PROFESSOR: Today I'm starting a new topic and that's always the occasion for putting things into perspective. Keep in mind what we were trying to do in the subject. We were trying to introduce several intellectual themes. The first, and absolutely the most important, is how do you design a complex system? We think that's very important because there's absolutely no way this department could exist the way it does, making things like that, hooking up internets and so forth. Those are truly complex systems.

And if you didn't have an organized way of thinking about complexity, they're hopeless. So the kinds of things we're interested to teach you about are just hopeless if you can't get a handle on complexity. So that's by far the most important thing that we've been thinking about.

We've been interested in modeling, and controlling physical systems. I hope you remember the way we chased the robot around the lab, and that was the point there. We've thought about augmenting physical systems by adding computation, I hope you've got a feel for that. And we're going to start today thinking about how do you build systems that are robust.

So just in review, so far-- you've already seen most of this-- so far we've taught you about abstraction, hierarchy, and controlling complexity starting primarily by thinking about software engineering. Because that's such a good pedagogical place to start. We introduced the idea of PCAP, and that has continued throughout the rest of the subject, then we worried about how do you control things. We developed ways of modeling so that you could predict the outcome before you actually built the system. That's crucial. You can't afford to build prototypes for everything, it's just not

economical.

And so this was an exercise in making models, figuring out how behaviors relate to the models, and trying to get the design done in the modeling stage rather than in the prototyping stage, and you built circuits. This had to do with how you augment a system with new capabilities, either hardware or software.

Today what I want to start to think about is, how do you with uncertainty? And how do you deal with things that are much more complicated to plan? So the things that we will do in this segment are things like mapping. What if we gave you a maze-- you, the robot. What if we gave the robot a maze and didn't tell them the structure of the maze? How would it discover the structure? How would it make a map? How would it localize? What if you had a maze-- to make it simple, let's say that I tell you what the maze is. But you wake up-- you're the robot, you wake up, you have no idea where you are. What do you do? How do you figure out where you are? That's a problem we call localization.

And then planning. What if you have a really complicated objective? What's the step-by-step things that you could do to get there? Those are the kinds of things we're going to do, and here's a typical kind of problem. Let's say that the robot starts someplace, and say that it has something in it that lets it know where it is, like GPS. And it knows where it wants to go. Making a plan is not very difficult, right? I'm here and I want to go there, connect with a straight line. And that's what I've done here. The problem is that unbeknownst to the robot, that path doesn't really work. So on the first step, he thinks he's going to go from here to here in a straight line. The blue represents the path that the robot would like to take, but then on the first step the sonars report that they hit walls. And those show up as the black marks over here.

So already it can see that it's not going to be able to do what it wants to do. So it starts to turn and it finds even more places that don't work. Try again. Try again. Notice that the plan now is, well I don't know what's going on here, but I certainly can't go through there, so I'm going to have to go around it. Keep trying. Keep

trying. Notice the plan. So he's always making a plan that sort of make sense. He's using for each plan the information about the walls that he's already figured out. And now he's figured out, well that didn't work. So now back track, try to get out of here.

AUDIENCE: Is he backtracking right now? Or is he--

PROFESSOR: Well, he's going forward. He's making a forward plan. He's saying, OK now I know all these walls are here, and I'm way down in this corner, how do I get on the other side of that wall. Well given the information that I know, I'm going to have to go around the known walls.

So my point of showing you this is several fold. First off it's uncertain. You didn't know at the outset just how bad the problem was. So there's no way to kind of pre-plan for all of this. Secondly, it's a really hard problem. If you were to think about structuring a program to solve that problem, in a kind of High School programming sense-- if this happens then do this, if this happens then do this-- you would have a lot of if statements, right? That's just not the way to do this.

So what we're going to learn to do in this module is think through much more complicated plans. We're going to be looking at the kind of plans like shown here that just are not practical for, do this until this happens, and then do this until this happens, and then while that's going on, do this. It's just not going to be practical, that's the idea.

So the very first element, the thing that we have to get on top of first, is how to think about uncertainty. And there's a theory for that and the theory is actually trivial, the theory is actually simple, except that it's mind-boggling weird that nobody can get their head around it the first time they see it. It's called Probability Theory. As you'll see in a minute, the rules are completely trivial. You'll have no trouble with the basic rules. What you will have trouble with-- unless you're a lot different from most people-- the first time you see this theory it's very hard to imagine exactly what's going on. And it's extremely difficult to have an intuition for what's going on. So the theory is going to give us a framework then for thinking about uncertainty. In

particular, uncertainty sounds uncertain.

What we would like to do is make precise statements about uncertain situations. Sounds contradictory, but we'll do several examples in lecture and then you'll do a lot more examples in the next week. So that you learn exactly what that means. We would like to draw reliable inferences from unreliable observations. OK, you have a lot of experience with unreliable observations, right? The seminars don't tell you the same thing each time. That's what we'd like to deal with. We would like to be able to take a bunch of different individually, not all that reliable observations, and come up with a conclusion that's a lot more reliable than any particular observation. And when we're all done with that what we'd like to do is use this theory to help us design robust systems. Systems that are not fragile. Systems that are not thrown off track by having a small feature that was not part of the original formulation of the problem.

So that's the goal. And what I'd like to do is start by motivating it with the kind of practical thing to get you thinking. So here's the game, Let's Make a Deal. I'm going to put 4 LEGO bricks in a bag. OK. LEGO bricks, you've seen those probably. Bag. The LEGO bricks are white or red. There's only going to be 4, and you're not going to know how many of each there is. Then you get to pull one LEGO brick out, and if you pull a red one out, I'll give you 20$. The hitch is you have to pay me to play this game.

