[SQUEAKING]

[RUSTLING]

[CLICKING]

DAVID SONTAG: So today's lecture is going to continue on the lecture that you saw on Tuesday, which was introducing you to causal inference. So the causal inference setting, which we're studying in this course, is a really simplistic one from a causal graphs perspective. There are three sets of variables of interest-- everything you know about an individual or patient, which we're calling x over here; and intervention or action-- which for today's lecture, we're going to suppose that it's either 0 or 1, so a binary intervention. You either take it or don't-- and an outcome y.

And what makes this problem of understanding the impact of the intervention on the outcome challenging is that we have to make that inference from observational data, where we don't have the ability-- at least not in medicine, we typically don't have the ability to make active interventions. And the goal of what we will be discussing in this course is about how to take data that was collected from a practice of medicine where actions or interventions were taken, and then use that to infer something about the causal effect. And obviously, there are also randomized control trials where one intentionally does randomize, but the focus of today's lecture is going to be using observational data, or ready collected data, to try to make these conclusions.

So we introduced the language of potential outcomes on Tuesday. Potential outcomes is the mathematical framework for trying to answer these questions. Then with that definition of potential outcomes, we can define the conditional average treatment effect, which is the difference between Y1 and Y0 for the individual Xi.

So you'll notice here that I have patients, so treating the potential outcome as a random variable in case there might be some stochasticity. So sometimes, maybe if you were to give someone a treatment, it works, and sometimes it doesn't. So that's what the expectation is accounting for. Any questions before I move on?

So with respect to this definition of conditional average treatment effect, then you could ask, well, what would happen in aggregate for the population? And you can compute that by taking the average of the conditional average treatment effect over all of the individuals. So that's just this expectation with respect to, now, p of x.

Now, critically, this distribution, p of x, you should think about as the distribution of everyone that exists in your data. So some of those individuals might have received treatment 1 in the past. Some of them might have received treatment 0. But when we ask this question about average treatment effect, we're asking, for both of those populations, what would have been the effect-- what would have been the difference about [INAUDIBLE] they received treatment 1 minus had they received treatment 0?

Now, I wanted to take this opportunity to start thinking a little bit bigger picture about how causal inference can be important in a variety of societal questions, and so I'd like to now spend just a couple of minutes thinking with you about what some causal questions might be that we urgently need to answer about the COVID-19 pandemic. And as you try to think through these questions, I want you to have this causal graph in mind. So there is the general population. There is some action that you want to perform, and the whole notion of causal inferences assessing the effective action on some outcome of interest. So in trying to give the answer to my-- various answers to my questions of what are some causal inference questions of relevance to the current pandemic, I want you to try to frame your answers in terms of these Xs, Ts, and Ys.

It's also, obviously, very hard to answer using the types of techniques that we will be discussing in this course, and partly because the techniques that I'm focusing on are very much data driven techniques. That said, the general framework that I've introduced on Tuesday for covariate adjustment of, come up with a model and use that model to make a prediction, and the assumptions that underlie that in terms of, well, where's that model coming from, if you're fitting the parameters from data, having to have common support in order to be able to have any trust in the downstream conclusions. Those underlying assumptions and the general premises will still hold, but here, obviously, when it comes to something like social distancing, they're complicated network effects.

And so whereas up until now, we've been making the assumption of what was called SUTVA-- it was a assumption that I probably didn't even talk about in Tuesday's lecture. But intuitively, what the SUTVA assumption says is that each of your training examples are independent of each other. And that might make sense when you think about, give a patient a medication or not, but it certainly doesn't make sense when you think about social distancing type measures, where if some people social distance, but other people don't, it has obviously a very different impact on society. So one needs a different class of models to try to think about that, which have to relax that SUTVA assumption.

So those were all really good answers to my question, and in some sense, now-- so there's the epidemiological type questions that we last spoke about. But the first few set of questions about, really, how does one treat patients who have COVID are the types of questions that only now we can really start to answer now, unfortunately, because we're starting to get a lot of data in the United States and internationally. And so for example, my own personal research group, we're starting to really scale up our research on these types of questions.

Now, one very simplified example that I wanted to give of how a causal inference lens can be useful here is by trying to understand case fatality rates. So for example, in Italy, it was reported that 4.3% of individuals who had this condition passed away, whereas in China, it was reported that 2.3% of individuals who had this condition passed away. Now, you might ask, based on just those two numbers, is something different about China?

For example, might it be that the way that COVID is being managed in China is better than in Italy? You might also wonder if the strain of the disease might be different between China and Italy? So perhaps there were some mutations since it left Wuhan.

