PETER SZOLOVITS: OK. Today's topic is differential diagnosis. And so I'm just quoting Wikipedia here. Diagnosis is the identification of the nature and cause of a certain phenomenon. And differential diagnosis is the distinguishing of a particular disease or condition from others that present similar clinical features.

So doctors typically talk about differential diagnosis when they're faced with a patient and they make list of what are the things that might be wrong with this patient. And then they go through the process of trying to figure out which one it actually is. So that's what we're going to focus on today.

Now, just to scare you, here's a lovely model of human circulatory physiology. So this is from Guyton's textbook of cardiology. And I'm not going to hold you responsible for all of the details of this model, but it's interesting, because this is, at least as of maybe 20 years ago, the state of the art of how people understood what happens in the circulatory system. And it has various control inputs that determine things like how your hormone levels change various aspects of the cardiovascular system and how the interactions between different components of the cardiovascular system affect each other.

And so in principle, if I could tune this model to me, then I could make all kinds of pretty good predictions that say if I increase my systemic vascular resistance, then here's what's going to happen as the rest of the system adjusts. And if I get a blockage in a coronary artery, then here's what's going to happen to my cardiac output and various other things.

So this would be terrific. And if we had this kind of model for not just the cardiovascular system, but the entire body, then we'd say, OK, we've solved medicine. Well, we don't have this kind of model for most systems. And also, there's this minor problem that if I give you this model and say, "How does this relate to a particular patient?", how would you figure that out? This has hundreds of differential equations that are being represented by this diagram. And they have many hundreds of parameters.

And so we were joking when we started working with this model that you'd really have to kill the patient in order to do enough measurements to be able to tune this model to their particular physiology. And of course, that's probably not a good practical approach.

We're getting a little better by developing more non-invasive ways of measuring these things.

But that's moving along very slowly. And I don't expect that I or maybe even any of you will live long enough that sort of this approach to doing medical reasoning and medical diagnosis is actually going to happen.

So what we're going to look at today is what simpler models are there for diagnostic reasoning. And I'm going to take the liberty of inflicting a bit of history on you, because I think it's interesting where a lot of these ideas came from.

So the first idea was to build flowcharts. Oh, and by the way, the signs and symptoms, I've forgotten if we've talked about that in the class. So a sign is something that a doctor sees, and a symptom is something that the patient experiences. So a sign is objective. It's something that can be told outside your body. A symptom is something that you feel. So if you're feeling dizzy, then that's a symptom, because it's not obvious to somebody outside you that you're dizzy, or that you have a pain, or such things.

Normally, we talk about manifestations or findings, which is sort of a super category of all the things that are determinable about a patient. So we'll talk about flowcharts, models based on associations between diseases and these manifestations. Then there are some issues about whether you're trying to diagnose a single disease or a multiplicity of diseases, which makes the models much more complicated whether you're trying to do probabilistic diagnosis or definitive or categorical. And then we'll talk about some utility theoretic methods. And I'll just mention some rule-based and pattern-matching kinds of approaches.

So this is kind of cute. This is from 1973. And if you were a woman and walked into the MIT Health Center and complained of potentially a urinary tract infection, they would take out this sheet of paper, which was nicely color-coded, and they would check a bunch of boxes. And if you hit a red box, that represented a conclusion. And otherwise, it gave you suggestions about what further tests to do.

And this was essentially a triage instrument. It said, does this woman have a problem that requires immediate attention? And so we should either call an ambulance and take them to a hospital, or is it something where we can just tell them to come back the next day and see a doctor, or is it in fact some self-limited thing where we say, take two aspirin, and it'll go away. So that was the attempt here.

Now, interestingly, if you look at the history of this project between the Beth Israel Hospital and

Lincoln Laboratories, it started off as a computer aid. So they were building a computer system that was supposed to do this. And then in-- but you can imagine, in the late 1960s, early 1970s, computers were pretty clunky. PCs hadn't been invented yet. So this was like mainframe kinds of operations. It was very hard to use. And so they said, well, this is a small enough program that we can reduce it to about 20 flow sheets-- 20 sheets like this, which they proceeded to print up.

And I was amused, because in the-- around 1980, I was working in my office one night. And I got this splitting headache. And I went over to MIT medical. And sure enough, the nurse pulled out one of these sheets for headaches and went through it with me and decided that a couple of Tylenols should fix me. But it was interesting. So this was really in use for a while.

Now, the difficulty with approaches like this, of which there have been many, many, many in the medical world, is that they're very fragile. They're very specific. They don't take account of unusual cases. And there's a lot of effort in coming to consensus to build these things. And then they're not necessarily useful for a long time.

So MIT actually stopped using them shortly after my headache experience. But if you go over to a hospital and you look on the bookshelf of a junior doctor, you will still find manuals that look kind of like this that say, how do we deal with tropical diseases? So you ask a bunch of questions, and then depending on the branching logic of the flowchart, it'll tell you whether this is serious or not.

And the reason is because if you do your medical training in Boston, you're not going to see very many tropical diseases. And so you don't have a base of experience on the basis of which you can learn and become an expert at doing it. And so they use this as a kind of cheat sheet.

I mentioned that the association between diseases and symptoms is another important way of doing diagnosis. And I swear to you, there was a paper in the 1960s, I think, that actually proposed this. So if any of you have hung around ancient libraries, libraries used to have card catalogs that were physical pieces of paper, cardboard. And one of the things they did with these was each card would be a book.

And then around the edges were a bunch of holes, and depending on categorizations of the book along various dimensions, like its Dewey decimal number, or the top digits of its Library of Congress number or something, they would punch out holes in the borders. And this

allowed you to do a kind of easy sorting of these books.

