HERBERT GROSS: Hi, our lesson today concerns functions. And in the certain manner of speaking, this is perhaps where a course in calculus should usually begin. After all, what will be studying is a relationship between certain variables. This is what we mean by a function in general. Or at least, in particular, with our emphasis on real variables and graphs that we were talking about. And what we would like to do now is to motivate the concept of function from a more general point of view in terms of the language of sets that we have talked about and read about in our supplementary notes. Let's then begin with a basic general definition of what a function is, after which we will specialize and talk about functions of a real variable for the remainder of the lecture.

We begin with a definition written this way. A function f from A to B-- see, it's written this way. It means f is a function from A to B means that f is a rule, which assigns to each element in set A, an element in set B. Now of course, this may sound like just an empty bunch of words. But if we come here and look a little bit at a picture, we get the idea as to what's going on. Namely, we may visualize our sets as circles. And what does the function do? The function assigns to each element of A, an element of B. In fact, notice the geometric wording again. This element is mapped into this element by f. This element is mapped into this element by f. And this element is mapped into this one.

And while we're looking at this, perhaps it would be a good chance to emphasize not to read more into a definition than what's already there. You see when we say that f assigns to each element of A, what we mean is we will not allow something like this to happen. We will not say, for example, let f send this element into this one and this one. This becomes non well-defined. It becomes ambiguous. In other words, we do not want to have to make the value judgment as to which of two things we are going to look back at as being what A maps into. And why that's the case we'll mention as our course proceeds.

Also, notice that when we say that each element of A is assigned to an element of B, two things are implied here. First of all, notice that we do not insist that all of B be used up. You see, in other words, here's some surplus B's over here, which are not used up by A's with respect to f. And secondly, whereas we prohibit the same A from having two different images in B, we do not prohibit one B from being the image of two different elements in A. In other words, notice that even though both of these elements here are assigned to the same element in B with respect to f, there certainly is no violation of our definition that each element here was assigned to an element here.

By the way, this leads in the literature to three different terms that we should define right now. One of these is the set A itself and that's called the domain of f. You see this is what f is defined on. The domain is the set A over here. Often abbreviated-- well, there are many different abbreviations. Sometimes one just writes D-O-M with an f or D-O-M with a subscript. Or a capital D with a subscript f. But we will adapt to these notations as we need them. But for our purposes all the domain of f is, it's the set A. It's the set on which f operates to assign values, ok. And then the companion to that is the set B. And B, almost again in graphic terms, is called the range of f.

And by the way, as you look at the range of f, you may get the feeling that somehow or other, this little element over here is kind of out of place. That maybe somehow or other let me just circle this. Let me call this the set C. Somehow or other, you get the feeling that C describes much better what f does to A than B. Because you see, each of the elements in C is used up as being the image of at least one element in A. Somehow or other you see, we could have deleted this element B without too much loss of continuity. And to get around this we interject still a third definition, C. It's called the image of f. You see, image verses range. Image is the part that's actually used up by f with the mapping. And sometimes this is even abbreviated as follows. You use the same f as one uses for the function and then a capital A in parentheses to indicate what? This is a set of all things of the form f of x where x is in A. In other words, again, the standard notation if we want to look at it this way.

If we call this element here a and this element here b, to indicate that a was mapped into b by the function f or the rule f, we often write that as f of a equals b.

Now in many places, including our text, there is no distinction made between the image and the range. And the reason for this is that in many situations, the image and the range turn out to be quite the same thing. In fact, if they do turn out to be the same thing, we call that a rather special type of function. Namely, the function from A to B is said to be onto if its image equals its range. Now again, in terms of words, that doesn't say very much. The image equals the range. What does that mean, image equals the range? Perhaps the best way to see this is by means of an example.

The example I have in mind is this. And this happens many, many times in mathematics. One begins with a set. In this case, I picked A to be the set consisting of the numbers 1, 2, and 3. Frequently, one defines a function explicitly on A without any regard to a second set B.

For example, in this illustration I've said, let f of a be 4a for each a in set capital A. Well, what are the elements of capital A? They are 1, 2, and 3. So according to this recipe, what do we have?

