DAVID SONTAG: So today's lecture is going to be about causality. Who's heard about causality before? Raise your hand. What's the number one thing that you hear about when thinking about causality? Yeah?

AUDIENCE: Correlation does not imply causation.

DAVID SONTAG: Correlation does not imply causation. Anything else come to mind? That's what came to my mind. Anything else come to mind? So up until now in the semester, we've been talking about purely predictive questions. And for purely predictive questions, one could argue that correlation is good enough. If we have some signs in our data that are predictive of some outcome of interest, we want to be able to take advantage of that. Whether it's upstream, downstream, the causal directionality is irrelevant for that purpose.

Although even that isn't quite true, right, because Pete and I have been hinting throughout the semester that there are times when the data changes on you, for example, when you go from one institution to another or when you have non-stationary. And in those situations, having a deeper understanding about the data might allow one to build an additional robustness to that type of data set shift. But there are other reasons as well why understanding something about your underlying data generating processes can be really important.

It's because often, the questions that we want to answer when it comes to health care are not predictive questions, their causal questions. And so what I'll do now is I'll walk through a few examples of what I mean by this. Let's start out with what we saw in Lecture 4 and in Problem Set 2, where we looked at the question of how we can do early detection of type 2 diabetes. You used Truven MarketScan's data set to build a risk stratification algorithm for detecting who is going to be newly diagnosed with diabetes one to three years from now.

And if you think about how one might then try to deploy that algorithm, you might, for example, try to get patients into the clinic to get them diagnosed. But the next set of questions are usually about the so what question. What are you going to do based on that prediction? Once diagnosed, how will you intervene? And at the end of the day, the interesting goal is not one of how do you find them early, but how do you prevent them from developing diabetes? Or how do you prevent the patient from developing complications of diabetes?

And those are questions about causality. Now, when we built a predictive model and we

introspected at the weight, we might have noticed some interesting things. For example, if you looked at the highest negative weights, which I'm not sure if we did as part of the assignment but is something that I did as part of my research study, you see that gastric bypass surgery has the biggest negative weight. Does that mean that if you give an obese person gastric bypass surgery, that will prevent them from developing type 2 diabetes?

That's an example of a causal question which is raised by this predictive model. But just by looking at the weight alone, as I'll show you this week, you won't be able to correctly infer that there is a causal relationship. And so part of what we will be doing is coming up with a mathematical language for thinking about how does one answer, is there a causal relationship here? Here's a second example. Right before spring break we had a series of lectures about diagnosis, particularly diagnosis from imaging data of a variety of kinds, whether it be radiology or pathology.

And often, questions are of this sort. Here is a woman's breasts. She has breast cancer. Maybe you have an associated pathology slide as well. And you want to know what is the risk of this person dying in the next five years. So one can take a deep learning model, learn to predict what one observes. So in the patient in your data set, you have the input and you have, let's say, survival time. And you might use that to predict something about how long it takes from diagnosis to death.

And based on those predictions, you might take actions. For example, if you predict that a patient is not risky, then you might conclude that they don't need to get treatment. But that could be really, really dangerous, and I'll just give you one example of why that could be dangerous. These predictive models, if you're learning them in this way, the outcome, in this case let's say time to death, is going to be affected by what's happened in between.

So, for example, this patient might have been receiving treatment, and because of them receiving treatment in between the time from diagnosis to death, it might have prolonged their life. And so for this patient in your data set, you might have observed that they lived a very long time. But if you ignore what happens in between and you simply learn to predict y from X, X being the input, then a new patient comes along and you predicted that new patient is going to survive a long time, and it would be completely the wrong conclusion to say that you don't need to treat that patient. Because, in fact, the only reason the patients like them in the training data lived a long time is because they were treated.

And so when it comes to this field of machine learning and health care, we need to think really carefully about these types of questions because an error in the way that we formalize our problem could kill people because of mistakes like this. Now, other questions are ones about not how do we predict outcomes but how do we guide treatment decisions. So, for example, as data from pathology gets richer and richer and richer, we might think that we can now use computers to try to better predict who is likely to benefit from a treatment than humans could do alone.

But the challenge with using algorithms to do that is that people respond differently to treatment, and the data which is being used to guide treatment is biased based on existing treatment guidelines. So, similarly, to the previous question, we could ask, what would happen if we trained to predict past treatment decisions? This would be the most naive way to try to use data to guide treatment decisions. So maybe you see David gets treatment A, John gets treatment B, Juana gets treatment A. And you might ask then, OK, a new patient comes in, what should this new patient be treated with?

And if you've just learned a model to predict from what you know about the treatment that David is likely to get, then the best that you could hope to do is to do as well as existing clinical practice. So if we want to go beyond current clinical practice, for example, to recognize that there is heterogeneity in treatment response, then we have to somehow change the question that we're asking. I'll give you one last example, which is perhaps a more traditional question of, does X cause y? For example, does smoking cause lung cancer is a major question of societal importance.

Now, you might be familiar with the traditional way of trying to answer questions of this nature, which would be to do a randomized controlled trial. Except this isn't exactly the type of setting where you could do randomized controlled trials. How would you feel if you were a smoker and someone came up to you and said, you have to stop smoking because I need to see what happens? Or how would you feel if you were a non-smoker and someone came up to you and said, you have to start smoking? That would be both not feasible and completely unethical.

And so if we want to try to answer questions like this from data, we need to start thinking about how can we design, using observational data, ways of answering questions like this. And the challenge is that there's going to be bias in the data because of who decides to smoke and who decides not to smoke. So, for example, the most naive way you might try to answer this question would be to look at the conditional likelihood of getting lung cancer among smokers

and getting lung cancer among non-smokers. But those numbers, as you'll see in the next few slides, can be very misleading because there might be confounding factors, factors that would, for example, both cause people to be a smoker and cause them to receive lung cancer, which would differentiate between these two numbers.

And we'll have a very concrete example of this in just a few minutes. So to properly answer all of these questions, one needs to be thinking in terms of causal graphs. So rather than the traditional setup in machine learning where you just have inputs and outputs, now we need to have triplets. Rather than having inputs and outputs, we need to be thinking of inputs, interventions, and outcomes or outputs. So we now need be having three quantities in mind. And we have to start thinking about, well, what is the causal relationship between these three?

