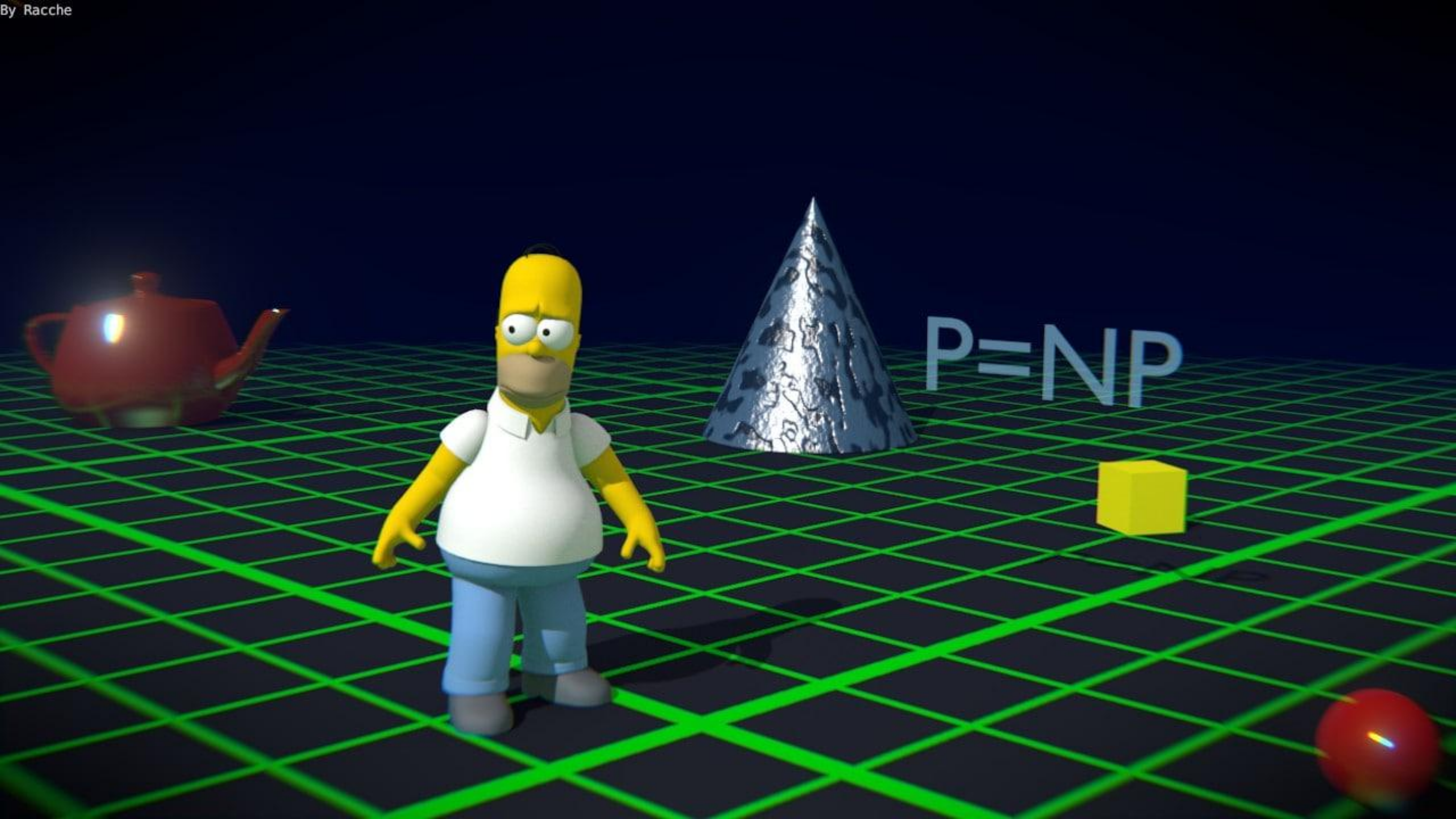
PCA

Week 07 - Day 02

Principal Component Analysis

Do you remember the kernel trick for SVM?


By Racche
P=NP

Project features

in a higher dimensional space

2D -> 3D

(x1, x2) -> (x1**2, x2**2, x1*x2)

PCA is the opposite!

Project features

in a lower dimensional space

3D -> 2D

(x1, x2, x3) -> (2*x1 + x2 - x3, x1 - x2 + x3)

New features

=

Linear combinations of old features

Why PCA?!

