Agenda

→ Bias Variance of KNN

→ Time & Space Complexity

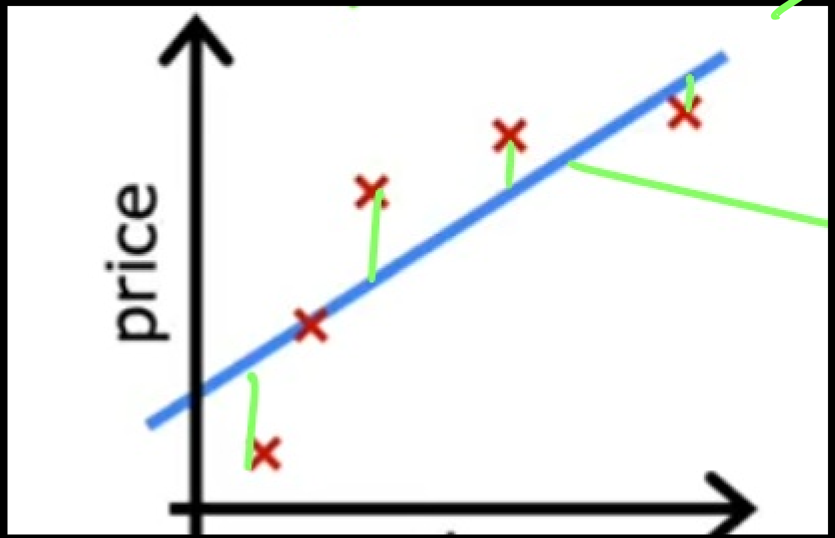
→ KNN Imputation

→ A : Underfitting

B : Overfitting

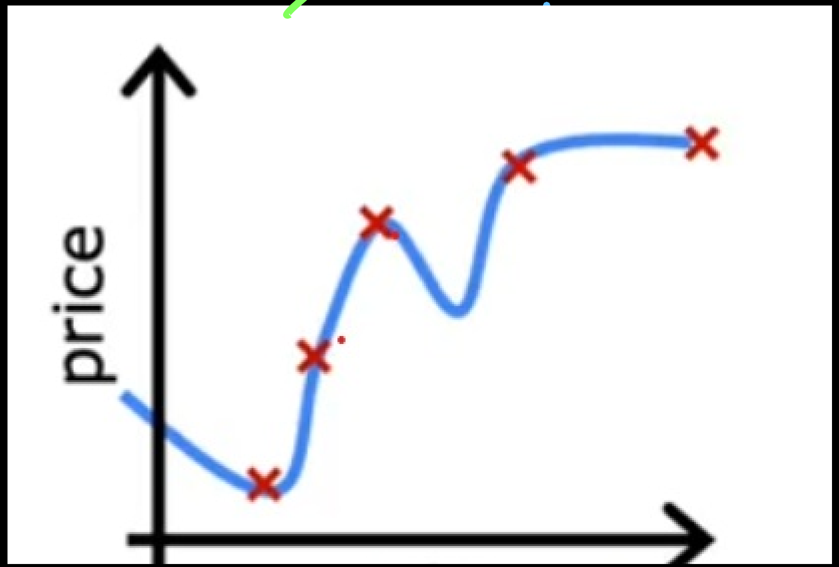
$$y = w_1 x + w_0$$

$$y = w_1 x + w_{12} x^2 + w_{13} x^3 + w_{14} x^4 + w_0$$



(1)

