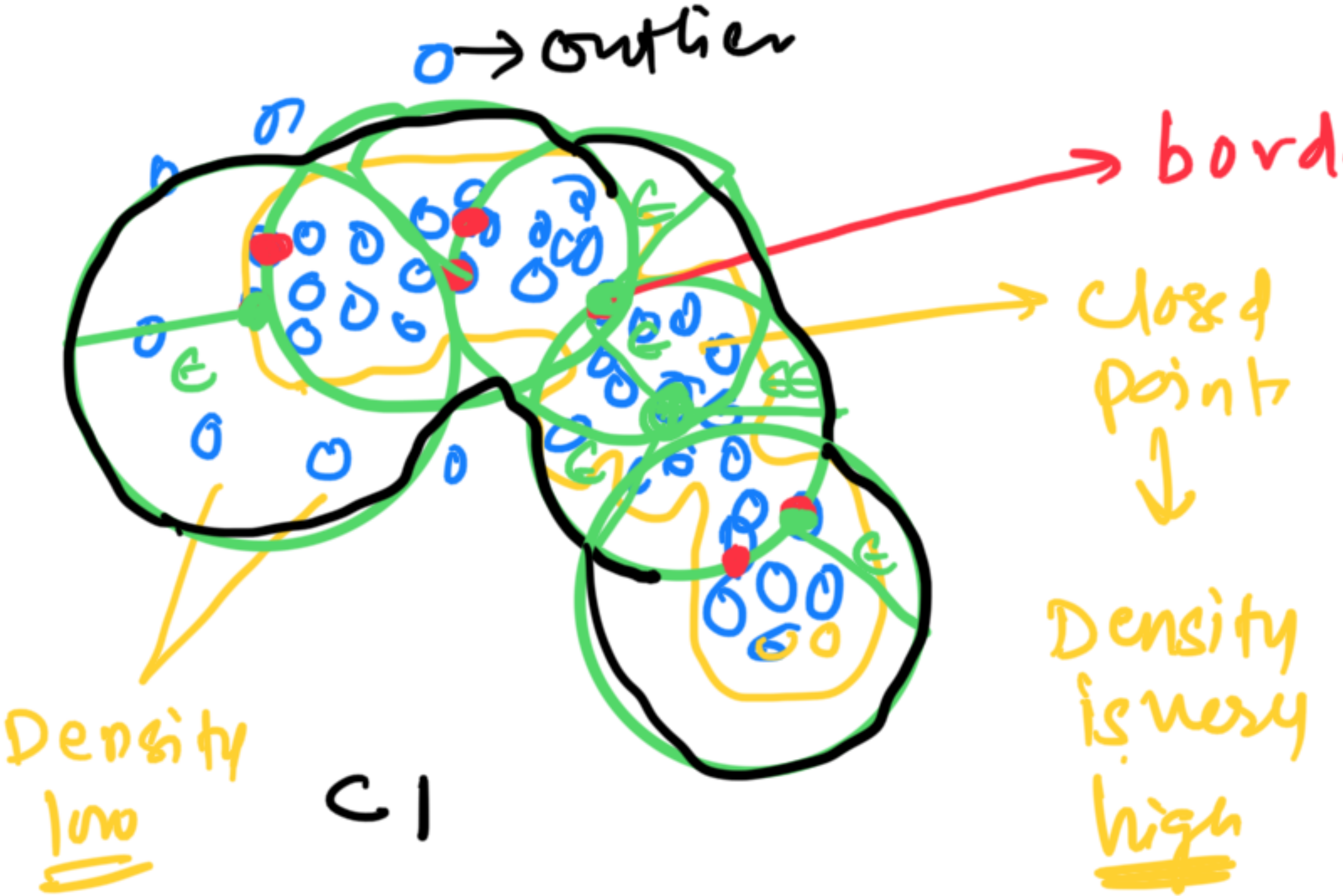
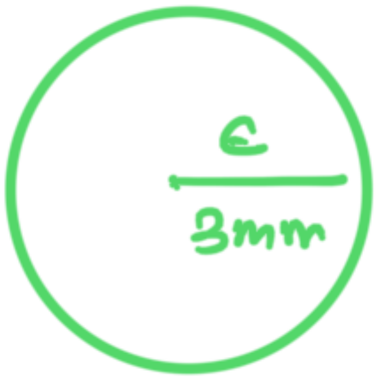
# DBSCAN Clustering



ε = 3mm (user)



1) hierarchy

2) distance

3) density — closeness → radius

↓

draw one circle ← ε

ε = radius

within

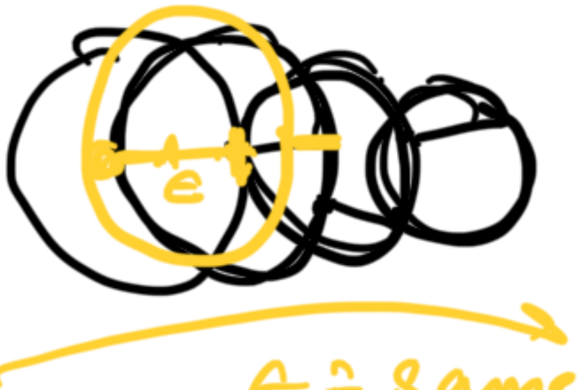
cluster

## Cluster

1) **Cluster points** → density
   - → ε draw ⨀
   - → points within ⨀

2) **Border points** → extension
   - → point on ⨀ boundary
   - → draw ⨀ with radius ε
   - → repeat this process to cover max data points