Well, f of 1 is 4. f of 2, see, it's what? 4 times 2 is 8. And f of 3. That's what? 4 times 3 is 12. Let me now invent a new set B. And the elements of B, as you could probably guess, are going to be 4, 8, and 12. And now I look at my function f from A to B.

Notice that in this case the function from A to B uses up all of B. In fact, in terms of a diagram, you see here's A. Has what? 1, 2, and 3. And here's B, which is made up of 4, 8, and 12.

Now, what does have f do? It maps 1 into 4. It maps 2 into 8. And it maps 3 into 12. What's happened here? A has not only mapped into B, but all of B is used up in this. In other words, notice then that the range of B and the image of B in this particular case happen to be the same. This will happen in every single case where we start with a set A and define a function on A. Namely, we see what f of x is for each x in A, take the collection of all those images, call that set B, and then you see by default so to speak, B will be both the range and the image of f. And it's in this sense that in most textbook examples that we deal with, we need not make any distinction between the range of the function and the image of the function. But roughly speaking then, just to keep things straight, a function is called onto if the entire range is used up in the mapping. But if there are elements of the range which are not used up as images, then the function is simply called into, or not onto. But at any rate, this is the concept of what we mean by onto.

Now, a second feature that one talks about with functions which in no way is connected with onto, this, but which is a very important independent feature is something which is called a 1:1 function. Let's look at that for a moment too.

For example, let's suppose I have a function f defined on my set A. The question is, f maps a1 into a particular element of B and it maps a2 into a particular element of B.

Now, there are two possibilities that can happen. One is that f of a1 and f of a2 will be different elements of b. In other words, what will happen is, is that two distinct elements of A will have distinct images in B.

On the other hand, it's possible that the two different elements of A have the same element of the same image in B.

Now, by and large, whereas nothing is wrong when this happens, it does cut down our operating speed to some extent when it does. Because you see, frequently to study a particular function, we may want to look at the image rather than the domain. And somehow or other, you see if two different elements can map into the same element in the image, then you see when we look at the image we have no way of knowing which of these two elements we're talking about. So in other words then, if it should turn out that no two different elements in A can have the same image in B, in other words, notice what this thing here says. You see, if what? This is the image of a1. This is the image of a2. It says what? If a1 and a2 have the same image, then they must be the same element.

Now if that happens, and again as I show you over here, it doesn't have to happen. If that happens, the function is called 1:1. For example, here is a picture I've drawn in which a function is 1:1.

By the way, I've drawn this picture so that my function f is both 1:1. Meaning what? That no two elements in A have the same image in B. And secondly, it also happened to be onto here. Namely, no element of B was left out by f. Now that wasn't crucial.

For example, if I do this notice now that the function from A to B is still 1:1. No two different elements in A have the same image in B. But now you see the function is no longer onto because there happened to be elements in B which are not mapped into under F by elements of A.

Now, what is nice of course is that if a function happens to be both 1:1 and onto, notice that we can induce a new function, which I'll call g from B to A by essentially reversing the arrowheads here. You see, if the function is both 1:1 and onto, by reversing the arrowheads, instead of getting a function from A to B, I do get a function from B to A. This function is called the inverse function and will play a very important role in much of our course which follows.

The important point to notice, however, is that if the function is not both 1:1 and onto, you cannot reverse the arrowheads, believe it or not. Well, you say, I can reverse them, can't I? Why can't I reverse them over here? And the answer is well, look it. If we include these being in here, suppose we reverse the arrowheads now. Look at B.

What is the domain of g? Well, for B to be the domain, every element of B has to be assigned to something in A by g. But look at these two elements over here, g doesn't act on those at all. In other words, if the original function is not onto, then when you reverse the arrowheads you haven't defined the new function on your whole domain here.

In another sense, if the function was not 1:1 when you started. In other words, suppose this happened. So f was not 1:1. Now you see when you try to reverse your arrowheads, notice that the element here in B is assigned to two different elements in A. And we agreed that we wouldn't allow that to happen.

OK, so far so good. Notice that that particular part of our course has nothing to do with real variables and the like. Meaning when we're talking about sets they can be sets of arbitrary numbers. Now what I'd like to do is zero in, on our specific calculus of a single variable course. And let's go back to our old friend who somehow or other has made an appearance in every lecture that we've had so far. Let's go back to s equals 16t squared.

