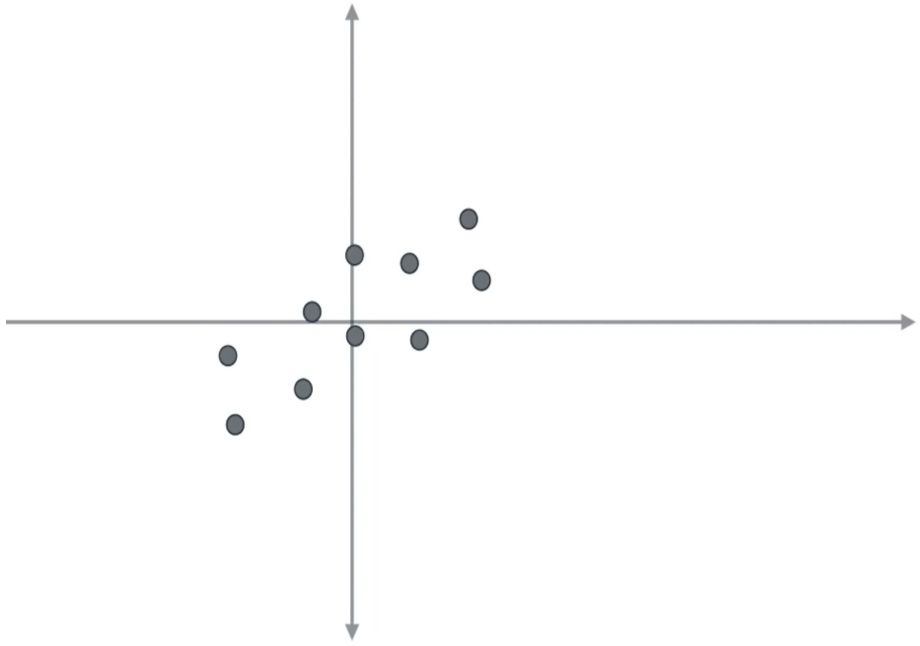
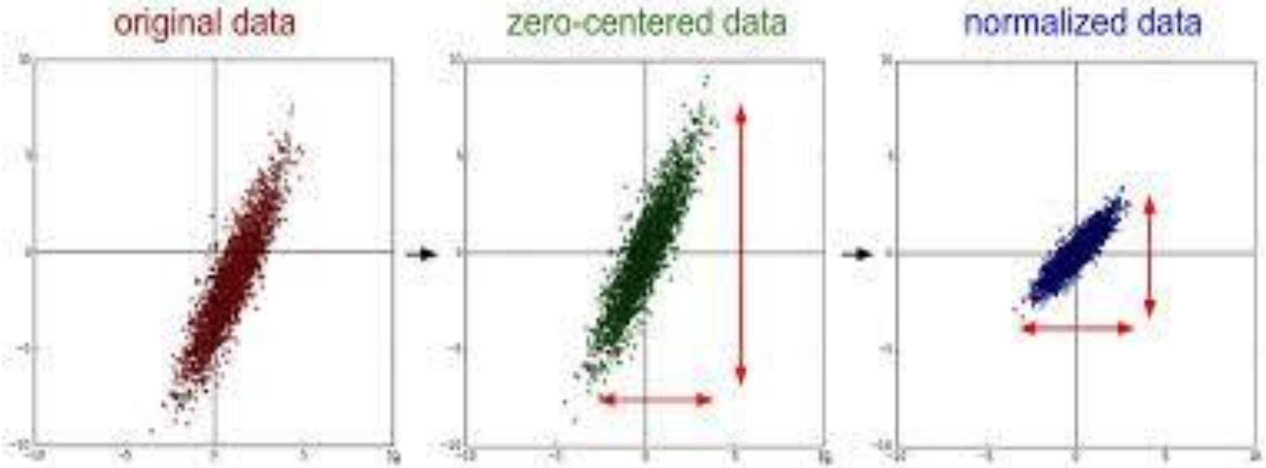
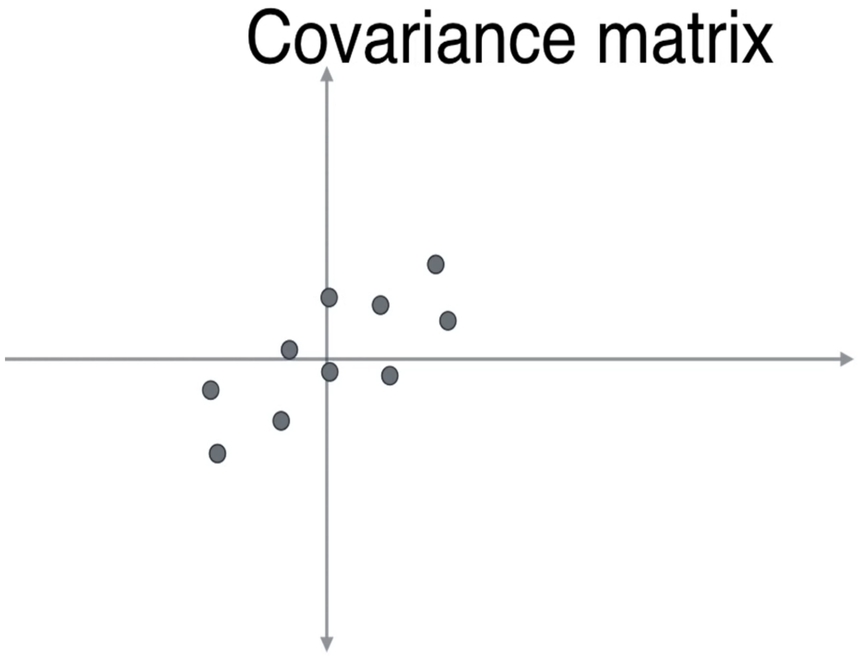
Principal Component Analysis

Why Centering?





Find m by n covariance matrix

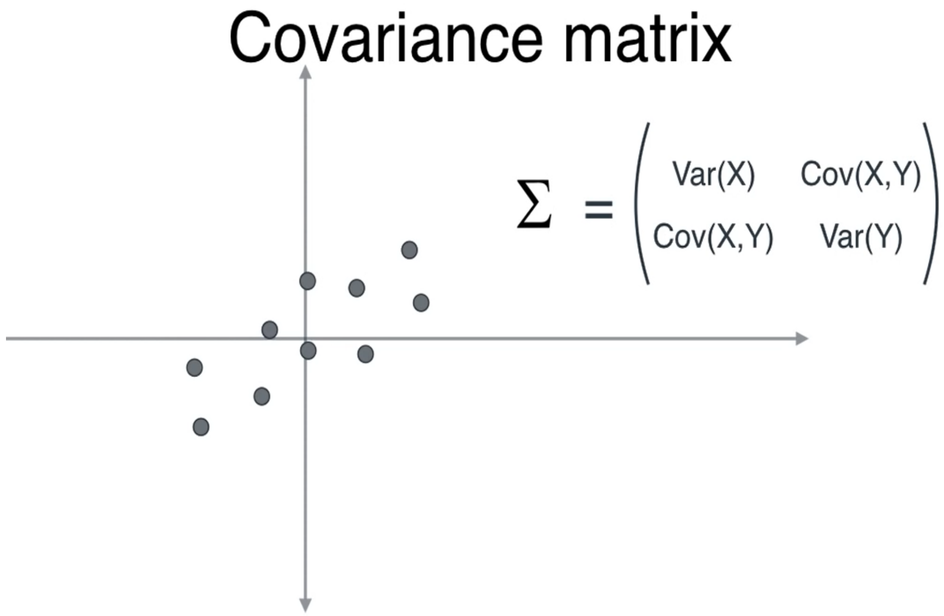


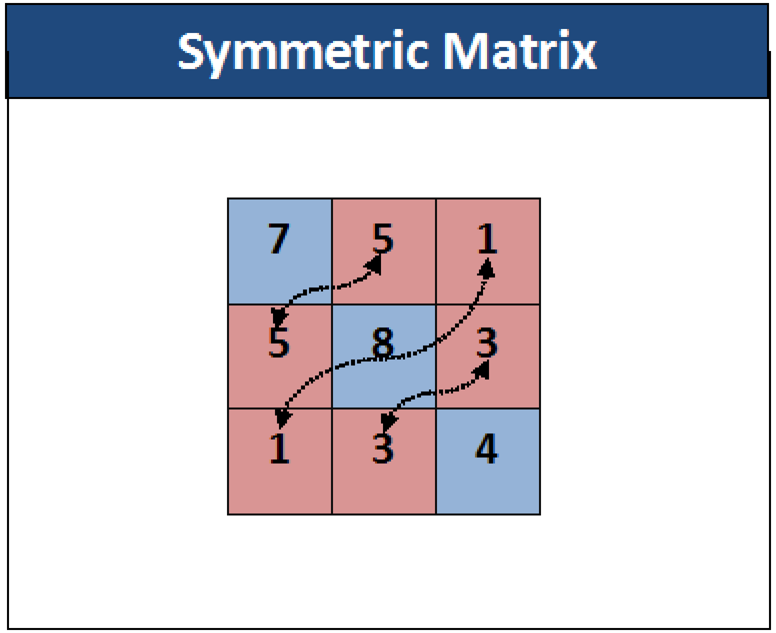
$$cov_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Symmetric Matrix

$$\begin{bmatrix} 7 & 5 & 1 \\ 5 & 8 & 3 \\ 1 & 3 & 4 \end{bmatrix}$$

Find m by n covariance matrix

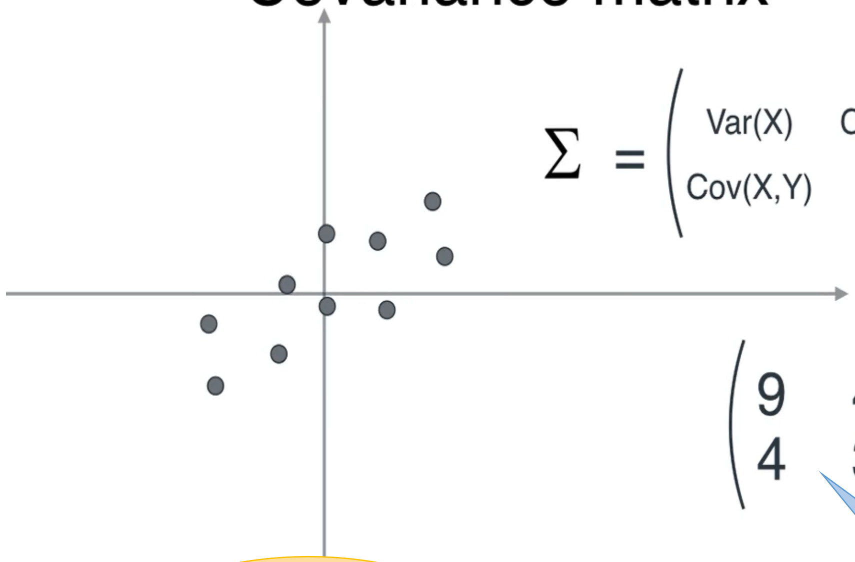




Find m by n covariance matrix

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Covariance matrix

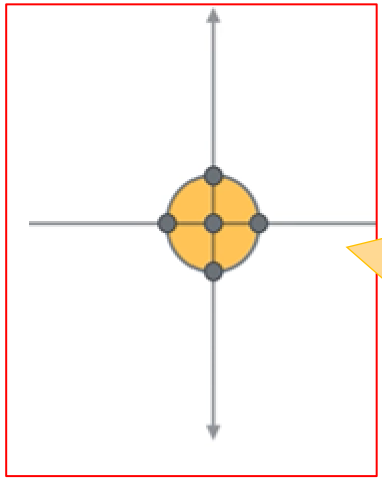


$$\Sigma = \begin{pmatrix} \mathrm{Var(X)} & \mathrm{Cov(X,Y)} \\ \mathrm{Cov(X,Y)} & \mathrm{Var(Y)} \end{pmatrix}$$

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

Symmetric Matrix

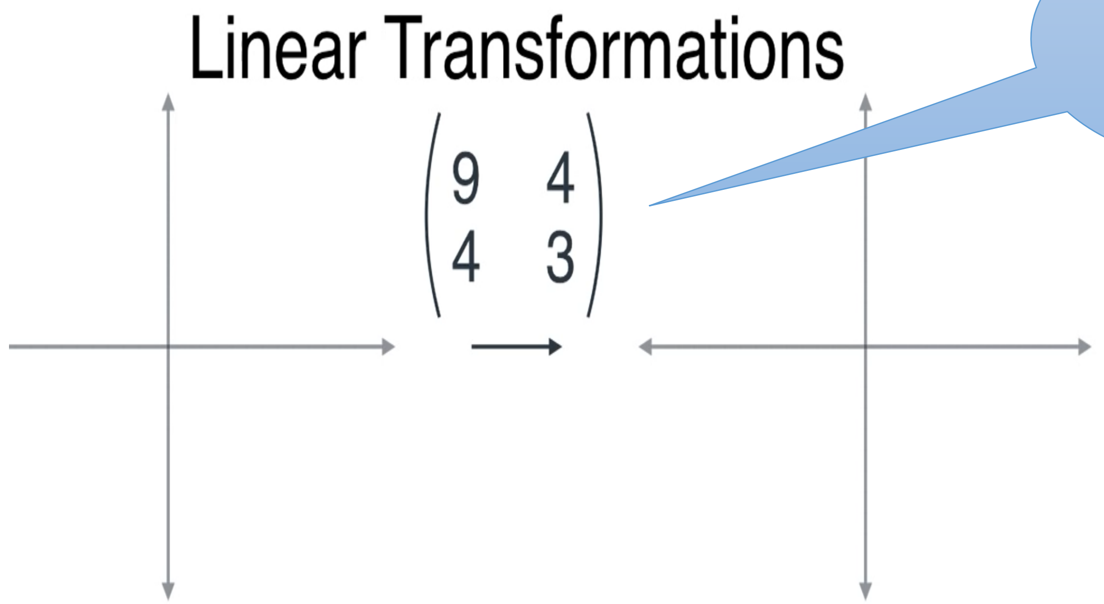
$$\begin{bmatrix} 7 & 5 & 1 \\ 5 & 8 & 3 \\ 1 & 3 & 4 \end{bmatrix}$$



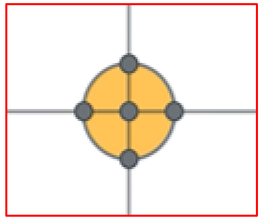
Say our raw data is like this and we will try to improve its presentation

Through Cov Matrix we found some magic numbers!

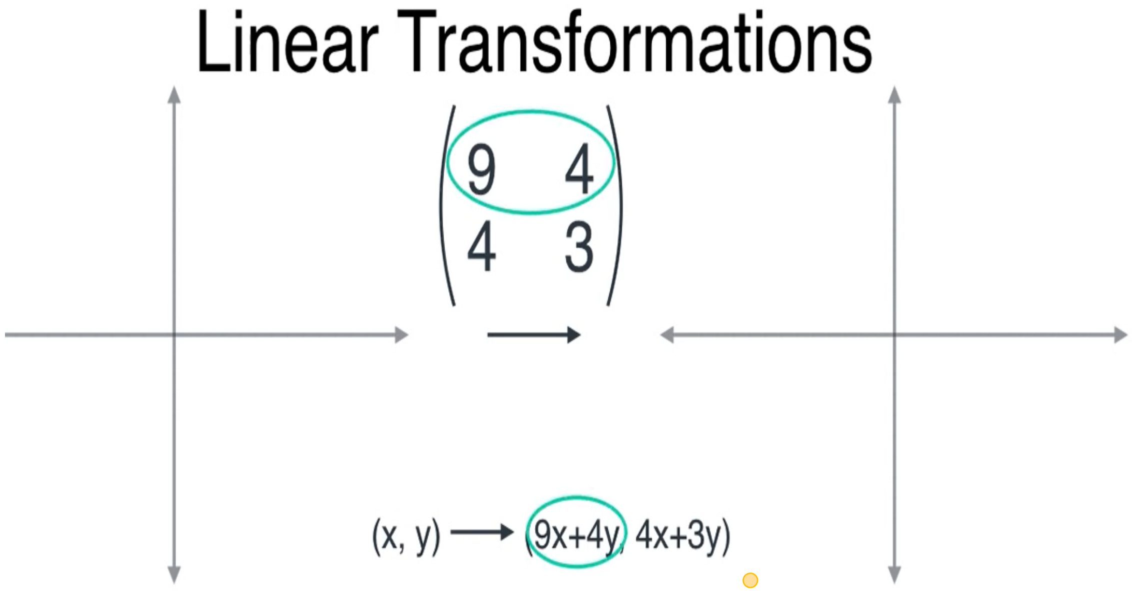
Transformation

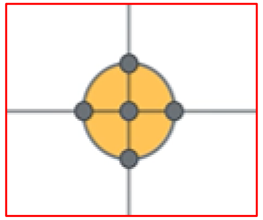


Magic numbers!



