**D2-S1 Audio**

Aug 25, 2025, 9:50 AM

**D2-S1 Audio**

(0:00) So how many of you have heard this term before, (0:04) correlation is not causation? (0:07) When, what was the context? (0:09) Today's class is going to be a little bit lecture (0:11), next I cover the basics about the concept (0:15) and then we'll have a puzzle or a game session. (0:20) What do you think about this? (0:21) Is this true? (0:23) No, why not? (0:24) But every time people have, there's data to support this. (0:28) You're making this claim based on data (0:31) and if you eat chicken, you get a good amount.

(0:36) Is this true? (0:37) No. (0:38) Data supports it. (0:41) Don't we rely on data? (0:42) No, no.

(0:46) Hmm? (0:47) Just because you eat such a chicken. (0:49) Data, that is true. (0:51) We rely on data these days.

(0:59) 50% of it's true. (1:01) 50% even if the data supports it? (1:04) Yeah. (1:07) Okay, good.

(1:09) What about this? (1:10) So this graph basically shows that countries (1:14) who consumes more chocolate are more likely (1:17) to have more number of Nobel laureates in their, (1:22) so Switzerland has the most number of Nobel laureates. (1:26) Is this true? (1:28) Maybe. (1:29) I can share the data with you guys later on (1:32) if you don't believe the data.

(1:35) Hey, this is. (1:39) Scotland's not even on there. (1:41) So now.

(1:43) How much chocolate can they eat in Scotland? (1:45) Probably none. (1:47) Not much. (1:49) So where is Scotland? (1:52) Yeah, so the highest is Switzerland, (1:55) then Germany, or maybe UK, Ireland, Norway.

(2:03) Okay, so you don't guys, you don't believe this. (2:07) What do you think about this? (2:10) So another, now I have a graph. (2:12) I have a data to support it, (2:15) and it says number of people who drown (2:18) by falling into a swimming pool (2:19) is correlated with number of films Nicolas Cage appears in.

(2:25) So this basically means, (2:28) so more films Nicolas Cage appears in, (2:31) there are more people drowning (2:33) or falling in the swimming pool. (2:35) Should we just ban him? (2:37) Yeah. (2:39) Right, so that people can be saved? (2:45) Okay.

(2:46) This is an interesting one. (2:48) It says, quit going to the lecture, (2:51) only use the outside resources, (2:53) and that would lead your grades to go up. (2:58) No? (2:59) Yes? (3:01) Okay, so if you quit right now this class, (3:05) and then I quiz you later, your grades are gonna go up.

(3:09) You'll get the highest grade (3:10) of all the people sitting here. (3:15) What? (3:15) Like time. (3:17) Time? (3:18) Okay, how does time factors in here? (3:21) Because if you could go into the lecture, (3:23) obviously you're doing something else in your time.

(3:27) But if it's just today, (3:28) then you really just don't grow or learn. (3:32) Okay, so this is not true. (3:36) Okay.

(3:37) You guys are very smart in studios. (3:40) You pay attention in class, you come to class. (3:42) Very good.

(3:44) What about this one? (3:48) So the UFO sighting in South Carolina (3:51) correlates with number of people (3:53) successfully climbing the Mount Everest. (3:56) No? (3:58) Come on, I show you guys data. (4:00) I show you guys graph.

(4:02) I support it with data. (4:04) You don't believe me? (4:06) That's not the cause of it. (4:08) Oh, what could be the cause of it? (4:16) Huh? (4:17) Time and place? (4:19) Okay, how's trains? (4:22) How does time and place relate to this? (4:29) Just a coincidence? (4:34) You're onto something.

(4:39) When do you think people typically go to climb Mount Everest? (4:43) Winter time? (4:45) Summer, right? (4:47) Summertime. (4:48) When are people more likely to be outside in South Carolina? (4:51) Winter? (4:53) Summer, right? (4:55) We're more likely to say, oh, that might be a UFO. (4:59) So it's just a correlation.

