Data Analysis for Biologists

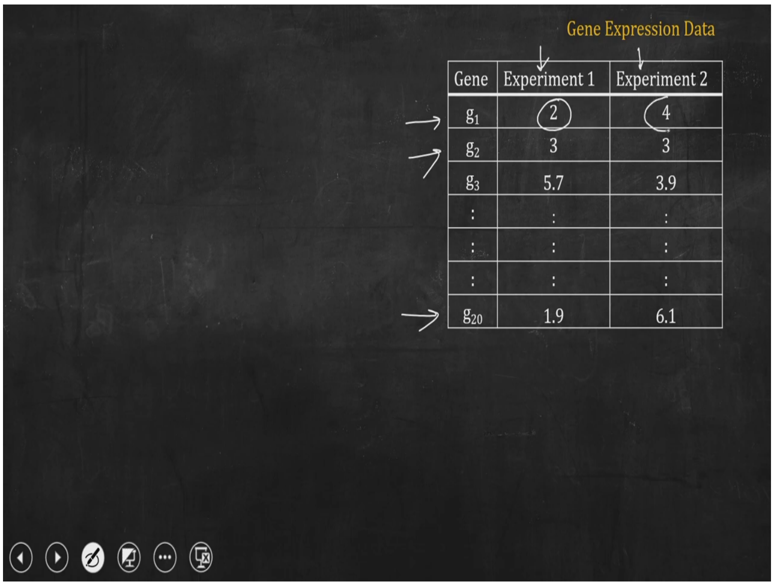
Professor Biplab Bose
Department of Bioscience and Bioengineering
Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology, Guwahati

Lecture 08

Vectors and Vector Operations

Hello everyone, welcome back. In this course, we will use lots of data analysis techniques, which are based on linear algebra, starting from regression to principal component analysis. And when in your future, we will learn R, you will see that many times we convert data sets into vectors and matrices, and then we use techniques of linear algebra to analyse them, that makes our life easy and R has inbuilt tools to do that. So, we have to learn the basic bone things of the linear algebra. So, in this series of lectures from now, we will discuss some elementary concepts which will be used in our data analysis course for from taken from linear algebra. So, let us start today, vectors.

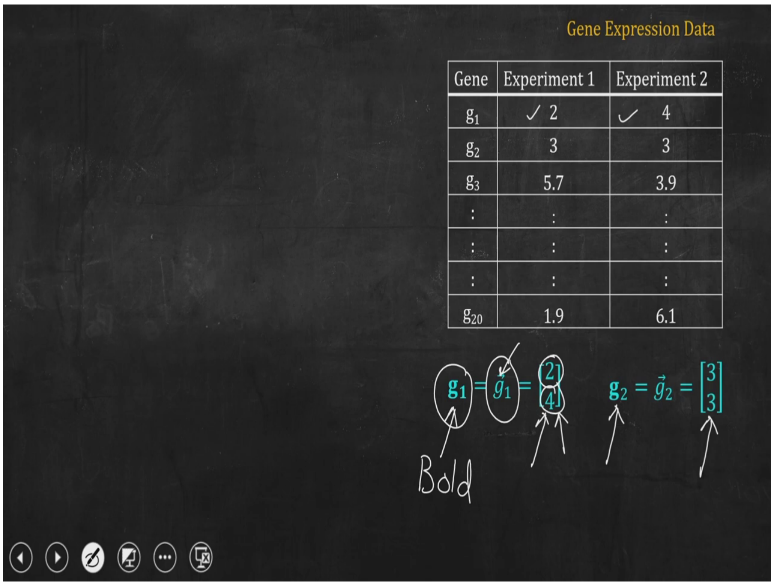
(Refer Slide Time: 01:22)



I have a data set here taken from a gene expression experiment. So, I have two experiments, experiment 1 and experiment 2 usually these types of gene expression experiments like microarray or RNA seq experiments, you have large number of data… genes for example, 200-2000 genes as add at a time, but for clarity here, I have taken only 20 g20.

So, I have 20 genes and they are data, fold change data for two experiments. Now, suppose I want to represent these data in a bit different way, I want to represent these each of these data for each of these genes as a vector, what do I mean by that? Let us take an example. So, for g1 and g2, I will show you the example how I can convert the data for g1 and g2 into vectorial form. For example, g1 in experiment 1 I have its fold changes 2 and in experiment 2, its fold changes 4. So, what I will do? I will stack these two data together and put them in a square bracket just like this.

(Refer Slide Time: 02:30)



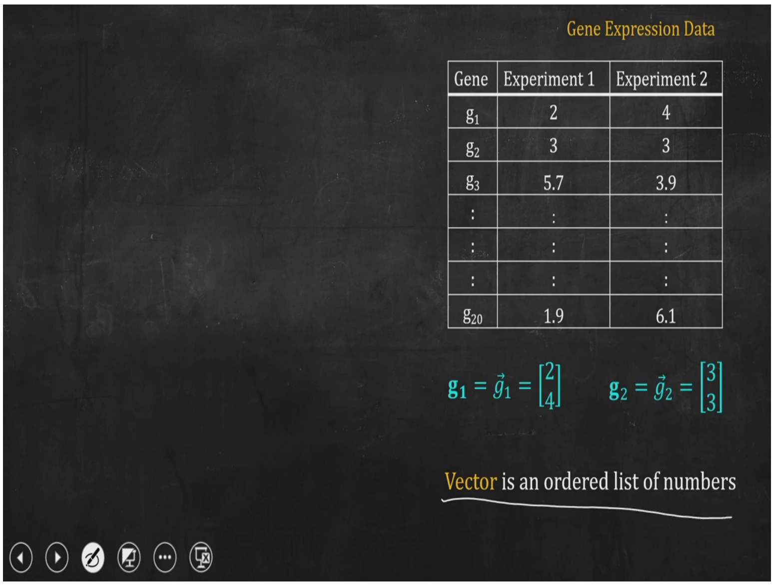
$$g_1 = \vec{g_1} = [2\ \ 4]$$

$$g_2 = \vec{g_2} = [3\ \ 3]$$

So, what I have done here? I have stacked 2 and 4 taken from this table and stacked them over each other and placed under a square bracket. So, this is the vectorial representation of the data. So, I have created a vector g1 and notice g1 I have written in bold or sometimes you can put an arrow to represent that this is a vector.

Similarly, for g2 again I have written this in bold, and what I have done? I have stacked the data for experiment 1 and experiment 2 together inside a square bracket. So, what I have got? Forget about the square bracket what I have done, I have created a list, ordered list, why it is ordered? Because the first row, the first element of this list is coming from the first experiment and the second one is coming from the second experiment, so it is an ordered list.

(Refer Slide Time: 03:32)

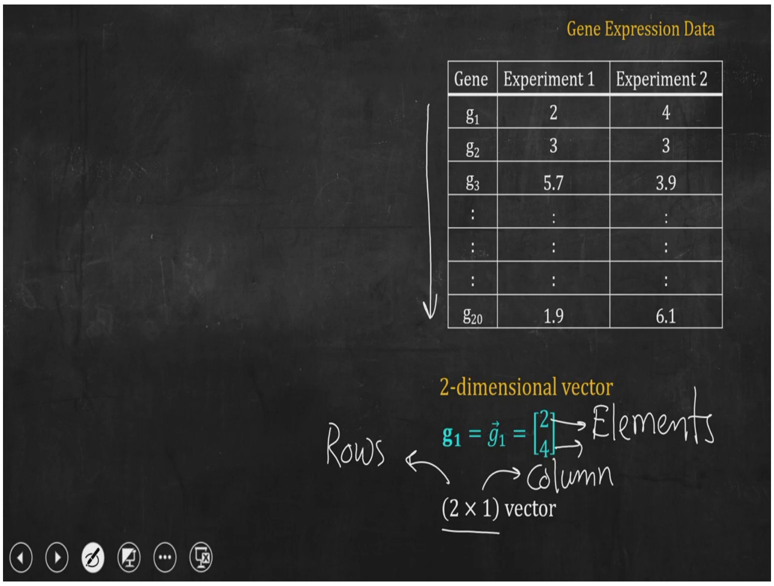


$$g_1 = \vec{g_1} = [2\ \ 4]$$

$$g_2 = \vec{g_2} = [3\ \ 3]$$

So, in a way a vector is an ordered list of numbers as simple as that.