1. Visualization during EDA (!!!)
2. Reduce multicollinearity
3. Manage dataset where columns > rows


ALVEAR
SOLERA 1927
PEDRO XIMÉNEZ
Denominación de Origen Protegida
MONTILLA-MORILES
BARBARESCO
DENOMINAZIONE DI ORIGINE CONTROLLATA E GARANTITA
RABAJA'
RISERVA
2011
PRODOTTE 16.524 BOTT.
N° 10411
IMBOTTIGLIATO ALL'ORIGINE DAI
PRODUTTORI del BARBARESCO
S.a.c. - BARBARESCO - Italia
PRODOTTO IN ITALIA
ALC. 14.5% BY VOL. - CONTAINS SULFITES - 750 ML
RED WINE - ESTATE BOTTLED - PRODUCT OF ITALY
SELECTED & IMPORTED BY: V.I.A.S. IMPORTS L.T.D.
NEW YORK, N.Y.
GARY FARRELL
2015
Russian River Valley
Chardonnay
Sonoma County
Russian River Selection
CHATEAU
VIGNELAURE
2015
COTEAUX D'AIX EN PROVENCE
IRON HORSE
ESTATE BOTTLED
SPARKLING WINE
Wedding Cuvée
2013

Color

Alcohol content

Year

Region

Density

Nutty aroma

etc.

1.5 * Color +

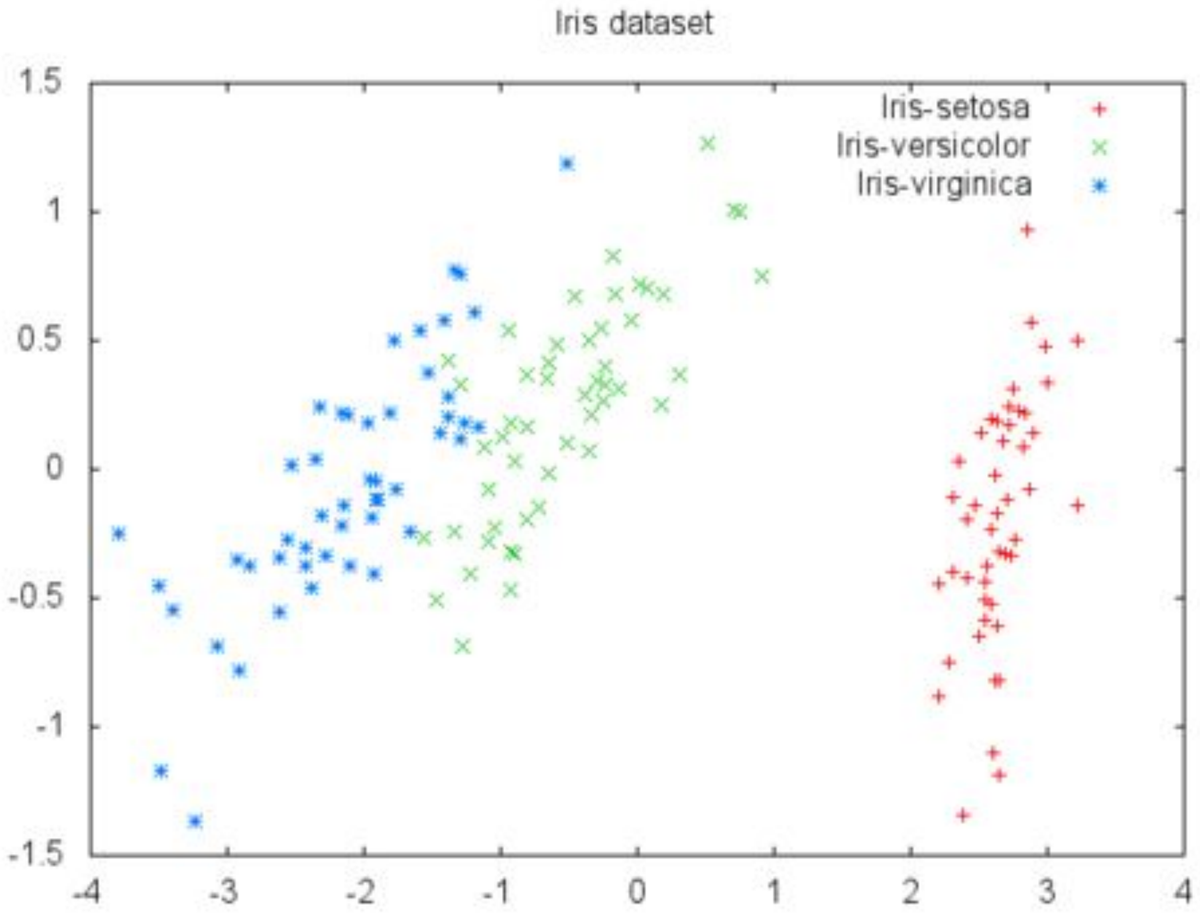
2.7 * Alcohol content +

0.2 * Year +

0.3 * Region +

0.001 * Density +

0.00001 * Nutty aroma +


Iris dataset
Iris-setosa
Iris-versicolor
Iris-virginica
1.5
1
0.5
0
-0.5
-1
-1.5
-4
-3
-2
-1
0
1
2
3
4