(5:01) But we are missing out this information (5:03) that it's summertime. (5:05) That's the common thing (5:06) which has not been plotted on this graph. (5:09) So I'm showing you the data.

(5:11) I'm showing you guys the graph, (5:12) but I'm withholding some information (5:15) which is a very important key that you are overlooking here. (5:22) What about this? (5:25) So sale of organic food per year increases (5:29) and that correlates with the number of people (5:32) getting diagnosed with autism. (5:36) Not true? (5:38) No? (5:40) Probably is true.

(5:42) Okay. (5:44) What about this one? (5:46) Yeah. (5:47) This is definitely true, right? (5:49) Yeah.

(5:50) The more ice cream you eat, (5:52) there are gonna be more shark attacks. (5:54) Definitely, there is a causation. (5:57) So the summer is the cause.

(5:58) That is causing the increase in the ice cream sale. (6:02) Summer is causing maybe sharks to come out more. (6:05) And that's why we say that, oh, there is a correlation.

(6:07) And people tend to miss out the correlation as causation. (6:11) It's not a cause, you are right. (6:13) You're absolutely right.

(6:15) They're correlated, but there's no causation (6:19) between the ice cream sales going up (6:21) that is not causing the increase in shark attacks. (6:26) It is just a correlation. (6:29) If you want to understand is there a causation, (6:32) there is no causation there.

(6:33) They are correlated because there's one variable (6:36) which is the summertime heat. (6:40) So you are absolutely right. (6:41) There is no causation in this.

(6:47) Maybe it will make sense when we look at the other examples. (6:51) So this is again the same thing. (6:54) So you see with the summertime, (6:55) the ice cream sales are going up.

(6:58) Summer doesn't cause the ice cream sales to go up. (7:01) So whatever data I have shown you guys, (7:04) why does it show those type of trends? (7:09) Like why do people plot those two different variables (7:12) that we have seen in future in the earlier slides (7:15) and we saw some sort of trend? (7:17) So this habit or this trend is known as data drenching. (7:22) When we don't understand what the variables in our data mean, (7:26) so imagine we have hundreds of variables.

(7:28) Rather than understanding what are the meaning (7:31) of those variables, we take those data (7:33) and then put it to some machine learning, some classifier (7:37) and try to come up with some stats. (7:41) And at times, some weird combination (7:44) of those variables would make sense, (7:46) just in the case of UFO sighting (7:49) and people going to Mount Everest. (7:51) It just made sense.

(7:54) So this practice is called data drenching (7:57) where we have thousands of variables in our data set. (8:01) We just rather than understanding our data set altogether, (8:04) we just start to compare every variable (8:07) with every other variable just randomly. (8:11) And so what is the best way to go instead of doing this (8:15) if you don't understand the data? (8:17) The first thing is to come up with a hypothesis.

(8:21) Why are we using this data? (8:23) What do we wanna achieve with this data? (8:25) So have a hypothesis and try to test our hypothesis (8:28) with whatever data set that we have been given. (8:35) And why does these things work? (8:37) Because typically there is lack of causal connections (8:40) and people use correlations as a mean (8:43) to connect these things and assume (8:46) that there is a causation in a given trend set. (8:51) I'm going to skip these things for the sake of time.

(8:55) Let's just dig into correlation. (8:57) The correlation is a relation (8:59) which exists between any phenomena, (9:01) anything or any variable. (9:03) It's a mathematical or a statistical variable (9:06) and it occurs together.

(9:07) So if two variables tends to increase together, (9:11) we say they are correlated. (9:13) Increase together, decrease together, (9:15) we say they are correlated. (9:17) And there are three different types of correlations.

(9:20) There's positive correlation, negative correlation (9:24) and zero correlation. (9:25) Any guesses what would be a positive correlation? (9:32) Positive means going up, yeah. (9:34) So positive basically means if two variables (9:37) are going up together in the same trend.

