[1] Hello. I'm Eric Hehner, from the computer science department, of the University of Toronto. I want to talk to you about probability. And since this is a video, I remind you that you can pause at any time you want to think about what I just said, and replay anything you want to hear again. You don't have to watch it straight through. [2] Let's start with a really simple probability problem. I have two children. At least one is a girl. What is the probability that the other one is also a girl? It's so simple it's like asking what's 2 plus 2. So everyone should see immediately that the answer is [3] one-third. But they don't. People argue about the answer. And they can't see the difference between this problem and [4] this one: I have two children. The older one is a girl. What is the probability that the younger one is also a girl? It's [5] one-half. How about [6] this problem: I have two children. The one named Pat is a girl. What is the probability that the other one, whose name is Chris, is also a girl? And are there hidden assumptions that we make when answering these questions? [7] Are we assuming that children are distinct things? [8] Are we assuming there are exactly two genders? [9] Are we assuming half the population of children consists of girls? [10] And on and on. It seems to me probability is not well understood, and that's why casinos and insurance companies do so well.

[11] The probability perspective in the title of my talk is really a combination of four perspectives. I'll say what Bayesian probability is, and then I'll talk a little about the connection between information and probability, and I'm going to claim that they are the same thing. And then I want to change probability questions from an argument about why the probability should be one thing or another into a simple calculation, and that's the formalist perspective. To do that, we have to have the right formalization, and that's programming. To show how it all works, I'll solve the famous problem of two envelopes. There are dozens of papers over the past fifty years, in probability journals and philosophy journals, about this problem. They all argue about why the probability should be one thing or another. But I just calculate the answer.

[12] Apparently, probabilists are divided into two camps. The old camp, and still the majority, are the frequentists. [13] They say that the meaning of probability is as follows: if you run an experiment a large number of times, the proportion of times that some event occurs is the probability of the event. But what about experiments that cannot be run many times? Like the gender of my children, for example. [14] Or the probability that a nuclear war will occur, or that a meteor will shatter the earth. We can't run those experiments a lot of times. [15] In the other view, probability is not a property of an event, but a measure of your ability to predict an event, and it depends on what you know, and it changes when you learn something new about the event. The Bayesian approach gives a way of updating a probability, but it has a [16] problem called prior probabilities, which means how do you get started. You need initial probabilities to update. I can solve that, or rather, make that problem go away by using

[17] information theory. According to information theory, [18] you have a set of messages m sub i, and [19] each message has a probability p sub i. [20] The information content of a message is minus the base 2 logarithm of its probability. [21] The average information content is called the entropy. In probability theory, they don't talk about messages; they talk about [22] events; but we can define the information content of an event just the same way. I don't really care whether we say message or event or something else, what I want is [23] the relationship between information and probability. *I* equals minus log p. And the unit of information is the bit.