→ High bias & low variance

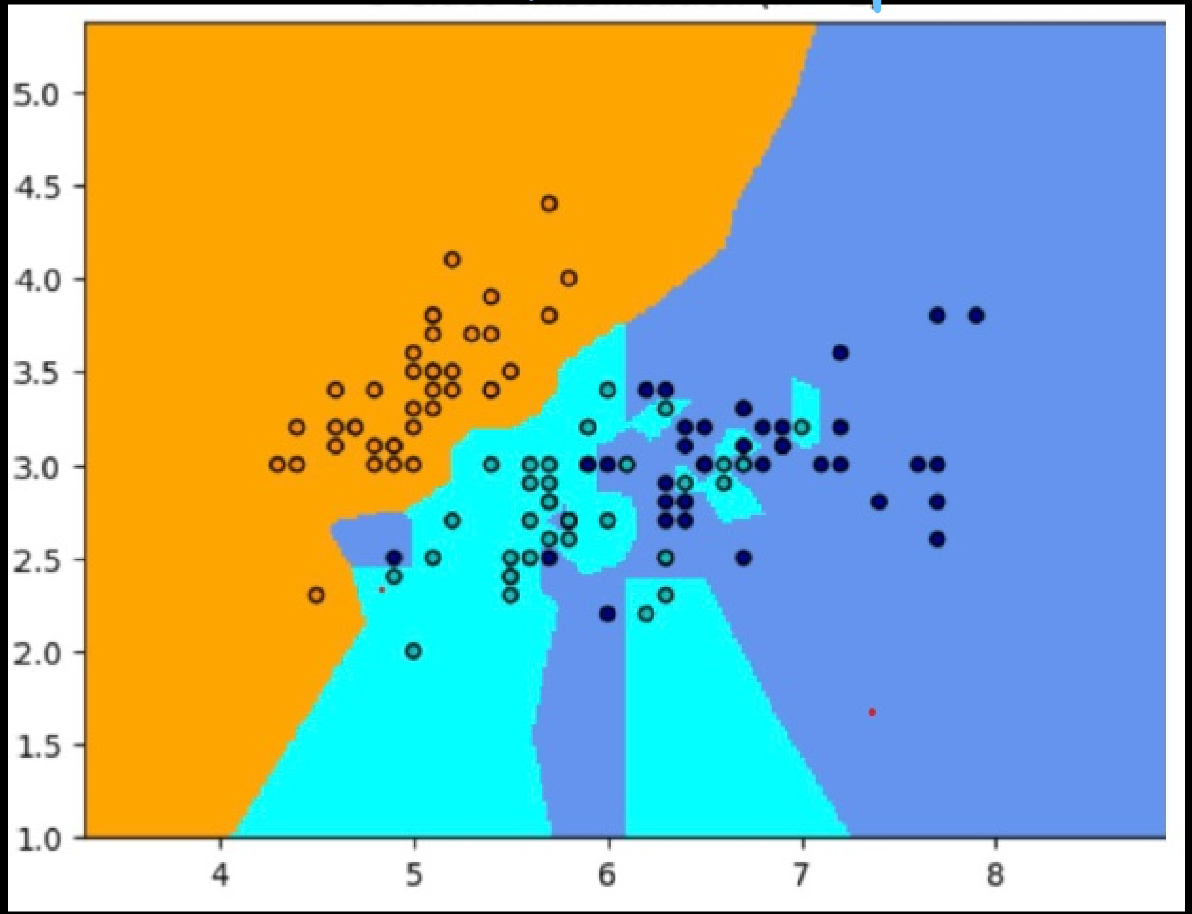


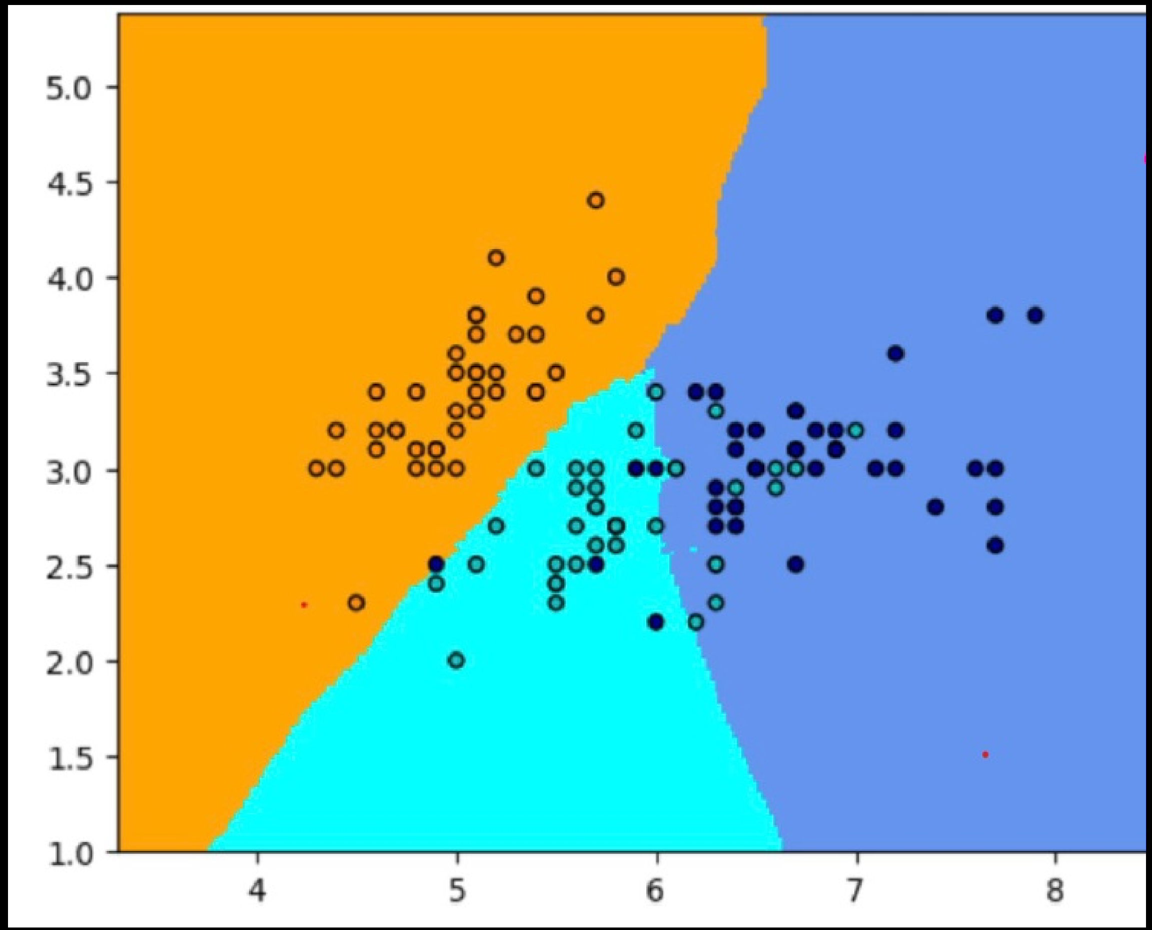
(2)

→ low bias & High variance

Decision boundary is more complicated

Decision boundary is simple





(1) Overfitting

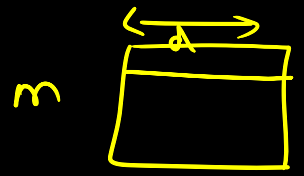
(2) Underfittin

K = 1

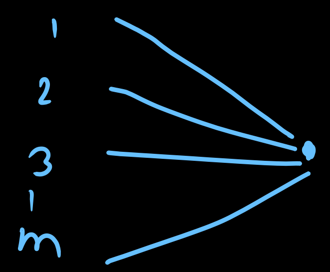
K = 15

K-NN

Algorithm



x_q



1. Compute the dist. of test point x_q from every training point → $O(md)$
2. Take the top-K Sort closest distance → $O(m \log m)$
3. For classification → Find the majority with K class