So the question is, how much are you willing to pay me to play the game? So, I need a volunteer. I need somebody to take 4 LEGOs and not let me see, OK, please, please. I want you to put 4 LEGOs, only four. They can be white or red. If you have LEGOs in your pockets that are a different color, don't use them. You're allowed to know what the answer is but you're not allowed to tell me, or them. So OK well come over here. So bag, LEGOs, hide, put some number in. Oh, no no no. Wait, wait, wait. Put them back, put them back. I'm not supposed to see either. OK, I'll go way. OK, 4. OK, so we'll close the bag, right? And I'll call you back later, but it'll be nearer the end of the hour. So here's 4 LEGOs, sort of sounds like 4 LEGOs, it's more than one. OK, so how much would you be willing to pay me to play the

game?

AUDIENCE: 5$.

PROFESSOR: 5$ -- Can I get more? I want to make money. Can I get a higher bid? More than 5$..

AUDIENCE: $9.90.

PROFESSOR: How much?

AUDIENCE: $9.90.

PROFESSOR: $9.90, very interesting. Can I get more than $9.90?

AUDIENCE: $9.99 and a half.

PROFESSOR: $9.99 and a half? Magic number.

AUDIENCE: 10$.

PROFESSOR: 10$. Can I hear even a penny more? A penny more?

AUDIENCE: I'll offer a penny more. You just have to go to the bag.

PROFESSOR: I thought we were being very careful and letting them not know.

AUDIENCE: No, no, no. Aren't you going to put 4 white blocks in all the time?

PROFESSOR: I didn't do it. That person did it. It wasn't me. I'm innocent. I'm completely fair. Yeah?

AUDIENCE: Are we imagining that you are equally as likely to put any number of blocks in? So, are we able to say that she's more likely to put it all white? Because that just changes how you calculate it.

PROFESSOR: OK, that's an interesting question. We need a model of a person. That's tricky. OK, I have another idea. Two more volunteers. OK, volunteer, volunteer. Here's the experiment one person will hold the bag up high, so that the other person can't see it, and the other person-- I didn't look in, notice I'm being very careful, I'm very

honest, right? Except for the X-ray vision, which you don't know about. Everything is completely fair. And the little window in the back you don't know about that either. So, one person holds it up so the other person can't see in, the other person grabs a LEGO and pulls it out and lets everybody see that LEGO.

It was intended to make it hard to see in. OK, red one. OK, that's fine, so we're done.

AUDIENCE: We each should get $20, right?

PROFESSOR: No, no, no. This was a different part of the bet. No, no, no, no, no. Thank you, thank you. Now how much would you pay me to play the game?

AUDIENCE: With that one out?

PROFESSOR: No, we'll put that one back. OK, so this one came out, it was red. Now without looking, I'm going to stick it back in. OK, so we pulled it out. So what do we know? We know there's at least 1 red. OK, now what are you willing to pay to play the game?

AUDIENCE: 5$..

PROFESSOR: 5$.. Yes? 5$..

AUDIENCE: $4.99.

PROFESSOR: $4.99? Wait a minute. Should you be willing to pay more or less? I got it up to 10$. Should you be willing to pay more or less now?

AUDIENCE: More.

PROFESSOR: More, why? The same. More. The same.

AUDIENCE: You're insured that there's at least 1 red block.

PROFESSOR: I know that there's at least 1, but didn't I know that before? No. The person could have been e that first person could have loaded it, because I was giving her a cut. I

didn't talk about this before. This is not a set-up. So I want to vote. How many people would give me less than 10$? I'm going to give you [UNINTELLIGIBLE] first. 10 to 12. Let's see, 13 to 15, 16 to 18, more than 18. So how many people would give me-- you're only allowed to vote once. Keep in mind that I'm more likely to choose you if you vote high. Right? Vote high. So how many people would give me less than $10 to play the game? A lot, I would say 20%. How many people would give me between $10 and $12? A lot smaller, 5%. How many people would give me between $13 and $15? Even smaller, 2%. How many people would give me between $16 and $18? Wait, these numbers are not going to add up to 100%.

OK, we'll learn the theory for how to normalize things in a minute. OK, so we're down to about 1%. How many people would give me more than $18? One person. Thank you, thank you. So that's 1 in 200 or 0.05%.

OK, so what I'd like to do now is go through the theory that's going to let us make a precise calculation for how much a rational person-- not to say that you're not rational-- but how much a rational person might be willing to pay. So that was the set up, then we'll do the theory, then we'll come back at the end of the hour and see how many people I would have gypped-- made money, or whatever.

OK, so we're going to think about probability. And the first idea that we need is set theory. Because we're going to think about experiments having outcomes, and we're going to talk about the outcomes being an event. An event is any describable outcome from an experiment. So for example, what if the experiment were to flip 3 coins in sequence. An event could be head, head, head. And you could talk about was the outcome head, head, head. The event could be head, tail, ahead. The event could be 1 head and 2 tails. The event could be the first toss was a head. So the idea is there's sets that we're rethinking about. And we're going to think about events as possible outcomes being members of sets.

There's going to be a special kind of event that we're especially interested in, and that is an atomic event. By which we mean finest grain. Finest grain is kind of amorphous idea. What it really means is for the experiment at hand, it doesn't seem

to make sense to try to slice the outcome into two smaller units. You keep slicing them down until slicing them into a smaller unit won't affect the outcome.