Only now, we're not going to repeat the same old stuff that we did before with it. We're now going to get slightly more sophisticated. Namely, when we talk about s equals 16t squared, what problem was being done here? You are assuming that there is no air resistance. An object is being held above the ground. You release the object and the distance s that the object falls in feet after t seconds is given by s equal 16t squared.

Now if we think about that for a while, we realize that that does not tell the whole picture. Obviously, the s equals 16t squared applies only to the time in which that object is falling. Perhaps what we should have said was this, that until you release the object it doesn't fall any distance at all. Then from the instance you release it, it starts to fall a distance s given by 16t squared. Not forever, but until it hits the ground. Let's call t sub g the time at which this thing hits the ground. You see this recipe that we called s equals 16t squared is not in effect forever. It's in effect only when t is between 0 and t sub g. And by the way, hopefully once the object hits the ground it won't fall any further. In other words, for any time after t sub g, the distance that it's fallen is 16t sub g squared. Meaning this is the distance that its fallen when it hits the ground and it stays there.

If we wanted to graph this, you see, and notice how we are refining our previous result. The graph is not this, you see, the graph is what? The distance is 0 until t equals 0. Then the distance that it falls increases up till the time the object hits the ground. And then it levels off like this. And by the way, in terms of making a few asides, notice that this curve here does represent a 1:1 function. Namely, if you pick two different times in this strip, you have two different distances. Two different times cannot yield the same distance.

As opposed to the fact, let's call this t1 and t2. As opposed to the fact that once the thing hits the ground, our function is no longer 1:1. In fact, the any two values of t once the thing has hit the ground, we have the same s value. In other words, what we're saying is what? That once the object hits the ground, it really makes no difference what t is, s is still going to be 16t sub g squared.

Now that was just an aside. The reason I mentioned this is to motivate a very important type of domain that takes place when we deal with functions of real numbers. In most cases, when we do a physical experiment it's over some time interval. We put something into effect and say, let's measure it for one hour. Or let's measure if from now until 3 o'clock tomorrow. In other words, in general, whereas a domain of a function can be anything we want it to be, in most real life laboratory situations, our domain happens to be a connected interval, whatever that means intuitively. In fact, let's try to talk about that in more detail.

In other words, a very special type of domain that one uses when one talks about functions of a real variable. They are called intervals. Written as sets, if a is less than b, we talk about what? The set of all x which greater than a and less than b. By the way, that's called the open interval from a to b. It's written this way with parentheses. The set of all elements from less than b and greater than a inclusively is called the closed set or the closed interval from a to b. And it's written this way. And pictorially, you can't tell these apart.

Namely, if this is a and this is b, both of these pictorially are what? An interval as we think of it intuitively. Namely, it's this stretch. But in one case, the endpoints a and b are included. And in the other case, the endpoints are excluded. They're included in the closed interval. They're excluded in the open interval. And again, notice that since a point has no thickness, we have no way of telling just by looking at the figure which of these two is meant unless we draw in the appropriate diagram this way.

By the way, notice also that an interval can be half open and have closed. I mean, for example, one could talk about how about including the left endpoint but excluding the right endpoint. See, why couldn't we talk about something like this? In which case we would have written the half open half closed interval this particular way. Now again, this is all notation. It's things that you can memorize. Things that are emphasized in the text. But the thing that I wanted to try to have you see from the lecture is why we concentrate so heavily on the things called intervals.

It's because in most situations when we deal with functions of a real variable, our so-called input, is usually defined on some continuous interval. All times from such and such to such and such.

Now, by the way again, notice that the picture-- just as we've been talking about before. The picture comes in handy. Namely, 1/2 being in the open interval from 0 to 1 does not need a picture to interpret it. Namely since 1/2 is greater than 0 but less than 1, by definition 1/2 is in this interval. On the other hand, by use of a picture, I think it becomes rather easy to visualize what it is that we're saying when 1/2 is in this particular interval.

