# Machine Learning
# Week 8
# Section 2
# Dimensionality Reduction

--------------------------------------------

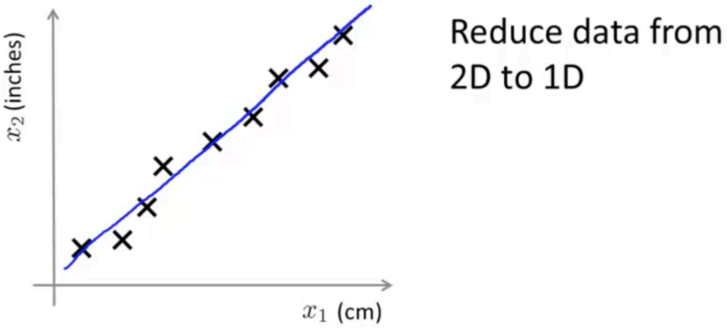
In this module, we look at an introduction to Principal Component Analysis and see how it can be used for data compression to speed up learning algorithms as well as for visualization of complex datasets.

## Motivation 1: Data Compression

Lets look at another type of unsupervised learning problem called dimensionality reduction.

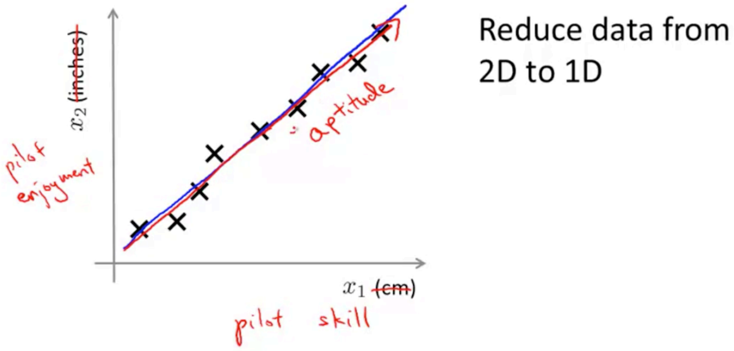
There are many good uses of dimensionality reduction, one is data compression.

Dimensionality Reduction / Data Compress example: Say for some data set we had a length feature that was in inches and one in centimeters that represented the same value, instead of having two separate features we can reduce it to one value to measure the length.



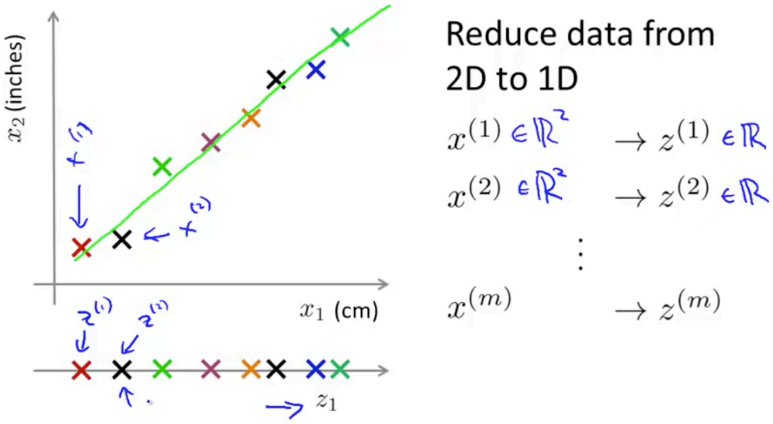
Reduce data from
2D to 1D

Another example in which two features don't necessary hold the same value but correlate an attribute could be pilot skill & enjoyment which could be indicating pilot aptitude:



Reduce data from
2D to 1D

Mathematically we do the following projection:



Reduce data from
2D to 1D

$x^{(1)} \in \mathbb{R}^2 \rightarrow z^{(1)} \in \mathbb{R}$

$x^{(2)} \in \mathbb{R}^2 \rightarrow z^{(2)} \in \mathbb{R}$

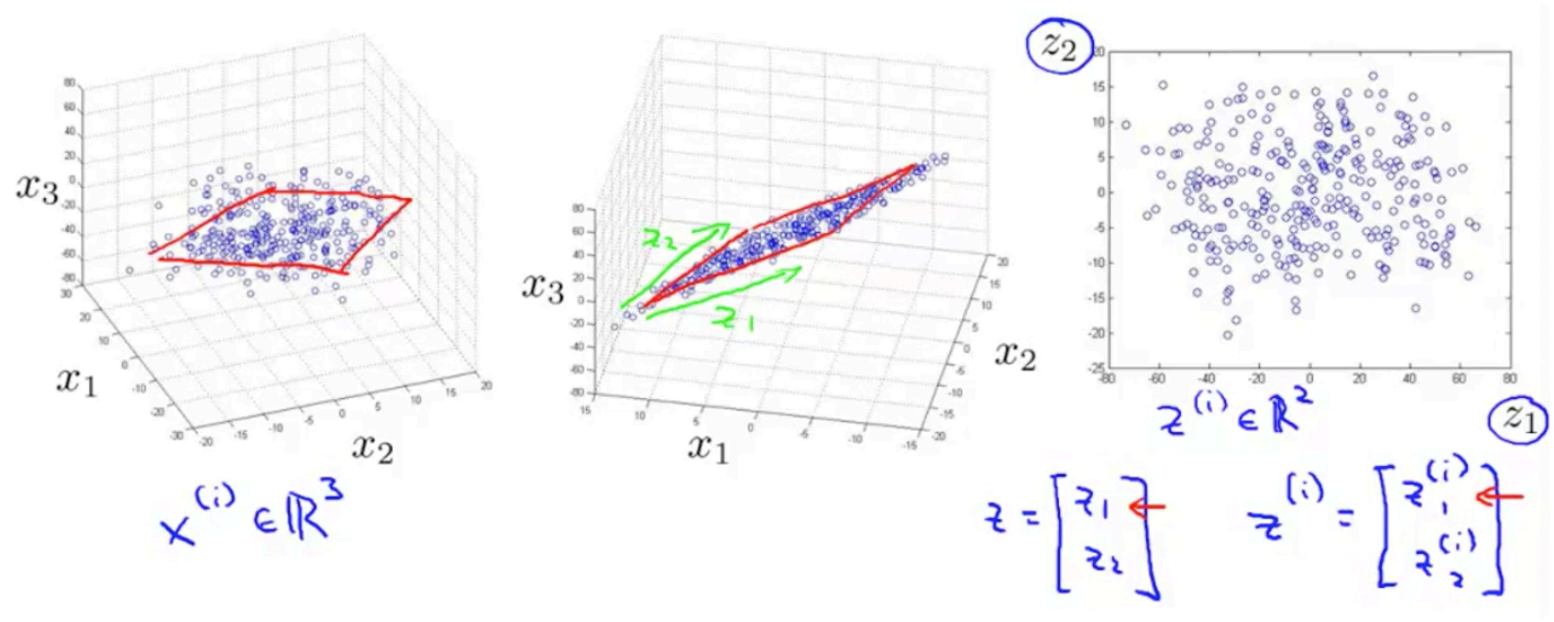
$\vdots$

$x^{(m)} \rightarrow z^{(m)}$

This halves the memory / space requirement for storing our data and allows for faster computation.

We can also reduce much higher dimensional data (1,000D -> 100D), but because of human limitations lets look at 3D -> 2D.

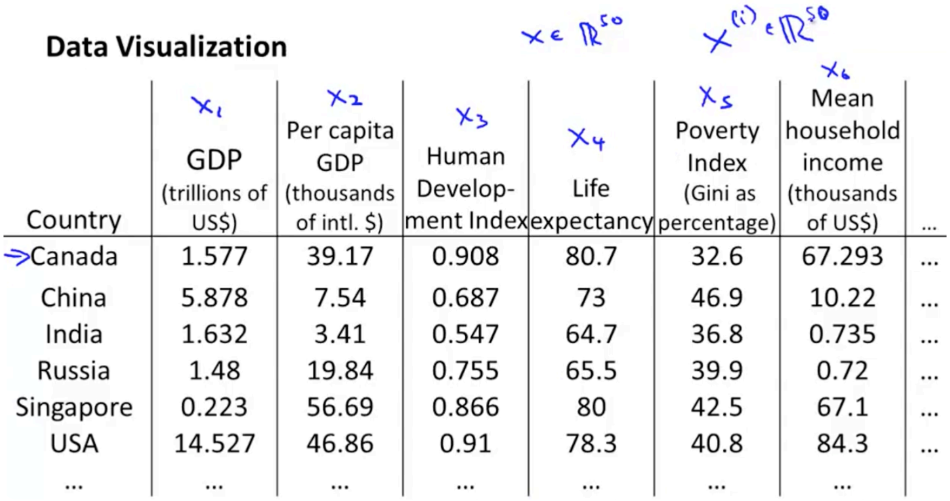
Maybe most of our data lies on a plane and can perform the following projection:



# Motivation 2: Visualization

A second application of dimensionality reduction is data visualization.