So for example, in the coin toss experiment, I might think that there are 8 atomic events. Head, head, head, head, head tail, head, tail, head, head, tail, tail, blah, blah, blah. So I've ignored some things like, it took 3 minutes to do the first flip, and it took 2 minutes to do the second one. Right? That's the art of figuring out what atomic units are. So for the class of problems that I'm thinking about, those things can be ignored so I'm not counting them. But that's an art, that's not really a science. So you sort of have to use good judgment when you try to figure out what are the atomic events for a particular experiment. Atomic events always have several properties, they are always mutually exclusive.

If I know the outcome was atomic event 3, then I know for sure that it was not atomic event 4. And you can see that these events up here don't have those properties, right? So the first toss-- here's an event head, head, head, which is not mutually exclusive with the first toss was a head. So atomic that events have to be mutually exclusive.

Furthermore, if you list all of the atomic events, that set has to be collectively exhaustive. Collectively exhaustive? What buzz words? OK, that means that you've exhausted all possibilities when you've accounted for the collective behaviors of all the atomic events. And we have a very special name for that because it comes up over, and over, and over again. The set of atomic events, the maximum set of atomic events, is called the sample space.

So the first thing we need to know when we're thinking about probability theory, is how to chunk outcomes into a sample space. Second thing we need to know are the rules of probability. These are the things that are so absurdly simple, that everybody who sees these immediately comes to the conclusion that probability theory is trivial, they then don't do anything until the next exam, and then they don't have a clue what we're asking. Because it's subtle, it's more subtle than you might think. Here's the rules, probabilities are real numbers that are not negative. Pretty easy.

Probabilities have the feature that the probability of the sample space is 1. That's really just scaling. That's really just telling me how big all the numbers are. So if I enumerate all the possible atomic events, the probability of having one of those as the outcome of an experiment, that probability is 1. Doesn't seem like I said much, and I'm already 2/3 of the way through the list. Yes?

AUDIENCE: Doesn't that just mean that something happened?

PROFESSOR: Something happened, yes. And we are going to say that this certain event has probability 1. All probabilities are real, all probabilities are bigger than 0, and the probability of the certain event-- written here as the universe, the sample space-- the probability of some element in the sample space is 1. The only one that's terribly interesting is additivity. If the intersection between A and B is empty, the probability of the union is the sum of the probabilities of the individual events. Astonishingly, I'm done. And this doesn't alter the fact that people are still, to this day, doing fundamental research in probability theory.

There are many subjects in probability theory, including many highly advanced graduate subjects, all of which derive from these three rules. It's absurd how un-intuitive things can be given such simple beginnings. Just as an idea. So you can prove all of the interesting results from probability theory-- you can prove all results from probability theory with these three rules, and here's just one example. If the intersection of A and B were not empty, you can still compute the probability of the union, it's just more complicated than if they were empty, if the intersection were empty. Generally speaking, the probability of the union of A and B, is the probability of A plus the probability of B, minus the probability of the intersection. And you can sort of see why that ought to be true, if you think about a Venn diagram.

If you think about the odds of having A in the universe-- the universe is the sample space-- probability of having sum event A, the probability of having sum event B, the probability of their intersection. If you were to just add the probability of A and B, you doubly count the intersection. You don't want to double count it, you want to count it once. So you have to subtract one off. So that's sort of what's going on.

OK, as I said the theory is very simple. But let's make sure that you've got the basics first. So experiment, I'm going to roll a fair, 6-sided die. And I'm going to count as the outcome the number of dots on the top surface, not surprisingly. Find the probability that the roll is odd, and greater than 3 You have 10 seconds.

OK, 10 seconds are up. What's the answer? (1), (2), (3), (4) or (5)? Raise your hands. Excellent, wonderful. The answer is (1). The way I want you to think about that is in terms of the theory that we just generated because it's useful for developing the answers to more complicated questions. In terms of the theory, what we will always do, the process that always works, is enumerate the sample space. What's that mean? That means identify all of the atomic events. The atomic events here are the faces that show are 1, 2, 3, 4, 5, 6. Enumerate the sample space. And then find the event interest.

So here the event was a compound event. The result is odd and greater than 3. Odd, well that's 1, 3, 5, shown by the check marks. Bigger than 3, that's the bottom 3 check marks. If it's going to be both, then you have to look where there's overlap and that only happens for the outcome 5. Since there's only 1, and so fair meant that these probabilities were the same. If you think through the fundamental axioms of probability, if they're equal, they're all non-negative real numbers, and they sum to 1, then they are all 1/6. So the answer is 1/6, right? OK, that was easy.

The rule that is most interesting for us, happens not surprisingly to also be the one that people have the most trouble with. Not excluding the people who originally invented the theory. The theory goes back to Laplace. A bunch of people back then who were absolutely brilliant mathematicians, and still it took a while to formulate this rule. It was formulated a guy named Bayes.

Bayes' theorem gives us a way to think about conditional probability. What if I tell you, in some sample space, B happened? How should you relabel the probabilities to take that into account? Bayes' rule is trivial, it says it if I know B happened, what is the probability that A occurs, given that I know B happens? And the rule is, you find the probability of the intersection.

AUDIENCE: How do you do that?

PROFESSOR: We'll do some examples. So we need to find the probability of the intersection, and then we have to find the probability of B occurring, and then we normalize-- a word I used before, and that's exactly what we need to do to that distribution-- we normalize the intersection by the probability of B.