(Refer Slide Time: 03:41)



$$g_1 = \vec{g_1} = [2\ \ 4]$$

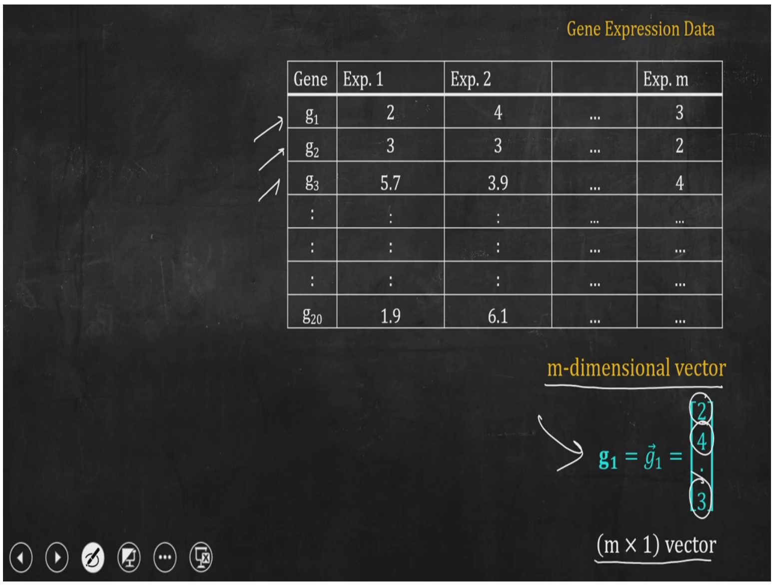
Now, look into the dimension of this vector. This dimension of this vector is 2, why? Because I have two elements these are elements. So, I have two elements coming from these two experiments and it has 2 rows this vector has 2 rows and 2 columns. So, 2 into 1 is 2. So, the

2-dimensional system 2-dimensional vector, when I will draw this vector and the data points these 2 issues of dimension will be much more clear.

Here many times we represent this 2-dimensional vector as 2 into 1 vector also we write it as a 2 into 1 vector what do I mean by 2 into 1? I mean I have 2 rows and I have 1 column. So, do the 2 into 1 vector g1 g2 both of them are 2 into 1 vector. So, in this way for all these 20 data points, all these 20 genes rather, all these 20 genes I can represent the data of all these 20 genes as 20 vectors, each of these vectors will be 2-dimensional vector.

So, now if I increase the number of experiments in this example, I have two experiments, suppose I have treated cells with 2 different doses of drugs and then compare the change in gene expression with respect to untreated one. So, I have 2 experimental conditions. Now, suppose I have treated cells with 10 different doses of the drug or 10 different drugs. So, then I have 10 experiments.

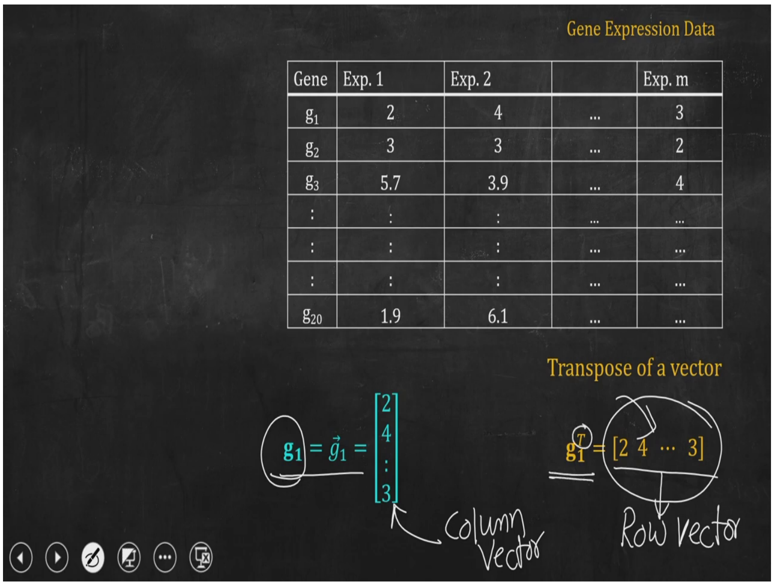
(Refer Slide Time: 05:22)



$$g_1 = \vec{g_1} = \begin{vmatrix} 2 \\ 4 \\ \vdots \\ 3 \end{vmatrix}$$

So, as a whole suppose I have m experiments. So, then my data table will look like this one starting from experiment 1 experiment 2, and then experiment m. Now, I can represent now, the data of each gene g1, g2, g3 like that as a m dimensional vector, the way I have shown here. So, what I have done now? I have now stacked m number of data for g1 in 1 single vector. So, the number of rows for these vectors is now m the number of columns is 1, so it is a m into 1 vector. So, total number of elements in these vectors m into 1 equal to m.

(Refer Slide Time: 06:07)



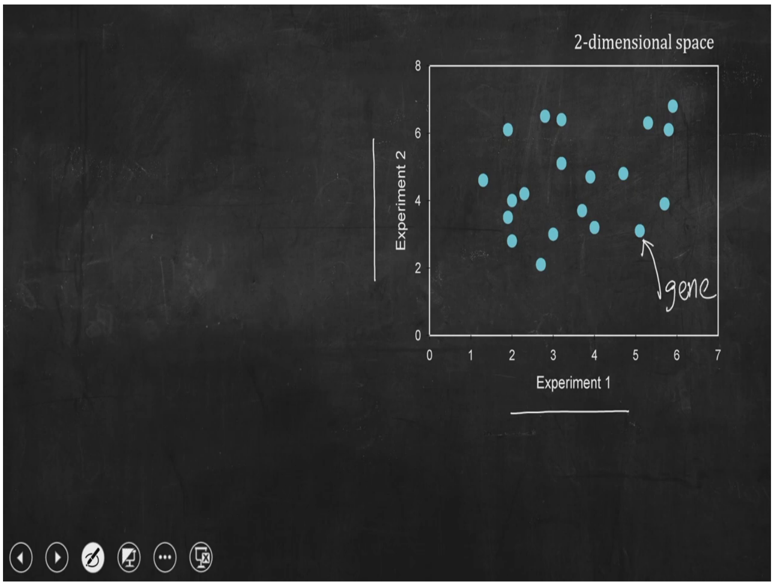
$$g_1^T = [2\ \ 4\ \ :\ \ 3]$$

Now, are these vectorial representation many a time called column vectors, this one where I have stacked one after another is called column vector, you can easily understand why it is called a column vector because is just 1 single column. Now, what if I write the same data in a row. For example, the way I have shown here, so side by side I have arranged, so then it will be called a row vector.

So, this will be called a row vector so, it is a matter of convenience how you are representing, you can represent the same data as column vector, you can represent the same data as row vector as an when you require. Now, if you notice this column vector g1, if I flip it, so that the column becomes row, then I get this row vector.

So, this flipping from column to row and the reverse one from row to column, if I flip the vector, this flipping of vector is called transposition. So, that means this vector which is a row vector is a transpose of my g1 vector, that is why it is written as g1 superscript T transpose of g1 vector, it is very simple, when you will discuss about matrix we will again come back to the transpose concept and but in vector it is so, easy to do, you simply convert the column vector and the row vector or the row vector into column vector, keeping the relative position of the elements same.