Again, notice when somebody says does 0 belong to this interval? Notice that subtlety about open and closed, point versus dot. Namely, 0 does not belong to the open interval from 0 to 1, but it does belong, for example, to the closed interval from 0 to 1. Because what is the basic difference between these two? In this one, the endpoints are not included. In this one, the endpoints are included.

Now, a companion to interval is a very important building block of this course. It's something called a neighborhood. Now, in terms of a definition, a neighborhood isn't a very exciting thing. A neighborhood of a point c, a neighborhood of x equal c is simply an interval which contains c inside. You want c to be inside the interval.

Now what does that mean intuitively? Well, what it means is pick any interval which has c inside. Maybe we can go from this point to this point. This would be called a neighborhood of c.

By the way, you may notice I've drawn this as an open interval. The idea is that we really want c to be inside the interval. We do not want the situation where c is one of the endpoints. And whereas we'll talk about this in more detail later, the important point is that in many of our investigations in calculus we will want to study what's happening just before we get to a certain point and just after we leave that point. And somehow or other, if we let that point be at the very end of our interval, we have no information. For example, if that point is the left endpoint, we don't know what's happening before we get to the point. If it's the right endpoint, we don't know what's happening afterwards. And that's why you'll find in the textbook that a neighborhood is defined to be an open interval, which contains c. In other words, we want to make sure that c is in the interior here.

By the way, in many cases it turns out algebraically to be easier if this happens to be what we call a symmetric neighborhood. In other words, if c is in the middle. We won't go into that right now, but if c happens to be in the middle that's called a symmetric neighborhood. In fact, another way of writing that is to say what? Pick some definite distance h and what do you write down? You write down c minus h to c plus h and that puts c right in the middle of this particular interval. And you see, the idea here is that when you're looking at what's happening to a function, you may lose symmetry.

For example, in this particular graph that I've drawn, notice that at this particular point I've marked off equal intervals on both sides of l. But notice that when I come down here, they do not project onto equal intervals on either side of c. In other words, if this had been a straight line. Frequently what we do in a case like this is we say well look it. If we're interested in seeing what happens near c, why don't we just-- this is non-symmetric. Why don't we just take the smaller of these two widths and see what happens in the symmetric part? In other words, if the neighborhood is not symmetric, we can always make it symmetric. And so there really isn't that much to worry about in that particular respect.

But why are we interested in neighborhoods in the first place? And the answer is that in many cases what we're going to be doing is studying what's happening near a particular point c and want to know what's happened just before and what's happening just after.

The next important concept that's connected with neighborhoods is the idea of a deleted neighborhood. And that in turn, is very strongly connected with 0/0.

For example, consider the function f of x which is x squared minus 9 over x minus 3. If we let x equal 3, if the input is 3, notice that the output becomes 9 minus 9 over 3 minus 3, or 0/0.

On the other hand, if x is any number whatsoever except x equals 3, no harm is done with this as an input. Consequently, what one is talking about now is the only time you get that 0/0 form is when x is 3.

What happens if you're in a neighborhood of 3, but not equal to 3 itself? You see what I'm driving at here is pick any number x in this interval other than 3 itself. And notice that f of x can be written this way. As long as x is not equal to 3, we can cancel x minus 3 from numerator and denominator. Remember, we can't divide by 0. And now we see what? That as long as x is not equal to 3, f of x is perfectly well defined. And consequently, this is what motivates the concept of a deleted neighborhood. Namely, everything is fine in this whole neighborhood except for 3 itself. So to avoid that unpleasantry, let's just delete that point. And that's called a deleted neighborhood.

And you see what we do when the neighborhood is deleted, we're still going to talk about what? How close you are to that point. And by the way, this brings us to another very fascinating aspect of what's going on between our geometry and our arithmetic. Do you really talk about the distance between numbers? I mean, is 7 near 3? And the guys says, well, what do you mean, is 7 near 3? Well, I would say here that 7 is very near to 3. When you say that 7 is near 3, you certainly don't mean close to in the geometric sense. You mean the difference between them is small.