That's an interesting rule. It's the kind of thing we're going to want to know about. We're going to want to know-- OK, I'm a robot. I'm in a space. I don't know where I am. I have some a priori probability idea about where I am, so I think I'm 1/20 likely to be here, I'm 1/20 likely to be there, et cetera, et cetera. And then I find out the sonars told me that I'm 0.03 meters -- no it can't be that small, 0.72 meters from a wall. Well, how do I take into account this new information to update my probabilities for where I might be? That's what this rule is good for. So here's a picture. The way to think about the rule is if I condition on B, if I tell you B happened, that's equivalent to shrinking the universe -- the universe U, the square. That's everything that can happen. Inside the universe, there's this event A and it does not occupy the entire universe.

There is a fraction of outcomes that belong logically in not A. OK? That's the part that's in U but not in A. Similarly with B. Similarly there's some region, there's some part of the universe where both A and B occur, the intersection of the two occurred.

So what Bayes' theorem says is, if I tell you B occurred, all this part of the universe outside of B is irrelevant. As far as you're concerned, B's the new universe. Notice that if B is the new universe, then the intersection-- which is the part where A occurred-- is bigger after the conditioning then it was before the conditioning. Before the conditioning the universe was this big, now the universe is this big. The universe is smaller, so this region of overlap occupies a greater part of the new universe. Is that clear? So when you condition, you're really making the universe smaller, And the relative likelihood of things that are still in the universe, seem bigger.

So what's the conditional probability of getting a die roll greater than 3, given that it was odd? Calculate, you have 30 seconds. This is three times harder.

OK, what's the probability of getting a die roll greater than 3, given that the die role was odd? Everybody raise your hands. And it's a landslide, the answer is (2). You roughly do the same thing we did before, except now the math is incrementally harder because you have to do a divide. So we think about the same two events, the event that it is odd and the event that it's bigger than 3, and now we ask the question. If it were odd, what's the likelihood that it's greater than 3? Before I did the conditioning, what was the likelihood that it was bigger than 3?

AUDIENCE: 1/6

PROFESSOR: Nope. 1/2. So bigger than 3 is 4, 5, or 6 -- right? There are 3 atomic units there. There are 6 atomic units to start with. They are equally likely. So before I did the conditioning, the event of interest had a probability of a 1/2. After I do the conditioning, I know that half of the possible samples didn't happen. The universe shrank. Instead of having a sample space with 6, I now have a sample space with 3.

Similarly the probability law changed. So now the event of interest is bigger than 3, but bigger than 3 now only happens once. So what I need to do is rescale my probabilities. Remember the scaling rule, one of the fundamental properties of probability. The scaling rule said the sum of the probabilities must be 1. After I've conditioned, the sum of the probabilities is a 1/2. That's not good. I've got to fix it.

So the way to think about Bayes' rule is, if all I know is it the universe got smaller, how should I redo the scaling? Well if all I've told you is that the answer is odd, then there are three possibilities. Before I told you that the answer was odd, they were equally likely. After I tell you that they're odd, has it changed the fact that they're equally likely? No. They're still equally likely even under that new condition. I haven't changed their individual probabilities. So they started out equally likely, they're still equally likely, they just don't sum to 1 anymore. Bayes' rule says, make them sum to 1.

OK, so the way I make this sum, sum to one is to divide by 1/2. If you divide six by 1/2, you get 1/3. Notice that the probability that it's bigger than 3 went from 1/2 to a

1/3. It got smaller. It could have gone either way. So, think about what happens when the world shrinks, when the universe gets smaller, when I tell you that B happened. Well when I tell you that B happened, then I ask you whether A happened, here I'm showing a picture that in the original universe A and B sort of covered the same amount of area. By which I mean, they're about equally likely.

Before I did the conditioning, the probability of A was about the same size as the probability of B. What happens when I condition? Well, when I condition now the universe is B. But notice the way I've drawn them, there's very little overlap. So now when I condition on B, the odds that I'm in A seems to have got smaller. Rather than being of equal probability, as I show here, after the conditioning the relative likelihood of being event A is smaller than it used to be. But that's entirely because of the way I rigged the circles. I could have rigged the circles to have a large amount of overlap. Then when I condition, it seems as though it's relatively more likely that I'm in the event A. That's what we mean by the conditioning.

The conditioning can give you un-intuitive insight. Because when you condition, probabilities can get bigger or littler. And that's something that sort of at a gut level, we all have trouble dealing with.

OK, so that's the fundamental ideas, right? We've talked about events. Three axioms of probability that are completely trivial. One, not quite so trivial rule, which is Bayes' rule. In order to apply it, there's two more things we need to talk about. The first is, notation. We could do the entire rest of the course using the notation that I showed so far, drawing circles on the blackboard, it would work. It would not be very convenient.

So to better take advantage of math, which is a very concise way to write things down, we will define a new notion which is a random variable. Random variable is just like a variable, except shockingly, it's random. So where we would normally think about a variable represents a number, a random variable represents a distribution. So we could, for example in the die rolling case, we could say the sample space has 6 atomic events, and I could think about it as 6 circles. Circles

wouldn't pack all that well. 6 squares inside the universe, right? Because they are mutually exclusive, and collectively exhaustive, so if I started with a universal that looked like that, I would have this one would be the probability that the number of dots was 1, 2, 3, it has to fill up by the time I've put 6 of them in there. And they have to not overlap.

A more convenient notation is to say, OK, let's let X represent that outcome. So I can label the events with math. I can say, there's the event X equals 1, the event X equals 2, the event X equals 3, and it just makes it much easier to write down the possibilities, then to try to draw pictures with Venn diagrams all the time. So all we're doing here is introducing a mathematical representation for the same thing we talked about before. But among the things that you can do, after you've formalized this, so you can have a random variable then it's a very small jump to say you can have a multi-dimensional random variable.