So for those of you who have taken more graduate level machine learning classes, you might be familiar with ideas such as Bayesian networks. And when I went to undergrad and grad school and I studied machine learning, for the longest time I thought causal inference had to do with learning causal graphs. So this is what I thought causal inference was about. You have data of the following nature-- 1, 0, 0, 1, dot, dot, dot. So here, there are four random variables. I'm showing the realizations of those four binary variables one per row, and you have a data set like this.

And I thought causal inference had to do with taking data like this and trying to figure out, is the underlying Bayesian network that created that data, is it X1 goes to X2 goes to X3 to X4? Or I'll say, this is X1, that's X2, x3, and X4. Or maybe the causal graph is X1, to X2, to X3, to x4. And trying to distinguish between these different causal graphs from observational data is one type of question that one can ask. And the one thing you learn in traditional machine learning treatments of this is that sometimes you can't distinguish between these causal graphs from the data you have.

For example, suppose you just had two random variables. Because any distribution could be represented by probability of X1 times probability of X2 given X1, according to just rule of conditional probability, and similarly, any distribution can be represented as the opposite, probability of X2 times probability of X1 given X2, which would look like this, the statement that one would make is that if you just had data involving X1 and X2, you couldn't distinguish between these two causal graphs, X1 causes X2 or X2 causes X1.

And usually another treatment would say, OK, but if you have a third variable and you have a

V structure or something like X1 goes to x2, X1 goes to X3, this you could distinguish from, let's say, a chain structure. And then the final answer to what is causal inference from this philosophy would be something like, OK, if you're in a setting like this and you can't distinguish between X1 causes X2 or X2 causes X1, then you do some interventions, like you intervene on X1 and you look to see what happens to X2, and that'll help you disentangle these directions of causality.

None of this is what we're going to be talking about today. Today, we're going to be talking about the simplest, simplest possible setting you could imagine, that graph shown up there. You have three sets of random variables, X, which is perhaps a vector, so it's high dimensional, a single random variable T, and a single random variable Y. And we know the causal graph here. We're going to suppose that we know the directionality, that we know that X might cause T and X and T might cause Y. And the only thing we don't know is the strength of the edges.

All right. And so now let's try to think through this in context of the previous examples. Yeah, question?

AUDIENCE: Just to make sure-- so T does not affect X in any way?

DAVID SONTAG: Correct, that's the assumption we're going to make here. So let's try to instantiate this. So we'll start with this example. X might be what you know about the patient at diagnosis. T, I'm going to assume for the purposes of today's class, is a decision between two different treatment plans. And I'm going to simplify the state of the world. I'm going to say those treatment plans only depend on what you know about the patient at diagnosis.

So at diagnosis, you decide, I'm going to be giving them this sequence of treatments at this three-month interval or this other sequence of treatment at, maybe, that four-month interval. And you make that decision just based on diagnosis and you don't change it based on anything you observe. Then the causal graph of relevance there is, based on what you know about the patient at diagnosis, which I'm going to say X is a vector because maybe it's based on images, your whole electronic health record. There's a ton of data you have on the patient at diagnosis.

Based on that, you make some decision about a treatment plan. I'm going to call that T. T could be binary, a choice between two treatments, it could be continuous, maybe you're deciding the dosage of the treatment, or it could be maybe even a vector. For today's lecture,

I'm going to suppose that T is just binary, just involves two choices. But most of what I'll tell you about will generalize to the setting where T is non-binary as well. But critically, I'm going to make the assumption for today's lecture that you're not observing new things in between.

So, for example, in this whole week's lecture, the following scenario will not happen. Based on diagnosis, you make a decision about treatment plan. Treatment plan starts, you got new observations. Based on those new observations, you realize that treatment plan isn't working and change to another treatment plan, and so on. So that scenario goes by a different name, which is called dynamic treatment regimes or off-policy reinforcement learning, and that we'll learn about next week.

So for today's and Thursday's lecture, we're going to suppose you base on what you know about the patient at this time, you make a decision, you execute the decision, and you look at some outcome. So X causes T, not the other way around. And that's pretty clear because of our prior knowledge about this problem. It's not that the treatment affects what their diagnosis was. And then there's the outcome Y, and there, again, we suppose the outcome, what happens to the patient, maybe survival time, for example, is a function of what treatment they're getting and aspects about that patient.

So this is the causal graph. We know it. But we don't know, does that treatment do anything to this patient? For whom does this treatment help the most? And those are the types of questions we're going to try to answer today. Is the setting clear? OK. Now, these questions are not new questions. They've been studied for decades in fields such as political science, economics, statistics, biostatistics. And the reason why they're studied in those other fields is because often you don't have the ability to intervene, and one has to try to answer these questions from observational data.

For example, you might ask, what will happen to the US economy if the Federal Reserve raises US interest rates by 1%? When's the last time you heard of the Federal Reserve doing a randomized controlled trial? And even if they had done a randomized controlled trial, for example, flipped a coin to decide which way the interest rates would go, it wouldn't be comparable had they done that experiment today to if they had done that experiment two years from now because the state of the world has changed in those years.

Let's talk about political science. I have close colleagues of mine at NYU who look at Twitter, and they want to ask questions like, how can we influence elections, or how are elections

influenced? So you might look at some unnamed actors, possibly people supported by the Russian government, who are posting to Twitter or their social media. And you might ask the question of, well, did that actually influence the outcome of the previous presidential election? Again, in that scenario, it's one of, well, we have this data, something happened in the world, and we'd like to understand what was the effect of that action, but we can't exactly go back and replay to do something else.

So these are fundamental questions that appear all across the sciences, and of course they're extremely relevant in health care, but yet, we don't teach them in our introduction to machine learning classes. We don't teach them in our undergraduate computer science education. And I view this as a major hole in our education, which is why we're spending two weeks on it in this course, which is still not enough. But what has changed between these fields, and what is relevant in health care?

