TOM LEIGHTON: Today we're going to talk about the concept of independence. In probability, we say that an event A is independent of an event B if one of two conditions hold. First, if the probability of A given B is just the same as the probability of A or if B can't happen, namely the probability of B is 0.

In other words, A is independent of B if knowing that B happened doesn't change the probability that A is going to happen. So knowing that this event occurs doesn't influence the probability that A occurs. And there's a special case where they're independent because you know that B can't happen. If the probability of B happening is 0, then everything is independent of B.

Now, the typical example that gets used is when you flip two coins. So say we flip two fair, independent coins. And let's let B be the event that the first coin is heads and that means that the probability of B happening is 1/2, because we've assumed it's a fair coin, and we'll let A be the event that the second coin comes out heads. So we know the probability of A is 1/2 because it's fair.

And because they're independent, we can conclude that the probability of A given B is 1/2, which is the probability of A. In other words, seeing the result of the second coin doesn't tell you anything about the result of the first coin.

Now actually, when you flip two coins, it's not just always the case if they're independent. Can anybody think of an example where you can flip a pair of coins and they are dependent somehow, they're not independent? Yeah.

AUDIENCE: Well, if you have to get two heads and two tails?

TOM LEIGHTON: If you have to get two heads or two tails. Well, how would you have to get?

AUDIENCE: The probability of getting two heads should be 1/4 [INAUDIBLE].

TOM LEIGHTON: Well, then they would be independent in that case. Yeah.

AUDIENCE: If you glue the coins together.

TOM LEIGHTON: Yeah. I mean, this is a silly example, but I got two fair coins here. I could clip them together and now I flip them and odds are pretty good they're both going to be heads or both be tails. If you know what happened to the right coin, it will tell you what happened to the left coin.

Now, that's a pretty contrived example, but it is illustrative of what happens in practice. In practice, we assume independence even though there can be subtle dependencies and this could lead to trouble. In fact, we're going to give a lot of examples where it leads to trouble today and also for the rest of the course. Because we're always going to want to assume independence and when we do, we're going to get very nice results, but things aren't always independent in practice and establishing independence is a hard thing to do.

For that matter, while we're on the subject, we always talk about fair coins. You flip a coin and it's fair. You know, that's not always to either. There's actually a famous mathematician named Persi Diaconis who used to down the street at Harvard and he came and gave a talk one day at MIT in the math department and he's a probabalist. He does probability theory and is a very cool guy.

And so he flipped a coin, got a quarter from somebody in the audience and flipped it and he flip that I think 10 or 20 straight times all the way to the roof, caught it, turned it over. Every time it was heads. And he goes, now what's the probability of that happening? Well, you know, it's 1/2 to the 20th or whatever, not very likely. How could he always make it come out heads?

Well, Persi was an unusual guy and in fact, he'd spent months in the strobe lab over at Harvard practicing to make it always rotate seven times, three of them on the way up, one at the top, and then three down. And he could actually see how many rotations it had done to make sure it was seven, so it always came out heads.

Now, he is an unusual fellow. He was 1 of 10 people in the world that could do a perfect shuffle reliably on a deck of cards and that's a very hard thing to do. He said he had to practice 8 hours a day for over six months to be able to do it every time. In fact, he gave another talk at MIT where he came in and he made magic tricks, actually based on mathematics. And you would cut a deck, he would feel it like this and tell you where you cut, how many cards were in the part you picked up and then do his eight perfect shuffles, which is enough to return a normal 52-card deck back to its original order.

And then using this, he could play the game where pick any card, you stick it in, he feels where the card went, and then using mathematics, he could shuffle the deck eight times and make the card come out anywhere he wanted in the deck. So he had a lot going on upstairs too.

He had an interesting life history. He ran away from home as a young child and joined the traveling circus. And then somehow from there, he joined the faculty at Harvard. You know, there's an amazing story.

And actually your story about Persi is he was the first guy to get kicked out of casinos for card counting. He figured that out way before the MIT team and the movie *21.* Down in Puerto Rico, he used to play and then they finally figured him out and he got booted.

So back to independence, let's do another picture example. Say that my sample space looks like this and I've got two events, A and B and they look like this, so they're dis-joined. Are A and B independent? No. In fact what is the probability of A given B as I've drawn it?

AUDIENCE: 0.

TOM LEIGHTON: 0. Because if B occurs, you're outside of A. And so this does not equal the probability of A as long as it's not 0. So disjoint events don't imply that they're independent.

Now, what's the picture look like for them to be independent? What is the right picture to draw here? So I got my sample space and say I make this half the sample space be A. Well, then B to be independent, would look something-- I didn't quite draw it. I actually have it be 50-50.

So if a is 50% of S, like this half, then for A to be independent of B, A intersect B, this part, has to be 50% of B. Because the probability of A given B must equal the probability of A to be independent. So this would be a picture where they are independent.

Now, independent events are really nice to work with and in part because they have a very simple rule for computing the probability of an intersection of events and it's called the product rule for independent events. And that says that if A is independent of B, then the probability of A and B or A intersect B is just the product of their probabilities separately, the probability of A times the probability of B.

So let's prove this. And there's two cases, depending on whether or not B can happen, if the probability of B is 0 or not. So case 1 is B can't happen. The probability of B is 0. In this case, what's the probability of A and B? B can't happen. 0.

If B can't happen, then they both can't. You can't have both of them happening and that equals the probability of A times the probability of B because the probability of B is 0. So that case works.