In other words, the next thing that we have to talk about is how when we talk about being close to a point which is a geometric term, how do we talk about that algebraically? You see geometrically, how do you talk about the distance between x1 and x2? Well, if you're going this way, it's just what? x2 minus x1. If you're going the other way, the direct distance this way, it's x1 minus x2. In any event, the distance between these two points is just the magnitude of the difference of these two numbers. And that leads, you see, to the concept that's hit quite heavily in our text, and that is the concept of absolute value. Perhaps one of the most critical analytical geometric topics that we tackle in our early part of our course.

Analytically, we define the absolute value written with vertical bars here, x1 minus x2. The absolute value of x1 minus x2 to be the positive square root of x1 minus x2 squared. And in plain English all this says is what? See when you square and then take the positive square root, you haven't undone what you've done before. All you've done is what? If it's positive here you haven't changed anything. But if x1 minus x2 are negative, when you square it and extract the positive square root, all you've done is changed the sign just like you're supposed to. Let me give you an example.

Suppose you're faced with the absolute value of x minus 3 is less than 2. What does this say geometrically? Geometrically what it says is that x is within two units of 3. in terms of a picture, all you have to do now is draw in 3, mark off two units on either side, and for x to be within two units of 3, all you know is is that x has to be in here. In other words, look at how easily you can solve this particular problem.

On the other hand, you can always go back to the basic definition and say, wait a second. This means the positive square root of x minus 3 squared is less than 2. So I will square both sides. If you do that you get x minus 3 squared is less than 4. If you now collect terms and expand, you get x squared minus 6x plus 9 minus 4. That's what? Plus 5 is less than 0. This factors into x minus 1 times x minus 5 is less than 0. The only way the product of two numbers can be negative is if the factors have different signs. Since this is x minus 5 is less than x minus 1, this must be the smaller of the two. This must be the larger of the two. To say that x minus 1 is greater than 0 is the same as saying that x is greater than 1. To say that x minus 5 is less than 0 is the same as saying that x is less than 5. You put that all together and notice that even though it wasn't quite as comfortable, we can obtain the same answer algebraically as we can obtain geometrically.

In other words, our relationship between algebra and geometry remains the same.

Again, when you can draw the picture, it's worth a thousand words. If you can't draw the picture or you're suspicious about the picture, especially when it involves point versus dot, then what you do is resort to the analytic definition. These do not replace one another, they work hand in hand.

Finally, what we must talk about now is the arithmetic of functions. Can we combine functions to form functions? And the answer is yes.

First of all, in talking about the arithmetic of functions, what must we do? We must first, at least, define what it means for two functions to be equal. Well, for two functions to be equal, all we insist on is that first of all, they're defined on the same domain. And secondly, that for each input in the domain, each function gives you the same output. For example, suppose a is a set whose elements consists of 0 and 1. And suppose b is also the set whose elements consist of 0 and 1. One such function would be f. It maps 0 into 0. It maps 1 into 1.

Another function, which I'll call g-- see, f does what? It maps 0 into 0 and 1 into 1. What does g do? g maps 0 into 1 and 1 into 0. Notice that f and g are different. They both have the same domain. They both have the same image. But notice that element for element, they're not the same. Namely, f and g do different things to 0. f sends 0 into 0. g sends 0 into 1. So I can tell f and g apart. And because I can tell them apart, they're not equal. All right, so equality means I can't tell f from g.

Now the next kind of question is, how do you do arithmetic with f and g? Can I add two functions? Can I multiply two functions? Can I subtract two functions? And the answer again, turns out to be yes. And not only yes, but yes in a rather simple way. Let's again do this by means of examples.

Suppose we defined f of x to be 2x for all x in A. Namely, if A is 1, 2, 3, f of 1 will be 2, f of 2 will be 4, f of 3 will be 6.

Let's define another function on A, let's call it g. g of x will be x plus 1 for each x in A. In other words, g of 1 will be 2, g of 2 will be 3, g of 3 will be 4.

Now the point is, can I add f of x and g of x? Well, sure. f of x is 2x. g of x is x plus 1. So if I add these I get h of x is 3x plus 1.

In a similar way, could I have multiplied these two? Well, sure. Again, f of x is 2x, g of x is x plus 1. If I multiply these together, I get what? 2x times x plus 1, which is the same as 2x squared plus 2x. Now of course this probably doesn't look too smooth because there's no pictures here. All we're saying is this.

