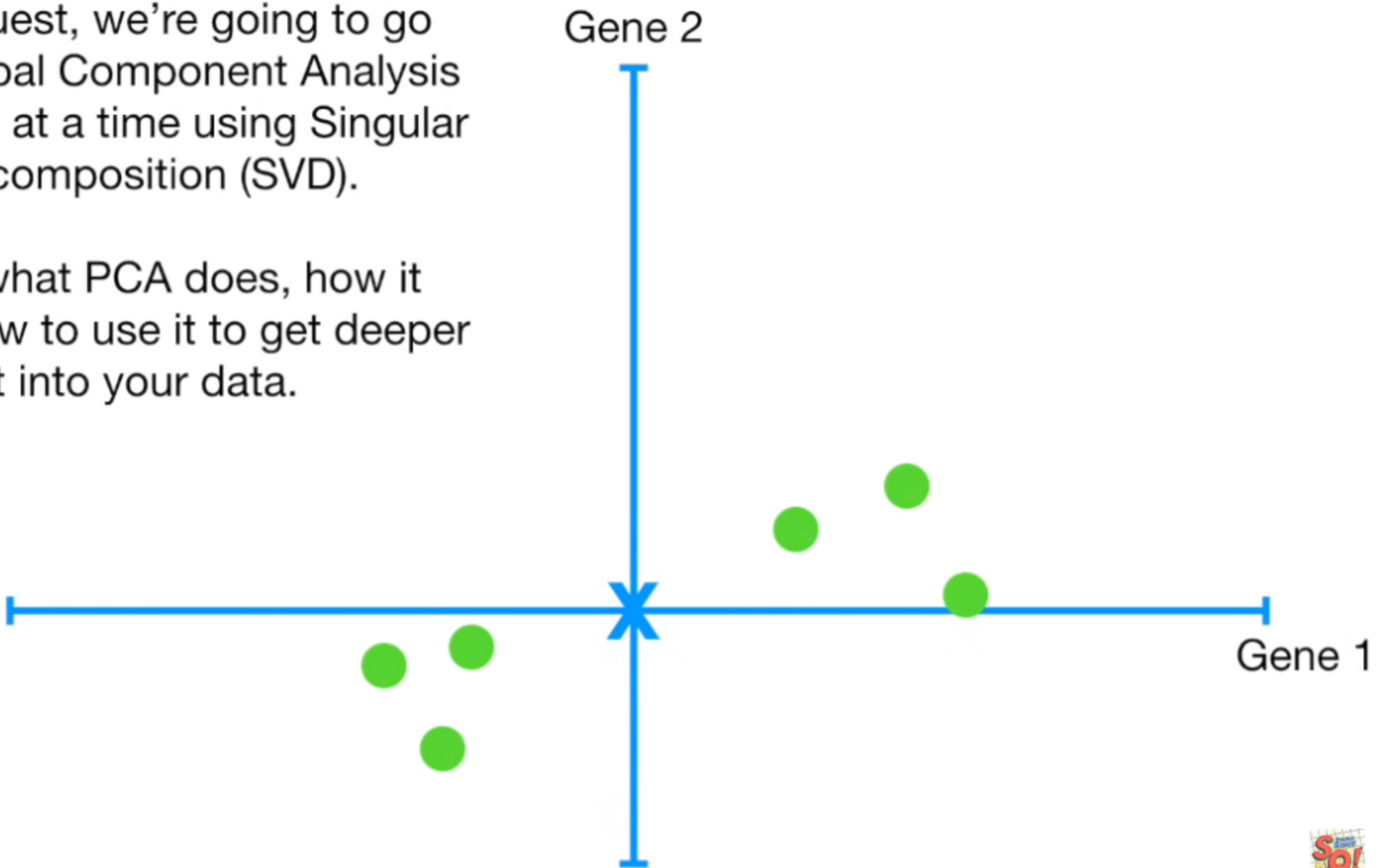
StatQuest: PCA, Step-By-Step!!!