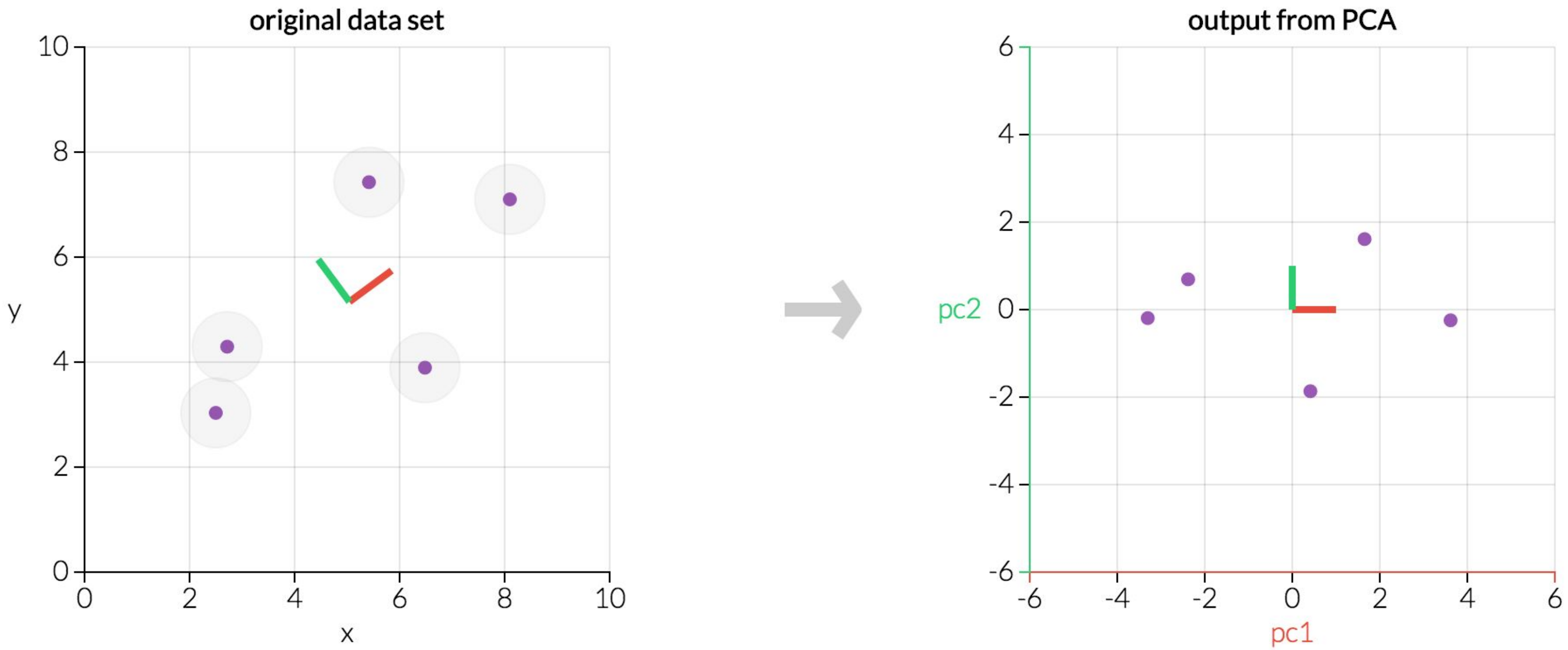
Principal Components

Principal Components

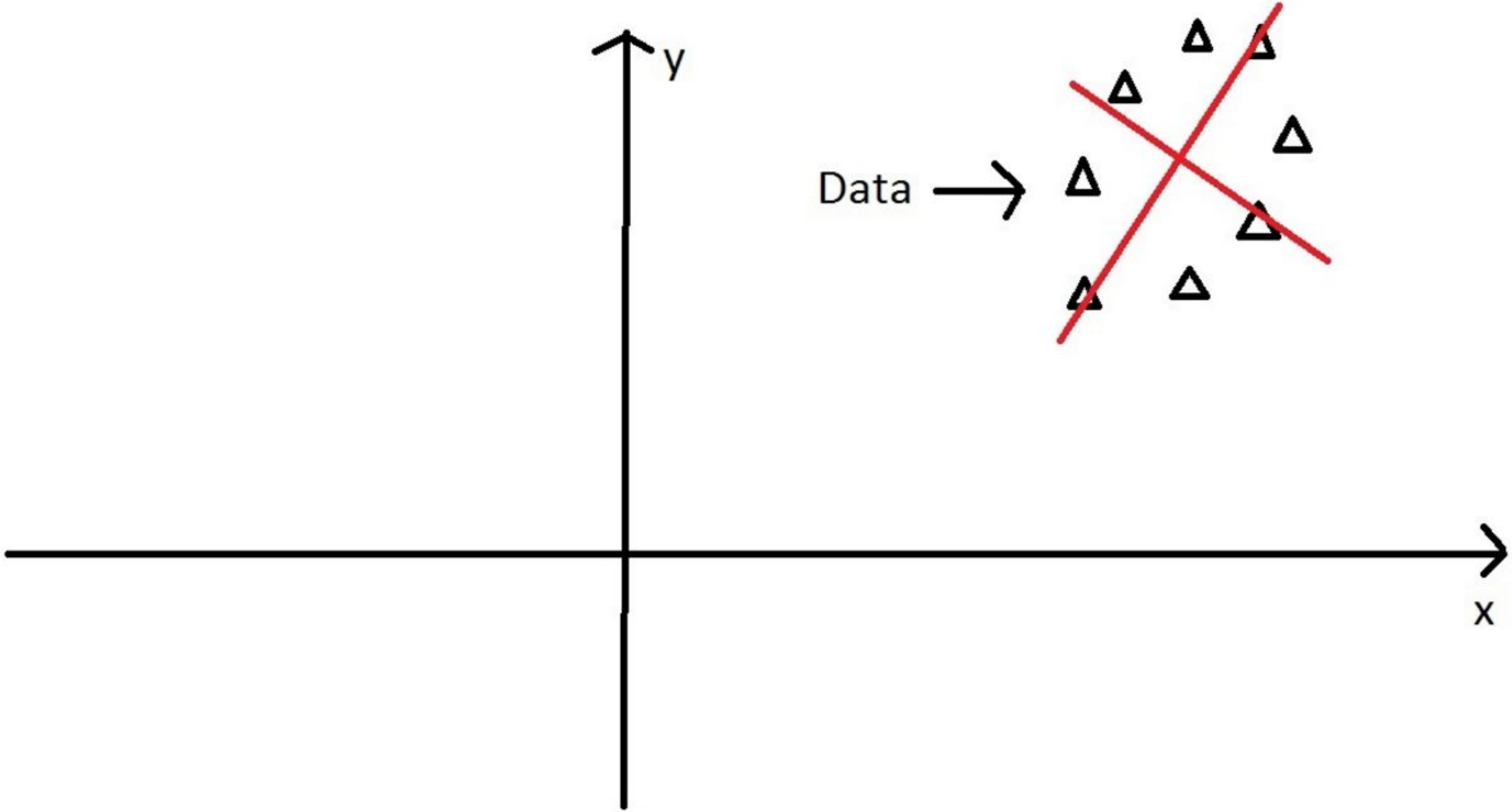
