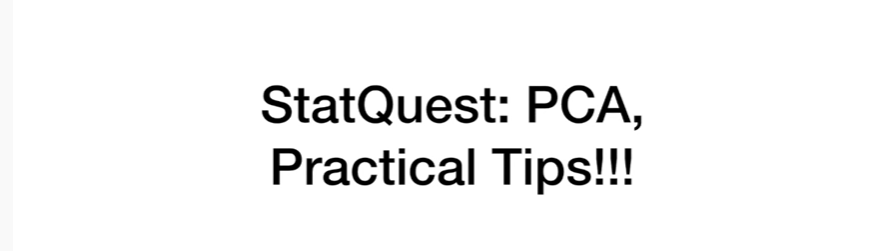
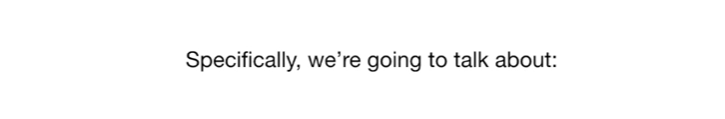
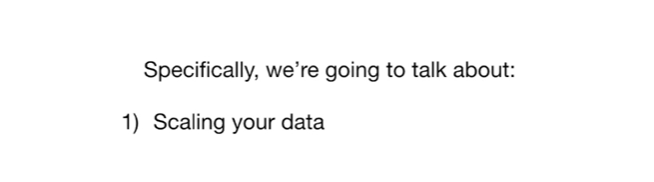
<https://www.youtube.com/watch?v=oRvgq966yZg&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=26>



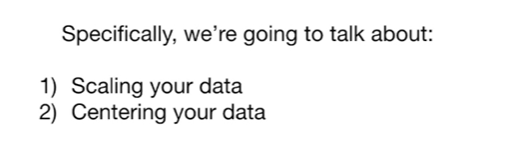
Today we're gonna be talking about pca and i'm gonna give you a few practical tips.



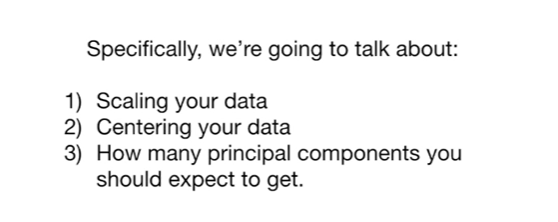
Specifically we're going to talk about :



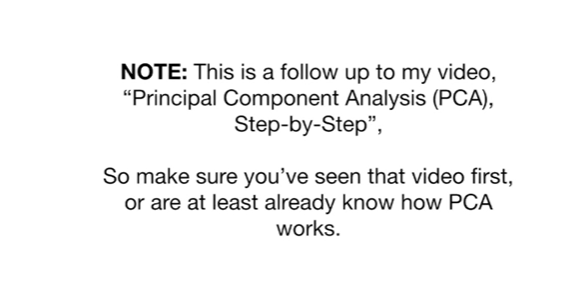
1. scaling your data



2. centering your data

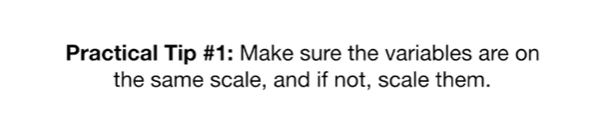


and 3. how many principal components you should expect to get.

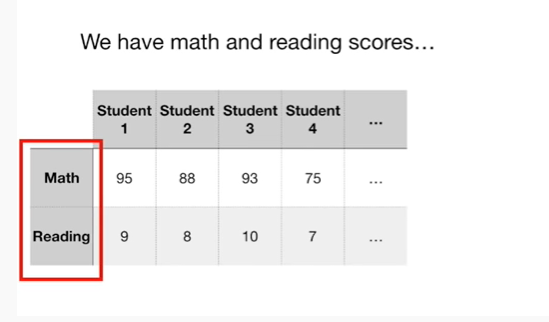


Note : this is a follow up to my video principal component analysis PCA step by step.

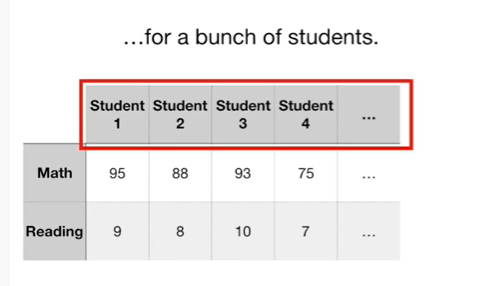
So make sure you've seen that video first or at least already know how PCA works.



Practical tip number 1 : make sure the variables are on the same scale and if not scale them.



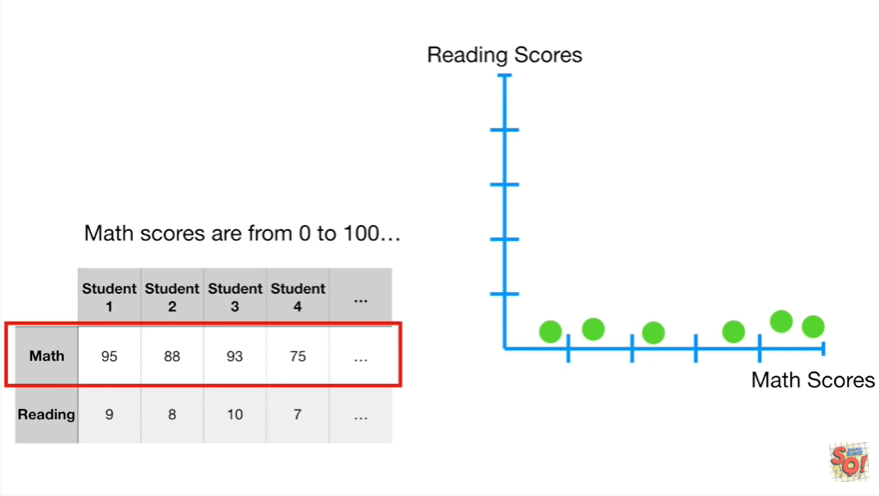
Here we have math and reading scores.



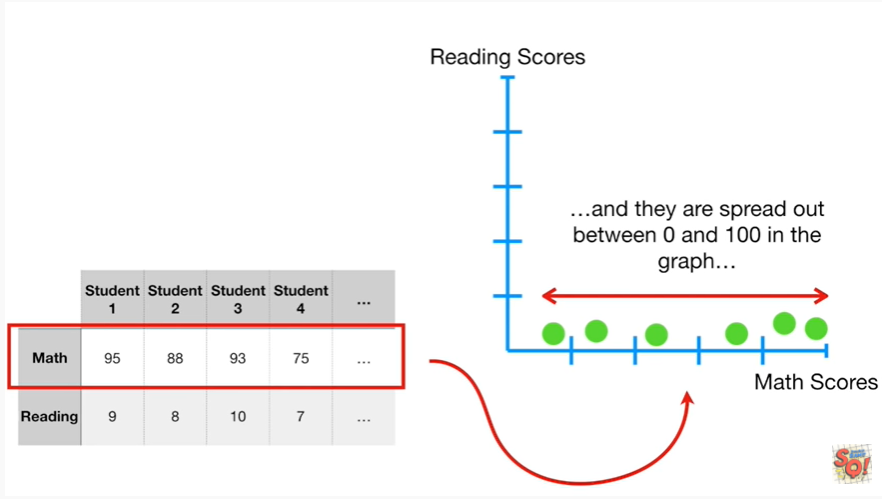
For a bunch of students.



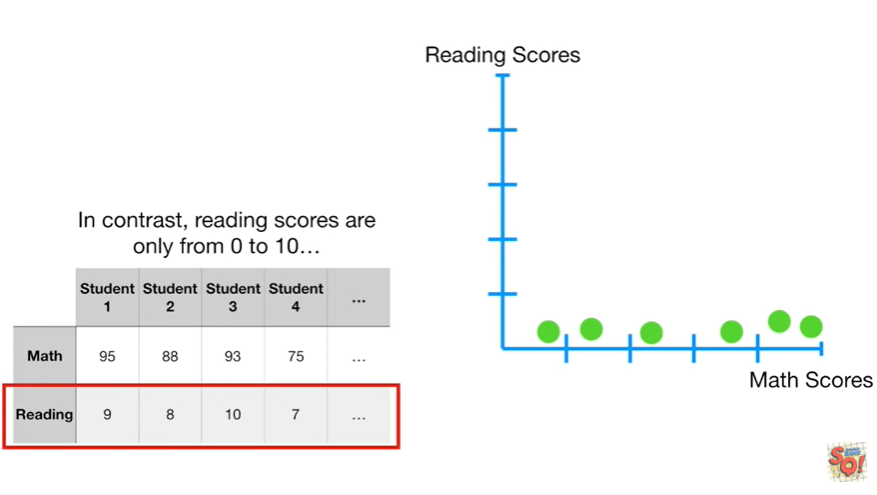
Let's plot the data.



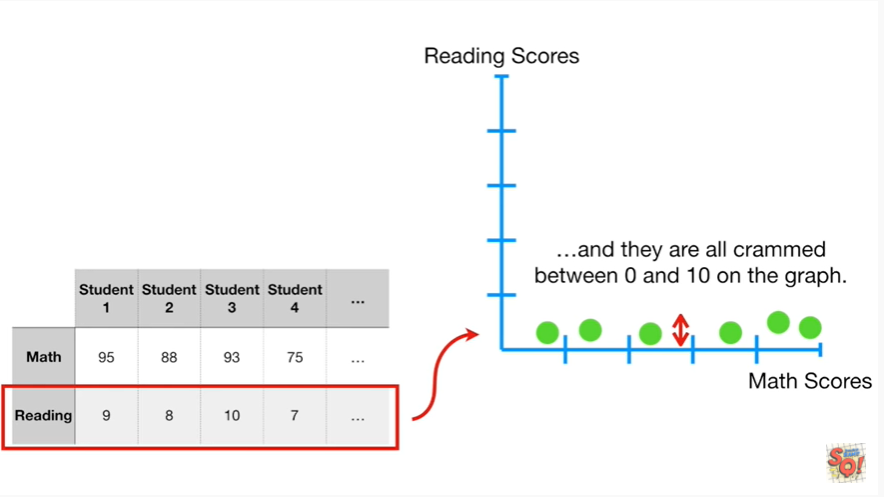
Math scores are from 0 to 100



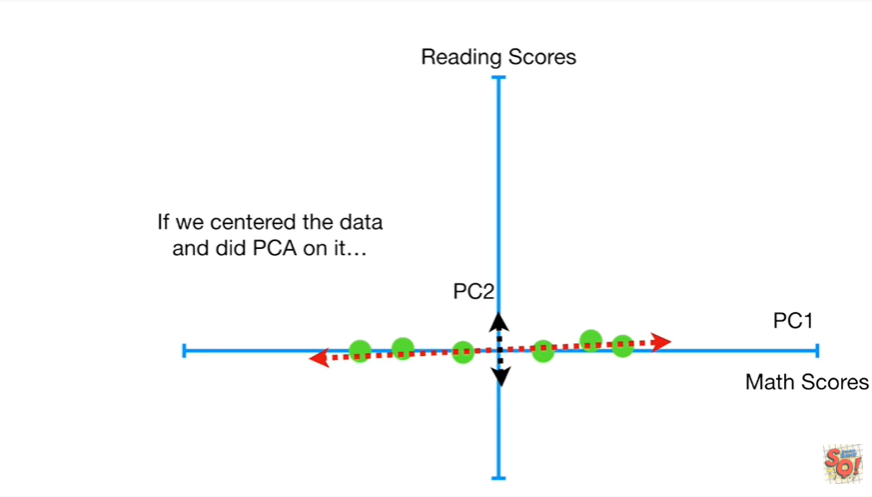
and they are spread out between 0 and 100 in the graph.



