Feature Selection

By,
Divesh R. Kubal
Data Scientist, eClerx Services
Center of Excellence – Machine Learning

Table of Contents

Introduction

- Statistical Model:
- How many features to have in a model?
- Prediction accuracy Vs. Model Interpretability

• Less number of features	More number of features
• Easy to interpret • Less likely to over fit • Low prediction accuracy	• Difficult to interpret • More likely to over fit • High prediction accuracy

Feature Selection

- Performance of a typical ML model depends on the following things:
  - Choice of Algorithm
  - Feature Selection
  - Feature Creation/Feature Engineering
  - Model Selection
- FS is also known as variable selection

Feature Selection Methods

- Feature Selection methods are basically of three types:
  - Filter Methods
  - Wrapper Methods
  - Embedded Methods

Filter methods

- Also called as **Single Factor Analysis**.
- Calculate/measure the predictive power of each individual variable by using,
  - Correlation with target variable.
  - Chi-Square Test (Categorical Variable).

Correlation

- **Correlation**: The degree of relationship between the variables under consideration is measure through the correlation analysis.
- The measure of correlation called the correlation coefficient
- The degree of relationship is expressed by coefficient which range from correlation **(-1 $\leq$ r $\geq$ +1)**
- The direction of change is indicated by a sign.
- The correlation analysis enable us to have an idea about the degree & direction of the relationship between the two variables under study.

Correlation Example

Number of Study Hours	2	4	6	8	10
Number of Sleeping Hours	10	9	8	7	6

Correlation Example

X	Y	$(X-\overline{X})$	$(Y-\overline{Y})$	$(X-\overline{X})(Y-\overline{Y})$	$(X-\overline{X})^2$	$(Y-\overline{Y})^2$
2	10	-4	+2	-8	16	4
4	9	-2	+1	-2	4	1
6	8	0	0	0	0	0
8	7	+2	-1	-2	4	1
10	6	+4	-2	-8	16	1
$\sum X = 30$	$\sum Y = 40$	$\sum(X-\overline{X}) = 0$	$\sum(Y-\overline{Y}) = 0$	$\sum(X-\overline{X})(Y-\overline{Y}) = -20$	$\sum(X-\overline{X})^2 = 40$	$\sum(Y-\overline{Y})^2 = 10$

$\overline{X} = \frac{\sum X}{n} = \frac{30}{5} = 6$ and $\overline{Y} = \frac{\sum Y}{n} = \frac{40}{5} = 8$

$$r_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2 \sum(Y-\overline{Y})^2}} = \frac{-20}{20} = -1$$

Chi-Square Test – Observed Frequencies

- A good first step for these data is inspecting the contingency table of marital status by education
- The table displays the frequency distribution of marital status for each education category separately.

Marital Status by Education | n = 300

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

Chi-Square Test

- The numbers in this table are known as the **observed frequencies**. They tell us an awful lot about our data
  - there's 4 marital status categories and 5 education levels;
  - we succeeded in collecting data on our entire sample of n = 300 respondents (bottom right cell);
  - we've 84 respondents with a Bachelor’s degree (bottom row, middle);
  - we've 30 divorced respondents (last column, middle);
  - we've 9 divorced respondents with a Bachelor’s degree.

Chi-Square Test – Column Percentages

- Although our contingency table is a great starting point, it doesn't really show us if education level and marital status are related. This question is answered more easily from a slightly different table as shown below.

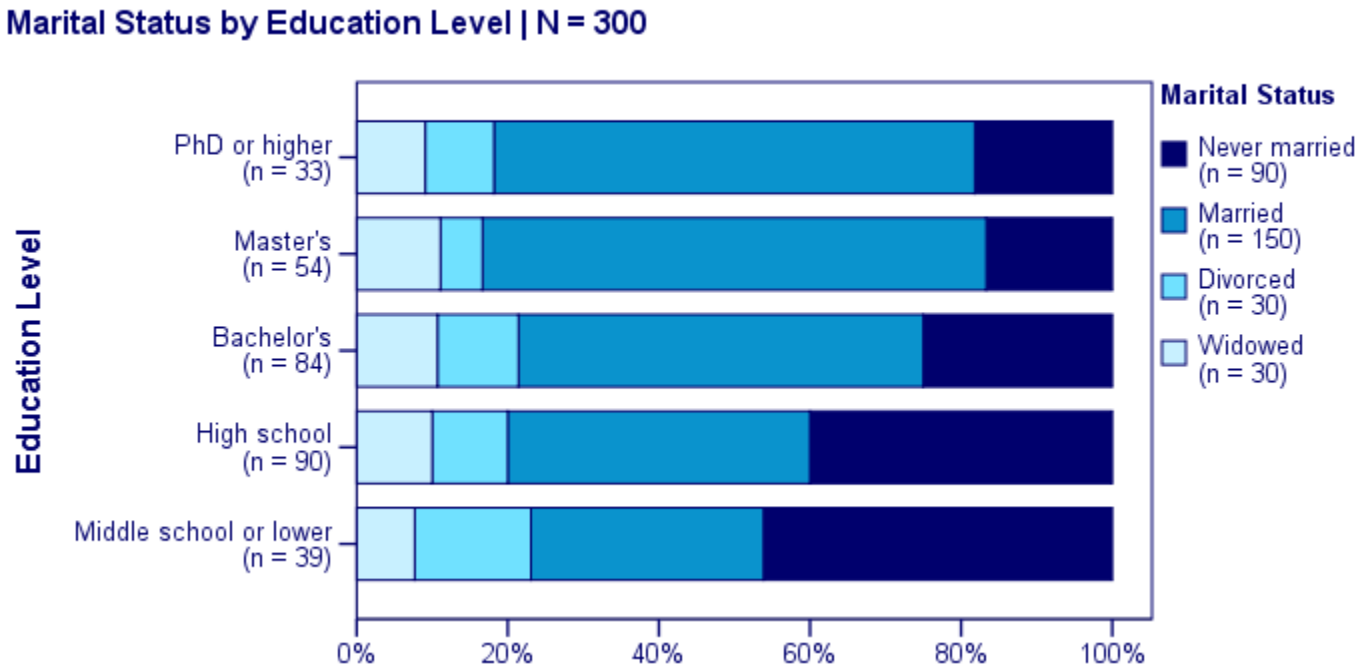
Marital Status by Education | n = 300

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	46%	40%	25%	17%	18%	30%
Married	31%	40%	54%	67%	64%	50%
Divorced	15%	10%	11%	6%	9%	10%
Widowed	8%	10%	11%	11%	9%	10%
Total	100%	100%	100%	100%	100%	100%

is marital status related to education level and - if so- how?

- If we inspect the first row, we see that 46% of respondents with middle school never married. If we move rightwards (towards higher education levels), we see this percentage decrease: only 18% of respondents with a PhD degree never married (top right cell).
  Reversely, note that 64% of PhD respondents are married (second row). If we move towards the lower education levels (leftwards), we see this percentage decrease to 31% for respondents having just middle school.
  In short, more **highly educated respondents seem to marry more** often than less educated respondents

Chi-Square Test – Stacked Bar Chart



Chi-Square Test – Stacked Bar Chart

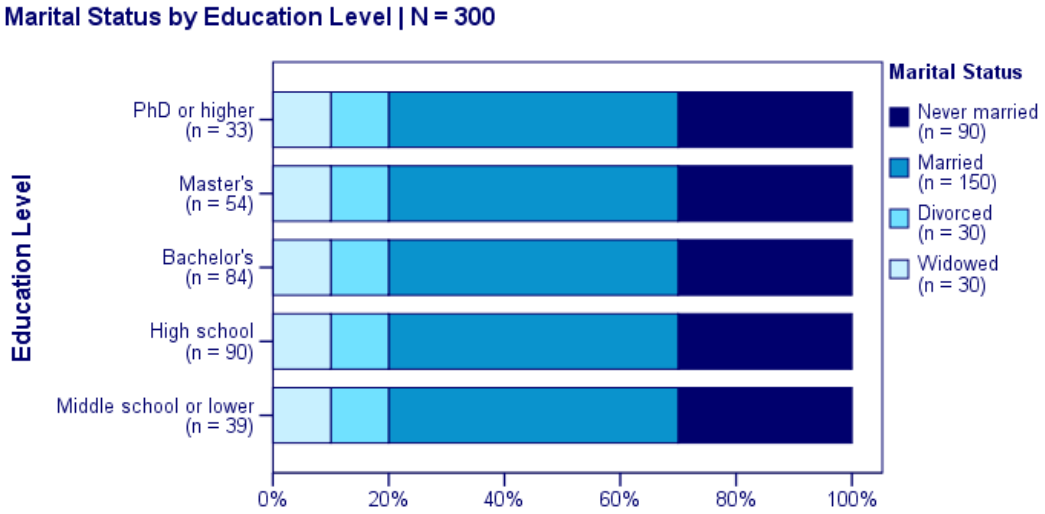
- If we move from top to bottom (highest to lowest education) in this chart, we see the dark blue bar (never married) increase. **Marital status is clearly associated** with education level: the lower someone's education, the smaller the chance he's married. That is: education "says something" about marital status (and reversely) in our sample. So what about the population?

Chi-Square Test – Null Hypothesis

- **two categorical variables are independent in some population.**
- Now, marital status and education are related - thus *not* independent- in our sample. However, we can't conclude that this holds for our entire population. The basic problem is that **sample outcomes usually differ from populations**
- So exactly **how strong is the relation** in our sample? And what's the probability -or significance level- of finding it if the variables are (perfectly) independent in the entire population?

Chi-Square Test – Statistical Independence

- **Statistical independence means that the frequency distribution of a variable is the same for all levels of some other variable.**



Chi-Square Test – Statistical Independence

- What does education “say about” marital status? Absolutely nothing! Why? Because the frequency distributions of marital status are identical over education levels: no matter the education level, the probability of being married is 50% and the probability of never being married is 30%. In this chart, education and marital status are **perfectly independent**. The hypothesis of independence tells us which frequencies we should have found in our sample: the expected frequencies.

Expected Frequencies

- **Expected frequencies are the frequencies we expect in our sample if the null hypothesis holds.**

Expected Frequencies for Perfectly Independent Variables

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	11.7	27.0	25.2	16.2	9.9	90.0
Married	19.5	45.0	42.0	27.0	16.5	150.0
Divorced	3.9	9.0	8.4	5.4	3.3	30.0
Widowed	3.9	9.0	8.4	5.4	3.3	30.0
Total	39.0	90.0	84.0	54.0	33.0	300.0

These expected frequencies are calculated as

$$eij = \frac{oi.\ o.j}{N}$$

where

- eij is an expected frequency;
- $oi.$ is a marginal column frequency;
- $o.j$ is a marginal row frequency;
- N is the total sample size.

Test Statistic

The chi-square test statistic is calculated as

$$\chi^2 = \Sigma \frac{(oij - eij)^2}{eij}$$

=SUM(B24:F24)

	A	B	C	D	E	F
1	**Observed Frequencies**					
2		Middle school or lower	High school	Bachelor's	Master's	PhD or
3	Never married	18	36	21	9	
4	Married	12	36	45	36	
5	Divorced	6	9	9	3	
6	Widowed	3	9	9	6	
7		39	90	84	54	
8						
9	**Expected Frequencies (Variables Perfectly Independent)**					
10		Middle school or lower	High school	Bachelor's	Master's	PhD or higher
11	Never married	11.7	27	25.2	16.2	
12	Married	19.5	45	42	27	
13	Divorced	3.9	9	8.4	5.4	
14	Widowed	3.9	9	8.4	5.4	
15		39	90	84	54	
16						
17						
18	**Chi-Square Points = (Observed - Expected) ^2 / Expected**					
19		Middle school or lower	High school	Bachelor's	Master's	PhD or higher
20	Never married	3.39	3.00	0.70	3.20	
21	Married	2.88	1.80	0.21	3.00	
22	Divorced	1.13	0.00	0.04	1.07	
23	Widowed	0.21	0.00	0.04	0.07	
24		7.62	4.80	1.00	7.33	
25						
26	**Chi-Square Statistic = Sum Chi-Square Points**					
27	23.57					
28						

- So $\chi^2 = 23.57$ in our sample. This number summarizes the difference between our data and our independence hypothesis. Is 23.57 a large value? What's the probability of finding this? Well, we can calculate it from its sampling distribution.

Chi-Square Test – Degrees of Freedom

- We'll get the significance level we're after from the chi-square distribution if we give it 2 numbers:
  - the χ2 value (23.57) and
  - the degrees of freedom (df).
  - The degrees of freedom is basically a number that determines the exact shape of our distribution. It's calculated as
  - df=(i−1)·(j−1)
- Where,
  - i is the number of rows in our contingency table and
  - j is the number of columns
  - so in our example
  - df=(5−1)·(4−1)=12.
- And with df = 12, the probability of finding $\chi^2 \geq 23.57 \approx 0.023$.* This is our 1-tailed significance. It basically means, there's a 0.023 (or 2.3%) chance of finding this assocation in our sample if it is zero in our population
- **"An association between education and marital status was observed, $\chi^2(12) = 23.57$, $p = 0.023$."**

Wrapper Methods

- Uses combination of predictors/features and finds out the best combination.
- Instead of finding best feature, it finds out the best feature combination.
- Predictive power of the variables is evaluated jointly.
- Set of variables that perform the best.
- Techniques:
  - Forward Selection
  - Backward Selection
  - Recursive Feature Elimination

Optimal Number of Features?

- An ideal model should do justice to both: **good prediction yet not overly complex to interpret and use.**
- One way to is to select the best set of features.
  - Subset Selection
  - Shrinkage
  - Dimension Reduction

Forward Selection

- Start with a null model.
- Add predictors to the model **one at a time**. Choose the best model among the results for each k-based on RSS.
- If a variable is retained, it never drops from the model.

	Subset Selection	Forward Stepwise
One variable	X1	**X1**
Two Variables	X1 X3	**X1** X2
Three Variables	X1 X3 X4	**X1 X2** X4
Four Variables	X1 X2 X3 X5	**X1 X2 X4** X5

Forward Selection

- Selection is constrained as a variable **that is in the model never drops.**
- So less candidate models for selection: 1+p(p+1)/2
- P=10
- Subset selection: Over a million models.
- Stepwise: 211 models.

Backward Selection

- It is the reverse of forward: Start with all the predictors and then drop one at a time. Finally select the best model.

	Backward Stepwise	Forward Stepwise
	X1 **X2** X3 X4 X5	**X1**
	X1 X3 **X4** X5	**X1** X2
	X1 **X3** X5	**X1 X2** X4
	X1 **X 5**	**X1 X2 X4** X5
	X1	**X1 X2 X4 X3 X5**

Backward Selection

- Computational power requirement is similar to that of Forward.
- Selection is made through RSS or Deviance.

Embedded Method

- Inbuilt variable selection methods (without one having to select/reject feature)
- Regularization – Controls the value of the parameter. Not so important variables are given very low weight (close to zero).
- Techniques:
  - Lasso and Ridge Regression
  - Also known as **Shrinkage Method.**

Embedded Methods (Shrinkage)

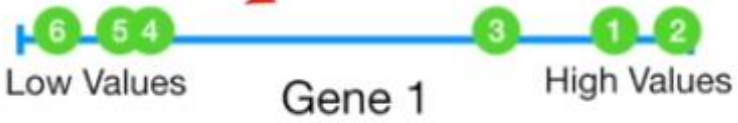
- Regularized regresion models – A technique that regularize the estimates or shrink the coefficients towards zero.
- Slight modification to least square estimation.

