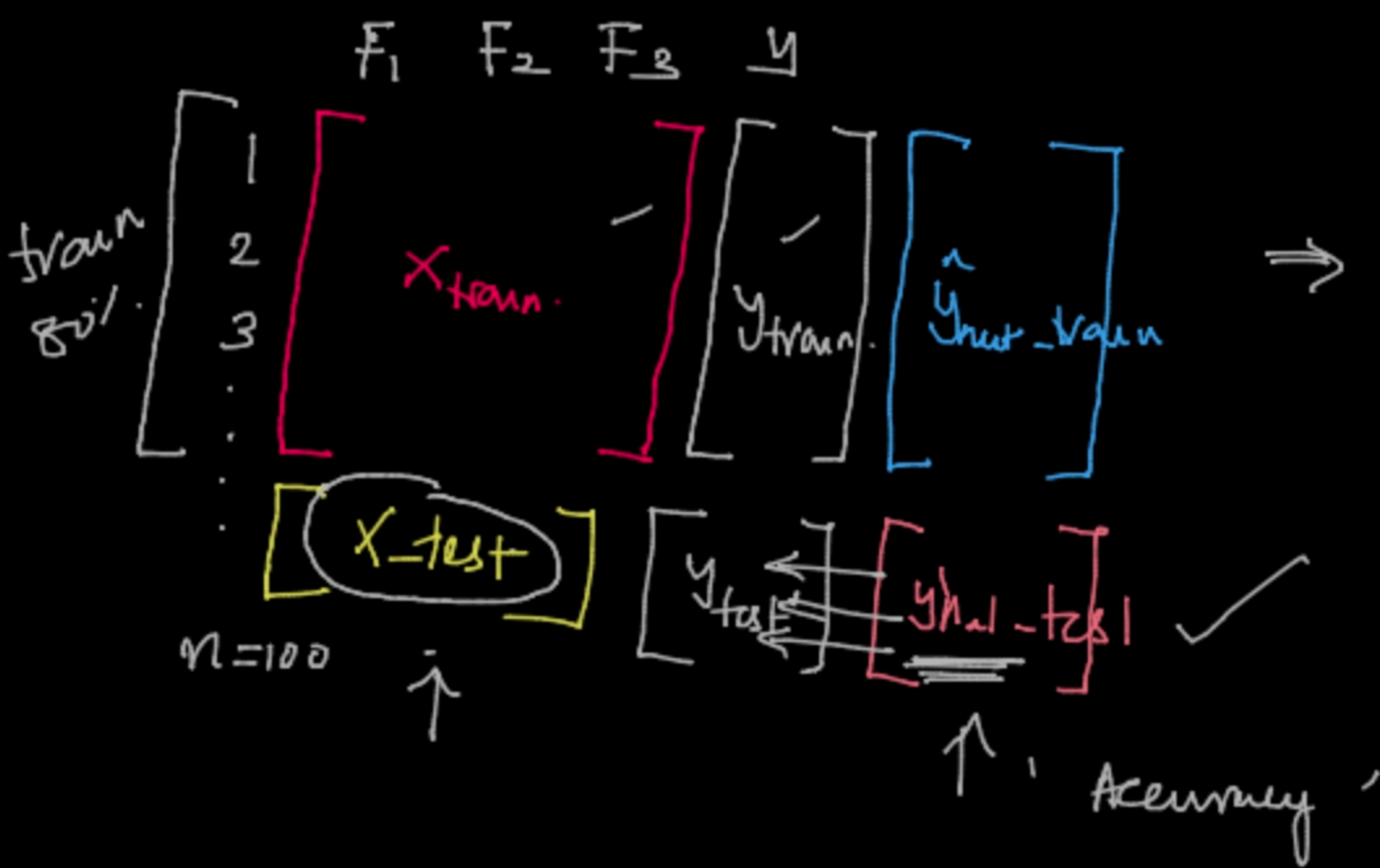
Data we have



Train_test_split (

model = logReg ← 80% →

model.fit(X_train, y_train)