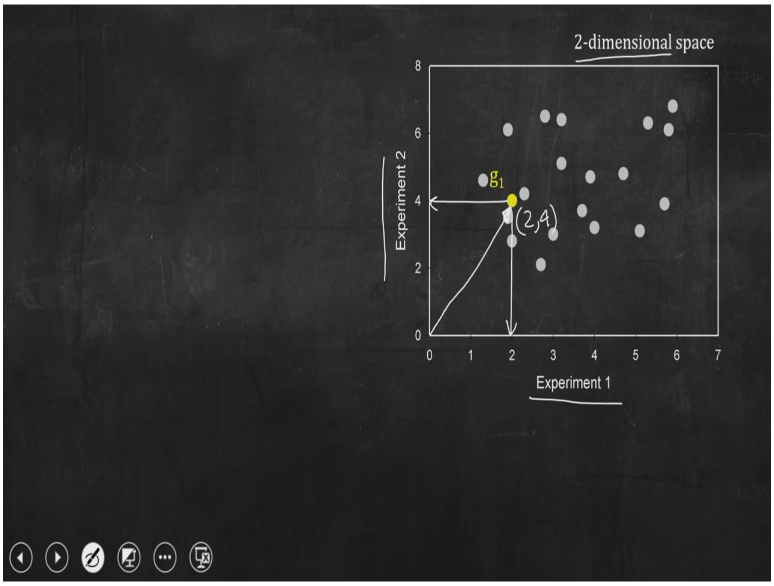
(Refer Slide Time: 07:52)

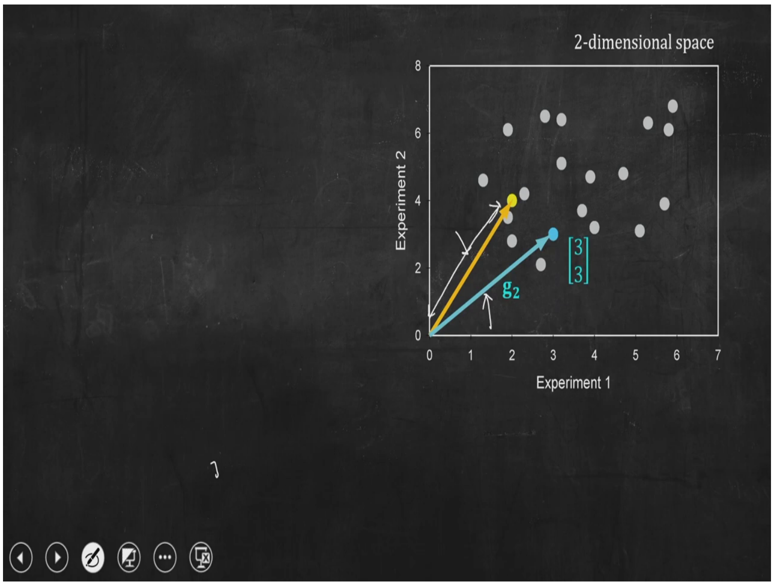


Now, in your physics class, you must have gone to vectors and you must be remembering that we have learned that vectors has magnitude and direction. Till now, what I have discussed, there is no issue of those arrows that we have seen in our physics textbook for vectors. We have just a least, it least inside a square bracket.

So, what is the connection of representation of that data in that vector format, and the arrows having a magnitude and direction that we know or associated with vectors, this is obvious connection between them to understand that what I have done here I have plotted the data. So, I have two experiments. So, in the horizontal axis, I have experiment 1 in the vertical axis, I have experiment 2, and each of these data points, each of these data points is a gene.

(Refer Slide Time: 08:48)





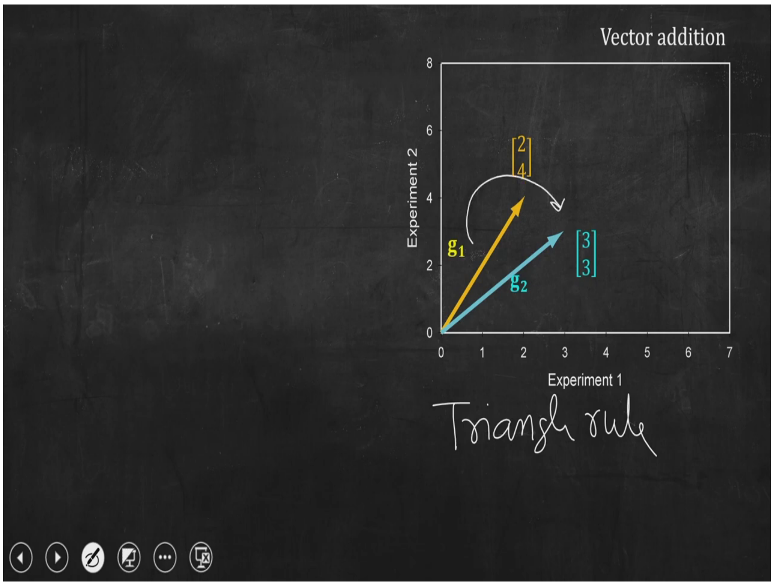
So, this yellow dot, which is this is 2 and this is 4. So, this is 2-4 coordinate, this is my gene 1. So, this is my gene, 1 is the result in the scatterplot, many a time you may have to show the result in this way also. Now, I can represent my g1 vector in the space. Remember, this is a 2-dimensional space because I have experiment 1 in horizontal axis and experiment 2 in the vertical axis.

Now, what I can do? I can draw an arrow from the origin up to this point Euler point, that will be a vector. And that is what I have done in this diagram. So, this yellow line, this yellow arrow is my g1 vector and you can represent that also as 2-4 vector. Similarly, take the g2 vector, that is for the gene 2.

So, that will be this blue line, this blue arrow. So, now you must be able to connect the vector that you have learned in your physics course, which has an arrow shape and has a magnitude and a direction You can easily see here these both of vectors has magnitude which is the length and they have directions.

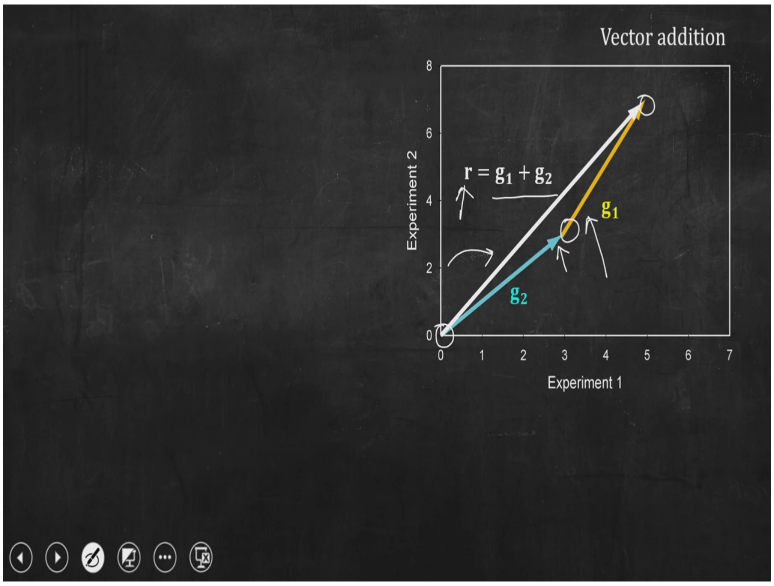
Now, once I have the vectors then I should be able to do operations of vector on these two vectors. So, these two gene data are now my vectors I can do all vector operations on that. So, I will… what we will discuss in this lecture? We will discuss the basic operations which are very useful for our course.

(Refer Slide Time: 10:29)



So, the first thing we learn is how to add these 2 vectors vector addition. So, what I will do here I will use the triangle rule, so to do that, I will ship this g1 vector in such a way that g1 and g2 get arranged tail to head, that means the tail of g1 will be at the head of g2. And if you remember vector concept, we know that, we can shift the vector anywhere in the space as long as the direction and the magnitude the length of the arrow does not change.

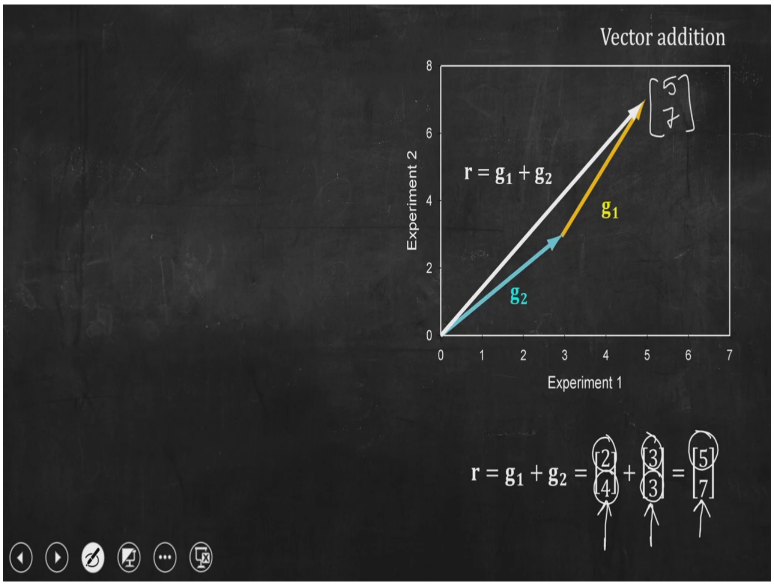
(Refer Slide Time: 11:15)



So, what I will do, I will shift g1 here and its tale is at the tip or head of g2. And then I have drawn a line a new vector white color vector from origin up to the tip of g1, this new vector, the white color 1 is the summation of these two vectors g1 and g2 and I am representing that by R. So, this is the geometric representation or graphical representation of vector addition.

