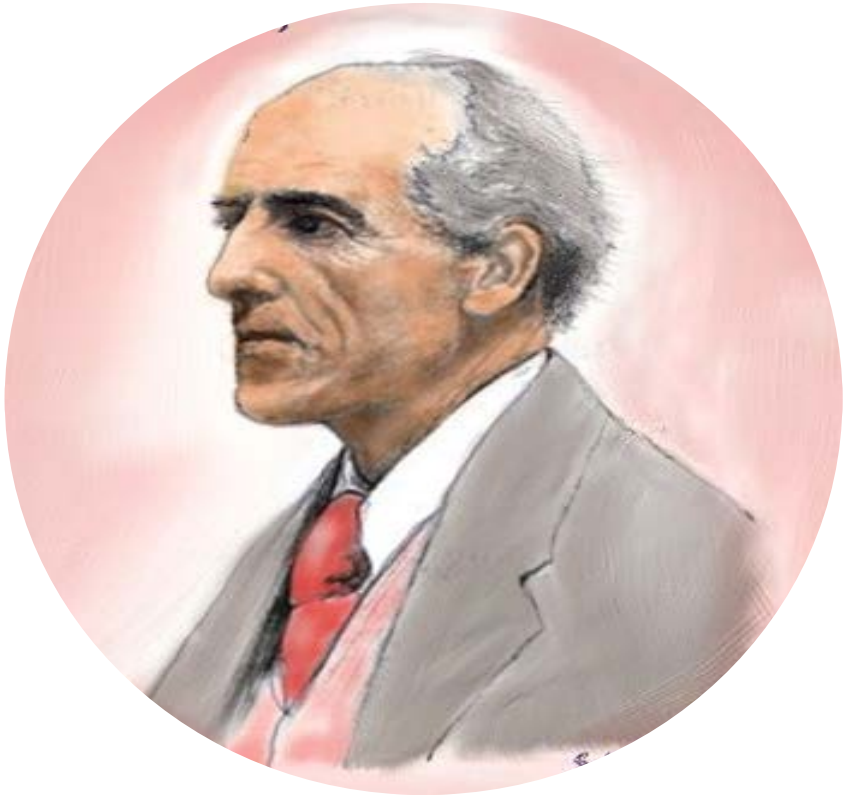
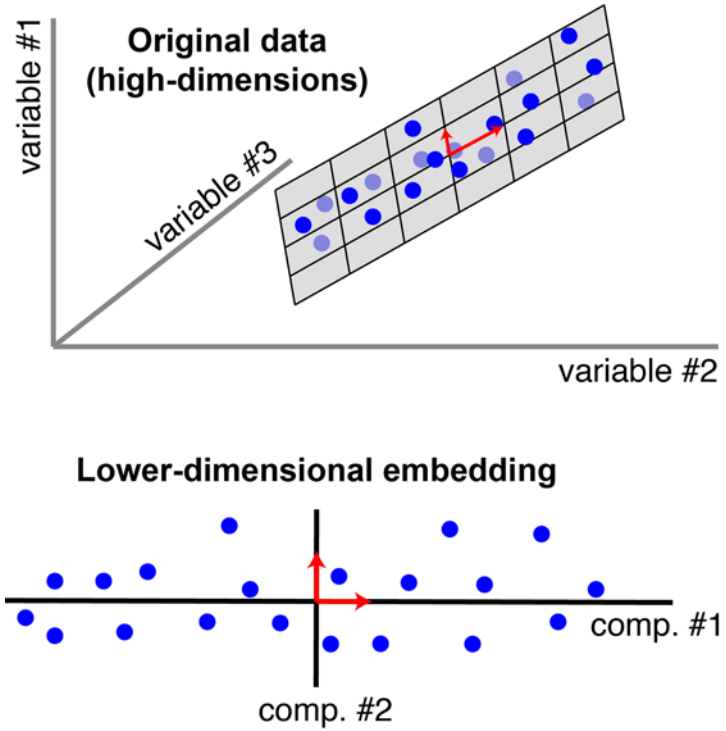
Principal Component Analysis



Karl Pearson, Father of Mathematical Statistics(1857 – 1936)

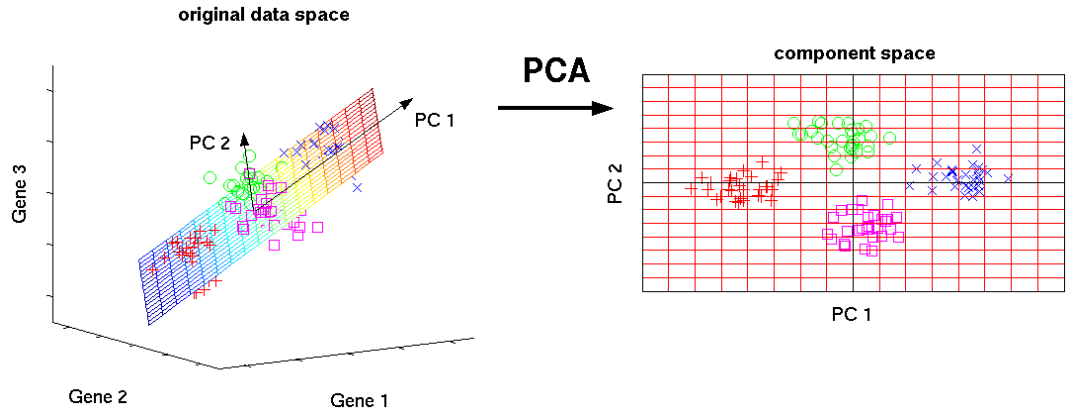
- Invented **PCA in 1901**.
- Rediscovered it multiple times in many fields like **Principal Axis theorem** in mechanics.