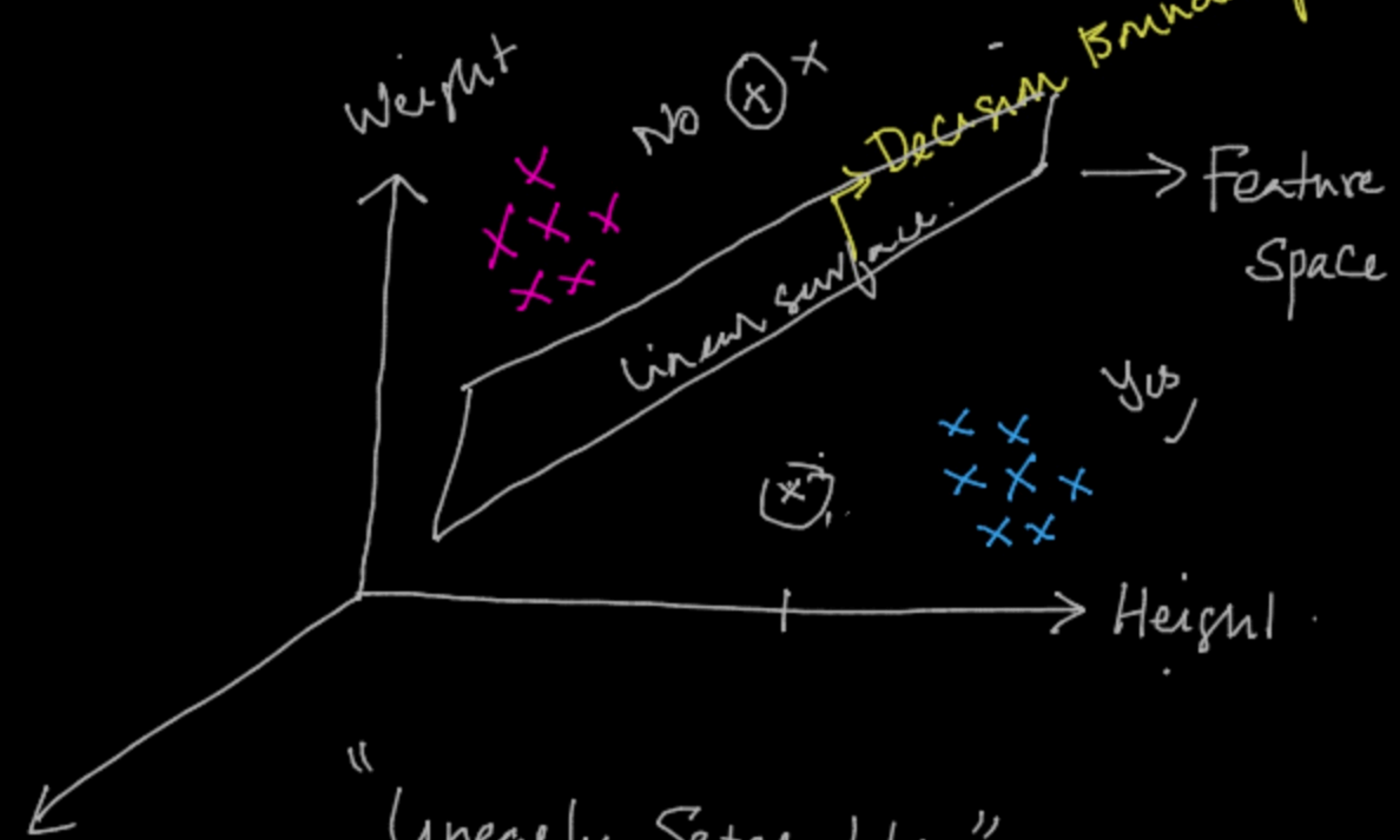
model.predict(X_train) ⇒

model.predict(X_test) ⇒

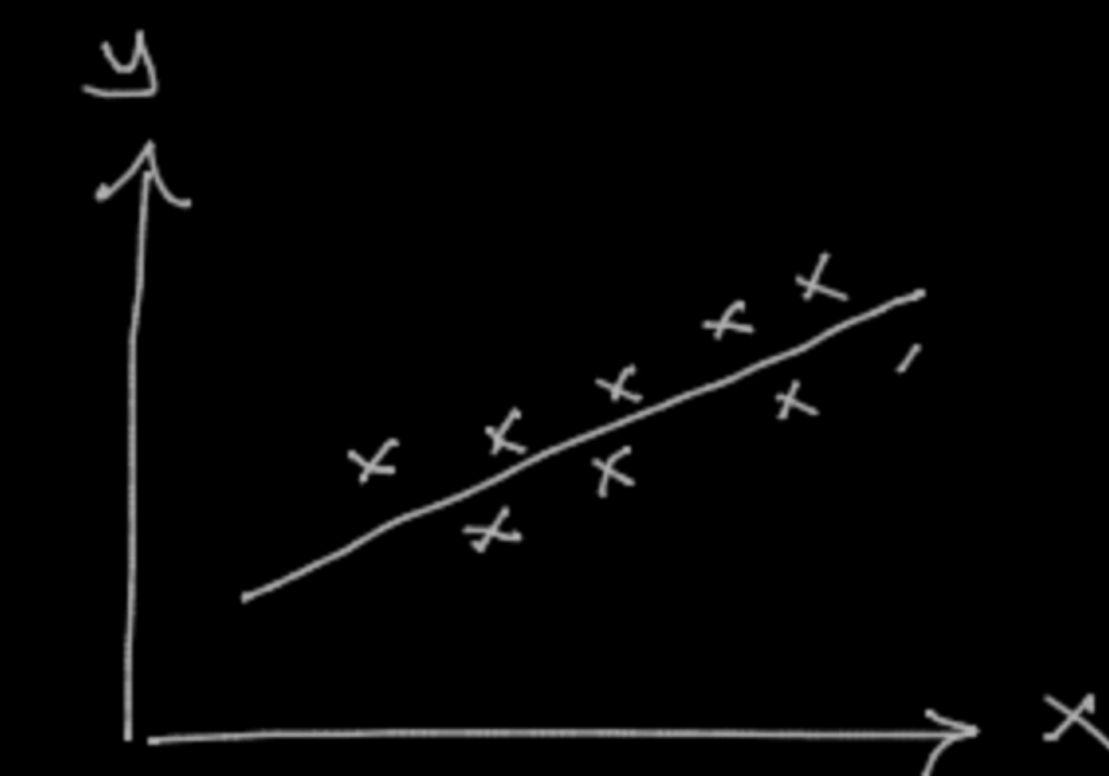
Unseen

logistic Regression – "Linear Model"

Classification → "Partitioning the feature space into pure regions assigned to each class"