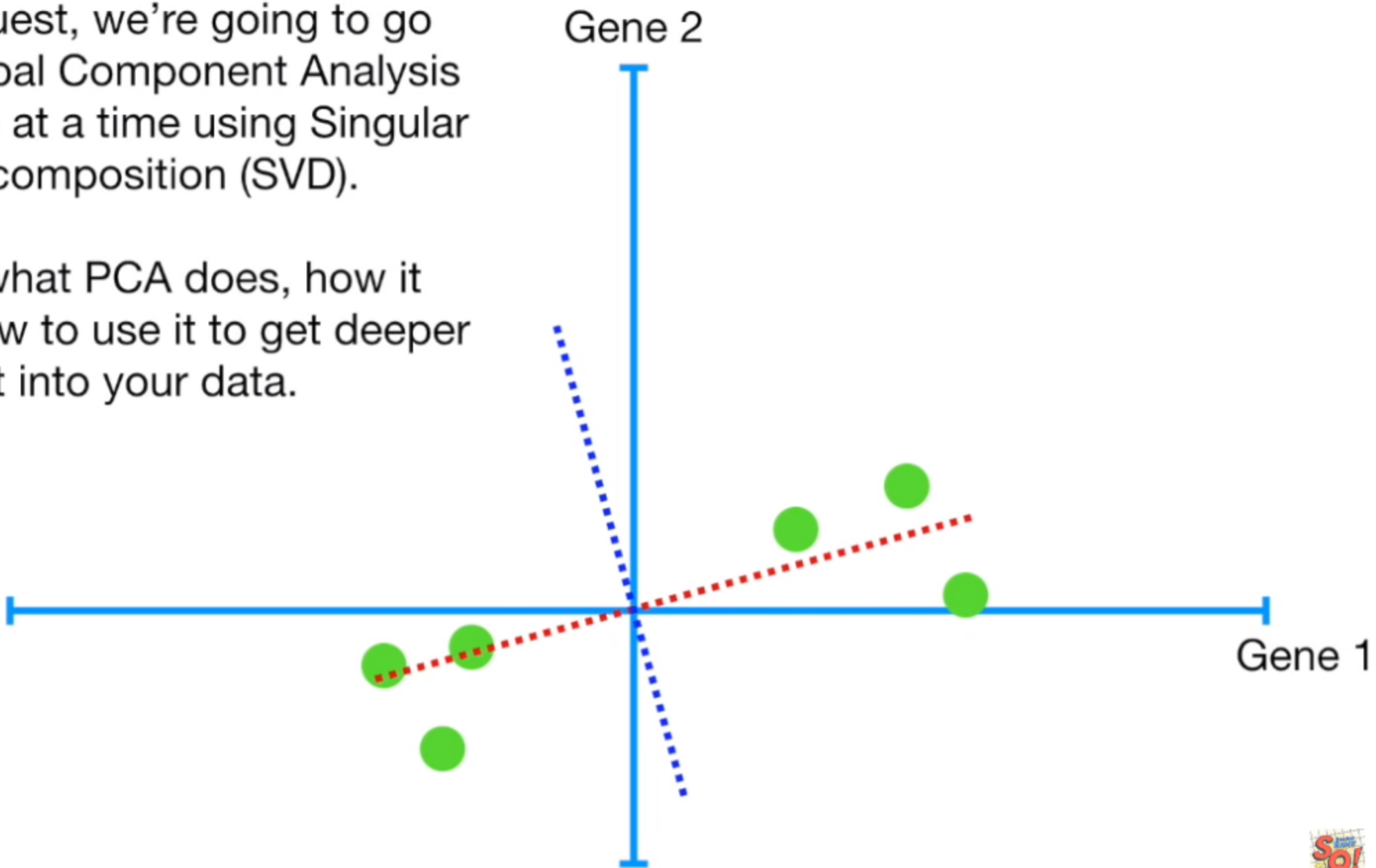
In this StatQuest, we're going to go through Principal Component Analysis (PCA) one step at a time using Singular Value Decomposition (SVD).

You'll learn what PCA does, how it does it, and how to use it to get deeper insight into your data.