When Claude Shannon invented information theory in 1948, probability theory was already well developed. There were already textbooks and courses on probability. But bits and bytes were unknown. So he defined information in terms of probabilities. But now, people know what bits and bytes are better than they know what probability is. We buy megabytes at the store, and we send them over the internet, and we wait for downloads. So we have a physical sense of what information is. So now, [24] maybe it makes more sense to define probability in terms of information, like this. Actually, whenever two quantities can be converted each one into the other, then they measure the same thing on different scales. [25] Like temperature in fahrenheit and celsius. Another example is [26] energy and temperature, which are related by the Boltzmann constant. And [27] Einstein's famous equation says that energy and mass measure the same thing because they are convertible each to the other. So I want to say [28] information and probability measure the same thing. I don't know what to call that thing they both measure. Maybe I'll just call it information, and say probability is another way to measure information. For celsius and fahrenheit, and for energy and temperature, and for energy and mass, the relationship between them is linear, so there's really no good reason for having two scales, except historical. But the relationship between bits and probability is not linear, so there may be a reason to have both scales. Suppose *I* bits and probability p are related this way, and suppose [29] J bits and probability q are related this way. Then we combine them by [30] adding the bits, or by multiplying the probabilities. And if you're really old like me, you know that a slide rule is a device that converts multiplications into additions, just like this formula. So we've got [31] probability and information, and the information has [32] a unit: the bit. For conversion, I need a unit of probability, which I'll call [33] the chance, which is really just a name for the number 1. And there's a [34] third scale for measuring this same quantity, and that's the number of states that a system can be in. And [35] here's the conversion chart for converting chance, state, and bits to each other. I think it's easiest to just look at some [36] points on these scales. Let's look at the middle line first. If there are 2 states, like a coin, then it takes 1 bit to say which state it's in, or will be in if you're predicting, and that's one-half probability for each state. On the top line, if the system has only 1 state, then it takes 0 bits to say which state it's in, and it's probability 1, or certain, that it's in that state. By the way, information doesn't just come in an integer number of bits. [37] If you know something is a decimal digit, and that's all you know about it, then when you find out which digit it is, you have just learned approximately 3.322 bits of information. If you want to store that information in a real, physical memory, you'll have to round that up to 4 bits. More generally, n states, and no other information, means log n bits, and it means probability 1 over n. [38] And states don't just come in integer numbers, either, although in any physical system, it will be an integer number of states. [39] And here's a teaser: negative information corresponds to a fraction of a state, and to a superdistribution of probability. But I won't talk about that.

[40] I have to get to the main part of the talk, which is about formalization. It's really just the scientific way to solve problems, and we learned it in high school. [41] Here's an example. Amanda is 164 cm tall. This is 8 cm more than 3 times her height at birth. Find her height at birth. First we formalize. That means [42] introducing a variable for each quantity in the problem, so *a* for Amanda's height now, and *b* for her height at birth, and then [43] expressing the given information as a binary expression: *a* equals 164 and *a* equals 8 plus 3 times *b*. When you formalize, you try to represent the given information as closely as possible. There should be nothing clever there because you want to be sure that you've represented the given information accurately. Then you [44] calculate. That means using algebraic laws. During the calculation, you don't care what the variables stand for. It doesn't matter that you're talking about Amanda's height. [45] And then you unformalize, which just means translating the answer back into the problems terms. A thousand years ago, philosophers would argue about what her height at birth was, and each would have reasons why their answer is right. But today we don't argue; we just calculate. Probabilists don't seem to have discovered this paradigm yet; they argue, and I think it's because they haven't discovered a good formalism yet.

[46] Finally, let's get to probability. It's a real number between 0 and 1. [47] And a distribution is an expression whose value is a probability, and whose sum is 1. [48] For example, if n and m are variables over the positive natural numbers, 1, 2, 3, and so on, then 2 to the power minus n minus m is a distribution, because [49] for each value of n and m, 2 to the power minus n minus m tells you the probability that n and m have those values, and if you add up all those numbers you get

1. [50] If the expression does not add up to 1, you can normalize it by dividing by its sum and get a distribution. [51] For example, if we let n and m vary over the natural numbers, 0, 1, 2, and so on, then 2 to the minus n minus m adds up to 4, so we divide by 4 to get a distribution. I'm only dealing with discrete distributions, so that's all the probability theory we need for this talk.

[52] Now we need a formalism to formalize probability problems. A lot of those problems talk about some activity. Maybe there is a sequence of events. Maybe some events are conditional upon other events. Maybe there's a repetition of events. Describing activities formally is just what programmers do. So here [53] are some programming notations. The simplest is [54] ok, which means do nothing. Some people call it the empty statement. Some people call it skip. [55] This is the assignment, variable x gets the value of expression e. [56] Here's an if, or conditional composition. [57] This is sequential composition, and [58] parallel composition, and [59] a loop construct. [60] I'm not advocating these notations or any particular programming notations; I just need a sample of programming notations to show you how to formalize probability problems.