(9:39) So if you see a linear line in this direction, (9:42) that's a positive correlation. (9:44) Negative, so this is what a positive correlation mean. (9:47) So the variable associated with the small amount, (9:50) it changes, basically these variables (9:54) are moving together at the same time.

(9:58) Negative correlation basically mean (10:00) if one is increasing, the other is decreasing. (10:03) So let's take an example of the ice cream sale (10:06) and the shark attack. (10:07) The negative correlation would be (10:09) the more number of ice creams you eat, (10:11) there are less number of shark attacks.

(10:13) The positive correlation would be (10:16) the more number of ice creams you eat, (10:18) there are more shark attacks. (10:20) The third is zero correlation, (10:22) which basically means there is no association (10:25) between any of the variables that you are dealing with. (10:28) And correlation, the range for correlation (10:30) is minus one to one, (10:33) because it covers both the positive zero (10:36) and the negative range.

(10:38) Now let's talk about causality. (10:41) So causality, as we understand, (10:43) it's a relation between two given variables (10:45) where you can establish relations such as A causes B. (10:50) And we see causality in our day-to-day lifestyle. (10:54) And this is a basic term which you guys have (10:56) clearly understood that correlation is not causation.

(11:00) So if we say that younger drivers (11:03) have high probability of being in accident, (11:06) does that mean that age is a cause of accident? (11:10) No, is it? (11:12) We should stop giving younger kids cars. (11:16) Only people above 50 should be able to drive. (11:19) Anybody less than 50 should not drive.

(11:22) No? (11:25) Okay. (11:25) Okay. (11:29) So there are two things in causation, cause and effect.

(11:34) So something or someone that produces an effect, (11:37) result or a condition is known as the cause. (11:41) The effect is a change that results (11:44) when something is done or happens to it. (11:48) So can you think of anything about any causation (11:52) in day-to-day life? (11:54) So human beings are inherently have this causal mentality.

(12:01) Okay, how about this? (12:03) If I spend five minutes on this slide, (12:06) I stare five minutes on this slide, (12:09) that will cause me to not have the awesome puzzles (12:14) that I have prepared for you guys. (12:15) You won't be able to do those puzzles. (12:20) It is causal, right? (12:22) Very simple.

(12:26) Anybody wanna give me any cause, (12:30) causal thought that they had this morning? (12:33) It's counterfactuals. (12:35) So counterfactuals are hypothetical scenarios. (12:39) So you might be thinking, (12:41) oh, what if I got up and the alarm rang the first time? (12:45) I would have gotten ready at the right time.

(12:48) Those are counterfactuals, which are hypothetical. (12:51) It didn't happen, but you are just assuming (12:54) like, oh, what if it happened? (12:55) You must be paranoid to death. (12:57) What if you did not take so long to get up? (13:00) I would be on time.

(13:03) So another term with causality (13:05) is those counterfactual analysis or hypothetical scenarios. (13:09) So whenever you wanna look at correlations, (13:12) which the graphs that we saw, (13:14) you can think about what if the ice cream sale, (13:16) they actually did not go up, (13:18) would there still be shark attacks? (13:21) That's a counterfact, yes, right? (13:23) So the ice cream sale is not causing the shark attacks (13:28) as soon as you start thinking about the counterfactual (13:31) in this scenario, right? (13:39) So just because two events or properties are correlated, (13:43) it does not mean one causes the other. (13:45) So going to the hospital is positively correlated (13:48) with dying, but it does not mean (13:51) that just because you go to the hospital (13:52) that will cause you to die.

(13:54) That's a strong claim, right? (13:56) So there's no causation in there. (13:58) The more firefighters at the fire (14:00) is positively correlated with the amount of damage (14:03) done to the building, but that does not mean (14:06) that the firefighters are causing the damage. (14:12) I'm gonna skip these.

(14:14) Okay, fun part. (14:17) I'll give chance to each group (14:20) or I have a couple of puzzles maybe. (14:28) I think, okay, each group can get one puzzle at least.