But if you dig a little bit deeper, you see that, if you plot case fatality rates by age group, you get this plot that I'm showing over here. And you see that if you compare Italy, which is the orange, to China, which is blue, now stratified by age range, you see that for every single age range, the percentage of deaths is lower in Italy than in China, which would seem to be a contradiction with what we saw-- with the aggregate numbers, where we see that the case fatality rate in Italy is higher than in China. And so the reason why this can happen has to do with the fact that the populations are very different.

And by the way, this paradox goes by the name of Simpson's paradox. So if you dig a bit deeper, you see then that, if you're to look at, well, what is the distribution of individuals in China and Italy that have been reported to have COVID, you see that, in Italy, it's much more highly weighted towards these older ages. And if you then combine that with the total number of cases, you get you get to these discrepancies, so it now fully explains these two numbers and the plot that you see.

Now if we're to try to think about this a bit more formally, we would try to formalize it in terms of following causal graph. And so here, we have the same notions of X, T, and Y, where X is the age of an individual who has been diagnosed with COVID. T is now country, so we're going to think about the intervention here as transporting ourselves from China to Italy, so thinking about changing the environment altogether. And Y is the outcome on an individual level basis. And so the formal question that one might want to ask is about a causal impact of changing the country on the outcome Y.

Now, for this particular causal question, this causal graph that I'm drawing here is the wrong one, and in fact, the right causal graph probably has an edge that goes from T to X. In particular, the distribution of individuals in the country is obviously a function of the country, not the other way around. But despite the fact that there is that difference in directionality, all of the techniques that we've been teaching you in this course are still applicable for trying to ask a causal question about the impact of intervening on a country, and that's really because, in some sense, these two distributions, at an observational level, are equivalent.

And if you want to dig a little bit deeper into this example-- and I want to stress this is just for educational purposes. Don't read anything into these numbers-- I would go to this Colab notebook after the course. So all of this was just a little bit of set up to help frame where causal inference shows up and some things that we've been thinking and really very worried and stressed about ourselves personally recently. And I want to now shift gears to starting to get back to the course material, and in particular, I want to start today's more theoretical parts of the lectures by returning to covariate adjustment, which we ended on and Tuesday.

In covariate adjustment, one-- we'll use a machine learning approach to learn some model, which I'll call F. So you could imagine a black box machine learning algorithm, which takes as input both X and T. So X are your covariates of the individual that are going to receive the treatment, and T is that treatment decision, which for today's lecture, you can just assume is binary 01, and uses those together now to predict the outcome Y.

Now, what we showed on Tuesday was that, under ignorability, where ignorability, remember, was the assumption of no hitting confounding, then the conditional average treatment effect could be defined as just a difference-- could be could be computed as the expectation of Y1 now conditioned on T equals 1, so this is the piece that I've added in here, and minus the expectation of Y0 now conditioned on T equal 0. And it's that conditioning which is really important, because that's what enables you to estimate Y1 from data where treatment 1 was observed, whereas you never get to observe Y1 in data when treatment 0 was performed. So we have this formula, and after fitting that model F, one could then use it to try to estimate CATE by just taking that learned function, plugging in the number 1 for the treatment variable in order to get your estimate of this expectation, and then plugging in the number 0 for the treatment variable when you want to get your estimate of this expectation.

Taking the difference between those then gives you your estimate of the conditional average treatment effect. So that's the approach, and what we didn't talk about so much was the modeling choices of what should your function class be. So this is going to turn out to be really important, and really, the punchline of the next several slides is going to be a major difference in philosophy between machine learning and statistics, and between prediction and causal inference.

So let's now consider the following simple model, where I'm going to assume that the ground truth in the real world has that the potential outcome YT of X, where T, again is the treatment, is equal to some simple linear model involving the covariates X and the treatments T, the treatment T. So in this very simple setting, I'm going to assume that we just have a single feature or covariate for the individual, which is there age. I'm going to assume that this model doesn't have any terms with an interaction between X and T, so it's fully linear in X and T.

So this is an assumption about the true potential outcomes, and what we'll do over the next couple of slides is think about what would happen if you now modeled Y of T, so modeling it with some function F, where F was, let's say, a linear function versus a nonlinear function, if F took this form or a different form. And by the way, I'm going to assume that the noise here, epsilon t, can be arbitrary, but that it has 0 mean. So let's get started by trying to estimate what the true CATE is, or Conditional Average Treatment Effect, for this potential outcome model.

Well, just by definition, the CATE is the expectation of Y1 minus Y0. We're going to take this formula, and we're going to plug it in for the first term using T equals 1, and that's why you get this term over here with gamma. And the gamma is because, again, T is equal to 1. We're also going to take this, and we're going to plug it in for, now, this term over here, where T is equal to 0. And when T is equal to 0, then the gamma term just disappears, and so you just get beta X plus epsilon 0.