=

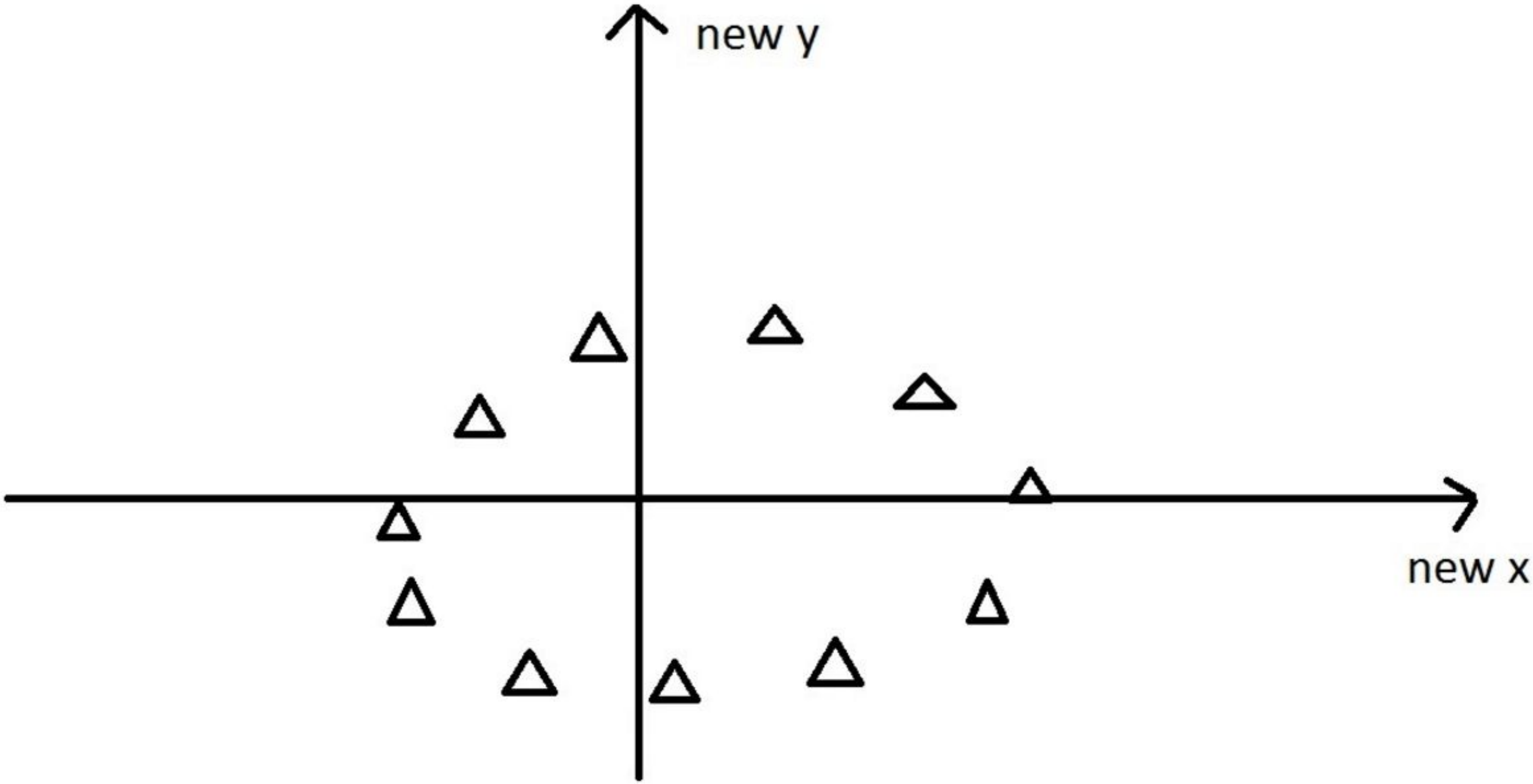
The new axis

Simple case: 2D -> 2D

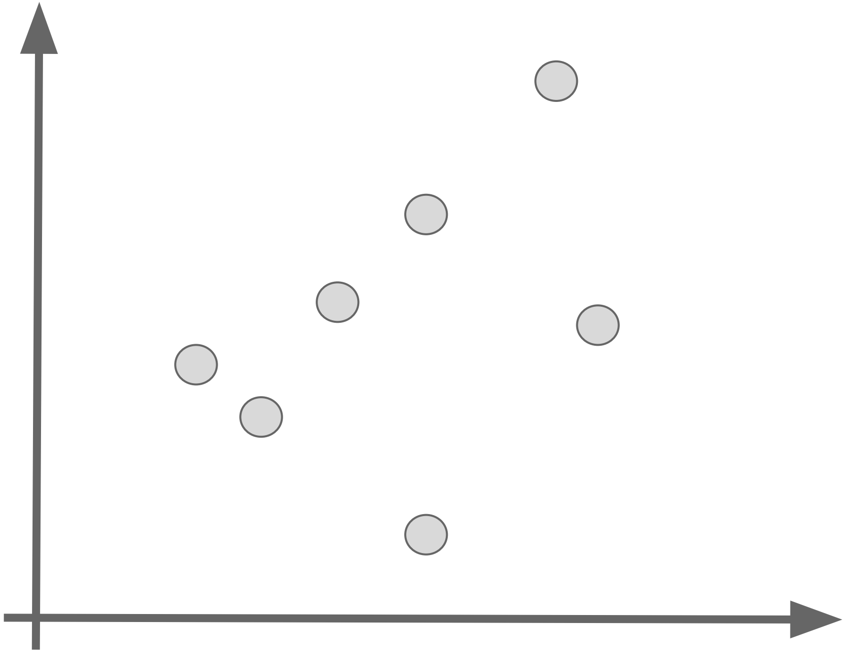