Well, the traditional way in which these questions were asked in statistics were ones where you took a huge amount of domain knowledge to, first of all, make sure you're setting up the problem correctly, and that's always going to be important. But then to think through what are all of the factors that could influence the treatment decisions called the confounding factors. And the traditional approach is one would write down 10, 20 different things, and make sure that you do some analysis, including the analysis I'll show you about in today and Thursday's lecture using those 10 or 20 variables. But where this field is going is one of now having high dimensional data.

So I talked about how you might have imaging data for X, you might have the whole entire patient's electronic health record data facts. And the traditional approaches that the statistics community used to work on no longer work in this high dimensional setting. And so, in fact, it's actually a really interesting area for research, one that my lab is starting to work on and many other labs, where we could ask, how can we bring machine learning algorithms that are designed to work with high dimensional data to answer these types of causal inference questions? And in today's lecture, you'll see one example of reduction from causal inference to machine learning, where we'll be able to use machine learning to answer one of those causal inference questions.

So the first thing we need is some language in order to formalize these notions. So I will work within what's known as the Rubin-Neyman Causal Model, where we talk about what are called potential outcomes. What would have happened under this world or that world? We'll call Y 0,

and often it will be denoted as Y underscore 0, sometimes it'll be denoted as Y parentheses 0, and sometimes it'll be denoted as Y given X comma do Y equals 0. And all three of these notations are equivalent.

So Y is 0 corresponds to what would have happened to this individual if you gave them treatment to 0. And Y1 is the potential outcome of what would have happened to this individual had you gave them treatment one. So you could think about Y1 as being giving the blue pill and Y0 as being given the red pill. Now, once you can talk about these states of the world, then one could start to ask questions of what's better, the red pill or the blue pill? And one can formalize that notion mathematically in terms of what's called the conditional average treatment effect, and this also goes by the name of individual treatment effect.

So it's going to take as input Xi, which I'm going to denote as the data that you had at baseline for the individual. It's the covariance, the features for the individual. And one wants to know, well, for this individual with what we know about them, what's the difference between giving them treatment one or giving them treatment zero? So mathematically, that corresponds to a difference in expectations. It's a difference in expectation of Y1 from Y0. Now, the reason why I'm calling this an expectation is because I'm not going to assume that Y1 and Y0 are deterministic because maybe there's some bad luck component. Like, maybe a medication usually works for this type of person, but with a flip of a coin, sometimes it doesn't work.

And so that's the randomness that I'm referring to when I talk about probability over Y1 given Xi. And so the CATE looks at the difference in those two expectations. And then one can now talk about what the average treatment effect is, which is the difference between those two. So the average treatment effect is now the expectation of-- I'll say the expectation of the CATE over the distribution of people, P of X. Now, we're going to go through this in four different ways in the next 10 minutes, and then you're going to go over it five more ways doing your homework assignment, and you'll go over it two more ways on Friday in recitation.

So if you don't get it just yet, stay with me, you'll get it by the end of this week. Now, in the data that you observe for an individual, all you see is what happened under one of the interventions. So, for example, if the i'th individual in your data set received treatment Ti equals 1, then what you observe, Yi is the potential outcome Y1. On the other hand, if the individual in your data set received treatment Ti equals 0, then what you observed for that individual is the potential outcome Y0.

So that's the observed factual outcome. But one could also talk about the counterfactual of what would have happened to this person had the opposite treatment been done for them. Notice that I just swapped each Ti for 1 minus Ti, and so on. Now, the key challenge in the field is that in your data set, you only observe the factual outcomes. And when you want to reason about the counterfactual, that's where you have to impute this unobserved counterfactual outcome. And that is known as the fundamental problem of causal inference, that we only observe one of the two outcomes for any individual in the data set.

So let's look at a very simple example. Here, individuals are characterized by just one feature, their age. And these two curves that I'm showing you are the potential outcomes of what would happen to this individual's blood pressure if you gave them treatment zero, which is the blue curve, versus treatment one, which is the red curve. All right. So let's dig in a little bit deeper. For the blue curve, we see people who received the control, what I'm calling treatment zero, their blood pressure was pretty low for the individuals who were low and for individuals whose age is high. But for middle age individuals, their blood pressure on receiving treatment zero is in the higher range.

On the other hand, for individuals who receive treatment one, it's the red curve. So young people have much higher, let's say, blood pressure under treatment one, and, similarly, much older people. So then one could ask, well, what about the difference between these two potential outcomes? That is to say the CATE, the Conditional Average Treatment Effect, is simply looking at the distance between the blue curve and the red curve for that individual. So for someone with a specific age, let's say a young person or a very old person, there's a very big difference between giving treatment zero or giving treatment one. Whereas for a middle aged person, there's very little difference.

So, for example, if treatment one was significantly cheaper than treatment zero, then you might say, we'll give treatment one. Even though it's not quite as good as treatment zero, but it's so much cheaper and the difference between them is so small, we'll give the other one. But in order to make that type of policy decision, one, of course, has to understand that conditional average treatment effect for that individual, and that's something that we're going to want to predict using data.

Now, we don't always get the luxury of having personalized treatment recommendations. Sometimes we have to give a policy. Like, for example-- I took this example out of my slides, but I'll give it to you anyway. The federal government might come out with a guideline saying

that all men over the age of 50-- I'm making up that number-- need to get annual prostate cancer screening. That's an example of a very broad policy decision. You might ask, well, what is the effect of that policy now applied over the full population on, let's say, decreasing deaths due to prostate cancer? And that would be an example of asking about the average treatment effect.

So if you were to average the red line, if you were to average the blue line, you get those two dotted lines I show there. And if you look at the difference between them, that is the average treatment effect between giving the red intervention or giving the blue intervention. And if the average human effect is very positive, you might say that, on average, this intervention is a good intervention. If it's very negative, you might say the opposite. Now, the challenge about doing causal inference from observational data is that, of course, we don't observe those red and those blue curves, rather what we observe are data points that might be distributed all over the place.

Like, for example, in this example, the blue treatment happens to be given in the data more to young people, and the red treatment happens to be given in the data more to older people. And that can happen for a variety of reasons. It can happen due to access to medication. It can happen for socioeconomic reasons. It could happen because existing treatment guidelines say that old people should receive treatment one and young people should receive treatment zero. These are all reasons why in your data who receives what treatment could be biased in some way. And that's exactly what this edge from X to T is modeling.

