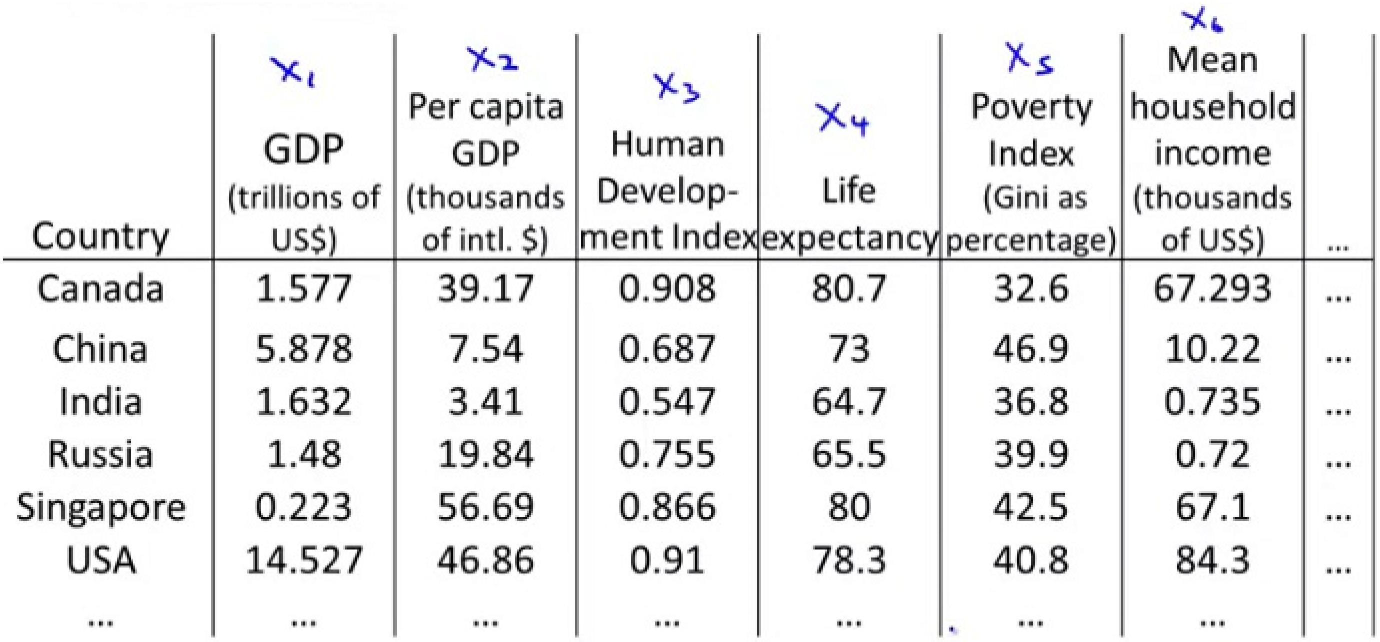
Motivation II: Visualization

For a lot of machine learning applications, it really helps us to develop effective learning algorithms, if we can understand our data better, If there is some way of visualizing the data better, and so, dimensionality reduction offers us, often another useful tool to do so.

Data Visualization

Example: Let's say we've collected a large data set of many statistics and facts about different countries around the world.

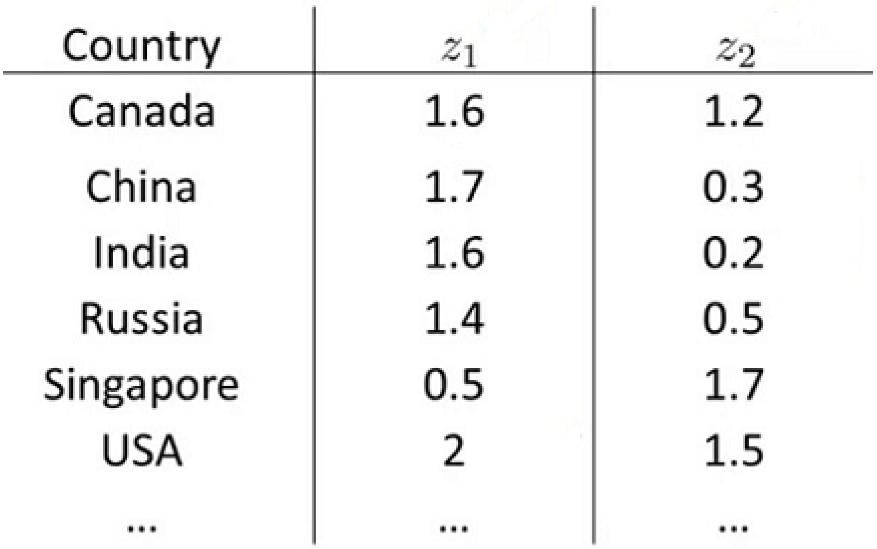


Country	x_1 GDP (trillions of US$)	x_2 Per capita GDP (thousands of intl. $)	x_3 Human Development Index	x_4 Life expectancy	x_5 Poverty Index (Gini as percentage)	x_6 Mean household income (thousands of US$)	...
Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	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	...
...	...	...	...	...	...	...	