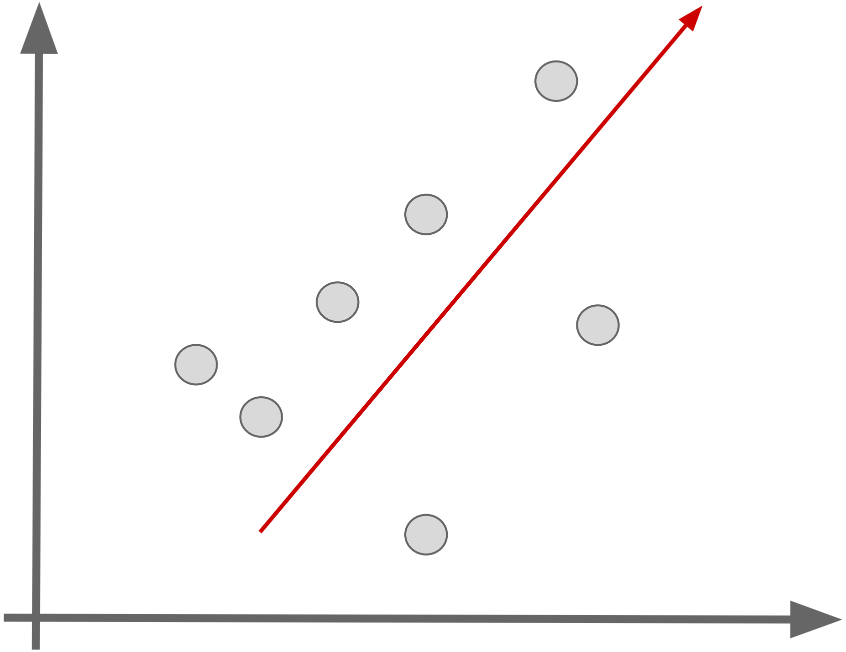
original data set
output from PCA
10
8
6
4
2
0
y
x
pc2
pc1
-6
-4
-2