Case 2 is the probability of B is bigger than 0. In that case, we have the probability of A and B, A intersect B, well, from the definition, is the probability of B times the probability of A given B. We did that last time. And by independence, this is just the probability of A because A is independent of B, so we're done.

In fact, many texts will define independence by this product rule. Many texts will say that A and B are independent if this is true. And it's equivalent, it turns out. We won't prove that here, but if you use this as the definition, then you can derive our definition as a result. So this is an equivalent definition of independence.

Another nice fact about independent events is that it's a symmetric relationship. It's called the symmetry of independence. That says that if they A is independent of B, then the reverse is true. B is independent of A. Now, we won't prove that. It's actually easier to see that it's true if this were the definition of independence because A intersect B is the same as B intersect A and multiplication is commutative. So it's easier to see it if we had used that definition.

So because of this we often just say A and B are independent because it doesn't matter which order you're taking them in. All right, any questions about the definition so far? All right. Let's do some examples.

Let's say I have two independent fair coins. And I'm going to have the event A be the situation when the coins match, both heads, both tails. And B is going to be the event that the first coin is heads. And I want to know, are A and B independent? Are those independent events?

Well, what's the first answer to this? I mean, A is event the coins match. B tells me what the first coin was. So the first inclination here is that these are dependent events because I know something about the first coin, so that might tell me something about the probability they match. There could be some dependence here.

Now, in fact, because it's set up, they're independent and we can check that by just doing the calculation, computing the probability of A given B. Maybe I can do that. I'll do that here. The probability of A given B is, well, the condition that they're going to match given that the first point is heads means it's the same as the second coin being heads. This is the probability the second coin is heads and that's just 1/2 because it's a fair coin and independent of the first one.

Now, the probability of A, by itself, the events the coins match, what's that? How much is that? What's the probability the coins match?

AUDIENCE: [INAUDIBLE].

TOM LEIGHTON: 1/4 plus 1/4. I've got 1/4 chance of heads, heads 1/4 chance of tails, tails, so it's 1/2, so it works out. The probability of A given B equals the probability of A. They're both 1/2. So A and B are independent events because that's just the definition even though it looked like there might have been some dependence lurking around here.

Now, this example that I just did is a little misleading. The intuition they probably are dependent actually is good intuition in this case because if I don't have fair coins, they are dependent. All right. So in particular, let's look at what happens if the probability of a heads is p and the probability of tails is 1 minus p for both coins.

So let's compute the probability of A given B. What is it in this case? Well, it's the probability the second coin is heads. What's that? p because both of them are heads with probability, p. They're independent still. The two coins are independent.

And now let's look at the probability that the coins match. Well, it's a probability of heads, heads and the probability of tails, tails. Heads, heads is p times p. Tails, tails is 1 minus p squared. So to independent, I need this to equal that or to have the probability of B be 0.

So A and B are independent if and only if-- the first case is probability B is 0, which means that p equals 0, or that has to equal this. So p would have to equal 1 minus 2p plus 2p squared, just square that out there. So let's solve this. That happens if and only if 0 equals 1 minus 3p plus 2p squared. That's true if and only if 0 equals-- I factor this-- it's 1 minus 2p times 1 minus p and that's if and only if p is 1/2 or p is 1, two roots.

So if the coins are always heads, they're independent. If they're always tails, the events are independent or if they're fair coins, these two events are independent. But anything else, they're not independent anymore. Any questions? And now you can sort of see if the coins are likely to be tails and the first one comes up heads, that should influence the probability the coins match. It should change.

Questions? All right. So there's a nice application of this to getting an edge in ultimate Frisbee. Now, when you're playing ultimate, you've got to decide who gets the Frisbee first. And sometimes you don't have a coin to flip, call heads or tails, but you do have the Frisbee.

Now, you could flip the Frisbee and call right side up or not, but the problem is the Frisbee is known not to be a fair coin. When you toss it up in the air, it's likely to wind up on, I guess, the curved edge down. So that wouldn't be fair to call heads or tails.

So the standard solution is to flip the two Frisbees at the same time or one Frisbee twice and somebody calls same or different, that the two Frisbees both come up on the same way or they come up different ways and then if you called it right, you get to start with a Frisbee. And the idea behind this is that that simulates a fair coin, that the probability that they're the same is 50-50.

What do you think. Is that a fair way to decide who starts first? Yeah.

AUDIENCE: No.

TOM LEIGHTON: No. Yeah, that's right. It's not. Now, it is in the case when the coin was fair, but we know the Frisbee is not fair. And in fact, you can see this from this probability. This is the probability of a match, which is fine at p equal 1/2, but in fact, if you analyze this equation, you find out its minimum value is at p equals 1/2 and as p starts moving away from 1/2 towards 0 or to 1, it gets bigger. And we know that for Frisbees, p is not 1/2. This means that the probability of a match is better than 50%.

So if you're ever playing ultimate, always call same because you're going to have a better than 50-50 chance of getting to start with the Frisbee. It's not a fair example. There is another example of how to make a fair coin from a biased coin to an unbiased coin in homework, ways of doing this that are fair. Because often you have biased random numbers and you want to get unbiased or maybe you got a fair coin and you want to make something that comes up heads with probability 1/3. How do you actually do that in a way that works? Any questions on that?

The next example is from the first OJ Simpson trial. How many people here know who OJ Simpson is? OK, so he's still pretty famous. Now, as you probably know then he was a famous football player. Back when I was a kid, he was a famous college player, then he was a famous pro player and then he was an actor, famous actor.