So if you've got a bunch of cards together when people were returning their books and you pulled a bunch of cards. And you wanted to find all the math books. So what you would do is you'd stick a needle through the hole that represented math books, and then you shake the pile, and all the math books would fall out because they had punched.

So somebody seriously proposed this as a diagnostic algorithm. And in fact, implemented it. And was trying to even make money on it. I think this was an attempt at a commercial venture, where they were going to provide doctors with these library cards that represented diseases. And the holes now represented not mathematics versus literature, but they represented shortness of breath versus pain in the left ankle versus whatever. And again, as people came in and complained about some condition, you'd stick a needle through that condition and you'd shake, and up would come the cards that had that condition in common.

So one of the obvious problems with this approach is that if you had two things wrong with you, then you would wind up with no cards very quickly, because nothing would fall out of the pile. So this didn't go anywhere.

But interestingly, even in the late 1980s, I remember being asked by the board of directors of the *New England Journal of Medicine* to come to a meeting where they had gotten a pitch from somebody who was proposing essentially exactly this diagnostic model, except implemented in a computer now and not in these library cards. And they wanted to know whether this was something that they ought to get behind and invest in. And I and a bunch of my colleagues assured them that this was probably not a great idea and they should stay away from it, which they did.

Well, a more sophisticated model is something like a Naive Bayes model that says if you have a disease-- where is my cursor? If you have a disease, and you have a bunch of manifestations that can be caused by the disease, we can make some simplifying assumptions that say that you will only ever have one disease at a time, which means that the values of that node D form an exhaustive and mutually exclusive set of values.

And we can assume that the manifestations are conditionally independent observables that depend only on the disease that you have, but not on each other or not on any other factors. And if you make that assumption, then you can apply good old Thomas Bayes's rule.

This, by the way, is the Reverend Bayes. Do you guys know his history? So he was a nonconformist minister in England. And he was not a mathematician, except I mean, he was an amateur mathematician. But he decided that he wanted to prove to people that God existed. And so he developed Bayesian reasoning in order to make this proof.

And so his argument was, well, suppose you're completely in doubt. So you have 50/50 odds that God exists. And then you say, let's look at miracles. And let's ask, what's the likelihood of this miracle having occurred if God exists versus if God doesn't exist? And so by racking up a bunch of miracles, you can convince people more and more that God must exist, because otherwise all these miracles couldn't have happened.

So he never publish this in his lifetime, but after his death one of his colleagues actually presented this as a paper at the Royal Society in the UK. And so Bayes became famous as the originator of this notion of how to do probabilistic reasoning about at least fairly simple situations, like in his case, the existence or nonexistence of God. Or in our case, the cause of some disease, the nature of some disease.

And so you can draw these trees. And Bayes's rule is very simple. I'm sure you've all seen it.

One thing that, again, makes contact with medicine is that a lot of times, you're not just interested in the impact of one observable on your probability distribution, but you're interested in the impact of a sequence of observations. And so one thing you can do is you can say, well, here is my general population.

So let's say disease 2 has 37% prevalence and disease 1 has 12%, et cetera. And now I make some observation. I apply Bayes's rule. And I revise my probability distribution.

So this is the equivalent of finding a smaller population of patients who have all had whatever answer I got for symptom 1. And then I just keep doing that. And so this is the sequential application of Bayes's rule. And of course, it does depend on the conditional independence of all these symptoms.

But in medicine, people don't like to do math, even arithmetic much. And they prefer doing addition rather than multiplication, because it's easier. And so what they've done is they said, well, instead of representing all this data in a probabilistic framework, let's represent it as odds. And if you represent it as odds, then the odds of some disease given a bunch of symptoms, given the independence assumption, is just the prior odds of the disease times the conditional

odds, the likelihood ratio of each of the symptoms that you've observed.

So you've just got to multiply these together. And then because they like adding more than multiplying, they said, let's take the log of both sides. And then you can just add them.

And so if you remember when I was talking about medical data, there are things like the Glasgow Coma score, or the APACHE score, or various measures of how badly or well a patient is doing that often involve adding up numbers corresponding to different conditions that they have.

And what they're doing is exactly this. They're applying sequentially Bayes's rule with these independence assumptions in the form of logs rather than multiplications, log odds, and that's how they're doing it.

Very quickly. Somebody in a previous lecture was wondering about receiver operator characteristic curves. And I just wanted to give you a little bit of insight on those. So if you do a test on two populations of patients-- the red ones are sick patients. The blue ones are not sick patients. You do some test. What you expect is that the result of that test will be some continuous number, and it'll be distributed something like the blue distribution for the well patients and something like the red distribution for the ill patients.

And typically, we choose some threshold. And we say, well, if you choose this to be the threshold between a prediction of sick or well, then what you're going to get is that the part of the blue distribution that lies to the right is the false positives and the part of the red distribution that lies to the left is the false negatives. And often people will choose the lowest point at which these two curves intersect as the threshold, but that, of course, isn't necessarily the case.

Now, if I give you a better test, one like this, that's terrific, because there is essentially no overlap. Very small false negative and false positive rates. And as I said, you can choose to put the threshold in different places, depending on how you want to trade off sensitivity and specificity.

And we measure this by this receiver operator characteristics curve, which has the general form that if you get a curve like this, that means that there's an exact trade-off for sensitivity and specificity, which is the case if you're flipping coins. So it's random.

And of course, if you manage to hit the top corner up there, that means that there would be no

overlap whatsoever between the two distributions, and you would get a perfect result. And so typically you get something in between. And so normally, if you do a study and your AUC, the area under this receiver operator characteristics curve, is barely over a half, you're pretty close to worthless, whereas if it's close to 1, then you have a really good method for distinguishing these categories of patients.