So all I've done so far is plug in the Y1 and Y0 according to the assumed form, but notice now that there's some terms that cancel out-- in particular, the beta X term over here cancels out with a beta X term over here. And because epsilon 1 has a 0 mean, and epsilon 0 also has a 0 mean. The only thing left is that gamma term, and expectation of a constant's obviously that constant. And so what we conclude from this is that the CATE value is gamma. Now, the average treatment effect, which is the average of CATE over all individuals X, will then also be gamma, obviously.

So we've done something pretty interesting here. We've started from the assumption that the true potential outcome model is linear, and what we concluded is that the average treatment effect is precisely the coefficient of the treatment variable in this linear model. So what that means is that, if what you're interested in is causal inference, and suppose that we were lucky enough to know that the true model were linear, and so we attempted to fit some function F, which had precisely the same form, we get some beta hats and some gamma hats out from the learning algorithm, all we need to do is look at that gamma hat in order to conclude something about the average treatment effect. No need to do this complicated thing of plugging in to estimate CATEs. And again, the reason it's such a trivial conclusion is because of our assumption of linearity.

Now, what that also means is that, if you have errors in learning-- in particular, suppose, for example, that you are estimating your gamma hat wrongly, then that means you're also going to be getting wrong your estimates of your conditional and average treatment effects. There's a question here, which I was lucky enough to see, that says, what does gamma represent in terms of the medication? Thank you for that question.

So gamma is-- literally speaking, gamma tells you the conditional average treatment effect, meaning if you were to give the treatment versus not giving the treatment, how that affects the outcome. Think about the outcome of interest being the patient's blood pressure, there being potential confounding factor of the patient's age, and T being one of two different blood pressure measurements. If gamma is positive, then it means that treatment 1 is more-- treatment 1 increases the patient's blood pressure relative to treatment 0. And if gamma is negative, it means that treatment 1 decreases the patient's blood pressure relative to treatment 0.

So in machine learning-- oh, sorry, there's another chat. Thank you, good. So in machine learning, I typically tell my students, don't attempt to interpret your coefficient. At least, don't interpret them too much. Don't put too much weight into them, and that's because, when you're learning very high dimensional models, there can be a lot of redundancy between your features.

But when you talk to statisticians, often they pay really close attention to their coefficients, and they try to interpret those coefficients often with the causal lens. And when I first got started in this field, I couldn't understand why are they paying attention to those coefficients so much? Why are they coming up with these causal hypotheses based on which coefficients are positive and which are the negative? And this is the answer.

It really comes down to an interpretation of the prediction problem in terms of the feature of relevance being a treatment, that treatment being linear with respect to the potential outcome, and then looking at the coefficient of the treatment as telling you something about the average treatment effect of that intervention or treatment. Moreover, that also tells us why it's often very important to look at confidence intervals, so one might want to know, we have some small data set, we get some estimate of gamma hat, but what if you had a different data set? So what happens if you had a new sample of 100 data points? How would your estimated gamma hat vary?

And so you might be interested, for example, in confidence intervals, like a 95% confident interval that says that gamma hat is between, let's say, 1 and, let's say maybe, 0.5 with probability 0.95. That'll be an example of a confidence interval around gamma hat. And such a confidence interval then gives you confidence-- a confidence interval around the coefficients, then gives you confidence intervals around the average treatment effect via this analysis.

So the second observation is what happens if the true model isn't linear, but we hadn't realized that as a modeler, and we had just assumed that, well, the linear model's probably good enough? And maybe even, the linear model gets pretty good prediction performance? Well, let's look at the extreme example of this. Let's now assume that the true data generating process, instead of being just beta X plus gamma T, we're going to add in now a new term, delta times X squared.

Now, this is the most naive extension of the original linear model that you could imagine, because I'm not even adding any interaction terms like 10 times XT. So no interaction terms involving treatment and covariate. Treatment is still-- the potential outcome is still linear in treatment. We're just adding a single nonlinear term involving one of the features.

Now, if you compute the average treatment effect via the same analysis we did before, you'll again find that our treatment effect is gamma. Let's suppose now that we hadn't known that there was that delta X squared term in there, and we hypothesized that the potential outcome was given to you by this linear model involving X and T. And I'm going to use Y hat to denote that that's going to be the function family that we're going to be fitting.

So we now fit that beta hat in gamma hat, and if you had infinite data drawn from this true generating process, which is, again, unknown, what one can show is that the gamma hat that you would estimate using any reasonable estimator, like a least squared estimator, is actually equal to gamma, the true ATE value, plus delta times this term. And notice that this term does not depend on beta or gamma. What this means is, depending on delta, your gamma hat could be made arbitrarily large or arbitrarily small.

So for example, if delta is very large, gamma hat might become positive when gamma might have been negative. And so your conclusions about the average treatment effect could be completely wrong, and this should scare you. This is the thing which makes using covariate adjustments so dangerous, which is that if you're making the wrong assumptions about the true potential outcomes, you could get very, very wrong conclusions.

