actually do this forever without stopping? It's actually not hard to see that that's not the case, it will stop at most "N" iterations. "N" is, in general, these domain's fairly large. In our experiments we show that we actually do much better than "N", but we'll come back to that.

Arguably, the key result is that whenever Greedy sanitize terminates, we can actually [inaudible 00:56:58] the number of true positives we leak. True positives in the machine learning sense in terms of simulated adversary and is bound as a function essentially of our cost benefit trade off, the C over L, which is key. So the more we care about the leaked sensitive entities, the more tightly we're going to bound this, which means basically the more iterations you're going to run this algorithm for, before essentially the noise is going to overpower the signal.

Now, you can imagine as a function of adversary's budget, the other aspect that you can consider is that this isn't enough. The adversary, if they have infinite budget, they'll actually inspect everything, true positives as well as everything else. So you need something else to meaningfully bound exactly what the adversary would do. A meaningful baseline is, what if the adversary just picks a random subsample, essentially ignoring the fact that they're using a machine learning algorithm. Let's call this the baseline. So we showed that essentially when budget is larger than the minimum amount of budget once they start going beyond sort of the positive count. At that point, they can do essentially no better, and not much better than just randomly sub-sampling. So the use of machine learning, the value of machine learning, for the adversary, is pretty minimal at this point.

Evaluation, we used four data sets, two medical, two non-medical, and our goal here is to suppress names, which is one of the reasons we could use things like n-run data set and newsgroup data set because they actually contain names, which we can label. I won't specifically get into the exact classifiers we use. CRF is one of the best ones for this task, so that's one of the main ones we used, but a few others we used as well.

Going back to the question of true positives. We have a bound that's nice, how does it work in practice? It works much better than the bound even would suggest, this is what the left part shows you, L over C. Basically, it says that you don't need to care about sensitive entities that much before you're not leaking much in the way of true positives. So that's a strong result. The baseline, the

dotted lines here, you see the horizontal, that's if you only do one iteration, which is what people would typically do now. So having multiple iterations is enormously helpful, but as you'll see in a second, you don't need many of those, you only need a few.

So, here's the number of iterations on the left. Again, this is a function of L over C. If you notice, the numbers on the horizontal axis, which is the number of iterations, are small. The data sets are in thousands, but we're only running it for five iterations at the most, usually a lot less than that, on average. That's very good news. Essentially what this says is that, it doesn't take a whole lot of iterating before basically you get so much noise that there's not much value in learning. That's what this is telling you.

What you see on the right is the fraction of the data you actually release. So that's the positive side. You want to share as much of this data as possible, and horizontal axis, again, is this L over C factor. As you go to the right, you care more about not leaking sensitive entities. So here is where you see the baselines are starting to do really badly and the interesting thing is the rightmost approach is our approach, and it's interesting that you're suppressing a whole lot. So essentially what it's saying is that you're balancing this pretty well, you're suppressing very little, but what you're suppressing is actually really important stuff. You're not really over-redacting too much.

To summarize, we've developed a iterative method for sanitizing data at scale. It uses formal machine learning based attack model in the loop, and allowed us to improve approach of better machine learning approaches for entity resolution as kinds of tasks like this come up. This is actually something I didn't really spend a lot of time with, but because we're using essentially the attacks that are simulated in the loop, if somebody comes up tomorrow and says, "I have a better attack model that is able to re-identify these things much better," we can actually throw that into the loop and this algorithm gets improved as the new attacks get developed. Which is really nice property.

It enables formal guarantees about privacy. Of course, the guarantees are within the framework we use, they're not broader than that, but our framework at least allows us to state these things precisely. It's extremely effective in practices, as I've showed. Thank you.

Eric Horvitz: So just to start things off, I'm curious, two things came to mind. Actually, several things, but I'll start with the first two. If you look at HIPPA law and HIPPA regulatory activity, and the way they specify different levels of access, there are different patterns of what is allowed to different kinds of researchers, given to different kinds of IRB approval, and so on. Not that we have the level of names versus no names, it's like name, zip code, age, you have these sets, constellations, of features and it'd be kind of interesting to take your method to one of these standard, templatized patterns of levels of access and see what

protection we take at level one, level two. They actually have names, but have you looked at that at all?

Yevgeniy V.: A few things. Not a specialist in HIPPA specifically, but per your question, there's sort of several kinds of things-

Eric Horvitz: Right.

