PRINCIPAL COMPONENT ANALYSIS (PCA)

Taking a picture

Taking a picture

Taking a picture

Draft: Sharing is strictly Prohibited!

Taking a picture

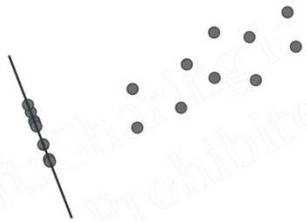
Taking a picture

Taking a picture

Taking a picture

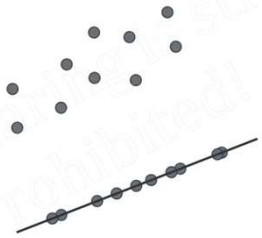
Draft. Confidential. Prohibited strictly

Dimensionality Reduction



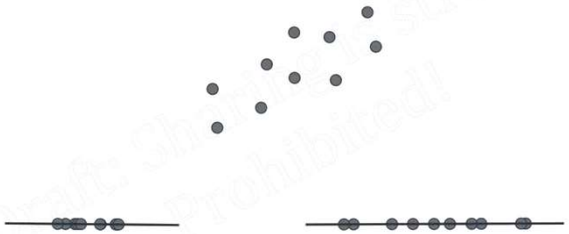
Draft: Sharing is strictly Prohibited!

Dimensionality Reduction



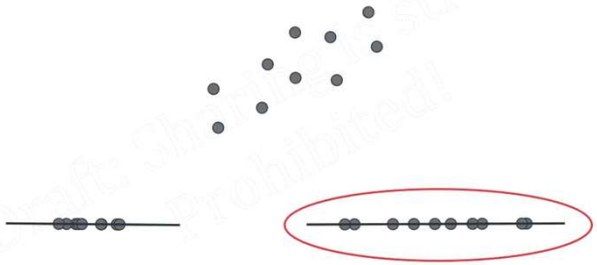
Draft: Sharing is strictly Prohibited!

Dimensionality Reduction



Draft: Sharing is strictly Prohibited!

Dimensionality Reduction



Draft: Sharing is strictly Prohibited!

Housing Data

Size
Number of rooms
Number of bathrooms

Schools around
Crime rate

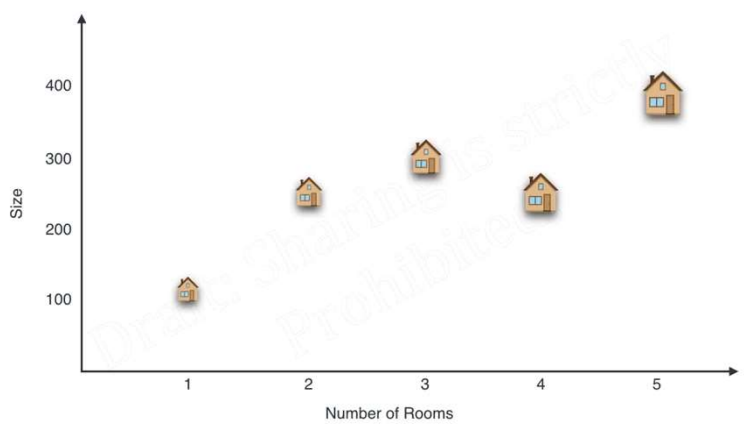
Draft: Sharing is strictly Prohibited!

Housing Data

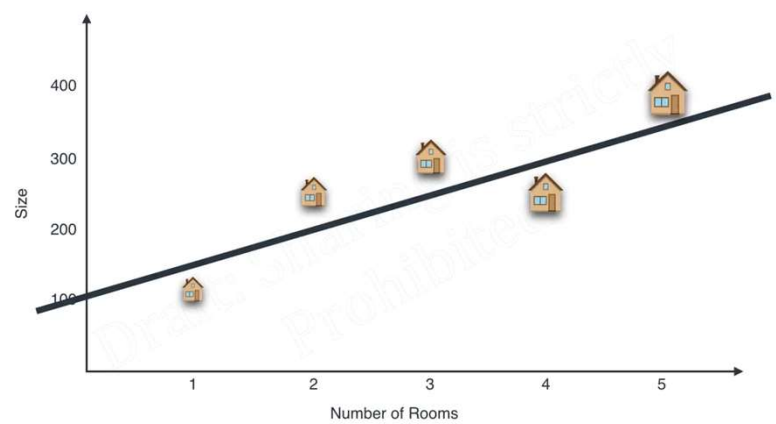
Size
Number of rooms → Size feature
Number of bathrooms

Schools around
Crime rate → Location feature

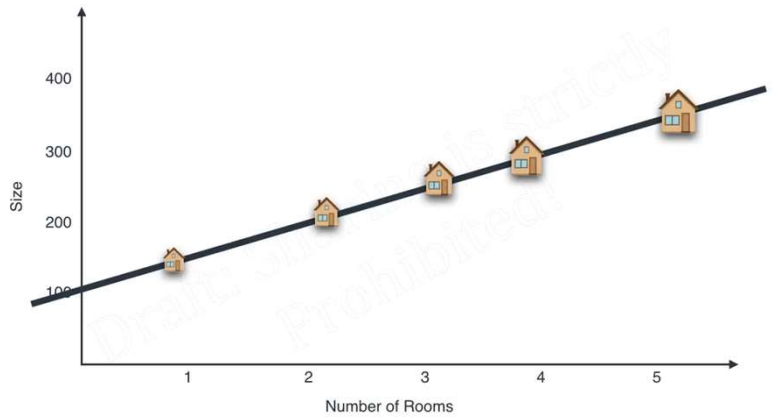
Draft: Sharing is strictly Prohibited!

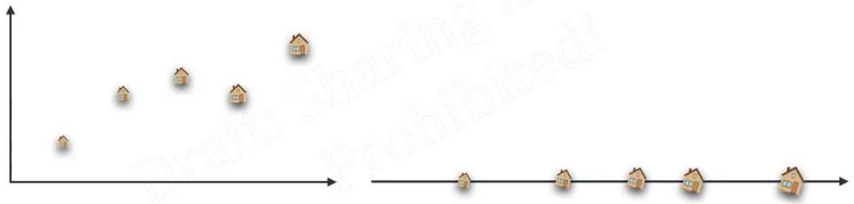


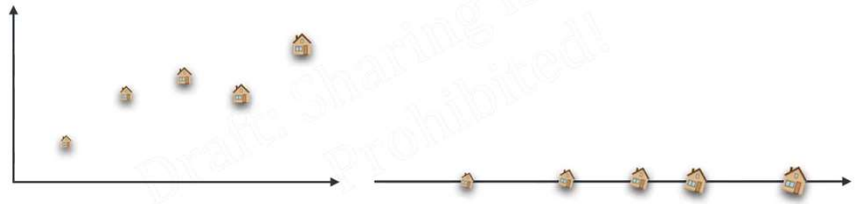
Draft: Sharing is strictly Prohibited!



Draft: Sharing is strictly Prohibited!







Draft: Sharing is strictly Prohibited!

Housing Data

5 dimensions

Size
Number of rooms
Number of bathrooms
Schools around
Crime rate

2 dimensions

Size feature

Location feature

Draft: Sharing is strictly Prohibited!

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

Let's start with a simple data set.

Draft: Sharing is strictly Prohibited!

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

We've measured transcription of two genes, gene 1 and gene 2…

Draft: Sharing is strictly Prohibited!

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

…in 6 different mice.

Draft: Sharing is strictly Prohibited!

	Student 1	Student 2	Student 3	Student 4	...
Math	95	88	93	75	...
Reading	96	79	98	81	...

the samples could be students in high school and the variables could be test scores in math and reading…

	Business 1	Business 2	Business 3	...
Market Cap	9.5 million	88 million	93 million	...
# Employed	960	79	98,000	...

…or the samples could be businesses and the variables could be market capitalization and number of employees.

	Sample 1	Sample 2	Sample 3	Sample 4	...
Variable 1	10	11	8	3	...
Variable 2	6	4	5	3	...

…think of the mice as individual samples…

	Sample 1	Sample 2	Sample 3	Sample 4	...
Variable 1	10	11	8	3	...
Variable 2	6	4	5	3	...

…and the genes as variables that we measure for each sample.

Draft: Sharing is strictly Prohibited!

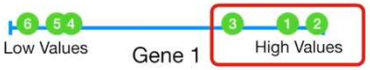
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

If we only measure 1 gene, we can plot the data on a number line…

Draft. Sharing is strictly Prohibited.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

Mice 1, 2 and 3 have relatively high values…



Draft. Sharing is strictly Prohibited.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

…and mice 4, 5 and 6 have relatively low values.

Draft. Sharing is strictly Prohibited.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

Even though it's a simple graph, it shows us that mice 1, 2 and 3 are more similar to each other than they are to mice 4, 5 6.

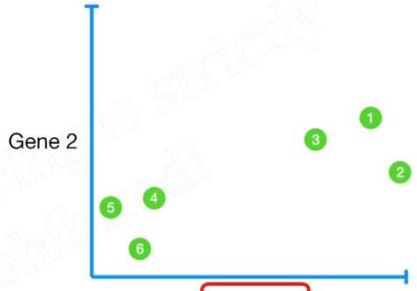
Draft. Sharing is strictly Prohibited.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

If we measured 2 genes…

Draft: Sharing is strictly Prohibited!

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



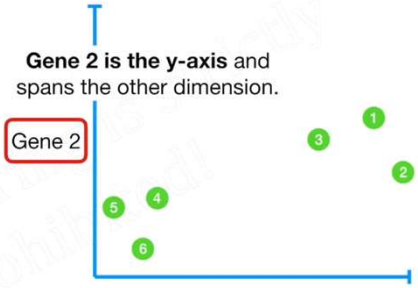
Gene 1

Gene 1 is the x-axis and spans one of the 2 dimensions in this graph.

Draft: Sharing is strictly Prohibited!

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

Gene 2 is the y-axis and spans the other dimension.

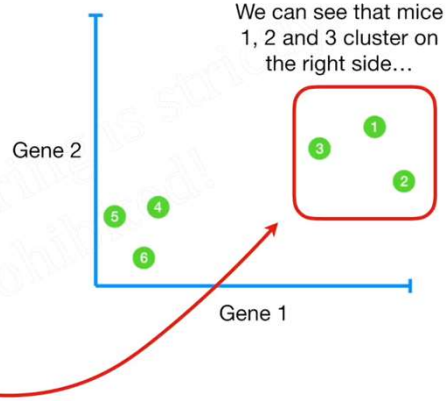


Gene 1

Draft: Sharing is strictly Prohibited!

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

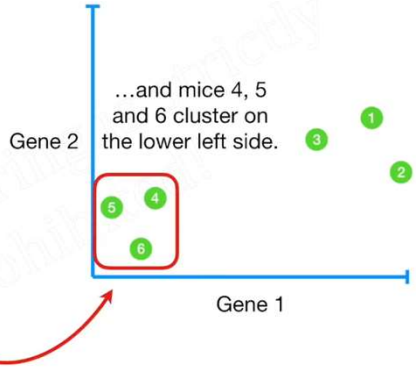
We can see that mice 1, 2 and 3 cluster on the right side…



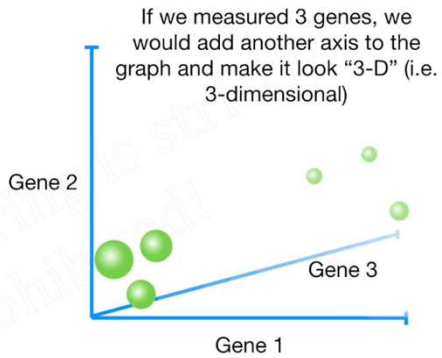
Gene 1

Draft: Sharing is strictly Prohibited!

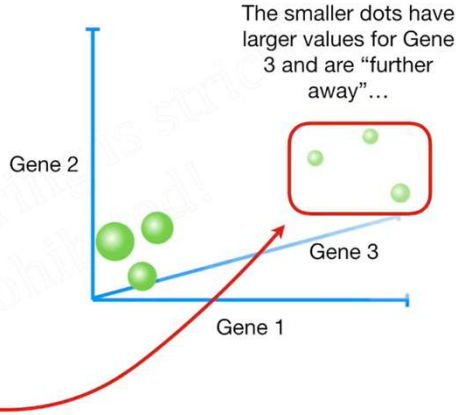
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



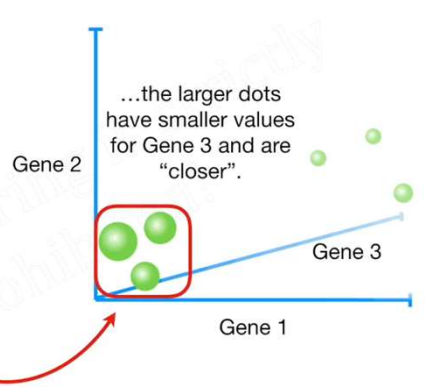
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

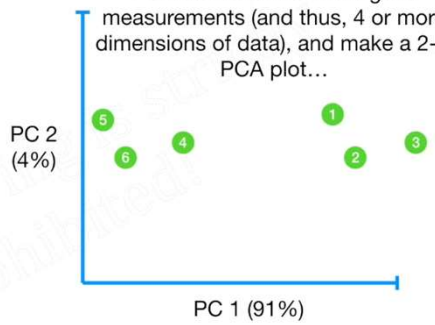


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

If we measured 4 genes, however, we can no longer plot the data - 4 genes require 4 dimensions.

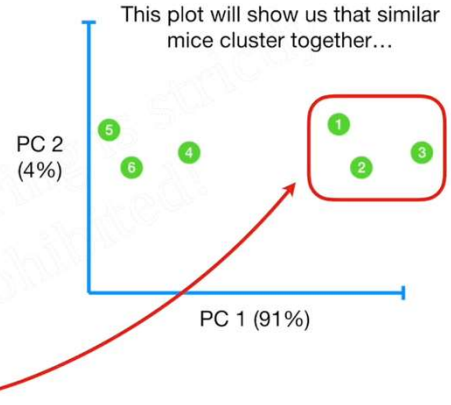
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

So we're going to talk about how PCA can take 4 or more gene measurements (and thus, 4 or more dimensions of data), and make a 2-D PCA plot…



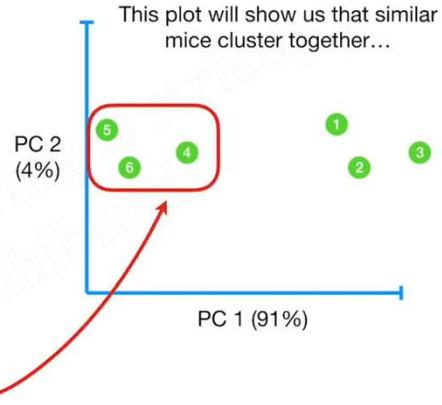
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

This plot will show us that similar mice cluster together…



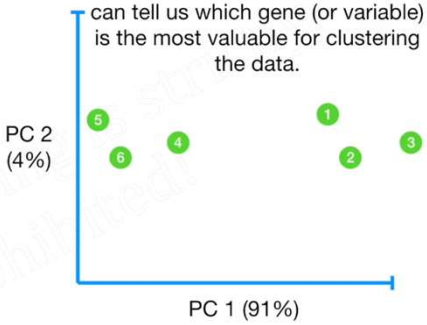
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

This plot will show us that similar mice cluster together…



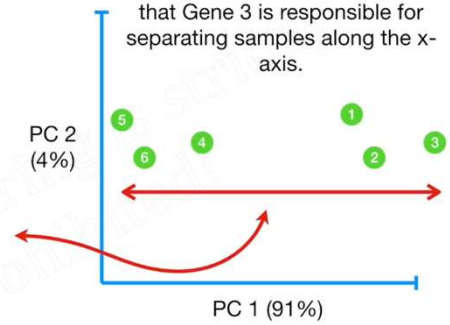
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

…We'll also talk about how PCA can tell us which gene (or variable) is the most valuable for clustering the data.



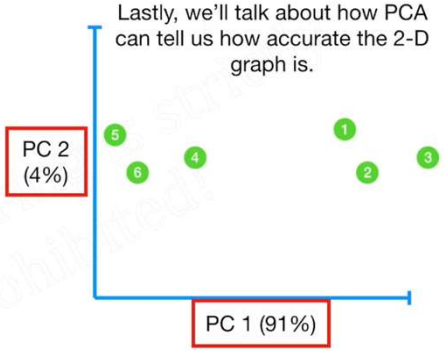
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

For example, PCA might tell us that Gene 3 is responsible for separating samples along the x-axis.



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

Lastly, we'll talk about how PCA can tell us how accurate the 2-D graph is.

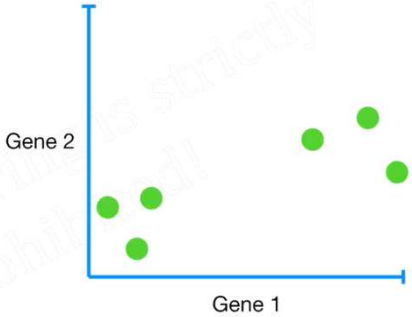


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

To understand what PCA does and how it works, let's go back to the dataset that only had 2 genes…

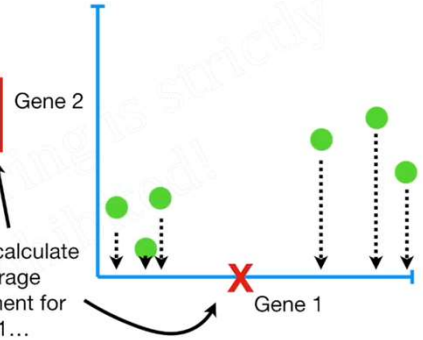
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

We'll start by plotting the data…



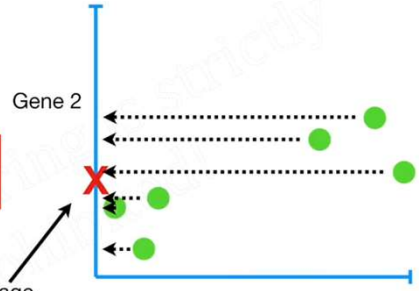
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

Then we'll calculate the average measurement for Gene 1…



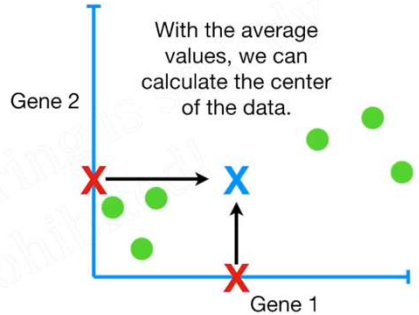
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

…and the average measurement for Gene 2.



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

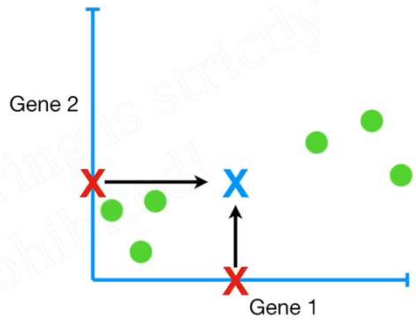
With the average values, we can calculate the center of the data.



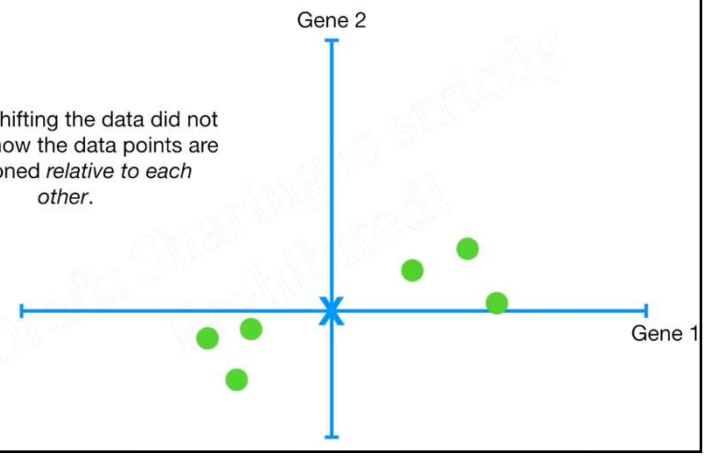
Draft: sharing is strictly prohibited!

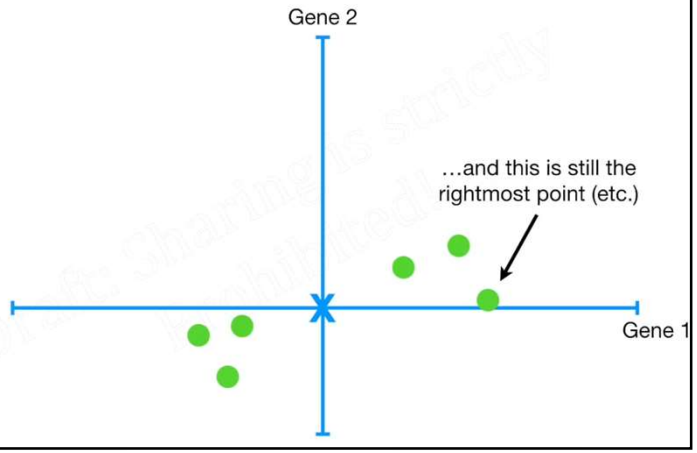
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

From this point on, we'll focus on what happens in the graph; we no longer need the original data…



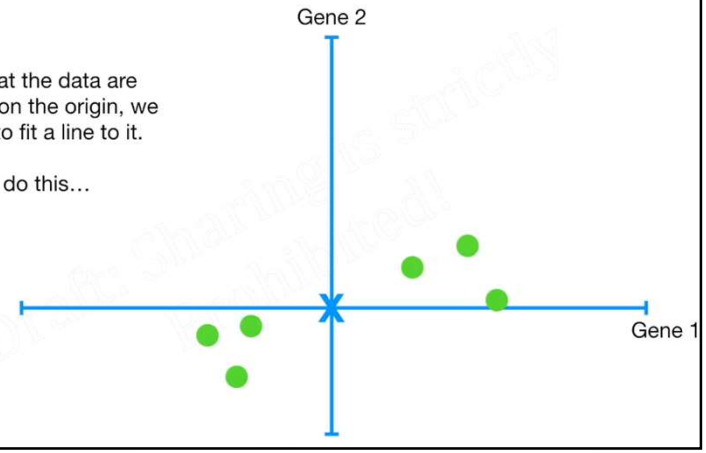
NOTE: Shifting the data did not change how the data points are positioned *relative to each other*.



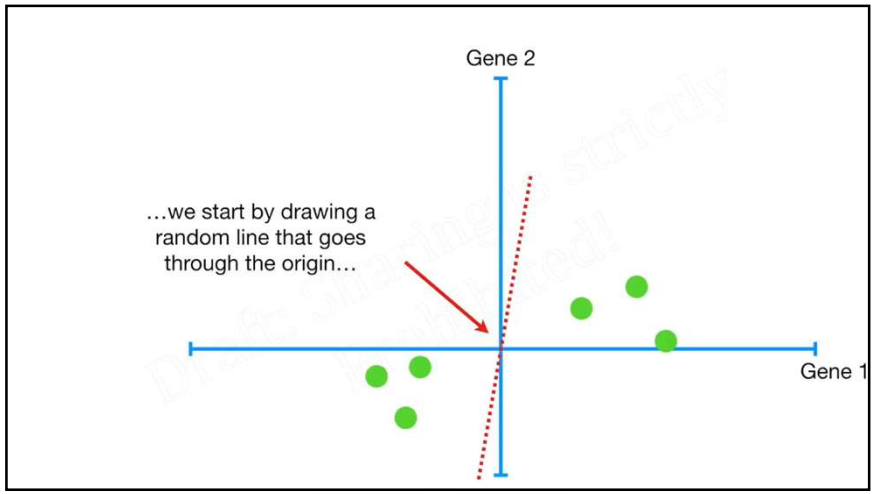


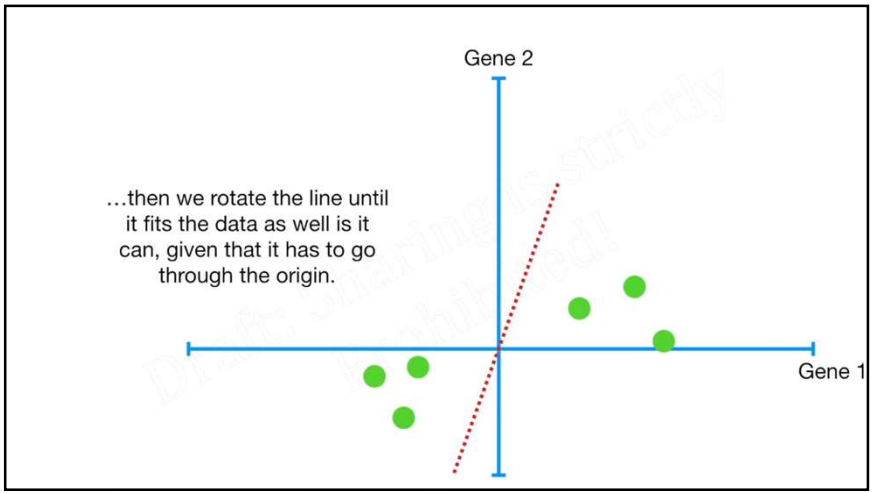
Now that the data are centered on the origin, we can try to fit a line to it.