Let us say we have a large dataset table of countries economic and population metrics, how can we visualize this data?

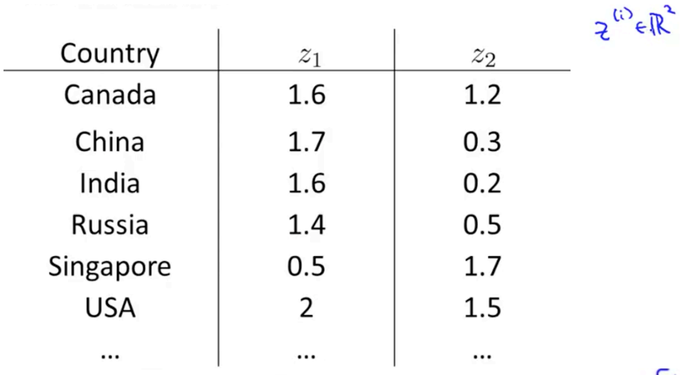


**Data Visualization**

$x \in \mathbb{R}^{50}$ $x^{(i)} \in \mathbb{R}^{50}$

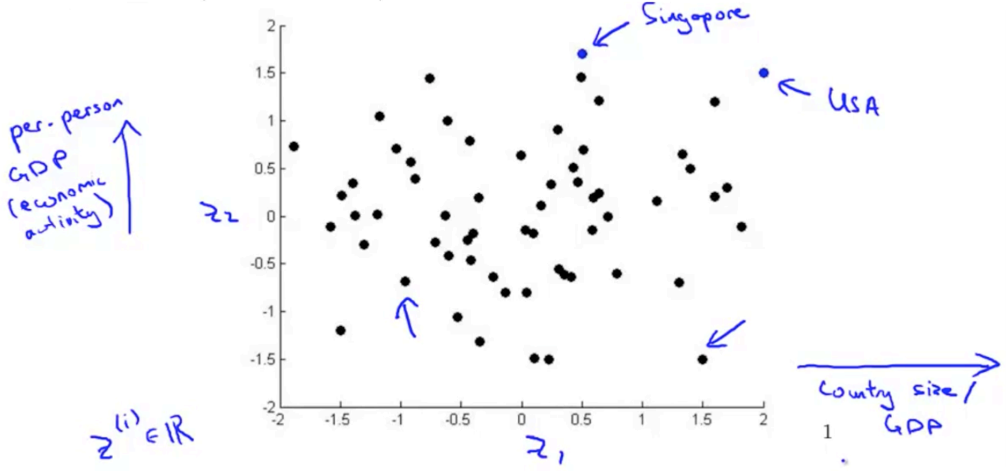
| Country | $x_1$<br>GDP<br>(trillions of US$) | $x_2$<br>Per capita GDP<br>(thousands of intl. $) | $x_3$<br>Human Development Index | $x_4$<br>Life expectancy | $x_5$<br>Poverty Index<br>(Gini as percentage) | $x_6$<br>Mean household income<br>(thousands of US$) | … |
|---|---|---|---|---|---|---|---|
| Canada | 1.577 | 39.17 | 0.908 | 80.7 | 32.6 | 67.293 | … |
| China | 5.878 | 7.54 | 0.687 | 73 | 46.9 | 10.22 | … |
| India | 1.632 | 3.41 | 0.547 | 64.7 | 36.8 | 0.735 | … |
| Russia | 1.48 | 19.84 | 0.755 | 65.5 | 39.9 | 0.72 | … |
| Singapore | 0.223 | 56.69 | 0.866 | 80 | 42.5 | 67.1 | … |
| USA | 14.527 | 46.86 | 0.91 | 78.3 | 40.8 | 84.3 | … |
| … | … | … | … | … | … | … | |

Maybe we can come up with a different feature representation as follows in which we have a pair of numbers z1,z2 which summarize our 50 features, we can plot these countries in R2 (Reduce the data from 50D to 2D)



$z^{(i)} \in \mathbb{R}^2$

| Country | $z_1$ | $z_2$ |
|---|---|---|
| Canada | 1.6 | 1.2 |
| China | 1.7 | 0.3 |
| India | 1.6 | 0.2 |
| Russia | 1.4 | 0.5 |
| Singapore | 0.5 | 1.7 |
| USA | 2 | 1.5 |
| … | … | … |



It is up to us to interpret what z1,z2 are / represent and can be somewhat ambiguous.

These types of visualizations may help us more succinctly capture what the main dimensions of variations are for our data.

In the next video we'll start to develop a specifically algorithm called Principal Component Analysis (