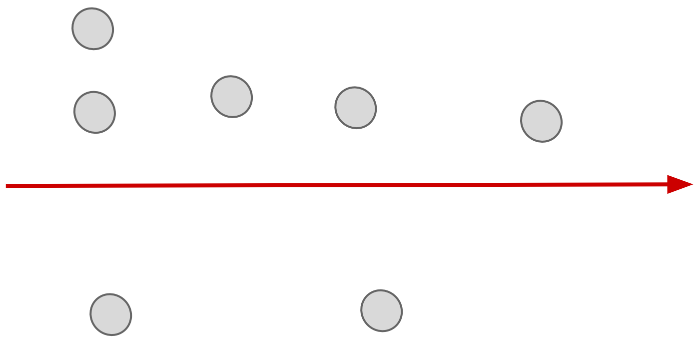

y
Data
x

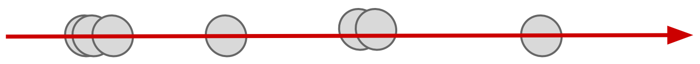

new y
new x

Example 2D->1D





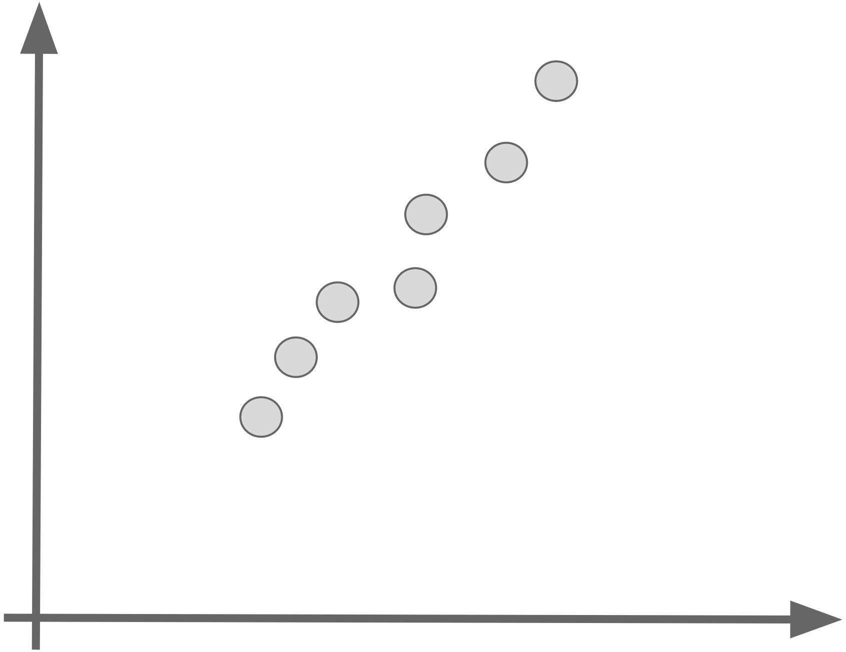




How to choose the axis
- Intuition -