Yevgeniy V.: ... that you can go ... Kinds of approaches you can get at HIPPA. One is essentially a rule based approach, which is what I think you are alluding to, this is the safe harbor approach where you suppress, at this level, these particular attributes, at this level, these other particular attributes. This is really designed for structure data because it's presumed, to begin with, you know exactly what those attributes are. This is already a bit step removed from HIPPA because it's unstructured and so even discovering those attributes is inherently hard. So, in HIPPA, those attributes that we're talking about, things like quasi-identifiers, for example, and safe harbor, these are actual identifiers. Even knowing where the names are of the patients here is a priori [inaudible 01:02:56]. In HIPPA, this is already actually a challenge. [crosstalk 01:03:01]

Eric Horvitz: I guess what I was saying was, you take those will based policies, which apply to structured data, and you have frames like names. So you say, "Okay, I'm going to deal with names per my discovery methodology here and attack and protect." You also have things like addresses, and zip codes, and ages, you can take some of these specifications and map it to this kind of work in unstructured data, and then call it proxies for protecting at level one, level two, level three.

Yevgeniy V.: Yeah. What I wanted to say is, once this problem is solved, so the next step would be exactly what you're suggesting, right?

Eric Horvitz: I see.

Yevgeniy V.: So this first step is just identifying where those things live in the data. Just another part to it, so there's safe harbor is one option, the other option is basically risk based. You have the expert certification that the data is low enough risk to be shared. This really is much more in the spirit of the risk based because if you go by something like safe harbor, [inaudible 01:04:00] clear that you would legally be able to do this because again, you can't guarantee that you've removed all the names.

Eric Horvitz: Right, interesting. [crosstalk 01:04:08] question was, there's a bunch of machine learning going on these days with unstructured medical data. It'd be interesting to know what kind of hits you take in a particular diagnostic challenge, for example, a predictive challenge, at some level of protection. [inaudible 01:04:25] trades in terms of, for example, AUC curve going down as you get better and better at protecting against X, Y, and Z. So as your risk goes down, what does the AUC do?

Yevgeniy V.: Yeah.

Eric Horvitz: To actually do a study maybe for a specific entity or diagnostic challenge.

Yevgeniy V.: Yeah, it's a good question. This is just focused on names, but there have been other studies that are related to this looking at the different things. Also, this is just a very limited snapshot of the research. We've done things like also cost sensitive classification, as well, that allow us to kind more finely map out these trade offs.

Eric Horvitz: Now we have some questions. Yeah?

Speaker 15: [inaudible 01:05:06] small. Just to clarify that.

Eric Horvitz: What's small?

Speaker 15: The hit you take in detection of other types of entities, is small. So there was a study that came out of Cincinnati Children's where they were re-running the medication extraction program from Texas, the [Medex 01:05:21] method. They showed that just after a single iteration, which is where you get a lot of this, the AUC for the detection of medications, the actual drugs and the dosing, actually didn't change. So, it's usually not ... You don't get a lot of confusion between clinical terms and potential identifiers in the medical records.

Eric Horvitz: Yeah, that's actually a very interesting comment. It's related to it, but not exactly what I was getting at. I was getting at this idea, if you have an aspirational target which is not in the data anywhere, like the probability this patient will come back and be readmitted for a condition of heart failure in 30 days, and you have access to all this data and you have the algorithms chug along and they're considering effectively thousands of variable. Let's say every single word being present or absent in the document. Then you say, "Well and that's protect in this way," it'd be interesting to see if there's an effect and to discover what the terms were that were actually useful in giving you that extra lift for which you remove them, give you kind of a loss of accuracy in a diagnostic task or predictive task, which is not necessarily based in entities.

Yevgeniy V.: I think I understand what you're saying better now and the short answer is [crosstalk 01:06:39]-

Eric Horvitz: Sorry to be so garbled before.

Yevgeniy V.: No, well, it's okay, I just misunderstood.

Eric Horvitz: Yeah.

Yevgeniy V.: The short answer is, we just haven't looked at that-

Eric Horvitz: Yeah.

Yevgeniy V.: ... specifically for this kind of data.

Eric Horvitz: Yeah.

Yevgeniy V.: This is definitely something that's on our plate.

Eric Horvitz: Yeah? We have a question-

Speaker 16: Hi, so I'm trying to understand how to interpret your privacy guarantee. For example, differential privacy basically says that your expected utility changes very little depending on whether or not you're in the data set. What does your privacy guarantee mean and how is that related to some other approaches that people use?