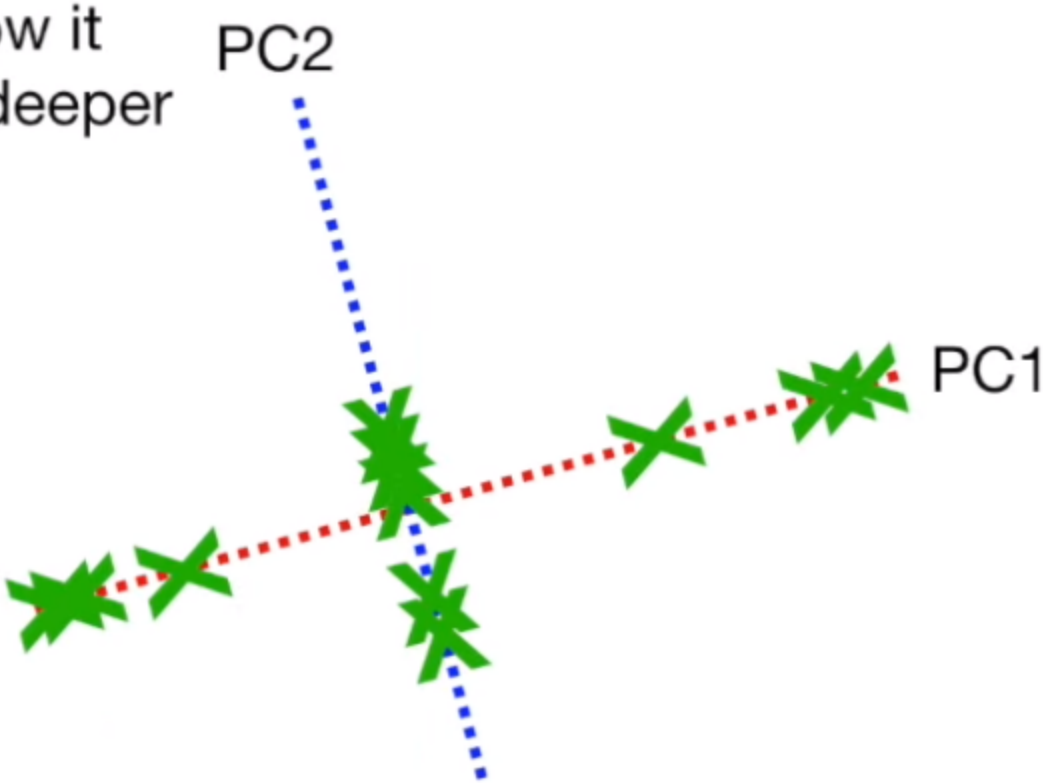
In this StatQuest, we're going to go through Principal Component Analysis (PCA) one step at a time using Singular Value Decomposition (SVD).

You'll learn what PCA does, how it does it, and how to use it to get deeper insight into your data.