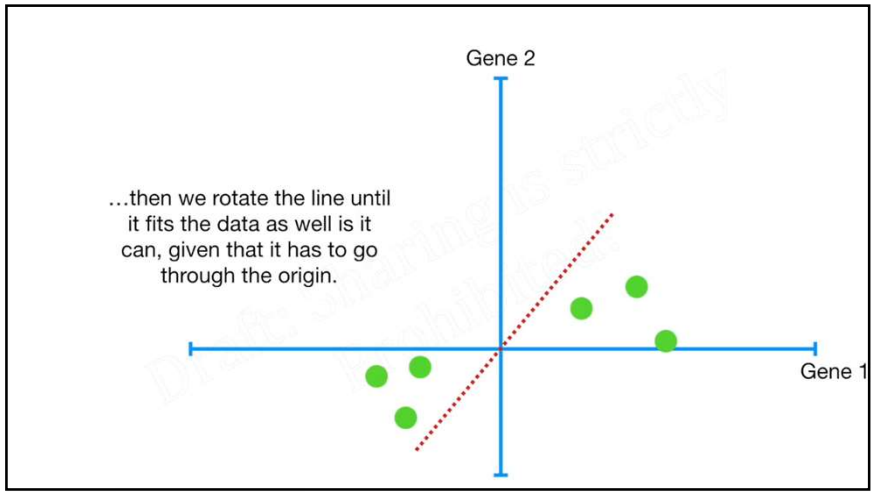
To do this…

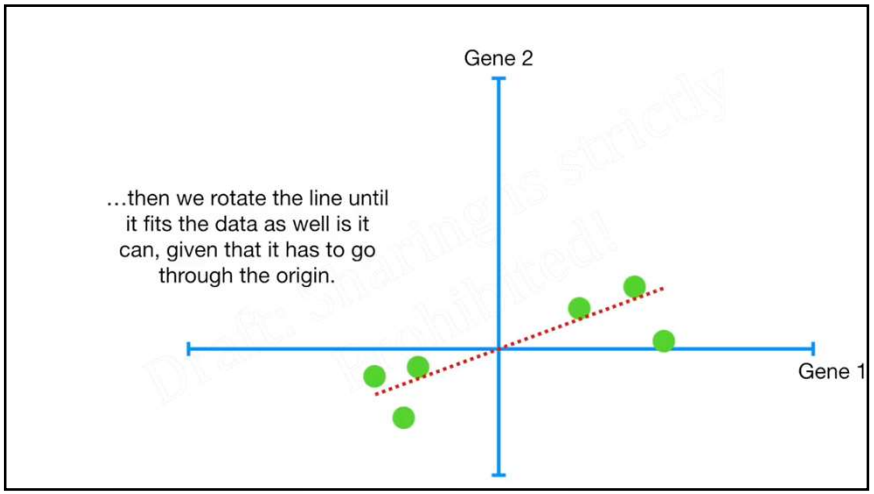


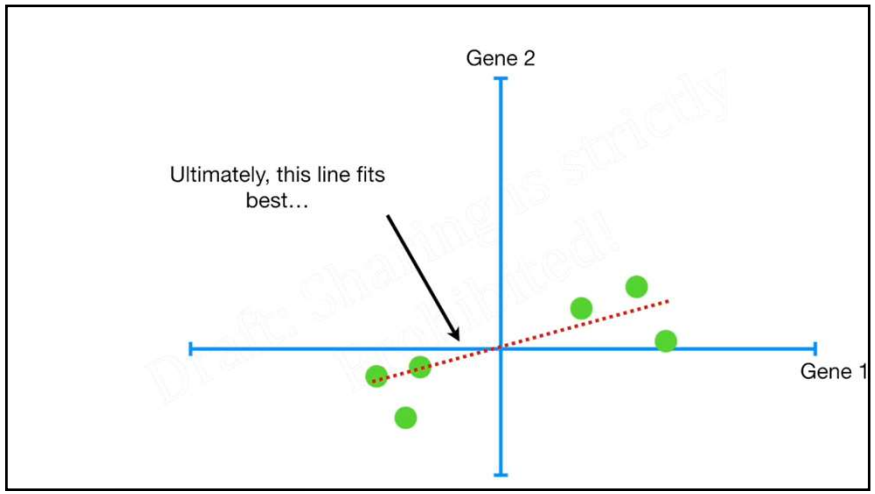
Draft: Sharing is strictly prohibited!

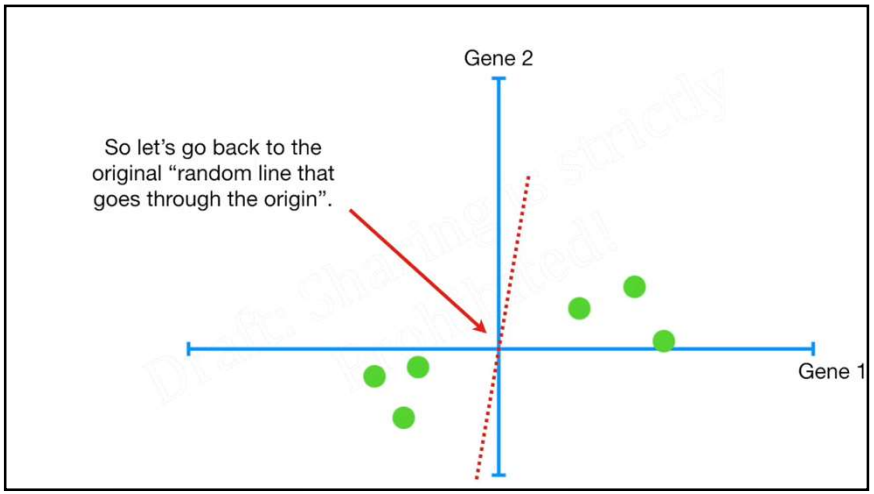

Gene 2
…we start by drawing a random line that goes through the origin…
Gene 1

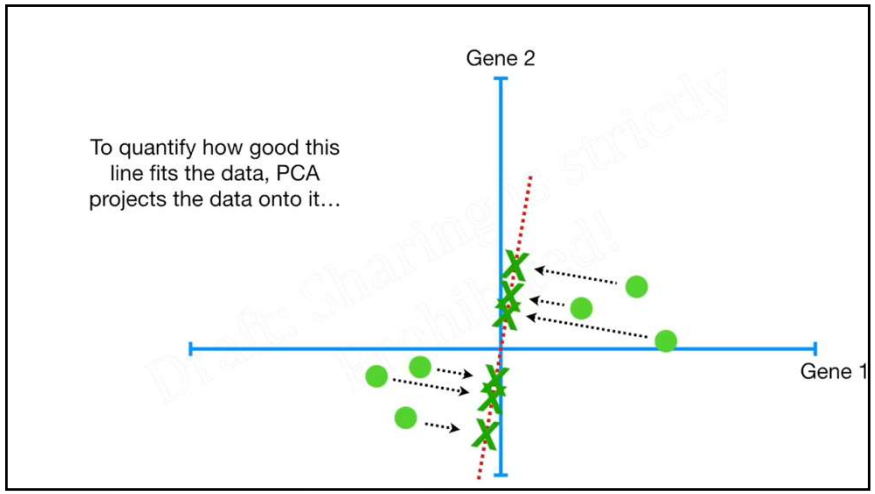

Gene 2
…then we rotate the line until it fits the data as well is it can, given that it has to go through the origin.
Gene 1

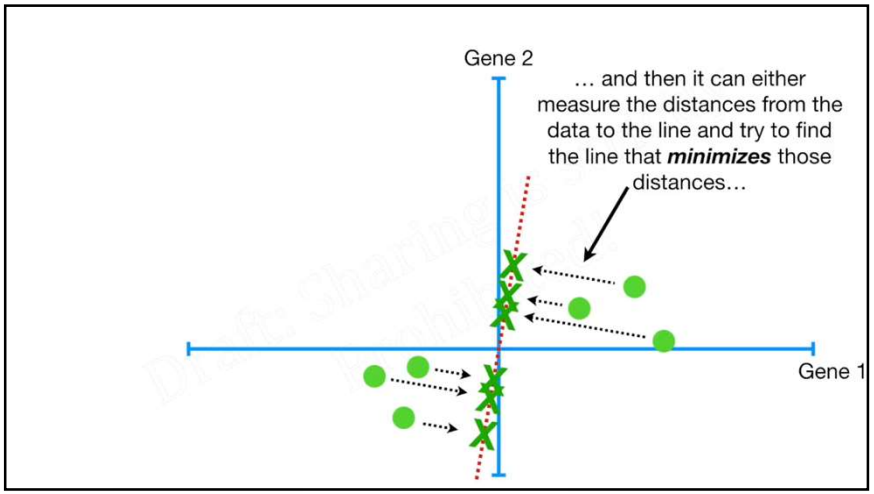

Gene 2
…then we rotate the line until it fits the data as well is it can, given that it has to go through the origin.
Gene 1

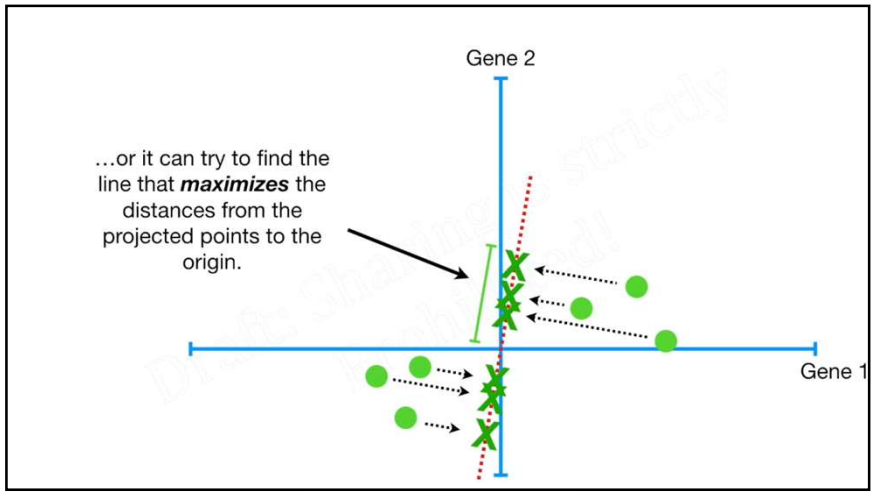

Gene 2
…then we rotate the line until it fits the data as well is it can, given that it has to go through the origin.
Gene 1

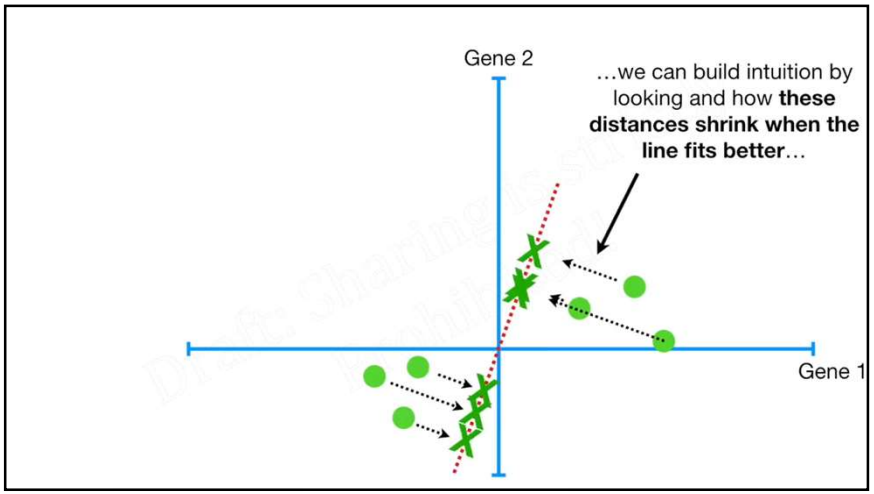

Gene 2
Ultimately, this line fits best…
Gene 1

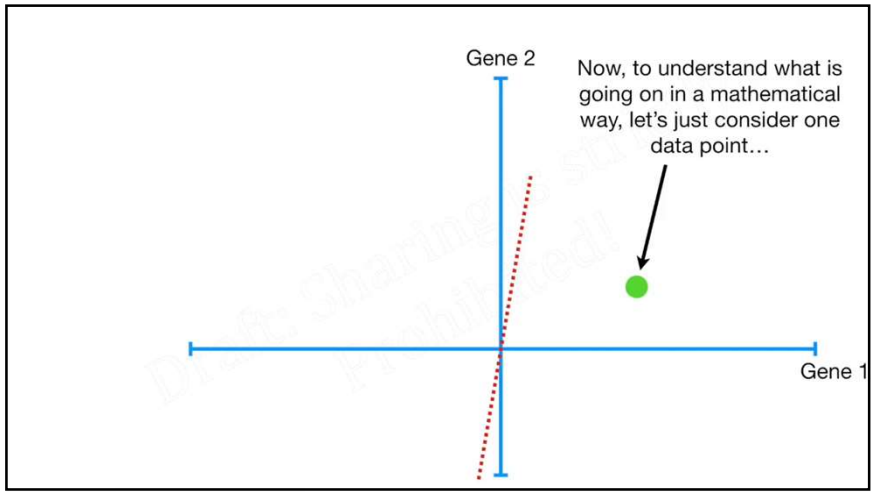

Gene 2
So let's go back to the original "random line that goes through the origin".
Gene 1

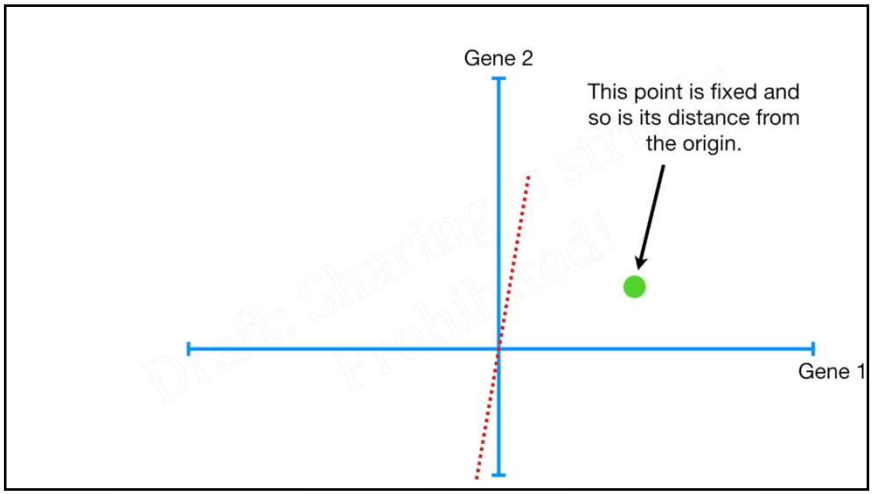

Gene 2
To quantify how good this line fits the data, PCA projects the data onto it…
Gene 1

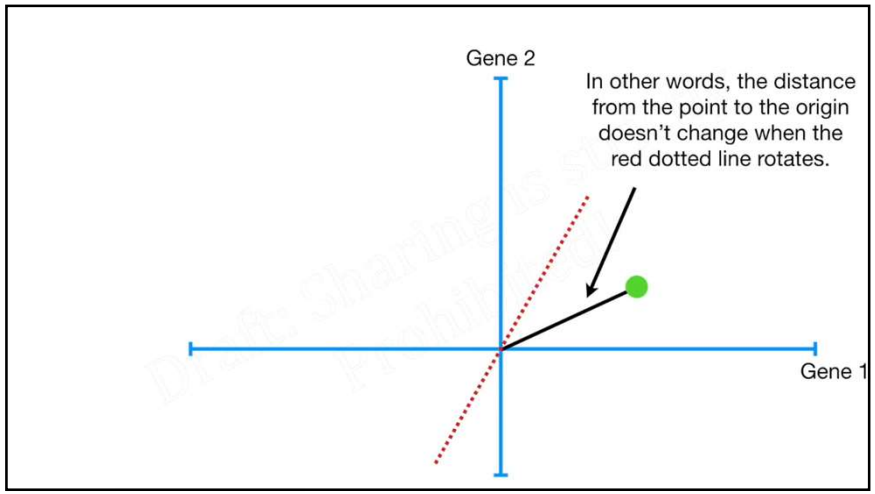

Gene 2
… and then it can either measure the distances from the data to the line and try to find the line that ***minimizes*** those distances…
Gene 1

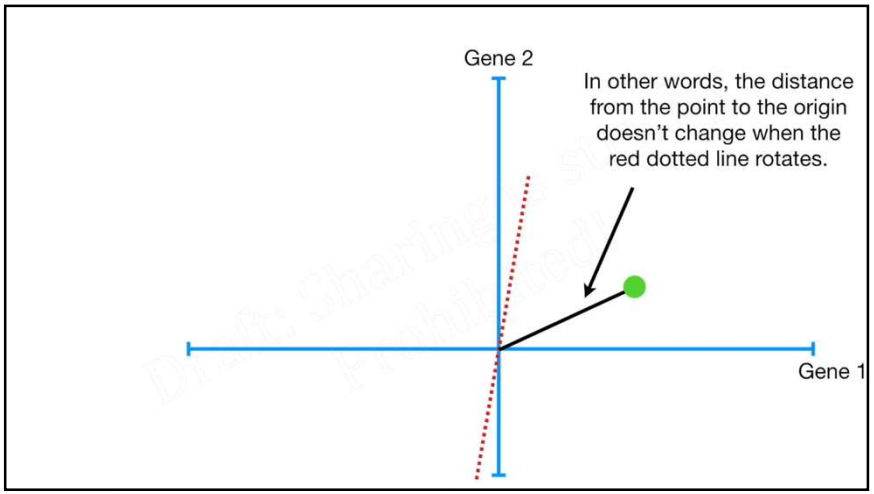

Gene 2
…or it can try to find the line that ***maximizes*** the distances from the projected points to the origin.
Gene 1

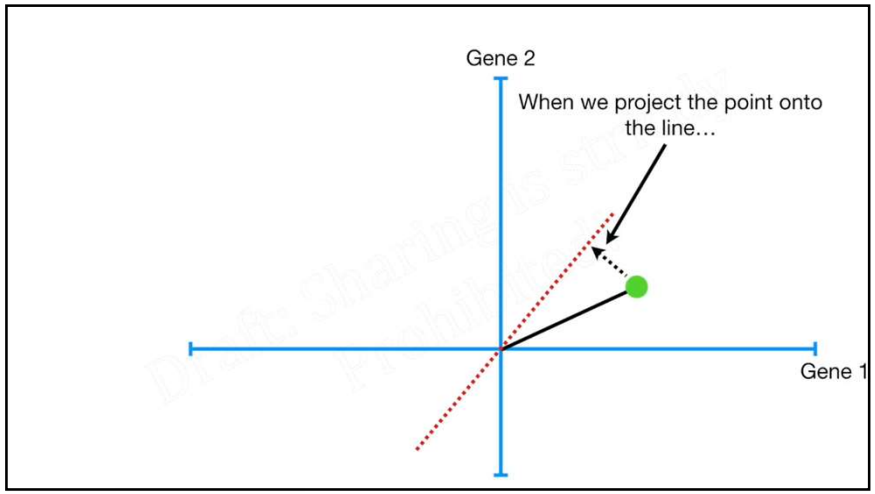

Gene 2
…we can build intuition by looking and how **these distances shrink when the line fits better**…
Gene 1

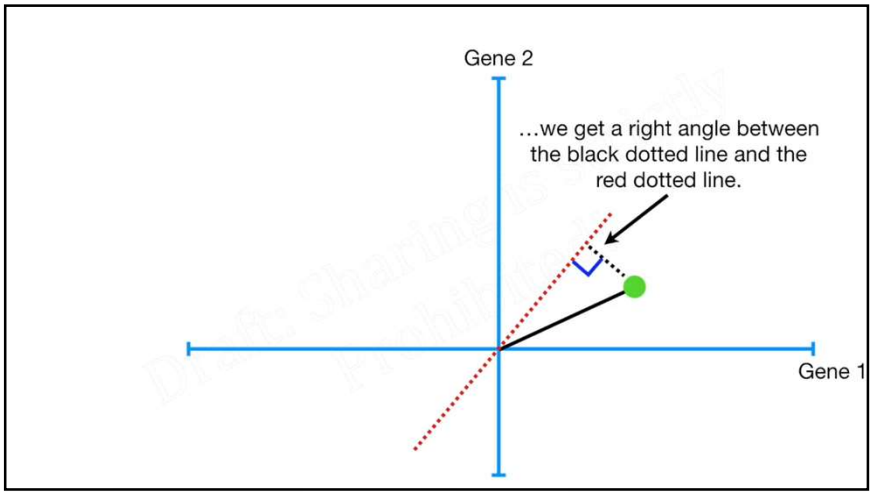

Gene 2
Now, to understand what is going on in a mathematical way, let's just consider one data point…
Gene 1

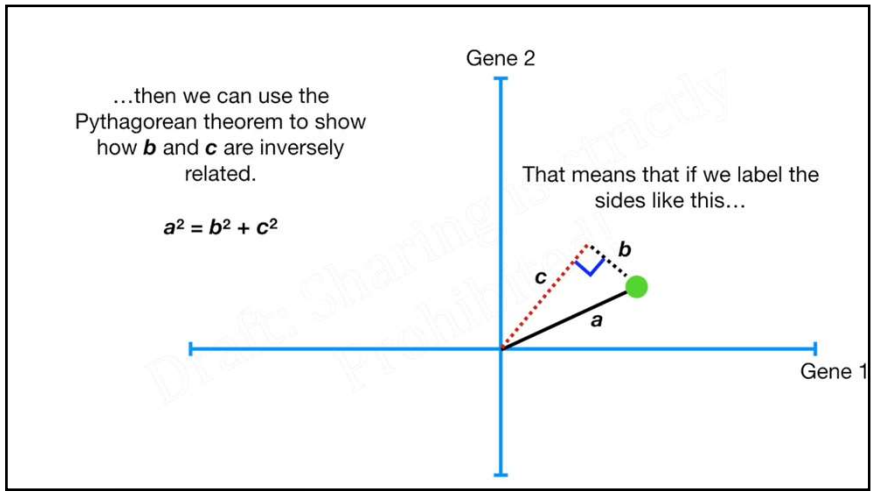

Gene 2
This point is fixed and so is its distance from the origin.
Gene 1

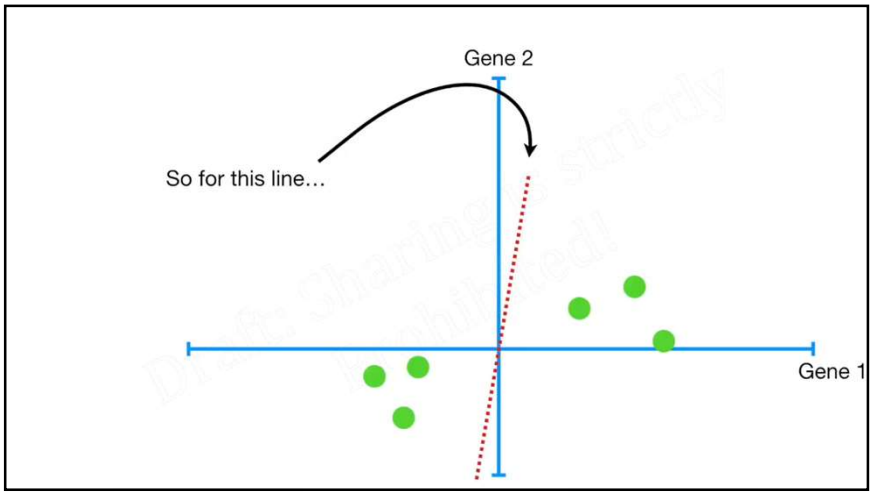

Gene 2
In other words, the distance from the point to the origin doesn't change when the red dotted line rotates.
Gene 1