So because of that, one typically wants to live in a world where you don't have to make many assumptions about the form, so that you could try to fit the data as well as possible. So here, you see that there is this nonlinear term. Well, obviously, if you had used some nonlinear modeling algorithm, like a neural network or maybe a random forest, then it would have the potential to fix that nonlinear function, and then maybe we wouldn't get caught in this same trap. And there are a variety of machine learning algorithms that have been applied to causal inference, everything from random forests and Bayesian additive regression trees to algorithms like Gaussian processes and deep neural networks. I'll just briefly highlight the last two.

So Gaussian processes are very often used to model continuous valued potential outcomes, and there are a couple of ways in which they can be done. So for example, one class of models might treat Y1 and Y0 as two separate Gaussian processes and fit those to the data. A different approach, shown on the right here, would be to treat T as an additional covariate, so now you have X and T as your features and fit a Gaussian process for that joint model.

When it comes to neural networks, neural networks had been used in causal inference going back about 20, 30 years, but really started catching on a few years ago with a paper that I wrote in my group as being one of the earliest papers from this recent generation of using neural networks for causal inference. And one of the things that we found to work very effectively is to use a joint model for predicting the causal effect, so we're going to be learning a model that takes-- an F that takes, as input, X and T and has to predict Y. And the advantage of that is that it's going to allow us to share parameters across your T equals 1 and T equals 0 samples.

But rather than feeding in X and T in your first layer of your neural network, we're only going to feed in X in the initial layer of the neural network, and we're going to learn a shared representation, which is going to be used for both predicting T equals 0 and T equals 1. And then for predicting when T is equal to 0, we use a different head from predicting T equals 1. So F0 is a function that concatenates these shared layers with several new layers used to predict for when T is equal to 0 and same analogously for 1. And we found that architecture worked substantially better than the naive architectures when doing causal inference on several different benchmark data sets.

Now, the last thing I want to talk about for covariate adjustment, before I move on to a new set of techniques, is a method called matching, that is intuitively very pleasing. It's a very-- would seem to be a really natural approach to do causal inference, and at first glance, may look like it has nothing to do with covariate adjustment technique. What I'll do now is I'm going to first introduce you to the matching technique, and then I will show you that it actually is precisely identical to covariate adjustment with a particular assumption of what the functional family for F is. So not Gaussian processes, not deep neural networks, but it'll be something else.

So before I get into that, what is matching as a technique for causal inference? Well, the key idea of matching is to use each individual's twin to try to get some intuition about what their potential outcome might have been? So I created these slides a few years ago when President Obama was in office, and you might imagine this is the actual President Obama who did go to law school. And you might imagine who might have been that other president? What President Obama have been like had he not gone to law school, but let's say, gone to business school?

So if you can now imagine trying to find, in your data set, someone else who looks just like Barack Obama, but who, instead of going to law school, went to business school, and then you would then ask the following question. For example, would this individual have gone on to become president had he gone to law school versus had he gone to business school? If you find someone else who's just like Barack Obama who went to business school, look to see did that person become president eventually, that would in essence give you that counterfactual. Obviously, this is a contrived example because you would never get the sample size to see that.

So that's the general idea, and now, I'll show it to you in a picture. So here now, we have to covariates or features-- a patient's age and their Charleson comorbidity index. This is some measure of how many-- what types of conditions or comorbidities the patient might have. Do they have diabetes, do they have hypertension, and so on?

And notably, what I'm not showing you here is the outcome Y. All I'm showing you are the original data points and what treatment did they receive. So blue are the individuals who received the control treatment, or T equals 0, and red are the individuals who received treatment 1.

So you can imagine trying to find nearest neighbors. For example, the nearest neighbor to this data point over here is this blue point over here, and so if you wanted to know, well, what we observed, some Y1, for this individual, we observed some Y0 for this individual. And if you wanted to know, well, what would have happened to this individual if they had received treatment 0 instead of treatment 1, well, you could just look at what happened to this blue point and say, that's what would have happened to this red point, because they're very close to each other.

Any questions about what matching would do before I define it formally? Here, I'll-- yeah, good, one question. What happens if the nearest neighbor is extremely far away? That's a great question.

So you can imagine that you have one red data point over here and no blue data points nearby. The matching approach wouldn't work very well. So this data point, the nearest neighbor, is this blue point over here, which intuitively, is very far from this red point. And so if we were to estimate this red point's counterfactual using that blue point, we're likely to get a very bad estimate, and in fact, that is going to be one of the challenges of matching based approaches. It's going to work really well in a high dimensional setting where you can imagine-- sorry, in a large-- it's going to work very well in a large sample setting, where you can hope that you're likely to observe a counterfactual for every individual. And it won't work well you have very limited data, and of course, all this is going to be subject to the assumption of common support.