Let's just for example have a 2-space. X and Y, for example. So now we can talk very conveniently about situations that factor. So, for example when I think about flipping 3 coins, I can think about that as a multivariate random variable in three dimensions. One dimension represents the outcome of the first die-- the first coin toss. Another dimension is the second, the third dimension is the third. So there is a very convenient way of talking about it, and we have a more concise notation. We say, OK let V be the outcome of the first die roll, or whatever. Let W be the second one, and then we can think about the joint probability distribution, in terms of the multi-dimensional random variable.

So we have the random variable defined by V and W. We will generally to try to make things easy for you to know what we're trying to talk about, we'll try to remember to capitalize things when we're talking about random variables, and then we'll use the small numbers to talk about events. So this notation would represent the probability that V took on the value little v, and W took on the value little w. We'll see examples of this in a minute. So the idea is-- you don't need to do this, it's just a convenient notation to write more complicated things concisely.

Now a concept that's very easy to talk about, now we have random variables, is reducing dimensionality. And in fact, we will constantly reduce dimensionality of complicated problems that are represented by multiple dimensions, to smaller dimensional problems. And we'll talk about two ways of doing that. The first is what we will call marginalizing. Marginalizing means, I don't care what happened in the other dimensions. So if I have a probability rule that told me, for example, about the outcome of one toss a fair die, and a second toss of a fair die, and if I tell you the joint probability space for that, right? So I would have 6 outcomes on one dimension, 6 outcomes on another dimension, let's say they're all equally likely. I have 36 points altogether, if they're all equally likely, then my probability law is a joint distribution. The joint distribution has 32 non-zero points and each point has height of. I said the right thing, right. 36 is what I meant to say. My brain is telling me that I might not have said that. I meant 36. So if I have 36 equally likely events, how high is each one? 1/36.

OK, so the joint probability space for two tosses of a fair 6-sided die, is this 6-by-6 space. And I may be interested in marginalizing. Marginalizing would mean, I don't care what the second one was. OK well, how do you infer the rule for the first one from the joint, if I don't care what the second one was, well you sum out the second.

So if I have this 2-space that represented the first and the second. So, say its X and Y, for example. So, I've got 6 points that represent 1, 2, 3, 4, 5, 6. And then 6 this way, that sort of thing, except now I have to draw in tediously all of the others, right? So you get the idea. Each one of the X's represents a point with probability 1/36, and imagine direction that they're all in straight lines. Now if I didn't care what is the second one, how would I find the rule for the first one, well I just sum over the second one. So, say I'm only interested in what happened in the first one, well I would describe all of the probabilities here to that point. I would sum out the one that I don't care about. That's obvious, right? Because if I marginalized these X's that all represent the number 1/36 have to turn into a single dimension axis, which is just X, and they have to be 6 numbers that are each how high? 1/6, right?

So the way I get 6 numbers that are each 1/6, when I started with 36 numbers that

were each 1/36 is use sum. OK, so that's called marginalization.

The other thing that I can do is condition. I can tell you something about the sample space and ask you to figure out a conditional probability. So I might tell you what's the probability rule for Y conditioned on the first one being 3? OK. Mathematically that's a different problem, that's a re-scale problem, because that's Bayes' rule.

So generally if I carved out by conditioning some fraction of the sample space, the way you would compute the new probabilities would be to re-scale. So there's two operations that we will do. We will marginalize, which means summing out. And we will condition, which means re-scale. OK.

So give some practice at that, let's think about a tangible problem. Example, prevalence and testing for AIDS. Consider the effectiveness of a test for AIDS. This is real data. Data from the United States. So imagine that we take a population, representative of the population in the United States, and classify every individual as having AIDS or not, and being diagnosed according to some test as positive or negative. OK, two dimensional. The two dimensions are what was the value of AIDS, true or false? And what's the value of the test, positive or negative? So we've divided the population into four pieces. And by using the idea of relative frequency, I've written probabilities here.

So what's the probability of choosing by random choice an individual that has AIDS and tested positive. OK, so that's 0,003648, et cetera. So I've divided the population into four groups. Multidimensional, multidimensional random variable. OK. The question is, what's the probability that the test is positive given that the subject has AIDS? I want to know how good the test is. So the first question I'm going to ask is, given that the person has AIDS what's the probability that the test gives a true answer? You've got 60 seconds. This is harder. Some people don't think it's harder.

So what's the probability that the test is positive, given that the subject has AIDS? Is it bigger than 90%? Between 50% and 90%? Less than 50%? Or you can't tell from the data? Everybody vote, and the answer is 100% correct. Wonderful. So let me make it harder. Is it between 90% and 95%? Or between 95% and a 100%?

AUDIENCE: 95% and a 100%

PROFESSOR: 95%. Is it between 95% and 97%, or 97% and 100%? OK, sorry. This is called marginalization. I told you something about the population that lets you eliminate some of the numbers. So if I told you that the person has AIDS, then I know I'm in the first column. That's marginalization. I gave you new information. I'm saying the other cases didn't happen. I've shrunk the universe, it used to have 4 groups of people, now it has 2 groups of people, I used Bayes' rule. I need to re-scale the numbers so that they add to 1. So these 2 numbers, the only 2 possibilities that can occur-- after I've done the conditioning, no longer add to 1. I've got to make them add to 1. I do that by dividing by the probability of the event that I'm using to normalize.