Next topic. What does it mean to be rational? I should have a philosophy course here.

AUDIENCE: Are you talking about pi?

PETER SZOLOVITS: Sorry.

AUDIENCE: Are you talking about pi? Pi is--

PETER SZOLOVITS: Pi is irrational, but that's not what I'm talking about.

Well, so there is this principle of rationality that says that what you want to do is to act in such a way as to maximize your expected utility. So for example, if you're a gambler and you have a choice of various ways of betting in some poker game or something, then if you were a perfect calculator of the odds of getting a queen on your next draw, then you could make some rational decision about whether to bet more or less, but you'd also have to take into account things like, "How could I convince my opponent that I am not bluffing if I am bluffing?" and "How could I convince them that I'm bluffing if I'm not bluffing?" and so on.

So there is a complicated model there. But nevertheless, the idea is that you should behave in a way that will give you the best expected outcome. And so people joke that this is Homo economicus, because economists make the assumption that this is how people behave. And we now know that that's not really how people behave. But it's a pretty common model of their behavior, because it's easy to compute, and it has some appropriate characteristics.

So as I mentioned, every action has a cost. And utility measures the value or the goodness of some outcome, which is the amount of money you've won, or whether you live or die, or quality adjusted life years, or various other measures of utility-- how much it costs for your hospitalization.

So let me give you an example. This actually comes from a decision analysis service at New

England Medical Center Tufts Hospital in the late 1970s. So this was an elderly Chinese gentleman whose foot had gangrene. So gangrene is an infection that usually people who have bad circulation can get these. And what he was facing was a choice of whether to amputate his foot or to try to treat him medically. To treat him medically means injecting antibiotics into his system and hoping that his circulation is good enough to get them to the infected areas.

And so the choice becomes a little more complicated, because if the medical treatment fails, then, of course, the patient may die, a bad outcome. Or you may have to now amputate the whole leg, because the gangrene has spread from his foot up the foot, and now you're cutting off his leg. So what should you do? And how should you reason about this?

So Pauker's staff came up with this decision tree. By the way, decision tree in this literature means something different from decision tree in like C4.5. So your choices here are to amputate the foot or start with medical care. And if you amputate the foot, let's say there is a 99% chance that the patient will live. There's a 1% chance that maybe the anesthesia will kill him.

And if we treat him medically, they estimated that there is a 70% chance of full recovery, a 25% chance that he'd get worse, a 5% chance that he would die. If he got worse, you're now faced with another decision, which is, do we amputate the whole leg or continue pushing medicine? And again, there are various outcomes with various estimated probabilities.

Now, the critical thing here that this group was pushing was the idea that these decisions shouldn't be based on what the doctor thinks is good for you. They should be based on what you think is good for you. And so they worked very hard to try to elicit individualized utilities from patients.

So for example, this guy said that having your foot amputated was worth 850 points on a scale of 1,000 where being healthy was 1 and being dead was 0.

Now, you could imagine that that number would be very different for different individuals. If you asked LeBron James how bad would it be to have your foot amputated, he might think that it's much worse than I would, because it would be a pain to have my foot amputated, but I could still do most of the things that I do professionally, whereas he probably couldn't as a star basketball player.

So how do you solve a problem like this? Well, you say, OK, at every chance node I can calculate the expected value of what happens here. So here at it's 0.6 times 995, 0.4 times 0. That gets me a value for this decision.

Do the same thing here. I compare the values here and choose the best one. That gives me a value for this decision. And so I fold back this decision tree.

And my next slide should have-- yeah, so these are the numbers that you get. And what you discover is that the utility of trying medical treatment is somewhat higher than the utility of immediately amputating the foot if you believe these numbers and those utilities, these probabilities and those utilities.

Now, the difficulty is that these numbers are fickle. And so you'd like to do some sort of sensitivity analysis. And you say, for example, what if this gentleman valued his living with an amputated foot at 900 rather than 850. And now you discover that amputating the foot looks like a slightly better decision than the other.

So this is actually applied in clinical medicine. And there are now thousands of doctors who have been trained in these techniques and really try to work through this with individual patients.

Of course, it's used much more on an epidemiological basis when people look at large populations.

AUDIENCE: I have a question.

PETER SZOLOVITS: Yeah.

AUDIENCE: How are the probabilities assessed?

PETER SZOLOVITS: So the service that did this study would read the literature, and they would look in databases. And they would try to estimate those probabilities. We can do a lot better today than they could at that time, because we have a lot more data that you can look in.

But you could say, OK, for people-- men of this age who have gangrenous feet, what fraction of them have the following experience? And that's how these are estimated.

Some of it feels like 5%. OK. So I just said this.

And then the question of where do you get these utilities is a tricky one. So one way is to do the standard gamble, which says, OK, Mr. Szolovits, we're going to play this game. We're going to roll a fair die or something that will come up with some continuous number between 0 and 1, and then I'm going to play the game where either I chop off your foot, or I roll this die and if it exceeds some threshold, then I kill you. Nice game.

So now if you find the point at which I'm indifferent, if I say, well, 0.8, that's a 20% chance of dying. It seems like a lot. But maybe I'll go for 0.9, right? Now you've said, OK, well, that means that you value living without a foot at 0.9 of the value of being healthy. So this is a way of doing it. And this is typically done.

Unfortunately, of course, it's difficult to ascertain the problem. And it's also not stable. So people have done experiments where they get somebody to give them this kind of number as a hypothetical, and then when that person winds up actually faced with such a decision, they no longer will abide with that number. So they've changed their mind when the situation is real.