3) Noise or outliers



ε = 3mm

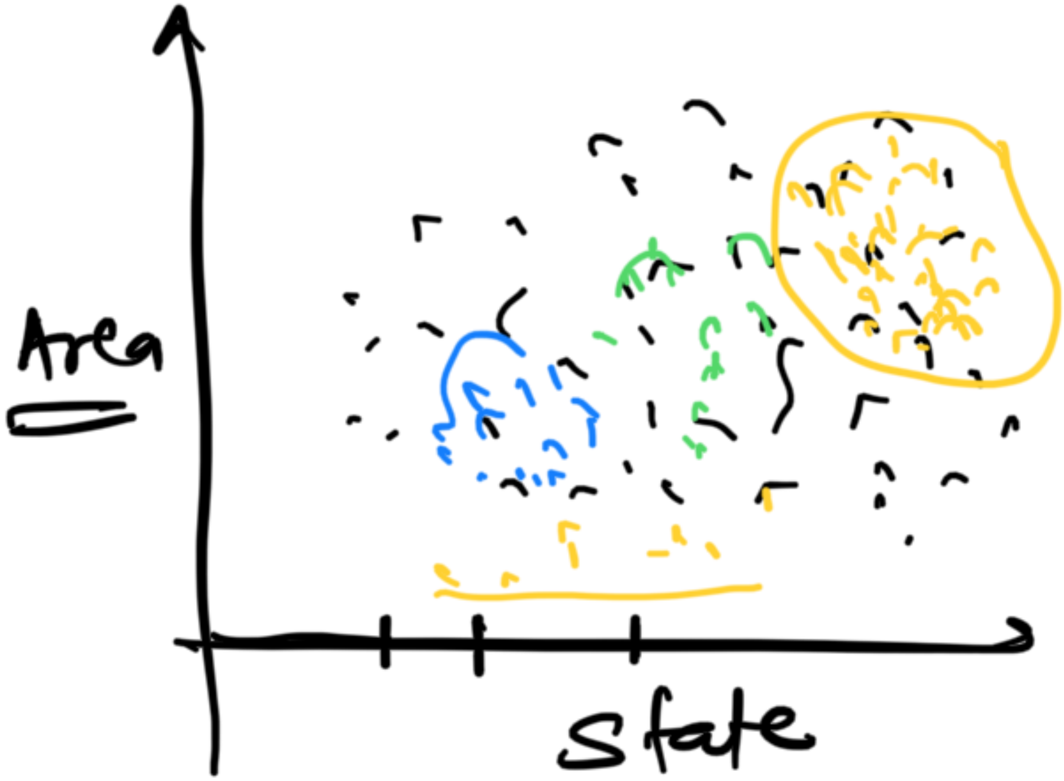
3mm / 5mm

DBSCAN → density based algorithm

DBScan → Density Based Spatial Clustering of Application with Noise

App → statewise density of languages
↳ DBSCAN

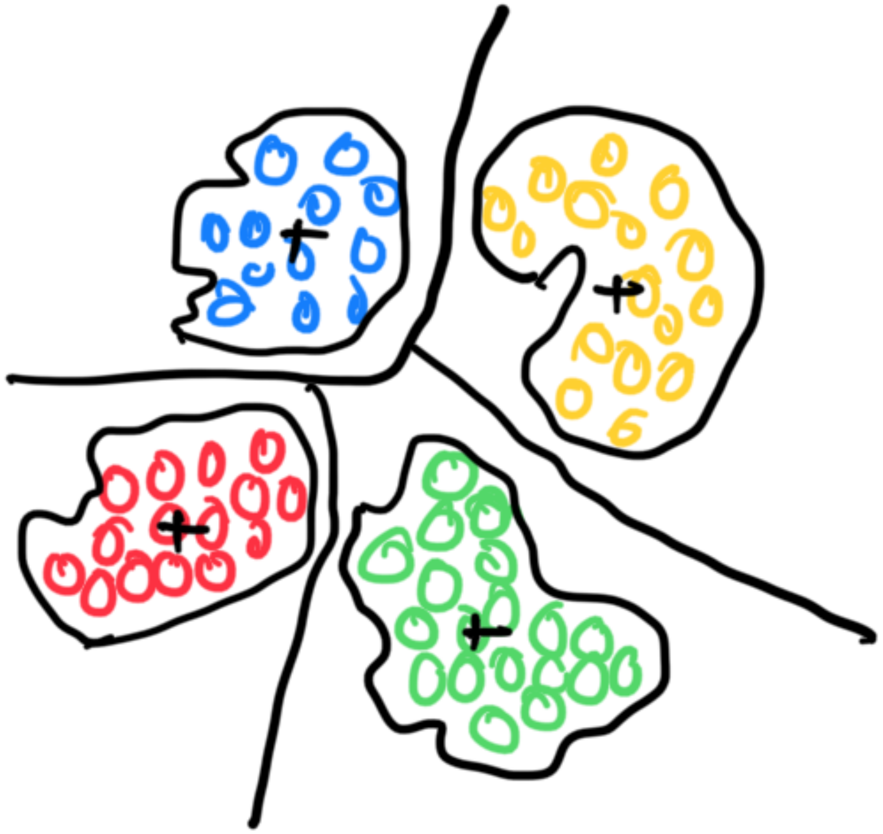
— Language → 5 → cluser
↳ Marathi
→ Hindi
Bengali
Tamil
Gujrati



Density = number of points within the radius of circle

($\epsilon$)

Basic Idea → Cluster the dense region



DBSCAN → Idea of density

↓

low/high

+

K Means

↳ Avg — mean

↳ Medoid → median ✓

↓

centroid

↓

K-medoid Algorithm

Values

Min - Max

⟶ Mode → Mode

$\boxed{\text{K-Mode}}$

df. describe()

$\underbrace{\text{min, max}}$, r, v

Range → 0-100 ⇒ mean

0, 200, 789, 450, 15, ⇒ medoid.