(14:31) So group one, what was the group one? (14:34) Which was, who was group one? (14:37) Okay, so this question is for you. (14:41) Discuss with your group whether or not (14:43) you think the following correlations (14:45) are also causal relations. (14:47) So there's a positive correlation between age and income.

(14:51) Do you think this is causation? (14:54) Third, second option that you have is (14:57) there's a positive correlation between the house size (15:00) and the value of the house. (15:01) So the bigger the house, the more the value of the house. (15:06) There is a negative correlation between distance you drive (15:09) and the amount of gas in your tank.

(15:14) So which among these do you think are causal? (15:20) So you have these correlations. (15:23) Are these correlations actually causal? (15:27) And which one of these? (15:31) So the first one says there's a positive correlation (15:34) between age and income, (15:36) which basically means the older you get, (15:39) the more money you're gonna earn. (15:40) That's the positive correlation.

(15:42) That's what correlation is telling us. (15:45) Does that mean that age is causing us to get more money? (15:50) No, so not causal right there. (15:53) The second one is there's a positive correlation (15:56) between the house size and the value of the house.

(15:59) So the bigger the house, (16:01) the more money you have to pay for it. (16:06) That's the correlation. (16:08) Is that a causation? (16:10) Does the size of the house causes (16:14) the sale price of the house to go up? (16:18) No, why not? (16:21) You guys are right.

(16:22) That's absolutely correct. (16:24) The third one is causal. (16:26) But when you say causal, (16:27) you should also mention that (16:29) given these other things are not true.

(16:32) So given that there's no leakage in the gas tank, (16:35) everything is perfect, the car is running fine, (16:38) then yes, this would be the more distance you drive (16:41) will cause you to have, (16:43) I mean, the amount of gas in your tank will go down. (16:46) So that's a negative correlation. (16:48) And at the same time, it's a causation.

(16:53) So there's this term called common causal variable. (16:56) So it occurs when two events or measurements are correlated. (17:01) And the assumption is made that one causes the other.

(17:05) So there is some hidden variable (17:07) which we have not seen in the data, (17:09) which is leading us to have this correlation. (17:13) So basically our belief says X causes Y, (17:16) but there might be some variable Z. (17:18) So if you remember the case of ice cream sale (17:20) and shark attacks, the common variable is the summer. (17:24) So summer is causing both the increase in the ice cream sales (17:27) and the increase in the shark attacks.

(17:30) Some of the examples could be that Bob notices (17:33) every time he has a temperature, he does not feel well. (17:37) And the reason he gives is (17:39) because he has a high body temperature (17:41) that is actually causing him to not feel well. (17:45) Then Bob jumps into the ice bath, (17:48) concluding that if he lowers his body temperature, (17:52) he's gonna feel better.

(17:54) But he's missing out, he may be sick because of a virus (17:58) or maybe have some other type of a disease or a condition. (18:01) So the common cause here is the virus (18:04) and not just not relating the body temperature (18:08) with not feeling well. (18:11) Puzzle three, I had puzzle two, (18:14) I think it's been somewhere, but okay.

(18:16) We'll assume this is puzzle two, (18:18) but we'll go with group two. (18:21) Ready? (18:22) Okay. (18:24) So the puzzle three says, can't you see the flaw? (18:29) So a study from University of Pennsylvania, (18:31) which was published in May 1999, (18:35) is in the issue of Nature Magazine.

(18:37) So Nature Magazine on a journal, it's a scientific journal. (18:40) So if you ever wanna go online, look for nature, (18:44) you're gonna get a lot of scientific articles (18:46) and papers in that journal or magazine. (18:50) So in that magazine, this study was published, (18:53) which said that babies younger than two years old (18:57) who slept with the light on, (18:59) were at increased risk of developing myopia, (19:03) which is nearsightedness or shortsightedness (19:06) later in the childhood.