AUDIENCE: But it's nice, because there are two feet, right? So you could run this experiment and see.

PETER SZOLOVITS: They didn't actually do it. It was hypothetical. OK.

Next program I want to tell you about, again, the technique for this was developed as a PhD thesis here at MIT in 1967. So this is hot off the presses. But it's still used, this type of idea.

And so this was a program that was published in the *American Journal of Medicine,* which is a high impact medical journal. I think this was actually the first sort of computational program that journal had ever published as a medical journal.

And it was addressed at the problem of the diagnosis of acute oliguric renal failure. Oliguric means you're not peeing enough. Renal is your kidney. So this is something's gone wrong with your kidney, and you're not producing enough urine.

Now, this is a good problem to address with these techniques, because if something happens to you suddenly, it's very likely that there is one cause for it. If you are 85 years old and you have a little heart disease and a little kidney disease and a little liver disease and a little lung disease, there's no guarantee that there was one thing that went wrong with you that caused all these.

But if you were OK yesterday and then you stopped peeing, it's pretty likely that there's one thing that's gone wrong. So it's a good application of this model. So what they said is there are 14 potential causes. And these are exhaustive and mutually exclusive.

There are 27 tests or questions or observations that are relevant to the differential. These are cheap tests, so they didn't involve doing anything either expensive or dangerous to the patient. It was measuring something in the lab or asking questions of the patient.

But they didn't want to have to ask all of them, because that's pretty tedious. And so they were trying to minimize the amount of information that they needed to gather in order to come up with an appropriate decision. Now, the real problem, there were three invasive tests that are dangerous and expensive, and then eight different treatments that could be applied.

And I'm only going to tell you about the first part of this problem. This 1973 article shows you what the program looked like. It was a computer terminal where it gave you choices, and you would type in an answer. And so that was the state of the art at the time.

But what I'm going to do is, god willing, I'm going to demonstrate a reconstruction that I made of this program. So these guys are the potential causes of stopping to pee-- acute tubular necrosis, functional acute renal failure, urinary tract obstruction, acute glomerulonephritis, et cetera. And these are the prior probabilities.

Now, I have to warn you, these numbers were, in fact, estimated by people sticking their finger in the air and figuring out which way the wind was blowing, because in 1973, there were not great databases that you could turn to.

And then these are the questions that were available to be asked. And what you see in the first column, at least if you're sitting close to the screen, is the expected entropy of the probability distribution if you answered this question. So this is basically saying, if I ask this question, how likely is each of the possible answers, given my disease distribution probabilities?

And then for each of those answers, I do a Bayesian revision, then I weight the entropy of that resulting distribution by the probability of getting that answer. And that gets me the expected entropy for asking that question. And the idea is that the lower the expected entropy, the more valuable the question. Makes sense.

So if we look, for example, the most valuable question is, what was the blood pressure at the

onset of oliguria? And I can click on this and say it was, let's say, moderately elevated.

And what this little colorful graph is showing you is that if you look at the initial probability distribution, acute tubular necrosis was about 25%, and has gone down to a very small amount, whereas some of these others have grown in importance considerably.

So we can answer more questions, we can say-- let's see. What is the degree-- is there proteinuria? Is there protein in the urine? And we say, no, there isn't. I think we say, no, there isn't. 0.

And that revises the probability distribution. And then it says the next most important thing is kidney size. And we say-- let's say the kidney size is normal. So now all of a sudden functional acute renal failure, which, by the way, is one of these funny medical care categories that says it doesn't work well, doesn't explain to why it doesn't work well, but it's sort of a generic thing.

And sure enough. We can keep answering questions about, are you producing less than 50 ccs of urine, which is a tiny amount, or somewhere between 50 and 400? Remember, this is for people who are not producing enough. So normally you'd be over 400. So these are the only choices.

So let's say it's moderate. And so you see the probability distribution keeps changing. And what happened in the original program is they had an arbitrary threshold that said when the probability of one of these causes of the disease reaches 95%, then we switch to a different mode, where now we're actually willing to contemplate doing the expensive tests and doing the expensive treatments. And we build the decision tree, as we saw in the case of the gangrenous foot, that figures out which of those is the optimal approach.

So the idea here was because building a decision tree with 27 potential questions becomes enormously bushy, we're using a heuristic that says information maximization or entropy reduction is a reasonable way of focusing in on what's wrong with this patient. And then once we focused in pretty well, then we can begin to do more detailed analysis on the remaining more consequential and more costly tests that are available.

Now, this program didn't work terribly well, because the numbers were badly estimated, and also because of the utility model that they had for the decision analytic part was particularly terrible. It didn't really reflect anything in the real world. They had an incremental utility model that said the patient either got better, or stayed the same, or got worse. And obviously in that

order of utilities, but they didn't correspond to how much better he got or how much worse he got. And so it wasn't terribly useful.

So nevertheless, in the 1990s, I was teaching a tutorial at a Medical Informatics conference, and there were a bunch of doctors in the audience. And I showed them this program.

And one of the doctors came up afterwards and said, wow, it thinks just the way I do. And I said, really? I don't think so. But clearly, it was doing something that corresponded to the way that he thought about these cases. So I thought that was a good thing.

All right. Well, what happens if we can't assume that there's just a single disease underlying the person's problems? If there are multiple diseases, we can build this kind of bipartite model that says we have a list of diseases and we have a list of manifestations. And some subset of the diseases can cause some subset of the symptoms, of the manifestations.