And then he was accused of murdering his wife in a gory knifing and a friend of his wife's. And ultimately, the jury found him not guilty, but pretty much everybody in the country thought he did it. He looked really guilty. And it was a big media event, one of the first big trial events on TV. And so all the proceedings were on TV and everybody watched them. We'd all go home to watch the OJ hearing. It was amazing.

Now, during the indictment proceedings, there was a huge dispute over what independence was and does it matter. The issue arose when the prosecution witness claimed that only 1 in 200 Americans had a certain blood type that matched the blood type found at the scene of the crime, which was alleged to be OJ's blood. And this was during the indictment and back then DNA tests took a long time and they weren't ready yet. And the witness presented the following facts and this was the crime lab guy, the police guy.

He said that 1 in 10 people, roughly, matched type O blood. And that 1 in 5 people matched the Rh factor positive. And that 1 in 4 people match a certain kind of marker, which I don't remember what it was. We'll just call it marker XYZ, some other factor of the blood. And then this conclusion was that this means that 1 in 200 match all three factors.

And this seems reasonable because there's 1/10 of the people have O, if 15 of them have positive Rh factor and then 1/4 of all of those have this marker, that's 1 in 200. Now, it's important because OJ's blood and the blood at the crime scene both matched all three. So the implication, of course, is that OJ is looking like the guy who did it. And the question was, well, is the 1 in 200 really true? We can sample these three in the populations and see they're true, but is 1 in 200 really true?

Now, it would be if, in fact, we verified that 1/5 of the type O people have positive and 1/4 of the O positive people have the XYZ marker. But well, we don't necessarily know that unless we go figure that out. If you assume they're independent, then it would be true. The product rule will tell us that if you assume they're independent.

So during the trial, a special math defense counsel showed up, not part of the normal defense team, but he was brought in as a mathematician and lawyer and he crosses the police guy on the stand. And he asked the police guy, the lab guy if it is known that these three factors are independent. Well, the poor police lab guy never heard the word independent before, didn't know what it meant and the defense counsel proceeded to crucify him on the stand. And then in the end, all he could say was, look, we just get these things and we multiply them. That's what we're supposed to do.

It was a little scary. The actual transcript-- you can still get it-- is a little scary. The same problem arises today with DNA testing. Only there, you've got lots of these things and you multiply them all together and you get probabilities like one in many billion probability of a match.

Now, there's probably a higher level of science going on with DNA testing, but it's even harder to really establish independence. If you assume it, fine. The math works out great. You just multiply them together. But how do you know it's really true? How do you know that maybe a lot of people that have those four markers and DNA don't happen to just have the fifth also, but it really is totally unrelated.

And to know that for sure, you got to test hundreds of millions of people, which we really haven't done yet, and not just a few guys in Detroit to be able to conclude independence of 1 in a billion probabilities.

So for us, this is a lot easier. In the classroom, we assume independence and we'll keep doing that left and right, but it doesn't mean it's true in reality. In fact, in the last week of class. We'll talk about how false assumption of independence on mortgage failures led to the subprime mortgage disaster in the recession. It was all because of some mathematics mistakes that people made.

Now, this example raises the question of, what does independence mean when you have more than two events? We defined independence when there is two events, but here there's three. And so to be careful, we got to actually define dependence among more than two events and in this case, we talk about the events as being mutually independent. So let me define that.

So if I've got events A1, A2, up to An, we say they are mutually independent if, and this is a little complicated notation, but for all i and for all sets j that are subsets of the events, but not including i, then the probability that the i-th event occurs given that all the events in the subset occurred, is the same as the probability of the i-th event occurring by itself. Or there's a special case where the chance the other events occur is 0.

In other words, a collection of events is mutually independent if any knowledge about any of the rest of the events, happening or not, does not influence the event you're looking at for each of those events. So no information about any of the other markers the blood influences the i-th marker for any i. The probabilities are unchanged.

Now, there's an equivalent definitions based and the product rule. Let me show you that version because that's easier to work with usually. This is the product rule form and it says that A1, A2, up to An are mutually independent if for any subset of the events the probability of each of those events in the subset happening, all them happening, is simply the product of their individual probabilities.

So independence means that if you want the probability of a bunch of events occurring, just multiply them out individually. And that follows for independence or it could be the definition of independence, depending on how you want to do it. So either of these are good enough for you to use as a definition or a result for independence. And so the blood guy, of course, is just multiplying them out because they're assumed to be independent, so it's OK that way.

Let's do an example. So for example, say we have three events. A1, A2, and A3 are mutually independent if, these are the things you have to check, probability A1 and A2 is just the probability of A1 times the probability of A2. Then you'd check that the probability of A1 and A3 is the product of their probabilities, A1 and A3. And you'd check the probability of A2 and A3 is the product of their probabilities.

And there's one more thing to check. What's that? All of them. The probability of all of them is the product of each of them together here. So if you want to show the three events are mutually independent, these are the four things you check. That's one way to do it, which is the case of the blood typing in the situation.

All right. Let's do an example. Well, for example, if I flip three unbiased, mutually independent coins. The probability of two of them being heads is 1/4. The probability of three being heads is 1/8 and so forth. Let's do a trickier example. This is a question that was on the final exam a few years ago and a lot of the class missed it. So now we'll do it here.