So one question's about how does that translate into high dimensions? The short answer-- not very well. We'll get back to that in a moment.

Can a single data point appear in multiple matchings? Yes, and I will define, in just a moment, how and why. It won't be a strict matching.

Are we trying to find a counterfactual for each treated observation, or one for each control observation? I'll answer that in just a second. And finally, is it common for medical data sets to find such matching pairs? I'm going to reinterpret that question as saying, is this technique used often in medicine? And the answer is, yes, it's used all the time in clinical research despite the fact that bio statisticians, for quite a few years now, have been trying to argue that folks should not use this technique for reasons that you see shortly.

So it's widely used. It's very intuitive, which is why I'm teaching it. And it's going to fit into a very general framework, as you'll see in just a moment, which I'll give you the natural solution for the problems that I'm going to raise. So moving on, and then I'll return to any remaining questions.

So here, I'll define one way of doing counterfactual inference using matching, and it's going to start, of course, by assuming that we have some distance metric d between individuals. Then we're going to say, for each individual i, let's let j of i be the other individual j, obviously different from i, who is closest to i, but critically, closest but has a different treatment. So where Ti is different from Tj, and again, I'm assuming binary, so Tj is either 0 or 1.

With that definition then, we're going to define our estimate of the conditional average treatment effect for an individual is whatever their actual observed outcome was. This, I'm going to give for an individual that actually received treatment 1, so it's Y1, and the reason-- it's Yi minus the imputed counterfactual corresponding to T is equal to 0. And the way we get that computed counterfactual is by trying to find that nearest neighbor who received treatment 0 instead of treatment 1 and looking at their Y.

Analogously, if T is equal to 0, then we're going to use the observed Yi, now over here instead of over there because it corresponds to Y0. And where we need to impute Y1-- capital Y1, potential outcome Y1-- we're going to use the observed outcome from the nearest neighbor of individual i who received treatment 1 instead of 0. So this, mathematically, is what I mean by our matching based estimator, and this also should answer one of the questions which was raised, which is, do you really need to have it matching, or could a data point be matched to multiple other data points?

And indeed, here, you see the answer to that last question is yes, because you could have a setting where, for example, there are two red points here. And I can't draw blue, but I'll just use a square for what I would have drawn as blue. And then everything else very far away, and for both of these red points, this blue point is the closest neighbor. So both of the counterfactual estimates for these two points would be using the same blue point, so that's the answer to that question.

Now, I'm just going to rewrite this in a little bit more convenient form. So I'll take this formula, shown over here, and you can rewrite that as Yi minus Yji, but you have to flip the sign depending on whether Ti is equal to 1 or 0, and so that's what this term is going to do. If Ti is equal to 1, then this evaluates to 1. If Ti is equal to 0, this evaluates to minus 1. Flips the sign.

So now that we have the definition of CATE, we can now easily estimate the average treatment effect by just averaging these CATEs over all of the individuals in your data set. So this is now the definition of how to do one nearest neighbor matching. Any questions?

So one question is, do we ever use the metric d to weight how much we would, quote, unquote, "trust" the matching? That's a good question. So what Hannah's asking is, what happens if you have, for example, very many nearest neighbors, or analogously, what happens if you have some nearest neighbors that are really close, some that are really far? You might imagine trying to weight your nearest neighbors by the distance from the data point, and you could imagine even doing that-- you can even imagine coming up with an estimator, which might discount certain data points if they don't have nearest neighbors near them at all by the corresponding weighting factor.

Yes, that's a good idea. Yes, you can come up with a consistent estimator of the average treatment effect through such an idea. There are probably a few hundred papers written about it, and that's all I have to say about it. So there's lots of variants of this, and they all end up having the same theoretical justification that I'm about to give in the next slide.

So one of the advantages of matching is that you get some interpretability. So if I was to ask you, well, what's the reason why you tell me that this treatment is going to work for John? Well, someone can respond-- well, I used this technique, and I found that the nearest neighbor to John was Anna. And Anna took this other treatment from John, and this is what happened for Anna. And that's why I conjecture that, for John, the difference between Y1 and Y0 is as follows.

And so then, that can be criticized. So for example, a clinician who has some domain expert, can look at Anna, look at John, and say, oh, wait a second, these two individuals are really different from one another. Let's say the treatment, for example, had to do with something which was gender specific.

Comparing two individuals which are of different genders are obviously not going to be comparable to one other, and so then the domain expert would be able to reject that conclusion and say, nuh-uh, I don't trust any of these statistics. Go back to the drawing board. And so type of interpretability is very attractive.

The second aspect of this, which is very attractive is that it's a non-parametric method, non-parametric in the same way that neural networks or random forest are non-parametric. So this does not rely on any strong assumption about the parametric form of the potential outcomes. On the other hand, this approach is very reliant on the underlying metric. If your distance function is a poor distance function, then it's going to give poor results. And moreover, it could be very much misled by features that don't affect the outcome, which is not necessarily a property that we want.