Mean = 5, 3, 4, 2, 1

$$\frac{5+3+4+2+1}{5}$$

=

Medoid = 100, 99, 5, 7

## DBSCAN - Concepts

1. Core points - data points → within ⊙
   - → cluster points
   - → Radius (∈) ⇒ ∈ value should be same

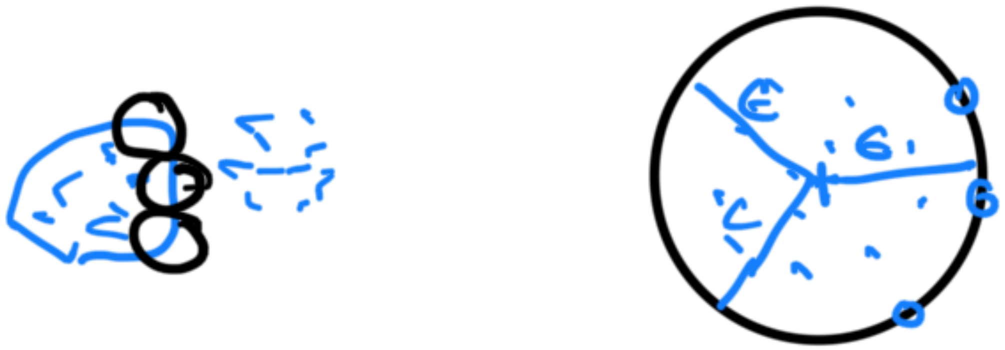
2. Border points - Neighborhood points

Next start for $\epsilon$ ⊙

3. Noise — Outliers → Discards $\{-1\}$

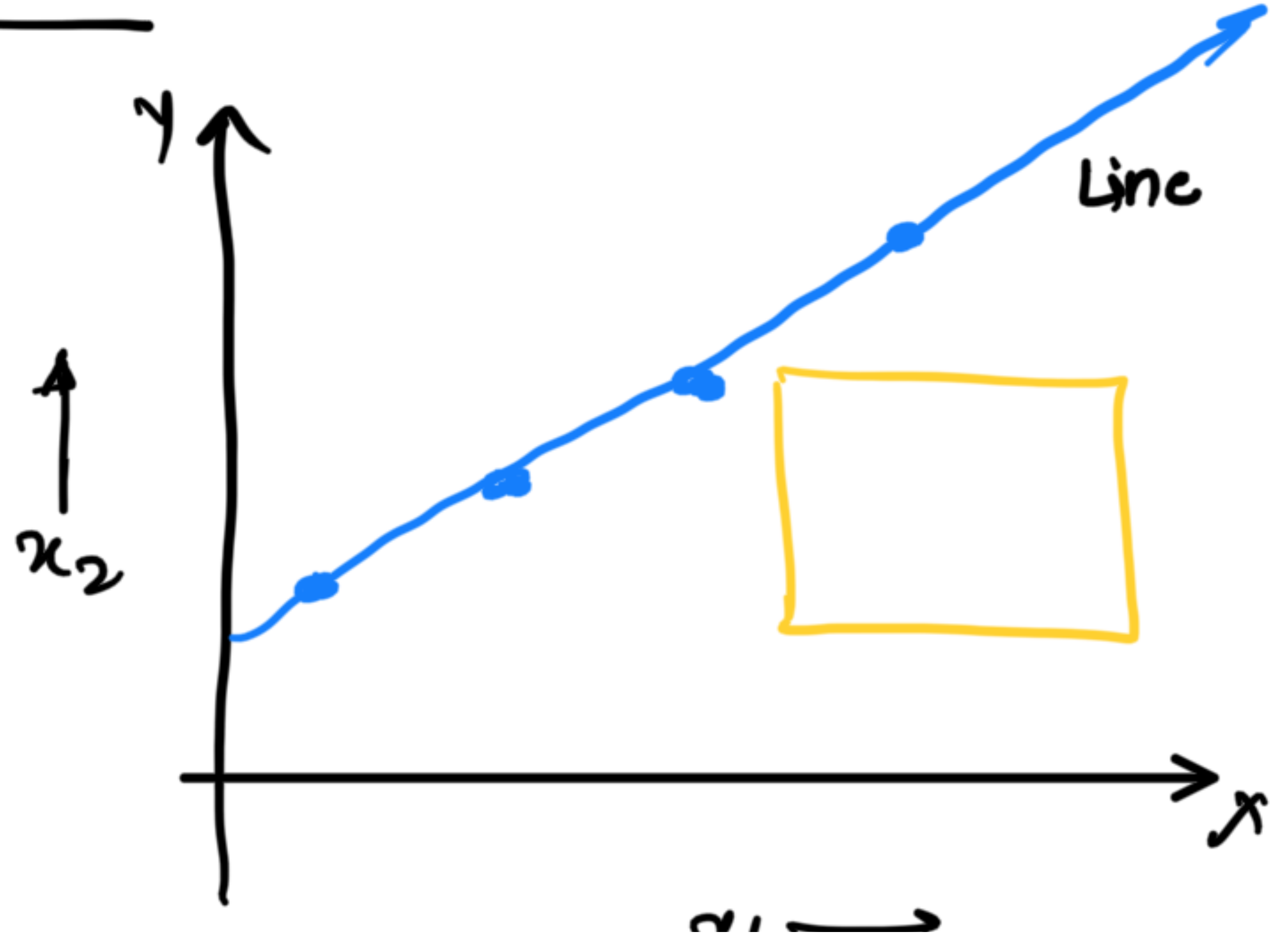
↑

DBSCA
K-Means

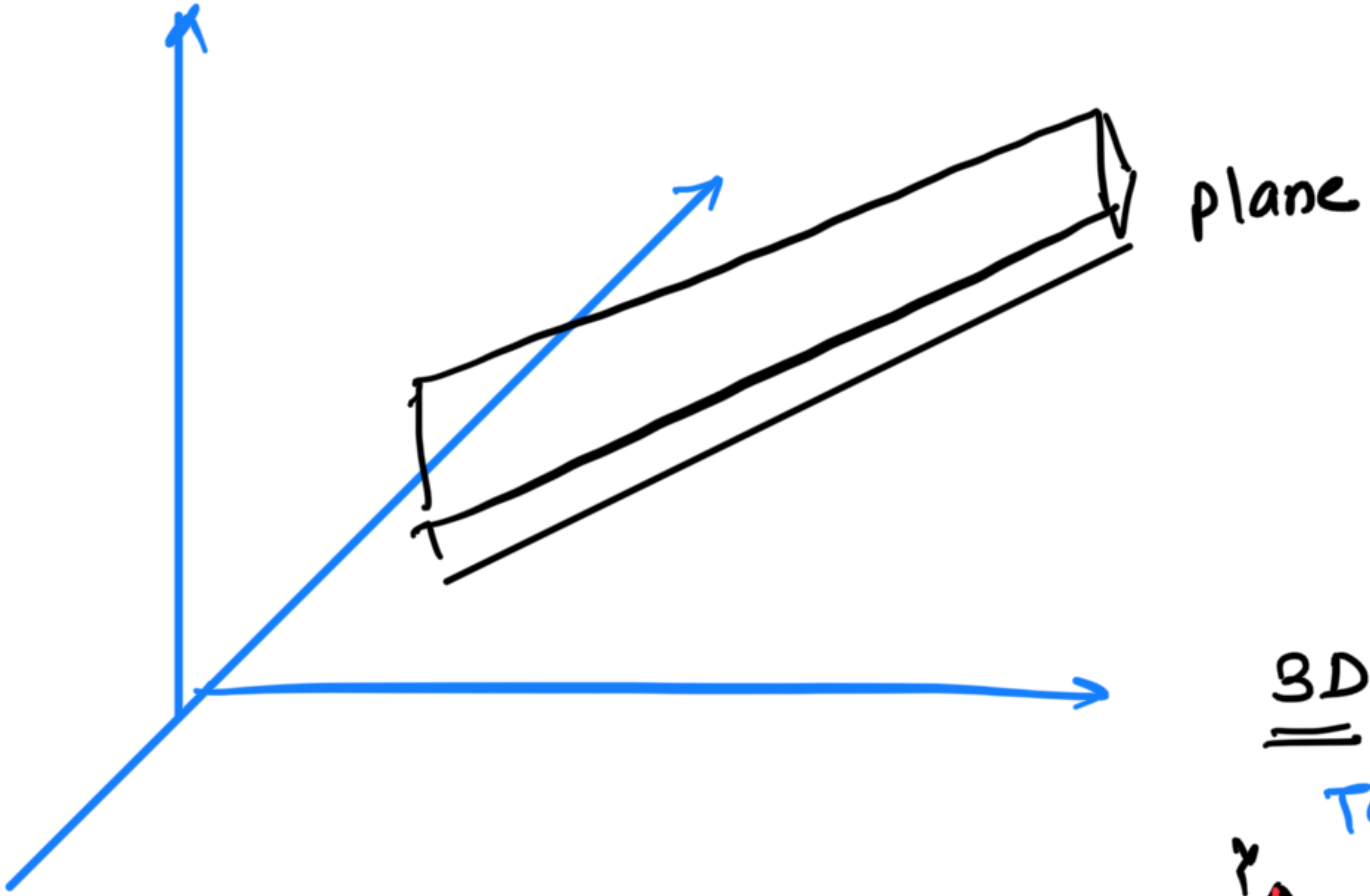


## Dimension Reduction

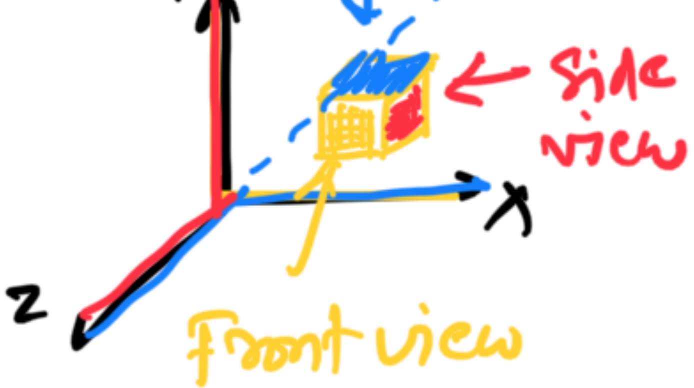