So the sum of these two probabilities is something, whatever it is 0.003700. So I divide each of those probabilities by that sum, that's just Bayes' rule. And I find out that the answer is the probability that the test is positive-- given that person has AIDS, the probability that the test is positive is 0.986. Good test? Good test? 98%. I won't say that. 98%. is a good test right? Not that today is an appropriate day to talk about the outcomes of tests and 98%. But, I won't mention that. OK, so good test. The accuracy of the test is greater than 98%. Quite good.

New question. What's the probability that the subject has AIDS given that the test is positive?

Everybody vote. (1), (2), (3), (4). Looks 100%. OK, the answer is less than 50%. Why is that? Well that's another marginalization problem, but now we're marginalizing on a different population. This is how you can go awry thinking about probability. The 2 numbers seem kind of contradictory. Here I'm saying that the test came out positive and I'm asking does the subject have AIDS. It's still marginalization. I'm still throwing away 2 of the conditions, two fractions of the population, I'm only thinking about 2. I still have to normalize so that the sums come out 1, but the numbers are different. Yes?

AUDIENCE: [INAUDIBLE PHRASE].

PROFESSOR: Thank you. Because my brain's not working. OK, I've been saying marginalization and I meant uniformly, over the last five minutes, to be saying conditioning. OK, so I skipped breakfast this morning, my blood sugar is low, sorry. Thank you very much. I should have been saying conditioning. Sorry. OK, so backing up. OK I conditioned on the fact that the person had AIDS, and then I conditioned on the fact that the test came up positive. In both cases I was conditioning. In both cases I was doing Bayes' rule.

Please ignore the person who can't connect his brain to his mouth. So, here because the conditioning event has a very different set of numbers from these numbers, the relative likelihood that the subject has AIDS is small. So even though the test is very effective in identifying cases that are known to be true, it is not very effective in taking a random person from the population and saying the test was positive, you have it. OK, those are very different things and the probability theory gives us a way to say exactly how different those are. Why are they so different?

The reason they're different is that other word. Because the marginal probabilities are so different. And that is because the population is skewed. So the fact that the test came out positive, is offset at least somewhat by the skew in the population. So the point here is actually marginalizing. If I think about how many people in the population have AIDS, that means I'm summing on the columns, rather than conditioning. And what you see is a very skewed population. And that's the reason you can't conclude from the test, whether or not this particular subject has the disease or not because the population is so skewed. So this was intended to be an example of conditioning versus marginalization and how you think about that in a multi-dimensional random variable. Yes?

AUDIENCE: Don't you sum [UNINTELLIGIBLE] in order to do Bayes' rule?

PROFESSOR: In order to condition on has AIDS, you need to sum has AIDS. And then you use that number. Yes? That's right.

AUDIENCE: So how are they different?

PROFESSOR: One of them has a [UNINTELLIGIBLE] and the other one doesn't. So when we did Bayes' rule, we did the marginalization here, but then we use that summed number to normalize the individual probabilities by scaling, by dividing. So that the new sum, over the new smaller sample space is still one. So your point 's right. So regardless of whether we're conditioning or marginalizing, we still end up computing the marginals. it's just that in one case were done, and in the other case we use that marginal to re-scale OK?

So I said, we could just use set theory and we're done. We'll in fact use random variables because it's simpler. That's one of the two other things we need to do which are non-essential, it just makes our life easier. And the other non-essential thing that we will do is represent it in some sort of a Python structure. So we would like to be able to conveniently represent probabilities in Python. The way we'll do that, is a little obscure the first time you look at it. But again, once you've done it a few times it's a very natural way of doing it, otherwise we wouldn't do it this way. How are we going to represent probability laws in Python? The way we'll do it, since the labels for random variables can be lots of different things-- so for example, the label in the previous one was in the case of the subject having AIDS or not, the label was true or false. The label for the test was positive or negative. So in order to allow you to give symbolic and human meaningful names to events we will use a dictionary as the fundamental way of associating probabilities with events.

So, we'll represent a probability distribution by a class-- what a surprise, by a Python class-- that we will call DDist which means discrete distribution. DDists want to associate the name of an atomic event which we will let you use any string, or in fact any-- I should generalize that. You can use any Python data structure to identify an atomic event. And then we will associate that using a Python dictionary, with the probability. So what we will do when you instantiate a new discrete distribution, you will-- the instantiation rule, you must call it with a dictionary. A dictionary is a thing in Python that associates one thing with another thing, I'll give an example in a minute. And the utility of this is that you'll be able to use as your atomic event a string, like

true or false, a string like positive or negative, or something more complicated like a tuple. And I'll show you an example of where you would want to do that in just a second.

So the idea is going to be you establish a discrete distribution by the unique method called the dictionary. The dictionary is just a list of keys which tell you which event that you're trying to name the probability of. Associated with a number, and that number is the probability. And this shows you that there's one extremely interesting method, which is the Prob method. The idea is that Prob will tell you what is the probability associated with that key. If it doesn't find the key in the dictionary, I'll tell you the answer is 0.

We do that for a specific reason too, because a lot of the probability spaces that we will talk about, have lots of 0's in them. So instead of having to enumerate all of the cases that are 0 we will assume that if you didn't tell us a probability, the answer was 0.

OK so this is the idea. I could say use the disk module in lib 601 to create the outcome of a coin toss experiment. And I have a syntax error. This should have had a squiggle brace. A dictionary is something that in Python-- So I should have said something like this-- dist.DDist of squiggle. Sorry about that, that should've said squiggle, I'll fix it and put the answer on the website. Head should be associated with the probability 0.5 and tail should be associated with the probability 0.5. End of dictionary, end of call. Sorry, I missed the squiggle. Actually what happened was, I put the squiggle in and LaTeX ate it. Because that's the LaTeX, anyway. It's sort of my fault. The dog ate my homework. LaTeX ate my squiggle, it's sort of the same thing.

