MITOCW | Lec-11

PATRICK WINSTON: Ladies and gentlemen, the Romanian national anthem.

I did not ask you to stand, because I didn't play it as a symbol of Romanian national identity.

But rather, to celebrate the end of the Cold War, which occurred about the time that you were born.

Before that, no one came to MIT from Eastern Europe.

But since that time, we've been blessed by having in our midst Lithuanians, Estonians, Poles, Czecs, Slovaks, Bulgarians, Romanians, Slovenians, Serbs, and all sorts of people from regions of the world formally excluded to us.

Believe me, you are all welcome in our house.

Almost all, that is to say.

Because you may recall that Romania is the traditional home of vampires.

And since the end of the Cold War, vampires have had new vectors for emerging from their traditional places and penetrating into the world at large.

You may have vampire in your suite, or on your floor.

And it's important to know how to recognize them, and take the necessary precautions.

So if you have this concern, I would expect that the first thing you would do would be to look at some data concerning the characteristics of vampires.

So there's a little database of samples of individuals who have been determined to be vampires and not vampires.

And our task today-- and what you'll understand how to do by the end of the hour-- is to use data like this to build a recognition mechanism that would help you to identify whether someone is a vampire or an ordinary person.

So this is a little different from the kind of problem we worked with neural nets.

Right?

So what's the most conspicuous difference between this data set and anything you could think to work on with nearest neighbors, which we studied last time.

Katie, do you have any thoughts about why it would be difficult to use nearest neighbors with data like this?

The question mark is there because this is MIT, and a lot of people are completely nocturnal.

So you can't tell whether they cast a shadow or not.

We want to take that into account.

So what's different about this from the electrical cover data set?

STUDENT: [INAUDIBLE] PATRICK WINSTON: Could you use the nearest neighbor technique to identify vampires with this data?

STUDENT: [INAUDIBLE] PATRICK WINSTON: So obviously-- Yes, Lana?

STUDENT: [INAUDIBLE] STUDENT: You cannot really quantify-- PATRICK WINSTON: Oh, that's the problem.

This is not numerical data.

This is symbolic.

So we're not saying that your ability to cast a shadow is 0.7.

You either cast a shadow, down cast a shadow, or we can't tell.

It's a symbolic result.

So problem number one we have to face with data of this kind is that it's not numeric.

And there are other characteristics, as well.

For example, it's not clear that all of these characteristics actually matter.

So some characteristics don't matter.

And a corollary to that is that some characteristics do matter, but they only matter part of the time.

And finally, there's the matter of cost.

Some of these tests may be more expensive to perform than others.

For example, if you wanted to determine whether someone casts a shadow, you'd have to go to the trouble of getting up during daylight.

That might be an expensive operation for you.

You'd have to go find some garlic and ask them to eat it.

That might be expensive.

So some of these tests might be expensive relative to other tests.

But once you realize that we are talking in terms of tests, and not a vector of real values, then what you do is clear.

You build yourself a little tree of tests.

So who knows how this problem will turn out?

But you can imagine a situation where you have one test up here which might have three outcomes.

And one but only one of those outcomes might require you to perform another test.

And only when you've created the tree of tests that look like this are you finished.

So given this set of tests and a set of samples, the question becomes, how do you arrange the tests in a tree like that so as to do the identification that you want to do?

So since we're talking about identification, it's not surprising that this kind of tree is called an identification tree.

And there's a tendency-- and I may slip into it myself-- to call this a decision tree.

But a decision tree is a label for something else.

This is an identification tree.

And the task is to create a good one.

So what is a good one versus a not so good one?

What characteristic would you like for a decision tree-- for an identification trade to have, if you're going to call it good identification tree?

What do you think, Krishna?

What would be a good characteristic?

STUDENT: Maybe the minimum number of levels?

PATRICK WINSTON: Yeah.

He said minimum number of levels.

What's another way you could say what a good one is?

