Module 3 Video 2: Causal Mistakes in Data

[00:00:00] Second mistake. It's really common in data journalism. But getting a lot less common but important to learn about from an ethics and equity point of view is the causal fallacy. When is data causal and when is data predictive? In the machine learning age on the at the A.I. age, it's really important to remember that almost all of the tools that we have are predictive tools because that's mostly what businesses and corporations want to know.

[00:00:34] What is the next movie that you're likely to click on in the Netflix algorithm? The Netflix algorithm doesn't care why you're going to watch your next movie. They just want to know how to predict which one. And the way that you build an analysis to predict what you're going to do. Is really different than the way you build analysis that's going to say why you do what you're going to do. So there's a couple of possible relationships in terms of variables. So let's say X is related to Y. There's the confounder model, so there can be one thing. Where there's this other variables, Z, which is causing both X and Y. And that's why X is related to Y. So that is our smoking example. Money is related to both purchasing cigarettes and having a longer life expectancy. So the more cigarettes. It's not actually causing, the more life expectancy. The more money Z is causing both of those.

[00:01:46] Another model that's important for you to understand is called the mediation model. And that's where X is causing Z and Z is causing why. So, for example, if we give somebody more money and they suddenly have a higher life expectancy, does X (having more money) cause Y (having a life expectancy that's higher)? No, X does not cause Y, X causes Z, which causes Y. So having more money causes, for example, eating better food, which in turn causes Y, which is having a longer life expectancy. So those are two very different ways to explain the possible relationship between X and Y.

[00:02:40] The confounder model and the mediator model. And if you tell a data story without thinking about whether you're talking about a confounder model or a mediator or model, you are telling very, very different stories. So you need to think about are you talking about a causal model or a predictive model? And if you're talking about a causal model. Are you doing a confounder model or a mediator model?

[00:03:13] Now, these are very advanced statistical concepts, so I'm not too worried about you retaining and learning. OK, what exactly how do I build a confounder model? How do I build a media model? As a data journalist basically, you need to be able to ask your data or your data source or the source of your research. Is this research or data causal or predictive? And is it a confounder model or a mediator model?

[00:03:44] You need to know how to ask the questions. You don't need to know how to do the statistical analysis. We're just gonna look at an example again from a story about cash transfer programs. So right now, especially in the age of Coronavirus, cash transfers getting a lot of attention. Should we just give people money? Is a basic universal income, actually what's going to help society get past some of the challenges that it's experiencing? You might want to write a story about this, and we don't know yet how that's gonna go in the whole world. So you're probably gonna have to write a story that focuses on a couple of examples that have been implemented.

[00:04:25] One of these is a cash transfer program for young mothers that was used to help prevent low birth weight babies. So did this work? Yes, and no. When writing a data story, there are the relationship between the cash transfer and the weight of the babies. And there's also this other variable, Z, which is here at the top of our picture in our slide. And that is, as we mentioned before, the food or nutrition that the mother eats, whether or not you're going to control for that and deciding whether or not the cash transfer works really matters. So here is a chart of the results of whether or not the mothers had low birth weight babies, including Z in the model. So this is a mediator model. So we have controlled for changes in what, the mother age. You can see from time 1 to times 3, there's very little change. Time 2 there's a bit of an up swing, but by time 3, that's gone back down. So direct cash transfers did not really affect the probability of a mother having a low birth weight baby. If we control for Z what the mother ate. So that's if we use a mediator model.

[00:05:54] This is without a mediator model. So this is a confounder model and we can see that the results are very, very different. The probability of the mother having a low birth weight baby goes down quite a bit. And that means that we are likely to read into a story that says the project really worked. Maybe cash transfers are a great idea and maybe they're the solution that we're looking for. And again, when we're embedding equity and ethics into the model, into our data journalism stories, one of these answers is not correct. And one of these answers is not incorrect. They just held two different stories. It depends on what question you're trying to answer. Do you want to know whether the direct cash transfer had an effect and you don't care how? So you just want a predictive answer. Does cash transfer predict a decrease in low birth weight babies? Or do you want to know an answer to the question, does the cash transfer change what a mother eats? And does that change her probability of low birth weight, baby? The answer to that is probably not.

[00:07:12] Two correct answers to two very different questions. And again, as a data journalist, I'm not expecting that you're going to know how to do all of these statistics and build these models. But you do need to know how to ask the questions to the people who are providing you with the data or with the results, because otherwise it's very easy to make the mistake of a causal fallacy using predictive data in a causal way.