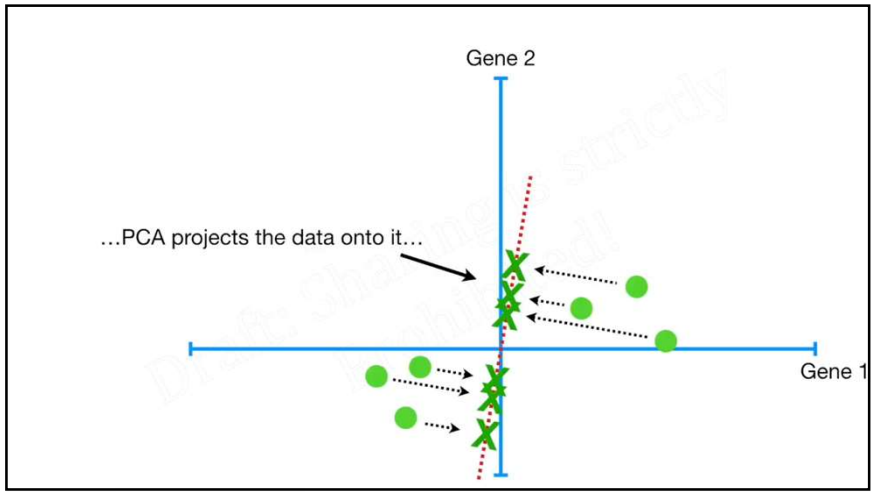

Gene 2
In other words, the distance from the point to the origin doesn't change when the red dotted line rotates.
Gene 1

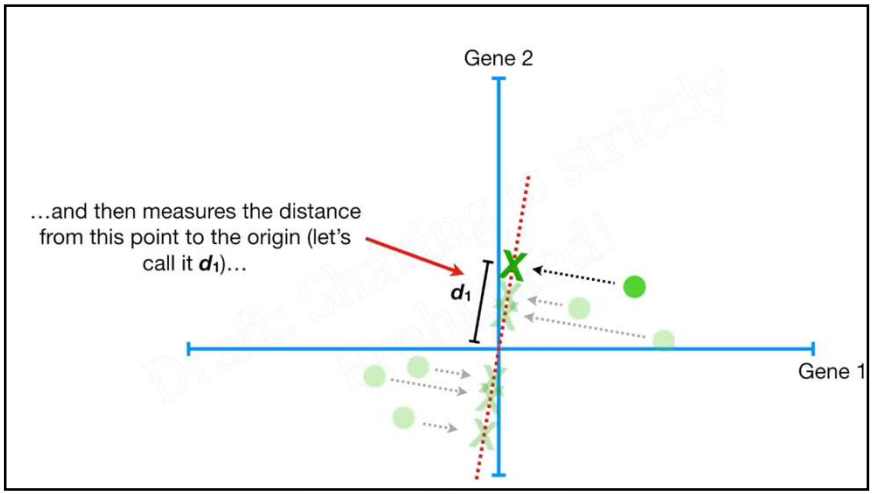

Gene 2
When we project the point onto the line…
Gene 1


Gene 2
…we get a right angle between the black dotted line and the red dotted line.
Gene 1


Gene 2
…then we can use the Pythagorean theorem to show how **b** and **c** are inversely related.
$a^2 = b^2 + c^2$
That means that if we label the sides like this…
b
c
a
Gene 1

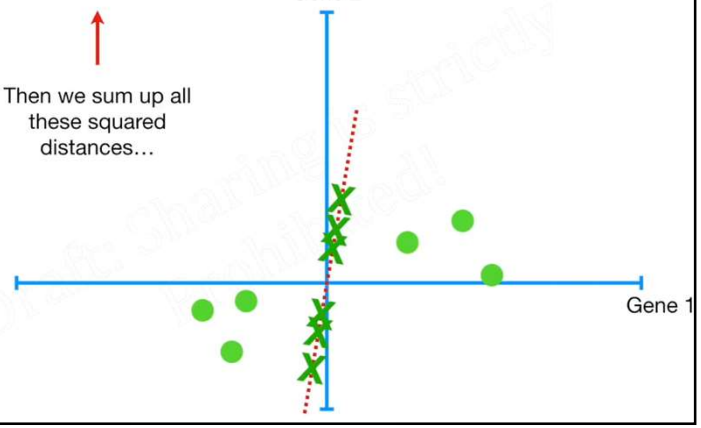

Gene 2
So for this line…
Gene 1


Gene 2
…PCA projects the data onto it…
Gene 1

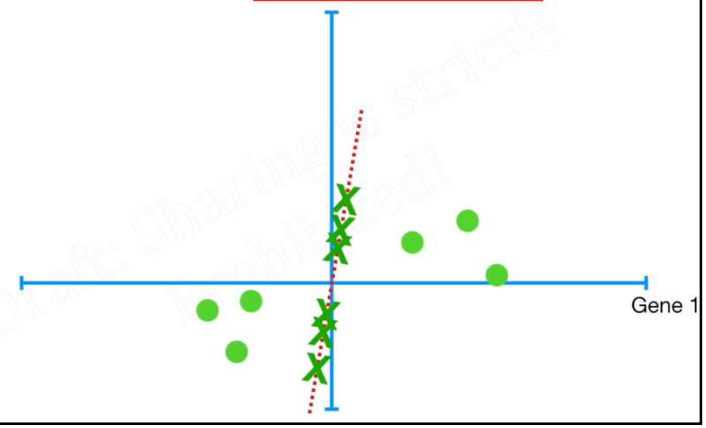

Gene 2
…and then measures the distance from this point to the origin (let's call it d_1)…
d_1
Gene 1

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$

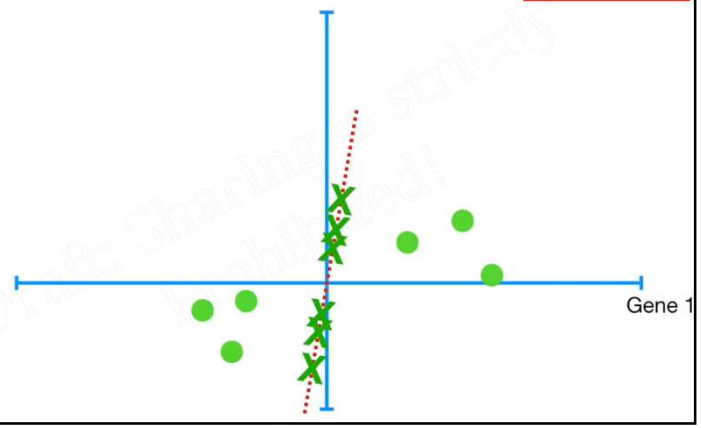
Then we sum up all these squared distances…



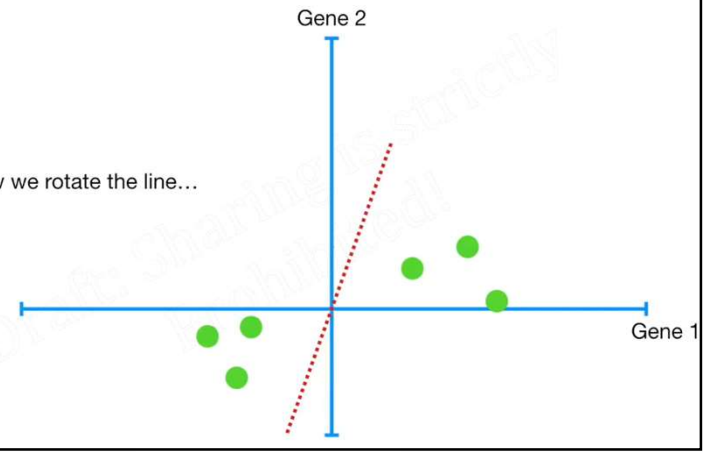
$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances



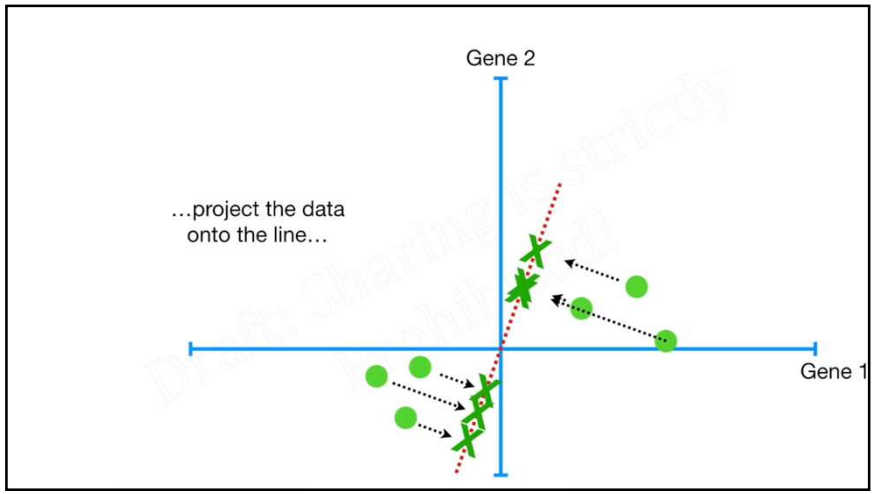
$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)

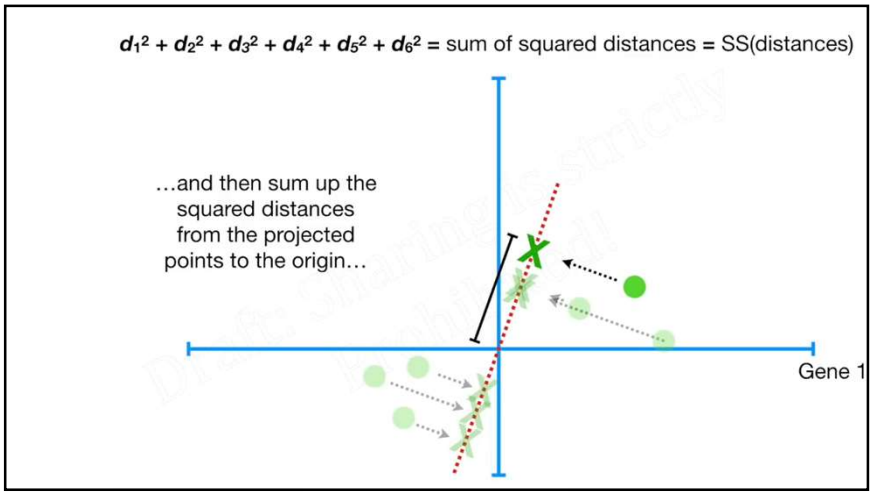


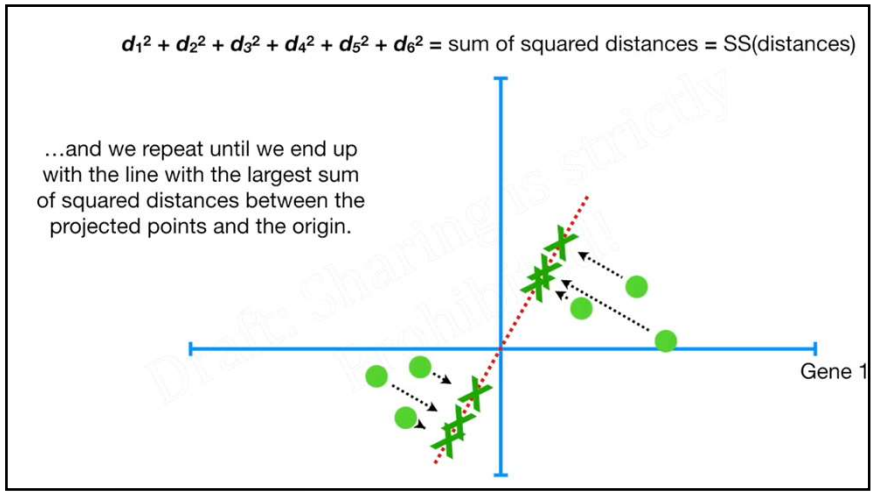
Now we rotate the line…

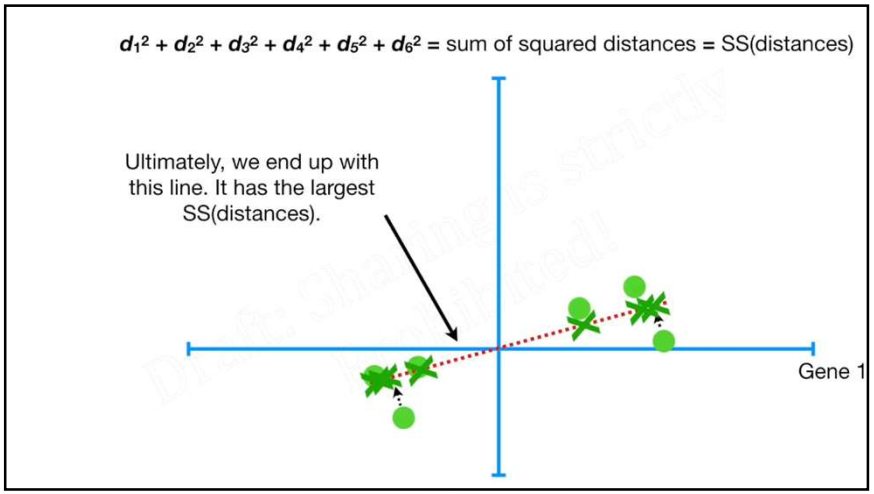


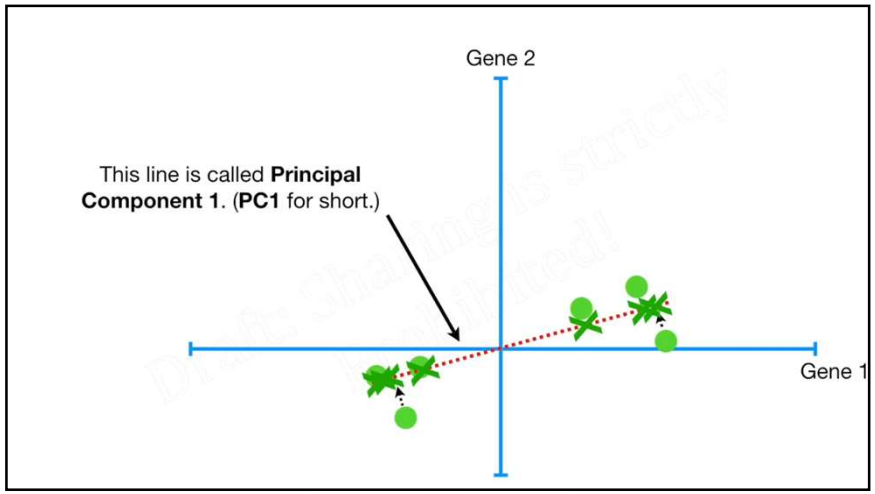
Draft. Sharing is strictly prohibited!

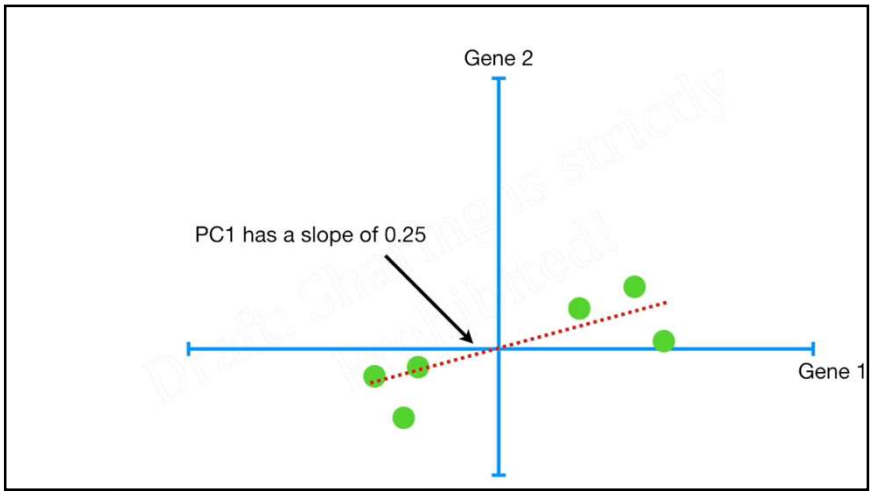

Gene 2
…project the data
onto the line…
Gene 1

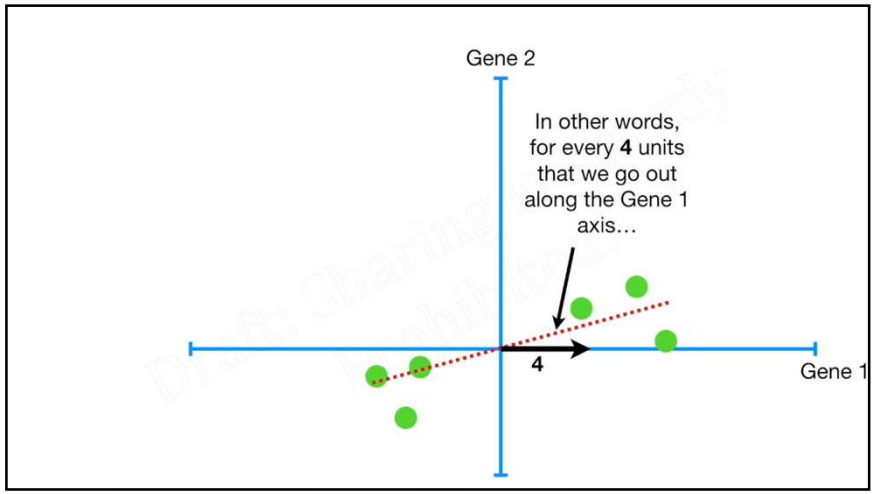

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)
…and then sum up the
squared distances
from the projected
points to the origin…
Gene 1

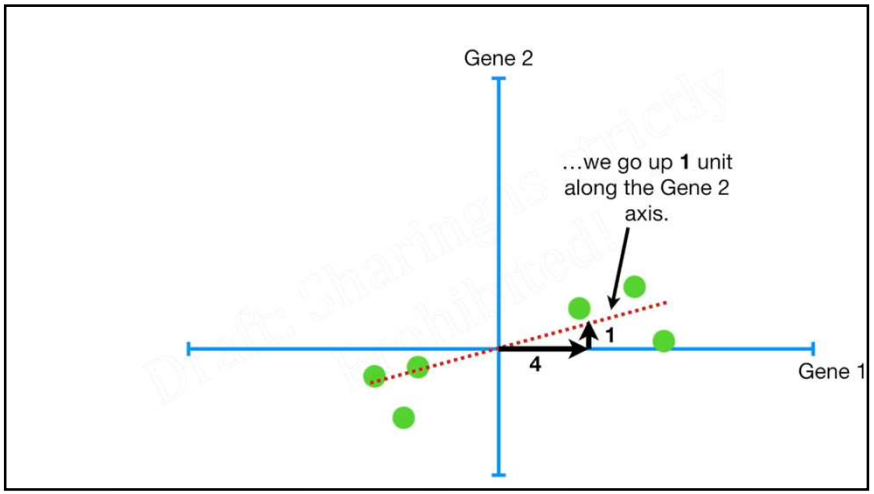

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)
…and we repeat until we end up
with the line with the largest sum
of squared distances between the
projected points and the origin.
Gene 1

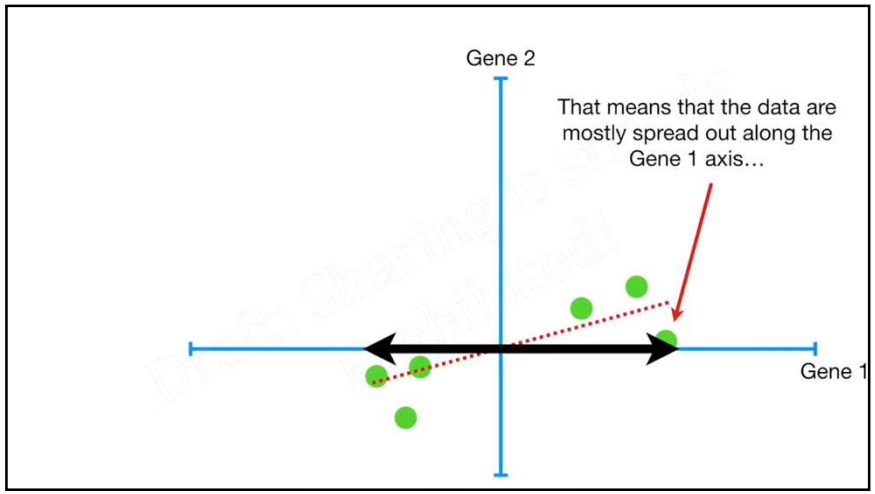

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)
Ultimately, we end up with
this line. It has the largest
SS(distances).
Gene 1

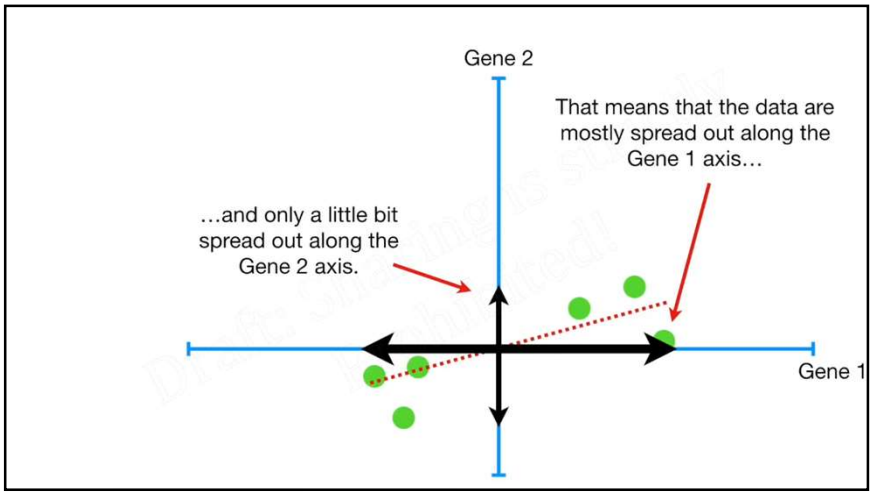

Gene 2
This line is called **Principal Component 1**. (**PC1** for short.)
Gene 1

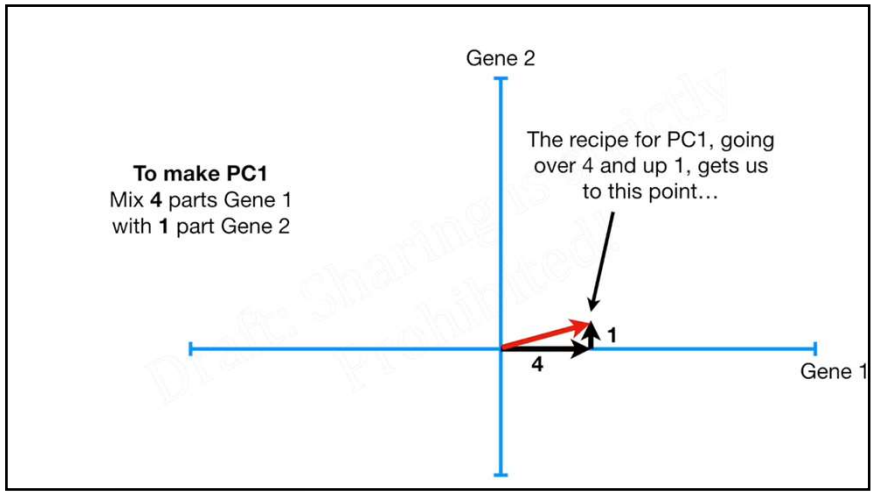

Gene 2
PC1 has a slope of 0.25
Gene 1

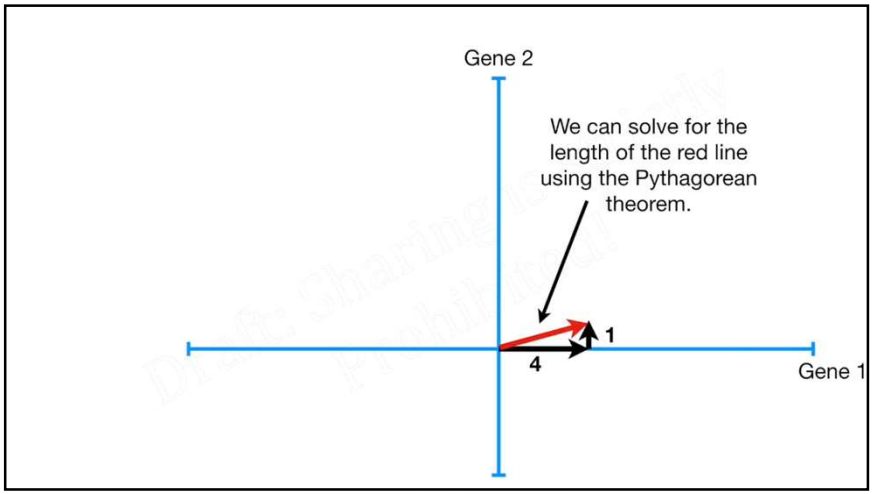

Gene 2
In other words, for every **4** units that we go out along the Gene 1 axis…
4
Gene 1