But for each of those people, you might want to know, well, what would have happened if they had gotten the other treatment? And that's asking about the counterfactual. So these dotted circles are the counterfactuals for each of those observations. And by the way, you'll notice that those dots are not on the curves, and the reason they're not on the curve is because I'm trying to point out that there could be some stochasticity in the outcome. So the dotted lines are the expected potential outcomes and the circles are the realizations of them.

All right. Everyone take out a calculator or your computer or your phone, and I'll take out mine. This is not an opportunity to go on Facebook, just to be clear. All you want is a calculator. My phone doesn't-- oh, OK, it has a calculator. Good. All right. So we're going to do a little exercise. Here's a data set on the left-hand side. Each row is an individual. We're observing the individual's age, gender, whether they exercise regularly, which I'll say is a one or a zero, and what treatment they got, which is A or B. On the far right-hand side are their observed

sugar glucose sugar levels, let's say, at the end of the year.

Now, what we'd like to have, it looks like this. So we'd like to know what would have happened to this person's sugar levels had they received medication A or had they received medication B. But if you look at the previous slide, we observed for each individual that they got either A or B. And so we're only going to know one of these columns for each individual. So the first row, for example, this individual received treatment A, and so you'll see that I've taken the observed sugar level for that individual, and since they received treatment A, that observed level represents the potential outcome Ya, or Y0.

And that's why I have a 6, which is bolded under Y0. And we don't know what would have happened to that individual had they received treatment B. So in this case, some magical creature came to me and told me their sugar levels would have been 5.5, but we don't actually know that. It wasn't in the data. Let's look at the next line just to make sure we get what I'm saying. So the second individual actually received treatment B. They're observed sugar level is 6.5. OK.

Let's do a little survey. That 6.5 number, should it be in this column? Raise your hand. Or should it be in this column? Raise your hand. All right. About half of you got that right. Indeed, it goes to the second column. And again, what we would like to know is the counterfactual. What would have been their sugar levels had they received medication A? Which we don't actually observe in our data, but I'm going to hypothesize is-- suppose that someone told me it was 7, then you would see that value filled in there. That's the unobserved counterfactual.

All right. First of all, is the setup clear? All right. Now here's when you use your calculators. So we're going to now demonstrate the difference between a naive estimator of your average treatment effect and the true average treatment effect. So what I want you to do right now is to compute, first, what is the average sugar level of the individuals who got medication B. So for that, we're only going to be using the red ones. So this is conditioning on receiving medication B.

And so this is equivalent to going back to this one and saying, we're only going to take the rows where individuals receive medication B, and we're going to average their observed sugar levels. And everyone should do that. What's the first number? 6.5 plus-- I'm getting 7.875. This is for the average sugar, given that they received medication B. Is that what other people are getting?

AUDIENCE: Yeah.

DAVID SONTAG: OK. What about for the second number? Average sugar, given A? I want you to compute it. And I'm going to ask everyone to say it out loud in literally one minute. And if you get it wrong, of course you're going to be embarrassed. I'm going to try myself. OK. On the count of three, I want everyone to read out what that third number is. One, two, three.

ALL: 7.125.

DAVID SONTAG: All right. Good. We can all do arithmetic. All right. Good. So, again, we're just looking at the red numbers here, just the red numbers. So we just computed that difference, which is point what?

AUDIENCE: 0.75.

DAVID SONTAG: 0.75? Yeah, that looks about right. Good. All right. So that's a positive number. Now let's do something different. Now let's compute the actual average treatment effect, which is we're now going to average every number in this column, and we're going to average every number in this column. So this is the average sugar level under the potential outcome of had the individual received treatment B, and this is the average sugar level under the potential outcome that the individual received treatment A. All right. Who's doing it?

AUDIENCE: 0.75.

DAVID SONTAG: 0.75 is what?

AUDIENCE: The difference.

DAVID SONTAG: How do you know?

AUDIENCE: [INAUDIBLE]

DAVID SONTAG: Wow, you're fast. OK. Let's see if you're right. I actually don't know. OK. The first one is 0.75. Good, we got that right. I intentionally didn't post the slides to today's lecture. And the second one is minus 0.75. All right. So now let's put us in the shoes of a policymaker. The policymaker has to decide, is it a good idea to-- or let's say it's a health insurance company. A health insurance company is trying decide, should I reimburse for treatment B or not? Or should I simply say, no, I'm never going to reimburse for treatment because it doesn't work well?

So if they had done the naive estimator, that would have been the first example, then it would

look like medication B is-- we want lower numbers here, so it would look like medication B is worse than medication A. And if you properly estimated what the actual average treatment effect is, you get the absolute opposite conclusion. You conclude that medication B is much better than medication A. It's just a simple example to really illustrate the difference between conditioning and actually computing that counterfactual.

OK. So hopefully now you're starting to get it. And again, you're going to have many more opportunities to work through these things in your homework assignment and so on. So by now you should be starting to wonder, how the hell could I do anything in this state of the world? Because you don't actually observe those black numbers. These are all unobserved. And clearly there is bias in what the values should be because of what I've been saying all along. So what can we do?

Well, the first thing we have to realize is that typically, this is an impossible problem to solve. So your instincts aren't wrong, and we're going to have to make a ton of assumptions in order to do anything here. So the first assumption is called SUTVA. I'm not even going to talk about it. You can read about that in your readings. I'll tell you about the two assumptions that are a little bit easier to describe. The first critical assumption is that there are no unobserved confounding factors. Mathematically what that's saying is that your potential outcomes, Y0 and Y1, are conditionally independent of the treatment decision given what you observe on the individual, X.

Now, this could be a bit hard to-- and that's called ignorability. And this can be a bit hard to understand, so let me draw a picture. So X is your covariance, T is your treatment decision. And now I've drawn for you a slightly different graph. Over here I said X goes to T, X and T go to Y. But now I don't have Y. Instead, I have Y0 and Y1, and I don't have any edge from T to them. And that's because now I'm actually using the potential outcomes notation. Y0 is a potential outcome of what would have happened to this individual had they received treatment 0, and Y1 is what would have happened to this individual if they received treatment one.