   ↳ $O(K)$

↳ $O(md + m\log m + \cancel{K})$

↳ $\boxed{O(md + m\log m)}$ ← For each test sample

$O(nmd + nm\log m)$

At test time

$$\hat{y}^{(i)} = \sigma\left(z^{(i)}\right)$$

Single test sample

$$= \sigma\left(w x^{(i)} + w_0\right)$$

$$\boxed{w_1 x_1^{(i)} + w_2 x_2^{(i)} - - w_d x_d^{(i)} + w_0}$$

$O(d)$ ← logistic

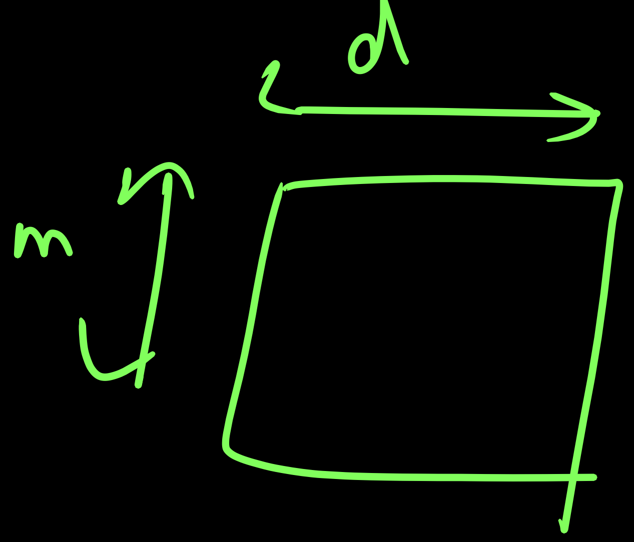
$O(md + m\log m)$ ← KNN

1) Logistic — $O(dn)$

2) KNN — $O((md + m \log m) \cdot n)$

Space Complexity

↳ $O(md)$



Categorical data

→ Label Encoding

↳ Ordinal data

Eg: High School, Madyamik, Graduate, Post graduate

→ OHE

→ Nominal data

Eg: Apple, Orange

↳ In 2 categories

M	0
F	1

One hot

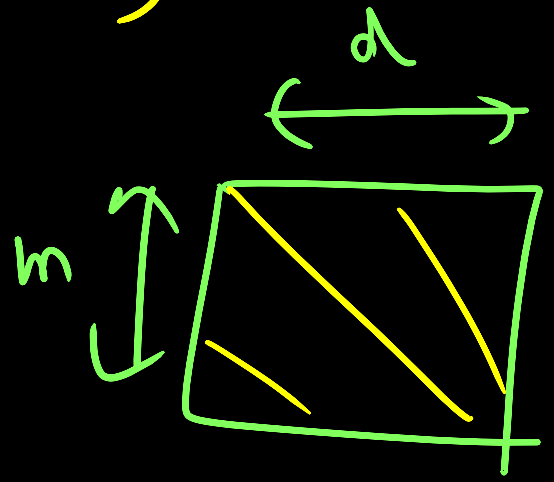
Blood group	A	B	AB	O
A	1	0	0	0
B	0	1	0	0
AB	0	0	1	0
O	0	0	0	1

Time & Space Complexity

Training time

Time $\rightarrow O(1)$

Space $\rightarrow O(md)$



Testing time

Distance

$x^{(1)}$ $x^{(2)}$

1> Manhattan (L1 norm) $\rightarrow \sum_{i=1}^{d} \left| x_i^{(1)} - x_i^{(2)} \right|$

2> Euclidean (L2 norm) $\rightarrow \left[\sum_{i=1}^{d} \left(x_i^{(1)} - x_i^{(2)} \right)^2 \right]^{1/2}$