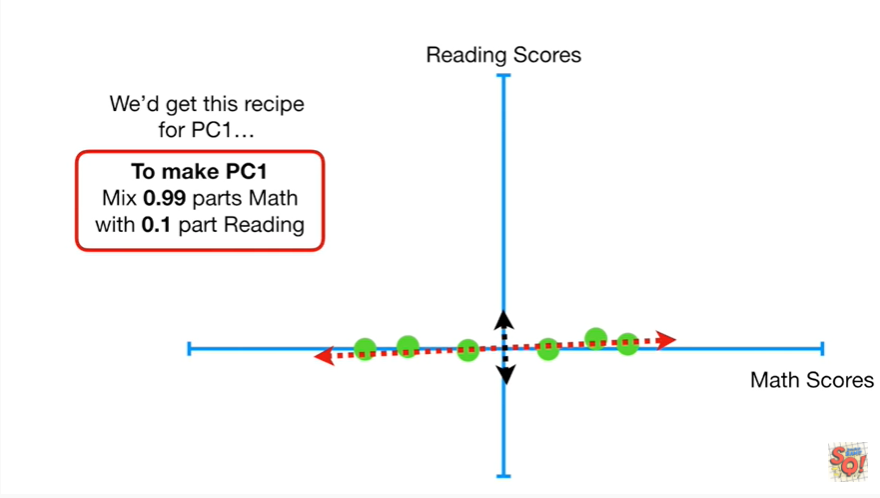
In contrast reading scores are only from 0 to 10



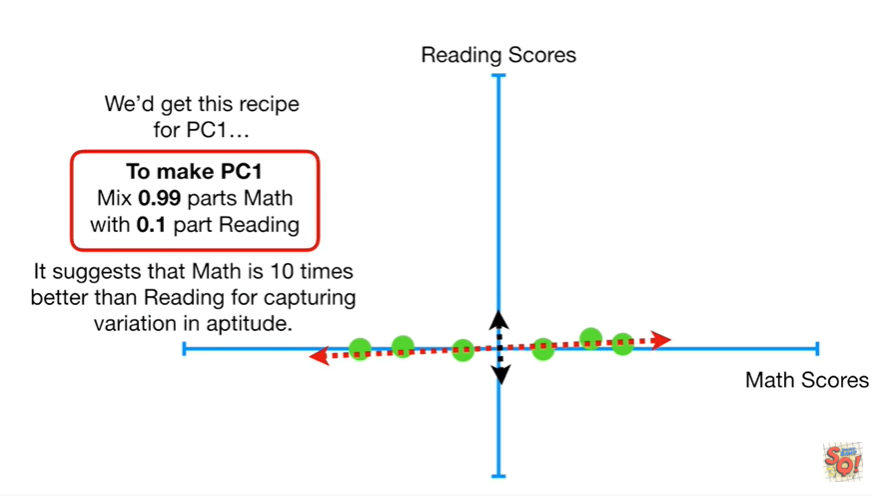
and they are all crammed between 0 and 10 on the graph.



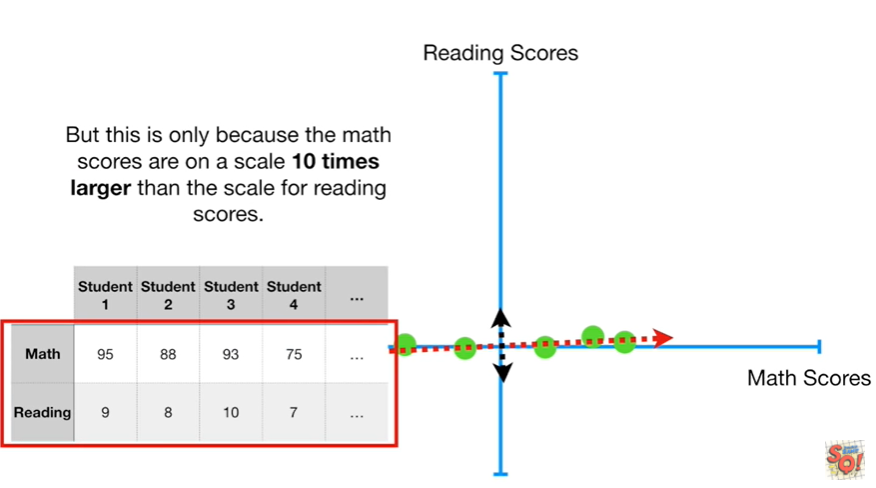
If we centered the data and did PCA on it



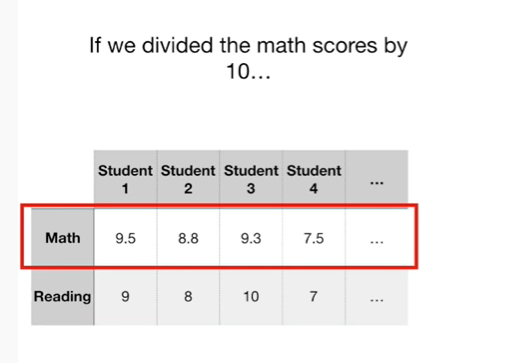
we get this recipe for PC 1 to make PC 1 makes 0.99 parts math with zero point one part reading



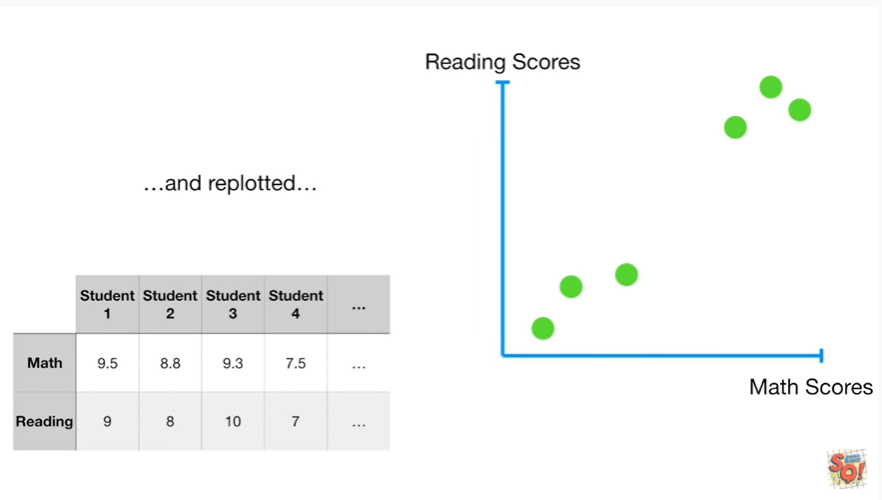
It suggests that math is 10 times better than reading for capturing variation in aptitude.



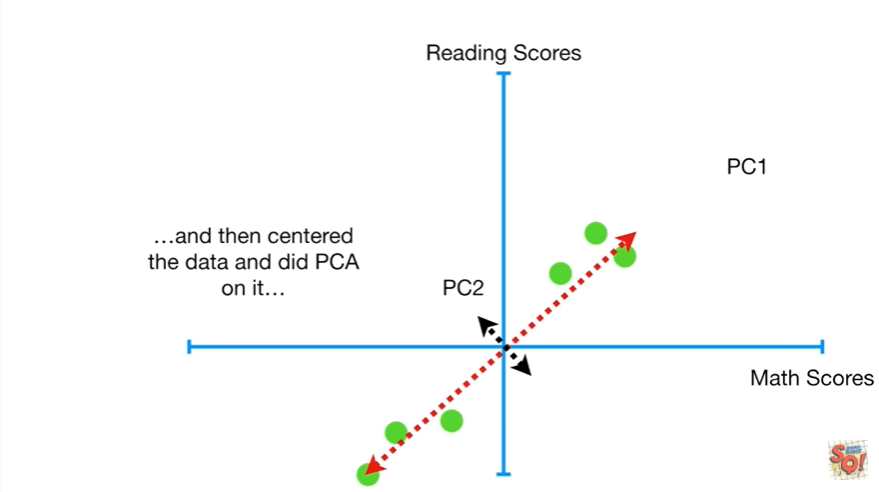
But this is only because the math scores are on a scale ten times larger than the scale for reading scores.