So having defined a distribution, then I can ask what's the probability of the event head? The answer is a half. The probability of event tail? The answer is a half. The probability of event H? There is no H. The answer 0. That's what I meant by sparsity. If I didn't tell you what the probability is, we assume the answer is 0.

Conditional probabilities are a little more obscure. What's the conditional probability

that the test gives me some outcome given that I tell you the status of whether the patient has, or doesn't have AIDS? OK, well conditionals-- you're going to have to tell me which case I want to condition on. So in order for me to tell you the right probability law you have to tell me does the person have AIDS or not. So that becomes an argument. So we're going to represent conditional probabilities as procedures. That's a little weird. So the input to the procedure, specifies the condition.

So if I want to call the procedure and find out what's the distribution for the tests, given that the person has AIDS? Then I would call, test given AIDS of true. So if AIDS is true, return this DDist, otherwise return this DDist. So it's a little bizarre but think about what it has to do. If I want to specify a conditional probability, I have to tell you an answer. And that's what the parameter is for. So the way that would work is illustrated here having defined this as the conditional distribution I could call it by saying what is the distribution on tests given that AIDS was true? And the answer to that is the DDist. Or if I had that DDist, which would be this phrase, I could say what's then the probability in that new distribution that the answer is negative? Then I would look up the dot prob method within the resulting conditional distribution, and look up the condition negative.

And finally the way that I would think about a joint probability distribution, is to use a tuple. Joint probability distributions are multi-dimensional, tuples are multi-dimensional. So for example, if I wanted to represent this multi-dimensional data, I might have the joint distribution of AIDS and tests. OK that's a 2-by-2. AIDS can take on 2 different values, true or false. And tests can take on 2 different values, positive or negative. So there's 4 cases.

The way I would specify a joint distribution would be create a joint distribution starting with the marginal distribution for AIDS and then using Bayes' rule tell me the two different conditional probabilities given AIDS. And that then will create a new joint distribution that whose DDist is a tuple. So in this new joint distribution, AIDS and tests, if AIDS is false, and test is negative-- so false negative is this number-- the probability associated with tuple is that number. Is that clear?

So I'm going to construct joint distributions by thinking about conditional probabilities. So I have a simple distributions which are defined with dictionaries. I have conditional probabilities which are defined by procedures. And I have joint probabilities which are defined by tuples. OK, so that's the Python magic that we will use and a lot of the exercises for Week 10 have to do with getting that nomenclature straight. It's a little confusing at first, I assure you that by the time you've practiced with it, it is a reasonable notation. It just takes a little bit of practice to get onto it, much like other notations.

OK where are we going with this? What we would like to do is solve that problem that I showed at the beginning of the hour. So we would like to know things like, where am I? So the kind of thing that we're going to do is think about where am I based on my current velocity and where I think I am, odometry-- which is uncertain, it's unreliable-- versus for example where I think I am based on noisy sensors. OK so that's like two independent noisy things. Right? The odometry you can't completely rely on it. You've probably run into that by now. The sonars are not completely reliable. So there are two kinds of noisy things. How do you optimally combine them? That's where we're heading. So the idea is going to be here I am, I think I'm a robot, I think I'm heading toward a wall, I'd like to know where am I.

So the kinds of data that we're going to look at are things like, I think I know where I started out. Now my thinking could be pretty vague. It could be, I have no clue so I'm going to assume that I'm equally likely anywhere in space. So I have a small probability of being many places. That just means that my initial distribution is very broad. But then I will define where I think I am by taking into account where I think I will be after my next step. So I think I'm moving at some speed. If I were here, and if I'm going at some speed I'll be there. So we will formalize that by thinking about a transition. I think that if I am here at time T, I will be there at time T plus 1. And I'll also think about, what do I think the sonars should've told me. If I think I'm here, what would the sonars have said? If I think I'm here, what would the sonars have said? And we'll use those as a way to work backwards in probability, use Bayes' rule. To say, I have a noisy idea about where I will be if I started there. I have a

noisy idea of what the sonars would have said, if I started there. But I don't know where I started. Where did I start? That's the way we're going to use the probability theory.

So for example, if I thought I was here and if I thought I was going ahead 2 units in space per unit in time, I would think that the next time I'm here. But since I'm not quite sure where I was maybe I'll be there, and maybe I'll be there, but there's very little chance that I'll be there. That's what I mean by a transition model. It's a probabilistic way of describing the difference between where I start and where I finish in one step.

Similarly, we'll think about an observation model. If I think I'm here, what do I think the sonars would have said. Well I think I've got some distribution that it's very likely that they'll give me the right answer, but it might be a little short it might be a long. Maybe it'll make a bigger error. So I'll think about two things. Where do I think I will be based on how I'm going? And where do I think I'll be based on my observations? And then we'll try to formalize that into a structure that gives me a better idea of where I am.

That's the point of the exercises next week when we won't have a lecture. So this week we're going to learn how to do some very simple ideas with modelling probabilities. With thinking about these kinds of distributions. And the idea next week then is going to be incorporating it into a structure that will let us figure out where the robot is in some sort of an optimal sense.

So thinking about optimal -- let's come back to the original question. How much would you pay me to play the game? OK, we had some votes. They didn't add up to 1. What should I do to make them add up to 1? Divide by the sum. Right? Look at all of you know already, right? So you now know all this great probability theory.