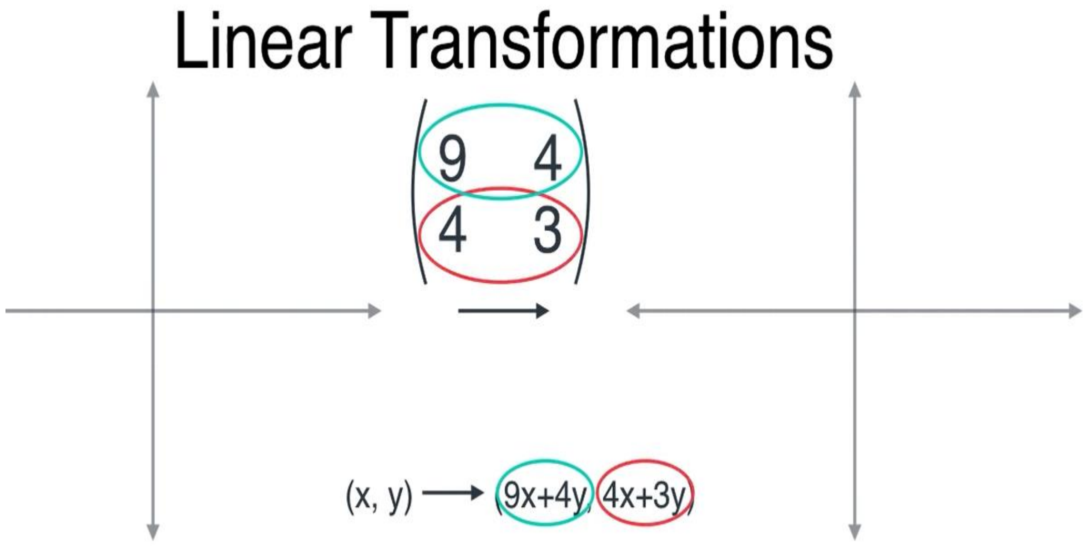
Transformation through what?

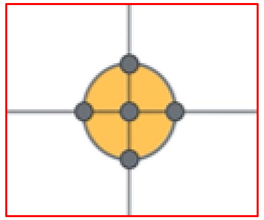




Transformation need some inspiration !

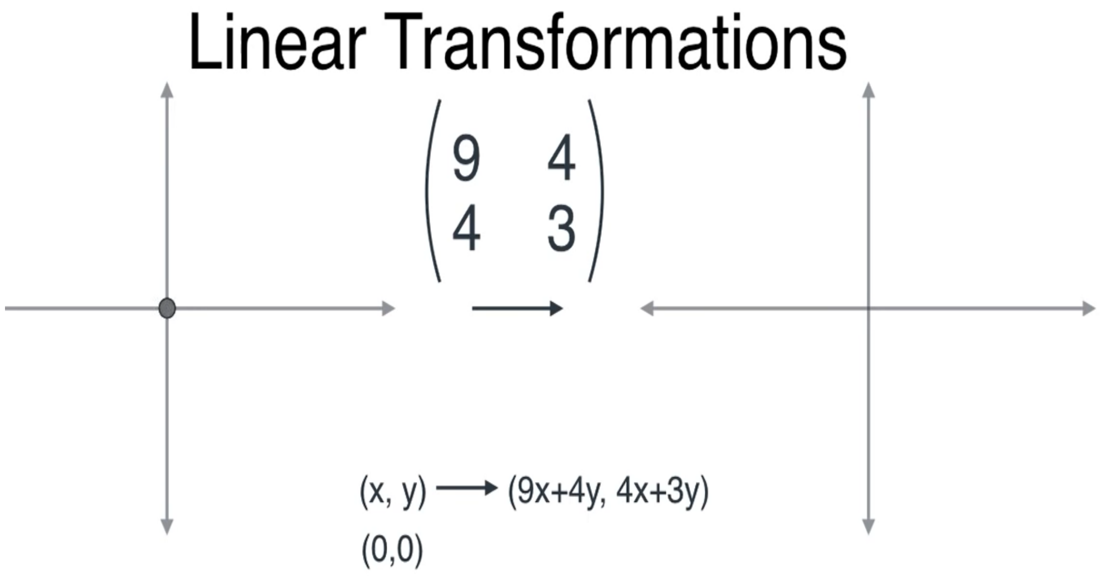
Transformation through what?

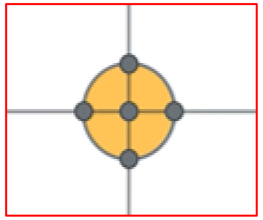


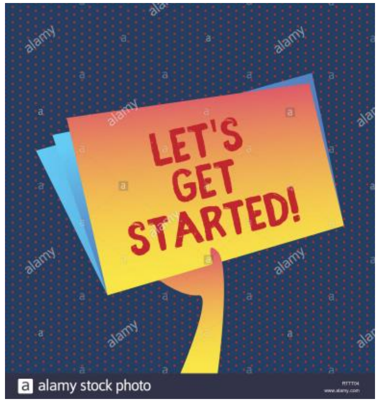


Transformation need some inspiration !

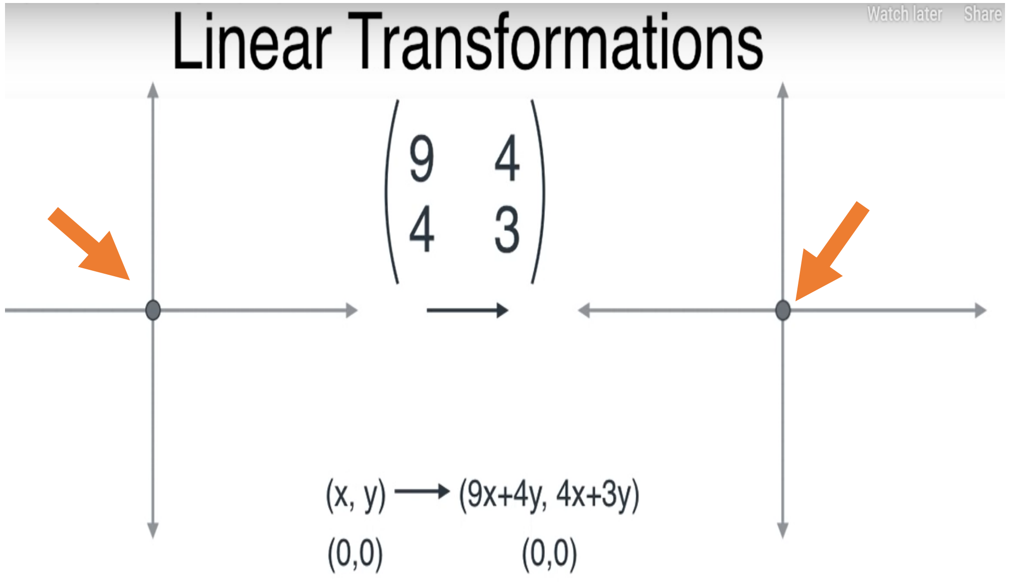
Transformation begins: where (0,0) will go?

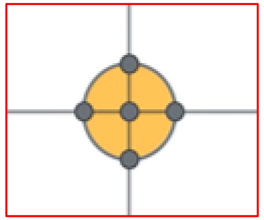




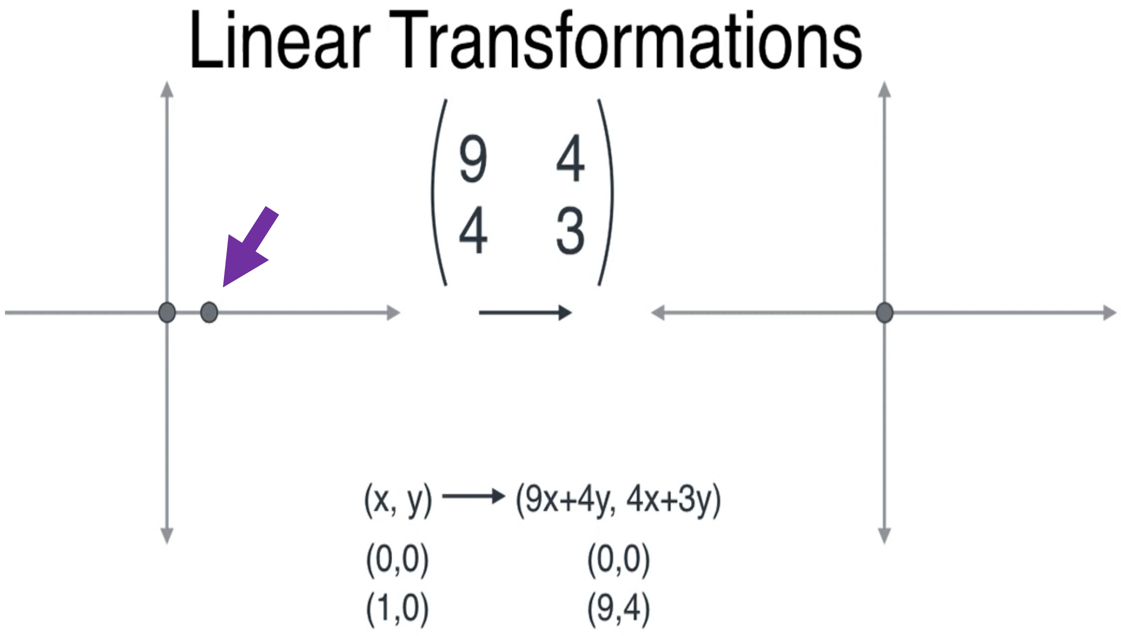


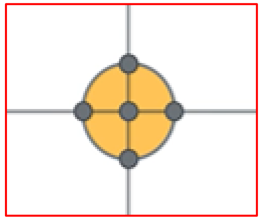
Transformation begins: where (0,0) will go?



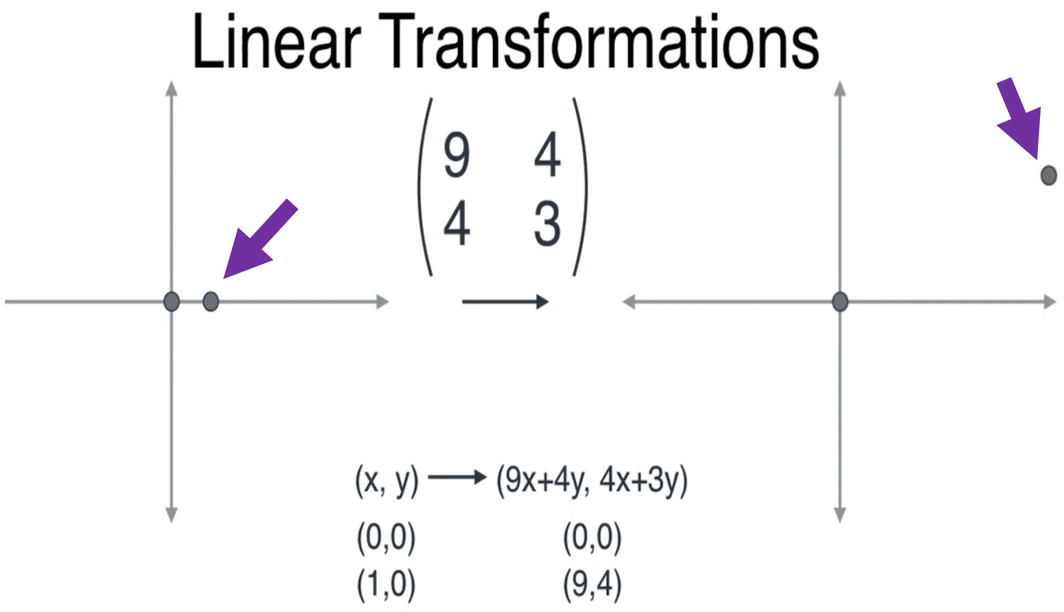


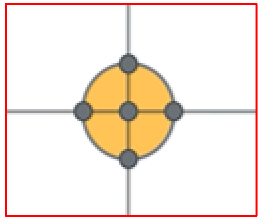
Transformation begins: where (1,0) will go?



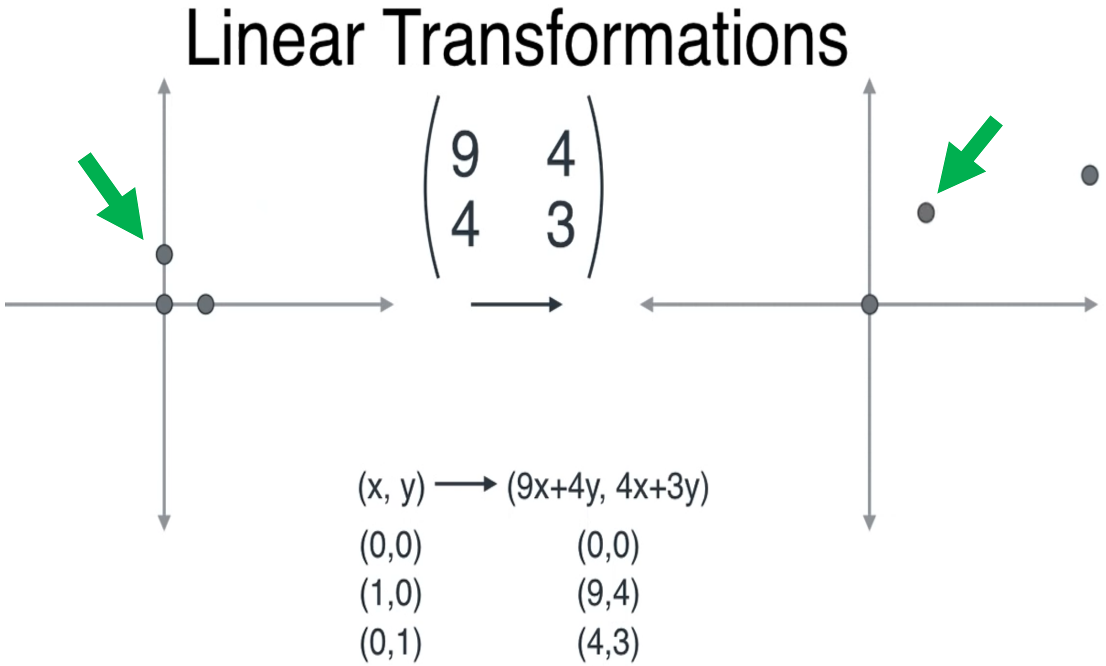


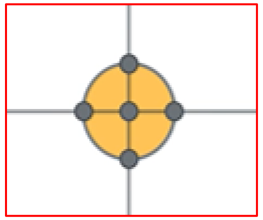
Transformation begins: where (1,0) will go?



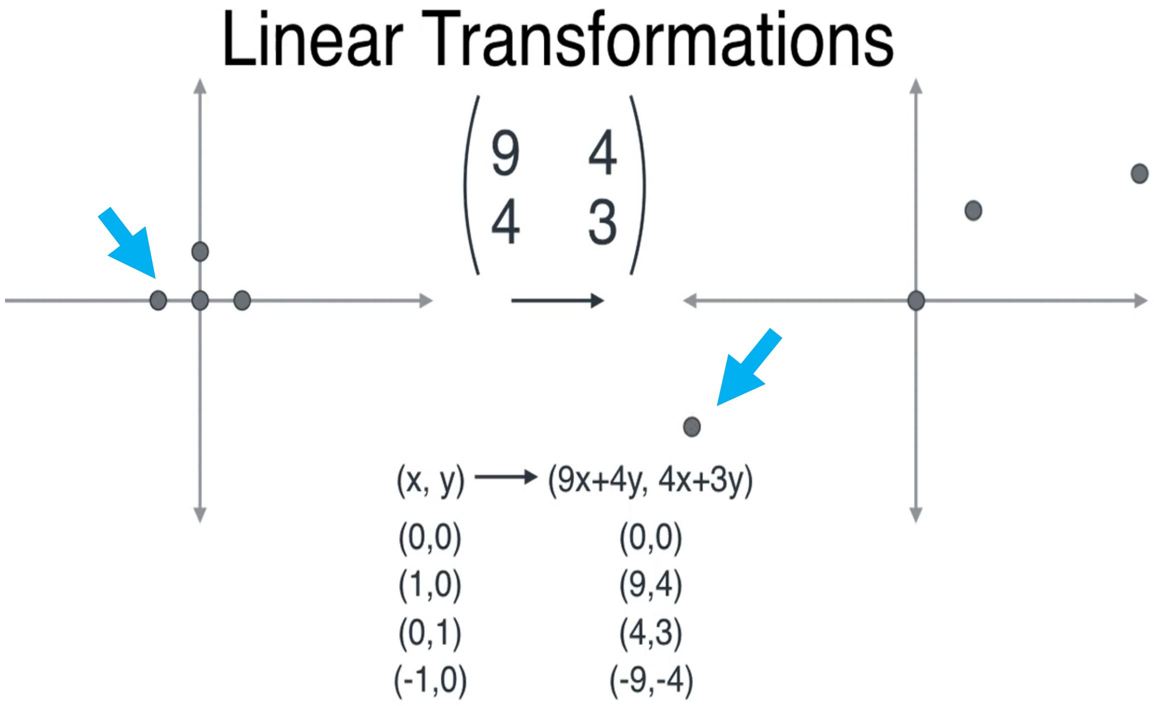


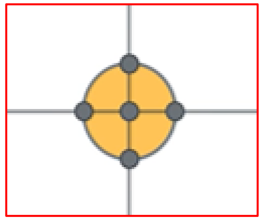
Transformation begins: where (0,1) will go?