3> Minkowski

$$\left[\sum_{i=1}^{d} \left| x_i^{(1)} - x_i^{(2)} \right|^p \right]^{1/p}$$

4> Cosine Similarity

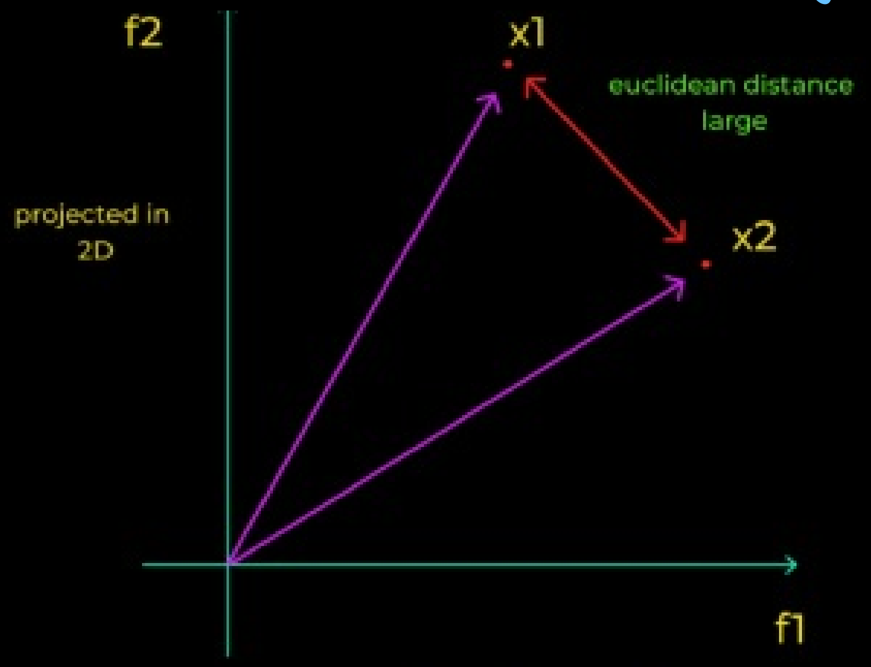
$p = 1 \longrightarrow$ Manhattan

$p = 2 \longrightarrow$ Euclidean

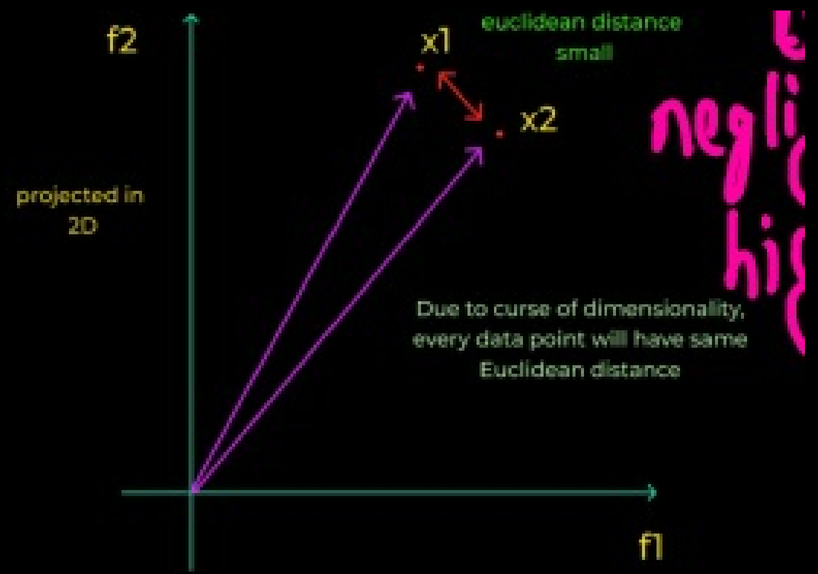
$$\text{Cosine similarity}\left(x^{(1)}, x^{(2)}\right) = \frac{x^{(1)} \cdot x^{(2)}}{\|x^{(1)}\| \, \|x^{(2)}\|}$$