And so the manifestations depend only on the diseases that are present, not on each other. And therefore, we have conditional independence. And this is a type of Bayesian network, which can't be solved exactly because of the computational complexity. So a program I'll show you in a minute had 400 or 500 diseases and thousands of manifestations. And the computational complexity of exact solution techniques for these networks tends to go exponentially with the number of undirected cycles in the network. And of course, there are plenty of undirected cycles in a network like that.

So there was a program developed originally in the early 1970s called Dialog. And then they got sued, because somebody owned that name. And then they called it Internist, and they got sued because somebody owned that name. And then they called it QMR, which stands for Quick Medical Reference, and nobody owned that name.

So around 1982, this program had about 500 diseases, which they estimated represented about 70% to 75% of major diagnoses in internal medicine, about 3,500 manifestations. And it took about 15 man years of manual effort to sit there and read medical textbooks and journal articles and look at records of patients in their hospital.

The effort was led by a computer scientist at the University of Pittsburgh and the chief of medicine at UPMC, the University of Pittsburgh Medical Center, who was just a fanatic. And he got all the medical school trainees to spend hours and hours coming up with these databases.

By 1997, they had commercialized it through a company that had bought the rights to it. And

they had-- that company had expanded it to about 750 diagnoses and about 5,500 manifestations. So they made it considerably larger. Details are-- I've tried to put references on all the slides.

So here's what data in QMR looks like. For each diagnosis, there is a list of associated manifestations with evoking strengths and frequencies. So I'll explain that in a minute.

On average, there are about 75 manifestations per disease. And for each disease-- for each manifestation in addition to the data you see here, there is also an important measure that says how critical is it to explain this particular symptom or sign or lab value in the final diagnosis.

So for example, if you have a headache, that could be incidental and it's not that important to explain it. If you're bleeding from your gastrointestinal system, that's really important to explain. And you wouldn't expect a diagnosis of that patient that doesn't explain to you why they have that symptom.

And then here is an example of alcoholic hepatitis. And the two numbers here are a so-called evoking strength and a frequency. These are both on scales-- well, evoking strength is on a scale of 0 to 5, and frequency is on a scale of 1 to 5. And I'll show you what those are supposed to mean.

And so, for example, what this says is that if you're anorexic, that should not make you think about alcoholic hepatitis as a disease. But you should expect that if somebody has alcoholic hepatitis, they're very likely to have anorexia. So that's the frequency number. This is the evoking strength number. And you see that there is a variety of those.

So much of that many, many years of effort went into coming up with these lists and coming up with those numbers. Here are the scales. So the evoking strength-- 0 means nonspecific. 5 means its pathognomonic. In other words, just seeing the symptom is enough to convince you that the patient must have this disease.

Similarly, frequency 1 means it occurs rarely, and 5 means that it occurs in essentially all cases with scaled values in between. And these are kind of like odds ratios. And they add them kind of as if they were log likelihood ratios. And so there's been a big literature on trying to figure out exactly what these numbers mean, because there's no formal definition in terms of you count the number of this and divide by the number of that, and that gives you the right

answer. These were sort of the impressionistic kinds of numbers.

So the logic in the system was that you would come to it and give it a list of the manifestations of a case. And to their credit, they went after very complicated cases. So they took clinical pathologic conference cases from *The New England Journal of Medicine.* These are cases selected to be difficult enough that doctors are willing to read these. And they're typically presented at Grand Rounds at MGH by somebody who is often stumped by the case. So it's an opportunity to watch people reason interactively about these things.

And so you evoke the diagnoses that have a high evoking strength from the giving manifestations. And then you do a scoring calculation based on those numbers. The details of this are probably all wrong, but that's the way they went about it. And then you form a differential around the highest scoring diagnosis.

Now, this is actually an interesting idea. It's a heuristic idea, but it's one that worked pretty well. So suppose I have two diseases. D1 can cause manifestations 1 through 4. And D2 can cause 3 through 6.

So are these competing to explain the same case or could they be complementary? Well, until we know what symptoms the patient actually has, we don't know. But let's trace through this.

So suppose I tell you that the patient has manifestations 3 and 4. OK. Well, you would say, there is no reason to think that the patient may have both diseases, because either of them can explain those manifestations, right? So you would consider them to be competitors.

What about if I add M1? So here, it's getting a little dicier. Now you're more likely to think that it's D1. But if it's D1, that could explain all the manifestations, and D2 is still viewable as a competitor.

On the other hand, if I also add M6, now neither disease can explain all the manifestations. And so it's more likely, somewhat more likely, that there may be two diseases present. So what Internist had was this interesting heuristic, which said that when you get that complementary situation, you form a differential around the top ranked hypothesis. In other words, you retain all those diseases that compete with that hypothesis.

And that defines a subproblem that looks like the acute renal failure problem, because now you have one set of factors that you're trying to explain by one disease. And you set aside all

of the other manifestations and all of the other diseases that are potentially complementary. And you don't worry about them for the moment. Just focus on this cluster of things that are competitors to explain some subset of the manifestations.

And then there are different questioning strategies. So depending on the scores within these things, if one of those diseases has a very high score and the others have relatively low scores, you would choose a pursue strategy that says, OK, I'm interested in asking questions that will more likely convince me of the correctness of that leading hypothesis. So you look for the things that it predicts strongly.

If you have a very large list in the differential, you might say, I'm going to try to reduce the size of the differential by looking for things that are likely in some of the less likely hypotheses so that I can rule them out if that thing is not present. So different strategies. And I'll come back to that in a few minutes.

So their test, of course, based on their own evaluation was terrific. It did wonderfully well. The paper got published in *The New England Journal of Medicine,* which was an unbelievable breakthrough to have an AI program that the editors of *The New England Journal* considered interesting.