Principal Component Analysis

	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

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



	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…

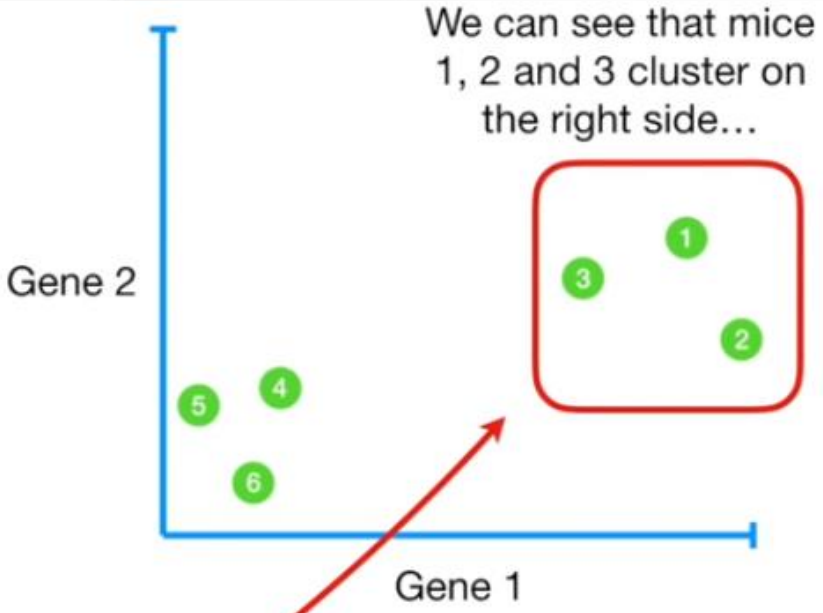
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

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

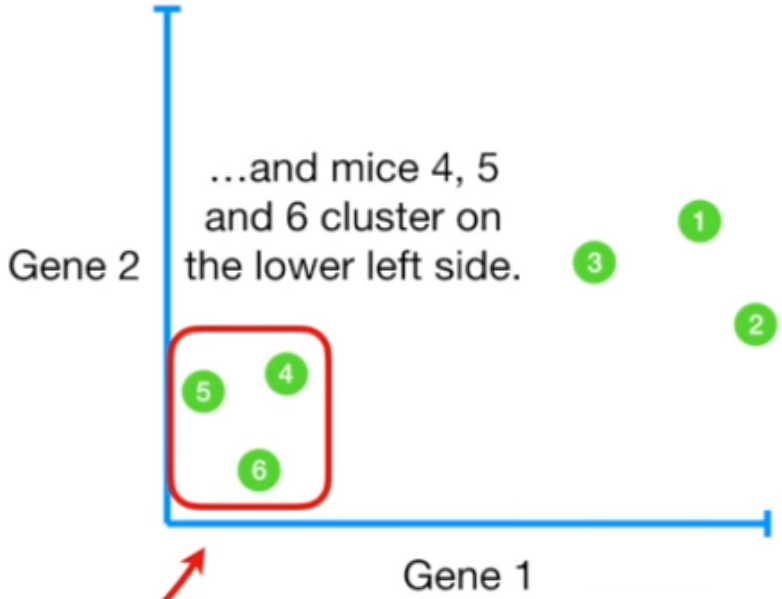
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

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.

	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

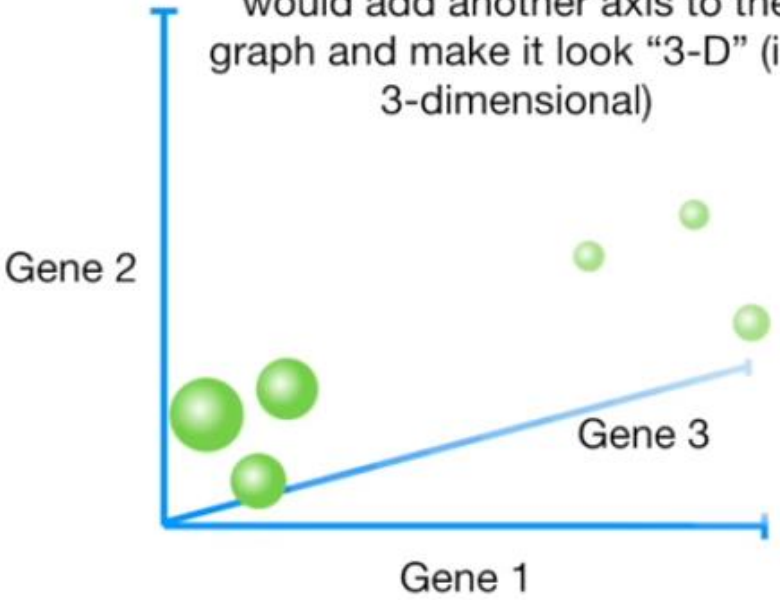


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



	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

If we measured 3 genes, we would add another axis to the graph and make it look "3-D" (i.e 3-dimensional)

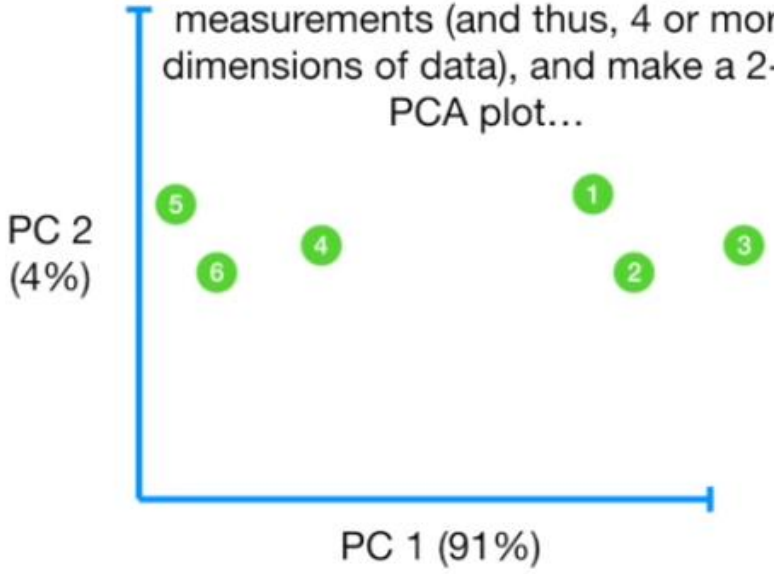


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

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

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

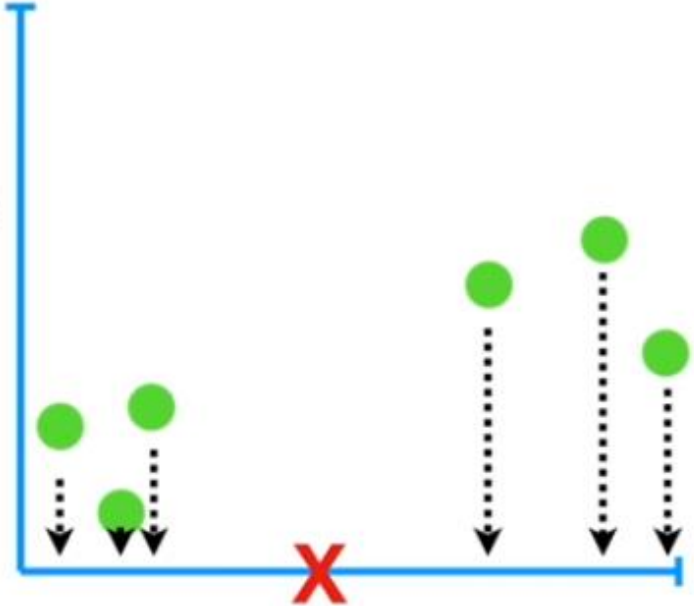


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

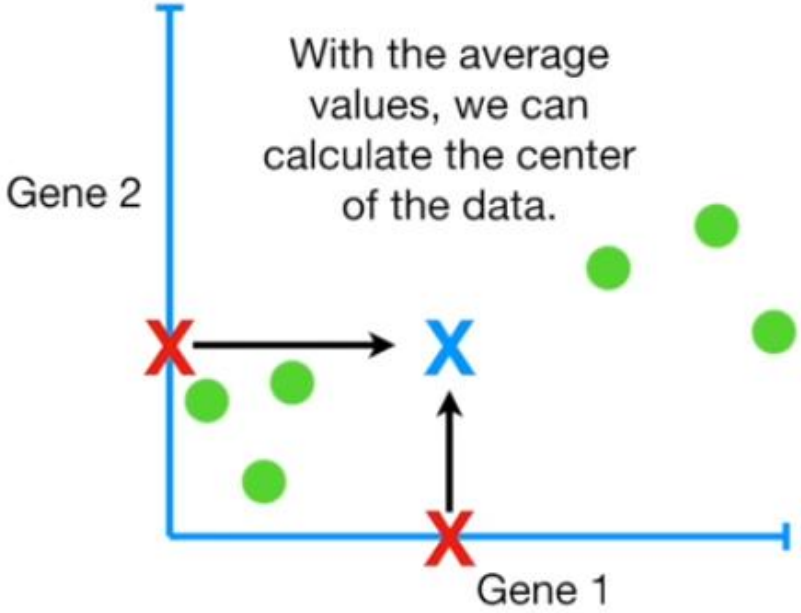
To understand what PCA does and how it works, let's go back to the dataset that only had 2 genes…

	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

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

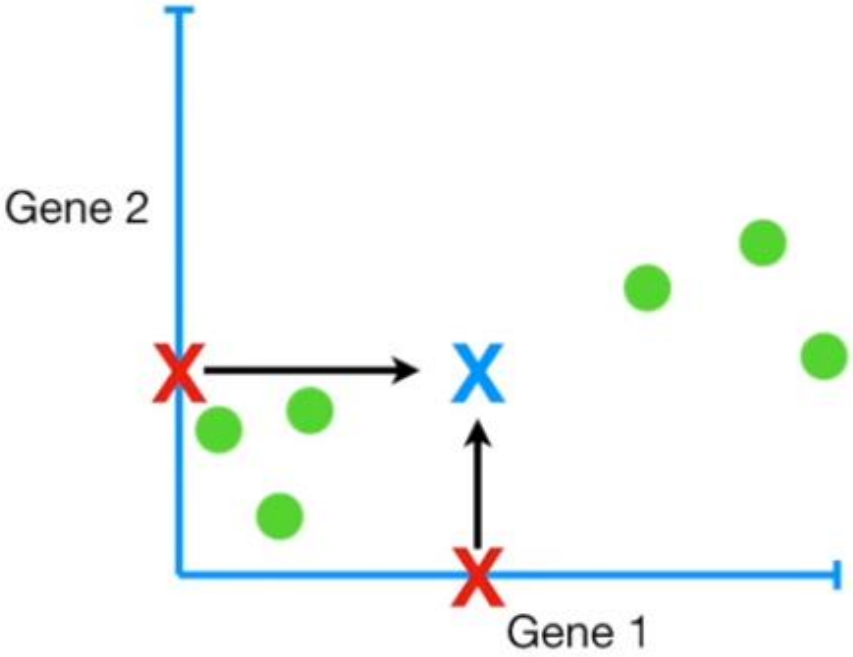


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

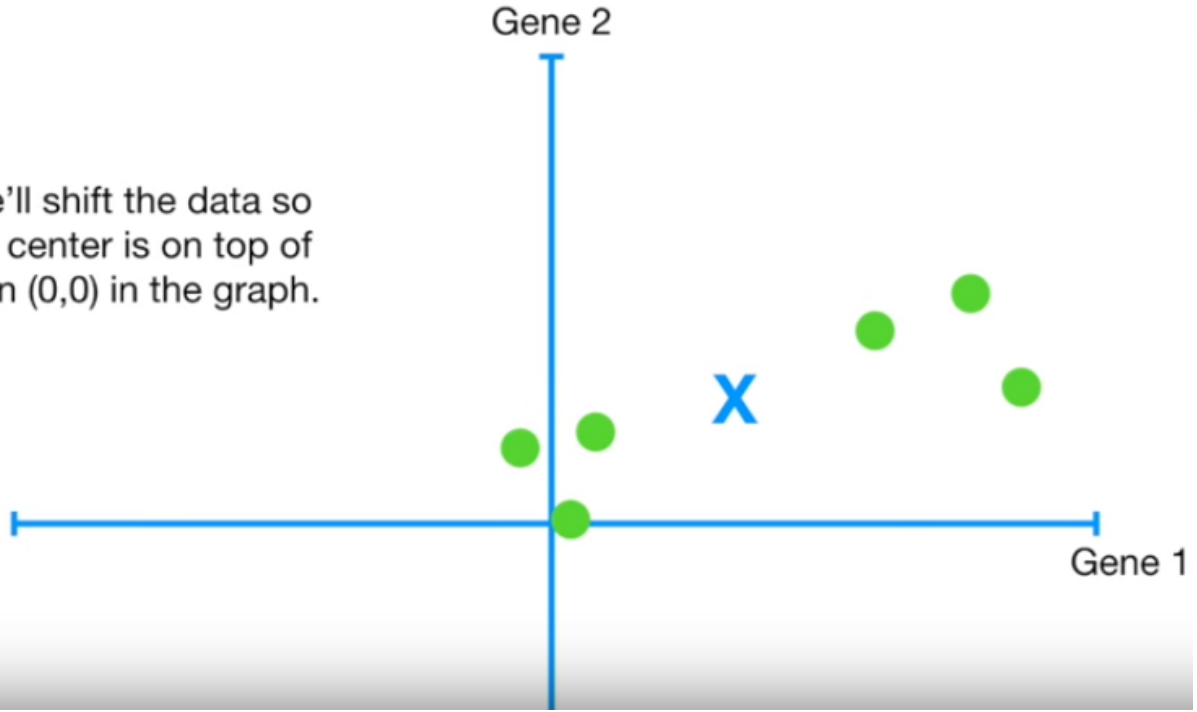


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

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



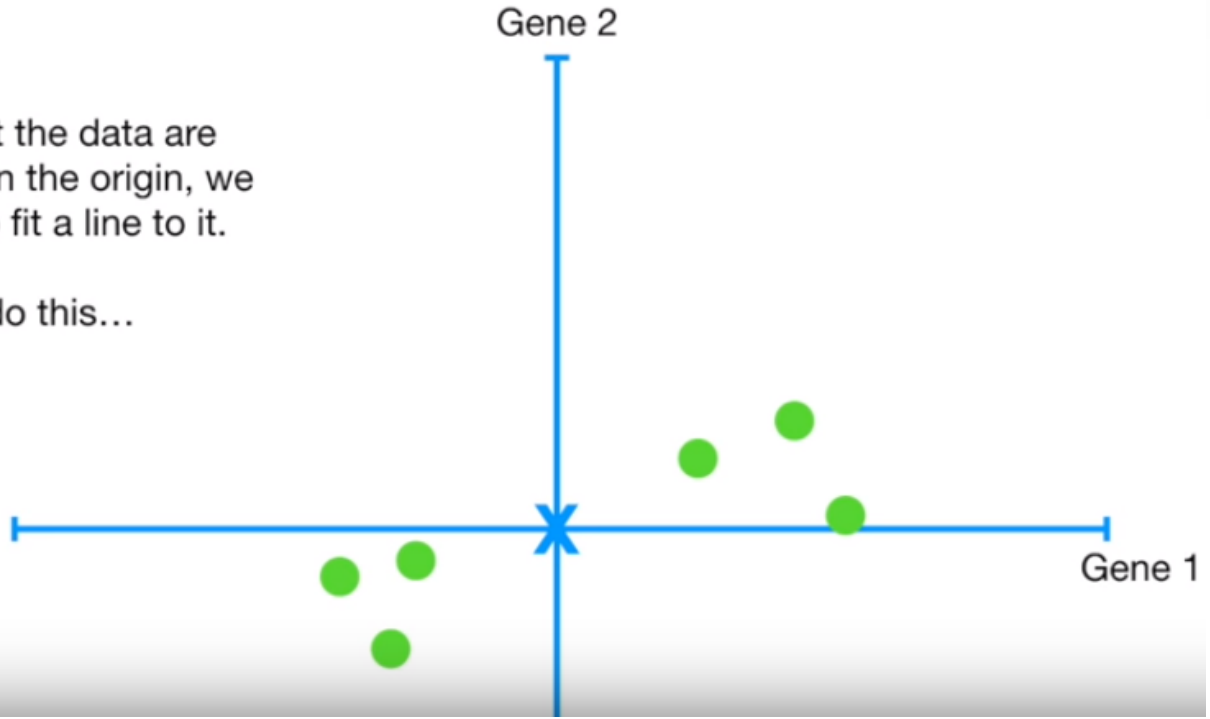
Now we'll shift the data so that the center is on top of the origin (0,0) in the graph.