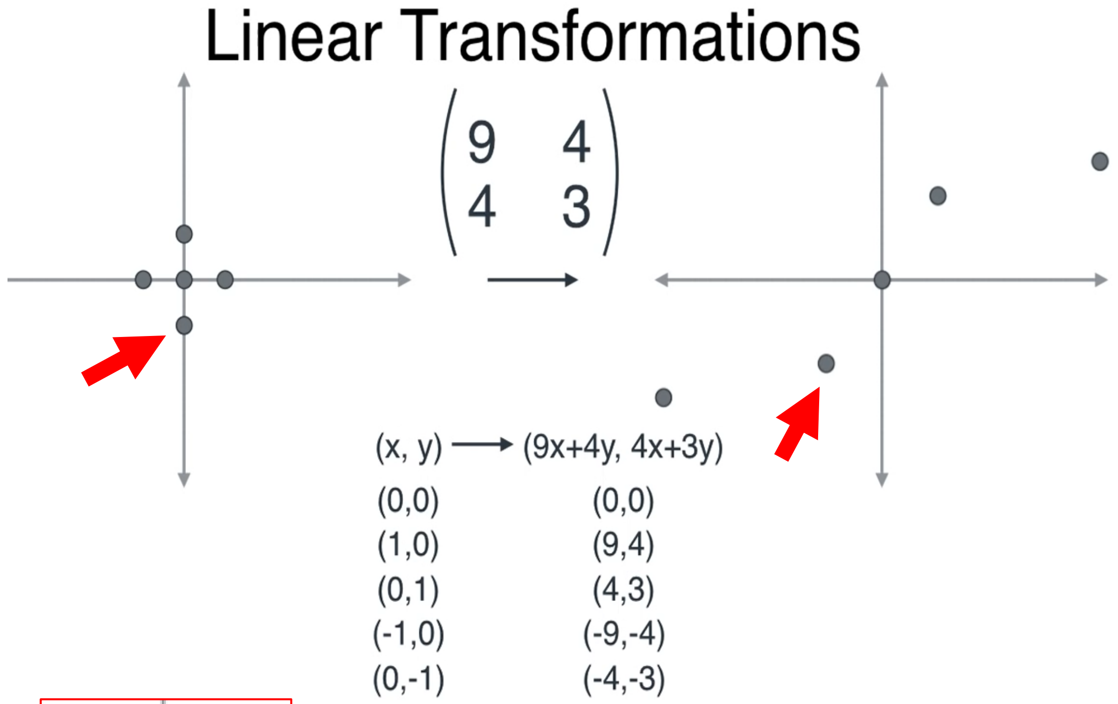


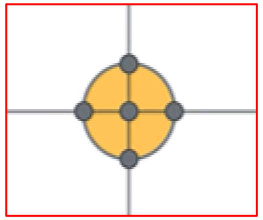
Transformation begins: where (-1,0) will go?

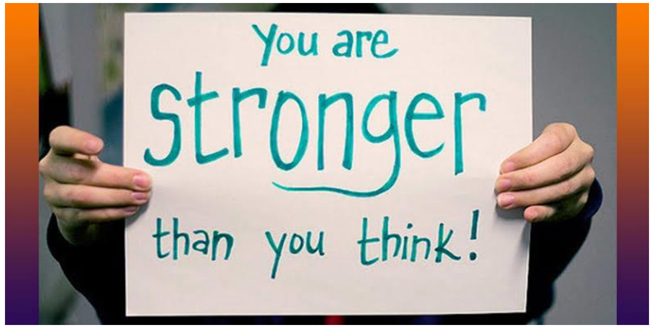




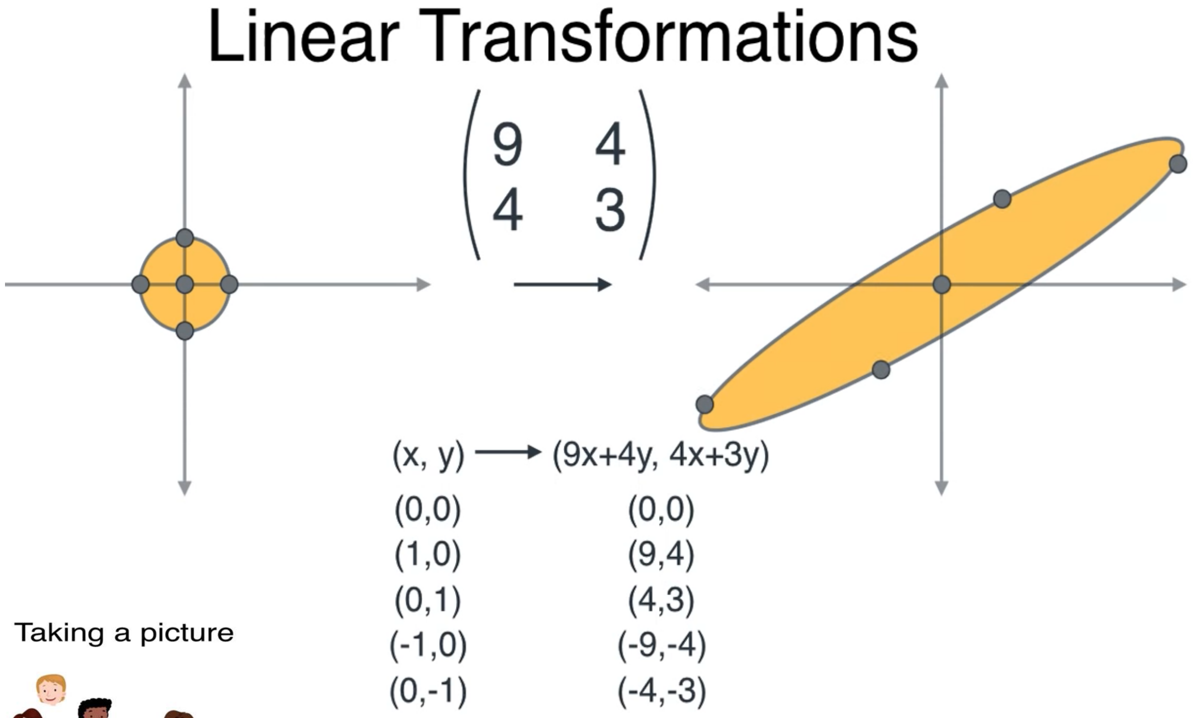
Transformation begins: where (0,-1) will go?





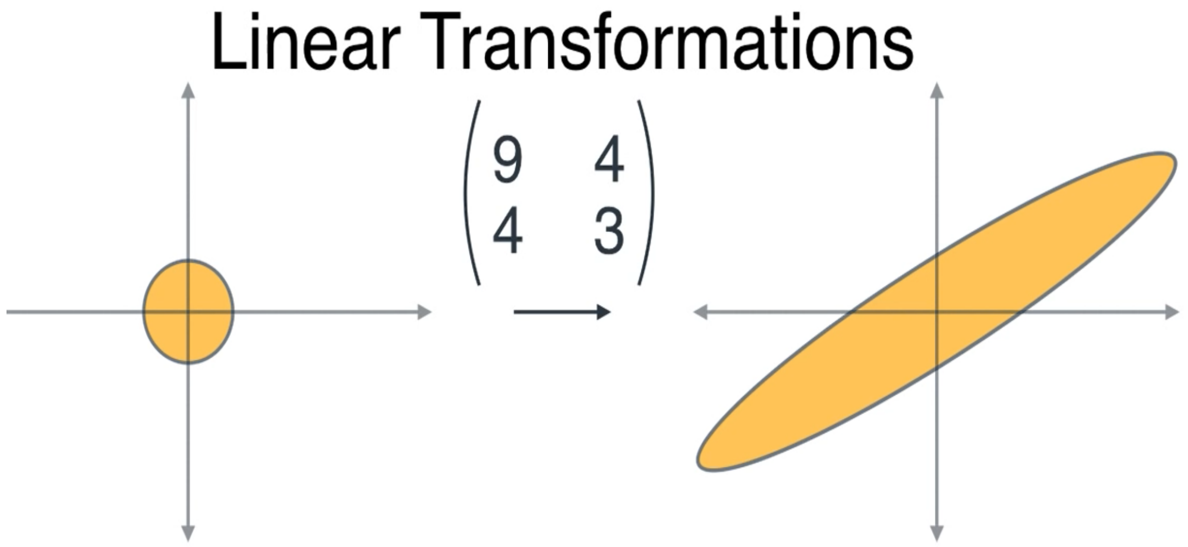


It’s done….wow!

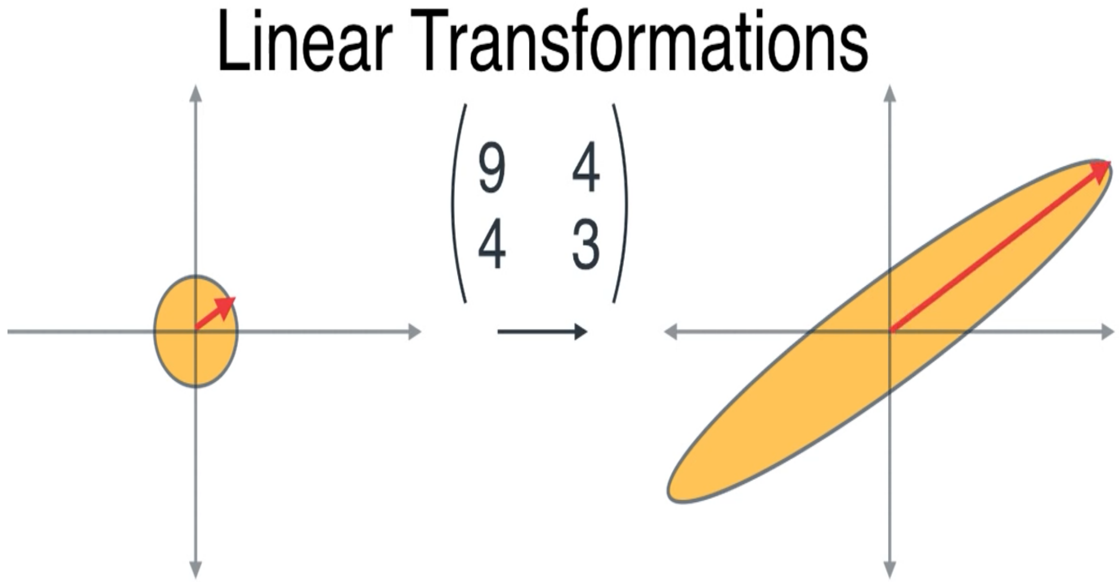


Taking a picture

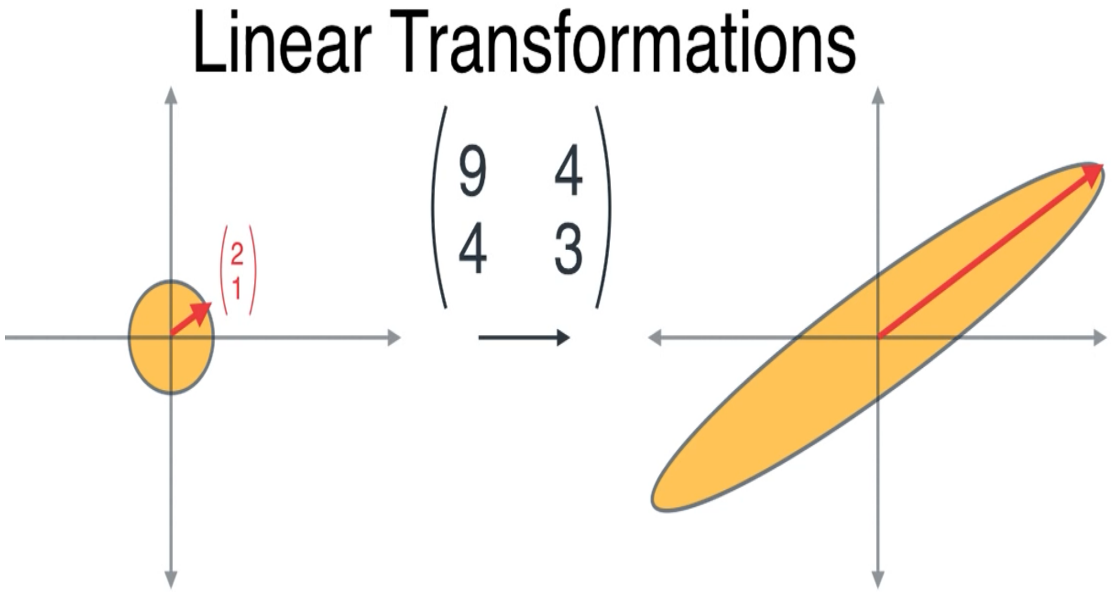
How it had happened?



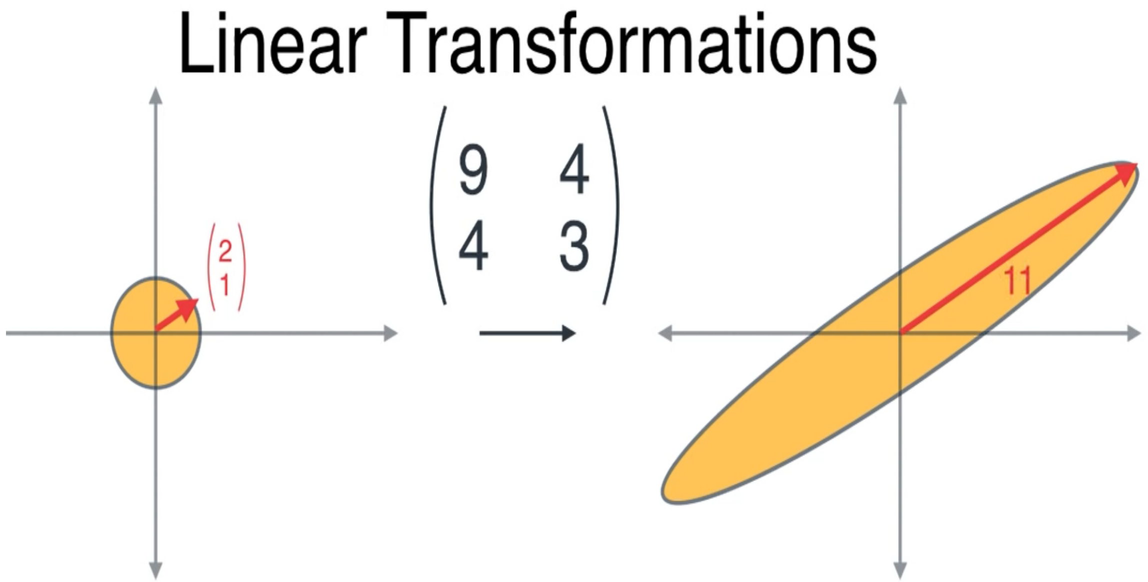
We found a vector (red colored)