Yevgeniy V.: It's different. The privacy guarantee, what I'm talking about is the number ... Is specifically within the machine learning context. It's essentially how much the adversary can gain as a function of budget, if you will. So we have this dichotomy as a function of budget here. So how much do they gain, how much added value does machine learning have for them? Those are the kind of property guarantees, they're orthogonal to the things that, for example, in differential privacy you hear about. Differential privacy guarantees are much stronger than what we're claiming, for example.

Speaker 16: Okay.

Yevgeniy V.: But really different.

Speaker 16: Okay.

Eric Horvitz: [inaudible 01:07:44] do you have a comment? Question?

Speaker 3: Is there a way in which you can rank or rate the loss of different things that you may lose in privacy? Or are all of the losses of equal value? For example, some might be more, some might be less, and so might be able to kind of play a game in order to try to only limit losses to the more important terms or something like that.

Yevgeniy V.: That's a good point. Right now, we're basically think of all sensitive entities homogenous. In practice, what we would do is we'd take just names first, using this approach, and that would be some L over C. Then you take something else like SSN and it would be a different L over C. So we'd just, in this language, would be playing different games for the different sensitive entities to capture the fact that they of course have different trade offs. There might be a better, cleaner way to do it within one framework, but I don't know.

Eric Horvitz: Yeah, the interesting thing for me, [inaudible 01:08:51] medical is companies like Microsoft, and I assume Google, and Amazon, would love to share their internal data sets under compliance with their end user licensing agreements with researchers and scientists, but it's often challenging to do so in a way that protects the consumer base. Having some methodology for cleaning the data in a way that reduces risk would make that more possible. We're hoping to see some cross industry initiatives on closing the data divide between companies, and academia, and NGOs by coming up with methods like that. One more question, yeah?

Speaker 15: [crosstalk 01:09:33] Eric, this one's for you.

Eric Horvitz: Oh, okay.

Speaker 15: That's first time I've heard you say something like this.

Eric Horvitz: Well we're transitioning to open discussion, so maybe we'll both take this question as a dual-

Speaker 15: Okay, so what would it take for a method like this, or any type of a sanitization method, to convince the Microsoft lawyers to allow the data to go out the door? Because I would love to work on that technology.

Eric Horvitz: You mean to enable that or get the access to the data?

Speaker 15: Both.

Yevgeniy V.: [inaudible 01:09:58] collaborator in this work, by the way. [crosstalk 01:10:00]

Eric Horvitz: Oh, okay. You are ... Let me see in the paper here-

Speaker 15: I'm the last one, I'm the one that asked you for the Pokemon Go data set.

Eric Horvitz: Oh, okay. We've had several requests like that. For one, we'd have to probably change our EULAs because the EULAs say we do not give out data of any kind to outside of Microsoft. So if you trust Microsoft, feel free to use the service, and it's similar to what other companies do. You can imagine us saying, if you opt in, your data will be ... Your name and age will be detectable with some probability. Some people say, "Oh, that's okay with me, I want to donate to science and behavioral studies." The other approach to that Microsoft has done, and I think we've been leading, at least in the industry, in this, because I don't know of other companies that have done the same thing.

Three times we had RFPs where we invited the academic community to apply to be part of a project and a program to come into the tent under license and use sanitized logs, and then it'd be a year of workshops, and gatherings, and people showed their studies and published their results. But it was in accordance with

the end user licensing agreement because they're now "Microsoft internal" by signing a document because the EULAs say, "Microsoft or its contractors and coworkers," and so on.

Speaker 15: [inaudible 01:11:38] identify information [inaudible 01:11:40]? Is it subject to EULA?

Eric Horvitz: I believe so, but you can imagine that if it was really, truly so easy change to make. Because our lawyers are very bright people and they say, "Oh, well in that case, we'll just change the sentence here and [inaudible 01:11:58] everybody run away from, more than they are already, from Bing to Google." That was supposed to be a joke there. Actually, Bing is doing very well, by the way, as I say for the live stream these days. Okay. Well, thank you very much. [crosstalk 01:12:14]

So now we have about 10 minutes of open discussion before we break, on healthcare in general. To any of the authors or co-authors in presence, or on any topic today that we discussed. So I thought some interesting discussions for this group would be the question I asked one of the speakers today, how we choose topics to work on, what makes a topic interesting and valuable. Where value is particularly a focus of this workshop. I was looking at things that I worked on in the past and I just noticed that I tend to get really excited about, even if they're not so sophisticated, if the upside is really big in terms of ... I often like looking at the value or the cost of a certain handling or mishandling of a disease entity.