Say I flip three fair, mutually independent coins and my events are going to be A1 is the event coin 1 matches coin 2. The second event, A2, is the event that coin 2 matches coin 3. And the third event, A3, is the event that coin 3 matches coin 1.

And the question was, are these three events mutually independent? Prove your answer. Let's try to figure that out. The coins, of course, are mutually independent, but what about these events? So let's start doing it. What's the probability one of the events occurring? Well, you got to get the two coins at hand to match, so that's the probability of a heads, heads plus the probability of a tails, tails. That's 1/4 plus 1/4 equals 1/2.

Now, the probability of Ai and Aj, i and j are 1 to 3, they're different, but what is a way of characterizing that case? Say event 1 occurred and event 2 occurred, how would I characterize that? Yeah.

AUDIENCE: All the same.

TOM LEIGHTON: All of them. Yeah. All of the coins are the same because if A1 and A2 occur, I know 1 matches 2 a 2 matches 3. If A1 and A3 happen, 1 matches 2 and 1 matches 3, so they're all the same and the same for A2 and A3. If 2 matches 3 and 3 matches 1, they're all the same. So this is the same as saying all three coins are the same. It could all be heads or all be tails.

And that's an 8 plus 8, which is 1/4 and that means equals the probability of Ai times the probability of Aj, which is what I need for independence. And then they said they're done. They are independent, the three events. You like that answer? What's missing?

The last case. They didn't check the last case and we got to do that to have mutual independence. So let's look at that. The last case is probability A1 intersect A2 intersect A3. What is the probability that all three events occur?

Well, the coins all have to match, right? If all the coins match, all three events occur, right? And what's the probability all 3 coins match? 1/4, just the same as this, is 1/4. Does that equal probability of A1 times the probability of A2 times the probability of A3?

What's that? 1/8. This is 1/8. They are not equal. They are not mutually independent events. All right? Any questions about that? It might well be something like this on the final this year, a good, decent chance.

So if you start going along, looks like they're independent, but you forget to check that last case, which shows they're not mutual independent. So you've got to check for all pairs and all subsets of events for mutual independence. Any questions about that?

Now, this is actually an interesting example because in this case, all pairs were independent and when that happens, we give that a special name and it's called pairwise independence, not too surprising. And that can be useful because there's many times where you do get pairwise independence, but not mutual independence. So let me give you that definition.

So a collection of events A1 through An are said to be pairwise independent if for all i and j, where i doesn't equal j, Ai and Aj are independent. Now, as we saw in this example, in this example, it was pairwise independence because the probability of Ai and Aj equaled the probability of Ai times the probably of Aj. For any pair, it was true. But it doesn't imply mutual independence. So pairwise does not imply mutual. Mutual would imply pairwise because it's true for every subset of events.

All right. So let's go back for OJ and see what would have happened. What can you say about the probability of a blood match for a random person if you only knew that these factors were pairwise independent? Say you only knew that. You didn't know they were mutually independent, but you knew they were pairwise independent in the population. What's the best you can say about the probability a random person matches that blood profile, an upper bound on the probability? Yeah.

AUDIENCE: 1 in 50.

TOM LEIGHTON: 1 in 50. Yeah. So what you can say is 1 in 50, but nothing better. So let's see why 1 in 50 works. So let's let M1 be the event you match here, M2 be the event you match their, and M3 be the event you match that. The probability you match all three is upper bounded by the probability you match the first two because matching all three is a subset of this.

Pairwise independence means that this is true. This equals the probability of matching the first times the probability of matching the second. The probability of matching the first is 1/10, probably of matching the second is 1/5, so this is 1/50. And you picked the best two. You could have picked these two and said it was at most 1/20 or those two and said it's at most 1/40. But you were clever and said, OK, I'm going to take these two and use that as my upper bound, which is 1/50.

And it might well be that 1 in 50 people match all three. That can well be. Because maybe whenever you're O positive, you have marker XYZ. That's possible, potentially, unless we find out otherwise.

What if I tell you can't assume any independence at all? What can you say about the probability of a blood match here for a random person? Yeah.

AUDIENCE: 1/10.

TOM LEIGHTON: What is it?

AUDIENCE: 1/10.

TOM LEIGHTON: 1/10. Because if they match all three, they match this and that probability is 1/10, so it's at most 1/10. And it could be that everybody who's O is O positive and has XYZ. So unless you have more information, that's the best you can say. It might well be that's the answer. Any questions about that?

So the assumptions really matter. The more independence you assume, the better bounds and the probability you get of a match. It's a little bit unrelated to this, but there was another mathematics dispute at the OJ trial. It turned out the that OJ had been beating up Nicole on a fairly regular basis and there were police records because after he'd beat her up, she'd go in and complain to the police.

And the prosecution wanted this evidence admitted at the trial because if the guy is a wife beater, it makes you think that maybe he killed her. And the defense lawyers argued against admitting that evidence because it wasn't tied to the actual murder scene in any way and they argued it would be prejudicial to the jury because, of course, if the jury hears that OJ was beating her, they might be more likely to include to convict him for murdering her.

Now, they got the math council again to argue that the reason you shouldn't admit this is because the probability that you kill your wife, that's K, given that you batter your wife, that's B, is 1 in 2,000. I would have guessed it was higher, but the evidence did show that. And so they said, look, there's only a 1 in 2,000 chance that this evidence of wife beating is relevant and therefore, it should not be admitted because there's a pretty decent chance if the jury hears this, they're going to convict him.

That's a pretty good argument. And usually that kind of thing, you exclude it. Yeah.