Here's A and all you're saying is that if you add f and g, what do you get? If A is 1, 3x plus 1 is 4. 2 times 3 plus 1 is 7. 3 times 3 plus 1 is 10. In other words, in this case, our image, if I want to call it b, would look like this. This would be the sum of the two functions f and g.

And similarly, for the product I could do the same kind of a thing. In other words, I can just arithmetically, since both the output of f and the g machines are real numbers, and the sum of two real numbers is a real number, I can add and multiply functions to form functions.

But there's one other important way of combining functions in calculus. A way which is very, very important and one which we may not have seen too much of before. And so let me close our lecture for today with an emphasis on that particular topic.

It's called composition of functions. And to see what composition of functions means think of a particular example where maybe the f machine f of x is 2x. In other words, think of it this way. We run x through the f machine, the output will be 2x. Now we run the output of the f machine into the g machine.

Now what does the g machine do? If x is the input, x plus 1 means 1 more than the input. The g machine always adds on 1 to the input to give you the output. Well, if the input is 2x, the output will be 2x plus 1.

Notice that these two together can be thought of as being one function machine, which I'll call the q machine. In other words, what happens for the q machine is what? x runs into the f machine, the output of the f machine becomes the input of the g machine, and the output of the g machine is then the output of the q machine. The q machine is sort of build with component parts here. And the reason that this is very, very important is that this comes up in calculus all the time, where the first variable is related to the second variable, the second variable has a definite relationship to the third variable, and we now want to relate the first variable to the third variable.

And the way we write that-- and I guess this is hard to see. This is not an O over here, it's a little circle, like a dot, and it's called the composition of g and f. It's not gof. Its g circle f.

And the q machine is what? You write it this way and maybe if you look at the picture you can see exactly what's happening here. You apply f to x and then apply g to the result. In other words, just looking at this picture it becomes rather apparent that q of x is just 2x plus 1. Notice you see, that the domain of the q machine is the same as the domain of f. The input of the q machine is what goes into the f machine. The output, the image of the q machine, is the image of the g machine. In other words, just this particular thing.

Now this type of function combination called composition, is a very intricate thing. It depends on the order in which you do these things. This is a rather interesting point. For example, when you add two numbers, a and b, it makes no difference in which order you add them. On the other hand, when you divide two numbers, a and b, the quotient does depend on the order in which you divided them. Well, the same thing is true here. Let's call p the function which starts with the g machine followed by the f machine. And as this lecture wears on, I think maybe there's a reason for making this circle look like an O. Maybe this is starting to look a little bit like fog at this time. It's not. All I want you to see is that what we do now is we start-- see, what are we going to do here now? We're going to start with the g machine first, then the f machine.

In other words, what happens now? If the input is x, the output of the g machine is one more than the input. That would make the input x plus 1 to the f machine.

What does f do? Remember, f doubles. f doubles the input. So the output here would be what? Twice x plus 1. In other words, what would p of x look like? If x goes into the p machine, what comes out is twice x plus 1, or 2x plus 2.

On the other hand, when we put the f and the g machine in the other order and formed q of x, what was the output? Let's go back here and look. The output was 2x plus 1. In other words, do you build a different function machine by interchanging the f and the g? And the answer is yes. In other words, what I'd like you to see for concluding this part is that whereas everything was pretty straightforward up until now, the most important new concept, one which was not so intuitive is the one that's called the composition of functions. It's the one that occurs all the time in related rates problems. We'll be using it over and over again in this course. And all I want you to see is that first of all when you use the composition of functions, what you get depends on the order in which you combine them. And that secondly, and most importantly, that neither of these is the same as this. That combining two functions is not the same as multiplying the outputs of two different functions.

Well, I think that's sort of enough of a mouthful for one sitting. Our main aim today was to introduce functions, the language that we're going to be using the rest of the way. Because after all if we don't have the vocabulary established it will not be second nature to talk about the concepts. Starting next time and beyond we will deal more with specific calculus contexts. But until next time, goodbye.

ANNOUNCER: Funding for the publication of this video was provided by the Gabriella and Paul Rosenbaum Foundation.