| $x_1$ | $x_2$ |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 6 |
| 4 | 8 |

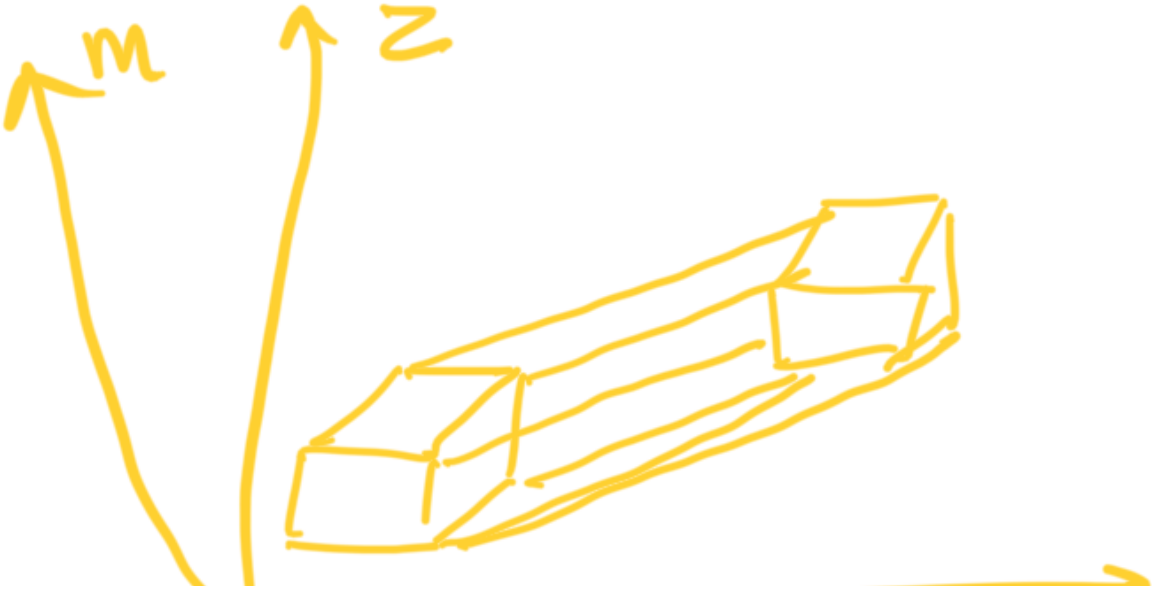


## 2D ✓



## 3D ✓





10, 100, 1000 → visualization high dimension is very difficult

↓

Complexity

↑

Imagination

## Dimension Reduction

Ex 1

- Documents → multiple page
  (words)
  ↳ multiple words
  ↳ multiple letter

Alphabets ⇒ $\cancel{26} \times \boxed{\underline{300} \times \underline{15} \times \underline{30} \times \underline{6} \,/\! \times 12}$

= $\cancel{x}$ ⋮ ?