So I had $35 billion a year for CHF, congestive heart failure care, and the fact that congestive heart failure will affect about 10% of us if things continue the way they do, per no breakthroughs and so on, over 65 years of age. And the fact that when you have diagnosed congestive heart failure, you have a 10% per year mortality rate. Or, hospital acquired infections, they're in the top seven, I believe of all causes of death in the United States, believe it or not. I have the stats on the tip of my tongue, but some large fraction of patients get sick in the hospital with something they didn't bargain for.

If I even asked this audience, how many people here know people, or family members, or friends, that went to the hospital for complicated procedure X and it went well, and the whole family's excited, and then you hear about this other thing going on, this hospital acquired infection. I bet you in this room some people have lost family and friends that way, not from the initial entity, but the side effect or the inadvertent infection.

In that department, there are, if you believe the various, several, studies, close to 750 people dying per day in the US because of preventable medical errors. That's human cognition and hospital workflow related challenges and opportunities for us to solve with various kinds of computational safety nets. Of course, the big one we all know about, about 100 deaths in the US per day, people just trying to drive a car from A to B, and about 1,000 lifetime disabling

injuries per day. In a healthcare session, you don't think about automatic braking systems and the attention of drivers, but it's a huge public health disaster, but I think it's about 4 million deaths to date in the US from automobiles, cars, since we started driving.

I often say when I think about politics, and Donald Trump, and his campaign, there are various ways to make America safer and there are more obvious ways than what are being talked about in the press. With that, I'll open it up to any discussion or conversation. Yeah?

Speaker 12: I had a general question, and this kind of ties back into the invited talk about the social justice, or the social work, and incarceration rates, where people could be let out by judges. There was interest in making the machine learning, the very black boxy algorithmic design ... Not opaque, make it transparent, and make it available for laymen people to understand. I would think if looking at the future of machine learning and artificial intelligence in healthcare in the next five or 10 years, that's going to be a major fundamental challenge to try to make these big data, deep learning models transparent and trustworthy.

The real question is, how do you bridge that gap of providing systems like Watson that are scanning millions, and millions, and millions of health records, and proving diagnoses or recommendations, and however that works is a complete mystery of magic. How do you translate that barrier and how does the industry and the profession translate that barrier to make these artificial intelligence solutions less opaque? Any authors can ...

Eric Horvitz: Right. I have my reaction to that, but let's open it up to the audience.

Speaker 3: I want to-

Eric Horvitz: Here.

Speaker 3: Interpretability of decisions is certainly one important aspect of deploying these systems in society, but I wanted to ask all the people who gave ... Speakers in this session, or maybe even the [inaudible 01:17:20], are there other such principles that we are to be thinking about for this area? For example, earlier there was some discussion of autonomy in the sense of where doctors or so forth, experts, give up some level of autonomy because you have a AI agent saying, "This is what you should do," and whether they would willingly accept that or whether you want to play along with them to some degree because you want to respect their opinions as well.

There's clearly also a sense of whether some of these decisions are perhaps causing more harm than good. So, are there a set of principles that need to be obeyed as we start deploying these systems in society? In addition to interpretability. I know that our panels have been put together to think about a series of principles by which AI systems could be deployed in society, but this

might ... The other thought then, is this like when we do human subject experiment and we have to go through our institutional review board and so forth, is there an AI review board that would need to say, "We certify this is interpretable, it does all these things," and so now we can deploy them?

Eric Horvitz: I'll add a third one, which is, when you have a system that's in healthcare, but in other areas as well, safety critical areas, when you have a system that's doing reasoning and decision making, it's actually influencing the world that it's studying, that it was built to make decisions about. Its influence is in ripples of its own effects being understood and ingested by the models in a way that makes sense. I can give you examples of where it is not. That's a third kind of principle. You're getting at the idea of what I would call loosely or summarized as best practices for different sectors when it comes to machine learning and intelligence. This is something that a number of people are thinking about. There's an effort called the Partnership in AI to benefit people in society.