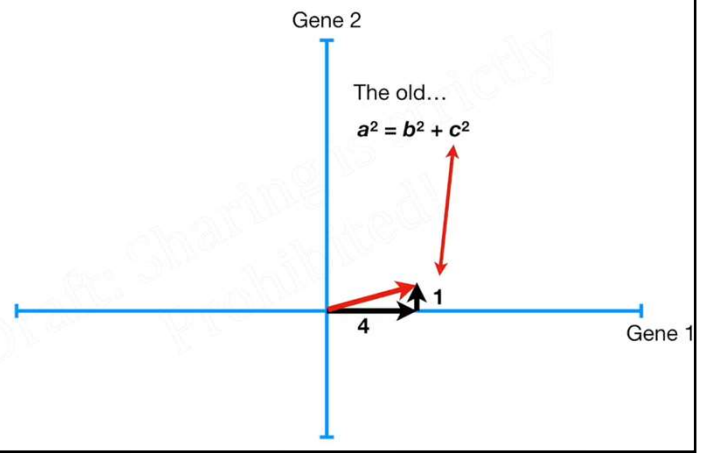

Gene 2
…we go up **1** unit along the Gene 2 axis.
1
4
Gene 1

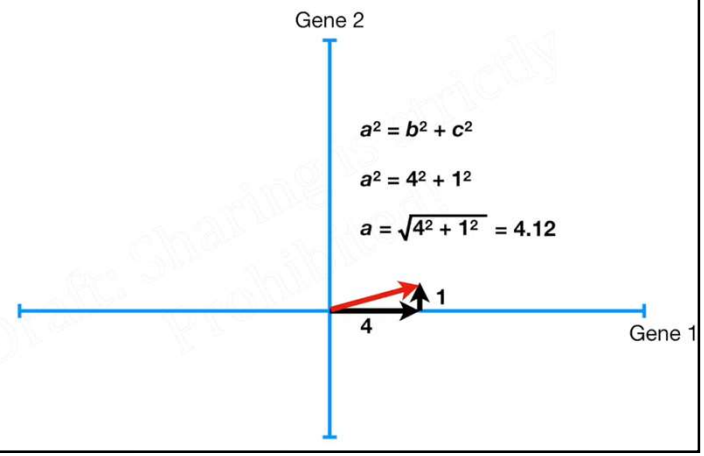

Gene 2
That means that the data are mostly spread out along the Gene 1 axis…
Gene 1

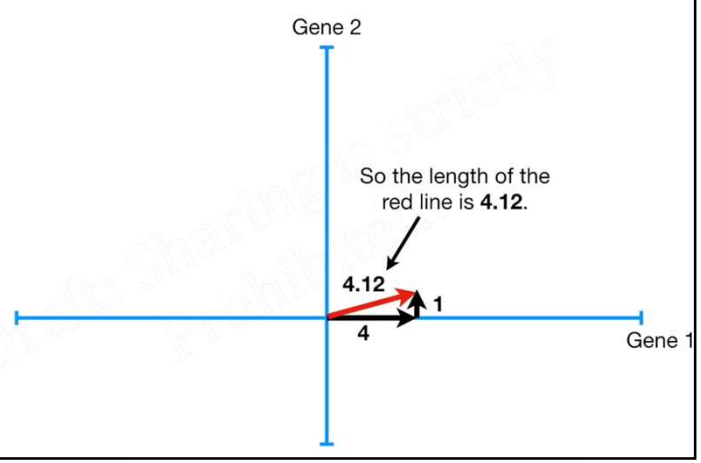

Gene 2
That means that the data are mostly spread out along the Gene 1 axis…
…and only a little bit spread out along the Gene 2 axis.
Gene 1

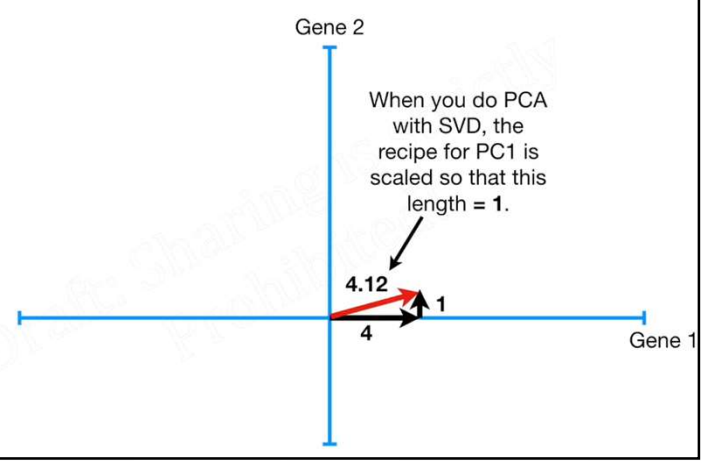

Gene 2
To make PC1
Mix 4 parts Gene 1
with 1 part Gene 2
The recipe for PC1, going over 4 and up 1, gets us to this point…
1
4
Gene 1


Gene 2
We can solve for the length of the red line using the Pythagorean theorem.
1
4
Gene 1


Gene 2
The old…
$a^2 = b^2 + c^2$
1
4
Gene 1


Gene 2
$a^2 = b^2 + c^2$
$a^2 = 4^2 + 1^2$
$a = \sqrt{4^2 + 1^2} = 4.12$
1
4
Gene 1

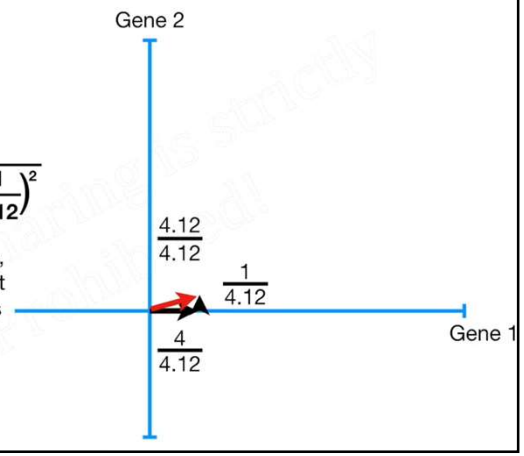

Gene 2
So the length of the red line is **4.12**.
4.12
1
4
Gene 1

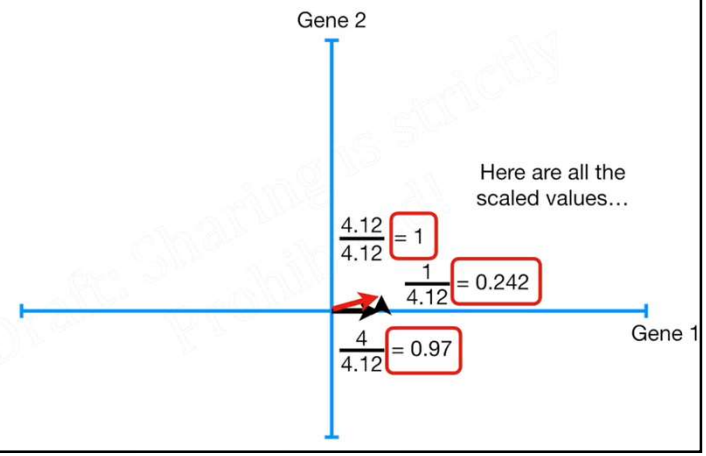

Gene 2
When you do PCA with SVD, the recipe for PC1 is scaled so that this length = **1**.
4.12
1
4
Gene 1

$$\frac{4.12}{4.12} = \frac{\sqrt{4^2 + 1^2}}{4.12} = \sqrt{\left(\frac{4^2 + 1^2}{4.12^2}\right)}$$

$$= \sqrt{\left(\frac{4}{4.12}\right)^2 + \left(\frac{1}{4.12}\right)^2}$$

For those of you keeping score, here's the math worked out that shows that all we need to do is divide all 3 sides by **4.12**.

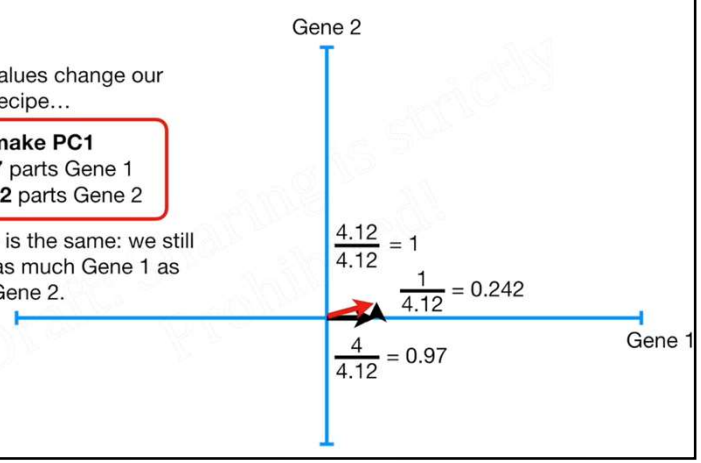




The new values change our recipe…

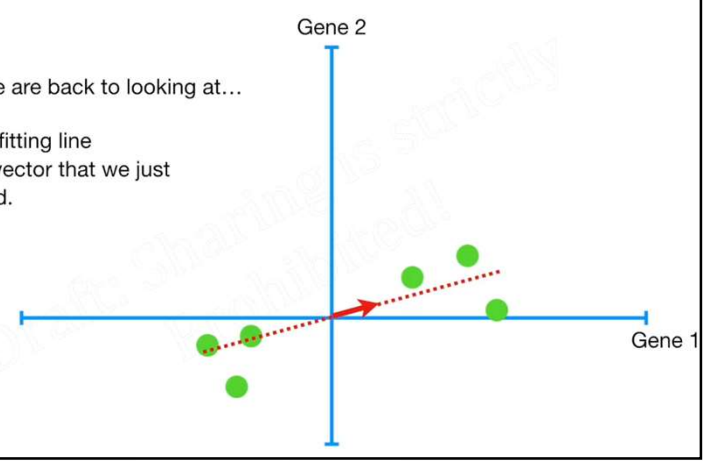
To make PC1
Mix **0.97** parts Gene 1 with **0.242** parts Gene 2

…but the ratio is the same: we still use 4 times as much Gene 1 as Gene 2.

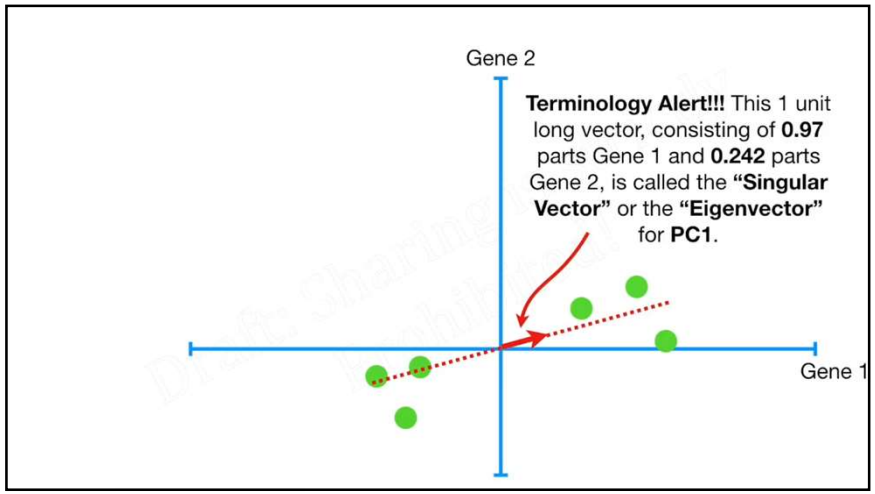


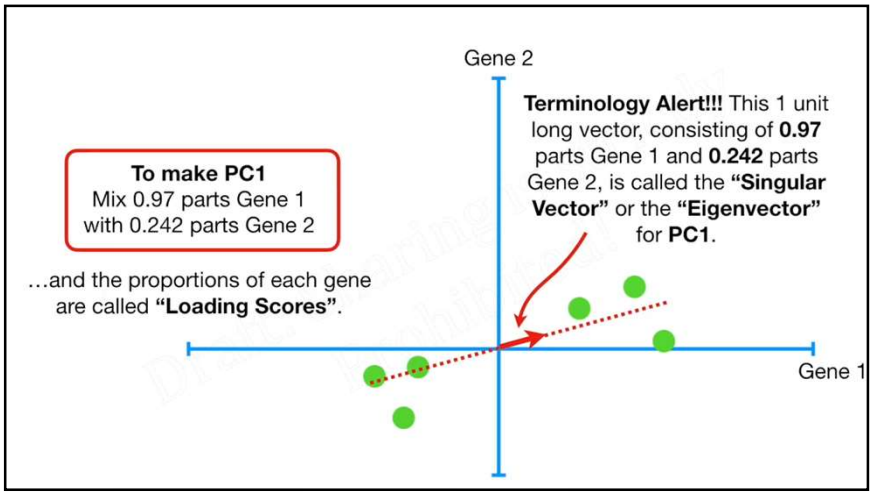
So now we are back to looking at…

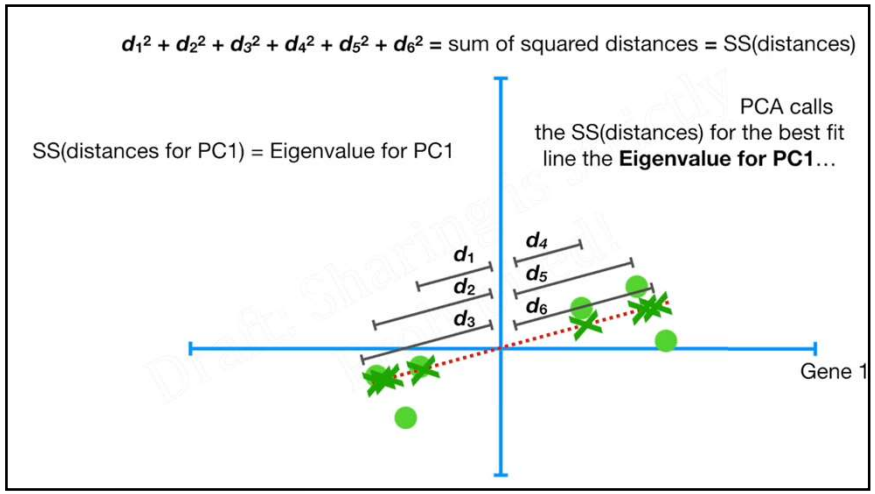
- The data
- The best fitting line
- The unit vector that we just calculated.

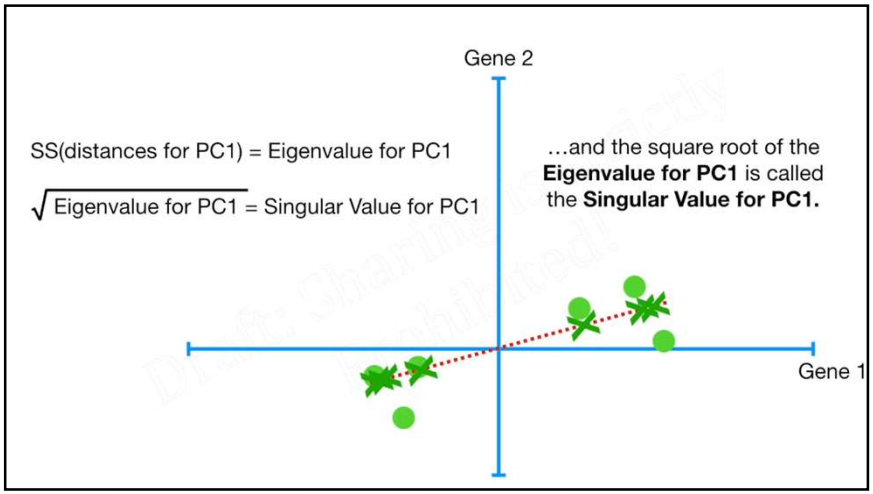


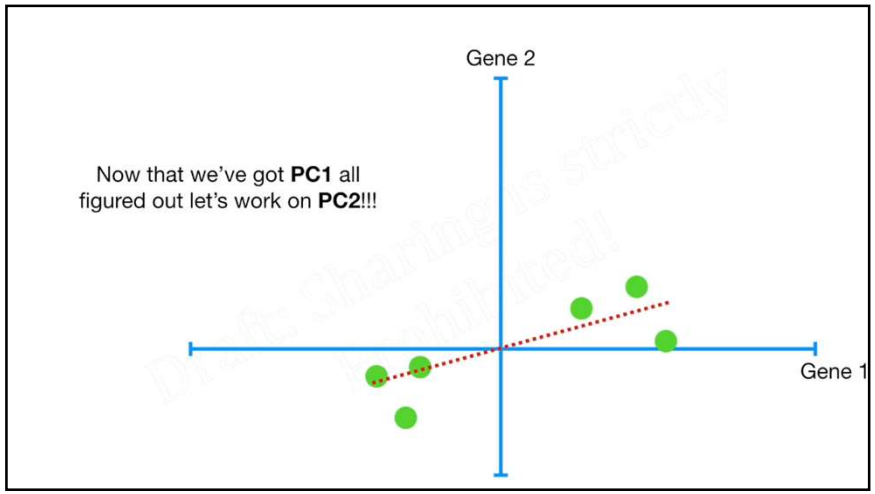
Draft: Sharing is strictly prohibited!

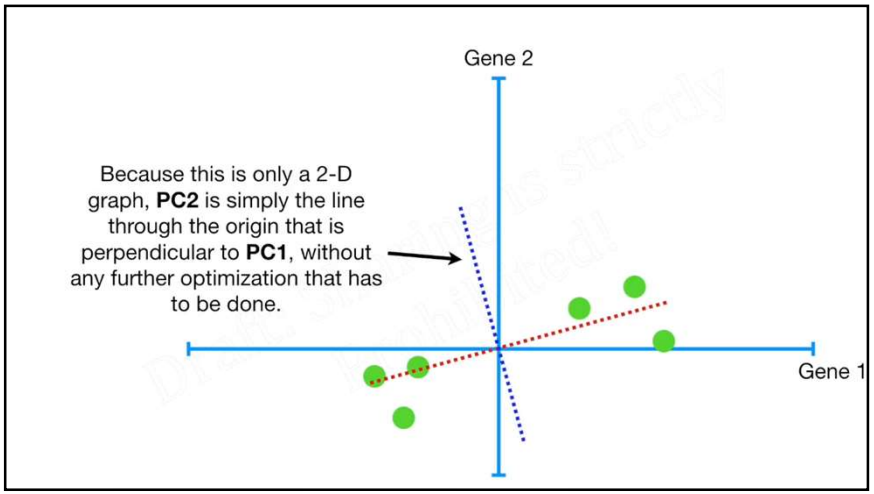

Gene 2
Terminology Alert!!! This 1 unit long vector, consisting of 0.97 parts Gene 1 and 0.242 parts Gene 2, is called the "Singular Vector" or the "Eigenvector" for PC1.
Gene 1

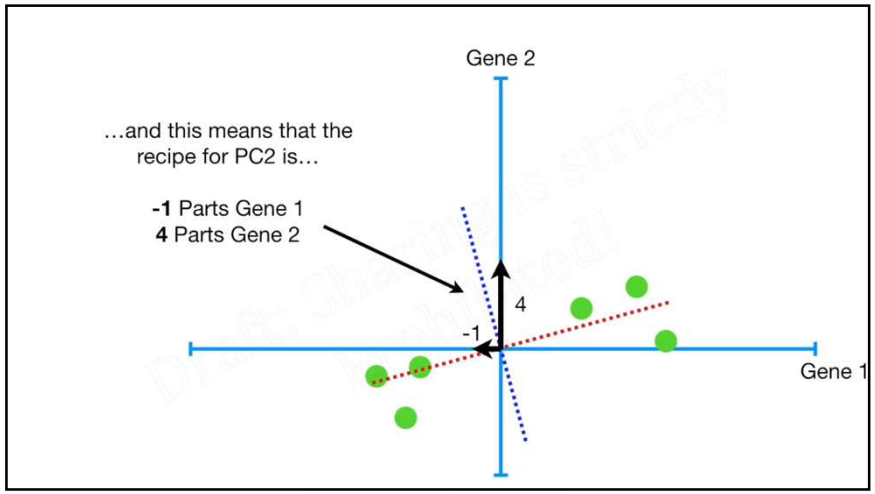

Gene 2
To make PC1
Mix 0.97 parts Gene 1 with 0.242 parts Gene 2
…and the proportions of each gene are called "Loading Scores".
Terminology Alert!!! This 1 unit long vector, consisting of 0.97 parts Gene 1 and 0.242 parts Gene 2, is called the "Singular Vector" or the "Eigenvector" for PC1.
Gene 1

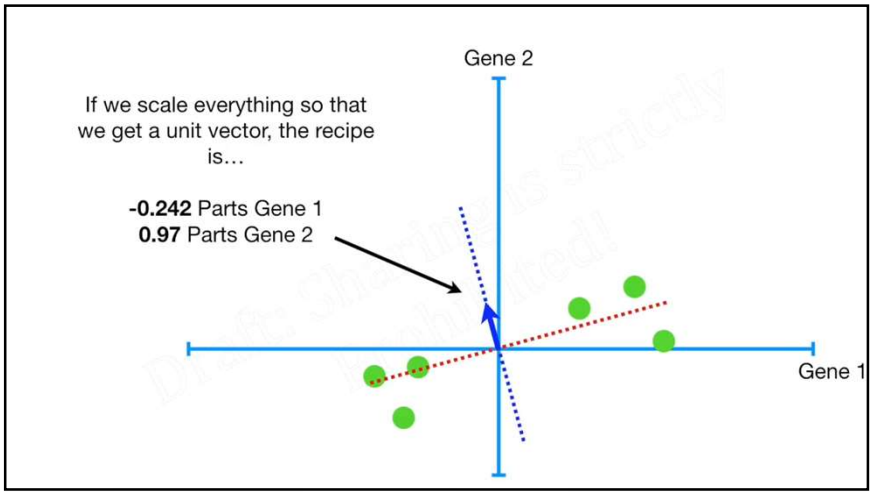

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)
SS(distances for PC1) = Eigenvalue for PC1
PCA calls the SS(distances) for the best fit line the Eigenvalue for PC1…
d_1
d_2
d_3
d_4
d_5
d_6
Gene 1

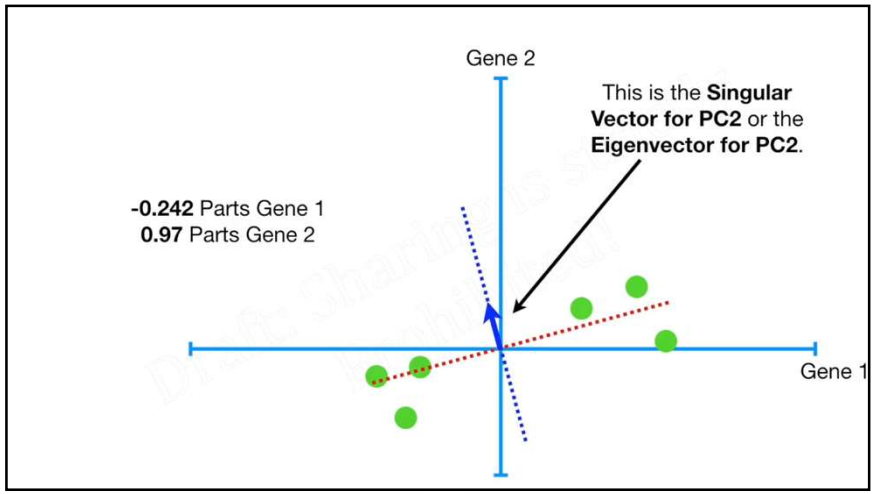

Gene 2
SS(distances for PC1) = Eigenvalue for PC1
$\sqrt{\text{Eigenvalue for PC1}}$ = Singular Value for PC1
…and the square root of the Eigenvalue for PC1 is called the Singular Value for PC1.
Gene 1