↑ Huge number

↑ words ↓ features

Dataset

| | the | for | sunset | - | - | . | . | - - | - | . | . | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 → | | | | | | | | | | | | |
| P2 → | | | | | | | | | | | | |
| P3 → | | | | | | | | | | | | |
| . | | | | | | | | | | | | |
| Pn → | | | | | | | | | | | | |

Lakhs

Difficult task

High dimensional data

Image ⇒ ?

daily man words

P1 P2 P3 ···P15

Ex. 2 Image

450
↓
× 300

Bank ⇒ SBI →
↓
105 → centers in Country

25 → States

10 → cities

10 → city

→ In each center

↪ 10,000

105 × 25 × 10 × 10000 ⇒ Customers ?

↓

Huges of data

Eg. Images → Traffic Control ⇒ Cameras → 10L

↓

24×7

↓

1 Lac

10L × 24 × 7 × 1 Lac

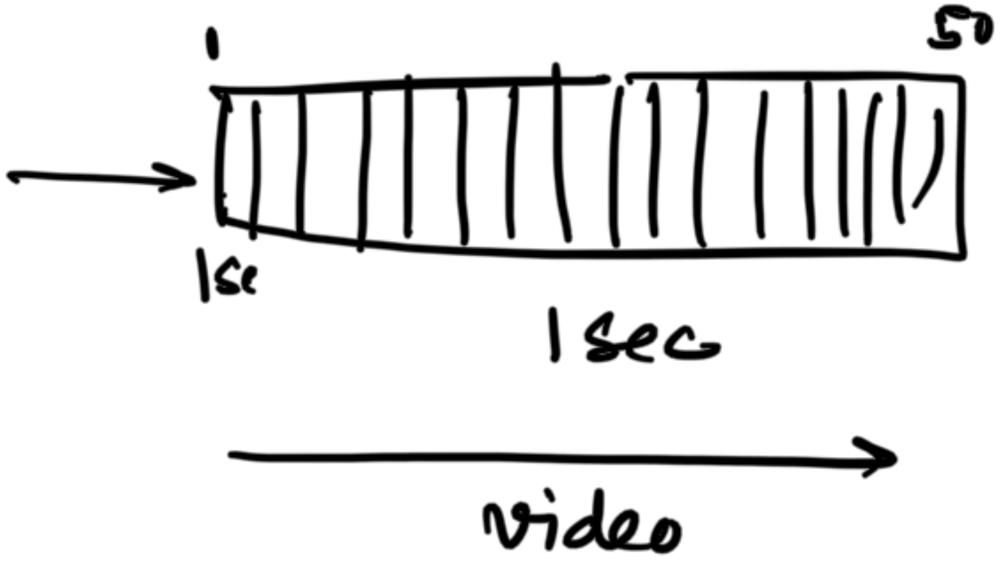
= Image dataset

= |

1 Camera → 1 day → 1 Lac

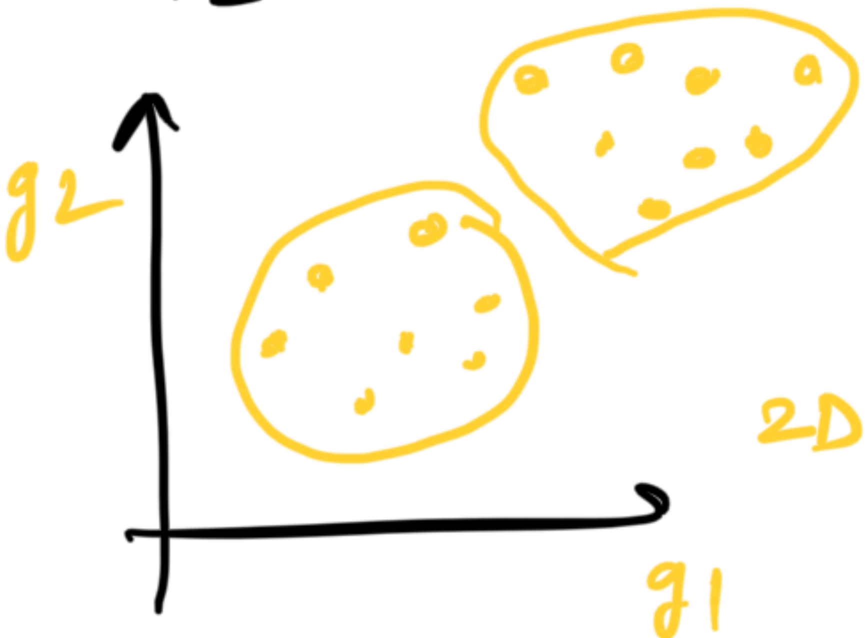
7 days → 7lac

10L × 7Lac ⇒ 70lacs

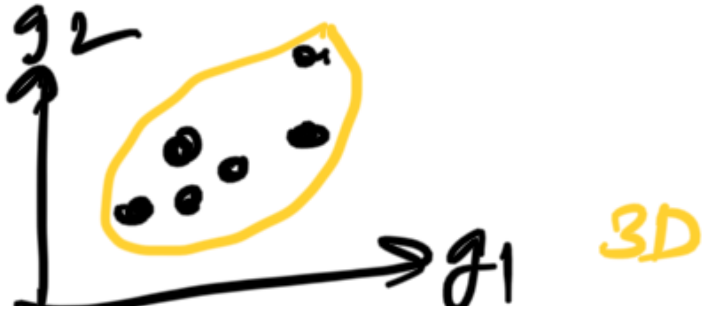


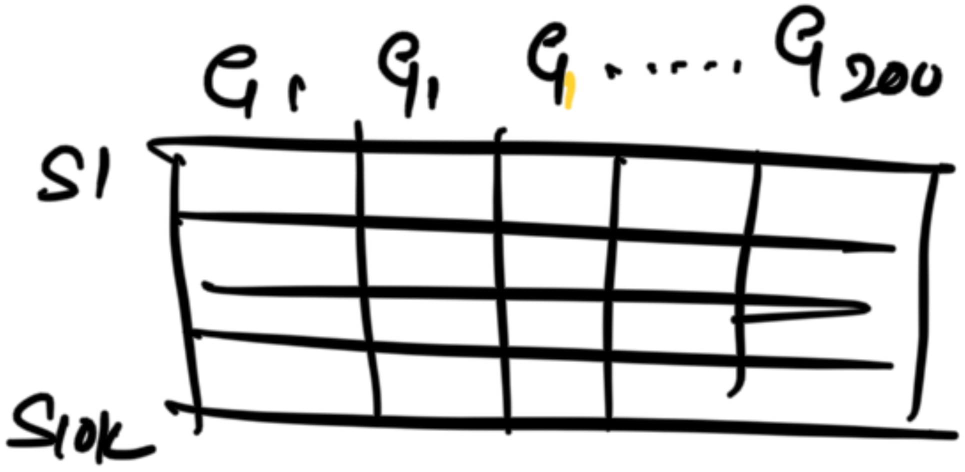
Eg. Genetics → Gen. → $G_1, G_2$