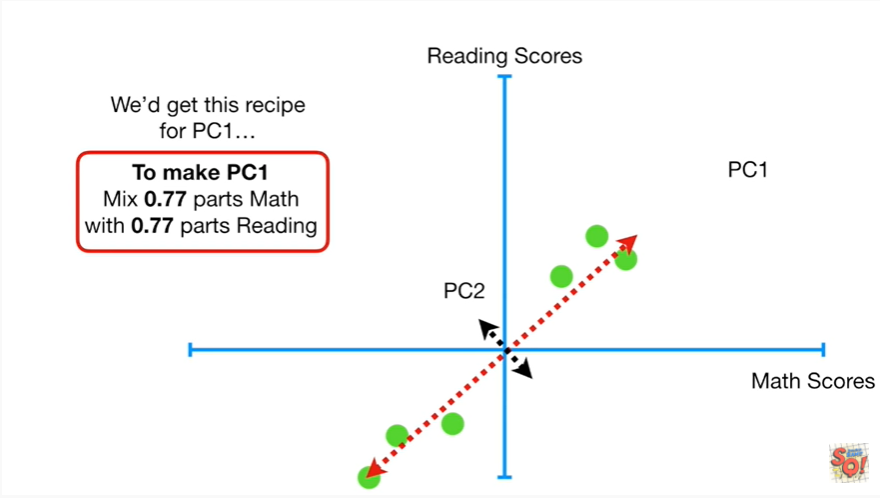
If we divided the math scores by 10



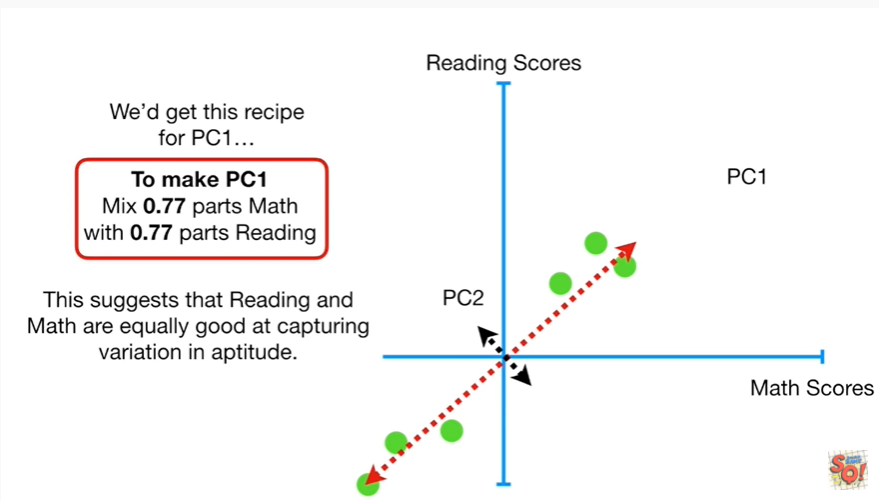
and re plotted



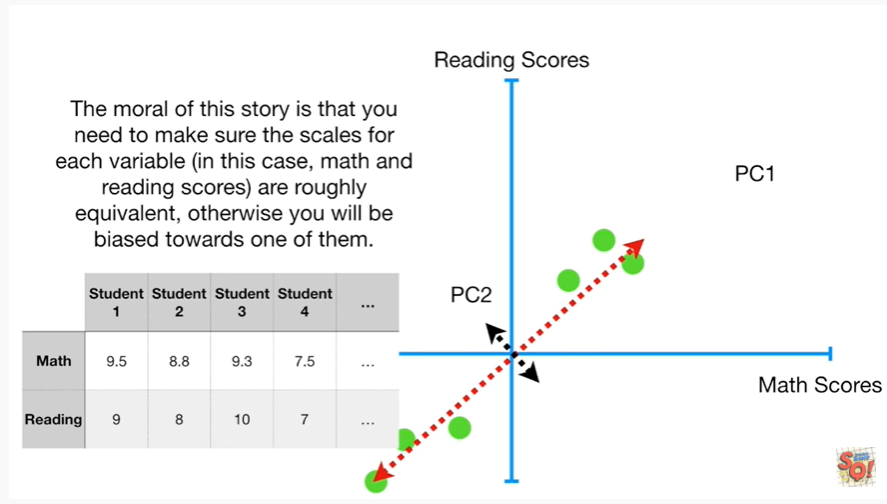
and then Center the data and did PCA on it.



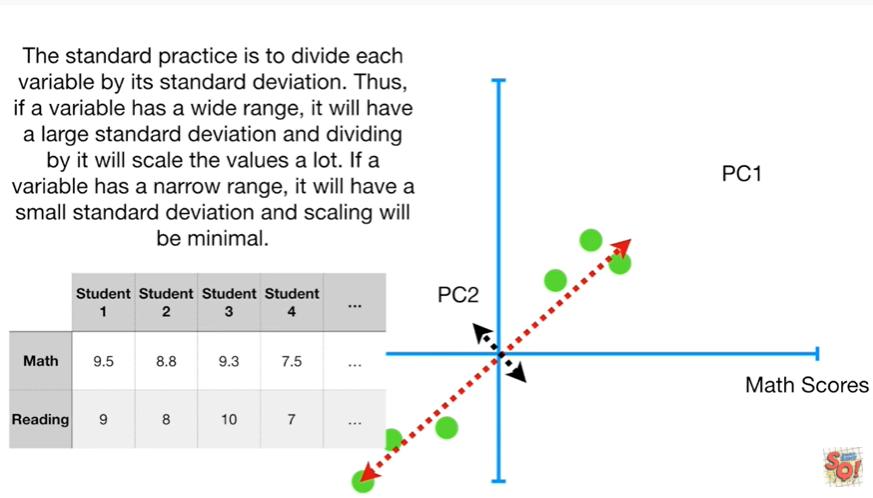
We'd get this recipe for PC 1 to make PC 1 mix 0.77 parts math with zero point seven seven parts reading.



This suggests that reading and math are equally good at capturing variation in aptitude.

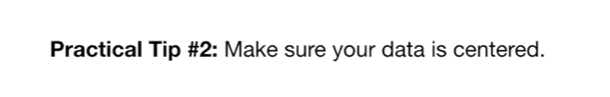


The moral of this story is that you need to make sure the scales for each variable (in this case math and reading scores) are roughly equivalent otherwise you will be biased towards one of them.

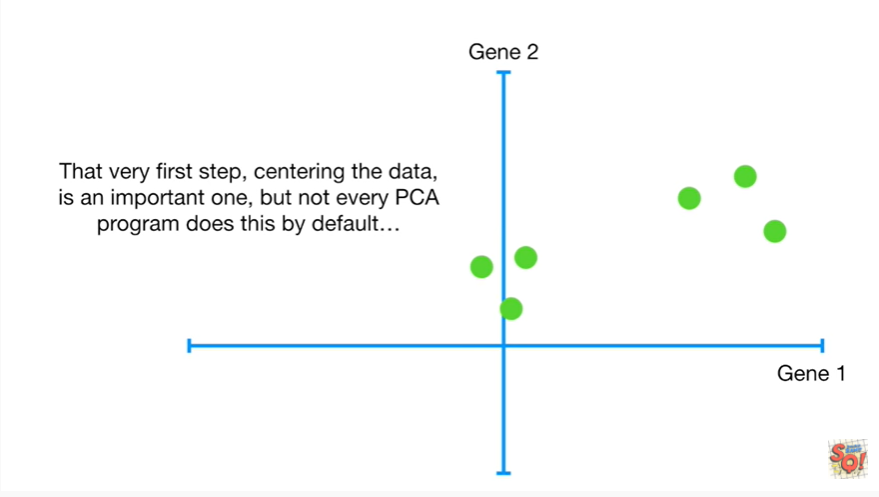


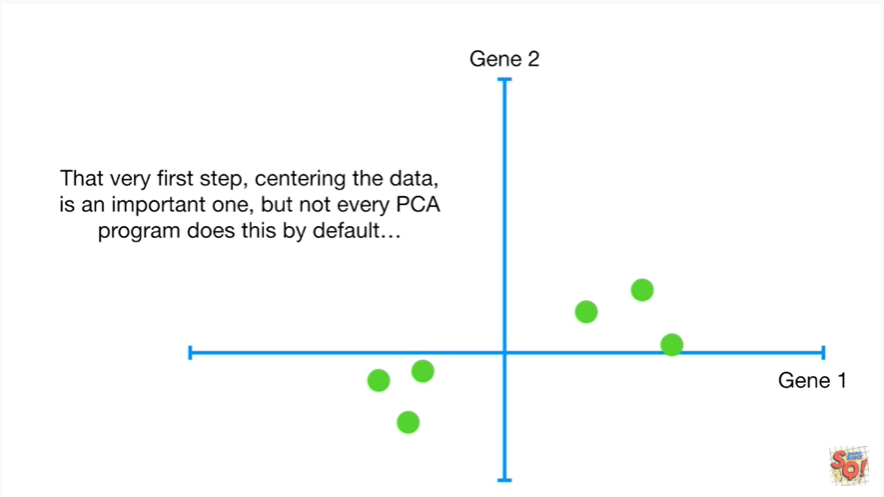
The standard practice is to divide each variable by its standard deviation.

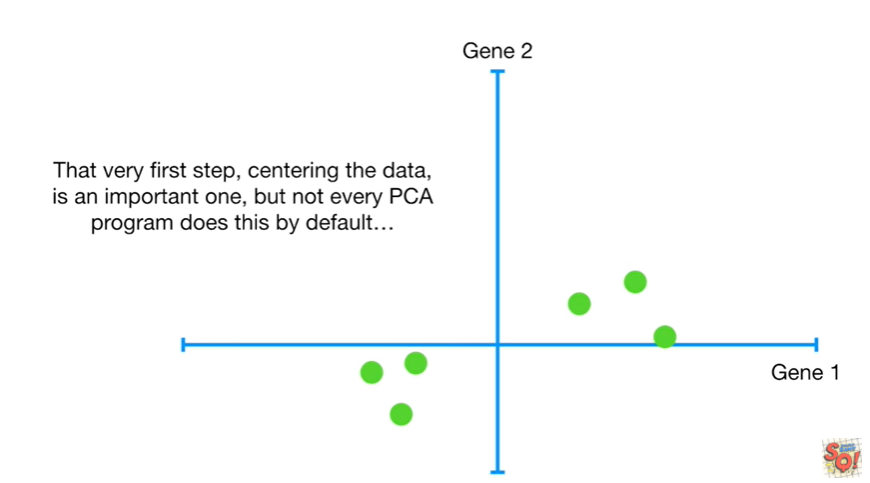
Thus if a variable has a wide range it will have a large standard deviation and dividing by it will scale the values a lot if a variable has a narrow range it will have a small standard deviation and scaling will be minimal.



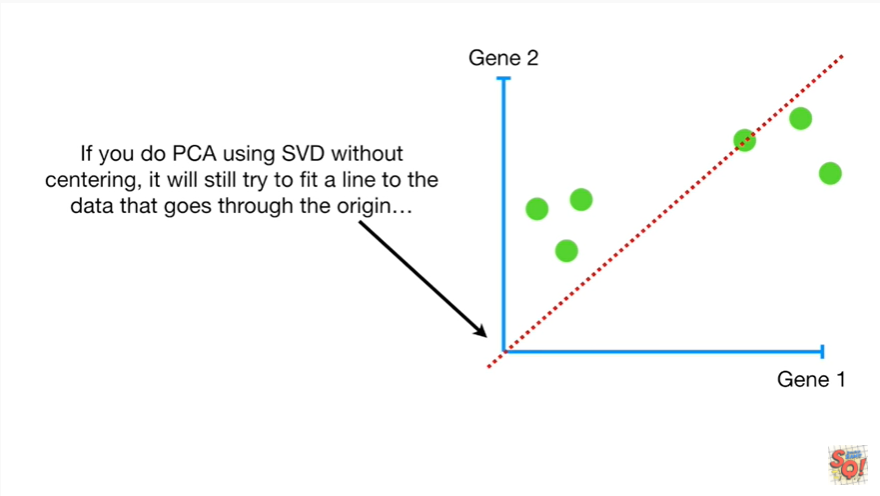
Practical tip number 2 : make sure your data is centered.