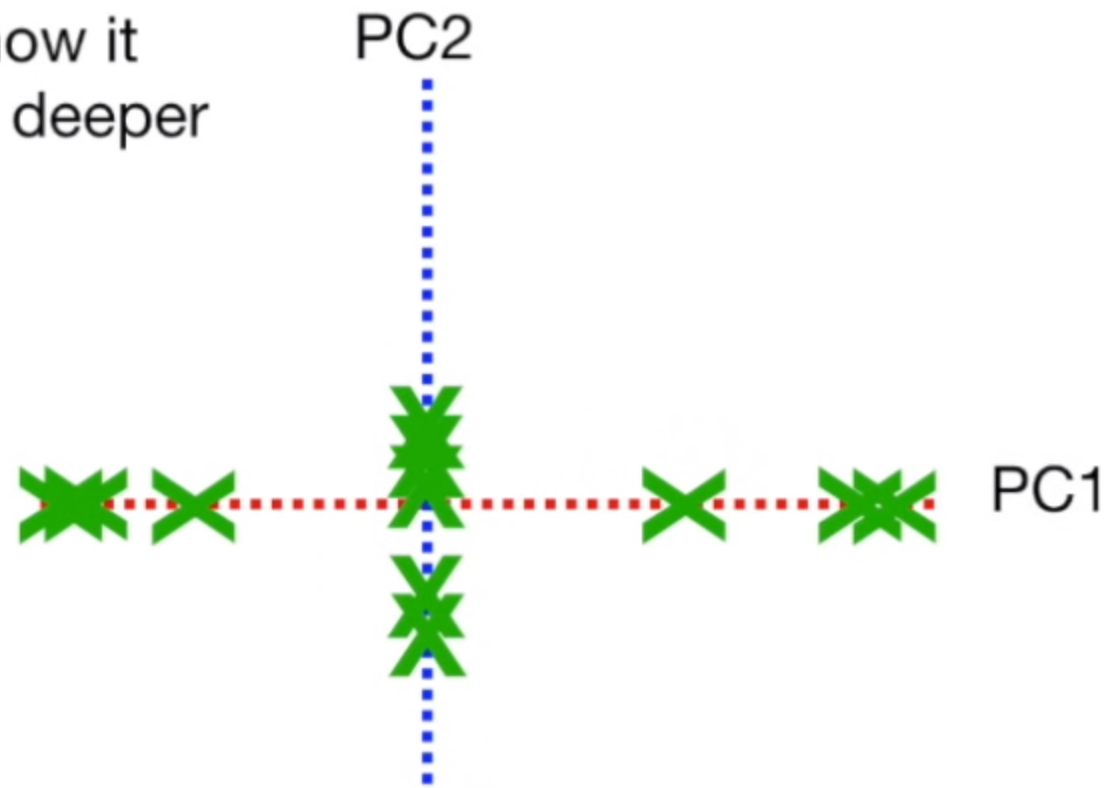
In this StatQuest, we're going to go through Principal Component Analysis (PCA) one step at a time using Singular Value Decomposition (SVD).

You'll learn what PCA does, how it does it, and how to use it to get deeper insight into your data.



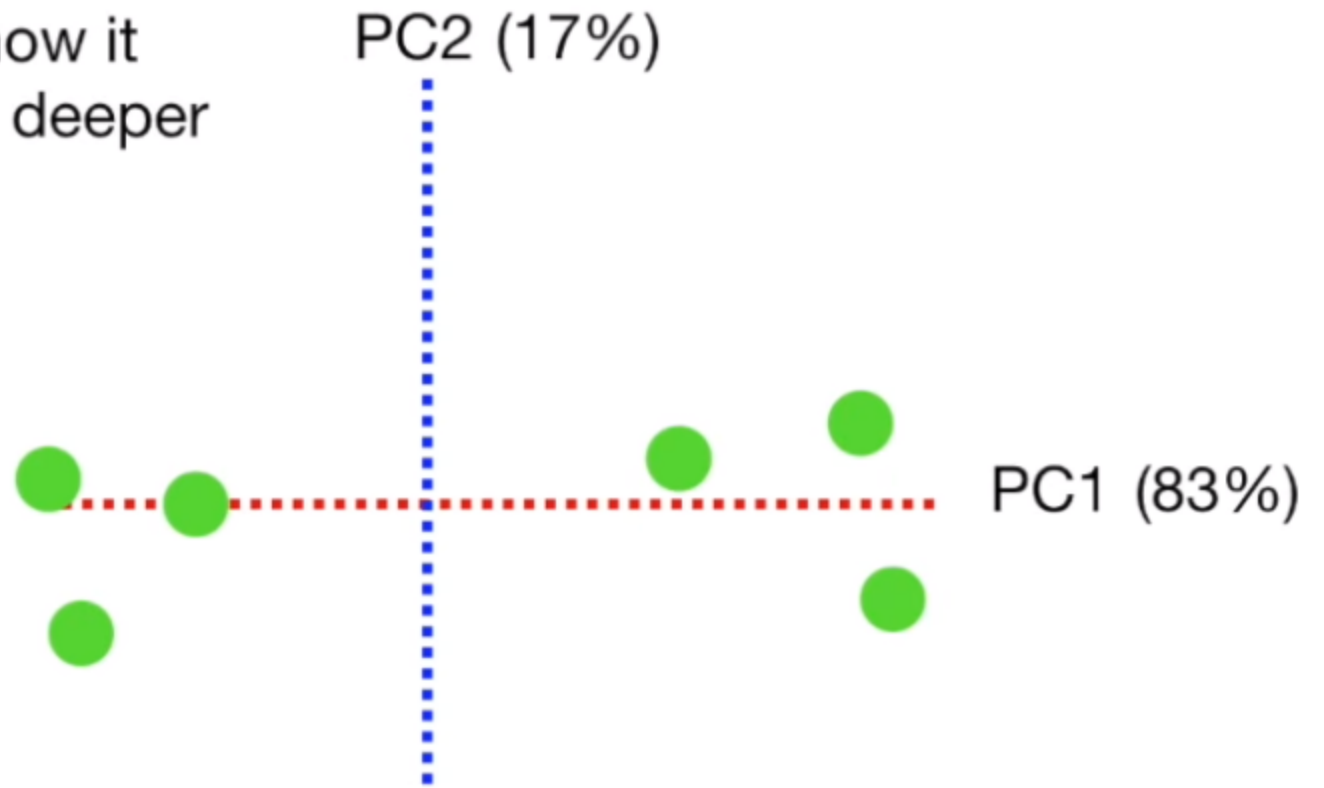
In this StatQuest, we're going to go through Principal Component Analysis (PCA) one step at a time using Singular Value Decomposition (SVD).