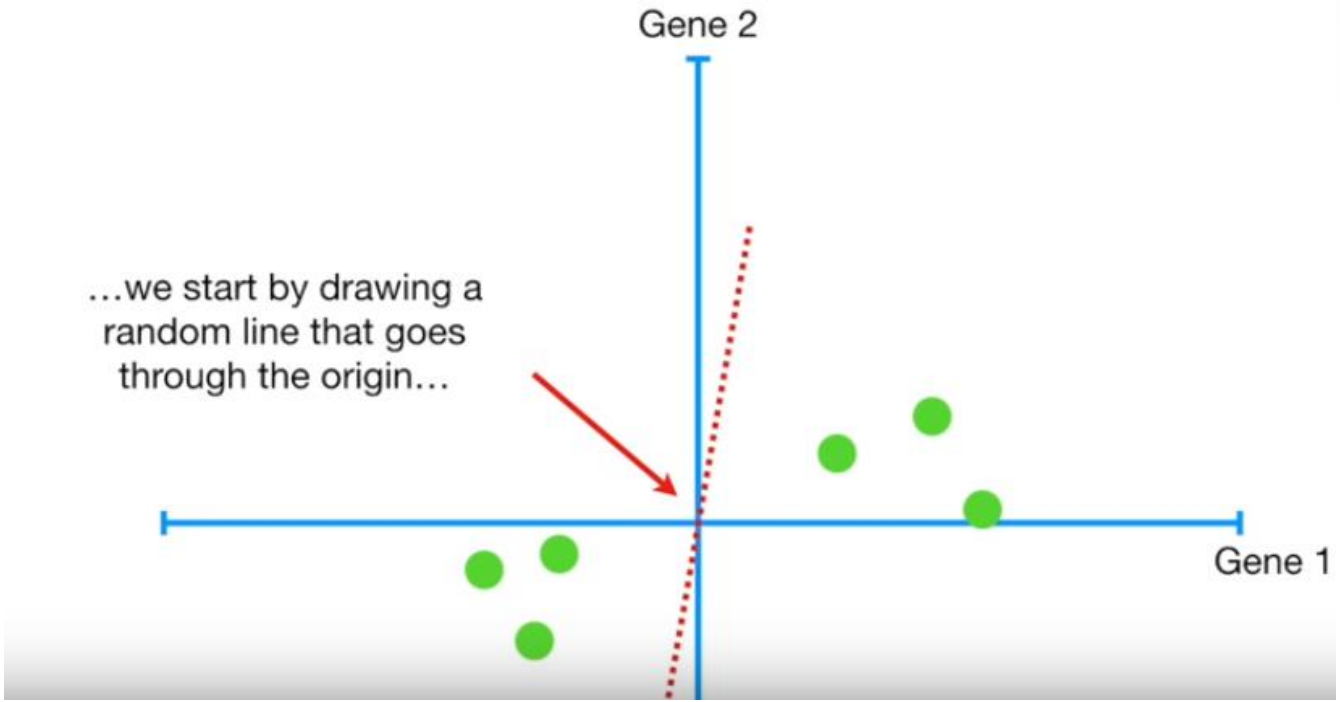


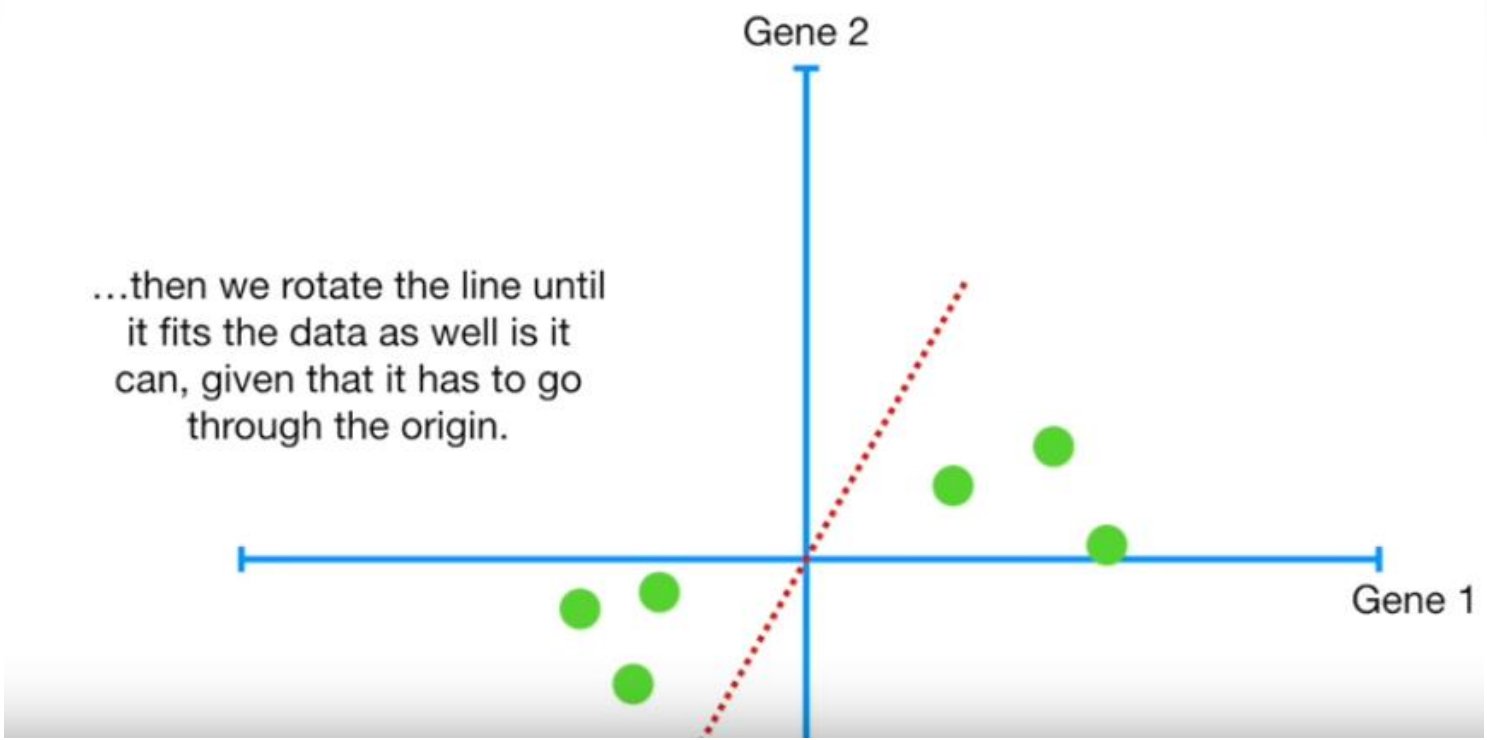
NOTE: Shifting the data did not change how the data points are positioned *relative to each other*.

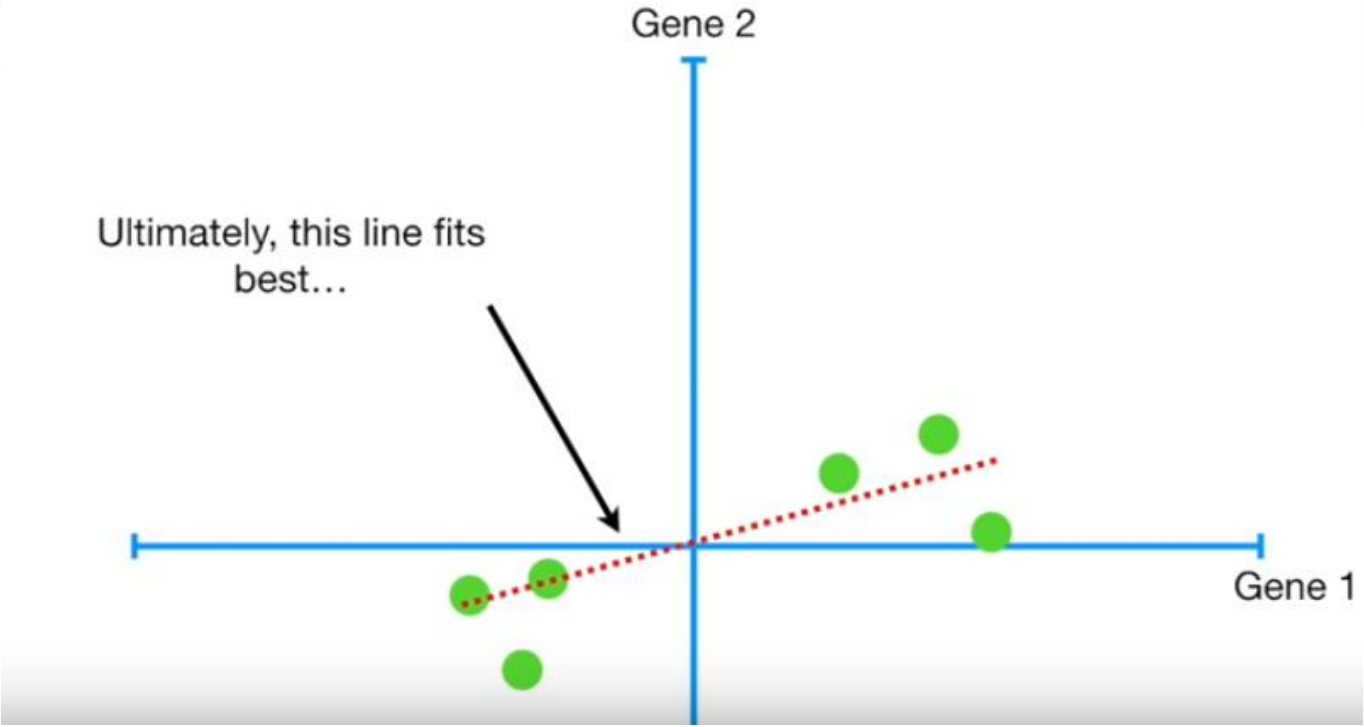
Now that the data are centered on the origin, we can try to fit a line to it.

