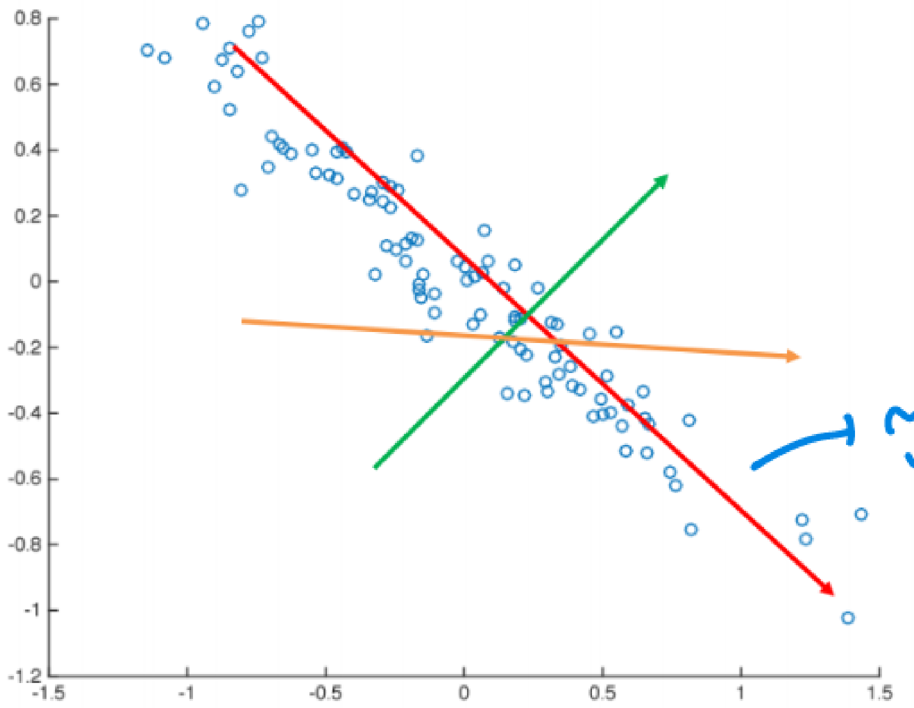
# Principal Component Analysis

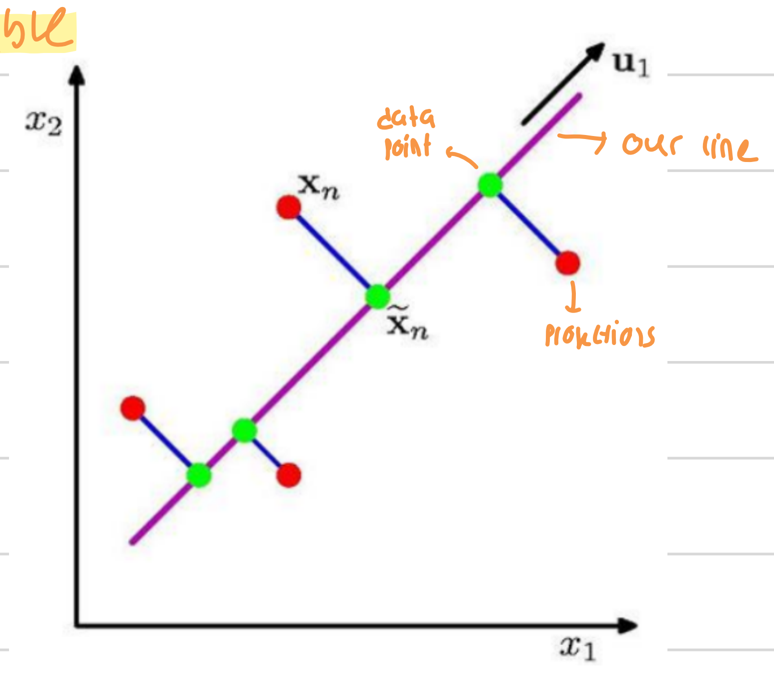
- Method for finding direction in high dimensional data that contain information } Dimension Reduction
- · unsupervised
  · non-parametric
- examine interrelations among set of variables
- Goal :- ∘ reduce the dimension of dataset while preserving the most important pattern or relationships between variables
  ∘ by finding new set of variables containing most sample's information