What is PCA ?



- PCA stands for **Principal component analysis**.
- **Unsupervised** learning algorithm.
- Method for figuring out how to represent a **complex collection of data** into a **less complex way.**

Principal Component Analysis



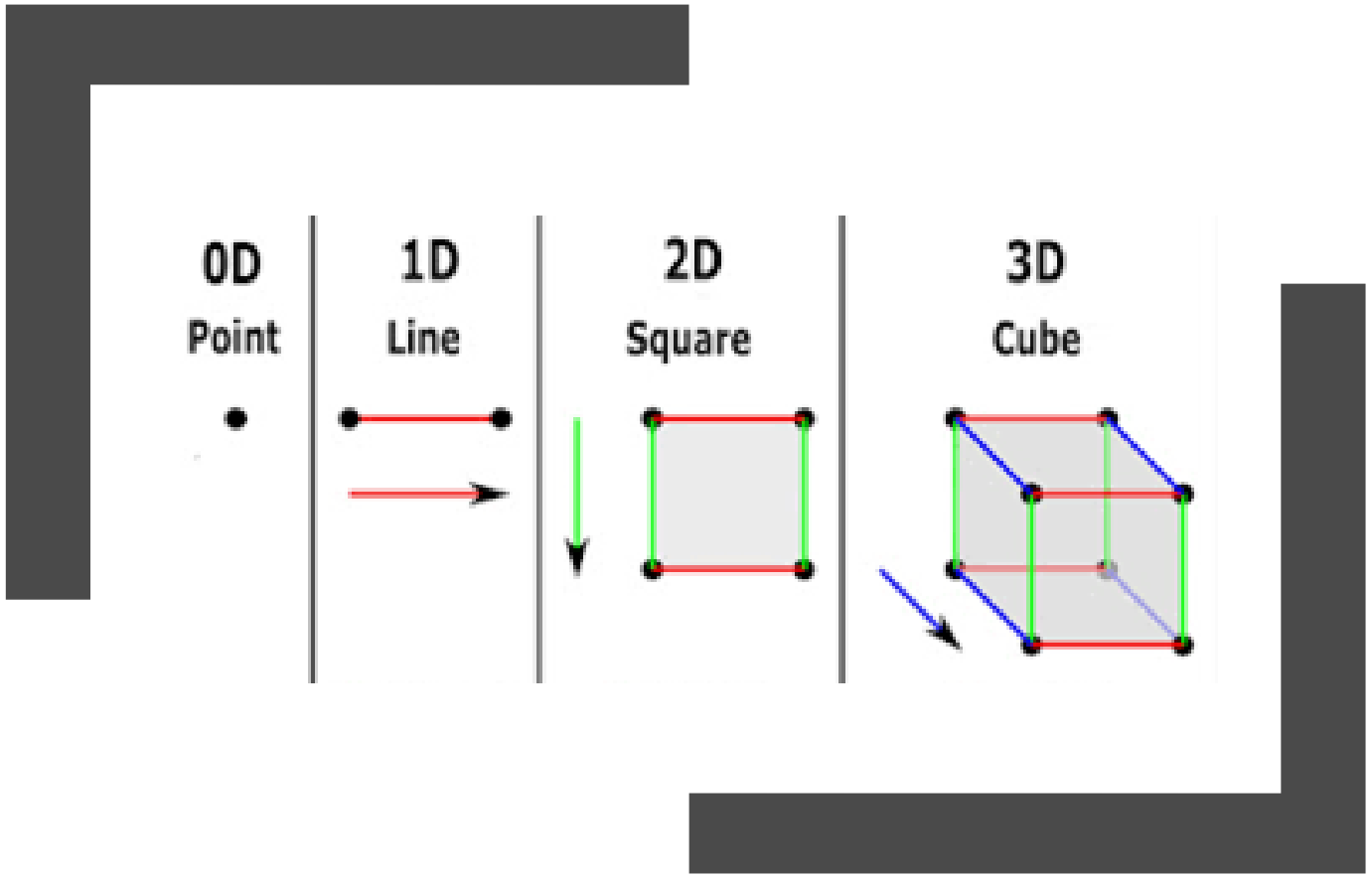
Flattening a dataset with lot of dimensions into 2 or 3 dimensions

- It is a technique to **emphasize variation** and bring out **strong patterns** in a dataset.
- It is often used to make data easy to **explore** and **visualise.**

It is a **dimensionality reduction** method.