(19:08) So the kids who are younger than two years of old (19:11) who slept with light on, (19:13) were more likely to develop shortsightedness (19:16) when they grew up. (19:19) However, so in the current study of about 1200 children (19:25) conducted by Ohio State University researchers, (19:28) they found there is no relation (19:30) between the nighttime lighting (19:32) and the development of the shortsightedness (19:34) or nearsightedness. (19:35) It does not matter if you sleep in a dark room, (19:38) light room, or you have a nightlight on.

(19:42) So what the researchers did find, however, (19:45) was a strong link between nearsighted parents (19:48) and nearsighted children. (19:52) Is this true? (19:53) There's something wrong with this. (19:59) So the third point.

(20:08) So this is basically talking about (20:10) the data drenching part that we saw, (20:13) that we heard that whatever data we had, (20:15) we just throw it into the system (20:17) and asking the system to make analysis for us. (20:23) So we need to start. (20:25) Okay.

(20:26) So the question is, (20:28) it says you have to figure out the flaw in this study. (20:34) Okay. (20:35) So the first assumption, (20:36) the first time they did study in 99, (20:39) they claimed or they got this result (20:42) that kids who slept with nightlight on (20:44) were more likely to develop shortsightedness (20:47) when they grew up.

(20:48) That's the first analysis of the result they got. (20:52) Then recently they found out there is no such relation (20:56) between the nightlight on (20:58) or it doesn't matter if the kid is sleeping (21:01) in the dark room, light room. (21:04) There's no relation between the light (21:06) and the kid developing shortsightedness.

(21:10) You have a question? (21:15) It's the same, I think, probably. (21:23) And then the researchers said (21:25) they had one analysis from both of these studies (21:29) that there was a strong link (21:31) between the nearsighted parents. (21:33) So parents who are nearsighted or shortsighted (21:36) are more likely to have kids who are shortsighted.

(21:40) They saw that link. (21:43) Is that a true link? (21:48) Or there's something missing in this third? (21:51) Here's the answer. (21:53) So the researchers noticed (21:55) that the nearsighted parents (21:57) were more likely to use nightlights in their kid's room.

(22:03) You didn't say that. (22:04) You didn't say that. (22:08) And then the researchers also said (22:10) that they think it may be due to parents (22:12) or own poor sightedness, (22:15) eyesightedness that they're keeping the lights on.

(22:18) And they also said that they found genetics. (22:21) So you guys were onto something (22:23) because he said there was neuropathological waste. (22:27) They might play a role in causing myopia.

(22:31) Did you guys do get a point? (22:33) Yeah, this was a tough one. (22:36) This is really tough one. (22:41) Okay, so the next thing, (22:43) which is the oversimplification on multiple causes.

(22:46) So that's for group. (22:48) It's gonna come for group three. (22:50) So maybe group three needs to pay more attention.

(22:53) So this policy occurs more often in media. (22:58) So there'll be multiple causes of something (23:00) that you're watching on TikTok or Instagram reel (23:03) or some posts that you're scrolling. (23:05) And you only get to see or read or hear (23:08) one side of the story.

(23:10) But however, there might be multiple causes (23:12) or might be multiple events (23:14) that might have led to whatever news article (23:17) or a news clipping or a video reel (23:19) or a post that you're reading. (23:21) So sometimes you have heard statements like, (23:24) you will do better at work or school (23:26) if you have a good breakfast. (23:28) So I'm assuming that you guys had a good breakfast.

(23:31) I love biscuit and you guys love biscuit. (23:33) Even if you did not, you guys love biscuits. (23:35) So you're gonna excel in this puzzles that I'm giving you.

(23:39) So this may be true in general, (23:42) this hypothesis that if you do good, (23:45) if you do better at work or school, (23:47) you had a good breakfast, that's the reason. (23:50) But then might be multiple causes (23:52) for you doing good in school (23:54) that you guys are better prepared, (23:56) you are not sleeping, you can see the screen. (24:00) So you are well motivated, you're all healthy.