|  | $G_1$ | $G_2$ |
|---|---|---|
| S1 |  |  |
| ⋮ |  |  |
| S50 |  |  |



|  | $G_1$ | $G_2$ | $G_2$ |
|---|---|---|---|
| S1 |  |  |  |





→ Dimensions more

1) meaningful visulization

2) how to extract that meaning informations

Lung Cancer ⇒ ?

Tumor →

> 10000

↓

2,50,000 → each cell.

↑

staging

Tum

2,50,000 × 1000.

↓

→ solⁿ ⇒ Dimension Reduction

1) Convert your high dimensional data
into lower dimensional data

a) Visualize the data

b) Analyze the data

Dimension Reduction → loss of relevant inform.

minimum loss of relev. info

## Approaches to DR -

1) **Feature Selection**

- Select subset of existing features

select subset of existing features

2) <u>Model Regularization</u> ✓

- L2 <u>reduces</u> → dimensionality

effective → L2

actual → L1 ✓

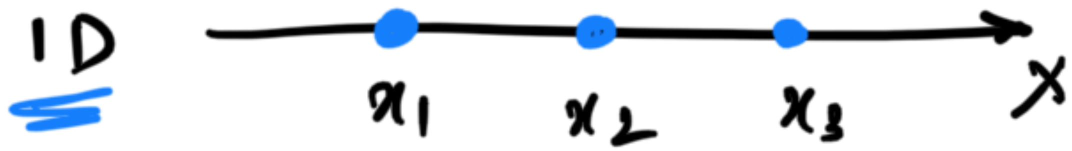
3) <u>Combining of existing feature</u> into smaller no. of new features

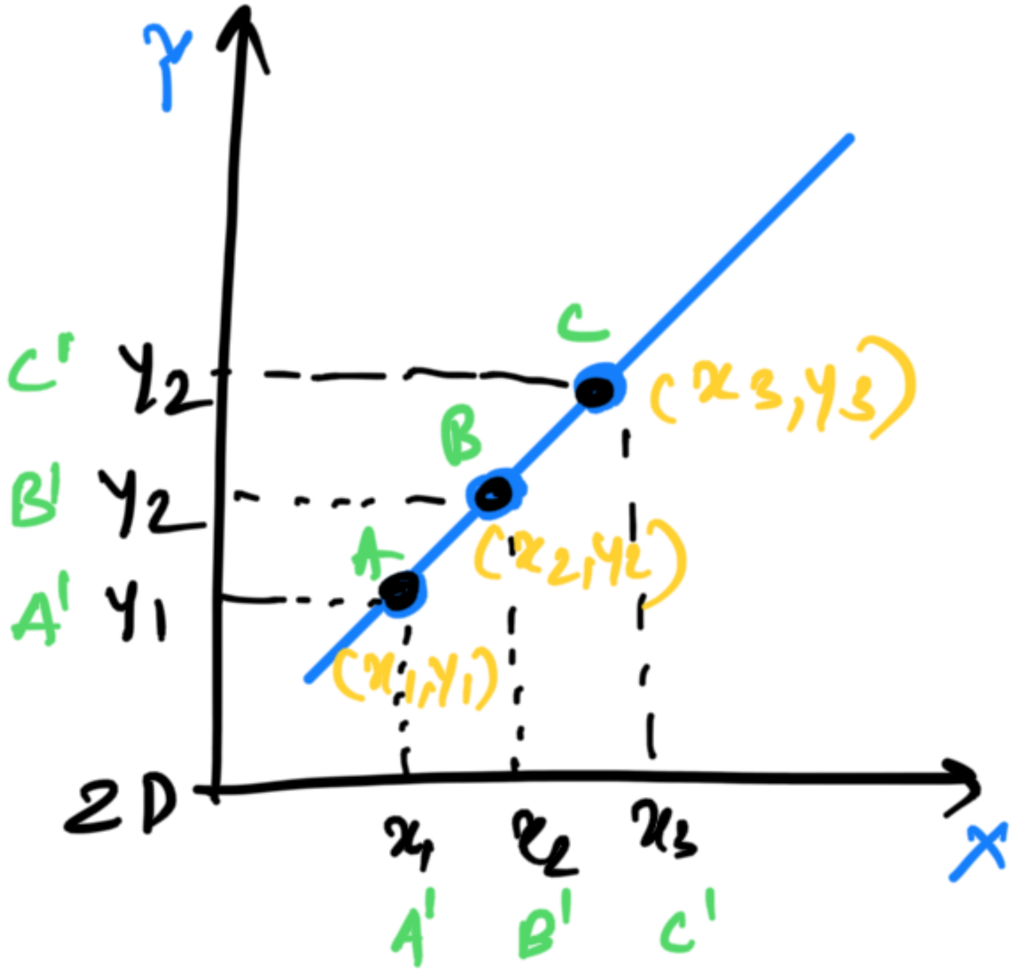
- Projection (Linear combination)