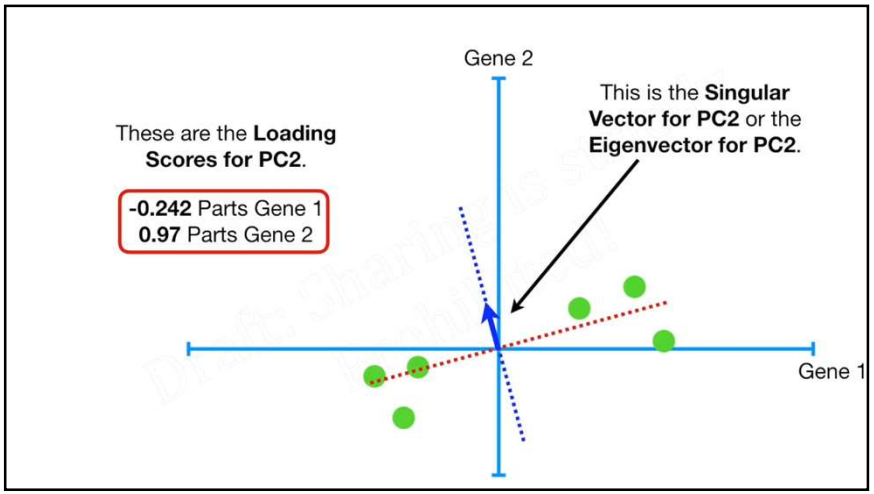

Gene 2
Now that we've got **PC1** all figured out let's work on **PC2**!!!
Gene 1

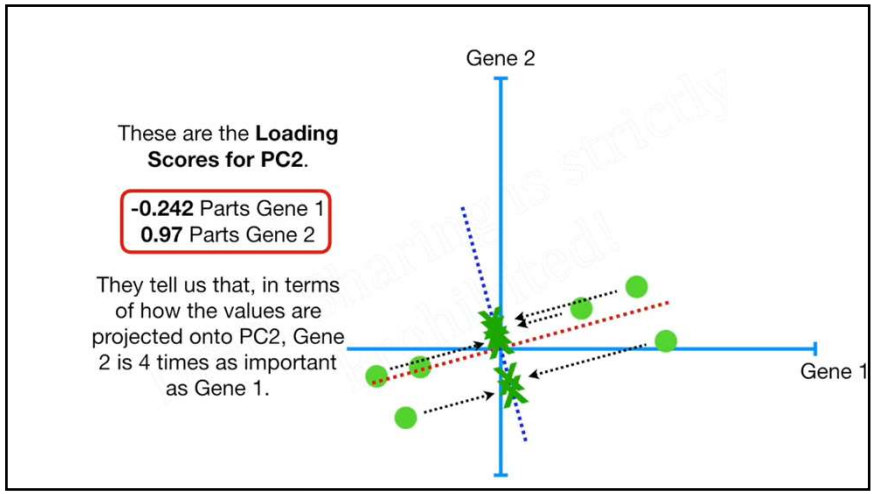

Gene 2
Because this is only a 2-D graph, **PC2** is simply the line through the origin that is perpendicular to **PC1**, without any further optimization that has to be done.
Gene 1

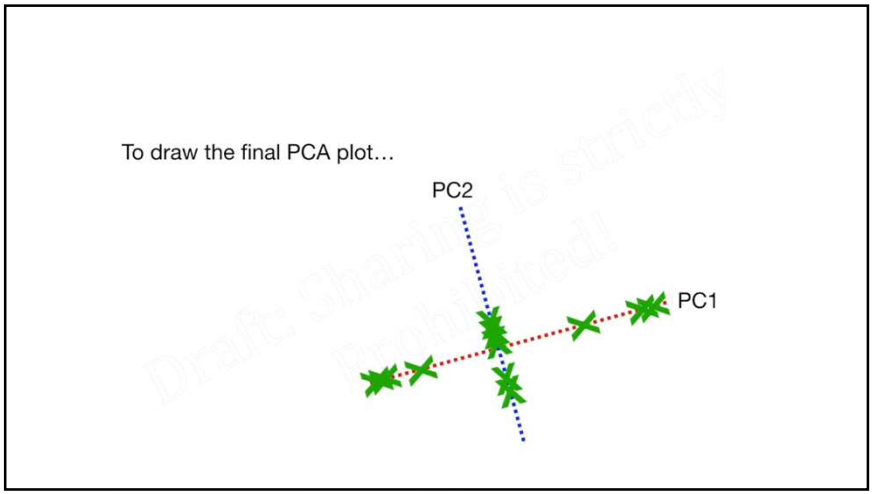

Gene 2
…and this means that the recipe for PC2 is…
-1 Parts Gene 1
4 Parts Gene 2
4
-1
Gene 1


Gene 2
If we scale everything so that we get a unit vector, the recipe is…
-0.242 Parts Gene 1
0.97 Parts Gene 2
Gene 1

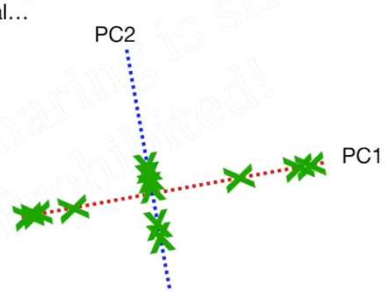

Gene 2
This is the **Singular Vector for PC2** or the **Eigenvector for PC2**.
-0.242 Parts Gene 1
0.97 Parts Gene 2
Gene 1


Gene 2
These are the **Loading Scores for PC2**.
-0.242 Parts Gene 1
0.97 Parts Gene 2
This is the **Singular Vector for PC2** or the **Eigenvector for PC2**.
Gene 1


Gene 2
These are the **Loading Scores for PC2**.
-0.242 Parts Gene 1
0.97 Parts Gene 2
They tell us that, in terms of how the values are projected onto PC2, Gene 2 is 4 times as important as Gene 1.
Gene 1

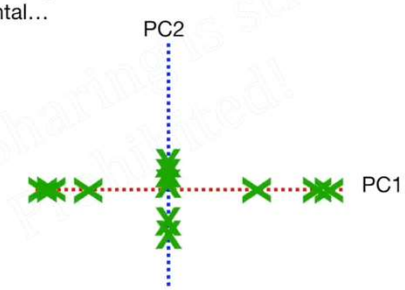

To draw the final PCA plot…
PC2
PC1

We simply rotate everything so that PC1 is horizontal…



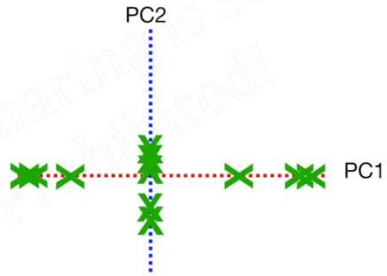
Draft: Sharing is strictly prohibited!

We simply rotate everything so that PC1 is horizontal…



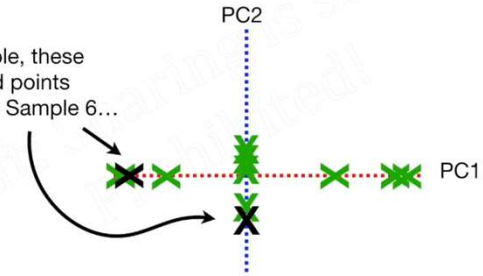
Draft: Sharing is strictly prohibited!

…then we use the projected points to find where the samples go in the PCA plot.

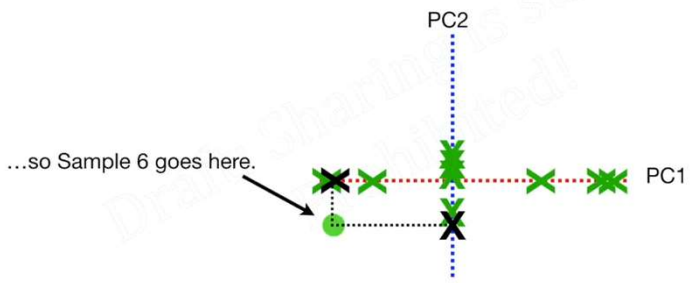


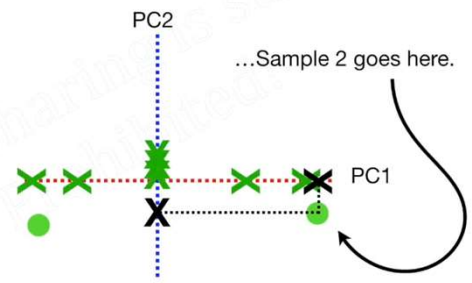
Draft: Sharing is strictly prohibited!

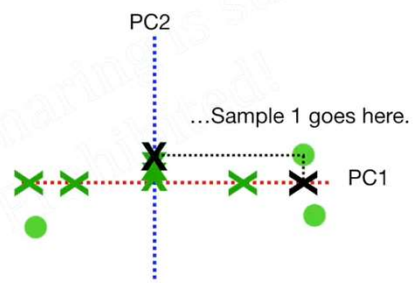
For example, these projected points correspond to Sample 6…

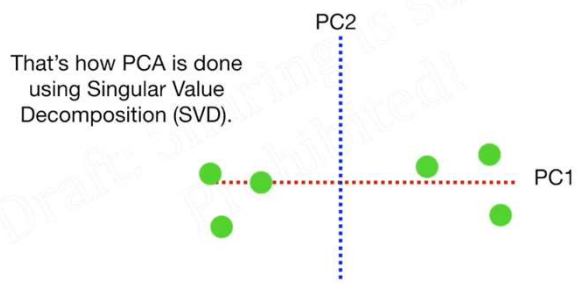


Draft: Sharing is strictly prohibited!



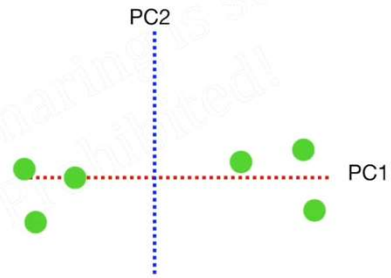






Draft: Sharing is strictly prohibited!

OK - one last thing before we dive into a slightly more complicated example…

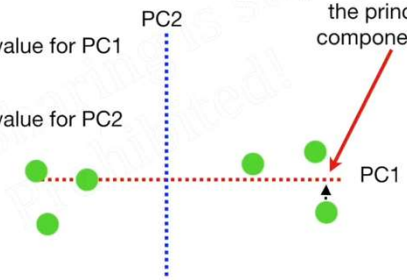


Remember the eigenvalues?

SS(distances for PC1) = Eigenvalue for PC1

SS(distances for PC2) = Eigenvalue for PC2

We got those by projecting the data onto the principal components…

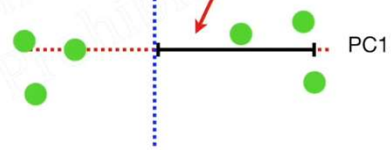


Remember the eigenvalues?

SS(distances for PC1) = Eigenvalue for PC1

SS(distances for PC2) = Eigenvalue for PC2

..measuring the distances to the origin…



Remember the eigenvalues?

SS(distances for PC1) = Eigenvalue for PC1

SS(distances for PC2) = Eigenvalue for PC2

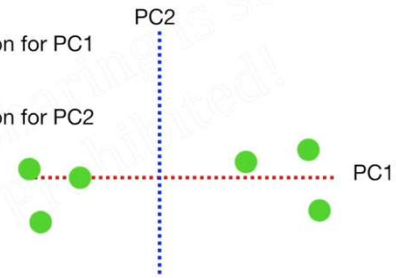
…then squaring and adding them together.

Draft: Sharing is strictly prohibited!

We can convert them into variation around the origin (0, 0) by dividing by the sample size minus 1 (i.e. n - 1).

$$\frac{\text{SS(distances for PC1)}}{n - 1} = \text{Variation for PC1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1} = \text{Variation for PC2}$$

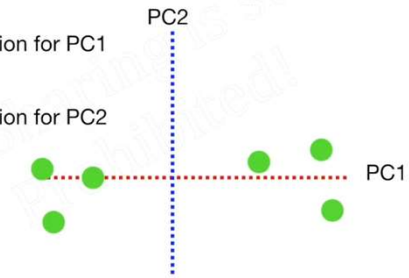


Draft: Warnings strictly prohibited!

For the sake of the example, imagine that the Variation for **PC1 = 15**, and the variation for **PC2 = 3**.

$$\frac{\text{SS(distances for PC1)}}{n - 1} = \text{Variation for PC1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1} = \text{Variation for PC2}$$



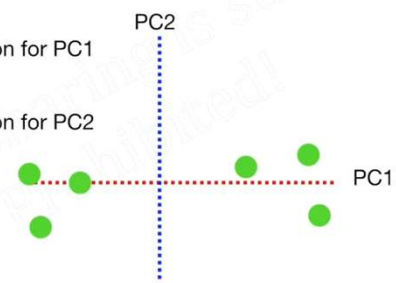
Draft: Warnings strictly prohibited!

For the sake of the example, imagine that the Variation for **PC1 = 15**, and the variation for **PC2 = 3**.

That means that the total variation around both PCs is **15 + 3 = 18**…

$$\frac{\text{SS(distances for PC1)}}{n - 1} = \text{Variation for PC1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1} = \text{Variation for PC2}$$



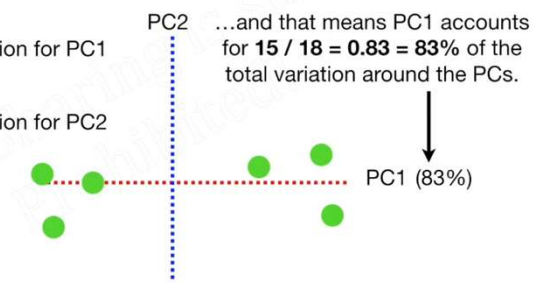
Draft: Warnings strictly prohibited!

For the sake of the example, imagine that the Variation for **PC1 = 15**, and the variation for **PC2 = 3**.

That means that the total variation around both PCs is **15 + 3 = 18**…

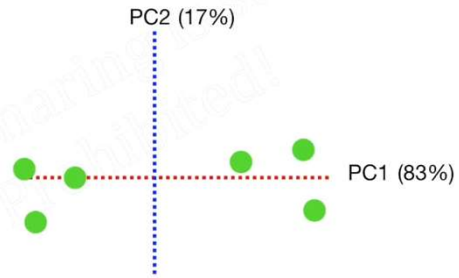
$$\frac{\text{SS(distances for PC1)}}{n - 1} = \text{Variation for PC1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1} = \text{Variation for PC2}$$

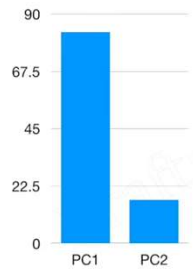


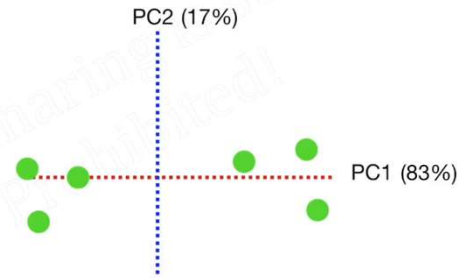
Draft: Warnings strictly prohibited!

PC2 accounts for **3 / 18 = 0.17 = 17%** of the total variation around the PCs.



A **Scree Plot** is a graphical representation of the percentages of variation that each PC accounts for.

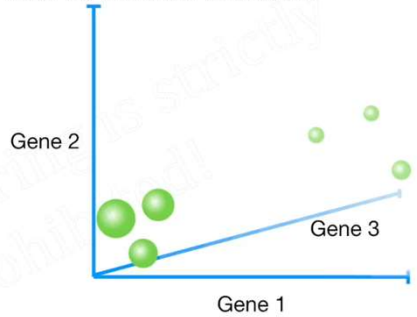




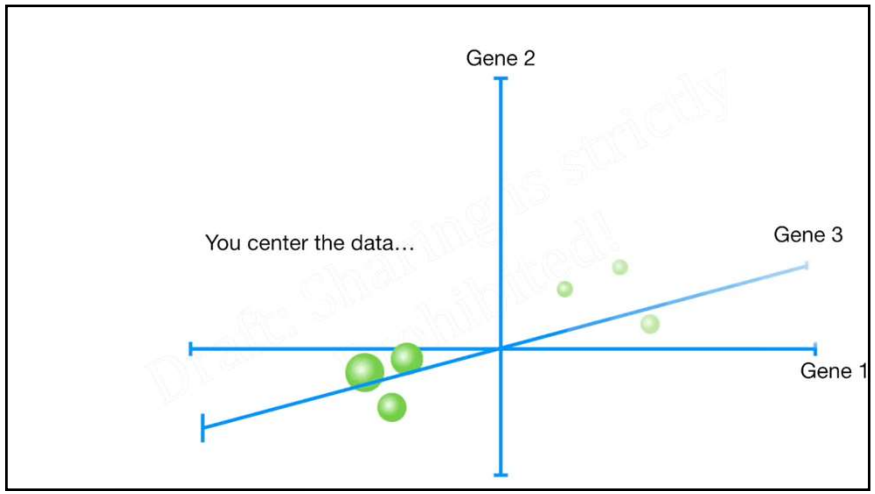
OK - Now let's quickly go through a slightly more complicated example

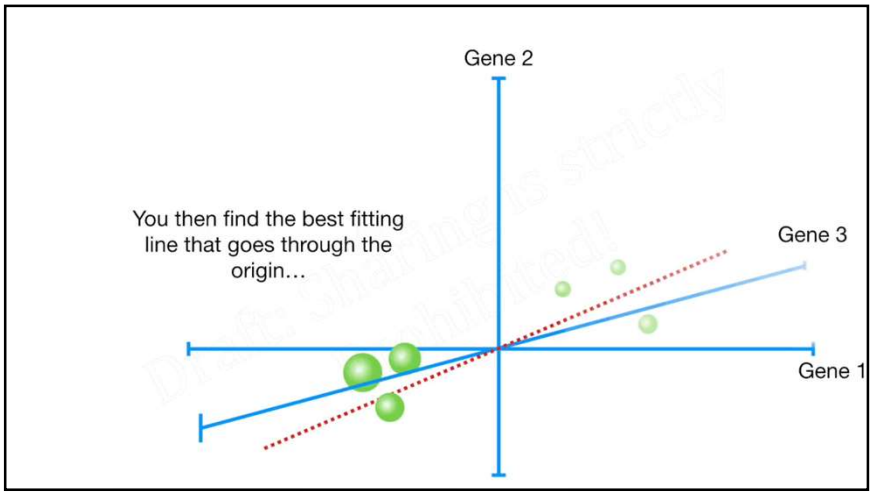
PCA with 3 variables (in this case, that means 3 genes) is pretty much the same as 2 variables…

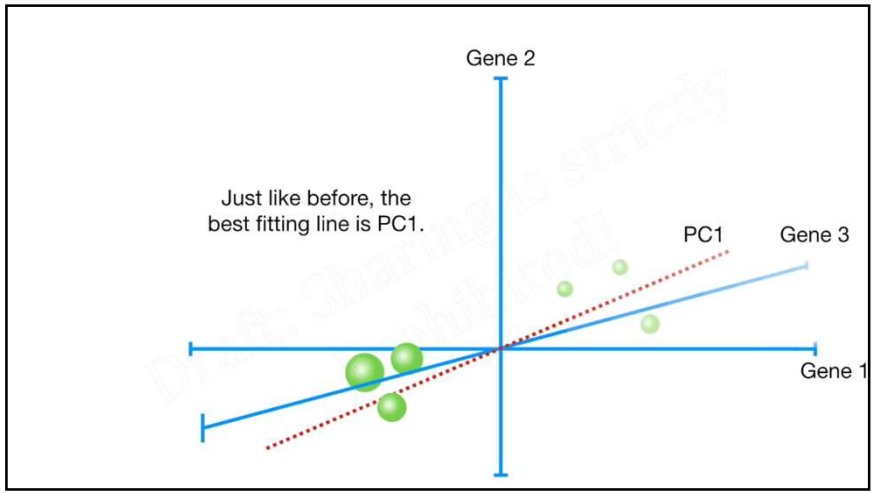
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

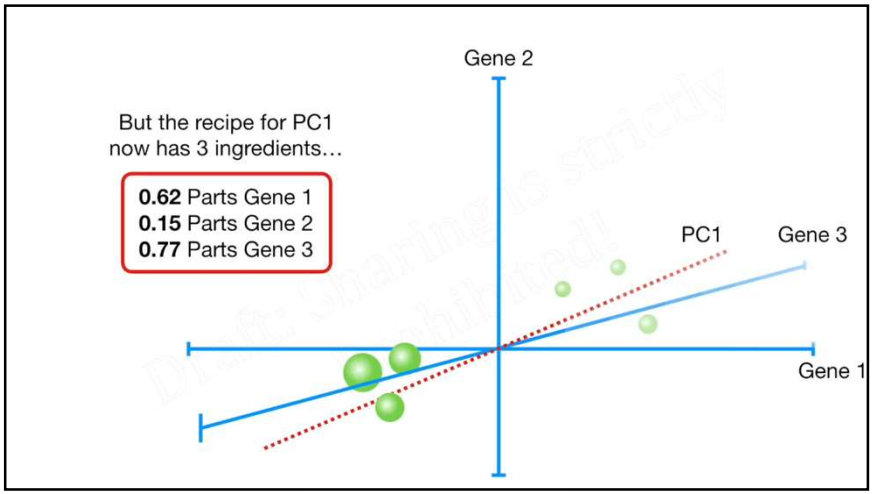


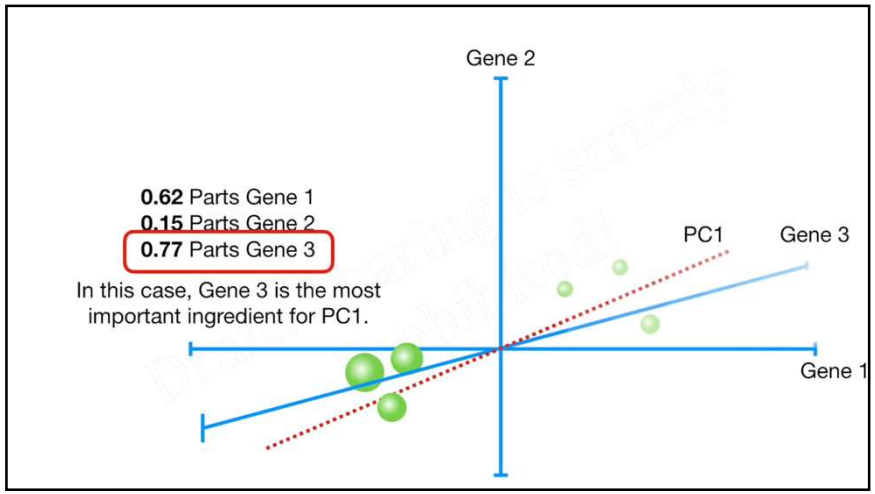
Draft: Sharing is strictly Prohibited!