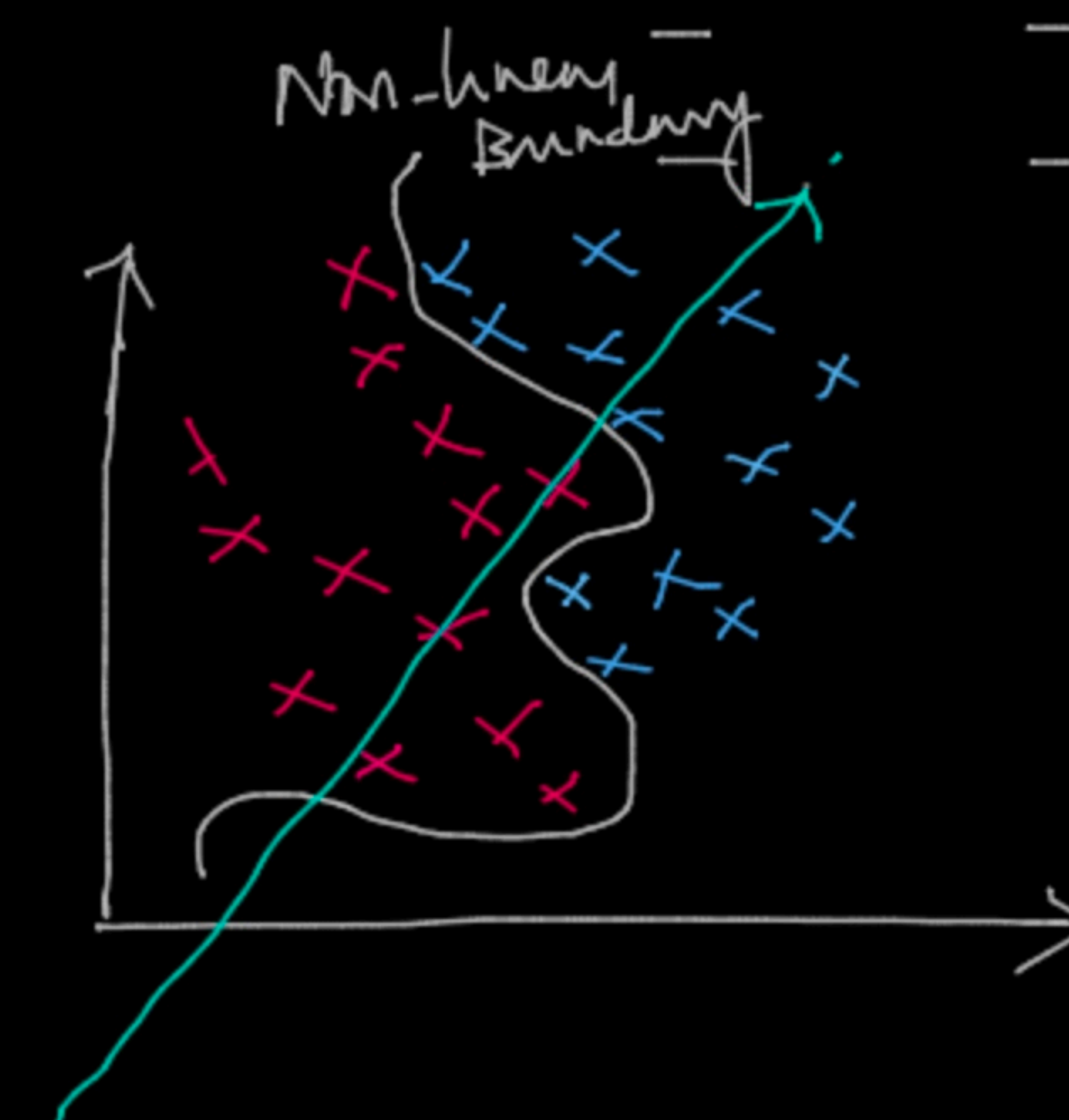


"Linearly Separable"



→ Dimensionality

Chd	Wc	BMI	Height (x_1)	Weight (x_2)	Select
			—	—	✓
				—	✗



Clustering

— unsupervised ✓

— No 'y' Value.

— Clusters of similar Data.

$y - \hat{y} =$ Regre

$y \; \hat{y} \Rightarrow$ Accuracy.

Good cluster ?