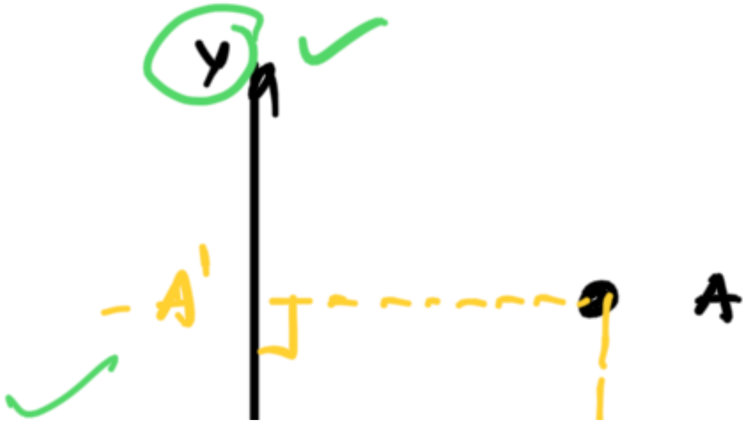
- LDA, <u>PCA</u>, Kernel PCA

## ③ Combining of existing features into new feature

### Projection

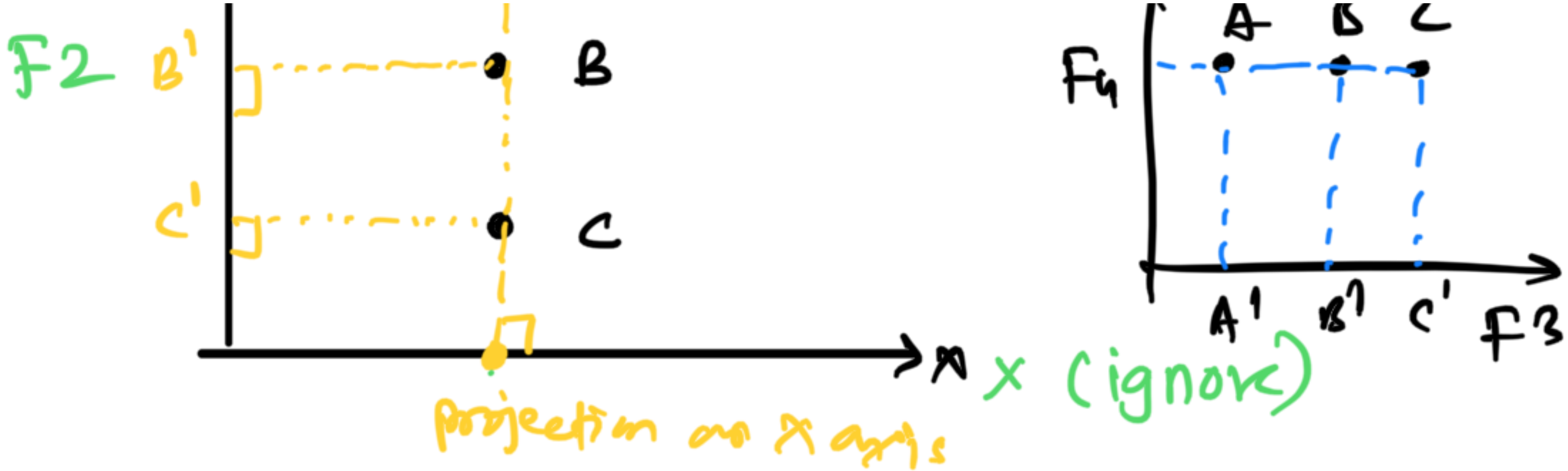






Linear Discriminant Analy

LDA

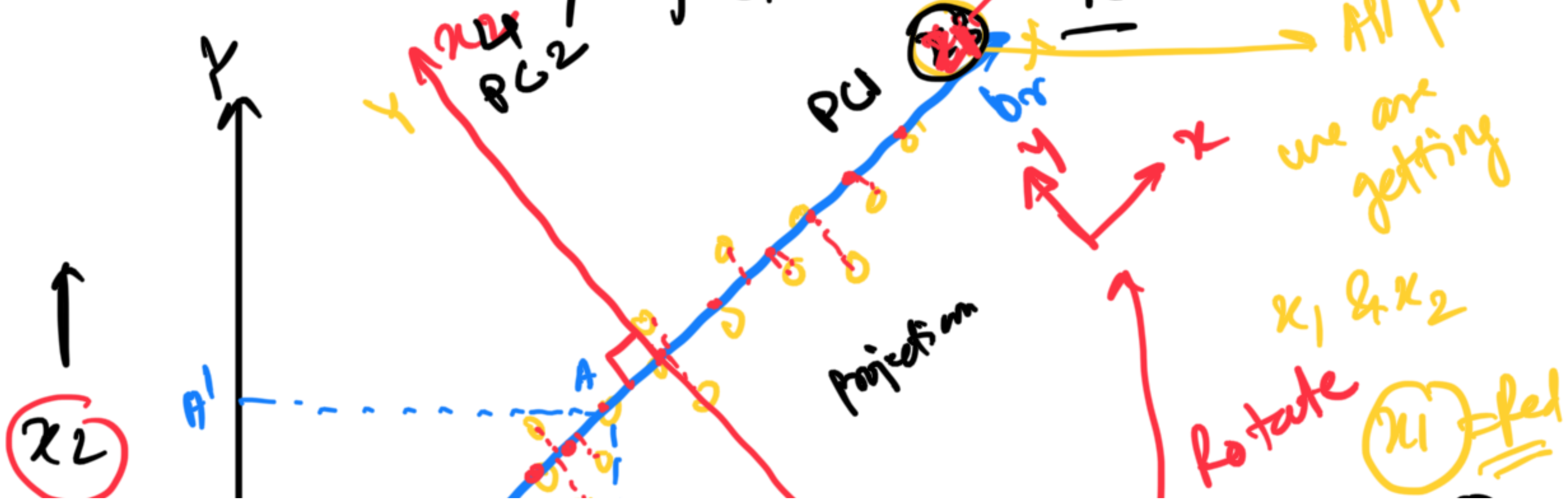


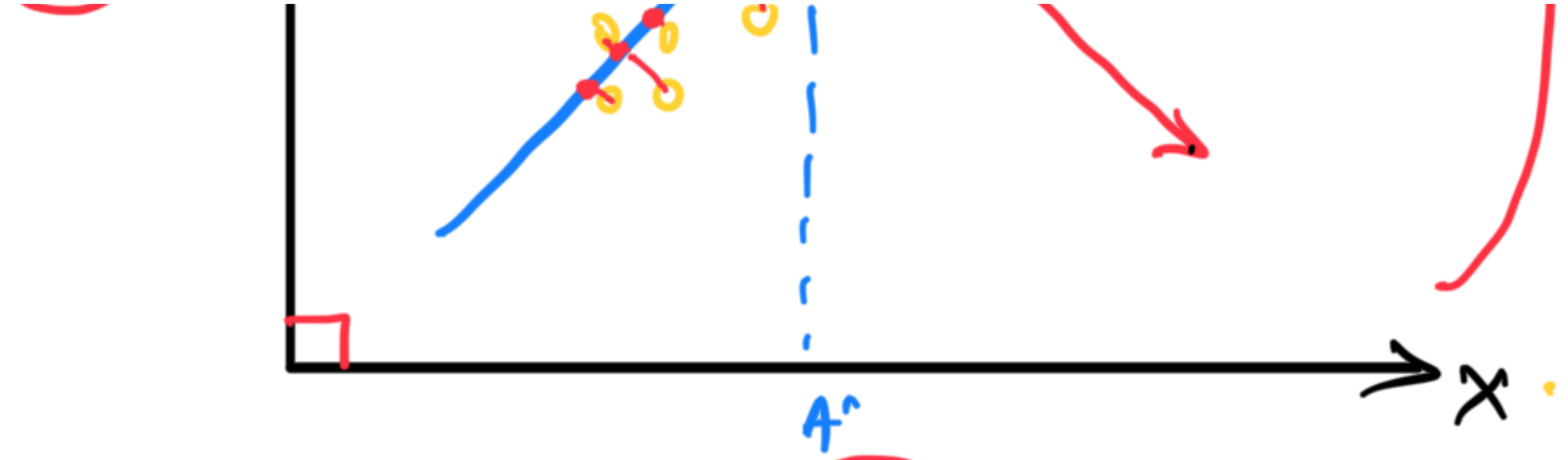
F1 (Ignore)
F4 ✗

F3
F2 } ✓

PCA - Principal Component Analysis

— linear projection — 90°

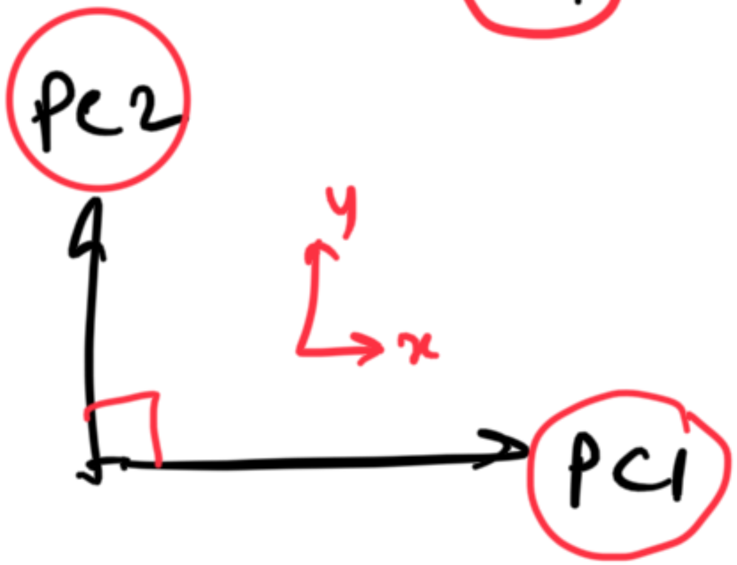