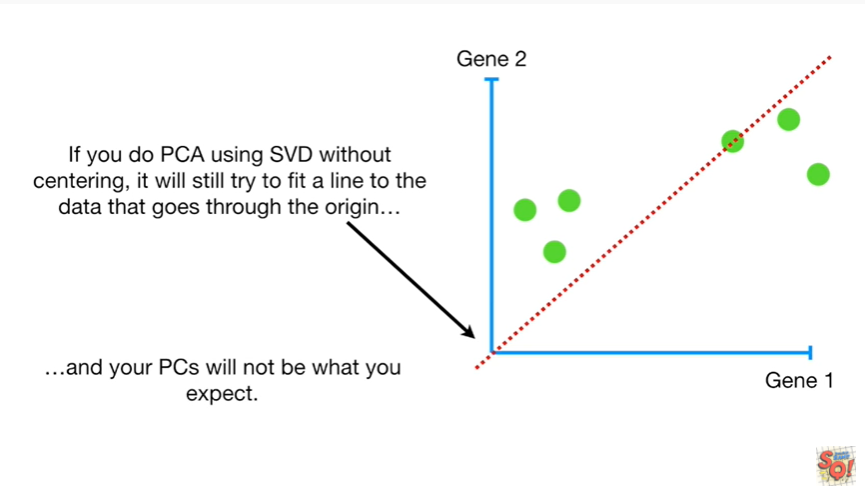




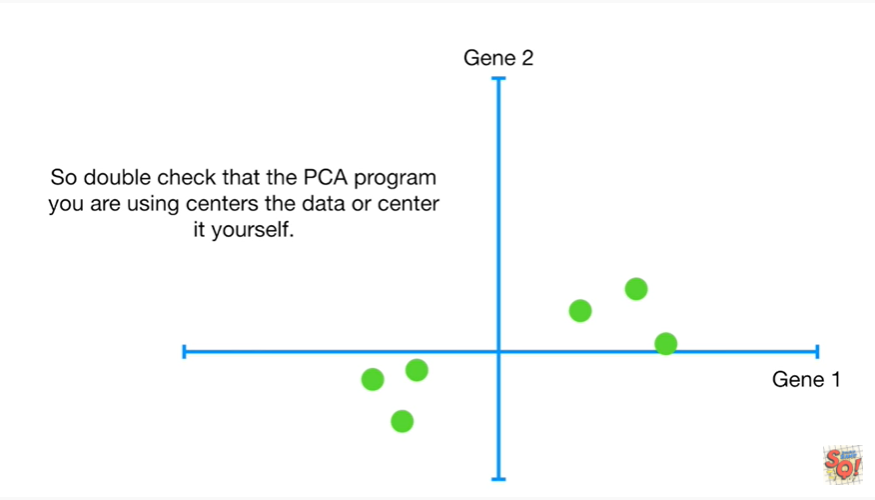
The very first step centering the data is an important one but not every PCA program does this by default



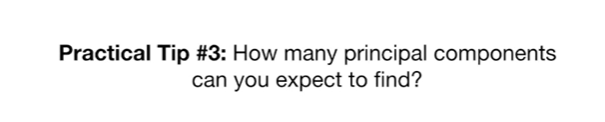
If you do PCA using SVD without centering it will still try to fit a line to the data that goes through the origin



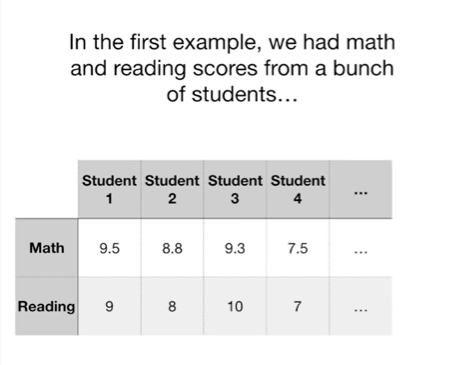
and your pcs will not be what you expect.



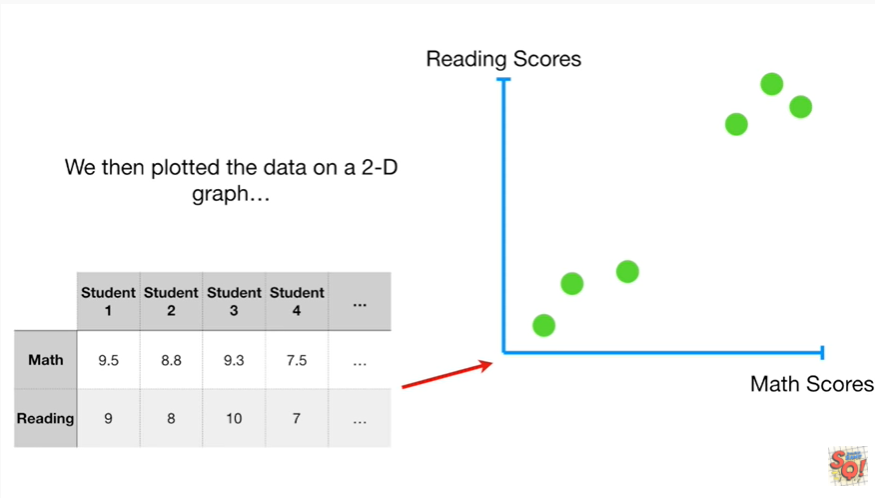
So double-check that the PC a program you are using centers the data or center it yourself.



Practical tip number 3 : how many principal components can you expect to find ?



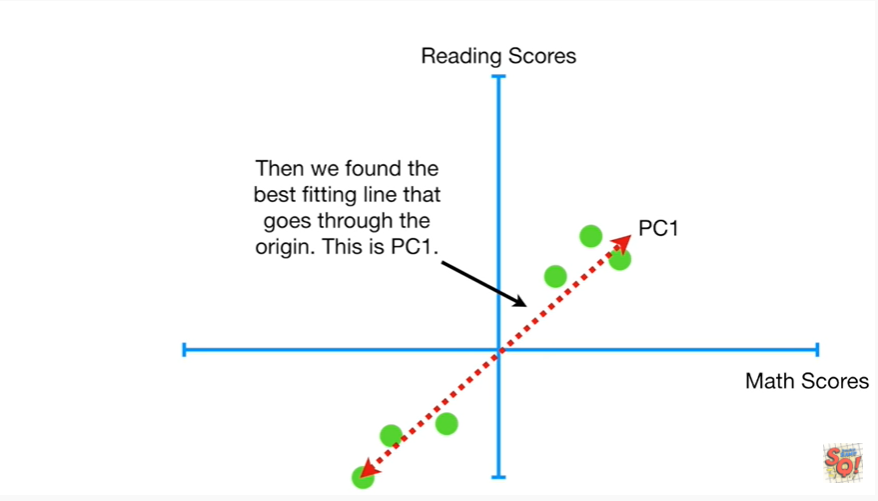
In the first example we have math and reading scores from a bunch of students



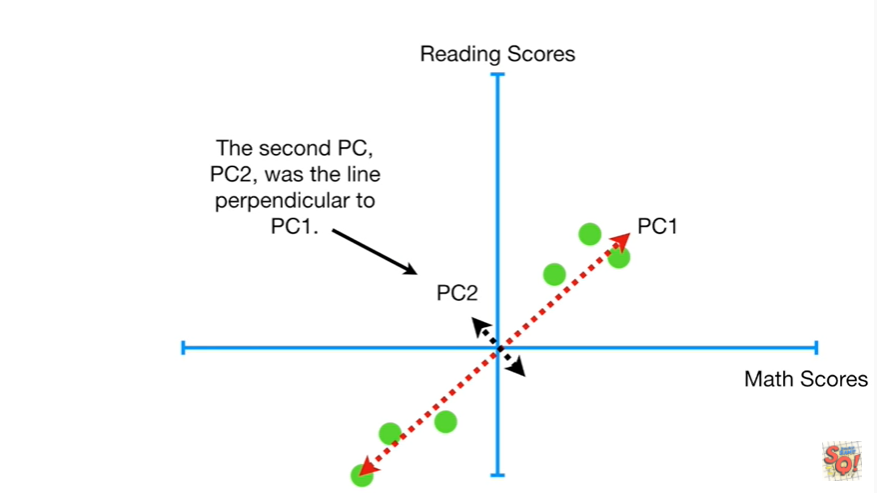
We then plotted the data on a two-dimensional graph.



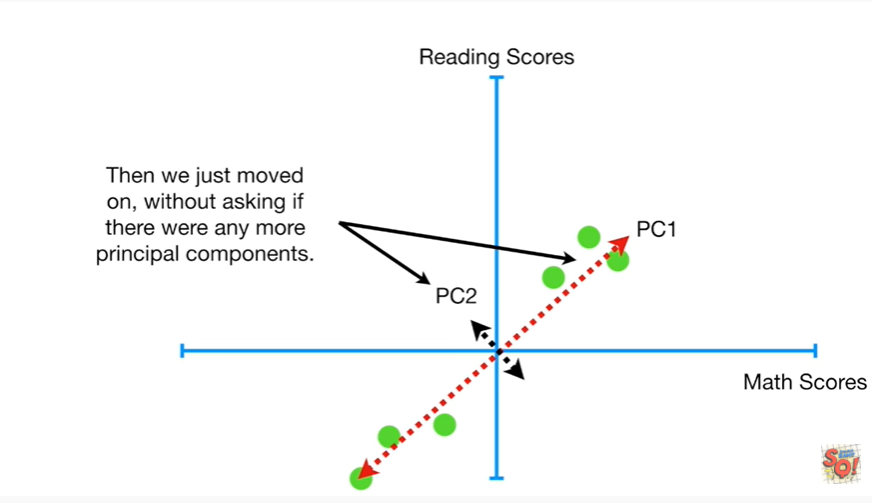
We then center the data.



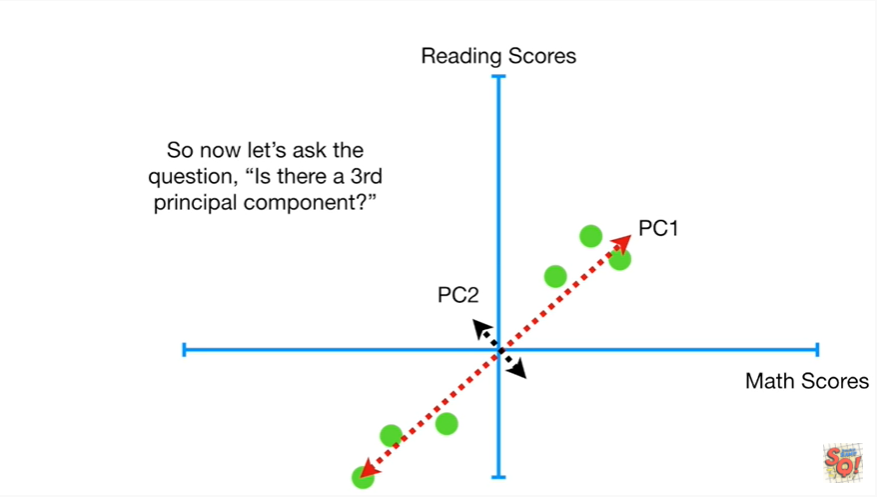
And then we found the best fitting line that goes through the origin this is PC 1.



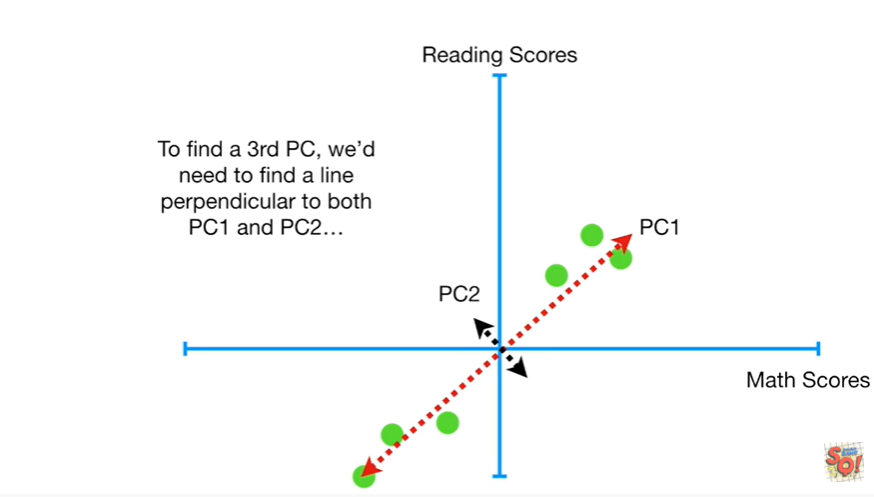
The second PC PC 2 was the line perpendicular to PC 1.