Each test costs something, right?

So what's another way of thinking about what a good tree would look like?

STUDENT: Minimum cost.

PATRICK WINSTON: The minimum cost.

And if they all have the same cost, then it's the number of tests.

So overall, what you like is a small tree rather than a big one.

So you might be able to take your sample data and divide it up, so that at the bottom of the tree, at the leaves, all of the sets that are produced by the tests are uniform, homogeneous.

We'd like that tree to be the simplest possible tree you can find, not some big complicated one that also divides up all the data into uniform subsets.

By uniform subset-- at the bottom of the tree, you have all of the vampires together, and all the non-vampires together.

So you'd like a small tree.

So why not just go all the way and do British Museum, and calculate all possible trees?

Well, you can do that, but it's one of those NP problems.

And as you know, NP problems suck in general.

And so you don't want to do that.

You want to have some kind of heuristic mechanism for building a small tree.

And we want a small tree because-- Why do we want a small tree?

Because of the cost.

but there's another, more important reason why we want a small tree.

Let me give you a hint.

It's Occam's Razor.

The simplest explanation is often the best explanation.

So if you have a big, complicated explanation, that's probably less good than a simple, small explanation.

Occam's Razor.

Spelled so many ways it doesn't matter how I spell it.

And that's good, because I can't spell.

So how are we going to go about finding the best possible arrangement of those four tests in a tree like that?

Well, step one will be to see what each test does with the data.

And by the way, before I go a step further, you know and I know that this is a sample data set that's very small, suitable for classroom manipulation.

You'd never bet your life on a data set this small.

We use it only for classroom illustration.

But imagine that these rows are multiplied by 10.

So instead of eight samples, you've got 80.

Then you might begin to believe the results that are produced.

So I'm just going to pretend that each one of those represents 10 other samples that I haven't bothered to show.

But we can work with this one in the classroom, because it's pretty small.

And we can say, well, what does this shadow test do?

Well, the shadow test divides the sample population into three groups.

There's the I Don't Know group of people who are nocturnal.

There are the people who do cast the shadow, the Yes people.

And the people who do not cast a shadow, the No people.

So if I look at those rows up there and see which ones are vampires, it looks to me that if there's no shadow cast-- there's only one that doesn't cast a shadow-- and that is a vampire.

So that's a plus over there.

Vampire.

Now, if we look at the ones who do cast a shadow, all those are not vampires.

They're all OK.

And now there're 8.

Three are vampires.

So that means that two of these must be vampires.

And I've got three, four, five, six so far.

So there must be two left.

So that's the way the shadow test divides up the data.

Now let's do garlic.

Vampires traditionally don't eat garlic.

I don't know why.

So we look at the garlic test, and we see that all of the Nos-- well, there're three Yeses, and they all produce a No answer.

So if somebody eats garlic, they're not vampires.

That means the three vampires must be over here.

Then there are two left.

So that's what the garlic test does.

See what we're trying to do?

We're trying to look at all these tests to see which one we like best on the basis of how it divides up the data.

So now we've got complexion.

And there are three choices for this.

You can have an average complexion.

But a lot of vampires, in my experience, are rather pale.

So pale is a possibility.

And then the other option is that just after gorging themselves with blood, they tend to get a little red in the face.

So we'll have a ruddy over here.

Once again, we have to go back to our data set to see how this test divides things up.

So there are three ruddies, and one's a No, one's a No, and one's a Yes.

So two Nos and a Yes.

Two Nos and a Yes.

Now we can try for pale complexion people.

There are only two of those.

A No and a No.

That must mean that there are two pluses over here, because there are three vampires altogether.

Two, four, six, seven, eight, nine.

Eight, sorry.

Eight.

Only eight.

Just one more to go, and that's the accent.

Historically, vampires go to great length to protect their accent and not betray their origins.

But nevertheless, we can expect that if they've just arrived-- if they're just in from Transylvania, part of Romania-- they may still have an accent.

