# **Motivation**

## **Data Compression**

Chart, scatter chart

Description automatically generated  
Why these points do not fall on the straight line ??  
Since the inches and cm is rounded off to their respective nearest inches and cm.

Chart

Description automatically generated   
Just to summarize, if we allow ourselves to approximate the original data set by projecting all of my original examples onto this blue line in the graph, then I need only one number, I need only real number to specify the position of a point on the line, and so what I can do is therefore use just one number to represent the location of each of my training examples after they've been projected onto that blue line.

* So this is an approximation to the original training self because I have projected all of my training examples onto a line.
* But now, I need to keep around only one number for each of my examples.
* The learning algorithm runs quickly even the more interesting application of this data compression rather than reducing the memory or disk space requirement for storing the data.

Chart

Description automatically generatedWe are going to project it onto 2D. So, I've projected this data so that all of it now lies on this 2D surface. As you can see all the data lies on a plane, because we've projected everything onto a plane, and so what this means is that now I need only two numbers, z1 and z2, to represent the location of point on the plane.

Graphical user interface, chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Text

Description automatically generatedText, letter

Description automatically generated

Suppose if there are 50 features, then how to visualize this data 🡪 not possible

## **Visualization**

# **Principal Component Analysis PCA**

## **Principal Component Analysis Problem Formulation**

Chart, line chart

Description automatically generated  
PCA tries to find a lower dimensional surface(i.e red line) onto which to project the data so that the sum of squares of the these little blue line segment is minimized.   
These blue lines are known as projection error.

Chart, line chart

Description automatically generated

With the red line 🡪 projection error are less.  
With the magenta line 🡪 projection error are more.

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generatedLets draw u(1) and u(2) vectors and extending these vectors will result in a 2D surface/plane on to which I am going to project my data.

Chart

Description automatically generated with medium confidence

Formal definition 🡪 We are going to find the set of vectors u(1), u(2), .. u(k). And then we are going to project the data onto the linear subspace spanned by this set of k vectors.

Informal definition 🡪 Finding k directions instead of just one direction onto which to project the data.

In PCA, we want to find away to project the data so as to minimize the sort of projection distance(i.e projection error) which is the distance between the points and the projection(green points)

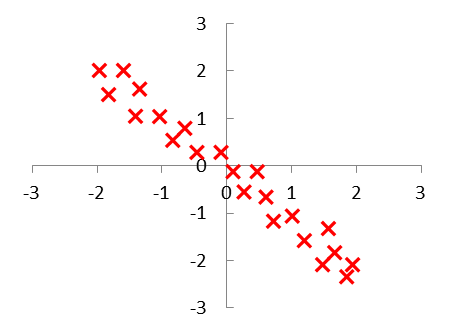
Chart, line chart

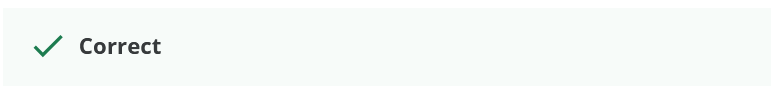
Description automatically generated  
Graphical user interface, application

Description automatically generated

PCA is not Linear Regression and even though at some cosmetic level they might look related but these are actually very different algorithm.

Graphical user interface, text, application

Description automatically generated with medium confidenceA picture containing text

Description automatically generated

# **Applying PCA**

# **Machine Learning Mastery**

Graphical user interface, text, application

Description automatically generated  
Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

As the number of dimension increases 🡪 accuracy decreases which is a curse.

# **References**

<https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/>

<https://www.ritchieng.com/machine-learning-dimensionality-reduction/>

<https://www.youtube.com/watch?v=OFyyWcw2cyM&t=1s>