[61] First, we generalize from binary to probability, with true being 1 and false being 0. Now we can put a probability after the word if, and after the word while. And the other generalization is to [62] define each program as a distribution of the final state. ok becomes a one-point distribution of the final state. If the variables' initial values are x and y, and their final values are x prime and y prime, this says the final values equal the initial values with probability 1, and they equal any other values with probability 0. Assignment is also a one-point distribution. Conditional, sequential, and parallel composition take distributions and compose them into a new distribution. And we'll see the loop construct later. There are lots of [63] laws that can be proven, but I'll just show you one: the substitution law, which says that an assignment followed by a distribution equals the same distribution but with a substitution.

[64] Here's an example. if one-third — that means with probability one-third, x gets 0, and with the remaining probability two-thirds, x gets 1. Following that, if x is 0, then — with probability a half add 2 to x, and with probability the other half add 3. And if x wasn't 0, then with probability a quarter add 4, and with the remaining probability three-quarters add 5. It's just a silly example to [65] show some calculation. Replacing the assignments and ifs by their definitions gives this. You can check it if you're fast; otherwise check them later; they're all in a paper on my web site that this talk comes from. [66] Now replace the semicolon by its definition and get this. Now we don't have any programming notations left, [67] we do the sum, and we get this. Now suppose the final value of x is 2. Then x-prime equals 2 is true, or 1, and x-prime equals 3 is false, or 0, and so on. So this is 1 sixth plus 0 sixths plus 0 sixths plus 0 halves, so the sum is one sixth. That's the probability that the final value of x is 2. The probability that x ends at 3 is a sixth. The probability that x ends at 4 is 0. And so on. The probabilistic program is telling us the probability distribution of the final state. — [68] [69] In general, after executing S, the average value of number expression e is S semicolon e. [70] And, after executing S, the probability that binary expression b is true is S semicolon b. [71] Probability is just the average value of a binary expression. For example, [72] if we follow this little program with [73] x, and follow the rules of calculation, we get [74] four and two-thirds, so that's the average value of x. If we follow it with [75] x greater than 3 we get two-thirds, so that's the probability that x will be greater than 3. Ok, that's enough theory for now.

[76] Let's do an example. Monty Hall is a game show host, and in this game there are three doors. A prize is hidden behind one of the doors. The contestant chooses a door. Monty then opens one of the doors, but not the door with the prize behind it, and not the door the contestant has chosen. Monty asks the contestant whether the contestant would like to change their choice of door, or stay with their original choice. What should the contestant do? What *we* should do is to formalize the given information. To start with [77] let variable p say where the prize is, so assign to p a random choice of door 0 or door 1 or door 2. That's not quite formal yet. [78] This says give p a value of 0 or 1 or 2, each one with probability one-third. [79] Now let variable c say which door the

contestant chooses, 0 or 1 or 2, each with probability one-third. [80] Now it's Monty's turn to choose a door. If the contestant has chosen the door with the prize behind it, then Monty chooses either of the other two doors, each with probability one-half. The circle plus is addition modulo 3. If the contestant didn't choose the door with the prize behind it, then there's only one door left for Monty to choose. Now the contestant can either stick with the choice already made, or switch to another door. Let's say the contestant [81] sticks with the choice they already made. Now we ask [82] whether the contestant has chosen the door with the prize. We don't argue about it. We don't start a blog and get everyone's opinion. We just apply the definitions, and [83] calculate the answer, and it's one-third. Now suppose [84] on this line, instead of sticking with their choice, the contestant [85] decided to switch. Again, don't argue, just calculate, and you get two-thirds. That means the contestant should switch to increase their chance of winning. A lot of people have spent a lot of words trying to explain why this probability should be two-thirds. To me, that's like trying to explain why 264 plus 179 should be 443. It's just the result of a calculation.