But when you will do the data analysis, you have to do it numerically. So, let us check how I should add these 2 vectors g1 and g2.

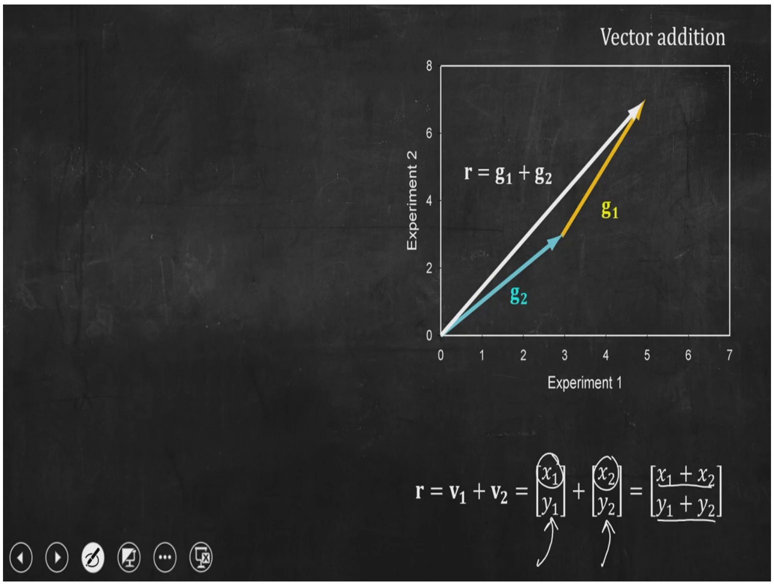
(Refer Slide Time: 12:02)



$$r = g_1 + g_2 = [2\ \ 4] + [3\ \ 3] + [5\ \ 7]$$

And doing that is very simple, I have to add g1 and g2 I know individual vector one is 2, 4 another one is 3, 3. So, take the first element of both the vectors first element of both the vector are 2 and 3, sum them, so I get the first element of the new vector obtain by addition that is 5. Similarly, summed the second element or the second-dimension element or both a vector those are 4 and 3, sum them together, so I get 7. So, my vector new vector obtained by adding g1 and g2 vectors is 5, 7. So, that means this one is 5, 7.

(Refer Slide Time: 12:50)

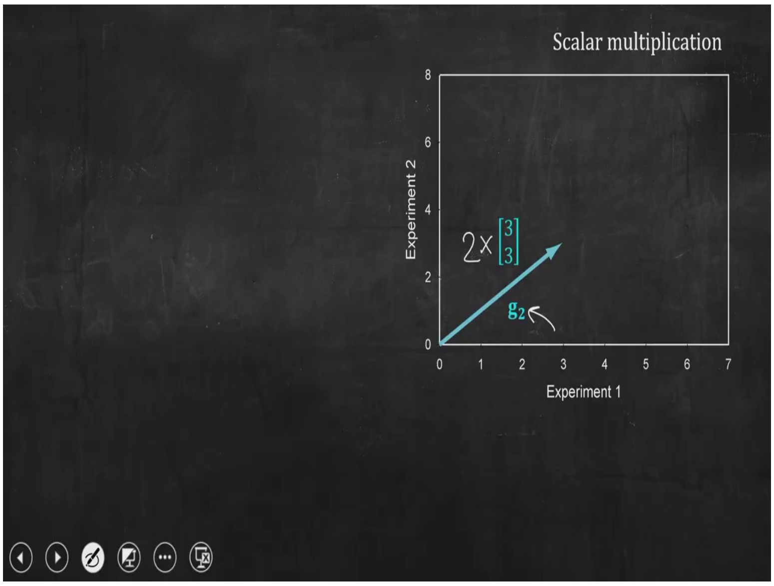


$$r = v_1 + v_2 = [x_1 \; y_1] + [x_2 \; y_2] = [x_1 + x_2 \; y_1 + y_2]$$

Now, in vector multiplication can be tricky. The first 1 that we will learn is called scalar multiplication, scalar multiplication means any normal number 2, 3, 3.6, 4.8 all these or minus 2 something like that these are scalar. Whereas vector has multiple numbers in listed together. So, if I multiply a vector by a scalar number, so what do I get? For example, take the g2 vector again from my data set.

So, it is for gene 2. So, it has 3 and 3 in both the experiments, so, it is a 3, 3 vector, I want to multiply it by 2, 2 is a scalar. So, 2 into 3. So, what will happen? In this case, the vector will get stretched, but it should not move it from his direction. So, it will maintain its direction but it will get stretched longer, it become longer twice because I am multiplying by 2.

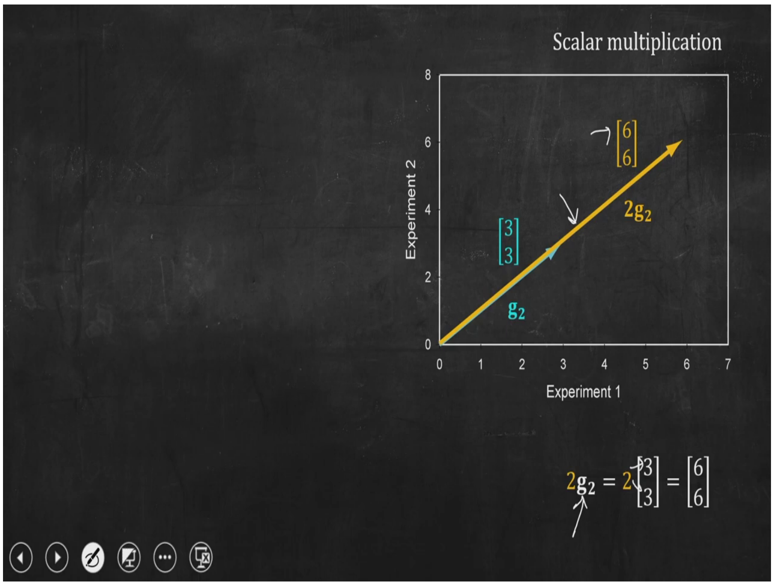
(Refer Slide Time: 13:38)



Now, in vector multiplication can be tricky. The first 1 that we will learn is called scalar multiplication, scalar multiplication means any normal number 2, 3, 3.6, 4.8 all these or minus 2 something like that these are scalar. Whereas vector has multiple numbers in listed together. So, if I multiply a vector by a scalar number, so what do I get? For example, take the g2 vector again from my data set.

So, it is for gene 2. So, it has 3 and 3 in both the experiments, so, it is a 3, 3 vector, I want to multiply it by 2, 2 is a scalar. So, 2 into 3. So, what will happen? In this case, the vector will get stretched, but it should not move it from his direction. So, it will maintain its direction but it will get stretched longer, it become longer twice because I am multiplying by 2.

(Refer Slide Time: 14:37)

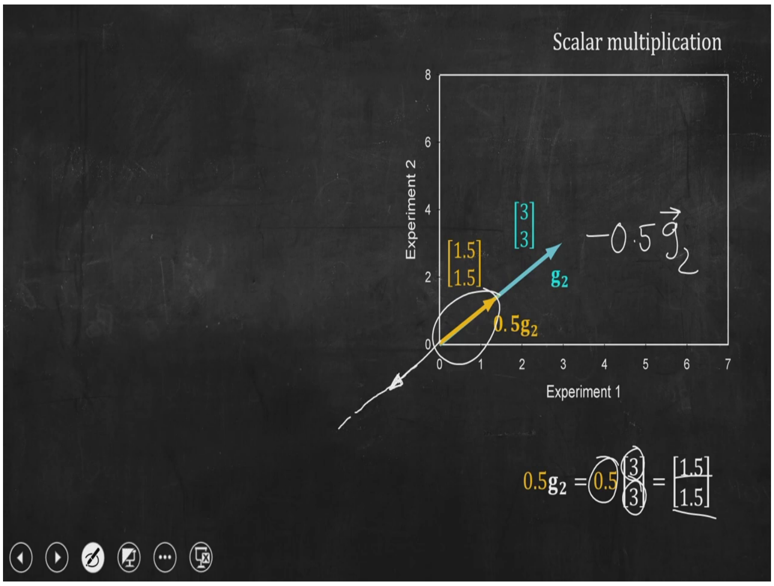


$$2g_2 = 2[3\ \ 3] = [6\ \ 6]$$

So, this is what has been shown here. So, this yellow arrow is my new vector obtained by scalar multiplication of g2 by 2 and obviously, it is 6 and 6. So, what I have done? I have multiplied g2 by a scalar 2 and then, to do that, what I have done? I have multiplied both the element by 2, so I have got a new vector 6 and 6.