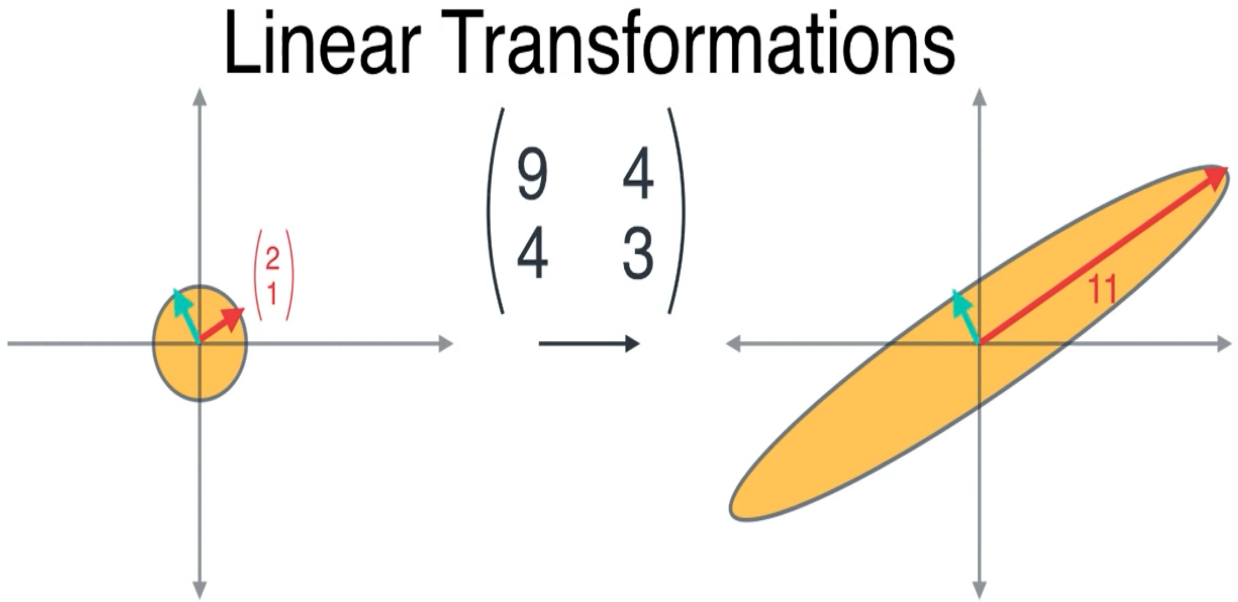
Direction of red colored vector



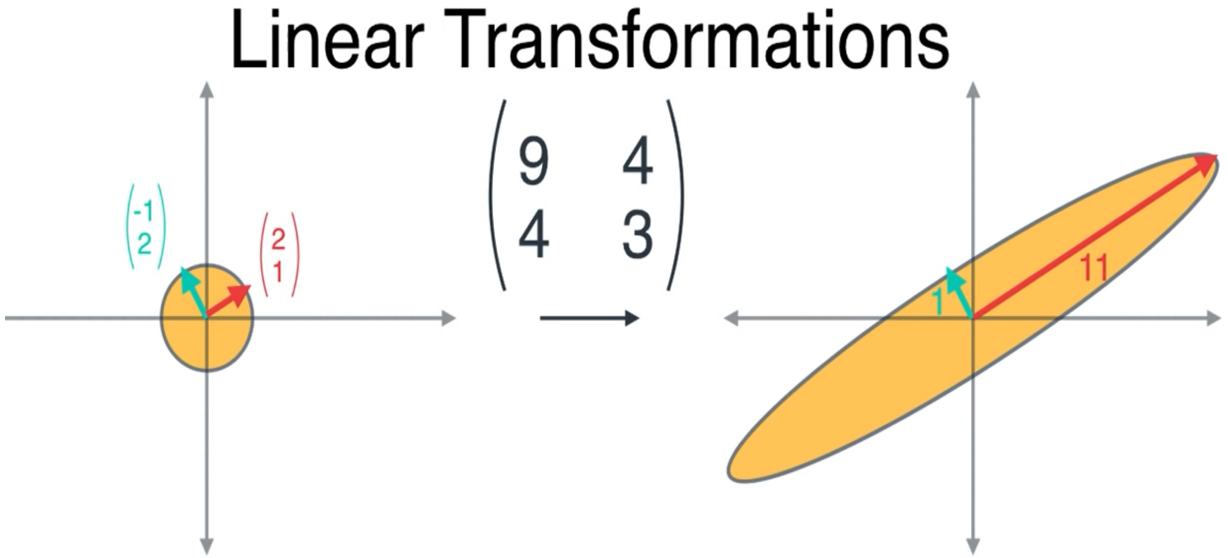
And, found magnitude of that vector



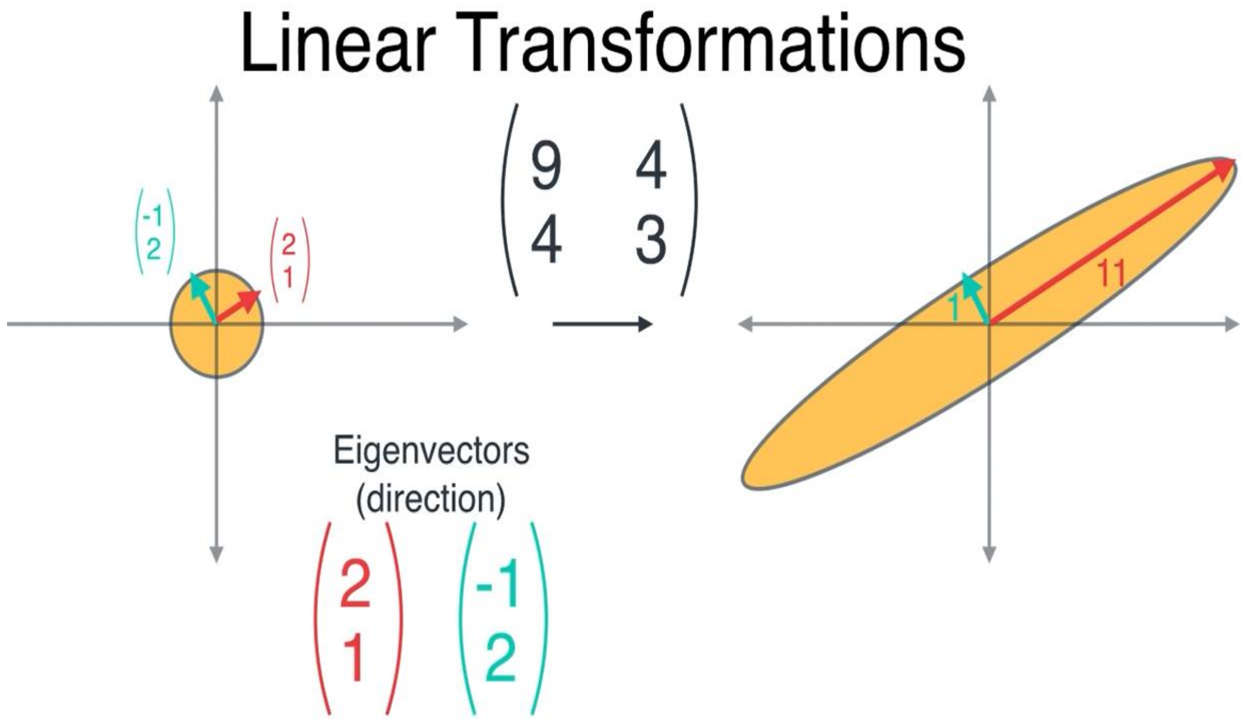
Vector dimension and magnitude



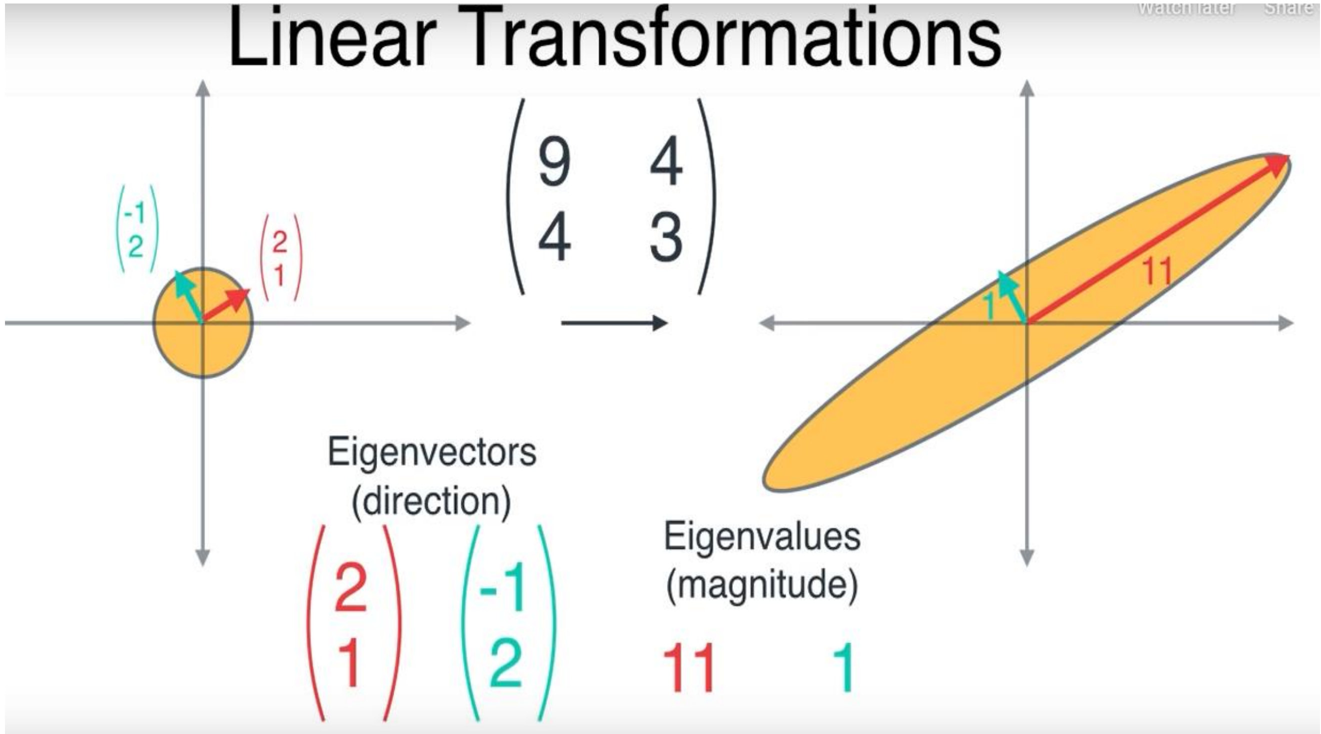
We found dimension of another vector (green) and its magnitude



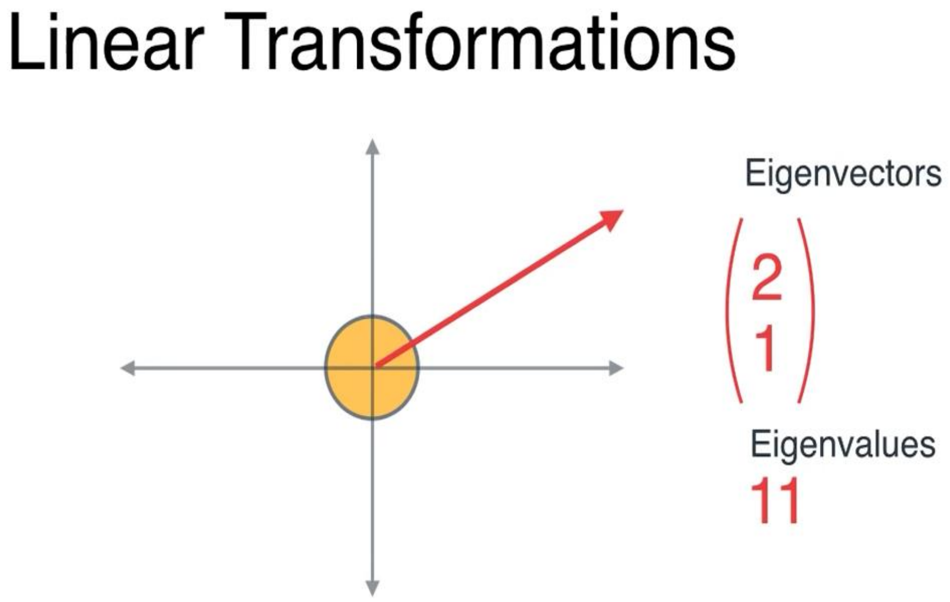
Directions are called eigen vectors



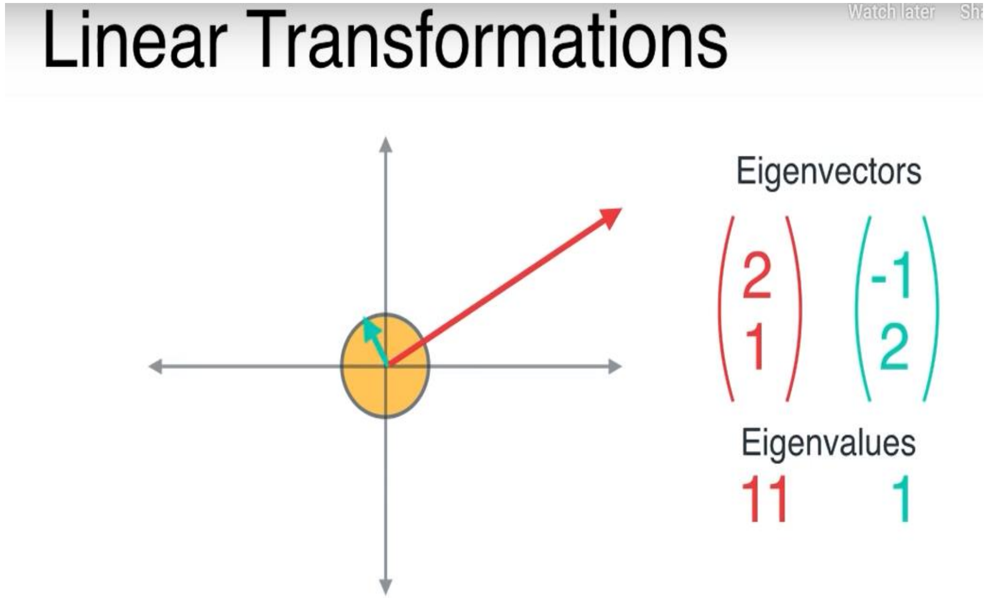
Magnitude is called eigen value



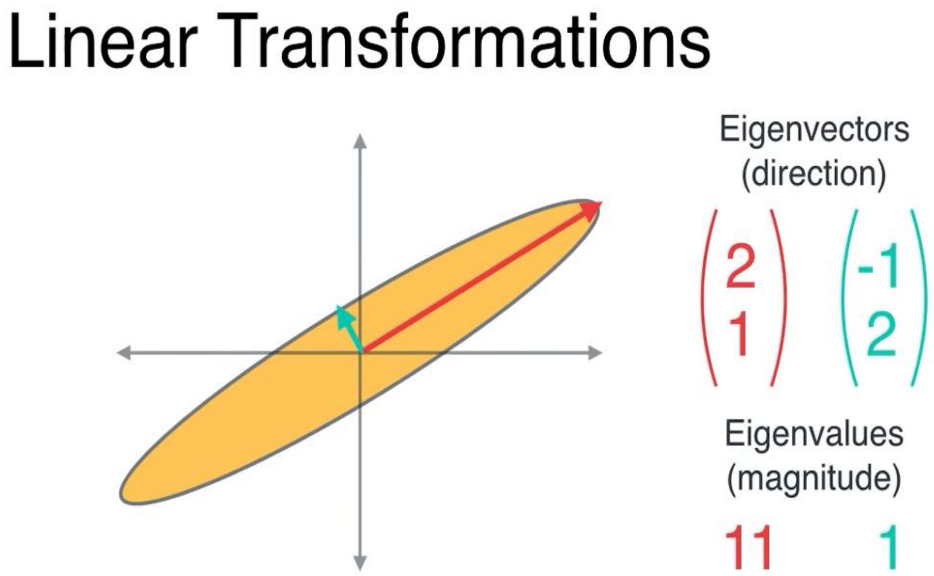
Vector associated with highest eigen value is called Principal Component