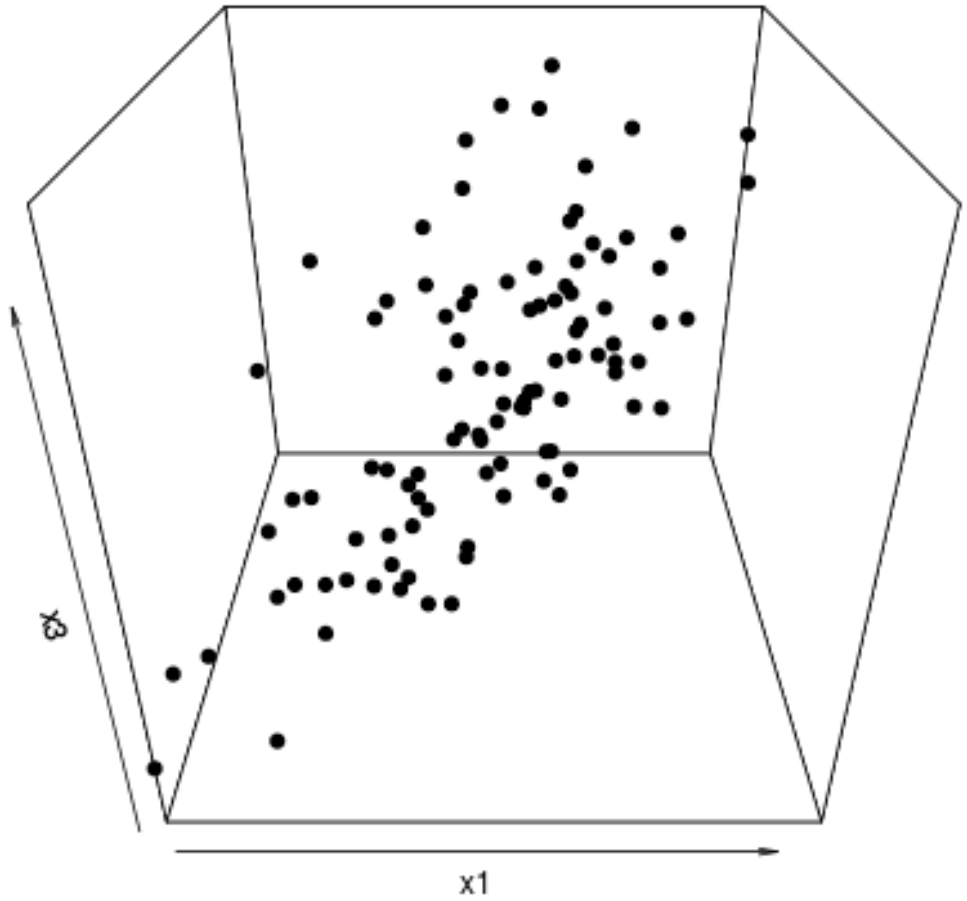
What is dimensionality reduction?

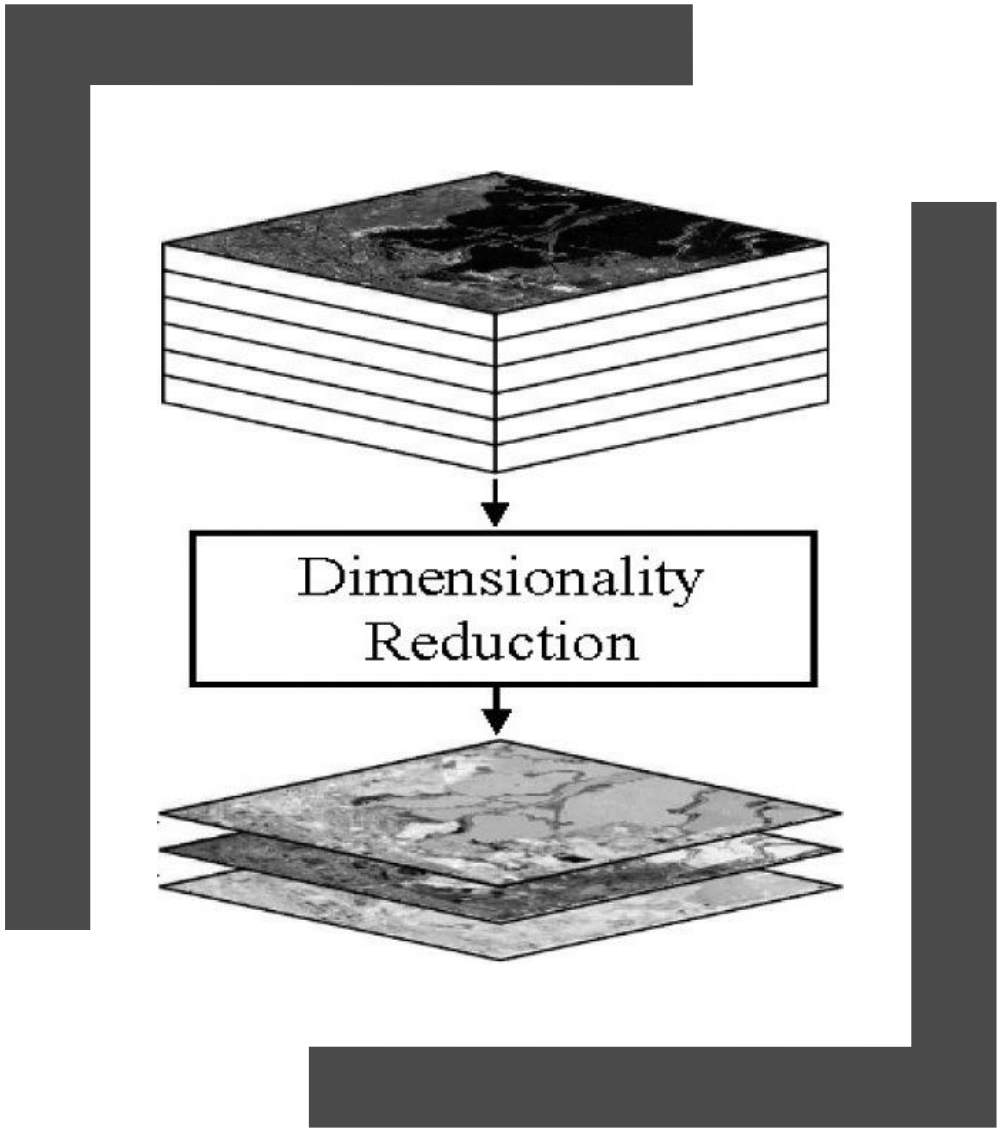
Lets look at the dimensions:



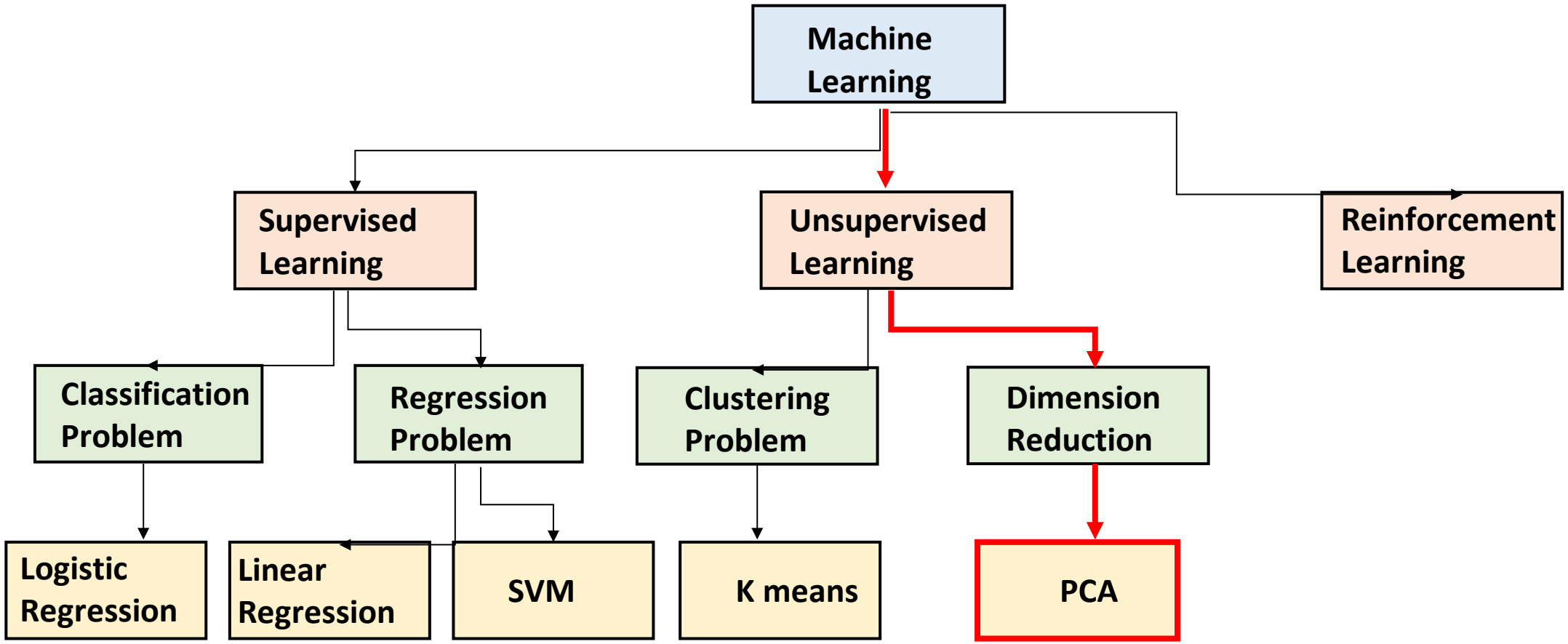
- Other dimensions cannot be drawn

Graphical Visualisation



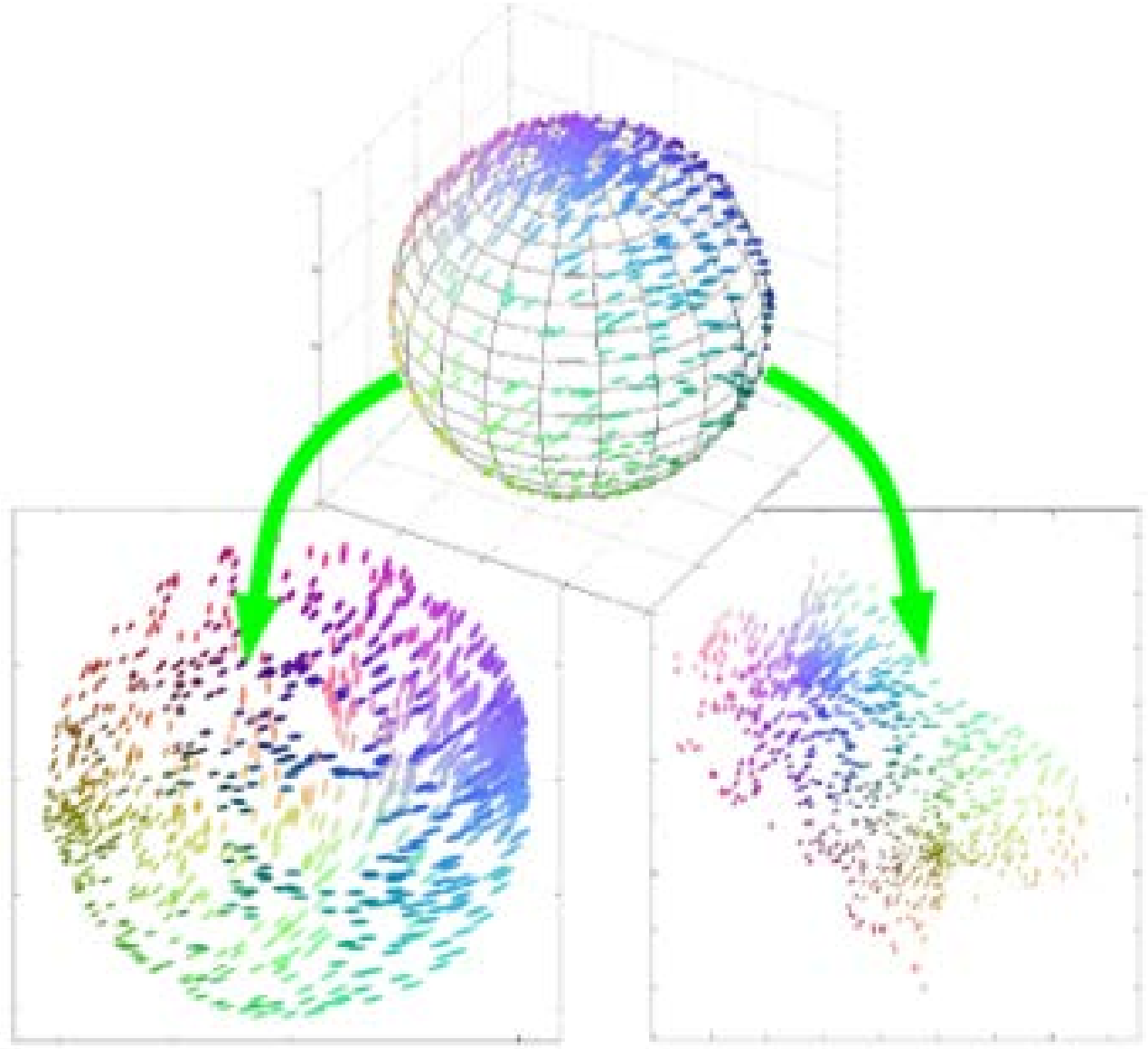


- It is reducing the **dimensions/Features** in a data.
- Through this, we can **visualise** and **plot** a data easily.


Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Classification Problem