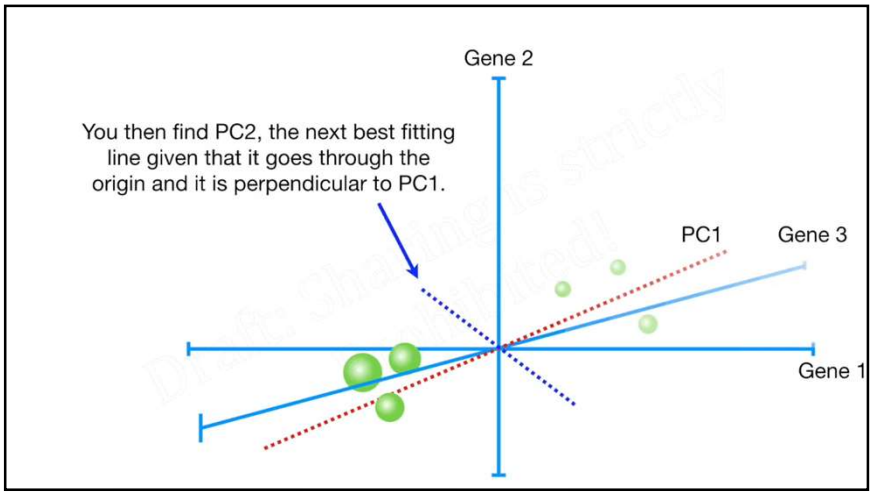

Gene 2
You center the data…
Gene 3
Gene 1

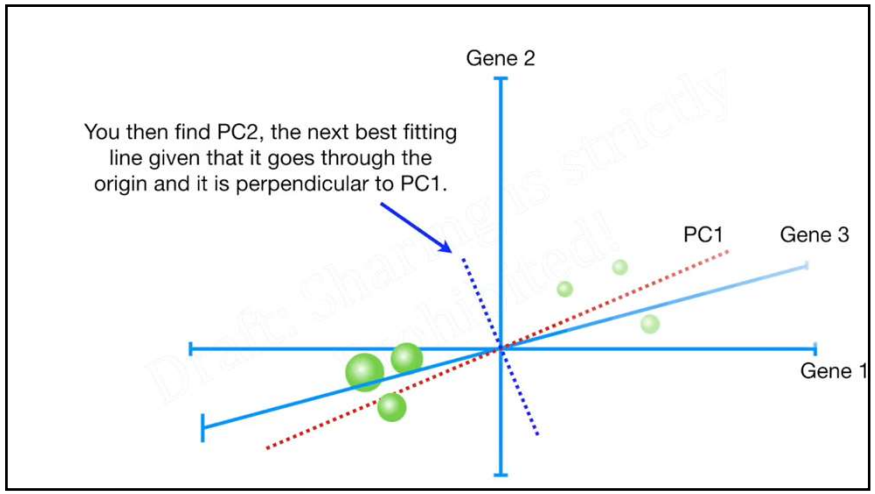

Gene 2
You then find the best fitting line that goes through the origin…
Gene 3
Gene 1

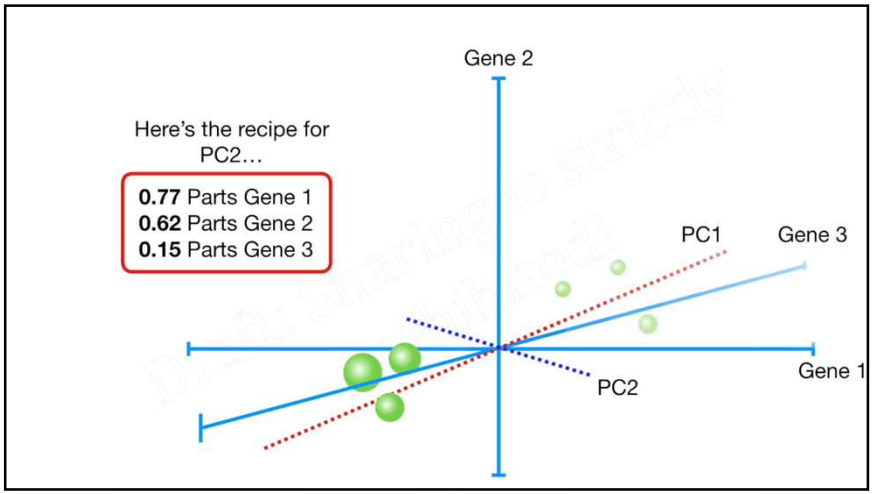

Gene 2
Just like before, the best fitting line is PC1.
PC1
Gene 3
Gene 1

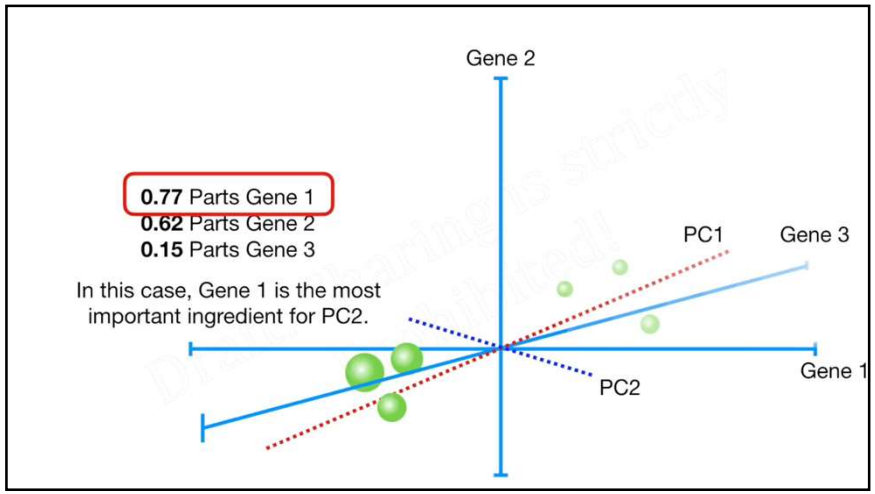

Gene 2
But the recipe for PC1 now has 3 ingredients…
0.62 Parts Gene 1
0.15 Parts Gene 2
0.77 Parts Gene 3
PC1
Gene 3
Gene 1

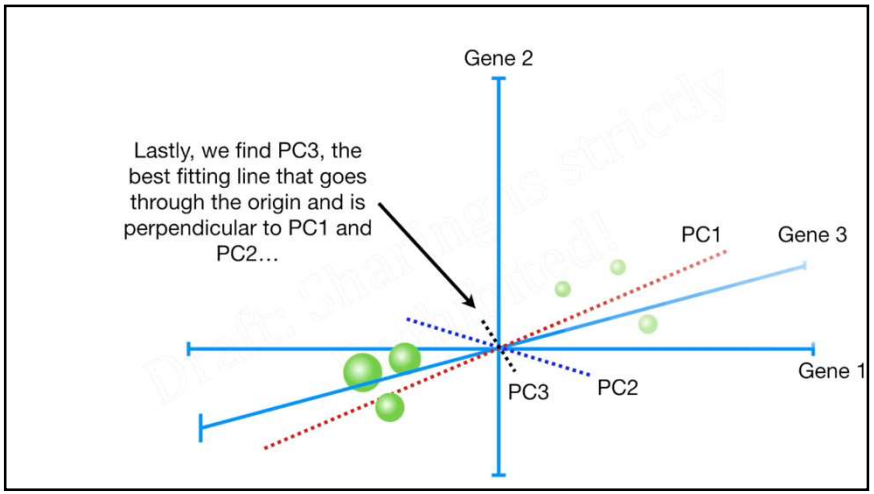

Gene 2
0.62 Parts Gene 1
0.15 Parts Gene 2
0.77 Parts Gene 3
In this case, Gene 3 is the most important ingredient for PC1.
PC1
Gene 3
Gene 1

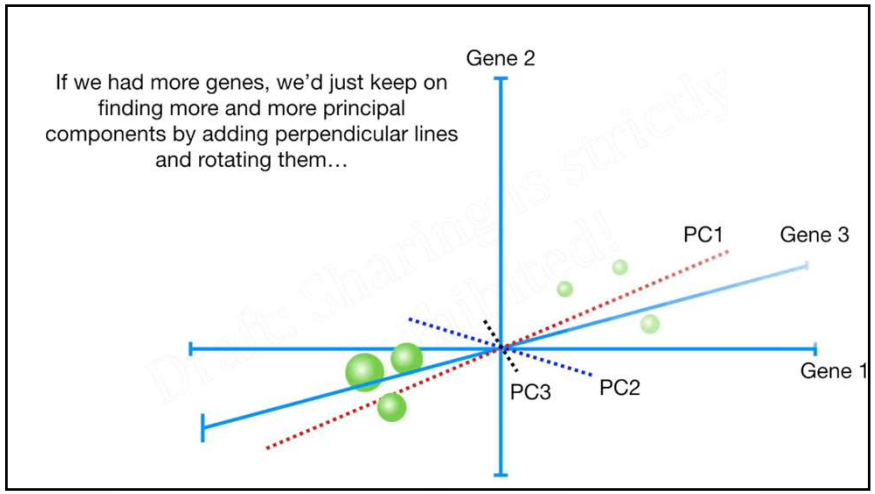

Gene 2
You then find PC2, the next best fitting line given that it goes through the origin and it is perpendicular to PC1.
PC1
Gene 3
Gene 1

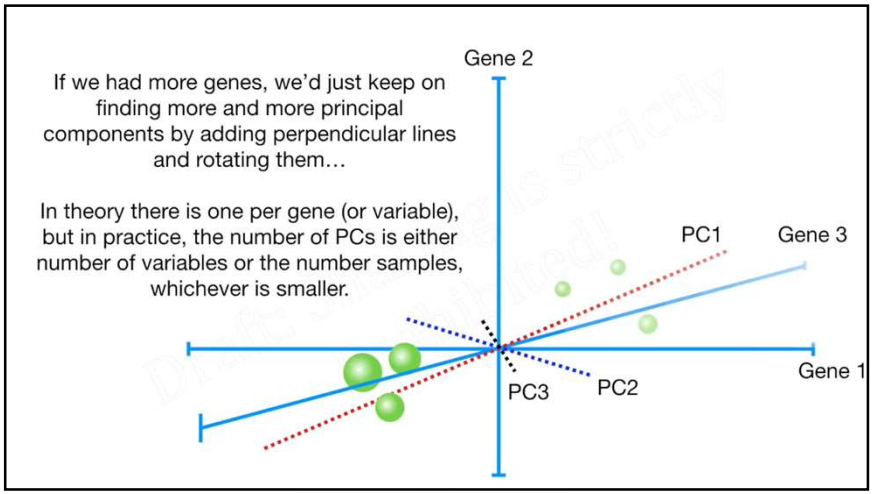

Gene 2
You then find PC2, the next best fitting line given that it goes through the origin and it is perpendicular to PC1.
PC1
Gene 3
Gene 1


Gene 2
Here's the recipe for PC2…
0.77 Parts Gene 1
0.62 Parts Gene 2
0.15 Parts Gene 3
PC1
Gene 3
PC2
Gene 1

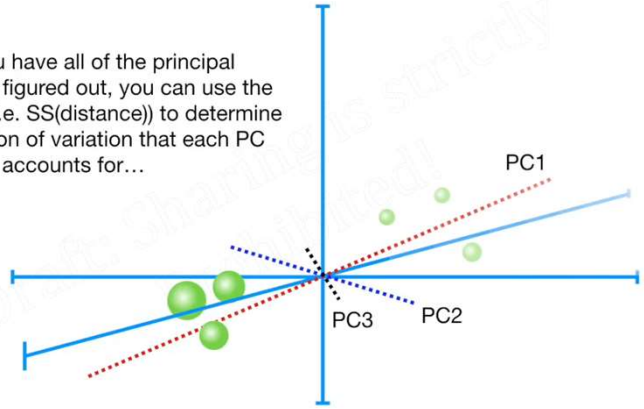

Gene 2
0.77 Parts Gene 1
0.62 Parts Gene 2
0.15 Parts Gene 3
In this case, Gene 1 is the most important ingredient for PC2.
PC1
Gene 3
PC2
Gene 1


Gene 2
Lastly, we find PC3, the best fitting line that goes through the origin and is perpendicular to PC1 and PC2…
PC1
Gene 3
PC3
PC2
Gene 1

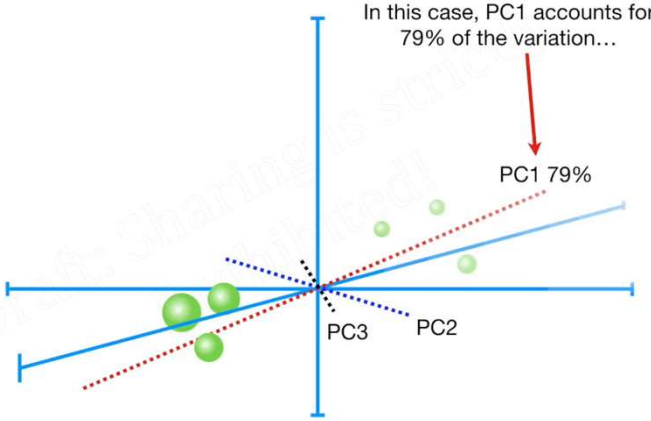

Gene 2
If we had more genes, we'd just keep on finding more and more principal components by adding perpendicular lines and rotating them…
PC1
Gene 3
PC3
PC2
Gene 1


Gene 2
If we had more genes, we'd just keep on finding more and more principal components by adding perpendicular lines and rotating them…
In theory there is one per gene (or variable), but in practice, the number of PCs is either number of variables or the number samples, whichever is smaller.
PC1
Gene 3
PC3
PC2
Gene 1

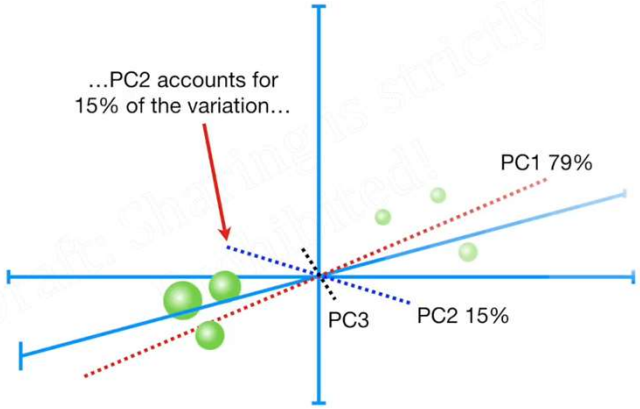
Once you have all of the principal components figured out, you can use the eigenvalues (i.e. SS(distance)) to determine the proportion of variation that each PC accounts for…



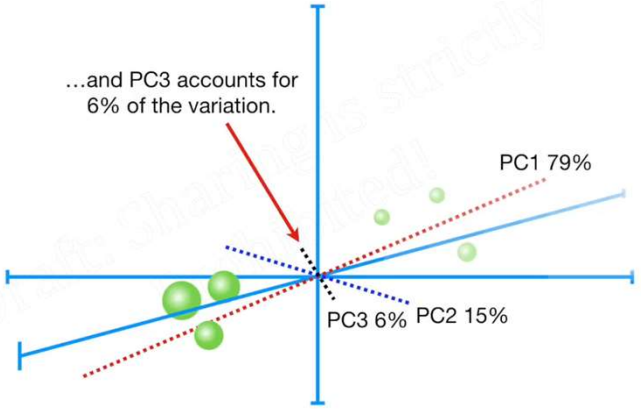
In this case, PC1 accounts for 79% of the variation…



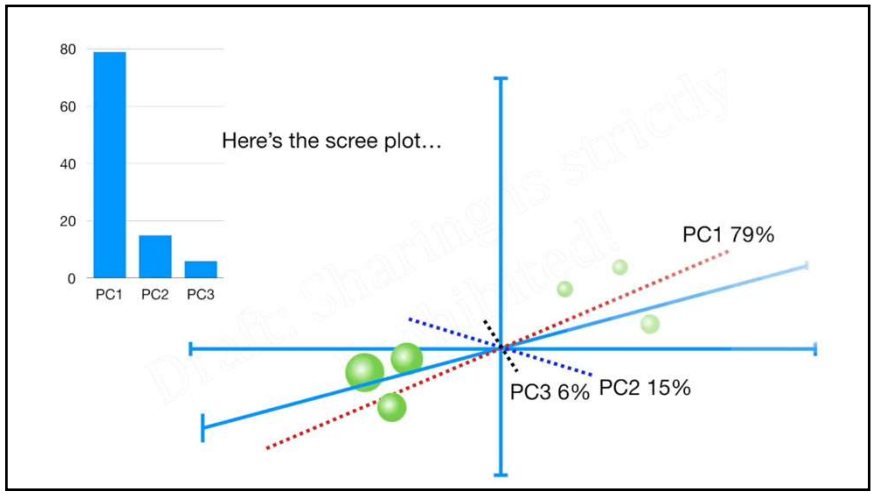
…PC2 accounts for 15% of the variation…

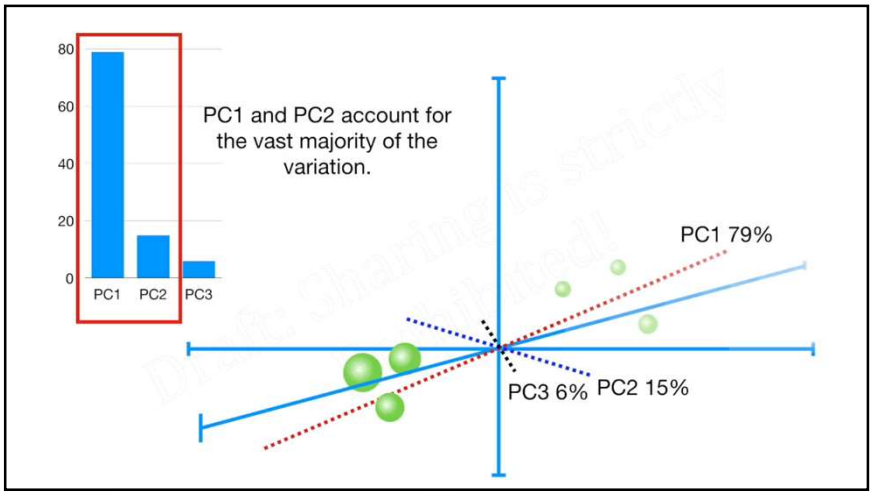


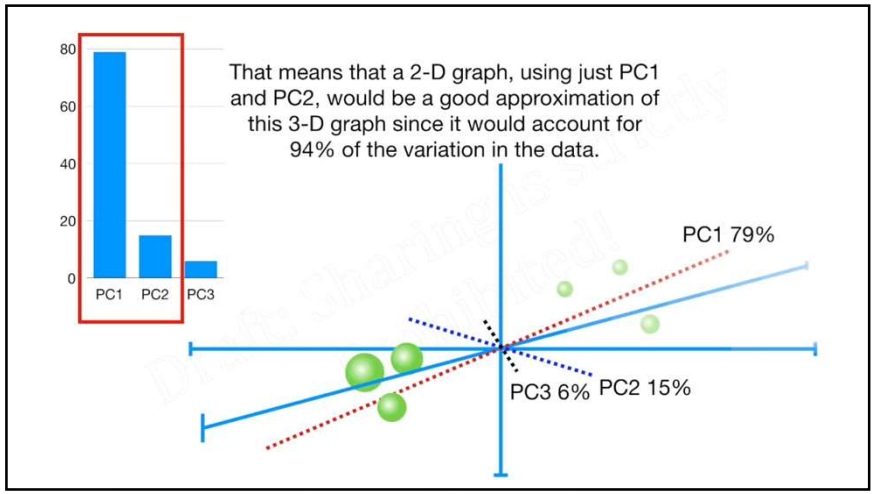
…and PC3 accounts for 6% of the variation.

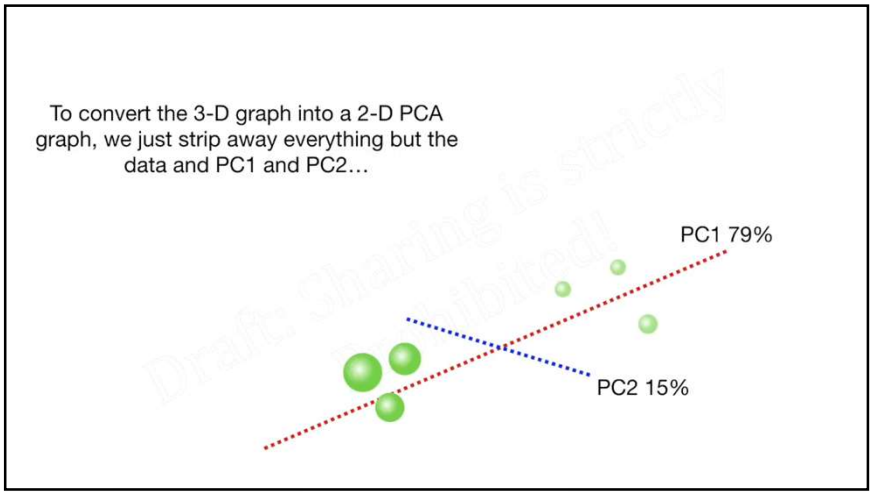


Draft: Sharing is strictly prohibited!

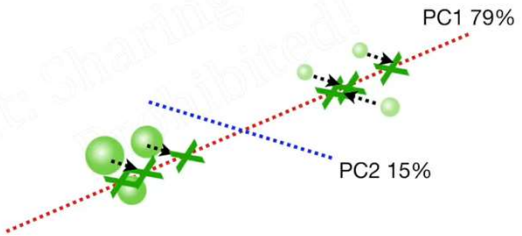

80
60
40
20
0
PC1
PC2
PC3
Here's the scree plot…
PC1 79%
PC3 6%
PC2 15%


80
60
40
20
0
PC1
PC2
PC3
PC1 and PC2 account for the vast majority of the variation.
PC1 79%
PC3 6%
PC2 15%


80
60
40
20
0
PC1
PC2
PC3
That means that a 2-D graph, using just PC1 and PC2, would be a good approximation of this 3-D graph since it would account for 94% of the variation in the data.
PC1 79%
PC3 6%
PC2 15%

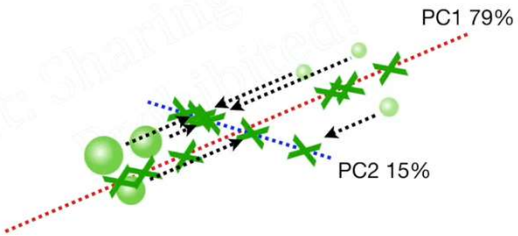

To convert the 3-D graph into a 2-D PCA graph, we just strip away everything but the data and PC1 and PC2…
PC1 79%
PC2 15%

Then project the samples onto PC1…



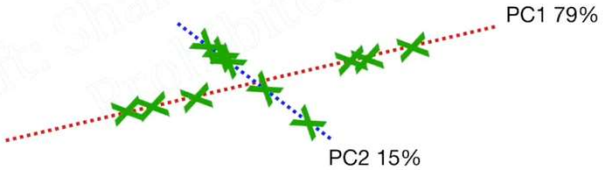
Draft: Sharing is strictly prohibited!

…and PC2.



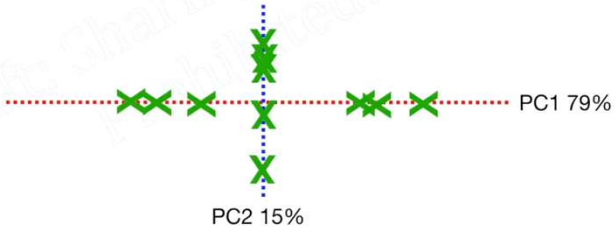
Draft: Sharing is strictly prohibited!

Then we rotate so that PC1 is horizontal and PC2 is vertical (this just makes it easier to look at).