Now, here's that final slide that makes the connection. Matching is equivalent to covariate adjustment. It's exactly the same. It's an instantiation of covariate adjustment with a particular functional family for F. So rather than assuming that your function F, that black box, is a linear function or a neural network or a random forester or a Bayesian regression tree, we're going to assume that function takes the form of a nearest neighbor classifier.

In particular, we'll say that Y hat of 1, the function for predicting the potential outcome Y hat 1, is given to you by finding the nearest neighbor of the data point X according to the data set of individuals that received treatment 1, and same thing for Y hat 0. And so that then allows us to actually prove some properties of matching. So for example, if you remember from-- I think I mentioned in Tuesday's lecture that this covariate adjustment approach, under the assumptions of overlap and under the assumptions of no hidden confounding, and that your function family for potential outcome is sufficiently rich that you can actually fit the underlying model, then you're going to get correct estimates of your conditional average treatment effect.

Now, one can show that a nearest neighbor algorithm is not, generally, a consistent algorithm. And what that means is that, if you have a small number of samples, you're going to be getting biased estimate. Your function F might, in general, be a biased estimate.

Now, we can conclude from that, that if we were to use one nearest neighbor matching for inferring average treatment effect, that in general, it could give us a biased estimate of the average treatment effect. However, in the limit of infinite data, one nearest neighbor algorithms are guaranteed to be able to fit the underlying function family. That is to say, that bias goes to 0 in the limit of a large amount of data, and thus, we can immediately draw from that literature and causal inference-- sorry, from that literature and machine learning to obtain theoretical results for matching for causal inference. And so that's all I want to say about matching and its connection to covariate adjustment. And really, the punchline is, think about matching just as another type of covariate adjustment, one which uses a nearest neighbor function family, and thus should be compared to other approaches to covariate adjustments, such as, for example, using machine learning algorithms that are designed to be interpretable.

So the last part of this lecture is going to be introducing a second approach for inferring average treatment effect that is known as the propensity score method, and this is going to be a real shift. It's going to be a different estimator from the covariate adjustment. So as I mentioned, it's going to be used for estimating average treatment effect. In problem set 4, you're going to see how you can use the same sorts of techniques I'll tell you about now for also estimating conditional average treatment effect, but that won't be obvious just from today's lecture.

So the key intuition for propensity score method is to think back to what would have happened if you had a randomized control trial. In a randomized control trial, again, you get choice over what treatment to give each individual, so you might imagine flipping a coin. If it's heads, giving them treatment 1. If it's tails, giving them treatment 0.

So given data from a randomized control trial, then there's a really simple estimator shown here for the average treatment effect. You just sum up the values of Y for the individuals that receive treatment 1, divided by n1, which is the number of individuals that received treatment 1. So this is the average outcome for all people who got treatment 1, and you just subtract from that the average outcome for all individuals who received treatment 0. And that can be easily shown to be an unbiased estimator of the average treatment effect had your data come from a randomized controlled trial.

So the key idea of a propensity score method is to turn an observational study into something that looks like a randomized control trial via re-weighting of the data points. So here's the picture I want you to have in mind. Again, here, I am not showing you outcomes. I'm just showing you the features X-- that's what the data points are-- and the treatments that were given to them, the Ts.

And the Ts, in this case, are being denoted by the color of the dots, so red is T equals 1. Blue is T equals 0. And my apologies in advance for anyone who's color blind.

So the key challenge when working with observational study is that there might be a bias in terms of who receives treatment 0 versus who receives treatment 1. If this was a randomized control trial, then you would expect to see the reds and the blues all intermixed equally with one another, but as you can see here, in this data set, there are very many more people who received-- very more young people who received treatment 0 than received treatment 1. Said differently, if you look at the distribution over X conditioned on T equals 0 in the data, it's different from the distribution over X conditioned on the people who receive treatment 1.

So what the propensity score method is going to do is it's going to recognize that there is a difference between these two distributions, and it's going to re-weight data points so that, in aggregate, it looks like, in any one region-- so for example, imagine looking at this region-- that there's roughly the same number of red and blue data points. Where if you think about blowing up this red data point-- here, I've made it very big-- you can think about it being many, many red data points of the corresponding weight. You look over here, see again roughly the same amount of red and blue mass as well. So if we can find some way to increase or decrease the weight associated with each data point such that, now, it looks like the two distributions, those who received treatment 1 and those who received treatment 0, look like they came from-- look like now they have the same weighted distribution, then we're going to be in business. So we're going to search for those weights, w, that have that property.