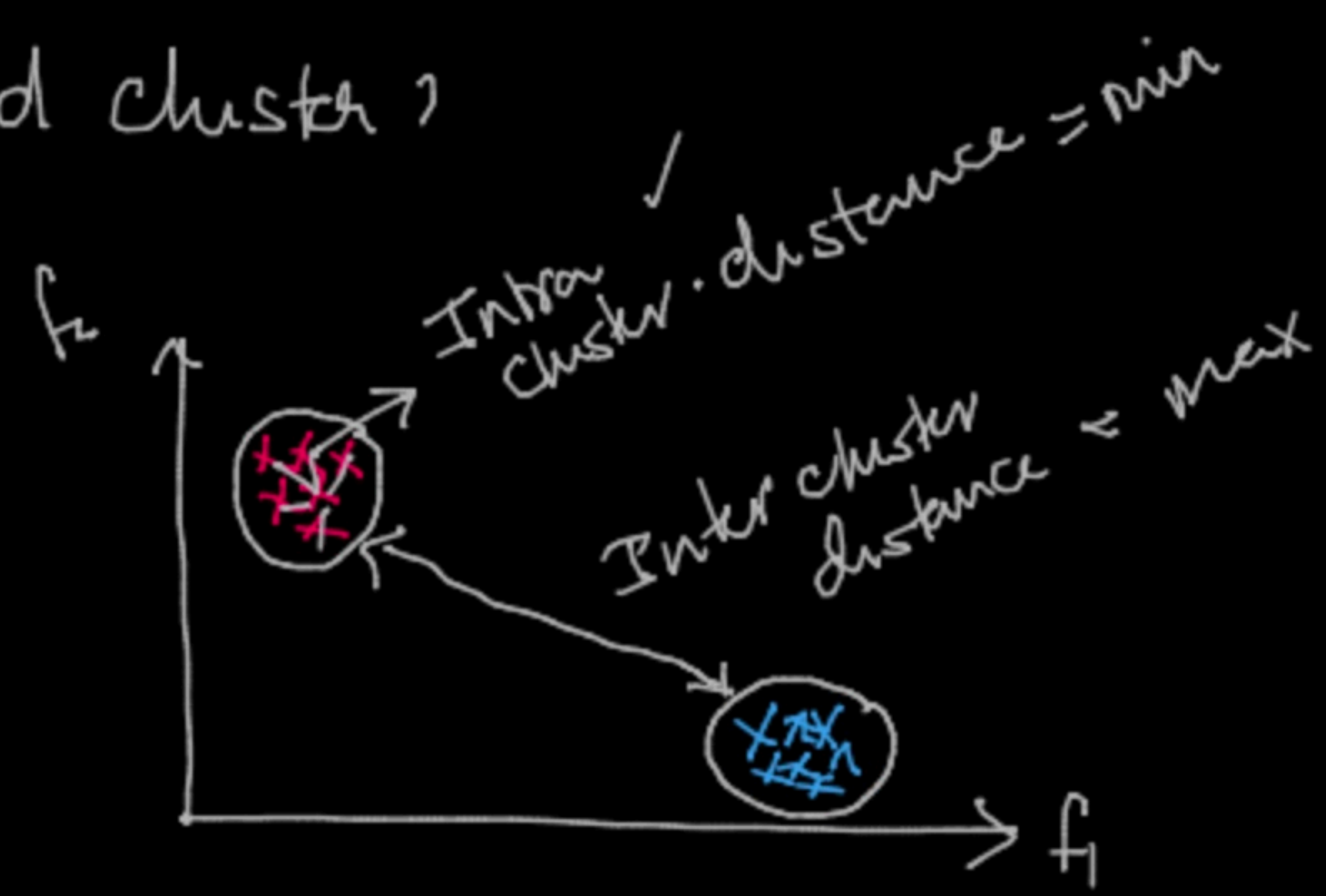


fig 1

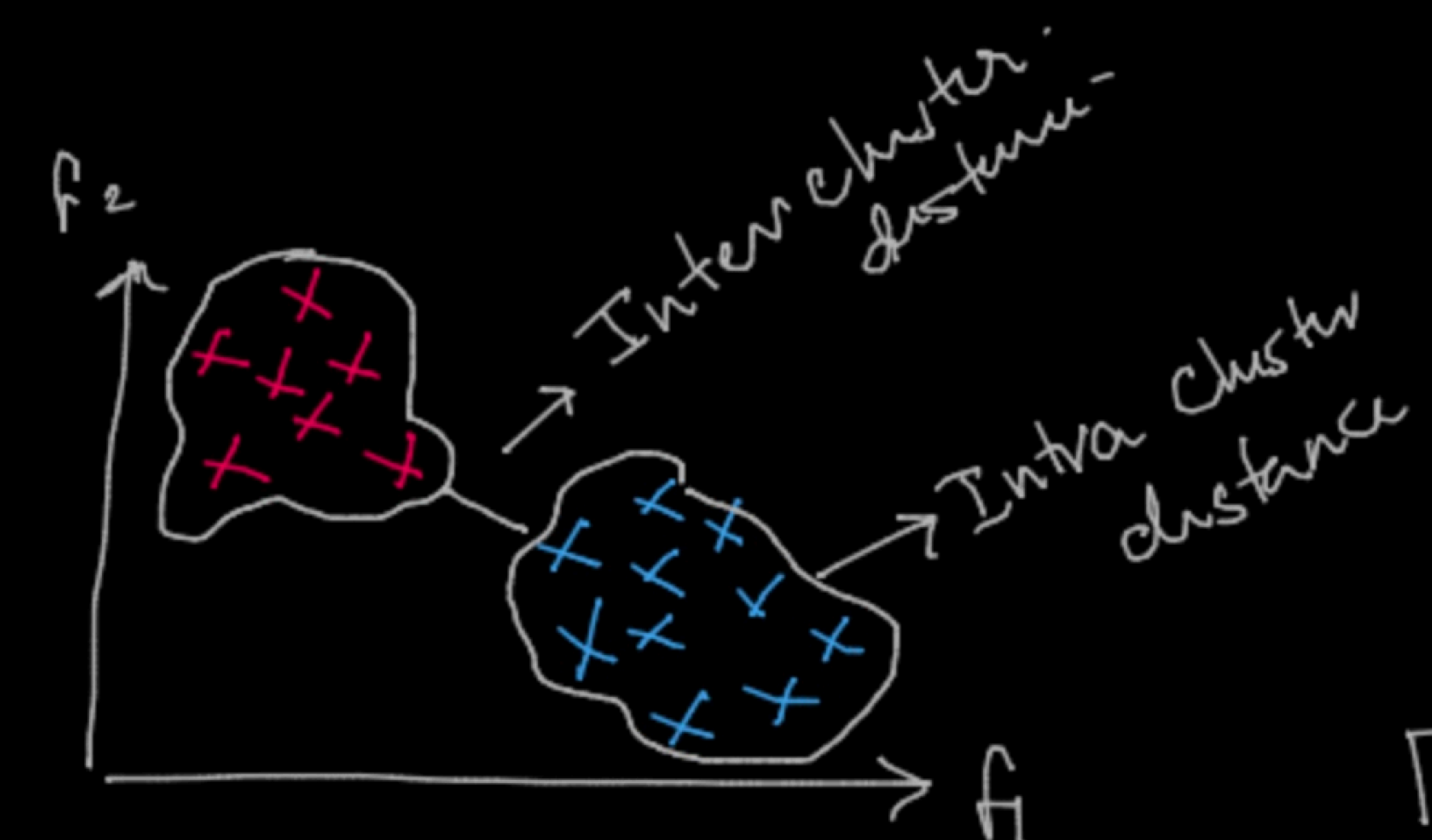


fig 2

Supervised
- Regression
- Classification

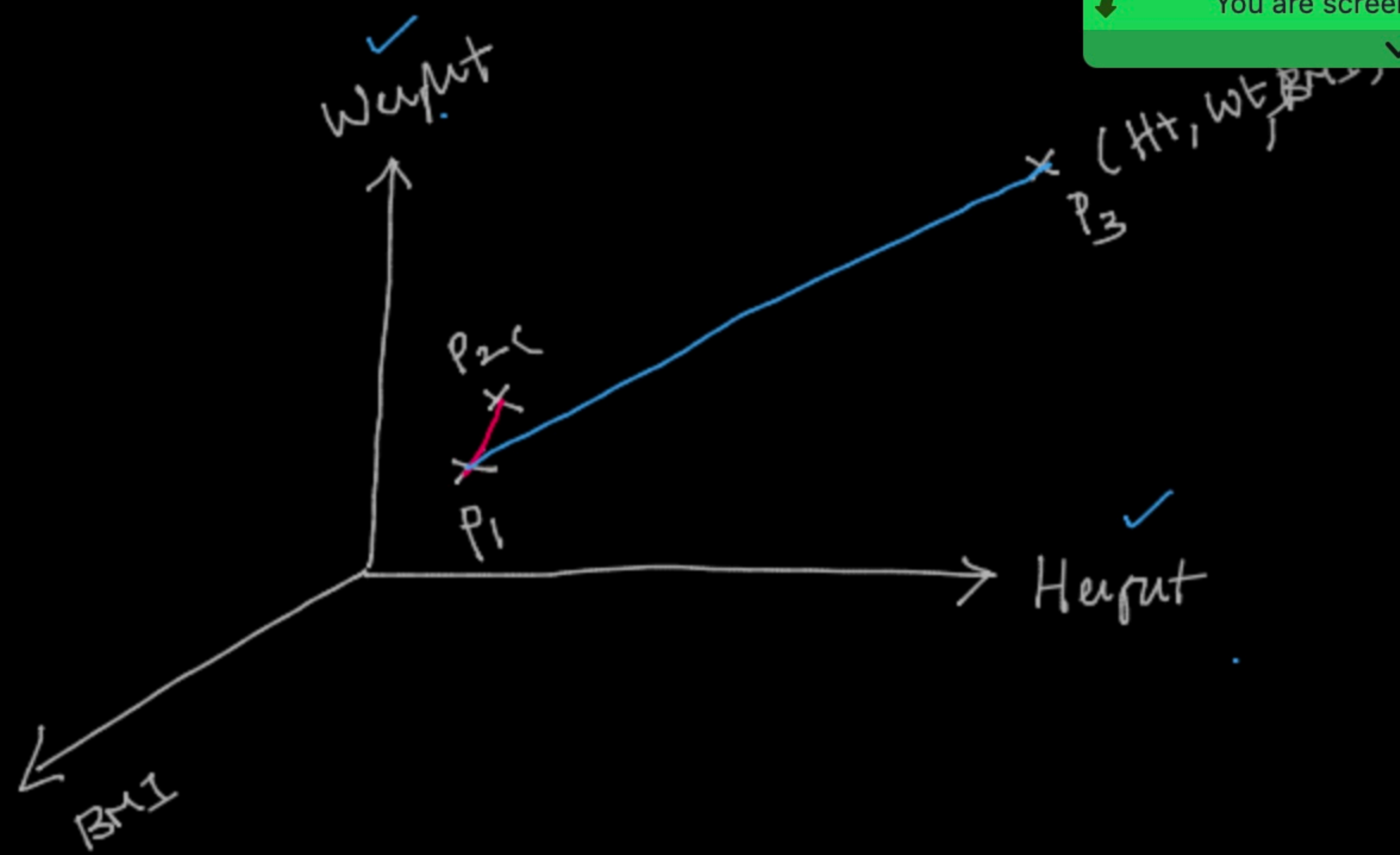
Unsupervised
- ✓ Clustering

Dunn Index $\Rightarrow \left[\frac{\text{Inter}}{\text{Intra}}\right]$

✓ 1 Minimum Intra cluster distance (WCSS)

— Within Cluster sum of squared distances.

✓ 2 Maximum Inter cluster distance



Distance $\Longleftrightarrow$ Similarity

$$\text{Distance} \propto \frac{1}{\text{Similarity}}$$

	f_1	f_2	f_3	f_4	
P_1	Nr	Caty	Nr	Nr	} Mixed Data.
P_2	Nr	Cat	Nr	Nr	
	Nr	Nr	Nr	Nr	} Numeric
	Nr	Nr	Nr	Nr	
	Cat	Cat.	Cat.	Cat	} only categorical
	Cat	Cat	Cat	Cat	