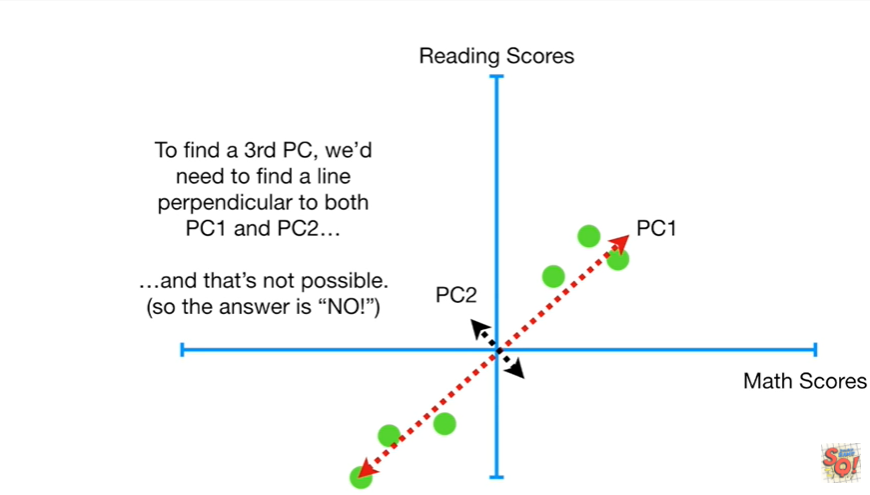
Then we just moved on without asking if there were any more principal components.



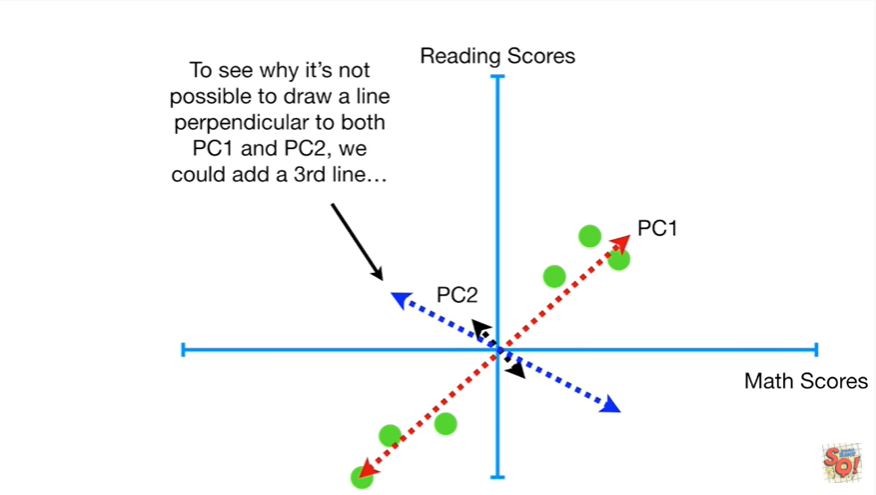
So now let's ask the question is there a third principle component ?



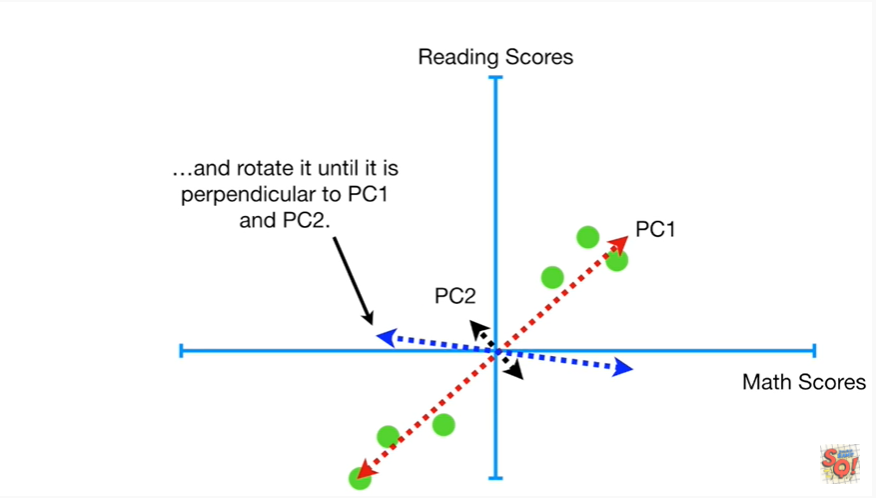
To find a third PC we need to find a line perpendicular to both pc1 & pc2



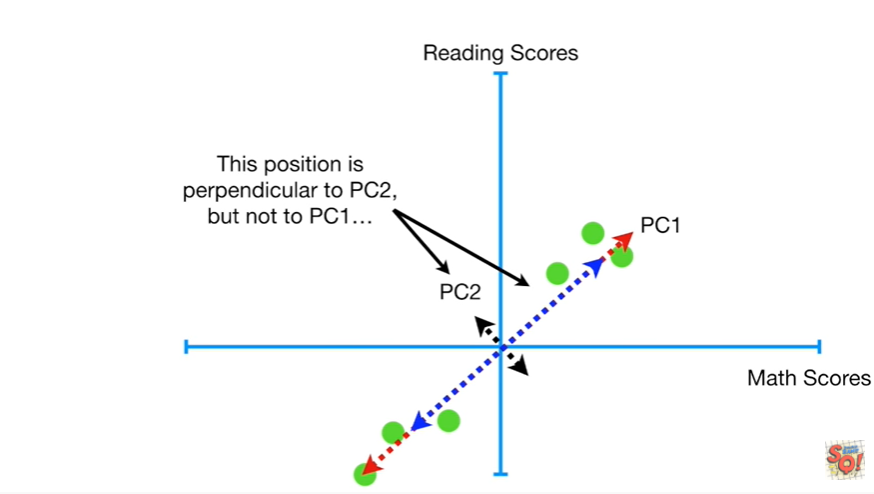
and in two dimensions that's not possible so the answer is no.



To see why it's not possible to draw a line perpendicular to both pc1 & pc2, we could add a third line



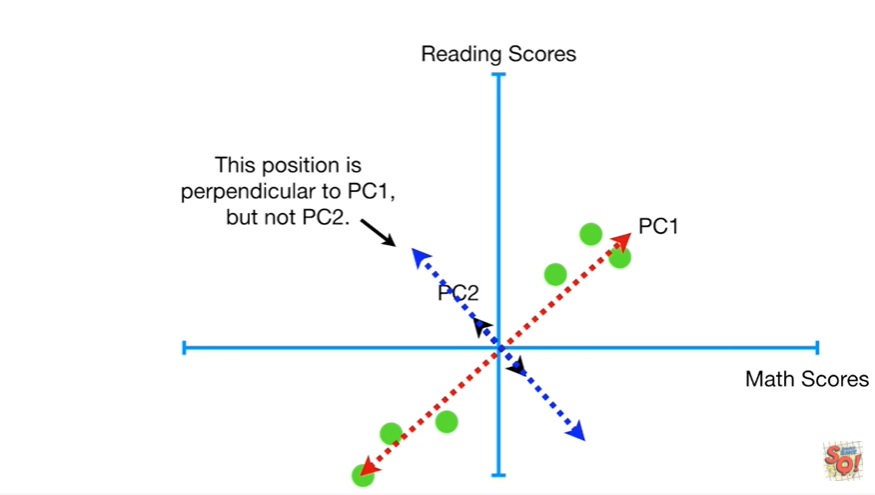
and rotate it until it is perpendicular to pc1 & pc2.



This position is perpendicular to pc2 but not PC1



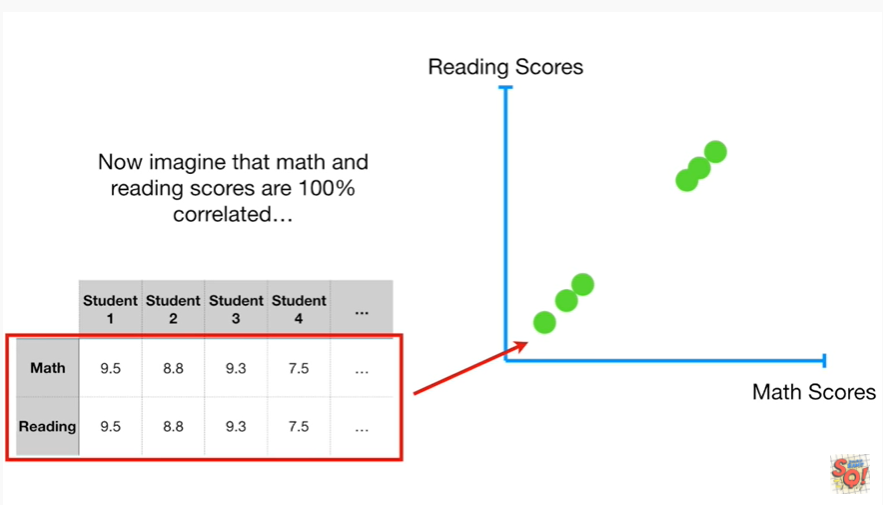
so we keep rotating.



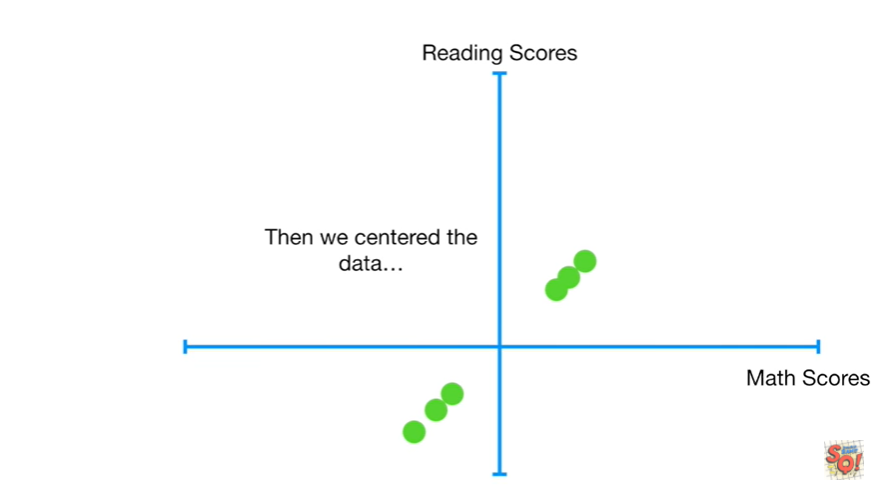
This position is perpendicular to PC1 but not PC2.