$\boxed{x_1} \longrightarrow$

$x_3 \Rightarrow \boxed{x_1}$

$x_1 \oplus x_2 \rightarrow \boxed{x_3'}$



2 comp

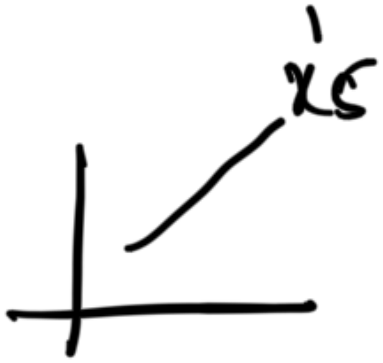
$x_1 \oplus x_2 \longrightarrow x_1' - \boxed{PC1}$

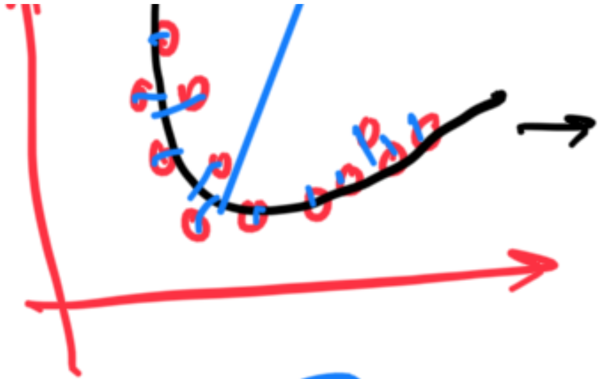
$x_3 \oplus x_4 \longrightarrow x_2' - \boxed{PC2}$

$x_1 + x_2 = \underline{x_1 + x_2}$

Day1 Math Day1Math

$x_1'$ vs $x_2'$





→ Polynomial → PCA
Regr

## ICA