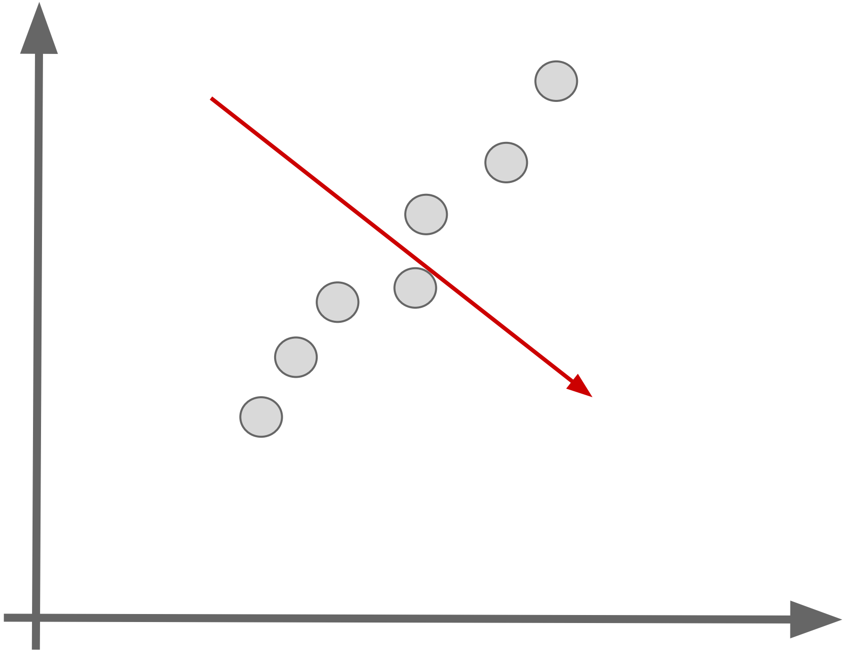
We want to “preserve”

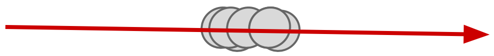
as more information as possible



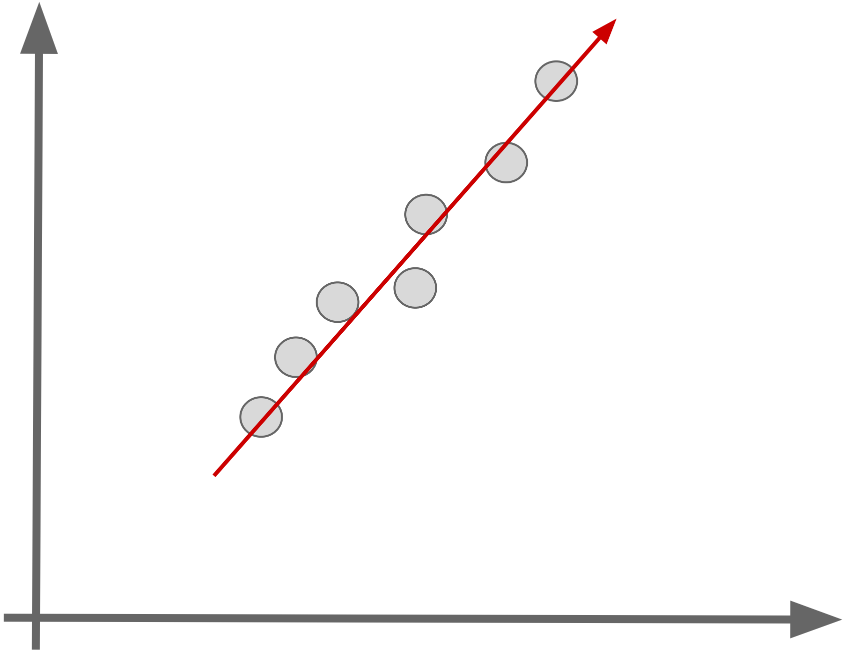
Example 1

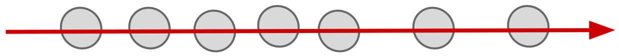
What happens?