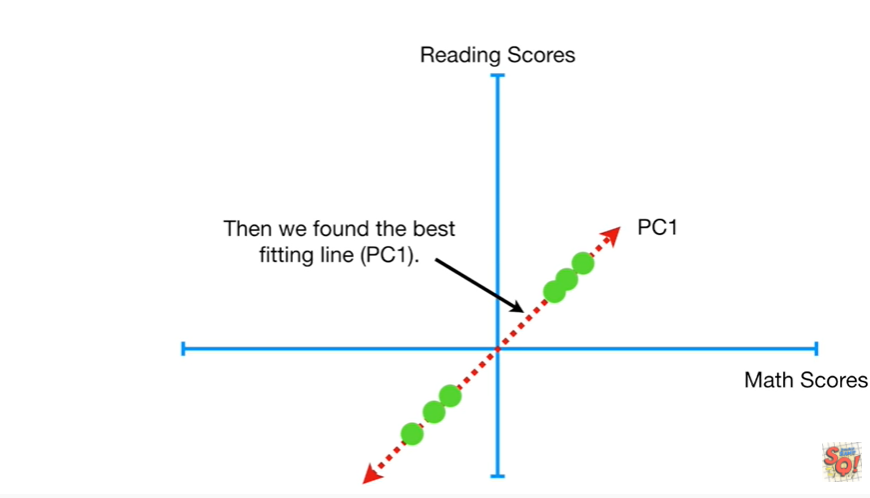
Thus when we measure two things math and reading in this case per sample ie per student at most we can have two principal components.



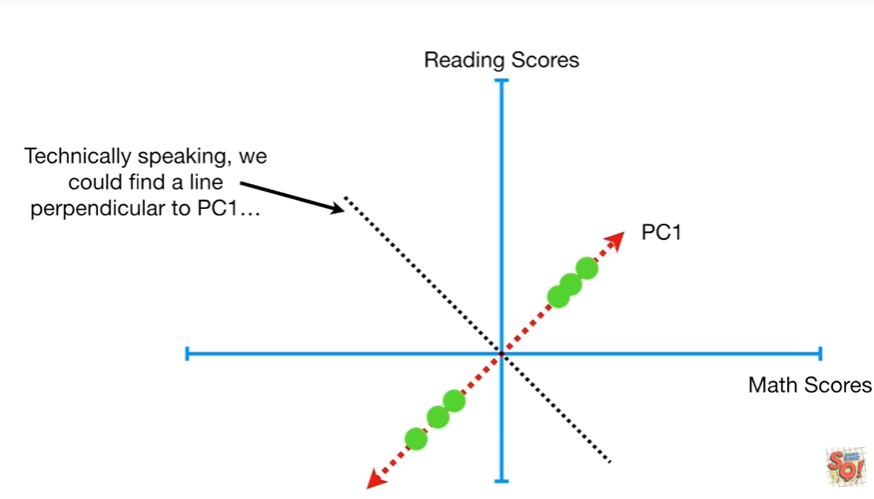
And now imagine that math and reading scores are 100% correlated



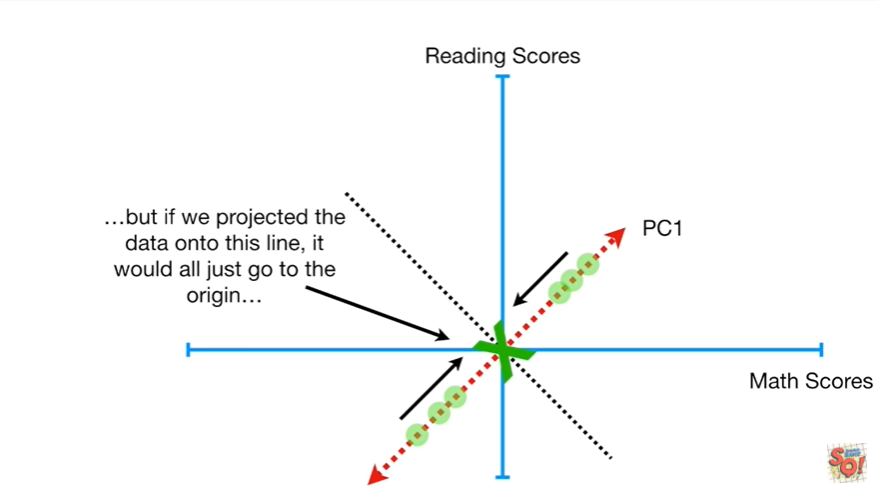
then we centered the data



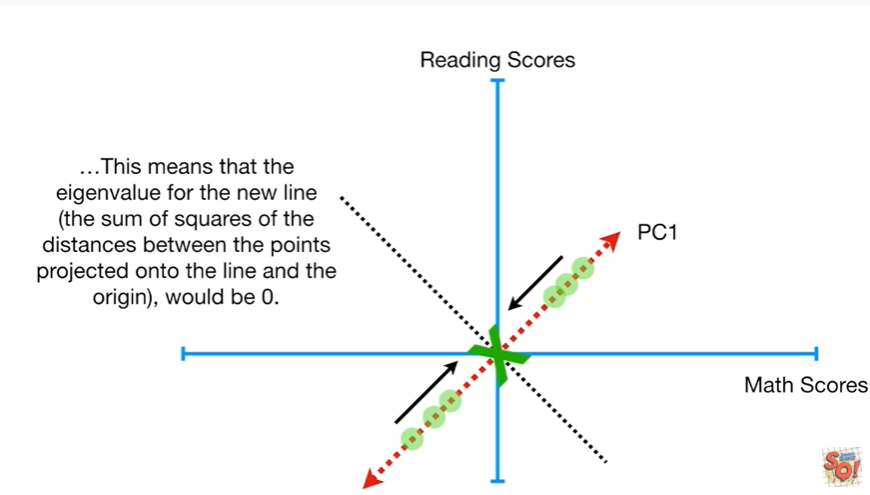
then we found the best fitting line principal component one.



Technically speaking we could find a line perpendicular to PC1



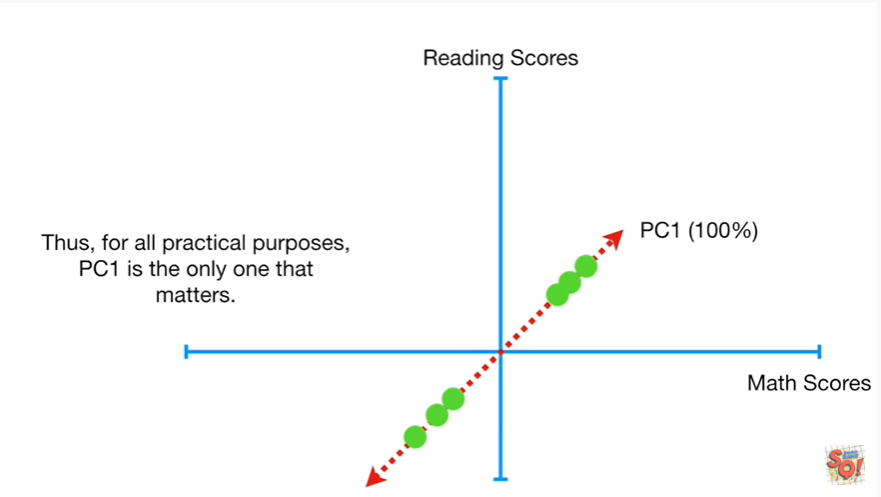
but if we projected the data on to this line it would all just go to the origin.



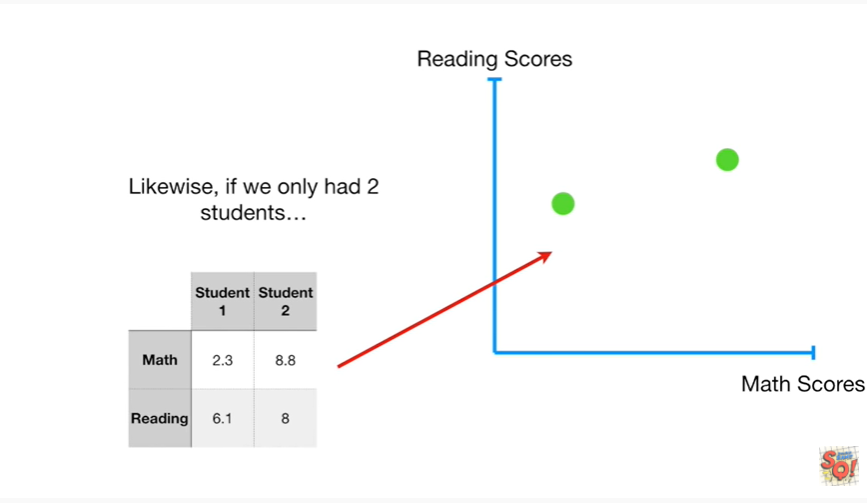
This means that the eigenvalue for the new line the sum of squares of the distances between the points projected onto the line and the origin would be zero.



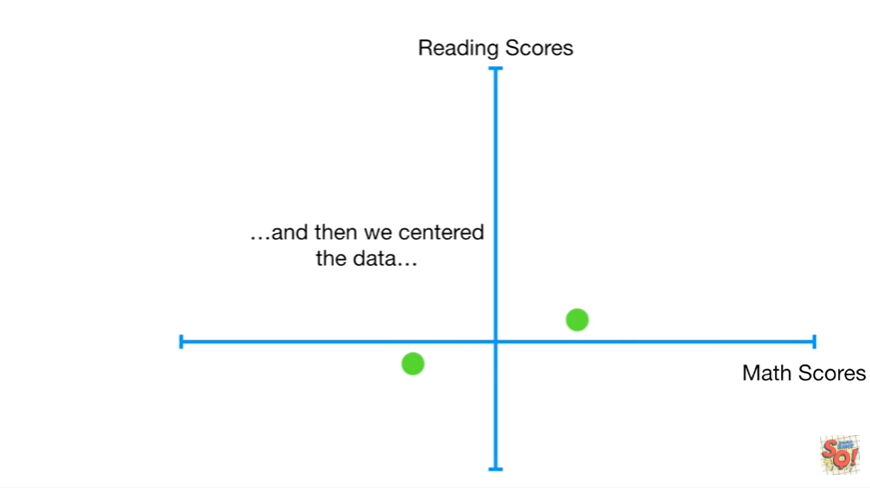
This means that pc1 accounts for 100% of the variation and the new line accounts for 0%.



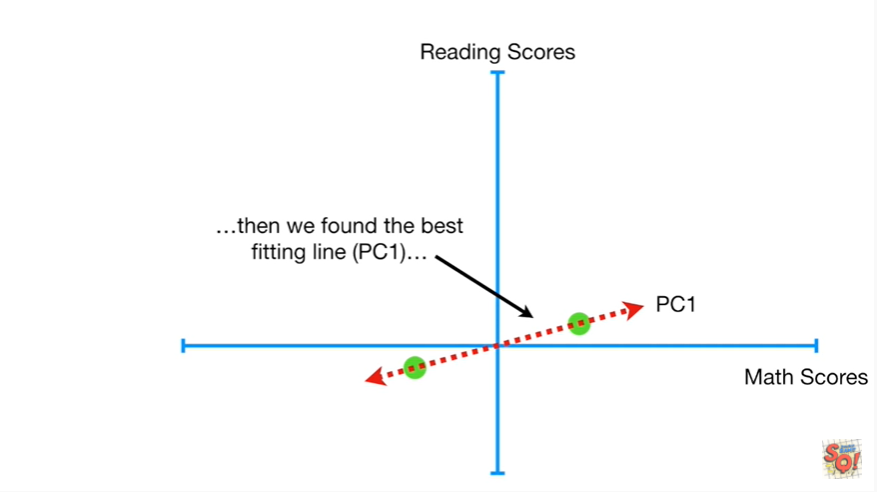
Thus for all practical purposes pc1 is the only one that matters.