So there's a normal, some still have a heavy accent, and some persist in having odd accents.

So let's see.

Accent.

Four of them, right at the top, have no accent.

Two Nos and a Yes.

Heavy accent.

Three of those.

A Yes and two Nos.

That means we must have a plus here.

3, 6, plus and a minus.

So we can look at this data and say, well, what will be the best test to use?

And the best test to use would surely be the one that produces sets here, at the bottom of the branches, that correspond to the outcomes of the test.

We're looking for a test that produces homogeneous groups.

So just for the sake of illustration, I'm going to suppose that we're going to judge the quality of the test by how many sample individuals it put into a homogeneous set.

So ideally, we'd like a test that will put all the vampires in one group and all the ordinary people in another group right off the bat.

But there are no such tests.

But we can add up the number of sample individuals who are put in to at least homogeneous sets.

So when we do that, this guy has 3 in a homogeneous set here.

A fourth.

But these are not a homogeneous set.

So the overall score for this guy will be 4.

This one, well, not quite as good.

It only puts 3 individuals in a homogeneous set.

This one here, 2 individuals into a homogeneous set.

Everybody else is all mixed up with some other kind of person.

And over here, how many samples are in a homogeneous set?

0.

So on the basis of this analysis, you would conclude that the ordering of the test with respect to their quality is left to right.

So the best test must be the shadow test.

So let's pick the shadow test first, see what we can do with that.

If we pick the shadow test first, then we have this arrangement.

We have question mark, and we have Yes, casts a shadow, and No, doesn't.

We have 3 minuses here.

We have a plus here.

And unfortunately, over here, we have plus, plus, minus, minus.

So we need another test to divide that group up.

Yes.

STUDENT: How did you get the 4 on the shadow test again?

Why was it 4?

PATRICK WINSTON: Well, if I look at the data and I see who-- the question is, what about that shadow test?

If you look at the shadow test, and you say, well, there are 4 question marks.

And if we look and see what kind of people belong to those 4 question marks, there are 2 vampires and 2 non-vampires.

That's why it's 2 pluses and 2 minuses.

STUDENT: No, I understand that.

The question is, how did you get to the score of 4?

PATRICK WINSTON: Oh, yeah.

The question is how did I get this number 4?

It has nothing to do this, because this is a mixed set.

In fact, I've got three guys in a homogeneous set here, and one guy in a homogeneous set here, and I'm just adding them up.

STUDENT: OK.

PATRICK WINSTON: So very simple classroom illustration.

Wouldn't work in practice.

Yes.

STUDENT: How do you adjust this for larger data sets where it's unlikely you're going to have any [INAUDIBLE]?

PATRICK WINSTON: The question is, how do I adjust this for larger data sets?

You're one step ahead.

Trust me, I'll be doing large data sets in a moment.

I just want to get the idea across.

And I don't want there to be any thought that the method we use for larger data sets has got anything magic about it.

OK, so we're off and running.

And now we have to pick a test that will divide those four guys up.

So we're going to have to work this a little harder, and repeat the analysis we did there.

But at least it'll be simpler, because now we're only considering 4 samples, not 8.

Just the 4 samples that we still have to divide up that have come down that left branch.

So I have the shadow test.

It has 3 outcomes.

We have the garlic test.

It has 2 outcomes.

Yes and No.

We have the complexion test.

There's 3 outcomes.

Average, pale, and ruddy.

And we have finally the accent test.

And that comes out to be either normal, heavy, or odd.

And now, it's a little awkward to figure out what the results are for this data set as shown.

So let me just strike out.

The ones that we're no longer concerned with, and limit our analysis to the samples for which the outcome of the shadow test is a question mark.

This is exactly the four people we still need to separate, right?

So switching colors, keeping the color the same.

We actually don't want to do the shadow test anymore, right?

Because we've already done that.

There's no point in doing that again.