$\hookrightarrow (-1, 1)$

Euclidean → Cannot be used with high dimension data



Due to low dimension, Euclidean distance between x1 & x2 is very large



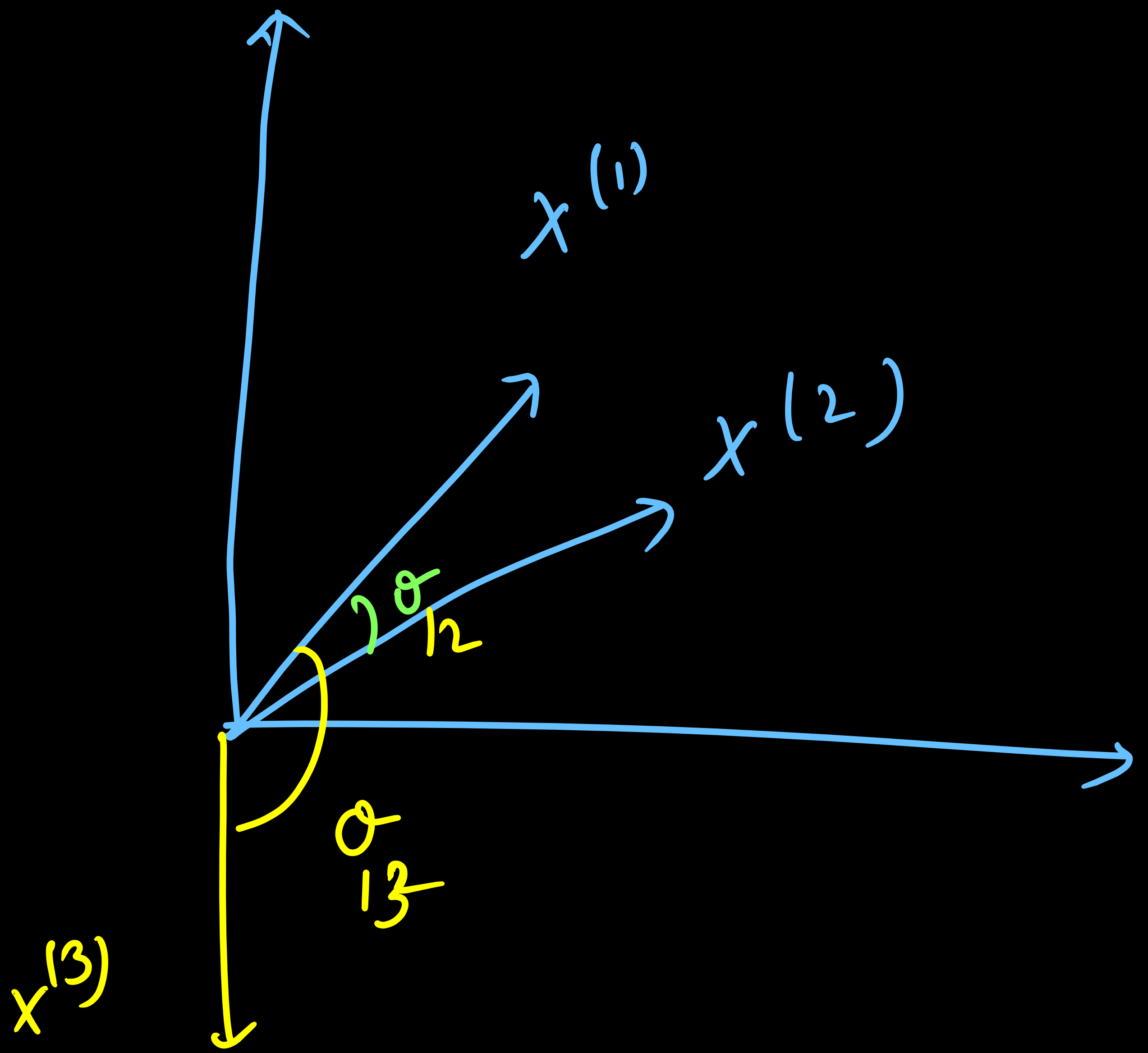
Euclidean distance cannot be used

Due to high dimension, Euclidean distance between x1 & x2 is very small