(24:05) So the normal belief, (24:07) if you just wanna put it in a hypothesis, (24:11) so belief could be A causes Z. (24:13) But what is happening in underlying is (24:15) that A is causing Z, then B is also causing Z, (24:19) some variable C is causing Z. (24:21) So there are multiple variables other than A (24:24) which are causing Z. (24:26) And sometimes we just oversimplify (24:28) and we just make our claim that A is the one (24:32) which is causing Z. (24:35) Okay, group three, raise your hand. (24:39) Where's group three? (24:41) Okay, all bronze. (24:43) So can you think of any more examples (24:45) of oversimplified cause? (24:49) And what other events may have many reasons for occurring? (24:55) So one example could be that the shark attacks (25:01) and the ice cream sale, let's say.

(25:03) So there are more people who are entering the water. (25:08) So there might be more sharks. (25:10) So maybe more people entering could cause it.

(25:13) Then the temperature could cause it. (25:15) Then maybe the migration pattern of sharks (25:19) coming to the beaches could cause the shark attacks. (25:22) So there are like multiple causes for the shark attacks.

(25:25) Ready? (25:27) Okay. (25:30) So these are the examples that I came up with. (25:33) So how fast you read a book causes (25:35) could be your reading abilities, (25:37) your book, length of the book (25:38) and the reading environment you're in.

(25:40) Then how long you spend time working on your homework, (25:44) the better you'll do. (25:46) So it could depend on how lengthy the homework is, (25:49) how difficult the homework is, (25:50) your level of understanding of the concepts, (25:52) your work environment and the amount of assistance (25:54) you might have received while you're doing the homework. (25:59) Okay, so this concept is for group four.

(26:03) So I hope you guys are paying attention. (26:05) So this is called bidirectional cause. (26:08) So when two events are result of a bidirectional causation, (26:12) one event causes another while other event causes the first.

(26:17) So they could be X causes Y, but it could be both ways. (26:21) So X is causing Y as well at the same time, (26:24) Y is causing X. (26:26) So some examples could be the number of lions in Kenya (26:30) affects the number of gazelles in Kenya. (26:33) So lions eat gazelles.

(26:34) But it is also true that the number of gazelles (26:37) also affect the number of lions in Kenya. (26:40) So if the lions don't have food, (26:42) they're probably gonna die with lack of food. (26:45) So this will lead to increase and decrease (26:48) of population in lion, (26:49) and will also cause the increase and decrease (26:52) in the population of gazelles.

(26:54) So this is called the predator-prey model, (26:56) which is more like a bidirectional causal relations. (27:00) You probably can request my puzzle. (27:03) Can you think of any examples of bidirectional causes? (27:09) So what are some of the advantages and disadvantages (27:12) that we learned from the correlation is non-causation? (27:16) So just doing merely correlation, it can save money, (27:21) it can save time, (27:22) you don't have to pay much attention to it.

(27:24) It can allow us to make these predictions (27:26) like food causing the increase in population, (27:29) and we can use the existing data. (27:32) But some of the disadvantages of just relying on correlation (27:36) is that we cannot establish cause and effect relationship. (27:40) They are prone to inaccurate reporting.

(27:43) As we just said, he rightly mentioned (27:46) that it's not a cause that the population grew up. (27:50) There might be confounding variables. (27:52) So confounding variables basically means (27:54) there might be some common variables (27:56) causing both the other variables, (27:59) which are cause and effect, (28:01) to go up and down at the same time.

(28:03) And in the correlation studies, (28:05) we cannot manipulate the data, (28:07) which basically means we cannot run studies (28:10) and say, what if I change this? (28:13) How would that affect my XYZ result in future? (28:18) So detection lies, (28:20) distinction lies in the nature of the relationship. (28:23) So correlation versus causation. (28:25) So correlation implies association.

(28:27) There might be some relation between two variables. (28:31) It tells us about the predictive capability (28:33) or a future analysis, (28:36) but it does not imply causation. (28:39) And causation implies direct influence.

(28:41) If we say something is a cause, (28:43) we know for sure one entity (28:45) would be affecting the other entity. (28:49) And that's it. (28:50) I hope you guys had fun.