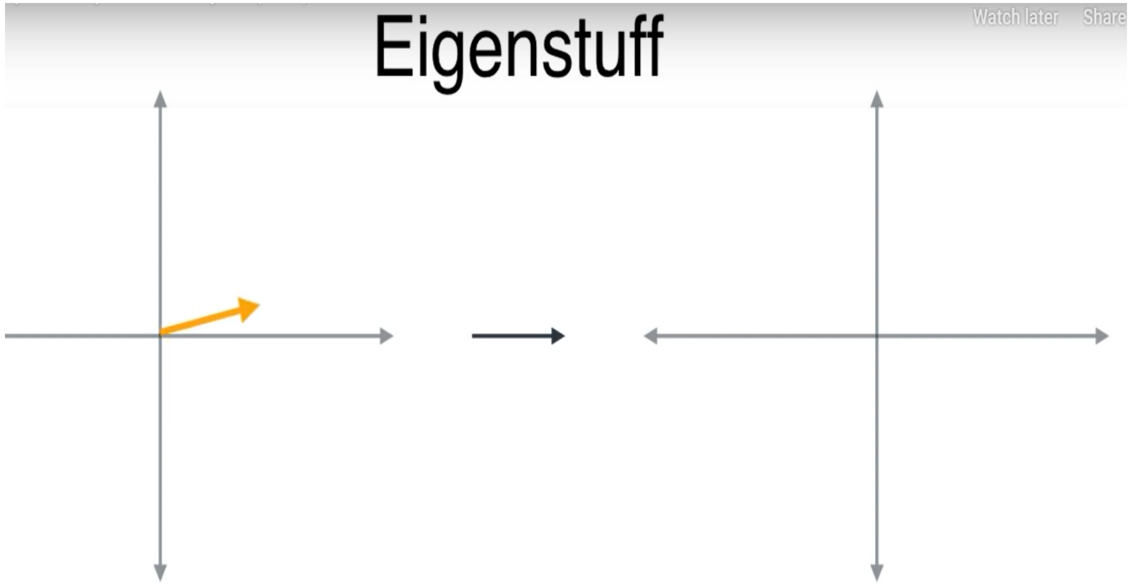
Eigen Values and Eigen Vectors of data



Eigen vectors show direction, values show magnitude

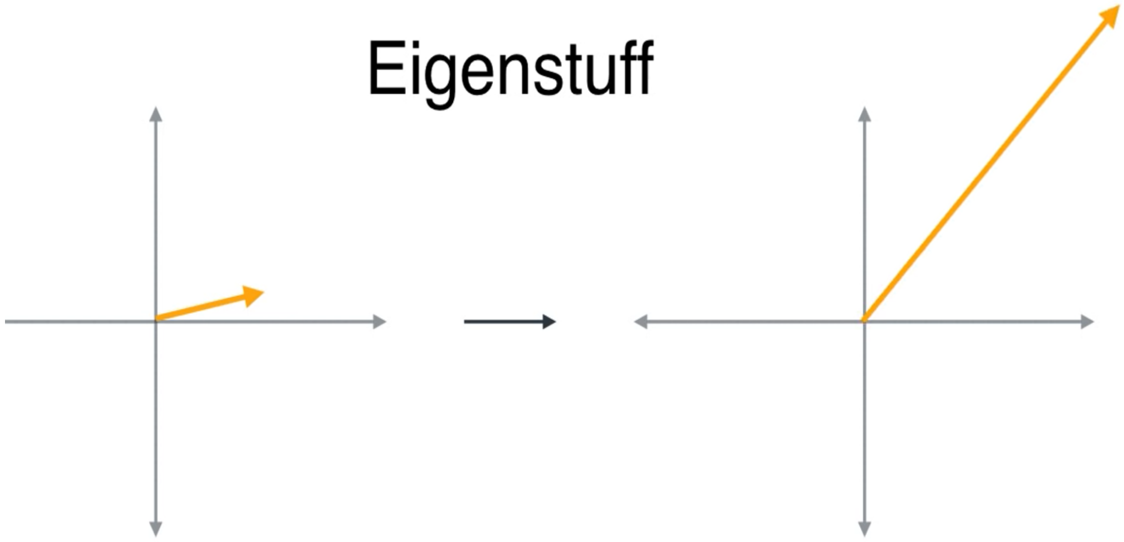


A vector other than eigen vector



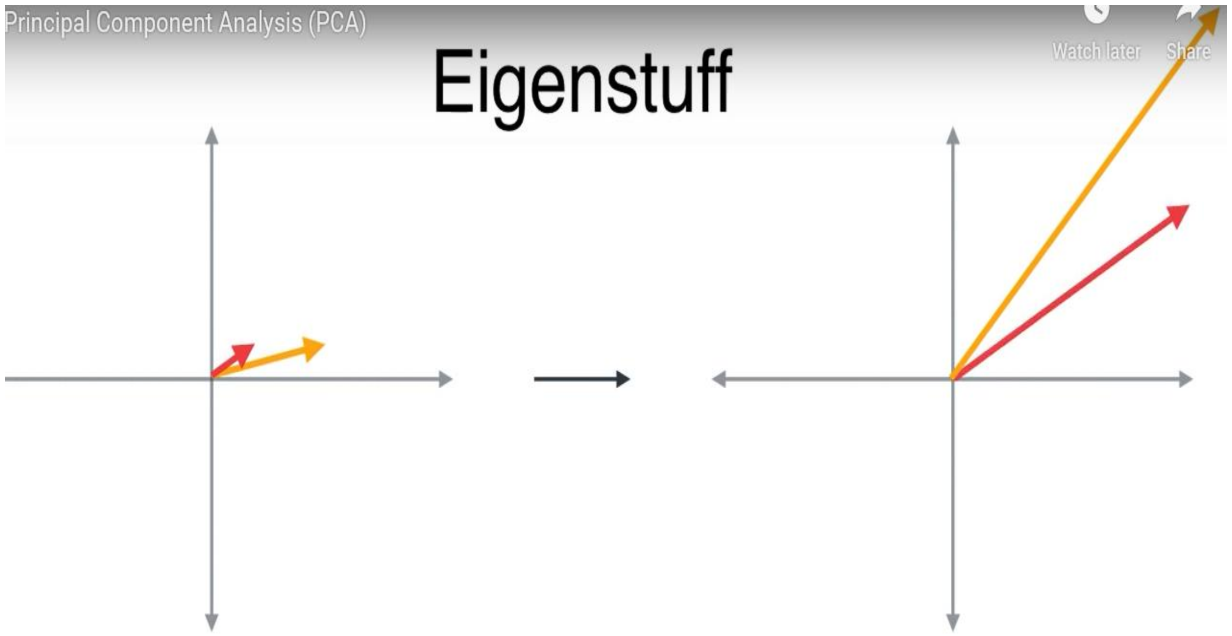
This vector will have changed direction in transformed space

A vector other than eigen vector

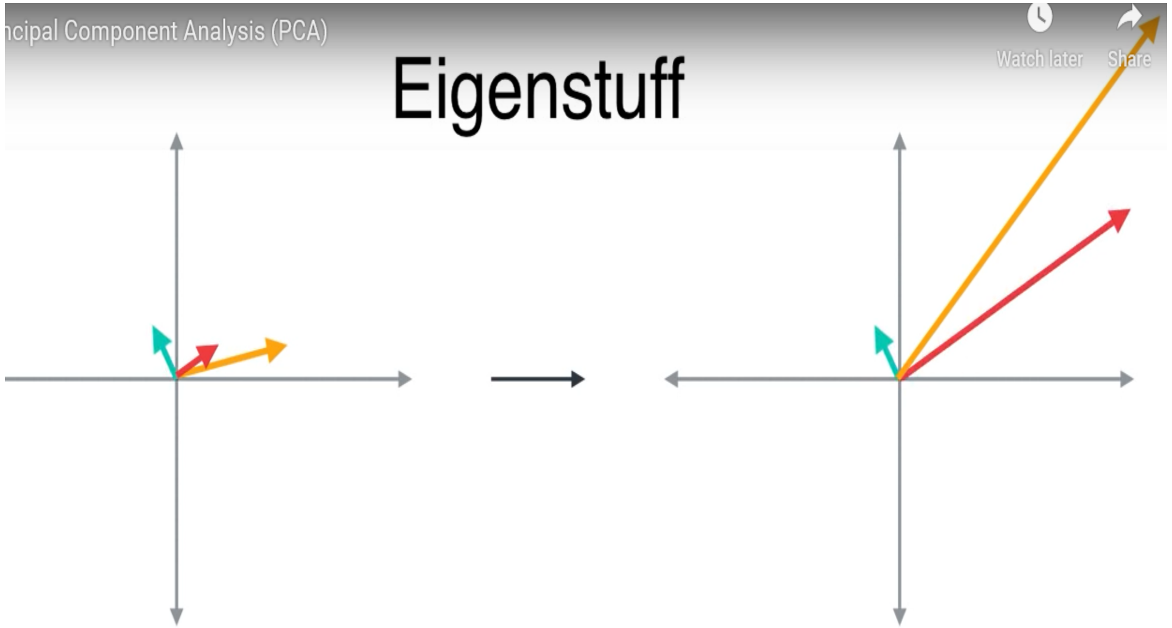


This vector will have changed direction in transformed space

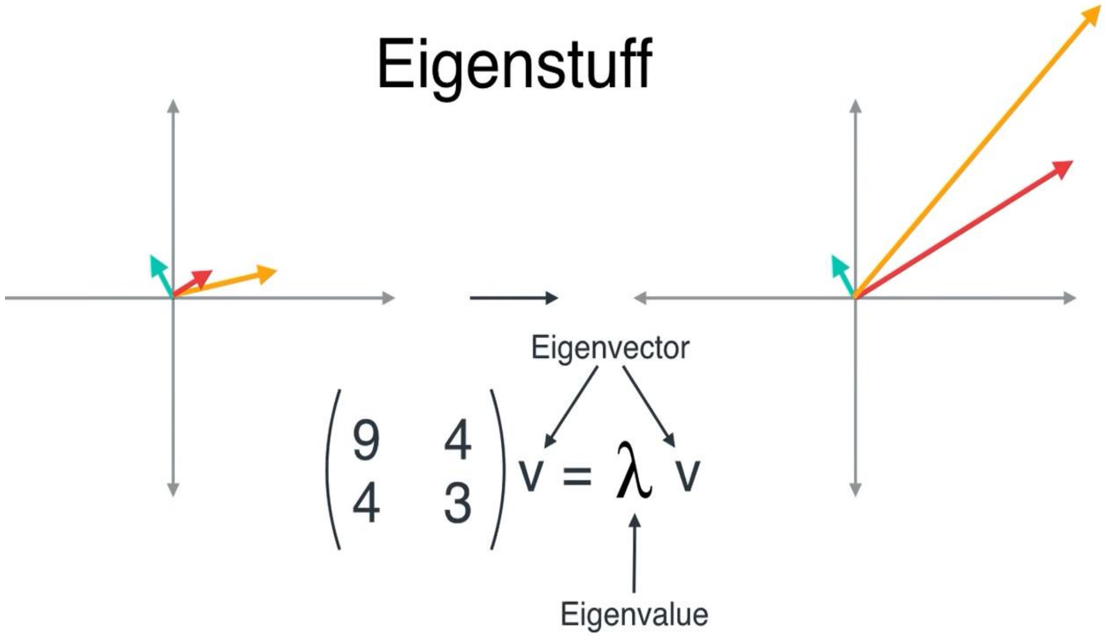
Eigen vector (red) versus any other vector yellow)



Red and green are eigen vectors



The Characteristic Equation



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

Quadratic Equation having two roots

Eigenvalues

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

Characteristic Polynomial

$$\begin{vmatrix} x-9 & -4 \\ -4 & x-3 \end{vmatrix} = (x-9)(x-3) - (-4)(-4) = x^2 - 12x + 11$$

$$= (x-11)(x-1)$$

Eigenvalues 11 and 1

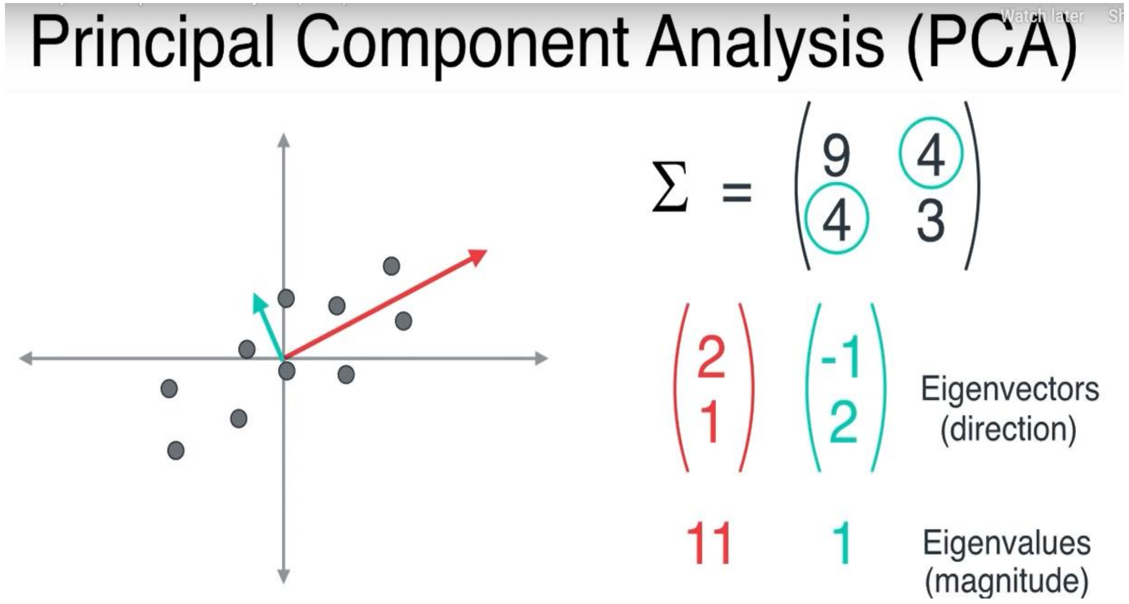
How 2,1 and -1, 2 are calculated?

Eigenvectors

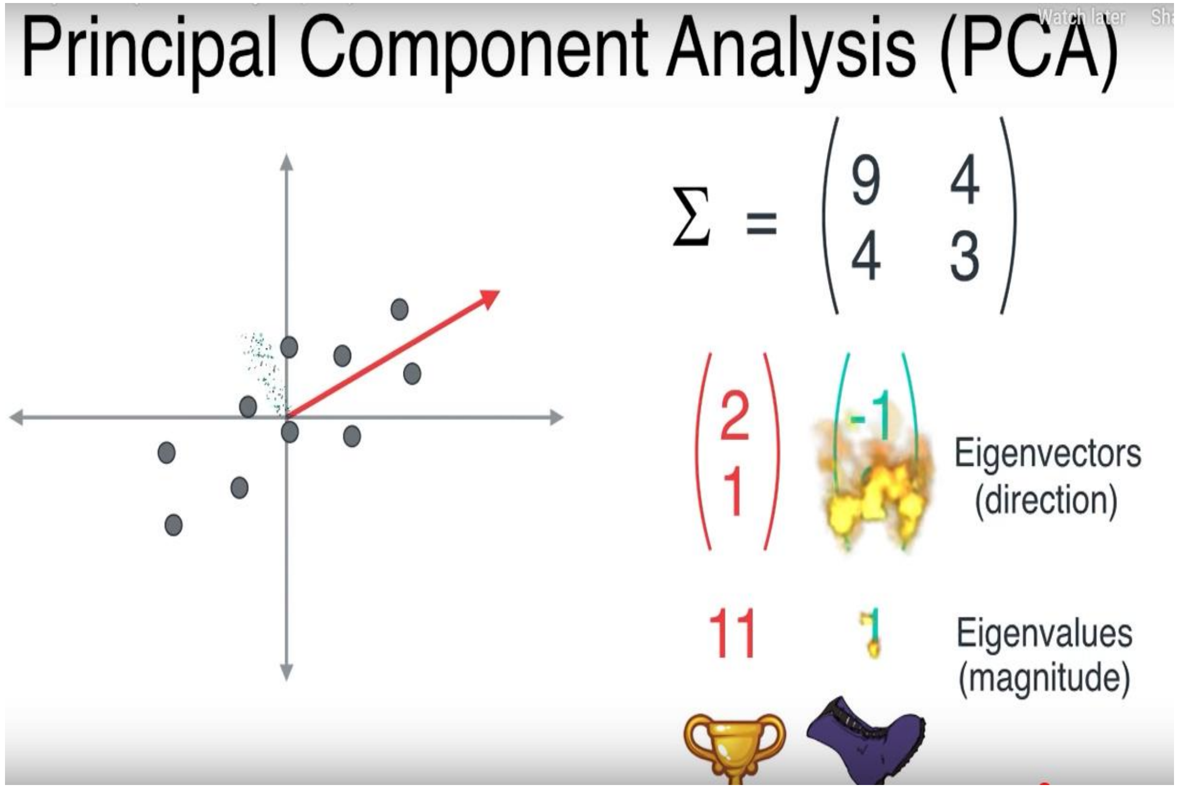
$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 11\begin{pmatrix} u \\ v \end{pmatrix} \qquad \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 1\begin{pmatrix} u \\ v \end{pmatrix}$$

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

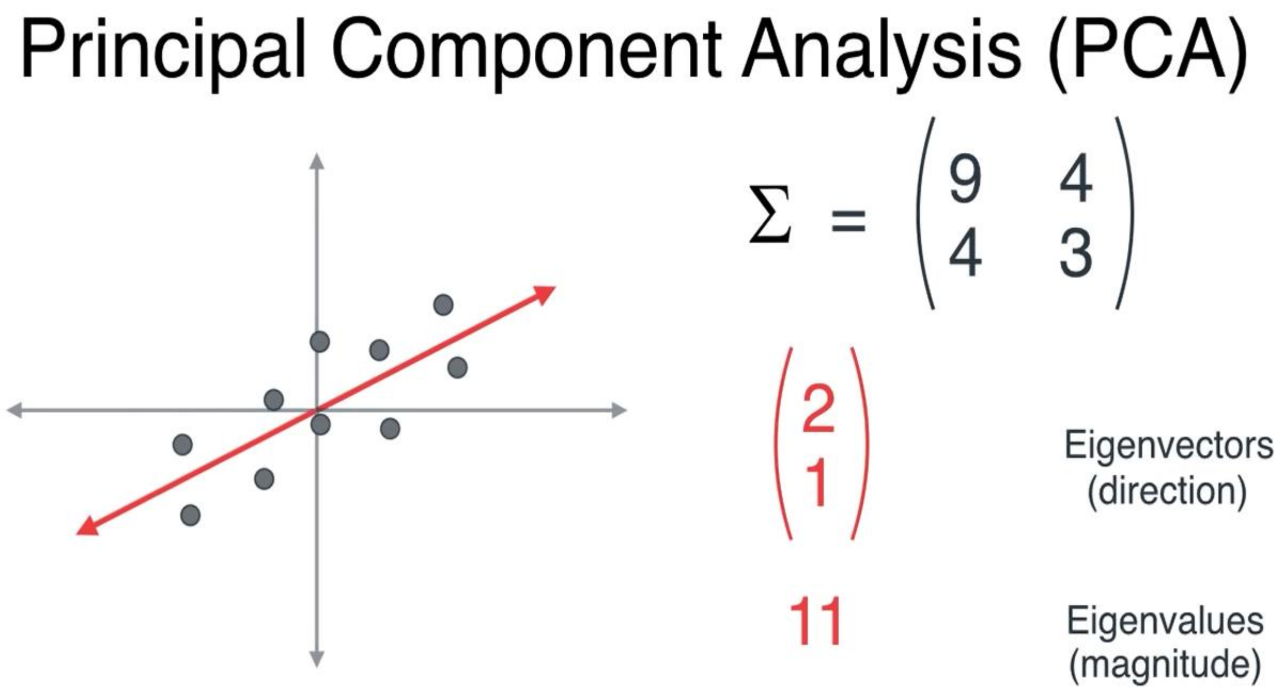
Look at the highest eigen value (11)