We don't have to look at that.

It's already done all the division of data that it can.

So the garlic test.

Well, let's see.

Garlic.

2 Yeses, 2 Nos.

The Yeses produce Nos and the Nos produce Yeses.

So if the person does eat garlic, they're OK.

And if they don't eat garlic, bad news-- they're vampires.

Well, that looks like a pretty good test.

But just for the sake of working it all out, let's try the others.

Complexion.

2 Ruddies, a Yes, and a No.

1 pale, and that's a No.

1 pale, and that's a No.

And we must have 1 average, and sure enough, that's a Yes.

Now we can do accent, the one on the far right, and look at how that measures up against the people who are still under consideration as samples.

Accent.

Let's see.

2 Nones, a Yes and a No.

No Heavies.

2 Odds, a Yes and a No.

All right.

So now we can do the same thing we did before, and just say, for sake of classroom illustration, how many individuals are put into a homogeneous sets.

And here we have 4.

And here we have 2.

And here we have 0.

So plainly, the garlic test is the test of choice.

So we go back over here, and we've completed the work that we needed to do.

So that's the garlic test.

And that produces 2 pluses.

Let's see.

Eats garlic, Yes.

Eats garlic, No.

I guess the pluses go over here like so.

And these are the two ordinary people.

And we're done with our task.

And now you can quickly run off and put this into your PDA, and forever be protected against the possibility that one of those vampires got out in the flood of people that came in from Eastern Europe.

Except what do we do a large data set?

Well, the trouble is, a large data set's not likely to produce-- if you have a large data set, no test is likely to put together any homogeneous set right off.

So you never get started.

Everything would be 0.

Every test would say, oh it doesn't put anybody into homogeneous sets.

So you're screwed.

You need some other, more sophisticated way of measuring how disordered this data is.

Or how disordered these sets are that you find at the bottom of the tree branches.

That's what you need.

You need a way of measuring disorder of these sets that you find at the bottom of these branches, so you can find a kind of overall quality to the test based on your measurement of disorder.

Now, the first heuristic of a good life is, when you have a problem to solve, ask somebody who knows the answer.

It's the least amount of work.

It's not even as hard going to Google.

So who would you ask about ways of measuring disorder in sets?

There are two possible answers.

STUDENT: You could just do entropy.

PATRICK WINSTON: What?

STUDENT: Find the entropy of the set.

PATRICK WINSTON: Who studies entropy?

STUDENT: Probability.

PATRICK WINSTON: What kind of classes?

STUDENT: Physics.

STUDENT: Thermodynamics.

PATRICK WINSTON: Thermodynamics!

The thermodynamicists are good at measuring disorder, because that's what thermodynamics is all about.

Entropy increasing over time, and all that sort of stuff.

There's another equally good answer.

STUDENT: Statisticians?

PATRICK WINSTON: Statisticians.

Perhaps, but it's not the second best answer.

It's actually not even the best answer.

That's the best answer.

What's your name?

STUDENT: Leo.

PATRICK WINSTON: Oh, yeah.

[LAUGHTER] PATRICK WINSTON: Leonardo has got his finger on it.

The information theorists are pretty good at measuring disorder, because that's what information is all about, too.

So we might as well borrow a mechanism for measuring the disorder of a set from those information theory guys.

So what we're going to do is exactly that.

Let's put it over here, so we'll have it handy when we want to try to measure those things.

The gospel according to information theorists is that the disorder, D, or some set is equal to-- now let's suppose that this is a set of binary values.

So we have positives and then we have negatives.

Pluses and minuses.

But pluses, they don't go very well in an algebraic equation, because they might be confused with adding.

So I'm going to say P and N. And then it'll be the total, which is P plus N. We only have two choices, positive and negative.

So the disorder of set, according those guys, is equal to minus the number of positives over the total number, times the log to the base 2 of the positives over the total, minus the negatives over the total, times the log 2 of the negatives over the total.