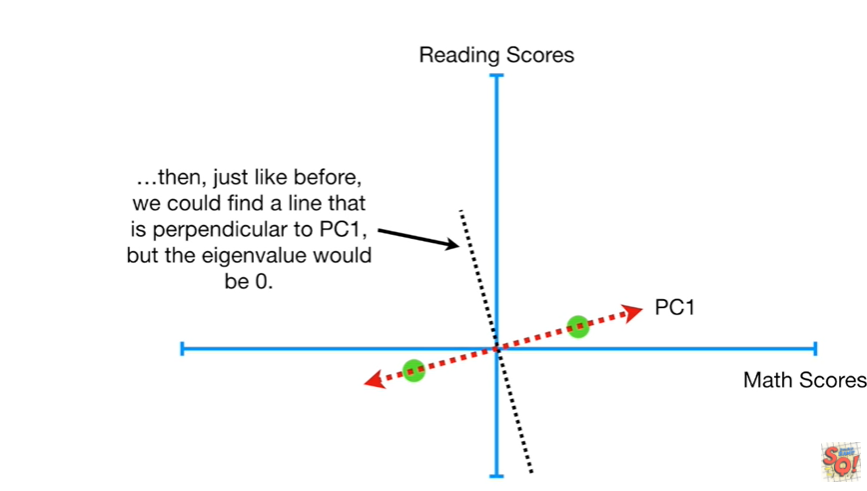
Likewise if we only had two students



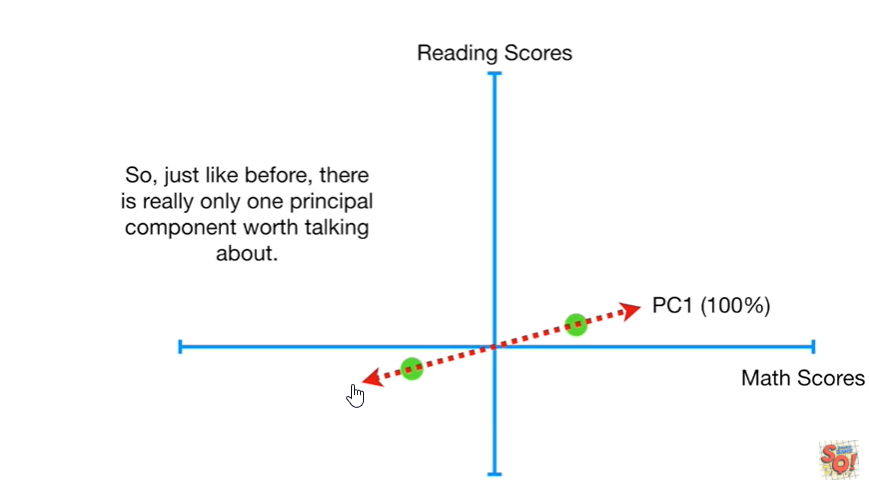
and then we centered the data



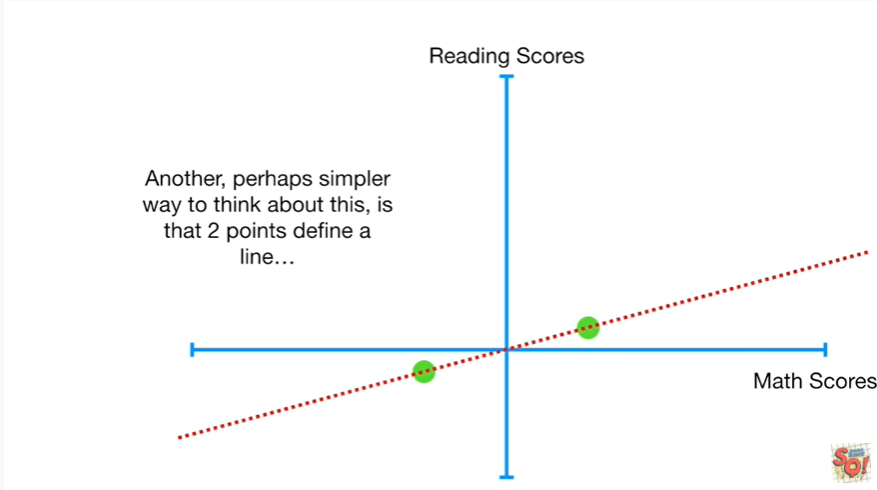
and then we found the best fitting line PC1



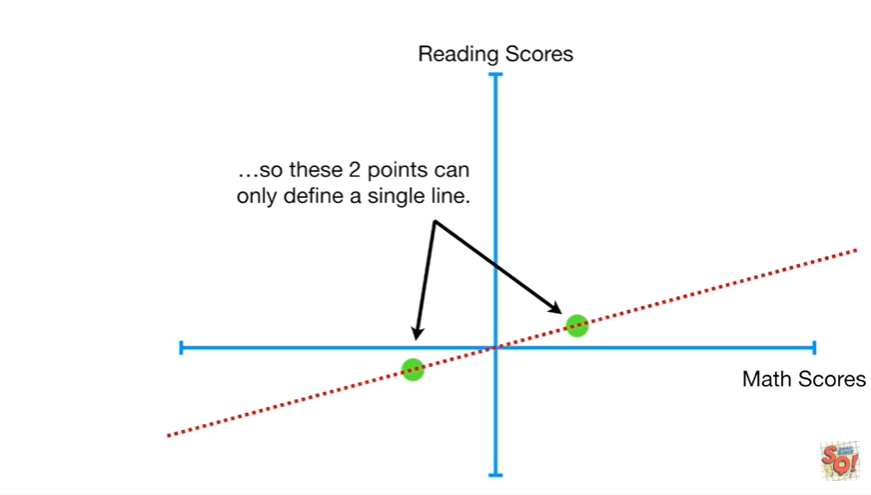
then just like before we could find a line that is perpendicular to PC1 but the eigen value would be zero.



So just like before there is really only one principle component worth talking about.



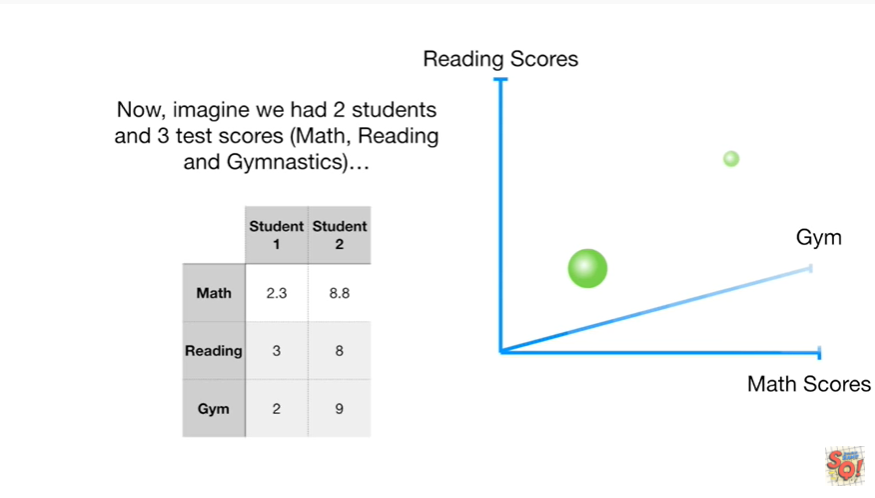
Another perhaps simpler way to think about this is that two points define a line



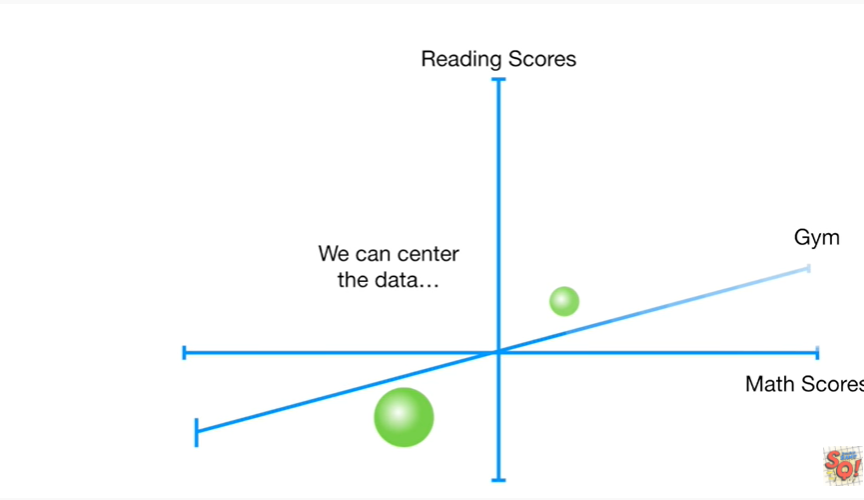
so these two points can only define a single line.



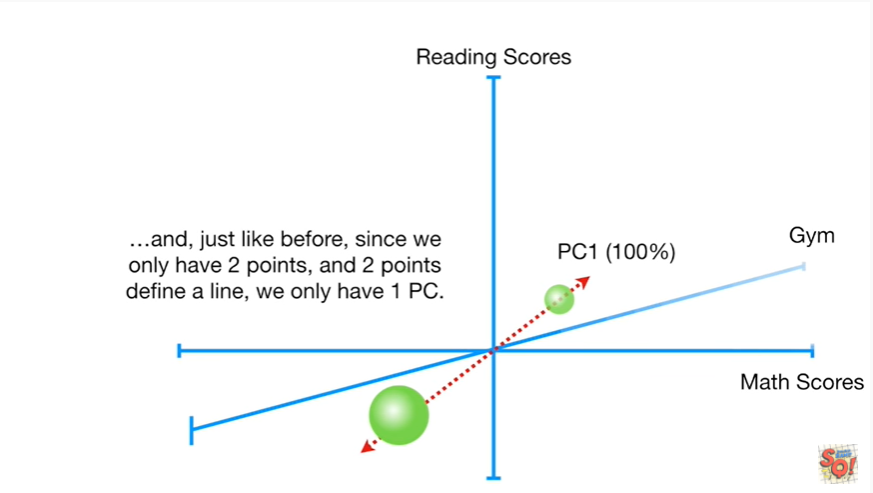
In order to define a plane something with two axes we need at least three points.