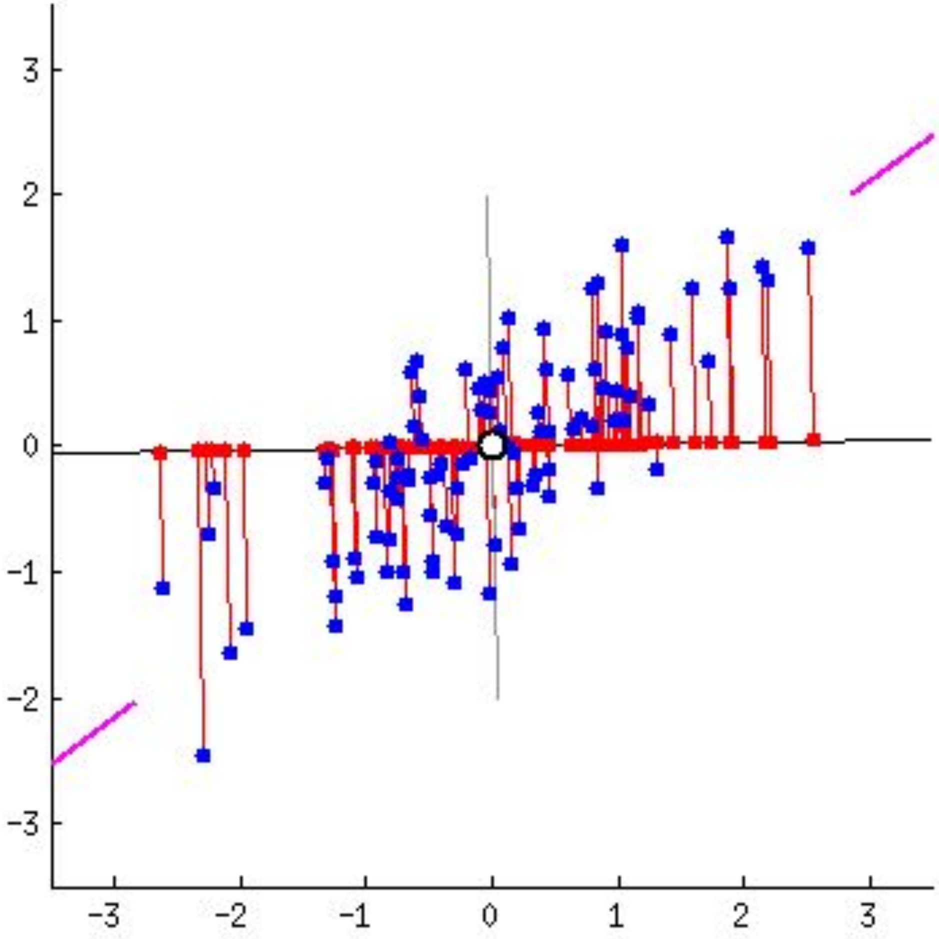


Example 2





Which one preserves more information about the original representation?


3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

More information

=

Explained Variance

How to choose the axis - Algorithm -

Covariance Matrix

=

Variance in the diagonal,

Covariance everywhere else

	Feature 1	Feature 2			Feature n
Feature 1	**var**				
Feature 2	**covar**	**var**			
	covar	**covar**	**var**		
	covar	**covar**	**covar**	**var**	
Feature n	**covar**	**covar**	**covar**	**covar**	**var**

What does a correlation matrix of uncorrelated features/axis look like?

Goal

high numbers in the diagonal,

small numbers everywhere else

Find Eigenvectors+Eigenvalues of the covariance matrix

Principal components = Eigenvectors associated with the largest eigenvalues

Easy, isn't it?

12 = 2x2x3

Matrix = (eigenvalue1, eigenvector1),

(eigenvalue2, eigenvector2)

Many mathematical objects can be understood better by breaking them into constituent parts, or finding some properties of them that are universal, not caused by the way we choose to represent them.

For example, integers can be decomposed into prime factors. The way we represent the number 12 will change depending on whether we write it in base ten or in binary, but it will always be true that $12 = 2 \times 2 \times 3$.

From this representation we can conclude useful properties, such as that 12 is not divisible by 5, or that any integer multiple of 12 will be divisible by 3.

Much as we can discover something about the true nature of an integer by decomposing it into prime factors, we can also decompose matrices in ways that show us information about their functional properties that is not obvious from the representation of the matrix as an array of elements.

One of the most widely used kinds of matrix decomposition is called eigen-decomposition, in which we decompose a matrix into a set of eigenvectors and eigenvalues.

https://stats.stackexchange.com/a/140579

https://deeplearning4j.org/eigenvector

What to remember

PCA

=

plotting in lower dimensional space

(usually 2d for visualization)

New features

=

linear combinations of old features

Best axis

=

Keep more information possible

(highest "explained variance")

```
PCA(n_components=2).fit_transform(X)
```

Useful for:

EDA,

removing multicollinearity,

manage strange datasets

It's hard (impossible?)

to interpret the new axis!