Those negatives look a little worrisome, because you think, well, maybe this thing can go negative.

But that's not going to be true, right?

Because these ratios are all less than 1, and the logarithm of something that's less than 1 is negative.

So we're OK.

So that's a lovely way of measuring disorder.

And then we ought to draw a graph of what that curve looks like.

And what we're going to graph it against is the ratio of positives to the total number.

So that's going to be an axis where we go from 0 to 1.

So let's just find a couple of useful values.

And by the way, it pays to pay attention to these curves, because if you pay attention to this stuff, you can work the quiz questions on this very rapidly.

Otherwise, we see people getting out their calculators and quickly becoming both lost and screwed.

OK so let's see.

Let's suppose that the number of positives is equal to the number of negatives.

So we've got a completely mixed-up set.

It has no bias in either direction.

So in that case, if P over T is equal to 1/2, then this is equal to minus 1/2, times the logarithm of 1/2.

And I guess, since they're both the same, we can multiply by two.

And what's that value?

[INAUDIBLE], what does that calculate out to?

STUDENT: Minus [INAUDIBLE] PATRICK WINSTON: Minus [INAUDIBLE].

Well, with a minus sign, you just turn the argument upside down, so it's log(2).

So what's log(2)?

Logarithm of base 2 of 2?

1!

So this whole thing is-- STUDENT: 1.

PATRICK WINSTON: 1.

So [INAUDIBLE], in her soft way, says, well, let's see.

2 times 1/2.

That cancels out.

The minus, that flips the arguments so it's log to the base 2 of 2, and that's 1.

So this whole thing, You work out the algebra, it gives you 1.

So that's cool.

So right here in the middle where they're equal, we get a value of 1.

Next thing we need to do is let's calculate what happens if P over T is equal to 1.

That is to say, everything is a positive.

Any guesses?

Maybe 10, 20, minus 15?

Let's work it out.

So if P over T equal 1, that would be minus 1 times the log to the base 2 of 1.

What's that?

STUDENT: [INAUDIBLE] PATRICK WINSTON: A 0?

Oh, yeah.

Because 2 raise to the 0 is one.

So this part is 0.

Now, what about this other part?

If everything's a P, then nothing's an N.

So we've got 0.

And we can quit already.

Well, not quite.

We ought to work it out.

Log 2 to the base 2 of 0.

What's that?

STUDENT: [INAUDIBLE] PATRICK WINSTON: Who?

Minus infinity?

Uh oh.

0 times minus infinity is What I didn't get that when I was in high school.

Finally, 1801 makes a difference.

Finally.

What's the answer.

We're interested in the limit as N over T goes to 0, right?

And when you have a deal like this, what do you do?

You use that famous rule, that we all mispronounce when we see it written, right?

We use the good old El Hospital's rule.

OK, it's L'Hopital.

L'Hopital's Rule.

You have to differentiate the-- I guess we differentiate this guy as a ratio or something, and see what happens when it goes to 0.

And what we get when we use L'Hopital's Rule is that, oh thank God, this is still zero.

So now we know that we have a point up there and a point down there.

So now we've got three points on the curve, and we can draw it.

It goes like that.

No, it doesn't go like that.

It's obviously a Gaussian, right?

Because everything in a nature is a Gaussian.

Can you put that laptop away, please?

Everything in nature is a Gaussian, so it looks like this.

That right?

No, actually, not everything in nature is a Gaussian.

And in particular, this one isn't a Gaussian either.

It looks more like one of those metal things they used to call quonset huts.

That's what it looks like.

Boom, like so.

So that is the curve of interest.

Now, did God say that using this way of measuring disorder was the best way?

No, Got has not indicated any choice here.

We use this because it's a convenient mechanism, it seems to make sense, but in contrast to the reason it's used information theory, it's not the result of some elegant mathematics.

It's just a borrowing of something that seems to work pretty well.