The Partnership in AI, PAI for short, was, I would say co-authored by researchers at the major IT companies starting with ... It was Microsoft, Facebook, Amazon, Google, and IBM, and then Apple joined up with this group. It'll be a bigger group soon, but the idea was, companies get together with non-profits, with civil liberties groups, to create a non-profit that would be arrayed around best practices for the field more globally, but also by sector. These are the kinds of questions that I think will be asked for which partners, and colleagues, and academics will be relied upon to help answer through workshops and white papers, and so on. So I think this is going to be happening as early as this coming year in terms of this being stood up. So it's a really good question as to what are best practices when it comes to AI entering an area like healthcare and what it means.

My other comment was on the trustworthiness and so on, is that in many cases, human decision making is so poor in healthcare, even by experts, that there's reason to trust even a classically validated test and train validation that shows how powerful a data-centric model can be in making a prediction. It's often the case that they don't have to explain things necessarily to end users or patients, or even single doctors, but that experts looking at the model should be able to inspect and understand what's happening. There's a dramatic case that Rich [inaudible 01:21:32] talks about, about ... It was published 1996 AI journal, Greg Cooper lead the study at Pittsburgh, of predicting death, patients at high risk of death, who have been diagnosed with pneumonia.

It turns out that most patients do fine with pneumonia, they get over it, but we hear about people even in their 50's, and 40's, and young people, dying after a bout of pneumonia. So a classifier was built to try to predict which patients should go to the ICU and get special care immediately versus be sent home with some antibiotics, which is a standard procedure. When the model was looked at, it had a great AUC, it performed well, it was fabulous, everyone was excited. But they were looking at this model and someone noticed looking at this, the

rules that come out of the model, that if a patient has asthma, they're low risk for dying from pneumonia. And doctors said, "Well, that's kind of funny, I don't buy that. Why would that be in the large data from Pittsburgh hospitals?"

It turned out that was only seen because it was a transparent, scrutinizable, linear model, or else he wouldn't have seen that right away, popping out of the data set. It turned out that the reason that was in the model, and it's a scary example, and it's well worth us thinking deeply about this. The reason that that erroneous rule, which would have killed patients if it was in practice in that model was because those patients that had asthma were considered so critical and so high risk, they were removed before the data analytic pipeline could capture them in this data set. So the data set itself ends up with a blind spot in a very important way, that was only seen through expert scrutinizing and transparency of the level that it had. Yeah.

Speaker 17: [inaudible 01:23:38]

Eric Horvitz: Oh, sorry. Your hand wasn't very high, but you have the microphone though, so you have the power.

Speaker 13: [inaudible 01:23:45] a few things answering [inaudible 01:23:48] on the principles that we should look at. In a European project where I'm involved, we are talking about the art principles of AI, ART, accountability, responsibility, and transparency. In that responsibility refers mostly to ourselves, what is our responsibility on not only what we can do, but why we are doing it and what's the impact on it. Accountability and transparency go kind of together as making it more transparent, you'll get more. If it translated to the [inaudible 01:24:23] that is something which we have been discussing in IEEE initiative for ethically aligned design, which I believe several of you are involved as well.

There we have looked as well, it's not only that as we are developing mechanisms for better diagnostics or automatic diagnostics, and so on, that should be going together with also a change on the way we perform medicine. Also, are we train our medical practitioners so as machines are more and more doing better diagnosis than people, the education of our doctors should shift not so much on the, let's say, the details of the diagnostic, but much more on the communicating. So it is much more than just the algorithmic approach, we have to look at it in a much broader societal setup. [inaudible 01:25:31], but on the other hand, also all the responsibility we have as the ones who are developing those algorithms to create the society in which those algorithms are really applied for good and not just to upset the setup.

Eric Horvitz: That's fabulous. The ART, is that acronym, that initiative, among AI researchers and scientists? Or is that part of the European Parliamentary guidelines and so on?

Speaker 13: [inaudible 01:26:05] the Parliament, but we are starting [inaudible 01:26:10]. That's where we are talking about the ART.

Eric Horvitz: Yeah.

Speaker 13: [inaudible 01:26:21] acronym.

Eric Horvitz: Part of this ... It's a beautiful acronym, it translates well across the languages. It's very interesting to think about even just getting sensitive to the stories. Clinical medicine is all about being exposed to these stories, to these situations, versus the first two years we look at data sets, and symptom tables, and so on. The actual experiences with these stories, like the pneumonia story, the story about even being sensitive to ask the question, if the whole United States, for example, is arrayed at minimizing penalties for 30 day readmission rates, what happens to patients who come in at 45 days even sicker?