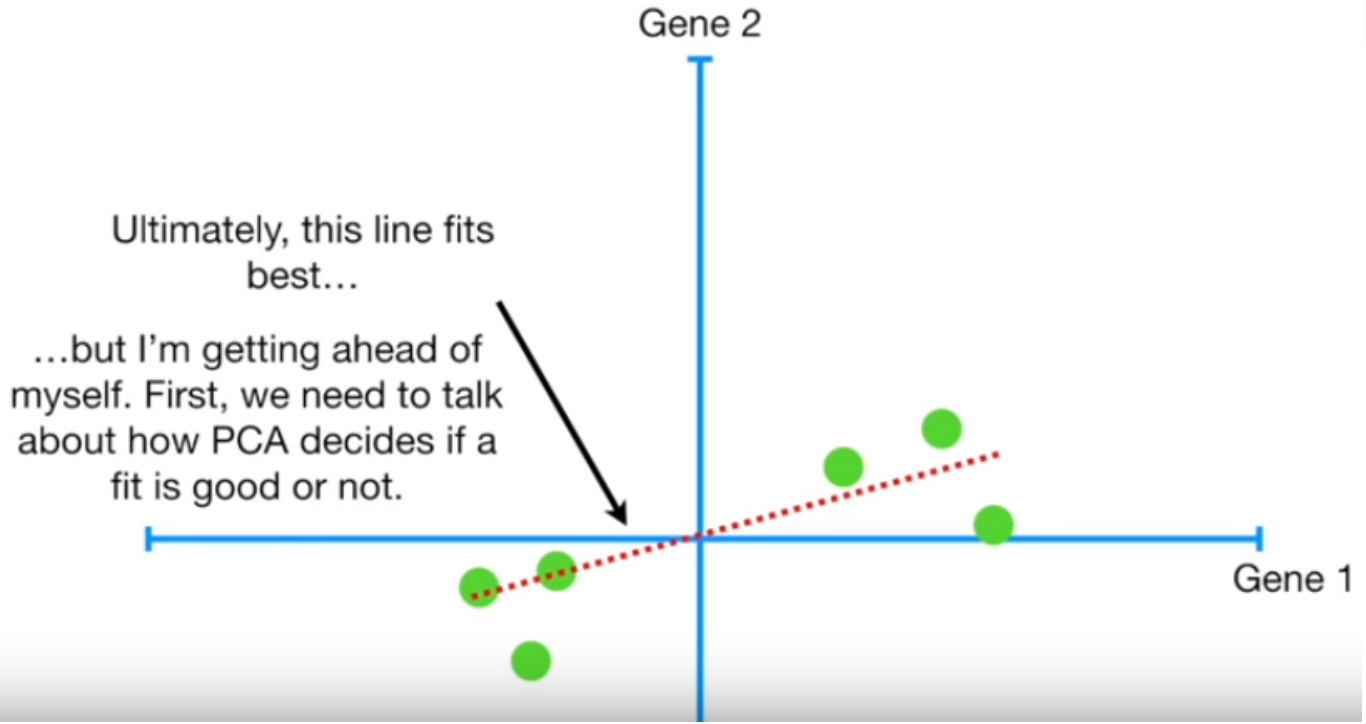
To do this…

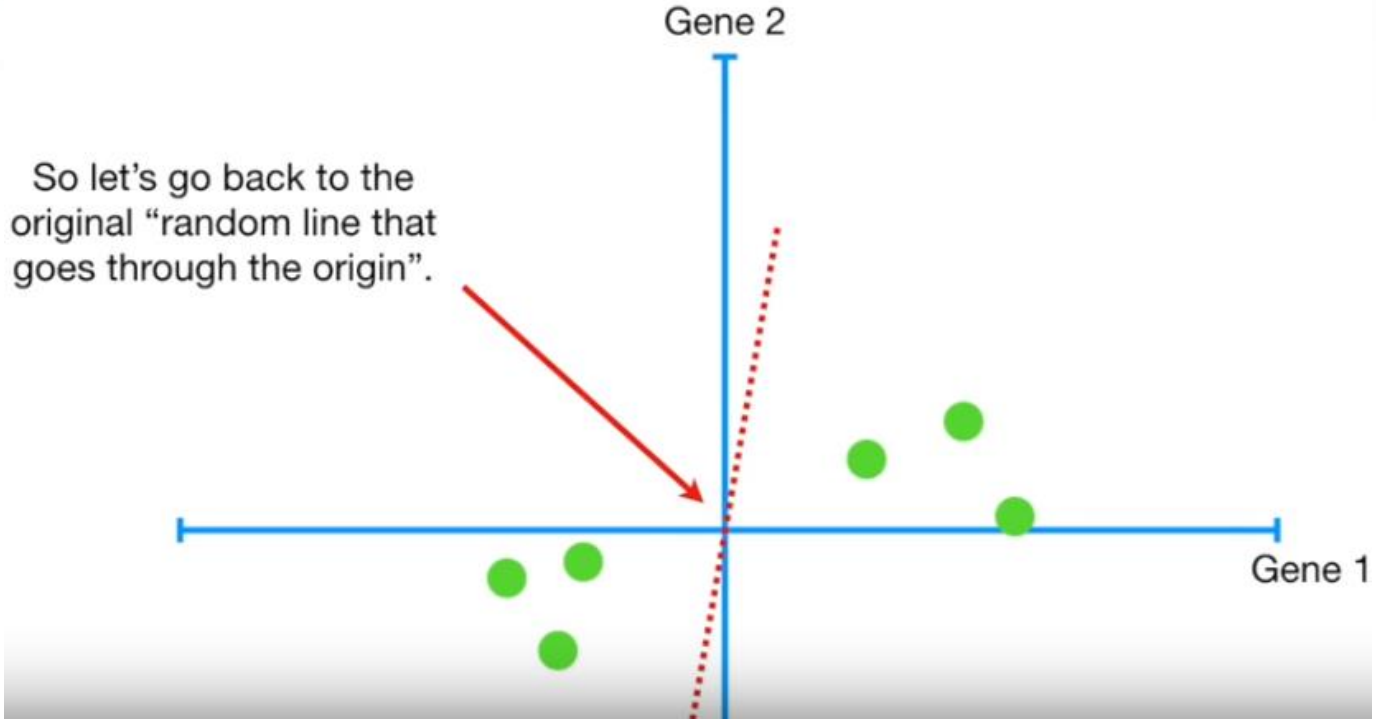


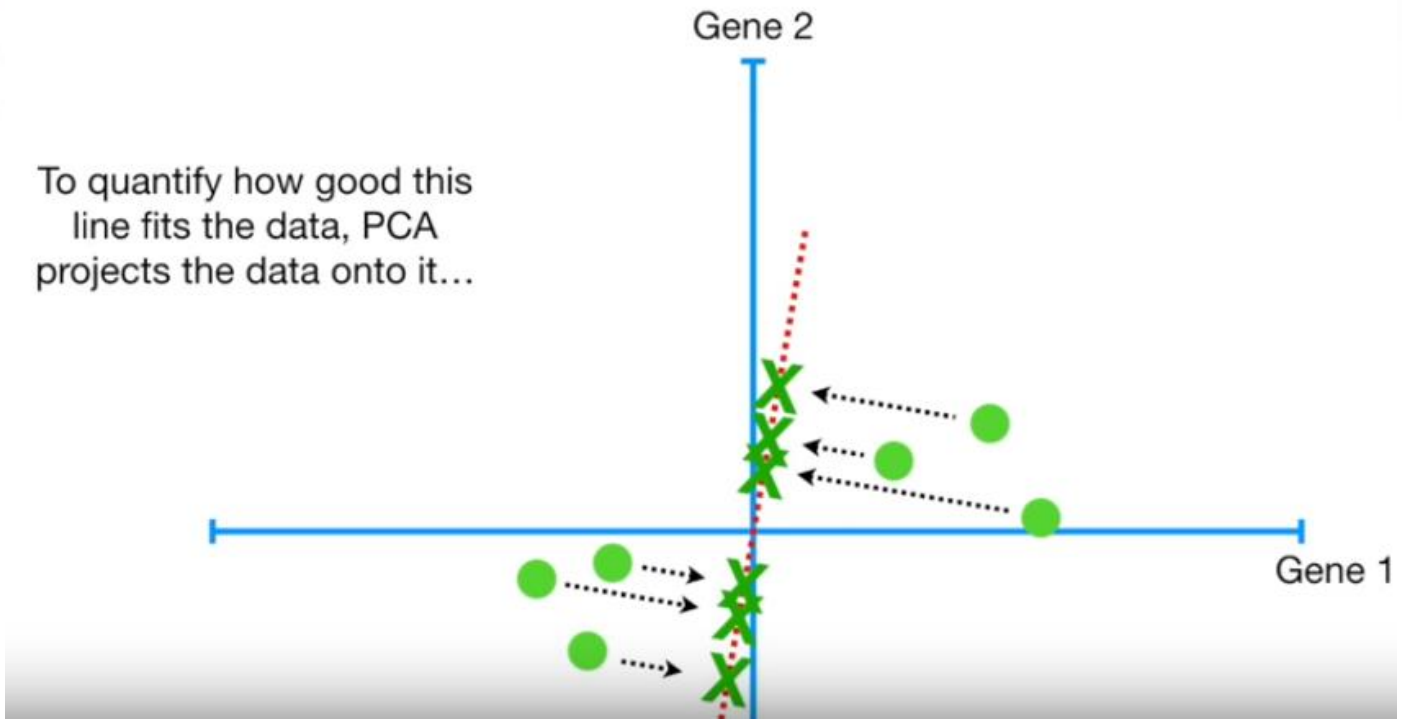

Gene 2
…we start by drawing a random line that goes through the origin…
Gene 1

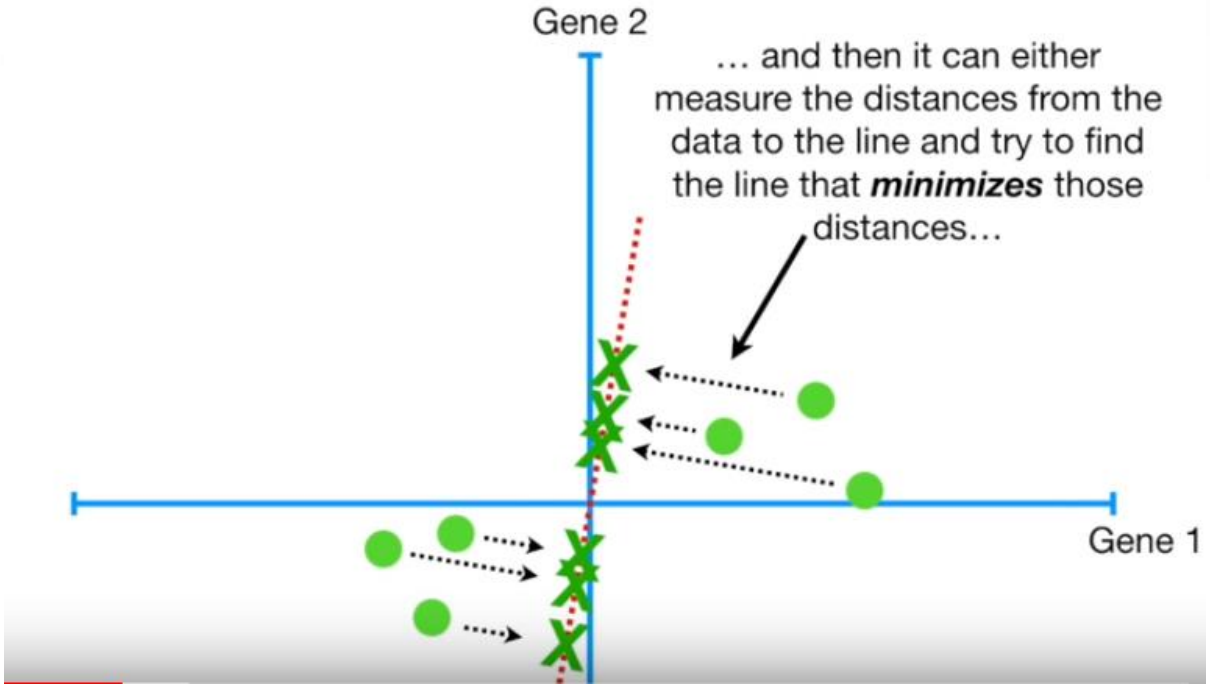

Gene 2
…then we rotate the line until it fits the data as well is it can, given that it has to go through the origin.
Gene 1

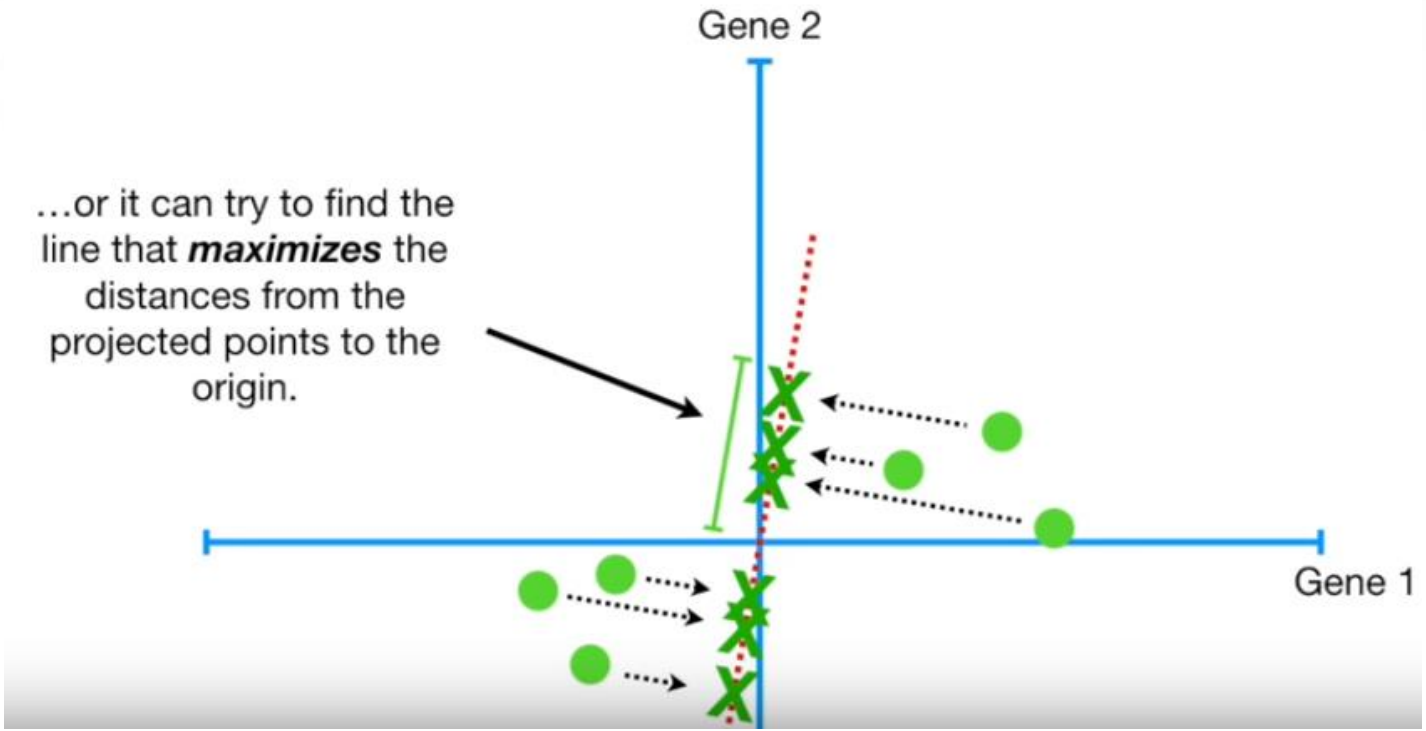

Gene 2
Ultimately, this line fits best…
Gene 1

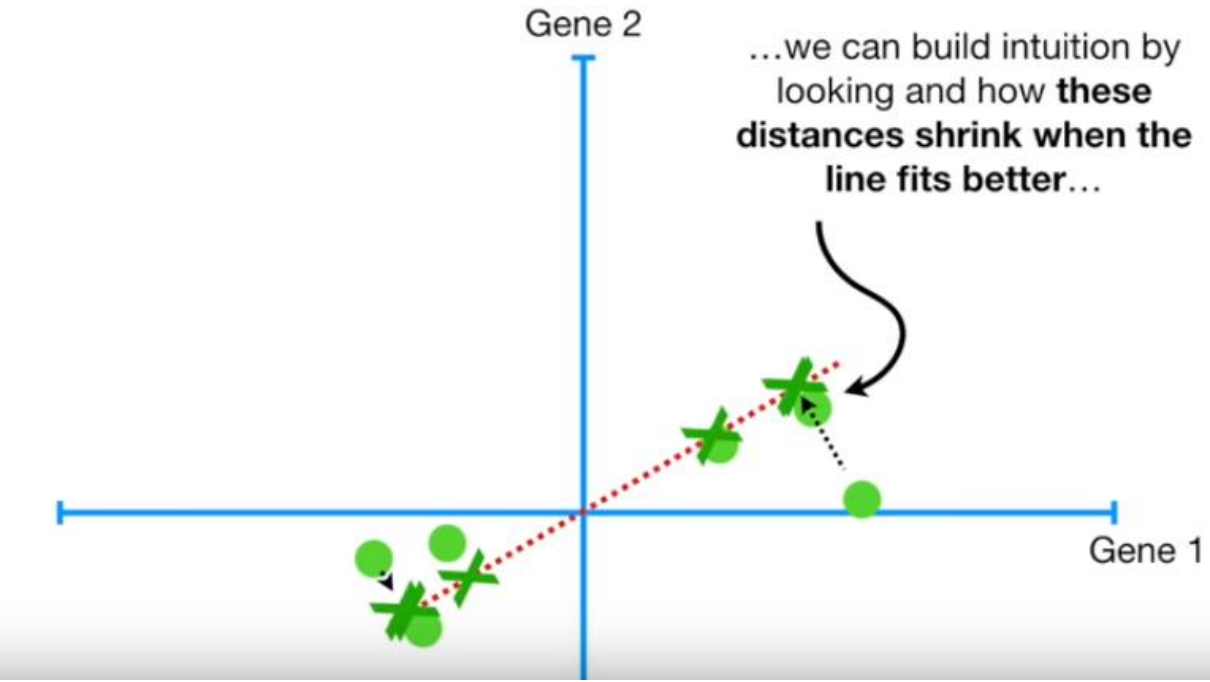

Gene 2
Ultimately, this line fits best…
…but I'm getting ahead of myself. First, we need to talk about how PCA decides if a fit is good or not.
Gene 1

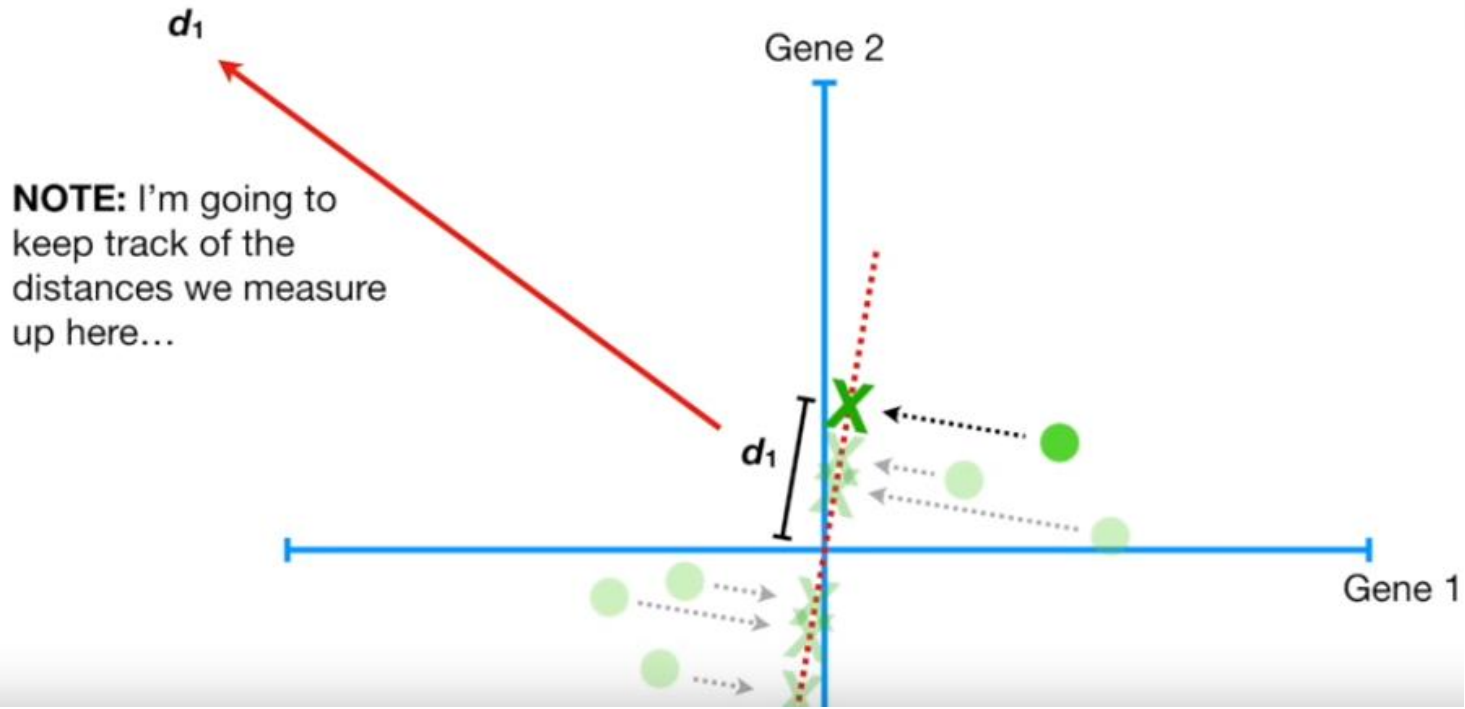

Gene 2
So let's go back to the original "random line that goes through the origin".
Gene 1