Regression Problem
Clustering Problem
Dimension Reduction
Logistic Regression
Linear Regression
SVM
K means
PCA

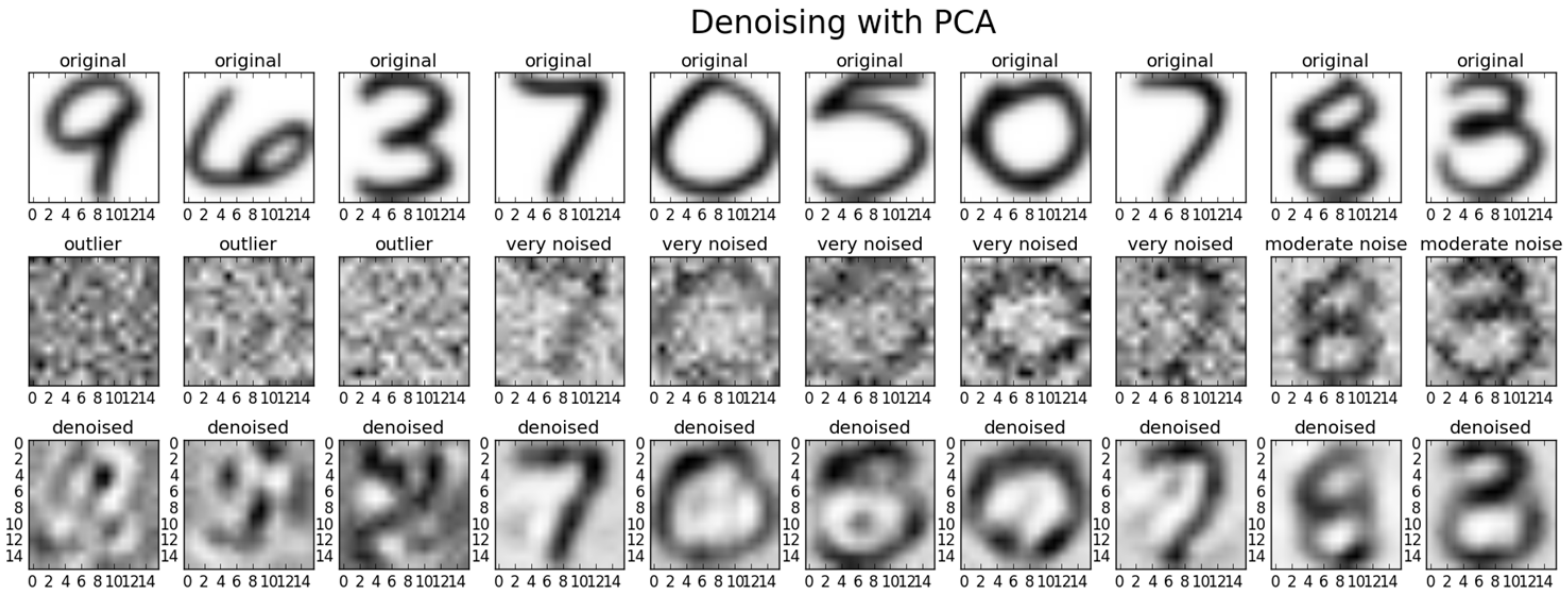
Why PCA?

- It helps **visualize high-dimensional data**.
- It **reduces noise** and finally makes other algorithms to work better.