Now, unfortunately, it didn't hold up very well. And so there was this paper by Eta Berner and her colleagues in 1994 where they evaluated QMR and three other programs. DXplain is very similar in structure to QMR. Iliad and Meditel are Bayesian network, or almost naive Bayesian types of models developed by other groups.

And they looked for results, which is coverage. So what fraction of the real diagnoses in these 105 cases that they chose to test on could any of these programs actually diagnose? So if the program didn't know about a certain disease, then obviously it wasn't going to get it right.

And then they said, OK, of the program's diagnoses, what fraction were considered correct by the experts? What was the rank order of that correct diagnosis among the list of diagnoses that the program gave? The experts were asked to list all the plausible diagnoses from these cases. What fraction of those showed up in the program's top 20? And then did the program have any value added by coming up with things that the experts had not thought about, but that they agreed when they saw them were reasonable explanations for this case?

So here are the results. And what you see is that the diagnoses in these 105 test cases, 91%

of them appeared in the DXplain program, but, for example, only 73% of them in the QMR program. So that means that right off the bat it's missing about a quarter of the possible cases.

And then if you look at correct diagnosis, you're seeing numbers like 0.69, 0.61, 0.71, et cetera. So these are-- it's like the dog who sings, but badly, right? It's remarkable that it can sing at all, but it's not something you want to listen to.

And then rank of the correct diagnosis in the program is at like 12 or 10 or 13 or so on. So it is in the top 20, but it's not at the top of the top 20. So the results were a bit disappointing. And depending on where you put the cut off, you get the proportion of cases where a correct diagnosis is within the top end. And you see that at 20, you're up at a little over 0.5 for most of these programs.

And it gets better if you extend the list to longer and longer. Of course, if you extended the list to 100, then you reach 100%, but it wouldn't be practically very useful.

AUDIENCE: Why didn't they somehow compare it to the human decision?

PETER SZOLOVITS: Well, so first of all, they assumed that their experts were perfect. So they were the gold standard. So they were comparing it to a human in a way.

AUDIENCE: Yeah.

PETER SZOLOVITS: OK. So the bottom line is that although the sensitivity and specificity were not impressive, the programs were potentially useful, because they had interactive displays of signs and symptoms associated with diseases. They could give you the relative likelihood of various diagnoses. And they concluded that they needed to study the effects of whether a program like this actually helped a doctor perform medicine better.

So just here's an example. I did a reconstruction of this program. This is the kind of exploration you could say. So if you click on angina pectoris, here are the findings that are associated with it. So you can browse through its database. You can type in an example case, or select an example case.

So this is one of those clinical pathological conference cases, and then the manifestations that are present and absent, and then you can get an interpretation that says, OK, this is our differential. And these are the complementary hypotheses. And therefore these are the manifestations that we set aside, whereas these are the ones explained by that set of

diseases. And so you could watch how the program does its reasoning.

Well, then a group at Stanford came along when belief networks or Bayesian networks were created, and said, hey, why don't we treat this database as if it were a Bayesian network and see if we can evaluate things that way? So they had to fill in a lot of details.

They wound up using the QMR database with a binary interpretation. So a disease was present or absent. The manifestation was present or absent. They used causal independence, or a leaky noisy-OR, which I think you've seen in other contexts. So this just says if there are multiple independent causes of something, how likely is it to happen depending on which of those is present or not. And there is a simplified way of doing that calculation, which corresponds to sort of causal independence and is computationally reasonably fast to do.

And then they also estimated priors on the various diagnoses from national health statistics, because the original data did not have prior data-- priors. They wound up not using the evoking strengths, because they were doing a pretty straight Bayesian model where all you need is the priors and the conditionals.

They took the frequency as a kind of scaled conditional, and then built a system based on that. And I'll just show you the results. So they took a bunch of *Scientific American* medicine cases and said, what are the ranks assigned to the reference diagnoses of these 23 cases? And you see that like in case number one, QMR ranked the correct solution as number six, but their two methods, TB and iterative TB ranked it as number one. And then these are attempts to do a kind of ablation analysis to see how well the program works if you take away various of its clever features.

But what you see is that it works reasonably well, except for a few cases. So case number 23, all variants of the program did badly. And then they excused themselves and said, well, there's actually a generalization of the disease that was in the *Scientific American* medicine conclusion, which the programs did find, and so that would have been number one across the board. So they can sort of make a kind of handwavy argument that it really got that one right.

And so these were pretty good. And so this validated the idea of using this model in that way.

Now, today you can go out and go to your favorite Google App store or Apple's app store or anybody's app store and download tons and tons and tons of symptom checkers. So I wanted to give you a demo of one of these if it works.

OK. So I was playing earlier with having abdominal pain and headache. So let's start a new one. So type in how you're feeling today.

Should we have a cough, or runny nose, abdominal pain, fever, sore throat, headache, back pain, fatigue, diarrhea, or phlegm? Phlegm? Phlegm is the winner. Phlegm is like coughing up crap in your throat.

AUDIENCE: Oh, luckily, they visualize it.

PETER SZOLOVITS: Right. So tell me about your phlegm. When did it start?

AUDIENCE: Last week.

PETER SZOLOVITS: Last week? OK.

I signed in as Paul, because I didn't want to be associated with any of this data.

So was the phlegm bloody or pus-like or watery or none of the above?

AUDIENCE: None of the above.

PETER SZOLOVITS: None of the above. So what was it like?

AUDIENCE: I don't know. Paul?

PETER SZOLOVITS: Is it any of these colors?

AUDIENCE: Green.

PETER SZOLOVITS: I think I'll make it yellow. Next.