And because you already know what treatment the individual has received, it doesn't make sense to talk about an edge from T to those values. That's why there's no edge there. So then you might wonder, how could you possibly have a violation of this conditional independence assumption? Well, before I give you that answer, let me put some names to these things. So we might think about X as being the age, gender, weight, diet, and so on of the individual. T might be a medication, like an anti-hypertensive medication to try to lower a patient's blood

pressure. And these would be the potential outcomes after those two medications.

So an example of a violation of ignorability is if there is something else, some hidden variable h, which is not observed and which affects both the decision of what treatment the individual in your data set receives and the potential outcomes. Now it should be really clear that this would be a violation of that conditional independence assumption. In this graph, Y0 and Y1 are not conditionally independent of T given X. All right. So what are these hidden confounders? Well, they might be things, for example, which really affect treatment decisions.

So maybe there's a treatment guideline saying that for diabetic patients, they should receive treatment zero, that that's the right thing to do. And so a violation of this would be if the fact that the patient's diabetic were not recorded in the electronic health record. So you don't know-- that's not up there. You don't know that, in fact, the reason the patient received treatment T was because of this h factor. And there's critically another assumption, which is that h actually affects the outcome, which is why you have these edges from h to the Y's.

If h were something which might have affected treatment decision but not the actual potential outcomes-- and that can happen, of course. Things like gender can often affect treatment decisions, but maybe, for some diseases, it might not affect outcomes. In that situation it wouldn't be a confounding factor because it doesn't violate this assumption. And, in fact, one would be able to come up with consistent estimators of average treatment effect under that assumption. Where things go to hell is when you have both of those edges. All right.

So there can't be any of these h's. You have to observe all things that affect both treatment and outcomes. The second big assumption-- oh, yeah. Question?

AUDIENCE: In practice, how good of a model is this?

DAVID SONTAG: Of what I'm showing you here?

AUDIENCE: Yeah.

DAVID SONTAG: For hypertension?

AUDIENCE: Sure.

DAVID SONTAG: I have no idea. But I think what you're really trying to get at here in asking your question, how good of a model is this, is, well, oh, my god, how do I know if I've observed everything? Right?

All right. And that's where you need to start talking to domain experts. So this is my starting place where I said, no, I'm not going to attempt to fit the causal graph. I'm going to assume I know the causal graph and just try to estimate the effects. That's where this starts to become really irrelevant. Because if you notice, this is another causal graph, not the one I drew on the board.

And so that's something where, really, talking with domain experts would be relevant. So if you say, OK, I'm going to be studying hypertension and this is the data I've observed on patients, well, you can then go to a clinician, maybe a primary care doctor who often treats patients with hypertension, and you say, OK, what usually affects your treatment decisions? And you get a set of variables out, and then you check to make sure, am I observing all of those variables, at least the variables that would also affect outcomes? So, often, there's going to be a back and forth in that conversation to make sure that you've set up your problem correctly.

And again, this is one area where you see a critical difference between the way that we do causal inference from the way that we do machine learning. Machine learning, if there's some unobserved variables, so what? I mean, maybe your predictive accuracy isn't quite as good as it could have been, but whatever. Here, your conclusions could be completely wrong if you don't get those confounding factors right. Now, in some of the optional readings for Thursday's lecture-- and we'll touch on it very briefly on Thursday, but there's not much time in this course-- I'll talk about ways and you'll read about ways to try to assess robustness to violations of these assumptions. And those go by the name of sensitivity analysis.

So, for example, the type of question you might ask is, how would my conclusions have changed if there were a confounding factor which was blah strong? And that's something that one could try to answer from data, but it's really starting to get beyond the scope of this course. So I'll give you some readings on it, but I won't be able to talk about it in the lecture. Now, the second major assumption that one needs is what's known as common support. And by the way, pay close attention here because at the end of today's lecture-- and if I forget, someone must remind me-- I'm going to ask you where did these two assumptions come up in the proof that I'm about to give you.

The first one I'm going to give you will be a dead giveaway. So I'm going to answer to you where ignorability comes up, but it's up to you to figure out where does common support show up. So what is common support? Well, what common support says is that there always must be some stochasticity in the treatment decisions. For example, if in your data patients only

receive treatment A and no patient receives treatment B, then you would never be able to figure out the counterfactual, what would have happened if patients receive treatment B.

But what happens if it's not quite that universal but maybe there is classes of people? Some individual is X, let's say, people with blue hair. People with blue hair always receive treatment zero and they never see treatment one. Well, for those people, if for some reason something about them having blue hair was also going to affect how they would respond to the treatment, then you wouldn't be able to answer anything about the counterfactual for those individuals. This goes by the name of what's called a propensity score. It's the probability of receiving some treatment for each individual.

And we're going to assume that this propensity score is always bounded between 0 and 1. So it's between 1 minus epsilon and epsilon for some small epsilon. And violations of that assumption are going to completely invalidate all conclusions that we could draw from the data. All right. Now, in actual clinical practice, you might wonder, can this ever hold? Because there are clinical guidelines. Well, a couple of places where you'll see this are as follows.

First, often, there are settings where we haven't the faintest idea how to treat patients, like second line diabetes treatments. You know that the first thing we start with is metformin. But if metformin doesn't help control the patient's glucose values, there are several second line diabetic treatments. And right now, we don't really know which one to try. So a clinician might start with treatments from one class. And if that's not working, you try a different class, and so on. And it's a bit random which class you start with for any one patient.

In other settings, there might be good clinical guidelines, but there is randomness in other ways. For example, clinicians who are trained on the west coast might be trained that this is the right way to do things, and clinicians who are trained in the east coast might be trained that this is the right way to do things. And so even if any one clinician's treatment decisions are deterministic in some way, you'll see some stochasticity now across clinicians. It's a bit subtle how to use that in your analysis, but trust me, it can be done.

So if you want to do causal inference from observational data, you're going to have to first start to formalize things mathematically in terms of what is your X, what is your T, what is your Y. You have to think through, do these choices satisfy these assumptions of ignorability and overlap? Some of these things you can check in your data. Ignorability you can't explicitly check in your data. But overlap, this thing, you can test in your data. By the way, how? Any

idea? Someone else who hasn't spoken today.