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

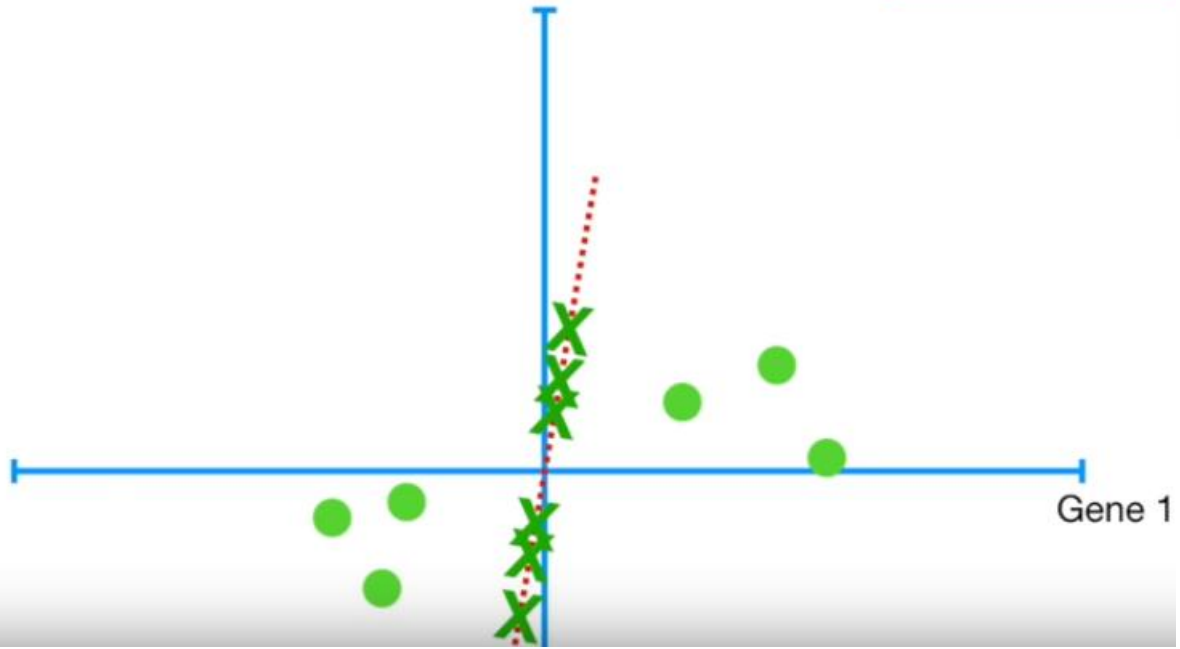

Gene 2
… and then it can either measure the distances from the data to the line and try to find the line that ***minimizes*** those distances…
Gene 1

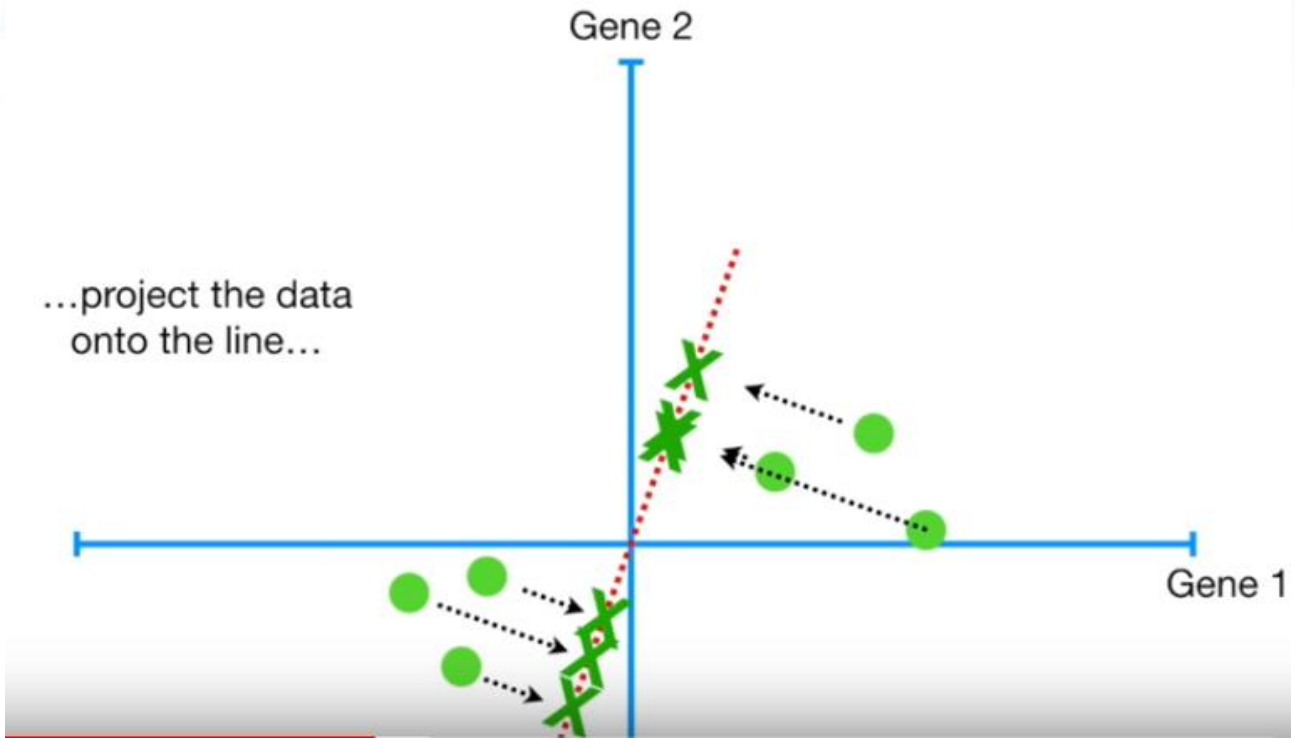

Gene 2
…or it can try to find the line that ***maximizes*** the distances from the projected points to the origin.
Gene 1


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

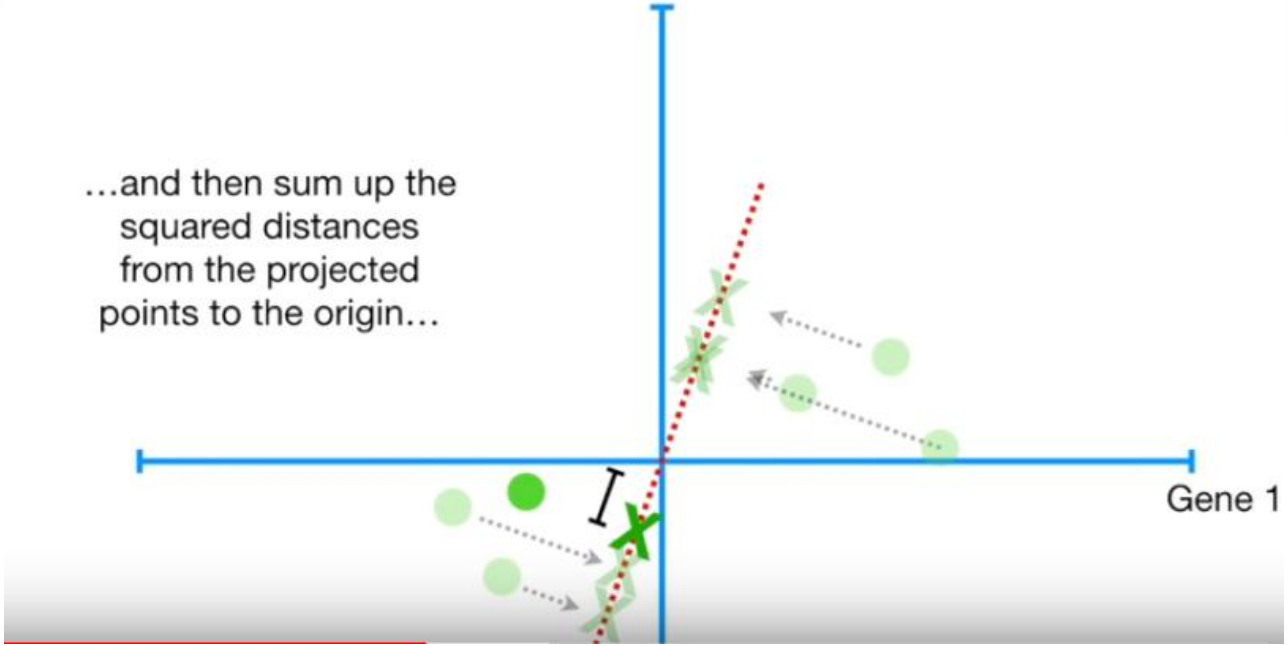

d_1
Gene 2
NOTE: I'm going to keep track of the distances we measure up here…
d_1
Gene 1

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



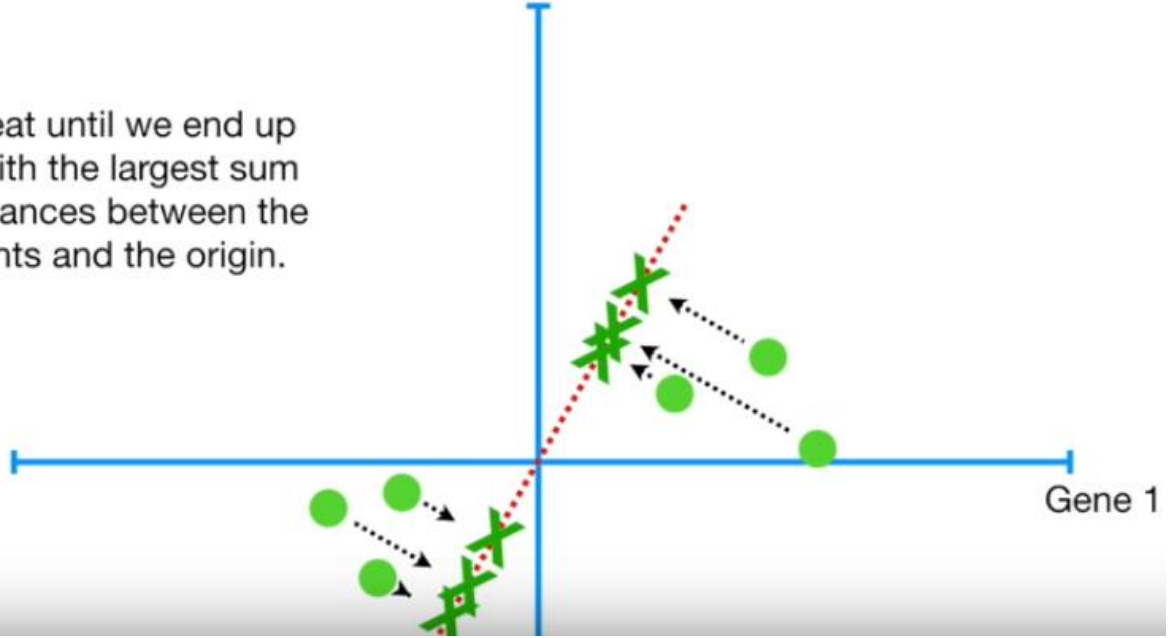

Gene 2
…project the data
onto the line…
Gene 1

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

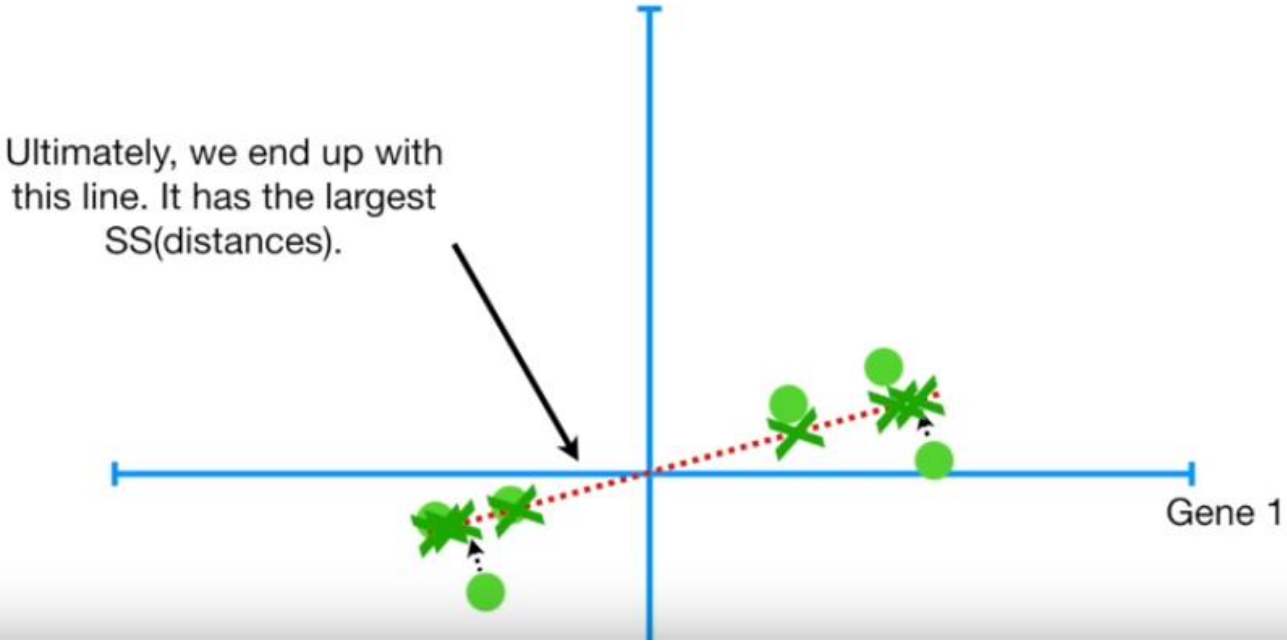


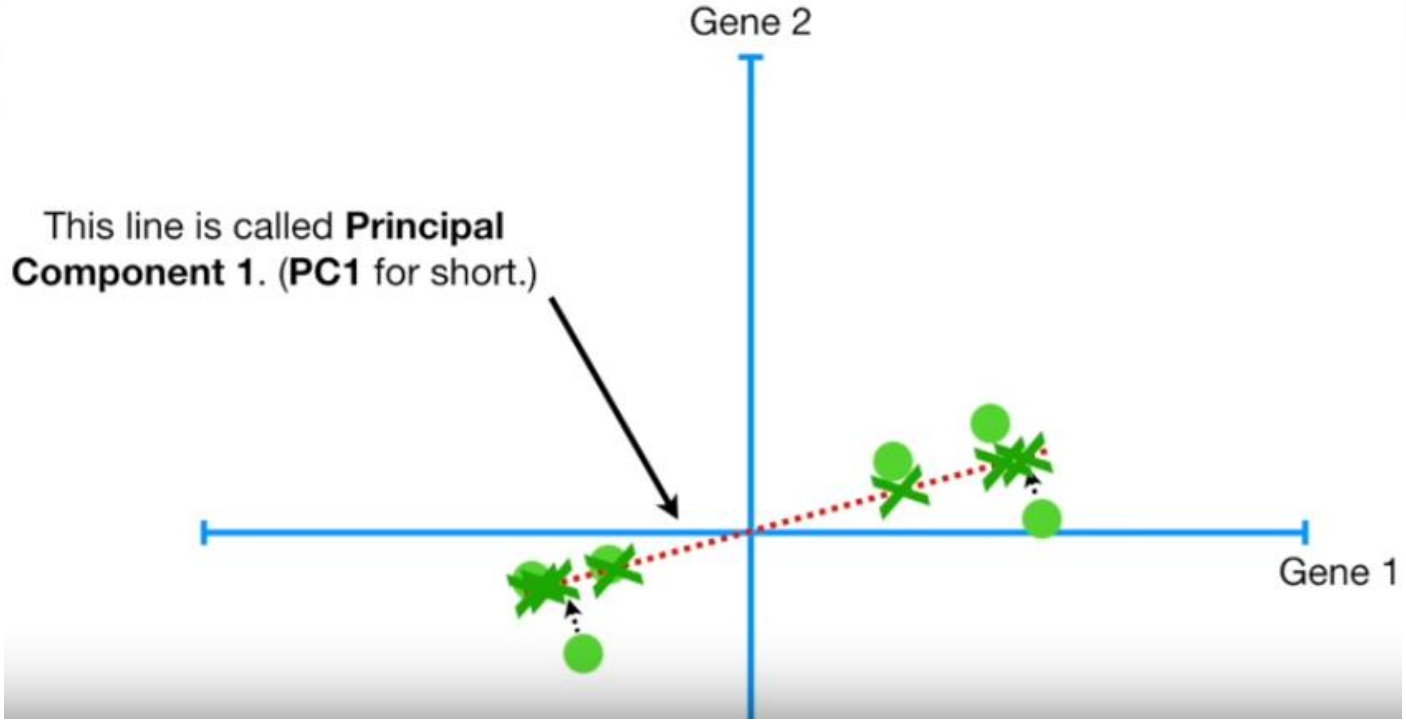
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = SS(\text{distances})$$