Does it happen in the morning, midday, evening, nighttime, or a specific time of year?

AUDIENCE: Specific time of year.

AUDIENCE: Yeah. Specific time of year.

PETER SZOLOVITS: Specific time of year.

And does lying down or physical activity make it worse?

AUDIENCE: Well, it's generally not worse. So that's physical activity.

PETER SZOLOVITS: Physical activity.

How often is this a problem? I don't know. A couple times a week maybe.

Did eating suspect food trigger your phlegm?

AUDIENCE: No.

PETER SZOLOVITS: I don't know. I don't know what a suspect food is.

AUDIENCE: [INAUDIBLE] food.

PETER SZOLOVITS: Yeah. This is going to kill most of my time.

AUDIENCE: Is it getting better?

PETER SZOLOVITS: Is it improving? Sure, it's improving.

Can I think of another related symptom? No.

I'm comparing your case to men aged 66 to 72. A number of similar cases gets more refined.

Do I have shortness of breath? No. That's good. All right.

Do I have a runny nose? Yeah, sure. I have a runny nose. It's-- I don't know-- a watery, runny nose.

AUDIENCE: Does it say you've got to call [INAUDIBLE]?

PETER Well, I'm going to stop, because it will just take-- it takes too long to go through this, but you

SZOLOVITS: get the idea. So what this is doing is actually running an algorithm that is a cousin of the acute renal failure algorithm that I showed you. So it's trying to optimize the questions that it's asking, and it's trying to come up with a diagnostic conclusion.

Now, in order not to get in trouble with things like the FDA, it winds up wimping out at the end, and it says, if you're feeling really bad, go see a doctor. But nevertheless, these kinds of things are now becoming real, and they're getting better because they're based on more and more data.

Yeah.

AUDIENCE: [INAUDIBLE]

PETER SZOLOVITS: Well, I can't get to the end, because we're only at 36%.

[INTERPOSING VOICES]

Yeah. Here. All right. Somebody--

AUDIENCE: Oh, I think I need your finger.

PETER SZOLOVITS: Oh. OK. Just don't drain my bank account.

So *The British Medical Journal* did a test of a bunch of symptom checkers, of 23 symptom checkers like this about four years ago. And they said, well, can it on 45 standardized patient vignettes can it find at least the right level of urgency to recommend whether you should go to the emergency room, get other kinds of care, or just take care of yourself. And then the goals were that if the diagnosis is given by the program, it should be in the top 20 of the list that it gives you. And if triage is given, then it should be the right level of urgency.

The correct diagnosis was first in 34% of the cases. It was within the top 20 in 58% of the cases. And the correct triage was 57% accurate.

But notice it was more accurate in the emergent cases, which is good, because those are the ones where you really care. So we have-- OK. So based on what he said about me, I have an upper respiratory infection with 50% likelihood. And I can ask what to do next.

Watch for symptoms like sore throat and fever. Physicians often perform a physical exam, explore other treatment options, and recovery for most cases like this is a matter of days to weeks. And I can go back and say, I might have the flu, or I might have allergic rhinitis. So that's actually reasonable. I don't know exactly what you put in about me.

AUDIENCE: What is the less than 50?

PETER SZOLOVITS: What is what?

AUDIENCE: The less than 50.

[INTERPOSING VOICES]

AUDIENCE: Patients have to be the same demographics.

PETER SZOLOVITS: Yeah. I don't know what the less than 50 is supposed to mean.

AUDIENCE: It started with 200,000 or so.

PETER SZOLOVITS: Oh, so this is based on a small number of patients. So what happens, of course, is as you slice and dice a population, it gets smaller and smaller. So that's what we're seeing. OK. Thank you.

OK. So two more topics I'm going to rush through. One is that-- as I mentioned in one of the much earlier slides, every action has a cost. It at least takes time. And sometimes it induces potentially bad things to happen to a patient.

And so people began studying a long time ago what does it mean to be rational under resource constraints rather than rational just in this Home economicus model.

And so Eric Horvitz, who's now a big cheese guy, he's head of Microsoft Research, but used to be just a lowly graduate student at Stanford when he started doing this work. He said, well, utility comes not only from what happens to the patient, but also from the reasoning process from the computational process itself.

And so consider-- do you guys watch *MacGyver?* This is way out of date. So if MacGyver is defusing some bomb that's ticking down to zero and he runs out of time, then his utilities take a very sharp drop at that point. So that's what this work is really about, saying, well, what can

we do when we don't have all the time in the world to do the computation as well as having to try to maximize utility to the patient?

And Daniel Kahneman, who won the Nobel Prize a few years ago in economics for this notion of bounded rationality that says that the way we would like to be rational is not actually the way we behave, and he wrote this popular book that I really like called *Thinking, Fast and Slow* that says that if you're trying to figure out which house to buy, you have a lot of time to do it, so you can deliberate and list all the advantages and disadvantages and costs and so on of different houses and take your time making a decision. If you see a car barreling toward you as you are crossing in a crosswalk, you don't stop and say, well, let me figure out the pluses and minuses of moving to the left or moving to the right, because by the time you figure it out, you're dead.

And so he claims that human beings have evolved in a way where we have a kind of instinctual very fast response, and that the deliberative process is only invoked relatively rarely. Now, he bemoans this fact, because he claims that people make too many decisions that they ought to be deliberative about based on these sort of gut instincts. For example, our current president. But never mind.

So what Eric and his colleagues were doing was trying really to look at how this kind of meta level reasoning about how much reasoning and what kind of reasoning is worth doing plays into the decision making process. So the expected value of computation as a fundamental component of reflection about alternative inference strategies.