Numeric

1. Euclidean Distance ✓
2. Manhattan distance ✓
3. Minkowski distance ✓
4. Mahalanobis dist

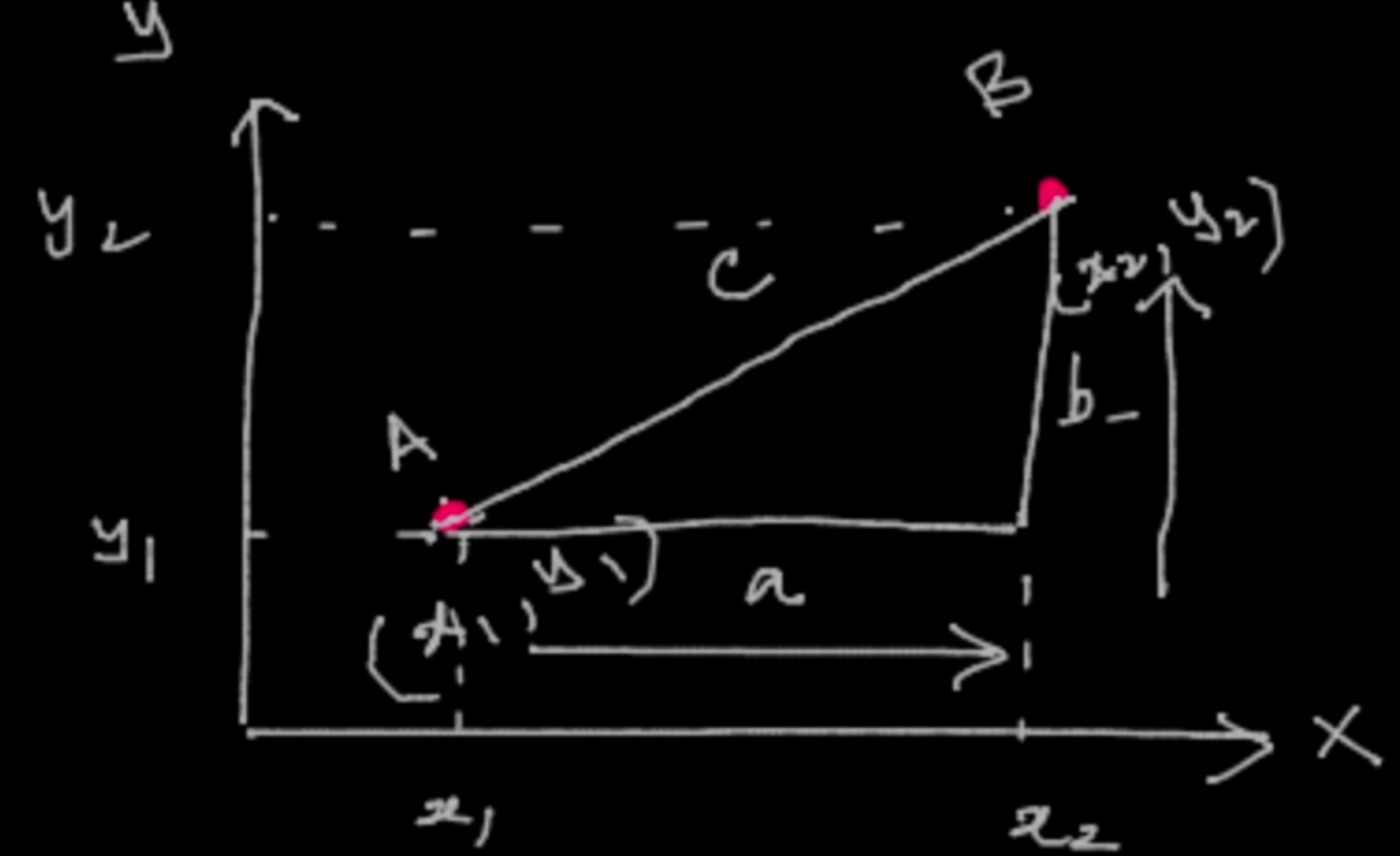
Categorical

1. Binary Euclidean.
2. Simple Matching Co-efficient
3. Jaccard's dist

Mixed

1. Gower's dissimilarity Index

Numeric



1. Euclidean Distance

$$c = \sqrt{a^2 + b^2 + c^2 + \cdots}$$

$$= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 + \cdots}$$

"As the crow flies"

2. Manhattan distance

$$c = a + b.$$

"Taxi Distance"

3. Minkowski distance.

$$c = \left[(x_2 - x_1)^p + (y_2 - y_1)^p + \cdots\cdots\right]^{1/p}.$$

$p = 1$; Manhattan.