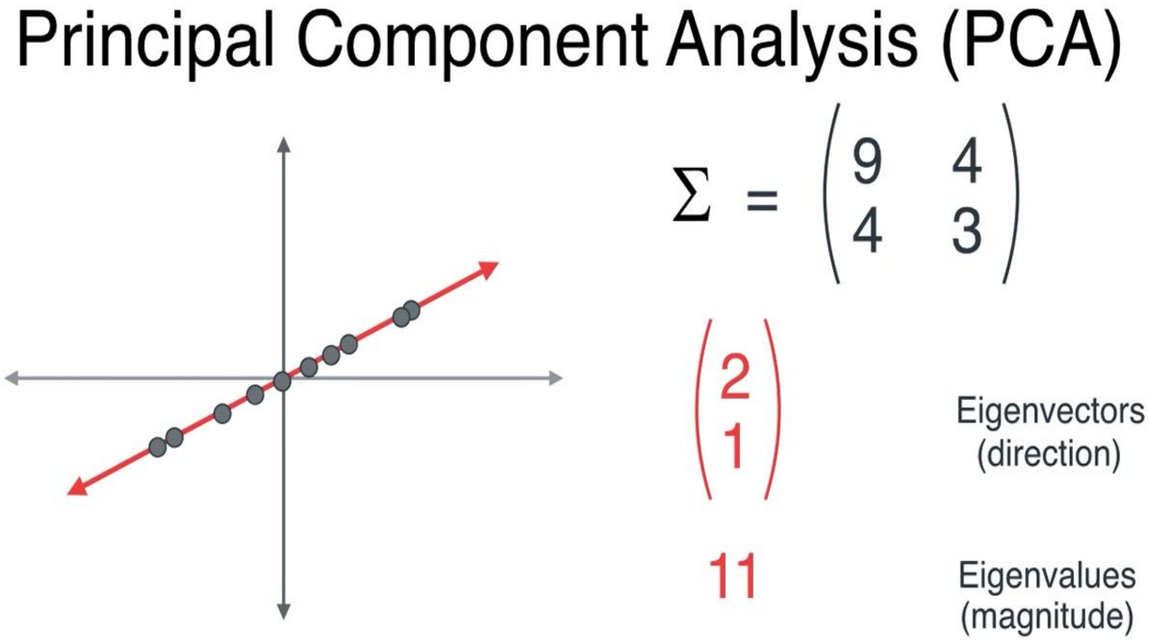
Erase/drop eigen vector associated with lesser eigen value (1, in green)



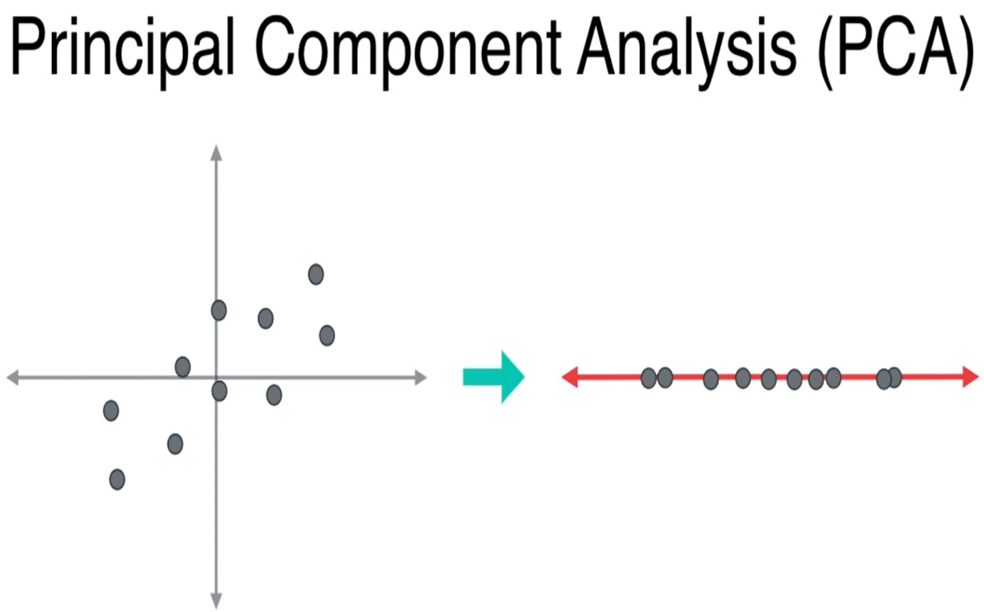
Stretch the Principal Eigen vector so as to cover entire data



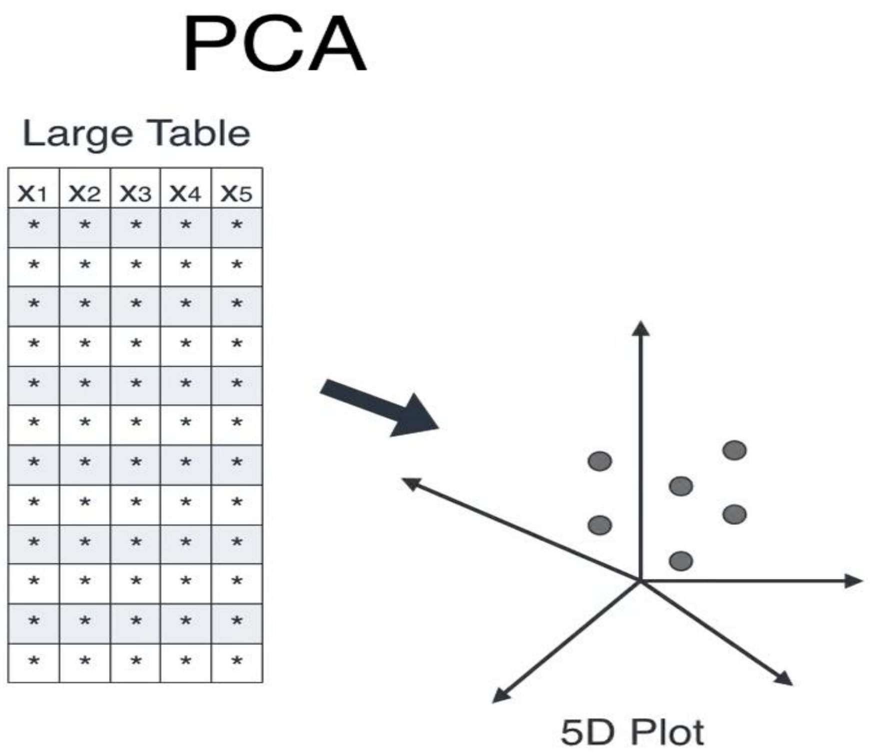
Data points projected on Principal Eigen vector



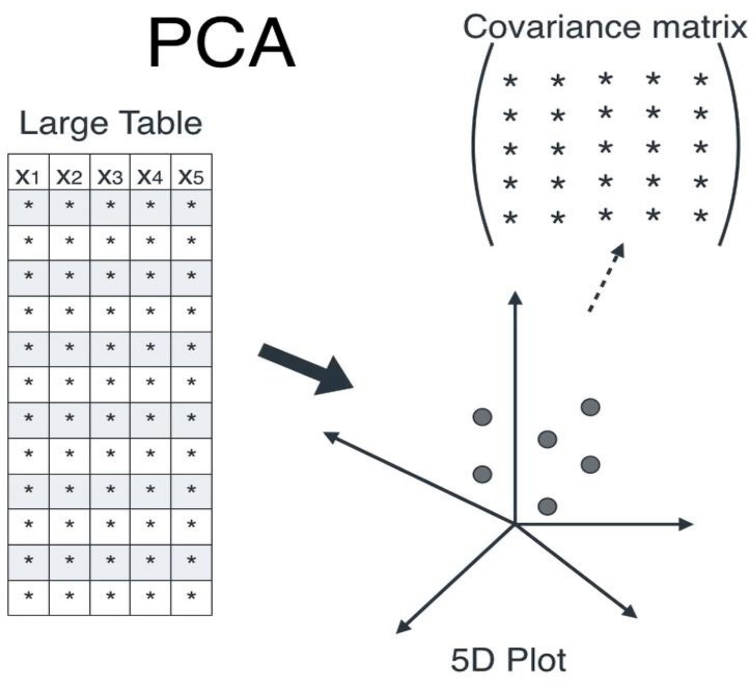
Transformed data



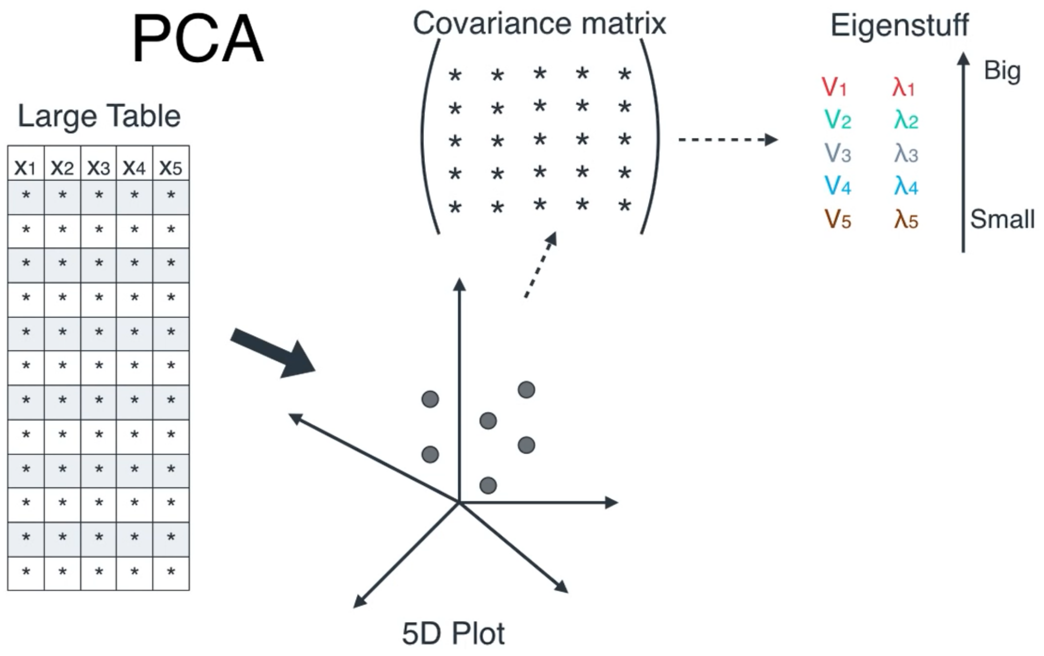
Five columns case



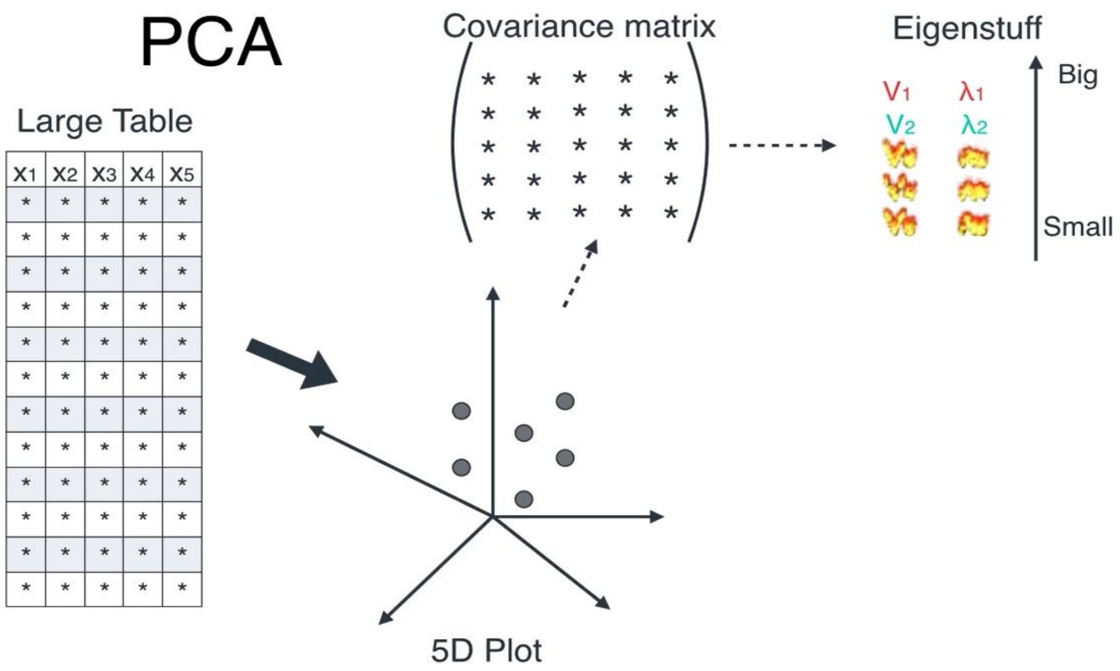
Covariance Matrix



Eigen Values

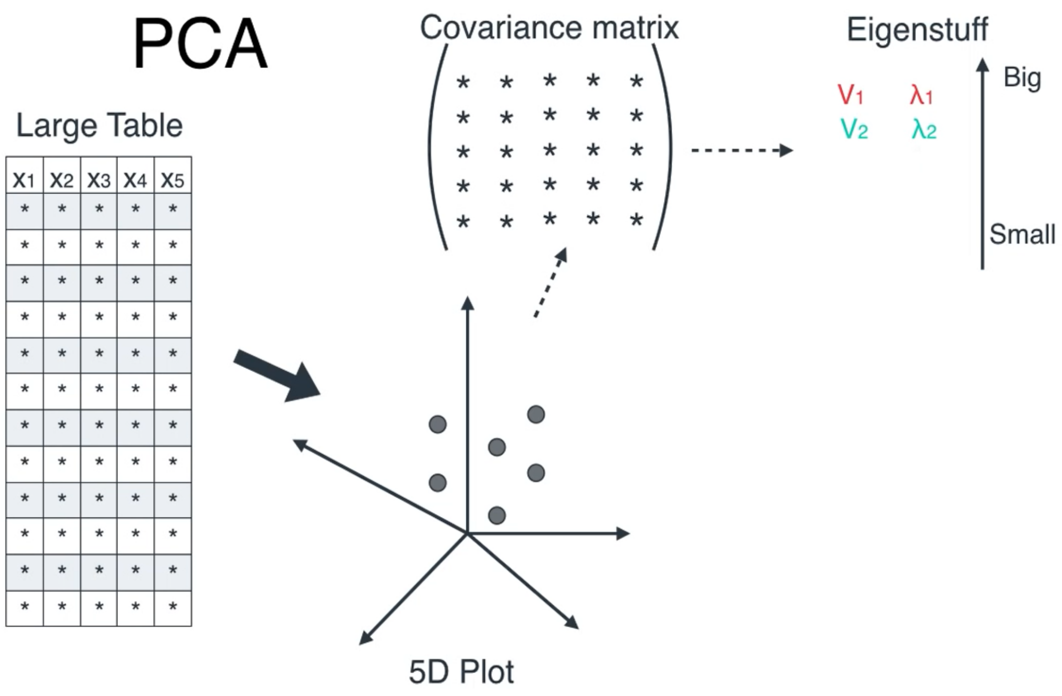


Select Top 2

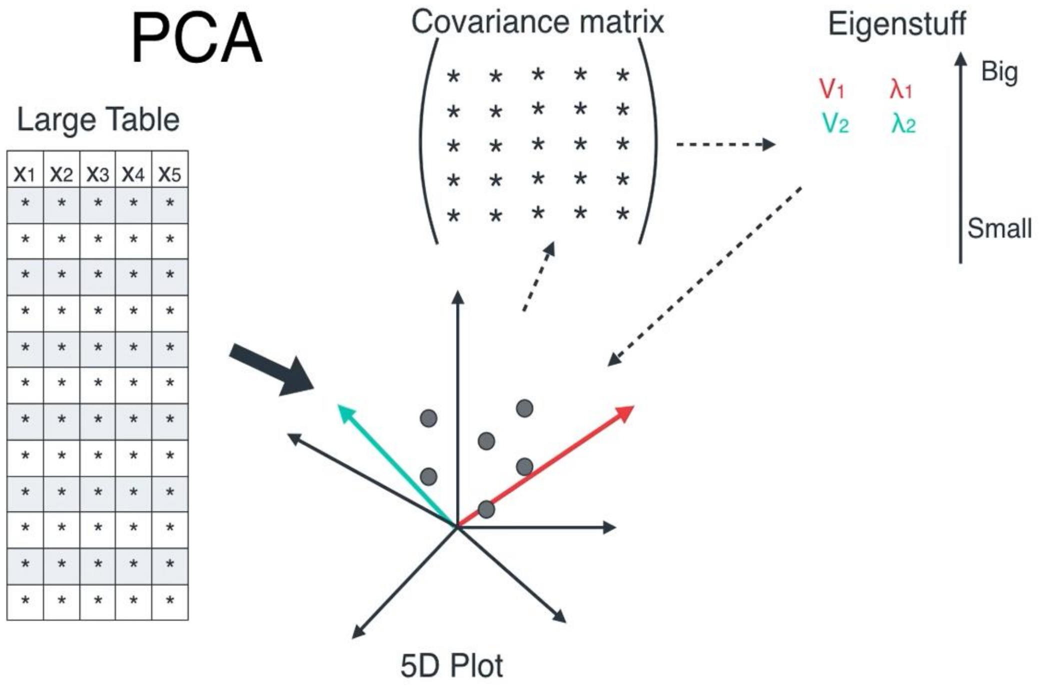


PCA
Large Table
x1 x2 x3 x4 x5
Covariance matrix
Eigenstuff
V1 λ1
V2 λ2
Big
Small
5D Plot