So for example, I mentioned that QMR had these alternative questioning methods depending on the length of the differential that it was working on. So that's an example of a kind of meta level reasoning that says that it may be more effective to do one kind of question asking strategy than another. The degree of refinement, people talk about things like just-in-time algorithms, where if you run out of time to think more deliberately, you can just take the best answer that's available to you now.

And so taking the value of information, the value of computation, and the value of experimentation into account in doing this meta level reasoning is important to come up with the most effective strategies. So he gives an example of a time pressure decision problem where you have a patient, a 75-year-old woman in the ICU, and she develops sudden breathing difficulties. So what do you do?

Well, it's a challenge, right? You could be very deliberative, but the problem is that she may die because she's not breathing well, or you could impulsively say, well, let's put her on a mechanical ventilator, because we know that that will prevent her from dying in the short term, but that may be the wrong decision, because that has bad side effects. She may get an infection, get pneumonia, and die that way. And you certainly don't want to subject her to that risk if she didn't need to take that risk.

So they designed an architecture that says, well, this is the decision that you're trying to make, which they're modeling by an influence diagram. So this is a Bayesian network with the addition of decision nodes and value nodes. But you use Bayesian network techniques to calculate optimal decisions here. And then this is kind of the background knowledge of what we understand about the relationships among different things in the intensive care unit. And this is a representation of the meta reasoning that says, which utility model should we use? Which reasoning technique should we use? And so on. And they built an architecture that integrates these various approaches.

And then in my last 2 minutes, I just want to tell you about an interesting-- this is a modern view, not historical. So this was a paper presented at the last NeurIPS meeting, which said the kinds of problems that we've been talking about, like the acute renal failure problem or like any of these others, we can reformulate this as a reinforcement learning problem.

So the idea is that if you treat all activities, including putting somebody on a ventilator or concluding a diagnostic conclusion or asking a question or any of the other things that we've contemplated, if you treat those all in a uniform way and say these are actions, we then model the universe as a Markov decision process, where every time that you take one of these actions, it changes the state of the patient, or the state of our knowledge about the patient. And then you do reinforcement learning to figure out what is the optimal policy to apply under all possible states in order to maximize the expected outcome.

So that's exactly the approach that they're taking. The state space is the set of positive and negative findings. The action space is to ask about a finding or conclude a diagnosis. The reward is the correct or incorrect single diagnosis.

So once you reach a diagnosis, the process stops, and you get your reward. It's finite horizon because they impose a limit on the number of questions. If you don't get an answer by then, you lose. You get a minus one reward.

There is a discount factor so that the further away a reward is, the less value it has to you at any point, which encourages shorter question sequences. And they use a pretty standard Q learning framework, or at least a modern Q learning framework using a double deep neural network strategy.

And then there are two pieces of magic sauce that make this work better. And one of them is that they want to encourage asking questions that are likely to have positive answers rather than negative answers. And the reason is because in their world, there are hundreds and hundreds of questions. And of course, most patients don't have most of those findings. And so you don't want to ask a whole bunch of questions to which the answer is no, no, no, no, no, no, no, no, no, because that doesn't give you very much guidance. You want to ask questions where the answer is yes, because that helps you clue in on what's really going on.

So they actually have a nice proof that they do this thing they call reward shaping, which basically adds some incremental reward for asking questions that will have a positive answer. But they can prove that an optimal policy learned from that reward function is also optimal for the reward function that would not include it. So that's kind of cool.

And then the other thing they do is to try to identify a reduced space of findings by what they call feature rebuilding. And this is essentially a dimension reduction technique where they're co-training. In this dual network architecture, they're co-training the policy model. It's, of course, the neural network model, this being that 2010s.

And so they're generating a sequence, a deep layered set of neural networks that generate an output, which is the m questions and the n conclusions that can be made. And I think there's a soft max over these to come up with the right policy for any particular situation.

But at the same time, they co-train it in order to predict a number of-- all of the manifestations from what they've observed before. So it's using-- it's learning a probabilistic model that says if you've answered the following questions in the following ways, here are the likely answers that you would give to the remaining manifestations. And the reason they can do that, of course, is because they really are not independent. They're very often co-varying. And so they learn that covariance, and therefore can predict which answers are going to get yes answers, which questions are going to get yes answers. And therefore, they can bias the learning toward doing that.

So last slide. So this system is called REFUEL. It's been tested on a simulated data set of 650 diseases and 375 symptoms. And what they show is that the red line is their algorithm. The yellow line uses only this reward reshaping. And the blue line is just a straight reinforcement learning approach.

And you can see that they're doing much better after many fewer epochs of training in doing this. Now, take this with a grain of salt. This is all fake data. So they didn't have real data sets to test this on. They got statistics on what diseases are common and what symptoms are common in those diseases.

And then they had a generative model that generated this fake data. And then they learned from that generative model. So of course it would be really important to redo the study with real data, but they've not done that. This was just published a few months ago.

So that's sort of where we are at the moment in diagnosis and in differential diagnosis. And I wanted to start by introducing these ideas in a kind of historical framework. But it means that there are a tremendous number of papers, as you can imagine, that have been written since the 1990s and '80s that I was showing you that are essentially elaborations on the same themes. And it's only in the past decade of the advent of these neural network models that people have changed strategy, so that instead of learning explicit probabilities, for example, like you do in a Bayesian network, you just say, well, this is simply a prediction task.

And so we'll predict the way we predict everything else with neural network models, which is we build a CNN, or an RNN, or some combination of things, or some attention model, or something. And we throw that at it. And it does typically a slightly better job than any of the previous learning methods that we've used typically, but not always.

OK. Peace.