I started this talk with [86] the problem of my two children. To formalize, [87] let c and d be the genders of my two children. They are variables over two values, so they are binary, and [88] let's say girl is 1 and boy is 0. We start off by saying [89] that at least one is a girl. That's a binary expression, but in general binary expressions are not distributions, so we have to [90] normalize it. And then we want to know whether the other one is also a girl, so that's [91] c and d, which says they are both girls. That's my formalization. [92] Here's my calculation. It's not long, and not hard, and the answer is one-third, and you can check it later. Now let's [93] change the problem. This time c is the older child, and we learn that she is a girl. And we ask whether the younger child d is also a girl, and the calculation says one-half. [94] I asked you whether we are assuming that children are distinct, unlike raindrops in a barrel. It's the formalization that answers the question. We used two variables, c and d, so [95] yes, we assumed they are distinct. With raindrops in a barrel, you can't point to one and say: that one. But we can say how many there are by knowing the weight or volume of a raindrop, and the weight or volume of water in the barrel. So we formalize [96] with a variable that says how many raindrops there are. Let's say they're in a thimble, because a barrel seems a bit big for two raindrops. Let n be the number of acidic raindrops. [97] We learn that n is at least 1. And we ask if n is 2. That's the formalization. [98] The calculation says one-half. [99] I asked if we assumed that children come in exactly two genders. [100] Yes, we did, by making our variables be two-valued. But there could be any number of subgenders of girl and boy. [101] Did we assume that half the population of children are girls? You may be surprised, but [102] the answer is no. Probability one-half for girl and one-half for boy does not mean half the population are girls and half are boys. It means we have no idea whether a particular child is a girl or boy. That was the information perspective that I told you about earlier. We don't even know if there are any other children, other than my two. But if we find out that there are other children, [103] and that one-third of them are girls, that will affect the answer. [104] We can say that c is a girl with probability one-third like this, and similarly [105] d is a girl with probability a third. And [106] the problem says that at least one is a girl, and asks [107] if they are both girls, and [108] the calculation says one-fifth. Ok, it's a bit long, but if you look at the right side, it's just applying definitions and summing, so it's not complicated, and it can be automated. The point is always the same: you don't need to think, or reason, or argue about it. You just formalize the problem, and calculate the answer.

Up to now, we could always calculate the answer because we haven't had any loops. [109] Here's our first example with a loop. A two-position switch is flipped some number of times. At each time there is probability 1/2 of continuing to flip, and probability 1/2 of stopping. What is the probability that the switch ends in its initial state? The program [110] looks like this. While coin shows head, do flip switch. Well, that's not fully formalized yet, but the point is there's a loop. And when you have a loop, you have to make a [111] hypothesis about what the distribution is. And when you have a hypothesis, you can [112] replace the loop with constructs we already have. In this case, the while loop becomes: if coin shows head, then flip switch and repeat the hypothesis, otherwise you're done. And we already know about if and semicolon and ok, so we know all we need to prove or disprove this equation. Now let's finish formalizing. The coin has two sides, so [113] coin shows head is probability one-half. For the switch, we can introduce a binary variable, and [114] flip switch is this. Now all we need is a hypothesis, and it's not obvious what that should be. So I'll just [115] show it to you. If the switch ends in its initial state, ok is true, which is 1, so that's probability two-thirds. If the switch ends in the opposite state, ok is 0, and that's probability one-third. And [116] here's the proof. If you make the wrong hypothesis, you find out it's wrong because the proof attempt fails. And the great thing is that the way the proof attempt fails tells you how to fix the hypothesis. But I don't have time to show you that.

[117] Here's an example that will amaze you. Start p off at 1. Now flip a coin. If it's head, double p. Keep flipping and doubling as long as you get a head. Stop the first time you see a tail. And I want to know how many times p gets doubled, so I'll include a [118] counter or time variable. Well, [119] here's the distribution. It says t-prime is greater than or equal to 0, meaning time doesn't go backwards. And the final value of p is a power of 2. And t-prime says what power of 2, so variable t is counting the doublings. And the last part says that p ends at 1 with probability a half, and at 2 with probability a quarter, and at 4 with probability an eighth, and so on. There's nothing surprising so far. To prove it, we need a hypothesis for the loop, and [120] here it is. As usual, the hypothesis looks a lot like the final distribution, so it's not too hard to find, but sometimes it takes a trial or two. This equation is trivial to prove; just use the substitution law, replacing p with 1 and t with 0 in the hypothesis. And the hypothesis is [121] easy enough too — just a two-step proof. "While a half" becomes "if a half", then the loop body, followed by the loop hypothesis, else the loop execution is finished. Now we just replace if and semicolon and ok with their definitions. [122] If a half becomes division by 2 in both the then part and the else part. [123] The then part is just two substitutions, replacing t with t plus one and replacing p with p times 2. [124] The else part, which is ok, becomes t prime equals t times p prime equals p. Then we simplify.