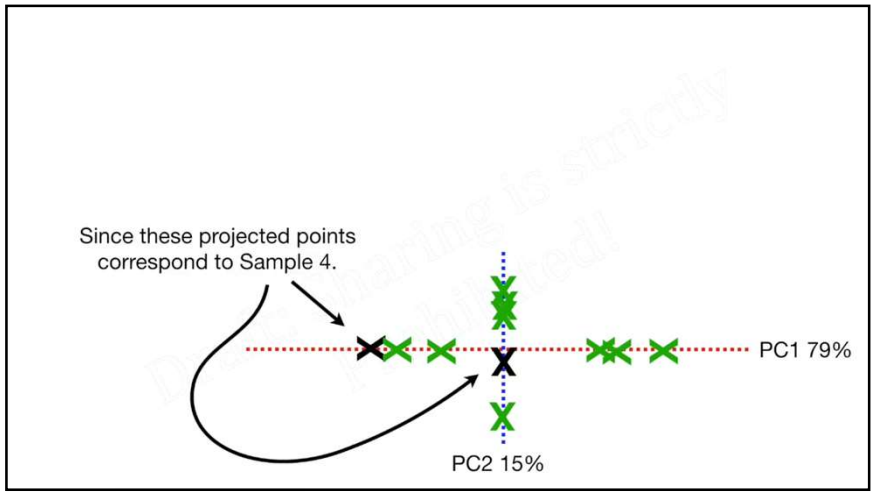


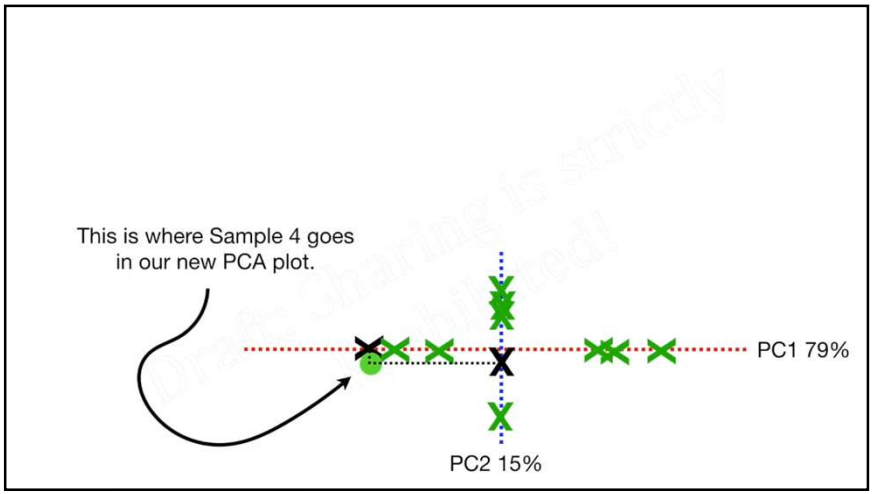
Draft: Sharing is strictly prohibited!

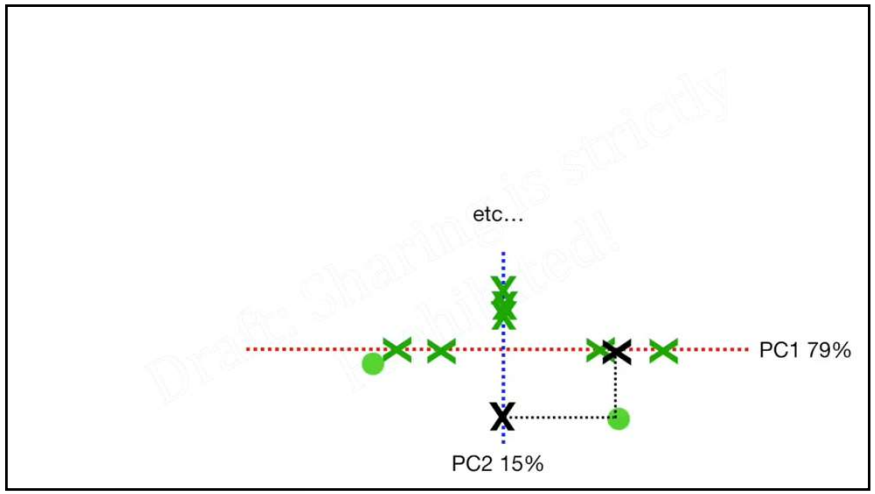
Then we rotate so that PC1 is horizontal and PC2 is vertical (this just makes it easier to look at).

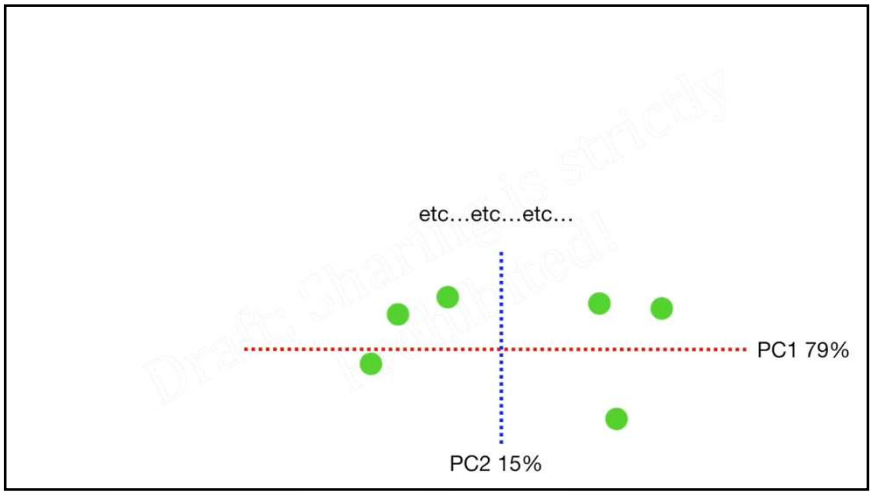


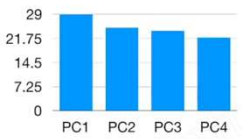
Draft: Sharing is strictly prohibited!


Since these projected points
correspond to Sample 4.
PC1 79%
PC2 15%


This is where Sample 4 goes
in our new PCA plot.
PC1 79%
PC2 15%

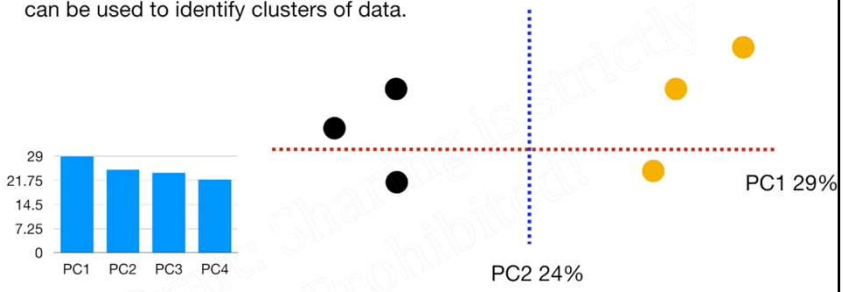

etc…
PC1 79%
PC2 15%


etc…etc…etc…
PC1 79%
PC2 15%

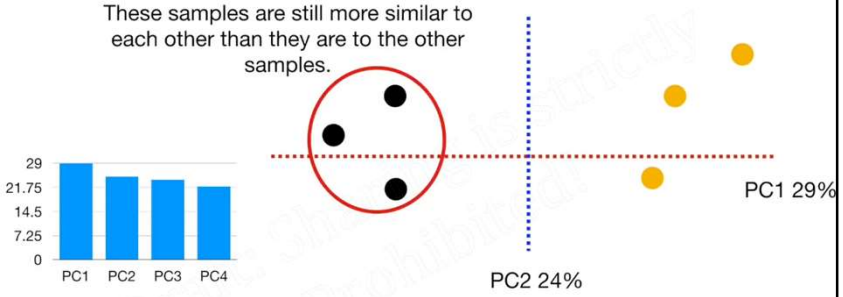


NOTE: If the scree plot looked like this, where PC3 and PC4 account for a substantial amount of variation, then just using the first 2 PCs would not create a very accurate representation of the data.

However, even a noisy PCA plot like this can be used to identify clusters of data.

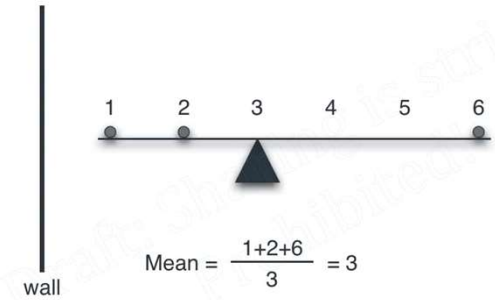


These samples are still more similar to each other than they are to the other samples.



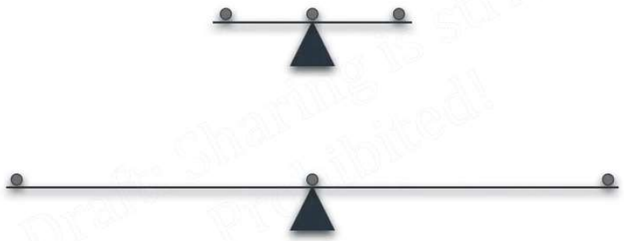
MATHEMATICAL WAY

Mean



$$\text{Mean} = \frac{1+2+6}{3} = 3$$

Variance



Variance

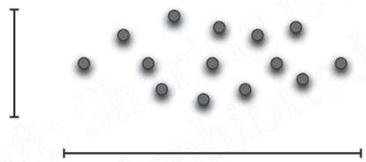
$$\text{Variance} = \frac{1^2+0^2+1^2}{3} = 2/3$$

$$\text{Variance} = \frac{5^2+0^2+5^2}{3} = 50/3$$

Variance?

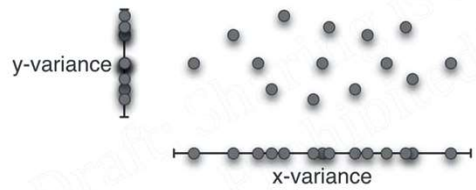
Draft: Sharing is strictly Prohibited!

Variance?



Draft: Sharing is strictly Prohibited!

Variance?



Draft: Sharing is strictly Prohibited!

Variance?

Draft: Sharing is strictly Prohibited!

Variance?

$$\text{x-variance} = \frac{2^2+0^2+2^2}{3} = 8/3$$

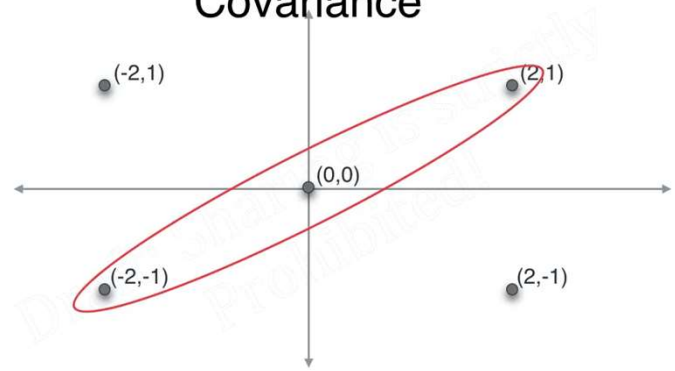
Draft: Sharing is strictly Prohibited!

Variance?

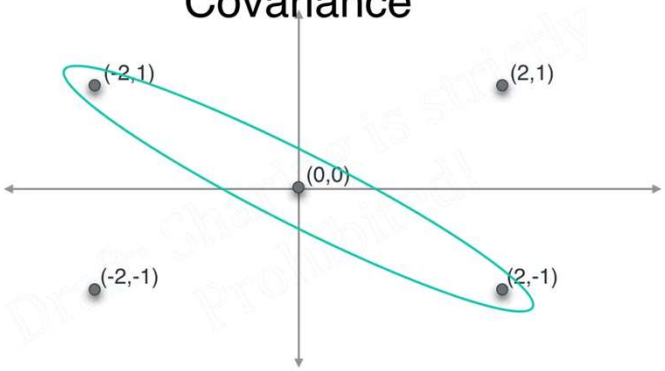
$$\text{x-variance} = \frac{2^2+0^2+2^2}{3} = 8/3$$

$$\text{y-variance} = \frac{1^2+0^2+1^2}{3} = 2/3$$

Covariance

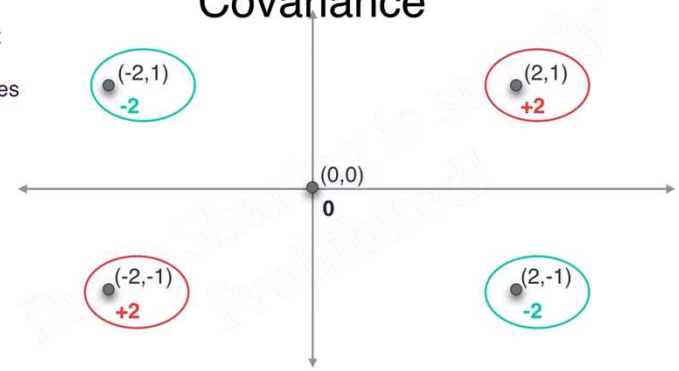


Covariance

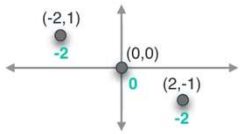


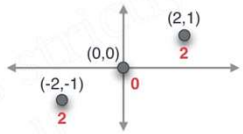
Covariance

Product of coordinates



Covariance

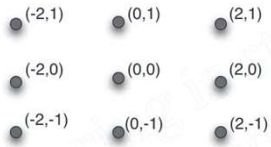




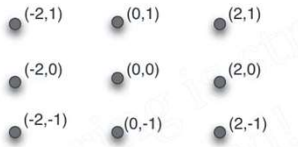
$$\text{covariance} = \frac{(-2) + 0 + (-2)}{3} = -4/3$$

$$\text{covariance} = \frac{2 + 0 + 2}{3} = 4/3$$

Covariance



Covariance



$$\text{covariance} = \frac{-2 + 0 + 2 + 0 + 0 + 0 + 2 + 0 + -2}{9} = 0$$

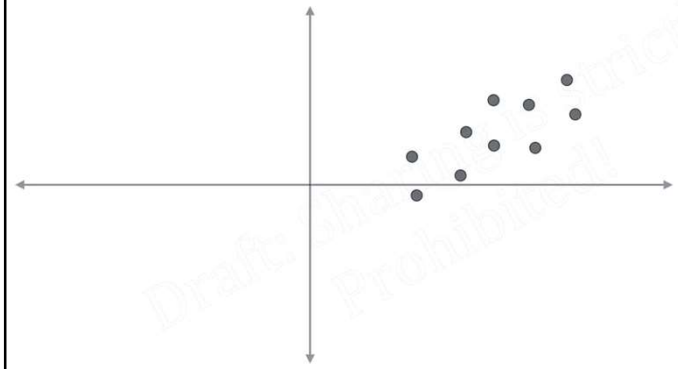
Covariance

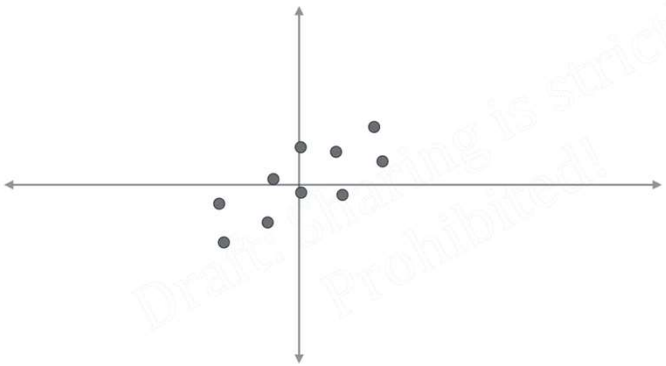
negative covariance

covariance zero (or very small)

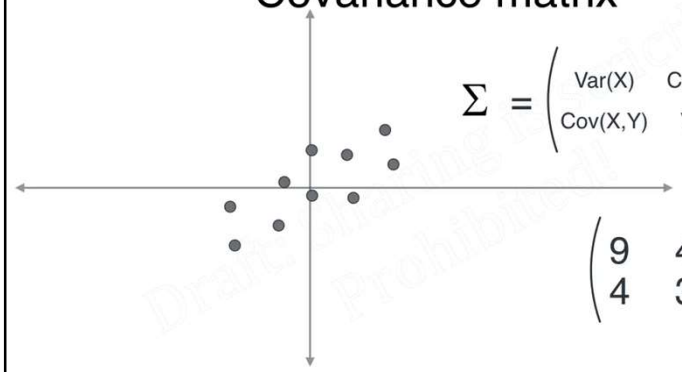
positive covariance

Draft: Sharing is strictly Prohibited!





Covariance matrix



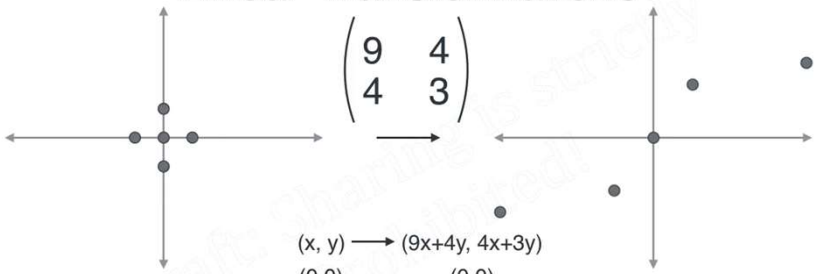
$$\Sigma = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix}$$

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

Draft: Sharing is strictly Prohibited!

Linear Transformations

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

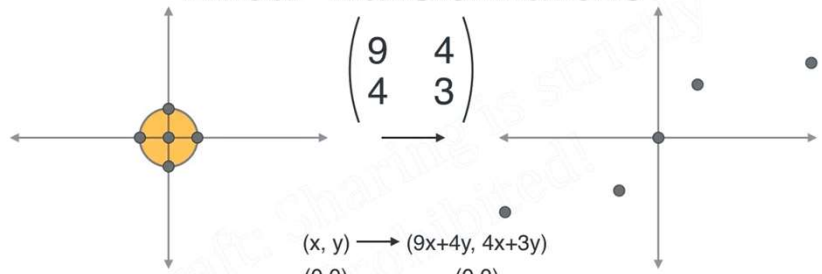


(x, y) ⟶	(9x+4y, 4x+3y)
(0,0)	(0,0)
(1,0)	(9,4)
(0,1)	(4,3)
(-1,0)	(-9,-4)
(0,-1)	(-4,-3)

Draft: Sharing is strictly Prohibited!

Linear Transformations

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

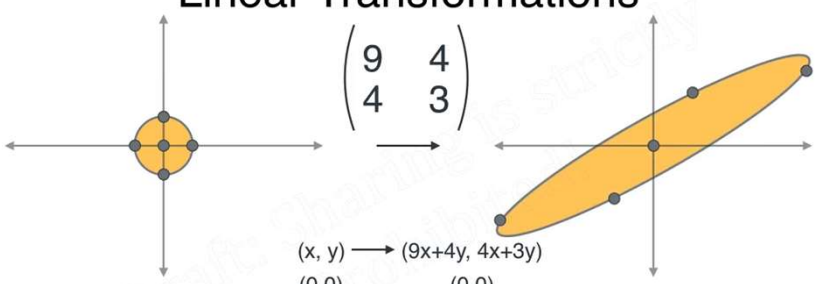


(x, y) ⟶	(9x+4y, 4x+3y)
(0,0)	(0,0)
(1,0)	(9,4)
(0,1)	(4,3)
(-1,0)	(-9,-4)
(0,-1)	(-4,-3)

Draft: Sharing is strictly Prohibited!

Linear Transformations

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

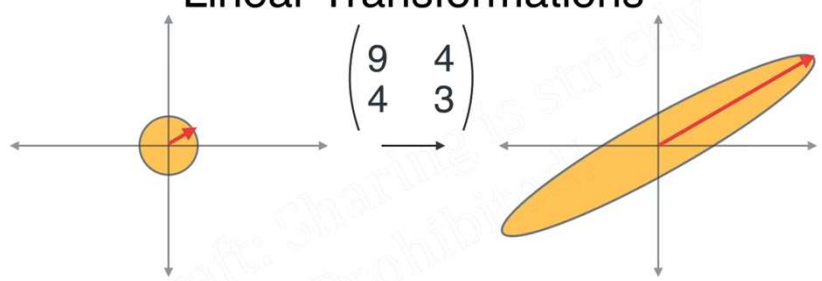


(x, y) ⟶	(9x+4y, 4x+3y)
(0,0)	(0,0)
(1,0)	(9,4)
(0,1)	(4,3)
(-1,0)	(-9,-4)
(0,-1)	(-4,-3)

Draft: Sharing is strictly Prohibited!

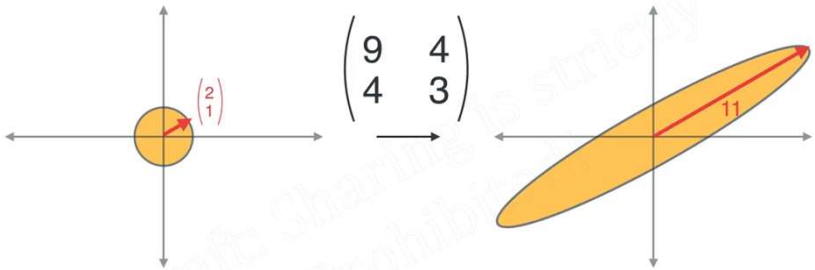
Linear Transformations

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$



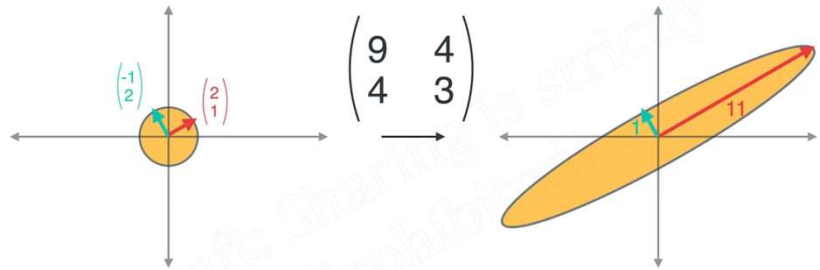
Draft: Sharing is strictly Prohibited!

Linear Transformations



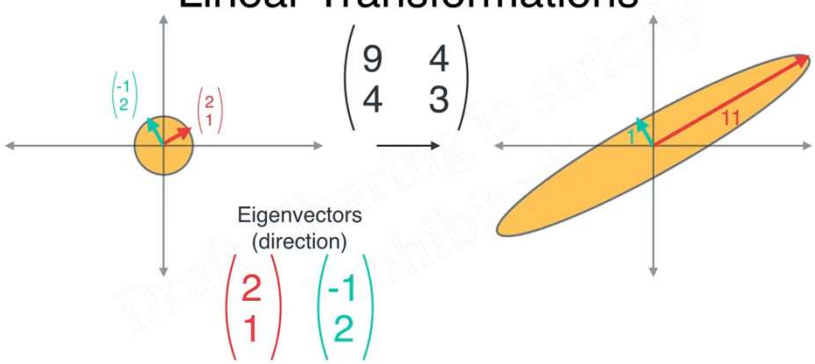
Draft: Sharing is strictly Prohibited

Linear Transformations



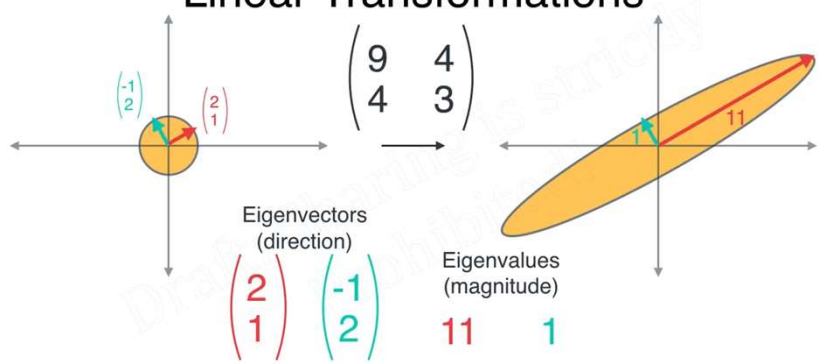
Draft: Sharing is strictly Prohibited

Linear Transformations



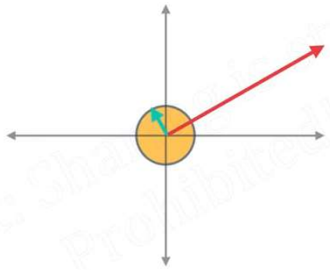
Draft: Sharing is strictly Prohibited

Linear Transformations



Draft: Sharing is strictly Prohibited

Linear Transformations



Eigenvectors

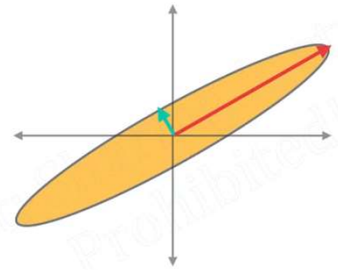
$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$

Eigenvalues

11 1

Draft: Sharing is strictly Prohibited!

Linear Transformations



Eigenvectors (direction)

$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$

Eigenvalues (magnitude)

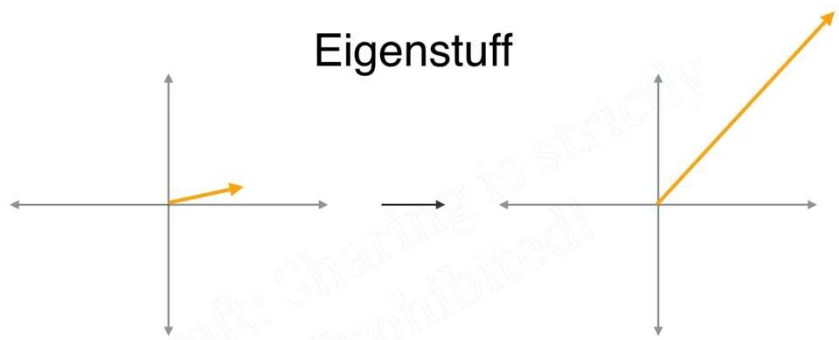
11 1

Draft: Sharing is strictly Prohibited!