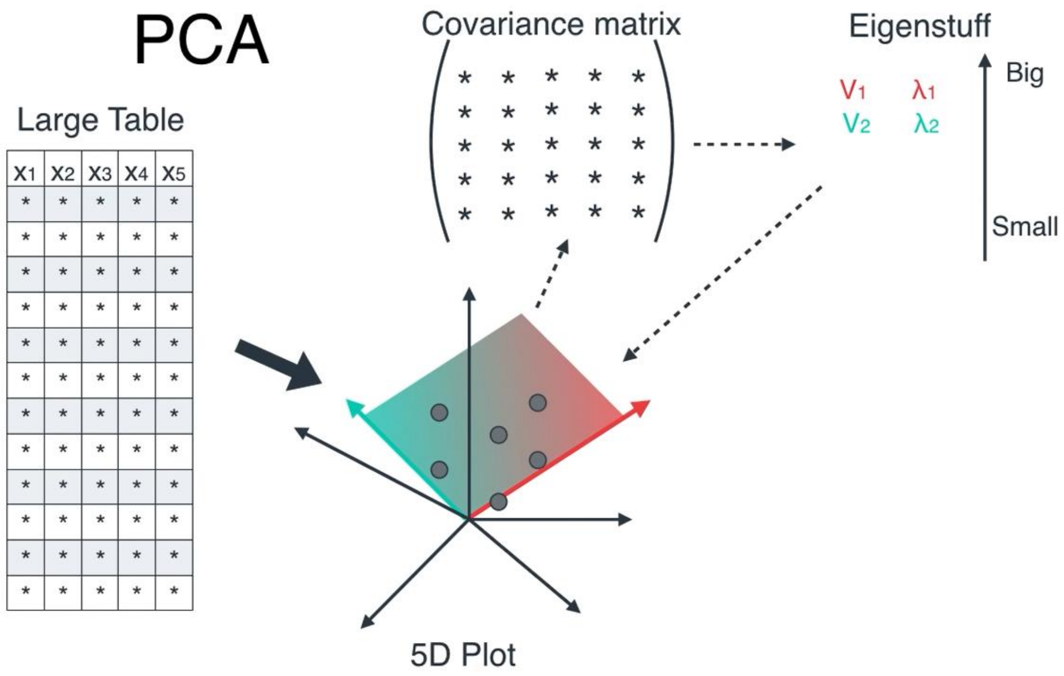
Top 2 Eigen Values only



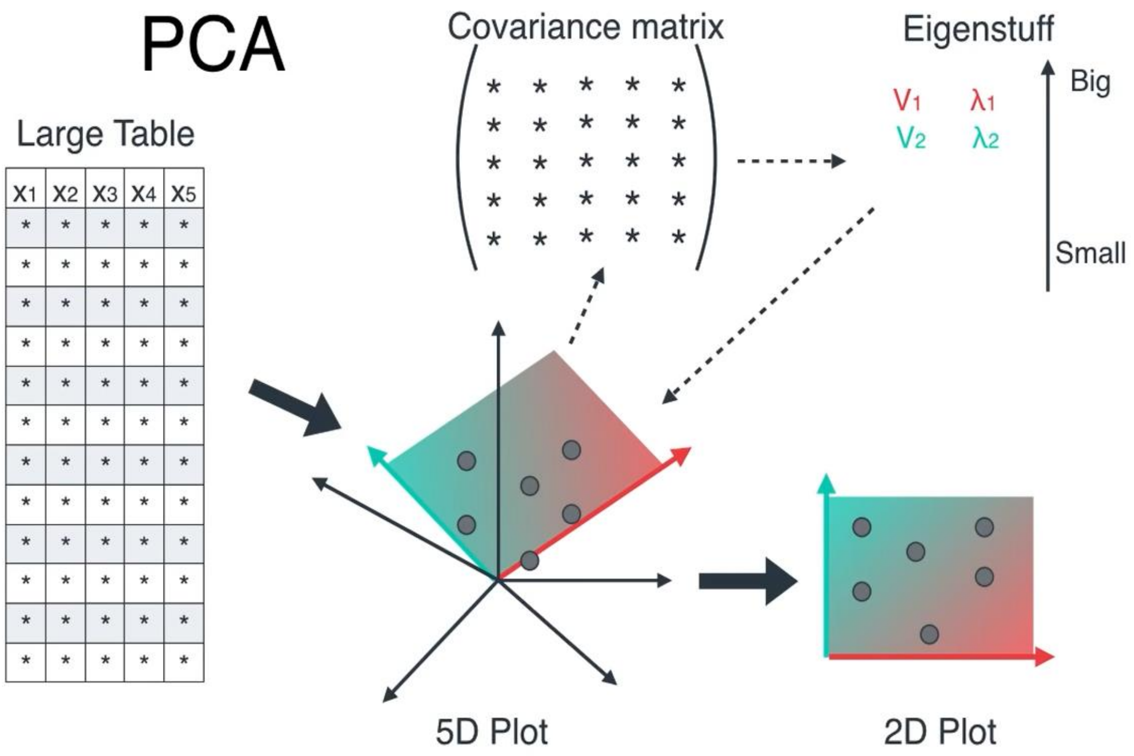
Find 2 Eigen vectors



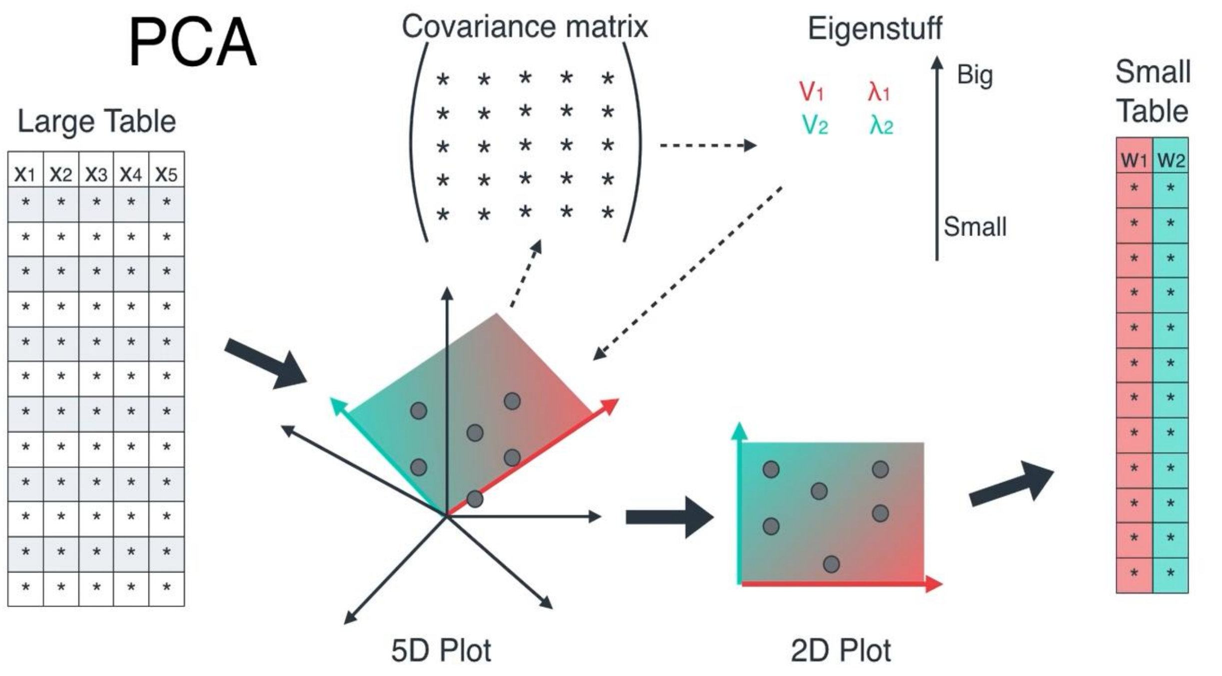
Data Projection



5D to 2D


PCA
Covariance matrix
Eigenstuff
Big
Small
V_1 λ_1
V_2 λ_2
Large Table
X_1 X_2 X_3 X_4 X_5
5D Plot
2D Plot

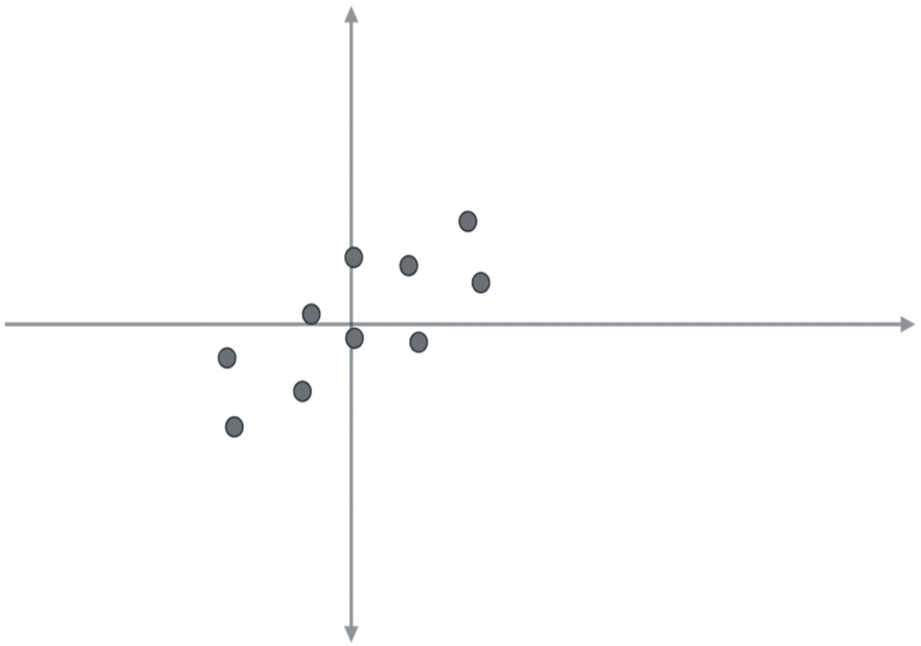
New Data in 2D space



Original Raw Data and Centered Data

A	B	C
	x	y
	2.5	2.4
	0.5	0.7
	2.2	2.9
	1.9	2.2
	3.1	3
	2.3	2.7
	2	1.6
	1	1.1
	1.5	1.6
	1.1	0.9
Mean =	1.81	1.91

D	E
x-mean	y-mean
0.69	0.49
-1.31	-1.21
0.39	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01



In our script/experiment, we had SCALED the raw data

Original Raw Data and Scaled Data

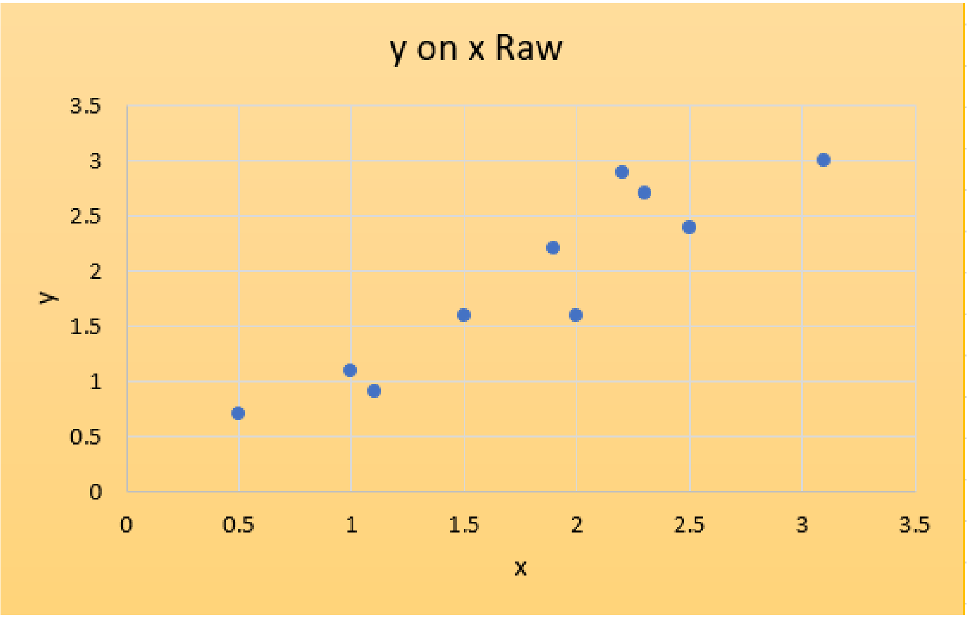
R16

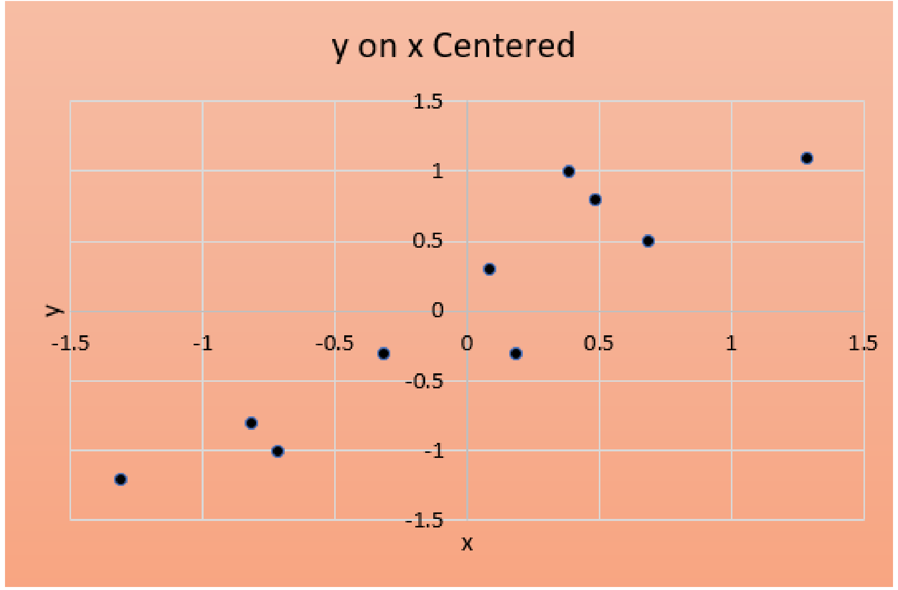
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1			Raw Data					normalize/min-max normalize			scale/standardize			scale sqrt10
2			x	y		(x - mean)^2		x	y		x	y		x
3		1	2.5	2.4		0.4761		0.769231	0.73913		0.878745	0.578857		0.926278808
4		2	0.5	0.7		1.7161		0	0		-1.66834	-1.42942		-1.758587303
5		3	2.2	2.9		0.1521		0.653846	0.956522		0.496682	1.169527		0.523548892
6		4	1.9	2.2		0.0081		0.538462	0.652174		0.114619	0.342589		0.120818975
7		5	3.1	3		1.6641		1	1		1.642872	1.287661		1.731738642
8		6	2.3	2.7		0.2401		0.692308	0.869565		0.624036	0.933259		0.657792197
9		7	2	1.6		0.0361		0.576923	0.391304		0.241973	-0.36622		0.255062281
10		8	1	1.1		0.6561		0.192308	0.173913		-1.03157	-0.95689		-1.087370775
11		9	1.5	1.6		0.0961		0.384615	0.391304		-0.3948	-0.36622		-0.416154247
12		10	1.1	0.9		0.5041		0.230769	0.086957		-0.90422	-1.19315		-0.95312747
13		mean	1.81	1.91	sum	**5.549**								
14		sd	0.785211	0.846496	div10	0.5549								
15		min	0.5	0.7	div9	0.616555556								
16		max	3.1	3	sqrt10	0.744916103								
17		range	2.6	2.3	**sqrt9**	**0.785210517**								