Any of those curves would work just about the same, because all we're doing with it is measuring how disordered a set is.

So one thing to note here is that in this situation, where we're dealing with two choices-- P and N, positives and negatives-- we get a curve that maxes out at one.

And notice that it kind of gets up there pretty fast.

In fact, if you're down here at 2/3, are you're up here, this is about 0.9.

So it gives you a large number for quite a bit of that area in the middle.

So that, unfortunately, still doesn't tell us everything we need to know.

That tells us how to measure a disorder in one of these sets.

But we want to know how to measure the quality of the test overall.

So we need some mechanism that says, OK, given that this test produces three different sets, and we now have a measure of the disorder in each of these sets, how do we measure the overall quality of the test?

Well, you could just add up the disorder.

Let's write that down, because that sounds good.

So you can say that the quality of a test is equal to some sum over the sets produced.

And what we're going to do is we're going to add up the disorder of each of those sets.

I'm almost home, except that this means we're going to give equal weight to a branch that has almost nothing down it-- we're going to give the same weight to that as a branch that has almost everything going down it.

So that doesn't seem that make sense.

So one final flourish is we're going to weight this sum according to the fraction of the samples that end up down that branch.

So it's, as usual, easier to write it down than to say it.

So we're going to multiply that times the number of samples in the set, divided by the number of samples handled by test.

So if half the samples go down a branch, and if that branch has a certain disorder, then we're going to multiply that disorder times 1/2.

All right.

So now let's see how it works with our sample problem.

Well, here is our sample data.

And we didn't need anything fancy for it.

But let's pretend it was a large data set.

Well, let's see.

What would we do?

Well, go down this way, there are 4 samples down that direction.

That's half of the total number of samples.

So whatever we find down there, we're going to multiply by 1/2.

This one we're going to multiply by 3/8.

And this one we're going to multiply by 1/8.

Now, what do we actually find at the bottom of these things?

Well, here's a homogeneous set.

Everything's the same.

So we go to that curve and say, what is the disorder of a homogeneous set?

It's zero.

Let's see, they're all the same.

I guess that means it's 0 over there.

So the disorder of this set of three samples is zero.

The disorder of this set of one sample, all the same, is zero.

The disorder of this set-- well, let's see.

Half of the samples there are plus, and half are minus, so we go over to our curve, and we say, what's the disorder of something with equal mixture of pluses and minuses?

And that's one.

So the disorder of this guy is one.

So now we've got 1/2 times 1, and 3/8 times 0, 1/8 times 0.

So the quality of this particular test, as determined by the disorder of the sets it produces, is 1/5.

0.5.

Let's do this one.

So we have 3/8 coming down this way, 5/8 coming down this way.

3/8 is multiplied by the disorder of a set of uniform things.

That's disorder 0.

So this guy over here, let's see.

That's 2/5 and 3/5 multiplied-- You know, this is one of those deals where if you look at the curve, you're pretty close to the middle.

And that curve goes all the way up to about 0.9 there.

So you can kind of just look at this, and eyeball it, and say, well, whatever it is, the overall, this is going to be something multiplied times 5/8.

Something like 0.9 times 5/8.

So let's just say, for the sake of discussion, that that's going to be about 0.6, which is within a hundredth, I think, of being right.

Just kind of guessing.

OK, well now we're on a roll.

Here, we have 3/8 coming down this branch, 3/8 coming down this branch, 1/4 coming down this branch.

This is 0.

And this is one of those deals where these two are about 0.9.

So it looks like it's going to be 3/8 plus 3/8 is 3/4.

Times about 0.9.

So that's going to turn out to be about 0.7.

So one last go here.

3/8, 3/8, and 1/4.

Oh, that's interesting.

Because these two are what we got contributed up to that 0.7.

This one is 0.4 times-- this is evenly divided, so that's going to have disorder of 1.

So that's going to be 0.25 bigger than the number we got over here.

So that's going to end up being about 0.95.