AUDIENCE: Where did that number come from?

TOM LEIGHTON: They got some study and some experts to come in and say that for every 2,000 wife beaters, only one of them actually kills his wife. Now, what do you suppose the prosecution argued back? They actually argued back very effectively, because that's a tough argument to get by. Yeah.

AUDIENCE: What's the probability that you kill your wife in the first place, that could be 100 times larger than usual.

TOM LEIGHTON: Well, that's a good point. So maybe the probability of killing your wife not knowing B, I hope is pretty small, probably that's very small, but I don't know. But in any case, this thing you're going from, say it's 1 in 1 million to 1 in 2,000, 1 in 2,000 is still too small to be used as evidence that OJ did it.

AUDIENCE: Frequency he did it.

TOM LEIGHTON: Frequency, they didn't get into that because I guess he'd done it a bunch, but that's a good point. It could be there's multiple beatings is higher. Maybe that's 1 in 200 then. In fact, that may be the case because I think there's probably they say because if you do it once, you do it multiple times. So there's not much more to be gaining there.

There's a critical piece of information we've left out of our conditional probabilities here. In fact, the most glaring piece of all of evidence. What's missing here? What haven't we factored in? Yeah.

AUDIENCE: The probability of B.

TOM LEIGHTON: The probability of B, that's the battering. Battering, I don't know what it is, probably a large number. Defense would argue it's large, I guess, but it shouldn't matter that much.

AUDIENCE: The probability that he actually beat her, given that she threatened him?

TOM LEIGHTON: Well, there's that, but they have police-- well, that's true. They didn't see him doing it, but let's say that they had good evidence that he did it and defense wasn't arguing that he didn't really beat her. The key thing we're missing here is Nicole wound up dead. She was dead. And there's another stat here that the prosecution argued.

So they argued this fact. The probability the husband kills his wife, given that he batters her and she wound up dead, that somebody murder her is bigger than 1/2. So here M is somebody murdered the wife. Here, the husband beats her. Now, the conditional probability that he killed her is bigger than 1/2 and that's a whopper. Now, it's very relevant.

The probability he killed her just given that he beat her is only 1 in 2,000, but if you add the fact, which is very relevant in this case, that the wife was murdered, this is now very compelling. Now, in fact, they should have really compare this to probability he kills her given that she's dead. And so that would determine now the relevance of the battering, the wife beating. That's what they should have done, but they didn't. They got this far and they had that and the judge said, I'm letting it in. So it came in at that point.

But this would be the right comparison, I think. Because you look at the probability that you killed her given that she's dead, but now the additional information, the wife battering, how does that change the probability? And it probably changes it materially. So it's all a little gory, but it's interesting to see how mathematics played out in this kind of environment. Yeah.

AUDIENCE: Are we supposed to assume that he did kill his wife?

TOM LEIGHTON: Yes, and they assumed that, but when you decide whether or not to admit evidence, if it's prejudicial, you've got to have a really good grounds to get it in. Like if the evidence is going to make the jury think he did it, then you really got to argue the evidence is relevant somehow. There's material information and that's what the fight was about. A 1 in 2,000 relevance isn't going to cut it. 1 in 2, that's probably pretty relevant. And that will be the grounds on which the judge makes his decision. But yeah, you assume he didn't do it.

All right. Back to independence. So the last example today is derived from a famous paradox and has several actually important applications in computer science. And this problem is known as the birthday problem or the birthday paradox. It's a paradox because it sort of has a surprising answer. Probably a lot of you have seen this before in some form or another.

In the birthday problem, there are N birthdays and typically we're going to look at the case where N is 365, the days of the year, and there is M people. And for example, know maybe there's 100 people here. And what we want to know is, what is the probability that two or more people have the same birthday.

For example, how many people think there's at least a 50% chance that a pair of you in the audience here have the same birthday? That's good. How many people think there's a better than 90% chance? A few of you. All right. How many people think there's a better than a 99% chance that there's a pair of matching birthdays? A couple left.

How many think it's better than a 99.9% chance? We've got one, two. You guys are going to be stubborn. Another one. All right. How many people think it's more than 99.999% chance? Actually it's six 9's. It's incredible. It is a virtual certainty.

So let's see. In fact, the chance that you're all different is about 1 in 3 million chance that you're all different. And we're going to see why that's true here. But to do that, we're going to need to make two important assumptions. Any ideas about what assumptions you're going to need? Yeah.

AUDIENCE: Birthdays are uniformly distributed.

TOM LEIGHTON: Birthdays are uniformly distributed. Any other ideas? Yes.

AUDIENCE: He stole my answer.

TOM LEIGHTON: Oh, he stole yours. What else are you going to need to assume? Yeah.

AUDIENCE: All birthdays are independent of each other.

TOM LEIGHTON: Yeah. Mutually independent. We're going to need that as well. Now, in actuality, neither is true in reality. It's well known that birthdays tend to follow seasonal patterns and they're related to major events.

Now, do you all remember the big blackout that hit the Northeast several years ago? Do you remember that? Well, it turns out, this is a true fact, there were a lot of babies born nine months later. In fact, they had a name. They're called blackout babies. If you were born in that period in the Northeast and there's all these news stories about the life of the blackout babies.

And the same thing happens after cold snaps in the winter and you get a blizzard or this kind of a thing. Nine months later, you get babies. In fact, I had a personal experience with this. Well, my son was born on October 18, 1996. And on the day he was born, we're going to the hospital and it was a zoo.