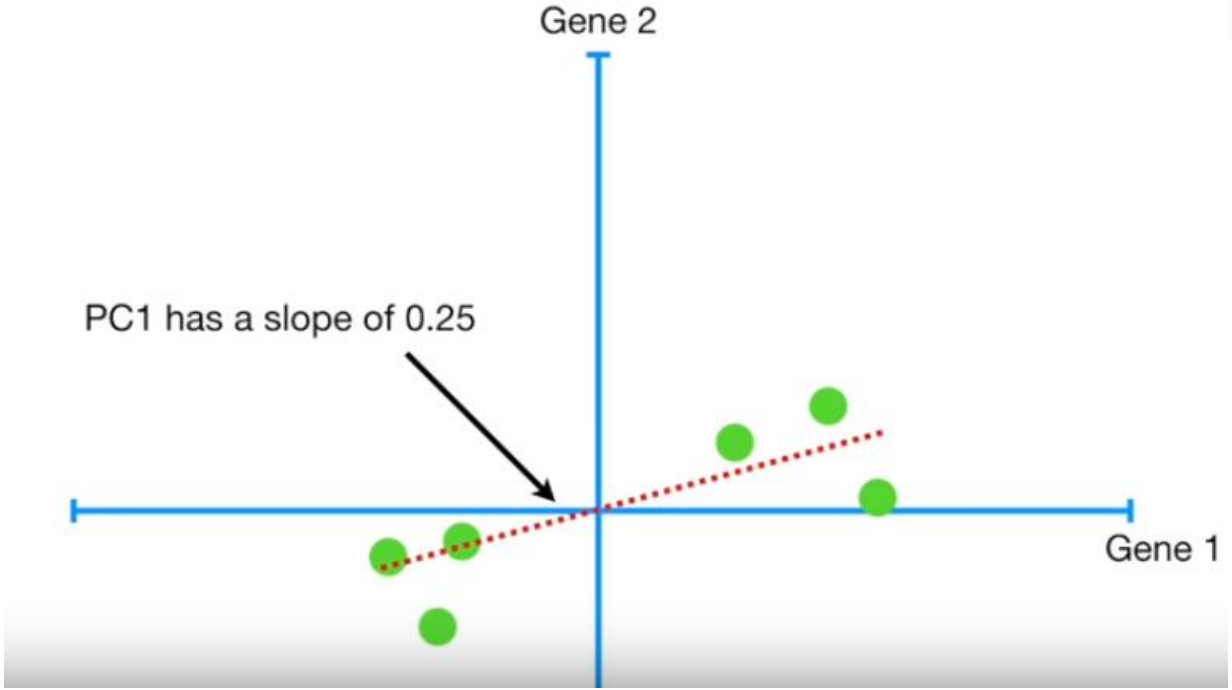
…and we repeat until we end up with the line with the largest sum of squared distances between the projected points and the origin.



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



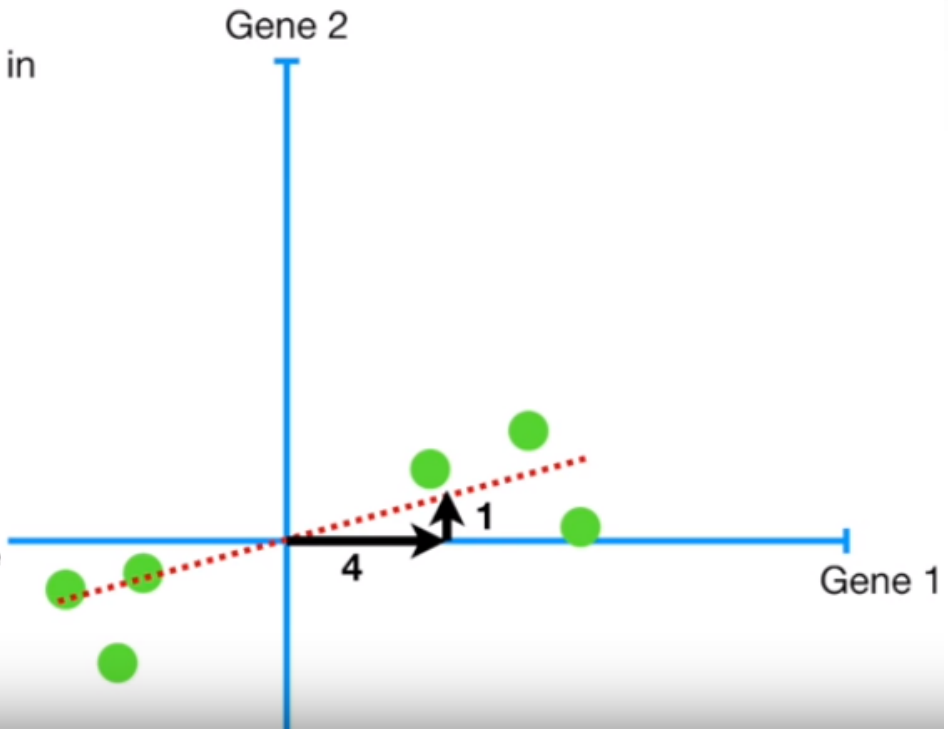

Gene 2
This line is called **Principal Component 1**. (**PC1** for short.)
Gene 1

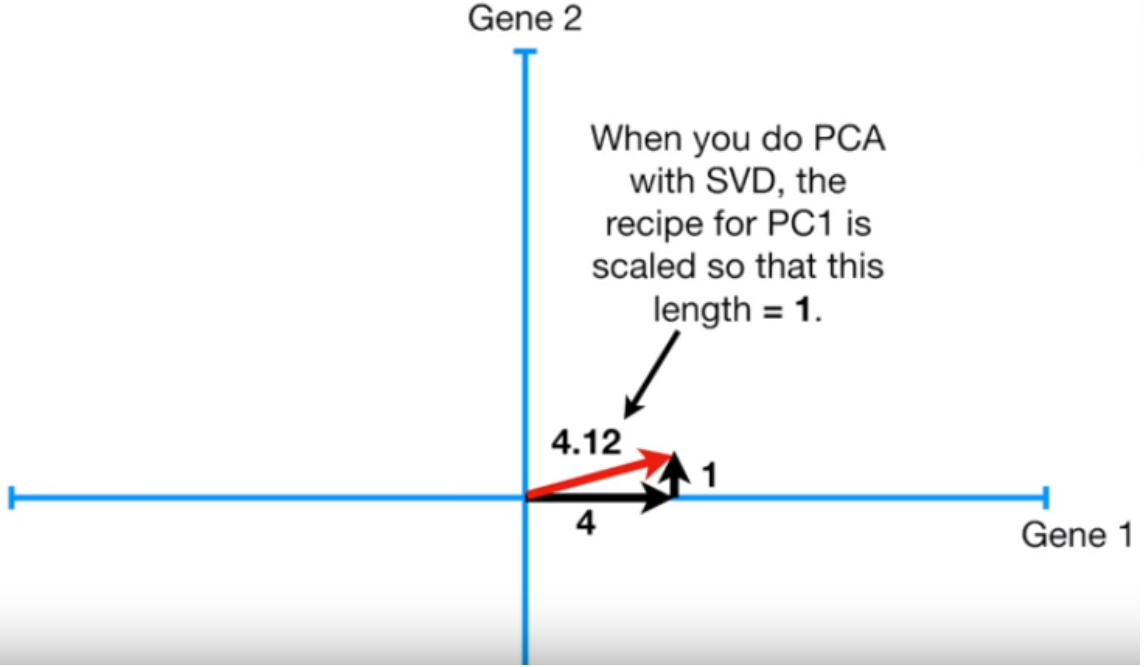

Gene 2
PC1 has a slope of 0.25
Gene 1

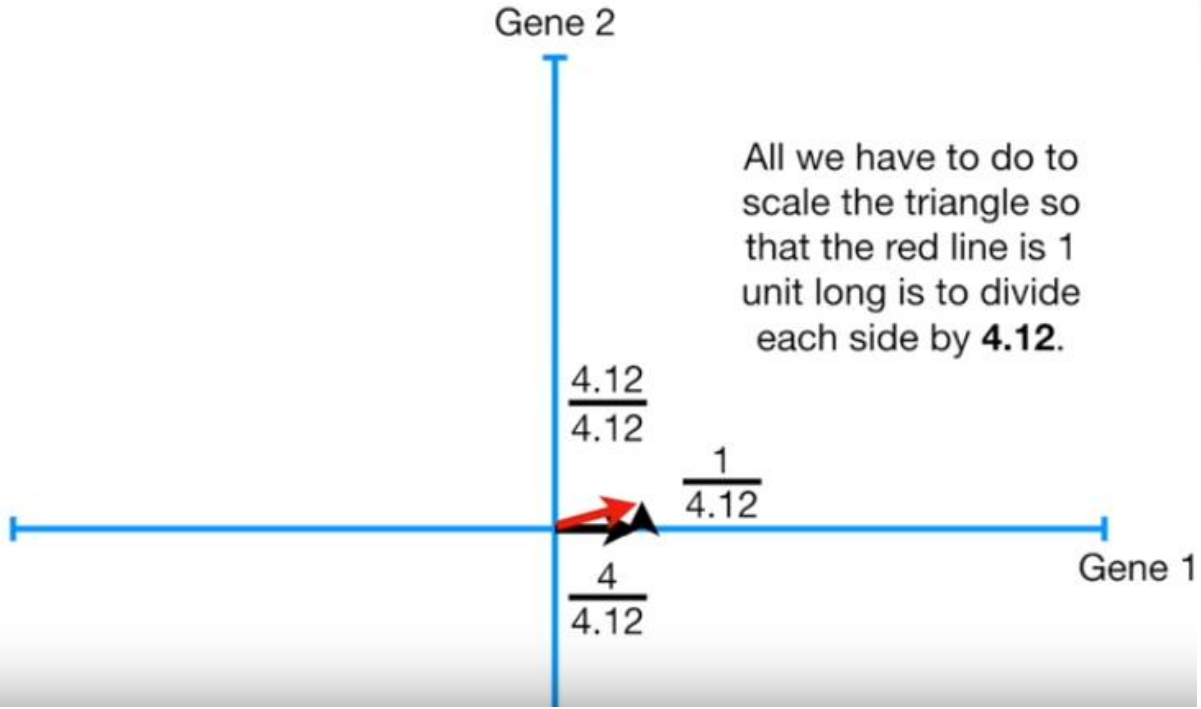
One way to think about PC1 is in terms of a cocktail recipe…

To make PC1
Mix **4** parts Gene 1
with **1** part Gene 2

The ratio of Gene 1 to Gene 2 tells you that Gene 1 is more important when it comes to describing how the data are spread out..



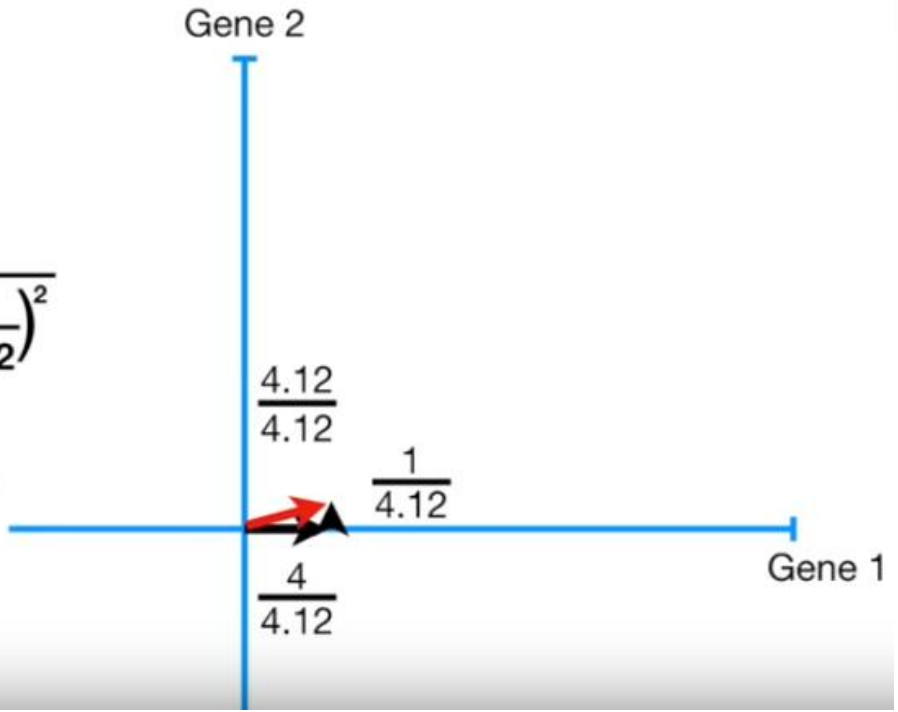

Gene 2
When you do PCA with SVD, the recipe for PC1 is scaled so that this length = **1**.
4.12
1
4
Gene 1

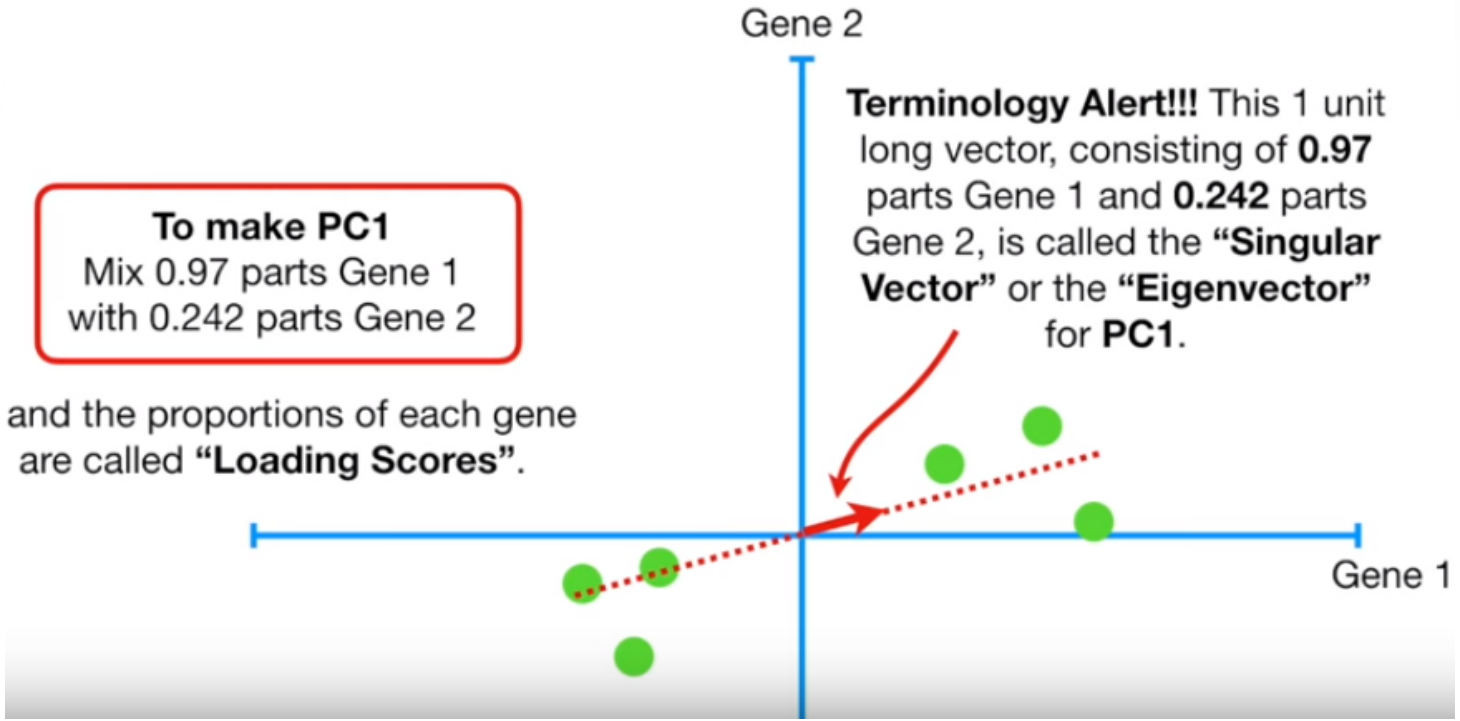

Gene 2
All we have to do to scale the triangle so that the red line is 1 unit long is to divide each side by **4.12**.
4.12 / 4.12
1 / 4.12
4 / 4.12
Gene 1

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

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

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**.