Now, to find the [125] average value of t-prime, just write the distribution, followed by t. Then [126] simplify, and it's 1, which says that [127] on average, the loop body is executed once. That's not amazing. But let's look at [128] the average value of p-prime. We write [129] the distribution, followed by p. Then [130] simplify, and it's infinity. On average, p is doubled once, and on average, its final value is infinity. I think that's amazing.

[131] Finally, we can do the famous problem of two envelopes. Imagine that I have two envelopes, [132] and each one contains some money. Let's start with an easy version of the problem and say each envelope contains an integer amount from 1 to a hundred dollars. You choose an envelope. You can look in it if you like, and see how much you've got. And now you can keep what you have, or switch. What should you do? This isn't hard: [133] if you have $50 or less, you should switch. That's the best strategy for ending up with the most money. Now, [134] I'm changing the upper limit on the amount to something enormous. Is it still true that you should switch if you see less than half that amount? How about if I [135] remove the upper limit altogether? When you look in the envelope, you are going to see a finite amount, so that's less than half of infinity dollars. Does that mean you should always switch? Seems odd, doesn't it? But that's what [136] Blaise Pascal thought, and he was a great mathematician. [137] Now I'm going to change the problem again. This time the envelopes contain positive rational amounts, and one contains twice as much as the other. This is the famous version. What should you do? Well, [138] here's one way to argue: Let the amount you see be x. If you switch, then with probability a half the other envelope has twice as much and you gain x, and with probability a half the other envelope has half as much and you lose x over 2. So that's an [139] expected gain of x over 4, which is [140] positive, so you should switch. According to this argument, it doesn't matter how much is in the envelope you chose first, so you don't even have to look in it. You just choose an envelope, and then you switch envelopes. It's

crazy. I mean, you could [141] make the same argument again and switch back. [142] Here's another argument: Let the amount in the other envelope, the one you didn't choose, be y. If you switch, then with probability a half, y is half of what you're holding and you lose y, and with probability a half y is twice what you're holding, and you gain y over 2. That's an expected loss, so you should *not* switch. Both arguments sound good, but they come to opposite conclusions. [143] Here's yet another version of the problem. I start off with one dollar, then I repeatedly flip a coin, doubling the amount each time I get head, and I stop the first time I get tail. That's the loop with the amazing average. And I use it to determine how much I put in one envelope, and I put twice that much in the other envelope. I tell you that's what I'm doing, but you don't see me do it and you don't know what the amounts are. So how does that affect what you do? Well, you might argue [144] that it doesn't affect what you do at all, because all you know is that there's twice as much in one envelope as in the other, same as before. For 50 years there have been papers in probability journals and philosophy journals making these arguments and other similar arguments about this problem, and these arguments are all wrong. [145] What they need to do is stop arguing, and start formalizing and calculating. [146] Let the amounts in the envelopes be x and y. [147] Choosing an envelope is formalized as: if a half then z gets x else z gets y. If you decide to stick with that envelope, [148] that's ok. And if you want to know how much you've got, [149] on average, then follow this with z. And the [150] calculation says, on average, you have x plus y over 2. And if you [151] decide to switch, you've still got x plus y over 2. Or, instead of making two calculations, [152] we can ask if the switch value is more than the stick value, and we find that if x and y are unequal, the probability is a half. Or, we might want to know [153] just what is the difference between the switch value and the stick value, and we find that on average it's 0. I'm not showing the calculations, but they're easy, and you can do them yourself later, or read them in the paper. All those calculations were in the absence of any information about how x and y are chosen, and without any strategy. A [154] strategy means you look in the envelope, and if you see less than some strategic amount s, then you switch, and otherwise you stick. [155] In the first version of the problem, x and y were between 1 and 100, and that makes the average profit look like this, which is at its maximum when s equals 50, [156] and that maximum is 12 and a half dollars profit.