So to do that, we need to introduce one new concept, which is known as the propensity score. The propensity score is given to you by the probability that T equals 1 given X. Here, again, we're going to use machine learning. Whereas in covariate adjustment, we used machine learning to predict Y conditioned on X comma T-- that's what covariate adjustment did-- here, we're going to be ignoring Y altogether.

We're just going to take X's input, and we're going to be predicting T. So you can imagine using logistic regression, given your covariates, to predict which treatment any given data point came from. Here, you're using the full data set, of course, to make that prediction, so we're looking at both data points where T equals 1 and T equals 0. T is your label for this.

Then what we're going to do is given, that learned propensity score-- so we take your data set. You, first, learn the propensity score. Then we're going to re-weight the data points according to the inverse of the propensity score.

And you might ask, this looks familiar. This whole notion of re-weighting data points, this whole notion of trying to figure out which, quote, unquote, "data set" a data point came from, the data set of individuals who receive treatment 1 or the data set of individuals who receive treatment 0-- that sounds really familiar. And it's because it's exactly what you saw in lecture 10, when we talked about data set shift. In fact, this whole entire method, as you'll develop in problem set 4, is a special case of learning under data set shift.

So here, now, is the propensity score algorithm. We take our data set, which have samples of X, T, and Y where Y, of course, tells you the potential outcome corresponding to the treatment T. We're going to use any machine learning method in order to estimate this model that can give you a probability of treatment given X.

Now, critically, we need a probability for this. We're not trying to do classification. We need an actual probability, and so if you remember back to previous lectures where we spoke about calibration, about the ability to accurately predict probabilities, that is going to be really important here. And so for example, if you were to use a deep neural network in order to estimate the propensity scores, deep networks are well known to not be well calibrated. And so one would have to use one of a number of new methods that have been recently developed to make the outputs of deep learning calibrated in order to use this type of technique.

So after finishing step 1, now that you have a model that can allow you to estimate the propensity score for every data point X, we now can take those and estimate your average treatment effect with the following formula. It's 1 over n of the sum over the data points, where the data points corresponding to the treatment 1 of Yi-- that part is identical to before. But what you see now is we're going to divide it by the propensity score, and so this denominator, that's the new piece here. That's the inverse of the propensity score is precisely the weighting that we were referring to earlier, and the same thing happens over here for Ti equals 0.

Now, let's try to get some intuition about this formula, and I like trying to get intuition by looking at a special case. So the simplest special case that we might be familiar with is that of a randomized control trial, where because you're flipping a coin, and each data point either gets treatment 0 or treatment 1, then the propensity score is precisely, deterministically equal to 5. So let's take this now. No machine learning done here. Let's just plug it in to see if we get back the formula that I showed you earlier for the estimate of the average treatment effect in a randomized control trial.

So we plug that in over there. This now becomes 0.5, and plug that in over here.

This also becomes 0.5.

And then what we're going to do is we're just going to take that 0.5. We're going to bring that out, and this is going to become a 2 over here, and same, a 2 over here. And you get to the following formula, which is-- if you were to compare to the formula from a few slides ago, it's almost identical, except that a few slides ago over here, I had 1 over n1, and over here, I had 1 over n0.

Now, these two are two different estimators for the same thing, and the reason why you can say they're the same thing is that, in a randomized control trial, the number of individuals that receive treatment 1 is, on average, n over 2. Similarly, the number of individuals receiving treatment 0 are, on average, n over 2. So if you were to-- that n over 2 cancels out with this 2 over n is what gets you a correct estimator. So this is a slightly different estimator, but nearly identical to the one that I showed you earlier, and by this argument, is a consistent estimator of the average treatment effect in a randomized control trial.

So any questions before I try to derive this formula for you? So one student asks, so the propensity score is the, quote, unquote, "bias" of how likely people are assigned to T equals 1 or T equals 0? Yes, that's exactly right.

So if you were to imagine taking an individual where this probability for that individual is, let's say, very close to 1, it means that there are very few other people in the data set who receive treatment 1. They're a red data point in a sea of blue data points. And by dividing by that, we're going to be trying to remove that bias, and that's exactly right.

Thank you for that question. Are there other questions? I really appreciate the questions via the chat window, so thank you.

So let's now try to derive this formula. Recall the definition of average treatment effect, and for those who are paying very close attention, you might notice that I removed the expectation over Y1. And for this derivation that I'm going to give you, I'm going to suppose-- I'm going to assume that a potential outcomes are all deterministic because it makes the math easier, but is without loss of generality.

So the average treatment effect is the expectation, with respect to all individuals, of the potential outcome Y1 minus the expectation with respect to all individuals of the potential outcome Y0. So this term over here is going to be our estimate of that, and this term over here is going to be our estimate of this expectation. So naively, if you were to just take the observed data, it would allow you to compute-- if you, for example, just averaged the values of Y for the individual who received treatment 1, that would give you this expectation that I'm showing on the bottom here. I want you to compare that to the one that's actually needed in the average treatment effect.