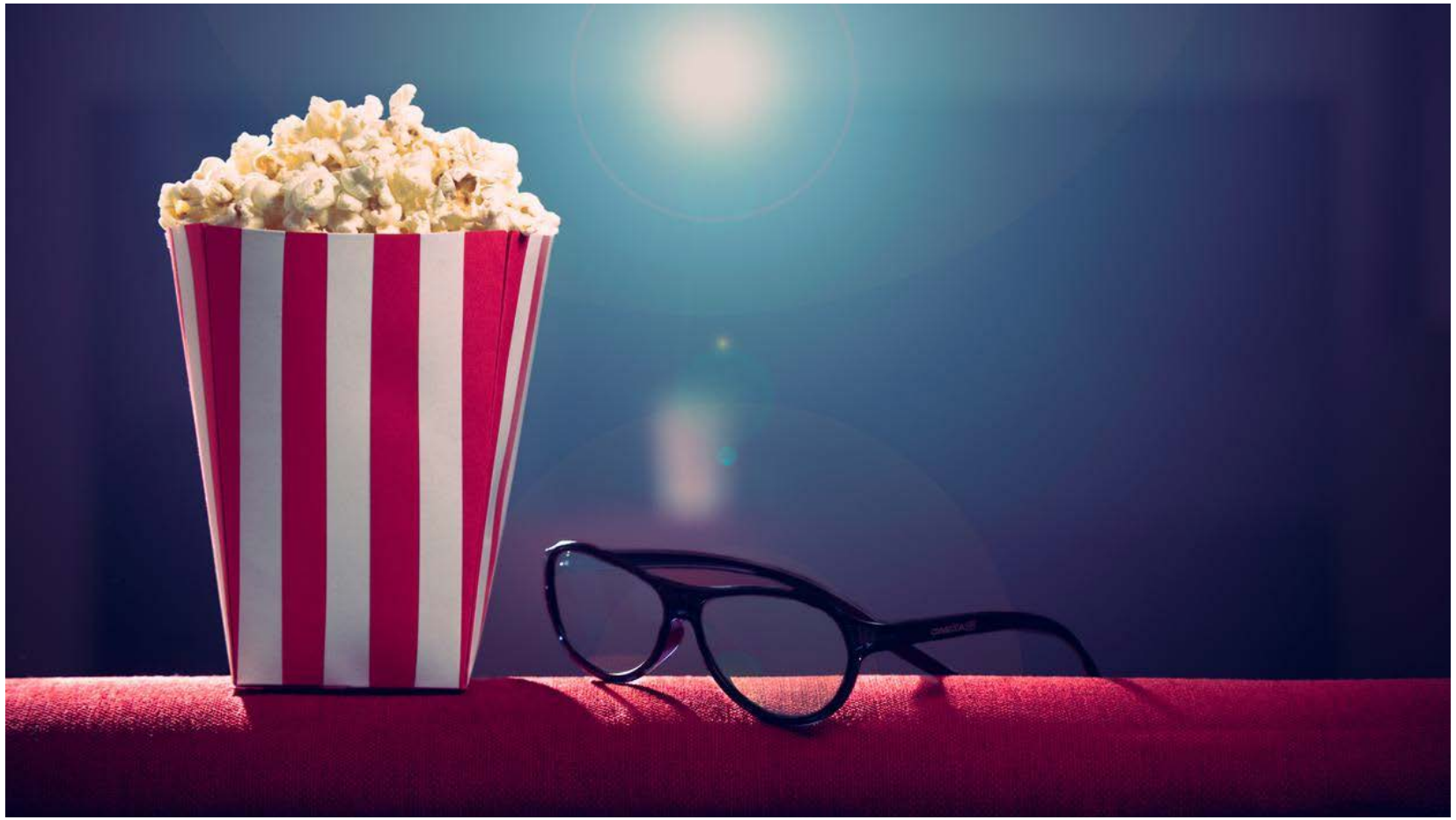


Principal Component Analysis

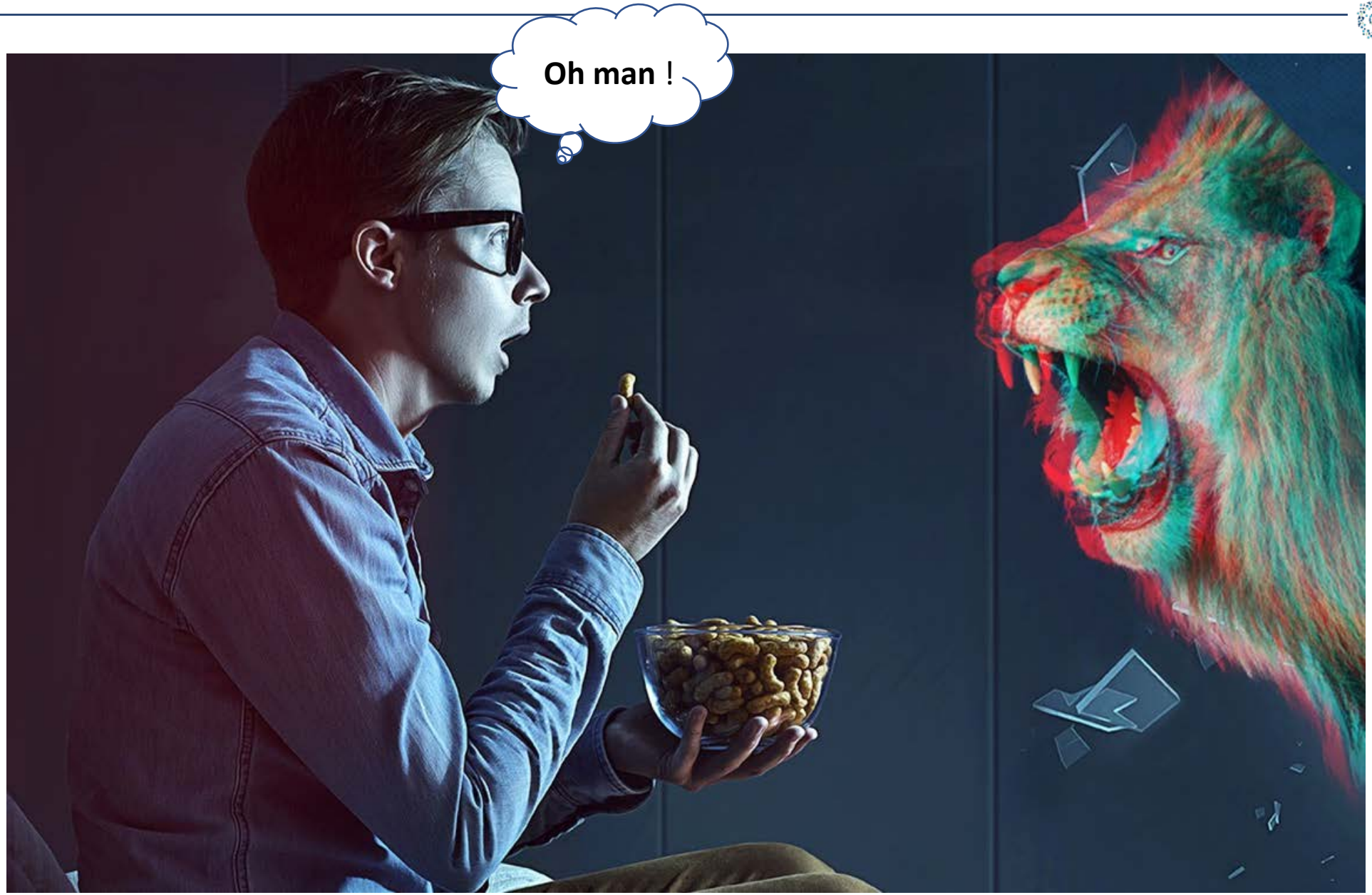
- It helps in **removing noise** in the data.



Everybody loves to watch movies, Right?



Would it be 3D or 2D?


Oh man !

