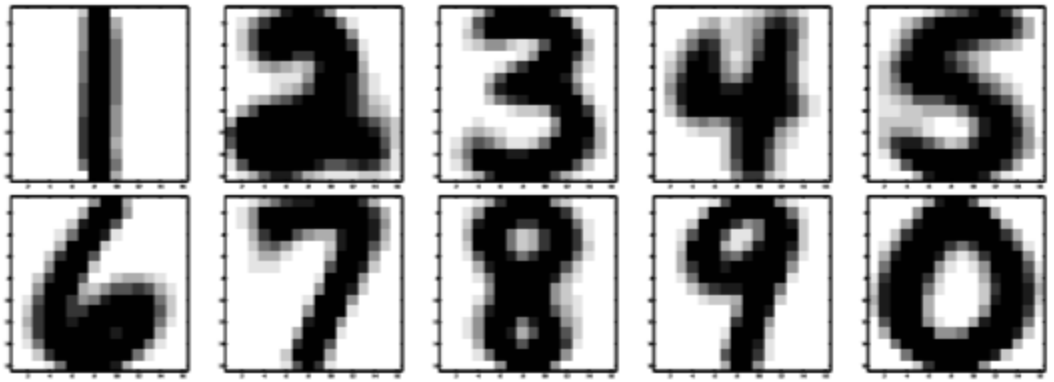
Principal Component Analysis (PCA)

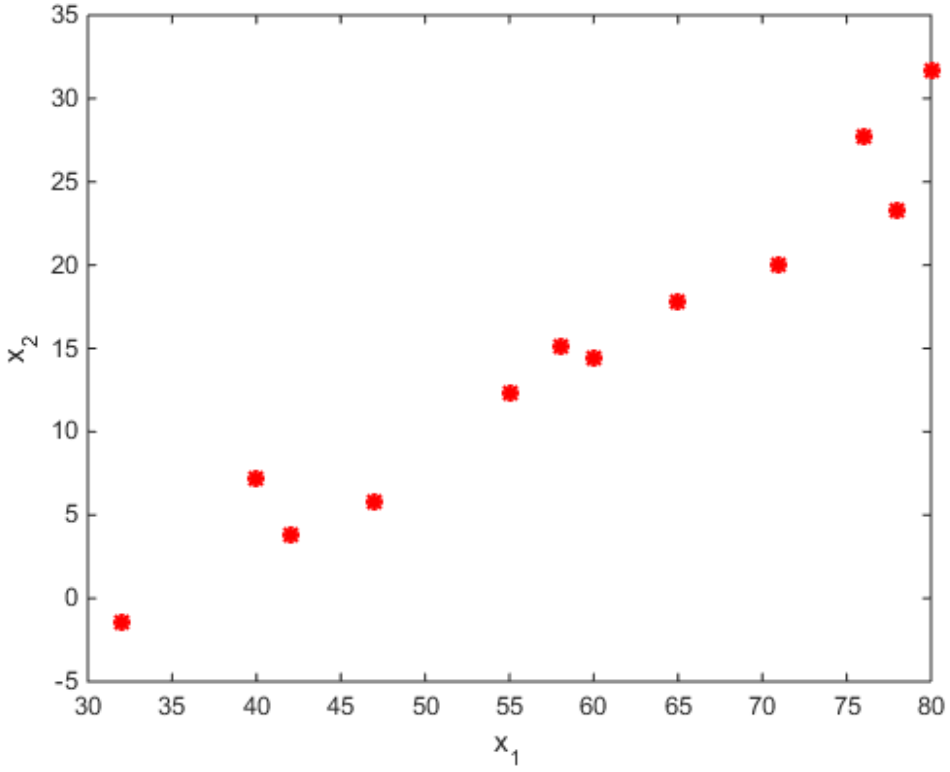
Why PCA?

- PCA has to do with reducing the dimension of multivariate data.
- Datasets are usually high dimensional.
- Homework problems:
  - Heathcare Studies: Predict patients compliance based on **90** attributes.
  - Digit Recognition: Classify the digits into one of 10 classes based on 16x16 (= **256**) pixel-images.
- Any realistic data has a high number of dimension.
  - Any text processing application can potentially deal with **billions** of words.



Why Reducing the Dimensionality?

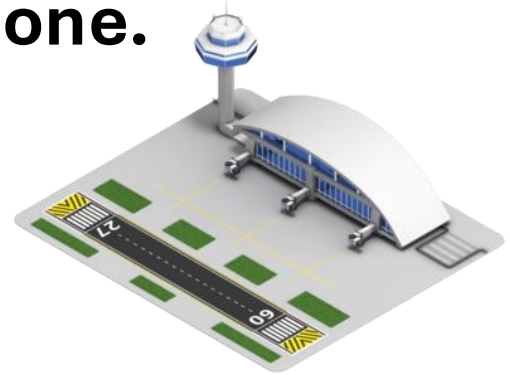
- **The true dimensionality of data is often lower than the observed dimensionality.**
- Example: You get a data set from a weather center that's collected over a period of 12 months at a particular region.
  - Data has the following format : (x_1, x_2)
  - What's the dimension of the data?
  - It turns out that x1 is temperature in Fahrenheit and x2 is temperature in Celsius.
  - So your data has only 1 dimension.



Why Reducing the Dimensionality?

- Example: You get a data set from a monitoring agency with the following attributes:
  - x_1: num. of traffic accidents
  - x_2: num. of school closures
  - x_3: num. of delayed flights
  - x_4: num. of wild fires
  - x_5: num. of patients with heat stroke

- Although, at the surface these all seem like different attributes, there's a single factor that can explain lots of these observations: temperature!
- **A machine learning algorithm should look for the single variable that counts for the others rather than looking at every individual one.**



Why Reducing the Dimensionality?

- Example: Handwritten digits in MNIST data set contains 16x16 images where each pixel can have a value of 0 or 1.
- This will result in 2^{256} possible events.
- However, many of these results will never happen, and true dimensionality is much smaller.

- **Data sets may have redundant attributes that don't contribute much to learning algorithms.**
- **Using the original representation 'wastes' the machine learning algorithm on the outcomes that will never happen.**