So thanks god our answer is the same as we got with our simple classroom measurement of disorder.

Except this is measuring how disordered stuff is, we want the small number, not the big number.

So once again, based on this analysis, you'll be sure to pick the shadow cast, because 0.5 is less than 0.6, which is less than 0.7, which is less than 0.95.

So that accent test is really horrible.

Don't use it.

Just because somebody has a heavy accent doesn't mean they're a vampire.

In fact, most vampires have worked very hard on their accent, as I mentioned before.

All right, so now we know that we're still going to pick the shadow test as our first go.

So that's good.

Now, let's see if we can repeat the exercise with our second selection, the one we have to have to pick those guys apart.

And this is going to be easier, because there are fewer things to work with.

Ooh, wow, look.

That's 0.

That's 0.

That's 1/2.

That's 1/2.

So the disorder of this guy is 0.0.

So this is 1/4, 1/4, 1/2, 0, 0.

1/2 times 1.

Ooh, that's 0.5.

That was easy.

How about this one?

Oh, he says 1.

Let's see.

That's 1.

That's 1.

That's 1/2.

That's 1/2.

Yeah, it is one.

So sure enough, the answer also comes out to be the same as before, when we did our just simple intuition exercise.

So I don't know.

Christopher, is this all about using information theory?

STUDENT: No.

PATRICK WINSTON: No, no, no.

See, it's not about the math.

It's about the intuition.

And the intuition is that you want to build a tree that's as simple as possible.

And you can build a tree that's as simple as possible if you look at the data, and say, well, which test does the best job of splitting things up?

Which test does the best job of building subsets underneath it that are as homogeneous as possible?

So all this information theory, all this entropy stuff, is just a convenient mechanism for doing something that is intuitionally sound.

OK?

It's not about information theory.

It's about a sound intuition.

Oh, by the way.

Does this kind of stuff ever get used in practice?

10s of thousands of times.

This is a winning mechanism that's used over and over again, even when the data is numeric.

How would it work if it's numeric data?

Well, let's think about that for a little bit.

So let's suppose that we have an opportunity.

We're an EMT or something, we work in the infirmary.

What do they call it these days?

Something else.

But anyhow, you work in that kind of area, and you have the opportunity to take people's temperature.

And so over time, you've accumulated some data on the temperature of people.

And maybe you've found that there's a vampire here at about 102.

There's a normal person here, about 98.6.

But then they're scattered around.

Some people have fevers when they come in.

So the question is, is there a way of using numerical data-- things that you can put real numbers-- is there a way of using that with this mechanism?

And the answer is yes.

You just say, is the temperature greater than or less than some threshold?

And that gives you a test, a binary test, just like any of these other tests.

[? Krishna? ?] Right?

But where would I put the threshold?

I suppose I could just put it at the average value.

But that might not be the place that does the best job of splitting the samples into homogeneous groups.

Christopher?

STUDENT: So you run this numerical analysis on different places with different thresholds.

PATRICK WINSTON: So you try different places, he says.

And he's right.

Because this is a computer, this is our slave.

We don't care how much it works to figure out the right threshold.

So what we do is we say, well, maybe the threshold's halfway between those two guys, or halfway between those two guys, or those two guys, or those two guys, or those two guys.

So we can try one less threshold than we have samples.

And we don't care if there are 10,000 samples, because this is a computer, and we don't care if it works all night.

So that's how you find the threshold for a numeric test.

By the way, I assured you earlier on you would never use the same test twice.

Is that true for this?

Yes, you would still never use the same test twice.

But what you might do is you might use a different threshold on the same measurement the next time around.

So when you start having numerical data, you may find yourself using the same test with the same axis but with a different value.

All right.

So now that we have this, then we can go back and compare how this method would look when we put it up against the sort of stuff we were talking about last time, with the electrical covers.

So with the electrical covers, we had a situation like this.

I don't know.