$$c = (x_2 - x_1) + (y_2 - y_1)^1$$

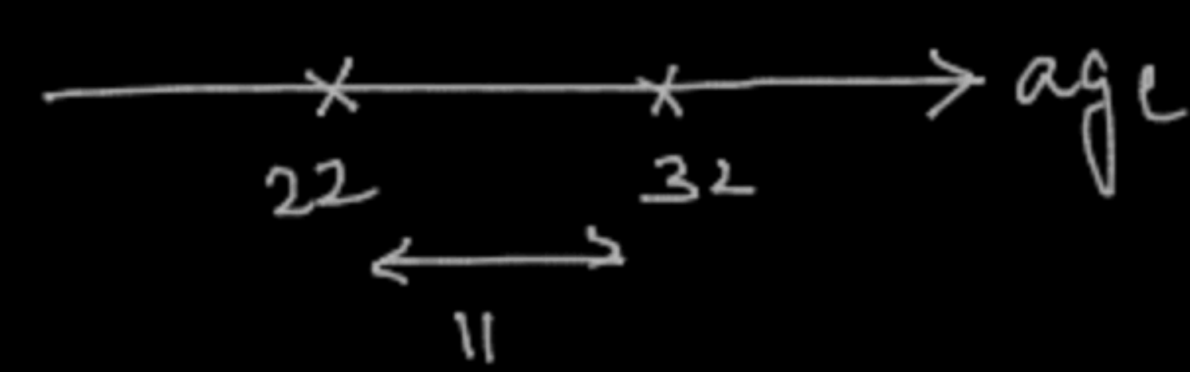
$$a + b$$

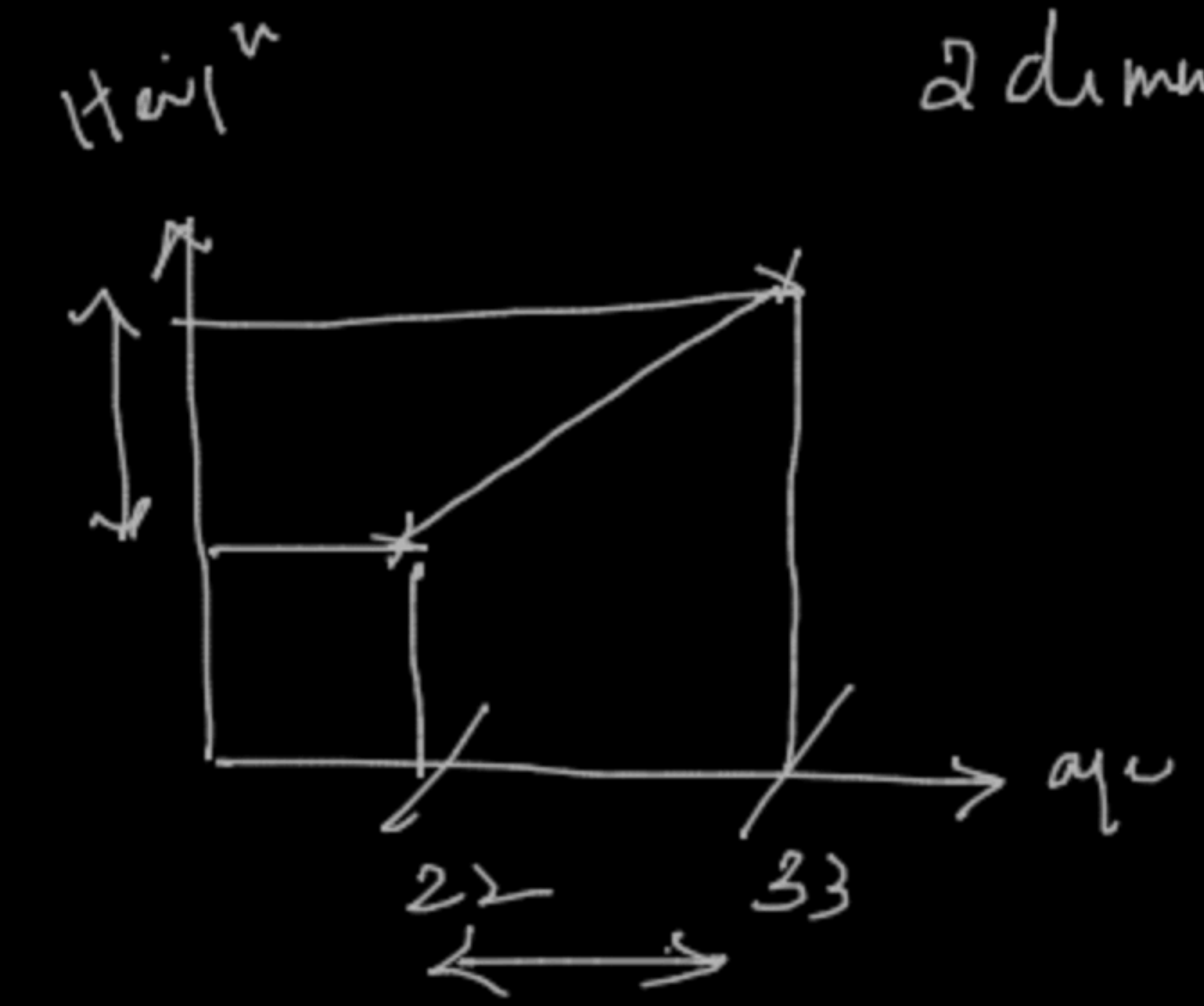
$p = 2$; Euclidean

$$c = \left[(x_2 - x_1)^2 + (y_2 - y_1)^2\right]^{1/2}$$

$$=$$

2 dimensi





$$= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

	age
P_1	22
P_2	33

11 yrs

	age	Height
$P_1 \rightarrow$	22	150
$P_2 \rightarrow$	33	167

You are screen sharing Stop Share

Talking: Geethika

← Numeric →

Age Income height Wt

P_1

P_2

P_3

Euclidean

Manhattan

← Categorical →

Married Manager Smoker gender

P_1

P_2

P_3

Binary Euclidean

Simple matching coeff

Jaccard's

	Id
	Age
P_1	–
P_2	–
P_3	–

← Mixed Data →