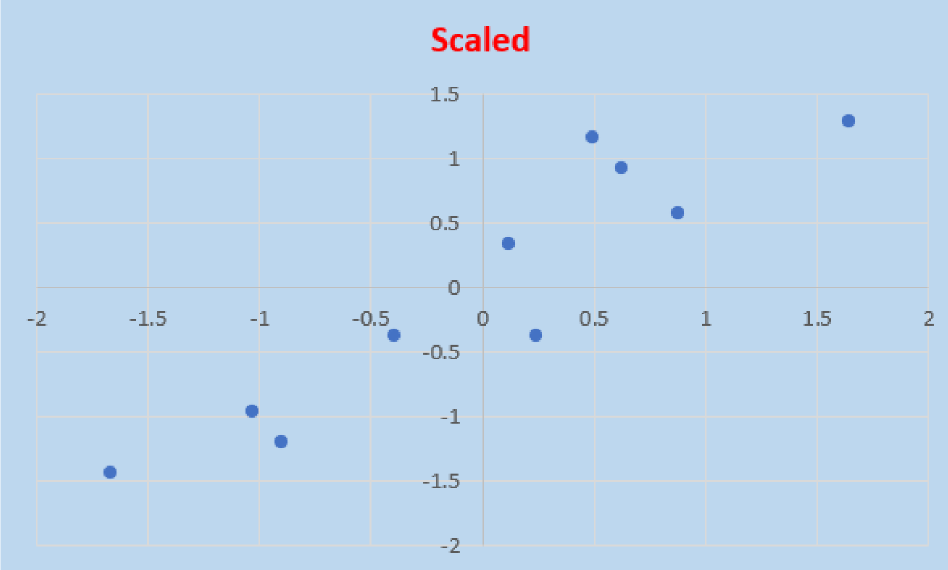
	x	y
count	10.000000	10.000000
mean	1.810000	1.910000
std	0.785211	0.846496
min	0.500000	0.700000
25%	1.200000	1.225000
50%	1.950000	1.900000
75%	2.275000	2.625000
max	3.100000	3.000000

	0	1
0	0.926279	0.610169
1	-1.75859	-1.50674
2	0.523549	1.23279
3	0.120819	0.36112
4	1.73174	1.35731
5	0.657792	0.983741
6	0.255062	-0.386025
7	-1.08737	-1.00865
8	-0.416154	-0.386025
9	-0.953127	-1.25769

Raw, Centered & Scaled







Covariance Matrix

F	G	H	I	J	K
	(x-mean)(y-mean)		(x-mean)(x-mean)		(y-mean)(y-mean)
	0.3381		0.4761		0.2401
	1.5851		1.7161		1.4641
	0.3861		0.1521		0.9801
	0.0261		0.0081		0.0841
	1.4061		1.6641		1.1881
	0.3871		0.2401		0.6241
	-0.0589		0.0361		0.0961
	0.6561		0.6561		0.6561
	0.0961		0.0961		0.0961
	0.7171		0.5041		1.0201
	5.539		5.549		6.449
cov(x,y)	**0.615444444**	**cov(x,x)**	**0.616555556**	**cov(y,y)**	**0.716555556**
			Cov Matrix		
				x	**y**
			x	**0.61656**	**0.615444444**
			y	**0.61544**	**0.716555556**

Eigen Values and Eigen Vectors

$eigenvalues =$

explained_variance_ - NumPy object array

	0
0	2.13992
1	0.0823008

$eigenvectors =$

pca2_egvct - NumPy object array

	0	1
0	-0.707107	-0.707107
1	-0.707107	0.707107

Eigen vector 1

Eigen vector 2

cov(x,y)	0.615444444	cov(x,x)	0.616555556	cov(y,y)	0.716555556
			Cov Matrix		
				x	y
			x	0.61656	0.615444444
			y	0.61544	0.716555556

Transformed Data

T19

	A	B	C	D	E	F	G	H	I	J	K	L	M	Q	R	S
1		Raw Data						from python		eigvalues-->	2.13992	0.082301	2.222221			
2		x	y		x_scale	y_scale		x_trans	y_trans		eignV 1	eignV 2			manual multiplication	
3	1	2.5	2.4		0.93	0.61	1	-1.09	-0.22		-0.707107	-0.70711		1	-1.08643	-0.22352
4	2	0.5	0.7		-1.76	-1.51	2	2.31	0.18		-0.707107	0.707107		2	2.308938	0.178081
5	3	2.2	2.9		0.52	1.23	3	-1.24	0.50					3	-1.24192	0.501509
6	4	1.9	2.2		0.12	0.36	4	-0.34	0.17					4	-0.34078	0.169919
7	5	3.1	3		1.73	1.36	5	-2.18	-0.26					5	-2.18429	-0.26476
8	6	2.3	2.7		0.66	0.98	6	-1.16	0.23					6	-1.16074	0.230481
9	7	2	1.6		0.26	-0.39	7	0.09	-0.45					7	0.092605	-0.45332
10	8	1	1.1		-1.09	-1.01	8	1.48	0.06					8	1.482108	0.055667
11	9	1.5	1.6		-0.42	-0.39	9	0.57	0.02					9	0.567227	0.021305
12	10	1.1	0.9		-0.95	-1.26	10	1.56	-0.22					10	1.563288	-0.21536
13				mean	0.00	0.00					0.9629646	0.03704				
14										cum sum=	0.963	1				

pca2_egvct - NumPy object array

	0	1
0	-0.707107	-0.707107
1	-0.707107	0.707107

explained_variance_ - NumPy object array

	0
0	2.13992
1	0.0823008

X1 - NumPy object array

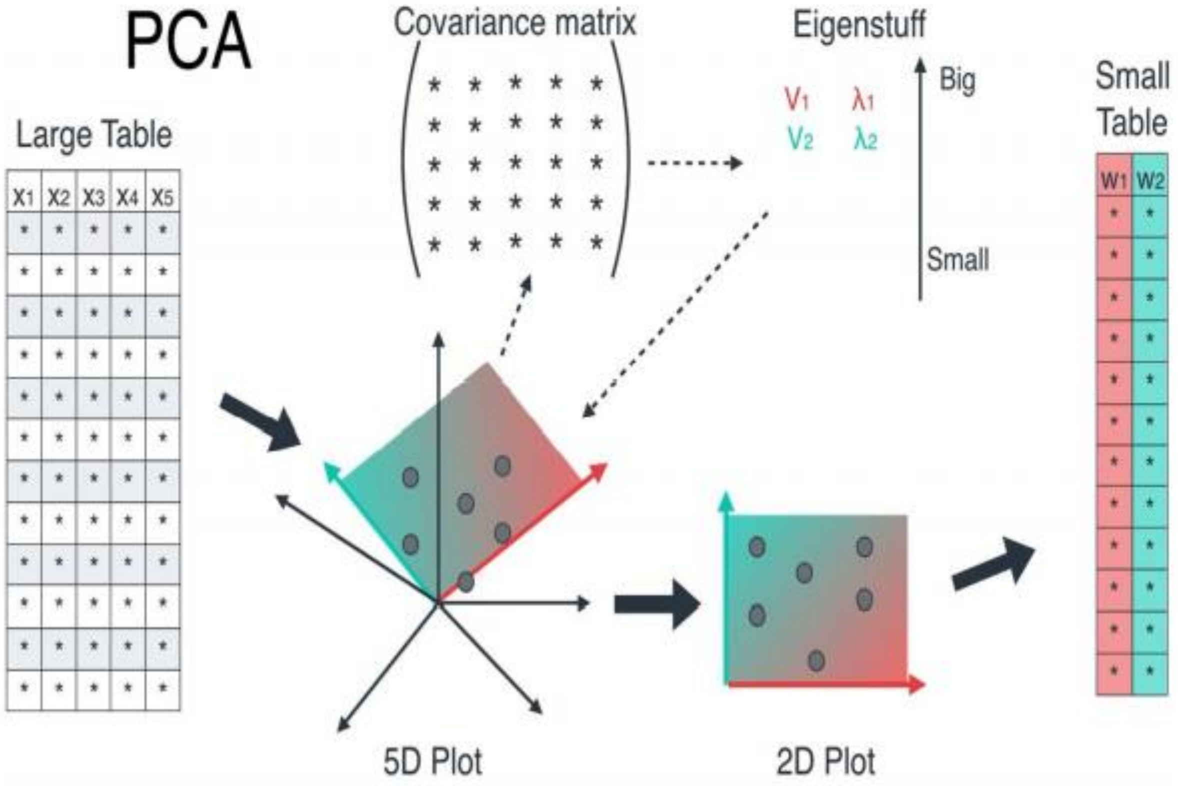
	0	1
0	-1.08643	-0.223524
1	2.30894	0.178081
2	-1.24192	0.501509
3	-0.340782	0.169919
4	-2.18429	-0.264758
5	-1.16074	0.230481
6	0.0926047	-0.453317
7	1.48211	0.0556667
8	0.567226	0.0213045
9	1.56329	-0.215361

```
In [10]: var
Out[10]: array([0.96296464, 0.03703536])
```

```
In [11]: var1=np.cumsum(np.round(pca.explained_variance_ratio_, decimals=4)*100)

In [12]: var1
Out[12]: array([ 96.3, 100. ])
```

Mathematical Treatment



Step 1: Centering

Step 2: Covariance Matrix

Step 3: Use Characteristic Equation det of ($A - \lambda I = 0$) for setting up equation for unknown λ (eigen values)

Step 4: Find eigen vectors

Step 5: Transform centered data by Matrix Multiplication. [M1 = centered data (10by3); M2 = Eigen Vectors (3by3)

Eigen Values

$$A = \begin{bmatrix} 8 & -6 & 2 \\ -6 & 7 & -4 \\ 2 & -4 & 3 \end{bmatrix}$$

characteristic eqn:

$$\det.[A - \lambda I] = 0$$

First find $[A - \lambda I]$

$$\begin{bmatrix} 8 & -6 & 2 \\ -6 & 7 & -4 \\ 2 & -4 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} (8-\lambda) & -6 & 2 \\ -6 & (7-\lambda) & -4 \\ 2 & -4 & (3-\lambda) \end{bmatrix}$$

det. of above: (i) follow $+, -, +$
(ii) start from NW corner (or can go down ☺)
(iii) hide that row & column

$$+ (8-\lambda)[(7-\lambda)\times(3-\lambda) - (-4\times-4)] - (-6)[(-6)\times(3-\lambda) - (-4\times 2)] + 2[(-6)\times(-4) - 2\times(7-\lambda)]$$

$$\boxed{\lambda^3 - 18\lambda^2 + 45\lambda = 0}$$

$\lambda = 0, \lambda = 3, \lambda = 15$

Eigen Vectors

This is same as λI

For $\lambda = 0$

$$[A - \lambda][x] = 0$$

$$\begin{bmatrix} (8-\lambda) & -6 & 2 \\ -6 & (7-\lambda) & -4 \\ 2 & -4 & 3-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Keeping $\lambda = 0$

$$\begin{bmatrix} 8 & -6 & 2 \\ -6 & 7 & -4 \\ 2 & -4 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$8x - 6y + 2z = 0 \quad -①$$
$$-6x + 7y - 4z = 0 \quad -②$$
$$2x - 4y + 3z = 0 \quad -③$$

Let's solve ① & ② You can take any two ☺
($+, -, +$ applicable)

$$+\frac{x}{(-6)\times(-4) - (7\times 2)} - \frac{y}{8\times(-4) - (-6)\times 2} + \frac{z}{8\times 7 - [-6\times(-6)]}$$

$$\frac{x}{24 - 14} - \frac{y}{-32 - (-12)} + \frac{z}{56 - 36}$$

$$\frac{x}{10} - \frac{y}{-20} + \frac{z}{20}$$

Divide by 10 $\quad \frac{x}{1} + \frac{y}{2} + \frac{z}{2}$

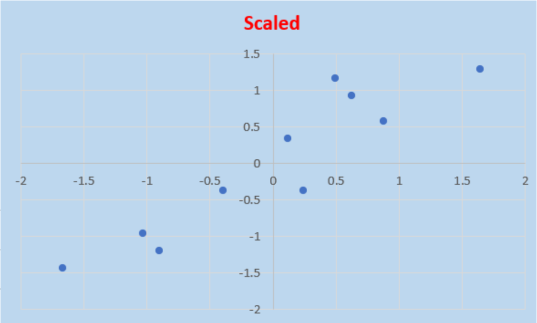
for $\lambda = 0$ eigen vector PC3

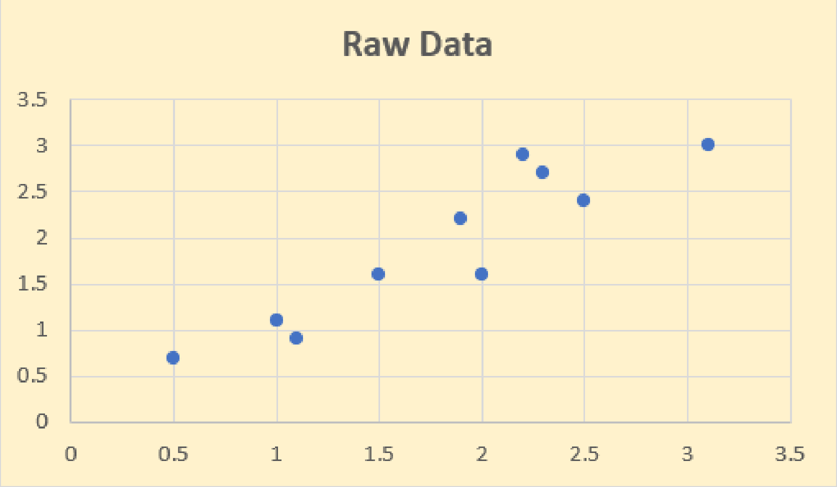
$$\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$$ ☺

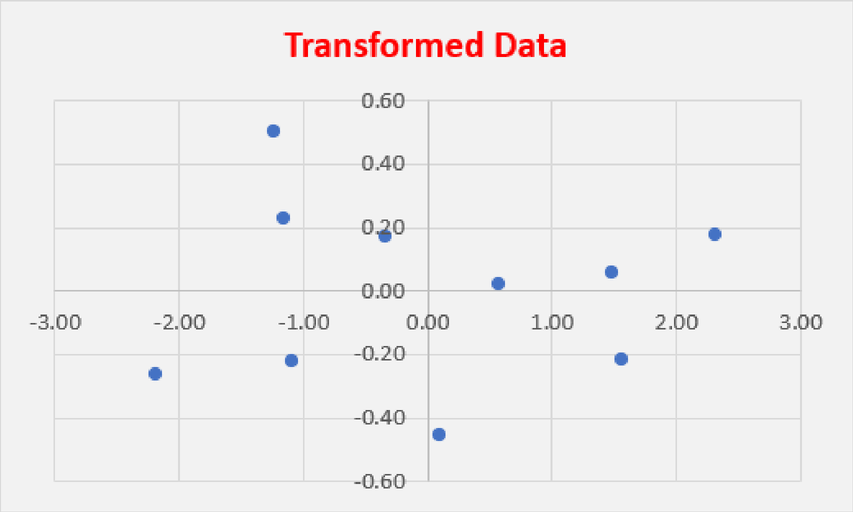
Transformed Data

Taking a picture

	Raw Data				
	x	y		x_trans	y_trans
1	2.5	2.4	1	-1.09	-0.22
2	0.5	0.7	2	2.31	0.18
3	2.2	2.9	3	-1.24	0.50
4	1.9	2.2	4	-0.34	0.17
5	3.1	3	5	-2.18	-0.26
6	2.3	2.7	6	-1.16	0.23
7	2	1.6	7	0.09	-0.45
8	1	1.1	8	1.48	0.06
9	1.5	1.6	9	0.57	0.02
10	1.1	0.9	10	1.56	-0.22







$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$