The maternity ward was totally full. We had to go at some other wing of the hospital. And babies were popping out all over the place. And I asked, what is going on? Why don't you have enough room for all the mothers here?

And they said, oh, it's all the blizzard babies. And I go, what? And they go, well, remember the blizzard of '96? It's like, oh yeah. I remember. Yeah. It was nine months prior is the big blizzard and so it's all the blizzard babies coming.

So they're not uniform. They're all different probabilities here, but we're going to assume they're equally likely.

Now, independence is also not true, in general. What's one way that birthdays might not be independent? What is it?

AUDIENCE: Twins.

TOM LEIGHTON: Twins. So if they're twins, they have the same birthday. Now, there's other ways. In fact, my only sibling, my brother, has the same birthday I do, but I'm two years older, so we weren't twins. Now, you say, what are the odds of that? Well, 1 in 365, you think.

Well, one day I'm in middle school, about the age you start thinking about these things, and you get the idea to count back nine months from your birthday. Probably some of you have done that. And I did that and that's my dad's birthday. I was like, oh. May is not 1 in 365. It's like, Happy Birthday. I don't know.

Anyway, I almost needed to go into therapy after that, you know. So now you all got to count back nine months from your birthday. Anybody whose birthday is on September 30 or October 1, nine months back is New Year's Eve. That's dangerous. So in reality, birthdays are not independent and they are not randomly distributed, but we're going to assume that because we're going to use this same analysis for computer science problems where things are, hopefully, more independent and random.

Now, we're going to do an experiment to see how many people it takes us to get a pair of matching birthdays. So I'm going to run through people in order in the rows here, get your birthday and we're going to record and we're going to see how far we go until there's a match in that group. So I will write up the months here. And we'll start with my birthday is October 28.

So let's go right across. What yours?

AUDIENCE: April 1.

TOM LEIGHTON: April 1. OK. We won't embarrass you here. OK, who's next? What's your birthday?

AUDIENCE: I'm sorry. September 2.

TOM LEIGHTON: September 2. All right. Yours.

AUDIENCE: June 1.

TOM LEIGHTON: June 1. OK. We'll come back.

AUDIENCE: April 8.

TOM LEIGHTON: What is it?

AUDIENCE: April 8.

TOM LEIGHTON: April 8. All right.

AUDIENCE: November 20.

TOM LEIGHTON: November 20.

AUDIENCE: June 12.

TOM LEIGHTON: June 12.

AUDIENCE: December 29.

TOM LEIGHTON: December 29.

AUDIENCE: [INAUDIBLE].

TOM LEIGHTON: What is it?

AUDIENCE: June 14.

TOM LEIGHTON: June 14. Ooh, I almost got one there. That one's close. All right. What's yours?

AUDIENCE: March 6.

TOM LEIGHTON: March 6.

AUDIENCE: May 2.

TOM LEIGHTON: May 2.

AUDIENCE: 17th of November.

TOM LEIGHTON: November 17. Close again.

AUDIENCE: August 4.

TOM LEIGHTON: August 4.

AUDIENCE: July 25.

TOM LEIGHTON: July 25. I don't think we'll get to 100 here, hopefully. Yeah, what's yours?

AUDIENCE: October 30.

TOM LEIGHTON: What is it?

AUDIENCE: October 30.

TOM LEIGHTON: October 30. Got close.

AUDIENCE: July 6.

TOM LEIGHTON: July 6. All right.

AUDIENCE: February 25.

TOM LEIGHTON: February 25.

AUDIENCE: May 21.

TOM LEIGHTON: May what? 21st of May.

AUDIENCE: May 30.

TOM LEIGHTON: May 30. You guys fooled me. What have you got?

AUDIENCE: January 12.

TOM LEIGHTON: January 12. All right.

AUDIENCE: July 14.

TOM LEIGHTON: July 14. OK.

AUDIENCE: April 30.

TOM LEIGHTON: April 30.

AUDIENCE: March 13.

TOM LEIGHTON: March 13. All right. Did I get--

AUDIENCE: October 7.

TOM LEIGHTON: October 7.

AUDIENCE: October 8.

TOM LEIGHTON: Ah, you guys. OK. Did I get you?

AUDIENCE: September 15.

TOM LEIGHTON: September 15.

AUDIENCE: November 9.

TOM LEIGHTON: November 9. All right.

AUDIENCE: July 15.

TOM LEIGHTON: July 15. Close.

AUDIENCE: September 3.

TOM LEIGHTON: September 3. You guys are killing me here.

AUDIENCE: February 6.

TOM LEIGHTON: February 6.

AUDIENCE: October 26.

TOM LEIGHTON: OK.

AUDIENCE: November 2.

TOM LEIGHTON: November 2.

AUDIENCE: January 23.

TOM LEIGHTON: January 23.

AUDIENCE: September 27.

TOM LEIGHTON: You guys are going to set a record for sure here. This isn't the way it's supposed to go.

AUDIENCE: December 30.

TOM LEIGHTON: December 30.

AUDIENCE: December 28.

TOM LEIGHTON: Ah, come on, guys. What is the probability of going this long here? Yeah.

AUDIENCE: September 22.

TOM LEIGHTON: September 22.

AUDIENCE: July 30.

TOM LEIGHTON: July 30.

AUDIENCE: The 24th of August.

TOM LEIGHTON: 24th August. I'm going to have to ask the same person to tell me twice here to get a match. We got over there now?

AUDIENCE: April 6.