Lets Consider data from a movie rating system

- Here the movie ratings from different users are **instances**, and the various movies are **features**.

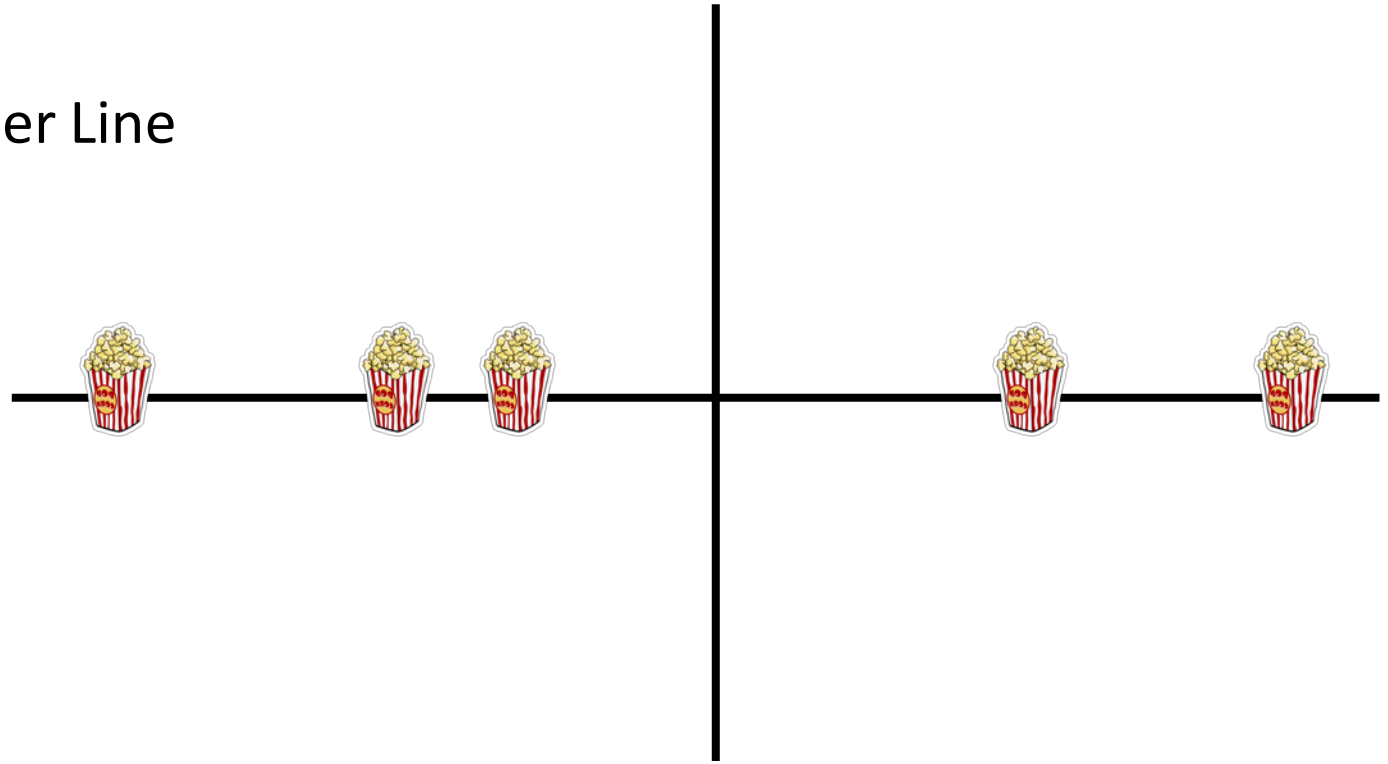
	X 1	X 2	X 3	X 4
Y 1	100	98	80	70
Y 2	0	20	40	60
Y 3	140	100	120	80
Y 4	330	450	480	520