Conclusion : Euclidean distance fails when dimension is high

Cosine Similarity $\rightarrow$ Very well for high dimension data

↳ NLP



→ Imputation

33

29 34

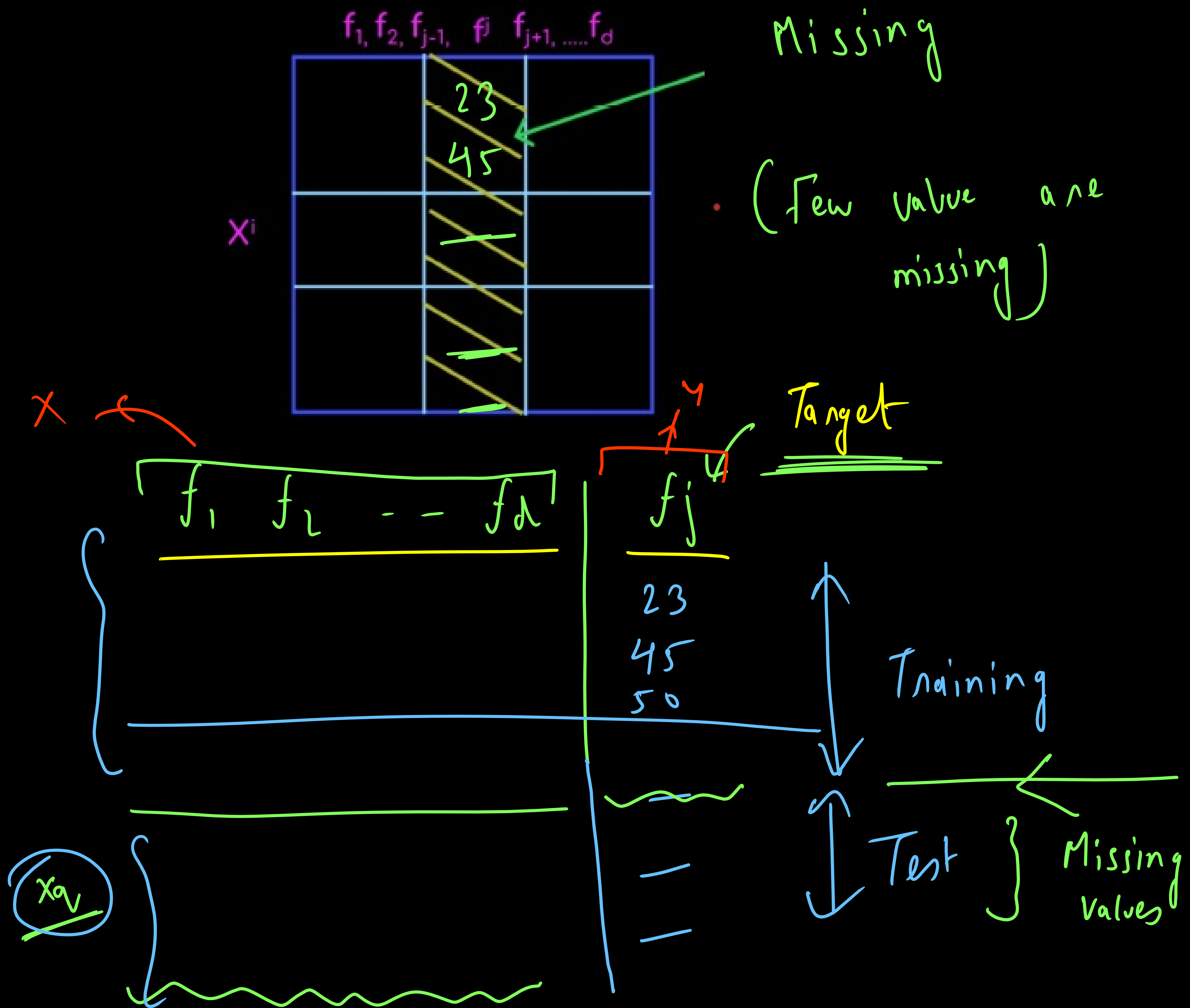
Gender	Class	Age	Fruits
M	L		
F	H	—	
M	Med		
F			
M		—	
F		—	