[157] Let's look at the case where there's twice as much money in one envelope as there is in the other. If you don't look in the envelope, it doesn't matter if you stick or switch. Even if you look, if you don't have any strategy, then it doesn't matter if you stick or switch. So let's [158] put in a strategy. To choose a good value for s, we need to know how x is chosen. So [159] I start with 1 dollar, and I keep flipping and doubling until I see a tail. Now it says that if s is 0, which means always stick, the profit is 0, and if s is infinity, which means always switch, the profit is also 0, and if s is anything in between, the profit is one quarter. Amazingly, it doesn't matter what you choose for s, as long as you choose some nonzero, finite value, the profit will be 25 cents on average. Fifty years of informal reasoning have never figured this out, but it's a simple calculation. Now suppose [160] that instead of doubling, we keep tripling the amount in one envelope, and then put triple that in the other. The strategy is to switch if you see less than 3 to the power n for some n. And the profit looks like this. As n increases, the profit increases, as long as n remains finite. If n is infinite, the profit is 0. So choose a really large, but finite value of n.

[161] If you have a coin, and the only thing you know about it is that is has two sides and flipping gives you one of them, then each side has probability a half. But if you look at the coin carefully, you might see that it's not perfect; maybe it's a tiny bit bowl shaped, and that affects the probabilities. Maybe it's denser in the middle, and that can affect the probabilities. This example is how to make probability one-half even if the coin is imperfect. [162] Flip the coin twice. If the outcomes differ, use the first outcome. If the outcomes are the same, repeat the experiment, until the two outcomes differ, and then use the first outcome of the first pair that differ. [163] Here's what we want: probability a half for each side. [164] Here's what we get: a pair of flips, x and y, that differ,

with probability a half that it's head first and tail second, and probability a half that it's the other way round. And that can be [165] simplified. And I'll call it [166] R. So [167] here's the program. Flip the coin, and with some probability p it's head, and with probability 1 minus p it's tail. Flip it again. If the two flips are equal, repeat, and otherwise you're done. Or, is [168] this the program? In the first program, you flip twice and see if they're equal, then you flip twice more, and see if they're equal, and so on. In the second one, you flip twice and see if they're equal, then you flip once more and see if it equals the previous flip, then once more and see if it equals the previous flip, and so on. When you just say something in English, or any natural language, it may seem clear, but different people can understand it differently. And they might not know that they understand it differently. Each person says it's clear, but then they argue past each other about whether it works or not. When you write a program, you make your understanding completely unambiguous. As it turns out, the [169] first one can be proven, and it's not hard. But the [170] second one can't be proven. Actually the second one is equivalent to a single flip, with probability p, not a half. [171] And if you want to know how long it takes, [172] you just put t gets t plus 1 in the body of the loop, and you can prove this complicated timing distribution that depends on p, and calculate [173] the average time.

[174] Conclusions. [175] I talked a little about Bayesian probability, which changes as you learn more about the situation. [176] And I suggested that probability and information are just different scales for measuring the same thing. [177] Then I suggested that probabilists should start behaving more like scientists, by formalizing, calculating and unformalizing. [178] And finally I brought in some formal methods of programming, which treats programs as mathematical expressions, and in this case as probability distributions. [179] So we can use programs to formalize probability problems, and calculate the resulting distributions. Well, that's the end of my talk. I hope it was worth watching. If you're interested, my paper titled a Probability Perspective, on my website, has lots more information and examples.