So just think back to the previous example. You have this table of these X's and treatment A or B and then sugar values. How would you test this?

AUDIENCE: You could use a frequentist approach and just count how many things show up. And if there is zero, then you could say that it's violated.

DAVID SONTAG: Good. So you have this table. I'll just go back to that table. We have this table, and these are your X's. Actually, we'll go back to the previous slide where it's a bit easier to see. Here, we're going to ignore the outcome, the sugar levels because, remember, this only has to do with probability of treatment given your covariance. The Y doesn't show up here at all. So this thing on the right-hand side, the observed sugar levels, is irrelevant for this question. All we care about is what goes on over here.

So we look at this. These are your X's, and this is your treatment. And you can look to see, OK, here you have one 75-year-old male who does exercise frequently and received treatment A. Is there any one else in the data set who is 75 years old and male, does exercise regularly but received treatment B? Yes or no? No. Good. OK. So overlap is not satisfied here, at least not empirically. Now, you might argue that I'm being a bit too coarse here.

Well, what happens if the individual is 74 and received treatment B? Maybe that's close enough. So there starts to become subtleties in assessing these things when you have finite data. But it is something at the fundamental level that you could start to assess using data. As opposed to ignorability, which you cannot test using data. All right. So you have to think about, are these assumptions satisfied? And only once you start to think through those questions can you start to do your analysis.

And so that now brings me to the next part of this lecture, which is how do we actually-- let's just now believe David, believe that these assumptions hold. How do we do that causal inference? Yeah?

AUDIENCE: I just had a question on [INAUDIBLE]. If you know that some patients, for instance, healthy patients, are not tracking to get any treatment, should we just remove them, basically?

DAVID SONTAG: So the question is, what happens if you have a violation of overlap? For example, you know that healthy individuals never receive any treatment. Should you remove them from your data set? Well, first of all, that has to do with how do you formalize the question because not

receiving a treatment is a treatment. So that might be your control arm, just to be clear. Now, if you're asking about the difference between two treatments-- two different classes of treatment for a condition, then often one defines the relevant inclusion criteria in order to have these conditions hold.

For example, we could try to redefine the set of individuals that we're asking about so that overlap does hold. But then in that situation, you have to just make sure that your policy is also modified. You say, OK, I conclude that the average treatment effect is blah for this type of people. OK? OK. So how could we possibly compute the average treatment effect from data? Remember, average treatment effect, mathematically, is the expectation between potential outcome Y1 minus Y0.

The key tool which we'll use in order to estimate that is what's known as the adjustment formula. This goes by many names in the statistics community, such as the G-formula as well. Here, I'll give you a derivation of it. We're first going to recognize that this expectation is actually two expectations in one. It's the expectation over individuals X and it's the expectation over potential outcomes Y given X. So I'm first just going to write it out in terms of those two expectations, and I'll write the expectations related to X on the outside. That goes by name of law of total expectation.

This is trivial at this stage. And by the way, I'm just writing out expectation of Y1. In a few minutes, I'll show you expectation of Y0, but it's going to be exactly analogous. Now, the next step is where we use ignorability. I told you I was going to give that one away. So remember, we said that we're assuming that Y1 is conditionally independent of the treatment T given X. What that means is probability of Y1 given X is equal to probability of Y1 given X comma T equals whatever-- in this case I'll just say T equals 1. This is implied by Y1 being conditionally independent of T given X.

So I can just stick n comma T equals 1 here, and that's explicitly because of ignorability holding. But now we're in a really good place because notice that-- and here I've just done some short notation. I'm just going to hide this expectation. And by the way, you could do the same for Y0-- Y1, Y0. And now notice that we can replace this average human effect with now this expectation with respect to all individuals X of the expectation of Y1 given X comma T equals 1, and so on. And these are mostly quantities that we can now observe from our data.

So, for example, we can look at the individuals who received treatment one, and for those

individuals we have realizations of Y1. We can look at individuals who receive treatment zero, and for those individuals we have realizations of Y0. And we could just average those realizations to get estimates of the corresponding expectations. So these we can easily estimate from our data. And so we've made progress. We can now estimate some part of this from our data.

But notice, there are some things that we can't yet directly estimate from our data. In particular, we can't estimate expectation of Y0 given X comma T equals 1 because we have no idea what would have happened to this individual who actually got treatment one if they had gotten treatment zero. So these we don't know. So these we don't know. Now, what is the trick I'm planning on you? How does it help that we can do this?

Well, the key point is that these quantities that we can estimate from data show up in that term. In particular, if you look at the individuals X that you've sampled from the full set of individuals P of X, for that individual X for which, in fact, we observed T equals 1, then we can estimate expectation of Y1 given X comma T equals 1, and similarly for Y0. But what we need to be able to do is to extrapolate. Because empirically, we only have samples from P of X given T equals 1, P of X given T equals 0 for those two potential outcomes correspondingly.

But we are going to also get samples of X such that for those individuals in your data set, you might have only observed T equals 0. And to compute this formula, you have to answer, for that X, what would it have been if they got treatment equals one? So there are going to be a set of individuals that we have to extrapolate for in order to use this adjustment formula for estimate. Yep?

AUDIENCE: I thought because common support is true, we have some patients that received each treatment or a given type of X.

DAVID SONTAG: Yes. But now-- so, yes, that's true. But that's a statement about infinite data. And in reality, one only has finite data. And so although common support has to hold to some extent, you can't just build on that to say that you always observe the counterfactual for every individual, such as the pictures I showed you earlier. So I'm going to leave this slide up for just one more second to let it sink in and see what it's saying.

We started out from the goal of computing the average treatment effect, expected value of Y1 minus Y0. Using the adjustment formula, we've gotten to now an equivalent representation, which is now an expectation with respect to all individuals sampling from P of X of expected

value of Y1 given X comma T equals 1, expected value of Y0 given X comma T equals 0. For some of the individuals, you can observe this, and for some of them, you have to extrapolate. So from here, there are many ways that one can go. Hold your question for a little while.