Eigenstuff

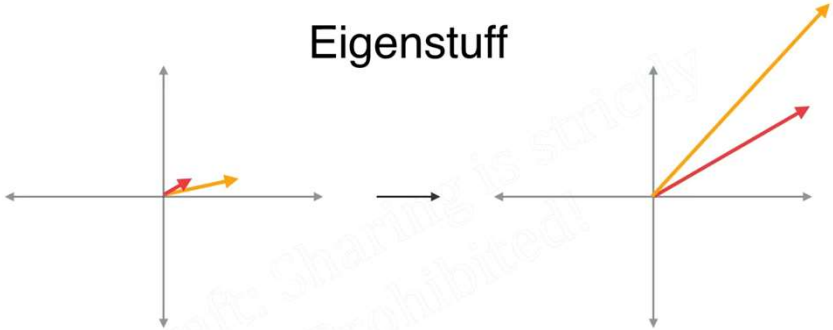
Draft: Sharing is strictly Prohibited!

Eigenstuff



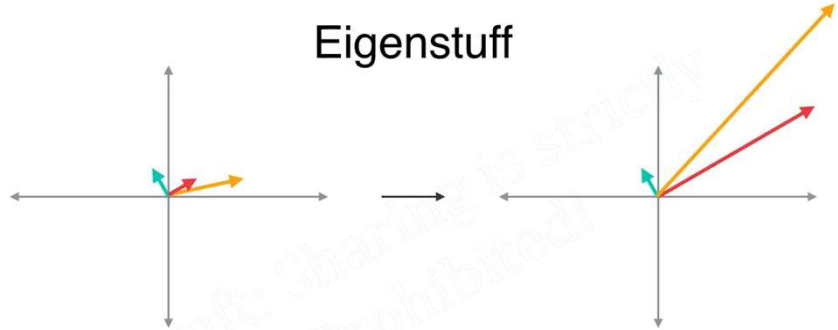
Draft: Sharing is strictly Prohibited!

Eigenstuff



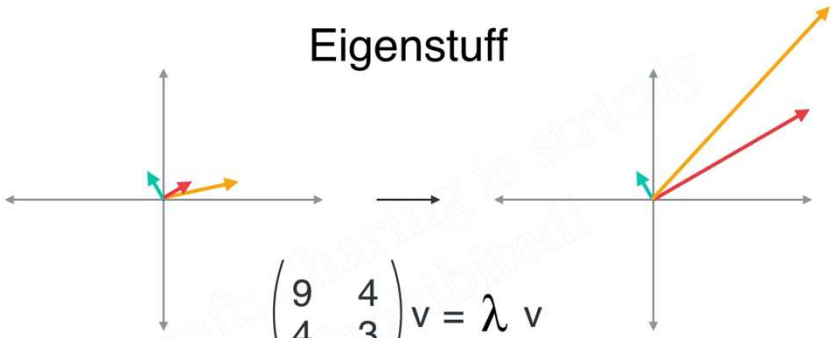
Draft: Sharing is strictly Prohibited!

Eigenstuff



Draft: Sharing is strictly Prohibited!

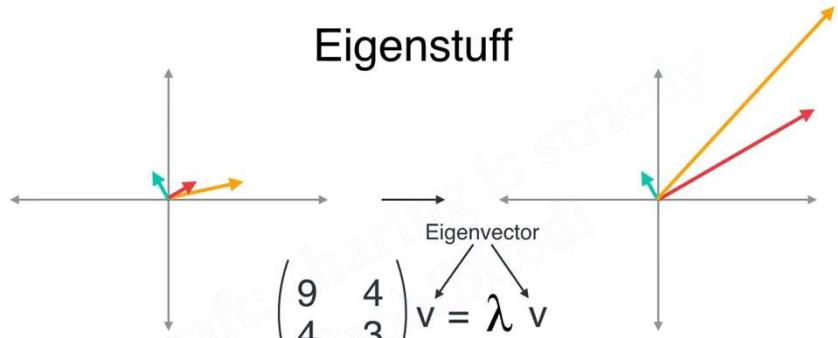
Eigenstuff



$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix} v = \lambda\, v$$

Draft: Sharing is strictly Prohibited!

Eigenstuff

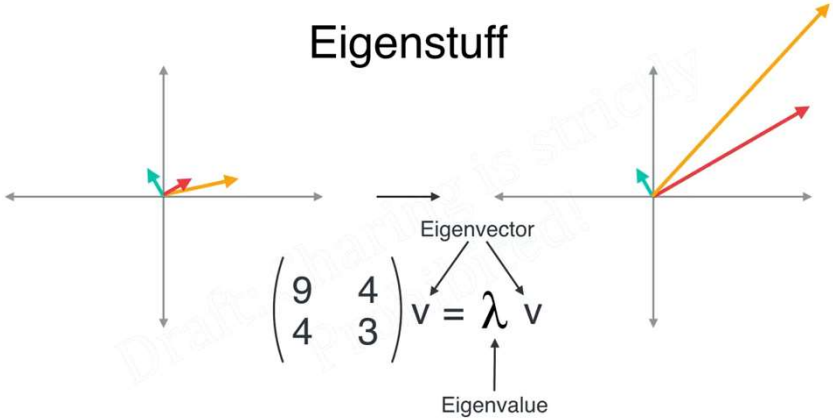


Eigenvector

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix} v = \lambda\, v$$

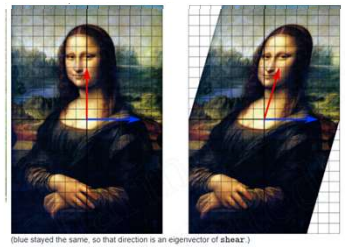
Draft: Sharing is strictly Prohibited!

Eigenstuff



Eigen Vector & Eigen Value

A simple example is that an eigenvector **does not change direction** in a transformation:



(blue stayed the same, so that direction is an eigenvector of **shear**.)

The Mathematics Of It

For a square matrix **A**, an Eigenvector and Eigenvalue make this equation true:

$$Av = \lambda v$$

Matrix *Eigenvector* *Eigenvalue*

We will see how to find them (if they can be found) soon, but first let us see one in action:

Example: For this matrix

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}$$

an eigenvector is:

$$\begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

with a matching eigenvalue of 6

Let's do some matrix multiplies to see what we get.

Av gives us:

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} -6\times1+3\times4 \\ 4\times1+5\times4 \end{bmatrix} = \begin{bmatrix} 6 \\ 24 \end{bmatrix}$$

λv gives us :

$$6\begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \\ 24 \end{bmatrix}$$

Yes they are equal! So $Av = \lambda v$ as promised.

Notice how we multiply a **matrix** by a **vector** and get the same result as when we multiply a **scalar** (just a number) by that **vector**.

How do we find these eigen things?

We start by finding the **eigenvalue**: we know this equation must be true:

$$Av = \lambda v$$

Now let us put in an identity matrix so we are dealing with matrix-vs-matrix:

$$Av = \lambda Iv$$

Bring all to left hand side:

$$Av - \lambda Iv = 0$$

If **v** is non-zero then we can solve for λ using just the determinant:

$$|A - \lambda I| = 0$$

Draft: sharing is strictly prohibited

Let's try that equation on our previous example:

Example: Solve for λ:

Start with $|A - \lambda I| = 0$

$$\left|\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right| = 0$$

Which is:

$$\begin{vmatrix} -6-\lambda & 3 \\ 4 & 5-\lambda \end{vmatrix} = 0$$

Calculating that determinant gets:

$$(-6-\lambda)(5-\lambda) - 3\times4 = 0$$

Which then gets us this Quadratic Equation:

$$\lambda^2 + \lambda - 42 = 0$$

And solving it gets:

$$\lambda = -7 \text{ or } 6$$

And yes, there are **two** possible eigenvalues.

Example (continued): Find the Eigenvector for the Eigenvalue λ = 6:

Start with:

$$Av = \lambda v$$

Put in the values we know:

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 6\begin{bmatrix} x \\ y \end{bmatrix}$$

After multiplying we get these two equations:

$$\begin{aligned} -6x + 3y &= 6x \\ 4x + 5y &= 6y \end{aligned}$$

Bringing all to left hand side:

$$\begin{aligned} -12x + 3y &= 0 \\ 4x - 1y &= 0 \end{aligned}$$

Either equation reveals that **y = 4x**, so the **eigenvector** is any non-zero multiple of this:

$$\begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

And we get the solution shown at the top of the page:

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} -6\times1+3\times4 \\ 4\times1+5\times4 \end{bmatrix} = \begin{bmatrix} 6 \\ 24 \end{bmatrix}$$

... and also ...

$$6\begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \\ 24 \end{bmatrix}$$

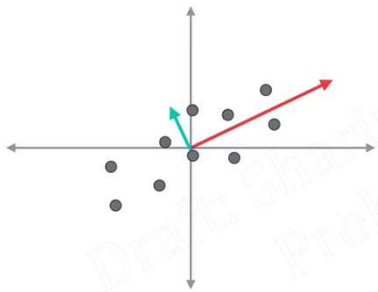
So $Av = \lambda v$

Eigenvectors

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 11\begin{pmatrix} u \\ v \end{pmatrix} \qquad \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 1\begin{pmatrix} u \\ v \end{pmatrix}$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

Principal Component Analysis (PCA)



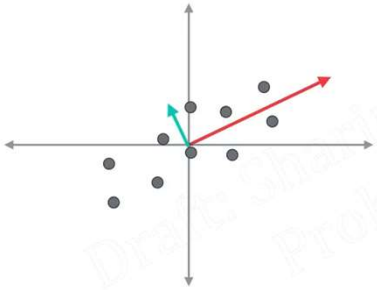
$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$ Eigenvectors (direction)

11 1 Eigenvalues (magnitude)

Draft: Sharing is strictly Prohibited

Principal Component Analysis (PCA)



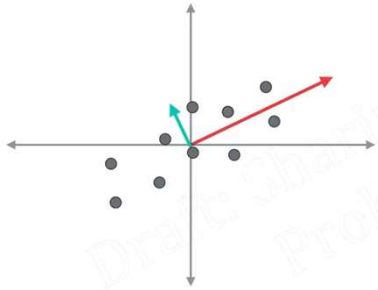
$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$ Eigenvectors (direction)

11 1 Eigenvalues (magnitude)

Draft: Sharing is Prohibited

Principal Component Analysis (PCA)



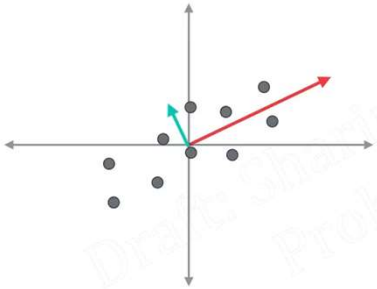
$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$ Eigenvectors (direction)

11 1 Eigenvalues (magnitude)

Draft: Sharing is Prohibited

Principal Component Analysis (PCA)



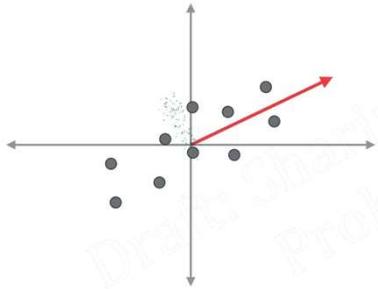
$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$ Eigenvectors (direction)

11 1 Eigenvalues (magnitude)

Draft: Sharing is Prohibited

Principal Component Analysis (PCA)



$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

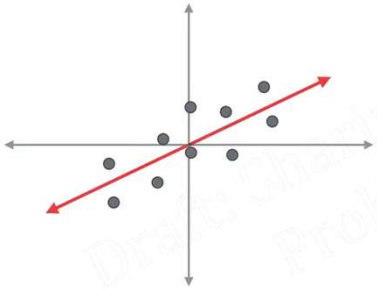
Eigenvectors (direction)

11

Eigenvalues (magnitude)

Draft: Sharing is Prohibited

Principal Component Analysis (PCA)



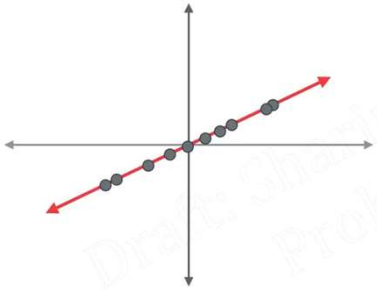
$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$ Eigenvectors (direction)

11 Eigenvalues (magnitude)

Draft: Sharing is strictly Prohibited!

Principal Component Analysis (PCA)



$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$ Eigenvectors (direction)

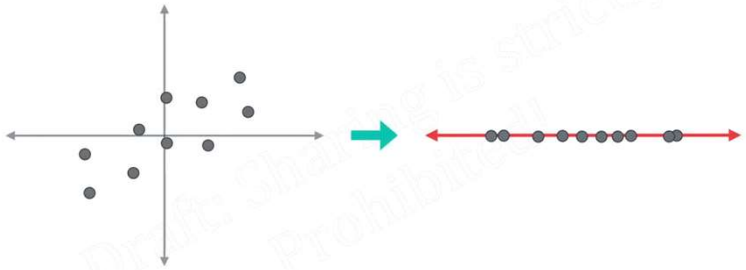
11 Eigenvalues (magnitude)

Draft: Sharing is strictly Prohibited!

Principal Component Analysis (PCA)

Draft: Sharing is strictly Prohibited!

Principal Component Analysis (PCA)



Draft: Sharing is strictly Prohibited!

PCA

Large Table

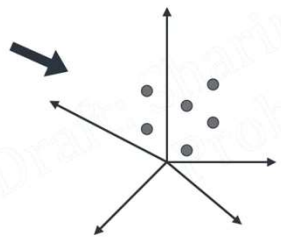
X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

Draft: Sharing is strictly Prohibited!

PCA

Large Table

X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*



5D Plot

Draft: Sharing is strictly Prohibited!

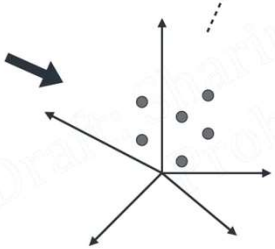
PCA

Covariance matrix

$$\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}$$

Large Table

X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*



5D Plot

Draft: Sharing is strictly Prohibited!

PCA

Covariance matrix

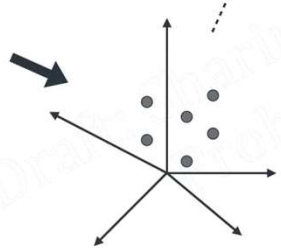
$$\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}$$

Eigenstuff

V_1	λ_1
V_2	λ_2
V_3	λ_3
V_4	λ_4
V_5	λ_5

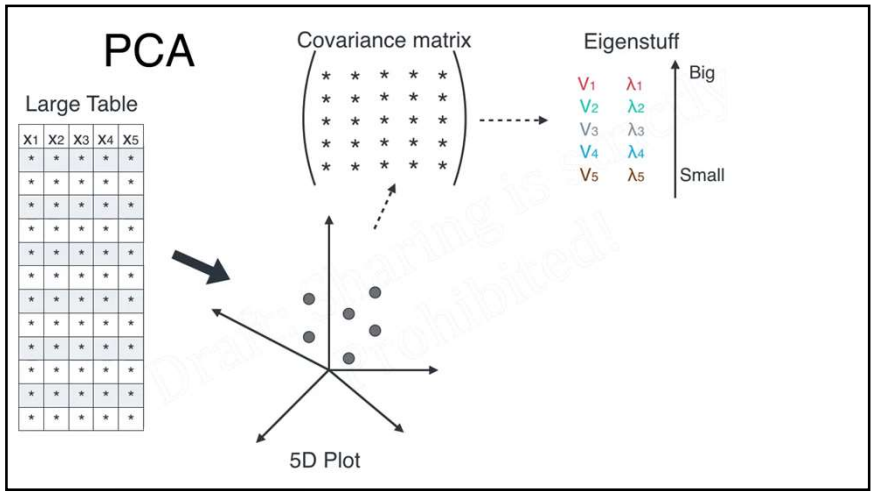
Large Table

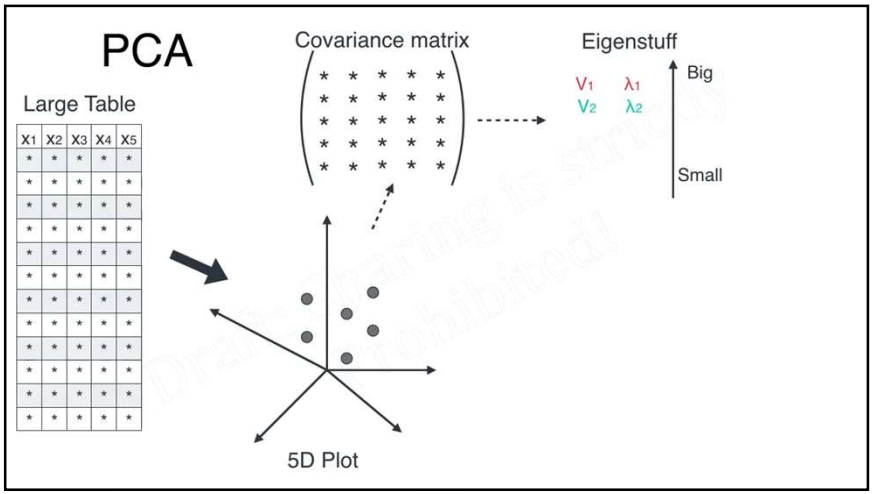
X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

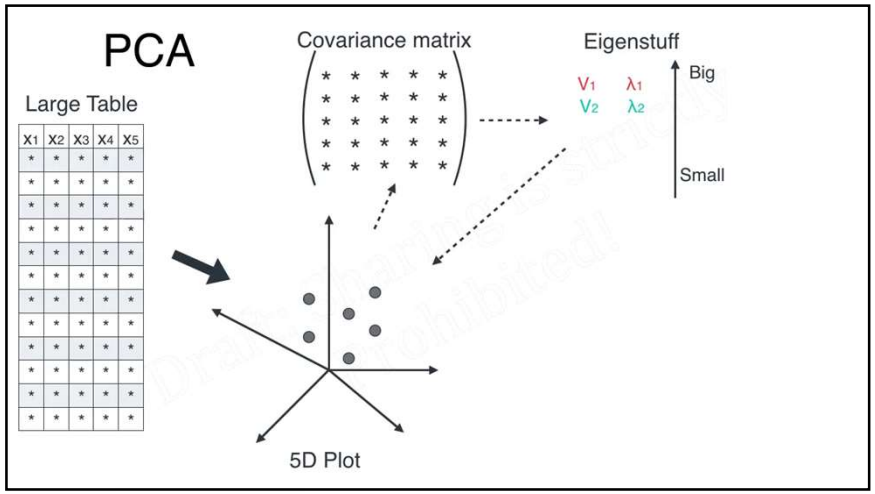


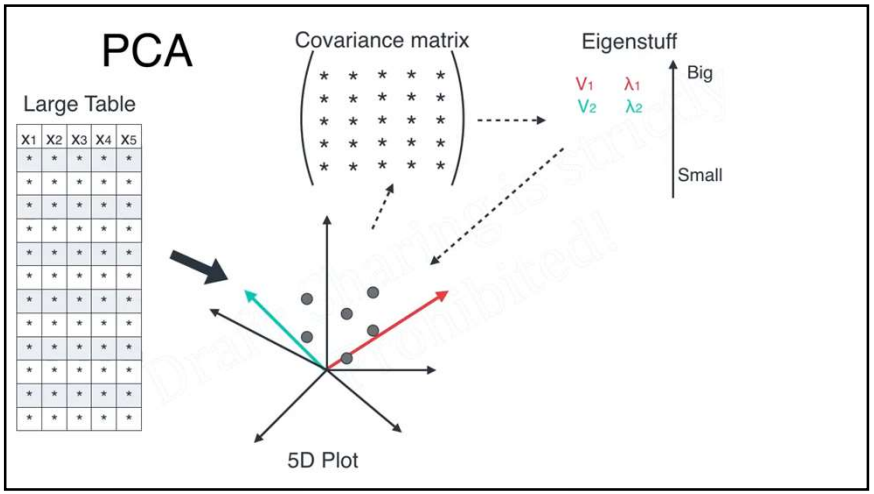
5D Plot

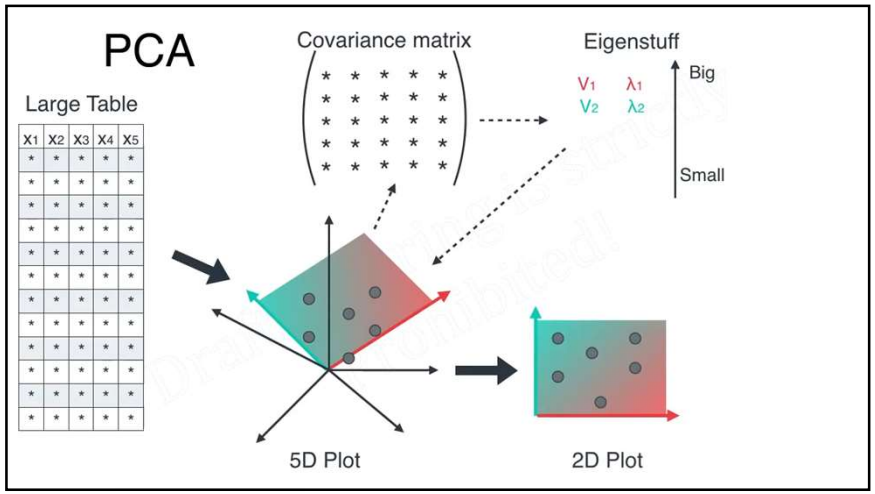
Draft: Sharing is strictly Prohibited!

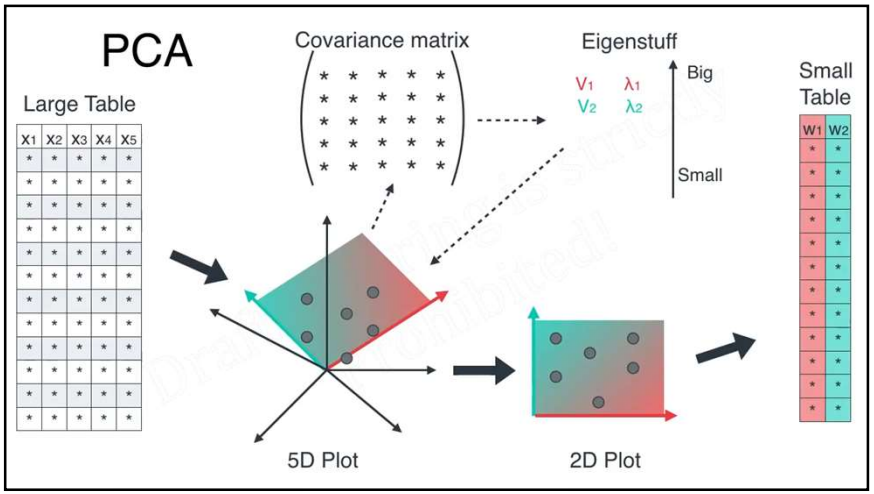

PCA
Large Table
X1 X2 X3 X4 X5
Covariance matrix
Eigenstuff
V1 λ1
V2 λ2
V3 λ3
V4 λ4
V5 λ5
Big
Small
5D Plot


PCA
Large Table
X1 X2 X3 X4 X5
Covariance matrix
Eigenstuff
V1 λ1
V2 λ2
Big
Small
5D Plot


PCA
Large Table
X1 X2 X3 X4 X5
Covariance matrix
Eigenstuff
V1 λ1
V2 λ2
Big
Small
5D Plot


PCA
Large Table
X1 X2 X3 X4 X5
Covariance matrix
Eigenstuff
V1 λ1
V2 λ2
Big
Small
5D Plot





THANK YOU

Youtube.com/c/Luis Serrano
https://www.mathsisfun.com/algebra/eigenvalue.html
StatQuest with Josh Starmer