- So x_1 = GDP, x_2 = Per capita GDP, $\ldots$, x_6 = Mean household income
- Say we have 50 features per country $x^{(i)} \in \mathbb{R}^{50}$

So is there something we can do to try to understand our data better?, we've given this huge table of numbers. How do we visualize this data? If you have 50 features, it's very difficult to plot 50-dimensional data. What is a good way to examine this data?

Using dimensionality reduction, instead of each country being represented by a 50-dimensional feature vector, come up with a different feature representation (z values) which summarize these features:



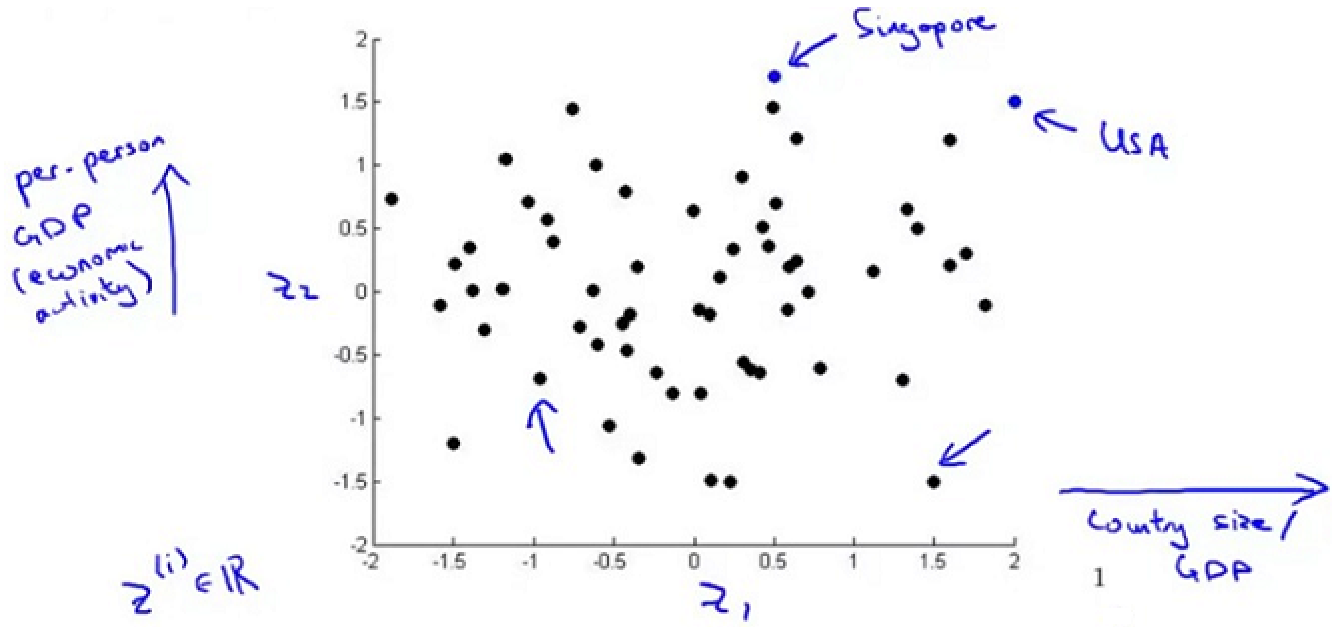
Country	z_1	z_2
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5
...	...	...

This gives us a 2-dimensional vector:

- Reduce 50D -> 2D
- Plot as a 2D plot

Typically we don't generally ascribe meaning to the new features (so we have to determine what these summary values mean)

- Example:
  - May find x horizontal axis (z_1) corresponds to overall country size/economic activity
  - and y axis may be the per-person well being/economic activity (z_2)



- So despite having 50 features, there may be two "dimensions" of information, with features associated with each of those dimensions
  - It's up to us to asses what of the features can be grouped to form summary features, and how best to do that (feature scaling is probably important)
- Helps show the two main dimensions of variation in a way that's easy to understand

Video Question: Suppose you have a dataset $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ where $x^{(i)} \in \mathbb{R}^n$. In order to visualize it, we apply dimensionality reduction and get $\{z^{(1)}, z^{(2)}, \ldots, z^{(m)}\}$ where $z^{(i)} \in \mathbb{R}^k$ is k-dimensional. In a typical setting, which of the following would you expect to be true? Check all that apply.

- $k > n$

> $k \leq n$

- $k \geq 4$

> $k = 2$ or $k = 3$ (since we can plot 2D or 3D data but don’t have ways to visualize higher dimensional data)