Now, suppose I multiply the same vector with a fraction, a whole number or a number greater than 1 will give me stretching of the vectors. So, if I multiply with a number which is smaller than 1, then my vector will get squished. So, suppose I multiply with 0.5 half, so I am multiplying g2 by half. So, what will happen?

(Refer Slide Time: 15:30)



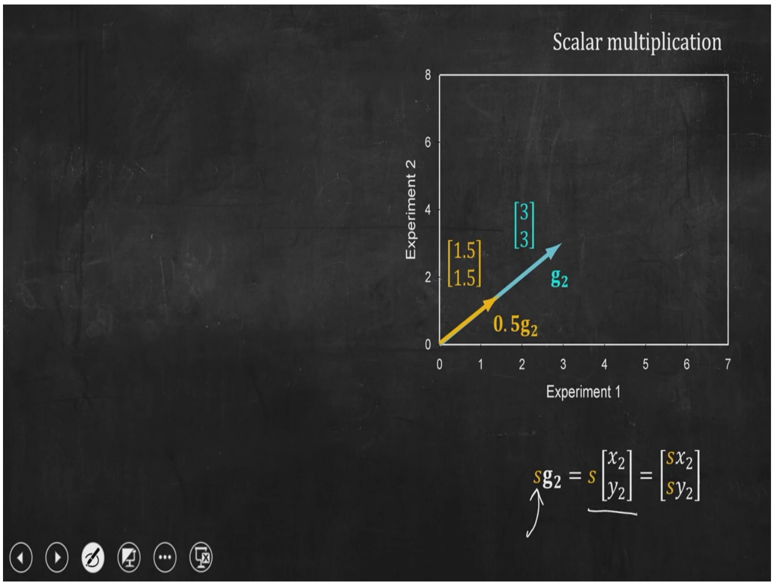
$$0.5\,g_2 = 0.5[3\ \ 3] = [1.5\ \ 1.5]$$

Now, I multiply 3 and 3 by half, that is 0.5. So, I get 1.5 and 1.5. So, this new yellow arrow is my new vector, obtained by scalar multiplication of g2 by 0.5. Now, what if I multiply by a negative number, just think over it is, yes again it will be a stretching or you know compression of the vector. But now the direction will flip.

If I multiply by suppose if I multiply by minus 0.5 or half g2, I will put a arrow to represent vector because I cannot make it bold here in the screen. So, then what will happen? I will get a vector which will be in this direction, remember it will be on the same straight line, but its direction is now flipped, because I have multiplied by minus, so minus number.

So, it is minus is giving the direction plus and minus are for direction, whereas the scalar value is how much it will be stretched, or how much it will be squeezed. So, it will get squeezed, but it will get flipped in the opposite direction. In this particular example, where we are talking about a fold change in gene expression, a negative value does not make any sense, physical sense, but just to explain what will happen, I have shown this example.

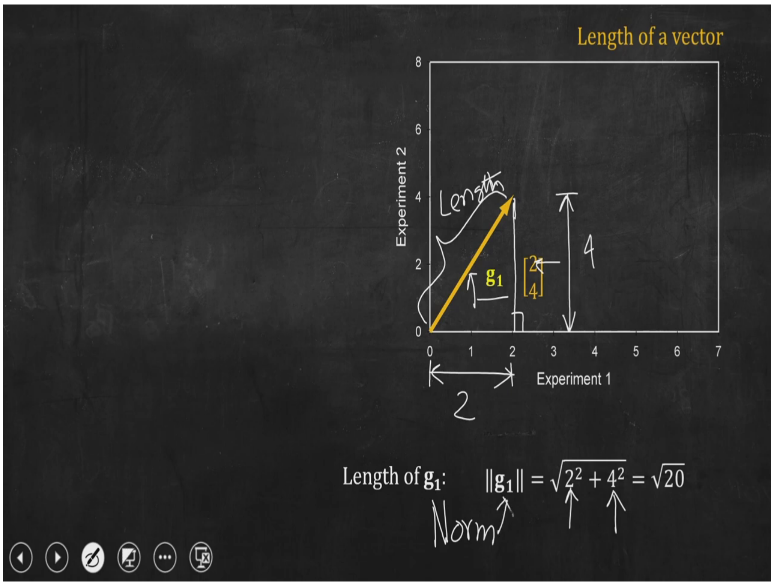
(Refer Slide Time: 16:56)



$$sg_2 = \begin{vmatrix} s\,x_2 \\ s\,y_2 \end{vmatrix}$$

So now, in general, what I am doing? If I have to do a scalar multiplication, I take a scalar and I multiply that vector, so each element, each row again multiplied with that scalar.

(Refer Slide Time: 17:08)



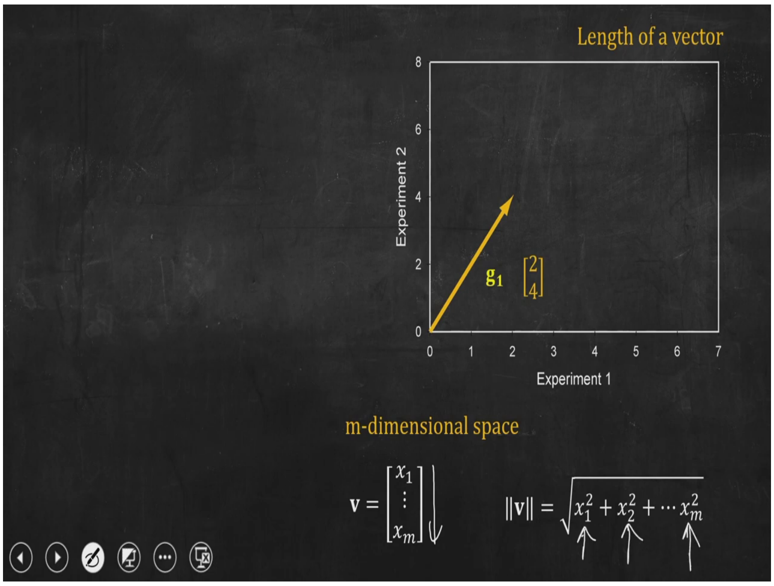
$$||g_1|| = \sqrt{2^2 + 4^2} = \sqrt{20}$$

Now, we have learned addition, we have learned scale multiplication. Now, what we will learn is to how to get the length of a vector. Remember, when I am doing multiplication, the length of the vector is changing not its direction. So, how do I measure the length of a vector, take the example for g1, which is 2, 4, how can I measure the length?

By length I mean, this distance, this is the length, length of g1 vector. So, how can I get it? I can simply use the geometry. So, I can draw a line which is vertical. So, what is this distance This distance is 2 because this is 2, 4. So, this horizontal distance is 2, whereas this vertical distance is 4.

So, simply by Pythagorean theorem, what I know this distance, this yellow line, length of the yellow line will be square root of 2 square plus 4 square. So, the length of my vector g1 for gene 1 is square root of 20 square root of 2 square plus 4 square. So, usually the length of a vector is also called a norm. So, length of a vector is also called a norm of that vector. And we represent them like this, where the vector is placed between 2 double line or 2 single lines.