Why wasn't that on the map that if you have a simple rule that you're optimizing, you might be pushing the poorly filled balloon on this side and have it pop on this side. We need to be thinking globally about even how these systems can be gamed, how they can creating gaming situations, how adversarials can attack them. For example, this is a simple adversarial model, to reduce your penalties in a hospital, I'll have you come back on the 33rd day. Just keep on using the oxygen at home. Yeah? Oh, I'm sorry, you're next, yeah. Then we'll pass it up here. Yeah.

Yevgeniy V.: Okay. I have a couple of quick comments. The first comment is actually tied into a number of examples that you have made, which is in modeling the common assumption you make as a closed world assumption. Once you deploy a model or anything that's based on the model in the world, the world is no longer closed from a perspective of this model. So a lot of these examples that you mentioned are really violations of the closed world assumption. Another one in security, a very sort of intuitive example. Let's say you deploy a really good security approach in one place, so the attackers could attack some other places as a consequence because now the other place is easier to attack. These are just violations of the closed world assumption.

Eric Horvitz: Yeah.

Yevgeniy V.: Another quick comment is about transparency. So there's sort of multiple modes of transparency and I think it's distinct from interpretability. So interpretability is, first of all, interpretable to whom, at what level of education? But transparency is a broader issue. For example, open source, creating things that are actually open source, makes it potentially transparent because it can be inspected and evaluated by other people, even if it's not necessarily interpretable.

Eric Horvitz: One comment I'll just make is that you said that all these ... [crosstalk 01:29:10] That was the ethical thing to do. [inaudible 01:29:14] just this comment for a second here. I love that you said that all these examples violate the closed world assumption. I was going to say, the reality of the open world violates the closed world assumption. I'm sorry, go ahead.

Speaker 18: Another example that I think [inaudible 01:29:33], getting back to your presentation on the wandering, Alzheimer's patients, they forget a lot, so the idea is to have this watch lasting for 21 days before it has to be charged. Why? I would say, actually, my experience, my mother has Alzheimer, is if there's something that she does every day, she will stick to it. If it lasts more than a few days, she will not have a clue what to do anymore. So having a watch that you have to recharge every day is better than one that you do 21 days.

You're solving a problem which I think, "Well, is it really a problem?" And we're doing that far too often. We have an idea about, what's the solution? Without actually looking at, what is the real world problem? Not what's our problem, but do we think is a problem, but what's the actual problem? Then looking at that and see what we have to do to solve that problem. This is one of those tiny things and there are many, many of us think where we think we're doing very clever things, and then we're actually not. Getting back into interpretability, we can actually do some very clever machine learning, we don't have to explain that. An expert doctor is not going to explain exactly what he did to a patient, he [inaudible 01:30:58] give some reasons that a patient will accept, that should be acceptable, not true, it should be something which is something that actually is close to what he did, but not what he did. So it's a different type of problem.

Eric Horvitz: Yeah. You could imagine though that there are approaches to ... Back to this question about trustworthiness and so on, where in medicine or criminal justice, per this morning's lecture, that at least a data set is available to be scrutinized by experts upon request, and they should hammer on this to sort of figure things out. Even if it can't be available to everybody per adversaries and so on. Our last comment because we have a coffee break in process as we speak.

Speaker 15: Just to follow up on that discussion, I actually don't think you can do this because for the purposes of malpractice and having documentation of what the physician actually did, you need to have that decision making process and you need to have appropriate documentation of what actually happened. If you don't have that documentation, then it's actually going to be problematic for the health care institutions and the physician that did this because they're going to be subject to all sorts of lawsuits at this point because they're not documenting appropriately.

Eric Horvitz: Per my understanding of malpractice, it often gets down to best practices. Did this doctor follow best practices? If not, often goes to decision analysis, actual cost benefit analysis under uncertainly, so you can imagine that a probabilistic

model being used would be admitted, and its characterization, as either a best practice or it's characterizable.

Speaker 15: Yeah. It will play on best practice, but Millennium Pharmaceuticals lost this battle when they tried to hide some of their decision making methods when they were doing genetic based decision making. When they tried to hide it, people said, "What exactly are you making recommendations for?" They said, "That's proprietary, we're not going to tell you," and then they ended up losing in court, and that information had to be made public at that point.

Eric Horvitz: Well, this sounds like a great coffee break story to continue. Thanks to all the speakers and the discussions.