So types of causal inference methods that you will have heard of include things like covariance adjustment, propensity score re-weighting, doubly robust estimators, matching, and so on. And those are the tools of the causal inference trade. And in this course, we're only going to talk about the first two. And in today's lecture, we're only going to talk about the first one, covariate adjustment. And on Thursday, we'll talk about the second one. So covariate adjustment is a very natural way to try to do that extrapolation. It also goes by the name, by the way, of response surface modeling.

What we're going to do is we're going to learn a function f, which takes as an input X and T, and its goals is to predict Y. So intuitively, you should think about f as this conditional probability distribution. It's predicting Y given X and T. So T is going to be an input to the machine learning algorithm, which is going to predict what would be the potential outcome Y for this individual described by feature as X1 through Xd under intervention T.

So this is just from the previous slide. And what we're going to do now are-- this is now where we get the reduction to machine learning-- is we're going to use empirical risk minimization, or maybe some regularized empirical risk minimization, to fit a function f which approximates the expected value of YT given capital T equals little t. Got my X. And then once you have that function, we're going to be able to use that to estimate the average treatment effect by just implementing now this formula here.

So we're going to first take an expectation with respect to the individuals in the data set. So we're going to approximate that with an empirical expectation where we sum over the little n individuals in your data set. Then what we're going to do is we're going to estimate the first term, which is f of Xi comma 1 because that is approximating the expected value of Y1 given T comma X-- T equals 1 comma X. And we're going to approximate the second term, which is just plugging now 0 for T instead of 1. And we're going to take the difference between them, and that will be our estimator of the average treatment effect.

Here's a natural place to ask a question. One thing you might wonder is, in your data set, you actually did observe something for that individual, right. Notice how your raw data doesn't show up in this at all. Because I've done machine learning, and then I've thrown away the

observed Y's, and I used this estimator. So what you could have done-- an alternative formula, which, by the way, is also a consistent estimator, would have been to use the observed Y for whatever the factual is and the imputed Y for the counterfactual using f.

That would have been that would have also been a consistent estimator for the average treatment effect. You could've done either. OK. Now, sometimes you're not interested in just the average treatment effect, but you're actually interested in understanding the heterogeneity in the population. Well, this also now gives you an opportunity to try to explore that heterogeneity. So for each individual Xi, you can look at just the difference between what f predicts for treatment one and what X predicts given treatment zero. And the difference between those is your estimate of your conditional average treatment effect.

So, for example, if you want to figure out for this individual, what is the optimal policy, you might look to see is CATE positive or negative, or is it greater than some threshold, for example? So let's look at some pictures. Now what we're using is we're using that function f in order to impute those counterfactuals. And now we have those observed, and we can actually compute the CATE. And averaging over those, you can estimate now the average treatment effect. Yep?

AUDIENCE: How is f non-biased?

DAVID SONTAG: Good. So where can this go wrong? So what do you mean by biased, first? I'll ask that.

AUDIENCE: For instance, as we've seen in the paper like pneumonia and people who have asthma, [INAUDIBLE]

DAVID SONTAG: Oh, thank you so much for bringing that back up. So you're referring to one of the readings for the course from several weeks ago, where we talked about using just a pure machine learning algorithm to try to predict outcomes in a hospital setting. In particular, what happens for patients who have pneumonia in the emergency department? And if you all remember, there was this asthma example, where patients with asthma were predicted to have better outcomes than patients without asthma.

And you're calling that bias. But you remember, when I taught about this, I called it biased due to a particular thing. What's the language I used? I said bias due to intervention, maybe, is what I-- I can't remember exactly what I said.

[LAUGHTER]

I don't know. Make it up. Now a textbook will be written with bias by intervention. OK. So the problem there is that they didn't formulize the prediction problem correctly. The question that they should have asked is, for asthma patients-- what you really want to ask is a question of X and then T and Y, where T are the interventions that are done for asthmatics.

So the failure of that paper is that it ignored the causal inference question which was hidden in the data, and it just went to predict Y given X marginalizing over T altogether. So T was never in the predictive model. And said differently, they never asked counterfactual questions of what would have happened had you done a different T. And then they still used it to try to guide some treatment decisions. Like, for example, should you send this person home, or should you keep them for careful monitoring or so on? So this is exactly the same example as I gave in the beginning of the lecture, where I said if you just use a risk stratification model to make some decisions, you run the risk that you're making the wrong decisions because those predictions were biased by decisions in your data.

So that doesn't happen here because we're explicitly accounting for T in all of our analysis. Yep?

AUDIENCE: In the data sets that we've used, like MIMIC, how much treatment information exists?

DAVID SONTAG: So how much treatment information is in MIMIC? A ton. In fact, one of the readings for next week is going to be about trying to understand how one could manage sepsis, which is a condition caused by infection, which is managed by, for example, giving broad spectrum antibiotics, giving fluids, giving pressers and ventilators. And all of those are interventions, and all those interventions are recorded in the data so that one could then ask counterfactual questions from the data, like what would have happened if this patient had they received a different set of interventions? Would we have prolonged their life, for example?

And so in an intensive care unit setting, most of the questions that we want to ask about, not all, but many of them are about dynamic treatments because it's not just a single treatment but really about a service sequence of treatments responding to the current patient condition. And so that's where we'll really start to get into that material next week, not in today's lecture. Yep?

AUDIENCE: How do you make sure that your f function really learned from the relationship between T and the outcome?

DAVID SONTAG: That's a phenomenal question. Where were you this whole course? Thank you for asking it. So I'll repeat it. How do you know that your function f actually learned something about the relationship between the input X and the treatment T and the outcome? And that really gets to the question of, is my reduction actually valid? So I've taken this problem and I've reduced it to this machine learning problem, where I take my data, and literally I just learn a function f to try to predict well the observations in the data.

And how do we know that that function f actually does a good job at estimating something like average treatment effect? In fact, it might not. And this is where things start to get really tricky, particularly with high dimensional data. Because it could happen, for example, that your treatment decision is only one of a huge number of factors that affect the outcome Y. And it could be that a much more important factor is hidden in X. And because you don't have much data, and because you have to regularize your learning algorithm, let's say, with L1 or L2 regularization or maybe early stopping if you're using deep neural network, your algorithm might never learn the actual dependence on T.