Whereas over here, it's an expectation with respect to individuals that received treatment 1, up here, this was an expectation with respect to all individuals. But the thing inside the expectation is exactly identical, and that's the key point that we're going to work with, which is that we want an expectation with respect to a different distribution than the one that we actually have. And again, this should ring bells, because this sounds very, very familiar to the data set shift story that we talked about a few lectures ago.

So I'm going to show you how to derive an estimator for just this first term, and the second term is obviously going to be identical. So let's start out with the following. We know that p of X given T times p of T is equal to p of X times p of T given X.

So what I've just done here is use two different formulas for a joint distribution, and then I've divided by p of T given X in order to get the formula that I showed you a second ago. I'm not going to attempt to erase that. I'll leave it up there.

So the next thing we're going to do is we're going to say, if we were to compute an expectation with respect to p of X given T equals 1, and if we were to now take the value that we observe, Y1, which we can get observations for all the individuals who received treatment 1, and if we were to re-weight this observation by this ratio, where remember, this ratio showed up in the previous bullet point, then what I'm going to show you in just a moment is that this is equal to the quantity that we actually wanted. Well, why is that? Well, if you expand this expectation, this expectation is an integral with respect to p of X conditioned on T equals 1 times the thing inside the brackets, and because we know that p of-- because we know from up here that p of X conditioned on T equals 1 times p of T equals 1 divided by p of T equals 1 conditioned on X is equal to p of X, this whole thing is just going to be equal to an integral of p of X times Y1, which is precisely the definition of expectation that we want. So this was a very simple derivation to show you that the re-weighting gets you what you need.

Now, we can estimate this expectation empirically as follows, the estimate that we're going to now sum over all data points that received treatment 1. We're going to take an average, so we're dividing by the number of data points that received treatment 1. For p of T equals 1, we're just going to use the empirical estimate of how many individuals received treatment 1 in the data set divided by the total number of individuals in the data set. That's n1 divided by n. And for the denominator, p of T equals 1 conditioned on X, we just plug in, now, the propensity score, which we had previously estimated.

And we're done. And so that, now, is our estimate for the first term in the average treatment effect, and you can do that now loosely for Ti equals 0. And I've shown you the full proof of why this is an unbiased estimator for average treatment effect.

So I'm going to be concluding now, in the next few minutes. First, I just wanted to comment on what we just saw. So we saw a different way to estimate the average treatment effect, which only required estimating the propensity score.

In particular, we never had to use a model to predict Y in this approach for estimating the average treatment effect, and that's a good thing and a bad thing. It's a good thing because, if you had errors in estimating your model y, as I showed you in the very beginning of today's lecture, that could have a very big impact on your estimate of the average treatment effect. And so that doesn't show up here.

On the other hand, this has its own disadvantages. So for example, the propensity score is going to be really, really affected by lack of overlap, because when you have lack of overlap, it means there's some data points where the propensity score is very close to 0 or very close to 1. And that really leads to very large variance in your estimators. And a very common trick which is used to try to address that concern is known as clipping, where you simply clip the propensity scores so that they're always bounding away from 0 and 1. But that's really just a heuristic, and it can, of course, then lead to biased estimates of the average treatment effect.

So there's a whole family of causal inference algorithms that attempt to use ideas from both covariate adjustment and inverse propensity weighting. For example, there's a method called doubly robust estimators, and we'll try to provide a citation for those estimators in the Scribe notes. And these doubly robust estimators are a different family of estimators that actually bring in both of these techniques together, and they have a really nice property, which is that if either one of them fail, you still get valid estimates of average treatment effect.

I'm going to skip this and just jump to the summary now, which is that we've presented two different approaches for causal inference from observational data-- covariate adjustment and propensity score based methods. And both of these, I need to stress, are only going to give you valid results under the assumptions we outlined in the previous lecture-- for example, that your causal graph is correct; critically, that there's no unobserved confounding; and second, that you have overlap between your two treatment classes. And third, if you're using a non-parametric regression approach, overlap is extremely important, because without overlap, your model's undefined in regions of space. And thus, as a result, you have no way of verifying if your extrapolations are correct, and so one has to use trust in the model, which is not something we really like.

And in propensity score methods, overlap is very important because if you don't have that, you get inverse propensity scores that are either-- which are infinite and lead to extremely high variance estimators. So in the end of this slide, which are already posted online, I include some references that I strongly encourage folks to follow up on. First references to two recent workshops that have been held in the machine learning community so that you can get a sense of what the latest and greatest in terms of research in causal inference are, two different books on causal inference that you can download for free from MIT, and finally, some papers that I think are really interesting, particularly of interest, potentially, to course projects.

So we are at time now. I will hang around for a few minutes after lecture, as I would normally. But I'm going to stop the recording of the lecture.