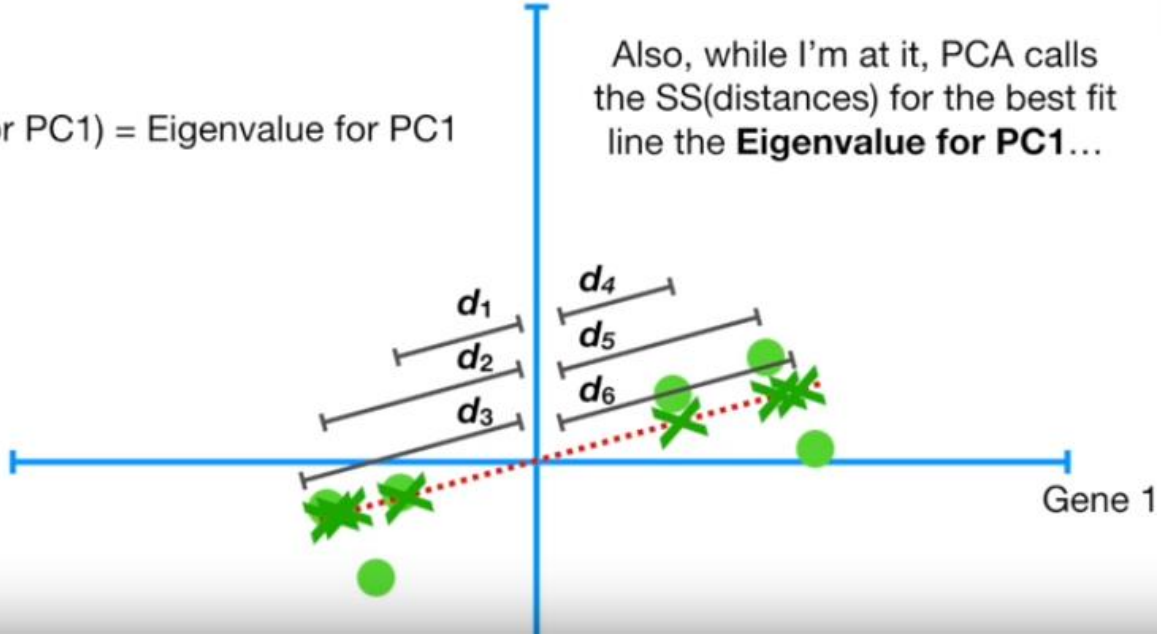



Gene 2
Terminology Alert!!! This 1 unit long vector, consisting of 0.97 parts Gene 1 and 0.242 parts Gene 2, is called the "Singular Vector" or the "Eigenvector" for PC1.
To make PC1
Mix 0.97 parts Gene 1 with 0.242 parts Gene 2
and the proportions of each gene are called "Loading Scores".
Gene 1

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

SS(distances for PC1) = Eigenvalue for PC1

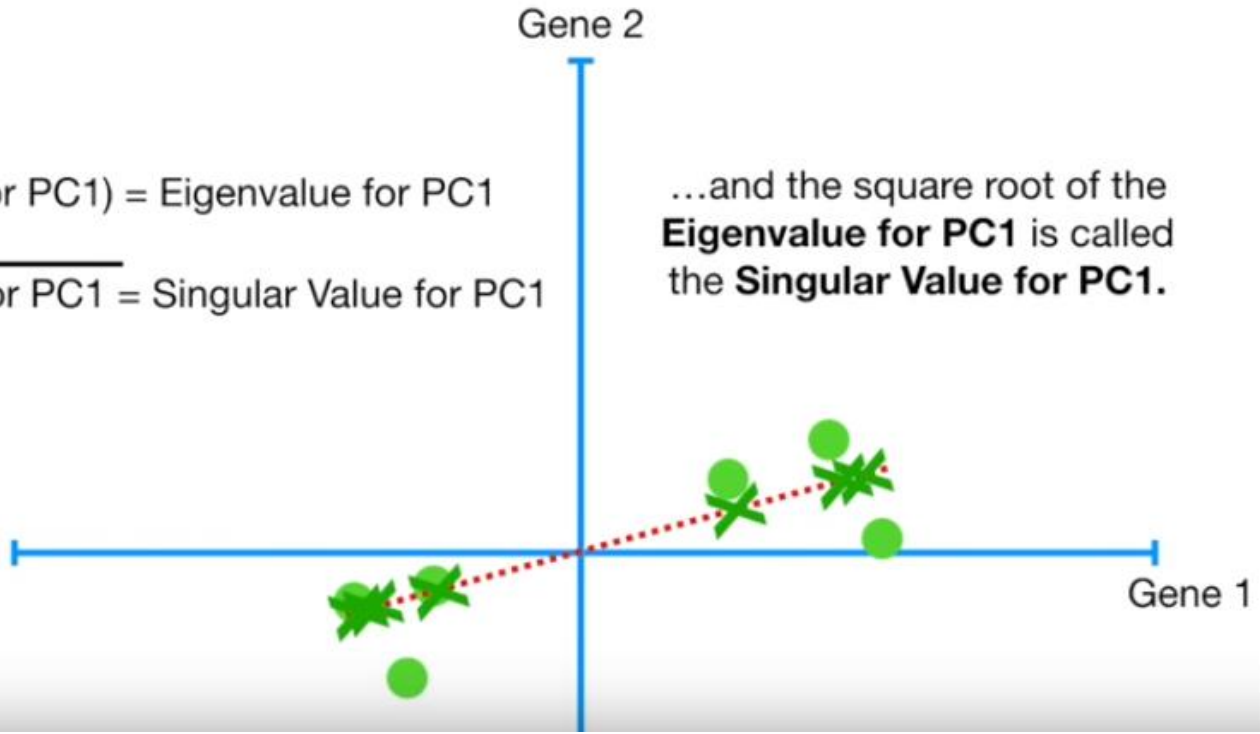
Also, while I'm at it, PCA calls the SS(distances) for the best fit line the **Eigenvalue for PC1**…

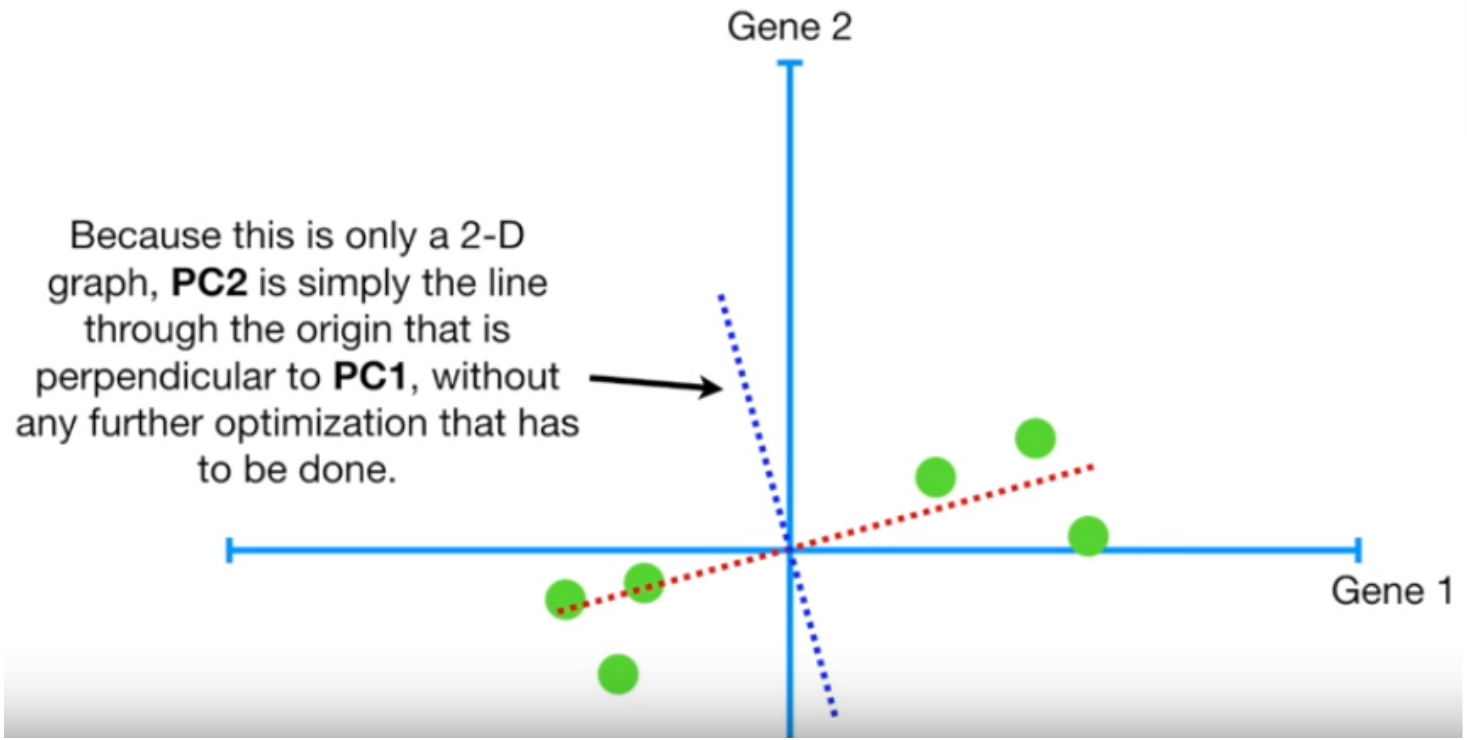


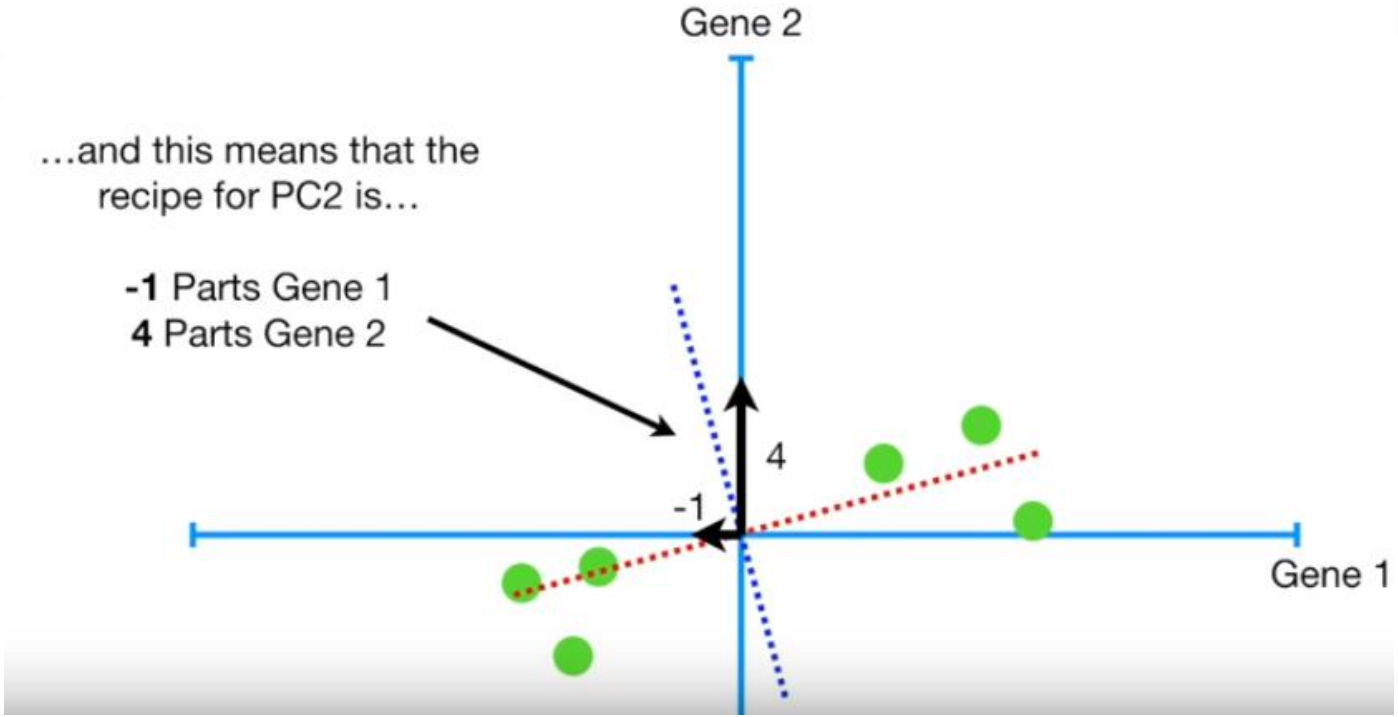
SS(distances for PC1) = Eigenvalue for PC1

$\sqrt{\text{Eigenvalue for PC1}}$ = Singular Value for PC1

…and the square root of the **Eigenvalue for PC1** is called the **Singular Value for PC1.**



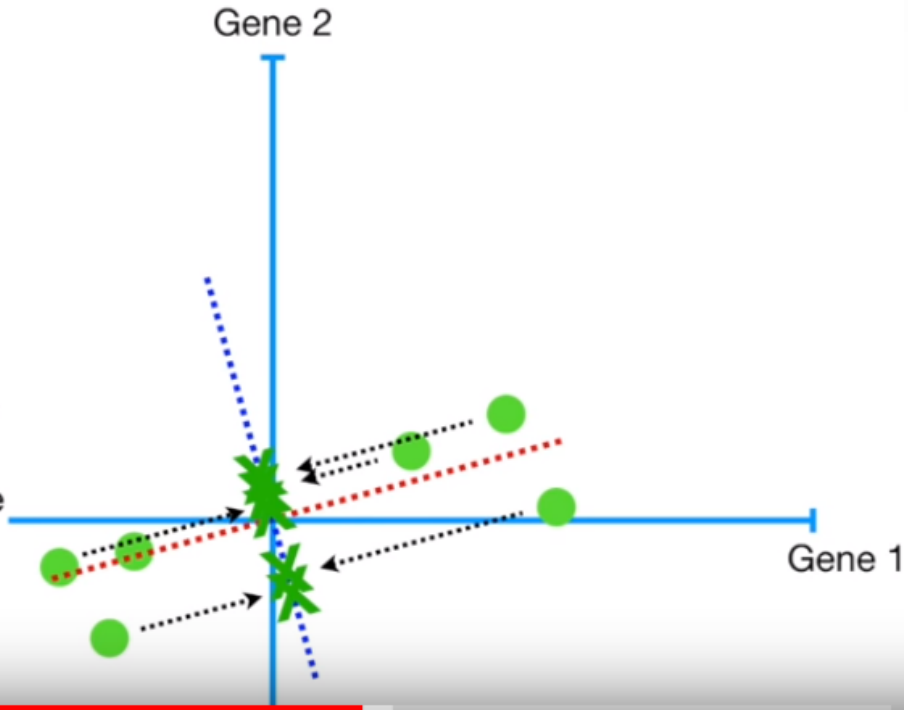

Gene 2
Because this is only a 2-D graph, **PC2** is simply the line through the origin that is perpendicular to **PC1**, without any further optimization that has to be done.
Gene 1


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

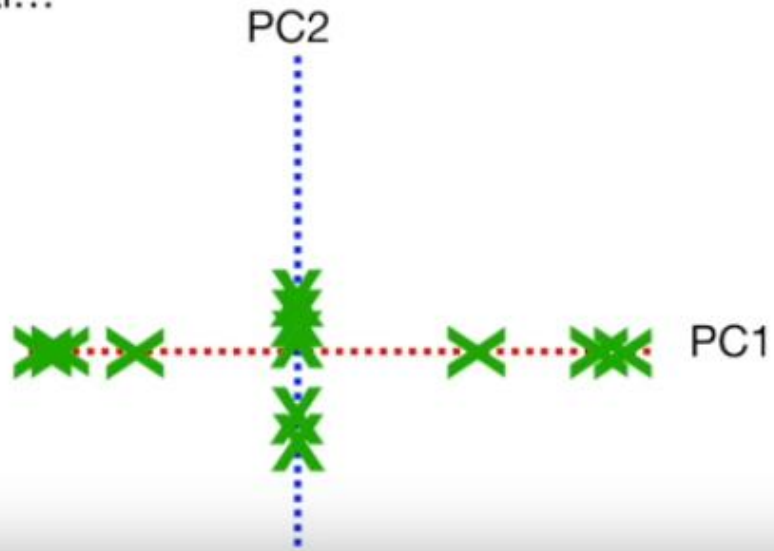
These are the **Loading Scores for PC2**.