So the question is can we use probability theory to come up with a rational way of thinking how much it's worth? Most of you thought that it's worth less than 10$. OK, so how do we think about this? How do we use the theory that we just generated to come up with a rational decision about how much that's worth? OK, thinking about

the bet quantitatively, what we're going to try to do is think about it with probability theory.

There are 5 possibilities inside the bag. Originally there could have been 4 white, or 3 white and 1 red, or 2 and 2, or 1 and 3, or 0 and 4. That was the original case. You didn't know. I didn't know. They were thrown into the bag over here. We didn't know. How much would that game-- how much should you be willing to pay to play that game? Someone asked how many white ones and how many red ones did the person put in the bag? I don't have a clue, right? We need a model for the person.

Since I don't have a clue, one very common strategy is to say all these things I know nothing about let's just assume they're all equally likely. So that's called maximum likelihood, when you do that. There's other possible strategies. I'll use the maximum likelihood idea just because it's easy. So I have no idea. Let's just assume that here's all of the conditions that could have happened. The number of red that are in the bag could have been 0, 1, 2, 3, or 4. I have no idea how the person chose the number of LEGO parts. So I'll assume that each of those cases is 1/5 likely, since there's 5 cases.

OK now I'll think about what's my expected value of the amount of money that I'll make if the random variable S, which is the number of red things that are in the bag was s which is either 0, 1, 2, 3, or 4. OK, if there are 0, how much money do you expect to make? None. If there are 4 reds, how much money would you expect to make? $20 If there are 2 reds, you would expect to make 10 $. Everybody see that.

I'm trying to think through a logical sequence of steps for thinking about how much is it worth to play the game. So this is the amount of money that you would expect given that the number of red in the bag, which you don't know, were 0, 1, 2, 3, or 4. That's this row. What's the probability, what's the expected value of the amount of money you would get., and that happens? Well I have to use Bayes' rule. What I need to do is I have to take this probability times that amount to get that dollar value. So over here, in the event that there are 4 reds in the bag, I'm expecting to get $20 but that's only 1/5 likely. Right? Because there don't have to be 4 reds in

the bag. So I multiply the 1/5 times the $20, and I get 4$.

So my expected outcome for this trial is 4$. Here, I'm expecting to make 10$ if I knew that there was 2 reds in the bag. But I don't know that there's 2 reds in the bag, there's a 1/5 probability there's 2 reds in the bag. So 1/5 of my expected amount of money which is 10$ is 2$.

So then in order to figure out my expected amount money I just add these all up, marginalizing. And I get the [UNINTELLIGIBLE] 4 plus 3 is 7 plus 2 is 9 plus 1 is 10. So this theory says it if I can regard the person who put the LEGOs in the bag as being completely random, I should expect to make 10$ on the experiment.

So that means you should be willing to pay 10$. Because on average, you'll get back 10$. If you wanted to make a profit you ought to be willing to pay 9$. Right? Because then you would pay 9$ expecting to get 10$. If you really would like to make a loss, right? Then you should pay $11. Yeah?

AUDIENCE: Why do we assume that these events are equally likely?

PROFESSOR: Completely arbitrary. So there's theories, more advanced theories, for how you would make that choice. So for example if in your head you thought that the person just took a large collection of LEGO parts and reached in, then you would think that the number of red and white might depend on the number that started out in the bin. But I don't think that's probably true, right. The person was probably looking at them and saying, oh throw in one red, through in one white. So you need a theory for doing that, and I'm saying that in the absence of any other information let me assume that those are equally likely and see what the consequence of that would be. The consequence of assuming that is that I should expect to get 10 $ back.

What happens if you pull out a red? As we did. How does that affect things? Well it increases the bottom line. I start out again with the assumption that all 5 cases are equally likely. Now I have to ask the case, how likely is it that the one that we pulled out was red? Well it's not very likely that the one that I pulled out was red, if they were all white. The probability of that happening is 0. What's the probability if there

were 2 that the person pulled out a red? Well 2 of them were red, 2 of them were white, 2 out of 4 cases would have showed this case of pulling out a red. So this line then tells me how likely is it that the red was pulled. OK. Then what I want to do is think about what's the probability that I pulled out a red, and there was 0, 1, 2, 3, or 4. So I multiply 1/5 times 0/40, 0/20, 1/5 times 1/4 to get 1/20, 1/5 times 2/4 you get 2/20.

So those are probabilities of each individual event happening. But they don't sum to 1. So then the next step I have to make them sum to 1. So the sum of these is a 1/2. So I make them sum to one this way. So now what's happened is it's relatively more likely 4 out of 10, that this case happened, than that case. I know for sure, for example, that there's not 4 whites. The probability of 4 whites is 0-- 0 out of 10

So what I've done is I've skewed the distribution toward more red by learning that there's at least 1, I now know that I know additional information. These were not equally likely. In fact, the ones with more red were relatively more likely. So if I compute this probability times that expected amount, I now get a much bigger answer for the high number of reds. So I still get 0 just like I did before for this case, because there's no reds in the bag. But now it's much more likely that they're all red, because I know there was at least 1 red. And then the answer comes out $15.

So my overall assessment, don't go to Vegas. You could have made a lot more money by offering $13. Because on average, you should've expected to make $15. OK, so what I wanted to do by this example is go through a specific example of how you can speak quantitatively about things that are uncertain. And that's the theme for the rest of the course.

6.01SC Introduction to Electrical Engineering and Computer Science

Spring 2011