TOM LEIGHTON: April 6.

AUDIENCE: October 16.

TOM LEIGHTON: October 16.

AUDIENCE: Did ask how many--

AUDIENCE: September 3.

TOM LEIGHTON: September 3. All right. Very good. All right. Let's count and see how many we got here. 1, 2, 3, 4, 5, 6, 7, 8. 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42. That is a record. So it took 42 people to get a match.

Now it turns out that for N equals 365, the magic number for N is 23, that by 23 people, we got a 50-50 chance. In fact, the probability of a match on 23 people is 0.506. It's a little bit better than 50-50 chance at 23. Now, maybe we should figure out. It's too late for homework to figure out what the chances are of going this long without a match. That maybe worth figuring that out.

Now, it may seem surprising at first that 23 people is enough to have a 50/50 chance because the chance of any pair matching is 1 in 365, by our assumption. And that's small, but there's lots of pairs of people and every pair of people have a chance to match and that's why 23 turns out to be enough to get to 50-50.

Now, we're going to do the analysis for general M and N to the figure out the probability of a match if there's M people and N birthdays. There's lots of ways to do it. The easiest is to sort of well, we'll draw the sample space. It will be too big to draw the whole thing, but we can sort of model the sample space and then look at the sample points.

So you've got the first person and there's N birthdays here, so it could be anywhere from January 1 out to December 31 and in general this will be N. And then you have the second person and they have N possibilities for their birthday. And you take the tree down M levels to the very last person here.

So each node has degree N and there's M levels on this tree. So the sample space is the set of all n-tuples b1, b2, to bm, these are the birthdays where every value of bi is between 1 and N. So a sample point is all the birthdays of the M people.

How many sample points are there here? Remember how to count these things? Number of leaves on an N-ary tree of depth M or you can think of it this way. I've got N choices for each bi and there's M of them.

AUDIENCE: [INAUDIBLE].

TOM LEIGHTON: So what's the number of sample points?

AUDIENCE: N to the M.

TOM LEIGHTON: N to the M. Because N choices here, N choices here, N choices there, so you have N times N times N M times. And what's the probability of each outcome? For a set of possible birthdays, what's its probability? What's the probability of b1, b2, bM?

So the probability of a sample point. What's the probability that the first person has birthday b1, the second has b2, and the N-th has bM? Remember that? Yeah.

AUDIENCE: 1 over N to the M.

TOM LEIGHTON: 1 over N to the M because each edge is probability of 1 over N and the paths are length M, so you've got 1 over N to the M-th power. Probability of the first birthday matching is 1 in N times 1 in N times 1 in N. And this actually makes sense because I've got N to the M sample points, each a probability 1 over N to the M. So they all add up to 1, which is good.

What kind of sample space is this where this happens where all the probabilities are the same?

AUDIENCE: Uniform.

TOM LEIGHTON: Uniform. Makes it very easy to work with. All we got to do now is just count the number of sample points where there's a matching birthday and then we multiply by that one probability 1 over N to the M.

Now, it turns out that rather than counting the number of sample points where there's a matching birthday, it's easier to count the number of sample points for all the birthdays are different. And this is often the case when you're doing a counting problem, it's easier to count the opposite of what you're after. That can be the case and it is the case here. So we're going to do that.

So let's count how many sample points are all different birthdays, so no pair of bi's is the same. Let's do that. How many choices are there for b1? 365 or N. Let's do this in terms of N because we're going to use this for general N.

How many choices for b2? N minus 1. Given you are the first one, you can't match it. And then N minus 2 all the way over to the last one is N minus M plus 1. And this is a formula you should all remember. That's just N factorial over N minus M factorial. You did this sort of stuff a couple weeks ago with counting sets and probability is really-- a lot of it's about counting.

So now we can compute the probability that all the birthdays are different. It's just adding up all the sample points of which there's n factorial over N minus M factorial and multiply by the probability of each one, which is 1 over N to the M. All right. So we've actually now answered the question. This is the probability that all the birthdays are different.

The only problem is, it's not so clear what the answer is to actually compute this or how fast it grows. So if I wanted to get a closed form for this without the factorials, what do I do? What do I use? Stirling's formula.

So let's remember that. It says that N factorial is asymptotically equal to square root 2 pi N times N over e to the N. And that is accurate within 0.1% when N is at least 100. So not only is it asymptotically equal, it's right on track for a reasonable size N.

Now, I won't drag you through all the calculations. I used to actually try plugging that formula in for here and here and then going through all the calculations, but we won't do it in class. It's in the text. But I will tell you where that winds up. It's not hard, you've just got to do the calculation.

So this is means the probability that all birthdays are different turns out to be asymptotically equal to e to the N minus M plus 1/2 times the natural log of N over N minus M minus M. And that's accurate to within 0.2%, if N and N minus M are large, larger than 100. So in fact, it's almost equal.

And now you could plug in N equals 365 and M equals 100. So if you do that, in fact, if somebody has a calculator, we should plug in, what do we have, 42. You should plug in M equals 42 and see what the probability is. But if M is 100, the chance that we're all different, this equals 3.07 dot, dot, dot times 10 to the minus 7. And we should check for M equals 42. My guess is it's pretty small, but I don't know. We'll have to check that.

AUDIENCE: 0.0859.

TOM LEIGHTON: Great. So a 9% chance of having 42 people all miss is a 9% chance. So we were little unlucky. That won't happen very often. But when you go from 42 to 100, it gets really small. 1 in 3 million or so. Now, if N is 365 and M is 23, the probability comes out to be about 0.49, so about 50-50, they're all different.