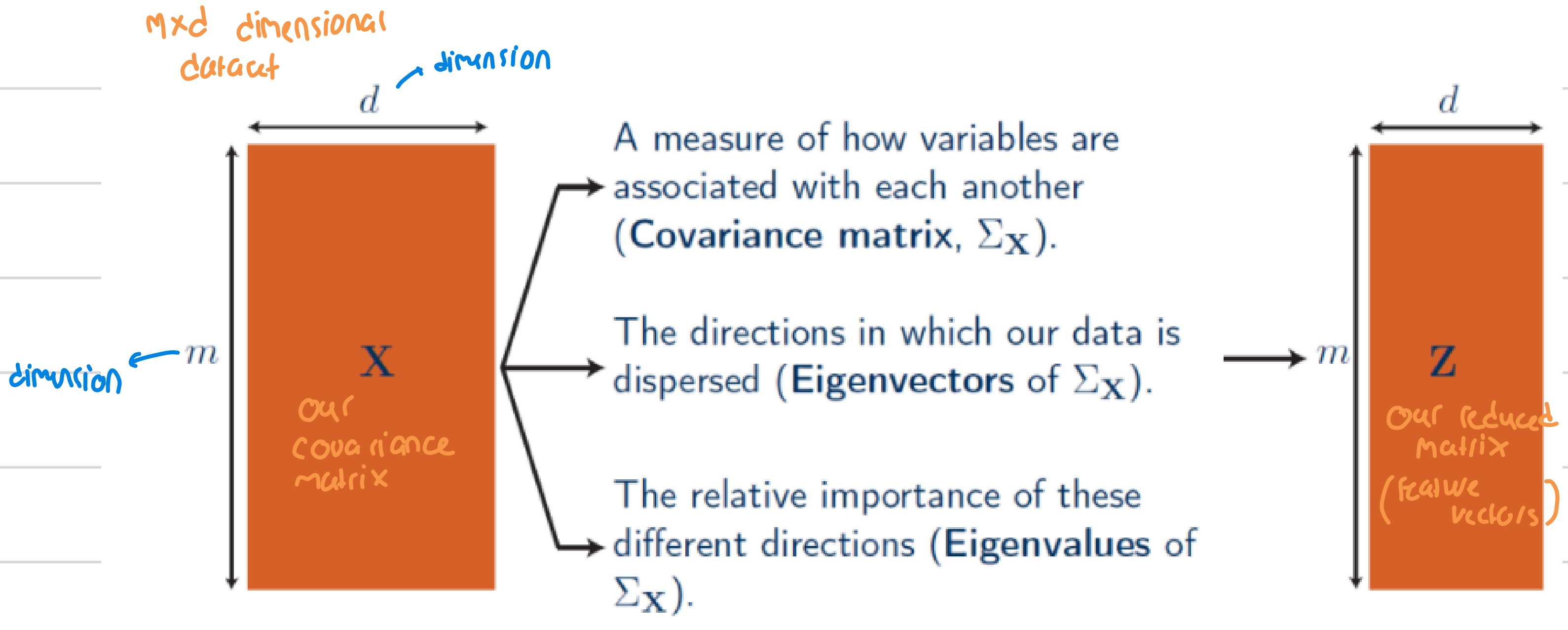
Which direction gives you the maximum variance?
Green / Orange / Red?

## Choose a line that fits the data, so that :-

- the projection maximizes variance/variable of projected data (purple line)
- minimizes mean squared distances between :-
  · data point &
  · projections (blue line)



### Principal Component Analysis



A measure of how variables are associated with each another (**Covariance matrix**, $\Sigma_{\mathbf{X}}$).

The directions in which our data is dispersed (**Eigenvectors** of $\Sigma_{\mathbf{X}}$).

The relative importance of these different directions (**Eigenvalues** of $\Sigma_{\mathbf{X}}$).

PCA linearly combines our variables and allows us to drop projections that are less informative.

## ① Eigenvectors & Eigenvalues :-

- directions of our spread w/ most info.
- each variable/feature has an eigenvectors &
- each eigenvectors have Eigenvalues :-
  - How much information carried
- Rank our eigenvectors by eigenvalues, from low to high
  $PC_1$ = highest, $PC_2$ = 2nd highest, . . .

## ② Feature Vectors :-

- Matrix that has eigenvectors of components we want to keep
- where dimension reduction happen
- keep or discord eigenvector w/ lower significance

## ③ Recast the Data along Principal Component Axes

- to recast our data from the original axes to Principal Component using our feature vector

$$\Rightarrow \text{Final Data Set} = \underbrace{\text{Feature Vector}^T}_{\text{transposed Feature vector}} * \underbrace{\text{Original Data Set}^T}_{\text{transposed og Data set}}$$