If we consider the first column(Movie 1):

	X 1	X 2	X 3	X 4
Y 1	100	98	80	70
Y 2	0	20	40	60
Y 3	140	100	120	80
Y 4	330	450	480	520

After plotting our graph would look like this:

1 – Dimensional(1 D) = Number Line



Lets consider two columns(Movie1 & Movie2):

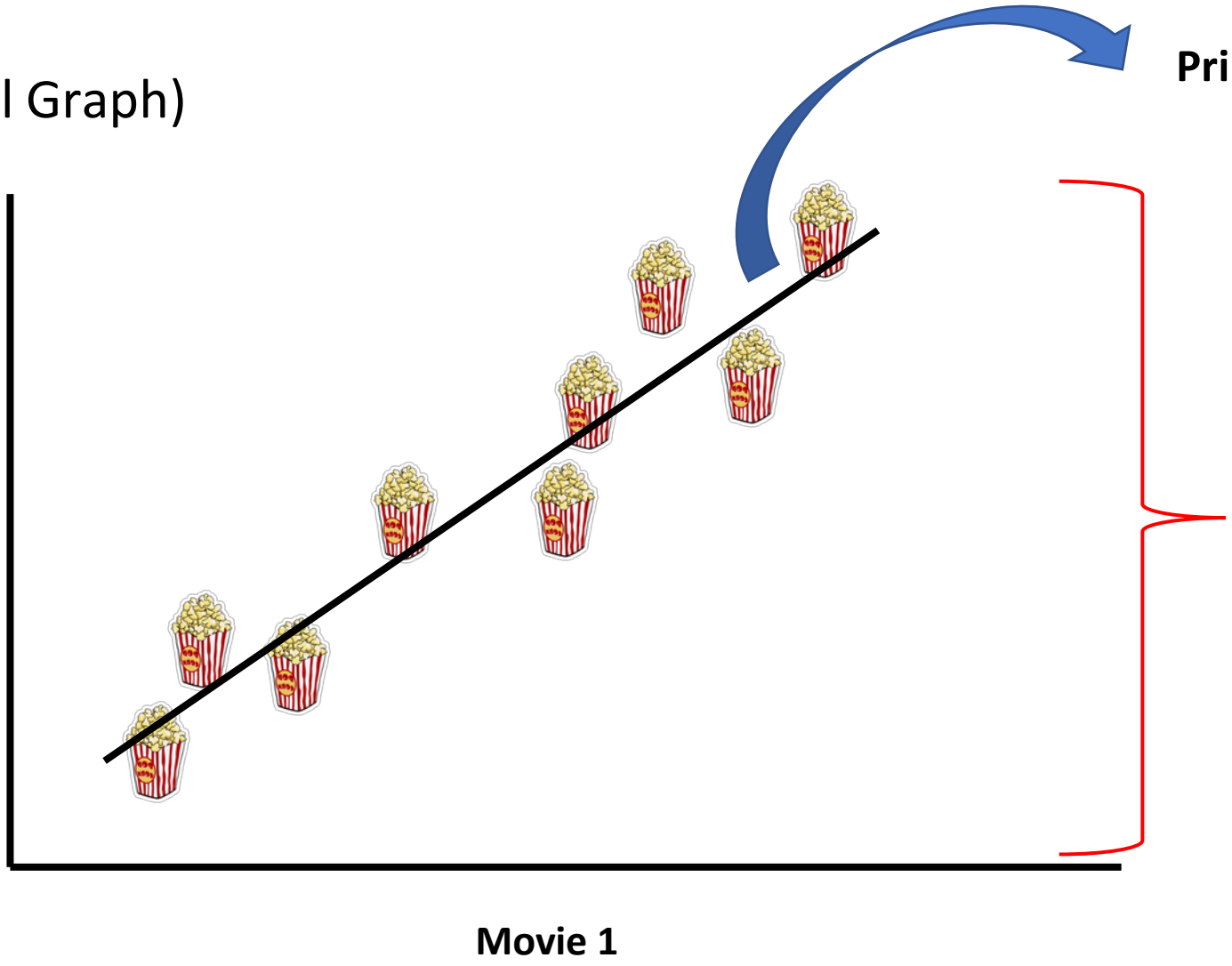
Movies (features) →

Users (instances) ↓

	Movie 1	Movie 2	Movie 3	Movie 4
Show 1	100	98	80	70
Show 2	0	20	40	60
Show 3	140	100	120	80
Show 4	330	450	480	520

After plotting our graph would look like:

2 - D(Normal Graph)



- This shows Movie 1 and Movie 2 are **correlated.**
- The dots are spread out along a diagonal line

What is Correlation?