You'll learn what PCA does, how it does it, and how to use it to get deeper insight into your data.



In this StatQuest, we're going to go through Principal Component Analysis (PCA) one step at a time using Singular Value Decomposition (SVD).

You'll learn what PCA does, how it does it, and how to use it to get deeper insight into your 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

Let’s start with a simple data set.

	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…

	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.

	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

NOTE: If you are not into mice and genes…

	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.

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

For example, 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	…

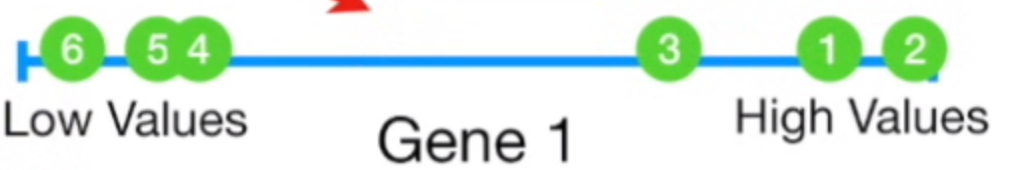
…or the samples could be businesses and the variables could be market capitalization and number of employees.

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

Ok, now we're back to mice and genes.

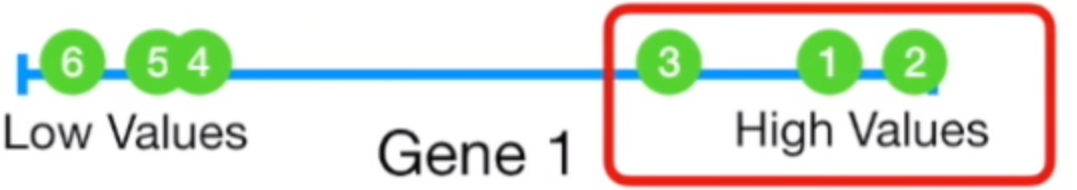
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

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…

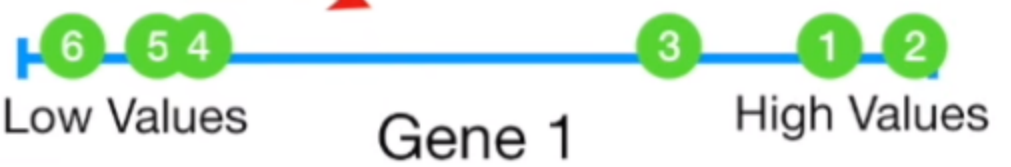


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

…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

…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.