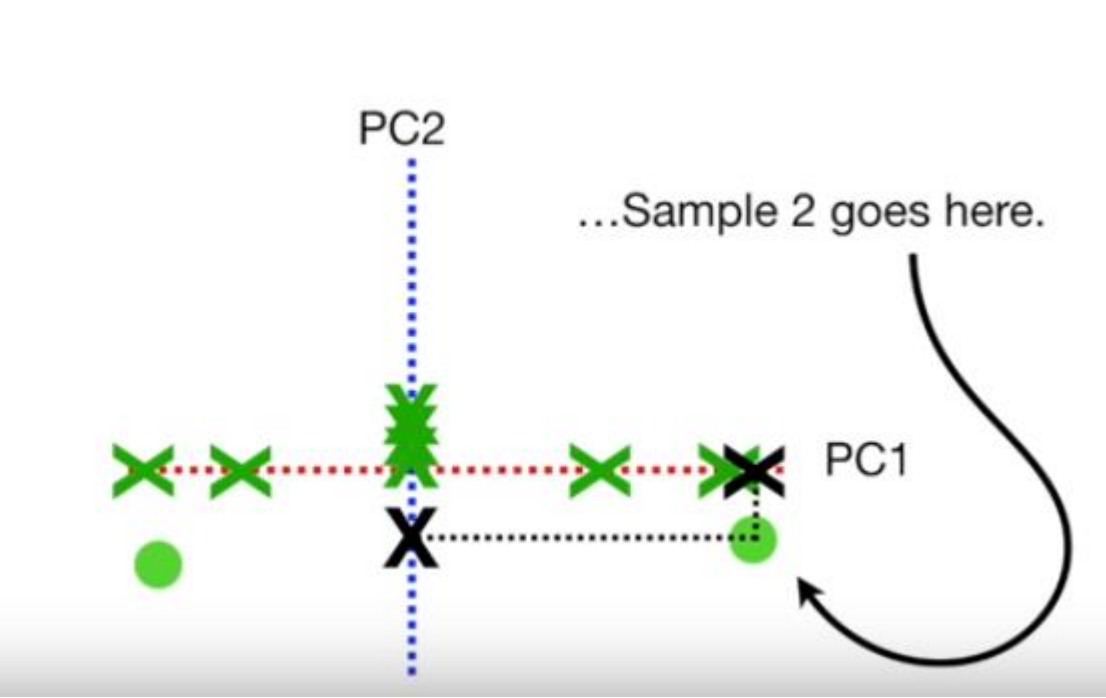
-0.242 Parts Gene 1
0.97 Parts Gene 2

They tell us that, in terms of how the values are projected onto PC2, Gene 2 is 4 times as important as Gene 1.



We simply rotate everything so that PC1 is horizontal…

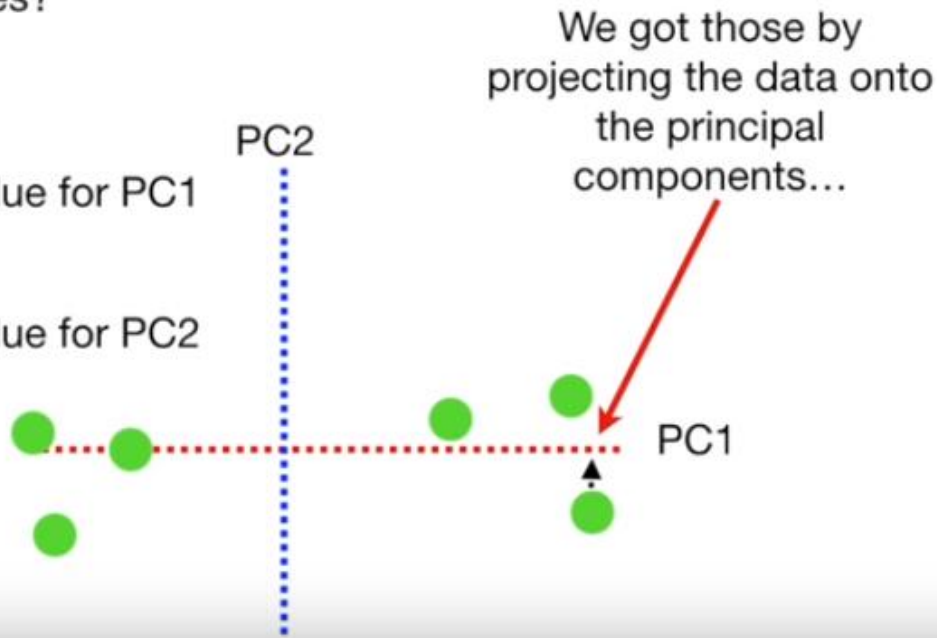



PC2
…Sample 2 goes here.
PC1

Remember the eigenvalues?

SS(distances for PC1) = Eigenvalue for PC1

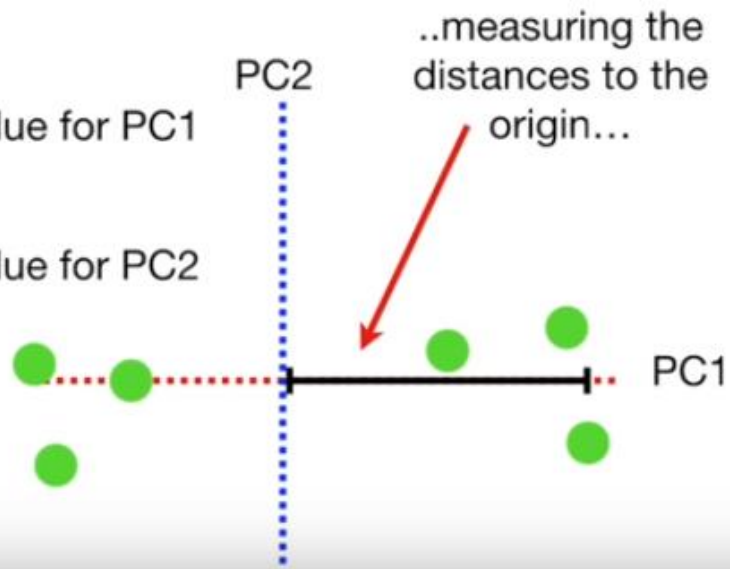
SS(distances for PC2) = Eigenvalue for PC2



Remember the eigenvalues?

SS(distances for PC1) = Eigenvalue for PC1

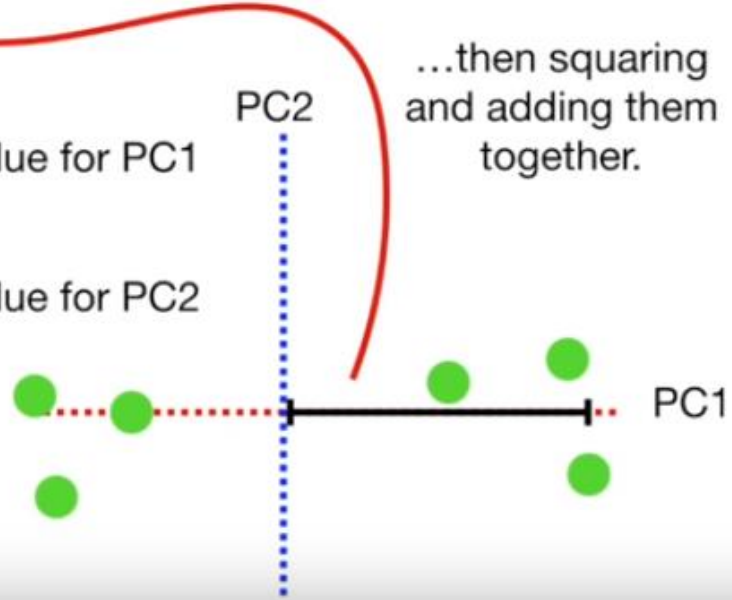
SS(distances for PC2) = Eigenvalue for PC2



Remember the eigenvalues?

SS(distances for PC1) = Eigenvalue for PC1

SS(distances for PC2) = Eigenvalue for PC2

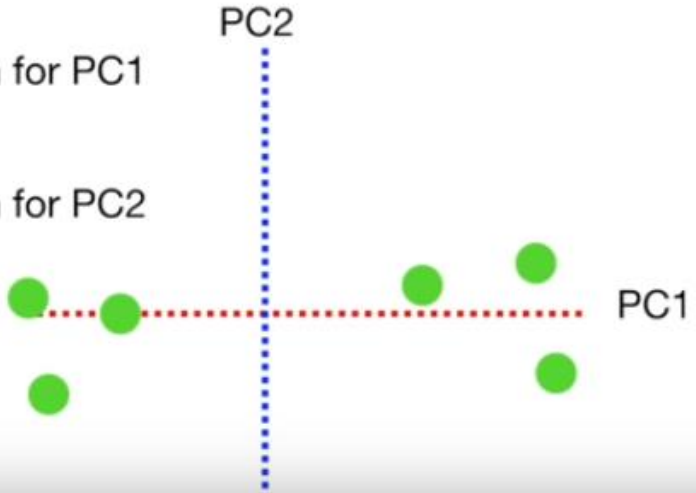


…then squaring and adding them together.

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

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

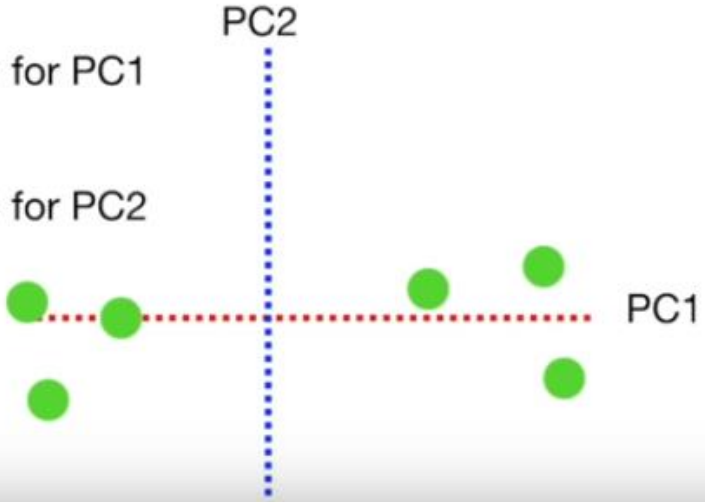
$$\frac{\text{SS(distances for PC2)}}{n - 1} = \text{Variation for PC2}$$



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

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

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



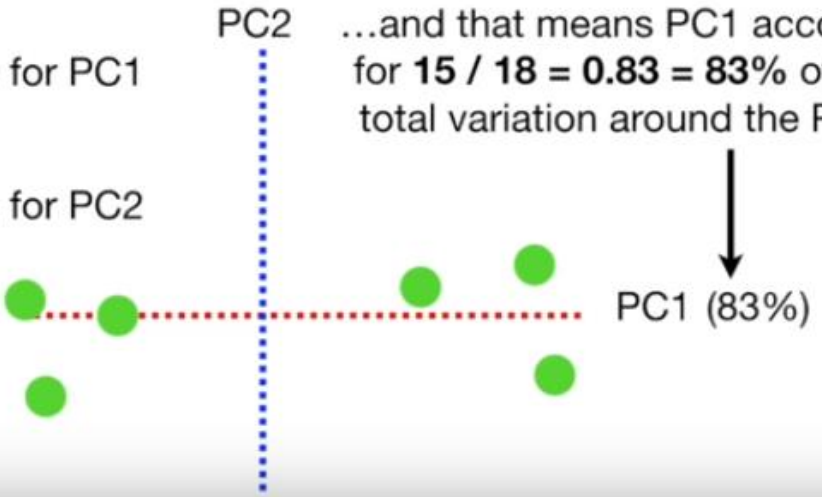
For the sake of the example, imagine that the Variation for **PC1 = 15**, and the variation for **PC2 = 3**.

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

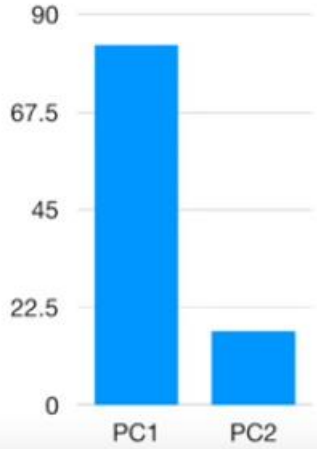
$$\frac{\text{SS(distances for PC1)}}{n - 1} = \text{Variation for PC1}$$

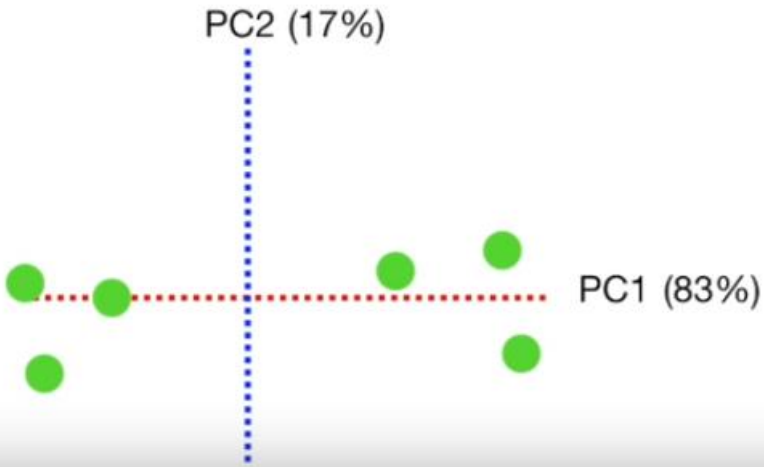
$$\frac{\text{SS(distances for PC2)}}{n - 1} = \text{Variation for PC2}$$

…and that means PC1 accounts for **15 / 18 = 0.83 = 83%** of the total variation around the PCs.



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





Dimensionality Reduction

- In dimensionality reduction, one does not drop any variable/do not select the variable from set of variables.
- In dimensionality reduction, one actually combines the variables into meaningful one.
- Techniques:
  - Principal Component Analysis
  - Factor Analysis