(Refer Slide Time: 18:51)

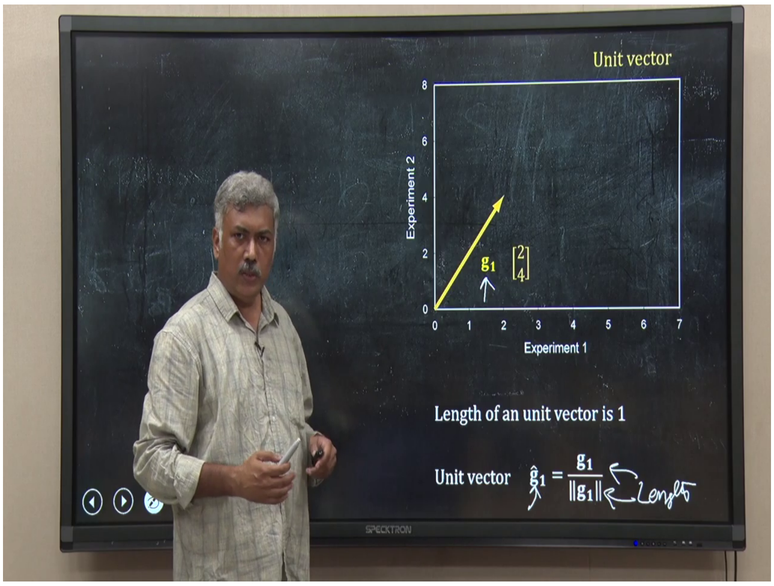


$$v = \begin{vmatrix} x_1 \\ : \\ x_m \end{vmatrix}$$

$$||v|| = \sqrt{x_1^2 + x_2^2 \cdots x_m^2}$$

Now, if I have m dimensional data, the g1 g2 the experimental data that I am showing here in the plot are 2-dimensional, I cannot plot m dimensional data, I cannot have plotted data of where I have 10-dimension or 5-dimension something like that, but you can imagine what will happen in m dimension, so m dimensional vector will have m element, m rows. So, the normal length of that vector V will be square root of summation of x1 square, x2 square to xm square, where x1, x2, xm are the elements in the vector.

(Refer Slide Time: 19:26)



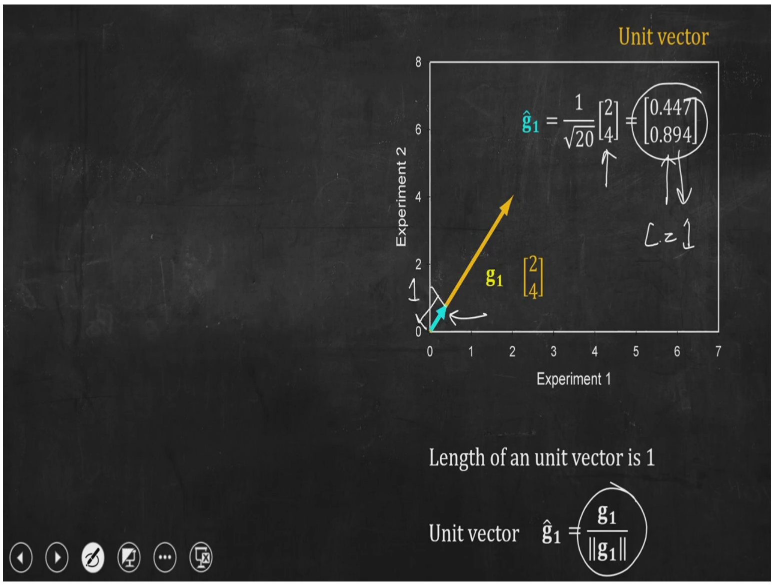
$$\hat{g}_1 = \frac{g_1}{\|g_1\|}$$

Now, once I have got this length of a vector, I can actually define something very useful. That is called unit vector. See, the most of the time we are actually not bothered about the length of the vector, but rather direction of the vector. In many cases in data analysis also, you will realise that I am actually not bothered about the length, but rather of the magnitude of the vector but rather the direction of the vector or rather the direction of the data point in the space.

So, now what we can do? We can normalise that vector, the original vector with respect to its length, so that I will get a new vector whose length will be 1, but it will retain its direction. So, listen carefully what I am trying to do, I want to create a new vector, which will have the unit length, but its direction will be same as the original vector.

So, I am doing normalisation with respect to the length of the original vector. So, if you have give me g 1, which is 2, 4 vector coming from my data, then the unit vector for that which can be written as g1 hat to represent that is a unit vector is a normal convention, what I will do? I will divide this g1 vector, original vector g1 by the norm, or the length, length of this vector.

(Refer Slide Time: 20:58)



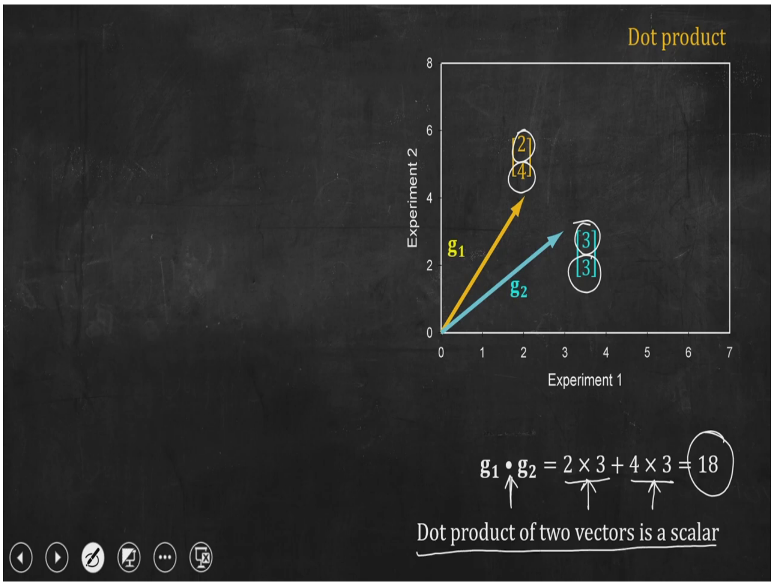
$$\hat{g}_1 = \frac{g_1}{\|g_1\|}$$

So, let me do it for this particular example. So, the length of my vector g1 was square root of 20. Just to slide back we did the calculation right, and the original factories 2, 4. So, I will multiply or rather divide 2, 4 by square root of 20, because that is my definition here. So, by doing that, I get a new vector of 0.447 and 0.894.

Now, if you do use the formula to calculate the length of a vector, you will find the length of this will be equal to 1, as I have truncated the data, it will not be exactly 1, but it will be close to 1. So, I have drawn this vector here, you can see, this vector has now unit length, the length of this vector is 1, but it is in the same direction as the original vector g1.

So, in this way many a time in data analysis, what we do? That we convert unit vectors for each of these data points, so, then there all lengths are same, but the directions are different. So, then it becomes very easy to compare one vector with another vector.

(Refer Slide Time: 22:10)



$$g_1 + g_2 = 2\times3 + 4\times3 = 18$$

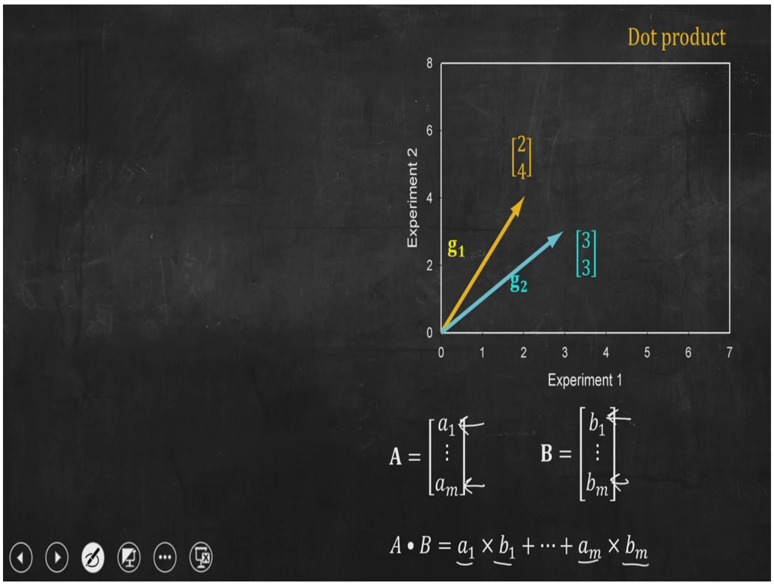
Now, previously we discuss about scalar product scalar multiplication, what I am doing in that? I am multiplying a vector with a scalar. Now, cannot I multiply a vector with a vector? Yes, I can do that, there are many types of multiplication, where one vector is multiplied with another vector, and they have individually different meanings. What I will discuss in this lecture is what is called dot product because this is mostly used in data analysis.