→ Mean or median

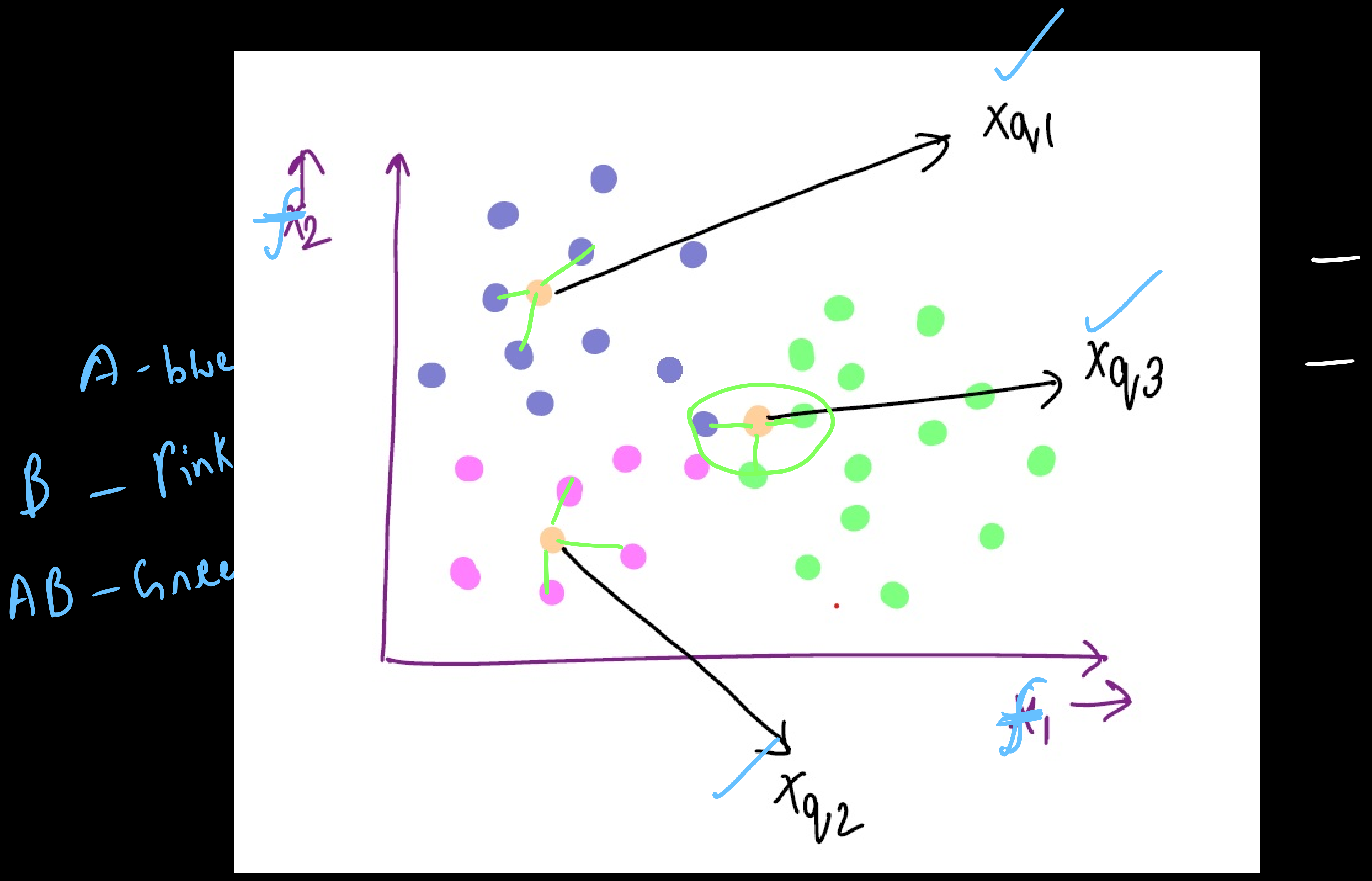
→ Entire data median

→ Female median Age } Grouped median
Male median Age

M	45
F	—
M	50
✓ F	23
M	55
✓ F	35
M	—

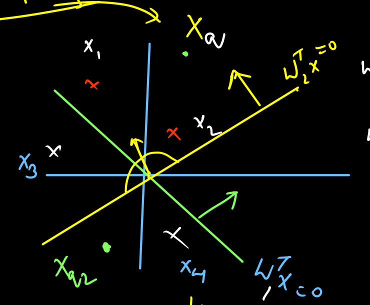


f1, f2, fj-1, fj fj+1, ... fd
23
45
Xi
Missing
(Few value are missing)
X
Y
Target
f1 f2 - - fd
fj
23
45
50
Training
Test
Missing values
xq



f2
A - blue
B - pink
AB - Green
xq1
xq3
xq2
f1

test point → LSH (Locality Sensitive Hashing) (p d)



$W^T x_1 > 0$

$W^T x_2 > 0$

	W_1	W_2
$x_q \to$	1	1

$W_1 x_q$ ↑ $W_2 x_q$ ↑

x_{q2}	0	0

	W_1	W_2
$x_1 \to$	$W_1^T x_1 > 0$	$W_2^T x_1 > 0$
$x_2 \to$	$W_1^T x_2 > 0$	$W_2^T x_2 > 0$
$x_3 \to$	$W_1^T x_3 < 0$	$W_2^T x_3 > 0$
$x_4 \to$	$W_1^T x_4 < 0$	$W_2^T x_4 < 0$

	W_1	W_2
x_1	1	1
x_2	1	1
x_3	0	1
x_4	0	0

$W^T x > 0 \to 1$

$W^T x < 0 \to 0$

hashmap Key		Value
(1, 1)	→	$x_1, x_2, \ldots, x_{10}, x_{20}, x_{50}$
(0, 1)	→	x_3
(0, 0)	→	x_4
(1, 0)	→	x_5

K

$(1,1) \to x_q$

$(0,0) \leftarrow x_{q2}$

2^2 $\boxed{2^p}$

$O(md + m \log m)$ $\boxed{O(n'd)}$ $+ O(pd)$

$O(md)$
↓
All training data

→ SCANN
→ FAISS

→ Hash table

Key Value

{ ______ : ______

}