It might learn just to throw away T and just use X to predict Y. And if that's the case, you will never be able to infer these average treatment effects accurately. You'll have huge errors. And that gets back to one of the slides that I skipped, where I started out from this picture. This is the machine learning picture saying, OK, a reduction to machine learning is-- now you add an additional feature, which is your treatment decision, and you learn that black box function f. But this is where machine learning causal inference starts to differ because we don't actually care about the quality of predicting Y.

We can measure your root mean squared error in predicting Y given your X's and T's, and that error might be low. But you can run into these failure modes where it just completely ignores T, for example. So T is special here. So really, the picture we want to have in mind is that T is some parameter of interest. We want to learn a model f such that if we twiddle T, we can see how there is a differential effect on Y based on twiddling T. That's what we truly care about when we're using machine learning for causal inference.

And so that's really the gap, that's the gap in our understanding today. And it's really an active area of research to figure out how do you change the whole machine learning paradigm to recognize that when you're using machine learning for causal inference, you're actually interested in something a little bit different. And by the way, that's a major area of my lab's research, and we just published a series of papers trying to answer that question. Beyond the

scope of this course, but I'm happy to send you those papers if anyone's interested.

So that type of question is extremely important. It doesn't show up quite as much when your X's aren't very high dimensional and where things like regularization don't become important. But once your X becomes high dimensional and once you want to start to consider more and more complex f's during your fitting, like you want to use deep neural networks, for example, these differences in goals become extremely important.

So there are other ways in which things can fail. So I want to give you here an example where-- shoot, I'm answering my question. OK. No one saw that slide. Question-- where did the overlap assumptions show up in our approach for estimating average treatment effect using covariate adjustment? Let me go back to the formula. Someone who hasn't spoken today, hopefully. You can be wrong, it's fine. Yeah, in the back?

AUDIENCE: Is it the version with the same age in receiving treatment B and treatment B?

DAVID SONTAG: So maybe you have an individual with some age-- we're going to want to be able to look at the difference between what f predicts for that individual if they got treatment A versus treatment B, or one versus zero. And let me try to lead this a little bit. And it might happen in your data set that for individuals like them, you only ever observe treatment one and there's no one even remotely like them who you observe treatment zero. So what's this function going to output then when you input zero for that second argument? Everyone say out loud. Garbage? Right?

If in your data set you never observed anyone even remotely similar to Xi who received treatment zero, then this function is basically undefined for that individual. I mean, yeah, your function will output something because you fit it, but it's not going to be the right answer. And so that's where this assumption starts to show up. When one talks about the sample complexity of learning these functions f to do covariate adjustment, and when one talks about the consistency of these arguments-- for example, you'd like to be able to make claims that as the amount of data grows to, let's say, infinity, that this is the right answer-- gives you the right estimate. So that's the type of proof which is often given in the causal inference literature.

Well, if you have overlap, then as the amount of data goes to infinity, you will observe someone, like the person who received treatment one, you'll observe someone who also received treatment zero. It might have taken you a huge amount of data to get there because treatment zero might have been much less likely than treatment one. But because the

probability of treatment zero is not zero, eventually you'll see someone like that. And so eventually you'll get enough data in order to learn a function which can extrapolate correctly for that individual.

And so that's where overlap comes in in giving that type of consistency argument. Of course, in reality, you never have infinite data. And so these questions about trade-offs between the amount of data you have and the fact that you never truly have empirical overlap with a small amount of data, and answering when can you extrapolate correctly despite that is the critical question that one needs to answer, but is, by the way, not studied very well in the literature because people don't usually think in terms of sample complexity in that field. That's where computer scientists can start really to contribute to this literature and bringing things that we often think about in machine learning to this new topic.

So I've got a couple of minutes left. Are there any other questions, or should I introduce some new material in one minute? Yeah?

AUDIENCE: So you said that the average treatment effect estimator here is consistent. But does it matter if we choose the wrong-- do we have to choose some functional form of the features to the effect?

DAVID SONTAG: Great question.

AUDIENCE: Is it consistent even if we choose a completely wrong function or formula?

DAVID SONTAG: No.

AUDIENCE: That's a different thing?

DAVID SONTAG: No, no. You're asking all the right questions. Good job today, everyone. So, no. If you walk through that argument I made, I assume two things. First, that you observe enough data such that you can have any chance of extrapolating correctly. But then implicit in that statement is that you're choosing a function family which is powerful enough that it can extrapolate correctly. So if your true function is-- if you think back to this figure I showed you here, if the true potential outcome functions are these quadratic functions and you're fitting them with a linear function, then no matter how much data you have you're always going to get wrong estimates because this argument really requires that you're considering more and more complex non-linearity as your amount of data grows.

So now here's a visual depiction of what can go wrong if you don't have overlap. So now I've taken out-- previously, I had one or two red points over here and one or two blue points over here, but I've taken those out. So in your data all you have are these blue points and those red points. So all you have are the points, and now one can learn as good functions, as you can imagine, to try to, let's say, minimize the mean squared error of predicting these blue points and minimize the mean squared error of predicting those red points. And what you might get out is something-- maybe you'll decide on a linear function. That's as good as you could do if all you have are those red points.

And so even if you were willing to consider more and more complex hypothesis classes, here, if you tried to consider a more complex hypothesis class than this line, you'd probably just over-fitting to the data you have. And so you decide on that line, which, because you had no data over here, you don't even know that it's not a good fit to the data. And then you notice that you're getting completely wrong estimates. For example, if you asked about the CATE for a young person, it would have the wrong sign over here because they flipped, the two lines.

So that's an example of how one can start to get errors. And when we begin on Thursday's lecture, we're going to pick up right where we left off today, and I'll talk about this issue a little bit more in detail. I'll talk about how, if one were to learn a linear function, how one could actually, under the assumption that the true potential outcomes are linear, how one could actually interpret the coefficients of that linear function in a causal way under the very strong assumption that the two potential outcomes are linear. So that's what we'll return to on Thursday.