So, we will discuss dot product between two vectors. And again, I have taken the same example of gene 1 gene 2 vector. So, these yellow and blue lines are gene 1 and g2, and I want to multiply g1 and g2 vectors. So, what I will do? What I will do is, that I will take the first element of g1 and multiply that with the first element of g2, same row coming from the same experiment in our data example.

Then I will sum that with multiplication of the second element of g1 with the second element of g2. So, that is what I have done here. So, these are 2 and 3 are the first element of the both the vectors, I multiplied them together, and 4 and 3 is the second element of my both the vectors g1 and g2, I multiplied them and sum them. So, 2 into 3 plus 4 into 3, I get 18.

So, this is the dot product of g1 and g2, and how we write it, we write it with a big dot between these this vector, we do not put a simple dot or a crossed to represent that. Because, here we have to put a big dot to represent that this is a dot product. Now, if you notice, g1 and g2 are vectors, but when I do the… do the dot product of them what I get? I get a scalar. So, dot product of 2 vectors gives me a scalar quantity.

(Refer Slide Time: 24:17)



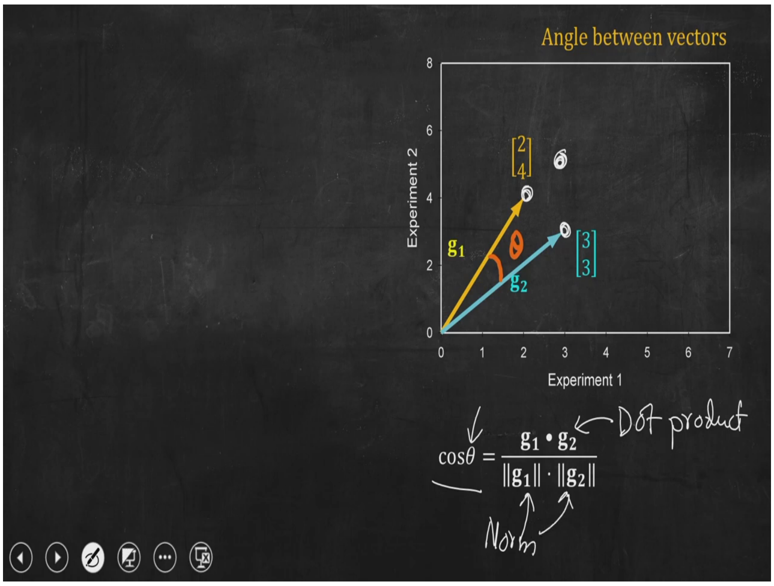
$$A = \left[a_1 \ldots a_m\right]$$

$$B = \left[b_1 \ldots b_m\right]$$

$$A \cdot B = a_1 \times b_1 + \cdots + a_m \times b_m$$

Now, what will happen if I have m dimensional data, in that case, each vector will have m element m rows starting from suppose A and B starting from a1 to am and for vector b it start with b1 up to bm. So, the principle of dot product will remain same. Earlier I had two elements two rows, now I have m rows. Now, I have to multiply m rows. So, first row a1 and b1 will be multiplied, last row am and bm will get multiplied and in between of these sums we will have the other rows multiplication, and I will get the dot product.

(Refer Slide Time: 24:56)



$$cos\Theta = \frac{g1 \cdot g2}{||g1|| \cdot ||g2||}$$

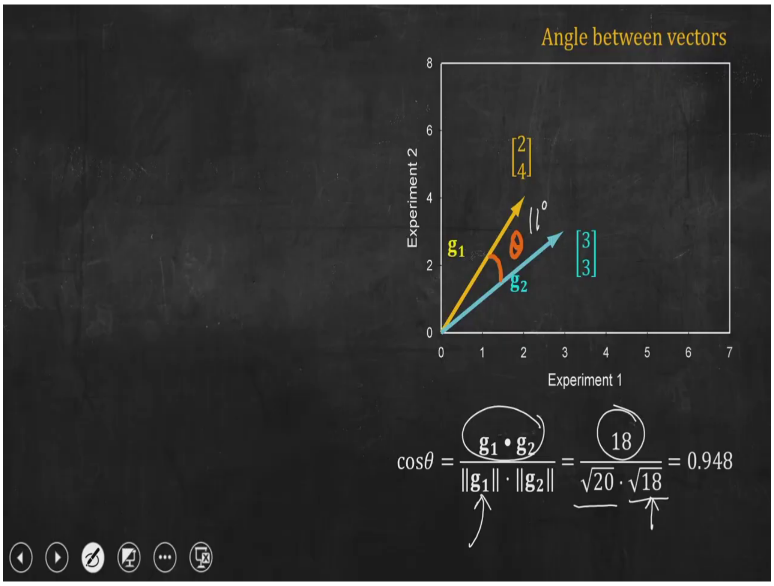
Now, this dot product of 2 vectors is very useful to calculate angle between two vectors. Now, you may wonder, why do I need to measure angle between two vectors am I doing geometry or physics here in this course, No. So, suppose I want to know I have multiple data points suppose I have these data points here, I have this data point here and represented by vectors I have another data point here.

So, I can draw a another vector for that. So, now angle between any pairs of vector can tell me how close they are they are with respect to each other in the space. So, that is why many a time measuring angle between vectors is very critical. So, for example, in this case, I want to measure the angle between g1 and g2 and suppose that angle is theta.

So, how do I do that? You can prove it, but you do not need to go into details of that I have written down the cos theta, where theta is the angle between g1 and g2 can be shown to be equal to the dot product, remember this is dot product, this is dot product of g1 and g2 divided by the norm or the length of g1 and g2 they are scalar. So, they are more simple multiplication of scalar.

Now, g1 and g2 dot product is a scalar we are dividing that by norm two norms, norms are scalar. So, essentially, we are dividing scalar by a scalar and cost theta obviously will be a scalar quantity I should get a number. So, let me apply this rule for this particular data set.

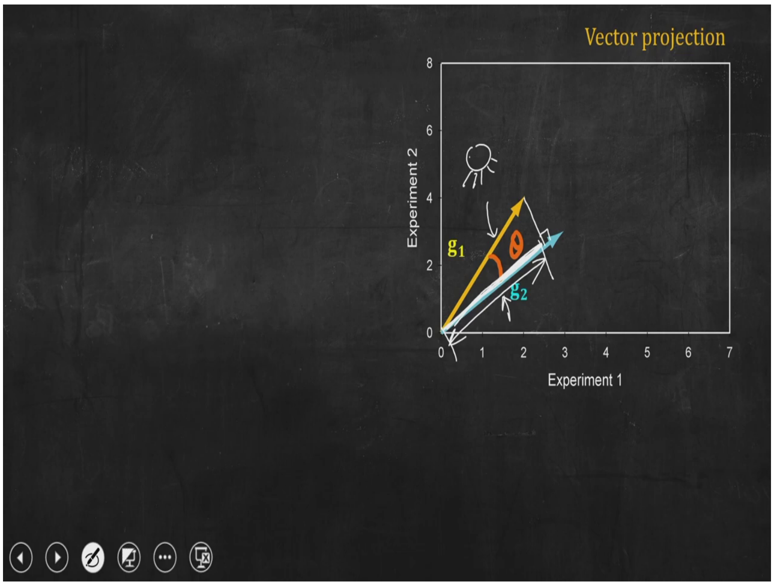
(Refer Slide Time: 26:41)



$$\cos\Theta = \frac{g1 \cdot g2}{||g1|| \cdot ||g2||} = \frac{18}{\sqrt{20}\cdot\sqrt{18}} = 0.948$$

So, what do I get? Just a few slides back we have calculated the dot product. So, that is a team length of g1 the norm of g1 square root of 20 and the length of g2 is square root of 18. And if you just do a simple arithmetic I get cos theta is equal to approximately 0.948 And if you do the inverse of that possibility will be around 10.6 or 10.7 or roughly 11 degree. So, this is maybe roughly 11 degree. So, in this way, I can use the dot product and the norm and the ratios to calculate the angle between 2 vectors. And as I said many times that angle is a very good measure for closeness of data points and for other analysis also.