Now. For general M and N, we'd like to know when do you get to the 50-50 point? We'd like to derive an equation for M in terms of N where the probability of being all different is about 1/2.

All right. So let's do that. So as long as we assume-- and this will turn out to be true-- that M is a little o of N to the 2/3 and remember little o means it grows slower than N to the 2/3. Then we can simplify that expression in asymptotic notation.

And when you do it, I won't drag it through on the board. It's also in the text, it turns out to be much simpler. It's just e to the minus M squared over 2N. So I take that thing up there and I assume that M is growing less fast than the 2/3 power of N and that whole upper expression reduces down to M squared over 2N. Everything else goes to 0 in the exponent. Doesn't matter.

Now, if I set this to be 1/2, I can solve this to find out what M has to be to make that be 1/2. All right. So this will be true if and only if minus M squared over 2N is equal to the natural log of 1/2. And that's true. Take the minus sign, put it inside to make a log of 2, multiply by 2N. That's true if M squared equals 2N natural log of 2.

And now I can solve for M really easily. That's true if and only if M equals the square root of 2 natural log of 2N, which is about 1.177 square root of N. So for general N, you get a 50% probability of having a matching birthday when M is in this range, pretty close to 1.2 square root of N.

Now, this square root N phenomenon, this thing here, that's what's known as the birthday principle. It says if you've got roughly square root of N randomly allocated items into N boxes or bins or birthdays, there's a decent chance two of the items will go into the same bin if the randomly allocated. In this case, the bins are the possible days of the year that we put each person into for their birthday. Any questions about that? Yeah.

AUDIENCE: M and N are like numbers like they're defined up there or does it mean to say M equals [INAUDIBLE]?

TOM LEIGHTON: Yeah. So here I looked at a special case where N was 365, M was 100, but we can imagine them as arbitrary numbers that could be getting large. And so over here and I say M is little o of N to the 2/3, I mean, well, M equals square root of N would qualify. Square root of N is little o of N to the 2/3. So as long as M is not growing too fast, I can simplify that expression up there, which is what I did.

And then we go back and we find, in fact, the square root of N the right answer and that is little o of N to the 2/3. And I have to use a different argument if I assumed M was bigger, which I didn't do. I didn't drag it for that. But I would have to go check that case.

So we can think of general is M and N as being arbitrary variables and potentially growing. M can be a function of N. And in fact, when M is the square root function of N, then we got a 50% chance of a match.

Now, the birthday principle comes up all over the place in computer science and it's worth remembering. For example, the generic form for this is when you have a hash function. Let's say I have a hash function, h, from a large set of items into a small set of items. For example, say I'm computing digital signatures. This is the space of all messages, this is the space of all 1,000-bit digital signatures, and h is a digital signature outcome.

Say I'm doing memory allocations. So all the things I might be sticking into a register, here's all the places it could go. Here's all the registers. Error checking. This is all the garbled messages in the world. This is the set of messages that make sense, all handled by functions, random kind of functions often.

Now, what you worry about when you're hashing is collisions. Let me define that. We say that x collides with y if the hash of x equals the hash of y, but x and y are different. For example, say you're looking at digital signatures. You would not want the signature for a $100 check to your mom to match your signature for $100,000 check to Boris. Because that would be bad because then Boris could come in and take that check to your mom for $100, converted to a $100,000 check to him and the signature is authentic if there's a collision in the signatures.

So very important when you're doing hash functions and in many applications, you don't want collisions because all the whole thing start breaking. Memory allocation. You don't want to assign two things in the same place. Error correction. There's only one answer you want to get out at the end.

Now, from the pigeon hole principle, you know if this set is bigger than that set, there is going to be a collision. That's what the pigeon hole principle says. Two guys will get mapped to the same thing. However, often in practice what we care about is a subset L prime of L that's pretty small because the set of messages we really assign is pretty small compared to all 1,000-bit signatures that are possible.

And what you'd like is that for this smaller set of messages, you might want to assign, they all get mapped one to one. And the birthday principle says life is not so nice. So let me write that down then we'll be done. All right. So the birthday principle says that if S is at least 100, L prime is a subset of L that is at least the square root of S.

So the cardinality of the things you want to hash is bigger than 1.2 square root the cardinality of S. And if the values of the function h on L prime are randomly chosen, uniform, and mutually independent, then there's at least a 50% chance, so with probability at least 1/2, there's a collision. There exists an x and a y such that x does not equal y-- and these are in L prime-- but h of x equals h of y.

All right. The proof is not hard, it's just we more or less did it. You just plug in the cardinality of L prime for M and the cardinality of S for N. And it's bad news because it means it doesn't take very many messages, just square root the number of signatures to get a collision. You'd hope you could get that you could have L prime be as big as S and that somehow they'd all go one to one, that everybody in this room would have a different birthday. That is not how it works if things are random, which is the case you usually like to have.

Now, this technique is used to crack cryptographic protocols and it's called the birthday attack based on the birthday principle. So what you do is, you get a bunch of messages that are encrypted and pretty soon you find two that get maybe encrypted the same way. And once you have that, now you can go back and crack the crypto system. For example, you break schemes like RSA with a birthday attack if this space is not big enough and that's one reason why now RSA, the keys have thousands of digits because otherwise you can use attacks like this and crack them more easily.

Any questions about that? OK. Very good. We're done for today.