We had samples that were places like this, and we had a division of the space that look pretty much like that.

Not quite exactly in the right spots, but pretty close.

So these are the decision boundaries for the situation where we are using nearest neighbors to divide up the data.

What would the decision boundaries look like if these were four different kinds of things, and we were using this kind of mechanism?

And maybe there's a lot of samples all clustered around places like that.

What would the decision boundaries look like?

Would they be the same as this?

god, I hope not.

Why?

Because what we're going to do is we're going to use a threshold on each axis.

So therefore, the decision boundaries are going to be parallel to one axis or the other.

So we might decide, for example-- Oh, shoot.

I think I'll draw it again, because it'll get confused if I draw it over the other one.

So it looks like this.

And that's how nearest neighbors does it.

But a identification tree approach will pick a threshold along one axis or the other.

Let's say it's this axis.

It's only got one choice there.

So it's going to put a line there.

And now, what's the next thing it does?

Well, it still has these two different kinds of things to separate.

We're going to assume we've got four different kinds of things.

So it's going to say, oh!

I've Come down the negative side, so I need a threshold on the remaining data.

And these are the only two things that are now remaining.

So my only choice is to put a threshold in there.

Now I guarantee this, absolutely guaranteed-- on the quiz, somebody-- presumably somebody who doesn't go to lectures-- will draw that line all the way across.

And that's desperately wrong.

Because we've already divided this data set in half.

Now the choice of what we do over here is governed only by the remaining samples that we see, these two.

And so the threshold is going to go in there like that.

So that's what happens when you go back.

This is used 10s of thousands of times.

Always used.

What are the virtues of it?

Number one, you don't use all the tests.

You use only the test that seem to be doing some useful work for you.

So that means that you do a better job, because your measurement technique is simpler.

And it costs less, because you're not going to the expense of doing all of the testing.

So it's a real winner.

But you know what?

Some classes of people-- not scientists, but I mean people like doctors and stuff.

They don't like to look at these tress.

They're kind of rule-oriented.

So they look a tree like this for determining what kind of thyroid disease you have, and it would have maybe 20 or so tests in it of various kinds of hormones, like thyroxine and this and that.

And they say, ah, we can't deal with that.

So we have to work with them.

So what we do is we convert the tree into a set of rules.

How do we convert the tree into a set of rules?

Oops, wrong one.

Go away, go away.

Here's what I want.

Yeah, good.

How would we convert this tree into a set of rules?

It's straightforward.

[INAUDIBLE], what do we do?

STUDENT: You'd basically just look down each branch-- PATRICK WINSTON: You'd basically just go down each branch to a leaf.

So you say, for example, here's one rule.

If shadow equals question mark, and garlic equals oh, [INAUDIBLE] want to choose No.

Doesn't eat garlic.

No.

I think I'll say Yes.

Yes.

That changes the answer.

Then if it eats garlic, it's not a vampire, right?

That's one of four possible rules, because there are four leaf nodes.

Now, almost done.

We are done, except for one thing.

We can actually take these four rules, and start thinking about how to simplify them.

You can ask questions like, if I have a rule that tests both the shadow and the garlic, do I actually need both of those antecedents?

And the answer is, in many cases, no.

And in particular, in this case, no.

Because if we look at our data set, what we discover is that in the event that we're talking about a shadow question mark-- oh, I guess I had a better choice the other way.

Oh, no.

If you look at the garlic, all the garlics-- Yes, Yes, and Yes-- it turns out that the answer is no, independent of what the shadow condition is.

So we can look at the rules, and in some cases, we'll discover that our tree is a little bit more complicated than it needs to be.

We can actually get rid of some of the clauses.

So in the end, we can develop a very simple mechanism based on good old fashioned rule-based behavior, like you saw almost in the beginning of the subject, that does the job.

And now, without any royalty, you're all free to put this in your PDA and use it to protect yourself in the days to com, especially since Halloween's just around the corner.