Age Income Gender Married height Wt Smoker

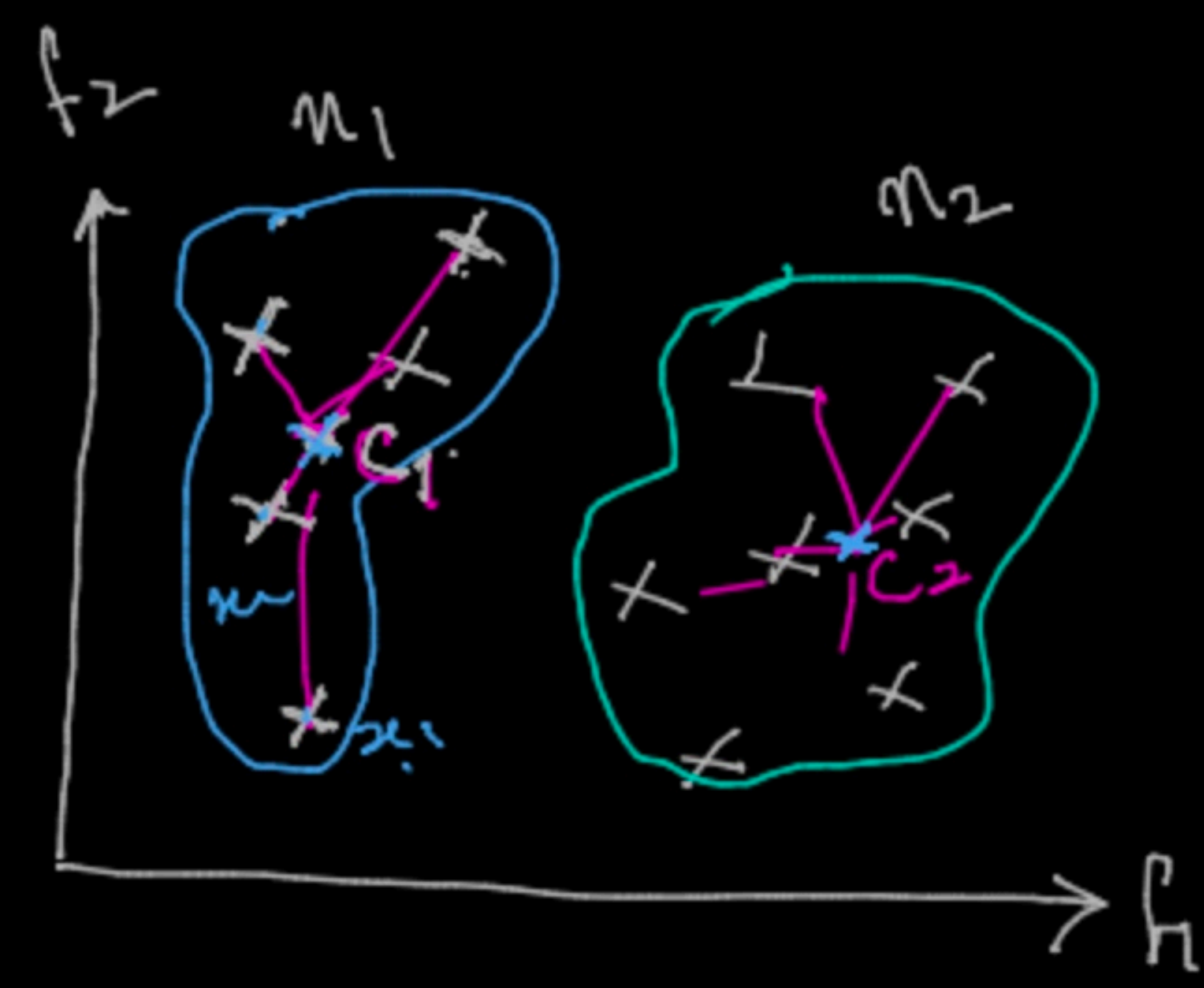
P_1

P_2

P_3

Gower's dissimilarity Index

7/7

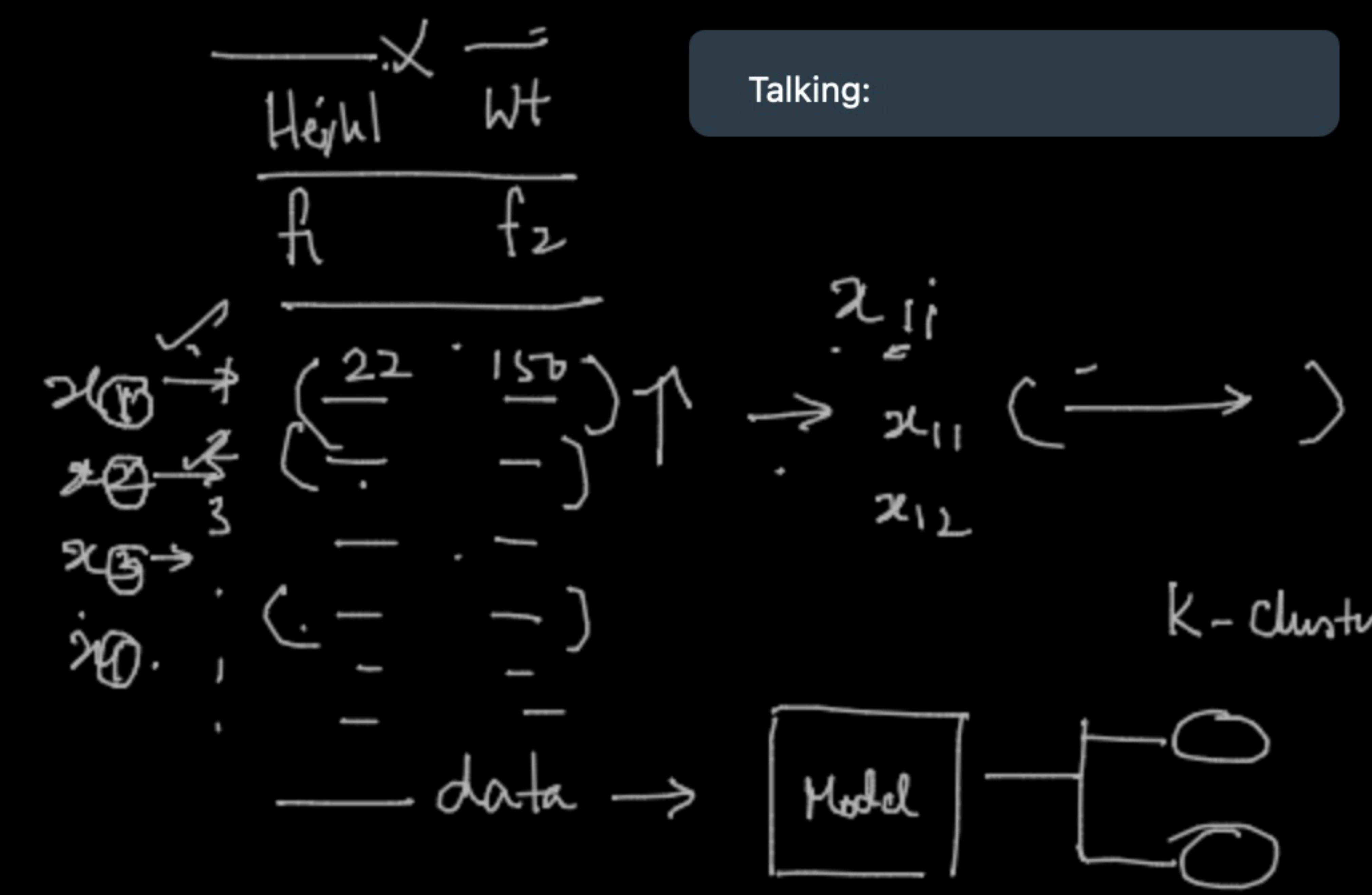


K-Means Algorithm

↳ No. of clusters.

— "Hyperparameter"

We decide what value to specify



$k \rightarrow$ No. of clusters

Task → Find 'k' centroids s.t the WCSS is minimum

$$WCSS = \underbrace{\sum_{i=1}^{n_1} (x_{1i} - c_1)^2}_{WCSS_1} + \sum_{i=1}^{n_2} (x_{2i} - c_2)^2 + \sum_{i=1}^{n_3} (x_{3i} - c_3)^2 + \cdots \sum_{i=1}^{n_k} (x_{ki} - c_k)^2$$

$$WCSS\downarrow = \sum_{j=1}^{K} \sum_{i=1}^{n_j} (x_{ji} - c_j)^2 \Rightarrow$$

'NP Hard Problems'

Approximation → Lloyd's Approximation

$$= WCSS.$$