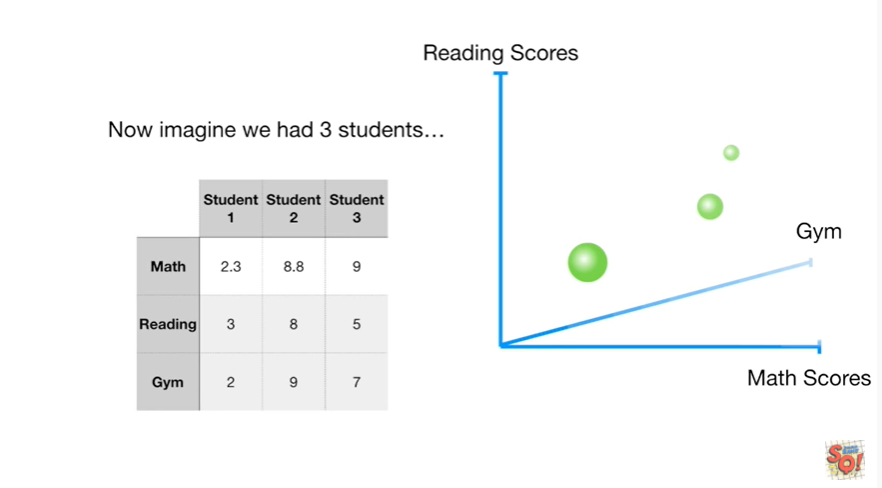
Now imagine we had two students and three test scores (math, reading and gymnastics)



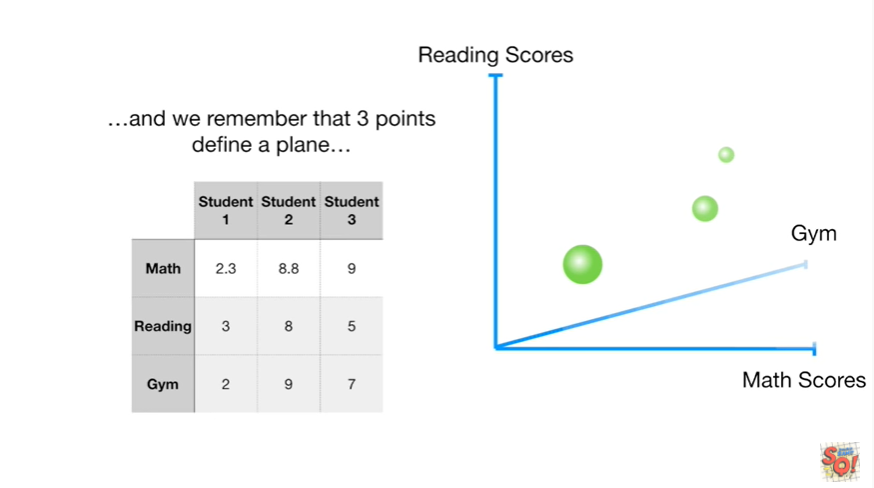
we can Center the data



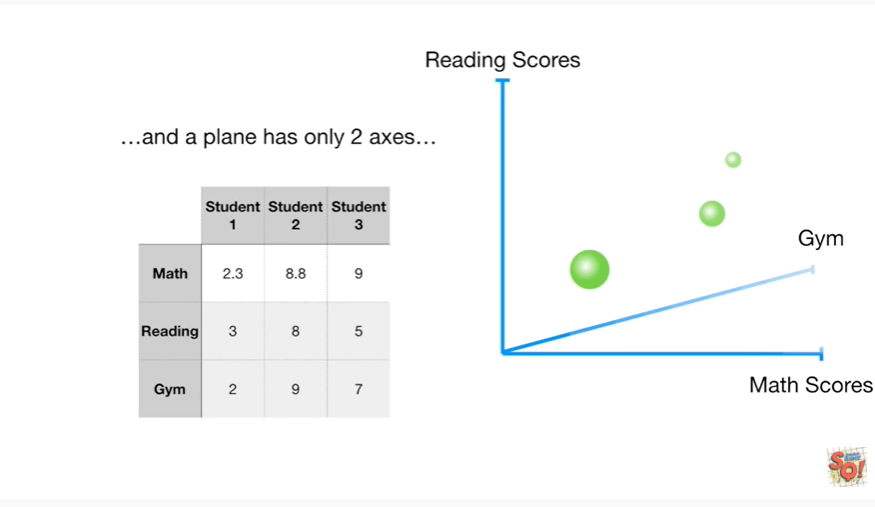
and just like before since we only have two points and two points to final line we only have one PC.



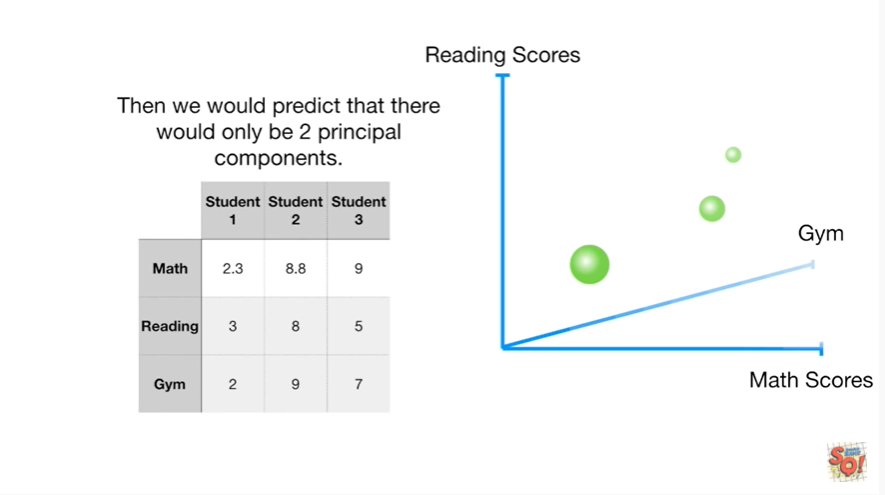
Now imagine we had three students



and we remember that three points define a plane



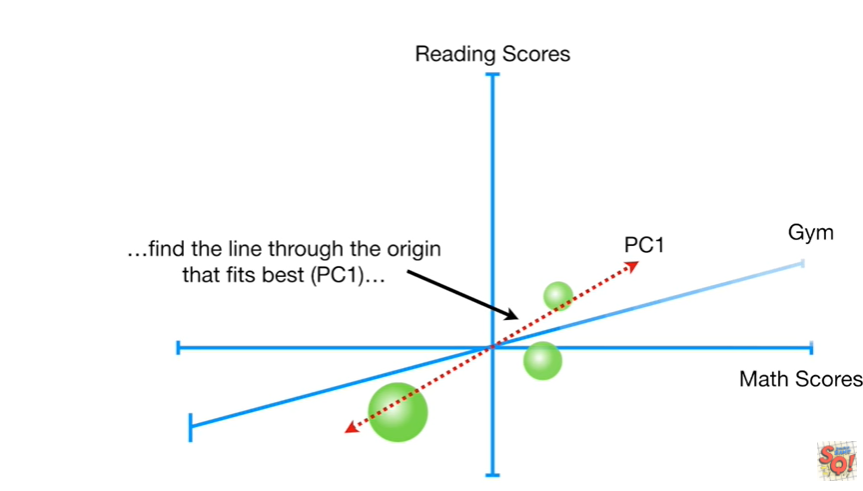
and a plane has only two axes.



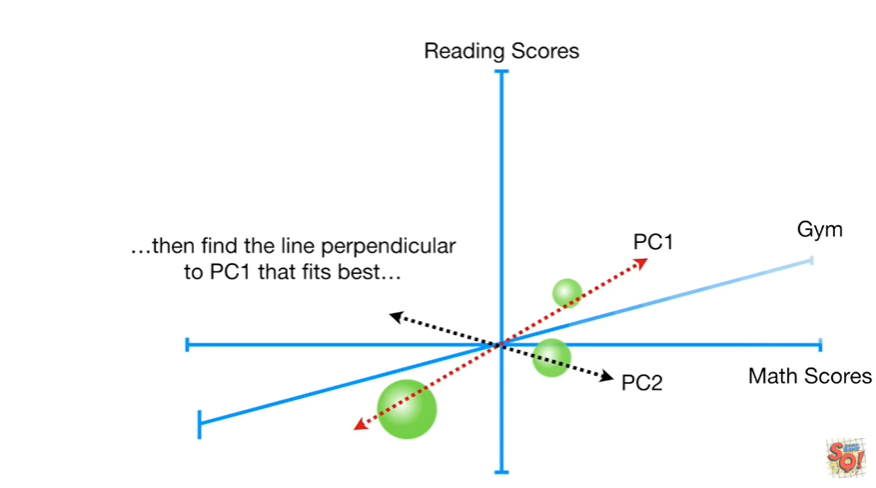
Then we would predict that there would only be two principal components.



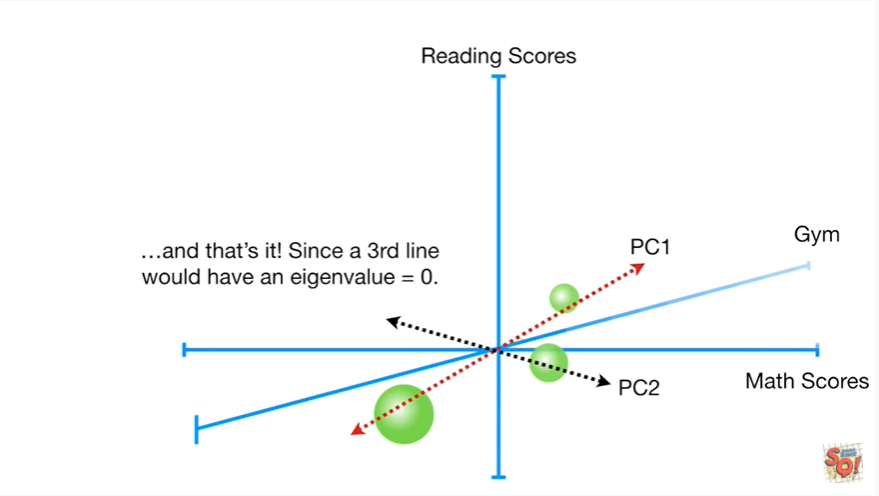
So we can Center the data



find the line through the origin that fits best PC1

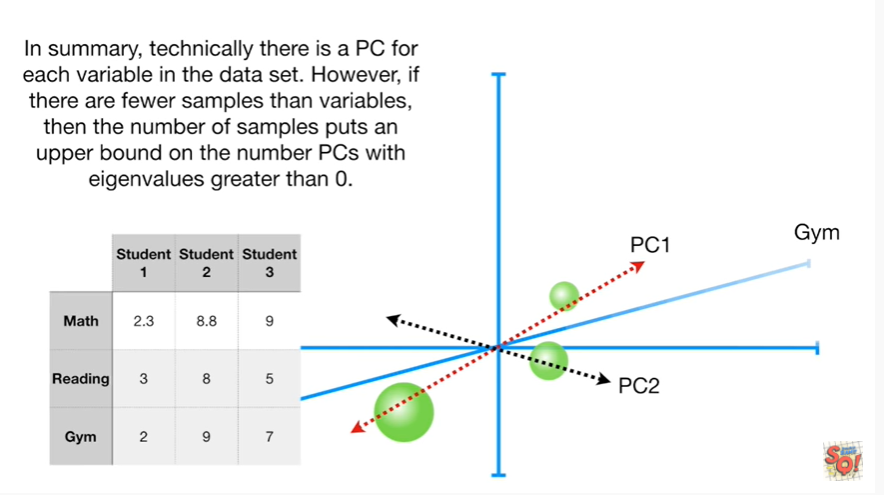


then find the line perpendicular to PC1 that fits best



and that's it since a third line would have an eigen value equal to zero.

There's no PC3.



In summary technically there is a PC for each variable in the data set.

However if there are fewer samples than variables then the number of samples puts an upper bound on the number of PCs with eigen values greater than zero.