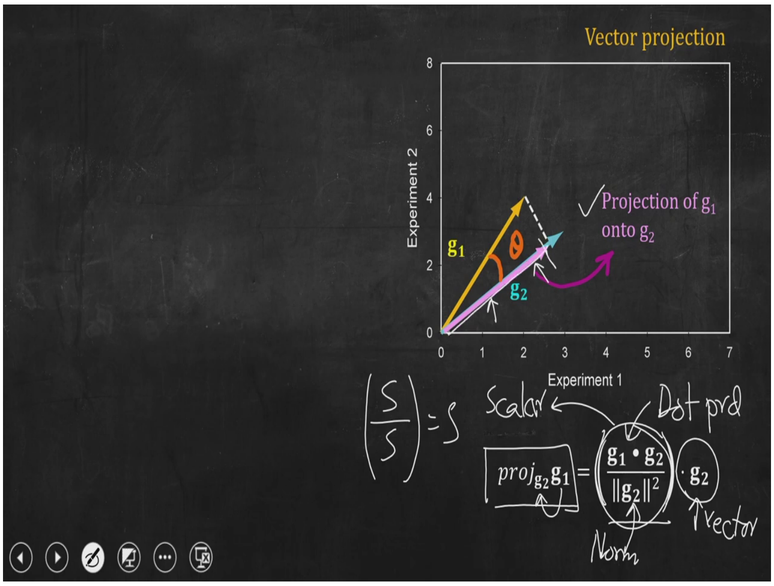
(Refer Slide Time: 27:27)



Now, one important idea which is recurrently used in data analysis, particularly regression is that projection of a vector on another. What do I mean by projection? Suppose, what I want that I have a light here and I want to see the shadow of this yellow vector g1 on my vector g2. So, that is the projection of g1 on g2.

So, how can I draw it? I can simply draw a line from the tip of g1 on g2 which will be perpendicular to g2. So, then this part which I am marking in white must be the shadow of g1 on g2. So, in mathematical term linear algebra what I will say this white thing is white line which is alone the vector g2 is the projection of g1 onto g2. Now, remember we are not bothered about just any line, we are talking about vectors. So, projection can also be a vector. So, that vector will have a length this one, but it will have the direction of g2.

(Refer Slide Time: 28:58)



$$\text{proj}_{g2}g1 = \frac{g1 \cdot g2}{||g2||^2} \cdot g2$$

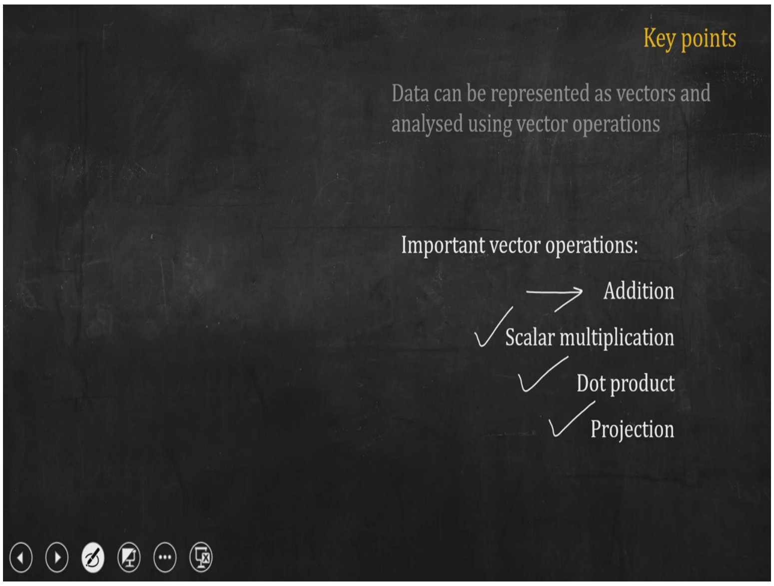
So, let me draw it cleanly. So, this pink line is the projection of g1 on g2. Now, I want to calculate that vector, I want to calculate what will be the projection of g1 on g2, here you can use the idea of angle between 2 vectors and the dot product and what you will end up is the projection of g1 onto g2 please notice how we write it.

So, projection of g1 onto g2, g2 is subscript. So, that will be equal to the dot product, dot product of g1 and g2 divided by the square of norm or length or magnitude of g2 into the g2 vector remember this is vector. Now, look carefully what we have in these first bracket thing? g1 dot g2 dot products, dot products are scalar I am dividing by square of norm length, the square of length, length is scalar square of that length is also scalar. So, I have scalar versus scalar in the first part, so, that is also a scalar that means this whole thing is a scalar quantity.

Now, you are multiplying g2 which is the vector by a scalar quantity. So, you are now getting another vector, which is a projection vector of g1 onto g2. So, that vector is the pink one, in this particular example, the scalar quantity is such, this scalar quantity is such that the g2 got squished, it becomes smaller and its length has supposed to reduce like this.

In some case, it may get stretched also it depends upon the vector g1 and g2 their relative size and other did and the angle between them also. So, projection is nothing but actually converting one vector and moving it on to the other one.

(Refer Slide Time: 31:19)



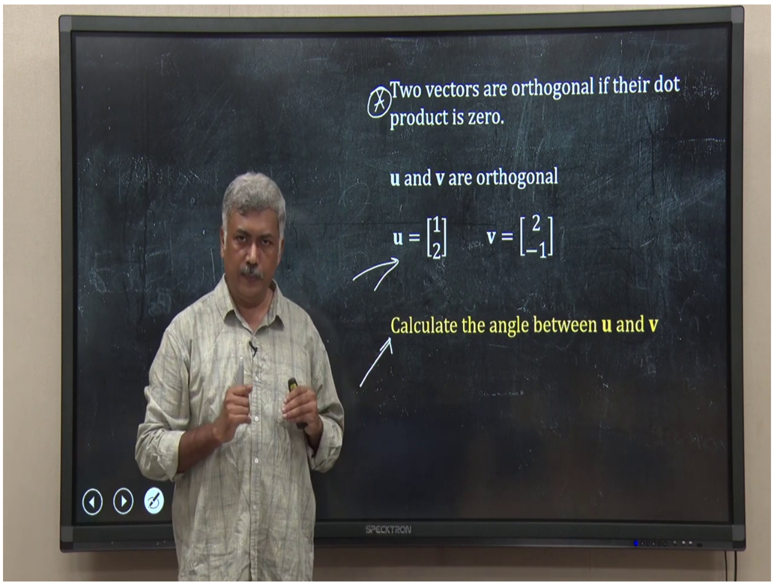
So, that is all for this lecture, what we have learned in this lecture, let me jot them down. The first thing that we have learned in this lecture is that, I can many a time represent my data as vectors. As we have seen in this discussion today that we have a genic expression data, data for each gene is represented as a vector.

So, I have vectors for each gene, repeatedly I have used two vectors for gene 1 and gene 2, if you have 2000 genes in that experiment, you will have 2000 vectors. Now, once you have your data in vectorial form as vectors, then you can use vector operations and vector arithmetic on those vectors to analyze the data and extract some meaning out of it.

And so, what we have learned in this lecture in today's lecture are few basic operations of vectors which will be very useful for our data analysis course, we have learned how to add vectors, then we have learned scalar multiplication, as multiplication are of different types are possible for the vectors.

So, we have learned only dot product because that is very useful for our work. Then from dot product, we can also know the angle between two vectors. And eventually, we use the concept of angle and dot product to get the projection of one vector onto another vector. That is all for this lecture. Before I leave you, I leave you with a problem to solve.

(Refer Slide Time: 32:44)



$$u = [1 \;\; 2]$$

$$v = [2 \;\; -1]$$

So, suppose I have two vectors u and v, I have written them 1, 2, and 2 minus 1 and it is said that they are orthogonal. What do I mean by orthogonal? Two vectors are orthogonal when they are dot product is 0. So, what I am telling you is that I have two vectors, u and v. And it has been told that they are orthogonal that means their dot product is 0. What do you have to calculate? You have to calculate the angle between these two vectors, u and v. Think about it is very simple to solve. Try it yourself without consulting any book or tutorials. Till then, happy learning.